Introduction to machine learning

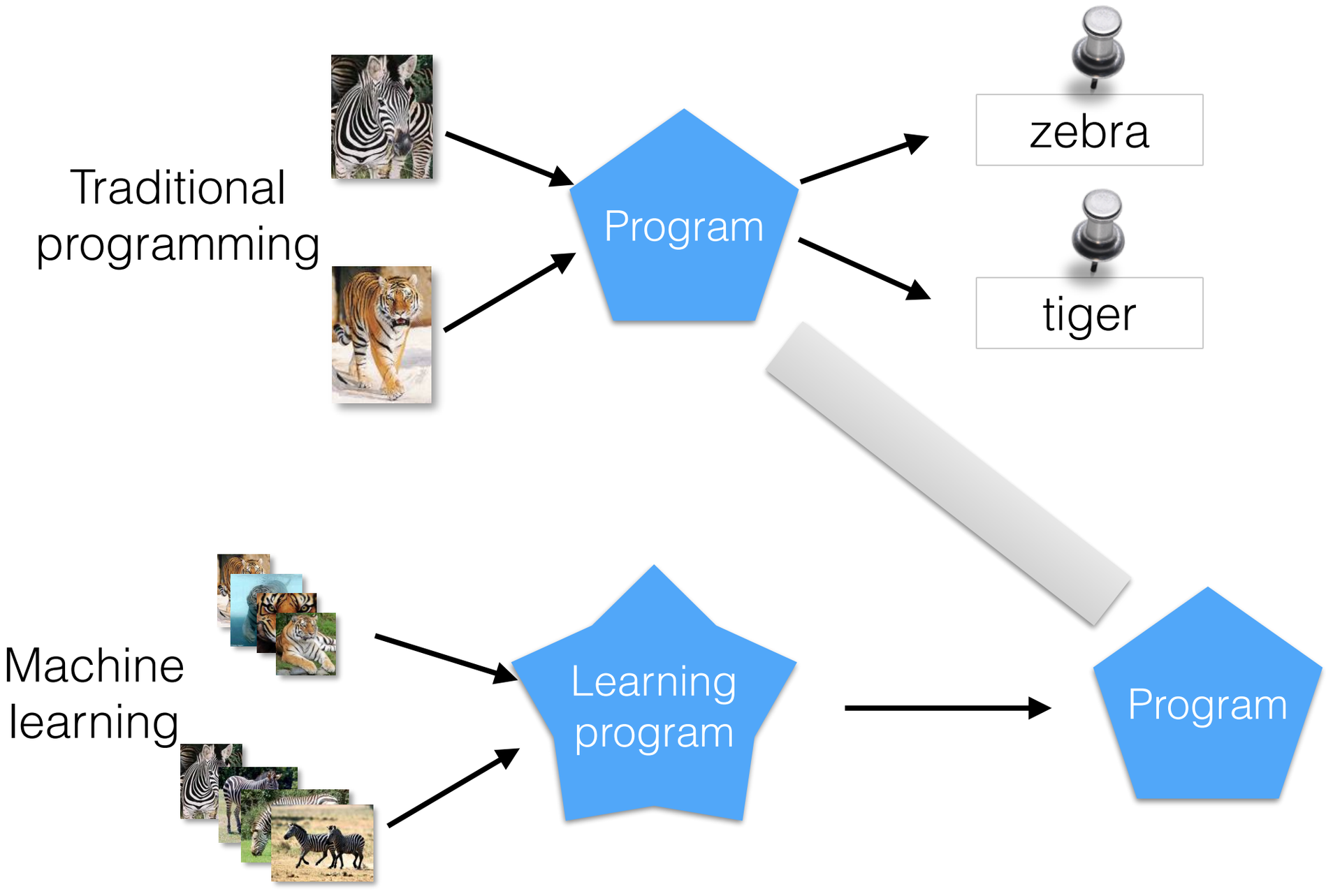
TANG Gen

Outline

- What is machine learning?
- Unsupervised learning
- Supervised learning
  - GLM
  - Naive Bayes
  - Decision trees and forests
- Dimension reduction
- Statistics vs. machine learning

What is machine learning?

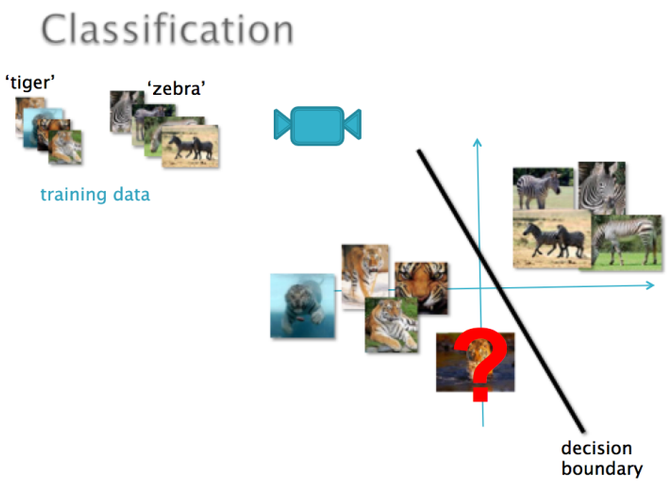
/A simple description

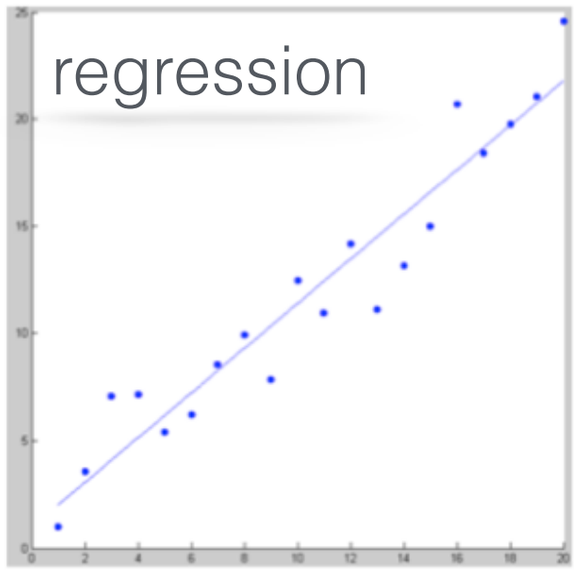


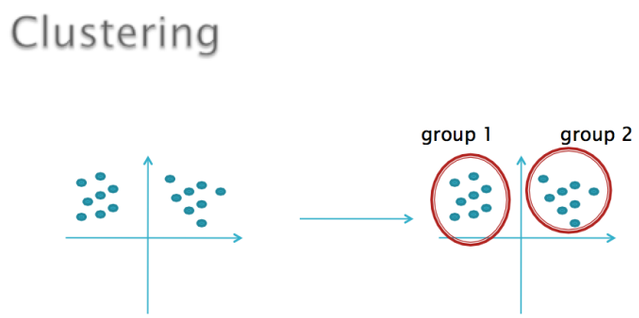
What is machine learning?

/Formalized definition

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed
  - According to the types of data: unsupervised and supervised (semi-supervised)
  - According to the purpose: cluster, regression and classification







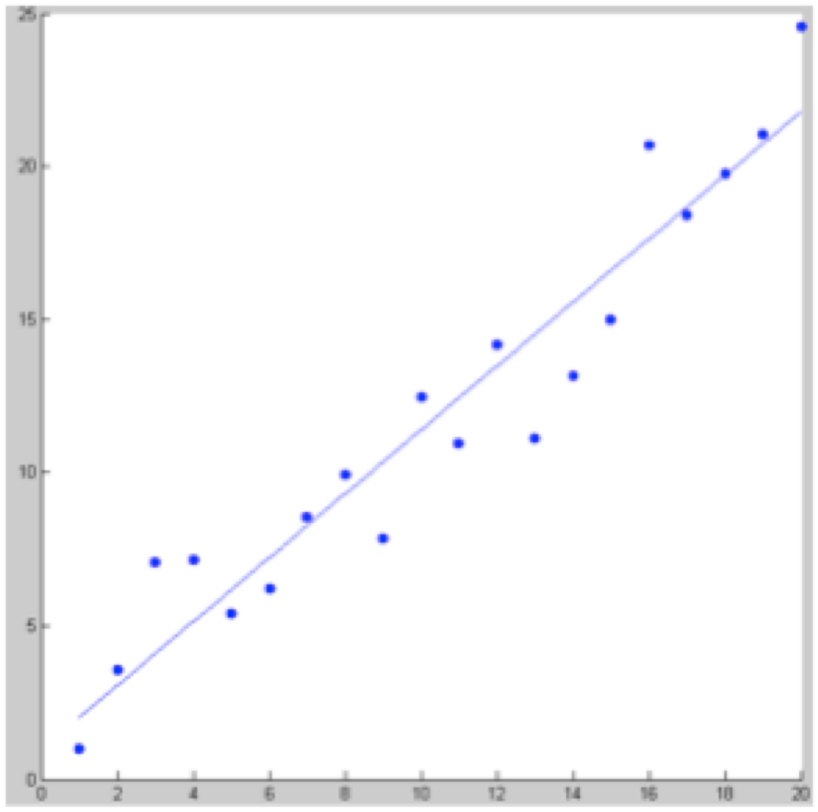
What is machine learning

/Supervised vs. unsupervised

- Supervised learning: Given $(x_i, y_i), i = 1, \ldots, n$, learn a function $f : X \to Y$.
  - Categorical Y: classification
  - Continuous Y: regression
- Unsupervised learning: Given only $(x_i), i = 1, \ldots, n$, can we infer the underlying structure of X?

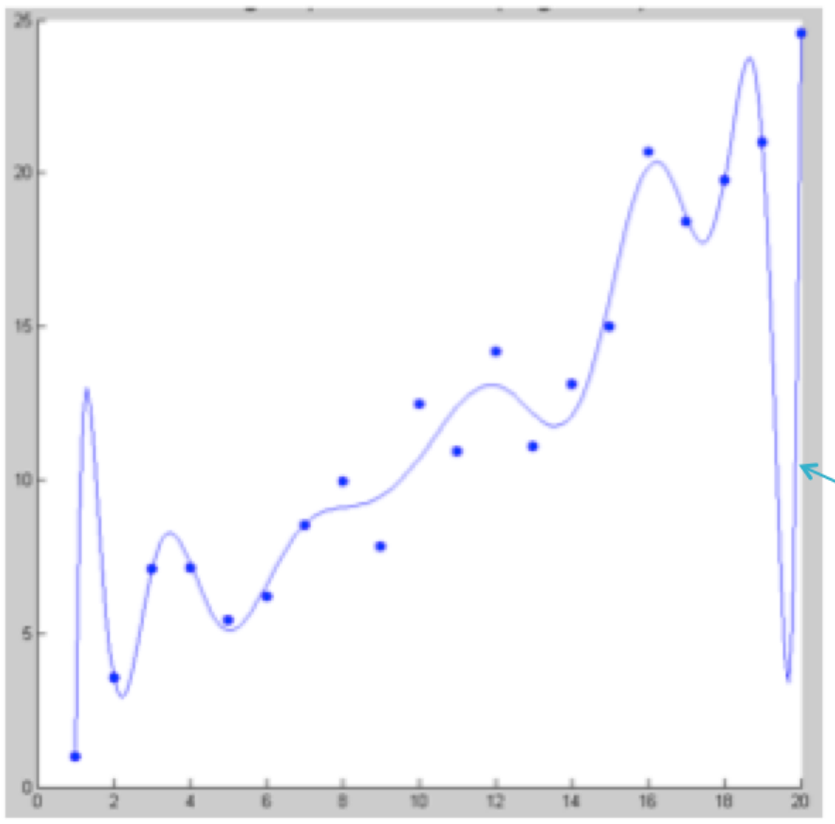
What is machine learning?

/Overfitting



linear model:

$a = w_0 + w_1 * F$



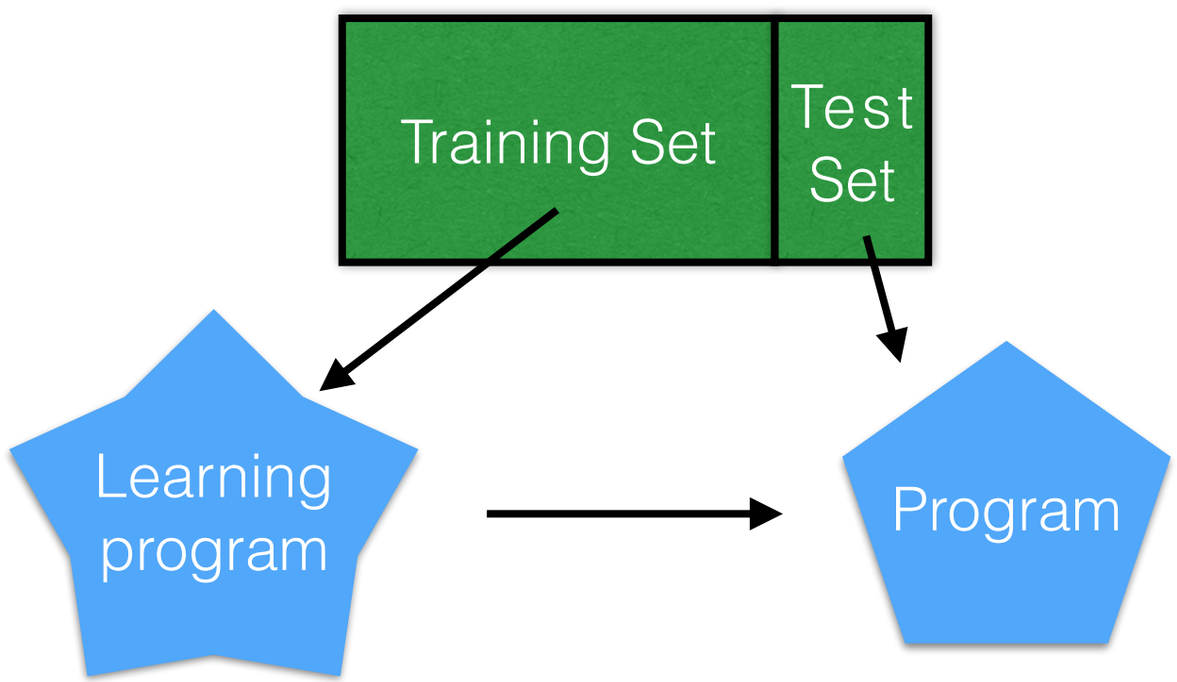
degree 15

overfitting!

What is machine learning?

/Cross-validation

- In order to overcome the problems of overfitting, we use cross validation technique to do model selection
  - The main idea is to divide data into two parts: training set and test set;
  - We use training set to estimate parameters of model and then use test set to estimate the quality of model



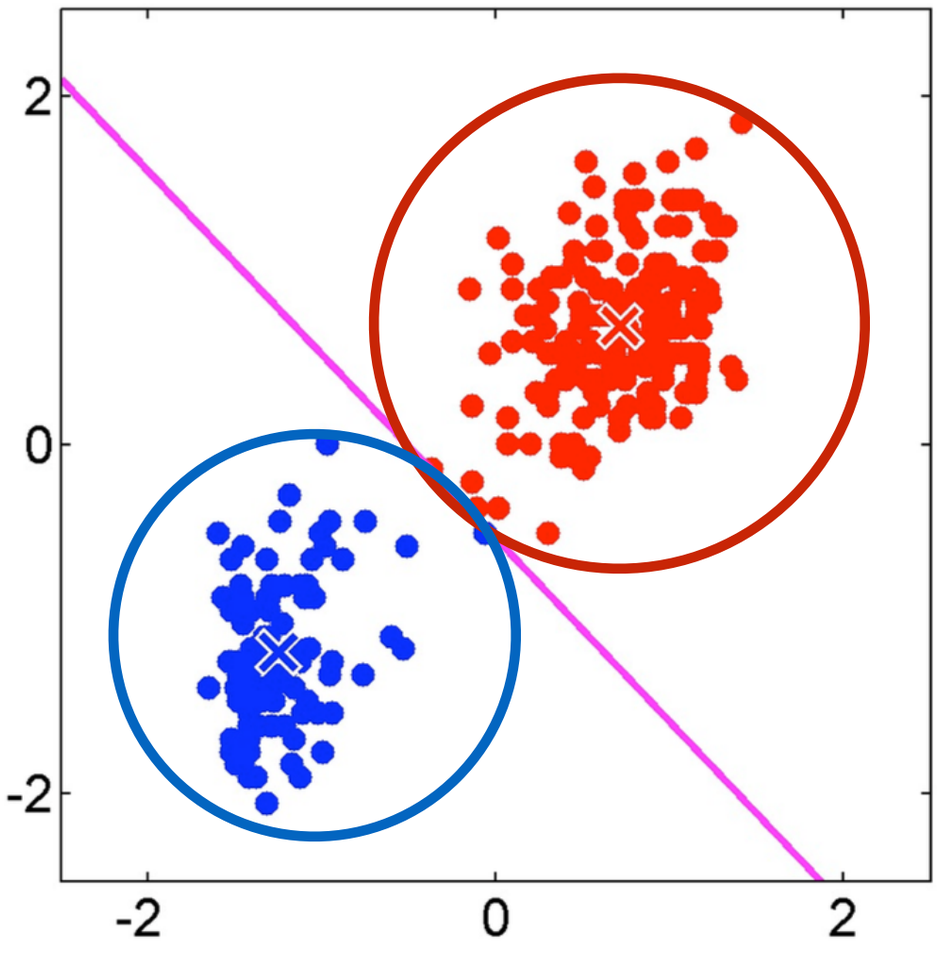
Unsupervised algorithms

/Clustering

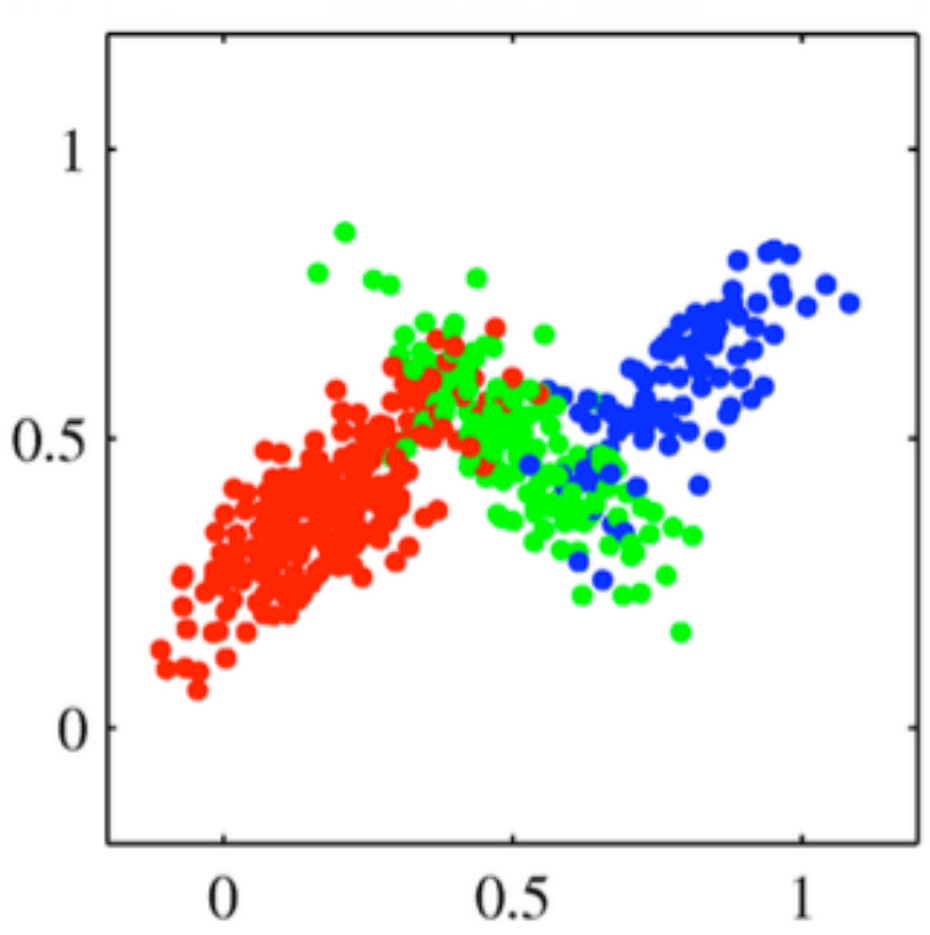
- Utility of unsupervised algorithms:
    - Raw data cheap. Labeled data expensive
    - Save memory/computation
    - Reduce noise in high-dimensional data
    - Useful in exploratory data analysis
    - Often a pre-processing step for supervised learning

Unsupervised learning

/Clustering



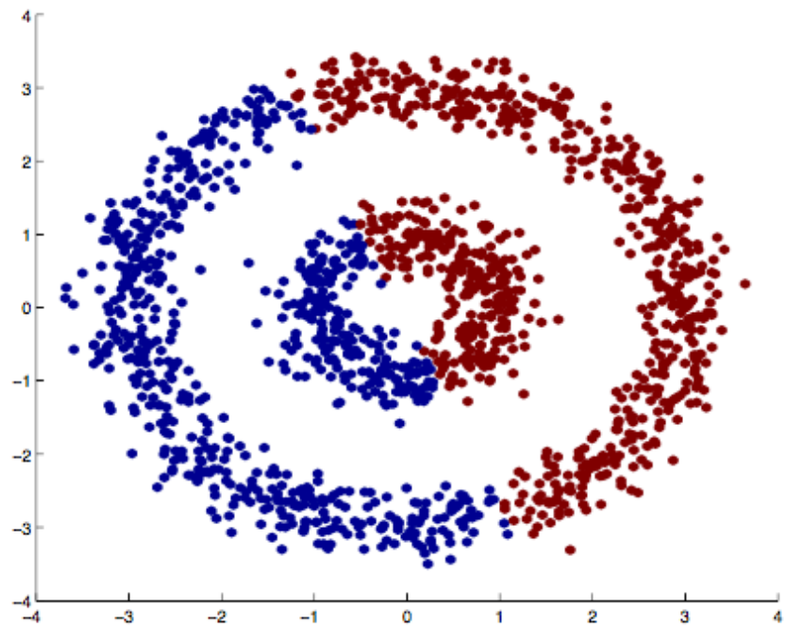
K-means

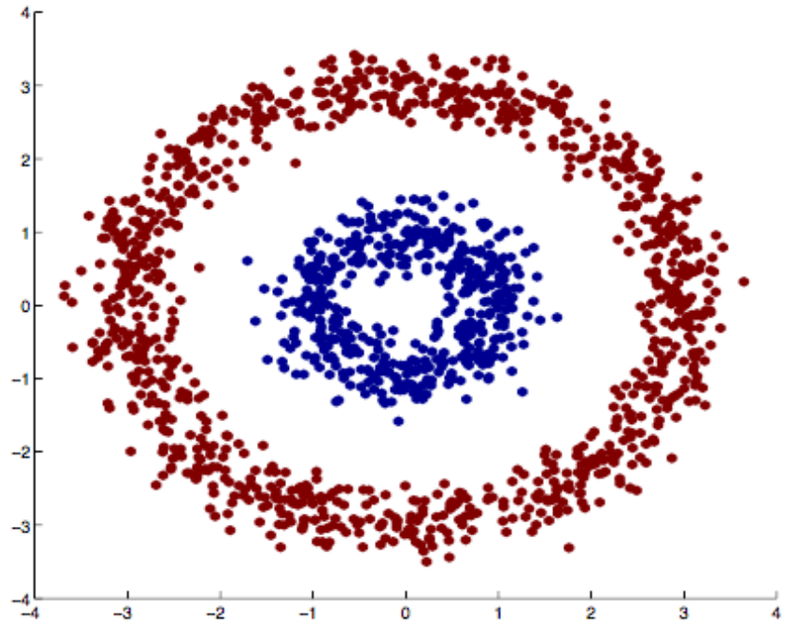


GMM

Unsupervised learning

/Clustering





K-means, K=2

Spectral clustering

Supervised algorithms

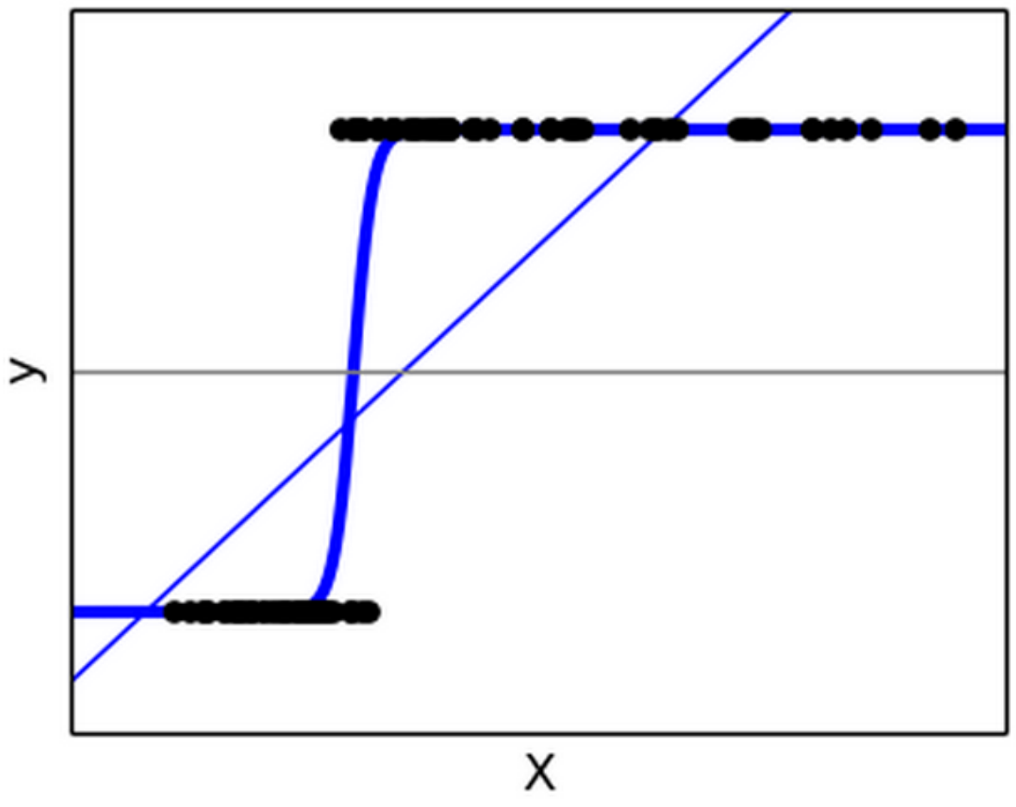
/Generalized linear models

- GLMs are the most used algorithms in real world
- Usually they are used to binary classification or linear regression
  - Logistic regression (Kernel regression)
  - Support vector machine
  - Lasso and ridge regression (overcome overfitting of linear regression)
  - Streaming regression

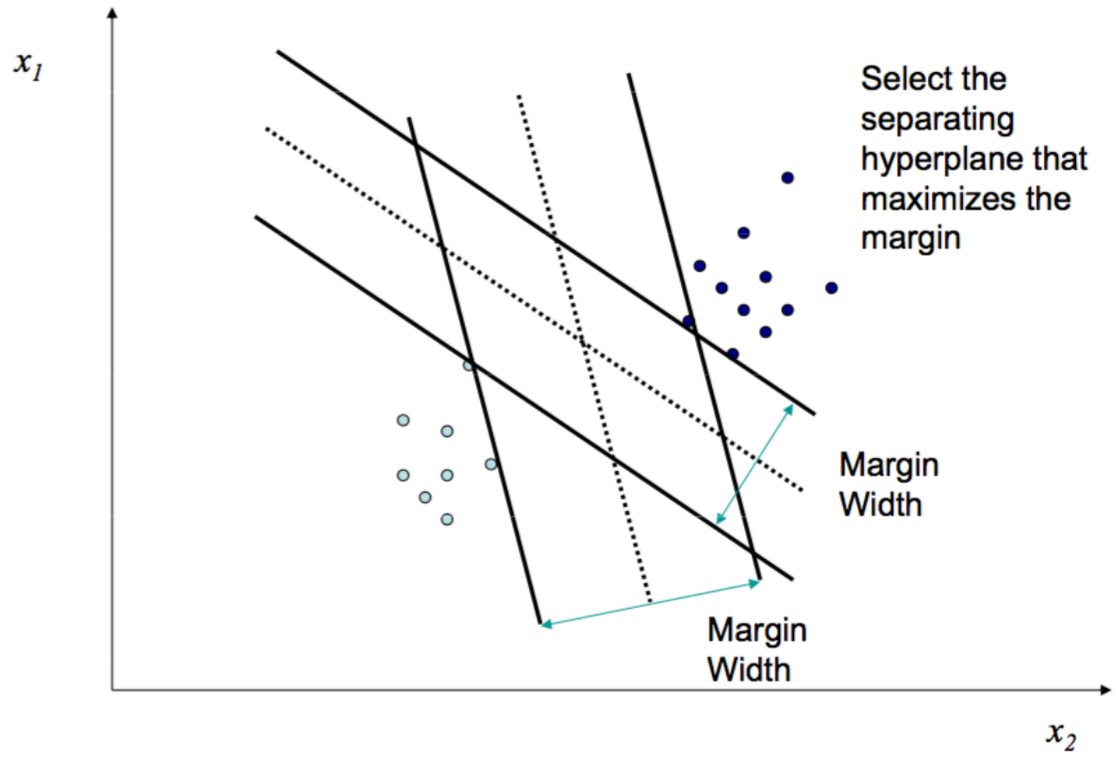
Supervised algorithms

/Generalized linear model

Logistic regression



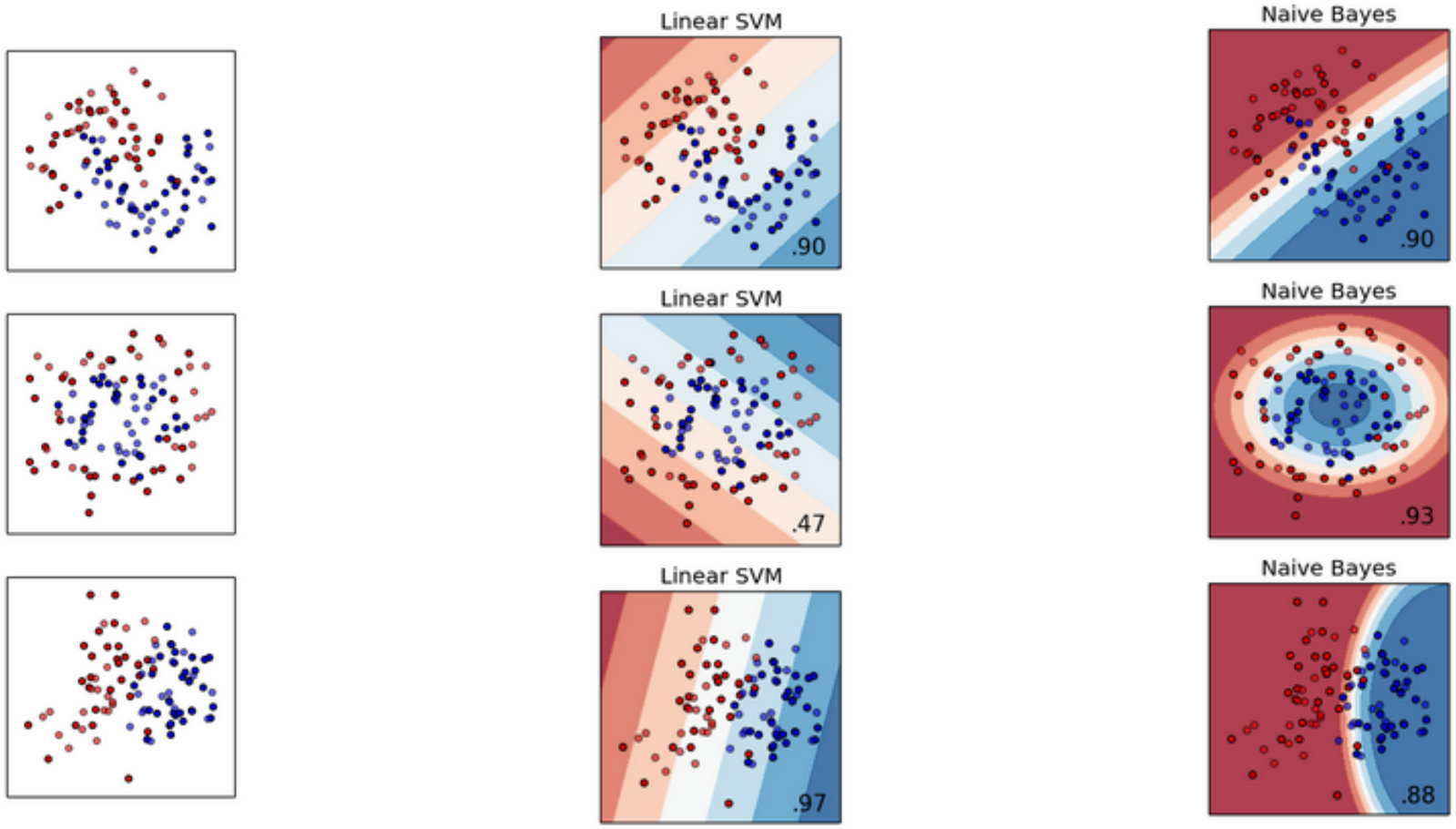
SVM



Supervised algorithms

/Naive Bayes

- Naive Bayes is a probabilistic model
- It is typically used for document classification

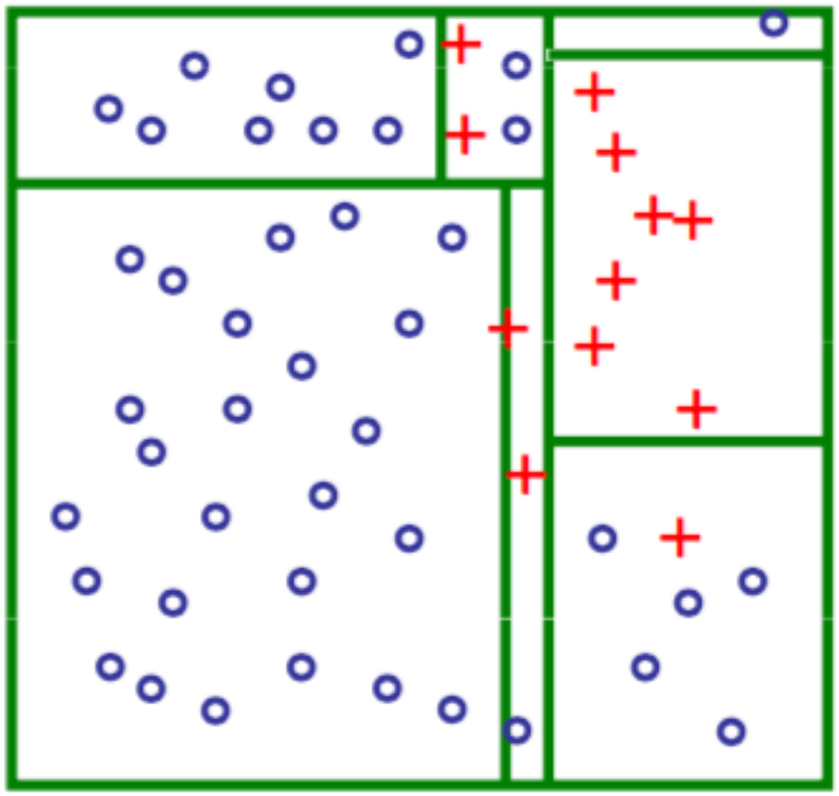


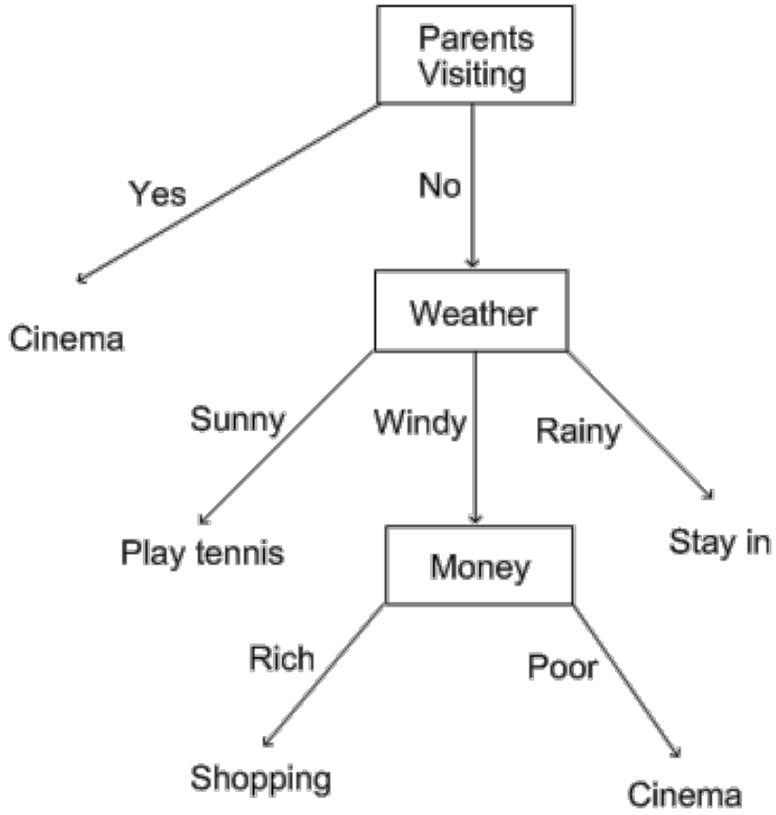
Supervised algorithms

/Decision trees and forests

- Decision trees and forests are popular methods for the machine learning tasks of classification and regression

- Decision trees are widely used since they are **easy to interpret**



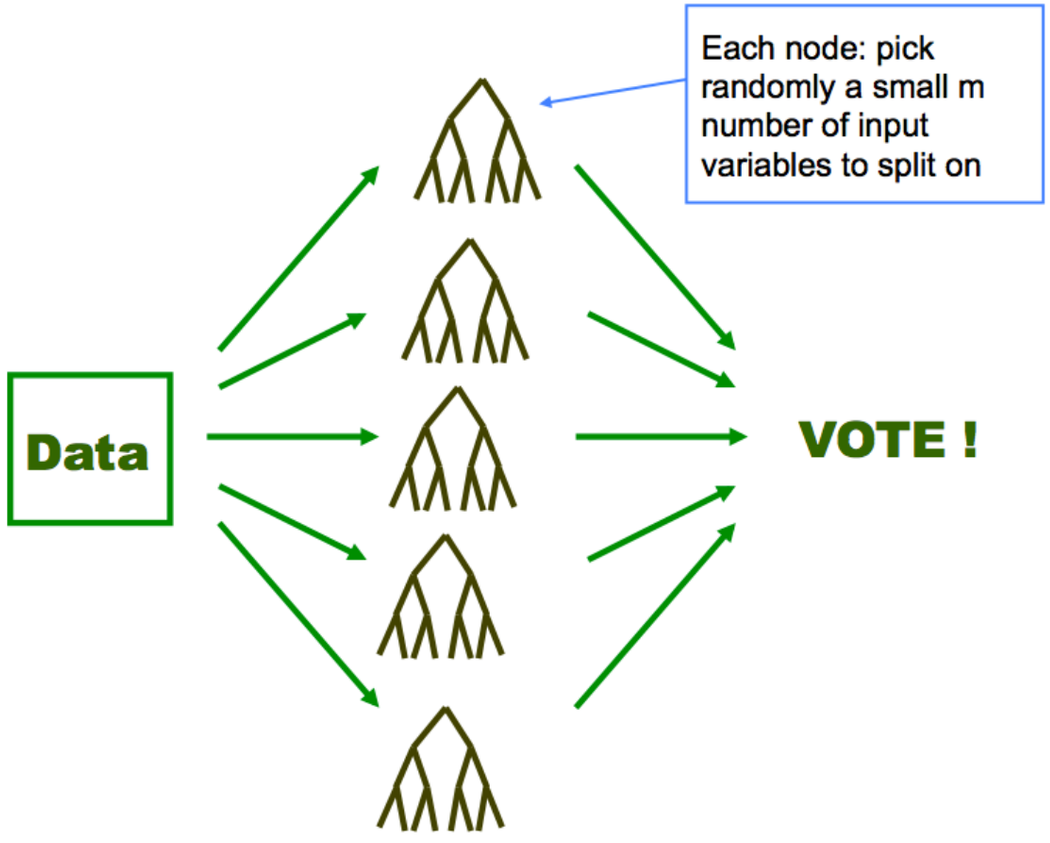


Supervised algorithms

/Decision trees and forests

- Forests is an ensemble of learning algorithm which creates a model composed of a set of decision trees

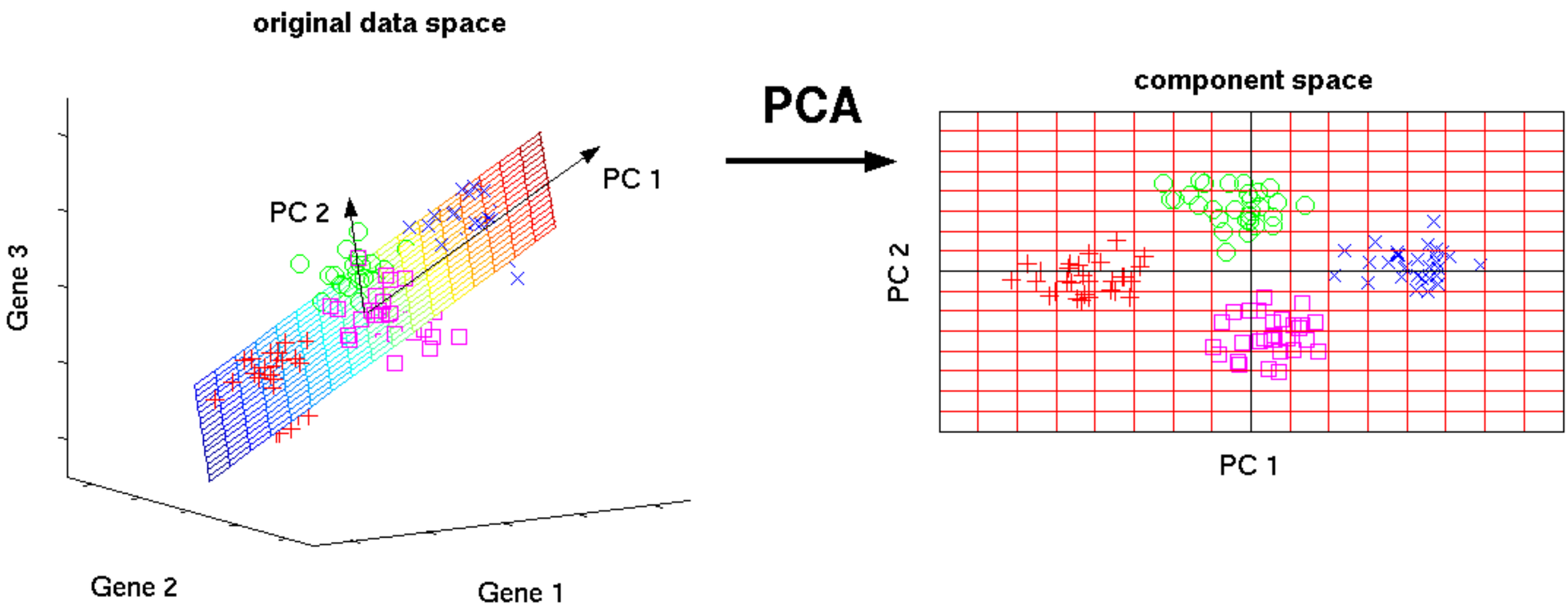
- Forests' algorithms are very powerful to predict, but not easy to interpret



Dimension reduction

/PCA

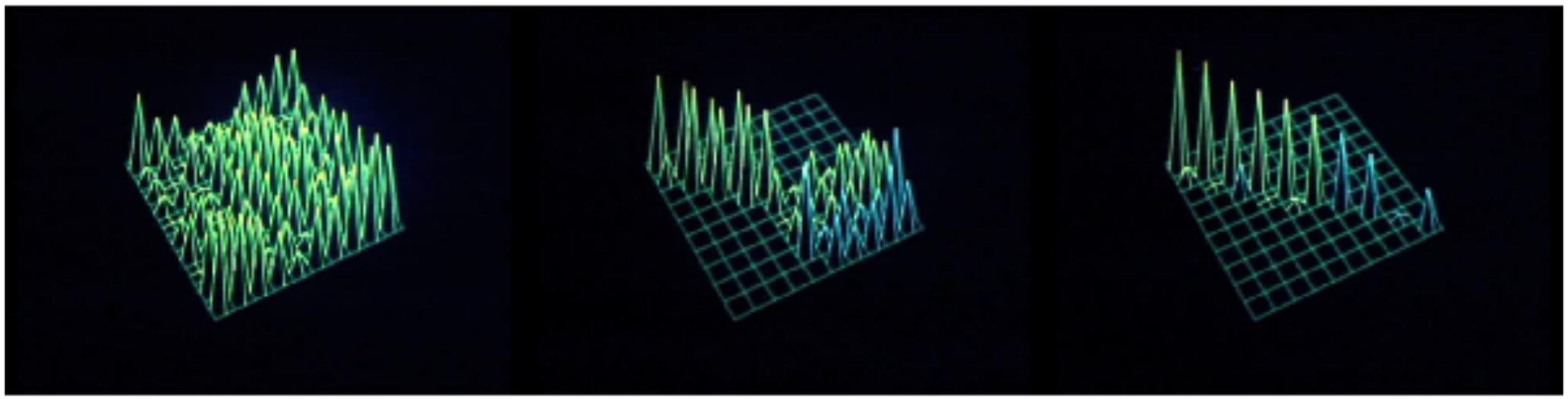
- Principal component analysis
- Data visualization or dimension reduction



Dimension reduction

/SVD

- Singular value decomposition
- Noise reduction or feature selection



Statistics vs. machine learning

/Larry wasserman's blog

Statistics	Machine learning
Estimation	Learning
Classifier	Hypothesis
Data point	Example/Instance
Regression	Supervised learning
Classification	Supervised learning
Covariate	Feature
Response	Label

And more important: Statisticians use R and machine learners use Python

Tips for MLlib

- All the algorithms that we discussed above are all included in MLlib
- Pay attention to the parallelism level of RDD passed to MLlib's algorithm. It can largely influence the computation speed
- Pay attention to shuffle volume which could to too large for a cluster
- If it is possible, put the data in memory which can accelerate the speed

Case studies

/MLlib in real world

- Collaborative filtering (ALS) for music recommendation

- Show a progression of code rewrites, converting a Hadoop-based app into efficient use of Spark