CERC2016

Collaborative European Research Conference

Cork Institute of Technology – Cork, Ireland

23 - 24 September 2016

www.cerc–conference.eu

Proceedings

Extended - Abstracts and selected Full Papers

Version 1 - last changes 23. September 2016

Editors :
Udo Bleimann
Bernhard Humm
Robert Loew
Ingo Stengel
Paul Walsh

ISSN : 2220 - 4164

Hints :
- Topics in the table of content are linked to page where that abstract begins.
- The page number at the end of all pages is linked to begin of table of content.

Table of Content

Keynote »Standards or Excellence in Project Management ?« 12
Prof. Dr. Nino Grau

Extended Abstracts

Chapter 1
Business

Suitable Financing in the Context of a High-Tech (Semiconductor) Start-up 14
Mahbub, Maaskant, Wright

A Port-Centric Logistic Supply Chain for Container Traffic: The Case for the Port of Cork Company? 15
O'Callaghan, Wright

Galloping Technology: What risks? The Customer Service & Engagement Dynamic. 16
Ryan W., Wright

Chapter 2
Business Information Systems

The Application of Real Time Analytics Technologies for the Predictive Maintenance of Industrial Facilities in Internet of Things (IoT) Environments 20
Rieger, Regier, Stengel, Clarke

Practical Relevance and Operational Scenarios of Semantic Wikis 26
Coym, Ertelt, Bleimann

Industry 4.0 national and international 30
Ehlers, Bleimann, Loew

Chapter 3

Applied Computing

Prototype proof of concept for a mobile agitation tracking system for use in elderly and dementia care use cases 34
Healy, Keary, Walsh

FindR-TB: A cloud–based tool for antibiotic resistance prediction in Mycobacterium tuberculosis 38
Walsh, Mac Aogáin, Lawlor

Towards lifecycle management of clinical health care environments 40
Kowohl, Engel, Walsh, Keary, Fuchs, Hemmje

Eye-Tracking in Computer-Based Simulation in Healthcare Training 43
Currie, Bond, McCullagh, Black, Finlay

Chapter 4

Computing

Harassment detection: a benchmark on the #HackHarassment dataset 50
Bastidas, Dixon, Loo, Ryan J.

Large-Scale Biomedical Data Integration and Data Mining: a Multiplex Network-based Approach 54
Wang, Zheng

Towards a Cross Industry Standard Process for Big Data Supporting Big Data Applications in Virtual Research Environments 56
Berwind, Bornschlegl, Hemmje, Kaufmann

Social Network Support for an Applied Gaming Knowledge Management Ecosystem 60
Salman, Becker, Fuchs, Heutelbeck, Hemmje

Improved Data Centre Network Management Through Software-Defined Networking 62
Sherwin, Sreenan

Towards Synchronizing Heterogeneous Data Sources and Information Visualizations 65
Danowski, Bornschlegl, Schmidt, Hemmje

Comparison of Machine Learning Algorithms in Classifying Segmented Photographs of Food for Food Logging 68
McAllister, Zheng, Bond, Moorhead

Chapter 5

Bioinformatics

Towards a molecular diagnostic for bacterial infection in newborn babies 72
Forster, Dantoft, Kropp, Dickinson, Smith, Walsh, Ghazal

Metamorphosis: Changing the shape of genomic data using Kafka 75
Lawlor, Walsh

A Generalized Multi-Network Framework for Disease Gene Progression Prioritization 77
Browne, Wang, Zheng, Lópezac, Noriega

Measuring the Matches: Genome Alignment Benchmarking 80
Lu, Michel, Baranov, Walsh

The Moorepark Grass Growth model, an user-friendly model for farmer and researcher 87
Ruelle, Hennessy, Shalloo, Walsh, Manning, Wietrich, Delaby

An mHealth platform for supporting self-management of chronic conditions 89
Zheng, Patterson, Nugent, McCullagh, Donnelly, Cleland, McDonough, Boyd, Black

Development of a Bioinformatics Pipeline to Identify Biomarkers of Infection 92
Mullins, Konstantinidou, Walsh

An approach for anonymization of sensitive clinical and genetic data based on Data Cube Structures 96
Ntalaperas

Chapter 6

Humanities / Social

Designing narrative artefacts - the lives of women in eighteenth century Cork 102
Dempsey, Green, Wilson

Exploring new collaborative approaches in engagement and education towards the creation and enhancement of future STEM careers. 107
Delaney, Vakaloudis, Green, Hogan, Kameas, Zaharakis

Workplace Well-Being: Mental and Psychological Health and Safety of Non-Unionised and Unionised Employees 111
O'Mahony, Wright

Gender, Entrepreneurial Self-Efficacy, and Entrepreneurial Career Intentions: Implications of Female Role Models for Entrepreneurship Education. 113
O'Brien

Playing in the museum: shaping a definition of digitally mediated playful experiences 116
Heffernan, Delaney, Green

Meeting in the Middle: A collaborative and multidisciplinary approach to developing an empowerment course for women survivors of gender-based violence. 120
Davis, Kenny

A Serious Game to Dissuade Adolescents from Taking Performance Enhancing Drugs 123
Cunningham

A User Engagement framework for Universal Inclusion 126
Vakaloudis, Magennis, Hayes, O'Flynn, Delaney

Building capacity when assessing non-formal and informal learning through collaborative networks for Recognition of Prior Learning 129
O'Leary, Ledwith

Selected Full Papers

Chapter 7

Applied Computing

A Clinical Decision Support System for Personalized Medicine 132
Idelhauser, Humm, Beez, Walsh

Chapter 8

Bioinformatics

MetaPlat: A Cloud based Platform for Analysis and Visualisation of Metagenomics 148
Walsh, Lu, Konstantinidou, Bekaert, Zheng, Wang, Browne

List of Authors

Baranov	80
Becker	60
Bekaert	148
Black	43, 89
Bond	43, 68
Boyd	89
Clarke	20
Coym	26
Currie	43
Dantoft	72
Delaby	87
Dempsey	102
Dixon	50
Ehlers	30
Ertelt	26
Forster	72
Ghazal	72
Green	102, 107, 116
Healy	34
Hemmje	40, 56, 60, 65
Heutelbeck	60
Humm	132
Kameas	107
Keary	34, 40
Konstantinidou	92, 148
Kropp	72
Ledwith	129
Loo	50
Lópezac	77
Mac Aogáin	38
Mahbub	14
McAllister	68
McDonough	89
Moorhead	68
Noriega	77
Nugent	89
O’Callaghan	15
O’Leary	129
Patterson	89
Bastidas	50
Beez	132
Berwind	56
Bleimann	26, 30
Bornschlegl	56, 65
Browne	77, 148
Cleland	89
Cunningham	123
Danowski	65
Davis	120
Delaney	107, 116, 126
Dickinson	72
Donnelly	89
Engel	40
Finlay	43
Fuchs	40, 60
Grau	12
Hayes	126
Heffernan	116
Hennessy	87
Hogan	107
Idelhauser	132
Kaufmann	56
Kenny	120
Kowohl	40
Lawlor	38, 75
Loew	30
Lu	80, 148
Maaskant	14
Magennis	126
Manning	87
McCullagh	43, 89
Michel	80
Mullins	92
Ntalaperas	96
O’Brien	113
O’Flynn	126
O’Mahony	111
Regier	20

Rieger	20
Ryan J.	50
Salman	60
Shalloo	87
Smith	72
Stengel	20
Walsh	34, 38, 40, 72, 75, 80, 87, 92, 132, 148
Wietrich	87
Wright	14, 15, 16, 111
Zheng	54, 68, 77, 89, 148
Ruelle	87
Ryan W.	16
Schmidt	65
Sherwin	62
Sreenan	62
Vakaloudis	107, 126
Wang	54, 77, 148
Wilson	102
Zaharakis	107

Preface

In today's world, which has recently seen fractures and isolation forming among states, international and interdisciplinary collaboration has become an important source of progress. Collaboration is a rich source of innovation and growth and it is the goal of the Collaborative European Research Conference (CERC 2016) to foster collaboration among friends and colleagues across disciplines and nations within Europe. CERC emerged from a long-standing cooperation between the Cork Institute of Technology, Ireland and Darmstadt University of Applied Sciences, Germany. CERC has grown to include more well-established partners in Germany (Hochschule Karlsruhe and FernuniversitÃ¤t Hagen), United Kingdom, Greece, Spain, Italy, and many more.

CERC is truly interdisciplinary, bringing together new and experienced researchers from science, engineering, business, humanities, and the arts. At CERC researchers not only present their findings as published in their research papers. They are also challenged to collaboratively work out joint aspects of their research during conference sessions and informal social events and gatherings.

To organize such an event involves the hard work of a number of people. Thanks go to the international program committee and my fellow program chairs, particularly to Prof Dr Udo Bleimann and Prof Dr Ingo Stengel for organizing the conference and the review process. Dr Robert Loew put a great effort into preparing the conference programme and proceedings. Prof Udo Bleimann was invaluable for local organization. Thanks also to Dr Brendan Murphy and Tim Horgan for supporting CERC.

Dr Paul Walsh

General Conference & Programme Co-Chair, CERC2016

Cork, September 2016

Keynote
Standards or Excellence in Project Management ?

Prof. Dr. Nino Grau
Technische Hochschule Mittelhessen - University of Applied Sciences, Germany
e-mail: Nino.Grau@wi.thm.de

How do standards for project management support project managers in their everyday work? Are they critical success factors or do they hinder project manager`s creativity on the way to excellence in project management?

Let us have a look on standards like ISO 21 500 and on IPMA PEM Project Excellence Modell.

Short CV Prof. Dr. Nino Grau

He holds degrees in computer science (University of Erlangen) and industrial engineering (Technical University Munich) and received his doctorate with a thesis about decision making in teams.

His main area of interest is project and process management. He was a member of the board and deputy president of GPM the German Project Manager Association. GPM is MA (Member Association) of IPMA the International Project Management Association. IPMA is the oldest international umbrella association for project management associations and repre-sents over 60 MAs with over 60 000 individual members Worldwide.

As a Vice President of IPMA Grau started the IPMA activities concerning the development of the new ISO 21500 standard and has been Head of Delegation representing IPMA at ISO. He was responsible for the IPMA “Young Crew” and “Standards and Awards” within IPMA.

As a professor at the Technische Hochschule Mittelhessen University of Applied Sciences in Friedberg he has been responsible for introducing the first postgraduate degree course in project management in Germany, which started in September 2002.

Chapter 1

Business

Suitable Financing in the Context of a High-Tech (Semiconductor) Start-up

Mahbub Akhter, Pleun Maaskant and Angela Wright*
Tyndall National Institute, University College Cork, UCC, Cork Institute of Technology, CIT*
Corresponding author: mahbub.akhter@tyndall.ie

Keywords: Start-up, Entrepreneur, Venture Capital.

This research investigates the financing options available to a "High-Tech Start-up" when a technical person seeks to start a business from innovations emerging from his or her daily work. In seeking to understand the different factors which might influence the choice of a particular type of financial source, a qualitative research approach was considered to be the most appropriate one. As s part of this qualitative approach, a number of interviews with entrepreneurs, CEOs and CTOs have been conducted. Findings from these people with financial experiences are studied, analysed and compared, to select a suitable financing option for a photonics start-up originating from the work inside a research institute.

Considering the high amount of capital required to start a semiconductor start-up, private equity based Venture Capitalist (VC)s are thought to be the most appropriate funding source for such a type of high-tech start-up. This research explores the different aspects of VC funding and the creation of favourable environments to attract more VC investment in the area of semiconductor start-ups. It is found that, the VCs bring much more to the table than merely finance. The role of the VCs in the start-up extends way beyond the traditional financial aspect, with their experience and expertise in different areas, they become deeply involved with the managing and mentoring of the firms they finance. With all their positive influences, the Venture Capitalists contribute significantly to the success of a high-tech start-up. As Filipov (2011) states that some presently highly successful companies such as Apple Computers, Cisco Systems, Genentech, Intel, Microsoft, Netscape, and Sun Microsystems were initially financed by venture capital organisations (Filipov, 2011: 5).

It is concluded that VCs are not the only solution for financing a high-tech start-up throughout all its stages of development. Also, the interaction between entrepreneurs and venture capitalists is not immune to conflicts; both sides need to promote mutual understanding in order to bring success to the "high tech start-up". The role of the support infrastructure and the formation of the proper team in terms of required qualifications and experience has also been studied.

References

Filipov, G., N. 2011, 'Does Venture Capital Contribute to the Success of Startups?' B. Sc. Thesis, Erasmus University, Holland

A Port-Centric Logistic Supply Chain for Container Traffic: The Case for the Port of Cork Company?

Captain Kevin O'Callaghan, Dr Angela Wright
Department of OPD
CIT – Cork Institute of Technology, Ireland
e-mail: angela.wright@cit.ie

Keywords: Port-Centric Logistics, Ports, Cluster Containerisation, Shipping

As a consequence of globalisation, port performance and integrated port logistics have become increasingly important. Ireland as an Island nation depends on ports for the import and export of 99% of its goods. The only way that Ireland can remain competitive within worldwide markets is to have an efficient logistics supply chain. This collaborative research was undertaken in conjunction with the port of Cork and CIT, and seeks to examine the feasibility of introducing a port-centric logistics supply chain for containers in the Port of Cork. A port-centric logistics supply chain is a futuristic logistics process, whereby the port becomes a logistics hub rather than the traditional role as a simple gateway pier for the transfer of cargo. Currently, there is a dearth of academic research literature examining the case for port-centric logistics in Ireland.

A post-positivistic qualitative research methodology is applied in this new research to investigate the feasibility of a port-centric logistics supply chain for containers in the Port of Cork. This approach is considered the best one to address the aims and objectives of the study. The empirical data was gathered by undertaking semi-structured in-depth interviews with port customers, port management, logistics academics, and a senior government official. In total 10 interviews were undertaken.

Positive findings from this research indicate that both industry and the Port of Cork itself support the implementation of a port-centric logistics hub. Another key finding is that there is the potential for a future partnership between the Port of Cork and Cork International Airport, potentially providing a major boost for the Munster area and a significant collaboration for the future of the region. This research has also identified that the implementation of a port-centric logistics hub will reduce the national carbon footprint.

The detailed findings of this research will support the case for the future implementation of a port-centric logistics hub in the Port of Cork and future collaborations.

Galloping Technology: What risks? The Customer Service & Engagement Dynamic.

William Ryan, Dr Angela Wright
Department of OPD
CIT – Cork Institute of Technology, Ireland
e-mail: angela.wright@cit.ie

Keywords: Customer Services, Customer Engagement Strategy, Customer Loyalty, SMEs, Customer Management, Technology

Traditionally, customer service was very 'mechanistic' but, with the development of customer engagement, a transformation occurred; companies recognize the value of interacting with their customers, and customers need and want interaction. This interaction enables organisations to understand their customers, learn more about their needs and what may be required to satisfy them. The aim of this process is to achieve and enable increased customer loyalty. The result is that organisations are engaging with their customers in a timely fashion, and this change has been facilitated by some major advances in technology, namely the advancement of the internet, social media, and the arrival of many and varied new mobile communication channels and platforms.

A modern reality is that internet usage is continuing to expand rapidly, with global internet usage increasing from 29.2% in 2010, to 40.7% in 2014, while usage in Ireland has increased from 69.9% in 2010, to 79.7% in 2014 (World Bank, 2015). This change is also being replicated in the case of mobile phone usage, with three-quarters of the world's inhabitants now having access to a mobile phone. The number of mobile subscriptions in use worldwide, both pre-paid and post-paid, has grown from fewer than 1 billion in 2000 to over 6 billion in 2012, of which nearly 5 billion is in developing countries (World Bank, 2012).

The mobile phone usage figures are proof that the customer landscape has been altered, and with this change comes vast opportunity. Organisations are now embracing these changes, and having effective, timely two-way relationships that were not possible in the past. These new interactions are delivering immense benefits to companies in terms of customer loyalty, product development, and increased sales opportunities. On a more cautious note, these advancements also come with new risks and new challenges for organisations in the form of negatively critical word of mouth. Customers now have the power to highlight poor performance or poor quality to a mass audience through social media, blogs, and forums, with immediate effect. Customers are now becoming more empowered, and are putting the basics of traditional business to the test. Current changes in customer services, and an organisation's ability to successfully employ customer engagement, will largely determine to what degree a company is successful. Consequently, the researchers consider that this is an opportune time to research dynamic field.

This research study is an exploratory examination of customer service in an Irish context. This research will endeavour to evaluate the impact of customer service on corporate strategy and on an organisation's performance in the context of this macro world. The research seeks to determine if technological change and advancements have impacted on customer services and, especially, customer engagement, causing it to evolve, and to examine the effect, if any, these changes have had on businesses. The research explores the concept of 'loyal customers', the availability of 'best

practice' models to aid organisations employing these concepts within the context of customer services, and to evaluate if components of customer loyalty models are transferrable from large organisations to small medium enterprises.

This research on customer services focuses specifically on the concepts and developments of customer engagement, within an Irish perspective. Adopting an interpretivist's paradigm, a qualitative research process was applied involving semi-structured face to face interviews with ten senior experts on customer services and customer engagement, including senior management of global organisations, chief executive officers, experts with worldwide customer services experience, company directors, and individuals from diverse commercial and sporting organisations. These leaders all operate in roles that have direct customer services interaction, and hold positions that have an influence over the strategy and the future direction of customer services within their organizations. Data analysis was carried out using grounded theory, and due consideration was given to ethical, reliability and validity issues. The resultant views, whilst varied, presented seven core themes that will be outlined in full in the paper.

Findings from this research reveal that customer services is having a significant impact on the corporate strategies of companies in today's dynamic technological business environment, and can be the difference between organisational success or failure. Technology has positively impacted customer engagement where it is now 'the lever' to gain closer relationships with customers. This research has identified a direct correlation between the success of a customer services strategy, and the implementation of an effective customer engagement program.

In the absence of a holistic customer engagement model for 'best practice' to assist organisations in implementing a customer engagement strategy, this study has subsequently developed a model to fill this gap in academic literature, and provide organisations with a model that is built in stages, allowing them to progress to 'best in class' status. This research had also identified transferrable components of the new model for practice to small and medium enterprises.

References

World Bank, (2015). "Internet Usage", available at, http://data.worldbank.org/indicator/IT.NET.USER.P2, retrieved January 26, 2016, 08.45.

World Bank, (2012). "Mobile Phone Access Reaches Three Quarters of Planet's Population", available at, http://www.worldbank.org/en/news/press-release/2012/07/17/mobile-phone-access-reaches-three-quarters-planets-population, retrieved January 26, 2016, 10.07.

Chapter 2

Business Information Systems

The Application of Real Time Analytics Technologies for the Predictive Maintenance of Industrial Facilities in Internet of Things (IoT) Environments. A Discussion on Appropriate Research Methodologies

Thomas Rieger †, Stefanie Regier ††, Ingo Stengel ††, Nathan Clarke †
† School of Computing and Mathematics, Plymouth University, United Kingdom
†† Karlsruhe University of Applied Sciences, Germany
e-mail: thomas.rieger@plymouth.ac.uk

Keywords: PhD Thesis, Research Methodology, Predictive Maintenance

Software systems for the analysis of large amounts of data (Big Data Analytics) are applied in more and more areas. The subject of Predictive Maintenance (PdM) using Big Data Analytics systems is present in industrial manufacturing processes, especially in combination with Internet of Things (IoT) environments. Until now, such systems have been using mainly established batch-oriented approaches. New technologies in the area of real time/streaming analytics open up entirely new technological opportunities (Gupta, et al., 2016). These new opportunities might lead to new and even better models in the area of PdM.

The project discussed here deals with remapping existing models for PdM in IoT environments, e.g. condition based or reliability based models, using the above-mentioned new technological opportunities, and with identifying new models and model combinations. New models need to be defined. A comparison of these with established models using big data will outline possible improvements.

The aim of this abstract is to illustrate the discussion that helps to find a suitable research methodology, since a correct scientific approach differs from a practical software developer approach significantly.

The abstract starts with a brief introduction to the subject area and the goals. By adapting the work packages to selected research methodologies it will then illustrate how to find the correct research methodology. The objective of this abstract is to raise the question if one of the methods demonstrated might be valid.

From an IT perspective, industrial automation systems are about managing systems and processes. Communication is focused on process-, production-, and product data. Until now storing the data for later analysis has not been a priority (Schaeffer, 2015).

Due to intelligent devices, more and more data is generated in industrial systems. Connecting these devices in IoT environments makes it possible to store and analyse this data. As a consequence the use of Big Data Analytics for storage and analysis makes sense. “Internet of Things (IoT) will comprise billions of devices that can sense, communicate, compute and potentially actuate. Data streams coming from these devices will challenge the traditional approaches to data management and contribute to the emerging paradigm of big data.” (Zaslavsky, et al., 2013).

A relevant field of application for the analysis of such data is Predictive Maintenance (PdM). The main purpose of PdM is to anticipate unscheduled outages of machines and plants. Maintenance

costs have the potential to be reduced. The availability of production capacities will be more predictable (Eoda, 2014).

New technologies in the area of real time/streaming analytics make it possible to process large amounts of data in real time. Immediate results, tendencies and trends are thus available. Continuous data streams can be permanently analysed.

Research methodology

After the topic of the research project was clear the following questions are about the research design and how to find the correct research methodology. Formulating a precise research question is an important first step. According to Shuttleworth (2015) a general research question will usually be based around 'why' or 'how' a certain phenomenon is happening.

The project discussed here deals with new technologies in the area of real time/streaming analytics. The focus lies on PdM in industrial processes. Large amounts of data which are generated at high frequencies are necessary for the effective use of Big Data Technologies. These amounts of data can be found in the industrial sector in IoT environments (Zaslavsky, et al., 2013). This leads to the research question "How can new technological opportunities in the area of real time/streaming analytics improve Big Data Applications for PdM in industrial IoT environments?"

When the research question is defined, the following question arises: How to answer the research question in a scientifically sound approach? Following the relevant literature scientific-theoretical approaches of quantitative, qualitative and mixed studies can be found. Applying the classification profile for study designs defined in literature (Döring, et al., 2016) two potential alternatives in the cases of the scientific-theoretical approach and the study groups can be considered (Figure 1, alternatives are on a grey background).

Qualitative or quantitative Approach

Table 1 gives an overview of the differences between qualitative and quantitative approaches (Xavier, 2012). It seems that the quantitative approach fits better to the described research topic than the qualitative approach.

To carry out a quantitative research approach research hypotheses have to be defined. An extensive analysis needs to be carried out using structured data collection methods. This will help to investigate the defined hypothesis. The collected quantitative (numerical) data needs to be analysed statistically (Döring, et al., 2016).

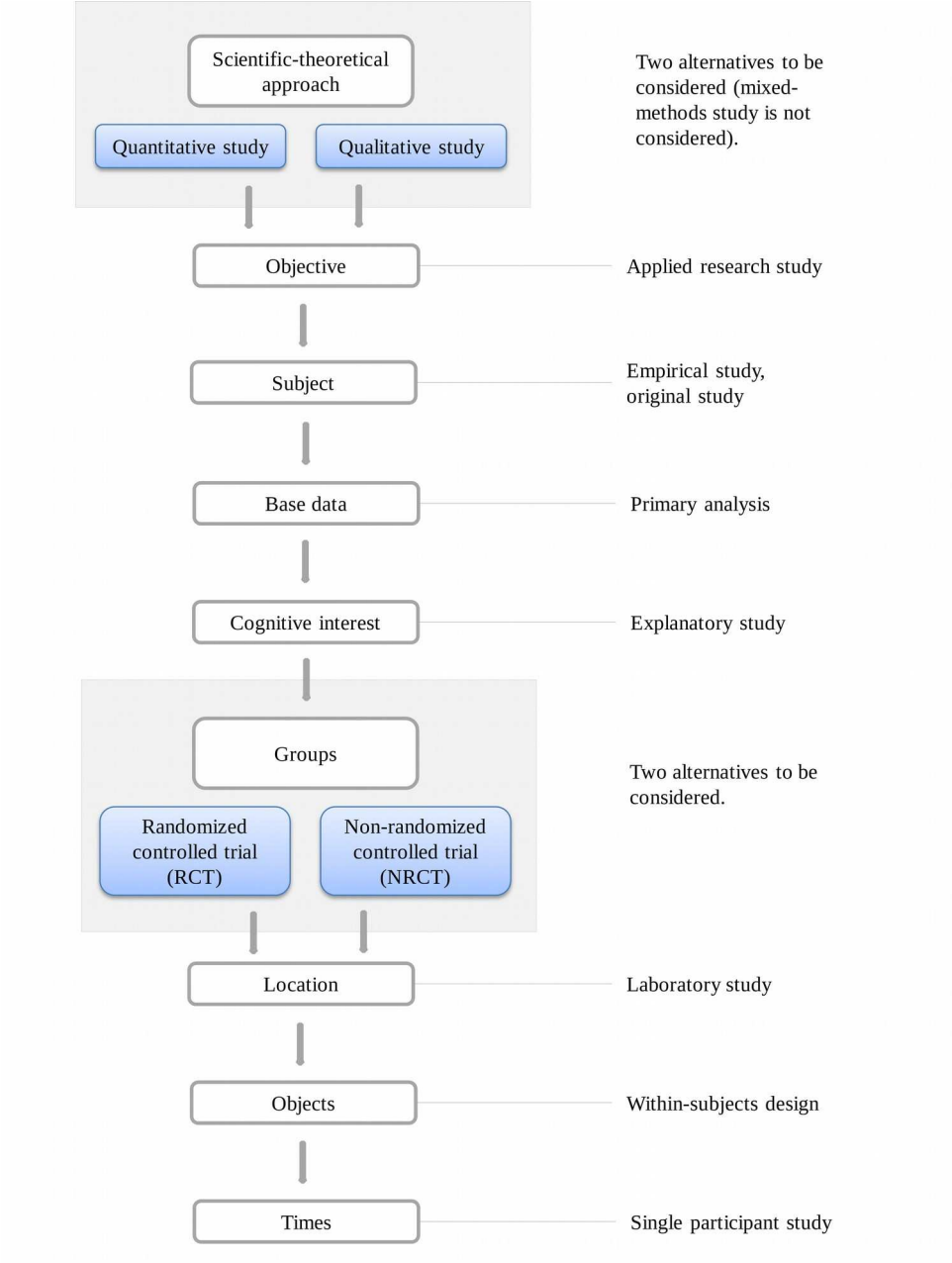


Figure 1: Study design approach according to (Döring, et al., 2016)

Criteria	Qualitative Research	Quantitative Research
Purpose	To understand & interpret social interactions.	To test hypotheses, look at cause & effect, & make predictions.
Group Studied	Smaller & not randomly selected.	Larger & randomly selected.
Variables	Study of the whole, not variables.	Specific variables studied
Type of Data Collected	Words, images, or objects.	Numbers and statistics.
Form of Data Collected	Qualitative data such as open- ended responses, interviews, participant observations, field notes, & reflections.	Quantitative data based on precise measurements using structured & validated data-collection instruments.
Type of Data Analysis	Identify patterns, features, themes.	Identify statistical relationships.
Objectivity and Subjectivity	Subjectivity is expected.	Objectivity is critical.
Role of Researcher	Researcher & their biases may be known to participants in the study, & participant characteristics may be known to the researcher.	Researcher & their biases are not known to participants in the study, & participant characteristics are deliberately hidden from the researcher (double blind studies).
Results	Particular or specialized findings that is less generalizable.	Generalizable findings that can be applied to other populations.
Scientific Method	Exploratory or bottom–up: the researcher generates a new hypothesis and theory from the data collected.	Confirmatory or top-down: the researcher tests the hypothesis and theory with the data.
View of Human Behavior	Dynamic, situational, social, & personal.	Regular & predictable.
Most Common Research Objectives	Explore, discover, & construct.	Describe, explain, & predict.
Focus	Wide-angle lens; examines the breadth & depth of phenomena.	Narrow-angle lens; tests a specific hypotheses.
Nature of Observation	Study behavior in a natural environment.	Study behavior under controlled conditions; isolate causal effects.
Nature of Reality	Multiple realities; subjective.	Single reality; objective.
Final Report	Narrative report with contextual description & direct quotations from research participants.	Statistical report with correlations, comparisons of means, & statistical significance of findings.

Table 1 - QUALITATIVE VERSUS QUANTITATIVE RESEARCH (Xavier, 2012)

An example for a hypothesis could be "This engine is going to fail within a time frame of [X] hours" or "This in-service engine will last [X] more days before it fails" or (comparing batch-oriented PdM to real-time PdM). "The downtime of a machine is reduced from [Y] to [X] hours using the same PdM-Modell/Algorithm with real-time approaches instead of batch-oriented approaches". It is thus necessary to make possible improvements measurable. But, how could valid measurability be achieved?

Industrial environments are diverse and have very different characteristics (e.g. heavy industry vs. pharmaceutical industry). The same is true for IoT systems in these industrial sectors. In order to perform valid measurements and evaluations, a precise application case and a precise environment have to be defined. Application cases in this project use different PdM models (e.g condition-based, model-based, reliability-centred, multi-threshold or continuous-time, mathematical models or empirical approaches). The decision was made to focus on a few selected models, such as remaining lifetime using sensor-based degradation models (Kaiser, et al., 2009).

As the project is mainly data-driven, the criterion for the selection of an environment is the availability of valid measurement data, including as much historical data as possible. There is a choice between generated data, open data repositories or real life data. Current research outlines that open data repositories like the Turbofan Engine Degradation Simulation Data Set (Saxena, et al., 2008) are a good choice because they normally contain comprehensive metadata in addition to existing massive outcome data. The metadata often includes occurred error events and additional information for validation, labelling and classification. The data sets can be used for development of prognostic algorithms (Saxena, et al., 2008). Just like real life data, data from an open data repository is only valid for a specific case of application. Thus, the whole project must be focused

on this application case exclusively. The title of the project and the research question have to be extended by the specific application case.

Once the application case and the environment have been defined, the question of measurability remains. For this purpose Key Performance Indicators (KPIs) will be applied. Even technological objectives, such as "faster availability of results" or "reduction of unscheduled outages" can be defined as KPIs. Ways to identify KPIs include the categorisation via success factors (Parmenter, 2015). Success factors could be, for example, the cost savings in EUR, or the reduction of unproductive time in hours. The following table (Table 2) shows an example for this:

CHARACTERISTIC	UNIT	MEASURE	TYPE
Maintenance cost	EURO	Reduction in %	quantitative
Machine damage	EURO	Reduction in %	quantitative
Spare part storage	EURO	Reduction in %	quantitative
Machine outage	Time	Reduction in %	quantitative
Overtime expenditure	EURO	Reduction in %	quantitative
Machine service life	Time	Increase in %	quantitative
Profits	EURO	Increase in %	quantitative

Table 2 - Success factors according to (Eoda, 2014)

Defining groups

The second open question according to Figure 1 is about the creation and treatment of analysed groups. There are three alternative approaches to be considered: randomized controlled trial (RCT), non-randomized controlled trial (NRCT), and non-experimental study (Döring, et al., 2016).

Non-experimental studies (or descriptive studies) only cover found variations and no experimental variations of independent variables. Non-experimental studies are only very conditionally suitable for the evaluation of a causality hypothesis (Döring, et al., 2016).

A real experiment occurs if randomization or random assignment of groups is performed. An experiment without randomization is referred as a quasi-experiment. If and how groups can be defined for the described project is a topic for discussions (Döring, et al., 2016).

Conclusions

Once the subject of a research project has been found, it is important to define the goal, the major research questions and the approach. This is indispensable in order to ensure a scientifically sound approach and the generating of valid results. The approach already has to be defined at the beginning of the project in order to avoid wrong turns in the course of work.

This abstract describes the journey towards finding a suitable approach for the research project. It has been demonstrated which approaches have been taken for this purpose and which issues still need to be clarified.

References

S. Gupta; S. Saxena, Packt Publishing Ltd. (2016). Realtime Big Data Analytics, First Edition February 2016

M. Shuttleworth, Explorable.com (2015), How to write a Research Paper, 2015

A. Zaslavsky, C. Perera, D. Georgakopoulos (Jan 2013), Sensing as a Service and Big Data, URL: https://arxiv.org/abs/1301.0159, last accessed 08/08/2016

K. A. Kaiser, N. Z. Gebraeel, Predictive Maintenance Management Using Sensor-Based Degradation Models, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, August 2009

Eoda GmbH, Whitepaper Predictive Maintenance (mit R), eoda, Februar 2014,
URL: https://www.eoda.de/files/Use_Case_Seiten/Whitepaper/Predictive_Maintenance_mit_R.pdf , last accessed 08/08/2016

N. Döring, J. Bortz, Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer, 5. Auflage 2016

A. Saxena and K. Goebel (2008). Turbofan Engine Degradation Simulation Data Set, NASA Ames Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA, URL: http://ti.arc.nasa.gov/project/prognostic-data-repository, last accessed 08/08/2016

C. Schaeffer, Big Data + The Internet of Things = Big Manufacturing Opportunity. Vantive Media CEO & Blogger, 2015 URL: http://www.crmsearch.com/internetofthings.php, last accessed 08/08/2016

D. Parmenter, Key Performance Indicators: Developing, Implementing, and Using Winning KPIs, John Wiley & Sons, Third Edition 2015

Xavier Unversitiy Library, QUALITATIVE VERSUS QUANTITATIVE RESEARCH, 2012
URL: http://www.xavier.edu/library/students/documents/qualitative_quantitative.pdf, last accessed 08/08/2016

Practical Relevance and Operational Scenarios of Semantic Wiki

Matthias Coym, Philipp Ertelt, Udo Bleimann
University of Applied Science Darmstadt, Department of Computer Science, Germany
e-mail: matthias.coym@h-da.de, philipp.ertelt@h-da.de

Keywords: knowledge, semantic, wiki

Introduction

To keep up with the top contestants on the market, knowledge management is a crucial topic for today's companies. Every organizations resolution is to improve the way it does business. Achieving this goal is strongly dependent on experience and knowledge of the company and its employees.

One approach of knowledge management in a business is the deployment of a wiki. A wiki is an online encyclopedia with articles or entries which can be accessed by the employees. The articles can provide information about different procedures or workflows in the company. Futhermore, it can contain information about customers, projects and experiences. A main fact of a wiki is that it is mostly used as an open encyclopedia. That means employees may be able to not just read, but create new and edit existing articles. This is an agile way to improve and extend the platform. It can be used to concentrate the knowledge in just one system to reduce redundancy and to avoid outdated information, always supposing that the platform is maintained regularly. New employees can use the existing information to have a better start, in contrast, the company keeps the knowledge of employees who leave the company if it is documented it in the wiki.

A wiki does not have to be set in a business case, but can also be used to make information about a certain area accessible to anyone. The most popular example is the free internet encyclopedia platform Wikipedia[1].

A way to extend a wiki platform is the usage of semantic. Semantic describes the relationship between words or phrases in a context. A subject can be connected to an object using a predicate, e.g., university has students. The idea to connect subjects and objects is to give information about relations which are maybe not visible on first sight. This information is called metadata and contain information about characteristics. A semantic network is a representation model of semantic. Figure 1 shows a semantic network where subjects and objects are linked. Each link has a description to define the relation between subjects and objects.

In a semantic wiki the relationships between articles can be filled with information about the data, i.e., it has an underlying knowledge model. The combination of semantic and a wiki can be valuable if used in the right way. This work is about the practical relevance and in which operational scenarios a semantic wiki is useful.

1 www.wikipedia.org

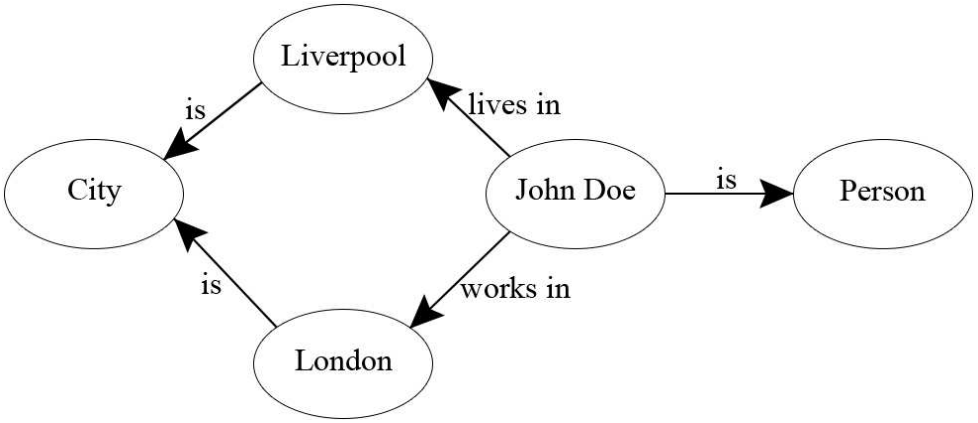


***Figure 1**: Example of a simple semantic network*

Related Research

Several semantic wikis have been developed in recent years. The semantic wiki systems can have very different purposes and functions like personal knowledge management. One of the first semantic wiki systems is the PlatypusWiki, an enhanced Wiki Wiki Web platform (Tazzoli et al., 2004). It uses the Resource Description Framework (RDF), a standard for metadata, and Web Ontology Language (OWL) vocabulary for knowledge representation. It is implemented in Java following the Model View Controller pattern. The focus of the Wiki is set to the creation of RDF data.

Völkel et al. developed an extension for the MediaWiki system called Semantic MediaWiki (Völkel et al., 2006). The extension provides the ability to semantically annotate articles in a wiki. It has also shown that the performance of a semantic wiki has to be considered, since a wiki can become slower with the amount of metadata in the wiki.

Salzburg Research developed a Semantic Wiki called IkeWiki (Schaffert, 2006). IkeWiki has been primarily developed as software for ontology engineering. The wiki data is stored in a Postgres database. It contains two types of content editors. A WYSIWYG editor for novice users as well as a traditional text editor.

Since the Web 2.0 is spread by blogs, wikis, image/audio/video sharing websites, the user behaviour has changed dramatically. Users want to share information and participate among other web users. Therefore, tools are necessary to help users finding the information they search for. Wikipedia has proven to be successful for information interchange. To extend the functionality of a Wiki, the concept of a semantic wiki was born, mixing the features of a wiki and the technologies of a semantic web. Buffa et al. have published in (Buffa, 2008) the SweetWiki. A semantic wiki that combines the advantages of a wiki with the semantic web. Buffa et al. have shown that the use of semantic web concepts can improve navigation, search and usability of a wiki.

In 2005, Krötzsch et al. present a technique to allow knowledge processing in a computer assisted way for Wikipedia (Krötzsch, 2005). The technique called typed links is a simple method for rendering large parts of the Wikipedia platform, so that it is machine readable. Using the technique the user can query intelligently the knowledge database and find related topics of the searched keyword. The method also impacts hardly usability and performance of the wiki.

As aforementioned, a wiki can have different purposes and functions. Oren et al. introduce the prototype of SemperWiki in (Oren, 2006). The wiki is especially for personal and organisational knowledge management. Traditional tools for personal knowledge managements like todo lists or files are very common. But these analogue methods are not automated and cannot be searched digitally. Hence, a digital tool with the requirements to save information, but also provide support finding and reminding information is needed. The advantages of wikis are that they are simple to use and collaborative access is easy to manage, but provide only very limited support for finding and reminding. Oren et al. show how a semantic wiki using metadata annotations improve information access by structured navigation and queries.

Another approach by Tobias Kuhn is the AceWiki (Kuhn, 2008). AceWiki is a prototype of a semantic wiki using the Attempto Controlled English (ACE) for representing its content. ACE is a subset of the English language with limited grammar and fixed formal semantics, i.e., ACE looks like English but has a very restricted syntax. The focus of AceWiki is to improve knowledge aggregation and information representation. The main difference compared to other semantic wikis is, it does not use annotations or metadata for formal statements, but they are the main content of the wiki. An evaluation shows that AceWiki is easy to learn and content can be added quickly by following the three design principles: naturalness, uniformity, and strict user guidance.

KawaWiki is a semantic wiki based on RDF templates and was introduced by Kawamoto et al. in 2006 (Kawamoto, 2006). Describing that for end users it is not easy to create a semantic wiki page to share information from scratch without knowledge of the RDF/OWL syntax, the KawaWiki provides RDF templates to support the author process. The RDF templates used by KawaWiki validate the consistency of the wiki and the created pages. The wiki can generate forms from templates to support the user creating a page and filling it with information. The developed RDF template system by Kawamoto et al. assists expert and non-expert users with a validation checking mechanism to guarantee that the consistency of the wiki is always given.

Methodology

To show how a semantic wiki might be beneficial used, we first tried to understand why semantic wikis are not often used. To do this we looked at the technological aspect of semantic wiki as well as the impact of semantic on the usage and maintenance of the wiki.

Practical Relevance and Operational Scenarios

The practical usage of a wiki system is depending on its purpose. A wiki can be used by a single person as personal knowledge management up to a knowledge management system of a big company to concentrate experience and knowledge in one system, or even a public system like Wikipedia accessible by anyone. A wiki system is mostly used to collaboratively collect and share information to a certain group of users.

A semantic wiki has the benefit over a normal wiki that the underlying semantic net is machine comprehensible. The added meta data enhances the search functionality and can be queried like a database to make use of otherwise unusable connections to automatically generate diagrams and lists.

There are several drawbacks to the use of semantic wiki as well.

There is a conflict between the collaboration aspect and “free for all” approach of a wiki and the care and focus needed for a semantic web. If each writer freely adds relations between pages, the resulting semantic net is supposable complex and contains redundancy. To avoid this a kind of higher administration and/or restriction needs to be set in place, which supposable reduces the motivation to add content or use the semantic features. The addition and maintenance of meta data is extra work for the writer and the use of queries to generate content automatically is not trivial. It is furthermore not enough to just add meta data and relations. A benefit is only achieved, if the semantic data can then be used for non-trivial concrete use cases. This means that not every possible relation is also beneficial to add and advance planning is needed.

Our assumption is therefore that a semantic wiki works best with a wiki of following criteria:

- The content of the wiki is about one specific topic
- Wiki pages correlate to specific objects
- There are multiple cases of the same semantic relations

Conclusion and Future Work

The effort to set up a semantic wiki is reasonable there are also several plug-ins to enhance an existing wiki with semantic. The extra data stored in a wiki can affect the runtime of the wiki, but overall the technical barrier can still be considered low.

A real barrier lies in the conception of beneficial use cases for the semantic data, in the overhead of maintenance for the writers and in the need of a global administration.

The next planned step is to verify our assumption and develop an approach to enhance an existing wiki with semantics based on a concrete use case.

References

Tazzoli, R., Castagna, P., Campanini, S.E.: Towards a Semantic WikiWikiWeb.In: 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan (2004)

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. (2006). Semantic Wikipedia. In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM, New York, NY, USA, 585-594.

Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management.(2006)

Buffa, Michel, et al. "SweetWiki: A semantic wiki." *Web Semantics: Science, Services and Agents on the World Wide Web* 6.1 (2008): 84-97.

Krötzsch, Markus, Denny Vr Denny Vrandecic, and Max Völkel. "Wikipedia and the semantic web-the missing links." *Proceedings of Wikimania 2005*. 2005.

Oren, Eyal, et al. Semantic wikis for personal knowledge management. In: *International Conference on Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2006. S. 509-518.

Kuhn, Tobias. Acewiki: A natural and expressive semantic wiki. *arXiv preprint arXiv:0807.4618*, 2008.

Kawamoto, Kensaku, Yasuhiko Kitamura, and Yuri Tijerino. "Kawawiki: A semantic wiki based on rdf templates." *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE, 2006.

Industry 4.0 national and international

Nils Ehlers, Robert Loew, Udo Bleimann
University of Applied Sciences Darmstadt, Department of Computer Science, Germany
e-mail: nils.ehlers@h-da.de, robert@loew.com

Keywords: Knowledge, Industry 4.0, Internet of Things, Industrial Internet Consortium

At the beginning we will explain the term ”Industry 4.0” and the German point of view. In a next step the differences between several countries will be analyzed in more detail.

Germany

International compared Germany is a high-wage region. This leads to the key question if it is possible to maintain and to enhance a production site given the international competition. In contrast to many other industrial countries, Germany was still able to maintain its number of employees in the production sector. The middle class industry plays an important role in Germany. In addition, the small and medium enterprises are the most innovative industries in Germany and helped therefore to minimize the impact of the financial crisis for the German economy, compared to many other countries. The development and the integration of new technologies and processes played a major role in that case. “Industry 4.0” describes the fourth industrial (r)evolution, after the steam machine in 1712, the product line in 1913 and the digital revolution in 1969. Since it will play such an important role for Germany as well as for the rest of the world, it is important for Germany as a production country to maintain and enhance its position.

The development of mechanism with water and steam is called the first industrial revolution. This revolution led to a strong increase in the development of technology, productivity and science.

The second industrial revolution was the development of the production line. Besides new research and knowledge orientated industries, industries in sectors like electrical engineering, chemical industry, mechanical engineering and optical industry took place in Germany. This was also the time when Germany became the leading country in the mechanical engineering industry. This position helped Germany to develop and to use the production line to increase efficiency and to reduce headcount.

The third revolution, also called the digital revolution, describes the digitalization and the related change in industry as well as in many other areas of life the end of the 20th century. Communication played a major role at that time and increased steadily. In 1986 already 20% of the communication was digitalized, 1993 two third and in 2000 already 98% of the whole communication in the world. The digitalization of information processes as well as communication processes led to an information explosion. The global telecommunication capacity as well as the information memory capacity per head increased between 1986 and 2007 about 23-28% per year. These are impressing numbers, especially in comparison with global economy growth of around 3-6% in the same time period.

The fourth revolution is characterized by the interaction between the industrial production with modern information and communication technologies. The German definition of the fourth revolution, based on the “Bundesministerium für Bildung und Forschung”, is: “The future project

Industry 4.0 aims to make the German industry ready for the production in the future." One characteristic is the high level of individualization of the products with the precondition of flexible production for huge amounts of products. Known as a smart factory, a vision of a production environment, production sites as well as logistical systems interact almost fully without the intervention of humans. As a technical basis, cyber-physical systems (CPS) will support the vision under the term "Internet of Things" (IoT). CPS will also allow to produce smart products which have their own memory with all of their characteristics, production processes and determinations. In addition, smart products are also able to exchange that information between other smart products as well as with machines. But not only smart products are able to communicate with each other; also machines will be able to exchange information. This communication is called M2M communication and allows machines to make their own decisions during the whole production process. This communication technology is essential for the effective use of Industry 4.0 and is part of the Internet of Things (IoT) and the Internet of Services. Both terms combined are called the Internet of Things and Services. The relation between IoT and CPS is neither completely clarified nor clearly separated because both concepts were researched and developed simultaneously but they have always been closely related. The term IoT is widely used in America whereas CPS is more common in Europe. Nevertheless, the goal of the total connection und communication of everything is the same for both technologies. Also customers as well as business partners are directly involved in the business processes and the value added processes. The production is also linked with high quality service. With intelligent monitoring and decision processes, companies as well as whole value added processes should be controlled and optimized in real time. Now is the time for Germany to directly shape the fourth industrial revolution. Nevertheless, it is important to keep Industry 4.0 as a research project since it is only depicting a guideline.

Europe

One could assume that under the umbrella of the European Union a centralized guideline for the implementation of Industry 4.0 exists. There are several programs on European level, nevertheless do many European countries have their own initiatives with own goals and funding programs. Besides the German Industry 4.0 project, the EU as well as several countries do have some activities related to the production of the future. On European level a high impact initiative has been started by the European Institute of Innovation & Technology (EIT) with the name "Industry 4.0: Powering Europe". This initiative is under German lead and has the goal to rollout cyber physical platforms in smart factories all over Europe. In addition there is the public-private partnership "Factories of the Future" that focusses on the support of small and medium enterprises enabling success on the global market. Several other countries have also their own activities: Finland wants to spend for his "Industrial Internet Program" around 100 million Euro till 2019. Austria wants to invest 250 million Euro for his Industry 4.0 program. Also France, Great Britain and the Netherlands have started with their own industry 4.0 programs.

USA

In March 2014 the Industrial Internet Consortium (IIC) was founded in the USA. More than 100 companies are participating, like General Electric, IBM, Intel, AT&T and Cisco, to work on the technology for the production of the future. Also important German companies are participating at the consortium, like Siemens and Bosch. Since many IT companies are included, the main focus is on IoT and the interconnection of future factories. The vision of IIC is more or less similar with the goal of Industry 4.0. Production should be more efficient and value added processes should be

optimized. Furthermore, the companies want a higher availability of the machines as well as a high level of individualized production.

China

The Chinese government wants to build up intelligent production systems on the level of omnipresent information systems and communication systems. This is visualized in their "Advanced Manufacturing Technology Roadmap" which is depicting every step till 2050, including milestones for 2020 and 2030. The aim of China is to get less dependend on imported products with the help of technologies for the production, for instance in the automotive sector. This is also clearly visible in the five-year plan since "Advanced Manufacturing Equipment" is one of the areas with the highest priority.

In other countries, like Japan and South Korea, activities are still not clearly known. The visit of many delegations in Germany shows that the interest in this area is extremely high.

References

Toni Pierenkemper: Wirtschaftsgeschichte. Die Entstehung der modernen Volkswirtschaft. Berlin, 2009, S. 88

„The World's Technological Capacity to Store, Communicate, and Compute Information", Martin Hilbert and Priscila López (2011), Science, 332(6025), 60–65

„Videoanimation über The World's Technological Capacity to Store, Communicate, and Compute Information from 1986 to 2010"

Bernd Overmaat, World Wide Wettlauf - Industrie 4.0 in Europa/Asien/USA, https://engineered.thyssenkrupp.com/world-wide-wettlauf-industrie-4-0-aktivitaeten-in-europe-asien-usa/

Was ist Industrie 4.0? In: www.plattform-i40.de. Accessed 9. April 2016.

Chapter 3

Applied Computing

Prototype proof of concept for a mobile agitation tracking system for use in elderly and dementia care use cases

Michael Healy, Alfie Keary, Paul Walsh
Department of Computer Science
CIT – Cork Institute of Technology, Ireland
e-mail: michael.healy2@mycit.ie; alphonsus.keary@cit.ie; paul.walsh@cit.ie

Keywords: Vision, Agitation, Affective, Dementia, CMAI

Introduction

The Mobile Agitation Tracker (MAT) project aims to develop and evaluate an initial set of algorithms that can detect restlessness and agitation generated aggression in elderly patients and dementia care subjects. Cognitively impaired elderly subjects often show signs of being in an agitated state. The Merriam Webster medical dictionary defines agitation as a "state of excessive psychomotor activity accompanied by increased tension and irritability" (Merriam Webster, 2016). People who suffer from agitation are often tense, confused, irritable and also possibly showing signs of aggressive behaviour.

One of the main symptoms of agitation is restlessness and irritability. This is characterised by showing an inability to remain in a restful state of being. Initially MAT has been developed to track a person's state of restlessness and irritability. In order to implement this in an everyday setting, MAT uses innovative vision based analytics to track a subjects facial expressions in real-time. The MAT algorithms evaluate video frame based facial expression data and use identified feature data to determine whether an in-camera subject may be experiencing restlessness or irritability. MAT also has the potential to be used in more advanced machine learning and data analytics applications typically for research purposes on elderly care.

Related research

MAT has been developed to address the quality of life aspects experienced by cognitively impaired elderly subjects. A study conducted on 408 nursing home residents found that there is a direct link between agitated behaviours and cognitive function (Jiska Cohen-Mansfield, 1990). The study used the Cohen-Mansfield Agitation Inventory (CMAI) (Cohen-Mansfield, 1991) which is used to score 29 different agitation states on a 7-point scale. These 29 states fall under three categories;

1. Aggressive behaviour,
2. Physically nonaggressive behaviour,
3. Verbally agitated behaviour.

The CMAI describes some of the physically non-aggressive behaviours such as pacing, stealing, trying to get to a different place, handling items inappropriately and general restlessness. The report also found that "cognitively impaired patients were significantly more physically non-aggressive than patients with low cognitive impairment" (Jiska Cohen-Mansfield, 1990).

CMAI is a complex evaluation process to be carried out by trained professionals and it has been used as the foundation for the MAT project from a clinical perspective. The examples listed below (relating to typical CMAI based visual observations of a subject) have been the basis for the initial set of vision based algorithms to be developed for MAT.

- Agitation may be abusive, aggressive or repetitive behaviour towards self or others such as shouting, head banging, staring and repetitive mouth opening and closing.
- Agitation may be appropriate behaviour performed with inappropriate frequency, such as excessive head movement, rocking, trying to get up from a chair or out of bed constantly, and asking the same questions continually.
- Agitation may be inappropriate according to social standards for the specific situation, as in spitting at a person, hitting a person or taking clothes off in a common room.

Research approach

The first version of MAT has been developed to address the following two typical use cases.

Use Case #1: Restlessness Detection.

- MAT monitors the head movements of an in-camera subject for a given time period to detect if the subject may be in a restless state or not. This typically relates to a subject that may be in bed or confined to a chair that is demonstrating excessive head movement which may be any mixture of head rolling, head banging of an object, excessing tilting or drooping of the head. MAT also provides a report for a subject's carer for on-going analytics on head movement based agitated behaviours.

Use Case #2: Aggression Detection.

- The second MAT use case is more advanced and involves the monitoring and tracking of specific facial landmark points from the in-camera facial frame images of the subject. This use case is closely related to the CMAI visual observation one outlined above and is designed to detect if a subject is showing signs of a violent outburst or abusive behaviour. This version of MAT is able to detect facial expression images that demonstrate that the subject is showing an aggressive or angry facial expression

MAT also incorporates and provides for a subject's carer to receive cloud based text alerts to their mobile phone using configurable mobile application functionality built into the system. Both of the above use cases may be configured to incorporate mobile alert functionality and can also provide for the review and analysis of previous alerts notified by MAT.

Preliminary results

MAT has originated from a final year computer science project and has been supervised by researchers from the Cork Institute of Technology, SIGMA group (Cork Institute of Technology (CIT), 2016). The system has been tested in a lab based environment for both the vision and cloud based alert aspects of the software. Both use cases have been tested extensively within the lab but due to ethical, security and privacy consideration the MAT platform has not reached a stage where

it is ready for real-world testing. Also during the testing phase the following limitations were identified in relation to the current versions of the detection algorithms. For example, performing an analysis of a subject's mouth to distinguish normal speaking from shouting or screaming is challenging and needs further work. Also this could be enhanced in a future version of MAT by incorporating voice analytics into the detection algorithm. The algorithm for finding common facial features that could be used to identify aggression in subjects is also at a very early stage and could be improved using affective computing techniques (Picard, 1997), facial expression databases and machine learning algorithms.

The following screen shot demonstrates the lab based testing of the real-time tracking capabilities of the MAT platform.

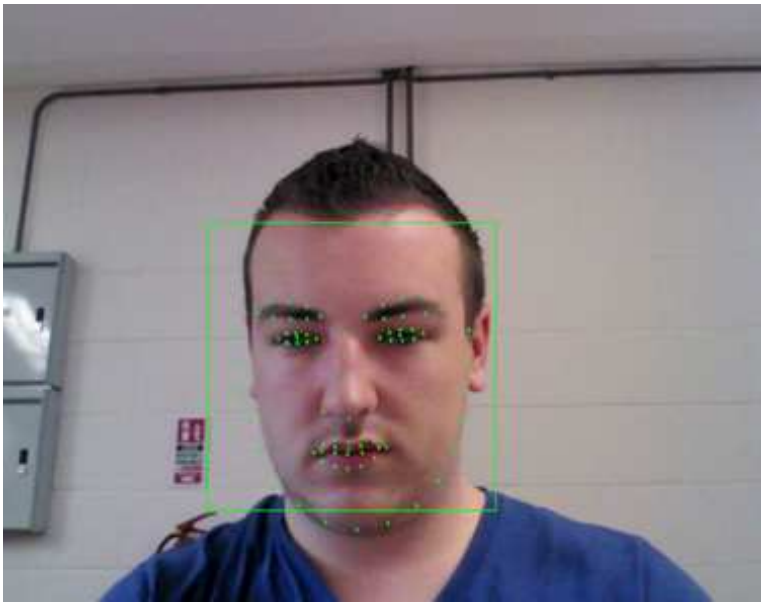


Figure 1. Example of facial landmarks generated by MAT

Research – Current and next stage

The Cork Institute of Technology is the lead partner in the development of an affective computing based platform called SenseCare (SenseCare EU Partnership, 2016). SenseCare is a first of its kind and will uniquely create a cloud based affective computing operating system capable of processing and fusing multiple sensory data streams to provide cognitive and emotional intelligence for AI connected healthcare systems. The underlying sensory layer of the SenseCare platform involves the fusion of data streams from multiple affective sensory devices such as vision, wearables, audio and EEG devices. These data streams are then processed by a fusion and machine learning layer in the SenseCare platform that delivers cognitive and affective state data on a subject.

MAT has the potential to be incorporated into the lower layers of the SenseCare platform as it can provide raw facial landmark data that can be fused with the other affective based sensory adaptors configured for the SenseCare platform. Currently MAT provides data sets for each subject's monitoring period in a CSV file format which could be used to populate SenseCare data repositories or act as a feature set for SenseCare related machine learning classification algorithms.

Conclusions

MAT is at an early stage of development but the next version of the software will see increased focus and implementation of other aspects of the CMAI such as analysing the subject's body movement as mentioned in the visual observations above. This would see the MAT project addressing further aspects of CMAI process and would greatly improve accuracy in detecting agitation. Also machine learning algorithms (possibly SenseCare based) could be used to analyse both existing data and data created by MAT. This analysis would provide advanced insights into agitation detection for elderly care and future research.

Work is also expected to start on the development of a MAT based adaptor for the SenseCare platform. This work will lead to further development of the existing MAT algorithms with a specific SenseCare related focus.

A video demonstration of M.A.T's head tracking capabilities can be seen on this YouTube video https://youtu.be/hZn3PuGha1w

References

Cohen-Mansfield, J., 1991. Cohen-Mansfield Agitation Inventory. [Online]

Available at: www.dementia-assessment.com.au/symptoms/cmai_manual.pdf [Accessed 28 July 2016].

Cork Institute of Technology (CIT), 2016. Software Intelligence and Machine Intelligence Applications group (SIGMA). [Online] Available at: http://sigma.cit.ie/ [Accessed 28 July 2016].

Jiska Cohen-Mansfield, M. S. M. a. A. S. R., 1990. "Dementia and Agitation in Nursing Home Residents", s.l.: American Psychological Association.

Merriam Webster, 2016. Merriam Webster Medical Dictionary. [Online]

Available at: http://www.merriam-webster.com/medical/agitation [Accessed 28 July 2016].

Picard, R., 1997. Affective Computing. s.l.:MIT Press Cambridge, MA, USA ©1997.

SenseCare EU Partnership, 2016. Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare). [Online] Available at: http://www.sensecare.eu/ [Accessed 28 July 2016].

FindR-TB: A cloud–based tool for antibiotic resistance prediction in *Mycobacterium tuberculosis*

Paul Walsh, Micheál Mac Aogáin, Brendan Lawlor . . . .
Department of Computer Science
CIT – Cork Institute of Technology, Ireland

Department of Clinical Microbiology, School of Medicine,
TCD – Trinity College Dublin.
e-mail: m.macaogain@tcd.ie

Keywords: Cloud computing, Microbiology, Antibiotic resistance, Tuberculosis

The emergence of affordable DNA sequencing technology holds promise for infectious disease diagnostics. Access to whole-genome sequence (WGS) data of pathogenic micro-organisms is now within reaching distance of hospital diagnostic laboratories and promises to bring about positive disruptive changes in diagnostic microbiology. However, without 'point and click' applications addressing computational bottlenecks in the analysis of complex bacterial DNA sequences, clinical adoption of genome-based molecular diagnostics will not be realised.

Mycobacterium tuberculosis (TB) is a global infectious disease that affects approximately 9 million people every year leading to over a 1 million deaths. An estimated 450,000 TB infections exhibit drug resistance. The basis of drug resistance is underpinned by genetic mutations in the TB genome and is therefore predictable given access genomic sequence data. The application of WGS analysis to TB has gained traction as an aid in the diagnosis of TB promising gains in antibiotic resistance detection times and more accurate disease transmission mapping of infected individuals.

Here we report a new development in our cloud-based bacterial DNA management software; the FindR-TB resistance predictor. This tool is integrated into our cloud-based microbial DNA analysis platform (SimplicityTM) and functions in the prediction of antibiotic resistance in *M. tuberculosis* from in-putted user sequence data. In an initial characterisation of our tool we have tested it on genomic data from 100 *M. tuberculosis* isolate genomes allowing us to document the sensitivity and specificity of the tool by using WGS data from phenotyped strains with characterised antibiotic sensitivities.

Diagnostic test statistics were generated for 100 strains allowing us to gauge the potential of this tool as a diagnostic aid for clinical application; i.e. prediction of antibiotic resistance profiles of TB strains and subsequent guidance of therapeutic choices. Resistance to first line antibiotic agents for the treatment of TB including isoniazid, rifampicin, ethambutol and streptomycin was detected with a sensitivity of 91%, 98%, 96% and 100% respectively, and corresponding specificity of 97%, 90%, 73% and 100%. The first line agent pyrazinamide had a significantly lower sensitivity of 34% with a specificity of 100%. These results highlight the ability of our *in silico* predictions to be comparible with currently available diagnostic tests for antibiotic resistant TB while highlighting areas of discord where further characterisation of resistance mechanisms is required.

We plan to test our FindR-TB on a broader collation of TB isolates in order to hone the diagnostic test statistics and improve results. Future tools are currently in development for other pathogenic bacteria including *Clostridium difficile* and *Staphylococcus aureus.* Ultimately these tools will become cornerstones of the clinical microbiology lab as WGS technologies become increasingly integrated into the laboratory work-flows of modern medical diagnostics.

Towards lifecycle management of clinical health care environments

[1] Ulrich Kowohl, [2] Felix Engel, [3] Paul Walsh, [3] Alfie Keary,
[4] Michael Fuchs and [2] Matthias Hemmje

[1] FernUniversität in Hagen, Universitätstraße 1, Hagen, Germany,
Ulrich.Kowohl@Fernuni-Hagen.de
[2] Research Institute for Telecommunication and Cooperation, Dortmund, Germany,
{Fengel,Mhemmje}@FTK.de
[3] Cork Institute of Technology, Cork, Ireland
Paul.Walsh@cit.ie
[3] Wilhelm Büchner University of Applied Sciences, Pfungstadt, Germany mfuchs@ftk.de

Keywords: SenseCare, Patient Information Lifecycle Management, Patient Data Management, Assisted Living Sensory Data Fusion, Patient Data Management System, Open Archive Standard, Archive Centric Information Lifecycle

Affective computing (AF, cf. [1]) is the area of systems that can recognize, interpret, process, and simulate human affects. AC systems interpret the emotional state of humans and adapt its behaviour to them, giving an appropriate response for those emotions. In medicine, human affects is central to the work carried out by care givers. It is proposed that "understanding, reflecting on, engaging with and expressing emotion(s) are important aspects of medical practice".

Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is an EC co funded project, aiming to provide a new AC providing software services applied to the dementia care and connected health where enormous potentials and opportunities exist to provide assistance to medical professionals and patients on cognitive states and overall holistic well-being. The SenseCare solution will integrate data streams from multiple sensors. The potential exists to integrate holistic assessment data into multiple applications across connected healthcare domains. This paper adresses two challenges within the scope of the SenseCare project:

- How can sensory and fused sensory data that come into existence along a patient treatment be systematically managed, processed, analysed, annotated, packaged, and archived to enable its comprehensive long-term preservation, access, and re-use?
- How to represent and assemble provenance information in the sense of temporal bounded treatment data to enable traceability of treatment progress or traceability of decisions made?

We will present a conceptual design for a Patient Data Management System (PDMS) which manages patient data across an archive centric Information Lifecycle and aims at supporting access and reuse of data among usage scenarios.

Dementia is a long-term cognitive and psychologically debilitating disease. It reduces cognitive, brain and communication abilities. Several conditions like Alzheimer, Depression, Lewy Dementia or vascular Dementia are related to dementia. Treatment can be done with: drugs, brain training, psychological therapies (cf. [3]) in a care giving environment and lasts considerable time. During therapy, data will be generated: sensory data, clinical data, care plan/execution data and

communications. Data underlies a lifecycle that needs to be managed in compliance with regulatory like e.g. privacy laws. In SenseCare sensory data will be fused with other data and then analysed, to improve life and care quality for patients living at home with care assistance. We mainly address three use cases: ***Assisted Living Scenario*** deals with home living patients and their daily condition. With monitoring patients health conditions, care givers can be supported by a well being indicator so that planning can be adopted to patients needs. Caregivers can already prepare themselves according to patients well being or alert medical support in case of serious health incident indication. Patients well being trend analysis can indicate health incidents or necessary therapy adjustments. The ***Emotional monitoring*** during medical treatment enables monitoring patients affective signals. By having monitoring in place, therapists are able to judge behaviour in an improved way which improves diagnostic quality. In addition to successful tracking of dedicated therapy sessions reaction of patients regarding therapist can be tracked. This enables therapists to exchange therapy personal in order to have a best match between patient and therapist. Another scenario is ***Shared care giving Scenario*** – e.g. by family members. During treatment processes several groups of persons and actions are generating data. To enable caregivers to have a clear picture of the patient's condition and its history, data needs to be archived in searchable accessible and interpretable way. The goal is that the collection of information enables stakeholders to provide better medical treatment and diagnostic results. Possible incidents should be identified upfront which makes future diagnosis easier for the doctor since he has some data to work with.

The use cases showcase that various information comes into existence during a patient treatment that are of potential use for later exploitation, to increase a patient's wellbeing. We therefore propose the introduction of a *Patient Lifecycle Management system* (PLMS) supporting lifecycles of information. PLMS will use approaches from PDM/PLM and Patient data management systems. Existing PDMS are specialized EPR record archives. PLMS system extends the PDMS in order to support archive centric information lifecycles and integrates collaboration between stakeholders We will introduce a container based approach to introduce patient data structure (upper bill of information) and its instances (lower bill of information). Figure 1 displays one example instance of our PLMS with one patient.

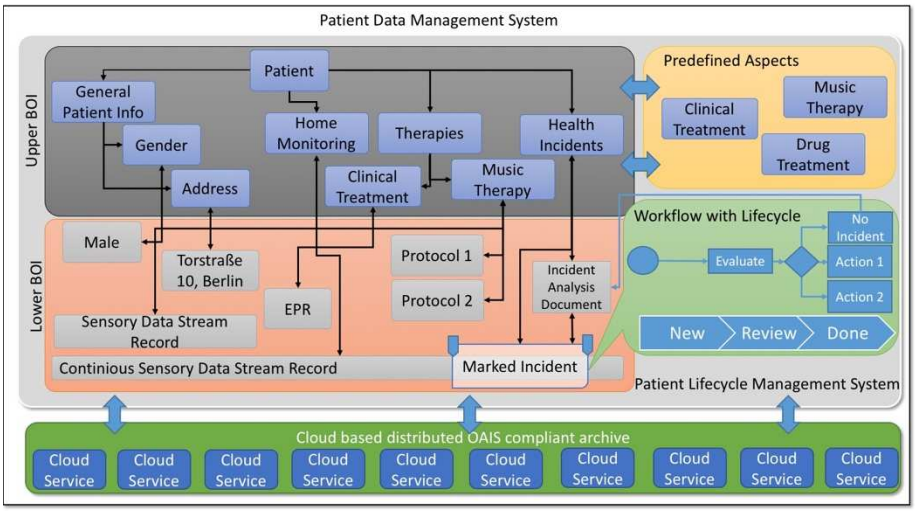


Figure 1. Novell Approach with one Patient Data container

The paper will give a brief overview about the SenseCare project and described SenseCare usage scenarios from a data management perspective. Next to standard data management planning activities we will describe the further need of data management approaches for:

- Relating EPR, collaboration, sensory data, fused sensory data as well as other clinical data
- Enhancing the ability to find already archived data and to speedup searching processes
- Introducing further approaches from PDM like workflows and workflow based lifecycles to optimize collaboration between stakeholders

For achieving this, we will outline that we will port approaches based in the Product Data Management and Product Lifecycle Management area to the domain of clinical treatment and management of patient data in clinical treatment usage scenarios. Purpose and status of the project related to data management will be shown and next steps will be presented

References

Tao J. and Tan T.: Affective Computing: A Review, Lecture Notes in Computer Science, Vol. 3784, pp 981-995 (2005)

Nawroth, C., Schmedding, M., Brocks, H., Kaufmann, M., Fuchs, M., & Hemmje, M. (2015). Towards Cloud-Based Knowledge Capturing Based on Natural Language Processing. Procedia Computer Science, 68, 206-216

Cayton, Harry, Nori Graham, and James Warner. Dementia: Alzheimer's and Other Dementias: The'at Your Fingertips' Guide: The Fully Updated and Comprehensive Reference Book for Alzheimer's and Other Forms of Dementia. Class Publishing Ltd (2002)

Regnier, Victor. Design for assisted living: Guidelines for housing the physically and mentally frail. John Wiley & Sons, 2003.

Ambaruch, B., Shandwick, W.: PTC's PLM Momentum Accelerates as Automotive Industry Evolves, PTC, BusinessWire. URL: http://files.shareholder.com/downloads/PMTC/0x0x524572/4a74a130-3109-42b8-9446-a0abb97b5e58/PMTC_News_2011_2_14_General.pdf last accessed 2016.03.01

Tekale S. K., Nandedkar V. M. PLM Implementation with Creation of Multiple Item using SOA Technology in Team center. IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE), ISSN(e) : 2278-1684, ISSN(p) : 2320–334X, PP : 62-66

Fretschner, R., Bleicher, W., Heininger, A., & Unertl, K. (2001). Patient data management systems in critical care. Journal of the American Society of Nephrology, 12(suppl 1), S83-S86. (2001)

Mitev, N. and Kerkham, S., 2001. Organizational and implementation issues of patient data management systems in an intensive care unit. Journal of Organizational and End User Computing, 13(3), p.20. (2001)

CCSDS: Reference model for an open archival in-formation system (OAIS) (2009)

Engel, F., Klas, C., Brocks, H., Kranstedt, A., Jäschke, G., & Hemmje, M. Towards supporting context-oriented information retrieval in a scientific archive based information lifecycle. Proceedings of Cultural Heritage online. Empowering users: an active role for user communities, (pp. 135-140). Florence, Italy. (2009)

Brocks, H., Kranstedt, A., Jäschke , G., & Hemmje, M.. Modelling context for digital preservation. Stud. Comput. Intell., 197–226 (2010)

Crompton, S., Giaretta, D., Matthews, B., Engel, F., Brocks, H., Shaon, A., & Marelli, F. SCIDIP-ES: A Sustainable Data Preservation Infrastructure to Support OAIS Conformant Archives. APA/C-DAC International Conference on Digital Preservation and Development of Trusted Digital Repositories. New Dehli, India. (2014)

Lagoze, C., Van de Sompel, H., Nelson, M.L., Warner, S., Sanderson, R. and Johnston, P., 2008. Object re-use & exchange: A resource-centric approach. arXiv preprint arXiv:0804.2273. (2008)

Eye-Tracking in Computer-Based Simulation in Healthcare Training

Jonathan Currie, Raymond R. Bond, Paul McCullagh, Pauline Black, Dewar Finlay
School of Computing and Mathematics, Jordanstown Campus
Computer Science Research Institute
Ulster University, United Kingdom
e-mail: currie-j@email.ulster.ac.uk

Keywords: Eye-Tracking, Visual Attention, Simulation-Based Training, Task Performance

Introduction

Patient safety is a critical area of concern within healthcare and medical errors are a well-known problem that can have fatal ramifications (Kohn et al. 2000). Lack of knowledge and skill with clinical tasks and procedures, as well as decision-making can be significant factors with many of the errors that are reported in healthcare (Zhang et al. 2002). Many healthcare tasks can be simulated using computer and web technology for training purposes and provide trainees (students and practicing) with a way to improve or maintain their knowledge and skills (Persson et al. 2014; Cant & Cooper 2014). The concept of visual attention during a task has been tested in medical and healthcare task studies (O'Meara et al. 2015; Zheng et al. 2011; Breen et al. 2014) with an aim of finding discriminative differences between competency levels. The study of this previous work led us to hypothesise that eye tracking metrics exclusively have a relationship with specific task performance and can discriminate between performance level. We found the duty of patient monitoring with interpreting vital signs monitors in nursing earmarked in the literature for improvement in available simulation-based training. We sought to use eye-tracking with the task of nurses interpreting simulated patient vital signs from a monitor. The objective was to determine if eye-tracking technology can be used to develop biometrics for automatically classifying the performance of nurses whilst they interact with computer-based simulations.

Methods

For the study a total of 47 nurses were recruited, with 36 nursing students (TG - Training Group) and 11 coronary care nurses (QG - Qualified Group). Each nurse interpreted five simulated vital signs scenarios whilst 'thinking-aloud'. We recorded the participant's visual attention (eye tracking metrics [ETMs]), verbalisation, confidence level (1-10, 10=most confident). Interpretations were given a score out of ten to measure performance. Data analysis was used to find patterns between the ETMs and performance. Multiple linear regression was used to predict performance (score) using ETMs.

The five scenarios were designed by an expert nurse with validation was provided by three colleagues of similar expertise. The scenarios were designed to be within the expected ability of the average undergraduate student but also with assessment criteria that would likely provide fully trained and qualified nurses a higher score.

The criteria assessed participants' verbal responses at three levels and scores were allocated according to:

Basic performance: identification of abnormalities in the presented vital signs.

Mid-range performance: identification of why the abnormalities occurred based on their knowledge and understanding of the presenting condition outlined in the case scenarios.

High-range performance: identification of and decision-making with regard to the immediate interventions required to stabilise the patient.

Performance scores were then put into classes according to expert advice:

- 0-5 = low
- 6-7 = medium
- 8-10 = high

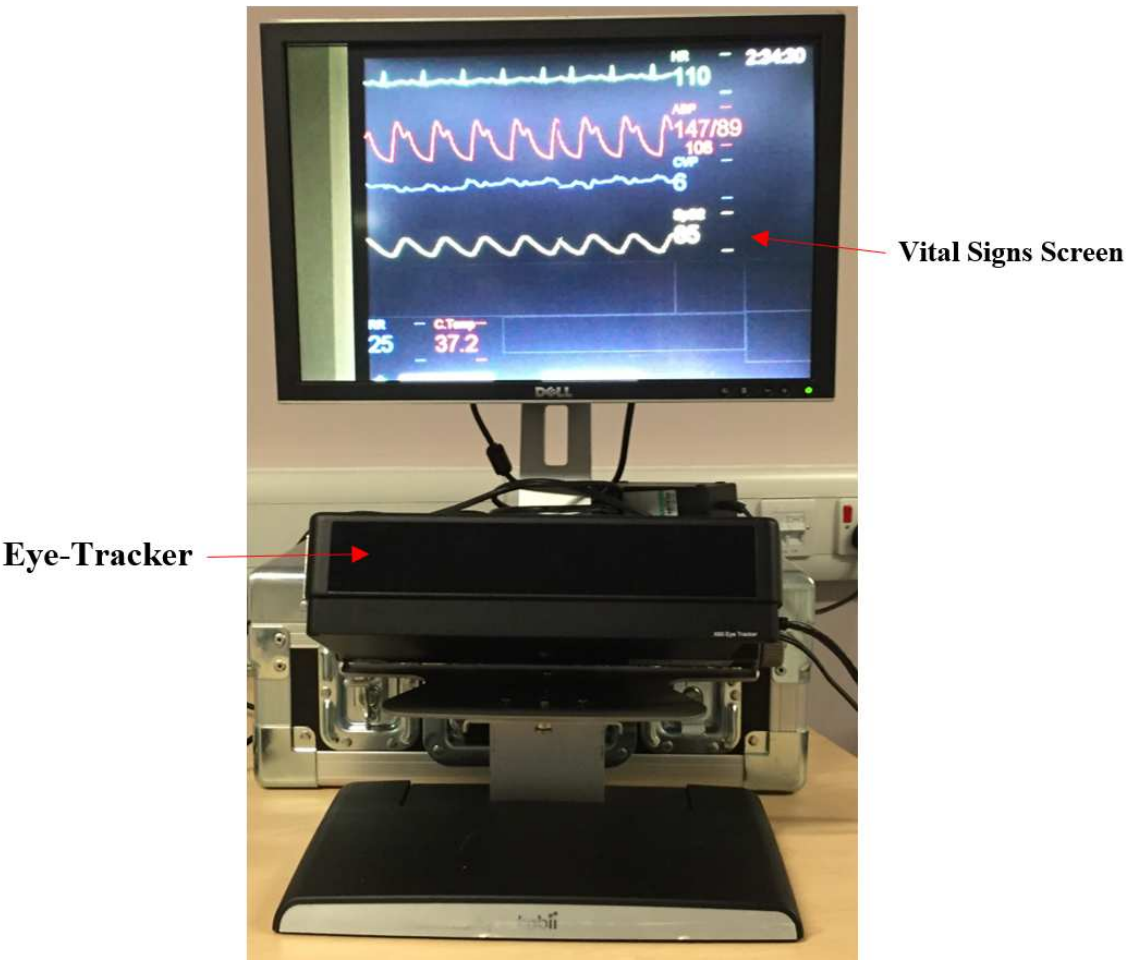


Figure 1. Participant Point of View During Interpretations

Figure 2. Simulated Vital Signs Screen with Eye Tracking Areas of Interest with Variable Names Highlighted

Results

The QG scored higher than the TG (6.85±1.5 vs. 4.59±1.61, p=<0.0001) and reported greater confidence (7.51±1.2 vs. 5.79±1.39, p=<0.0001).

Class	TG	QG	All
Low (0-5)	129 (72%)	10 (18%)	139 (59%)
Medium (6-7)	45 (25%)	25 (45%)	70 (30%)
High (8-10)	6 (3%)	20 (36%)	26 (11%)

Table 1. Interpretation Performance Results in Classes (Count and %)

20 ETMs discriminated between classes of performance (low, medium, high). Regression using ETMs (62 independent variables) was shown to adequately predict score (adjusted R2=0.80, p=<0.0001).

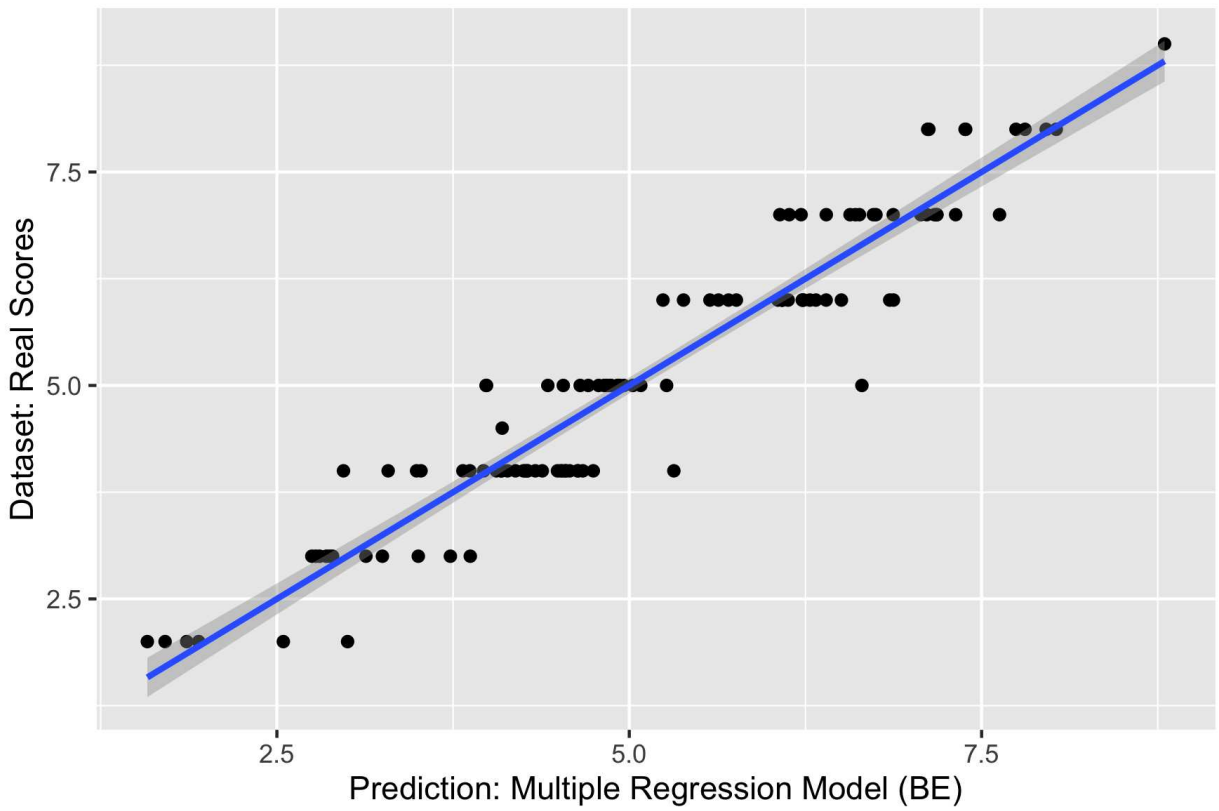


Figure 3. Participant Scores Vs Predicted Scores Using Subset of Eye Tracking Metrics As Predictors

Conclusions and Future Work

Thedata collected, specifically the data analysis of ETMs,has shown that visual attention and the level ofperformance for this specific task (measured byperformance score) are not independent of each other. Put differently, ETMs exclusively could be used topredict a person's performance when reading vital signs. Explicitly, we can see that specific ETMs are statisticallysignificant in discriminating between classes ofperformances for interpreting vital signs. With furtherrecruitment of participants (especially high classperformances), we could potentially reveal more ETMsthat are discriminatory between low, medium and highclass performance. Further research, including statisticaltechniques like PCA are required to refine the regressionmodels and the optimal level of accuracy for predictingperformance using the ETMs. At present, we canconclude that there is a relationship between ETMs and performance that can be seen but it is unclear to what extent ETMs can predict performances.

Future newly designed studies will include:

1. Using eye-tracking during a much more complex simulated procedure, such as cardio angiography. We want to test the assumption (Zheng et al. 2011) that trainees and experts have different patterns of visual attention during surgical tasks that classify them. During this, we will also look at the link between visual attention and attentional capacity (Weaver 1949) of the task performer by provided them with added stimulus/distraction.

2. Other future work will progress the study described above, with eye-tracking to discriminate performance whilst an experimental group is presented a changed layout of on-screen vital signs. This work will look at the influence of information hierarchy on interpretation performance as the experimental group is presented the same scenarios but with the most concerning vitals emphasized on screen (positioning). We wish to test if this provides any advantage to the experimental group in their performance. If so, it could make an argument for automatic prioritizing of vital signs in accordance to the specific patient (through machine learning algorithms) in future equipment to reduce poor interpretations and patient monitoring.

References

Breen, C.J., Bond, R. & Finlay, D., 2014. An evaluation of eye tracking technology in the assessment of 12 lead electrocardiography interpretation. Journal of electrocardiology, 47(6), pp.922–9. Available at: http://www.sciencedirect.com/science/article/pii/S0022073614003148.

Cant, R.P. & Cooper, S.J., 2014. Simulation in the Internet age: the place of web-based simulation in nursing education. An integrative review. Nurse education today, 34(12), pp.1435–42. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84906073568&partnerID=tZOtx3y1 [Accessed July 19, 2015].

Kohn, L.T., Corrigan, J.M. & Donaldson, M.S., 2000. To Err is Human: Building a Safer Health System, National Academies Press (US). Available at: http://www.ncbi.nlm.nih.gov/pubmed/25077248 [Accessed July 31, 2016].

O'Meara, P. et al., 2015. Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper. International emergency nursing, 23(2), pp.94–9. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84926483577&partnerID=tZOtx3y1.

Persson, J. et al., 2014. Evaluating interactive computer-based scenarios designed for learning medical technology. Nurse Education in Practice, 14(6), pp.579–585. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84918813292&partnerID=40&md5=ab8235ae36ae3a782c634b9afbf51abb.

Weaver, W., 1949. THE MATHEMATICS is closely linked with the concept of information. , 181(1), pp.11–15.

Zhang, J., Patel, V.L. & Johnson, T.R., 2002. Medical error: is the solution medical or cognitive? Journal of the American Medical Informatics Association?. JAMIA, 9(6 Suppl), pp.S75-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419424&tool=pmcentrez&rendertype=abstract.

Zheng, B. et al., 2011. Surgeon's vigilance in the operating room. The American Journal of Surgery, 201(5), pp.673–677. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0002961011001140.

Chapter 4

Computing

Harassment detection: a benchmark on the #HackHarassment dataset

Alexei Bastidas, Edward Dixon, Chris Loo, John Ryan
Intel
e-mail: edward.dixon@intel.com

Keywords: Machine Learning, Natural Language Processing, Cyberbullying

Introduction

Online harassment has been a problem to a greater or lesser extent since the early days of the internet. Previous work has applied anti-spam techniques like machine-learning based text classification (Reynolds, 2011) to detecting harassing messages. However, existing public datasets are limited in size, with labels of varying quality. The #HackHarassment [1] initiative (an alliance of tech companies and NGOs devoted to fighting bullying on the internet) has begun to address this issue by creating a new dataset superior to its predecssors in terms of both size and quality. As we (#HackHarassment) complete further rounds of labelling, later iterations of this dataset will increase the available samples by at least an order of magnitude, enabling corresponding improvements in the quality of machine learning models for harassment detection. In this paper, we introduce the first models built on the #HackHarassment dataset v1.0 (a new open dataset, which we are delighted to share with any interested researcherss) as a benchmark for future research.

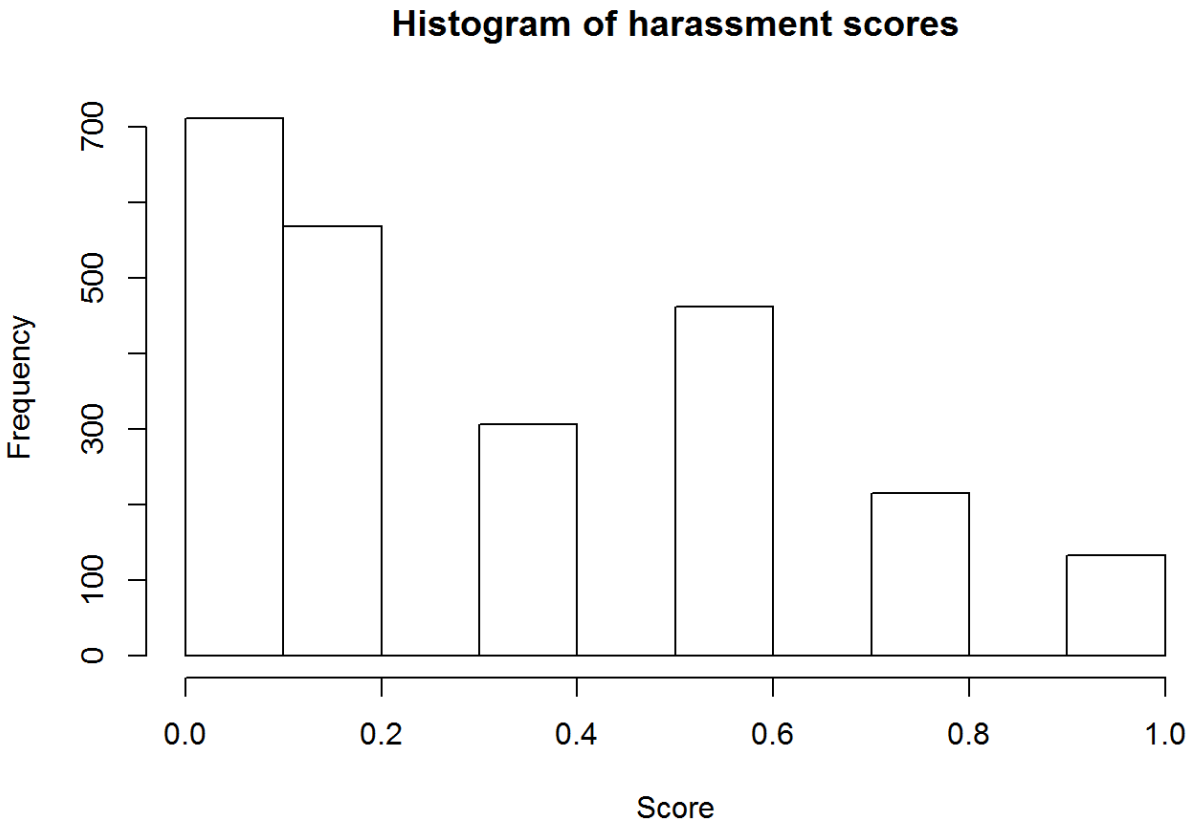
Related Work

Previous work in the area by Bayzik 2011 showed that machine learing and natural language processing could be successfully applied to detect bullying messages on an online forum. However, the same work also made clear that the limiting factor on such models was the availability of a suitable quantity of labeled examples. For example, the Bayzick work relied of a dataset of 2,696 samples, only 196 of which were found to be examples of bullying behaviour. Additionally, this work relied on model types like J48 and JRIP (types of decision tree), and k-nearest neighbours classifiers like IBk, as opposed to popular modern ensemble methods or deep neural-network-based approaches.

Methodology

Our work was carried out using the #HackHarassment Verison 1 dataset, the first iteration of which consists exclusively of Reddit posts. An initially random selection of posts, in which harassing content occured at a rate of between 5% and 7% was culled of benign content using models training on a combination of existing cyberbullying datasets (Reynolds 2001, also “Improved cyberbullying detection through personal profiles). Each post is labelled independently by at least five Intel Security Web Analysts. (a post is considered “bullying” if it labelled as such by 20% or more of the human labelers - as shown in the following histogram, a perfect consensus is relatively rare, and so we rate a post as “harassing” if 20% - 2 of our 5 raters - consider it to be harassing). This is a relatively balanced dataset, with 1,280 non-bullying/harassing posts,, and 1,118 bullying/harassing examples.

1 "Hack Harassment." 2016. 26 Jul. 2016 <http://www.hackharassment.com/>



All pre-processing, training and evaluation was carried out in Python, using the popular SciKit-Learn library [2] (for feature engineering and linear models) in combination with Numpy [3] (for matrix operations), Keras [4] and TensorFlow [5] (for models based on deep neural networks - DNNs).

For the linear models, features were generated by tokenizing the text (breaking it aparting into words), hashing the resulting unigrams, bigrams and trigrams (collectiojns of one, two, or three adjacent words) and computing at TF/IDF for each hashed value. The resulting feature vectors were used to train and test Logistic Regressioin, Support Vector Machine and Gradient Boosted Tree models, with 80% of data used for training and 20% held out for testing (results given are based on the held-out 20%).

For the DNN-based approach, a similar approach was taken to tokenization, both bigram and trigram hashes were computed; these were one-hot encoded, and dense representations of these features were learned during training, as per Joulin 2016.

2 "scikit-learn: machine learning in Python — scikit-learn 0.17.1 ..." 2011. 29 Jul. 2016 <http://scikit-learn.org/>

3 "NumPy — Numpy." 2002. 29 Jul. 2016 <http://www.numpy.org/>

4 "Keras Documentation." 2015. 29 Jul. 2016 <http://keras.io/>

5 "TensorFlow — an Open Source Software Library for Machine ..." 2015. 29 Jul. 2016 <https://www.tensorflow.org/>

The FastText model used is a python implenmentation of the model described in "Bag of Tricks for Efficient Text Classification." [6] . For the text encoding, bigrams and trigrams are used. 20% of the data was held out for testing.

The Recurrent Character Level Neural Network model consists of 2 GRU layers of width 100 followed by a Dense Layer of size 2 with softmax on the output, Between each of the layers batch normailization is performed. The optimiser used was rmsprop. For data preperation each of characters was onehot encoded and each sample was truncated/padded to 500 charcters in length. 20% of the data was held out for testing.

Results

Model	Precision (Harassing)	Recall (Harassing)
Gradient Boosted Trees (Scikit-Learn)	0.80	0.71
Bernoulli Naive Bayes	0.54	0.30
FastText	0.60	0.78
Recurrent Character Level Neural Network	0.71	0.73

Conclusions

We have presented the first results on a new open cyberbullying/harassment dataset. While our models clearly demonstrate a degree of ability to discriminate between the content classes, the achieved precision in particular falls far short of our ambitions for #HackHarassment.

Over the coming months, we'll massively expand the size of our labelled dataset, and review our labelling methodology, anticipating that a larger dataset will facilitate more accurate models. We look forward both to the availability of a larger dataset, and to seeing the development of classifiers that improvement on our work, and welcome partners able to contribute either in terms of expanding the dataset or improving the modelling.

6 "fastText" 2015. 22 Jul. 2016 <https://github.com/sjhddh/fastText>

References

Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on 18 Dec. 2011: 241-244.

Bayzick, Jennifer, April Kontostathis, and Lynne Edwards. "Detecting the presence of cyberbullying using computer software." (2011): 1-2.

Joulin, Armand et al. "Bag of Tricks for Efficient Text Classification." arXiv preprint arXiv:1607.01759 (2016).

Improved Cyberbullying Detection Through Personal Profiles http://doc.utwente.nl/80761/

Large-Scale Biomedical Data Integration and Data Mining: a Multiplex Network-based Approach

Haiying Wang, Huiru Zheng
School of Computing and Mathematics, Computer Science Research Institute
University of Ulster, N.Ireland, UK
e-mail: {hy.wang, h.zheng}@ulster.ac.uk

Keywords: Data integration, multiplex networks

Progress in medical sciences and molecular biology has led to the accumulation of tremendous amounts of biomedical data. In addition to traditional clinical data such as physical examinations, electrocardiogram (ECG) recordings and medical images, biomolecular data including DNA, RNA, protein sequences, and their two-dimensional (2D) and three-dimensional (3D) structures are being accumulated at an exponential rate in in both quantity and diversity, providing unique opportunities to study biological systems at different levels (Ritchie et al., 2015). For example, as a large-scale, collaborative effort led by the National Institute of Health, The Cancer Genome Atlas (TCGA) has collected massive, high quality information generated from various molecular levels for over 30 types of human cancer offering a valuable resource to accelerate our understanding of the molecular basis of human cancers (Cancer Genome Atlas Research, 2013).

Analysing such volume of biomedical data characterised by the presence of large amounts of high-dimensional data and a variety of fuzzy and noisy forms represents not only an enormous challenge but significant opportunity. Over the past decades, a wide range of computational approaches have been proposed and developed. Examples include iCluster (Shen et al., 2009), Similarity Network Fusion (SNF) (Wang et al., 2015), and novel statistical approaches (Pineda et al., 2015). In this study, an integrative, multiplex network-based approach was introduced to explore heterogeneous biomedical data (Mucha et al., 2010, Wang et al., 2016). For each given dataset, a similarity network will be constructed, in which each node corresponds to a subject and each edge represents the similarity between a pair of subjects derived from the given dataset. By introducing a coupling between networks to account for potential correlations between different data sources, a new multiplex network-based framework for integrative analysis of heterogeneous datasets, which has potential to extract homogeneous patterns shared by datasets, will be developed and implemented.

The system has been tested on the identification of the subtypes of glioblastoma multiforme (GBM) and breast invasive carcinoma from three omics data, i.e. i.e. mRNA expression, DNA methylationa, and miRNA expression. It has been shown that the subtypes identified are closely correlated with clinical variables as shown in Fig. 1 and a high level of concordance indicated by the value of Normalized Mutual Information (NMI) has been achieved in comparisons to state-of-the-art techniques (NMI > 0.8).

The preliminary results (Wang et al., 2016) have demonstrated that the proposed method represents a useful alternative network-based solution to large scale, integrative analysis of heterogeneous biomedical data and a significant step forward to the methods already in use in which each type of data are treated independently. It has several advantages. For example, the framework has the ability to correlate and integrate multiple data levels in a holistic manner to facilitate our understanding of the pathogenesis of disease. It provides a flexible platform to integrate different types of patient data, potentially from multiple sources, allowing discovering complex disease patterns with multiple facets. Currently we are applying the techniques to extract patterns associated with Alzheimer's disease based on integration of longitudinal clinical data and MRI images.

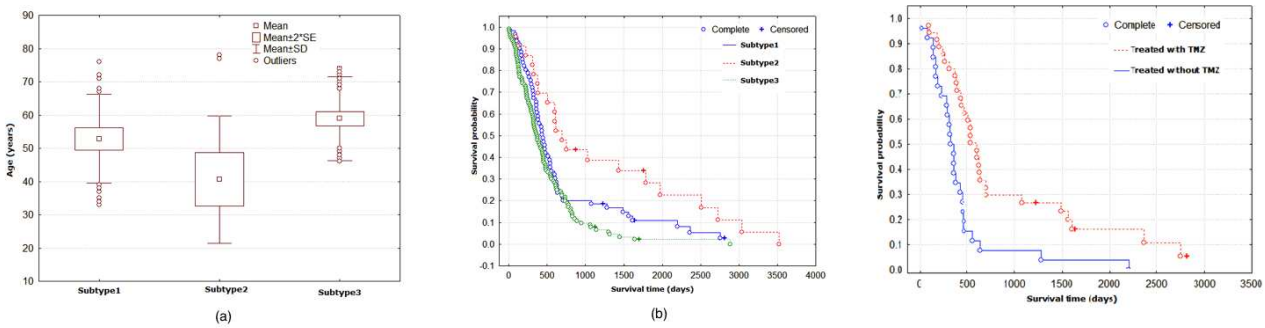


Figure 1. The correlation with clinical variables: (a) A statistically significant difference in terms of the average age was observed across 3 GBM subtypes (ANOVA test, $p < 0.0001$) with Subtype2 is closely associated with younger patients; (b) Survival times are significantly different among three GBM subtypes with patients in Subtype 2 having a more favorable prognosis; and (c) Survival analysis of GBM patients for treatments with temozolomide (TMZ) in Subtype1. A significantly increased survival time was observed (Cox log-rank test, $p < 0.005$).

References

Cancer Genome Atlas Research Network (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 2013, 497(7447), pp.67-73

Mucha, P., Richardson, T., Macon, K., Porter, M., and Onnela, J. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, vol. 208, pp. 876-878.

Pineda, S., Real, F.X., Kogevinas, M., Carrato, A., Chanock, S.J., Malats, N., et al. (2015). Integration Analysis of Three *Omics* Data Using Penalized Regression Methods: An Application to Bladder Cancer. *PLoS Genet*, 11(12): e1005689.

Ritchie, M. D., Holzinger, E. R., Li, Pendergrass, S. R., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16, pp.85–97

Shen, R., Olshen, A., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009, 25, pp. 2906–2912.

Wang, B., Mezlini, A.M., Demir, F., Flume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11, 333–337.

Wang, H., Zheng, H., Wang, J., Wang, C., and Wu, F. (2016) Integrating omics data with a multiplex network-based approach for the identification of cancer subtypes. *IEEE Transactions on Nanobioscience*, in press.

Towards a Cross Industry Standard Process to support Big Data Applications in Virtual Research Environments

K. Berwind, M. Bornschlegl, M. Hemmje
Faculty of Mathematics and Computer Science
University of Hagen, Germany
e-mail: kevin.berwind@fernuni-hagen.de, marco-xaver.bornschlegl@fernuni-hagen.de, matthias.hemmje@fernuni-hagen.de

M. Kaufmann
Engineering & Architecture
Lucerne University of Applied Sciences and Arts
e-mail: m.kaufmann@hslu.ch

Keywords: Big Data, CRISP4BigData, IVIS4BigData, Reference Model for Big Data Management, Virtual Research Environments

Introduction

The volume of data in enterprise data centres increases between 35 and 50 percent each year (Beath et al. 2012). Through new wide ranging developments in the area of information- and sensor technology, beside increasing computational processing capabilities, especially in terms of calculation intensive and analytics methods will be fostered. The collected mass data stem from the internal and external corporate division and consist of unstructured data such like text data, processing information, (database-) tables, graphics, videos (Beath et al. 2012), e-mails, feeds and sensor data (Freiknecht 2014).

The acquisition of decisional relevant mass data is known as “Big Data”. At the beginning, the term was defined by companies which had to handle an amount of fast growing data. The information technology research and advisory company Gartner Inc. defines the term as follows “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making” (Gartner 2015). In contrast the market research company IDC defines Big Data as follows: “Big Data technologies as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis” (Carter 2011).

The analysis of data is a process which could be managed by analysis process models like CRISP-DM (Cross Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, and Assess), or the KDD (Knowledge Discovery in Databases). Due to the new requirements (e.g. Meta data enrichment, archiving, automatic expert finder) on analysis, following, the new analysis process model called Cross Industry Standard Process for Big Data (abbreviated CRISP4BigData) will be offered.

Cross Industry Standard Process for Big Data Reference Model

CRISP4BigData is based on Kaufmann's reference model for Big Data Management (Kaufmann 2016), Bornschlegl´s IVIS4BigData reference model (Bornschlegl et al. 2016). IVIS4BigData was evaluated during an AVI 2016 Workshop in Bari. The overall scope and goal of the workshop was to achieve a road map, which can support the acceleration in research, education and training activities by means of transforming, enriching, and deploying advanced visual user interfaces for managing and using eScience infrastructures in VREs. In this way, the research, education and training road map will pave the way towards establishing an infrastructure for a visual user interface tool suite supporting VRE platforms that can host Big Data analysis and corresponding research activities sharing distributed research resources (i.e., data, tools, and services) by adopting common existing open standards for access, analysis and visualization. Thereby, this research helps realizing a ubiquitous collaborative workspace for researchers which is able to facilitate the research process and its Big Data Analysis applications (Bornschlegl et al. 2016).

CRISP4BigData is an enhancement of the classical Cross Industry Standard Process for Data Mining (CRISP-DM) developed by the CRISP-DM consortium, existing of DaimlerChrysler (later Daimler-Benz), SPSS (later ISL), NCR Systems Engineering Copenhagen und OHRA Verzekeringen en Bank Groep B.V., with the target to handle the complexity of Data-Mining projects (Chapman et al. 2000).

The CRISP4BigData reference (cf. Figure 1: CRISP4BigData Reference Model Version 1.0) model is based on Kaufmann`s 5 phases of Big Data Management such as "Data Collection, Management & Curation", "Analytics", "Interaction & Perception", "Insight & Effectuation", and "Knowledge-based Support". Within each of this phases are process elements implemented which handle special instructions und tasks to deal with Big Data and Analysis processes and projects.

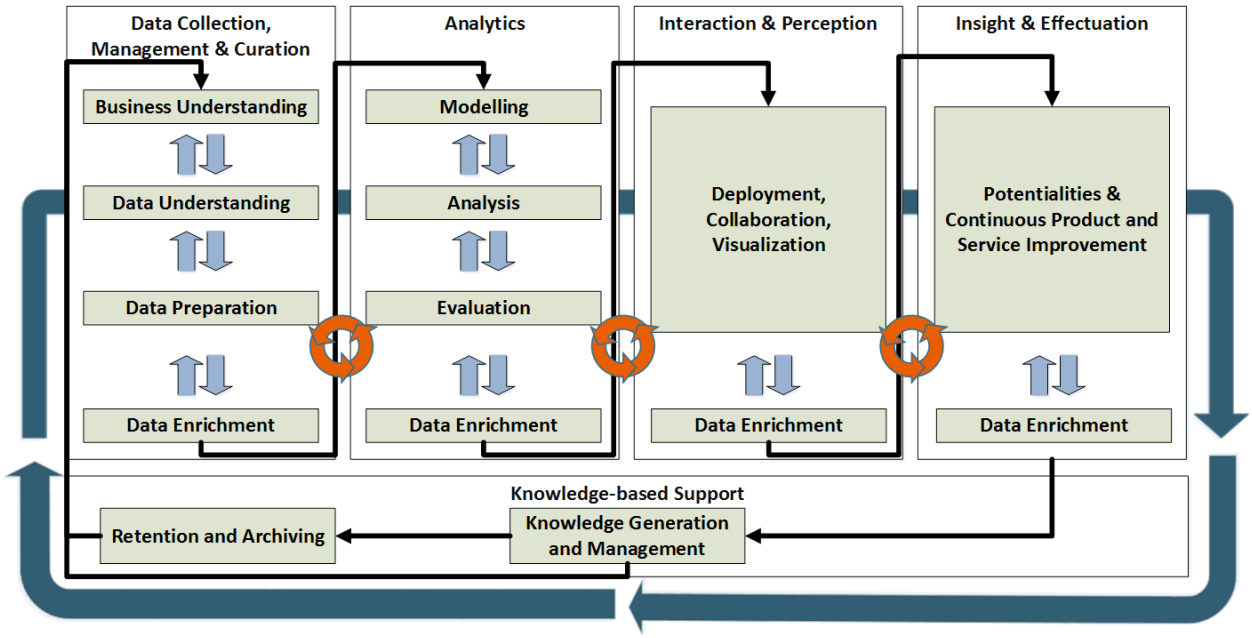


Figure 1: CRISP4BigData Reference Model Version 1.0

Business Understanding describes the goals and problem statements, derived from a project plan to develop a targeted deployment of Big Data Analysis Methods. The process element Data Understanding deals with the collection of internal and external data, the description of data types and source systems. The Data Preparation manages the preparation of the Data over a transformation or Data Cleansing (or update, enlarge, and reduce) to increase the data quality. All Data Enrichment process elements in the CRISP4BigData have the task to enrich the data base with useful Meta Data according to the whole process, to get insight of process information to increase the quality of the process or to update the Knowledge-based Support phase with information such as: Who uses which data or analysis methods? – Who is a knowledge carrier? Where is the data from? Where is the data used?

The Modelling element describes the development and implementation of a statistical or mathematical model based on an adequate data model. The Analysis element is based on the statistical or mathematical model implemented in the Modelling element and describes the deployment of a Big Data analysis method or algorithms to analyse the available data (model). The Evolution rates the result of the analysis and the process. The phase evaluates also the precision, usefulness, novelty, and significance of the result. The Deployment, Collaboration, and Visualization element deals with the deployment, visualization of the analysis result and manages the distribution of the right result to the right persons (collaboration). The Potentialities & Continuous Product and Service Improvement element steers the management of potentialities and cares about a continuous improvement process to work out some new opportunities for the company. The process element Knowledge Generation and Management manages that the generated insights and knowledge do not get lost. The Retention and Archiving step manages a long-term retention and archiving of the data, it manages also the classification of hot, cold and warm data (tiers). Finally, the circulation around the whole phases shows that the CRISP4BigData reference model is not an on-time process. It is partially useful or needed to repeat the whole process to deal with new information, obtained during the project or to improve the whole process.

Summary and Outlook

The evaluation of the offered model will be done by defined use cases part of the EC co-funded project *SenseCare* (SenseCare 2016). The specific use case is based on an unstructured data set of XML-based medical documents (circa 40 gigabyte) and is going to be transferred in a suitable ontology.

Therefore the process model is going to be implement within a cluster setup, based on virtual machines located in the EGI Federated Cloud infrastructure and on local virtual machines located at servers at the University of Hagen. The cluster is based on several Apache and Apache Hadoop components.

The CRISP4BigData will be additional implement with an Apache based workflow-engine to automate and describe the whole analysis process. The description of the analysis process based on CRISP4BigData is going to implement with a graphical user interface.

Acknowledgements & Disclaimer

This publication has been produced in the context of the *SenseCare* project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690862. However, this paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

References

Beath, C., Becerra-Fernandez, I., Ross, J., Short, J. (2012): Finding Value in the Information Explosion. In MITSloan Management Review. Available online at http://sloanreview.mit.edu/article/finding-value-in-the-information-explosion/, checked on 5/25/2015.

Bornschlegl, M. X., Berwind, K., Kaufmann, M., Hemmje M. (2016): Towards a Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments, Proceedings on the International Conference on Internet Computing (ICOMP): 78-81. Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Bornschlegl, M. X., Manieri, A., Walsh, P., <atarci, T., Hemmje. L. H. (2016). Road Mapping Infrastructures for Advanced Visual Interfaces Supporting Big Data Applications in Virtual Research Environments. In Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16), Paolo Buono, Rosa Lanzilotti, and Maristella Matera (Eds.). ACM, New York, NY, USA, 363-367. DOI=http://dx.doi.org/10.1145/2909132.2927471

Carter, P. (2011): Future Architectures, Skills and Roadmaps for the CIO. Edited by IDC. IDC. Available online at http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf, updated on 2011, checked on 7/1/2016.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.; Wirth, R. (2000): CRISP-DM 1.0. Step-by-step data mining guide. Edited by CRISP-DM Consortium. Available online at ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf, checked on 6/9/2016.

Freiknecht, J. (2014): Big Data in der Praxis. Lösungen mit Hadoop, HBase und Hive; Daten speichern, aufbereiten, visualisieren. München: Hanser.

Gartner, Inc. (2015): Gartner - IT-Glassary Big Data. Available online at http://www.gartner.com/it-glossary/big-data/, checked on 8/1/2015.

Kaufmann, M. (2016). Towards a reference model for big data management. Research report, Faculty of Mathematics and Computer Science, University of Hagen, retrieved July 15, 2016 from: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00000583

SenseCare (2016). SenseCare - Sensor Enabled Affective Computing for Enhancing Medical Care, checked on 08/30/2016.

Social Network Support for an Applied Gaming Knowledge Management Ecosystem

Munir Salman [1,4], Jana Becker [2], Michael Fuchs [3], Dominic Heutelbeck [2], Matthias Hemmje [1]
[1] FernUniversität in Hagen, Faculty for Multimedia and Computer Science, Hagen, Germany
[2] Forschungsinstitut für Telekommunikation und Kooperation e.V. (FTK), Dortmund, Germany
[3] Wilhelm Büchner University of Applied Science, Pfungstadt, Germany
[4] Munich Center for the Economics of Aging (MEA)
at the Max Planck Institute for Social Law and Social Policy
e-mail: Munir.Salman@fernuni-hagen.de

Keywords: Social Networking, Knowledge Management and Transfer, Applied Gaming, Digital Ecosystem

The European (EU)-based industry for non-leisure games (Applied Games, AGs or formally called serious games) is an emerging business. As such it is still fragmented and needs to achieve critical mass to compete globally. Nevertheless, its growth potential is widely recognized and even suggested to exceed the growth potential of the leisure games market. To become competitive the relevant applied gaming communities (applied game researchers, developers, customers and players) require support by fostering the generation of innovation potential. The European project Realizing an Applied Gaming Ecosystem (RAGE) (Netvision 2015) is aiming at supporting this challenge. RAGE will help to seize these opportunities by making available an interoperable set of advanced Applied Game (AG) technology assets, as well as proven practices of using asset-based AGs in various real-world contexts. As described in (Salman, Heutelbeck, et al. 2015), RAGE will finally provide a centralized access to a wide range of applied gaming software modules, relevant information, knowledge and community services, and related scientific documents, taxonomies, media, and educational resources within an online community portal called the RAGE Ecosystem. This Knowledge Management Ecosystem (KM-ES) is based on an integrational, user-centred approach of Knowledge Management (KM) and Innovation Processes in the shape of a service-based implementation (Becker et al. 2015). This will support information, User Generated Content (UGC), and knowledge sharing, as well as persistency of social interaction threads and Know-how transfer within Social Networking Sites (SNSs) and Groupware Systems (GWSs) that are connected to the RAGE Ecosystem. In this paper, we will review the integration of several Social Networking Platforms (SNPs) (e.g., LinkedIn and Twitter), as well as Groupware Systems (GWSs) such as the Scientific Publication and Presentation Platforms (SPPs) Mendeley (Netvision 2016b) and SlideShare (Netvision 2016d), along with GitHub software repository (Netvision 2016a) and the StackExchange network (Netvision 2016c) into the RAGE KM-ES. The integration approach of SNPs, SPPs, and GWSs into the RAGE KM-ES based on KM and Innovation Processes was described in (Salman, Fuchs, et al. 2016), (Salman, Star, et al. 2015) and (Salman, Becker, et al. 2016). Furthermore, this paper reviews related authentication, authorization, access, and information integration challenges in the context of the RESTful web service architecture (Benson & Grieve 2016) and (Gusmeroli et al. 2013). Our evaluation demonstrates that such integration of several SNPs and GWSs into the RAGE KM-ES is a feasible framework, on the one hand, to enhance KM and Know-how transfer in an Applied Gaming Ecosystem (AGE) and on the other hand to reduce the fragmentation among applied gaming communities.

References

Becker, J. et al., 2015. Realizing an Applied Gaming Ecosystem: Towards Supporting Service-based Innovation Knowledge Management and Transfer. In Proceedings of the 4th International Conference on Games and Learning Alliance, GALA 2015. GALA Conference. Rome, Italy.

Benson, T. & Grieve, G., 2016. The FHIR RESTful API. In Principles of Health Interoperability. Cham: Springer International Publishing, pp. 349–359. Available at: http://link.springer.com/10.1007/978-3-319-30370-3_19 [Accessed July 28, 2016].

Gusmeroli, S., Piccione, S. & Rotondi, D., 2013. A capability-based security approach to manage access control in the Internet of Things. Mathematical and Computer Modelling, 58(5–6), pp.1189–1205.

Netvision, 2016a. Build software better, together. GitHub. Available at: https://github.com [Accessed July 26, 2016].

Netvision, 2016b. Free reference manager and PDF organizer | Mendeley. Available at: https://www.mendeley.com/ [Accessed July 28, 2016].

Netvision, 2016c. Hot Questions - Stack Exchange. Available at: http://stackexchange.com/ [Accessed July 26, 2016].

Netvision, 2015. RAGE 2016. RAGE. Available at: http://rageproject.eu/ [Accessed February 26, 2016].

Netvision, 2016d. SlideShare.net. www.slideshare.net. Available at: http://www.slideshare.net [Accessed July 26, 2016].

Salman, M., Becker, J., et al., 2016. Enhancing Knowledge Management and Transfer in an Applied Gaming Ecosystem. In The 16th European Conference on Knowledge Management. ECKM 2016.

Salman, M., Fuchs, M., et al., 2016. Integrating Scientific Publication into an Applied Gaming Ecosystem. In Proceedings of the 17th European Conference on Knowledge Management, ECKM 2016. ECKM Conference. Northern Ireland, Belfast: Global Science & Technology Forum (GSTF). Available at: https://www.dropbox.com/s/olhvd0abqmdoqaq/CGAT_Proceeding_Paper_6.pdf?dl=0 [Accessed April 21, 2016].

Salman, M., Star, K., et al., 2015. Towards Social Media Platform Integration with an Applied Gaming Ecosystem. In SOTICS 2015,The Fifth International Conference on Social Media Technologies, Communication, and Informatics. pp. 14–21. Available at: zotero://attachment/573/ [Accessed November 21, 2015].

Salman, M., Heutelbeck, D. & Hemmje, M., 2015. Towards Social Network Support for an Applied Gaming Ecosystem. In Proceedings of the 9th European Conference on Games Based Learning ECGBL 2015. The 9th European Conference on Games Based Learning ECGBL 2015. Steinkjer, Norway, pp. 721–728.

ACKNOWLEDGEMENTS AND DISCLAIMER

This publication has been produced in the context of the RAGE project. The project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 644187. However, this paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Improved Data Centre Network Management Through Software-Defined Networking

J. Sherwin
Department of Computing
CIT – Cork Institute of Technology, Ireland
e-mail: jonathan.sherwin@cit.ie

C. J. Sreenan
Department of Computer Science
UCC – University College Cork, Ireland
e-mail: cjs@cs.ucc.ie

Keywords: Software-Defined Networking, Data Centre Networks, Network Management

Research in the area of Computer Networking has seen a surge in activity in recent years with the advent of Software-Defined Networking (SDN), and with the OpenFlow (McKeown et al., 2008) protocol in particular. SDN is the application of programmatic control over the resources and functions of a network in order to make the network more dynamically configurable to match the requirements of users and applications. OpenFlow is an open, standard protocol that is commonly used to implement SDN between network switches and a centralised controller, allowing the controller to install flow-rules on those switches in order to implement network policy. Network switches then simply forward data packets according to the flow-rules that have been installed.

SDN offers significant potential benefit for all aspects of network management – e.g. configuration, monitoring, performance and utilisation, and security. SDN has been applied in campus, data centre, wireless, sensor and mobile networks. Our interest is in performance, monitoring and analysis of flow-rules in Data Centre Networks (DCNs). Particularly relevant work done by other researchers includes: Planck (Rasley et al., 2014), where DCN switches are monitored for congestion, and flows are re-routed to achieve fine-grained traffic engineering; a study of latency in flow-rule change requests made by the controller being applied within the network switches (Kuźniar et al., 2015); Procera (Hyojoon and Feamster, 2013), which can translate event-driven network policies into flow-rule configurations for switches; NetSight (Handigol et al., 2014), which records information about all packets that match a subset of flow-rules in a network; and OFRewind (Wundsam et al., 2011), where a proxy between controller and switches records and replays sequences of flow-rule management messages. For a more detailed survey of research and identification of open issues in SDN for DCN management, see our technical report (Sherwin and Sreenan, June 2016).

We plan to create a number of components of a flow-rule management system for a DCN. The first component will target reducing flow setup delay caused by the latency of hardware switches in processing flow-rule changes as described in (Kuźniar et al., 2015). It will do this by initially forwarding flows through pre-configured tunnels originating and terminating at software switches, then setting up the appropriate flow-rules on hardware switches to establish a non-tunnelled path, and verifying that path is operational before rerouting the flow to the non-tunnelled path. An emulated testbed environment has been set up as shown in Figure 1. The authors of (Kuźniar et al., 2015) have kindly provided a script to mimic the delays of specific models of hardware switch using a software switch, and we are currently writing additional scripts to verify that we can

reproduce their results in our emulated environment. Once verified, the testbed will be the environment in which we develop and test our flow-rule management components. We plan to extend the testbed by adding real hardware switches to further validate our work.

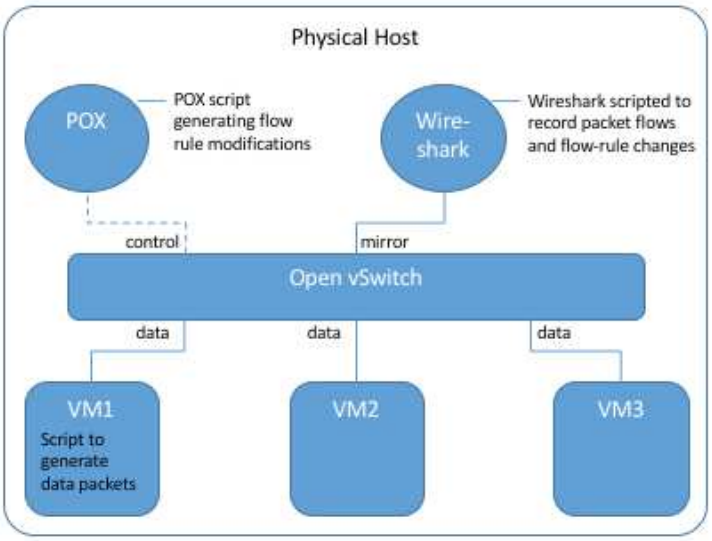


Figure 1: OpenFlow Test-bed Environment

DCNs can be comprised of thousands of switches carrying millions of flows. The flows often carry data belong to multiple tenants of the data centre. The data centre operator will have service level agreements with tenants that include minimum performance criteria. The network usually contains middle-boxes such as firewalls and load-balancers. Most data centres have a high level of server virtualisation, and an important feature facilitated by this is virtual machine (VM) migration. In order to be useful in a DCN, the components we develop will need to be scalable, low-latency, multi-tenant, middle-box friendly, and support VM mobility. To our knowledge, research efforts to date have not addressed all of these requirements for the functionality that our planned components will target.

Our research methodology is to develop prototype implementations targeting specific functionality, and to evaluate them in an emulated environment. This approach is commonly used for computer network research. For some components, we will be able to compare performance with other researchers' results for their attempts at providing similar functionality, for other components we will be providing functionality that has not been available up to now. For some of our planned components, we see opportunities to apply optimisation techniques for better performance, or less resource overheads. For all components, we will verify their usefulness in a DCN based on the criteria listed above.

In conclusion, our work is still at an early stage. We have reviewed SDN research from the last number of years and identified open issues in the area of DCN management. We have selected a few performance, monitoring and configuration issues on which to focus our own efforts. Having created a testbed environment, we are currently verifying it before we implement and evaluate components designed to address our selected issues.

References

Handigol, N., Heller, B., Jeyakumar, V., Mazi, D. and McKeown, N. (2014) 'I know what your packet did last hop: using packet histories to troubleshoot networks', Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation, USENIX Association, pp. 71-85.

Hyojoon, K. and Feamster, N. (2013) 'Improving network management with software defined networking', Communications Magazine, IEEE, 51(2), pp. 114-119.

Kuźniar, M., Perešíni, P. and Kostić, D. (2015). What you need to know about SDN flow tables. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S. and Turner, J. (2008) 'OpenFlow: enabling innovation in campus networks', SIGCOMM Comput. Commun. Rev., 38(2), pp. 69-74.

Rasley, J., Stephens, B., Dixon, C., Rozner, E., Felter, W., Agarwal, K., Carter, J. and Fonseca, R. (2014) 'Planck: Millisecond-scale monitoring and control for commodity networks'. SIGCOMM 2014 - Proceedings of the 2014 ACM Conference on Special Interest Group on Data Communication, pp. 407-418.

Sherwin, J. and Sreenan, C. J. (2016) Software-Defined Networking for Data Center Managment - A Survey and Open Issues Technical Report UCC-CS-2016-24-06, Dept. of Computer Science, University College Cork, Ireland, June 2016.

Wundsam, A., Levin, D., Seetharaman, S. and Feldmann, A. 'OFRewind: enabling record and replay troubleshooting for networks', Proceedings of the 2011 USENIX Annual Technical Conference, USENIX Association, pp. 29-29.

Towards Synchronizing Heterogeneous Data Sources and Information Visualizations

C. Danowski, M. Bornschlegl, B. Schmidt, M. Hemmje
Department of Mathematics and Computer Science
University of Hagen, Germany
Department of Geodesy
Bochum University of Applied Sciences, Germany
e-mail: christian.danowski@studium.fernuni-hagen.de

Keywords: Synchronization-enabled 3D Information Visualization (IVIS), IVIS Reference Model, Mediator-Wrapper Architecture

The modern World Wide Web has evolved to be *the* interface for various aspects of daily life, e.g. to query information (Campesato & Nilson, 2010). Typically, users have a certain Information Need. To satisfy it they execute so-called Information Behaviour, for instance by browsing and analysing relevant information returned by an Information Retrieval system. Eventually, users achieve insight. The whole process of analysing relevant information requires cognitive effort. An Information Visualization (IVIS) system may help to explore and perceive information in a cognitive efficient manner by offering an appropriate visual interface to interact with the data content (Spence, 2014). Particularly, interactive 3D visualizations are a promising solution to enhance user cognition, since navigation in a 3D virtual depiction space resembles navigating 3D space in daily life.

To derive cognitive efficient information visualizations from abstract data, the most proven reference model is the *IVIS Reference Model*, which defines three user-configurable transformations to convert *Raw Data* over *Data Tables* and *Visual Structures* to final *Views*, which act as a visual interface for users (Card et al, 1999). However, when the *IVIS Reference Model* was designed, the *Raw Data* component comprised more or less static data. But what about dynamic *Raw Data*, where the content of derived visualizations may change over time? In modern society and research, content can change rapidly. E.g., from the data point of view, sensor networks or simulation programs may produce new/changing content frequently, which, in turn, would require derived information visualizations to be updated. Concerning the user side, keywords like “Web 2.0” and the “Participative Web” indicate that users demand for more empowerment with respect to application interaction, -configuration as well as data generation and -modification (OECD, 2007). The latter aspects clearly have an impact on the workflow of the IVIS Reference Model. If the user is allowed to modify existing information or even contribute to the generation of new information, this change has to be reflected to the actual data sources storing the information.

With regard to current challenges in the field of Big Data, Bornschlegl et al. (2016) recommend an extension of the IVIS reference model, called *IVIS4BigData Reference Model*, applying the Mediator-Wrapper architectural pattern to semantically integrate data content from multiple heterogeneous data sources, which may then be used as input for the classical *IVIS Reference Model*. Although this extension considers Big Data challenges, it still neglects the explicit consideration of dynamic *Raw Data* and user-triggered data modifications. In consequence, this research approach suggests an explicit synchronization mechanism that keeps data sources and derived information visualizations synchronized and consistent. The existing *IVIS Reference Model*

does not comprise the ability to persist modified visualization content. Hence, our approach proposes an additional refinement of the *IVIS Reference Model*, which is depicted in Figure 1.

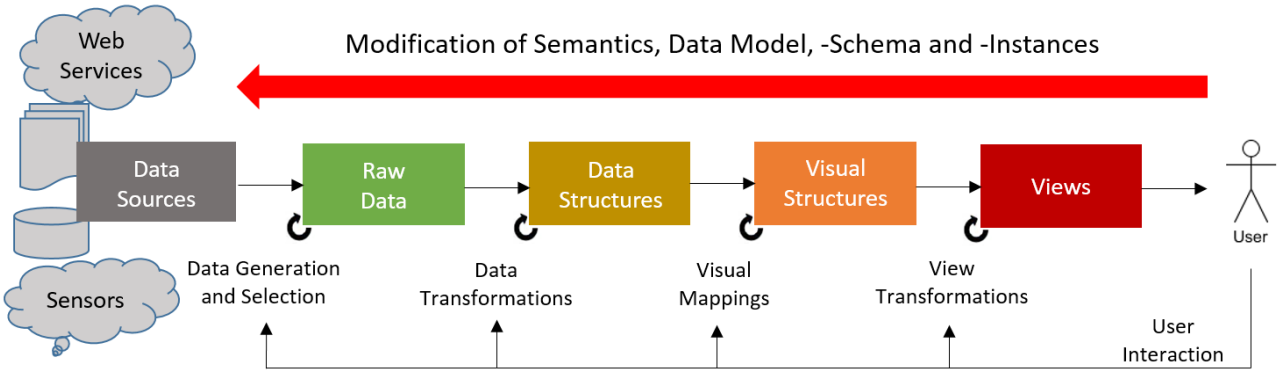


Figure 1 *Extension of the IVIS Reference Model towards dynamic Raw Data and user-executed modification*

Compared to the classical *IVIS Reference Model*, the component *Data Sources* has been added to indicate that there might be multiple heterogeneous data sources, from which dynamic *Raw Data* can be acquired. The *Data Sources* component is accompanied by a new transformation called *Data Generation and Selection*. Content from *Data Sources* might be selected or dynamically generated (e.g. by simulation or computation algorithms) and the output of this transformation is then used as dynamic *Raw Data* input for the classical *IVIS Reference Model*. Another aspect is the enablement of the user to modify several aspects of the visualization pipeline, as indicated by the red backwards arrow placed at the top. Hereby, the focus of the presented research approach is to allow the modification of data instances (e.g. change properties of existing instances or create completely new instances).

Using this augmented IVIS Reference Model as a foundation, the presented research approach aims to develop a prototypical Web-based synchronization-enabled 3D IVIS infrastructure that realizes the following three core functionalities:

- **Server Side Information Visualization:** A user (client) requests server-side application data. During the retrieval process the server transforms the requested (abstract) data into an appropriate cognitive efficient information visualization.
- **Runtime Information Visualization:** The Runtime component comprises any user interactions in an existing visualization. In particular, users may request additional data from within the visualization or modify the displayed data content. In the latter case, these modifications have to be persisted within the server-side data sources.
- **Synchronization:** To maintain a consistent state between data sources and any delivered visualization, any server-side data changes must be broadcast to all affected clients.

Due to the fact, that relevant domain data might be distributed to several heterogeneous (concerning syntax and semantics) data sources, this infrastructure is based on the Mediator-Wrapper architectural pattern that provides a unified access interface for users but manages data from diverse

data sources (Wiederhold, 1992). In accordance to Bornschlegl et al. (2016), these architectural considerations make the IVIS infrastructure future-proof in the context of current Big Data challenges, where relevant data might be distributed over several data sources with different access interfaces. To develop a contemporary prototypical Web-based synchronization-enabled 3D IVIS infrastructure, suitable modern (Web) technologies have been identified, including the WebSocket protocol for full-duplex bidirectional client-server communication (Wang et al., 2013) and ISO-certified 3D real-time format X3D/X3DOM (Web3D Consortium, 2016) (Behr et al., 2009) to offer an interactive client-side 3D visual interface.

The next steps would be to prototypically implement the described prototypical IVIS infrastructure and evaluate it against the mentioned three core functionalities *Server Side Information Visualization*, *Runtime Information Visualization* and *Synchronization* (which will be finished in time with respect to the date of the CERC conference). The described IVIS infrastructure may, in addition, serve as a verification of the *IVIS4BigData Reference Model* proposed by Bornschlegl et al. (2016), since it is based on the Mediator-Wrapper concept and provides a Web-based visualization pipeline to produce cognitive efficient information visualizations.

References

Behr, J., Eschler, P., Jung, Y. & Zöllner, M. (2009), X3DOM: a DOM-based HTML5/X3D integration model, in: *Proceedings of the 14th International Conference on 3D Web Technology*, ACM, 2009; pp. 127-135.

Bornschlegl, M., Berwind, K., Kaufmann, M., Engel, F., Walsh, P., Hemmje, M., Riestra, R. (2016), IVIS4BigData: A Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments, Advanced Visual Interfaces 2016, AVI 2016.

Campesato, O. & Nilson, K. (2010), Web 2. 0 Fundamentals: with AJAX, Development Tools, and Mobile Platforms, Jones & Bartlett Learning.

Card, S., Mackinlay, J. & Shneiderman, B. (1999), Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers.

OECD (2007), Participative Web and User-Created Content Web 2.0, Wikis and Social Networking, OECD Publishing.

Spence, R. (2014), Information Visualization: An Introduction, Springer, Berlin/Heidelberg

Wang, V., Salim, F. & Moskovits, P. (2013), The Definitive Guide to HTML5 WebSocket, Apress.

Web3D Consortium (2016), Why Use X3D. Available at: http://www.web3d.org/x3d/why-use-x3d [Accessed 29.06.2016].

Wiederhold, G. (1992), Mediators in the Architecture of Future Information Systems, IEEE Computer Magazine, Issue 04/1992, pp. 38-49.

Comparison of Machine Learning Algorithms in Classifying Segmented Photographs of Food for Food Logging

Patrick McAllister[1], *Huiru Zheng[1], Raymond Bond[1], Anne Moorhead[2]
[1] School of Computing & Engineering, [2] School of Communication
Ulster University, Shore Road, Newtownabbey
{mcallister-p2@email.ulster.ac.uk}, {h.zheng, rb.bond, a.moorhead}@ulster.ac.uk

Keywords: obesity, food logging, machining learning, classification

Obesity is increasing globally and is a major cause for concern (WHO, 2016). The main cause of obesity is a result of a high calorie/ fat diet and when the energy is not burned off through exercise, then much of the excess energy will be stored as fat around the body. Obesity is a serious threat to an individual's health as it can contribute to a range of major chronic conditions such as heart disease, diabetes, and some cancers (National Institutes of Health, 1998). Food logging is a popular dietary management method that has been used by individuals to monitor food intake. Food logging can include the use of text or images to document intake and research has shown that food intake monitoring can promote weight loss (Wing, 2001).

There has been much research in using computer vision algorithms to classify images of food for food logging. Computer vision methods can offer a convenient way for the user to document energy intake. The motivation for this work is to inform the development of an application that would allow users to use a polygonal tool to draw around the food item for classification. This work explores the efficacy classifying segmented items of food instead of entire food images.

This work explores machine learning (ML) techniques and feature extraction methods to classify 27 food categories with each category containing 100 segmented images. The image dataset used for this work comprises of 27 distinct food categories gathered from other research. (Jontou et al, 2009; Bossard et al, 2014). Non-food items contained in the images were removed to promote accurate feature selection (Figure 1).

Figure 1. Example of segmented food image.

Global and local feature types were extracted from the food image dataset; BoF with Speeded-Up-Robust-Features (SURF), BoF with colour features, and LBP (local binary pattern). SURF and colour features were extracted using bag of features (BoF) method to compute features for each image. A number of ML classifiers were used in this work; Sequential Minimal Optimisation (SMO, PolyKernel), Naïve Bayes (Multinomial), Neural Network (single layer, 100 nodes, 1000 epochs), and Random Forest (300 trees). Combinations of local and global features were used with ML algorithms.

Ten-fold cross validation was used to evaluate each experiment. Percentage accuracy was used to initially assess the performance of each ML algorithm. Matlab (vR2016a) was used to import and extract features from the image dataset and to export feature vectors as a CSV file. Weka (3.7.13) was used to import feature vectors and to apply ML algorithms on the feature sets. A series of classification experiments were completed using BoF with SURF and BoF with colour. The visual vocabulary used in each BoF model was changed (500 visual word increments) in each experiment to record changes in accuracy.

Table 1 and table 2 lists the results of these experiments. Further experiments were completed combining SURF and colour features. SURF and colour feature visual word sizes that achieved the highest accuracy in the previous experiments were concatenated e.g. feature length 500 achieved highest accuracy for neural network classification using colour features, and 500 for SURF, these were combined. Table 3 lists the combination percentage accuracy results.

Visual Words	Naïve Bayes (MN)	SMO	Neural Network	Random Forest
500	32.85*	42.15*	46.70*	43.33*
1000	32.07	41.07	45.48	42.26
1500	31.67	41.22	43.74	39.30
2000	31.67	41.78	43.11	38.96
2500	31.11	41.67	42.96	37.85
3000	30.89	40.89	41.56	37.22
3500	30.85	40.96	42.52	36.67
4000	30.63	40.52	39.78	35.93
4500	30.30	40.63	41.22	35.41
5000	29.85	41.70	42.04	35.41

Table 1. Percentage accuracy results of BoF with Colour features using 10-Fold cross validation.

Visual Words	Naïve Bayes (MN)	SMO	Neural Network	Random Forest
500	44.44	56.22	59.67*	46.11*
1000	45.19*	55.93	57.70	42.81
1500	45.04	56.07	57.00	42.07
2000	44.63	55.85	57.41	41.96
2500	44.33	57.15*	56.19	40.74
3000	44.48	55.89	55.81	40.07
3500	44.26	56.44	56.26	40.22
4000	43.37	56.74	56.44	39.81
4500	43.56	56.22	55.15	40.41
5000	42.96	55.51	55.74	39.00

Table 2. Percentage accuracy results of BoF with SURF features using 10-Fold cross validation.

Feature Combination	Naïve Bayes	SMO	Neural Network	Random Forest
BoF-SURF +BoF- Colour + LBP	50.48	68.29	71.77*	56.77

Table 3. Percentage accuracy results combing BoF colour and SURF features with LBP features.

The experiments focused on using an image dataset that was manually segmented to remove non-item foods or irrelevant food items from the images. Results show that using a Neural Network achieved the highest accuracy with 71.77% accuracy when combining BoF-SURF and BoF-colour features with LBP. Future work will include using other feature selection methods such as segmentation fractal texture analysis (SFTA) (Costa, 2012) gray level co-occurrence matrix (GLCM), and also exploring the use of other ML algorithms such as convolutional neural networks feature extraction and classification, and also multiclass classification methods (one vs one, one vs rest). Attribute selection methods will also be incorporated to select strongest features for classification. As well as using other ML algorithms, more food categories will be added to the food dataset.

References

Who.int. (2016). WHO | Controlling the global obesity epidemic. [online] Available at: http://www.who.int/nutrition/topics/obesity/en/ [Accessed 31 Jul. 2016].

National Institutes Of Health, 1998. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: the evidence report. Obesity Research, 6, p.51S–209S. Available at: http://www.ncbi.nlm.nih.gov/books/NBK2003/.

Wing, R.R. & Hill, J.O., 2001. Successful weight loss maintenance. Annu Rev Nutr, 21, pp.323–341. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11375440.

Taichi Joutou & Keiji Yanai, 2009. A food image recognition system with Multiple Kernel Learning. 2009 16th IEEE International Conference on Image Processing (ICIP), pp.285–288. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5413400.

Bossard, L., Guillaumin, M. & Van Gool, L., 2014. Food-101 - Mining discriminative components with random forests. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 446–461.

Costa, A.F., Humpire-Mamani, G. & Traina, A.J.M., 2012. An efficient algorithm for fractal analysis of textures. In Brazilian Symposium of Computer Graphic and Image Processing. pp. 39–46.

Chapter 5

Bioinformatics

Towards a molecular diagnostic for bacterial infection in newborn babies

Forster T, Dantoft W, Kropp K, Dickinson P, Smith CE, Walsh P, Ghazal P
Division of Infection and Pathway Medicine, University of Edinburgh, UK
NSilico Life Sciences, Cork, Ireland
e-mail: thorsten.forster@ed.ac.uk

Keywords: bacterial infection, host immune response, microarray, sequencing, software

Sepsis is a life-threatening immune response to infectious agents, in severe cases the unconstrained immune system will escalate to a 'cytokine storm' causing organ failure and death. Not much is known about the neonatal immune system (Liu et al., 2012; Madsen-Bouterse et al., 2010), and in clinical practice, newborn babies (neonates) are treated with broadband antibiotics on suspicion of bacterial infection. This poses risks if no actual bacterial infection is present, and current diagnostics (clinical symptoms, blood culture assays) are slow und of low sensitivity with respect to neonatal treatment practice.

We hypothesized that the human host immune response at the level of transcribed genes can be used to develop a quick and reliable diagnostic assay that identifies from patient blood if a bacterial infection is present.

To develop this approach beyond the level of simply looking at a gene list, translating an assay into clinical use requires interdisciplinary expertise in clinical medicine, host and pathogen biology, data science and computing software and infrastructure. In response, the ClouDX-I (EU FP7 IAPP) project was initiated to support this multidisciplinary approach. We employ a suite of biomedical and computational approaches to address our hypothesis, broadly, this involves the use of microarray transcription arrays, next generation sequencing, and multiple analytical and computational solutions. Specifically, we first generate a clinical study for collecting data on blood samples of ~300 neonates and use microarray transcription array technology to develop a gene-expression based classifier that can correctly predict presence of bacterial infection. We validate this classifier on independent datasets for neonates and bacterial infections in other age groups and ethnicities. We also validate this classifier on an assay platform closer to clinical application (QuantiGene). In parallel to this host-specific research, we use next generation sequencing on whole-blood derived clinical isolates of bacteria from individual patients to determine genomic links between host immune response and pathogen. We separately investigate gene expression immune responses to bacterial infection in human neonates to those in mice in order to establish if a mouse pup model system for bacterial infection can enable further biomedical intervention studies. Using unexpected outcomes from the original neonatal microarray study data, we also investigate if host responses to bacterial infection can be distinguished not only from healthy controls, but also from host responses to viral infections. Analytically and computationally, we investigate requirements for and performance of machine learning algorithms in diagnostic classification; we also develop the use of cloud computing and next generation sequencing pipelines for enabling this research.

We established the computational requirement for the number of gene predictors to be between 4 and 19 for good predictions (Khondoker et al., 2010), and that the use of a simple ROC based classifier combined with a z-score transformation of all individual studies allows for classification to generate predictions across multiple studies and platforms. We identified a 52-gene classifier of innate immune, adaptive immune and metabolism-related genes that can predict with very high sensitivity and specificity the presence of a bacterial infection from RNA circulating in whole blood (Smith et al., 2014). This is true for both the study population used in generating the classifier as well as independent samples collected at a later time and in unrelated research studies with different patient demographics. We also demonstrate good performance of this classifier gene set on QuantiGene assays containing only the predictive genes in question (as opposed to whole genome microarrays). We were able to establish that a mouse pup model for bacterial infection may be suitable for researching bacterial infection in neonatal humans, with the mouse pup model capturing the innate immune and metabolic response of neonates, although not an adaptive immune response (Dantoft et al., 2016). Following up unexpected results from the original neonatal microarray study, we were able to identify 8 potential viral infection cases in our study population, previously undetected by clinical diagnosis and providing a potential gene signature for neonatal immune responses to viral infection. Using infrastructure and software developed for this research (Walsh et al., 2013), sequencing analysis of bacteria obtained from whole blood of neonates generated draft genome sequences of six different *Staphylococcus epidermis* clones (Walsh et al., 2015) and strain differences currently being investigated for mutation within a clinical ward setting.

In conclusion, we have identified and independently validated a promising host-based classifier gene set for clinical diagnosis of neonates presenting with symptoms of infection and early signs of sepsis. In parallel to generating a biomedical research outcome, we have also identified laboratory, analysis and computational processes that aid in establishing a clinical outcome from the base ingredients of gene transcription and sequencing study data. However, to complete our research objectives, we are now (a) pursuing a large UK neonatal multi-centre study to evaluate the clinical use of this classifier, and (b) in further academic-industry collaboration bundle up diverse approaches into a methodological and software pipeline that can be used for generating diagnostic tests for other human diseases.

References

DANTOFT, W., TOROW, N., FORSTER, T, CRAIGON, M., PÄGELOW, D., ZHANG, K., WALSH, P., FULDE, M., SMITH, C.L., DICKINSON, P., HORNEF, M.W. and GHAZAL, P. 2016. Recapitulation of human neonatal immune signatures in newborn infected mice. Nature Medicine [In Review]

KHONDOKER, M. R., BACHMANN, T. T., MEWISSEN, M., DICKINSON, P., DOBRZELECKI, B., CAMPBELL, C. J., MOUNT, A. R., WALTON, A. J., CRAIN, J., SCHULZE, H., GIRAUD, G., ROSS, A. J., CIANI, I., EMBER, S. W., TLILI, C., TERRY, J. G., GRANT, E., MCDONNELL, N. & GHAZAL, P. 2010. Multi-factorial analysis of class prediction error: estimating optimal number of biomarkers for various classification rules. J Bioinform Comput Biol, 8, 945-65.

LIU, L., JOHNSON, H. L., COUSENS, S., PERIN, J., SCOTT, S., LAWN, J. E., RUDAN, I., CAMPBELL, H., CIBULSKIS, R., LI, M., MATHERS, C., BLACK, R. E. & UNICEF, C. H. E. R. G. O. W. A. 2012. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet, 379, 2151-61.

MADSEN-BOUTERSE, S. A., ROMERO, R., TARCA, A. L., KUSANOVIC, J. P., ESPINOZA, J., KIM, C. J., KIM, J. S., EDWIN, S. S., GOMEZ, R. & DRAGHICI, S. 2010. The transcriptome of the fetal inflammatory response syndrome. Am J Reprod Immunol, 63, 73-92.

SMITH, C. L., DICKINSON, P., FORSTER, T., CRAIGON, M., ROSS, A., KHONDOKER, M. R., FRANCE, R., IVENS, A., LYNN, D. J., ORME, J., JACKSON, A., LACAZE, P., FLANAGAN, K. L., STENSON, B. J. & GHAZAL, P. 2014. Identification of a human neonatal immune-metabolic network associated with bacterial infection. Nat Commun, 5, 4649.

WALSH, P., BEKAERT, M., CARROLL, J., MANNING, T., KELLY, B., O'DRISCOLL, A., LU, X., SMITH, C., DICKINSON, P., TEMPLETON, K., GHAZAL, P. & SLEATOR, R. D. 2015. Draft Genome Sequences of Six Different Staphylococcus epidermidis Clones, Isolated Individually from Preterm Neonates Presenting with Sepsis at Edinburgh's Royal Infirmary. Genome Announc, 3.

WALSH, P., CARROLL, J. & SLEATOR, R. D. 2013. Accelerating in silico research with workflows: a lesson in Simplicity. Comput Biol Med, 43, 2028-35.

Metamorphosis: Changing the shape of genomic data using Kafka

B Lawlor, P Walsh
Department of Computer Science
CIT – Cork Institute of Technology, Ireland
e-mail: brendan.lawlor@mycit.ie

Keywords: Bioinformatics, Big Data, Cloud Computing

A bioinformatic researcher, when seeking novel ways to process genomic data, will naturally look for a standarad reference genomic database to test their new ideas against. Perhaps the best-known genomic reference is NCBI's Blast[1]. The word 'Blast' can refer both to the database and to the search algorithm used to access it, and in fact this ambiguity highlights an underlying problem for the aforementioned bioinformatic researcher. The shape of the Blast genomic data is set by the Blast search algorithm. To make this data available in its basic form – for example to researchers who wish to test other search algorithms or process the entire database contents - it must be downloaded, processed and reformatted into flat FASTA files. In either form – algorithm-specific proprietory format or more than 800GB of text files - the genomic reference data is not in good shape.

Genomic data is big data. It has high volume and increasingly high velocity (more database entries, more query sequences). The format in which that data is maintained matters. It should be robust, scalable and amenable to parallelized access. It should also be available in a way that stimulates innovation. If it is only available through BLAST, then it will only be used in ways that BLAST permits. If instead it is open to exhaustive, streamed, distributed and parallelized access, then novel applications of the data will result.

This paper explores the use of Apache Kafka[2] - a technology that was not conceived as a database but as a messaging system - for general purpose, efficient, distributed, stream-ready storage and retrieval of genomic data. It offers scalable, durable, distributed storage, with high-speed and flexible access. We consider how this technology could emerge as a primary format for storing genomic sequences, directly accessible by general purpose programs, but from which secondary algorithm-specific formats could be derived where needed. We present a Proof of Concept implementation of a Kafka Genomic Database and measure its properties.

The wider software community is exploring improved ways of storing big data that avoid the bottlenecks of centralized transactional databases. A central feature of such systems is the re-interpretation of system state as 'events'. 'Event sourcing'[3] is an increasingly popular and scaleable solution to maintaining the state of a system without creating database bottlenecks[4]. Reference genomes can be seen as the system state of many bioinformatics applications, and a Kafka-based log of the adds, updates and deletes of genomic sequences could offer a powerful implementation of this concept.

It is interesting to note that Event Sourcing as described above is analogous to transaction logs – the history of actions maintained by traditional database management systems to guarantee ACID properties[5].

References

[1] Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

[2] Wang, Guozhang, et al. "Building a replicated logging system with Apache Kafka." Proceedings of the VLDB Endowment 8.12 (2015): 1654-1655.

[3] Martin Fowler. Event sourcing. http://martinfowler.com/eaaDev/EventSourcing.html, 2005.

[4] Betts, Dominic, et al. "Exploring CQRS and Event Sourcing: A journey into high scalability, availability, and maintainability with Windows Azure." (2013).

[5] Gray, Jim, and Andreas Reuter. Transaction processing: concepts and techniques. Elsevier, 1992.

A Generalized Multi-Network Framework for Disease Gene Progression Prioritization

Fiona Browne [1], Haiying Wang [1], Huiru Zheng [1], Juan Carlos Leyva López[2], Jesús Jaime Solano Noriega [2]
[1] School of Computing and Mathematics
Ulster University, Northern Ireland
[2] Universidad de Occidente
Blvd. Lola Beltrán y Blvd. Rolando Arjona, Culiacán, Sinaloa, México
e-mail: f.browne@ulster.ac.uk

Keywords: outranking methods, multi-objective evolutionary algorithms, multi-network analysis

Background

Identification of disease progression and disease genes associated with physiological disorders is a fundamental task in the analysis of complex diseases. Diseases are often polygenic and multifactorial in nature and can present with different genetic perturbations in patients [[1]]. Recent years have witnessed the systematic investigation of complex disease through the application of high-throughput experimental technologies and the development of centralized databases. These experimental "omic" platforms target the comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). This research has been essential in (i) advancing the knowledge of biological systems (2) understanding, predicting, diagnosing and monitoring diseases (3) discovering biomarkers and (4) identifying drug targets. However, translation of relevant prognostic markers identified by such experiments into clinical tools for personalized patient treatment has been slow due to issues such as the reproducibility and validity of findings across studies, unfocused study design and inappropriate selection and application of statistical techniques (Dupuy & Simon 2007) (Simon 2005). In addition, experimental validation of large lists of candidate genes lists returned from experimental techniques are time-consuming and expensive with data obtained from individual platforms insufficient to fully characterize complex biological systems. The challenge we now face is the development of efficient methodologies to prioritize these gene lists and address the limitations of single platforms. To this end, combining experimental results from multiple "omic" platforms can help to uncover latent biological relationships evident only through integration of measurements across multiple biochemical domains (Wanichthanarak et al. 2015) (Hoadley et al. 2014). Network analysis has proved to be a very effective approach in modelling complex connections among diverse types of cellular components such as genes and metabolites.

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686785/#B1

Proposed Framework

In this research, we present a generalized framework (Figure 1), which integrates diverse "omic" data such as mRNA, gene co-expression, Gene Ontology (GO) semantic similarity and tissue-specific data using a consensus network combination approach. A Case Study based on Ovarian Cancer (OV) disease progression is presented to demonstrate the application of the framework. OV is associated with substantial mortality and morbidity, this disease is difficult to detect at its earliest stage, therefore, uncovering gene signatures at an early stage is an essential task. The framework as illustrated in Figure 1 consists of a number of stages. (A) extracting RNA-Seq OV gene expression and clinical information, (B) identification of differentially expressed genes (DEGs) between early stage and late stage tumour samples, (C) construction of DEG septic networks, (D) analysis of the DEG specific networks to identify stage associated hub genes and their interactors, (E) novel outranking-based algorithm. The method is based on ELECTRE III (Roy 1991) to create a fuzzy outranking relation and then, it uses a multi-objective evolutionary algorithm (López et al. 2016) to exploit the outranking relation and to derive a prioritized list of candidate disease genes. (F) evaluation of hub genes including the development of a hub gene-based classier model to distinguish OV stages.

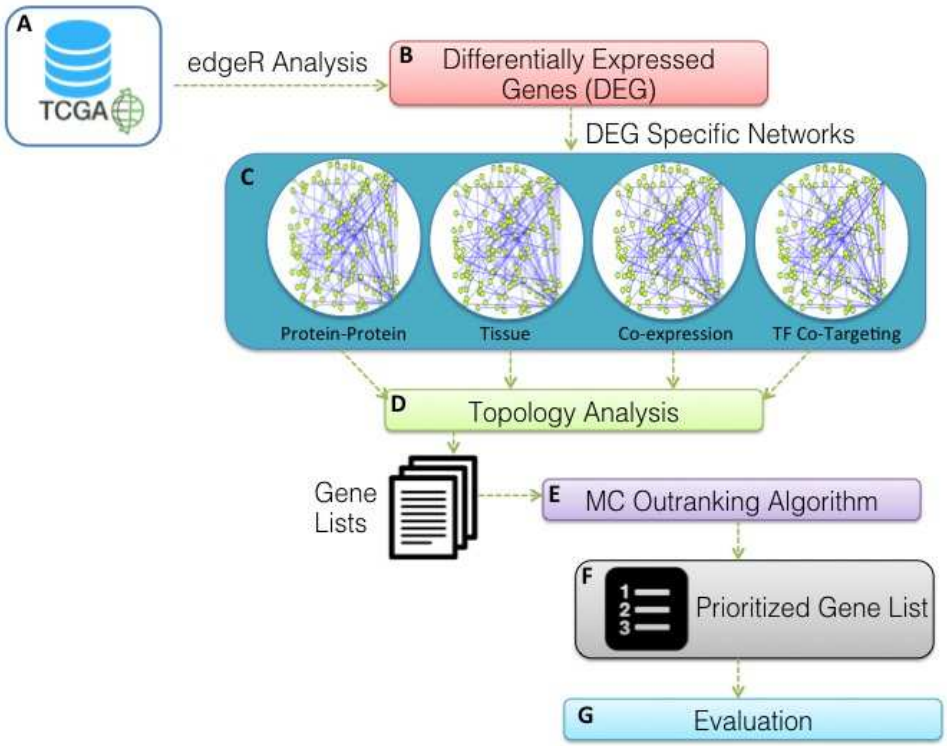


Figure 1, Overview of the Generalised Framework For Disease Progression Gene Prioritization

Preliminary Results

The proposed framework integrated data from four networks: protein-protein, co-expression, tissue and transcription factor co-targeting. Topological analysis based on network degree and betweenness was performed for the individual networks. Lists of genes with high degree and betweenness were selected from each network and then ranked using the novel outranking-based algorithm. A total of 11 genes were identified using the ranking approach. Using the Online Mendelian Inheritance in Man (OMIM) Morbid Map dataset, which lists diseases and gene associations from published studies. A total of 35 known ovarian genes were identified of which 2 genes from the ranked list namely Erbb2 and BRAC1 both of which have been implicated in the carcinogenesis and prognosis of ovarian cancer. Gene Ontology (GO) and pathway analysis was applied to investigate the biological implications of the integrated prioritized disease gene candidate list. Using the GO Biological Process and the DAVID system (Huang et al. 2008) a statistical over representation test was performed using the prioritized genes ($p < 0:05$ Bonferroni corrected), 79 GO processes were identified including ovulation cycle and ovulation cycle process. Furthermore, 16 significant KEGG pathways were identified including endometrial cancer, MAPK signaling pathway, and pathways in cancer. These initial results demonstrate the biological significance of a network based approach in integrating “omic” data.

Conclusions

Network analysis is an effective approach in modeling the complexity of human disease. Furthermore, networks have been instrumental in uncovering how complexity controls disease manifestations, prognosis, and therapy. Our initial studies have illustrated how applying a network approach, diverse “omic” data can be integrated and provides encouraging results for biologically important tasks such as ovarian cancer disease gene prioritization.

References

Dupuy, A. & Simon, R.M., 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute, 99(2), pp.147–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17227998 .

Hoadley, K.A. et al., 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4), pp.929–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25109877.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1), pp.44–57. Available at: http://www.nature.com/doifinder/10.1038/nprot.2008.211 .

López, J.C.L. et al., 2016. Exploitation of a Medium-Sized Fuzzy Outranking Relation Based on Multi-objective Evolutionary Algorithms to Derive a Ranking. http://dx.doi.org/10.1080/18756891.2016.1204122.

Roy, B., 1991. The outranking approach and the foundations of electre methods. Theory and Decision, 31(1), pp.49–73. Available at: http://link.springer.com/10.1007/BF00134132

Simon, R., 2005. Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. Journal of Clinical Oncology, 23(29), pp.7332–7341. Available at: http://www.jco.org/cgi/doi/10.1200/JCO.2005.02.8712

Wanichthanarak, K., Fahrmann, J.F. & Grapov, D., 2015. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomarker insights, 10(Suppl 4), pp.1–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4562606&tool=pmcentrez&rendertype=abstract

Measuring the Matches: GENOME ALIGNMENT BENCHMARKING

Xiangwu Lu [1,*], Audrey M.Michel [2], Pavel V. Baranov [2] and Paul Walsh [3,*]
[1] Department of Computing, Cork Institute of Technology, Bishopstown, Cork, Ireland
[2] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland
[3] NSilico Life Science, Bishopstown, Cork, Ireland
Email: xiang.lu@mycit.ie, paul.walsh@nsilico.com

Keywords: bioinformatics genome, benchmarking

Sequence alignment is a key bioinformatics technique for identifying those sequences of DNA, RNA, and proteins that are similar due functional, structural, or evolutionary relationships between the sequencesi. Of interest in this paper is the identification of messenger RNA (mRNA), which is type of RNA molecules that conveys genetic information from DNA to molecular machinery known as ribosomes, where they specify the amino acid sequence of the protein products of gene expression. Being able to identify mRNA accurately is a key step for understanding a wide range of cellular processes. Of particular interest is the identification of ribosomal protected mRNA fragments which are typically 30 base pairs long (ribo-seq). However, most software tools that have been developed for alignment focus on identifying much longer sequences of nucleotides.

Therefore in the absence of suitable benchmarks, the choice of tool for detecting much shorter ribosomal protected RNA is haphazard. This study will provide comprehensive analysis of alignment tools for short sequences and will serve as a useful guide for researchers on the best techniques to use for short read based research projects. We consider the alignment of ribo-seq data to a human genome of interest, for both simulated and real ribo-seq data. For the assessment of accuracy and throughput using simulated ribo-seq reads, we separated our genome alignment benchmarking into intra-exon mapping and splice junction mapping which we present separately. For real ribo-seq data we focused on the distribution of the aligned reads to the different genomic regions where the expectation is that the majority of reads should be aligned to CDS regions.

For the intra-exon benchmarking, we sampled 30 bp sequences that were fully contained within exons. The intra-exon benchmarking differs from our transcriptome benchmarking with respect to the indexes, strand specificity and reference sequence size. Fifteen short read alignment tools were used in the analysis and the overall results are recorded.

None of the tools mapped 100% of the simulated ribo-seq reads to their original sampled location when there were 0 single nucleotide polymorphisms (SNPs [1]) in the reads and 0 mismatches allowed. Many of the tools, however, did have 99.9% accuracy under these conditions: bowtie, bowtie2, gsnap, rum and tophat.

[1] Single nucleotide polymorphisms are the most common type of genetic variation among organisms. Each SNP represents a difference in a single DNA building block, called a nucleotide.

For some tools there is a strong correspondence between the number of single nucleotide polymorphisms (SNPs) and a mismatch constraint. Bwa provides a good example of this: 95% of the simulated query reads were aligned to the original sampled locations when there were 0 SNPs and 0 mismatches permitted. However, no query reads were aligned when there was 1 SSNP in the reads and 0 mismatches permitted. Compare these results to those of soap2 where ~ 94% of reads were accurately mapped to their original sampled location when 1 SSNP was in the reads but the mismatch allowed parameter was set to 0. With respect to the mapping of reads with INDELs, smalt, bwa, star and tophat had the highest mapping accuracy in that order.

One aspect we wished to investigate was the time required to align different size datasets, i.e. the scalability of the tools. Many of the tools had a similar throughput when mapping a dataset of 1 million reads compared to 10 million reads which indicates that the tools scale up reasonably well. Also, the majority of the tools do not exhibit any great differences in throughput as the number of SNPs/INDELs and mismatches allowed are varied.

Purely in terms of alignment throughput, bfast, bwa and soap2 had the highest throughput, although their performance needs to be examined with respect to accuracy (see section Genome intra-exon alignment accuracy and throughput). Note that while we recorded the throughput of each tool, we did not record the memory footprint of each tool. Indeed a tool may achieve higher throughputs at the expense of higher memory usage. In practice the amount of memory space used by a tool, particularly for large genomes, may be an important consideration.

Overall, bowtie (-v mode) had the highest combined accuracy and throughput for alignments of simulated ribo-seq reads to the intra-exon only regions of the human genome when 0 or 1 mismatches were allowed. When INDELs were introduced into the reads, bwa achieved the highest accuracy and highest throughput.

The numbers of TPs, FPs and FNs for each tool for our genome intra-exon benchmarking are provide. Under 0 SSNP and 0 mismatches conditions, bowtie (-v mode), bwa, gsnap, rum and tophat all had very high true positive rates with 0% FPs and 0% FNs, with bowtie (-v mode) topping the list in terms of TPs.

The alignment of ribo-seq data to a eukaryotic genome with alternative splicing poses additional challenges. In particular, we wished to assess the accuracy of aligning short 30 bp sequences that span splice junctions. Indeed, many of the tools that we benchmarked for transcriptome alignment do not have the functionality for mapping across splice junctions. As only 4 of the tools (gsnap, rum, star and tophat) mentioned splice junction mapping capability in their documentation, we carried out our splice junction benchmarking on these 4 tools only and the overall results were recorded.

Overall the results of our study provides a comprehensive guide for researchers on the best tools to use for short read alignment to reference genomes.

Additional Material

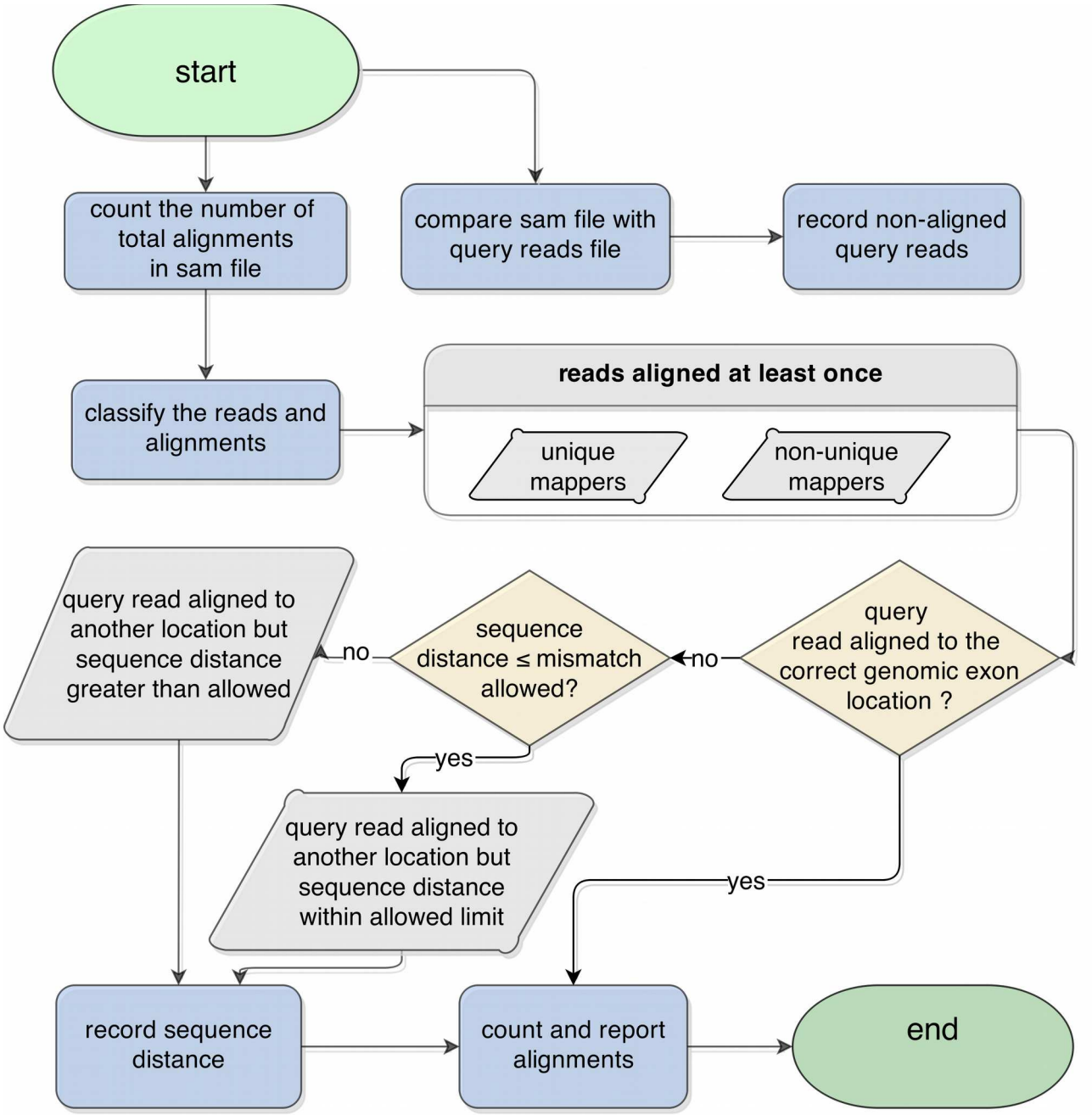


Figure 1. The workflow designed for genomic exon benchmarking

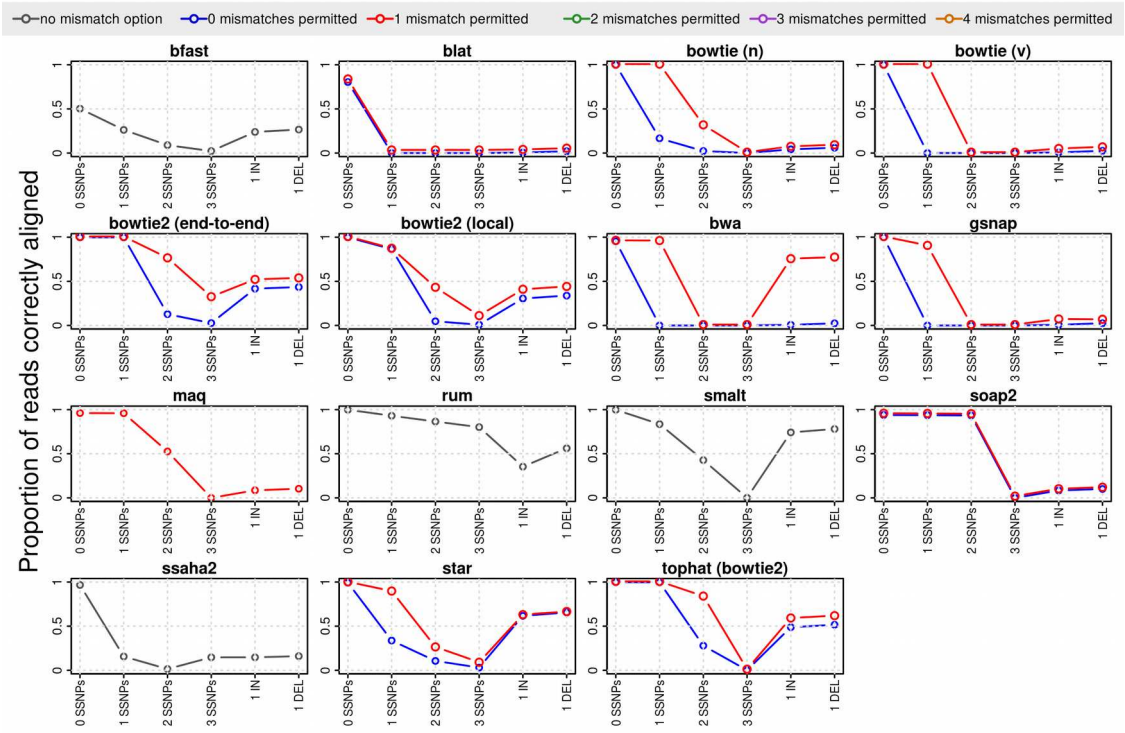
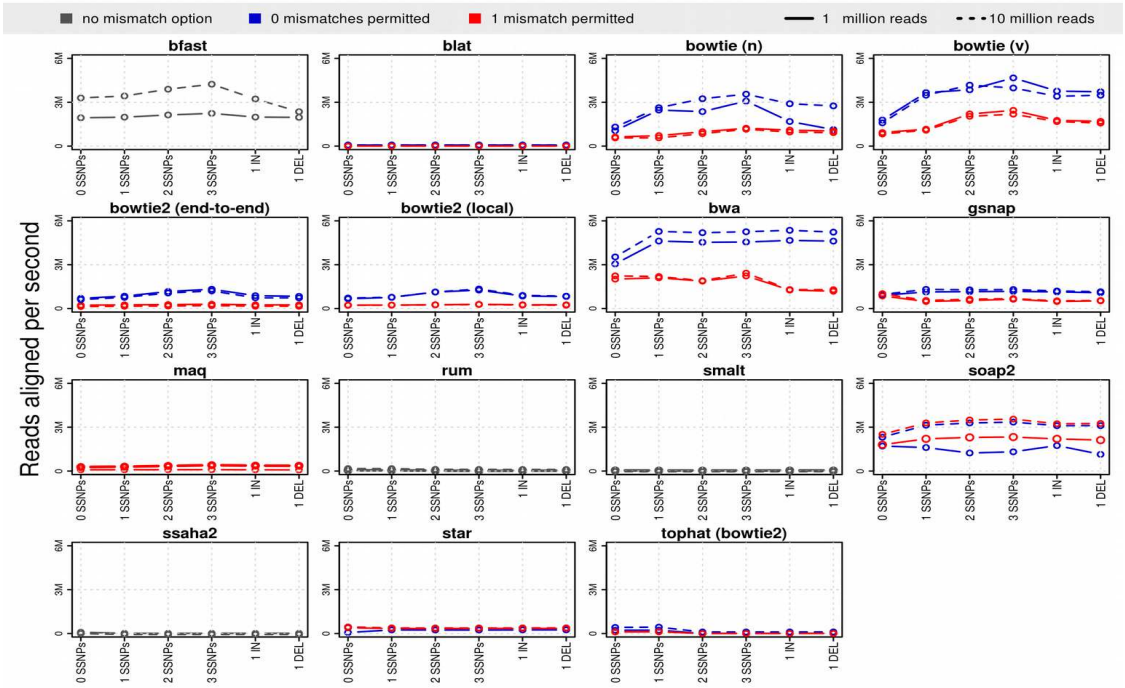
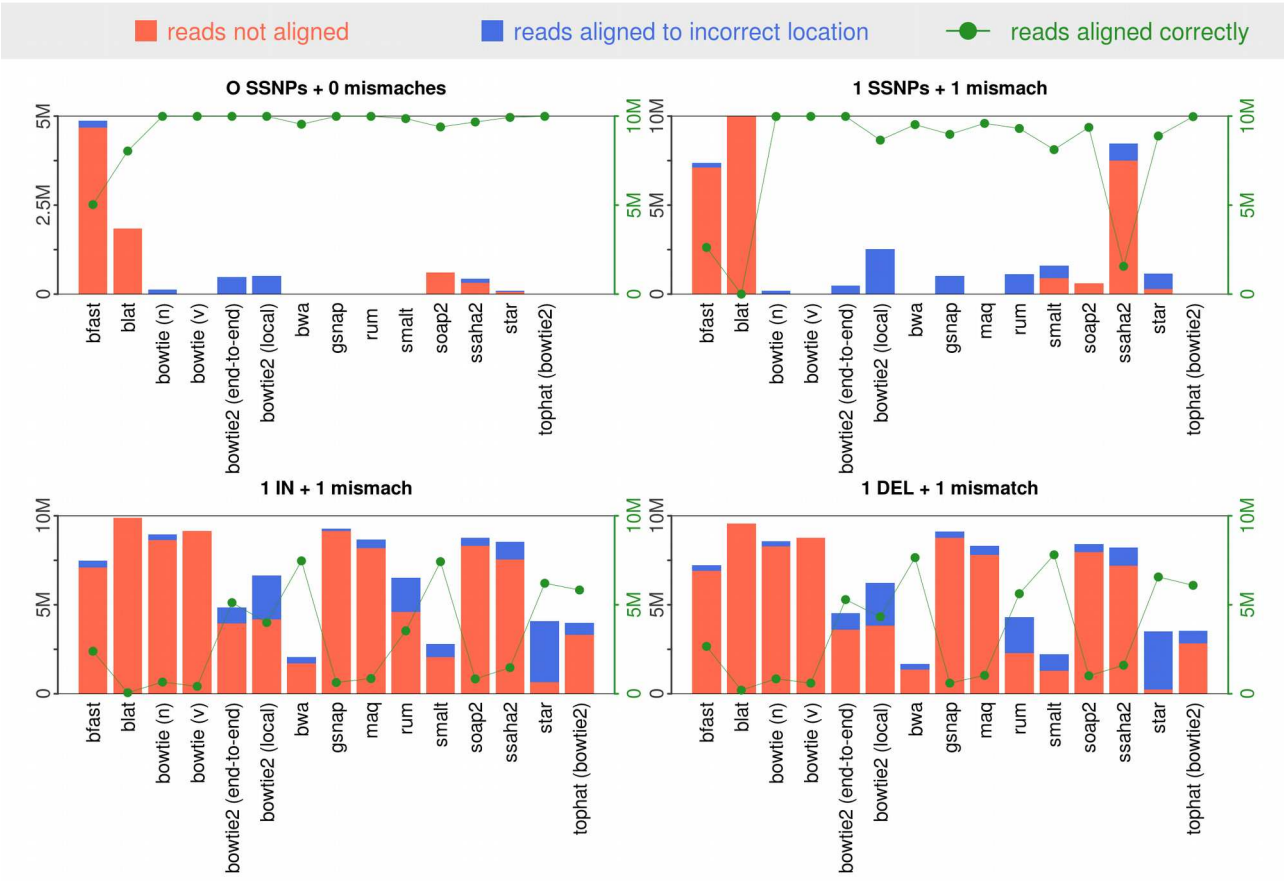


Figure 2. The accuracy of each short read alignment tool when evaluated for genome exon alignment using different SSNP/INDEL and mismatch metrics (See Additional Table S18 for details)



Additional Figure S3. The comparison of the genome alignment throughput of each short read alignment tool. The throughput when mapping a dataset of 1 million reads compared to mapping a dataset of 10 million reads when evaluated for genome exon alignment using different SSNP/INDEL and mismatch metrics (see Additional Table S19 for details). "M" in the Y-axis indicates million.

Additional Figure S4. The accuracy and throughput of each short read alignment tool when evaluated for genome exon alignment using different SSNP/INDEL and mismatch metrics (See Additional Table S20 for details). The size of the pie chart is proportional to the throughput achieved by the alignment tool while the green slicing of the pie chart represents the alignment accuracy.



Additional Figure 5. The true positives (TPs), false positives (FPs) and false negatives (FNs) of each short read alignment tool when evaluated for genome exon alignment using different SSNP/INDEL and mismatch metrics (See Additional Table S21 for details). The green lines represent the number of reads out of 10 million that were mapped to the original sampled location (TPs). The orange bars represent the number of reads not aligned (FNs) and the blue bars represent the number of reads aligned to another location where the sequence distance was greater than the mismatch allowed metric (FPs – see "Alignments to locations other than the original sampled location" in Methods). Note that in the plot of "0 SNPs + 0 mismatches", the scale of left Y- axis for the bars is from 0 to 5 million and other Y-axis scales are from 0 to 10 million.

References

Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, 218–23

Weiss,R.B. and Atkins,J.F. (2011) Molecular biology. Translation goes global. Science, 334, 1509–10.

Michel,A.M. and Baranov,P. V (2013) Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip. Rev. RNA, 4, 473–90.

Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet., 15, 205–13.

Steitz,J.A. (1969) Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature, 224, 957–64.

Lareau,L.F., Hite,D.H., Hogan,G.J. and Brown,P.O. (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments.Elife, 3, e01257.

O'Connor,P.B.F., Li,G.-W., Weissman,J.S., Atkins,J.F. and Baranov,P. V (2013) rRNA:mRNA pairing alters thelength and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics, 29, 1488–91.

Guydosh,N.R. and Green,R. (2014) Dom34 rescues ribosomes in 3' untranslated regions. Cell, 156, 950–62.

Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63.

Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851–8.

Dobin,A., Davis,C. a, Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R.

Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7.

Funding

Cork Institute of Technology [Rísam PhD Sholarships and MetaPlat [RISE EU H2020 project to XWL]; Science Foundation Ireland [12/IA/1335 to AMM and PVB]; and the ClouDx-i [MetaPlat [RISE EU H2020 project to PW].

The Moorepark Grass Growth model, an user-friendly model for farmer and researcher

E. Ruelle [1], D. Hennessy [1] L. Shalloo [1], P.Walsh [2], T. Manning [2], S. Wietrich [2] and L. Delaby [3]
[1] Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, Co. Cork;
[2] Department of Computer Science CIT – Cork Institute of Technology, Ireland
[3] INRA, AgroCampus Ouest, UMR 1348, Physiologie, Environnement and Génétique pour l'Animal et les Systèmes d'Elevage, F-35590 Saint-Gilles, France
e-mail: elodie.ruelle@teagasc.ie

Keywords: model, grass growth, graphical interface

In temperate climates grass provides the cheapest and highest quality feed for dairy cows (Dillon et al., 2005). However, even though a temperate climate allows grass growth throughout the year, grass growth is highly seasonal and depends heavily on climatic conditions. Management of pasture (such as fertilisation, grazing severity) is also an important factor influencing grass growth. Decision support models and tools can provide the farmer with information to better manage his grass and so increase grass growth and grass utilisation. However to be used by a farmer, models and model inputs have to be as simple as possible and must be assessable through a user-friendly graphical interface. This paper presents the Moorepark Grass Growth model (MGGM) a model capable of predicting grass growth and nitrogen utilisation across farms in Ireland.

Material and Methods

The MGGM is a dynamic mechanistic model developed in C++ describing the grass growth and the N fluxes of a paddock. The model is run with a daily time step simulating soil N mineralisation, immobilisation, water fluxes, grass growth, N uptake and grass N content. The model is driven by a daily potential growth curve depending on the radiation and the total green biomass. To calculate the actual daily growth, this potential growth is then multiplied by parameters depending on environmental conditions (temperature, water in the soil and radiation) and a parameter depending on the availability of the mineral N in the soil compared to the demand for N associated with the potential grass growth. The availability of the N in the soil depends on the mineral N in the soil, the proportion of the N usable by the plant (depending on the time of the year and the heading date) and the N demand to grow one kg of biomass. The N dilution curve represents the decrease of the N needed for one kg of DM growth with the increase of the accumulated biomass. Animal deposition of nitrogen is randomised on the surface of the paddock in the form of nitrogen and faeces and is simulated through a one day-grazing sequence.

Different simulations have been run to show the impact of different mineral N fertilisation in grazing conditions across different weather years. These simulations have been tested on 2 different weather conditions in Ireland (Co. Cork) and in France (Normandy). Two different mineral N applications were tested 0 or 300 kg of N per ha per year in Ireland. Due to the lack of rain and high temperature in summer in Normandy, the annual mineral N fertilization has been reduced by 25% through the removal of Autumn fertilisations. The size of the paddock is 1 ha and a total of seven (France) or eight (Ireland) grazing events have been simulated.

Result and discussion

The results are presented in Figure 1. The average grass grown by the animal at 0 kg of N has been of 6,014 kg of DM in the French weather and 8,388 with the Irish weather. When at 300 kg of N the grass grown go at 8,710 for the French weather and 13,452 for the Irish weather. Consequently, the model is capable of reacting to different weather conditions and is highly sensitive to different levels of mineral N fertilisation (Figure 1). The lack in rainfall in France in 2008 led to a poor growth in summer and autumn. The response to N mineral level of fertilisation conforms with the literature (Whitehead, 1995).

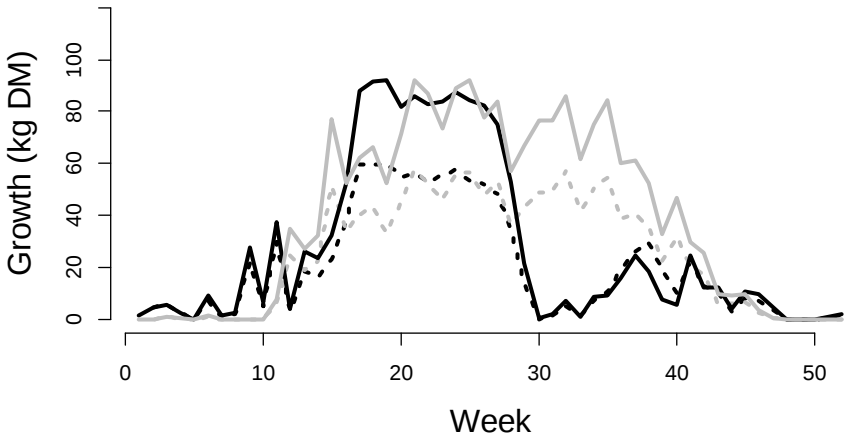


Figure 1. Net growth for the French 2008 weather (black) and the Irish 2009 weather (grey) at the level 0 (dotted line) and 300 (continuous line) of mineral N fertilisation.

The development of a graphical interface for this model will facilitate greater penetration at both farm and research levels. There will be two main applications of the new user friendly tool, depending on the user. The farmer will generally use it to predict changes in grass growth in the following week to 10 days while the researcher will use it around the effect of different management regimes on various environmental and economic outputs.

Conclusion and further work

The MGGM is a user friendly model with basic and simple inputs. The model has been capable of reacting in a sensible manner to different weather and mineral N fertilisation events. In the future the MGGM will be included in a whole farm dairy model (the Pasture Base Herd Dynamic Milk model (Ruelle et al., 2015), permitting the prediction of the impact of different climatic conditions, grazing management and fertilisation practices on various components of grass based systems of milk production.

References

Dillon, P., J. R. Roche, L. Shalloo, and B. Horan. (2005). "Optimizing financial returns from grazing in temperate pastures," in: *Proceedings of the XX International Grassland Congress (ICG 2005). Cork Satellite.* pp. 131.

Ruelle, E., L. Shalloo, M. Wallace, and L. Delaby. (2015). "Development and evaluation of the pasture-based herd dynamic milk (PBHDM) model for dairy systems," *European Journal of Agronomy,* vol. 71, pp 106-114.

Whitehead, D. C. (1995). "Grassland nitrogen," CAB international, Wallingford, UK, pp.416.

A mHealth platform for supporting self-management of chronic conditions

Huiru Zheng[1], Timothy Patterson[1], Chris D. Nugent[1], Paul J. McCullagh[1] , Mark P. Donnelly[1], Ian Cleland[1] , Suzanne McDonough[2] , Adele Boyd[1], Norman D. Black[1]
[1] School of Computing and Mathematics, [2] School of Health Sciences
University of Ulster, Shore Road, Jordanstown, Newtownabbey, Co. Antrim, BT37 0QB
e-mail: {h.zheng, t.patterson, cd.nugent, pj.mccullagh, mp.donnelly, i.cleland, a.boyd, s.mcdonough, nd.black}@ulster.ac.uk

Keywords: self-management, mHealth, chronic conditions

Background

Self-management is an crucial approach for people with chronic conditions in daily management of their conditions. Mobile health (mHealth) has been used to an effective tool for self-management of various chronic conditions, for example, stroke, chronic pain, diabetics and early dementia. It can provide basic information about the related disease, set up individual's self-management goal, assist in adopting a self-management care plan, lifestyles and intervention, and monitoring the progress in reaching the goal. However, most of mHealth applications are specific designed and lack of scalability. In this paper we present our current work in the development of a generic self-management mHealth platform, KeepWell, which can be readily extended for specific chronic conditions. The evaluation of KeepWell on self-management of the chronic obstructive pulmonary disease (COPD) will be presented and how our proposed design may be adapted for a specific condition will be demonstrated. The project is funded by Invest Northern Ireland (RD0513844).

Methods

The design of the generic mHealth platform KeepWell is built upon previous research on home based self management and is in collaboration with clinical experts. Functional requirements for the KeepWell platform align with the National Institute for Clinical Excellence (NICE) guidelines on COPD management (NICE, 2010) and Stroke Management (NICE, 2013) and the conceptualization of self-management for early stage Dementia by Martin et al. (2012) and can be summarised into five main functions of the self-management, i.e. activity planning, self-reporting, educational information, facilitation of social engagement and device measurement. The platform consists of four main components (Patterson et al, 2014): (1) an Android app that facilitates consumption of educational material, goal setting and display of health metrics; (2) health-tracking devices that enable the capture and quantification of measurements where state-of-the-art off-the-shelf devices such as smart wristband, wireless blood pressure monitor and wireless scales are used; (3) servers that provide storage and perform computation; and (4) a web-based portal. The goal setting is personalised according to each individual's condition and health status. This can be set with the support from clinicians. The tracking devices can easily be extended or added. The web-based portal enables the management of educational resources and self-report including the symptoms and workouts. These four components realise the five main functions of the self-management, i.e. activity planning, self-reporting, educational information, facilitation of social engagement and device measurement.

Results

The system has been developed and implemented. Currently a suite of Witherings devices are used for healthcare measurements. A single testing phase of a beta version of the platform has been conducted by members of the project team (N = 9) for approximately one week. After that, the usability of the KeepWell mobile application for COPD was tested by eight clinicians on a one off occasion.

Clinicians were contacted by one of the research team to arrange a time and date for an individual meeting. The meetings took place across Northern Ireland at the clinician's place of work or at the Ulster University. Before the participants began, the procedure was explained fully and they had a chance to ask questions. Clinical opinion was obtained by asking eight clinicians (six working in the area of COPD management across Northern Ireland; and two working in a research environment) to complete a series of tasks on KeepWell (see Table 1 below). Throughout the tasks, clinicians were permitted to consult with a printed manual, which explained the operation of the app, as needed.

Task	Purpose
Clinician to complete the following steps • Put activity tracker together and position on wrist. • Launch app on mobile phone device. • Set a step goal. • Self-report a workout or Borg. • View progress within app including daily/weekly. • Create a reminder using date and time. • Change/edit reminder date and time.	To demonstrate how to wear and use the wearable device. To demonstrate overall function of app and highlight specific features. To obtain clinician opinion on usefulness of features.
Clinician to view educational materials	Check format of the included materials.
Clinician to review suitability of self-report workouts included in KeepWell.	Check suitability of workouts included.
Clinician to complete questionnaire.	Quantitative and qualitative feedback

Table 1. *Tasks completed during the clinical evaluation. PR=pulmonary rehabilitation.*

Ethical approval was not required as the study was conducted under the auspices of Personal and Public Involvement that is the involvement of expert clinician as specialist advisers in the design of the KeepWell app.

The clinicians were asked to verbalise what they thinking about, looking at, doing and feeling throughout the process. They provided valuable knowledge and expertise based on their experience of COPD management in the development of the KeepWell platform. No information was

collected that could identify the clinicians involved, as the sole purpose was to evaluate usability of the KeepWell app.

After the 'think aloud' study had finished, users were asked if they have any suggestions for how the app could be improved or any additional comments they wished to make and were asked to complete a usability questionnaire. Any qualitative comments or any problems during the completion of each task were noted.

Overall the participants reported high levels of usability of the app and the suggestions have informed the improvement of the design (McDonough et al, 2016).

Summary

The KeepWell self-management mHealth platform generically implemented to accommodate three chronic conditions, and it is extensible to goal-setting, measurement, educational information, self-reporting and feedback components are generically implemented to accommodate three chronic conditions. The further work will be carried out to improve the usability of the KeepWell for COPD self-management and testing of KeepWell for stroke and dementia self-management.

References

National Institute for Clinical Excellence (NICE) (2010). Chronic Obstructive Pulmonary Disease Management of Chronic Obstructive Pulmonary Disease in Adults in Primary and Secondary Care. Technical Report April 2007, NICE, 2010.

National Institute for Clinical Excellence (NICE) (2013). Stroke Rehabilitation Long-Term Rehabilitation after Stroke. Technical Report April 2007, NICE, 2013.

S. M. McDonough, A. Boyd, T. Patterson, P. McCullagh, I. Cleland, C. Nugent, M. Donnelly, H. Zheng and N Black(2016), Development of mobile phone app in collaboration with Chronic Obstructive Pulmonary Disease clinical experts, ICDVRAT2016, to be presented.

T. Patterson, I. Cleland, C.D. Nugent, N.D. Black, P. McCullagh, H. Zheng, M.P. Donnelly, and S. McDonough (2014), Towards a Generic Platform for the Self-Management of Chronic Conditions. In Bioinformatics and Biomedicine (BIBM), 2014, IEEE International Conference on, pages 44–47, Belfast, Northern Ireland, 2014. IEEE.

Development of a Bioinformatics Pipeline to Identify Biomarkers of Infection

D. Mullins [1,2], N. Konstantinidou [1], R.D. Sleator [3], B. Kelly [1], P. Walsh [1*]

[*]paul.walsh@nsilico.com
[1] NSilico Life Science Ltd., Cork
[2] School of Microbiology, University College Cork, Cork, Ireland
[3] Department of Biological Sciences, Cork Institute of Technology, Bishopstown, Cork, Ireland

Keywords: Bioinformatics; Pipeline; Pathogen; Biomarker

In 2013, over 3 million children under 5 years of age died due to infectious diseases worldwide (Liu et al. 2015). Neonatal sepsis alone accounted for half a million deaths (Liu et al. 2015). Antibiotics are the main treatment used for neonatal sepsis, but with the rapid evolution of antibiotic resistance in sepsis pathogens, novel therapeutic agents are urgently required (Simonsen et al. 2014). Identification of virulence factors as potential therapeutic targets is likely to aid in the development of new treatments to reduce the death toll from neonatal sepsis (Allen et al. 2014).

The in silico identification of pathogen specific biomarkers is an obstacle for researchers lacking the required bioinformatics skills. The goal of the software pipeline developed in this study was, therefore, to provide a solution that would accelerate the study of microbial infections. Some aspects of this pipeline already exist in various other software, including NSilico's Simplicity™ (Walsh et al. 2013). The task was to further this, automating more genomic tools with one user-friendly interface. The result is a greater simplification of complex genomic studies for researchers and clinicians: providing a comprehensive comparison pipeline for the identification of clinically relevant biomarkers of infection.

Development of the pipeline required a thorough review and comparison of the available tools and their individual capabilities. With an automated pipeline as the objective, open-source tools which can be used locally from a Linux command-line were favoured. Visualisation steps of the pipeline required tools with a graphical user interface. Two bacterial genomes provided for this research were isolated and sequenced from mouse pups in a study of neonatal sepsis and identified as Listeria monocytogenes and Salmonella enterica serovar Typhi (Glaser et al. 2001; McClelland et al. 2001). An opportunity to draw biological inferences from the analysis of these pathogens is a useful indicator of the pipeline's potential.

The overall pipeline can be divided into two consecutive but separate components (Figure 1). These may be used in conjunction with each other, or independently for their individual purposes. The function of the first is the assembly and annotation of a bacterial genome from its raw sequence data, similar to that performed by Simplicity™ (Walsh et al. 2013). FASTQ format bacterial sequence files from an Illumina MiSeq desktop sequencer were assembled using SPAdes (Bankevich et al. 2012) or Velvet (Zerbino & Birney 2008) tools for de novo assembly. This step was followed by the application of several tools, including Ragout (Kolmogorov et al. 2014) and ABACAS (Assefa et al. 2009), for improving the resulting contiguous sequences by reference-genome-assisted assembly. Automatic genome annotation of the assembled genome was carried out by either Prokka (Seemann 2014) or RASTtk (Brettin et al. 2015). These annotators work by

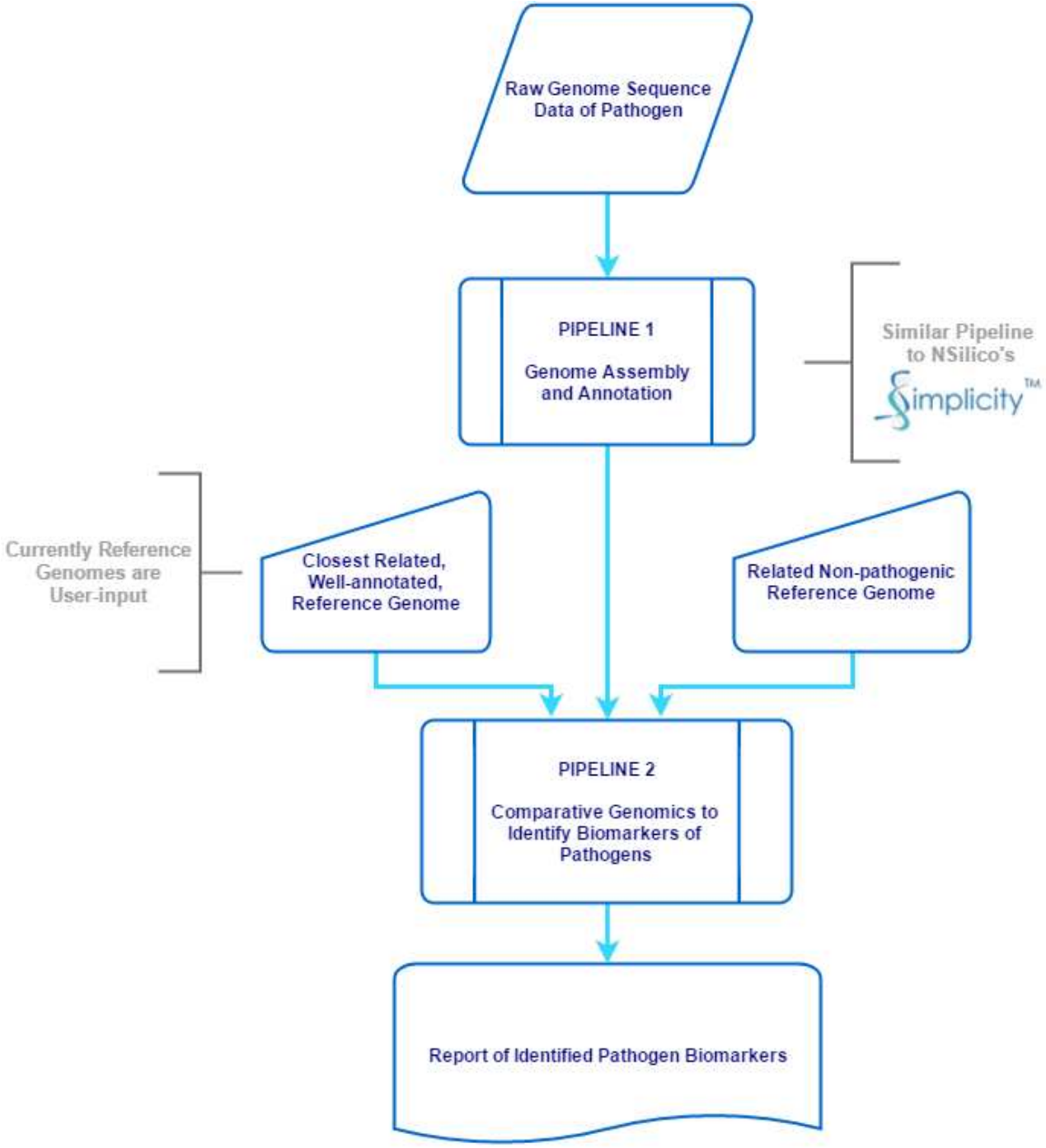


Figure 1. An automated pipeline to accelerate the discovery of pathogen biomarkers from the raw genome sequence data of a bacterium. Pipeline 1 takes genome sequence data from a clinical setting and outputs an assembled and annotated genome. One or more of such genomes may be used as input to Pipeline 2. In Pipeline 2 they are compared with related reference genomes, enabling the identification of pathogen biomarkers.
This flow chart was created at https://www.draw.io/.

identifying open-reading frames and annotating the functions of those matched with the database genes. The proteome obtained from this was further searched for specific virulence factors, antibiotic resistance mechanisms and disease-related loci, which may function as targets for disease targeting and treatment.

The second pipeline provides a homology-based comparative genomics analysis. Several annotated genomes are required as input, which may be assembled and annotated using the first pipeline or derived from an external source. The two genomes analysed here, L. monocytogenes and S. enterica, differ widely, being Gram-positive and negative respectively. Due to the large-scale differences between the pathogens, comparisons for the identification of specific biomarkers may not be useful. Instead, several closer related reference genomes, including non-pathogenic strains and those from different ecological niches, were selected for each pathogen bacterium. These genomes were obtained from the NCBI Genome Resource (https://www.ncbi.nlm.nih.gov/genome/). More distant species, outgroups, were required to reveal both core and unique genes for each group of genomes.

This input was then used for the construction of a 16S rRNA phylogenetic tree by the maximum likelihood method using the RAxML tool (Stamatakis 2014). This informed the evolutionary relationship between the genomes. The method employed for constructing the tree was similar to that used by Case et al. (2007), and involved building the tree from the consensus sequence of all 16S rRNA genes in each genome. A multiple genome alignment visualisation tool, such as Mauve (Darling et al. 2004), revealed large scale evolutionary events. Pan-genome analysis, with Roary (Page et al. 2015), was an effective method for finding pathogen-specific genes. Protein networks illustrate the pathways and biological processes where the pathogenicity-related proteins are involved. A pathway analysis for comparison of cell efflux systems, reveals potential targets for drugs such as efflux inhibitors (Poole 2012).

In this research, the substance of an automated pipeline for comparative genomics has been established in detail. It can provide useful information for any sequenced bacterial pathogen compared with similar and avirulent strains. With regards to neonatal sepsis, future use of the pipeline to compare many different pathogens implicated in this disease may highlight a prevalent virulence method between them. The pipeline is modular in nature, allowing for both advanced user-customisation and simplified use with default settings. A researcher or clinician using this pipeline would currently be required to provide appropriate reference genomes and an outgroup for input to the comparative genomics pipeline. Automation of this step, based on the particular genomes input, would prevent the need for in-depth knowledge of phylogenetics. The output of the pipeline will also be collated and presented to the user in an easily interpretable manner. Future directions will include a user-friendly web interface for the ease of use of the pipeline. Complete automation of the pipeline, once achieved, will greatly increase the power of this tool for clinical settings. Reproducibility and traceability and incorporated into this pipeline will further enhance the power of this in silico analysis.

References

Allen, R.C. et al., 2014. Targeting virulence: Can we make evolution-proof drugs? Nature Reviews Microbiology, 12(4), pp.300–308. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84893786916&partnerID=40&md5=6efffe6c0b1ea23c2761489df8157c7a.

Assefa, S. et al., 2009. ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics, 25(15), pp.1968–1969. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-67650735586&partnerID=40&md5=f83f960dd387f51e9a03a02c59ada26c.

Bankevich, A. et al., 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), pp.455–477. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84860771820&partnerID=40&md5=f36e6b5ca20bc9b547985e6750e9e79f.

Brettin, T. et al., 2015. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific Reports, 5. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84938841197&partnerID=40&md5=e9654b5105207240abcda82ee39b0b00.

Case, R.J. et al., 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and Environmental Microbiology, 73(1), pp.278–288.

Darling, A.C.E. et al., 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7), pp.1394–1403. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-3543051830&partnerID=40&md5=6956967b687ca5de895d50720417b588.

Glaser, P. et al., 2001. Comparative genomics of Listeria species. Science, 294(5543), pp.849–852. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035955375&partnerID=40&md5=26cf45584dc8f59a5b01bc78ccc5188f.

Kolmogorov, M. et al., 2014. Ragout - A reference-assisted assembly tool for bacterial genomes. Bioinformatics, 30(12), pp.I302–I309. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84902459237&partnerID=40&md5=d06ed37c5dcf32a85b3fa4cd72bfe524.

Liu, L. et al., 2015. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: An updated systematic analysis. , 385(9966), pp.430–440.

McClelland, M. et al., 2001. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature, 413(6858), pp.852–6. Available at: http://dx.doi.org/10.1038/35101614.

Page, A.J. et al., 2015. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), pp.3691–3693. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84947793096&partnerID=40&md5=9d8bd28e8238471bee72359051c0509d.

Poole, K., 2012. Efflux-mediated antimicrobial resistance. In Antibiotic Discovery and Development. Department of Microbiology and Immunology, Queen's University, Kingston, ON, Canada, pp. 349–395. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84929574738&partnerID=40&md5=5d7fc0b88f9e9793d7f5435744f9e232.

Seemann, T., 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), pp.2068–2069. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84901417347&partnerID=40&md5=548a55adcb4e1f2ad39b09d7fb1a25e4.

Simonsen, K.A. et al., 2014. Early-onset neonatal sepsis. Clinical Microbiology Reviews, 27(1), pp.21–47. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84891669485&partnerID=40&md5=b03c6854569ce8f5c1393e6d1305a8a2.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), pp.1312–1313. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84899553363&partnerID=40&md5=3c7be5d3c4ad2b61fde609bc54835b6d.

Walsh, P., Carroll, J. & Sleator, R.D., 2013. Accelerating in silico research with workflows: A lesson in Simplicity. Computers in Biology and Medicine, 43(12), pp.2028–2035. Available at: http://dx.doi.org/10.1016/j.compbiomed.2013.09.011.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), pp.821–829. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-43149115851&partnerID=40&md5=5f76329a247cda3e6d8c34393080742c.

An approach for anonymization of sensitive clinical and genetic data based on Data Cube Structures

Ntalaperas Dimitrios
Ubitech Ltd, Greece
e-mail: dntalaperas@ubitech.eu

Keywords: e.g. Anonymization, Security, Algorithms

Introduction

As Electronic Health Records (EHR) are adopted by an ever growing number of clinical health care providers, institutes can easily interchange information electronically, which can then be statistically analyzed, thus increasing research output. However, interchange of data concerning clinical trials and genetic data are subject to protection laws, which inhibit the disclosure of sensitive patient information. To this extent, data should be anonymized, not only to ensure that sensitive data is not eavesdropped during transmission, but also to ensure that the party legitimately receiving the data cannot have access to sensitive personal data.

In this paper, we present a methodology for transforming data corresponding to patient clinical and genetic data to an equivalent data set that contains the informational content of the original data set regarding the medical data, but from which all information regarding personal data has been filtered out. The methodology builds upon the Data Cube approach as this has been adopted by the Linked2Safety[1]consortium (Perakis et al. 2013) combined with cell-suppression and perturbation techniques (Forgó et al. 2012) which ensure that the produced data cube cannot be reversed engineered, while also retaining the same statistical characteristics as the original data set.

Theory

A dataset containing medical data can have mixed information with sensitive data being present at various locations of the source files. Moreover, the informational content may be in semi-structured or unstructured format (e.g. free text). An anonymized dataset must fulfil the following criteria:

1. Any information of personal data must be filtered out
2. Information that can be used to reverse engineer the anonymized data must be filtered out. If, for example, a combination of phenotypic characteristics is extremely rare and this data is present in the anonymized set, an eavesdropper can use this information to identify a patient.
3. Any statistical analysis performed on the anonymized data should yield the same results as for the original set.

[1] http://www.linked2safety-project.eu/

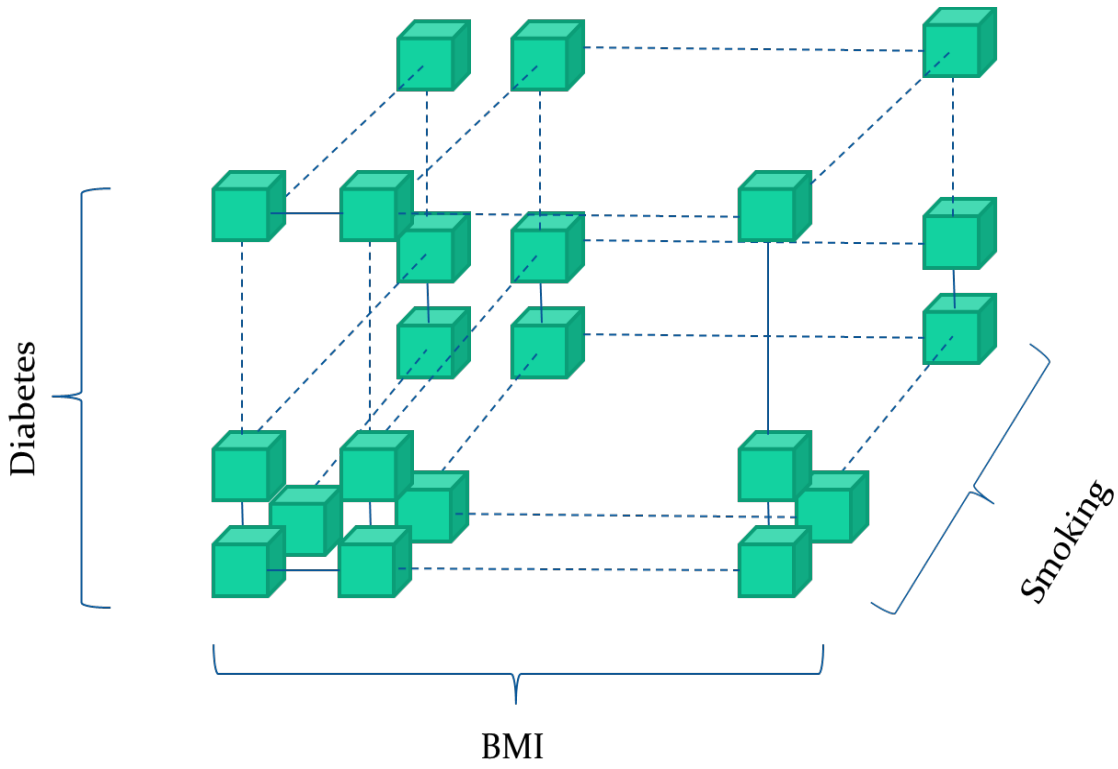


Figure 1. Example of a Data Cube Structure.

A data structure that can fulfil the above critirea a Data Cube. A Data Cube as an $m_1 x m_2 x ... x m_n$ array, with n being the number of selected variables and m_i being the number of distinct values the variable indexed by i can take. Each cell contains the number of total counts with the combination of values defined by the index of the Data Cube. Figure 1 depicts an example of a three-dimensional Data Cube. In this example, each variable is supposed to be categorical (take values from a district set of integers). The value *0* for BMI for example can correspond to low BMI, value *0* for diabetes to low blood sugar levels and so on. If the Data Cube is named *DC*, then the cell *DC[i][j][k]*, will have the total count of patients having BMI equal to *i,* Smoking equal to *j* and Diabetes equal to *k*. Since the cube contain only aggregate data, it fulfils criterion 1. Moreover, all data correlations of the original dataset concerning these variables are not altered, thus criterion 2 is satisfied.

A special case is that of *degenerate* data cubes, that can arise from non-categorical data. If a variable is not categorized (e.g. has continuous values) and a data cube is constructed naively, then it may be the case that each distinct value will correspond to a very small number of patients. In this case, the data cube will have a size of the order of the size of the original data set with each containing a small number. Criterion 2 in this case will be violated.

The case of degenerate cells is handled by adopting a combination of cell-suppression and perturbation techniques as well as imposing categorization on the data. Cell-suppression means that cells with a low value are discarded from the dataset. By discarding these cells distinguishing non-existent combinations from combinations with a low count becomes impossible. Perturbation imposes an extra layer of security by imposing a random, low amplitude noise to each cell so they are no longer guaranteed to contain the exact number of patients sharing the characteristics

described by the cell. The distribution of the noise can always be chosen as to not alter the statistical characteristics of the original dataset (Antoniades et al. 2012).

Imposing categorization means the definition of a mapping schema which groups values to categories. Such a schema must always be provided as input to the Data Cube Algorithm along with the source data.

Implementation and Results

The above general techniques presented above were implemented in the context of the SAGE-CARE[2] project to create anonymized clinical (phenotypical) data and genetic data recording the expression of genes of each patient. The methodology pipeline is depicted in Figure 2. The component at first accepts the two files that contain the source data; a .csv file that contains phenotypical data with each row corresponding to each patient and a .tsv file that contains genetic data with each row corresponding to a gene and each column to a patient. The .tsv file contains the patient code at the top cell of each column thus facilitating the linking of patients between the two datasets. The component constructs an intermediate structure which aligns the two data set so that operations, such as aggregation which is required for data cubes, can be easily performed.

At the second state, the component receives a mapping file in .xml format which contains the variables and genes for which a data cube is to be created. The mapping file also contains information for grouping values to categories; the categorization can be performed either by using value ranges or by using enumerations (see Fehler: Referenz nicht gefunden for a sample mapping file). The data are transformed according to the categorization schema and are then fed as input to the Data Cube Creation component. The Data Cube Creation component performs aggregation on the data (see Listing for pseudocode describing the algorithm).

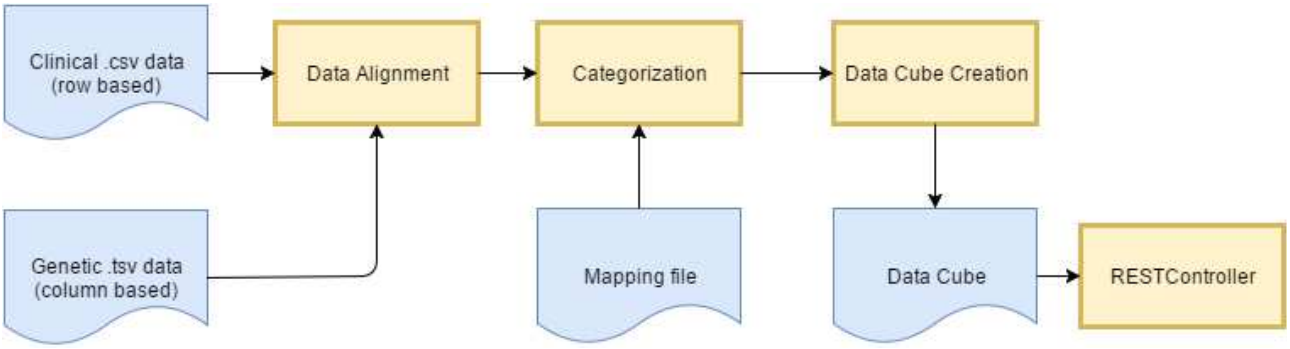


Figure 2. Methodology Pipeline

[2] http://www.sage-care.eu/

```
<?xml version="1.0" encoding="UTF-8"?>

<root>

    <phenos>

        <pheno name="gender" />

        <pheno name="height" round="1d" />

    </phenos>

    <genes>

        <gene name="155060" round="3d" />

    </genes>

</root>
```

Listing 1. Sample mapping file. Variable gender is defined without a mapping, which implies that the data are already categorized regarding gender values. Variable height is to be rounded to first digit which means that values within 10 cms will be grouped together. Similar for the gene with entrez code 155060 the expression counts within 1000 will be grouped together

```
foreach line
        computeunique_key
        foreach cell in line_cells
                cell_field++
                value[cell_field] = valueOf(cell)
                value[cell_field] = convertInRange(value, ranges[cell_field])
        Cube[unique_key][indexOf(value[cell_field])]++
foreachunique_key
        for i=1 to num_of_fields
                perturbe(Cube[unique_key][i])
                if Cube[unique_key][i] <= threshold
                        discard (Cube[unique_key][i])
```

Listing 2. Data Cube Creation Algorithm

Conclusions

In this paper a methodology for ensuring the anonymization of data present in the context of the SAGE-CARE project was presented. The methodology is based on peer reviewed structures and techniques that are validated both technically and legally thus providing a framework for the required level of data security.

Future work consist of exploring ways to align data between heterogeneous data sources that are not covered by the mapping scheme presented here. Under consideration is the adoption of OpenRefine[3], which allows the filtering, faceting and noise extraction on data, as well as the definition of a common mapping ontology in a similar fashion that was followed by the Linked2Safety consortium.

References

K.Perakis et. al. (2013). "Advancing Patient Record Safety and EHR Semantic Interoperability", in: IEEE International Conference on Systems, Man, and Cybernetics, IEE, 2013.

N. Forgó, M. Góralczyk, and C. Graf von Rex. (2012) "Security issues in research projects with patient's medical data", in: 12th Int. IEEE Conf. on Bioinformatics & Bioengineering (BIBE), IEEE, 2012

A. Antoniades et. al. (2012). "The effects of applying cell-suppression and perturbation to aggregated genetic data", in: 12th Int. IEEE Conf. on Bioinformatics & Bioengineering (BIBE), IEEE, 2012.

[3] http://openrefine.org/

Chapter 6

Humanities / Social

Designing narrative artefacts – the lives of women in eighteenth century Cork

J. Dempsey, P. Green, A. Wilson
Department of Media Communications
CIT – Cork Institute of Technology, Ireland
e-mail: jenny.dempsey@mycit.ie

Keywords: narrative, design, material culture, heritage

Introduction

This paper introduces the initial phase of a research project which is aligned with the development of a new heritage centre in Cork City entitled Nano Nagle Place. The centre is dedicated to communicating the life story and legacy of Honora (Nano) Nagle. Part of this story involves presenting a portrait of the city of Cork against a backdrop of political and social life in eighteenth century Ireland. The topical content of the research is embedded within this history and focusses specifically on the impact of the social and political environment at the time on the lives of women.

Academically, the project investigates the way in which narrative, when integrated with visual design and material culture, can support the goal of producing engaging and memorable designs for visitors or remote users of the heritage site. The project therefore contributes to the emerging study of narrative in art and design (Anderson, 2011, Hadjilouca et al., 2013) by accounting for the way in which visual media can be employed to allow audiences participate in the recovery of stories from material objects. This approach to narrative in design is typically influenced by a constructivist orientation in psychology (Bruner, 1986, 1990) and underlies recent studies that contribute transmedial narrative (Herman, 2004), cognitive narratology (Ryan, 1991, Herman, 2003, 2010, Fludernik, 2003), and to some extent multimodal narrative (Page, 2010).

Approach & Methodology

The project has access to various onsite facilities at Nano Nagle Place as indicated in Figure 1. These include a permanent exhibition space, a temporary exhibition space, an education room, rooms dedicated to the delivery of educational content and workshop activities, as well as outdoor spaces which are networked with high capacity digital communications. There will also be opportunities for the project to connect with audiences through the centre's outreach educational programme which will support and augment the curriculum in primary and secondary schools in the region. The research project is therefore related to learning but more explicitly it aims to engage the audience with eighteenth century historical information.

The development of the research was initially organised into four discreet phases: 1) data collection; 2) design investigations; 3) prototype/artefact development; and 4) the review and presentation of overall findings. Each stage involves its own cycle of investigation which results in a report of the findings related to that phase. While each stage can be regarded as discreet in terms of its objectives there is also significant overlap between the stages. For example, while stage one was intentionally dominated by activities such as: the review of historical material directly associated with Nano Nagle and the Presentation Sisters; site visits to relevant archives and museums; desk reviews of

literature on narrative theory; there was also some visualisation exercises and paper prototyping of ideas as information was being collected and evaluated.

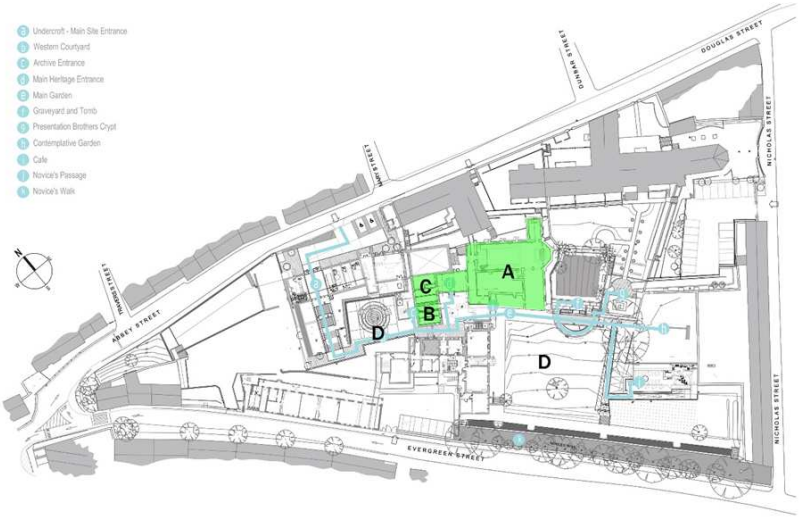


Figure 1. The site plan for Nano Nagle Place (approx. 3.8 acres) which includes potential target locations for the research onsite. The area marked A represents the location of the permanent exhibition; B is the proposed site of the temporary exhibition space; C is the education room and workshop spaces for young visitors; D represents potential outdoor spaces also available for use by the research project.

The first stage resulted in a body of factual information which provided a snap shot of eighteenth century Cork when Nano Nagle was establishing her schools for catholic children. The factual content helped established the conditions under which women from different social and economic standing lived. It was possible from this information to imagine how women of different standing may have experienced Cork in the 1700s and this provided the basis for the development of 'fictional' content for prototype designs which could then be tested with key stakeholders involved in the design of the heritage site and exhibition. The prototypes were presented in early June 2016 to stakeholders who included members of Nano Nagle Place Heritage Committee, and the JCA architects responsible for the design and layout of the overall site and exhibition spaces.

Prototypes

There are two prototypes presented in this section which emerged from the process outlined above. The first of these is a design for an interactive book and the second explored using picture cubes that could be reconfigured to provided different narrative outcomes for the user to consider.

The first prototype was intended to be located in the section of the permanent exhibition presenting Cork City as the mise en scène for the (hi)story of Nano Nagle's philanthropic work. Within this world of eighteenth century city life five fictional female characters from different social standings were developed and a vignette of their lives was illustrated. These fictional accounts were directly drawn from the historical data accumulated from stage one. One character was selected and a book (Figure 2) was designed around her which could be interacted with in a number of ways. The structure of the book itself proposed to take the user through a set of events associated with the purchase of material for the design of a dress. This operated as a conventional book with the plot unfolding as the user flips through the pages. This facilitated the story of the character's journey

through the city offering the user an introduction to the 'factual' city by taking them to shops in specific streets in the city. The book however also promised to offer options for accessing, and thereby directly experiencing, elements of the material culture of the time through the physical attributes of the book itself. The elements considered included the materials used in the making of the book, or physical elements that were part of the story but which could be removed and explored by the user. Examples included a piece of cloth fashionable and of the time, or patterned paper used for handwritten letters. Some consideration was also give as to how new media technologies could be integrated with the objects to provide further options for interaction and feedback.

Figure 2. Paper prototype of interactive book with embedded material elements. Considerations were given to how a narrative account could be facilitated by the materials themselves or extended by digitally augmenting the elements.

A second prototype (Figure 3) explored the idea of designing blocks that would allow users to assemble images into narrative sequences. Only certain coded sequences would offer the user with the reward of a story which could be recovered in a number of ways. By being digitally augmented certain sequences could automatically activate an audio account or unlock further clues in a block encouraging the user to 'discover' stories. This strategy identified well with some contemporary uses of narrative in art and design where the designer is delegated the role of providing narrative material to the audience, who then construct their own account of what is they think is going on. In practice however the range of potential narrative options produced significant design problems which made it questionable as a solution for engaging visitors.

Figure 3. Story Blocks

Conclusion

While these approaches outlined above failed on many levels they should be understood as early stage iterations in an overall design methodology which has narrative at its centre. These design solutions are intended to hook the user into a cognitive process of recovering a story from the designed material artefact. Part of the design process has come to involve judgments about what kind of gaps should be left for the user to fill in. What role the materiality of the object plays in the provision of narrative content is also a question that has emerged from the process of practically developing the prototypes above. Currently the project is exploring some of these ideas through prototyping a whimsical multi-modal book modelled on the idea of a 'time traveling guide' to 18th century Cork for ladies.

Figure 4. Prototype ideas for a whimsical multi-modal book supporting engaging access to history of eighteenth century Cork.

References

ANDERSON, S. Forward. In: HONIBALL, N. & HADJILOUCA, D., eds. Narrative in Practice Design Symposium Catalogue, 21 May 2011 2011 Central Saint Martins College of Art & Design. University of the Arts London, 27.

BRUNER, J. S. 1986. Actual minds, possible worlds, Cambridge, Mass., Harvard University Press.

BRUNER, J. S. 1990. Acts of meaning, Cambridge, Mass., Harvard University Press.

FLUDERNIK, M. 2003. Natural narratology and cognitive parameters. In: HERMAN, D. (ed.) Narrative theory and the cognitive sciences. Stanford, Calif.: CSLI Publications.

HADJILOUCA, D., HONIBALL, N. & NAGASAWA, Y. Introduction. In: NINA HONIBALL, D. H., ed. NiP 2013 Creative Symposium, 21 May 2011 2013 Central Saint Martins College of Art & Design. Hato Press, 18.

HERMAN, D. 2003. Narrative theory and the cognitive sciences, Stanford, Calif., CSLI Publications.

HERMAN, D. 2004. Towards a Transmedial Narratology. In: RYAN, M.-L. (ed.) Narrative across media : the languages of storytelling. Lincoln: University of Nebraska Press.

HERMAN, D. 2010. Narrative theory after the second cognitive revolution. In: ZUNSHINE, L. (ed.) Introduction to cognitive cultural studies. Baltimore: Johns Hopkins University Press.

PAGE, R. E. 2010. New perspectives on narrative and multimodality, New York, Routledge.

RYAN, M.-L. 1991. Possible worlds, artificial intelligence, and narrative theory, Bloomington, Indiana University Press.

Exploring new collaborative approaches in engagement and education towards the creation and enhancement of future STEM careers.

Dr Kieran Delaney [1], Dr Alex Vakaloudis [1], Paul Green [1], Trevor Hogan [1], Prof. Achilles Kameas [2], Dr. Ioannis D. Zaharakis [2],
[1] Cork Institute of Technology, Ireland, [2] Computer Technology Institute & Press, Greece
Email: kieran.delaney@cit.ie

Keywords: STEM

Introduction

European countries need as potential employees youngsters who are able to think creatively. They must apply new knowledge in an effective way, become continuously competitive in a highly demanding working environment through constant self-monitoring and thus be able to stand up for all the challenges that work based learning brings. The ability to switch efficiently between different disciplines (such as STEM [1]) depends on effectively processing the various forms of information based on clearly defined target outcomes, expanding a broad repertoire of ICT communication, problem-solving and decision-making skills, and using the collective knowledge represented in networks, based on working environments.

These are becoming urgent challenges in Europe and globally as a growing deficit in 21st century skills needs to be overcome. A recent report, the 'Future of Jobs' released by the Word Economic Forum [1] highlights the need for cross-industry and public-private collaboration; this envisages "partnerships between multiple businesses, educational institutions and accreditation providers can result in an overall increase in the quality of the talent pool, at lower costs and with greater social benefits". Reaching this goal is a significant challenge requiring new innovative solutions. This paper presents an approach to leverage novel new technologies, such as Ubiquitous Computing (UbiComp) [2], Mobile Computing (MobiCom) and the Internet of Things (IoT), in combination with the use of Communities of Practice (CoP) [2] to deliver these much needed solutions.

The objective of the research is to seek to expand the scale of student transdisciplinary projects by enabling the creation of a much broader range of more ambitious collaborative projects. These multi-disciplinary "projects of common purpose" would be meta-level initiatives that build upon the existing landscape, employing connective engagement models, new technology and tools that currently do not exist in order to achieve this. Students' participating to create, implement and disseminate transdisciplinary 'Challenge' projects would increase their empathic intelligence and problem-solving skills. They would also achieve a very strong understanding of how to collaborate, create and produce to real-world requirement and learn deeply how to operate effectively in dynamic environments.

[1] STEM: science, technology, engineering, and mathematics

[2] A community of practice (CoP) is a group of people who share a common craft or set of skills and share knowledge of the executive of these skills. The concept was proposed in 1991 by Jean Lave and Etienne Wenger. It should be noted that Communities of practice are not new phenomena; they have existed for as long as people have been sharing their learnt experiences through reports and storytelling.

Figure 1. High level overview of the elements present in a Community of Practice for STEM topics, which uses problem-based learning approaches.

Concept and Approach

In education, numerous programs exist to promote STEM activities; many take the form of project-based science and innovation competitions. This creates a landscape that talented STEM students will navigate without great difficulty; however, this is a fragmented picture for many students and a difficult one for teachers. The competitive nature of these projects acts to limit involvement and most operate annually to relatively short timelines. This effectively resets these programs every year. Importantly, an emphasis upon STEM skillsets means that there is often a significant deficit in creative enablers in most of these programs.

Many successful examples do exist of project-based collaborations built by multi-disciplinary student teams that transcend this. Children in coding clubs (e.g. CoderDojo) mentor each other using programs like Scratch and have subsequently self-organized to build teams to compete in robotics competitions. Undergraduates in engineering courses collaborate with business students to create new product prototypes; often these projects emerge from undergraduate work and evolve towards competitive programs, and may then be reinvented as postgraduate projects.

Our research envisages the creation of a continuum effect that connects these types of projects together through a suite of tools that enables students of different disciplines, ages, background and abilities to work together on transdisciplinary projects. Along these lines, this initiative is designing meta-level solutions to link school, community and third-level activities together and should foster a model that looks to strongly broaden impacts from the current cohort (e.g. elite students competing in STEM competitions) to the entire ‘education’ population. The project will also provide tools, particularly technology, to support this; however, the creation of this toolset must empower the stakeholder group and this will be considered carefully. New technology is often rejected, so approaches where participants can help co-create the toolkit must be prioritised.

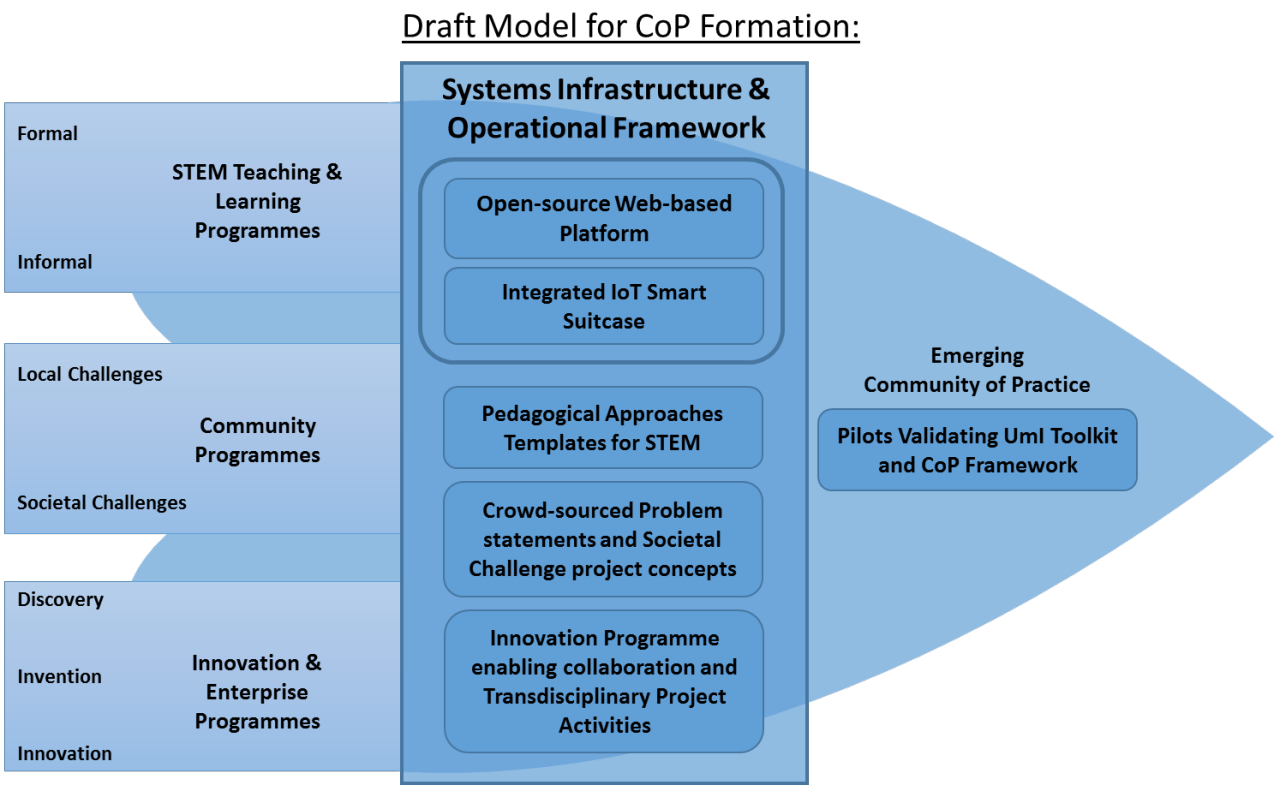


Figure 2. A Model for the Implementation of a Community of Practice: bringing Education, Industry and Communities together by using new enabling technologies and novel models of engagement.

The most effective way to do this will be to build these CoPs within active clusters where a body of knowledge already exists. These CoPs will need to foster cultures of innovation, investigate the drivers of success that can sustain key activities in the face of limited resources and high levels of turnover. [3, 4, 5]. A novel CoP framework will be presented to that explores new technologies and new models of engagement to create syntheses between STEM education, the arts and enterprise-driven and social innovation.

Conclusion

The central approach of this research will be multi-disciplinary project-based learning, with strong emphasis on increasing empathic intelligence and problem-solving skills. Student projects will reflect societal challenges, crowd-sourced from industry and communities in the innovation programmes and address challenging, complex issues; a mix of technology, arts and social science students will need to assemble to address them. Teachers will frame templates mapping the pedagogical requirements to project team profiles so that the integrity of assessments will be maintained. Students across many locations and disciplines will then self-organise to address the problems. A key innovation will be to foster scalable, collaborative internship projects operating across a range of host companies, mentored coherently by teachers from different disciplines and educational organisations.

Acknowledgements

This work is funded in part by the H2020 project UmI-Sci-Ed [710583]: Exploiting Ubiquitous Computing, Mobile Computing and the Internet of Things to promote Science Education

References

1. The 'Future of Jobs' report: http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf [accessed 12/08/2016]

2. Weiser, M. (1993). Ubiquitous Computing. Computer, 26(10), 71–72.

3. Lave, Jean; Wenger, Etienne (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge: Cambridge University Press. ISBN 0-521-42374-0.; first published in 1990 as Institute for Research on Learning report 90-0013

4. Wenger, Etienne (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press. ISBN 978-0-521-66363-2.

5. Wenger, Etienne; McDermott, Richard; Snyder, William M. (2002). Cultivating Communities of Practice (Hardcover). Harvard Business Press; 1 edition. ISBN 978-1-57851-330-7.

About UmI-Sci-Ed

UmI-Sci-Ed: Exploiting Ubiquitous Computing, Mobile Computing and the Internet of Things to promote Science Education [Project: 710583] Funded under H2020-SEAC-2015-1: Innovative ways to make science education and scientific careers attractive to young people

Many exciting new technologies are emerging, like Ubiquitous Computing (UbiComp), Mobile Computing (MobiCom) and the Internet of Things (IoT); in the following, we shall refer to them collectively as UMI. UMI technologies attempt to revolutionize everyday life by "allowing computers themselves to vanish into the background". UMI applications are presumed as a possible next generation computing environment in which each person is continually interacting with hundreds of nearby wirelessly interconnected devices: as a result, radical new uses of portable information technology emerge based on "the nonintrusive availability of computers throughout the physical environment, virtually, if not effectively, invisible to the user". These technologies are so modern and powerful that they can be both an educational means and end, thus fostering innovation and supporting promising scientific careers.

The broad aim of the project is to investigate the introduction of UMI technologies in education. By carefully exploiting state of the art technologies in order to design educational tools and activities, the project aims to offer novel educational services, implement innovative pedagogies and enhance students' and teachers' creativity, socialisation and scientific citizenship. We intend to put these technologies in practice, so as to enhance the level of science, technology, engineering and mathematics (STEM) education young girls and boys are receiving and at the same time make attractive the prospect of pursuing a career in domains pervaded by UMI. The orientation of UMI-Sci–Ed is entrepreneurial and multidisciplinary in an effort to raise young boys' and girls' motivation in science education and to increase their prospects in choosing a career in pervasive, mobile computing and IoT. We aim to develop an open fully integrated training environment that will offer 14-16 year old students and their teachers an open repository of educational material, educational scenarios, training material and activities, social tools to support communities of practice, entrepreneurship training, showcases, self-evaluation online tools, mentoring, and content and information management.

Workplace Well-Being: Mental and Psychological Health and Safety of Non-Unionised and Unionised Employees

Peter O'Mahony, Dr Angela Wright
Cork Institute of Technology, Ireland.
Dept. of Organisational and Professional Development. School of Business, Bishopstown, Cork.
Email: peteromahony@gmail.com, angela.wright@cit.ie

Keywords: Mental Health, Work Place, Guarding Minds at Work, Occupational Health, Psychosocial Risk, Workplace Well-being

Abstract: Imagine working in a highly productive environment in which you feel safe, respected and valued; the work is challenging; the demands of the job are reasonable; you have work-life balance; and your employer supports your involvement in your work and interpersonal growth and development. This is what is known as a mentally healthy workplace, (Canadian Centre for Occupational Health and Safety, 2012).

Twenty-three percent of workers in Europe are reported to have low levels of well-being (Ardito et al., 2012). Organisations, therefore, need to examine the operating environment so that they can begin to improve the psychological health of their employees (Mowbray, 2014). Failure to assess and resolve psychological health and safety risks in the workplace can create significant employee issues in an environment that requires "psychological competencies such as judgement, knowledge, interpersonal cooperation and emotional self-regulation. These psychological tools and skills flourish only in work environments that nurture and support their development and use, and minimize psychosocial factors in the work environment that can serve to undermine them" (Gilbert & Samra, 2010: para 2). While work can play a vital role in a person's quality of life as it provides them with a source of income and a platform for broader social advancement, it can also impact on a person's health as a result of risk factors present in the workplace which may cause injury, work-related illness or could potentially result in long-term health implications (Ardito et al., 2012). Consequently, mental and psychological health problems in the workplace can have a significant effect on both the employee and the organisation, in turn impacting productivity and competitiveness.

This research therefore seeks to explore the mental and psychological health and safety of employees within a multi-national pharmaceutical company based in the Republic of Ireland so as to address some of these pertinent issues. The participating company is listed in the top one hundred companies on the 2015 Fortune 500 list (Time Inc., 2015).

A quantitative case study was conducted in this study to gain a wider understanding of the mental and psychological health and safety of non-unionised verses unionised employees, specifically within this multi-national pharmaceutical company. One hundred and sixty-seven employees and seven senior managers were surveyed to explore their perspectives of psychological health and safety risks in the workplace across thirteen different psychosocial factors which are known to have either positive or negative effects on the psychological health and well-being of an employee.

To conduct this study, the research utilized online resources provided by Guarding Minds @ Work, an initiative which has been developed by the Canadian based Centre for Applied Research in Mental Health and Addiction (CARMHA). By using two online survey methods, this study tested a hypothesis which assumed that there would be more psychological health and safety risks among

unionised employees, while also investigating employee's experience of discrimination, bullying or harassment, and unfair treatment in the workplace. In order to test this theory, the following alternative hypothesis (H1) was tested:

'If a union exists within a workplace, then there will be a difference between the combined psychological health and safety concerns for non-unionised and unionised employees which will indicate that there is a greater risk to the psychological health and safety of unionised employees'.

Differences in the perspectives of all employees and senior managers were also compared with respect to the psychological health and safety of employees in the workplace. Statistical analyses in the form of t-tests were applied to the data obtained from ninety-seven non-unionised and sixty-three unionised employees to examine the level of psychological health and safety risks present within these groups.

The key findings from this study reveal that the psychological health and safety of unionised employees is more at risk in this workplace; in fact, unionised employees reported greater concerns across eleven out of the thirteen psychosocial factors. Furthermore, this research found that more unionised employees report previous experience of bullying or harassment in the workplace. It is also evident from this study that senior managers are significantly underestimating the psychological health and safety risks which are impacting employee's psychological health and safety in the workplace.

Considering that physically and psychologically healthy engaged employees maximise business efficiency and profitability, enabling organisations to achieve its goals and objectives, this study is pertinent and relevant to businesses who seek to protect the psychological health and safety of employees. As a result of dynamic workplaces, employees are now expected to be more flexible and resilient to changing organisational priorities as companies attempt to compete in challenging environments. Such challenges place a significant amount of pressure on employees which can inherently increase the risks to their physical and psychological health, safety, and well-being. Therefore, the main findings and recommendations for future practice and future research resulting from this study provide valuable insights not only for the participating multi-national company, but also for policy makers, employers, management, human resource departments, trade union representatives and professional or organisational development institutions.

References

Ardito, C., d'Errico, A., Leombruni, R., & Pacelli, Lia. (2012). Health and well-being at work: A report based on the fifth European Working Conditions, Dublin: European Foundation for the Improvement of Living and Working Conditions.

Canadian Centre for Occupational Health and Safety. (2012). Health & Safety Report, Vol. 10, No. 9, available at http://www.ccohs.ca/newsletters/hsreport/issues/2012/09/ezine.html, accessed January 9th at 6pm.

Gilbert, M, & Samra, J. (2010). "Psychological Safety and Health in the Workplace", Good Company, American Psychological Association for Organisational Excellence, Vol. 4, No. 5, May 12th.

Mowbray, D. (2014). "Psychologically healthy workplaces", The WellBeing & Performance Group, available at http://www.mas.org.uk/uploads/articles/Psychologically_healthy_workplaces_September_2014.pdf

Time Inc. (2015). Fortune 500, available at http://www.timeinc.com/experiences/the-fortune-500/, accessed January 6th, at 9pm.

Gender, Entrepreneurial Self-Efficacy, and Entrepreneurial Career Intentions: Implications of Female Role Models for Entrepreneurship Education

Ciara Lavelle O'Brien
Department of Management and Enterprise
CIT – Cork Institute of Technology, Ireland
E-mail: ciara.lavelleobrien@gmail.com

Keywords: entrepreneurship, self-efficacy, career intentions

Objectives

The objective of the research activity is to further the understanding of how entrepreneurship contributes to economic growth, competitiveness and social wellbeing.

Within the context of entrepreneurship education and entrepreneurial role models, this research explores the impact of gender. Guided by self-efficacy, entrepreneurship and entrepreneurship education theory three main research objectives are proposed. Firstly, to examine and assess how entrepreneurship education is delivered from primary school level education to third level education in an Irish context. Secondly, to explore the role and input of female role models in the entrepreneurship education system. Thirdly, to assess the effects of female role models on the entrepreneurial self-efficacy and career intentions of students.

Prior Work

This research draws on scholarly work from gender studies, entrepreneurship education research, role model theory and self-efficacy studies and includes writers such as Ajzen, I.; Bandura, A.; Henry, C.; Foss. L.; Kickul, J.; Wilson, F.; Marlino, D. and Barbosa, S. D.

In advanced market economies, women own 25% of all businesses and the number of women-owned businesses in Africa, Asia, Eastern Europe, and Latin America are increasing rapidly (Estes, 1999; Jalbert, 2000). This data reinforces the value of conducting further research into female entrepreneurship. (Women Owned Businesses, 2004).

The concept of female entrepreneurship is gaining traction in the academic world as well as the practitioner community. Indeed, Henry et al's (2012) review of extant gender and entrepreneurship literature's demonstrates a recent and significant proliferation of female entrepreneurship empirical research.

The Global Entrepreneurship Monitor (GEM), which provides useful international comparative information on entrepreneurship, reflects the difficulties for entrepreneurship which Ireland has experienced in recent years.

According to GEM 2013, Irish men are 1.9 times more likely than Irish women to be an early stage entrepreneur, with rates of early stage entrepreneurs at 12.1% for men and 6.4% for women. This has been steadily improving and although now level with the EU- 28 average of 1.9:1 and slightly higher than the OECD average of 1.7:1, it still shows untapped potential amongst female entrepreneurs in Ireland.

In 2014, the GEM survey confirmed again that entrepreneurial activity is mostly performed by men. More men than women are actively planning and starting new businesses in Ireland (2.1:1). Of the 20,400 individuals who started a business in Ireland in 2014, 14,400 are men (70.5%) and 6,000 are women (29.5%).

This situation is not unique to this country. Across G7 countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) a similar trend applies. In order to address this issue, the world's leading industrialised nations cited women's economic empowerment as a top global priority in a Joint leader's declaration presented on 8 June 2015 at the Group of Seven (G7) summit in Schloss Elmau, Germany and agreed the following common principles to support women's entrepreneurship:

- Make girls and women aware of the possibility of becoming entrepreneurs and actively encourage them to transform their ideas into business cases – starting at early stages, e.g. in schools, vocational training and universities by promoting tailored information.
- Counter gender stereotyping, and develop specific measures for girls to enrol in and complete education in STEM (Science, Technology, Engineering, and Mathematics) disciplines early on.
- Make successful female entrepreneurs more visible, e.g. as role models for new generations of women founders.
- Address the specific needs of female entrepreneurs, e.g. by providing them with tailored information, skills, mentoring and coaching and promoting networks for women entrepreneurs.
- Facilitate women entrepreneurs' access to finance, e.g. alternative sources of funding as well as the banking system, to technology and innovation and to domestic and international markets.

The declaration highlights the importance of catapulting women's entrepreneurship as a key driver of innovation, growth and jobs and provides further rationale for the need for research in this area.

Prior research has been conducted on entrepreneurial self-efficacy, gender roles in entrepreneurship as well as a large amount of research attention on entrepreneurship education. There does however seem to be a gap in combining these theories together to recognise the effect that entrepreneurship education has on female students in particular as well as examining the effect of female role models on the entrepreneurial mind-set and career intentions of students.

With regard to the methodology, the use of George Kellys (1955) repertory grid technique (RGT) in entrepreneurship research has been minimal thus far (Klapper 2015). This addresses another gap in the research.

The outcomes of this research will provide insights into how entrepreneurship is delivered in the modern classroom as well as the impact of female role models on the entrepreneurial mind-set and career intentions of students.

Approach

Extensive work has been conducted in an attempt to examine entrepreneurship education but has been inconclusive (Cox et al., 2002). This may be linked to methodological issues, specifically the outcome measures used in many studies, such as student satisfaction and performance, as these are insufficient indicators (Cox et al., 2002). For the purpose of this study a qualitative approach will be adopted as it has the potential to increase the validity, depth, richness and creativity of entrepreneurship education research. Qualitative research encompasses a variety of methods that can be applied in a flexible manner, to enable respondents to reflect upon and express their views. Qualitative data typically captures what OECD (2009) refers to as hard (actions taken such as start-up of a new venture) and soft (attitudes, intentions, cognitions etc., such as self-efficacy, intention to become an entrepreneur).

This research will explore the use of George Kellys (1955) repertory grid technique (RGT), the methodological tool of Personal Construct Theory (PCT). One of the main aims of PCT is to understand people's unique views of the world by exploring their thoughts, feelings and beliefs (Cooper, 1998). Data will be collected mainly by methods such as semi-structured interviews, participant observation and repertory grids, which are a methodological tool of George Kelly's Personal Construct Theory (1959).

Implications

This research will have practical as well as policy implications. From a policy perspective, the research will inform entrepreneurship, education and gender policy development. From a practical perspective, this research will provide empirical evidence of the role and impact of female role models across the education system.

Value

Due to the research deficit, it is anticipated that the ongoing study will make a substantial contribution to both academic knowledge and practice by expanding the entrepreneurship literature. The study will also highlight essential aspects of the discipline which warrant further investigation in the future. From a research perspective this research will also bring about a better understanding of how role models are constructed in the minds of entrepreneurship students.

References

Cox, L., Mueller, S., & Moss, S. (2002). The impact of entrepreneurship education on entrepreneurial self-efficacy. International Journal of Entrepreneurship Education, 1, 2.

Global Entrepreneurship Monitor Report (2013), Paula Fitzsimmons, Colm O'Gorman 2011, [Online] Available at: https://www.enterprise-ireland.com/en/Publications/Reports-Published-Strategies/GEM-Reports/GEM-Report-2013.pdf

Global Entrepreneurship Monitor Report (2014), Slavica Singer, Jose Ernesto Amoros, Daniel Moska Arreola 2014, [Online] Available at: http://www.babson.edu/Academics/centers/blank-center/global-research/gem/Documents/GEM%202014%20Global%20Report.pdf

Playing in the museum: shaping a definition of digitally mediated playful experiences

Denise Heffernan [1], Dr Kieran Delaney [1], Paul Green [2]
[1] Nimbus Centre, [2] Department of Media Communications
CIT – Cork Institute of Technology, Ireland
Email: denise.heffernan@mycit.ie

Keywords: museum, play, embodied interaction

Introduction

Museums have always been early adopters of new technology and in recent years, museums have increasingly embraced gameful and playful design (Rugged Rovers; Designing the Pen, 2014). However, museums traditionally have a preference for a didactic approach and emphasize learning over personal experience. Digital technologies have been embraced typically with the aims of democratising knowledge, contextualising information, and ultimately boosting visitor numbers. Research on promoting play within the museum has been focused on educational goals (Taylor et al., 2015). Little consideration has been given to designing for the intrinsic value of play; play itself.

Play is a form of understanding who we are, what surrounds us and how to engage with others (Sicart, 2013); the social context for playing is particularly important in creating memorable experiences. Play is not outcome-driven but focused on the process. It is inherently driven by curiosity. It encourages exploration and experimentation, results in discoveries, is often spontaneous, and is not concerned with goals. An environment can encourage and promote play just as it can hinder play. Playfulness is distinctly different to game-based activities. The adoption of gameful methods don't encourage open-ended play, but a specific goal-based learning interaction.

Emerging technologies that underpin the Internet of Things provide an opportunity to explore a space without the limits associated with typical interactive devices found in the museums setting. We speculate that the potential for a connected network of embedded digital interventions could allow the visitor to playfully investigate, explore and re-interpret their museum experience. In this paper we explore the design of playful, digital meditated experiences. We describe the design process that is being undertaken in this research. We discuss our play matrix as a tool that helps define playful activities. We briefly review digitally augmented museum experience, and finally, we reflect on the importance of this research and future directions.

Related Work

Research within the field of interaction design (IxD) and human computer interaction (HCI) has produced a variety of technological solutions for the museum context which are mobile, interactive, and reusable. Projects such as the MeSch Project (Material EncounterS with digital Cultural Heritage) aim to provide a series of tools that will allow cultural heritage professionals design and implement interactive experiences without the need for "expert" developers; supporting the creation of an open [playful] community of cultural heritage professionals driving and sharing a new generation of interactive physical/digital smart exhibits. (Petrelli & Lechner, 2014).

While the education goals of museums are still paramount, there has been a shift from the didactic model to a more adaptive, socially inclusive dialogical approach. This coupled with the emerging DIY approach afforded through projects like MeSch, moves the museum from teacher to facilitator; "a place where visitors can create, share, and connect with each other around content" (Simon, 2010). A place where playful exploration becomes central to a more engaging and memorable experience.

Research approach

The research has adopted a mixed methods approach. The focus of this initial investigative stage is to define a hierarchy of playful criteria and sensory modalities most appropriate to play. This involved a survey of literature as well as a review of current playful and technological practices employed by museums.

We began by defining initial play criteria through a literature review of play theories. This formed the basis of our play matrix, a tool to identify observable playful behaviour. The play matrix helped to observe and measure play activities. We tested the matrix in museums and other public spaces, including playgrounds where overtly playful behaviour is found. Observations using our play matrix, noted that play activities often in their appropriation of the space was increasingly tactile and social. Sticks become swords. Climbing frames become mountains. Acquaintances became trusted allies. The evaluation of the matrix allowed dominant and new criteria to emerge and began to reveal a hierarchy of sensory modalities which could be mapped back to our playful criteria. Emerging sensory modalities will be used as a method of interpreting how IoT technologies should be built to genuinely enhance opportunities for playfulness.

Self-directed	Imaginative	Experimentation	Social	Collaborative	Discovery	Challenge	Body Language
Minimal direction from museum/staff/adults	Separate from real world restrictions	Using the space/object for something other than intended	Engage with others in own group/family unit	Asking another to join activity	Questioning	Requires application of self	Open - Welcoming, friendly, Inclusive
Freedom of choice (leaving the prescribed route etc)	Role playing	Observation, watching and learning	Engage with strangers	Combine with another to finish/start an activity	Looking beyond what is presented	Does not cause anxiety	Active
Choose to engage, not forced	Prescribing roles to another	Test boundaries & limitations	Communication outside of expected interactions	Including others in an activity	Searching	There is some resolution to the challenge	Animated
Stopping/Starting at own discretion	Engrossing	Freedom from externally imposed rules (e.g quiet in the museum)	Laughing	Joining others in their movement through the space	Finding, prescribing new meaning	Problem solving	Inquisitive, picking up objects, getting closer to an object
Non instruction led activities	Fantasy/ make believe elements		Cross-generational engagement	Cooperation with others	Attention to the activity rather than outcome		
	Active engagement				Curiosity		

Figure 1. Example of initial play matrix

A review of current practices in museums was also undertaken through a review of literature and in-the-field observations. We found that technological solutions often cut visitors off from the social interactions that define and shape the museum experience. The museum visit is shaped and enhance through our encounters with those whom we visit with and encounter within the museum (Falk & Dierking, 2012). However, there are many examples of technology that hinder social, engaging interactions: the popular use of audio guides can isolate the visitor, actively discouraging talking and engaging with one another; a kiosk directs attention towards itself and away from artefacts on display, and interactive interfaces and touchscreens that are led by a single user. Designing for playful engagement could break down the barriers often associated with current technological solutions and, importantly, create a more memorable social experience.

Discussion

The possibilities offered by emerging IoT technologies and playful design can help support the museum to offer new experiences, even to an audience already familiar with its content. There is opportunity to take advantage of visitors' physical experience with cultural heritage and to integrate technology into it instead of creating a parallel and detached digital experience. It also provides the opportunity to create a social and collaborative experience with those we visit with and encounter in the museum. Research not specific to exhibition design has shown how embedded technologies can be used to create, meaningful, collaborative, social experiences (Speed, 2010). Technology can be used to encourage playful, exploratory modes of engagement, and to provide an experience-centric not learning-focused museum visit.

By focusing on designing for the social, exploratory nature of play could support a more meaningful, social experience for some within museums when compared to other activities such as tours or presentation material that are prescribed by the museum curator. We have begun to outline design sensitivities, the next steps are to evaluate and validate these design sensitives and further explore what opportunities and limitations do IoT technologies have in the implementation and support of playful spaces in complex public environments such as the museum.

Future Direction

By focusing on designing for the social, exploratory nature of play could support a more meaningful, social experience for some within museums when compared to other activities such as tours or presentation material that are prescribed by the museum curator. We have begun to outline design sensitivities, the next steps are to evaluate and validate these design sensitives and further explore what opportunities and limitations do IoT technologies have in the implementation and support of playful spaces in complex public environments such as the museum.

References

Cooper Hewitt Smithsonian Design Museum. (2014). Designing the Pen. [online] Available at http://www.cooperhewitt.org/new-experience/designing-pen/ [Accessed 28 Nov. 2015]

Falk, J. & Dierking, L. (2012). Museum Experience Revisited. Walnut Creek, CA, USA: Left Coast Press.

Lechner, M. & Petrelli, D. (2014) The Mesch project: reusing existing digital resources in the creation of novel forms of visitor's experiences in CIDOC - the International Committee for Documentation, Annual Conference - Access and Understanding: Networking in the Digital Era. Dresden

Preloaded Games. (n.d.). Rugged Rovers. [online] Available at http://preloaded.com/games/rugged-rovers/ [Accessed 28 Nov, 2015]

Sicart, M. (2014). Play Matters. MIT Press.

Simon, N. (2010). The Participatory Museum. Santa Cruz. Left Coast Press, Inc.

Speed, C. (2010). An Internet of Old Thing. Digital Creativity, 21(4), 235–246.

Taylor, R., Bowers, J., Nissen, B., Wood, G., Chaudhry, Q., Wright, P. (2015). Making Magic: Designing for Open Interactions in Museum Settings in Creativity and Cognition Conference, Glasgow. pp. 313-322

Meeting in the Middle: A collaborative and multidisciplinary approach to developing an empowerment course for women survivors of gender-based violence.

S. Davis, B. Kenny
Hincks Centre for Entrepreneurship Excellence,
Department of Management and Enterprise
CIT – Cork Institute of Technology, Ireland
e-mail: sarah.davis@cit.ie

Keywords: Collaborative learning, Gender-based violence, Women's empowerment

Introduction

The purpose of this paper is to explore the challenges faced in the creation and piloting of a training programme for women survivors of gender-based violence in Ireland. The training programme included employability and entrepreneurship skills and combined this with elements of personal-development to improve self-confidence and self-efficacy. Empowerment of women who had experienced gender-based violence was the goal of the Daphne III project called NEW START which ran from 2014-2016 in seven EU countries.

Gender-based violence (GBV), also commonly known as domestic violence (McWilliams and McKiernan 1993), is a complex and wide spread phenomenon. In Ireland, over 12,500 people annually, 9,448 women with 3,068 children received support and/or accommodation from a domestic violence service (SAFEIreland 2016). Women who have experienced GBV tend to have low self-esteem and low self-efficacy. This paper draws on the literature to develop an understanding of the collaborative process in curriculum development and seeks to reconcile this with women's empowerment literature and an understanding of GBV survival and recovery. The Wheel of Power and Control (Goodman and Fallon 1994) a commonly-used model to explain GBV to those seeking assistance and support is reviewed. An understanding of power and empowerment is then sought (Charmes and Wieringa 2003) and finally, the educational literature in the areas of andragogy, heutagogy (Blaschke 2012) and transformational learning (Mezirow 1991) is explored.

Design/Methodology/Approach

This paper outlines the collaborative development (Voogt, Westbroek et al. 2011) and implementation of a ten week pilot programme that was delivered in Ireland in 2016 to eight women who had previously experienced domestic violence. The EU project requirement was to combine elements of life-coaching and elements of mentoring or employability training with personal development training and, to produce and test training materials suitable for use with women survivors. A qualitative emergent enquiry approach (Keegan 2009) was taken (by the collaborating partners in Cork Institute of Technology and YANA (You Are Not Alone) Domestic Violence Project) to the multidisciplinary programme development and implementation. Qualitative data in the form of notes from training the trainers and notes from the pilot sessions debriefing were analysed. Thematic analysis was used to identify the main themes from this qualitative data.

Quantitative data, in the form of three standard instruments for self-esteem (Rosenberg 1965), self-efficacy (Schwarzer and Jerusalem 2010) and entrepreneurial self-efficacy, were gathered in a pre- and post-test with the women (N=7) who completed the pilot programme. The women were all between 35 and 65 years of age and all were current clients with YANA, a north Cork domestic violence project with outreach services. All had attended the 'Pattern Changing' programme (Goodman and Fallon 1994) or had received personal development training prior to undertaking the NEW START programme. The aim of this quantitative test data was to establish whether the training had been effective in empowering the women by improving their self-esteem, self-efficacy and entrepreneurial self-efficacy.

Findings

The findings of this research show that there is benefit in extending the training available to survivors of GBV by bringing in outsider trainers and by covering mainstream employability and entrepreneurship training materials. Also, for this cohort of women during this pilot, the benefit was mostly in the training process rather than in the mastering of the material covered. Similar training was provided by the other European partners. (*– awaiting final report to insert overall result summary*) In general, demand for entrepreneurship in the context of starting up a business was limited, but entrepreneurial skills were relevant.

Research Implications/Limitations

The findings are based on a single implementation of the programme. The researchers were actively involved in all stages of the project. However, the results obtained were in line with those found by the six other European partners contributing to the NEW START program. However, due to the limited number (N=7) who received training in Ireland and only one group implementation, no generalisable conclusions can be drawn. The programme developed should be tested with further groups to establish reliability of this means of empowerment.

Originality/Value

The added value of the Irish NEW START course contribution is that it is a programme that directs itself to the subject of 'moving on' from being defined by the experience of GBV. This programme was in effect a transitional bridge between the safe and known environment of an outreach provider and the greater world. It provided a space where the women could still check-in while also checking out other wider options.

References

Blaschke, L. M. (2012). "Heutagogy and lifelong learning: A review of heutagogical practice and self-determined learning." The International Review of Research in Open and Distributed Learning **13**(1): 56-71.

Charmes, J. and S. Wieringa (2003). "Measuring women's empowerment: an assessment of the gender-related development index and the gender empowerment measure." Journal of Human Development **4**(3): 419-435.

Goodman, M. S. and B. C. Fallon (1994). Pattern changing for abused women: An educational program, Sage Publications.

Keegan, S. (2009). "Emergent inquiry." Qualitative Market Research: An International Journal **12**(2): 234-248.

McWilliams, M. and J. McKiernan (1993). Bringing it out in the open: domestic violence in Northern Ireland, HM Stationery Office.

Mezirow, J. (1991). Transformative dimensions of adult learning, ERIC.

Rosenberg, M. (1965). "Rosenberg self-esteem scale (RSE)." Acceptance and commitment therapy. Measures package **61**.

SAFEIreland (2016). The State We Are In In 2016: Towards a Safe Ireland for Women and Children. Athlone, Co. Westmeath, SAFE Ireland.

Schwarzer, R. and M. Jerusalem (2010). "The general self-efficacy scale (GSE)." Anxiety, Stress, and Coping **12**: 329-345.

Voogt, J., H. Westbroek, A. Handelzalts, A. Walraven, S. McKenney, J. Pieters and B. de Vries (2011). "Teacher learning in collaborative curriculum design." Teaching and Teacher Education **27**(8): 1235-1244.

A Serious Game to Dissuade Adolescents from Taking Performance Enhancing Drugs

Larkin Cunningham
Department of Computing
CIT – Cork Institute of Technology, Ireland
e-mail: larkin.cunningham@cit.ie

Keywords: Serious Games, Education, Performance Enhancing Drugs

Can a serious game that invites players to explore the morality of performance enhancing drugs (PEDs) from several perspectives (athlete, federation, wider society) be effective in changing the attitudes of athletes and non-athletes to PEDs?

Performance enhancing drugs (PEDs) are a major issue not just in professional sports, but also in amateur and recreational sports, including the use of androgenic anabolic steroids (AAS) in gym culture (Hartgens and Kuipers 2004). Studies have shown that between 7% and 12% of young athletes would choose to dope using an illegal substance if it guaranteed improved performance, could not be detected and did not adversely affect lifespan (Bloodworth *et al*. 2012; Connor *et al*. 2013).

The abuse of non-prescribed PEDs, such as AAS, is a public health concern that requires the implementation of education programmes aimed at adolescents (Tahtamouni *et al*. 2008); research has shown that such programmes can be effective in reducing the intentions of adolescent athletes to use supplements (DesJardins 2002).

Michael and Chen (2005) offer the qualified definition of serious games as "game[s] in which education (in its various forms) is the primary goal, rather than entertainment", though they do concede that games intended primarily for entertainment purposes can contain serious game elements. Serious games have seen rapid growth for more than a decade in both academia and industry, in areas such as education and cultural heritage, well-being and health care, and they have huge potential in improving achievements in these areas and others (Laamarti *et al*. 2014).

One existing serious game, Play True Challenge, was developed for the 2010 Youth Olympic games (World Anti-Doping Agency 2016). It is a small-scale cartoon-like 2D adventure game featuring a sport similar to steeplechase at its core. Players learn about banned substances like erythropoietin (EPO), steroids and stimulants; they can take these in the game and discover the consequences later, such as being dropped from the athletics team.

A new serious game is under development, aimed primarily at adolescents, but also older athletes and non-athletes, to inform them about the consequences of taking PEDs. It will be more extensive than WADA's Play True Challenge and use updated technology to explore a wider range of moral grey areas of PED usage. Its primary contribution will be an examination of the efficacy of a serious game in bringing about attitudinal change with respect to athlete doping and the tolerance of wider society to the use of PEDs.

Using Unreal Engine, the engine used in commercial successes like Gears of War and Tom Clancy's Rainbow Six Vegas, the game is being developed using an immersive 3D environment in the first-

person perspective. Learning outcomes of the game are being informed by literature of the type presented in this short paper, as well as other educational materials on the positive and negative effects of PEDs. A number of popular game mechanics are being employed, such as object interaction and collection, goal completion, challenges, puzzles, achievements and dynamic interactive dialogue.

Ultimately, the game is intended to be a persuasive one. Ian Bogost, in his book Persuasive Games (2007), defines procedural rhetoric as “the art of persuasion through rule-based representations and interactions, rather than the spoken word, writing, images, or moving pictures”. His interest lies in how videogames argue about the way systems work in the real world outside the videogame. A persuasive game on the subject of PEDs would, therefore, through the medium of the 3D first-person game under development, represent the reality of doping in sport, how it can be nefarious and be protected by corruption, what the daily reality of doping can be like for an athlete, the devastating long-term effects it can have on personal health and reputation.

The main protagonist in the game (the primary player character) is an investigative journalist piecing together the story of doping in sport for an exposé article. He interacts with a number of adolescents with hopes of future stardom, existing athletes and retired athletes (see Fig. 1 for an example of this interaction through dynamic dialogue). The game will feature some flashbacks from the perspective of the athletes, where you as the player will make choices, such as whether to take a banned substance or not, and see how this affects the athlete’s story in the present. This will have a dynamic effect on the final exposé article.

The game will be modular in nature, allowing for it to be customized for different sports if required. Characters can be swapped for others, dialogue can be edited separately from the game engine code, it can be localized for different languages, and so on.

Development of the game is underpinned by the established Game Object Model (GOM) version II theoretical framework (Amory 2007), which uses the object-oriented programming metaphor. The model describes a number of interfaces to game elements that can be abstract (pedagogical and theoretical constructs, such as critical thinking, emotiveness, goal formation and goal completion) or concrete (design elements, such as story, plot, backstory, graphics and sound). These are then mapped to the intended learning outcomes and the game mechanics.

Aside from outlining the background of this project, this short paper is also a call for collaboration. The development of a persuasive videogame set in an immersive 3D environment is an ambitious and multifaceted one. Researchers from a variety of backgrounds, such as pedagogy, psychology, sport science, digital arts, game studies and other areas could contribute.

In conclusion, there is an obvious problem with PEDs, made all the more topical by recent scandals, such as in athletics (Russia banned from athletics in 2016 Olympics), tennis (Maria Sharapova banned for two years) and mixed martial arts (Brock Lesnar testing positive after earning $2.5 million for a fight at UFC 200). Education on the subject of PEDs and other supplements has been shown to be effective. Serious games are a growing force in education and have the potential to greatly improve educational outcomes. A persuasive serious game can show how systems in the real world work, preparing the adolescents who play the game for the choices they will face and be aware of the consequences of their choices.

Fig. 1 – A screenshot showing dialogue from an early prototype

References

Amory, A. (2007) 'Game Object Model Version II: A Theoretical Framework for Educational Game Development', *Educational Technology Research and Development*, 55(1), 51–77.

Bloodworth, A.J., Petróczi, A., Bailey, R., Pearce, G., McNamee, M.J. (2012) 'Doping and supplementation: the attitudes of talented young athletes', *Scandinavian Journal of Medicine & Science in Sports*, 22(2), 293–301.

Bogost, I. (2007) *Persuasive Games: The Expressive Power of Videogames*, MIT Press: Cambridge, MA, USA.
Connor, J., Woolf, J., Mazanov, J. (2013) 'Would they dope? Revisiting the Goldman dilemma', *British Journal of Sports Medicine*, 47(11), 697–700.

DesJardins, M. (2002) 'Supplement use in the adolescent athlete', *Current Sports Medicine Reports*, 1(6), 369–373.

Hartgens, F., Kuipers, H. (2004) 'Effects of Androgenic-Anabolic Steroids in Athletes', *Sports Medicine*, 34(8), 513–554.

Laamarti, F., Eid, M., Saddik, A.E. (2014) 'An Overview of Serious Games', *International Journal of Computer Games Technology*, 2014.

Michael, D.R., Chen, S.L. (2005) *Serious Games: Games That Educate, Train, and Inform*, Course Technology / Cengage Learning: Boston, MA, USA.

Tahtamouni, L.H., Mustafa, N.H., Alfaouri, A.A., Hassan, I.M., Abdalla, M.Y., Yasin, S.R. (2008) 'Prevalence and risk factors for anabolic-androgenic steroid abuse among Jordanian collegiate students and athletes', *The European Journal of Public Health*, 18(6), 661–665.

World Anti-Doping Agency (2016) Play True Generation [online], available: https://www.wada-ama.org/en/play-true-generation [accessed 25 Jul 2016].

A User Engagement framework for Universal Inclusion

A User Engagement framework for Universal Inclusion
Alex Vakaloudis, Mark Magennis, Martina Hayes, Jane O'Flynn, Kieran Delaney
Nimbus Centre. CIT – Cork Institute of Technology, Ireland
e-mail: mark.magennis@mail.com, martina.hayes@mycit.ie, {janeoflynn,alex.vakaloudis, kieran.delaney}@cit.ie

Keywords: User Engagement, social inclusion, cross-sectoral project management

Introduction

Numerous innovative products, albeit pioneering, have not been successful because they have not being adopted by their end-users as they disregard their functional or interfacing needs. Likewise, the wishes of end-users are not easily or effectively communicated to researchers and industry innovators who can make use of technological advances so as develop prototypes and eventual products to correspond to these needs. The solution is to create a sustained process that engages at an early stage the cross-sectoral characteristics of a project and involve all stakeholders in fruitful and effective collaborations. In this paper we summarise the development of User Engagement Frameworks design to achieve this in a sustained manner.

Related work

There is a huge increase in the involvement of communities of people in cross-sectoral projects. There are significant drivers in the academic world (national civic engagement targets) [1][6], the corporate world (emphasis on CSR benefits) [4][5] and the public sphere (push for more effective forms of public engagement) [2][3].

Concepts and Approach

The aim of this effort is to provide the tools so as to support the management of cross-sectoral collaborative projects where at least one partner is an academic institution and the other partner(s) may be from the academic, corporate or public sectors. The end-users need to be involved in all stages of the project lifecycle from requirements elicitation through to specification of user-experience to proof of concept and prototype trialling. Likewise industrial stakeholders encompass a pivotal role to realistic product development and possible commercialisation.

Apart from offering project management, the main objectives are to improve project outcomes by reducing the risk of one or both partners not achieving their goals, to save time and cost by allowing partners to store and re-use successful project processes and supporting materials and to enable organisations to develop their project management capabilities to benefit similar future projects.

Methodology

We used interviews and focus groups to ascertain each stakeholder's requirements for how a resulting collaborative platform should work and what functionality it should offer. The unique challenges in cross-sectoral engagements are:

- Finding the right opportunities and partners based on their skillsets, availability and compatibility
- Understanding each other's aims and expectations and achieving a shared vision of the goals of the project
- Understanding what each partner can contribute and under which conditions
- Organising activities, working together effectively and managing interactions

A platform to meet these challenges consists of the following functional modules

- User Management Module to manage the profiling attributes of participating organisations and participants and assign appropriate access rights for each project
- Dialogue Module to implement messaging among team members' preferred channels (e.g. email, SMS, WhatsApp, Facebook) in order to schedule and realise (e.g. interaction activities such as discussions, meetings and interviews and the creation of demos/prototypes of any level of maturity. Furthermore to give a mechanism for everyone outside academia to express needs and research ideas.
- Project Management Module to create a project agenda, allocate tasks including any documents (e.g. the informed consent of the participants) and organise activities with relevant information such as aims, date, time and location, required participant demographics, abilities/disabilities, digital skill level, use of assistive technologies, level of compensation offered, etc.
- Dissemination Module to provide metrics on the user engagement.

Case Studies

Two case studies implementing this work will be described:

ASPIRE (Application to Support Participant Involvement in Research) is a programme for universal inclusion targeting projects that involve people with disabilities. The intention of ASPIRE is to support their inclusion in research, co-design, social campaigns and other activities, regardless of their physical, social or economic abilities. Apart from identifying the functional modules for platform to fulfil these needs, ASPIRE highlighted the cyber-physical characteristics of such interaction and identified the importance of multiple and novel interaction methodologies such as smart/wearable devices and smart surfaces.

Melu is a web-based platform for cross-sectoral projects that enables detailed and flexible member profiling with skills, capabilities, conditions and requirements in order to achieve optimum matching for collaborations and activities. Apart from managing projects, Melu aims to build a knowledge base of procedures/materials to improve the formation of successful collaborations.

Conclusions and future work

This is a system that takes the pain out of cross-sectoral project management, improves project outcomes, reduces the risk of one or both partners not achieving their goals and saves time by allowing partners to save and re-use successful project processes and supporting materials. There is scope for additional future work. This refers to the broad adoption of IoT technologies by understanding and overcoming practical and cultural barriers and the detailed customisation of the interface depending on the micro-profiling of participants,

Acknowledgements

ASPIRE was developed as part of funding by the Irish National Council under the New Foundations programme in collaboration with NCBI.

References

1. ViPi (Virtual Portal for Interaction and ICT Training for People with Disabilities) KA3 LLL project Lifelong Learning program, subprogramme KA3 ICT (511792-LLP-1-2010-1-GR-KA3-KA3NW). https://www.vipi-project.eu.

2. e-LIFE (e-Learning Virtual Center for People with Physical Disabilities) European Union, Lifelong Learning Programme 2007-2013, Leonardo Da Vinci Sub-Programme, Transfer of Innovation Code: 2011-1-RO1-LEO05-15322. http://www.elife-eu.net.

3. University of Cambridge. Inclusive Design Toolkit. http://www.inclusivedesigntoolkit.com

4. IDC, Worldwide Quarterly Mobile Phone Tracker http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37

5. http://www.doro.ie/

6. J.Ye, G. Stevenson, and S. Dobson (2014). KCAR: A knowledge-driven approach for concurrent activity recognition. Pervasive and Mobile Computing, 201

Building capacity when assessing non-formal and informal learning through collaborative networks for Recognition of Prior Learning

P. O'Leary
Cork Institute of Technology, Ireland
e-mail: phil.oleary@cit.ie

A. Ledwith
Adult & Continuing Education
University of Limerick, Ireland

Keywords: Recognition of Prior Learning, Non-formal and informal learning, values, beliefs, assumptions

The current European policy climate is intended to create conditions where Recognition of Prior Learning (RPL), can flourish within the formal learning system, in turn supporting lifelong learning across the European Qualifications Frameworks.

However, RPL is a complex brief to deliver in practice. Academic assessors often find that assessing non-formal and informal learning is difficult and time consuming as there is so much variety within each RPL case. Similarly, the individual story of the RPL candidate presented for assessment is unique to themselves; the language used, evidence presented, the reflective element within the assessment piece and how they position their own case against a set of standards will be different each time. Such diversity does little to reassure academic assessors who commonly hesitate when evaluating RPL cases.

Literature reports cultural acceptance of RPL across an institution is a critical factor cultivating conditions for pedagogic agency for RPL (Cooper and Harris, 2013). This research takes a novel approach to supporting RPL provision by exploring the values, beliefs and assumptions of assessors and candidates in order to better understand the perspectives of both actors (Friesen, 2011).

This research in progress, has taken a constructivist grounded theory approach to explore the various perspectives at play, through a study of the values, beliefs and assumptions of academic assessors and candidates for RPL (Charmaz, 2006, Bryman, 2012). The setting for the research is within one third level institution in Ireland. Semi-structured interviews were held in 2015 with 31 academic assessors. In the current year, 27 past candidates of RPL have been interviewed using the same questions around values and beliefs. Aspects of Bernstein's theories have provided the theoretical framework to support analysis of data, including the idea of the classification of knowledge, the totally pedagogised society, the field of recontextualisation where knowledge is repositioned to facilitate entry to the formal learning system through assessment for RPL (Bernstein, 2000). Bernstein's theories are about language, its codes and rules, and how the structure of communication influences an individual's identity. Inherent within RPL are tensions arising from the repositioning of every-day working knowledge into a form which is suitable for validation by the formal learning system. Comparison of the data arising from the two sets of participants has allowed for the identification of the various perspectives at play within RPL.

Results reveal that the academic assessor looks first to maintain the standards, however they are willing, where merited, to step towards supporting the candidate's claim, revealed as the act of 'balancing'. Results also show that candidates possess values, beliefs, and assumptions that strongly defend their reasons for applying, and the validity behind the basis of their case. The challenge for the formal learning system is to harness this insight to deliver more effective RPL, one where the various perspectives are acknowledged. It is envisaged that this research will build a set of recommendations around supporting the confidence and capability of the main actors within RPL.

This research concludes that building capacity through collaborative networks will support assessors and RPL candidates and nurture capability for assessing non-formal and informal learning in practice. The recommendation of the Council of Europe in December 2012, calling for the realisation of RPL provision within the formal learning system by 2018, behoves further research into the socio-cultural aspects around RPL, as these are critical to its successful adoption to mainstream provision.

References

BERNSTEIN, B. 2000. *Pedagogy, sybolic control and identity; Theory research and critique,* Lanham, MD, Rowman & Littlefield.

BRYMAN, A. 2012. *Social Research Methods,* Oxford, Oxford University Press.

CHARMAZ, K. 2006. *Constructing Grounded Theory: a Practical Guide through Qualitative Analysis,* London, Sage.

COOPER, L. & HARRIS, J. 2013. Recognition of prior learning: exploring the 'knowledge question'. *International Journal of Lifelong Education,* 32**,** 447-463.

FRIESEN, N. 2011. Endword: Reflections on research for an emergent field *In:* J. HARRIS, M. BRIER & C. WIHAK (eds.) *Researching the Recognition of Prior Learning; International Perspectives.* Leicester: NIACE.

Chapter 7

Applied Computing

Full Papers

A Clinical Decision Support System for Personalized Medicine

Johannes Idelhauser, Bernhard G. Humm, Ulrich Beez
Department of Computer Science
Hochschule Darmstadt – University of Applied Sciences
Haardtring 100, 64295 Darmstadt, Germany
e-mail: johannes.idelhauser@stud.h-da.de, {bernhard.humm / ulrich.beez}@h-da.de

Paul Walsh
NSilico Lifescience Ltd.
Bishopstown, Co. Cork, Ireland
e-mail: paul.walsh@nsilico.com

Keywords: Clinical Decision Support System, Electronic Health Record, Information Retrieval

Abstract: Given the rapidly growing number of medical publications and resources, consultants face challenges in keeping up-to-date with current research and patient care best practices. This paper presents the concept and prototypical implementation of a Clinical Decision Support System (CDSS) for personalized medicine. It satisfies information needs of consultants at the point of care by integrating secondary medical resources based on a concrete patient's Electronic Health Record (EHR). Particular focus has been on the usability of the CDSS allowing consultants to quickly and intuitively gather relevant information with minimal system interaction. An initial assessment of the CDSS by medical professionals indicate its benefit.

1 Introduction

Given the rapidly growing number of medical publications and resources, consultants face challenges in keeping up-to-date with current research and patient care best practices. Ongoing research and new treatment options require practitioners to keep up-to-date to ensure good treatment outcomes and prevent malpractice lawsuits (Marchant and Lindor, 2013). Physicians' information needs at the point of care range from refreshing and confirming knowledge over logistical questions like drug dosage to teaching and personal learning (Maggio et al., 2014).

Personalized medicine aims at tailoring medical decisions, practices, interventions or products to the individual patient based on their predicted response or risk of disease (Academy of Medical Sciences, 2015). While tailoring the treatment to individual patients is common practise in medicine, the term has been recently used for informatics approaches in medicine that use large amounts of patient data, particularly genomics data, for selecting appropriate therapies.

This paper presents the concept and prototypical implementation of a Clinical Decision Support System (CDSS) (Kawamoto et al., 2005) for personalized medicine. It satisfies information needs of consultants at the point of care by aggregating and integrating primary and secondary medical resources based on a concrete patient's Electronic Health Record (EHR), thus paving the way for personalized medicine. The CDSS is intended to be integrated into an EHR application to be used at the point of care.

The remainder of this paper is structured as follows. Section 2 specifies the problem statement in terms of requirements. Section 3 is the core of the paper describing the concept and prototypical implementation of the CDSS. Section 4 evaluates our approach. Section 5 compares our approach with related work. Section 6 concludes the paper and indicates future work.

2 Problem Statement

Having consulted extensively with clinicians involved in the treatment of melanoma, we have identified the following requirements:

1. Functional Requirements
    1. Relevant: The CDSS shall satisfy the consultants' information demand with relevant, helpful and latest medical information.
    2. Personalized: The information provided shall be tailored to the medical condition of a particular patient.
    3. Pro-active: The CDSS shall offer information pro-actively without additional data entry by the user.
    4. Easily comprehensible: shall provide a quick overview of all information available as well as the possibility to easily acquire more detailed information where needed.
    5. Workflow: The CDSS shall not interfere with the consultant's EHR workflow.
2. Non-Functional Requirements
    1. Usable: The CDSS shall be intuitive to use and self-explanatory.
    2. Low response time: The response time for all interactions with the CDSS shall be less than 1s.
    3. Extensible: The ongoing extension of the CDSS with new information sources shall be facilitated with moderate implementation effort.

3 A Clinical Decision Support System for Personalized Medicine

3.1 User Interaction Model

Physicians' information needs at the point of care include refreshing and confirming knowledge and logistical questions, e.g., medication dosage, idea generation and personal learning (Maggio et al., 2014). To satisfy these needs, physicians tend to access primary literature in the form of abstracts and full-text as well as secondary literature in the form of summaries and reviews (Maggio et al., 2013). In order to provide an intuitive way for physicians and other health professionals to access this information, the proposed CDSS leverages several information services that each try to satisfy different information needs. Those information services are organized in web page panels that the users can customize and fit to their needs by deciding which service panels should be displayed and which should be hidden. Additionally the order and size of the panels can be adapted to the user's individual needs while the resulting layout is persisted over different sessions for the individual user.

3.1.1 Literature Service

One of the main services in the CDSS is the literature service to find and display relevant primary medical literature that is related to the patient at hand (Figure 1).

Figure 1. Interaction concept to display related medical publications (literature service)

The literature service displays automatically generated filters to quickly navigate the literature search results. The filters are displayed on the left whereas the medical literature is shown on the right. For each medical publication its title, journal and publication date is displayed. In the context of evidence-based medicine (EBM), publications with a high degree of evidence are to be prefered in patient care (Hung et al., 2015). As such, publications that are reviews or clinical trials are shown with a marker indicating their publication type. This also aligns with a study from 2013 that logged and analysed data queries in a hospital and found that "[a]lmost a third of the articles [...] accessed were reviews." (Maggio et al., 2013). For quick orientation and relevance assessment, terms that appear in the patient's EHR are highlighted in the literature service. To help the literature relevance assessment process, a teaser text is displayed when hovering the mouse pointer over the eye icon after each publication title. In order to give the users a way to give feedback on the relevance of a publication and improve the literature search, icons with a thumbs-up and a thumbs-down are provided.

3.1.2 Evidence-Based Medical Recommendations

Evidence-based medicine describes the assessment and use of the best available research for decision making in patient treatment and diagnosis. This is done by focusing on well-designed, conducted research with a strong level of evidence like systematic reviews or randomized controlled trials (Hung et al., 2015). Existing web-based clinical decision support systems specialize on providing evidence-based treatment guidelines and summaries written by medical experts that reflect the current state of research. Obst et al. (2013) assessed the use of such a service named UpToDate.com and suggested that physicians often have little time and that navigating the service could at times be time consuming. They also noted that the summaries were sometimes written confusingly and that it took too long to quickly grasp the desired information.

The EBM recommendation service (Figure 2) therefore queries different secondary information sources for patient-related evidence-based summaries and reviews and extracts the most important sections that describe the patient's issue. If the users wish to see the extracted sections in context, they can follow the links and visit the original text.

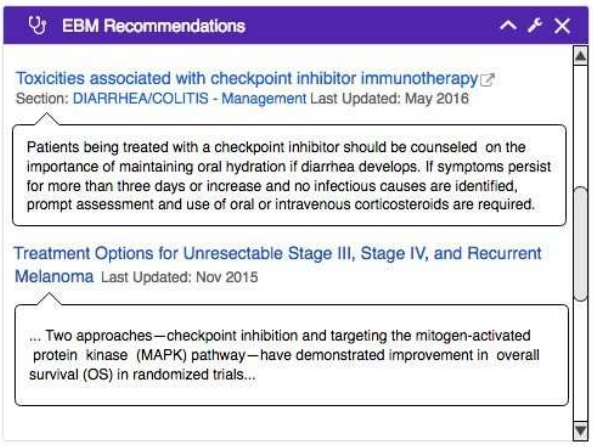


Figure 2. EBM recommendations service

3.1.3 Drug Information Service

Other important information in patient care is material on drugs and their interactions "at the point of drug prescribing" (Rahmner et al., 2012). Therefore, the drug information service provides information normally available in medication package leaflet inserts and secondary decision support services in a more accessible and structured way (Figure 3, left). The provided information includes dosage data for different age groups and pre-filled calculators to compute the correct dosage based on the age and weight of the patient. Other information consists of warnings, adverse effects, pregnancy, pharmacology, administration guidelines, material for patient education and pill images and prices. Selecting a drug for displaying can be done in an autosuggest-supported field that ranks already prescribed medication higher, but allows also searching for medication not yet prescribed.

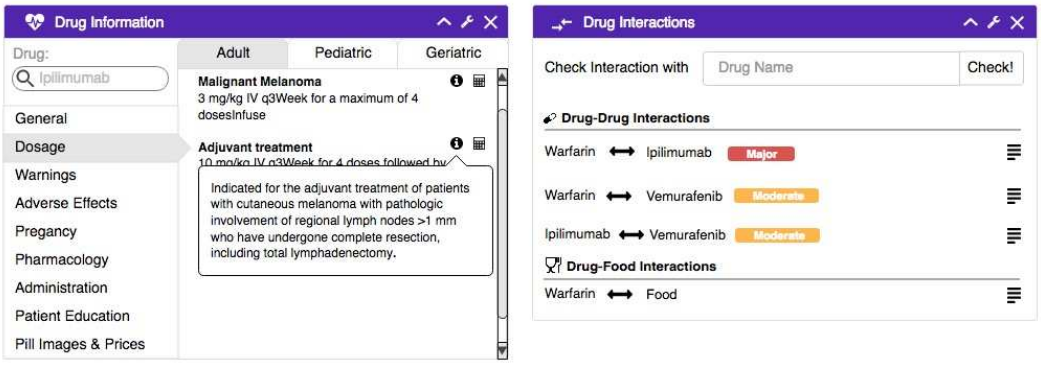


Figure 3. Drug information service

As physicians indicated they wanted to see automatically generated alerts for severe drug interactions and adverse effects (Rahmner et al., 2012), an alert is displayed prominently (Figure 3, top) . For more information on how to manage the interaction or alternative drugs, an appropriate link is provided. Non-severe drug interactions as well as drug-food interactions are displayed in an own panel where the users have the possibility to check interaction with other, not yet prescribed drugs (Figure 3, right).

To "make drug information more searchable" (Rahmner et al., 2012) and for example allow checking if a patient's symptom could be drug related, an adverse effects panel is introduced

(Figure 4). It automatically identifies drug-related comorbidities that are registred in the EHR but also allows searching for symptoms not yet recorded. An option to read more provides information on how to manage this effect, when it will occur or how long it will last.

Figure 4. Searchable Adverse Effects

3.1.4 Clinical Trials Service and News Service

Especially in cancer care the participation in clinical trials is an option for patients of all clinical stages as new findings lead to the development of many new drugs (National Cancer Institute, 2016). A clinical trials service is therefore introduced that searches for nearby clinical trials fitting the patient (Fiure. 5, left).

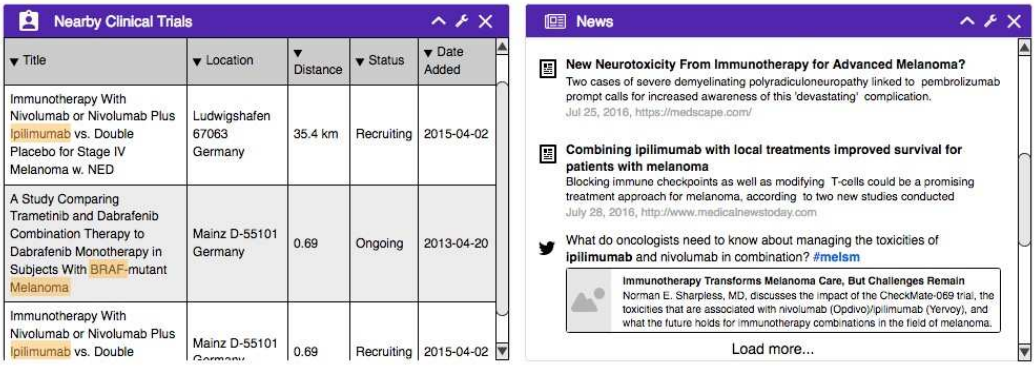


Figure 5. Clinical trials locator service and news service

Finally, a news service provides recent news on treatments, drugs, legislative information or other scientific breakthroughs that can be related to the current EHR (Figure 5, right).

3.2 Information Sources

In order to provide the CDSS services, several potential information sources were identified. The selection of information sources is a critical part in any CDSS application as they are the basis for the trust given them by physicians. As a patient's health might also depend on this information, the quality and correctness of the information is obligatory. This section introduces a summary of identified potentially relevant sources for CDSS.

3.2.1 Literature Service

There are an abundant amount of databases and search services for the health domain. Table 1 lists only a few of them that would be an option to use for the literature service.

Name	Description	API	Access	Size
Google Scholar	Search engine for scientific publications of all fields. Automatically crawls many journals.	no	commercial	estimated at 160 million articles
Ovid	Science search platform that includes many databases, including MEDLINE.	?	subscription	?
PubMed	Search engine mainly accessing MEDLINE database and focused on health topics. Query expansion by use of MeSH ontology.	yes	public & free	> 24.6 million records, about 500,000 new records each year
ScienceDirect	Website with access to large database of scientific publications from many fields.	yes	free (abstracts), subscription (full-text)	12 million records from 3,500 journals and 34,000 eBooks
Scopus	Database with abstracts and citations from many academic journals and many scientific fields, not focused on health topics.	yes	paid subscription	~55 million records
Springer API	Access to all Springer published journals, also includes BioMedCentral open-access publications.	yes	partly free, partly subscription	~2,000 journals and >6,500 books per year, access to >10 million online documents

Table 1. Literature data sources

3.2.2 Evidence-based Medical Recommendations

As evidence-based reviews or summaries should be written by medical experts to ensure quality and reliability, most services require a commercial licence or paid subscription to access them. Only a few public and free sources could be found in the scope of this thesis (Table 2).

Name	Description	API	Access	Volume
BMJ Best Practice	Evidence-based information to offer step-by-step guidance on diagnosis, prognosis, treatment and prevention.	yes	subscription	?
DynaMedPlus	Evidence-based clinical overviews and recommendations. Content updated daily. Also offers calculators, decision trees and unit and dose converters.	yes	subscription	> 3,200 topics and > 500 journals
EBMeDS	Platform-Independent web service CDSS with EBM module	yes	commercial	
Medscape / eMedicine	Largest clinical knowledge base available freely. Articles updated yearly. Also available as mobile application.	no	free, registration required	~6,800 articles
Physician Data Query	Cancer database from the U.S. *National Cancer Institute*. Contains peer-reviewed information on cancer treatment in the form of summaries for patients and professionals.	no	public	Only cancer domain
UpToDate	Popular evidence-based POC tool for a wide range of disciplines but targeted on internal medicine. Extensive peer-review process to ensure accurate and precise recommendations.	yes	subscription, some articles free	~8,500 topics

Table 2. EBM resources

3.2.3 Drug Information

There are few public resources for drug information that are also accessible via an API. Table 3 shows services that could be used in the context of the CDSS proposed in this work. Other not listed services include DrugBank (accessible over RxNav), ResearchAE.com, SIDER, NDF-RT, Epocrates or the OpenFDA API.

As Peters et al. (2015) stated, the data quality of some public resources might be a problem in clinical settings. They compared the two public drug interaction services DrugBank and NDF-RT and found a limited overlap between their drug interaction information. The commercial service used to compare against provided better coverage of the test data then both of the free services

combined. Assuming the commercial drug information services provide a better data quality, one would have to select one of those as data source. Additionally, as of September 2016, the public data source NDF-RT removed the drug interaction information from their service.

Name	Description	API	Access	Drug Information	Drug Interactions	Adverse Events	Drug Announcements/Recalls
DailyMed	Website by U.S. National Library of Medicine (NLM), provides high quality and up-to-date drug labels. Updated daily by FDA. Documents use structured XML format.	yes	public & free	✓	✓	✓	
MedlinePlus Connect	Service by NLM, provides unstructured natural language drug information/labelling and health topic overviews	yes**	public & free	✓*		✓*	
Medscape	Many clinical information resources available over website or mobile app. Articles updated yearly.	no	free, registration required	✓	✓	✓	
RxNav	Provides access to different drug resources like *RxNorm*, *NDF-RT* and *DrugBank*. Drug normalisation over different codes and systems by using *RxNorm*, drug interactions from *DrugBank*.	yes	public & free		✓		
Wolters Kluwer Clinical Drug Information	Commercial drug information APIs including interaction, adverse effects, indications and mapping to RxNorm.	yes	commercial	✓	✓	✓	

Table 3. Drug information sources

3.2.4 Clinical Trials

Several information sources for clinical trials were identified (Table 3).

Name	Description	API	Access	Type	Country
ClinicalTrials.gov	Trial registry from US National Institute of Health. 39% are U.S. only trials.	no*	Public & free	Register	Worldwide with a focus on U.S. (39%)
EU Clinical Trials Register	Clinical Trials Register for trials in the EU	no*	Public & free	Register	European Union
German Clinical Trials Register	German clinical trial register. Also imports trials from clincaltrials.gov that are located in Germany	no*	Public & free	Register	Germany
WHO International Clinical Trials Registry Platform	Search portal to central database with links to original records. Regular fetch of trials from currently 16 data providers, including sources mentioned earlier	no*	Public & free	Search Service	Worldwide

Table 4. Potential clinical trials sources

As the clinical trials' location plays a vital role in assessing their relevance, a service is needed that is not solely focused on one or a few countries. As the WHO registry platform seems to aggregate data from many national registers and other services, the choice would most likely fall on the WHO service.

3.2.5 News

Many online services offer medical news, e.g. MedScape, ScienceDaily or Medical News Today. Usually their content is not accessible over an API and there can also be legal issues prohibiting any "unauthorized copy, reproduction, distribution, publication, display, modification, or transmission of any part of this Service" (ScienceDaily, 2016).

An alternative to this might be the online social networking service Twitter which is used by many medical news sites and professionals.

3.3 Software Architecture

The CDSS is intended to be integrated into an EHR application. The application is organized in a three layer architecture where each layer consists of components that encapsulate logically separable units (Figure 5).

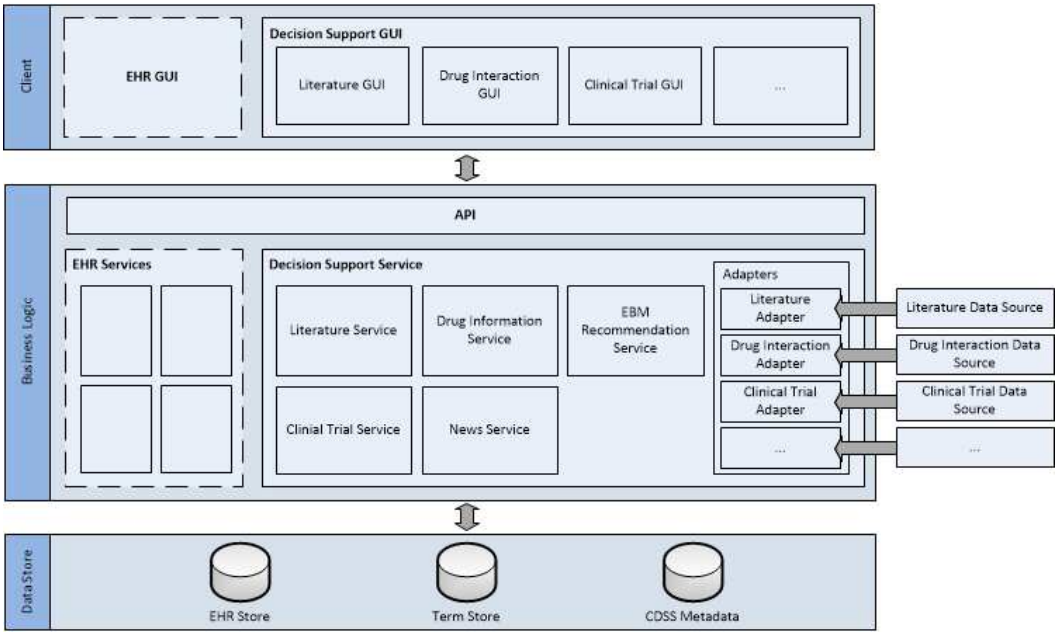


Figure 5. System architecture

On the client side, the Decision Support GUI is implemented alongside the EHR GUI. The different CDSS service are subcomponents of a single decision support module.

The business logic of the Decision Support System consists of the different service components as well as different adapters to connect to the information sources. The services store and use data from the EHR and the Term Store and persist data in the CDSS Metadata store.

3.4 Prototype Implementation

The literature service and drug information service have been implemented prototypically as part of an EHR application for melanoma treatment (Humm and Walsh, 2015) (Beez et al., 2015). The application is implemented in C# using .NET and MS SQL Server on the server side, and in HTML5 / CSS / JavaScript on the client side, using Bootstrap and AngularJS. As data source for the literature service PubMed was selected. Drug interaction data is acquired by querying DrugBank

over the RxNav drug interaction API. In the following sections, we describe the implementation of the literature service in more detail. See Figure 6 for the detailed architecture.

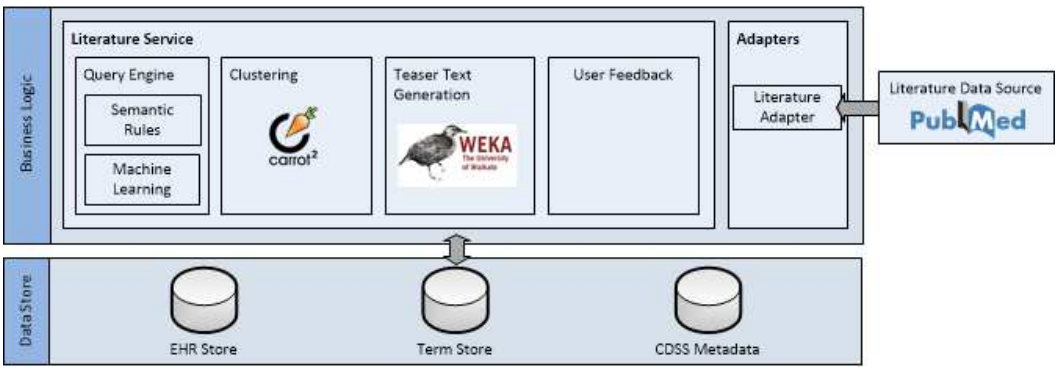


Figure 6. Literature service architecture

3.4.1 Query Engine

As PubMed is selected as the source for the literature service, we need a query in the literature source's query language to find and display publications. To generate the query from the EHR, two strategies are employed: use of semantic rules for creating queries from EHR attributes and machine learning.

Semantic rules

From the ca. 100 attributes that are currently used in the EHR application, not all are helpful for getting personalized literature suggestions for a concrete patient (Figure 7).

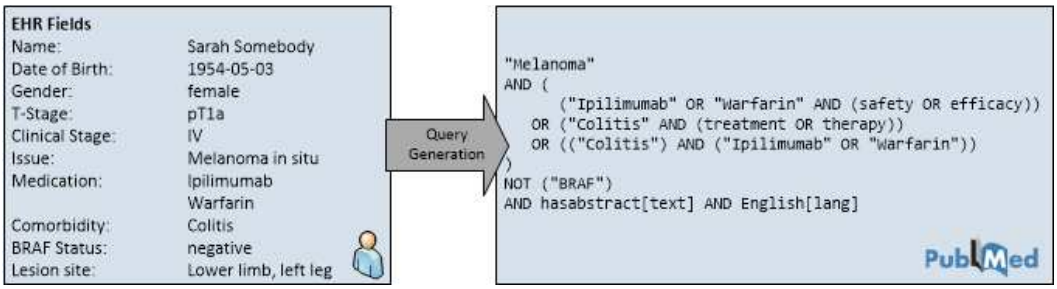


Figure 7. Sample query generation from an EHR

Fields with no relevance like the patient's name are omitted in query generation whereas relevant fields like the issue, medications or comorbidities are included using rules. One rule to search for publications that address the safety or efficacy aspects of one of the medications prescribed, combines all medication with an "OR" and adds "(safety OR efficacy)" to the medication subquery. Another rule combines the comorbidities field with the medication to search for drug-related adverse effects and their treatment. To ensure data quality and only search for recent literature, restrictions are added to the query like the "hasabstract[text]" to only show publications that contain an abstract.

Machine Learning Approach

The presented rule-based approach returns good results in various cases as fields like comorbidity, medication and issue are always relevant. However, in using a rules approach only a predefined field of possible questions can be answered and other fields like age, gender or the clinical stage could also in certain contexts return relevant literature. Additionally, the above mentioned terms "safety" and "efficacy" are only one example of additional query terms. Other terms that are not captured in the semantic rules could also become relevant. We therefore apply a query expansion method by automatically refining queries based on user feedback. This is done using a machine learning approach. Examples for additional query terms from the current prototype implementation include "adverse effects" and "drug therapy". Also, other attributes may become relevant in the future, e.g., the publication type "Review".

To facilitate machine learning, Accord.NET's [1] one-against-all multi label Support Vector Machine (SVM) is used to predict suitable search terms. Training data consists of user feedback data on the relevance of a literature given a specific EHR. As input vector, a bag-of-words (BoW) is built from all EHR terms. The output labels are extracted from the literature using the EHR's medical ontology.

Both strategies are used in parallel and their results are combined using an heuristic approach. The final query is passed to the literature adapter that retrieves the publications from PubMed and parses the retrieved XML document to an internal object representation. Publications are then ranked utilizing their PubMed rank in combination with explicit and implicit user feedback.

3.4.2 Clustering

Different users are interested in different topics and want to have different questions answered. A one-fit-for-all ranking is, therefore, not sufficient. To accommodate this, the result set is clustered to allow filtering the publications according to different semantic criteria. Example filter labels include "Ipilimumab-induced Colitis", "Adverse Effects" and "Overall Survival" (Figure 8).

Figure 8. Partial screenshot of automatically generated filter labels

For clustering the publications, the open source search engine clustering server "Carrot2" is used. It utilizes the specialized "Lingo" algorithm that automatically creates meaningful cluster labels from the publications' titles and abstracts (Osiński et al., 2004). The generated cluster labels are then filtered using the EHR application's medical ontology as well as a custom whitelist.

[1] http://accord-framework.net

3.4.3 Teaser Text Generation

If a publication title catches a consultant's attention, he may choose to read a teaser text in order to easily assess its relevance. This teaser text should be about three sentences long and contain the conclusion of the publication. Usually, abstracts consist of at least 500 or more words and not all of them have their conclusion displayed in a structured form. Therefore, for publications for which the conclusion is not explicitly marked, a machine learning algorithm is employed to predict the concluding sentences. The implementation uses the libSVM algorithm and the open source machine learning software Weka.

3.4.4 User Feedback

There are two kinds of user feedback gathered in this application, active and passive. Active feedback is generated by the users clicking on thumbs-up or thumbs-down icons in the client. Passive feedback is gathered by logging and interpreting all clicks on a publication. All feedback is stored in the CDSS Metadata store and is used for the ranking of publications and clusters as well as for creating training data for machine learning.

4 Evaluation

We compare concept and the prototypical implementation with the requirements stated in Section 2. Requirement 1.2 (personalized) and 1.3 (pro-active) are obviously met since the CDSS information is displayed automatically based on the EHR currently being handled. Also, requirement 1.5 is met, since the CDSS panels can be separated from the EHR dialogs. Assessing the Requirement 1.1 (relevant), 1.4 (easily comprehensible) and 2.1 (usable) is less obvious and needs feedback from users. Therefore, we have conducted an initial survey with 4 medical students and 1 resident physicians.

Users were given the task to prepare for a multidisciplinary team (MDT) meeting using the different CDSS modules. Subsequently, they were to fill out a questionnaire to assess the usability aspects and relevance of the CDSS services implemented in the prototype as well as in the interaction concept.

Initial usability observations indicate positive feedback. Positively feedback mentioned was the display of the teaser text for assessing literature relevance, the drug interaction information, searchable adverse effects as well as the automatically generated filters. The usability test also revealed some weak points that were, subsequently, improved. For example, the teaser text's icon position was initially not prominent enough and was easily overlooked. Therefore, its position was moved after the literature's title and a legend was added (Figure 1). However, the survey was only an initial one and relevance assessment by medical students might be skewed due to their lack of clinical experience. In the future, a comprehensive survey with physicians working in the melanoma domain is planned.

When opening EHR for the first time initial loading of literature data may take up to 15 seconds. However, as this happens asynchronously while the consultant is working with EHR this will not interfere with the EHR workflow. As soon as all data is loaded and the physician enters the CDSS, each interaction is less than 500 ms which meets Requirement 2.2 (low response time) clearly. With caching strategies this initial loading time may also be reduced.

Concerning Requirement 2.3 (extensible), the component-based system architecture and the use of adapters to access information sources enables the extension of the CDSS and implement other decision support modules with moderate implementation effort. For a new data source one would have to implement an additional adapter which results in about 100 lines of code (LOC). For a new decision support module the service on the server and the client GUI would have to be implemented. For the client this would require about 300 LOC. The server implementation depends on the module's complexity, e.g. the literature service is ~2,500 LOC whereas the interaction service is ~200 LOC.

5 Related Work

The idea of CDSS is not new and there exist many services accessible over the browser and/or smartphone applications. Their scope ranges from drug information, drug interaction and diseases to guidelines, pill identification and alternative medications. Example services include UpToDate, Epocrates, MedScape and First Databank. Often, the integration of theses services into an EHR system consists of providing a standard search field that enables the users to search and visit the CDSS service's main webpage. Some EHR systems like Athenahealth's EHR include context-sensitive drug monographs that provide information like dosing, adverse effects and other key safety data directly in the EHR [2]. However, there are systems that include include the patient's context in the CDSS search. The integration of UpToDate into various EHR systems provides such a future by displaying an info button in various EHR locations and providing an automatic search [3]. However, this function delivers standard UpToDate results pages and problems mentioned earlier in this work like confusingly written texts and difficulties navigating the long summaries remain (Obst et al., 2013).

Finland's Evidence based Medicine electronic Decision Support (Nyberg, 2012) is a platform-independent online service that accepts structured EHR data as input and returns clinical decision support data like links to guidelines, therapeutic suggestions, clinical reminders and alerts. These rules can be created by experts in a web-based editor and scoped per organization or globally. It can also populate forms and calculators with patient specific data. They do not include a literature search service but provide other services like the drug alerts that are similar to services presented in this work. Additional services like the knowledge assistant would be relevant in the context of this work and could later be integrated into our proposed CDSS.

Alternative approaches for the task of finding literature fitting to a patient's case has been described in different publications. Perez-Rey et al. (2012) propose a visual browser extension that allows the user to select a subset of extracted search terms from a natural language medical record. These selected terms will then be used to search PubMed. However, the selection of search terms is not automatic or pro-active as the user has to interact with the application to build the search.

[2] http://www.athenahealth.com/enterprise/epocrates/clinical-decision-support

[3] http://www.uptodate.com/home/hl7

Soldaini et al. (2015) propose a CDSS that tries to find fitting medical literature for medical case reports instead of EHRs. They consider the natural language case report as the query and apply query reformulation techniques like query reduction by identifying medical terms and expansion by using pseudo relevance feedback to build the search. They try to best answer the case report (~60 words) by providing relevant literature. In contrast to this work they do not use PubMed as search engine but use a local search server.

To the best of our knowledge, no system integrates a literature search system into an EHR to display relevant medical literature. Similarly, we believe there is no implementation of a patient-specific search service for clinical trials.

6 Conclusions and Future Work

Personalised medicine offers great promises for consultants' decision making, potentially resulting in improved patient treatment. Numerous medical information sources are already available which can be utilized, and they are growing rapidly. However, personalized medicine has not yet widely used in day-to-day clinical practise. With the concept and a prototypical implementation of a Clinical Decision Support System (CDSS) for personalized medicine presented in this paper, we intend to make a contribution towards this direction.

Particular focus has been on the usability of the CDSS allowing consultants to quickly and intuitively gather relevant information with minimal system interaction. A number of artificial intelligence (AI) techniques have been used to tailor information to the patient's medical situation and the consultant's information demand.

It is planned to integrate the CDSS presented into a commercial EHR application suite for melanoma treatment. Towards this end, future work is required. Additional information services as presented in the interaction concept need to be implemented. A comprehensive analysis of the CDSS with consultants in the field needs to take place resulting in potential improvements of the concept and the implementation. Then, a trial phase with real patient data is planned to extend the data base for machine learning. We intend to publish insights from these analyses.

May this work eventually help consultants improve patient care.

7 Acknowledgements

This work was funded by the European Commission, Horizon 2020 Marie Skłodowska-Curie Research and Innovation Staff Exchange, under grant no 644186 as part of the project SAGE-CARE (SemAntically integrating Genomics with Electronic health records for Cancer CARE).

References

Academy of Medical Sciences (2015), Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education (Technical report). Academy of Medical Sciences. May 2015. Retrieved 24/8/2016.

Beez, U., Humm, B.G. and Walsh, P. (2015), "Semantic AutoSuggest for Electronic Health Records", 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 760-765.

Humm, B.G. and Walsh, P. (2015), “Flexible yet Efficient Management of Electronic Health Records”. 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 771-775.

Hung, B. T., Long, N. P., Hung, L. P., Luan, N. T., Anh, N. H., Nghi, T. D., … Hirayama, K. (2015), “Research Trends in Evidence-Based Medicine: A Joinpoint Regression Analysis of More than 50 Years of Publication Data”, PLoS ONE, 10(4).

Kawamoto, K, Houlihan, C.A., Balas, E.A. and Lobach, D.F. (2005), "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success", BMJ? British Medical Journal, 330(7494), 765.

Maggio, L. A., Cate, O. T., Moorhead, L. L., Van Stiphout, F., Kramer, B. M., Ter Braak, E., Posley, K., Irby, D. and O'Brien, B. C. (2014), “Characterizing physicians' information needs at the point of care”, Perspectives on Medical Education (Perspect Med Educ), 33(5), 332-42.

Maggio, L. A., Steinberg, R. M., Moorhead, L., O'brien, B. & Willinsky, J. (2013). "Access of primary and secondary literature by health personnel in an academic health center: implications for open access". Journal of the Medical Library Association (J Med Libr Assoc), 101(3), 205-12.

Marchant, G. E. and Lindor, R. A. (2013), “Personalized medicine and genetic malpractice”, Genetics in Medicine (Genet Med), 15(12), 921-2.

National Cancer Institute (2016). Melanoma Treatment. URL: http://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq#section/_885 (visited on 11/09/2016).

Nyberg, P. (2012), “EBMeDS Clinical Decision Support. EBMeDS White Paper”, URL: http://www.ebmeds.org/www/EBMeDS %20White%20Paper.pdf (visited on 12/09/2016).

Obst, O., Hofmann, C., Knüttel, H. and Zöller, P. (2013), “‘Ask a question, get an answer, continue your work!’ – Survey on the use of UpToDate at the universities of Freiburg, Leipzig, Münster and Regensburg”, GMS Medizin—Bibliothek—Information, 13(3), 26.

Osiński, S., Stefanowski, J. and Weiss, D. (2004), “Lingo: Search results clustering algorithm based on singular value decomposition”. In Intelligent information processing and web mining, 359-368, Springer Berlin Heidelberg.

Perez-Rey, D., Jimenez-Castellanos, A., Garcia-Remesal, M., Crespo, J., & Maojo, V. (2012), “CDAPubMed: a browser extension to retrieve EHR-based biomedical literature”, BMC Medical Informatics and Decision Making, 12, 29.

Peters, L. B., Bahr, N. and Bodenreider, O. (2015), “Evaluating drug-drug interaction information in NDF-RT and DrugBank”, Journal of Biomedical Semantics, 6, 19.

Rahmner, P. B., Eiermann, B., Korkmaz, S., Gustafsson, L. L., Gruvén, M., Maxwell, S., … Vég, A. (2012), “Physicians’ reported needs of drug information at point of care in Sweden”, British Journal of Clinical Pharmacology, 73(1), 115–125.

ScienceDaily (2016). Terms and Conditions of Use. URL: https://www.sciencedaily.com/terms.htm (visited on 11/09/2016).

Soldaini, L., Cohan, A., Yates, A., Goharian, N. and Frieder, O. (2015), “Retrieving medical literature for clinical decision support”, European Conference on Information Retrieval, 538-549, Springer International Publishing.

Chapter 8

Bioinformatics

Full Papers

MetaPlat: A Cloud based Platform for Analysis and Visualisation of Metagenomics Data

Nina Konstantinidou [1], Paul Walsh [1*], Xiangwu Lu [2], Michaël Bekaert [1], Brendan Lawlor [1,2], Brian Kelly [1], Huiru Zheng [3], Fiona Browne [3], Richard Dewhurst [4], Rainer Roehe [4], Haiying Wang [3]

[*] paul.walsh@nsilico.com
[1] NSilico Life Science Ltd., Cork, Ireland
[2] Department of Computer Scince, Cork Institute of Technology, Cork, Ireland
[3] Computer Science Research Institute, School of Computing and Mathematics, University of Ulster at Jordanstown, Co. Antrim, N. Ireland, BT37 0QB
[4] Future Farming Systems, Scotland's Rural College, Edinburgh, United Kingdom

Keywords: Metagenomics, visualisation, cloud computing, sequencing, Simplicity, Bos taurus, rumen

Abstract: In recent years, the number of projects producing very large quantities of sequence data for analysis of the community structure and environmental interactions of microbes, has increased dramatically. Yet, the depth of analysis and visualisation performed is very superficial, reflecting an inefficient use of available information and financial resources. To address these deficiencies, this paper proposes the development of a fast and accurate analytic platform to fully analyse and visualise metagenomic datasets. The platform is illustrated with a use case based on metagenomic data derived from cattle (Bos taurus) intestinal digesta. By focusing on advanced hardware and software platforms, providing training and integrating diverse expertise, this project will pave the way for optimal usage of metagenomic datasets, thus reducing the number of animals required for such studies. For the livestock production industries, our approach will lead to an increase in resource use efficiency coupled with cost reductions, which will have positive effects on both financial viability and climate change consequences.

1 Introduction

Bos taurus (cattle) are ruminants, a group of mammals which also include sheep and goats. The value of domesticated ruminants comes from their ability to convert human inedible forages into high-quality, high-protein food products. This ability results from fermentation by symbiotic microorganisms in the rumen. These microorganisms confer the ability to breakdown complex polysaccharides such as cellulose and hemicellulose, that are not susceptible to mammalian enzymes. Thus, symbiosis with gut bacteria provides the host animal with nutrients, such as bacterial proteins and volatile fatty acids that are important nutrition contributors. The microbiota of the rumen, as well as other segments of the intestines, has a paramount role in cattle performance, production efficiency, immunity and health [1]. Alongside the beneficial effects of the rumen microbes, they also have a series of potential negative consequences including the biohydrogenation of polyunsaturated fatty acids (hence ruminant fat is more saturated than monogastric fat [2]) and production of methane. Ruminant methane emissions are the largest component of livestock greenhouse gas (GHG) emissions, which are estimated to make up around 11% of total anthropogenic GHG emissions [3].

Understanding the structure of the intestinal microbial community is crucial to establish links with physiological effects in the host [4]. Metagenomic sequencing, based on high-throughput analysis, offers unparalleled coverage and depth in determining the structure and dynamics of the intestinal microbial community. A key challenge, in leveraging this genomic sequencing technology, is identifying rumen microbial profiles that may be associated with phenotypic traits. Many researc8/h

groups including Mizrahi [5], Santos et al. [6], Weimer and Kohn [7], Wallace et al. [8] as well as Roehe et al. [9] have investigated the symbiotic microorganisms in the rumen because of their ecologically important traits such as feed conversion efficiency, methane production, and more recently discovery of microbes and enzymes that enable fermentation of biomass for biofuel production. A key challenge now is to identify, analyse in silico and visualise rumen microbial profiles, which are associated, and potentially predictive of these traits.

In order to build and visualise predictive models for the cattle phenotypic traits, based on rumen microbiota, both phenotype and genotype data need to be collected and cleaned. The metagenomic sequence fragments are then assembled and sorted against known annotation databases before further downstream analysis and visualisation of these data in a secure and traceable manner. Thus, to investigate microbiota in the context of genetic effects or dietary changes designed to alter methane production or feed conversion efficiency, the careful development of a sophisticated research software platform is necessary for complete analysis and visualisation of metagenomic data.

The first step, in building such a system, is the provision of a cloud-based research infrastructure that allows researchers to load and link controlled experimental data. For example, cattle identity and genotype, feed quantities and composition can be examined in the context of animal performance measurements, feed efficiency and methane output. To this end, we propose MetaPlat, a cloud-based research infrastructure for metagenomics analysis of rumen microbiota, illustrating the system that allows researchers to rapidly load, manage, analyse and visualise genome and phenome datasets along with related metadata.

In order to reduce the redundancy, the data will be characterised for assurance and will be pre-processed to remove adapters and repeating sequences using open tools. The system supports reproducibility and tracks the phenotypic information associated with each sequence, including its origin, quality, taxonomical position and any associated host genome, with the production of automated reports and visualisations of the metagenomic analyses via Simplicity's platforms. Simplicity is a powerful, yet easy-to-use bioinformatics tool which enables non-specialists to reliably run bioinformatics pipelines in a simple, safe, speedy, secure and reproducible manner.

2 Metagenomics

2.1 Introduction

In the recent years, the number of studies producing huge amounts of metagenomic data has increased dramatically [2013-2014: 182 publications, 2014-2015: 226 publications]. But the depth of analysis and visualisation is superficial, representing an inefficient use of the available information and financial resources necessary to obtain biological samples [10].

Metagenomics (also referred to as community genomics) is a useful approach for the analysis of the microorganisms that cannot be cultured in the laboratory, but which can be directly extracted from their environment. The development of metagenomics stemmed from the ineluctable evidence that as-yet-uncultured microorganisms represent the vast majority of organisms on earth. This evidence was derived from the analyses of 16S ribosomal RNA (rRNA) gene sequences amplified directly from the environment, an approach that avoids bias imposed by culturing and has led to the discovery of vast new lineages of microbial life. Although the view of the microbial world is still limited by analysis of 16S rRNA genes alone, such studies yielded a substantial advance in the

phylogenetic description and profiling of community membership. Metagenomics based on high-throughput sequencing offers unparalleled coverage and depth in determining microbial dynamics as long as the analytic resources and visualisation tools are available. These studies have been applied to a number of organisms including humans, but in this work we focus on Bos taurus.

2.2 In silico Analysis of Metagenomic Data

Careful development of a software platform with comprehensive visualisation tools for more complete in silico analyses of metagenomic data is essential for further advances in understanding of relationships between the gut microbiota, dietary changes, methane production and feed conversion efficiency. Currently, some generic sequence databases, such as DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) and National Center for Biotechnology Information (NCBI), are the only sources of reference sequences for the classification of newly identified species, based on their nucleic acid sequences. However, there has been no systematic attempt to classify uncategorised groups or reduce the level of redundancy produced, despite the general effort/wish to curate quality sequence data, which is a central objective in MetaPlat.

A related problem is the lack of new algorithms, approaches and visualisation tools (for example, comprehensive phylogenetic trees) to deal with the huge number of sequences generated via sequencing of bacterial metagenomes. Usually, assignment of a sequence to an organism is performed by sequence similarity, which is not optimal if the sequences diverge greatly or not enough. For example, if two or more sequences are identical, the sequence/species assignment is often given by the first sequence name encountered by the algorithm, which can be ambiguous and even missing, producing very biased and non-reproducible results and interpretations. In addition to lack of reproducibility and inadequate visualisation of results, significantly low speed and scalability of the analysis are now concerns for the field, not to mention the lack of skilled bioinformaticians. Thus, robust computational approaches are required to distil big data derived from metagenomics analysis down to clear visualisation of the gut microbiota community structure, while supporting and enhancing reproducibility.

2.4 Related Work

Research by Brulc et al. [11] and Hess et al. [12] were seminal in the development of rumen metagenomics, but their studies involved a few animals. Sequencing costs are reducing rapidly and the ability to sequence either multiple 16S libraries or whole rumen metagenomes has advanced to the point where data analysis capability and acquisition of phenotypes are now major constraints, rather than the cost of sequencing.

Almost all tools currently available assign a sequence to an organism by sequence similarity after sequence alignment (Table 1). This is a computationally intensive and time consuming task, and these tools are not optimised for parallel or cloud execution, nor for visualisation. Some tools also rely on scriptic languages and algorithms which are not optimised. The lack of new algorithms, optimised approaches and visualisation tools, to deal with the huge number of generated sequences, increases bias and contributes to the production of non-reproducible results [10]. Representation of results for dissemination is also an important challenge. Today, only a single tool used by almost all studies, Krona [13], provides a valuable interactive generic hierarchical data browser, but does not convey all degrees of information and visualisation.

Software	Reference	Access	Comments
RDP suit	[14]	http://rdp.cme.msu.edu/	Single sequence request, BLAST alignment based
PhyloSift	[15]	https://phylosift.wordpress.com/	NGS data, Phylgeny-driven classification
MGTAXA		http://mgtaxa.jcvi.org/	NGS data, Phylgeny-driven classification
PhymmBL	[16]	http://www.cbcb.umd.edu/software/phymmbl/	Discontinued, BLAST alignment based
MG-RAST	[17]	http://metagenomics.anl.gov/	Online only, NGS data, BLAST alignment based
Metaxa2	[18]	http://microbiology.se/software/metaxa2	Perl scripts, NGS data, BLAST alignment based
MetaPhlAn	[19]	http://huttenhower.sph.harvard.edu/metaphlan/	NGS data, BLAST alignment based
Parallel-META	[20]	http://www.computationalbioenergy.org/parallel-meta.html	NGS data, BLAST alignment based
MetaPhyl	[21]	http://www.sc.ucr.edu/~tanaseio/metaphyl.html	Training data set required, k-mer based
CLcommunity		http://www.chunlab.com/software_clcommunity_about/	Licenced, NGS data, BLAST alignment based
QIIME	[22]	http://qiime.org/	Performs microbiome analyses
Cytoscape	[23]	http://www.cytoscape.org/	Produces molecular interaction networks
STRING	[24]	http://string-db.org/	Produces functional protein association networks

Table 1. Online tools for the analysis of metagenomic data

3 Materials and Methods

To address the challenges outlined in the previous section, we will present a methodology that will interlink important research elements in the metagenomics domain: Biological samples (3.1), Big data management, characterisation and curation (3.2), Reference database distribution (3.3), In silico search algorithms (3.4), Phylogeny-aware classification (3.5), Machine learning models (3.6), New visualisations (3.7), Integrated software pipeline development using high-performance computing (HPC) platforms (3.8) and Reproducibility (3.9). The platform will be developed based on a Simplicity pipeline [25].

3.1 Biological Samples

In the first stage, biological samples will be collected and sequenced in parallel to reference databases preparation and algorithm development. Rumen fluid samples will be collected using a naso-gastric sample tube system at Scotland's Rural College (SRUC). DNA isolation will be carried

out with a bead-beating procedure according to Yu and Morrison [26] and 16S amplicons will be prepared using Caporaso primers [27]. Sequencing of both 16S amplicons (in parallel) and total genomic DNA will be accomplished using Illumina sequencers (MiSeq and HiSeq).

3.2 Big Data Management, Characterisation and Curation

Characterisation of sequenced data will facilitate better understanding of the bacterial communities collected from B. taurus rumen. Accurate classification of new sequences entails the formal and reliable analysis of the complete set of reference sequences. In order to reduce the redundancy, a system will be developed to keep track of each sequence origin, taxonomical position, associated bacterial genome (gene set), quality and redundancy status. This will involve the development of a system allowing automatised update of the databases and the public accessibility of the resulting up-to-date dataset. Another system will be designed to provide homogenised and high quality data, by analysing and visualising the results in a fully integrated manner (Figure 1).

3.3 Reference Database Distribution

To create a reference database, data from the publicly available reference databases such as GreenGenes [http://greengenes.lbl.gov/cgi-bin/nph-index.cgi] will be used in the metagenomics pipeline (Figure 1). This reference database will be updated automatically via a relevant system allowing automated update of the data. This system will be designed to provide homogenised and high quality data, by visualising and analysing the sequenced data in a fully integrated manner.

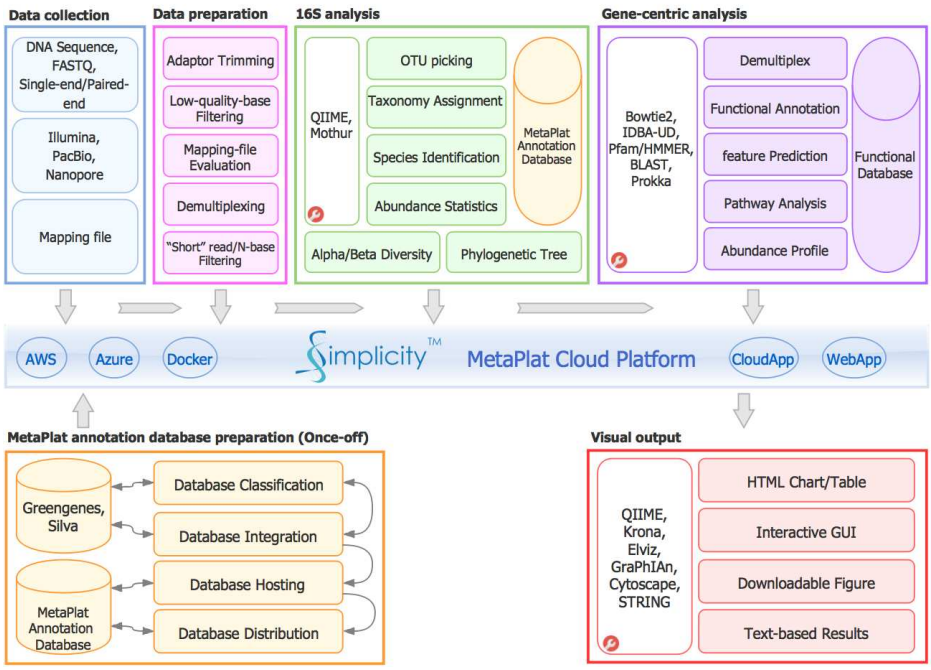


Figure 1. Illustration of a high level Big data reference model based on cloud technology. Big data and Simplicity's services and analysis will be stored in a cloud. This will reduce the required space for the analysis of large datasets by the users. Diverse visualisations will be integrated into the metagenomics pipeline of Simplicity to enhance classification and interpretation of biological data.

3.4 In silico Search Algorithms

Metagenomic samples contain reads from a huge number of organisms. For example, in a single gram of soil, there can be up to 18,000 different types of microbes, each with different genomic background. Usually, classical binning approaches are used for grouping reads or contigs and assigning them to operational taxonomic units (OTUs). Based on Mande et al. [28] review, we will develop a new algorithmic approach, facilitating accuracy, speed, reproducibility and visualisation of the analysis that will subsequently be implemented in a High-Performance Computing (HPC) infrastructure. We will build on the known and trusted classification algorithms to implement a comprehensive pipeline of algorithms for the integrated analysis and visualisation of microbial populations in Simplicity's metagenomics pipeline.

3.5 Phylogeny-aware Classification

Phylogeny awareness is an important element for the annotation and classification of the sequences. The main objective of this task is to develop a new phylogeny-aware pipeline, within Simplicity's metagenomics platform, in order to classify sequencing data. This algorithm will facilitate annotation of each sequence based on phylogenetic information available online.

3.6 Machine Learning Models

Machine learning models are useful for the prediction of previously unclassified sequences. Currently, greedy and hierachical clustering are the most frequently used machine learning algorithms for metagenomic data analysis [29]. In this study, a new machine learning based classification model will be proposed and implemented into the Simplicity's metagenomics pipeline. To train, test and evaluate this model, the curated data with phylogeny-aware classification will be utilised in the pipeline.

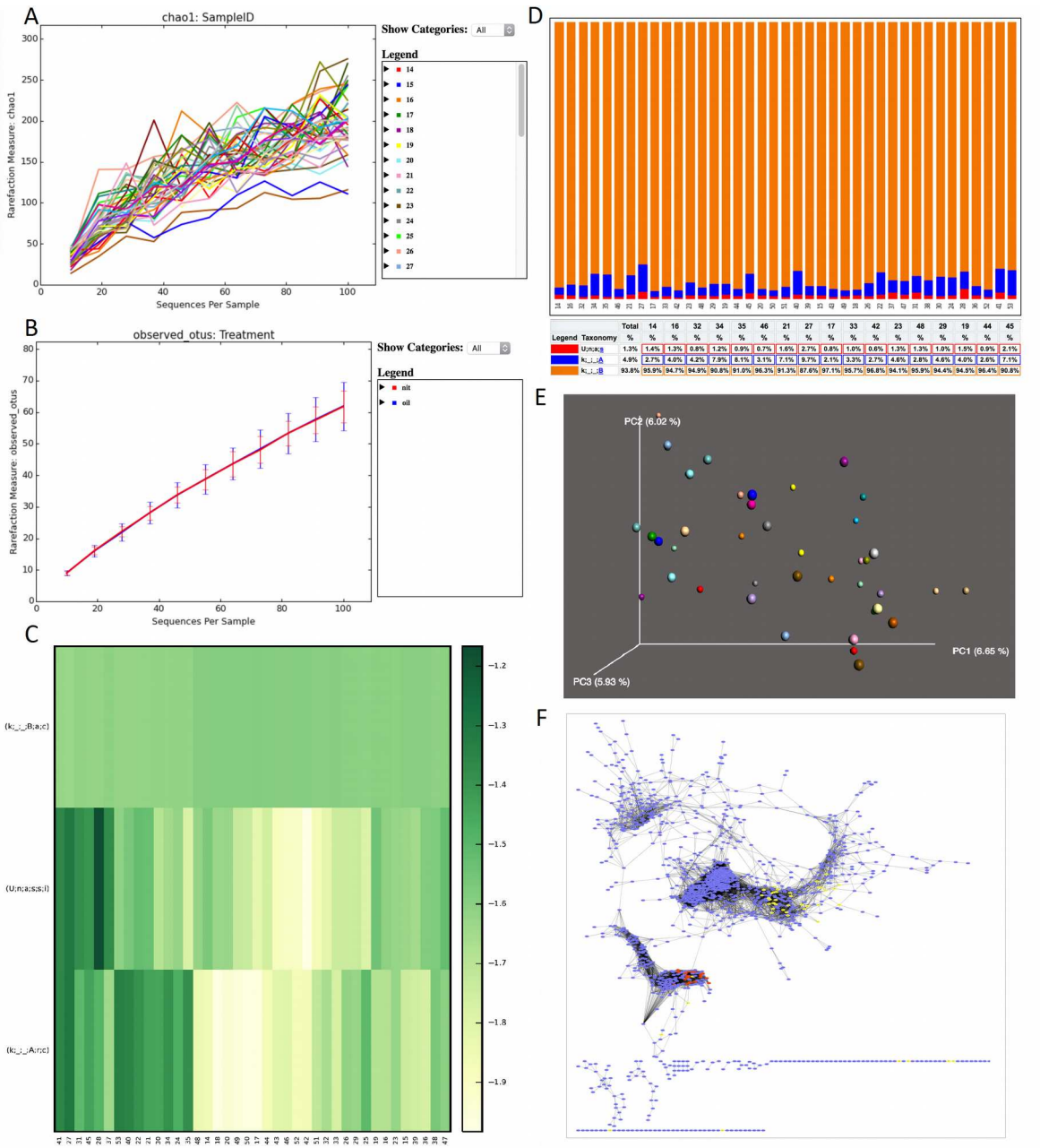


Figure 2. Preliminary QIIME analyses of a dataset from Teagasc, Ireland (A-E) and Cytoscape analysis of a dataset from Scotland's Rural College (F). (A) Alpha (α) diversity of the species richness within the community is calculated using Chao1 metric and represented as a line chart with QIIME (B) Alpha (α) diversity is measured using Shannon index. The Shannon index denotes the entropy of the observed OTU abundances and accounts for both richness and evenness. (C) Heatmap reflects the matrix of OTU abundance in each sample with meaningful taxonomic identifiers for each OTU. (D) Bar chart illustrates the taxa assignments for each sample within communities. (E) Principal Coordinates Analysis (PCoA) plot demonstrates the hierarchical clustering by transforming the distance matrix into three dimensional (3D) coordinates using beta (β) diversity metrics. The similarity of samples is shown by smaller distance between the dots. (F) Co-abundance network analysis using Cytoscape. With the threshold of 0.95, the yellow nodes represent microbial genes associated with feed conversion efficiency. Red nodes illustrate methane emission-related microbial genes which were identified in Scotland's Rural College (SRUC) data [9].

3.7 New Visualisations

Visualisation of the metagenomic data is crucial for understanding the microbial diversity in the gut. To assign biological meaning to the in silico metagenomic analysis, Krona [13] and additional visualisation tools will be integrated into the Simplicity (Figure 1). For example, QIIME (Quantitative Insights Into Microbial Ecology) [22] and Elviz [30] are designed to analysed 16S rRNA gene sequences of microbial communities and for interactive visualisation of metagenome assemblies respectively. These tools will be incorporated into the metagenomics pipeline of Simplicity to visualise microbial data through PCoA plots, bar charts (QIIME) (Figure 2) and bubble plots (Elviz). To reveal the evolutionary lineages of the microbes in the gut, phylogenetic analysis of metagenomic data needs to be conducted using additional visualisation aspects. Due to the latter, GraPhlAn [31], a compact graphical representation tool for metadata, will be integrated into the Simplicity's pipelines.

Visualisation of protein-protein interaction (PPI) networks can reveal biological functions of the proteins and the molecular processes they are involved. To carry out network analysis, Cytoscape [23] and STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) [24] can additionally be employed for a graphical output of PPI networks and visualisation of functional protein associations respectively. For paradigm, using Cytoscape PPI network, the preliminary analysis of Scotland's Rural College (SRUC) data highlighted two distinct sets of interesting genes (Figure 2). Figure 2F shows that the microbial genes linked with feed conversion efficiency (yellow nodes) and those related with methane emission (red nodes) were clustered into two different clusters, indicating functional differences. Even though diverse visualisation tools will be added to the metagenomics pipeline, Simplicity will maintain its easy-to-use features. Importantly, the reproducibility of the analyses and visualisations will be secured via Docker technology (see below section 3.9 Reproducibility).

3.8 Integrated Software Pipeline Development using High-Performance Computing Platforms

For the development of Simplicity's metagenomics pipeline, a formal software development approach with a standard software engineering model will be applied. Software development will be initiated with careful analysis of the requirements, by engaging with multidisciplinary teams of the researchers including microbiologists, molecular biologists, clinicians, bioinformaticians and developers. The core effort will be to develop thread safe extensible bioinformatics algorithms for high-performance computing (HPC) architectures using ISO/IEC 12207, an international standard for software life-cycle processes, which defines best practices for developing and maintaining software. A process will be put in place to identify the most scalable parallel computing architectures and the appropriate functional partitioning strategies to achieve the most optimal solution by maximising data locality and minimising inter-process communication.

3.9 Reproducibility

Reproducibility is an important aspect of the research but difficult to achieve while using in silico techniques. We propose to address this using Docker technology [32], which will facilitate reproducibility by encapsulating a complete environment with scripts, libraries, databases, system tools and tool dependencies. This approach will provide reproducibility and transparency in experimental methodology, observation and collection of data through the development of web-based tools by facilitating collaboration, storage and integration of new data and visualisations. Data will be shared and archived using digital information objects interlinked with internal and

external resources, in a structured and machine readable serialization mechanism, as measured against ISO 14721:2012 (OAIS), by adding the required provenance and descriptive. For instance, domain information (metadata) for ensuring results may be shared and reproduced over long term data life cycles. In order to support digital preservation by paving the way towards reproducibility, previous results on reproducibility of bioinformatics analyses by means of archiving research data in archival packages, are compliant with ISO 14721:2012 (Open Archival Information System – OAIS).

4 Discussion, Summary and Future Work

In this paper we have highlighted key challenges to be addressed in the metagenomics domain and highlighted our proposed solution, MetaPlat. MetaPlat will integrate complementary innovative activities, including new visualisation options, to address the limitations in previous work, and thus underline the overall metagenomics pipeline development plan within Simplicity. Currently, integration and evaluation of metagenomics analysis and visualisation tools has been initiated in collaboration with microbiologists and bioinformaticians. Thus, the following innovations, outlining metagenomics data analysis pipeline, are noteworthy:

Development of non-redundant sequence databases, that will encapsulate all currently available good quality metadata (species and sub-species/strain, number of sequence), as well as all previously unclassified sequences, using a machine learning approach. This will include a procedure to update the data with newly released information and a public dissemination of the resulting database. This database will be incorporated within Simplicity's metagenomics pipeline, a new platform focusing on reproducible metagenomics analysis in silico.

Development of accurate and reproducible classification algorithms, optimised for parallel and cloud execution with open source code. These algorithms will be organised into services and pipelines, collectively leading to the development of Simplicity's metagenomics pipeline, a revolutionary bioinformatics tool for in silico analysis of complex metagenomic data.

Real-time or time-efficient comparisons and analyses of large datasets with phylogeny-aware classification, but also quantitative and functional analyses (when possible) will also be carried out using Simplicity's metagenomics comparison pipeline that is able to compare hundreds of bacteria.

Production of statistical and visual representations in a standard report, after data analysis with Simplicity's metagenomics pipeline, can maximise microbiome information and convey biological meaning to current visualisations.

Simplicity's metagenomics pipeline will be able to perform standard steps of high quality analyses and comparisons of public databases as well as user-own dataset with a few clicks automatically, in an integrated environment (e.g. Galaxy [33] web-based platform).

It is envisaged, that the Simplicity metagenomics pipeline output, addressing key objectives of the MetaPlat project, will revolutionize gut microbiome analysis. This is important for advancing our understanding of gut microbiota to benefit humans, animals and the environment. These benefits include improved animal health and productivity, reduced methane emissions and improved targeting of dietary supplements. Future work will focus on the development of machine learning and visualisation models for metagenomics analysis in order to integrate them into the next versions of Simplicity's metagenomics pipeline.

Acknowledgements

This research is supported by Horizon 2020 MSCA-RISE-2015 Marie Skłodowska-Curie Research and Innovation Staff Exchange Project (ID: 690998), www.metaplat.eu.

References

1. Beever, D.E.: The Impact of Controlled Nutrition During the Dry Period on Dairy Cow Health, Fertility and Performance. Anim Reprod Sci. 96, 212-226 (2006).

2. Dewhurst, R.J.: Targets for Milk Fat Research: Nutrient, Nuisance or Nutraceutical? Journal of Agricultural Science 143, 359-367 (2005).

3. Gerber, P.J., Hristov, A.N., Henderson, B., Makkar, H., Oh, J., Lee, C., Meinen, R., Montes, F., Ott, T., Firkins, J., Rotz, A., Dell, C., Adesogan, A.T., Yang, W.Z., Tricarico, J.M., Kebreab, E., Waghorn, G., Dijkstra, J., Oosting, S.: Technical Options for the Mitigation of Direct Methane and Nitrous Oxide Emissions from Livestock: a Review. Animal 7, 220-234 (2013).

4. Oskoueian, E., Abdullah, N., Oskoueian, A.: Effects of Flavonoids on Rumen Fermentation Activity, Methane Production, and Microbial Population. Biomed Res Int. 2013, 349129 (2013).

5. Mizrahi, I.: The Role of the Rumen Microbiota in Determining the Feed Efficiency of Dairy Cows. Springer 14, 203-210 (2011).

6. Santos, V.C. , Ezequiel, J.M.B. , Homem Junior, A.C. , Pinheiro, R.S.B.: Quantification of Ruminal Microbiota and Production of Methane and Aarbonic Dioxide from Diets with Inclusion of Glycerin. Arq Bras Med Vet Zootec. 67, 1678-4162 (2015).

7. Weimer, P.J., Kohn, R.A.: Impacts of Ruminal Microorganisms on the Production of Huels: How Can We Intercede from the Outside? Appl Microbiol Biotechnol. 100, 3389-98 (2016).

8. Wallace, R.J., Rooke, J.A., McKain, N., Duthie, C-A., Hyslop, J.J., Ross, D.W., Waterhouse, A., Watson, M., Roehe, R.: The Rumen Microbial Metagenome Associated with High Methane Production in Cattle. BMC Genomics 16, 839 (2015).

9. Roehe, R., Dewhurst, R., Duthie, C-A., Rooke, J.A., McKain, N., Ross, D.W., Hyslop, J.J., Waterhouse, A., Watson, M., Wallace, R.J.: Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. PLoS Genetics 12, e1005846 (2016).

10. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. PLoS Med. 2, e124 (2005).

11. Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., Coutinho, P.M., Henrissat, B., Nelson, K.E., White, B.A.: Gene-centric Metagenomics of the Fiber-adherent Bovine Rumen Microbiome Reveals Forage Specific Glycoside Hydrolases. Proc Natl Acad Sci USA 106, 1948-1953 (2009).

12. Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M.: Metagenomic Discovery of Biomassdegrading Genes and Genomes from Cow Rumen. Science 331, 463-467 (201).

13. Ondov, B.D., Bergman, N.H., Phillippy, A.M.: Interactive Metagenomic Visualization in a Web Browser. BMC Bioinformatics 12, 385 (2011).

14. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive Bayesian Classifier for Rapid Assignment of RRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73, 5261–5267 (2007).

15. Darling, A.E., Jospin, G., Lowe, E., Matsen IV, F.A., Bik, H.M., Eisen, J.A.: PhyloSift: Phylogenetic Analysis of Genomes and Metagenomes. PeerJ. 2, e243 (2014).

16. Brady, A., Salzberg, S.: PhymmBL Expanded: Confidence Scores, Custom Databases, Parallelization and More. Nat Methods 8, 367 (2011).

17. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A.: The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. BMC Bioinformatics 9, 386 (2008).

18. Bengtsson-Palme, J., Hartmann, M., Eriksson, K.M., Pal, C., Thorell, K., Larsson, D.G.J., Nilsson, R.H.: METAXA2: Improved Identification and Taxonomic Classification of Small and Large Subunit rRNA in Metagenomic Data. Mol Ecol Resour. 15, 1403-1414 (2015).

19. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C.: Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes. Nat Methods 9, 811–814 (2012).

20. Su, X., Pan, W., Song, B., Xu, J., Ning, K.: Parallel-META 2.0: enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization. PLoS One 9, e89323 (2014).

21. Tanaseichuk, O., Borneman, J., Jiang, T.: Phylogeny-based Classification of Microbial Communities. Bioinformatics 30, 449-456 (2014).

22. Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G., Knight, R.: Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. Curr Protoc in Bioinformatics, Chapter 10, Unit 10.7 (2011).

23. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498-2504 (2003).

24. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: STRING 8--a Global View on Proteins and Their Functional Interactions in 630 Organisms. Nucleic Acids Res. 37, D412-416 (2009).

25. NSilico Life Science Ltd.: SimplicityTM Version 1.5, Cork, Ireland (2015).

26. Yu, Z., Morrison, M.: Improved Extraction of PCR-quality Community DNA from Digesta and Fecal Samples. Biotechniques 36, 808-812 (2004).

27. Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J., Smith, G., Knight, R.: Ultrahigh-throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms. ISME J. 6, 1621-1624 (2012).

28. Mande, S.S., Mohammed, M.H., Ghosh, T.S.: Classification of Metagenomic Sequences: Methods and Challenges. Brief Bioinform. 13, 669-681 (2012).

29. Soueidan, H., Nikolski, M.: Machine Learning for Metagenomics: Methods and Tools. Review Metagenomics 1, 1–19 (2016).

30. Cantor, M., Nordberg, H., Smirnova, T., Hess, M., Tringe, S., Dubchak, I.: Elviz - Exploration of Metagenome Assemblies with an Interactive Visualization Tool. BMC Bioinformatics 16, 130 (2015).

31. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., Segata, N.: Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn. PeerJ. 3, e1029 (2015).

32. Boettiger, C.: An Introduction to Docker for Reproducible Research. ACM SIGOPS Operating Systems Review 49, 71-79 (2015).

33. Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team: Galaxy: a Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences. Genome Biol. 11, R86 (2010).

Conference Awards

Best Paper / Presentation Award

presented to :

Lawlor, Walsh

Metamorphosis: Changing the shape of genomic data using Kafka

ISSN : 2220 - 4164