DATA SCIENCE: EVOLUTIONARY APPROACH AND ITS CHALLENGES

Amit Sharma
Department of Computer Science & IT,
Kathua Campus
University of Jammu, J&K

ABSTRACT: "Data Science" is a new buzzword that describes a discipline that really isn't all that new. At its core, most consider data science to be applied mathematics and statistics, principles and concepts that have been critical in driving business, engineering, and science for ages. Data science is the new amalgam of applied mathematics with the "big data" hype of the last decade, in particular the technology and finance sector. In my opinion, data science is incredibly difficult to hire for. To be a data scientist, you first and foremost need to have a huge array of expertise that spans business acumen as well as technical and academic proficiency with machine learning. This paper explores the evolution of data science and its challenges. Our goal is to encourage data related researchers to transfer their focus towards this new science.

KEYWORDS: Machine Learning, Data Science, Big Data, Hadoop, R, Python, HDFS

I. INTRODUCTION

First, the term "data science" is a misnomer with respect to what most people consider endeavours classified as such. Fundamentally, "science" is about formalizing a hypothesis given a reasonable set of observations and assumptions, designing an experiment around that hypothesis, testing it and analysing the data generated through that process to either confirm or falsify the hypothesis. Therefore, "data" is simply a natural by-product of science. Very rarely are things labelled as data science actually scientific. Rather, data science most often refers to the tools and methods used to analyse large amounts of data. As such, the discipline is an amalgamation of many bits from other areas of research. For tools, the influence primarily comes from computer science, where issues of algorithmic efficiency and storage scalability form the main focus. For analysis, however, the influences are much more varied. Modern methods are borrowed from both the so-called hard sciences (physics, statistics, graph theory) and the social sciences (economics, sociology, political sciences, etc). Specific classes of techniques that are naturally interdisciplinary are also very popular, such as machine learning.[1]

Basically we can deduce that Data Science is the practice of: Asking questions (formulating hypothesis), answers to which solve known problems or unearth unknown solutions that in turn drive business value, Defining the data needed or working with an existing data set and employing tools (computer science based) to collect, store and explore such data generally in huge volume & variety (often more than 1 TB and 1000s of dimensions), Identifying the type of analysis to be done to get to the answers and performing such analysis by implementing various algorithms/tools (statistics based), often in a distributed and parallel architecture, Communicating the insights gathered from the analysis in the form of simple stories/visualizations/dashboards (the Data Product) that a non-data scientist can understand and build conversation out of it. (It should be kept in mind that a product can also be an piece of code that is internal to a company and is used by various departments. The presentation, maintenance, scalability, etc of the code are then the product features, which is often not practiced in many organizations), Building a higher level abstraction, analysing and taking actions on new data as they are fed to the system.

II. Historical Background

In the year 1974, Peter Naur publishes "*Concise Survey of Computer Methods*" in Sweden. The book is a survey of contemporary data processing methods that are used in a wide range of applications. It is organized around the concept of data as defined in the "*IFIP Guide to Concepts and Terms in Data Processing*", which defines data as "*a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process*."[2] The Preface to the book tells the reader that a course plan was presented at the IFIP Congress in 1968, titled "*Datalogy, the science of data and of data* *processes and its place in education*,“ and that in the text of the book, the term ‘*data science*’ has been used freely. Naur offers the following definition of data science: “*The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences*.”[3]

In 1977, The International Association for Statistical Computing (IASC) was founded as a Section of the ISI. “*It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge*.” In 1996 *Members of the International Federation of Classification Societies (IFCS)* meet in Tokyo for their biennial conference. For the first time, the term “*data science*” is included in the title of the conference (“Data science, classification, and related methods”). The IFCS was founded in 1985 by six country- and language-specific classification societies, one of which, The Classification Society, was founded in 1964. The aim of these classification societies has been to support the study of “*the principle and practice of classification in a wide range of disciplines and research in problems of classification, data analysis, and systems for ordering knowledge*” (IFCS), and the “*study of classification and clustering (including systematic methods of creating classifications from data) and related statistical and data analytic methods*” [4]. The classification societies have variously used the terms data analysis, data mining, and data science in their publications. It made data science well known to the circles of researchers and distinguished it from other data analysis terms such as classification which are not broad as data science. This helped gradually make data science as an independent field.

The role of data science started to become more apparent at the end of the 1990’s as data bases grew larger. This was highlighted very eloquently by Jacob Zahavi in December 1999 in his article “*Mining Data for Nuggets of Knowledge*”. Conventional statistical methods work well with small data sets. In the 2000’s publications about data science started to appear at an increasing rate, though they are mainly academic. Journals and books on data science became more common and attracted interest among through the researchers.

2009 was a great year for the data science Yangyong Zhu and Yun Xiong, two researchers from the Research Center for Datalogy and Data Science, Shanghai, China declared in their publication *“Introduction to Datalogy and Data Science*” [5], that the data science was a new science, distinctly different from natural science and social science.

III. DATA SCIENCE AS A VENN DIAGRAM

In the year 2010, Drew Conway gave a Venn-diagram definition [1] which clearly illustrates the key components of data science.

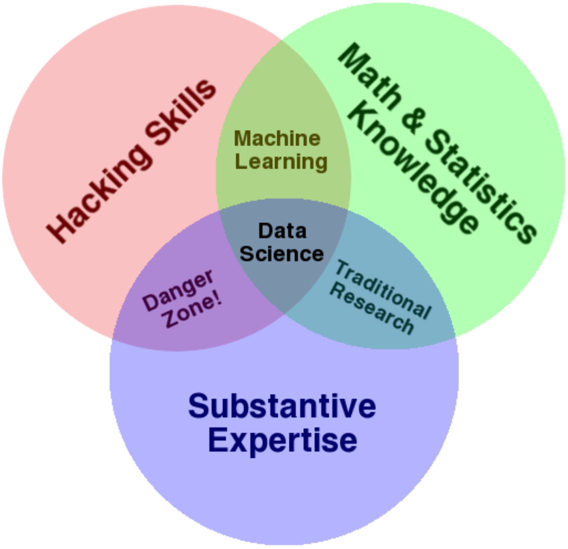

Figure 1. Venn diagram

- ***Hacking skills:*** A data scientist must have the ability to extract and structure data. To do so, he/she must possess advanced programming abilities to manipulate data and apply algorithms.
- ***Math and Statistics Knowledge:*** To extract meaning from large volumes of data, a data scientist must have knowledge of at least some basic level of mathematics and statistics, since most data science techniques involve statistical computation and modelling.
- ***Substantive Expertise:*** Since the fundamental aim of data science is to build knowledge, it must build upon previous knowledge bases and discoveries. This requires that the data scientist must have a large amount of experience at his/her disposal, so that the best results can be obtained from the new data.

Data Science is right there at the middle, combining the skills of Hacking, Substantive Expertise, and Math/Statistics Knowledge. I especially like the way it highlights the danger of applying statistical tools (including R) to an applied problem without a rigorous statistical background. [6] For Conway, the centre of the diagram is *Data Science*. There's some controversy over what the bottom circle means, all I can say, is if Conway meant something other than what I would call domain knowledge (e.g. physics), he chose the name *Substantive Expertise* very poorly. So assuming domain knowledge is at least part of what he meant, the idea is that a physicist, say, would have expertise in physics and math/stats knowledge, but lack hacking knowledge. *Machine Learning* experts tend to apply algorithms without an understanding of the domain they're analysing. And the people who can program and know their field but have no way to tell a statistically significant result from one arising from sheer coincidence are dangerous; they can arrive at some drastically wrong solutions.

IV. TECHNOLOGIES USED IN DATA SCIENCE

i. Map Reduce

The term Map Reduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce job is always performed after the map job. [7]

ii. Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is a versatile, resilient, clustered approach to managing files in a big data environment. HDFS is not the final destination for files. Rather, it is a data service that offers a unique set of capabilities needed when data volumes and velocity are high. Because the data is written once and then read many times thereafter, rather than the constant read-writes of other file systems, HDFS is an excellent choice for supporting big data analysis.[8]

iii. Advanced Text Analytics

Advanced Text analytics can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology.[9]

iv. Large Scale data programming Languages

Programming languages that can be used with large data sets (especially big data) in an efficient manner. These were underdeveloped or completely absent before data science appeared. Some of the languages now we are using are R-language, Python, Scala, Jethro etc.

v. Alternative database Structures

Databases for archiving, querying and editing big data using parallel computing technologies. Few important alternative databases are Hbase, Cassandra, MongoDB, Google Big Table, etc.

V. CHALLENGES

Some Important challenges one can encountered in the field of Data Science are

a. Poor-quality data such as: dirty data, missing values, inadequate data size, and poor representation in data sampling.
b. Lack of understanding/lack of diffusion of data mining techniques in academic arenas.
c. The lack of good literature on important data mining topics and techniques.
d. (Academic institutions) have trouble accessing commercial-grade software at reasonable costs.
e. Data variety - trying to accommodate data that comes from different sources and in a variety of different forms (images, geo data, text, social, numeric, etc.).
f. Data velocity - online machine learning requires models to be constantly updated with new, incoming data.
g. Dealing with huge datasets, or 'Big Data,' that require distributed approaches.
h. Coming up with the right question or problem
i. Remaining objective and allowing the data to lead you, not the opposite. Preconceived notions can be dangerous, but luckily it is in our power to resist them.[10]

VI. CONCLUSION

Since data science become the need of the hour. This field would rule the entire world of computation near future. Various techniques and algorithms are being developed and are in use. When you hear people going on about the Internet of Things, this is where that fits in. Where in the past common datasets have included things like sales/purchase data or click stream data, more and more you will see data scientists asked to writing value from sensor-generated data from manufacturing lines, retail environments, vehicles, even offices. A lot of this data will be time series based and have its own set of unique problems. Tools emerging to make things that are difficult today much easier.

We already see this happening with Business Intelligence tools, and open source libraries in the R and Python communities. Data science and quantitative methods becoming distributed throughout roles rather than concentrated in a single role or department. This comes along with the point above. If there are tools that provide the power of Python or Spark with the ubiquity and simplicity of something like Excel, and companies really believe in more quantitative approaches, there will be people in HR, sales, manufacturing, finance, etc. doing work that looks a lot like things data scientists are doing today. Which is to say that if you learn data science today, your job title 20 years from now might not be “Data Scientist” but I’m confident your skills will still be relevant? If it interests you, go learn without fear!

VII. REFERENCES

1. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
2. IFIP guide to concepts and terms in data processing, North-Holland Publ. Co., Amsterdam, 1971.
3. Peter Naur: Concise Survey of Computer Methods, 397 p Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974 ISBN/Petro celli 0-88405-314-8, 1975.
4. Zacharias Voulgaris: Data Scientist: The Definitive Guide to Becoming a Data Scientist, Technics Publications, 01-May-2014, ISBN-10: 193550469X,ISBN-13: 978-1935504696.
5. Yangyong Zhu, Yun Xiong. Introduction to Datalogy and Data Science 2009. http://www.datascience.cn/en/viewpoint.aspx.
6. http://blog.revolutionanalytics.com/2010/10/the-data-science-venn-diagram.html
7. https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/
8. http://www.dummies.com/programming/big-data/hadoop/hadoop-distributed-file-system-hdfs-for-big-data-projects/
9. http://searchbusinessanalytics.techtarget.com/definition/text-mining
10. http://info.salford-systems.com/blog/bid/305673/9-Data-Mining-Challenges-From-Data-Scientists-Like-You
11. https://www.quora.com/What-is-the-future-of-data-science-1