RIGACODINGSCHOOL

Machine Learning
with Python Anaconda Data Science Distribution

python™

ANACONDA®

VALDIS SAULESPURĒNS

- Izglītība: Maģistra grāds datorzinātnēs
- Pieredze programmēšanā: 20+ gadi
- Pythons + Datu Analīze: 10+ gadus
- Specialitāte: grafu teorija sociālo tīklu analizēšanā
- Hobiji: prāta spēles, riteņbraukšana, šahs

Contact: valdis.s.coding@gmail.com

Data Lake or Data Swamp?

Buzzword bingo

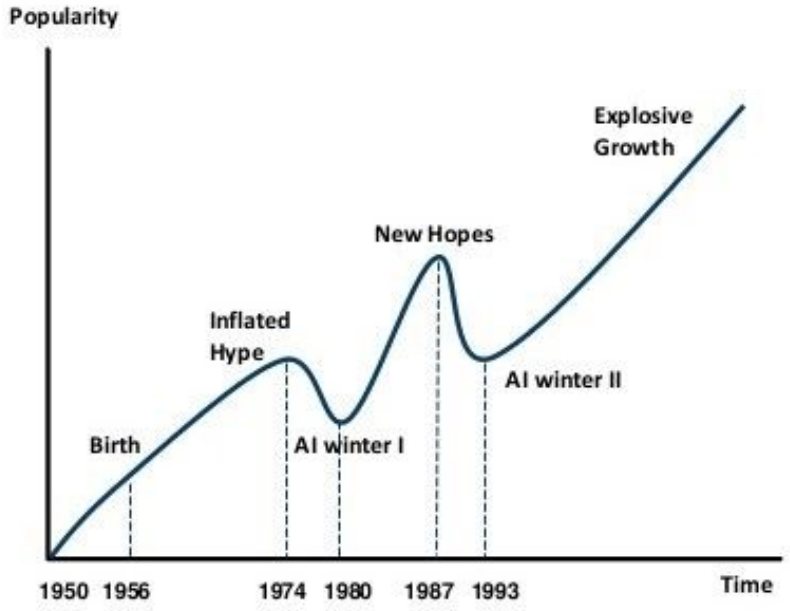
- Big Data
- Data Mining (datizrace)
- Machine Learning – subset of AI
- Data Science – statistics

- Big Data or Pokemon
- https://pixelastic.github.io/pokemonorbigdata/

Random Forests	Neural Network	Reinforcement Learning	Supervised Learning	Cognitive Computing
Caffe	Support Vector Machine	Artificial Intelligence	Python	Cloud
Unstructured Data	Bot	DATA SCIENCE BUZZWORD BINGO (free square)	K-means	GPU
Spark	Data Wrangling	Deep Learning	Ensemble	Machine Learning
Keras	Tensorflow	Big Data	Algorithm	Feature Engineering

Brief History of AI

AI HAS A LONG HISTORY OF BEING "THE NEXT BIG THING"...



Timeline of AI Development

- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM's Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world's champions

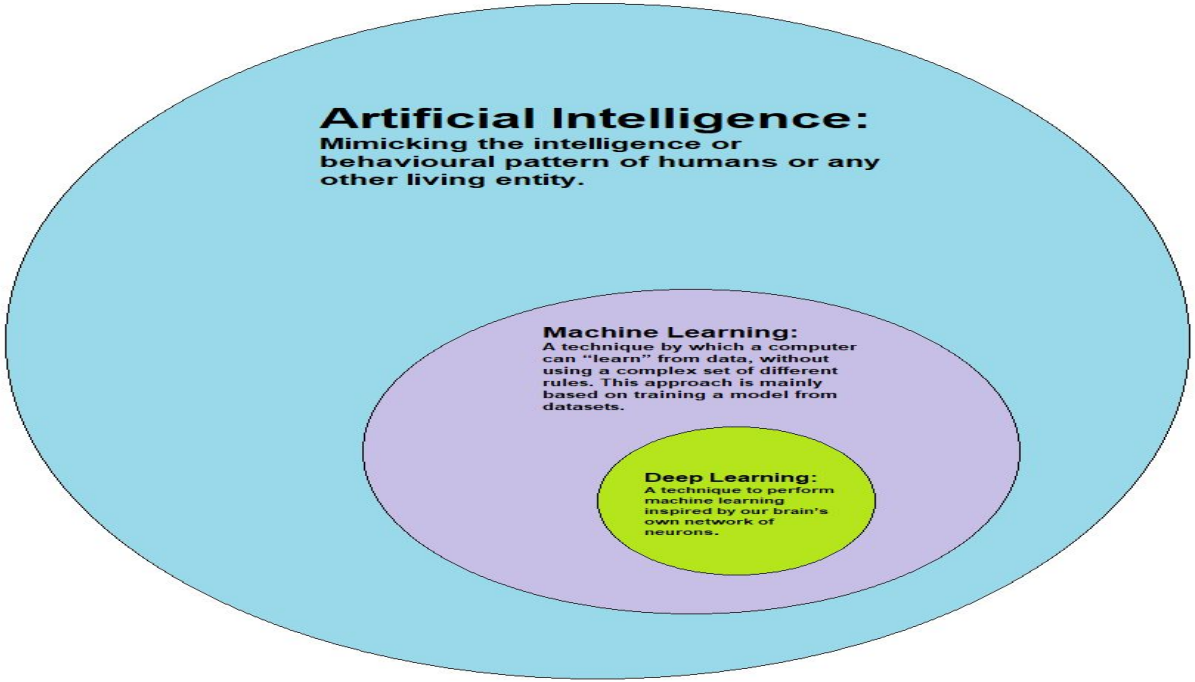
What is machine learning?

"Machine learning is the science of getting computers to act without being explicitly programmed." – Stanford ML course

"The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" – Carnegie Mellon University

"Machine learning can't get something from nothing…what it does is get more from less." – Dr. Pedro Domingo, University of Washington

AI, ML, Deep Learning



Brief History of Data Analysis

- What is this ? :)

Brief History of Data Analysis

- ~ 18,000BC – Uganda, Ishango Bone
- ~ 2400BC – Babylon abacus, libraries
- 300BC – 48AD – Library of Alexandria

Brief History of Data Analysis

- How about this modern recreation of a 2000 years old device?

Brief History of Data Analysis

- **~ 100-200AD**
  Antikythera Mechanism

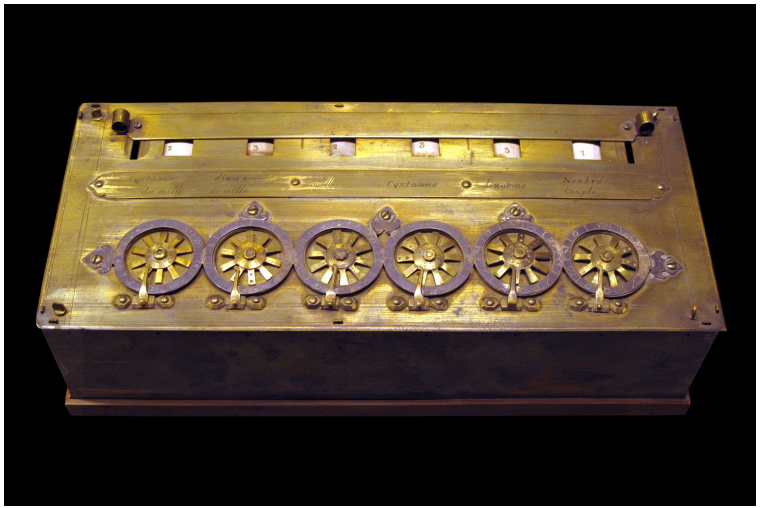
  Predicting:
  - Astrology
  - Astronomy
  - Olympics
  - Calendar

Brief History of Data Analysis

- **1642 Blaise Pascal's Pascaline**

Performs:

- Addition
- Subtraction
- Multiply/Divide using Add/Sub
- 1649 Royal Patent by Louis XIV

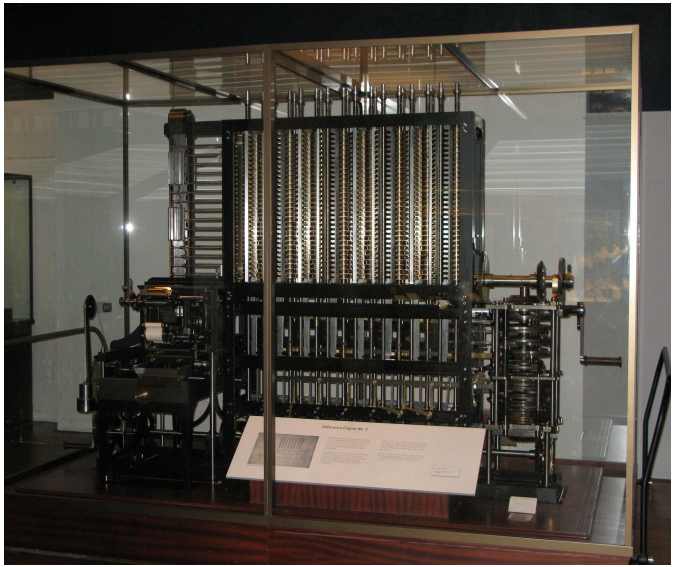


Brief History of Data Analysis

- **Charles Babbage**
  Difference Engine
- Ada Lovelace - first programmer?

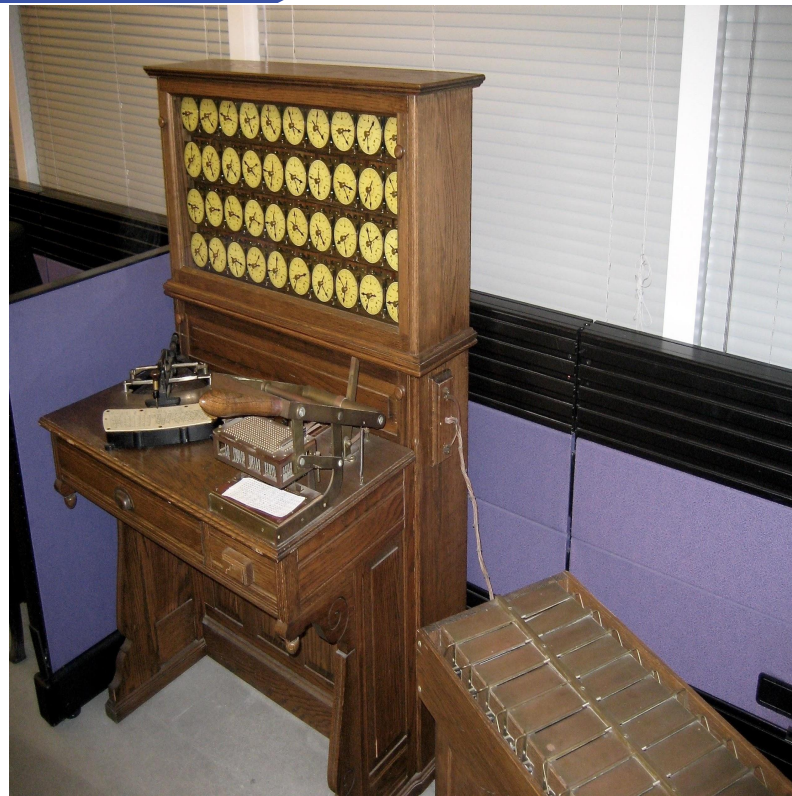
Performs:

- Arithmetic
- Derivation
- Power Series
- Curve Fitting



Brief History of Data Analysis II

- 1663 – London, J.Graunt mortality analysis
- 1865 – banker H. Furnese business intelligence
- 1880-90 US Census Hollerith Machine -> IBM
- 1928 – F. Pfleumer magnetic tape invention

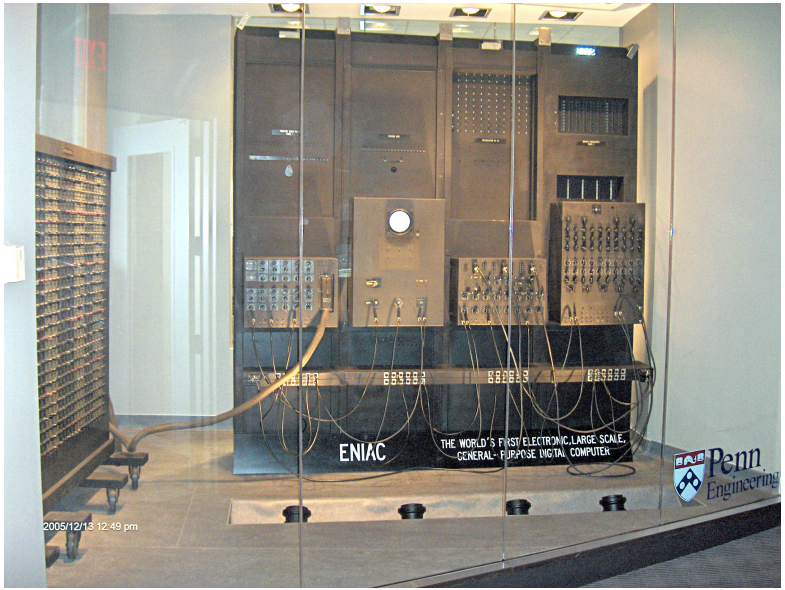


Brief History of Data Analysis

- 1940s - First General Purpose Electronic Computer ENIAC (Zuse mechanical)

- Turing complete
- von Neumann architecture

- most computers work the same today

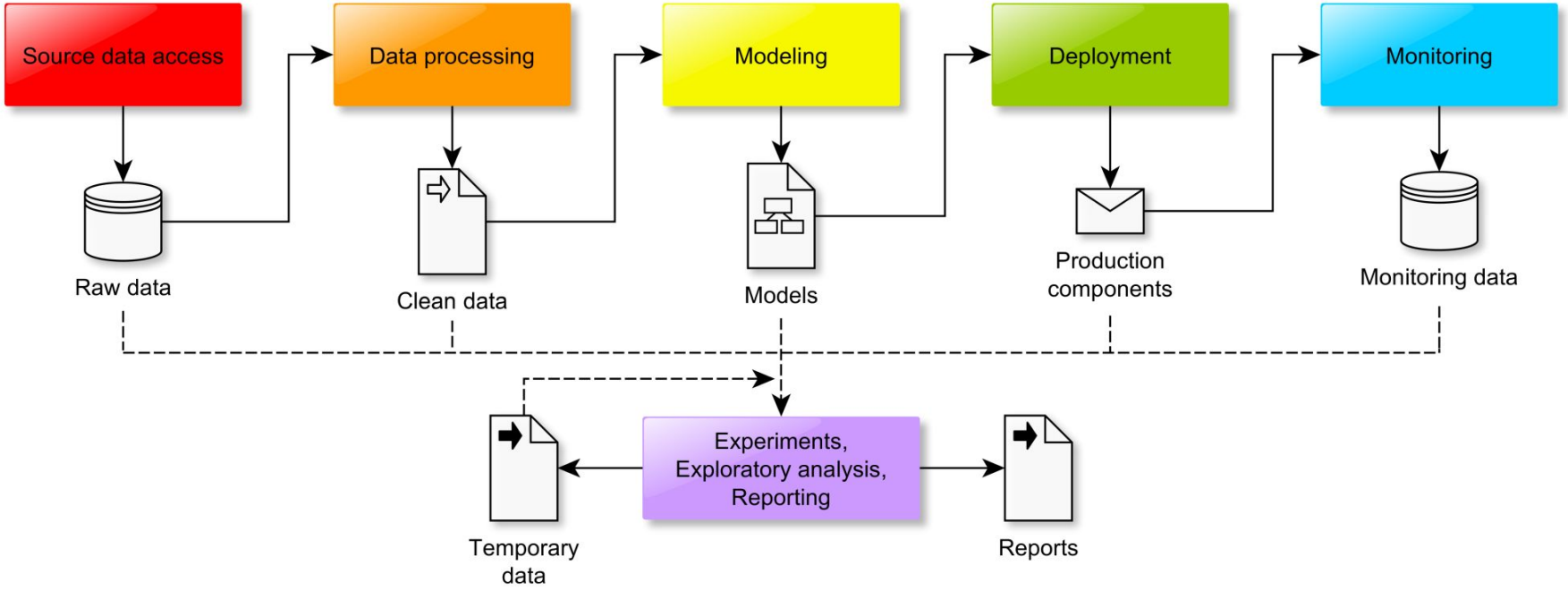


Brief History of Data Analysis

- 1950s - Flat Files
- 1958 – IBM's Luhn defines Business Intelligence
- 1960s - CODASYL
- 1970s – Codd's relational DBs -> SQL
- 1980s – Data Werehouses / Marts
- 2000s – Big Data / noSQL DBs

BIG DATA LANDSCAPE 2017

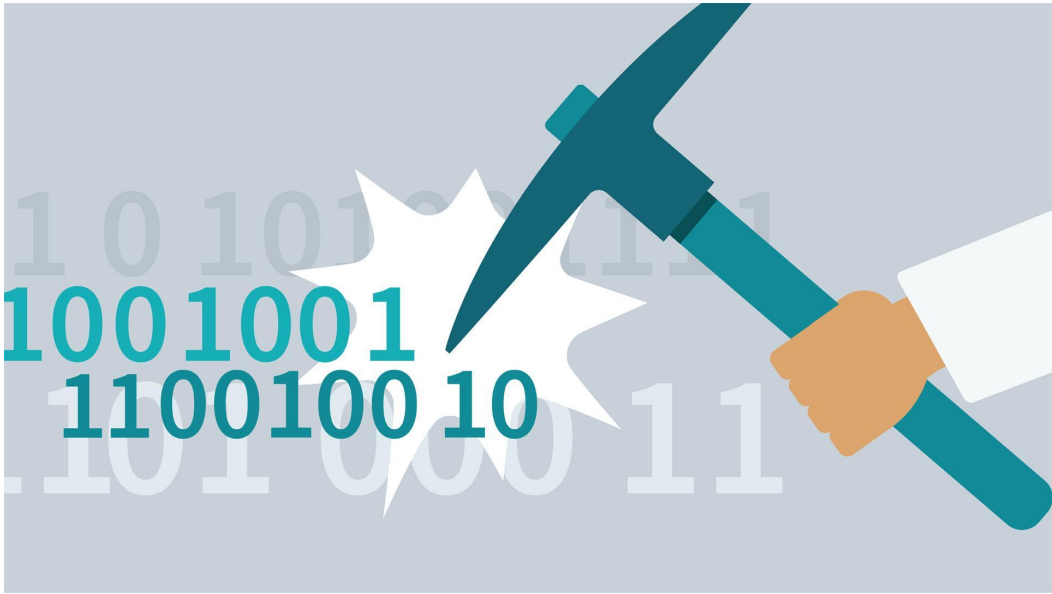
Full Analysis Framework



Data Analysis Sandbox

Data Mining

- Anomalies
- Classification
- Clusters
- Dimension Reduction
- Regression
- Relationship finding
- Summarization / Visualization

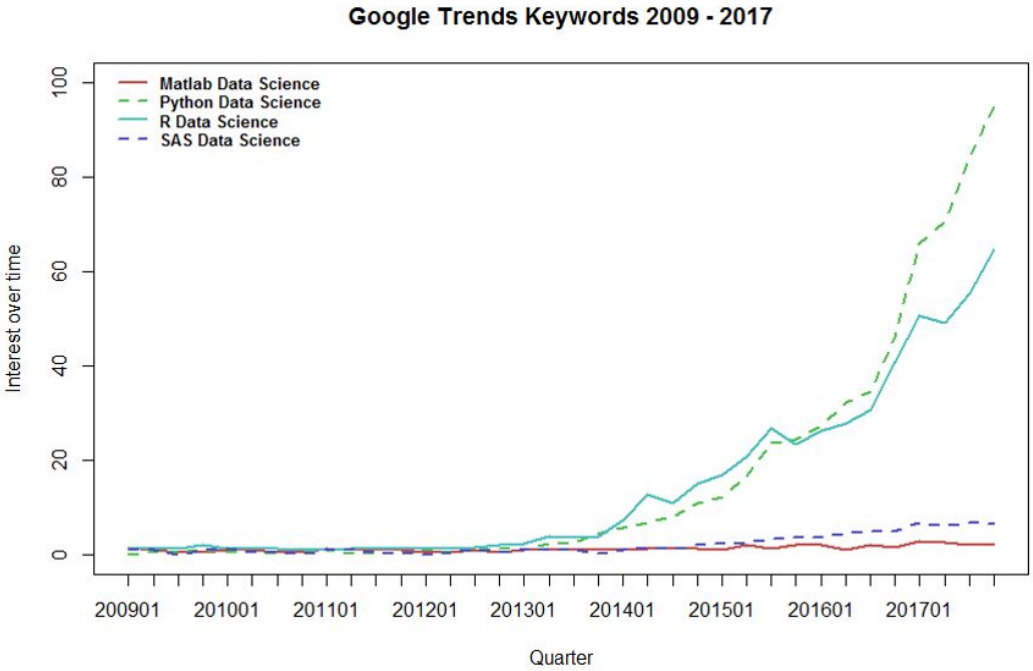


Building a Pipeline

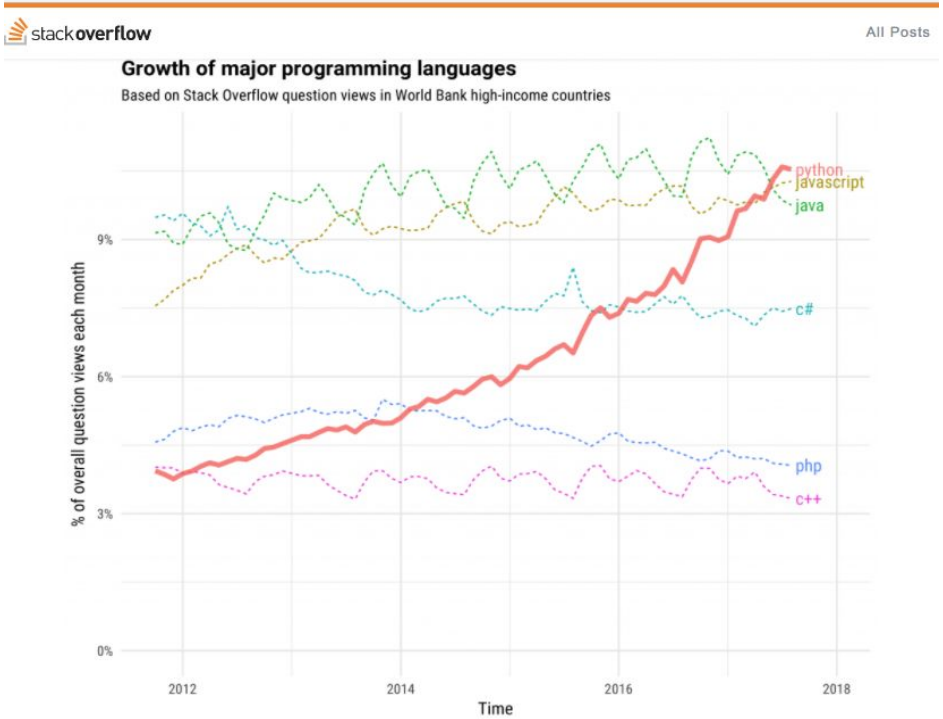
- Data Security
- Cleanup
- Organising Database
- Analysis
- Visualization – Dashboard

- Emphasis on Analysis less on Infrastructure

Why Python?


Google Trends Keywords 2009 - 2017
Matlab Data Science
Python Data Science
R Data Science
SAS Data Science
Interest over time
0
20
40
60
80
100
200901
201001
201101
201201
201301
201401
201501
201601
201701
Quarter

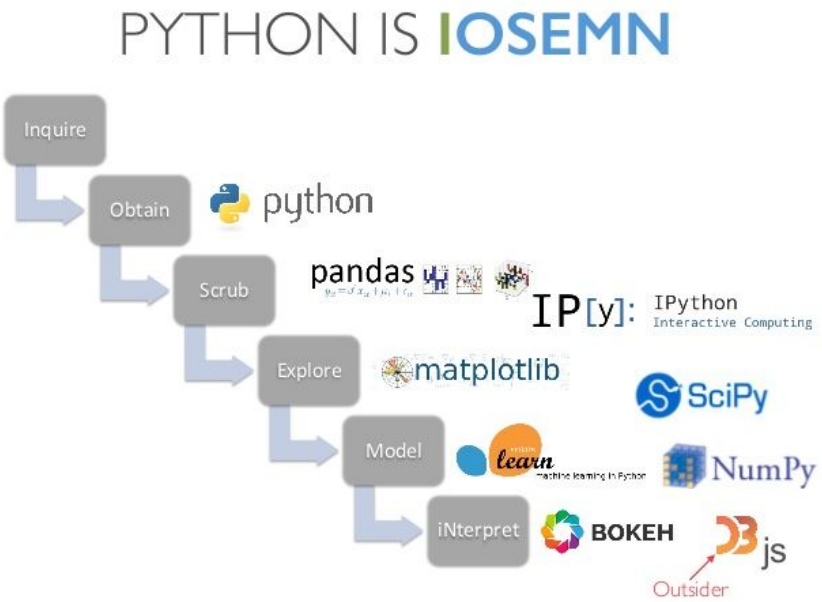
Why Python? Part 2



Why Python?

- Readable
- Interactive
- Interpreted
- Dynamic
- Object-oriented
- Portable
- High level
- Extensible in C++ & C

Python Ecosystem



Anaconda Distribution

ANACONDA POWERS THE MODERN ANALYTICS STACK

APP	Notebooks Embeddable Dashboards Data Services Visual Apps
VIZ	Plots Interactive Viz Big Data Maps & GIS 3D Streaming Graphs
STORYBOARD	Notebooks Interactive Exploration Visual Programming Data IDEs
ANALYTICS	Data Prep Stats ML & Ensembles Deep Learning Simulation & Optimization Geospatial Text & NLP Video/Image/Audio Mining Graph & Network
DATA	Hadoop & Hive Spark NoSQL DW & SQL Files & Web Services
HARDWARE	Servers Clusters GPUs & High End Workstations

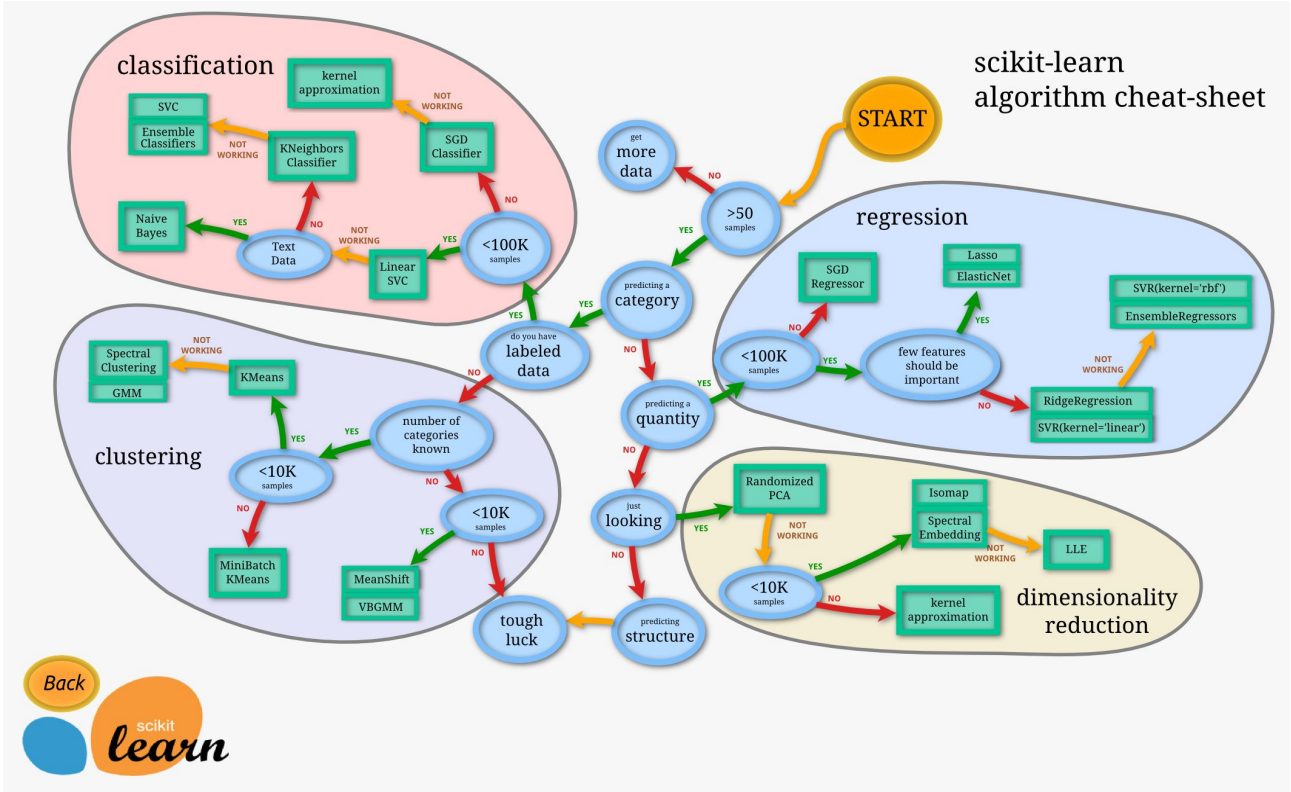
scikit-learn machine learning library

- Tools for data mining and data analysis
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license
-
- Source: https://github.com/scikit-learn/scikit-learn

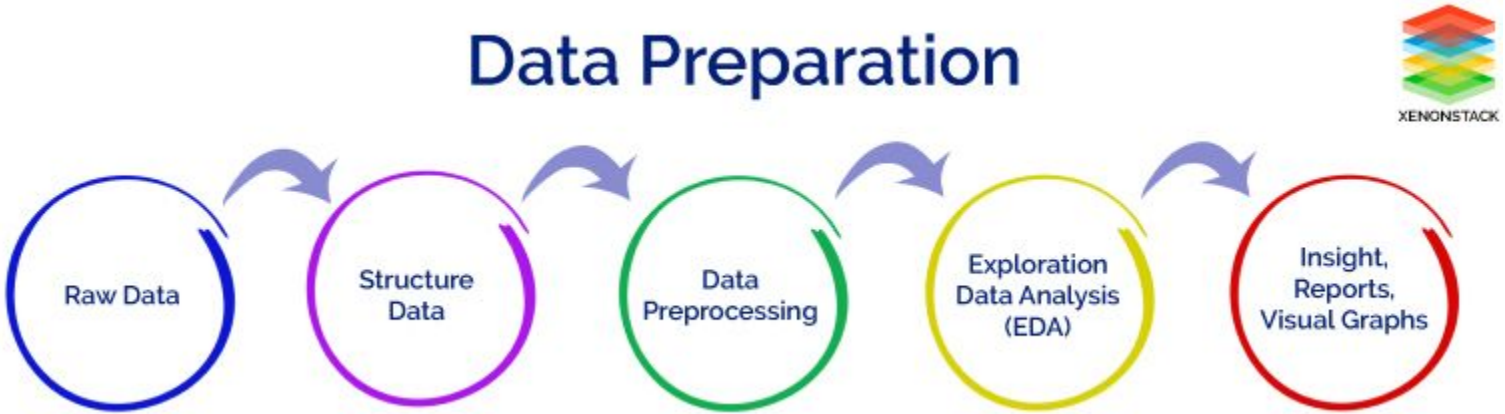
scikit-learn machine learning library

- Classification
- Regression
- Clustering
- Dimensionality Reduction
- Model Selection
- Preprocessing

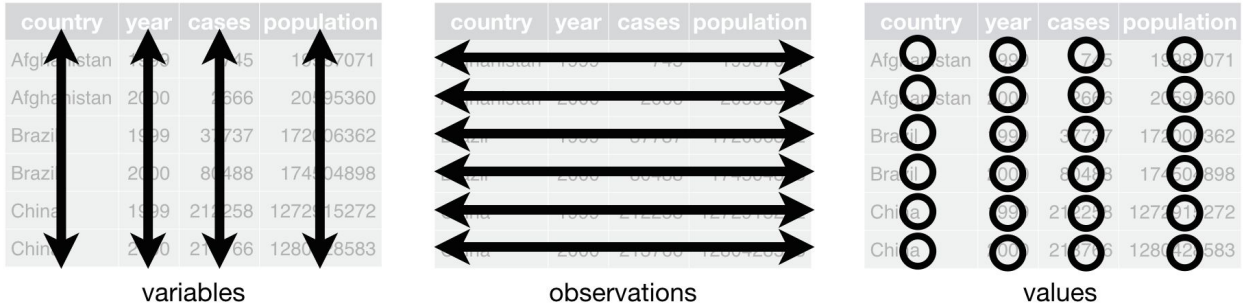
scikit-learn machine learning library



Data Cleaning / Wrangling



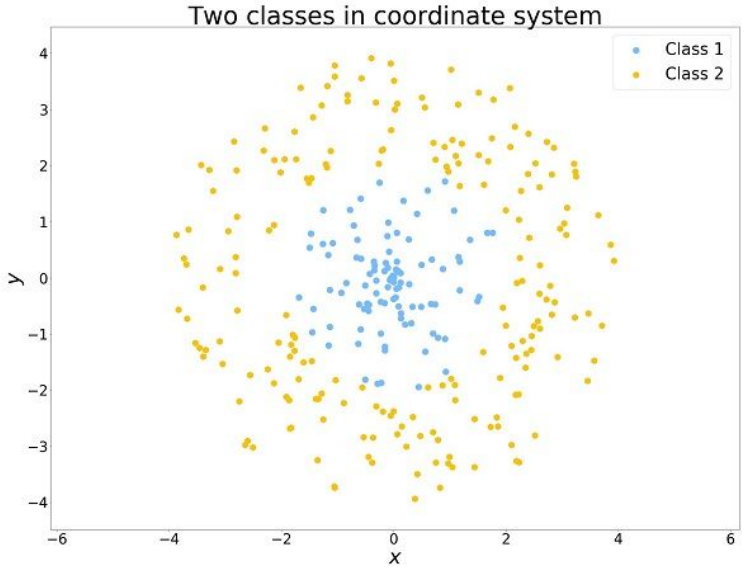
Tidy Data



1. Each variable -> column.
2. Each observation -> row.
3. Each type of obseration -> table.

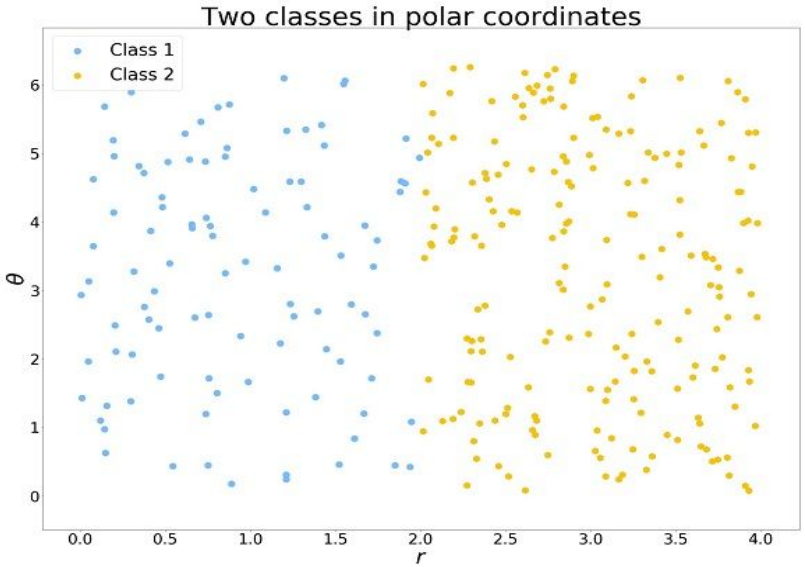
https://vita.had.co.nz/papers/tidy-data.pdf

Feature Engineering



Source: kdnuggets.com

Feature Engineering



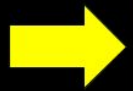
$$r = \sqrt{x^2 + y^2} \quad \theta = \arctan \frac{\tilde{y}}{x}$$

Source: kdnuggets.com

Feature Engineering

- Categorical Data -> Category Codes
- Categorical Data -> One-Hot Encoding (Only one True value)

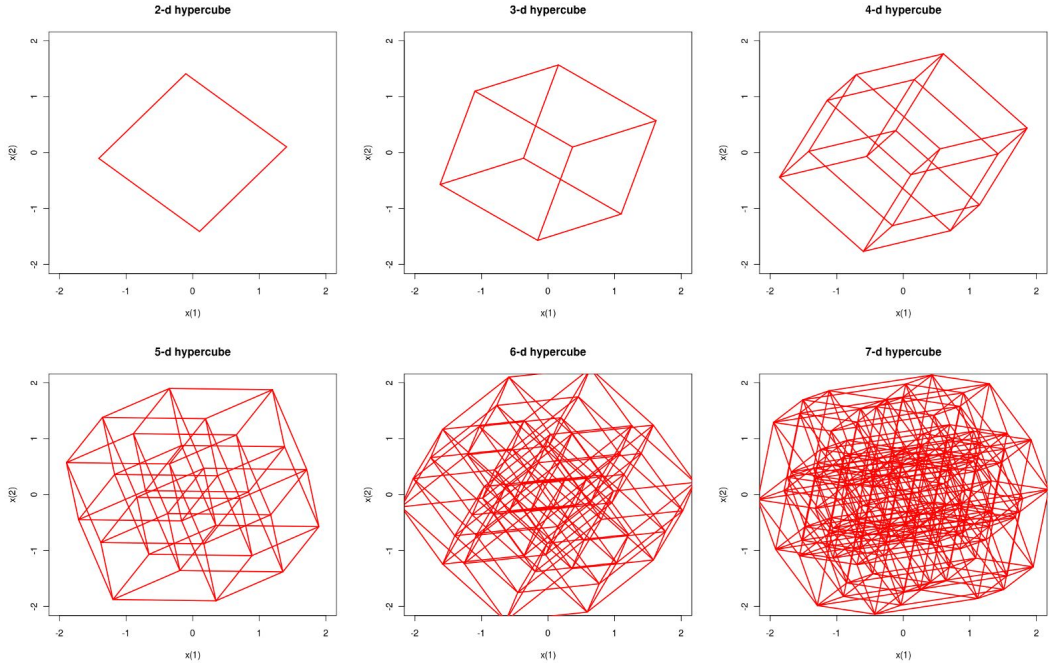
Color
Red
Red
Yellow
Green
Yellow



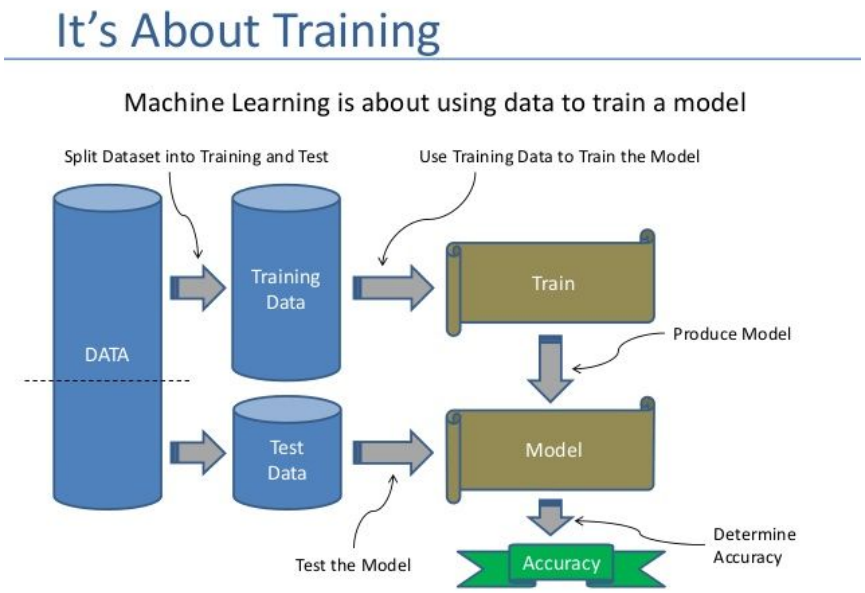
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Source: kaggle.com

PCA and Curse of Dimensionality



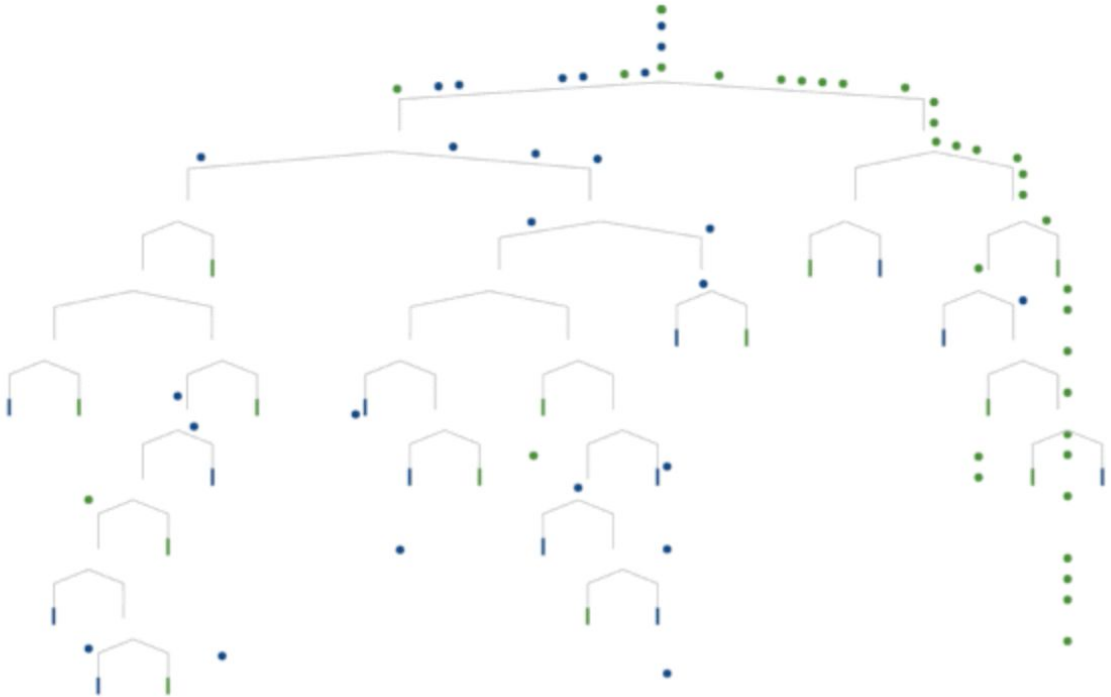
Training Data vs Testing Data



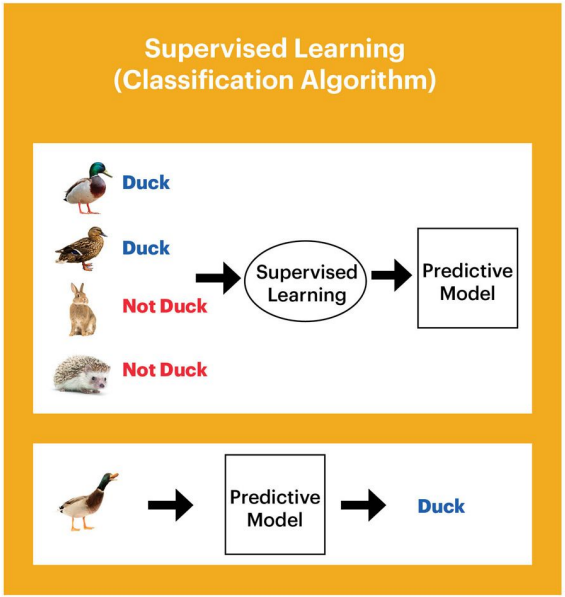
Cross Validation

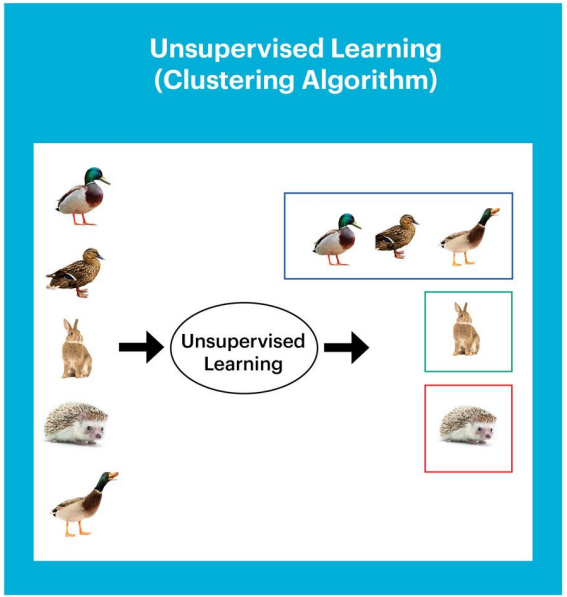
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Decision Trees Classificators



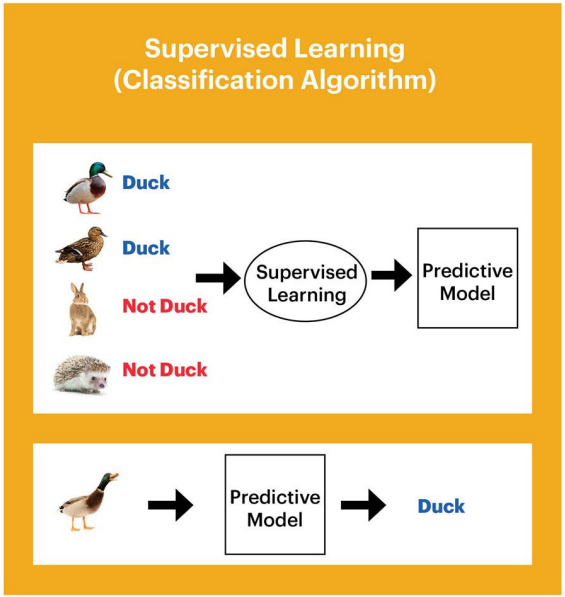
Supervised vs Unsupervised

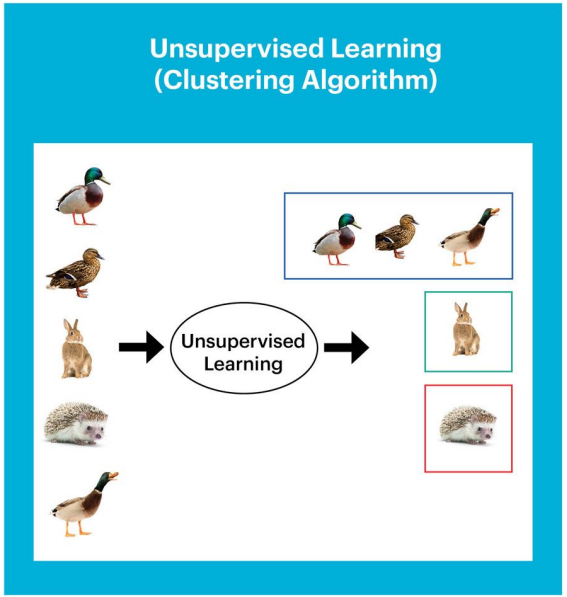




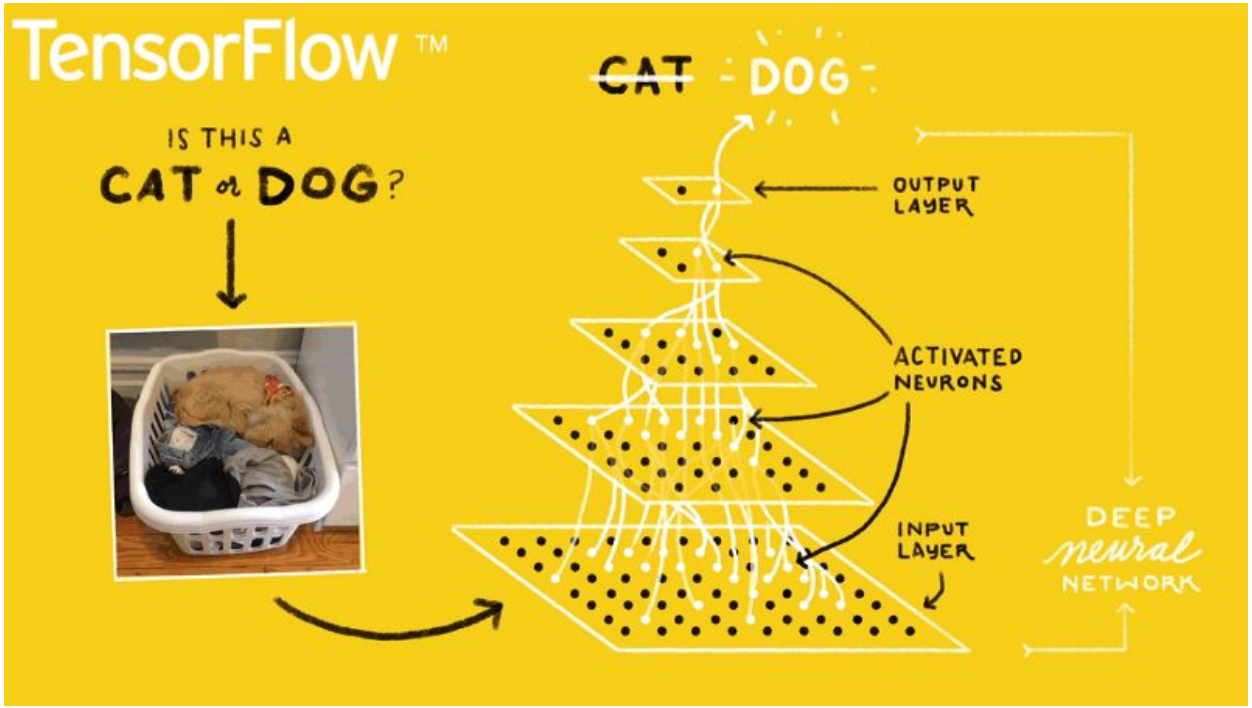
Western Digital.

Supervised vs Unsupervised





Deep Learning with Keras,TensorFlow



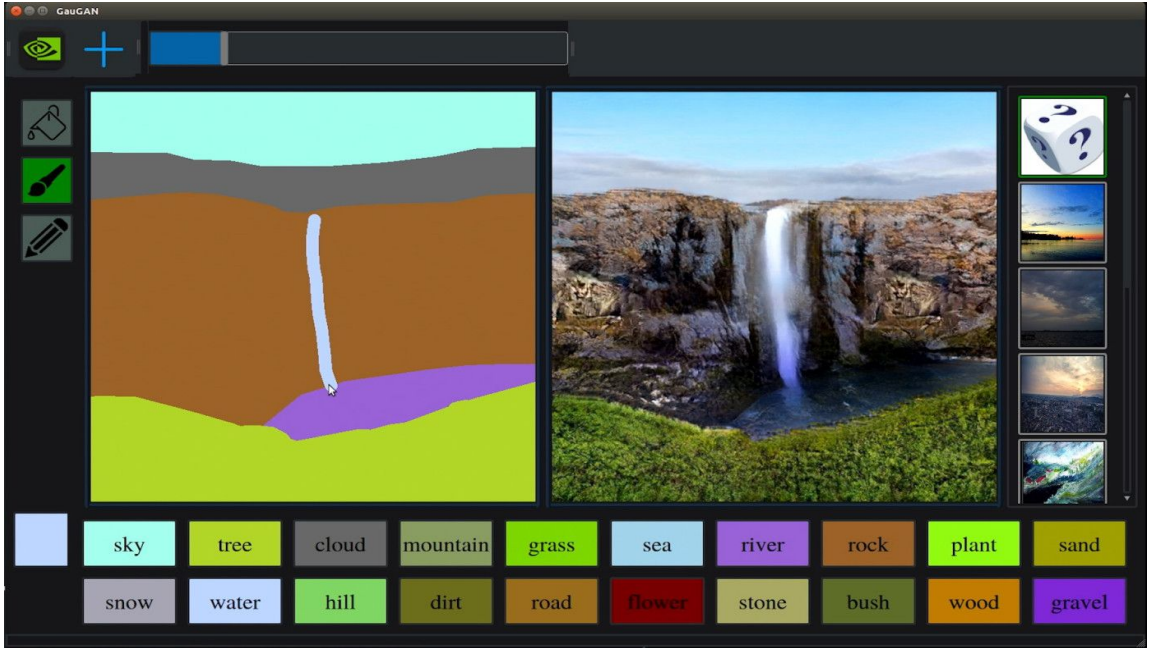
Deep Learning with Keras,TensorFlow

TensorFlow - The core open source library to help you develop and train ML models
https://www.tensorflow.org/
Changes Often! :(

Keras - Wrapper Library for TensorFlow(and some others)
https://keras.io/

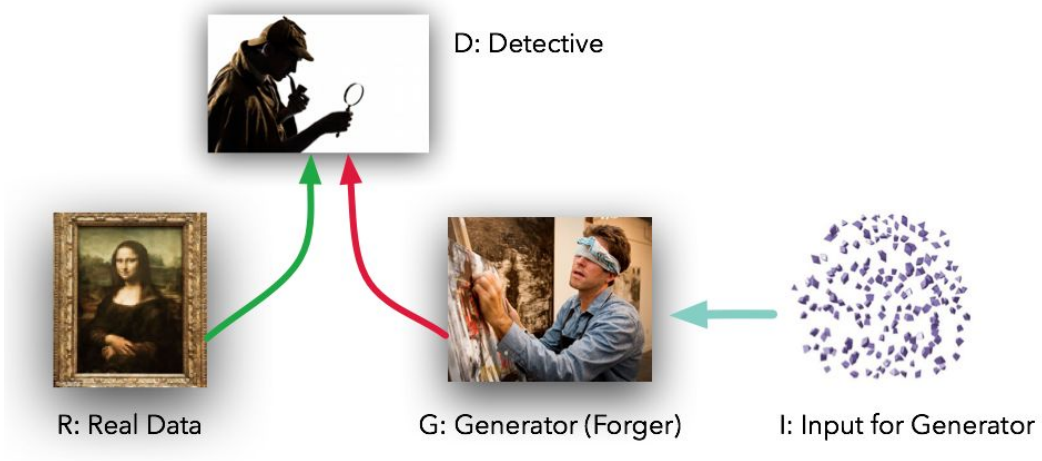
Example Image Classification Lab:
http://colab.research.google.com/github/ValRCS/RCS_Data_Analysis_Python_2019/blob/master/Keras_TensorFlow_Image_Recognition/keras_image_recognition_classifier_in_class.ipynb

New! GAN (Generative Adversarial Networks)



https://techcrunch.com/2019/03/18/nvidia-ai-turns-sketches-into-photorealistic-landscapes-in-seconds/

GAN - in 50 lines of Python

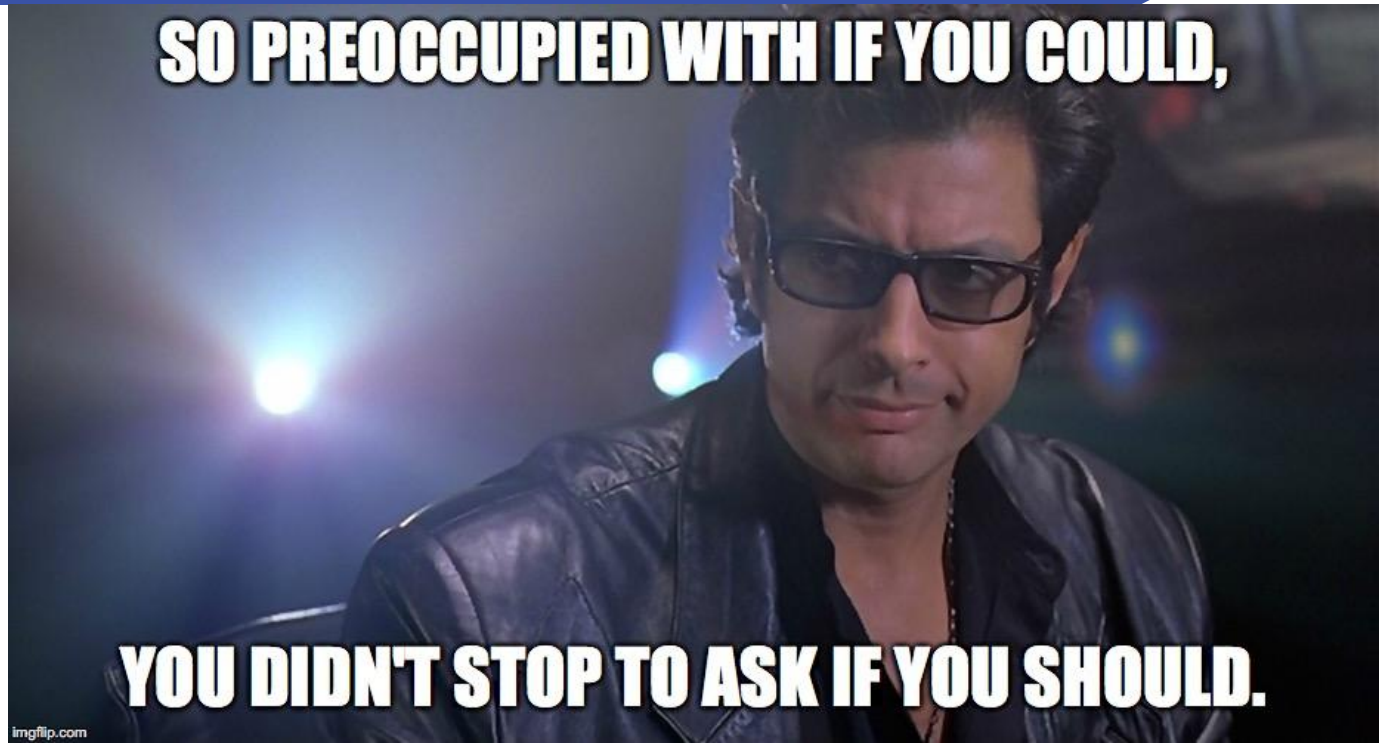


https://pytorch.org/
https://medium.com/@devnag/generative-adversarial-networks-gans-in-50-lines-of-code-pytorch-e81b79659e3f

Avoiding Pitfalls

- Overfitting
- Data Dredging / p-hacking
- https://xkcd.com/882/

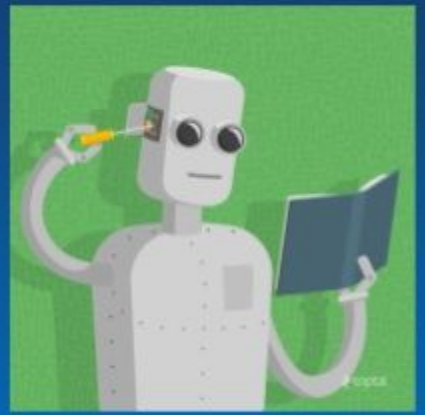
GDPR, ethics


SO PREOCCUPIED WITH IF YOU COULD,
YOU DIDN'T STOP TO ASK IF YOU SHOULD.
imgflip.com

Practical ML Applications

Machine Learning: Practical applications

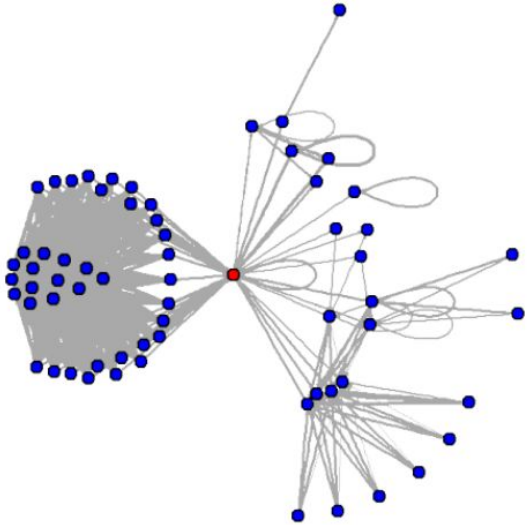
- Spam filtering
- Fraud detection
- Optical character recognition
- Medical diagnoses
- Anticipate customer needs
- Route requests efficiently
- Anticipate deals



Projects

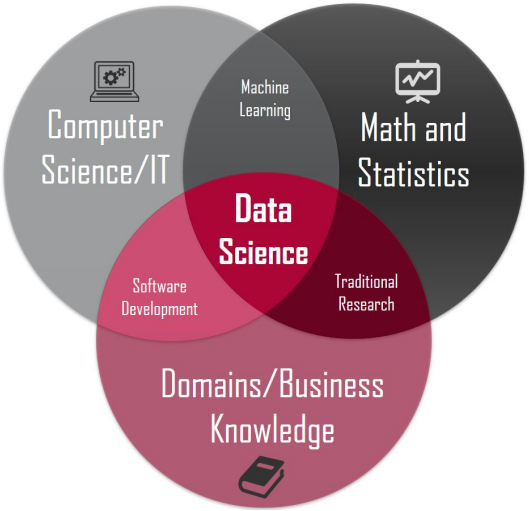
- Recommendation System / Churn Prediction
- Web Comments Sentiment Analysis
- Network Analysis (including blockchain)

- Visualizations with PowerBI(or Tableau)
- Dashboards with Dash/plotly

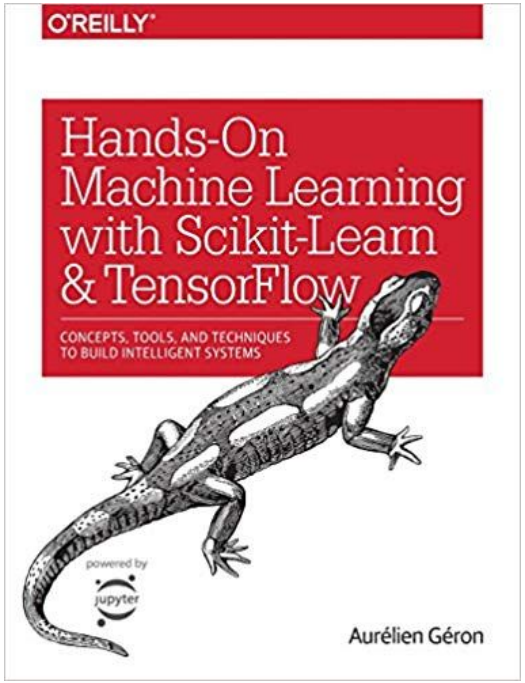


Goals

- Access structured/unstructured data
- Clean data
- Apply correct methods for analysis
- Visualize Results



Books



- General Python resources
- https://automatetheboringstuff.com/
- https://www.py4e.com/lessons

The BEST book on Python and Machine Learning
https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1491962291

Requirements

- Analytical / Logical Mind

- Helpful but NOT required knowledge:
- Comfortable in command line
- SQL
- Statistics

python™

ANACONDA®

- Helpful: a computer with a minimum of 8GB RAM
- https://www.anaconda.com/download/ (3.6+)

PALDIES!