Introduction to Machine Learning

DSCI6003

galvanize

- What is Machine Learning?
    - What?
    - Why?
    - ML vs. Statistics
- Types of Learning
    - Supervised
    - Unsupervised
    - Semi-Supervised (ML2)
    - Reinforcement
- Supervised Learning
    - 3 Components of a model
    - ML Process

- Lab: Predicting Interest Rates

> Field of study that gives computers the ability to learn without being explicitly programmed.

-Arthur Samuel circa 1959

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
>
> -Tom M. Mitchell

Machine learning is *NOT*:

- Hard coded logic by programmer: **if**s and **else**s...
- Predefined results: completely deterministic
- Burden is placed on programmer at design time
- Must anticipate all inputs to program, and react

Machine learning *is*:

- Automated knowledge acquisition through input
- Iterative improvement as more data is seen
- Adaptive Algorithms

Regression:

- Loan interest rate prediction
- Utilities: smart grid load forecasting
- Web: page traffic prediction
- Advertising CTR prediction

Classification:

- Spam Filtering and document classification
- Finance: Fraud detection and loan default prediction
- Sentiment Analysis: People like to do this with Tweets
- National Security: ??? PRISM!

Clustering:

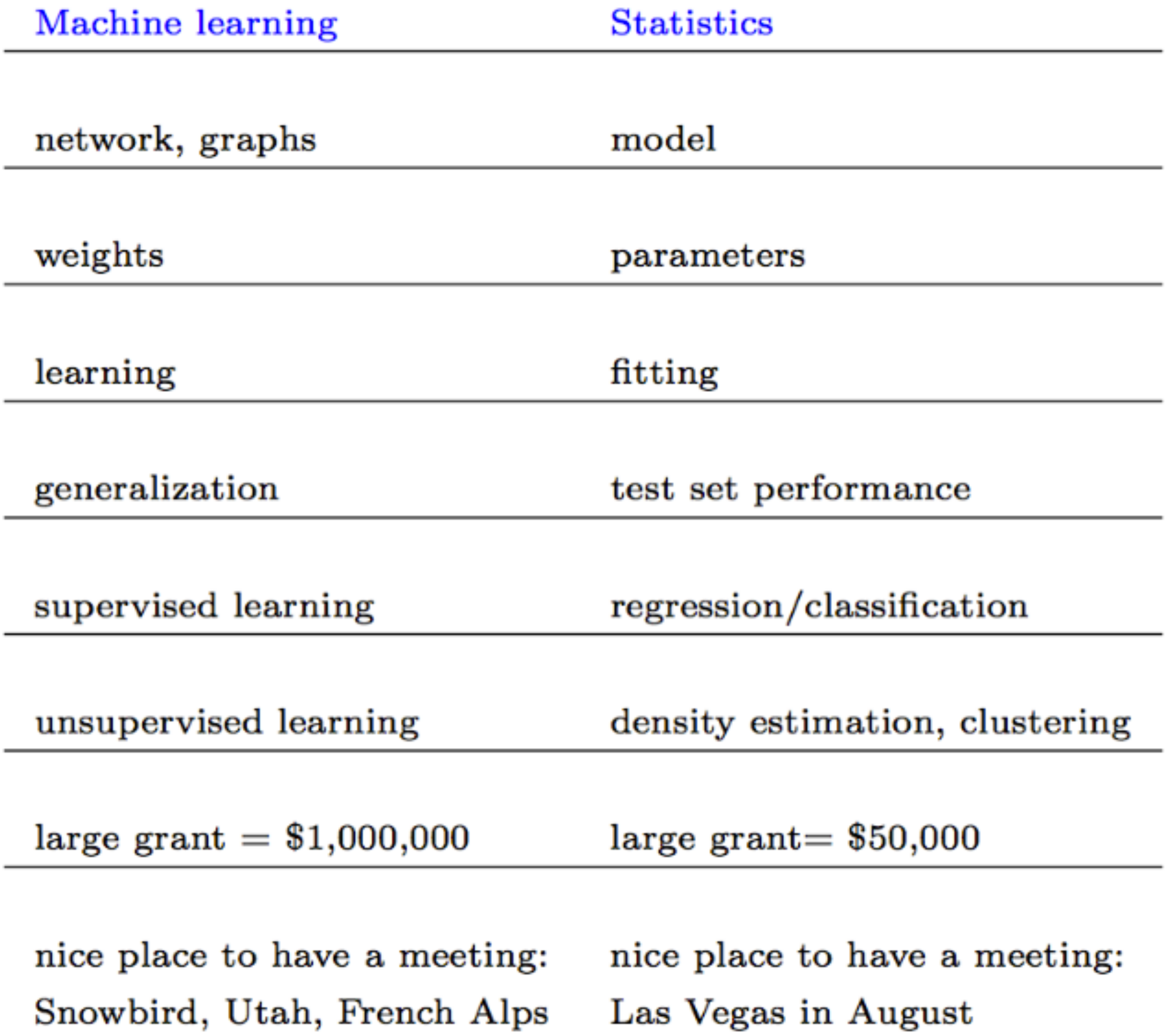
- Product Marketing: Cohort Analysis
- Oncology: Malignant cell identification
- Computer Vision: entity recognition
- Census: demographics analysis

https://www.kaggle.com/wiki/DataScienceUseCases

Quiz: Which category would Churn prediction fall into?

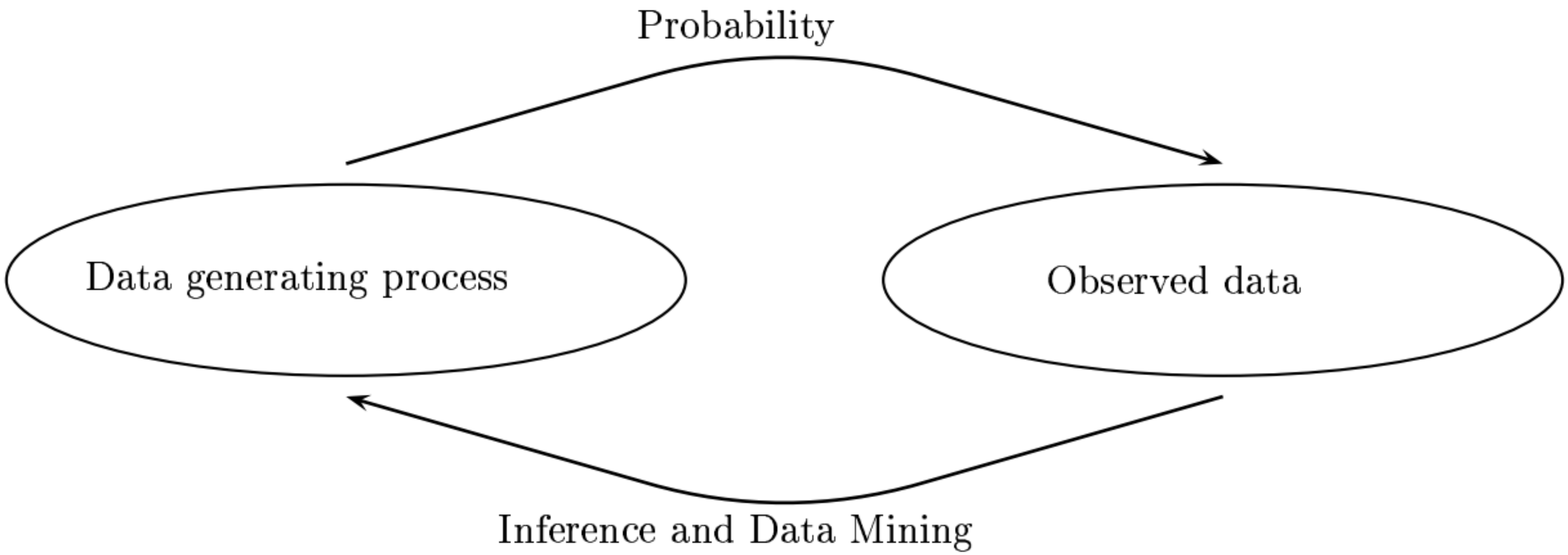
- Red: Clustering
- Blue: Classification
- Green: Regression
- Yellow: Other

MACHINE LEARNING VS. STATISTICS

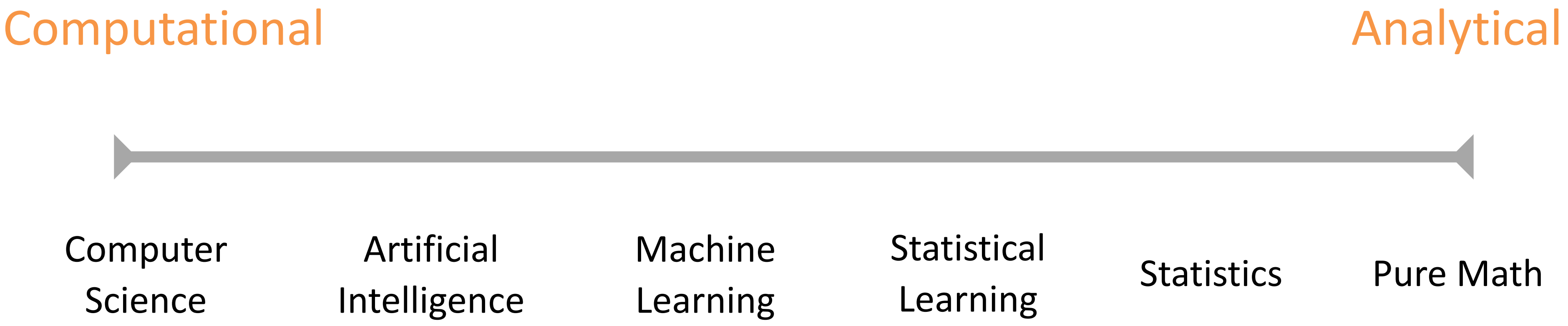


Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

http://datavu.blogspot.com/2014/08/statistical-modeling-vs-machine-learning.html


Probability
Data generating process
Observed data
Inference and Data Mining

The Spectrum of the Learning Arts



Computational

Analytical

Computer Science

Artificial Intelligence

Machine Learning

Statistical Learning

Statistics

Pure Math

TYPES OF LEARNING

How can computers learn??!?

Supervised Learning

- Training Data includes desired output

Unsupervised Learning

- Training Data does not include desired output

Semi-supervised Learning

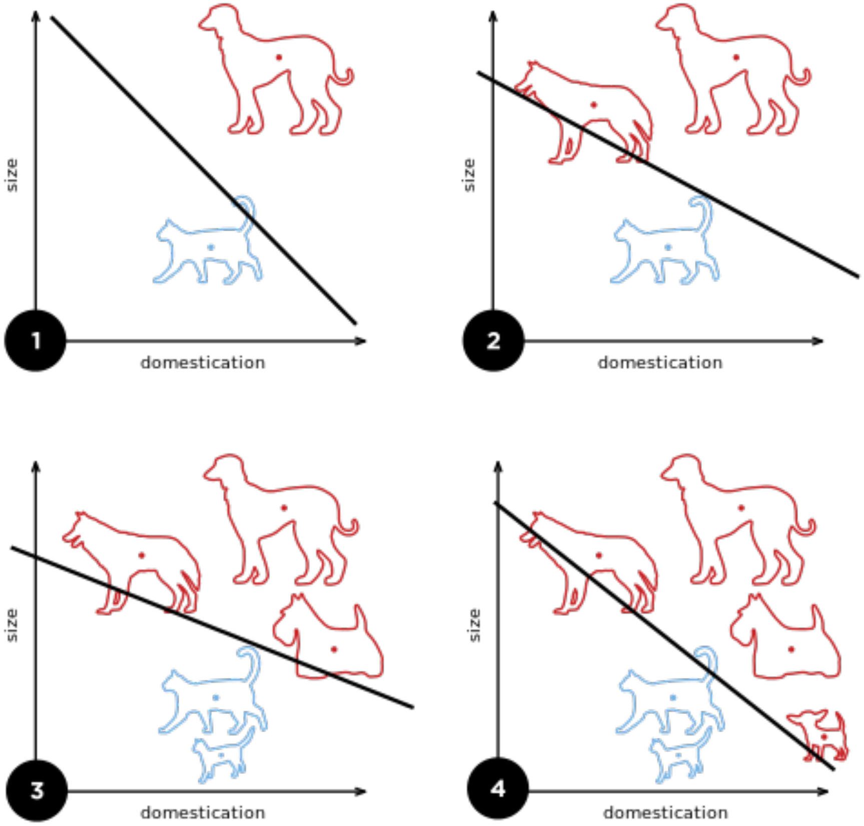
- Training Data includes a few desired outputs

Reinforcement Learning

- Rewards from sequence of actions

Supervised Learning

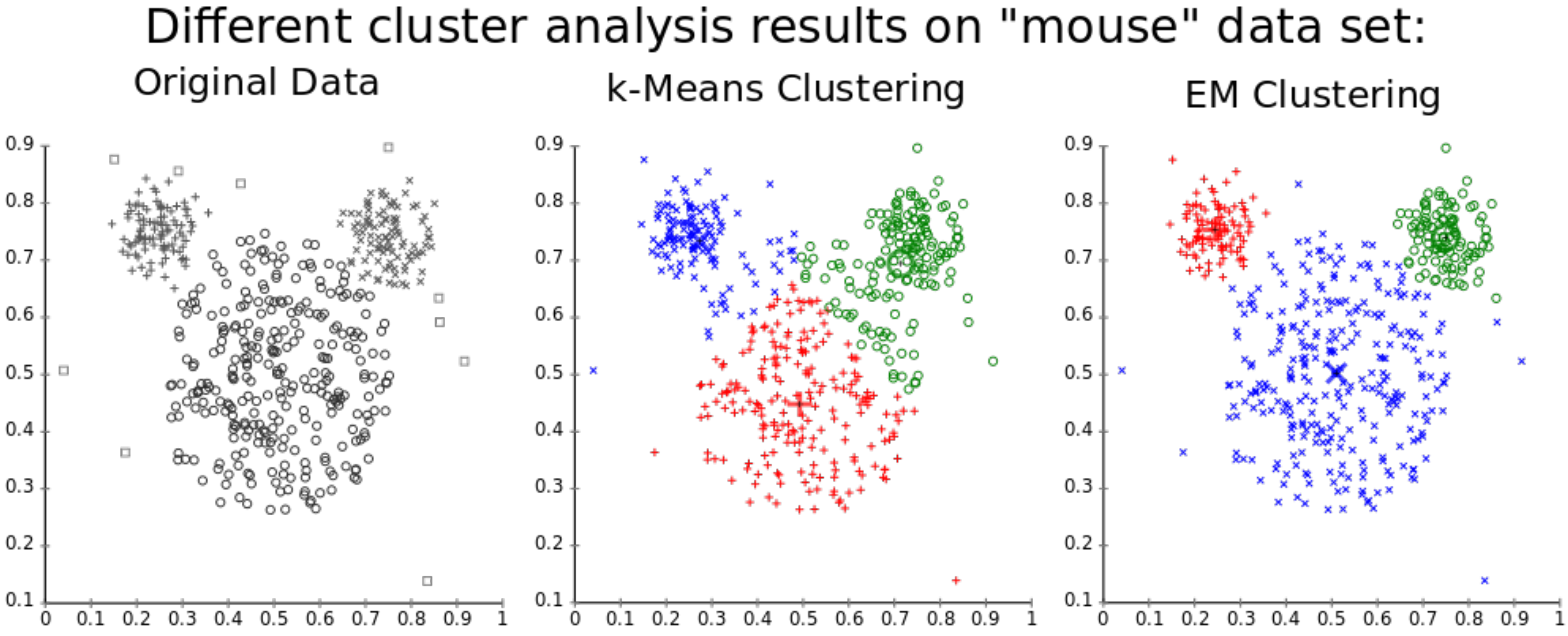
- Training Data includes desired output



Example: Spam email classifier

Unsupervised Learning

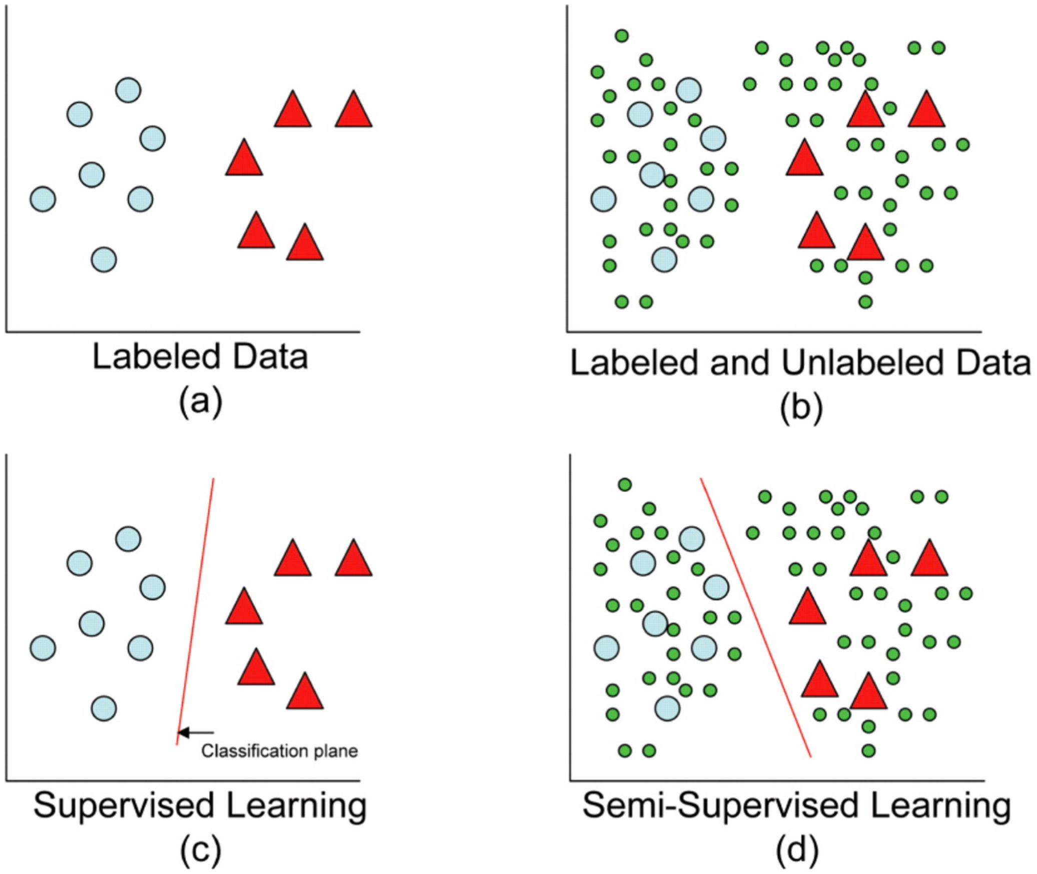
- Training Data does not include desired output



Example: Group users into cohorts based on habits

Semi-supervised Learning

- Training Data includes a few desired outputs



Labeled Data
(a)

Labeled and Unlabeled Data
(b)

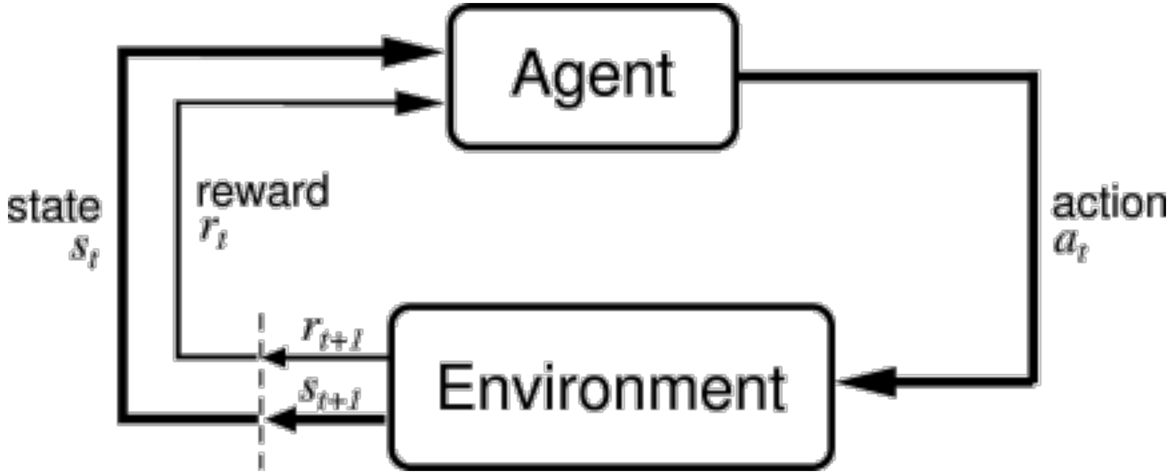
Supervised Learning
(c)

Semi-Supervised Learning
(d)

Example: What to do when you don’t having enough labels

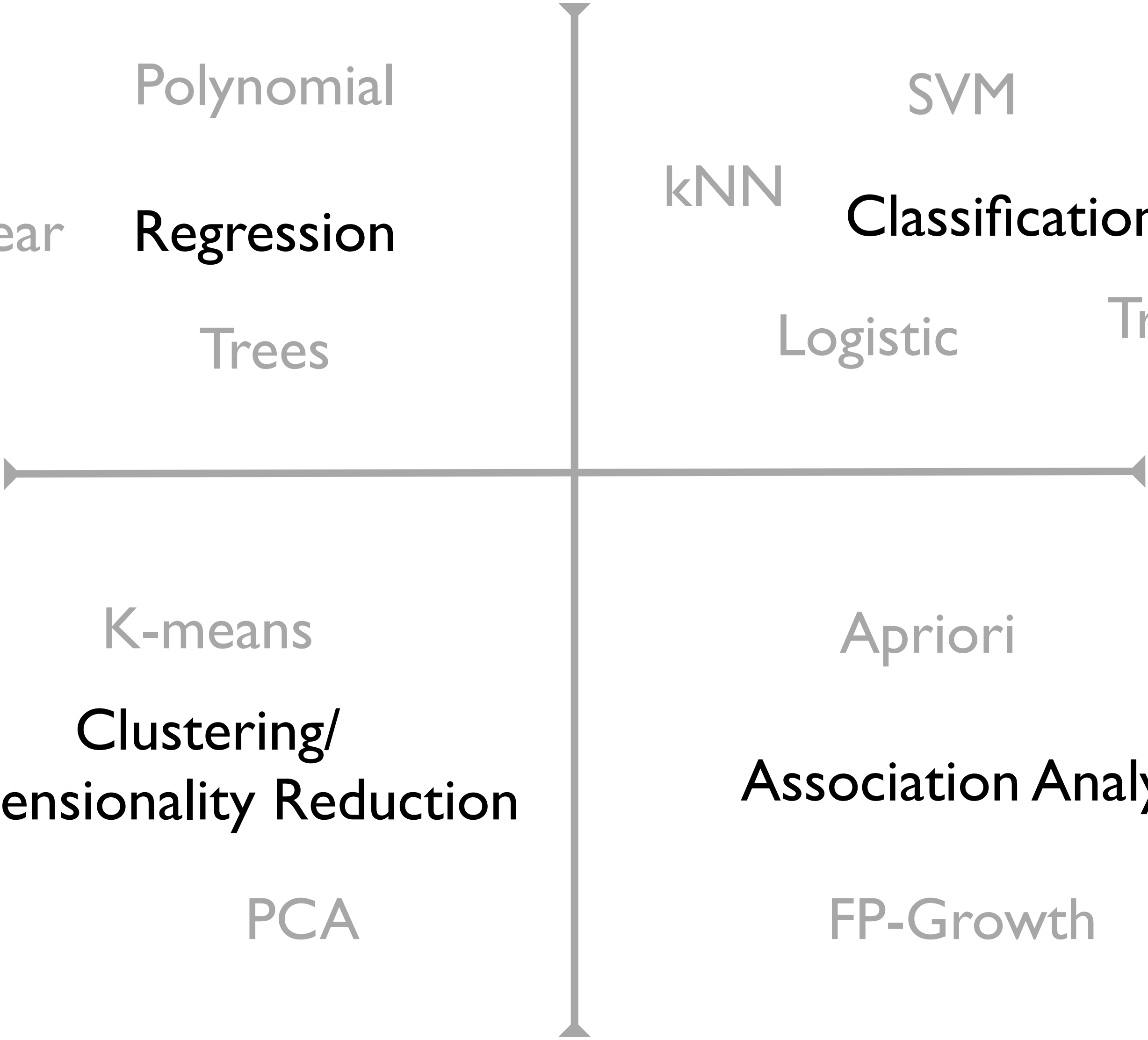
Reinforcement Learning (traditional AI)

- Rewards from sequence of actions



Example: Autonomous Video game Player

Labeled Data



Polynomial

Linear Regression

Trees

SVM

kNN Classification

Logistic

Trees

Continuous Target

Discrete Target

K-means

Clustering/
Dimensionality Reduction

SVD

PCA

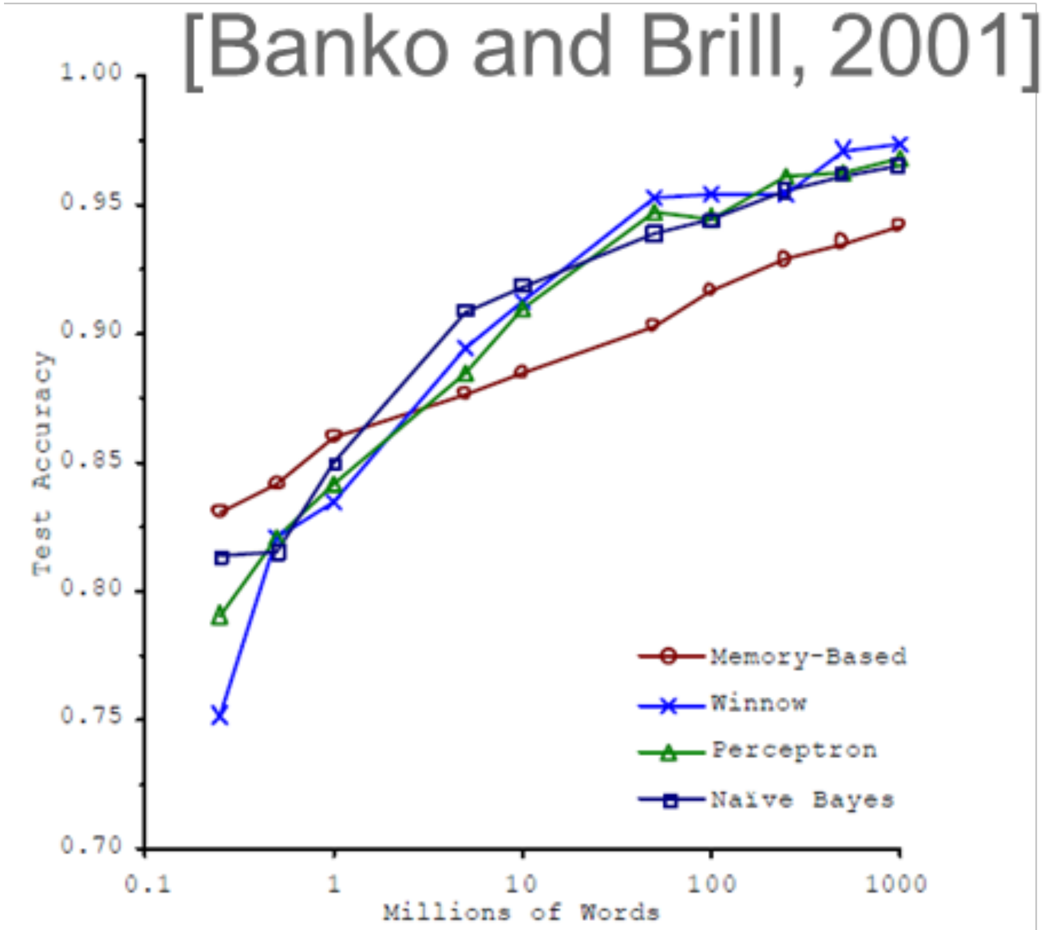
Apriori

Association Analysis

FP-Growth

Unlabeled Data

The Unreasonable Effectiveness of Data



[Banko and Brill, 2001]

Figure 1. Learning Curves for Confusion Set Disambiguation

Source: http://recsys.acm.org/more-data-or-better-models/

Quiz:

You are the Dean of a college and you need an automated way to send dissertations to the right departments…

Quiz: Which model will you use to avoid having to read all these theses?

- Red: Naive Bayes
- Blue: Linear Regression
- Green: Logistic Regression
- Yellow: Other (neural nets)

Quiz 2:

You are a realtor trying to find undervalued houses on the market…

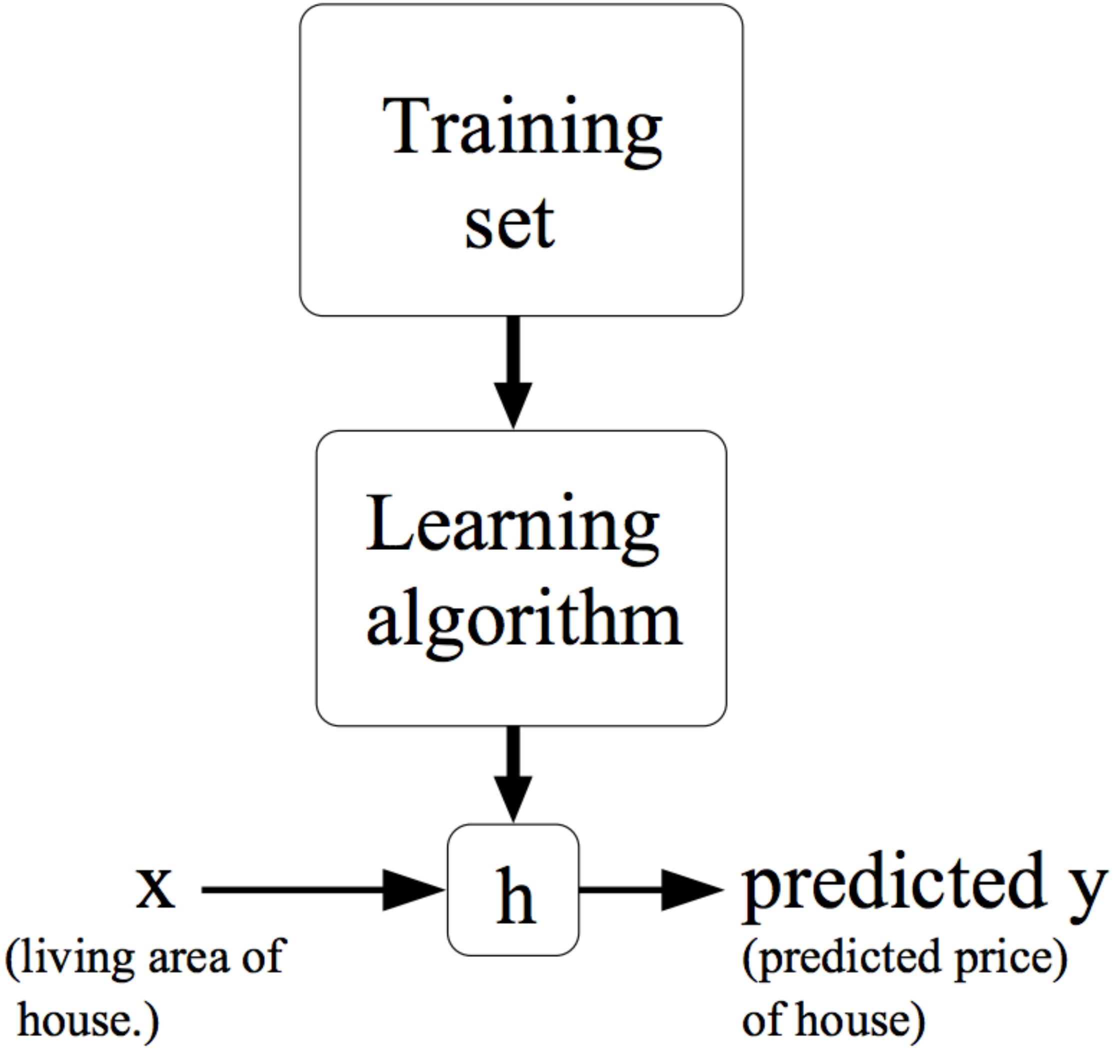
Quiz 2: Which model will you use to predict the market value of houses?

- Red: Naive Bayes
- Blue: Linear Regression
- Green: Logistic Regression
- Yellow: Other (neural nets)

INTRODUCTION TO SUPERVISED LEARNING

- Supervised Learning
  - 3 Components of a model
    - Hypothesis Function
    - Cost Function
    - Optimization Technique
  - Process
    - Labels vs. Features
    - Train
    - Test
    - Predict

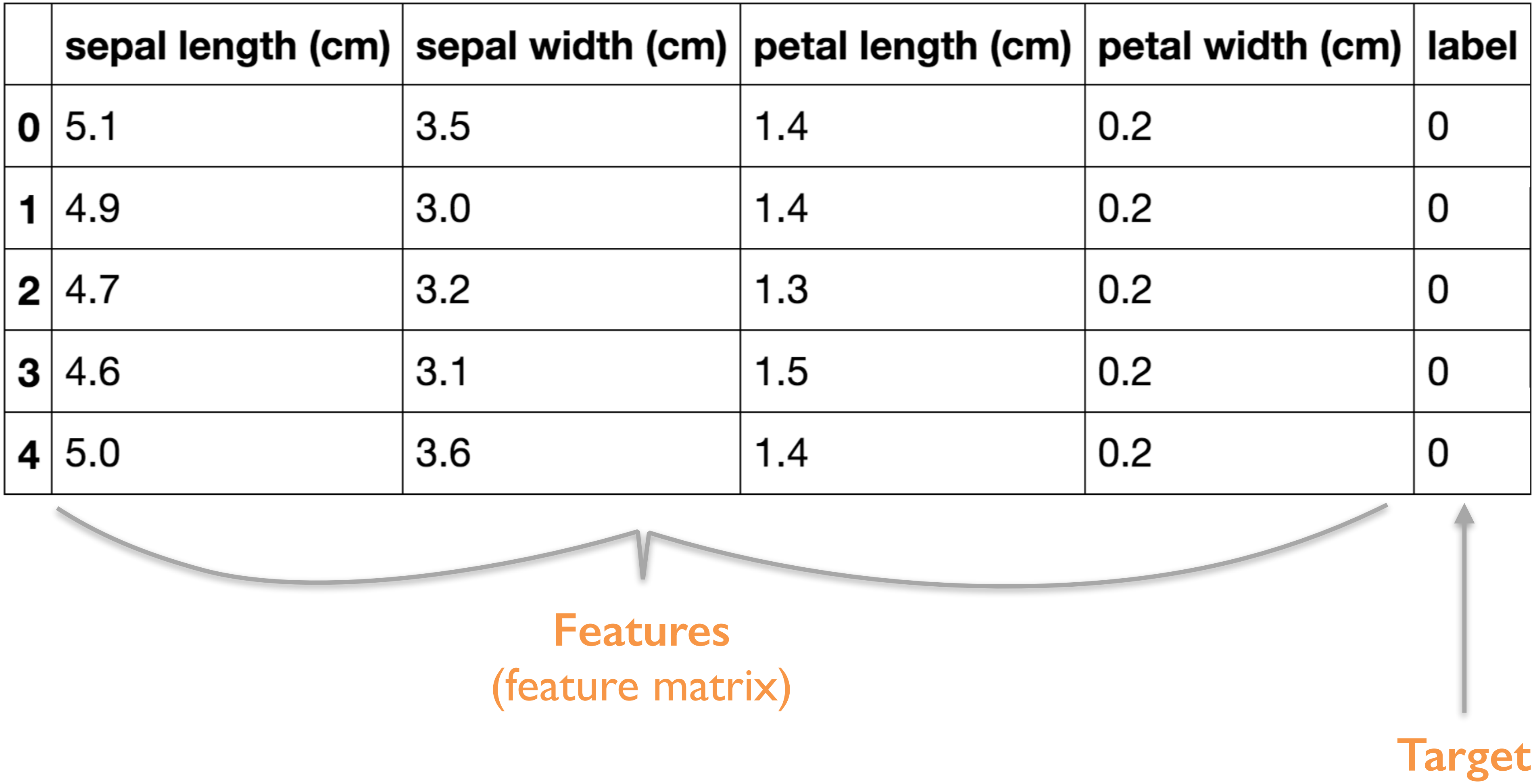
- Lab: Predicting Interest Rates



galvanize

source: Andrew Ng, CS229

Iris Dataset



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Features
(feature matrix)

Target

Train

Input: historical labeled data

+

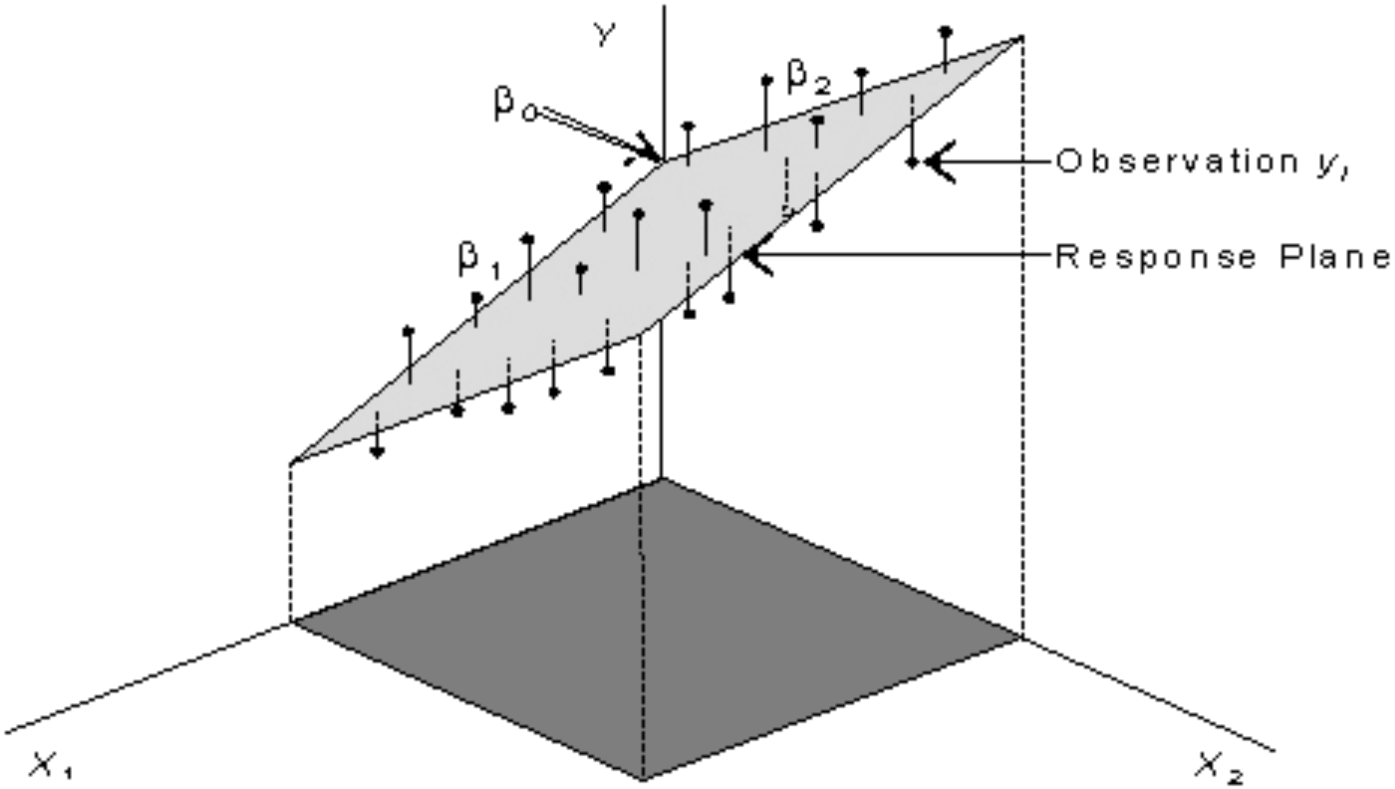
(hypothesis) function with unknown parameter values
<linear, logistic, etc.>

=

Predict

Output: parameter values

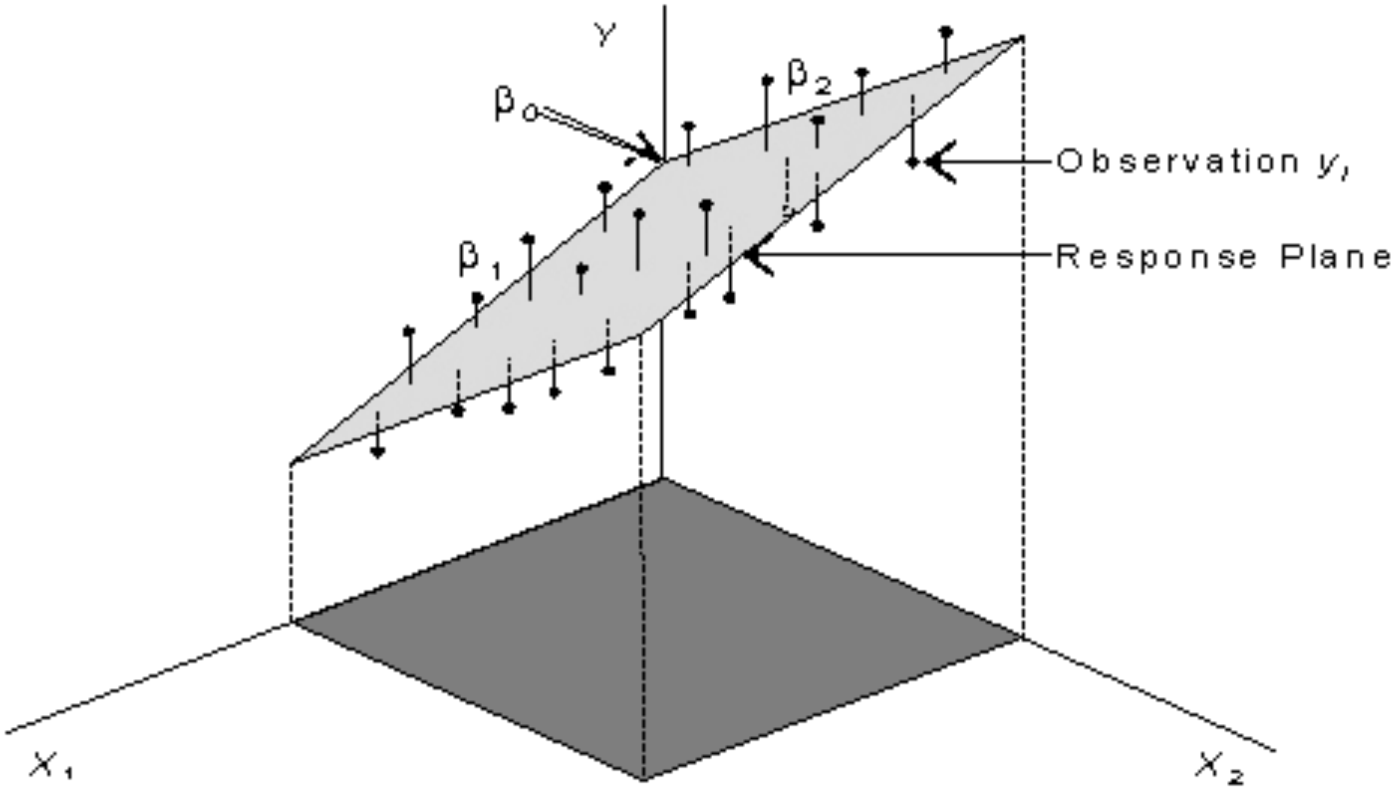
Multi Dimensional Regression



$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Hypothesis Function

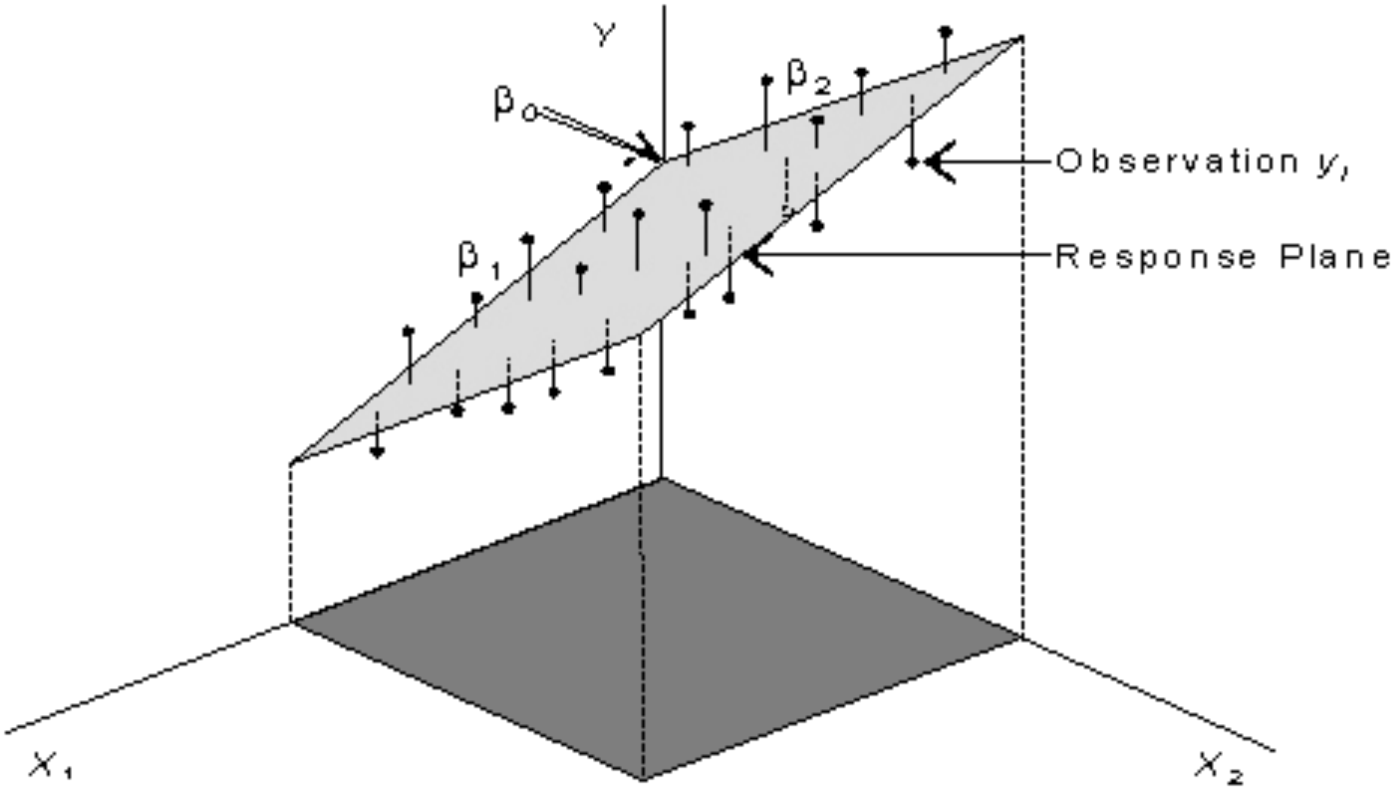
Multi Dimensional Regression



Parameters

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

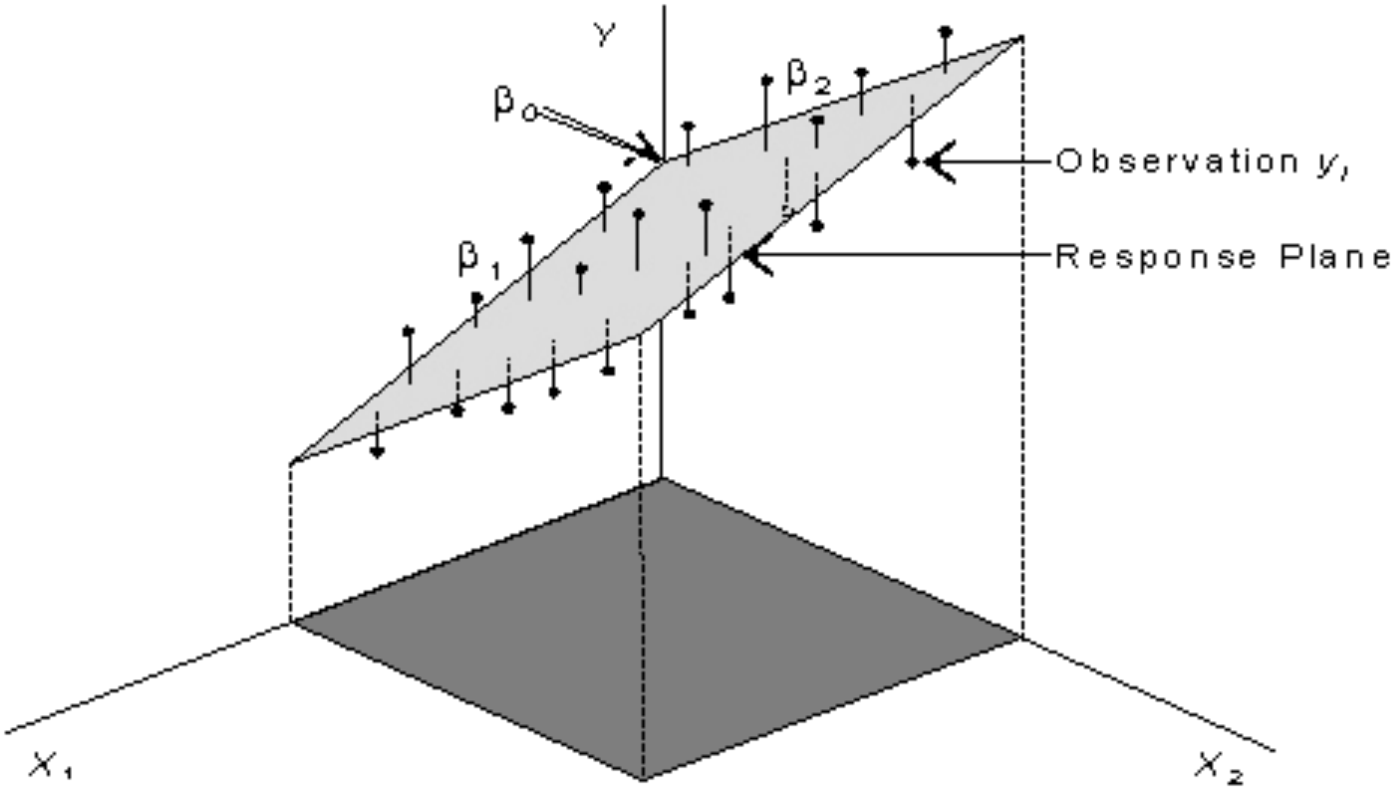
Multi Dimensional Regression



ML Gold

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Multi Dimensional Regression



Mine it!

$$\hat{y} = 0.32 + 1.45x_1 + 3.7x_2 + \ldots - 8.91x_n$$

What to learn an unknown target function f()

Input: labeled training set (xi, yi)

- yi = f(xi)

Output: hypothesis h() function “close” to f()

Many possible hypothesis families:

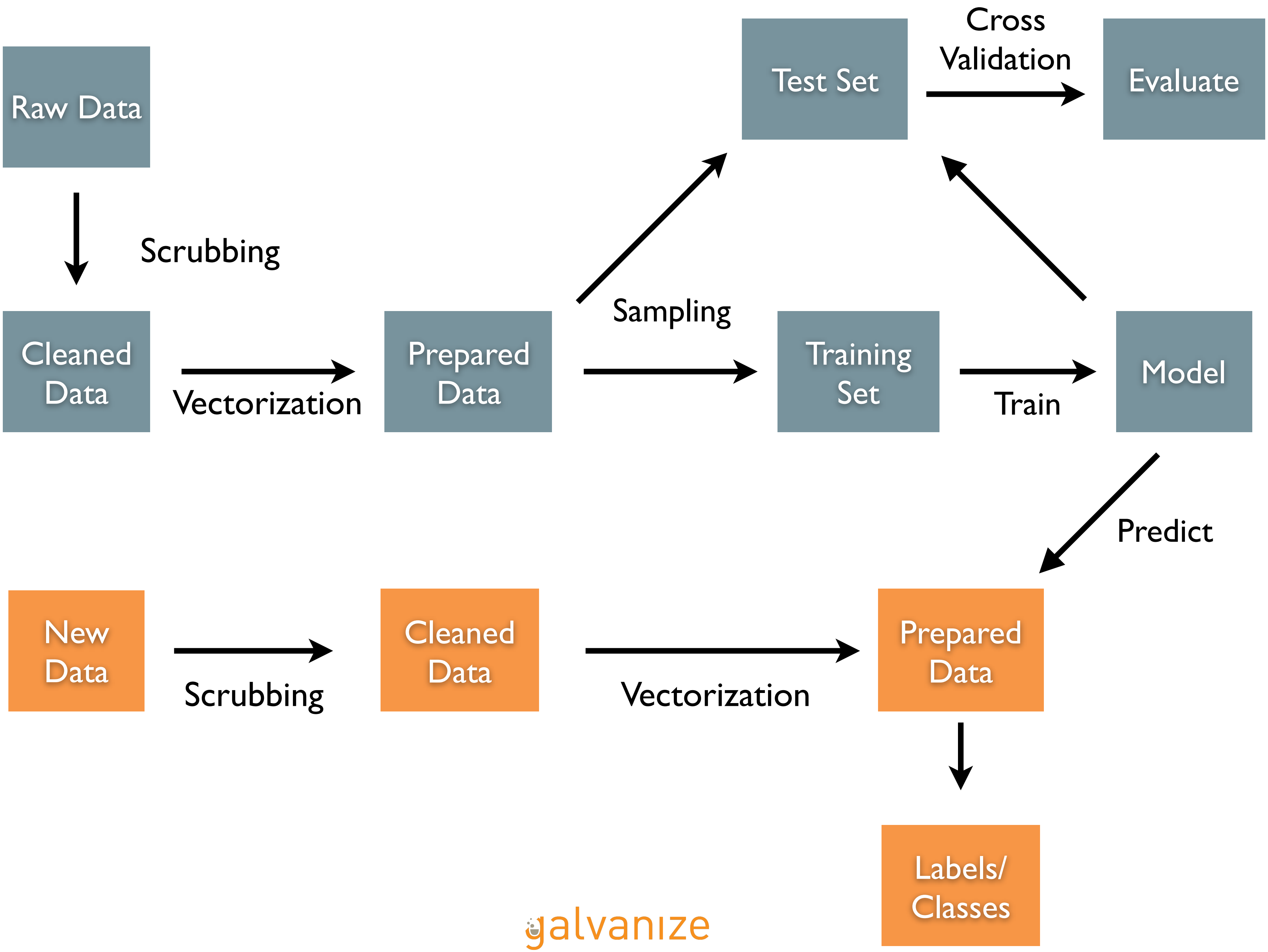
- Logistic
- Linear
- decision trees
- example-based (nearest neighbor)
- etc.

Throughout this class we will work to understand the answers the following questions:

- Which hypothesis space to choose?
- How do we measure goodness of fit?
- How do we balance goodness of fit with complexity?
- How do we make h() a good method?
- How do we pick the right kind of h()?
- How do we know if a good h() will predict well?

Throughout this class we will work to understand the answers the following questions:

- Which hypothesis space to choose?
  - Hypothesis Function
- How do we measure goodness of fit?
  - Cost Function
- How do we balance goodness of fit with complexity?
  - Regularization
- How do we make h() a good method?
  - Optimization
- How do we pick the right kind of h()?
  - Cross Validation
- How do we know if a good h() will predict well?
  - Hold-out Evaluation



Raw Data

Scrubbing

Cleaned Data

Vectorization

Prepared Data

Sampling

Test Set

Training Set

Cross Validation

Evaluate

Train

Model

Predict

New Data

Scrubbing

Cleaned Data

Vectorization

Prepared Data

Labels/ Classes

galvanize

LAB: PREDICTING LOAN RATES (WITH SCIKIT-LEARN)

galvanize