Scikit-learn: Machine Learning in Python

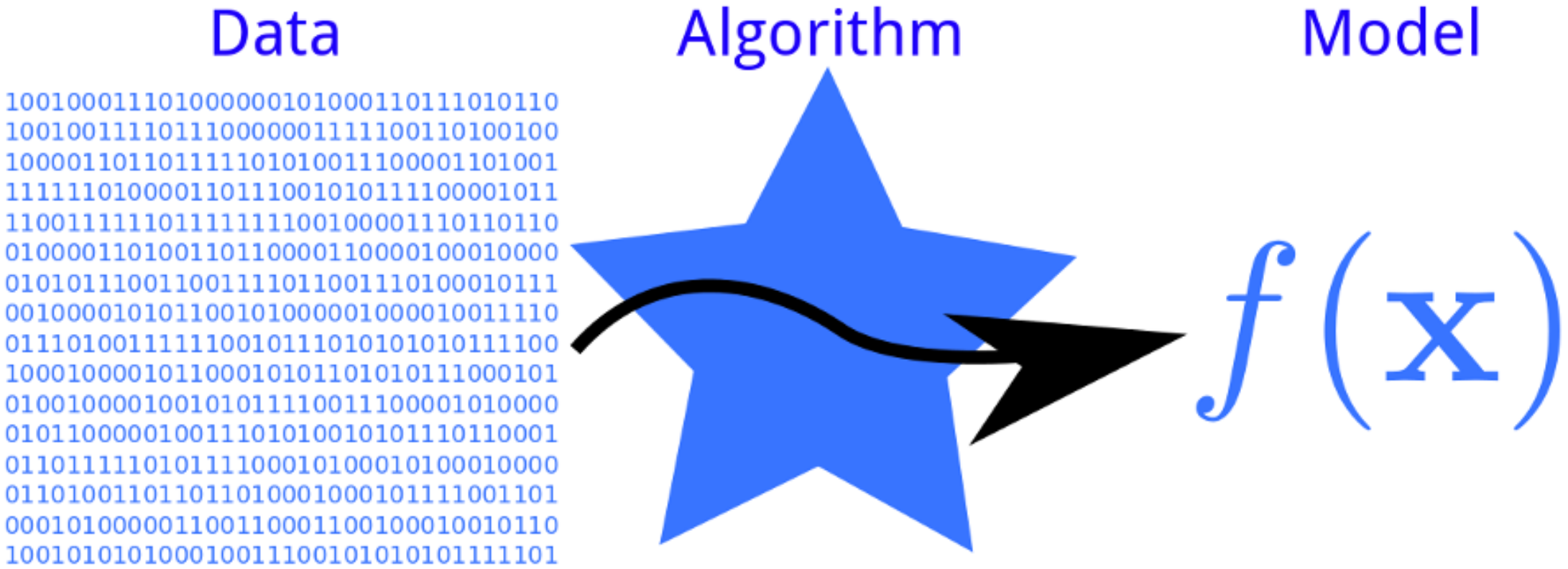
WEEK 2

Outline

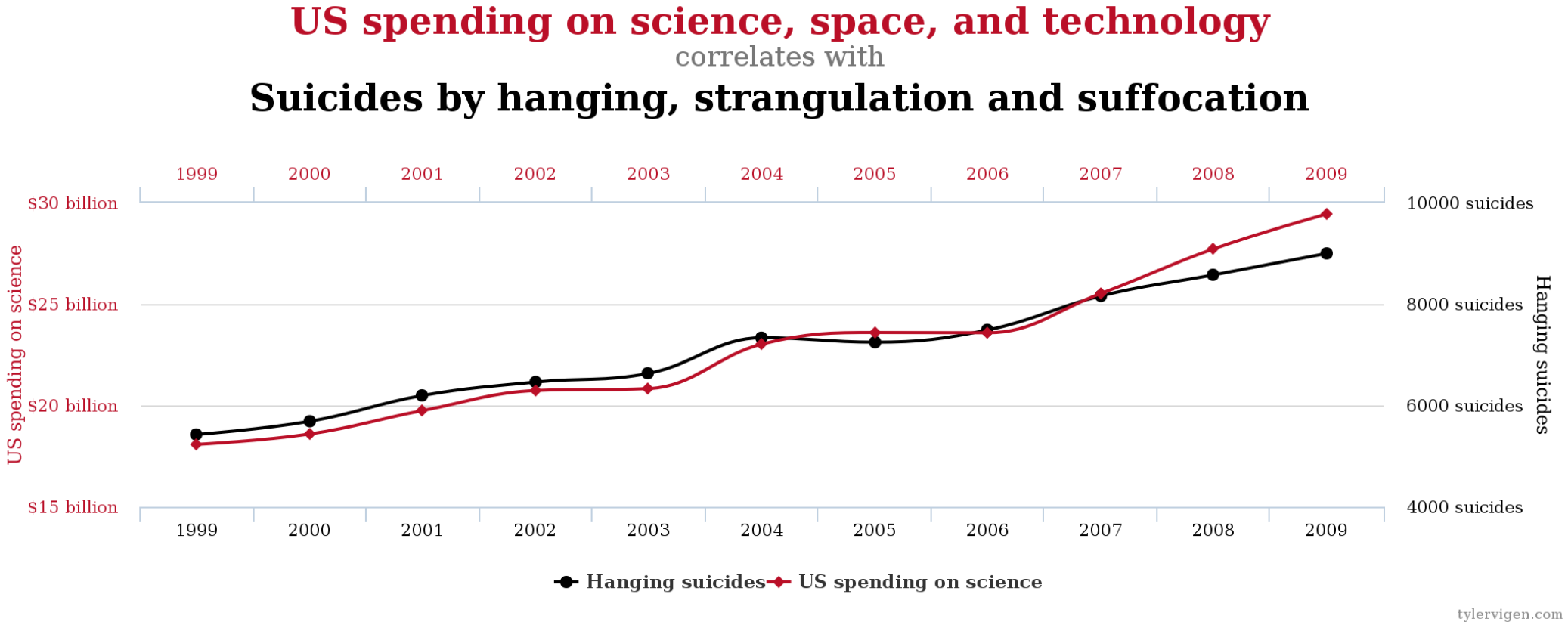
Machine Learning Introduction

Scikit-Learn ecosystem

What is machine learning?



Learnable or not? Causation or Correlation?



http://www.tylervigen.com/spurious-correlations

General Learning Models - Supervised

"Labeled" data

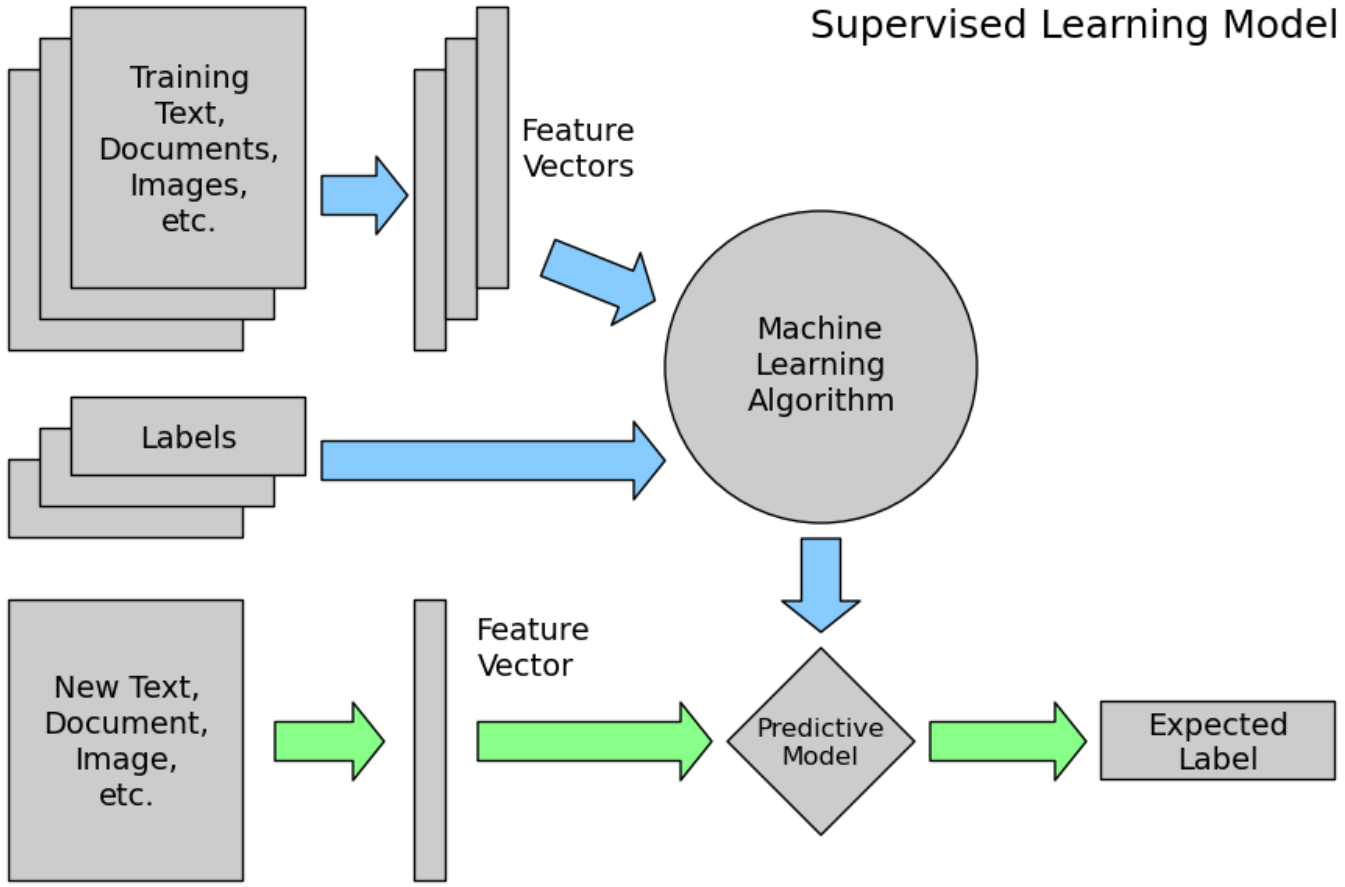
Feature Vector

+

Labels

=

Predictive Model



General Learning Models - Unsupervised

Unlabeled Data

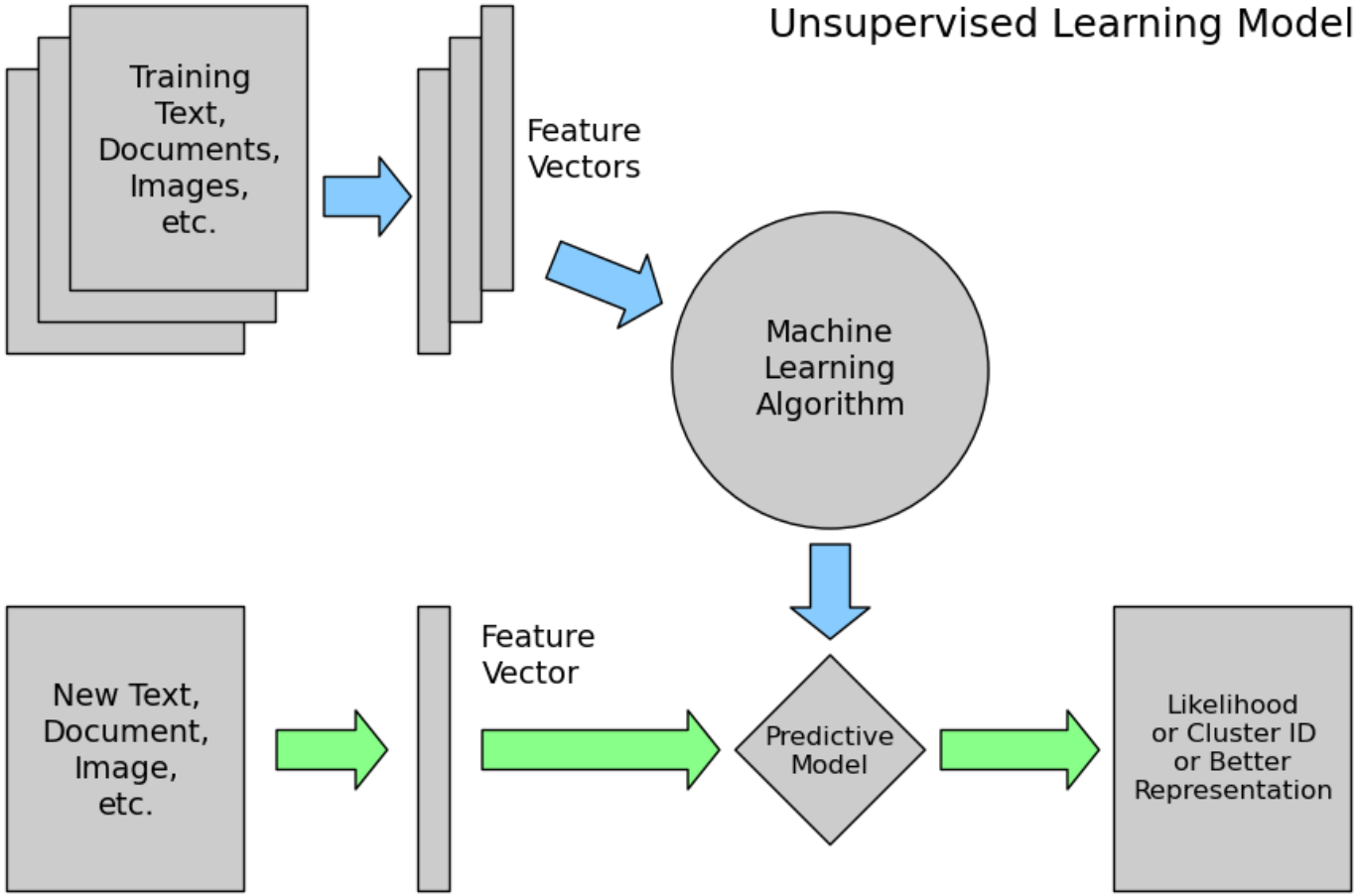
Feature Vectors

+

Intrinsic structure

=

Predicted Clusters



Part 1. Feature Extraction

This is the KEY

If the total prediction power is 100%.

Effort on feature engineering contribute to 80%

Effort on learning algorithm contribute to 20%

What are features?

Information that is useful to make prediction

Data types

1. Numeric
2. Text
3. Audio
4. Image
5. Video

Part 2. Learning Algorithms

Supervised (given X, Y)

Regression

Classification

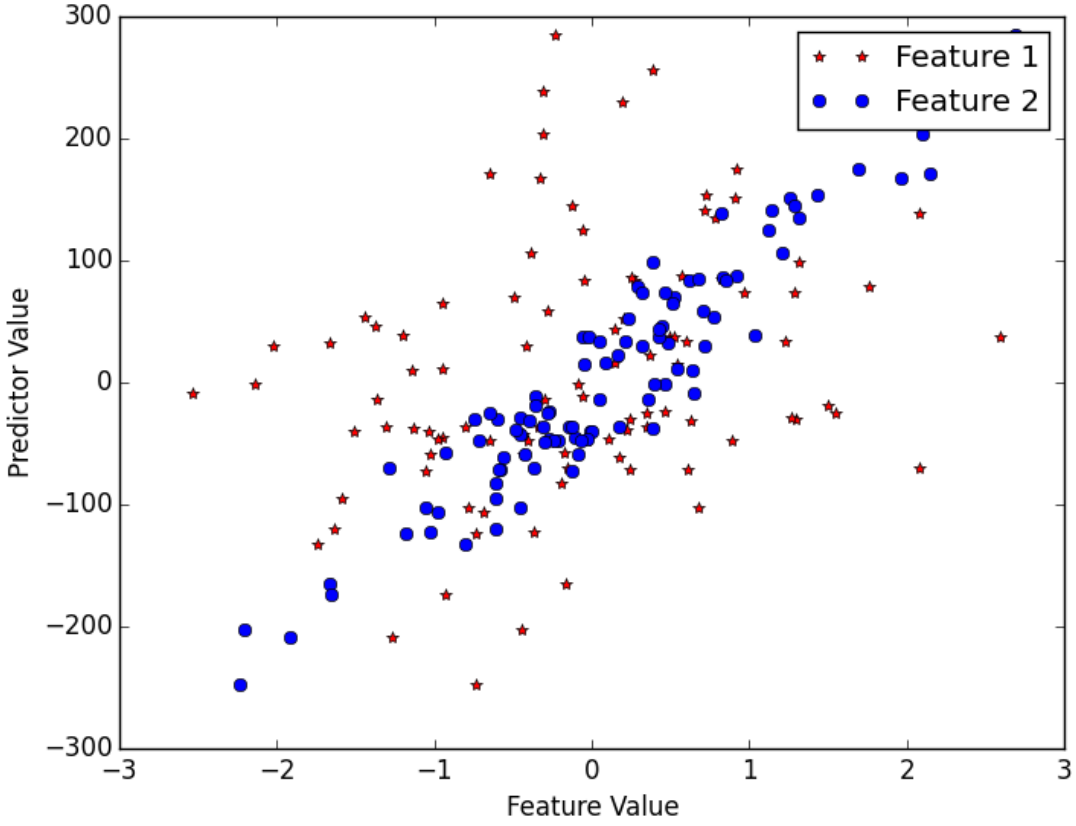
Unsupervised (given X, no Y)

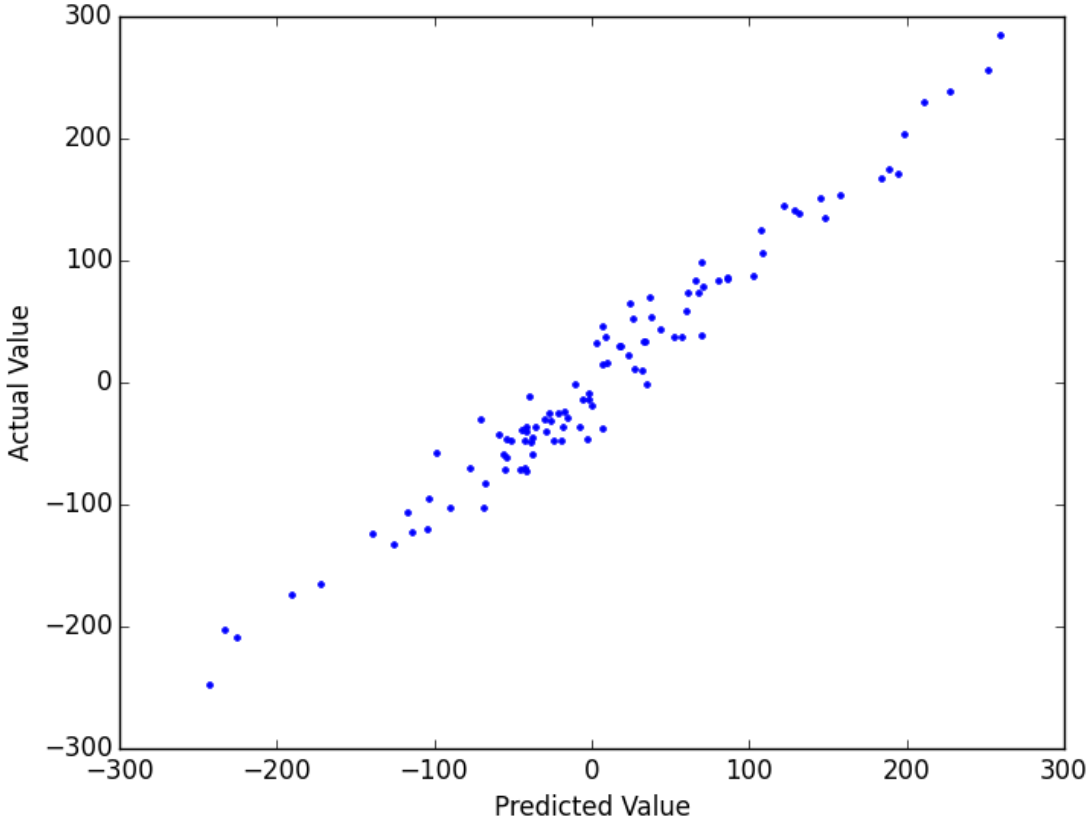
Clustering

Dimension Reduction

Outlier / anomaly detection

Goal Today: Regression with Scikit-Learn





Scikit Learn Syntax – Cross Validation

Cross

&

Validation