The Emerging Discipline of Data Science:

Principles and Techniques for Data-Intensive Analysis

Qatar Computing Research Institute (QCRI)
Doha, Qatar
October 21, 2015

Version: October 5, 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

Over the past two decades, Data-Intensive Analysis (Big Data Analytics) has emerged not only as a basis for the *Fourth Paradigm [1][3]* of engineering and scientific discovery but more broadly as a basis for discovery in most human endeavors. The roots of Data-Intensive Analysis are in Big Data (~2000) that, just emerging, are opening the door to profound change – to new ways of thinking, problem solving, and processing that in turn bring new opportunities and challenges.

Data-Intensive Analysis has produced significant results in areas from particle physics (e.g., Higgs Boson), to identifying and resolving sleep disorders using Fitbit data, to recommenders for literature, theatre, and shopping. More than 50 national governments have established data-driven strategies as policy directions as in Science and engineering [2] as well as in healthcare, e.g., US National Institutes of Health and President Obama's Precision Medicine Initiative for "Delivering the right treatments, at the right time, every time to the right person." The hope is that data-driven techniques will accelerate the discovery of cures to manage and prevent chronic diseases that are more precise and tailored to specific populations as well as being at dramatically lower cost.

With these amazing potential rewards what are the associated potential risks of recommending the wrong film, the wrong product, the wrong medical diagnoses, treatments, or drugs? Do we understand Data-Intensive Analysis to the extent that we can assign probabilistic measures of likelihood to such analytical results? With the scale and emerging nature of Data-Intensive Discovery, how do we estimate the correctness and completeness of analytical results relative to a hypothesized discovery question when the underlying principles and techniques may no longer apply? Given the potential risks and rewards of Data-Intensive Analysis and its breadth of application across conventional, empirical scientific and engineering domains as well as across most human endeavors we better get this right.

To better understand Data-Intensive Analysis and the significance of this challenge I examined over 30 Data-Intensive Analysis use cases that are at very large-scale - in the range where theory and practice may break. This talk presents results of this research related to defining Data Science as a body of *principles and techniques with which to measure and improve the correctness, completeness, and efficiency of Data-Intensive Analysis.* As with its predecessors, establishing this new Fourth Paradigm and the underlying principles and techniques of Data Science may take decades.

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] 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

Related Articles

[4] P. Faller, *"We Can't Rely on Machines"* an interview with Michael L. Brodie, ZHAW Impact 30/15, No. 30 September 2015. Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Switzerland. Republished ODBMS.ORG

[5] M. L. Brodie**,** *Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery*, in Shannon Cutt (ed.), Getting Data Right: Tackling The Challenges of Big Data Volume and Variety, O'Reilly Media, Sebastopol, CA, USA, June 2015

[6] M. L. Brodie, Doubt and Verify: Data Science Power Tools, KDnuggets, July 2015. Republished on ODBMS.org.

[7] Duggan and M. L. Brodie, "Hephaestus: Data Reuse for Accelerating Scientific Discovery," *CIDR 2015*, Jan. 2015.

[8] M. L. Brodie, "*Piketty Revisited: Improving Economics through Data Science – How Data Curation Can Enable More Faithful Data Science (In Much Less Time)*," *KDnuggets*, Oct. 2014.

[9] M. L. Brodie, "*The First Law of Data Science: Do Umbrellas Cause Rain?*," *KDnuggets*, Jun. 2014.