From the Expert: What is Data Science?
Data science is an emerging and challenge discipline that attempts to extract knowledge or findings of value from data. Ultimately, this enables the creation of data products or data-driven decisions. The rise of data science significantly impacts and changes the way that enterprises operate.

There is no consensus meaning for the term data science. Generally, at the heart of data science are the skills that integrate techniques and theories from multidisciplinary fields such as mathematics, statistics, and computer science (software engineering, database, data mining, artificial intelligence, visualization) with domain-level expertise.

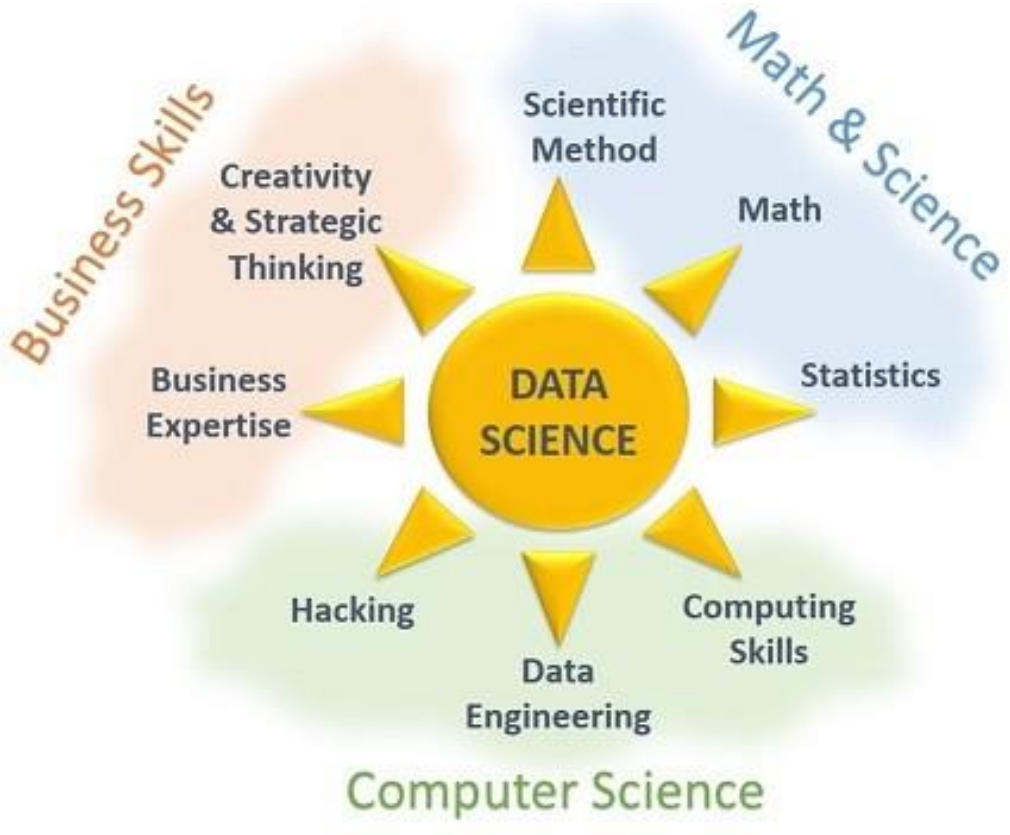

Figure 1: Data Science
(From: https://web.archive.org/web/20180822200023/http://www.datascienceguide.com/data-science.html)

According to Cloudera, the important characteristics of data products are 1) they build from data, 2) they generate more data, 3) new data can be used to improve existing data products.

Two well-known examples of data science products:

The recommendation engine from Amazon is an example of a data product that distinguishes Amazon from the traditional retail business. The recommender utilizes data collected from other users or products to understand uses preferences and drive more sales. Other data products include the finding of similar items and product advertising API's.

The search engine "PageRank" algorithm from Google combines text information on the web page with data outside the page such as the number of links pointing in/out from that page to markedly enhance the search results. Additional data products derived from the success of "PageRank" are Google Analytics and Adwords.

Other applications of data science:
Commerce: predict which shoppers will become repeat buyers, predict the risk of customer credit default, suggest further purchases to customers, and target advertisements to potential customers

Markets/Trading: prediction of future commodity prices on exchanges

Medicine: discover patterns in disease transmission, modeling hospital costs and efficiency

Social Sciences: suggestions of other people you may know, or shared traits, through modeling of social relationships

Linguistics: Find and impute missing words in a billion-word corpus

Weather Forecast: prediction of upcoming weather patterns for farmers

Engineering: detecting software plagiarism, locating mineral deposits

The Data Scientist
The practitioner who performs data science is known as a "data scientist". They use their skills and technology to work with data. Thorough understanding of the data is the key requirement in all fields of data science. To be successful in data science, a data scientist must know how to ask the right questions and know how to get the answers by applying the right expertise and tools to the questions at hand. Other important characteristics of data scientists include creativity and curiosity.

Why are Data scientists in Demand?
The amount of available data from government, trading data from the stock markets, satellite images, social media, e-mail and text messages, etc. has surged in recent years. Not only are more data generated, but data are also generated at a much faster rate than ever before. For example: Twitter processes hundreds of millions of messages daily, while Facebook users generate billions of comments and "likes".

Data are very valuable, so it should not be discarded. Fortunately, information storage technologies have kept pace while the price has decreased. Therefore, we now can afford to retain the data we produce. However, data that is stored has no useful value until it has been analyzed.

The success stories from the forerunner corporations such as Google, Yahoo, Facebook, and Amazon drive the demand for data scientists. These companies realized that data scientists are essential for their business success. Data scientists are value-adds to the organization.

Data scientists play critical roles from collecting the data to building the data products. The skill set that data scientists should have is displayed in Figure 2:

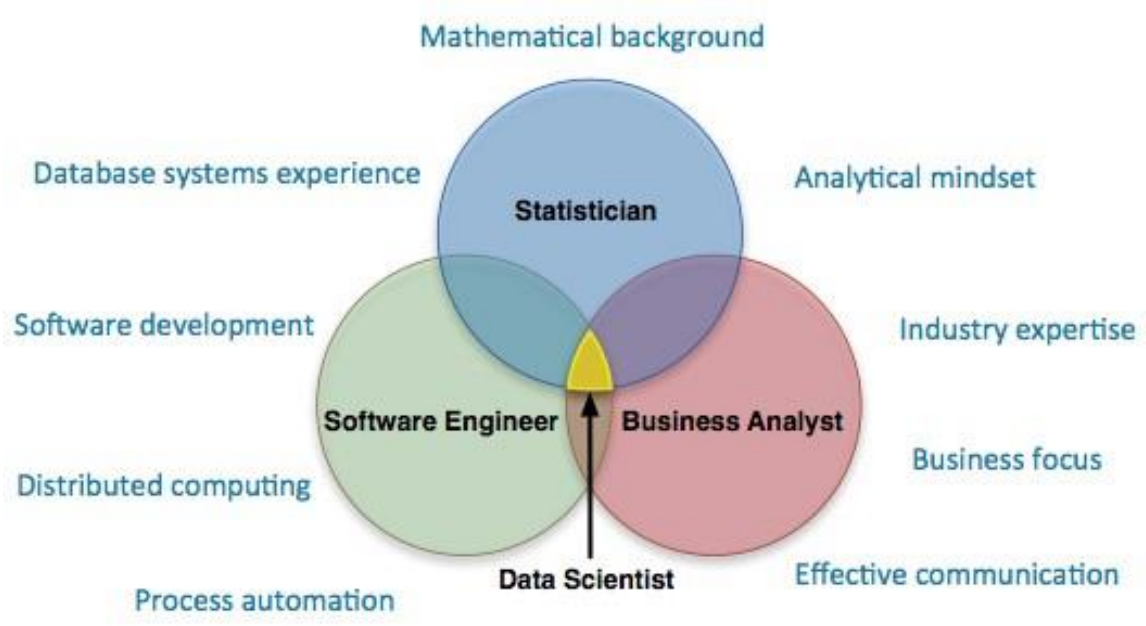

Figure 2: Data Scientist Skills

The role of the data scientist depends on the type of company, whether they are engineering driven, data-product driven (LinkedIn, Facebook) or commerce driven (amazon, Wal-Mart) etc. According to Patil (2011), the roles of data scientists fall into the following domains:

1. Decision Science and Business Intelligence

We all make decisions every day. Most of the decisions come from our intuition or rules of thumb, such as what should we prepare for dinner. However, complex questions that involve uncertainty, risk, multiple objectives, etc. may need more systematic methods to analyze the problem effectively; this is called "Decision Science". Data (from various

sources) have long been considered as the critical element for assisting in strategies for decision-making. After data analysis, metrics are needed to evaluate the choices, and finally reports are established to disseminate the findings.

2. Product and marketing analytics

Product analytics involve applications that are created to interact directly with customers, such as a product that suggests further purchases from Amazon: "People who viewed this item also viewed"; or a product that suggest friends or other connections, such as "People you may know". Many organizations attempt to build products in creative ways to attract new customers/users or retain old ones. However, these products should be tested for their effectiveness.

3. Fraud, abuse, risk and security

Challenges for data scientists include: what data needed to be collected, data processing time to respond to an incident promptly, developing a model to close the limitations of the system, segmenting out bad users, and preventing the future fraud or attacks.

4. Data services and operations

The primary responsibilities of this team are on data stores, databases (data structures, data schemas), data warehouses, and the stability of the system. Defining the metrics is not the goal of this team. However, collaboration with other teams such as decision science team is very important.

5. Data engineering and infrastructure

The main focus of this team is on the technologies and tools for data ingestion, processing data, querying large databases, managing complex data flows, etc. The wellknown Big Data technologies under "Hadoop" often come into play here (Mapreduce, Hive, Pig, Oozie, etc.) In addition, the team also involves in monitoring, deploying, and implementing the technologies for the system.

6. Organizational and reporting alignment

The function and structure of the data science team depends on many factors such as size, scale, and focus of the company. For example, in a startup company, a few people may have to do many tasks such as infrastructure, operations, analysis, and security. As the size of the organization grows, people do only specific tasks. Since, everyone in the company is involved with the data, it is essential for team members to learn and grow together by sharing experiences with others.

The data scientist should be able to work individually or collaborative as part of the team such as working with marketing and business analysts.

In order to get answers, the data scientists should know where to acquire the data, how to clean and transform it so that they can analyze the data and interpret the findings.

The critical question is how to employ the vast amounts of data, which may come from multiple sources, effectively?

It is important to mention that Data science does not necessarily have to deal with big data. Data Science may involve in solving complex data or even simple data with different perspectives.

Project Lifecycle

A typical data science project is an iterated process consisting of these steps:

1) Define a problem
2) Identify the desired outcome
3) Determine which data are needed
4) Evaluate possible solutions
5) Measure effectiveness
6) Make improvements
7) Communicate results

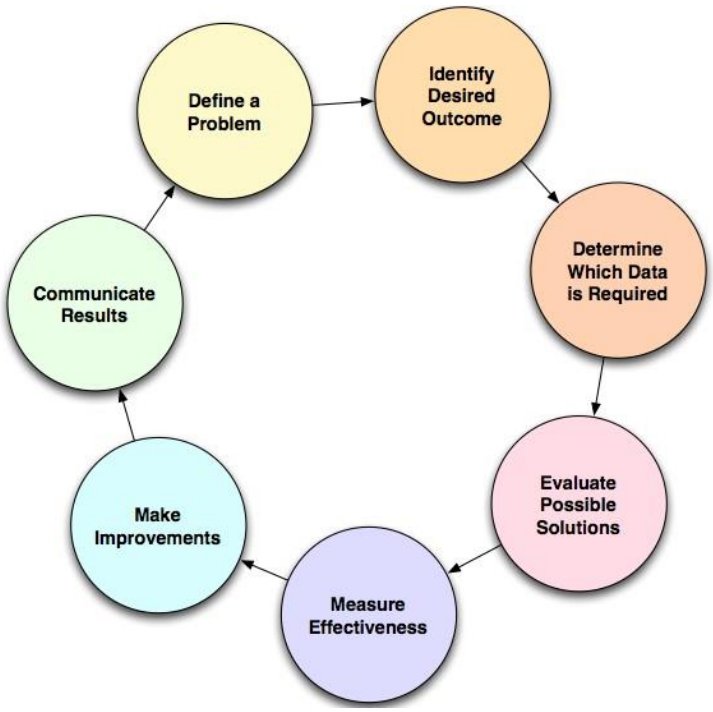

Figure 3: Data Science Project Lifecycle

1. Define a problem

In business, the problems frequently relate to cost, revenue, or customer behaviors. For example: customer services cost up 15% this year, customers do not renew their memberships, what type of movies have the highest watching percentage?

2. Identify the desired outcome
After the problem is defined, you then establish the preferred goal. For example, reduce the customer services cost by 10% at the end of the year, increase the membership renew rate by 20%, etc.

3. Determine which data is needed
You have to determine what the needed data are in order to solve the problem. Where can the data be obtained (internal, external sources)? What is the format or quality of the data?

4. Evaluate possible solutions
It is possible that there is more than one solution for the desired outcome. The first solution to pursue should be the simplest one. Usually, hypothesis testing and root cause analysis are used to evaluate the performance of the possible solutions.

5. Measure effectiveness
We are only interested to measure what are essential. To measure, we need to know what properties to measure and which methods can be used to compare the measurements.

6. Make improvements
Is there any room for improvement? This could be adding more data, check the validity of the hypothesis. After the solution is implemented, then evaluate it again.

7. Communicate the results
In business, dashboards are usually employed as tools for communicating the results. The results can be shown in the form of statistics, or visualizations. The data scientist must tell the story and insight in the findings. Telling the story should be concise and specific to the audience. All the technical terms should be avoid as much as possible.

Brief introduction to Big Data
Massive data are generated and collected by large corporations, institutes, and government, thus fostering the revolution of Big data. Big data is usually defined using three "V" aspects, namely: volume, velocity and variety.

Volume – huge amounts of data have been generated, such as data from mobile devices and on-line transactions

Variety – data is in diverse forms such as databases, email messages, word processing documents, spreadsheet, webpages, photo, video, and audio files

Velocity – incoming data with a very rapid speed, such as trading volumes from Wall Street

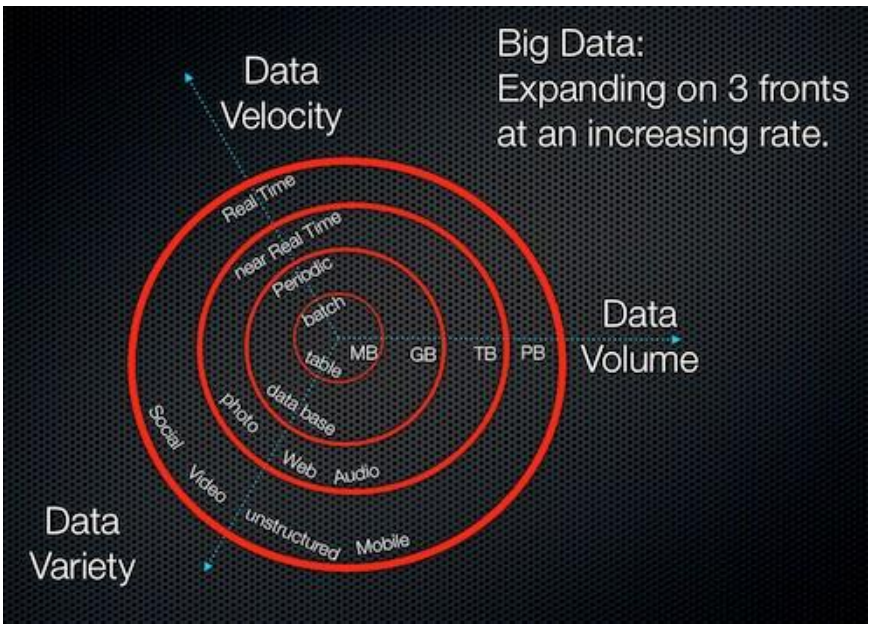

Figure 4: Three V's. Concept for Big data
(From: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data)

In short, Big data can be described as data that surpasses the capacity of traditional processing systems, moving with a very rapid rate, and cannot fit in conventional database architectures (Patil, 2011). As a result, we need different innovative tools to process this data. Tools and technology will be discussed in later contents.

The Big data revolution is closely tied to data science. The values of big data derive from data analytics and new product creation.

Ethics

Data science leverages the power of the data. The capability to discern the meaning from data is invaluable. For example, business can target products to the customer interests from their buying behaviors. Physicians can predict the future risk of patients for diseases such as cancer and Alzheimer's or their health status.

Along with these positive impacts, there are unexpected costs and liability if the implications of data usage are not handled appropriately. For example, Target predicted a teen's pregnancy from her buying habits and sent mailing advertisements for baby products. As a result, the teen pregnancy was revealed to her father after he saw these advertisements sent to her daughter, though he initially was furious at Target for their presumption.
As another example, recently a EU court ruled against Google for the "right to be forgotten", so that people of the EU can ask Google to delete their personally identifying and tracked information.

The data science technology is growing immensely, but the issue such as ethics is still lagging behind. As data analytics get smarter, issues such as privacy become of greater concern. What are the limitations in using data? What should be disclosed? Data science misconduct could do great damage to individuals and society. Therefore, it is important for the data science professionals to hold accountability, make ethical decisions, and maintain self-regulations in case of doubts.

The Association of Computing Machinery (ACM) attempt to address the issue regarding privacy with this resolution:

> "It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals, This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals."

The ethical standards for data science professionals are still at its early stage, but are being developed. One proposed draft for a code of conduct can be found here: https://web.archive.org/web/20160313135935/http://www.rosebt.com/blog/open-for-comment-proposed-data-science-code-of-professional-conduct. There are other related codes of conduct we will cover in MSDS640.

References

ACM Code of Ethics (October 6, 1992). Retrieved from https://www.acm.org/code-of-ethics.

Statistics:

Coursera (n.d.). *Statistics: making sense of data*. Retrieved from https://www.coursera.org/learn/basic-statistics

MITOPENCOURSEWARE (n.d.). *Statistical Thinking and Data Analysis*. Retrieved from https://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/

Data Science:

Loukides, M. (June 2, 2010). What is data science? Retrieved from http://radar.oreilly.com/2010/06/what-is-data-science.html

Patil, D.J. (2011). *Building Data Science Teams*. O'Reilly Media. Inc. Retrieved from http://www.oreilly.com/data/free/building-data-science-teams.csp (free ebook).