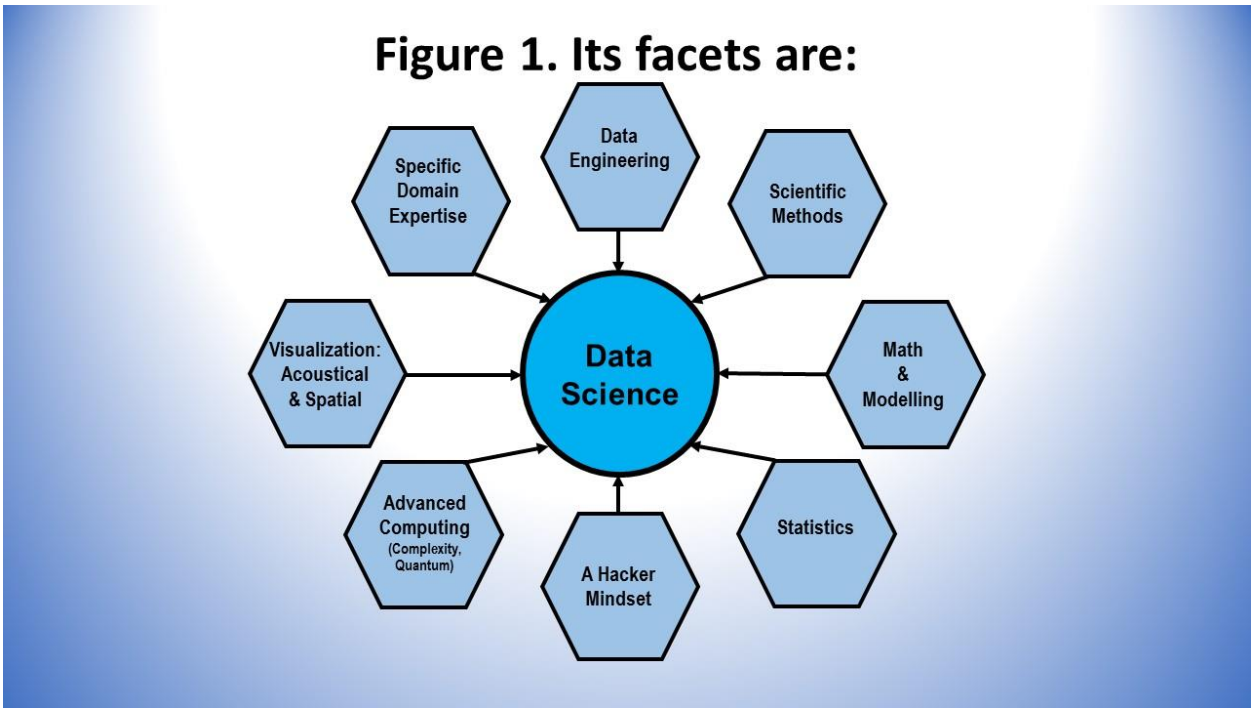
WHAT IS DATA SCIENCE AND HOW IS IT RELATED TO AI

Richard Bruno
CIRANO
March 2017
V1.5

Data science[1] covers the whole spectrum of data processing, not just the algorithmic or statistical aspects. It uses a combination of statistics, mathematics, programming, problem solving, capturing data in ingenious ways to look at things differently in order to cleanse, prepare, and align data for specified needs.

Data science deals with unstructured (*e.g. music, speech, pictures, video,*), the biggest pipeline as well as the fastest growth sector, and structured data (*e.g. data in relational data bases*). It is a field that comprises everything that is related to data cleansing, preparation, analysis and synthesis. As such, it is the umbrella of techniques used and disciplines applied to extract insights from data[1].

The most important facets of data science are shown in Figure 1.

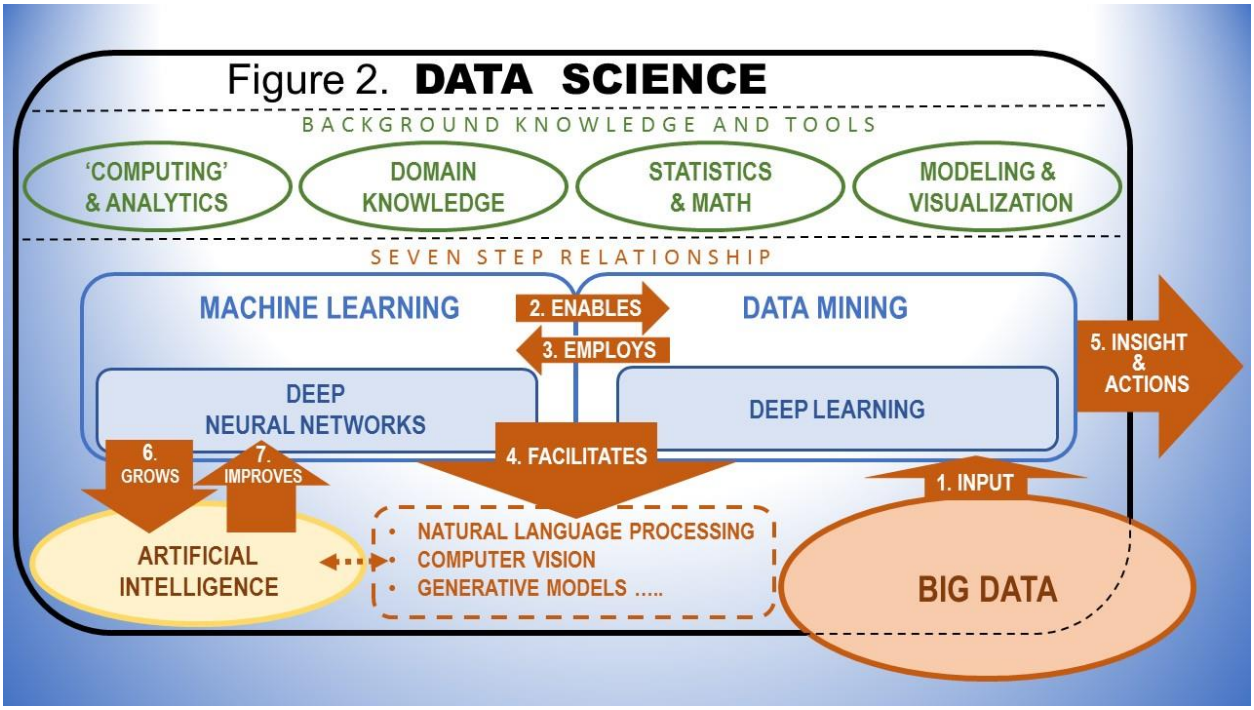


These include data engineering methods, well established scientific methods, mathematics, digital modelling, most of statistics, advanced computational methods like complexity analysis and reduction or

[1] An example of a data science project is the creation of Twitter profiling for computational marketing. It leverages big data, and is part of the viral marketing business and its growth strategy. This also includes automated high quality (*relevant and fact-based*) syndicated content generation (*e.g. digital publishing version 3.0*). Note that Data science overlaps significantly with or may wholly contain certain disciplines:

- **Computer science ("computing")**: overlaps significantly with data science via computational complexity, internet topology and graph theory, distributed architectures, data plumbing (optimization of data flows and in-memory analytics), data compression, computer programming, sensor processing and the streaming of real-time data a.o..
- **Statistics ("data analysis/analytics")**: including e.g. multivariate testing, cross-validation, stochastic processes, sampling, model-free confidence intervals (*but, e.g., not p-value or obscure tests of hypotheses that are subjects of the 'curse' of big data or very-sparse data tests*) are used in data science.
- **Machine learning** and **data mining**: data science indeed fully encompasses these two domains.
- **Operations research**: data science encompasses most of operations research (*i.e. that which is data based*) as well as any techniques aimed at optimizing decisions based on analyzing data.
- **Business intelligence**: every business intelligence aspect of designing/creating/identifying metrics and KPI's, creating database schemas, dashboard design and visualizations, and data-driven strategies to optimize decisions and ROI, is fully part of data science.

quantum computing, visual and acoustical signal processing and feature extraction, a hacker mindset[2] and having deep expertise in a disciple or domain.



The steps in Figure 2 are:

1. Input Big Data to Deep Learning and other algorithms for Data Mining
2. Have Machine Learning algorithms Enable Data Mining processes.
3. Data Mining Employs its results to modify Machine Learning algorithms
4. Steps 2&3 impact:
    a. Deep Learning in-production-use algorithms
    b. Deep Neural Network's prototype algorithms' architecture and parameters
    c. Both of these, in turn, Facilitate Natural Language Processing, Computer Vision, Generative Learning Models etc.
5. All of which result in new Insights (e.g. creation of Twitter profiling for computational marketing) & Actions (e.g. for fraud detection)
6. Thereafter Deep Neural Networks add new learning to Grow Artificial Intelligence ("AI") field
7. And, new research findings in AI Improves Deep Neural Networks.

Data scientists[3] can be found anywhere in the lifecycle of data science projects, that is, at the data gathering stage, or the data exploratory stage, all the way up to statistical modeling and maintaining of existing systems stages.

[2] Finding opportunities and therein being clever, ethical, seeking excellence and enjoyment out of deliberate behavior.

[3] The categories of data scientists are those with a:

- Strength in statistics: they develop new statistical theories for big data. They are e.g. expert in statistical modeling, experimental design, sampling, clustering, data reduction, confidence intervals, testing, modeling, predictive modeling and related techniques.
- Strength in mathematics: e.g. techniques of minimum complexity and high throughput to collect, analyze and extract value out of data.
- Strength in data engineering: e.g. Hadoop, database/memory/file systems optimization and architecture, Application Programming Interfaces, Analytics as a Service, optimization of data flows, data plumbing.
- Strength in machine learning / computer science: algorithms, the management of computational complexity

The term **Big Data**[4] refers to the handling of very large unstructured data sets in automated ways. This usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process within acceptable lapse times. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.

The processing of big data begins with the raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. Big data constantly inundates businesses and institutions on a day-to-day basis. It is a starting point to analyze information which can lead to better insights, decision making, process automation, predictions and/or actions.

Machine learning[5] is a part of Data Science. It involves the use of a set of algorithms that train on a large data sets (i.e. big data) to gain insights that result in predictions or take actions in order to optimize some

- Strength in business domain: e.g. ROI optimization, decision sciences, dashboards design, metric mix selection and metric definitions, high-level database design.
- Strength in natural languages: natural language processing
- Strength in production code development, software engineering (use of programming languages) and the use of levels of abstraction
- Strength in visualization: e.g. spatial and acoustical
- Strength in Geographic Information Systems: spatial data, data modeled by graphs, graph databases

[4] **Big data** "size" is a constantly moving target however, as of 2016, it ranges from a hundred or so terabytes (*10**12*) to a hundred or more petabytes (*10**15*) of data. Today each person in the OECD creates about 4TB (4*10**12)/month. By comparison, note that the largest storage devices for PCs are about 10 terabytes and that the human brain has a few billion neurons each containing, on average, 1000 connections. Simple math might indicate that the human brain only has the capacity to store a few terabytes however, neurons combine to help each other store multiple memories all at once thus increasing the brain's capacity to a few petabytes. Also note that a googol (*google*) of bytes is 10**100 bytes.

Some characteristics when thinking of big data are:

- Volume: size; this aspect is also related to processing capacity (*both of which are expanding at 'Moore's Law' rates*). Big data doesn't sample like an analog audio signal; it just observes and tracks what happens.
- Velocity: the speed at which data is generated and processed and the ability to process that data in real-time. Big data is more often than not available in real-time.
- Variety: data sets consisting of for example (1) Structured data: typically, text organized in conventional relational databases (2) Unstructured data: audio, sounds, photos, videos, text data like newspaper articles, etc. Big data not only draws from text, images, audio, video but it completes missing pieces through processes of data fusion.
- Veracity: refers to data accuracy and validity.
- Machine learning aspect: big data often doesn't ask why and this aspect simply detects patterns.
- Its digital footprint: big data is often a cost-free byproduct of digital interaction.

[5] **NLP** (Natural language processing) is simply the part of AI that has to do with language; written or oral.

Machine learning: Given some Artificial Intelligence ("AI") problem that can be described in discrete terms (e.g. out of a defined set of actions, which one is the right one), and given a lot of information about the world wherein the problem lies, machine learning is concerned with figuring out what is the "correct" action, without having the programmer program it in. To draw a distinction with AI, if one can write a very clever program that has human-like behavior, it can be AI, but unless its parameters are automatically learned from data, it's not machine learning. Here one designs algorithms (*like in **data mining***), but with an emphasis on prototyping algorithms for production mode, and designing automated systems (*e.g. bidding algorithms, ad targeting algorithms*) that automatically update themselves (*i.e. constantly train/retrain/update the training sets/cross-validate, and refine or discover new rules e.g. for fraud detection*) on a constant basis over time. Python is now a popular language for machine learning development. Core algorithms include, e.g., clustering and supervised classification, rule systems, and scoring techniques. Machine learning is, in a way, the connection between big data and artificial intelligence since it is the process of learning from data over time. However, it's not the only thing connecting those two together.

Data mining: is about designing algorithms to extract insights from rather large and potentially unstructured data (text mining). Techniques include pattern recognition, feature selection, clustering, supervised classification and encompasses a few statistical techniques (*though without the p-values or confidence intervals attached to most statistical methods being used*). Here the emphasis is on robust, data-driven, scalable techniques, without much interest in discovering causes or interpretability. Data mining thus has some intersection with statistics, and it is a subset of data science. Data mining is applied computer engineering, rather than a mathematical science. Data miners use open source software such as Rapid Miner.

Deep neural networks - deep learning are one kind of machine learning – data mining, respectively, that are very popular now. They involve a particular kind of mathematical model that can be thought of as a composition of simple blocks (*function composition*) where some of these blocks can be adjusted to better predict the final outcome.

set of problems or systems. An example are supervised classification algorithms which are used to classify potential clients into good or bad prospects for loan purposes based on historical data.

The word ***learning*** in machine learning means that the algorithms used depend on some data which is used first as a training set to fine-tune a model or algorithm parameters. When this model and the associated algorithms are automated, as in automated piloting or driver-less cars, it is called **deep learning**. It is the deep learning that 'does' the adaption or learning.

When there are masses of data it makes sense to bring in machines to help in processing and analysis. Once one can see what one think's is going to happen (*e.g. an outcome*), and one begins to receive feedback on what actually happens then one's model/algorithms can update automatically and become even better at predicting which actions might occur. Such predictive analytics, when dynamic, drive machine learning so that one's model is constantly becoming more and more accurate.

Artificial intelligence ("AI") is a subfield of data science which is concerned with solving tasks that are easy for humans, but hard for computers. This includes all kinds of tasks, such as planning, moving around in the world, recognizing objects and sounds, speaking, translating, performing social or business transactions, doing creative work (*e.g. making art or poetry*), etc. The sub-areas of AI are:

- Deduction, reasoning, problem solving
- Knowledge representation
- Planning
- Learning
- Natural language processing
- The representation and manipulation of motion and space
- Perception
- Creativity
- Social intelligence
- General intelligence….

Operations research: is about decision science and optimizing traditional business projects such as inventory management, supply chain, pricing etc. Here there is heavy use of Markov Chain models, Monte-Carlo simulations, queuing and graph theory. Big traditional existing companies use traditional operations research whilst most new and small companies use data science to handle pricing, inventory management or supply chain problems. Also, operations research problems can be solved by data science. As a data science based example, car traffic optimization is a modern example of an operations research problem solved with simulations, commuter surveys, sensor data and statistical modeling. In operations research one may or may not learn from data, one can actually not have any data and still model and optimize a problem (*e.g. Linear and Dynamic programming are two of the most used tools in operations research and one doesn't need data for them to work just the ability to correctly model the problem one is trying to solve*).

Data engineering: is the applied part of computer science. It is the implementation part that allows all sorts of data to be easily processed in-memory or near-memory, and to have it flow to and between end-users. A sub-domain is **data warehousing**, as this term is associated with static, siloed conventional data bases, data architectures, and data flows.

Quant: refers to data scientists working for Wall Street on problems such as high frequency trading. They use C++, Matlab and most have backgrounds in statistics, mathematical optimization, and industrial statistics.

Data analysis/analytics: is the new term used today for data and business statistics. It covers a large spectrum of applications including fraud detection, advertising mix modeling, attribution modeling, sales forecasts, cross-selling optimization (*retailing*), user segmentation, churn analysis, computing long-time value of a customer and cost of acquisition, and so on. Data Analytics, more specifically, is the science of examining raw data with the purpose of drawing conclusions about that information. It involves applying an algorithmic or methodological process to derive insights. It is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and/or disprove existing theories or models. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on the information one knows.