Data Mining: Opportunities and Pitfalls

Toon Calders

ULB
ECOLE POLYTECHNIQUE DE BRUXELLES

Outline

Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects (skip)
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models
- Conclusion

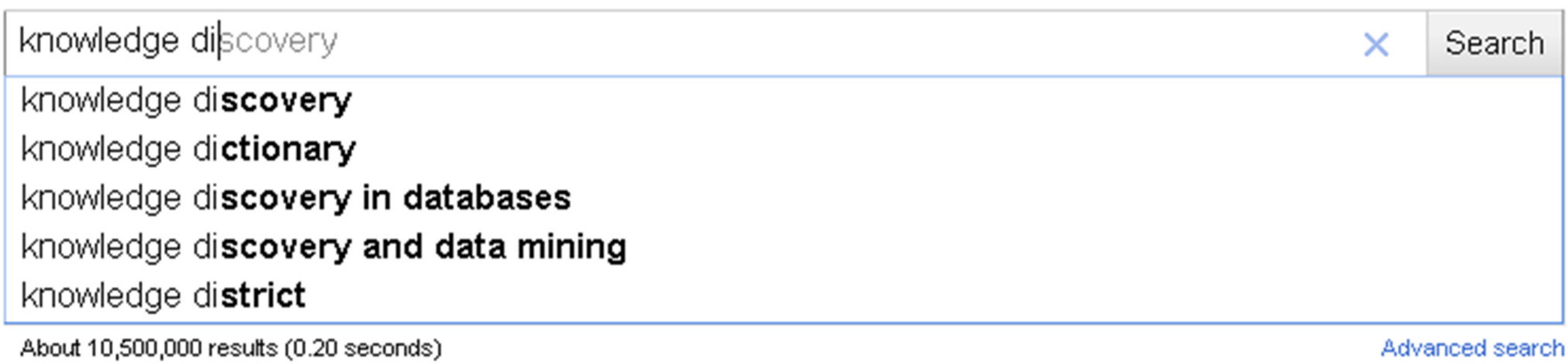
What is Data Mining?

- Data mining is the use of automatic techniques to "discover *knowledge*"
  - Data driven discovery
  - Making implicit knowledge explicit

- Data mining is part of the knowledge discovery process
  - Collecting data, Preprocessing, Mining, Visualizing, …

Unprecedented Opportunities

- **Example: n-grams dataset by Google**
  - **1,024,908,267,229 words of running text available online**
  - **All sequences up to 5 words appearing at least 40 times**
- **Applications:**
  - **auto-complete**



  - **Machine translation, auto-correction, …**
- **Statistically-based techniques rule**

Different Categories of Data Mining

- **Exploratory Analysis**
  - **Clustering**
  - **Outlier Detection**
  - **Association Rule / Frequent Pattern mining**

Exploratory Data Mining

110000000100000001000011001000000011000000000000010000100
111000000100000010000011000000000010100000000000010000000
111000000100000000000011000000000011000000000000010000000
010111000000110111011100100110011100011111100001000000000
000111000000110101011100000110011101011111100001000000000
111000000100000000000011000010000011001000000000010100000
110000000100000000000011000000000011000000000000010000000
111100101100001000000011111000100011000000001100011000000
000111000000110111011100000110011100011111000001000000000
000111011011110111111100000111111000111111100111001100000
000000100000000110000001111100000000000000001011000010000000
000000111011110011000000000000110000110000000011100110000
000000111011110011100000000000110000010000000011100110000
000100001011110011100001000000110000010001000011100110000
000000011010110010110110000000110001010000000011100110000
000000100000000100001010111100000000000000000011000110000000
000000100000000100000001111100000000000000000011000110000000
000000100010001000000001111100000000000000000010000111000000
001000100000000100000001011100000000000000000011000110000000
000111000000110111011100000110011100011111100001000000000
000111000000110111011100000110011100011111100001000010000
000000100000000100000001111100000000000000000011000110000000

111101111000000000000000010000000001000000000000000000000

111111111000000000000000000000000000000000000000000000000

111111111000000000000010001000000000000000000000000100000

111011111000000000000000000000000000000000000000000000000

111111111111111110000000100000000000000000000010010000000

000000111111111110000000000000000000000000000000000000000

000000111110111110001000000000000010000000000000000000000

000000101111111110000000000101000000000000000000000000000

000000111111111110000000000000000000000000000000000000000

000000111111101110000000000000000000000000000010010000000

000100110111111110000000000000000000000000000000000000000

000000000000000001111111111111111111111100000000000000000

000000000000000001111111111111111111111100000000000010000

000000000000000001101111100111111111111100000000100000000

001000000100000001111111111111111111111100000000000000000

000001000000000001111111111111111110111100000000000000000

000000000000000001111111011111111111111100000000000000000

000000000000000001111011111111111111111100000000000000000

000000000000000001111111111111110111111111111111111100000

000001000000100000000000000000000001111101111111111100000

000000000000100000000000000000000001111111111111111100000

000000001000000000010000100000000001111111111011111100000

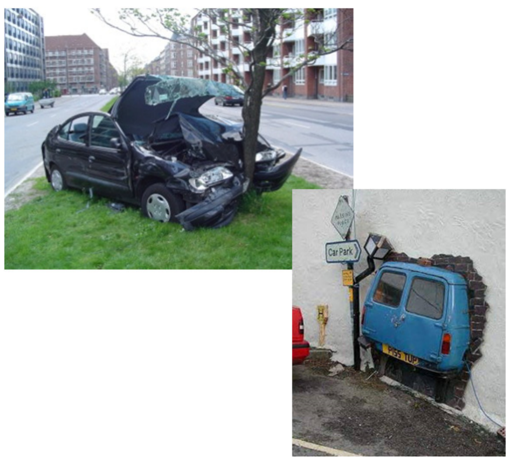
000010010000000000000000000101000001111011111111011100000

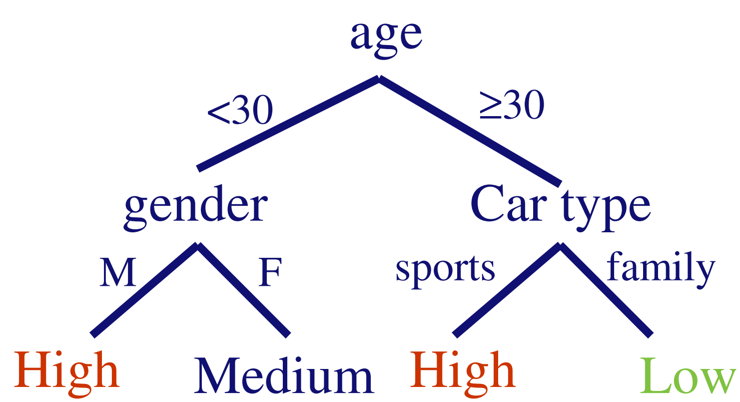
000000000000010000000000000000000001111111011111111100000

Different Categories of Data Mining

- **Predictive Modeling**
  - **Classification, Regression**

Example:





Other Types

- Some data types require special treatment
  - Text mining
  - Spatio-temporal data mining
  - Web mining
  - Sensor network mining
  - …
- Usually the algorithms fit into one of the previous categories, but require specialized algorithms

Examples of Data Mining in Policing

- Data Mining in policing
  - Predict crimes: type, location, time, time of the year, …
  - Learn characteristics of people that wear concealed weapons
  - Find patterns in crimes; e.g., sudden increase in burglaries in one particular area

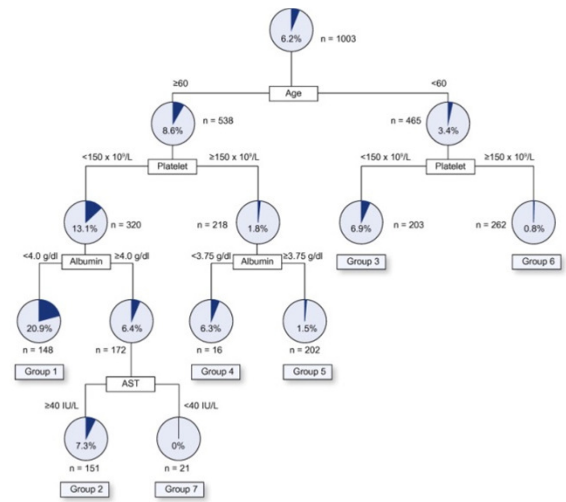
See, e.g.: R. van der Veer, H. T. Roos, A. van der Zanden. Data mining for intelligence led policing. In ACM SIGKDD Workshop on Data Mining Case Studies (2009)

Phases in a Data Mining Project

Data collection → Data warehouse → Selection, preprocessing

A	B	C	D
1	1	1	1
0	1	1	1
0	1	0	1

→ Data Mining - Modeling →



Evaluate → Deploy

Is Data Mining Statistics (or vice versa) ?

YES:

- **Goals are similar: inferred from data**
- **Many techniques in common (linear regression, EM)**

NO:

- **Two communities with very different methodologies:**
  - **Statistics: heavily based on assumptions and models**
  - **Data mining:**
    - **Ad-hoc methods; proof of the pudding is in the eating**
    - **Difficult to answer:**
      - **How much data needed?**
      - **Confidence interval?**

Outline

Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects
  - Data Collection & Pre-processing
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models
- Conclusion

Data Collection & Pre-Processing

- **Often performed before data mining**
  - **Impossible to influence data gathering**
  - **Observatory data mining**

- **Pre-processing**
  - **Careful! Do not introduce artifacts …**

Sanity check: if a new instance comes, can I apply the same preprocessing at the time I have to predict?

Introduction of Artifacts During Pre-processing

Case: classify fMRI-scans: THC / normal

- **Preprocessing: Independent Component Analysis for THC and normal separately**
- **100% predictive accuracy …**

Case: time series prediction

- **based on first X weeks of sales, predict sales in week X+1**
- **preprocessing: normalize complete timeseries**
- **Algorithm: if sales in weeks 1…X below 0: predict above 0**

Outline

Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects
  - Data Collection & Pre-processing
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models

- Conclusion

Data Quality Problems

Problem 1: the sample is biased

- Data collection often biased
  - Internet survey to poll election results
  - Alcohol tests mainly taken from young males
  - Tax declarations more thoroughly checked for suspicious profiles

Data observational: often unaware of the precise conditions of data collection

Sample Bias: Case Study

- Company designing medical appliances for assisting patients with coronary heart disease
  - Extend time they can stay at home
  - Increase quality of life

Research question: which person to give what appliance?

Data included a strong bias: more expensive device given only to people in a higher risk category

→ more expensive device seemingly worse

Cfr.: propensity modeling

Data Quality Problems

Problem 2: self-fulfilling prophecy

In many cases there is an interaction between data, action, and outcome.

- **Bank: model who will default to improve loan approval procedure**
  - **Only people that were accepted in the first place can default**
- **Who wears a concealed weapon?**
  - **We only know for those people that were checked**

Identify strata of instances that were treated similar.

Case: Self-Fulfilling Prophecy

- Funnygames.nl
  Task: which games to list in the "popular" listing?
  - If a game gets many hits, it gets into the list
  - If a game is in the list, it will get many hits

  First task:
  - model the effect of being in the list
  - How many hits would a game have gotten if it would have been in the list?

Outline

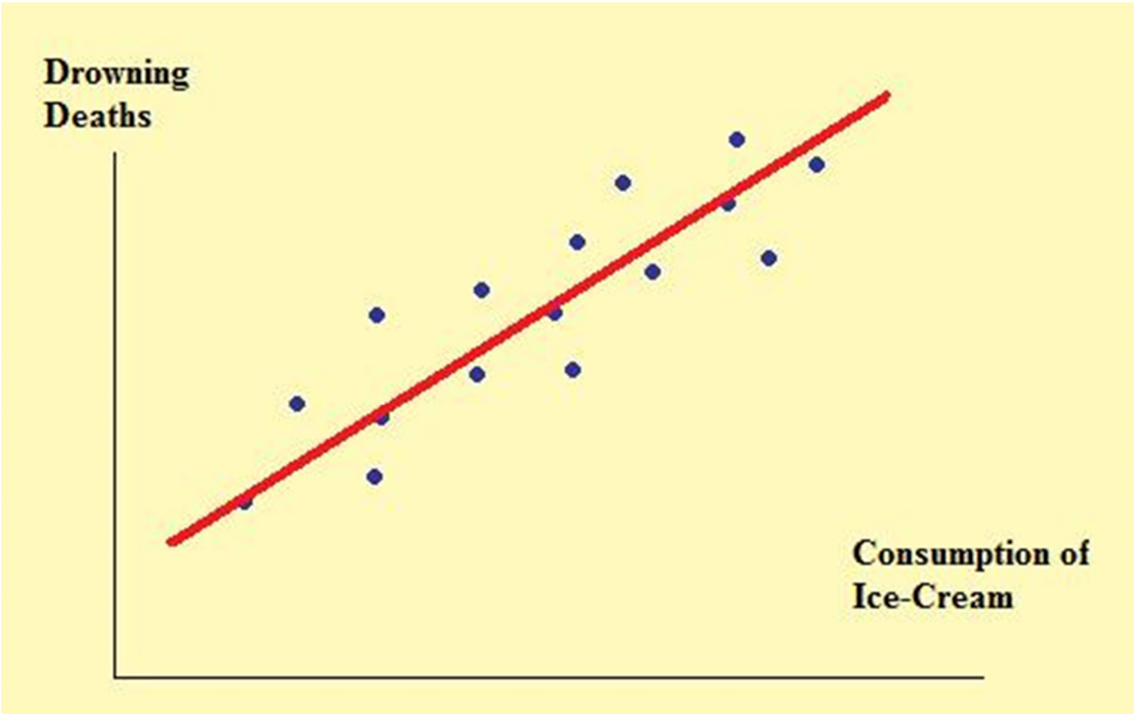
Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects
  - Data Collection & Pre-processing
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models

- Conclusion

Correlation ≠ Causality

- Diet Coke → Obesity
- Intensive Care → Death
- Drowning versus Ice Cream:



Correlation ≠ Causality

Performance → Trust or Trust → Performance?

Simpson's Paradox

- **Berkeley Case (1973)**

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Source: Wikipedia

False Discoveries

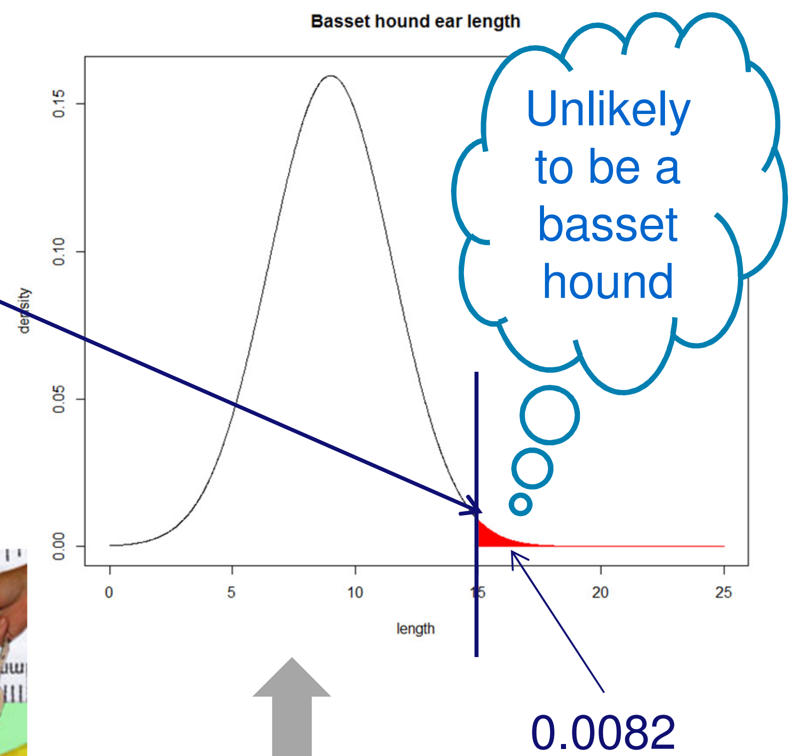
- **Statistical test:**
  1. **Formulate hypothesis**
  2. **Collect data**
  3. **Test hypothesis on the data**

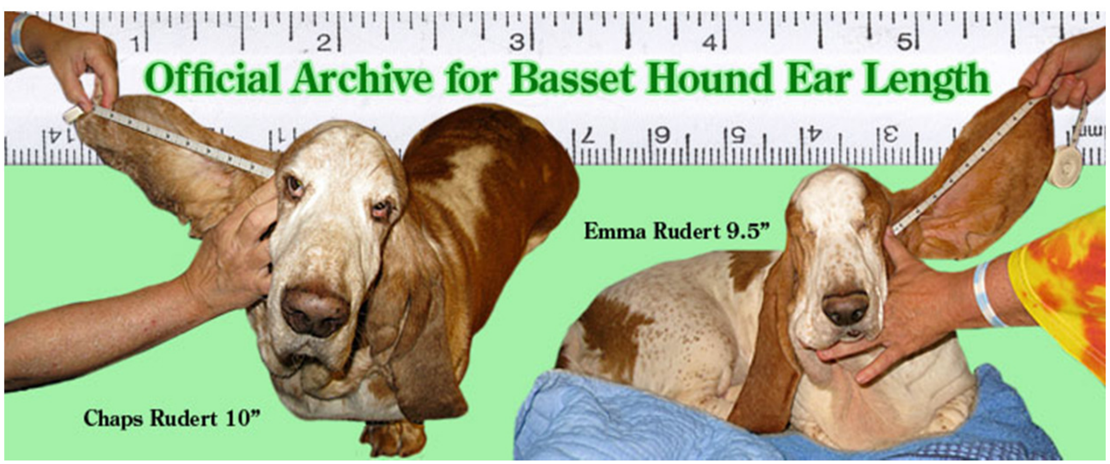
- **p-value expresses *how extreme* a finding is**
  - **“the chance of getting the observed outcome is p”**
  - **If p is very low: reject hypothesis**

Coin example

p-Value Illustrated

- **H0: the following animal is a Basset Hound**





False Discoveries

- **Example: is the coin fair? Toss 10 times:**

- **If the coin is fair, the probability of having 8 or more heads or 8 or more tails is approximately 11%**

False Discoveries

- **Example: is the coin fair? Toss 10 times:**

- **If the coin is fair, the probability of having 9 or more heads or 9 or more tails is approximately 2%**

False Discoveries

- Data mining:
  - Collect data
  - Generate hypothesis using the data

- Two important differences with statistical test
  - Data is not collected with the purpose to test hypotheses
  - Many hypotheses are generated and tested

- Hypotheses found by data mining do not have the same status as statistical evidence!
  - Cfr. Lucia de B.

Corrections for Multiple-Hypothesis Testing

- p-value =
  P[outcome as extreme as observed | M_0]
  - p-value expresses probability of false positive
- Suppose N hypothesis $H_1, \ldots, H_N$
  - Adjust significance level to α/N

- If all pass at level of significance α/N :

P[any of Hi is FP | M_0]

$$\leq P[\, H_1 \text{ is FP} \mid M_0 \,] + \ldots + P[\, H_N \text{ is FP} \mid M_0 \,]$$

$$\leq N\, \alpha/N = \alpha$$

Do we have to correct?

- Given a database D
  - Generate all association rules $X \rightarrow Y$ with minsup 10%, minconf 75%
  - Select the rule with the highest *lift* L

    (lift $X \rightarrow Y = P_{obs}[Y|X] \,/\, P_{obs}[Y]$)

Null model:

X and Y are independent

P-value = $P[\ \text{lift}(X \rightarrow Y) \geq L \mid X \text{ and } Y \text{ are independent}\ ]$

Can easily be approximated with a normal distribution

One test, so no correction needed?

What's the Problem?

- **We are using the same dataset for exploration and testing!**
  - **Very common mistake; even many research papers have this problem**

Example:

Random dataset

all entries: P[0]=P[1]=0.5

Explore: select first transaction

$P(\text{support}(0011010) \geq 1 \mid M_0)$

$\leq 1 - (127/128)^5 \approx 3.8\%$

A	B	C	D	E	F	G
0	0	1	1	0	1	0
1	0	0	1	1	1	1
1	0	1	0	1	1	1
0	0	1	0	0	0	0
1	1	1	0	0	1	1

Lucia de B

- Nurse in a Dutch hospital
  - Accused of murdering several patients and convicted
  - Statistical "evidence": probability of being involved in as many incidents as Lucia was: *1 out of 342 million*

- Statisticians soon started criticizing the method:

  "it is one of these problems that seem to arise all over the place: one sets up an hypothesis on the basis of certain data, and after that one uses the same data to test this hypothesis."

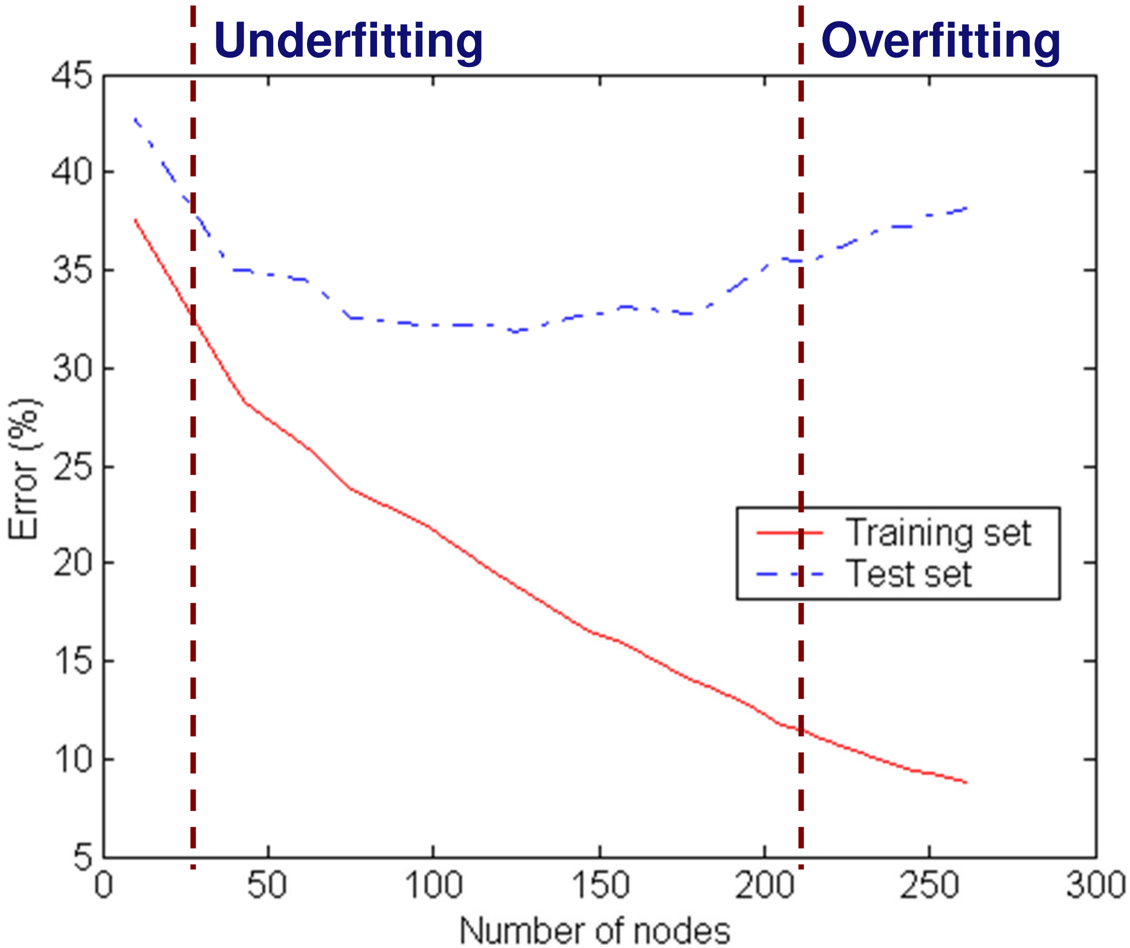
More information: R. Meester et al. On the (ab)use of statistics in the legal case against the nurse Lucia de B. Law, Probability and Risk Advance Access (2007)

Outline

Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects
  - Data Collection & Pre-processing
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models
- Conclusion

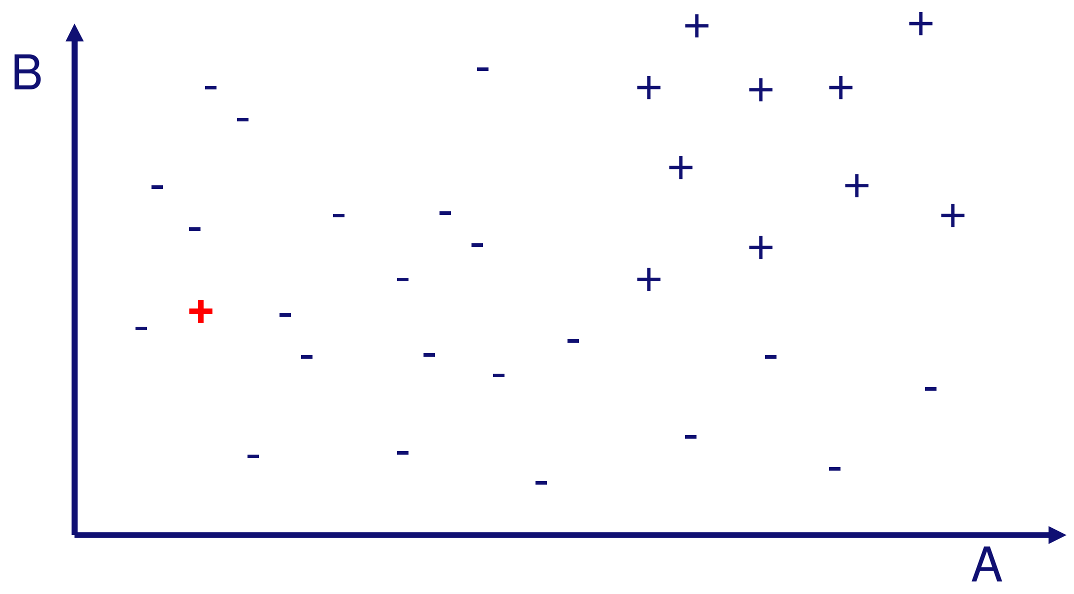
Overfitting



Underfitting: Model did not see enough data
Overfitting: Model learns peculiarities of input data

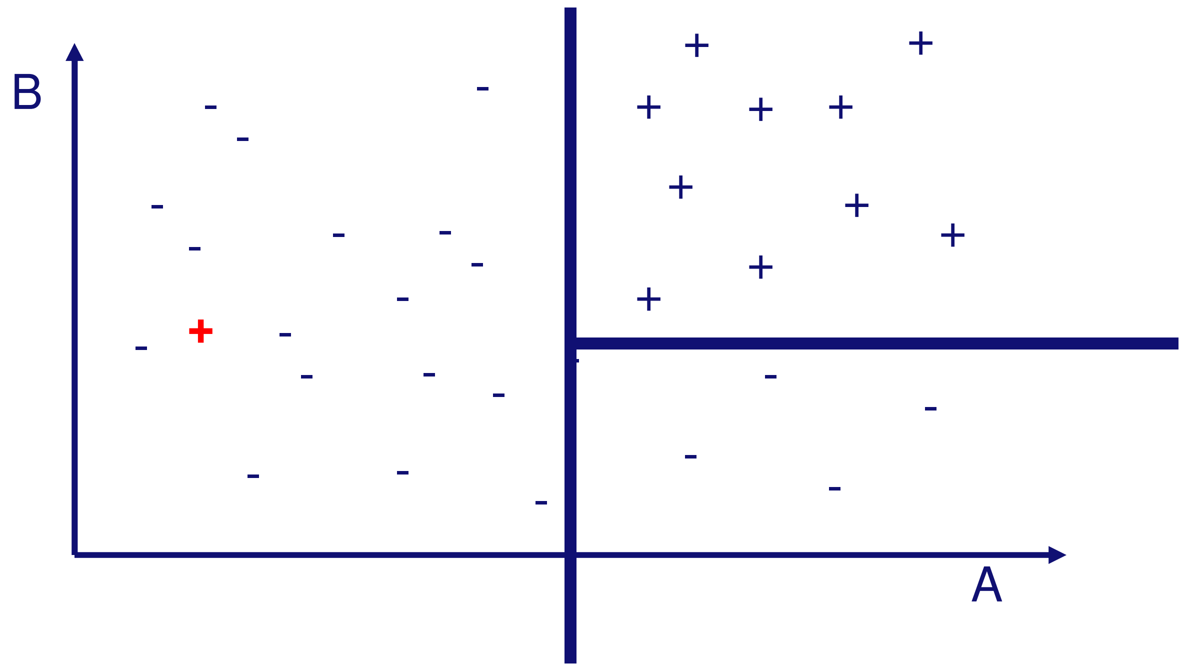
Overfitting Due to Noise

- **Two-dimensional data, class + or -**



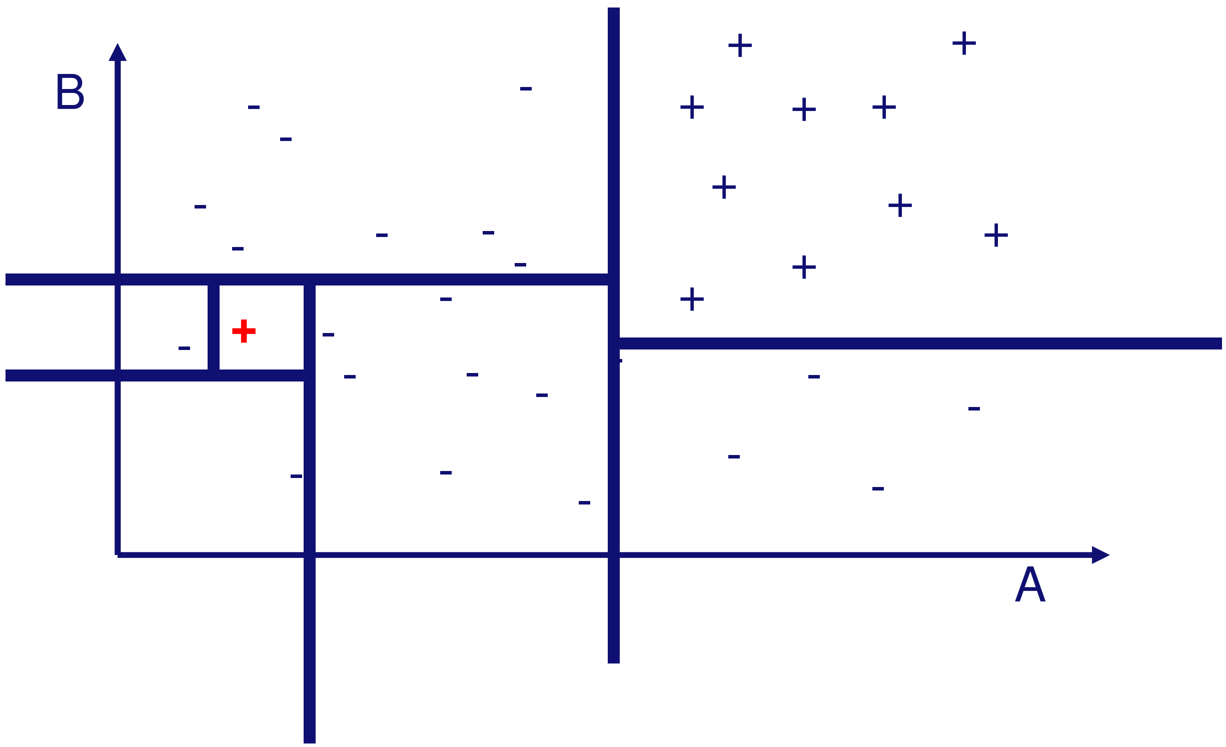
Overfitting Due to Noise

- **Good model**



Overfitting Due to Noise

- **Bad model with better training performance**



Performance of Classifiers

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n

Overfitting

- Not always that obvious
  - Optimize parameter outside cross-validation loop
  - Try different approaches to the data, select the best
- Do not test if there is over-fitting, but how much

Case study: food sales prediction

- Hypothesis: different types of products, depending on product a different prediction algorithm should be used
- Products clustered according to which prediction algorithm performed best
- Artificial gain of 1% → lots of $$$

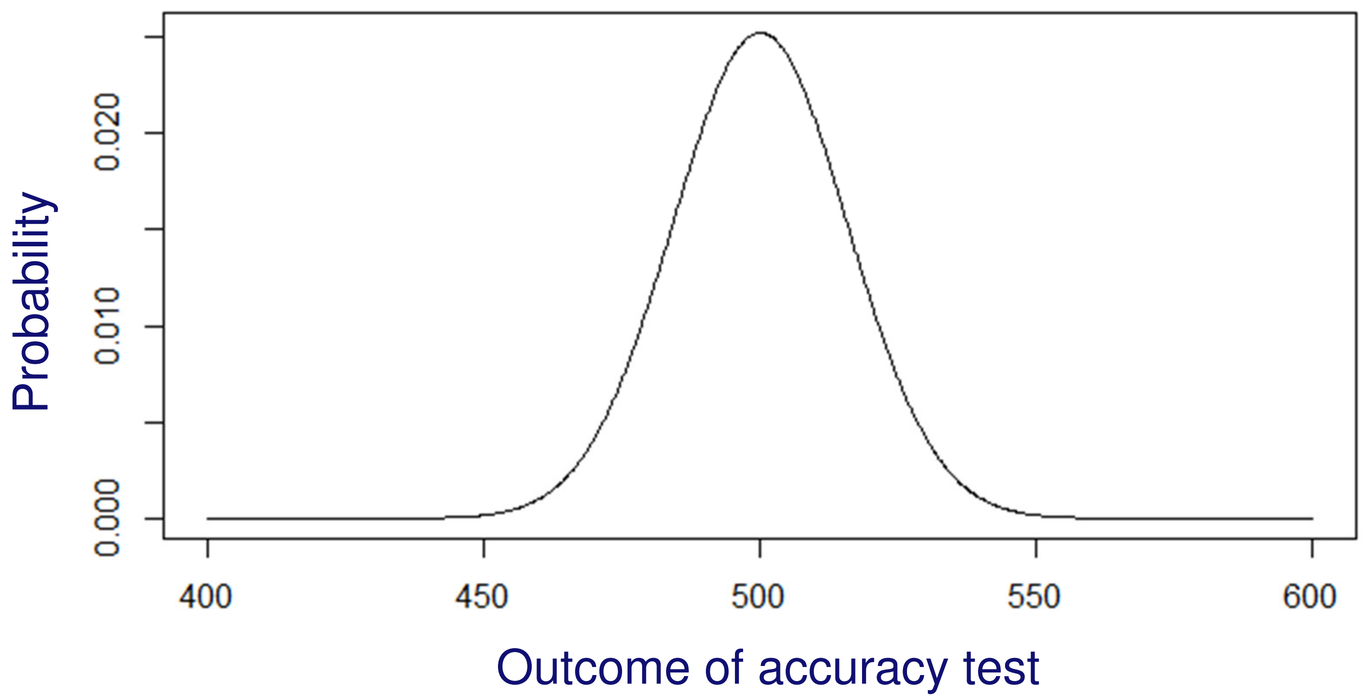
Overfitting Illustrated

- Test 3 types of classifiers:
  - Nearest Neighbor for 3, 5, 7, 9, 11, 13 neighbors, for Euclidian distance measure, and Cosine similarity
  - Decision tree
    - Binary split versus multi-split
    - Gini-index versus Information gain
  - Support vector machine
    - Parameter “c” varied: 1, 100, 1000, 10000

- All are tested with 10-fold cross-validation
  - Baseline performance is 50%
  - Best classifier is selected; accuracy = 90%

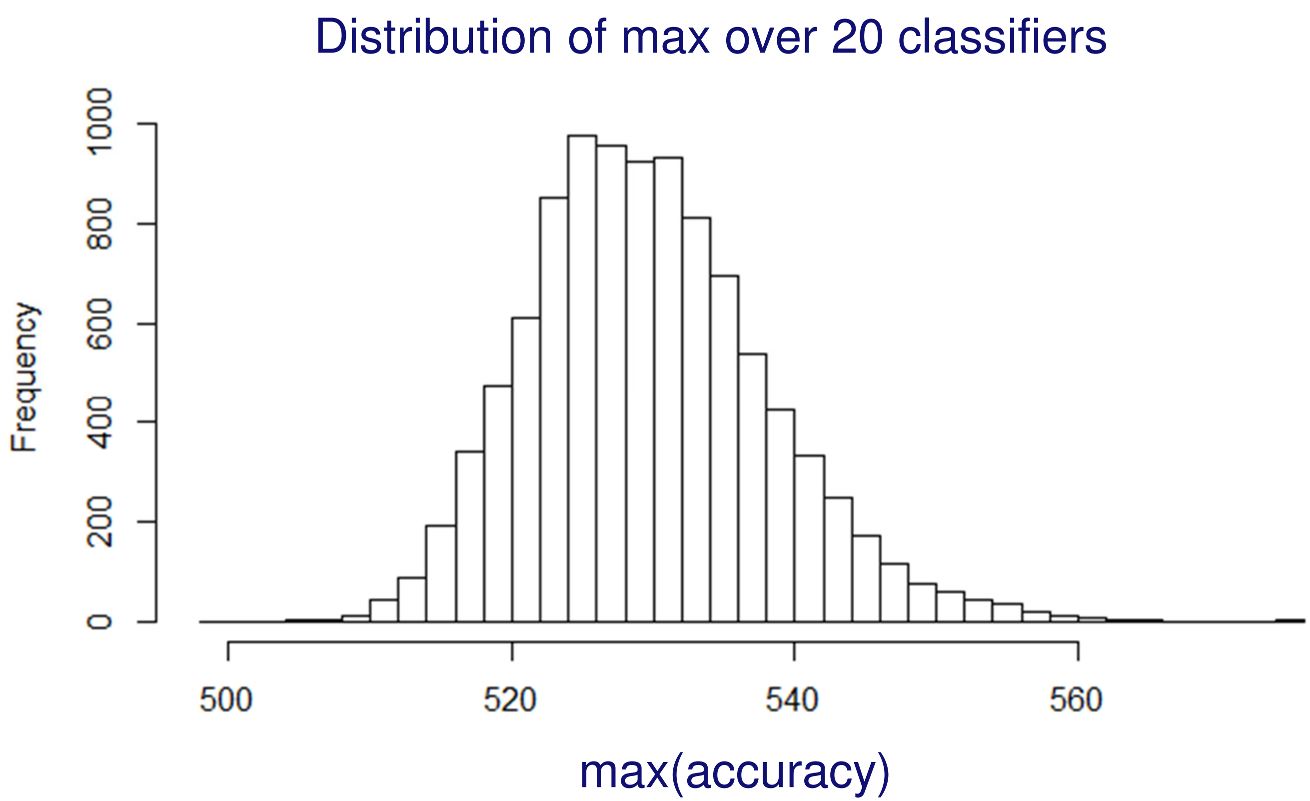
Overfitting Illustrated

- **Suppose all classifiers are random; i.e., E[accuracy]=50%**
  - **Distribution of the accuracy measurement on 1000 examples**



Overfitting Illustrated

- **Distribution of the maximum over the 20 classifiers**



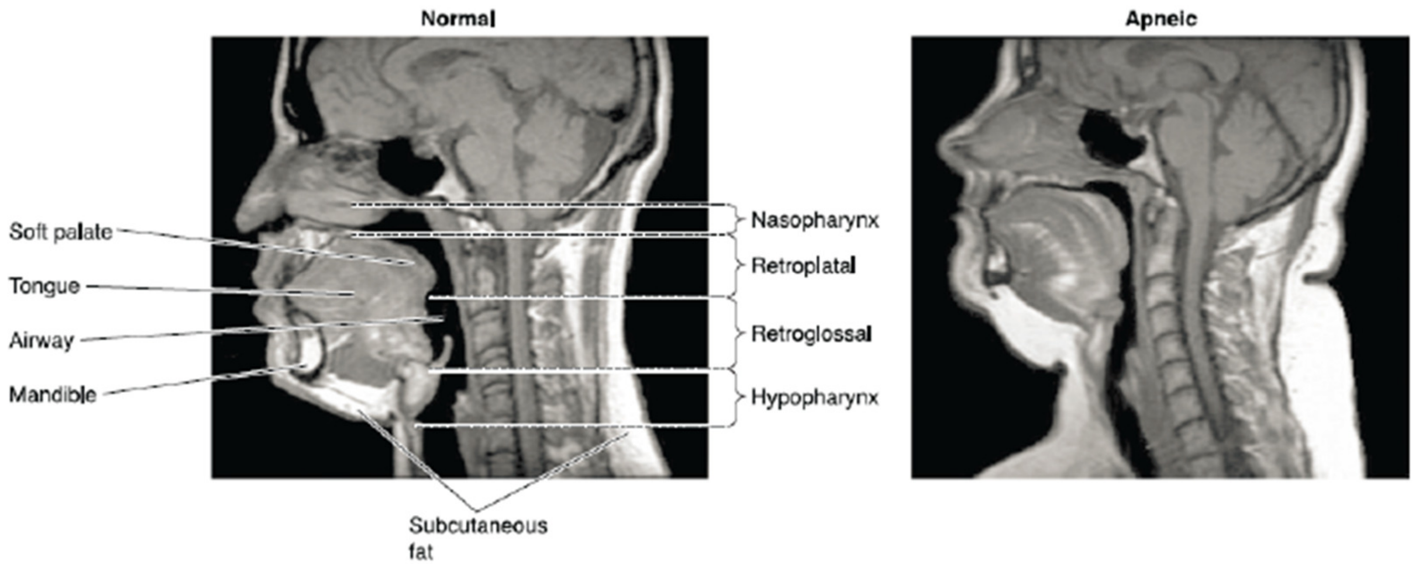
Systematic over-estimation of performance

Overfitting

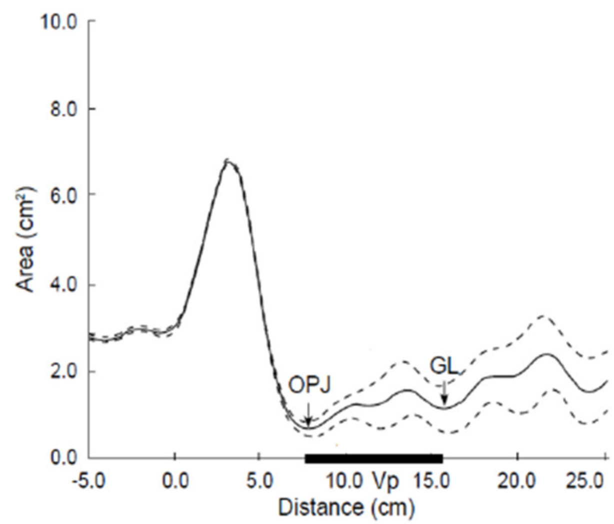
- Effect is amplified by:
  - High class imbalance
  - Small datasets
  - Many features/attributes
  - Learning methods with many parameters

- Unfortunately these data properties are very common
  - DNA data: usually huge data about few patients
  - Personalization: more and more models being built at individual customer level
  - Feature construction often part of learning process

Overfitting: Case Study



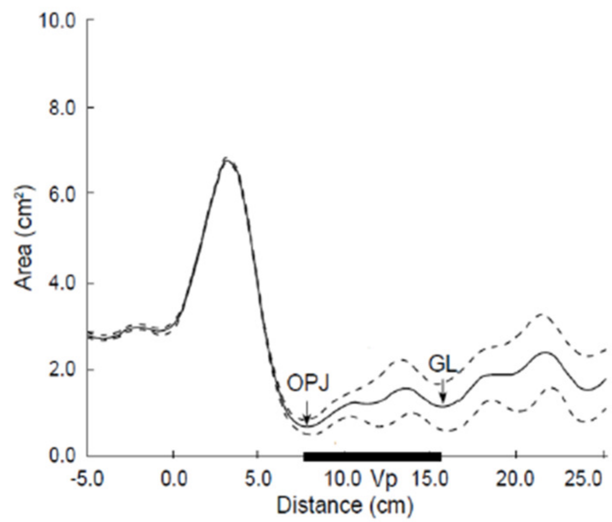
Using sound waves to
create cross-cut of throat



Overfitting: Case Study

- Volume in certain segments of the throat are very important
  - Exact region of importance, however, is unknown
- Feature extraction: for every possible region:
  - Compute the volume, add as a feature
  - Learn the best model; use cross validation
  - In the end select the region that gave the best performance

- Length of sequence: 100
  - 5000 segments

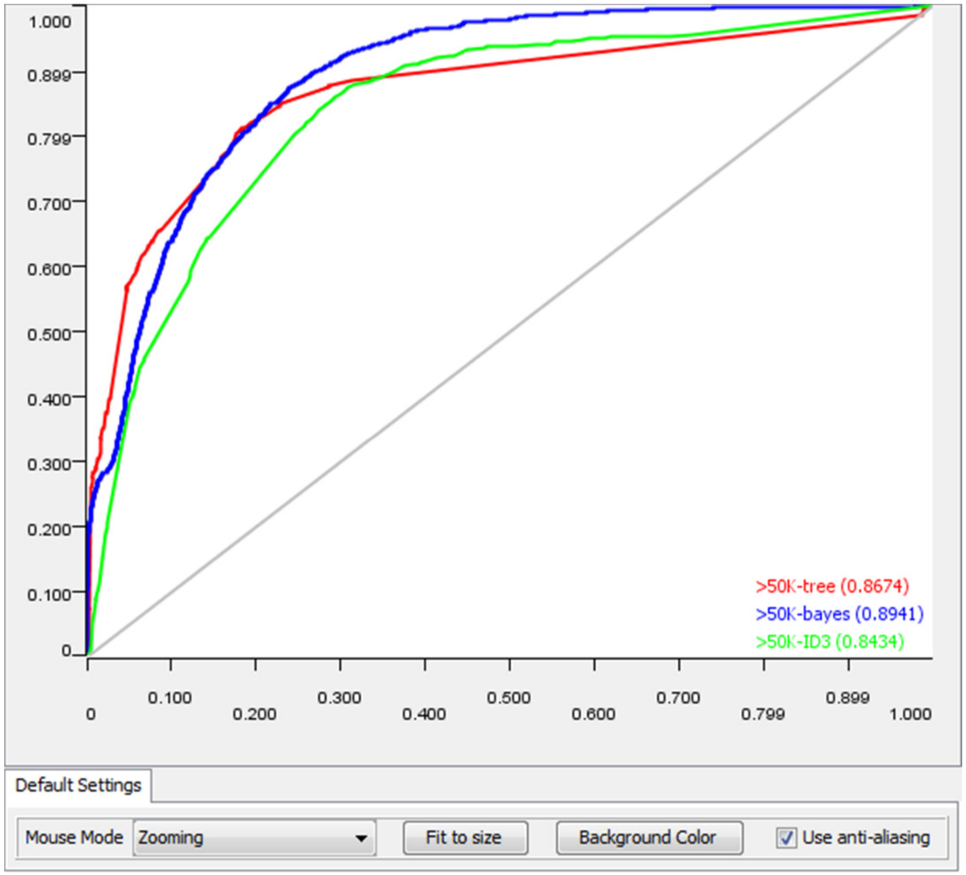


Outline

Overview of personal experiences

- What is Data Mining?
- Pitfalls experiences in data mining projects
  - Data quality problems
  - Interpretation problems
  - Over-fitting
  - Deploying models
- Conclusion

What is the Best Model?



Alg.	AUC	Acc
Dtree	0.87	86%
NB	0.89	83%
ID3	0.84	82%

Deploying Models

- Good model does not necessarily mean good prediction
  - Will person get cancer?
    - Smoker: increase risk
    - Red meat: increased risk
    - Alcohol: increased risk
  - Best model: always predict NO
    - Even the smoking, red meat eating, drinking person has more chance NOT to get cancer than a smoker

Deploying Models

- Credit scoring: who to give a loan?
  Data of a bank, with who defaulted
  Task: improve the banking algorithm
  - Had an excellent risk model
  - Did not improve the bank's predictions

- Note: AUC measures model quality not predictive accuracy. For specific tasks it is hardly useful.

Skip riddle

Riddle

A man and his son were in a car accident. The man died on the way to the hospital, but the boy was rushed into surgery. The surgeon said “I can't operate, for that's my son!” How is this possible?

Popular Answers

- One is the adoptive dad and the other is the biological dad
- The surgeon is the father's domestic partner, and the kid is a test tube baby
- The milkman became doctor after years of hard study
- God is the surgeon. God views all humans as his children
- The doctor is the church father …
  … who knows what happened when the boy was conceived

Popular Answers

- One is the adoptive dad and the other is the biological dad
- The surgeon is … l the kid is a test tube b…
- The milkman b… years of hard study
- God is the surg… humans as his children
- The doctor is the church father …
  … who knows what happened when the boy was conceived

What about the surgeon is the boy’s mother?

Objectivity of Data Mining

- Human judgment is influenced by circumstances
  - Gender stereotypes
- Opportunity for data mining
  - Based on data
  - Objective
- Is it really?

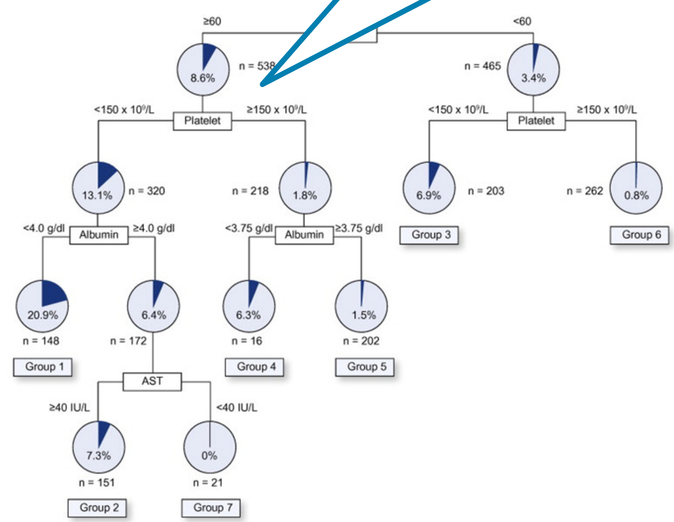
Discrimination and Data Mining

May be biased

Data

Models

Will be even more biased

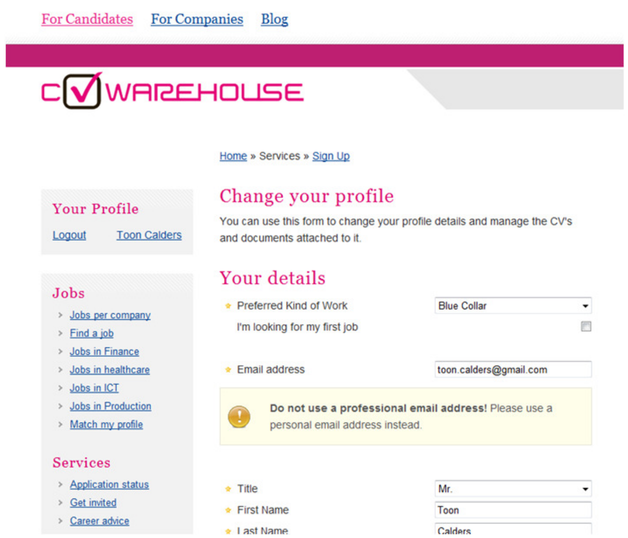


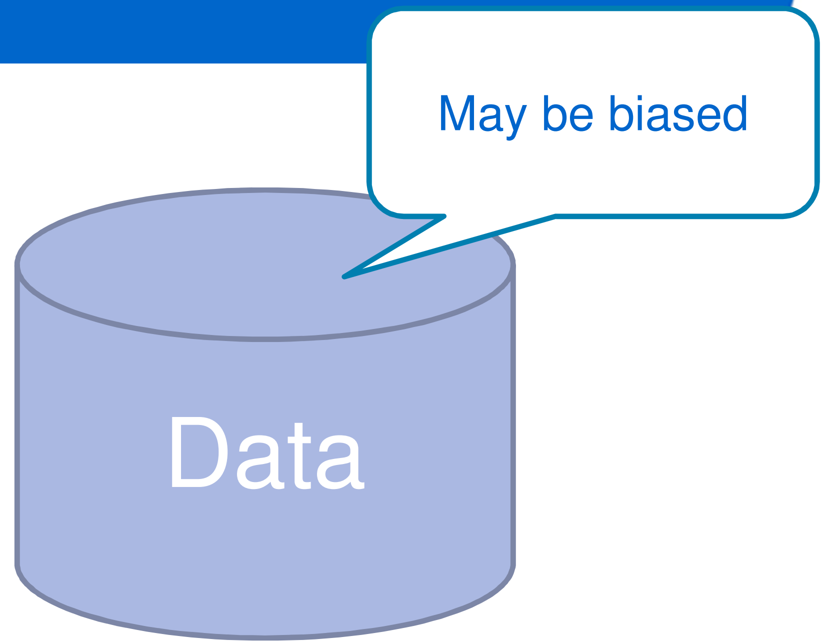
Integrate into a decision procedure

Legal/ethical restrictions

Bias in Data

- **Discrimination in data**
  - **E.g., job offering**





- **Selection bias in data**
  - **Company owns data about customers**
    - **Once accepted for loan/insurance**

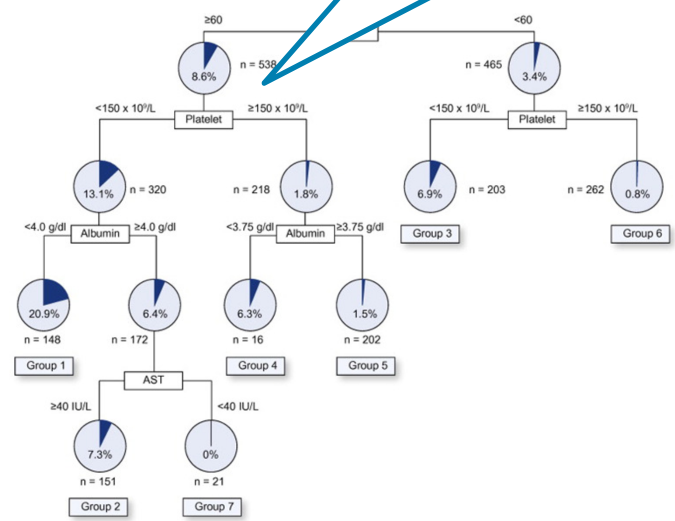
Bias in Models

- **Models *generalize* based on incomplete data**
  - **Make mistakes**
  - **Mistakes often asymmetric**

Gender	Drinks & drives	Likes to speed	High risk?
M	Y	Y	Y
M	N	Y	Y
M	Unknown		N
F	N	Y	Y
F	N	N	N
F	N	N	N

Will be even more biased

Models

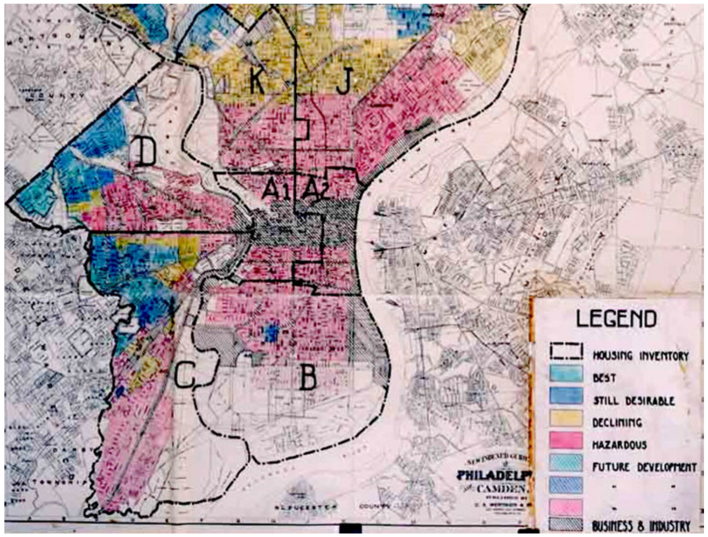


Redlining

Observation:

- **Just removing the *sensitive attributes* does not help**

- **Other attributes may be highly correlated with the sensitive attribute:**
  - **Gender ←→ Profession**
  - **Race ←→ Postal code**
  - **…**



Standard solution: Remove sensitive attributes

Example: Credit scoring dataset

% acceptance difference
males/females

Original data

	male	female
loan	3256	590
no loan	7604	4831

19%

Predictions

	male	female
loan	4559	422
no loan	6301	4999

31%

Predictions not based on gender

	male	female
loan	4134	567
no loan	6726	4854

28%

Legal / Ethical Restrictions

Legal/ethical restrictions

Integrate into a decision procedure

If lenders think that race is a reliable proxy for factors they cannot easily observe that affect credit risk, they may have an economic incentive to discriminate against minorities.

Thus, denying mortgage credit to a minority applicant on the basis of minorities on average-but not for the individual in question-may be economically rational.

But it is still discrimination, and it is illegal.

Source: "*Mortgage lending discrimination: a review of existing evidence.*" Report of *The Urban Institute*

Economic Incentives

Google ads discriminate against African-Americans: study

TU THANH HA
The Globe and Mail
Published Wednesday, Feb. 06 2013, 11:02 AM EST
Last updated Wednesday, Feb. 06 2013, 1:38 PM EST

- When Harvard computer scientist Latanya Sweeney types her name into Google, the search result yields an Instant Checkmate ad, titled “Latanya Sweeney Arrested?”

- Searching for “Kristen Lindquist” or “Jill Foley,” however, produced more neutral results

Source: The Globe and Mail (Feb. 6, 2013)

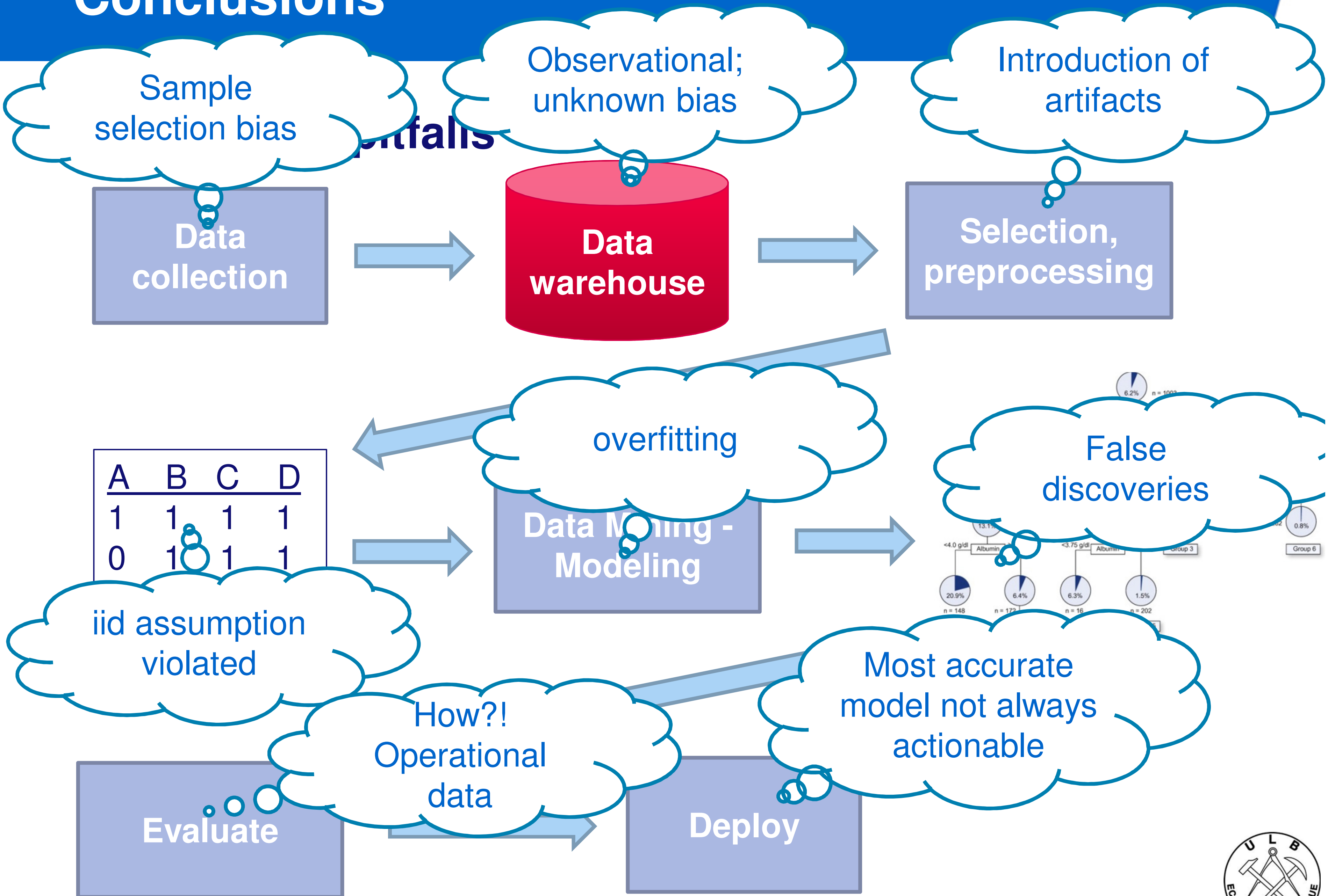
Skip Bayesian

Conclusions

- Data mining as a way to automatically derive models
  - Not as mature as statistics
  - Proof of the pudding is in the eating

- Many useful applications:
  - Spam detection
  - More efficient policing
  - Automatic model building
  - Tax evasion
  - …

Conclusions



Conclusions

- Bottom line:
  - Use common sense; do not fire your statistician
  - There is no golden bullet
  - Evaluate your models in a realistic setting

- As a data miner
  - Keep the scenario in mind
    - Are there biases in my data?
    - Can I apply the preprocessing to new instances?

- As a manager: do not trust the data miners
  - Separate evaluation from model selection

Thank You for Your Attention!