DS501: Machine learning, Part 1

Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

Worcester Polytechnic Institute

2014

WPI

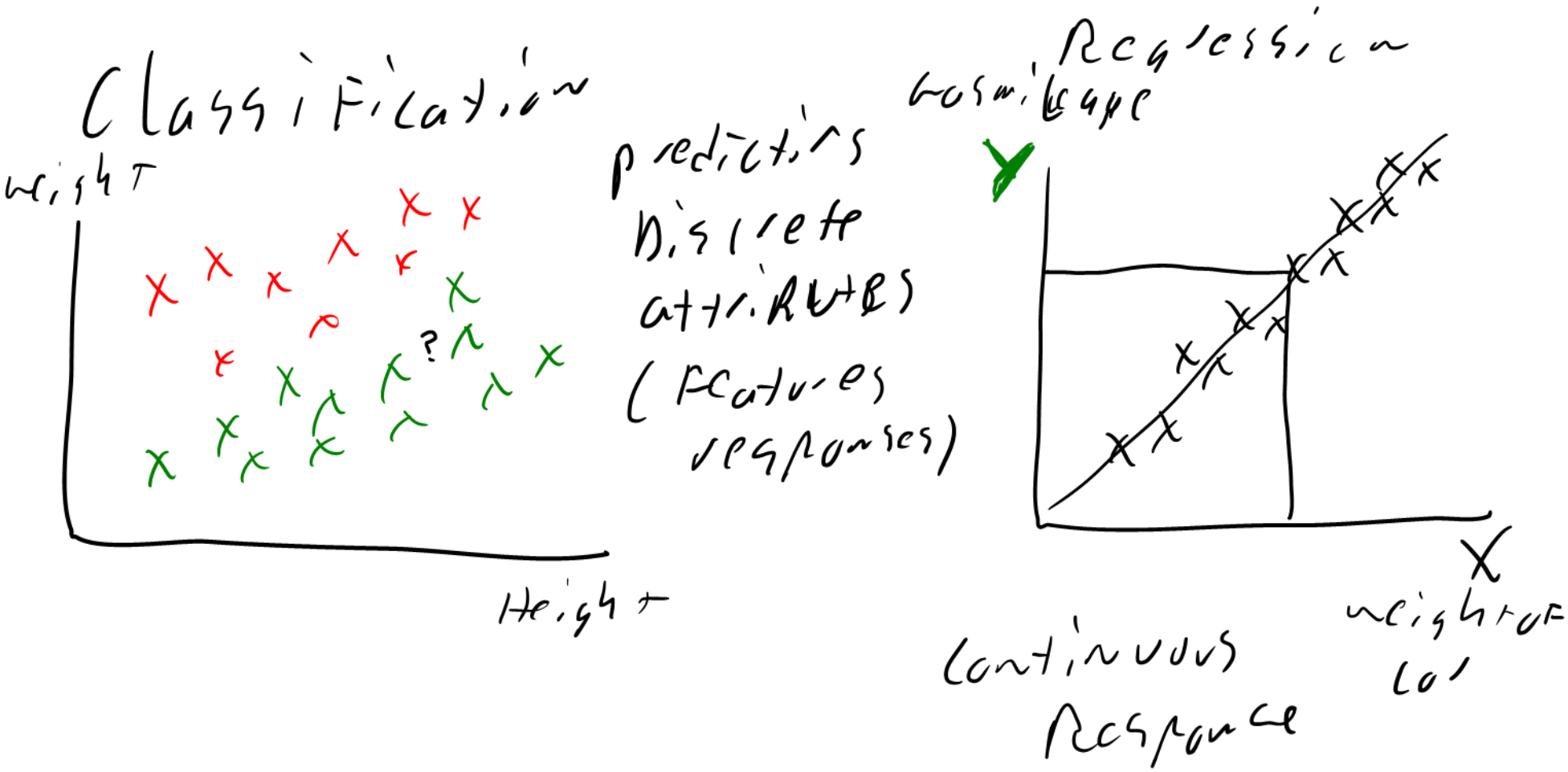
Learning **objectives** for this machine learning class.

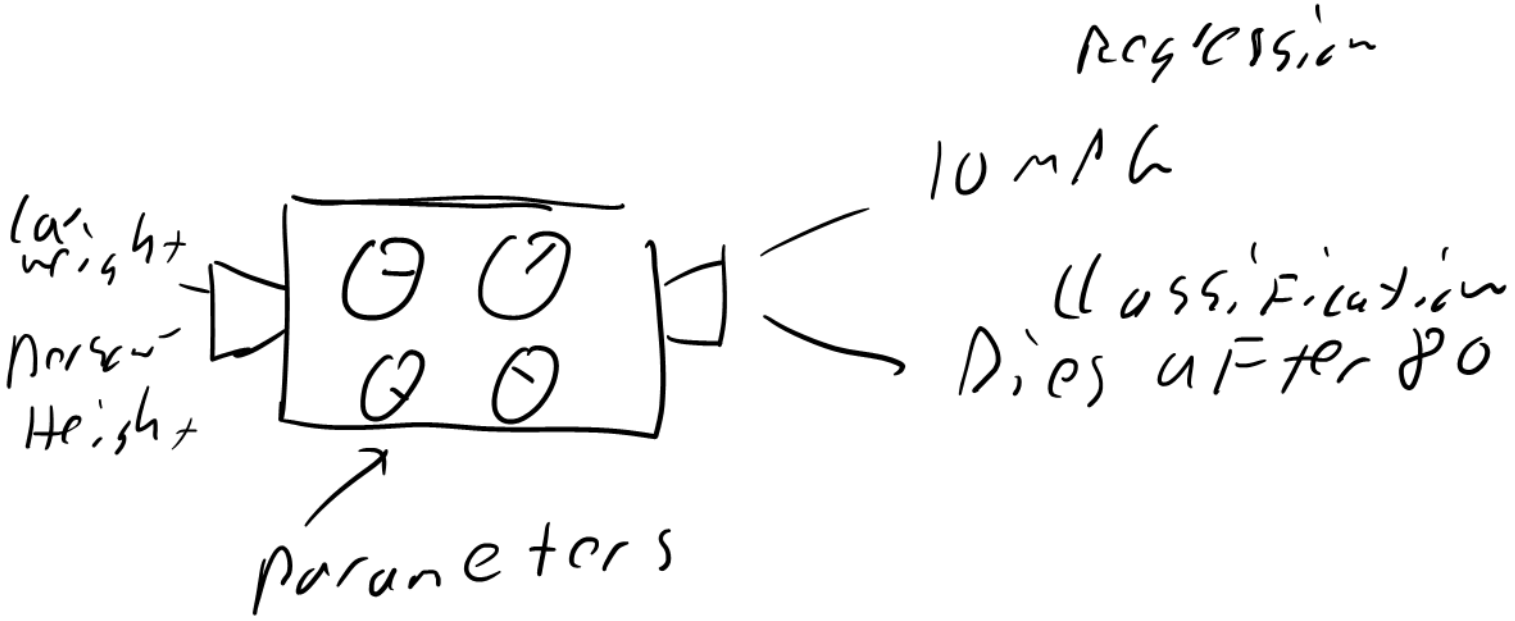
- Overview of machine learning
- Supervised classification.
  - K-nearest neighbors
  - Support vector machines
- Learn some Python packages, including:
  - scikit-learn

Inspired by: http://scikit-learn.org/stable/tutorial/statishttp://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.htmltical_inference/supervised_learning.html

WPI

Basic definitions: Classification vs. Regression



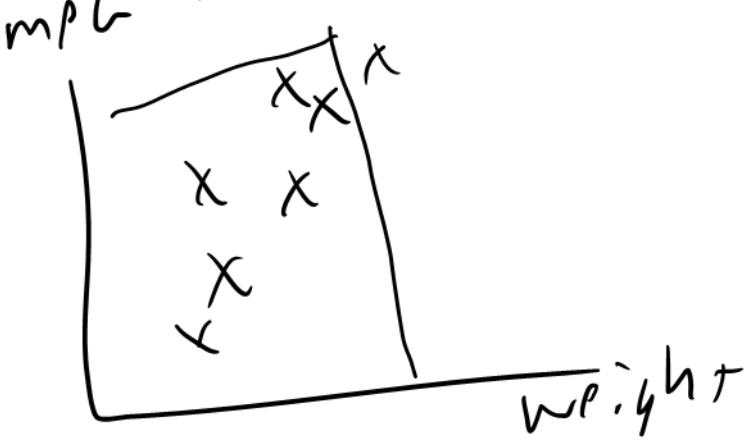


Data is used to set the dials

Basic definitions: Supervised vs. Unsupervised Learning

supervised

You have examples to learn from



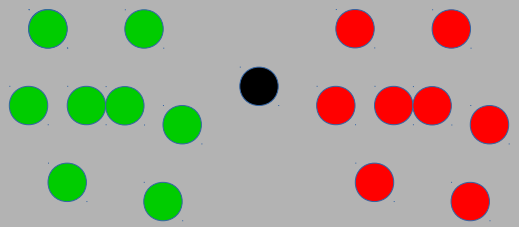
unsupervised

no examples to learn from!

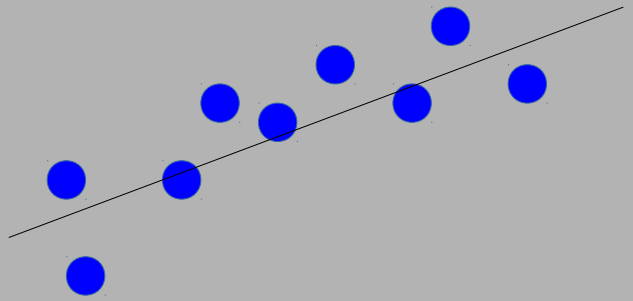
colors make it supervised

We have four days we will cover four topics.

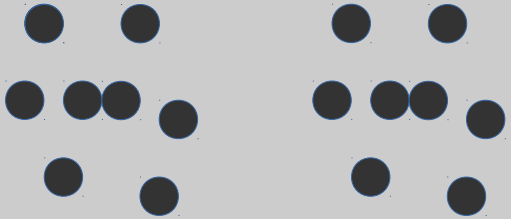
Supervised Classification



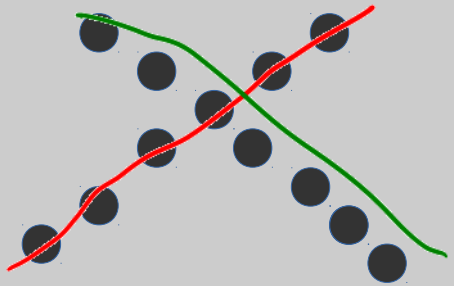
Supervised Regression



Unsupervised Clustering

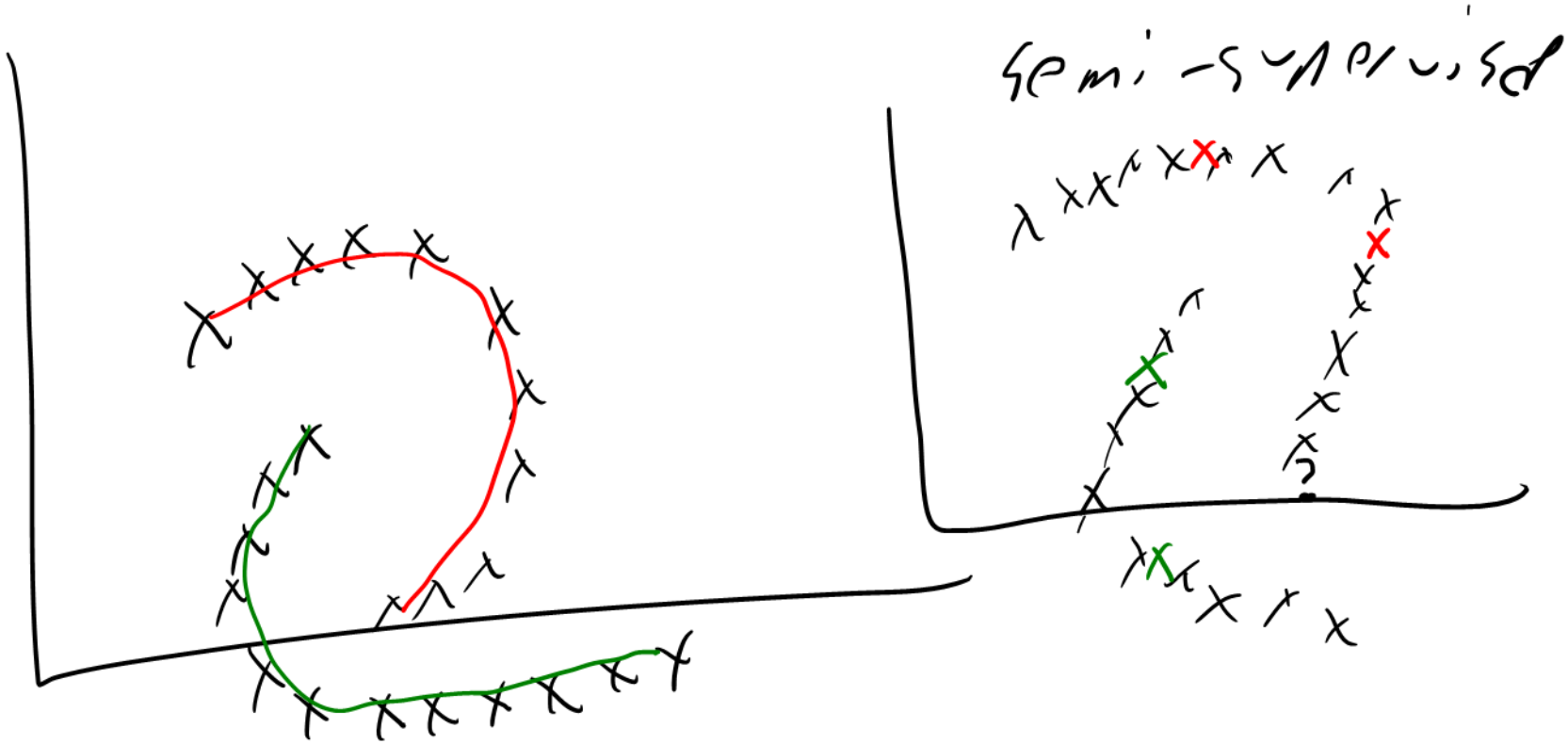


Dimension Reduction

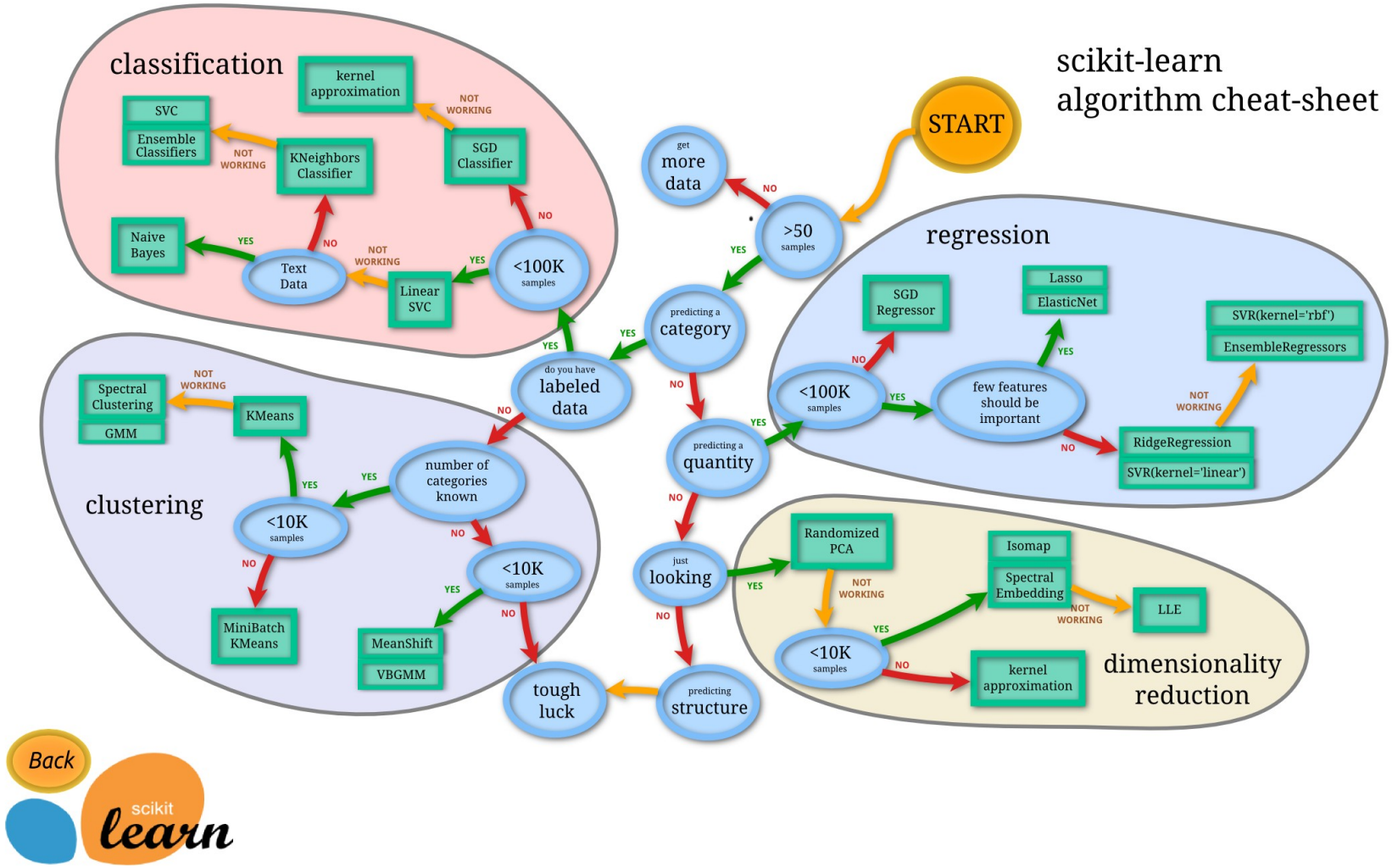


WPI

What about Unsupervised Regression?



scikit-learn



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

WPI

Iris data set

Features:

sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)

"Iris virginica" by Frank Mayfield - originally posted to Flickr as Iris virginica shrevei BLUE FLAG. Licensed under Creative Commons Attribution-Share Alike 2.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Iris_virginica.jpg#mediaviewer/File:Iris_virginica.jpg

Catagories:

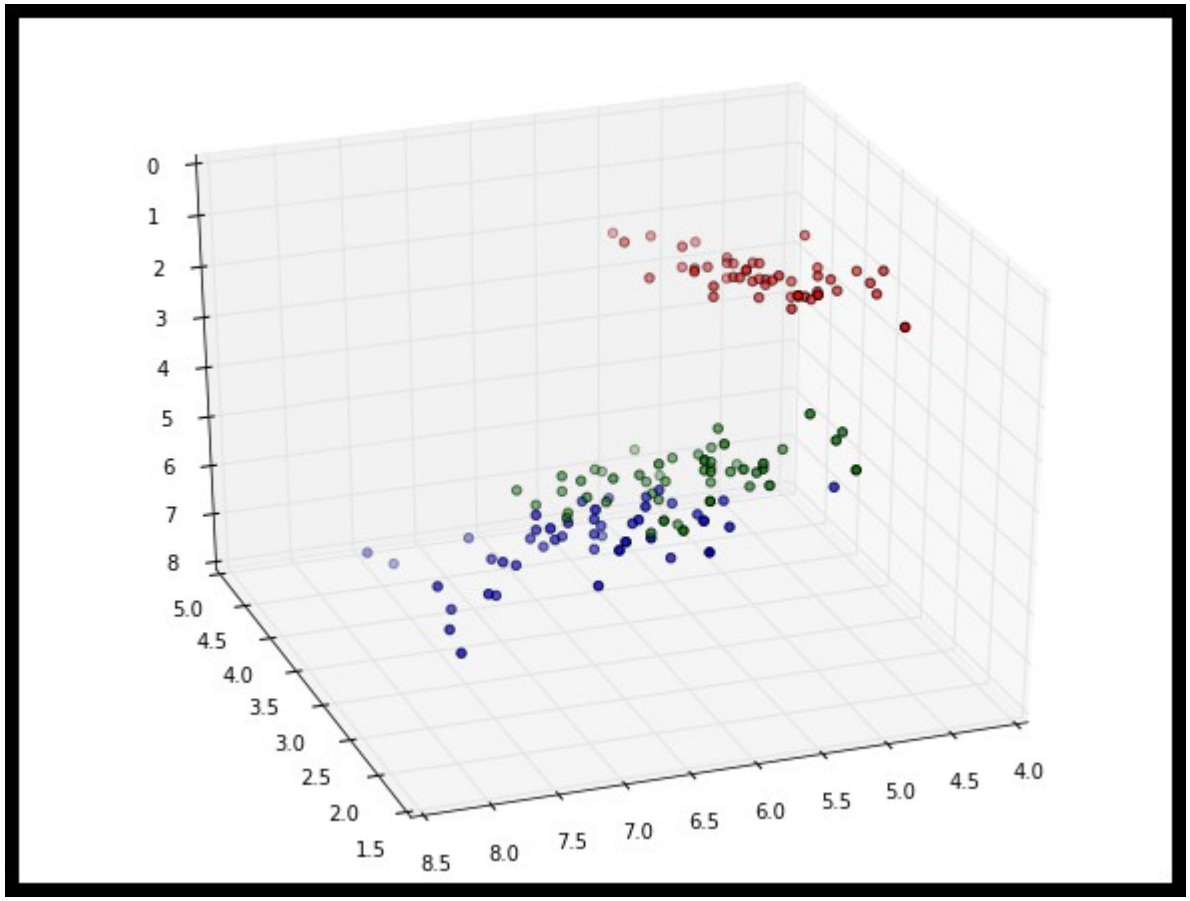
setosa
versicolor
virginica

"Kosaciec szczecinkowaty Iris setosa". Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg#mediaviewer/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg

"Iris versicolor 3". Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Iris_versicolor_3.jpg#mediaviewer/File:Iris_versicolor_3.jpg

WPI

Let's look at it in Python



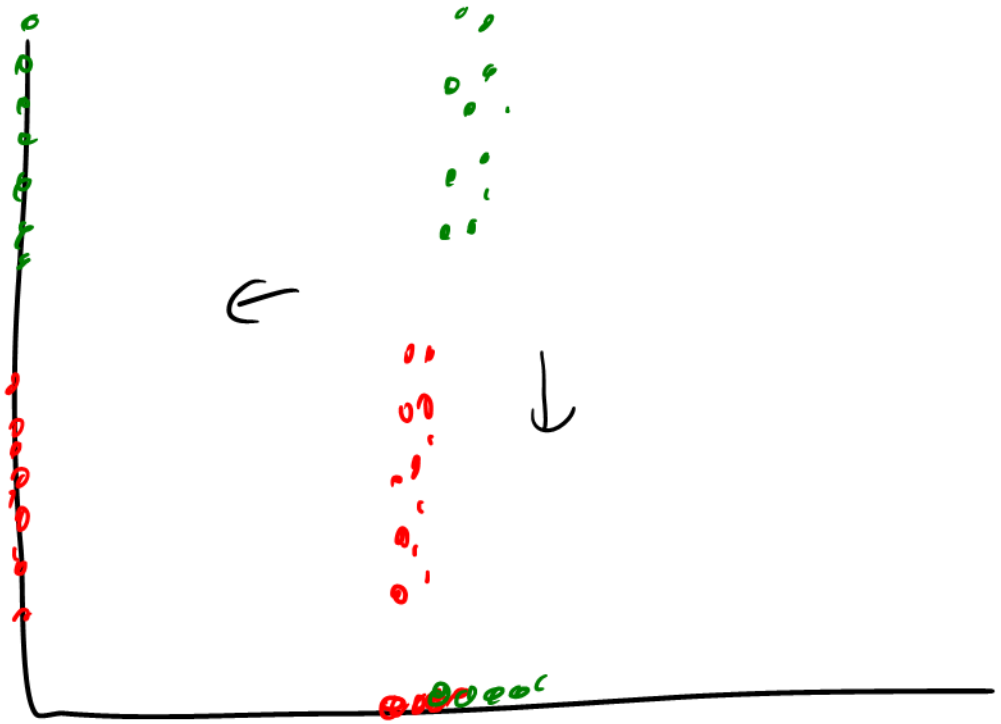
WPI

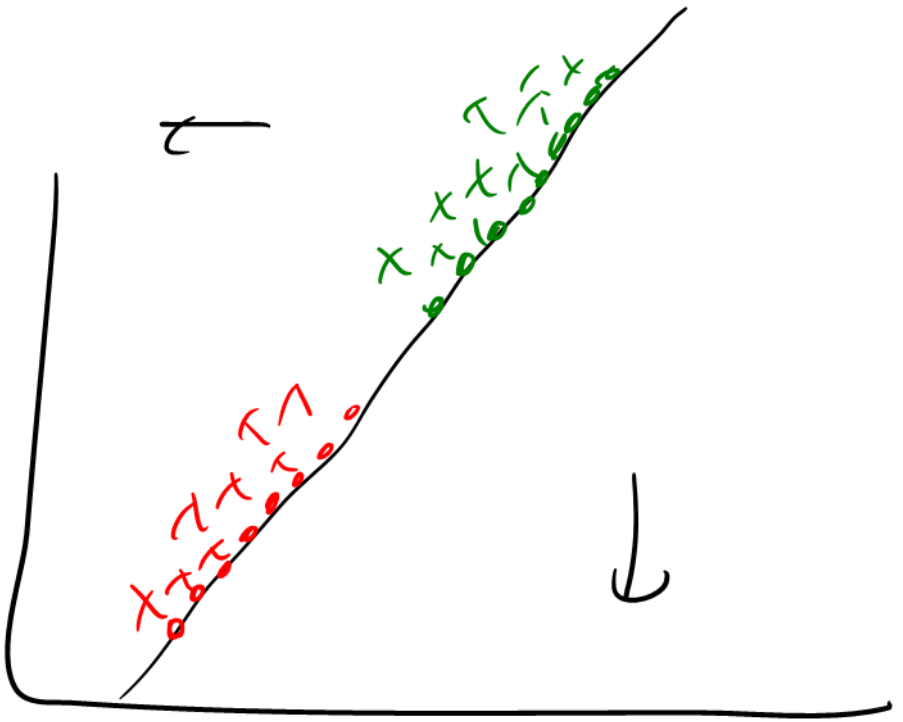
What is PCA?

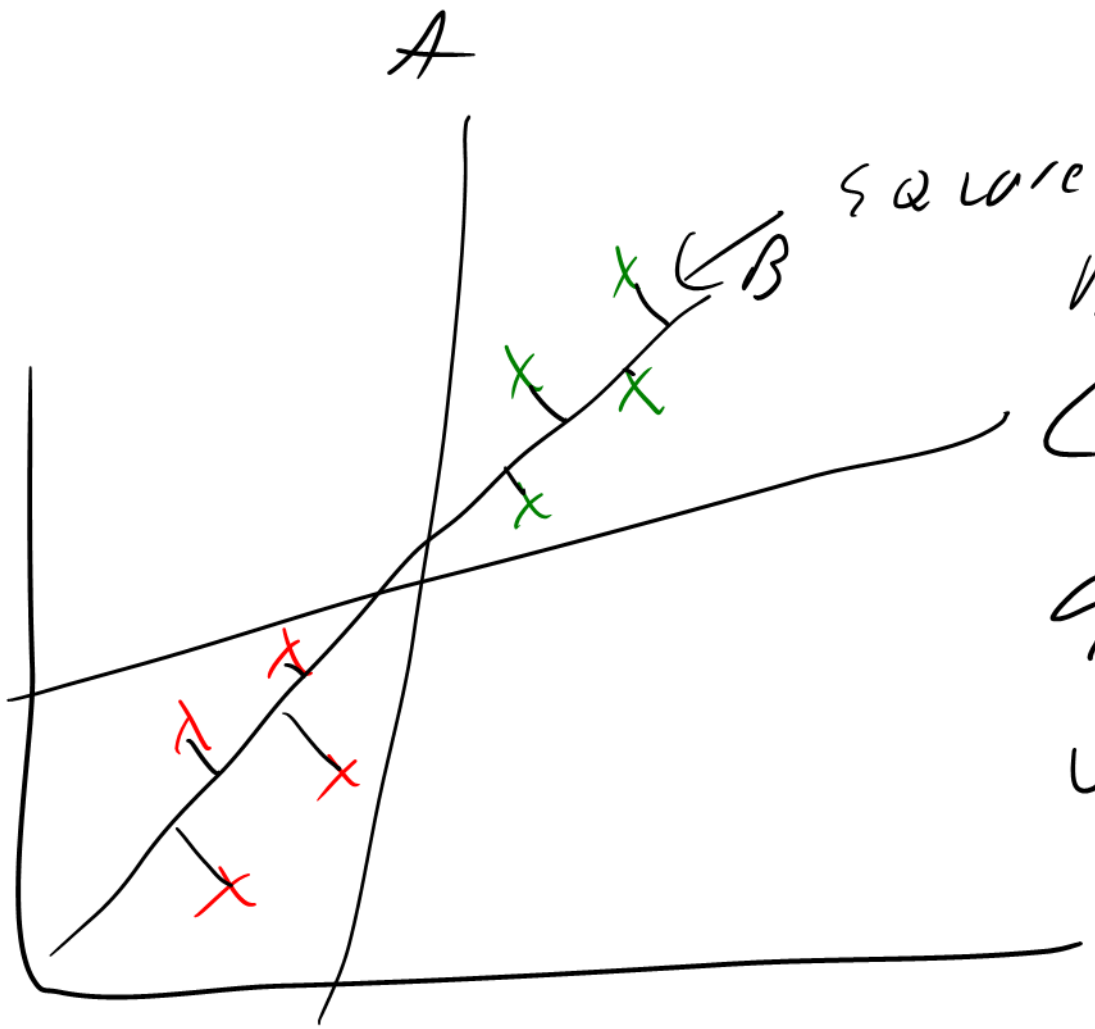
- Principle Component Analysis
  - Commonly used tool for visualization and data pre-processing.

Key idea of PCA

maximize the variance of the projection!

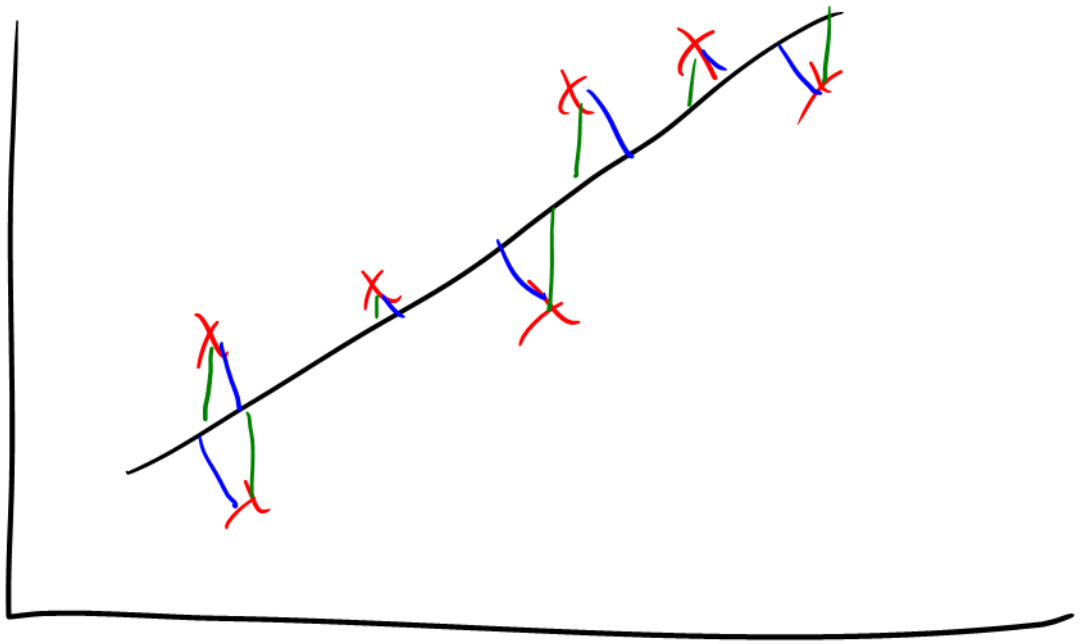




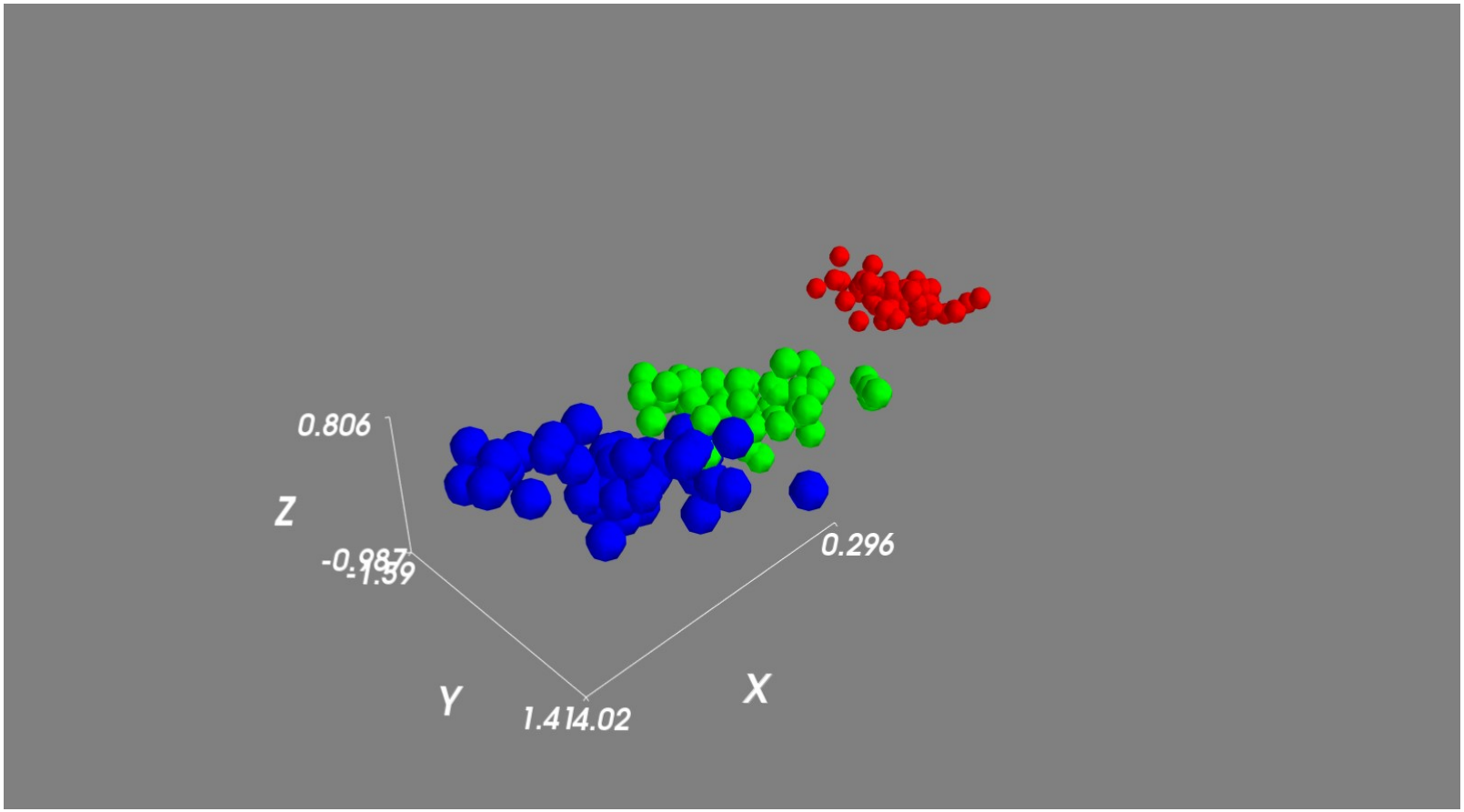


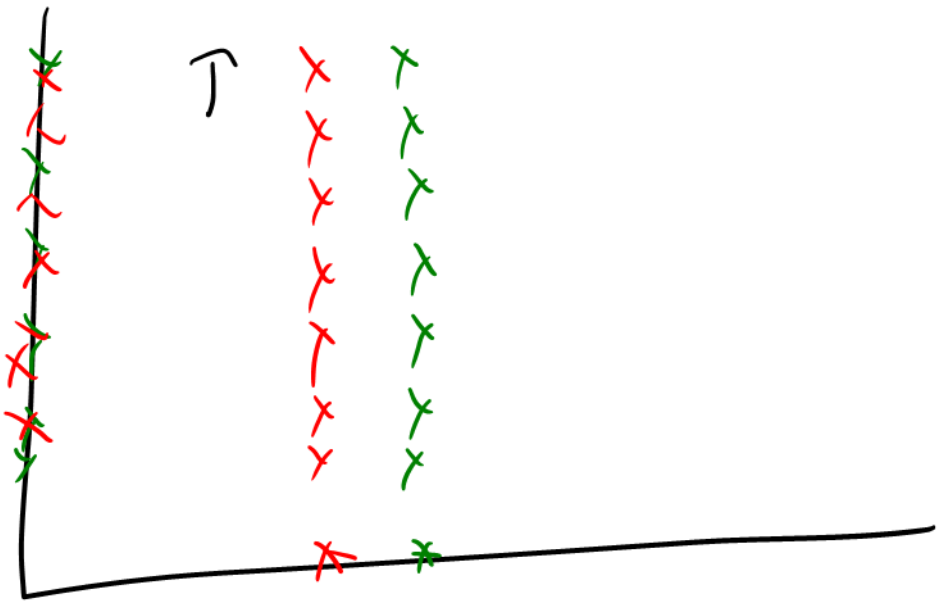
squares with the
Distance
C

Singular
value
Decomposition
SVD

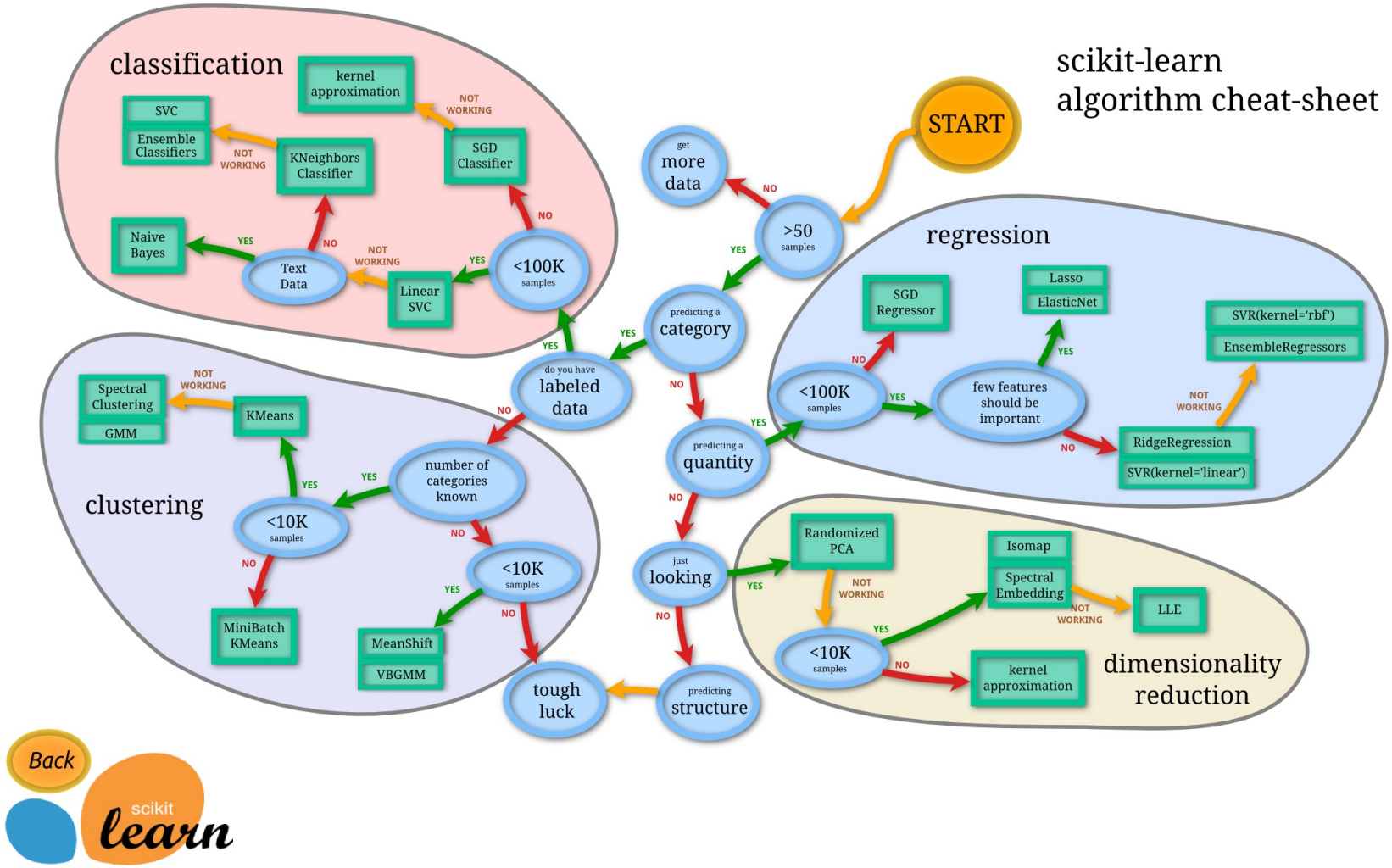


Lets take a look at it in Python





scikit-learn



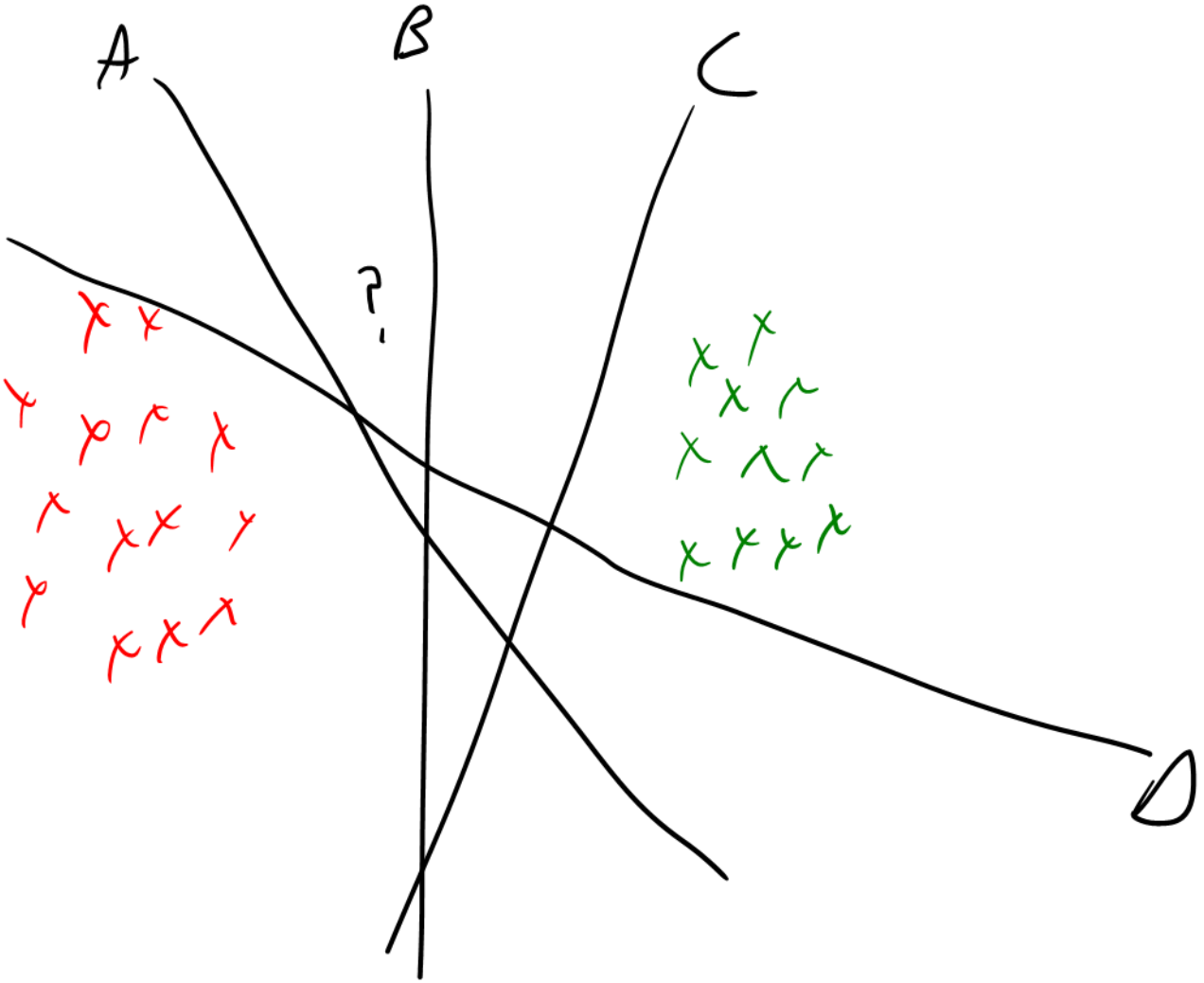
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

WPI

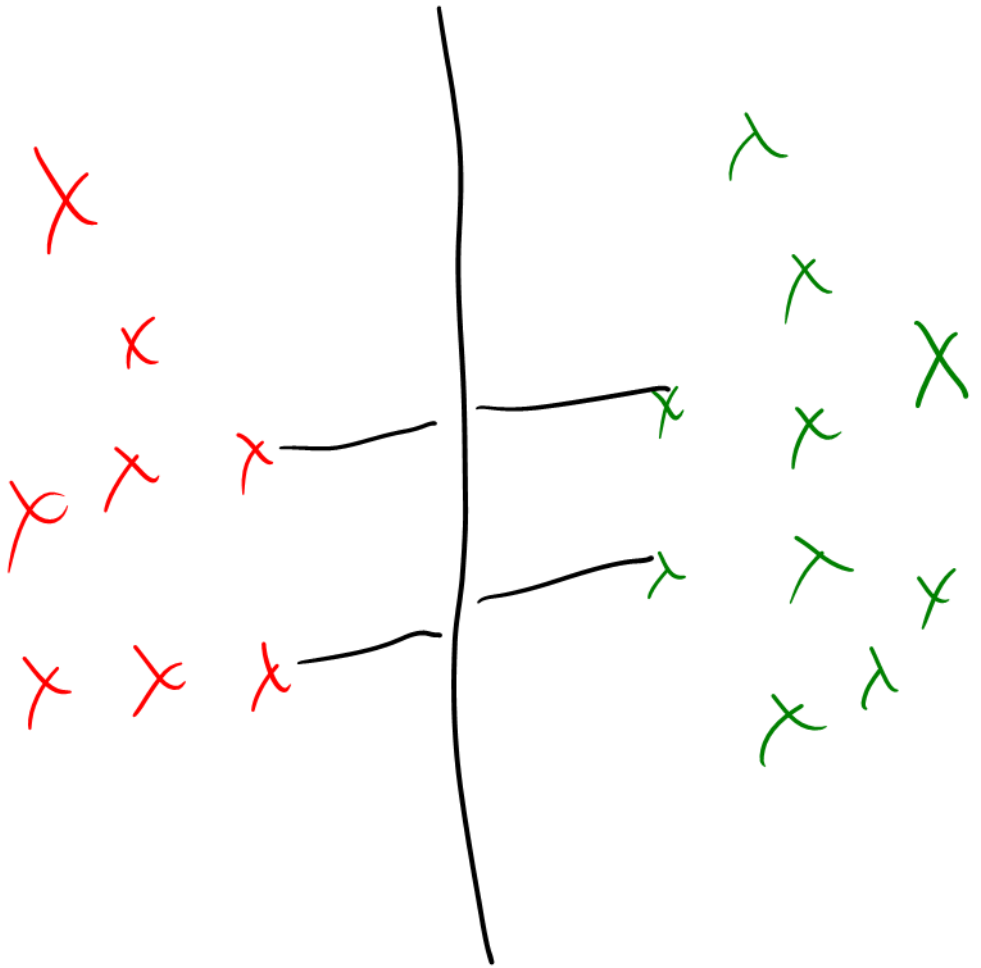
What is Linear Support Vector Machine (SVM)?

- Maximum margin classifier
  - Computes a linear “decision boundary” that splits the data into two regions.
  - Allows one to predict a classification of a point based upon which side of the decision boundary it lay on.

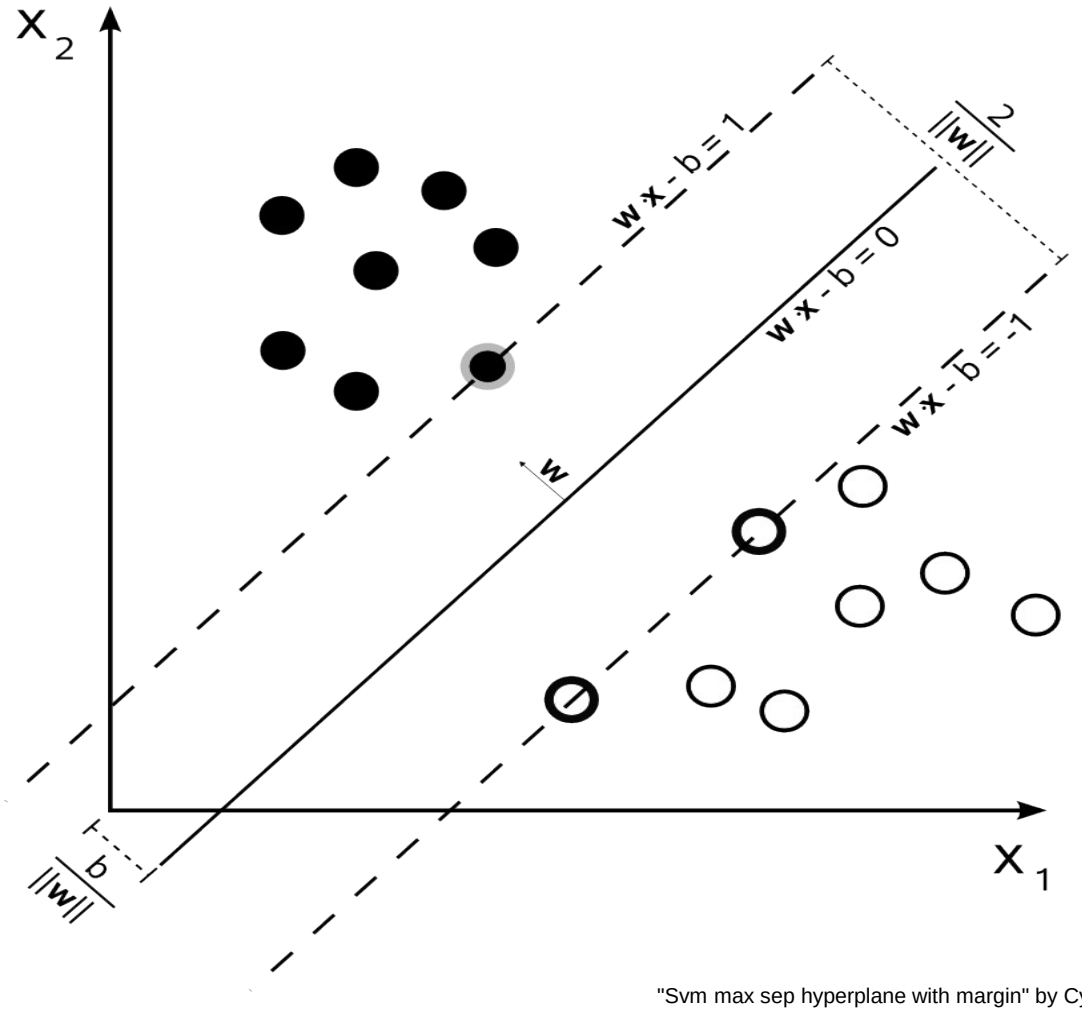
Let's derive SVM



maximize the minimum distance

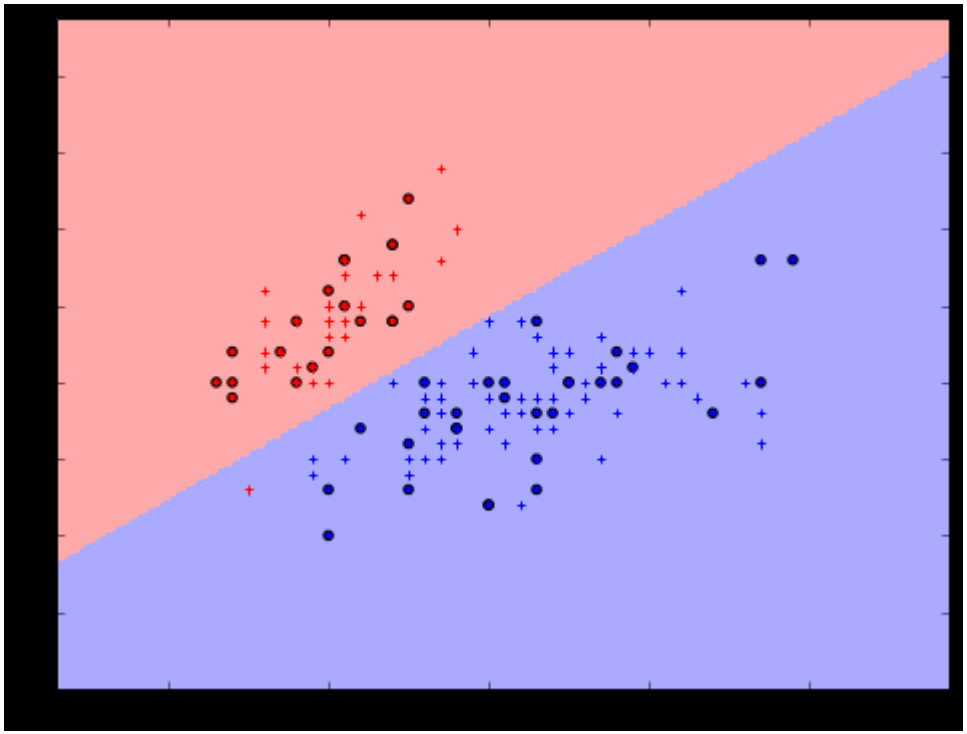


SVM



WPI

Let's look at SVM in Python



WPI

But wait! Training vs. testing!

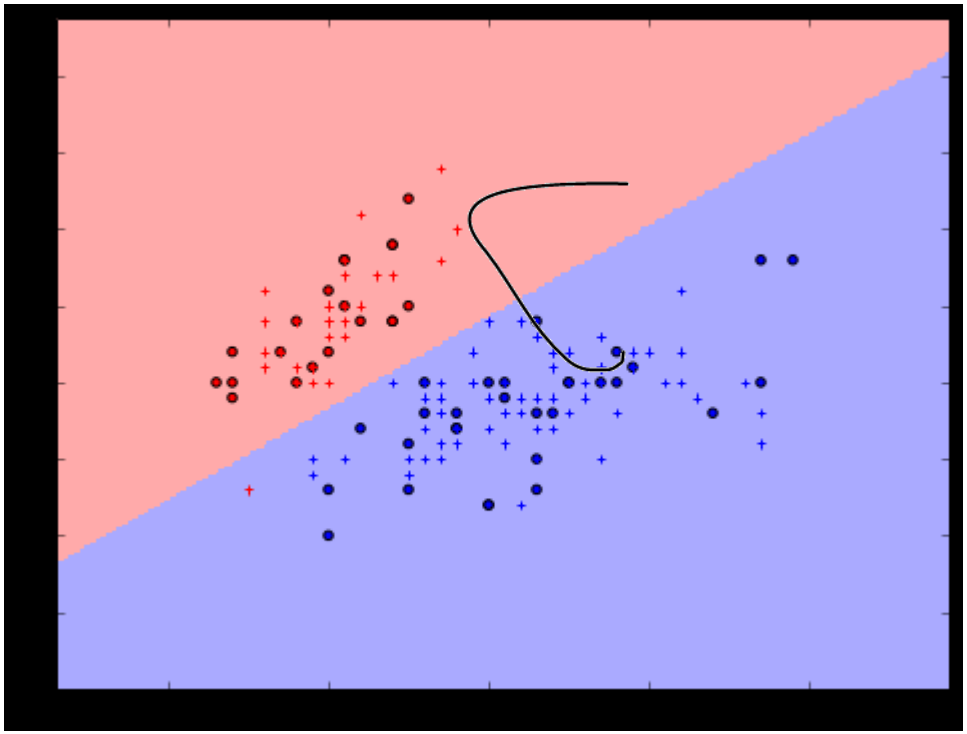
Some data For your algorithm

some data For testing your Algorithm

Fundamental

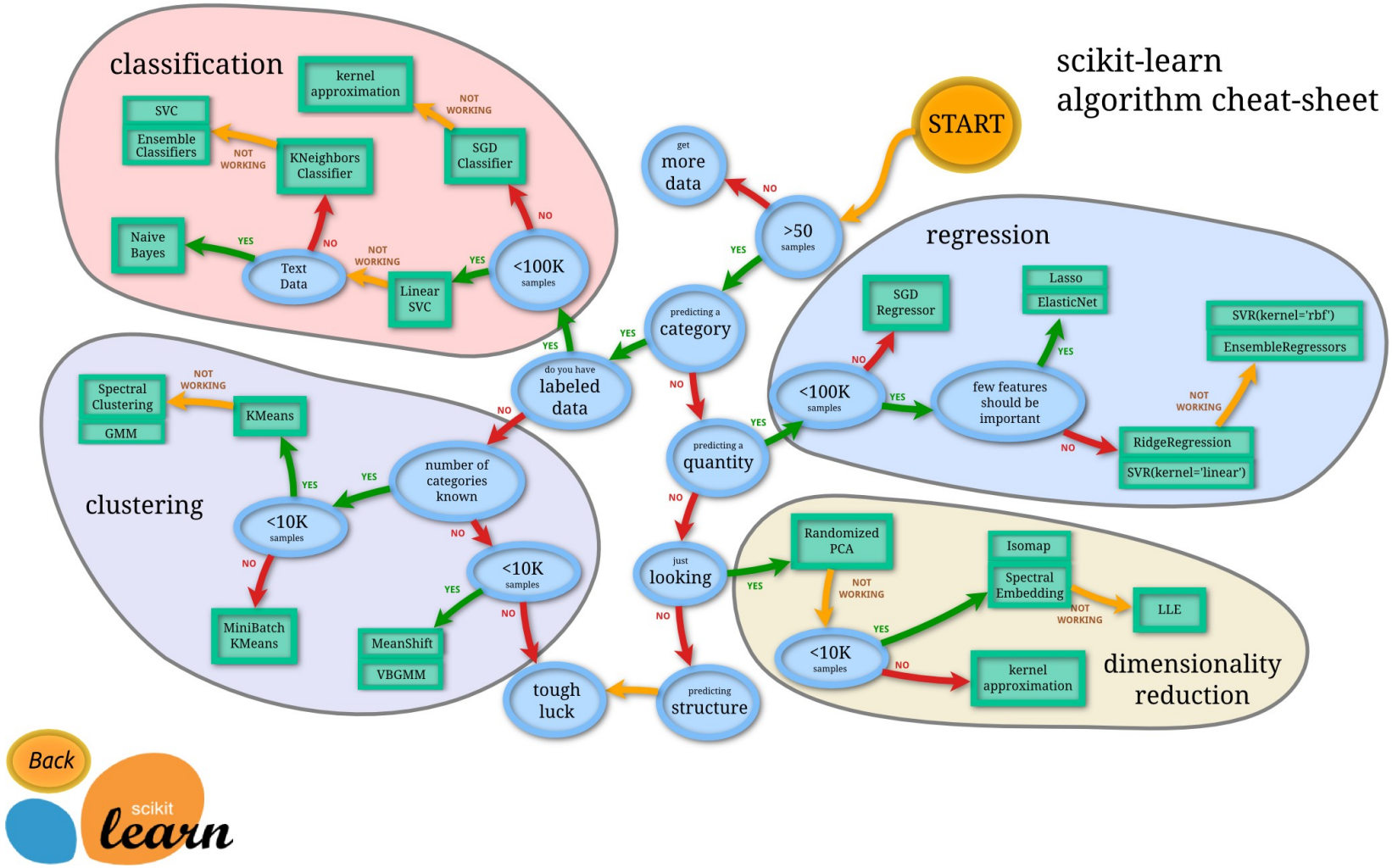
avoid over Fittings

Back to Python



WPI

scikit-learn



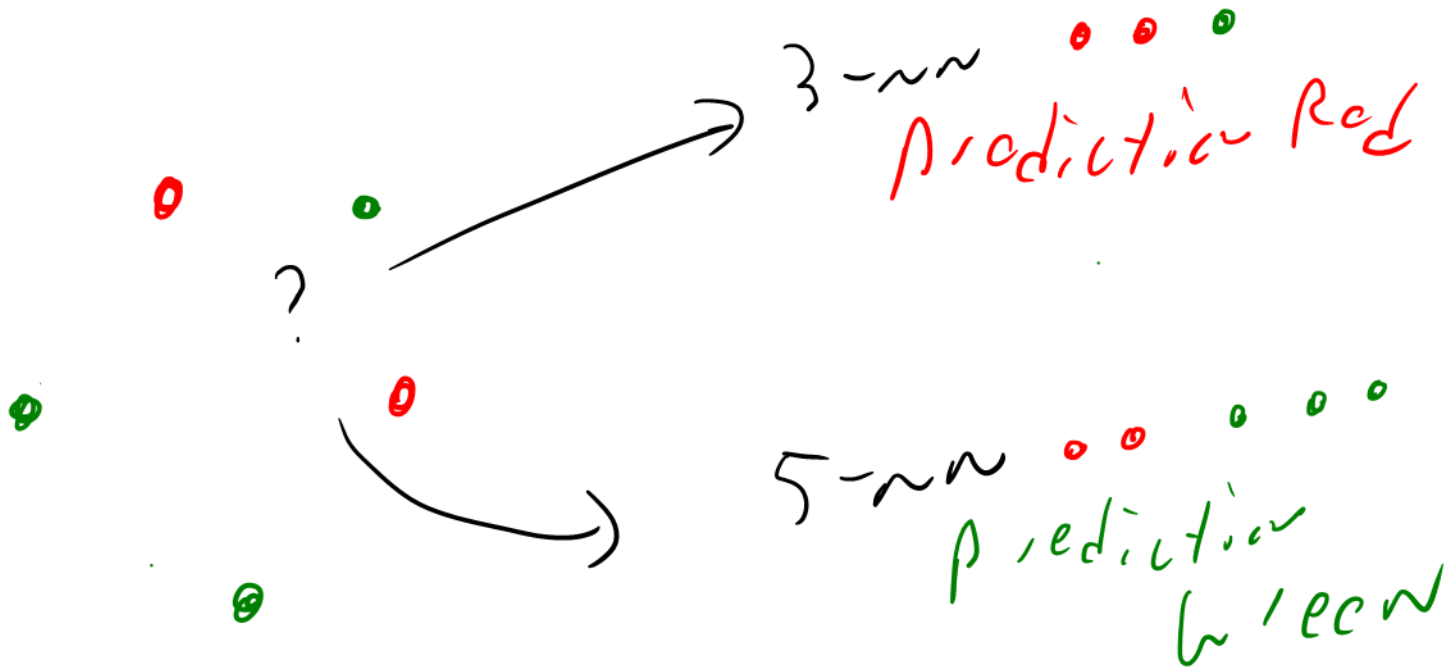
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

WPI

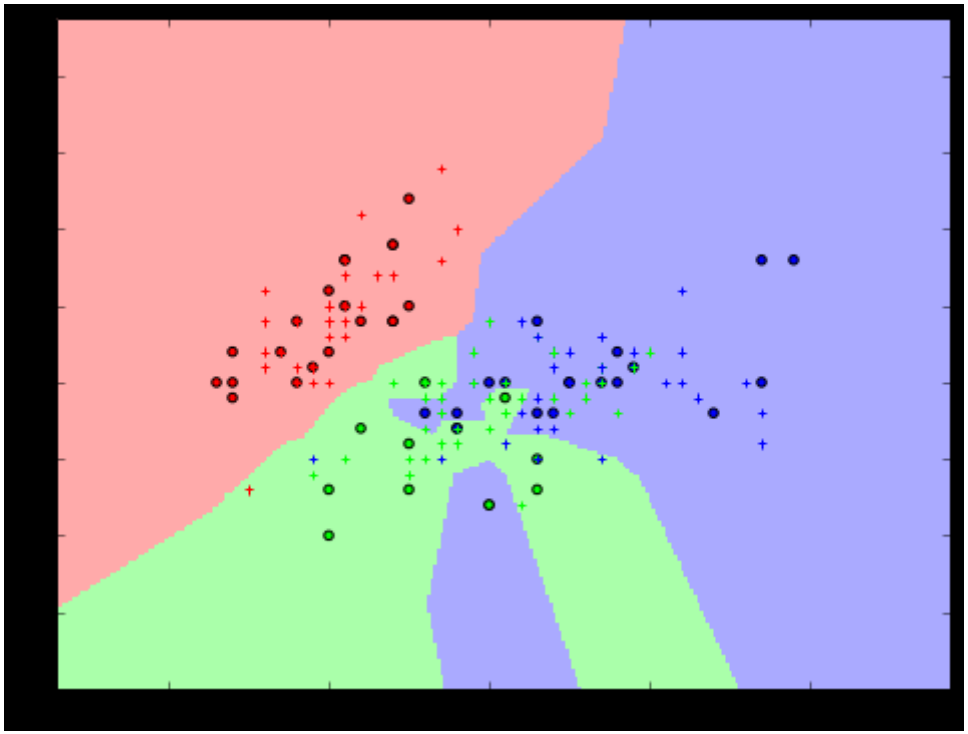
What is K-NN?

- K-nearest neighbors
- Another common classification algorithm
  - Perhaps the most common

Let's try an example



Let's try K-NN in Python



WPI