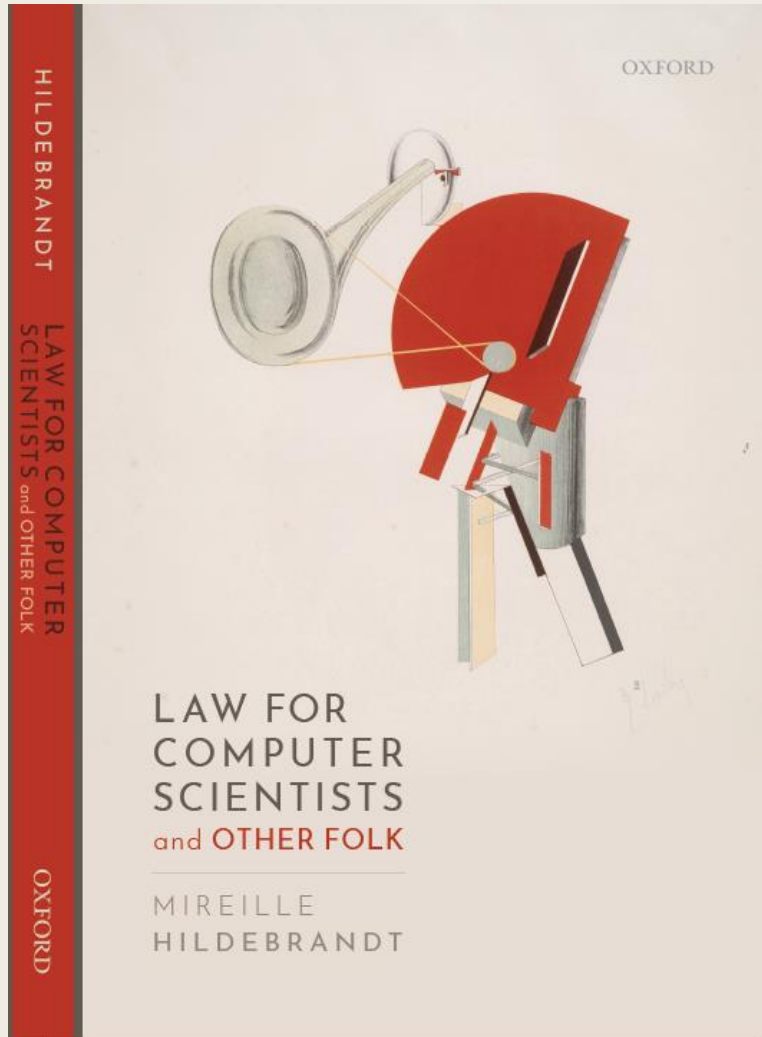




GROUND-TRUTHING IN THE EUROPEAN HEALTH DATA SPACE

Mireille Hildebrandt, FBA
PI **COHUBICOL** ERC ADG project



- My background: law, philosophy of technology
- Chair at Computer Science Department @Radboud University
- My research focus: implications of 'AI' for law and the rule of law

- Implications of 'AI' for law and the rule of law
 - Privacy, fairness – the usual suspects
 - **More important:**
 - **4R AI (robust, resilient, reliable, responsible)**
 - Involving methodological integrity and key questions such as:
 - how does design and use of AI shift power relationships?
 - the political economy of AI systems
 - relationship between doctor, patient, hospital, insurance, government

In the context of the ERC ADG I am:

- Investigating **claims made on behalf of** AI systems
- Investigating **the substantiation of such claims**
 - Mathematical verification, empirical validation, certification
 - Impact on the domain: gaps between requirements and specifications
 - Real-world impact (gap between specification and real-world goal)
- As to **training data** that stand for a ground truth:
 - 'ground truth' concerns real world issues:
it **cannot** be completely and finally computed/formalised
 - meaning that it can be computed/formalised **in different ways**

PROJECT PUBLICATIONS

Home

Get in touch

VOCABULARIES

WORKING PAPERS

TYPOLOGY OF LEGAL TECH

The Typology


How to use

FAQs & methodology

Typology of Legal Technologies

A Method – A Mindset

The Typology is a curated set of legal technologies (applications, scientific papers, and datasets) that we handpicked to demonstrate the potential impact on *legal effect* of different types of 'legal tech'. To understand how and why we created this, see the [FAQs & methodology](#) page.

- **Use the filters below** to find legal techs you are interested in. Click a system to view its full profile.
- **Compare systems** by clicking  on one or more systems (view the comparison at the bottom of this page).

SHOWING 30 TECHS

 RESET FILTERS

END-USERS	FUNCTIONALITY	CODE/DATA-DRIVEN	TYPE OF SYSTEM	
Any	Any	Either	<input checked="" type="radio"/> Any	<input type="radio"/> App
<input type="radio"/> Legislation	<input type="radio"/> Search	<input type="checkbox"/>	<input type="radio"/> Dataset	<input type="radio"/> Paper
Akoma Ntoso	Automatic Catchphrase Identification from Legal Court Case Documents (Mandal et al. 2017)	Blawx		
<input type="radio"/> Legislation	<input type="radio"/> Litigation	<input type="radio"/> Legislation		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		
Casetext	Catala	Chinese AI and Law dataset (CAIL2018)		
<input type="radio"/> Litigation	<input type="radio"/> ADM	<input type="radio"/> Litigation		
<input type="checkbox"/>	<input type="radio"/> Legislation	<input type="checkbox"/>		

PROJECT PUBLICATIONS

[Home](#)

[Get in touch](#)

VOCABULARIES ▾

WORKING PAPERS ▾

TYPOLOGY OF LEGAL TECH ▴

[The Typology](#)

[How to use](#)

[FAQs & methodology](#)

[Typology of Legal Tech](#) /

Chinese AI and Law dataset (CAIL2018)

Litigation: prediction of judgment

github.com/thunlp/CAIL/blob/master/README_en.md ↗

Main research: March 2022

CONTENTS

- [What does it claim to do?](#)
- [Substantiation of claims & potential issues](#)
- [Is it currently in use?](#)
- [The creators](#)
- [Jurisdiction](#)
- [License](#)

▾ What does it claim to do?

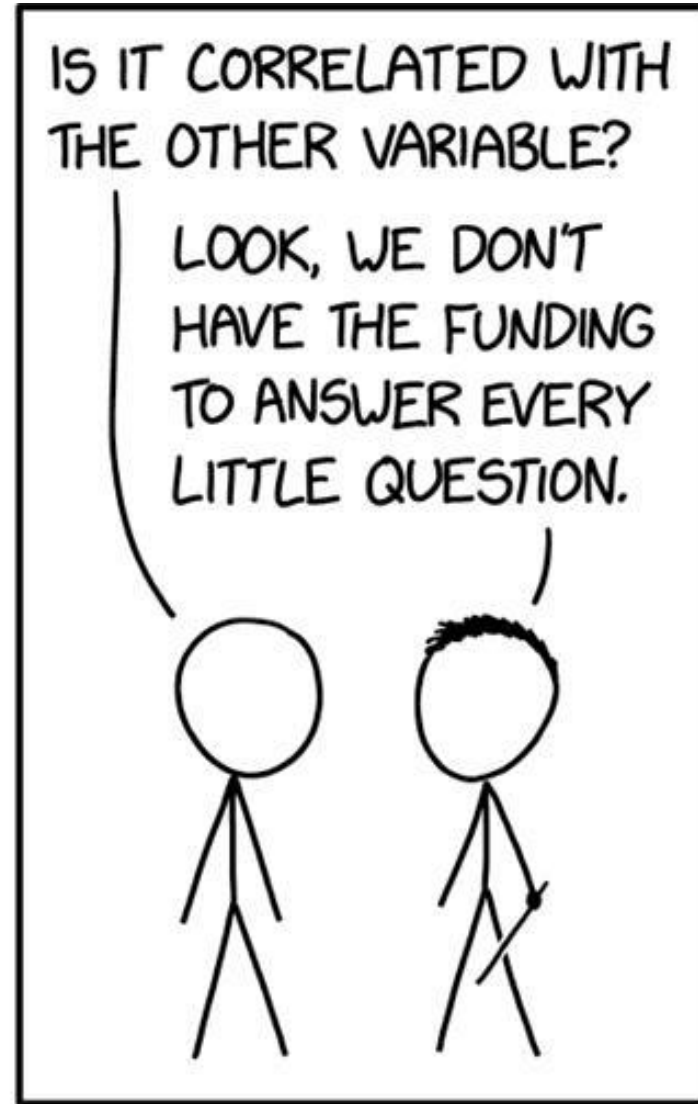
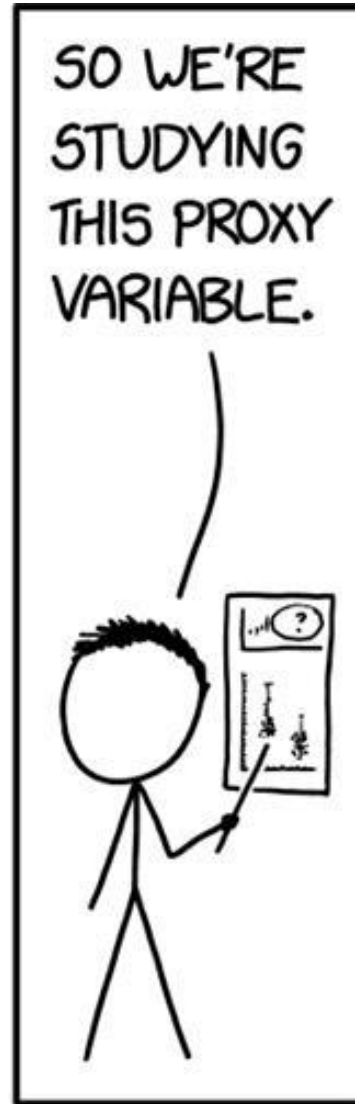
CAIL2018 is the Chinese AI and Law challenge dataset. It was created for the purposes of encouraging research into how machine learning can assist in the process of Legal Judgment Prediction (LJP). For the authors, LJP is about enabling machines to predict the outcome of legal cases by reference to the descriptions of fact set out in those cases. The dataset was released in 2018 as part of the CAIL2018 competition. The competition, which attracted more than 200 participants, focussed on how natural language processing improves performance in LJP tasks. It presented competitors with three subtasks. These were the (1) prediction of applicable law articles, (2) charges, and (3) prison terms by reference to the descriptions of facts for the cases forming part of the training data of the CAIL2018 dataset.

▾ AT A GLANCE ⓘ

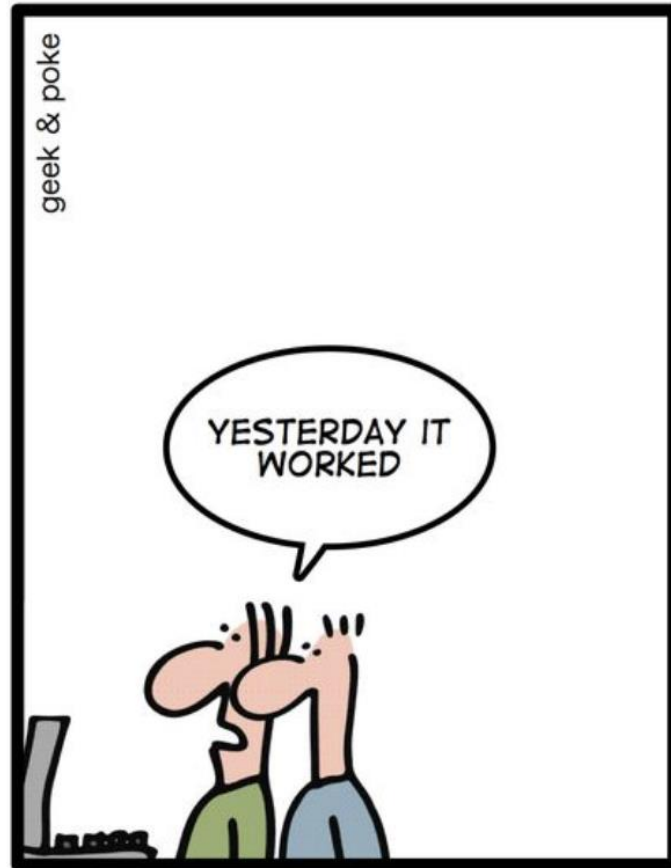
Intended users	<ul style="list-style-type: none"> ▪ Academics ▪ Software developers
Code- or data-driven	Data-driven
Form	Dataset (off-the-shelf)
Automation or support	<ul style="list-style-type: none"> ▪ Legal decision support ▪ Legal research strategy ▪ Legal strategy support
In use?	Unknown
Creators	Academics Details ⓘ
Access	<ul style="list-style-type: none"> ▪ Free download/web application

See our [methodology](#) for field definitions.

- Medical and health applications:
 - Diagnosis, treatment, research
 - Crucial role of **training & validation** data
 - Crucial role of **test** data
 - Assumptions of what counts as 'ground truth'
 - Implications of those assumptions



WHEN YOU HEAR THIS:



*YOU KNOW YOU'RE IN A
SOFTWARE PROJECT*

- Software, including what some like to call AI, is always running behind.
- Medical expert systems are stuck with the moment they were finalised
 - Medical ML can only be trained on past data
 - **Prediction is difficult, especially when it's about the future**

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji

Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender

Department of Linguistics
University of Washington

Amandalynne Paullada

Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

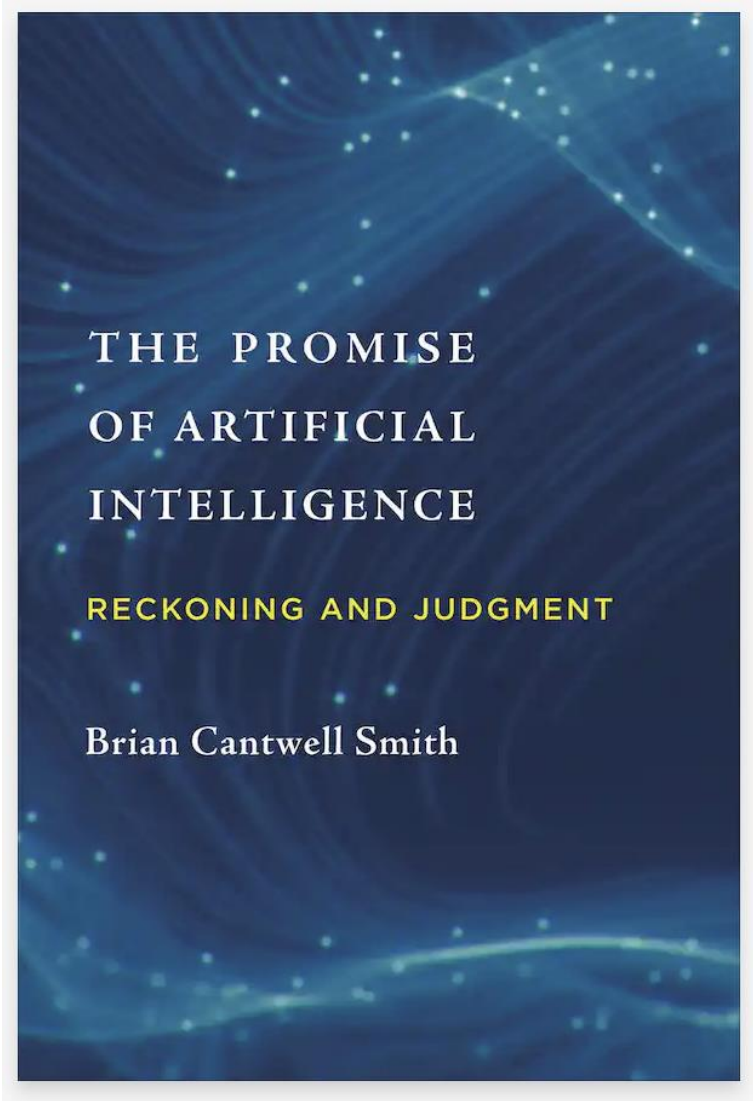
Abstract

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally “general” broad measures of progress they are set up to be.

- 35th Conference on Neural Information Processing Systems (NeurIPS 2021)
Track Datasets and Benchmarks.

1 Introduction

In the 1974 Sesame Street children's storybook *Grover and the Everything in the Whole Wide World Museum* [Stiles and Wilcox, 1974], the Muppet monster Grover visits a museum claiming to showcase "everything in the whole wide world". Example objects representing certain categories fill each room. Several categories are arbitrary and subjective, including showrooms for "Things You Find On a Wall" and "The Things that Can Tickle You Room". Some are oddly specific, such as "The Carrot Room", while others unhelpfully vague like "The Tall Hall". When he thinks that he has seen all that is there, Grover comes to a door that is labeled "Everything Else". He opens the door, only to find himself in the outside world.



What's next?

- European Health Data Space (EHDS)
- The constructive and/or approximate nature of 'the' ground truth
- Unsupervised , RI and iML
- Why ground truthing matters and how the EHDS fits in

What's next?

- **European Health Data Space (EHDS)**
- The constructive and/or approximate nature of 'the' ground truth
- Unsupervised , RI and iML
- Why ground truthing matters and how the EHDS fits in

European Health Data Space

- EU ambitions:
 - Creating EU data spaces in various domains
 - Starting with health data
 - *Repurposing data* for research & competitive innovation

European Health Data Space

Re-use of medical data:

- Van der Lei's (1991) first **Law of Medical Informatics**:
 - “don't use data for a purpose other than that for which it was collected”
 - examples:
 - data entry with purpose of obtaining insurance coverage or permission for testing
 - incentive structure that drives recording of data differs between MS
 - implications:
 - data are incomplete, incorrect and not interoperable
 - standardisation will be applied, potentially making them even more incorrect

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space'

Aims:

- to establish 'rules, common standards and practices, infrastructures and a governance framework for the primary and secondary use of electronic health data' (art. 1).
- Including the establishment of '**a mandatory cross-border infrastructure for the secondary use of electronic health data**' (art. 2(e)).

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space'

Defining:

- 'data quality' as 'the degree to which characteristics of electronic health data are suitable for secondary use' (art. 2(ad)),
- 'data quality and utility label' as 'a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset' (art. 2(ae))

European Health Data Space

■ 2020 proposed 'Regulation on the European Health Data Space', art. 33.1:

'Data holders shall make the following categories of electronic data available

for secondary use in accordance with the provisions of this Chapter:

- a. EHRs [electronic health record systems];
- b. data impacting on health, including social, environmental behavioural determinants of health;
- c. relevant pathogen genomic data, impacting on human health;
- d. healthrelated administrative data, including claims and reimbursement data;
- e. human genetic, genomic and proteomic data;
- f. person generated electronic health data, including medical devices, wellness applications or other digital health applications;

and many more.

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 33.3:

The electronic health data referred to in paragraph 1 shall cover **data processed for**

- the provision of health or care or
- for public health, research, innovation, policy making, official statistics, patient safety
- or regulatory purposes,

collected by

- entities and bodies in the health or care sectors,
- including public and private providers of health or care,
- entities or bodies performing research in relation to these sectors, and
- Union institutions, bodies, offices and agencies.'

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 34:

The purposes for which secondary use is permitted are **limited**,
their articulation is **very broad**, e.g. including:

- scientific research related to health or care sectors;
- development and innovation activities for products or services contributing to
 - public health or social security, training, testing and evaluating of algorithms
- and for providing personalised healthcare.

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 35:

Some purposes are **explicitly prohibited**, for instance

- taking decisions that are detrimental for a natural person,
- decisions that exclude certain groups from insurance or
- marketing to health professionals

It is, however, unclear how this could be **monitored and enforced**, knowing that the enforcement of purpose limitation in the context of the GDPR has been notoriously difficult

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 36-43:

The **governance of the EHDS** is attributed to **Health Data Access Bodies** that can issue

- **data permits** to access data to potential **data users**,
- provided a number of procedural and material conditions are fulfilled
 - including purpose limitation.

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 52:
- a 'crossborder infrastructure for secondary use of electronic health data' is set up
- by designated contact points in the MSs.

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 52:

Datasets available for cross-border access must contain **a metadata catalogue** that describes e.g.

- 'the source, the scope, the main characteristics, nature of electronic health data and
- **conditions for making** electronic health data **available**'.

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 57.1:

The European Commission will set up

- 'an **EU Datasets Catalogue**
- connecting the national catalogues of datasets established by
- the health data access bodies and other authorised participants'

European Health Data Space

- 2020 proposed 'Regulation on the European Health Data Space', art. 56:

Data made available through the health data access bodies **may** have

- a 'data quality and utility label',
- which is compulsory when processed
- 'with the support of Union or national public funding'.

European Health Data Space

The label must comply with the following elements (art. 56.3):

- (a) for **data documentation**: metadata, support documentation, data model, data dictionary, standards used, provenance;
- (b) **technical quality**, showing the completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;
- (c) for **data quality management processes**: level of maturity of the data quality management processes, including review and audit processes, biases examination;

European Health Data Space

- (d) **coverage**: representation of multi-disciplinary electronic health data, representativity of population sampled, average timeframe in which a natural person appears in a dataset;
- (e) **information on access and provision**: time between the collection of the electronic health data and their addition to the dataset, time to provide electronic health data following electronic health data access application approval;
- (f) **information on data enrichments**: merging and adding data to an existing dataset, including links with other datasets;

What's next?

- European Health Data Space (EHDS)
- The constructive and/or approximate nature of 'the' ground truth
- Unsupervised, RI and iML
- Why ground truthing matters and how the EHDS fits in

The constructive / approximate nature of 'the' ground truth

- Ground truth *condition sine qua non* for supervised ML
 - A proxy for **a state of affairs** in the real world or
 - A proxy for **a desirable state of affairs** in the real world
 - Labelled training data serve as ground truth
 - Labelling by medical experts, e.g. radiologists

The constructive / approximate nature of 'the' ground truth

- Ground truth is a *conditio sine qua non* for supervised ML
 - If the proxy is wrong or incomplete
 - the algorithm that is trained on the data will learn *incorrectly*
 - Even if the accuracy on the **validation data** is very high
 - We may expect accuracy **on test data** to be low then
 - In the case of medical science that may cause harm or death

The constructive / approximate nature of 'the' ground truth

- I. Ground truth *as a proxy* is a construction, meant to enable the system to learn
- II. Ground truth is *what the proxy stands for*, what it aims to approximate

The constructive / approximate nature of 'the' ground truth

- Ground truth is the result of hard work:
 - Collecting, cleansing and labelling training data is not obvious
 - It can be done in different ways, with **different trade-offs**
 - In terms of availability (low hanging fruit?)
 - In terms of noise
 - In terms of costs (low hanging fruit?)
 - In terms of complexity

The constructive / approximate nature of 'the' ground truth

- To remind us of the deliberate effort of 'preparing' the ground truth
- We should speak of 'ground truthing'
- Which also reminds us of its dynamic nature

The constructive / approximate nature of 'the' ground truth

- In the end the question is **whether the distribution of the training data**
- Is sufficiently ***similar to that of the test (= future) data***
 - We cannot ***train on*** future data
 - We cannot ***model*** future data, only past or present data

The constructive / approximate nature of 'the' ground truth

■ Objectivism:

- claims to truth that
- hide relevant assumptions and
- resist contestation

■ Objectivity:

- a well argued
- cross-disciplinary
- contestable
- construction of a ground truth

The constructive / approximate nature of 'the' ground truth

- The difference between **objectivity and objectivism**:
 - Connects with Popper's **falsification requirement** (philosophy of science)
 - Translates the idea of **deliberative & participatory** democratic practices
 - Informs the core of the rule of law: **contestability**

The constructive / approximate nature of 'the' ground truth

Peircing the veil of objectivism (Cabitza et al, 2020)

- more granular metrics to account for the choice of a particular ground truth
 - **how true** (distance between proxy and real world target)
 - **how reliable** (new metric for the degree of concordance of those who label)
 - **how informative** (new metric for correspondence between sample and population)

What's next?

- European Health Data Space (EHDS)
- The constructive and/or approximate nature of 'the' ground truth
- **Unsupervised, RI and iML**
- Reinforcement and interactive ML
- Why ground truthing matters and how the EHDS fits in

Unsupervised, RL and iML

Deep learning (ANN) used for unsupervised learning

- Where's the ground truth here?
- Can *the constructed nature* of what a system trains on in *supervised learning* be avoided by allowing the system to learn what is relevant instead of providing it with labels/examples/instructions?

Unsupervised, RL and iML

- Does the success of AlphaZero in complex games (Chess, Go) prove that the system is now learning directly from 'reality'?
- Does that imply a kind of UR-Ground-Truth that is available for training?
 - Without mediation?
 - Or is the training data a proxy here (deployed as ground truth)?

Unsupervised, RL and iML

Data is always, by definition, a proxy, it is never what it stands for

A trace (e.g. sensor data)

A representation (e.g. health records)

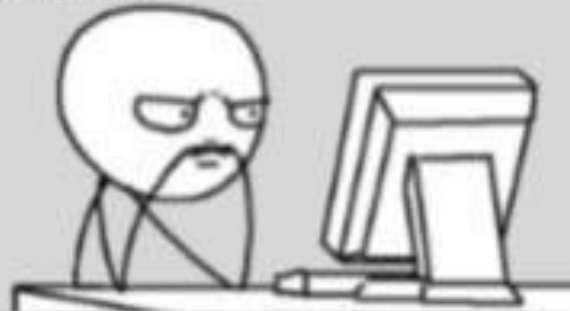
A translation (records, sensor data)

Unsupervised, RL and iML

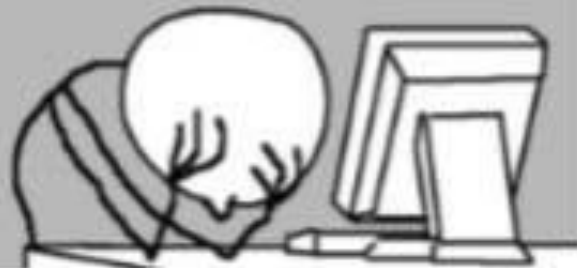
- **Large language models:**
 - **Generative Pre-Trained**
 - **N=ALL?**
- **Stochastic Parrots**
 - **Depends on what the parrot has been trained on**
 - **OPENAI is not very open about that**

Days before OpenAI

Developer coding
- 2 hours



Developer debugging
- 6 hours

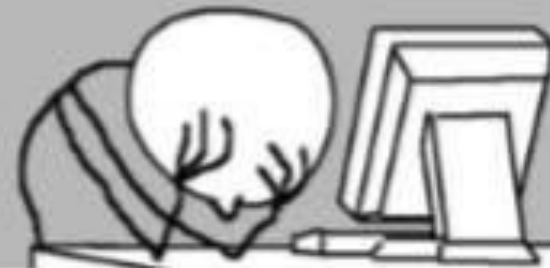


Days after OpenAI

ChatGPT generates
Codes - 5 min



Developer debugging
- 24 hours



Unsupervised, RL and iML

Ground truth is *hidden* in three 'design' decisions:

1. **what data** to use as a proxy for whatever it is one wants to achieve
2. how to **curate** the data (remove noise, structure, add, integrate and interoperationalise data)
3. **what 'learner'** to develop for training on the data

Unsupervised, RL and iML

AlphaFold:

- a method to reliably predict a protein's structure
- just from its sequence of amino acids

Claim:

- the ability to predict a protein's shape computationally from its genetic code alone
- as a **complementary alternative** to costly and time-consuming experimentation
- could help dramatically accelerate research

Unsupervised, RL and iML

AlphaFold gets its data from EMBL-EBI, a not-for-profit international institute that:

- is part of the European Molecular Biology Laboratory (EMBL)
- provides the infrastructure needed to share data openly and fairly in the life sciences
- curates the AlphaFold dataset in a way that allows
 - linking with other biological datasets
 - such as the Protein Data Bank Europe UniProt

Unsupervised, RL and iML

AlphaFold articulates its objective in terms of a Grand Challenge, we should therefore ask:

1. What problem does it actually solve?
 - Developing hypotheses of protein structure
2. What problem(s) does it not solve?
 - The need to test, experimentally, whether the hypothesis is correct
3. What problem(s) could it create?
 - Dual use

Unsupervised, RL and iML

AlphaFold deploys:

- a **transformer model** and an attention architecture
 - calculating dependencies between distant sequential data
- using **multiple sequence alignment statistics**,
 - detecting 'context' (complex interactions between sequence and environments)
- having moved from AlphaFold1 to AlphaFold2 with a leap in accuracy
 - dependencies can be modelled after **pre-training** and fine-tuned on smaller specific data

Unsupervised, RL and iML

As Marcu et al (2022) state:

- One limitation of approaches based on MSAs, such as AlphaFold2, is
 - that they are constrained by our current knowledge and data sets
1. Does adding molecular dynamics solve this issue?
 - remains restricted to known or abducted dynamics
 2. Does adding more and/or other data solve this issue?
 - remains restricted to available data

Unsupervised, RL and iML

The construction of the data and the models constitute (an approximation of) the ground truth:

- Testing on real world proteins will always be necessary
- Which may be far removed from the virtual laboratories of protein fold mappings

This in no way deflects from the incredible achievements, but reminds us of

- Potential real world implications
- Acknowledge that the data and the models are not what they stand for
- Need to always engage in real world testing

Unsupervised, RL and iML

reinforcement learning (RL) concerns

- systems that are built to **interactively learn from the environment they navigate**

however, we should note that

- the 'environment' is mostly mediated
 - formalisation in terms of states, actions and rewards
- if so, all caveats apply about **data not being what it stands for**

Unsupervised, RL and iML

Interactive machine learning (Holzinger 2016):

- algorithms that can interact with both
 - computational agents and
 - human agents [mh: 'oracles', 'prompt engineering'] and
- can optimize their learning behavior through these interactions

Note the difference with stochastic parrots, this explains why we are invited to improve ChatGPT with our 'prompt engineering activities'

How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, R. Andrew Taylor, David Chartash

doi: <https://doi.org/10.1101/2022.12.23.22283901>

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.



Abstract

Full Text

Info/History

Metrics

Preview PDF

ABSTRACT

Background ChatGPT is a 175 billion parameter natural language processing model which can generate conversation style responses to user input.

Objective To evaluate the performance of ChatGPT on questions within the scope of United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams, as well as analyze responses for user interpretability.

Reinforcement and interactive machine learning

‘it could be an interesting educational and knowledge assessment tool’

What's next?

- European Health Data Space (EHDS)
- The constructive and/or approximate nature of 'the' ground truth
- Supervised ML
- Unsupervised, RI and iML
- Why ground truthing matters and how the EHDS fits in

Why ground truthing matters

- Clearly, the key role of human domain expertise in iML testifies to the need to mitigate risks inherent in ground truthing,
- especially (though not only) in the case of unsupervised and reinforcement learning.
- This confirms potential problems with the interoperability of medical data across different jurisdictions and healthcare systems,
- due to the fact that data have often been collected and stored based on different purposes which render aggregation in a shared data space hazardous at least.
- This problem can be solved by standardisation, but then we will lose lots of information that 'sits' in different templates, configurations etc.

Where the EHDS fits in

The label must comply with the following element (art. 56.3):

- (a) for **data documentation**: metadata, support documentation, data model, data dictionary, standards used, provenance;
- (b) **technical quality**, showing the completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;
- (c) for **data quality management processes**: level of maturity of the data quality management processes, including review and audit processes, biases examination;

Where the EHDS fits in

- (d) **coverage**: representation of multi-disciplinary electronic health data, representativity of population sampled, average timeframe in which a natural person appears in a dataset;
- (e) **information on access and provision**: time between the collection of the electronic health data and their addition to the dataset, time to provide electronic health data following electronic health data access application approval;
- (f) **information on data enrichments**: merging and adding data to an existing dataset, including links with other datasets;

Where the EHDS fits in

- Peek & Pereira Rodriguez' (2018) caveats about:
 - Repurposing treatment data for medical research
 - Replacing clinical trials with big data research

Where the EHDS fits in

- Medical Data Science should:
 - Develop typologies of medical technologies
 - To **map and compare** their status, method, prevalence, availability (proprietary)
 - To **assess claims** made on behalf of them
 - In **terms of their substantiation** (actual and potential)
 - Explain to the EU legislature what **assumptions that inform the EHDS**
 - Are simply incorrect
 - May need qualification
 - Are largely ok



m.e.menair