CERN

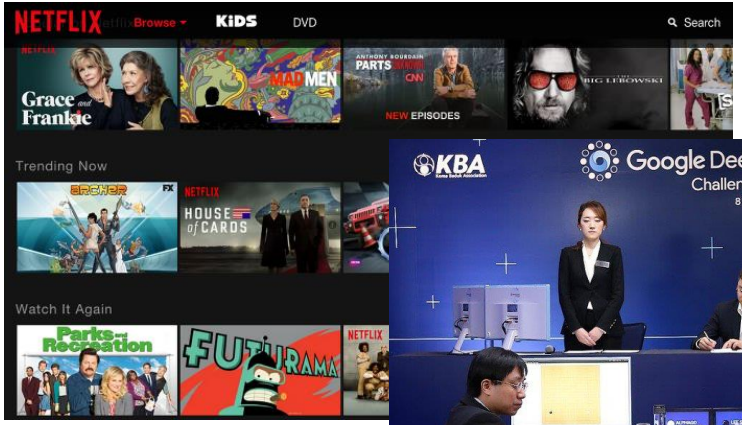
Machine Learning with Spark MLlib

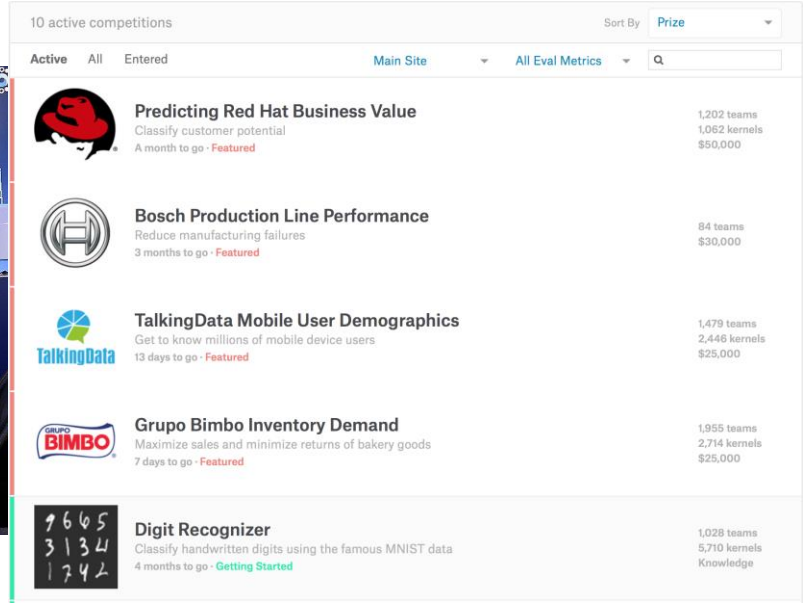
Manuel Martín Márquez
Antonio Romero Marin
Joeri Hermans
Hadoop Tutorials

Machine Learning (ML)

- ML is a branch of artificial intelligence:
  - Uses computing based systems to make sense out of data
    - Extracting patterns, fitting data to functions, classifying data, etc
  - ML systems can learn and improve
    - With historical data, time and experience
  - Bridges theoretical computer science and real noise data.

ML in real-life

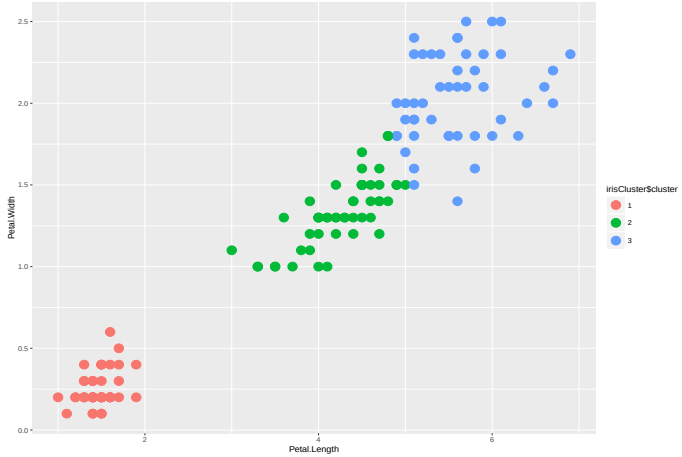




Supervised and Unsupervised Learning

- Unsupervised Learning
  - There are not predefined and known set of outcomes
  - Look for hidden patterns and relations in the data
  - A typical example: Clustering

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1



Supervised and Unsupervised Learning

- Supervised Learning
  - For every example in the data there is always a predefined outcome
  - Models the relations between a set of descriptive features and a target (Fits data to a function)
  - 2 groups of problems:
    - Classification
    - Regression

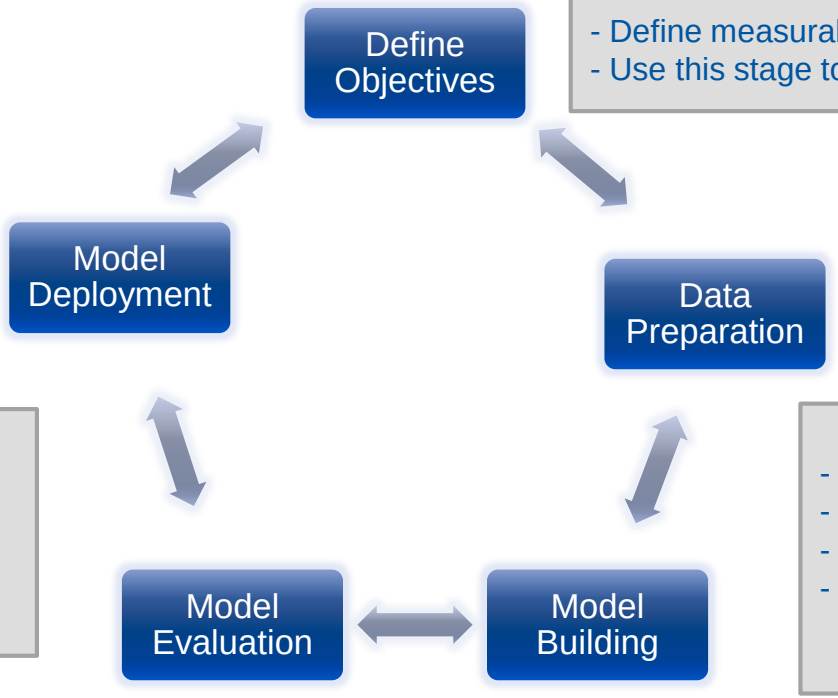
Supervised Learning

- Classification
  - Predicts which class a given sample of data (sample of descriptive features) is part of (**discrete value**).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

- Regression
  - Predicts continuous values.

Machine Learning as a Process



- Define measurable and quantifiable goals
- Use this stage to learn about the problem

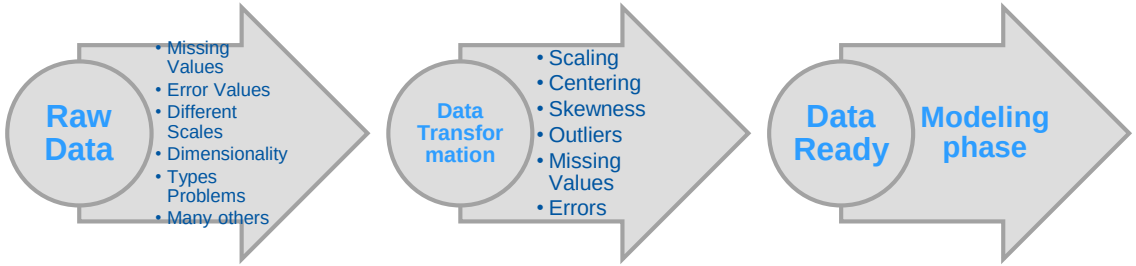
- Normalization
- Transformation
- Missing Values
- Outliers

- Data Splitting
- Features Engineering
- Estimating Performance
- Evaluation and Model Selection

- Study models accuracy
  - Work better than the naïve approach or previous system
- Do the results make sense in the context of the problem

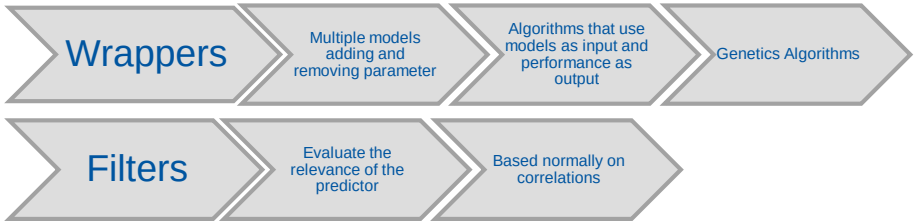
ML as a Process: Data Preparation

- Needed for several reasons
  - Some Models have strict data requirements
    - Scale of the data, data point intervals, etc
  - Some characteristics of the data may impact dramatically on the model performance
- Time on data preparation should not be underestimated



ML as a Process: Feature engineering

- Determine the predictors (features) to be used is one of the most critical questions
- Some times we need to add predictors
- Reduce Number:
  - Fewer predictors more interpretable model and less costly
  - Most of the models are affected by high dimensionality, specially for non-informative predictors



  - Binning predictors

ML as a Process: Model Building

- Data Splitting
  - Allocate data to different tasks
    - model training
    - performance evaluation
  - Define Training, Validation and Test sets
- Feature Selection (Review the decision made previously)
- Estimating Performance
  - Visualization of results – discovery interesting areas of the problem space
  - Statistics and performance measures
- Evaluation and Model selection
  - The ‘no free lunch’ theorem no a priory assumptions can be made
  - Avoid use of favorite models if NEEDED