COMP9321:
Data services engineering

Week 7: Introduction to Data Analytics

Term 1, 2023

By Mortada Al-Banna, CSE UNSW

Data Driven Organizations

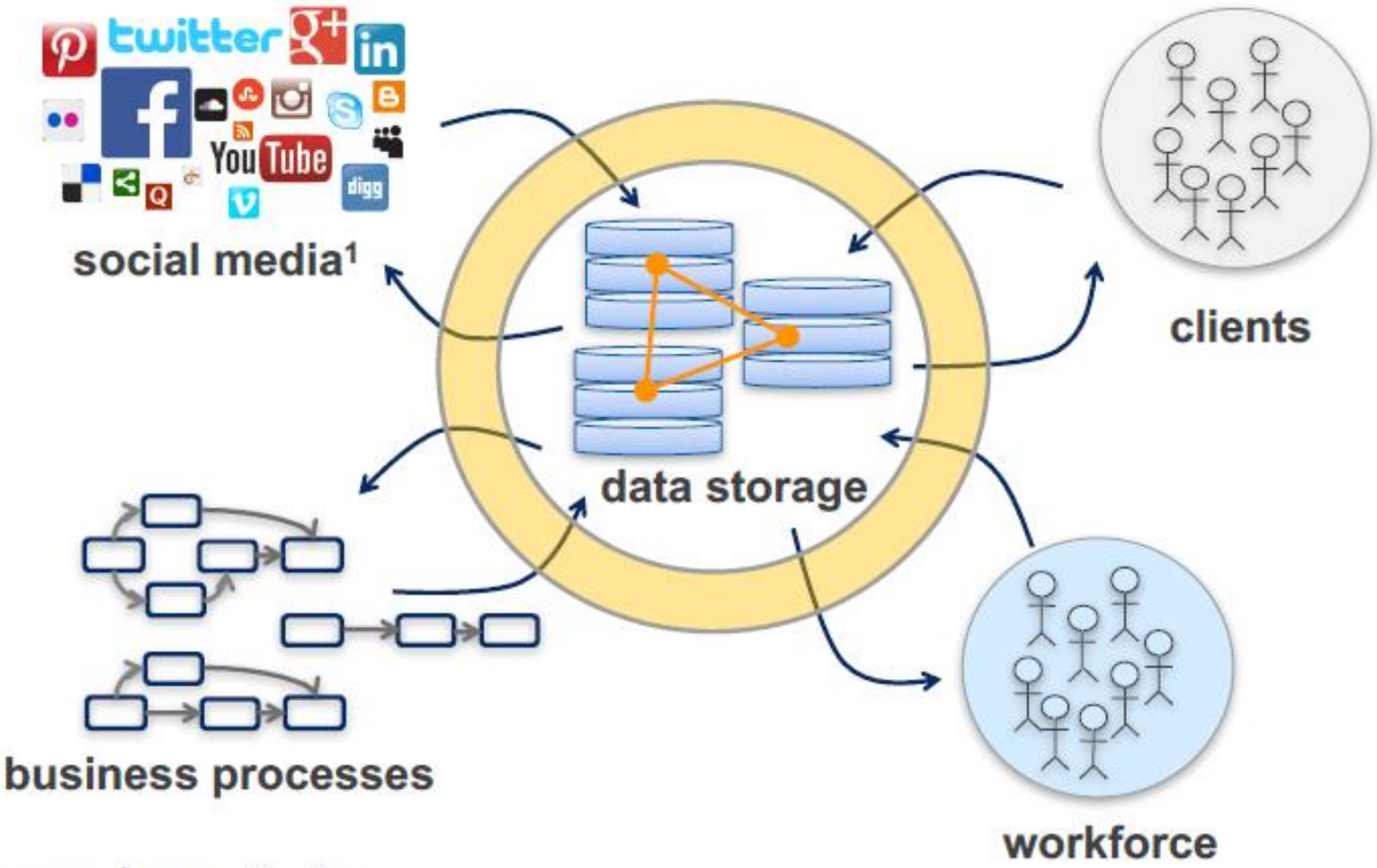


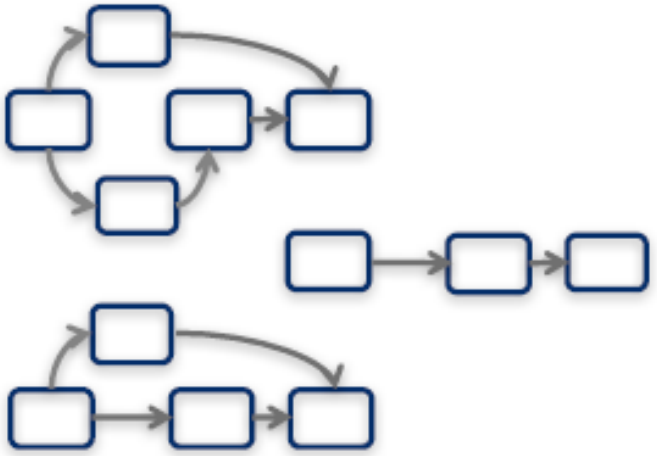
Image source: [1]commons.wikimedia.org

Data Driven Organizations and Data Analytics

- Product and service recommendation
- Customer support
- Dashboard and reporting services
- Customer engagement
- Promotions and deals
- Product and service customization
- Communication

Data Driven Organizations and Data Analytics

- Key process performance indicators
- Process execution predictions
- Decision making support services
- Process mining
- Dynamic process adaptation
- People to task assignment
- Compliance verification



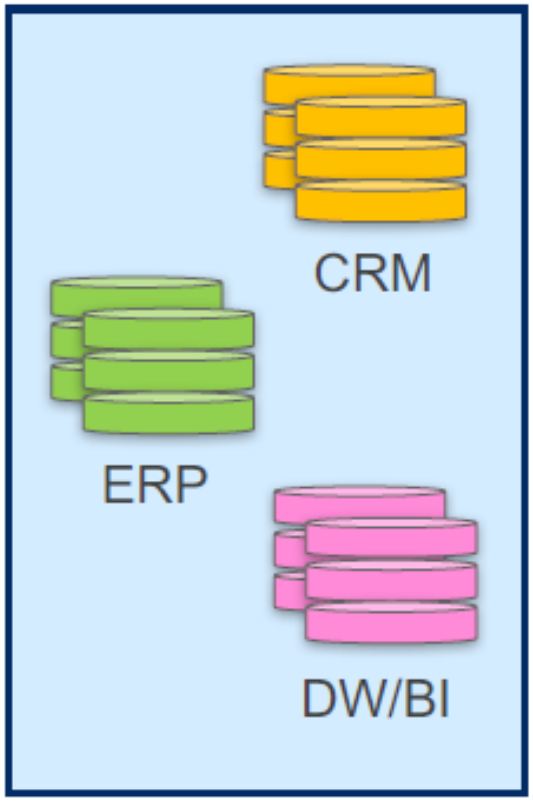
business processes

Data Driven Organizations and Data Analytics

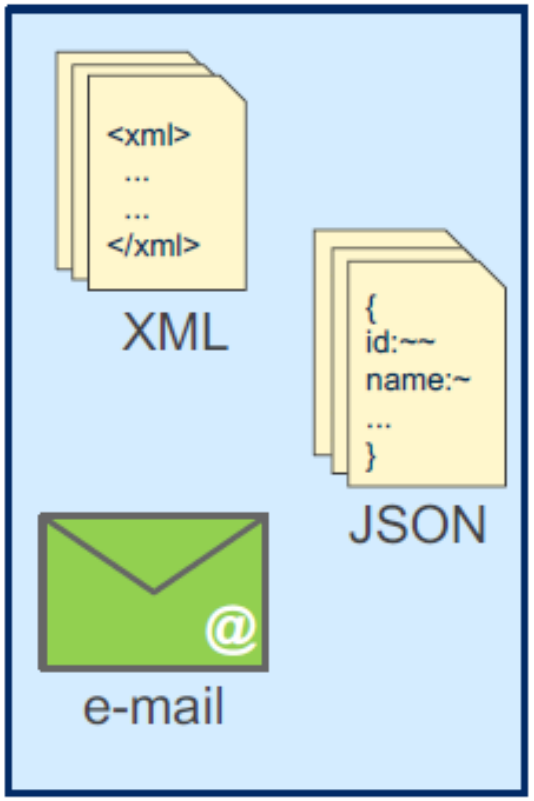
- Product and service advertisement
- Sentiment analysis
- Demographics analysis
- Virality
- Social network insights

social media[1]

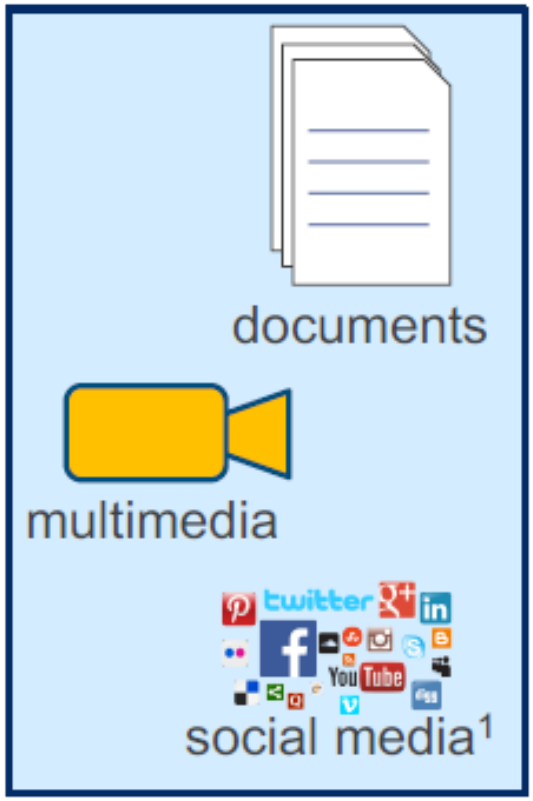
Data Used for Analytics



structured data



semi-structured data



unstructured data

Image source: [1]commons.wikimedia.org

Data Used for Analytics

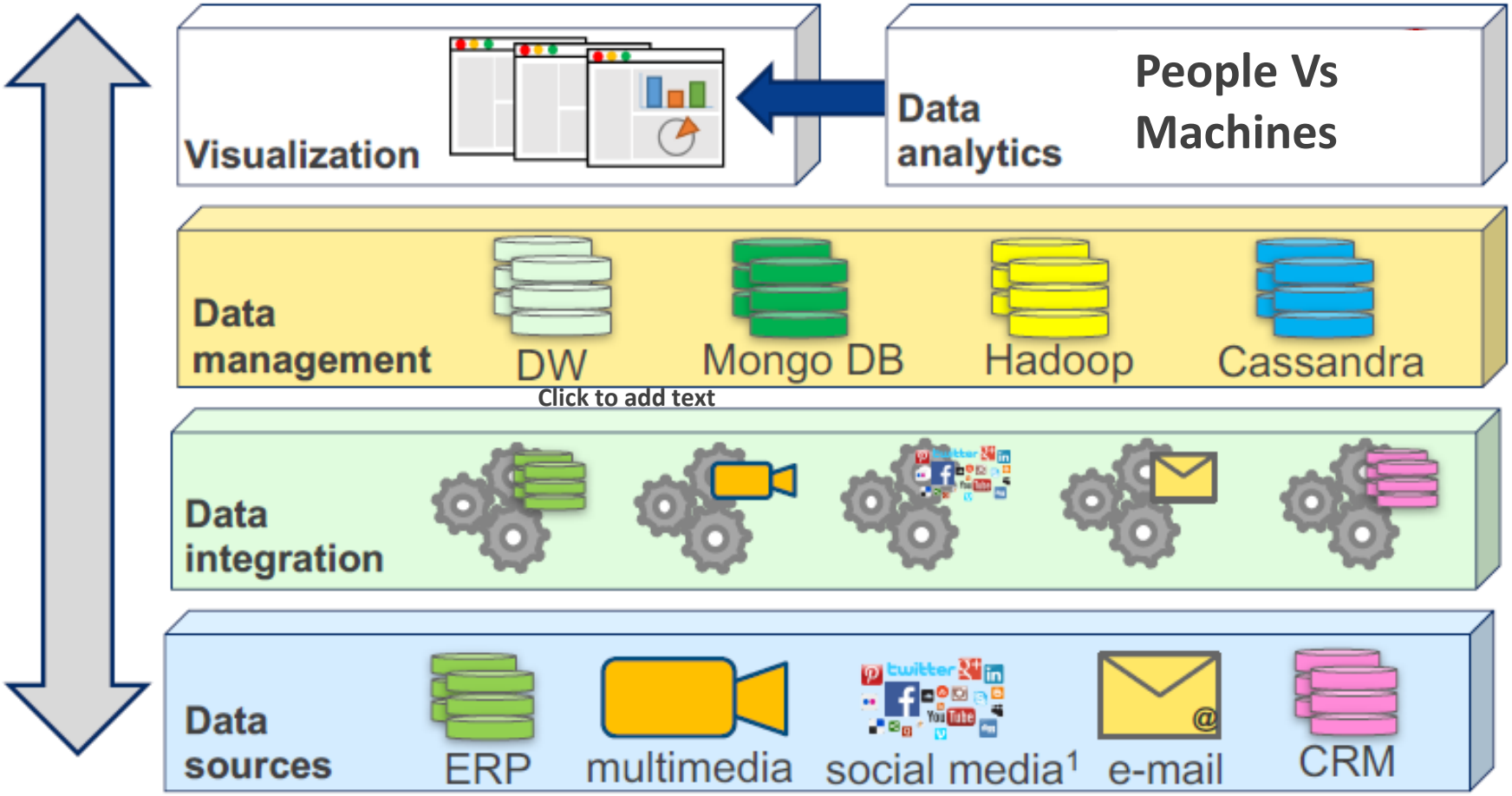
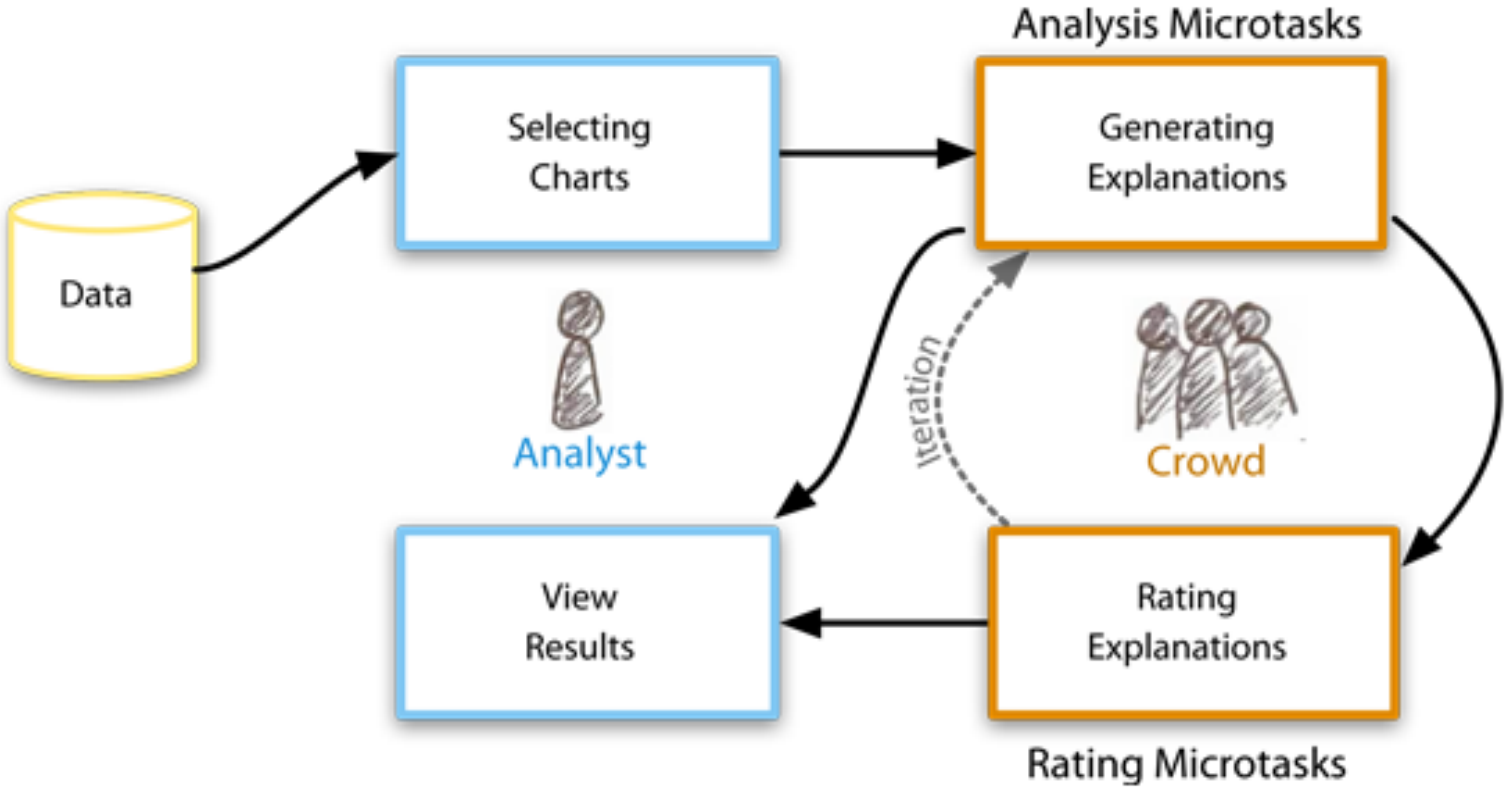


Image source: [1]commons.wikimedia.org

Types of Analytics

- **Descriptive Analytics** tells us what happened in the past and helps a business understand how it is performing by providing context to help stakeholders interpret information.
- **Diagnostic Analytics** takes descriptive data a step further and helps you understand why something happened in the past.
- **Predictive Analytics** predicts what is most likely to happen in the future and provides companies with actionable insights based on the information.
- **Prescriptive Analytics** provides recommendations regarding actions that will take advantage of the predictions and guide the possible actions toward a solution.

Crowdsourcing Data Analytics



http://vis.berkeley.edu/papers/CrowdAnalytics/

What is Machine Learning?

- Machine learning is an application of **artificial intelligence (AI)** that provides systems the ability to **automatically learn** and improve from experience without being explicitly programmed.
- Machine learning focuses on the development of "**computer programs that can access data and use it to learn for themselves**".

Useful Terminology

- Features: Distinctive treats that would help describe an item.
- Samples: Items to be processed (e.g., Classified). Samples could be documents, records, rows in a CSV file…etc. Basically, it is whatever you can describe with the distinctive traits (i.e., the features).
- Feature vector: an n-dimensional vector of distinctive treats to describe an item
- Feature Extraction: the preparation of the feature vector. Ideally transforming the data in a high dimensional space to a space of fewer dimensions.
- Training set: a set of data samples to discover predictive relationships

Useful Basic Statistics

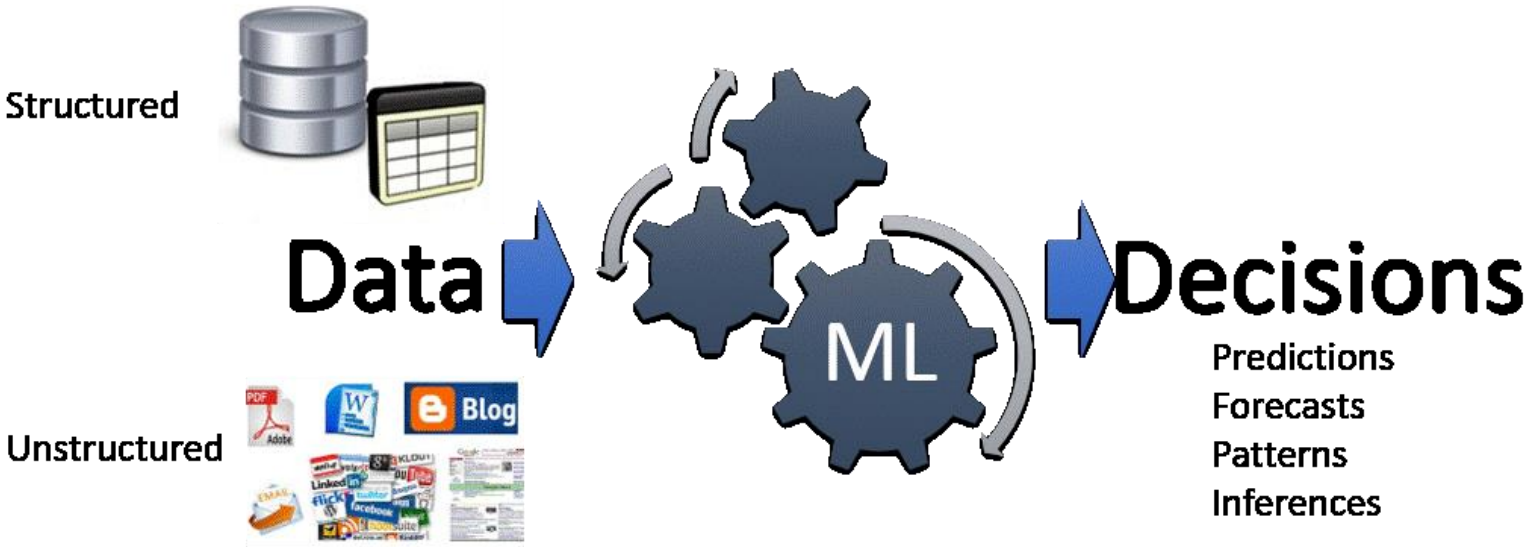
- **Mean**: The average of the dataset.
- **Median**: The middle value of an ordered dataset.
- **Mode**: The most frequent value in the dataset. If the data have multiple values that occurred the most frequently, we have a multimodal distribution.
- **Probability:** is the measure of the likelihood that an event will occur in a Random Experiment.
- **Bayes' Theorem:** describes the probability of an event based on prior knowledge of conditions that might be related to the event.
- **Range**: The difference between the highest and lowest value in the dataset.

https://towardsdatascience.com/machine-learning-probability-statistics-f830f8c09326

Useful Basic Statistics

- **Variance**: The average squared difference of the values from the mean to measure how spread out a set of data is relative to mean.
- **Standard Deviation**: The standard difference between each data point and the mean and the square root of variance.
- **Causality**: Relationship between two events where one event is affected by the other.
- **Covariance**: A quantitative measure of the joint variability between two or more variables.
- **Correlation**: Measure the relationship between two variables and ranges from *-1 to 1*, the normalized version of covariance.

Machine Learning for Data Analytics



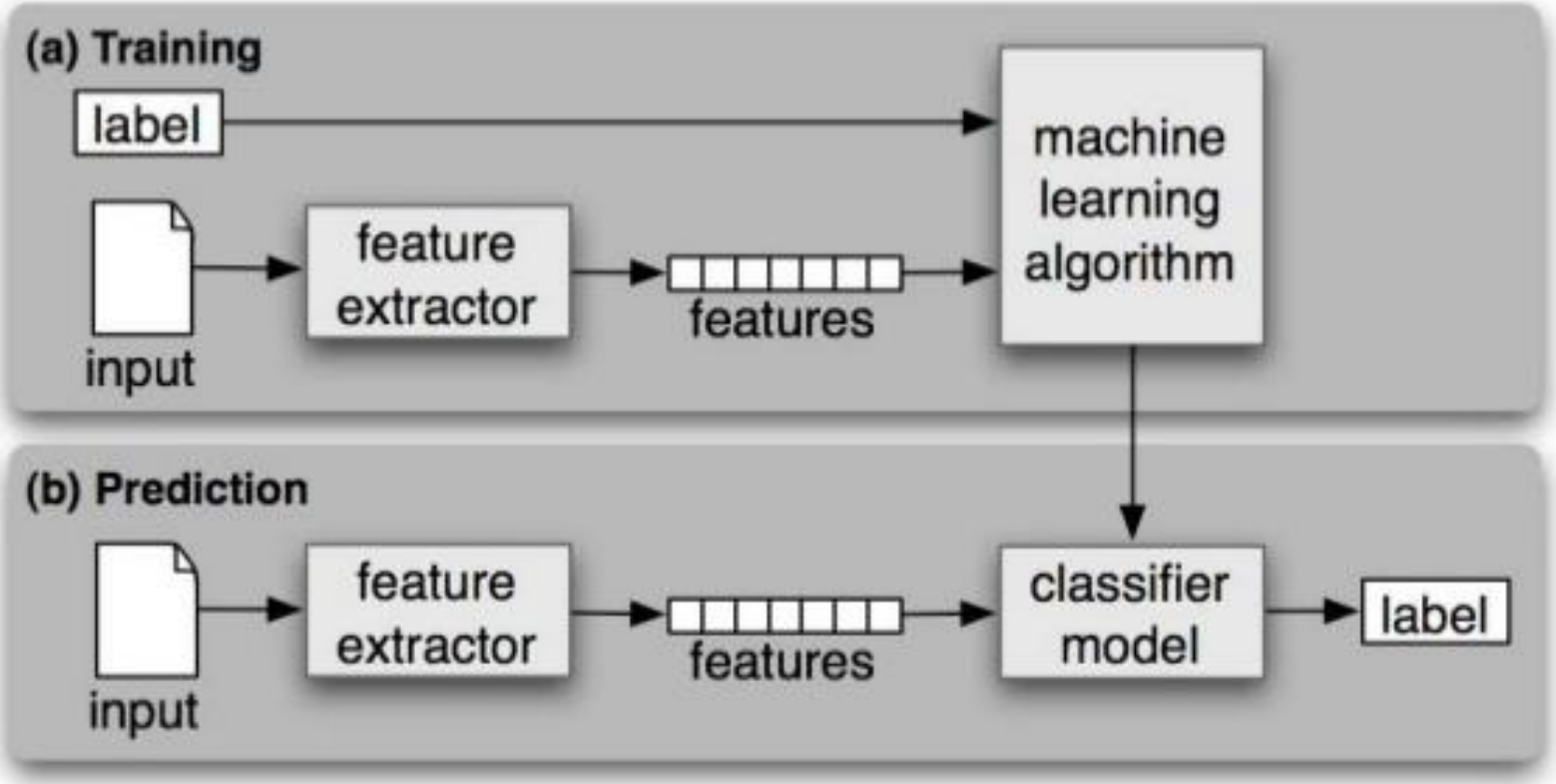
https://www.datasciencecentral.com/profiles/blogs/decoding-machine-learning-methods

Machine Learning for Data Analytics

1. **Prepare** your Data
2. **Define** and **Initialize** a Model
3. **Train** your Model (using your training dataset)
4. **Validate** the Model (by prediction using your test dataset)
5. Use it: **Explore** or **Deploy** as a web service
6. **Update** and **Revalidate**

https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Example of a General Flow



https://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937

What is an Apple?

Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc...

Features:
1. Sky Blue
2. **Logo**
3. Shape
etc...

Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc...

https://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937

Machine Learning Methods

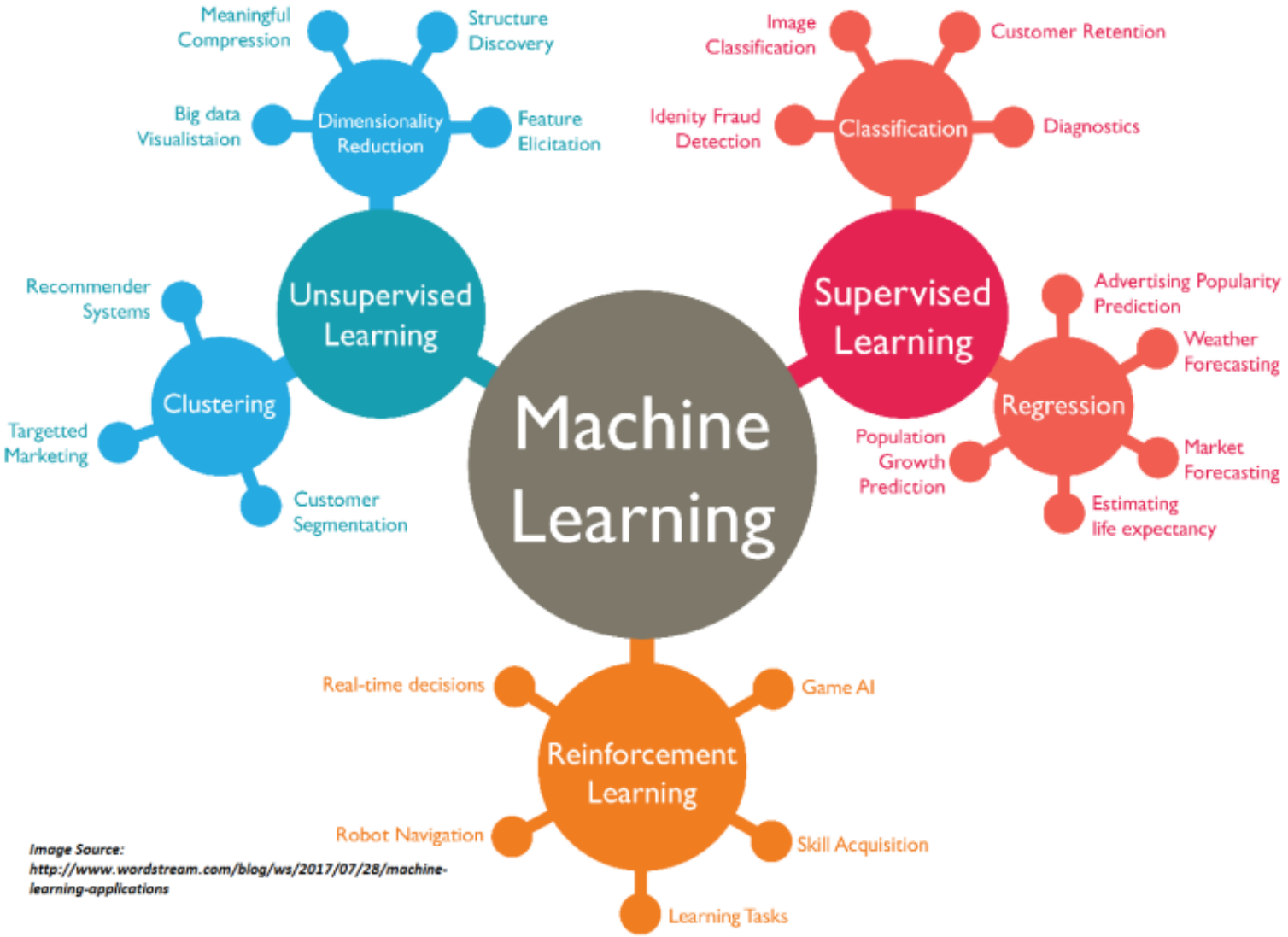
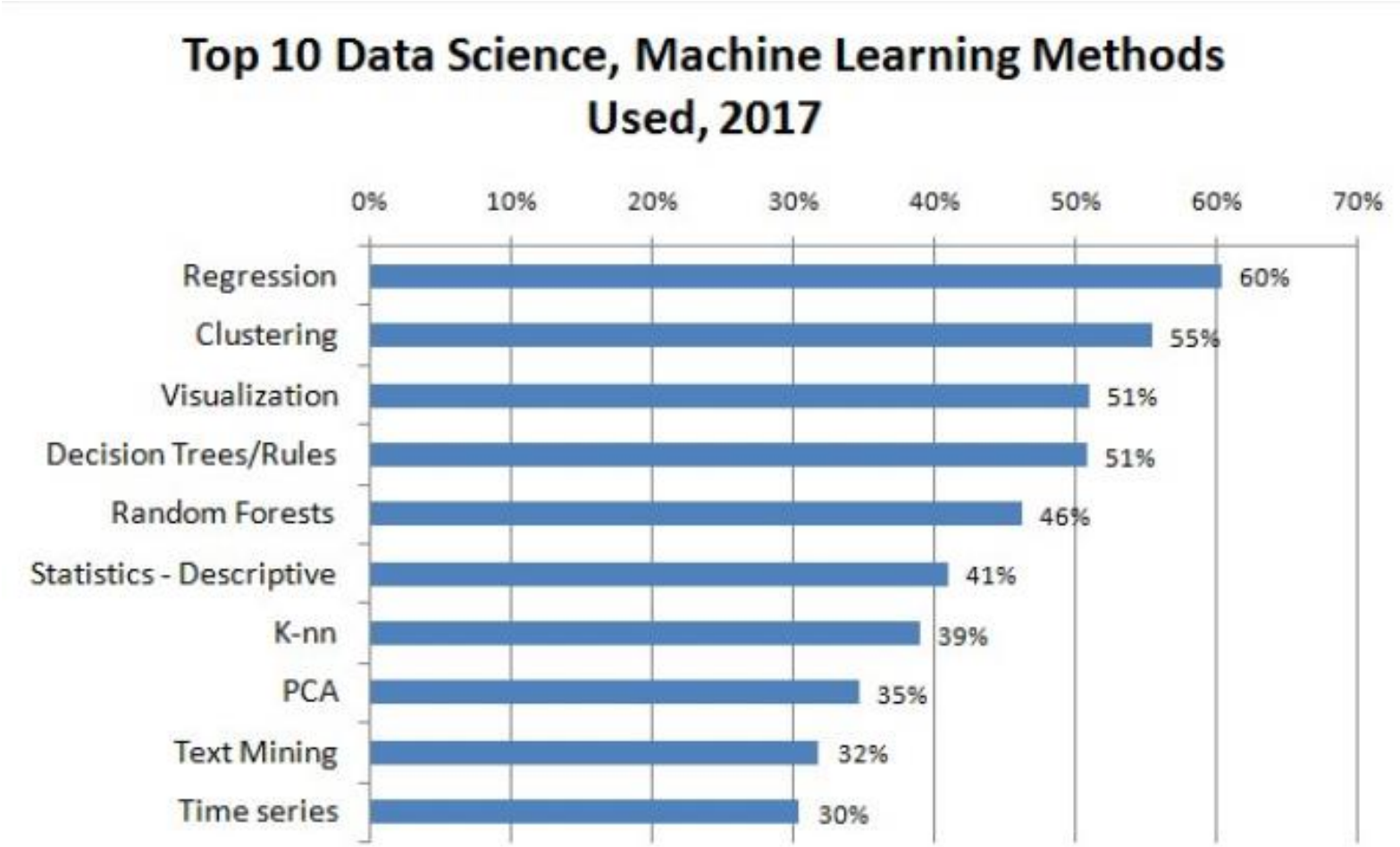


Image Source:
http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Machine Learning Methods

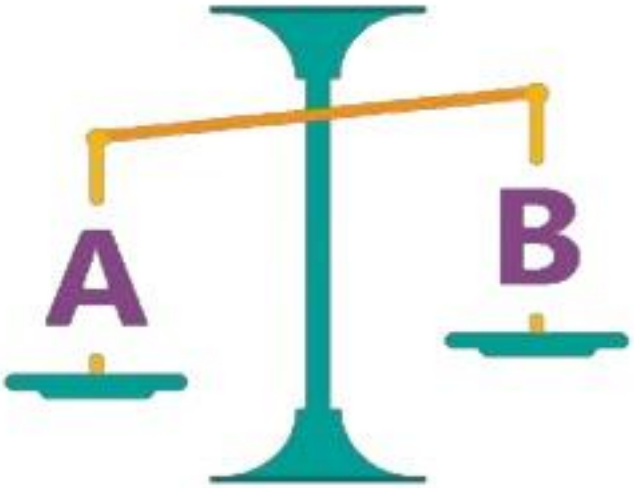


https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html

Questions Machine Learning Can Answer

1. Is this A or B?

Classification Algorithms

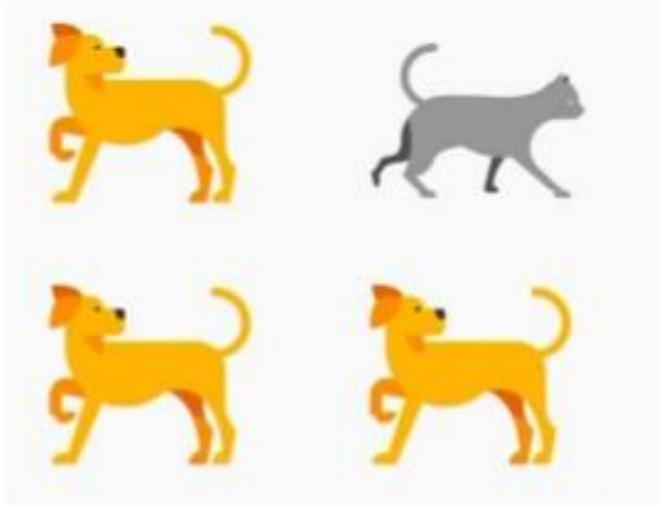


https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Questions Machine Learning Can Answer

2. Is this Weird?

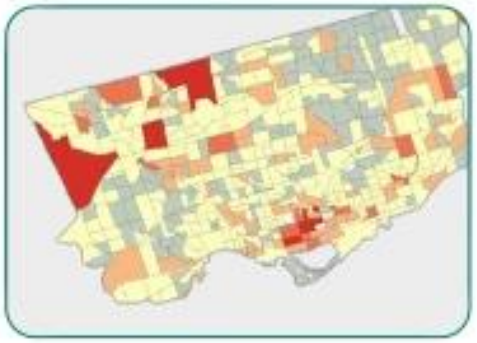
Anomaly detection algorithms



https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Questions Machine Learning Can Answer

3. How much? How many?



Regression algorithms

https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Questions Machine Learning Can Answer

4. How is this organized?

Clustering algorithms

https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Questions Machine Learning Can Answer

5. What should I do now?

Reinforcement learning algorithms

https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

Regression Analysis

https://www.slideshare.net/Simplilearn/linear-regression-analysis-linear-regression-in-python-machine-learning-algorithms-simplilearn?qid=c0cbd932-3ad1-45ec-b123-36805981982d&v=&b=&from_search=4

How Linear Regression Works

$$\hat{Y} = f(X) + \epsilon$$

X (input) = Assignment Results
Y (output) = Final Exam Mark
f = function which describes the relationship between X and Y
e (epsilon) = Random error term (positive or negative) with a mean zero (there are move assumptions for our residuals, however we won't be covering them)

https://hackernoon.com/supervised-machine-learning-linear-regression-in-python-541a5d8141ce

Linear Regression Example

Training Set

StudentID	Assignment_Mark (X)	Final_Exam_Mark (Y)
1292393	80	90
1823812	70	53
281823	63	74
....	...	.....
183823	58	63
238381	54	61

https://hackernoon.com/supervised-machine-learning-linear-regression-in-python-541a5d8141ce

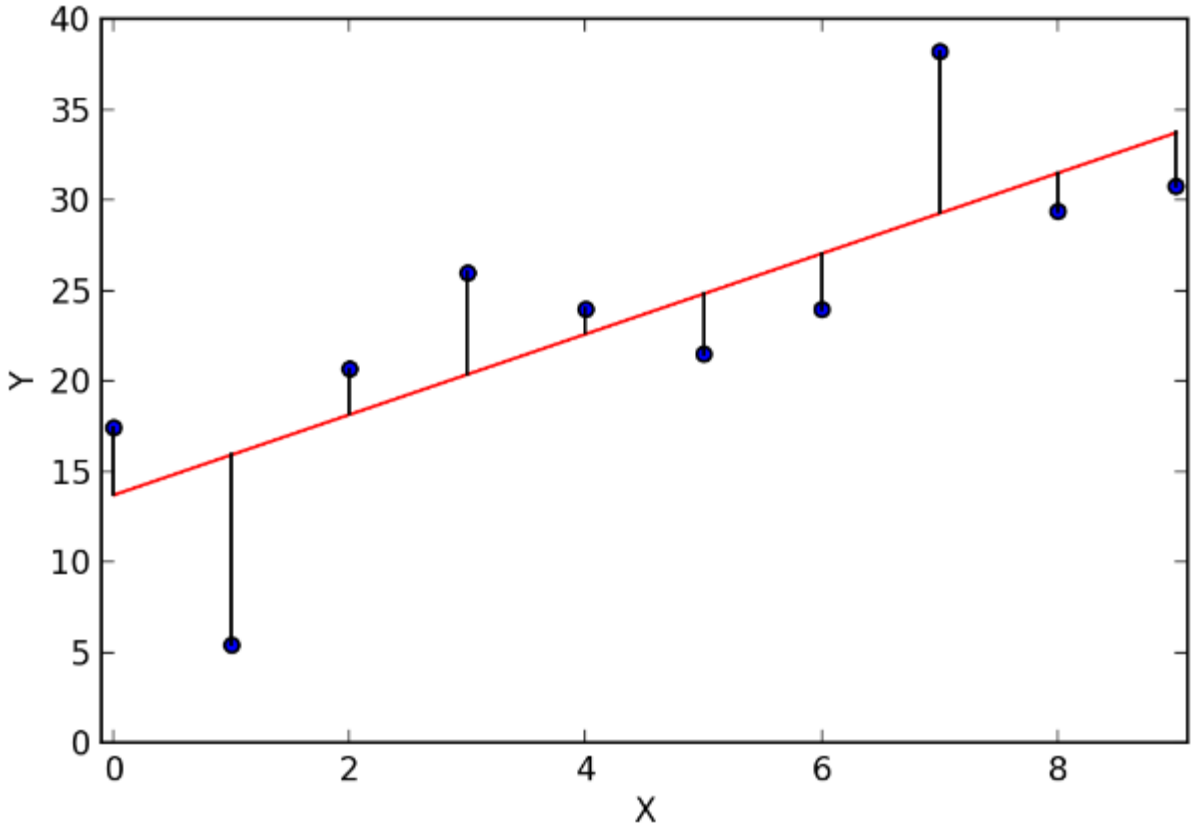
Linear Regression Example

Test Set

StudentID	Assignment_Mark (X)	Final_Exam_Mark (Y)
184712	80	???
937217	70	???
...	...	???
836162	63	???

https://hackernoon.com/supervised-machine-learning-linear-regression-in-python-541a5d8141ce

Linear Regression Example



Where Y is our Final Exam Mark, and X is our Assignment Mark

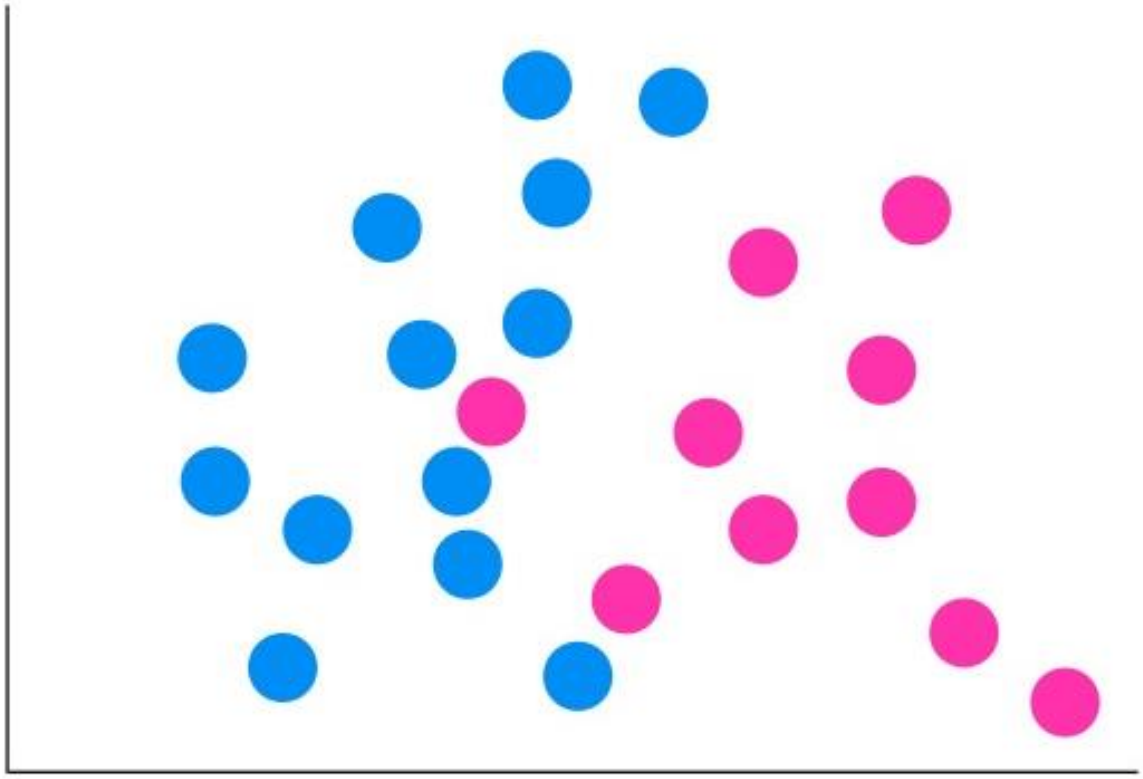
https://hackernoon.com/supervised-machine-learning-linear-regression-in-python-541a5d8141ce

Classification

- Supervised Learning
- You need the data labelled with the correct answer to train the algorithm
- Trained classifiers then can map input data to a category.

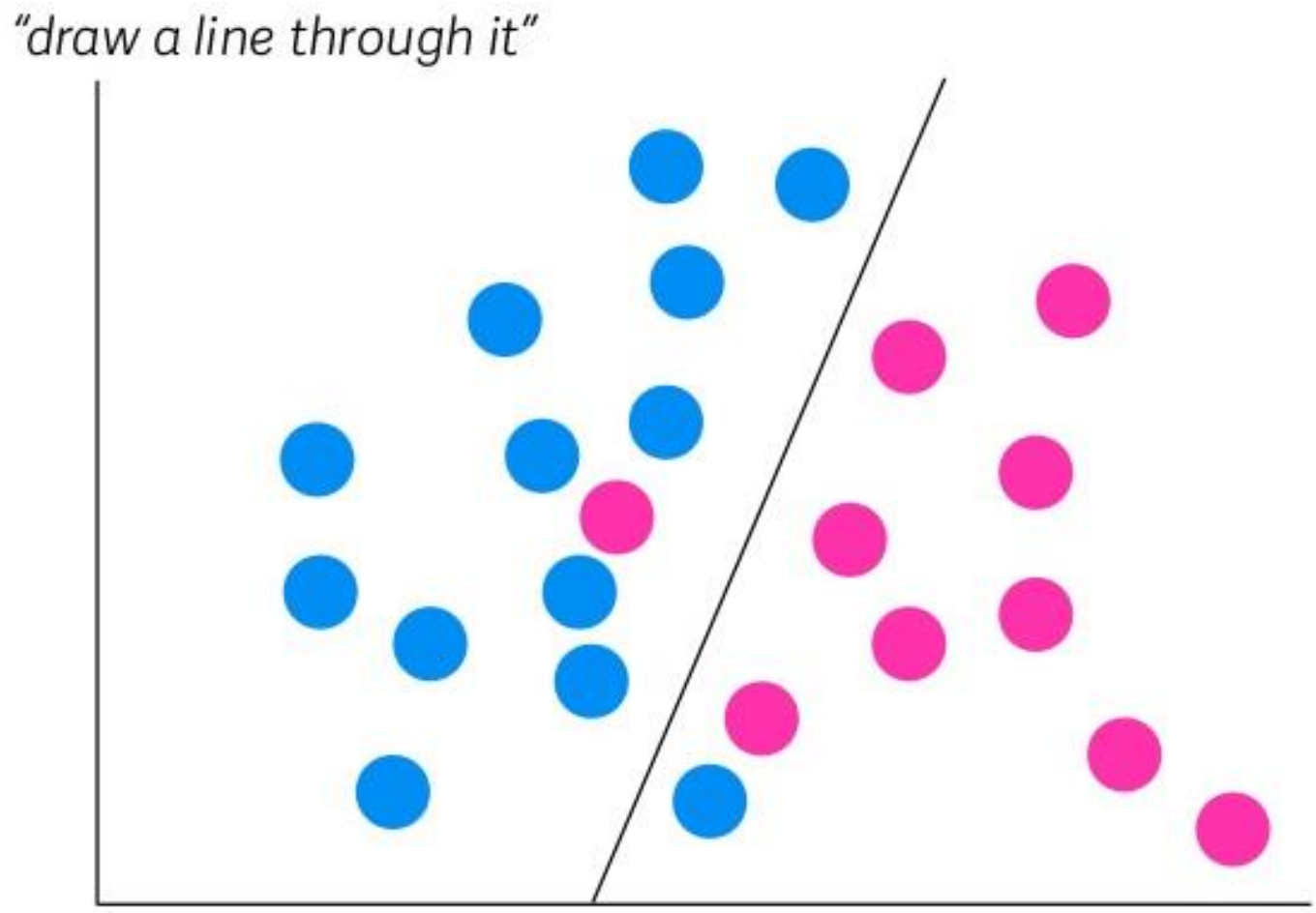
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Classification



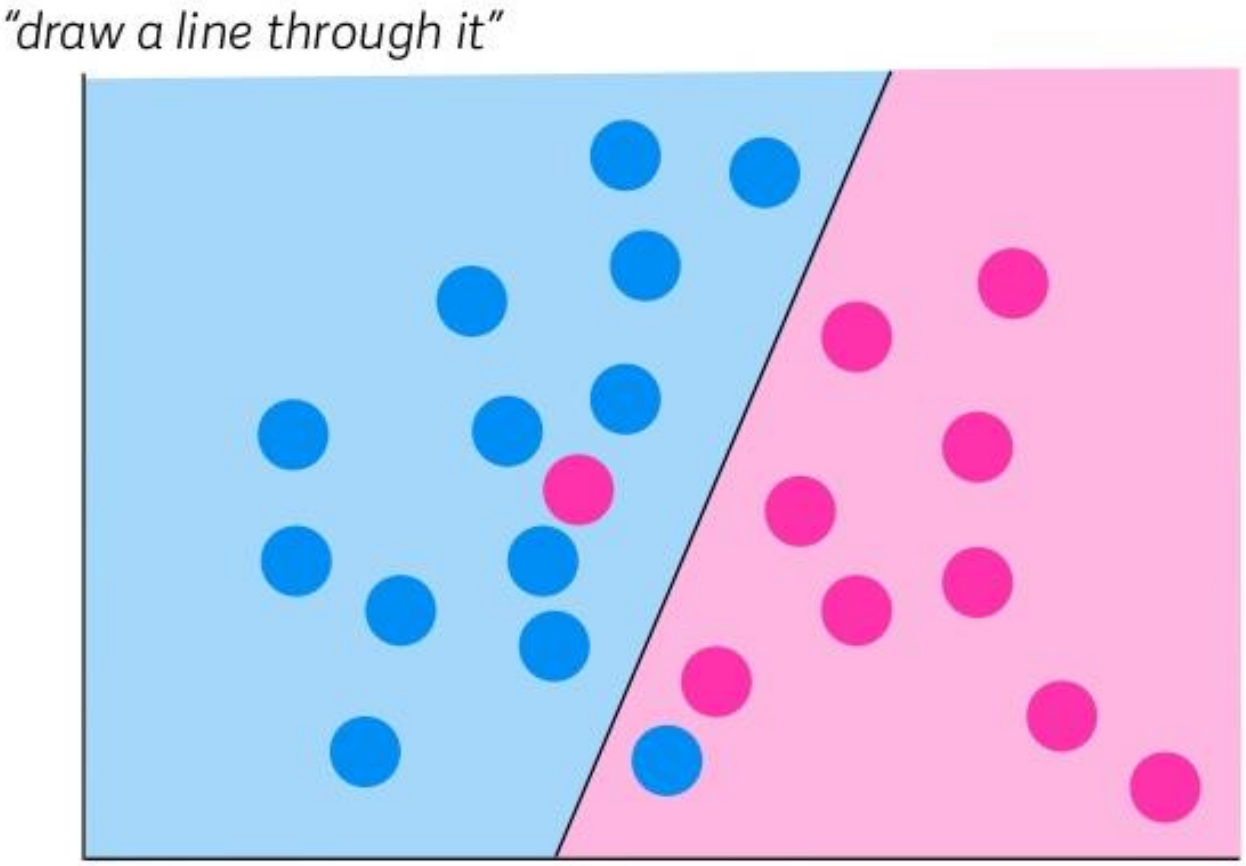
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Classification



https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Classification



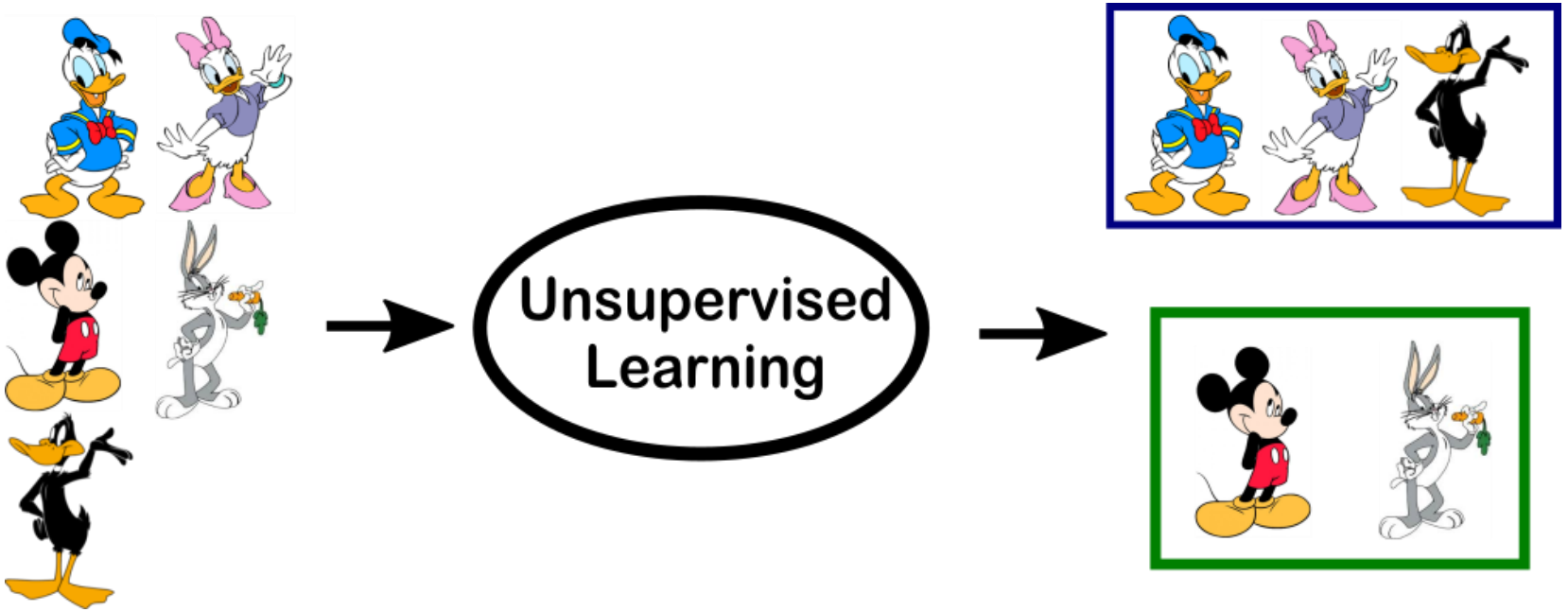
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Clustering

- Unsupervised Learning
- Automated grouping of objects into so called clusters
- Objects of the same group are similar
- Different groups are dissimilar

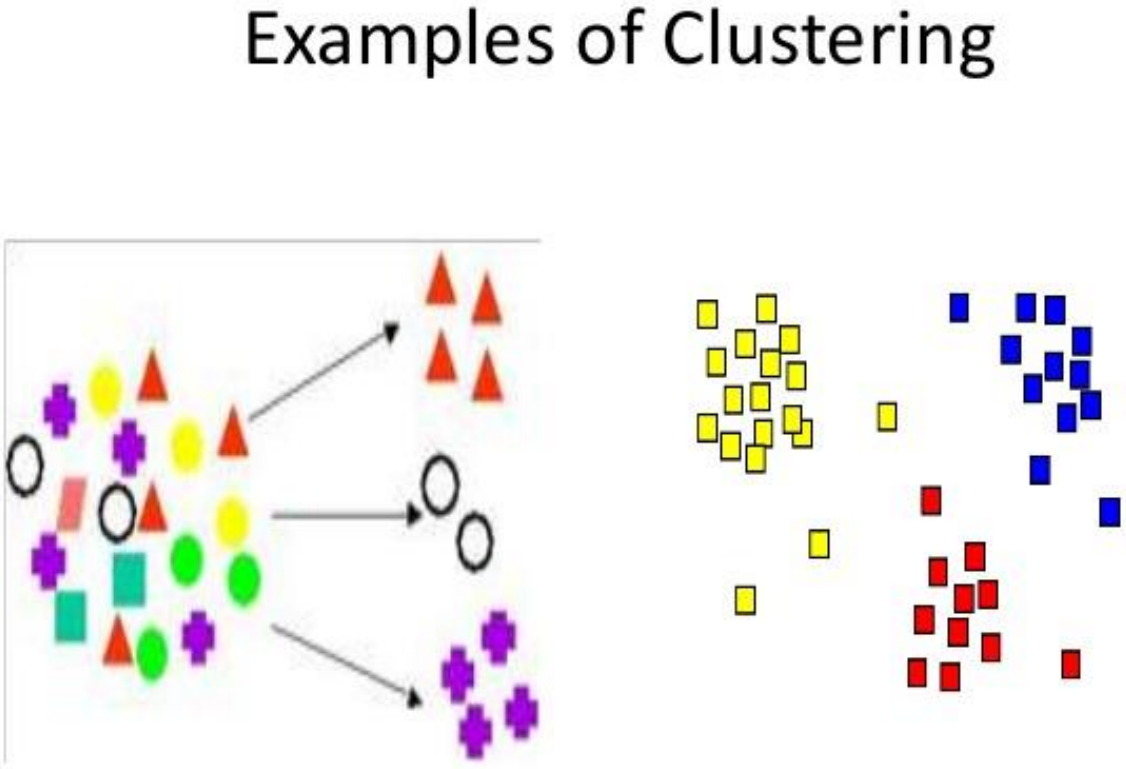
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Clustering



https://https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51

Clustering



https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Popular Machine Learning Tools

TensorFlow

Machine Learning Tools

Favorites
78

Stacks
495

I Use This

Fans 406 | Votes 52 | Jobs 246

scikit-learn

Machine Learning Tools

Favorites
13

Stacks
209

I Use This

Fans 153 | Votes 18 | Jobs 147

PredictionIO

Machine Learning Tools

Favorites
13

Stacks
31

I Use This

Fans 35 | Votes 4 | Jobs 0

Hacker News, Reddit, Stack Overflow Stats	3.89K 3.26K 32.7K	- 912 12.3K	422 114 181
GitHub Stats	*No public GitHub repository stats available*	30.5K 15K about 3 hours ago	11.4K 1.87K about 18 hours ago

Popular Machine Learning Tools

- TensorFlow
- scikit-learn
- PredictionIO

Further Reading and Useful Resources

- Book: Mastering Machine Learning with Scikit-Learn, Second Edition. Gavin Hackeling
- https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html
- https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0
- https://towardsdatascience.com/machine-learning-probability-statistics-f830f8c09326
- https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning
- http://gael-varoquaux.info/scikit-learn-tutorial/

Q&A