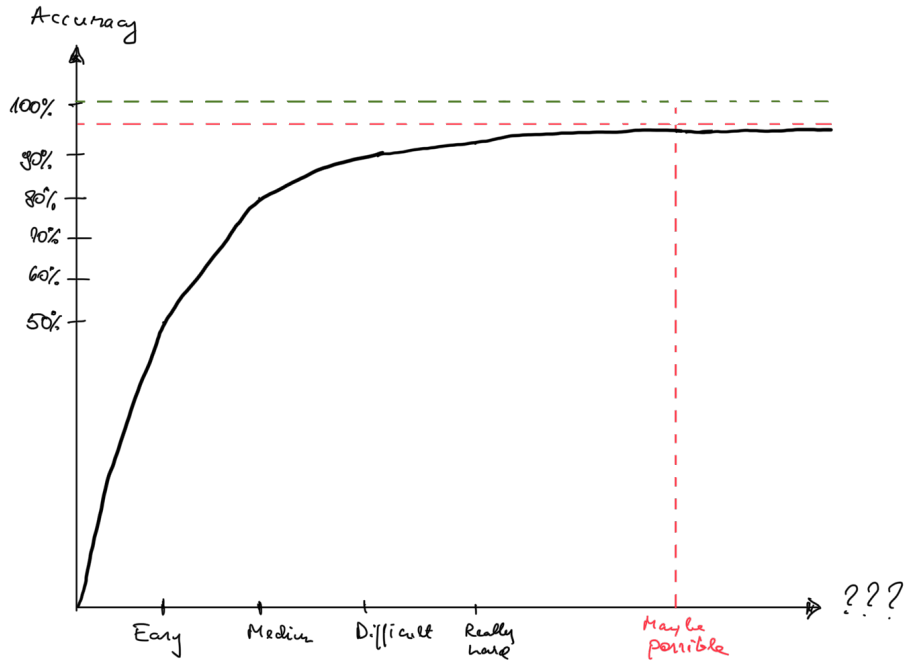
LTAT.02.004 Machine Learning II

Introduction

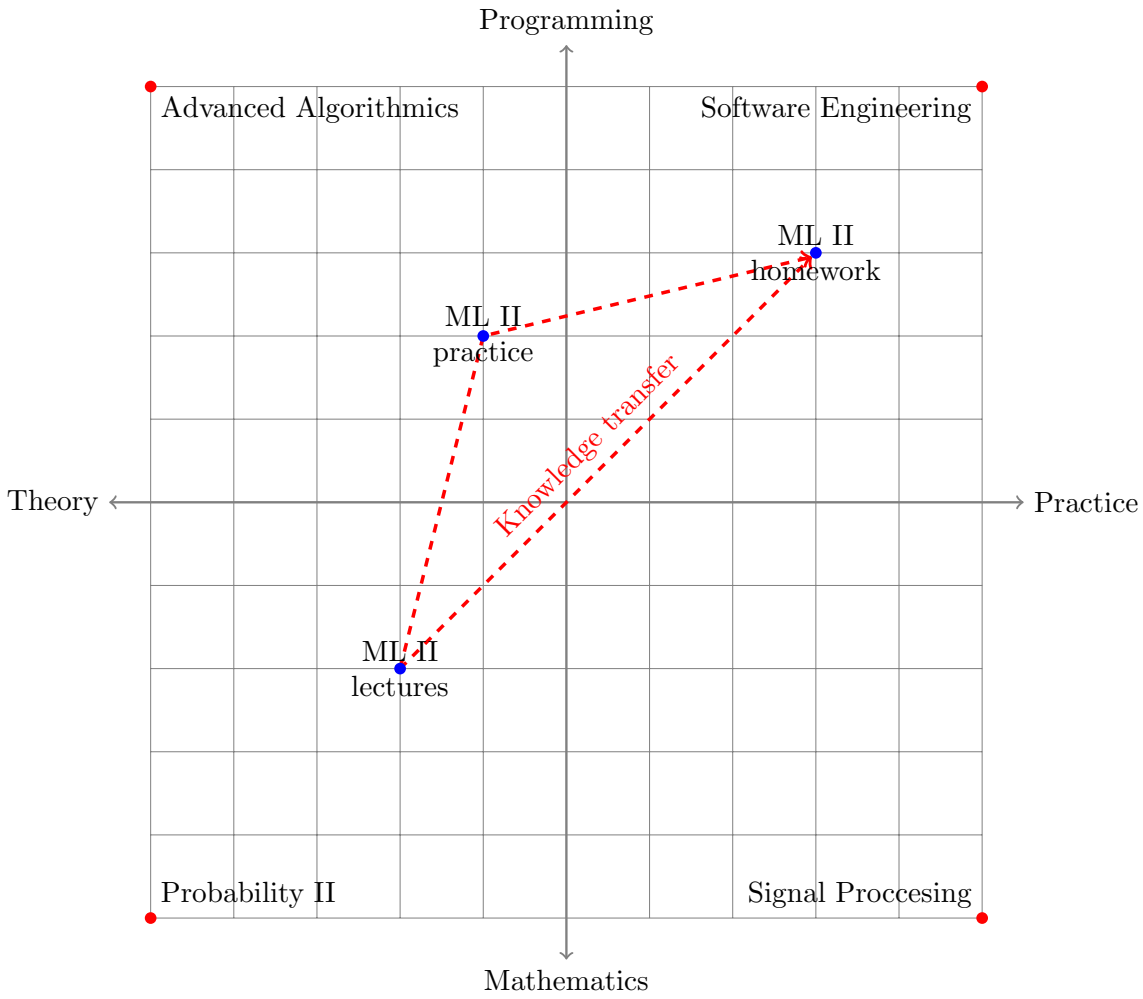
Sven Laur
University of Tartu

Why there are no self-driving cars?



- ▷ Gathered data, Background knowledge
- ▷ Computational resources, Learning algorithms
- ▷ Uncanny valley, Public perception

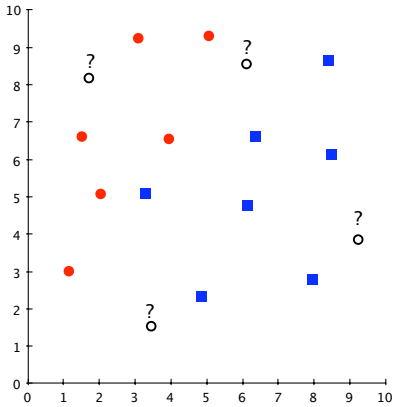
What is this course about?

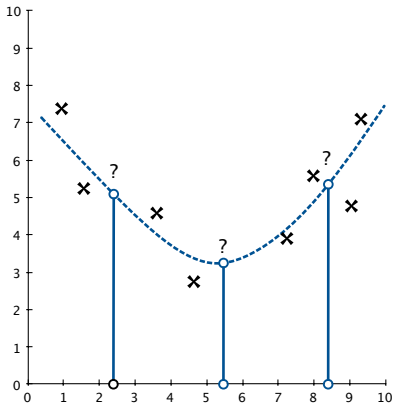


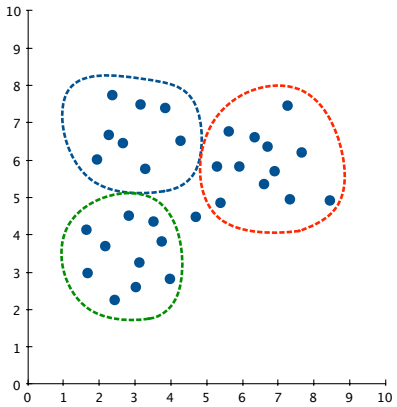
Course plan

- ▷ **Performance evaluation**
  $\rightsquigarrow$ absolute risk, relative risk, CLT, crossvalidation, bootstrap
- ▷ **Probabilistic modelling**
  $\rightsquigarrow$ frequentism, bayesianism, statistical tests, confidence intervals
- ▷ **Sequence models**
  $\rightsquigarrow$ Markov chain, Hidden Markov Mode, belief propagation
- ▷ **Multivariate normal distribution**
  $\rightsquigarrow$ Models behind linear regression and linear time series analysis
- ▷ **Affine transformations**
  $\rightsquigarrow$ Blind Source Separation, PCA, LDA, ICA, NMF
- ▷ **Model based clustering**
  $\rightsquigarrow$ Hierarchical clustering, k-means, gaussian mixture model
- ▷ **Expectation-maximisation algorithm**
  $\rightsquigarrow$ weighted data, gaussian mixture model
- ▷ **Expectation-maximisation algorithm for sequential models**
  $\rightsquigarrow$ Hidden Markov Model, Kalman filter

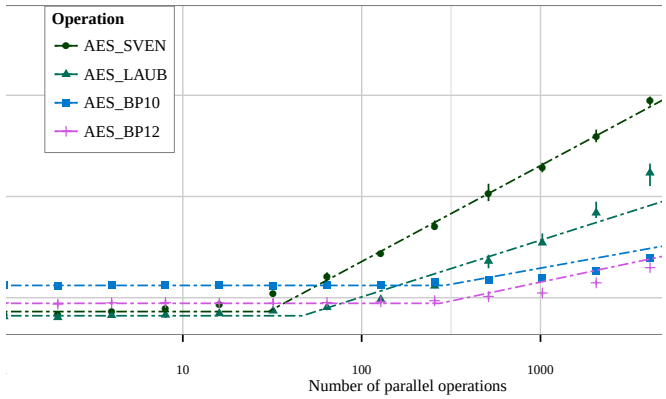
Four basic tasks in machine learning

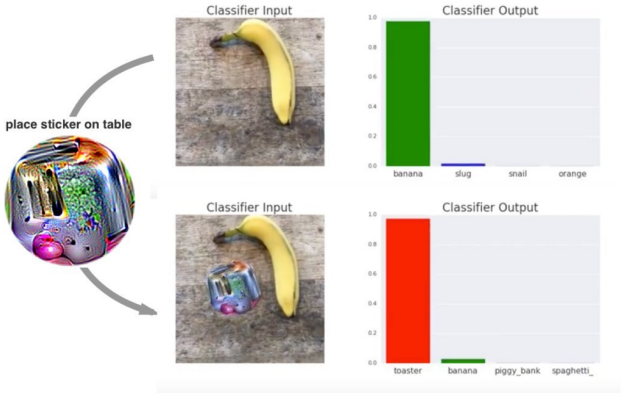


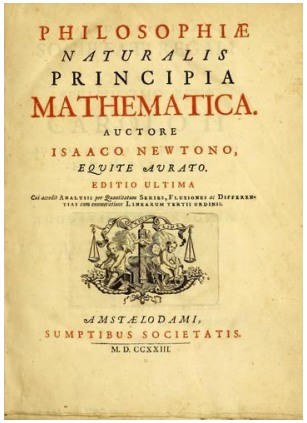


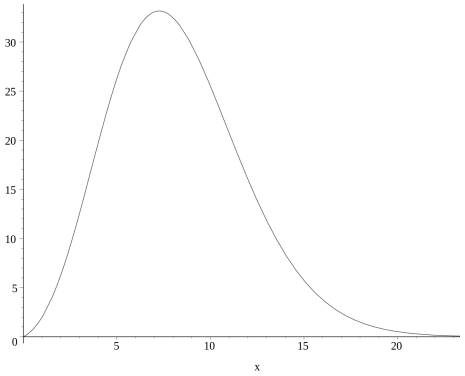


Basic issues you have to solve

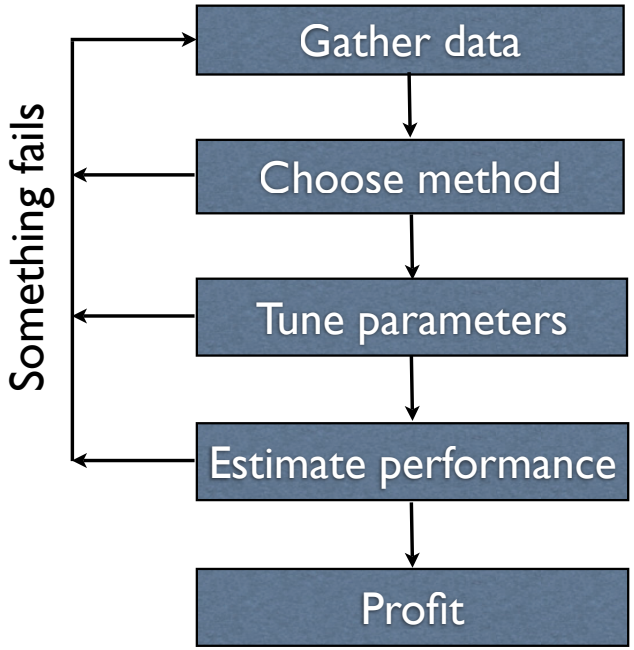








Main inference prodedure



Usually no machine learning method works on real data without tweaking

- ▷ The signal might be missing form the data
- ▷ The method uses wrong features for its predictions

Data gathering is programming

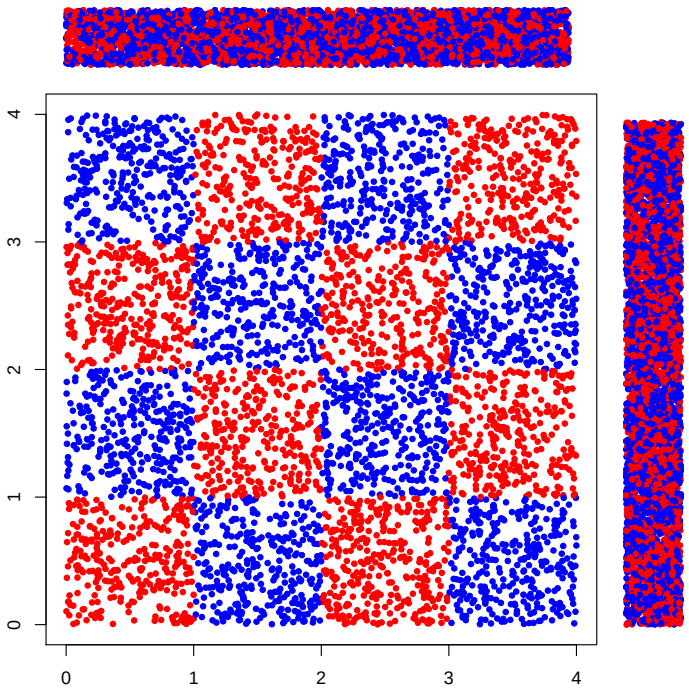
How to increase volume

- ▷ Self-labelled data vs manual labelling
- ▷ Symmetries and data augmentation
- ▷ Consistency and data augmentation
- ▷ Backgroung knowledge and data augmentation

Software engineering practices

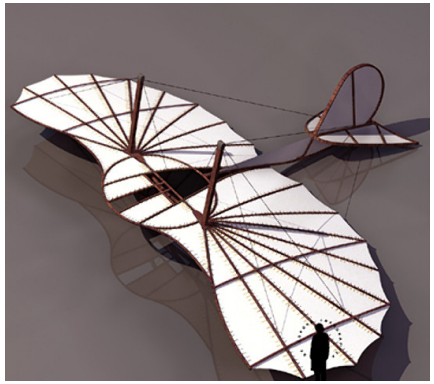
- ▷ Version control $\rightsquigarrow$ Data and model versioning
- ▷ Code differences (blame) $\rightsquigarrow$ Model comparison
- ▷ Bugtracking $\rightsquigarrow$ Model diagnostics & Abalation studies
- ▷ Tests $\rightsquigarrow$ Dedicated testsets
- ▷ Automatic testing $\rightsquigarrow$ Robustness analysis and Advdersarial learning

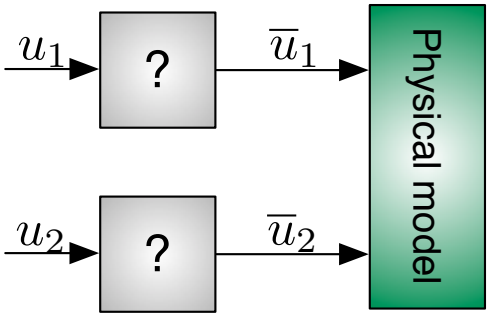
Features are more important than method



The signal is completely lost if we observe a single feature: x-coordinate or y-coordinate. By knowing both features the pattern is clearly visible.

Do not learn what you already know!





Sometimes we know the overall structure of the model

- ▷ In robotics the effect of actuators can be expressed directly
- ▷ Sometimes we know some governing rules form previous studies

In such cases, learning the entire model with machine learning is wasteful

- ▷ Locate the parts of the model that are undefined
- ▷ Use machine learning to find missing links