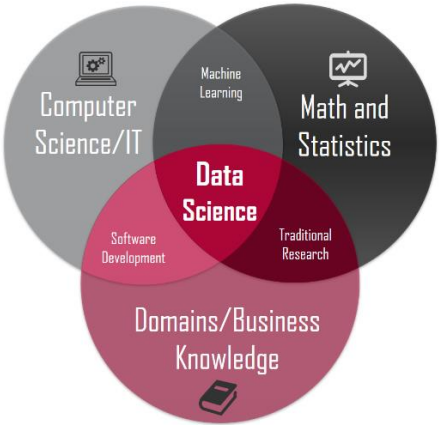
Chapter 1: Introduction

1.1 What Is Data Science?

- **Data science**:
  - The fusion of three components: Computer Science, Mathematics and Statistics, and domain/business knowledge.
  - Involves the collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon.

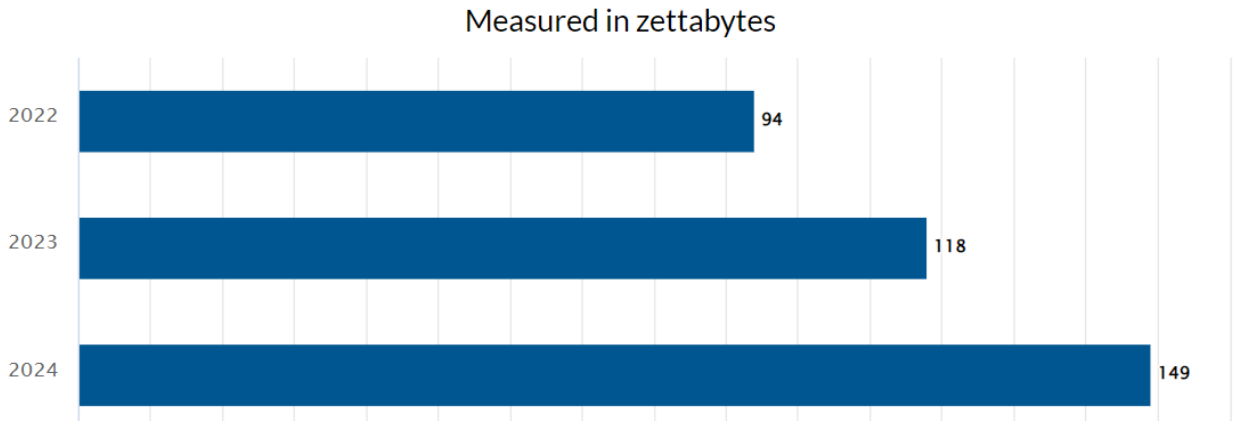


- Data sizes:

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

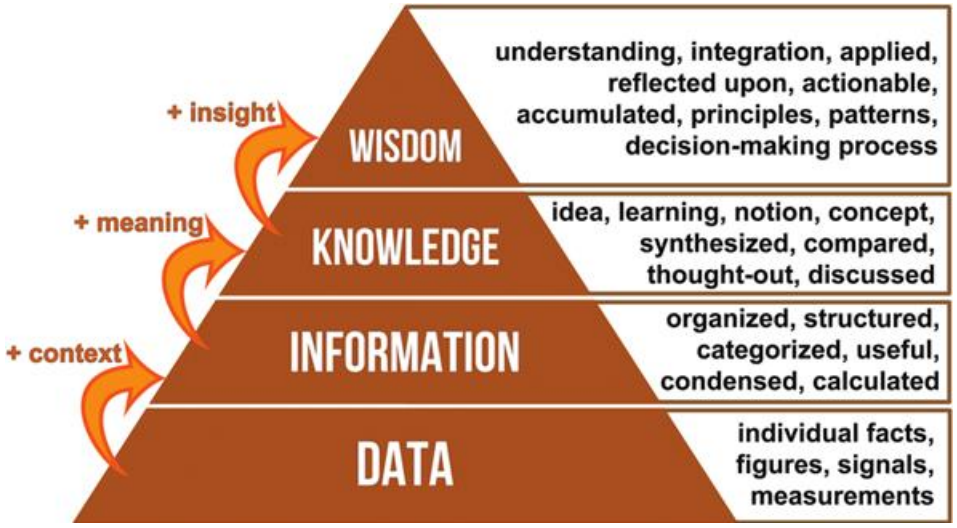
- We are drenched in **data**, and so many of our problems need to be solved using large amounts of data existing at personal and societal levels.

- Projected data volumes:

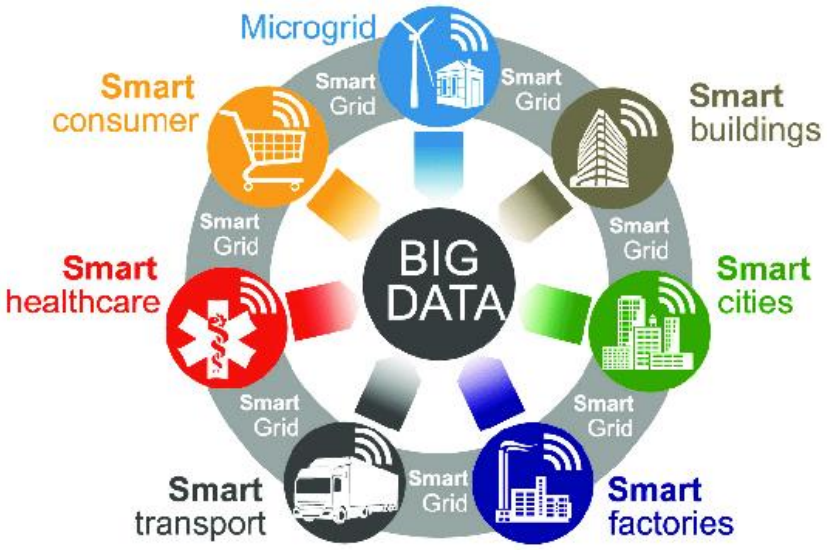


- **Data** vs. **Information** vs. Knowledge vs. **Wisdom**

    - Information consists of data, but data is not necessarily information.
    - Wisdom is knowledge, which in turn is information, which in turn is data, but, for example, knowledge is not necessarily wisdom.
    - Wisdom is a subset of knowledge, which is a subset of information, which is a subset of data.

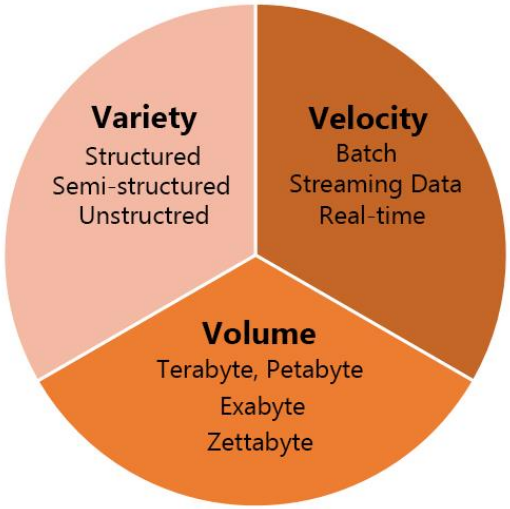


- **Big data** is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

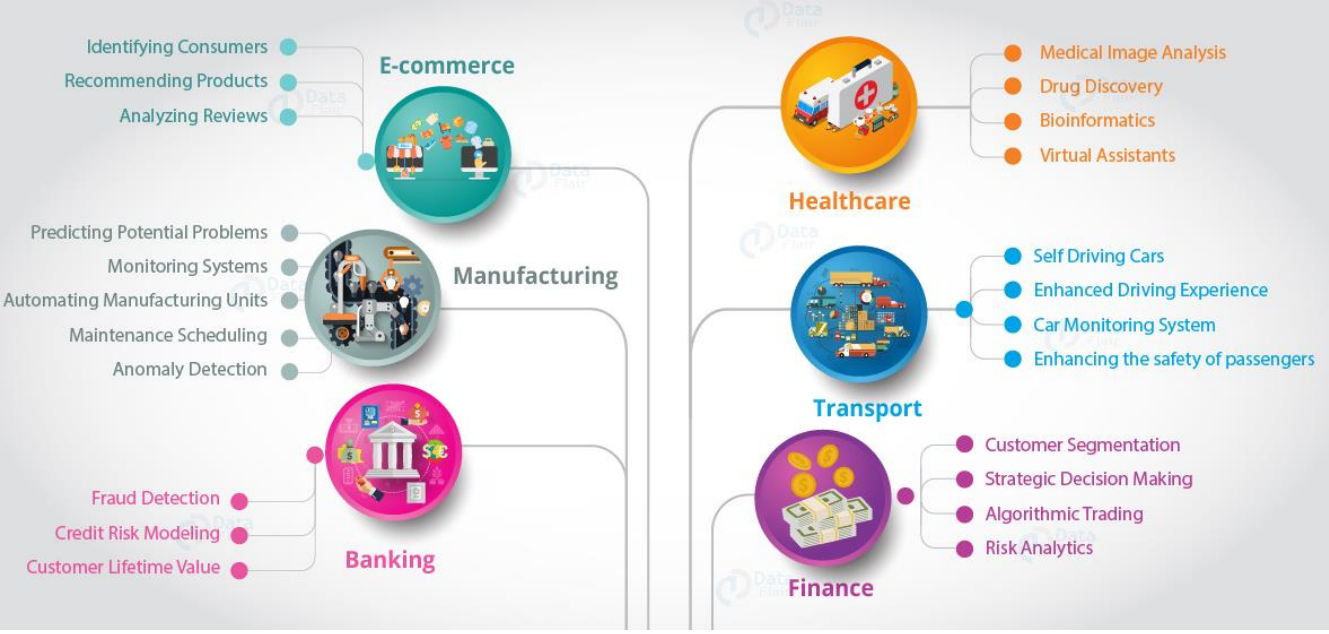


- Why is data science so important?
    - We have a lot of data and continue to generate a staggering amount of data at an unprecedented and ever-increasing speed
    - Analyzing **big data** wisely necessitates the involvement of competent and well-trained practitioners, and analyzing such data can provide actionable insights.

- The 3V Model:
    1. **Velocity**: The speed at which data is accumulated.
    2. **Volume**: The size and scope of the data.
    3. **Variety**: The massive array of data and types (structured and unstructured).



1.2 Where Do We See Data Science?



- Social media activity, mobile interactions, server logs, real-time market feeds, customer service records, transaction details, and information from existing databases combine to create a rich and complex conglomeration of information that experts must tackle.

- Data scientists capture and analyze new sources of data, build predictive models and run real-time simulations of events to obtain the information necessary to make accurate **predictions**.

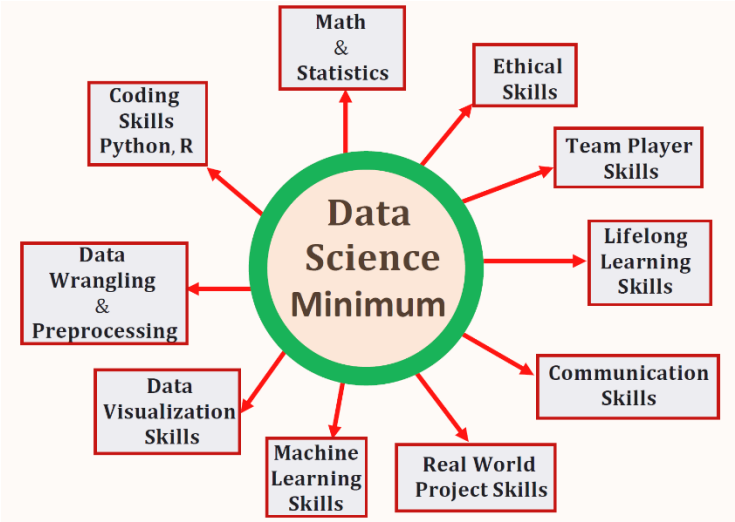
- Banks can manage their resources efficiently, and furthermore make smarter decisions through fraud detection, management of customer data, risk modeling, real-time predictive analytics, customer segmentation, etc.

- With the addition of technologies like the Internet of Things (IoT), data science has enabled the companies to predict potential problems, monitor systems and analyze the continuous stream of data.

- With the introduction of autonomy to vehicles through reinforcement learning, vehicle manufacturers are able to create intelligent automobiles. Furthermore, industries can create better logistical routes with the help of data science.

- In the medical image analysis, data science has created a strong sphere of influence for analyzing medical images such as X-rays, MRIs, CT-Scans, etc. Previously, doctors and medical examiners would have to manually search for clues in the medical images.

1.6 Skills for Data Science



- Needed skills:
  - Probability & Statistics
  - Calculus & Linear Algebra
  - Programming, Software Packages
  - Data Wrangling
  - Database Management
  - Data Visualization
  - Machine Learning, Deep Learning
  - Cloud Computing
  - Data Gathering

Hands-On Example 1.2: Analyzing Data

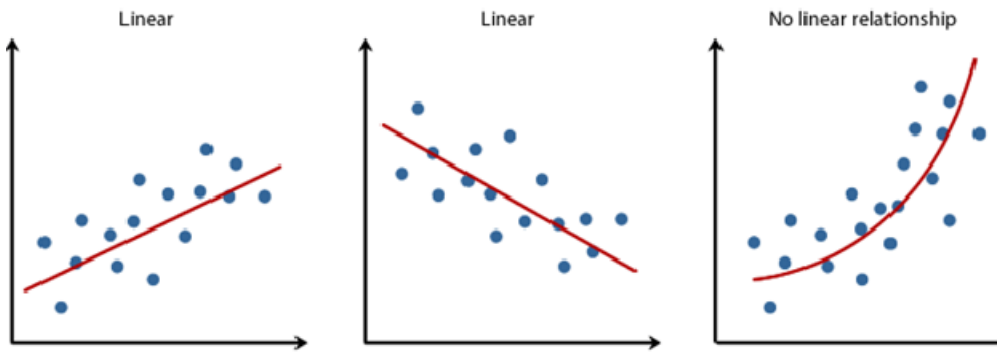
- Let's start with a data-driven problem, identify a data source, collect data, clean the data, analyze the data, and present our findings.

- We will use the dataset of average heights and weights for American women.

Observation	Height (inches)	Weight (lbs)
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164

- On average, how much increase can we expect in weight with an increase of one inch in height?

- We need a model that describes the differences in height and weight:



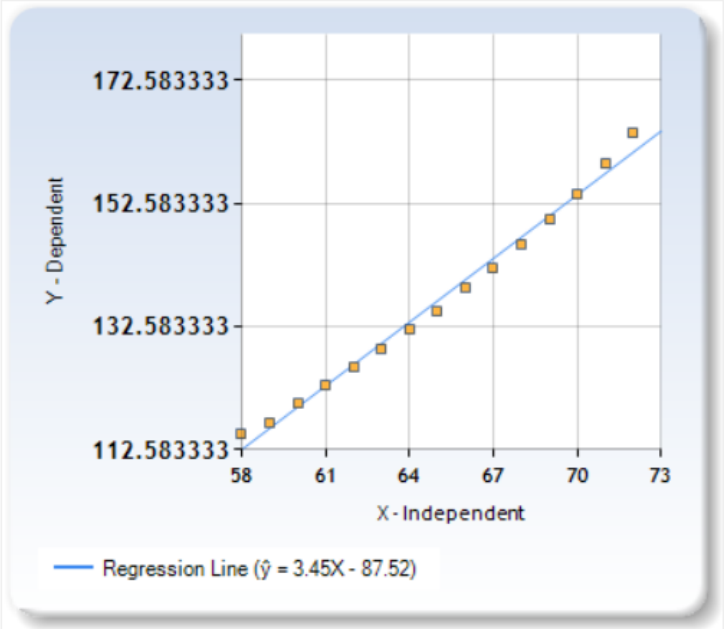
- A simple approach:
    - Compute the differences in height (72 − 58 = 14 inches) and weight (164 − 115 = 9 pounds), then divide the weight difference by the height difference, that is, 49/14 = 3.5.
    - **On average, one inch of height difference leads to a difference of 3.5 pounds in weight.**

- A problem with the above approach:
    - **The weight change with respect to the height change is not that uniform.**
    - On average, an increase of an inch in height results in an increase of less than 3 pounds weight for height between 58 and 65 inches (65 inches is the average)
    - For values of height greater than 65 inches, weight increases more rapidly (by 4 pounds mostly until 70 inches, and pounds for more than 70 inches).

- **Question**:
    - What would you expect the weight to be of an American woman who inches tall?
    - To answer this, we will have to extrapolate the data we have, i.e., build a model that predicts the relationship between **independent** and **dependent** parameters

- Let's see what **a linear regression** calculator predicts: (https://www.socscistatistics.com/tests/regression/default.aspx):

X Values	*Y* Values
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164
M: 65	M: 136.7333

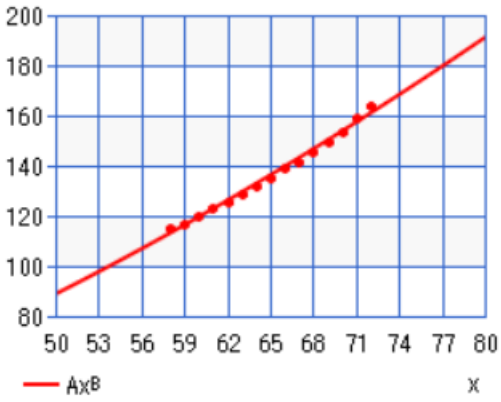


- Let's repeat the above but this time use a power regression tool to see what the **non-linear model** looks like (https://keisan.casio.com/exec/system/14059931777261):

Power regression: $y=Ax^B$

No.	x	y
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164

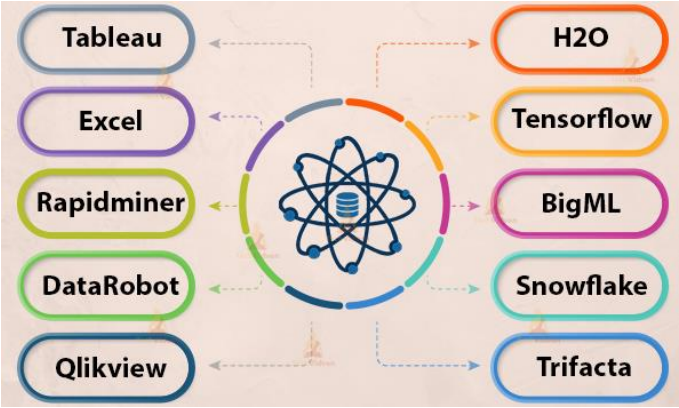
function	value
mean of x	64.85599887
mean of y	135.922013
correlation coefficient r	0.99719611
A	0.151807
B	1.629183



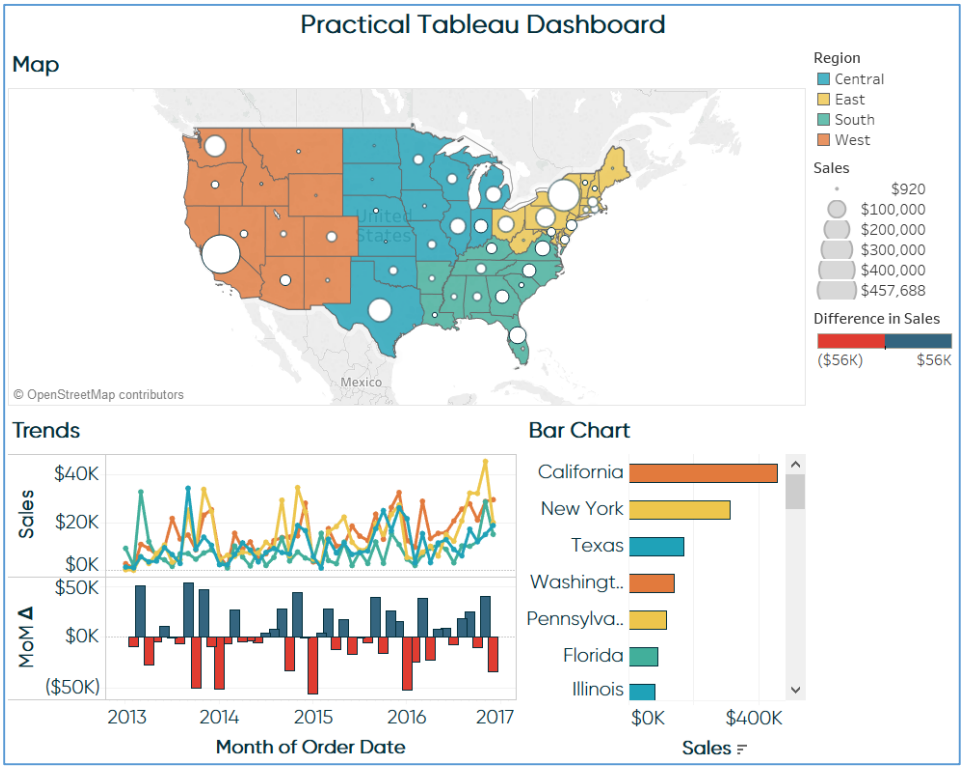
Guidelines for interpreting correlation coefficient r :

$0.7 < \|r\| \leqq 1$	strong correlation
$0.4 < \|r\| < 0.7$	moderate correlation
$0.2 < \|r\| < 0.4$	weak correlation
$0 \leqq \|r\| < 0.2$	no correlation

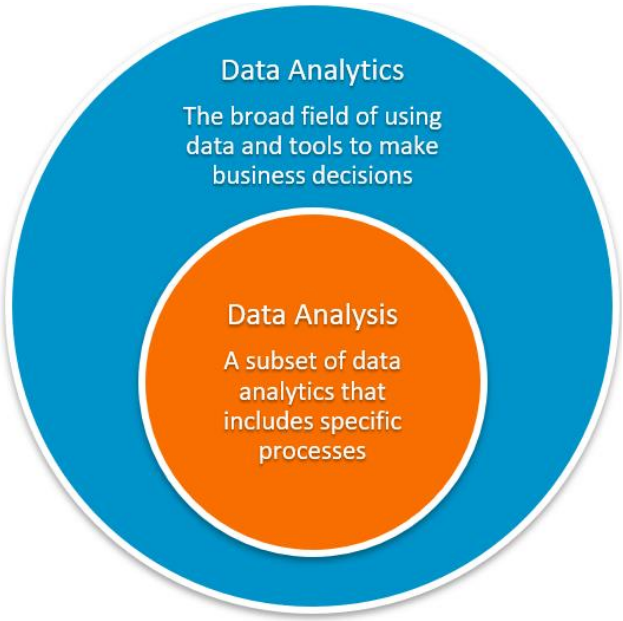
1.7 Tools for Data Science



- **Tableau**
  - Tableau is a popular tool for **data visualization** (**Qlikview** is another popular DV tool).
  - It transforms the raw data in such a format that it becomes easy to understand and use.
  - It does not require any knowledge of programming for using it so people from different backgrounds such as Business, Research, Industry, etc. are using it.



- **RapidMiner**:
    - Designed to support data mining, text mining, multimedia mining, machine learning (ML), business analytics, and predictive analysis.



- **Machine Learning** (ML) tools:
    - DataRobot, Tensorflow, H2O, BigML

- Cloud Storage:
    - SnowFlake (SQL)

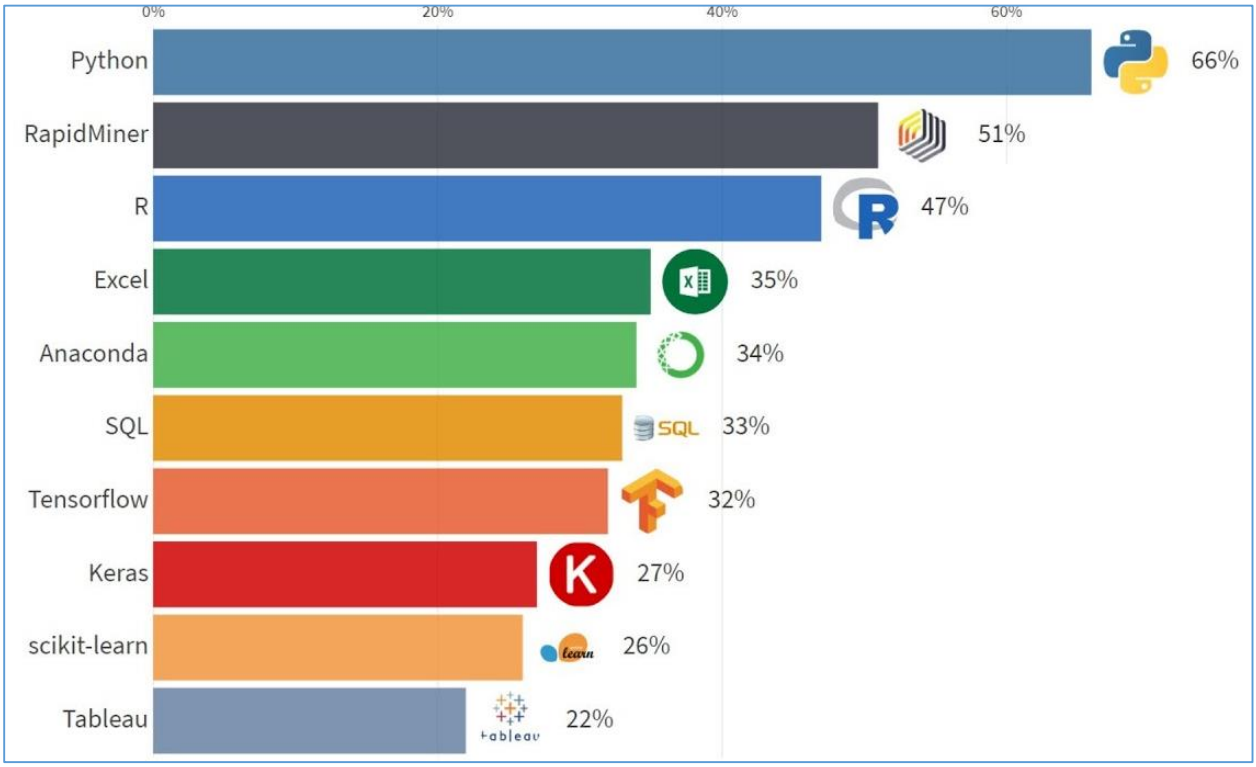
- Languages:
    - Python and R (**Anaconda**), Scala, Java, SQL, NoSQL

- **Data Wangling**:
    - The process of transforming raw data into a more usable form by applying various methods.

    - For better results, it is very important to convert this data in such a form that it becomes easier to manage, analyze, and draw useful conclusions out of it.

    - **Trifacta** is an open-source tool used for Data Wrangling. It enables the Data Scientists to completely focus on the actual analysis of the data by giving them already prepared data.

- Top Data Science Tools (as of 2019)



1.8 Issues of Ethics, Bias, and Privacy in Data Science

- Many of the issues related to **privacy**, **bias**, and **ethics** can be traced back to the origin of the data.

- How, where, and why was the data collected? Who collected it? What did they intend to use it for?

- More important, if the data was collected from people, did these people know that:
    - Such data was being collected about them
    - How the data would be used?

- Often those collecting data mistake availability of data as the right to use that data.

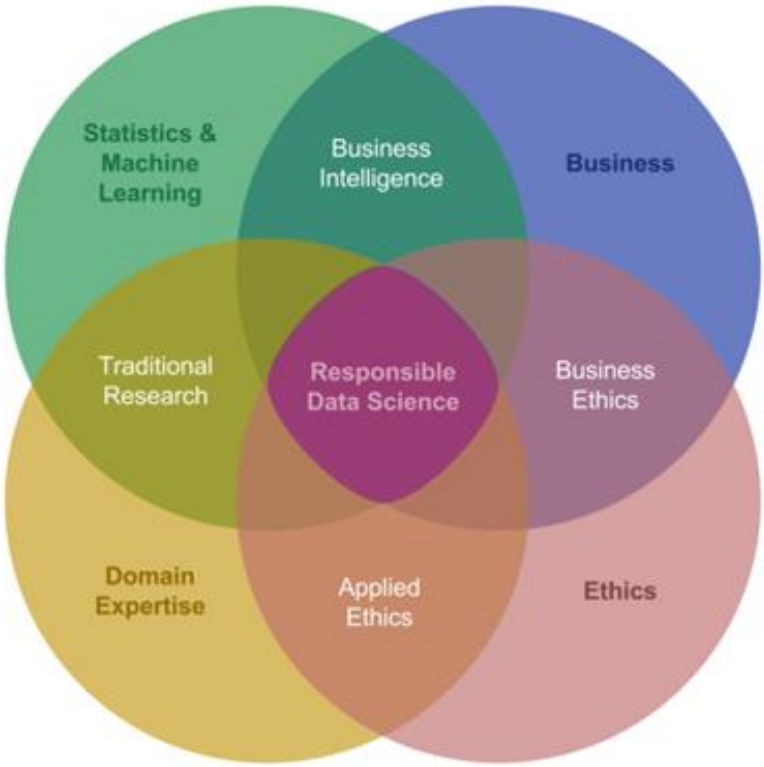
- For instance, just because data on a social media service such as Twitter is available on the Web, it does not mean that one could collect and sell it for material gain without the consent of the users of that service.

- **Cambridge Analytica** (2018)
    - Obtained data about a large number of Facebook users to use for political campaigning.
    - Those Facebook users did not even know that:
        - Such data about them was collected and shared by Facebook to third parties
        - The data was used to target political ads to them.

- For many years, various companies such as **Facebook** and **Google** have collected enormous amounts of data about and from their users in order not only to improve and market their products, but also to share and/or sell it to other entities for profit.

- As the old saying goes, “there is no free lunch.” So, when you are getting an email service or a social media account for “free,” ask why? As it is often understood, “if you are not paying for it, you are the product.”
    - For **Facebook**, each user is worth $158.
    - For **Google**, each user is worth $182.
    - For **Amazon**, each user is worth $733.

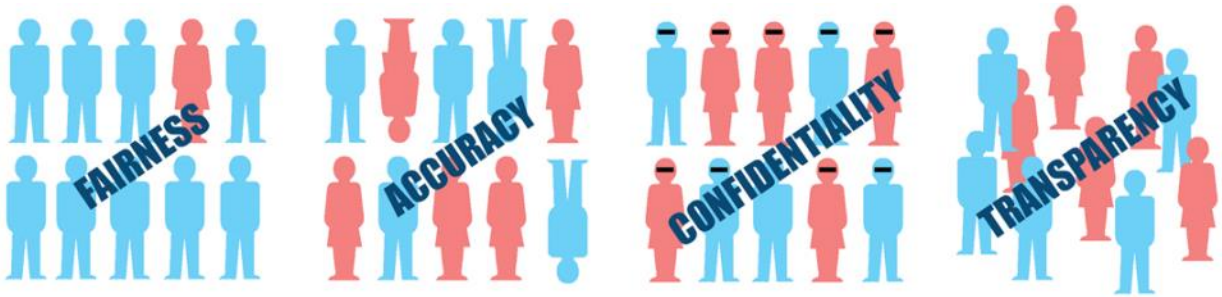
- **Responsible Data Science (RDS)**



- Observations:
  - Our mission is to transform the workforce so that people and organizations thrive, and one of our core theses is that we can do this by removing bias from the hiring process through machine learning and predictive analytics.

  - Responsibly deploy intelligent systems to help us make socially sensitive decisions like medical diagnosis, parole-granting, and hiring, we should consider **ethics** to be a critical competency of the contemporary data scientist.

- There's a growing body of work relating to **fairness, accountability, and transparency** in ML. At conferences like FATML, ML practitioners, researchers, and policymakers are sharing knowledge about how to design fair and interpretable machine learning algorithms when applied to socially sensitive decisions.



- **RDS** evolves around four main challenges: