Methods and tools for AI applications

Ellák Somfai
Dept. of AI, Faculty of Informatics, ELTE

2024/25 semester 1

Course Admin

Course instructors:

- Ellák Somfai (lecture) somfaiellak@inf.elte.hu
- Balázs Nagy (practicals) nagybalazs@inf.elte.hu

Meetings:

- lecture: Thu 10:15-11:45(?) North -1.62
- practicals: varies varies

Online tools:

- canvas (canvas.elte.hu)
- teams - if necessary

Grading:

- midterm + exam, details in Practicals
  quizzes

AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment

AI and ML

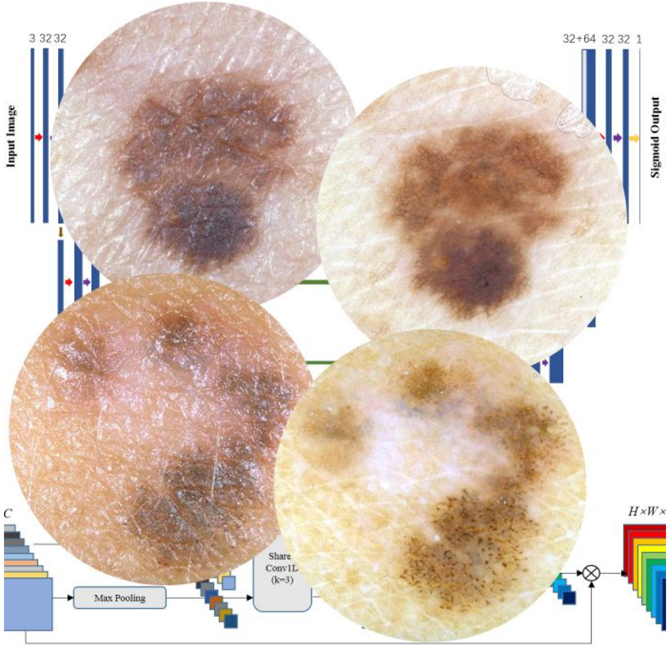
Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment

AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment



AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment

AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment

AI and ML

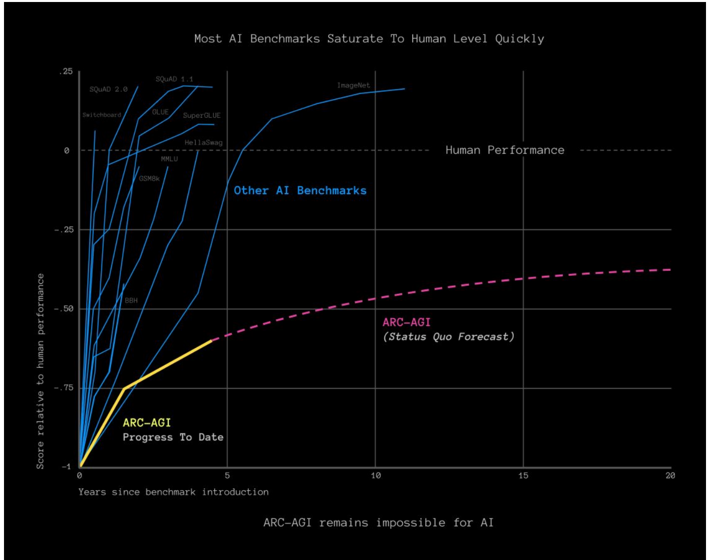
Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment
- eg. autonomous driving, medical diagnosis and treatment, intelligent robots with dexterity
- compare artificial vs human intelligence

AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment
- eg. autonomous driving, medical diagnosis and treatment, intelligent robots with dexterity
- compare artificial vs human intelligence



AI and ML

Artificial Intelligence:

- capability of machines to imitate intelligent human behavior
- eg. recognize visual scene, communicate with natural language, perform action in physical environment
- eg. autonomous driving, medical diagnosis and treatment, intelligent robots with dexterity
- compare artificial vs human intelligence

Machine learning:

- subfield of AI (currently the most important)
- learning without explicit programming, generalize from existing data
- supervised learning: using annotated (human labeled) data; regression, SVM, neural nets
- unsupervised learning: search patterns in unlabeled data; clustering, dimension reduction, anomaly detection
- reinforcement learning: learn from trial-and-error behavior using reward system

Need: model + data + computing infrastructure

Why need models? And Maths?

- E.Wigner: *“The Unreasonable Effectiveness of Mathematics in the Natural Sciences”* (1960)
  efficiency of the mathematical tools in describing physical phenomena

- R.Hamming: *“The Unreasonable Effectiveness of Mathematics”* (1980)
  more general discussion

“Everything should be made as simple as possible, but not simpler.” - A.Einstein

- Models based on Mathematics are essential to describe the world around us,
  including all aspects relevant to AI.

Purpose of this course: **provide a (review of the) mathematical tools needed to understand and perform research in AI.**

Including: calculus, linear algebra, probability, statistics.

Lecture outline

1. Introduction
2. Calculus revision
3. Linear algebra revision
4. Probability introduction
5. Statistics introduction
6. Generalized linear model
7. Bayesian models
8. Classification
9. Unsupervised learning, dimension reduction
10. Clustering and mixture models
11. Model design and validation

(topics might change)

How about data?

Wigner, Hamming inspired thinking:

- Norvig et al: “The Unreasonable Effectiveness of Data” (2009)
- Karpathy: “The Unreasonable Effectiveness of Recurrent Neural Networks” (2015)
- Zhang et.al: “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric” (2018)
- Parisi et al: “The (Un)Surprising Effectiveness of Pre-Trained Vision Models for Control” (2022)

Data

“Computers have promised us a fountain of wisdom but delivered a flood of data.”
“The amount of information in the world doubles every 20 months.”
(Frawley, Piatetsky-Shapiro, Matheus, 1991)

A proper dataset is more than a list of data points. We should have information about

- details about the data collection mechanism
- how reliable are the data points? Noise? Outliers?
- any missing values? Why?

Examples:

- ImageNet - test dataset for image classification, originally: 1.2M images, 1k categories
  is 91.1% (top1) accuracy enough?

- ISIC-2019 challenge dataset - images of skin lesions, 25k images, 8 categories (eg. melanoma or nevus)
  any “almost”-duplicates?

Data

Possible uses of data:

- **describe** - understand available data
  eg.: anomaly detection, eg. fraud prevention based on customer transaction data

- **predict** - “best guess” of something not (yet) available
  eg.: forecast gas / hydro power plant energy demand (forecast consumption, renewable generation)
  eg.: classification / object detection / segmentation using supervised training models

- **prescribe** - suggest best action
  eg.: autonomous agents (self-driving car, robot)
  eg.: queries for document content (search engines)
  recommender systems (Amazon, Netflix),
  content-sensitive ads (Google AdSense)

Taster

- how to fit a wide class of functions on data (explicitly; not just a line)
  what is overfitting / underfitting, how to reduce these?
  (general linear model)

- covid test: patient documentation say: sensitivity 93%, specificity 99%
  - what do these number say? which one is more relevant?
  (binary classification)

- melanoma thickness prediction on small datasets (own research)
  our benchmark: 71% (balanced accuracy)
  better than everyone else... except one paper: 85%
  we have shown they made a mistake - data leakage - real value would have been 68%
  (model design and validation)

“Machine learning is just glorified statistics.”

“When you’re fundraising, it’s AI.
When you’re hiring, it’s ML.
When you’re implementing, it’s logistic regression.” - anon