Computing for Data Analytics
CPSC 4800

Introduction

CPSC 4800, Nasim Taba

Langara.
THE COLLEGE OF HIGHER LEARNING

Introduce Yourself!

Ice Breaker

- What is your name?
- What is your background?
- Are you currently in Vancouver? If not, where are you joining from?
- Why did you choose this program?
- Are you currently working?

What will you learn in this lecture?

- What is data science?
- Who is a data analyst?
- Data science salary
- Data science/analyst roles and responsibilities
- Definition of data
- Different types of Data
- The Data Science Pipeline
- The list of applications to install for this course

What is data science?

Data science

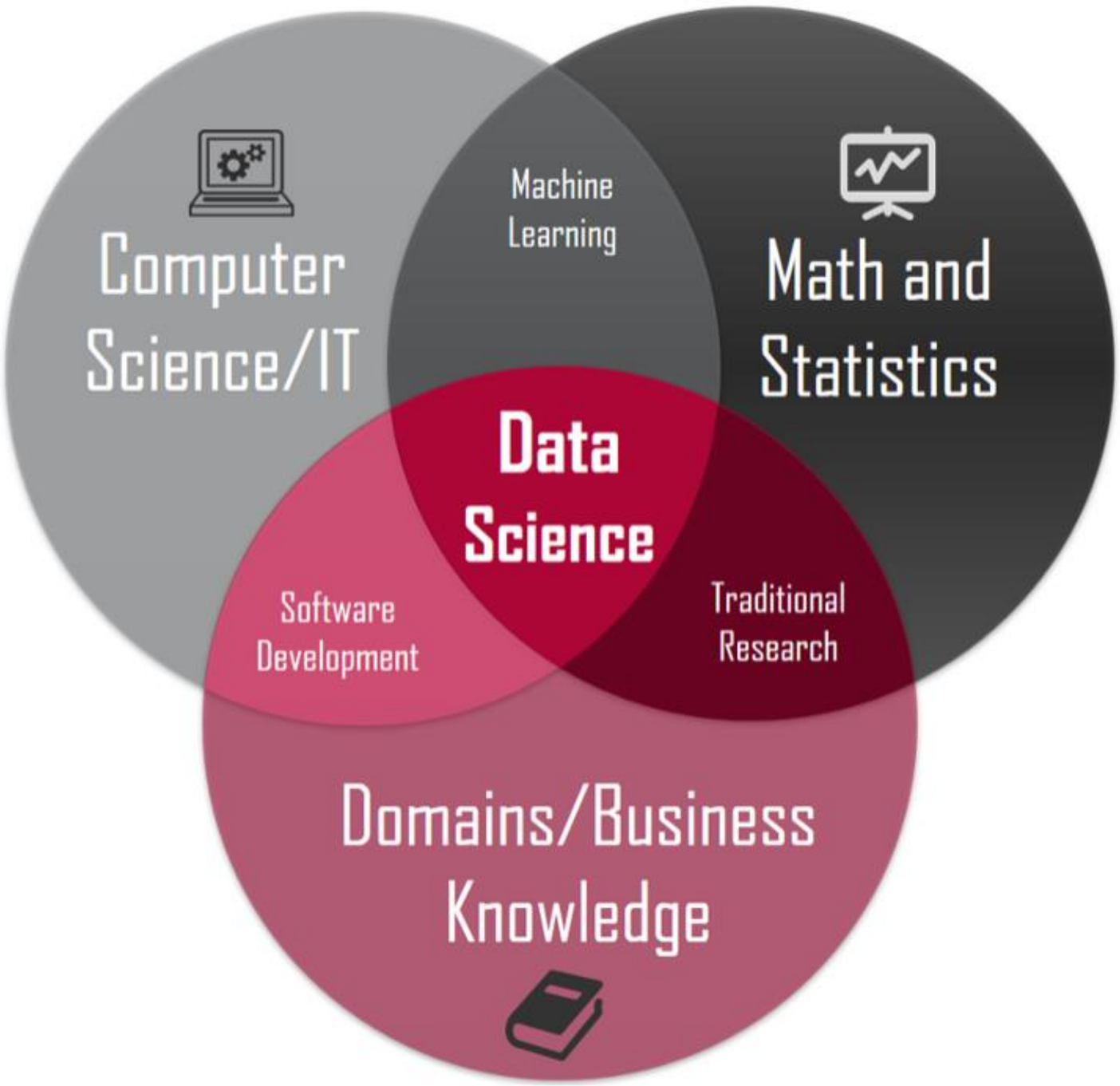
From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.[1][2] Data science is related to data mining and big data.

Josh Wills – 2012:

“Data Scientist: Person who is better at statistics than any software engineer and better at software engineering than any statistician.”


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

Source

Who is a data analyst?

A data analyst is an individual who is responsible to gather, investigate and represent data and filter out useful information from it.

Duties of a Data Analyst

- Conduct preliminary data analysis to assess the nature of data
- Conduct further analysis to extract meaningful knowledge
- Perform data mining and use querying languages (for e.g. SQL)
- Determine data configurations and patterns
- Represent data through graphs, charts and other representational techniques

Duties of a Data Scientist

- Model data through different data modelling techniques
- Make data projections and advice the relevant stakeholders
- Present findings through meetings, presentations, workshops and seminars
- Prepare the final reports on the basis of the analysis

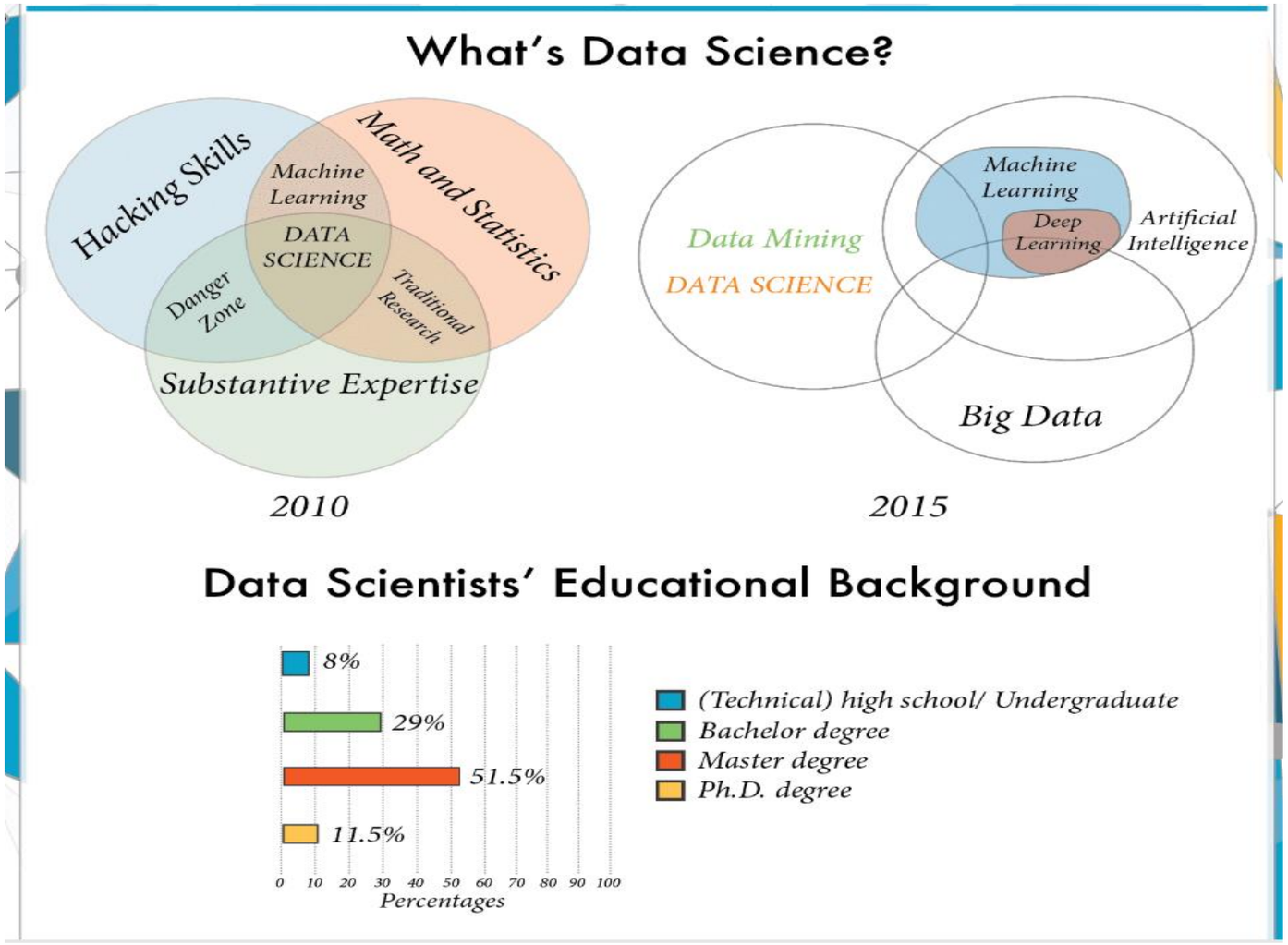
What are the required skills to become a data scientist?

- Technical Skills
  - Python, R programming
  - SQL programming
  - Machine Learning/AI
  - Data Visualization
- Non-Technical Skills
  - Intellectual curiosity
  - Business acumen
  - Communication skills

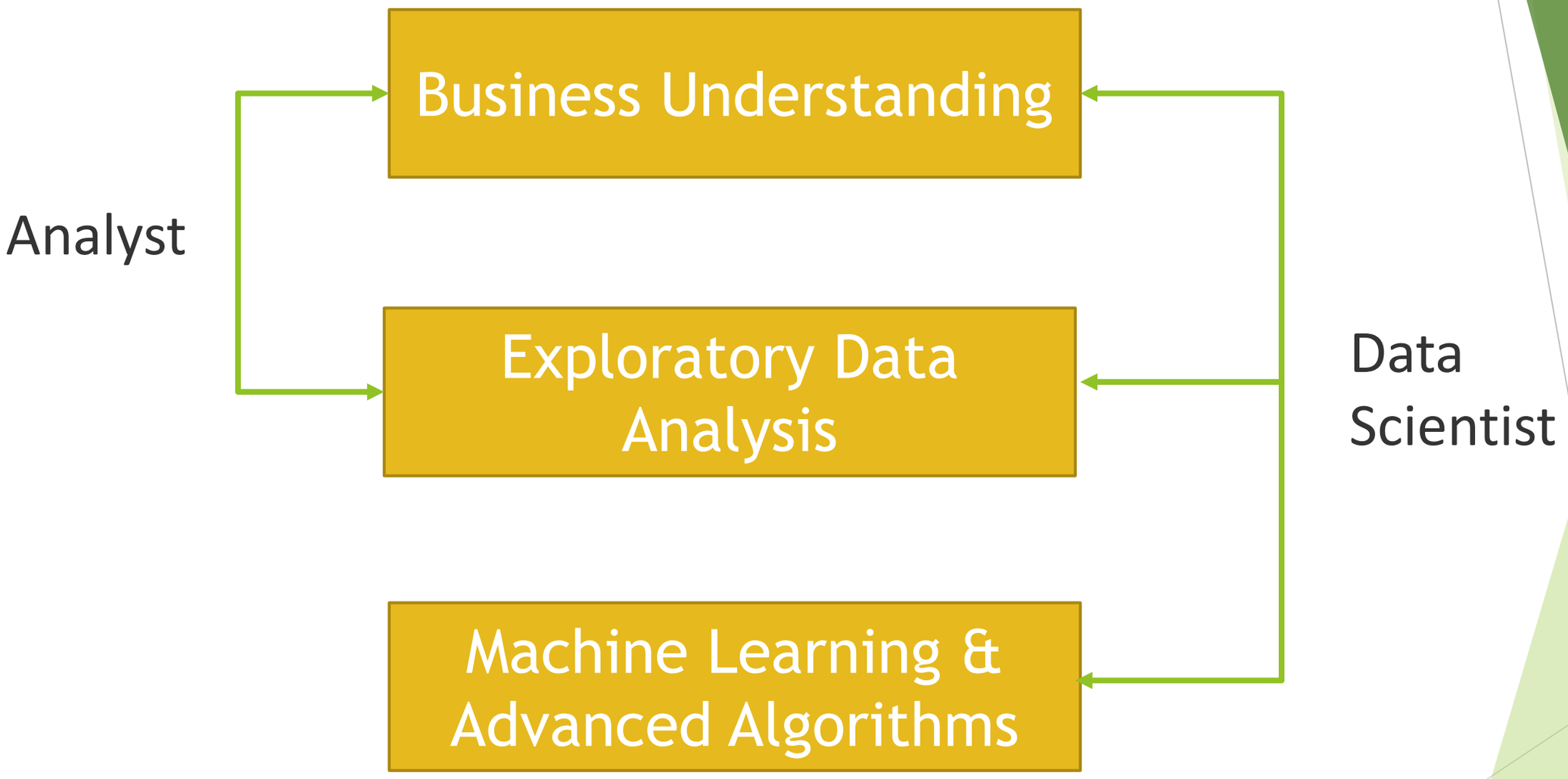
Source

A Visual Guide to Become a Data Scientist

- A visual guide to Becoming a Data Scientist in 8 Steps by DataCamp



Data Science Vs Data Analyst



Data Science vs Machine learning

Skills Needed for Data Scientists	Skills Needed for Machine Learning Engineers
Statistics	Computer science fundamentals
Data mining and cleaning	Statistical modeling
Data visualization	Data evaluation and modeling
Unstructured data management techniques	Understanding and application of algorithms
Programming languages such as R and Python	Natural language processing
Understand SQL databases	Text representation techniques

Source

Why do we need data scientist or data analyst?

How much can happen in a minute?

- $ 400 M sales on Alibaba
- 439, 000 page views on Wikipedia
- 194, 000 apps downloaded
- 31, 700 hours of music played on Pandora
- 38, 000 photographs uploaded to Instagram
- 4.1 Million searches on Google
- 139, 000 hours of video watched on Youtube
- 10 million ads displayed
- 3.3 million shares on Facebook

1 internet minute

- $ 400 M sales on Alibaba
- 439, 000 page views on Wikipedia
- 194, 000 apps downloaded
- 31, 700 hours of music played on Pandora
- 38, 000 photographs uploaded to Instagram
- 4.1 Million searches on Google
- 139, 000 hours of video watched on Youtube
- 10 million ads displayed
- 3.3 million shares on Facebook

Each of these activities generates

DATA

DATA

1 internet minute

- $ 400 M sales on Alibaba
- 439, 000 page views on Wikipedia
- 194, 000 apps downloaded
- 31, 700 hours of music played on Pandora
- 38, 000 photographs uploaded to Instagram
- 4.1 Million searches on Google
- 139, 000 hours of video watched on Youtube
- 10 million ads displayed
- 3.3 million shares on Facebook

1 internet minute

$ 400 M sales on Alibaba
439, 000 page views on Wikipedia
194, 000 apps downloaded
31, 700 hours of music played on Pandora
38, 000 photographs uploaded to Instagram
4.1 Million searches on Google
139, 000 hours of video watched on Youtube
10 million ads displayed
3.3 million shares on Facebook

DATA

4.1 Million searches on Google

Results returned
Results viewed
Results clicked

Alibaba
Wikipedia
Pandora
Instagram
Google
Youtube
Facebook

These companies and others are collecting

PetaBytes of data every minute

PetaBytes of data every minute

What does this mean?

1 PetaByte ~ 1000 TeraBytes

1 PetaByte ~ 1000 TeraBytes

This is a 1 TB hard disk drive

1 PetaByte ~ 1000 TeraBytes

1000s of such 1 TB drives are filled up every minute by data collected on the web!!

Understand Data Measurements

Value	Symbol	Name
1024	KB	Kilobyte
1024^2	MB	Megabyte
1024^3	GB	Gigabyte
1024^4	TB	Terabyte
1024^5	PB	Petabyte
1024^6	EB	Exabyte
1024^7	ZB	Zettabyte
1024^8	YB	Yottabyte

Source of Data Generation

Social Media

Sensors

Cell Phones

GPS

Purchase

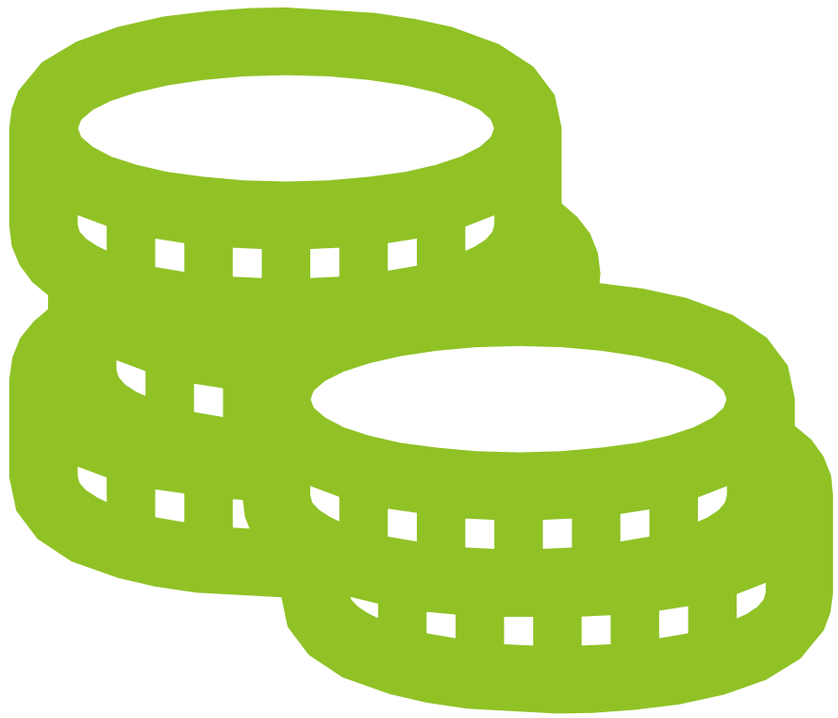
WWW

E-mails

Media streaming

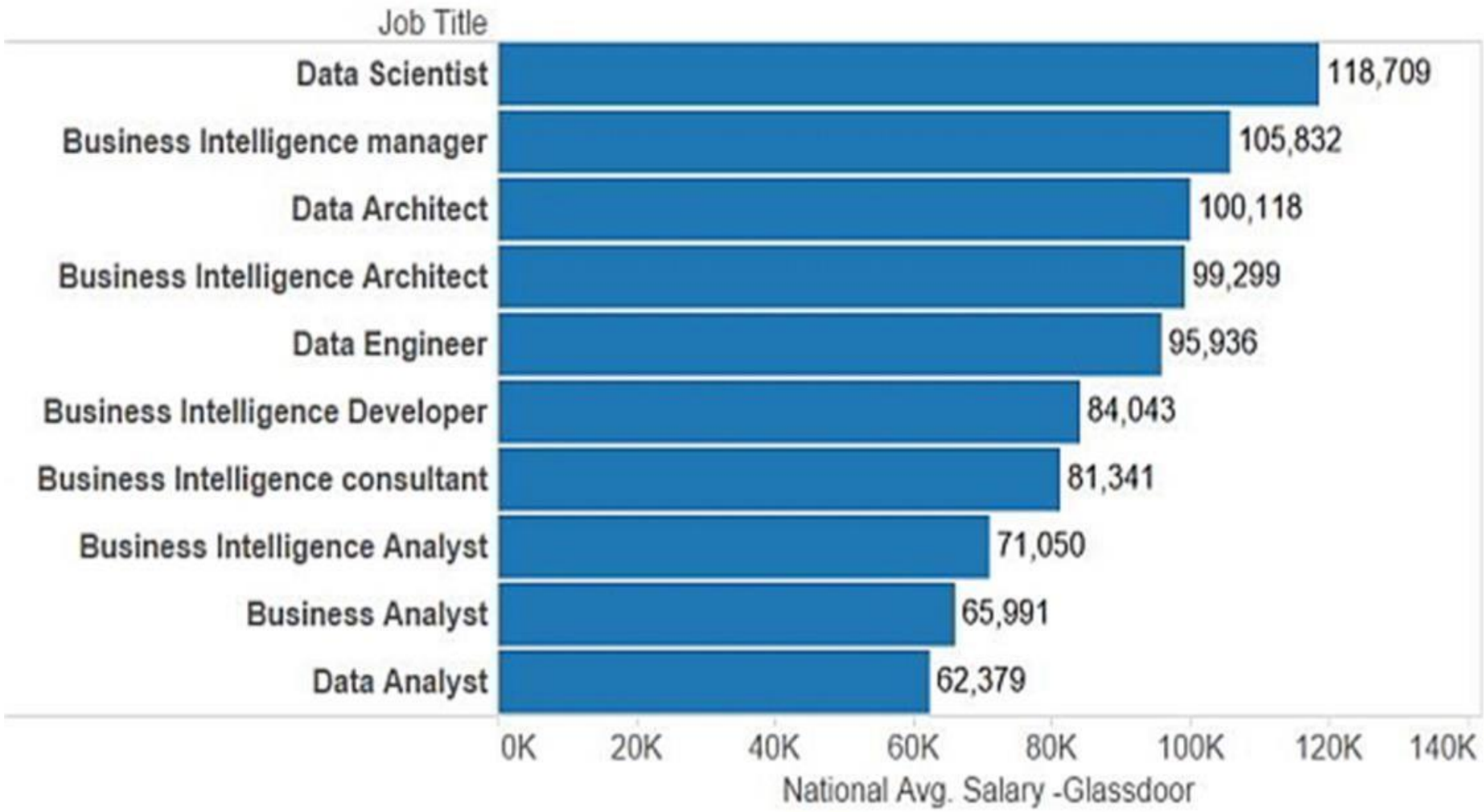
Healthcare

IOT



Data Science Salary

Popular Job Titles in data science & business intelligence by National Avg. Salary (in dollars)



Source

Are these the world's best jobs?

Ranking determined by work-life balance rating

WORLD ECONOMIC FORUM

Rank	Job	Salary
1	Data Scientist	$114,808
2	SEO Manager	$45,720
3	Talent Acquisition Specialist	$63,504
4	Social Media Manager	$40,000
5	Substitute Teacher	$24,380
6	Recruiting Coordinator	$44,700
7	UX Designer	$91,440
8	Digital Marketing Manager	$70,052
9	Marketing Assistant	$32,512
10	Web Developer	$66,040
11	RIsk Analyst	$69,088
12	Civil Engineer	$65,532
13	Client Manager	$71,120
14	Instructional Designer	$66,040
15	Marketing Analyst	$60,000
16	Software QA Engineer	$91,440
17	Web Designer	$53,848
18	Research Technician	$36,525
19	Program Analyst	$71,120
20	Data Analyst	$58,928
21	Content Manager	$60,960
22	Solutions Engineer	$92,456
23	Lab Assistant	$27,550
24	Software Developer	$80,000
25	Front End Developer	$75,000

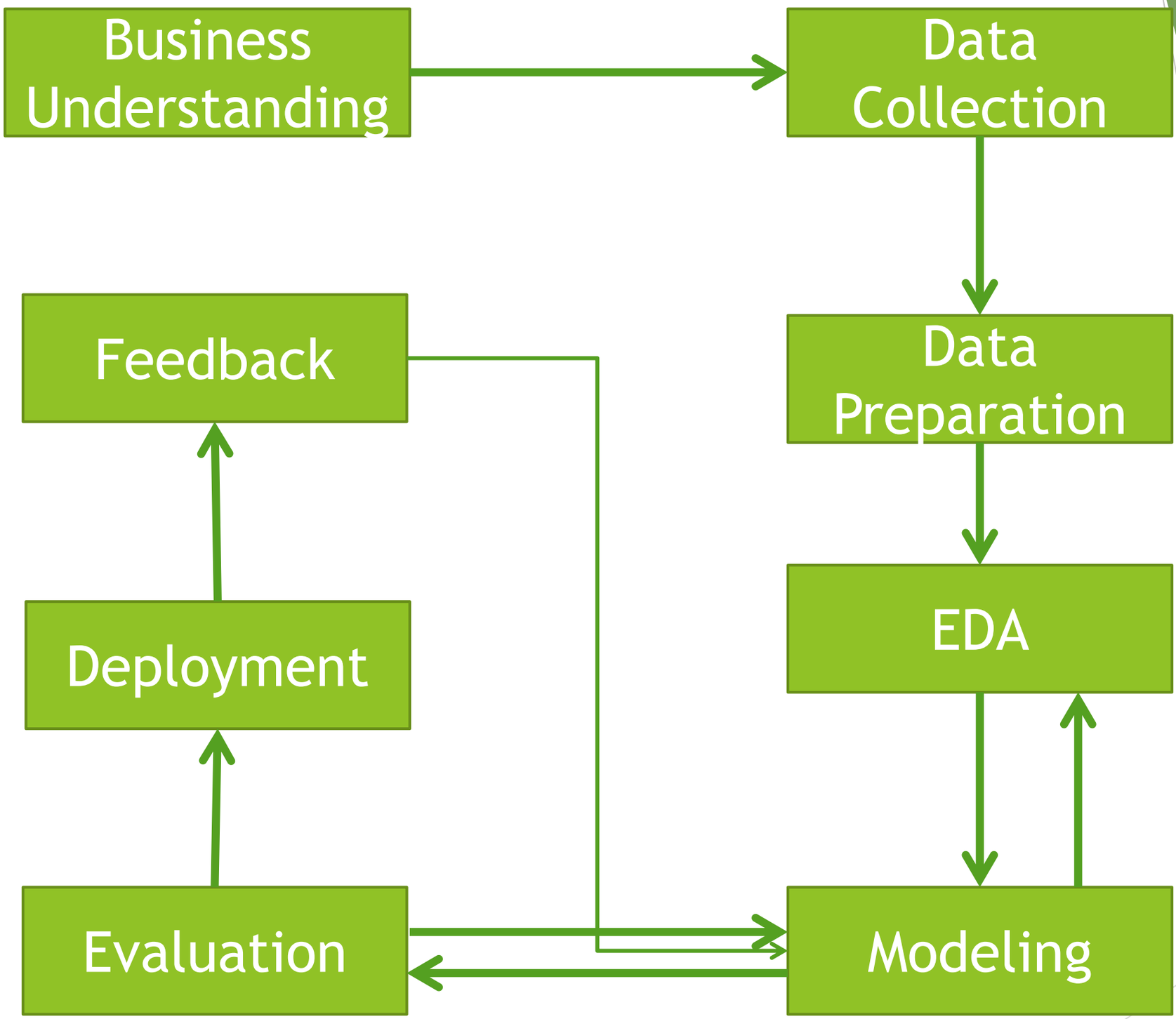
Source: Glassdoor.com

Data science Applications

Travel	Dynamic Pricing Predicting Flight Delay
Marketing	Customer Churn Predicting lifetime value of a customer Cross Selling
Healthcare	Disease prediction
Social Media	Digital Marketing Sentiment Analysis
Sales	Demand Forecasting Discount Offering
Automation	Self Driving Cars Pilotless Drones, aircraft
Credit & Insurance	Claim prediction Fraud and Risk detection

Data Analytics/Machine Learning Steps

- Step 1: Define your problem
- Step 2: Prepare your data
- Step 3: EDA
- Step 4: Feature Engineering
- Step 5: Model Building
- Step 6: Model Evaluation
- Step 7: Model Deployment
- Step 8: Present your results


Business Understanding
Data Collection
Data Preparation
EDA
Modeling
Evaluation
Deployment
Feedback

Source

What is EDA?

EDA: an approach to data analysis that aims at observing and summarizing the main characteristics of a dataset

EDA is more about observing data

EDA often makes use of visual methods (plots)

EDA may require iterations with data cleaning steps

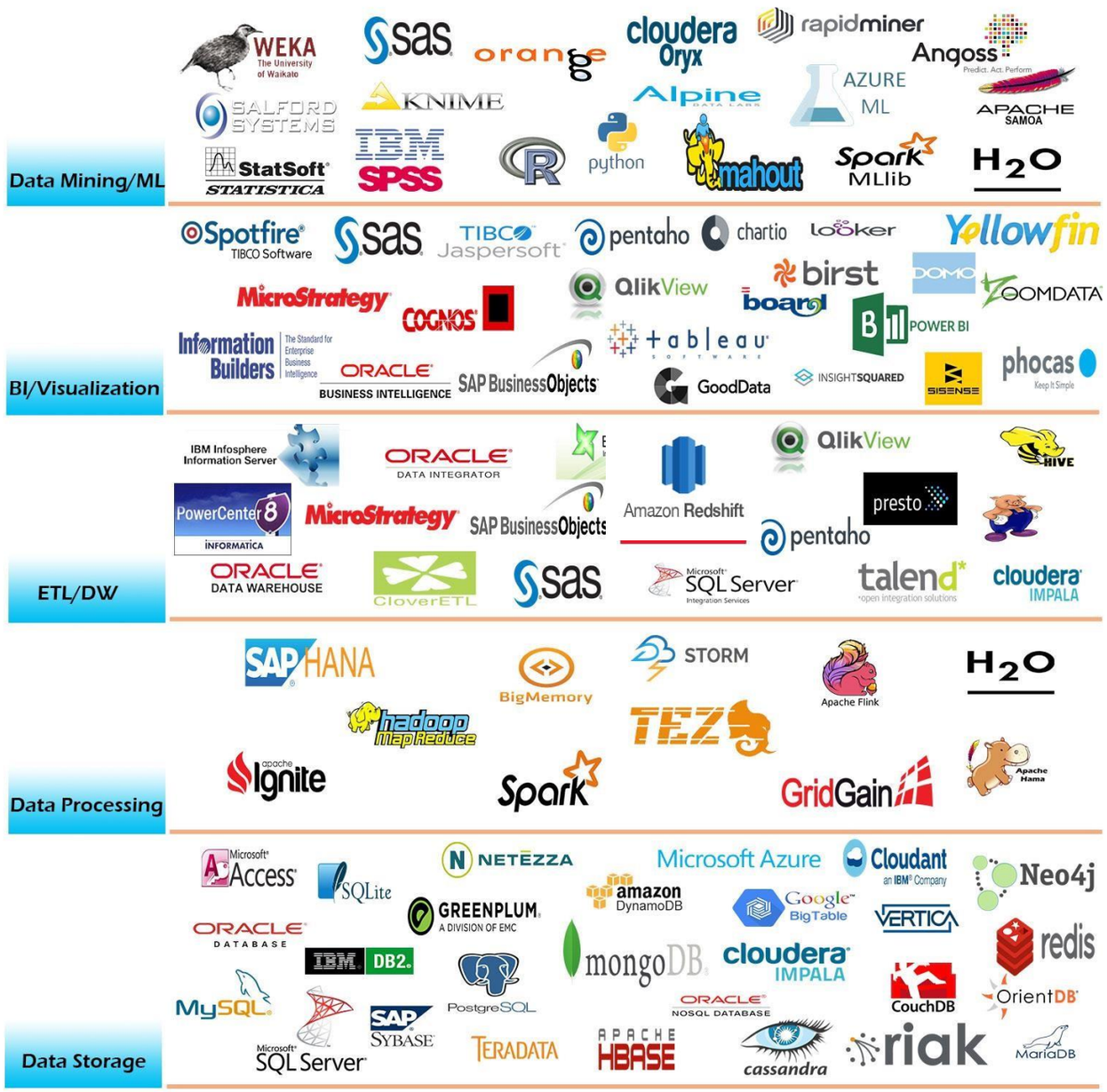
Data Definition Framework

Data Definition Framework

Data Format

Data Source	Structured	Unstructured
Internal	**Human-Generated** • Survey ratings • Aptitude testing **Machine-Generated** • Web metrics from Web logs • Product purchase from sales Records • Process control measures	**Human-Generated** • Emails, letters, text messages • Audio transcripts • Customer comments • Voicemails • Corporate video/communications • Pictures, illustrations • Employee reviews
External	**Human-Generated** • Number of Retweets, Facebook likes, Google Plus +1s • Ratings on Yelp • Patient ratings ratings **Machine-Generated** • GPS for tweets • Time of tweet/updates/postings	**Human-Generated** • Content of social media updates • Comments in online forums • Comments on Yelp • Video reviews • Pinterest images • Surveillance video

Tools related to Data



Case Study

Have you seen the Titanic Movie?

In 1912, the ship RMS Titanic struck an iceberg on its maiden voyage and sank, resulting in the deaths of most of its passengers and crew.

Titanic Dataset

Contains the details of a sample of passengers on board

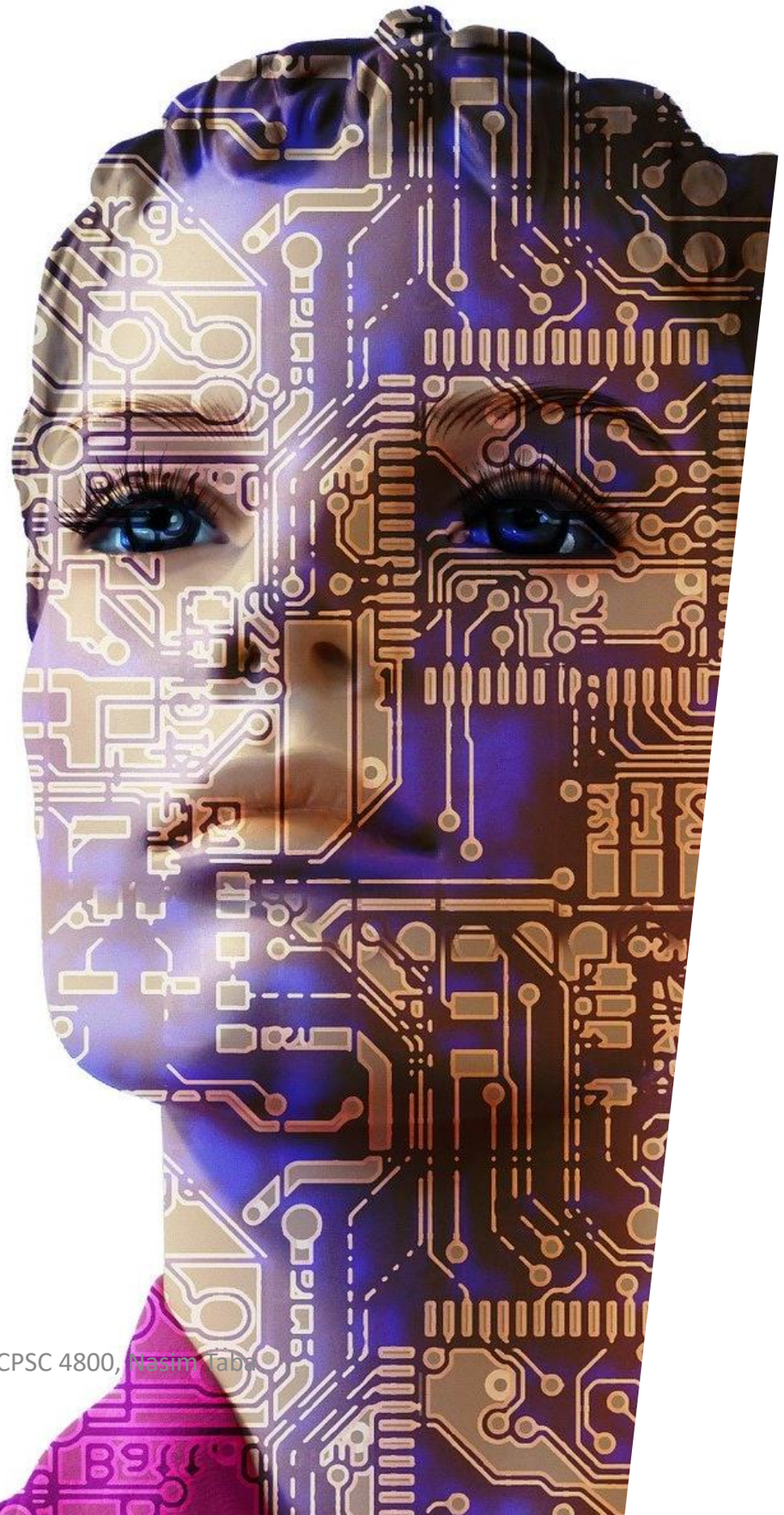
Reveal whether they survived or not

Titanic is a machine learning competition on Kaggle

What are the features?

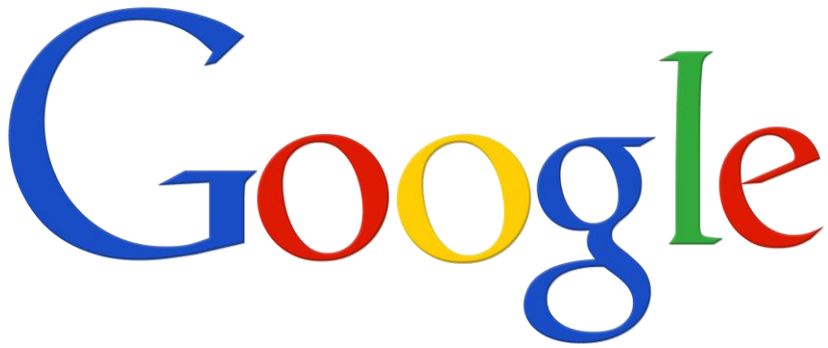
- Gender of the passenger
- Age of the passenger
- Socio-economic class (1 = Upper class; 2 = Middle class; 3 = Lower class)
- Number of siblings and spouses of the passenger aboard

What is Kaggle?



Kaggle is the home for data scientists

Kaggle Owned by Google



- A platform to compete with other data scientists
- Allows users to explore various datasets, and build models

Hypotheses

First

- Determine if the survival rate is associated to the class of passenger

Second

- Determine if the survival rate is associated to the gender

Third

- Determine if the survival rate is associated to the age

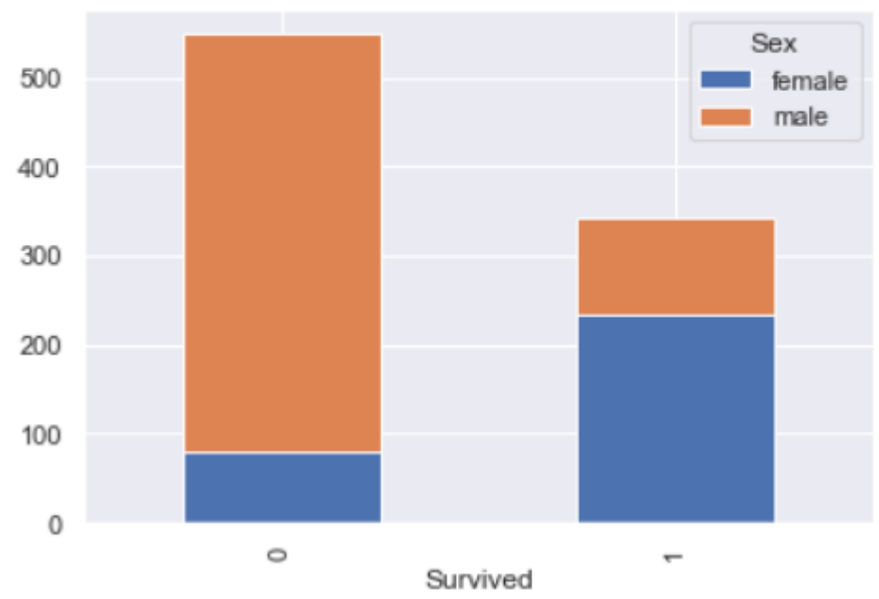
Let’s begin Exploring!

Gender - Survival

Gender	Survived
Female	74.2%
Male	18.89%

```
gender_table.plot(kind="bar", stacked=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21892a60c88>
```

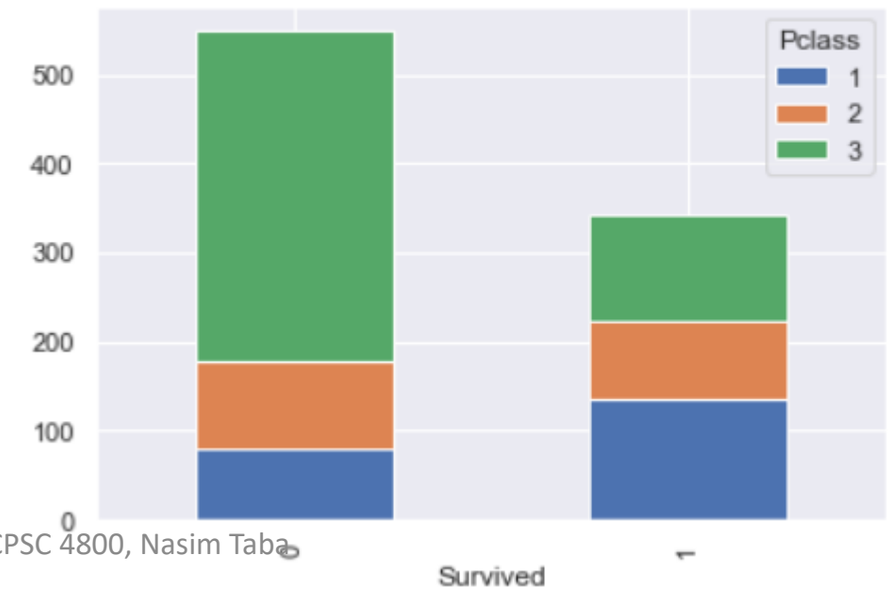


Pclass - Survival

Class	Survived
1	62.96%
2	47.28%
3	24.24%

```
table.plot(kind="bar", stacked = True)
```

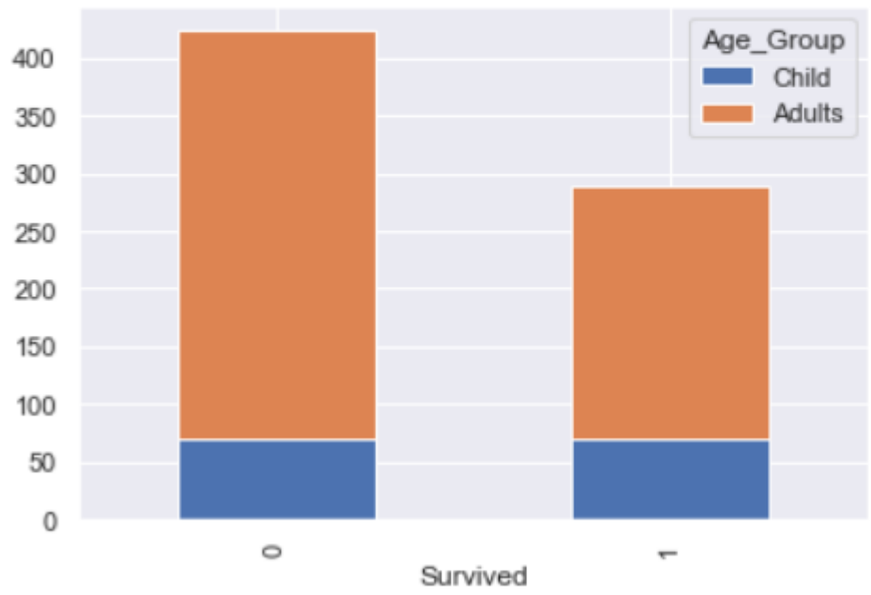
```
<matplotlib.axes._subplots.AxesSubplot at 0x218937ae438>
```



Age Group - Survival

Age Group	Survived
Child	59.03%
Adult	38.19%

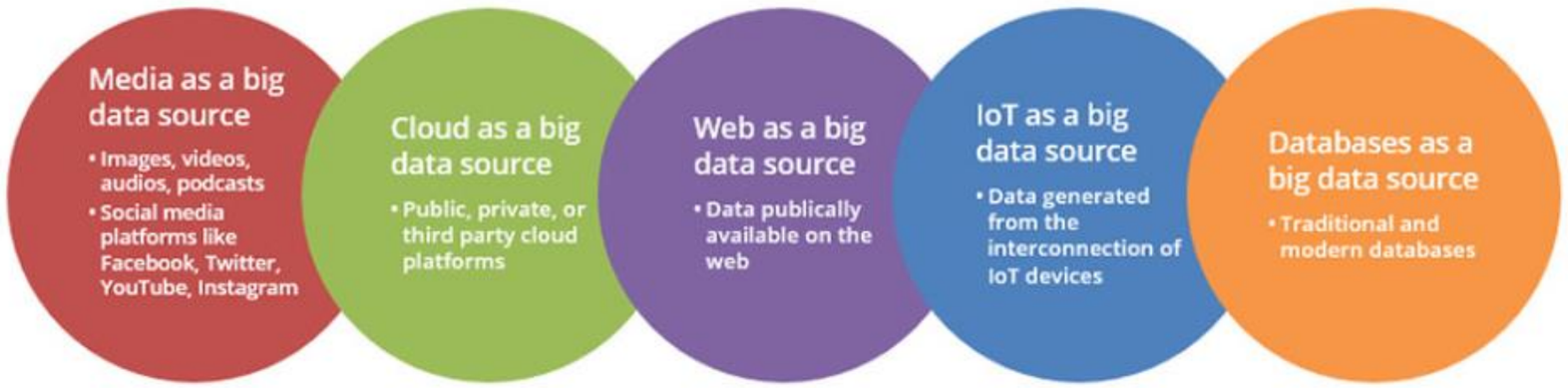
```
: table.plot(kind="bar", stacked = True)

: <matplotlib.axes._subplots.AxesSubplot at 0x21893350748>
```



Conclusion

Item #	Conclusion
1	There is a statistically significant relationship between survival rate and Cabin Class.
2	There is a statistically significant relationship between survival rate and age group.
3	There is a statistically significant relationship between survival rate and gender.

Sources of Data



MEDIA AS A DATA SOURCE

The most popular source of data

The fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making.

Includes social media as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction.

CLOUD AS A DATA SOURCE

Companies have moved ahead of traditional data sources by shifting their data on the cloud.

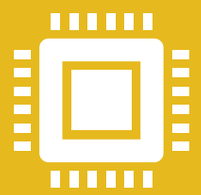
Cloud storage accommodates structured and unstructured data

Provides business with real-time information and on-demand insights.

The main attribute of cloud computing is its flexibility and scalability.

THE WEB AS A DATA SOURCE

The public web constitutes big data that is widespread and easily accessible.

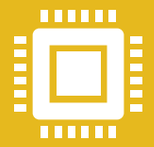


Data on the Web or ‘Internet’ is commonly available to individuals and companies alike.

Moreover, web services such as Wikipedia provide free and quick informational insights to everyone

THE IOT (INTERNET-OF-THINGS) AS A DATA SOURCE

Machine-generated content or data created from IoT constitute a valuable source of data.



This data is usually generated from the sensors that are connected to electronic devices.

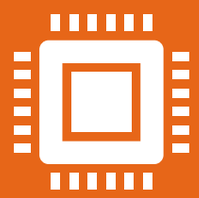
The sourcing capacity depends on the ability of the sensors to provide real-time accurate information.

With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, household appliances, and the like.

DATABASES AS A DATA SOURCE

Databases are deployed for several business intelligence purposes as well.

These databases can then provide for the extraction of insights that are used to drive business profits.



Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, PostgreSQL, and Amazon S3, among others.

Data Information Variables

- Quantitative variables
  - Numerical variables: counts, percent or numbers
- Categorial variables
  - Descriptions of groups or things, like “breeds of dog” or “voting preference”

Examples of Quantiative variables

- High school Grade Point Average (e.g. 4.0, 3.2, 2.1).
- Number of pets owned (e.g. 1, 2, 4).
- Bank account balance (e.g. $100, $987, $-42.
- Number of stars in a galaxy (e.g. 100, 2301, 1 trillion) .
- Average number of lottery tickets sold (e.g. 25, 2,789, 2 million).
- How many cousins you have (e.g. 0, 12, 22).
- The amount in your paycheck (e.g. $200, $1,457, $2,222).

Examples of Categorical Variables

- Class in college (e.g. freshman, sophomore, junior, senior).
- Party affiliation (e.g. Republican, Democrat, Independent).
- Type of pet owned (e.g. dog, cat, rodent, fish).
- Favorite author (e.g. Stephen King, James Patterson, Charles Dickens).
- Preferred airline (e.g. Southwest, Virgin, Quantas)
- Hair color (e.g. blond, brunette, black).
- Your race (e.g. Asian, Latino, black).

Questions

- What are the responsibilities of a data analyst?
- What are the skills required to be a data scientist/analyst?
- What are the examples of categorial variables in the titanic dataset?
- What are the examples of quantitative variables in the titanic dataset?
- What are the sources of data?
- What is the difference between structured and unstructured data?

References

Frey, B., Savage, D., & Torgler, B. (2011). Behavior under Extreme Conditions: The "Titanic" Disaster. The Journal of Economic Perspectives, 209-221.

Takis, S. (1999). Titanic: A Statistical Exploration. The Research Institute of Industrial Economics, 660-664.

Wikipedia - https://en.wikipedia.org/wiki/Data_science

https://www.kdnuggets.com/2018/05/simplilearn-9-must-have-skills-data-scientist.html

Datacamp: https://s3.amazonaws.com/assets.datacamp.com/blog_assets/DataScienceEightSteps_Full.png

https://www.mastersindatascience.org/careers/data-science-vs-machine-learning/

https://www.kdnuggets.com/2015/09/salaries-roles-data-science-business-intelligence.html

https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science