SAP CONCUR Data Day

Master Class: Getting Started with ML in Python

Training, evaluating and exporting a classification model

SAP CONCUR Data Day

Everaldo Aguiar

Principal Data Scientist

Data Science, Spend

Some Preliminaries

Preliminaries

- **What is Data Science?**
- **DS Applications**
- **DS Tasks**

- **Join #dataday19_intro_to_ml**
- Pre-download Miniconda (just in case)

Our Definition

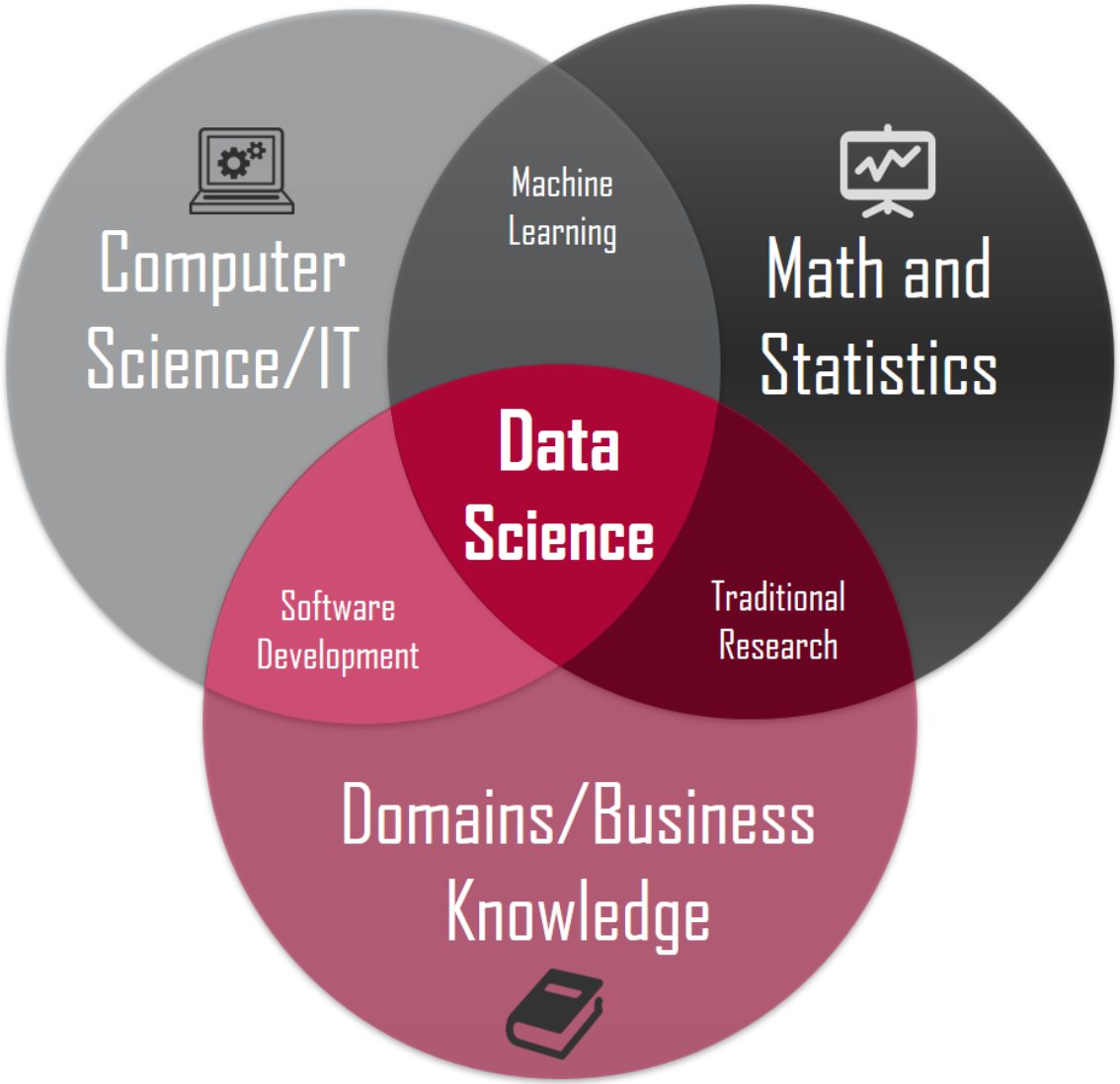
Data Science *is the art of extracting knowledge from large bodies of data.*

Two Key Features

1. Extracting Knowledge

2. Large Bodies of Data

It's a Confluence of Domains



Let’s look at a basic example of **Data Science**.

Rock, Paper, Scissors

Rock, Paper, Scissors

1

Rock-it:

Males have the tendency to produce rock on their first throw. If you are playing against one, try using paper

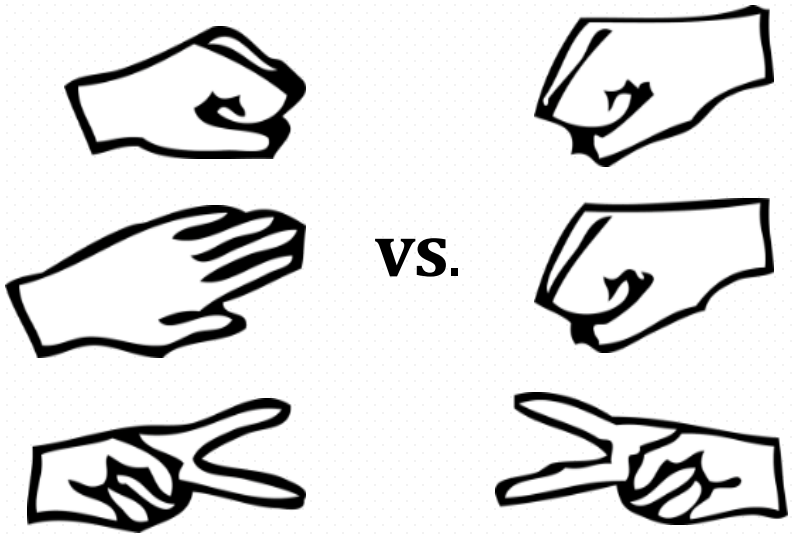
Based on: http://flowingdata.com/2010/07/30/how-to-win-rock-paper-scissors-every-time/

Rock, Paper, Scissors

2

Double on the Rocks:

When you see a two-Rock run, it is highly likely that your opponent's next move will be Scissors or Paper. People dislike being predicable. Counter with rock.



Based on: http://flowingdata.com/2010/07/30/how-to-win-rock-paper-scissors-every-time/

Rock, Paper, Scissors

3

Paper Please:

Paper is thrown the least in a match. Use it as an unexpected options.

Paper is thrown 29.6% of the time.

Rock is thrown 35.4% of the time.

Scissors is thrown 35% of the time.

Based on: http://flowingdata.com/2010/07/30/how-to-win-rock-paper-scissors-every-time/

Data Science Tasks

1 Descriptive Tasks

Here, the objective is to derive patterns
(correlations, trends, clusters, trajectories, and anomalies)

2 Predictive Tasks

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes.

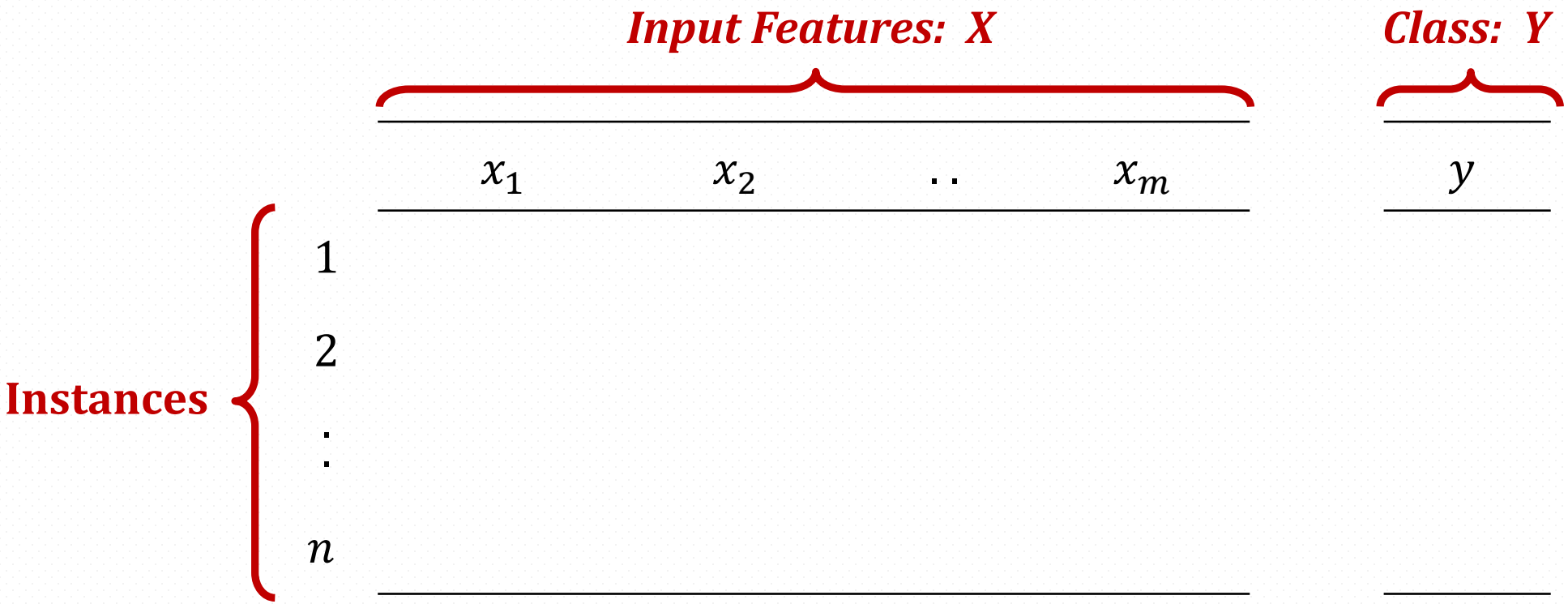
We'll be focusing on these

What's in *Data*?

What's in Data?

- Instance:
  - The "thing" to be classified, associated, or clustered.
  - Individual, independent example of target concept.
- Features:
  - Set of values that describe each instance.
  - Can be numeric, categorical, or discrete.
- Class:
  - A discrete value that "labels" individual instances.

Instances, Features, and Classes



Instances, Features, and Classes

Make	Cylinders	Length	Weight	Style
Honda	Four	150	1956	Hatchback
Toyota	Four	167.9	2280	Wagon
BMW	Six	176.8	2765	Sedan

Given car make, cylinders, length, and weight, learn a function for the body style.

Instances, Features, and Classes

Temperature	Wind Speed	Decision
80°	Low	Bike Day
40°	Low	Couch Day
60°	Medium	Couch Day
80°	High	Bike Day

Go out and bike or laze on the couch.

Supervised Learning

Temperature	Wind Speed	Decision
80°	Low	Bike Day
40°	Low	Couch Day
60°	Medium	Couch Day
80°	High	Bike Day

1. If wind-speed = high, *then* Bike Day
2. If wind-speed = medium, *then* Couch Day
3. If wind-speed = low and temp ≤ 40, *then* Couch Day
4. If wind-speed = low and temp > 40, *then* Bike Day

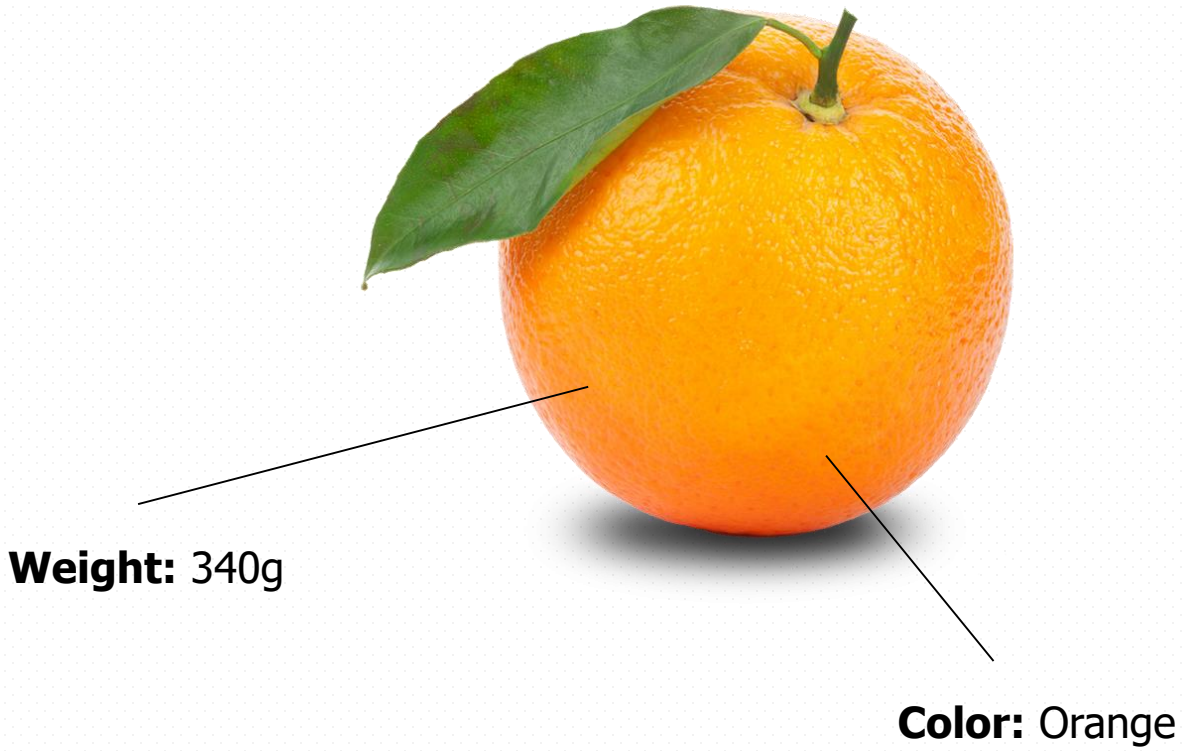
Now Consider this Problem

- An advertising company wants to group customers based on similarities and past behavior to help with their marketing efforts. There are no predefined labels.
- What is this?
  - An example of **Unsupervised Learning**.
  - No predefinition of groups, a.k.a. classes.
  - Find similarities in data based on features.
  - This is the simplistic view of clustering.

Features (in supervised learning)

Features (or attributes) are used to train an ML system.

They are properties of the things you are trying to learn about.

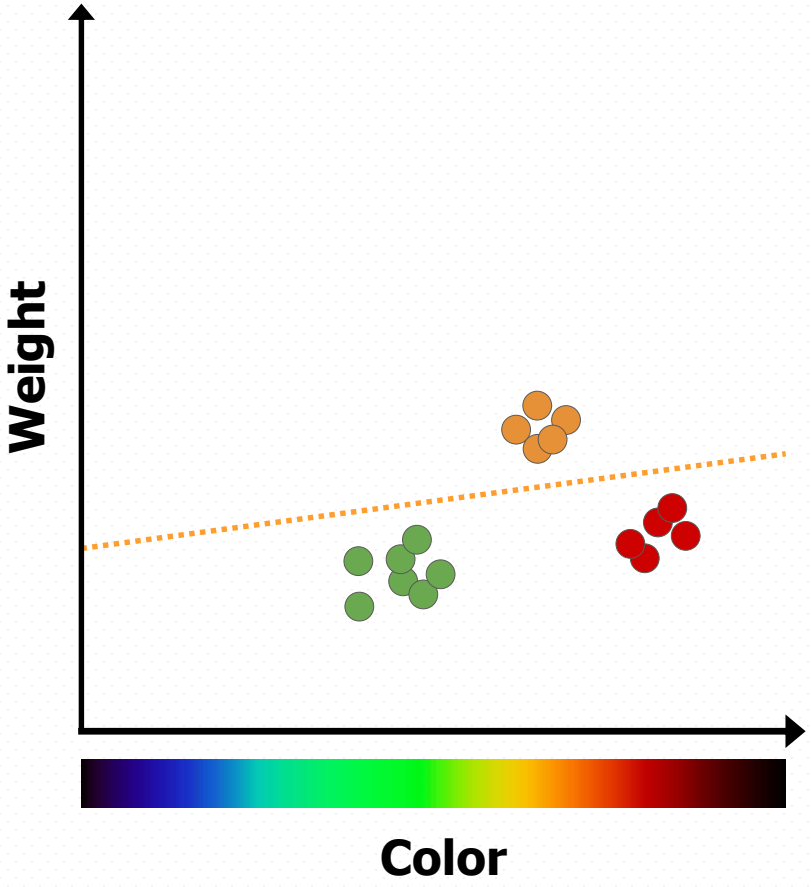


Features (in supervised learning)

Consider fruits. Features of fruits might be weight and color. **2 features would mean there are 2 dimensions.**

A 2D system may be plotted on a graph if features are represented numerically.

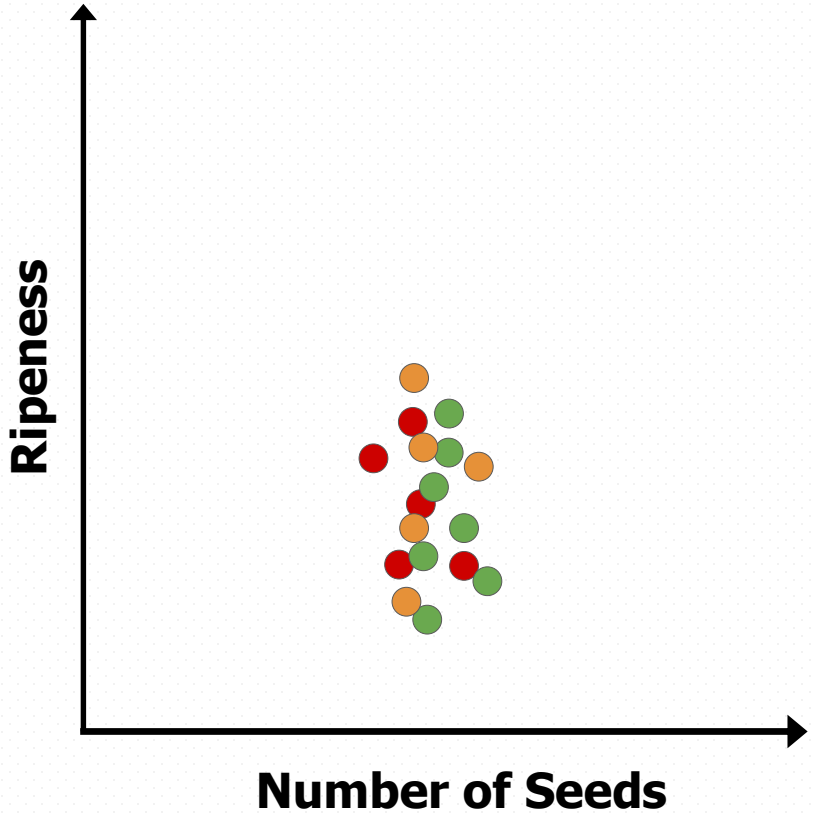
Plotting apples (red and green dots) and oranges (orange dots) weight and color.



Features (in supervised learning)

Choosing useful input features can have a big impact on the quality of the ML system.

Some features may not be useful enough to separate the data.



Supervised Learning

The ML program is provided with training data that is labeled. You specify how to categorize the example data. For example:

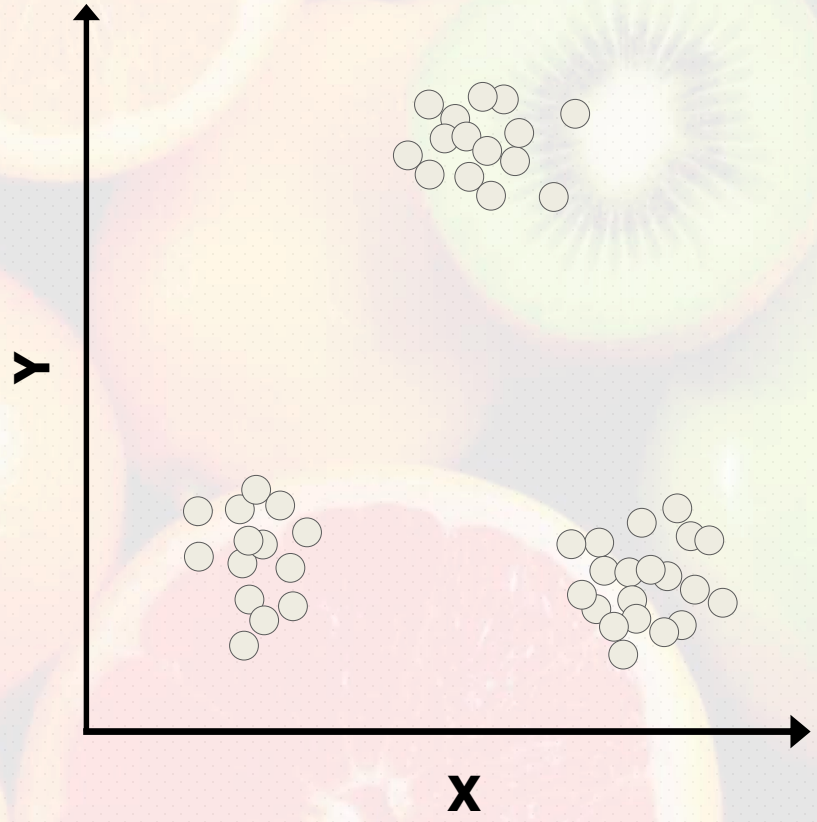
Color	Weight	Label
Red	200g	Apple
Orange	300g	Orange
Green	150g	Apple
…		

Unsupervised Learning

The ML program **must learn from an unlabeled data set**.

Imagine we had a clumps of points that represent 3 different things.

The program must learn that there are 3 distinct clusters and categorize them.



What is Machine Learning?

Machine Learning Definition

Field of study that gives computers the ability to ***learn without being explicitly programmed****.*

—Arthur Samuel (1959)

A Different Way of Doing Things

Traditional Approach

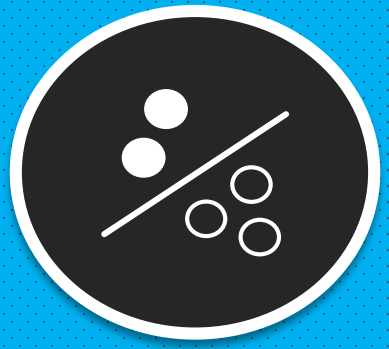
Program with **explicit rules** to follow:

```
if email contains n!@gerian prince:
  then mark spam;
if email contains …
if email contains …
```

ML Approach

Program that **learns from examples**:

```
do train:
  classify email;
  identify errors;
  adjust self to reduce error;
  if no errors:
    end
```

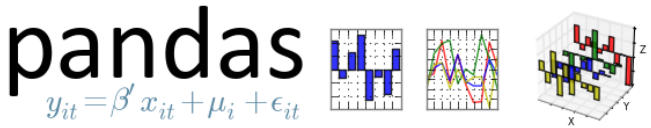


Let's train a classifier

Predicting Diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
4	183	0	0	0	28.4	0.212	36	1
5	162	104	0	0	37.7	0.151	52	1
2	197	70	99	0	34.7	0.575	62	1
13	158	114	0	0	42.3	0.257	44	1
0	162	76	56	100	53.2	0.759	25	1
0	57	60	0	0	21.7	0.735	67	0
0	101	76	0	0	35.7	0.198	26	0
1	90	62	18	59	25.1	1.268	25	0
5	117	86	30	105	39.1	0.251	42	0
1	124	74	36	0	27.8	0.1	30	0

What we'll be using



- Install Anaconda (Python 3): https://www.anaconda.com/distribution/#download-section
- Install the required packages
- In our terminal:

```
pip install scikit-learn==0.21.2
pip install pandas
pip install jupyter
pip install joblib
```

- Clone: https://github.com/anfibil/dataday19_intro_to_ml
- Start Jupyter and navigate to the `Classification Template` notebook