Michael Grossberg

Intro to Data Science

CS59969

Classification

Classification

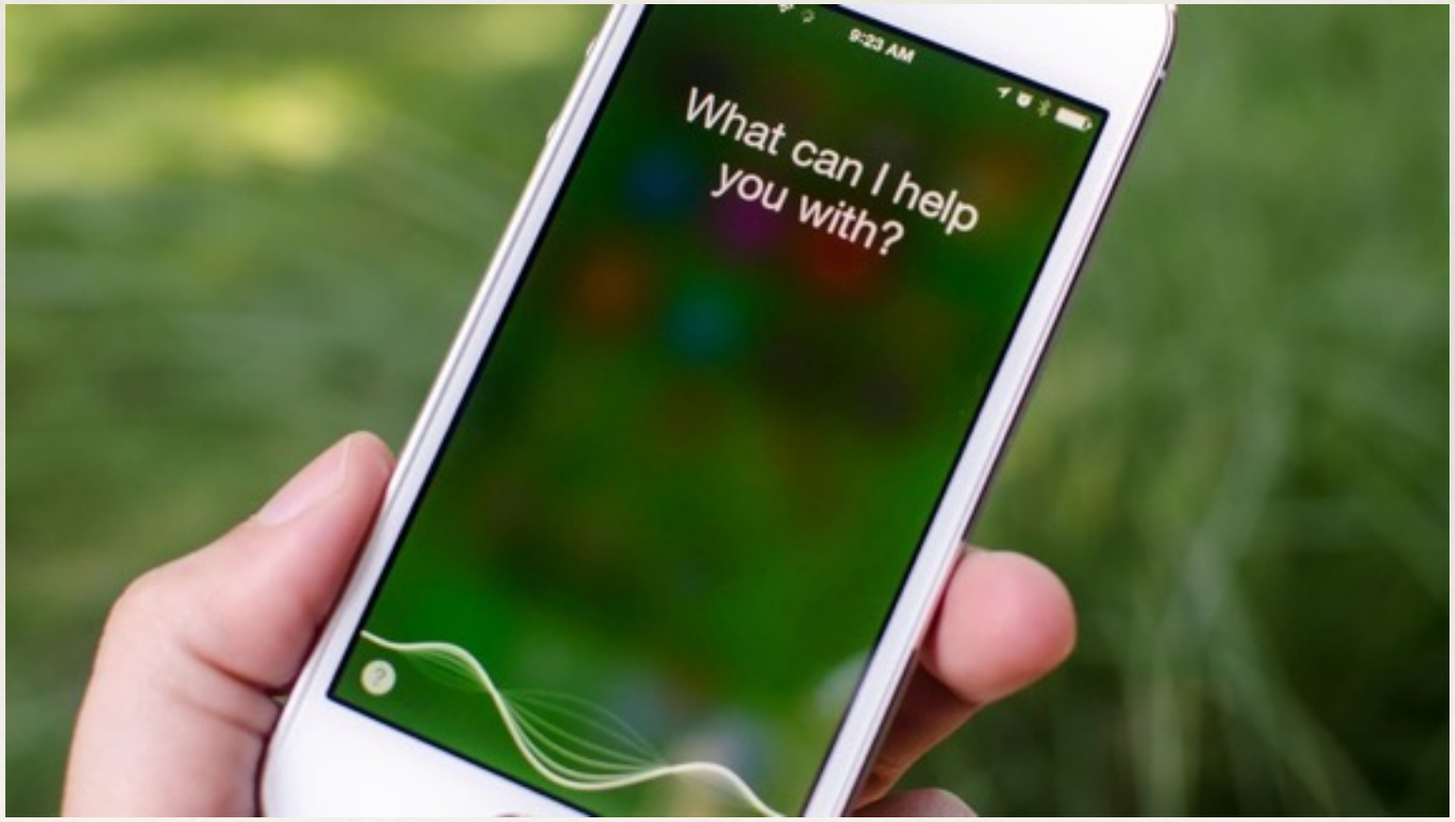
Example Problem MNIST

Assign Handwritten Digit
class 0,1,2, …, 9

0

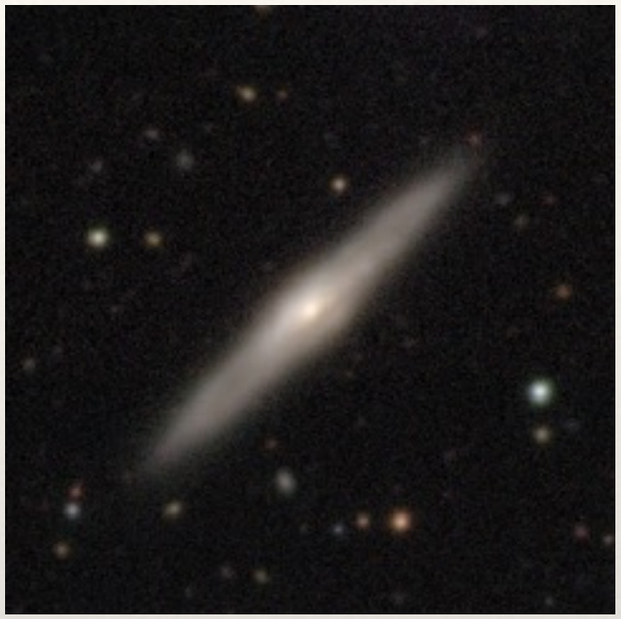
Speech Recognition

Sound ⟶ Word



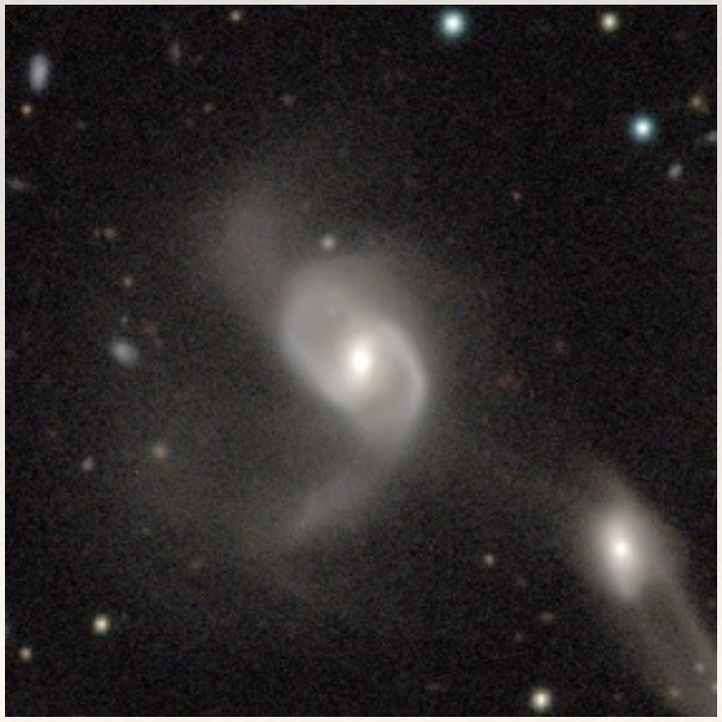
Automatic Identification Galaxies

Smooth

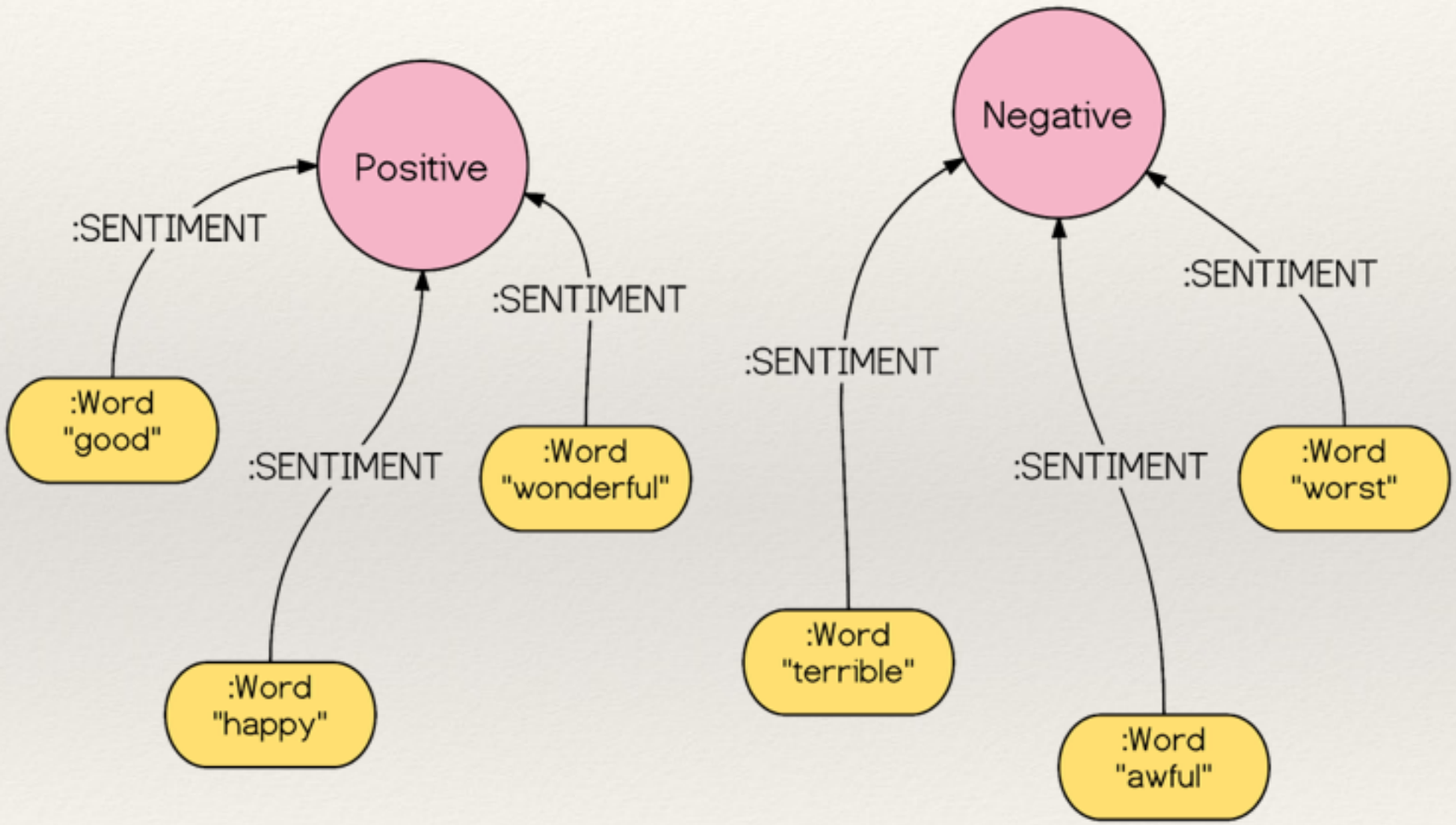


Disk

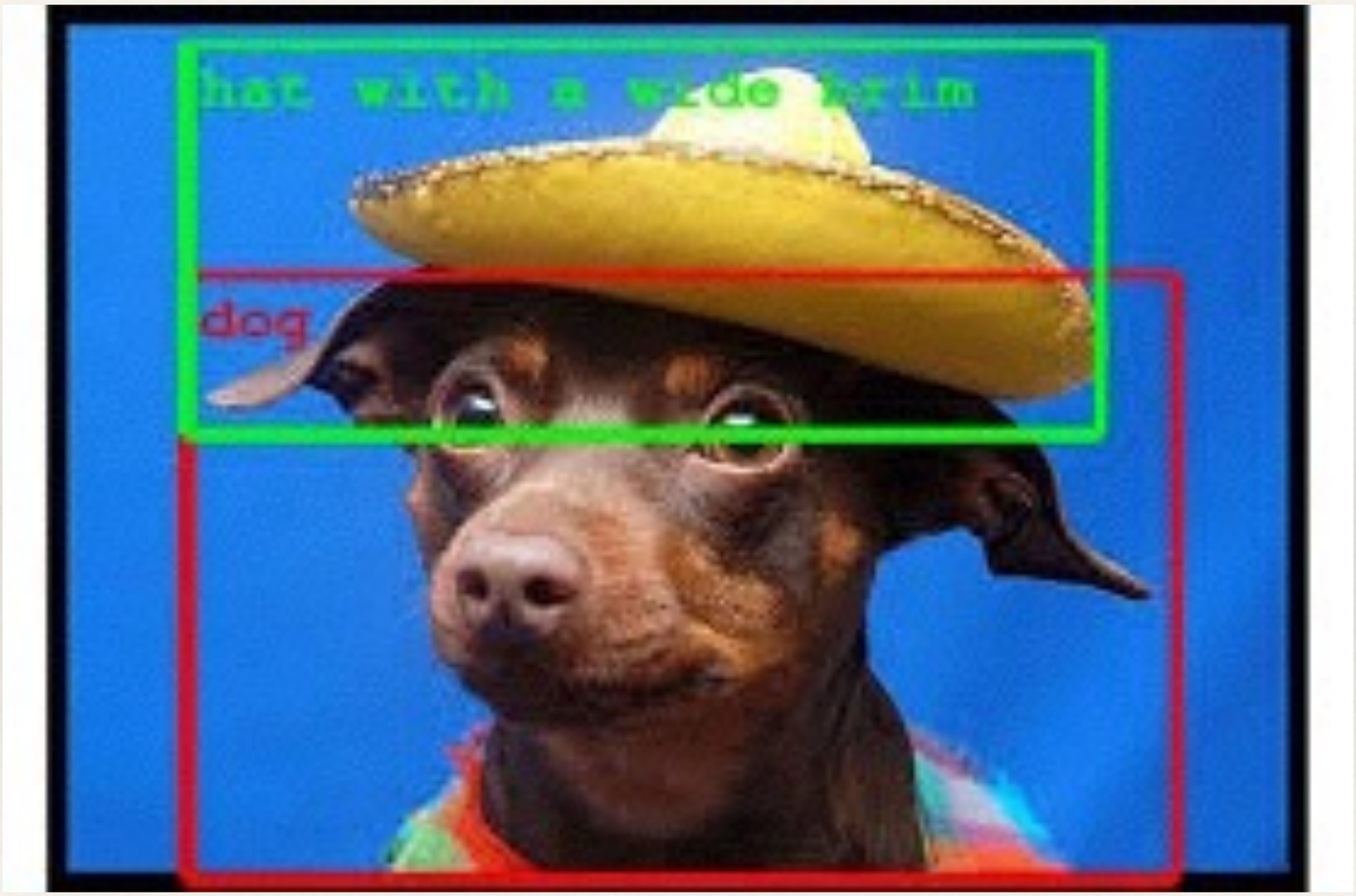
Spiral



Sentiment Analysis



Object Recognition

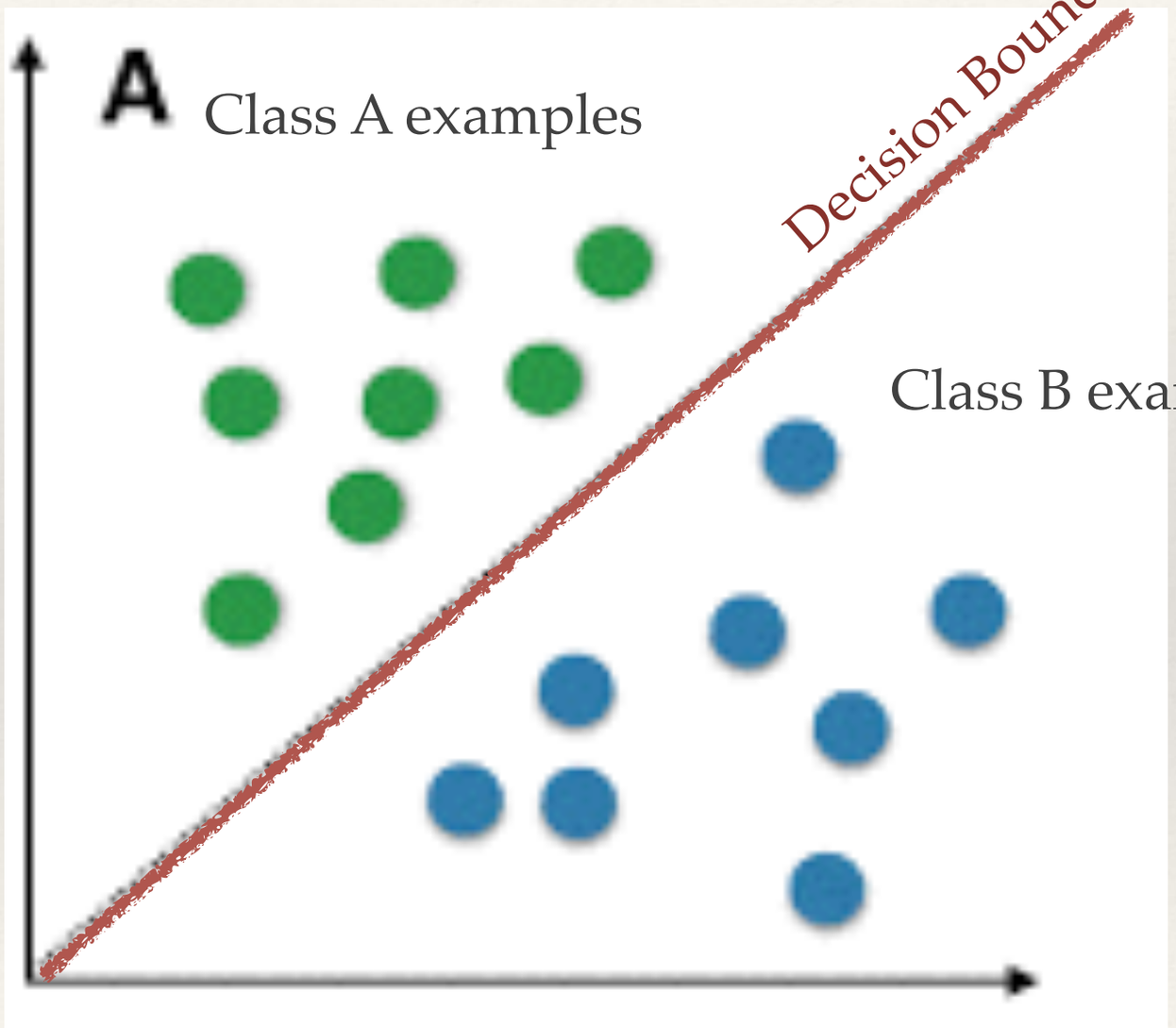


Object Recognition

Train

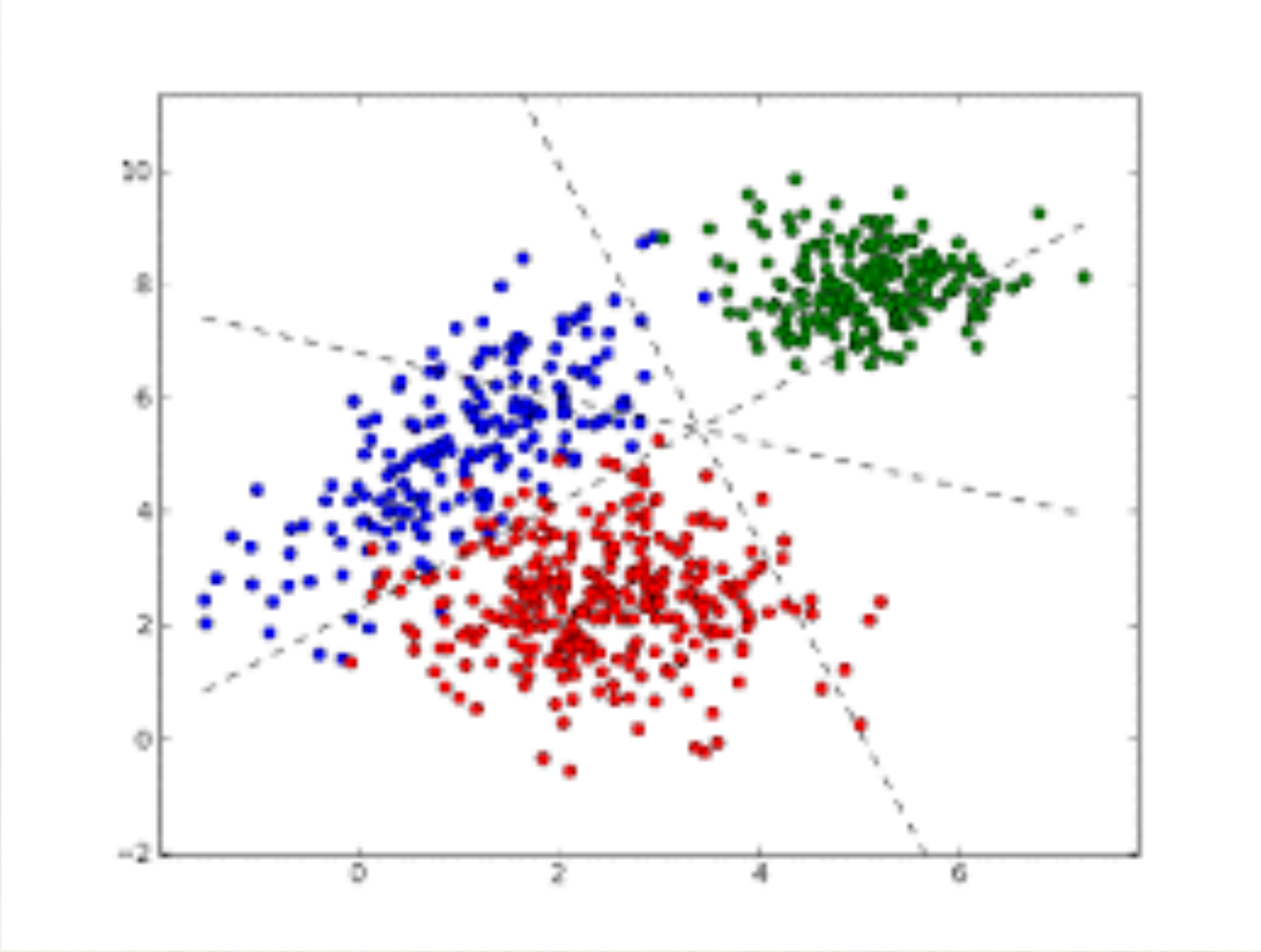
Test

Classification Part (2 class)

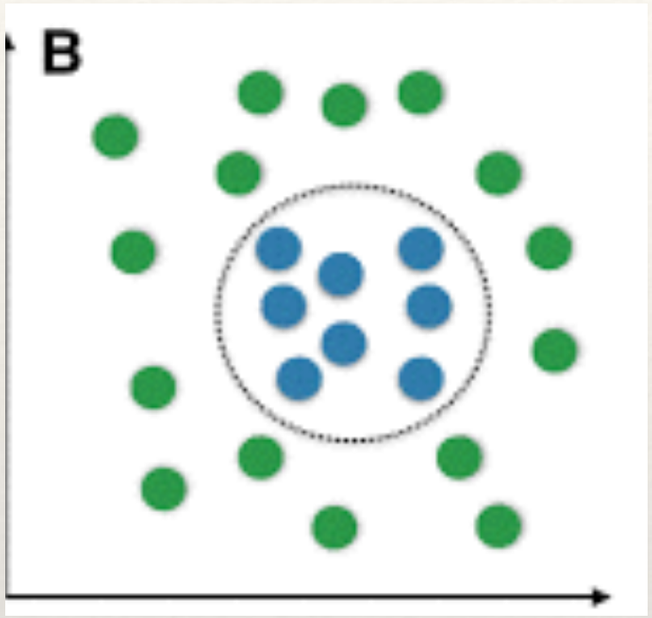


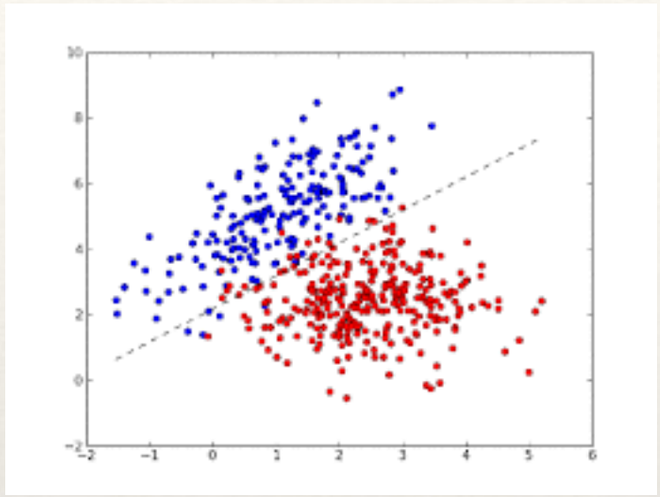
Goal Find a Decision Boundary

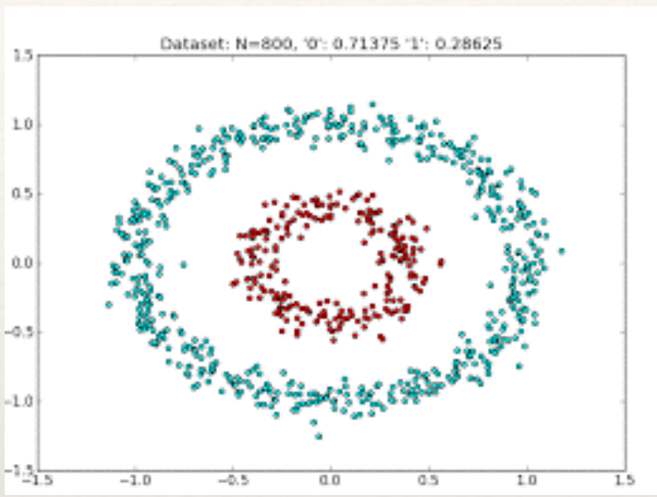
K-Class Similar

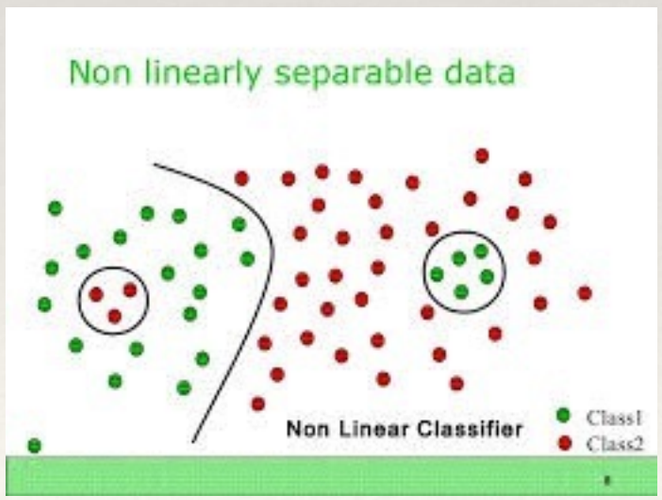


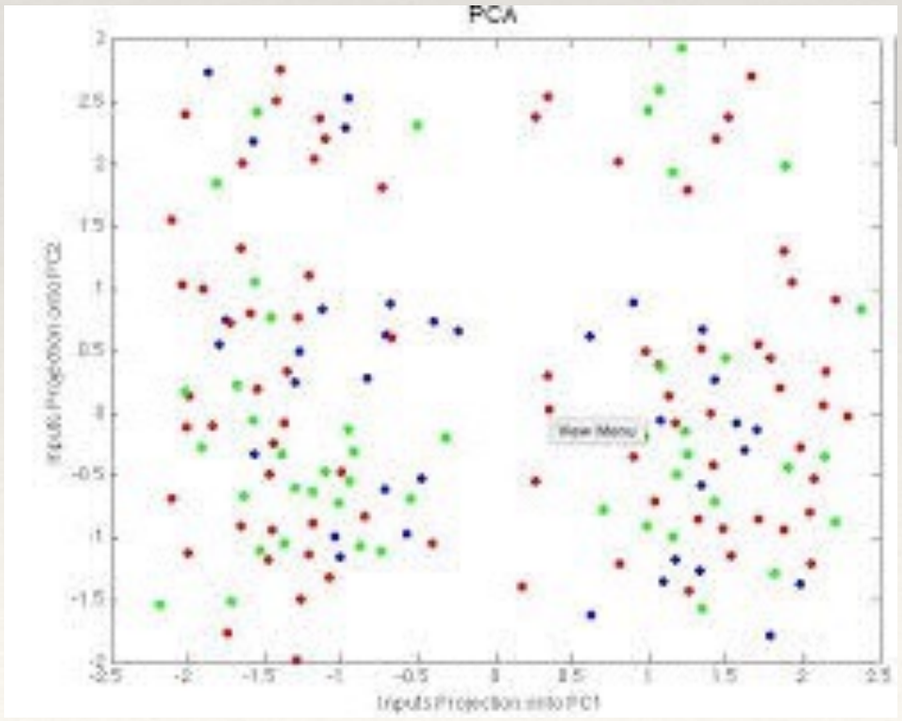
Not Always (Usually Not) So Easy



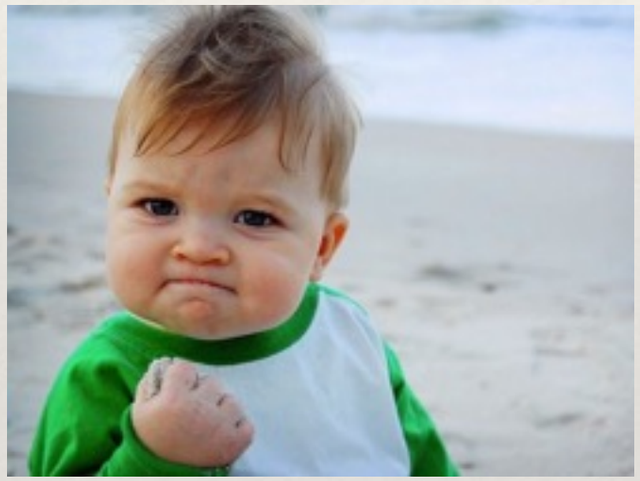


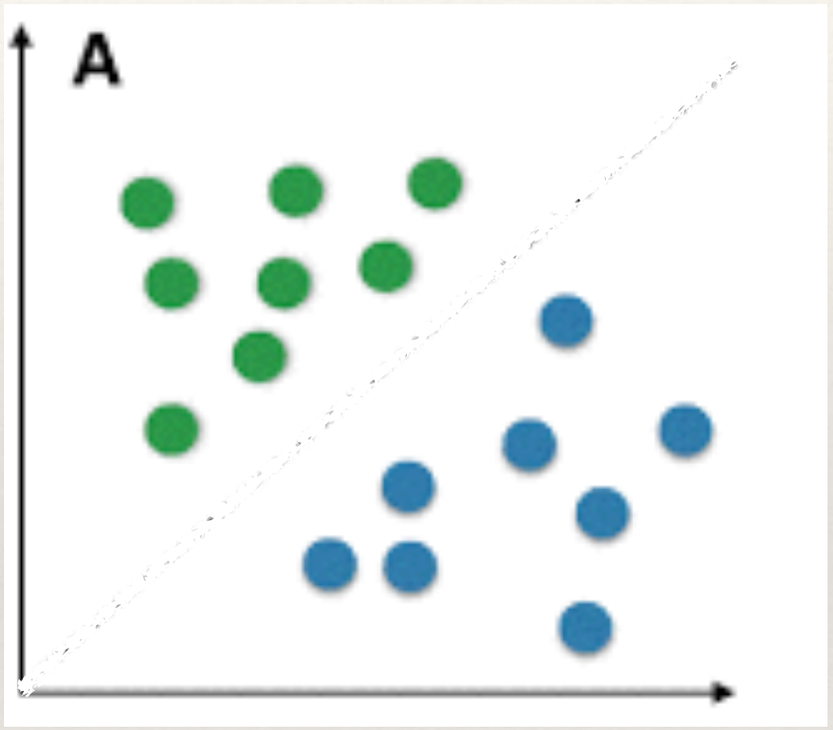






Feature Problem

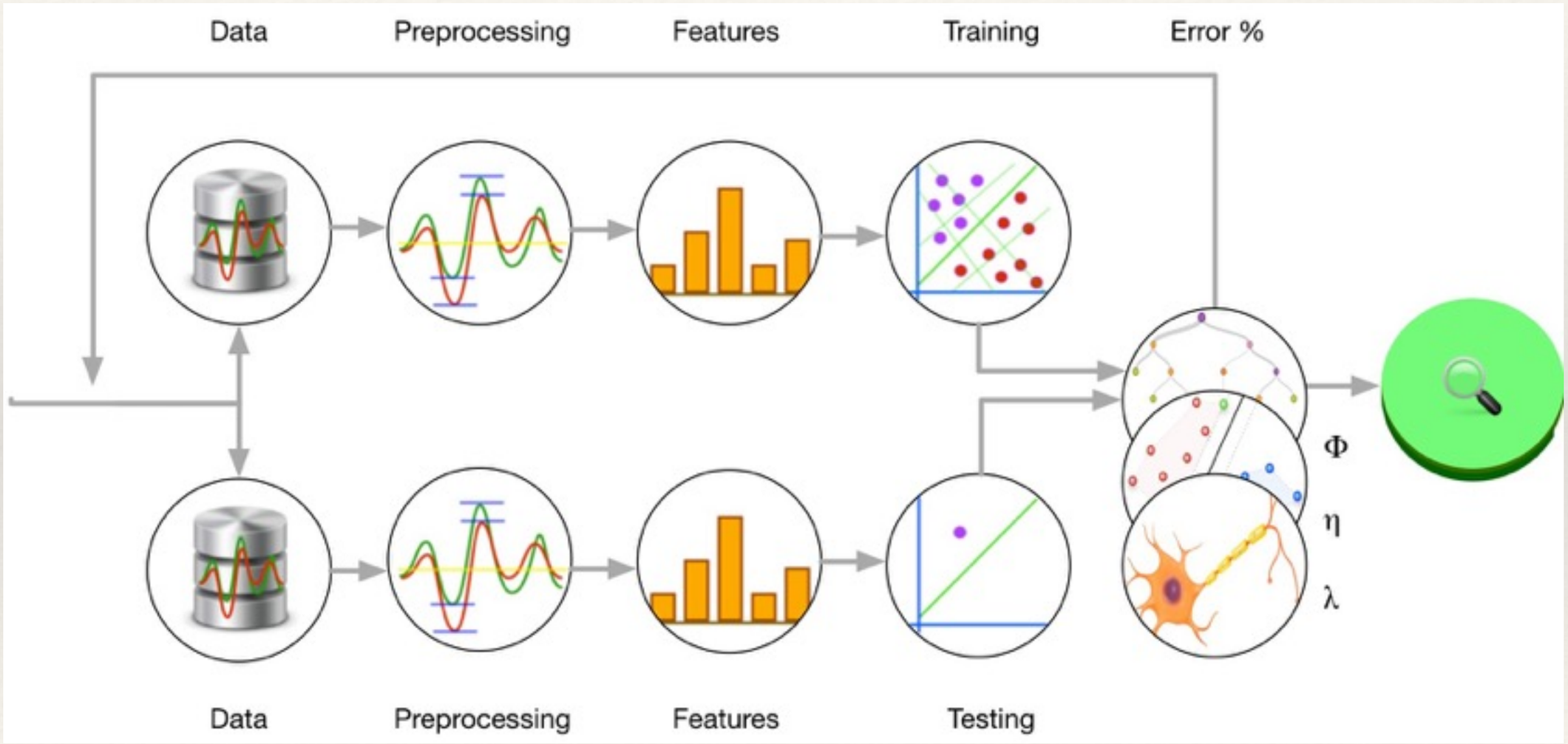




How do you go from this ….

to this?

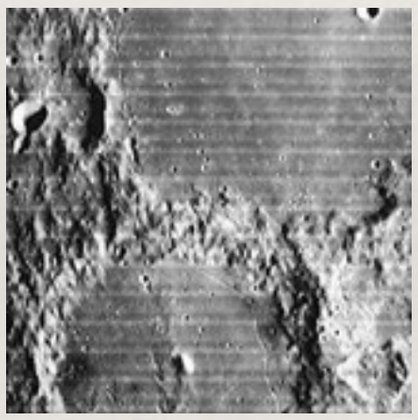
Usual Steps


Data
Preprocessing
Features
Training
Error %
Φ
η
λ
Data
Preprocessing
Features
Testing

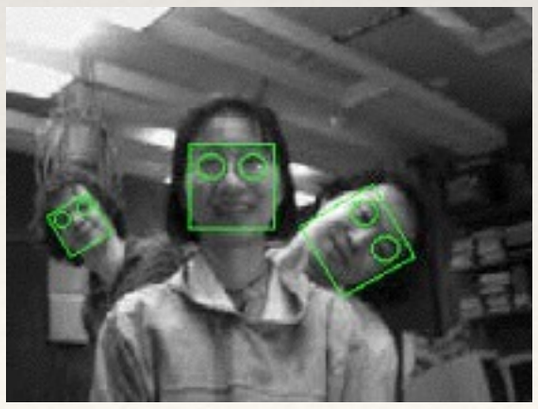
Preprocessing

Data Cleaning

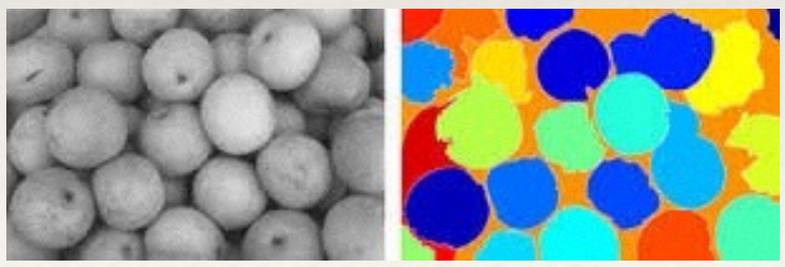
Drop bad rows?
Find bad columns?
Fill in missing data?



De-noise
De-Stripe



Localize
Rotate
Segment

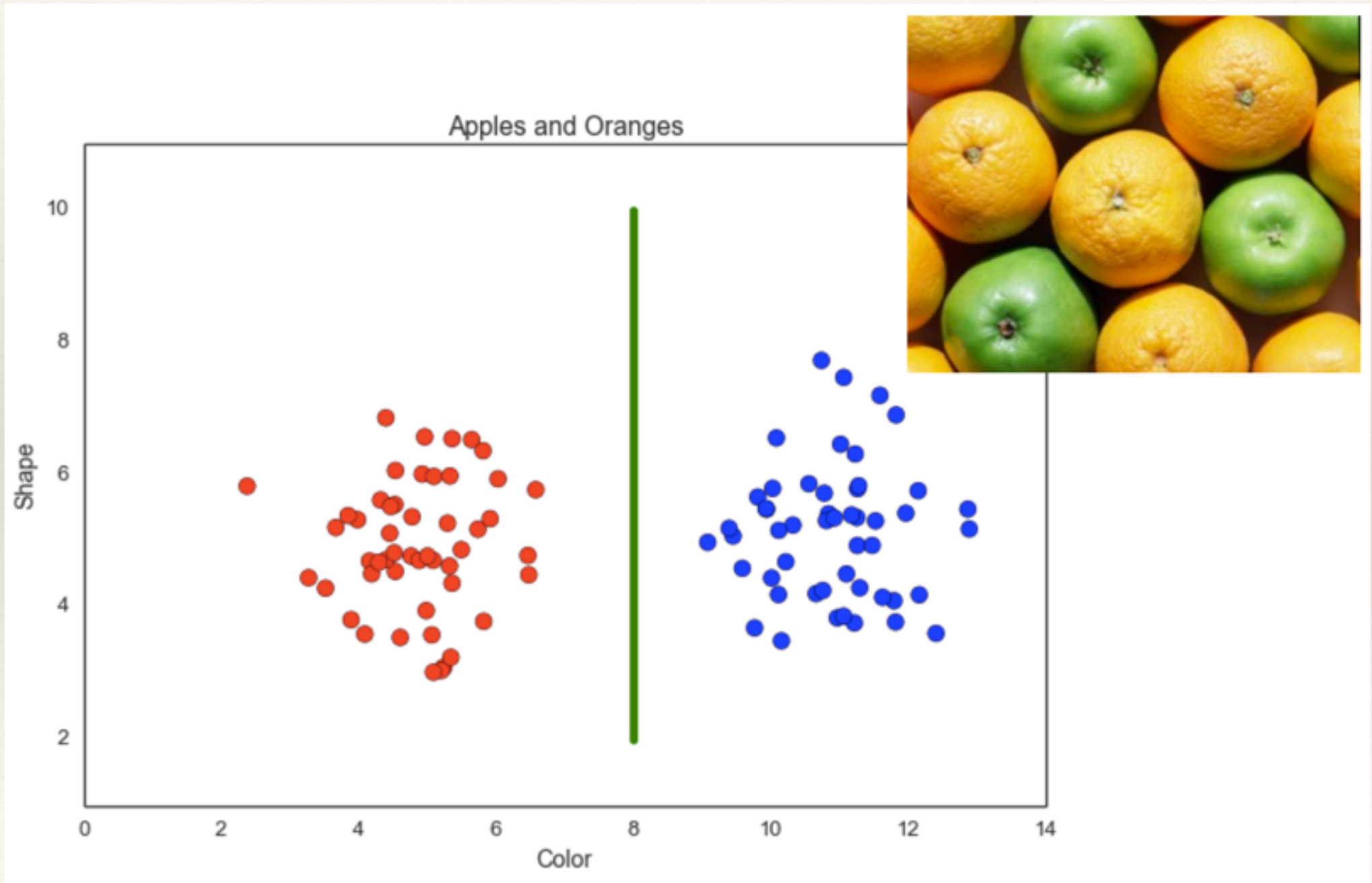


Discovering Good Preprocessing Easily 80% of Work

Finding Features (preprocessing?)

Bad Features

Good Features



Finding Good Features

Still Can be Alchemy….

Some Approaches:

- Dimension Reduction
  - PCA
  - ICA
  - Matrix Factorization
  - LDA
- Anova
- Lasso
- Deep Learning (feature learning)

Iris Data Set

Attribute Information:

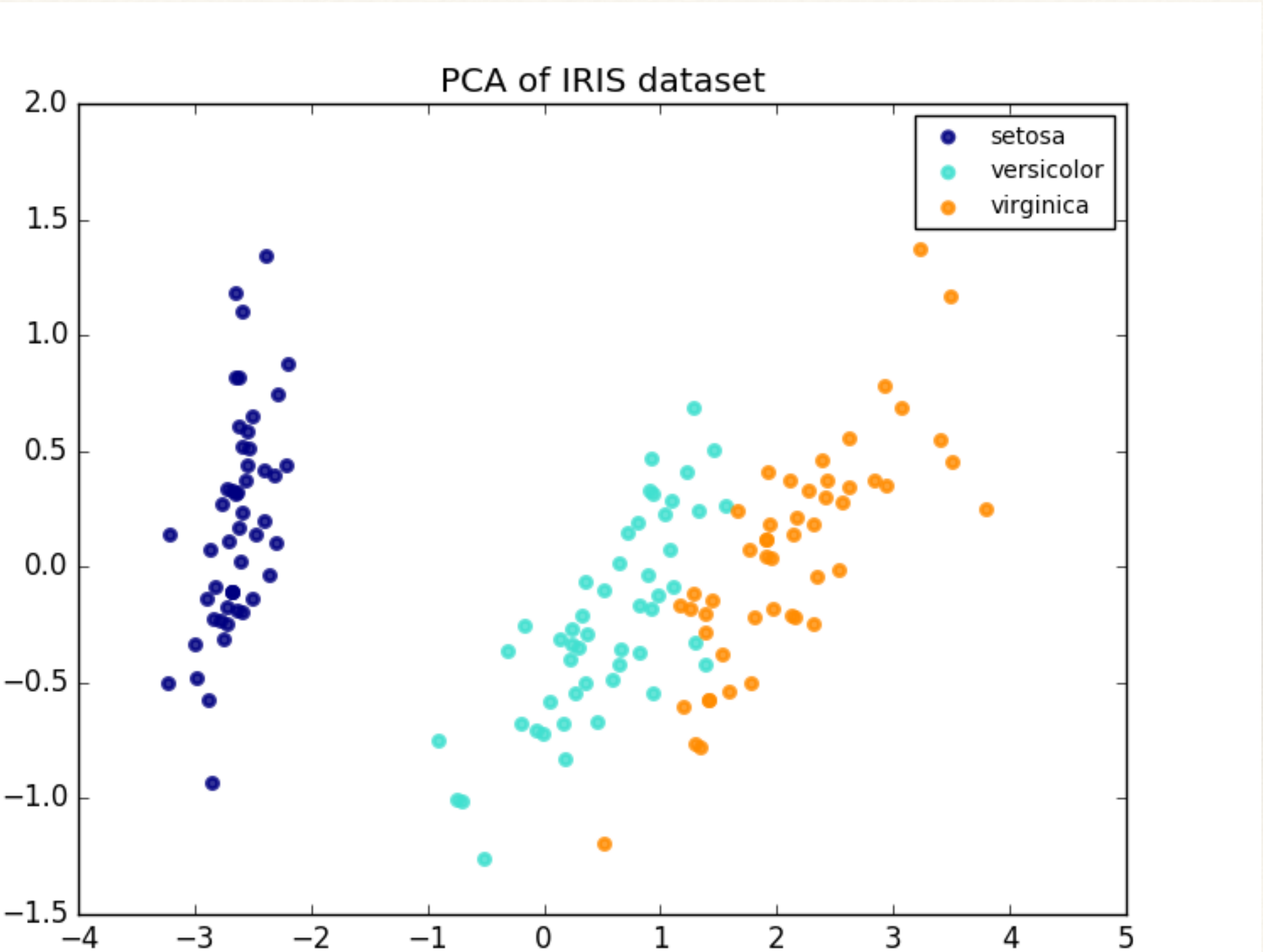
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

PCA


PCA of IRIS dataset
setosa
versicolor
virginica
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3
4
5

Classification: KNN

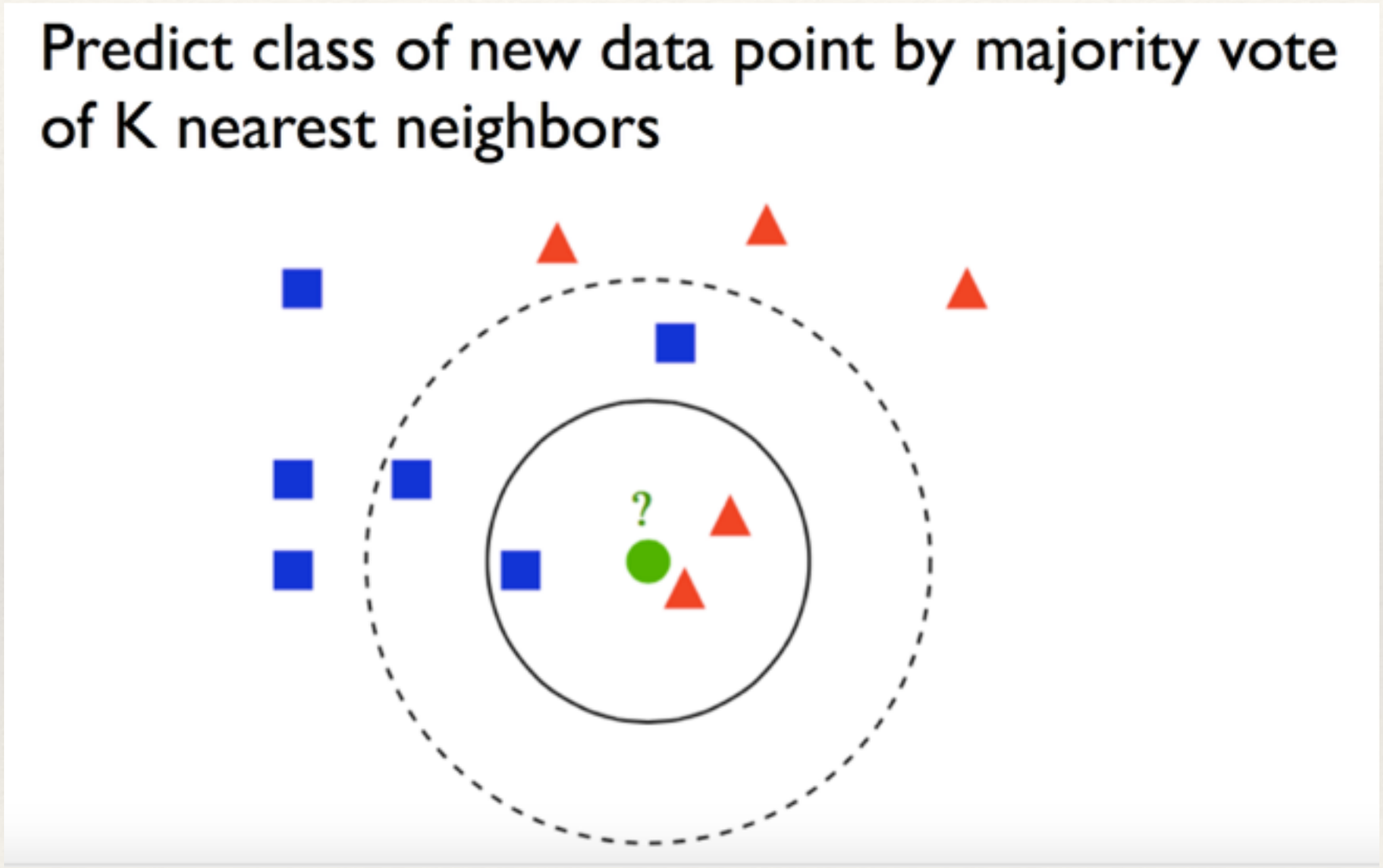
Two Simple ("Brain Dead") algorithms:

KNN

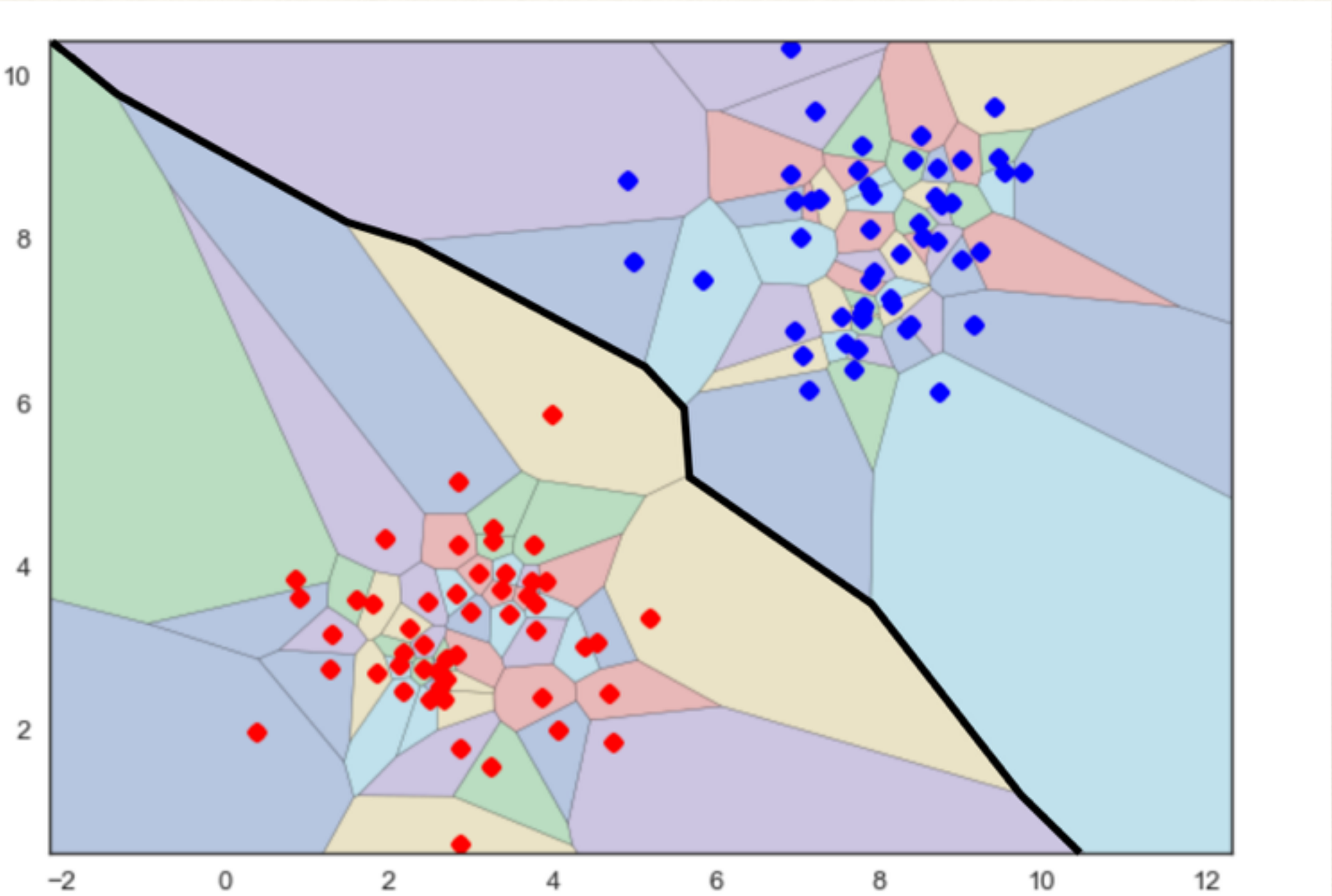
Linear Discriminant Analysis

Always Try this first!!!!

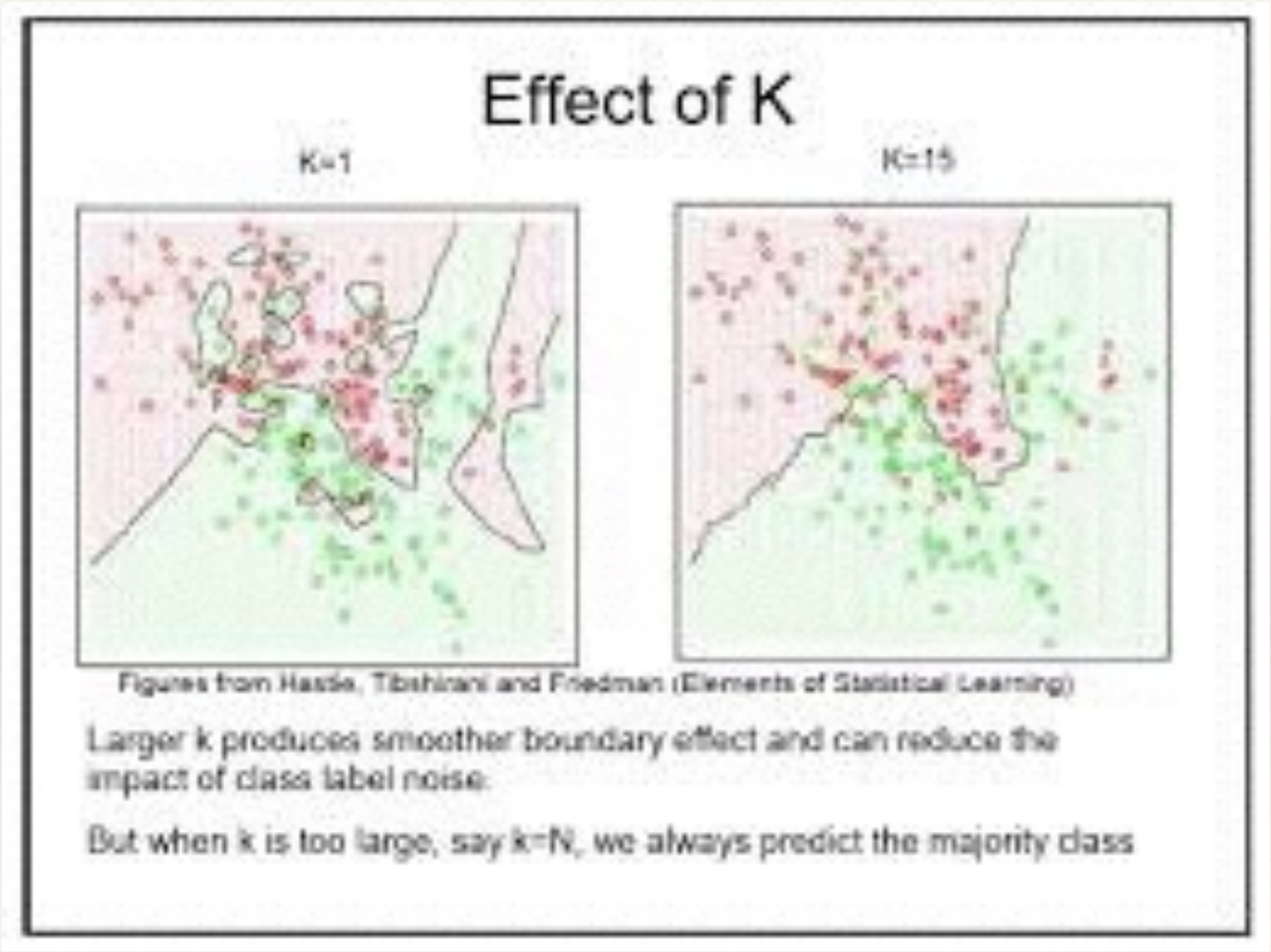
KNN


Predict class of new data point by majority vote of K nearest neighbors
?

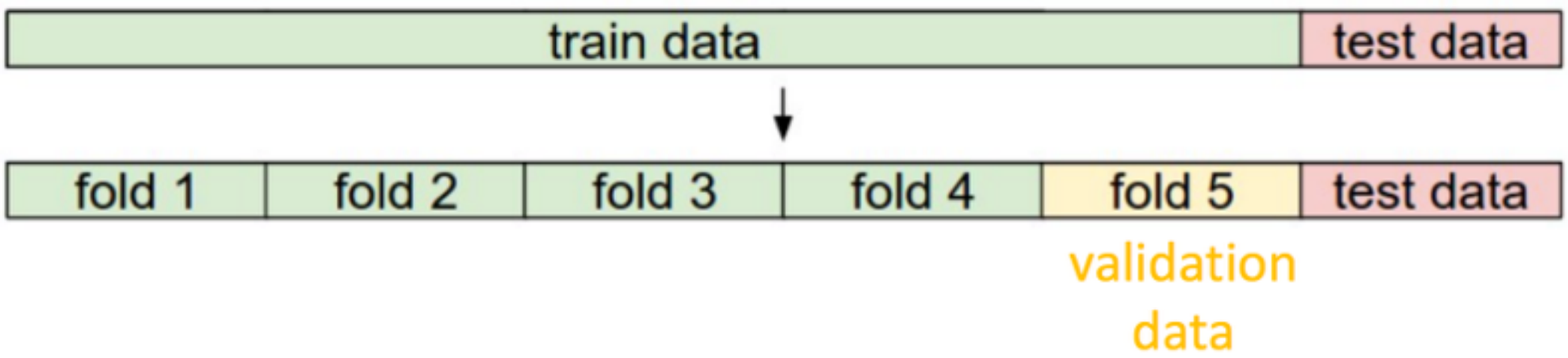
Voronoi Cells



Effect of K


Effect of K
K=1
K=15
Figures from Hastie, Tibshirani and Friedman (Elements of Statistical Learning)
Larger k produces smoother boundary effect and can reduce the impact of class label noise.
But when k is too large, say k=N, we always predict the majority class

Evaluate With Cross Validation



- Training data: train classifier
- Validation data: estimate (hyper) parameters (k)
- Test data: measure performance