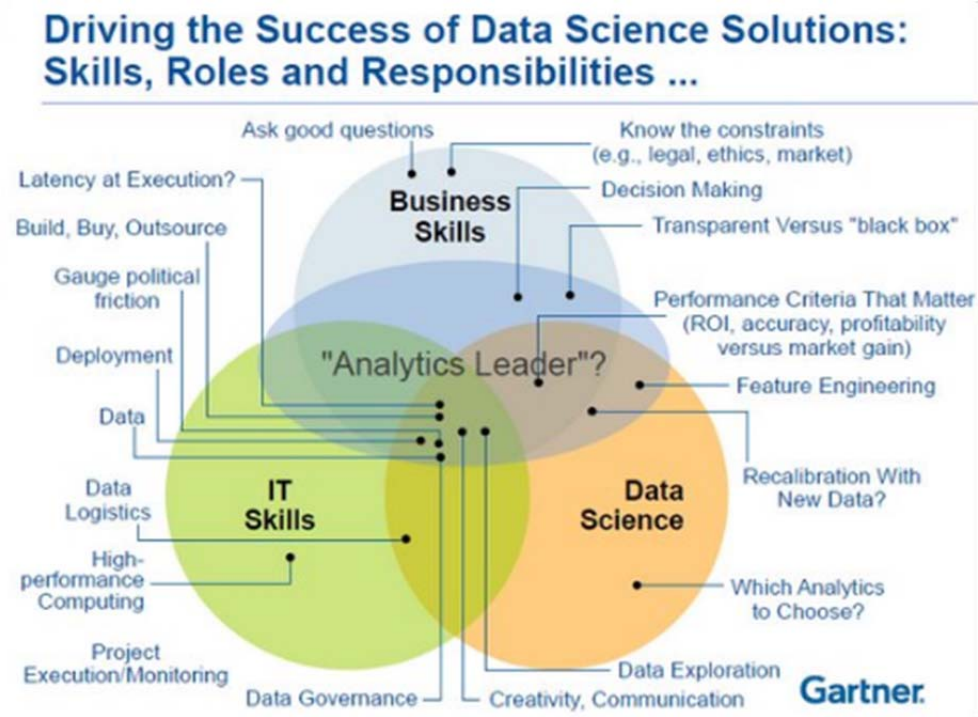
https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning

Difference between Machine Learning, Data Science, AI, Deep Learning, and Statistics

- Posted by Vincent Granville on January 2, 2017 at 8:30pm

In this article, I clarify the various roles of the data scientist, and how data science compares and overlaps with related fields such as machine learning, deep learning, AI, statistics, IoT, operations research, and applied mathematics. As data science is a broad discipline, I start by describing the different types of data scientists that one may encounter in any business setting: you might even discover that you are a data scientist yourself, without knowing it. As in any scientific discipline, data scientists may borrow techniques from related disciplines, though we have developed our own arsenal, especially techniques and algorithms to handle very large unstructured data sets in automated ways, even without human interactions, to perform transactions in real-time or to make predictions.



1. Different Types of Data Scientists

To get started and gain some historical perspective, you can read my article about 9 types of data scientists, published in 2014, or my article where I compare data science with 16 analytic disciplines, also published in 2014.

The following articles, published during the same time period, are still useful:

- Data Scientist versus Data Architect
- Data Scientist versus Data Engineer
- Data Scientist versus Statistician
- Data Scientist versus Business Analyst

More recently (August 2016) Ajit Jaokar discussed Type A (Analytics) versus Type B (Builder) data scientist:

- *The Type A Data Scientist can code well enough to work with data but is not necessarily an expert. The Type A data scientist may be an expert in experimental design, forecasting, modelling, statistical inference, or other things typically taught in statistics departments. Generally speaking though, the work product of a data scientist is not "p-values and confidence intervals" as academic statistics sometimes seems to suggest (and as it sometimes is for traditional statisticians working in the pharmaceutical industry, for example). At Google, Type A Data Scientists are known variously as Statistician, Quantitative Analyst, Decision Support Engineering Analyst, or Data Scientist, and probably a few more.*

- *Type B Data Scientist: The B is for Building. Type B Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data "in production." They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results). Source: click here.*

I also wrote about the ABCD's of business processes optimization where D stands for data science, C for computer science, B for business science, and A for analytics science. Data science may or may not involve coding or mathematical practice, as you can read in my article on low-level versus high-level data science. In a startup, data scientists generally wear several hats, such as executive, data miner, data engineer or architect, researcher, statistician, modeler (as in predictive modeling) or developer.

While the data scientist is generally portrayed as a coder experienced in R, Python, SQL, Hadoop and statistics, this is just the tip of the iceberg, made popular by data camps focusing on teaching some elements of data science. But just like a lab technician can call herself a physicist, the real physicist is much more than that, and her domains of expertise are varied: astronomy, mathematical physics, nuclear physics (which is borderline chemistry), mechanics, electrical engineering, signal processing (also a sub-field of data science) and many more. The same can be said about data scientists: fields are as varied as bioinformatics, information technology, simulations and quality control, computational finance, epidemiology, industrial engineering, and even number theory.

In my case, over the last 10 years, I specialized in machine-to-machine and device-to-device communications, developing systems to automatically process large data sets, to perform automated transactions: for instance, purchasing Internet traffic or automatically generating content. It implies developing algorithms that work with unstructured data, and it is at the intersection of AI (artificial intelligence,) IoT (Internet of things,) and data science. This is referred to as deep data science. It is relatively math-free, and it involves relatively little coding (mostly API's), but it is quite data-intensive (including building data systems) and based on brand new statistical technology designed specifically for this context.

Prior to that, I worked on credit card fraud detection in real time. Earlier in my career (circa 1990) I worked on image remote sensing technology, among other things to identify patterns (or shapes or features, for instance lakes) in satellite images and to perform image segmentation: at that time my research was labeled as computational statistics, but the people doing the exact same thing in the computer science department next door in my home university, called their research artificial intelligence. Today, it would be called data science or artificial intelligence, the sub-domains being signal processing, computer vision or IoT.

Also, data scientists can be found anywhere in the lifecycle of data science projects, at the data gathering stage, or the data exploratory stage, all the way up to statistical modeling and maintaining existing systems.

2. Machine Learning versus Deep Learning

Before digging deeper into the link between data science and machine learning, let's briefly discuss machine learning and deep learning. Machine learning is a set of algorithms that train on a data set to make predictions or take actions in order to optimize some systems. For instance, supervised classification algorithms are used to classify potential clients into good or bad prospects, for loan purposes, based on historical data. The techniques involved, for a given task (e.g. supervised clustering), are varied: naive Bayes, SVM, neural nets, ensembles, association rules, decision trees, logistic regression, or a combination of many. For a detailed list of algorithms, click here. For a list of machine learning problems, click here.

All of this is a subset of data science. When these algorithms are automated, as in automated piloting or driver-less cars, it is called AI, and more specifically, deep learning. Click here for another article comparing machine learning with deep learning. If the data collected comes from sensors and if it is transmitted via the Internet, then it is machine learning or data science or deep learning applied to IoT.

Some people have a different definition for deep learning. They consider deep learning as neural networks (a machine learning technique) with a deeper layer. The question was asked on Quora recently, and below is a more detailed explanation (source: Quora)

- *AI (Artificial intelligence) is a subfield of computer science, that was created in the 1960s, and it was (is) concerned with solving tasks that are easy for humans, but hard for computers. In particular, a so-called Strong AI would be a system that can do anything a human can (perhaps without purely physical things). This is fairly generic, and includes all kinds of tasks, such as planning, moving around in the world,*

recognizing objects and sounds, speaking, translating, performing social or business transactions, creative work (making art or poetry), etc.

- *NLP (Natural language processing) is simply the part of AI that has to do with language (usually written).*
- *Machine learning is concerned with one aspect of this: given some AI problem that can be described in discrete terms (e.g. out of a particular set of actions, which one is the right one), and given a lot of information about the world, figure out what is the "correct" action, without having the programmer program it in. Typically some outside process is needed to judge whether the action was correct or not. In mathematical terms, it's a function: you feed in some input, and you want it to to produce the right output, so the whole problem is simply to build a model of this mathematical function in some automatic way. To draw a distinction with AI, if I 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.*
- *Deep learning is one kind of machine learning that's very popular now. It involves a particular kind of mathematical model that can be thought of as a composition of simple blocks (function composition) of a certain type, and where some of these blocks can be adjusted to better predict the final outcome.*

What is the difference between machine learning and statistics?

This article tries to answer the question. The author writes that statistics is machine learning with confidence intervals for the quantities being predicted or estimated. I tend to disagree, as I have built engineer-friendly confidence intervals that don't require any mathematical or statistical knowledge.

3. Data Science versus Machine Learning

Machine learning and statistics are part of data science. The word *learning* in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes or supervised clustering. But not all techniques fit in this category. For instance, unsupervised clustering - a statistical and data science technique - aims at detecting clusters and cluster structures without any a-priori knowledge or training set to help the classification algorithm. A human being is needed to label the clusters found. Some techniques are hybrid, such as semi-supervised classification. Some pattern detection or density estimation techniques fit in this category.

Data science is much more than machine learning though. Data, in data science, may or may not come from a *machine* or mechanical process (survey data could be manually collected, clinical trials involve a specific type of small data) and it might have nothing to do with *learning* as I have just discussed. But the main difference is the fact that data science covers the whole spectrum of data processing, not just the algorithmic or statistical aspects. In particular, data science also covers

- data integration
- distributed architecture
- automating machine learning
- data visualization
- dashboards and BI
- data engineering
- deployment in production mode
- automated, data-driven decisions

Of course, in many organisations, data scientists focus on only one part of this process. To read about some of my original contributions to data science, click here.