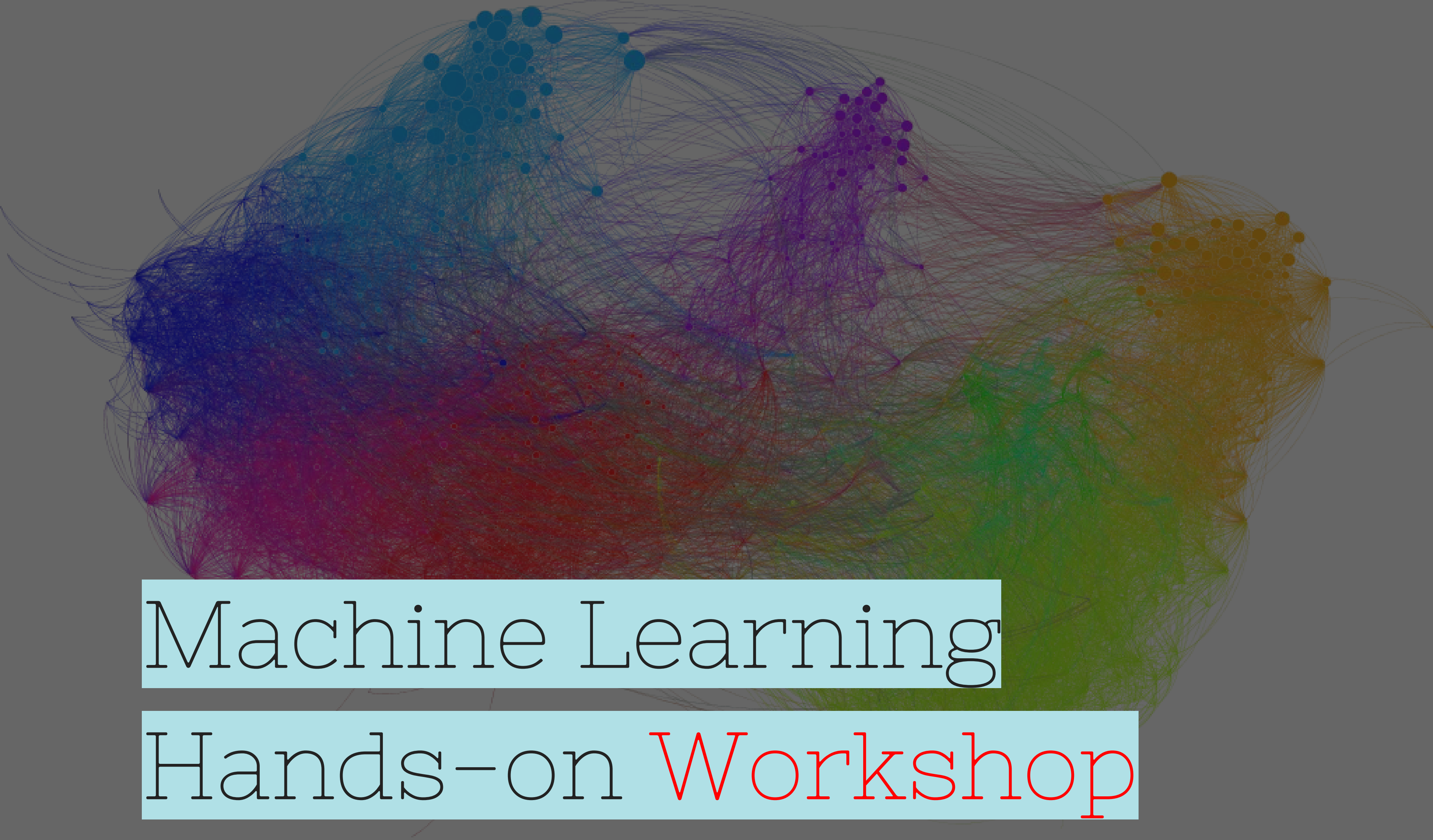
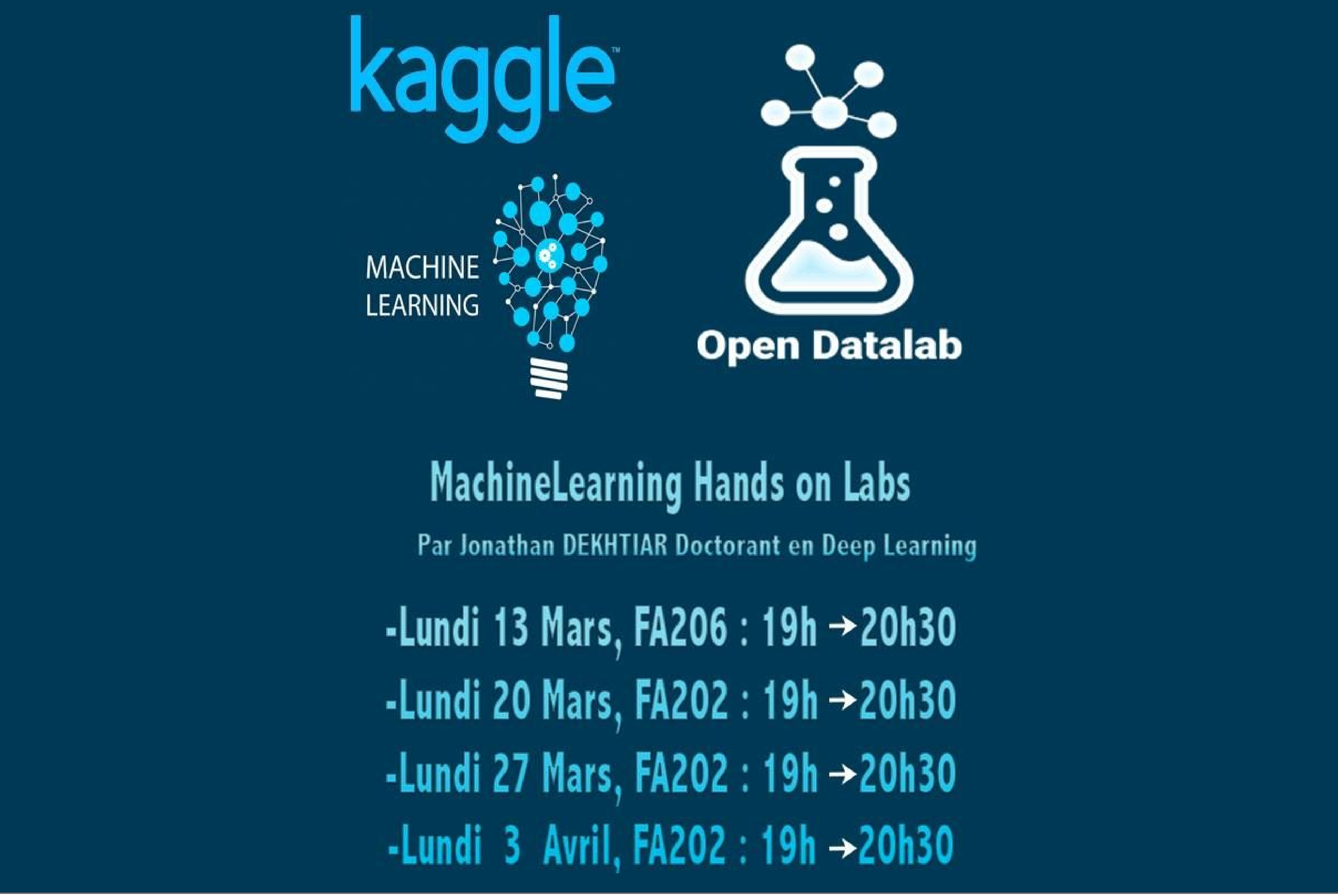


Machine Learning
Hands-on Workshop



kaggle

MACHINE LEARNING

Open Datalab

MachineLearning Hands on Labs

Par Jonathan DEKHTIAR Doctorant en Deep Learning

-Lundi 13 Mars, FA206 : 19h → 20h30

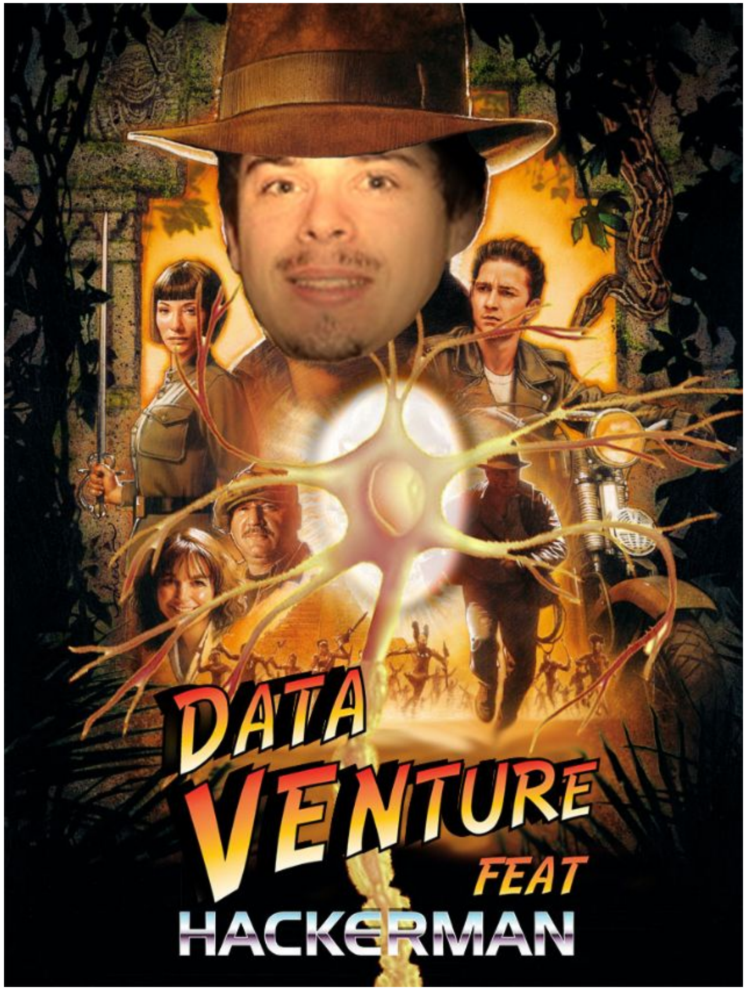
-Lundi 20 Mars, FA202 : 19h → 20h30

-Lundi 27 Mars, FA202 : 19h → 20h30

-Lundi 3 Avril, FA202 : 19h → 20h30

Data Venture

http://www.opendatalab.xyz



Data Venture

http://www.opendatalab.xyz

Président	Julien	Jerphanion
Vice-Président	Sylvain	Marchienne
Trésorière	Léna	Schofield
Secrétaire	Vincent	Dehaye
Resp. Com	Sevan	Garois
Resp. Setup	Valentin	Montupet
Resp. Scientifique	Sy Toan	Ngo
Resp. Talks	Gabriel	Hurtado
Resp. Projet	Benjamin	Rivière
Resp. Partenariat	Antoine	Jeannot

About me

Jonathan DEKHTIAR

2nd Year PhD. Student

Engineer @ UTC 2015
Machine Learning and Statistical Analysis

Keeping in touch’

Twitter: @Born2data

LinkedIn: https://www.linkedin.com/in/jonathandekhtiar/

Tech. Blog : https://www.born2data.com

RSS Feed: https://www.feedcrunch.io/@dataradar

Email: contact@jonathandekhtiar.eu and jonathan.dekhtiar@utc.fr

How to ML

https://www.docker.com

docker

Most efficient, lightweight, hacking-ready, versatile virtualization platform.

amazon webservices

Windows Azure

Google Cloud Platform

Jupyter Notebooks for Data Science

```
docker pull jupyter/datascience-notebooks #take a coffee
```

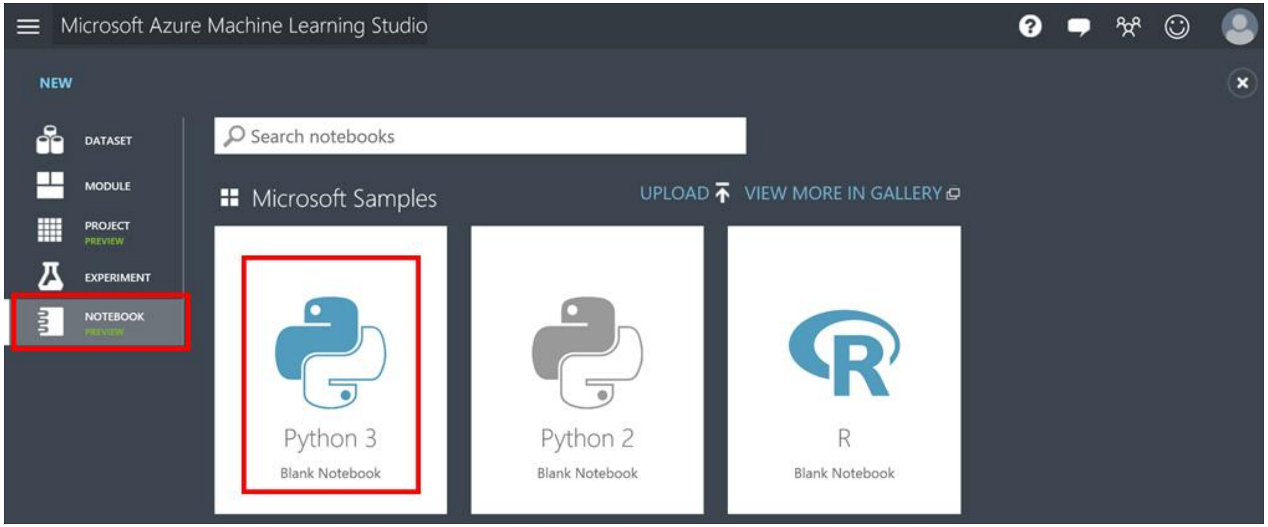
```
docker run -d -p 8888:8888 jupyter/datascience-notebook start-notebook.sh --NotebookApp.token=''
```

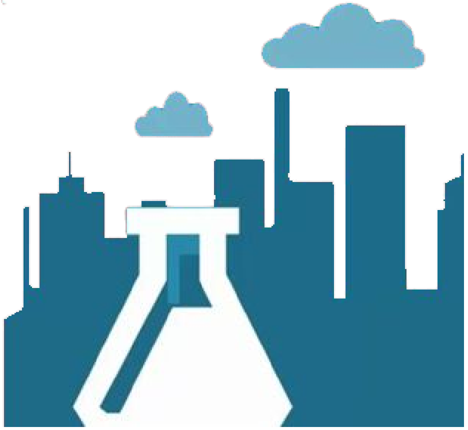
Open Browser on: http://localhost:8888/

How to ML ?

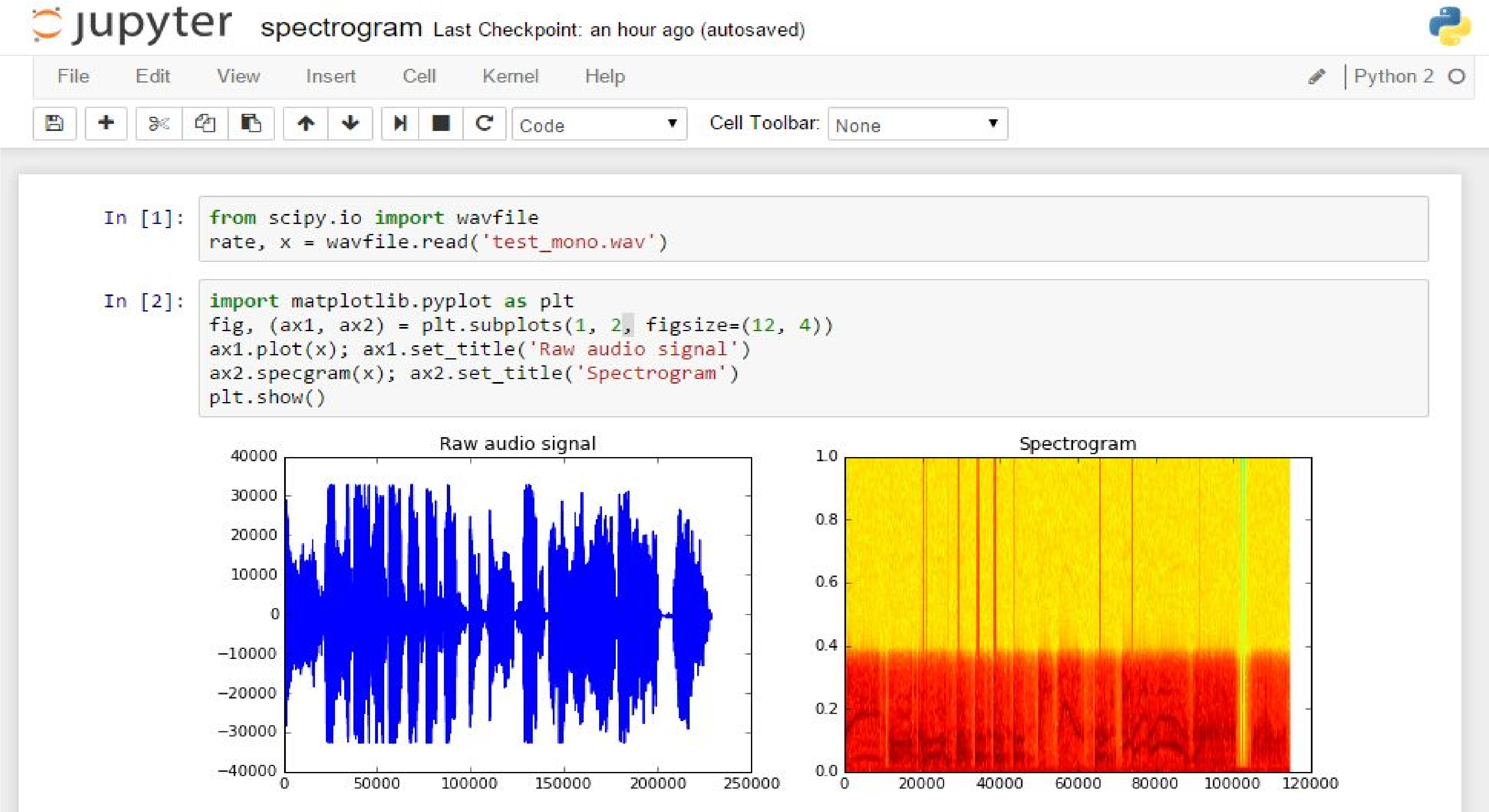
https://studio.azureml.net/

Free-Tier Usage.
Quite fast and scalable.
Cloud based on - Azure SaaS





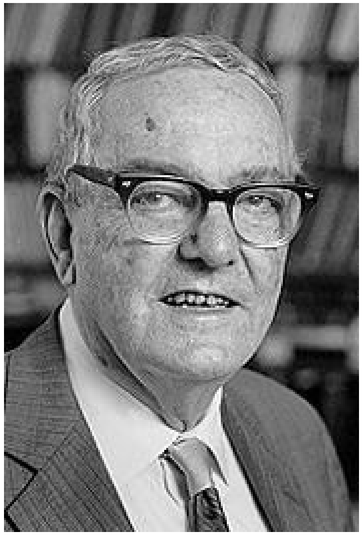
Jupyter Notebooks ?



In [1]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

In [2]:
```
import matplotlib.pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram')
plt.show()
```

Machine Learning



Herbert Simon

Turing Award 1975

Nobel Prize in Economics 1978

“Learning is any process by which a system improves performance from experience.”

“Machine Learning is concerned with computer programs that automatically improve their performance through experience. “

Why Machine Learning ?

- Develop systems that can automatically adapt and customize themselves to individual users.
    - Personalized environment, context-awareness, etc.

- Discover new knowledge from large databases (data mining).
    - Pattern Discovery, Influential Parameters, etc.

- Ability to mimic human and replace certain monotonous tasks - which require some intelligence.
    - like analyzing imagery data 2D/3D and comparing them

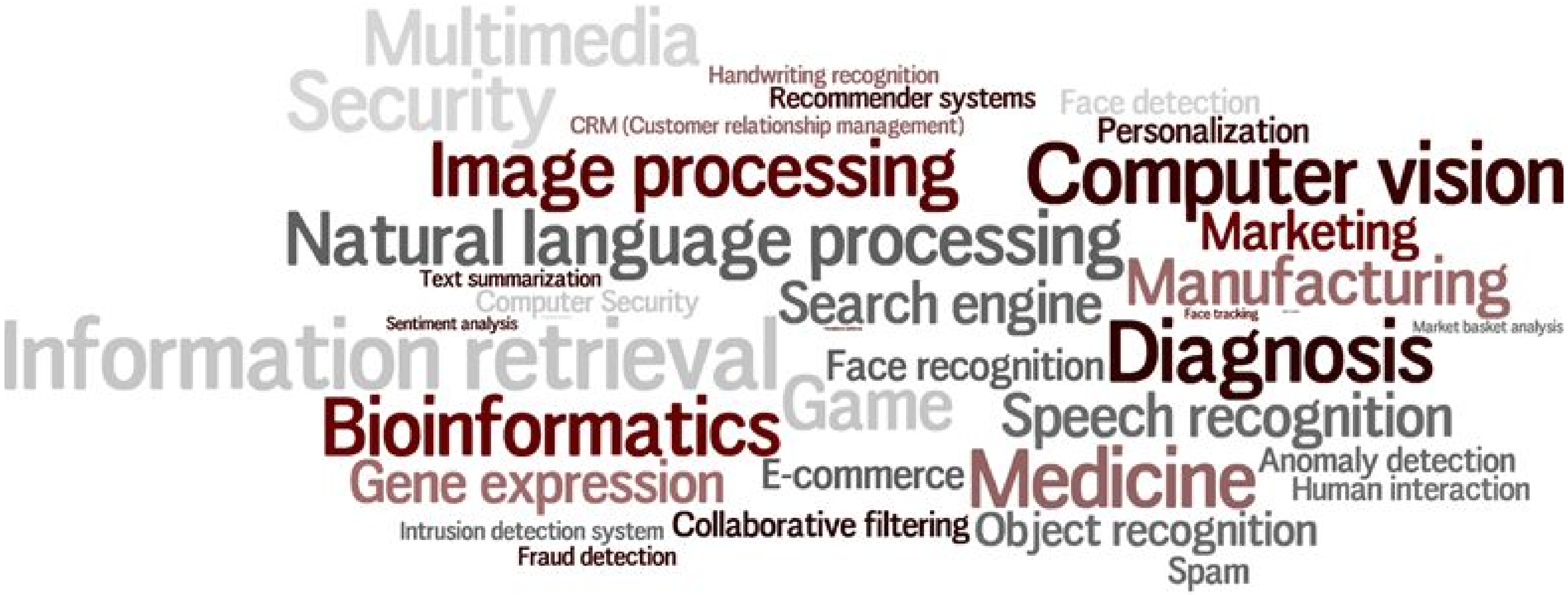
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (knowledge engineering bottleneck).

- Etc.

Why Now ?

- Flood of available data (especially with the advent of the Internet)
- Increasing computational power (GPU Computing, Moore's Law, etc.)
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries => Research fundings
- Increasing complexity of challenges induced by massive data and complex data
- etc.

Who needs/uses ML ?



What's the concept ?

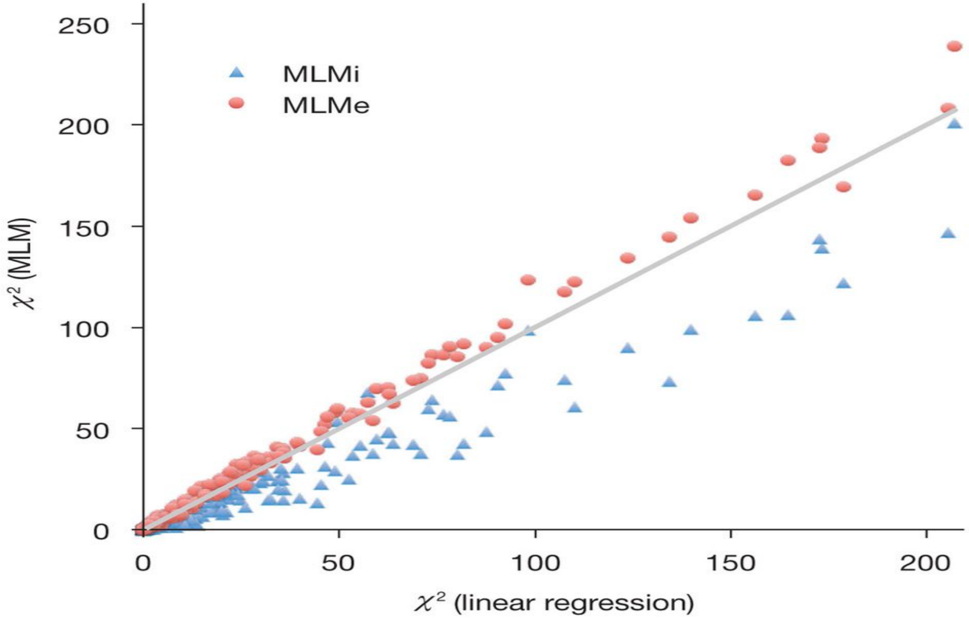
Learning

⇓

Improving over time with experience at **some task**.

1. Improve over task: T
2. Measure Performance with indicators: P
3. Based on Experience: E

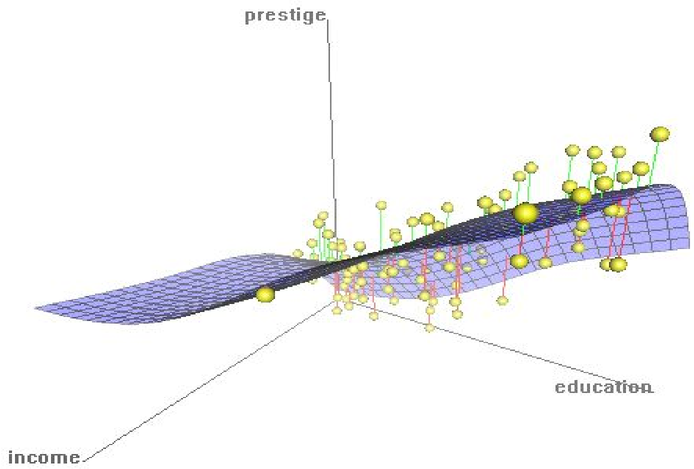
Regression



Objective:

To determine a parameter in function of a set of features.

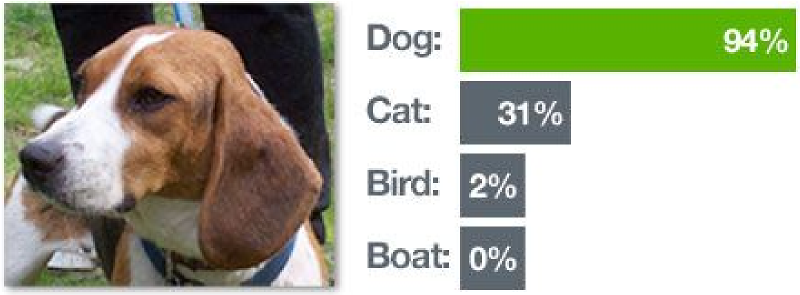
$$f(x_1, \dots, x_n) = y$$



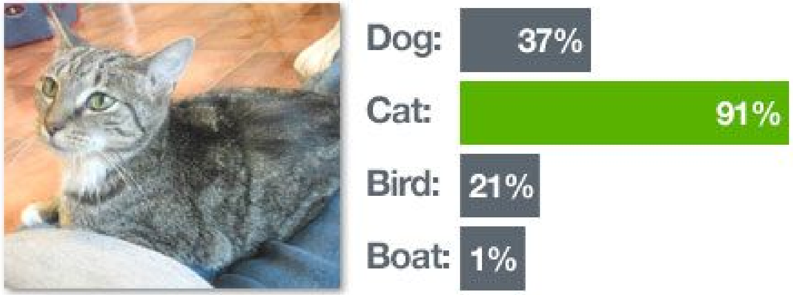
A few models:

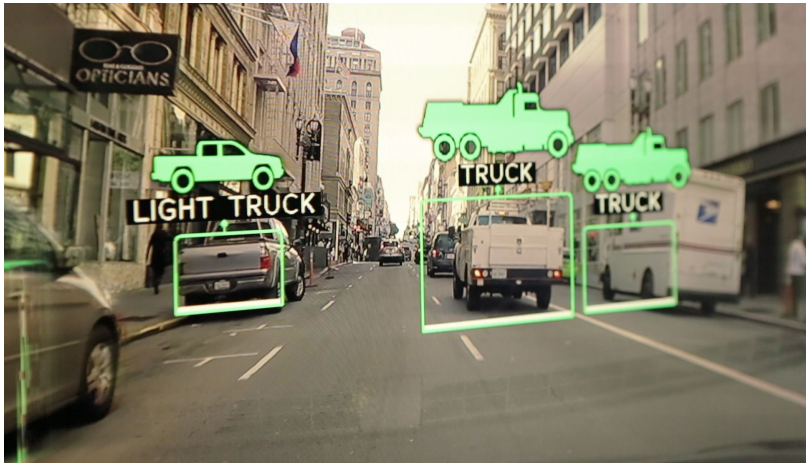
- Linear Regression
- Logistic Regression
- Quadratic Regression
- Bayesian Regression
- etc.

Classification



Objective:
To apply a label on a given set of features.



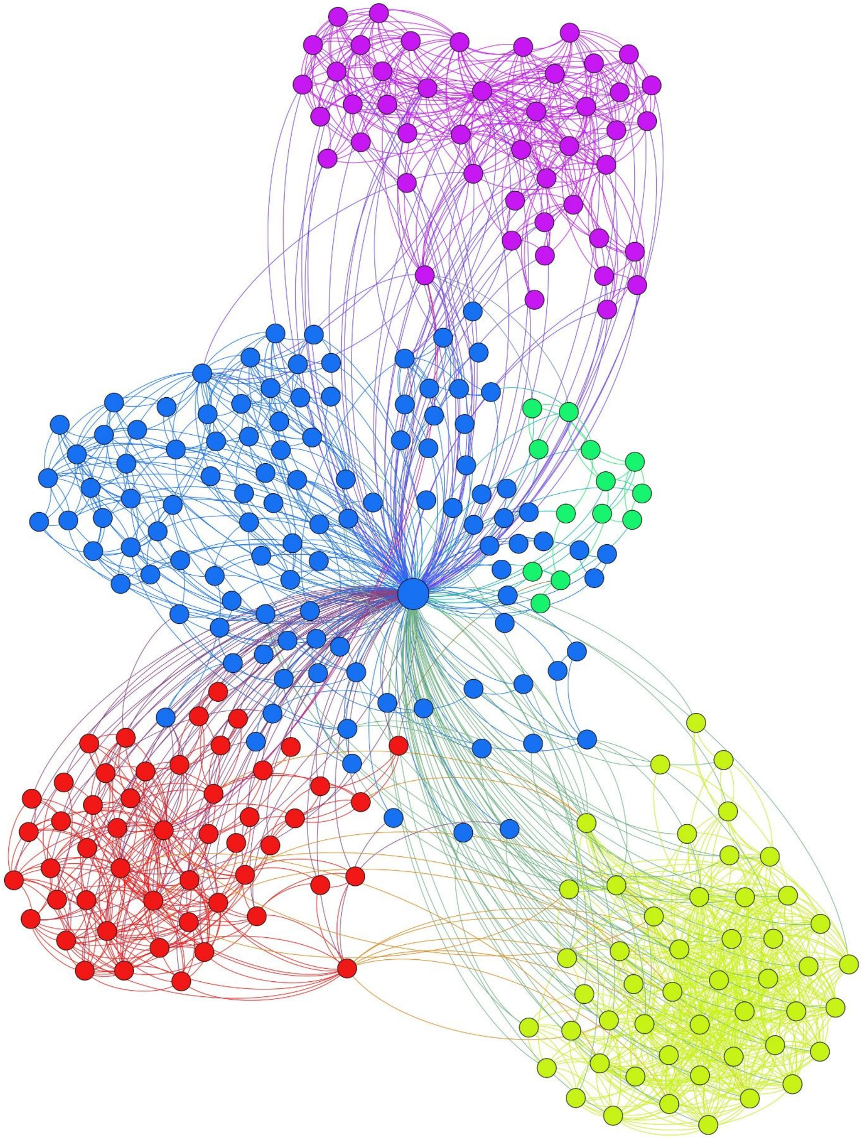


A few models:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Random Forest
- (Deep?) Neural Network (NN)
- etc.

Clustering

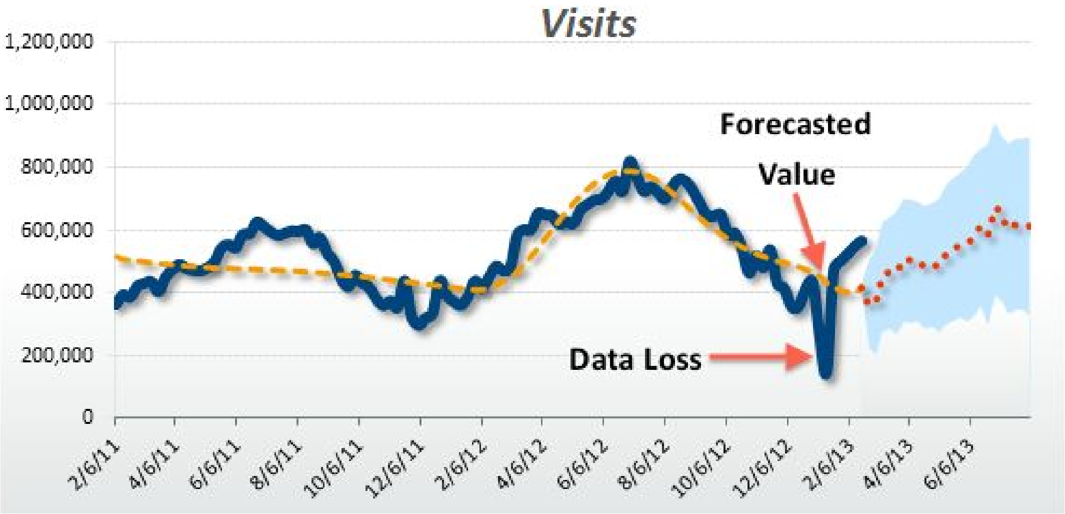
Objective:
Discover hidden groups or patterns inside the data



A few models:

- K-Means
- EM-Clustering
- Hierarchical Clustering
- Neural Networks (SOM & SOFM)
- etc.

Time Series Prediction



Objective:

Predict the next value(s) of a time series beforehand.

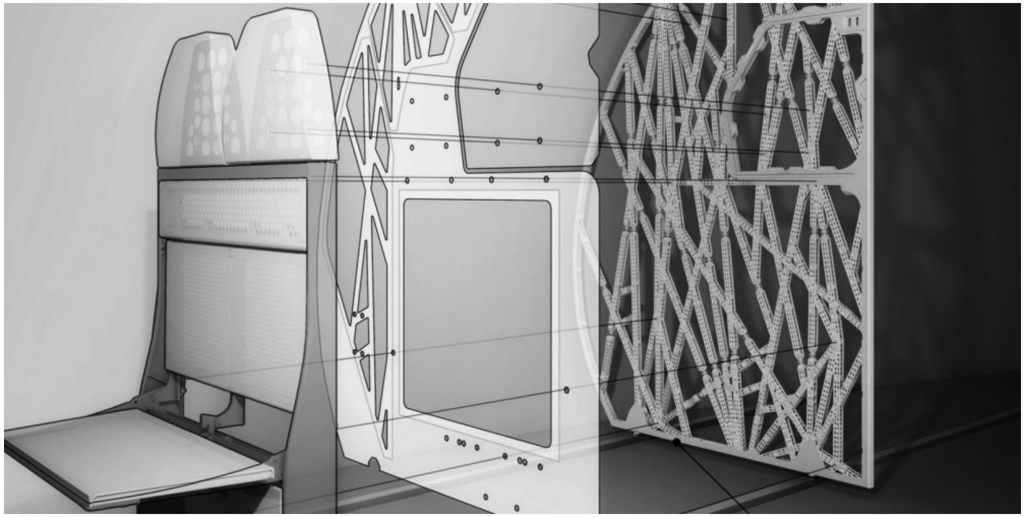
A few models:

- Recurrent Neural Networks (RNN)
- Long Short Term Memory Neural Networks (LSTM)
- etc.

Generative Models

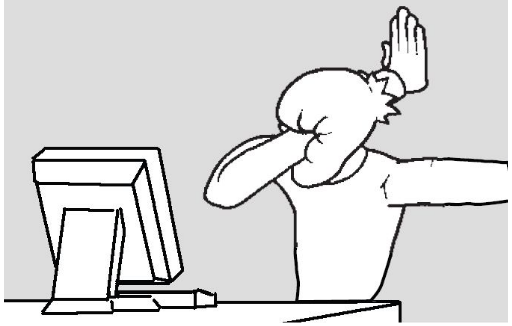
Objective:

Generate a completely new piece of data that looks like a human-made one.



A few models:

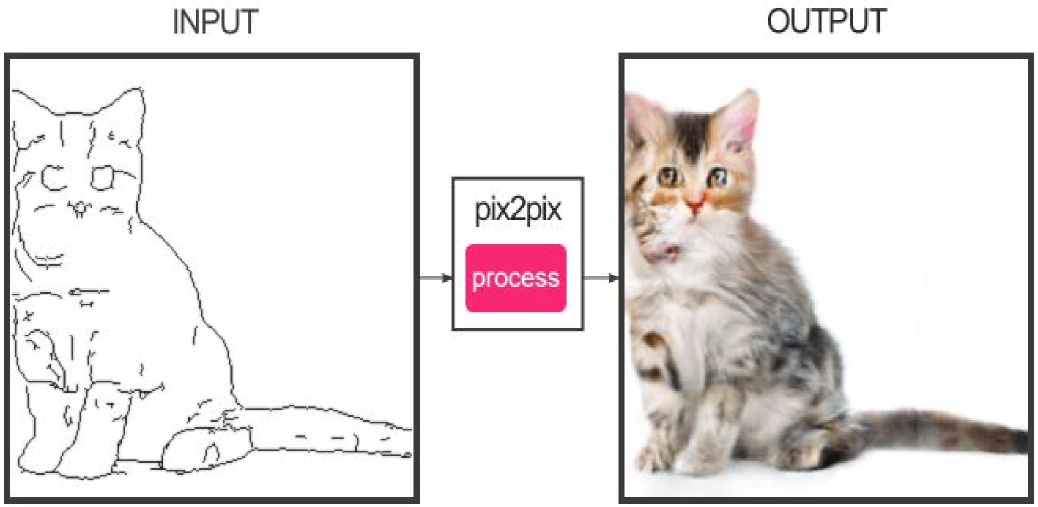
- Generative Adversarial Neural Network (GAN)
- RNN
- LSTM
- etc.



Generative Models

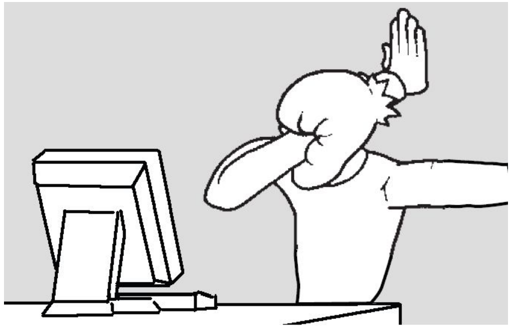
Objective:

Generate a completely new piece of data that looks like a human-made one.



A few models:

- Generative Adversarial Neural Network (GAN)
- RNN
- LSTM
- etc.



What about generalisation?

What about generalisation?

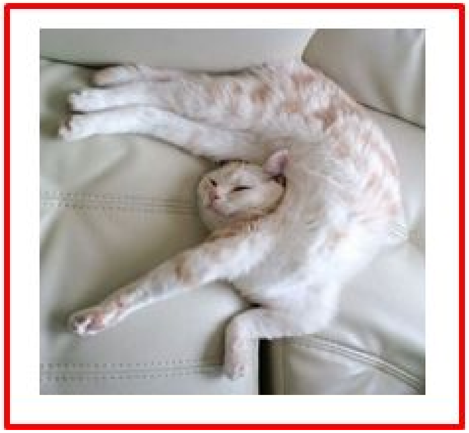
They are “**Cats**”…

Too easy Bro !
Make it way harder !

What about generalisation?

???

???



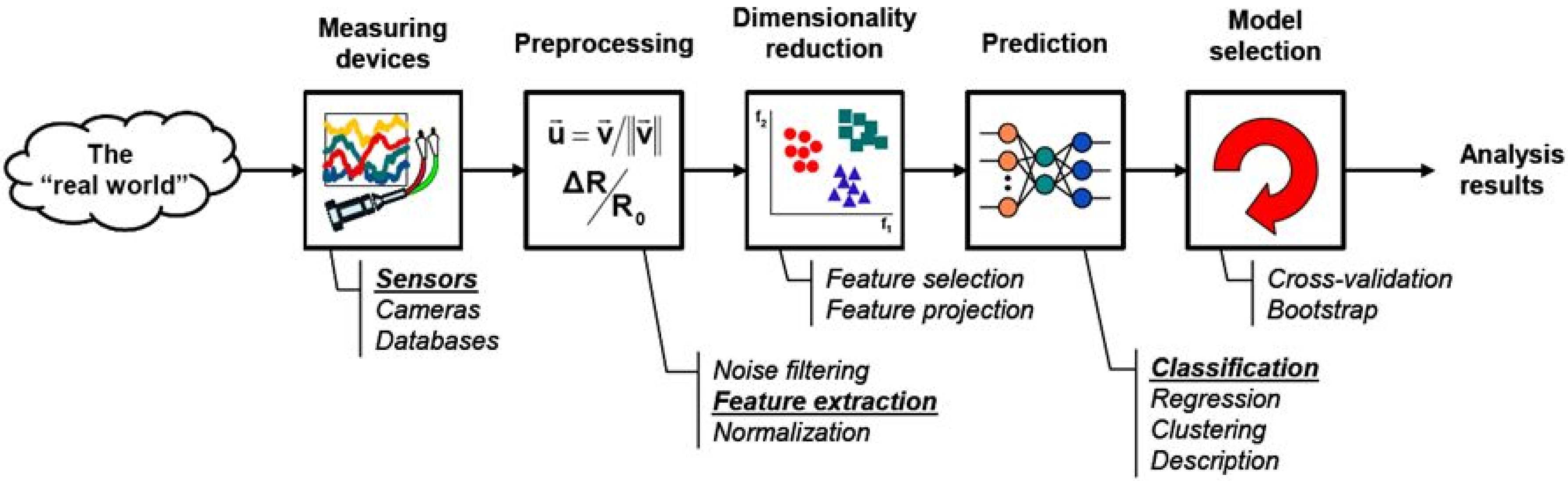
???

???

???

???

Learning workflow



The "real world"

Measuring devices

- **Sensors**
- Cameras
- Databases

Preprocessing

- Noise filtering
- **Feature extraction**
- Normalization

Dimensionality reduction

- Feature selection
- Feature projection

Prediction

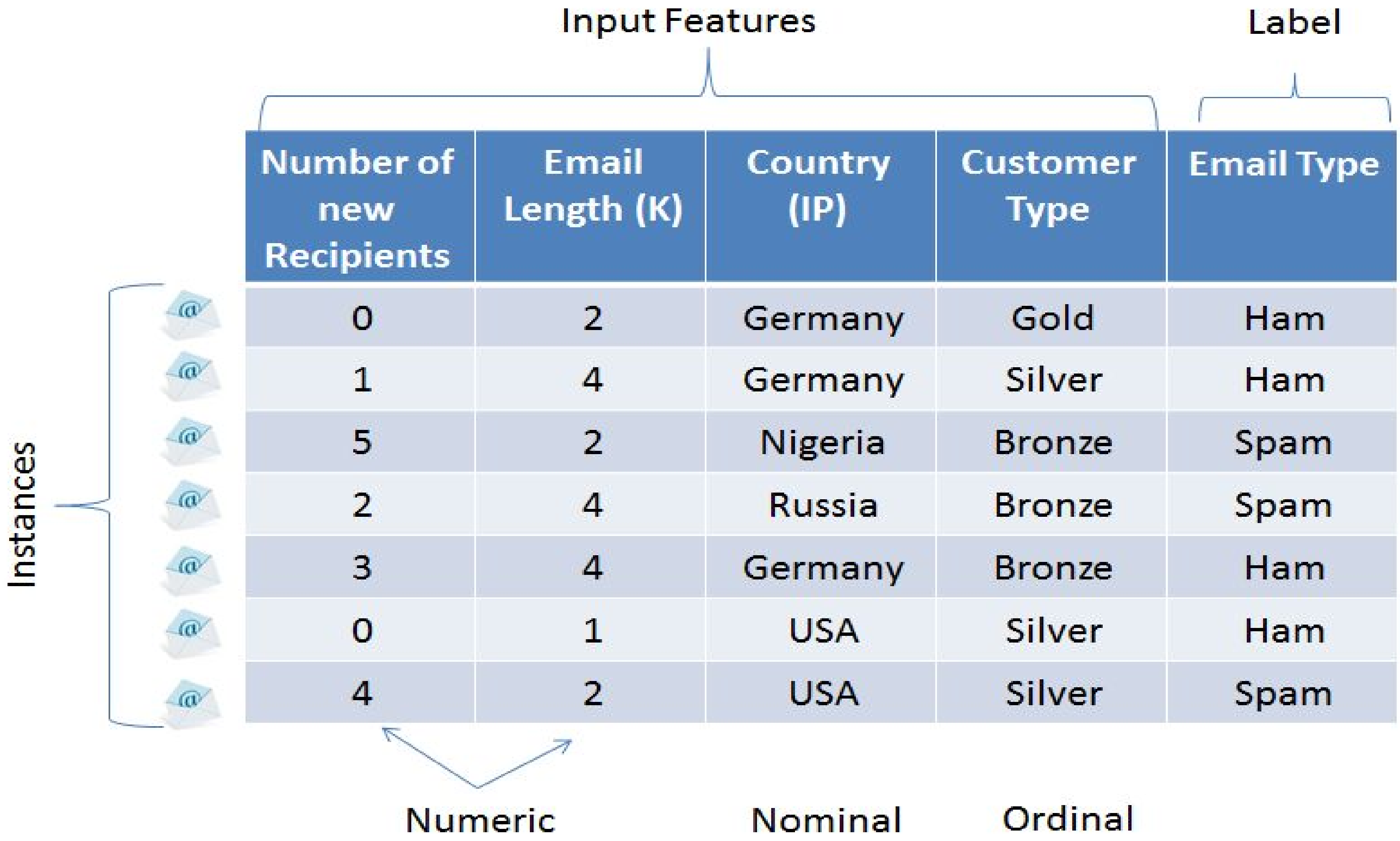
- **Classification**
- Regression
- Clustering
- Description

Model selection

- Cross-validation
- Bootstrap

Analysis results

Speaking about... data !

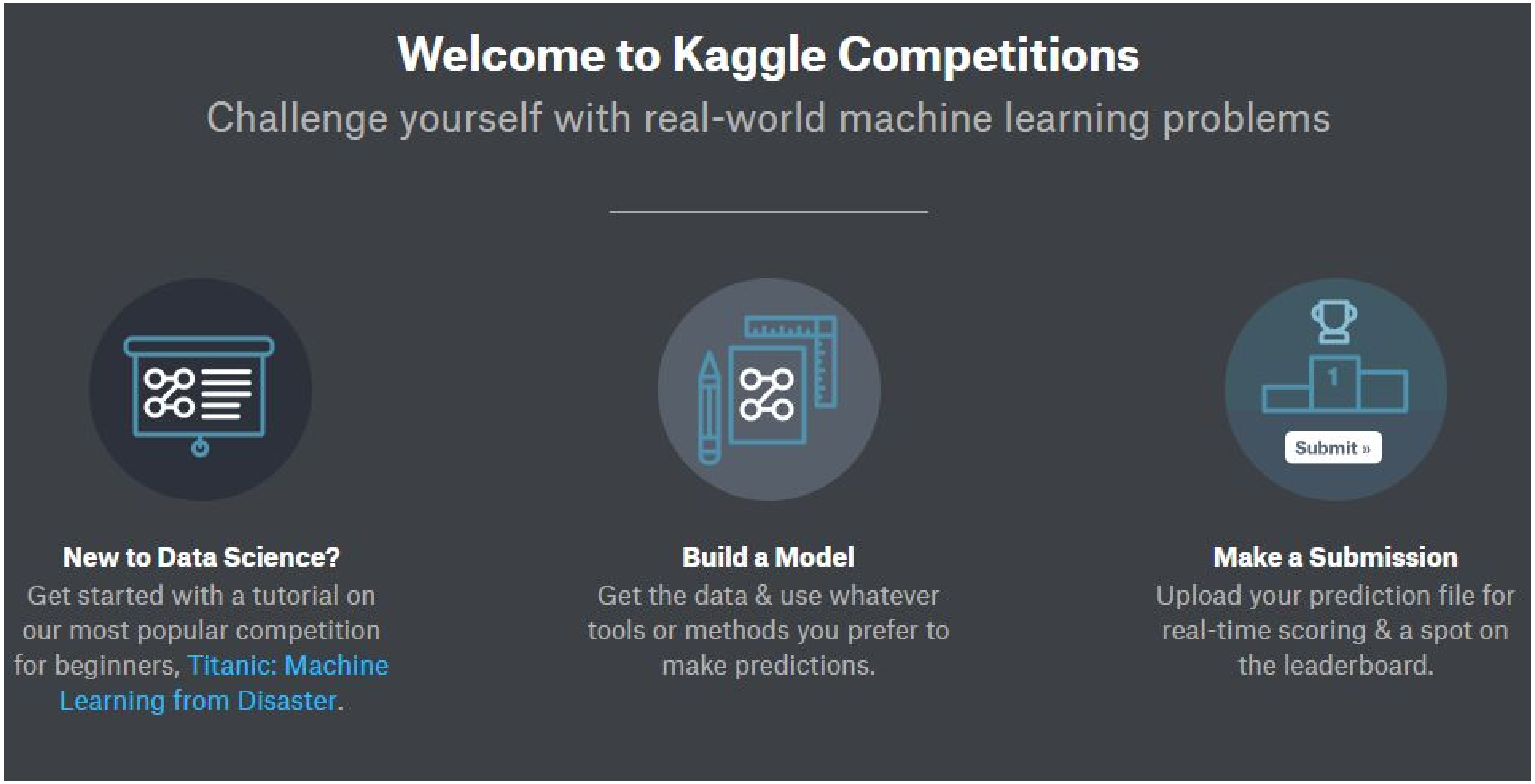


Input Features
Label
Instances
Numeric
Nominal
Ordinal

Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Kaggle

https://www.kaggle.com/



Welcome to Kaggle Competitions

Challenge yourself with real-world machine learning problems

New to Data Science?
Get started with a tutorial on our most popular competition for beginners, Titanic: Machine Learning from Disaster.

Build a Model
Get the data & use whatever tools or methods you prefer to make predictions.

Make a Submission
Upload your prediction file for real-time scoring & a spot on the leaderboard.

AirBnB

https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

Where will a new guest book their first travel experience?

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

In this recruiting competition, **Airbnb challenges you to predict in which country a new user will make his or her first booking**.

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

scikit learn

NumPy

bokeh

python™

SciPy

Hands-on Lab

dmlc XGBoost

https://goo.gl/O2PWM7