Infofarm

DATA SCIENCE COMPANY

Hands-on introduction to

Data Science

Infofarm
DATA SCIENCE COMPANY

Course Schedule

Day 1: Introduction

- Introduction to DataScience.
- Crash course in Python and Scientific Libraries.

Day 2: Data Exploration

- Data Cleaning and Visualization

Day 3: Data Analysis and Prediction

Day 4: Bring your own data

First: Installation!

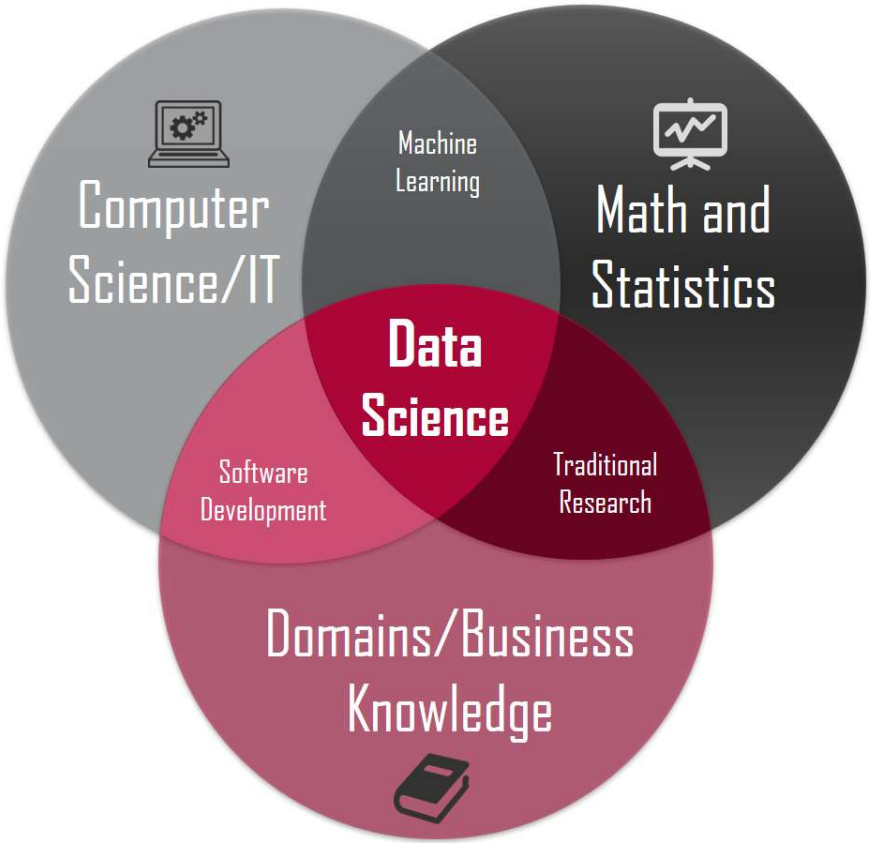
Very simple: Install Anaconda! It will give us the current Python interpreter 3.6 and most fo the packages that we will need.

Data Science

Definitions and concepts

Infofarm
DATA SCIENCE COMPANY

Data science – mix and match



Source: www.towardsdatascience.com

Infofarm
DATA SCIENCE COMPANY

Data Science

Data Science *is the art of handling data* ***using scientific methods, processes, algorithms and tools*** *in order to* ***extract knowledge*** *or business value from any* ***data*** *that is available,* ***structured or unstructured****.*

It is a concept ***unifying statistics, mathematics, computer science, machine learning and artifical intelligence*** *that is used to extract insights that are* ***predictive*** *or intelligent,* ***rather than descriptive****.*

Data science vs Big data and BI

Big Data

Problem domain of handling data that traditional systems have difficulties coping with because of the size, variety, speed, … of the data.

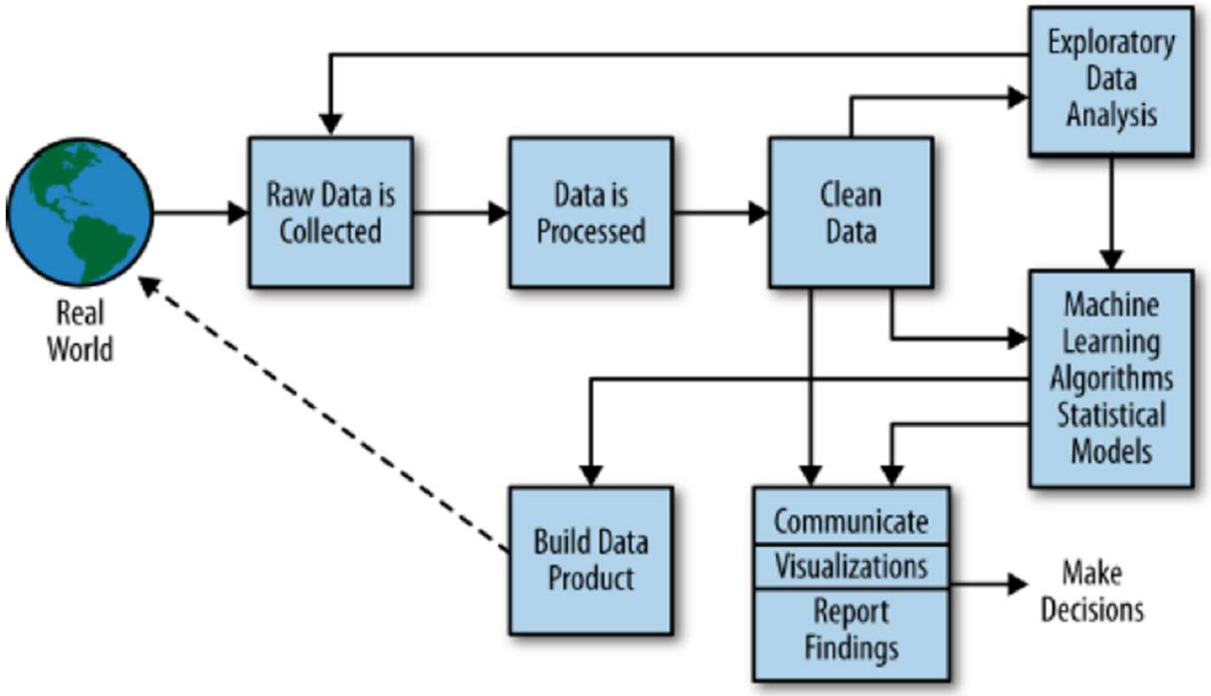
Data Science

Handling data using scientific methods and algorithms in order to extract knowledge from data. Used for developing intelligent systems and predictive models.

Business Intelligence

Gathering data in a structured matter for reporting and descriptive analytics.

Data Science Process

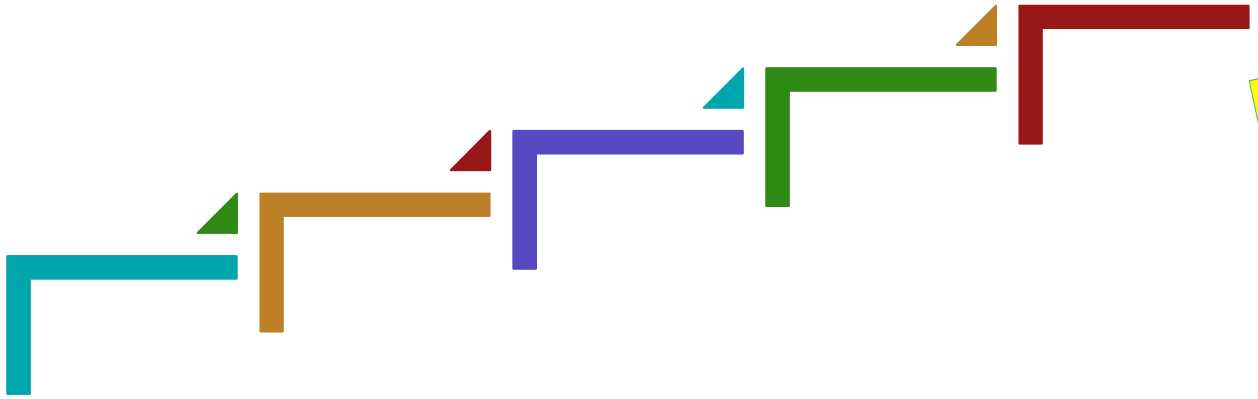


Source: O'Reilly – Hands-on machine learning with Scikit-Learn and Tensorflow

Infofarm

The Data Science maturity model

- Don’t run before you can walk: The **Data Science Maturity model**
  Each level builds on the quality of the underlying step. It’s science, not magic …



This is were the **hype** around Big Data and Data Science generates **unrealistic expectations**!

- Start off by simply **collecting** the data you need (type, quantity, quality)
- Then **report** on your current business (confirmative analysis)
- **Discover** new and valuable information (exploratory analysis)
- Build and test **prediction models** (predictive analysis)
- Steer your business based on advise output from your predictions (data-driven)

Source: O’Reilly – Hands-on machine learning with Scikit-Learn and Tensorflow

Infofarm
DATA SCIENCE COMPANY

The Data Science maturity model

Phase	Actions	Examples
Collect	Logging information Gathering data from different sources	Logging user actions on a website Using loyalty cards to id customers
Describe	Explorative Data Analysis Basic analytical functions Checking quantity and quality of data	Typical reporting Correlating data over sources
Discover	Finding correlations Building models	Finding similarly behaving customers
Predict	Building prediction models Formulating expectations for the future based on past info	Predict sales figures for a new product Predict whether a certain customer will or will not buy a certain product
Advise	Use prediction models to evaluate decision possibilities and pick the best	Target advertising to the right customer groups to optimize revenue

Collect

- Possibly multiple sources
  - Internal ‘clear’ data (excel sheets)
  - Internal ‘dark’ data (weblogs)
  - External data (social network data)
- Possibly different structures
  - Relational structured (RDBMS)
  - Structured, non-relational (NoSQL)
  - Unstructured (weblogs, social network)

Describe

Get to know your data by:

- Summary statistics
- Visualization
- Outlier detection
- Identify missing data

Describe

	sex	income	company	kids	age	changed	married	christmas_budget
1	male	43980.690	NA	2	42	three times	TRUE	402.09241
2	female	0.000	company6	3	47	once	FALSE	435.99572
3	female	21208.706	NA	2	NA	never	FALSE	317.48160
4	male	42111.054	company7	1	58	never	TRUE	487.62778
5	female	24457.319	company4	4	58	never	TRUE	764.70025
6	male	25376.901	company2	3	45	three times	TRUE	491.68038
7	female	2741.037	company7	3	21	twice	NA	289.25619
8	male	34384.994	company7	3	40	three times	FALSE	388.12018
9	male	40619.699	company9	2	29	once	FALSE	202.87002
10	female	80322.753	company7	4	42	never	TRUE	716.30891
11	male	51390.334	company5	4	21	three times	TRUE	399.64391
12	male	34633.316	company4	3	23	never	TRUE	279.12623
13	female	33271.236	company1	1	24	twice	TRUE	293.26005
14	male	54220.996	NA	1	45	twice	TRUE	416.76615
15	male	36063.802	company8	1	43	three times	FALSE	335.31159

Data cleansing

Source: www.indiamart.com

Infofarm

Cleaning data - Missing values

- Missing data affects some models more than others
- Even for models that handle missing data, they can be sensitive to it. Missing data for certain variables can result in poor predictions.
- Mssing data can be more common in production
- Missing value imputation can get very sophisticated

Cleaning data - Outliers

- What an outlier exactly is can be somewhat subjective
- Can be very common in multidimensional data
- Some models are less sensitive (more robust) to outliers than others. E.g. tree models are more robust that regression models.
- Outliers can be the result of bad data collection. Or they can be legitimate extreme values.
- Sometimes, outliers are the interesting data points we want to model (anomaly detection). E.g. Fraud detection

Discover

- Summary statistics (mean, stdev etc)
  - Percentiles can help identify the range for most of the data
  - Averages and medians can describe central tendency => Does this make sense business-wise?
  - Correlations can indicate strong relationships
- Visualize data
  - Box-plots can help identify outliers
  - Density plots and histograms show the spread of the data
  - Scatter plots can describe bivariate relationships

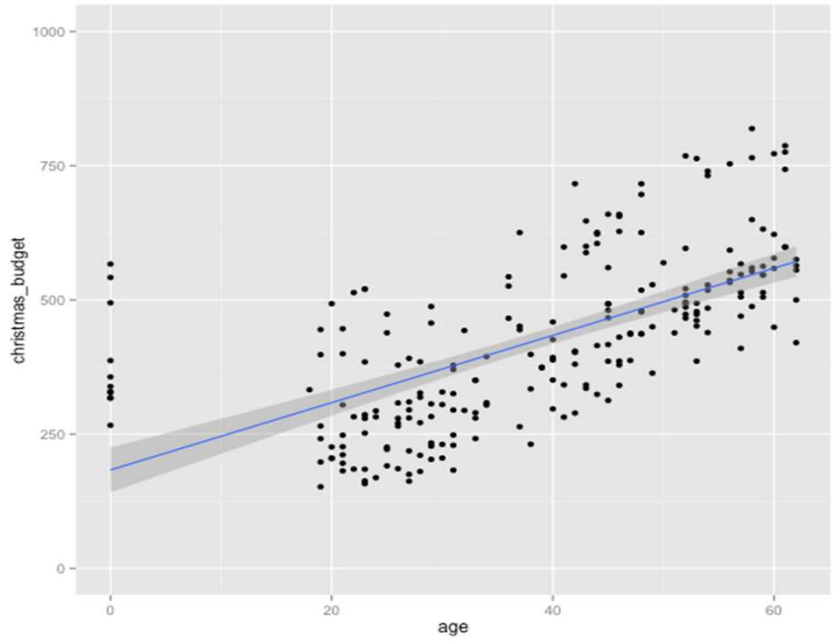
Discover

Correlations tell you that age and Christmas budget are related

	income	kids	age	christmas_budget
income	1.00000000	0.09158540	0.08338248	0.17018442
kids	0.09158540	1.00000000	0.01774910	0.48135056
age	0.08338248	0.01774910	1.00000000	0.63112796
christmas_budget	0.17018442	0.48135056	0.63112796	1.00000000

Discover

Visual insights rise: age might have an influence on Christmas budget



Infofarm
DATA SCIENCE COMPANY

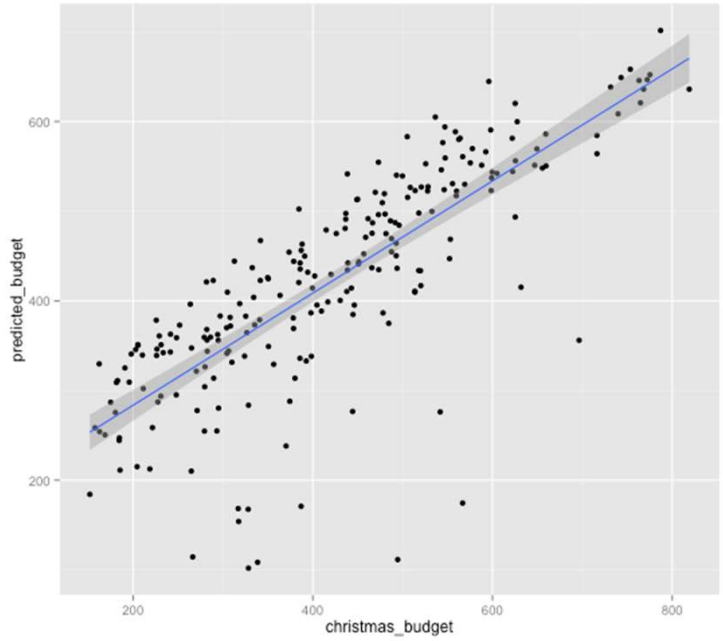
Predict

When understanding your data, you are finally in the phase to:

- Create prediction models
- Create predictions
- Score your models
- Find the best model on new data

Predict

A model was built using age, income and number of kids.
Predicted versus actual values are displayed.



Infofarm
DATA SCIENCE COMPANY

Advise

You now have your model to advise your customer what to do, in this case:

- Identify customers with biggest opportunities to increase sales
- Identify customer profiles who you want to prevent from churning, no matter what
- You know your business better than us, so we identify possibilities together!

Machine Learning

Definitions and concepts

Infofarm
DATA SCIENCE COMPANY

Definitions

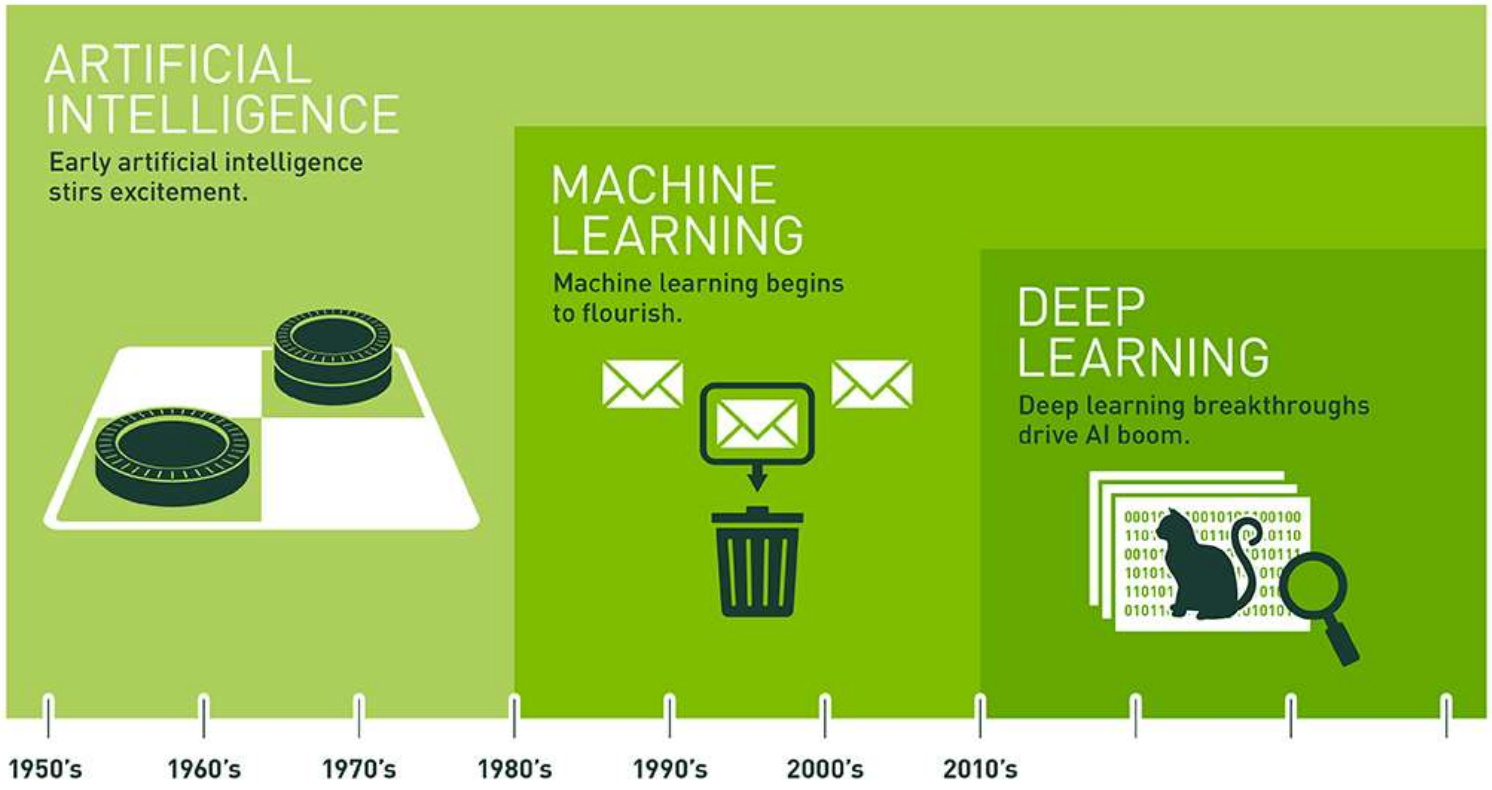
AI

Theory and development of computer systems that are able to perform tasks that would normally require human intelligence

Machine learning

Subset of AI that focuses on the ability of machines to receive a set of data and learn for themselves. They change algorithms (or their parameters) as they learn more about the information they are processing.

AI vs Machine learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: www.datasciencecentral.com

Infofarm
DATA SCIENCE COMPANY

What is an algorithm?

Infofarm
DATA SCIENCE COMPANY

Types of Machine Learning

- Supervised vs unsupervised learning
- Semisupervised learning
- Reinforcement learning

Supervised / Unsupervised Learning

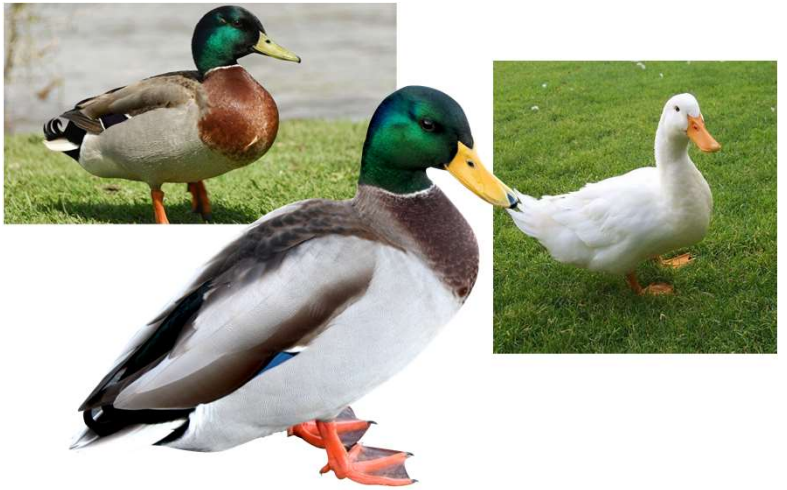
Supervised

- Predictive
- Learns by example
- Labeled data typically split into test set and training set
- More examples = better model
- Typical tasks: classification (spam filter) and prediction (oil price)

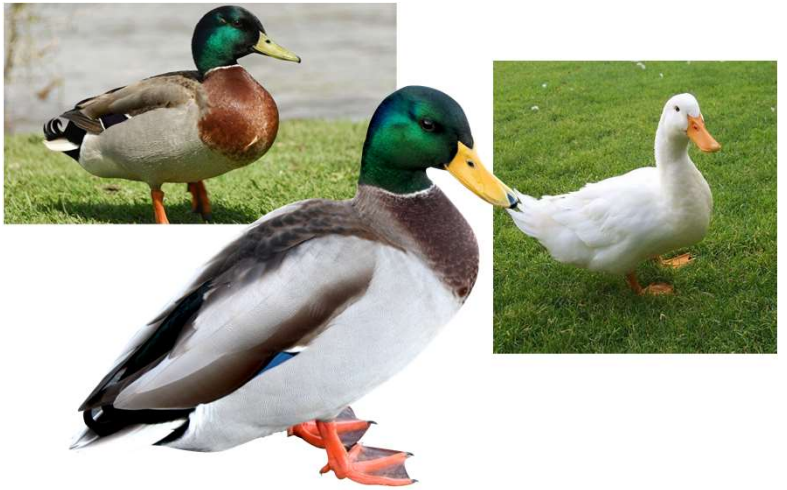
Unsupervised

- Descriptive
- No examples needed
- You provide labels after the fact
- Sometimes hard to see why the model makes certain decisions
- Typical tasks:

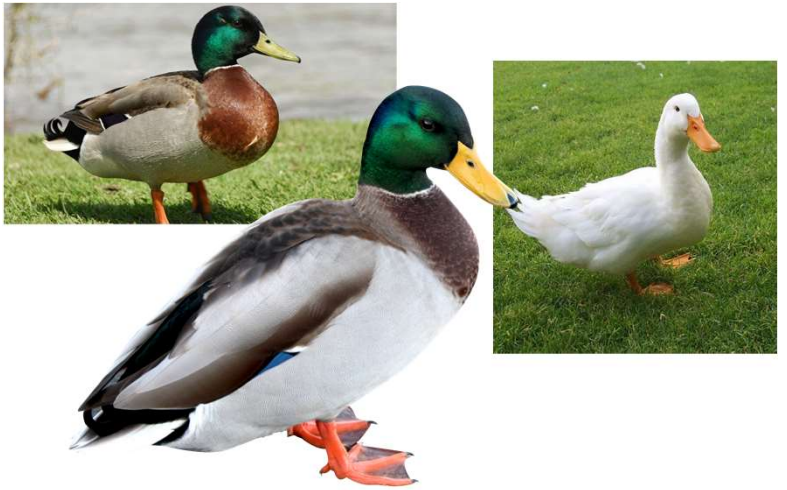
Supervised Learning



Supervised Learning

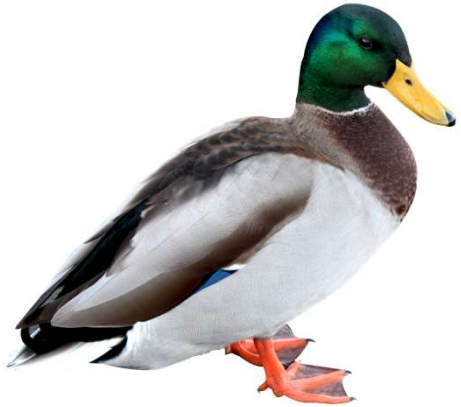
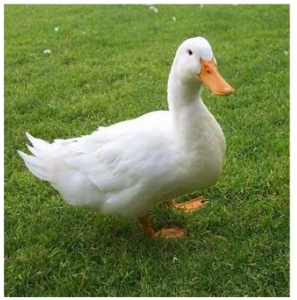
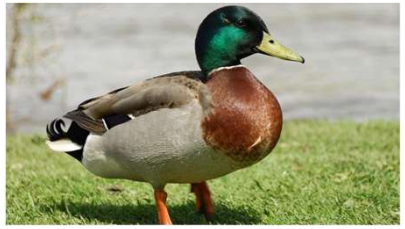


Supervised Learning

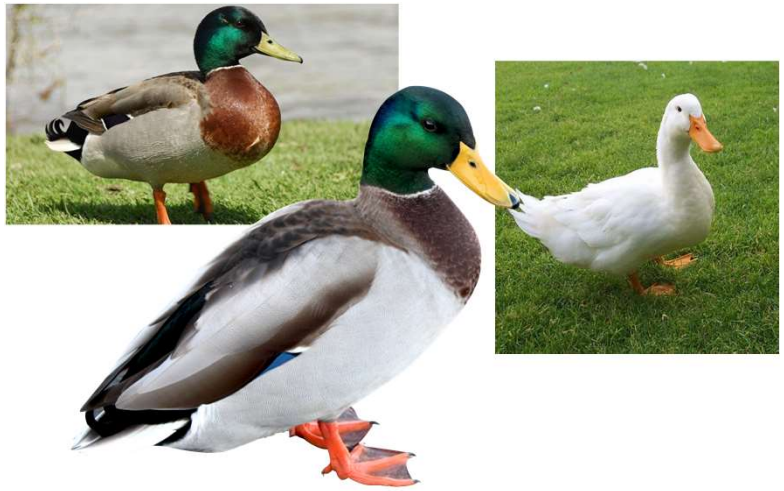


Unsupervised Learning



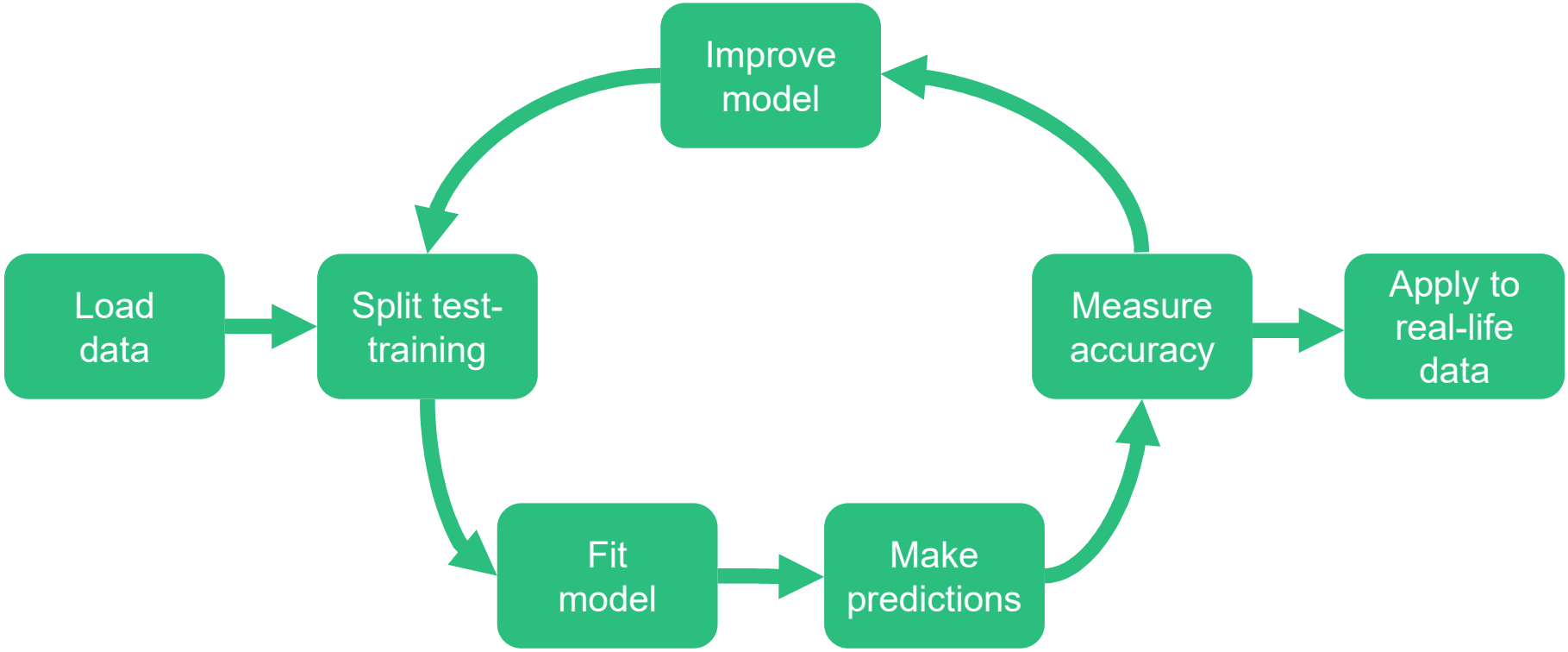

Unsupervised Learning



Most popular algorithms

Supervised Learning	Unsupervised learning
K-Nearest neighbors	K-means clustering
Linear regression	Hierarchical clustering Analysis
Logistic regression	Principal Component Analysis
Support Vector Machines	Kernel PCA
Decision Trees and Random forests	Apriori association rule learning
Neural networks (*)	Eclat association rule learning

Machine learning process



Infofarm
DATA SCIENCE COMPANY

Supervised learning

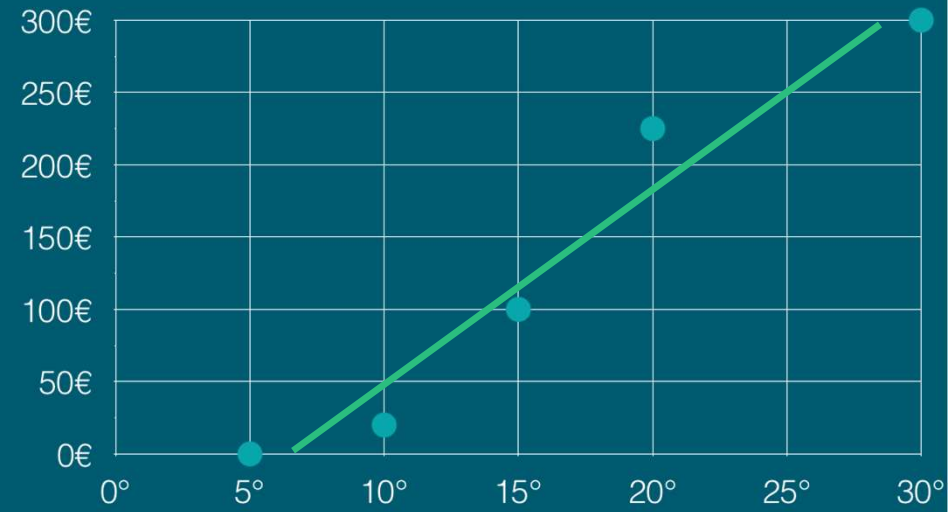
Algorithms and examples

Infofarm
DATA SCIENCE COMPANY

Regression

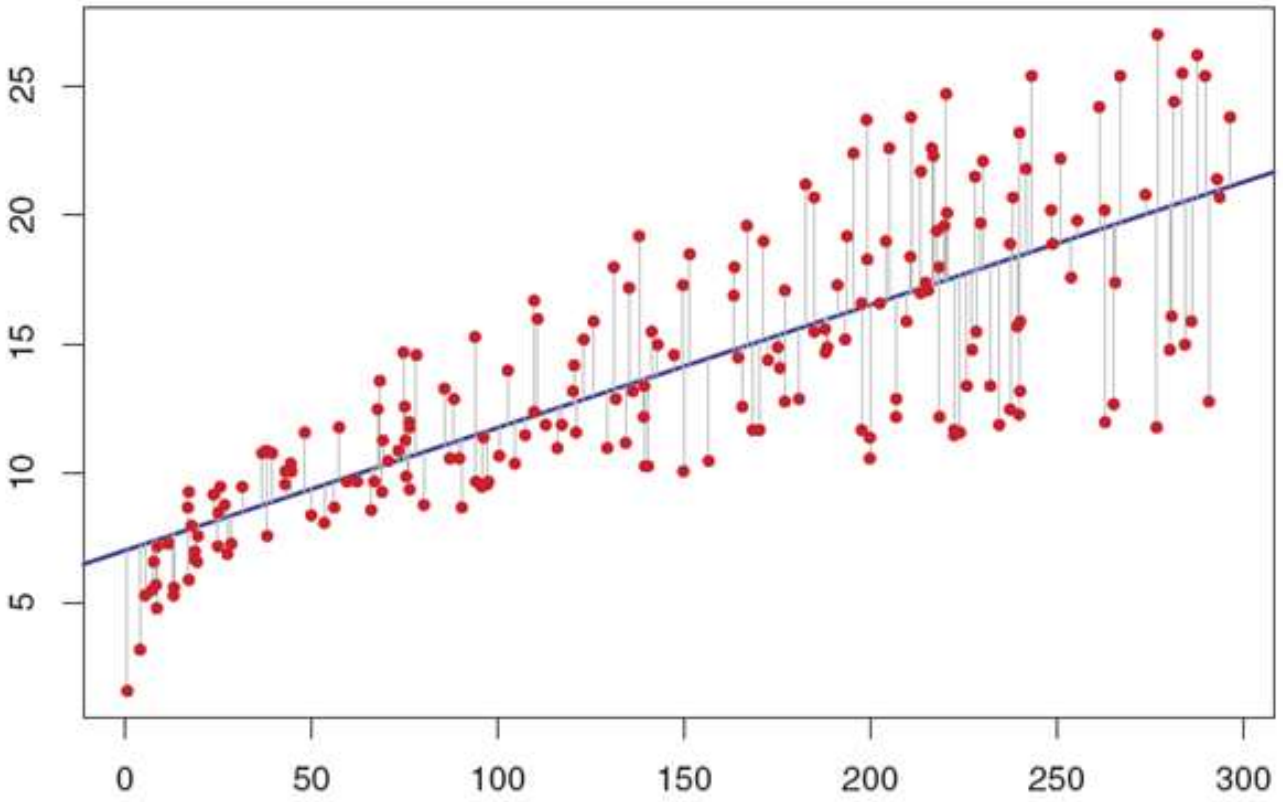
- Predict a continuous value based on historical examples

Temperature	Ice Cream Revenue
30°	300 €
15°	100 €
20°	225 €
10°	20 €
5°	0 €
25°	????



Infofarm
DATA SCIENCE COMPANY

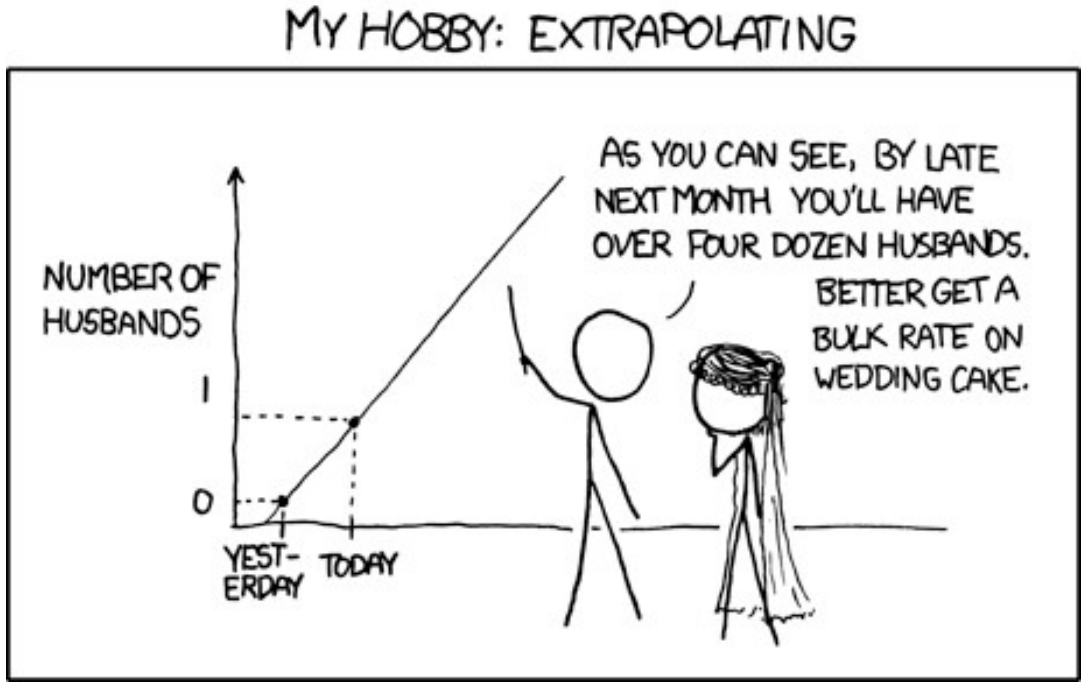
Linear regression



Source: www.towardsdatascience.com

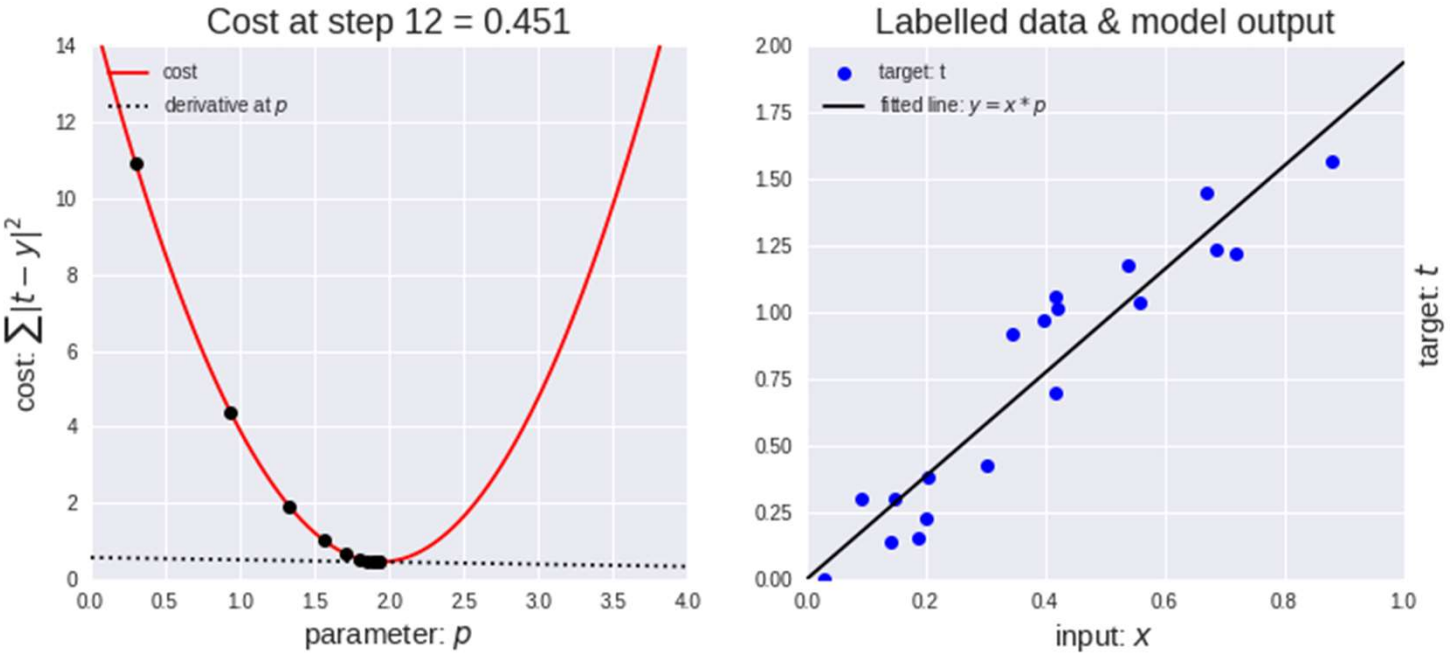
Infofarm
DATA SCIENCE COMPANY

Regression : example



Infofarm
DATA SCIENCE COMPANY

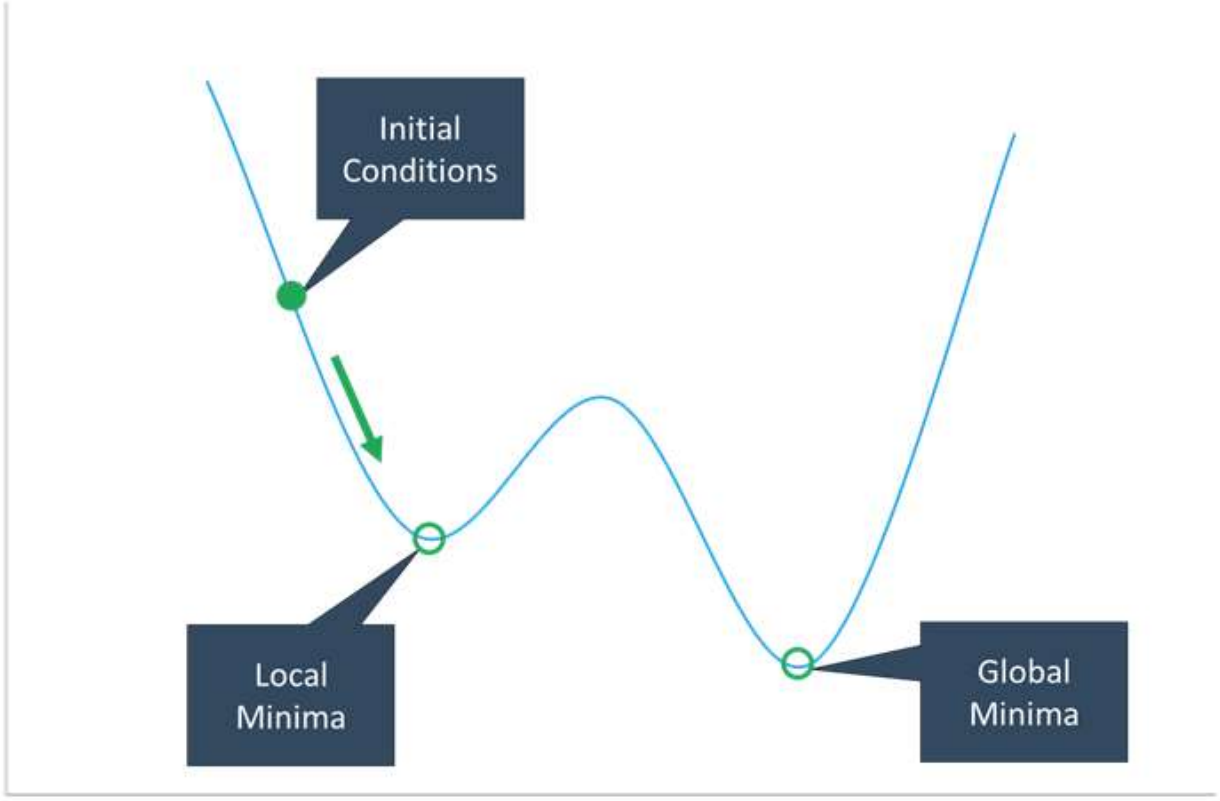
Gradient descent



Source: www.medium.com

Infofarm
DATA SCIENCE COMPANY

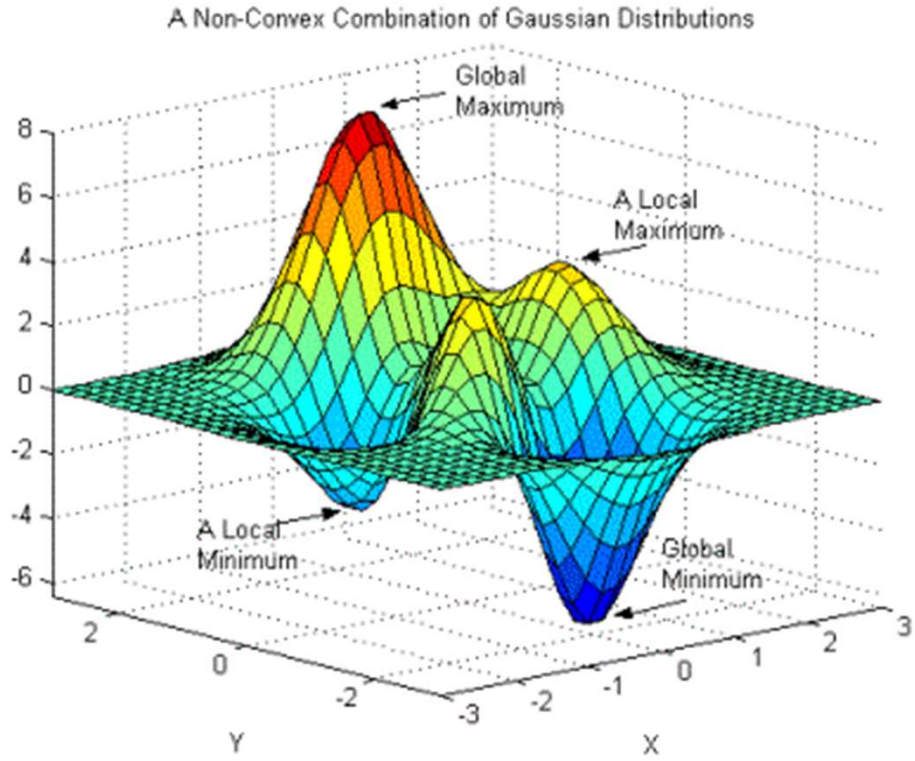
Local vs global optima



Source: www.engineerexcel.com

Infofarm
DATA SCIENCE COMPANY

Local vs global optima



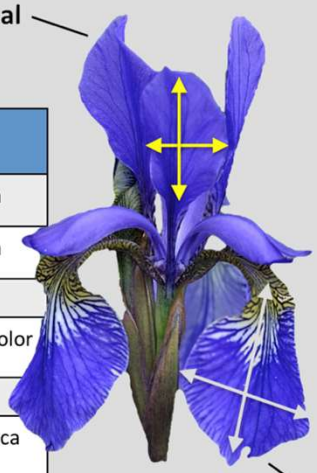
Source: www.lindo.com

Infofarm
DATA SCIENCE COMPANY

Classification

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

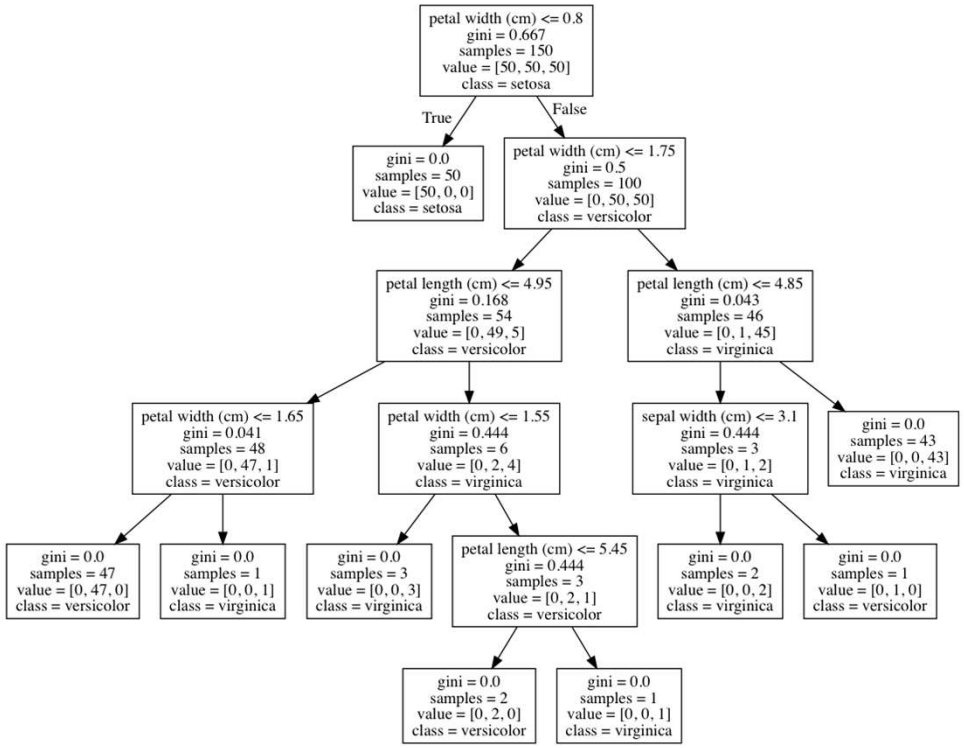


Features
(attributes, measurements, dimensions)

Class labels
(targets)

Infofarm
DATA SCIENCE COMPANY

Classification



Infofarm
DATA SCIENCE COMPANY

Classification

- Classifying objects to fit a certain category
- Typically done training on labeled data
- Algorithm builds *Decision Tree* for quick classification

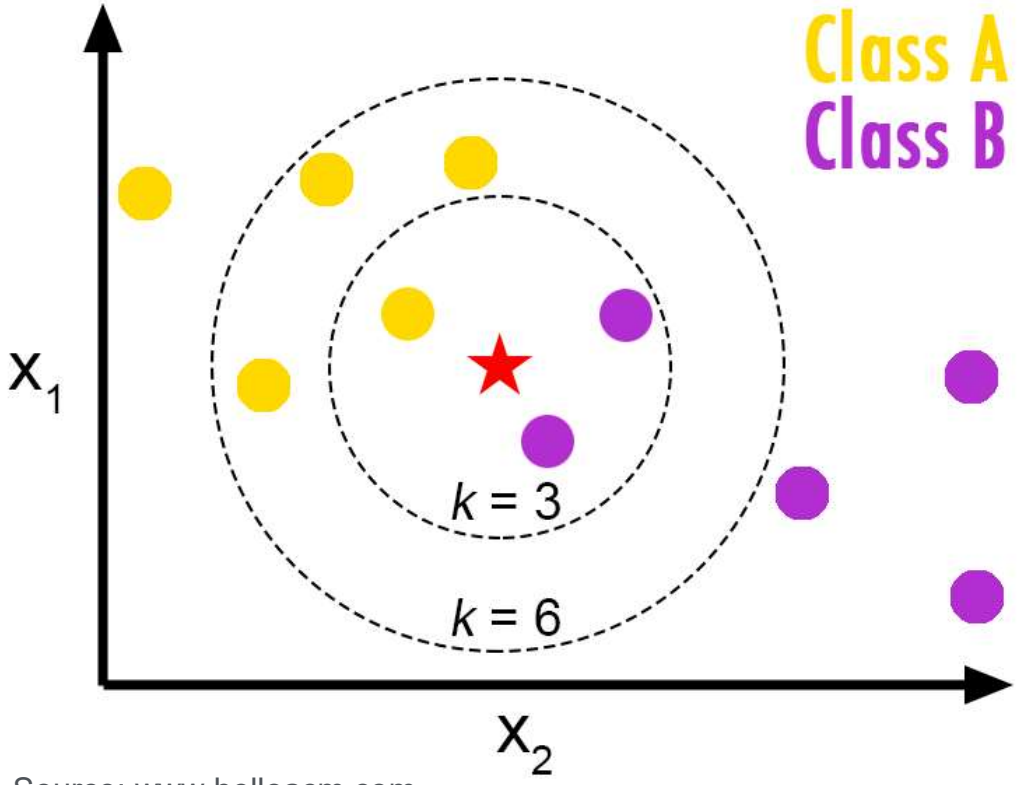
Examples

Labeling Spam

Categorizing text

Any case where you want to automatically label an item

K-Nearest Neighbor



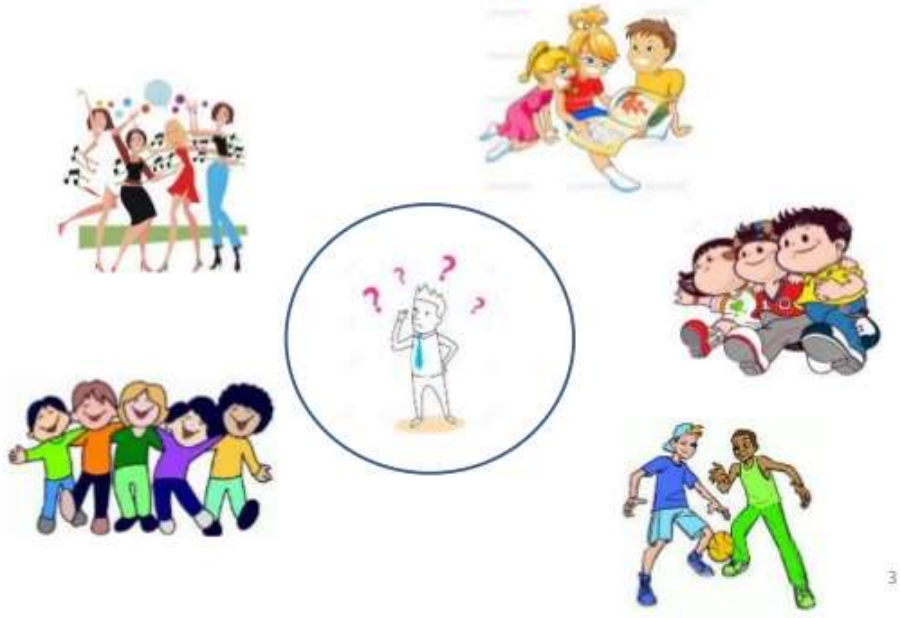
Source: www.helloacm.com

Infofarm
DATA SCIENCE COMPANY

KNN - example

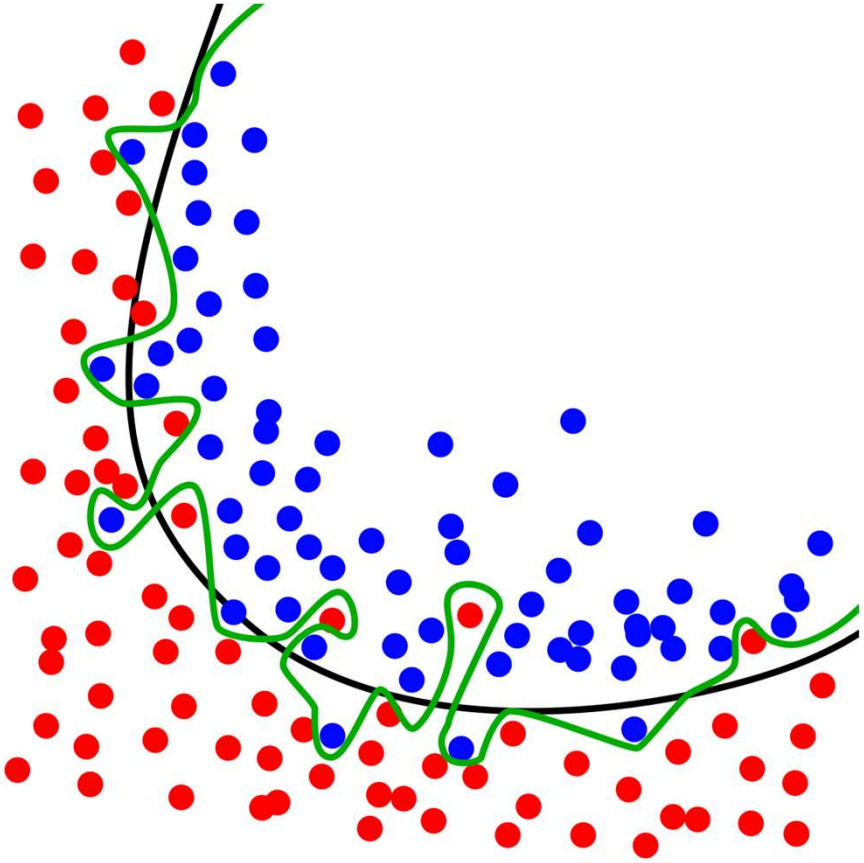
Simple Analogy..

- Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.



Infofarm
DATA SCIENCE COMPANY

Overfitting



Source: https://en.wikipedia.org/wiki/Overfitting

Infofarm
DATA SCIENCE COMPANY

Regression vs Classification

Classification

DISCRETE VALUES

- **Which** trip will I take?
- **What** animal is this?
- **Where** will I forward this mail?

Regression

CONTINUOUS VALUES

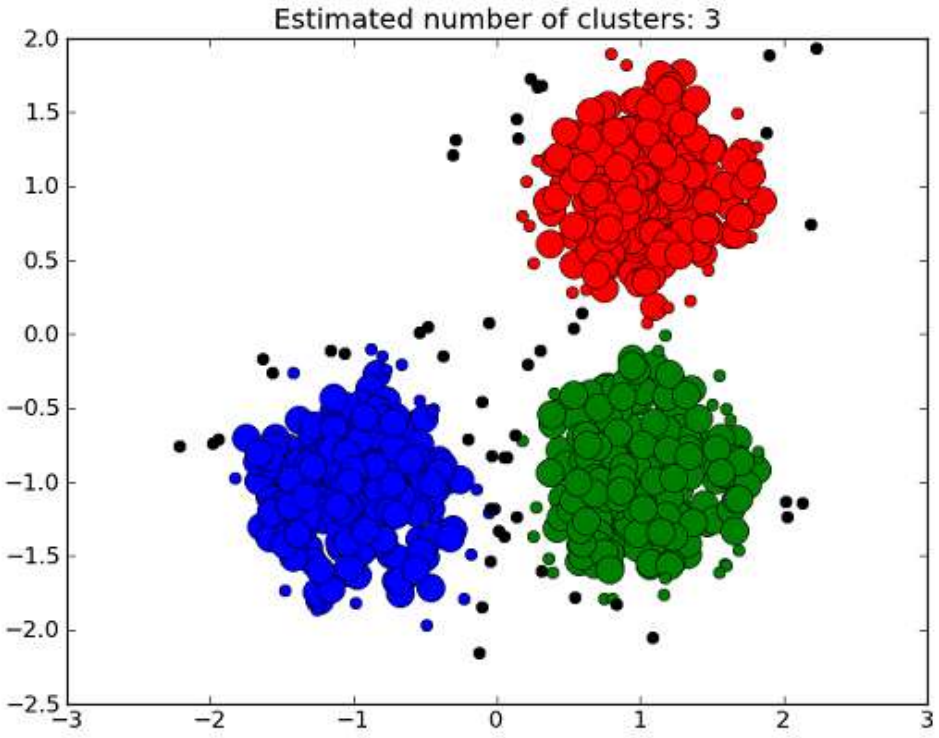
- **How** much will I earn?
- **How** high will this boy grow?
- **How** long will this take?

Unsupervised learning

Algorithms and examples

Infofarm
DATA SCIENCE COMPANY

Clustering



Source: www.cssanalytics.wordpress.com

Infofarm
DATA SCIENCE COMPANY

Clustering

- Calculating distances between items, and grouping them so they are more close to each other than the items in other groups
- K-means is the most well-known algorithm

Examples

Grouping Images

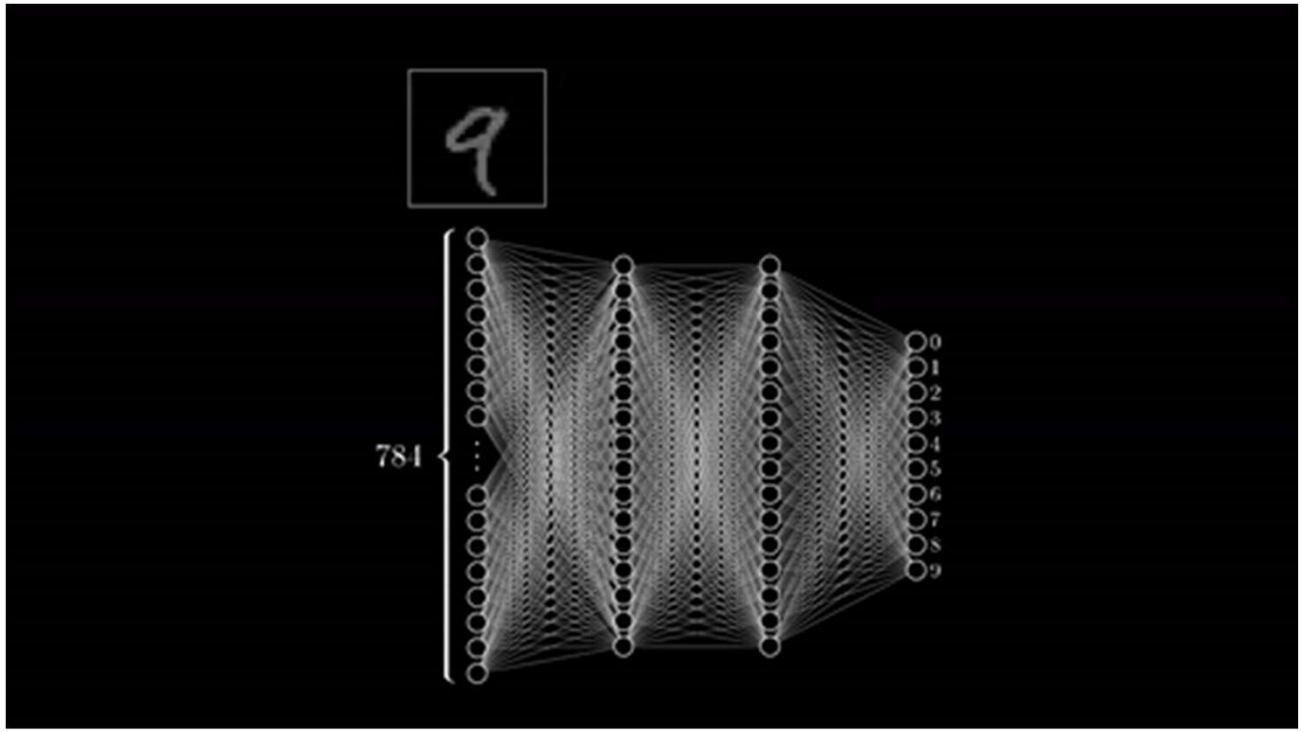
Customer Segmentation

Finding similar items (text grouping)

Fraud Detection

Predictions

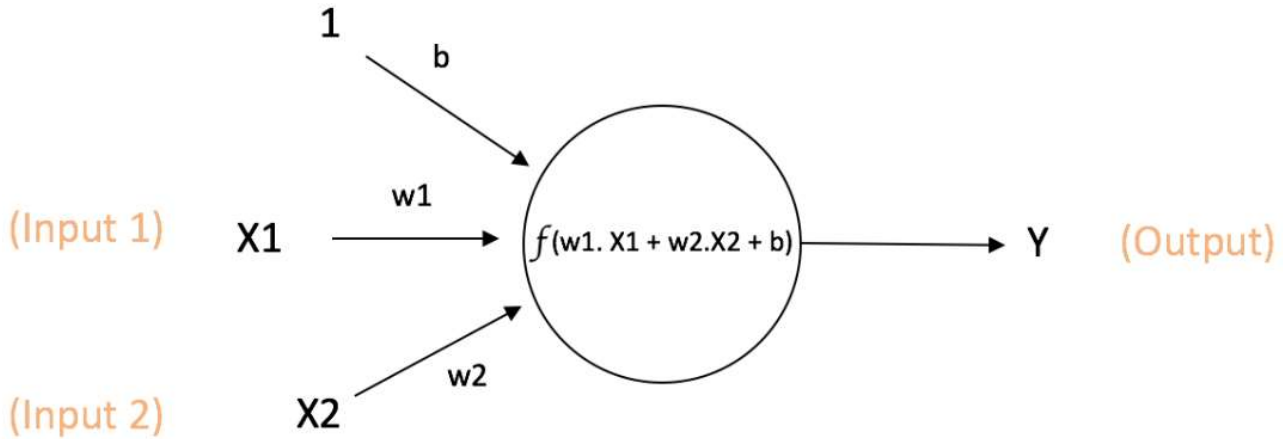
Neural networks



Source : 3Blue1Brown - But what *is* a Neural Network? | Chapter 1, deep learning

Infofarm
DATA SCIENCE COMPANY

Neural network - perceptron



Output of neuron = Y= f(w1. X1 + w2.X2 + b)

Infofarm
DATA SCIENCE COMPANY

Image Recognition: Clustering

Infofarm
DATA SCIENCE COMPANY

Image Recognition: Clustering

Infofarm
DATA SCIENCE COMPANY

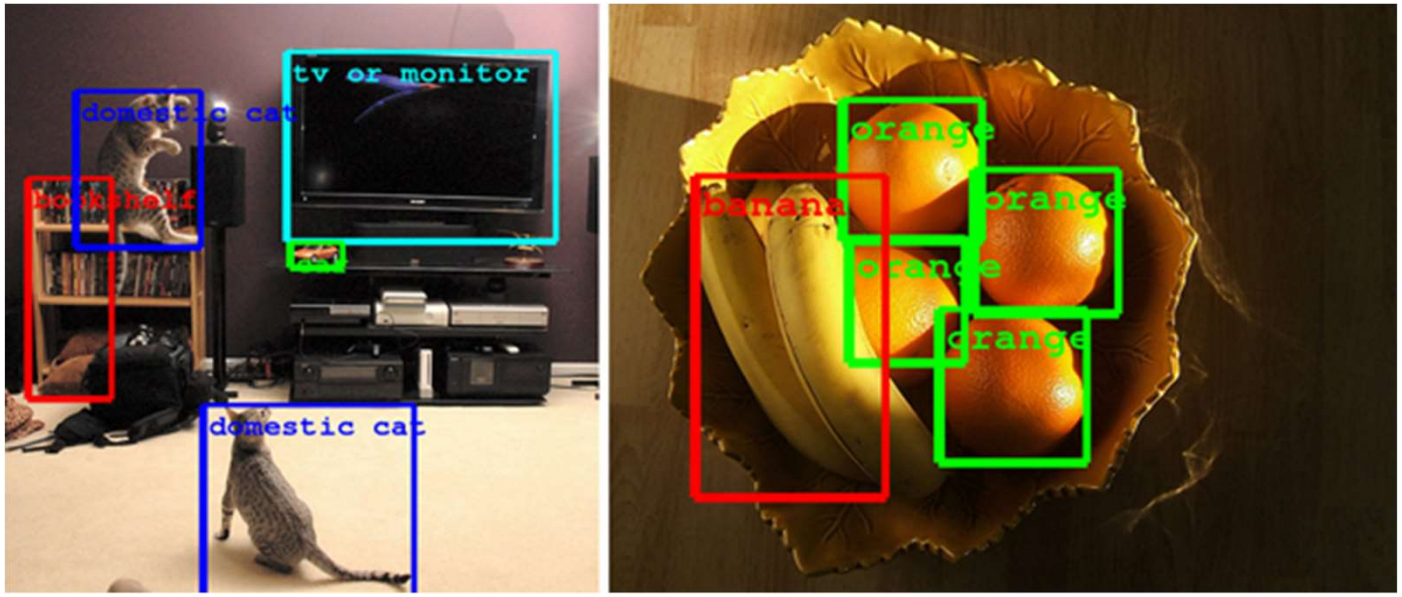







Image Recognition

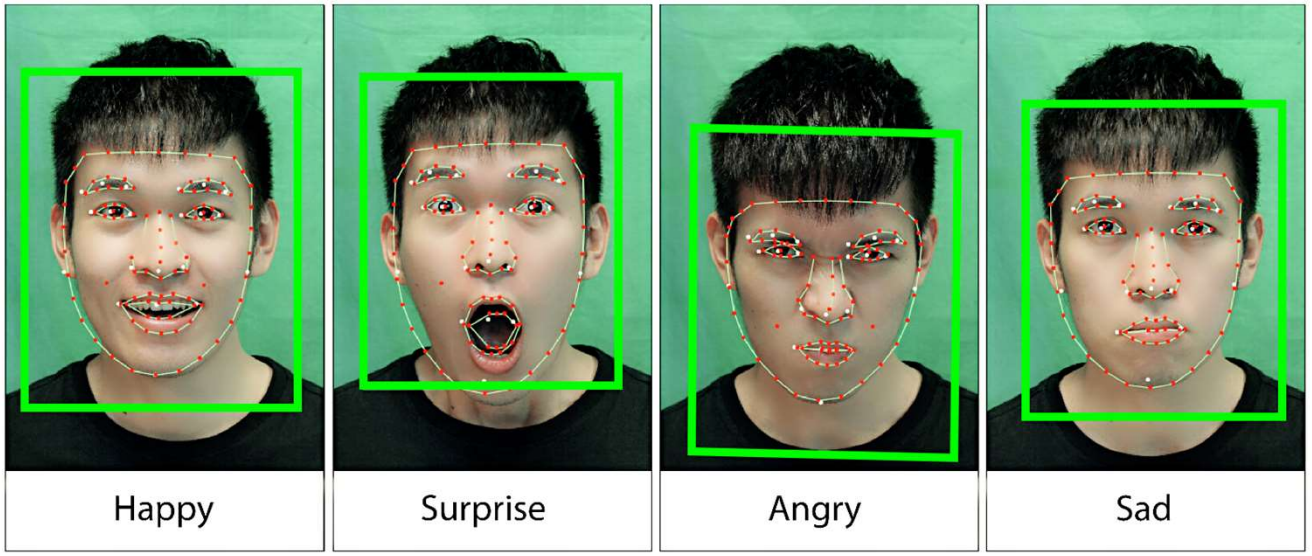
Object, Scene and Activity Detection



Infofarm
DATA SCIENCE COMPANY

Image Recognition

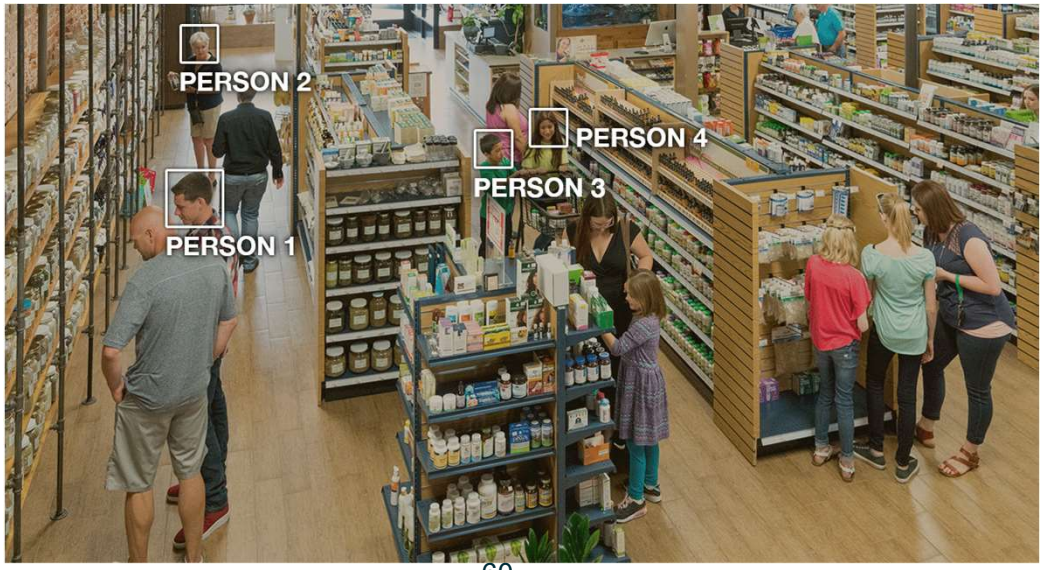
Classifying Images



Infofarm
DATA SCIENCE COMPANY

Video

- Everything you can do with regular images, you can do with video
- Added bonus: tracking



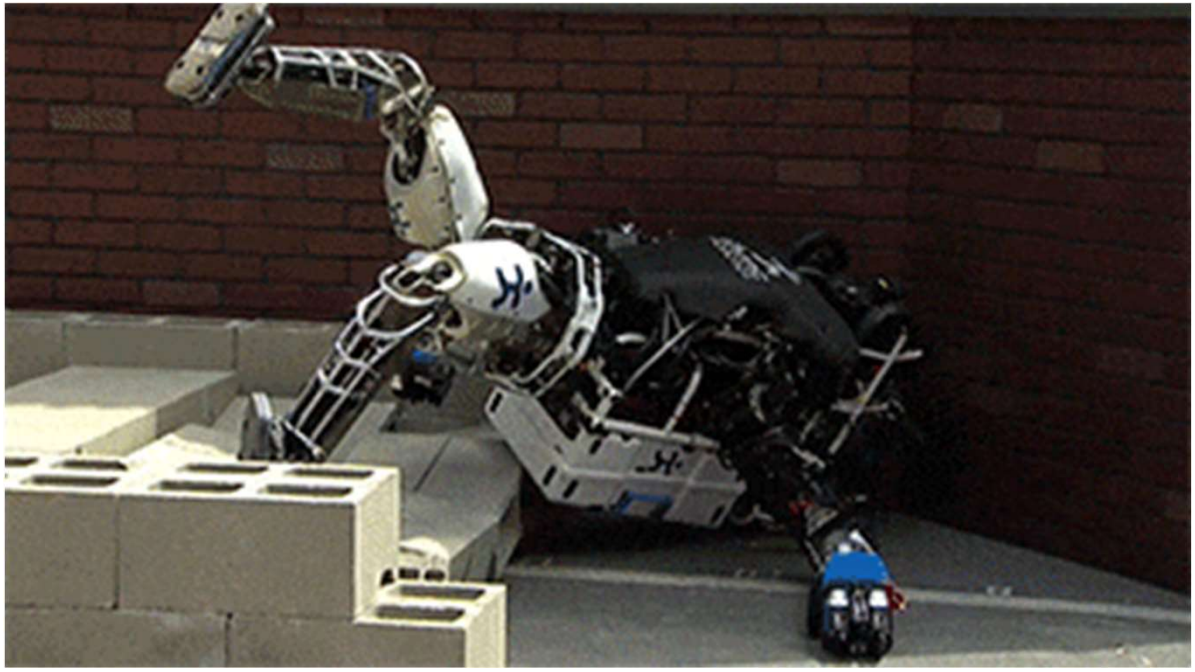
Infofarm
DATA SCIENCE COMPANY

Reinforcement learning

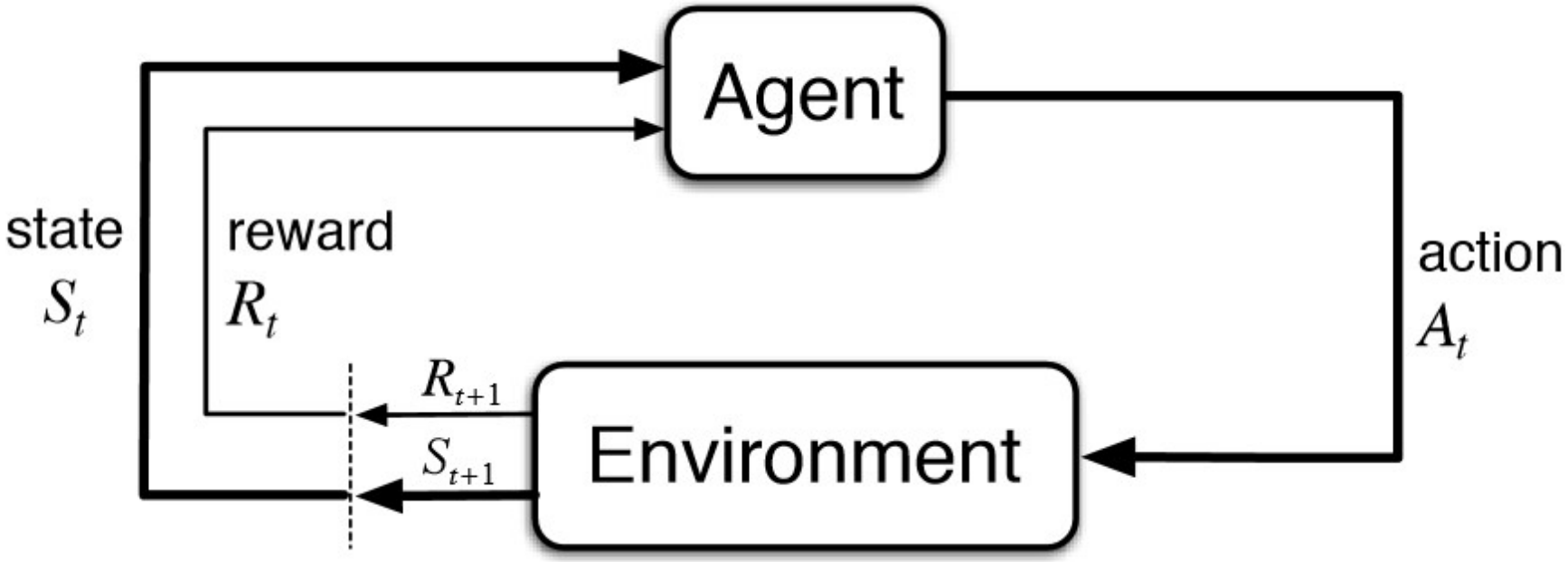
Concepts

Infofarm
DATA SCIENCE COMPANY

Reinforcement learning



Reinforcement learning



Infofarm
DATA SCIENCE COMPANY

Special Kid: Reinforcement Learning

- Combines Supervised and Unsupervised Learning
- Model learns itself doing a task (unsupervised)
- It is rewarded / penalised when the task goes well or wrong (supervised)
- Frequently used in simulations / AI modelling – Robotics and gaming

Reinforcement Learning example

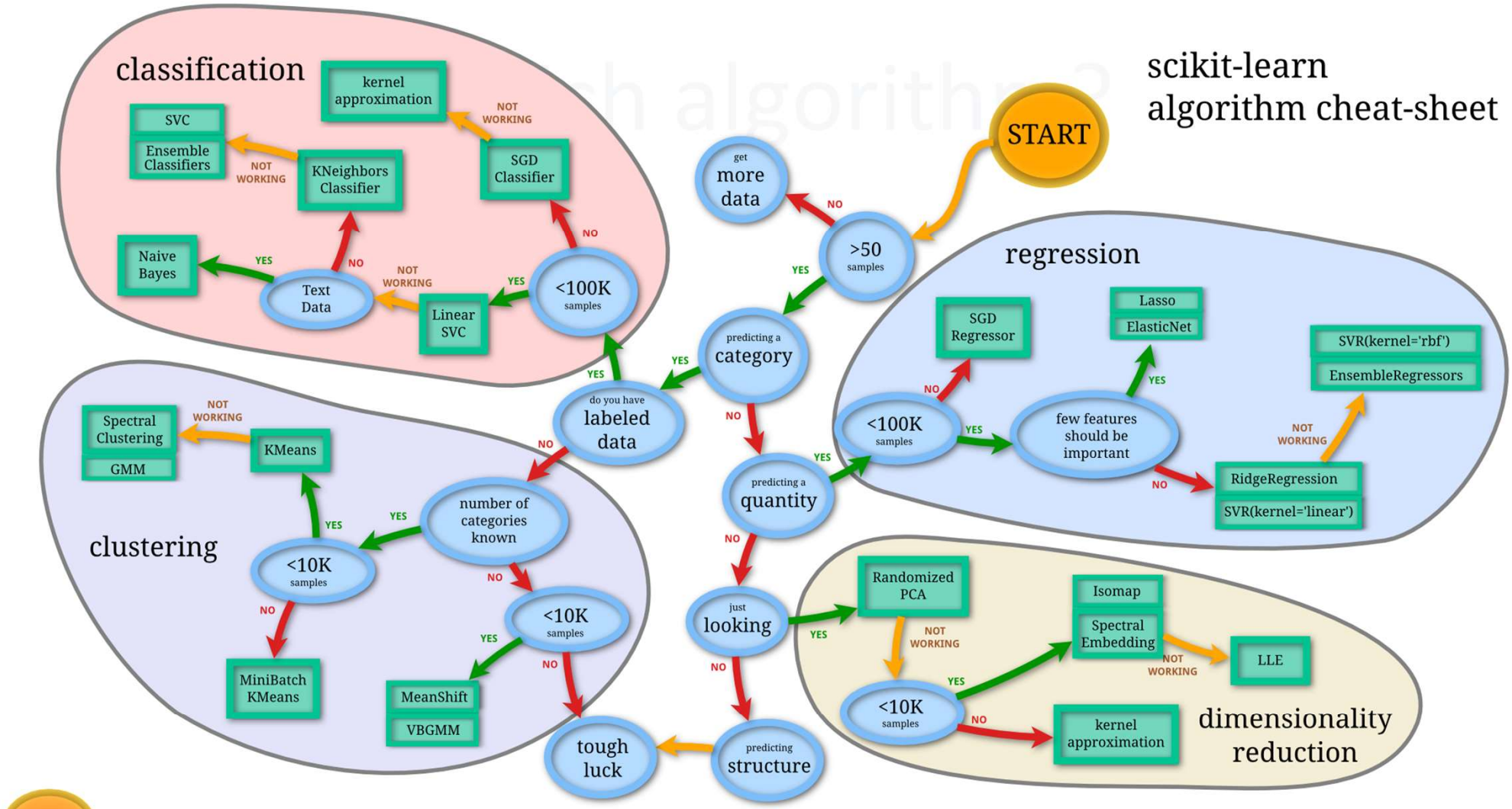
Infofarm
DATA SCIENCE COMPANY

Extra topics

Concepts

Infofarm
DATA SCIENCE COMPANY

scikit-learn algorithm cheat-sheet



Back

scikit learn

Data Science Skillset

Math & Statistics
Machine Learning
Decision Trees
Random Forest
Clustering
Optimization

Soft Skills
Interest in the business
Curiosity
Influence
Strategic Decision Making
Translate insights to actions

Programming Skills
Java, Scala, Python, …
R, Sas, SPSS, ...
SQL, NoSQL
Hadoop: Pig, Hive, ...
Spark
Cloud Systems

Visualisation
D3.js, Tableau, Kibana, …
Sell research to management
Communication
Visual Design

R and R Studio

- Open Source and free programming language and tool
- Statistical computing, data analyis, data mining
- Broad library of packages available (through CRAN
- Comparable to SAS and SPSS
- GUI with R Shiny

Python

- Programming language becoming more and more popular for data analytics and machine learning
- Easy for rapid prototyping because of dynamic typing and late binding
- Standard for Deep Learning

Python – Core Libraries

- **NumPy**

  Provides operations on n-dimensional arrays and matrices. Can do vectorization of mathematical operations which speeds up execution

- **SciPy**

  SciPy contains modules for linear algebra, optimization, integration, and statistics. The main functionality of SciPy library is built upon NumPy, and its arrays thus make substantial use of NumPy.

- **Pandas**

  Library for working with labeled series of data and dataframes (tables). Allows for easy data manipulation, aggregation and visualisation

Python – Visualisation Libraries

- **Matplotlib**

  Higly customizable visualisations in python. Used for line plots, scatter plots, bar charts, pie charts, stem plots, spectograms, …

- **Bokeh**

  Library focussing on interactive visualisations. Presents in browsers using data driven documents (d3.js)

Python – Machine Learning

- **Scikit-learn**

  The scikit-learn library exposes a concise and consistent interface to the common machine learning algorithms, making it simple to bring ML into production systems. The library combines quality code and good documentation, ease of use and high performance and is de-facto industry standard for machine learning with Python.

  Surprise

  Surprise stands for stands for Simple Python RecommendatIon System Engine. It provides various ready-to-use prediction and recommendation algorithms.

Python – Artificial Intelligence

- **OpenCV-Python**

  Python library originally written in C++ for (realtime) Computer Vision cases.

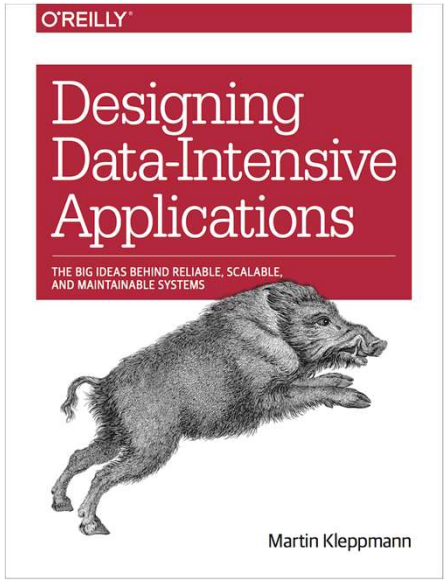
- **NTLK – Natural Language Toolkit**

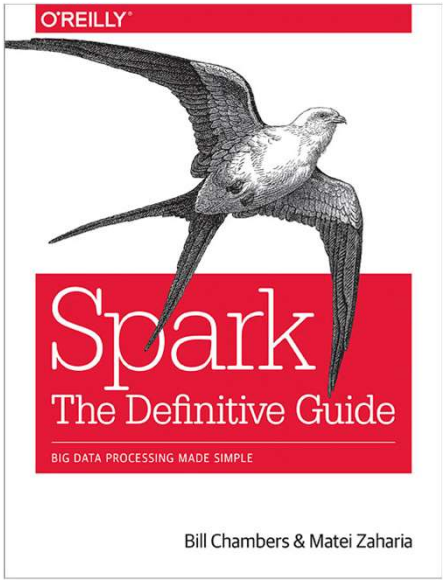
  NLTK allows a lot of operations such as text tagging, classification, and tokenizing, name entities identification, building corpus tree that reveals inter and intra-sentence dependencies, stemming, semantic reasoning.

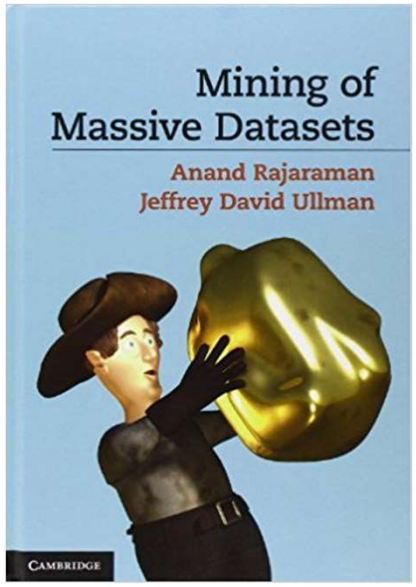
Cloud Options

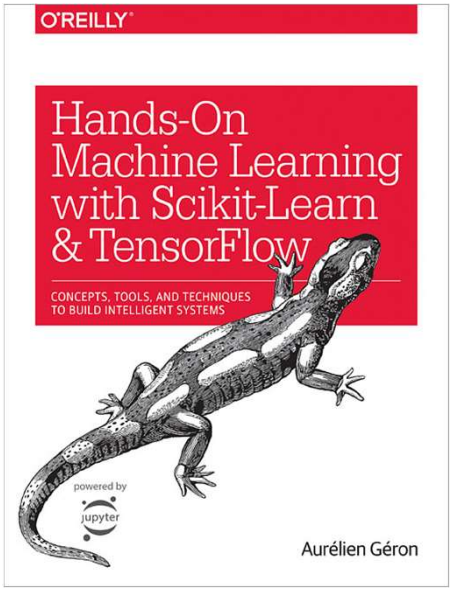
- Lots of pre-trained models and functions available
- More and more focus from cloud providers (Amazon, Microsoft, Google, IBM)
- For example: Amazon offering

Further Reading









Infofarm
DATA SCIENCE COMPANY