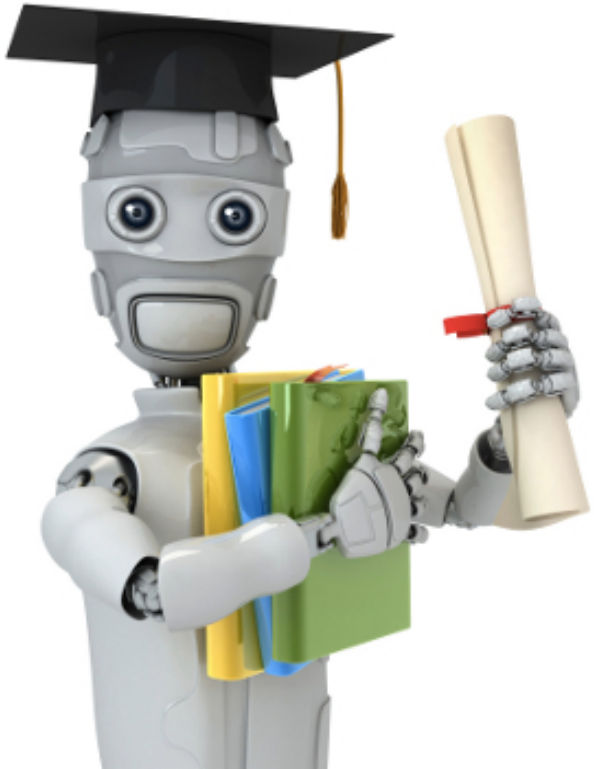
VERY BASIC OVERVIEW OF STATISTICS AND MACHINE LEARNING

INTRODUCTION TO DATA SCIENCE

ELI UPFAL



teaching datascience .org

MACHINE LEARNING – exciting!

MACHINE LEARNING – exciting!

STATISTICS - boring

MACHINE LEARNING – exciting!

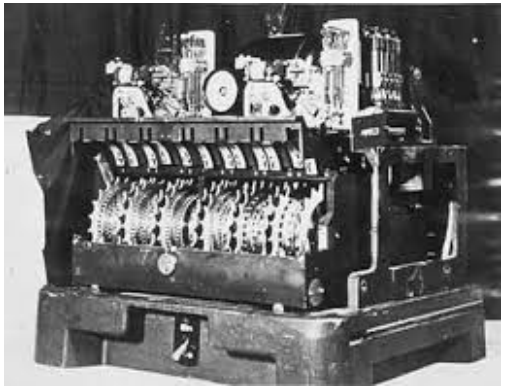
STATISTICS - boring

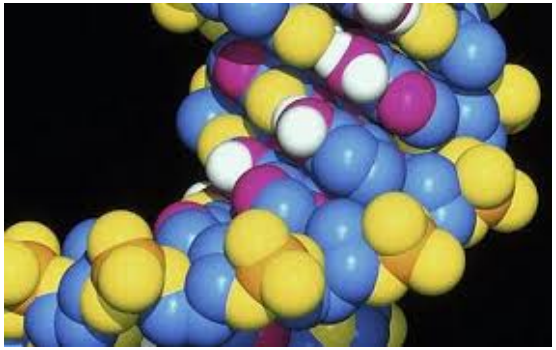
ACTULLY – not that different

EXTRACTING INFORMATION FROM DATA

Data

Analysis





Model

Predictions

"IT'S DIFFICULT TO MAKE PREDICTIONS, ESPECIALLY ABOUT THE FUTURE"

Example of ML predictions:

Observation: Millions of connections to web sites.

If you connect from IP 128.148.31.5, your computer type is …., your operating system is …., your mouse movements profile is …. , **then** your likely age is …. your likely disposable income is …. your likely gender is …., your likelihood of a purchase is…

Works very well in practice!

"IT'S DIFFICULT TO MAKE PREDICTIONS, ESPECIALLY ABOUT THE FUTURE"

Example of ML predictions:

Observation: Millions of connections to web sites.

If you connect from IP 128.148.31.5, your computer type is …., your operating system is …., your mouse movements profile is …. , etc. **then** your likely age is …. your likely disposable income is …. your gender is …., your likelihood of a purchase is…

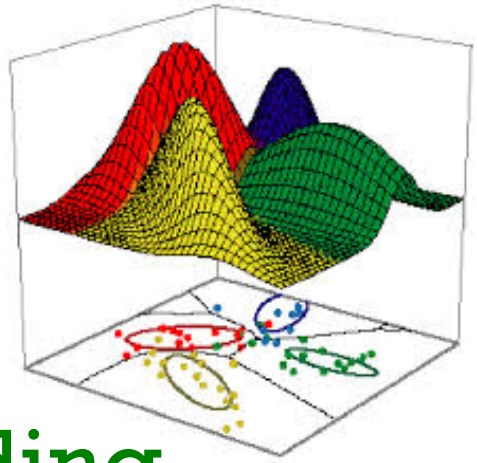
Works very well in practice!

Observation: Among flu patients, those who take medicine have a longer recovery

If you take medicine for flu **then** you have longer recovery.

Obviously wrong!

Just because a machine learning, data mining, or data analysis application outputs a result - it doesn't mean that it's right



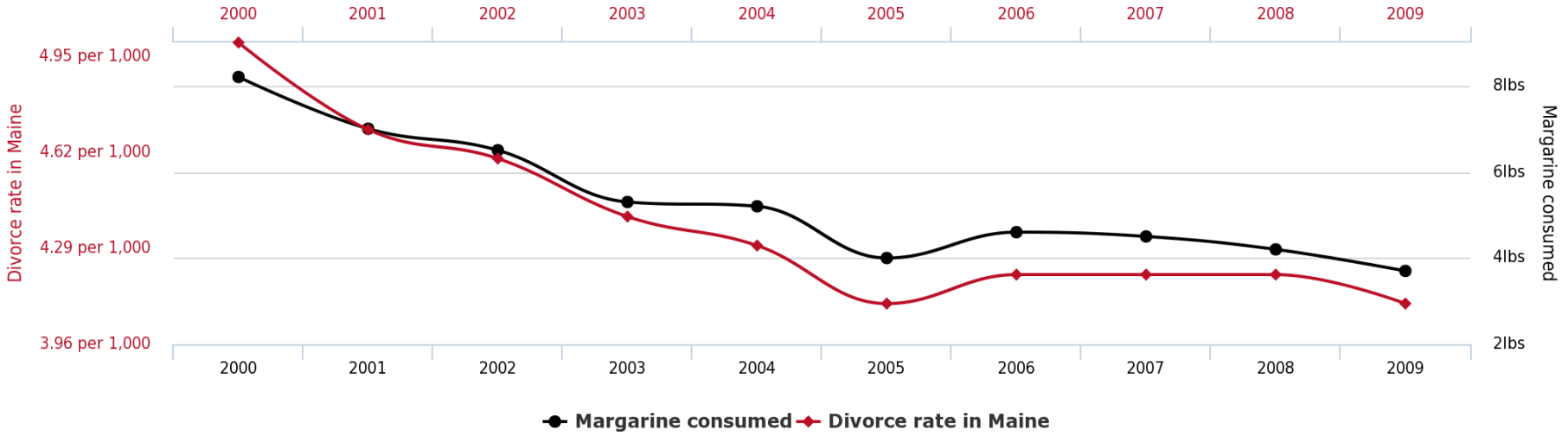
Data analysis is often misleading

Machine learning without statistical analysis is pure nonsense

Divorce rate in Maine

correlates with

Per capita consumption of margarine

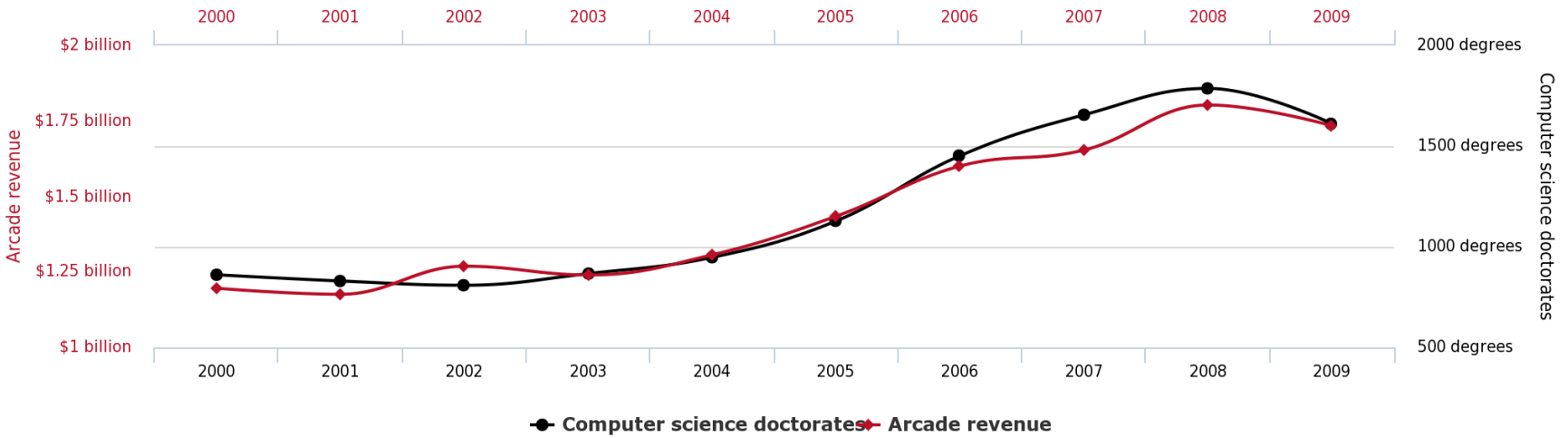


tylervigen.com

Total revenue generated by arcades

correlates with

Computer science doctorates awarded in the US

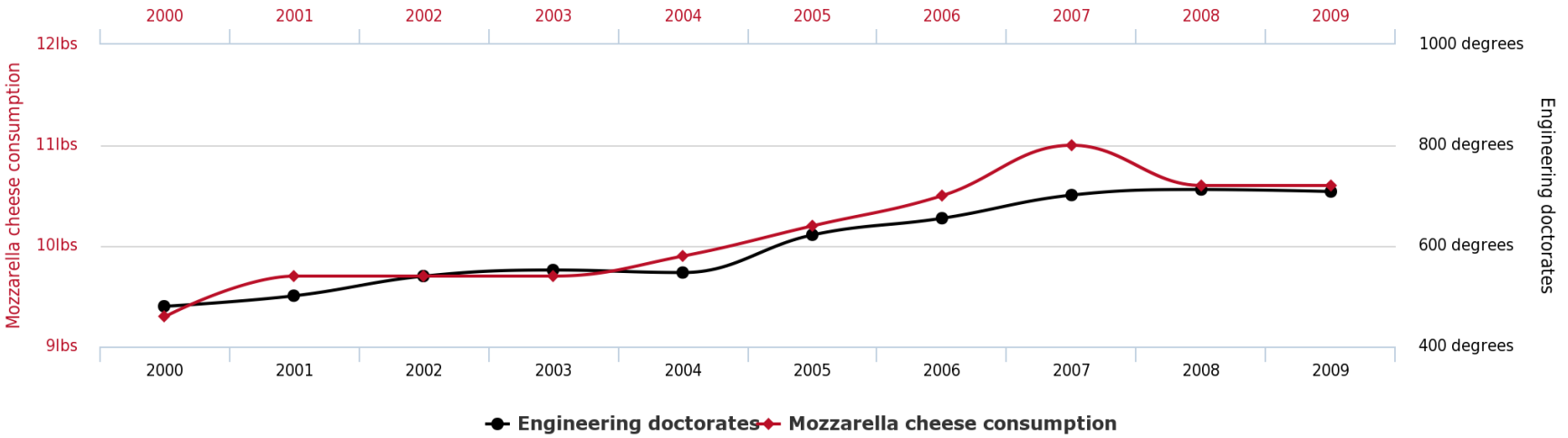


tylervigen.com

Per capita consumption of mozzarella cheese

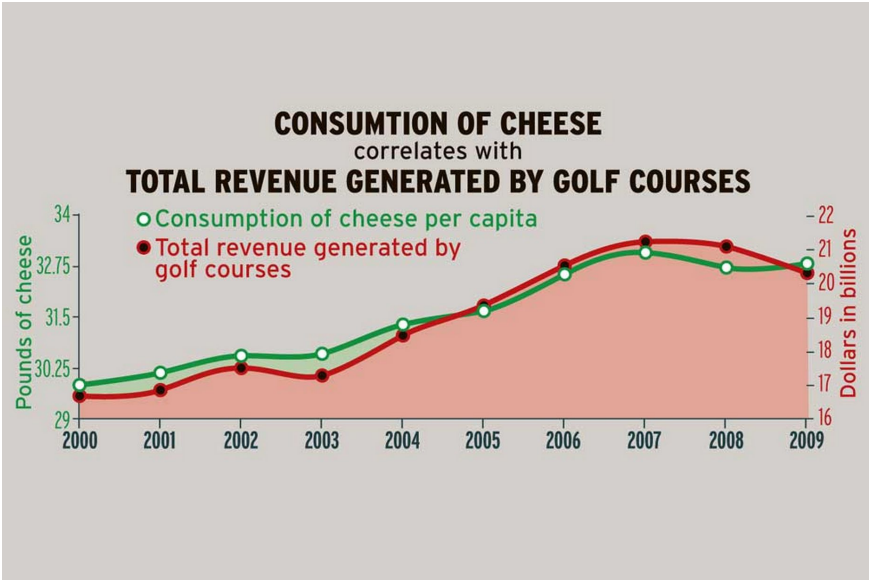
correlates with

Civil engineering doctorates awarded



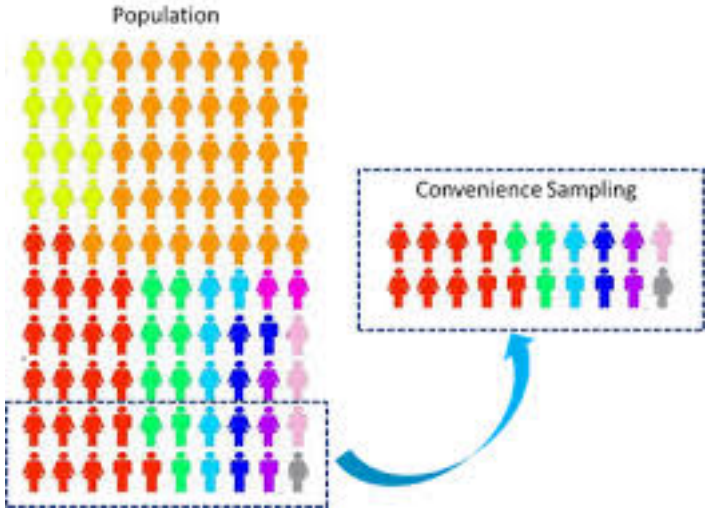
tylervigen.com

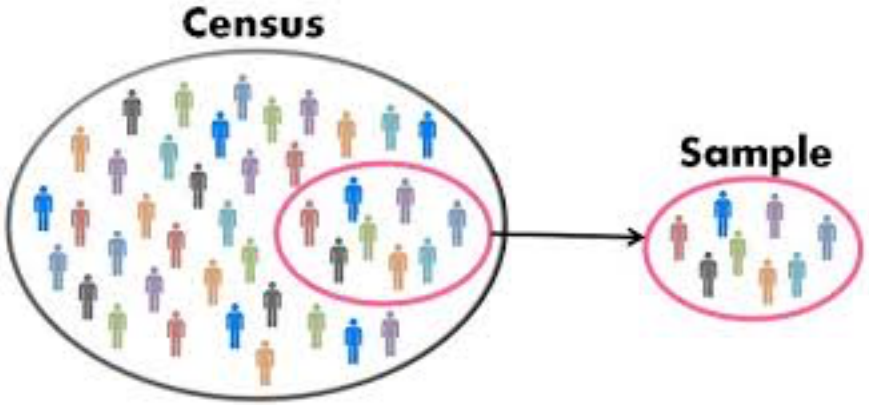
Which Correlation is more convincing?

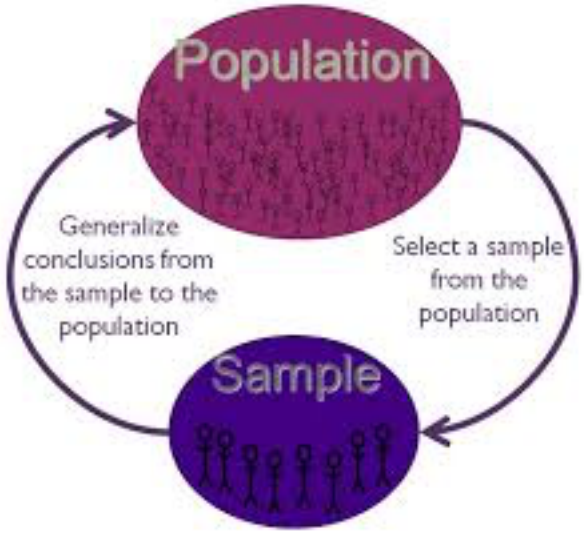


How do we distinguish between facts and coincident?

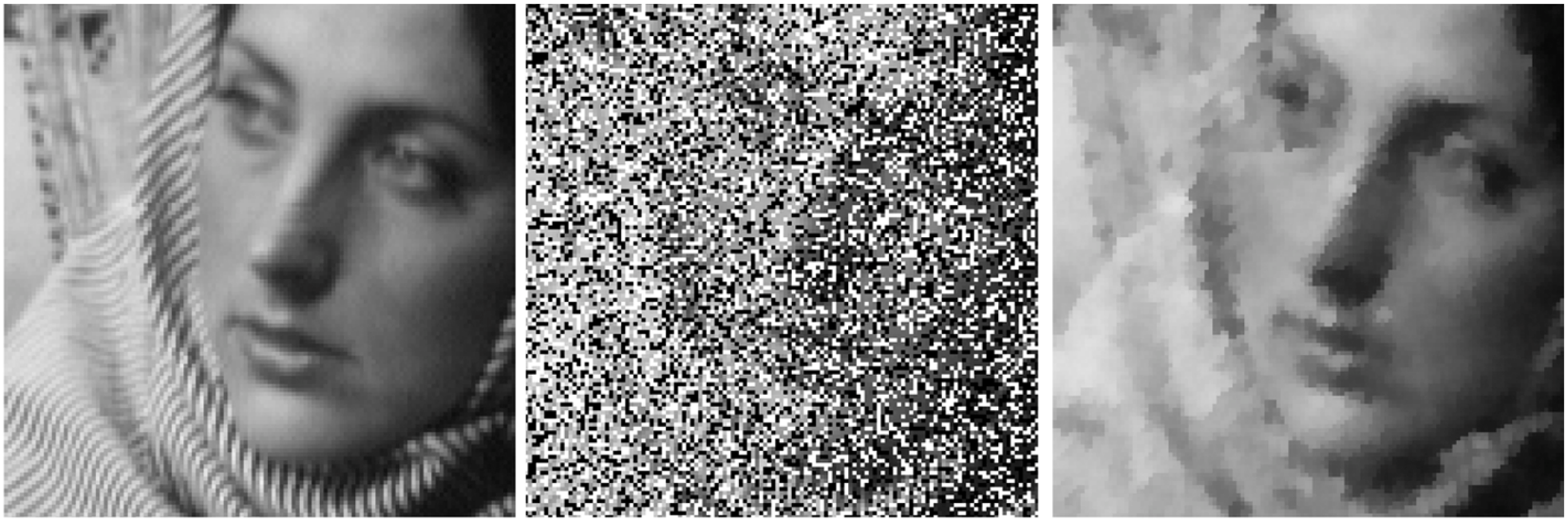
DATA: LEARNING FROM A TRAINING SET (SAMPLE)







THE DATA IS THR MODEL



(a) (b) (c)

STATISTICS VS. MACHINE LEARNING

First there was statistics:

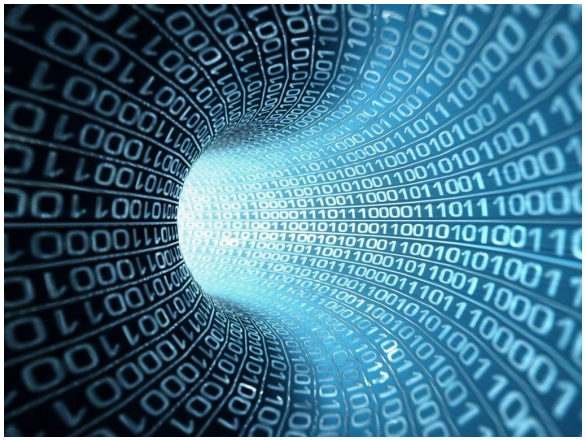
Strict criteria for when an hypothesis (''discovery'') is statistically significant

Strong assumptions, elaborate computation

Then came Computer Science:

Emphasize on efficient computation

Output best approximation, even if not certain



… And a lot of BIG data

With lucrative business applications.

GOAL OF THIS PART OF THE COURSE

Basic machine learning tools for data analysis

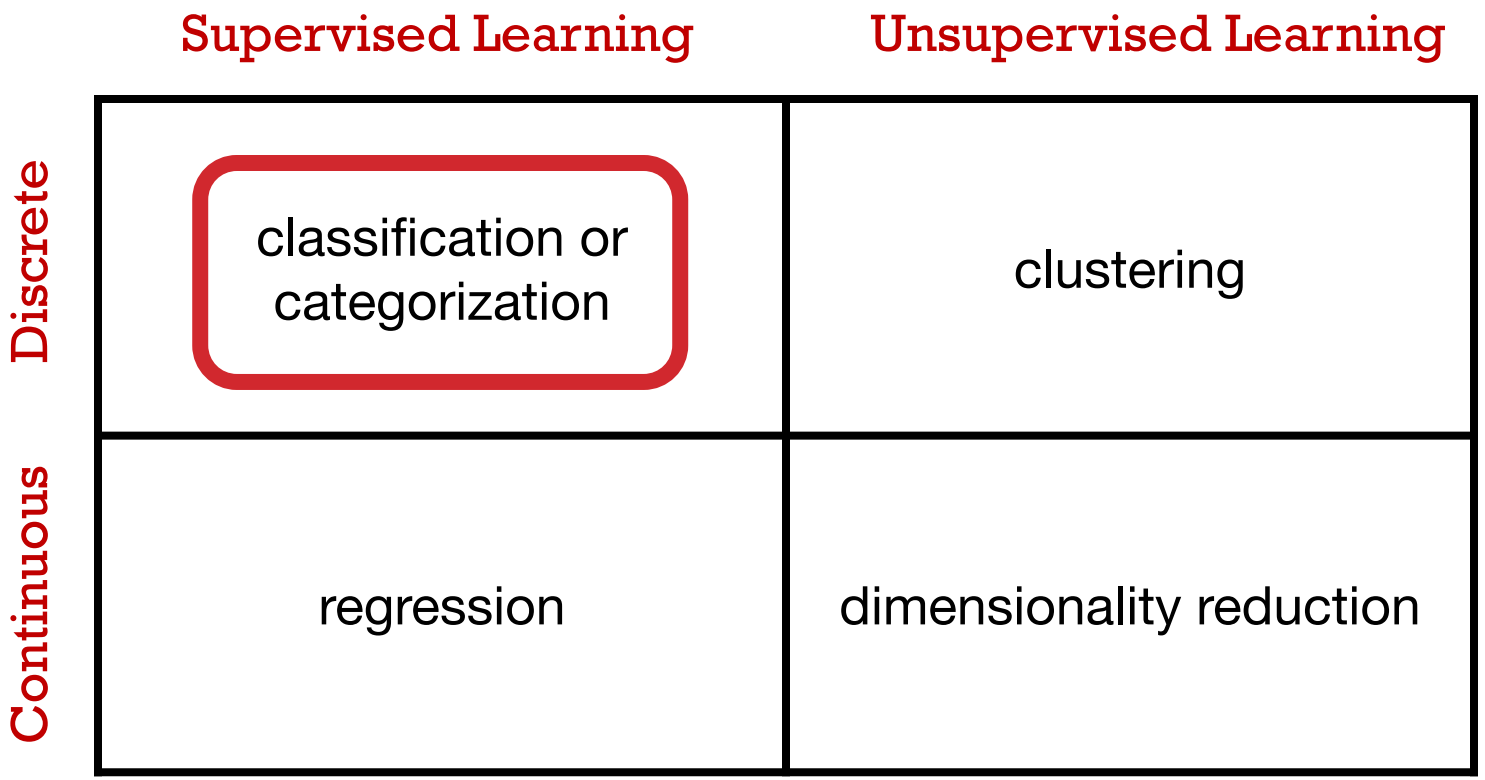
Very basic concepts in probability and statistics

Understanding the power and pitfalls of data analysis

MACHINE LEARNING PROBLEMS

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

MACHINE LEARNING PROBLEMS

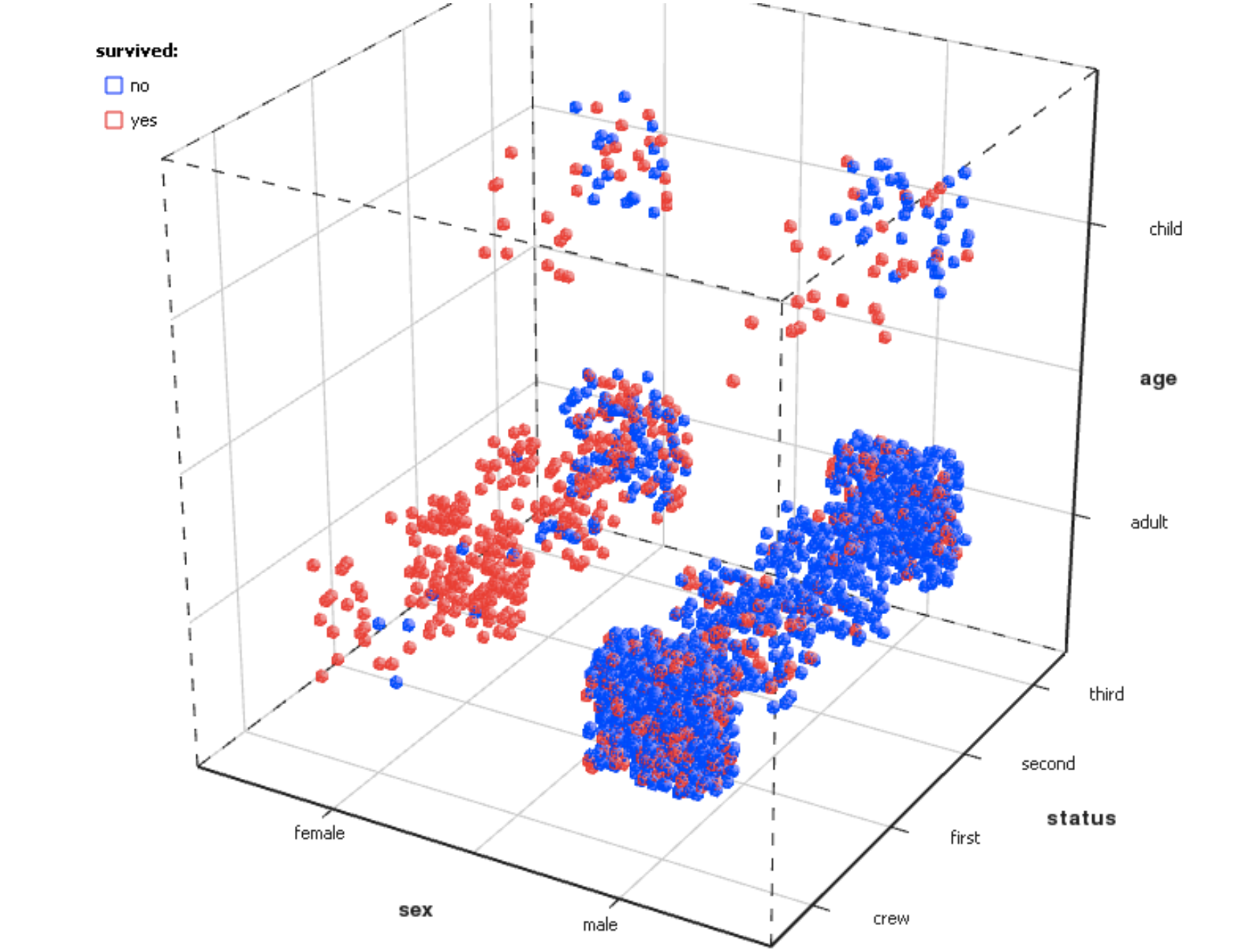


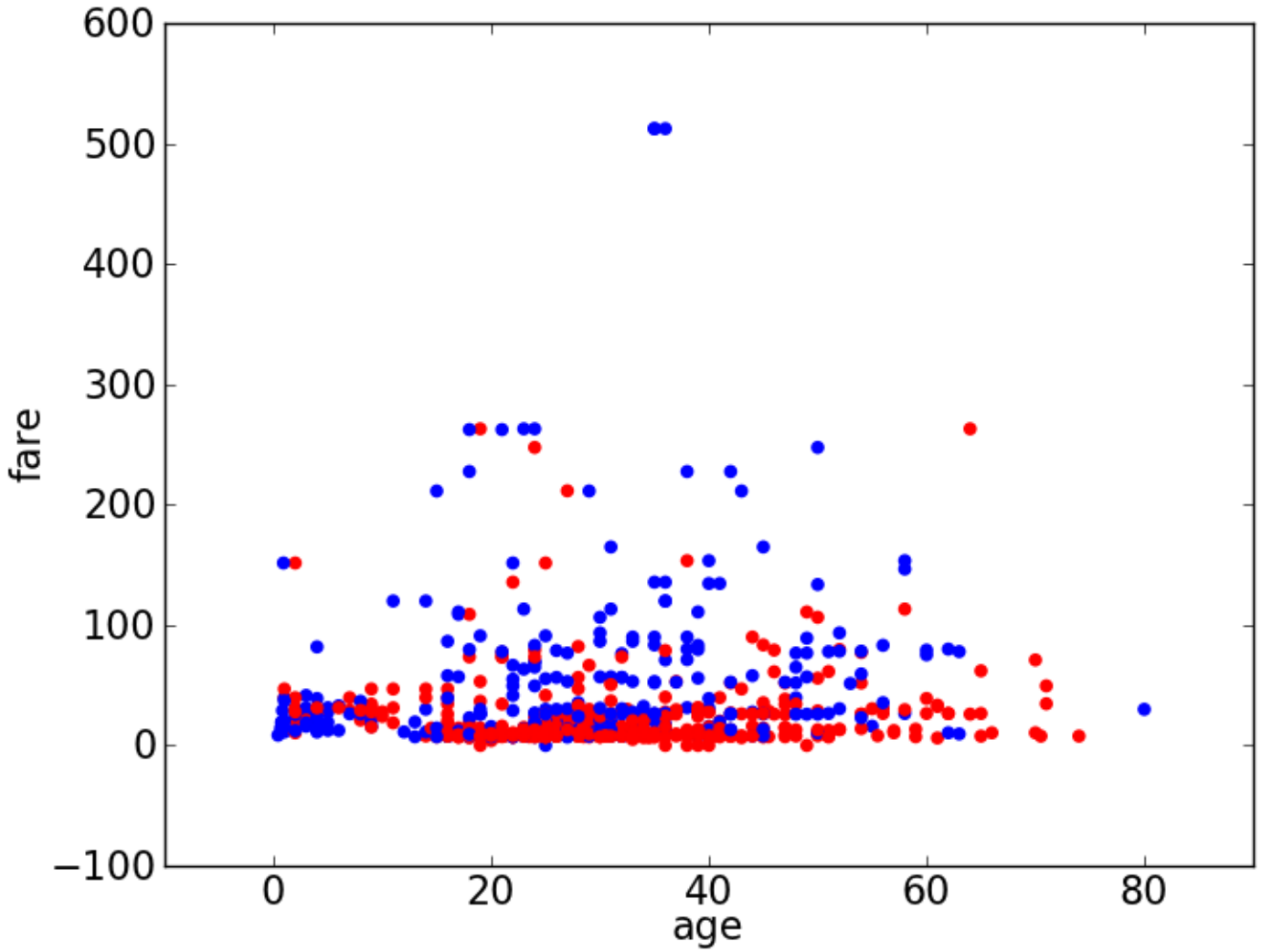
EXAMPLE: TITANIC DATASET

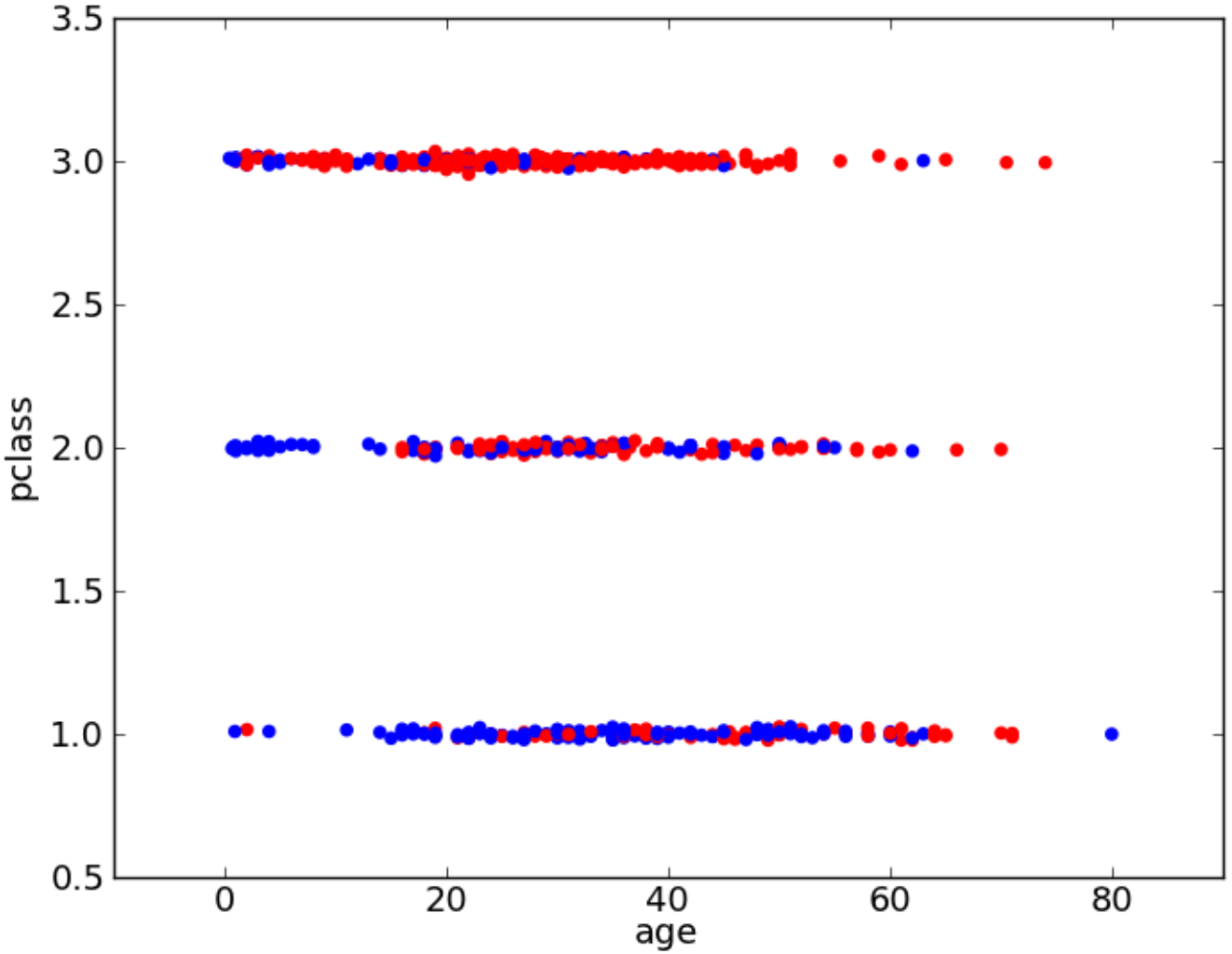
Label | Features

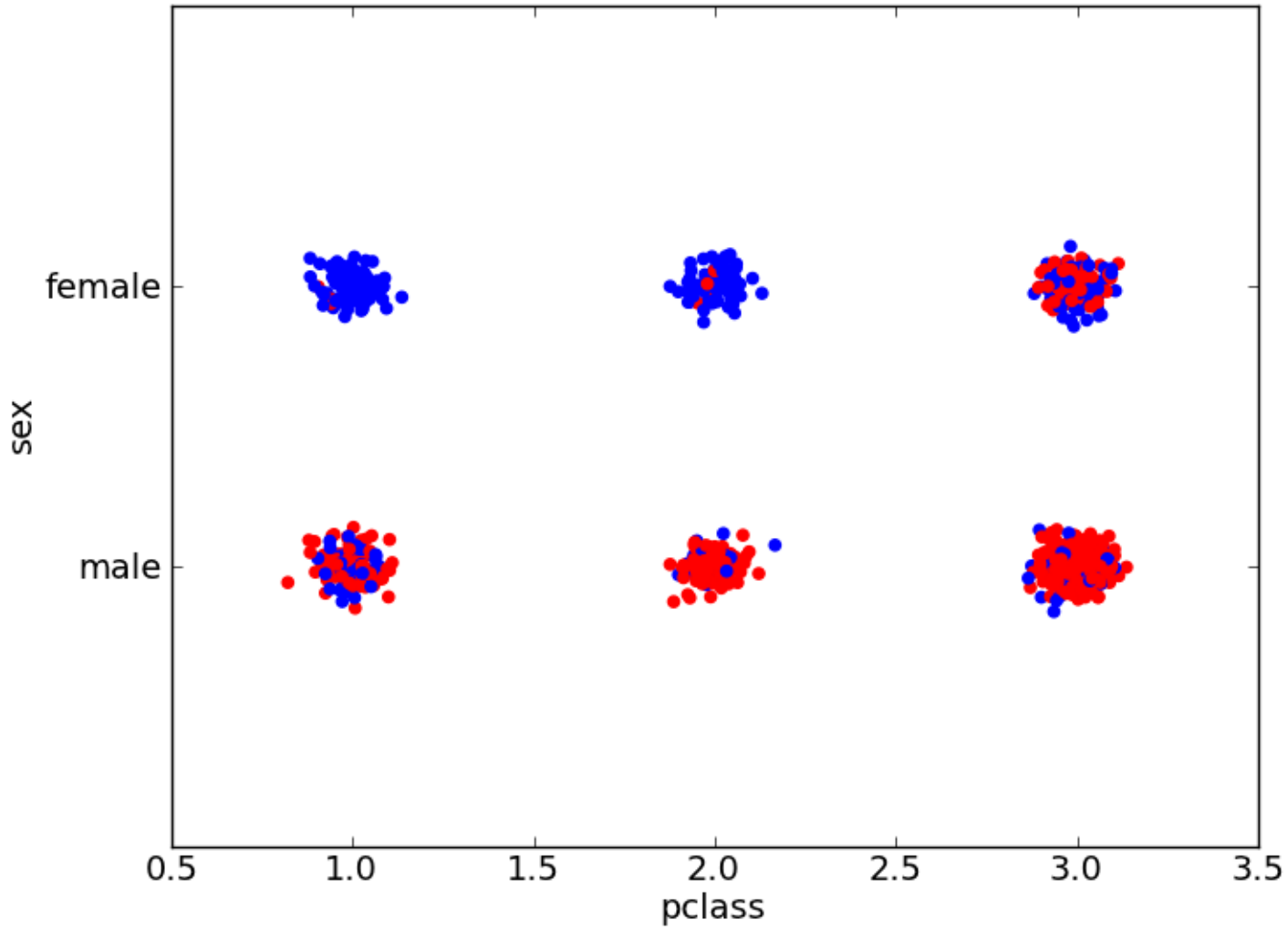
survived	pclass	sex	age	sibsp	parch	fare	cabin	embarked
0	3	male	22	1	0	7.25		S
1	1	female	38	1	0	71.2833	C85	C
1	3	female	26	0	0	7.925		S
1	1	female	35	1	0	53.1	C123	S
0	3	male	35	0	0	8.05		S
0	3	male		0	0	8.4583		Q
0	1	male	54	0	0	51.8625	E46	S
0	3	male	2	3	1	21.075		S
1	3	female	27	0	2	11.1333		S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

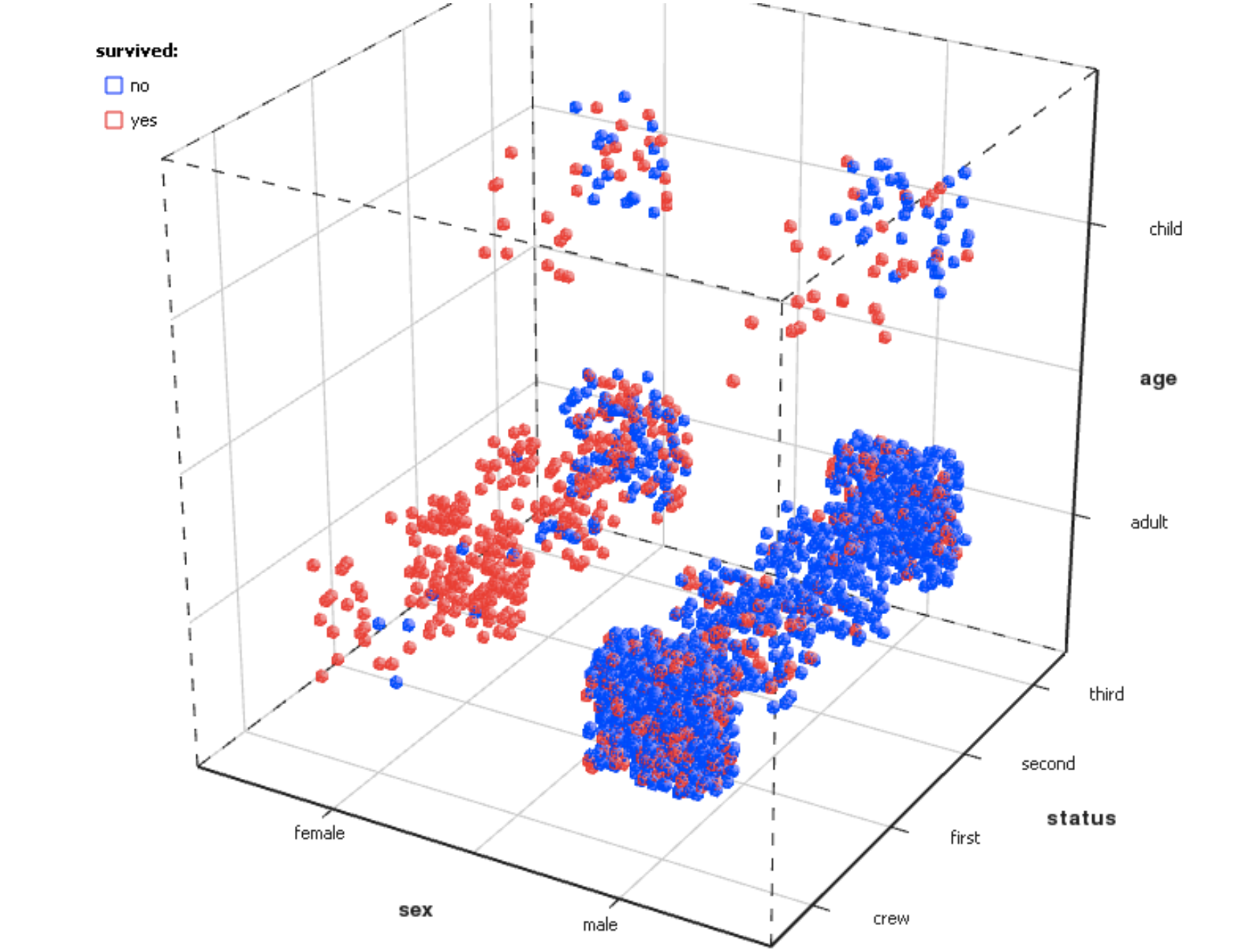
Can we predict survival from these features?


survived:
no
yes
child
age
adult
third
second
status
first
crew
female
sex
male

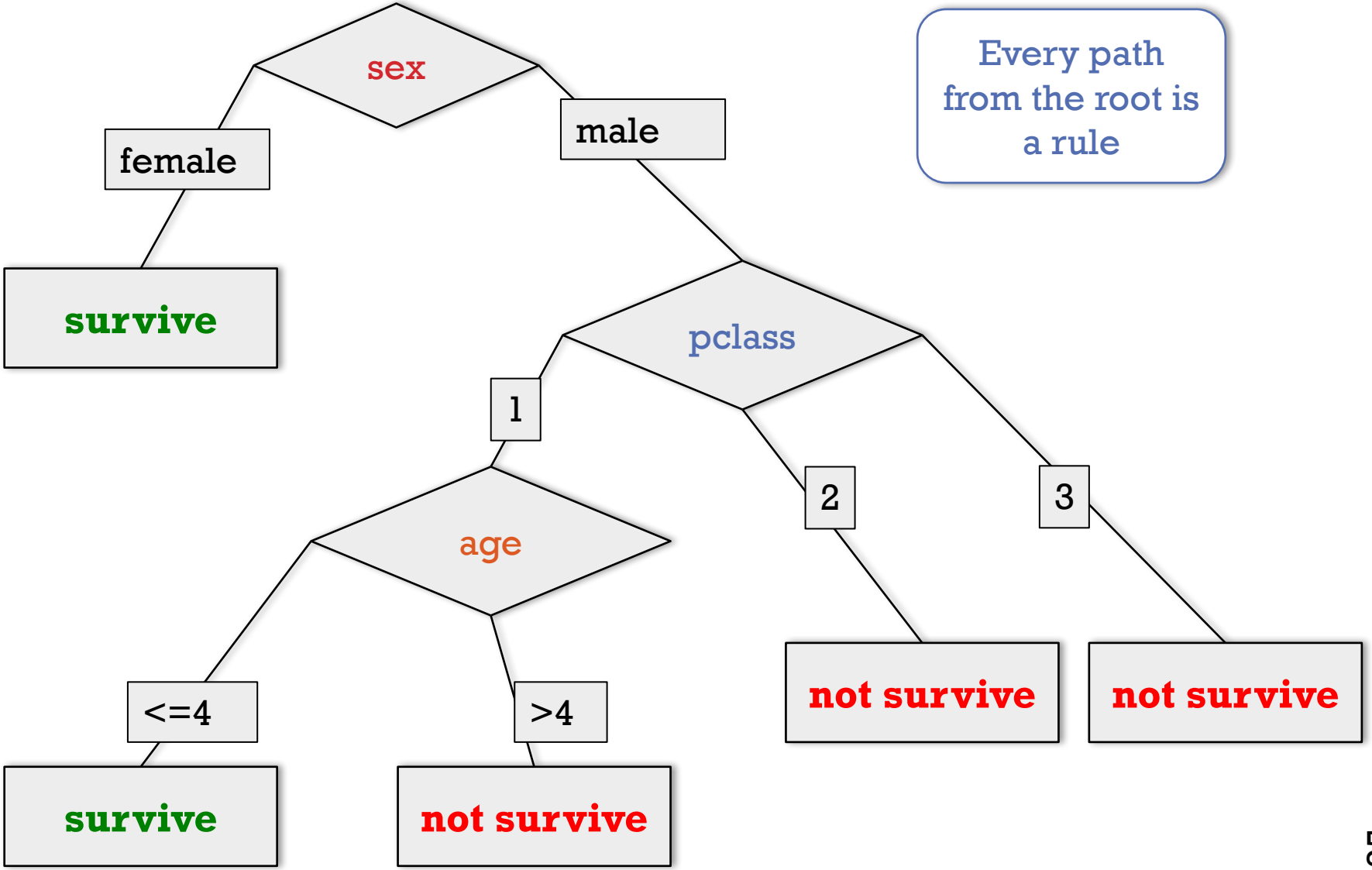

600
500
400
300
200
100
0
−100
fare
0
20
40
60
80
age


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


survived:
no
yes
child
age
adult
third
second
status
first
crew
female
sex
male

DECISION TREE



WHAT IS A CLASSIFIER

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

f() = “apple”

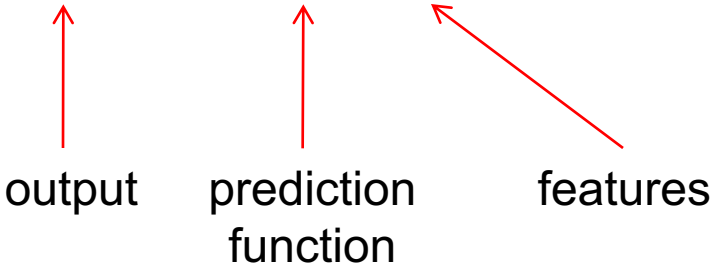
f() = “tomato”

f() = “cow”

Slide credit: L. Lazebnik

THE MACHINE LEARNING FRAMEWORK

$$y = f(x)$$

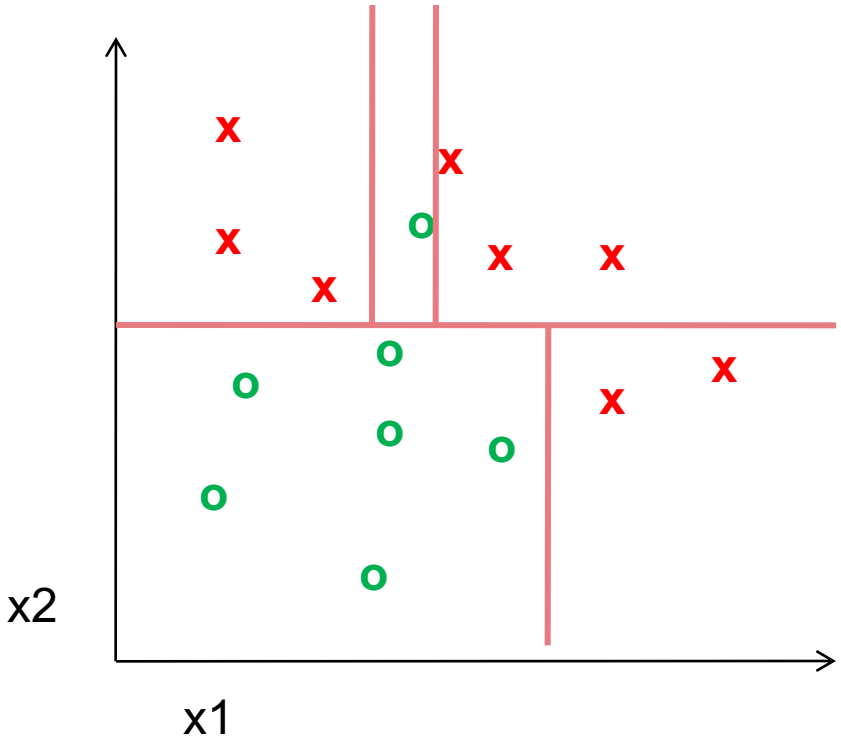


Training: given a *training set* of labeled examples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

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

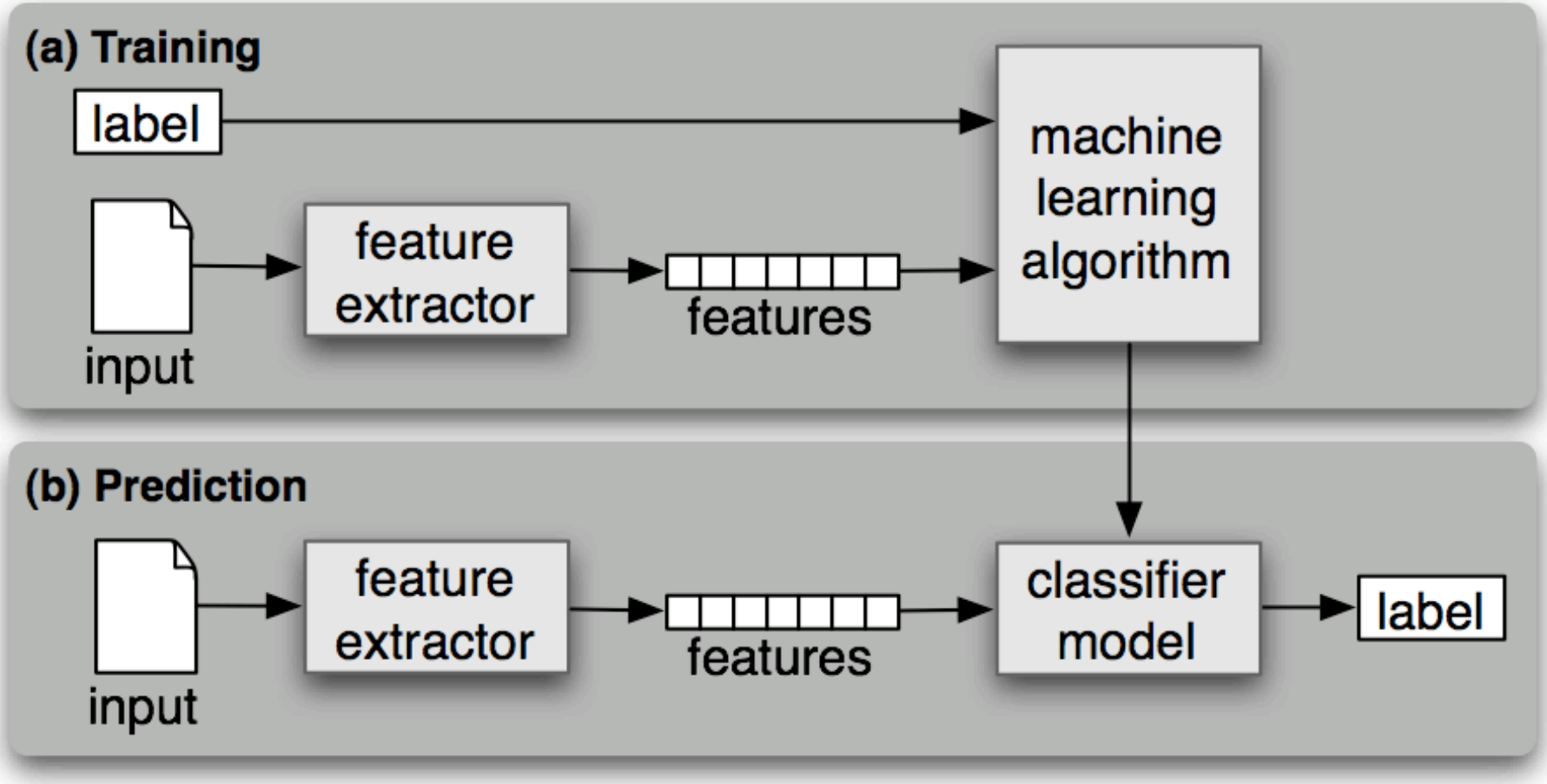
Evaluation

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

Optimization

- Build decision tree that maximize accuracy

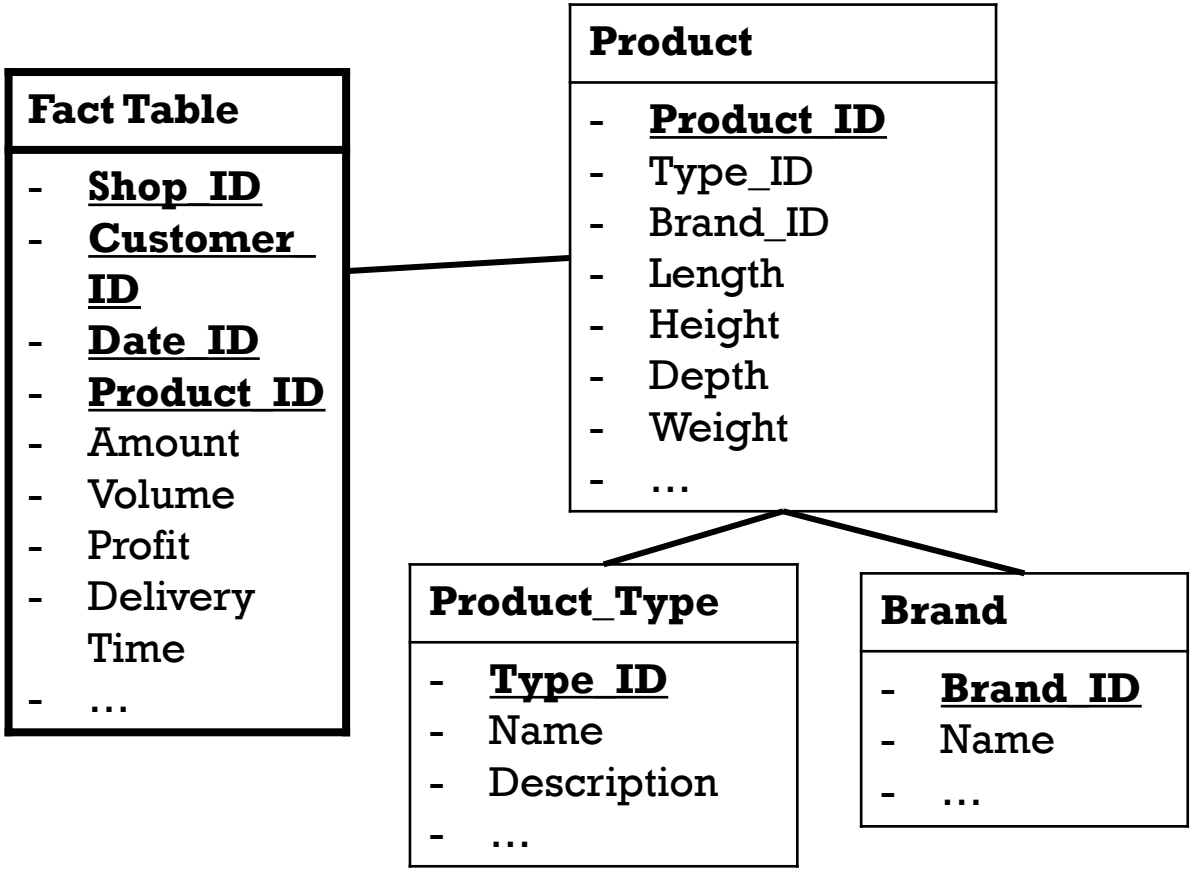
ML PIPELINE (SUPERVISED)



FEATURES

Fact Table

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

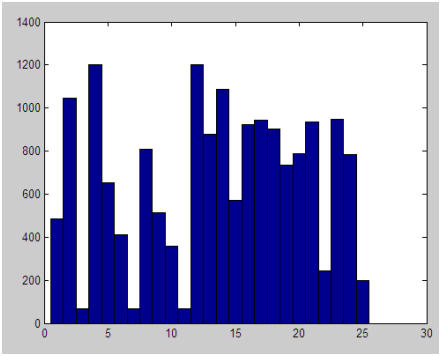


Custerme r State	Product Type	Product Weight	Volume (L*H*D)	Month	Delivery Time

IMAGE FEATURES

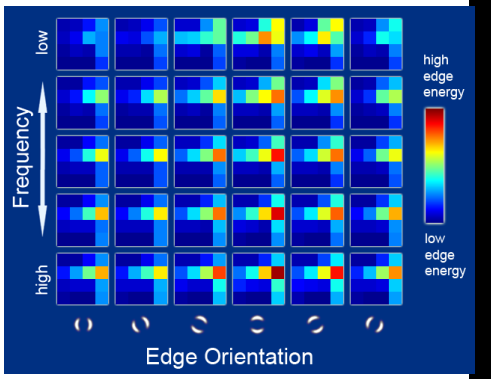
Raw pixels

Histograms



Geometric features – edge

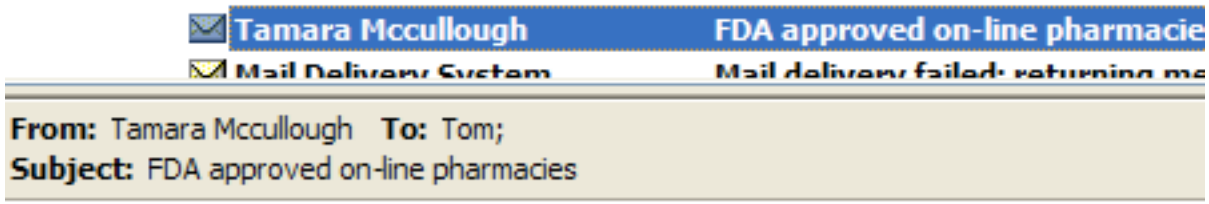
Corner,…



…

Slide credit: L. Lazebnik

TEXT FEATURES



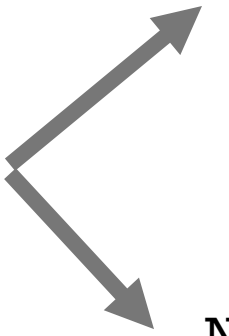
From: Tamara Mccullough **To:** Tom;
Subject: FDA approved on-line pharmacies

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, pro
dangerous pumps, exercises and surgeries.

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



Spam

Not Spam

Bag of Words

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

N-Grams

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