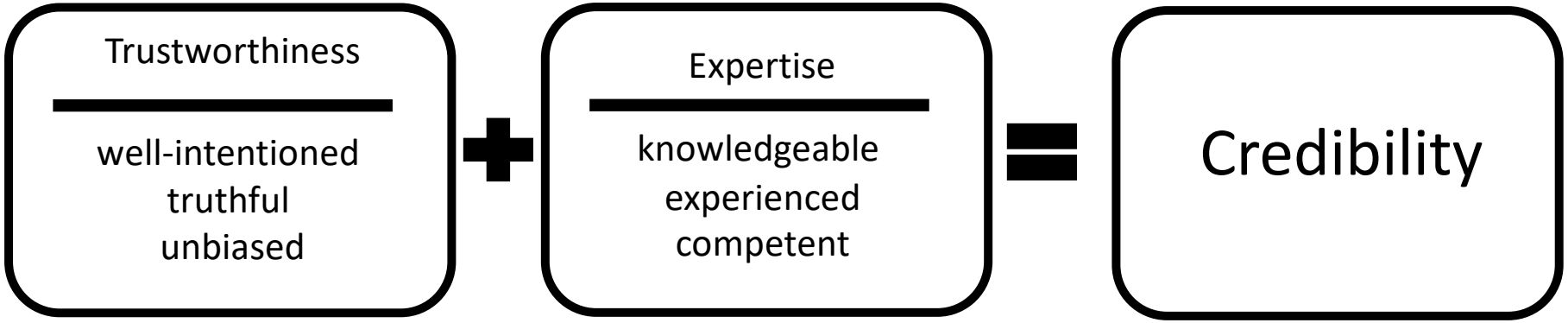
5. CLASSIFICATION METHODOLOGY

Example: Credibility Evaluation

Internet has become a primary data source

- Data is sometimes questionable, misleading or even erroneous
- Actions taken on the basis of incorrect data can have serious consequences



A classification problem!

Credible Page?

HealthDay
News for Healthier Living

search

Follow Us On

SIGN UP FOR OUR NEWSLETTER

Health Conditions | HealthDay TV | Wellness Library | HealthDay en Español | Physician's Briefing | License Our News

Alternative Therapies Widely Used for Autism

Study finds many parents use them alongside conventional treatments to try to manage symptoms

By Brenda Goodman
HealthDay Reporter

TUESDAY, Jan. 14, 2014 (HealthDay News) -- Nearly 40 percent of preschoolers with autism are getting some kind of complementary or alternative therapy for their condition, with nutritional supplements and special diets being the most common things parents try, a new study shows.

There are no medications currently approved specifically to treat autism spectrum disorders and its core symptoms of social and behavioral problems, according to the U.S. Centers for Disease Control and Prevention. Autism symptoms also include stomach upset and difficulty sleeping, among others.

So doctors and parents often rely on a variety of different, and sometimes unproven and unconventional, treatments to try to manage the wide variety of issues that can crop up.

Some experts had feared that parents might be turning to complementary therapies because they couldn't access recommended social or behavioral services or because they were trying to avoid conventional medicines, like vaccines.

But the study, published in the January issue of *Journal of Developmental and Behavioral Pediatrics*, found that wasn't the case.

Children in the study were 2 to 5 years old. Nearly all of the 453 children who had autism and another 125 with developmental disabilities were receiving the kinds of physical or behavioral therapy and other kinds of social services that are typically advised to help manage the condition. Many were also taking some kind of conventional medication for

HealthDay Video

As the price of healthy foods go up, so do blood sugar levels, study finds.

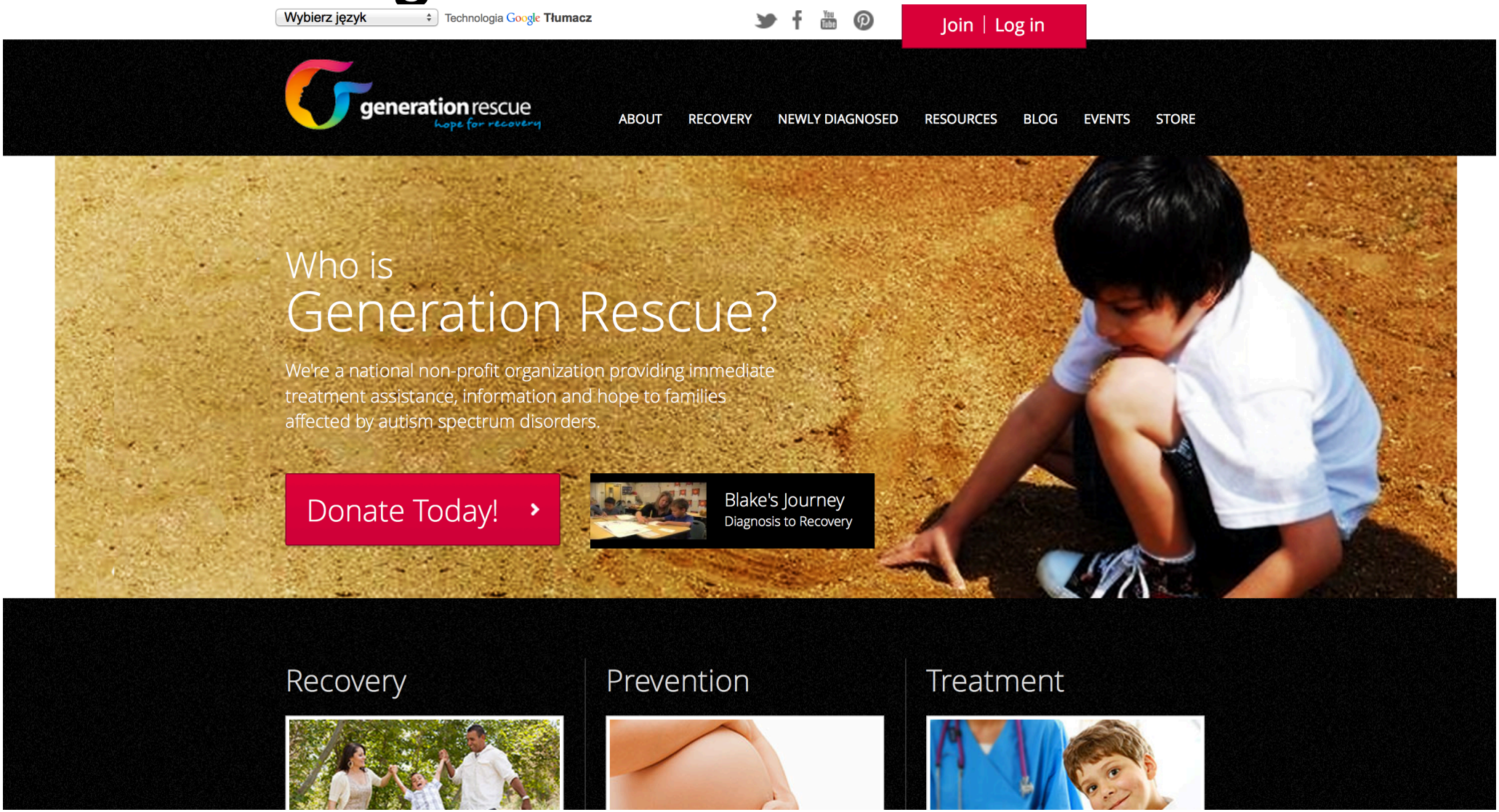
» watch this video

Advertisement

RELATED STORIES

- Video Games Might Help People With Dyslexia Learn to Read, Study Suggests
- Making Acupuncture Even Safer
- Preemie Birth Linked to Higher Insulin Levels in

Credible Page?



Classification Pipeline

Performing a data analysis project, such as for credibility evaluation …

… is more than knowing and using a classification algorithm

Main steps

1. Data collection and preparation
   - Domain knowledge and understanding essential
2. Model training, selection and assessment
   - Understanding potential and limits of machine learning essential

Data Collection and Preparation

1. Feature identification
2. Labelling
3. Discretization
4. Feature selection
5. Feature normalization

Model Training

1. Selecting performance metrics
2. Model selection
3. Organizing training and test data

DATA COLLECTION AND PREPARATION

1. Feature Identification

The first step is collecting data related to the classification task

- Definition of the attributes (or features) that describe a data item and the class label

Domain knowledge is needed

Feature Identification

Content

- Text and Images
- Appearance (design, ads)
- Domain type
- Sentiment (subjective, objective)
- #typos, #nouns, #words

Social

- #FB likes
- #G+ +1
- #tweets
- pagerank

Features

Different types of features

- Numerical (e.g., age, temperature ...)
- Ordinal (e.g., phone code ...)
- Categorical (e.g., student, weather ...)

Some classifiers require categorical features

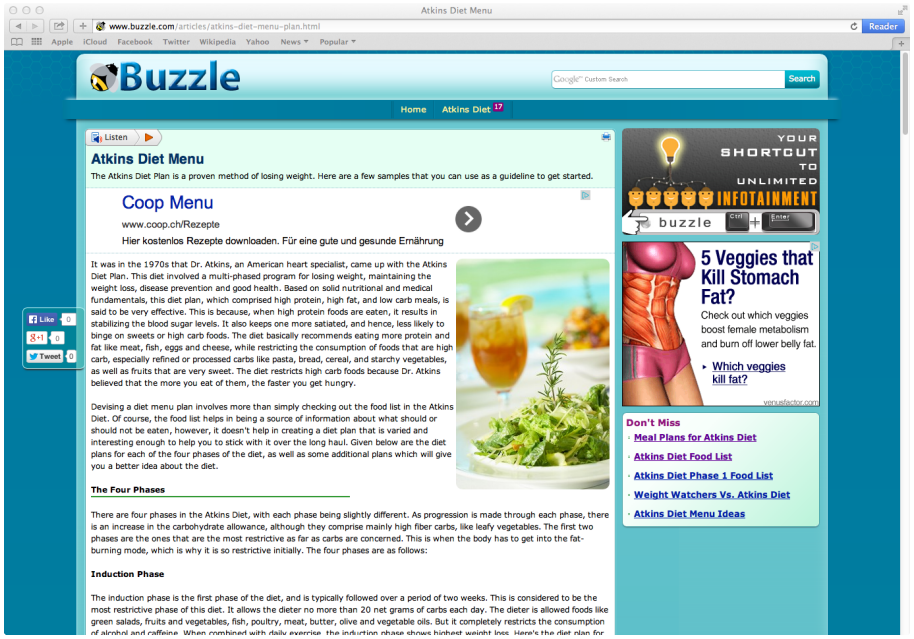
- Discretisation

2. Labelling

Collecting lot of data is easy

Labelling data is time consuming, expensive, difficult and sometimes even impossible

Expert in diets is needed



How to Obtain Labels?

1. Ask experts or do it yourself
   - Expensive, boring, low volume
2. Ask the crowd (crowd-sourcing)
   - Less expensive, popular, unreliable
3. Find some complementary information sources (distant learning)
   - Works in some cases, but …

Crowdsourcing

Is this webpage credible?

1. Submit task

Requester

4. Collect answers

C C ¬C C ¬C

CrowdFlower

clickworker.com

Crowdsourcing platform

amazon mechanical turk™
Artificial Artificial Intelligence

Crowd (workers)

2. Accept task

3. Return answers

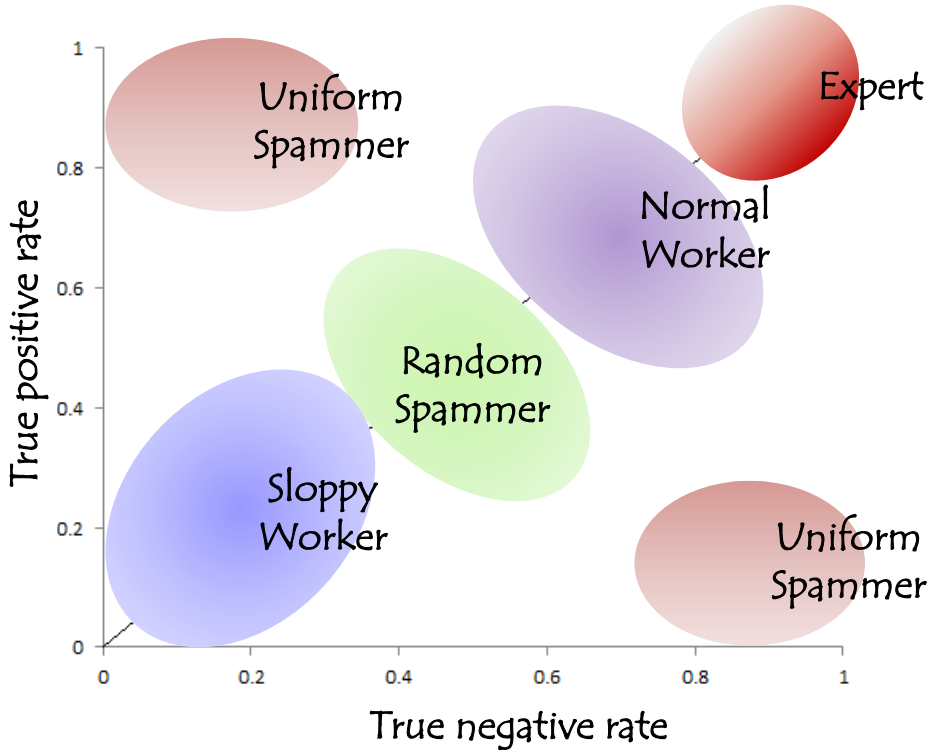
Different Types of Crowd-Workers

Truthful

- Expert
- Normal

Untruthful

- Sloppy
- Uniform spammer
- Random spammer



Answer Aggregation Problem

Crowd (workers)

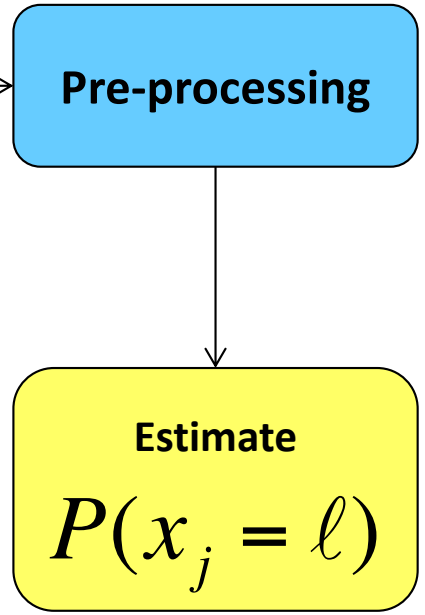
Worker	Webpage	Credible
W_1	www.diet.com	C
W_2	www.diet.com	¬C
W_3	www.diet.com	C
...	...	...

Aggregation

www.diet.com	C

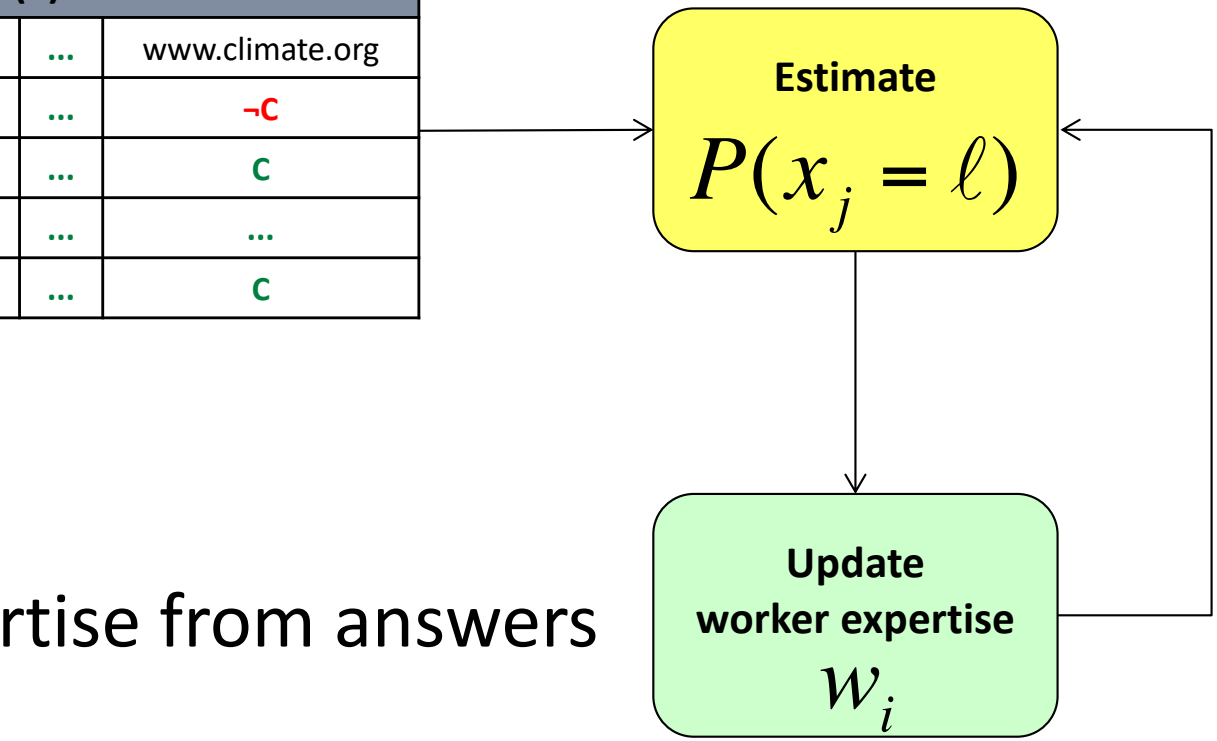
Non-Iterative Aggregation Algorithms

		Webpages (*x*)			
		www.diet.com	www.health.com	...	www.climate.org
Workers	W_1	C	C	...	¬C
	W_2	¬C	C	...	C
	...	...	...	...	...
	W_N	¬C	¬C	...	C



Iterative Aggregation Algorithms

		Webpages (*x*)			
		www.diet.com	www.health.com	...	www.climate.org
Workers	W_1	C	C	...	¬C
	W_2	¬C	C	...	C
	...	...	...	...	...
	W_N	¬C	¬C	...	C



Estimate worker expertise from answers

Majority Decision (MD)

Non-iterative aggregation algorithm

- No pre-processing step
- Estimate $P(x_j = \ell)$ as

$$P(x_j = \ell) = \frac{1}{N} \sum_{i=1}^{N} (1 \mid a_i(x_j) = \ell)$$

x_j webpage to label

N number of workers

ℓ label

$a_i(x_j)$ answer of worker *i* to webpage x_j

Honey Pot (HP)

Non-iterative aggregation algorithm

- Pre-processing step
  - Insert webpages $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$ for which the labels $\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_k$ are known
  - Remove workers and corresponding answers that fail at correctly labelling more than *m%* of webpages (either spammer or sloppy worker)
- Same decision rule as MD

Expectation Maximisation (EM)

Iterative aggregation algorithm

Iterates in two steps

1. E-Step: estimate the labels from the answers of workers
2. M-Step: estimate the reliability of workers from the consistency of answers

Expectation Maximisation (EM)

(E) step: estimate $P(x_j = \ell)$ as

$$P(x_j = \ell) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} (w_i \mid a_i(x_j) = \ell)$$

(M) step: update the expertise w_i as

$$w_i = \frac{1}{M} \sum_{j=1}^{M} (1 \mid a_i(x_j) = \arg\max_{\ell} P(x_j = \ell))$$

w_i expertise of worker *i*

M number of webpages to label

Expectation Maximisation (EM)

		Webpages (x)					w_i
		diet.com	health.com	climate.nasa.gov	climate.org	climatechange.net	
Workers (w)	W1	1	1	1	0	1	1
	W2	0	1	1	1	0	1
	W3	0	1	1	1	0	1
	W4	0	0	1	1	1	1

(E) step:

$P(x_j = \ell)$

P(x=0)	0.75	0.25	0	0.25	0.5
P(x=1)	0.25	0.75	1	0.75	0.5

$$P(x_j = \ell) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} (w_i \mid a_i(x_j) = \ell)$$

Expectation Maximisation (EM)

(M) step:

		Webpages (x)					
		diet.com	health.com	climate.nasa.gov	climate.org	climatechange.net	w_i
Workers (w)	W1	1	1	1	0	1	2/5
	W2	0	1	1	1	0	5/5
	W3	0	1	1	1	0	5/5
	W4	0	0	1	1	1	3/5

(E) step:

$P(x_j = \ell)$

P(x=0)	0.75	0.25	0	0.25	0.5
P(x=1)	0.25	0.75	1	0.75	0.5

$$w_i = \frac{1}{M}\sum_{j=1}^{M}(1 \mid a_i(x_j) = \arg\max_{\ell} P(x_j = \ell))$$

Expectation Maximisation (EM)

(M) step:

		Webpages (x)					
		diet.com	health.com	climate.nasa.gov	climate.org	climatechange.net	w_i
Workers (w)	W1	1	1	1	0	1	2/5
	W2	0	1	1	1	0	5/5
	W3	0	1	1	1	0	5/5
	W4	0	0	1	1	1	3/5

(E) step:

$P(x_j = \ell)$

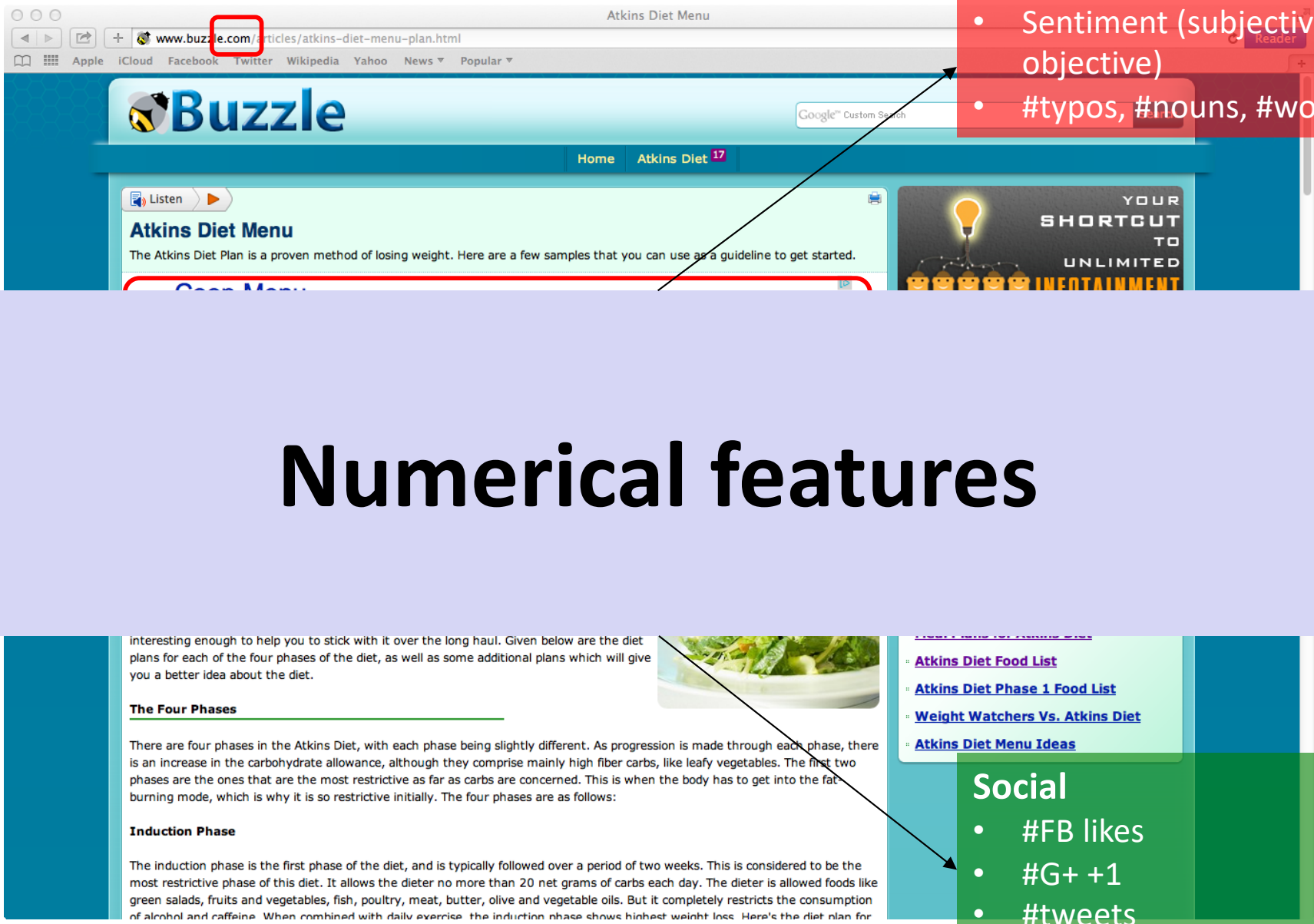
P(x=0)	0.867	0.2	0	0.133	0.667
P(x=1)	0.133	0.8	1	0.867	0.333

$$P(x_j = \ell) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} (w_i \mid a_i(x_j) = \ell)$$

Which data mining algorithm belongs to the Expectation-Maximization class?

A. Apriori
B. K-Means
C. Decision tree
D. None of the above

3. Discretization



Content

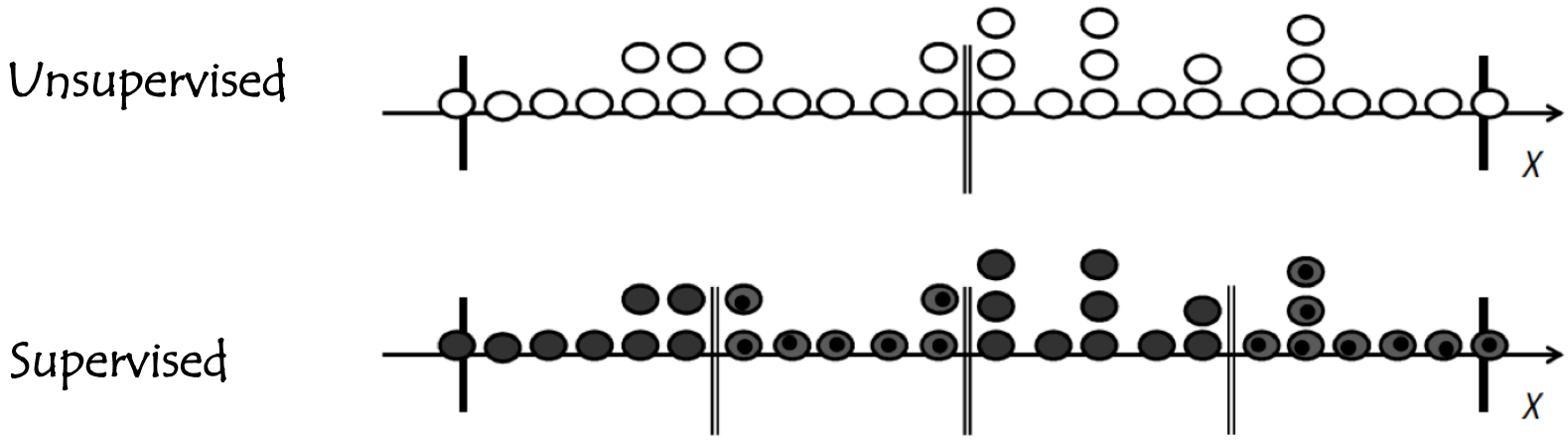
- Text and Images
- Appearance (design, ads)
- Domain type
- Sentiment (subjective, objective)
- #typos, #nouns, #words

Numerical features

Social

- #FB likes
- #G+ +1
- #tweets
- pagerank

Discretisation Methods

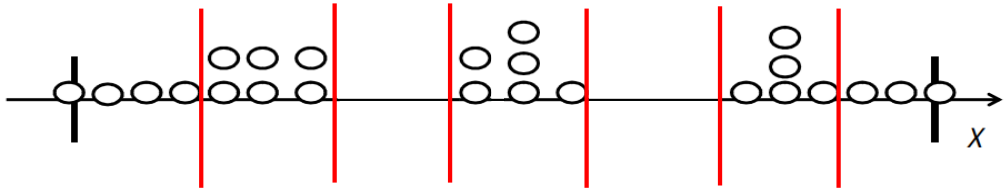


X is a numerical attribute

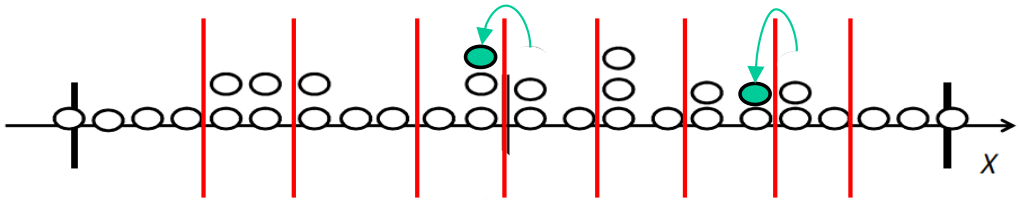
- Unsupervised case: no class information ➔ bins might confuse data from different clases
- Supervised case: discretization follows class boundaries

Unsupervised Discretisation

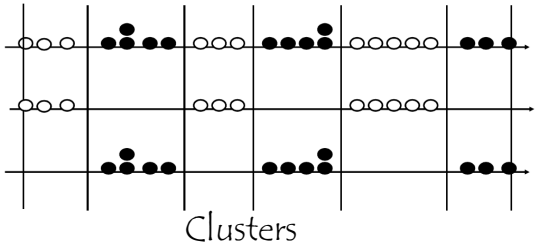
Equal width: divide the range into a predefined number of bins



Equal frequency: divide the range into a predefined number of bins so that every interval contains the same number of values



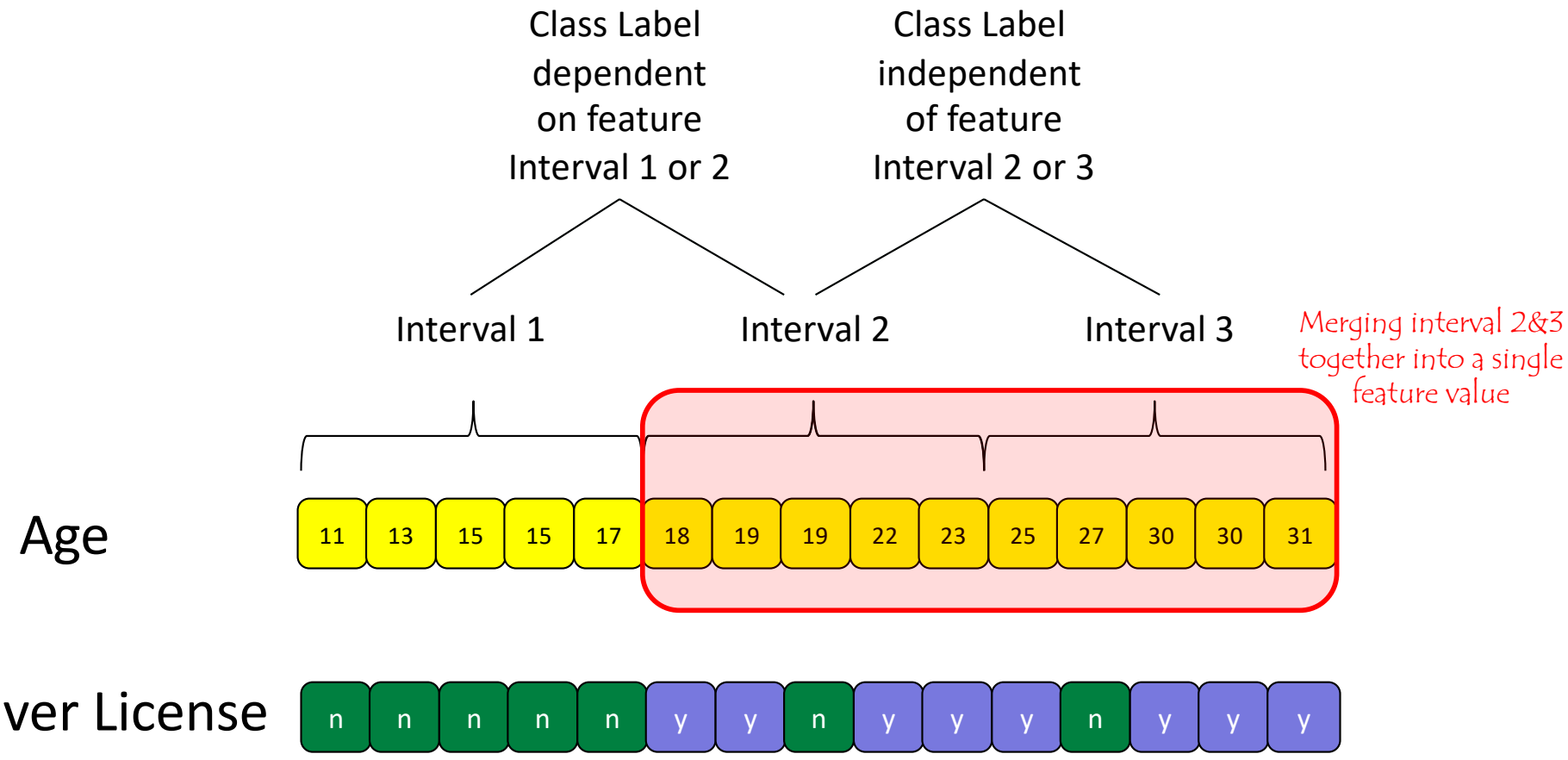
Clustering: Use any suitable clustering method for multi-dimensional data and assign one feature value per cluster



Multiple numerical attributes

Supervised Discretisation

Idea: if the class label does not depend on the choice among two (adjacent) intervals,
the separation of the intervals does not provide useful information to the classifier



Supervised Discretisation

Independence test: χ^2 statistics

	Class1	Class2	*sum*
Interval 1	$O_{11} = n_{11}$	$O_{12} = n_{12}$	R1
Interval 2	$O_{21} = n_{21}$	$O_{22} = n_{22}$	R2
sum	C1	C2	N

O_{ij} observed frequency
E_{ij} expected frequency

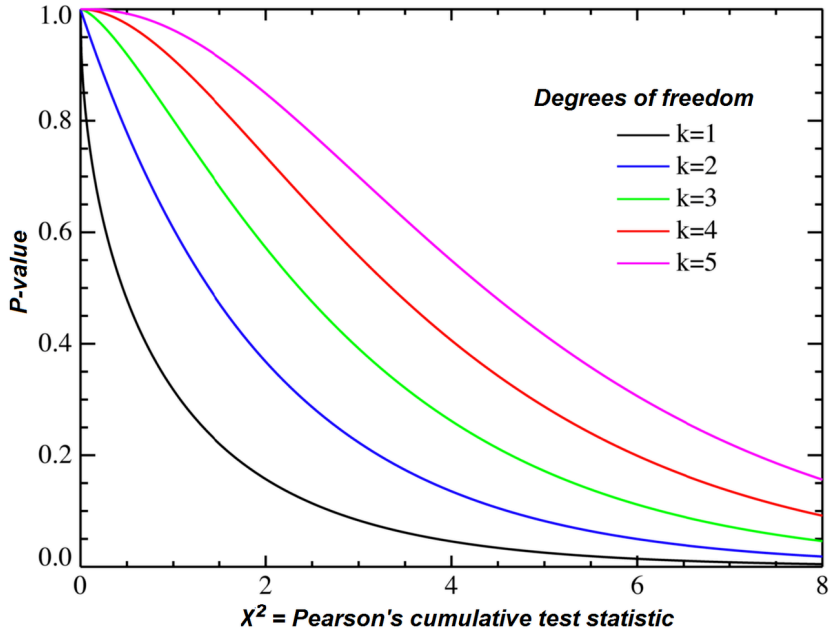
$$\chi^2 = \sum_{i=1,2} \sum_{j=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Supervised Discretisation

Null hypothesis: Assumes that the class label is independent of the feature intervals

If $P(\chi^2 \mid DF = 1) > 0.05$ (independent), merge the intervals

DF = degrees of freedom = (#rows-1)*(#cols



Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Example

Driving License?

observed	No	Yes	SUM
Interval 1	51	0	51
Interval 2	1	50	51
SUM	52	50	102
expected	No	Yes	
Interval 1	26.00	25.00	
Interval 2	26.00	25.00	
chisquare statistics	No	Yes	
Interval 1	24.04	25.00	
Interval 2	24.04	25.00	
chisquare	**98.0769231**		
P(chi | df = 1)	**4.0244E-23**		

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63

p-value of 4.0244e-23 is less than 0.05 → we reject the null hypothesis
there is a dependence → no merge!

Example

Driving License?

observed	No	Yes	SUM
Interval 2	49	51	100
Interval 3	50	51	101
SUM	99	102	201
expected	No	Yes	
Interval 2	49.25	50.75	
Interval 3	49.75	51.25	
chisquare statistics	No	Yes	
Interval 2	0.00131	0.00127	
Interval 3	0.00129	0.00126	
chisquare	**0.00512601**		
P(chi | df = 1)	**0.94292328**		

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63

p-value of 0.9429 is greater than 0.05 → we accept the null hypothesis
they are independent → we can merge

Supervised discretization …

A. Can only be applied after unsupervised equal-width discretization
B. Attempts to merge intervals where the class label does not depend on the attribute value distribution
C. Merges intervals if the value of the χ^2 statistics is above 0.05
D. Works only for binary class labels

4. Feature Selection

Content
- Text and Images
- Appearance (design, ads)
- Domain type
- Sentiment (subjective, objective)
- #typos, #nouns, #words

Are all these features relevant?

Social
- #FB likes
- #G+ +1
- #tweets
- pagerank

Feature Selection

Reducing the number of N features to an optimal subset of M features, M < N

There are $\binom{N}{M}$ possible subsets

Approaches

- Filtering: consider features as independent
- Wrapper: consider dependencies among features

Feature Selection: Filtering

Filtering: rank features according to their predictive power and select the best ones

Pros

- Independent of the classifier (performed only once)

Cons

- Independent of the classifier (ignore interaction with the classifier)
- Assumes features are independent

Ranking Features

χ^2 statistics (as before) for the n feature values

	Class c_1	Class c_2	*sum*
Value f_1			
Value f_2			
...			
Value f_n			
sum			

$P(\chi^2 \mid DF = n - 1)$ gives a rank measure

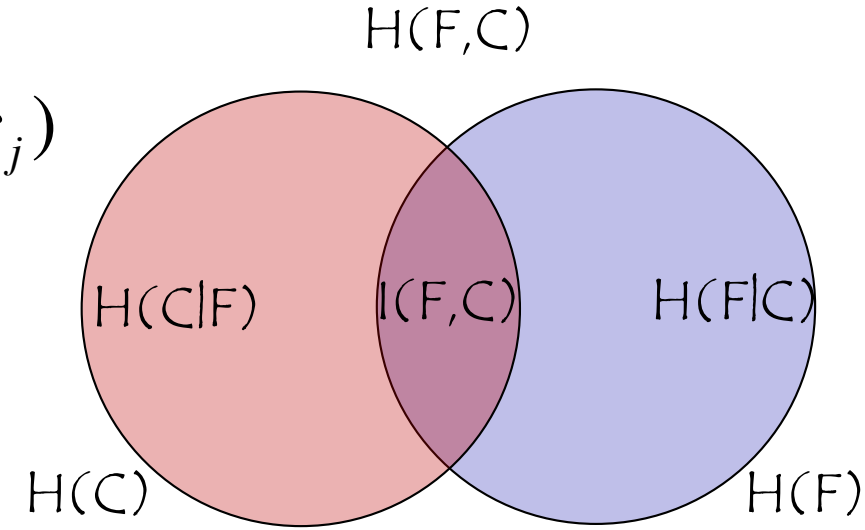
Information-theoretic approach

Mutual information between feature F and class label C

$$I(F;C) = H(C) - H(C \mid F) = H(F) + H(C) - H(F,C)$$

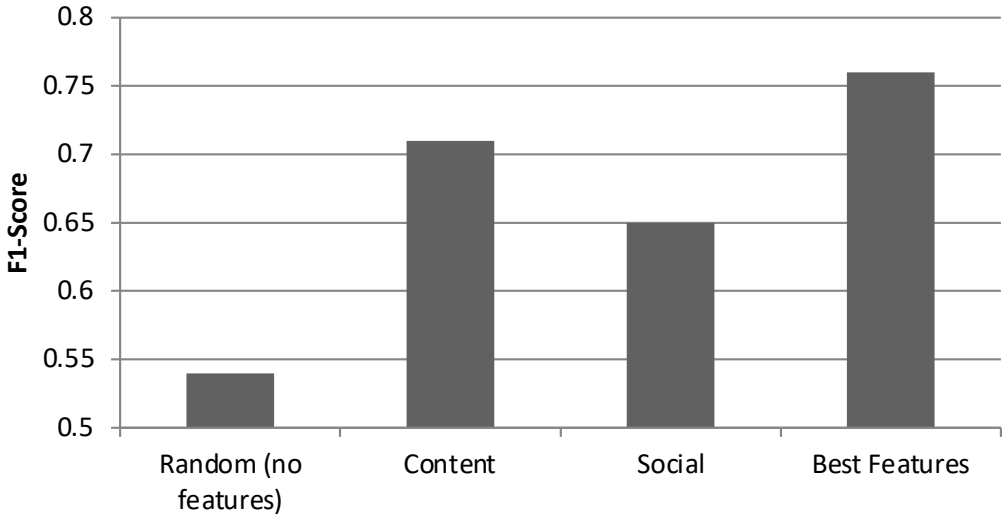
$$H(F) = -\sum_i P(f_i) \log_2 P(f_i)$$

$$H(F,C) = -\sum_i \sum_j P(f_i, c_j) \log_2 P(f_i, c_j)$$



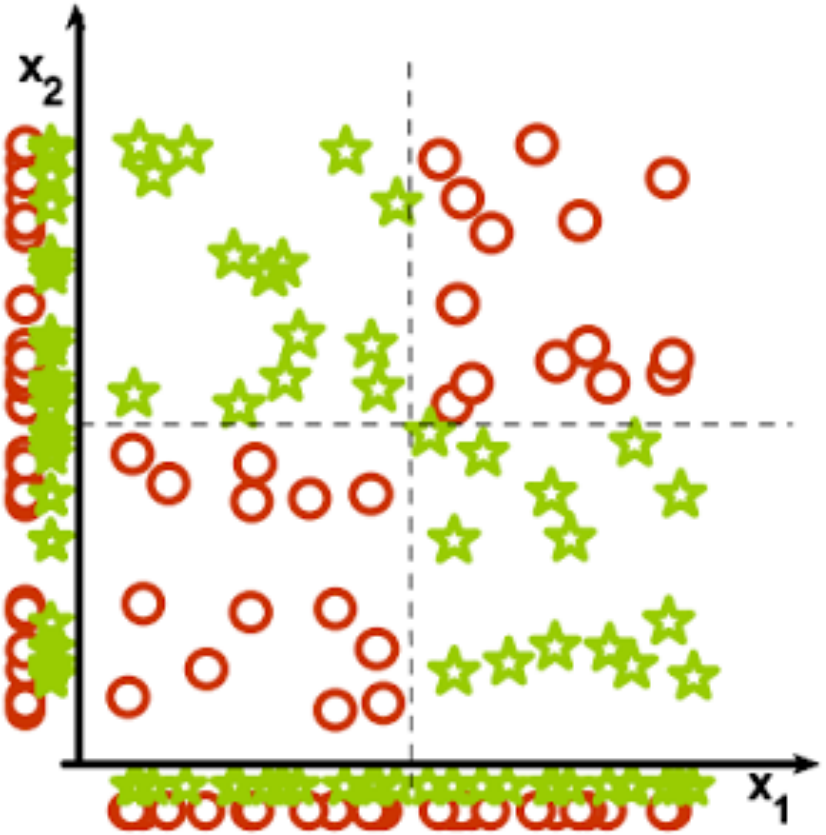
Small I(F,C) ➔ Feature has little predictive power!

Example: Credibility Features



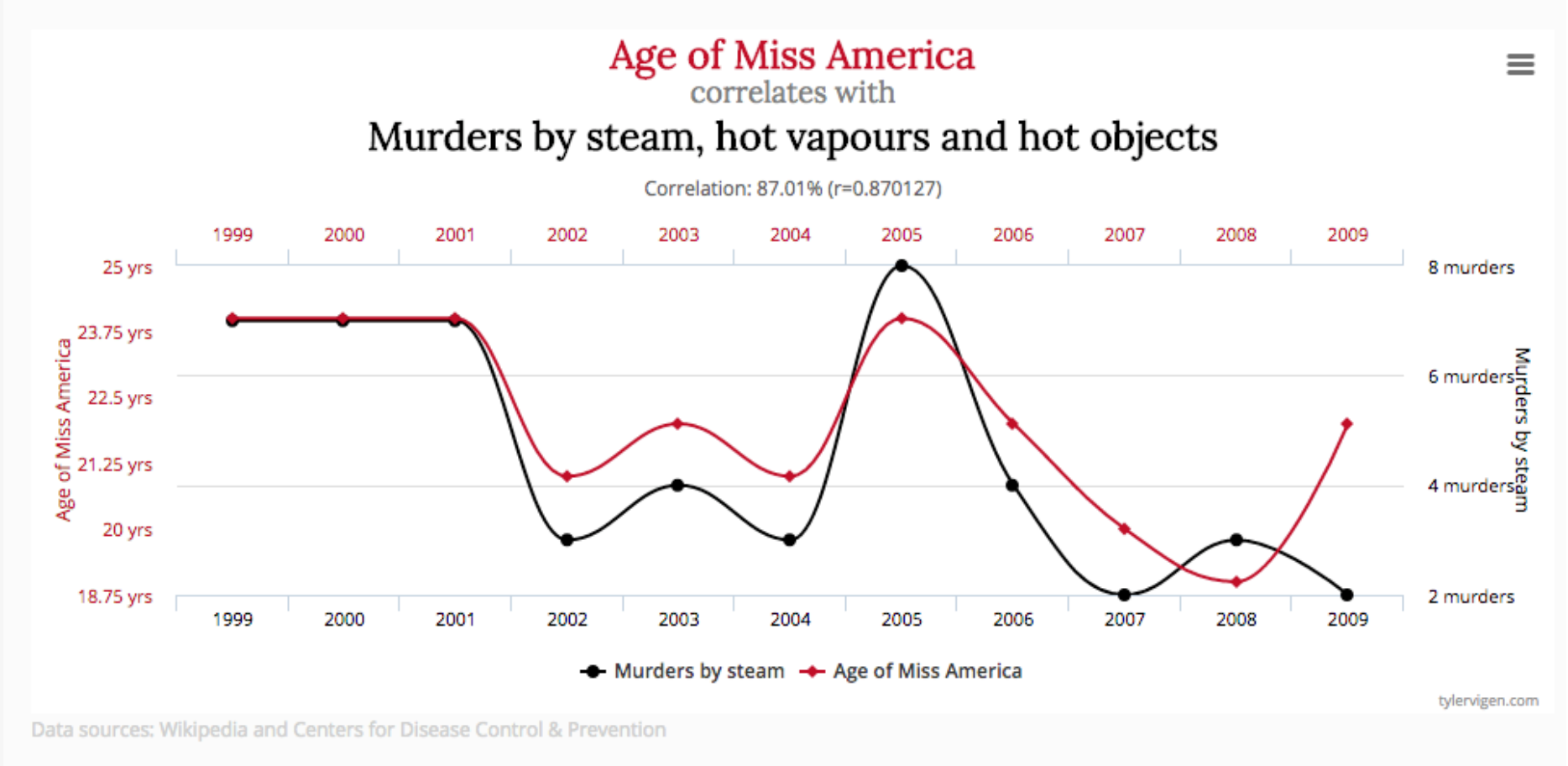
Selection of Features - Pitfalls

Collectively relevant features may look individually irrelevant



Selection of Features - Pitfalls

Correlation ≠ Causality



Many more examples: http://www.tylervigen.com/spurious-correlations

Selection of Features - Pitfalls

Beware of trusting correlations in a blind way

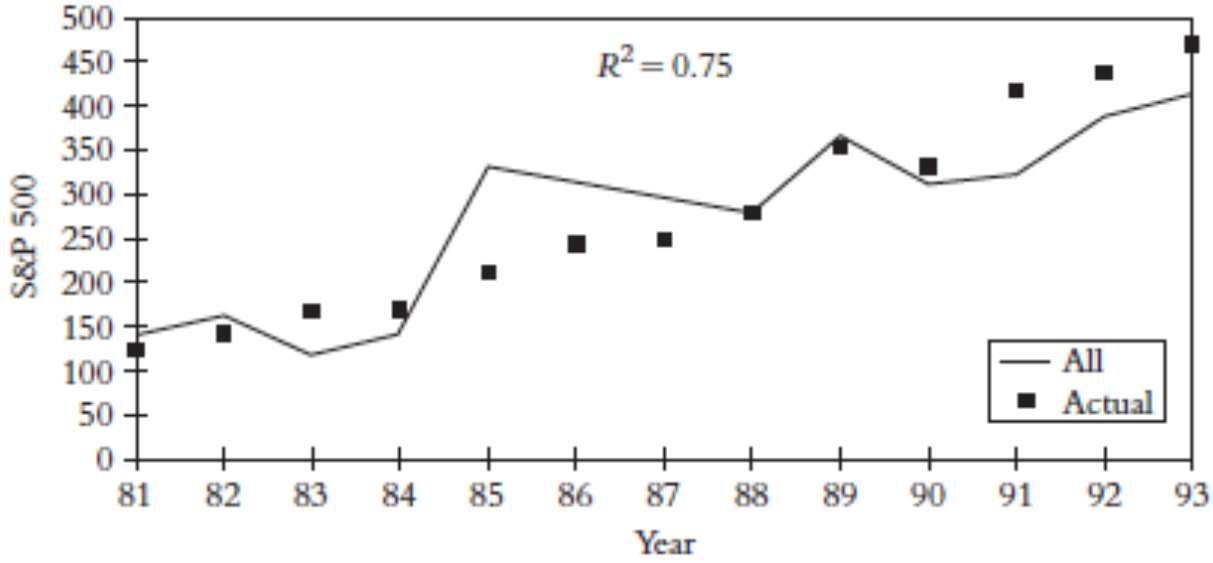


Figure 6.2 Overfitting the S&P 500: butter production in Bangladesh—a single variable that "explains" 75 percent of the S&P's returns.

http://m.shookrun.com/documents/stupidmining.pdf

Selection of Features - Pitfalls

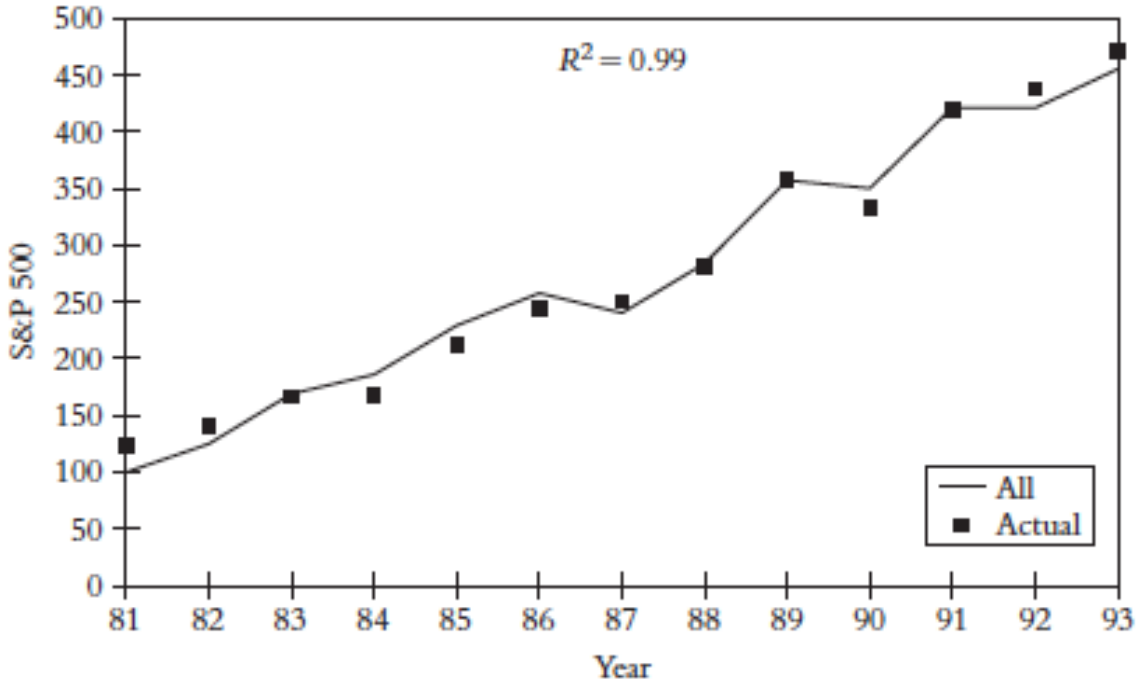


Figure 6.4 Overfitting the S&P 500: butter in Bangladesh and United States, plus U.S. cheese production, as well as sheep population in Bangladesh and United States. Now we're at 99 percent. You can do this as long as you can find data not perfectly correlated with butter, cheese, sheep, and so on. There is no shortage of that.

http://m.shookrun.com/documents/stupidmining.pdf

Feature Selection: Wrapping

Iteratively add features

- at each iteration create a classifier for each new feature and evaluate its performance
- Add best feature or stop when no further improvement

Pros

- Interact with the classifier
- No independence assumption

Cons

- Computationally intensive

The filtering approach to feature selection …

A. tests whether features are independent among each other
B. tests whether a feature is dependent on the class label
C. tests whether a feature is dependent on the classifier
D. eliminates strongly correlated features

5. Feature Normalisation

Some classifiers do not manage well features with very different scales

- # followers: 10,000,000
- # tweets: 300

Features with large values dominate the others, and the classifier tend to over-optimise them

Standardisation and Scaling

Standardisation: map to a normal distribution N(0, 1)

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}$$

- μ_i is the mean value of feature x_i
- σ_i is the standard deviation
- The new feature x_i' has mean 0 and standard deviation 1

Scaling: map to interval [0,1]

$$x_i' = \frac{x_i - m_i}{M_i - m_i}$$

- M_i and m_i are the maximal and minimal values of the feature x_i

Standardisation vs Scaling

Standardisation

- Assumes that the data has been generated by a Gaussian process (not necessarily true)

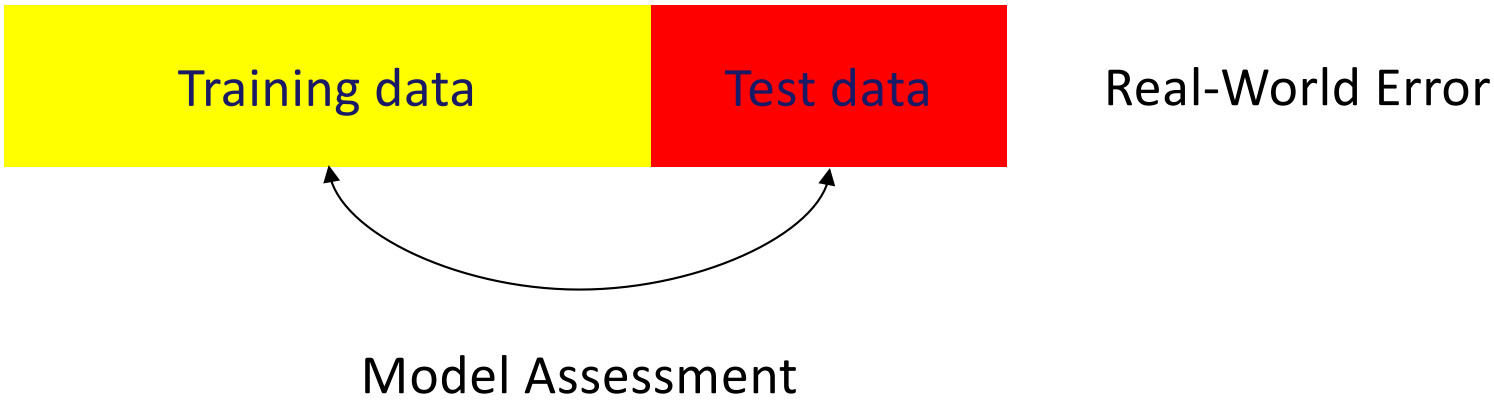
Scaling

- If the data has outliers, they scale the “normal” values to a very small interval

MODEL TRAINING, SELECTION AND ASSESSMENT

1. Choosing Performance Metrics

Model Assessment: Having chosen a model, estimate the prediction error on new data



Performance Metric for Binary Classification

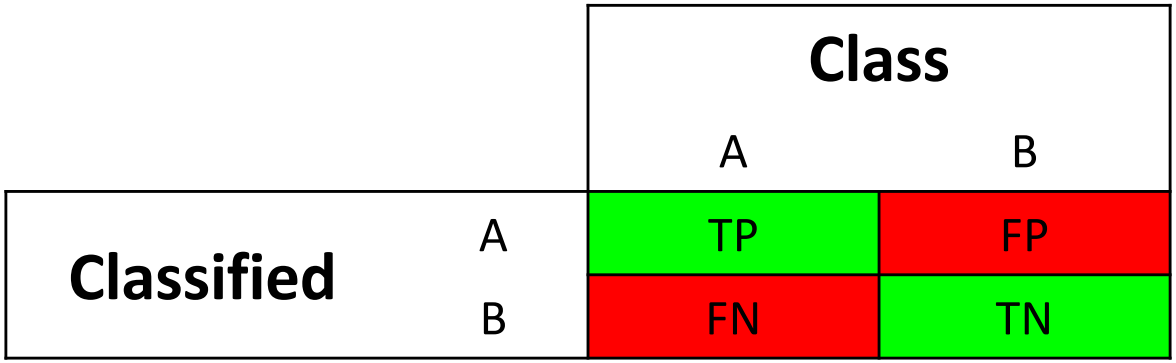
For categorical binary classification, the usual metrics consider four types of outcomes

Correct results

- True Positive (positive examples classified as positive)
- True Negative (negative examples classified as negative)

Incorrect results

- False Positive (negative examples classified as positive)
- False Negative (positive examples classified as negative)



Accuracy

$$A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$

Appropriate metric when

- Classes are not skewed
- Errors have the same importance

Accuracy - Pitfall

Classifier 1

		Class: Fraud	Class: ¬Fraud
Classified	Fraud	5	10
	¬Fraud	5	80

A = 85/100 = 0.85

Always ¬Fraud

		Class: Fraud	Class: ¬Fraud
Classified	Fraud	0	0
	¬Fraud	10	90

A = 90/100 = 0.90

Which is the “best” classifier?

Classifier 1		Class A	Class B
Classified	A	45	20
	B	5	30

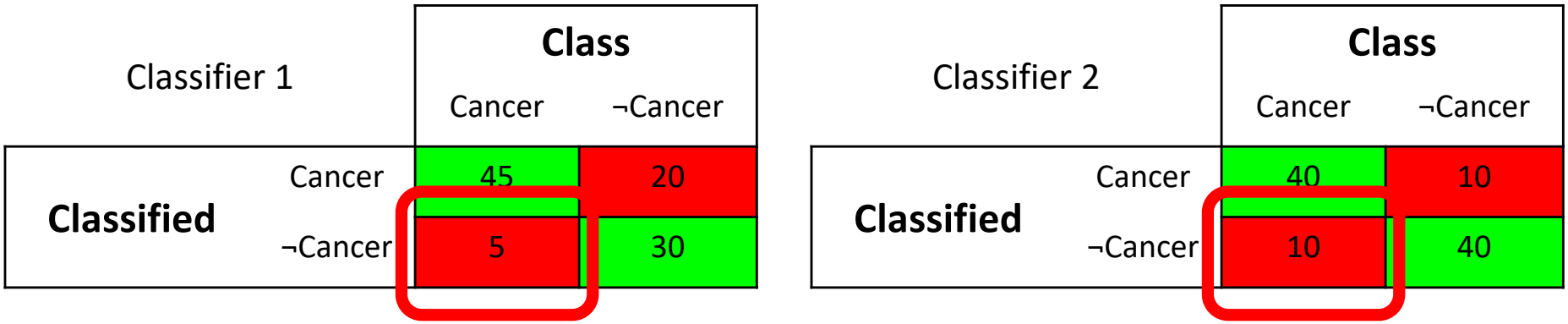
Classifier 2		Class A	Class B
Classified	A	40	10
	B	10	40

A. Classifier 1

B. Classifier 2

C. Both are equally good

Which is the “best” classifier?



A. Classifier 1

B. Classifier 2

C. Both are equally good

Precision and Recall

Precision

$$P = \frac{TP}{TP + FP}$$

Recall

$$R = \frac{TP}{TP + FN}$$

Precision and Recall: Example

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

$P_1=45/65=0.69$

$R_1=45/50=0.9$

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

$P_2=40/50=0.8$

$R_2=40/50=0.8$

Everybody has cancer		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$P = 50/100 = 0.5$

$R = 50/50 = 1$

F-score

Sometimes it’s necessary to have a unique metric to compare classifiers

F-score (or F1-score)

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

F-beta score

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P + R)}$$

F-Score: Example (beta = 1)

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

$F_1=2*(0.69*0.9)/(0.69+0.9)$

$= 0.78$

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

$F_2=2*(0.8*0.8)/(0.8+0.8)$

$=0.8$

Everybody has cancer		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$F=2*(0.5*1)/(0.5+1)=0.66$

F-beta-Score: Example (beta = 2)

Classifier 1		Class: Cancer	Class: ¬Cancer
Classified	Cancer	45	20
	¬Cancer	5	30

Classifier 2		Class: Cancer	Class: ¬Cancer
Classified	Cancer	40	10
	¬Cancer	10	40

$F_1=5*(0.69*0.9)/(4*0.69+0.9)$

$= 0.84$

$F_2=5*(0.8*0.8)/(4*0.8+0.8)$

$=0.8$

Everybody has cancer		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$F=5*(0.5*1)/(4*0.5+1)=0.83$

Considering Cost

The “cow case”

- Predict when cows are “in heat”
- Important for fertilization
- Observe different body parameters
- Predictor: never (correct 97% of the time!)

Solution: analyze different types of errors separately and associate (financial) cost and benefits

		Predicted class: yes	Predicted class: no
Actual class	yes	True positive: 30	False negative: -30
	no	False positive: -1	True negative: 1

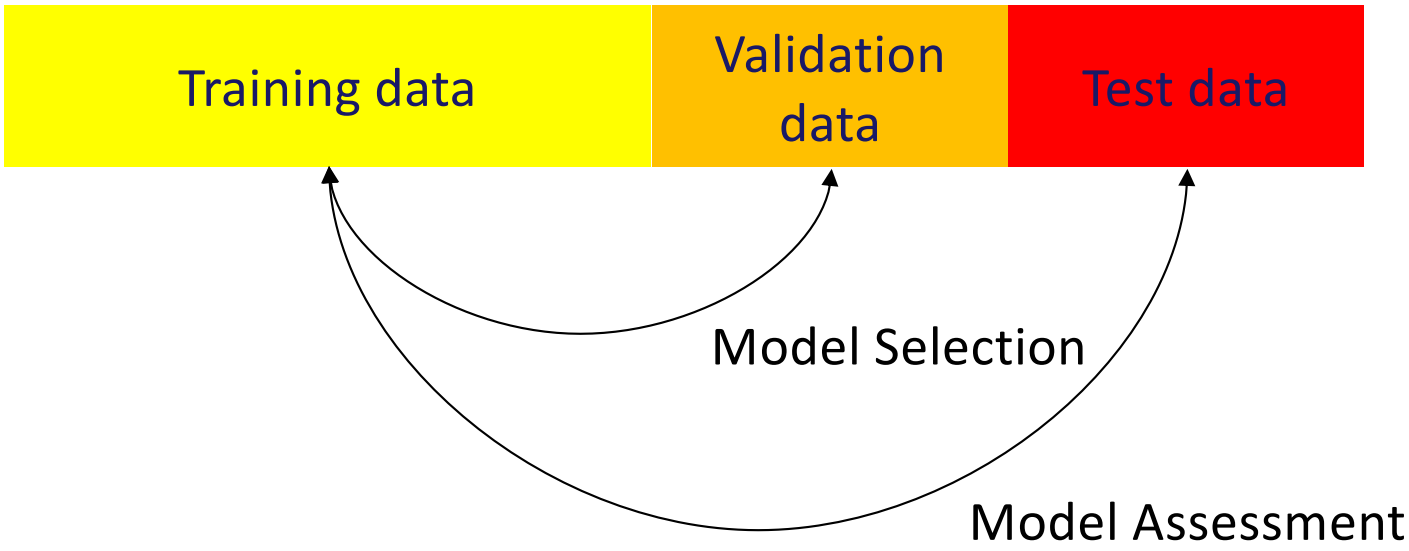
2. Model Selection

Usually a classifier has some parameters to be tuned

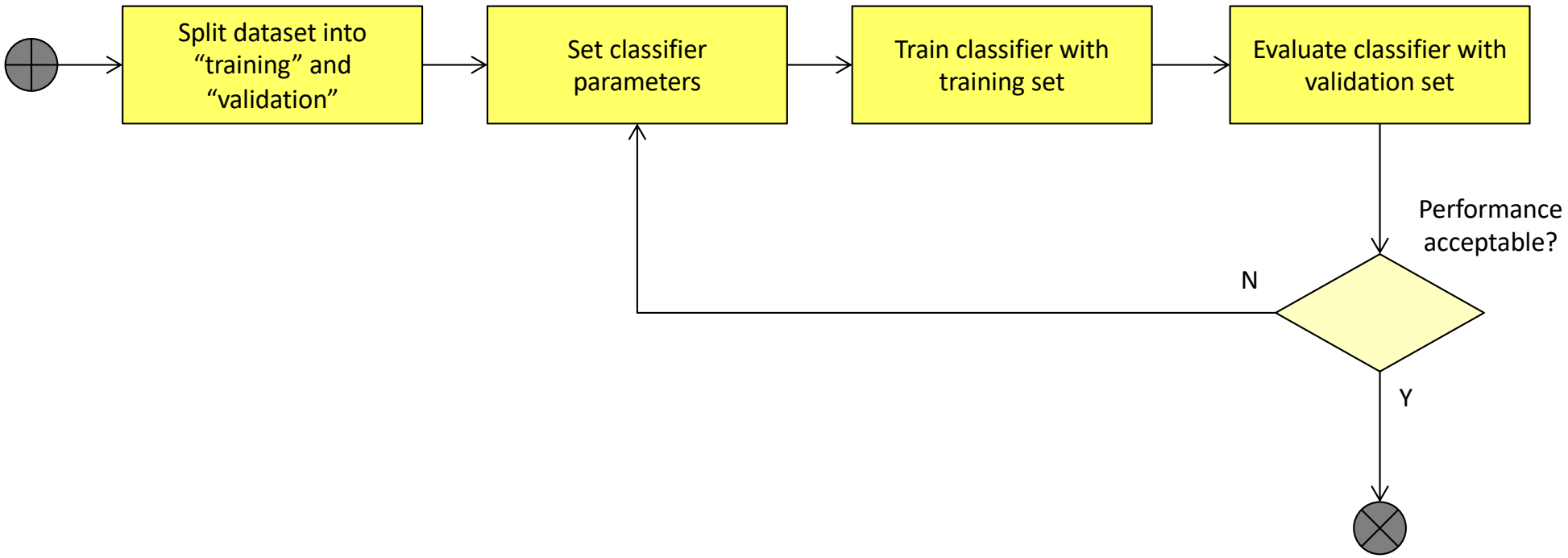
- Sampling strategy
- Regularisation factor
- Threshold
- Distance function
- Number of neighbours

Model Selection: Estimating performances of different models to select the best one

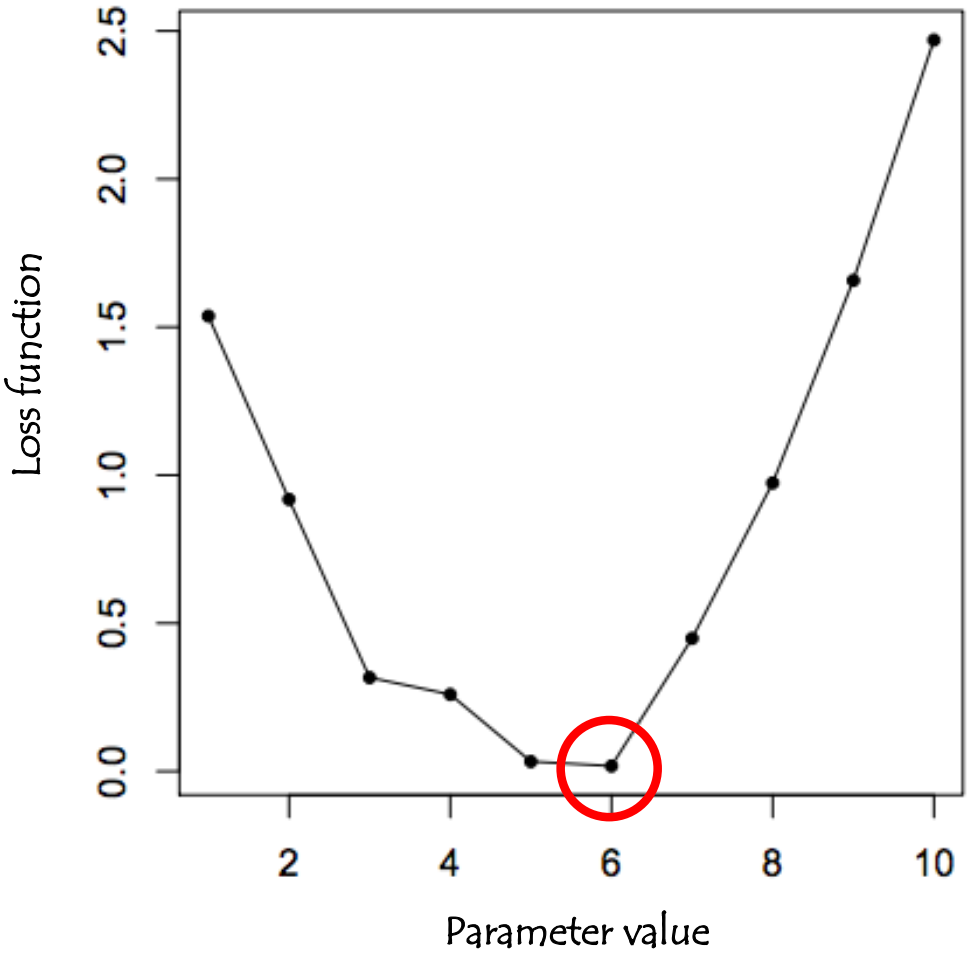
Model Selection



Model Selection



Model Selection



Loss Function (Error Function)

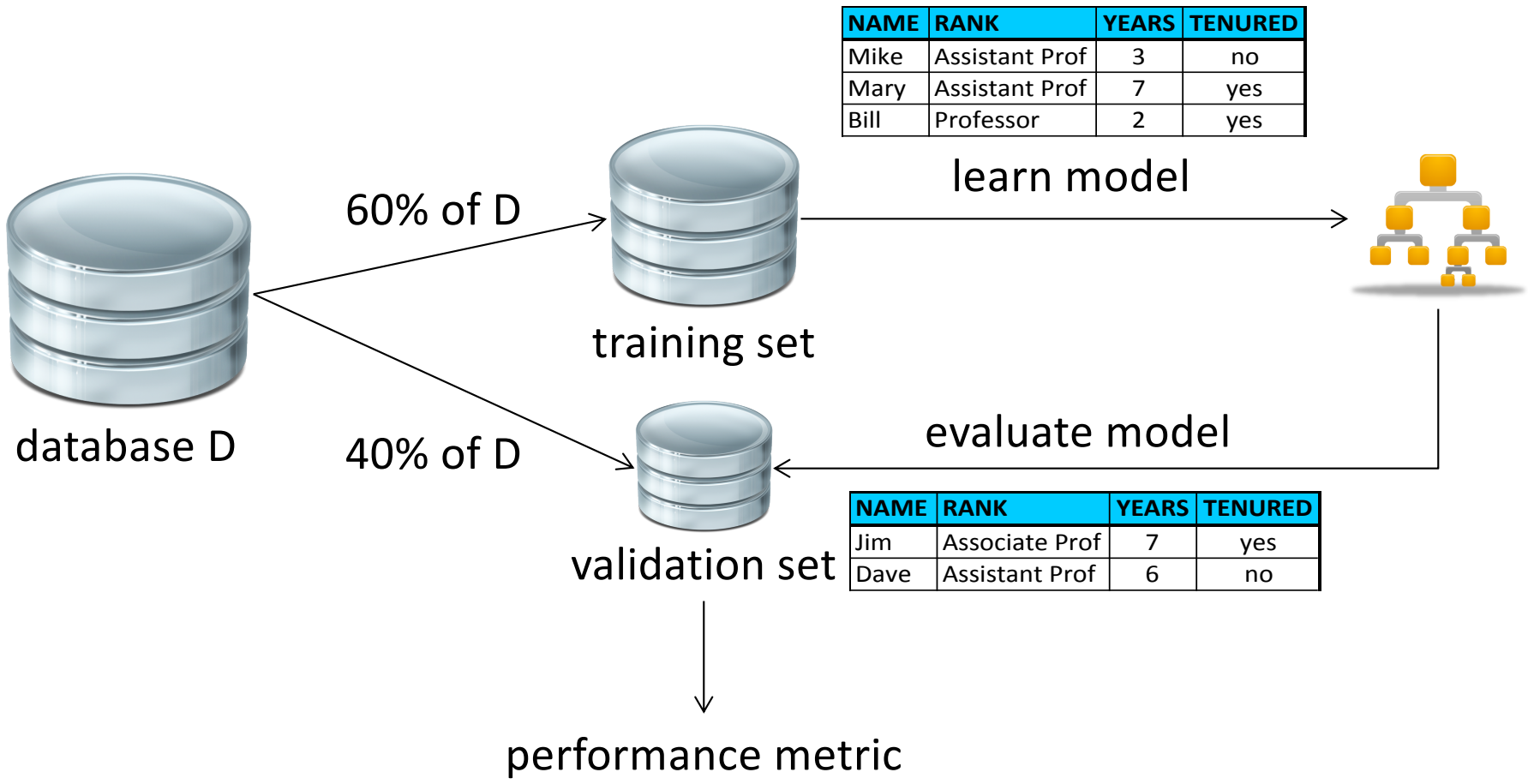
Categorical output

- 0-1 loss function: $J = \sum_{i=1}^{n} \#(y \neq f(x_i))$

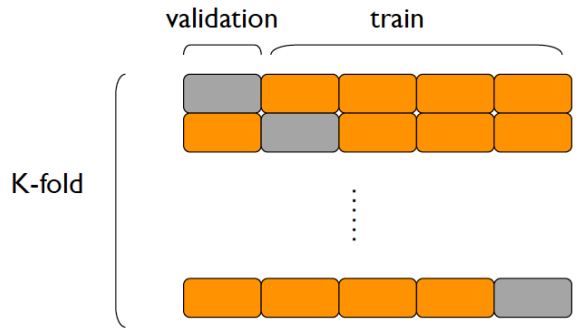
Real value output

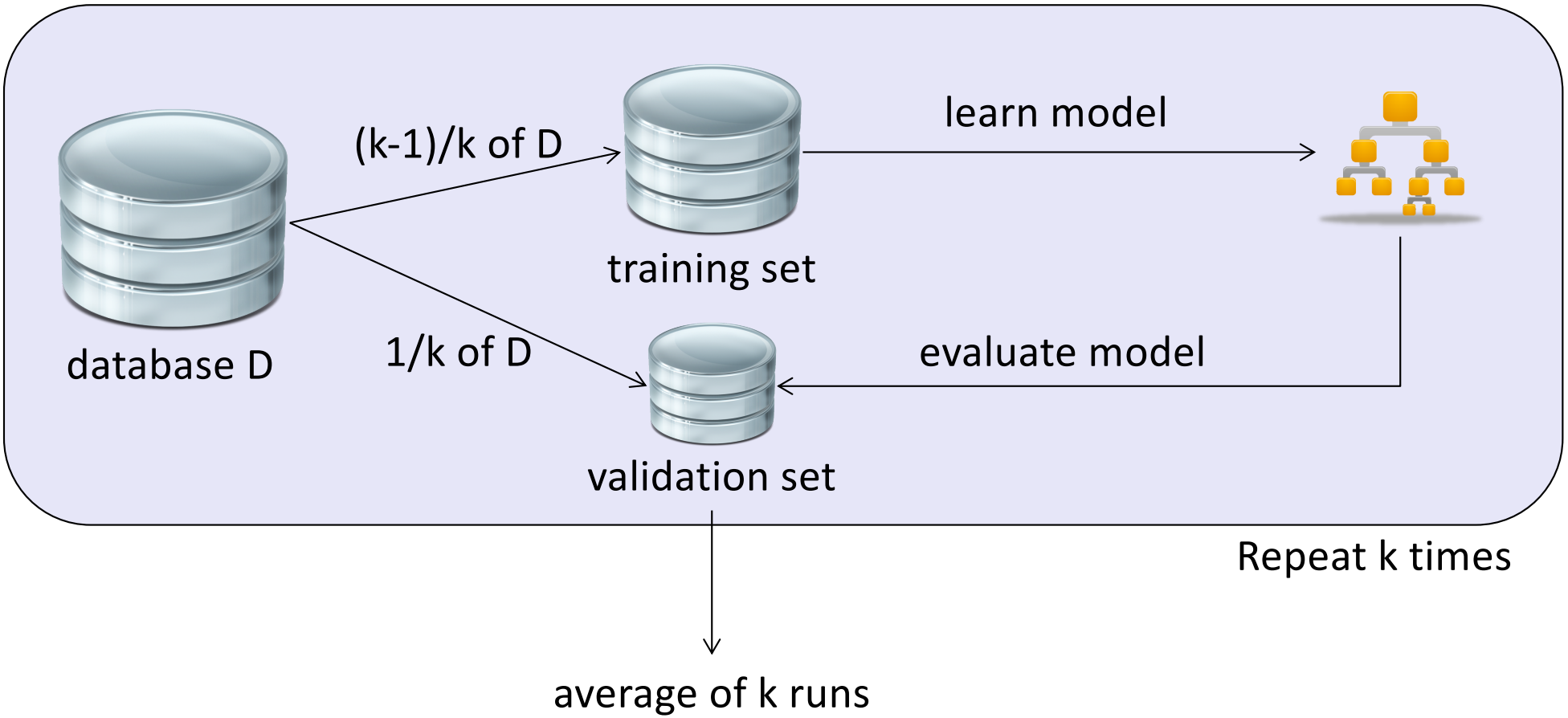
- Squared error: $J = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$
- Absolute error: $J = \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|$

3. Training and Validation Set

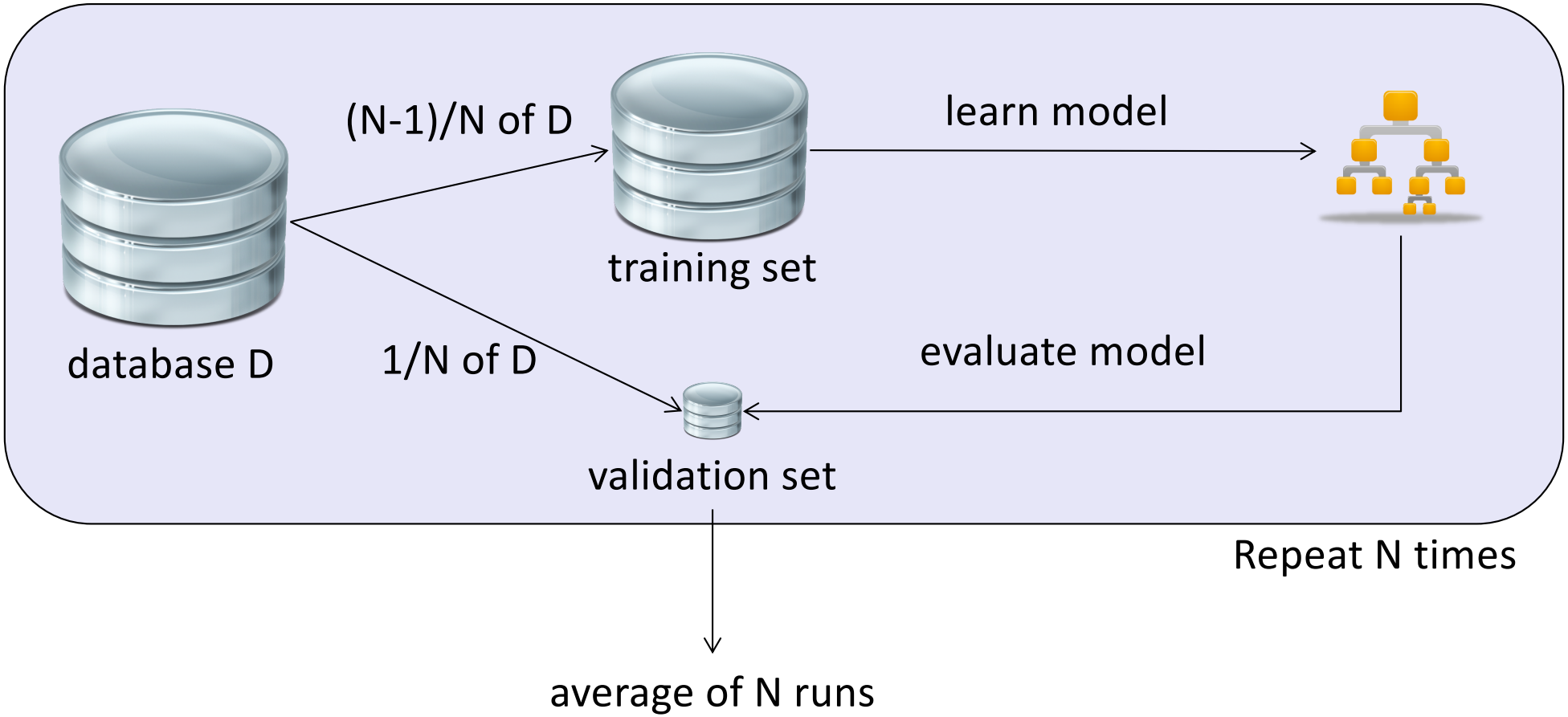


K-fold Cross Validation





Leave-one-out Cross Validation

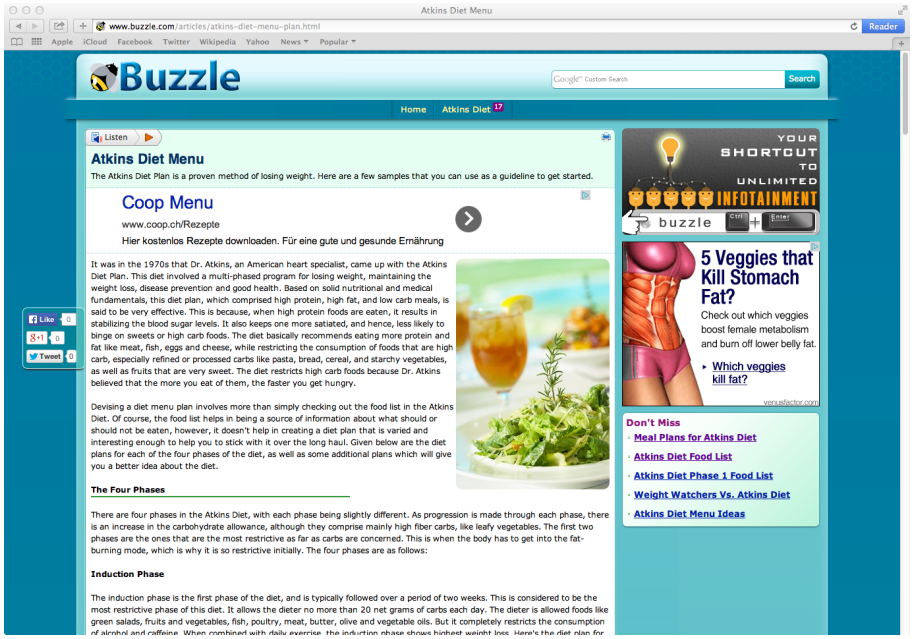


Skewed Distributions

Some class labels might be heavily skewed, e.g.

- Non-Fake pages 10000
- Fake pages 10

Rare data points missed in validation set



Fighting Skew

Stratification

- Select validation set as random sample, but assure that each class is (approximately) proportionally represented

Over- and Under-Sampling

- Including over-proportionally number from the smaller class (over-sampling)
- Including under-proportional number from larger class (under-sampling)

How Good is a Model?

Model is a function f_D that **estimates** a function

$$f: X^d \rightarrow Y \text{ with } y = f(X)$$

Evaluating the error

$$Err(f_D, T) = \frac{1}{|T|} \sum_{X \in T} (f_D(X) - y)^2$$

D = Training set from which the model is learnt

T = Validation set on which error is evaluated (test set)

Squared error measure

Training and Test Error

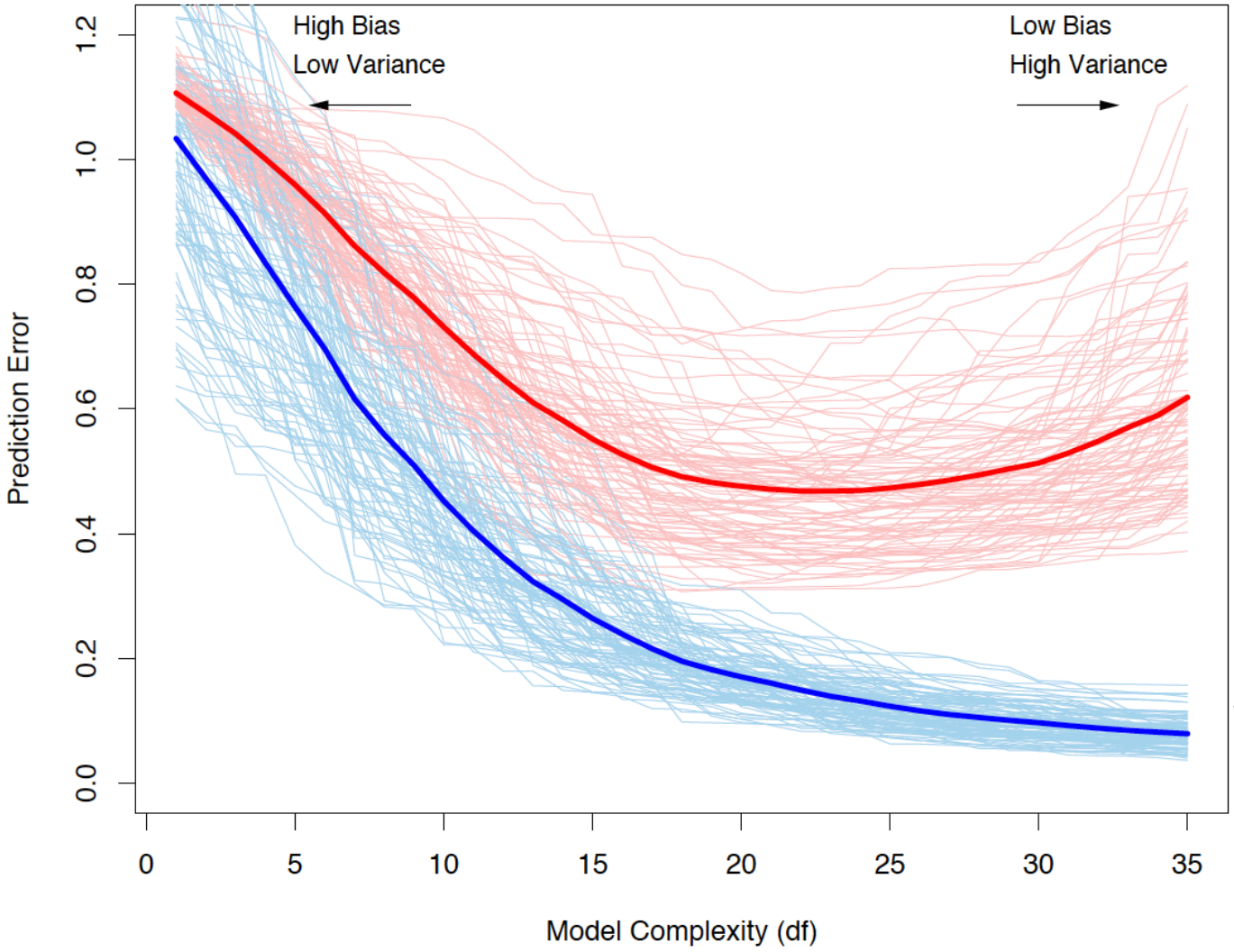
Evaluate error on training set D: **training error**

$$Err_{train} = Err(f_D, D)$$

Test model with an independent test set T: **test error**

$$Err_{test} = Err(f_D, T)$$

Comparing Training and Test Error



Test Error

Training Error

Expected Errors

Repeatedly evaluate error for different models generated from different training sets $D \in \mathcal{D}$ and corresponding test sets $T(D)$

Expected training error

$$EErr_{train} = E_{\mathcal{D}}[Err(f_D, D)] = \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} Err(f_D, D)$$

Expected test error

$$EErr_{test} = E_{\mathcal{D},T}\left[Err(f_D, T(D))\right]$$
$$= \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} Err(f_D, T(D))$$

Bias and Variance

The error can be rewritten as follows

$$EErr_{test} = Bias^2 + Variance$$

where

$$Bias = E_{\mathcal{D},T}[f_D(X) - y]$$

Deviation of predicted value from true value over all models

and

$$Variance = E_{\mathcal{D},T}[\left(f_D(X) - \bar{f}(X)\right)^2]$$

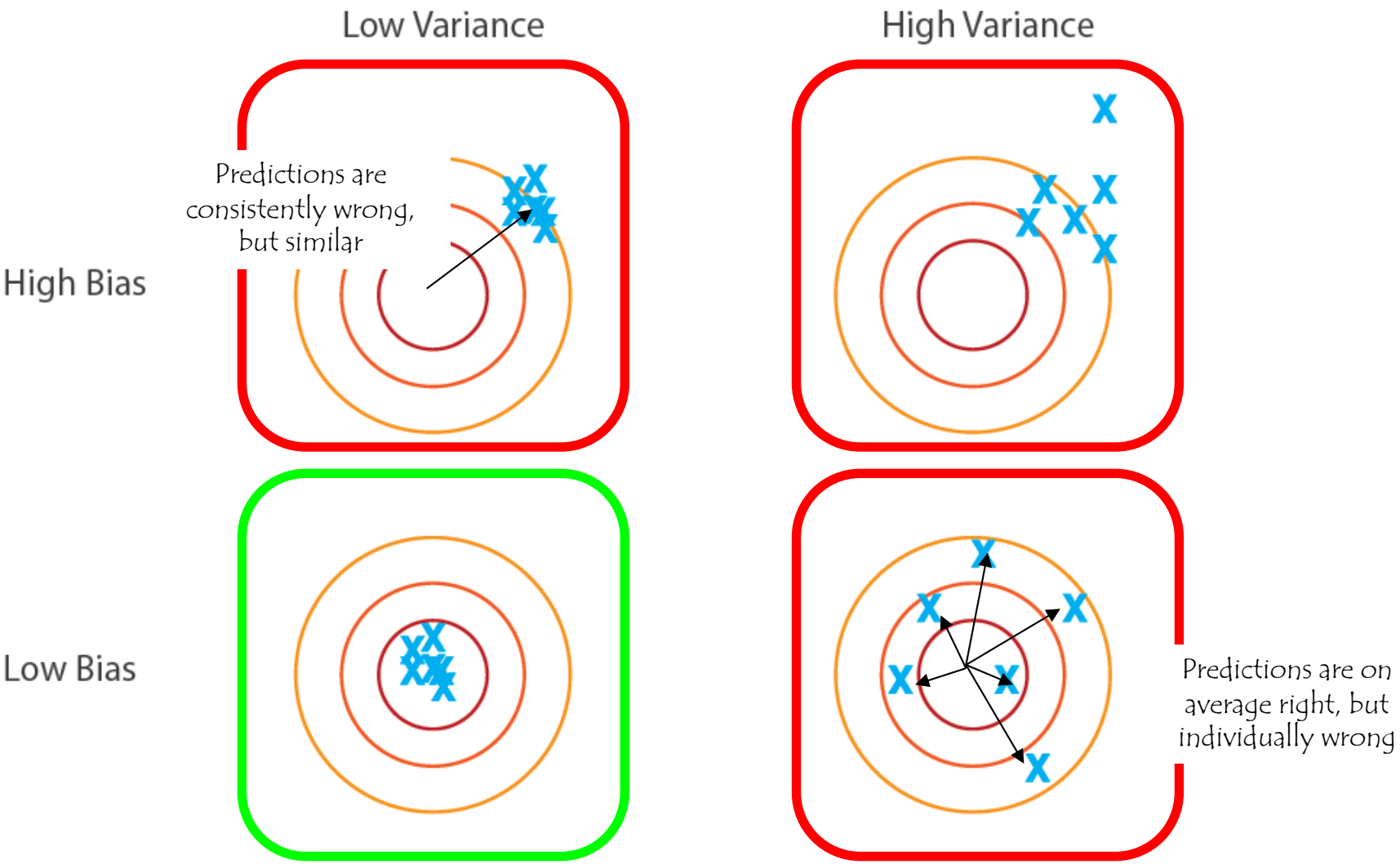
Deviation from average predicted value by all models

with

$$\bar{f}(X) = E_{\mathcal{D}}[f_D(X)]$$

Average predicted value

Bias and Variance



Bias / Variance and Model Complexity

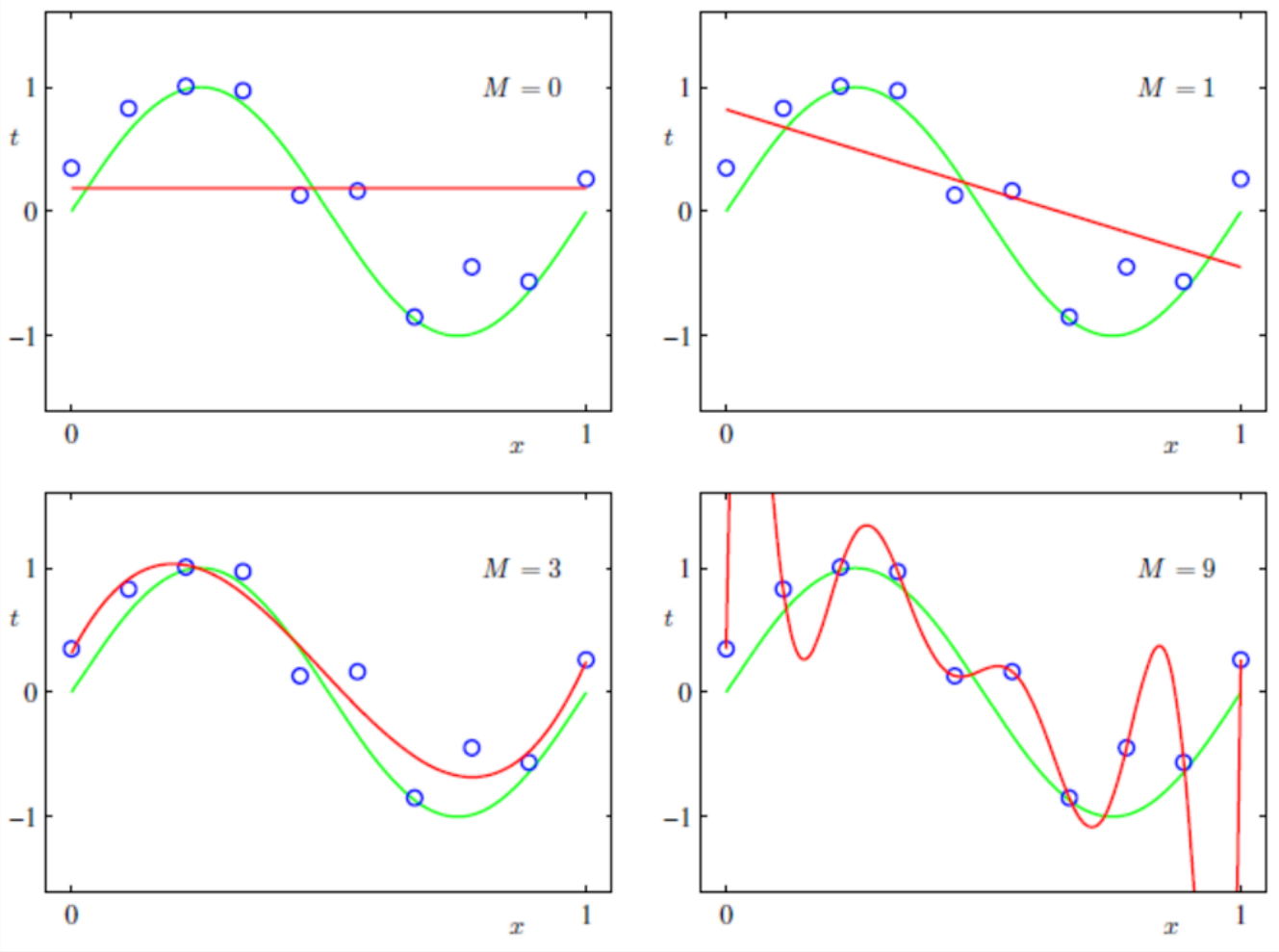
There is usually a bias-variance tradeoff caused by model complexity

Complex models (many parameters) usually have lower bias, but higher variance
→ **over-fitting**

Simple models (few parameters) have higher bias, but lower variance

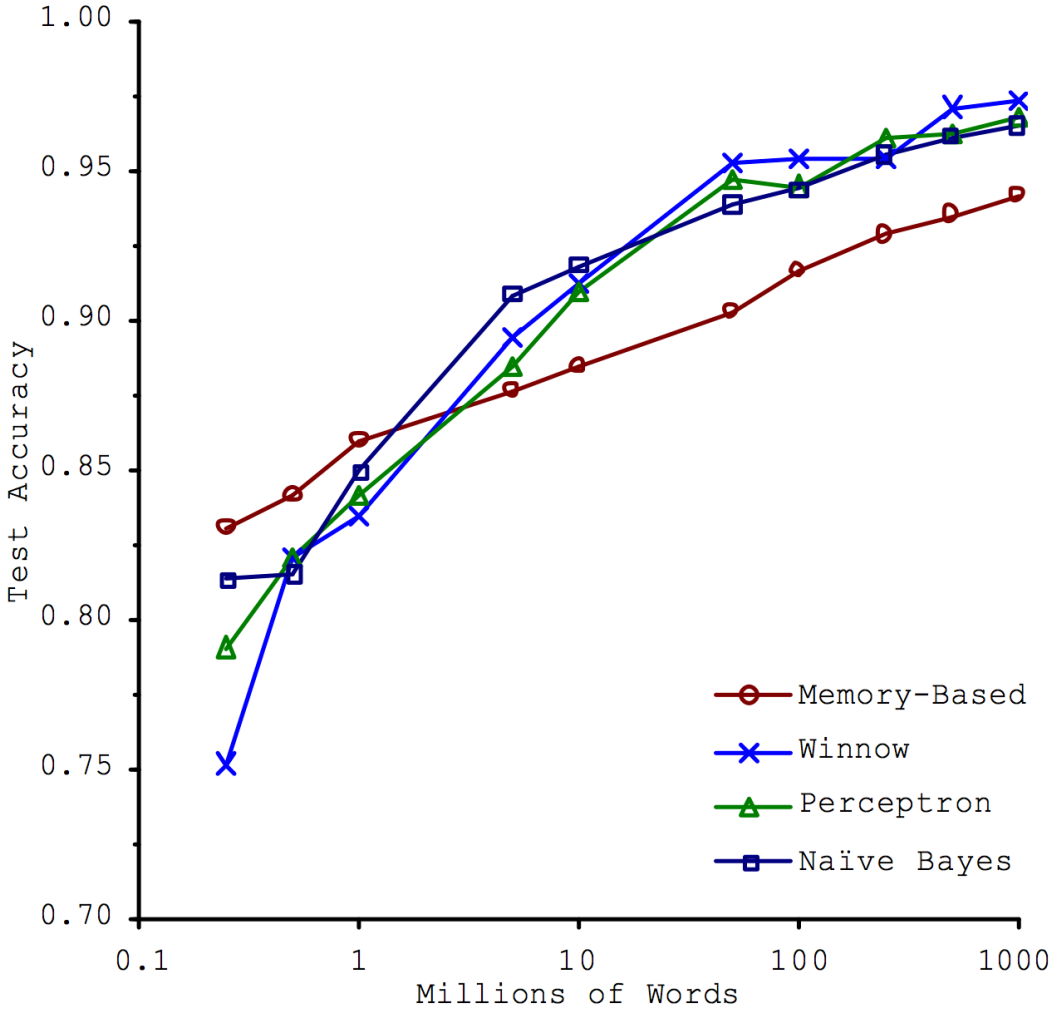
→ **under-fitting**

Example



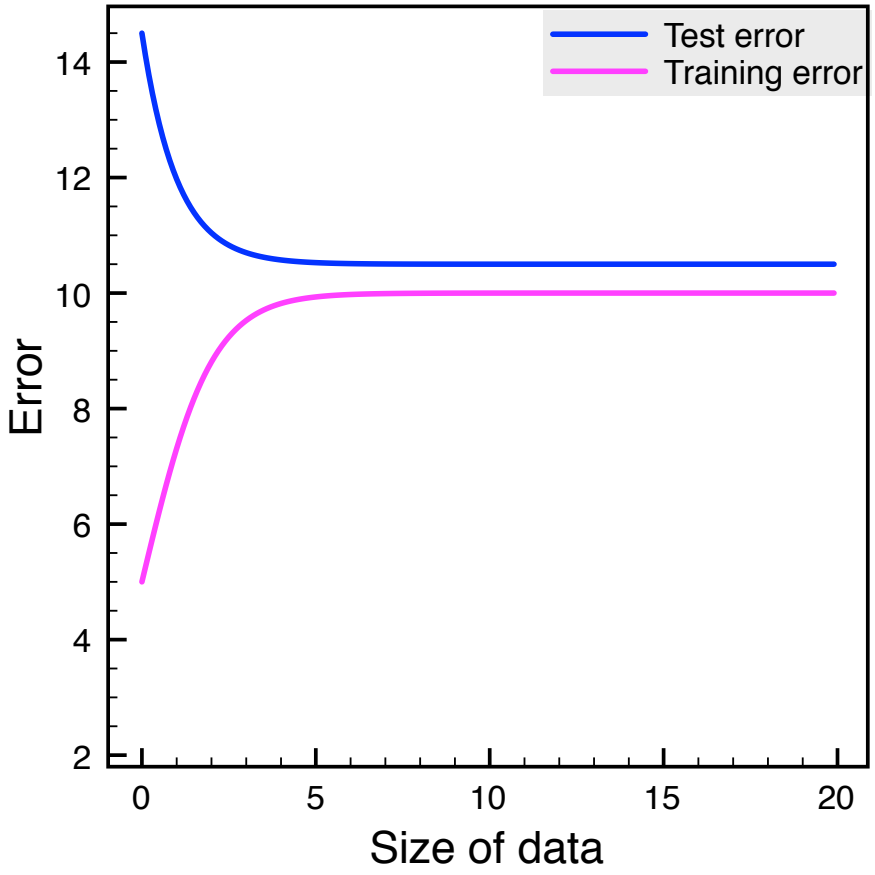
Fitting a data set to different orders of polynomials (Bishop, "Pattern Recognition and Machine Learning")

Does More Data Help?

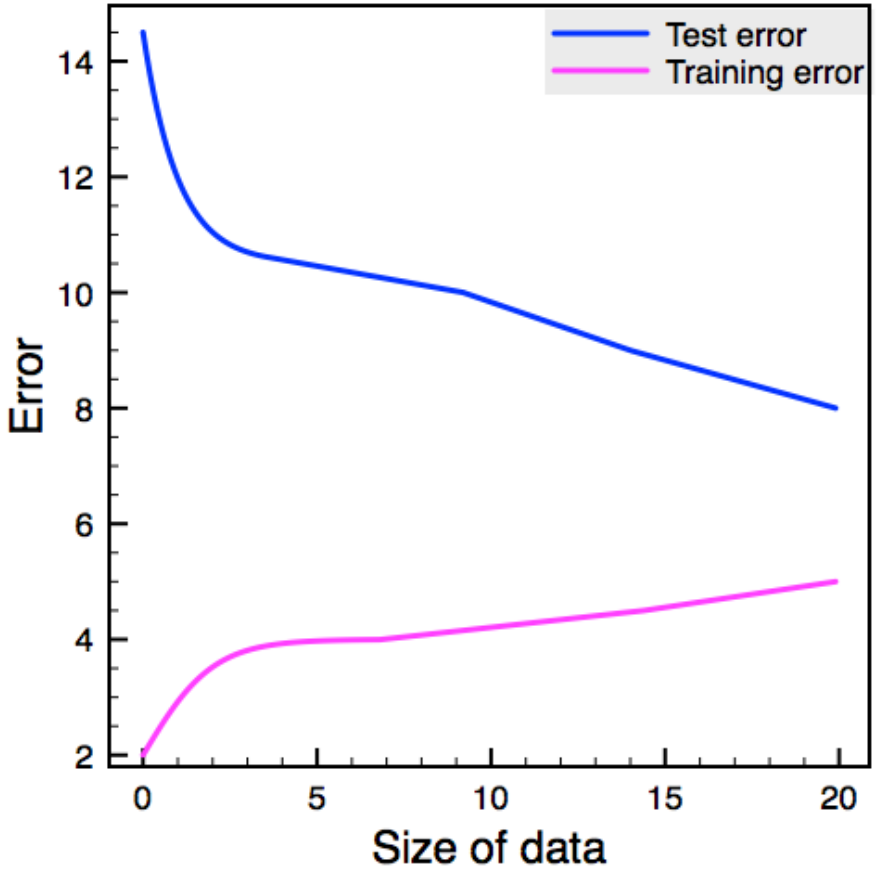


Data > Algorithms

Bias / Variance and Data Volume

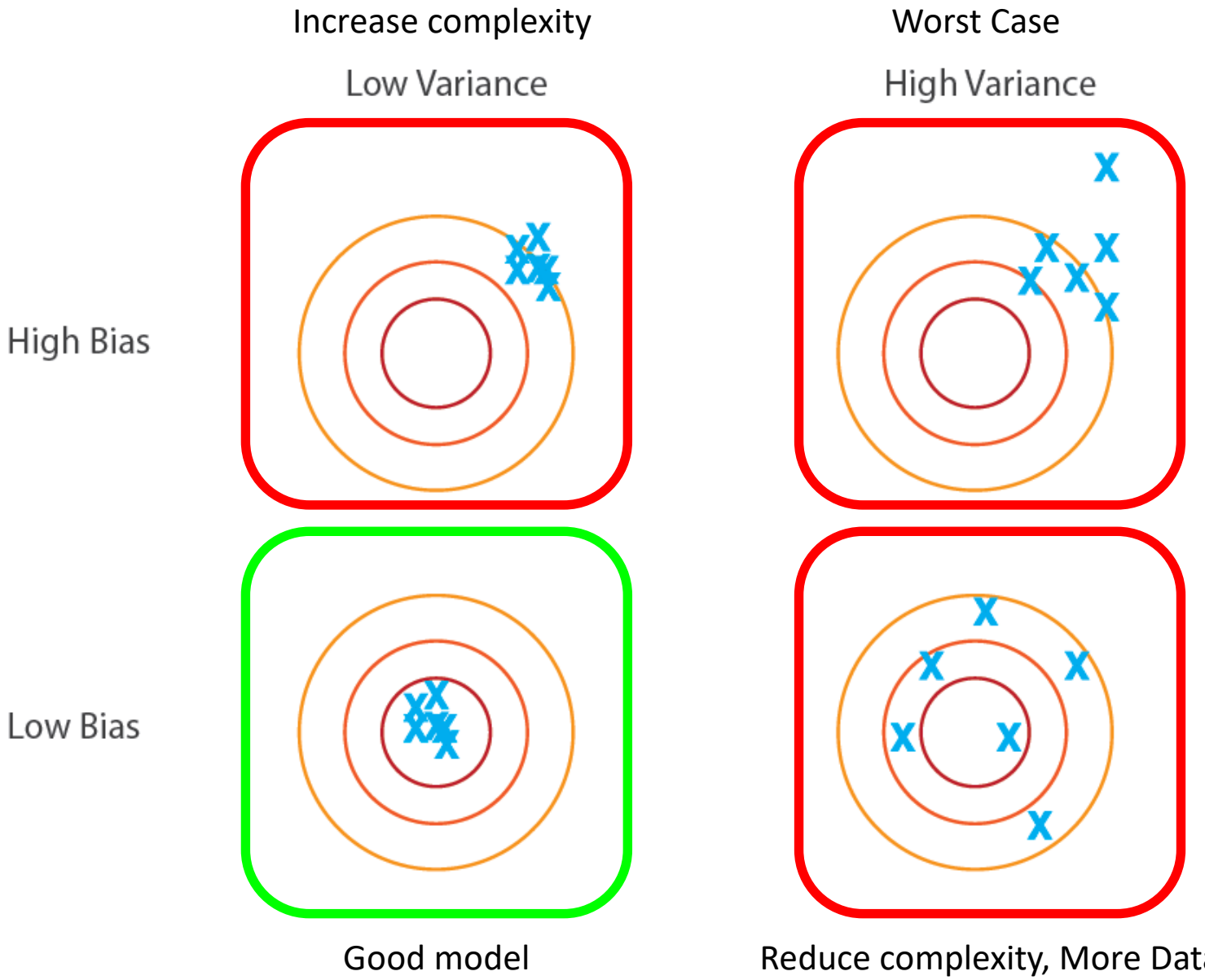


High bias



High variance

Bias and Variance



In k-fold cross-validation …

A. 1/k of the labelled data is used for training
B. The prediction is the average of k different models
C. The k validation sets are disjoint
D. The resulting error is an estimate of the quality of the classifier on real-world data

Which is wrong?

A. The lower model complexity, the higher bias
B. The higher model complexity, the higher variance
C. The higher the data volume, the higher the training error
D. The training error is always higher than the test error

References

Textbook

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001

Papers

- Banko, Michele, and Eric Brill. "Scaling to very very large corpora for natural language disambiguation." *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2001
- Chawla, Nitesh V. "Data mining for imbalanced datasets: An overview." *Data mining and knowledge discovery handbook*. Springer US, 2005. 853-867.
- http://m.shookrun.com/documents/stupidmining.pdf
- Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. "Web credibility: Features exploration and credibility prediction." *European Conference on Information Retrieval*. Springer Berlin Heidelberg, 2013.