DATA11002 Introduction to Machine Learning

Kai Puolamäki

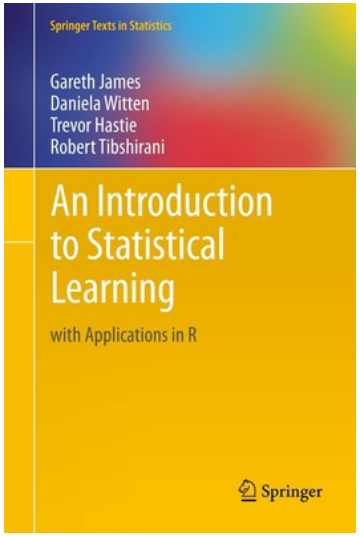
28 October 2020

Organisation and practicalities

- See the course Moodle page, https://moodle.helsinki.fi/course/view.php?id=41607

The lecture slides on this and follow-up lectures use partly material by Antti Ukkonen, Patrik Hoyer, Jyrki Kivinen, and Teemu Roos. Thanks!

Book



Available as a pdf file from http://www.statlearning.com/

We'll cover roughly the whole book except splines and generalized additive models (GAMs) – and include some additional Bayesian stuff

(See the Moodle for alternate source materials!)

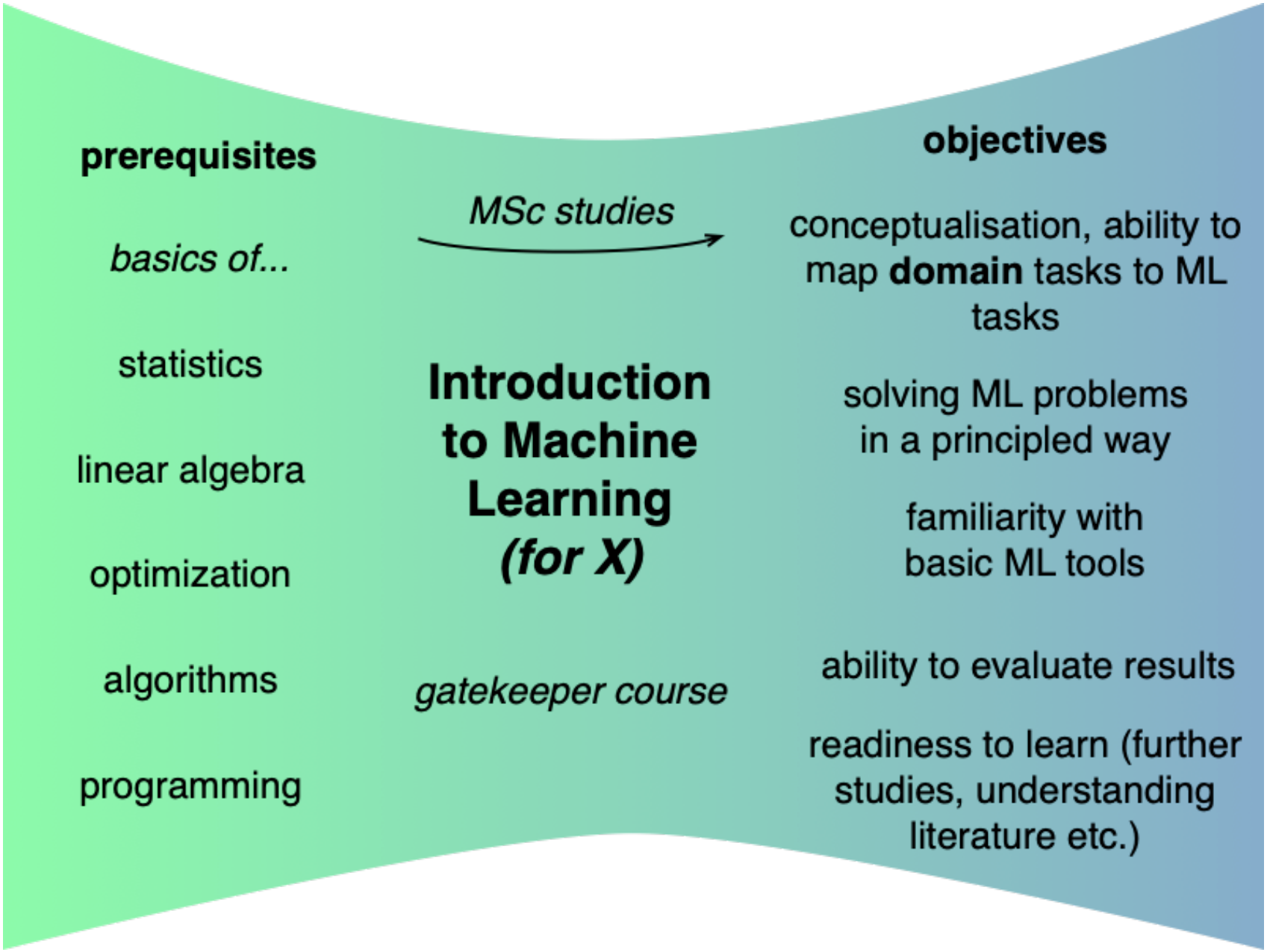
What is ML and why should you care?

- The core technology that underlies all of the big data and AI fuzz
- Something that guarantees you a well-paying job (or so we wish :)
- Prevalent and transformative methodology in sciences and humanities

Why in Helsinki?

- ML / AI is one of the strengths at University of Helsinki + Aalto University
- Helsinki Institute for Information Technology HIIT
- HiDATA - Helsinki Centre for Data Science
- Finnish Center for Artificial Intelligence FCAI
- Lots of jobs in private sector as well

Position in curriculum and learning objectives



Follow-up machine learning courses

- Advanced Course in Machine Learning
- Bayesian Machine Learning
- Computer Vision
- Introduction to Deep Learning
- Design and Analysis of Algorithms
- Information Retrieval
- Network Analysis
- Many seminars also have a strong machine learning flavour!

more info

Contents of the course

View 1

- ingredients of machine learning
  - what is machine learning? (overview)
  - tasks, model, data, statistics, algorithms, optimization
  - evaluation of algorithms
- supervised learning
  - regression, classification
- unsupervised learning
  - clustering, dimensionality reduction

View 2

- tasks
  - supervised learning (regression & classification)
  - unsupervised learning (clustering & dimensionality reduction)
- models
  - linear models
  - probabilistic modelling
  - algorithmic models (kNN, tree-based models etc.)
- statistics and evaluation
  - how to evaluate errors and confidence algorithms and optimization
  - greedy algorithms tree algorithms

Elements of ML task

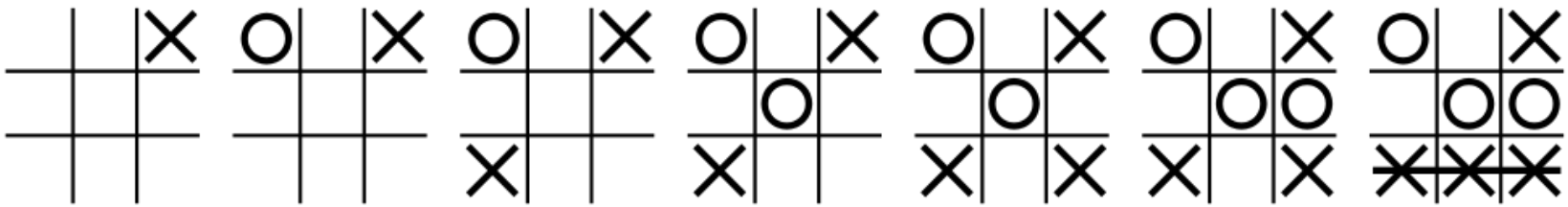
- Approach 1: download some software, run, have results!
- Approach 2: understand what you are doing

What is machine learning?

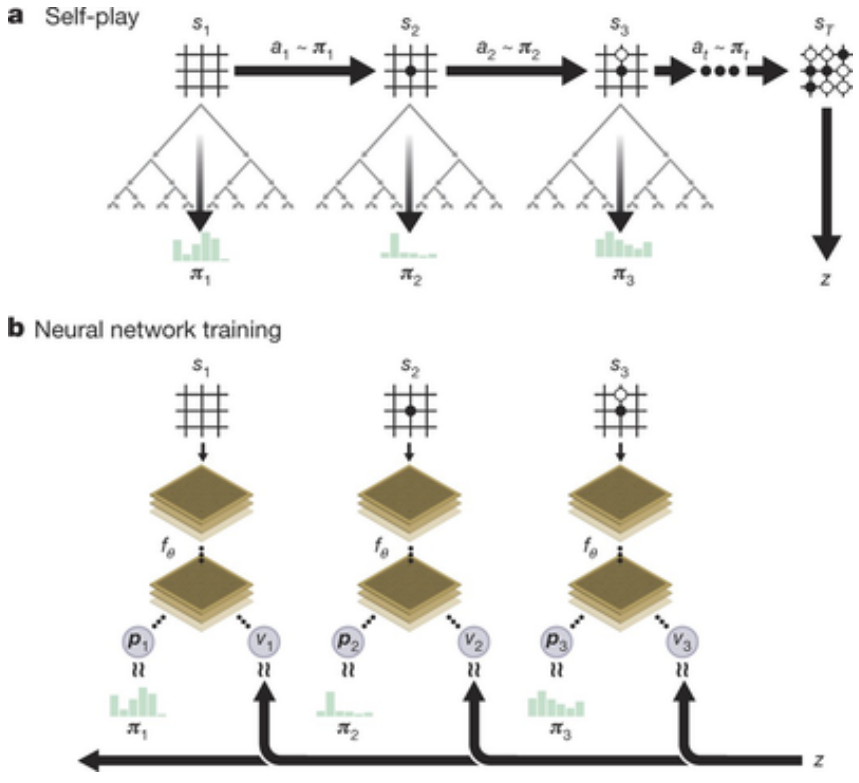
- **machine** = computer and computer progrem (in this course)
- **learning** = improving performance on a given task, based on experience / examples
- In other words:
  - instead of programming explicit rules, computer learns from examples.
  - often, computer will reach better performance than the programmer ever could!

Example 1: tic tac toe

- *Option A:* The programmer writes explicit rules, e.g., "if the opponent has two in a row, and the third position is free, place your mark there', etc
- *Option B:* Go through the game tree, choose optimally
- *Option C:* Let the computer try out various strategies by playing against itself and others, and noting which strategies lead to winning and which to losing (=**machine learning**)



Example 2: alpha go zero



- Computer plays against itself and learns (no data needed!)
- Figure from Silver et al. (2017) Nature.

Example 3: spam filter

- **Classify** email messages as spam or ham (=not spam) based on message features
- Example: SpamAssassin https://spamassassin.apache.org/
  - Uses rule outputs and features and (historically) Naive Bayes classifier, nowadays perceptron to adjust rule weights
  - Has to work in *adversarial world* where the spammers are adapting (trying to circumvent the filter)
- Analogous problem: detecting malware

Example 3: regression

- Predict precipitation
- Figures from Eronen, Puolamäki et al. (2010) Evolutionary Ecology Research (a, b)

Example 3: regression

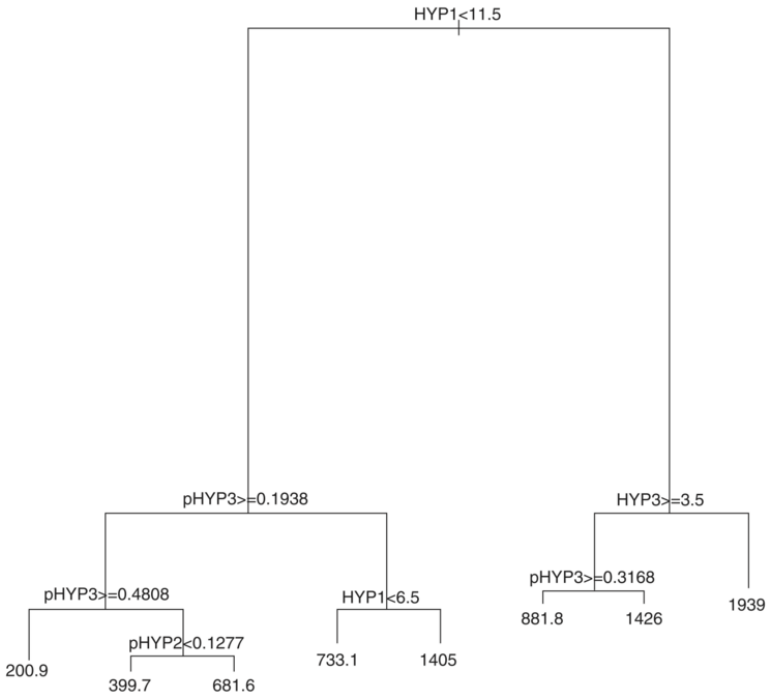
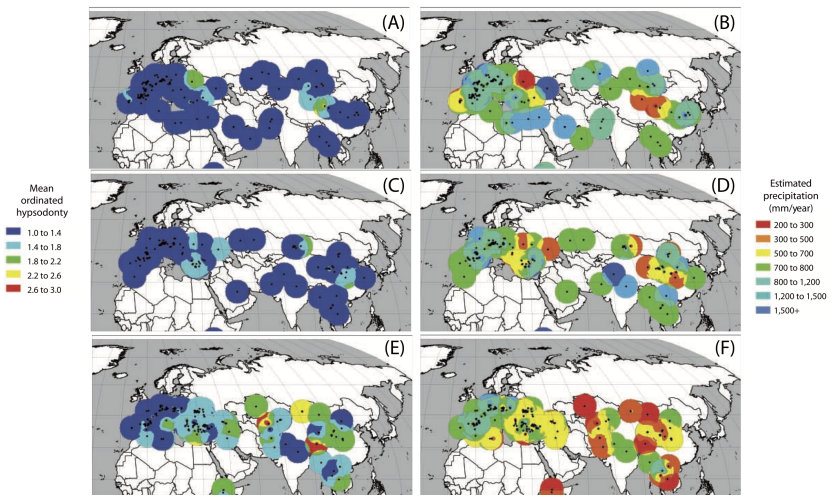


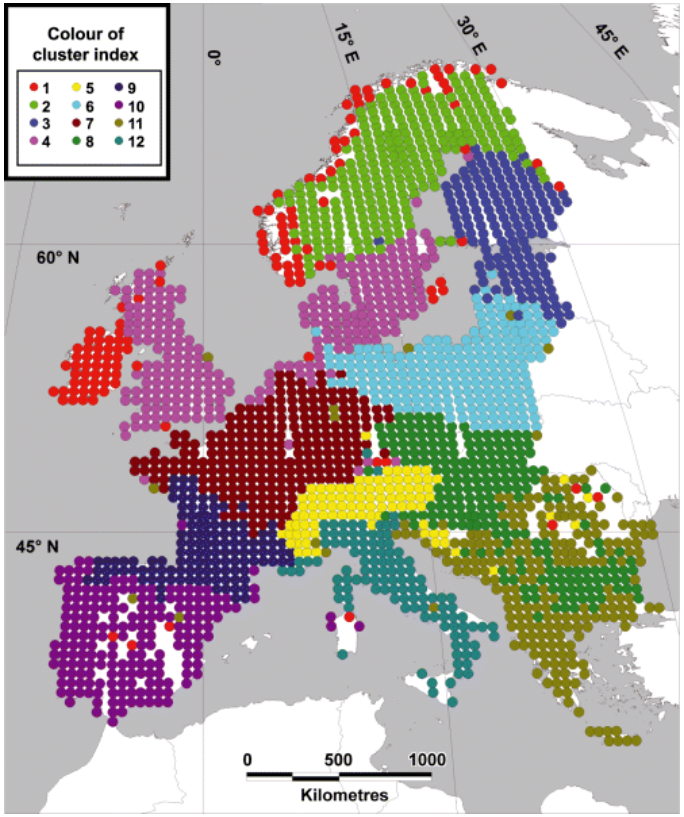
Fig. 1. Decision tree for annual precipitation using hypsodonty alone as regressor (see online Appendix 1 for other decision trees generated: evolutionary-ecology.com/data/2538A1.pdf).

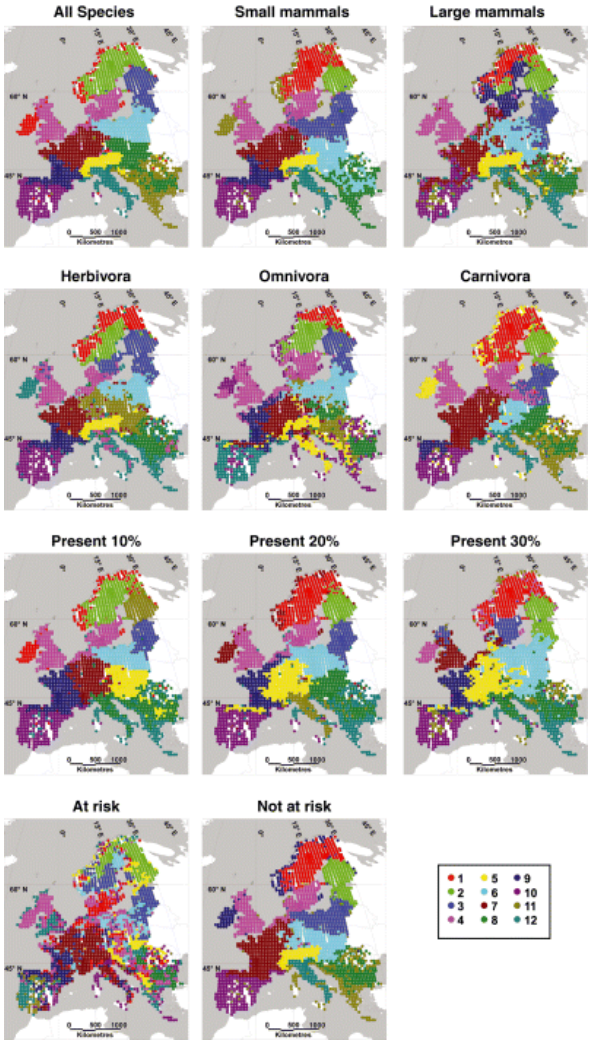


Example 4: clustering

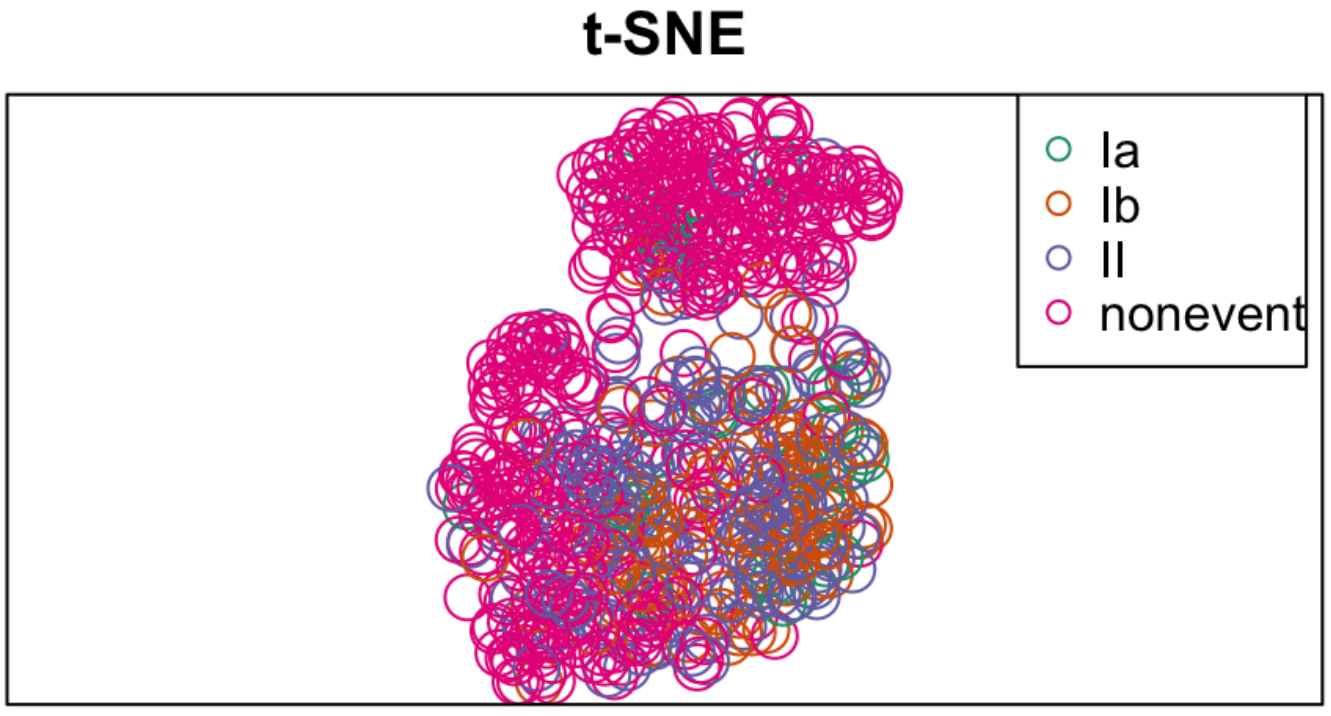
- Question: Can we find ecological communities?
- Question: What explains the communities?
- The 50 $\times$ 50 km map grids were grouped into clusters. Map grids within a cluster should occupy similar mammals.
- Figures from Heikinheimo et al. (2007) J. of Biogeography.

Example 4: clustering (unsupervised learning)





Example 5: t-SNE embedding of Hyytiälä measurements (unsupervised learning)



Problem setup

- One definition of ML: A computer program improves its **performance** on a **given task** with **experience** (i.e., **examples**, **data**).
- So we need to separate
  - *Task*: What is the problem that the program is solving?
  - *Performance measure*: How is the performance of the program (when solving the given task) evaluated?
  - *Experience*: What is the data (examples, **features**) that the program is using to improve its performance?

Related disciplines

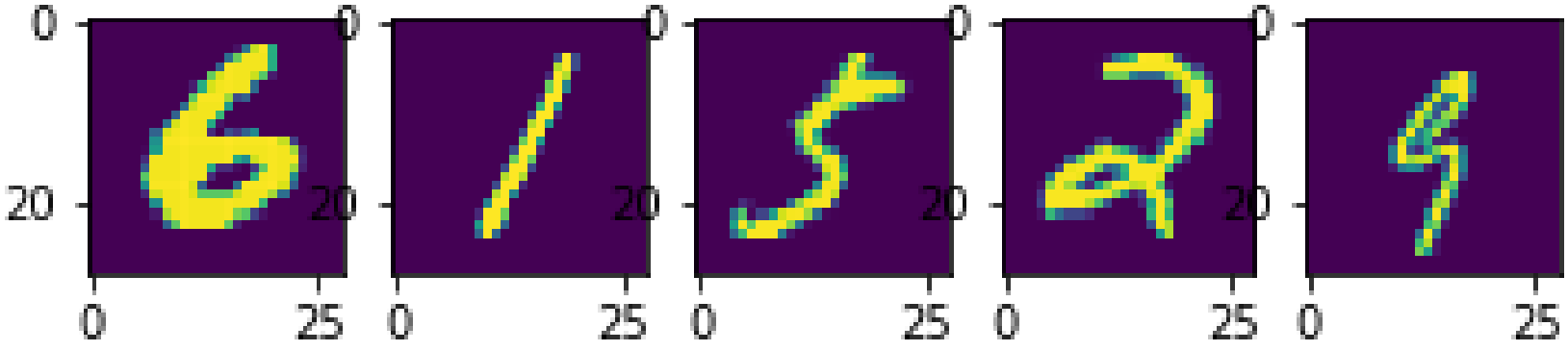
- Artificial intelligence (AI)
  - ML is a central technology that underlies AI
- Artificial neural networks, computational neuroscience
  - Modern ML grew from NN boom during the past millenea
- Pattern learning
  - recognizing objects and patterns, typically implements ML techniques
- Data mining
- Statistics
- Algorithms (computer science)
- Optimization

Deep learning

- "Family" of machine learning methods.
- Has become incredibly popular in the past few years.
- Yields very good results, e.g., in computer vision and speech recognition tasks.
- Heavily based on classical work on artificial neural network methods. (Fundamentally not really that novel.)
- Basic principles of ML covered in this course also apply to deep learning!

Example 6: MNIST digits and deep learning

- Machine learning software is already quite sophisticated
- It is relatively easy to do complex tasks that 10 years ago required PhD in computer science
- Example: Keras, TensorFlow, CNN, and mnist images (following example at https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py)
- MNIST: train CNN using 60000 28x28 pixel handwritten digits 0-9. Classification task: given a new digit, try to guess the number 0-9 it is supposed to be.



Example 6: MNIST digits and deep learning

- CNN (convolutional neural network): training time 13 minutes, accuracy 0.996 (fraction of numbers in test set classified correctly)
- The same task with some other classifiers:
  - logistic regression: training time 8 s, accuracy 0.925
  - random forest: training time 37 s, accuracy 0.969
  - k nearest neighbours (kNN): training time 1 s, accuracy 0.967
  - support vector machine (SVM): training time 33 s, accuracy 0.955
- Why neural networks often have good accuracies on images?
  - trained on large annotated data sets
  - relatively low level features (image pixels)
- Lots of applications

Issues: black box classifiers

(a) (b)

Figure 5: Adversarial examples generated for AlexNet [9].(Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example. All images in the right column are predicted to be an *"ostrich, Struthio camelus"*. Average distortion based on 64 examples is 0.006508. Plase refer to `http://goo.gl/huaGPb` for full resolution images. The examples are strictly randomly chosen. There is not any postselection involved.

Figure from Szegedy et al. (2013) arXiv:1312.6199 [cs.CV]

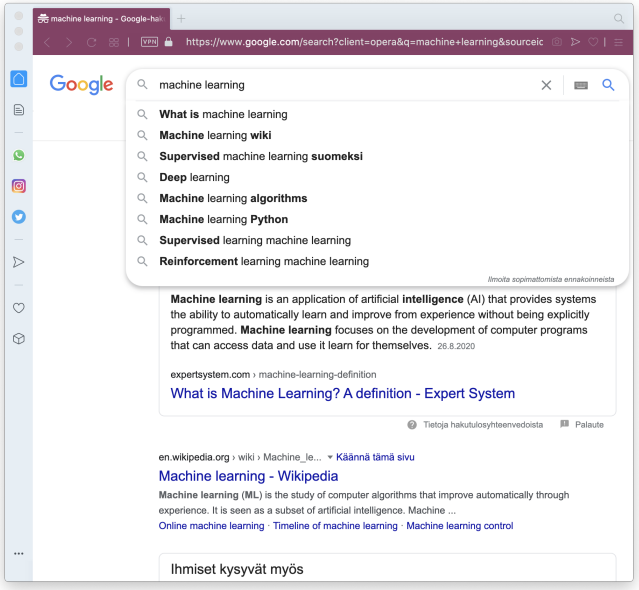
Example 7: self-driving cars

- Self-driving cars:
  - Sensors (radars, cameras) superior to humans
  - How to make the computer react appropriately to the sensor data?
  - Note: The sensors can be broken and deliver incorrect/broken data.
  - Adversarial attacks on computer vision systems!

Credit: Grendelkhan CC BY-SA 4.0

Example 8: search engines

- Ranking search results:
  - Various criteria for ranking results
  - What do users click on after a given search? Search engines can learn what users are looking for by collecting queries and the resulting clicks.
- Predictive queries
- Ad placement
  - Which ads to show to maximise profits?



Example 9: recommendation systems use collaborative filtering

- Amazon: "customers who bought X also bought Y..."
- Netflix: "based on your movie ratings you might enjoy..."
- Spotify: "discover weekly playlists"

Example 10: analysis of social networks

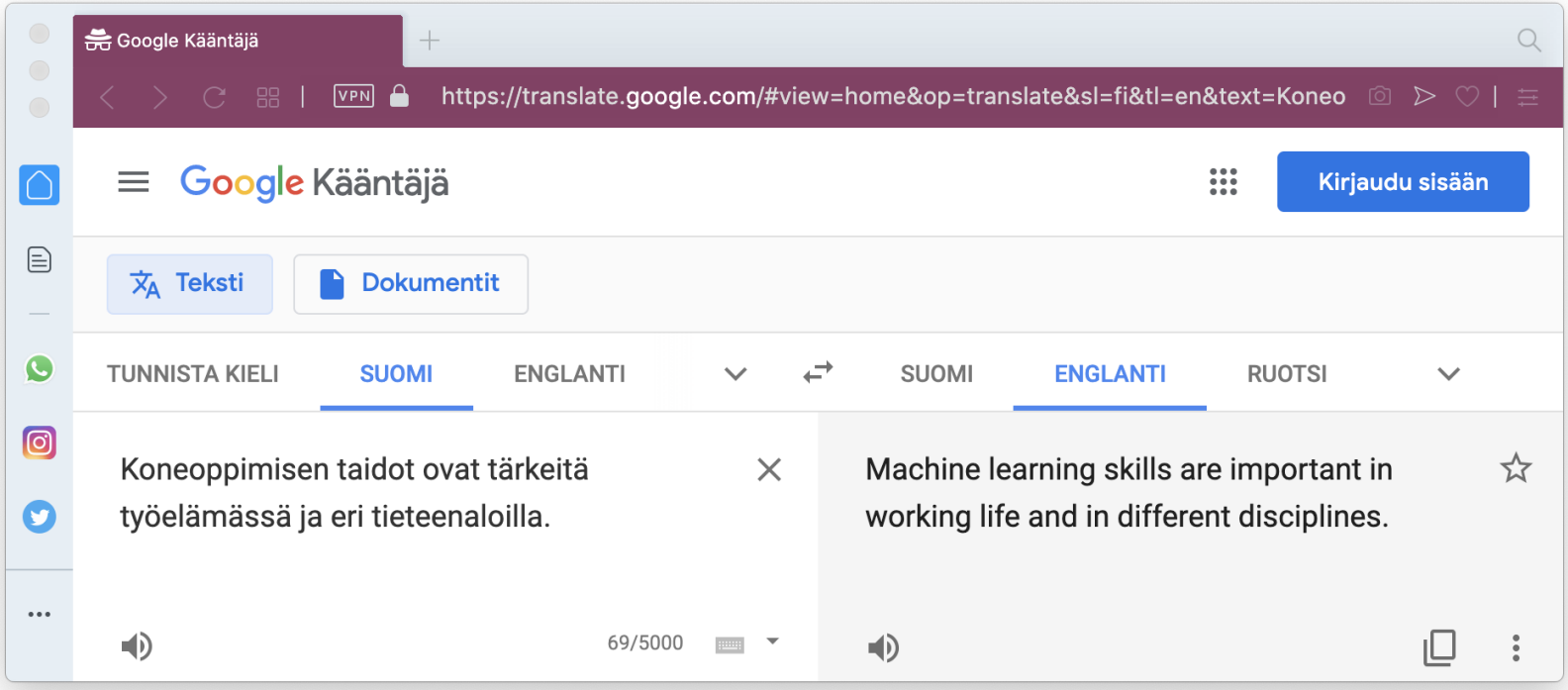
- Prediction of friends in Facebook, or prediction of who you'd like to follow on Twitter.
- Increase appeal of social network → more profit

Issues: privacy

- Users are surprisingly willing to sacrifice privacy to obtain useful services and benefits
- Regardless of what position you take on this issue, it is important to know what can and what cannot be done with various types information (i.e., what the dangers are)
- *Privacy-preserving data mining*
  - What type of statistics/data can be released without exposing sensitive personal information? (e.g., government statistics)
  - Developing data mining algorithms that limit exposure of user data

Example 11: machine translation

- Old: hard-code grammar etc.
- Traditionally: statistical machine translation
- Nowadays: deep learning based methods



Example 12: business analytics

- Find anomalous bank or credit card events
  - detecting anomalies and frauds is important
- How to make smart decisions?
  - how to price car insurance based on customer data?
  - how to use only "allowed" information (e.g., income is ok, gender is not)?
  - how to explain decisions?
- ...

Availability of data

- These days it is very easy to
  - collect data (sensors are cheap, much information digital)
  - store data (hard drives are big and cheap)
  - transmit data (essentially free on the internet)
- Result: everyone has large quantities of data
  - how to deal with it?

Example 13: mining chat and discussion forums

- Breaking news
- Tracking consumer sentiment about companies / products
- Detecting outbreaks of infectious disease (by mining search queries, next slide)

Issues: parable of big data

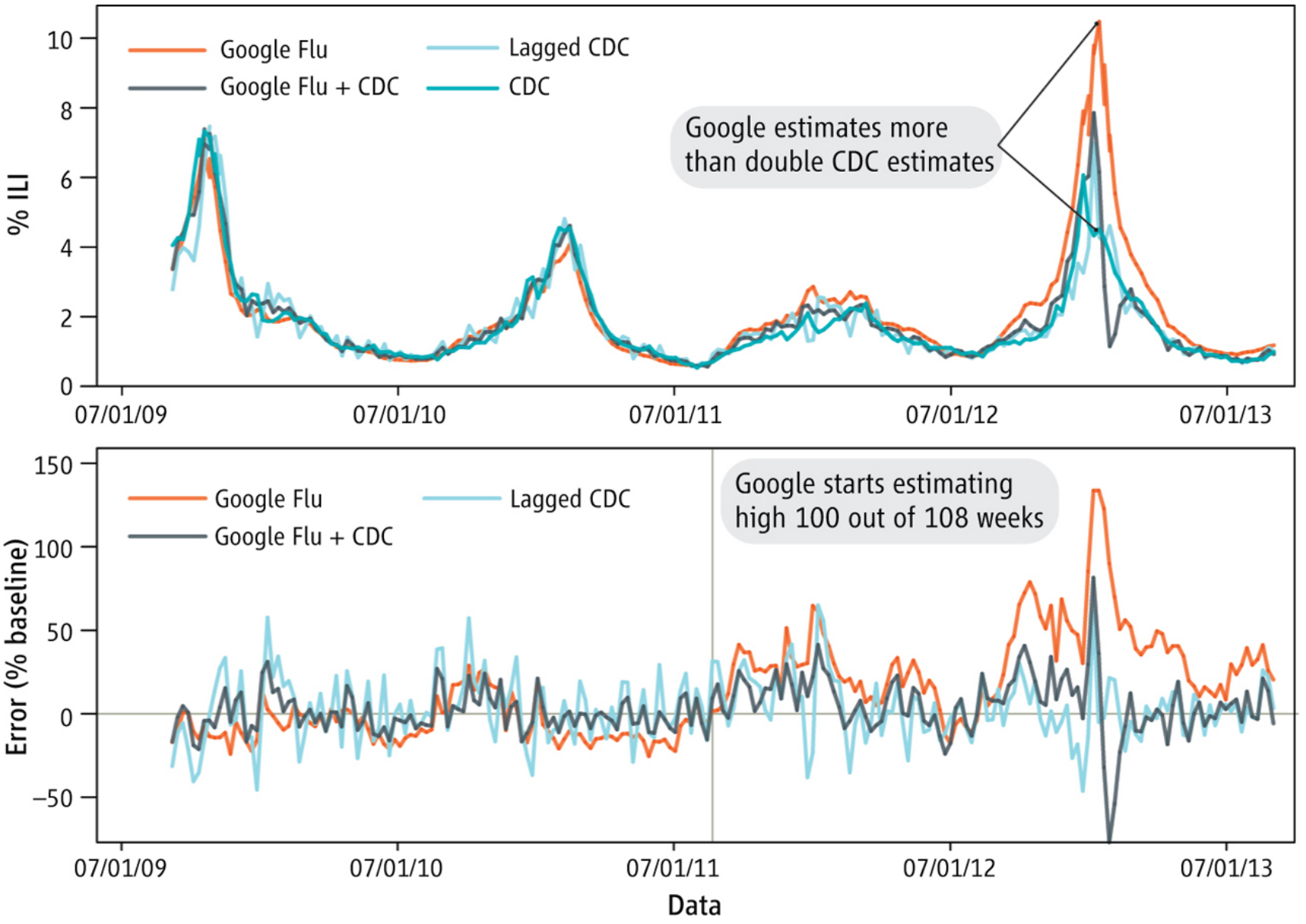


Figure from Lazer et al. (2014) Science

Resources: datasets

- UCI Repository:
  http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/
- Smartsmear: https://avaa.tdata.fi/web/smart/smear

Resources: ML journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Trans on Neural Networks and Learning Systems
- IEEE Trans on Pattern Analysis and Machine Intelligence
- Journals on Statistics/Data Mining/Signal
- Processing/Natural Language
- Processing/Bioinformatics/...

Resources: ML conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NeurIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on Artificial Neural Networks (ICANN)
- International Conference on AI & Statistics (AISTATS)
- International Conference on Pattern Recognition (ICPR)
- ...