MABEL Technology

Data Analysis and Machine Learning Using Python

Liran Ben Haim
liran@mabel-tech.com

MABEL Technology

Python and Data Analysis

- Python is a great programming language
  - Open source
  - Cross platform
  - Easy to learn and use
  - General purpose
    - Procedural
    - Functional
    - Object Oriented
- With tons of available packages you can write almost any program with python
- Many packages to manage data (local and remote) and analyze data

MABEL Technology

What can I do with Python?

- Administrative scripts
- Work with data
- Web development
- Mobile development
- Desktop applications
- Embedded and IOT
- Just about anything else you can think of

MABEL Technology

Python Packages

- Package is a collection of modules
- **Python has hundreds of modules** covering almost all areas
  - GUI
  - Scientific
  - Big data
  - Gaming
  - Tools
  - ….

MABEL Technology

Data Science

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.
- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.
- It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.
- Source: Wikipedia

MABEL Technology

Data Analysis

- Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.
- Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.
- In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation

MABEL Technology

Data Mining

- Data mining is the process of discovering patterns in large data sets involving methods at the intersection of:
  - Machine learning
  - Statistics
  - Database systems.

MABEL Technology

Machine Learning

- Machine Learning is the science of programming computers so they can *learn from data*
- *"field of study that gives computers the ability to learn without being explicitly programmed" (*Arthur Samuel, 1959)
- Example: spam filter - Machine Learning program that can learn to flag spam given examples of spam emails
- **Machine learning is a comprehensive approach to solving problems.** Its not a list of algorithms

MABEL Technology

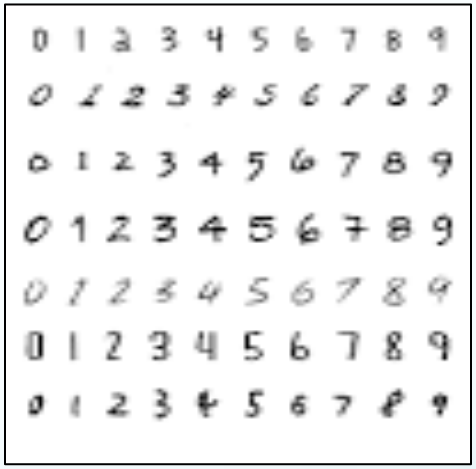
Data Mining vs Queries

- SQL Query:
  - Select * from customers where age>45
- Data mining:
  - Select ??? From ???
- How many items we sell every year? – Query
- Is there any dependency between items we sell? – data mining
- How many items we need to produce for black Monday – data mining

MABEL Technology

Why Learn?

- Learn it when you can't code it
  - Complex tasks where deterministic solution don't suffice
  - e.g. speech recognition, handwriting recognition
- Learn it when you can't scale it
  - Repetitive task needing human-like expertise (e.g. recommendations, spam & fraud detection)
  - Speed, scale of data, number of data points
- Learn it when you need to adapt/personalize
  - e.g., personalized product recommendations, stock predictions



MABEL Technology

Applications

- Fraud detection.
- Web search results.
- Real-time ads on web pages
- Credit scoring and next-best offers.
- Prediction of equipment failures.
- New pricing models.
- Network intrusion detection.
- Recommendation Engines
- Customer Segmentation
- Text Sentiment Analysis
- Predicting Customer Churn
- Pattern and image recognition.
- Email spam filtering.
- Financial Modeling

MABEL Technology

Supervised Learning

- Predicting using labeled data

A	B	C	Res
10	3	45	T
23	7	12	F
56	8	9	F
20	9	30	T

30	**4**	**24**	**?**

MABEL Technology

Unsupervised Learning

- Group unlabeled data using similarities
- Cluster analysis
- Extract interesting patterns

A	B	C	D
1	3	5	2
8	6	2	4
1	5	8	9
10	6	2	1
3	12	7	4
1	10	19	3
4	8	5	11

MABEL Technology

Reinforcement Learning

- Perform an action from experience
- Feedback received after a sequence of actions/predictions
- Games/Robots

MABEL Technology

CRISP-DM

- Cross Industry Standard Process for Data Mining

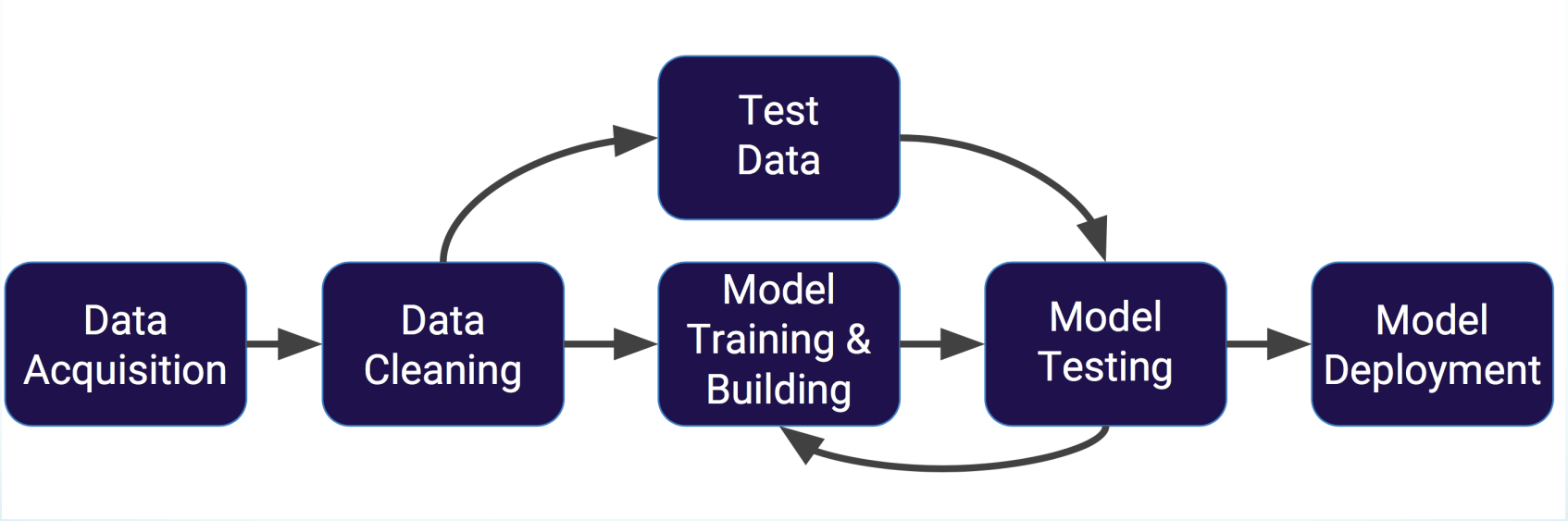
MABEL Technology

Terminology

- **Model** - a set of patterns learned from data.
- **Algorithm** - a specific ML process used to train a model.
- **Training data** - the dataset from which the **algorithm** learns the **model**.
- **Test data** - a new dataset for reliably evaluating model performance.
- **Features** - Variables (columns) in the dataset used to train the model.
  - **Independent variables** – but most of the time depends
- **Target variable** - A specific variable you're trying to predict.
- **Observations** - Data points (rows/records) in the dataset

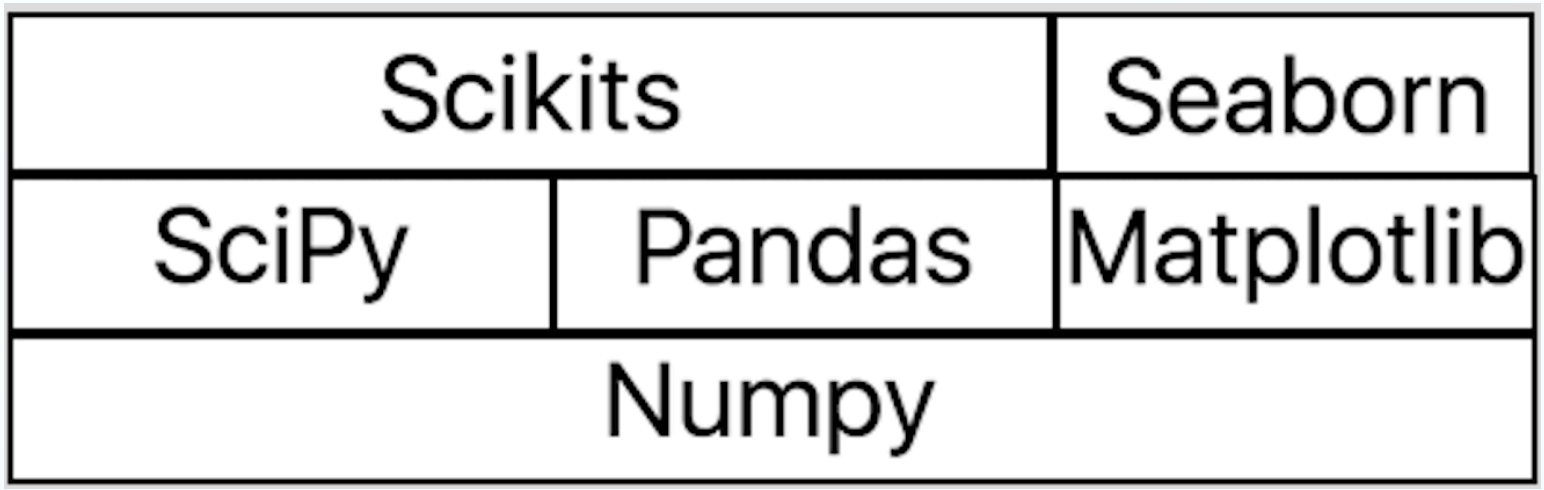
MABEL Technology

Machine Learning Process



MABEL Technology

Python Packages



MABEL Technology

Scikit Learn

- Every algorithm is exposed in scikit-learn via an “Estimator”
- import the algorithm:
  - from sklearn.linear_model import LinearRegression
- The process for each algorithm depends on its type

MABEL Technology

Deep Learning

- If we have a large amount of features, It is not easy to find out the features that really matter
- For example – we need to find a car in a picture:
  - The pixels – our features (320 * 200 = 64000 features!!!)
  - Some pixels are not important and some very
- Deep Learning uses methods to split the problem to layers
- The popular algorithm is Neural Networks

MABEL Technology

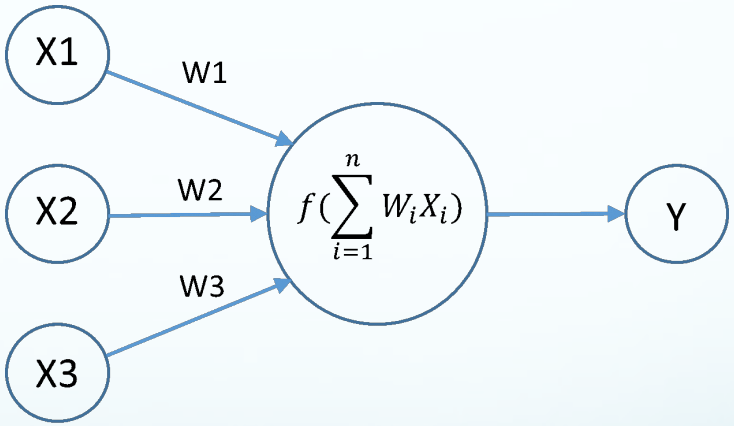
Our Brain

- 100,000,000,000 neurons
- Each neuron is connected to 10000 other neurons using synapses
- Simple decision uses 200-300 neurons
- Communication speed is measured in ms
  - Signals are varying
- 100THz , 10000 bit computer (at least)

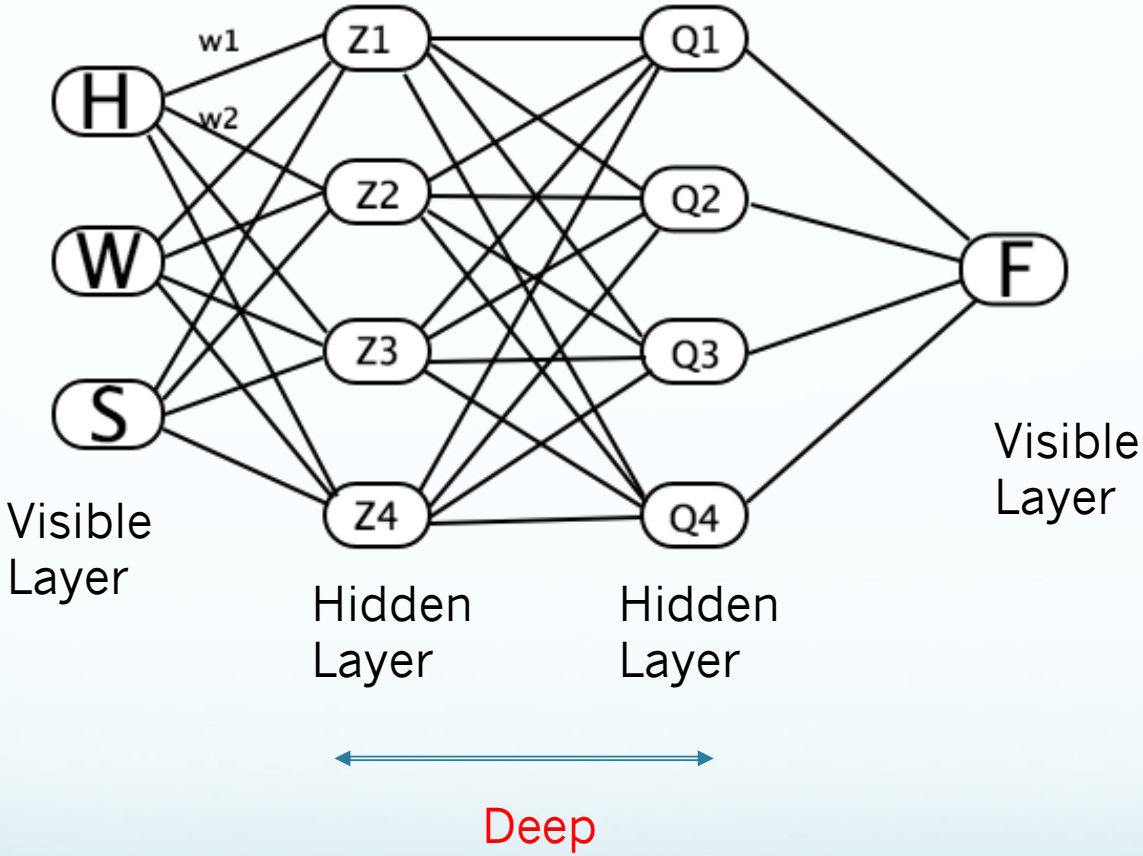
MABEL Technology

Neural Network

- Based on neurons
- Each neuron has N inputs with weights and one output
- Calculate
  - sum = $\sum_1^n (xi * Wi)$
  - Apply activation function
    - Output = f(sum)



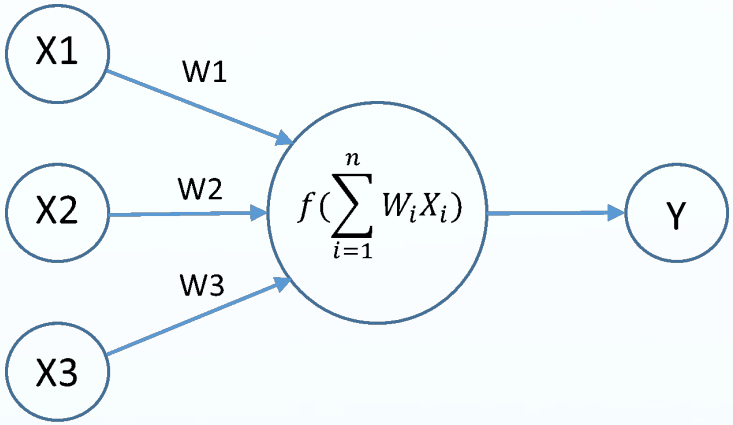
- NN can solve any ML problem, the only difference is the activation function


w1
w2
H
W
S
Z1
Z2
Z3
Z4
Q1
Q2
Q3
Q4
F
Visible Layer
Hidden Layer
Hidden Layer
Visible Layer
Deep

MABEL Technology

Single Neuron

- Can learn linear problem



- Identity activation function
  - F(x) = x
- Y = X1 * w1 + X2 * w2 + X3 * w3

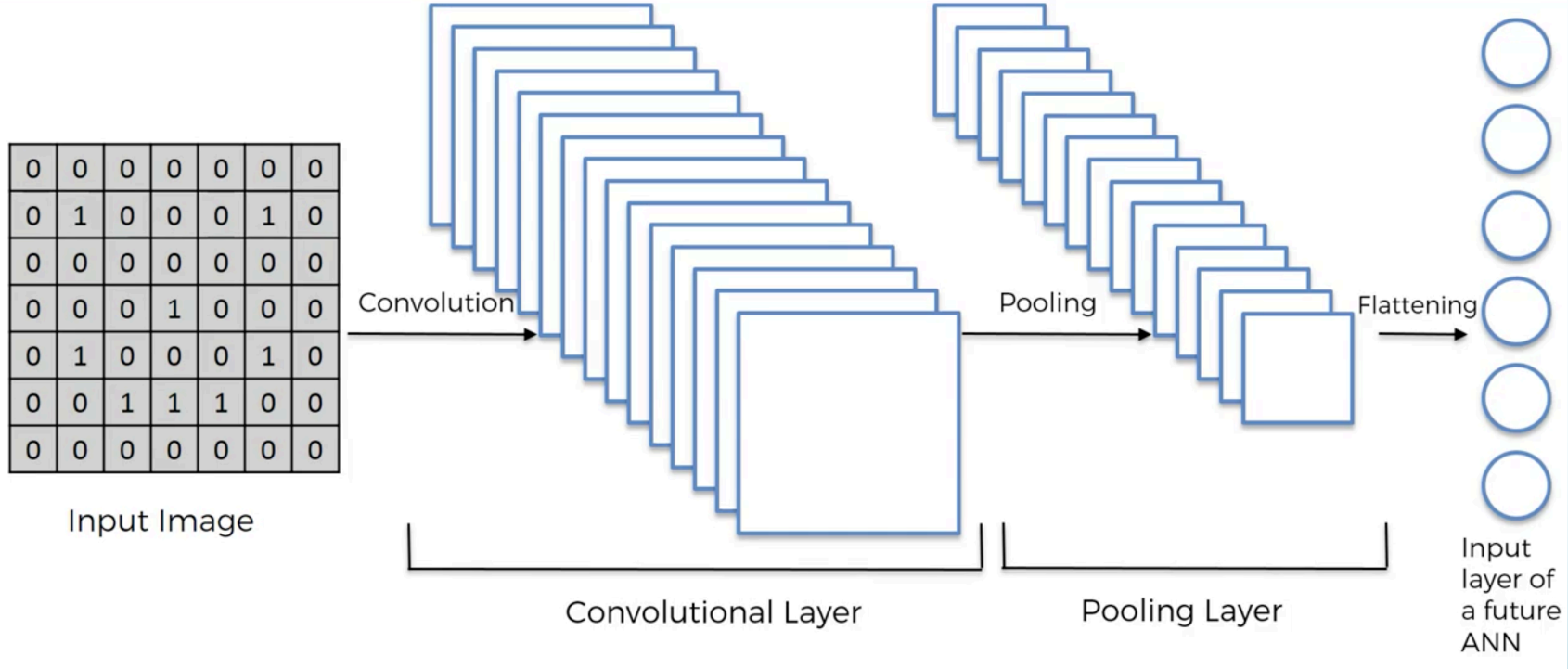
MABEL Technology

Other NN types

- Auto encoders
- SOM
- Convolutional
- Boltzmann machines
- Recurrent Neural Networks

MABEL Technology

Convolutional NN



http://www.cs.cmu.edu/~aharley/vis/conv/flat.html

MABEL Technology

Natural Languages Processing

- Processing text is a complex task
- We will want to:
  - Compile Documents
  - Featurize Them
  - Compare their features
- Tools:
  - String ops in python
  - Regular expressions
  - External packages

MABEL Technology

Libraries

- NLTK
- TextBlob
- Tensorflow
- Stanford CoreNLP
- Spacy
- Gensim
- Many more
  - See pypi