Data mining
(and machine learning)

Rasmus Pagh

Some figures for slide set of Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition.

Today's lecture

- What is data mining?
- Crash course in statistics
- Models for identifying patterns

- Examples of data mining models:
  - Decision trees
  - Association rules
  - Similarity and clustering
  - Non-negative matrix factorization

What is data mining?

- One definition:
  "Data mining is the (automated) exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data."

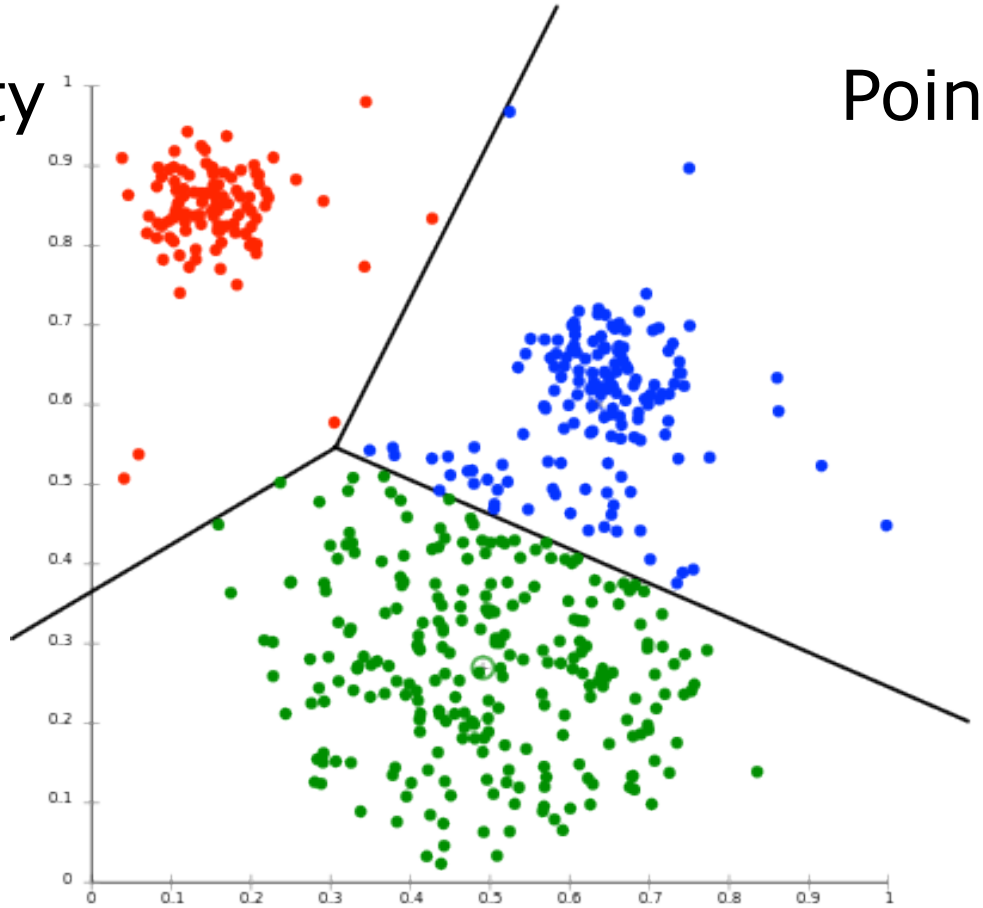
- Main differences from OLAP:
  - Little or no human interaction.
  - "Queries" are more weakly specified.

Example: Prediction and Targeting

- US retail chain, 365000 employees.
- Collects and buys data on their costumers, e.g.: Purchases made, coupons used, residence, salary estimate, marital status, ethnicity, real estate trades,…
- Growth idea, 2002: Special ads for women in mid-pregnancy. But how?

Example: Market segmentation

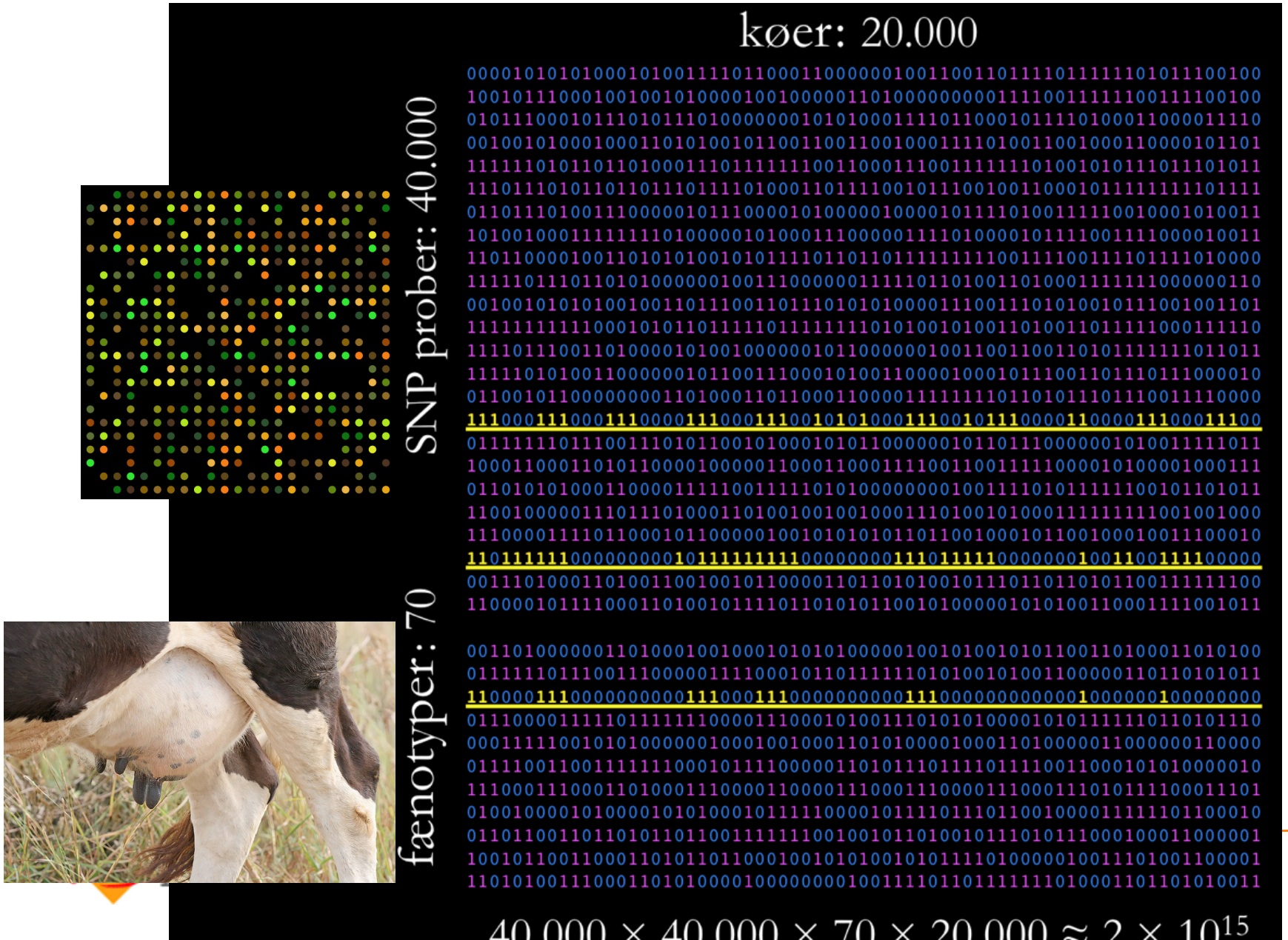
House equity



Point = customer

Monthly salary

Example: Scientific applications



Why data mining?

- "The secret of success is to know something that nobody else knows"
  - Aristotle Onassis

- In recent years data mining has moved more towards *machine learning*, which aims at *predictive models* of data. E.g.
  - Will I make money on this customer?
  - What is this customer likely to buy?
  - Will this patient respond well to treatment?

Crash course on statistics

- Basic premise:
  - Sufficiently many observations of a phenomenon will reveal cause-and-effect relationships.
  - More generally, will reveal correlations.
- Example:
  Hundred observations of dice throw:

```
10  7  8  11  12  8  10  8  8  12  8  9  5  7  6  3  8  8  7  6
8  8  11  11  8  7  9  11  5  10  7  7  10  5  10  7  4  12  8
11  12  8  9  8  9  6  7  4  6  4  12  10  8  6  11  9  5  7  10
7  9  11  10  8  8  7  5  6  12  6  7  12  7  9  10  3  5  5  7
8  4  7  8  7  7  7  4  10  7  5  9  9  9  11  6  11  7  9  6  10
```

How many observations are needed?

- Depends on whether the observations are independent.
  - Assume they are, otherwise things get complicated.
- If we see 7 throws of 12 in 100 tosses, what does that tell us about the probability of throwing a 12?

Normal approximation

- Common in statistics: Approximate random variable by normal distribution.
  - Easy to analyze probability to see a value close to the expected one.
  - Typically reports *confidence* in a certain maximum deviation.

Confidence	Deviation at most
0.7	$1.0364333894937896\,\sigma$
0.8	$1.2815515655446005\,\sigma$
0.9	$1.6448536269514727\,\sigma$
0.95	$1.9599639845400542\,\sigma$
0.99	$2.575829303548900 8\,\sigma$
0.995	$2.8070337683438041\,\sigma$

- If we observe something σ^2 times, the standard deviation is σ.

Normal approximation

- **Example**: We observed 7 throws of 12.
  - Standard deviation $\sigma = \sqrt{7} \approx 2.65$
  - With confidence 0.8 the deviation is at most 1.28 $\sigma \approx 3.4$ from expectation
  - I.e. we are 80% confident that 12s are thrown with probability in 3.6%-10.4%.

Confidence	Deviation at most
0.7	$1.0364333894937896\,\sigma$
0.8	$1.2815515655446005\,\sigma$
0.9	$1.6448536269514727\,\sigma$
0.95	$1.9599639845400542\,\sigma$
0.99	$2.575829303548900 8\,\sigma$
0.995	$2.8070337683438041\,\sigma$

- If we observe something σ^2 times, the standard deviation is σ.

Required number of samples

- How many samples n do we need to have confidence that n is within 10%=0.1 of the expected value?
  - For confidence .9, need about 270.

Confidence	Deviation at most	Required no. samples n for error at most εn
0.7	$1.0364333894937896\,\sigma$	$\frac{1.0741941708575853}{\epsilon^2}$
0.8	$1.2815515655446005\,\sigma$	$\frac{1.6423744151498164}{\epsilon^2}$
0.9	$1.644853626951472\,7\,\sigma$	$\frac{2.7055434540954146}{\epsilon^2}$
0.95	$1.959963984540054\,2\,\sigma$	$\frac{3.8414588206941260}{\epsilon^2}$
0.99	$2.575829303548900\,8\,\sigma$	$\frac{6.634896601021215}{\epsilon^2}$

Understanding observations

- What generated these?

```
10  7  8  11  12  8  10  8  8  12  8  9  5  7  6  3  8  8  7  6
8  8  11  11  8  7  9  11  5  10  7  7  10  5  10  7  4  12  8
11  12  8  9  8  9  6  7  4  6  4  12  10  8  6  11  9  5  7  10
7  9  11  10  8  8  7  5  6  12  6  7  12  7  9  10  3  5  5  7
8  4  7  8  7  7  7  4  10  7  5  9  9  9  11  6  11  7  9  6  10
```

- Would like a hypothesis, or *model*.
- Example models:
  - Data generated by throwing two dice.
  - Data generated by throwing two biased dice.
  - Data generated by hashing the first 100 words of a book.

Occam's razor, winner's curse

Occam's razor:
Among several possible explanations, the simplest one is preferable.

Winner's curse:

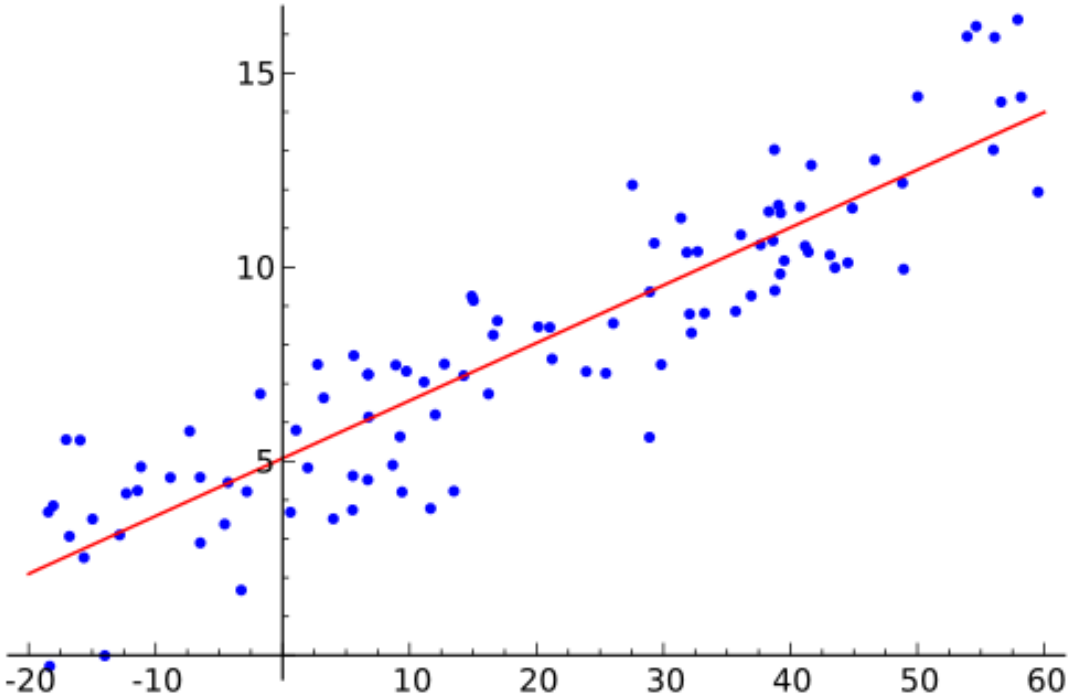
If we allow too many different candidate patterns (or models), we increase the probability that an observed pattern is due to chance.

Models used in data mining

Tons of them exist. Will talk about:

- Linear regression
- Decision trees
- Association analysis
- Clustering
- Matrix factorization (NMF)

Linear regression



Model: y=ax+b+”noise”
Typically: minimize mean squared error

Linear regression in d dimensions

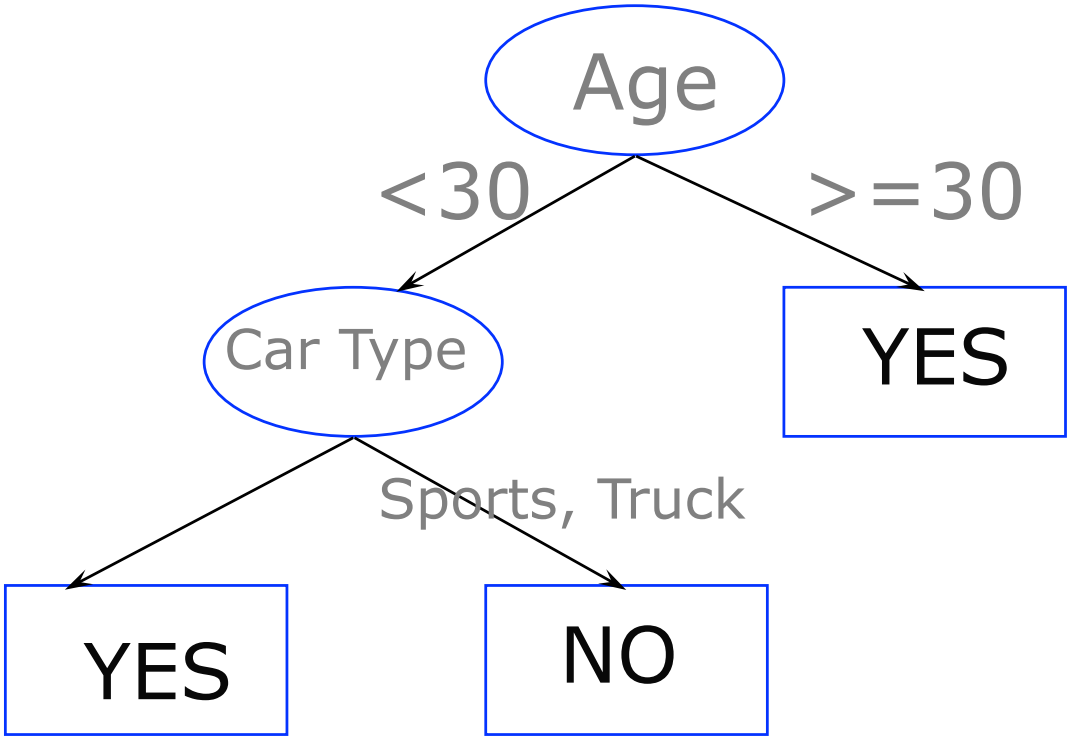
Model: $a_1x_1+a_2x_2+\ldots+a_dx_d=0+$"noise"

- Problem with Occam's razor for large d: Too many degrees of freedom (many models will be equally good)
- Tool: Regularization
  - Find the best model where $(a_1,\ldots,a_d)$ is "nice", e.g. $|a_1|+\ldots+|a_d|<1$.
  - How to apply: Far outside scope of class.
- E.g. Genome-Wide Association Studies

Decision trees

Used for categorization of data items/tuples.

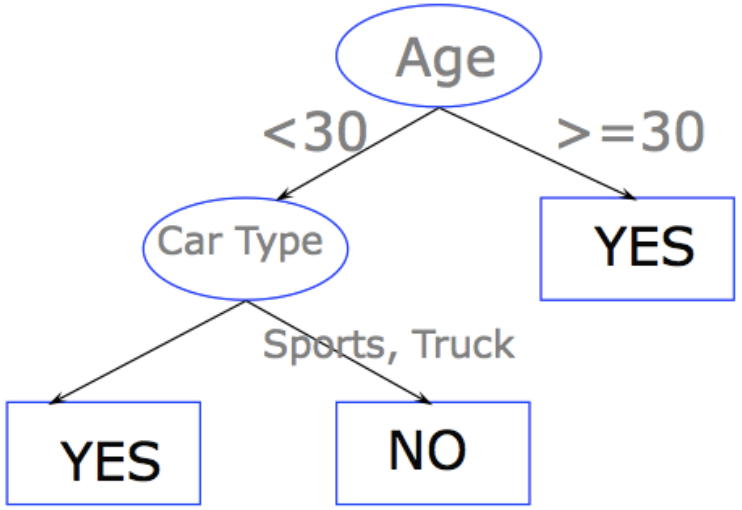
Example: Do I expect to make money by insuring this person's car?



Idea: Create automatically from historical data.

What decision tree is best?

- Splits known data is a "best possible" way
  - Outside of scope of this class.
- Ideally, need enough data along each path to be confident in the answer.
  - A too big/detailed decision tree will not achieve this ("overfitting").
- Cross-validation:
  - Build model based on part of the data.
  - Check validity on the other part.



Association analysis

Simplest case: Market basket analysis.

Association rules:

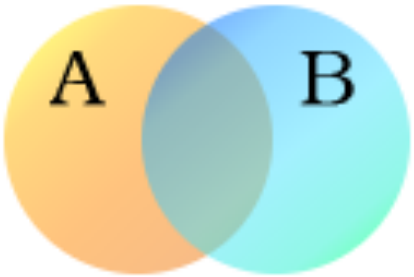
- {Pen} => {Milk}
  Support: 75%
  Confidence: 75%
- {Ink} => {Pen}
  Support: 100%
  Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Similarity measures

Idea: take into account the number of occurrences of an item to measure how "similar" the occurrence sets are.

Many possible measures:
Jaccard, lift, cosine, …

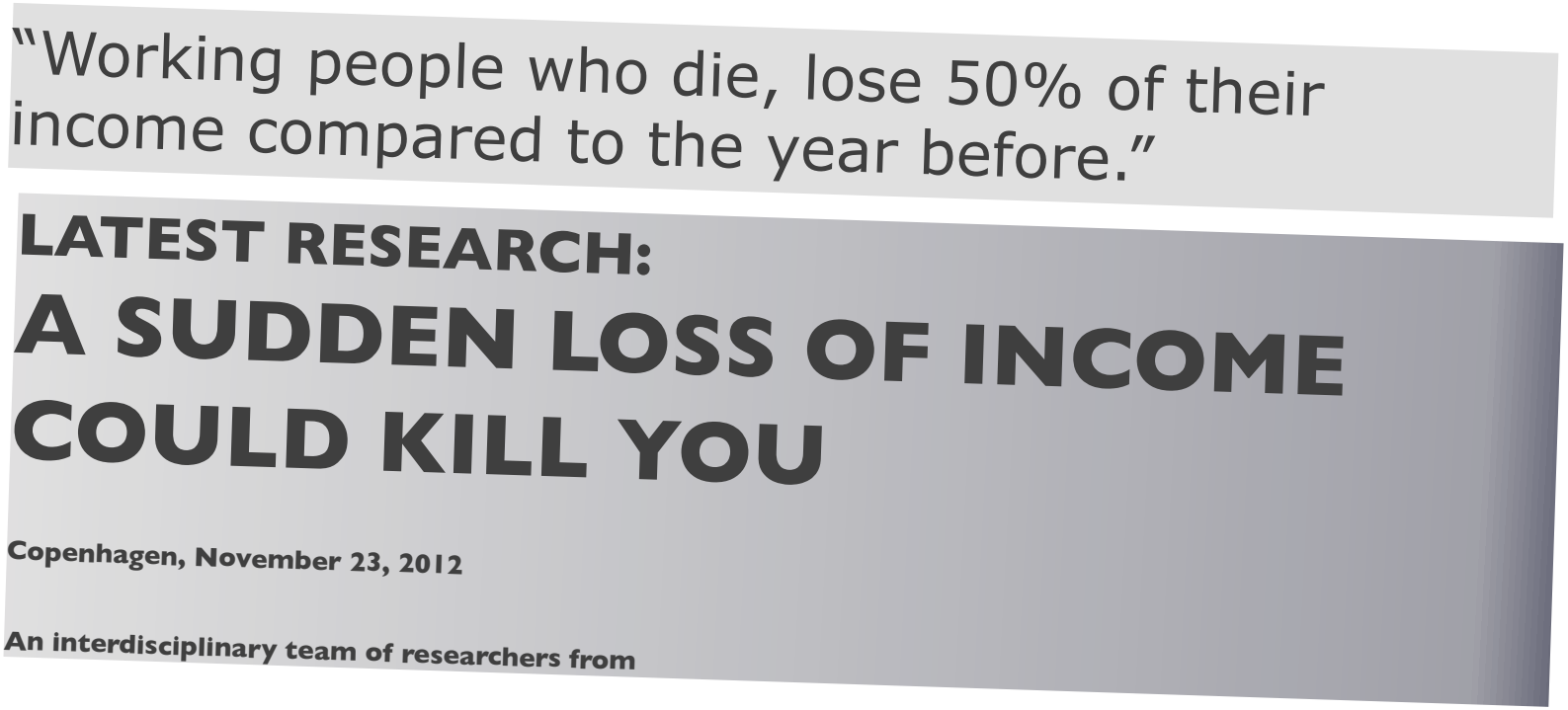


Notable uses:

- Similar documents (e.g. on the web)
- Users with similar preferences (collaborative filtering).

Correlation vs causality

Do not confuse correlation with a causal relationship!

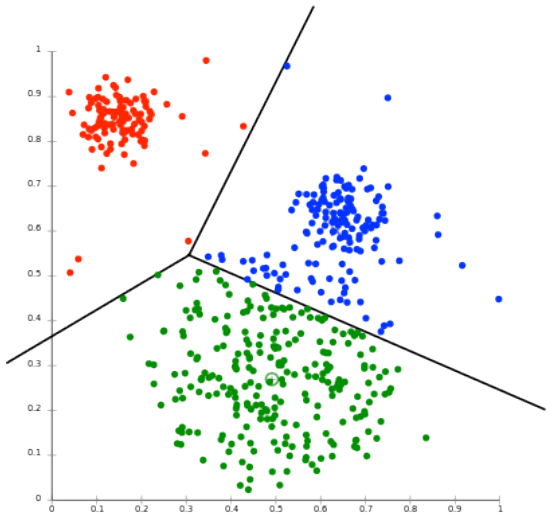


Clustering

Have:

- Data points (e.g. customers).
- Measure of similarity among any 2 customers.
- Metric clustering: Data = d-dimensional points.

Algorithms:
k-means,…

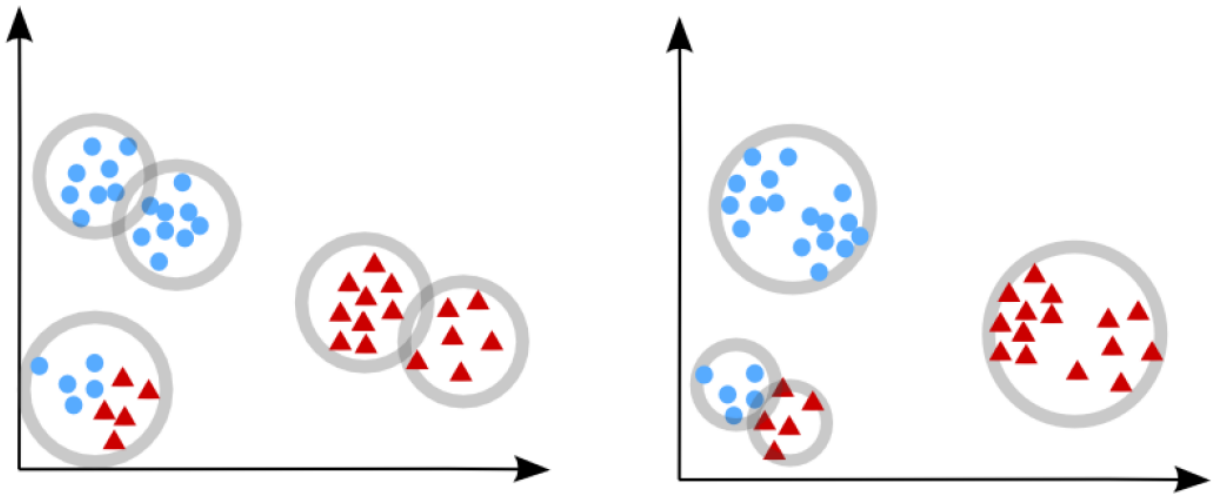


Want:

- To find groups of customers that are similar.

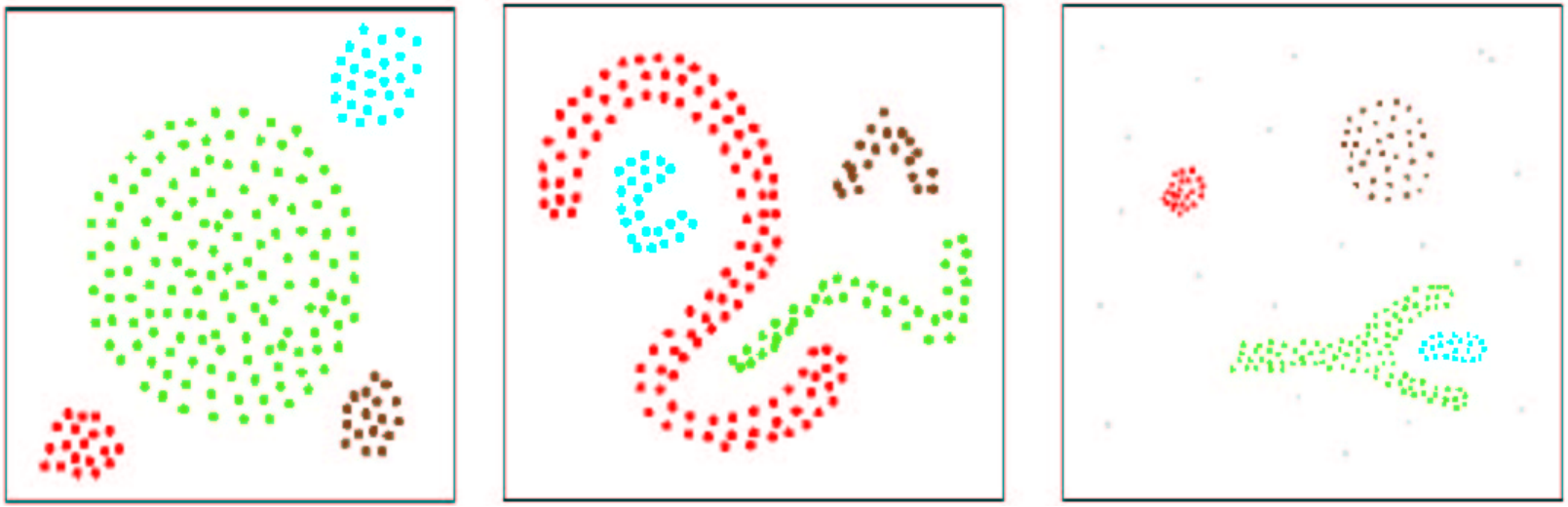
Clustering

- Supervised clustering:
  - Have grouped existing data points, want to put new data points in the right group.
- Semi-supervised clustering:
  - Have grouped **some** existing data points, want to put new data points in the right group.



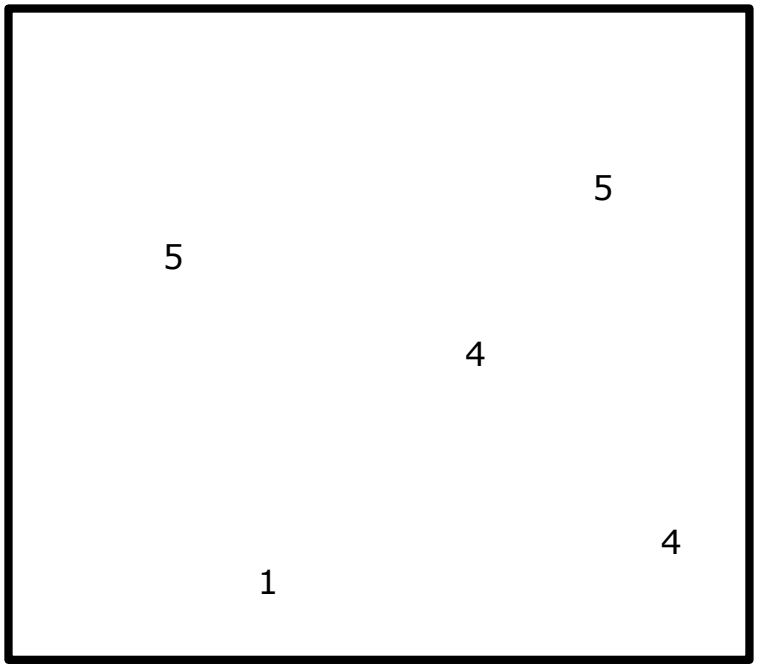
Clustering

- Soft clustering:
  - Each item can belong to several clusters, e.g. 60% cluster 1, 40% cluster 3.
- Manifold clustering:
  - Take point set "connectivity" into accounts when forming the clusters.



Matrix factorization

Goal: Predict values in a large matrix, where some entries are known:



Example: Movie ratings of users.

Matrix factorization

Model: There exist a small set of "basis" behaviors, and every data point is a convex combination of these.

- Non-negativity: A behavior can increase the probability of a choice, but not decrease it.

- Example:
  - I am 20% behavior A and 80% behavior B.
  - Behavior A does not like "E.T." (score 0), but behavior B rates it with a score of 5.
  - My score for "E.T." becomes $0.8 \cdot 5 = 4$.

Matrix factorization

- Last decade:
  Algorithms that **simultaneously** find basis behaviors and each data point's composition of basis behaviors, in order to minimize the model error.

Intended learning outcome

- After the course the students should be able to:
  - …
  - suggest an abstract model suitable for a given data mining task.

Data mining at ITU

- SDT course
  - taught by Julian Togelius, spring semesters
  - goes much more in depth with technical aspects of data mining.
- MSc/BSc thesis possibilities:
  - Julian's group supervise thesis projects with particular emphasis on data mining in data from computer games.
  - I supervise thesis projects in algorithmic foundations of data mining (MaDaMS project).

Next steps

- Exercises this afternoon:
  - Sample "trial exam" problem on data mining.
  - Solution will be made available later.
- Recall: Hand-in 4 due in 1 week!
- Next week: Infrastructures for big data.