Say hello to machine learning in Python (with scikit-learn)

Michal Mikolajczak (mkm0796@gmail.com)
Dominik Pieczynski (dominik.pieczynski@gmail.com)

Before we start...

The most important thing (WiFi):

AP: **Lawendowa**
Pass: **lawendowa2015**

Workshop material link:
https://github.com/PUTvision/MLWorkshopPyCon2018

Before we start...

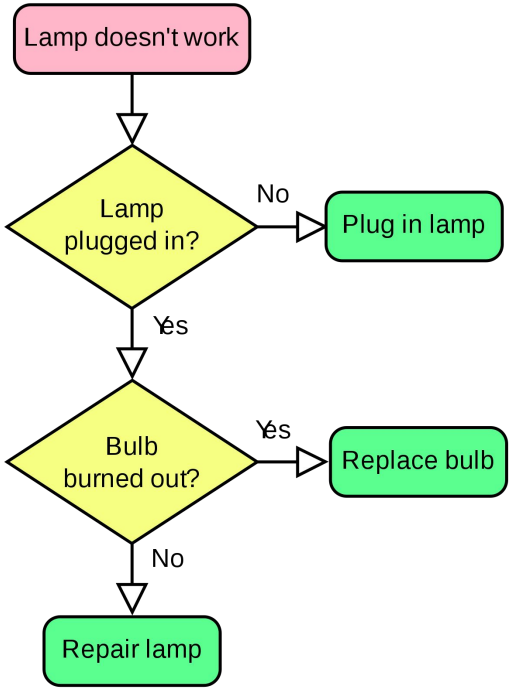
Important note: it is a basic level tutorial,and this workshop goal is for attendants to learn something (hopefully, as much as possible :))

Don't be afraid to ask questions - they are highly encouraged!

Workshop outline

1. Introduction - what is machine learning and how it can be useful?
2. Supervised learning - classification.
3. Supervised learning - regression.
4. Stand alone project.

Machine learning - what is it exactly?



Machine learning - what is it exactly?

Is this a dog?

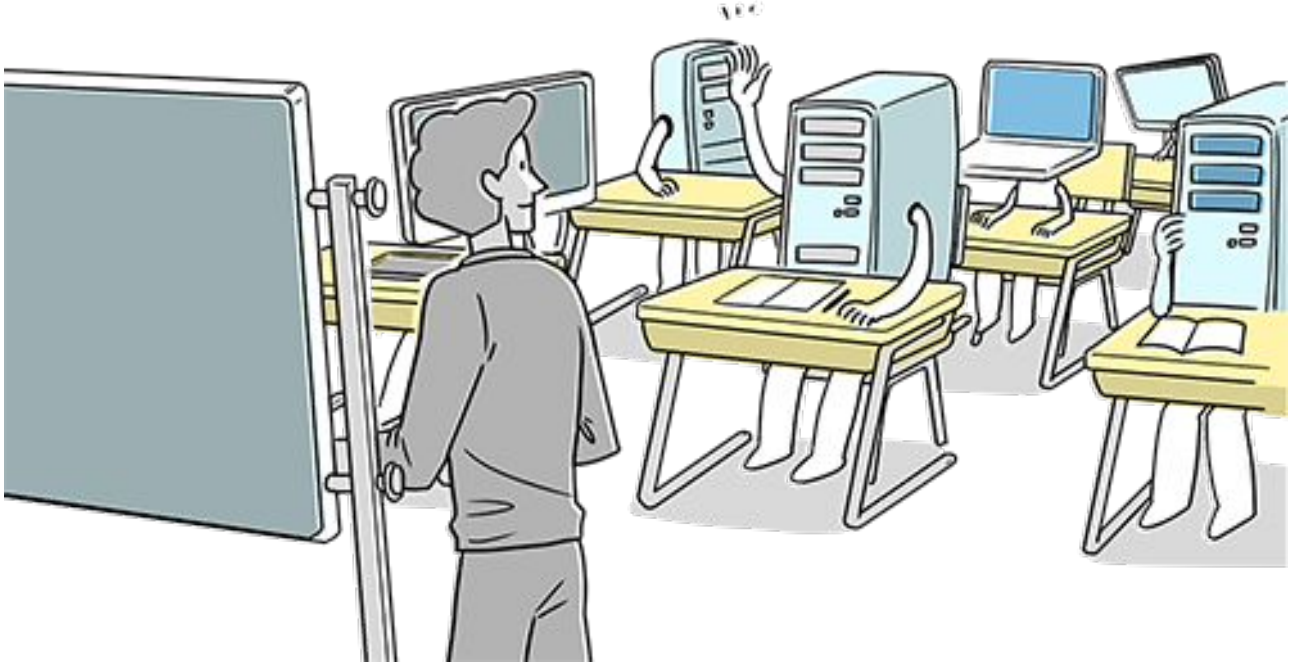
Machine learning - what is it exactly?

Is this a cat?

Machine learning - what is it exactly?

What about the computer?

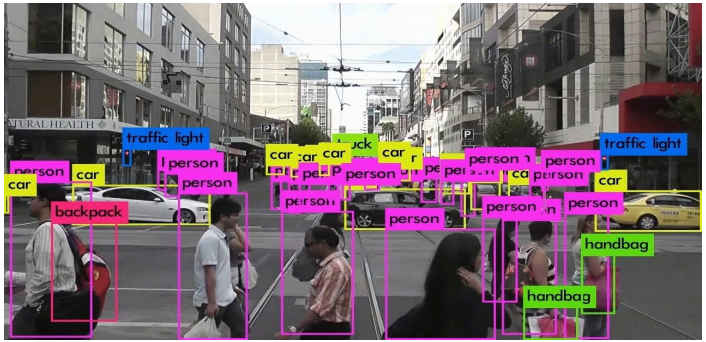
Machine learning - what is it exactly?

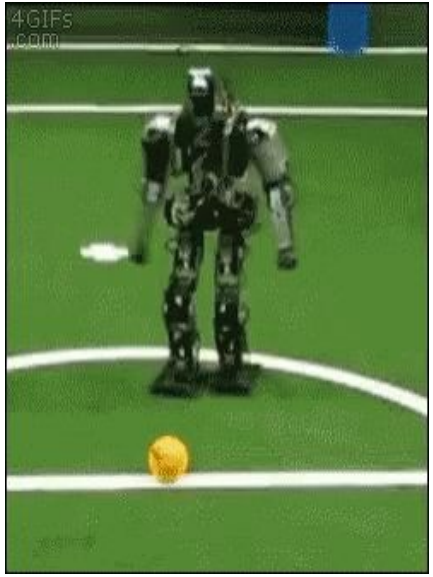


Machine learning - what is it exactly?

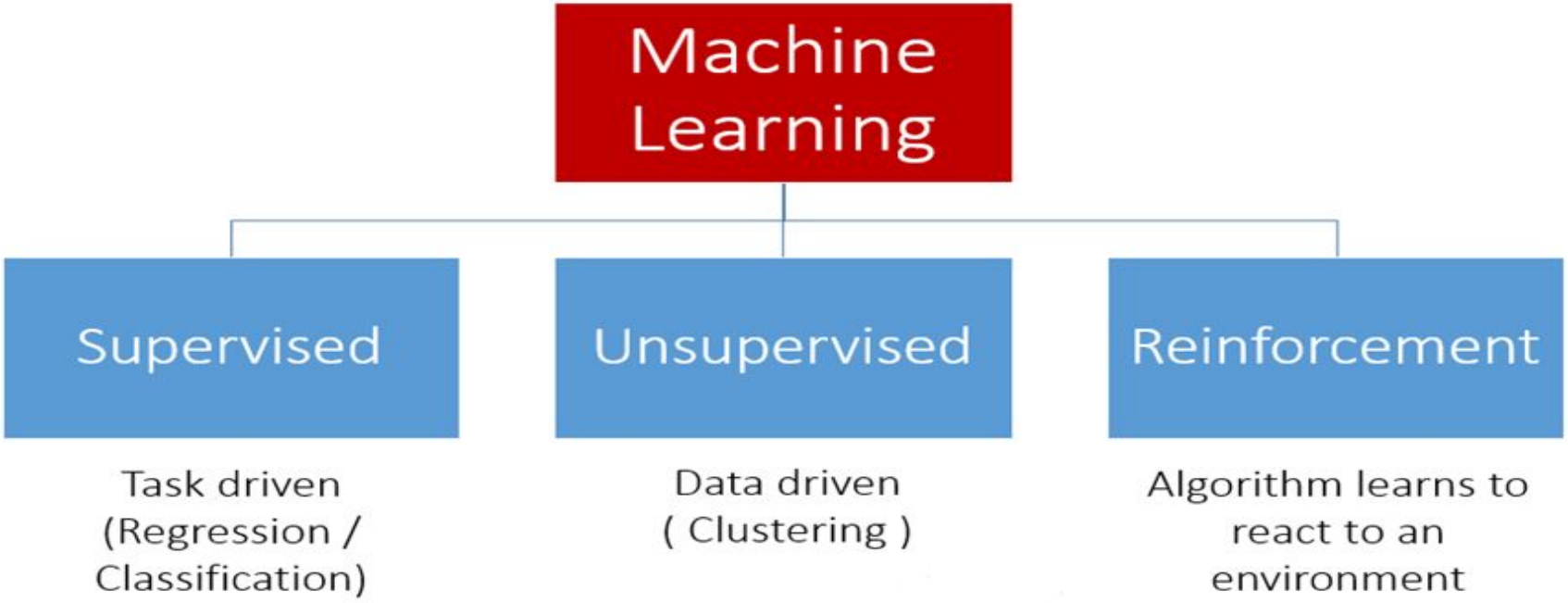
Machine learning is the science of programming computers so they can learn from data - without being explicitly programmed.

ML - how it can be useful?



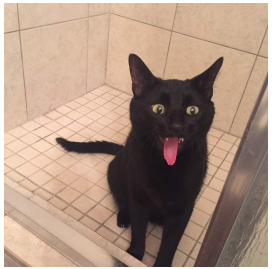


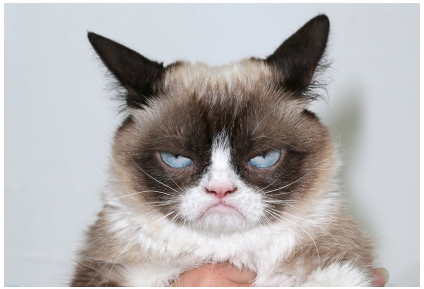
Types of machine learning



Supervised learning

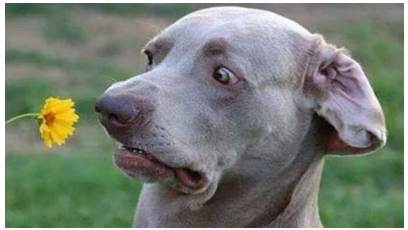
Cats



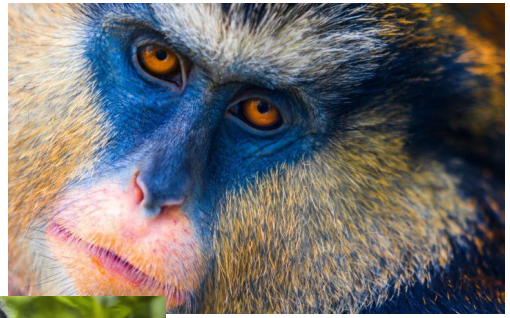
Dogs





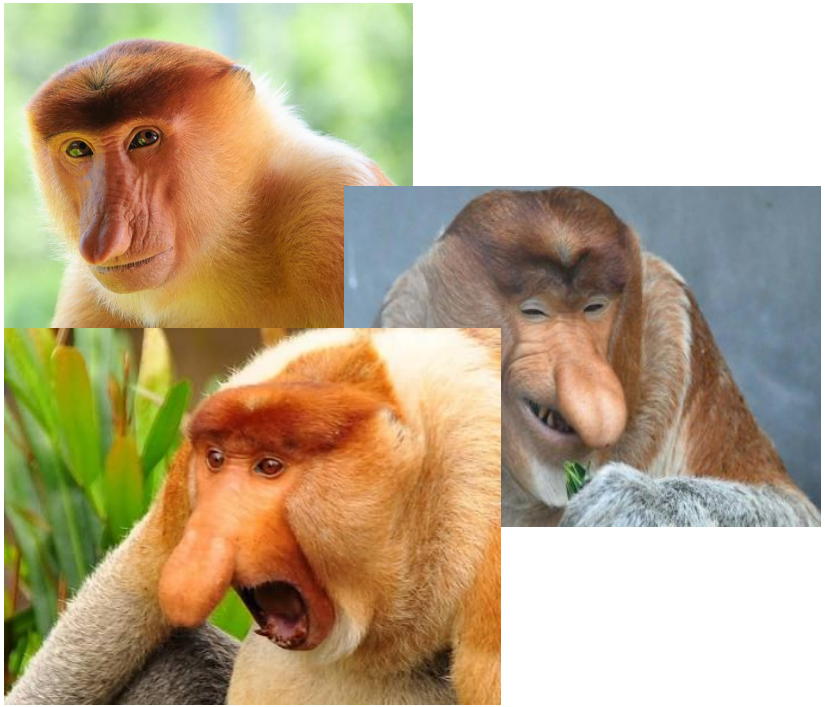

Unsupervised learning



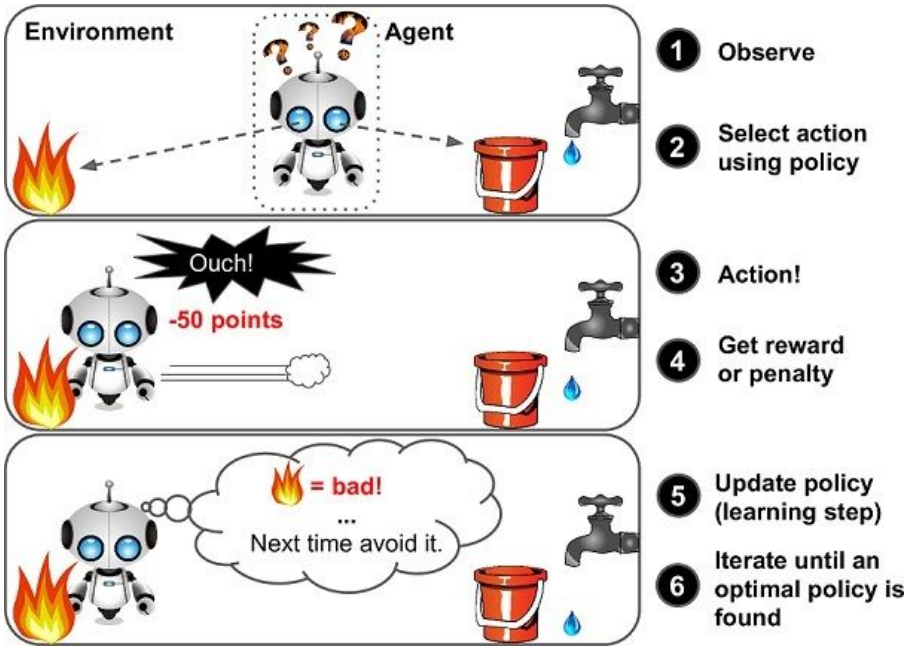




Unsupervised learning



Reinforcement learning



We will only work with simple examples of supervised learning

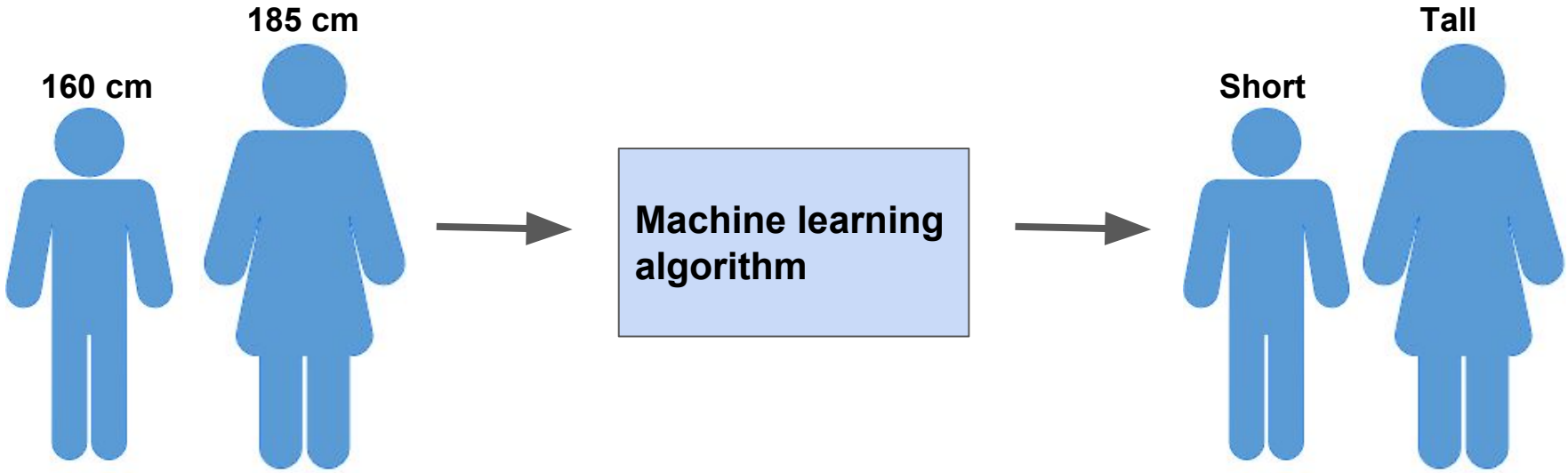
Why scikit-learn?

- Fast, efficient implementations of ML algorithms.
- Coherent, user friendly interfaces.
- Popular, actively maintained.

Workshop outline

1. Introduction - what is machine learning and how it can be useful?
2. Supervised learning - classification.
3. Supervised learning - regression.
4. Stand alone project.

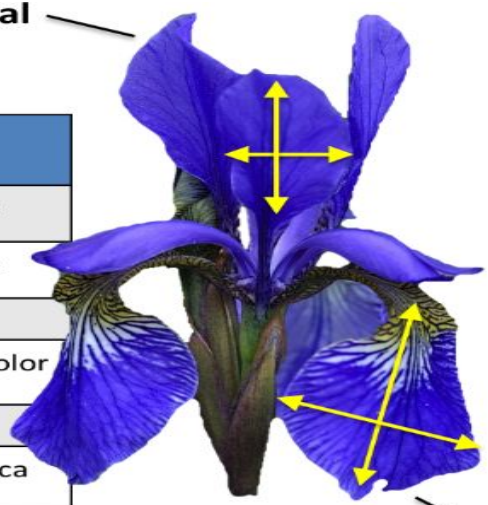
Classification problem



Data - features and labels

Samples
(instances, observations)

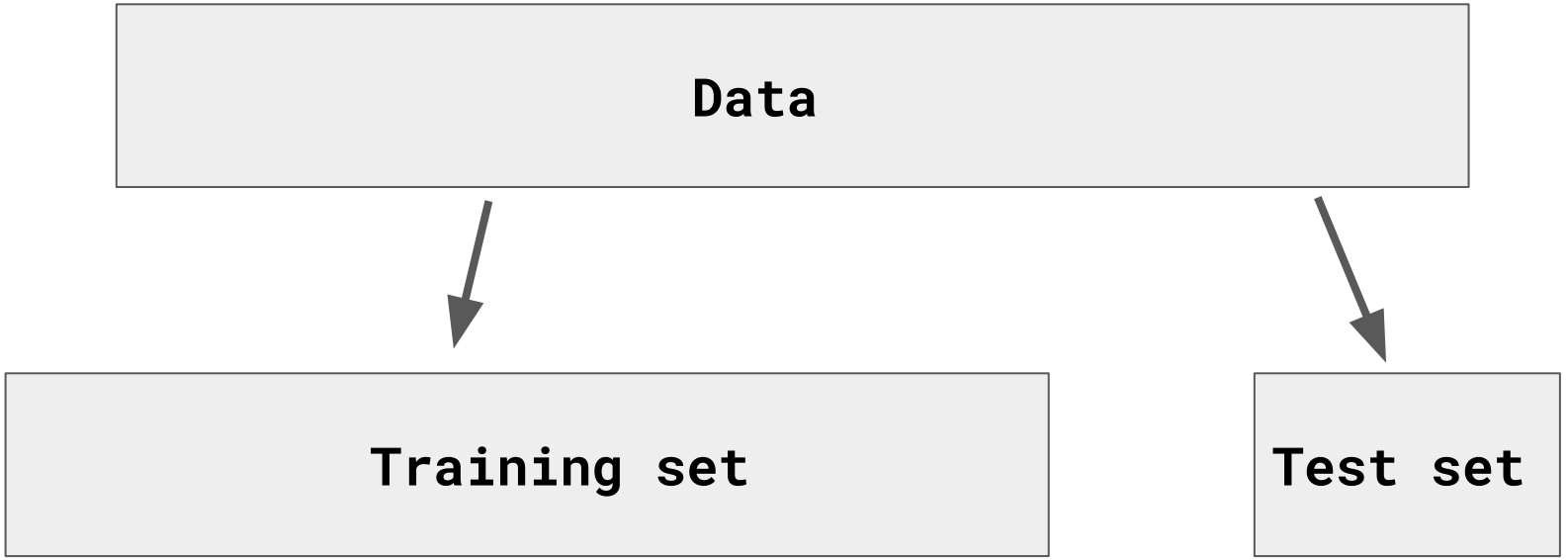
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Train/test splitting



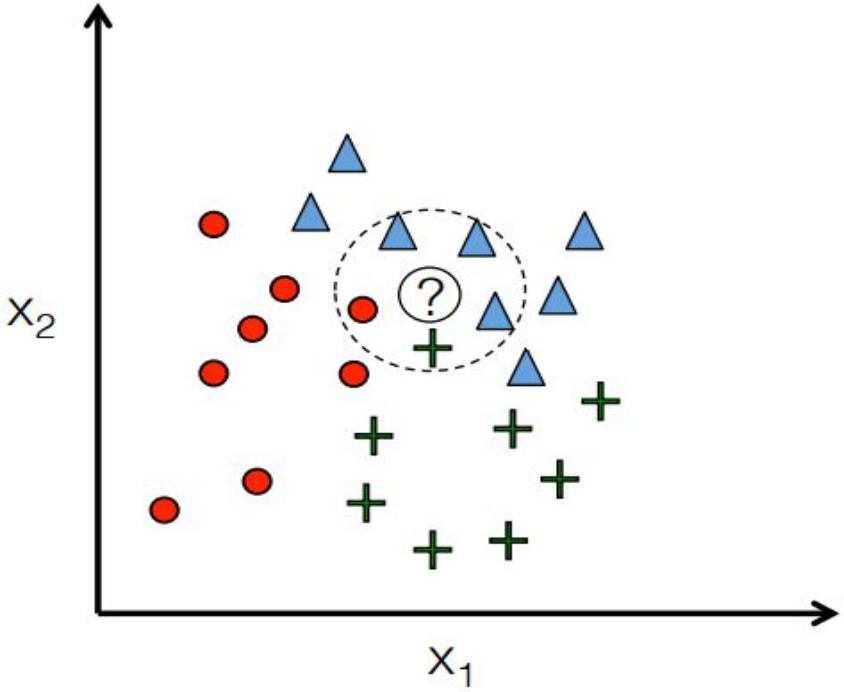
Used for model fitting, feature selection, parameter tuning.

Final tests, not used in training.

Train/test splitting

Performance on training data vs test data

K-Nearest Neighbors overview



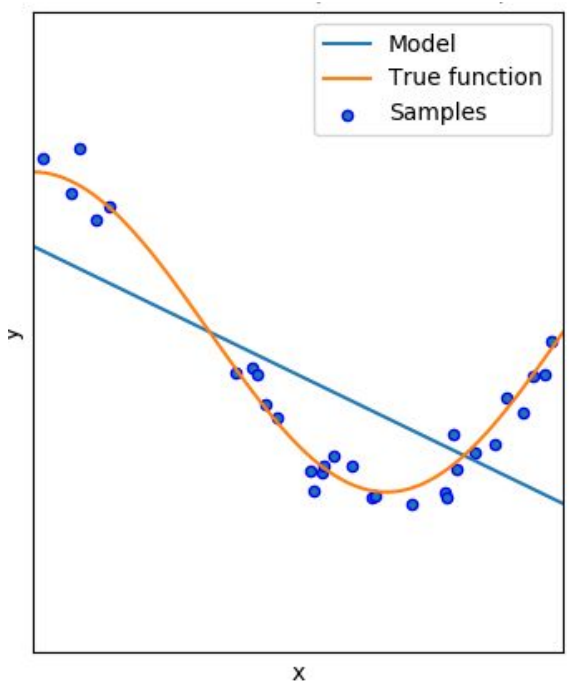
Neighborhood

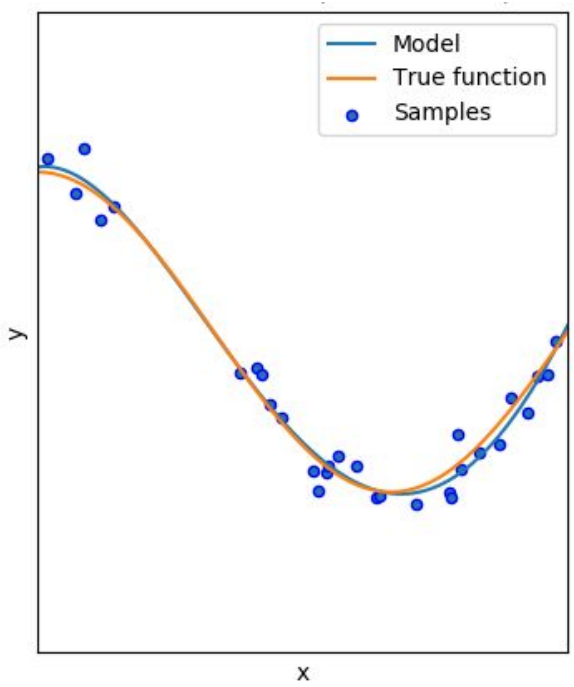
1 × +
1 × ●
3 × ▲

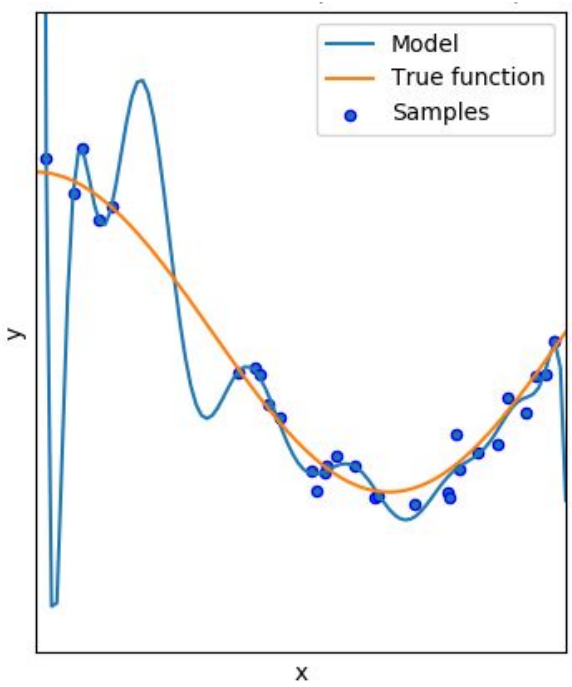
Prediction

? = ▲

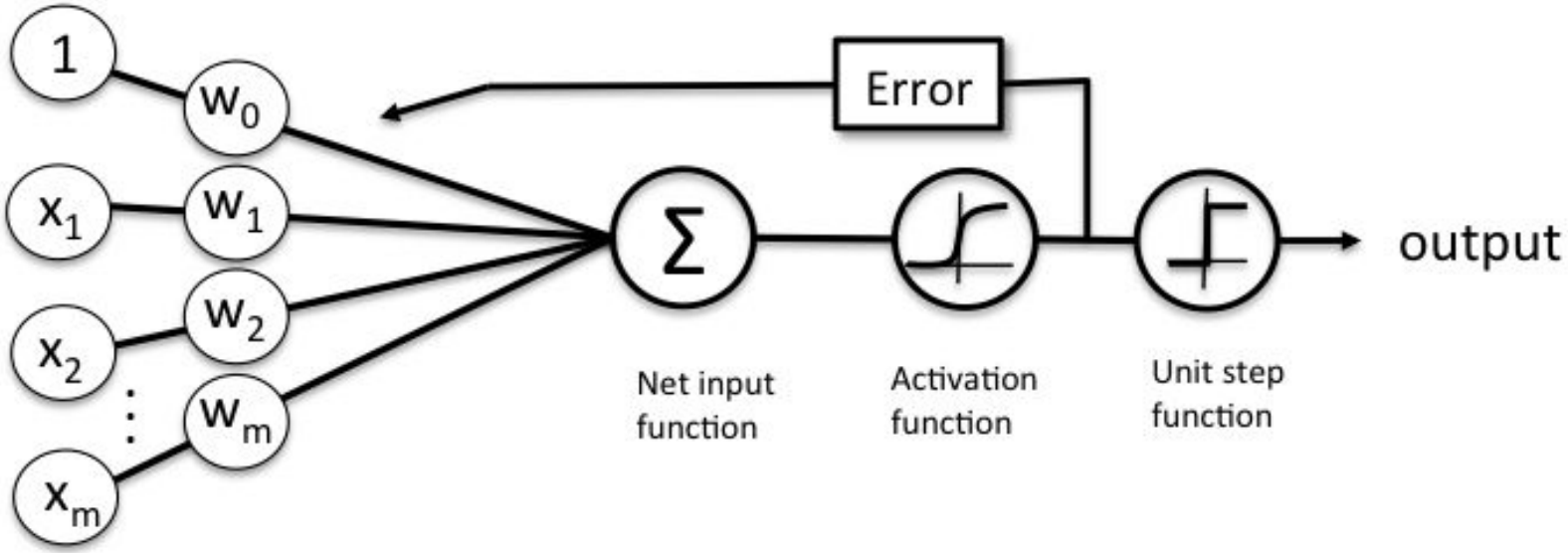
Underfitting and overfitting







Logistic regression overview



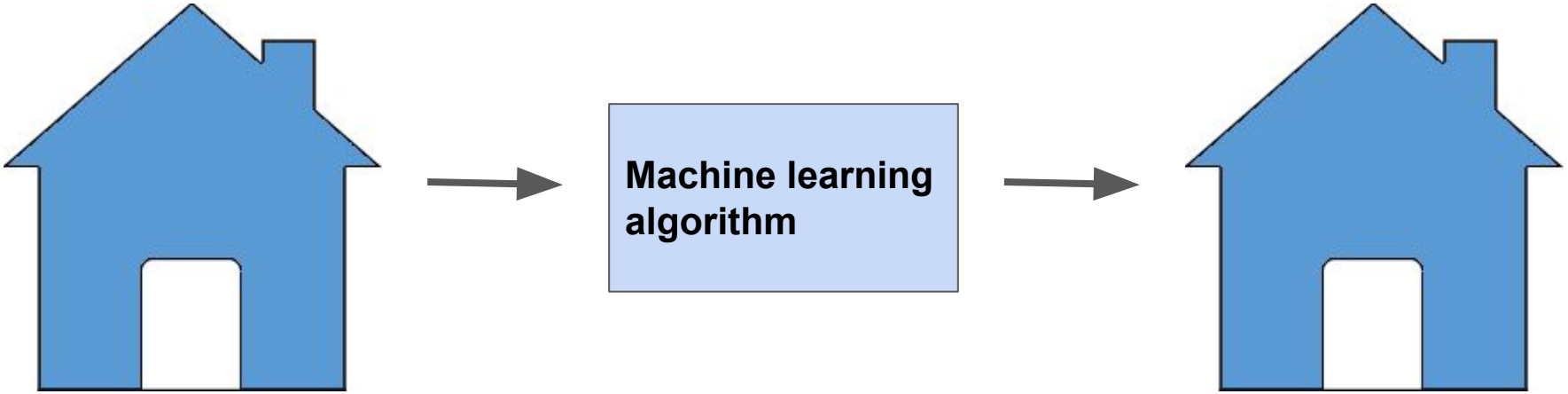
Workshop outline

1. Introduction - what is machine learning and how it can be useful?
2. Supervised learning - classification.
3. Supervised learning - regression.
4. Stand alone project.

Regression problem

Square meters: 50
Condition: good
Location: NY

Price: 100 000$
(Continuous value)



Data preparation - missing values

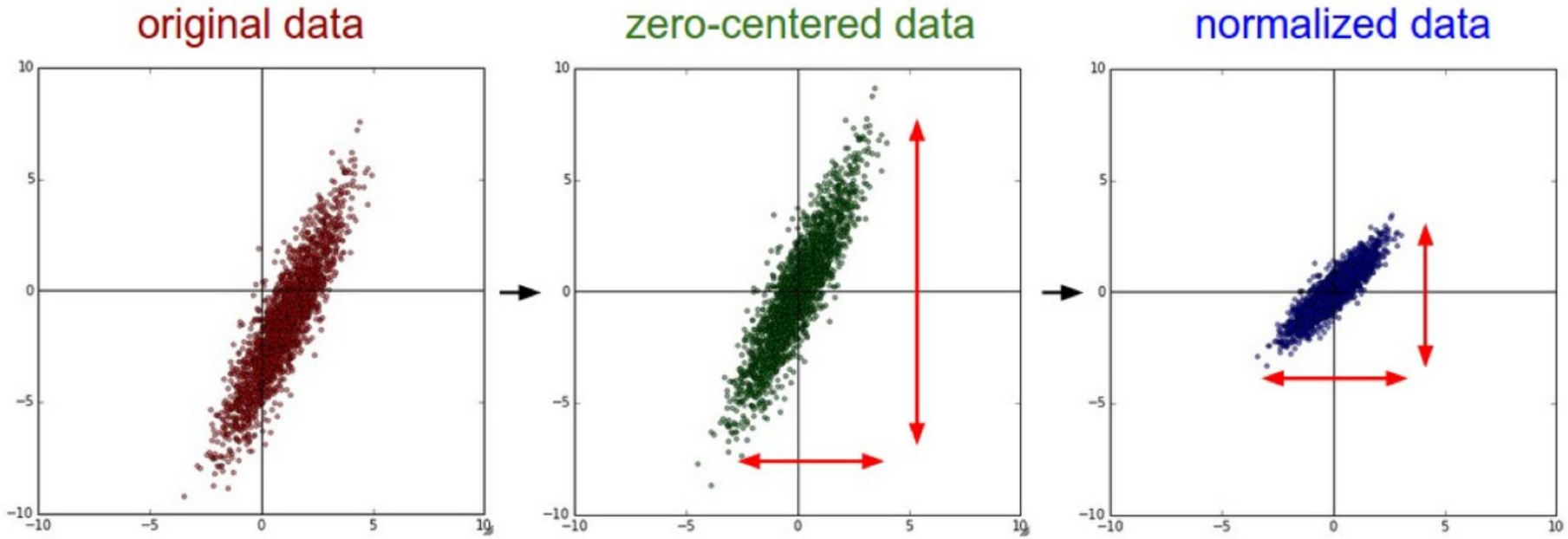
ID	Universe	Comics count	Vilains defeated	FanRating	HeroName
1	Marvel	237	13	9.1	IronMan
2	Marvel	189	16	N/A	Captain America
3	Marvel	165	N/A	8.4	Thor
4	Marvel	172	25	8.3	Hulk
5	Marvel	109	10	9.5	Dr. Strange
6	N/A	N/A	N/A	N/A	Batman
7	DC	45	4	7.9	Superman

Data preparation - categorical variables

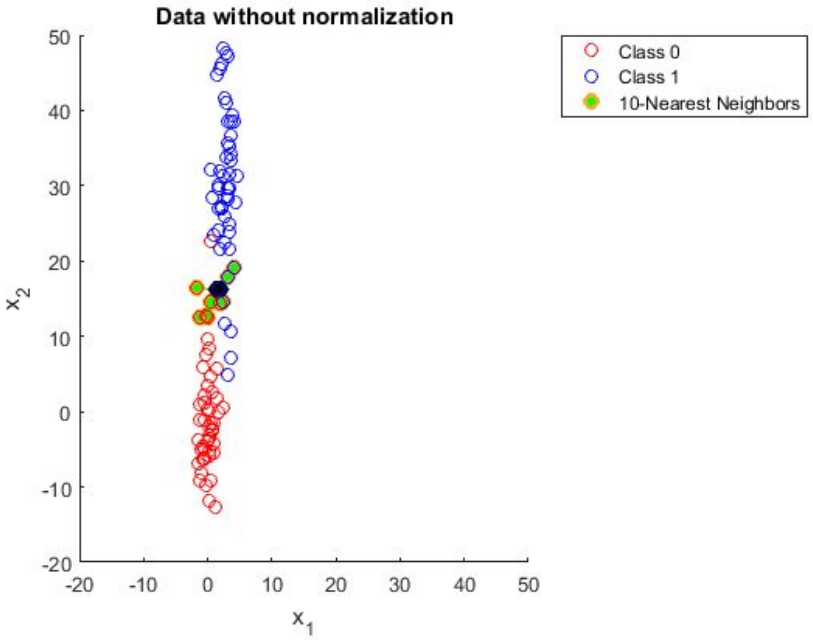
Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

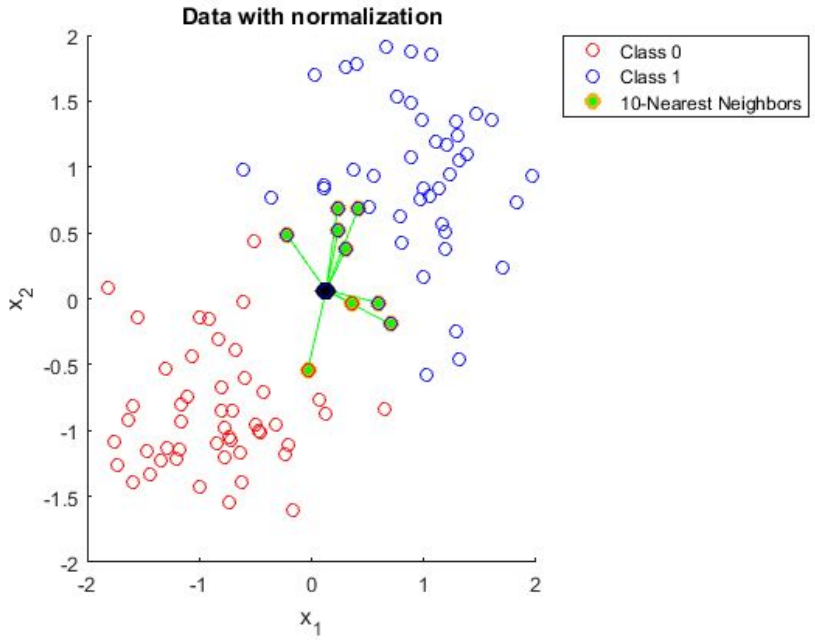
Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Data preparation - feature scaling

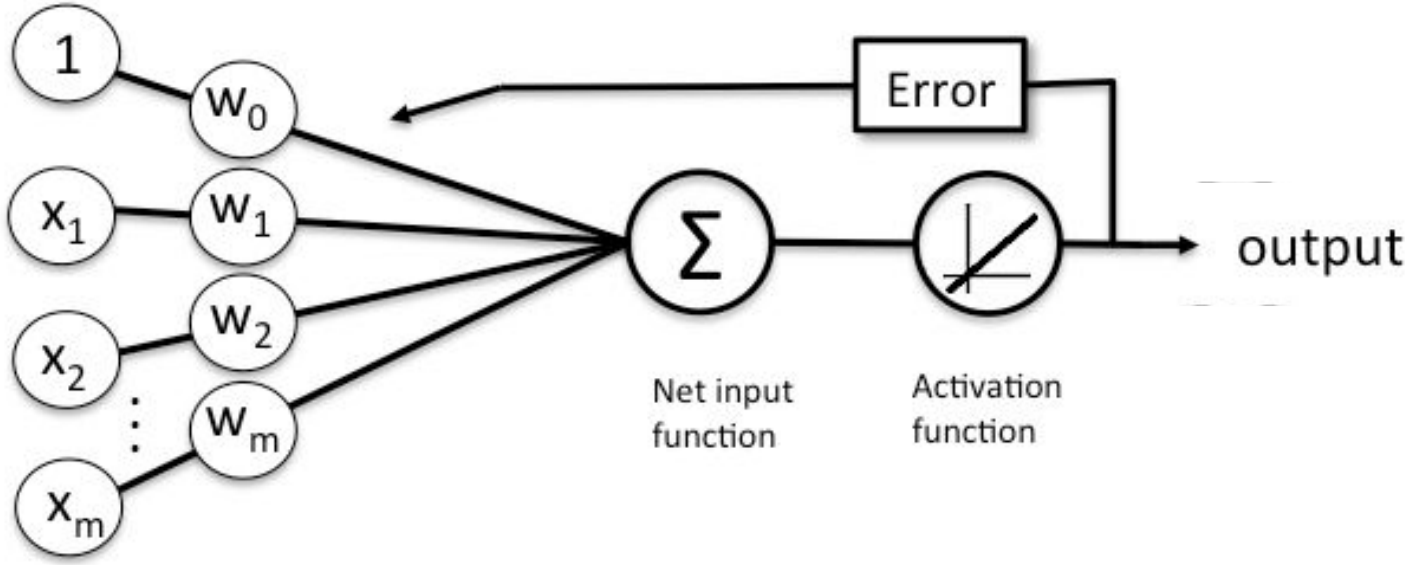


Data preparation - feature scaling





Linear regression overview



Workshop outline

1. Introduction - what is machine learning and how it can be useful?
2. Supervised learning - classification.
3. Supervised learning - regression.
4. Stand alone project.

Stand alone project

Appendix A - further resources

- Scikit-learn docs: **http://scikit-learn.org/stable/**
- Kaggle ML tasks and datasets: **https://www.kaggle.com**
- “Python Machine learning” by Sebastian Raschka, often available for free on: https://www.packtpub.com/packt/offers/free-learning
- Also, something from Google (other than search bar):
  - https://developers.google.com/machine-learning/crash-course/
  - https://www.tensorflow.org/tutorials/