

Machine Learning with Python - from Model to Web-Service

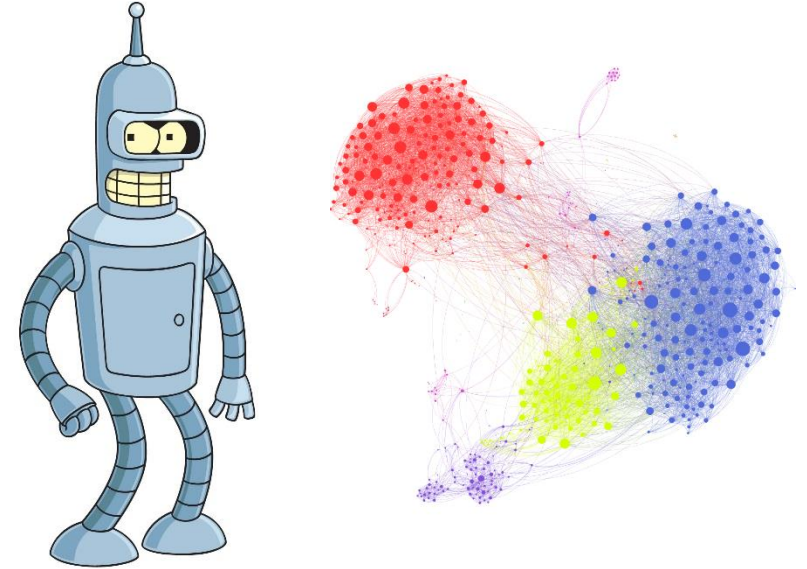
Aleksei Tiulpin,
Research Unit of Medical Imaging,
Physics and Technology
University of Oulu

Few words about me

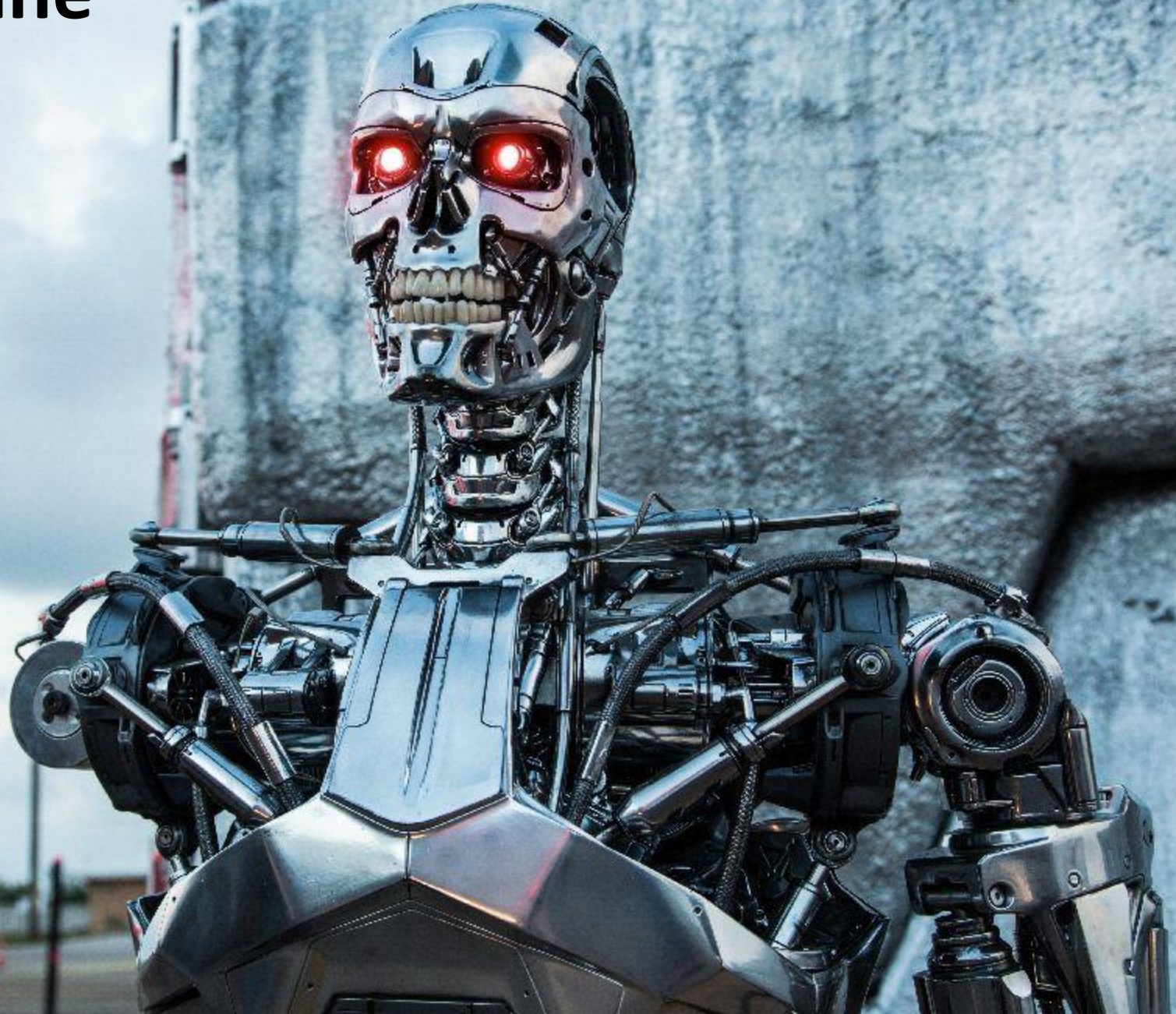
- I am ~23.93 years old 😊
- Have a degree in Applied Mathematics
- Doing PhD on novel quantitative methods for image analysis of the knee joint
- 4 years with Python
- <https://github.com/lext>

Outline

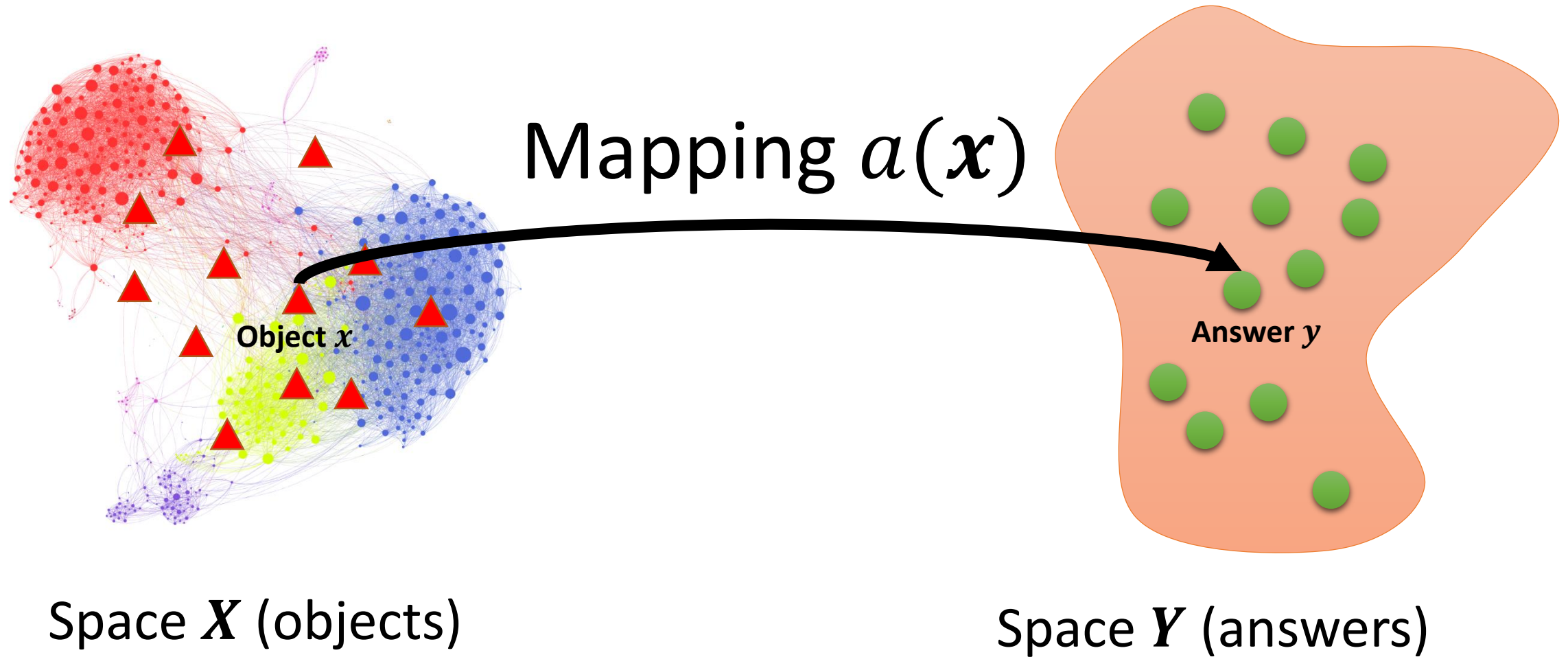
- What is Machine Learning
- Random Forest
- Python for Machine Learning: scikit-learn
- Case study: handwritten digit recognition
 - Implementation of the pipeline
 - Web-service with a pre-trained model



What is Machine Learning?



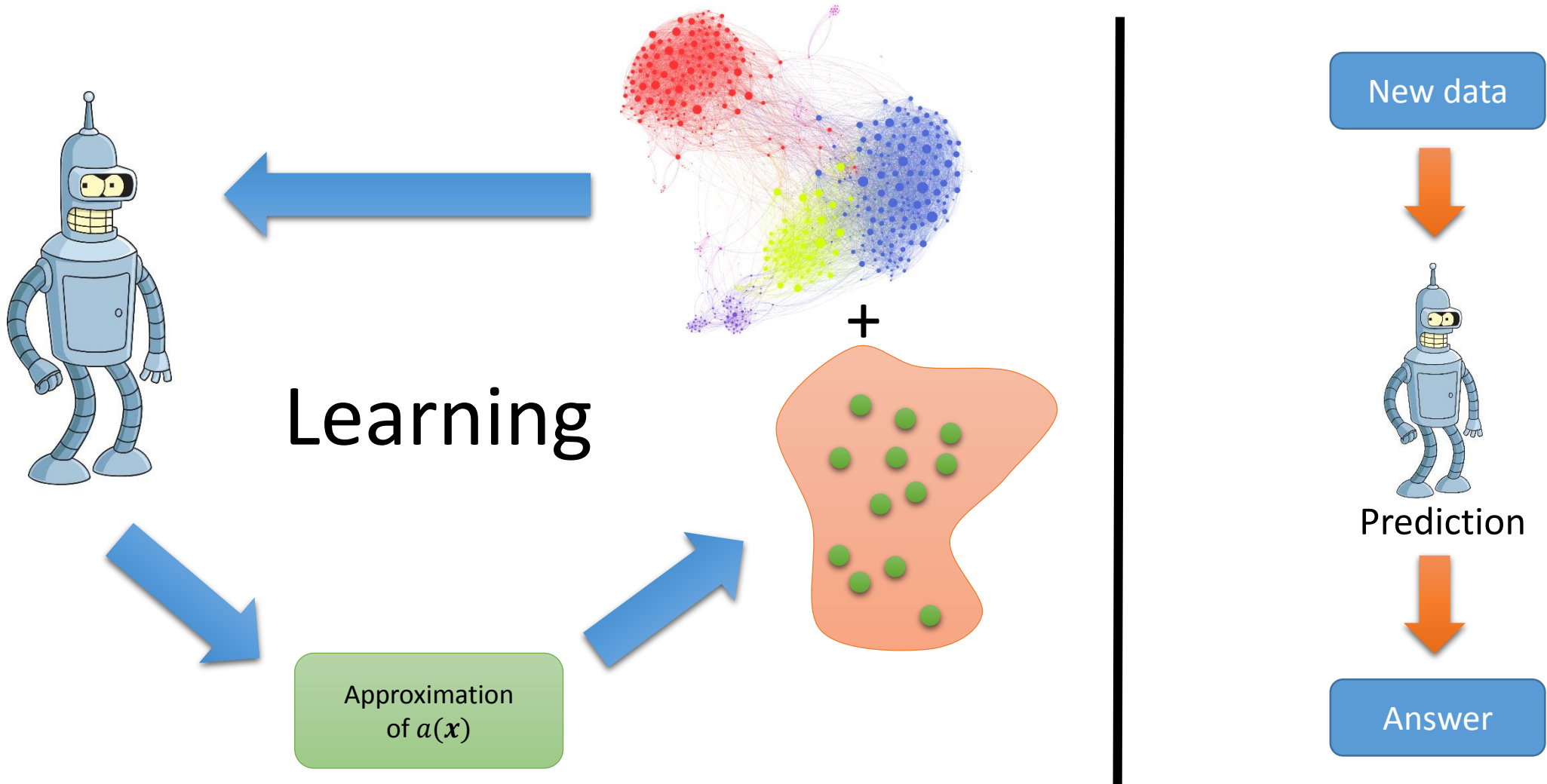
Machine Learning



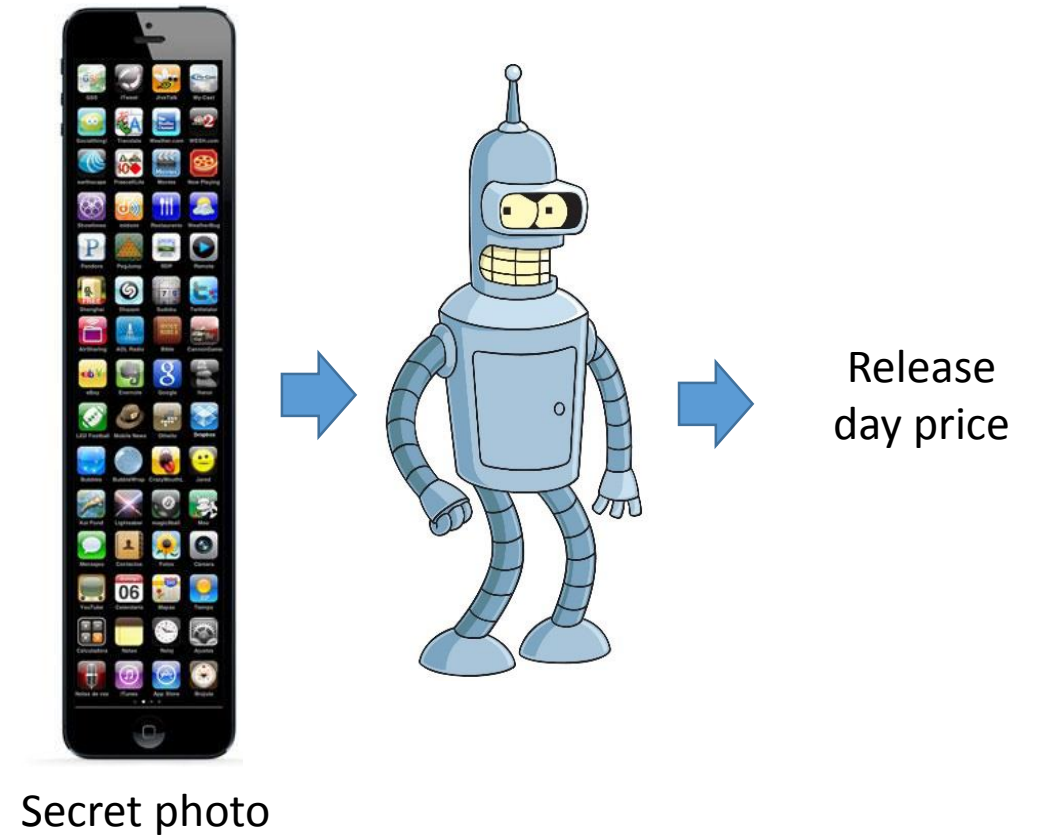
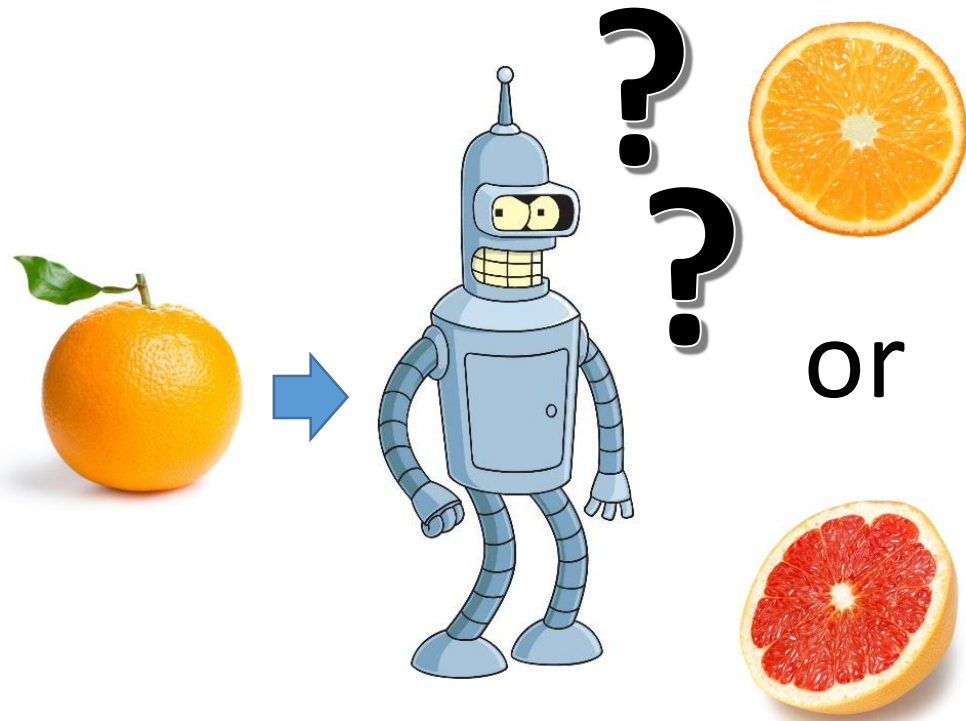
Types of learning

- **Supervised** – our training data comes from both X and Y
 - **Classification (This is what we are going to discuss)**
 - Regression
 - ...
- **Unsupervised** – training data comes only from X
 - Clustering
 - Density estimation
 - Topic modelling
- **Reinforcement** – similar to supervised, but too long to explain 😊
- ...

Supervised Learning

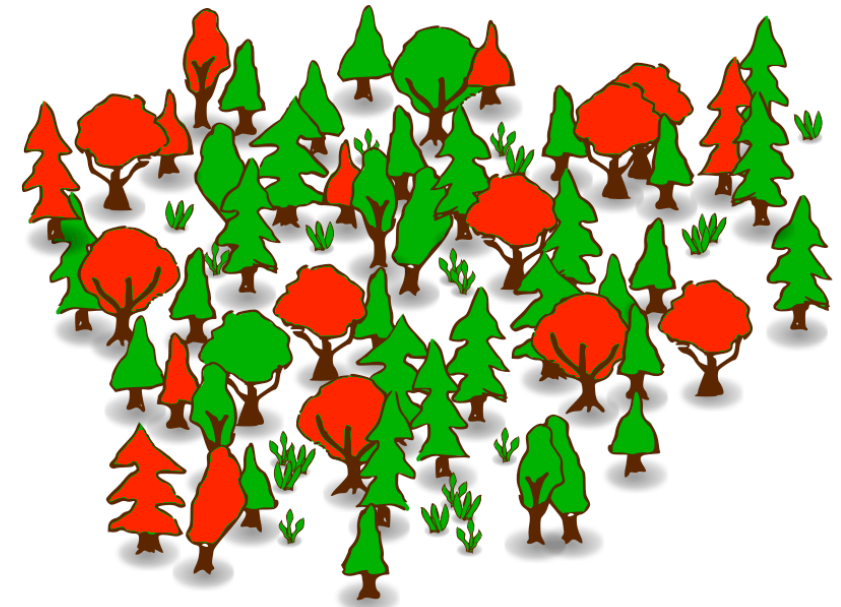


Classification vs. regression



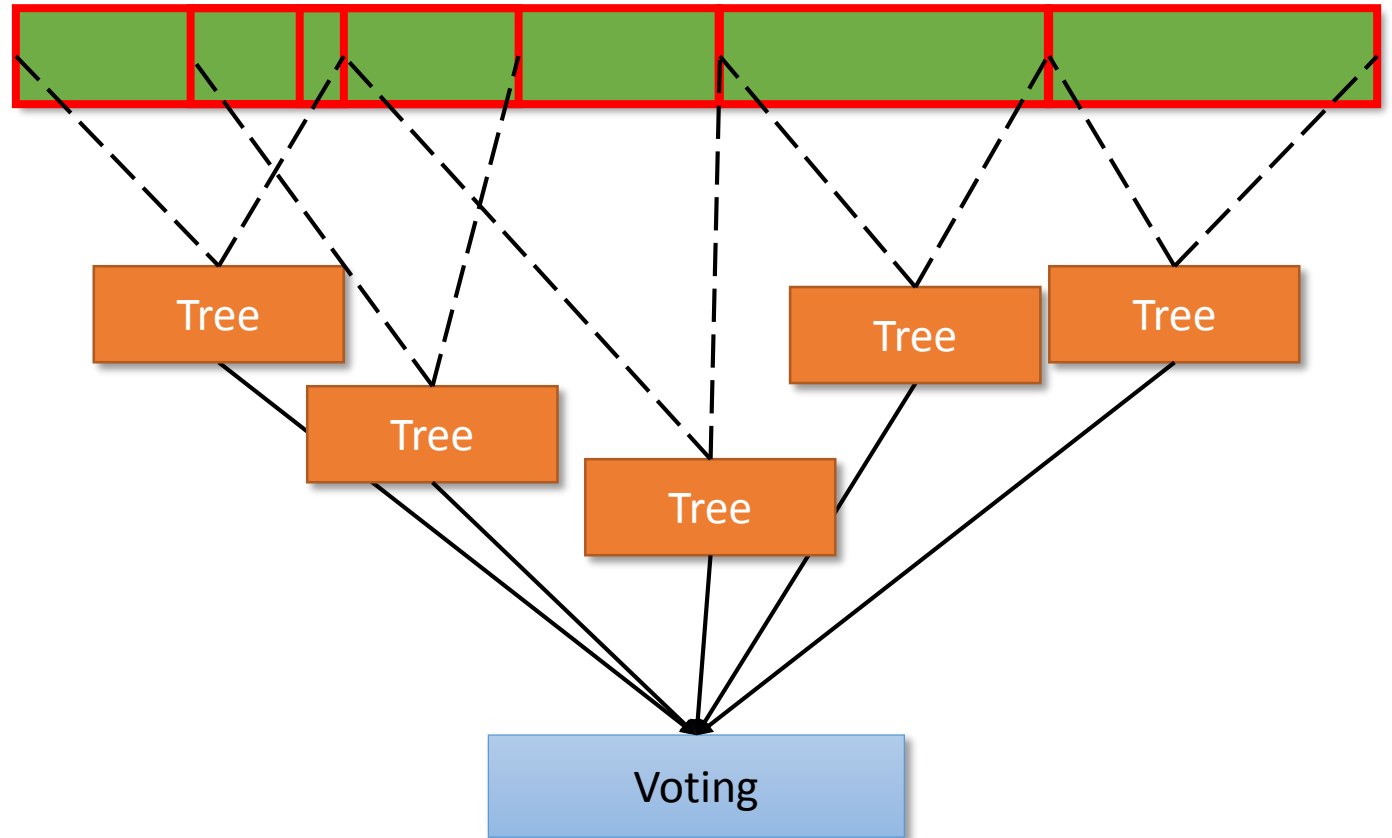
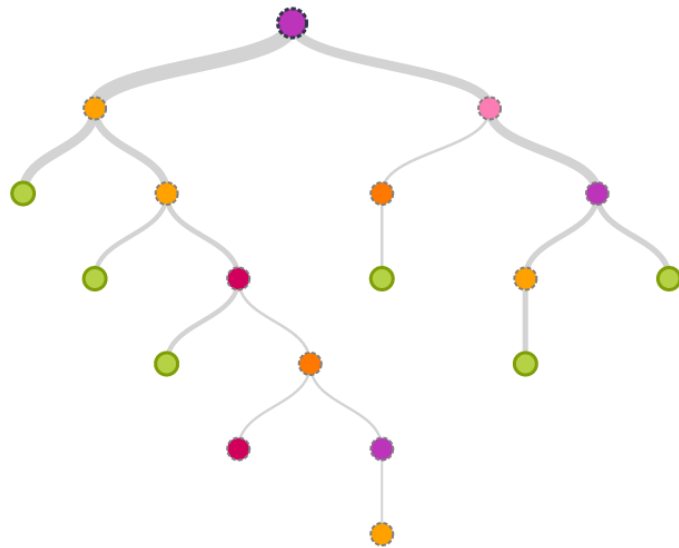
Random Forest

- Very powerful algorithm
 - Works for
 - **Classification**
 - Regression
 - ...
- Why forest?
 - Because **the main component is a tree**
- Why random?
 - Each tree is trained on the ***random subset*** of the data
- Each individual tree gives a prediction
 - Eventually they are combined
 - Using majority vote for example



Random Forest

- Random data splits can overlap
- Sampling is done with repetitions
- Trees are trained independently
- The algorithm is quite robust to outliers



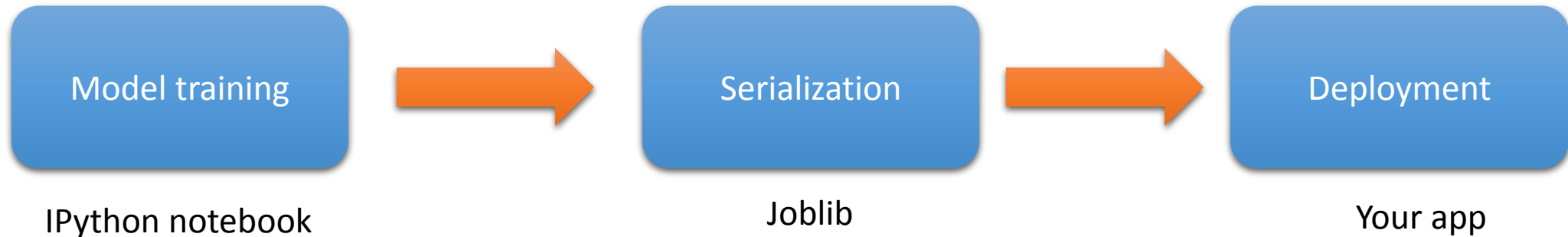
Scikit-learn: Machine Learning with Python

- Simple and efficient tools for data mining and data analysis
- Easy to use
- Built on NumPy, SciPy and matplotlib
- Open source and **commercially usable**
- Basic object – Estimator
 - They can be classifier, regressors, cluster estimators, etc
 - Basic method - .fit(..)
- Example:

```
>> clf = RandomForestClassifier(n_estimators=100)
>> clf.fit(X,y) # X - matrix data items, y - vector of labels
>> clf.predict(X_new) # X_new - new data
```

Pickling estimators

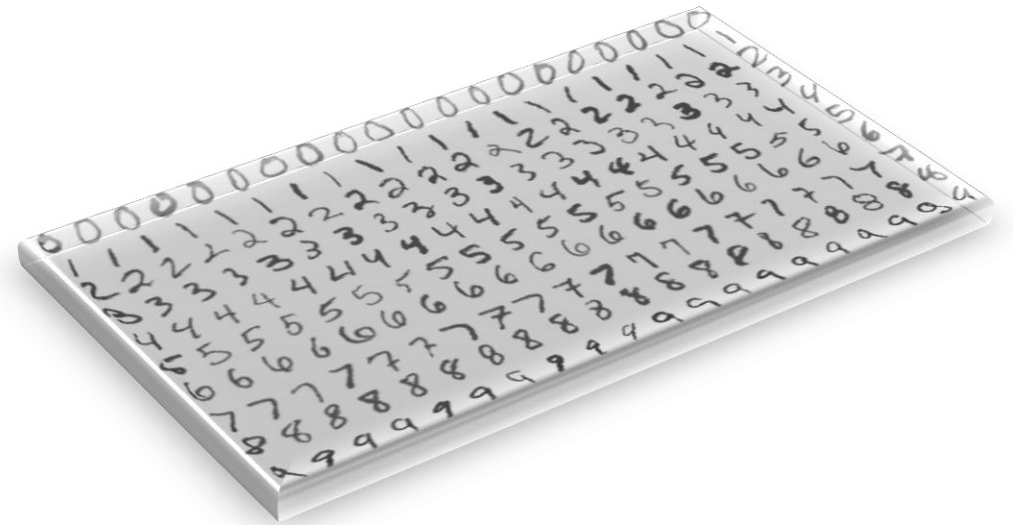
- Any object in scikit-learn can be serialized
- Joblib is recommended instead of pickle
 - More efficient for classifiers containing large NumPy arrays
 - Like Random Forest
- When the object is serialized we can load it from any Python application



Case study: MNIST digits recognition

- Dataset containing 60 000 training examples of handwritten digits
- 10 000 test examples
- 28x28 pixels pictures
- We will build a web service with Flask and scikit-learn
 - Drawing a digit in the browser window and getting the recognition result from the backend

MNIST digits:



Typical pipeline in Machine Learning

- Getting the data
- Train/test split
- Feature extraction
 - Getting a bunch of numbers which represent our data
- Model selection
 - Hyper-parameter search
- Testing

DEMO TIME

Machine Learning with Python - from Model to Web-Service

Aleksei Tiulpin,
Research Unit of Medical Imaging,
Physics and Technology
University of Oulu

Pictures used

- <http://www.supanova.com.au/wp-content/uploads/2012/09/bender-for-website.jpg>
- <https://s-media-cache-ak0.pinimg.com/originals/45/48/5a/45485ab0b9f02019c3d5df7dfcb46d92.png>
- <http://cdn1-www.comingsoon.net/assets/uploads/gallery/terminator-genisys-1407342180/terminator-genisys-018.jpg>
- http://blog.yhat.com/static/img/a_random_forest.png
- http://magazine.foxnews.com/sites/magazine.foxnews.com/files/Orange_1.jpg
- <http://www.26fruits.co.uk/images/full-fruit/grapefruit.jpg>
- <http://cdn.smosh.com/sites/default/files/bloguploads/funny-iphone-5-10-long.jpg>