» **Post Jobs** | Employer Login

DATAJOBS

Rarefied **talent** in big data technology and analytics

» **Home** » **Data Science Jobs / Analytics** » **Data Technology Jobs** » **About DataJobs.com** » **Big Data Knowledge Repo**

What is Data Science?

What is analytics? What is a data scientist?

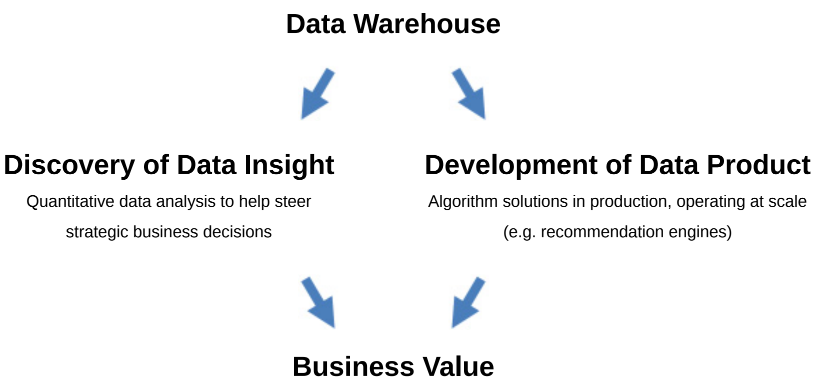
» Posted by *Frank Lo*

"We have lots of data – now what?"

(How can we unlock real value from our data?)

Data science is a multidisciplinary blend of **data inference, algorithmm development, and technology** in order to solve analytically complex problems.

At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it. Advanced capabilities we can build with it. Data science is ultimately about using this data in creative ways to generate business value:



Data Warehouse

Discovery of Data Insight

Quantitative data analysis to help steer strategic business decisions

Development of Data Product

Algorithm solutions in production, operating at scale (e.g. recommendation engines)

Business Value

Data science – discovery of data insight

This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviors, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions. For example:

- Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.
- Target identifies what are major customer segments within it's base and the unique shopping behaviors within those segments, which helps [illegible]de messaging to different market audiences.
- Proctor & Gamble utilizes tin[illegible]s models to more clearly understand future demand, which help plan for production levels more optimally.

How do data scientists mine out insights? It starts with data exploration. When given a challenging question, data scientists become detectives. They investigate leads and try to understand pattern or characteristics within the data. This requires a big dose of analytical creativity.

Then as needed, data scientists may apply quantitative technique in order to get a level deeper – e.g. inferential models, segmentation analysis, time series forecasting, synthetic control experiments, etc. The intent is to scientifically piece together a forensic view of what the data is really saying.

This data-driven insight is central to providing strategic guidance. In this sense, data scientists act as consultants, guiding business stakeholders on how to act on findings.

Data science – development of data product

Post or Tweet this Article

Share Tweet Share

Read about Big Data

- What is Data Science?
- What is Hadoop and NoSQL?
- State of Big Data Recruiting
- Big Data Salary: An Inside Look

Expert's Corner

For Big Data specialists only:

- Data Science – Math and Tech
- Data Warehousing – Platforms

Big Data Job Market

Currently in very high demand:

- Data Scientists
- Analytics Managers/Directors
- Business Analysts
- BI Developers
- Big Data Architects
- NoSQL DBAs
- Hadoop Experts

A "data product" is a technical asset that: (1) utilizes data as input, and (2) processes that data to return algorithmically-generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data. Here are some examples of data products:

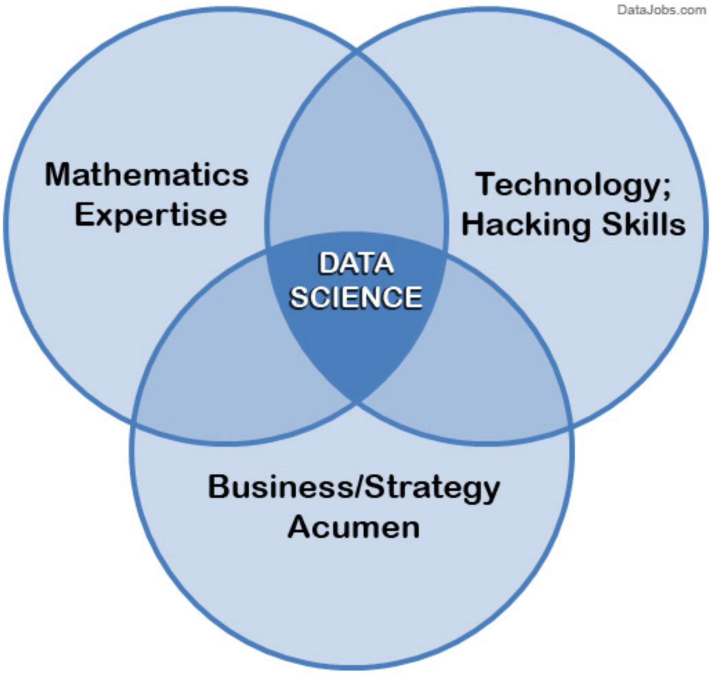
- Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you.
- Gmail's spam filter is data product – an algorithm behind the scenes processes incoming mail and determines if a message is junk or not.
- Computer vision used for self-driving cars is also data product – machine learning algorithms are able to recognize traffic lights, other cars on the road, pedestrians, etc.

This is different from the "data insights" section above, where the outcome to that is to perhaps provide advice to an executive to make a smarter business decision. In contrast, a data product is technical functionality that encapsulates an algorithm, and is designed to integrate directly into core applications. Respective examples of applications that incorporate data product behind the scenes: Amazon's homepage, Gmail's inbox, and autonomous driving software.

Data scientists play a central role in developing data product. This involves building out algorithms, as well as testing, refinement, and technical deployment into production systems. In this sense, data scientists serve as technical developers, building assets that can be leveraged at wide scale.

What is data science – the requisite skill set

Data science is a blend of skills in three major areas:



Mathematics Expertise

At the heart of mining data insight and building data product is the ability to view the data through a quantitative lens. There are textures, dimensions, and correlations in data that can be expressed mathematically. Finding solutions utilizing data becomes a brain teaser of heuristics and quantitative technique. Solutions to many business problems involve building analytic models grounded in the hard math, where being able to understand the underlying mechanics of those models is key to success in building them.

Also, a misconception is that data science all about statistics. While statistics is important, it is not the only type of math utilized. First, there are two branches of statistics – classical statistics and Bayesian statistics. When most people refer to *stats* they are generally referring to *classical stats*, but knowledge of both types is helpful. Furthermore, many inferential techniques and machine learning algorithms lean on knowledge of linear algebra. For example, a popular method to discover hidden characteristics in a data set is SVD, which is grounded in matrix math and has much less to do with classical stats. Overall, it is helpful for data scientists to have breadth and depth in their knowledge of mathematics.

Technology and Hacking

First, let's clarify on that we are *not* talking about hacking as in breaking into computers. We're referring to the tech programmer subculture meaning of hacking – i.e., creativity and ingenuity in using technical skills to build things and find clever solutions to problems.

Why is hacking ability important? Because data scientists utilize *technology* in order to wrangle enormous data sets and work with complex algorithms, and it requires tools far more sophisticated than Excel. Data scientists need to be able to code — prototype quick solutions, as well as integrate with complex data systems. Core languages associated with data science include SQL, Python, R, and SAS. On the periphery are Java, Scala, Julia, and others. But it is not just knowing language fundamentals. A hacker is a technical ninja, able to creatively navigate their way through technical challenges in order to make their code work.

Along these lines, a data science hacker is a solid algorithmic thinker, having the ability to break down messy problems and recompose them in ways that are solvable. This is critical because data scientists operate within a lot of algorithmic complexity. They need to have a strong mental comprehension of high-dimensional data and tricky data control flows. Full clarity on how all the pieces come together to form a cohesive solution.

Strong Business Acumen

It is important for a data scientist to be a **tactical business consultant**. Working so closely with data, data scientists are positioned to learn from data in ways no one else can. That creates the responsibility to translate observations to shared knowledge, and contribute to strategy on how to solve core business problems. This means a core competency of data science is using data to cogently tell a story. No data-puking – rather, present a cohesive narrative of problem and solution, using data insights as supporting pillars, that lead to guidance.

Having this business acumen is just as important as having acumen for tech and algorithms. There needs to be clear alignment between data science projects and business goals. Ultimately, the value doesn't come from data, math, and tech itself. It comes from leveraging all of the above to build valuable capabilities and have strong business influence.

What is a data scientist – curiosity and training

The Mindset

A common personality trait of data scientists is they are deep thinkers with *intense intellectual curiosity*. Data science is all about being inquisitive – asking new questions, making new discoveries, and learning new things. Ask data scientists most obsessed with their work what drives them in their job, and they will not say "money". The real motivator is being able to use their creativity and ingenuity to solve hard problems and constantly indulge in their curiosity. Deriving complex reads from data is beyond just making an observation, it is about uncovering "truth" that lies hidden beneath the surface. Problem solving is not a task, but an intellectually-stimulating journey to a solution. Data scientists are passionate about what they do, and reap great satisfaction in taking on challenge.

Training

There is a glaring misconception out there that you need a sciences or math Ph.D to become a legitimate data scientist. That view misses the point that data science is multidisciplinary. Highly-focused study in academia is certainly helpful, but doesn't guarantee that graduates have the full set of experiences and abilities to succeed. E.g. a Ph.D statistician may still need to pick up a lot of programming skills and gain business experience, to complete the trifecta.

In fact, data science is such a relatively new and rising discipline that universities have not caught up in developing comprehensive data science degree programs – meaning that no one can really claim to have "done all the schooling" to be become a data scientist. Where do[illegible]ch of the training come from? The unyielding intellectual curiosity of data scientists push them to be motivated a[illegible]acts, driven to self-learn the right skills, guided by their own determination.

Analytics and machine learning – how it ties to data science

There are a slew of terms closely related to data science that we hope to add some clarity around.

What is Analytics?

Analytics has risen quickly in popular business lingo over the past several years; the term is used loosely, but generally meant to describe critical thinking that is quantitative in nature. Technically, analytics is the "science of analysis" — put another way, the practice of analyzing information to make decisions.

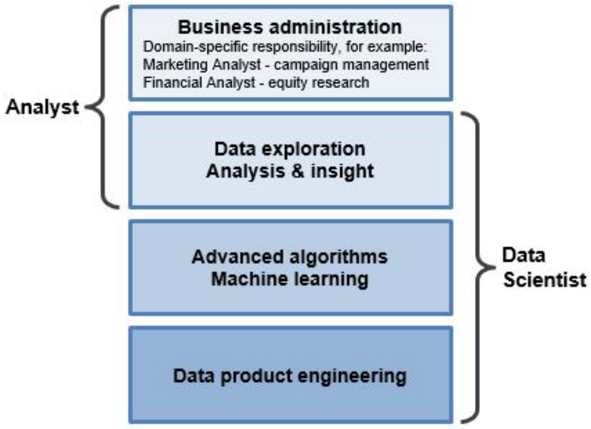
Is "analytics" the same thing as data science? Depends on context. Sometimes it is synonymous with the definition of data science that we have described, and sometimes it represents something else. A data scientist using raw data to build a predictive algorithm falls into the scope of analytics. At the same time, a non-technical business user interpreting pre-built dashboard reports (e.g. GA) is also in the realm of analytics, but does not cross into the skill set needed in data science. Analytics has come to have fairly broad meaning. At the end of the day, as long as you understand beyond the buzzword level, the exact semantics don't matter much.

What is the difference between an analyst and a data scientist?

"Analyst" is somewhat of an ambiguous job title that can represent many different types of roles (data analyst, marketing analyst, operations analyst, financial analyst, etc). What does this mean in comparison to data scientist?

- *Data Scientist:* Specialty role with abilities in math, technology, and business acumen. Data scientists work at the raw database level to derive insights and build data product.
- *Analyst:* This can mean a lot of things. Common thread is that analysts look at data to try to gain insights. Analysts may interact with data at both the database level or the summarized report level.

Thus, "analyst" and "data scientist" is not exactly synonymous, but also not mutually exclusive. Here is our interpretation of how these job titles map to skills and scope of responsibilities:



What is Machine Learning?

Machine learning is a term closely associated with data science. It refers to a broad class of methods that revolve around data modeling to (1) algorithmically make predictions, and (2) algorithmically decipher patterns in data.

- **Machine learning for making predictions** — Core concept is to use tagged data to train predictive models. *Tagged data* means observations where ground truth is already known. *Training models* means automatically characterizing tagged data in ways to predict tags for unknown data points. E.g. a credit card fraud detection model can be trained using a historical record of tagged fraud purchases. The resultant model estimates the likelihood that any new purchase is fraudulent. Common methods for training models range from basic regressions to complex neural nets. All follow the same paradigm known as *supervised learning*.
- **Machine learning for pattern discovery** — Another modeling paradigm known as *unsupervised learning* tries to surface underlying patterns and associations in data when no existing ground truth is known (i.e. no observations are tagged). Within this broad category of methods, the most commonly used are clustering techniques, which algorithmically detect what are the natural groupings that exist in a data set. For example, clustering can be used to programmatically learn the natural customer segments in a company's user base. Other unsupervised methods for mining underlying characteristics include: principal component analysis, hidden markov models, topic models, and more.

Not all machine learning methods fit r to the above two categories. For example, collaborative filtering is a type of recommendations algorithm with elements related to both supervised and unsupervised learning. Contextual bandits are a twist on supervised learning where predictions get adaptively modified on-the-fly using live feedback.

This wide-ranging breadth of machine learning techniques comprise an important part of the data science toolbox. It is up to the data scientist to figure out which tool to use in different circumstances (as well as how to use the tool correctly) in order to solve analytically open-ended problems.

What is Data Munging?

Raw data can be unstructured and messy, with information coming from disparate data sources, mismatched or missing records, and a slew of other tricky issues. Data munging is a term to describe the data wrangling to bring together data into cohesive views, as well as the janitorial work of cleaning up data so that it is polished and ready for downstream usage. This requires good pattern-recognition sense and clever hacking skills to merge and transform masses of database-level information. If not properly done, dirty data can obfuscate the 'truth' hidden in the data set and completely mislead results. Thus, any data scientist must be skillful and nimble at data munging in order to have accurate, usable data before applying more sophisticated analytical tactics.

Final word

For any company that wishes to enhance their business by being more data-driven, data science is the secret sauce. Data science projects can have multiplicative returns on investment, both from guidance through data insight, and development of data product. Though, hiring people who carry this potent mix of different skills is easier said than done. There is simply not enough supply of data scientists in the market to meet the demand (data scientist salary is sky high). Thus, when you manage to hire data scientists, nurture them. Keep them engaged. Give them autonomy to be their own architects in how to solve problems. This sets them up in the company to be highly motivated problem solvers, there to tackle the toughest analytical challenges.