

Data science: big data, data analytics algorithms & tools

Geneveva Vargas-Solar
French Council of Scientific Research, LIG
geneveva.vargas@imag.fr

<http://vargas-solar.com/data-centric-smart-everything/>

<https://classroom.google.com/c/MTQ4MzcwMjY1MDEz?cjc=5bz2tk6>

Slack channel: https://join.slack.com/t/colenationale-5jr8199/shared_invite/zt-hhf9euv7-bmp7Kn9LL68RyzdhJnbKxA





DATA SCIENCE

The representation of complex environments by rich data opens up the possibility of applying all the scientific knowledge regarding how to infer knowledge from data

Definition:

- Methodology by which actionable insights can be inferred from data
- Complex, multifaceted field that can be approached from several points of view: ethics, methodology, business models, how to deal with big data, data engineering, data governance, etc.

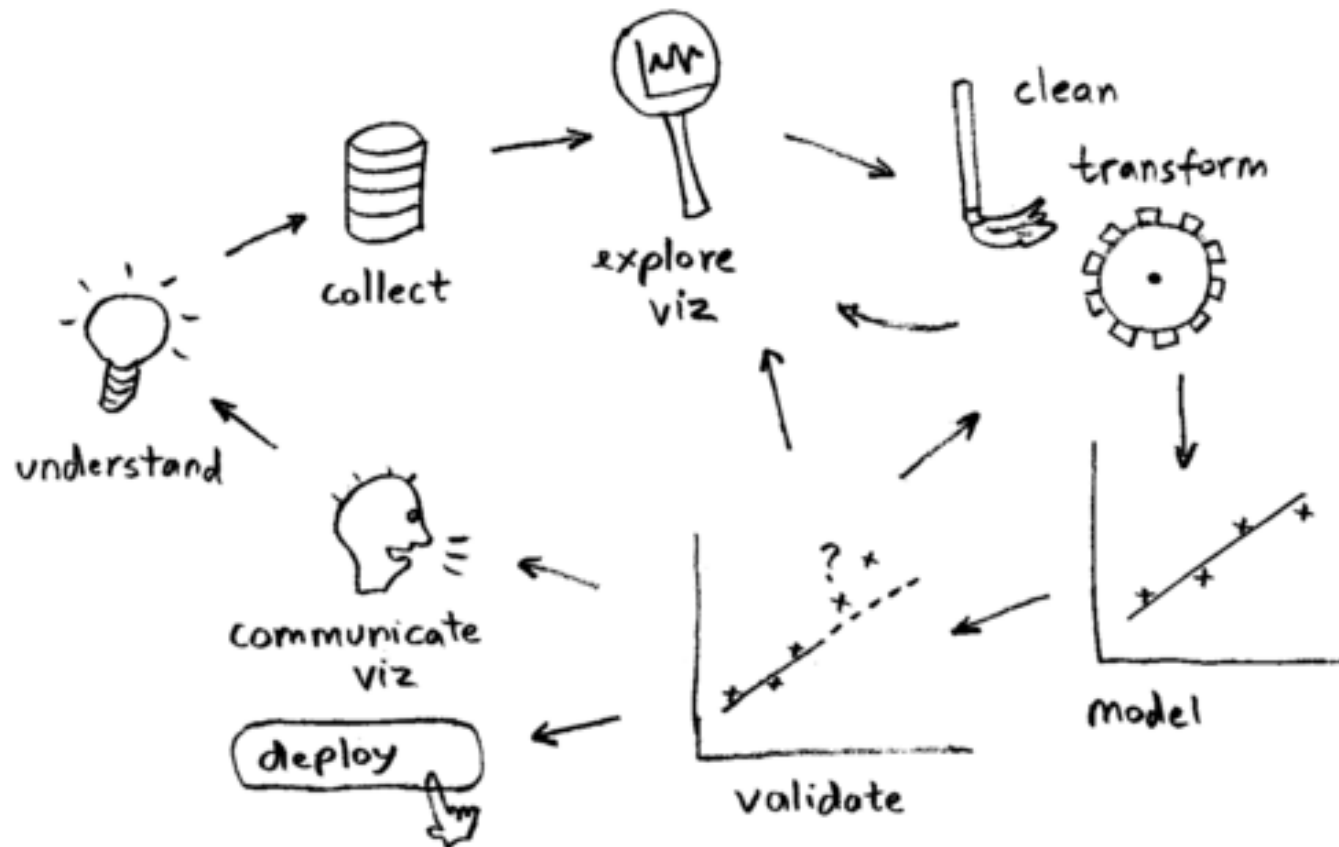
Objective:

- Production of beliefs informed by data and to be used as the basis of decision making
- N.B. In the absence of data, beliefs are uninformed and decisions are based on best practices or intuition

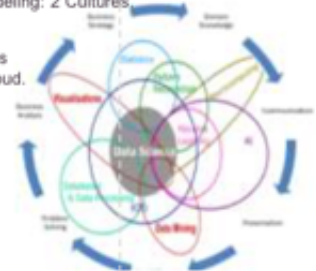
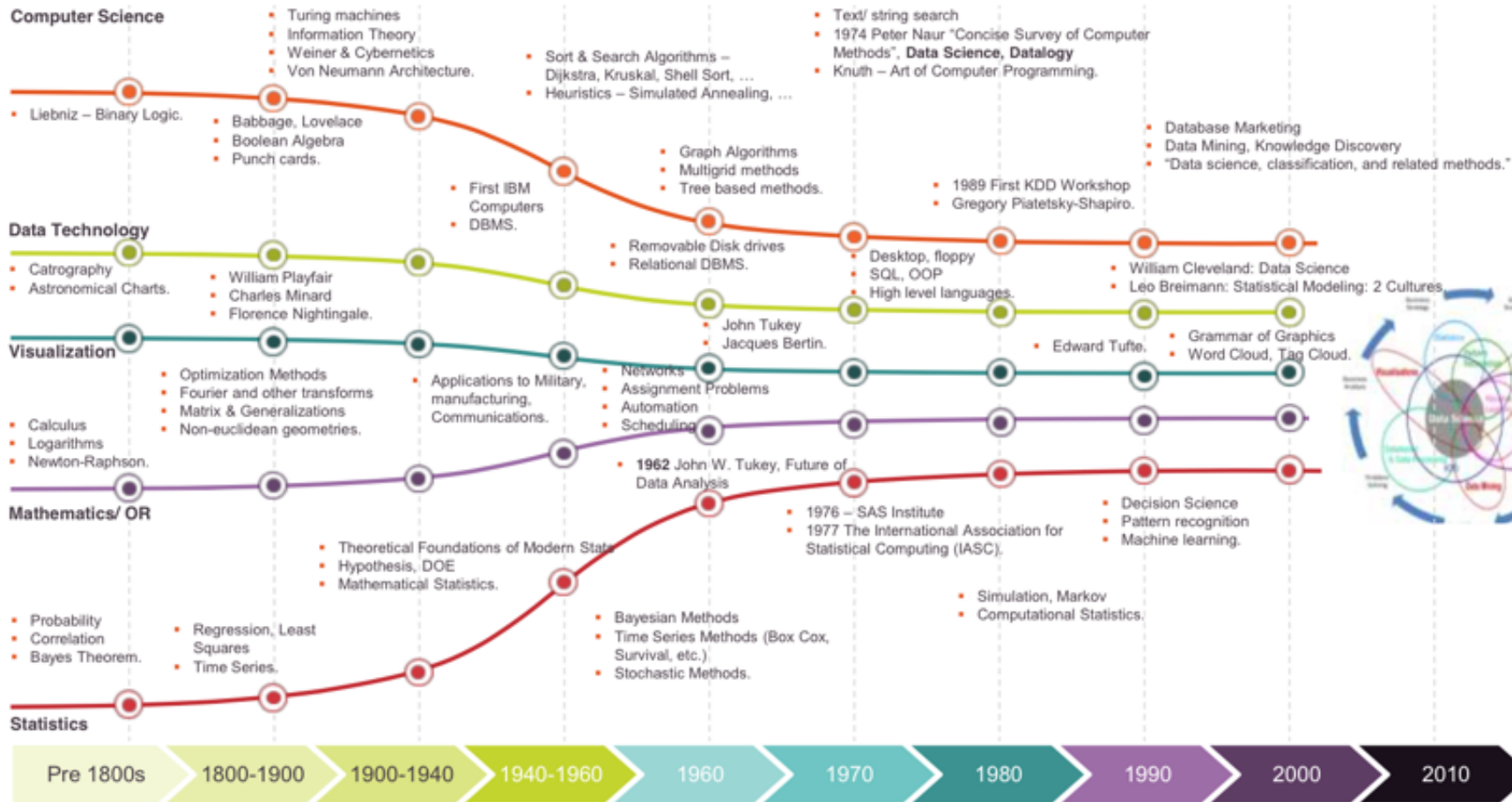
DATA SCIENCE STRATEGIES

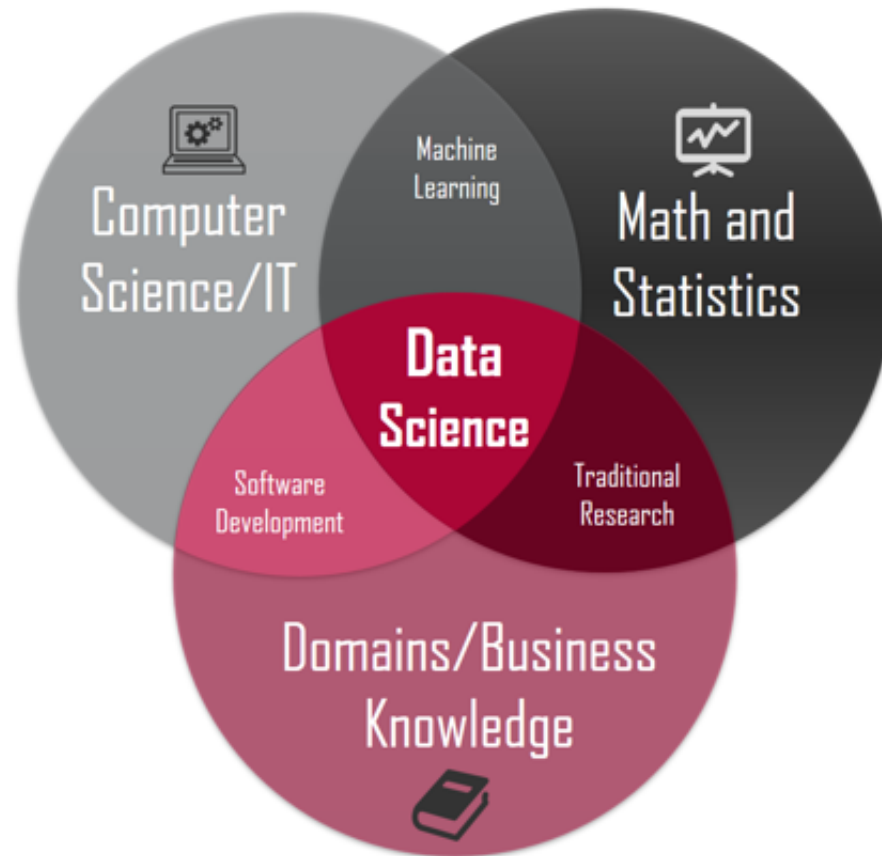
- **Probing reality:** data can be gathered by passive or by active methods. In the latter case, data represents the response of the world to our actions. *What is the best button size and colour? The best answer can only be found by probing the world.*
- **Pattern discovery:** datified problems can be analysed automatically to discover useful patterns and natural clusters that can greatly simplify their solutions. *For example, profile users is a critical ingredient in important fields as programmatic advertising or digital marketing.*
- **Predicting events:** predictive analytics allows decisions to be taken in response to future events. *For example, optimize the tasks planned for retail store staff during the following week, by analysing data such as weather, historic sales, traffic conditions, etc.*
- **Understanding people in the world:** understanding natural language, computer vision, psychology and neuroscience. In order to make optimal decisions, it is necessary to know the real processes that drive people's decisions and behaviour.

DATA SCIENCE WORKFLOW









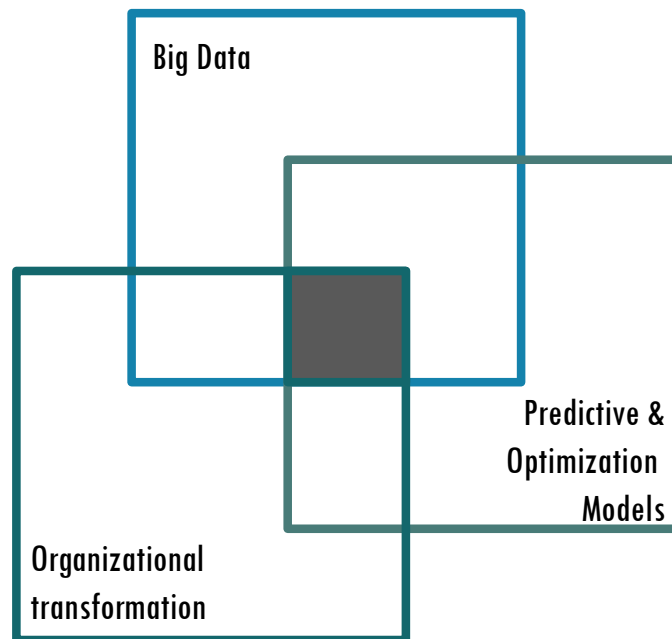
PRINCIPLE

Given lots of data

Discover patterns and models that are:

- **Valid:** hold on new data with some certainty
- **Useful:** should be possible to act on the item
- **Unexpected:** non-obvious to the system
- **Understandable:** humans should be able to interpret the pattern

CAPTURING VALUE FROM ADVANCED ANALYTICS

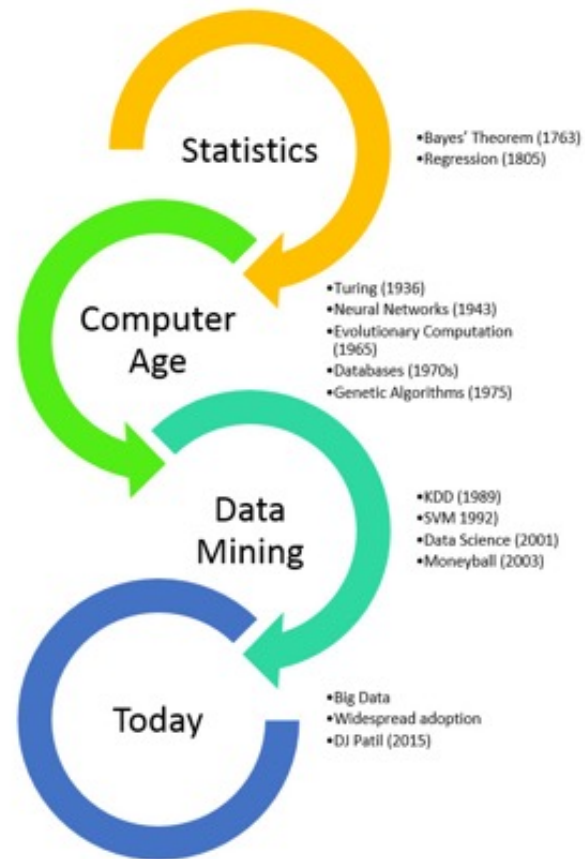


Based on three guiding principles

Decision backwards

Step by step

Test and learn



STATISTICS

Concepts:

- *Population*: collection of objects, items (“units”) about which information is sought.
- *Sample*: a part of the population that is observed.

Descriptive statistics: simplify large amounts of data in a sensible way presenting quantitative descriptions in a manageable way. It is a way to describe data.

- Applies the concepts, measures, and terms that are used to describe the basic features of the samples in a study.
- These procedures are essential to provide summaries about the samples as an approximation of the population.
- Together with simple graphics, they form the basis of every quantitative analysis of data.

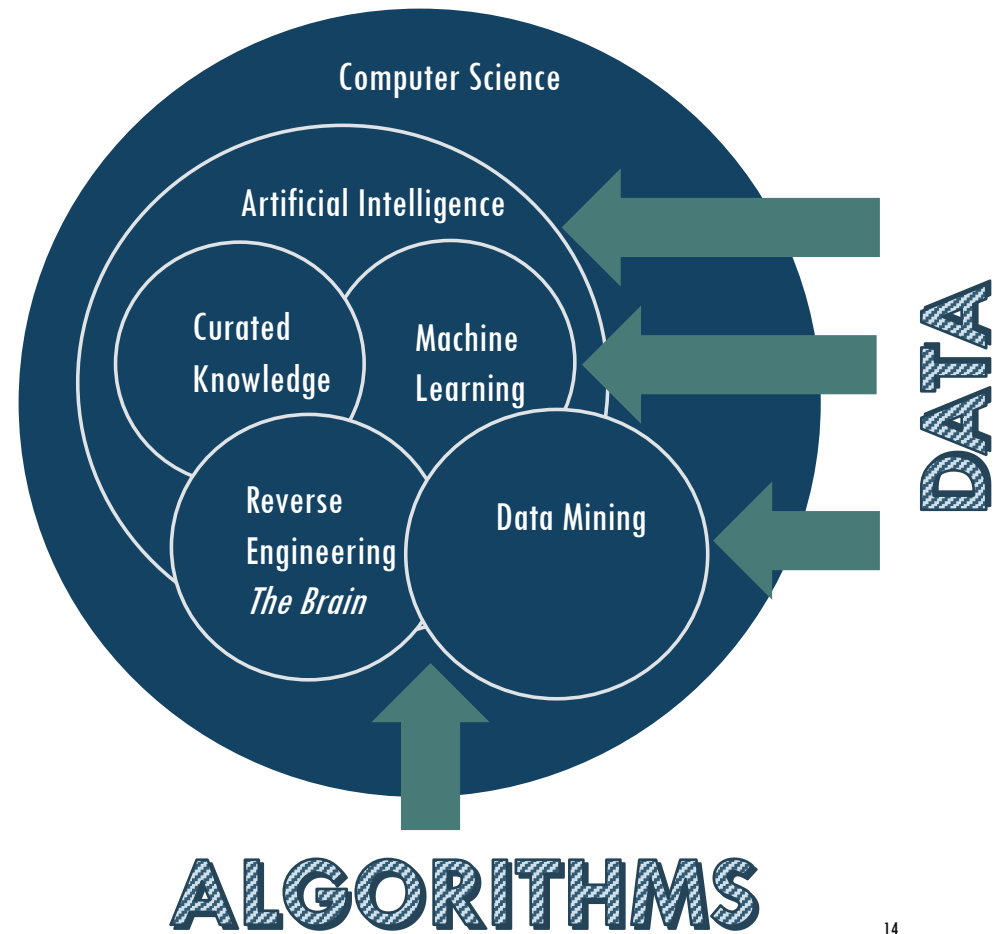
Inferential statistics: draws conclusions beyond the analysed data ; reaches conclusions regarding made hypotheses; aims at inferring characteristics of the “population” of the data.

DATA FOR STATISTICS

To describe the sample data and to infer any conclusion **data preparation** is required for generating statistically valid descriptions & conclusions

1. **Harvesting** from a file or obtained from sources (archives, data stores, sensors).
2. **Parsing** depends on what format the data are in for example plain text, fixed columns, CSV, XML, HTML, etc.
3. **Cleaning** : Survey responses and other data files are almost always incomplete. Sometimes, there are multiple codes for things such as, not asked, did not know, and declined to answer. And there are almost always errors.
4. **Building data structures**: store data in a data structure that lends itself to the analysis we are interested in.
 - If the data fit into the memory, building a data structure is usually the way to go.
 - If not, usually a database is built, which is an out-of-memory data structure.
 - Most databases provide a mapping from keys to values, so they serve as dictionaries.

IA & DATA MINING



DATA MINING CULTURES

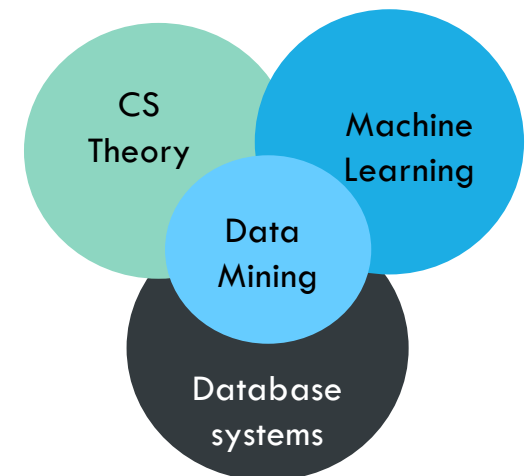
Data mining overlaps with:

- **Databases:** Large-scale data, simple queries
- **Machine learning:** Small data, Complex models
- **CS Theory:** (Randomized) Algorithms

Different cultures:

- To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
 - Result is the query answer
- To a ML person, data-mining is the **inference of models**
 - Result is the parameters of the model

In this class we will do both!



regression

Ordinary Least Squares Regression (OLSR)
Linear Regression
Logistic Regression
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Jackknife Regression

bayesian

Naive Bayes
Gaussian Naive Bayes
Multinomial Naive Bayes
Averaged One-Dependence Estimators (AOOE)
Bayesian Belief Network (BBN)
Bayesian Network (BN)
Hidden Markov Models
Conditional random fields (CRFs)

deep learning

Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders

regularization

Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least-Angle Regression (LARS))

decision tree

Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5 and C5.0 (different versions of a powerful approach)
CHI-squared Automatic Interaction Detection (CHAID)
Decision Stump
M5
Random Forests
Conditional Decision Trees

ensemble

Logit Boost (Boosting)
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (blending)
Gradient Boosting Machines (GBM)
Gradient Boosted Regression Trees (GBRT)
Random Forest

instance based

also called **case-based, memory-based**

k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)

clustering

Single-linkage clustering
k-Means
k-Medians
Expectation Maximisation (EM)
Hierarchical Clustering
Fuzzy clustering
DBSCAN
OPTICS algorithm
Non Negative Matrix Factorization
Latent Dirichlet allocation (LDA)

associated rule

Apriori
Eclat
FP-Growth

dimensionality reduction

Principal Component Analysis (PCA)
Principal Component Regression (PCR)
Partial Least Squares Regression (PLSR)
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Discriminant Analysis (LDA, MDA, QDA, FDA)

neural networks

Self Organizing Map
Perceptron
Back-Propagation
Hopfield Network
Radial Basis Function Network (RBFN)
Backpropagation
Autoencoders
Hopfield networks
Boltzmann machines
Restricted Boltzmann Machines

...and others

Support Vector Machines (SVM)
Evolutionary Algorithms
Inductive Logic Programming (ILP)
Reinforcement Learning (Q-Learning, Temporal Difference, State-Action-Reward-State-Action (SARSA))
ANCOVA
Information Fuzzy Network (IFN)
Bayes Rule



PROCESSING DATA COLLECTIONS

How many accidents are reported per day?

Percentage of use of available bicycles in downtown?



Is the number of car accidents related to seasons?

Which are the traffic bottle-neck regions in the city?

How will the use of bicycles will evolve in downtown during the summer of the next 5 years?

What type of cars are those that have more accidents in the highspeed roads?

Will increasing the parking cost reduce car traffic in the city and increase the amount of people using public transport?

Description

Modelling
(LU, QR, PCA=SVD, PARAFAC)

Association

Clustering
(Bayesian, hierarchical, k-means, CLARA, PAM)

Trend prediction

Classification
(Neural network, PLSDA, KNN, decision trees)

Regression
(PLSR, PCR)



PROCESSING DATA COLLECTIONS

Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program [ECCS 2009]

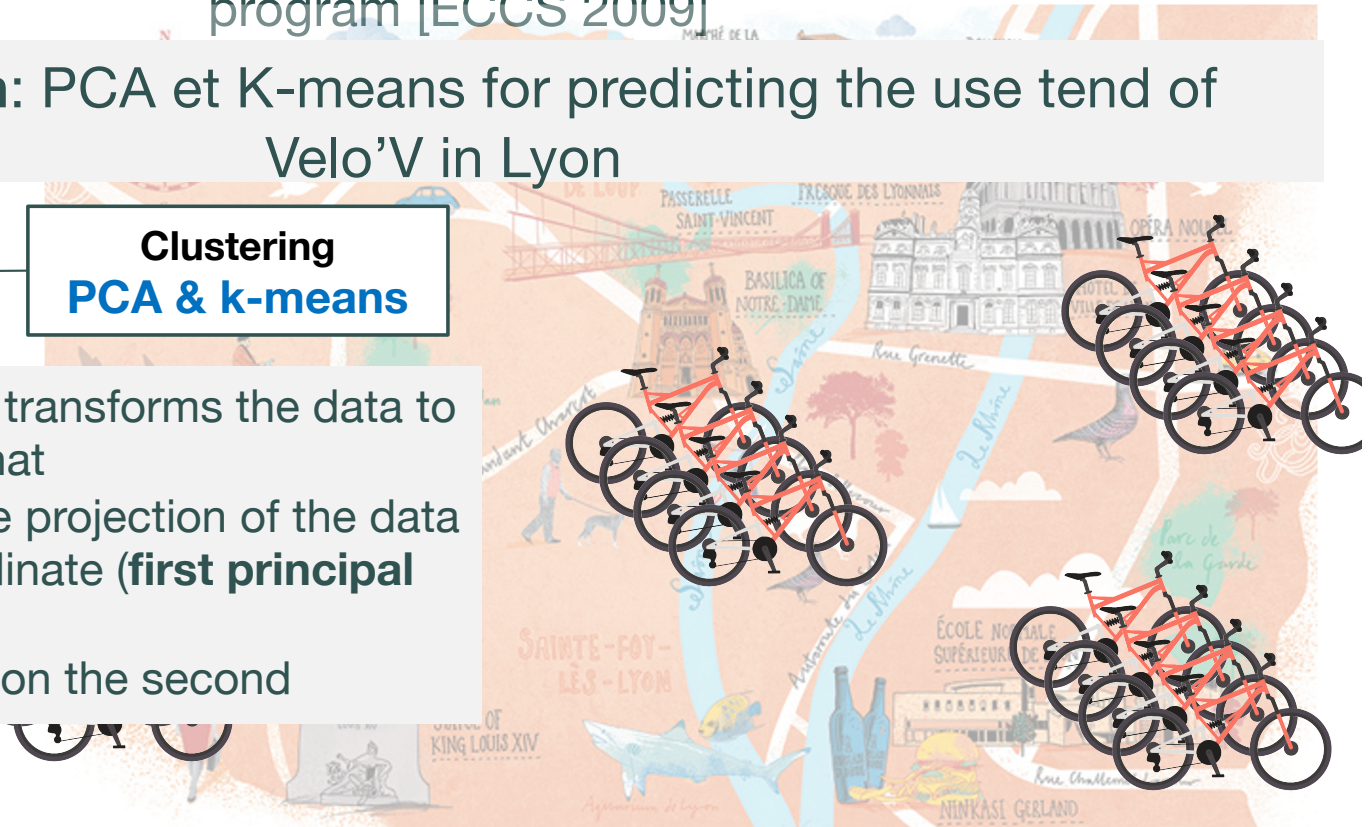
Contribution: PCA et K-means for predicting the use tend of Velo'V in Lyon

Description

Clustering
PCA & k-means

Orthogonal linear transformation transforms the data to a new coordinate system such that

- the greatest variance by some projection of the data comes to lie on the first coordinate (**first principal component**),
- the second greatest variance on the second coordinate, and so on.





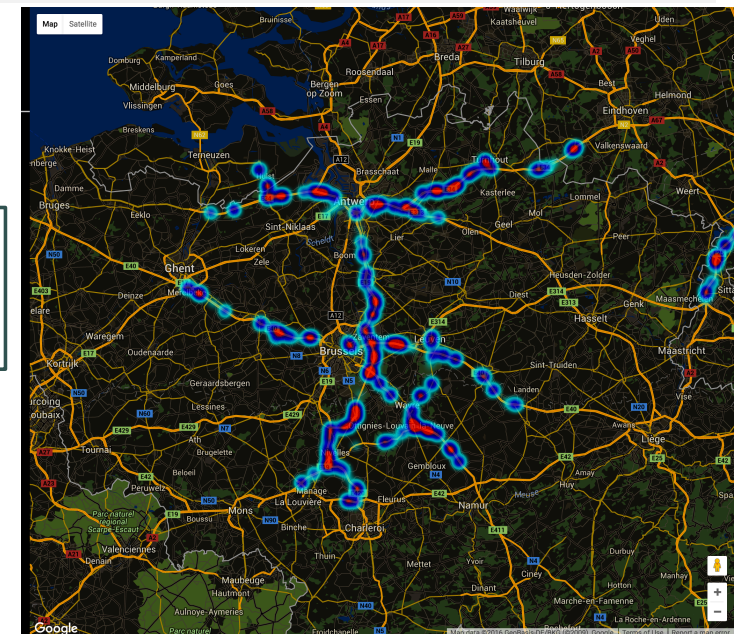
PROCESSING DATA COLLECTIONS

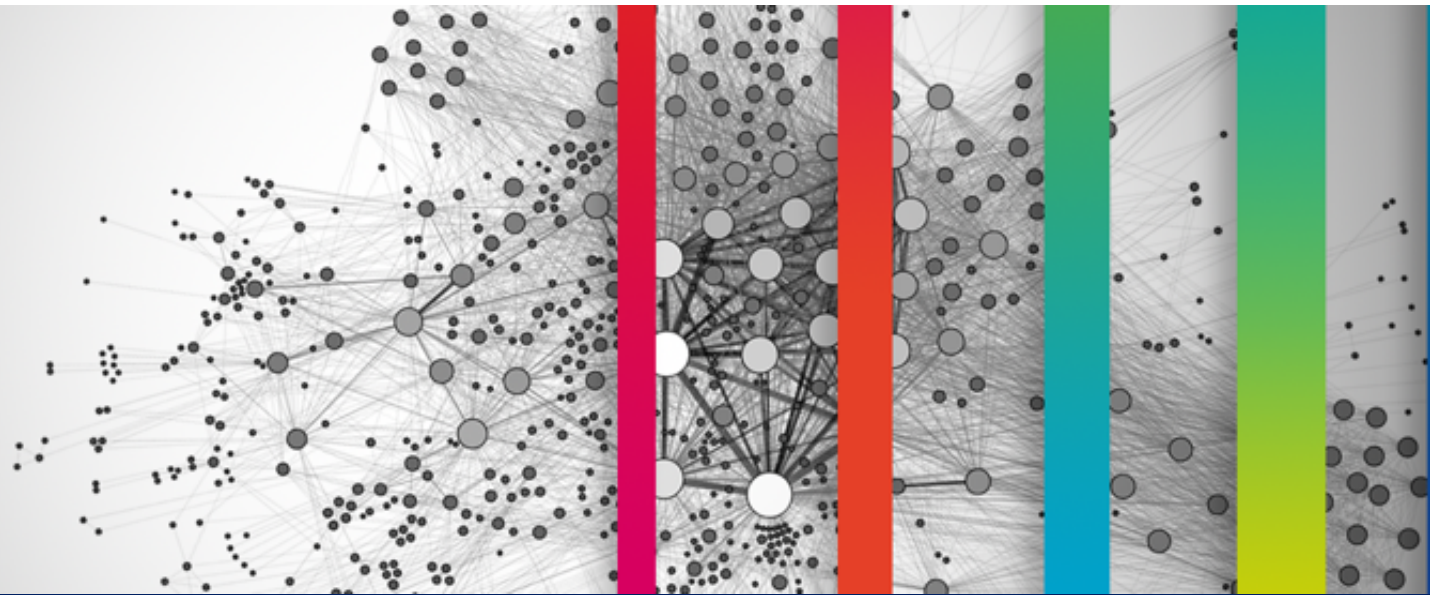
Inferring the Root Cause in Road Traffic Anomalies [IEEE Data Mining 2012]

Contribution: PCA for identifying traffic anomalies and thereby detecting problems in roads

Description

Modelling
Principal Component
Analysis (PCA)

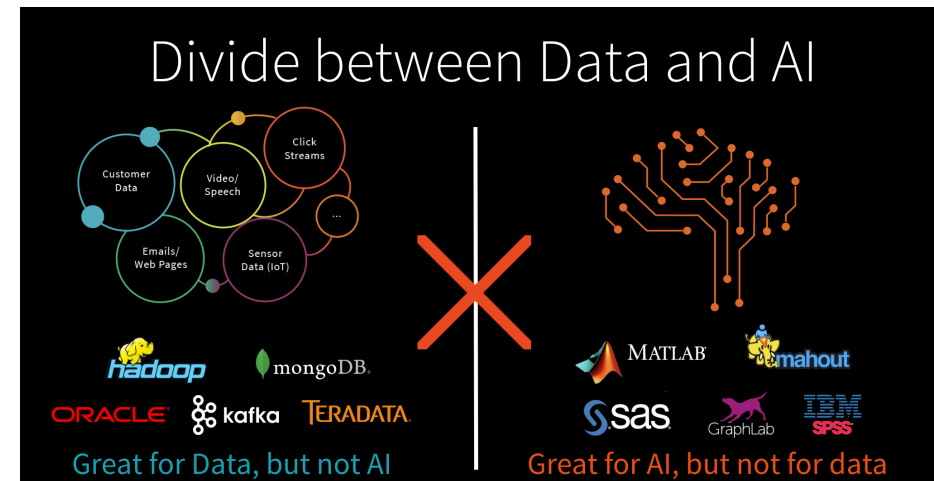




DATA SCIENCE TOOLBOXES

DATA CONSUMPTION PHILOSOPHIES

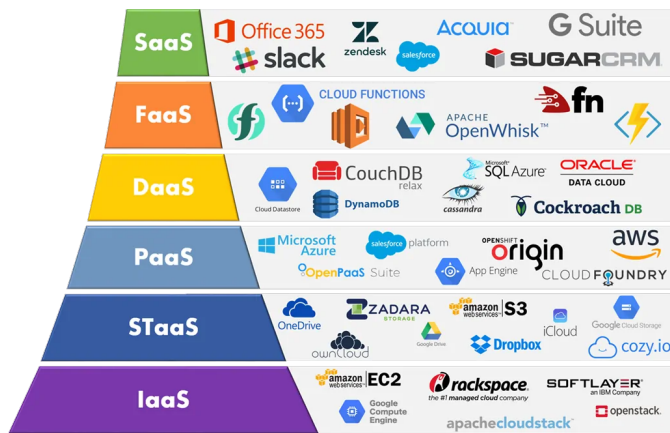
- Data loading
- In memory/cache/disk indexing
- Data persistence
- Query optimization
- Concurrent access
- Consistency and access control



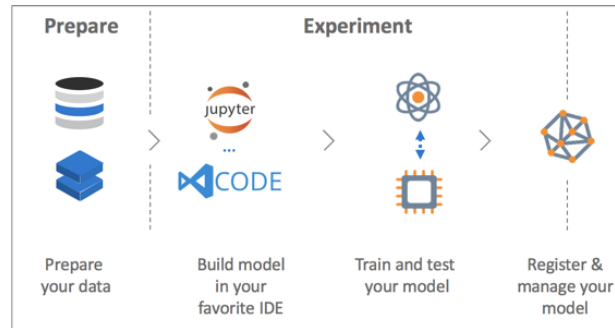
Functions must be revisited under less strong hypothesis to support the enactment of data science pipelines

ENACTMENT ENVIRONMENTS

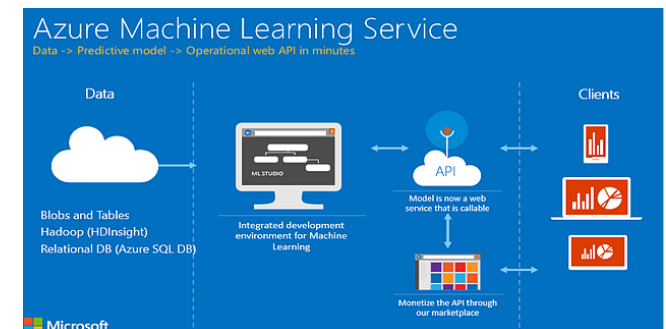
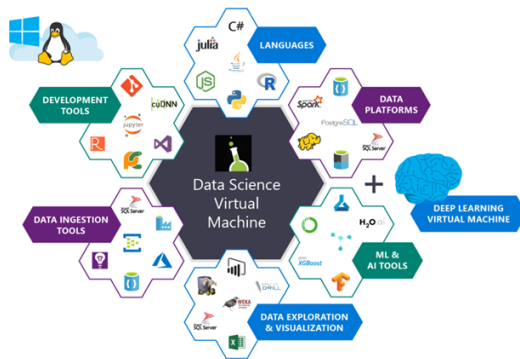
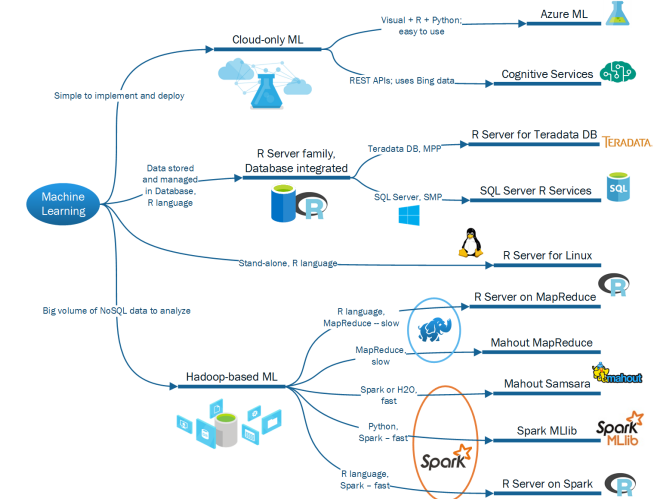
● Big Data Platforms & Stacks



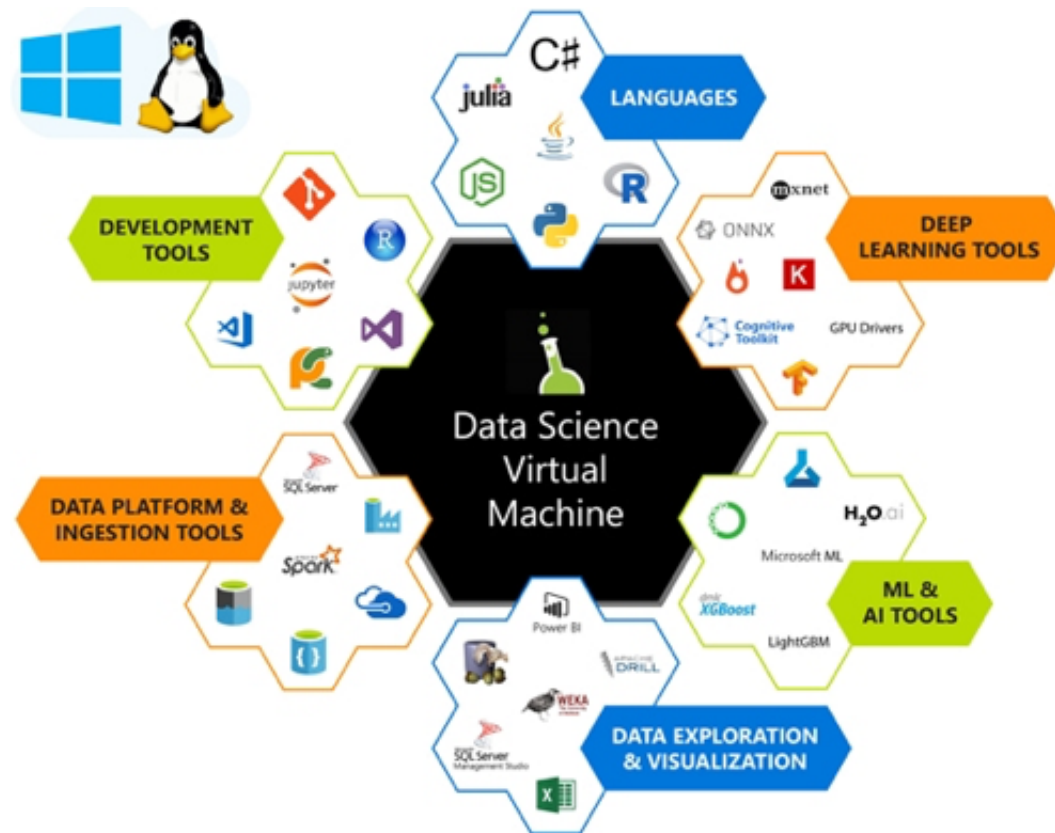
● WIDE environments



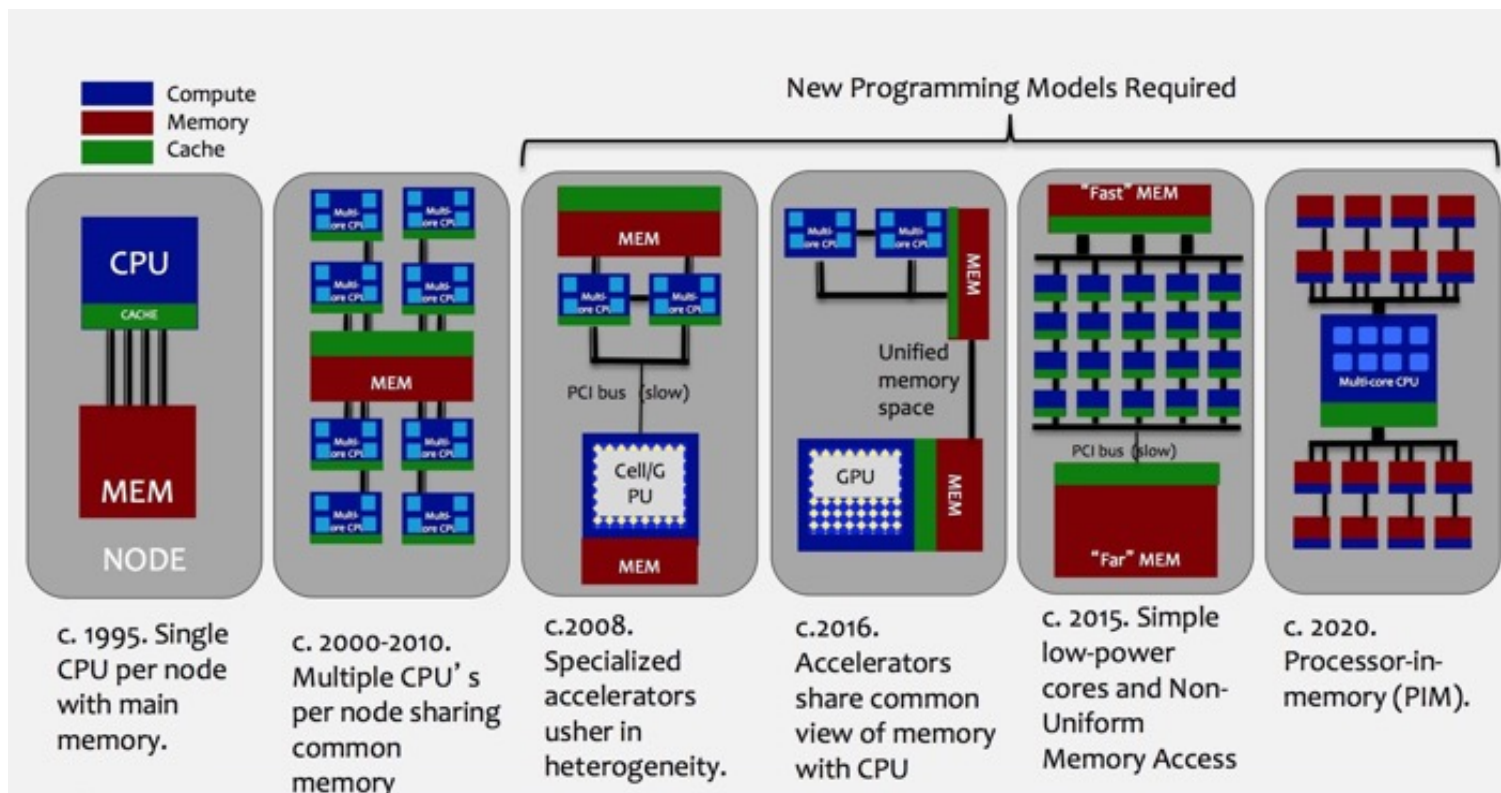
● Machine Learning Services



DATA SCIENCE VIRTUAL MACHINE



COMPUTING CAPACITY



WHICH TOOLS FOR DATA SCIENCE

[1]: data
> science
\$ toolbox

Programming language:

- Python one of the most flexible programming languages because it can be seen as a multiparadigm language
- Alternatives are MATLAB and R

Fundamental libraries for data scientists in Python: NumPy, SciPy, Pandas and Scikit-Learn

- *NumPy* provides, support for multidimensional arrays with basic operations on them and useful linear algebra functions.
- *SciPy* provides a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics.
- *Matplotlib* tools for data visualization
- *Sci-kit-Learn*: machine learning library with tools such as classification, regression, clustering, dimensionality reduction, model selection, and preprocessing.
- *Pandas*: provides high-performance data structures and data analysis tools for data manipulation with integrated indexing

DATA SCIENCE ECOSYSTEM & INTEGRATED DEVELOPMENT ENVIRONMENT

To get started on solving data-oriented problems, we need to set up our programming environment

Decide programming language version, whether to install a **data scientist ecosystem** by **individual toolboxes**, or to perform a **bundle installation**

- For example, Anaconda Python provides integration of all the Python toolboxes and applications for data scientists in a single directory

The **integrated development environment (IDE)** is an essential tool designed to maximize programmer productivity.

- The basic pieces of any **IDE** are three: the editor, the compiler, (or interpreter) and the debugger.
- Examples: PyCharm,⁹ WingIDE¹⁰, SPYDER (Scientific Python Development EnviRonment)



WEB INTEGRATED DEVELOPMENT ENVIRONMENT

Web-based IDEs were developed considering how not only your code but also all your environment and executions can be stored in a server.

- The server can be set up in a centre, such as a university or school
- Students can work on their homework either in the classroom or at home
- Students can execute all the previous steps over and over again, and then change some particular *code cell* (a segment of the document that may contain source code that can be executed) and execute the operation again.

For example IPython has been issued as a browser version of its interactive console: Jupyter

- Markdown (a wiki text language) **cells** can be added to introduce algorithms.
- It is also possible to insert Matplotlib graphics to illustrate examples or even web pages.
- Experiments can become completely and absolutely replicable.

WEB INTEGRATED DEVELOPMENT ENVIRONMENT

jupyter spectrogram (autosaved) Python 3

File Edit View Insert Cell Kernel Help

Markdown CellToolbar

Simple spectral analysis

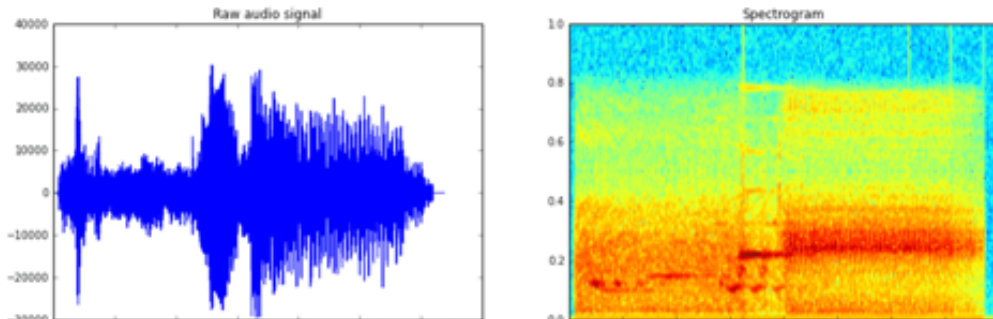
An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n \exp^{-\frac{j2\pi}{N}kn} \quad k = 0, \dots, N-1$$

```
In [2]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



The image shows two plots side-by-side. The left plot, titled 'Raw audio signal', is a time-domain waveform plot with a y-axis ranging from -30000 to 40000. The right plot, titled 'Spectrogram', is a frequency-time plot with a y-axis ranging from 0.0 to 1.0. The spectrogram shows a color-coded intensity of frequencies over time, with a prominent horizontal band of high intensity around 0.5 on the y-axis.