Mi a mesterséges intelligencia? – Alapvető MI technológiák

Dr. Toka László

Távközlési és Mesterséges Intelligencia Tanszék
Villamosmérnöki és Informatikai Kar
Budapesti Műszaki és Gazdaságtudományi Egyetem

2025. május 13.

Data Mining – Data Science

- Data Mining
  - The (semi)automatized extraction of new and useful non-trivial interrelation from large amounts of data
  - Basically: finding patterns
  - Aim: support decision-making, understand the data and the world better
- Data Science
  - The same(ish) but sounds more scientific :D
    - Research area
    - Non-relational data as well

Automatized?!

- Yes, we can automatize the following:
  - Extracting patterns
  - Inferring values
  - Predicting events
  - Validating hypotheses
  - …
- No, we might have to do the following:
  - Describe the task
  - Select method, statistics, machine learning
  - Interact with the modeling method
  - …

Machine learning

- Based on
  - Prejudice
  - Skepticism (a lot)
  - Statistics
  - Linear algebra
  - Pure luck :), because there is randomness
    - In the data
    - In the methodologies
- Rhymes with **ARTIFICIAL INTELLIGENCE**
  - Intelligence ← learning
  - Artificial intelligence ← machine learning

Learning

- Forming
- Maintaining
  - Replacing
  - Tuning
- Reinforcing

using

- Perception, sensing
- Thinking

- Knowledge
- Behavior
- Ability
- Skill
- Values
  - Preference
  - Strategy
  - Politics

Some basic tasks

Supervised Learning (Learning a label)

Basic tasks: classification, regression

- Given dataset
  - With entities and the label in question
- Searching for the relation ≈ f(X) → y
  - Or the *best* possible approximation
- Want to use that to deduce the label of others

Questions

- What should be considered as an attribute?!
- It seems true on the dataset but will it stay true?!
- Best how?!

Unsupervised Learning (Learning a representation)

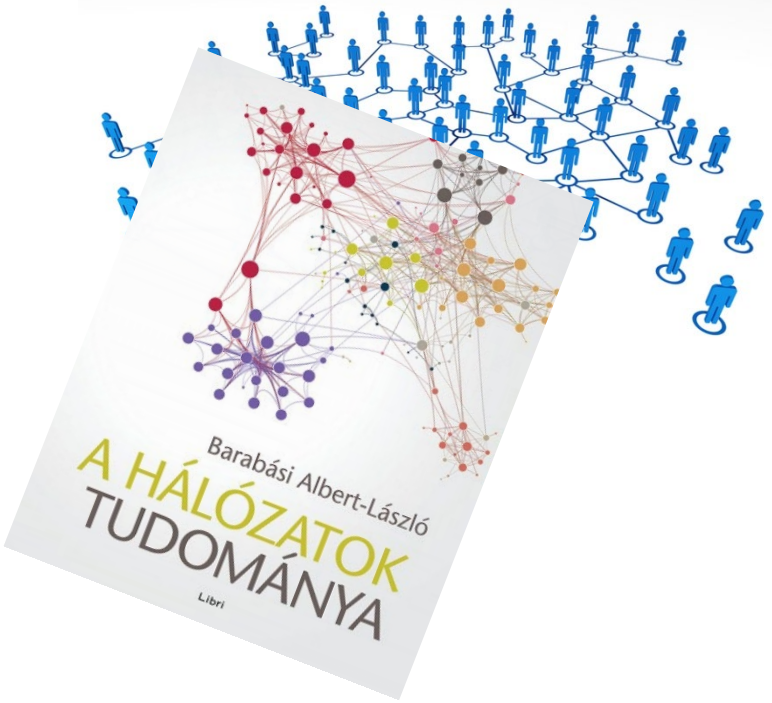
Basic tasks: clustering

- Sorting into groups
  - Similar to the same group
  - Differents to distinct groups
- A lot of ways to proceed, none is perfect

Questions:

- How to decide whether it is good enough?
- How to make a grouping better?
- Does the result mean anything at all?!

Networks



Graph

- Nodes and edges
- Values and functions
- Temporal property?

Questions

- Can an edge effect an attribute of a node?
- Can a neighbor effect an attribute of a node?
- Can a n's neighbor effect …?
- Can a n's n's … neighbor effect…?

Anomaly detection

Outliers

- Rare items
- Significantly different from majority
- Time series analysis

Questions

- What is normal?
- What difference is significant?
- Remove beforehand?

Some basic modeling methods

k-NN for classification

k nearest neighbors

- Get the k most similar data points from the training
- They vote by majority

Lazy modeling

- We did not build a model, it’s the data itself

Linear regression

Linear (combination)

- $f(\boldsymbol{x}) = w_0 + \sum w_i x_i$

Ordinary Least Squares

- Minimize the sum of squared errors
- The line (hyperplane) gives the prediction of any input

k-means clustering

Start from k random data points as the centers

- Assign each training sample to the closest center
- Calculate the new mean for each group, start over

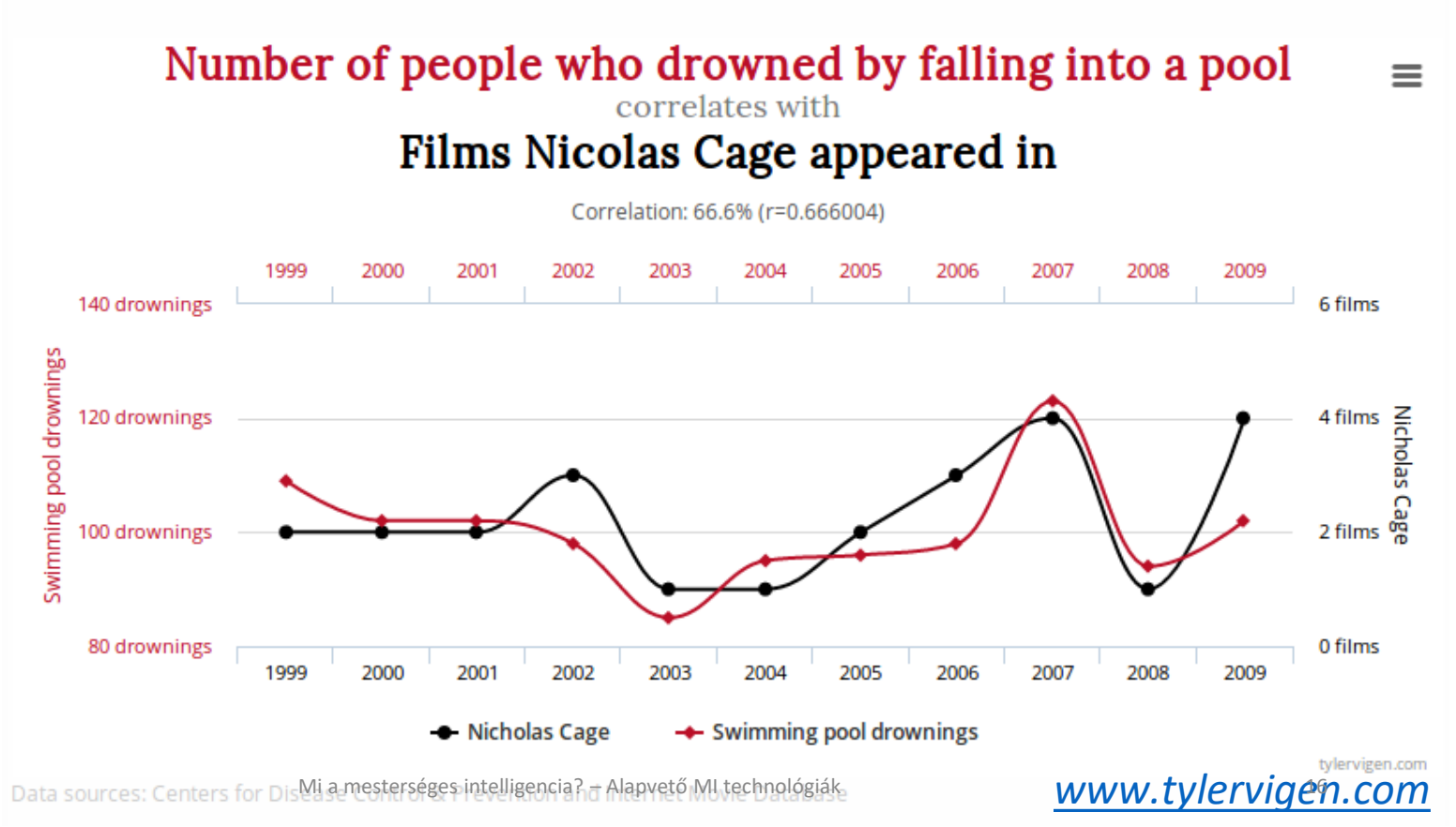
Iterative!

- …as a lot ML algorithms
- Continue as you want (until criteria)

Learning problems

Some problems one should be aware of

Correlation vs. Cause and effect



www.tylervigen.com

Eternal truth

Statement

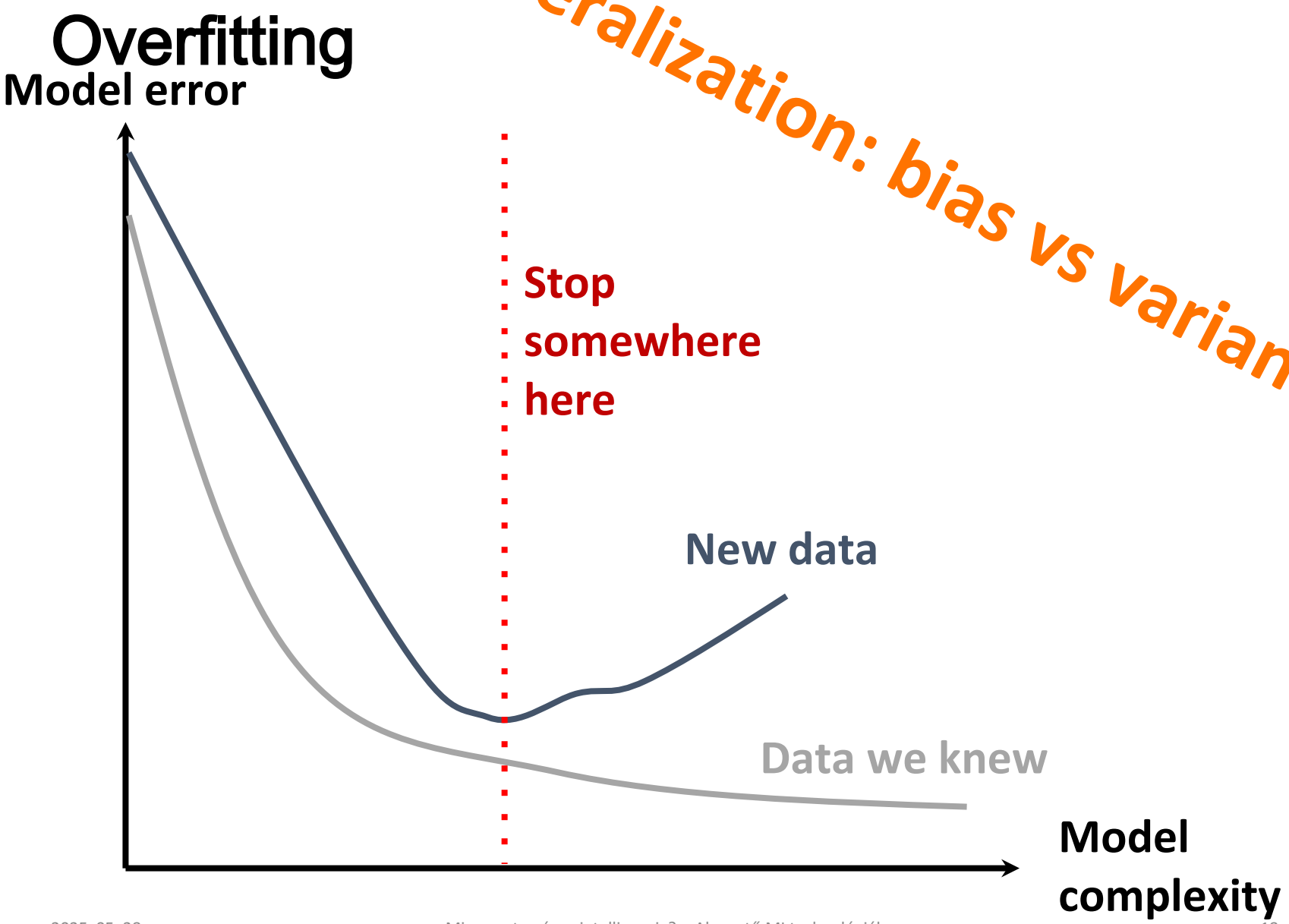
- A model will degrade over time
  - Especially when it comes to human nature
- this is true in most cases

Multiple levels

1. Build a model once
2. Build a model from time to time
3. Rebuild when it degrades
4. Reinforcement learning

The algorithm will imply how painful this is…

Generalization: bias vs variance



Bad metric

Typical: accuracy in an asymmetric case

“Our accuracy is 98%”

- Sound good, eh?!
  - Let’s assume
    - 100,000 entities
    - Yes-No label
    - Reality is around 1000 Yes-es
  - Let’s predict No for all, that is around 99% accuracy…
- Still good?

Bad features

- I know more about the relation than I tell
  - If the features and the label can be brought closer…
  - If this simplifies the relation, the model can be simpler
    - And almost always better
- I know more about the attributes than I tell
  - 1, 2, 3, 4, 5, 6, 7
  - Is 1 more similar to 2 than it is to 5?!
  - Are those numbers or categories (e.g., days of the week)?!

What is there to cope with?

- Data availability, content, freshness
- Legal environment
- Costs
- Business knowledge
- Deployment
- Handling and modeling data

Toka László

toka.laszlo@vik.bme.hu