MTAT.03.277 Research Seminar in Data Mining

How to pass this course

Sven Laur
University of Tartu
Why this course?

▷ I need to graduate but do not know how
  ◦ We introduce potential supervisors
  ◦ We introduce potential thesis topics
  ◦ We explain how academic research is done

▷ I am not sure whether a topic works out
  ◦ You can try out on small scale
  ◦ You get direct feedback on the plan
  ◦ You get at least 3 ETC for trying

▷ I am too lazy to start writing a thesis right now
  ◦ We provide intermediate deadlines
  ◦ The course structures your academic research
How to pass the course?

1. Topic conformation
2. THINK
3. Research plan defence
4. Deregistration
5. Final report
6. First draft
7. Presentation
8. Reviewing
9. Final report
10. First draft
11. Presentation
What are the topics?

- Theory
  - Advanced Statistics
  - Machine learning
- Academic research
  - Neuroscience
  - Bioinformatics
- Engineering
  - Big Data
  - Robotics
  - Medical data mining
- Natural language processing
- Medical data mining
- Geo data mining
- Machine learning
- Neuroscience
- Bioinformatics
- Big Data
- Robotics
- Medical data mining
What are the topics?

Object complexity

Training data size

Neuroscience
Bioinformatics
Medical records
NLP
Geo mining
Radar imaging
Robotics (SI)
Finance
Medical bills

MTAT.03.277 Research Seminar in Data Mining, How to pass this course, September 13, 2018
What are the topics?

- Description tasks
  - Bioinformatics
  - Neuroscience
- Prediction tasks
  - Medical data
  - Geo mining
  - Robotics (SI)
  - NLP
  - Finance
  - Radar imaging
What are the topics?

**Background knowledge**
- Bioinformatics
- Neuroscience
- Medical data

**Statistical knowledge**
- NLP
- Geo mining
- Robotics (SI)
- Radar imaging
- Finance
How do I contact supervisors?

https://www.cs.ut.ee/et/kontakt/arvutiteaduse-instituut

▷ Natural Language Processing
  ◊ Mark Fishel, Kairit Sirts

▷ Bioinformatics
  ◊ Dmytro Fishman, Jaak Vilo

▷ Neuroscience
  ◊ Raul Vicente, Ardi Tampuu

▷ Machine learning
  ◊ Meelis Kull

▷ Medical data mining
  ◊ Sven Laur, Sulev Reisberg, Jaak Vilo

▷ Big data
  ◊ Sherif Sakr

▷ Software engineering
  ◊ Ezequiel Scott, Dietmar Pfahl, Marlon Dumas
Topics
Medical data analysis

- EHIF billing data (7 years, 1M patients, ~100M records)
  - econometric
  - average price comparisons
  - disease trajectories
  - anomaly detection

- EGV epicrisis data (7 years, 50k patients, ~1M records)
  - data cleaning
  - fact extraction with NLP
  - disease trajectories
Engineering projects

▷ Regime changes in time series (Kappazeta)
  ◦ When the farmers harvest the crop?
  ◦ Current model does not work for drought years
  ◦ Transfer-learning task

▷ Analysis of drone telemetry (XXX)
  ◦ How long the drone can fly?
  ◦ Data is semi-secured
  ◦ Time-series analysis