INTERNATIONAL SERIES ON INFORMATION SYSTEMS AND MANAGEMENT IN CREATIVE eMEDIA

NUMBER 2016/2

the series

IS&M in Creative eMedia

Artur Lugmayr, Richard Seale, Andrew Woods,
Eunice Sari, and Adi Tedjasaputra (eds.)

Proceedings of the 9th Workshop on Semantic Ambient Media Experiences (SAME 2016b)

VISualization, emerging Media, and user-eXperience (Vis-MX)

The International Series on Information Systems and Management in Creative eMedia is advancing the knowledge of the use of informaiton systems and management in the wider field of creative eMedia industries. The series covers a wide range of media , such as television, publishing, digital games, radio, ubiquitous/ambient media, advertising, social media, motion pictures, online video, eHealth, eLearning, and other eMedia industries. The series is indexed in Scopus and available under open acces under: www.ambientmediaassociation.org/Journal

Printed by the
International Ambient Media Association (iAMEA)
Ihanakatu 7-9/A1
FIN-33100 Tampere
FINLAND
available through: www.ambientmediaassociation.org
and http://www.ambientmediaassociation.org/Journal

ISBN 978-952-7023-15-0 (PDF)

ISSN 2341-5584 (Print)
ISSN 2341-5576 (PDF)
ISSN 2341-6165 (CD-ROM)

International Series on Information Systems and Management in Creative eMedia

Nr. 2016/2

INTERNATIONAL SERIES ON INFORMATION SYSTEMS AND MANAGEMENT IN CREATIVE eMEDIA

Artur Lugmayr, Richard Seale, Andrew Woods,
Eunice Sari, and Adi Tedjasaputra (eds.)

Proceedings of the 9th Workshop on Semantic Ambient Media Experiences (SAME 2016b)

VISualisation, emerging Media, and user-eXperience (Vis-MX)

Number 2016/2
Perth, Western Australia, AUSTRALIA, 10th – 11th November 2016

Printed by the International Association for Ambient Media (iAMEA) Ry

Perth, WA, Australia 2016

Published and printed by the International Ambient Media Association (iAMEA) Ry, Ihanakatu 7-9/A1, FIN-33100 Tampere, Finland. The online version is available through www.ambientmediaassociation.org/Journal.

Cover image: © Artur Lugmayr

Printed and published by the International Ambient Media Association (iAMEA)
Ihanakatu 7-9/A1
FIN-33100 Tampere
FINLAND
ISBN 978-952-7023-15-0 (PDF)

ISSN 2341-5584 (Paperback)
ISSN 2341-5576 (PDF) Online: www.ambientmediaassociation.org/Journal
ISSN 2341-6165 (CD-ROM)

Preface

Digital and interactive technologies are becoming increasingly embedded in everyday lives of people around the world. Application of technologies such as real-time, context-aware, and interactive technologies; augmented and immersive realities; social media; and location-based services has been particularly evident in urban environments where technological and sociocultural infrastructures enable easier deployment and adoption as compared to non-urban areas. There has been growing consumer demand for new forms of experiences and services enabled through these emerging technologies. We call this ambient media, as the media is embedded in the natural human living environment. This workshop focuses on ambient media services, applications, and technologies that promote people's engagement in creating and re-creating liveliness in urban environments, particularly through arts, culture, and gastronomic experiences.

The 9th Semantic Ambient Media Experience (SAME) proceedings where based on the academic contributions to a two day workshop that was held at Curtin University, Perth, WA, Australia. The symposium was held to discuss visualisation, emerging media, and user-experience from various angles. The papers of this workshop are freely available through http://www.ambientmediaassociation.org/Journal under open access as provided by the International Ambient Media Association (iAMEA) Ry. iAMEA is hosting the international open access journal entitled "International Journal on Information Systems and Management in Creative eMedia", and the series entitled "International Series on Information Systems and Management in Creative eMedia". For any further information, please visit the website of the Association: http://www.ambientmediaassociation.org

The International Ambient Media Association (AMEA) Ry organizes the Semantic Ambient Media (SAME) workshop series, which took place in 2008 in conjunction with ACM Multimedia 2008 in Vancouver, Canada; in 2009 in conjunction with AmI 2009 in Salzburg, Austria; in 2010 in conjunction with AmI 2010 in Malaga, Spain; in 2011 in conjunction with Communities and Technologies 2011 in Brisbane, Australia; in 2012 in conjunction with Pervasive 2012 in Newcastle, UK; and in 2013 in conjunction with C&T 2013 in Munich, Germany; and in 2014 in conjunction with NordCHI 2014 in Helsinki, Finland. In 2015 we had no workshop, but have been collaborating with the SEACHI Workshop Smart Cities for Better Living with HCI and UX, which has been organized by UX Indonesia and held in conjunction with Computers and Human-Computer Interaction (CHI) in San Jose, CA, USA in 2016. In 2016 we organised a second workshop, which has been held at Curtin University, Perth, WA, Australia. The workshop proceedings are indexed in Scopus, and peer reviewed. For this edition of the workshop, 4 pages from 10 submitted papers made it into the final proceedings.

The workshop organizers present you a fascinating crossover of latest cutting edge views on the topic of ambient media, and hope you will be enjoying the reading. We also would like to thank all the contributors, as only with their enthusiasm the workshop can become a success. At last, we would like to thank the lovely organizing team at Curtin University, as well as we would like to acknowledge the TASS grant that supported the second day of the VisMX symposium.

The Editors Perth, WA, AUSTRALIA

Table of Contents

List of Contributors

Aleksandra Filipovska (Univ. of Western Australia, AUSTRALIA)
Beata Dawson (Curtin University, AUSTRALIA)
Andreas Holzinger (TU Graz, AUSTRIA)
Pauline Joseph (Curtin University, AUSTRALIA)
Yuliya V Karpievitch (Univ. of Western Australia, AUSTRALIA)
Henrik Bjerre Karsgaard (Networked Business Initiative, DENMARK)
Irina Kuznetsova (TU Graz, AUSTRIA)
Jan Futtrup Kjaer (Networked Business Initiative, DENMARK)
Lester Lasrado (Copenhagen Business School, DENMARK)
Shastri Laksham (Curtin University, AUSTRALIA)
Artur Lugmayr (Curtin University, AUSTRALIA)
Amit Rudra (Curtin University, AUSTRALIA)
Ravi Vatrapu (Copenhagen Business School, DENMARK)

Call for Papers

CALL FOR CONTRIBUTIONS:

Position Papers, Joint Project Ideas, Industry Talks, Demos, and Exhibits

VISualization, emerging Media, and user-eXperience Vis-MX
[:vis-em-ex:]

A Think-Tank and Forum for Collaboration,
Cross-Disciplinary Thinkers, and Visionaries

10th-11th November 2016
Curtin University – @ Curtin's HIVE
Perth, Australia

- proceedings published in Scopus indexed proceedings
- journal special issue in an ERA ranked publication venue
- the academic part of the symposium is the 9th Semantic Ambient Media Experience Workshop (SAME) which is held in conjunction with this event

---

Welcome to the 1st Western Australian Symposium on Visualization, Emerging Media, and User-Experience! The aim of the symposium is to lay ground for a yearly symposium and develop towards a leading forum for inter-disciplinary thinkers of industrial professionals, academics, and practitioners to discuss and present the latest emerging trends in media, visualization, human-computer interaction, and the use of new emerging technologies in industry-university cooperation in teaching and research.

The purpose of the symposium is to provide an inter-disciplinary event and gather a committed community to continue cooperating together beyond the event on other larger scale activities such as collaborative funding proposals, scientific events, industry-university collaborations, or simply other networking activities.

THEMES

The main themes of the symposium address the following major tracks. The aim is to attract visitors with backgrounds from industry, design, communication studies, computer science, or business:

- Virtual Reality, Augmented Reality, and Mixed Reality
- Information, Data, Knowledge, and Cultural Visualization
- User-Experience, and Human-Computer-Interaction
- Media Studies, Screen Arts, and Socio-Technological Environments
- Digital Media and New Technologies in Education
- Revenue Models, Digital Innovations, Policies, and Business Processes

REGISTRATION

The event if free of charge, however, active contributors have priority. Please register on the event website.

KEY-DATES
---

* 31st October
  submission of position papers, joint project ideas, industry talks, posters, demos and award project proposals

* 6th November
  acceptance notification

* 10th-11th November
  visemex symposium

* 27th November

full papers due for academic presenters
a selected set of high-quality full-papers will be published within a journal special issue in spring 2017

SUBMISSIONS & CONTRIBUTIONS

Please follow the submission guidelines on the website. Position papers are based on an abstract (minimum 250-300 words) but can extend towards max. 8 pages. Please use the same template when submitting a position paper, joint project idea, industrial talk, or demo. The submission system for submitting your contribution can be found on http://www.ambientmediaassociation.org/Submissions/2016VisMX/.

Several submissions will be published as part of a Scopus ranked conference proceeding after the conference. Selected high quality contributions will be published within an ERA listed journal special issue - International Journal of Web Based Communities (IJWBC).

SUPPORTED BY

Curtin University (in particular through a TASS grant)
Curtin's HIVE (Hub for Immersive Visualisation and eResearch)
Association for Information System SIG eMedia (AIS SIG eMedia)
Human Factors and Ergonomics Society of Australia (HFESA)
International Association for Ambient Media (iAMEA)

ORGANIZERS

A/Prof. Dr. Artur Lugmayr, Curtin University, AUSTRALIA
Dr. Andrew Woods, Curtin University, AUSTRALIA
Richard Seale, Curtin University, AUSTRALIA
Dr. Eunice Sari, HFESA CHISIG (WA) and UX Indonesia, AUSTRALIA
Adi Tedjasaputra, HFESA CHISIG (WA) and UX Indonesia, AUSTRALIA

Talk Abstracts

Unique Issues of a Contemporary Medium: Consumption Patterns of the Video Game Community

David Jian-Jia Cumming, Curtin University, Australia
david.jianjia.cumming@gmail.com

Video games are a relatively new entertainment medium in comparison to traditional media like television or print publishing. As such, their development has taken place in a more contemporary setting than other mediums, which is subjecting them to a different set of factors and complications. As a result, the news media and community that has formed around the video game industry has evolved in a very unique way. This ultimately affects how video game related news emerges, circulates and is consumed. This paper explores how the video game community consumes news and the factors that influence these consumption patterns and news delegation. This study focuses on a very unique aspect of video games by identifying how the video game community consumes news. There is a current lack in academic research on aspects of news consumption patterns of the video game community. There is also minimal academic research on video game news and journalism, despite the growth of the medium and well-publicised issues of the video game news media. To guide the investigation, the following research question has been formulated: What are the news consumption patterns and methods of the video game community? To answer this question, the following two hypotheses have be proposed: 1) conventional media outlets will, on average, be less popular in comparison to grassroot outlets due to a perceived lack of creditability in the former and a better sense of creditability in the latter; and 2) those more engaged in video games as an interest will consume gaming related news from a wider variety of outlets than those less engaged, because their desire to consume gaming related information is greater. To test the hypotheses and address the research question, seventy members of the video game community took a survey to gather information about their video game news consumption patterns and methods. The results were then analysed and then contextualised by interviews with three industrial representatives to help understand the results. The key purpose of this research to determine what factors influence the news consumption patterns and methods of the video game community. It was found that the video game community base their news consumption patterns on the need to fulfil niche interests. As a result, online news outlets are greatly favoured, as they are easily able to be set up and operated by enthusiasts. Thus, they are more numerous and able to cover more niche interests than traditional, offline outlets.

Resolution vs Stereoscopy in Visualisation of 3D Data

Joshua Hollick, Curtin HIVE, Australia
joshua.hollick@curtin.edu.au

Visualisation is often used to make patterns and relationships in data visible, often in a way that does not require domain specific knowledge. When visualising data there are not only several methods of presenting the data but also several presentation devices available. In this presentation we explore some of the trade-offs between high resolution 2D displays and lower resolution but stereoscopic displays.

Advocating Users in the Development of an Online Learning Platform for Adult Learners

Eunice Sari, UX Indonesia, Australia
eunice@uxindo.com

Novistiar Rustandi, Haruka Evolusi Digital Utama, Indonesia
novistiar@harukaedu.com

Martin Tjahjono, Haruka Evolusi Digital Utama, Indonesia
martin@harukaedu.com

Adult learners are a niche but significant type of users in the field of continuing education. Continuously balancing work, family and study, the adult learners have created a new direction how online continuing education should be developed. The development of an online education generally covers at least four main aspects, which are learner, learning content, learning technology and learning support to ensure the process of learning working properly. However, in reality, technology has often become the primary focus when designing an online education. This study has been run for a year with an Indonesian education technology startup supported by Google Launchpad Accelerator program. In this project, we have applied user-centred design methods to engage users, which is in this case adult learners and stakeholders in the process of co-designing of a new online learning platform and services in Indonesian context. Based on the quantitative and qualitative data, this paper describes the role of users and stakeholders in impacting the development of online learning content, strategy and platform as well as potentials and challenges in implementing user experience approach in a high-paced startup environment.

Title: Exploring the user experience of mobile e-commerce: a competitive study of WeChat in-app Shop and Mobile Taobao app

Danjing Joy Zhang, Curtin University, Australia
joy.zhang@curtin.edu.au

ICT initiatives are attending to provide micro e-business opportunities. However, only a few studies have been conducted to identify e-business implications of the mobile internet. This paper will investigate the user experience of two case studies: WeChat e-shop and Mobile Taobao representing fast pace developing and extending ecosystem beyond the geography boundaries. Mobile e-business has been attached great importance by enterprises, small business and individuals. Service providers need to understand the communication functionality affecting user behavior, and take effective measures to facilitate user experience. This paper will examine if the social-commerce and mobile payment take effective measures to facilitate user experience of mobile e-business apps from a customer perspective.

Although the potential uses and impacts of mobile e-commerce are emerging in many countries, China has led the hybrid of mobile internet and e-business development. According to a report issued by China Internet Network Information Center (CNNIC) in 2016, the number of mobile internet users in China has reached 527 million, and world largest online e-commerce market with 590 million (CNNIC, 2016). Faced with the populous market, both the technology company Tencent and e-business giant Alibaba have established mobile ecosystem with inclusion of in-app purchasing. Guo and Krogstie (2015) noted that WeChat was more than the mobile instant messaging by added the WeChat Ecosystem, a variety of mobile services, including mobile news, mobile games and mobile payment.

WeChat has incorporated e-commerce in order to generating revenue in 2014. In this aspect; the technology innovation allows businesses and individuals to directly market and sell to a highly engaged audience. As one of the actors of WeChat ecosystem, WeChat's allows anyone to set up a store inside WeChat. WeChat's innovative financial mechanisms and social networking function are integrated to facilitate the marketing and sales of mobile e-business. In comparison, Mobile Taobao is a transactional e-business ecosystem from online to mobile taking advantages of Alibaba's mature ecosystem of logistics, manufacture, suppliers and

exporters. This paper aims to identify the cultural and technological factors regarding to user experience of viewing product reviews, sharing and links to favored products and seek out third-party opinions. It also analyzed the differences on mobile e-business based on the following criteria, i.e., mobile payment, geography, and group function.

Considering the high adoption rate of WeChat and Taobao mobile app, survey and interviews will be conducted among Chinese immigrants in Australia who shop through these two apps. By interview and survey the user experience, the unique characteristics of these two digital platforms are the potential of the mobile e- business which can support the evolution of mobile ecosystem globally.

Cultural Heritage Visualization: Using Interactive Multimedia in Museum Environment

Beata Dawson

Ph.D. Candidate

Curtin University

Kent Street, Bentley, Perth,
Western Australia 6102

+61405064163

beata.dawson@postgrad.curtin.edu.au;
www.beata-dawson.site

Dr. Pauline Joseph

Lecturer

Curtin University

Kent Street, Bentley, Perth,
Western Australia 6102

+61892667180

p.joseph@curtin.edu.au;
http://oasisapps.curtin.edu.au/staff/profile/view/P.Joseph

ABSTRACT

Storytelling in museum environments can be 'materialized' using print and visual media formats. The use of graphical visualization techniques improves the interpretation in a narrative context, hence helps to convey information and deliver a better understanding of a story. This research was prompted by the rising use of interactive visualization techniques for 'storytelling' in museum environments.

KEYWORDS

Visualization, multimedia, interactive, storytelling, digital humanities, virtual museum, museum environment

1. INTRODUCTION

The use of multimedia to present and communicate information is a popular trend in our society. This is evident in museums that are increasingly using multimedia technologies, online multimedia exhibitions, and virtual museums to successfully showcase their collections and to tell stories about their physical objects. Using multimedia technologies enables stories about museum collections to be discoverable by anyone and from anywhere in the world.

Innovative options for 'storytelling' open new ways to communicate and convey stories and to share information. Digital technologies play an increasingly important function in the digital humanities discipline, in cultural heritage areas, and in the museum environment. These emerging technologies raise questions about the future roles of professionals in the museum and digital humanities disciplines. What skill sets do they need to apply these new techniques and methods? More importantly, how can they keep abreast of technological developments in the 21st century?

We answer these questions by sharing our experience in developing an interactive digital multimedia production to tell the story of the physical Markham car collection held by a state museum in Western Australia.

a. Related Works

In a larger context, this research work is touching the field of Digital Humanities (DH) (see e.g. [1] and [2]), investigating the knowledge process in the humanities domain. The research work in particular addresses the utilization of modern visualization technology in the domain of DH, as e.g. [3] and [4, pp. 31–36] to make efficient use of information visualization [5] and its related aesthetics [6]. It also describes how latest multimedia technology can be applied in museum contexts, as e.g. described in [7]. The content and storytelling aspects are based on research conducted by the second author Pauline Joseph published in the academic peer-reviewed journal [8].

2. DIGITAL MEDIA TECHNOLOGY IN MUSEUM ENVIRONMENTS

a. Visualization for Museum Collections

Using visual imagery is an excellent technique to convey and share information; and also to tell and comprehend stories. An "emerging class of visualization" combines "narratives with interactive graphics" [9, p. 1139], which supports storytelling in an efficient way [9, p. 1140]. Consequently, enabling the information or story to be more "comprehensible,

memorable and credible to the general public" [10, p. 19].

Skillful employment of visualization techniques enables a story to be told through the "graphical depiction of statistical information" [11]. Information visualization can also refer to "computer generated interactive graphical representations of information" [5], and its process [12, p. 387]. It can be functional and aesthetic [6].

Using just printed text without visual aids to communicate lots of information decreases comprehension of the messages being communicated. Hence, the inclusion of visual media is vital to survival in the current "information ocean" [13, p. 1].

This application especially follows Paivio's (1971) 'dual-coding hypothesis', "pictorial elements are easier to retrieve from memory". In Jacoby et al.'s view (1983), "the larger the number of sensory modes" to deliver a message is "the greater the likelihood of effective communication" [14, p. 378]. Furthermore, psychological research results indicate that, cognitively, information encoded in two forms – visual and verbal – is more easily retrieved from memory than single-format information [15, p. 921].

b. Multimedia Technology

Using multimedia technologies to present information in a number of ways, including through visualization, has gained popularity in recent times. Combining different media elements/components - such as textual information, photographs, pictures, graphics, drawings, video, audios, sounds, and animation – digitally into one tool, results in multimedia.

The visual elements in multimedia technologies support the accompanying verbal information, thereby facilitating quick comprehension of information and stories being communicated.

"Multimedia includes any presentation combining more than one format (…) within a single sensory modality (…) or across modalities" [15, p. 918]. The application of multimedia elements gives the benefit of showing "movement and sound along with visuals", and multimedia platforms give the opportunity to "stimulate real life situations" [16, p. 197]. Its use in education is 'commonplace', and now multimedia-use appears in a variety of other areas. Research findings indicate that "knowledge acquisition from multi-format sources has largely supported the effectiveness of multimedia relative to single-format learning", and its advantages are proven in the ease of understanding fact-based content and "comprehension of expository information" [15, pp. 918–919].

c. Virtual Exhibitions: Media Technology in Museums

Museum environments are increasingly incorporating multimedia technology in their presentations [7]. Multimedia elements are usually used alongside museum objects or collections to inform the visitors about the physical exhibits and enable engagement to enhance their experience with the collection.

An example of multimedia use in the museum environments is the virtual museum. Many online multimedia museum exhibitions (virtual museums) successfully work, so that their collections, and stories are discoverable for anyone and from anywhere in the world. However, some issues need to be considered when creating virtual exhibitions. First and foremost the visual consistency in the user interface (UI) design must be maintained. Subsequently, the digital information must be clear, immersive, easily navigable, and interactive to enable the effective comprehension of the content without difficulty.

3. The Story of the Markham car Collection

The story of the Markham car collection is an interactive digital multimedia production (DMMP). It tells the story of Percy Markham, an antique car collector, who wanted to leave a cultural heritage of antique cars for Western Australians. The story provides a historical account of how Markham sold his collection of 22 vintage and veteran cars for AUD 180,000 (much below the valued price) to the Western Australian Museum Board in 1969. However, in 1988 the Museum auctioned ten of these cars, which enraged the motoring community in Western Australia. Fortunately, one veteran car, the 1898 Star Vis-á-vis survived the auction and is displayed in a local motor museum.

There are future plans to offer the final DMMP to be displayed beside the exhibition of the 1898 Star at the museum for visitors to learn this story.

a. Description of the Production

The production of the digital story is based on Joseph's research article [8]. The story combines digitized and visualized information sources and different media

formats and modals into one complex multimedia production.

b. Overview of the Structure of the Production

The production is partially interactive and mainly guided. However, the audience can also navigate it themselves. It comprises authentic 360-degree panoramic pictures (Figure 1) and partial 180-degree panoramas (Figure 2).

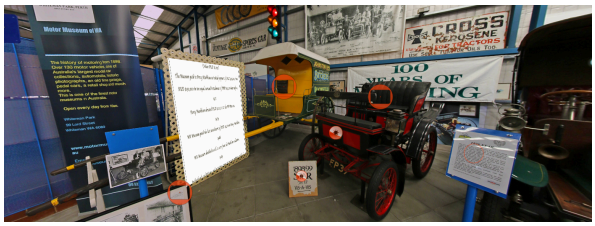

Figure 1. Motor Museum of Western Australia (authentic panorama).

Figure 2. Imaginary museum rooms (designed 180-degree panoramas).

These panoramas have been created as the virtual representation of a museum environment with various imaginary gallery showrooms. Each room represents a significant aspect and milestone in the history of the Markham car collection story.

c. Opening Scene

The opening scene displays five antique doors (Figure 3) that lead to the different gallery showrooms (scenarios of the story) in this story. The virtual museum visitors can visit each room by entering these doors; they also can 'walk' through the different scenarios by clicking on the interactive arrows/icons in each showroom. The order of the scenes is interchangeable, however, the story cannot be modified. Users will have a greater comprehension of the story if they follow the predetermined sequence of the rooms.

Figure 3. The five doors - lead to the different gallery showrooms.

d. Rooms and Levels of the Production

The first door is labeled with the numeric 'one' and with the prompt stating 'Go to the room 'STAR''. This door leads visitors to where the 1898 Star Vis-á-vis is introduced (Figure 4). This physical museum object is currently displayed at the Motor Museum of Western Australia on loan from the Western Australian Museum.

Figure 4. The room behind the first door - introduces the 1898 Star Vis-á-vis.

The second from the five doors (or the interactive button from the first room) leads to the room titled 'Markham family' (Figure 5) Here visitors 'meet' Mr. Percy Markham, the owner of the 1898 Star and are introduced to his family.

Figure 5. The room behind the second door - introduces the Markham family.

The visitor then, by clicking on an interactive arrow, seamlessly walks from the second into the third room, 'Antique Auto Museum' (Figure 6). Here, the visitors view an imaginary exhibition with 3D objects, photographs and moving slides of some cars and pictures of the museum in original conditions. Using the interactive icons they can watch short movies about the antique cars once displayed at Markham's private museum that was open to the public on weekends in the 1960s.

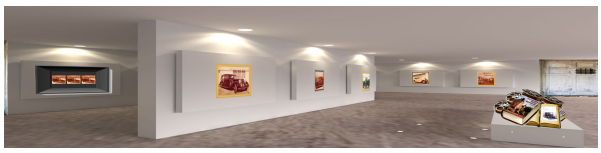

Figure 6. The room behind the third door - introduces Markham's Antique Auto Museum.

Visitors, using the interactive icon/arrow, walk to the fourth room (which is two half-rooms), 'WA Museum'. In the first part of this imaginary museum area (Figure 7), the story unfolds about the advantageous sale of the 22 Markham cars to the Western Australian Museum in 1969. In the second half-room (Figure 8), the story describes what happened between 1988 and 1989. This room can be entered from the opening scene by clicking on the fourth door's interactive button.

Figure 7. The room behind the fourth door - tells the story of what happened with the Markham cars between 1967 and 1969.

Figure 8. The second half of the room behind the fourth door - continues the story of what happened to the collection in 1988 and 1989.

e. Exiting the Virtual Tour: Final Imaginary Room

The final imaginary museum room, titled 'Christie's Auction' (Figure 9) behind the fifth door, (or following the arrow in the previous scenario) narrates the decision by the Western Australian Museum to auction ten cars in 1990. Photos of each of the ten cars published in Christie's catalog are presented.

Figure 9. The room behind the fifth door - narrates the WA Museum decision and the Christie's auction.

Leaving the 'Christie's Auction' imaginary room, the visitor can 'walk' to the Whiteman Park location and enter the front door of the Motor Museum of Western Australia, where the 1898 Star Vis-á-vis is currently displayed. Authentic 360-degree panoramas were employed to showcase this environment, where the visitors can learn the story about these cars.

The different types of media appearing within the scenarios can be read, listened to, or watched according to the users' preferences.

4. Development and Implementation

The project to develop an interactive multimedia storytelling production commenced in 2015. The story line for the project was borrowed from Dr. Joseph's research about the heritage of the Markham car collection. It is published as a peer-reviewed article titled 'Heritage of the Markham car collection: Estrangement from the West Australian motoring community' [8]. In this article, Joseph researched what happened with the Percy Markham car collection sold to the Western Australian Museum in 1967 to understand how the controversial auction of ten of these cars disappointed the motoring community. Both Joseph's meticulous account of what happened and her use of a storyline writing style to report her research findings made it suitable to select this story for the development of this DMMP.

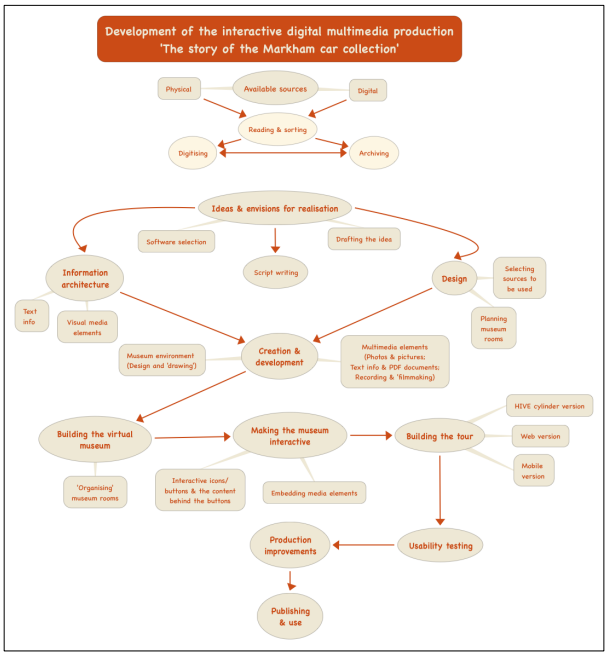

Figure 10: Stages of production development.

a. Technical Setup

A complex virtual museum environment was designed and developed where 'visitors' can discover what

happened to the prestigious Markham cars. Its intention is to offer a unique and immersive experience for the audience. The story of the car collection is told using text, pictures, photographs, audio and video panels. The production is aimed to provide users with an immersive and interactive experience. Hence, users have the free will to select the information they want to read, view, listen to or watch.

There are numerous stages in the development of the production with each of them requiring advanced technical skill sets, knowledge and experience (Figure 10). Further skills in archiving the research records and archives gathered, writing the storyline, producing the digital content and writing computer code were also required.

The production is currently in its final stages of the development. Next, usability testing is planned to employ eight to ten user experience (UX) experts. Then, these experts' feedbacks will be utilized to improve and finalize the production. Finally, the production will be promoted to tell the digital story of the Markham car collection for both live and virtual visitors.

b. From Available Source Materials to Building the Complete Tour

Information Management

Firstly, the available physical and digital records that were provided by the Markham family and Dr Joseph needed to be appraised and processed: read, sorted, archived, and digitized.

These sources were: aged photographs, old manuscripts, and original copies of archival records, paper-based newspaper clippings, magazines, books, digital photos, and panoramas.

Figure 11. Processing and archiving the paper-based research materials.

The archives collection provided by the Markham family is unique and has heritage value. Its preservation was an important aspect of the archival process (Figure 11).

Figure 12. Archiving the Markham records.

Considering the age of the different manuscripts, their protection from damage and deterioration was essential. Handling (organizing, managing) the archival materials required wearing cotton gloves to ensure their protection from salt, acid, and contamination. Also, the use of acid-free paper files and storage containers was vitally important (Figure 12). A simple descriptive catalog was created using Microsoft Excel to register and describe in detail each archival record. The Excel spreadsheet format allowed easy access for anyone to search, retrieve and access this archive register (Figure 13).

Folder	DATE	FROM	TO	KEYWORDS	DESCRIPTION	NOTE
A	1962-08-05	Endeavour Motor Museum, Lakemba	MARKHAMS	cars	napoleon, silver ghost, 1909, daimler	
A	1965-04-13			press	newspaper	
A	1967-11-02	MARKHAMS	David Brand Premier of WA	cars	antique car museum	*
A	1968-01-15	MARKHAMS		note	new address note	
A	1968-07-01	Rolls-Royce Owners Club of Australia	MARKHAMS	advert	advertising	
B	1969-03-20	MARKHAMS	David Brand Premier of WA	cars	antique car museum	
B	1969-03-26	David Brand Premier of WA	MARKHAMS	cars	antique car museum	*
B	1969-04-21	MARKHAMS	David Brand Premier of WA	cars	antique car museum	

Figure 13. Excerpt from the Markham archives' description.

Once the source archival records were registered and organized, the project focused on identifying suitable software to develop the MMDP.

Information Technology Issues

The key software selected to vivify the 'story' was the Kolor PanoTour Pro 2.5. It was primarily selected owing to its functionalities that are suitable for and capable of creating a museum environment and for building an immersive virtual tour (Figure 14). It provides functionality to insert different hotspots (e.g. navigation icons, buttons), which enabled the creation of the 'museum' to be interactive.

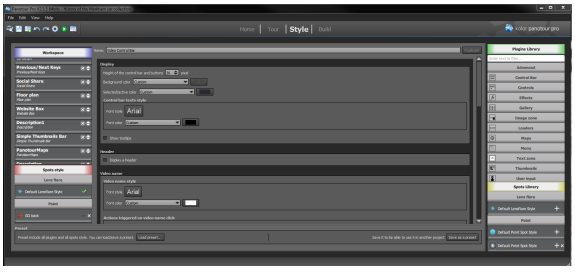

Figure 14. Kolor PanoTour Pro 2.5 user interface.

In addition, other computing software and applications, such as Adobe Photoshop, Audition, Premier Pro,

iMovie, Photos, and Sublime Text 2 were used in the production design and development (Table 1).

Table 1. Software used for design and development of the DMMP.

Software/ application	Used for/to	Required skills/knowledge
Adobe Photoshop CS6 Version 13.0.4 x64	Edit, improve quality of photos, documents; create 3D objects, and partial 180 degrees panoramas	Advanced
Adobe Audition CS6 Version 5.0.2	Editing sound recording	Professional
iMovie Version 10.1.3	Creating short clips and movies	Advanced
Photos (Apple) Version 2.0	Creating built-in clips/slides	Advanced
Adobe Premier Pro CS6 Version 6.0.5	Postproduction editing (audio and video)	Professional
Sublime Text 2 Version 2.0.2.	Editing HTML codes / sources (web publishing)	Professional
Adobe Acrobat X Pro Version 10.0.0	Creating PDFs	Basic
Kolor PanoTour Pro Version 2.5.5 64 bits	Develop the virtual tour	Professional

c. Developing the Story Elements

Scriptwriting

The next major step in developing the DMMP was the scriptwriting for the story. From the available and organized sources (e.g. Joseph's article, archival records, magazines, newspaper clips, photographs) the storyline had to be composed. As stated previously, the narrative was based on research article [8], and the supporting materials to visualize the story were used to compile the 'screenplays'. Additionally, the first author also conducted research to source for media elements (e.g. the video clip of the London to Brighton Rally 1938, where the 1898 Star Vis-á-vis also appears), and other photographs (e.g. "participants" in the story other than the Markham family) while scriptwriting the story.

From Storyboard to Storyline

Drawing the storyboard was the next stage in developing the production (Figure 15). During this juncture, decisions were made about where the different media elements, such as textual information, pictures, audio, video or film clips would be embedded in the digital story.

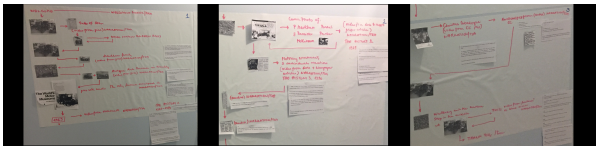

Figure 15. Drawing storyboard.

d. Digital Media Design

Design Approach

After the complete storyline implementation plan had been outlined, the next phase was to select and sort the appropriate pictures, and records. Lacking personal artistic ability and drawing skills, stock photos, pictures, and vectors (e.g. doors, rooms, objects) needed to be purchased to design the imaginary museum rooms. This decision was also made to afford a professional look and feel of a museum setting. From these elements and the available photographs, using Adobe software, 3D objects were created; and imaginary museum rooms were designed and developed.

The rooms were developed to provide an 180-degree panoramic view, with a size of 5048x1200 pixels to fit on the Cylinder screen located in Curtin University's HIVE (Hub for Immersive Visualisation and eResearch), in Perth, Western Australia. The full authentic spherical panoramas that were provided by Dr Joseph, developed with the assistance of A/Prof Paul Bourke (iVEC@UWA) did not require a new design, development, editing or modifications.

Improving the Quality of Media Elements

After designing the museum environment, another significant step was to edit and improve the quality of the aged original archival records, and photographs. Using the Adobe Photoshop software, the image quality of these deteriorated photographs was enhanced until its colour qualities appeared brighter, sharper and clearer.

Multimedia Work

The next crucial steps were to create the multimedia elements of the production. This process consisted of generating PDF documents, recording – cutting – editing voice-overs, making movie clips and short films.

The multimedia development (Figure 16) required advanced knowledge, understanding, and proficiency in the use of different software (e.g. Adobe Audition, Premier Pro, iMovie). These procedures were extremely time-consuming, as it required hours of patience to pay meticulous attention to the details (e.g. design principles and elements to enhance the quality of the archival materials, preciseness in the sound editing process, inserting layers of various media elements, aligning of image and sound components).

Figure 16. Multimedia work.

The multimedia work stage was followed by the next phase in the production development: building the virtual museum**.**

The production consists of eight authentic panoramic pictures (Whiteman Park, and the Motor Museum of WA), and thirteen partial panoramas (imaginary scenes and museum environment) (Figure 17). Every museum room represents a key point of the story, and the rooms are organized in chronological order.

Figure 17. PanoTour Pro in use.

Each room has multiple media elements embedded: e.g. text, audio, and video. These components assist telling the story at a particular time intended for the showroom.

5. User Experience and Engagement: Navigating through the Virtual Museum

a. Giving Life to the Tour

To animate the production and make the virtual museum experience engaging for users, interactive icons and navigation buttons were designed and positioned where required in the DMMP. Similarly, different media elements were embedded to make the story interactive. Navigating through the virtual museum, users can perform various actions, move backward and forward between the museum rooms or open links to a website. They can view digitized photographs/pictures, listen to audios, read information, read archived documents or watch videos to learn and understand the story of the Markham car collection.

Figure 18. Information and navigation screen in the production.

Finally, three forms (one tour for three platforms) of the virtual tour have been built to cater for the multiple computing interfaces in which the digital story will be presented to the users (Table 2).

Table 2. Features of the different versions of the virtual tour.

	Web	Cylinder	Mobile
Text information	✔	✔	✔
PDF document	✔	✔	✔
Pictures and photos	✔	✔	✔
Built-in clips	✔	✗	✗
Movie	✔	✔	✗

Audio	✔	✔	✔
Website	✔	✔	✗
Rotation	✔	✗	✔
VR	✔	✗	✔

Firstly, a full version of the production was built for a web interface (Figure 19). The PanoTour Pro software generates the tour in .html format. Sublime Text 2 software was used to modify and improve some of the source code. This version contains the full range of multimedia elements: textual information, pictures and photos, different audio components, short movies, built-in video clips, and PDF documents. The advantage of using the web version on mobile devices is - in the case of full panoramas, users wearing Virtual Reality (VR) headset can enjoy a VR environment.

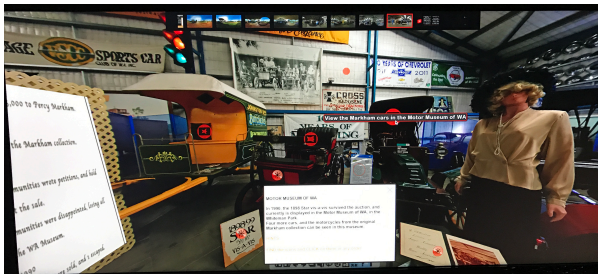

Figure 19. DMMP in use on the web interface

Secondly, the tour was developed and built for presentation on the wide Cylinder screen at the Curtin's HIVE (Figure 20). This second version of the production omitted built-in video clips. The cylinder interface allows users to have an immersive experience, with a more engaging and satisfying visual view of the story.

Figure 20. DMMP demo showcase on the Cylinder (Photograph was taken by Prof Erik Champion at GLAMVR16 symposium).

Thirdly, the tour was generated in a different file format (.pvt) so that it can be used on iOS and Android devices if the users download the free PanoViewer application (Figure 21). The disadvantage of this version is that some media components (e.g. built-in clips and movies) do not work correctly. However, its advantage is that users will be provided some form of VR experience by moving/rotating the mobile device, the users can look around as they walk through the museum rooms.

Figure 21. Using mobile application to interact with the story.

b. From Usability Testing to Publishing

It is important to understand user requirements when developing virtual museums. Such online interactive user interfaces need to be designed to be intuitive and captivating for the users to stay online and explore the DMMP and the story it conveys. Hence, conducting usability testing is vital before finalizing the production for public access.

Usability tests for this project will be conducted to investigate the interactive DMMP's usability and the users' expectations against four criteria: learnability, understandability, attractiveness, and satisfaction. Its aim is to have UX experts conduct an 'expert review' of the production and elicit their feedback. These expert testers' feedback will be applied to improve the DMMP before releasing it for public access for the next phase of the research goals.

c. Current Implementation Status

This interactive DMMP, *'The story of the Markham car collection',* is currently being reviewed by usability experts, it will then be refined before making it accessible to users. It will be released online and showcased on the Cylinder in Curtin's HIVE in mid-2017.

The final version of the DMMP will be used by Dawson, in her Ph.D. research, to investigate if museum visitors prefer discovering stories about the exhibited physical collections by reading about the object from printed sources like Joseph's (2016) article or by engaging with an interactive DMMP or both.

Dr. Joseph uses this DMMP as an educational resource for her information studies' students to explore lessons

learned from this empirical case study of the Markham car collection. Pedagogy topics concerning ethics, role of cultural institutions, and community engagement issues are discussed using this digital production. The storytelling design of the DMMP allows for easy use of this production for educational purposes both when teaching online and face-to-face. Other academics may see the potential in using this DMMP in an educational context for pedagogy reasons.

6. CONCLUSIONS

Applying 21st-century digital technologies is vital in the digital humanities discipline, especially in museum environments. The benefits of visualization and the use of multimedia presentations are to convey visually impressive stories or share information about museum collections. This improved type of 'communication' simplifies the way in which the users/museum visitors access the relevant information. Also, the rich visualization and the immersive experience afforded by DMMPs enable the fuller understanding of a story for the visitors.

We feel it is worthwhile for professionals in the digital humanities and possibly in the GLAM sector to equip some of their staff with DMMP development skill sets. It requires learning and engagement with different hardware, software, and applications. This investment and the implementation of the evolving technologies would yield success in showcasing their collections; improve community involvement and position museums as destination for tourist attractions.

a. Future Development

Future plans and challenges are to develop/engineer the content to achieve a more immersive Virtual Reality and/or Augmented Reality and/or Mixed Reality experience. Also, another goal is to apply the proximity beacon (Bluetooth Low Energy transmitter) technology (e.g. iBeacon, Eddystone, Estimote) on museum objects (e.g. cars, 1898 Star Vis-á-vis) to deliver the content (the story or part of the story) contextually and provide personalized experiences to users on location.

ACKNOWLEDGEMENTS

The first author expresses special thanks to her Ph.D. supervisors Dr. Pauline Joseph and Professor Erik Champion for the continuing assistance and encouragement through the production of the creative work.

The authors' sincere thanks go to the Markham family for providing access to Percy Markham's archival records and the family photographs. The authors thank Ronan Sulich, Director and Australian Representative from Christie's for granting permission to reproduce the photos of the Markham cars published in Christie's auction catalog. The authors are thankful to Thomas Benson-Lidholm, a motoring enthusiast, for providing expertise in historical motorcars that greatly assisted the research.

We are appreciative of the valuable technical assistance provided by Jesse Helliwell (HIVE), Adam Barrett and Joachim Strand (CPSU). We are grateful to Luke Worthington for lending voice to the story.

REFERENCES

[1] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp, *Digital Humanities*. Cambridge, USA: Massachusetts Institute of Technology (MIT), MIT Press, 2012.

[2] J. Drucker, D. Kim, I. Salehian, and Bushong, "Introduction to Digital Humanities." 2014.

[3] A. Lugmayr and M. Teräs, "Immersive interactive technologies in Digital Humanities: A review and basic concepts," in *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, Brisbane, Australia, 2015, pp. 31–36.

[4] A. Lugmayr, A. Greenfeld, A. Woods, and P. Joseph, "Cultural visualisation of a cultural photographic collection in 3D environments - Development of 'PAV 3D' (Photographic Archive Visualisation) In: Wallner G, Kriglstein S, Hlavacs H, Malaka R, Lugmayr A, Yang H-S, editors. Entertainment Computing," presented at the ICEC 2016: 15th IFIP TC 14 International Conference, September 28-30, 2016, Vienna, Austria, 2016, pp. 272–277.

[5] C. Chen, *Information visualization: beyond the horizon*, Second. London: Springer London, 2006.

[6] A. R. Gaviria, "When is visualization art? Determining the critical criteria," *Leonardo*, vol. 41, no. 5, pp. 479–482, 2008.

[7] F.-S. Hsu and W.-Y. Lin, "A multimedia presentation system using a 3D gesture interface in museums," *Multimed. Tools Appl.*, vol. 69, no. 1, pp. 53–77, 2014.

[8] P. Joseph, "Heritage of the Markham car collection: Estrangement from the West Australian Motoring Community," *Collect. J. Mus. Arch. Prof.*, vol. 12, no. 1, pp. 21–44, 2016.

[9] E. Segel and J. Heer, "Narrative visualization: telling stories with data," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1139–1149, 2010.

[10] K.-L. Ma, I. Liao, J. Frazier, H. Hauser, and H.-N. Kostis, "Scientific storytelling using visualization," *IEEE Comput. Graph. Appl.*, vol. 32, no. 1, pp. 12–19, 2012.

[11] J. Stikeleather, "How to tell a story with data," *Harv. Bus. Rev.*, 2013.

[12] C. Chen, "Information visualiation," *Issue Wiley Interdiscip. Rev. Comput. Stat. Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 387–403, 2010.

[13] J. Sviokla, "Swimming in data? Benefits of visualisation," *Knowledge Management*, 2009. .

[14] M. Dijkstra, H. Buijtels, and W. F. van Raaij, "Separate and joint effects of medium type on consumer responses: a comparison of television, print, and the Internet," *J. Bus. Res.*, vol. 58, no. 3, pp. 377–386, 2005.

[15] T. T. Brunyé, H. A. Taylor, D. N. Rapp, and Spiro Alexander B, "Learning procedures: the role of working memory in multimedia learning experiences," *Appl. Cogn. Psychol.*, vol. 2ö, no. 7, pp. 917–940, 2006.

[16] A. Gaur and K. K. Grover, "Application of multimedia in the industrial sector," *Comput. Control Eng. J.*, vol. 12, no. 4, pp. 197–200, 2001.

AUTHORS

Beata Dawson is a Ph.D. candidate at Curtin University (Media Culture and Creative Arts / Information Studies). She holds a Master's Degree in Information Management from Curtin University (2014), and a Bachelor Degree in Pedagogy from Eszterházy Károly Főiskola (currently Eszterházy Károly University of Applied Sciences), in Eger, Hungary (2001). Before she moved to Perth, Australia she partially completed her postgraduate degree in Library Informatics (2007), and had qualifications in Foreign Trade and Customs Administration (1993, 1994), as well.

She previously worked in foreign economics area, and as an educator, and is currently working in information management.

She has conducted research in educational history, child protection, drug prevention, and in the information field: in library history, and in online information retrieval.

Her current research is cross-disciplinary and related to communication technology, digital media, heritage and museum studies.

Pauline Joseph (Ph.D.) is a Lecturer in Records and Archives Management at the Department of Information Studies at Curtin University. Pauline studies how information is perceived and used in organizations and communities. With this focus in mind, she completed her Ph.D. at the University of Western Australia in 2011, studying how knowledge workers search for corporate information in the electronic document and records management systems (EDRMS). Her Ph.D. research is titled "EDRMS search behaviour: implications for records management practices".

Pauline's current research interests are about the sustainability of community-based information management practices using the motorsport community as a case study. An aspect of this research investigated how and why the motoring community became estranged with cultural institutions.

REVIEW OF MACHINE LEARNING ALGORITHMS IN DIFFERENTIAL EXPRESSION ANALYSIS

Irina Kuznetsova
Graz University of Technology
Graz, AUSTRIA
and
Inst. of Interactive Systems and Data Science, Harry Perkins Inst. of Medical Research
University of Western Australia
Perth, AUSTRALIA
i.kuznetsova@hci-kdd.org
http://hci-kdd.org

Yuliya V Karpievitch
Plant Energy Biology Centre of Excellence
and
Harry Perkins Inst. of Medical Research
University of Western Australia
Perth, AUSTRALIA
yuliya.karpievitch@uwa.edu.au

Aleksandra Filipovska
Harry Perkins Inst. of Medical Research
University of Western Australia
Perth, AUSTRALIA
aleksandra.filipovska@uwa.edu.au
https://www.perkins.org.au/our-people/laboratory-heads/filipovska-profile/

Artur Lugmayr
Visualisation and Interactive Media Lab. (VisLab)
Curtin University
Perth, AUSTRALIA
lartur@acm.org
http://www.artur-lugmayr.com

Andreas Holzinger
Inst. of Interactive Systems and Data Science
Graz University of Technology
Graz, AUSTRIA
a.holzinger@hci-kdd.org
http://hci-kdd.org

ABSTRACT

In biological research machine learning algorithms are part of nearly every analytical process. They are used to identify new insights into biological phenomena, interpret data, provide molecular diagnosis for diseases and develop personalized medicine that will enable future treatments of diseases. In this paper we (1) illustrate the importance of machine learning in the analysis of large scale sequencing data, (2) present an illustrative standardized workflow of the analysis process, (3) perform a *Differential Expression (DE)* analysis of a publicly available *RNA sequencing (RNA-Seq)* data set to demonstrate the capabilities of various algorithms at each step of the workflow, and (4) show a machine learning solution in improving the computing time, storage requirements, and minimize utilization of computer memory in analyses of RNA-Seq datasets. The source code of the analysis pipeline and associated scripts are presented in the paper appendix to allow replication of experiments.

KEYWORDS

Machine learning; big data; data mining; Next Generation Sequencing; Burrows-Wheeler transform; semiglobal alignment; clustering; biology; RNA-Seq

1. INTRODUCTION

Every living cell's genome is encoded in *DNA* (*Deoxyribonucleic Acid)* – a long sequence of nucleic acids, also called nucleotides, encoded in four letters: *adenine*, *thymine*, *cytosine*, and, *guanine*, abbreviated as *A*, *T*, *C* and *G* respectively. DNA provides recipes for making all active molecules in the cell, such as *RNA (Ribonucleic Acid)* and proteins [1, 2]. In RNA *T* is replaced by *uracil (U)* and proteins are made of twenty amino acids.

Information stored in the DNA is transcribed into RNA some of which are further translated into proteins, which are the main workhorses of the cell. These three steps are also referred to as genome, transcriptome and proteome, which make up multidimensional data.

Currently genome and transcriptome are analysed via *Next Generation Sequencing (NGS).* Sequencing is the process of transforming molecular information into a digital format. NGS allows sequencing the entire genome, specific regions of the genome, as well as epigenetic modifications of the genome (e.g. Methyl-C-seq, ChIP-seq) [3-6].

Sequencing RNA provides a snapshot of cellular gene expression, which is compared among various treatment groups or disease states to identify differentially expressed genes. *RNA sequencing (RNA-Seq)* reveals the order of the four nucleotides in short segments [7, 8]. These short segments are called *reads* and are stitched together algorithmically into a large genome sequence computationally [4, 6, 9]. There are two ways to sequence DNA/RNA fragments using single-end or paired-end sequencing. In single end sequencing DNA/RNA fragment is sequenced from one end only, whereas in pair-end sequencing fragment is sequenced from two opposing ends producing two reads per fragment.

NGS generates vast amounts of data, usually hundreds of gigabytes [5], commonly referred to as Big Data, and as researchers we aim to transform into Cognitive Big Data [10, 11] or other practical use cases as e.g [58] or [59]. Therefore automatic machine learning plays a crucial role in handling, interpreting, learning and visualising the big NGS datasets to produce easily understandable knowledge base.

Here we illustrate the importance of machine learning algorithms in analyzing big data and provide a specific example of analysis pipeline (also referred to as workflow) of NGS data. Our analysis pipeline consist of a standardised modular workflow where some modules are taken from the pipeline proposed by Partea et al. [12].

Within the scope of this paper, we show the utility of machine learning algorithms in identifying genes that are different among various conditions. We use two publicly accessible mouse RNA-Seq data sets available from the NCBI GEO database accession NCBI GEO database under the accession number GSE56933 [13] and under the accession number GSE60450 [14].

a. Research Goals

Here we provide a detailed review of the algorithms most widely used in RNA-Seq DE analysis. We showcase the necessity and requirements for each step in the analysis pipeline and deliver the minimal required knowledge about RNA-Seq DE analysis.

We show:

- practical impossibility of analysing NGS data without the use of computer algorithms
- standardized workflow of DE analysis
- utilization of publicly available data for analysis and new algorithm development

To achieve our goal, we first describe several potential pipelines. Next, we evaluate and selected software and last, we perform an illustrative performance analysis which can be used as a learning example.

2. Tools for Transcriptomic Data Analysis

Transcriptome studies using RNA-Seq reveal quantitative information about transcripts, which are fragments of RNA. Transcripts, in turn, show variation in patterns of gene expression at specific developmental time points, in various treatment conditions [9, 15-17]. As such, the most common goal in RNA-Seq analysis is finding differentially expressed genes across different conditions.

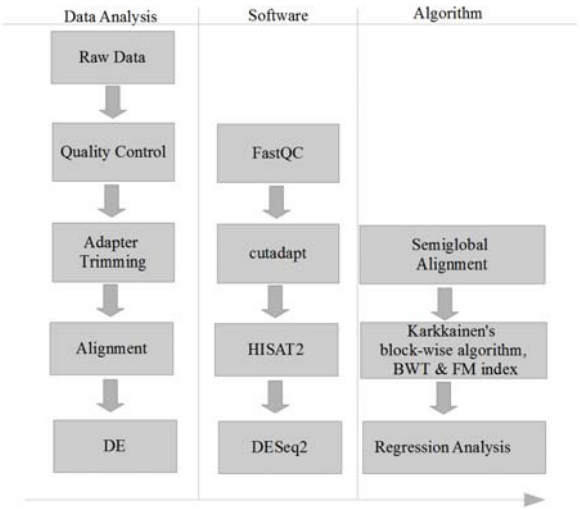

Figure 1. Data analysis workflow, software, algorithms.

A standard RNA-Seq workflow can be divided into the following steps:

1. Design of experiment: determine the biological question to study and estimate the sample size necessary to identify new knowledge with statistical significance
2. Perform the biological experiment using cells or animal models under different treatment conditions or genetic alterations
3. Obtain sequencing data [7]

4. Preprocess data: perform quality control, adapter trimming, and alignment
5. Analyze data: identify coverage at gene- or transcript- expression level, normalize, find which genes behave differently across conditions, visualize results, and validate results if needed

Our differentially expression analyses example is outlined in Figure 1. It includes quality control, adapter trimming, alignment, and differential expression. The tools we suggest to use are FastQC [18] (quality control), cutadapt [19] (adapter trimming), HISAT2 (alignment) [20], DESeq2 [21] R package (DE). There are alternative methods to perform most of the steps.

To understand the performance of the applied tools, we conducted a performance analysis such as algorithm execution time, memory (RAM) and CPU usage.

Table 1. Representation of tools that can be used for RNA-Seq data analysis.

		Task	Tools	Algorithm	Ref
Preprocessing	Quality control	Adapter trimming; poor quality bases elimination	Cutadapt	Semiglobal alignments	[19]
			FASTX-Toolkit	-	[22]
			Trimmomatic	Seed and extend followed by palindrome mode approach	[23]
			BBDuk	-	[24]
	Alignment	Alignment of reads to the reference genome	BWA	Burrows-Wheeler transform	[25]
			Bowtie 2	Burrows-Wheeler transform	[26]
			HISAT2	Karkkainen's blockwise algorithm	[20]
Analysis	Differential Expression	Identify differentially expressed genes	Deseq2 (R)	Negative binomial generalized linear models	[21]
			edgeR (R)	Negative binomial generalized linear models	[27]
			Ballgown (R)	Standard linear model-based comparison statistical test	[28]

3. PIPELINE SETUP

a. Data Description

We downloaded the GSE60450 [14] dataset from the *Gene Expression Omnibus (GEO).* These data is of gene expression changes in luminal and basal mammary glands of non-pregnant, pregnant and lactating mice was run on the Illumina HiSeq 2000 platform. Sequenced total RNA from the samples generated single-end reads of 100 *base pairs (bp)* in length.

The second GSE56933 [13], RNA-Seq dataset was extracted from the heart and liver tissues of 10-weeks old male mice and ran on the Illumina Genome Analyzer IIx. The resulting reads are single-end of 75 bp in length and contain adapter sequences. These data are selected to exemplify the necessity of adapter trimming algorithms that remove adapter sequences before the data is aligned to a reference genome. Since adapter sequences are not present in organism's genome reads that contain sequences will fail to correctly align to the genome.

In both datasets raw sequence information was saved in FASTQ file format. FASTQ file format is text-based with four lines of the file describing one read/sequence at a time. The first line is a sequence identifier, the second, is the sequence itself, the third line is no longer used for sequence identification and contains a +, and forth line contains sequence quality information [29].

b. Quality Control

Quality control is performed on the raw data to detect low quality bases, duplicates, PCR primers, or adapters in the reads. FastQC [18], an open-source software is widely used to investigate: (1) per base sequence quality, (2) per sequence quality scores, (3) per base sequence content, (4) per base GC content, (5) per base N content, (6) sequence length distribution, (7) duplication level, (8) overrepresented sequences, (9) adapter content, (10) kmer content, and (11) per tile sequence quality [18]. The quality summary provides information on possible artifacts in the raw data that can affect the next steps of the RNA-Seq analysis. The adapter content section of the FastQC report gives information about the adapter sequence observed in the data. Additionally, the overrepresented sequences

section of the FastQC report sometimes, but not always, gives more specific information of the possible adapter source. For example, Figure 2A shows a list of overrepresented sequences found in GSE56933 dataset, and indicates presence of the TruSeq adapter. The pattern of the adapter can be found in the Illumina adapter catalogue and should always be provided by the authors [30].

A. Overrepresented sequences before adapter trimming

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA	673591	2.695638929	TruSeq Adapter, Index 7 (97% over 36bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAA	394423	1.578438538	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAACA	96534	0.386318713	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAC	78364	0.313604322	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAACAA	60850	0.243515173	TruSeq Adapter, Index 7 (97% over 36bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAACA	42623	0.170572674	TruSeq Adapter, Index 7 (97% over 36bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAC	37795	0.151251536	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAATA	31804	0.127276197	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAGA	27342	0.109419751	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAG	26938	0.107802987	TruSeq Adapter, Index 7 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAT	26640	0.106610422	TruSeq Adapter, Index 7 (97% over 36bp)

B. Overrepresented sequences after adapter trimming

Sequence	Count	Percentage	Possible Source
	2017092	8.072185821	No Hit
A	934711	3.74061316	No Hit

Figure 2. FastQC HTML report overview of overrepresented sequences before (A) and after (B) trimming [18].

c. Adapter Trimming

Once the presence of adapters is identified, the next step is to remove/trim those adapters prior to read alignment to a reference genome.

To perform adapter trimming we used one of the commonly used trimming tools, cutadapt [19]. Cutadapt can remove adapters an error-tolerant way, which means that it will trim adapters even if errors were introduced during the sequencing. The adapter sequence can be fully or partially present in the read.

Cutadapt uses unit costs function to consider mismatches (a single nucleotide base substitution), insertions (one or more base pair are added to the sequence), or deletions (one or more base pairs are lost in the sequence) as a single error score. The best mapped sequences are those that have an overlap score maximized but are below the selected threshold error rate. If after passing this condition there are multiple options of mapped sequences, the mapping with the smallest error rate is selected. However, if by passing this condition there are still more than one mapped sequences, then the mapping of the adapter sequence to the most left position on the read is selected as the best match [19]. To validate that all adapters have been trimmed we run FastQC software on adapter trimmed data. Figure 2B shows a list of overrepresented sequences after trimming with cutadapt. Table 1 contains a list of adapter trimming software.

d. Alignment

The next step in the analysis pipeline is read alignment. For organisms for which genome sequence (reference genome) is available the reads are aligned to that organism's reference genome. For organisms without genome reference genome *de novo* assembly is required (de novo assembly is beyond the scope of this manuscript).

Although there are many programs that can perform sequence alignment not all of them are appropriate for NGS data, such as BLAST [31] and BLAT [32], as these tools were developed for low volume datasets.

Table 1 contains a list of aligners suitable for NGS sequencing data. The main goal of the alignment process is to correctly align and assemble a large number of short reads to a reference genome, which is a time consuming process, requires temporary disk storage, and intensive CPU usage. Majority of currently available alignment tools use algorithms that can be categorized into two groups: (1) based on hash tables and (2) based on suffix tree [33].

The hash table algorithm stores reads of the data in an array that can be retrieved with an index, where similar values are stored at the same location under the same

index. On the other hand, the suffix array keeps the suffixes of the data values in a tree-like manner, where identical copies of reads' suffixes are stored at the same path. Both approaches attempt to minimize execution time.

When selecting an algorithm to process the data one has to consider the following factors: (1) reference genome size, for example, human genome consists of approximately of three billion bp [34], which comes at a cost of time and space for its alignment as compared to a shorter genome, (2) the length of the individual read can range from 25 to 500 bp and the length affects the accuracy and speed of the alignment. The longer reads will take longer to align.

In our example we utilize *Hierarchical Indexing for Spliced Aligner (HISAT2)*[16] that uses the block-wise algorithm of Karkkainen in combination with the *Burrows-Wheeler Transformation (BWT)* and *Ferragina and Manzinin (FM)* index, data transformation and compression algorithms respectively [35-37]. HISAT2 [16, 20] can be used to align genes that have annotated splice sites, unlike Bowtie 2 [26], which is splice-unaware aligner. Figure 3 depicts the need for splice-aware alignment in higher level organisms.

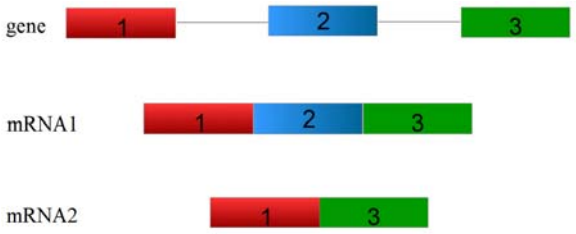

Figure 3. Alternative splicing. One gene with three introns is transcribed into two different mRNAs, one containing all three introns (mRNA1), and the second (mRNA2) with the second intron spliced out.

The BWT is a compression algorithm and *suffix array (SA)* is lexicographically sorted array that when combined create space-efficient index or the FM-Index, which uses a prefix as a search pattern [36]. Storage space requirements of SA can be reduced to small blocks that is the approach of Karkkainen's algorithm [37]. A detailed explanation of the BWT and FM-indexes is described in Langmead's tutorial *("Introduction to the Burrows-Wheeler Transform and FM Index" Langmead 2013*) and the block-wise algorithm is thoroughly explained in [37].

The output of the alignment software is a *SAM* format file which is a tab delimited file that contains the alignment information, for example, a read sequence that mapped to a genome, quality score of the alignment, genome mapping position (coordinate) [38]. The SAM files are large human readable text files , which require a lot of storage space, therefore they are commonly converted to a binary BAM format [38]. BAM files are accompanied by index files with extension *bai* to speed up access time.

All aligners will produce a short summary of the alignment such as a number of raw reads and the number of aligned reads. If the number of aligned reads is satisfactory, which is generally anywhere above 70% of raw reads, next step is to visualise the aligned reads. Genome browser such as *Integrative Genomics Viewer (IGV)* [39] or UCSC genome browser [40] are some of the best ways of confirming correct alignment.

IGV is a freely available Java based visualization platform that is run locally on a computer, and enables to explore fragments that are mapped to the reference genome. IGV interface enables one to perform wide range of tasks, such as zooming in to the region of chromosome/gene of interest with good resolution, coloring, or sorting [39]. Due to the need of installation on a local PC performance of IGV is faster than web-based genome browsers (such as the UCSC genome browser). IGV is a quick way to visualize the data, whereas researchers use UCSC browser to produce publication qualities images. UCSC browser allows for easy data sharing among multiple collaborators.

Figure 4 illustrates read coverage (raw counts) for the *EGF* gene that provides a general overview of the depth of sequencing coverage. Visualization of the entire genome coverage provides limited information due to data being highly condensed; zooming in at the level of a region or gene of interest will provide sufficient detail about the coverage.

e. Differential Expression

DE analysis investigates differences in RNA levels among various samples as readout of gene expression changes. In general, DE analysis takes aligned raw counts, subjects them to normalization to improve comparability across samples prior to estimating statistical significance of the gene expression change.

Several R packages exist that analyze RNA-Seq data for detecting differentially expressed genes across various conditions. The most popular are edgeR [26], DESeq2 [21] and Ballgown [16, 27]. While DESeq2

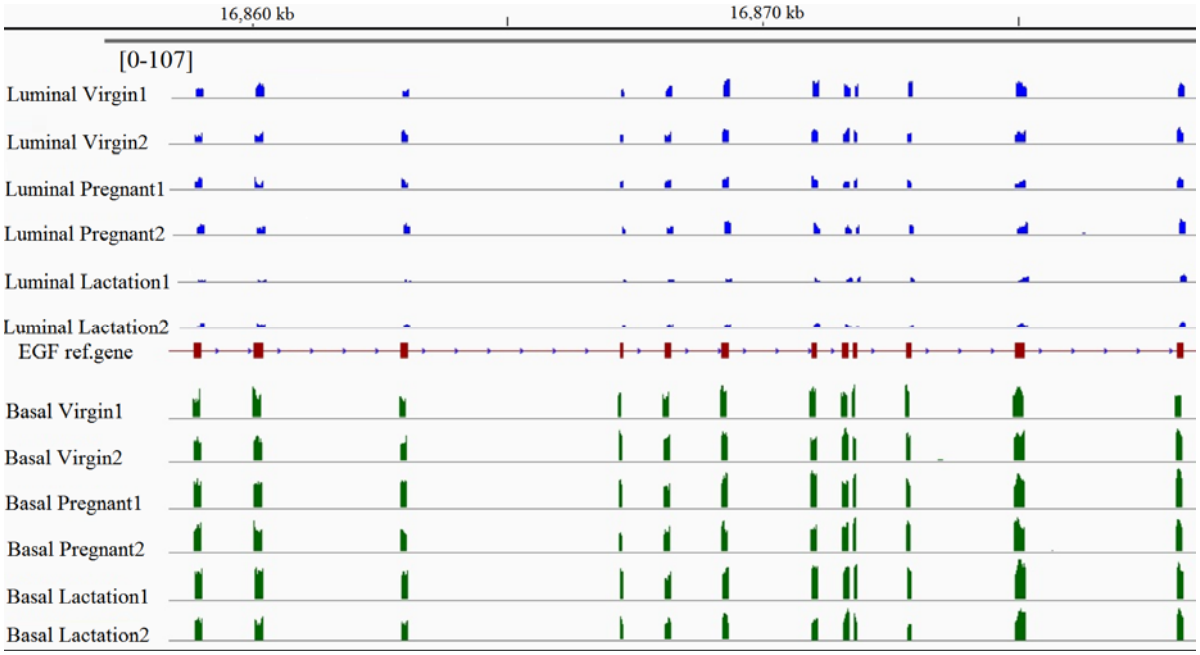

Figure 4. Coverage representation of aligned reads. IGV snapshot of EGF gene.

[21] and edgeR [26] are used for DE analysis based on gene annotations and are similar in performance, Ballgown [16, 27] is capable of analyzing both gene and transcript annotations. Both edgeR and DESeq2 packages take raw counts as input data [40]. Ballgown requires transcript assembly with StringTie to be performed, which stores results in *ctab* format as gene-, transcript-, exon- and junction-level expression measurements. All methods perform count normalization to the total number of reads (library size) to produce abundance estimates [16, 21, 26, 27].

Statistical analysis

Both DESeq2 and edgeR require at least three samples per treatment group, whereas it is suggested to use Ballgown with four or more replicates due to the linearity in the model-based analysis [16]. We chose to perform DE analysis of annotated genes without the need to identify novel isoforms. For this reason and the availability of three biological samples per variable the most suitable package for DE in our example was DESeq2. The raw counts matrix was extracted with the provided python script (http://ccb.jhu.edu/software/stringtie/dl/prepDE.py) and supplied and analyzed with DESeq2 [21] (Appendix). The choice of DESeq2 enabled us to do the analysis quickly, exemplifying that the choice of the program and algorithms can cater to the needs of the biological question and sample availability.

Visualization

Visualization is an important step in data analysis that allows for ease of data interpretation. There are many possible ways of presenting the results of expression data such as MA-plot, volcano plot, counts plot, or heatmap among others. *Clustering Image Map (CIM)* visualized as a heatmap shows differences and similarities across various conditions. Figure 4 shows the most differentially expressed genes of EGF receptor family.

Heatmap is a two-dimensional display of numbers as colors. Heatmap has rows that indicate genes, and columns that show different conditions or samples. Figure illustrates a heatmap of 11 pre-selected significantly expressed genes of EGF receptor family under three conditions (virgin vs pregnant vs lactation) for two cell types (luminal vs basal). The color key scheme (top left) shows the association of the numeric values to the color scheme. Here lowly expressed genes are shown in blue, and highly expressed genes are shown in red.

Most heatmap functions can perform hierarchical clustering of genes and samples, which are shown as a dendrogram on the left and top of the plot in Figure 5.

Here rows/genes that have similar gene expression patterns are grouped together, and in columns luminal and basal cells represent two separate clusters. Moreover, the pregnant and virgin conditions show similar pattern in gene expression as compared to lactation condition for luminal cell type. The fact that

gene_ID	name	baseMean	log2FC	lfcSE	stat	pvalue	padj
ENSMUSG00000020122	Egfr	2903.022	1.942	0.262	7.403	1.33E-13	2.13E-12
ENSMUSG00000028017	Egf	396.336	-4.816	0.953	-5.052	4.37E-07	2.95E-06
ENSMUSG00000082361	Btc	28.649	-3.978	0.594	-6.693	2.19E-11	2.71E-10
ENSMUSG00000029377	Ereg	618.594	6.173	0.508	12.154	5.49E-34	3.67E-32
ENSMUSG00000029999	Tgfa	314.596	2.198	0.298	7.382	1.56E-13	2.48E-12
ENSMUSG00000062991	Nrg1	2343.881	8.098	0.598	13.533	9.97E-42	9.15E-40
ENSMUSG00000060275	Nrg2	489.638	7.145	0.508	14.068	6.00E-45	6.27E-43
ENSMUSG00000029378	Areg	2044.000	-4.575	1.077	-4.249	2.15E-05	0.000106
ENSMUSG00000062312	Erbb2	971.805	-1.454	0.370	-3.924	8.70E-05	0.000376
ENSMUSG00000018166	Erbb3	2403.597	-3.993	0.414	-9.639	5.49E-22	1.83E-20
ENSMUSG00000062209	Erbb4	263.729	-4.984	0.323	-15.417	1.26E-53	2.02E-51

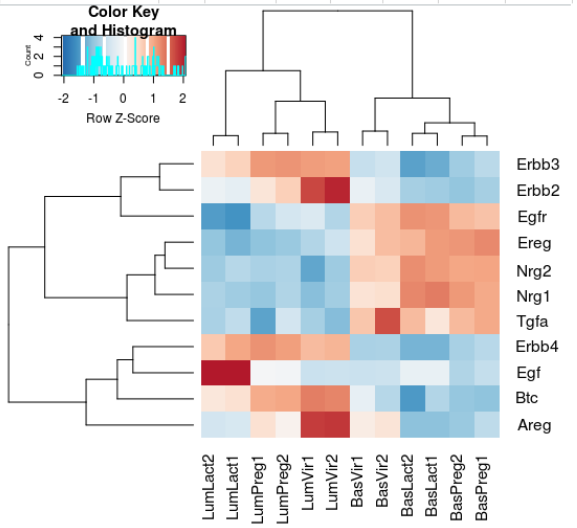

Figure 5. DESeq2 resulting table for EGF receptor family (top panel) and heatmap of the most differentially expressed genes of EGF receptor family (bottom panel). *Note: only three digits are shown after decimal point.

two replicates for each cell type/condition cluster together confirms good quality of our example data. Heatmaps allow one to employ different algorithms to clustering the data, which will effect the dendogram.

4. Presentation and Performance Analysis

Within this section, we describe a method for conducting a performance analysis for the standard DE analysis for RNA-Seq data. Table 2 illustrates the most popular tools in the domain, including the software versions that we have been applying as part of our DE analysis.

To evaluate the performance of the applied software tools, we utilized the performance counter tool (*perf*) [41]. Perf is a Linux based command line utility that provides performance information of the operating system, applications and hardware. *Perf stat* function is an excellent choice for software that can provide performance analysis.

Table 2. Software Tools Applied for a DE Analysis

Software Tool	Version	Ref
FastQC	v0.11.5	[18]
Cutadapt	v1.10	[19]
HISAT2	v2.0.4	[20]
DESeq2 (R)	v3.3.2	[21]
Perf (Linux)	v4.9.rc8.g810ac7b7	[41]

Table 3 illustrates performance parameters for FastQC, cutadapt, HISAT2 as *Instructions per Cycle (IPC)*, where results are represented as software and hardware related events. The task-clock, context-switches, cpu-migrations, page-faults are software related events, and the cycles, instructions are related to the hardware events. The task-clock shows the amount of time spend on the task. However, if there is parallel computing involved (for example, threads option in HISAT2) this number has to be devised on the CPUs involved. Context switch explains how many times the software switched of the CPU from one process/thread to another. CPU migration describes equality in a work load distribution across all cores. Finally, the page-faults occur when a program's virtual content have to be copied to the physical memory.

Our initial tests show, that HISAT2 (without threads option) is one of the most time-consuming steps (~25 min) in the analysis process, with the highest CPU usage (1.047 CPUs utilized). This can be explained due to the nature of the alignment process, where enormous amount of reads are required to be mapped the reference genome. Although HISAT2 utilizes Karkkainen's algorithm, which is time and space efficient, volume of the data that is required to be mapped to the reference genome is enormous and takes time to execute. One of the solutions to this problem is to apply threads (-p) option that enables to process the data in a parallel. We run HISTA2 with 18 threads which reduced the run time to 5 minutes. The number of cycles, which are an indicator for the number of instructions performed by software to produce final result is the highest for HISAT2 with thread parameter (~2.791 GHz), as expected, due to the fact that multiple parallel processes were executed.

Table 3. Perf Software Performance Analysis on the Example of FastQC, cutadapt, and HISAT2 utilizing the (SRR1552444) Data Set. Data is presented in Instructions per Cycle (IPC)

Performance Parameter	FastQC	Cutadapt	HISAT2	HISAT2 with thread –p 18	Measured
Software Events					
task-clock	3.8 (min) / 1.021 CPUs utilized	5.6 (min) / 0.995 CPUs utilized	25.9 (min) / 1.047 CPUs utilized	70.28 (min) / 14.534 CPUs utilized	
context-switches	0.094	0.002	0.752	4e-6	K/sec
cpu-migrations	0.006	0.000	0.001	0.011	K/sec
page-faults	0.302	0.019	3.6e-5	1.3e-5	K/sec
Hardware Events					
cycles	2.925	2.976	2.959	2.791	GHz
Total number of cycles	6.70122e+11	1.00997e+12	4.61275e+12	1.17684e+13	
instructions	1.14	1.27	0.91	0.48	insn per cycle
Total Run Time					
	3.7	5.7	24.8	4.8	minutes

6. CONCLUSIONS AND DISCUSSION

Within the scope of this paper we contributed with:

- evaluation of machine learning algorithms utilized for differential expression analysis of RNA-Seq data
- an example pipeline, which can be used to perform DE analysis (Figure 1)
- source code of a script-based pipeline

a. Advantages of Applying Machine Learning

The application of machine learning helps to analyze large volumes of data without loading it into the computer memory. The initial size of the SRR1552444.sra sample (GEO GSE60450 [14]) was 9.7 GB was narrowed to 1.1 MB of meaningful data, in other words we identified 7170 of significantly expressed genes with only 15 minutes software run time.

b. Source Code of the Script Based Workflow

Appendix A contains the source code and a practical guideline for people interested in conducting a similar kind of analysis. We provided and evaluated a script workflow of an analysis process to allow others repeating the experiment.

c. Generic Reference Pipeline and Workflow

To conduct the analysis, we applied publicly available datasets with the accession number GSE56933 [13]; and accession number GSE60450 [14] to describe the performance of software used in DE analysis of limited (up to three samples) number of samples.

The analysis process contains four steps:

- quality control
- adapter trimming
- alignment
- differential expression analysis

This is a standard workflow for DE analysis and will only vary due to selection of various algorithms for each step. For example, adapter trimming step may be omitted if the data has been already preprocessed in the past and was downloaded from GEO.

Here we show how to obtain data from publically accessible repositories can be used for analyses. One can also use such data to develop novel algorithms that can be benchmarked against existing ones without the need to produce more sequencing data. Algorithms described here provide efficient processing of big data but further improvements in speed, disk and RAM utilization are necessary to deal with larger and larger datasets.

d. Final Remarks

Big data coming from various omics platforms will only increase in size in the near future thus increasing the requirements for high performance analysis to gain understanding of the data. To process, visualize, and understand such omics big data we must apply machine learning algorithms. Interestingly, and depending on the complexity of the task, multiple algorithms are utilized and parametrized to improve performance. Within this paper, we gave an overview several widely used analysis

algorithms for NGS data, and indicated steps that assist in the analysis process.

Future work and development will require improved machine learning algorithms such as deep learning on distributed systems [42, 43]. In addition porting existing statistical methods from other omics platforms [44-48] as well as developing new ones [16, 28, 49] specific to NGS is necessary and will provide greater statistical certainty in the results.

In biomedical sciences, however, automatic machine learning approaches may suffer from insufficient number of training samples as a result of limited number of biological data sets, and in this instance *interactive Machine Learning (iML)* may be particularly useful [50, 51]. A grand challenge is to provide integrative machine learning approaches, i.e. the optimization of workflows and processes that are in-line with the main workflow of biomedical researchers, thereby increasing their capacity whilst reducing costs and improving efficiency [52, 53]. In this context usability gets a new meaning and an increasing importance, as experimental scientists often have limited skills in machine learning generally or in algorithms specifically – raising the need for multidisciplinary training the next generation of researchers in biology, bioinformatics and statistics [54].

Finally, bioinformatics software needs to be user-friendly and be accompanied by a comprehensive user manual. User-friendly, well documented software we provide to biologists will ease the discovery process in biological sciences to improve our understanding of diseases [55, 56] and transform our medical system into truly personalized medicine [57].

REFERENCES

[1] H. F. Lodish, Molecular cell biology, 4th ed. New York: W.H. Freeman, 2000, pp. xxxvi, 1084, G-17, I-36 p.

[2] C. Suzanne, "RNA splicing: introns, exons and spliceosome," Nature Education 2008.

[3] T. Stuart, S. R. Eichten, J. Cahn, Y. V. Karpievitch, J. O. Borevitz, and R. Lister, "Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation," Elife, vol. 5, Dec 02 2016.

[4] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, "The impact of next-generation sequencing on genomics," J Genet Genomics, vol. 38, no. 3, pp. 95-109, Mar 20 2011.

[5] H. P. Buermans and J. T. den Dunnen, "Next generation sequencing technology: Advances and applications," Biochim Biophys Acta, vol. 1842, no. 10, pp. 1932-1941, Oct 2014.

[6] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," Nat Rev Genet, vol. 13, no. 1, pp. 36-46, Nov 29 2011.

[7] B. Langmead. (2015). ADS1: Sequencing by Synthesis. Available: https://www.youtube.com/watch?v=IzXQVwWYFv4

[8] Y. Chu and D. R. Corey, "RNA sequencing: platform selection, experimental design, and data interpretation," Nucleic Acid Ther, vol. 22, no. 4, pp. 271-4, Aug 2012.

[9] C. Trapnell et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," Nat Protoc, vol. 7, no. 3, pp. 562-78, Mar 01 2012.

[10] A. Lugmayr, C. Scheib, and M. Mailaparampil, "Cognitive Big Data. Survey and Review on Big Data Research and its Implications: What is Really 'New'? Cognitive Big Data!," Journal of Knowledge Management (JMM), 2016.

[11] A. Lugmayr, C. Scheib, M. Mailaparampil, N. Mesia, and H. Ranta, "A Comprehensive Survey on Big Data Research and It's Implications - What is really 'new' in Big Data? It's Cognitive Big Data," presented at the Proceedings of the 20th Pacific-Asian Conference on Information Systems (PACIS 2016), 2016.

[12] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," Nat Rev Genet, vol. 16, no. 6, pp. 321-32, Jun 2015.

[13] A. Latorre-Pellicer et al., "Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing," Nature, vol. 535, no. 7613, pp. 561-5, Jul 28 2016.

[14] N. Y. Fu et al., "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival," Nat Cell Biol, vol. 17, no. 4, pp. 365-75, Apr 2015.

[15] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," Nat Rev Genet, vol. 10, no. 1, pp. 57-63, Jan 2009.

[16] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," Nat Protoc, vol. 11, no. 9, pp. 1650-67, Sep 2016.

[17] Y. Han, S. Gao, K. Muegge, W. Zhang, and B. Zhou, "Advanced Applications of RNA Sequencing and Challenges," Bioinform Biol Insights, vol. 9, no. Suppl 1, pp. 29-46, 2015.

[18] S. Andrews. A quality control tool for high throughput sequence data. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[19] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," EMBnet.journal, vol. volume 17, no. issue 1, pp. 10-12, 2011.

[20] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," Nat Methods, vol. 12, no. 4, pp. 357-60, Apr 2015.

[21] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome Biol, vol. 15, no. 12, p. 550, 2014.

[22] FASTX-Toolkit. Available: http://hannonlab.cshl.edu/fastx_toolkit/links.html

[23] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," Bioinformatics, vol. 30, no. 15, pp. 2114-20, Aug 01 2014.

[24] BBDuk. Available: http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/

[25] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, vol. 25, no. 14, pp. 1754-60, Jul 15 2009.

[26] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," Nat Methods, vol. 9, no. 4, pp. 357-9, Apr 2012.

[27] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," (in eng), Bioinformatics, vol. 26, no. 1, pp. 139-40, Jan 1 2010.

[28] A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek, "Flexible isoform-level differential expression analysis with Ballgown," bioRxiv, 2014.

[29] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," Nucleic Acids Res, vol. 38, no. 6, pp. 1767-71, Apr 2010.

[30] (2016). Illumina. Available: http://support.illumina.com/downloads/illumina-customer-sequence-letter.html

[31] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J Mol Biol, vol. 215, no. 3, pp. 403-10, Oct 05 1990.

[32] W. J. Kent, "BLAT--the BLAST-like alignment tool," Genome Res, vol. 12, no. 4, pp. 656-64, Apr 2002.

[33] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," Brief Bioinform, vol. 11, no. 5, pp. 473-83, Sep 2010.

[34] An Overview of the Human Genome Project. Available: https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

[35] M. Burrows and D. J. Wheeler, A Block-sorting Lossless Data Compression Algorithm (no. no. 124). Digital, Systems Research Center, 1994.

[36] G. M. P. Ferragina, "Opportunistic data structures with applications," presented at the Proceedings of the 41st Annual Symposium on Foundations of Computer Science, November 12 - 14, 2000.

[37] K. Juha, "Fast BWT in Small Space by Blockwise Suffix Sorting," (in English), Theoretical Computer Science, vol. 387, no. 3, pp. 249-257, November, 2007 2007.

[38] H. Li et al., "The Sequence Alignment/Map format and SAMtools," Bioinformatics, vol. 25, no. 16, pp. 2078-9, Aug 15 2009.

[39] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," Brief Bioinform, vol. 14, no. 2, pp. 178-92, Mar 2013.

[40] M. L. Speir et al., "The UCSC Genome Browser database: 2016 update," Nucleic

Acids Res, vol. 44, no. D1, pp. D717-25, Jan 04 2016.

[41] E. G. Stephane Eranian, Tipp Moseley, Willem de Bruijn. Tutorial. Available: https://perf.wiki.kernel.org/index.php/Tutorial

[42] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle, "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning," Genome Biol, vol. 18, no. 1, p. 67, Apr 11 2017.

[43] Y. V. Karpievitch and J. S. Almeida, "mGrid: a load-balanced distributed computing environment for the remote execution of the user-defined Matlab code," BMC Bioinformatics, vol. 7, p. 139, Mar 15 2006.

[44] Y. V. Karpievitch et al., "Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition," Bioinformatics, vol. 25, no. 19, pp. 2573-80, Oct 01 2009.

[45] Y. Karpievitch et al., "A statistical framework for protein quantitation in bottom-up MS-based proteomics," Bioinformatics, vol. 25, no. 16, pp. 2028-34, Aug 15 2009.

[46] Y. V. Karpievitch et al., "PrepMS: TOF MS data graphical preprocessing tool," Bioinformatics, vol. 23, no. 2, pp. 264-5, Jan 15 2007.

[47] T. Taverner et al., "DanteR: an extensible R-based tool for quantitative analysis of -omics data," Bioinformatics, vol. 28, no. 18, pp. 2404-6, Sep 15 2012.

[48] Y. V. Karpievitch, S. B. Nikolic, R. Wilson, J. E. Sharman, and L. M. Edwards, "Metabolomics data normalization with EigenMS," PLoS One, vol. 9, no. 12, p. e116221, 2014.

[49] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, "Normalization of RNA-seq data using factor analysis of control genes or samples," Nat Biotechnol, vol. 32, no. 9, pp. 896-902, Sep 2014.

[50] A. Holzinger, "Machine Learning for Health Informatics," vol. 9605, pp. 1-24, 2016.

[51] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," Brain Inform, vol. 3, no. 2, pp. 119-131, Jun 2016.

[52] A. Holzinger and I. Jurisica, "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions," vol. 8401, pp. 1-18, 2014.

[53] A. Holzinger, Biomedical Informatics: Computational Sciences meets Life Sciences. Norderstedt: BoD, 2012.

[54] A. Holzinger, M. Errath, G. Searle, B. Thurnher, and W. Slany, "From Extreme Programming and Usability Engineering to Extreme Usability in Software Engineering Education (XP+UE→XU)," vol. 2, pp. 169-172, 2005.

[55] S. B. Nikolic et al., "Serum metabolic profile predicts adverse central haemodynamics in patients with type 2 diabetes mellitus," Acta Diabetol, vol. 53, no. 3, pp. 367-75, Jun 2016.

[56] V. P. Andreev et al., "Label-free quantitative LC-MS proteomics of Alzheimer's disease and normally aged human brains," J Proteome Res, vol. 11, no. 6, pp. 3053-67, Jun 01 2012.

[57] N. Liang, C. A. Trujillo, P. D. Negraes, A. R. Muotri, C. Lameu, and H. Ulrich, "Stem cell contributions to neurological disease modeling and personalized medicine," Prog Neuropsychopharmacol Biol Psychiatry, May 30 2017.

[58] C. Helf and H. Hlavacs, “Apps for life change: Critical review and solution directions,” *Entertainment Computing*, vol. 14, pp. 17–22, 2016.

[59] Pogorelc, B. et al. 2012. Semantic ambient media: From ambient advertising to ambient-assisted living. *Multimedia Tools and Applications*. 58, 2 (2012), 399–425.

APPENDIX: SOURCE CODE AND SCRIPTS TO CONDUCT A SIMILAR KIND OF ANALYSIS

Herein we describe step by step procedure used to achieve final results of the DE analysis to make the analysis process available to others. The Unix shell is used to run majority of commands of described protocol, Python, and R. One sample is taken to show how to use these tools. We tried to present each step in a very precise way by avoiding one line complex commands

ANALYSIS PROCESS GUIDELINES FOR SOFTWARE INSTALLATION

NCBI SRA Toolkit:
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software
FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Catadapt: http://cutadapt.readthedocs.io/en/stable/installation.html
HISAT2: https://ccb.jhu.edu/software/hisat2/manual.shtml
Samtools: http://www.htslib.org/download/
StringTie: https://ccb.jhu.edu/software/stringtie/#install
Python (prepDE.py): http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual
R (Deseq2): source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")

ANALYSIS PROCEDURE

STEP 1: RAW DATA

Note: the first two steps are performed for GSE56933 dataset based on one sample (SRR1257444).

Download *sra* files. Unix shell is used to download the samples data from NCBI Geo database. The ftp directory of each sample should be supplied to *wget*. The samples will be saved to the directory from which the command is run. Run following command to download sample:

```
$ wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByExp/sra/SRX/SRX681/SRX681985/SRR1552444/ SRR1552444.sra
```

The next step is to convert *sra* to *fastq* file format. This can be done with *sra-toolkit* (see software installation section):

```
$ path_to_sratoolkit/fastq-dump –gzip –split-3 SRR1257444.fastq
```

STEP 2: QUALITY CONTROL

Quality control with *FastQC*:

```
$ fastqc -o SRR1257444.fastq.gz
```

Unzip file:

```
$ gunzip SRR1257444.fastq.gz
```

STEP 3: ADAPTER TRIMMING

Adapter trimming with *cutadapt*. Use only the prefix of the adapter sequence (TruSeq Index):

```
$ cutadapt -a  GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -o SRR1257444.fastq
adapt_tr_SRR1257444.fastq &> stat_SRR1257444.log
```

STEP 4: ALIGNMENT

Note: following steps are performed for data set from GSE60450, sample (SRR1552444). Map reads to the mouse reference genome:

```
$ hisat2 -p 18 --dta -x index/mouse_genome -U SRR1552444.fastq -S
aligned_SRR1552444.sam &> stat_SRR1552444.log
```

Sort and convert the *SAM* to *BAM*:

```
$ samtools sort -@ 18 -o SRR1552444.bam SRR1552444.sam
```

Generate *BAM* index file:

```
$ samtools index SRR1552444.bam
```

Assemble transcripts.

```
$ stringtie -p 18 -e -B -G musmusc.gtf -o outfolder/SRR1552444/SRR1552444.gtf -l
SRR1552444 SRR1552444.bam
```

Note: Important the *outfolder* holds all sample output from *stringtie.* For example, for the next sample the *stringtie* command line looks like:

```
$ stringtie -p 18 -e -B -G musmusc.gtf -o outfolder/SRR1552445/SRR1552445.gtf -l
SRR1552445 SRR1552445.bam
```

PYTHON

Follow an alternative differential expression workflow. Download *prepDE.py* from http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual. Note that *prepDE.py* script aims to extract raw count matrixes for gene-level measurements and transcript-level measurements. The goal of this paper is to show differentially expression analysis at gene-level, thus we use only *gene_level_raw_counts.csv* output file. The script should be saved to the same directory as the folder that holds *stringtie* output (same path as *outfolder*)

STEP 5: DIFFERENTIAL EXPRESSION ANALYSIS

R

Load *DESeq2* package:

```
$ R
R version 3.3.2 (2016-10-31)
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
library(DESeq2)
```

Set directory to a file. Load gene-level count matrix:

```
setwd(“path to gene_level_raw_counts.csv”)
file_count = read.csv("gene_level_raw_counts.csv",row.names=1)
countData = as.matrix(file_count)
```

Create phenotype data (multi-compariosn):

```
ctype=factor(c(rep('Luminal',6),rep('Basal',6)))
condition=factor(c(rep('Virg',2), rep('Pregn',2), rep('Lact',2)))
coldata <- data.frame(row.names=colnames(countData),ctype,condition)
```

Construct DESeq object:

```
dds = DESeqDataSetFromMatrix(countData = countData, colData = coldata, design =
~condition + ctype)
```

Filter law abundance genes:

```
dds = dds[ rowSums(counts(dds))>1,]
```

Differential expression analysis based on the Negative Binomial distribution:

```
dds = DESeq(dds)
res = results(dds, contrast=c("ctype","Basal","Luminal"))
```

Get significantly expressed genes (cut-off < 0.01)

```
signif = res[res$padj<0.01 & !is.na(res$padj),]
write.csv(as.data.frame(signif),file="DESeq2_signif_two_group_gene_level_001.csv")
```

TOWARDS A SUSTAINABLE DESIGN FOR MATURITY MEASUREMENT MARKETPLACE

Lester Lasrado [1,2]
[1]Centre for Business Data Analytics,
Copenhagen Business School
[2] Networked Business Initiative
lal.itm@cbs.dk

Ravi Vatrapu[1,2]
[1]Centre for Business Data Analytics,
Copenhagen Business School
[2]Westerdals Oslo ACT, Oslo, Norway,
vatrapu@cbs.dk

Henrik Bjerre Kærsgaard
Networked Business Initiative,
hkh@networkedbusiness.org
networkedbusiness.org

Jan Futtrup Kjaer
Networked Business Initiative,
jfk@networkedbusiness.org
networkedbusiness.org

ABSTRACT

In this research-in-progress paper, we propose a solution in form of an IT artefact to address both theoretical and practical challenges faced by maturity model designers. We identify and list out the existing challenges & criticisms of maturity models research through an extensive literature review, followed by semi-structured interviews with four maturity model designers. We also explore different motivations of building a maturity model, and using them further scope the boundaries of our solution.

KEYWORDS

Maturity model, Benchmarking, Design science

1. INTRODUCTION

The debate about rigor and relevance continues to exist in the Information Systems (IS) field ever since its inception [9, 55] and maturity model research is no different. Prior research has identified lack of applying scientific re-search methods in a rigorous manner and has called upon IS researchers to not to create elements of maturity models only from prior normative studies, but also validate them empirically. Considering the multitude of maturity models to increase, both academic and consultancy, researchers using self-assessment surveys would definitely face humongous practical challenges in producing empirically founded and validated maturity models. Recent literature in IS [52, 40, 44, 6, 32] have identified some future trends in the domain of maturity models research, especially the increasing academic and practitioner interests in maturity models [8] across multiple domains like business process management [49], e-government [44], and few others wherein the levels of maturity are well established. [40] rightly questions if this high quantity of maturity model literature translates to high quality. Interestingly the trend is stronger in the development of new maturity models of emerging technologies, also called as entities like web/social media [3, 20], Analytics [15], cloud [51], wherein the levels of maturity can be very uncertain and deemed speculative by the academic audience. [32] questions the maturity of such an entity under maturation, while many others [29, 52] questioning the empirical evidence behind these maturity models as well. In line with this we ask our first research question: *What are the current challenges of IS researchers designing maturity models? How can they be addressed?*

In the process of doing so, we reviewed prior literature on maturity models research, examined models by consultancies and interviewed four designers. During this process observed that maturity models, especially in case of emerging technologies is a super-set of the benchmarking [10] concept. Secondly we also discovered that IS literature on maturity model design has not covered some of the most practical challenges that a design product would face like competition for attention, limited exposure to targeted audience and lack of holistic ecosystem thinking by designers, thus risking the model of remaining unused and deemed irrelevant by practitioners. In this paper, we argue that maturity models developed by researchers with a purpose of benchmarking organizations should not only be designed for rigor, but also making the maturity model attractive and accessible for practitioners use. We further propose a solution in form of an IT artefact. Accordingly, we adopt design science approach for

design and evaluation. We restrict the scope of this artefact to researchers and consultants building a maturity model for a highly innovative and emerging phenomenon, wherein the dominant design and best practises are still being understood. In line with this our second research question is as follows: *What are the design principles of IT artefact that can successfully address some of the challenges?*

The rest of the paper is organised as follows. In section 2 we briefly explain the research method, with section 3 focussing on identifying the motivations to design and use maturity models. The next section highlights the challenges that have not been debated in prior literature. In section 5 we propose design requirements for the IT artefact, discuss the results obtained till now and finally in section 6 state our future research agenda.

2. RESEARCH METHOD

Write Figure 1 shows the research steps taken to undertake this task.

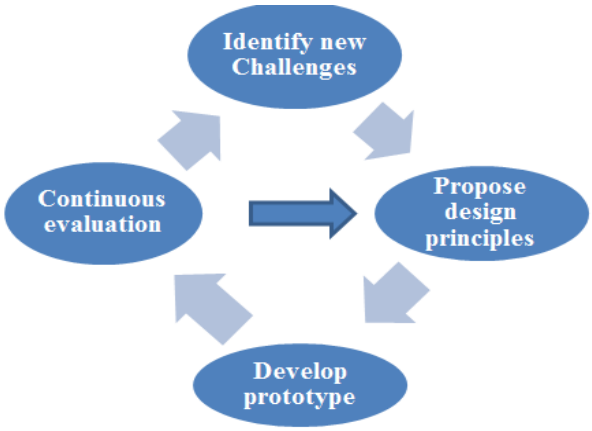

Figure 1. Four iterative research steps.

It comprises of 4 basic iterative steps, starting with identifying the challenges & criticisms of maturity (section 2). This was done through an extensive literature review of maturity models literature in IS, studying practitioner reports and semi-structured interviews with four maturity model designers i.e. two each from academia and consulting respectively (table 1). The next step was to propose design principles to overcome these challenges and develop a working IT artefact, which would be tested by involving the necessary stakeholders (section 3). This artefact is refined iteratively while being validated and improved continuously. The whole process is cyclic and iterative following the design science guidelines.

3. MOTIVATION TO DESIGN AND USE

Recent articles on maturity models by [52], [39] have stated that IS researchers have primarily focused on developing new maturity models, with a majority of them published as conceptual models without an instrument for measurement, thus questioning the relevance of maturity models to practice. Recent papers [29, 52] concluded that there are three identities of maturity models depending on the purpose of use and motivation behind its development.

- The identity portrays them as normative theories e.g. EDP [35],Intranet [14] that are predominantly grounded as process theories.
- The identity portrays them as “best practice guide” or “certification mechanism”, especially post the success of Capability maturity model [36, 37]. E.g. Business process maturity [49, 17], Healthcare analytics certification [23], etc.
- The third and final identity portrays a maturity model as a practical benchmarking tool, wherein organizations are classified and compared against each other using a scale of low to high maturity.

Within the given scope, we found that in majority of the maturity models an entity under maturation could be (1) benchmarked against a pre-defined standard or best practice e.g. e-Government [2, 31] or (2) be subjected to quantitative benchmarking against other organizations, e.g. BI maturity [41], social business maturity [18], etc. With maturity models representing stage-based evolution theories [40] among research community, in practice its application is diverse and lately benchmarking stands very high on the agenda.

In order to understand this better, we thoroughly investigated 12 maturity models with regards to motivation of designers and users as shown in table 1. With benchmarking as a subset we classify 9 out of the 12 models under this category (B, M). Taking inspiration from the categorization scheme [1], we classify maturity models into 6 groups:

Group 1 are models developed mainly by researchers through their own research, are mostly motivated from theoretical aspect, and the model may or may not have been implemented and validated through real life applications.

Group 2 are models developed mainly by academics but as part of an engaged scholarship project [48]. Models are implemented and validated through real life applications.

Table 1. Motivation of Designers & Users.

		Motivations to Design							Motivations of users				
G	SET	Designers	K	B	C	M	L	R	Users (Audience)	B	C	M	I
1	B,M	ITSM self- survey platform* [54]	©	Ᵽ					IS Researchers & IT Practitioners;				©
	~B,M	Social media maturity [25]	©						IS research community. No value for practitioners.				©
2	B,M	E- Government maturity* [2]	Ᵽ	©	Ᵽ				Danish Government organizations & IS research community.	Ᵽ	©	Ᵽ	
	B,M	Process Management Maturity [12]	Ᵽ	©					Hospital Management in Switzerland & the IS research community.	Ᵽ	©	Ᵽ	
3	B,M	Online Analytics Maturity Model* [21]	Ᵽ			©	Ᵽ		Free Online tool for everyone interested in analytics with no clear audience.				©
	B,M	Omni-channel Maturity* [24]		Ᵽ		©	Ᵽ	Ᵽ	Free Online tool for everyone interested in Omni-channel marketing				©
4	B,M	Social Business Maturity [18]	Ᵽ		©	Ᵽ			Decision makers (C suite, Department heads, IT managers).			Ᵽ	©
	B,M	Digital Maturity [53]	Ᵽ		©	Ᵽ			Decision makers (C suite, Department heads, IT managers).			Ᵽ	©
5	B,M	Customer Experience Maturity [38]				©	Ᵽ	Ᵽ	Free Online tool for everyone interested in online marketing.				©
	B,M	Adobe Analytics Maturity [5]				©	Ᵽ		Free Online tool for everyone interested in analytics with no clear audience.				©
6	~B,M	Capability Maturity Model (CMM) [36]	Ᵽ	Ᵽ	©			Ᵽ	Comprehensive tool built for software companies	Ᵽ	©		
	~B,M	CMM Integration [13]	Ᵽ	Ᵽ	Ᵽ			©	Extension of CMM for software companies for certification.	©	Ᵽ	Ᵽ	

© - Core or main motivation; **Ᵽ** - Peripheral or other motivations; **K** – Contribute to knowledge; **B** – Benchmarking (Internal & External); **C** – Drive Change or be a change agent; **M** – Marketing & brand value; **L** – Generating future leads; **R**– Generate revenue; **I** – Other Intrinsic motivations. *Indicates discussion/interview with the main author/designer that respective maturity model.

Group 3 is models developed by mostly consultants from personal opinion and judgment through experience in providing consultancy to organizations. These models may be from an individual capacity, sometimes funded by consultancies which may or may not be embarking on a real project.

Group 4 is models developed by large management consultancies in collaboration with renowned academic and research institutions. From the models analysed by us, we found that driving change or being a change agent was the core motivation of such high profile collaboration. Ideally these models are easily accessible through the internet and are read widely by the practitioners.

Group 5 is models developed by IT vendors, whose main business is selling IT products and services.

Group 6 consists of consortium driven projects and are usually very well planned and executed. They involve consortium of industries, the government and some large educational institutions.

In the model analysed by us CMM [36, 13] has clearly moved from driving change to a full-fledged certification industry generating revenue (R). We also classified some users as those looking for some certification and benchmarking (B) in order to drive an agenda of change (C) within an organisation. Above all, what we found most interesting was the participation of the users'; especially with group 1 & 3 was due to intrinsic motivation (I) – e.g. curiosity, fun, learning, helping a researcher, personal favour or something similar. Moreover, we found that benchmarking in form of a working IT artefact would be a requirement in order to make maturity models relevant especially for group 1, 2 & 3.

4. CHALLENGES AND CRITICISMS

Now that we established the motives, we look towards answering our first research question by identifying the main challenges of maturity models (Table 2) and then classify them into two groups. The first group consists of theoretical and design challenges that have been debated at length in prior IS literature, enough though solutions to solve many of these challenges are still satisfactory.

The theoretical challenges deals mostly with the dilemma on the identity of maturity models and this has debated for last 40 years from [26] to [44]. However, this debate is not purpose of this paper, hence we take a stand that maturity model is well accepted tool, both relevant to practice and research community and move

forward. The design challenges deals with the maturity model design, both the design process [7, 16, 44] as well as design principles for maturity model as a design product [40, 32]. This too has been discussed in detail over the last few years and listed in table 2.

Table 2. Challenges in MM research.

	Type of Challenges	
Already debated extensively	**Theoretical Challenges**	
	Lack of theoretical foundations with models adopting the design structure from Nolan and Gibson [35]and CMMI [13]	[44]
	One size fits all approach & non-acknowledgement of equifinality. Minimal evidence to prove improvements in maturity corresponds to higher benefits.	[27] [34]
	Design Challenges	
	Lack of empirical validation in selection of dimensions, predominantly conceptual.	[28] [31]
	Use of easy to measure, shallow & incomplete measures. Ambiguous interpretation of benchmarking – Explaining the final purpose of use a challenge. Unable to deal with variety, context and continuously changing environment.	[7] [33] [31]
	Developed in Isolation- Lack practical implications. Need for dashboards or similar IT artefact for comparison among respondents.	[54]
Not debated in IS	**Competition for attention**	
	In case of self-assessment, surveys are used as an instrument for benchmarking. The challenge of low response rates & survey fatigue is a huge challenge.	[54], [43] [30]
	Consulting firms are considered to be a central actor in the management fashion arena. Considering maturity measurement as one such fashion, competition for practitioner attention is a challenge for IS researchers.	[8]
	Too many generic maturity models both in the consulting and research world. Need to develop practical advice on selection of maturity models.	e.g. [50]
	Limited exposure to relevant context	
	Lack relationship with the intended audience, no follow up with audience. No reach to relevant practitioner audience, just conceptual models.	[25], [46]
	The time factor has been completed ignored by IS researchers	[20], [19]
	Lack of ecosystem thinking	
	Level of respondent's readiness to participate not looked at. No promotion thus no accessibility & applicability - does not reach practice.	[54], [9], [42]
	Service & support costs of maintaining the maturity instrument to keep the models from being outdated and relevant needs adoption and use in practise.	[10], [31], [13]

It is known that empirically founded maturity models are rare [29]. In order for it to be a common reality, there is need industry participation during the building, testing and validation stage. We use the classic "chicken and egg" analogy here i.e. empirical data for a tool like maturity model would require practitioner participation throughout the process. However except for [32] and [16] none of the procedure models acknowledge the fact it is important to involve stakeholders throughout the process of design and thereafter. We see this as a big research gap and classify them in the second group of challenges (table 2) as follows:

Competition for attention – This addresses the surge in the increasing number of models with fancy reports and artefacts measuring maturity that are easily available via a simple "google search", thus grabbing the time and attention of the practitioners, moulding their opinions before the researcher even decides to reach them.

Limited exposure to relevant context –The time taken by researchers to publish results took around 2 to 3 years' time as compared to their consulting counterparts. E.g. Social media business profile maturity assessment for Irish SME'S took 3 years from the initial conceptual model [20] to the assessment results [19]. The social business maturity assessment [18] on the other hand has published assessment for three years consecutively in the same period.

Lack of ecosystem thinking – Over and above the prior challenges discussed, the value proposition of participation in the whole maturity assessment exercise is not communicated by the researchers to relevant stakeholder throughout the process of development. Except in the case of few engaged scholarship projects, no evidence is seen that effort was put in communicating results beyond academic publications. Moreover, in case of self-assessment using surveys, none of the academic maturity models published look at respondent readiness to participate again.

Addressing the above three challenges would be of utmost importance for IS researchers in order to make the maturity model empirical founded and sustainable. Therefore, in line with the goals of this paper, the next section proposes a solution and subsequently evaluates a novel IT artefact to address these challenges.

5. PROPOSED SOLUTION

In our quest to find a solution for this we align our thinking through design theory [47] and thus consider a kernel theory to guide our solution i.e. theory of platform business [45] to address the lack of ecosystem thinking, limited explore to relevant context and address completion for attention. We further propose a set of design principles, develop and test our IT artefact called "maturity measurement marketplace". We claim that

this IT artefact would be suitable for maturity model designers of group 1, group 2 and group 3. The development of this IT artefact follows a design science approach as it gives a "methodological frame for creating and evaluating innovative IT artefacts" [22, 54]. We formulated the design requirements (req) for our artefact that we plan to adhere to during the entire design process.

Relevance - Req1: Our design should attract independent maturity model designers from group 1, 2 or 3 who would like to use self-assessment survey technique for benchmarking maturity of their targeted audience. Req2: The artefact should also attract and motivate actors in organisations (e.g. managers, CEO's, etc.) to use the artefact and drive change.

Rigor - Req3: The design should have a mechanism to filter and select only relevant actors i.e. both the designers and users. This exercise must involve rigorous analysis on the actor's ability to keep the marketplace credible and relevant as the same time. Req4: Most importantly, the data privacy of users (respondents) must be respected and their assessments protected from misuse at all stages.

Usability - Req5: The artefact should be easy to use and understand, navigation must easy and multi-language support must be provided.

Generalizability – Req6: The maturity model selection and implementation process should be easily mutable across the maturity models and must allow for easy reuse and replicability. Req7: Most importantly, the users (respondents) must be able to navigate and benchmark their maturity across the maturity models hosted on the IT artefact.

Addressing the design requirements (Req1, 2 and 3), we first propose the conceptual model of the artefact as shown in figure 1. In order to satisfy rigor and relevance requirements, the conceptual model has three filters: (i) **Credibility filter** - is the screening process of which maturity model would be allowed. (ii) Intermediary (P) – usually an actor that already has or intends to develop a working relationship with the user. (iii) Catalyst & validity filter – Intermediary acts both as a catalyst by promoting the maturity model and validates the user responses given their business relationship.

Who are these intermediaries? - In our design we consider management & digital consultancies, industry associations and IT Vendors (P). We have three strong reasons for doing so

1. All the three actors are interested in maturity model research.

2. Currently in case of self-assessment tools, there is no mechanism to check if the responses coming in are from respondents actually working for real organisations. E.g. online analytics maturity [21]. The presence of intermediaries would not only ease the process of reaching relevant users, but also make the process of collecting the empirical evidence more rigorous and verifiable.

3. The results from survey research rarely reaches the respondents, even if it does it is in an aggregated level after a long time. Winkler bridges this gap by providing immediate results and feedback to the respondents through their self-survey platform. We go a step further, by not only providing immediate and detailed results, but also an opportunity for the intermediary to get in touch with the users to interpret and study the results obtained.

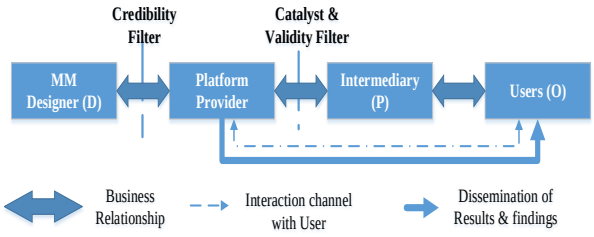

Figure 2. Conceptual design for IT artefact (Maturity measurement marketplace).

The design requirement 3 calls for attracting managers from organisations to adopt the tool. Prior literature on management fashion [4, 30] has positioned management consultants and industry associations are influencers who usually push new knowledge and drive change in the industry. We use this knowledge and in accordance with Req2, design the IT artefact in a way that the relationship of the platform with the end user will always be through the intermediary; however, the user can provide feedback to the IT artefact via an online feature satisfying Req5. This relationship among the actors in the IT artefact is very similar to the well-known supply chain concept of buyer-supplier-supplier triad relationship [11]. As of date, conceptual design for the IT artefact has been adopted and the artefact implemented, with the results and evaluations in table 3.

In addition to feedback via the IT artefact, interviews and talk out loud studies are being conducted with professionals to improve the visualizations, communication and major concerns. One such concern

among users is data privacy. In spite of taking all precautions and continuously evolving our data privacy rules to suit the needs of the users, it still is one of major concerns from recent evaluations. We believe that satisfying Req5, Req6 and Req7 and improving our communication would address the user concerns.

Table 3. Initial results after implementation.

<table>
<tr><th></th><th colspan="2">Digital Maturity score</th><th>Omni-channel Maturity [24]</th><th>DI Productivity index</th></tr>
<tr><td>D</td><td colspan="2">Academics & Consultants</td><td>Consultants</td><td>Association</td></tr>
<tr><td>P</td><td colspan="4">Partners that signed up till now. 90% are consultancies, Associations (2) and vendors (3)</td></tr>
<tr><td rowspan="2">O</td><td colspan="4">Over 900 sign-ups, while the users completing the process has been around 50% (roughly 430)</td></tr>
<tr><td colspan="2">900 (360)*</td><td>700 (350)**</td><td>50 (27)</td></tr>
<tr><td>Status</td><td colspan="2">3 iterations done (17 testers & 44 respondents)
Average time spent on the artefact reduced to 25 minutes from initial 1 hour</td><td>2 iterations done. New visualisation maturity implemented. Collecting user feedback.</td><td>No iterations.</td></tr>
<tr><td>O</td><td rowspan="2">Evaluations</td><td colspan="2">Well done and professional service (+)
Dashboard with survey results looks good, (-)
Sign up and Survey takes too much time (-)
What will you do with the data (-)</td><td>No Feedback yet</td></tr>
<tr><td>P</td><td>How can I interpret these results (-)</td><td>Still analysing feedback collected</td><td>No Feedback yet</td></tr>
</table>

*Total users starting (completed). **There are number of common users for both digital maturity and Omni channel maturity. D – MM Designer. P- Intermediary (Consultancy/Association –One having a working relationship with end user). O – End user(s)

The IT artefact is being developed over last 24 months while taking continuous inputs from intermediaries and users in the process. This evolution of the design and change in features would continue based on future evaluations and we clearly see us moving towards a multisided platform. However, this discussion is not within the scope of this research in progress paper and thus we state our future research plans in the next section.

6. FUTURE RESEARCH AGENDA

Considering the multitude of maturity models to increase [40], both academic and consultancy, researchers using self-assessment survey's would definitely face humongous practical challenges in producing empirically founded and validated maturity models. The completion of this research will produce design guidelines along with a working IT artefact to address the challenges faced by individual maturity model designers working on an emerging phenomenon (group 1, 2, and 3 in table 1). First, we anticipate that successful implementation and adoption of this IT artefact (Maturity measurement marketplace) would provide a set of design principles to build empirically founded maturity models. We also anticipate bridging the gap between academia and industry with regards to maturity models research. We scope our contribution carefully, stating that the above contributions would be relevant to those developing maturity models for the purpose of benchmarking in a domain that is still emerging.

Our initial intention was make a completed research paper, but we believe that the validation of such a tool is not complete with only developing the IT artefact and testing three maturity models for such a short period. Moreover, the maturity models themselves are still undergoing their own validation. Our future research agenda is therefore to validate if the maturity model designers have actually benefited from this IT artefact. From the current numbers the acceptance is visible, but the actual success of a maturity model is proved if it brings about a discussion on improvement among the targeted audience and this would take at least next year or so. During this period, we will host few other models as experiments and evaluate the impact on practise. We will also test the hypothesis that availability of empirical data during development process would produce more rigorous models as compared to theoretically and conceptually grounded models.

7. REFERENCES

[1] G. Anand And R. Kodali, "Benchmarking The Benchmarking Models", Benchmarking: An International Journal, 15 (2008), Pp. 257-291.

[2] K. V. Andersen And H. Z. Henriksen, "E-Government Maturity Models: Extension Of The Layne And Lee Model ", Government Information Quarterly, 23 (2006), Pp. 236-248.

[3] A. Back, Christopher, "Assessing Degrees Of Web-2.0-Ness For Websites: Model And Results For Product Websites In The Pharmaceutical Industry", Bled 2011 Proceedings. Paper 48 (2011).

[4] R. L. Baskerville And M. D. Myers, "Fashion Waves In Information Systems Research And Practice", Management Information Systems Quarterly, 33 (2009), Pp. 647-662.
[5] J. Bates, Is Your Company Healthy? Diagnose With The Data Analytics Maturity Model | Adobe, @Adobemktgcloud, Http://Myanalyticsscore.Com/, 2014.
[6] J. Becker, R. Knackstedt And J. Poeppelbuss, "Developing Maturity Models For It Management – A Procedure Model And Its Application", Business & Information Systems Engineering, 1 (2011), Pp. 213-222.
[7] J. Becker, R. Knackstedt And J. Pöppelbuß, "Developing Maturity Models For It Management", Business & Information Systems Engineering, 1 (2009), Pp. 213-222.
[8] J. Becker, B. Niehaves, J. Poeppelbuss And A. Simons, Maturity Models In Is Research, European Conference On Information Systems (Ecis) 2010 Proceedings, Paper 42, 2010.
[9] I. Benbasat And R. W. Zmud, "Empirical Research In Information Systems: The Practice Of Relevance", Mis Quarterly (1999), Pp. 3-16.
[10] R. C. Camp, Benchmarking: The Search For Industry Best Practices That Lead To Superior Performance, Benchmarking: The Search For Industry Best Practices That Lead To Superior Performance, Asqc/Quality Resources, 1989.
[11] T. Y. Choi And Z. Wu, "Triads In Supply Networks: Theorizing Buyer–Supplier–Supplier Relationships", Journal Of Supply Chain Management, 45 (2009), Pp. 8-25.
[12] A. K. Cleven, R. Winter, F. Wortmann And T. Mettler, "Process Management In Hospitals: An Empirically Grounded Maturity Model", Business Research, 7 (2014), Pp. 191-216.
[13] P. T. Cmmi Cmmi For Development, Version 1.3 (Cmu/Sei-2010-Tr-033). Software Engineering Institute, Carnegie Mellon University, 2010. Http://Resources.Sei.Cmu.Edu/Library/Asset-View.Cfm?Assetid=9661 (Retrieved On 16th Novemeber 2014), Software Engineering Institute, 2010.
[14] J. Damsgaard And R. Scheepers, "Managing The Crises In Intranet Implementation: A Stage Model", Information Systems Journal, 10 (1999), Pp. 131-149.
[15] T. Davenport, H And J. Harris, G., Competing On Analytics: The New Science Of Winning, Harvard Business Press, 2007.
[16] T. De Bruin, R. Freeze, U. Kaulkarni, M. Rosemann, B. Campbell, J. Underwood And D. Bunker, Understanding The Main Phases Of Developing A Maturity Assessment Model, Australasian Chapter Of The Association For Information Systems, 2005.
[17] T. De Bruin, Michaelbartmann, Drajola, Fkallinikos, Javison, Dwinter, Rein-Dor, Pbecker, Jbodendorf, Fweinhardt, C, "Towards A Business Process Management Maturity Model", Australasian Chapter Of The Association For Information Systems (2005).
[18] Delloitte And Mit-Sloan, "Mit Smr's Social Business Interactive Tool 2014", (2015).
[19] A. Duane And P. O'reilly, Social Media Adoption: Stages Of Growth, Paths Of Evolution And Dominant Problems, Proceedings Of The 2nd European Conference On Social Media 2015: Ecsm 2015, Academic Conferences Limited, 2015, Pp. 130.
[20] A. Duane And P. Oreilly, A Conceptual Stages Of Growth Model For Managing An Organization's Social Media Business Profile (Smbp), International Conference On Information Systems (Icis) 2012 Proceedings, 2012.
[21] S. Hamel, The Web Analytics Maturity Model, A Strategic Approach Based On Business Maturity And Critical Success Factors. Last Consulted January, Https://Digitalanalyticsmaturity.Org/, 2009, Pp. 2012.
[22] A. R. Hevner, S. T. March, J. Park And S. Ram, "Design Science In Information Systems Research", Mis Quarterly, 28 (2004), Pp. 75-105.
[23] Himss-Analytics, Delta Powered - Analytics Maturity Suite, Http://App.Himssanalytics.Org/Emram/Delta.Aspx#Dpaa, 2015.
[24] R. Houlind, Hvis Det Handler Om Mig, Sã¥ Kã¸Ber Jeg!, Omnichannel Institute, Http://Omnichannelmarketing.Dk/, 2015.
[25] H. Karkkainen, J. Jussila And J. Lyytikkä, Towards Maturity Modeling Approach For Social Media Adoption In Innovation, 4th Ispim Innovation Symposium, Wellington, New Zealand, 2011.
[26] J. L. King And K. L. Kraemer, "Evolution And Organizational Information Systems: An Assessment Of Nolan's Stage Model", Commun. Acm, 27 (1984), Pp. 466-475.
[27] W. King And T. Teo, "Integration Between Business Planning And Information Systems Planning: Validating A Stage Hypothesis", Decision Sciences, 28 (1997), Pp. 279-308.
[28] G. Lahrmann, F. Marx, T. Mettler, R. Winter And F. Wortmann, Inductive Design Of Maturity Models: Applying The Rasch Algorithm For Design Science Research, Service-Oriented Perspectives In Design

Science Research, Springer Berlin Heidelberg, 2011, Pp. 176-191.
[29] L. A. Lasrado, R. Vatrapu And K. N. Andersen, Maturity Models Development In Is Research: A Literature Review, Iris Selected Papers Of The Information Systems Research Seminar In Scandinavia 2015.Paper 6, 2015.
[30] D. Madsen And K. Slåtten, "The Role Of The Management Fashion Arena In The Cross-National Diffusion Of Management Concepts: The Case Of The Balanced Scorecard In The Scandinavian Countries", Administrative Sciences, 3 (2013), Pp. 110-142.
[31] D. Maheshwari And M. Janssen, "Measurement And Benchmarking Foundations: Providing Support To Organizations In Their Development And Growth Using Dashboards", Government Information Quarterly, 30 (2013), Pp. S83-S93.
[32] T. Mettler, "A Design Science Research Perspective On Maturity Models In Information Systems", Universität St. Gallen, St. Gallen, Switzerland, Technical Report Be Iwi/Hne/03 (2009).
[33] T. Mettler And P. Rohner, Situational Maturity Models As Instrumental Artifacts For Organizational Design, Proceedings Of The 4th International Conference On Design Science Research In Information Systems And Technology, Acm, 2009, Pp. 22.
[34] M. P. W. Mullaly, "If Maturity Is The Answer, Then Exactly What Was The Question?", International Journal Of Managing Projects In Business, 7 (2014), Pp. 169-185.
[35] R. L. Nolan And C. F. Gibson, Managing The Four Stages Of Edp Growth, Harvard Business Review January–February 1974, 1974.
[36] M. Paulk, B. Curtis, M. Chrissis And C. Weber, "Capability Maturity Model, Version 1.1", Ieee Software, 10 (1993), Pp. 18-27.
[37] M. C. Paulk, "Surviving The Quagmire Of Process Models, Integrated Models, And Standards", (2004).
[38] L. B. Petersen, R. Person And C. Nash, Connect: How To Use Data And Experience Marketing To Create Lifetime Customers, John Wiley & Sons, 2014.
[39] J. Pöppelbuß, B. Niehaves, A. Simons And J. Becker, "Maturity Models In Information Systems Research: Literature Search And Analysis", Communications Of The Association For Information Systems, 29 (2011), Pp. 505-532.
[40] J. Pöppelbuß And M. Röglinger, What Makes A Useful Maturity Model? A Framework Of General Design Principles For Maturity Models And Its Demonstration In Business Process Management, Ecis 2011 Proceedings.Paper 28, 2011.
[41] D. Raber, F. Wortmann And R. Winter, "Towards The Measurement Of Business Intelligence Maturity", Ecis 2013 Completed Research. Paper 95. (2013).
[42] M. Rosemann And I. Vessey, "Toward Improving The Relevance Of Information Systems Research To Practice: The Role Of Applicability Checks", Mis Quarterly (2008), Pp. 1-22.
[43] T. Schoenherr, L. M. Ellram And W. L. Tate, "A Note On The Use Of Survey Research Firms To Enable Empirical Data Collection", Journal Of Business Logistics, 36 (2015), Pp. 288-300.
[44] H. Solli-Sæther And P. Gottschalk, "The Modeling Process For Stage Models", Journal Of Organizational Computing And Electronic Commerce, 20 (2010), Pp. 279–293.
[45] K. S. Staykova And J. Damsgaard, A Typology Of Multi-Sided Platforms: The Core And The Periphery, Ecis 2015 Completed Research Papers. Paper 174., 2015.
[46] Z. Thompson And P. Booth, Social Media Within Business: Furthering The Maturity Model, Proceedings Of The 2nd European Conference On Social Media 2015: Ecsm 2015, Academic Conferences Limited, 2015, Pp. 445.
[47] V. Vaishnavi And W. Kuechler, ""Design Science Research In Information Systems". January 20, 2004, Last Updated October 23rd, 2013", Url: Http://Www. Desrist. Org/Design-Researchin-Information-Systems (2004).
[48] A. H. Van De Ven, Engaged Scholarship: A Guide For Organizational And Social Research, Oup Oxford, 2007.
[49] A. Van Looy, An Experiment For Measuring Business Process Maturity With Different Maturity Models, Twenty-Third European Conference On Information Systems (Ecis 2015), Ais Electronic Library (Aisel), 2015, Pp. 1-12.
[50] A. B. Van Looy, Manupoels, Geertsnoeck, Monique, "Choosing The Right Business Process Maturity Model", Information & Management, 50 (2013), Pp. 466-488.
[51] D. Weiss, Jonaszarnekow, Rüdigerand Schroedl, Holger, "Towards A Consumer Cloud Computing Maturity Model - Proposition Of Development Guidelines, Maturity Domains And Maturity Levels", Pacis 2013 Proceedings. Paper 211 (2013).
[52] R. Wendler, "The Maturity Of Maturity Model Research: A Systematic Mapping Study", Information And Software Technology, 54 (2012), Pp. 1317-1339.
[53] G. Westerman, M. Tannou, D. Bonnet, P. Ferraris And A. Mcafee, Digital Transformation - The Digital

Advantage: How Digital Leaders Outperform Their Peers In Every Industry, 2012.
[54] T. Winkler, J. Wulf And W. Brenner, "Selfsurvey: A Prediction-Based Decision Support Platform For Survey Research", Twenty-Third European Conference On Information Systems (Ecis), Münster, Germany, 2015 (2015).
[55] S. Xie, M. Helfert, A. Lugmayr, R. Heimgärtner And A. Holzinger, Influence Of Organizational Culture And Communication On The Successful Implementation Of Information Technology In Hospitals, International Conference On Cross-Cultural Design, Springer, 2013, Pp. 165-174.

Visualization as a Big Data Artefact for Knowledge Interpretation of Digital Petroleum Ecosystems

Dr. Shastri L. Nimmagadda

School of Information Systems
Curtin Business School

Curtin University

Kent Street, Bentley, Perth,
Western Australia 6102

shastri.nimmagadda@curtin.edu.au

Dr. Amit Rudra

School of Information Systems
Curtin Business School

Curtin University

Kent Street, Bentley, Perth,
Western Australia 6102

Amit.rudra@cbs.curtin.edu.au

ABSTRACT

In the current upstream business environment, we examine the risk involved in the petroleum exploration and field development. Many sedimentary basins worldwide possess hundreds of petroleum systems with thousands of oil and gas fields, geographically scattered. A significant amount of unstructured heterogeneous and multidimensional data are locked up in many industrial applications and knowledge domains. Our objective is to bring the relevant data together, integrate and visualize for adding values to the existing interpretation. We simulate a Big Data guided digital petroleum ecosystem (DPE) approach, a digital oil field solution, a new direction in the analysis of a total petroleum system (TPS), in which multiple sedimentary basins may have been grouped, inheriting an interconnectivity between the systems. The DPE is articulated in a framework, organizing variety of data associated with the elements and processes of complex petroleum systems and integrating their data dimensions and attributes. We develop an ontology based data warehousing and mining artefacts. We present warehoused metadata, with slicing and dicing of data views for visualization of new prospects in the investigating area. We further investigate the risk of exploratory drilling campaigns and how the integrated framework, with visualization and interpretation artefacts can holistically support the delivery of high-quality products and services.

KEYWORDS

Digital Petroleum Ecosystem; Big Data; Data Visualization; Interpretation; Knowledge Discovery.

1. INTRODUCTION

Data visualization [13] is the study of the visual representation and graphical depiction of data, meaning thereby, the information is explored and processed in different schematic forms, with variables in different units. Data fusion is a sort of visualization, in which we describe the data instances in different graphic visuals, in a way the knowledge extracted from the metadata is interpretable. The goal of data visualization is an information delivery effectively through graphical means. The designers [13] focus on visuals, and their functionality ensuring new insights of information, especially in sparse and complex data areas in a more intuitive way. Without losing the clarity and perception, we obtain the knowledge through visual representations and graphic displays of geological and geophysical (G & G) and exploration and production (E & P) knowledge. The knowledge-based data structures [2] are of ontology focus. For transmission of data views geographically, data are structured in XML codes [2]. As described in [11, 4], various visualization techniques motivate us interpreting multiple datasets including various association mining rules. An oil and gas exploration with an application of data visualization technique is given in [3]. The visualization of frontier areas of petroleum prospects is described in Australian sedimentary basin [6] contexts with 4-D seismic technologies. We examine various issues of exploitation of reservoirs, structural and strati-structural plays under different geological settings [3]. The risks of exploratory drilling, and under-explored areas are given in [1]. Aside from brief discussions on data structuring of the resources data, there is a limited literature available on ontology application in oil and gas industries, especially the implementation of DPE in the oil and gas industry. In this context, we introduce the Big Data paradigm

though as a hype [7], but emerging as a think-tank in the context of DPE in which the volume and variety characteristics play a dominant role in particular the unstructured data representation more explicitly through visualization and interpretation artefacts.

2. RESEARCH OBJECTIVES

We aim at analysing the digital ecosystems focusing the following data visualization and interpretation objectives:

1. How are the correlations, trends and patterns of data views extracted from metadata structures visualized and presented for interpretation?
2. How to present and visualize the explored data for knowledge discovery, extracting new value of information.
3. Whether the visualized data views extracted from metadata structures, can be interpreted using the existing interpretation procedures?

To share common understanding of the data structure among entities: It is one of the common goals in developing ontologies. Several websites contain geographically varying volumes of oil and gas data and information. If these websites share and publish similar ontological descriptions of the data entities, in a way the computer agents can extract and aggregate information more visually. The agents can use the aggregated information to answer user queries or as input data in other applications.

Models from several domains need to represent the notation in space and time: This representation includes the notions of time-intervals, points in time, relative measures of time, and so on. If one group of researchers develops such ontology in detail, others can reuse it in their domains and contexts. As an example, the domain knowledge acquired from a particular model of a particular conventional oil and gas field, may be reused in the same ecosystem for an unconventional field. Additionally, if a large ontology description is needed to be built, several existing ontologies can be integrated describing the portions in the larger domain.

The domain assumptions are made more explicit: Make necessary changes as per the interpretation and implementation and just in case the knowledge about the domain changes. Explicit specifications of domain knowledge are useful for researchers and investors who must learn what terms or entities/dimensions in the domain mean to earth science systems.

Existing known operational knowledge among entities or dimensions is separated from the undiscovered knowledge among the emerging conceptualized and contextualized entities or dimensions. To generate the domain knowledge, we use the following visualization and interpretation methodologies.

3. DATA VISUALIZATION METHOD

In this section, we use several visual tools, providing an effective means of communication, with highly developed 2D and 3D pattern-recognition capabilities that allow processing and perceiving the pictorial digital data efficiently. We summarize the data, highlighting the visualization trends. Unknown phenomena are uncovered through various kinds of graphical representation. Several visualization techniques use volumes and varieties of Big Data including spatial-temporal multidimensional datasets that exist in the DPE contexts. We exploit the following visualization methods:

One method of visualizing the results of data search is through scatter- plot display in a three-dimensional grid. A scatter plot is a visualization technique that shows each data point as a color sphere, or bubble. Scatter plots are referred to as bubbles in these visualizations. The size, shape and color of the bubble, each is used to represent a variable in the data. The three axes in the visualization ideally should be able to describe any dimension within the same data plot and should be able to be randomly selected and modified by the user. The user can choose to explore more fully those data points using visual methods or perhaps click on the data and see a traditional numerical display in a spreadsheet. The volume of data points can also be rotated to observe clusters in different areas. By preparing and presenting the data graphically, the user can uncover properties of the data quickly and easily detecting any patterns or deviations from expected results.

Bubble charts are other ways of presenting data, because they convert pages of hard-to-understand numerical and textual data into something that is easily comprehensible to analyze. Bubble plot is a simple example of the use of graphics to quickly convey information about the data what they describe. This bubble plot, representing bubble sizes, densities and trends, suggests several inferences such as structure, reservoir and production attributes, their strengths and magnitudes that can bring out the qualitative and quantitative properties of reservoirs.

The visual data are typically read in as a 3D block of data, called a volume. Workstations do read the data in chunks because of voluminous and variety of data. Each point of data in the selected area represents a physical location (i.e., an X, Y, Z position) in the 3D space

represented by that particular field. The value at each data point represents many attributes or properties (such as amplitude, phase, frequency, velocity). A collection of surrounding values in a given area is in turn identified (to a certain degree of probability) as the type of entity (or object) or a dimension at each geographic location. A value is assigned to each data point (or more typically to a range of data points) corresponding to a color attribute to display in that range of values or instances. All the points that fall within the selected range of values have the same color. In a 3D representation of the object, the colors smooth out to form layers. We use similar visuals [4, 8] for discovering numeric association rules among the structured resources business data. These rules assign various color attributes for visualization.

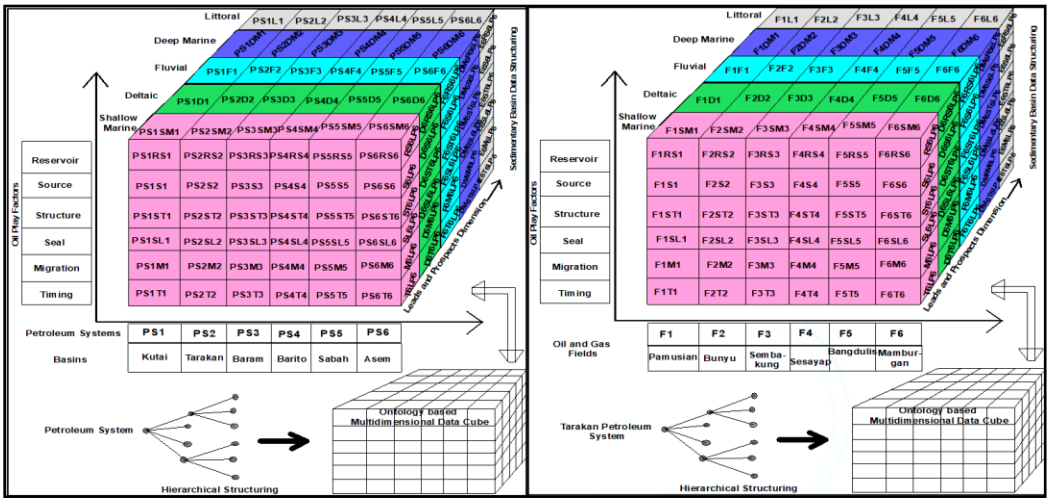

Figure 1: 3D Cube multidimensional representation of matured fields with reference to the elements of a petroleum system

We focus on data visualization, keeping in view its increasing demand in presenting the data views in various graphic visuals. The visualization explicitly facilitates the domain-specific data interpretation from fine-grained metadata volumes. Visualization is a display of multiple views, such as map, plot, chart, bubble plot views and how multiple dimensions and all the data patterns, trends and correlations are visualized in a single plot more explicitly. OLAP visualization is a presentation of metadata views from warehoused G & G and E & P data sources. As shown in Figure 1, we display both periodic and geographic dimensional data views for visual interpretation. The cloud computing systems (synonymous to network of systems in the present context) accommodating various data organizations distribute the data views and deliver products and services of good quality visuals to the data interpreters and oil and gas explorers. We use the visualization workflow as given in Figure 2 to process the graphic visuals for business analytics, delivering quality and interpretable information to variety of users. Business and data analytics and presentation of processed data in desired visuals are significant criteria. The new knowledge convey the information that needed in identifying the oil and gas prospective locales, but broadly in the DPE and TPS scales.

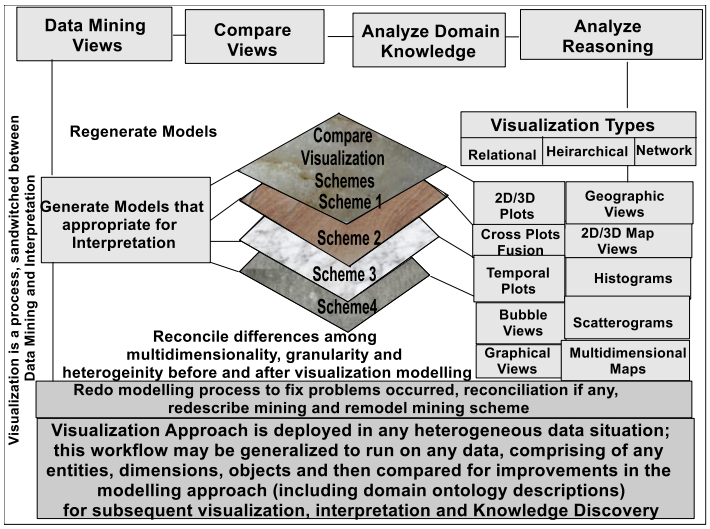

Figure 2: Visualization modelling – workflow

Information represented in the form of graphics, audio and video, makes use of spatial-temporal data and other multidimensional datasets from warehouse repositories. Visualization modelling has an impact in any application domain [8, 10] for which creative thinking, product ideation, and advanced business analytics are envisioned. Questions such as "what" to explain the "why" of engineering graphics, are incorporated within the design of visualization models. During data visualization process, data views, extracted from the warehoused metadata ensure with users' requirements so as to interpret them for knowledge discovery. As described in Figure 2, various visualization schemes are attempted to best fit the current interpretation and knowledge.

4. DATA INTERPRETATION METHODOLOGY

As a part of implementation in the present study, we propose several interpretation methodologies. We interpret the extracted data-views for evaluating the effectiveness of integrated framework and data models designed in different knowledge domains and contexts. The Cognitive Big Data as introduced in [7] is a new data interpretative research with several baseline scenarios and contexts. The rationale behind this approach is to bring together and integrate different contexts. Data analysis is crucial, in addition to testing the validity of data models, data warehousing and mining and the effectiveness of visualization. Qualitatively, the trends, patterns and correlations observed among various data events are interpreted in the knowledge enhancement domain. Besides, we describe the relevance, effectiveness, efficiency, impacts and sustainability criteria. Extent and duration of usage of data models and integrated framework including implementation of contextual, short- and long-term research outcomes among latitude and longitude dimensions are other interpretation objectives.

Data analysis and interpretation are meant for transforming the processed data into critical knowledge guaranteeing the research outcomes for descriptive analysis. The measure, consistency and effectiveness of multidimensional and heterogeneous data organization, modeling, mapping, data mining including effectiveness of data visualization are the other tests. Interpretation is either qualitative and or quantitative and the data patterns, trends and correlations interpreted, lead to the discovery of knowledge, implementing it in multiple domains. Interpretation made for evaluating the data models follows the criteria:

Relevance: methodologies, models, data mining, visualization are relevant to support the analysis in different application domains.
Effectiveness: achieved the research objectives.
Efficiency: within the available resources, to achieve maximum targets and goals.
Impacts: there is an immense impact in the implementation of data models in various application domains.
Sustainability: the scope of models, methodologies and implementation in other domains
Impact based evaluations are refined based on the use of our models in multiple domains and applications with the criteria:

1. ***Extent of use*** – how many stakeholders identified this approach and what degree outcome of research findings used

2. ***Duration and extent of usage*** – will the models, methodologies and implementations continue to be in multiple dimensions, such as geographical and periodic; to use in multiple countries and historical periods

a. Interpretation of Cross-sectional and Longitudinal Data Dimensions

Interpretative research is part of design science information science (IS) research. Data views represented in multidimensional views provide insights of interpretation and anticipated domain knowledge. Multidimensional metadata and their data views are interpretative in systems analysis and development scenarios. Big data in geographically spread countries and historical periods are significant in testing the current data models. Knowledge is built based on both short and long-term outcomes for interpretations. It is good idea evaluating the implementations of short and long term outcomes separately, so that we experience a fair assessment, examination of time-frame and resources needed for the projects and sustained impacts at different stages.

b. Contextual Implementation

Interpretation of results or implementation outcomes are possible in proper contexts, which include what outcomes are expected from current implementations, based on similar implementations that may have been made in previous visualization models.

c. Knowledge Modelling

Based on the domain application, interpretation objectives are chosen. But in the present context, we narrate methodologies as described in Figures 3 and 4. Knowledge is built based on the analysis and interpretations. Data mining rules focus on interpretation of attributes. The anomalies are deviations from the common rules or standardized or expected values. Interpretation and qualitative analysis of anomalies are the basis of building knowledge, such as attitude attributes of petroleum systems' elements and or processes. We perform the quantitative analysis by measuring the thicknesses of reservoirs and the areal extents of structures.

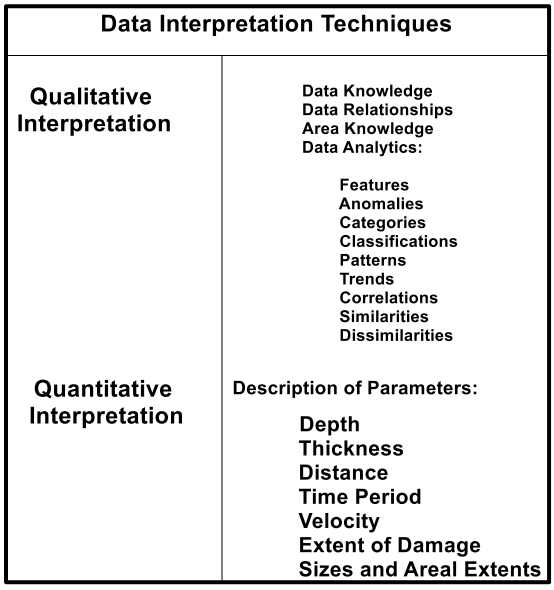

Figure 3: Interpretation methodologies

Data views extracted for visualization and interpretation are examined for anomalies and their evaluation for corroborating a particular model that achieved the knowledge presentation objectives. For this purpose different interpretation approaches are adapted for evaluating the qualitative and quantitative anomalies for knowledge discovery. We attempt several interpretation schemes for achieving the set of objectives of interpretation as narrated in Figures 3 and 4. Focusing on the number and type of property attributes, we reconcile the differences before and after the interpretation modelling, the best scheme in the current domain application. The domain ontologies and their modelling help facilitating both the visualization and interpretation schemes that ultimately provide new knowledge interpretation of prospective areas.

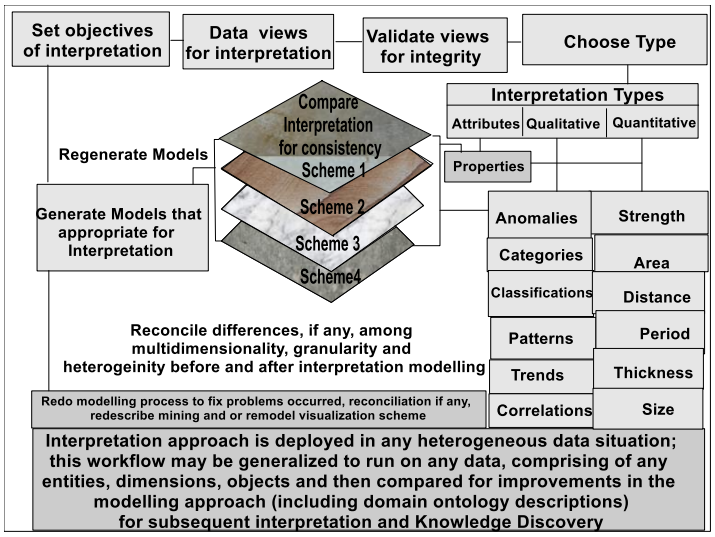

Figure 4: Data interpretation modelling – a workflow

We iterate the advantage of ontologies for representing the knowledge and modelling the domain knowledge. Ontology supports storage and manipulation of knowledge, including drawing inferences and making decisions. Mechanism of generalization and specialization including classification facilitate the semantics and fine tuning of knowledge representation. Selected data views consist of interesting patterns and trends, which may be descriptive and or predictive. The data attributes are either qualitative or quantitative. Attributes that depict the spatially and periodically varying properties are used for interpreting the data inferences in different knowledge domains and contexts. Bid Data as a cognitive approach support the inferences. We use data views for extracting domain knowledge for interpretation in support of knowledge-based systems.

As demonstrated in Figure 4, various data views are derived from metadata. Knowledge obtained in all case studies, is ensured with meaningful interpretations and implementation of metadata in different application domains. For example, an element of a petroleum system is found to be more productive and its areal extents discovered is large enough, such that similar strength of attributes is predicted in other fields of associated systems. The creation and discovery of knowledge play a decisive role on increased availability of knowledge from a system and effectiveness of the knowledge in its associated systems. Knowledge acquired in a system has an impact in perceiving knowledge of other related (or associated) domains, for which the data models described for mining, visualization and interpretation are effective.

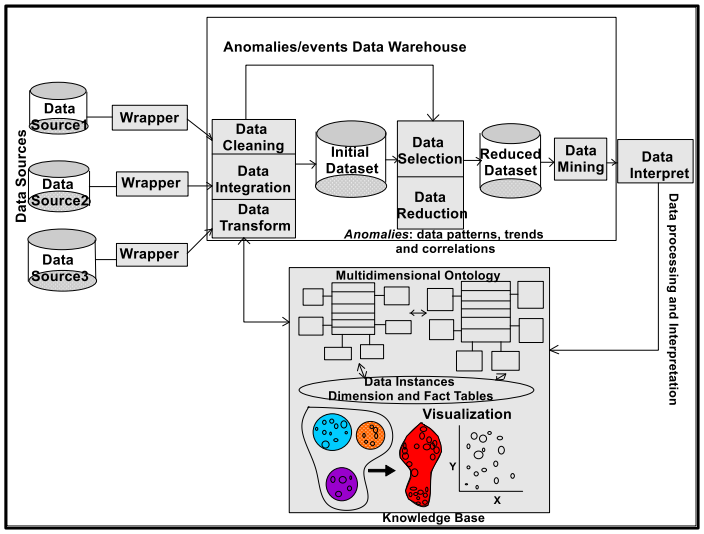

Figure 5: Knowledge building process model

A generalized knowledge process model as given in Figure 5 depicts a workflow from modelling of data sources to interpretation and then implementation of knowledge. Data exploration, prospecting, appraisal and development stages produce enormous amount of knowledge at multiple levels of systems' investigation and analysis, each level adding information for knowledge, interpretation and its analysis.

The current literature is not sufficient enough to interpret domain knowledge and its use in oil and gas industries. Though operating and service companies have their own proprietary interactive interpretation software, the interpretation methodologies are either developed based on user needs or problem solutions. But still, qualitative and quantitative interpretation methodologies are popularly used in many domains including oil and gas domains. Data views extracted from warehoused metadata, are validated for interpretation and the type of interpretation is chosen for its consistency and knowledge discovery. Data are either qualitatively or quantitatively interpreted based on the objectives of interpretation and project goals. Models computed from statistical mining are used for interpreting their properties, more often for qualitative interpretation. Implementation of the data models in different knowledge domains is carried out, addressing the issues of heterogeneity. Metadata that represent demographic, geographic and periodic data instances, is interpreted to have a meaningful information for decision support systems that assist in understanding system's behaviour, further analyse for future improvements. Merits and demerits of the systems and performances are analysed. Measures, strengths and anomalies of the properties of the attributed dimensions are analysed using different models and methodologies for interpretation. We analyse the domain knowledge and its limitations in the digital ecosystems.

5. MAPPING THE DOMAIN KNOWLEDGE

The warehoused metadata [5] are explored for implementing their data views in the strategic knowledge management. In the oil and gas industries, in particular exploration entities, "domain knowledge" is commonly used during interpretation of elements, processes and chains of petroleum systems and their ecosystems. Translating the existing knowledge with new knowledge, using the new attribute dimensions interpreted in multiple domains is a key focus. The data views extracted from warehoused metadata generate several prospective locales in the investigating area. As described in Figures 6a and 6b, several data attributes and their instances of geological structure, seismic data attributes are interpreted to identify new prospective areas. Structural highs, thicker isochrones- packs and their attribute visualizations when superimposed each other, can provide better areas for detailed drilling campaigns. The acoustic impedance attribute visualizations can contribute to better interpretation.

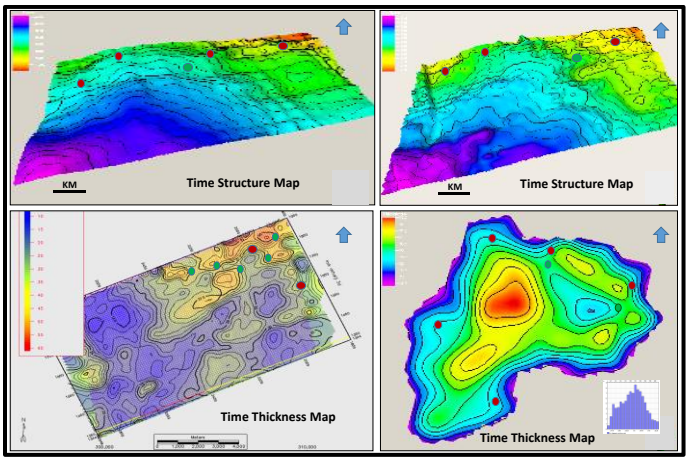

Figure 6a: Visualization of geological structure data-views

Use of "domain knowledge" in the context of petroleum systems, and oil and gas industries is uncommon and implicitly understood in spite of that several domain applications are involved in the upstream business environments. Though literature is available in this context in public domain, it is highly commercialized.

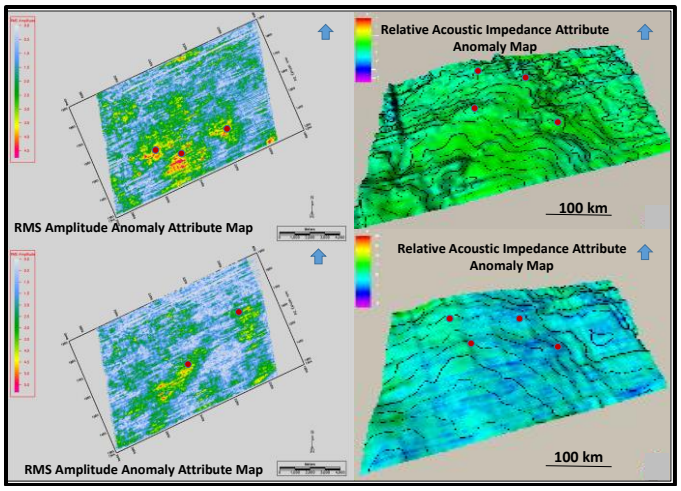

Figure 6b: Visualization of well-driven seismic attribute data-views

6. DIGITAL ECOSYSTEMS

The research objectives are further examined, exploring the scope of developing the visualization models for digital ecosystems of the other petroleum bearing basins worldwide. As demonstrated in [11, 12], several sedimentary basins described in Australia, India, Indonesia, Uganda, Middle East and USA are considered in the current research, keeping in view the intricacies and complexities of data sources in these regions. Our focus is on conventional, unconventional and matured oil and gas fields of these basins, exploring a scope of analysing huge volumes of heterogeneous and multidimensional data and their visualizations.

Australian sedimentary basins: There is an immense scope of integrating and analysing many sedimentary basins of Australian onshore and offshore basins, especially Western Australia, which produces more oil and gas deposits than elsewhere in Australia. Sedimentary basins possess heterogeneous and multidimensional data sources as shown in digital form [11]. Integrated framework and workflows explore the use of exploration and production datasets for risk minimizing the oil and gas business in Australia. Super Westralian basin, a total petroleum system (TPS), such as North West Shelf (NWS) possesses the shelf, slope and deep geological events, which appear to have a connectivity through phenomena of a digital ecosystem. Besides, this super basin has a multitude of sub-basins, each basin is associated with multiple petroleum systems, and each system with unknown or limited areal extents. Each petroleum system contains various oil and gas fields, with the hierarchical structuring of data dimensions and their associated attributes. North West Shelf in the Western Australia possesses unstructured heterogeneous and multidimensional data [11], we explore the scope of analysing their dimensions, attributes and instances in a warehousing environment with data mining and visualization opportunities exploiting untapped reservoirs of these basins.

Arab Gulf basins: Petroleum digital ecosystems and their embedded systems are described in the context of Arabian Gulf Basins (Middle Eastern onshore and offshore basins, [11], demonstrating the necessity of ontology structuring in the integrated workflows and their implementations. There is scope of specification of conceptualization and contextualization analysis in modelling and integrating multidimensional and heterogeneous data sources in Arab Gulf basins. Ecology, petroleum system and geomorphic systems cannot be isolated, which are otherwise are embedded, demonstrating an ecosystem, with multiple systems' connections. In this context, an integrated methodology proposed in Gulf basins enables us to understand the ecosystem phenomena through interconnected multiple digital ecosystems and their visualizations.

Indonesian basins: Indonesia is an island country with more than 30000 islands, with scattered sedimentary basins [9], in huge geographic regions and areal extents. They inherit volumes of multidimensional and heterogeneous data sources. Indonesian petroleum ontology (PO) descriptions can make good use of integrating oil and gas data sources of Indonesian sedimentary basins in different knowledge and application domains. These descriptions facilitate digital oil field solutions and visualizations in these regions.

***East African Rift System (EARS)*:** Several data sources exist within East African rifted basins. As a part of demonstrating the concept of digital petroleum Ecosystem (PDE), we identify ontology-based data warehousing associated with multiple petroleum systems, in the context of Albertine Graben (located in the Western Uganda, [10, 11]. Albertine Graben is considered as super basin with sub type basins. Each sub-basin consists of multiple petroleum systems and each is associated with multiple oil and gas fields. The super basin concept is simulated as an ecosystem, in which all petroleum systems and their embedded oil and gas fields have a connectivity. Volumes and varieties of data available in these basins facilitate the demonstration of emerging petroleum ontologies (PO) in the rift systems. The emerging visualizations have a further scope of analysing a large number of productive basins starting from southern Sudan in the north to the Malawi rift in the south-eastern parts of Uganda.

Unconventional Energy Scenarios in the USA: Several shale gas projects are developed [10] in the southern parts of USA to meet the demand energy resource. Because of growing demand of energy sources and the steady depletion of the current conventional oil and gas resources, there is an increasing quest for unconventional oil and gas business. Shale gas occurs within fractured shale reservoirs. Data sources associated with fractured shales do exist in many company situations, but unsuitable for an integrated framework because of the heterogeneity and multidimensionality. Problems associated with drilling and production, especially in the fracture development areas may be resolved with fracture visualizations, their associated subsurface lithologies (geological sense) and their connectivity.

Oil & Gas Deposits in the Indian Sub-continent: In the context of Indian sub-continent [11], there is immense scope of analysing different sedimentary basins for risk minimizing the exploration & production tasks in onshore and offshore regions. Data mining and visualization are used for building knowledge and effectively managing geographically scattered petroleum systems. The Cambay Basin, KG Basins, Cauvery Basin, Bombay Offshore Basins and several onshore basins of the North Eastern India, where several matured fields reported wealth of the data. These petroleum provinces and their linked data are suitable in designing and implementing the visualization modelling and knowledge interpretation methodologies.

7. CONCLUSIONS

Based on the visualization and interpretation of various data views of the PDE and TPS of multiple sedimentary basins, we have made the following conclusions:

1. Data visualization is a successful and widely used technology for viewing the resources data hidden under great depths. Data mining is an iterative process, implying that each time it refines the resultant data, a better visualization attribute is observed. This technology supports the Big Data cognitive approach.
2. Interpretation models drawn in the present studies are useful for the resources industries in terms of predicting the drillable exploratory locales.
3. It is recommended to use the data warehousing and mining technologies together with visualization and interpretation artefacts that supported by Big Data.
4. The map views are valuable tools for interpretation and implementation in the upstream business.
5. SQL and classical statistical mining are quite useful for mining and visualizing the multidimensional data, including geographical and periodic dimensions.
6. Heterogeneous data sources located in government and private enterprises, national and multinational companies can successfully be used in building multidimensional models.
7. Petroleum digital ecosystems (PDE) and petroleum information systems are digital oil field solutions in Big Data scale for various producing companies and multinational service companies.
8. There is immense scope of extending the current research application in worldwide sedimentary basins.

REFERENCES

[1] Castañeda G. O. J., Nimmagadda, S.L, Cardona Mora, A. P, Lobo, A, and Darke, K. (2012), On Integrated Quantitative Interpretative Workflows for interpreting structural and combinational traps for risk minimizing the exploratory and field development plans, *Bolivarian Geophysical Symposium* proceedings, held in Cartagena, Colombia.

[2] Erdman, M. and Studer, R. (2001), Heterogeneous Information Resources Need Semantic Access, Volume 36, Issue 3, Data and Knowledge Engineering, p. 317-335.

[3] Gilbert, R. Liu, Y. and Abriel, W. (2004), Reservoir modeling: integrating various data at appropriate scales, *The Leading Edge*, Vol. 23(8) (pp. 784-788), EAGE, The Netherlands.

[4] Han, J. and Cercone, N. (2000), Aviz: A visualization system for discovering numeric association rules. In: Terano. T, Liu. H, Chen A.L.P.

(eds.) *PAKDD* 2000, LNCS (LNAI) vol. 1805, pp. 269-280, Springer, Heidelberg.
[5] King, E. (2000), "Data Warehousing and Data Mining: Implementing Strategic Knowledge Management", 1st Ed, CTR Corporation, ISBN 1566070782, SC, USA.
[6] Longley, I.M. Bradshaw, M.T. & Hebberger, J. (2001), Australian petroleum provinces of the 21st century, in Downey, M.W. Threet, J.C. & Morgan, W.A (2001) Petroleum provinces of the 21st century, *AAPG Memoir, 74*, 287-317.
[7] Lugmayr, A. Stockleben, B. Scheib, C. M. Mailaparampil, M. Mesia, N. and Đanta, H. (2016), "A Comprehensive Survey on Big Data Research and It's Implications - What is really 'new' in Big Data? It's Cognitive Big Data," Proceedings of the 20th Pacific-Asian Conference on Information Systems (PACIS 2016), S.-I.C. Shin-Yuan Hung Patrick Y.K. Chau Ting-Peng Liang, ed.
[8] Marakas, M. G. (2003), "Modern Data Warehousing, Mining, and Visualization Core Concepts", Prentice Hall Pub.
[9] Nimmagadda, S.L, Dreher, H, Noventianto. A, Mustofa. A and Fiume. G. (2012), Enhancing the process of knowledge discovery from integrated geophysical databases using geo-ontologies, a paper presented and published *in the proceedings of Indonesian Petroleum Association (IPA)* conference, held in Jakarta, Indonesia.
[10] Nimmagadda, S. L. and Dreher, H. (2012), "On new emerging concepts of Petroleum Digital Ecosystem (PDE)", Journal *WIREs Data Mining Knowledge Discovery,* 2012, 2: 457–475 doi: 10.1002/widm.1070.
[11] Nimmagadda, S.L. (2015), Data Warehousing for Mining of Heterogeneous and Multidimensional Data Sources, Verlag Publisher, Scholar Press, OmniScriptum GMBH & CO. KG, p. 1-657, Germany.
[12] Nimmagadda, S.L. and Rudra, A. (2016), Big Data Information Systems for Managing Embedded Digital Ecosystems (EDE), in a book entitled "*Big Data and Learning Analytics in Higher Education: Current Theory and Practice",* Springer International, DOI: 10.1007/978-3-319-06520-5, ISBN: 978-3-319-06519-9, The Netherlands.
[13] Post, H.F., Gregory, M. N. and Bonneu, G. (2002), Data Visualization: The State of the Art, Research Paper TU Delft, EUROGRAPHICS 2002, The Netherlands.