Towards an educational model for data scientists in HEIs in South Africa.

Dr Patricia Rudo Chikuni

Data Curation Officer

Digital Library Services, UCT Libraries

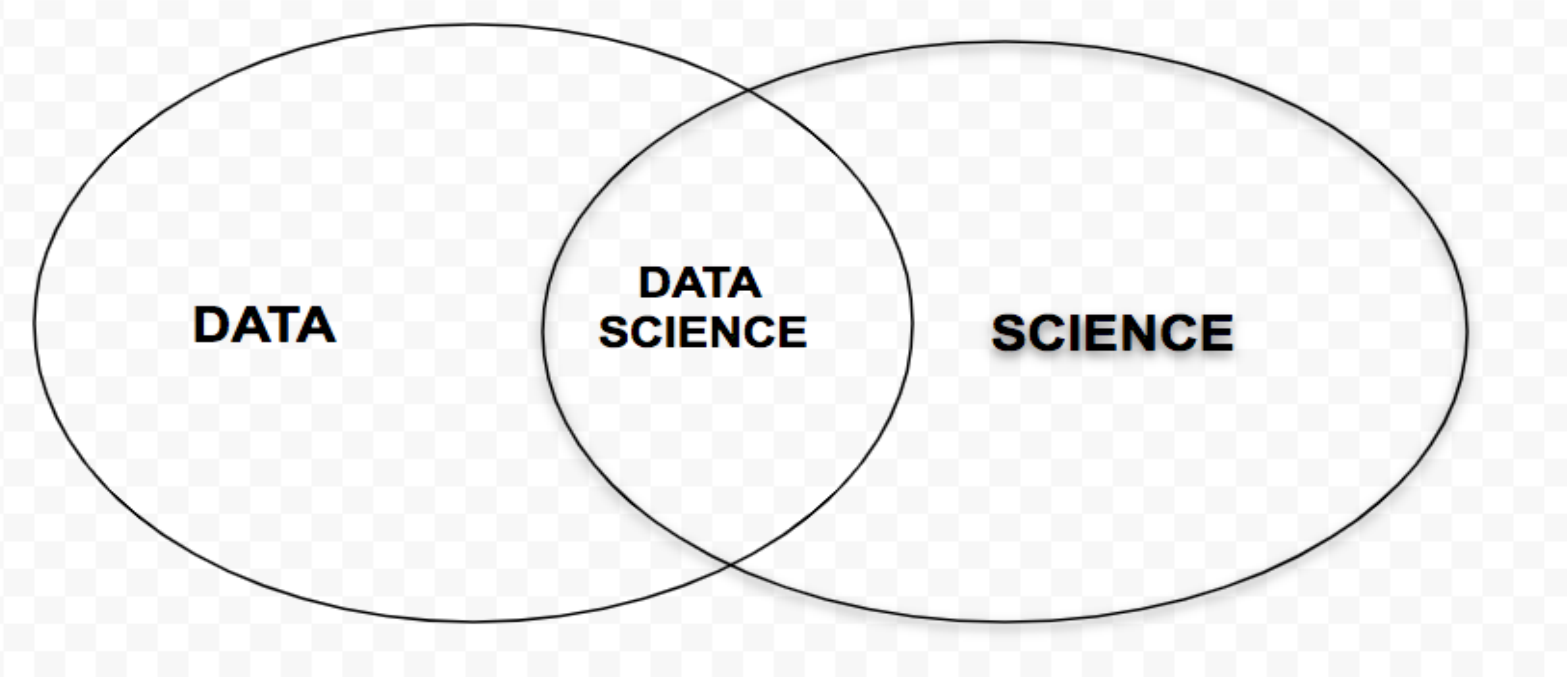
Chikuni, P. R. (2019) Towards an educational model for data scientists in Higher Education Institutions (HEIs) in South Africa: https://doi.org/10.25375/uct.7937408

Outline

- An introduction to Data Science
- Scientific research processes
- Changing workflows in doing research
- The emergent roles of data scientists in HEIs in South Africa
- The skills that data science will enhance for universities in South Africa

‘Data Science’

Data or Science?



https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

Why focus on science & not data?

DATA

Volume
Velocity
Tools
Python vs R

SCIENCE

Research Question
Gap
Adequacy of data
Relationships
Structure

Problem statement

The skills that are foregrounded in most data science programs are not enough to solve the types of problems that modern scientists face (Blei and Smith, 2017: 8689)

Very few studies discuss data science from the perspective of scientific research (Blei and Smith, 2017:8689). Misunderstanding reigns about the roles of DS(Harlan, Harris, Murphy & Valsman, 2012)

Whilst research has become more data driven, an issue that pervades many, if not all, scientific disciplines is that scientists cannot yet fully take advantage of their new data (Blei and Smith, 2017,p.8689)

RESEARCH QUESTIONS

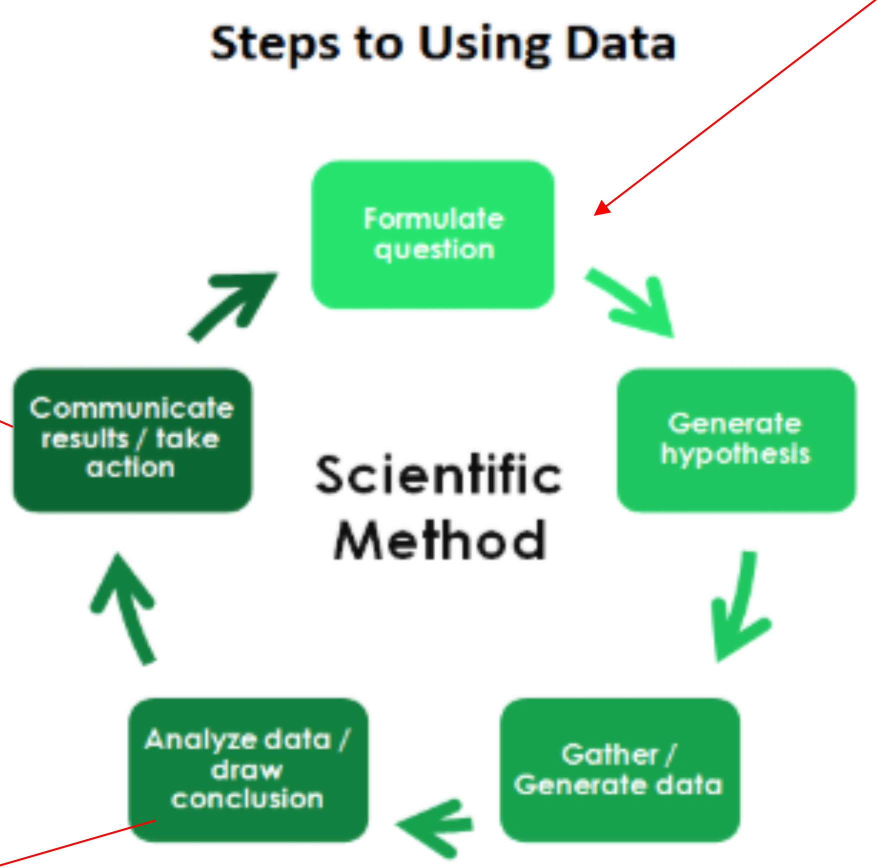
1. How are the emergent roles of data scientists conceptualized in the emergent data science curriculum in HEIs in South Africa?

2. What is the strategic role of a data scientist in the scientific research process.

Scientific research processes

Generating data using the conventional scientific method of enquiry

Formulate a Qsn?
Problem statement



Communicate the results

Generate a hypothesis that can be tested

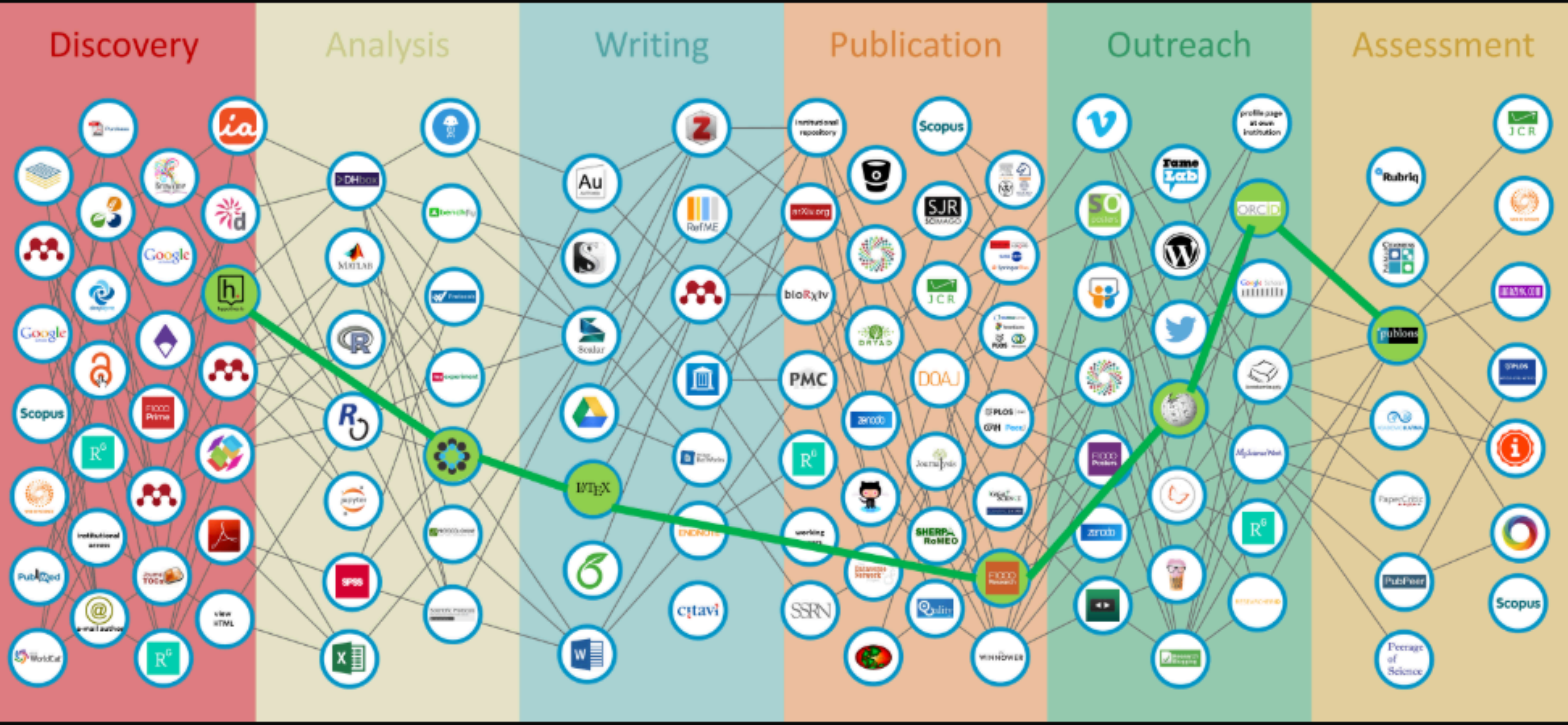
Analyze the data
Test the hypothesis
Draw results and conclusions

Understand the phenomenon through experiments, observations, measurements

Hypothetical open science workflow

Bosman, J. and Kramer, B. (2015) Innovations in scholarly communication: changing research workflows. Available at https://101innovations.wordpress.com/workflows

HYPOTHETICAL WORKFLOWS SUPERIMPOSED ON TOOL COMBINATIONS



Generating data in the context of the 4IR

- The 4IR consists of those “technological developments that blur the lines between the physical, digital and biological spheres...it integrates cyber-physical systems and the Internet of things, big data and cloud computing, robotics and artificial intelligence-based systems” (World Economic Forum, 2016)

- Across global society a diversity of new technologies – disruptive, constraining and enabling in complex ways - are changing the ways that we live and work

The research data lifecycle

THE RESEARCH DATA LIFECYCLE

1. Concept	Start by identifying the research problem
2. DMP	-A DMP helps to ensure that your data is safe and shareable and many funders now require detailed DMPs to be submitted as part of a research proposal. -Identify which data will be collected -How it will be organised, documented and stored along with quality assurance, storage and preservation plans
3. Data Collection	-Data types, sources, volume, and file formats, -Include descriptive, technical and administrative metadata and use open machine readable formats -Use logical file names -Keep data in raw format whenever possible to facilitate future reanalysis and analytical reproducibility
4.Data Pre-processing	-Use appropriate encryption or anonymisation methods and privacy protocols -Use version control and keep raw data separate from derived clean data -Record workflows for provenance and context
5. Data storage, access and collaboration	-Have systematic backup scheme. Storage method depends on size and nature of data, costs of storage, how the data will be used, time to transfer, who needs access and privacy concerns

McLean, Cameron (2016): The Research Data Lifecycle. figshare. Poster.https://doi.org/10.17608/k6.auckland.3102922.v1

6. Data repurposing	Ensure you have consent or legal rights to reuse data. Help others by choosing open licences and formats, structures that make it easy to combine data
7. Analysis and modelling	How computationally intensive are your analytical processes. Conduct analysis with a particular level of reuse in mind track versions of data and any processes used to generate them Keep an electronic lab notebook to record metadata that will later be packaged with final data that is stored, reused and shared.
8. Knowledge transfer, publishing and sharing	Establish copyright and licensing of data, give data a permanent unique identifier and publish in institutional/ discipline repositories. [ZivaHub-UCT's open data repository] or journal repositories. Cite and link your data in publications, also provide and create discovery metadata along with user documentation or links to provide the context needed to interpret the data.
9. Archive and long term management	How long the data should be accessible for? Consider preservation and curation issues, how and where the data will be stored and accessed. The need to migrate data to different formats.

The role of data scientists in the scientific research process

- Data scientists focus on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention.

- Data scientists value the **effective communication of the results of a data analysis** and of the understanding about the world that we glean from it.

- Choudhury (2010) described the importance of the new role of ‘data scientist’ as a person who possesses data management experience as well as domain specific knowledge, who can provide a human interface between the Library and e-Science (science that uses immense data sets).

Methodology

Findings

Preliminary findings

1. Data Science programs are being offered largely as Computer Science, Business and Mathematics or Statistics courses.
2. There is no evidence from the programs analysed that data science pays recognition to research data management as a skill
3. There is a clear influence of the private sector in funding data science courses-implications on the curriculum, discourse & vocabulary.
4. The focus with all the programs is to produce graduates who can work in business.

Preliminary findings

5. Data Science is an ambiguous term associated with producing Data analysts, Machine learning engineers, Data engineers and data scientists.

6. Data science within the context of universities is understood in the context of bridging a gap between scientific e-research processes and researchers.

7. Most data science jobs in universities are situated in Libraries with however differing nomenclature

Reflections from the findings

- Most courses are enrolling learners with a background in statistics and computing.
- Statistics provides the foundational techniques for analysing and reasoning about data
- Computational methods are also key, particularly when scientists face large and complex data and have constraints on computational resources, such as time and memory.
-
- Finally, there is the human angle, the reality that data science cannot be fully automated. Human judgement and deep disciplinary knowledge are necessary skills.

Reflections from the findings

- Some issues are philosophical and fuzzier eg. Mis-specified models of the world, difficulties in identifying causality from empirical data (Bleia and Smythd, 2017)

- Knowledge of the scientific research methods and techniques makes the Data Scientist profession different from all previous professions (Demchenko, 2016).

- Before you reach the algorithms there's an entire process of planning on the research problem (Zimbres, 2017).

The next generation of data scientists

Demands for new professions that should support all stages of the Research Data Lifecycle (RDL) from data production and input to data processing, storing and obtaining scientific results publishing and dissemination (Demchenko, 2016).

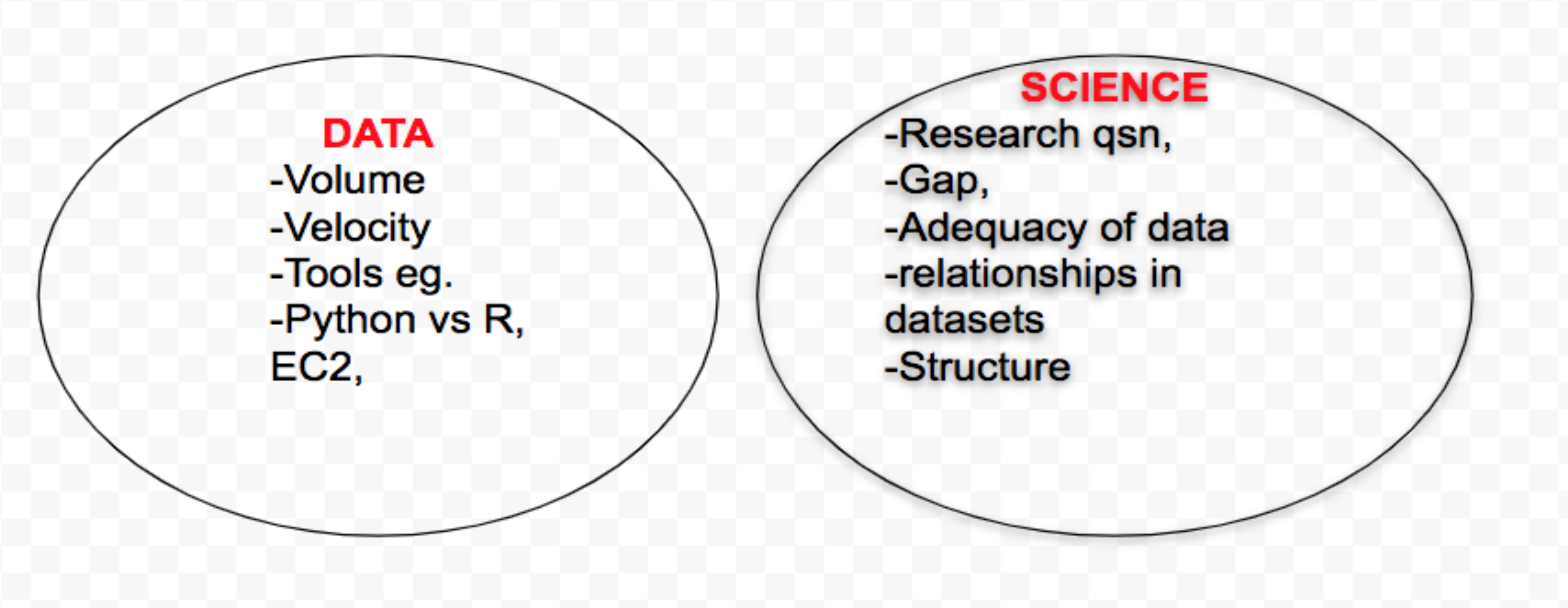
Data scientists must possess knowledge and obtain competencies and skills in data mining and analytics, information visualization and communication, as well as statistics engineering and computer science.

Next generation of data scientists must also acquire experiences in the specific research or industry domain of their future work and specialisation (Demchenko, 2016).

Remarks

- The underlying paradigm of big data-driven machine learning reflects the desire of deriving better conclusions from **simply analysing more data,** without the necessity of looking at theory and models.
- Is having simply more data always helpful? **More data** will not magically give you **better answers**.
- Data Science is fundamentally inter-disciplinary and the education of data scientists must reflect this.
- Is it possible to design a data science curriculum in view of changing technological revolutions.

Final remark



The long term impact of data science will be measured by the scientific questions we can answer with the data.

References:

Bosman, J. and Kramer, B. (2015) Innovations in scholarly communication: changing research workflows. Available at https://101innovations.wordpress.com/workflows/

Brown, D.J. (2009) International Council for Scientific and Technical Information (ICSTI) annual conference. Managing data for Science Information Services and use. Vol.29 No.4, pp.103-21

David M. Bleia and Padhraic Smythd (2017) Science and data science .University of California, Berkeley, CA, and approved June 16, 2017 August 15, 2017 | vol. 114 | no. 33 | 8689–8692

Demchenko, Y. (2016) EDISON data science framework to define the Data Science profession.https://www.kdnuggets.com/2016/10/edison-data-science-framework.html

Elo, S. and Kyngash, H. (2008) The qualitative content analysis process. Journal of Advanced nursing 62(1), 107-115 doi: 10.1111/j.1365-2648.2007.04569.x

Harris-Pierce, R., Liu, Yan Quan (2012) Is data curation education at library and information science schools in North America adequate? New library world, volume 113 Issue: 11/12. Pp. 598-613.

References:

Ijsbrand, A., Dunham, J. and Koers, H. (2013) Connecting Scientific articles with research data: new directions in online scholarly publishing. Data Science Journal 12. https://doi.org/10.2481/dsj.wds-043

Kalam, N. J. (2018) An analysis of job advertisements for research data management librarians.

Pertsas, V. and Costantopoulos, P. (2017) Scholarly ontology: modelling scholarly practoces. International Journal on Digital Libraries vol. 18 Issue 3. Pp. 173-190.

Zimbres, R. (2017). Putting the science back in data science Available at: https://www.kdnuggets.com/2017/09/science-data-science.html