XCCELERATE

Data Science / Machine Learning - Fundamentals

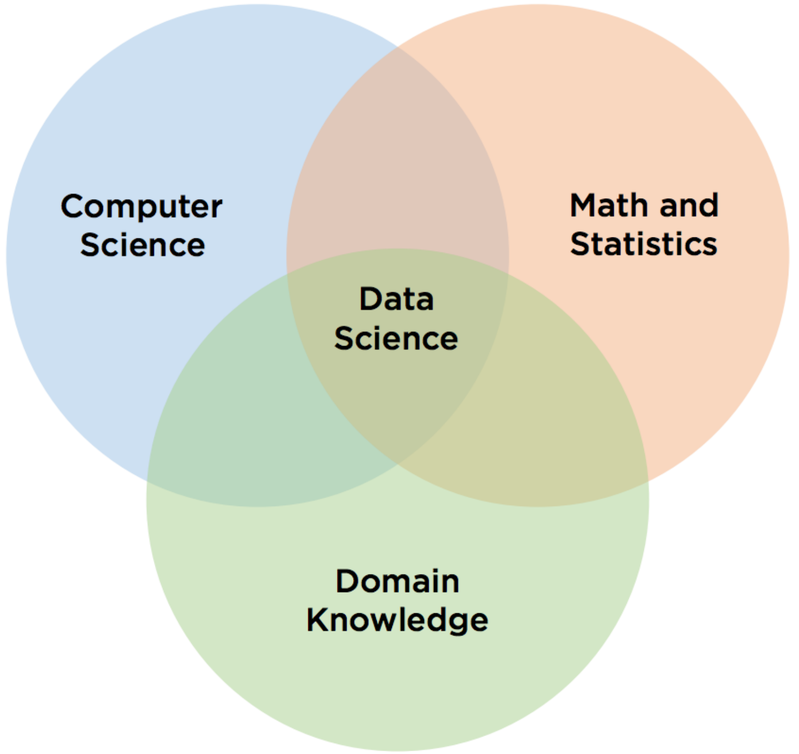
What is Data Science?

Joins statistics and programming in applied settings.

Analysis of diverse data

Deriving insights from data to make decisions or actionable steps.

These insights help business understand customers and competition.



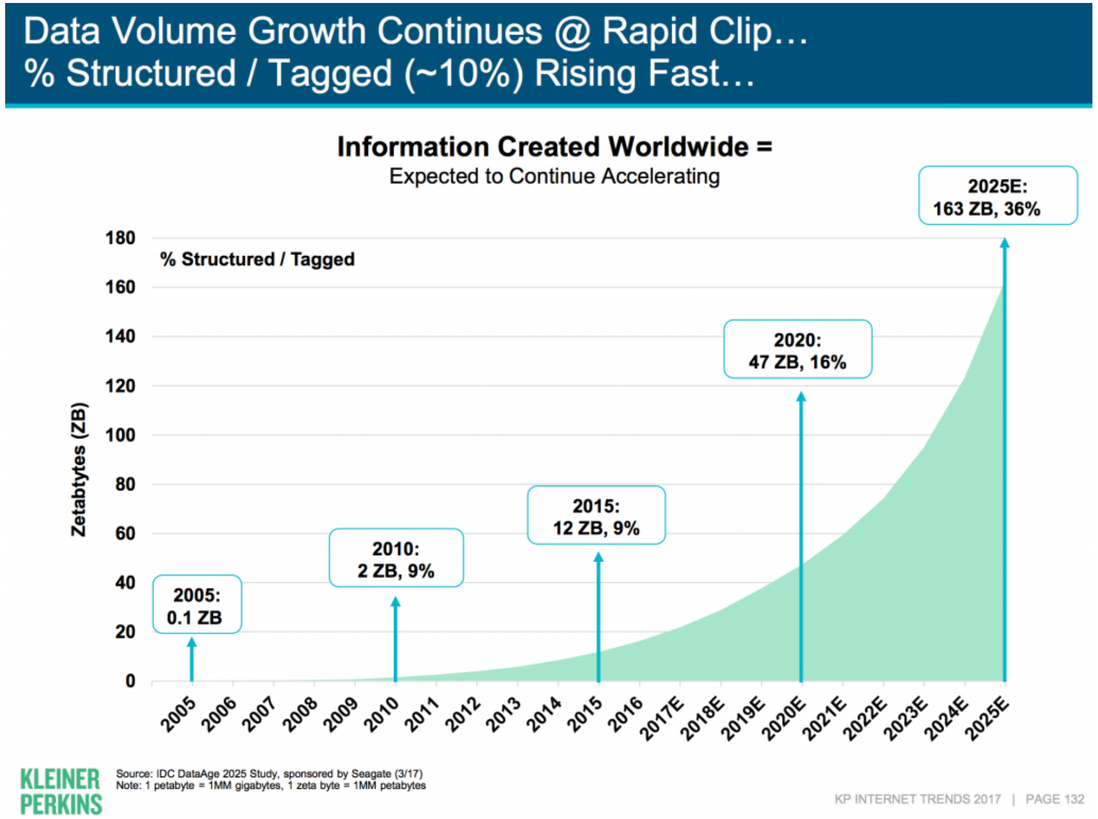
Future of Data Science?

Until 2005 - 130 Exabytes of Data

2010 – increase 10x – 1200 Exabytes of data

2020 – 40900 Exabytes of data

US: 140K – 190K people shortfall for data science and 1.5m managers who can use the data.



Summary of demand

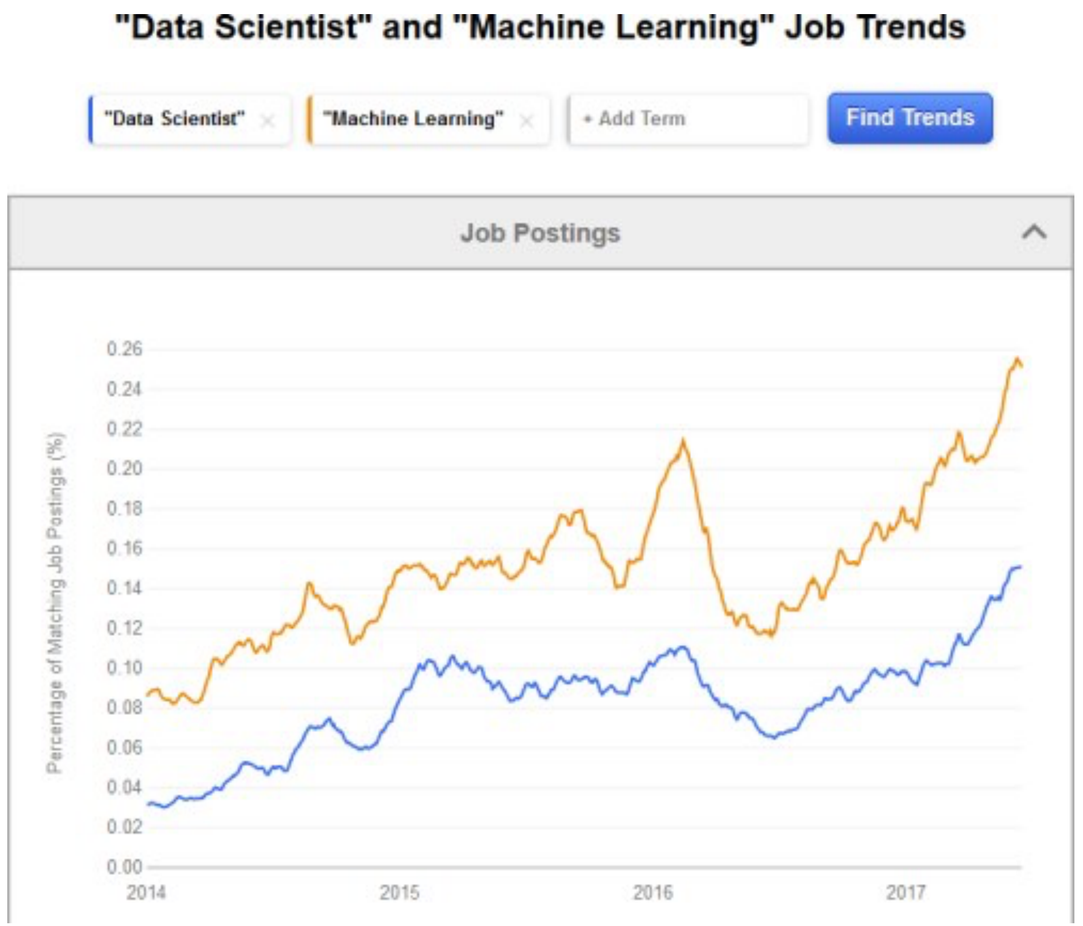
High demand for data science

Need for specialist and generalists.

Deriving insights from data to make decisions or actionable steps.

These insights help business understand customers and competition.

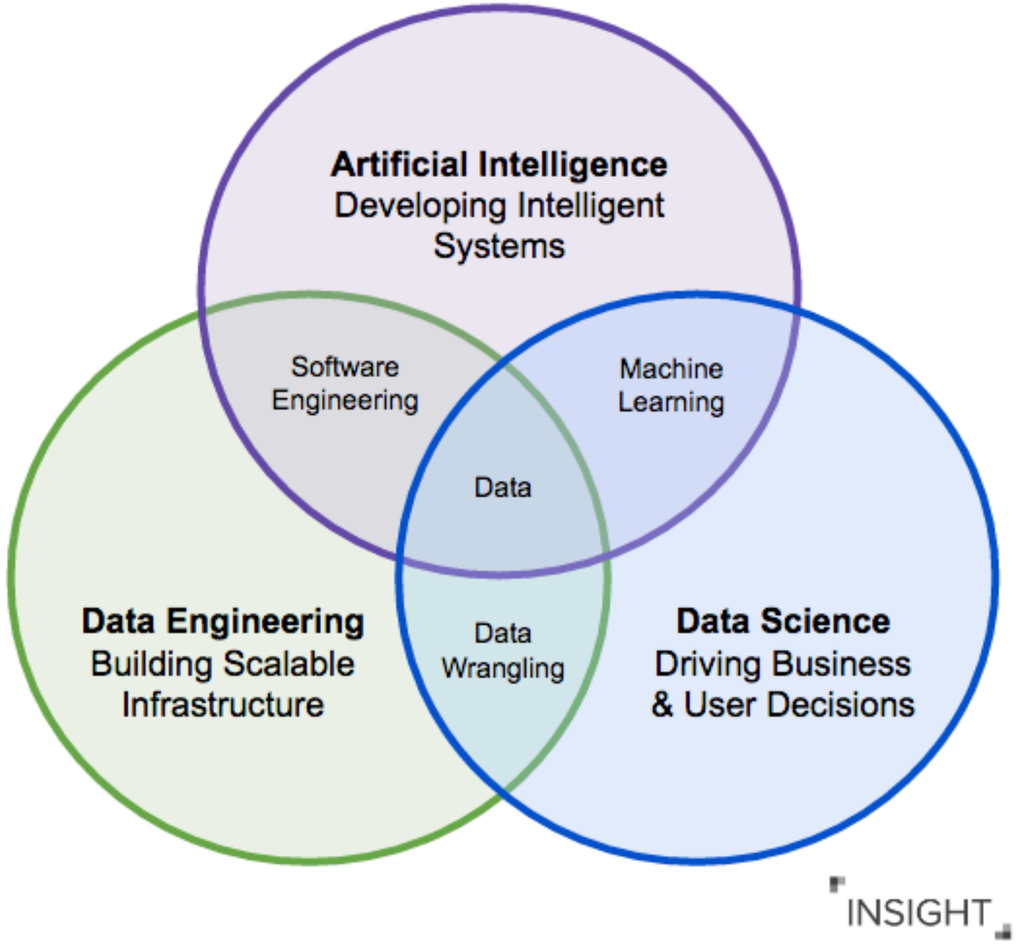
US: 140K – 190K people shortfall for data science and 1.5m managers who can use the data.



Source: KDNuggets

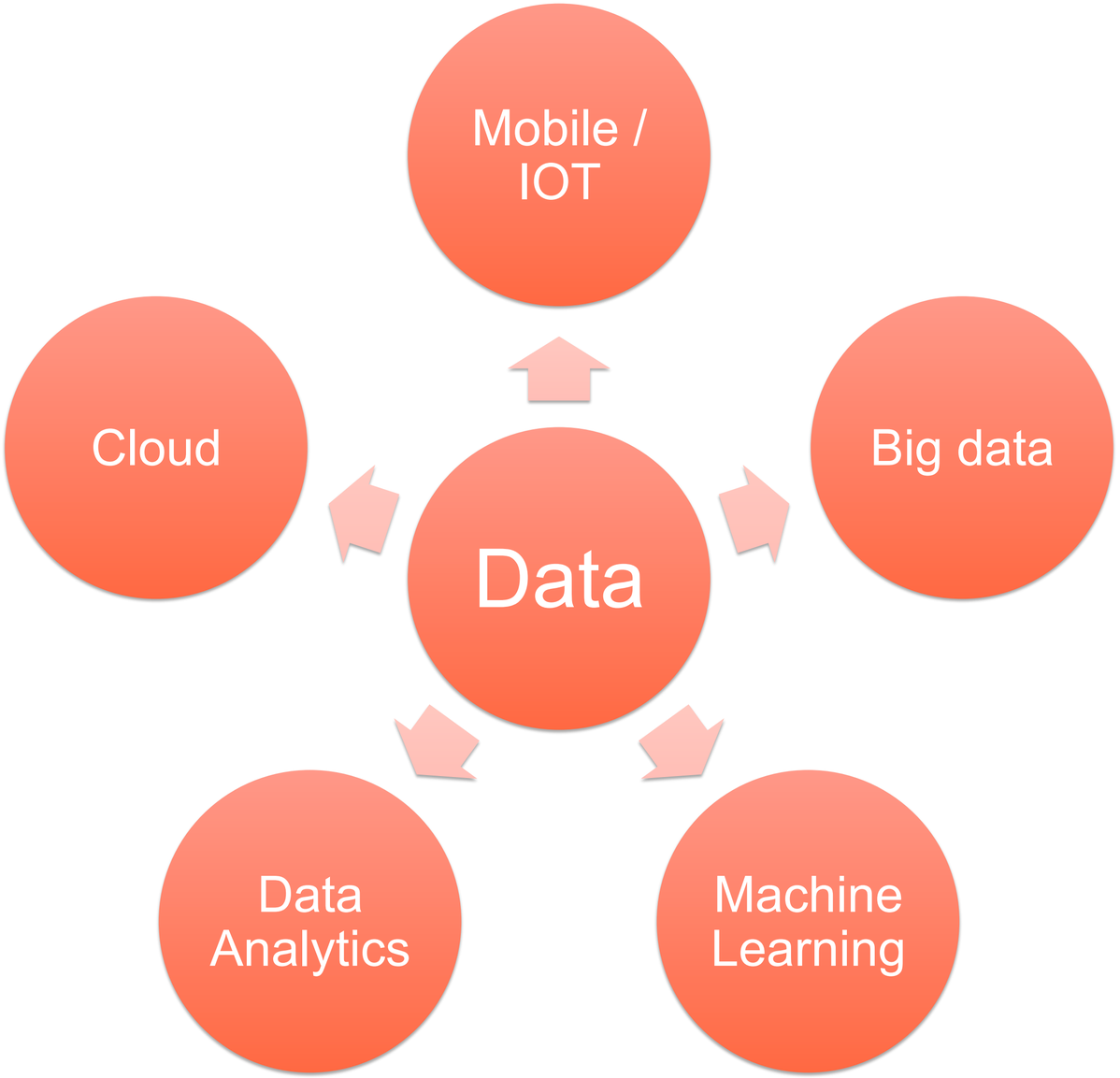
Data Science and AI

At the heart of every role is availability of data and be able to deal and manage data.

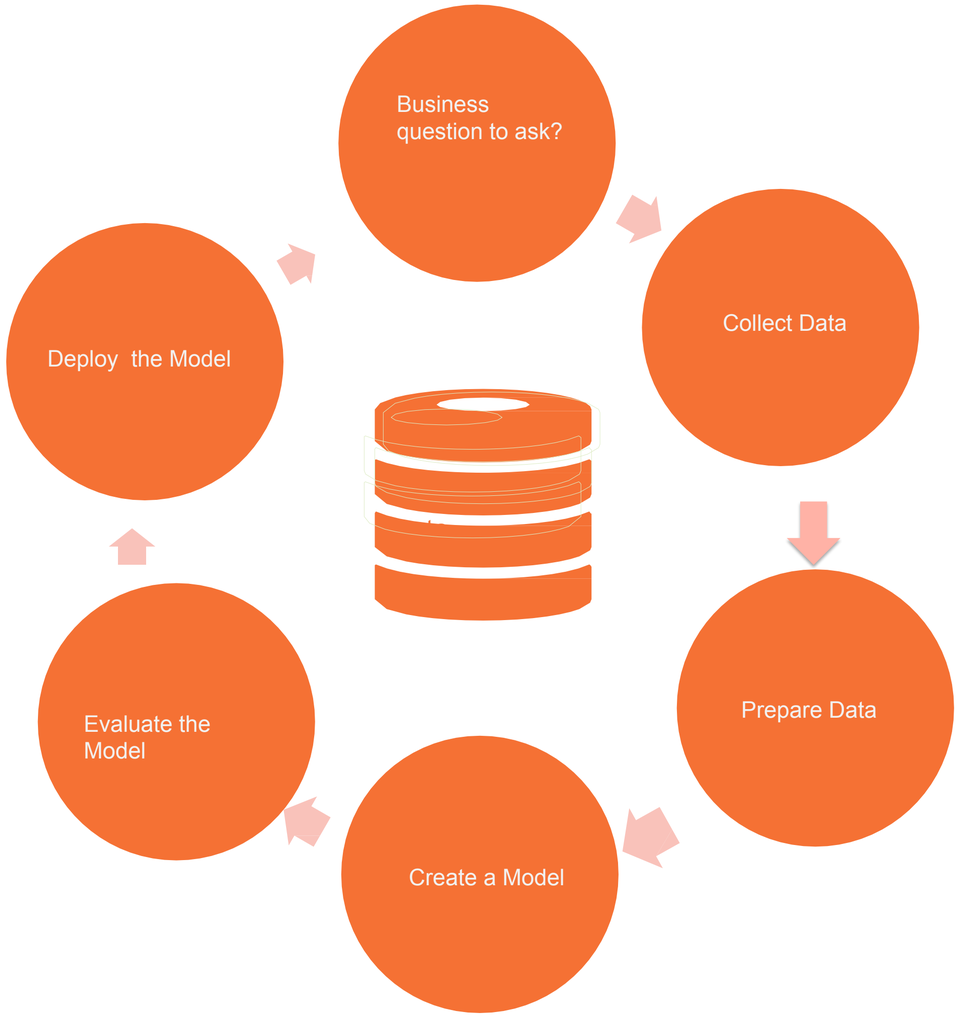


Why Data Science?

Importance



Data Science Process



Data Science Skills

Programming

Databases and process

Data Mining

Data Visualization

Statistics

Pattern recognition

Machine learning

Communication → Presentation → Domain knowledge → Creativity → Programming → Real-life practice

What do Data Scientists do?

Understand the business and ask the right questions
Identify key business variables
Define success metrics and business requirements?

Acquire data from relevant sources:
Data collection: ingest the data into analytic environment
Data exploration: explore the data to check quality and adequac

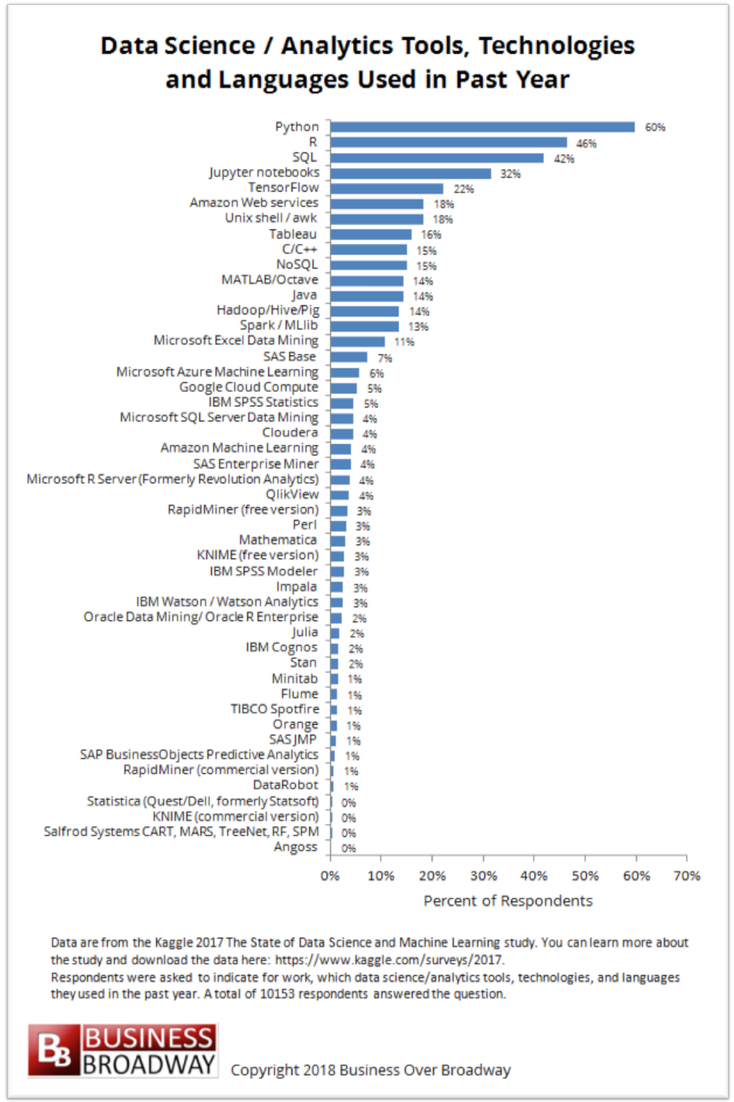
Modeling
Pre-processing: set up data pipeline to prepare data
Feature engineering: create features from raw data
Model evaluation: perform model fitting
Model selection: explore to find the optimal model

Operationalization:
Deploy model to production
Model consumption: use model to make predictions (scoring)
Business validation: To verify if business requirements are met.

Iterate

Data Science Tools

- **Python**
- **SQL**
- **Jupyter Notebooks**
- **Tensorflow**
- **Excel**
- **Tableau**
- **Azure**



Course Outline -

Fundamentals	Machine Learning Track	Business Intelligence Track	Data Engineering Track	Projects
Python Numpy and Pandas Visualization libraries API /JSON Web scraping Linear algebra, Matrices, Vectors Essentials / operations research Analysis Project	Fundamentals Statistics Machine Learning Feature Engineering, PCA Dimensionality reduction	BI Fundamentals Data cleaning in Excel Data analysis in Excel Visualizing in Excel. Project	Data Storage Database Management systems SQL Data Lakes Tableau Prep Tool	Collaborative Project Capstone Project
	Image recognition Recommender systems Text mining Projects	Introduction to Tableau Tableau for data cleaning and exploration Data mining with Tableau Dashboards/Visualization Project	Data Warehousing concepts Hadoop and Big data Google Big Query Azure Data Bricks Analytics	

Data sciences compared.

Compare

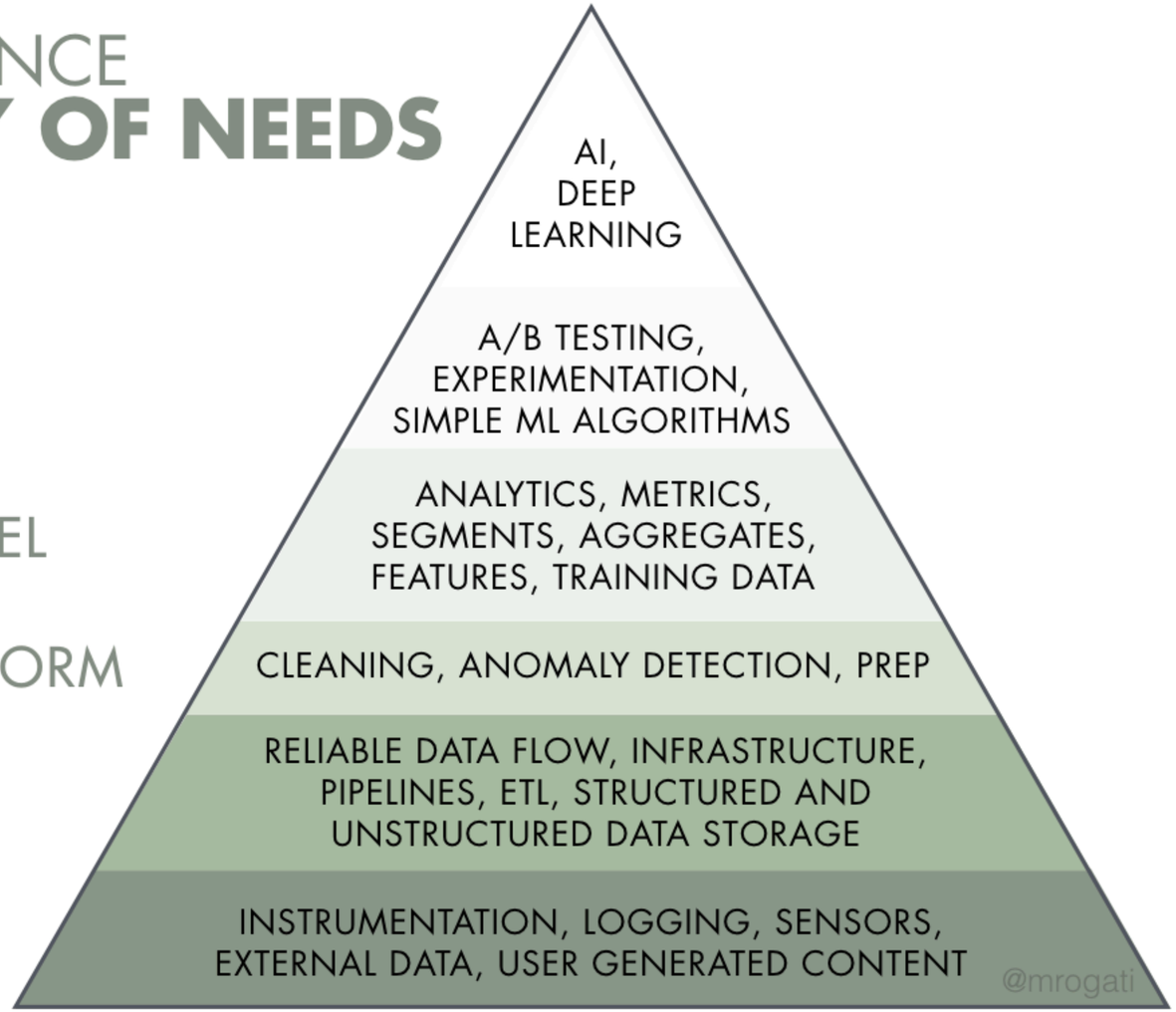
Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources	X			X
Build Data Pipelines	X			X
Process and Transform Data	X			X
Store Data	X			X
Dashboards/Reports		X	X	X
Exploratory Analytic		X	X	X
Statistical Modelin			X	X
Machine Learnin			X	X
Business Recommendations		X	X	X
Business Action			X	X

Challenges In data Science

- Dirty data (36% reported)
- Lack of data science talent (30%)
- Company politics (27%)
- Lack of clear question (22%)
- Data inaccessible (22%)
- Results not used by decision makers (18%)
- Explaining data science to others (16%)
- Privacy issues (14%)
- Lack of domain expertise (14%)
- Organization small and cannot afford data science team (13%)

Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: https://www.kaggle.com/surveys/2017. Respondents were asked, "At work, which barriers or challenges have you faced this past year? (Select all that apply)." A total of 10153 respondents were asked this questions.

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

- Source: HackerNoon

Data science on the Cloud

Data science in the Cloud

Amazon (AWS)

EC2

S3

Redshift (Data warehouse)

EMR(Elastic Map Reduce)

Microsoft

Azure Platform

Google

Google Cloud Platform

Google BigQuery

Machine Learning

What is Machine Learning?

Arthur samuel (1959) Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed.[1]

Tom M. Mitchell (1998)"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."[16]

Machine Learning

What is Machine Learning?

Algorithms that do the learning without human intervention.

Learning is done based on examples (aka dataset).

Goal:
learning function f: x [?] y to make correct prediction for new input data
Choose function family (logistic regression, support vector machines)
Optimize parameters on training data: Minimize Loss ∑(f(x) - y)2

How is it used today?

Company	How its using Machine Learning today?	
Yelp	**Yelp's machine learning algorithms help the company's human staff to compile, categorize, and label images more efficiently**	
Pinterest	**machine learning touches virtually every aspect of Pinterest's business operations, from spam moderation and content discovery to advertising monetization and reducing churn of email newsletter subscribers**	
Facebook Chatbots	**AI applications are being used at Facebook to filter out spam and poor-quality content** **That's because Messenger has become something of an experimental testing laboratory for chatbots**. computer vision algorithms that can "read" images to visually impaired people.	
Twitter	**Twitter's AI evaluates each tweet in real time and "scores" them according to various metrics** **Twitter's machine learning tech makes those decisions based on your individual preferences**	
Google Deepmind	perhaps most exciting for tech nerds – neural networks.**the DeepMind network, the "machine that dreams**	
EdgeCase	Edgecase hopes its machine learning technology can help ecommerce retailers improve the experience for users Edgecase plans to **leverage its tech to provide a better experience for shoppers who may only have a vague idea of what they're looking for**	
Baidu	Voice search: Deep Voice, **a deep neural network that can generate entirely synthetic human voices that are very difficult to distinguish from genuine human speech**	
Salesforce	**Salesforce Einstein allows businesses that use Salesforce's CRM software to analyze every aspect of a customer's relationship**	

Machine learning types

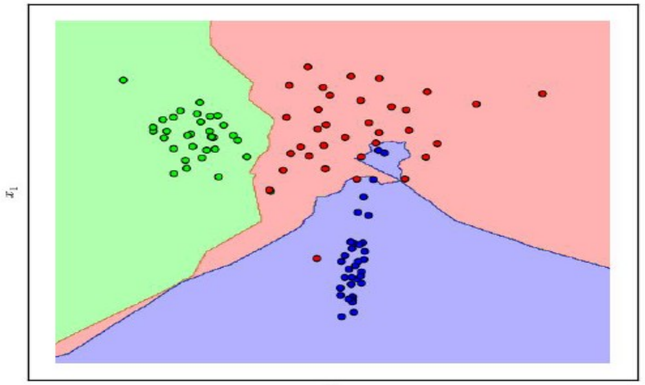
- Supervised Learning:
  - Look at some examples (labeled data) and find a way to predict future (unlabeled) examples
  - the target variable ("labels") contains the ground truth we want to predict
  - by comparing predictions with the ground truth, we know how well we're doing

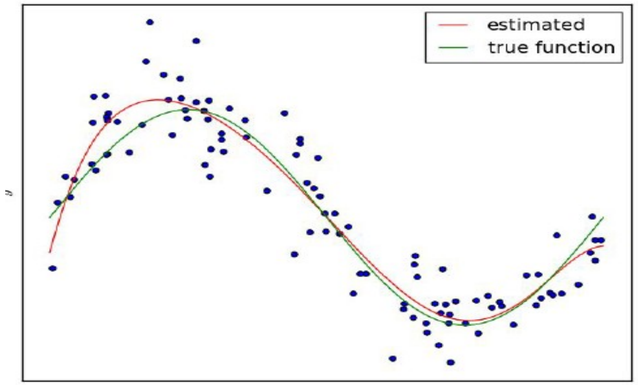
Summary: We are trying to predict a variable (called labels, target variable, response variable or dependent variable) using other variables (called features, explanatory variables, covariates, attributes or independent variables).

- regression algorithms predict a number (numeric target)
- classification algorithms predict a category (categorical target)
- Sometimes regression refers to a family of ML algorithms. For example, linear regression is a regression algorithm but logistic regression is a classification algorithm!

Classification and Regression

- **Classification**
  - Given labeled input data (with two or more labels), fit a function that can determine for any input, what the label is.
- **Regression**
  - Given continuous input data fit a function that is able to predict the continuous value of input given other data.





Supervised Learning

“Teach the model”, then with that knowledge, have it predict future instances.

- Use the **training data** to fit a model which is then to predict incoming input

Machine learning types

- Supervised Learning:

  Regression

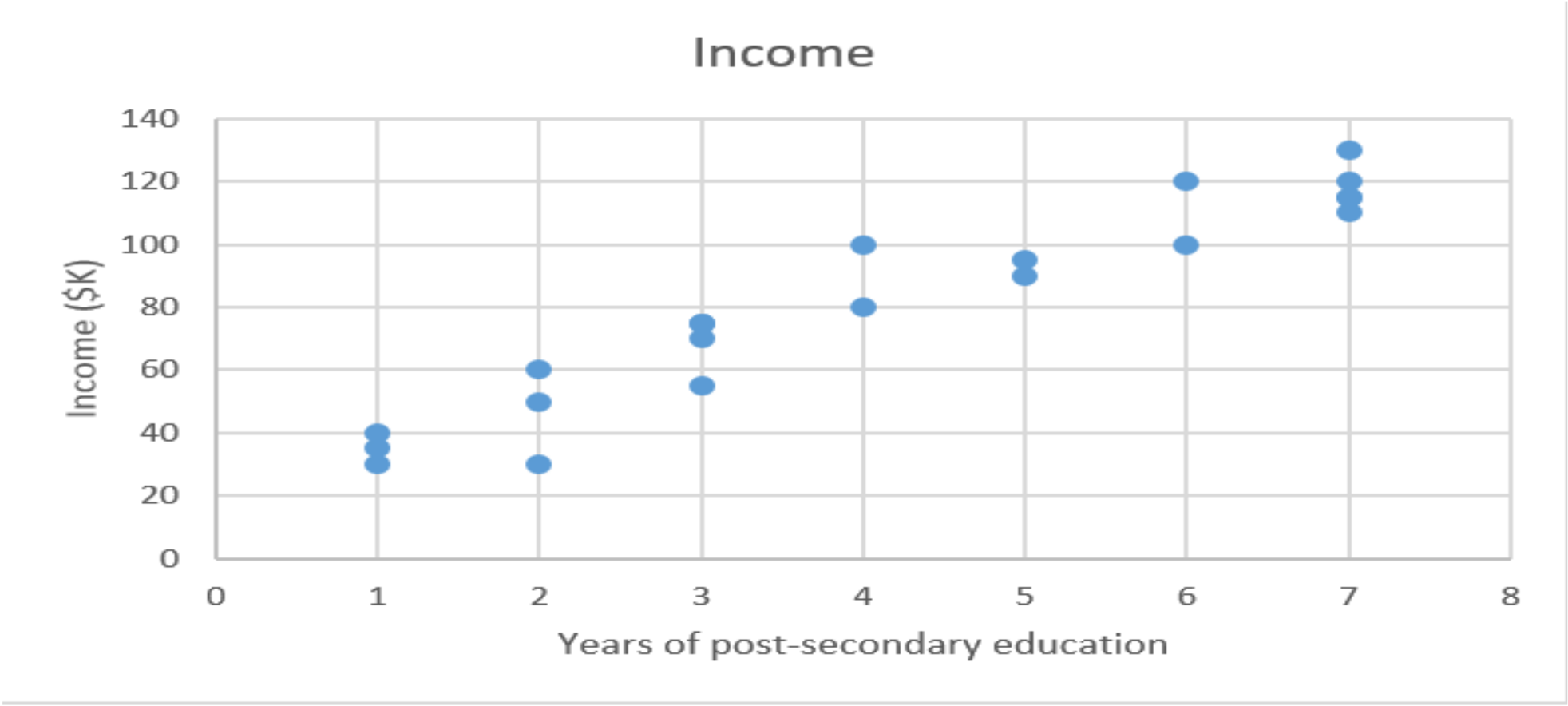
  - predict a continuous numerical value. *How much will that house sell for?*
  - *Predict income levels based on years of experience*

Machine learning types

- Supervised Learning
  - Regression
    - predict a continuous numerical value. *How much will that house sell for?*
    - *Predict income levels based on years of experience*

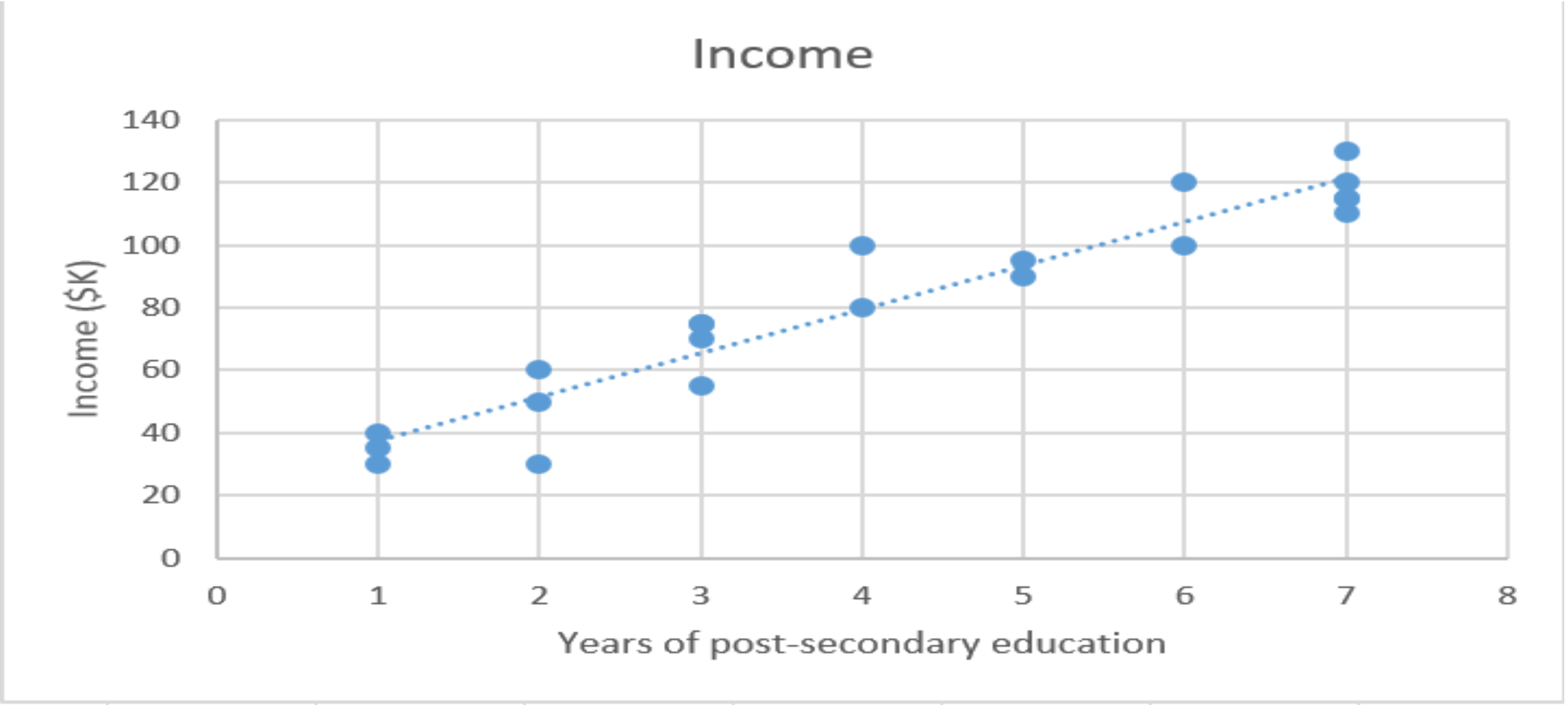
Machine learning types

- Supervised Learning – Linear regression



Machine learning types

- Supervised Learning – Linear regression

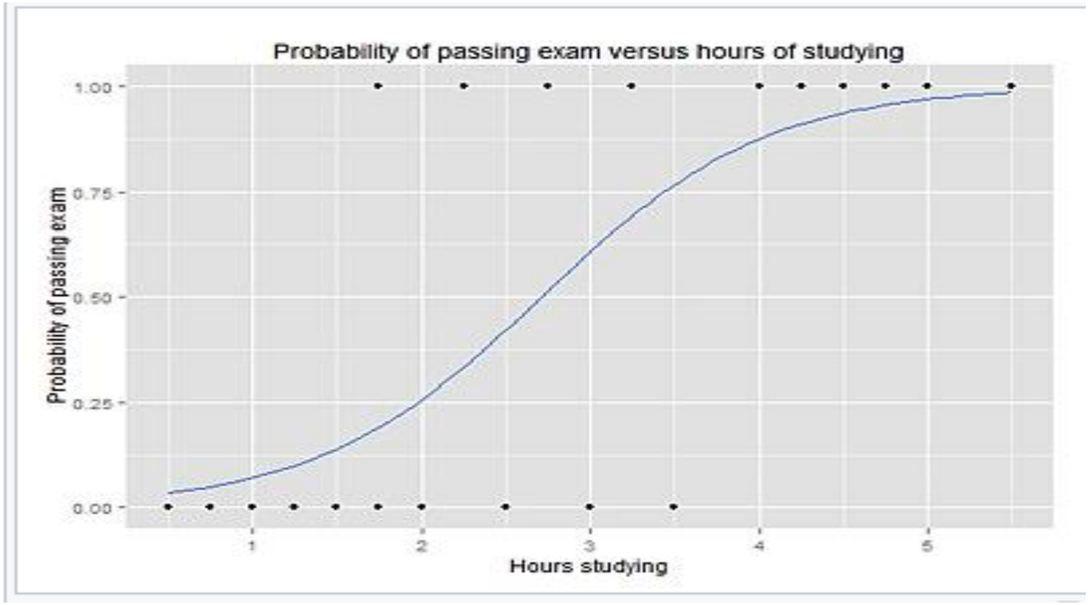


Machine learning types

- Supervised Learning
  - Classification
    - Is this email SPAM or NOT
    - *Will users click on that Ad?*
    - *Who is the person in the picture?*

Machine learning types

- Supervised Learning
  - Logistics regression
    - Is this email SPAM or NOT
    - *Will users click on that Ad?*

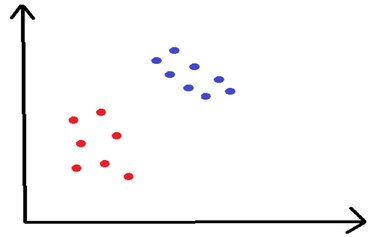


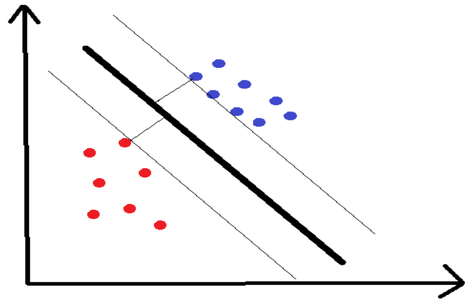
Machine learning types

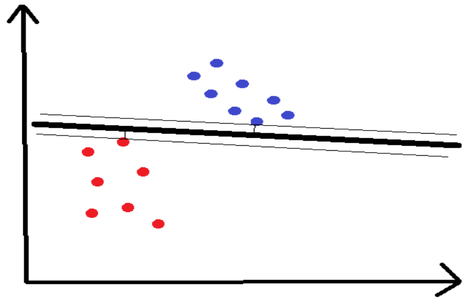
- Supervised Learning
  - Support vector machines(SVM)
    - Is this an image of a cat or a dog?
    - Is this review positive or negative?
    - Are the dots in the 2D plane red or blue?

Machine learning types

- Supervised Learning
  - Support vector machines(SVM)

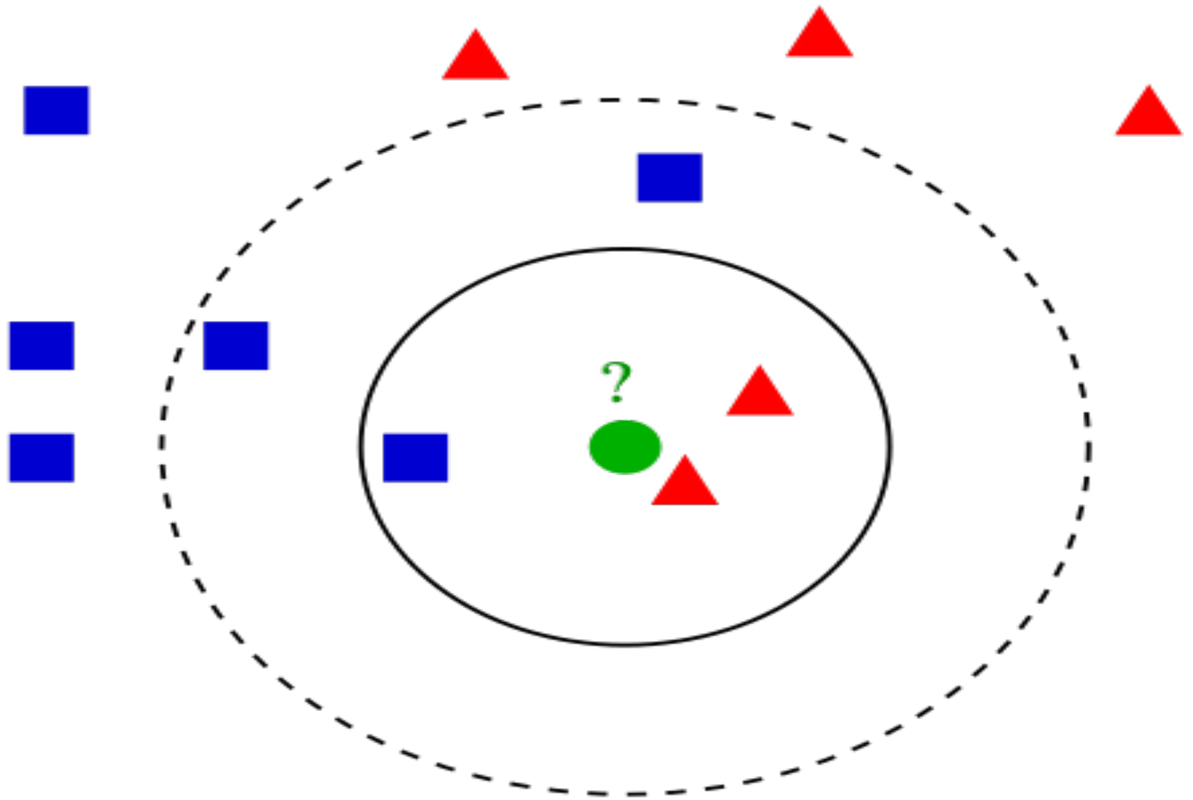






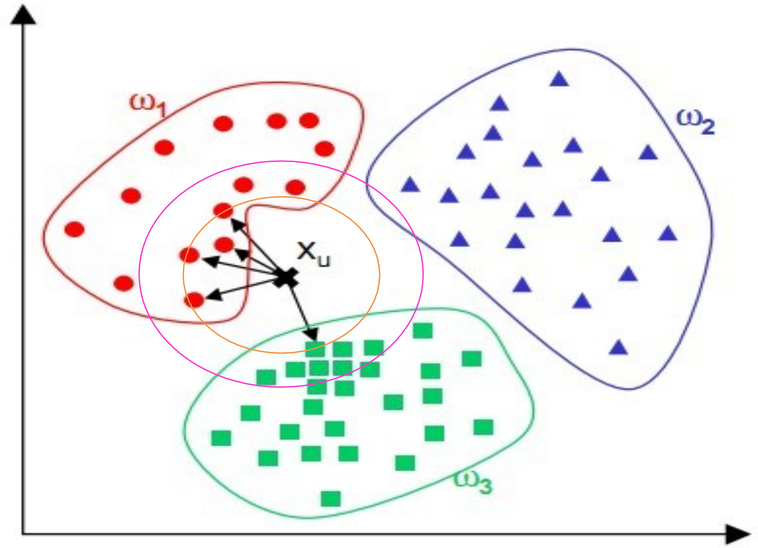
Machine learning types

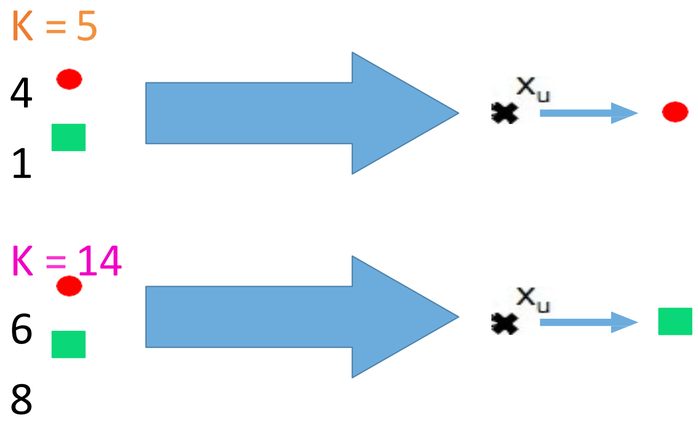
- Supervised Learning
  - *K-nearest neighbours(k-NN)*



K-Nearest Neighbor

In KNN, data points are categorized and when determining the category of a new data point, the K nearest points are used in this process.



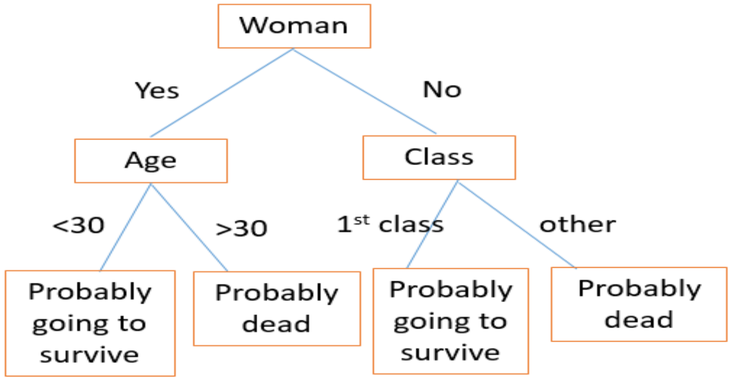


Machine learning types

- Supervised Learning
  - *Decision Trees, Random Forests*

A normal tree

A decision tree



K-Nearest Neighbors

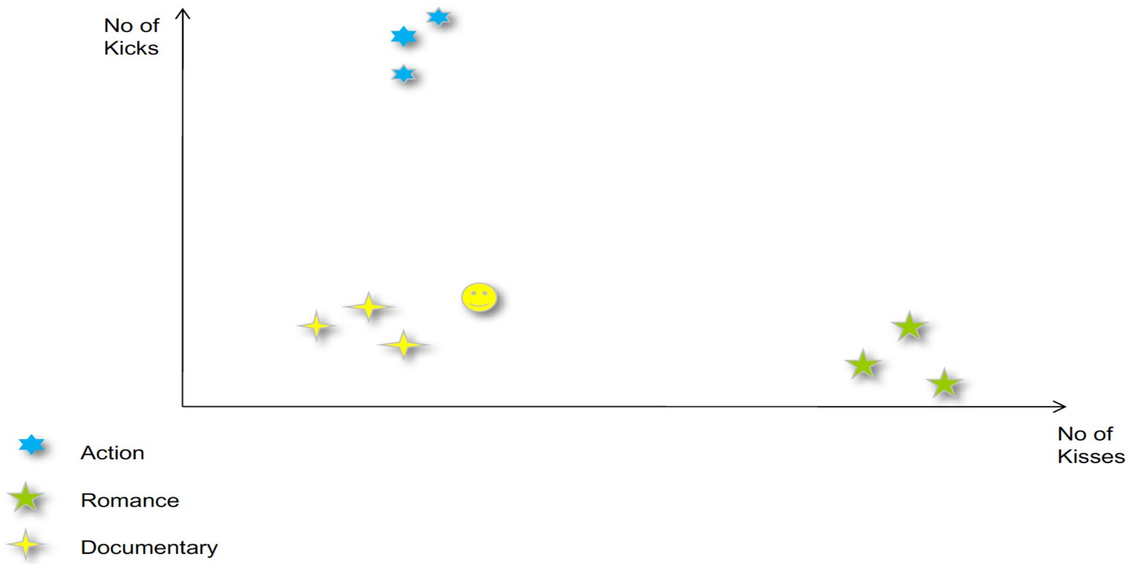
The distance of the points can be calculated with the following fomula:

- For points in two dimensional space:

$$d = \sqrt{(x_0 - y_0)^2 + (x_1 - y_1)^2}$$

- For points in n-dimensional space:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



K-Nearest Neighbors

- To classify with distance measurements
  - Normalize the numeric values so that different features may have comparable numeric values.
  - Calculate the distance between the Test Sample and all Trained Samples, choose the k-nearest Trained Samples, and retrieve their classes
- Determine the class of the test sample by choosing the majority class of the k-nearest Trained Samples.
- Intuition: finding k samples which are most similar to the sample to be classified, and use these samples' most common classification.

Machine learning types

- UnSupervised Learning
  - look at unlabeled data and find general patterns
  - more subjective and difficult to interpret

Machine learning types

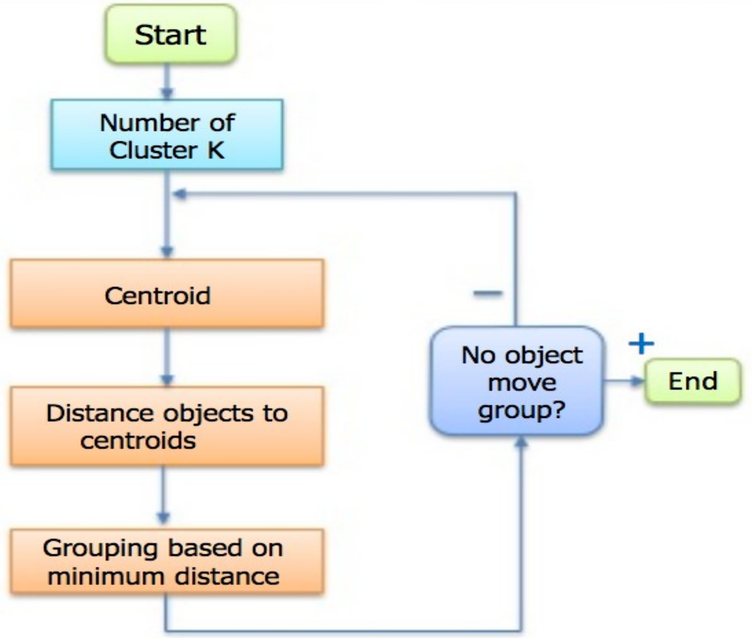
- UnSupervised Learning
  - *Clustering*
    - *K-means clustering*

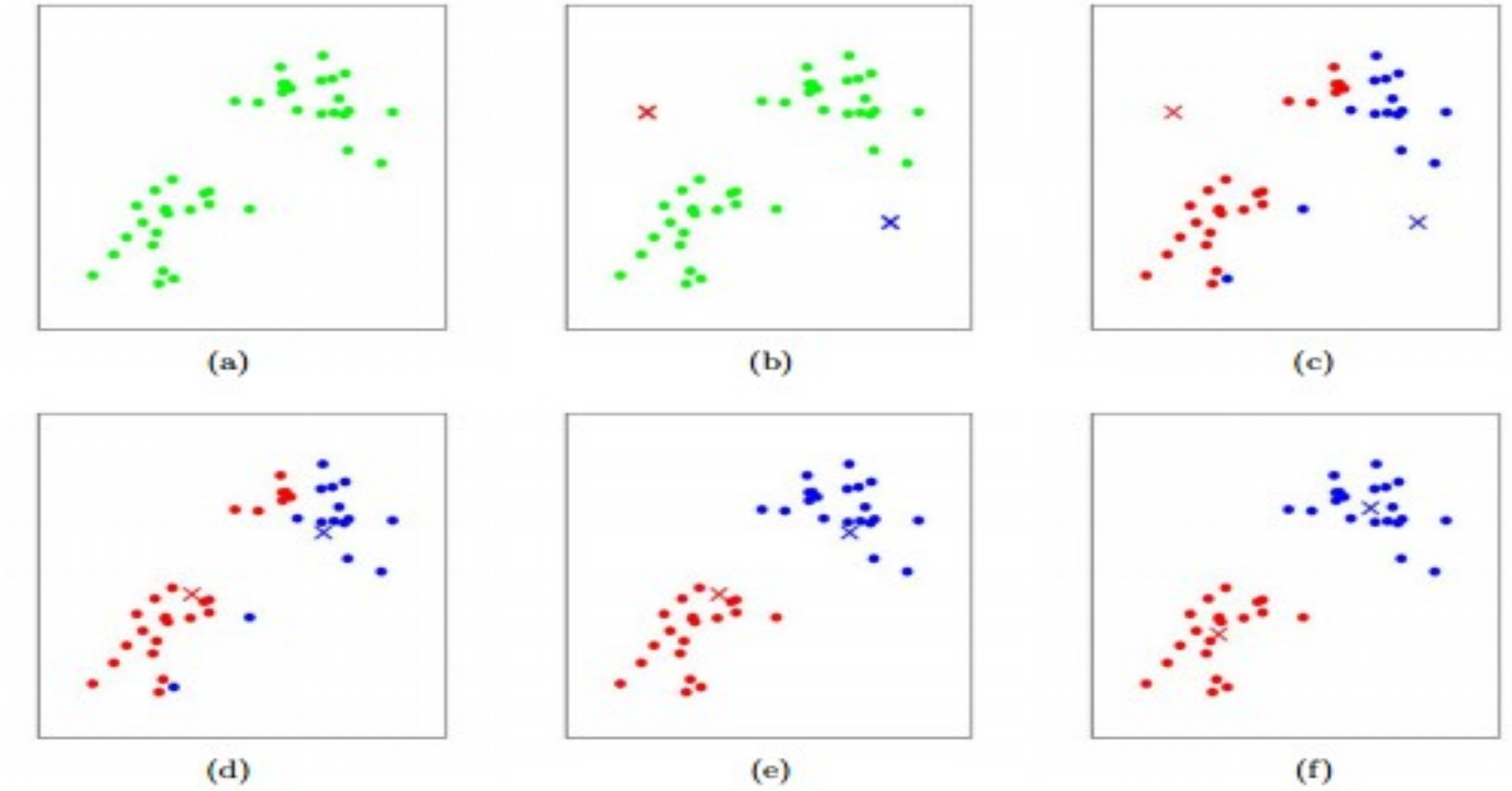
Clustering

- Clustering is a type of **unsupervised learning** that automatically form clusters (groups) of similar things. It differs from Classification in that there is NOT pre-defined classes.
- In Clustering, we try to put similar things in a cluster. But the similarity depends on how we measure it.
- The type of similarity measurement varies by the algorithms
  - K-Means
  - Two-steps
  - Self-Organizing Maps (Kohonen Map)

K-means

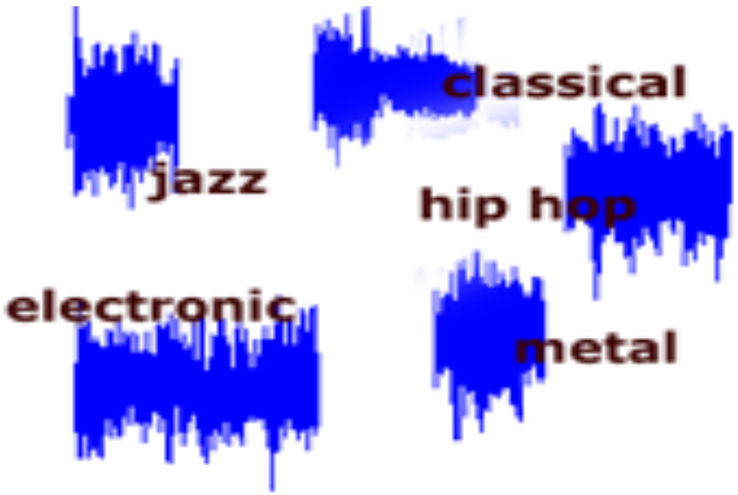
- K-means is a type of clustering algorithm that find k clusters for a given dataset. k is a user define parameter.
- Each cluster is described by a point named centroid, which is the center of the cluster
- The algorithm works like this:
    1. Create k centroid (often randomly)
    2. Each point in the dataset are assign to a cluster whose centroid is closest the point.
    3. Calculate the new centroid for each cluster by take the mean value of the points in the cluster.
    4. Iterate step 2 & 3 until none of the points change its cluster (or no centroid is changed).



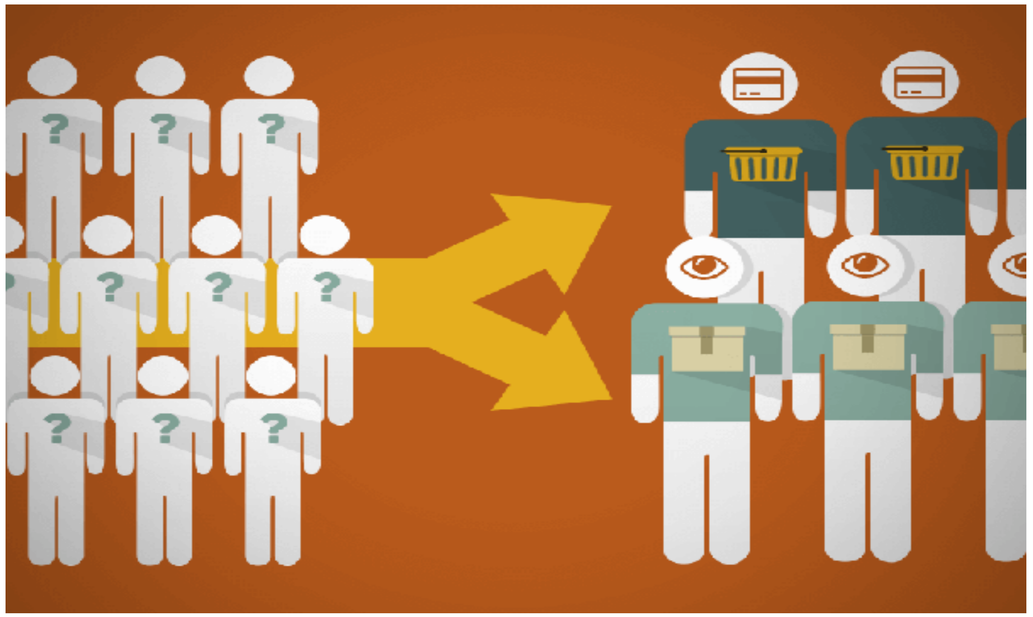


Source: http://stanford.edu/~cpiech/cs221/img/kmeansViz.png

Unsupervised learning examples



Music Categorization



Customer Segmentation

Other use cases for unsupervised Learning

- Pattern recognition
- Text clustering
- Object recognition
- Feature extraction
- Data dimensionality reduction

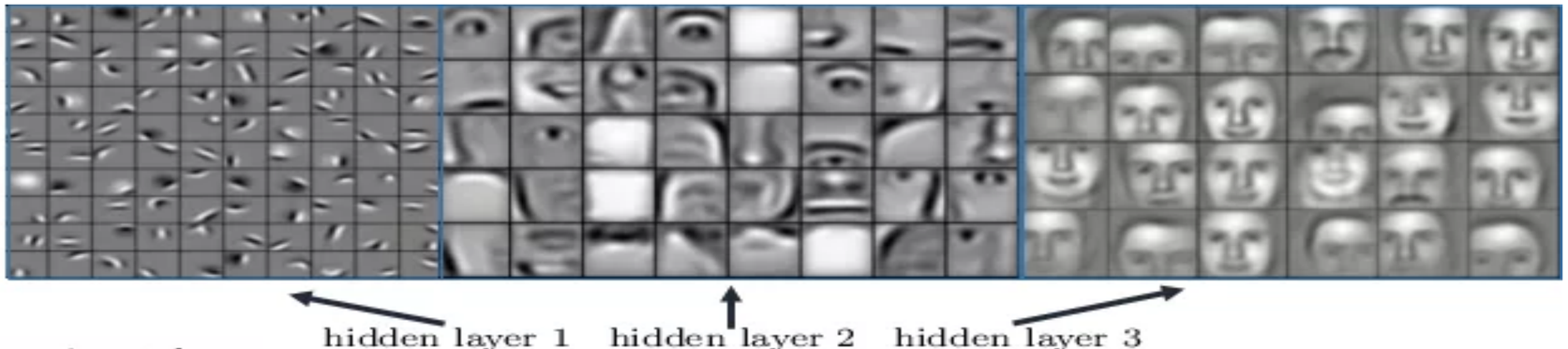
Exploratory data analyses

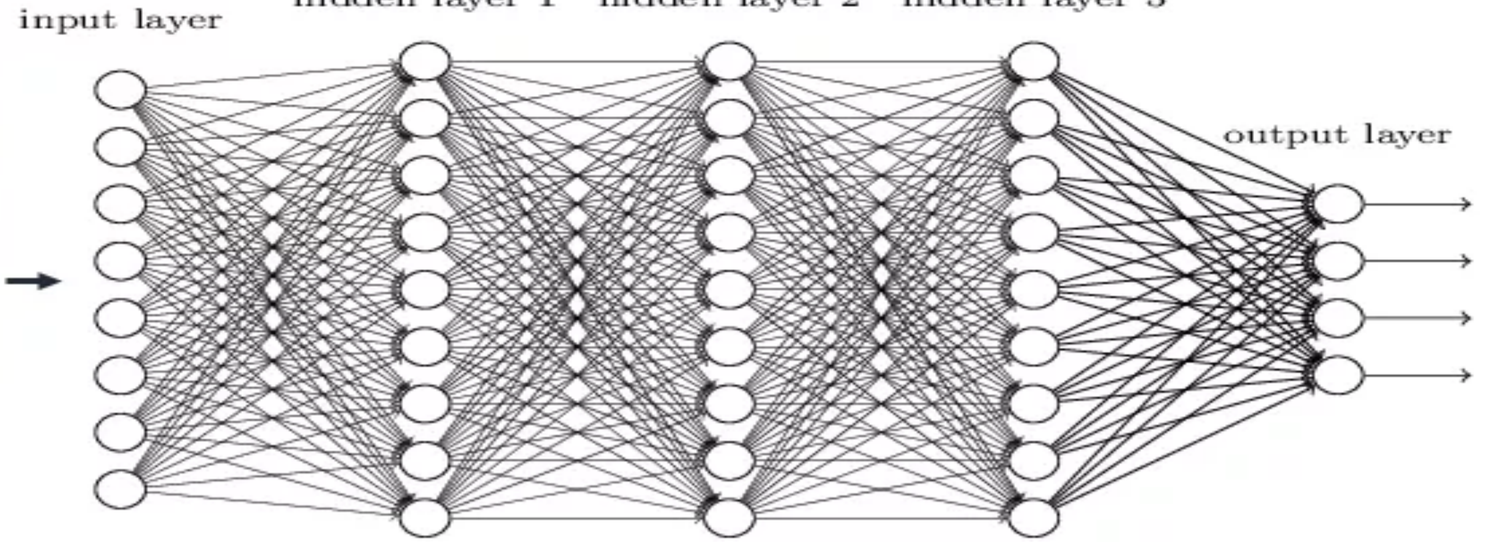
- Slice and dice data to get a feel for it
- Visualise the data looking at patterns
- Bivariate statistics and plots can hint at relationships between variables
- use statistical summaries to see if the data makes sense or if outliers are present
- examine missing values

- Visual EDA –
- scatter_matrix
- Head
- Info
- Describe
- Names , data types, Columns etc

Machine learning types

Deep neural networks learn hierarchical feature representations





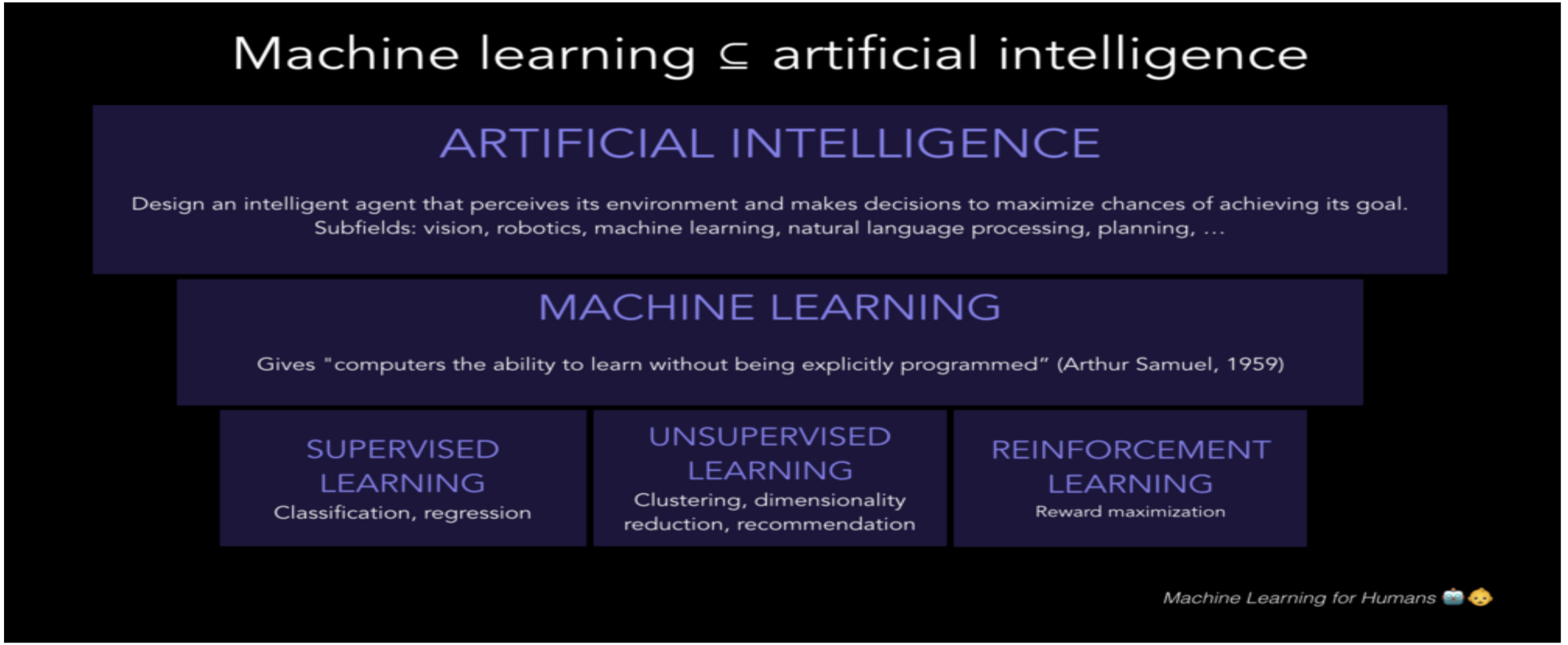
Deep Learning Uses

- Self-driving cars rely on deep learning for visual tasks like understanding road signs, detecting lanes, and recognizing obstacles.
- Deep learning can be used for fun stuff like art generation. A tool called neural style can impressively mimic an artist's style and use it to remix another image.
- Predicting molecule bioactivity for drug discovery
- Face and object recognition for photo and video tagging
- Powering Google search results
- Natural language understanding and generation, e.g. Google Translate
- The Mars explorer robot Curiosity is autonomously selecting inspection-worthy soil targets based on visual examination
- …and many, many, more.

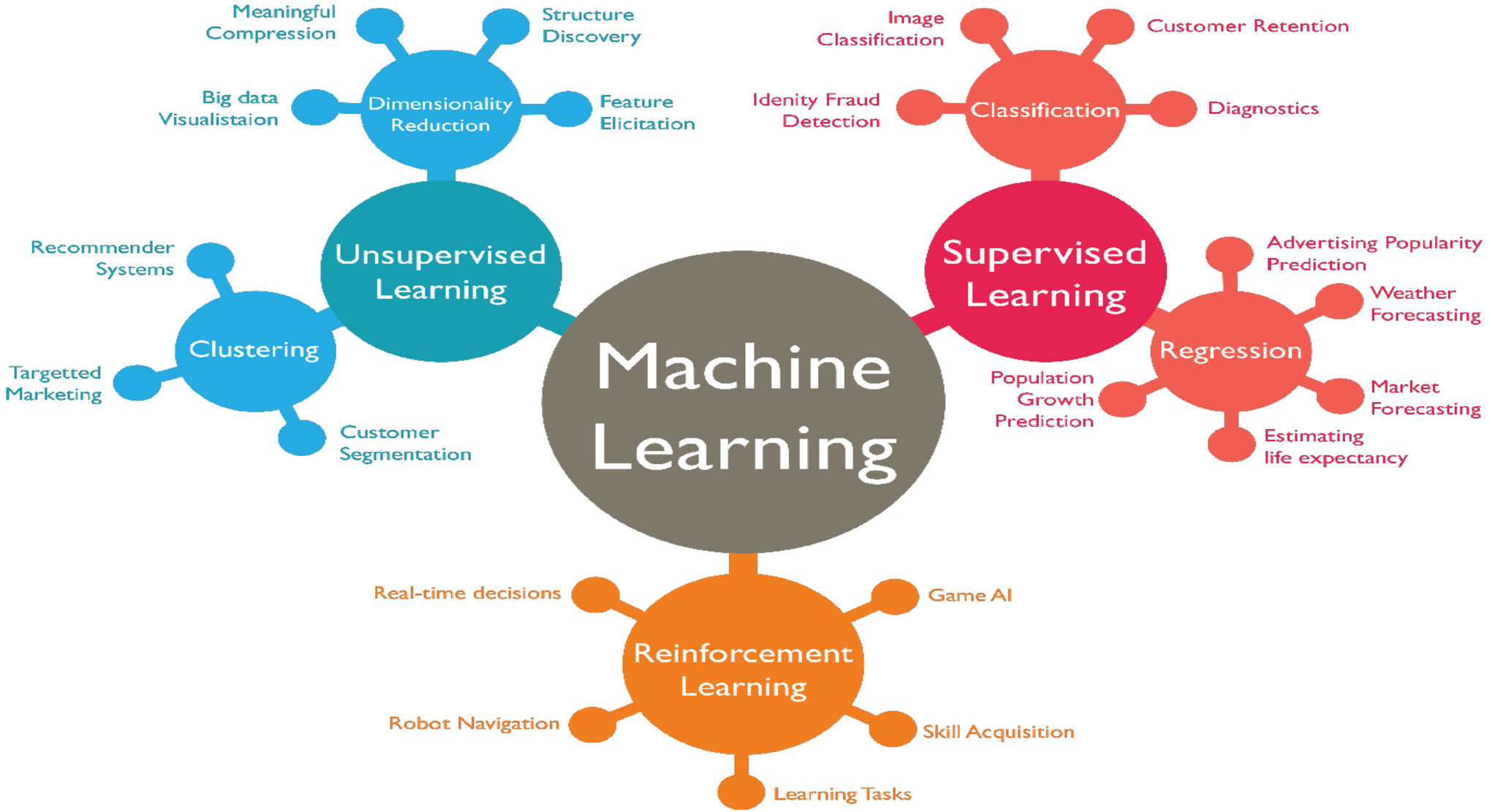
Machine learning types

- Re-inforcement Learning

Machine learning types



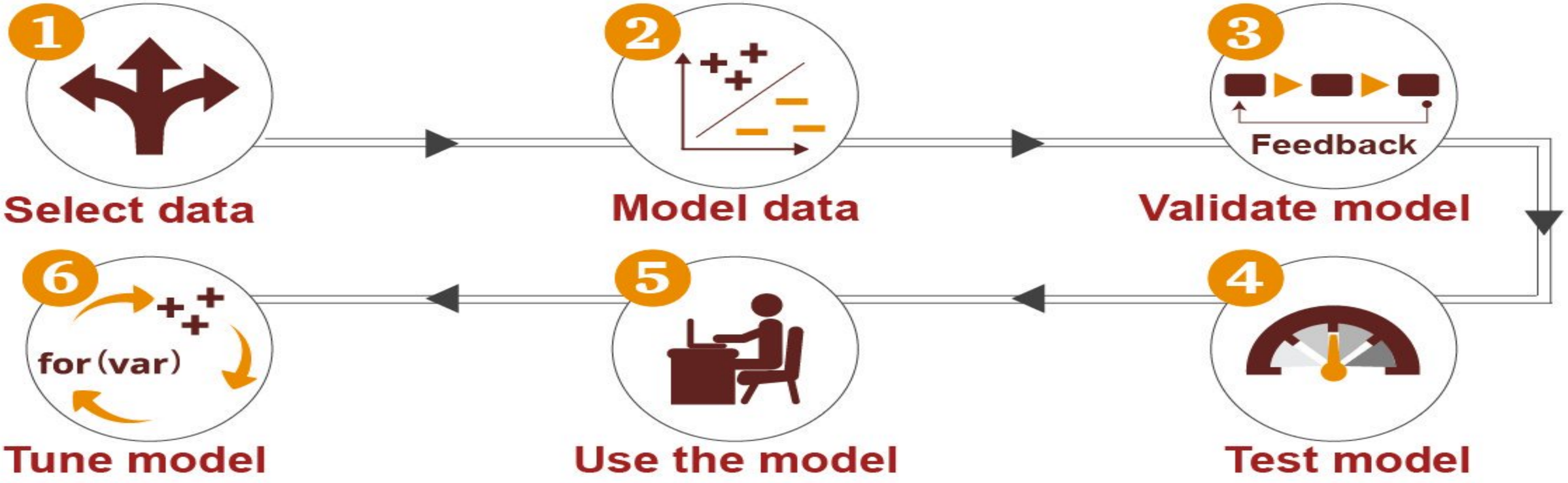
Machine learning types



Step to machine learning

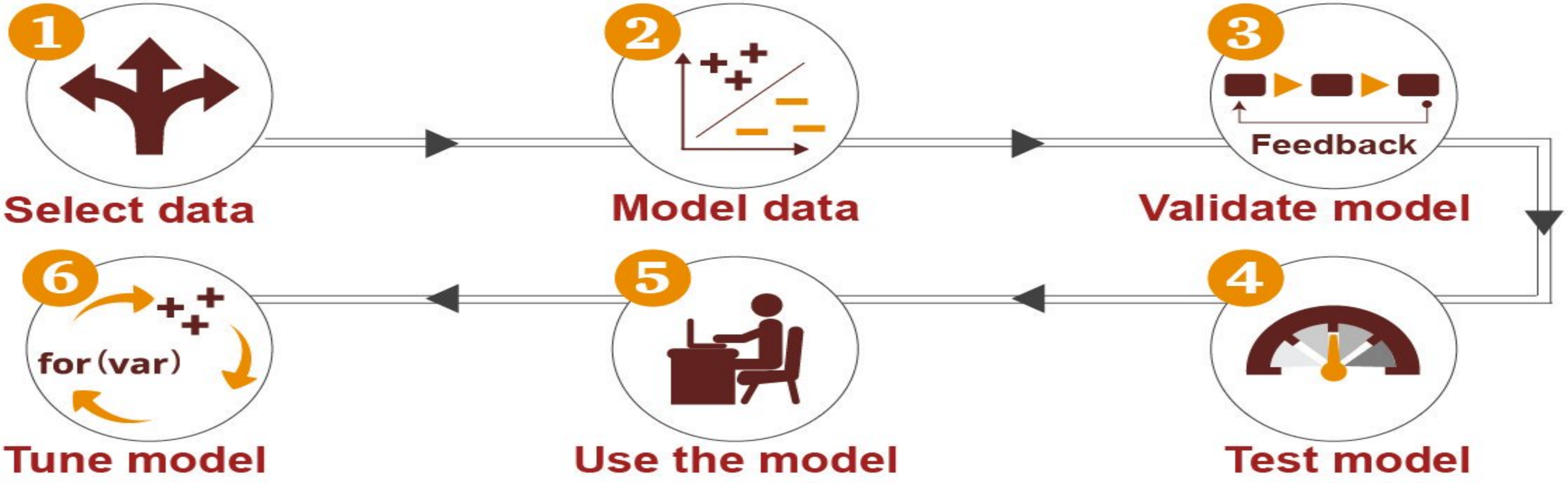
- Gathering data
- Preparing that data
- Choosing a model
- Training
- Evaluation
- Hyperparameter tuning
- Prediction.

How machine learning works



Source: pwc.com/nextintech

How machine learning works



Source: pwc.com/nextintech