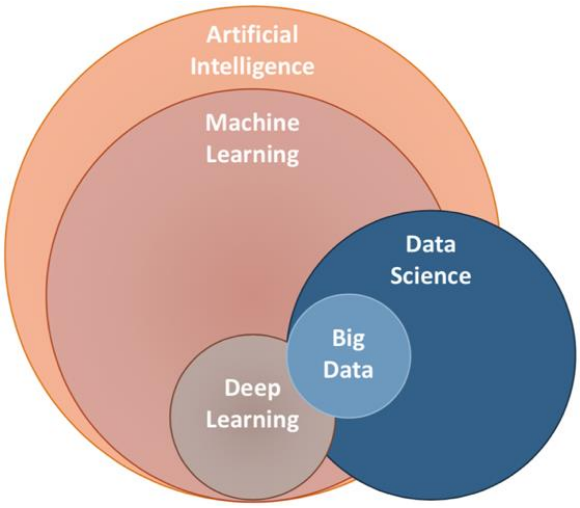
1.1 Introduction to Data Science

Data science is an interdisciplinary field that use scientific methods, processes, algorithms, programming skills, and systems to extract knowledge and insights from noisy, structured, and unstructured data, as well as to apply that knowledge and actionable insights to a variety of application domains.

Data science is a discipline that combines domain knowledge, programming abilities, and math and statistics knowledge to extract useful insights from data.



Artificial Intelligence (AI) is the *simulation of human intelligence* in machines [robots] that have been trained to think and act like humans. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

Machine Learning (ML) is a discipline of Artificial Intelligence and computer science that focuses on the use of data and algorithms to mimic the way *humans learn and improve accuracy over time*. Machine learning is a method of data analysis that automates analytical model building. ML is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

Deep Learning (DL) is an area of machine learning that deals with artificial neural networks (ANNs) that are inspired by the structure and function of the brain and use numerous layers of processing to *extract progressively higher-level features from data*. These neural networks attempt to simulate the behavior of the human brain and matching its ability to "learn" from large amounts of data.

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain and allowing it to "learn" from large amounts of data.

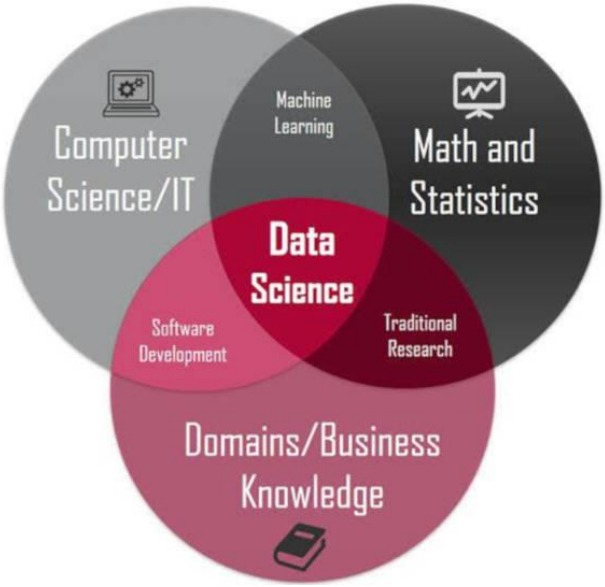
In essence, Data Science is all about:

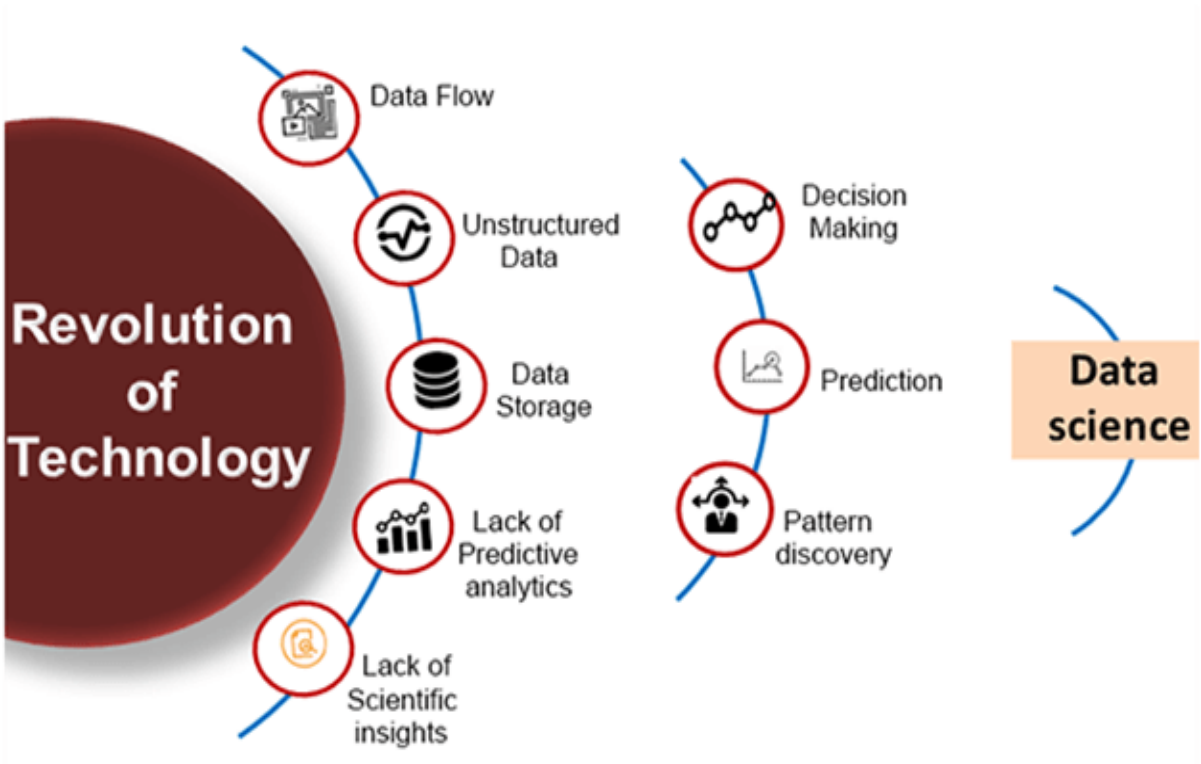
a) Asking the correct questions and analysing the raw data.

b) Modelling the data using various complex and efficient algorithms.

c) Visualizing the data to get a better perspective.

d) Understanding the data to make better decisions and finding the result.





Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

1.2 Introduction to Big Data

Big Data is a field that analyses computationally large and complex data sets to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. that are too large or complicated for typical data-processing application software to handle. Big data is a term for any collection or handling of data sets so large or complex that it becomes difficult to process them using traditional data management techniques such as RDBMS. Big Data calls for specialized techniques to extract the insights.

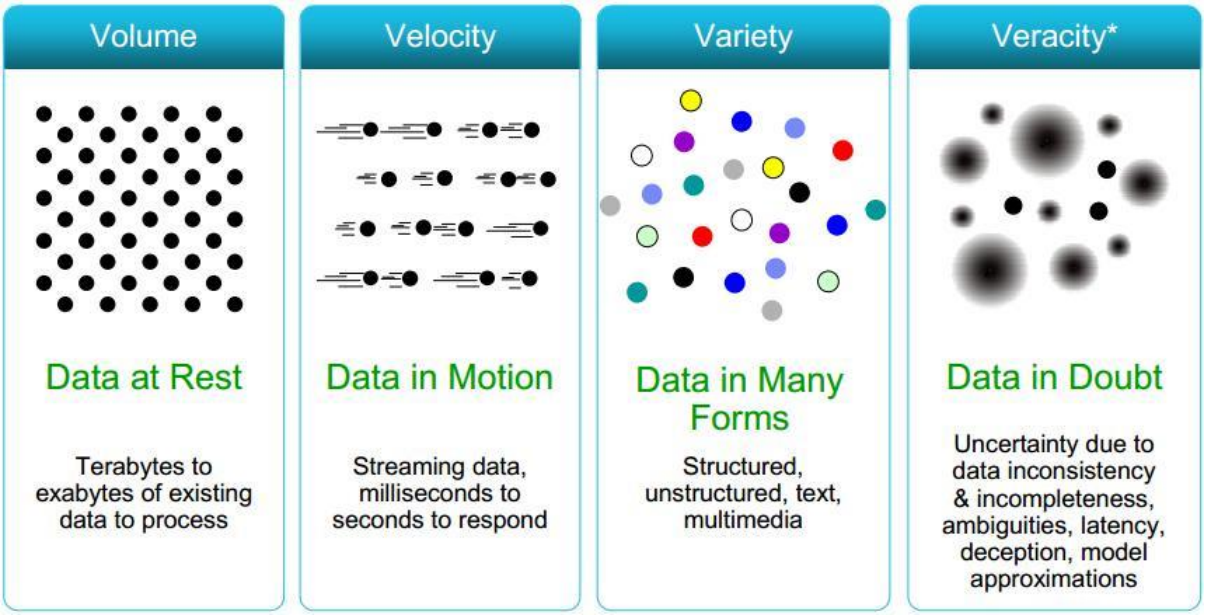
The characteristics of big data are referred to as the four Vs:

a) **Volume** - How much data is there?

b) **Velocity** - At what speed is new data generated?

c) **Variety** - How diverse are different types of data? [Structured, Unstructured, machine generated, Streaming, Graph-based data, etc…]

d) **Veracity** - How accurate is the data?



The challenges of handling Big Data can be seen in almost every aspect:

a) data capture is the process of extracting information from paper or electronic documents and converting it into data for key systems

b) data curation is the process of creating, organizing and maintaining data sets so they can be accessed and used by people looking for information

c) data storage how it is stored and how can it be efficiently used accessed

d) data search Data Science is used for internet search to provide relevant results and fast

e) data sharing is the practice of making data used for research available to other investigators.

f) data transfer to the secure exchange of large files between systems or organizations.

g) data visualization

is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, patterns etc . . in data.

1.3 Difference between Data Science and Big Data

DATA SCIENCE	BIG DATA
Data Science uses statistical techniques to analyse and deduce insights from raw data.	Big Data is a collection of data that can be analysed for Business purposes.
The idea behind Data Science is to identify patterns, discover relationships among data, and to make sense of the raw data.	Big Data is data that can be used to analyze insights which results in informed decisions and strategic business moves.
Data Science is an area / domain.	Big Data is a technique to collect, maintain and process the huge information.
It is about collection, processing, analyzing and utilizing of data into various operations. It is more conceptual.	It is about extracting the vital and valuable information from huge amount of the data.
It is a field of study just like the Computer Science, Applied Statistics or Applied Mathematics.	It is a technique of tracking and discovering of trends of complex data sets.
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e., by extracting only important information from the huge data within existing traditional aspects.
It is a super set of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics and many more techniques.	It is a subset of Data Science as mining activities which is in a pipeline of the Data science.
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
It broadly focuses on the science of the data.	It is more involved with the processes of handling voluminous data.
Data Science refers to Quality of data.	Big Data refers Quantity of data.
Tools include SAS, R, Python etc…	Tools include Hadoop, Spark, Flink etc…

Data science is an umbrella term that encompasses all the techniques and tools used during the life cycle stages of useful data. Data Science identifies patterns, discover relationships, and make sense of raw data. It also uses statistical techniques to analyze and deduce insights from raw data. *Big data* is a collection of data that can be analyzed for business purposes. It is the process of analyzing insights from data which result in informed decisions and strategic business moves.

1.4 Big Data and Data Science hype

The Hype related to Big Data and Data Science include:

a) There's a lack of definitions around the most basic terminology. What is "Big Data" anyway? What does "data science" mean? What is the relationship between Big Data and data science? Is data science the science of Big Data? Is data science only the stuff going on in companies like Google and Facebook and tech companies?
b) From the way the media describes it, machine learning algorithms were just invented last week, and data was never "big" until Google came along. This doesn't mean that there's not new and exciting stuff going on, but we think it's important to show some basic respect for everything that came before.
c) The hype is crazy - people throw around tired phrases like "Masters of the Universe" to describe data scientists?
d) Statisticians already feel that they are studying and working on the "Science of Data."
e) Although there might be truth in there, that doesn't mean that the term "data science" itself represents nothing, but of course what it represents may not be science but more of a craft.

1.5 Getting Past the Hype

We were simply facing the difference between academic statistics and industry statistics. Sure, there's is a difference between industry and academia. But does it really have to be that way? Why do many courses in school have to be so intrinsically out of touch with reality? Even so, the gap doesn't simply represent a difference between industry statistics and academic statistics. The general experience of data scientists is that, at their job, they have access to a larger body of knowledge and methodology, as well as a process, which we now define as the data science process, that has foundations in both statistics and computer science.

Around all the hype, in other words, there is a ring of truth: this is something new. But at the same time, it's a delicate, promising idea at real risk of being rejected prematurely.

Why Now?

We have massive amounts of data about many aspects of our lives, and, simultaneously, an abundance of inexpensive computing power. Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions - all this is being tracked online, as most people know.

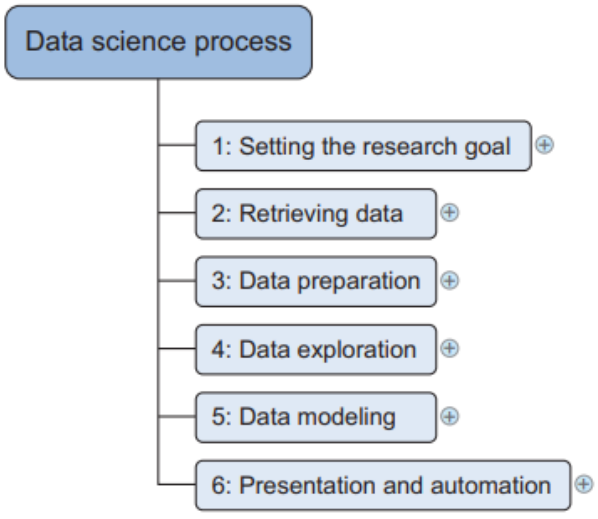
What people might not know is that the "datafication" of our *offline* behavior has started as well, mirroring the *online* data collection revolution. Put the two together, and there's a lot to learn about our behavior and, by extension, who we are as a species.

It's not just Internet data, though—it's finance, the medical industry, pharmaceuticals, bioinformatics, social welfare, government, education, retail, and the list goes on. There is a growing influence of data in most sectors and most industries. In some cases, the amount of data collected might be enough to be considered "big"; in other cases, it's not.

But it's not only the massiveness that makes all this new data interesting (or poses challenges). It's that the data itself, often in real time, becomes the building blocks of data products. On the Internet, this means Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, and so on. In finance, this means credit ratings, trading algorithms, and models. In education, this is starting to mean dynamic personalized learning and assessments coming out of places like Nptel and Coursera Academy. In government, this means policies based on data.

We're witnessing the beginning of a massive, culturally saturated feedback loop where our behavior changes the product, and the product changes our behavior. Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives. This wasn't true a decade ago. Considering the impact of this feedback loop, we should start thinking seriously about how it's being conducted, along with the ethical and technical responsibilities for the people responsible for the process. One goal of this book is a first stab at that conversation.

1.6 Data Science Process



1. Setting the research goal:

Data science is mostly used in the context of a company. When a company wants you to do a data science project, you must first create a project charter. This charter details what you'll be researching, how the firm will profit from it, what data and resources you'll need, a timeline, and deliverables.

2. Retrieving data:

The next stage is to gather information. In the project charter, you specified which data you require and where you may obtain it. This phase ensures that you can utilise the data in your software by verifying the data's existence, quality, and accessibility. Third-party firms can also supply data in a variety of formats, ranging from Excel spreadsheets to various sorts of databases.

3. Data preparation:

Data collection is an error-prone process; in this phase, you improve the data's quality and prepare it for usage in the following steps. *Data cleansing* removes erroneous values from a data source and inconsistencies among data sources, *Data integration* improves data sources by merging information from many data sources, and *Data transformation* guarantees that the data is in a format that can be used in your models.

4. Data exploration:

The goal of data exploration is to gain a better knowledge of your data. You want to know how variables interact with one another, how the data is distributed, and if there are any outliers. You mostly utilise descriptive statistics, visual approaches, and simple modelling to accomplish this. Exploratory Data Analysis (EDA) is a common shorthand for this step.

5. Data modeling or Model building:

Here, we employ models, domain knowledge, and insights about the data you discovered in the previous steps in this phase. You choose a technique from statistics, machine learning, operations research, and other domains. Selecting variables for the model, executing the model, and performing model diagnostics are all part of the iterative process of building a model.

6. Presentation and automation:

Finally, you report your findings to your company. These outcomes can take a variety of forms, from presentations to research reports. Because the business may want to leverage the insights you got in another project or enable an operational process to use the output from your model, you may need to automate the process execution.

1.7 Datafication

Datafication is a process of "*taking all aspects of life and turning them into data*." As examples, they mention that "Google's augmented-reality glasses datafies the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks." Datafication consider its importance with respect to people's intentions about sharing their own data. We are being datafied, or rather our actions are, and when we "like" someone or something online [web], we are intending to be datafied, or at least we should expect to be.

when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors, cameras, or Google glasses. *Once we datafy things, we can transform their purpose and turn the information into new forms of value.* who is "we" in that case? What kinds of value do they refer to? Mostly, given their examples, the "we" is the modelers and entrepreneurs making money from getting people to buy stuff, and the "value" translates into something like increased efficiency through automation.

1.8 The Current landscape of perspectives

Metamarket CEO Mike Driscoll's answer on "What is Data Science" in 2010:

- Data science is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.
- Data science is not merely hacking - because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics.
- Data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.
- Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.

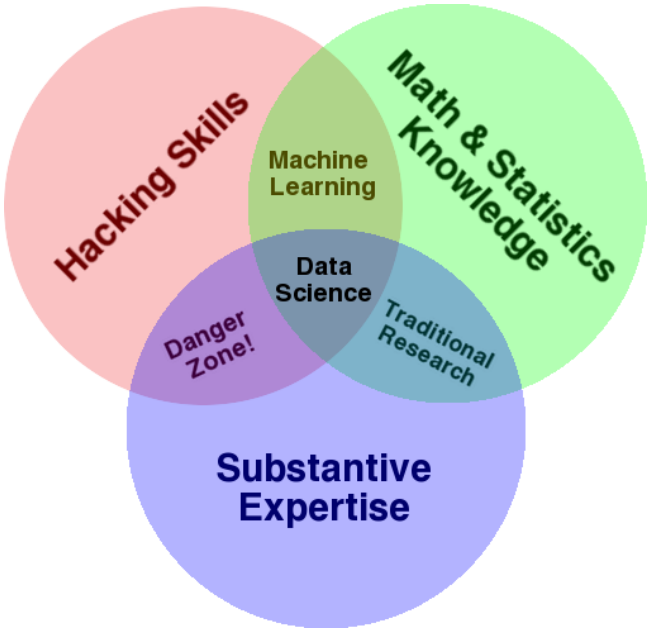
Nathan Yau's 2009 post on "Rise of the Data Scientist" includes:

- Statistics (traditional analysis you're used to thinking about)
- Data munging (parsing, scraping, and formatting data)
- Visualization (graphs, tools, etc.)

ASA President Nancy Geller's 2011 defends Statistics in Data Science:

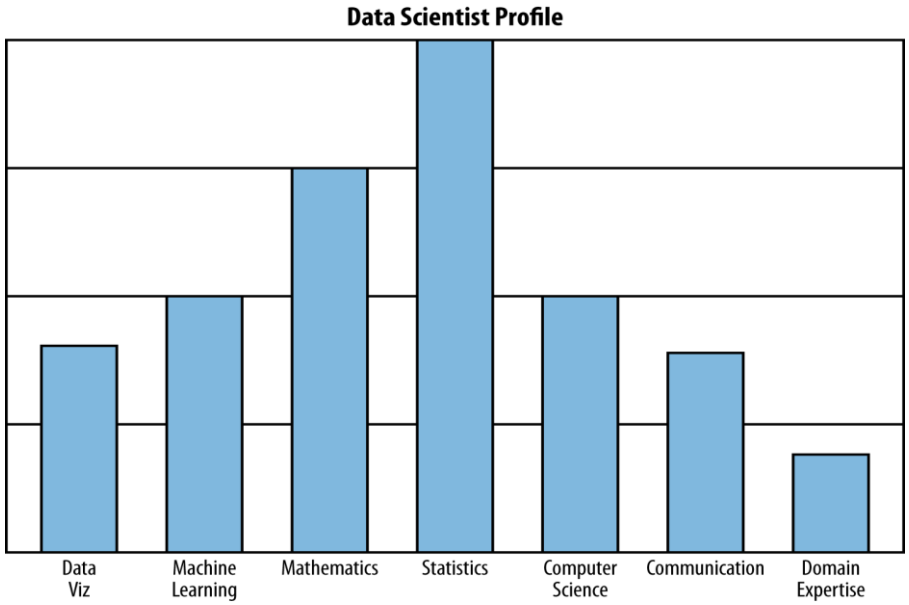
- *Statisticians* are the ones who make sense of the data overflow occurring in science, engineering, and medicine; that statistics provides methods for data analysis in all fields, from art history to zoology;

Driscoll refers to Drew Conway's Venn diagram of data science from 2010,

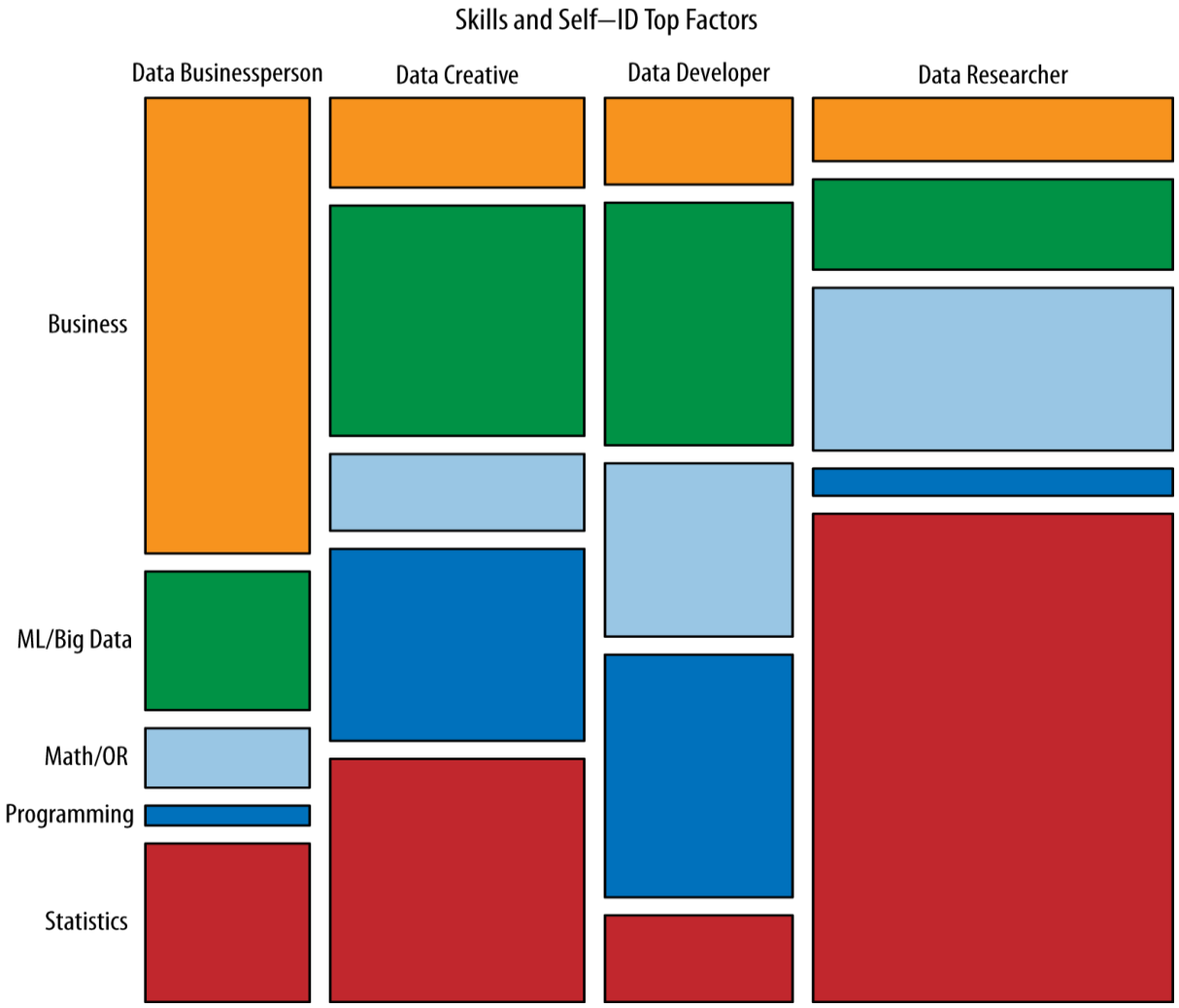


1.9 Skill sets needed

- Computer science
- Math
- Statistics
- Machine learning
- Domain expertise
- Communication and presentation skills
- Data visualization



Harlan Harris recently took a survey and used clustering to define subfields of data science,



1.10 Applications of Data Science

- Fraud and Risk Detection
- Healthcare
- Internet Search
- Targeted Advertising
- Website Recommendations
- Advanced Image Recognition
- Speech Recognition
- Airline Route Planning
- Gaming
- Augmented Reality