

The Dangers of Data Mining

Toon Calders
TU Eindhoven

TU / **e** Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Motivation for Data Mining: the Data Flood

- Huge amounts of data are available in digital form
 - Internet
 - IP Traffic logs
 - Scientific data
 - Customer profiles
 - ...
- No longer possible to analyze the data “manually”

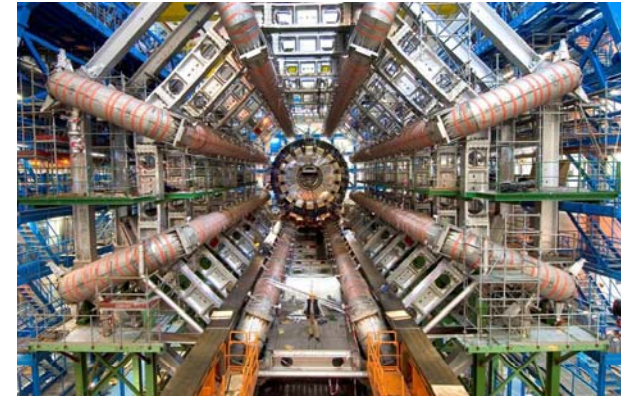


We are flooded by data but starving for information

Jiawei Han

Some Examples of Data Size

- **155 million websites on the Internet**
- **Large Hadron Collider produces 15 petabytes of data per year**
- **1 billion text messages sent every day (USA, 2007)**
- **294 billion emails sent per day (Radicati group, 2010)**



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**

A screenshot of a search engine's auto-complete feature. The search bar contains the text 'knowledge discovery'. Below the search bar, a dropdown menu displays several suggestions: 'knowledge discovery', 'knowledge dictionary', 'knowledge discovery in databases', 'knowledge discovery and data mining', and 'knowledge district'. The dropdown menu is enclosed in a light blue border. To the right of the search bar, there is a small 'X' icon and a 'Search' button.

About 10,500,000 results (0.20 seconds)

[Advanced search](#)

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

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, ...

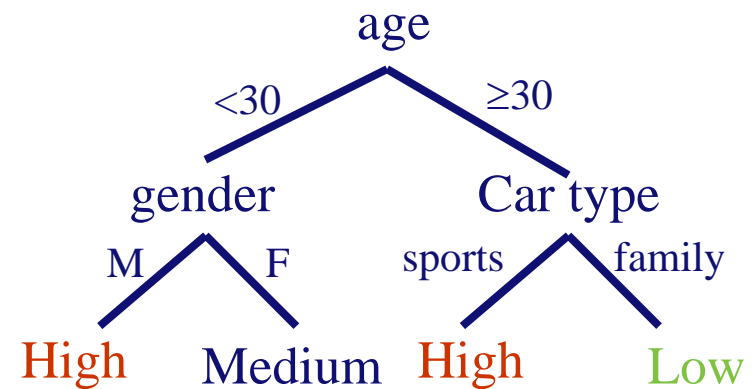
Can You See the Pattern?

```
110000000100000001000011001000000011000000000000010000100
11100000010000000100000110000000000001010000000000010000000
1110000001000000000000011000000000011000000000000010000000
010111000000110111011100100110011100011111100001000000000
000111000000110101011100000110011101011111100001000000000
1110000001000000000000011000010000011001000000000010100000
1100000001000000000000011000000000011000000000000010000000
111100101100001000000011111000100011000000011000110000000
000111000000110111011100000110011100011111000001000000000
000111011011110111111100000111111000111111100111001100000
0000001000000011000000111110000000000000001011000010000000
000000111011110011000000000001100001100000000111001100000
000000111011110011100000000001100000100000000111001100000
000100001011110011100001000001100000100010000111001100000
000000011010110010110110000001100010100000000111001100000
000000100000001000010101111000000000000000011000110000000
000000100010001000000011111000000000000000010000111000000
001000100000001000000010111000000000000000011000110000000
000111000000110111011100000110011100011111100001000000000
000111000000110111011100000110011100011111100001000010000
000000100000001000000011111000000000000000011000110000000
```


Example 1: Classification

- Learn a *model* based on *labeled* data.
- The model can be used for *prediction*.

Example:



Example 2: Clustering

- **Example:**

 [Advanced Search](#)
[Preferences](#)

[George W. Bush - Wikipedia, the free encyclopedia](#)

Open-source encyclopedia article provides personal, business and political information about the President, his policies, and public perceptions and ...

en.wikipedia.org/wiki/George_W_Bush - 459k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Bush \(band\) - Wikipedia, the free encyclopedia](#)

Bush was a post-grunge band from the UK, formed in 1992. Their debut album was the self-released Sixteen Stone in 1994. They have sold well over 10 million ...

[en.wikipedia.org/wiki/Bush_\(band\)](http://en.wikipedia.org/wiki/Bush_(band)) - 60k - [Cached](#) - [Similar pages](#) - [Note this](#)

[More results from en.wikipedia.org »](#)

[President of the United States - George W. Bush](#)

The Oval Office contains speeches and statements of President **Bush**, a description of policy priorities, biographies, and photo essays.

www.whitehouse.gov/president/ - 21k - [Cached](#) - [Similar pages](#) - [Note this](#)

[More results from www.whitehouse.gov »](#)

[Gavin Rossdale: gavinrossdalefans.com](#)

The former lead singer of **BUSH**, the platinum selling alt rock juggernaut, Gavin can now be seen UP CLOSE at this intimate Past Show. ...

gavinrossdalefans.com/ - 38k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Bush Furniture, Inc](#)

Bush designs and manufactures quality, ready to assemble, entertainment centers, TV stands, home office and business furniture.

www.bushfurniture.com/ - 26k - [Cached](#) - [Similar pages](#) - [Note this](#)

Example 3: Pattern Mining

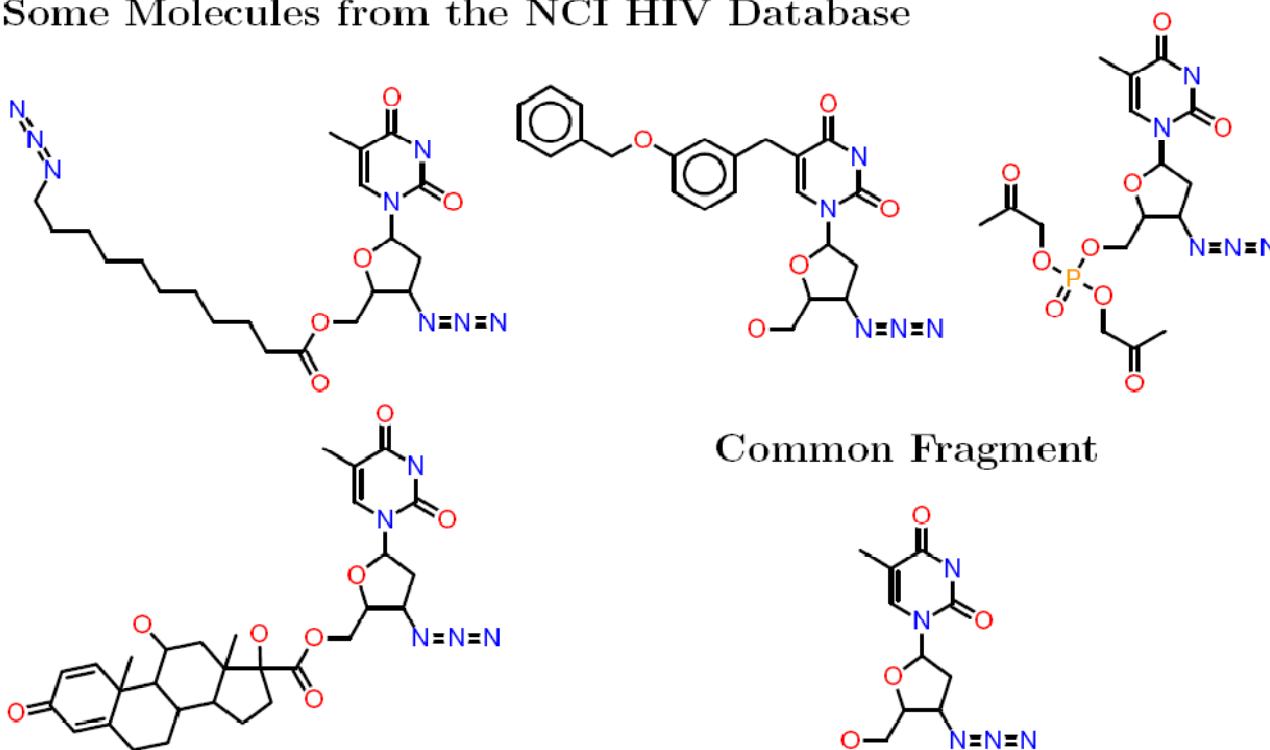
- Find regularities, trends, patterns that frequently occur in the data



Example 4: Pattern Mining

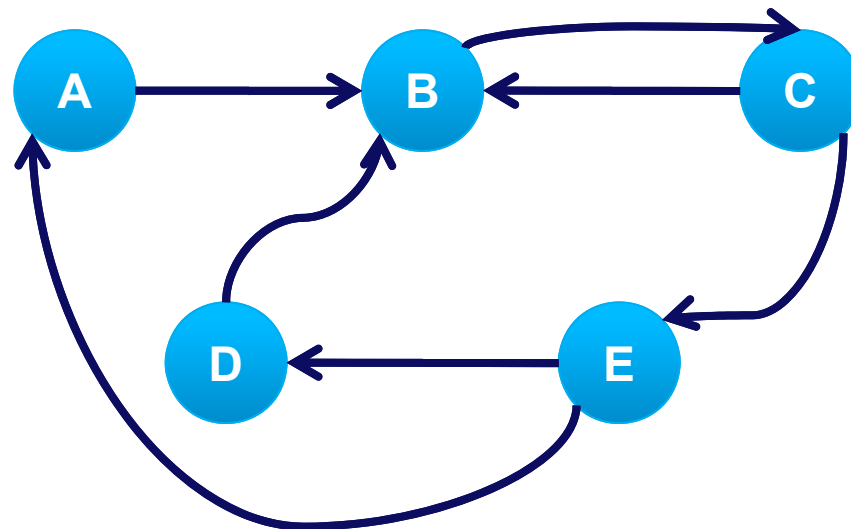
- Other example:

Some Molecules from the NCI HIV Database



Example 5: Importance of Webpages

- **Pagerank as used by google**
 - **Page structure implicitly holds importance of a page**
 - **Important pages are linked to by important pages**



Examples of Mining on the Web

- **Sentiment analysis of twitter messages**
 - market study
- **Analyze web-site visitors; what makes a customer leave your webpage?**
- **What type of news articles do you like?**
 - Personalized suggestions
- **Find influential persons in a community of bloggers**
 - Targets for viral marketing

Examples of Mining Customer Data

- **Which customers are likely to “churn”**
 - **Concentrate on these customers**
- **What kind of customers like a specific offer?**
 - **Up-lift modeling**
- **Which promotions to offer to a customer?**
 - **Products he/she likes**
 - **Products people with a similar profile like**

Examples of Data Mining in Policing

- **Datamining 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)

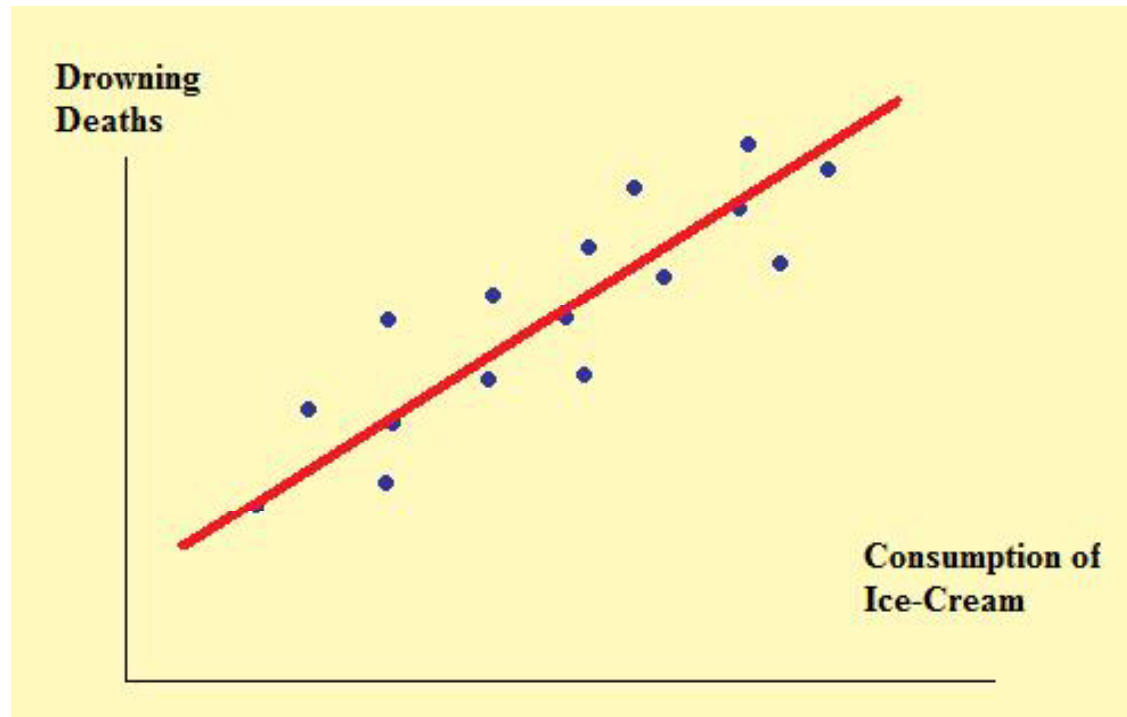
The Dangers of Data mining

However, the use of data mining also has some potential dangers:

- **Interpretation of results**
 - Implication is not causality
 - Simpson's Paradox
- **Privacy issues**
- **False discoveries**
- **Discriminating models**

Correlation \neq Causality

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



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%

Danger 1: Privacy Issues

- **First step in data mining:**
 - **Collect data**
 - **Buying history (Loyalty cards)**
 - **Blog posts, Twitter messages**
 - **Browsing behavior**
 - **Contest**
 - **Buy the data from a company**
 - **Survey data; e.g. CBS (average income/postal code)**
- **Combining different sources may lead to privacy issues**
 - **Information asymmetry**



Why is Facebook Worth 10bn\$?

Source: Facebook's Privacy Policy:

- **Information we collect when you interact with Facebook:**
 - **Site activity information, ...**
 - **“We may ask advertisers to tell us how our users responded to the ads we showed them”**
- **We allow advertisers to choose the characteristics of users who will see their advertisements and we may use any of the non-personally identifiable attributes we have collected (including information you may have decided not to show to other users)**

Danger 2: 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

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.**



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)

Danger 3: Discriminating Models

- Often we observe classifiers learn undesirable properties from data ...

[For Candidates](#) [For Companies](#) [Blog](#)



[Home](#) » [Services](#) » [Sign Up](#)

Your Profile

[Logout](#) [Toon Calders](#)

Change your profile

You can use this form to change your profile details and manage the CV's and documents attached to it.

Jobs

- > [Jobs per company](#)
- > [Find a job](#)

Your details

★ Preferred Kind of Work

I'm looking for my first job

Blue Collar



Classification

- Often we observe classifiers learn undesirable properties from data ...

Gender	Age	Diploma Candidate	Nationality	...	City	Required Diploma	Invite?
M	29	MSc. Math	Belgian	...	Antwerp	MSc. CS	y
F	49	BSc. CS	Turkish	...	Eindhoven	BSc. CS	n
M	32	HBO	Dutch	...	The Hague	-	y
...

[Home](#) » [Services](#) » [Sign Up](#)



Your Profile

Change your profile

You can use this form to change your profile details and manage the CV's

(Gender = F) and (Job_type = "full professor") \Rightarrow (Invite = No)

(Nationality = "Moroccan") or (Nationality="Turkish") \Rightarrow (Invite = No)

Jobs

- > [Jobs per company](#)
- > [Find a job](#)

Your details

★ Preferred Kind of Work

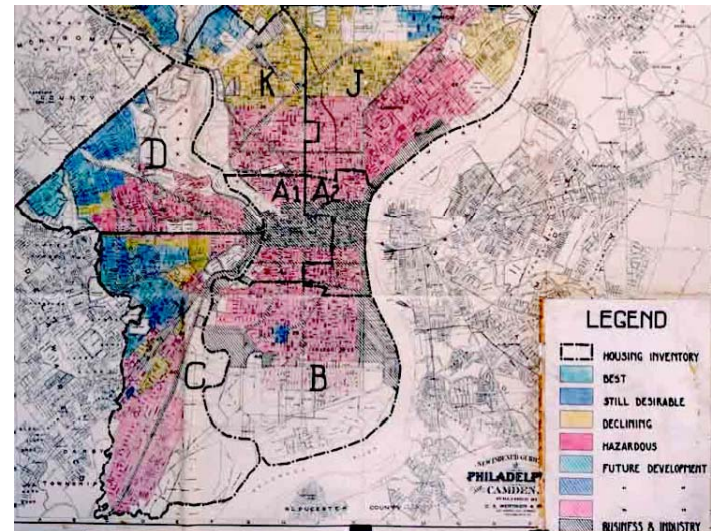
I'm looking for my first job

Blue Collar

Redlining

Observation:

- Just removing the *sensitive attributes* does not help
- Other attributes may be highly correlated with the sensitive attribute:
 - Gender \leftrightarrow Profession
 - Race \leftrightarrow Postal code
 - ...



Redlining

Example: Credit scoring dataset

Original data

	male	female
loan	3256	590
no loan	7604	4831

19%

Predictions using gender

	male	female
loan	4559	422
no loan	6301	4999

31%

Predictions without gender

	male	female
loan	4134	567
no loan	6726	4854

28%

What's Actually the Problem?

Are men better drivers?



"All the evidence points to young males having riskier driving habits than young females. Men between the ages of 16 and 25 are much more likely to be involved in accidents, or be cited for traffic violations."

Sam Belden, Insurance.com VP

- **It's nothing personal, it's just statistics!**

Problem 1: You'll Get Convicted

- **March 1, 2011 European Court of Justice ruling in Test-Achats (C-236/09):**

“Taking the gender of the insured individual into account as a risk factor in insurance contracts constitutes discrimination. The rule of unisex premiums and benefits will apply with effect from 21 December 2012.”

Quote: M.A. Turner, F. Skidmore

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*

Problem 2: Imbalance in Errors

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

Problem 2: Imbalance in Errors

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

- **2 new customers arrive:**
 - **Non-drinking, non-speeding male** → risk = 66%
 - **Drinking, speeding female** → risk is 33%

Discrimination-Aware Data Mining

- **A lot of our current work deals with this last problem**
 - **Identify discrimination in data**
 - **Remove discrimination from data**
 - **Learn non-discriminatory models from discriminatory data**
 - **Clean up discriminatory models**
- **Governmental partners:**
 - **Central Bureau of Statistics**
 - **WODC – Study Center of the Department of Justice**

Conclusion

- **Data mining = using automatic techniques to find patterns in data**
- **Many useful applications:**
 - **Spam detection**
 - **More efficient policing**
 - **Automatic model building**
- **However, also dangers!**
 - **Privacy issues**
 - **False discoveries**
 - **Discrimination**

Thank you for your attention!

**Many thanks to the collaborators of the project
“Discrimination-Aware Data Mining”**

