Data Mining & Matrices

Pauli Miettinen

Uni Saarland, summer 2015

max planck institut
informatik

Basic Info

- 6 credits
- Lectures:

  Tuesdays 12–14 @ 029, E1.5

- Tutorials:

  Wednesdays 14–16 @ 022, E1.4

- Web page:

  https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/teaching/ss15/data-mining-and-matrices/

- Email:

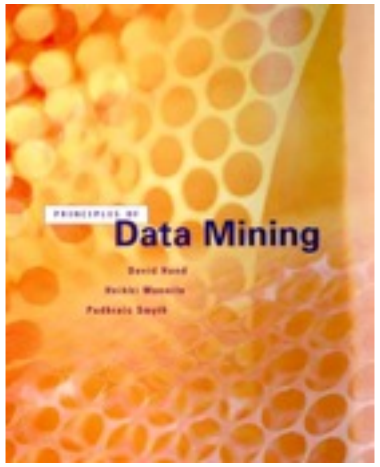
  dmm15@mpi-inf.mpg.de

What is Data Mining?

What is Data Mining?

“Data mining is the process of extracting hidden patterns from data.”

“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”



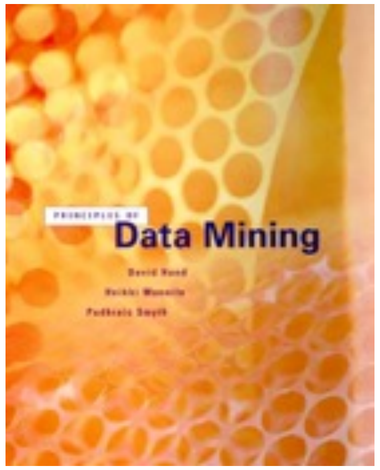
“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

“Data mining, in a broad sense, is the set of techniques for analyzing and understanding data.”

What is Data Mining?

“Data mining is the process of **extracting hidden patterns** from data.”

~~“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”~~

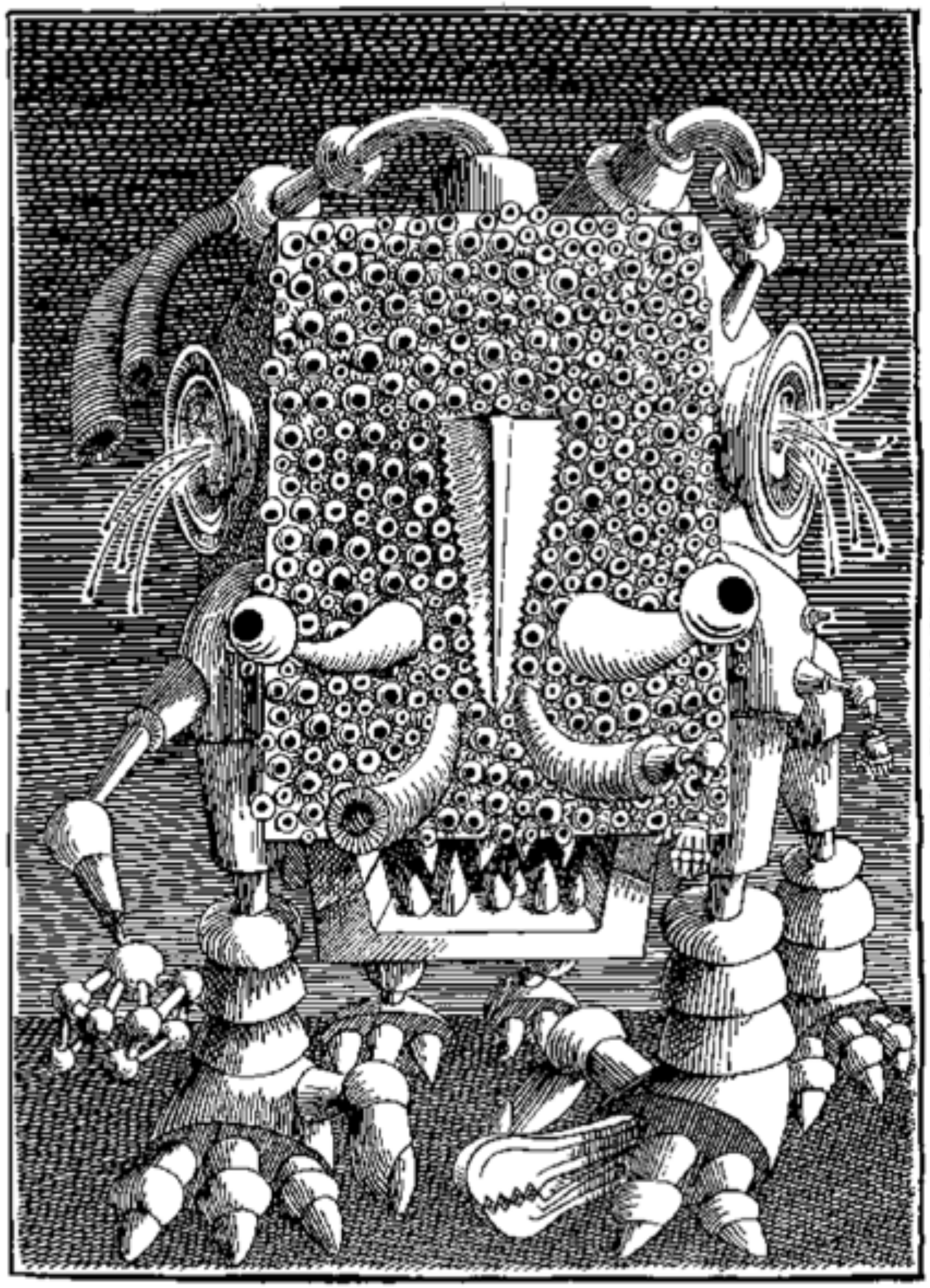


“Data mining is the **analysis** of (often large) observational data sets to find **unsuspected relationships** and to **summarize the data** in novel ways that are both **understandable** and **useful** to the data owner.”

“Data mining, in a broad sense, is the set of techniques for **analyzing** and **understanding** data.”

Why Data Mining?

Why Data Mining?



The ”PHT” Pirate wanted all information of the world. But before he realized most of it was useless, he was already buried under it.

—Stanisław Lem,
The Cyberiad

Data Mining Applications

≈ 5GB of climate data

Walmart

1 000 000 customer transactions per hour

To utilise this data, we need tools to analyse and understand it.

We need data mining.

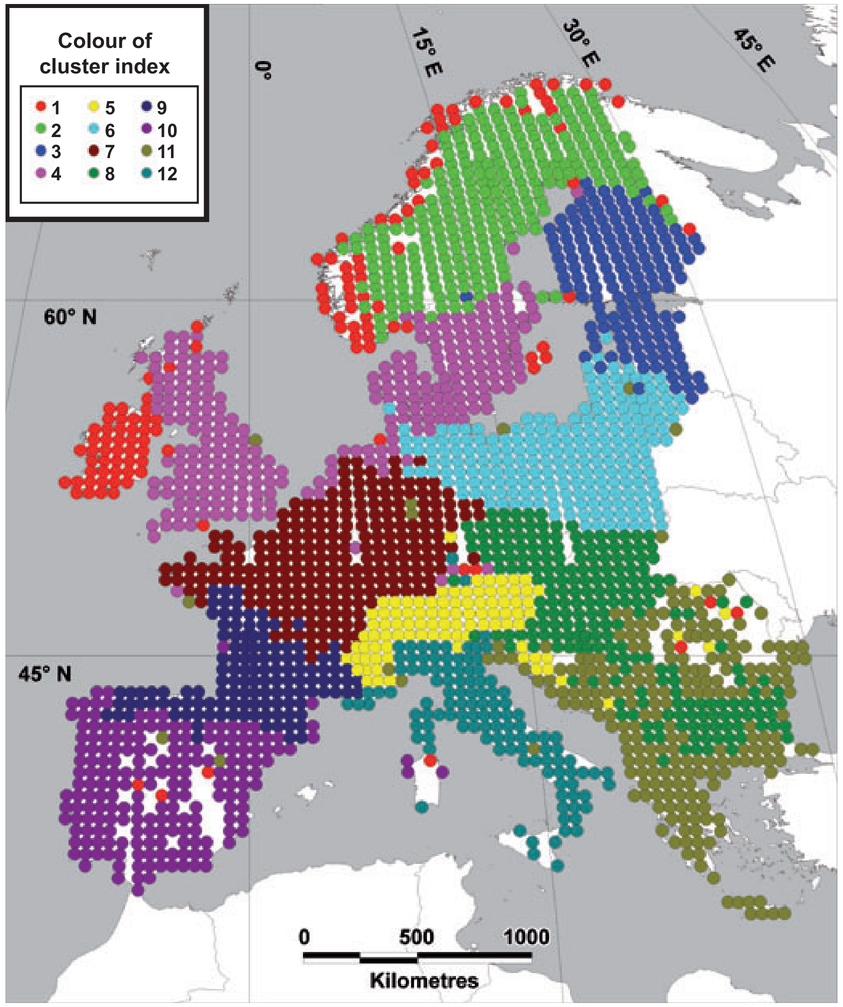
340 000

40 000 000

Data Mining Applications

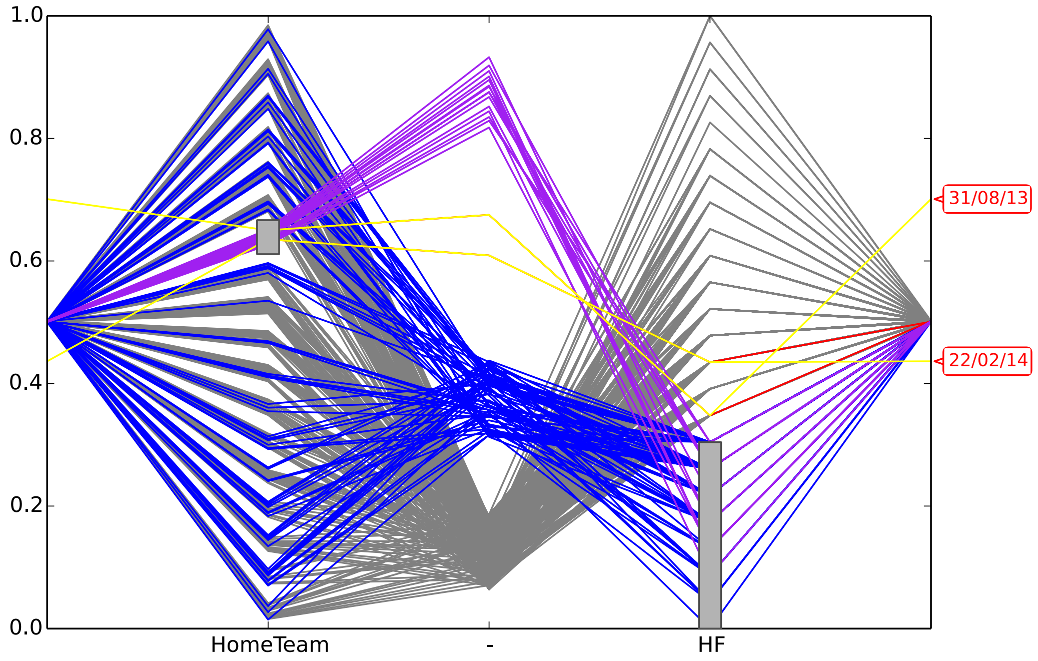
- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…
  - Who Barack Obama had to persuade to vote him?
  - Is there a problem in the International Space Station?

Example Results

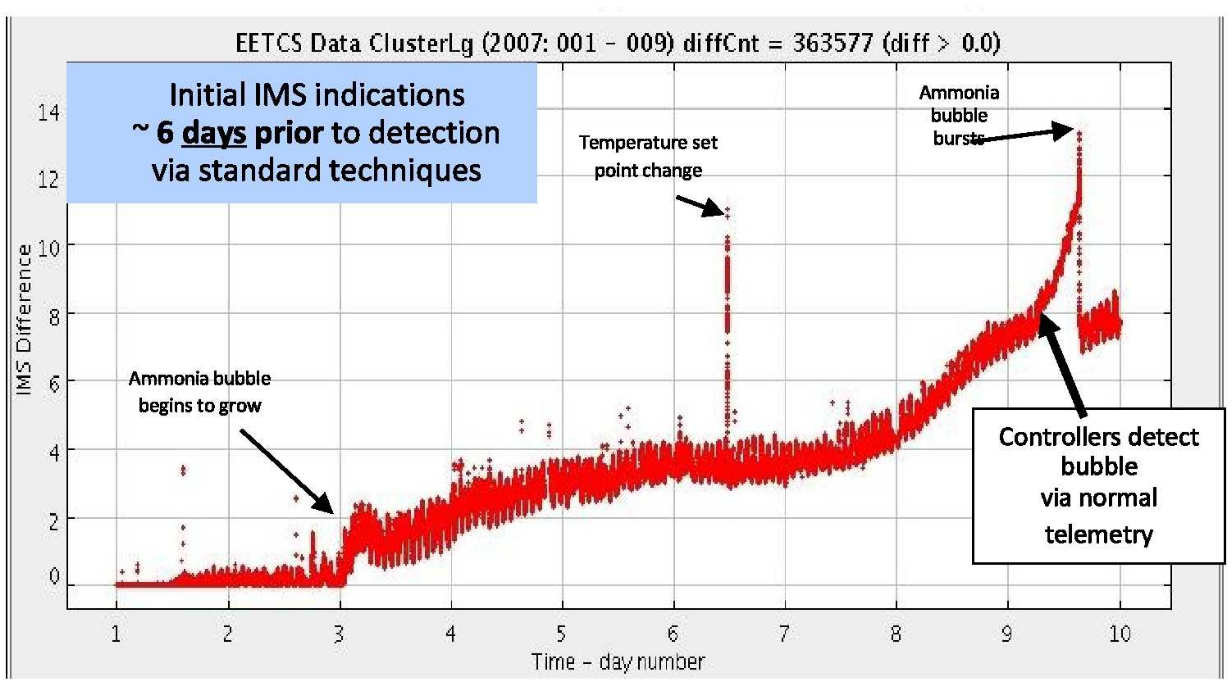


Areas with similar mammals

Heikinheimo, H., et al. (2007). Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6), 1053–1064.



If home team is M'gladbach, then home team doesn't commit more than 12 fouls



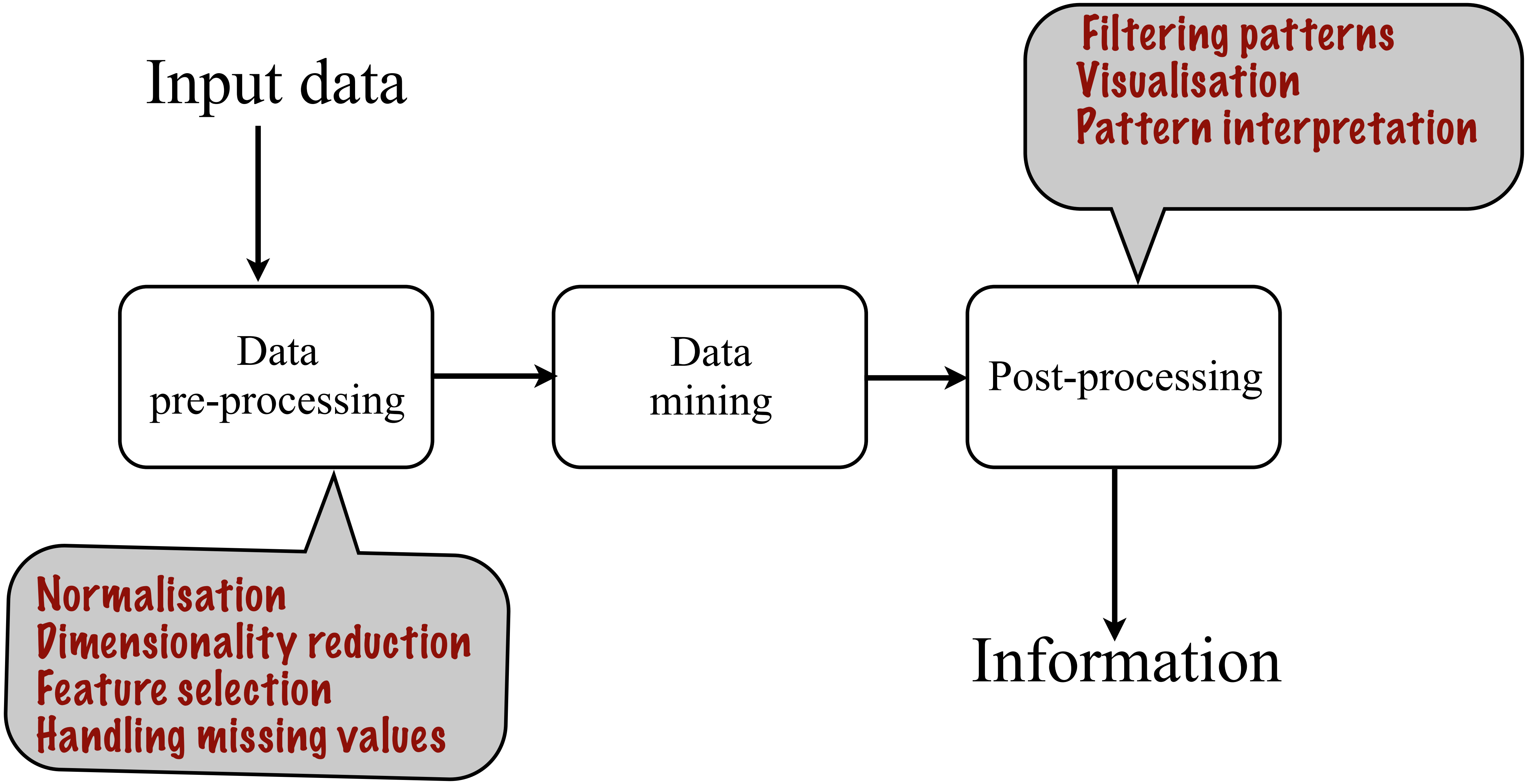
- In early January 2007, ISS Early External Thermal Control System developed an ammonia gas bubble
- Bubble noted by ISS controllers only ~9 hours before it "burst" and dissipated back into liquid

Ashok N. Srivastava: Data Mining at NASA: from Theory to Applications, KDD 2009

Data mining in practice

- Real world is a messy place
    - Real-world data is even messier
    - Data needs pre-processing
- Applications have (hopefully) domain experts
    - Domain knowledge should be incorporated
    - Domain experts should be able to interpret the results
        - Not too many results
        - Post-processing

The KDD process



Filtering patterns
Visualisation
Pattern interpretation

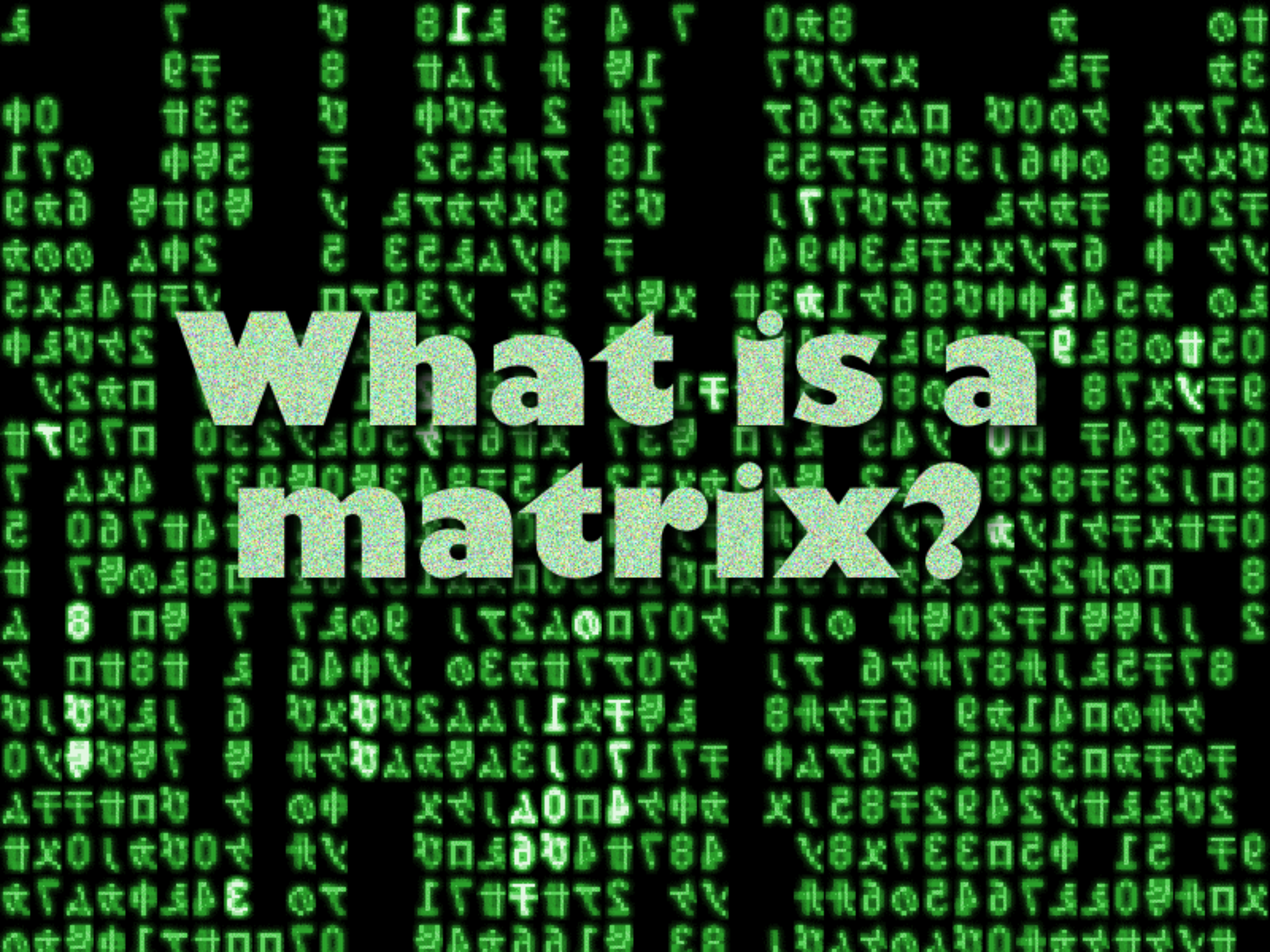
Normalisation
Dimensionality reduction
Feature selection
Handling missing values

Data pre-processing

- Garbage in, garbage out
- Many issues
  - What to do with missing values
    - Are missing values clearly marked?
  - What's the dimensionality vs. sample size
    - Anyway, which way the observations are?
- Do some features correlate with each other in an uninteresting way
  - Record ID and class label
- Is data type suitable for our algorithm
  - Binary, categorical, numerical
- And many, many more…

Post-processing

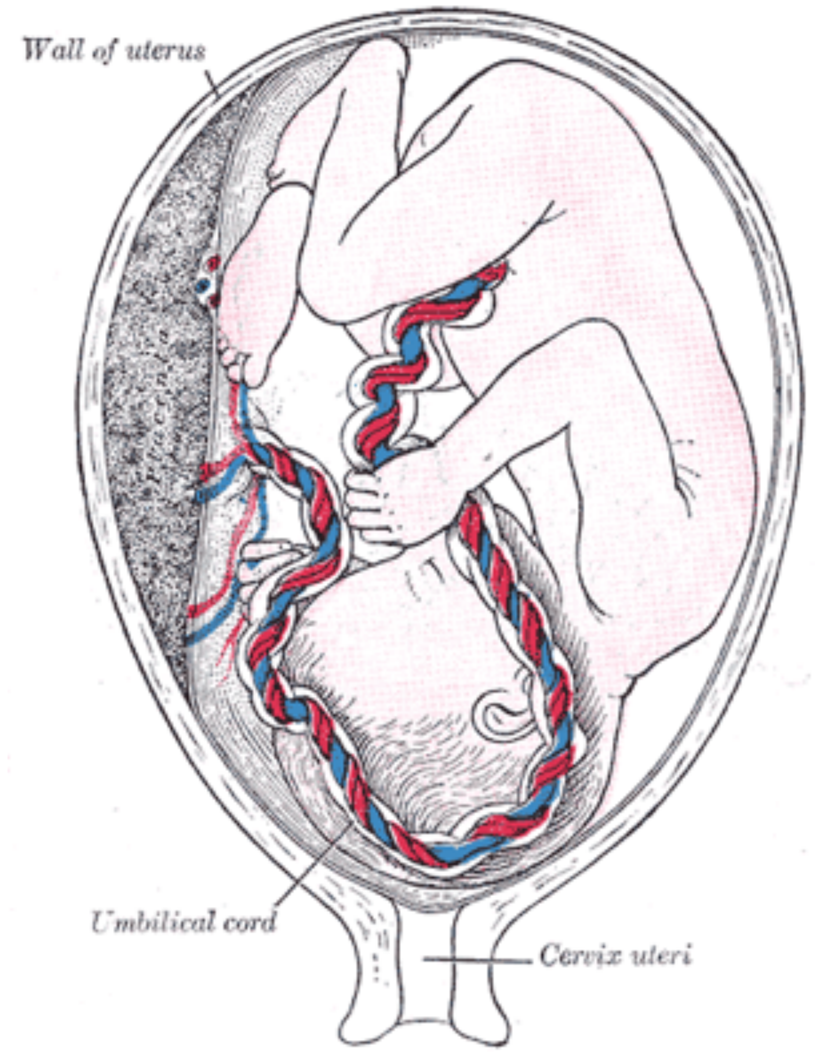
- Humans can only interpret so many results
    - Computers are a different thing
- Select top-*k* results
    - What criteria?
- Are the results significant?
    - Statistics
- Are the results meaningful?
    - Domain expert
- Visualisation
- Humans are great at finding patterns (even when they don’t exist)
    - Computers are a different thing



What is a matrix?

A womb?

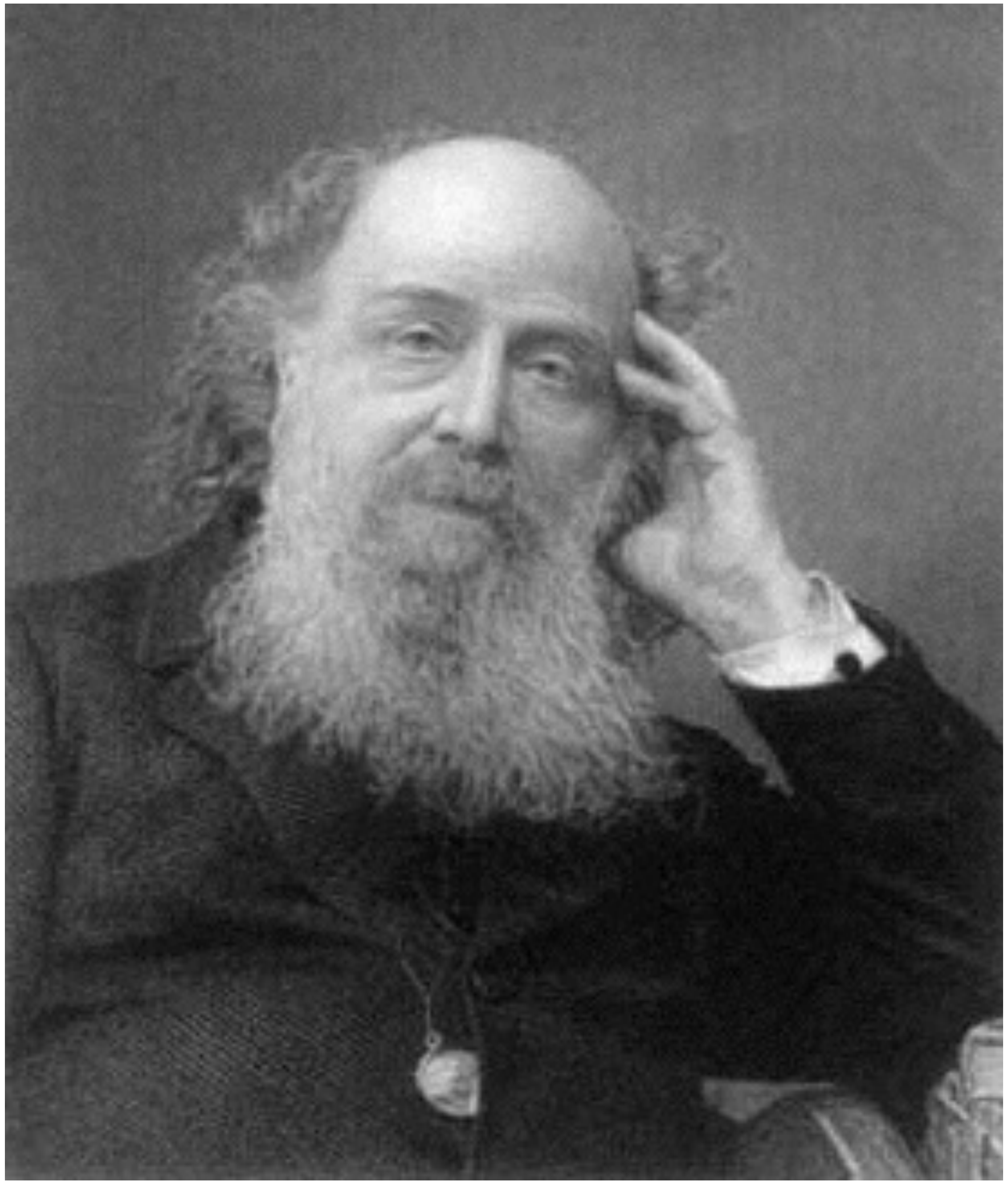
- *mater = mother*
- *matrix = pregnant animal*
- *matrix = womb*, also
  source, origin
- Since 1550s: *place or medium where something is developed*
- Since 1640s: *embedding or enclosing mass*



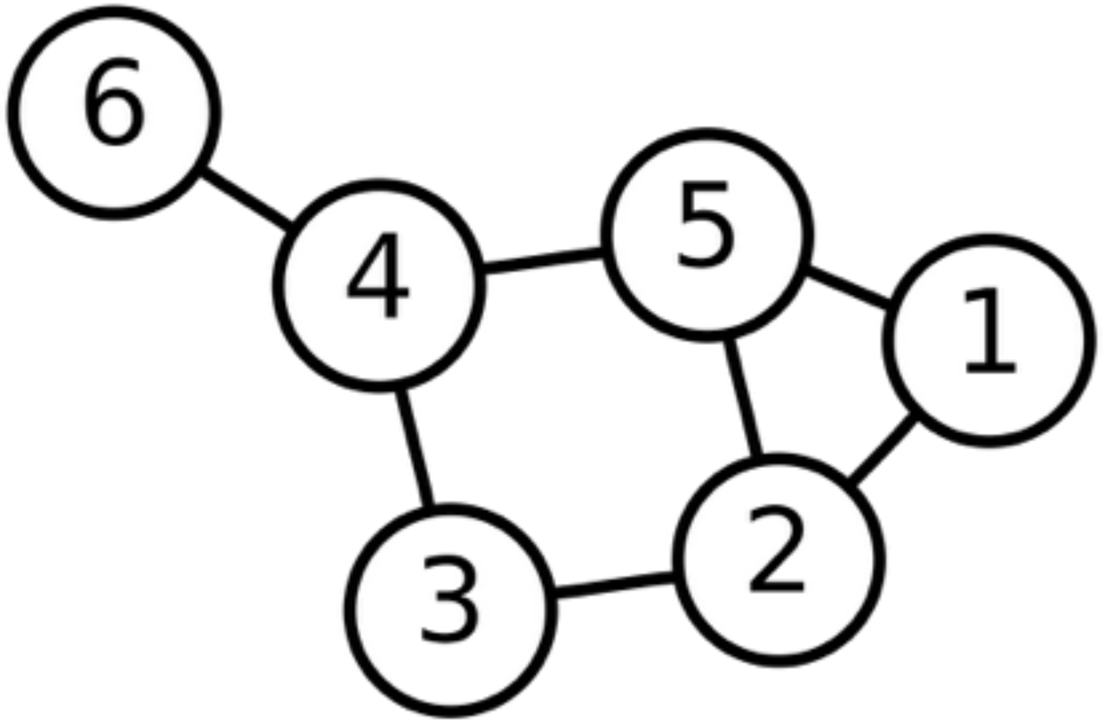
A womb

A rectangular array of numbers?

- *Rectangular arrays* were already known in ancient China in *rod calculus*
  - estim. 300 BCE
- Term *matrix* coined by J. J. Sylvester in 1850



A graph?



A graph

A system of linear equations?

$$3x + 2y + z = 39$$

$$2x + 3y + z = 34$$

$$x + 2y + 3z = 26$$

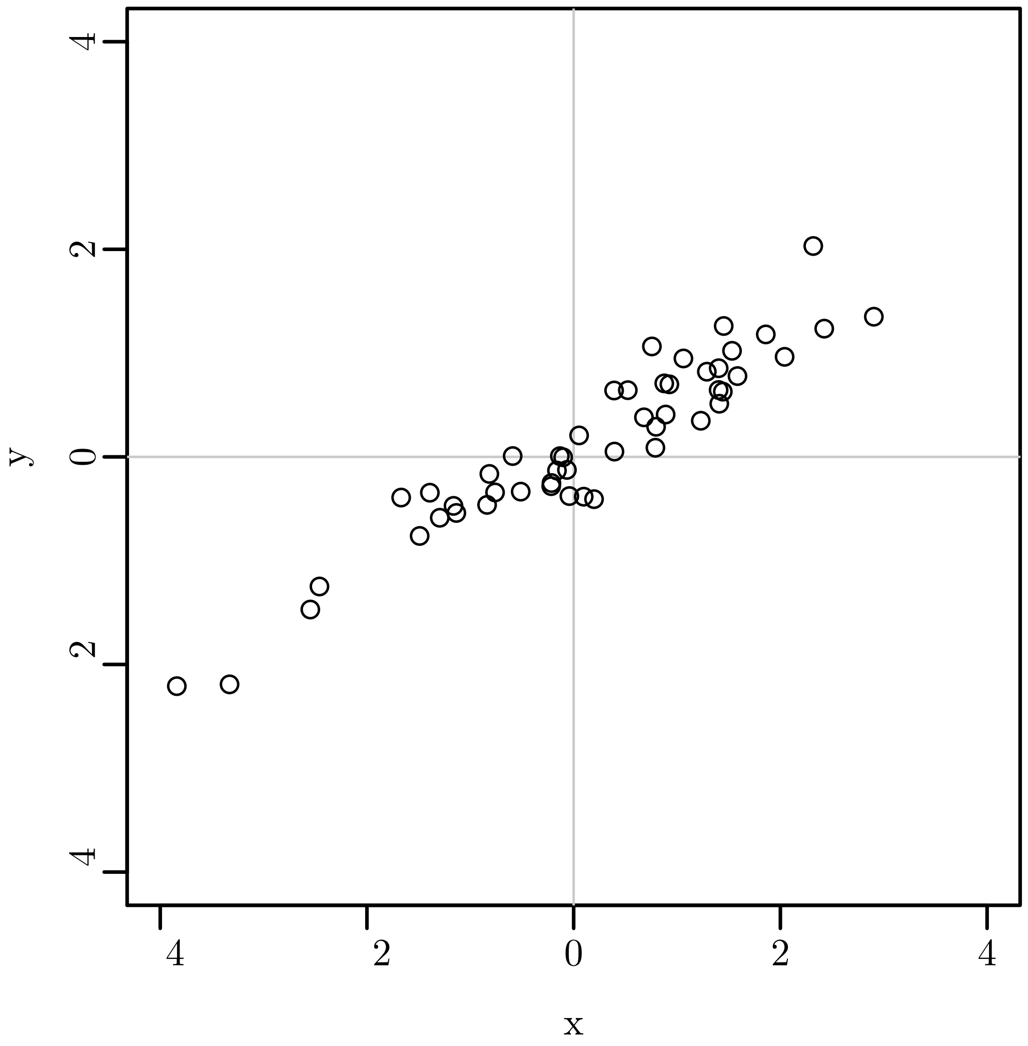
A system of
linear equations

A linear mapping?

$$\begin{aligned} f_1(x,y,z) &= 3x + 2y + z \\ f_2(x,y,z) &= 2x + 3y + z \\ f_3(x,y,z) &= x + 2y + 3z \\ f_4(x,y,z) &= x \end{aligned}$$

A linear mapping

A set of data points?



A set of data points

A brief history…

- First systems of linear equations solved in *Nine Chapters on Mathematical Art,* China, 200–100BCE
- Determinants were invented in 1683 in Japan (by Seki) and Europe (by Leibniz)
  - Further work by Cramer (1750), Laplace (1772), Lagrange (1773)
  - Term *determinant* was coined by Gauss (1801) but first used in its modern sense by Cauchy (1812)
- Jacobi (1830s), Kronecker, and Weierstrass (1850s) considered matrices in as linear transformations
- Caley (1858) published the first abstract definition of a matrix

Matrices and Data Mining

Matrices in data mining

$$\begin{array}{lccc} & \text{Bread} & \text{Butter} & \text{Beer} \\ \text{Anna} & 1 & 1 & 0 \\ \text{Bob} & 1 & 1 & 1 \\ \text{Charlie} & 0 & 1 & 1 \end{array}$$

Customer transactions

$$\begin{array}{lccc} & \text{Data} & \text{Matrix} & \text{Mining} \\ \text{Book 1} & 5 & 0 & 3 \\ \text{Book 2} & 0 & 0 & 7 \\ \text{Book 3} & 4 & 6 & 5 \end{array}$$

Document-term matrix

$$\begin{array}{lccc} & \text{Avatar} & \text{The Matrix} & \text{Up} \\ \text{Alice} & & 4 & 2 \\ \text{Bob} & 3 & 2 & \\ \text{Charlie} & 5 & & 3 \end{array}$$

Incomplete rating matrix

$$\begin{array}{lccc} & \text{Jan} & \text{Jun} & \text{Sep} \\ \text{Saarbrücken} & 1 & 11 & 10 \\ \text{Helsinki} & 6.5 & 10.9 & 8.7 \\ \text{Cape Town} & 15.7 & 7.8 & 8.7 \end{array}$$

Cities and monthly temperatures

Matrix decompositions in data mining

- A common goal in data mining is to find regularities (or patterns) in the data
  - Often, to summarize the data
- A *matrix decomposition* presents the data as a sum of “simple” elements, i.e. patterns
  - but there’s also other uses… *stay tuned!*

Learning objectives

- To know the most common/important matrix factorization methods
    - their advantages and disadvantages
    - their use in data mining
- To understand the theoretical foundation behind the techniques
- To be able to use the techniques to solve real-world data analysis problems

Organization

Staff

- Lecturer: Dr. Pauli Miettinen
- Tutors:
  - Sanjar Karaev (theor. assignments)
  - Saskia Metzler (prog. assignments)

Course Structure

- Lectures (almost) every week
- Pen-and-paper homeworks every second week
- Three hands-on assignments
- Final exam (oral/written TBD)

Course material

- Lecture slides (available on course homepage)
- David Skillicorn: *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. Chapman & Hall 2007
- Gene H. Golub & Charles F. Van Loan: *Matrix Computations*, 3rd ed. Johns Hopkins University Press 1996
- Jure Leskovec, Anand Rajaraman & Jeff Ullman: *Mining of Massive Datasets*, 2nd ed. Cambridge University Press 2015 (available online http://www.mmds.org)

D5 Databases and Information Systems

Data Mining and Matrices

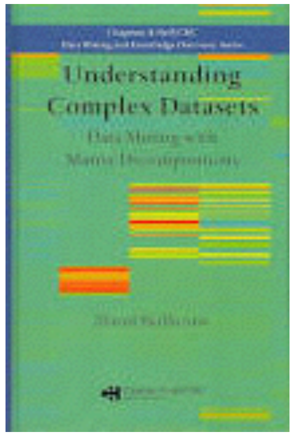
Organization

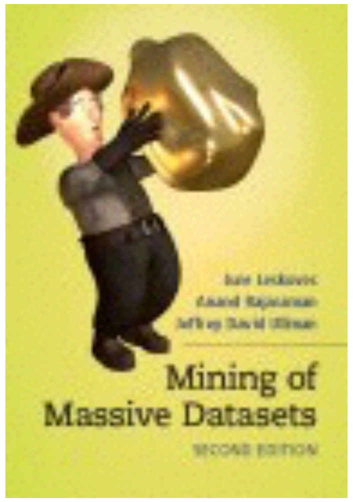
Staff

Time & Location

News

Course contents





Lectures

- Slides will be made available to the home pages
- Which one do you prefer:
  - Starts 12:00, no break, ends 13:30
  - Starts 12:00, 30min break, ends 14:00
  - Starts 12:15, no break, ends 13:45
  - Starts 12:15, 15min break, ends 14:00
  - Starts 12:30, no break, ends 14:00

To Pass the Course:

- Return acceptable solutions to at lest 50% of the homework questions
- Return a solution to all three hands-on assignments
    - At most one failed solution can be converted to a pass by doing extra homework
- Pass the final exam

Bonus Points

- Excellent solutions to hands-on assignments earn 1 bonus point each
- At least 75% acceptable homework solutions earn one bonus point
- At least 90% acceptable homework solutions earn one more bonus point
- Every bonus point improves a passing final exam grade by one third, to the maximum of one full point
  - Ex: 75% hw, 1 excellent, 2.3 from exam ⇒ 1.7 final grade

Homeworks

- Handed out every second week
- One week to do (with one exception)
- Return in written form by next week's lecture
  - By email to dmm15@...
  - By hand to the lecture
- LaTeX-prepared solutions preferred, legible hand-written ones accepted if returned personally to the lecture

Homeworks cont’d

- You can use computers (but must show sufficient details & intermediate steps)
- Discussing is OK, copying is not
- Acceptable answers don’t have to be fully correct
- 36 hw points in total
    - 50% ≥ 18, 75% ≥ 27, 90% ≥ 33
    - Some questions can award extra points

Homework Tutorials

- Tutorial meetings covering the homeworks are in the day after their due day
  - 6.5., 20.5., 3.6., 17.6., 1.7., and 15.7.
- Discussion on correct solutions & get to know your points

Hands-On Assignments

- Three hands-on assignments
  - Implementing and using methods from the lectures; analyzing results; understanding the process
- 3–4 weeks to complete
- Done using the R language by default

Hands-On Tutorials

- Every second week tutorial meetings discuss the hands-on assignments
    - 13.5., 27.5., 10.6., 24.6., 8.7., 22.7., 29.7.
- Help with problems, feedback from previous assignments, meeting with tutor & peers
    - Discussion is OK, copying is not
- Next week's lecture: Intro to R
    - Bring your laptop!

Exam

- Written or oral (TBD)
- Place TBA
- Time: I’m proposing the last lecture’s time, 28 July 2015
    - Otherwise either very early next week, or very late in summer

One more thing

- First homework is given out today, due 5 May
- You should be able to answer to all questions with prerequisite knowledge