Different Analytic Disciplines Compared to Data Science

Abstract

What are the differences between data science, data mining, machine learning, statistics, operations research, and so on?

In this work, there is a comparison of several analytic disciplines that overlap,and explanation of the differences and common denominators. Sometimes differences exist for nothing else other than historical reasons. Sometimes the differences are real and subtle. It is also provided some typical job titles, types of analyses, and industries traditionally attached to each discipline. Underlined domains are main sub-domains.

Keywords: Data Science,Data Engineering, Machine Learning, Data Mining, Big Data, Business Intelligence, Artificial Intelligence, Computer Science, Predictive Modeling, Statistics, Operations Research, Mathematics Optimization, Six Sigma, Actuarial sciences, Econometrics

Data Science

Job titles include data scientist, chief scientist, senior analyst, director of analytics and many more. It covers all industries and fields, but especially digital analytics, search technology, marketing, fraud detection, astronomy, energy, healthcare, social networks, finance, forensics, security (NSA), mobile, telecommunications, weather forecasts, and fraud detection.

Projects include taxonomy creation (text mining, big data), clustering applied to big data sets, recommendation engines, simulations, rule systems for statistical scoring engines, root cause analysis, automated bidding, forensics, and early detection of terrorist activity or pandemics. An important component of data science is automation, machine-to-machine communications, as well as algorithms running non-stop in production mode (sometimes in real time).For instance to detect fraud, predict weather or predict home prices for each home.

An example of data science project is the creation of the fastest growing data science Twitter profile, for computational marketing. It leverages big data, and is part of a viral marketing / growth hacking strategy that also includes automated high quality, relevant, syndicated content generation (in short, digital publishing version 3.0).

Unlike most other analytic professions, data scientists are assumed to have great business acumen and domain expertize -- one of the reasons why they tend to succeed as entrepreneurs. There are many types of data scientists, as data science is a broad discipline. Many senior data scientists master their art/craftsmanship and possess the whole spectrum of s0kills and knowledge; they really are the unicorns that recruiters can't find. Hiring managers and uninformed executives favor narrow technical skills over combined deep, broad and specialized business domain expertize - a byproduct of the current education system that favors discipline silos, while true data science is a silo destructor. Unicorn data scientists (a misnomer, because they are not rare - some are famous VC's) usually work as consultants, or as executives. Junior data scientists tend to be more specialized in one aspect of data science, possess more hot technical skills (Hadoop, Pig, Cassandra) and will have no problems finding a job if they received appropriate training and/or have work experience with companies such as Facebook, Google, eBay, Apple, Intel, Twitter, Amazon,

Zillow etc.

Data science overlaps with:

- Computer Science: computational complexity, Internet topology and graph theory, distributed architectures such as Hadoop, data plumbing (optimization of data flows and in-memory analytics), data compression, computer programming (Python, Perl, R) and processing sensor and streaming data (to design cars that drive automatically)

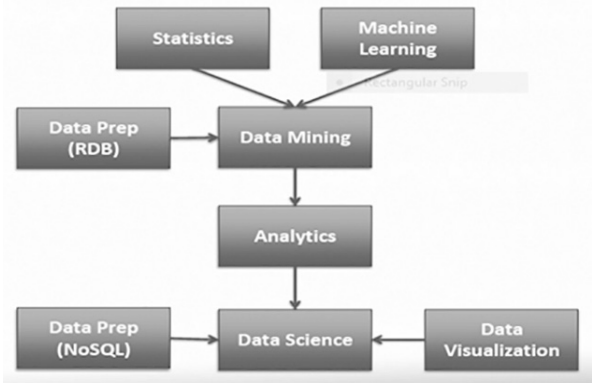
- Statistics: design of experiments including multivariate testing, cross-validation, stochastic processes, sampling, model-free confidence intervals, but not p-value nor obscure tests of the hypotheses that are subjects to the curse of big data

- Machine Learning and Data Mining: data science indeed fully encompasses these two domains.

- Operations Research: data science encompasses most of operations research as well as any techniques aimed at optimizing decisions based on analyzing data.

- Business Intelligence: every BI aspect of designing/creating/identifying great metrics and KPI is, creating database schemas (be it NoSQL or not), dashboard design and visuals, and data-driven strategies to optimize decisions and ROI, is data science.

Data Science Family relation is exhibited in the following figure:



The above figure indicate some of the main areas that are considered the most potential to data science.

Comparison withOther Analytic Disciplines

- Machine Learning: Very popular computer science discipline, data-intensive, part of data science and closely related to data mining. Machine learning is about designing algorithms (like data mining), but emphasis is on prototyping algorithms for production mode, and designing automated systems (bidding algorithms, ad targeting algorithms) that automatically update themselves, constantly train/retrain/update training sets/cross-validate, and refine or discover new rules (fraud detection) on a daily basis. Python is now a popular language for ML development. Core algorithms include clustering and supervised classification, rule systems, and scoring techniques. A sub-domain, close to Artificial Intelligence (see entry below) is deep learning.

Machine Learning	Data Science
Develop new (individual) models	Explain many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few relatively clean few datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action

- Data Mining: This discipline is about designing algorithms to extract insights from rather large and potentially unstructured data (text mining), sometimes called nugget discovery, for instance unearthing a massive Botnets after looking at 50 million rows of data. Techniques include pattern recognition, feature selection, clustering, and supervised classification and encompasses a few statistical techniques (though without the p-values or confidence intervals attached to most statistical methods being used). Instead, emphasis is on robust, data-driven, scalable techniques, without much interest in discovering causes or interpretability. Data mining thus have 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 and software such as Rapid Miner.

- Predictive Modeling: Not a discipline per se. Predictive modeling projects occur in all industries across all disciplines. Predictive modeling applications aim at predicting future based on past data, usually but not always based on statistical modeling. Predictions often come with confidence intervals. Roots of predictive modeling are in statistical science.

- Statistics. Currently, statistics is mostly about surveys (typically performed with SPSS software), theoretical academic research, bank and insurance analytics (marketing mix optimization, cross-selling, fraud detection, usually with SAS and R), statistical programming, social sciences, global warming research (and space weather modeling), economic research, clinical trials (pharmaceutical industry), medical statistics, epidemiology, biostatistics. In addition, government statistics. Agencies hiring statisticians include the Census Bureau, IRS, CDC, EPA, BLS, SEC, and EPA (environmental/spatial statistics). Jobs requiring a security clearance are well paid and relatively secure, but the well-paid jobs in the pharmaceutical industry (the golden goose for statisticians) are threatened by a number of factors - outsourcing, company merging, and pressures to make healthcare affordable. Because of the big influence of the conservative, risk-adverse pharmaceutical industry, statistics has become a narrow field not adapting to new data, and not innovating, losing ground to data science, industrial statistics, operations research, data mining, machine learning -- where the same clustering, cross-validation and statistical training techniques are used, albeit in a more automated way and on bigger data. Many professionals, who were called statisticians 10 years ago, have seen their job title changed to data scientist or analyst in the last few years. Modern sub-domains include statistical computing, statistical learning (closer to machine learning), computational statistics (closer to data science), data-driven (model-free) inference, sport statistics, and Bayesian statistics (MCMC, Bayesian networks and hierarchical Bayesian models being popular, modern techniques). Other new techniques include SVM, structural equation modeling, predicting election results, and ensemble models.

- Industrial Statistics. Statistics frequently performed by non-statisticians (engineers with good statistical training), working on engineering projects such as yield optimization or load balancing (system analysts). They use much applied statistics, and their framework is closer to six sigma, quality control and operations research, than to traditional statistics. Also, found in oil and manufacturing industries. Techniques used include time series, ANOVA, experimental design, survival analysis, signal processing (filtering, noise removal, and deconvolution), spatial models, simulation, Markov chains, and risk and reliability models.

- Mathematical Optimization. Solves business optimization problems with techniques such as the simplex algorithm, Fourier transforms (signal processing), differential equations, and software such as Matlab. These applied mathematicians are found in big companies such as IBM, research labs, NSA (cryptography) and in the finance industry (sometimes recruiting physics or engineer graduates). These professionals sometimes solve the exact same problems as statisticians do, using the exact same techniques, though they use different names. Mathematicians use least square optimization for interpolation or extrapolation; statisticians use linear regression for predictions and model fitting, but both concepts are identical, and rely on the exact same mathematical machinery: it's just two names describing the same thing. Mathematical optimization is however closer to operations research than statistics, the choice of hiring a mathematician rather than another practitioner (data scientist) is often dictated by historical reasons, especially for organizations such as NSA or IBM.

- Actuarial Sciences. Just a subset of statistics focusing on insurance (car, health, etc.) using survival models: predicting when you will die, what your health expenditures will be based on your health status (smoker, gender, previous diseases) to determine your insurance premiums. Also predicts extreme floods and weather events to determine premiums. These latter models are notoriously erroneous (recently) and have resulted in far bigger payouts than expected. For some reasons, this is a very vibrant, secretive community of statisticians that do not call themselves statisticians anymore (job title is actuary). They have seen their average salary increase nicely over time: access to profession is restricted and regulated just like for lawyers, for no other reasons than protectionism to boost salaries and reduce the number of qualified applicants to job openings. Actuarial sciences is indeed data science (a sub-domain).

- HPC. High Performance Computing, not a discipline per

se, but should be of concern to data scientists, big data practitioners, computer scientists and mathematicians, as it can redefine the computing paradigms in these fields. If quantum computing ever becomes successful, it will totally change the way algorithms are designed and implemented. HPC should not be confused with Hadoop and Map-Reduce: HPC is hardware-related, Hadoop is software-related (though heavily relying on Internet bandwidth and servers configuration and proximity).

- Operations Research. Abbreviated as OR. They separated from statistics a while back (like 20 years ago), but they are like twin brothers, and their respective organizations (INFORMS and ASA) collaborate. OR is about decision science and optimizing traditional business projects: inventory management, supply chain, pricing. They heavily use Markov Chain models, Monter-Carlo simulations, queuing and graph theory, and software such as AIMS, Matlab or Informatics. Big, traditional old companies use OR, new and small ones (start-ups) use data science to handle pricing, inventory management or supply chain problems. Many operations research analysts are becoming data scientists, as there is far more innovation and thus growth prospect in data science, compared to OR. Also, OR problems can be solved by data science. OR has a significant overlap with six-sigma (see below), also solves econometric problems, and has many practitioners/applications in the army and defense sectors. car traffic optimization is a modern example of OR problem, solved with simulations, commuter surveys, sensor data and statistical modeling.

- Six Sigma. It is more a way of thinking (a business philosophy, if not a cult) rather than a discipline, and was heavily promoted by Motorola and GE a few decades ago. Used for quality control and to optimize engineering processes (see entry on industrial statistics in this article), by large, traditional companies. They have a LinkedIn group with 270,000 members, twice as large as any other analytic LinkedIn groups including our data science group. Their motto is simple: focus your efforts on the 20% of your time that yields 80% of the value. Applied, simple statistics are used (simple stuff works must of the time, I agree), and the idea is to eliminate sources of variances in business processes, to make them more predictable and improve quality. Many people consider six sigma to be old stuff that will disappear. Perhaps, but the fundamental concepts are being solid and will remain: these are also fundamental concepts for all data scientists. You could say that six sigma is a much more simple if not simplistic version of operations research (see above entry), where statistical modeling is kept to a minimum. Risks: non-qualified people use non-robust black-box statistical tools to solve problems.I can result in disasters. In some ways, six sigma is a discipline more suited for business analysts (see business intelligence entry below) than for serious statisticians.

- Quant. Quant people are just data scientists working for Wall Street on problems such as high frequency trading or stock market arbitraging. They use C++, Matlab, and come from prestigious universities, earn big bucks but lose their job right away when ROI goes too south too quickly. They can also be employed in energy trading. Many who were fired during the great recession now work on problems such as click arbitraging, and optimization and keyword bidding. Quants have backgrounds in statistics (few of them), mathematical optimization, and industrial statistics.

- Artificial Intelligence.. The intersection with data science is pattern recognition (image analysis) and the design of automated (some would say intelligent) systems to perform various tasks, in machine-to-machine communication mode, such as identifying the right keywords (and right bid) on Google AdWords (pay-per-click campaigns involving millions of keywords per day). I also consider smart search (creating a search engine returning the results that you expect and being much broader than Google) one of the greatest problems in data science, arguably also an AI and machine learning problem. An old AI technique is neural networks, but it is now loosing popularity. To the contrary, neuroscience is gaining popularity.

- Computer Science. Data science has some overlap with computer science: Hadoop and Map-Reduce implementations, algorithmic and computational complexity to design fast, scalable algorithms, data plumbing, and problems such

as Internet topology mapping, random number generation, encryption, data compression, and steganography (though these problems overlap with statistical science and mathematical optimization as well).

- Econometrics. Why it became separated from statistics is unclear. So many branches disconnected themselves from statistics, as they became less generic and start developing their own ad-hoc tools. Nevertheless, in short, econometrics is heavily statistical in nature, using time series models such as auto-regressive processes. Also overlapping with operations research (itself overlapping with statistics!) and mathematical optimization (simplex algorithm). Econometricians like ROC and efficiency curves (so do six sigma practitioners, see corresponding entry in this article). Many do not have a strong statistical background, and Excel is their main or only tool.

- Data Engineering. Performed by software engineers (developers, architects or designers) in large organizations. Sometimes data engineering is performed by data scientists in small companies.This is the applied part of computer science, to power systems that allow all sorts of data to be easily processed in-memory or near-memory, and to flow nicely to (and between) end-users, including heavy data consumers such as data scientists. A sub-domain currently under consideration is data warehousing, as this term is associated with static andconventionaldata bases, data architectures, and data flows, threatened by the rise of NoSQL, NewSQL and graph databases. Transforming these old architectures into new ones (only when needed) or make them compatible with new ones, is a lucrative business.

- Business intelligence. Abbreviated as BI. Focuses on dashboard creation, metric selection, producing and scheduling data reports (statistical summaries) sent by email or delivered/presented to executives, competitive intelligence (analyzing third party data), as well as involvement in database schema design (working with data architects) to collect useful, actionable business data efficiently. Typical job title is business analyst, but some are more involved with marketing, product or finance (forecasting sales and revenue). They typically have an MBA degree. Some have learned advanced statistics such as time series, but most only use (and need) basic stats, and light analytics, relying on IT to maintain databases and harvest data. They use tools such as Excel (including cubes and pivot tables, but not advanced analytics), Brio (Oracle browser client), Birt, Micro-Strategy or Business Objects (as end-users to run queries), though some of these tools are increasingly equipped with better analytic capabilities. Unless they learn how to code, they are competing with some polyvalent data scientists that excel in decision science, insights extraction and presentation (visualization), KPI design, business consulting, and ROI/yield/business/process optimization. BI and market research (but not competitive intelligence) are currently experiencing a decline, while AI is experiencing a come-back. This could be cyclical. Part of the decline is due to not adapting to new types of data (e.g. unstructured text) that require engineering or data science techniques to process and extract value.

- Data Analysis. This is the new term for business statistics since at least 1995, and it covers a large spectrum of applications including fraud detection, advertising mix modeling, attribution modeling, sales forecasts, cross-selling optimization (retails), and user segmentation, churn analysis, computing long-time value of a customer and cost of acquisition, and so on. Except in big companies, data analyst is a junior role; 0these practitioners have a much more narrow knowledge and experience than data scientists, and they lack (and don't need) business vision. They are detail-oriented and report to managers such as data scientists or director of analytics, in big companies, someone with a job title such as data analyst III might be very senior, yet they usually are specialized and lack the broad knowledge gained by data scientists working in a variety of companies large and small.

- Business Analytics. Same as data analysis, but restricted to business problems only. Tends to have a bit more of a financial, marketing or ROI flavor. Popular job titles include data analyst and data scientist, but not business analyst (as already mentioned in business intelligence entry for business intelligence, a different domain).