Machine learning for health

-

Introduction

Xavier Tannier

2023

Who are we?

- Xavier Tannier
  - Professor in Computer Science at Sorbonne Université
  - Research at LIMICS
  - Natural Language Processing, Information extraction, machine learning… applied to health data

- Manon Chossegros

  Computer scientist, PhD student at Sorbonne Université, LIMICS
- Aniss Acherar

  Computer scientist, PhD student at Sorbonne Université, iPLesp
- Christel Gérardin

  Medical doctor, PhD student at Sorbonne Université, iPLesp

Who are you?

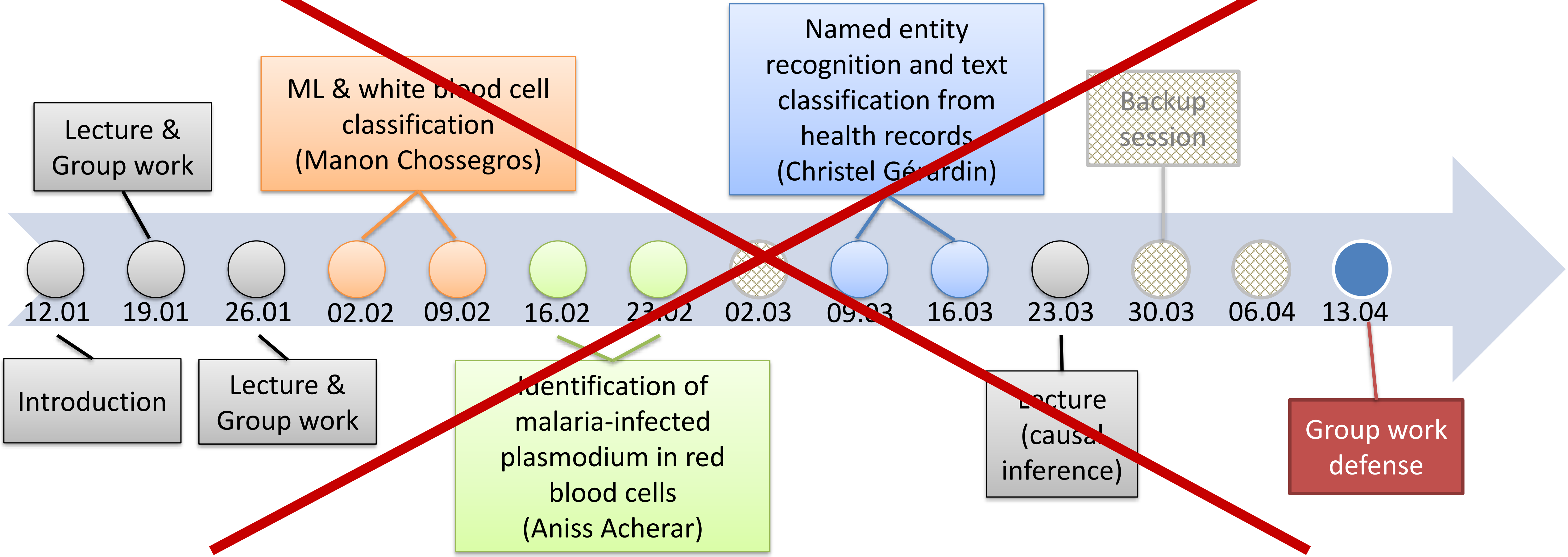
- Stats / M2A
- Why are you here?

- How familiar are you with
  - Regular statistics for health data?
  - Machine learning libraries?
  - Deep learning?
  - NLP?
  - Image processing?
  - Causality?

Organization & Grading

Organization

- 10 x 3-hour sessions



12.01

19.01

26.01

02.02

09.02

16.02

23.02

02.03

09.03

16.03

23.03

30.03

06.04

13.04

Introduction

Lecture & Group work

Lecture & Group work

ML & white blood cell classification (Manon Chossegros)

Identification of malaria-infected plasmodium in red blood cells (Aniss Acherar)

Named entity recognition and text classification from health records (Christel Gérardin)

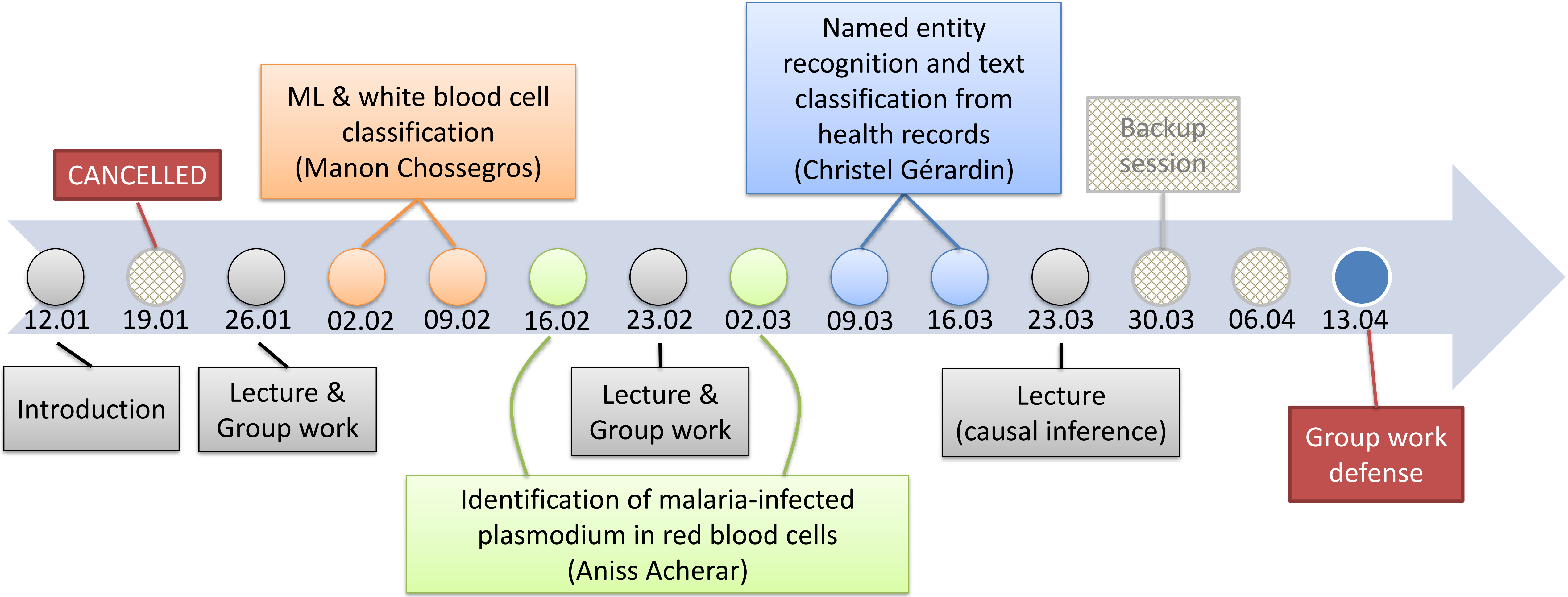
Lecture (causal inference)

Backup session

Group work defense

Re-Organization

- 10 x 3-hour sessions


CANCELLED
ML & white blood cell classification (Manon Chossegros)
Named entity recognition and text classification from health records (Christel Gérardin)
Backup session
12.01
19.01
26.01
02.02
09.02
16.02
23.02
02.03
09.03
16.03
23.03
30.03
06.04
13.04
Introduction
Lecture & Group work
Lecture & Group work
Lecture (causal inference)
Group work defense
Identification of malaria-infected plasmodium in red blood cells (Aniss Acherar)

Evaluation

- **35% -** graded "TME" (homework, deadline April 7th, upload code + report to moodle)
- **65% -** group work (report + defense April 14th)

Graded “TME”

- Choose one of the three topics presented during TMEs
- Explore it further, suggest one or several improvements
- Upload to *moodle*:
  - The **commented code**
  - **Slides or report** detailing your work
  - (No defense)

Group work

- Choose a topic from the list (next slide)
- **Only one group per topic**!
  Please organize yourselves, unless you prefer that I choose arbitrarily.
- Prepare a **15-minute talk** about this topic, as well as a **report (between 5 and 10 pages)**

⚠ The suggested link is just to help you get started,
not the only source to consult!

Topics for group work

"Machine learning, healthcare and …

Explainability & case-based reasoning

"The false hope of current approaches to explainable artificial intelligence in health care",
Lancet Digital Health, 2022

Missing data imputation

"Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty",
Artificial Intelligence in Medicine, 2022

Very high-dimensional data (esp. omics)

"Deep Learning in Omics Data Analysis and Precision Medicine",
Computational Biology, 2019

AutoML

"Automated machine learning: Review of the state-of-the-art and opportunities for healthcare",
Artificial Intelligence in Medicine, 2020

Reproducibility

"Reproducibility in Machine Learning for Health",
Science Translational Medicine, 2019

Transparency

"Transparency: Motivations and Challenges",
ICML, 2017

Ethics

"Ethics and governance of artificial intelligence for health",
WHO, 2021

"Diagnostic Médical et Intelligence Artificielle : Enjeux Ethiques",
Avis 141 du CCNE, nov. 2022

Environment

"L'impact environnemental du numérique en Santé",
Ministère des solidarités et de la santé, 2021

Equality of care & Digital gap

Didn't like my list of topics ? Propose your own.

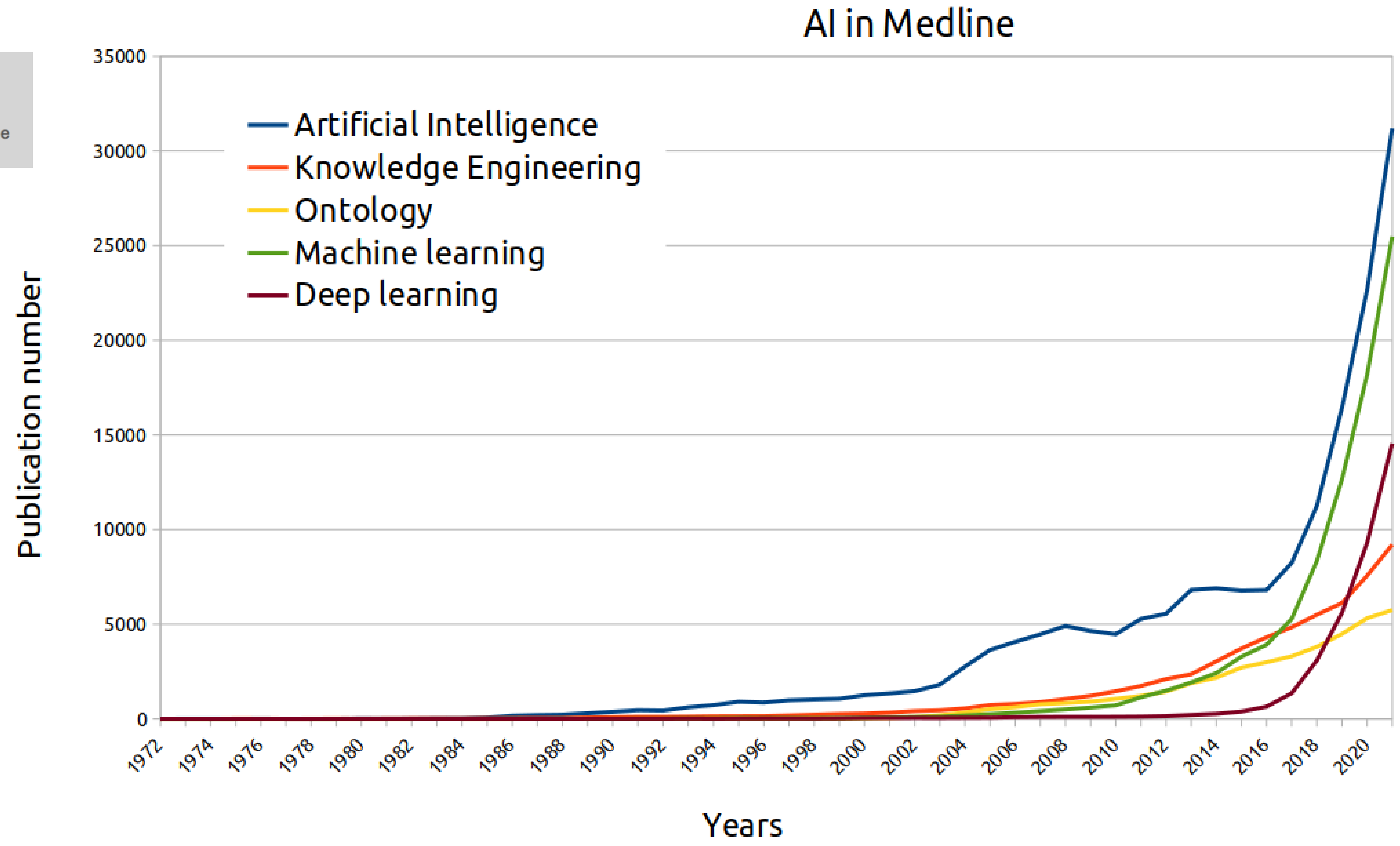
But ask me first, I might reject your idea.

"

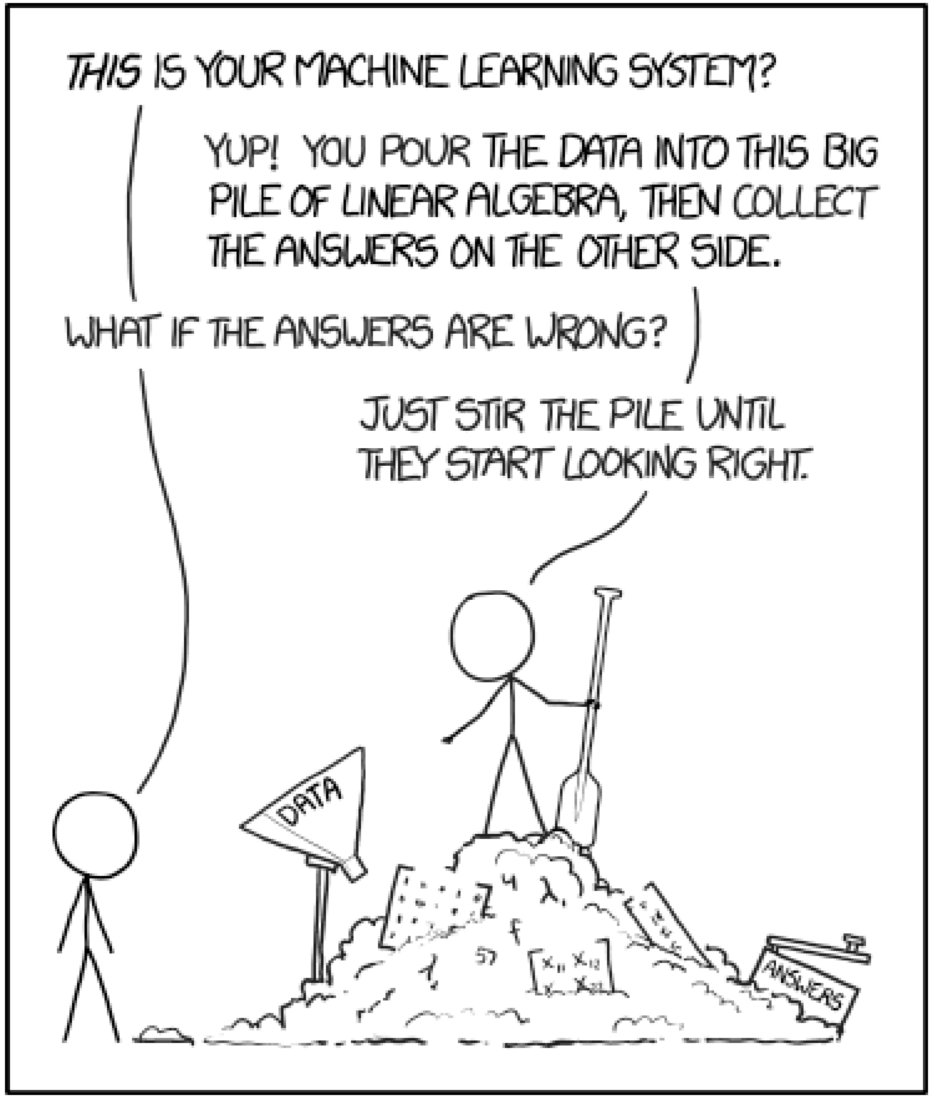
What is it about?

Machine learning is hype in health-related research

PubMed.gov
US National Library of Medicine
National Institutes of Health

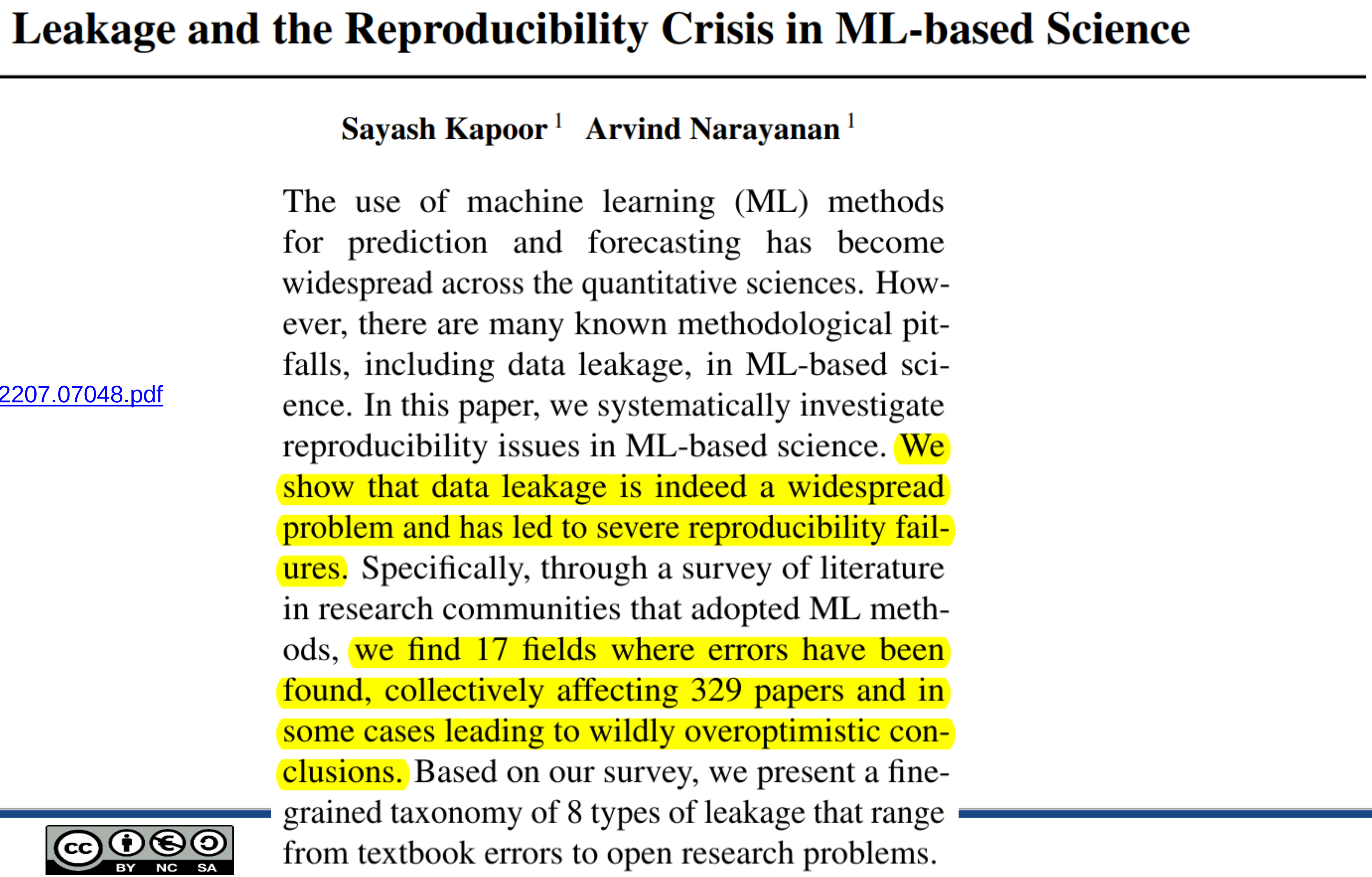


With real-life applications, we can't do anything


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838/

Yet, we do



Leakage and the Reproducibility Crisis in ML-based Science

Sayash Kapoor[1] Arvind Narayanan[1]

The use of machine learning (ML) methods for prediction and forecasting has become widespread across the quantitative sciences. However, there are many known methodological pitfalls, including data leakage, in ML-based science. In this paper, we systematically investigate reproducibility issues in ML-based science. We show that data leakage is indeed a widespread problem and has led to severe reproducibility failures. Specifically, through a survey of literature in research communities that adopted ML methods, we find 17 fields where errors have been found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. Based on our survey, we present a fine-grained taxonomy of 8 types of leakage that range from textbook errors to open research problems.

Jul.2022

https://arxiv.org/pdf/2207.07048.pdf

Yet, we do

Leakage and the Reproducibility Crisis in ML-based Science

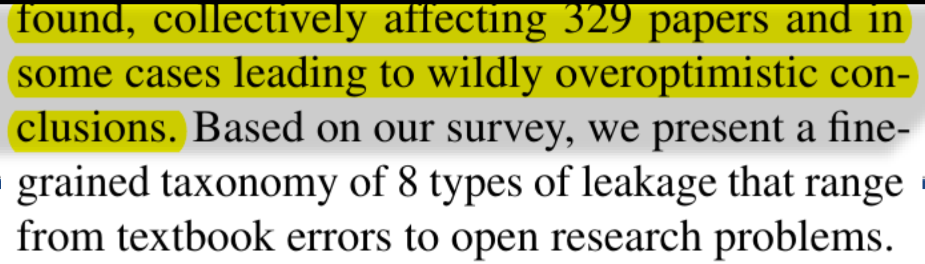
Sayash Kapoor [1] Arvind Narayanan [1]

The use of machine learning (ML) methods

Jul.2022

https://arxiv.org/pdf/2207.07048.pdf

We need people who know both machine learning and its best practices, and the challenges of medical data.



found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. Based on our survey, we present a fine-grained taxonomy of 8 types of leakage that range from textbook errors to open research problems.