The Emerging Discipline of Data Science:

Principles and Techniques for Data-Intensive Analysis

Keynote

2nd Swiss Workshop on Data Science – SDS|2015

Winterthur, Switzerland

12 June 2015

Version: June 12, 2015

By

Dr. Michael L. Brodie

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Cambridge, MA USA

The Emerging Discipline of Data Science:

Principles and Techniques for Data-Intensive Analysis

The Scientific Revolution (1550-1700) led to the increasing significance, potential, and risks of empiricism – 17th Century knowledge discovery - that in turn led over 400 years to the Scientific Method – *a body of principles and techniques for investigating phenomena, acquiring new knowledge, and correcting and integrating previous knowledge*[1].

The Computing Revolution (1940-1970) led to the increasing significance, potential, and risks of software – 20th Century knowledge work – including the software crisis (1968) that in turn led over 40 years to Software Engineering - a body of principles and techniques for *the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software*[1].

The Digital Revolution (1970-) with the emerging Digital Universe and Big Data Revolution (2000-) is leading to the significance, potential, and risks of data-intensive analysis – 21st Century knowledge discovery – that is leading to the need for Data Science – an emerging discipline currently in its infancy, analogous to the scientific method and software engineering in their revolutions. The importance of Data Science can be seen in the potential impact on the quality of lives of the US Government's Precision Medicine Initiative for "Delivering the right treatments, at the right time, every time to the right person."

This exploratory talk examines Data Science from data analysis to data-intensive analysis. Data analysis with roots in Babylonia (1700-1200 BCE) and India (1200 BCE) is applied in most human endeavors following well-established principles, e.g., statistics, and guidelines, e.g., the Cross-Industry Standard Process for Data Mining. The roots of data-intensive analysis are in Big Data (~2000) that, just emerging, is opening the door to profound change – to new ways of thinking, problem solving, and processing that in turn bring new opportunities and challenges. Since 2007, this *Fourth Paradigm [5]* of science is being applied to evidence/data-based analysis in most human endeavors.

The talk presents an emerging data-intensive analysis workflow that augments the previously dominant data analysis phase with an equally important and substantial data management phase and correspondingly augments the scope of Data Science. Through use cases we identify opportunities and challenges across the data-intensive analysis workflow and their requirements for *principles and techniques to measure and improve the correctness, completeness, and efficiency of data-intensive analysis.*

[1] Wikipedia

References

[1] A. J. G. Hey, S. Tansley, and K. M. Tolle, "The fourth paradigm: data-intensive scientific discovery," 2009.

[2] ACCELERATING DISCOVERY IN SCIENCE AND ENGINEERING THROUGH PETASCALE SIMULATIONS AND ANALYSIS (PetaApps), National Science Foundation, Posted July 28, 2008.

[3] Interview: Michael Brodie, leading database researcher, industry leader, thinker. SIGKDD Explor. Newsl. 16, 1 (September 2014), 57-63. DOI=10.1145/2674026.2674035 *http://doi.acm.org/10.1145/2674026.2674035* by Gregory Piatetsky (re-published)

[4] Jennie Duggan and Michael L. Brodie Hephaestus: Data Reuse for Accelerating Scientific Discovery, In Conference on Innovative Data Systems Research (CIDR) 2015, January 2015.

[5] Jim Gray on eScience: a transformed scientific method, in Tony Hey, Stewart Tansley, Kristin M. Tolle (Eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research 2009 ISBN 978-0982544204

[6] Michael L. Brodie and Jennie Duggan, *What versus Why. Towards Computing Reality*, Operational Database Management Systems, April 22, 2014.

[7] Michael L. Brodie, Accelerating Scientific Discovery: Efficacy, Efficiency, and Reuse of Big eScience Data, Univ. of New South Wales, Canberra; Univ. of Technology, Sydney; RMIT, Melbourne, September 2014

[8] Michael L. Brodie, Data Curation @ Scale: Tools and Techniques for the Emerging Discipline of Data Science, RMIT, Melbourne, September 16, 2014 (Presentation)

[9] Michael L. Brodie, *Laws and Limits of Data Science: The Next Decade*, Keynote, Analytics Week Conference 2014, Boston, MA November 7, 2014.

[10] Michael L. Brodie, Piketty Revisited: Improving Economics through Data Science: How Data Curation Can Enable More Faithful Data Science (In Much Less Time), *Tamr Featured Content, Insights* October 2014. Re-published *KDnuggets*.

[11] Michael L. Brodie, The First Law of Data Science: Do Umbrellas Cause Rain? Operational Database Management Systems and KDNuggets, June 10, 2014

[12] Michael L. Brodie, *The Other Side of Big Data*, Interview by Prof. Roberto V. Zicari, Editor, ODBMS Industry Watch, April 26, 2014

[13] Michael L. Brodie, *What is Big Data for?* Letter to the editor, The Economist, May 23, 2014