Data Mining
&
Matrices

Pauli Miettinen

Uni Saarland, summer 2017

Basic Info

- 6 credits
- Lectures:

  Mondays 14–16 @ 029, E1.5

- Tutorials:

  Wednesdays 10–12 @ 024, E1.4

- Web page:

  https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/teaching/ss17/data-mining-and-matrices/

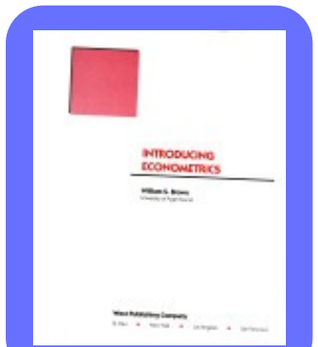
- Email:

  dmm17@mpi-inf.mpg.de

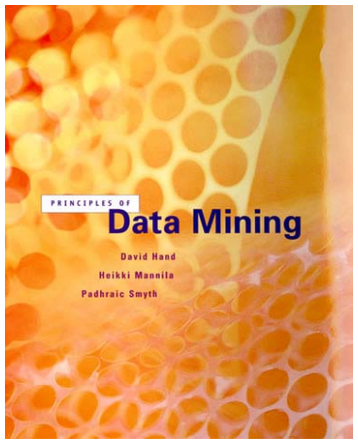
What is Data Mining?

What is Data Mining?

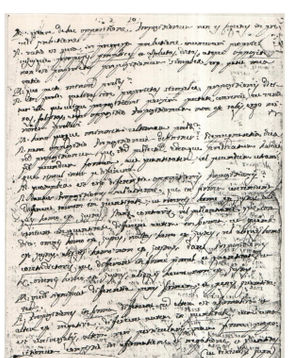
"Data mining is the process of extracting hidden patterns from data."



"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."

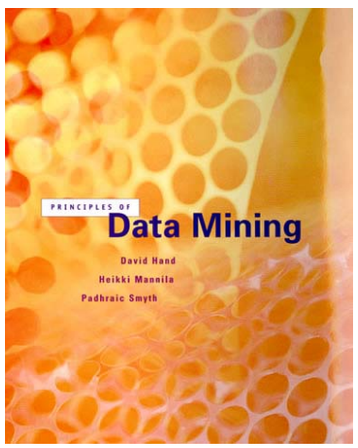


"Data mining, in a broad sense, is the set of techniques for analyzing and understanding data."

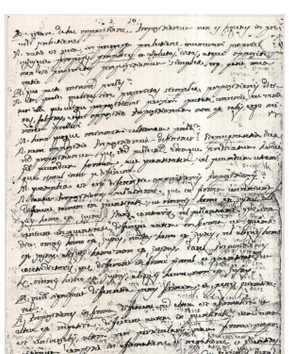
What is Data Mining?

"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."~~



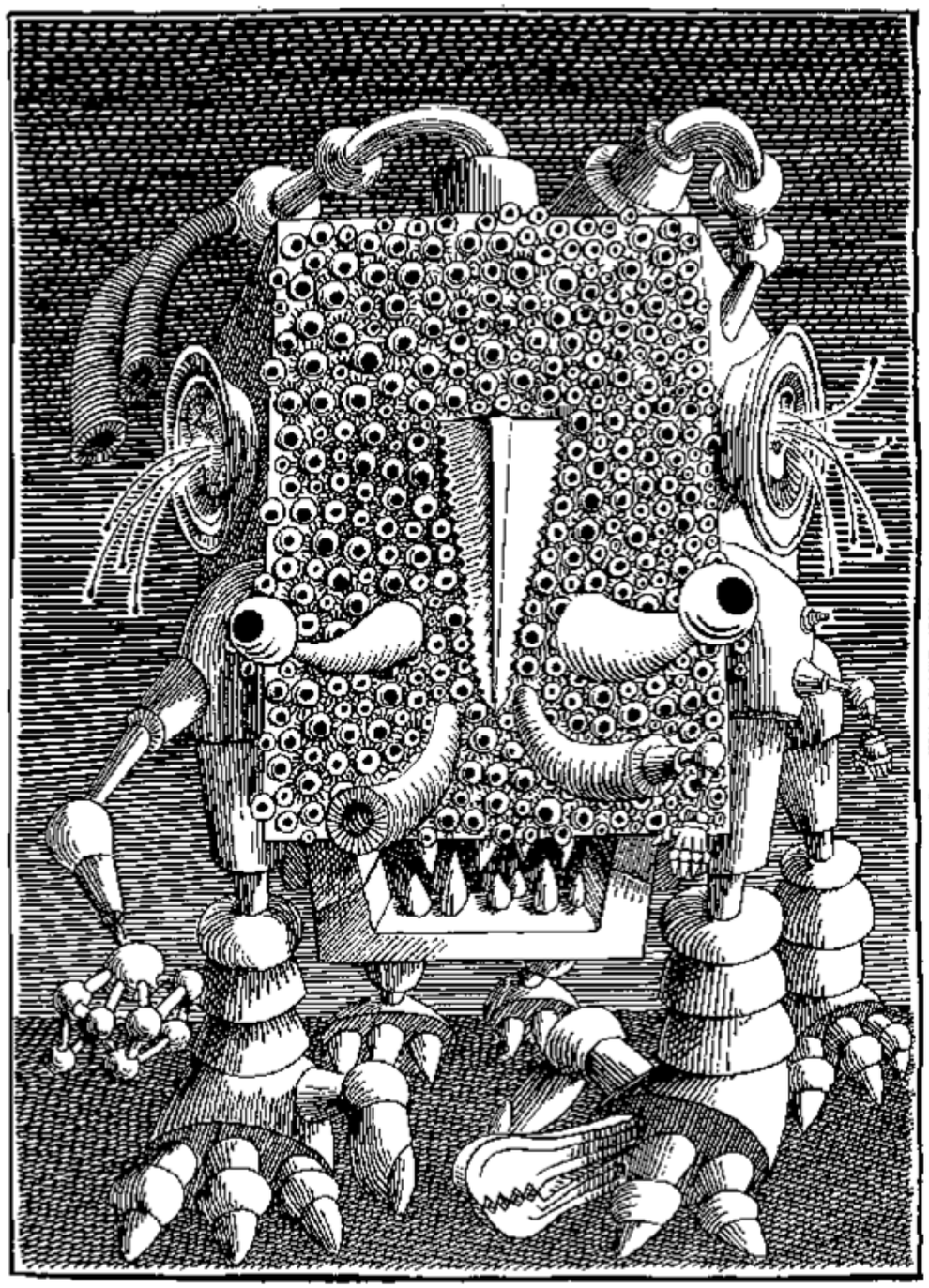
"Data mining is the **analysis** of (often large) observational data sets to find **unsuspected relationships** and to **summarize the data** in novel ways that are both **understandable** and **useful** to the data owner."



"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 realised most of it was useless, he was already buried under it.

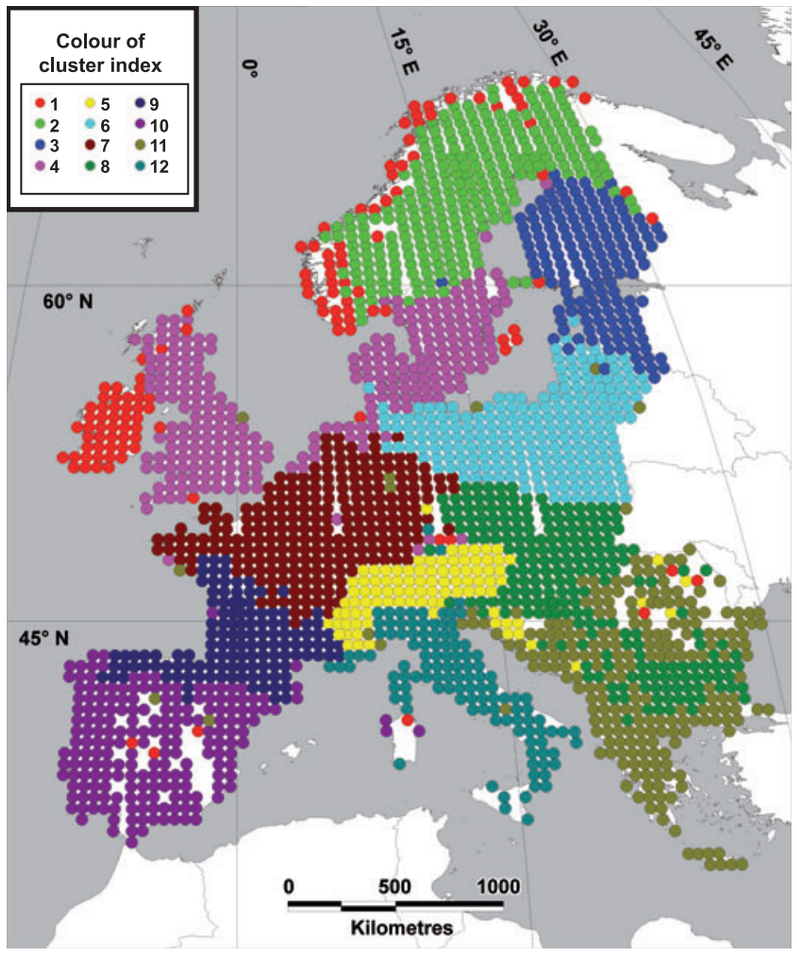
—Stanisław Lem,

The Cyberiad

Data mining applications

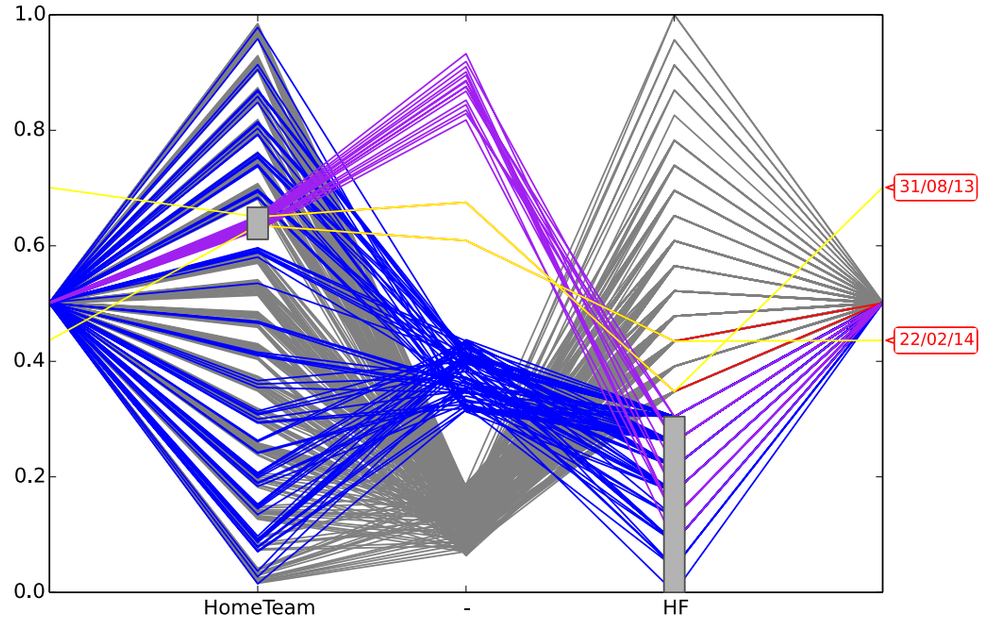
- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…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

Example results

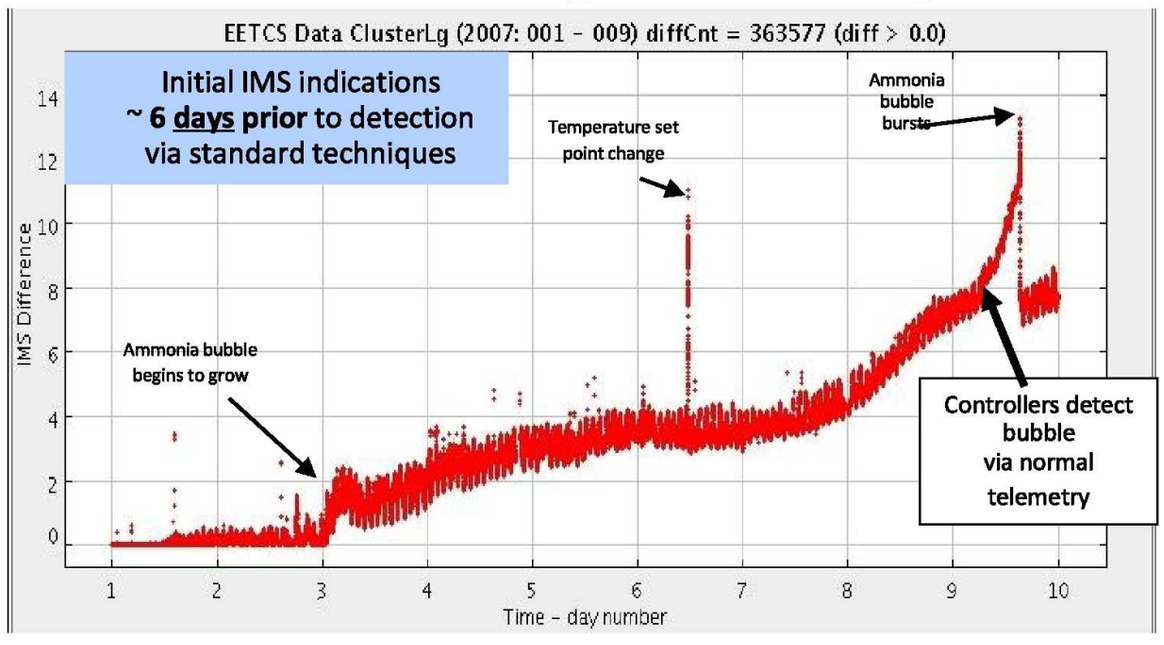


Areas with similar mammals

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

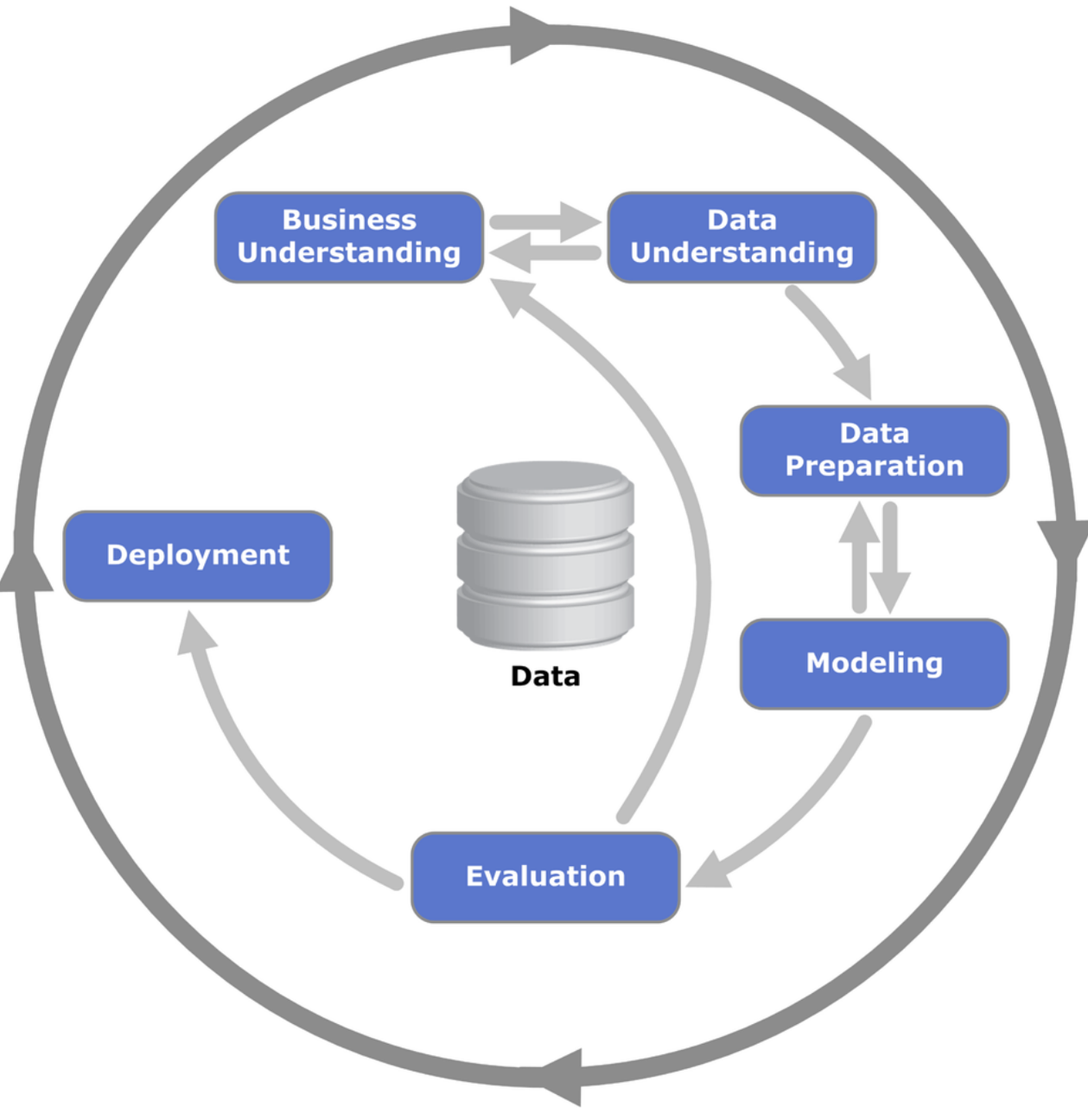
Ashok N. Srivastava: Data Mining at NASA: from Theory to Applications, KDD 2009

The KDD process

Input d

Data
pre-proce

Normalis
Dimensionalit
Feature se
Handling miss



ng patterns
sualisation
n interpretation

ng

n

"CRISP-DM Process Diagram" by Kenneth Jensen - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png#/

Data pre-processing

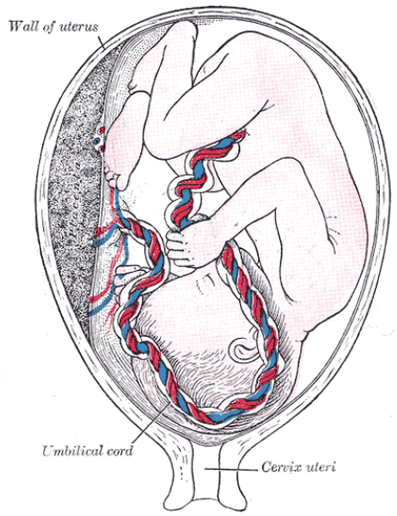
- Garbage in, garbage out
- Many issues
  - What to do with missing values
    - Are missing values clearly marked?
  - What's the dimensionality vs. sample size
    - Anyway, which way the observations are?
  - Do some features correlate with each other in an uninteresting way
    - Record ID and class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

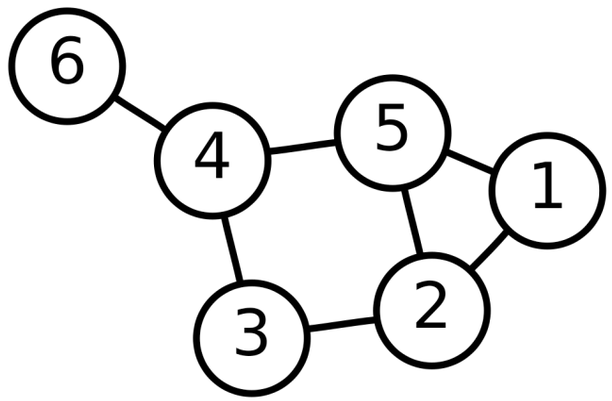
Post-processing

- 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?

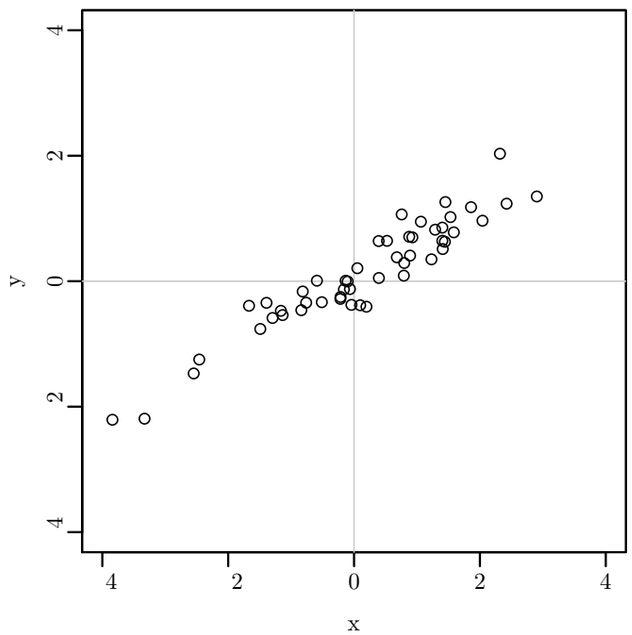
Which of these is a matrix?





$$3x + 2y + z = 39$$
$$2x + 3y + z = 34$$
$$x + 2y + 3z = 26$$

$$f_1(x, y, z) = 3x + 2y + z$$
$$f_2(x, y, z) = 2x + 3y + z$$
$$f_3(x, y, z) = x + 2y + 3z$$
$$f_4(x, y, z) = x$$



$$\begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix}$$

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 as linear transformations
- Caley (1858) published the first abstract definition of a matrix

Matrices and Data Mining

Matrices in data mining

	Bread	Butter	Beer
Anna	1	1	0
Bob	1	1	1
Charlie	0	1	1

Customer transactions

	Data	Matrix	Mining
Book 1	5	0	3
Book 2	0	0	7
Book 3	4	6	5

Document-term matrix

	Avatar	The Matrix	Up
Alice		4	2
Bob	3	2	
Charlie	5		3

Incomplete rating matrix

	Jan	Jun	Sep
Saarbrücken	1	11	10
Helsinki	6.5	10.9	8.7
Cape Town	15.7	7.8	8.7

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 summarise 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 factorisation 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

Learning objectives: theory

- Students understand how matrix decompositions and linear algebra can be used to solve and model data analysis problems
- Students understand the theory behind the most common matrix decomposition methods
- Students can prove (simple) theorems about the methods and present their proves to their peers
- Students have the basic knowledge to be able to understand new matrix decomposition methods themselves

Learning objectives: practice

- Students have working knowledge of the R statistical language and can use existing matrix decomposition implementations and implement new ones
- Students can apply matrix decomposition methods to real-world data analysis problems
- Students can analyse the results and present their analysis in a coherent written format

Organisation

Staff

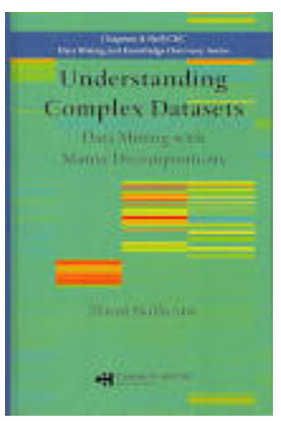
- Lecturer: Dr. Pauli Miettinen
- Tutors:
  - Sanjar Karaev (theory assignments)
  - Saskia Metzler (practical assignments)

Course structure

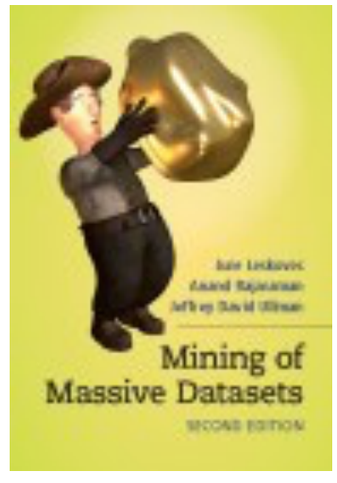
- Lectures (almost) every week
- Pen-and-paper problem sheets every second week & tutorial sessions in the following week
- Three hands-on assignments
- Final exam (written)

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)



Lectures

- Slides will be made available via the home page
- Which one do you prefer:
  - Starts 14:00, no break, ends 15:30
  - Starts 14:00, 30min break, ends 16:00
  - Starts 14:15, no break, ends 15:45
  - Starts 14:15, 15min break, ends 16:00
  - Starts 14:30, no break, ends 16:00

Practical assignments

- Three hands-on assignments
  - Implementing and using methods from the lectures; analysing results; understanding the process
- 3 weeks to complete
- Graded: failed, passed, excellent
- Done using the R language

Hands-on tutorials

- Every second week tutorial meetings discuss the hands-on assignments
  - 3 May, 17 May, 31 May, 14 June, 28 June, 12 July
  - Volunteer, but recommended!
- 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!

Problem sheets

- Handed out every second week
- Week-and-a-half to do (with one exception)
- Six problems per sheet
- To get marks for solved problems, **you must attend the tutorial session**

Theory tutorials

- At begin, you mark which problem's you've solved
  - **You must have the solutions written down and with you**
- Tutor chooses (randomly) students to present their solutions in the blackboard
  - One student per problem
  - Corrects & guides if there are issues

Theory tutorials cont’d

- You can use computers (but must show sufficient details & intermediate steps)
- Discussing is OK, copying is not
- You can mark an answer even if you know it’s not fully correct
  - Must show a significant progress towards solving, and you must always be ready to present your solution
- Marking problems you haven’t done leads into losing the mark for the problem *and* the tutor *will* check the rest of your solutions, and can remove marks from those as well

Theory tutorials

- Tutorial meetings covering the problem sheets are in the week after their hand-out
  - 10 May, 24 May(?), 7 June, 21 June, 5 July, and 19 July
- 24 May is MPI-INF SAB ⇒ place and maybe the time for the tutorial might have to be changed

To pass the course:

- Mark solved at lest 50% of the homework questions (18/36)
- Return acceptable 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

- You can earn at most 3 bonus points
  - Each increases the grade of a passed exam by one step (1/3)
    - E.g. with 1 bonus point, 1.7 turns into 1.3

Bonus point matrix

3 points	2 points	1 point
at least 33 marked problems and three excellent grades	at least 30 marked problems and two excellent grades	at least 30 marked problems
		at least two excellent grades
		at least 27 marked problems and at least one excellent grade

Exam

- Written
- Place TBA
- Time: (tentatively) the last lecture, 24 July
  - Otherwise very late in summer

One more thing

- First problem sheet is given out today, tutorial on 10 May (2½ weeks from now)
- You should be able to answer to all questions with prerequisite knowledge
  - **For this problem sheet, you must attend the tutorial and mark all problems solved to be eligible to sit in the final exam**

That’s all folks!

- Next week, no lecture (1st of May)
  - Saskia will give an introduction to the R language on next week’s Wednesday