COMP9321: Data services engineering

Week 6: Introduction to Data Analytics

Term 1, 2020

By Mortada Al-Banna, CSE UNSW

Data Driven Organizations

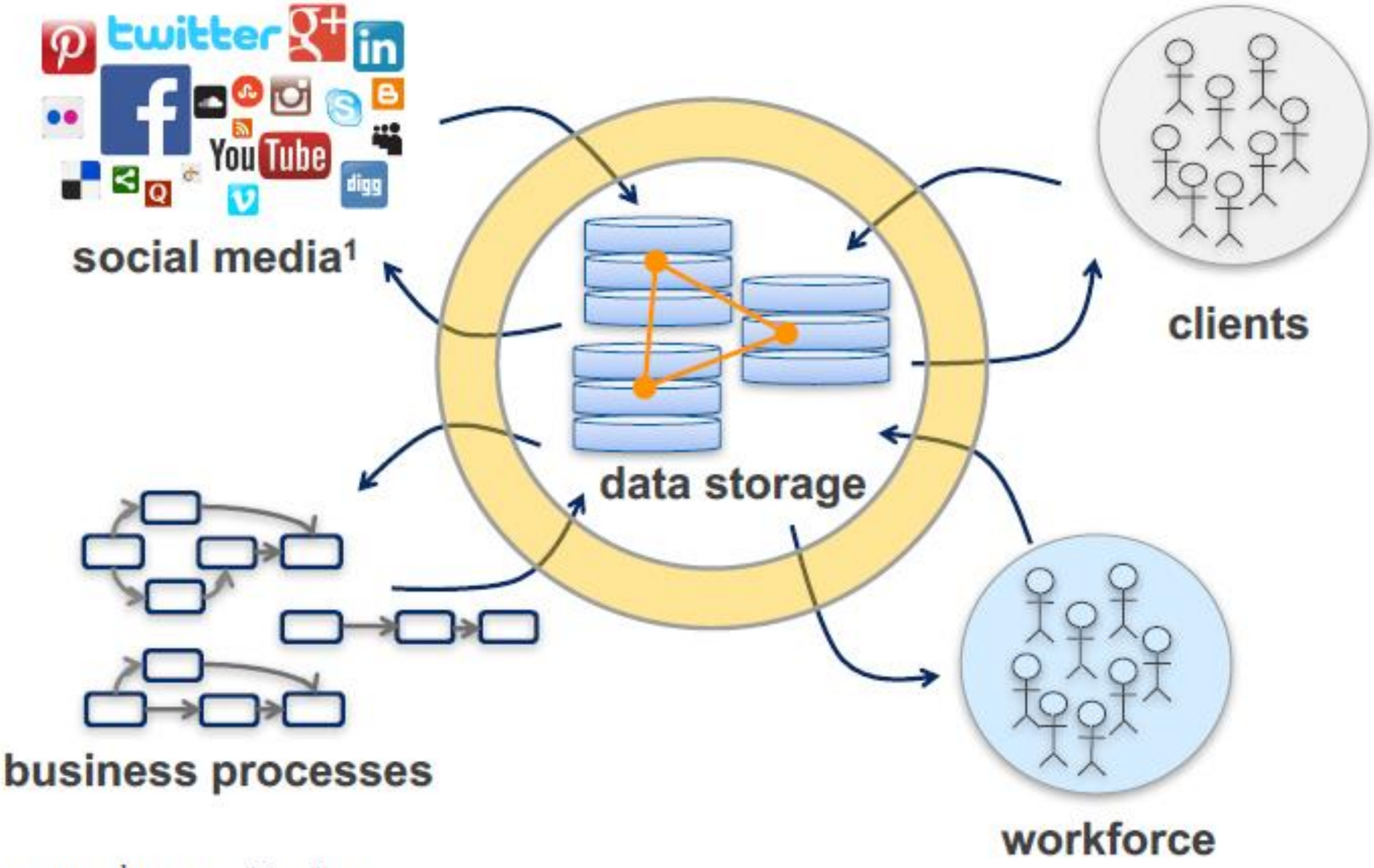


image source: [1]commons.wikimedia.org

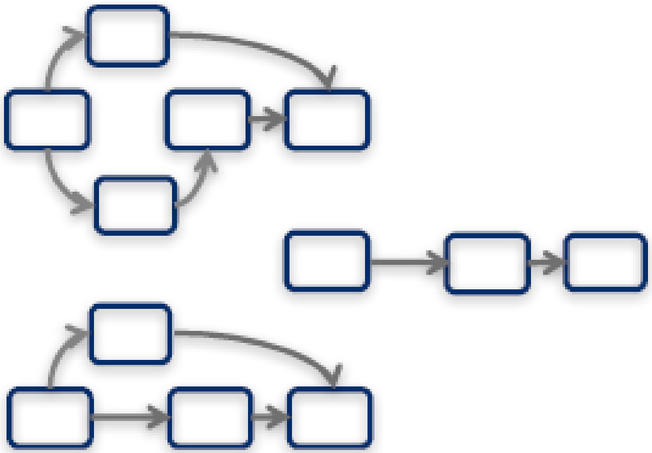
Data Driven Organizations and Data Analytics

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

Clients

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



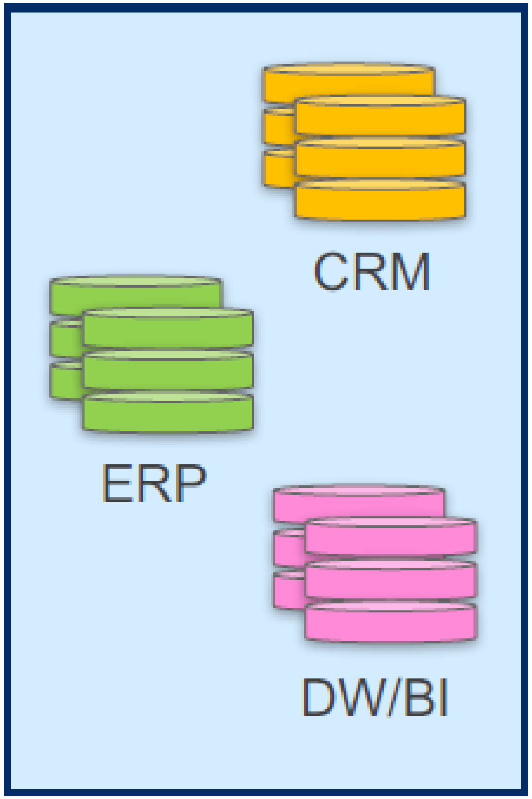
business processes

Data Driven Organizations and Data Analytics

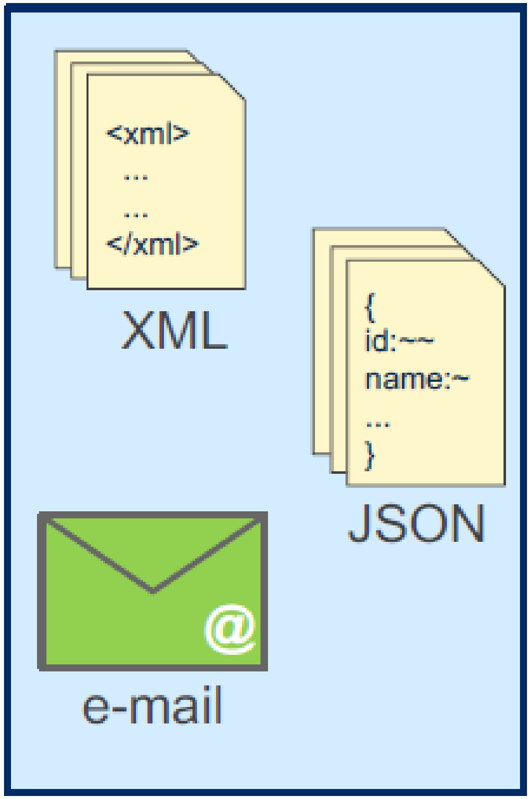
- Product and service advertisement
- Sentiment analysis
- Demographics analysis
- Virality
- Social network insights

social media[1]

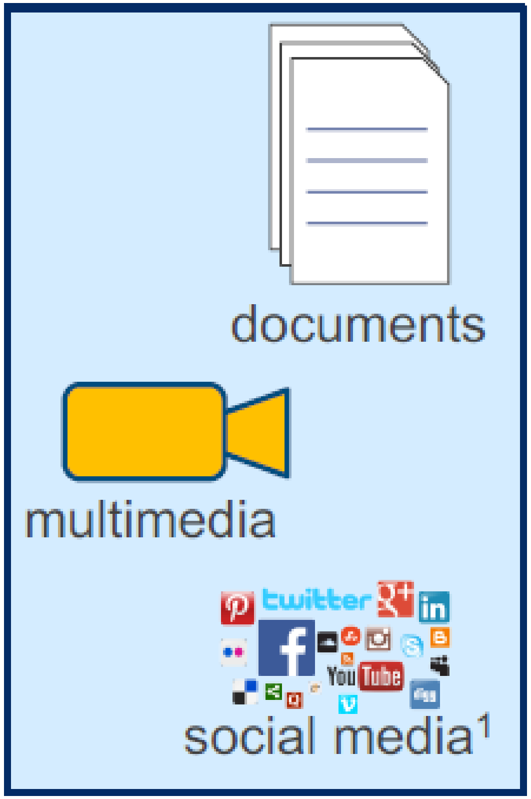
Data Used for Analytics



structured data



semi-structured data



unstructured data

image source: [1]commons.wikimedia.org

Data Used for Analytics

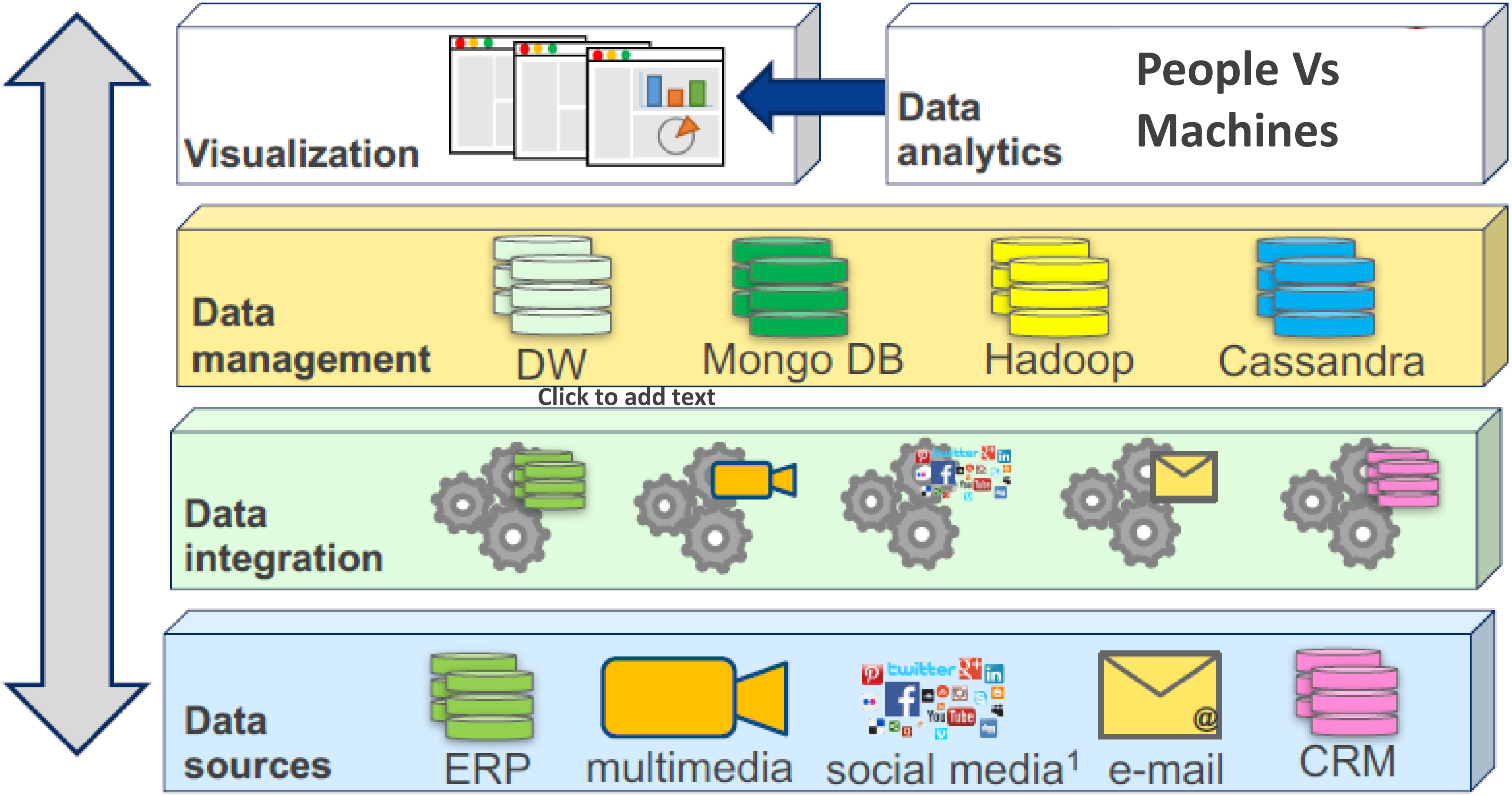
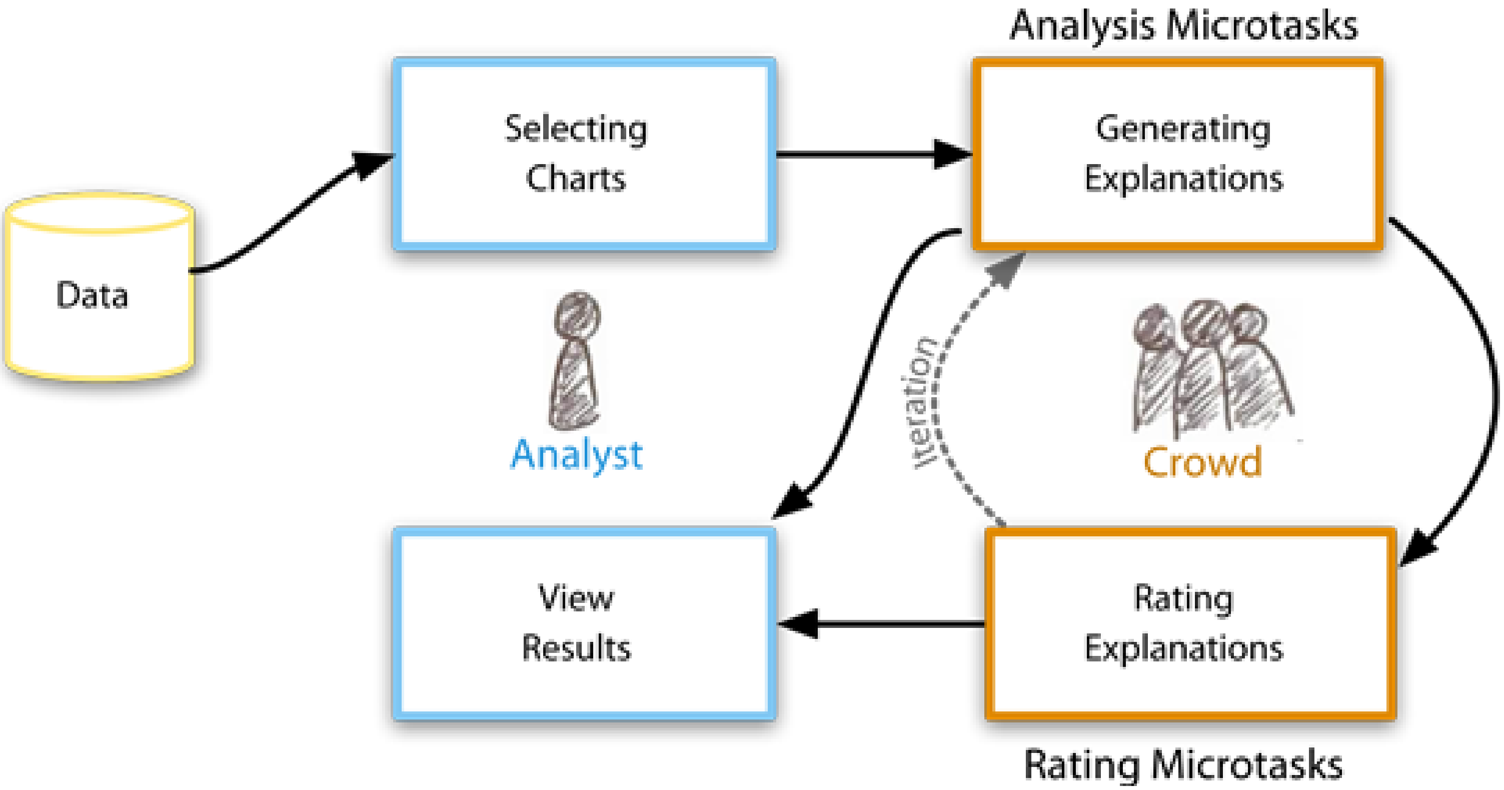


image source: [1]commons.wikimedia.org

Crowdsourcing Data Analytics

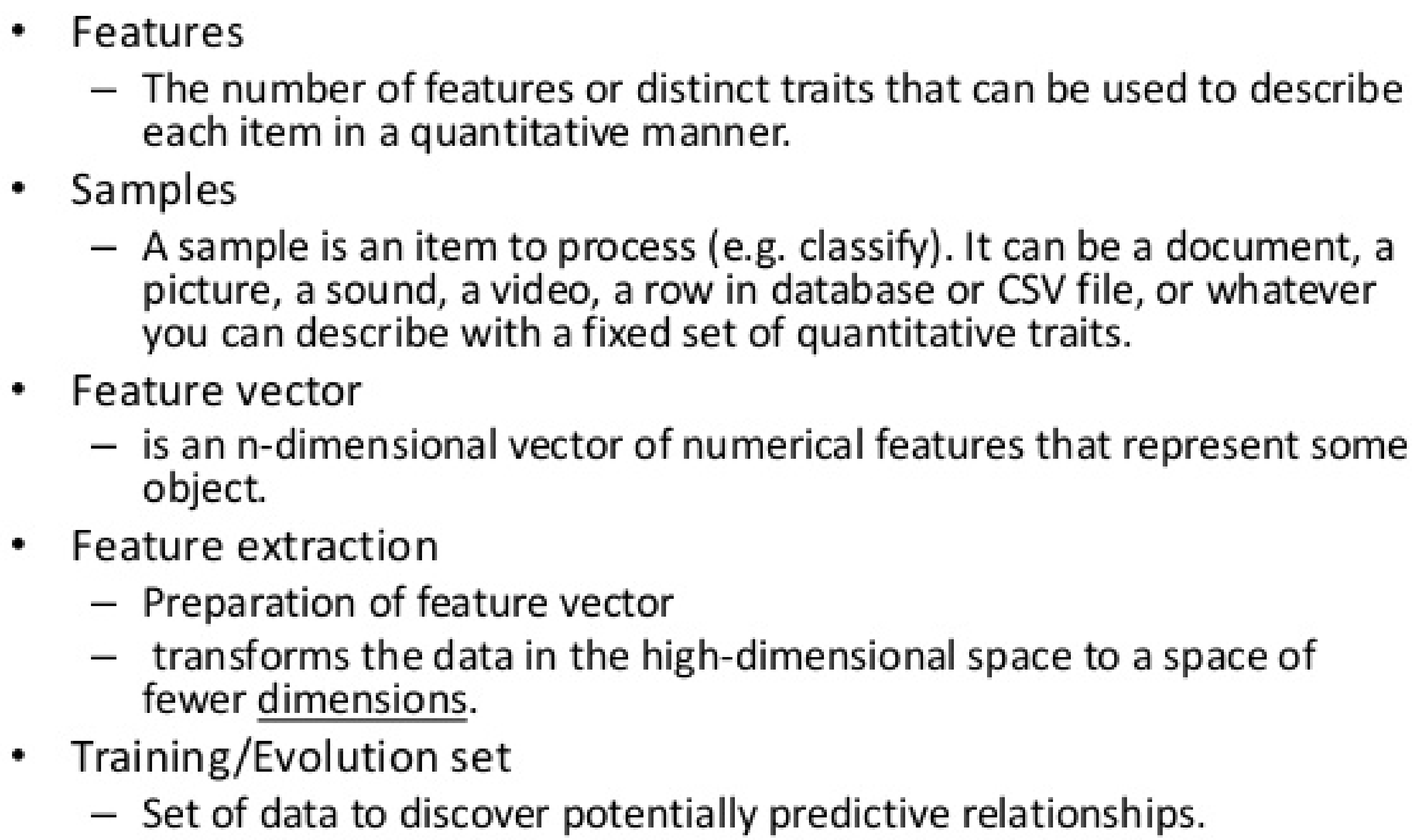


http://vis.berkeley.edu/papers/CrowdAnalytics/

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 learn for themselves**".

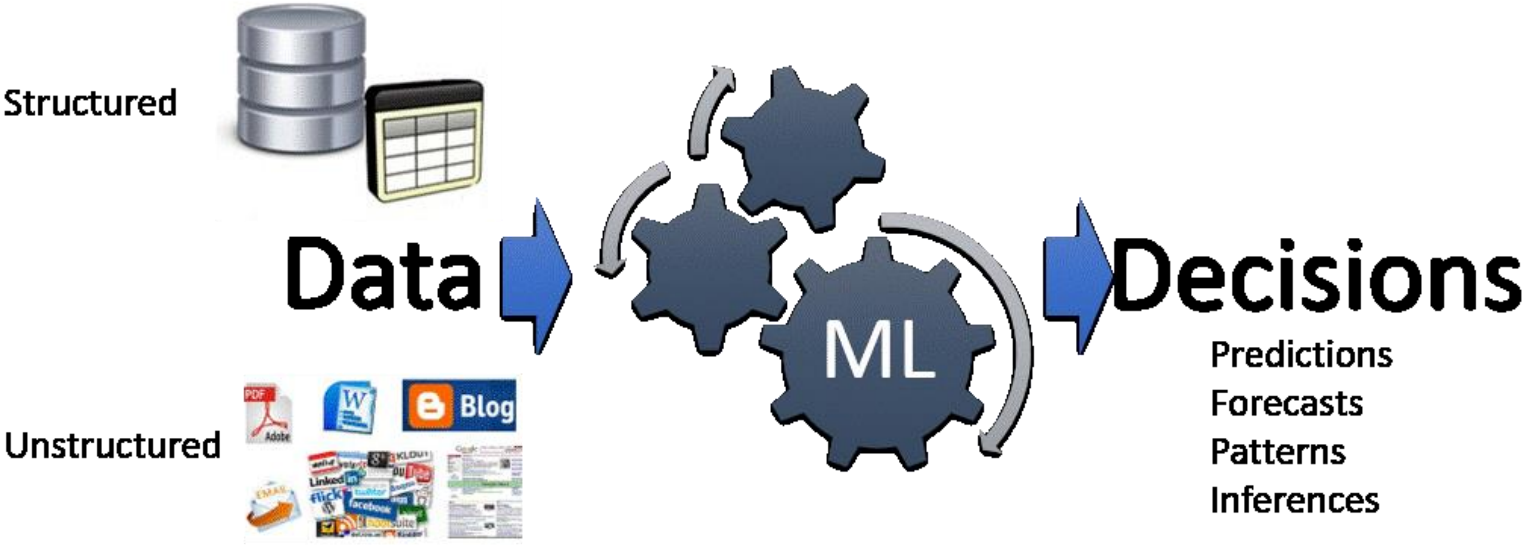
Useful Terminology



- Features
  - The number of features or distinct traits that can be used to describe each item in a quantitative manner.
- Samples
  - A sample is an item to process (e.g. classify). It can be a document, a picture, a sound, a video, a row in database or CSV file, or whatever you can describe with a fixed set of quantitative traits.
- Feature vector
  - is an n-dimensional vector of numerical features that represent some object.
- Feature extraction
  - Preparation of feature vector
  - transforms the data in the high-dimensional space to a space of fewer dimensions.
- Training/Evolution set
  - Set of data to discover potentially predictive relationships.

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

Machine Learning for Data Analytics



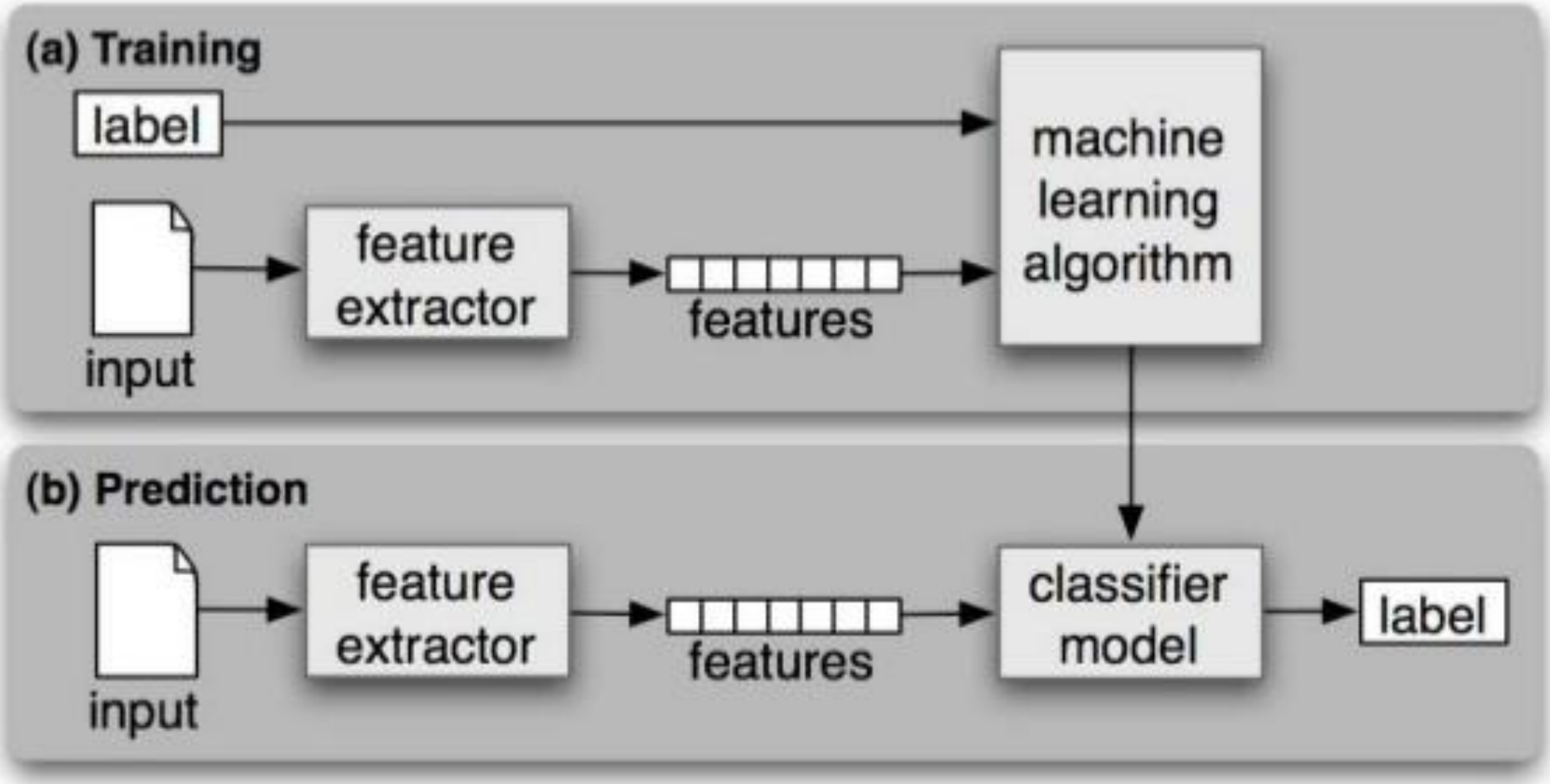
https://www.datasciencecentral.com/profiles/blogs/decoding-machine-learning-methods

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

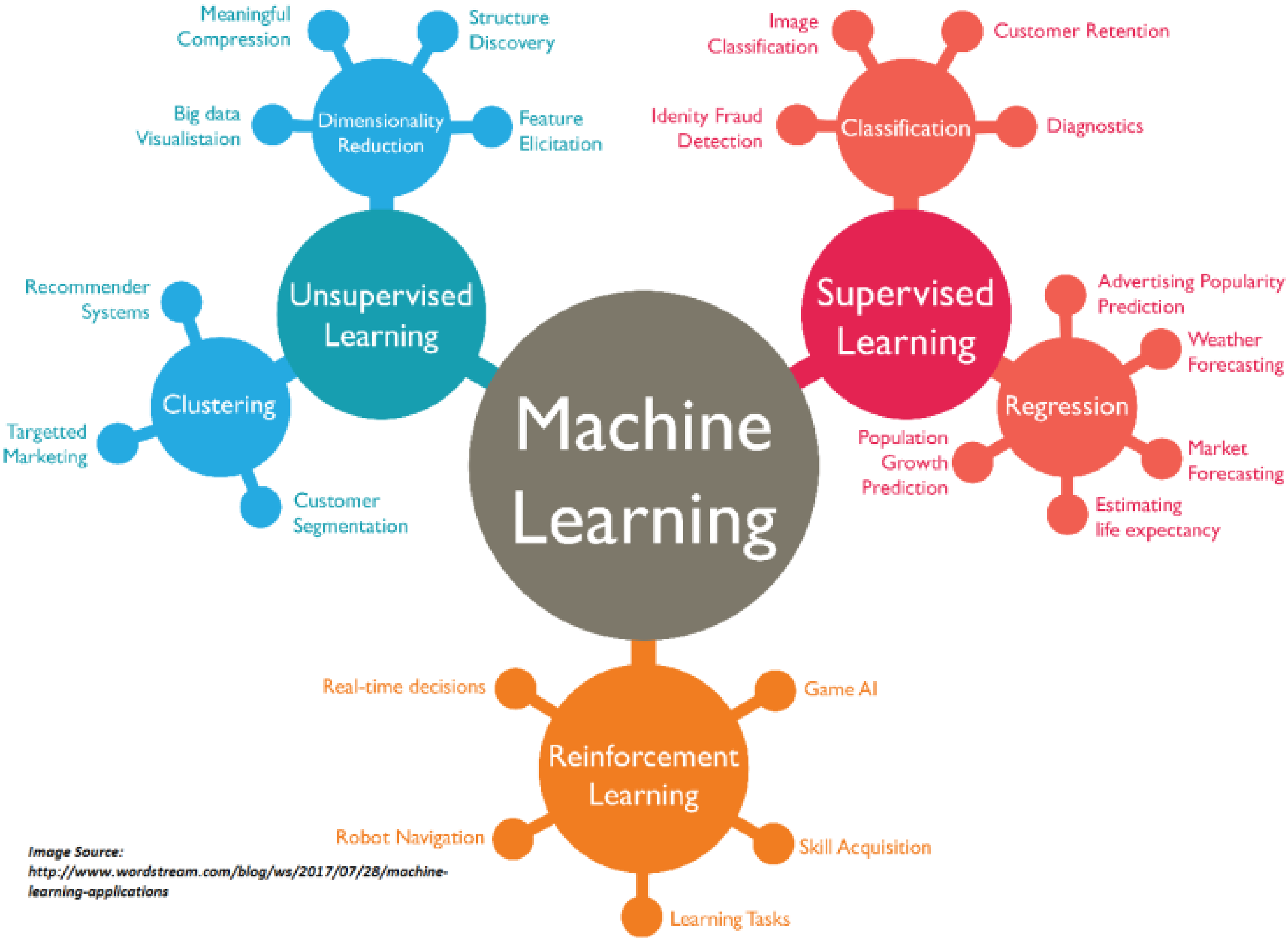
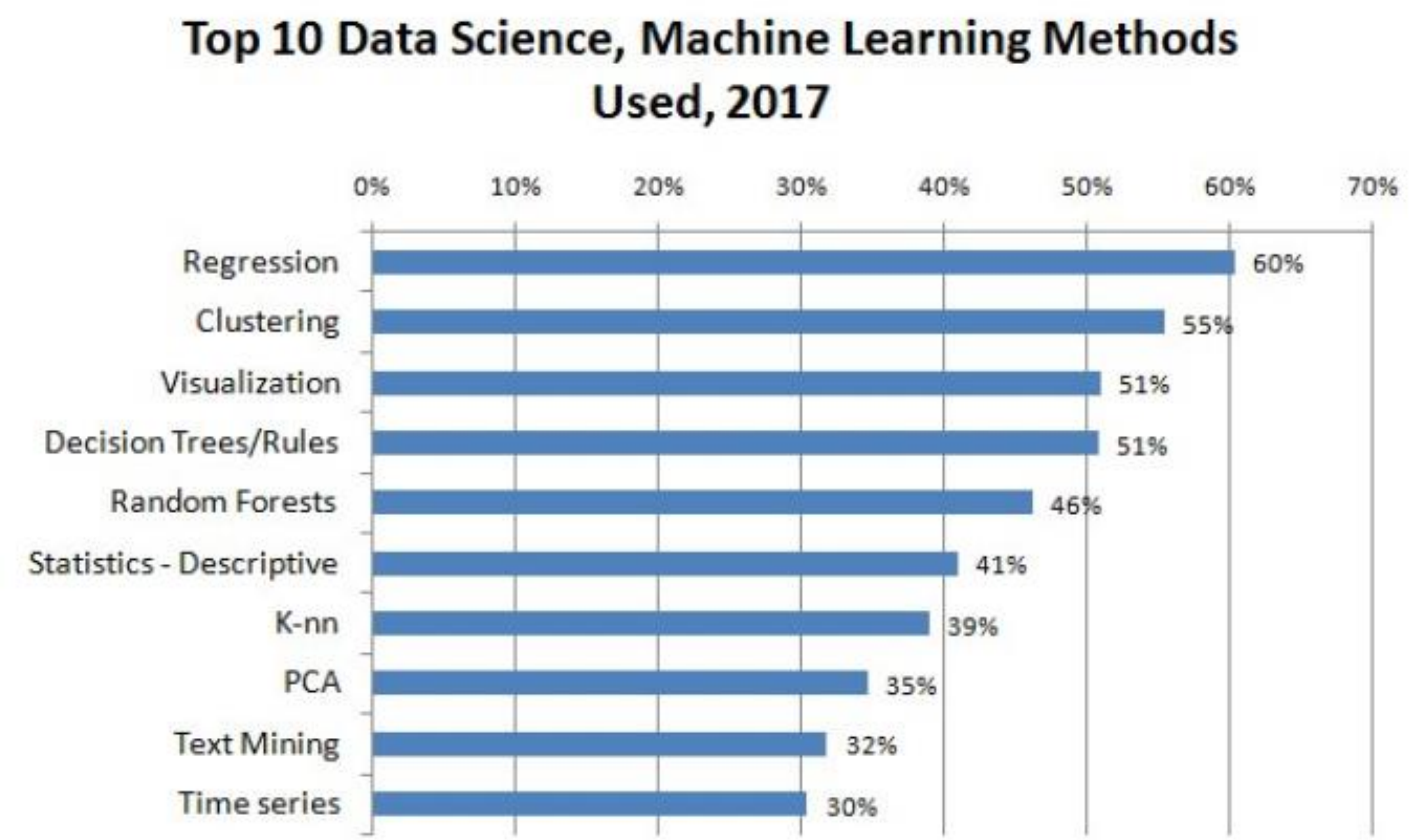


Image Source: http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

Machine Learning Methods

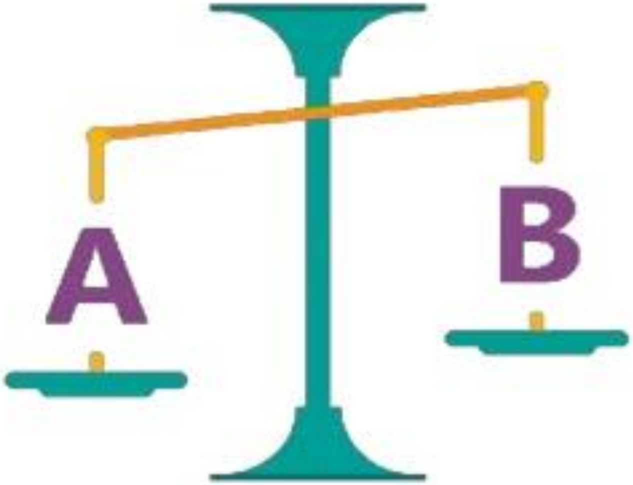


https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html

Questions Machine Learning Can Answer

1. Is this A or B?

Classification Algorithms

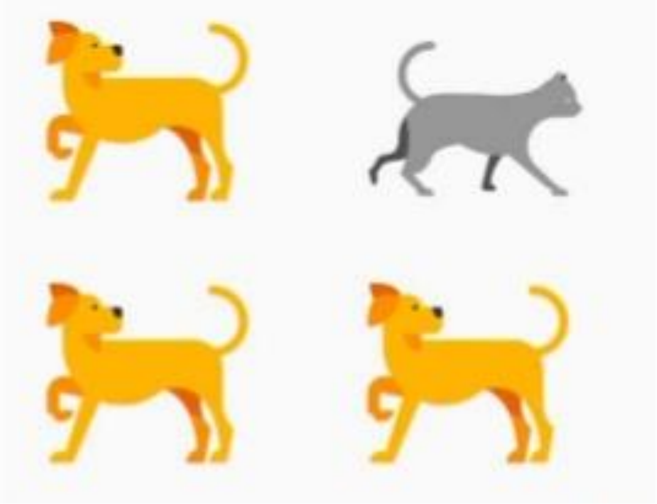


https://www.slideshare.net/SetuChokshi/machine-learning-101-74017810

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

Monday
72°
Tuesday
?

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

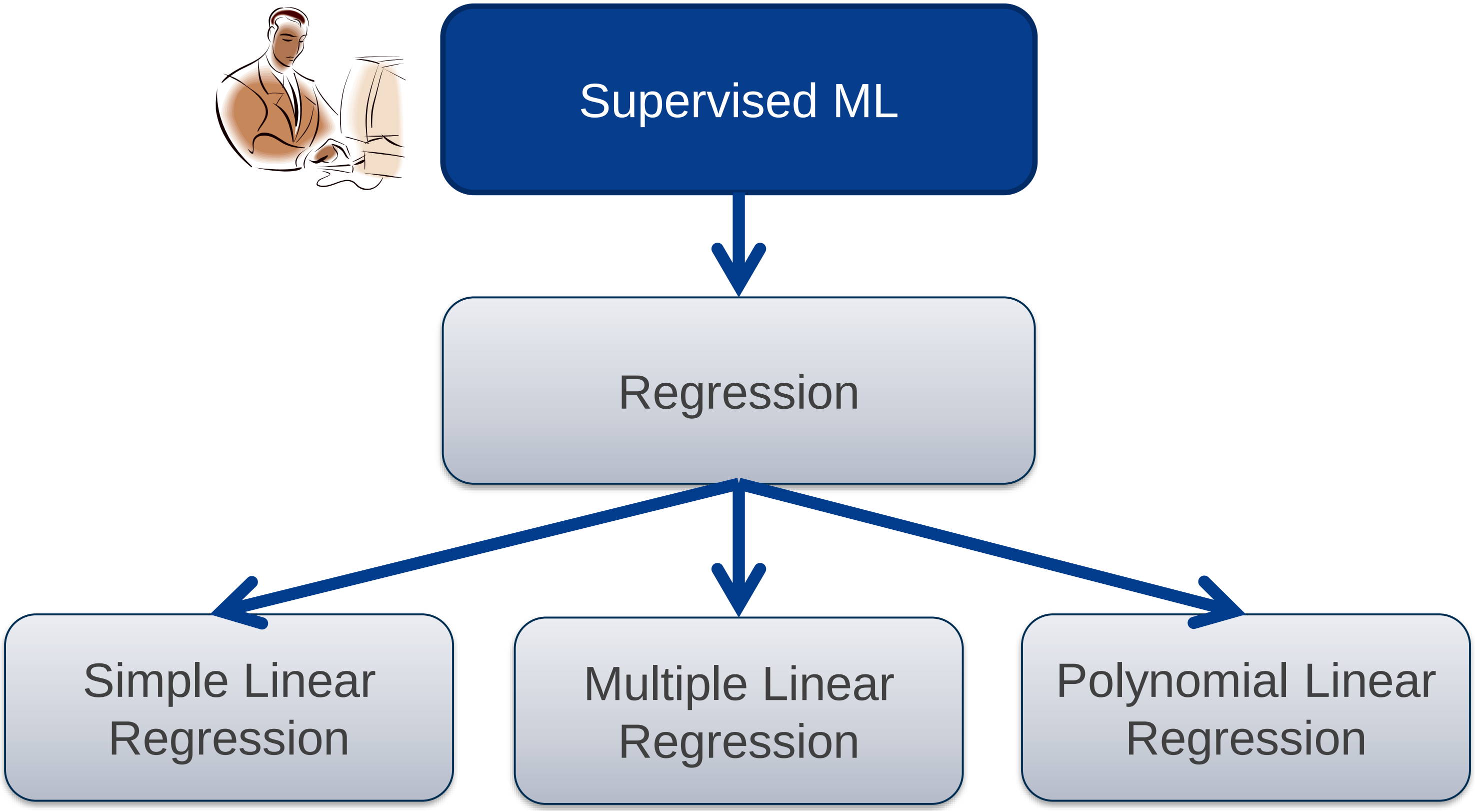
Questions Machine Learning Can Answer

5. What should I do now?

Reinforcement learning algorithms

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

Regression Analysis



Supervised ML

Regression

Simple Linear Regression

Multiple Linear Regression

Polynomial Linear Regression

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

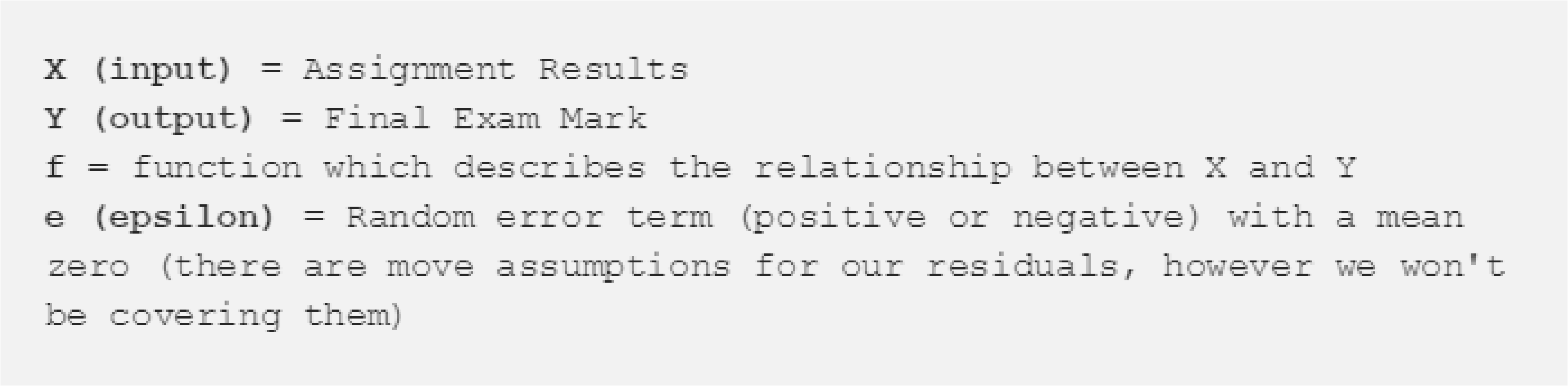
Linear Regression (terminology)

- **Independent Variables (features):** An independent variable is a variable that is manipulated to determine the value of a dependent variable. Simply, they are the features which we want to use to predict some given value of Y. It can be also called an explanatory variable
- **Dependent Variable(target):** The dependent variable depends on the values of the independent variable. Simply put, it is the feature which we are trying to predict. This can also be commonly known as a response variable.

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

How Linear Regression Works

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



X (input) = Assignment Results

Y (output) = Final Exam Mark

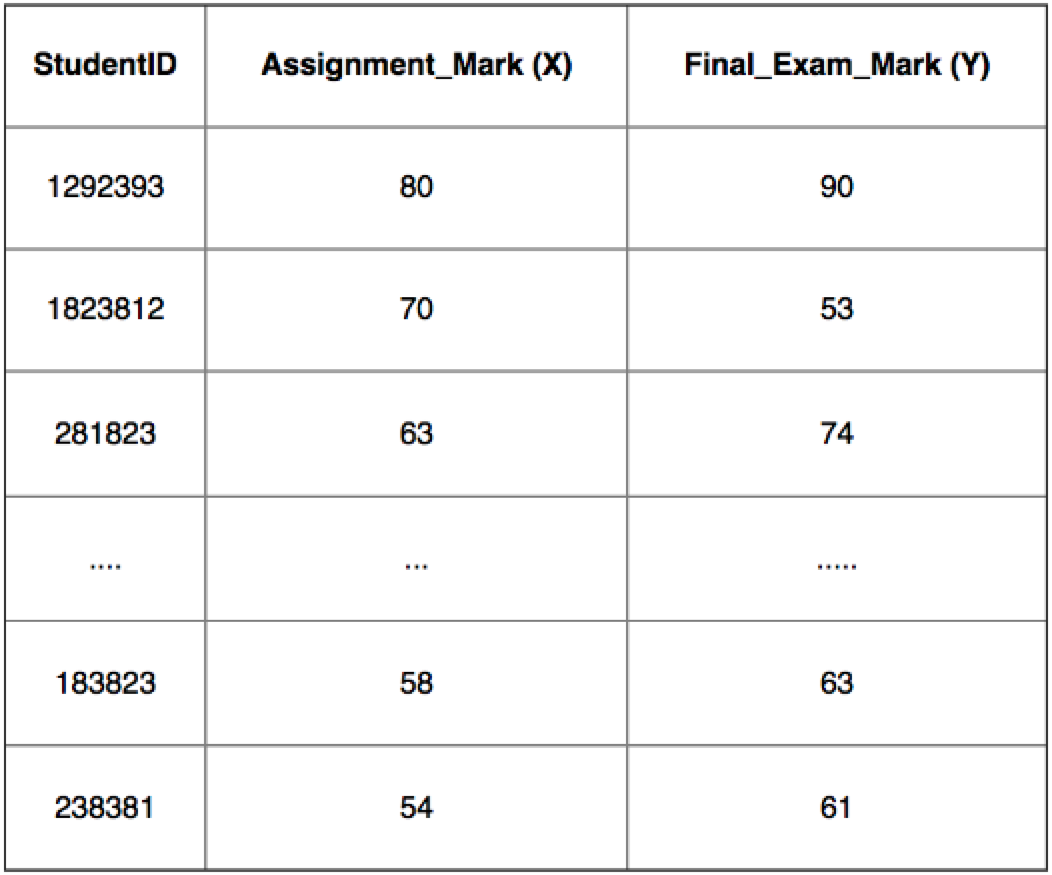
f = function which describes the relationship between X and Y

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

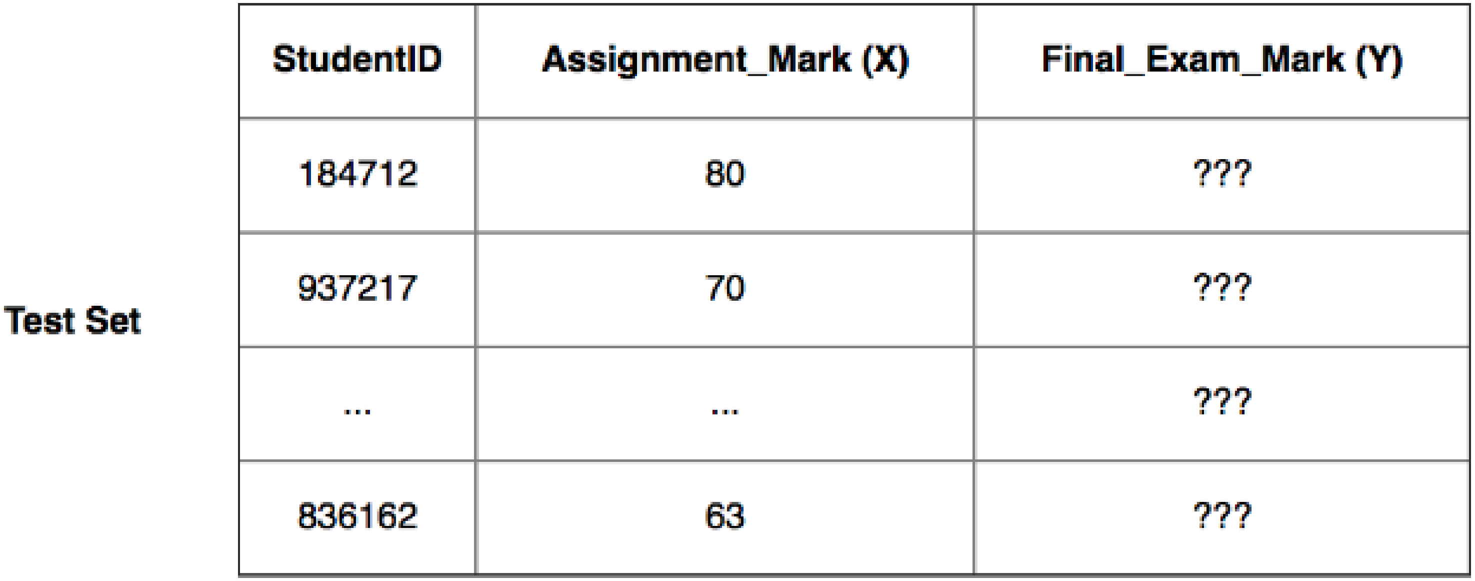
Training Set



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

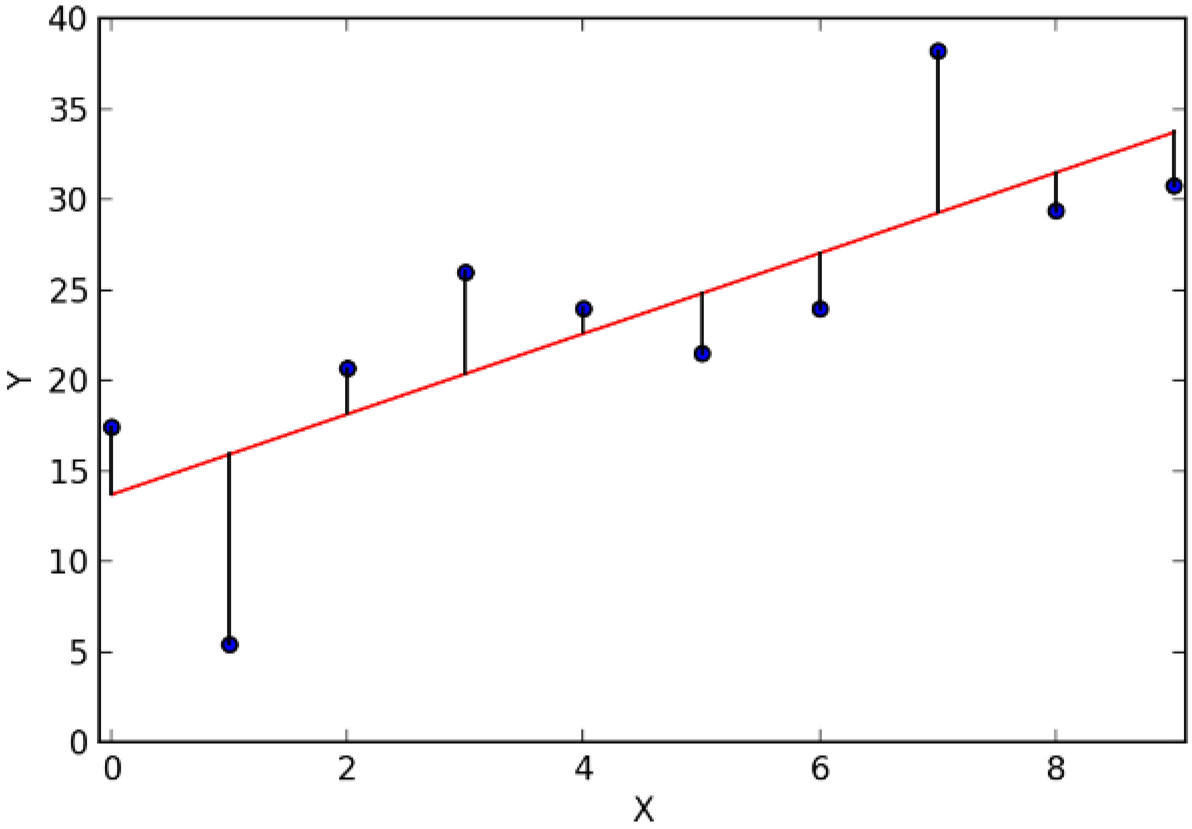


Test Set

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

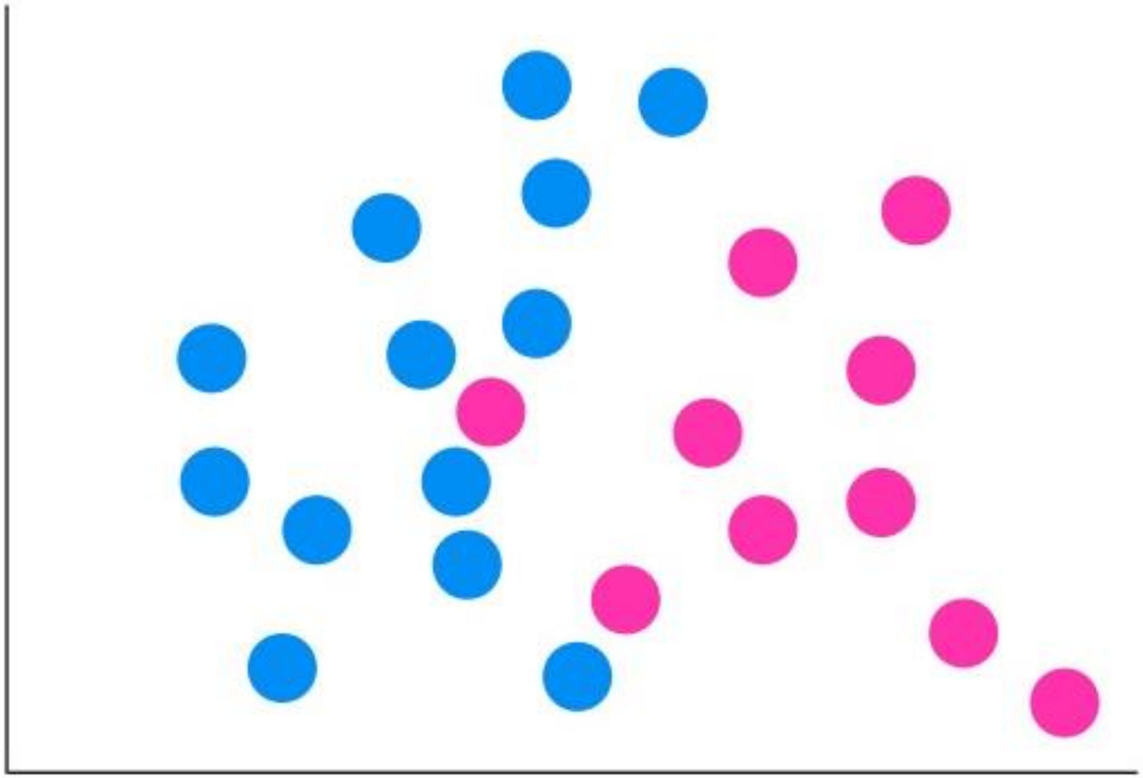
https://hackernoon.com/supervised-machine-learning-linear-regression-in-python-541a5d8141ce

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.

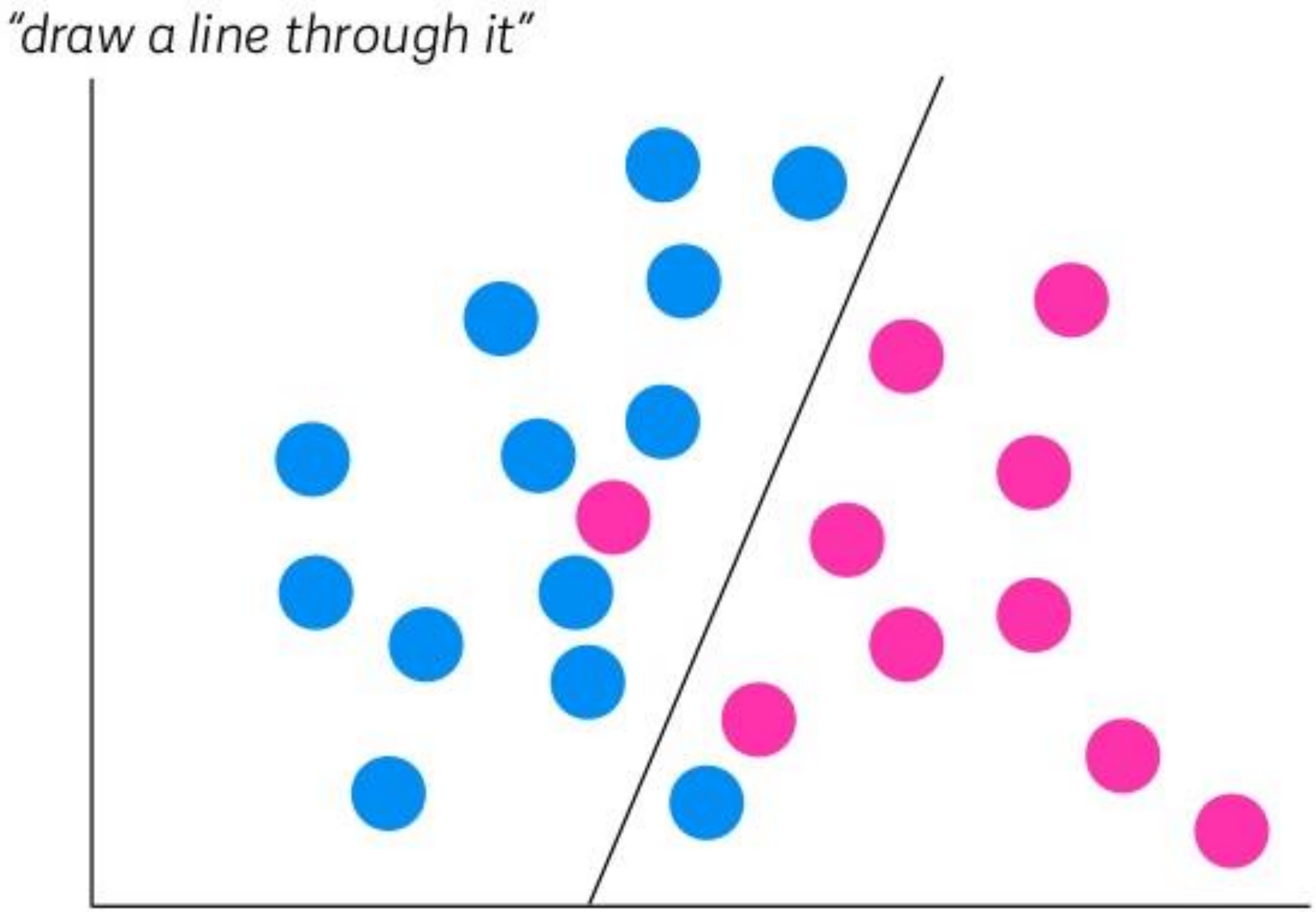
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

Classification



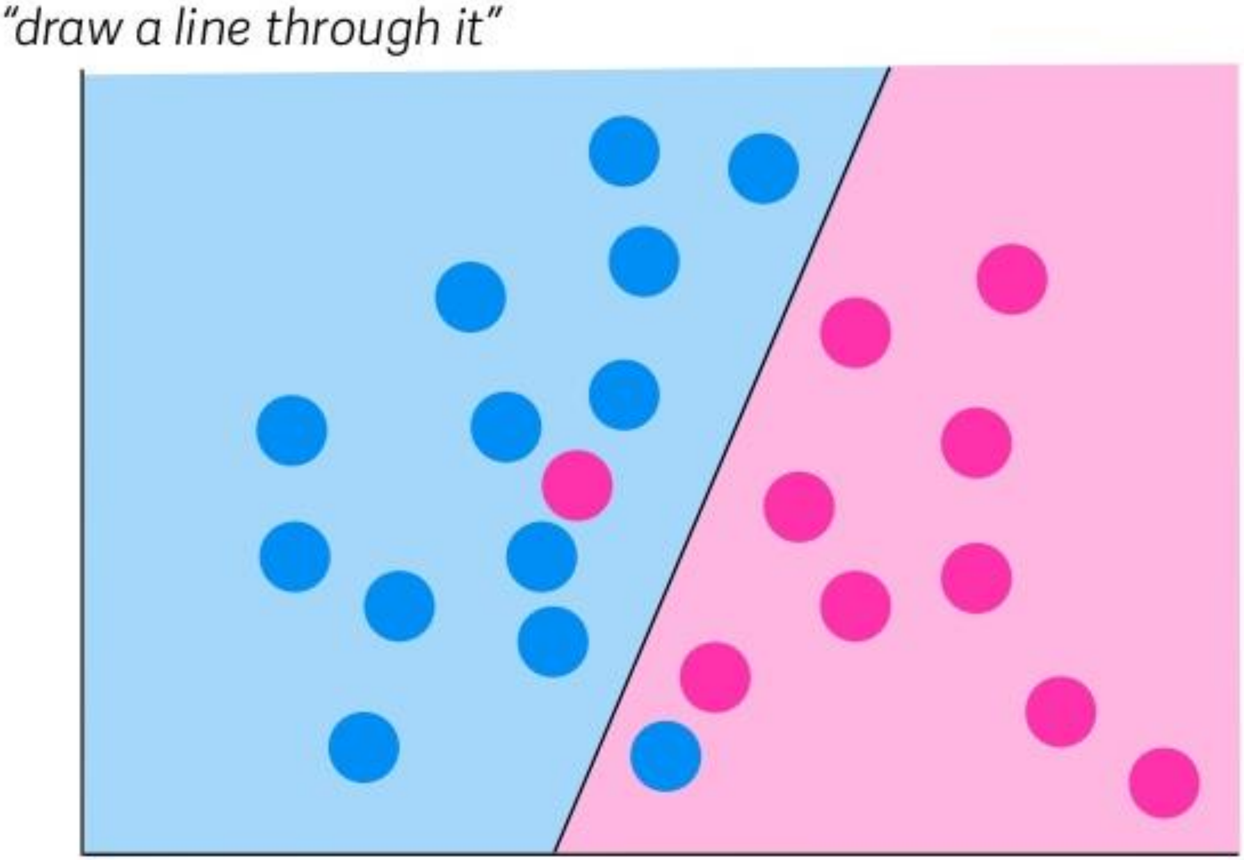
https://www.slideshare.net/brianjlange/machine-learning-in-5-minutes-classification

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

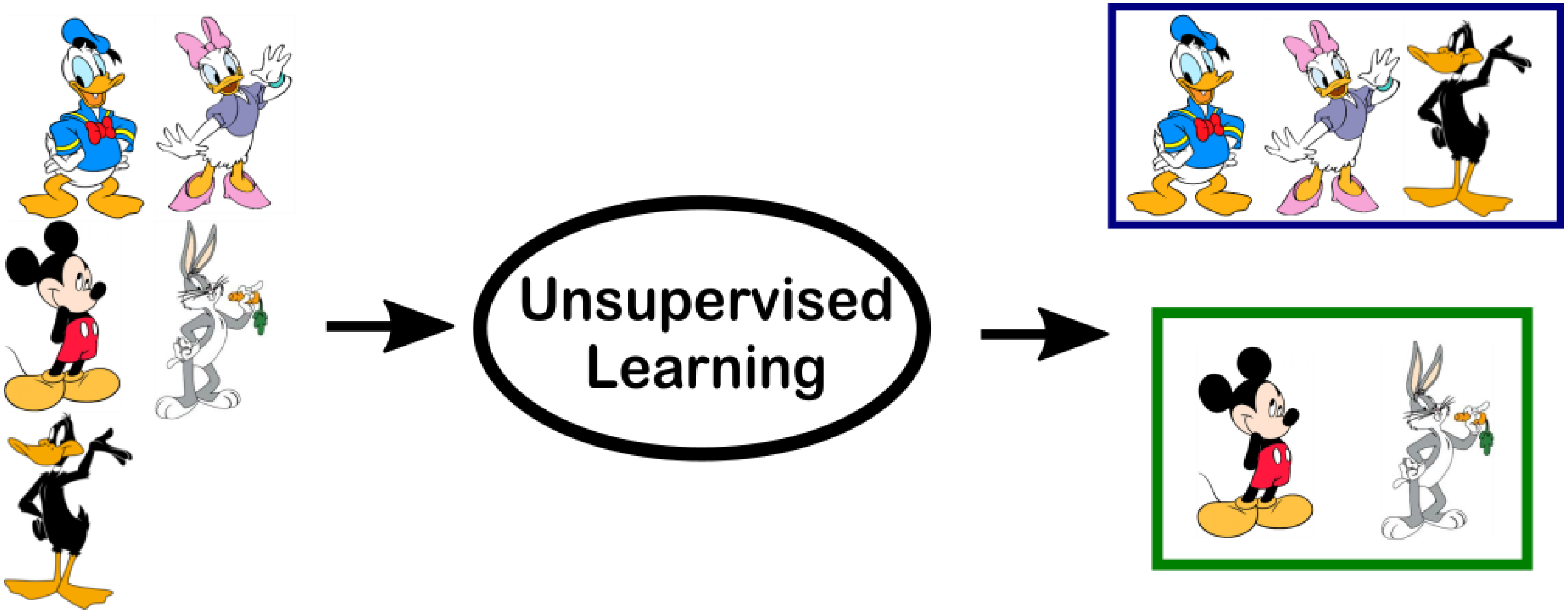
- Automated grouping of objects into so called clusters

- Objects of the same group are similar

- Different groups are dissimilar

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

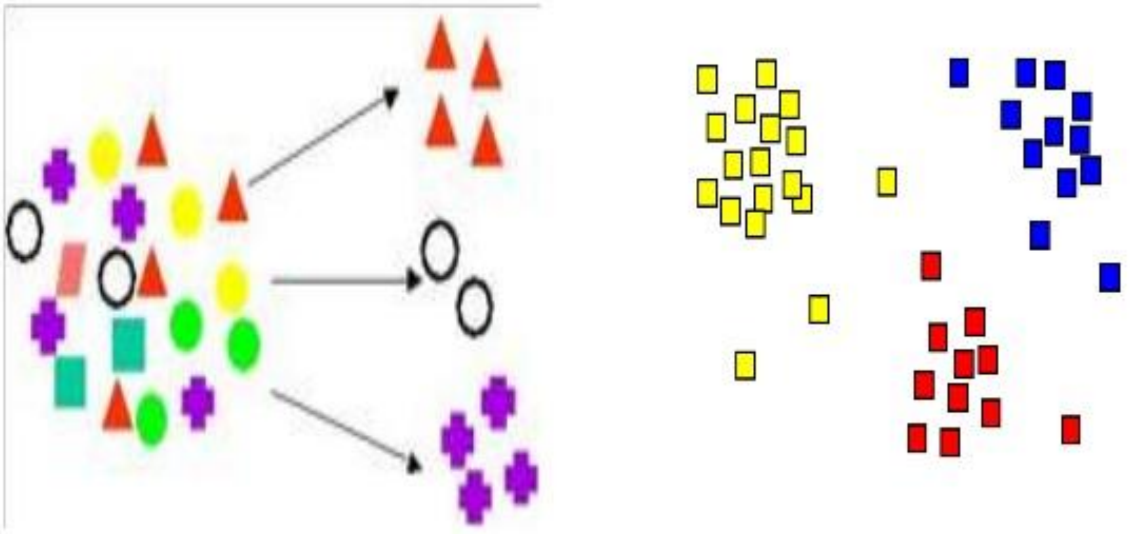
Clustering



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

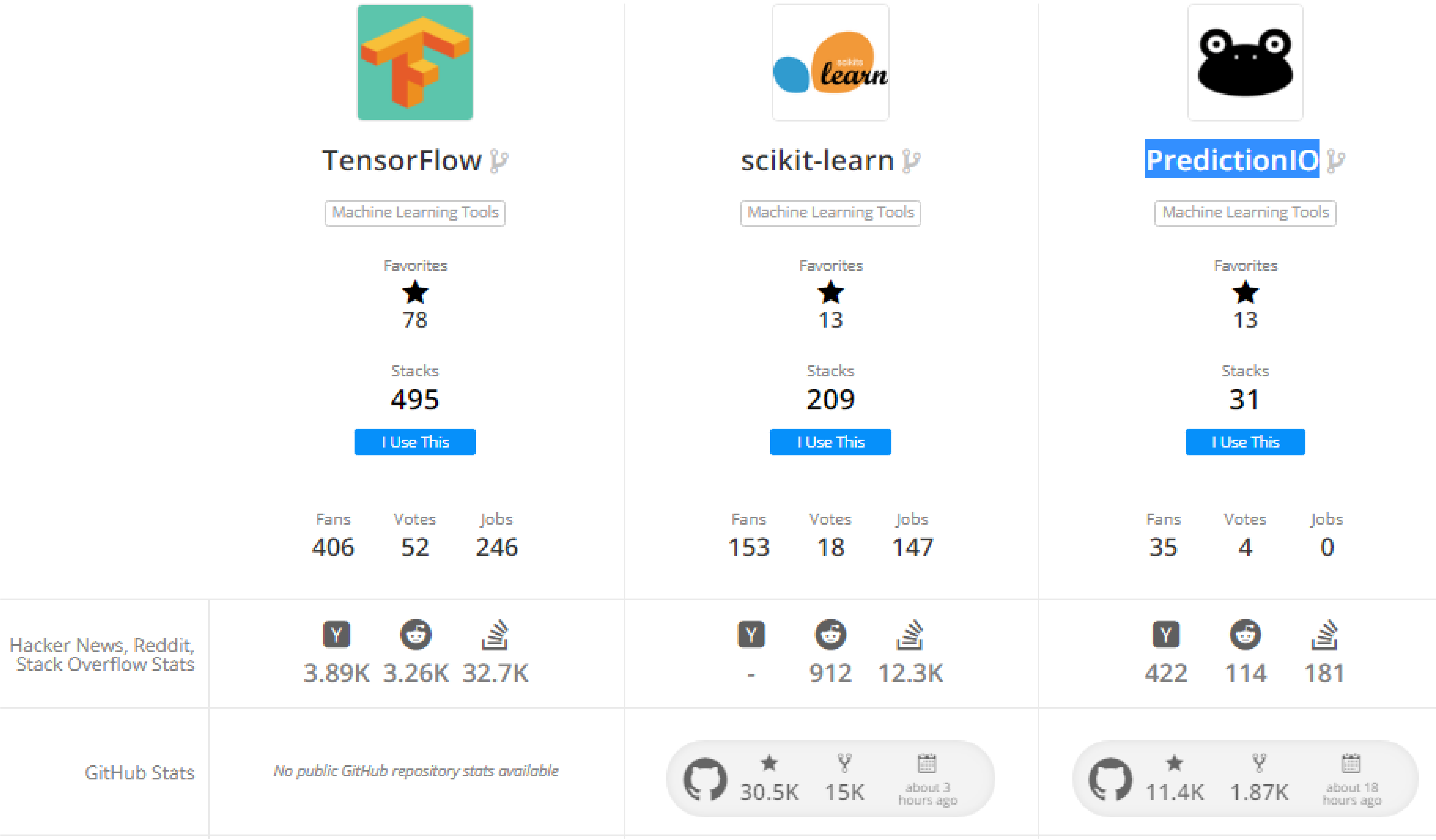
Clustering

Examples of Clustering



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

Useful Tools



	TensorFlow	scikit-learn	PredictionIO
	Machine Learning Tools	Machine Learning Tools	Machine Learning Tools
	Favorites ★ 78	Favorites ★ 13	Favorites ★ 13
	Stacks 495	Stacks 209	Stacks 31
	I Use This	I Use This	I Use This
	Fans 406 Votes 52 Jobs 246	Fans 153 Votes 18 Jobs 147	Fans 35 Votes 4 Jobs 0
Hacker News, Reddit, Stack Overflow Stats	Y 3.89K 3.26K 32.7K	Y - 912 12.3K	Y 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

Useful Tools

- TensorFlow
- scikit-learn
- PredictionIO

Further Reading and Useful Resources

- https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html
- https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4
- https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning
- http://gael-varoquaux.info/scikit-learn-tutorial/

Q&A