Predictive Data Mining and Big Data Analytics

Prof. dr. Bart Goethals

Advanced Database Research & Modelling

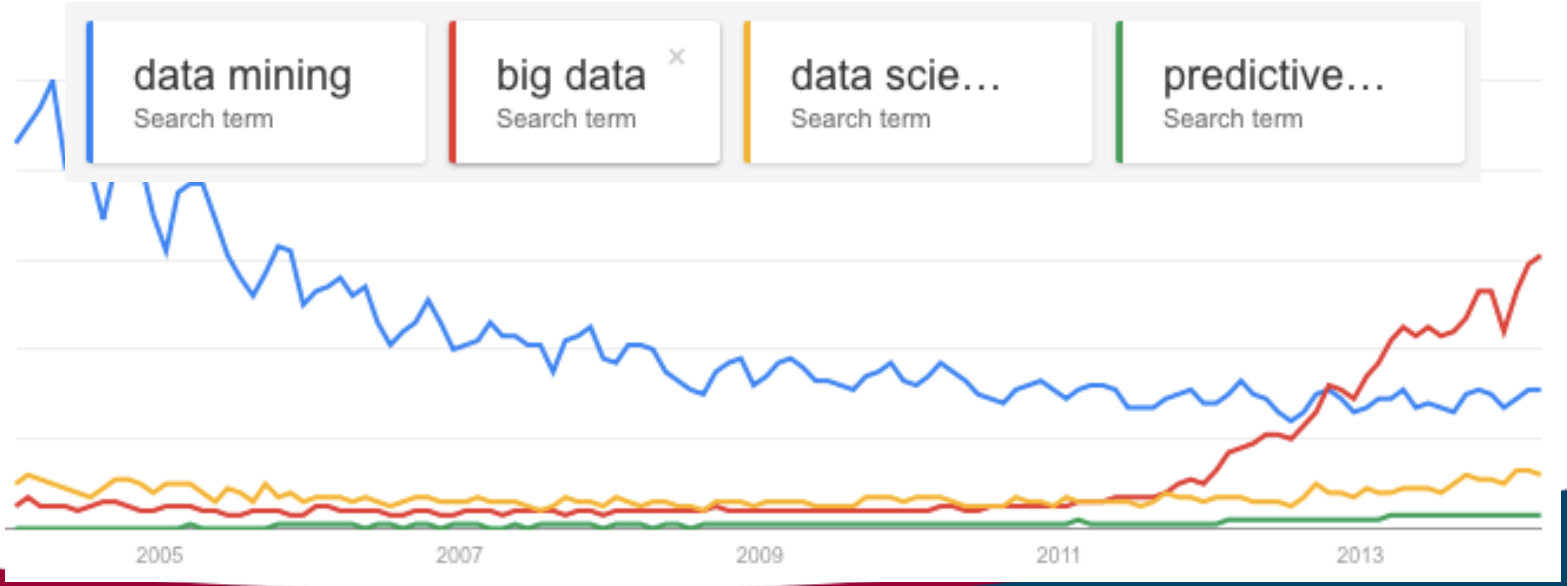
Department of Mathematics & Computer Science

Universiteit Antwerpen

Big Data or …

- Statistics
- Data Mining
- Knowledge Discovery in Data
- Analytics
- Data Science
- …

Big Data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it…
[Dan Ariely]



Universiteit Antwerpen

Big Data Goal

Goal is the same:

Find **useful** patterns or models in Data

Emphasis Changes:

Volume

Velocity

Variety

V…

Universiteit Antwerpen

Is Big a problem?

- Data can (not) be summarised (sampling)
  - Too much information lost for reasonable sizes
  - We need to find patterns that are useful and valid for all data
    - **Personalized** Recommendation
    - **Personalized** Advertising
    - **Rare** diseases
- Current methods do not scale or produce satisfactory results

Universiteit Antwerpen

Big Data Velocity & Variety & V…

- Twitter (> 12 Tb of tweet data daily)
- Facebook (> 25 Tb of log data daily)
- …

- Data can be structured, semi-structured, text, images, video, time series, click-streams, graphs or (social) networks, …

Applications?

- Predict voting behaviour based on Twitter (~20M tweets)
- Detect Fiscal Fraud based on network of ~7M transactions
- Recognise cyberpedophiles
- e-Health, predict rare diseases
- Mining train delays (sources)
- Personalised Advertising, Recommendation,Cross-selling, Product placement, Distribution planning
- …

Universiteit Antwerpen

What about the methods?

- Association-, Pattern Discovery
- Classification, Prediction, Regression
- Clustering
- Recommendation
- Exploration
- Summarization
- Visualization

Universiteit Antwerpen

Association-, Pattern Discovery

- Imagine a supermarket
- What sets of products frequently bought together?
- What products influence the sales of each other?

Universiteit Antwerpen

Challenge

- Number of potentially interesting patterns is larger than the number of particles in the universe

Universiteit Antwerpen

Association-, Pattern Discovery

- “75% of all customers that buy diapers also buy beer”

Universiteit Antwerpen

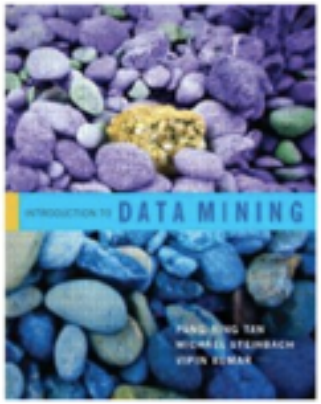
amazon.com

Hello. Sign in to get personalized recommendations. New customer? Start here.

Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search | Books

Books | Advanced Search | Browse Subjects | New Releases | Bestsellers | The New York Times® Bestsellers



See larger image

Share your own customer images

Publisher: learn how customers can search inside this book.

Introduction to Data Mining (Hardcover)

~ Pang-Ning Tan (Author), Michael Steinbach (Author), Vipin Kumar (Author)

(15 customer reviews)

List Price: ~~$105.00~~

Price: $80.80 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $24.20 (23%)

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Friday, December 4? Order it in the next 4 hours and 7 minutes, and choose **One-Day Shipping** at checkout. Details

Ordering for Christmas? To ensure delivery by December 24, choose **FREE Super Saver Shipping** at checkout. Read more about holiday sh

22 new from $75.00 15 used from $68.00

Frequently Bought Together

Price For All Three: $189.60

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

- **This item:** Introduction to Data Mining by Pang-Ning Tan
- Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems) by Eib
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Robert Tibshirani

Customers Who Bought This Item Also Bought

Different patterns for different data

- Patients, symptoms, diseases
- Movies, ratings, viewers
- Friends, Likes, Status Updates, Interactions
- Routes, Trucks, Packages, Distributors, Locations

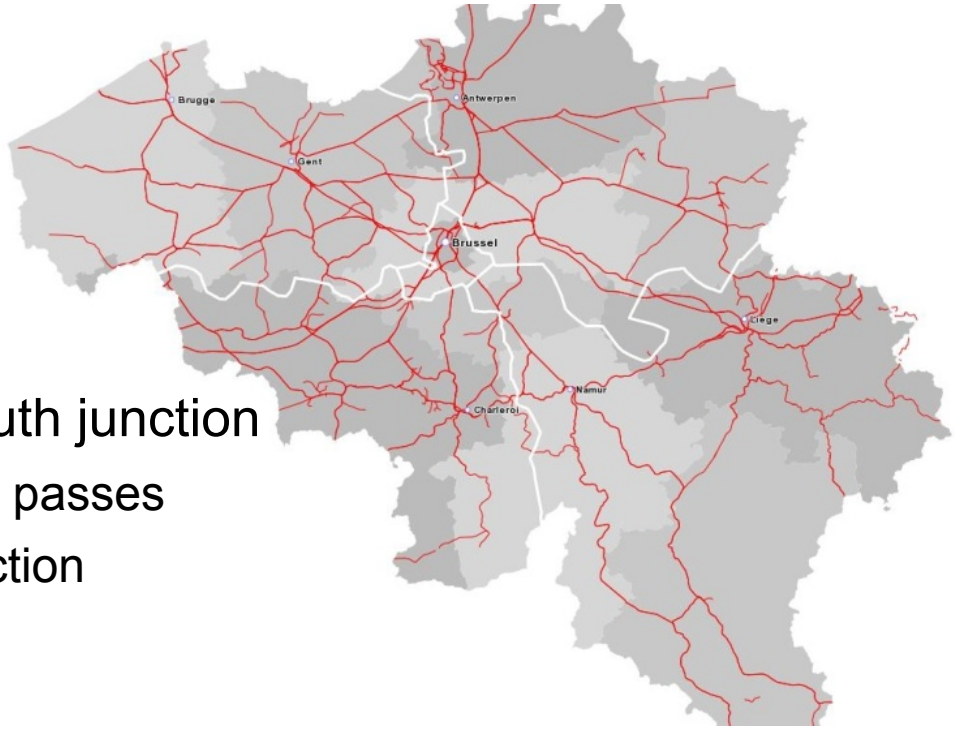
- Sequences, spatial, time series, graphs, multi-relations, RDF, …

Universiteit Antwerpen

INFRABEL
Right On Track

Universiteit Antwerpen

Mining train delays



- Very dense network
- Bottleneck in Brussels north-south junction
  - Approximately 40% of daily traffic passes through Brussels north-south junction
  - Only 6 tracks
- Star-shaped structure
  - 6 main lines come together in Brussels
- Many bifurcations, often on short distances
- Punctuality decreasing

INFRABEL Right On Track

Universiteit Antwerpen

Pattern mining

- Itemsets
  - Simplest pattern
  - Events that occur together
  - Itemsets applied to railway data
    - Moving time window of delayed trains
    - Itemset = trains that are together in a window
      (i.e. trains that are late within the same time frame)
    - Support = The number of windows an itemset can be found in
    - Frequent itemset = Itemsets with support ≥ threshold

INFRABEL Right On Track

Universiteit Antwerpen

example

Day 1	Time	Day 2	Time	Day 3	Time
…	…	…	…	…	…
Train A	07:05	Train B	07:05	Train A	07:05
Train B	07:06	Train G	07:06	Train B	07:06
Train E	07:07	Train F	07:07	Train D	07:07
Train G	07:08	Train E	07:08	Train E	07:08
Train H	07:09	Train H	07:09	Train G	07:09
Train K	07:10	Train J	07:10	Train I	07:10
Train N	07:11	Train M	07:11	Train K	07:11
…	…	…	…	…	…

- Support {B} is 12
  - 4 in day 1
  - 4 in day 2
  - 4 in day 3
- Support {B,E} is 6
  - 3 in day 1
  - 1 in day 2
  - 2 in day 3
- Support {B,E,G} is 4
  - 2 in day 1
  - 1 in day 2
  - 1 in day 3
- …

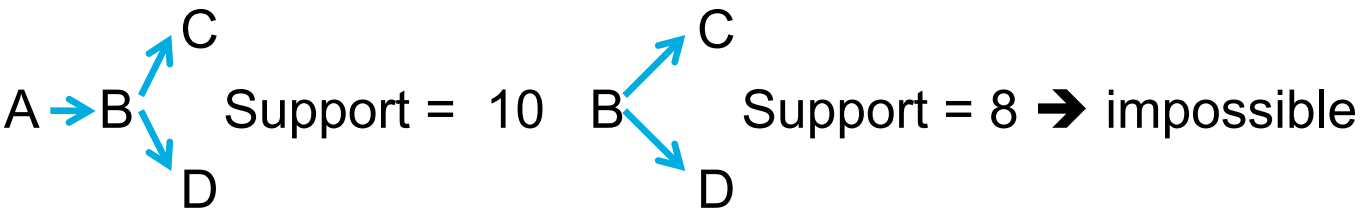
example

- Episodes
  - Frequent itemset with additional information
  - events often occur together in a specific order
  - Example:
    - Support Episode B→E→G is 3
      - No support on day 2
    - Support Episode B→E is 4
      B→G
      - Support day 2 is 1

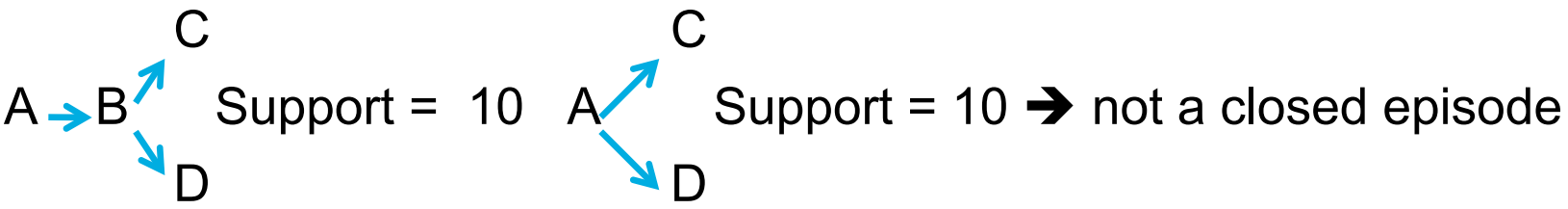
Day 1	Day 2	Day 3
…	…	…
Train A	Train B	Train A
Train B	Train G	Train B
Train E	Train F	Train D
Train G	Train E	Train E
Train H	Train H	Train G
Train K	Train J	Train I
Train N	Train M	Train K
…	…	…

Closed episodes

- Used to reduce output
- If pattern is frequent → sub-pattern is also frequent



  - If support pattern is equal to support sub-pattern → sub-pattern gives no extra information



  - An episode is closed if none of its super episodes have the same support.

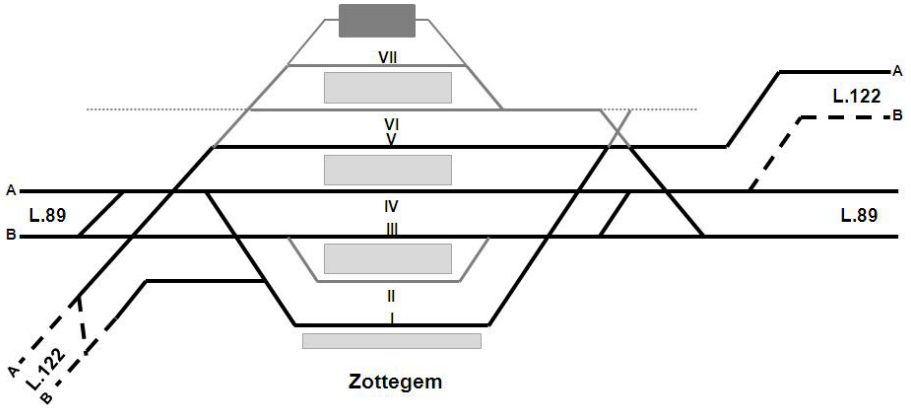
INFRABEL Right On Track

Universiteit Antwerpen

Experiment

- Approximately 1800 characteristic points
  - Stations
  - Bifurcations
  - Country borders
  - …
- Database with timestamps of trains in characteristic points

- Algorithm applied to station of Zottegem
  - Medium-sized station
  - Crossing of 2 lines (L.89 and L.122)
  - Intelligible infrastructure
  - #trains in January 2010 was 4412
  - Many peak-hour trains disrupting the regular schedule
- If applied over entire network, too much false patterns would arise
- 2 definitions of delay:
  - 3 minutes late
  - 6 minutes late



INFRABEL Right On Track Universiteit Antwerpen

Example results

- Window size is 1800 seconds (half hour)
- 4 different sizes of episodes
  - **Size(1,0) → 1 node, 0 edges**
  - Size(2,k) → 2 nodes, k edges (k can be 1 or 0)
  - Size(3,k) → 3 nodes, k edges
  - Size(4,k) → 4 nodes, k edges

Train ID	Route	Support	
		Delay ≥ 3'	Delay ≥ 6'
1867	Zottegem-Kortrijk	27000	14400
8904	Schaarbeek-Oudenaarde	28800	18000
8905	Schaarbeek-Kortrijk	27000	14400
8963	Geraardsbergen-Ghent-SP	25200	12600

Support divided by window size → Number of days train was delayed

3.3 Example results

- Window size is 1800 seconds (half hour)
- 4 different sizes of episodes
  - Size(1,0) → 1 node, 0 edges
  - **Size(2,k) → 2 nodes, k edges (K can be 1 or 0)**
  - Size(3,k) → 3 nodes, k edges
  - Size(4,k) → 4 nodes, k edges

Train ID	Episode Relation	Train ID	Support Delay ≥ 3'	Delay ≥ 6'
1867		8904	15079	-
1867	←	8904	13557	-
8904	→	8905	18608	9506
8905		8963	20580	10608
…				

Episode takes place on min(support / 1800) days

INFRABEL Right On Track

Universiteit Antwerpen

Example results

- Window size is 1800 seconds (half hour)
- 4 different sizes of episodes
  - Size(1,0) → 1 node, 0 edges
  - Size(2,k) → 2 nodes, k edges (K can be 1 or 0)
  - **Size(3,k) → 3 nodes, k edges**
  - **Size(4,k) → 4 nodes, k edges**

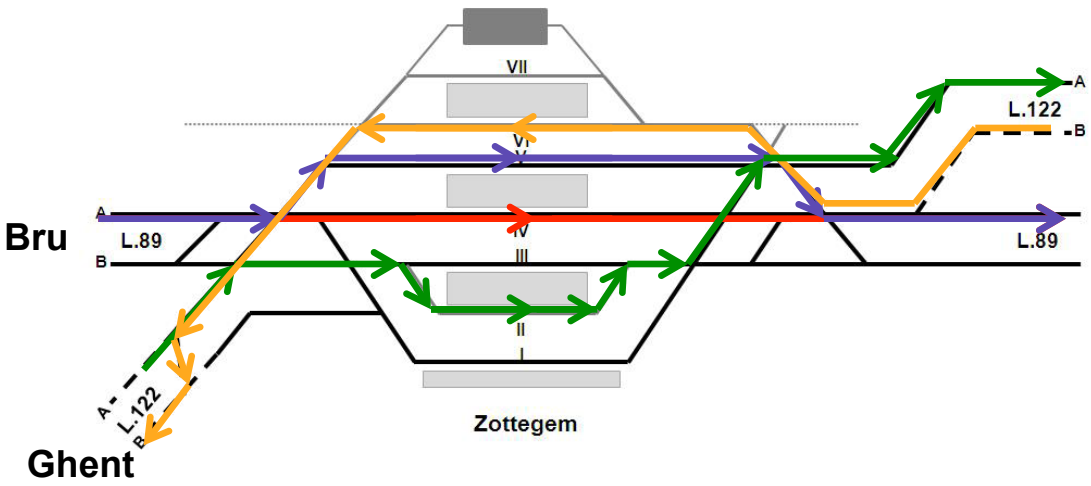
Size(3,2)	Support Delay ≥ 3’	Delay ≥ 6’	Size(4,3)	Support Delay ≥ 3’	Delay ≥ 6’
8904 → 8905, 8963	14358	7510	8904 → 1867, 8905, 8963	10024	6104

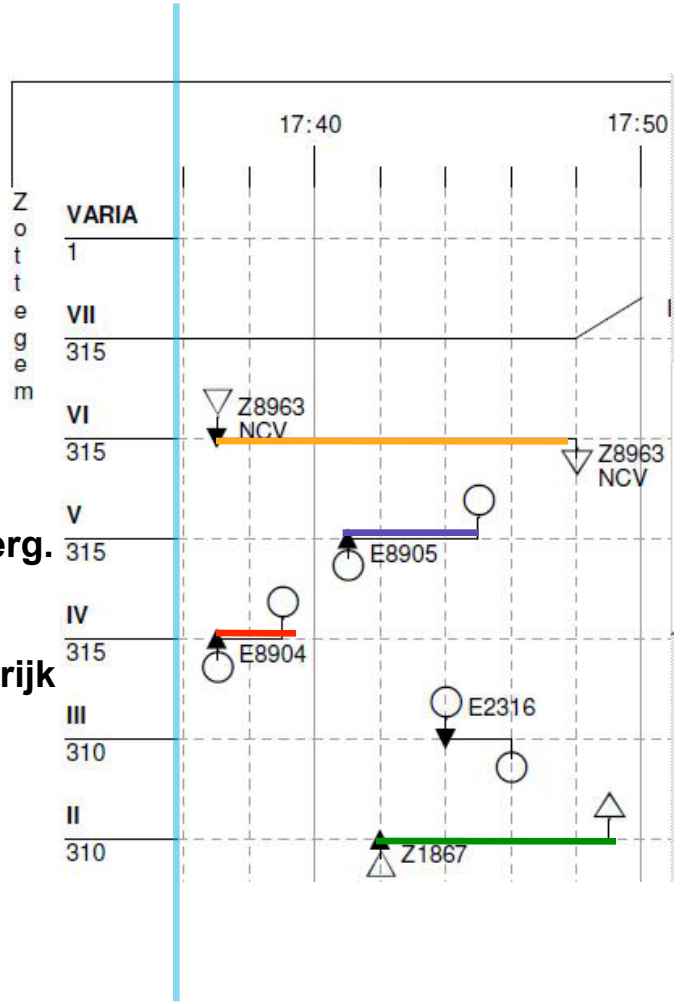
Episode takes place on min(support / 1800) days

3.4 Comparison with reality

8904 → 8905

8904 → 1867 / 8905 / 8963

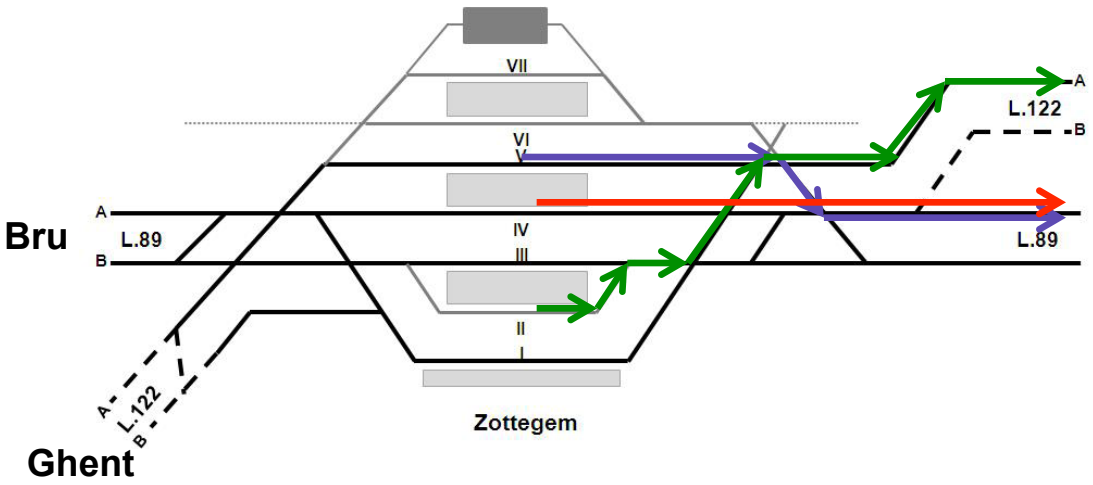


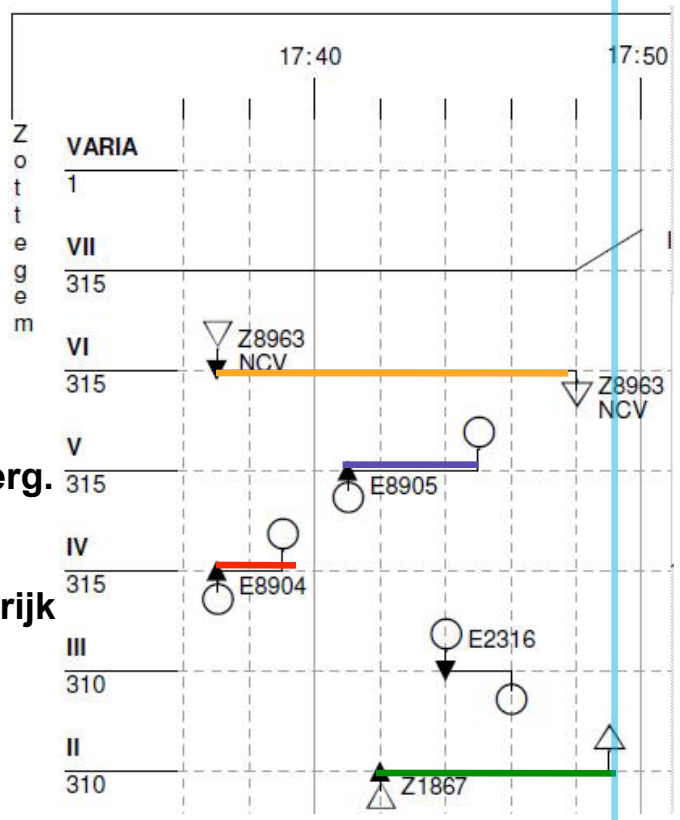


3.4 Comparison with reality

8904 → 8905

8904 → 1867 / 8905 / 8963





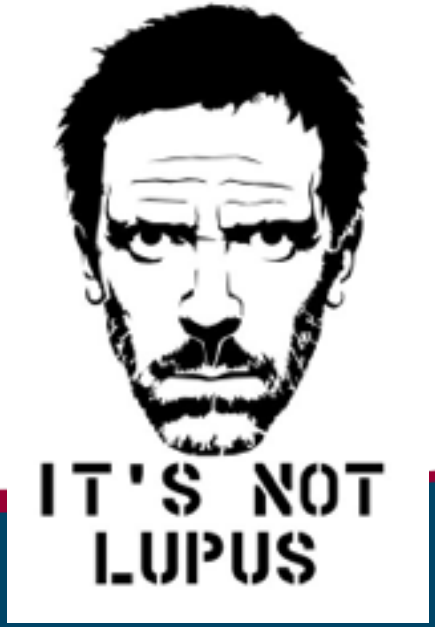
Classification / Prediction

- How to separate two classes of objects from each other

Universiteit Antwerpen

Rare diseases

- Neonatal heel prick used for detection of potential Medium-chain acyl-coenzyme A dehydrogenase deficiency
- Classify whether expensive genetic test is required
- Intensive Care, fast prediction of e.g. kidney failure



Universiteit Antwerpen

	Recall	Precision	F_1
CD & V	95.83%	100.00%	97.87%
Groen!	98.36%	88.24%	93.02%
LDD Nationaal	60.00%	50.00%	54.55%
N-VA	100.00%	94.59%	97.22%
OpenVLD	96.77%	98.36%	97.56%
PvdA	61.54%	100.00%	76.19%
SP.a	98.40%	97.62%	98.01%
Vlaams Belang	73.33%	100.00%	84.62%
M (macro)	85.53%	91.10%	88.23%
μ (micro)	95.00%	95.00%	95.00%

Table 4.7: Relevance measures for individual classes and aggregated variants in $V_{selected}$ for the second experiment.

Classification methods

- Pattern Based Classification
- Nearest Neighbour Classification
- Decision Trees
- Support Vector Machines
- Neural Networks
- Random Forsests
- Conditional Random Fields
- …

Universiteit Antwerpen

Recommendation methods

- A customer arrives on your web-shop: show her the product she doesn't know yet, but might be interested in
- For Any (online) shop!
  Famous example: Netflix
  (pattern mining is even used to produce new series: 'House of Cards')
- Methods:
  - Collaborative Filtering
    - Find most similar users/customers
      - similar buying behaviour
      - similar social network
      - …
    - Recommend objects specific for that group, but missing from target
  - Matrix Factorisation

Universiteit Antwerpen

Clustering: grouping similar things together

Universiteit Antwerpen

Conclusion

Universiteit Antwerpen

Garbage in - Garbage out

http://www.uantwerpen.be/bart-goethals
bart.goethals@uantwerp.be

Universiteit Antwerpen