OVERVIEW MACHINE LEARNING

INTRODUCTION TO DATA SCIENCE

TIM KRASKA

teachingdatascience.org

TOKENIZATION AND STEMMING

WORKING WITH TEXT

teaching datascience .org

TOKENIZATION

Input: "*Friends, Romans and Countrymen*"

Output: Tokens

- *Friends*
- *Romans*
- *and*
- *Countrymen*

A token is an instance of a sequence of characters

COMMON STEPS

- **Remove Stop Words (a, an, the, to, be, …)**
- **Normalization to terms**
  - **deleting periods:** U.S.A. → USA
  - **deleting hyphens:** *anti-discriminatory → antidiscriminatory*
  - **Abbreviations**: Massachusetts Institute of Technology → MIT
  - **Case-folding:** Meal → meal, Brown → brown
  - **Language-issues:** *Tuebingen, Tübingen → Tubingen*
  - **asymmetric expansion***: windows → window*
  - *...*
  - *Why is this a form of entity resolution? What examples above are problematic?*
- **Thesauri and soundex**
  - ***car* = *automobile*** ***color* = *colour***
- ***Stemming***

STEMMING

Reduce terms to their "roots" before indexing

"Stemming" suggest crude affix chopping

- language dependent
- e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

PORTER'S ALGORITHM

Commonest algorithm for stemming English

- Results suggest it's at least as good as other stemming options

Conventions + 5 phases of reductions

- phases applied sequentially
- each phase consists of a set of commands
- sample convention: *Of the rules in a compound command, select the one that applies to the longest suffix.*

TYPICAL RULES IN PORTER

sses → ss

ies → i

ational → ate

tional → tion

Weight of word sensitive rules

(m>1) EMENT →

- *replacement → replac*
- *cement → cement*

OTHER STEMMERS

Other stemmers exist, e.g., Lovins stemmer

- http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm
- Single-pass, longest suffix removal (about 250 rules)

Full morphological analysis – at most modest benefits for retrieval

Do stemming and other normalizations help?

- English: very mixed results. Helps recall for some queries but harms precision on others
  - E.g., operative (dentistry) ⇒ oper
- Definitely useful for Spanish, German, Finnish, …
  - 30% performance gains for Finnish!

MACHINE LEARNING

INTRODUCTION TO DATA SCIENCE

TIM KRASKA

teaching datascience .org

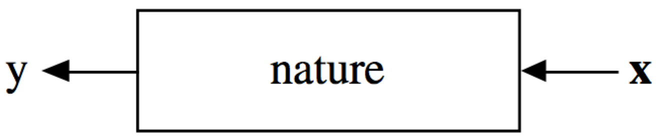
TITANIC DATASET

Label Features

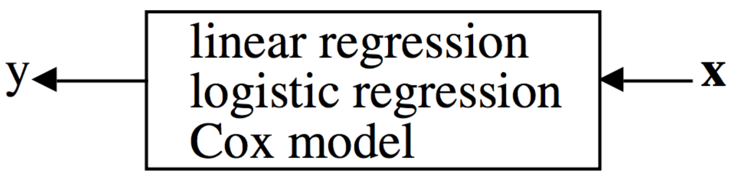
survived	pclass	sex	age	sibsp	parch	fare	cabin	embarked
0	3	male	22	1	0	7.25		S
1	1	female	38	1	0	71.2833	C85	C
1	3	female	26	0	0	7.925		S
1	1	female	35	1	0	53.1	C123	S
0	3	male	35	0	0	8.05		S
0	3	male		0	0	8.4583		Q
0	1	male	54	0	0	51.8625	E46	S
0	3	male	2	3	1	21.075		S
1	3	female	27	0	2	11.1333		S
1	2	female	14	1	0	30.0708		C
1	3	female	4	1	1	16.7	G6	S
1	1	female	58	0	0	26.55	C103	S
0	3	male	20	0	0	8.05		S

DIFFERENCE BETWEEN STATISTICS AND MACHINE LEARNING

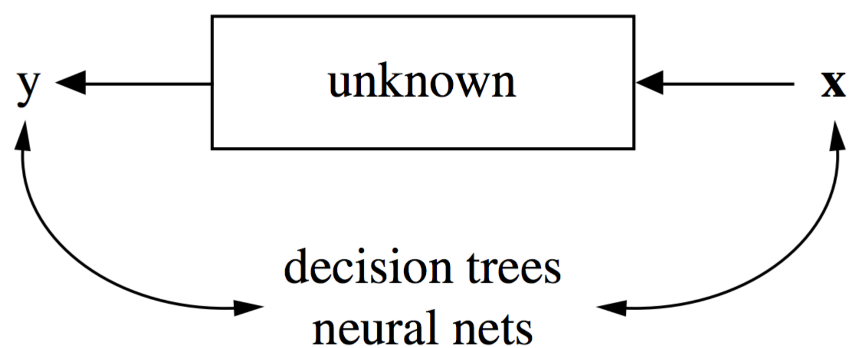
One view:



Emphasis on stochastic models of nature:

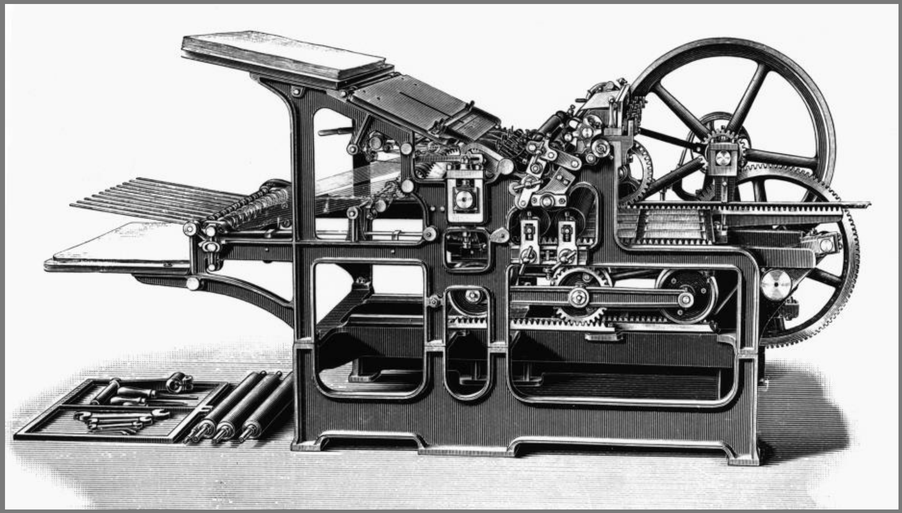


Find a function that predicts y from x: no model of nature implied or needed



Leo Breiman, Statistical Modeling: The Two Cultures, Statistical Science 16(3), 2001

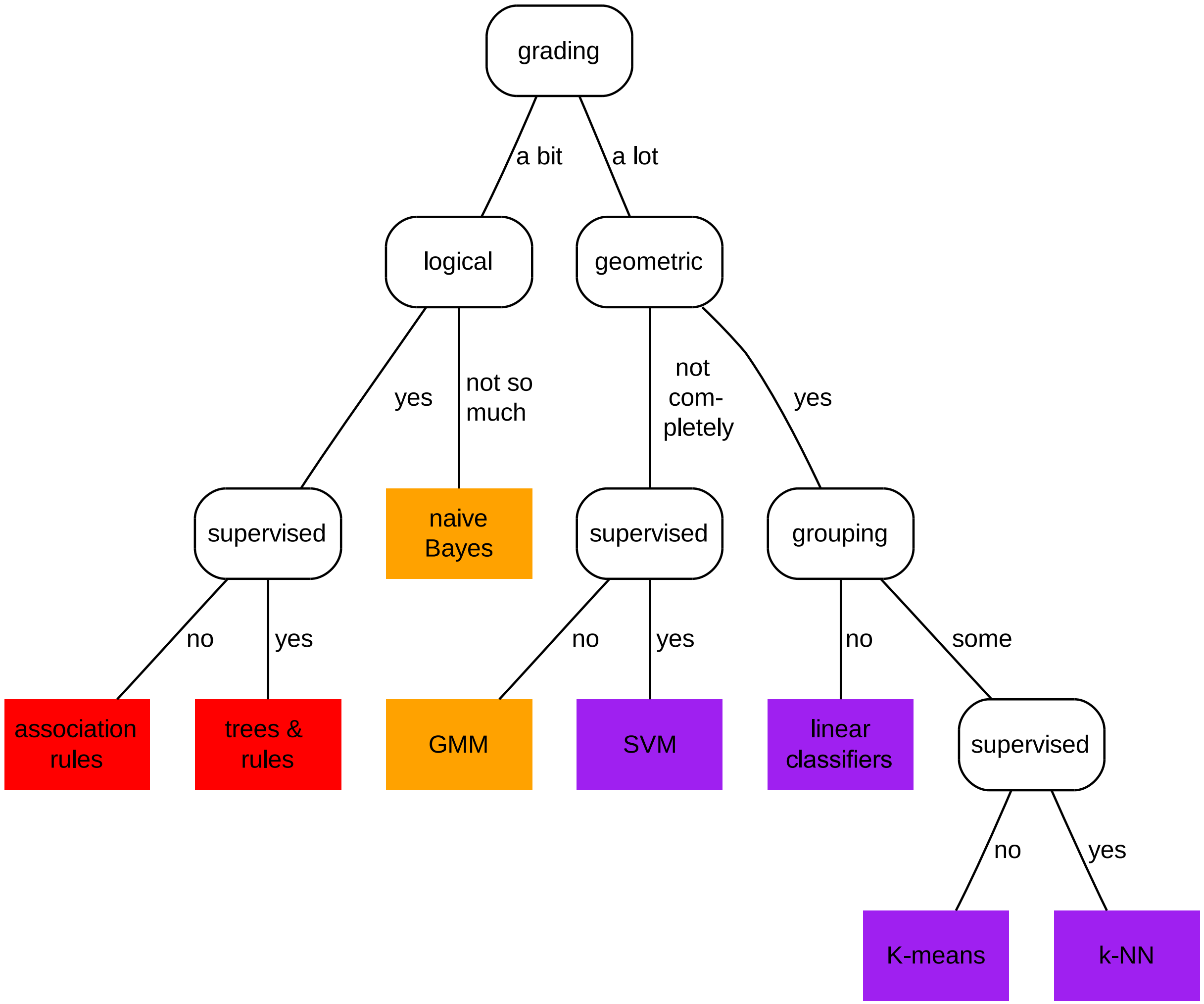

VS




MACHINE LEARNING PROBLEMS

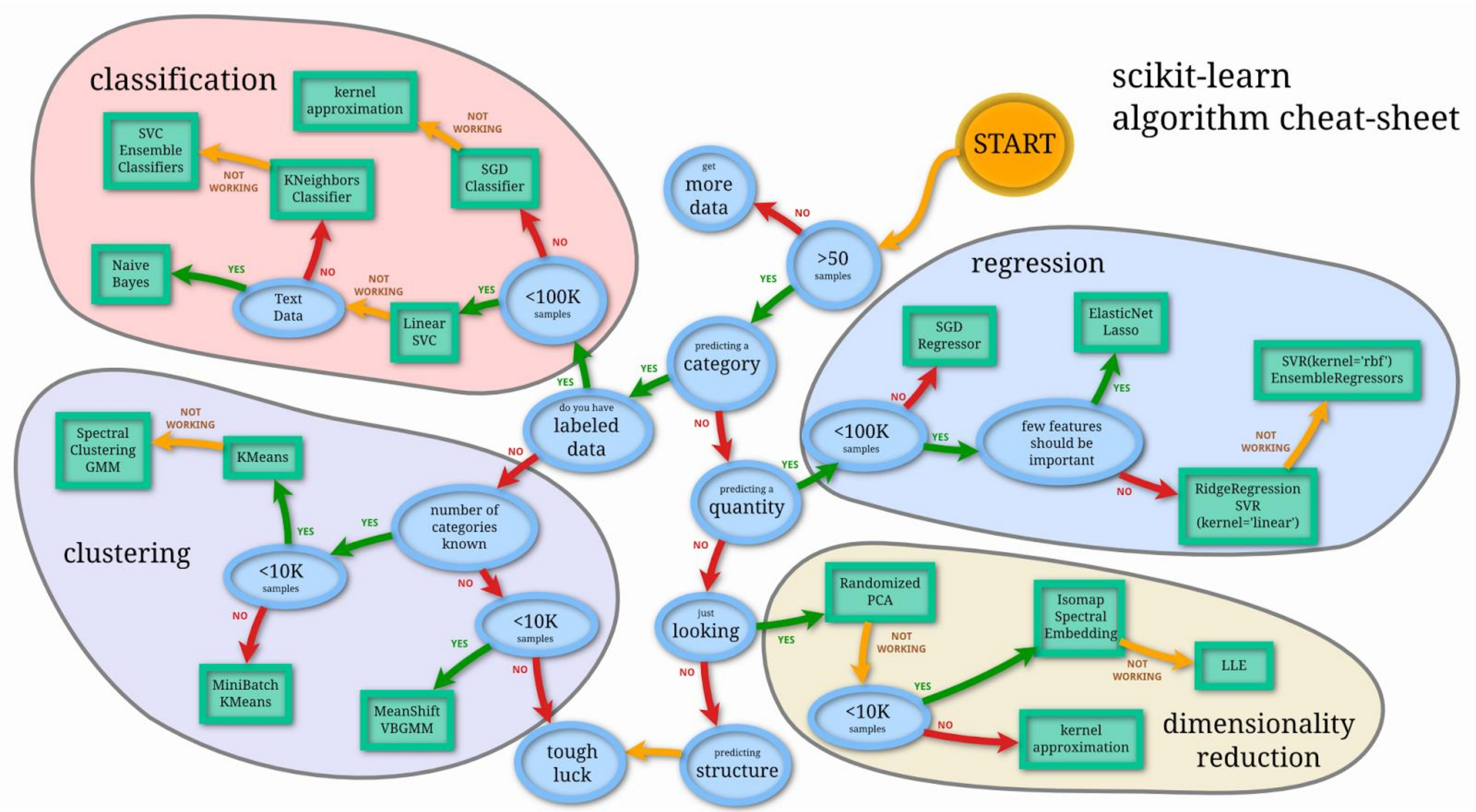
	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

ML on ML



[Peter Flach: Machine Learning]

(BE CAREFULLY)



MACHINE LEARNING PROBLEMS

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

CLUSTERING STRATEGIES

K-means

- Iteratively re-assign points to the nearest cluster center

Agglomerative clustering

- Start with each point as its own cluster and iteratively merge the closest clusters

Mean-shift clustering

- Estimate modes of PDF (i.e., the value x at which its probability mass function takes its maximum value)

Spectral clustering

- Split the nodes in a graph based on assigned links with similarity weights

As we go down this chart, the clustering strategies have more tendency to transitively group points even if they are not nearby in feature space

K-MEANS

Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, *c*:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|}\sum_{\vec{x} \in c} \vec{x}$$

Reassignment of instances to clusters is based on distance to the current cluster centroids.

- (Or one can equivalently phrase it in terms of similarities)

K-MEANS ALGORITHM

Select K random data points $\{s_1, s_2, \ldots s_K\}$ as centroids c_{j}.

Until clustering converges or other stopping criterion {

  For each data point x_i:

    Assign x_i to the closes centroid such that

      $dist(x_i, c_{\mathrm{j}})$ is minimal.

    For each cluster c_j, update the centroids

      $c_{\mathrm{j}} = \mu(c_j)$

}

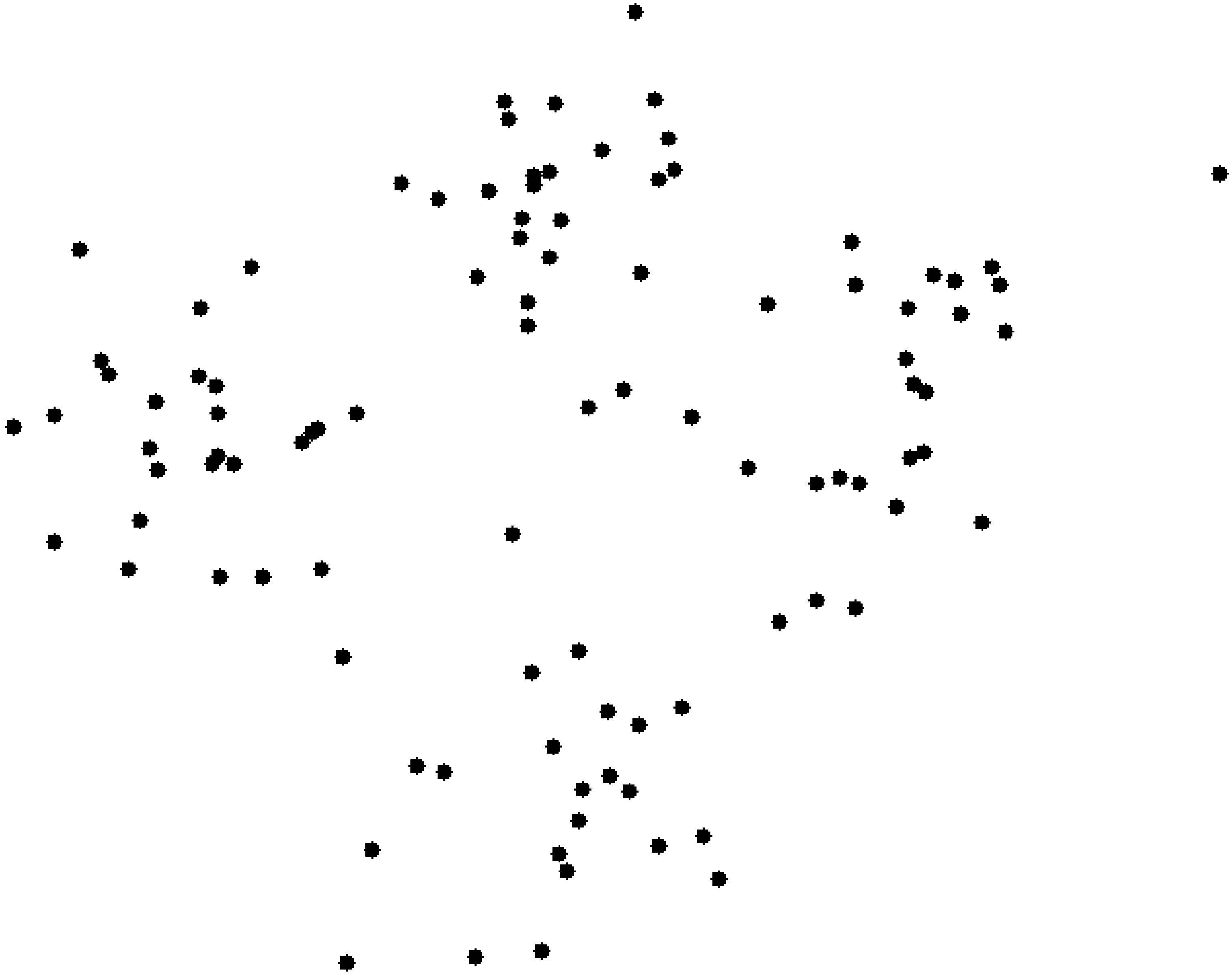
MORE FORMALLY

Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \le n$) $S = \{S_1, S_2, \ldots, S_k\}$:

$$\arg\min_{S} \sum_{i=1\ldots k} \sum_{x_j \in S_i} dist\left(x_j, c_i\right)$$

$$dist = \left(x_j - c_i\right)^2$$

$$c_i = \mu\left(S_i\right) = \frac{1}{|c|} \sum_{x_i \in S_i} x_i$$

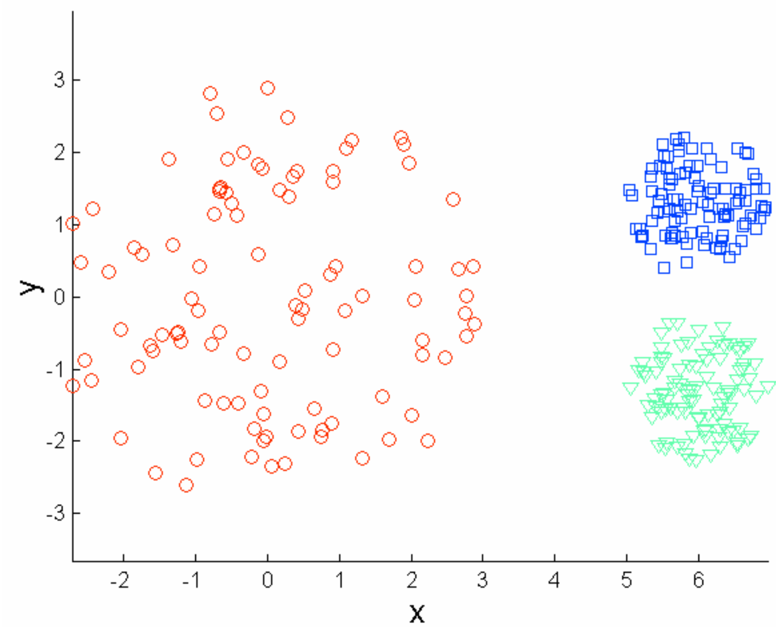


TERMINATION CONDITIONS

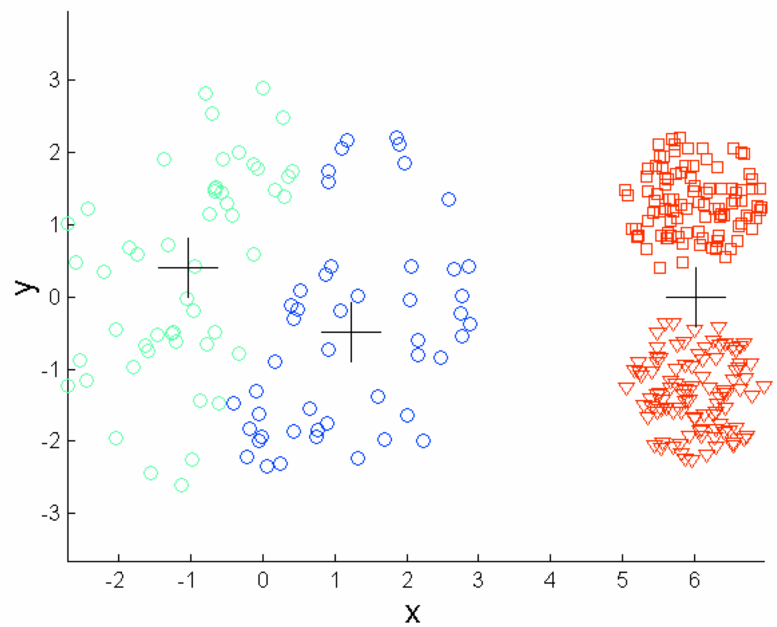
Several possibilities, e.g.,

- A fixed number of iterations.
- Partition unchanged.
- Centroid positions don’t change.

LIMITATIONS OF K-MEANS: DIFFERING DENSITY



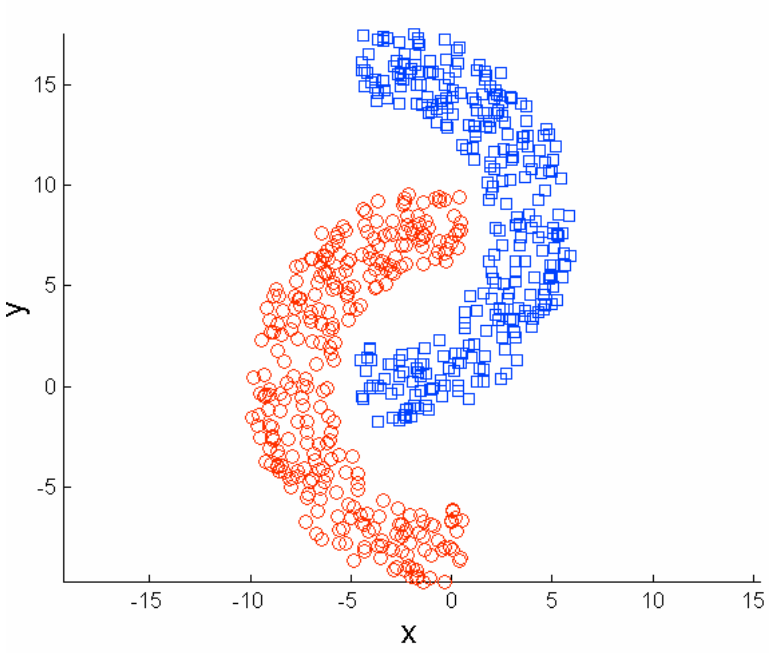
Original Points



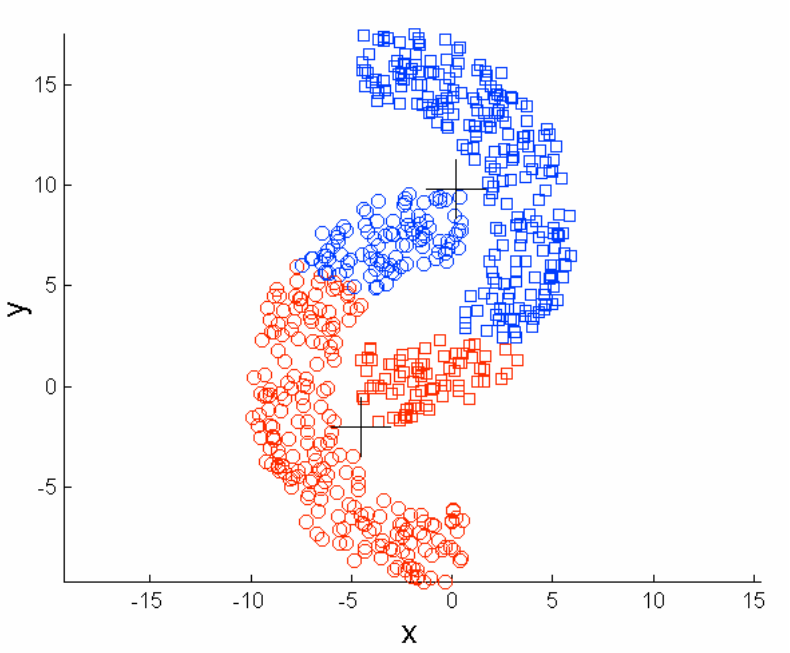
K-means (3 Clusters)

LIMITATIONS OF K-MEANS

Non-globular Shapes

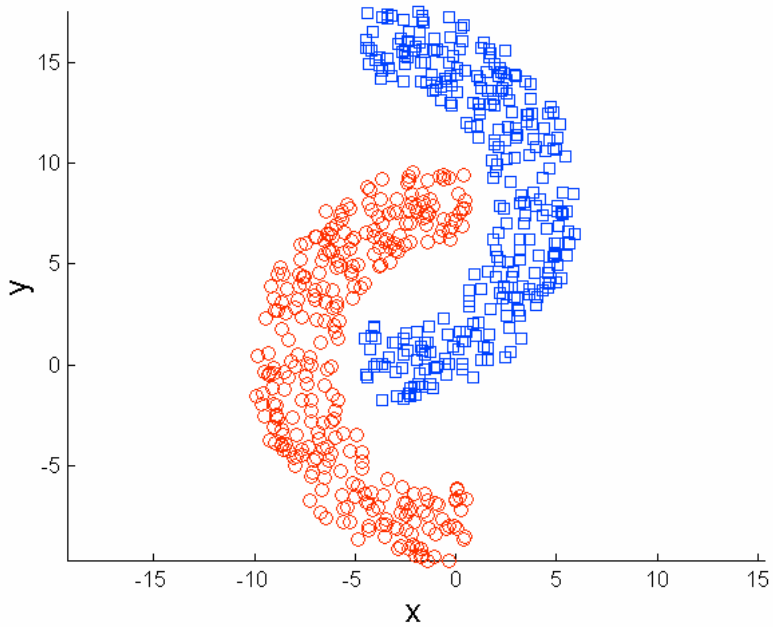


Original Points

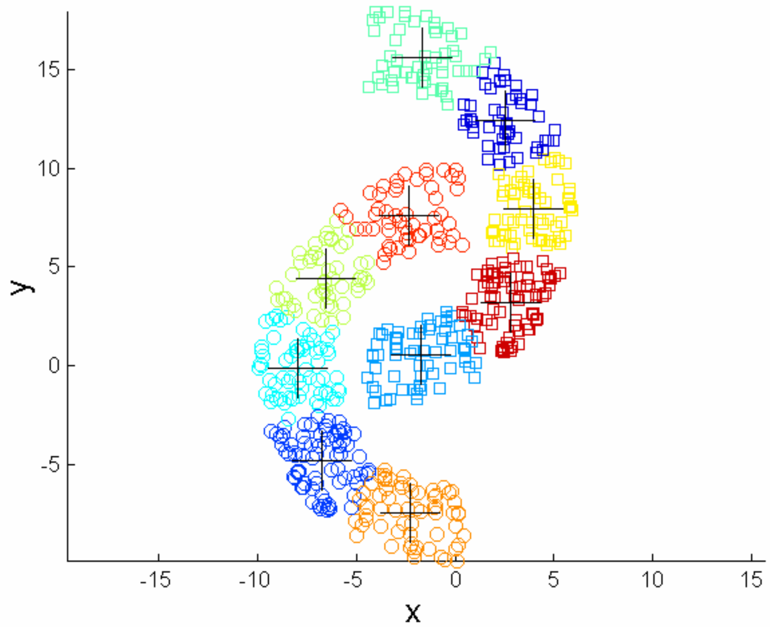


K-means (2 Clusters)

OVERCOMING K-MEANS LIMITATIONS



Original Points



K-means Clusters

Can you think of other ways to overcome the limitations?

CLUSTERING STRATEGIES

K-means

- Iteratively re-assign points to the nearest cluster center

Agglomerative clustering

- Start with each point as its own cluster and iteratively merge the closest clusters

Mean-shift clustering

- Estimate modes of PDF (i.e., the value x at which its probability mass function takes its maximum value)

Spectral clustering

- Split the nodes in a graph based on assigned links with similarity weights

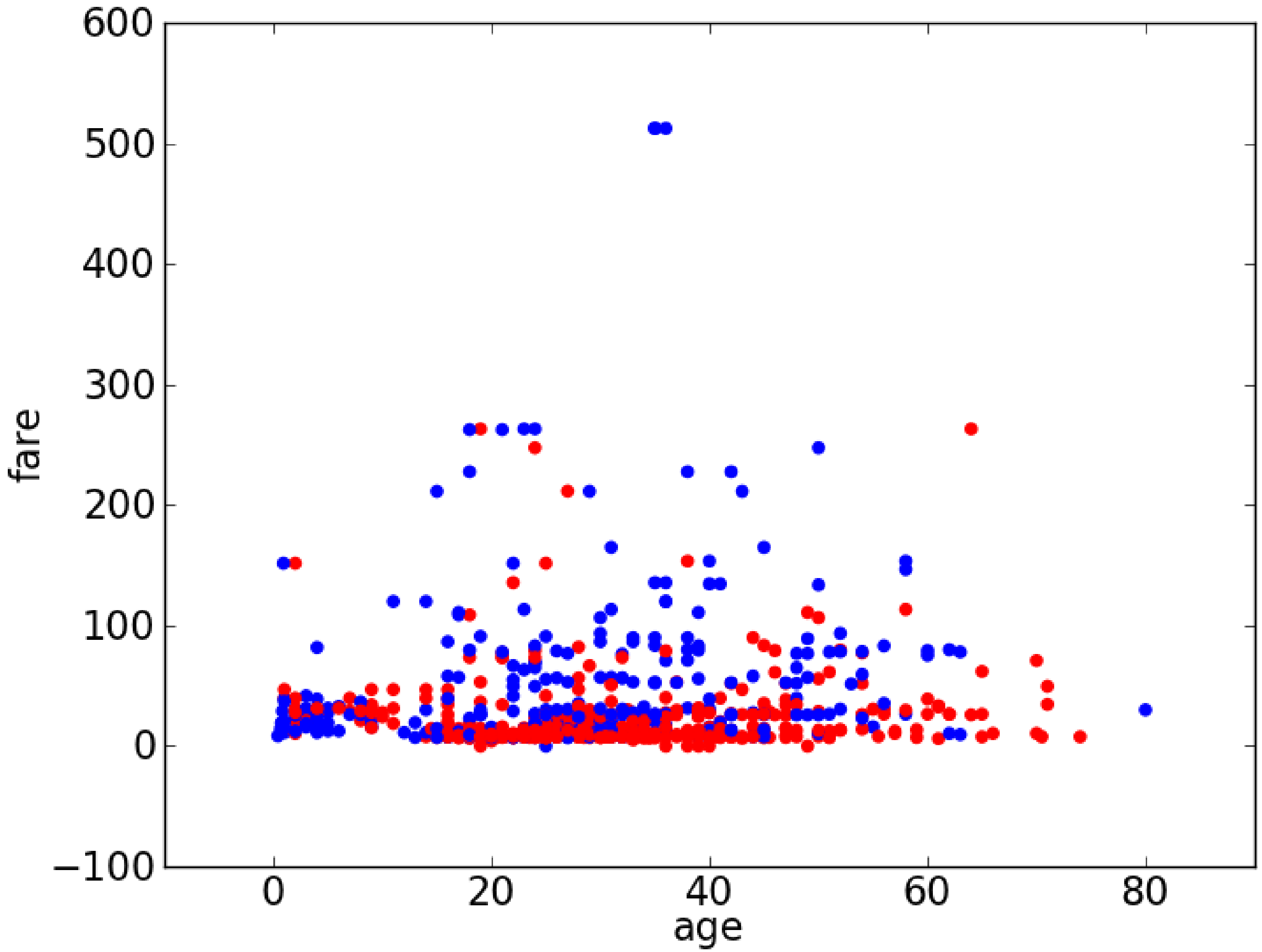
As we go down this chart, the clustering strategies have more tendency to transitively group points even if they are not nearby in feature space

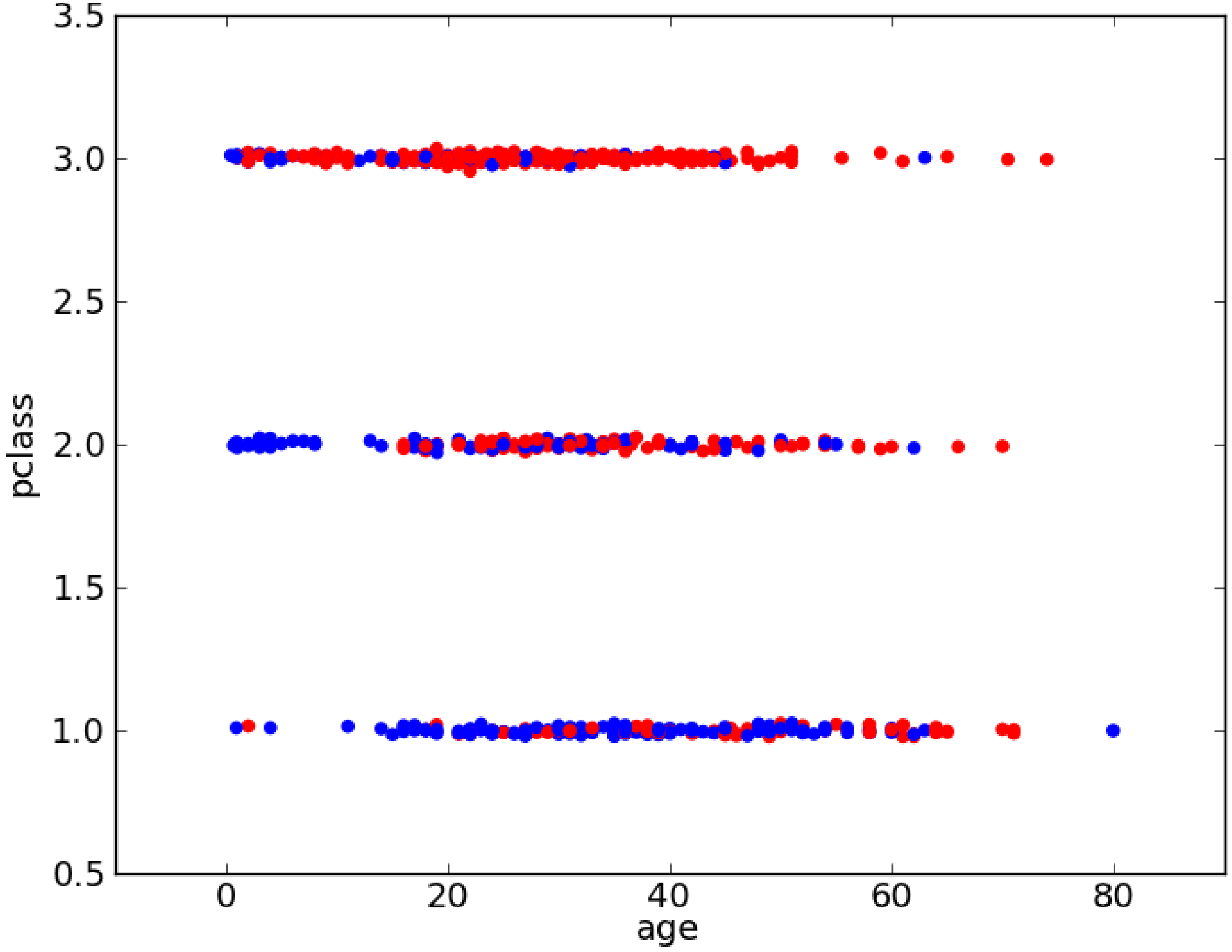
MACHINE LEARNING PROBLEMS

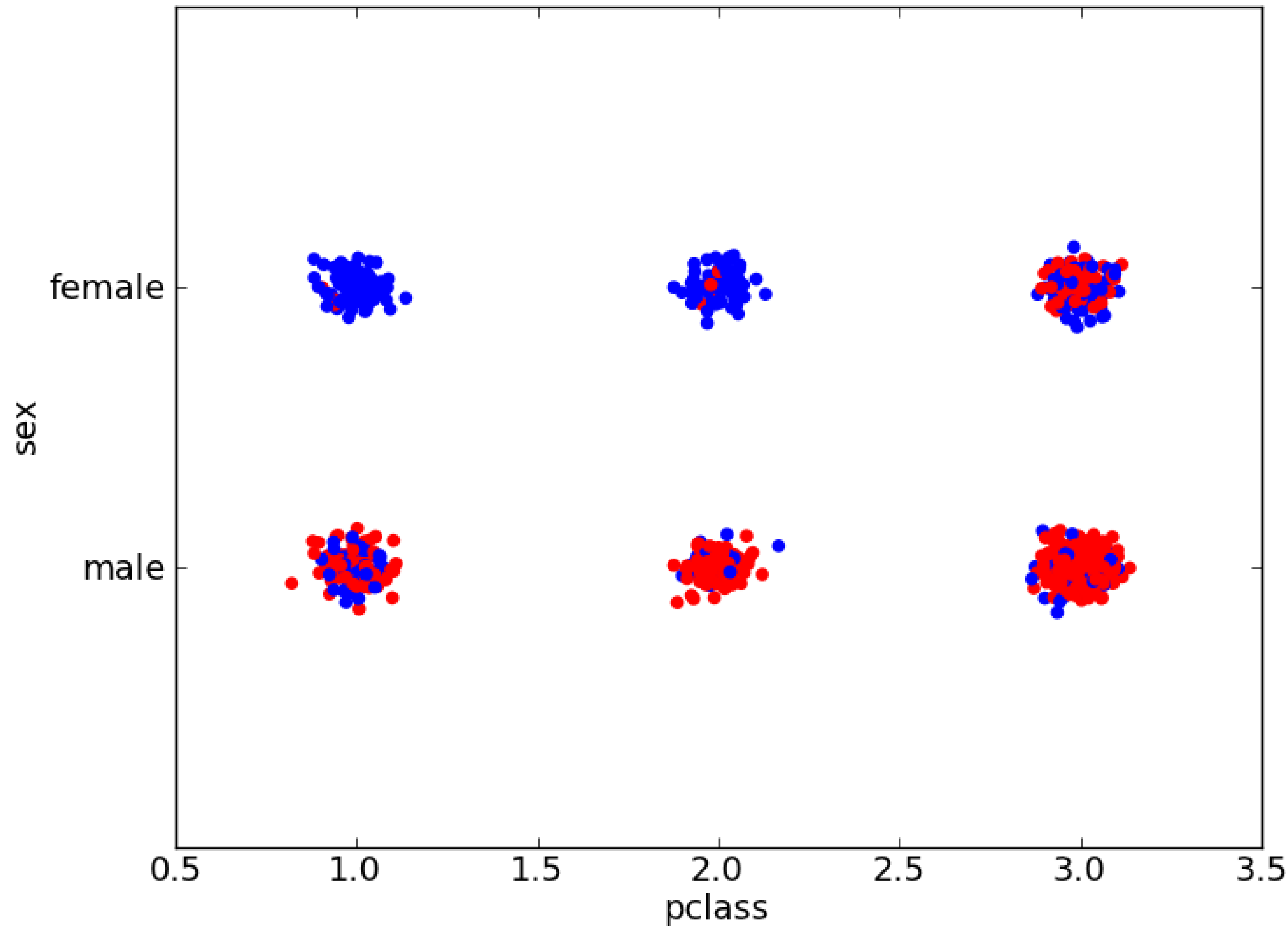
	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

TITANIC DATASET

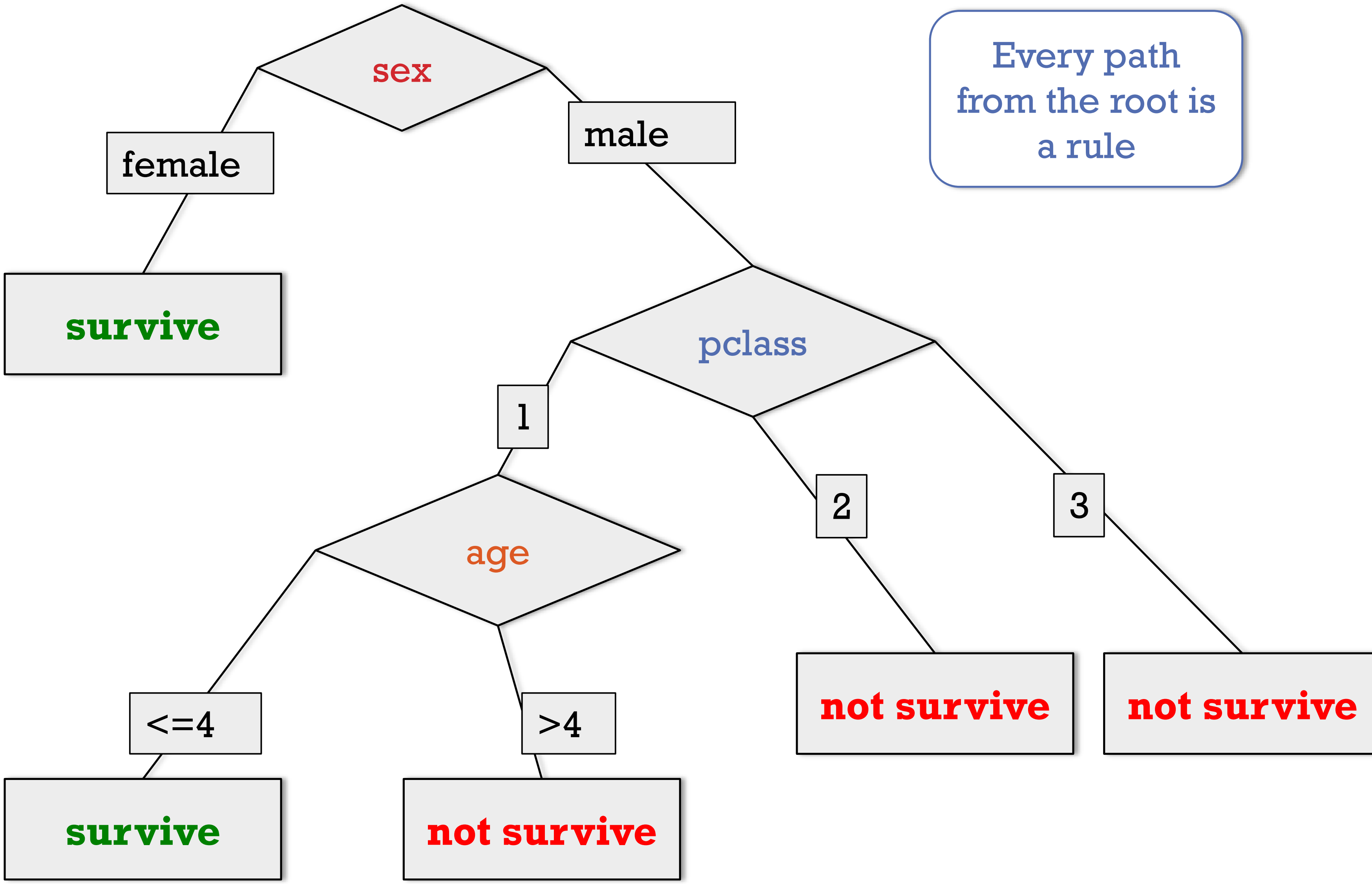
survived	pclass	sex	age	sibsp	parch	fare	cabin	embarked
0	3	male	22	1	0	7.25		S
1	1	female	38	1	0	71.2833	C85	C
1	3	female	26	0	0	7.925		S
1	1	female	35	1	0	53.1	C123	S
0	3	male	35	0	0	8.05		S
0	3	male		0	0	8.4583		Q
0	1	male	54	0	0	51.8625	E46	S
0	3	male	2	3	1	21.075		S
1	3	female	27	0	2	11.1333		S
1	2	female	14	1	0	30.0708		C
1	3	female	4	1	1	16.7	G6	S
1	1	female	58	0	0	26.55	C103	S
0	3	male	20	0	0	8.05		S


600
500
400
300
200
100
0
−100
fare
0
20
40
60
80
age


3.5
3.0
2.5
2.0
1.5
1.0
0.5
pclass
0
20
40
60
80
age


female
male
sex
0.5
1.0
1.5
2.0
2.5
3.0
3.5
pclass

DECISION TREE



Every path from the root is a rule

WHAT IS A CLASSIFIER

Apply a prediction function to a feature representation of an image/data-set to get the desired output:

f() = “apple”

f() = “tomato”

f() = “cow”

Slide credit: L. Lazebnik

THE MACHINE LEARNING FRAMEWORK

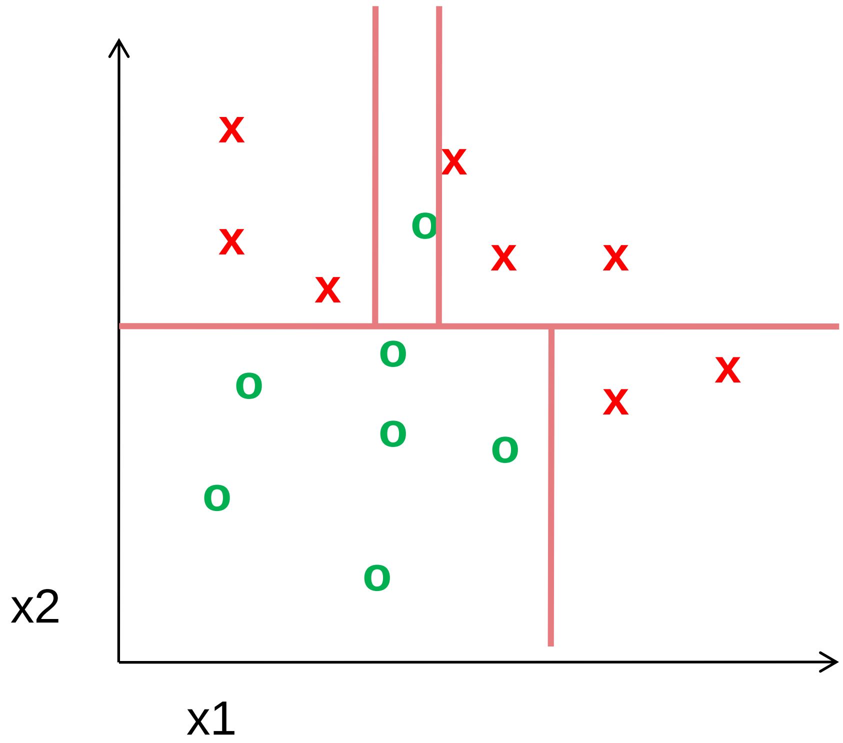
$$\underset{\text{output}}{\mathbf{y}} = \underset{\text{prediction function}}{\mathbf{f}}(\underset{\text{features}}{\mathbf{x}})$$

Training: given a *training set* of labeled examples $\{(x_1,y_1), \ldots, (x_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

Testing: apply f to a never before seen *test example* x and output the predicted value $y = f(x)$

Slide credit: L. Lazebnik

DECISION BOUNDARIES: DECISION TREES



CLASSIFIER OVERVIEW

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

RECAP: DECISION TREES

Representation

- A set of rules: IF…THEN conditions

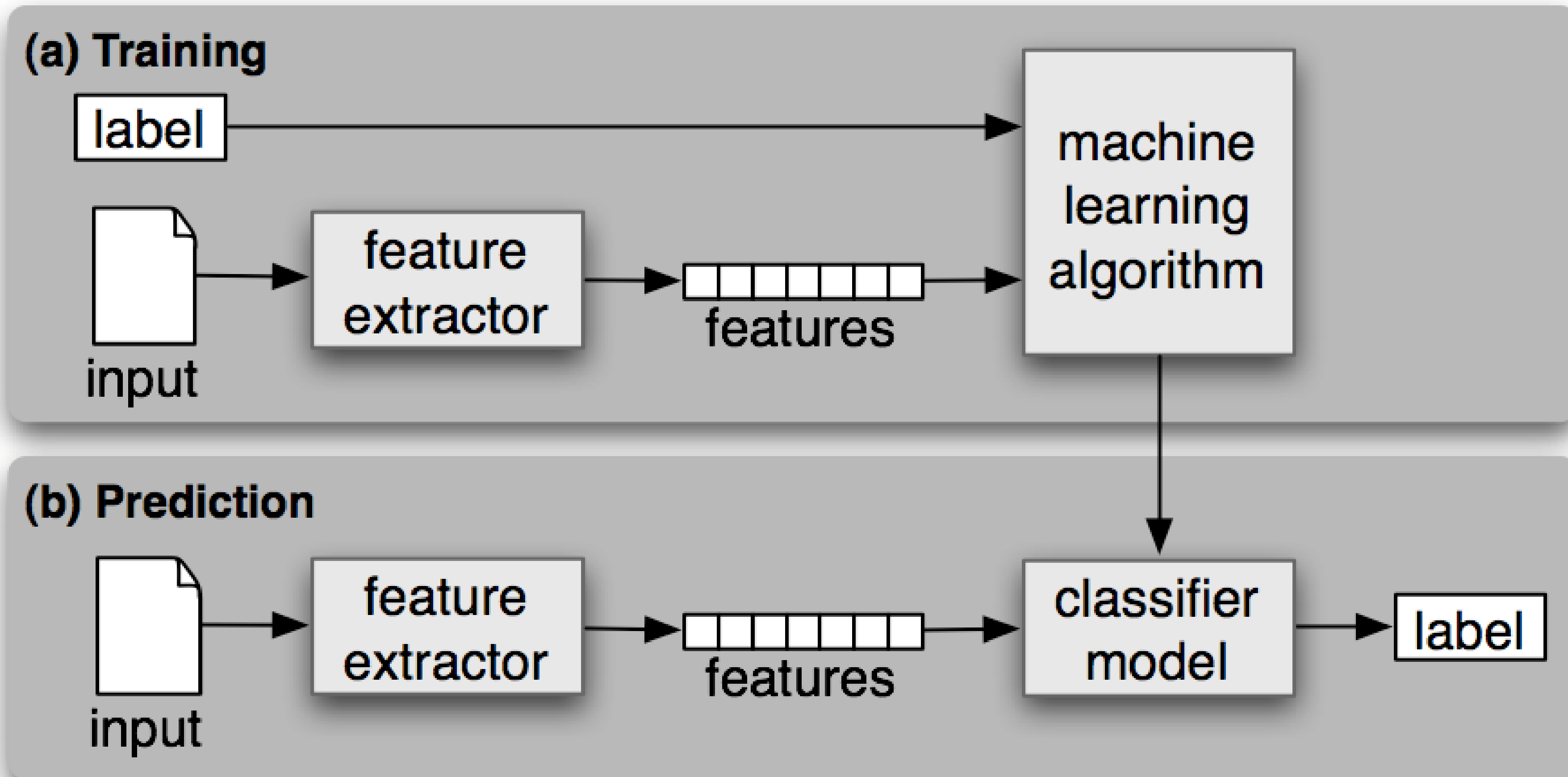
Evaluation

- coverage: # of data points that satisfy conditions
- accuracy = # of correct predictions / coverage

Optimization

- Build decision tree that maximize accuracy

ML PIPELINE (SUPERVISED)



FEATURES

Fact Table

- **Shop_ID**
- **Customer_ID**
- **Date_ID**
- **Product_ID**
- Amount
- Volume
- Profit
- …

Fact Table

- **Shop_ID**
- **Customer_ID**
- **Date_ID**
- **Product_ID**
- Amount
- Volume
- Profit
- Delivery Time
- …

Product

- **Product_ID**
- Type_ID
- Brand_ID
- Length
- Height
- Depth
- Weight
- …

Product_Type

- **Type_ID**
- Name
- Description
- …

Brand

- **Brand_ID**
- Name
- …

Custermer State	Product Type	Product Weight	Volume (L*H*D)	Month	Delivery Time

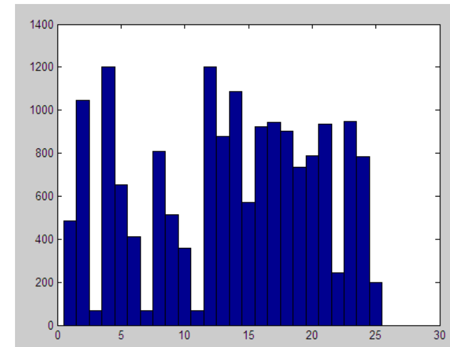
IMAGE FEATURES

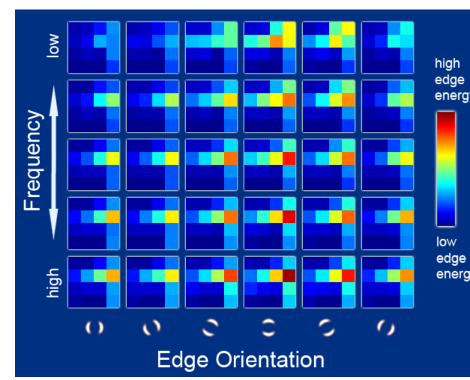
Raw pixels

Histograms

GIST descriptors

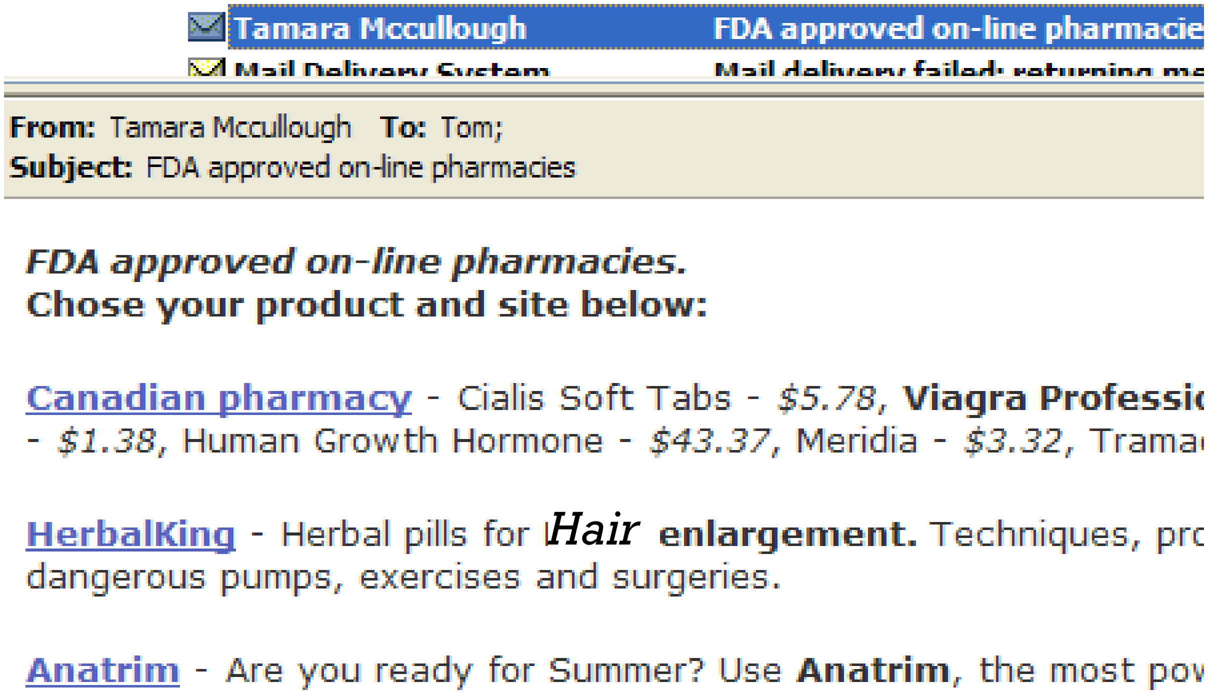
...





Slide credit: L. Lazebnik

TEXT FEATURES


Tamara Mccullough FDA approved on-line pharmacie
Mail Delivery System Mail delivery failed: returning me

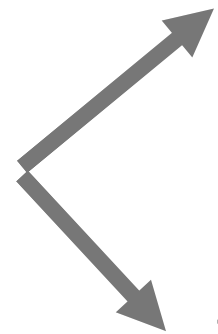
From: Tamara Mccullough **To:** Tom;
Subject: FDA approved on-line pharmacies

FDA approved on-line pharmacies.
Chose your product and site below:

Canadian pharmacy - Cialis Soft Tabs - *$5.78*, **Viagra Professi**
- *$1.38*, Human Growth Hormone - *$43.37*, Meridia - *$3.32*, Trama

HerbalKing - Herbal pills for *Hair* **enlargement.** Techniques, pr
dangerous pumps, exercises and surgeries.

Anatrim - Are you ready for Summer? Use **Anatrim**, the most pow



Spam

Not Spam

Bag of Words

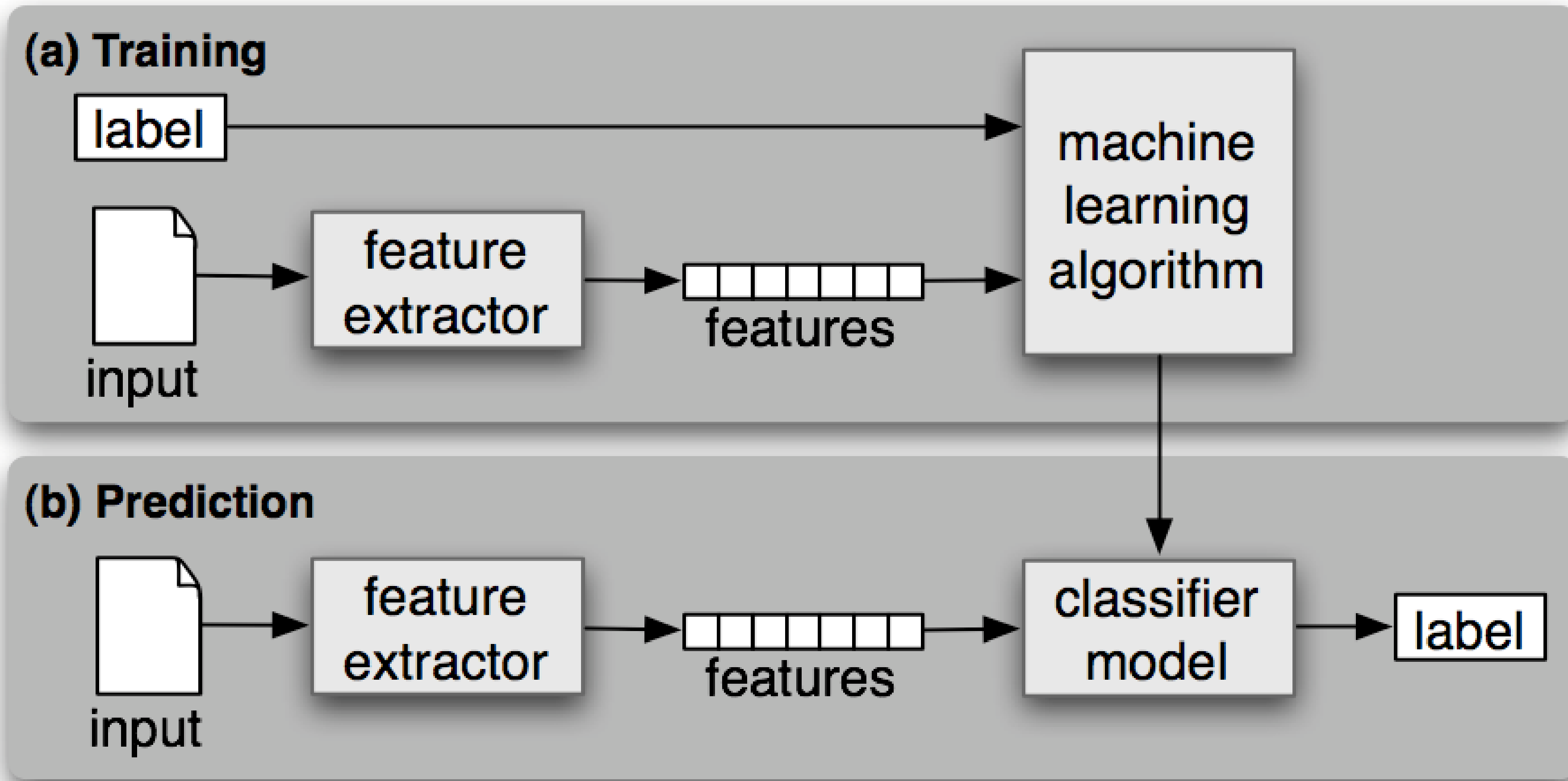
$$\begin{pmatrix} Viagra \\ Soft \\ Herbel \\ Pills \\ Are \\ \ldots \end{pmatrix}$$

N-Grams

$$\begin{pmatrix} herbel\ pills \\ pills\ for \\ for\ Hair \\ Hair\ enlargement \\ enlargement\ Techniques \\ \ldots \end{pmatrix}$$

FEATURE TO PREDICT UNEMPLOYMENT

ML PIPELINE (SUPERVISED)



MANY CLASSIFIERS TO CHOOSE FROM

Decision Trees

K-nearest neighbor

Support Vector Machines

Logistic Regression

Naïve Bayes

Random Forrest

Bayesian network

Randomized Forests

Boosted Decision Trees

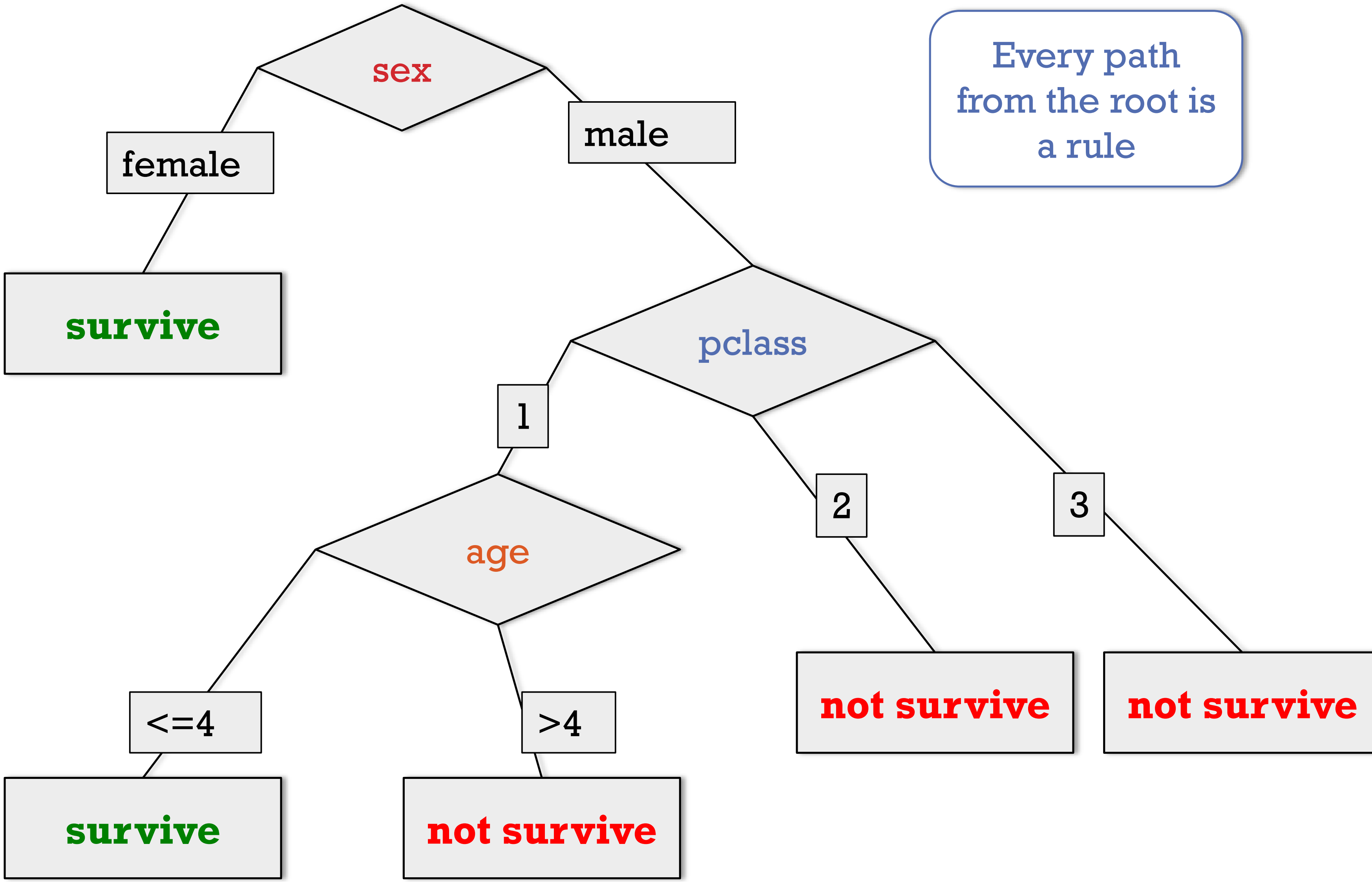
RBMs

….

Which is the best one?

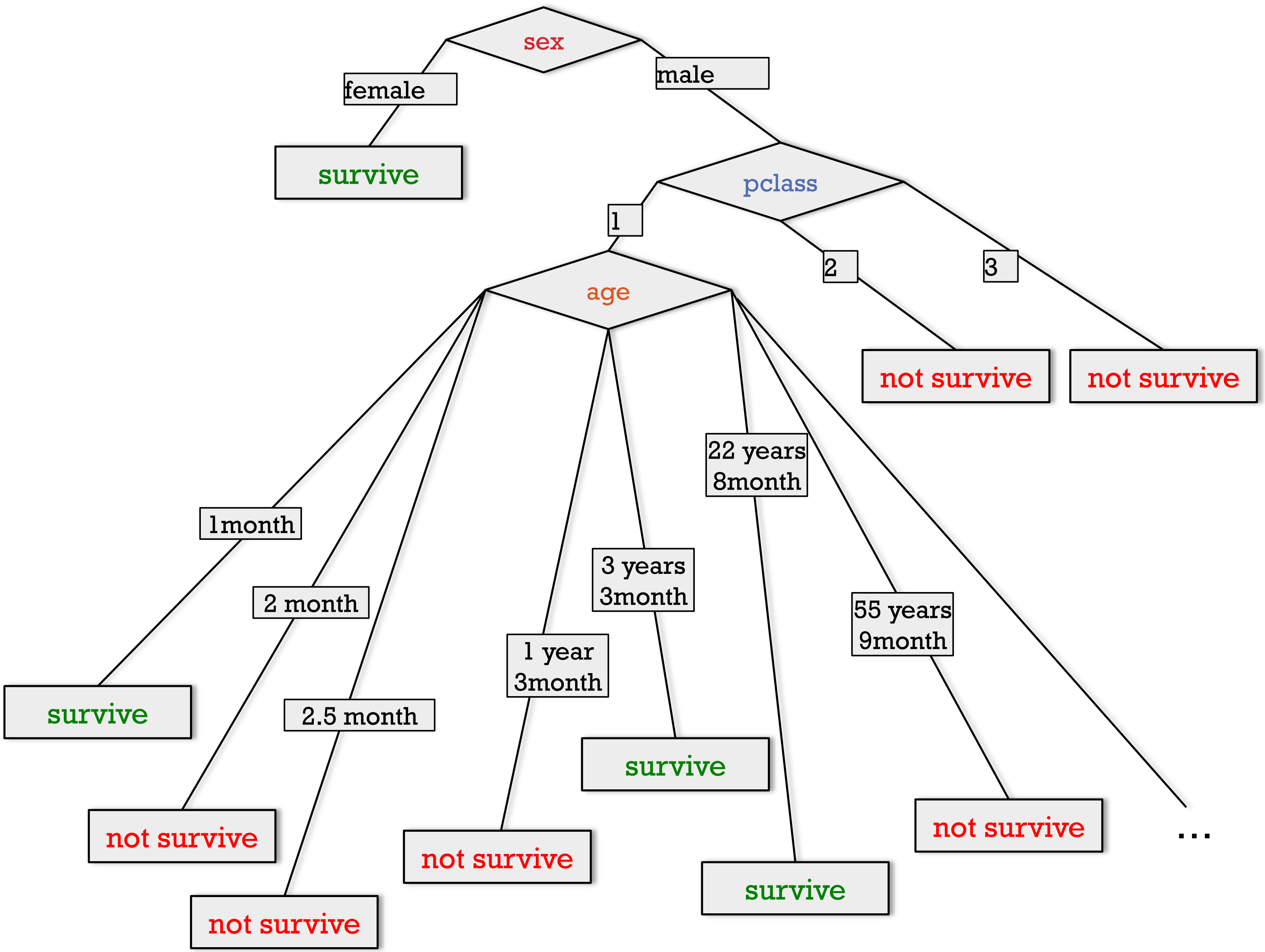
Slide credit: D. Hoiem

DECISION TREE



Every path from the root is a rule

WHAT ABOUT THIS TREE?



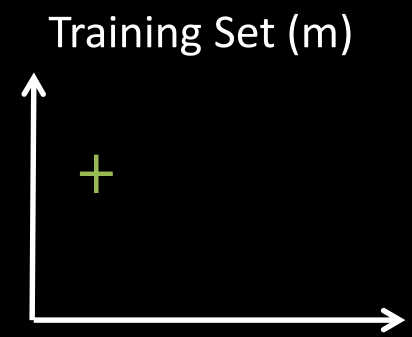
Machine Learning

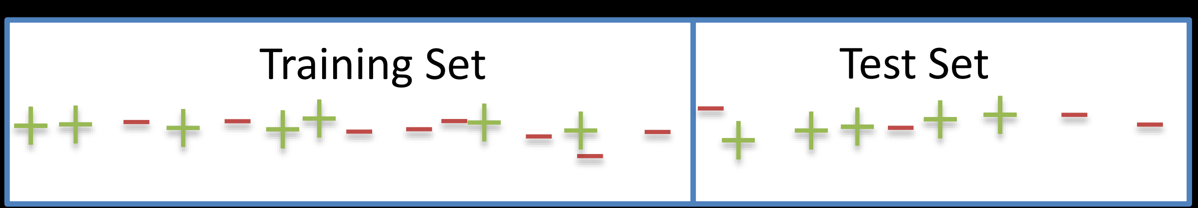
What if you model has an high error?

- Get more training examples
- Try smaller sets of features
- Try getting additional features
- Try adding polynomial features (kernels)
- Try decrease regularization
- Try increase regularization

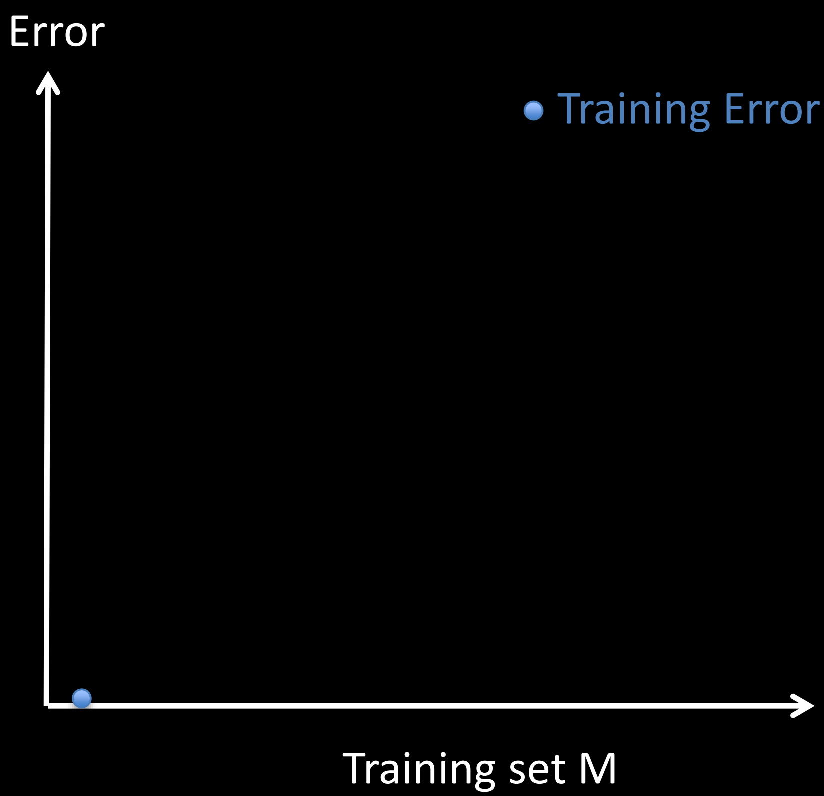
Bias and Variance

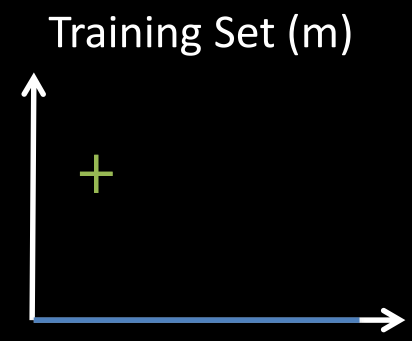
Bias and Variance

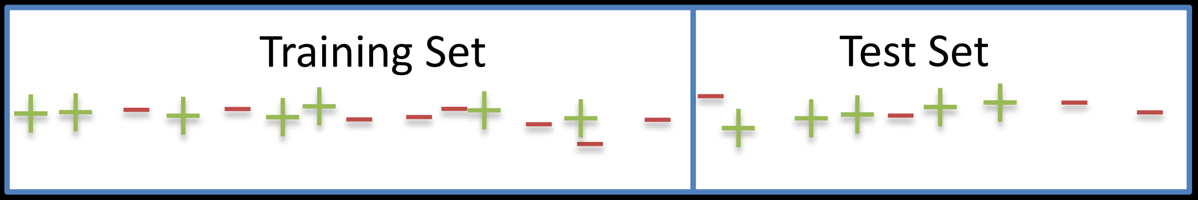




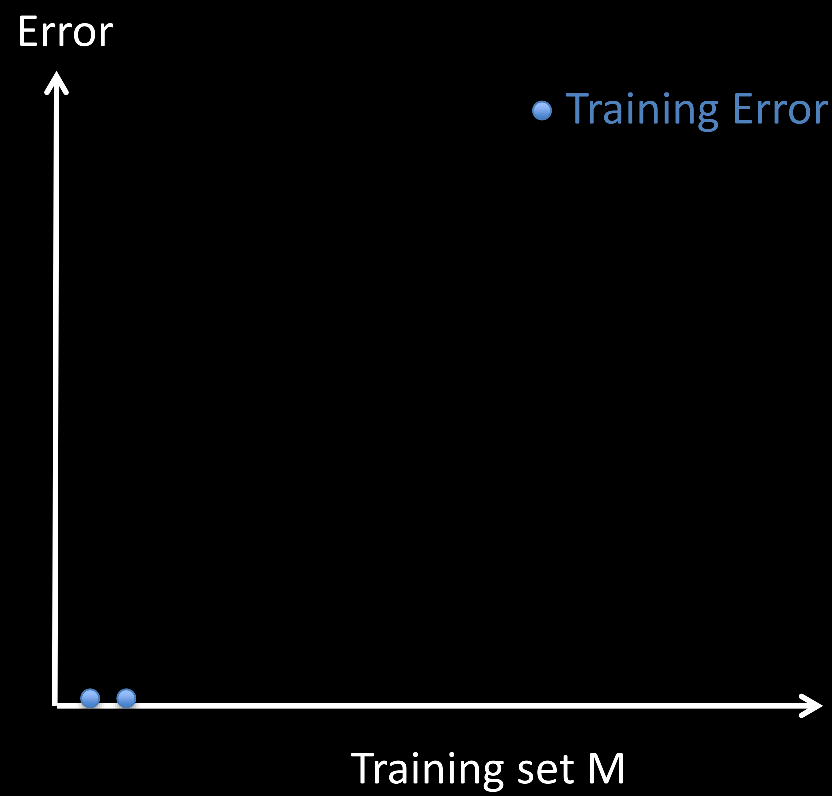
Bias and Variance

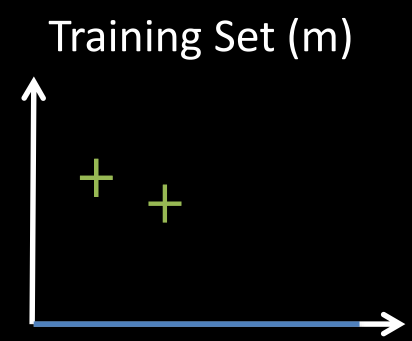


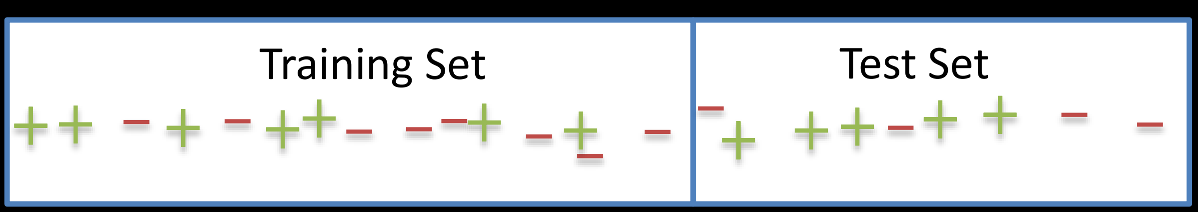




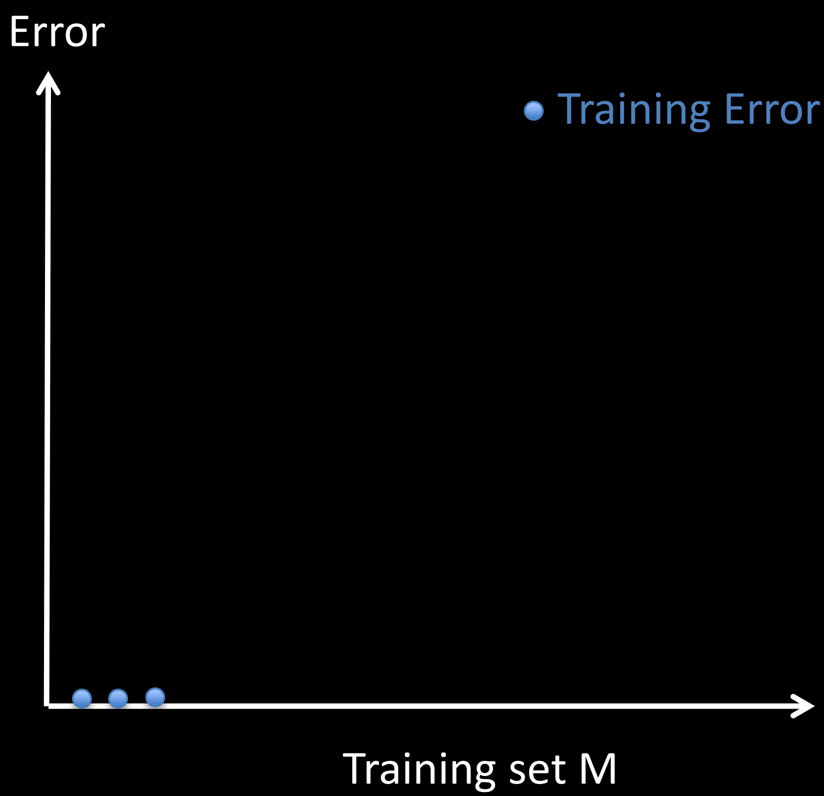
Bias and Variance

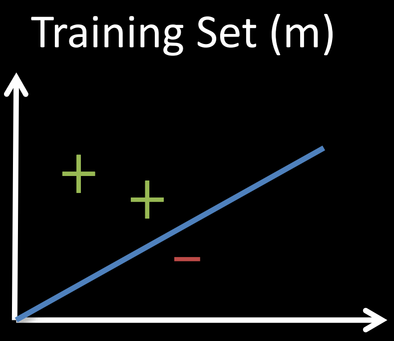


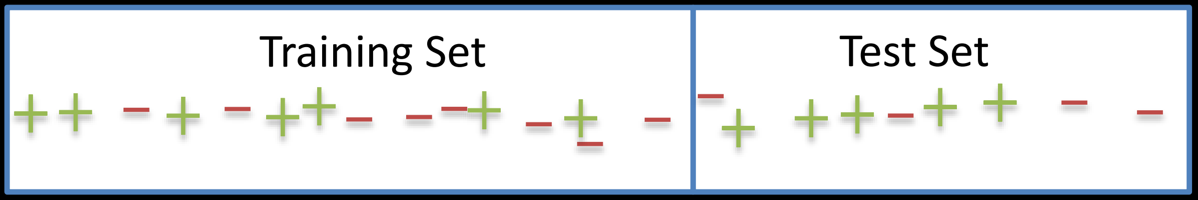




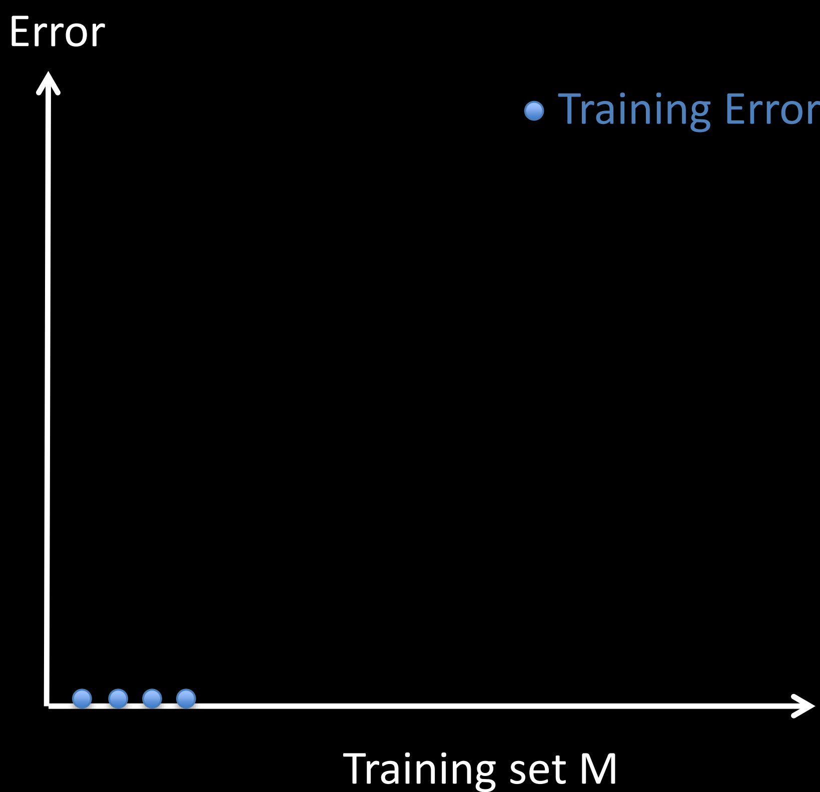
Bias and Variance

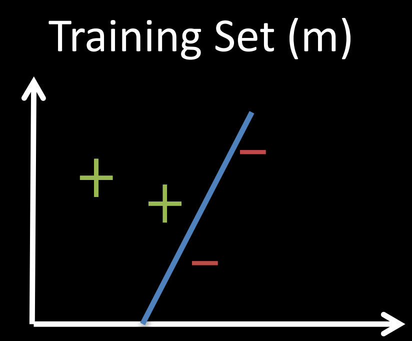


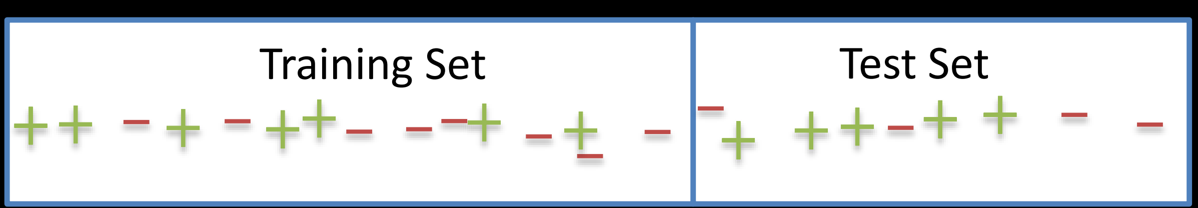




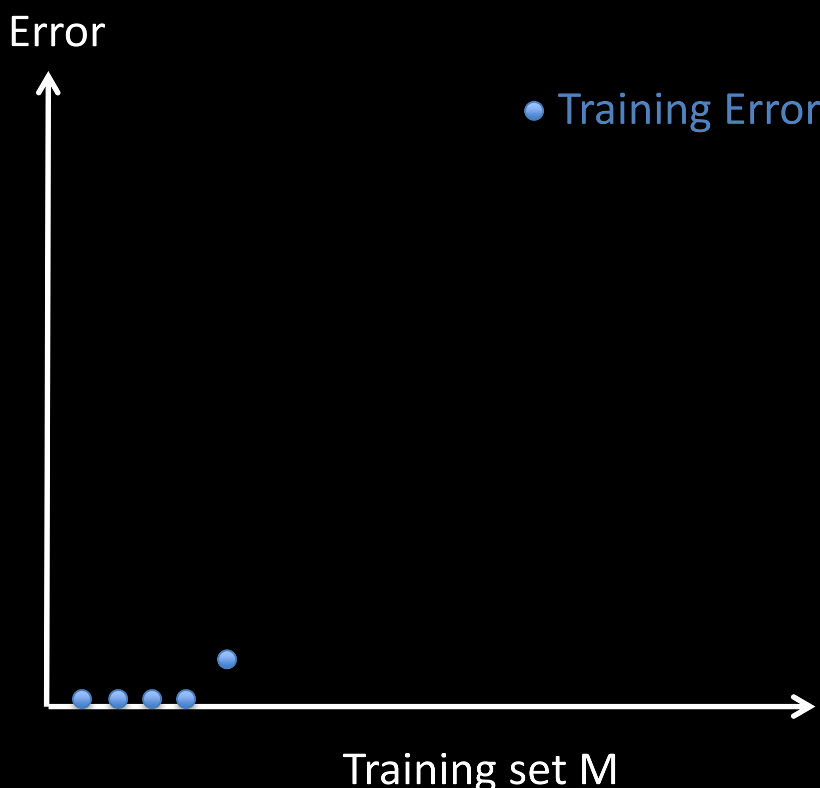
Bias and Variance

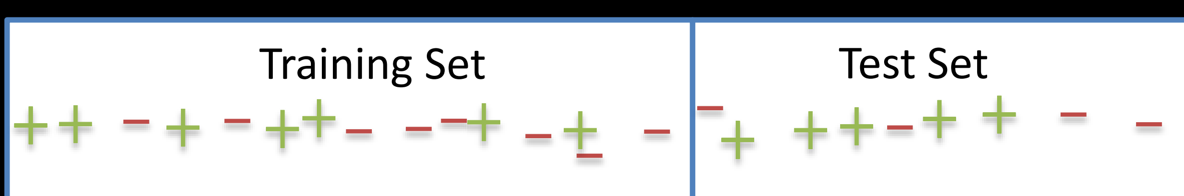




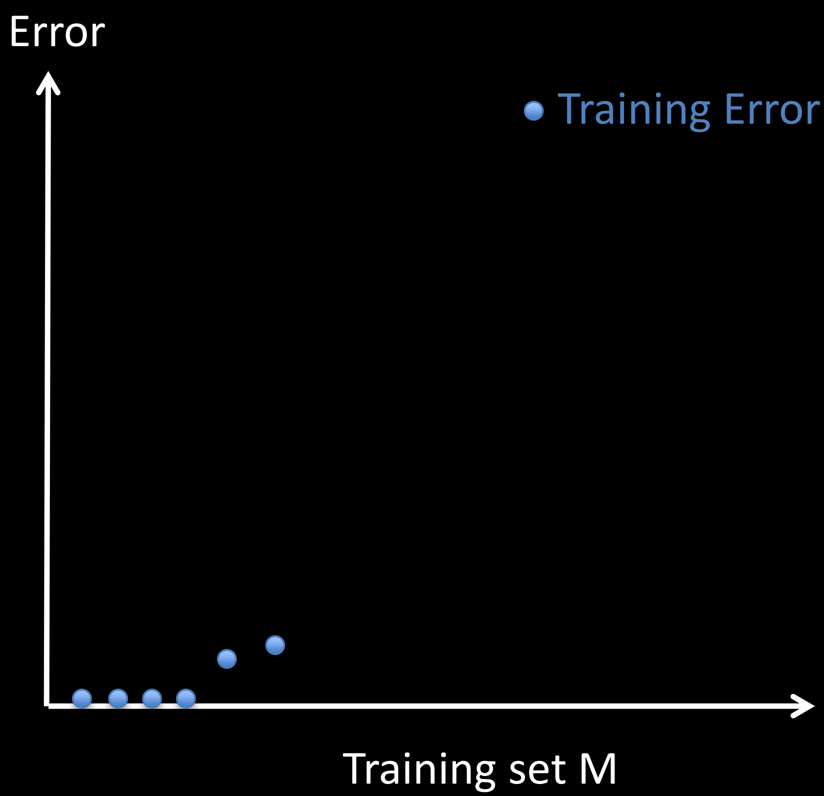


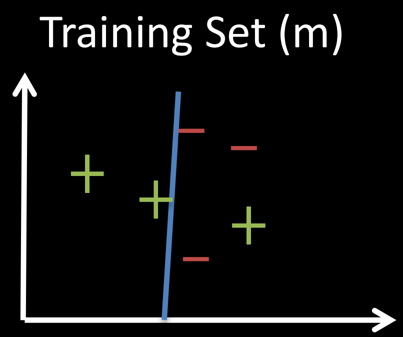
Bias and Variance

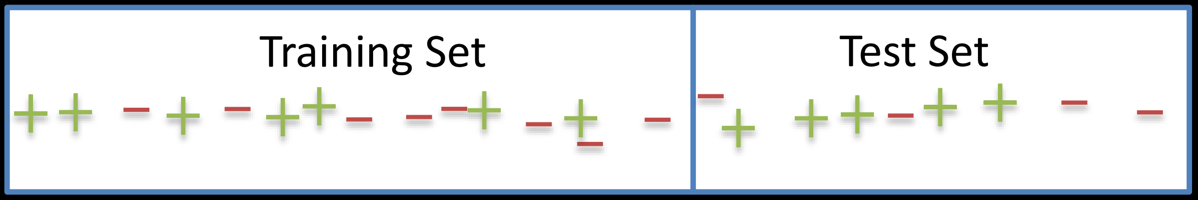




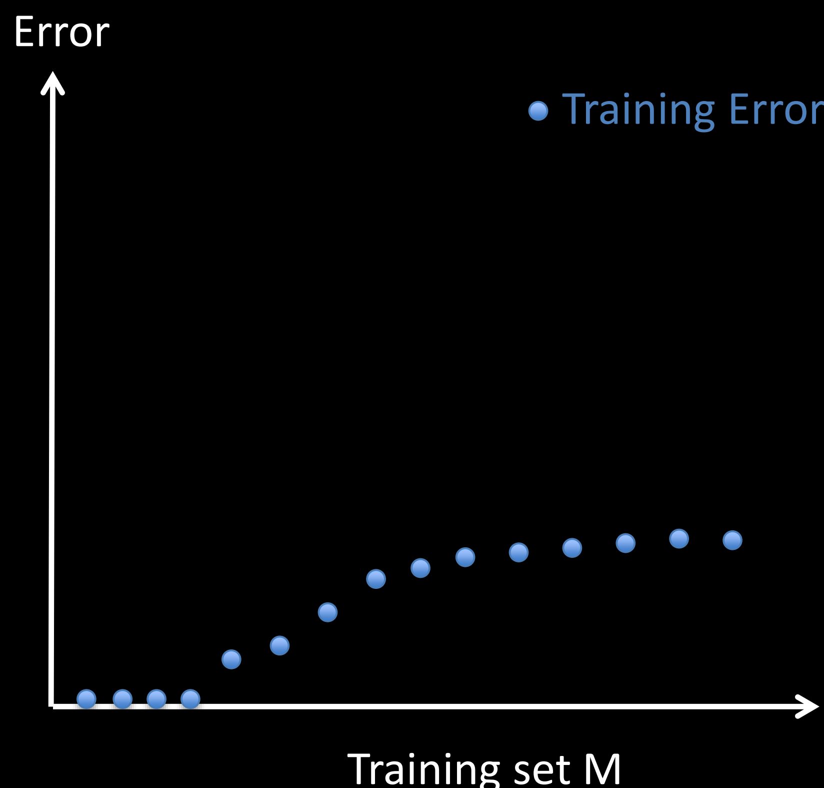
Bias and Variance

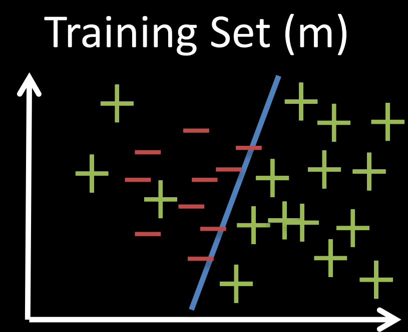


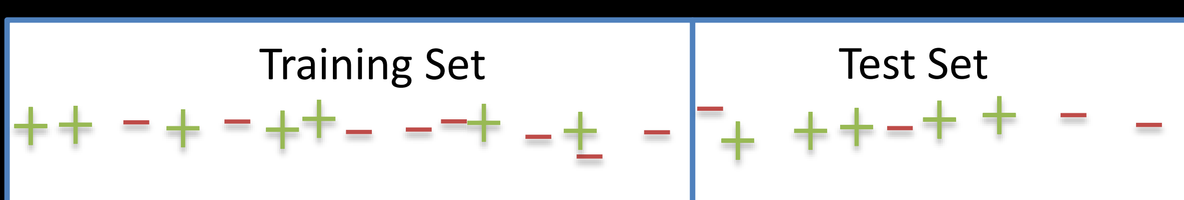




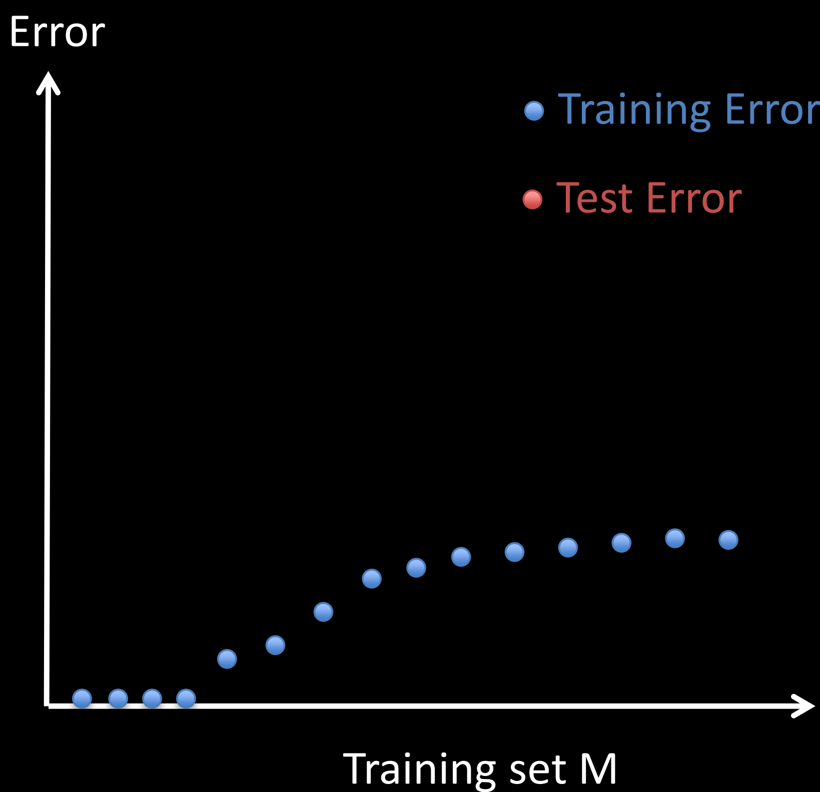
Bias and Variance







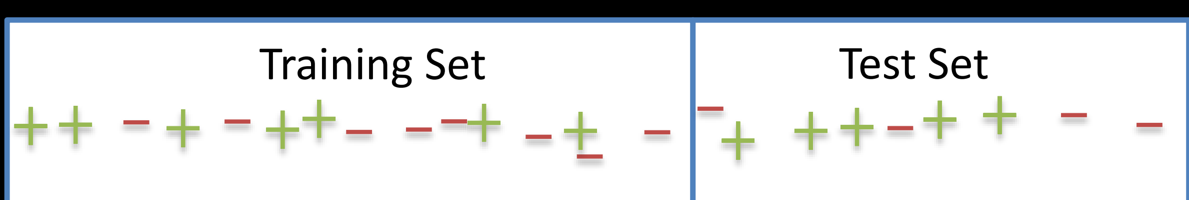
Bias and Variance



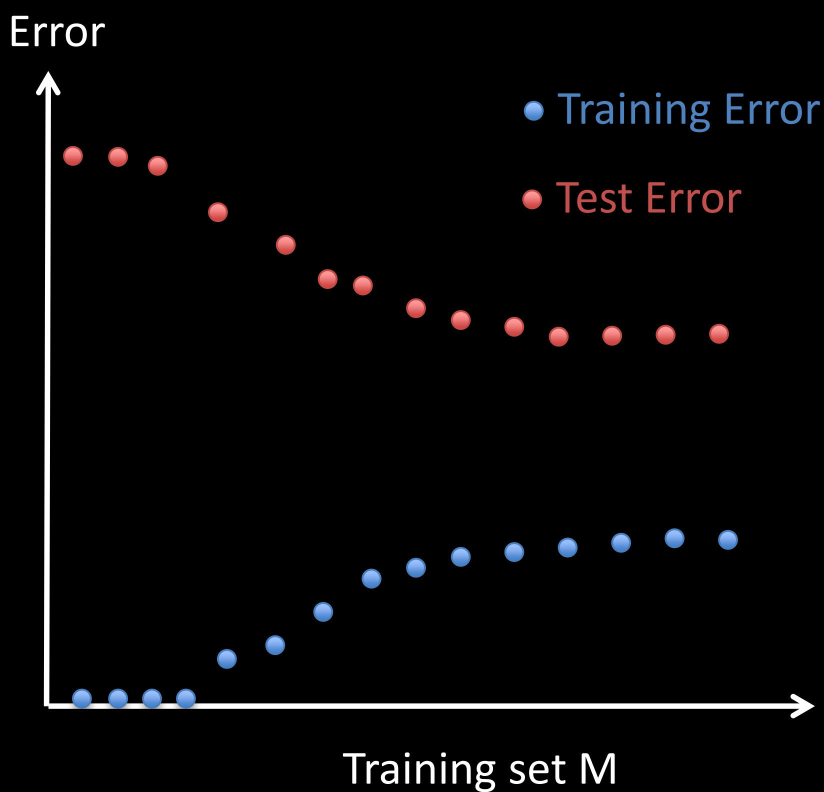
Clicker:

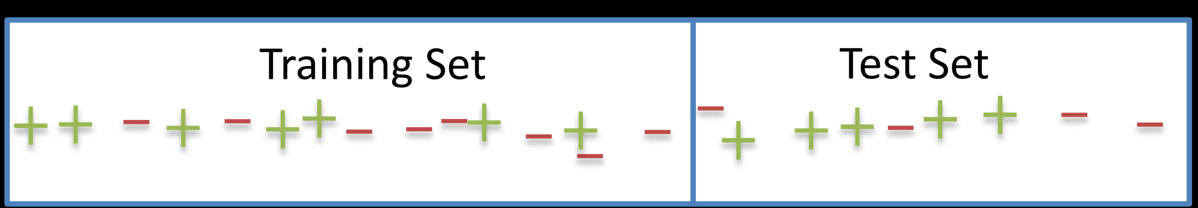
Training error

a) decreases with M
b) increases with M
c) stays constant



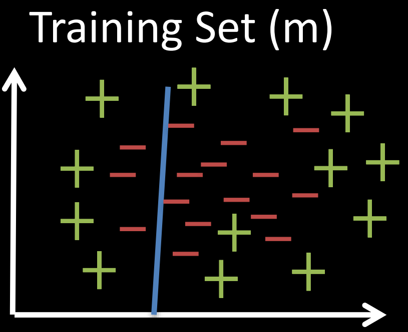
Bias and Variance



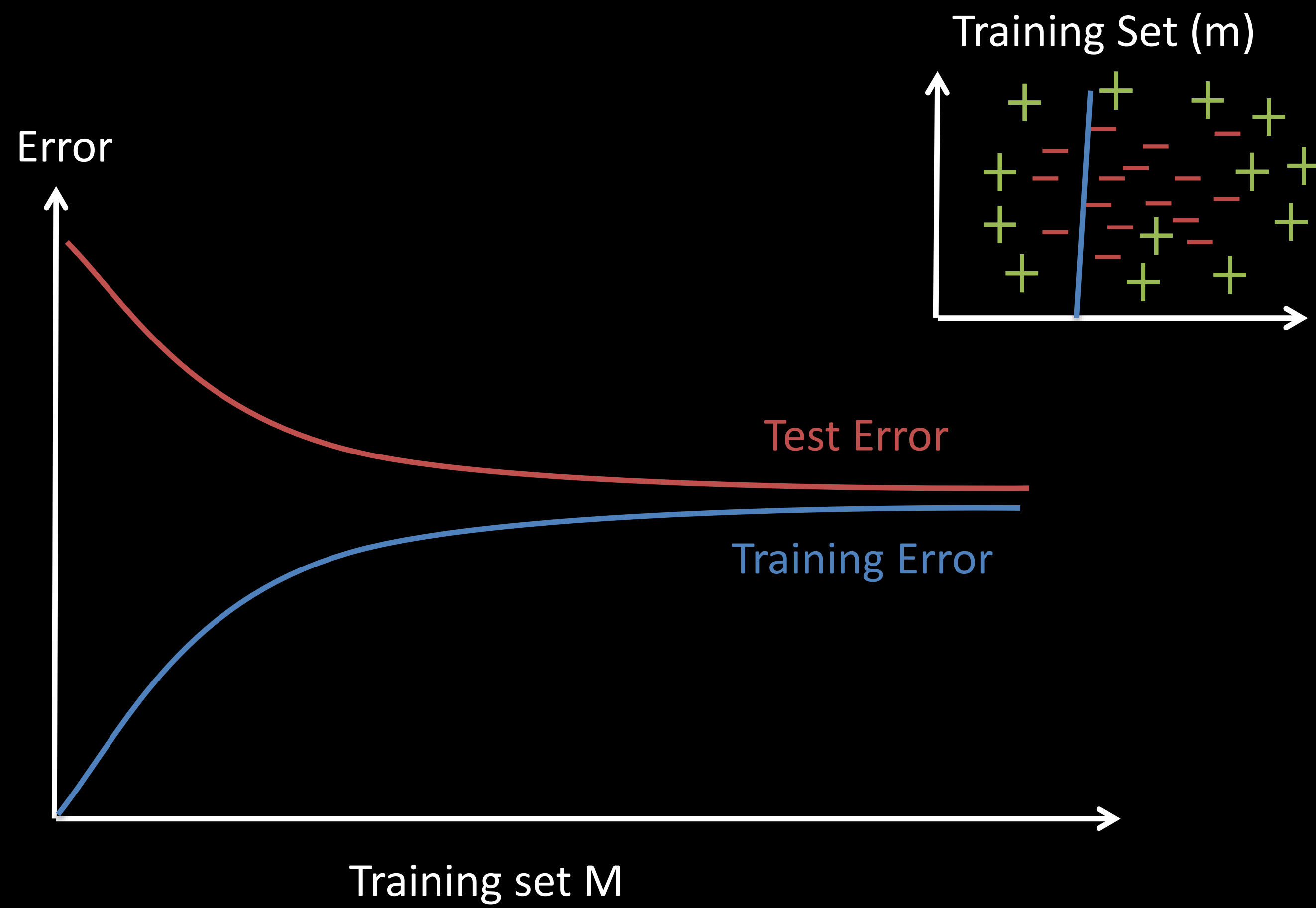


High Bias


Error
Test Error
Training Error
Training set M


Training Set (m)

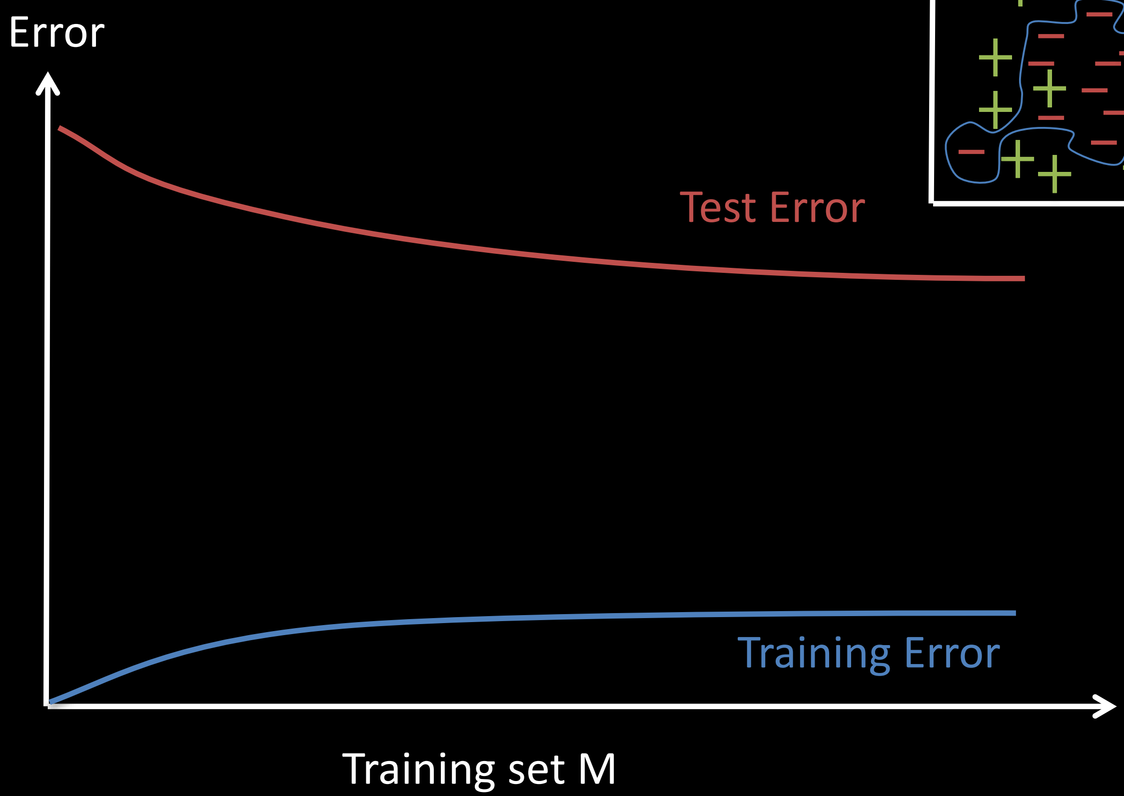
High Bias

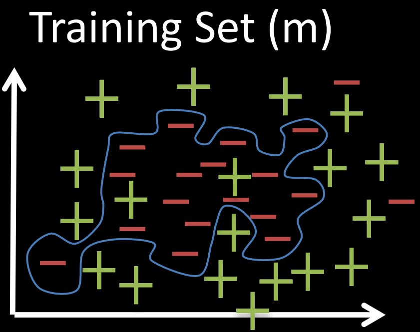


Clicker: If you have high-bias, does more data help?

a) No
b) Yes

High Variance





Overfitting

What if the knowledge and data we have are not sufficient to completely determine the correct classifier? Then we run the risk of just hallucinating a classifier (or parts of it) that is not grounded in reality, and is simply encoding random quirks in the data.

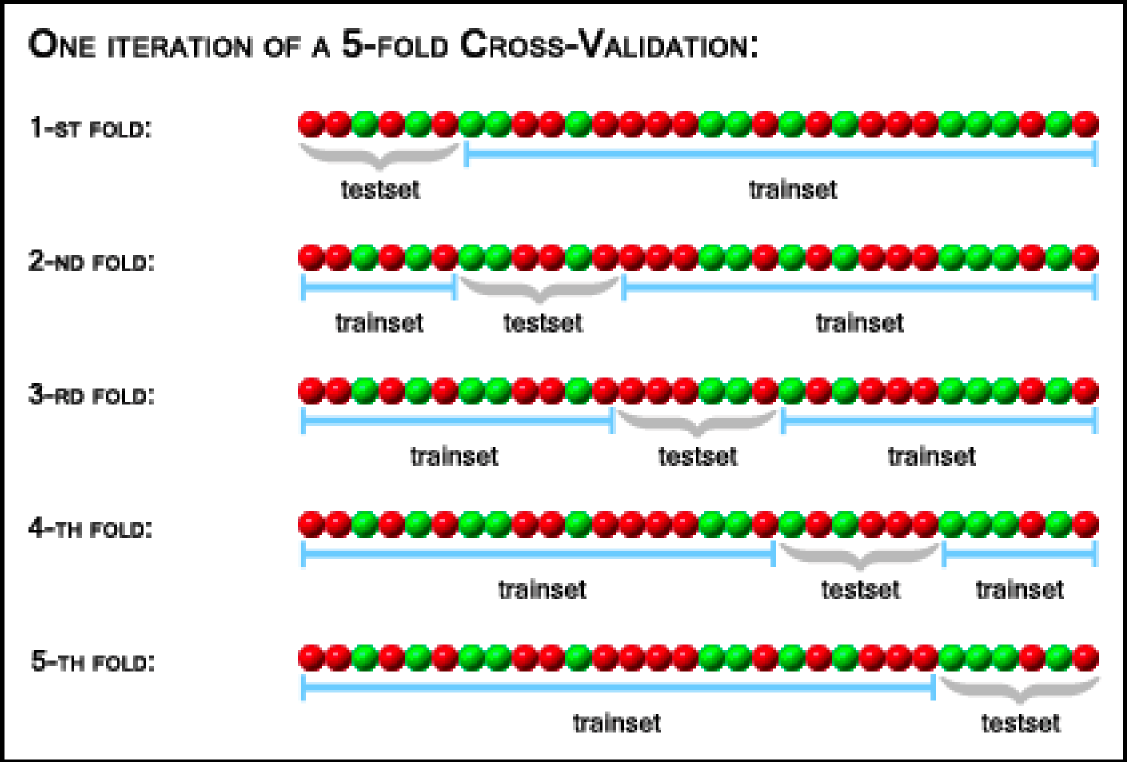
This problem is called *overfitting*, and is the bugbear of machine learning. When your learner outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both, it has overfit.

Pedro Domingos, A Few Useful Things to Know About Machine Learning, CACM 55(10), 2012

Cross-validation

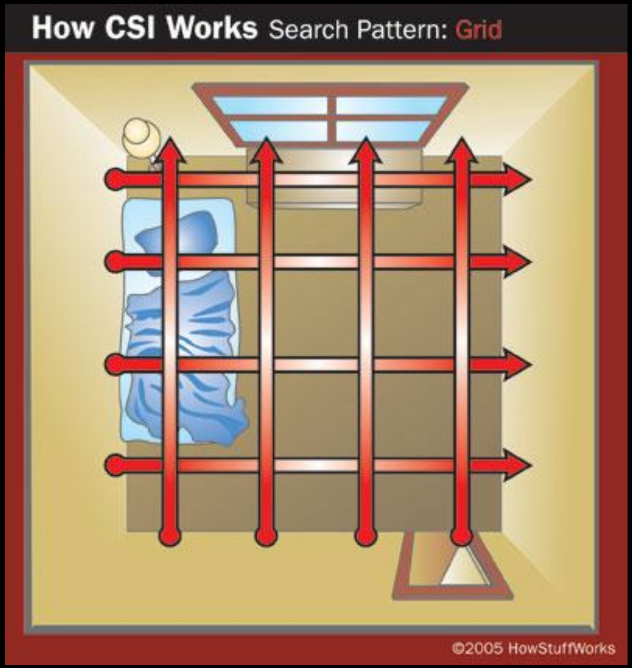
k-fold: split the data into k groups, train on every group except for one, which you test on.

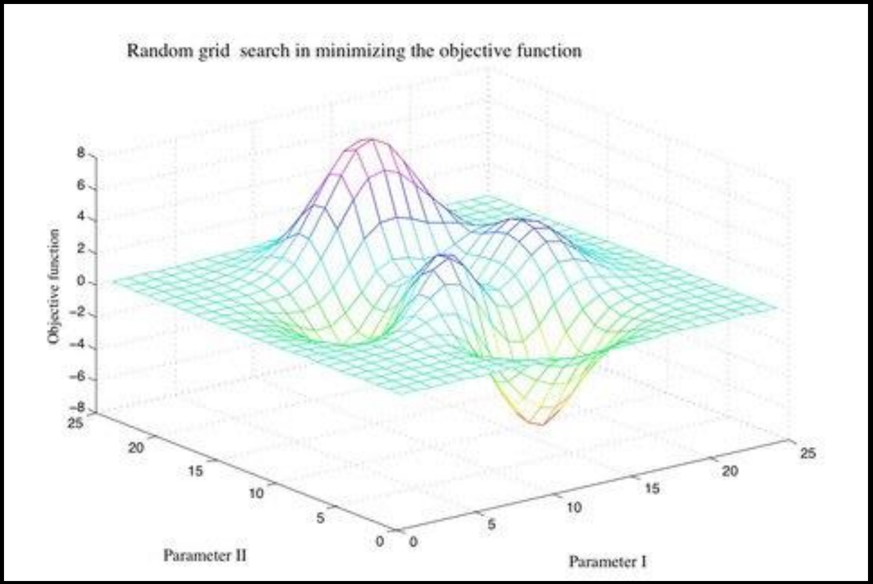
Repeat for all groups



Parameter Tuning

Grid Search





Many classifiers to choose from

- Decision Trees
- K-nearest neighbor
- Support Vector Machines
- Logistic Regression
- Naïve Bayes
- Random Forrest
- Bayesian network
- Randomized Forests
- Boosted Decision Trees
- RBMs
- ….

Slide credit: D. Hoiem