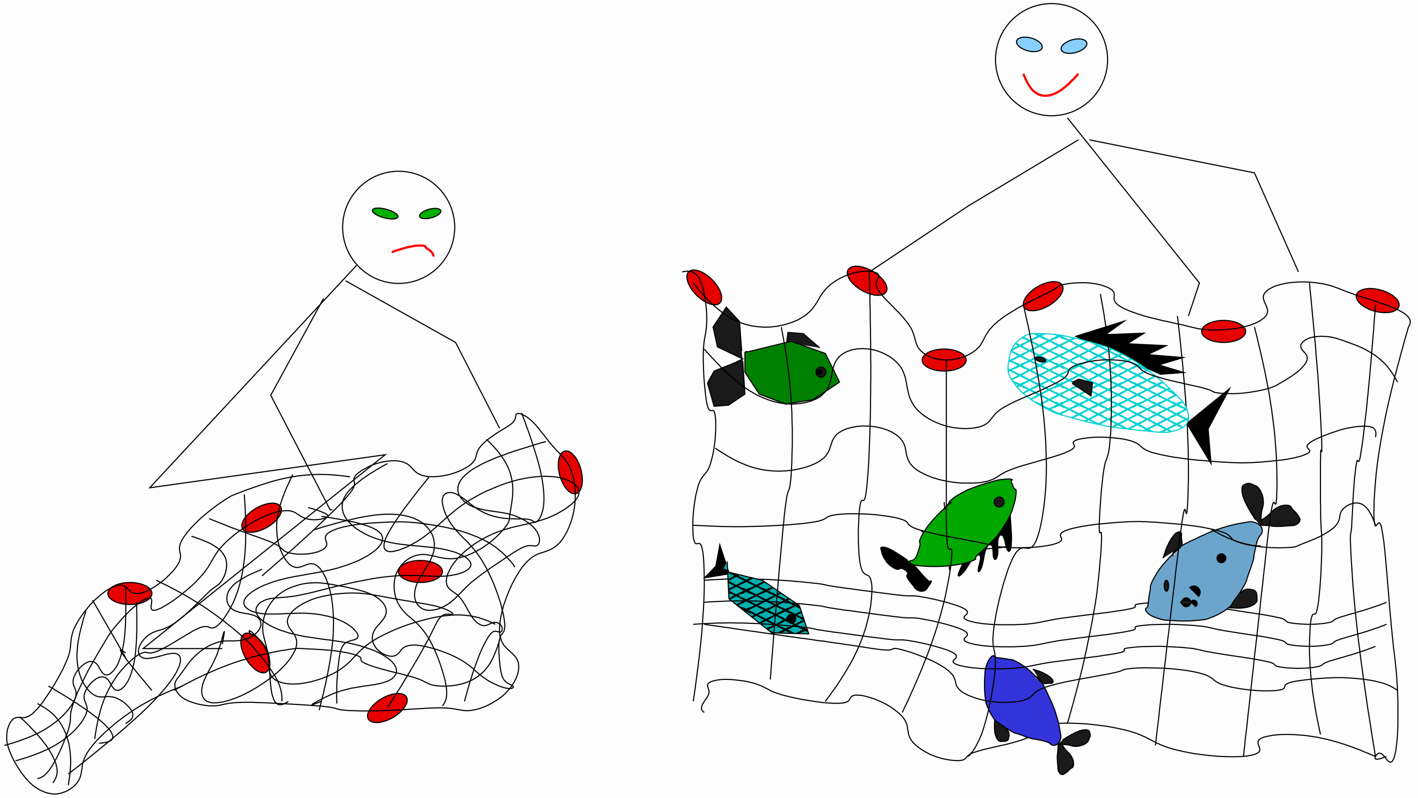
CS-E4650 Methods of Data Mining



I Course logistics, II Introduction to DM, III Preprocessing

Teaching staff

instructor: Wilhelmiina “**Wiki**” Hämäläinen (wilhelmiina.hamalainen@aalto.fi)

teaching assistants:
Egor Eremin egor.eremin@aalto.fi
Georgy Ananov georgy.ananov@aalto.fi
Hieu Nguen Khac hieu.nguyenkhac@aalto
Lai Khoa khoa.lai@aalto.fi
Paavo Reinikka paavo.reinikka@aalto.fi
Yinjia Zhang yinjia.zhang@aalto.fi
+ Vinh N’guyen helped with preparation

guest lecturers or visitors:
prof. Heikki Mannila and MSc Juho Rinta-Paavola

contact: course forum, please avoid email chaos!

Communication and course material

All course information available via **mycourses.aalto.fi** (MC):
`https://mycourses.aalto.fi/course/view.php?id=41020`

- announcements (all important announcements by MC!)
- lecture notes and external material
- link to the text book: **Charu C. Aggarwal: Data Mining: The Textbook**, Springer 2015
- exercise tasks and material
- link to **course forum** `https://mdm2023.zulip.aalto.fi/`

Ask during lectures and exercise sessions and in the course forum. **Please, use email only for personal matters** that you cannot ask elsewhere.

Advicing in zulip

Questions on exercises and homeworks

- ask under the right channel (e.g., “Exercise session 1”)
- give informative title to the stream (like task number)
- TAs’ reply questions during weekdays (+ other students can reply)
- no real-time responses (some delays)
- if you want a reply before weekend, ask before Thu 4pm latest

Other questions (lectures, general)

- like above, but the lecturer and TAs reply (also students can reply, if you know the answer! e.g., something told in MC)

Completing course

1. activite participation in exercise groups (5 sessions, max 5p)
2. submitting homeworks in groups of 2–3 students (5 tasks, max 10p)
3. final **exam** Wed 13.12. 13:00–16:00 (max 24p)
4. prerequisite test (max 1p) `https://plus.cs.aalto.fi/cs-e4650/2023/` (**deadline 18th Sep 2023**)

- the final grade is based on the sum of points (max 40)
- to pass the course one needs to get $\geq$ **50% of total points and $\geq$ 50% of the exam points**

Exercises and homeworks

Exercise tasks

- individual solution beforehand
- processing in small groups during sessions + presentation
- in exceptional/force majeure circumstances you can once return a solution report to the TAs instead

Homeworks (home assignments)

- done in groups of 2–3 students (but independent work, no AI tools unless specifically asked to use)
- at least 10 days time to solve
- submit before the deadline! (with -10% penalty can be 24h late)

Average workload (5 ects ≈ 135h)

- 34–36h [a] contact sessions (lectures and exercises)
- 20h preparation for exercises
- 20h graded homeworks (in groups)
- 40h self-studying (more if skipped lectures/sessions)
- 20h preparation for the exam

Important: Solve exercise tasks beforehand! (Best way to learn!)
Self-study every week! (read the book & other learning material)

[a] now allocated 1 extra lecture

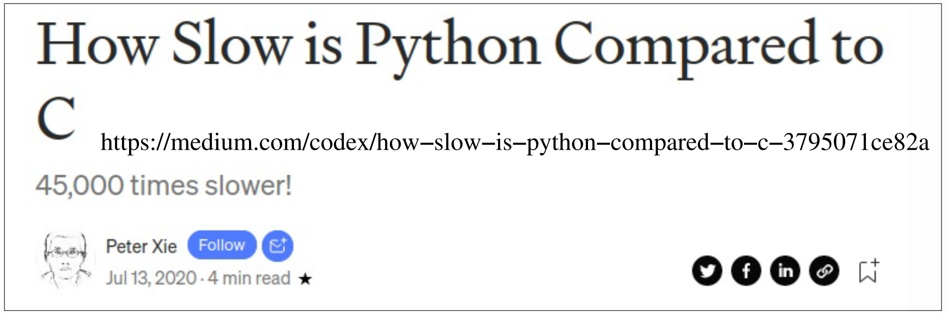
Learning goals

- Know fundamental data mining problems, pattern types and methods
- Know which methods to choose for a given problem or keywords to find more information
- Recognize when to expect computational problems and know some feasible strategies
- Understand importance of validation and know some approaches to validation
- Make programs that use or implement DM methods
- Utilize existing source code and tools in DM tasks
- Learn good DM practices

Meta-learning goals

Not actual learning goals, but **useful skills for data miners** that you are encouraged to learn!

- reading scientific papers related to DM
- writing efficient programs (and algorithms)
- managing many alternative tools or programming languages
- working in linux/unix environment
- learning critical thinking



How Slow is Python Compared to C

https://medium.com/codex/how–slow–is–python–compared–to–c–3795071ce82a

45,000 times slower!

Peter Xie

Jul 13, 2020 · 4 min read ★

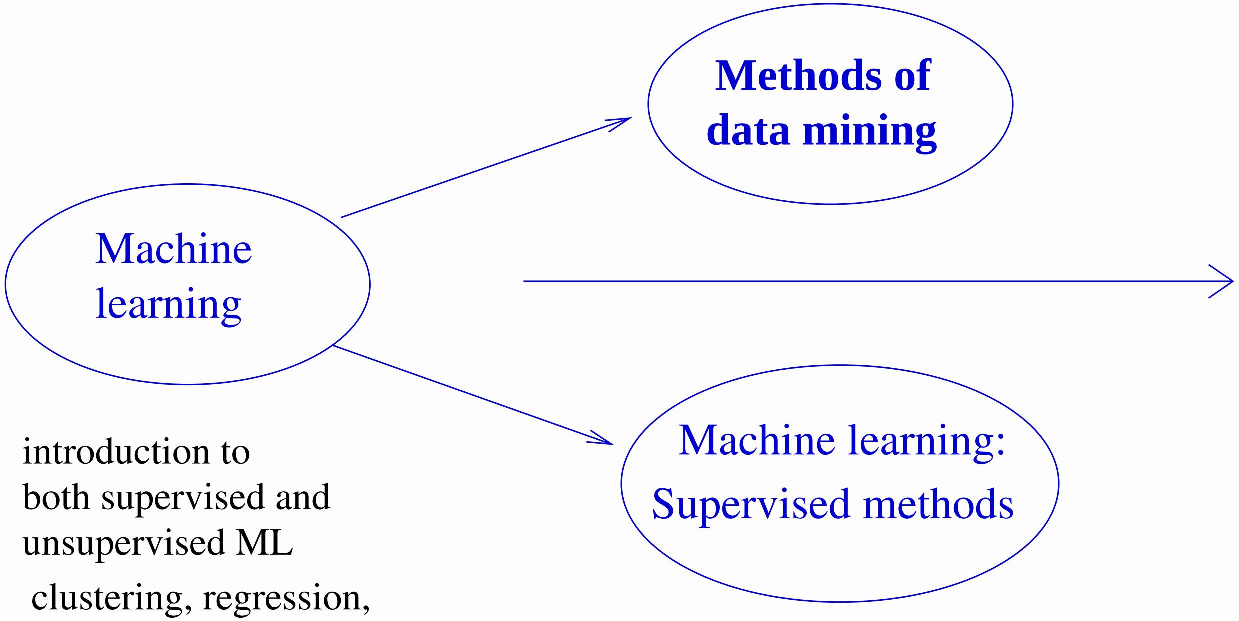
Syllabus

- Introduction to DM
- Data preprocessing
- Distance and similarity
- Clustering (extensions of K-family, hierarchical, spectral + evaluation)
- Association mining
- Graph mining
- Social network analysis
- Web mining and recommendation systems
- Text mining
- guests: Data randomization, Episode mining

Relationship to some other courses

pattern discovery

some unsupervised ML



Methods of data mining

Machine learning

Machine learning: Supervised methods

introduction to both supervised and unsupervised ML

clustering, regression, classification, ...

classification, linear models, ...

Special courses, e.g.,

Bayesian data analysis

ML: Advanced probabilistic methods

Complex networks

Research project in ML, DSC and AI

ML for bioinformatics

Statistical NLP

Kernel methods in ML

Deep learning

Prerequisites: Important!

1. Basic mathematics and statistics

- reading mathematical notations
- basic concepts of probability theory (distributions, conditional probability, independence, probability calculus)
- basic concepts of statistics (summary statistics like mean, median, variance, covariance, idea of statistical significance)
- basic matrix algebra (basic operations, some notion of eigenvalues and eigenvectors)

Prerequisites (cont'd)

2. Programming

- ability to process data and implement algorithms in some well-known programming language (Python, Java, C, C++, Matlab)

3. Algorithms and data structures

- reading pseudocode
- lists, trees, graphs etc.
- O-notation, NP-hardness
- basic algorithm strategies

Ask if you don't know something!

- utilize the **course forum**! It is most efficient!
  - channels for general/practical things, lectures and material, exercises, assigments
  - check extra clarifications, what others have asked and ask new questions
- ask during lectures
  `https://presemo.aalto.fi/mdm2023`
- take advantage of the exercise sessions
- read the textbook and extra materials
- make study groups with your colleagues
- use library and internet