How I failed machine learning in medical imaging

Responsible AI seminar – DTU

28th May 2021

Dr. Veronika Cheplygina

Joint work with Gaël Varoquaux

@drveronikach

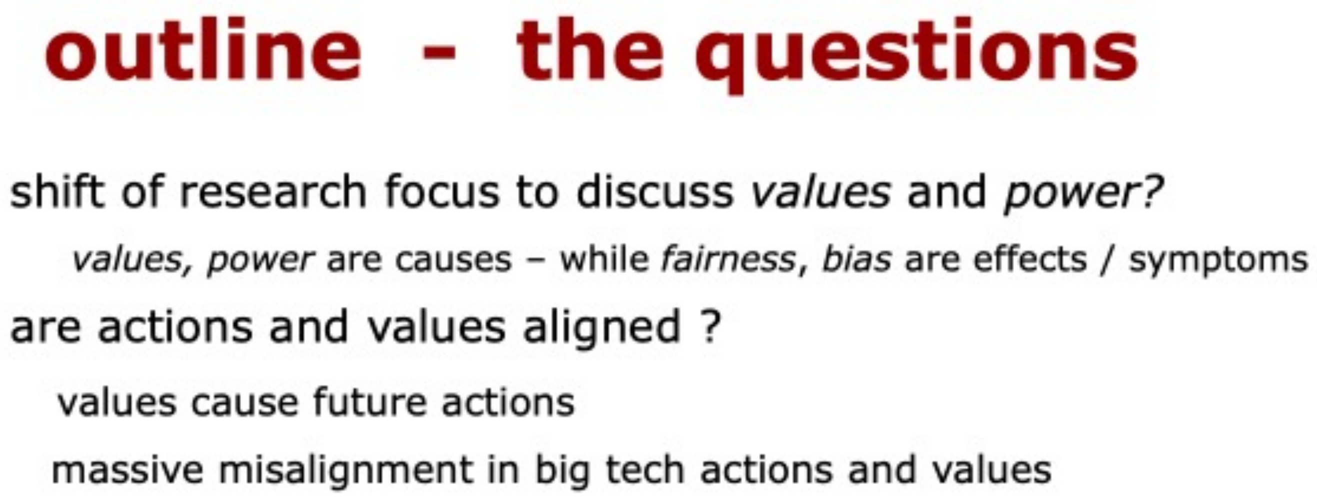
https://www.veronikach.com

Values & AI

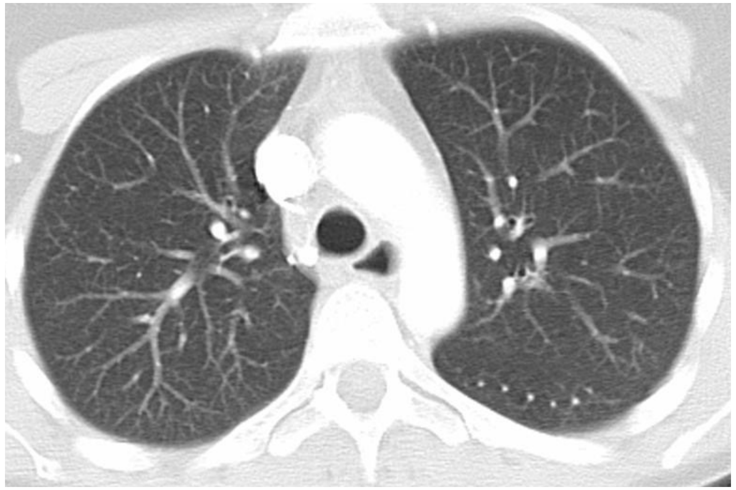
Thanks to Aasa, Melanie and Sune for organizing!

And Lars Kai Hansen for insightful start of seminar

Zooming into ML, medical imaging (diagnosis/segmentation) + my own perspective



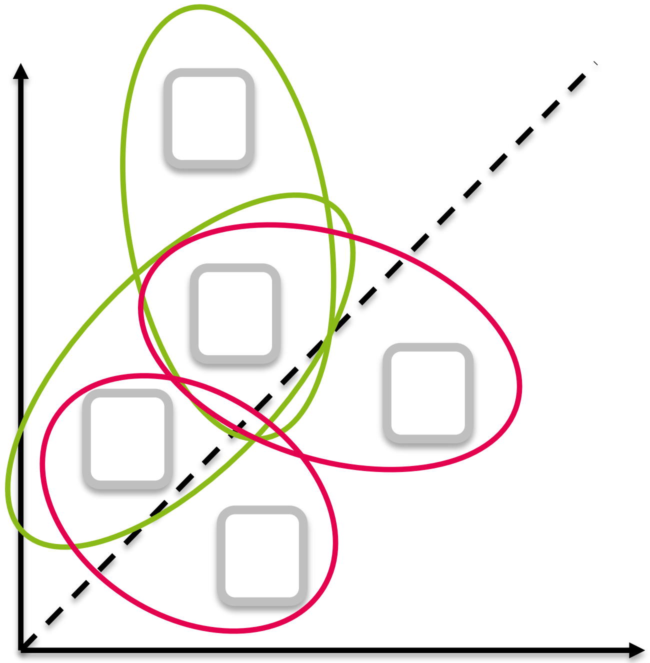
From Lars Kai Hansen’s talk



My perspective

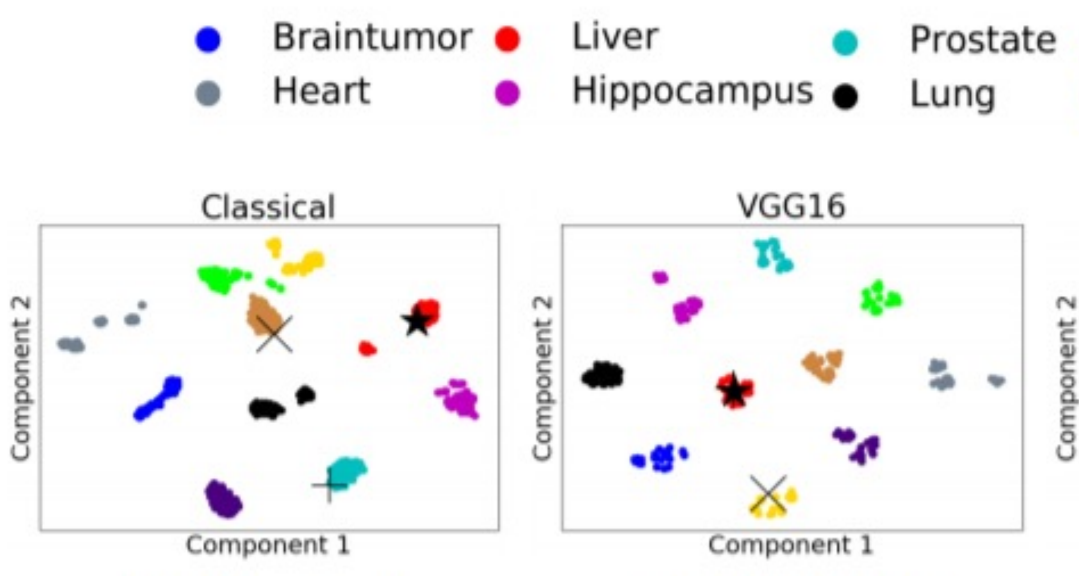
2010

- Pattern recognition
- Similarities of “bags” (multiple instance learning) / graphs



My perspective

- Similarities of datasets
  - Transfer learning
  - Meta-learning



(a) Classical (b) VGG16

(Work with Tom van Sonsbeek, Irma van den Brandt)

Cats or CAT scans: Transfer learning from natural or medical image source data sets?
V Cheplygina
Current Opinion in Biomedical Engineering 9, 21-27

My perspective

- Similarities of methods

Multiple instance learning: A survey of problem characteristics and applications
MA Carbonneau, V Cheplygina, E Granger, G Gagnon
Pattern Recognition 77, 329-353

Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis
V Cheplygina, M de Bruijne, JPW Pluim
Medical image analysis 54, 280-296

A survey of crowdsourcing in medical image analysis
SN Ørting, A Doyle, A van Hilten, M Hirth, O Inel, CR Madan, P Mavridis, ...
Human Computation 7, 1-26

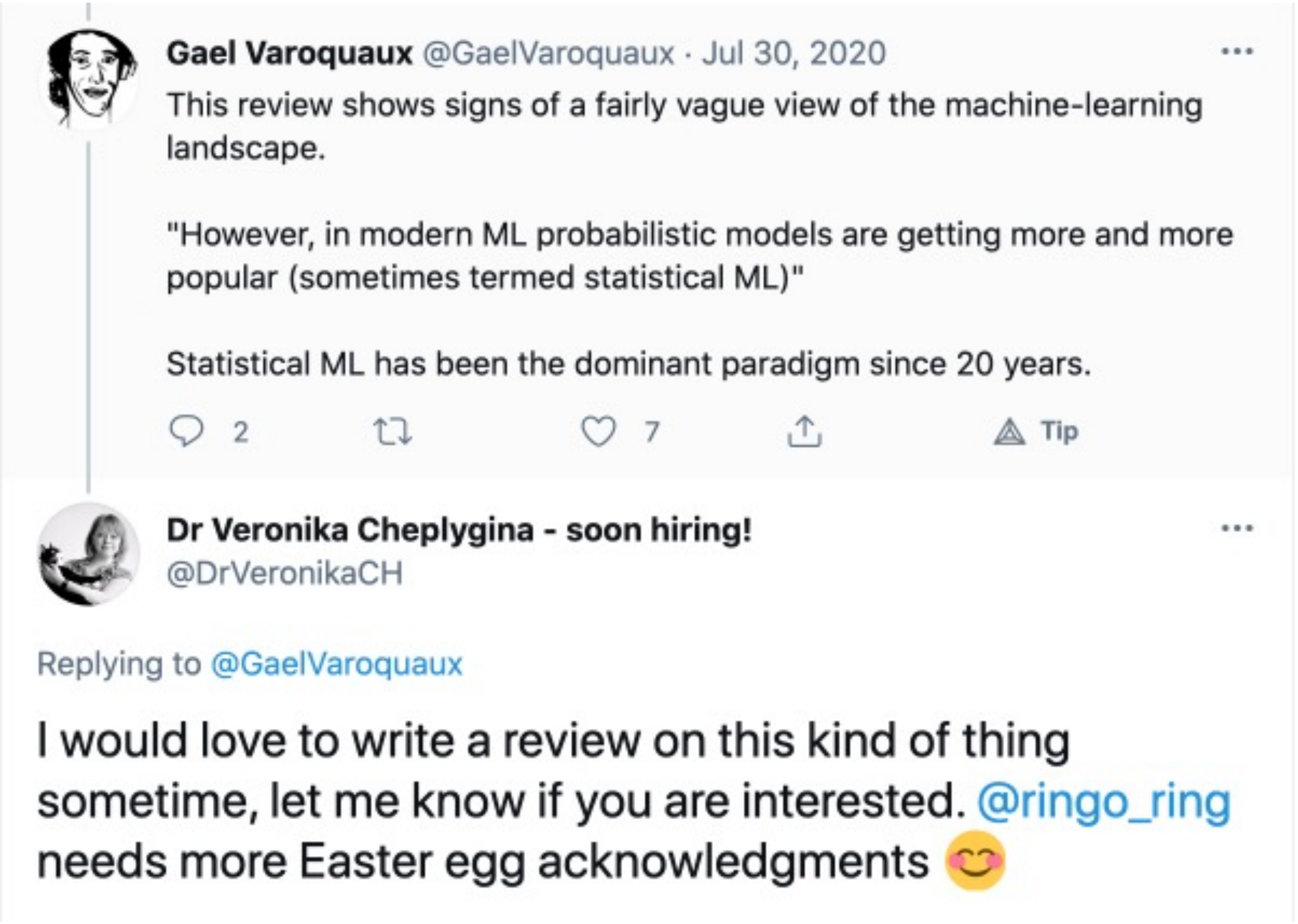
My perspective

- Who gets to do research and why
- I almost left research several times
  https://youtu.be/rdwQeH04OrY
  (Women in MICCAI 2020)
- CV of Failures / How I Fail

https://veronikach.com/category/how-i-fail/

My perspective

- Connecting over shared concerns
- Preprint!
- https://arxiv.org/abs/2103.10292



Outline

- Highlights from preprint
- Misalignment of values & actions
- Ideas how to do better

1

How I failed machine learning in medical imaging - shortcomings and recommendations

Gaël Varoquaux[*], Veronika Cheplygina[†]
[*]INRIA, France
[†]IT University of Copenhagen, Denmark

Disclaimer: this is a working paper, and represents research in progress. For missing references and other comments or questions, please email us at gael.varoquaux@inria.fr and vech@itu.dk

arXiv:2103.10292v1 [eess.IV] 18 Mar 2021

***Abstract*—Medical imaging is an important research field with many opportunities for improving patients' health. However, there are a number of challenges that are slowing down the progress of the field as a whole, such optimizing for publication. In this paper we reviewed several problems related to choosing datasets, methods, evaluation metrics, and publication strategies. With a review of literature and our own analysis, we show that at every step, potential biases can creep in. On a positive note, we also see that initiatives to counteract these problems are already being started. Finally we provide a broad range of recommendations on how to further these address problems in the future. For reproducibility, data and code for our analyses are available on https://github.com/GaelVaroquaux/ml_med_imaging_failures.**

I. Introduction

The great process in machine learning opens the door to many improvements in medical image processing [Litjens et al., 2017, Cheplygina et al., 2019, Zhou et al., 2020]. For example, to diagnose various conditions from medical images, ML algorithms have been shown to perform on par with medical experts [see Liu et al., 2019, for a recent overview]. Software applications are starting to be certified for clinical use [Topol, 2019, Sendak et al., 2020].

The stakes are high, and there is a staggering amount of research on machine learning for medical images, as many recent surveys show. This growth does not inherently lead to clinical progress. The higher volume of research can be aligned with the academic incentives rather than the needs of clinicians and patients. As an example, there can be an oversupply of papers showing state-of-the-art performance on benchmark data, but no practical improvement for the clinical problem.

In this paper, we explore avenues to improve clinical impact of machine learning research in medical imaging. After sketching the situation, documenting uneven progress, we study a number of failures we see in some medical imaging papers, which occur at different steps of the "publishing lifecycle":

- What data to use (Section III)
- What method to use and how to evaluate them (Section IV)
- How to publish the results (Section V)

In each section we first discuss the problems, supported with evidence from previous research as well as our own analyses of recent medical imaging work. We then discuss a number of steps to improve the situation, sometimes borrowed from related communities. We hope that these ideas will help shape a research community even more effective at addressing real-world medical-imaging problems.

II. It's not all about larger datasets

The availability of large labeled datasets has enabled solving difficult artificial intelligence problems, such as natural scene understanding in computer vision [Russakovsky et al., 2015]. As a result, there is widespread hope that similar progress will happen in medical applications: with large datasets, algorithm research will eventually solve a clinical problem posed as discrimination task. Few clinical questions come as well-posed discrimination tasks that can be naturally framed as machine-learning tasks.But, even for these, larger datasets have often failed to lead to the progress hoped for.

One example is that of early diagnosis of Alzheimer's disease (AD), which is a growing health burden due to the aging population. Early diagnosis would open the door to early-stage interventions, most likely to be effective. Hence, efforts have been dedicated to acquire large brain-imaging cohorts of aging individuals at risk of developing AD, on which early biomarkers can be developed using machine learning [Mueller et al., 2005]. As a result, there have been steady increases in the typical sample size of studies applying machine learning to develop computer-aided diagnosis of AD, or its predecessor, mild cognitive impairment, as visible in Figure 1a, built with a meta-analysis compiling 478 studies from 6 systematic reviews [Dallora et al., 2017, Arbabshirani et al., 2017, Liu et al., 2019, Sakai and Yamada, 2019, Wen et al., 2020, Ansart et al., 2020].

However, the increase in data size did not come with better diagnostic accuracy, in particular for the most clinically-relevant question, distinguishing pathological versus stable evolution for patients with symptoms of prodromal Alzheimer's (Figure 1b). Rather, studies with larger sample sizes tend to report worse prediction accuracy. This is worrisome, as these larger studies are closer to real-life settings. However, research efforts across time lead to improvements even on large, heterogeneous cohorts (Figure 1c), as studies published later show improvements for large sample sizes.

@drveronikach

https://www.veronikach.com

Values

@drveronikach

https://www.veronikach.com

Why do I/we do research?

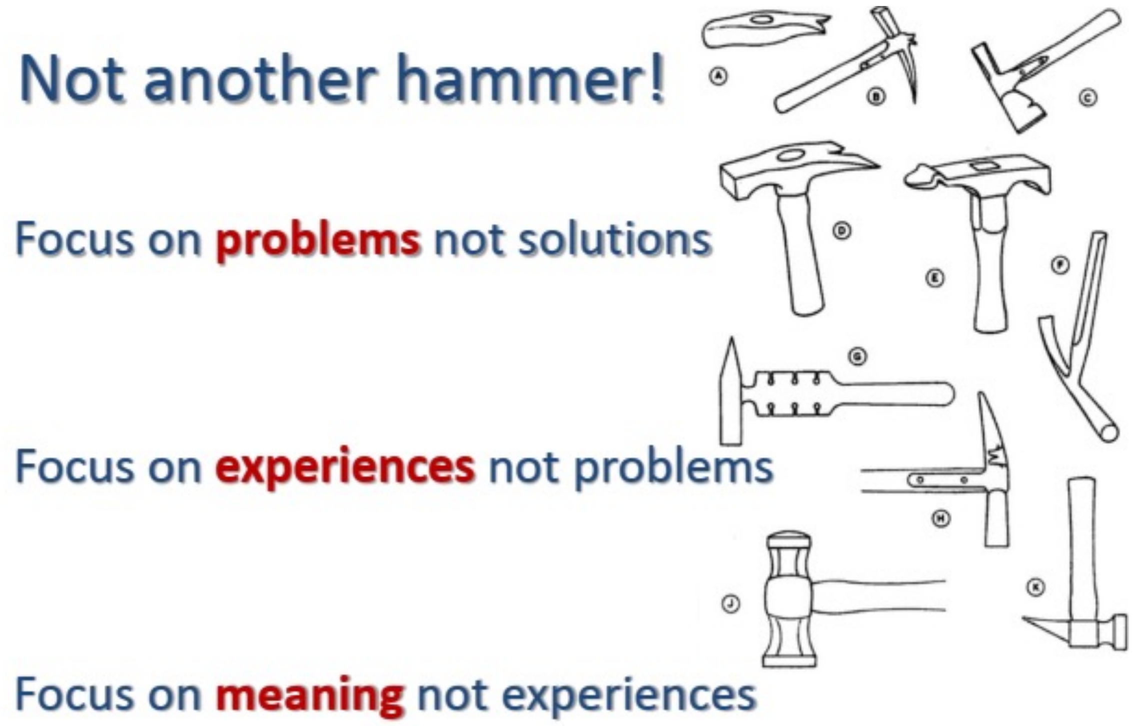
- Solve problems
- Help people
- Learn from experience

What should I research?

What are the biggest problems in the world? What are you working on?

What sentence in a textbook will your research change?

Don’t invent another hammer



Slide from Bob Williamson “Research problem choice”, MLCB summer school, Tübingen 2013

A simplistic view

Methods → Problems ↓	1	2	3	4	5	…
Recognize numbers	✓✓	✓				
Find photos of cats	✓		✓✓			
Diagnose lung cancer		✓✓	✓			
…..			✓	✓✓		
Next problem?	?	?	?	?	?	

Actions

@drveronikach

https://www.veronikach.com

Datasets are a reflection of reality

- Early diagnosis vs advanced disease
- “Hidden stratification”- pneumothorax & chest drain (AUC 0.94 vs 0.77)

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

Luke Oakden-Rayner,* Jared Dunnmon,* Gustavo Carneiro, and Christopher Ré

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7665161/

Datasets are a reflection of reality

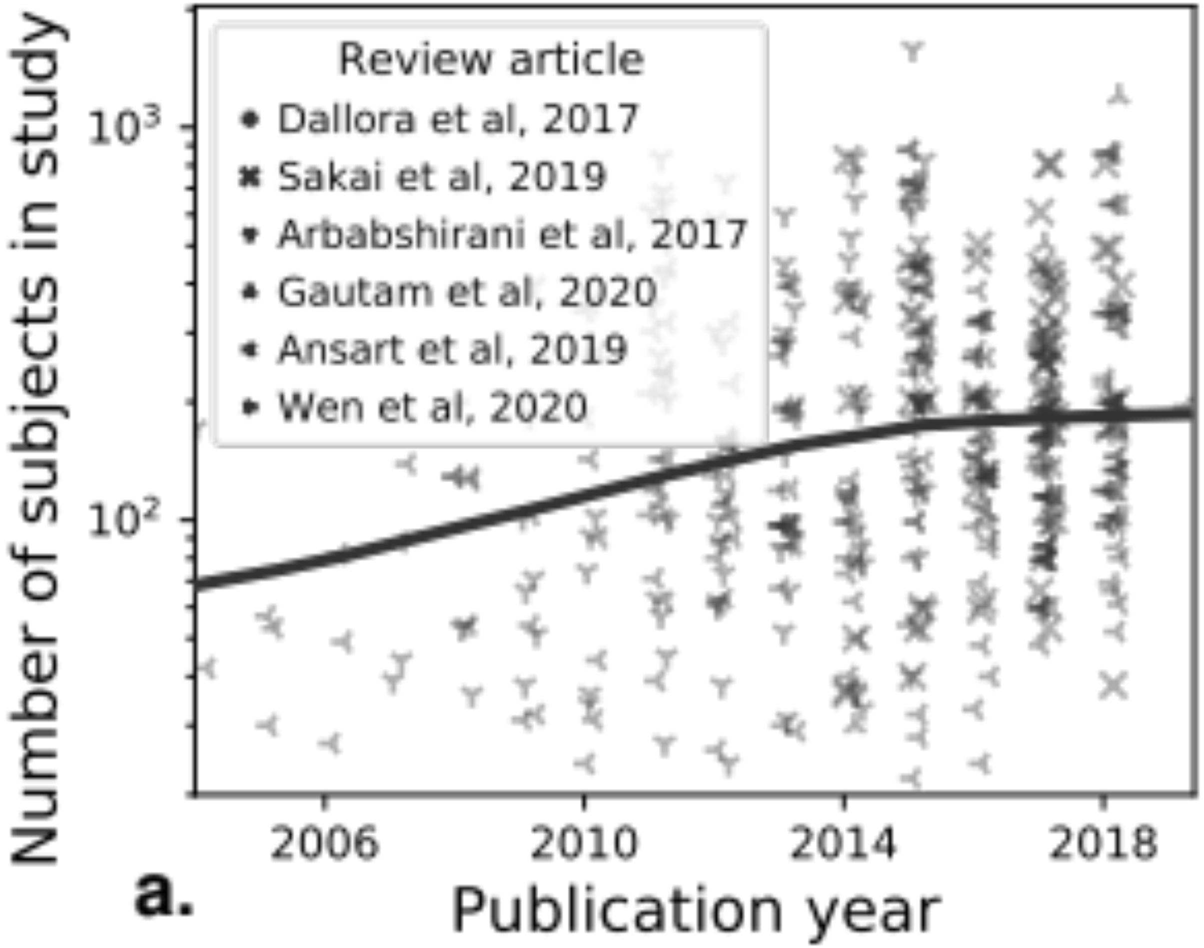
- Subset of population, even in large datasets

Test set	Training set	Atelectasis	Cardiomegaly	Consolidation
ChestX-ray14	**ChestX-ray14**	**0.8165**	**0.8998**	**0.8181**
	CheXpert	0.7850	0.8646	0.7771
	MIMIC-CXR	0.8024	0.8322	0.7898
CheXpert	ChestX-ray14	0.5137	0.5736	0.6565
	CheXpert	**0.6930**	**0.8687**	**0.7323**
	MIMIC-CXR	0.6576	0.8197	0.7002
MIMIC-CXR	ChestX-ray14	0.5810	0.6798	0.7692
	CheXpert	0.7587	0.7650	0.7936
	MIMIC-CXR	**0.8177**	**0.8126**	**0.8229**

Pooch, E. H., Ballester, P. L., & Barros, R. C. (2019). Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940*.

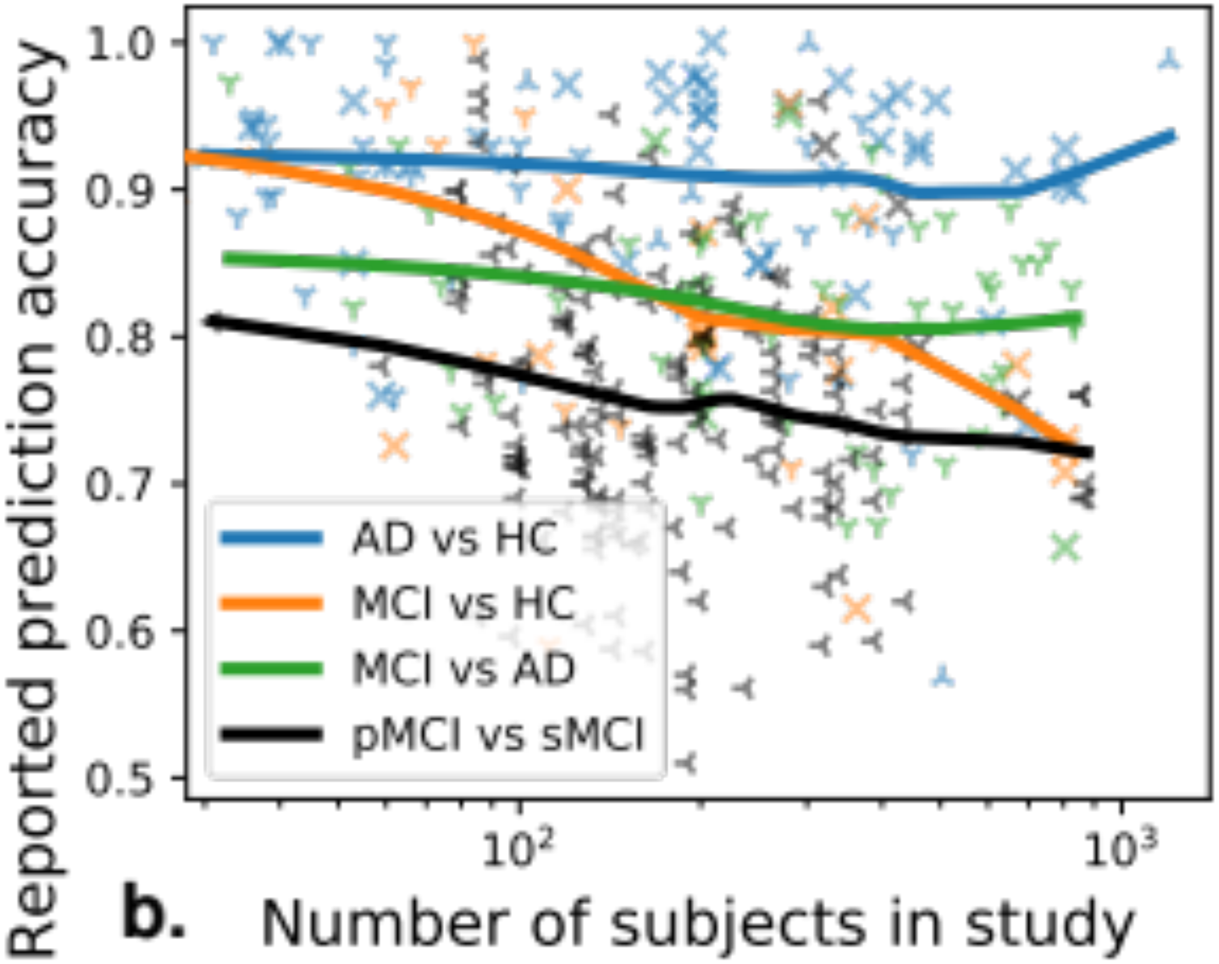
Larger datasets are not everything

- Limited growth of sample size



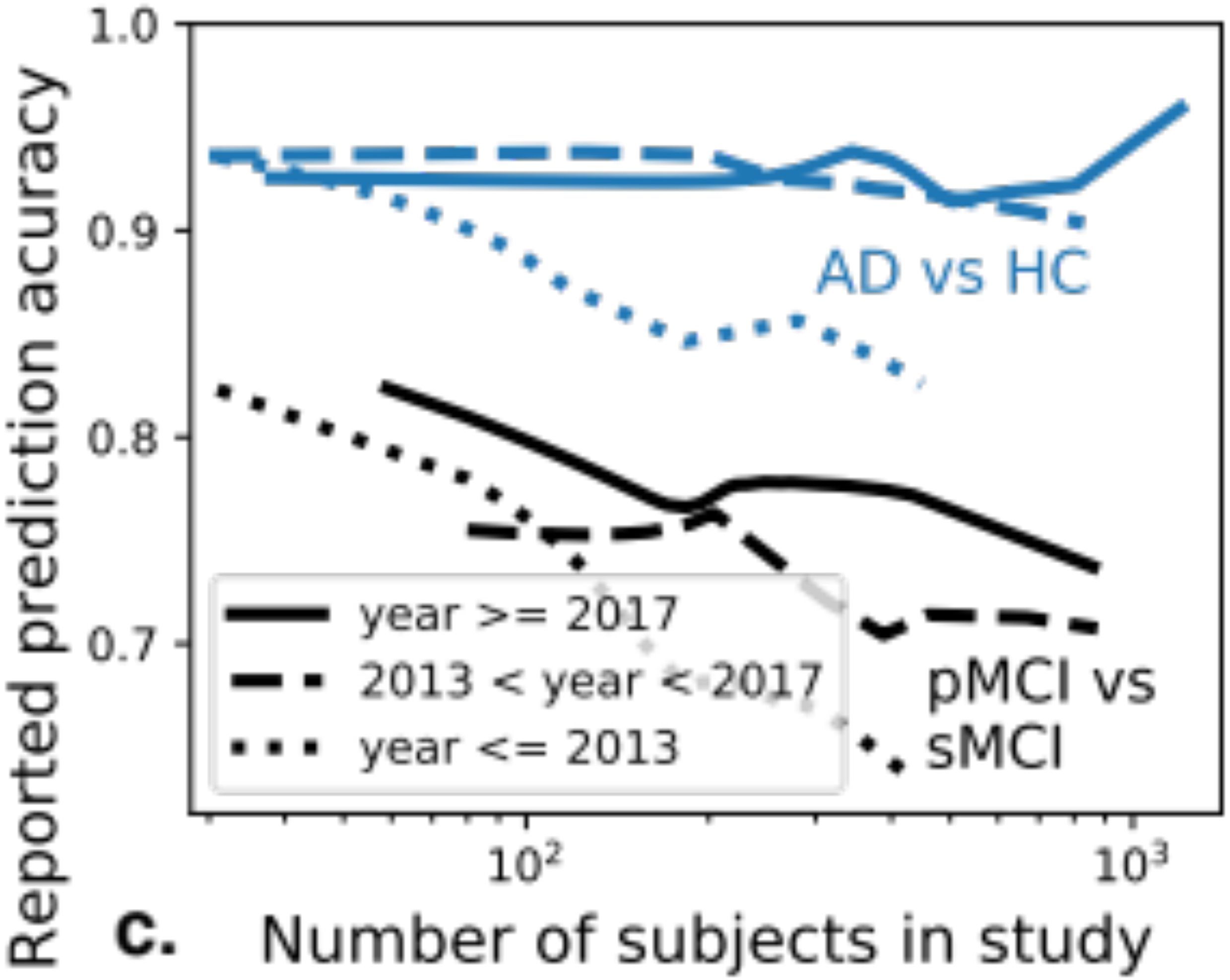
Larger datasets are not everything

- Larger test sets show earlier overfitting

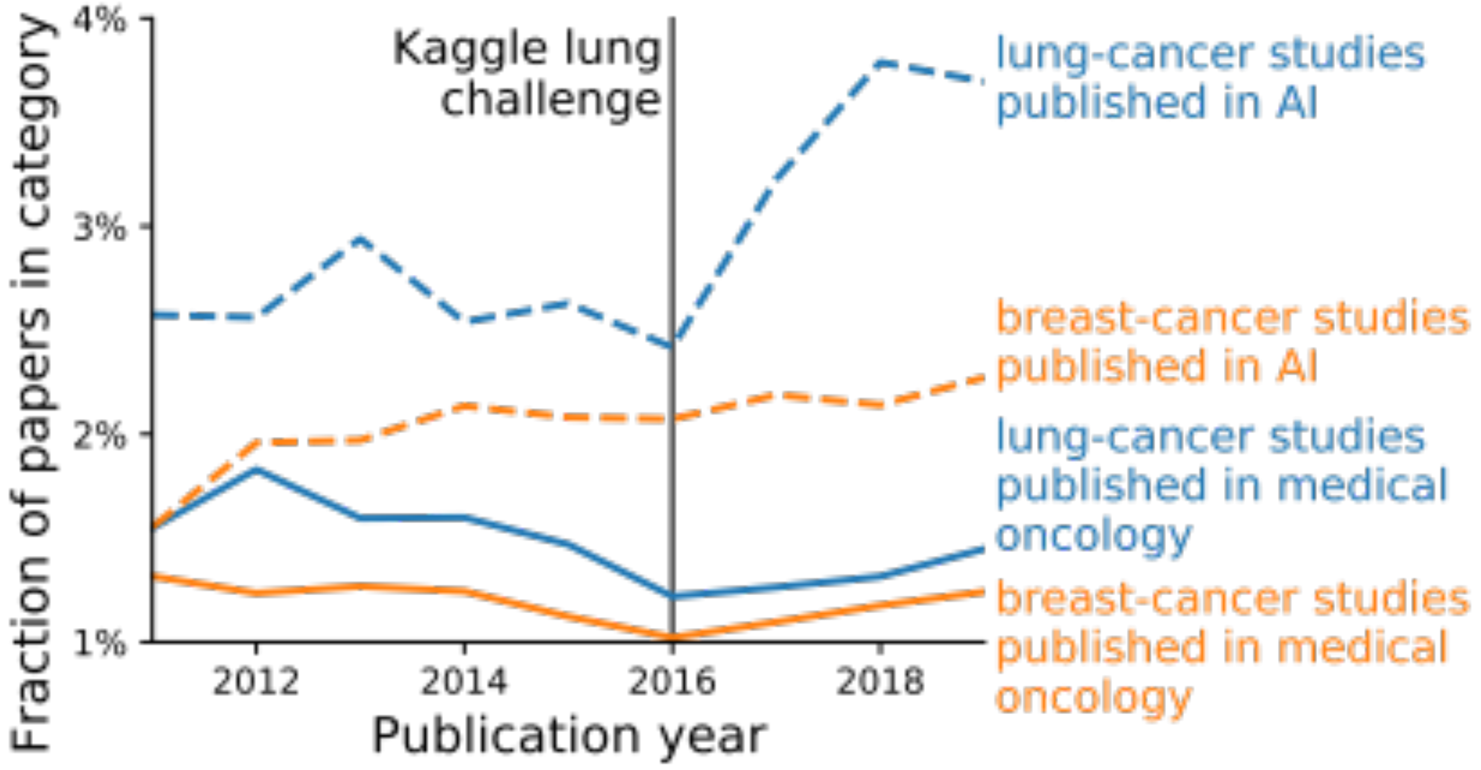


…but there is progress

- Better generalization in recent years



Large benchmarks change focus



Novelty

Needlessly complex methods

"Mathiness" / Proof by intimidation

Failure to identify sources of gains

Troubling Trends in Machine-learning Scholarship

ZACHARY C. LIPTON AND JACOB STEINHARDT

SOME ML PAPERS SUFFER FROM FLAWS THAT COULD MISLEAD THE PUBLIC AND STYMIE FUTURE RESEARCH.

https://dl.acm.org/doi/10.1145/3317287.3328534

State-of-the-art results

Baselines too simple, or not simple enough

Single focus on accuracy (or similar), variability often not considered

Statistical significance can be misunderstood

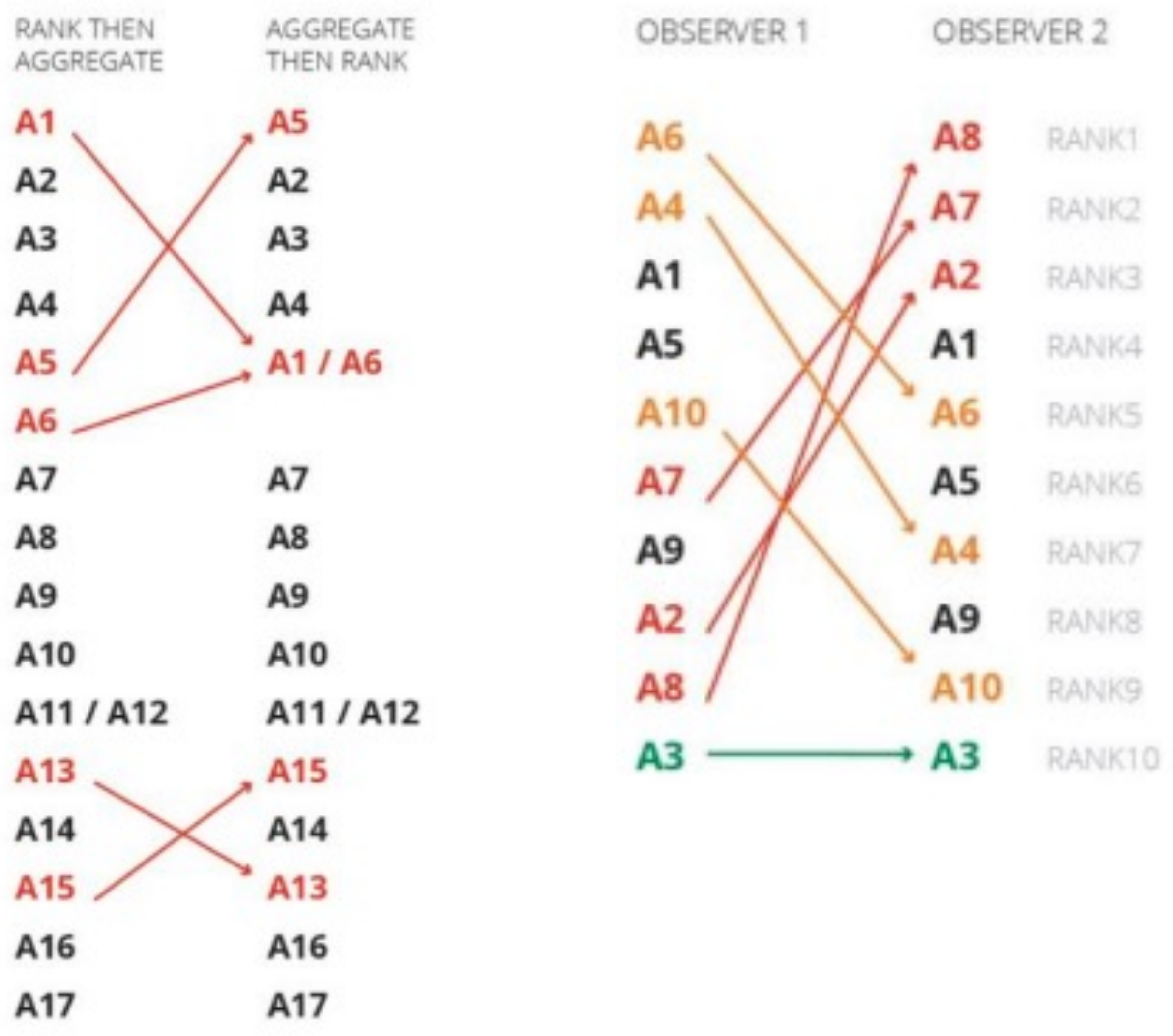
Statistical significance is not practical significance

@drveronikach

https://www.veronikach.com

State-of-the-art results

- Depends how you do the ranking
- Here: segmentation, multiple images with performance scores, and 2x ground truth
- Figures from Maier-Hein et al, https://arxiv.org/pdf/1806.02051.pdf

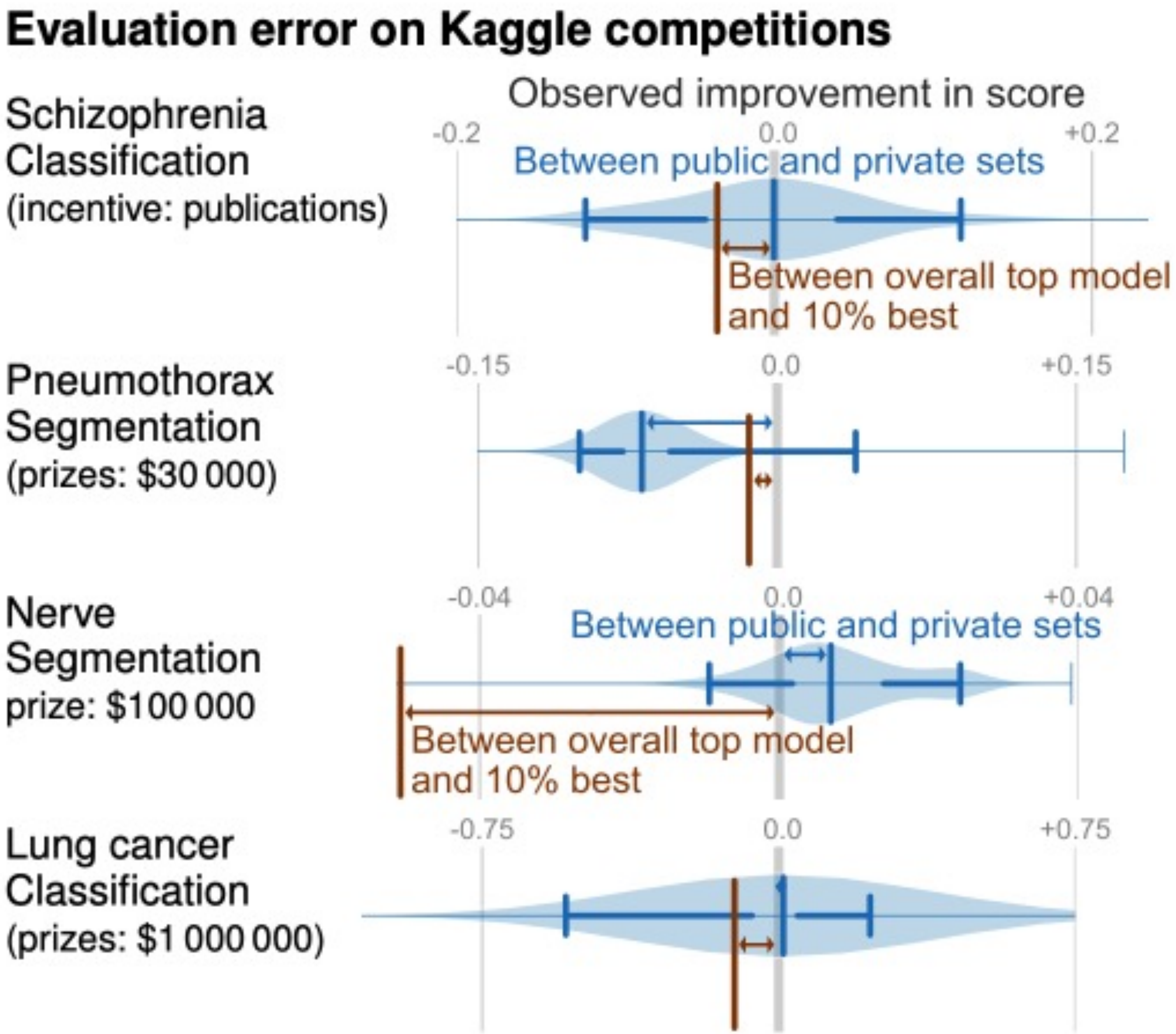


Overfitting

Public/private leaderboard differences (in blue)

Mean < 0 = private result worse

Top 10% gap vs noise in results (in brown)

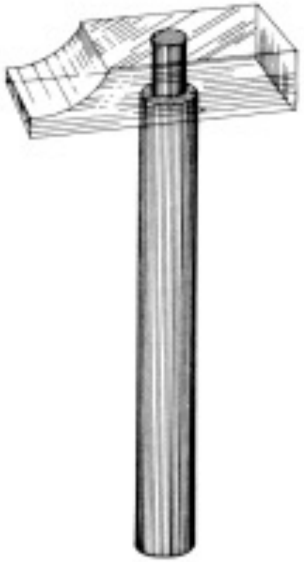


Where we are now

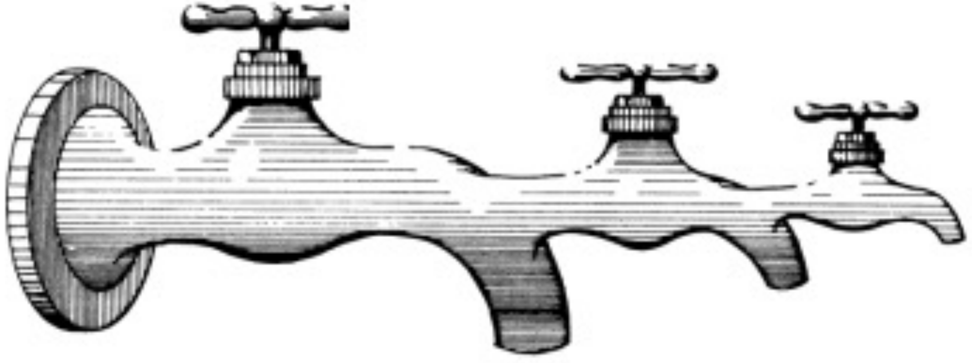
Where we are

Novelty, prestigious conference, …

Methods → Datasets ↓	1	2	3	4	5	…																																	
Cats	✓ ✓	✓																																					
Lung cancer	✓		✓ ✓																																				
…		✓ ✓	✓																																				
…			✓	✓ ✓																																			
…																																							



A6 — Marteau tordu. Sa forme spéciale lui permet d'atteindre aisément les clous les plus inaccessibles.



G3 — Robinet à débits différents. Économisez l'eau en choisissant le filet que vous voulez faire couler.

Lecture Bob Williamson "Research problem choice", MLCB summer school, Tübingen 2013

Where we are

De-democratization of AI https://arxiv.org/pdf/2010.15581.pdf

170K papers at 57 CS conferences – “large firms and elite universities increased participation since 2012”

Fueled by “compute divide”

The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research

Nur Ahmed*

Muntasir Wahed§

Where we are

Hardware lottery

Idea wins because of suitability of hardware/software

“Increasingly costly to stray off of the beaten path of research ideas”

The Hardware Lottery

Sara Hooker

Google Research, Brain Team
shooker@google.com

Abstract

Hardware, systems and algorithms research communities have historically had different incentive structures and fluctuating motivation to engage with each other explicitly. This historical treatment is odd given that hardware and software have frequently determined which research ideas succeed (and fail). This essay introduces the term hardware lottery to describe when a research idea wins because it is suited to the available software and hardware and *not* because the idea is superior to alternative research directions. Examples from early computer science history illustrate how hardware lotteries can delay research progress by casting successful ideas as failures. These lessons are particularly salient given the advent of domain specialized hardware which make it increasingly costly to stray off of the beaten path of research ideas. This essay posits that the gains from progress in computing are likely to become even more uneven, with certain research directions moving into the fast-lane while progress on others is further obstructed.

Where we are

“Grad student descent”
https://arxiv.org/pdf/1904.07633

“type of optimization scheme in which the task of model architecture or hyper-parameter search is assigned to several graduate students”

HARK Side of Deep Learning - From Grad Student Descent to Automated Machine Learning

Oguzhan Gencoglu
Top Data Science Ltd.
Helsinki, Finland
oguzhan.gencoglu@topdatascience.com

Mark van Gils
VTT Technical Research Centre of Finland Ltd.
Tampere, Finland
mark.vangils@vtt.fi

Esin Guldogan
Huawei Technologies
Tampere, Finland
esin.guldogan@huawei.com

Chamin Morikawa
Morpho Inc.
Tokyo, Japan
c-morikawa@morphoinc.com

Mehmet Süzen
Jülich, Germany
suzen@acm.org

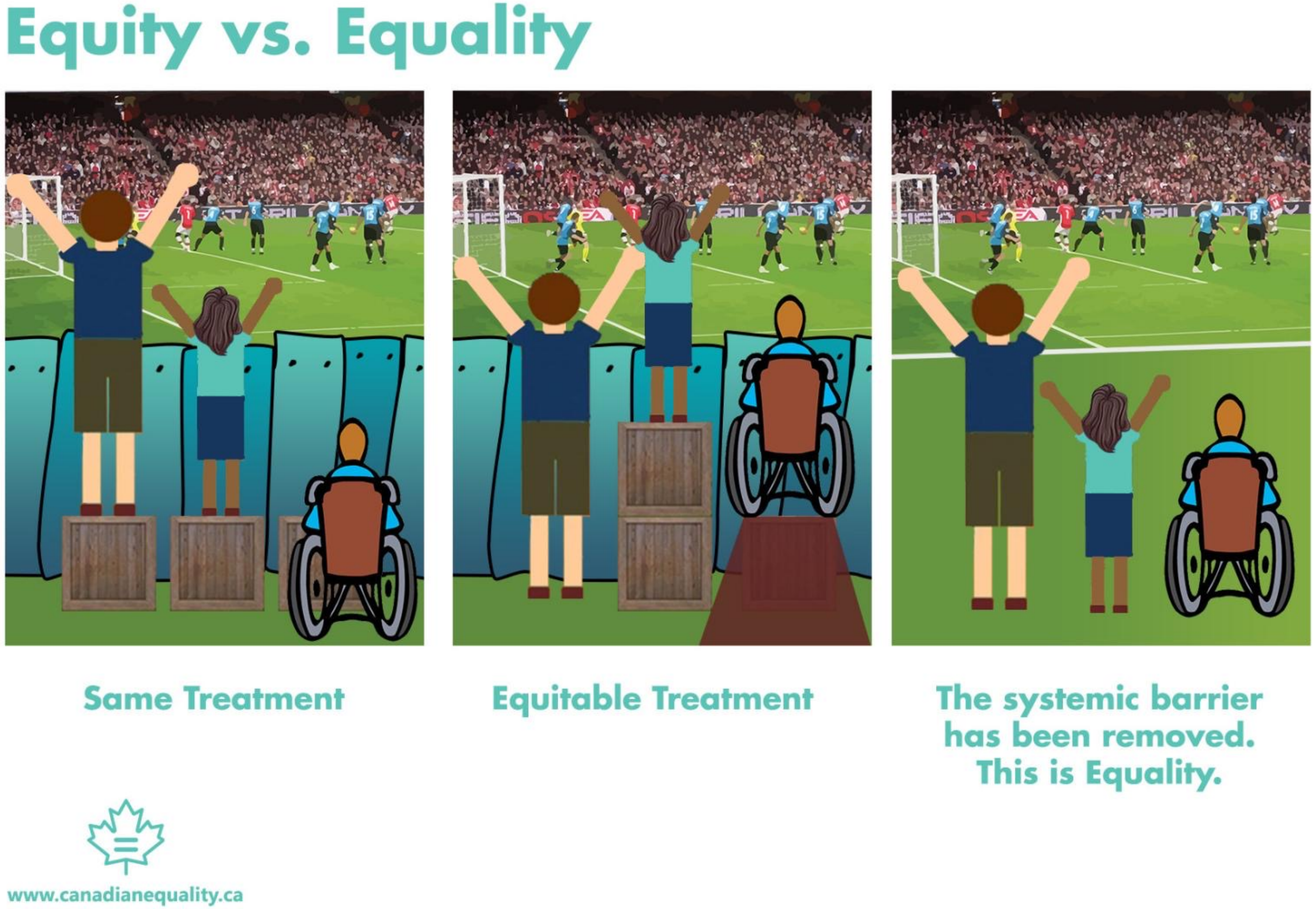
Mathias Gruber
Novozymes
Copenhagen, Denmark
mafg@novozymes.com

Jussi Leinonen
Bayer
Espoo, Finland
jussi.leinonen@bayer.com

Heikki Huttunen
Tampere University
Tampere, Finland
heikki.huttunen@tuni.fi

Where we are

Leaky pipeline



@drveronikach

https://www.veronikach.com

Doing better

@drveronikach

https://www.veronikach.com

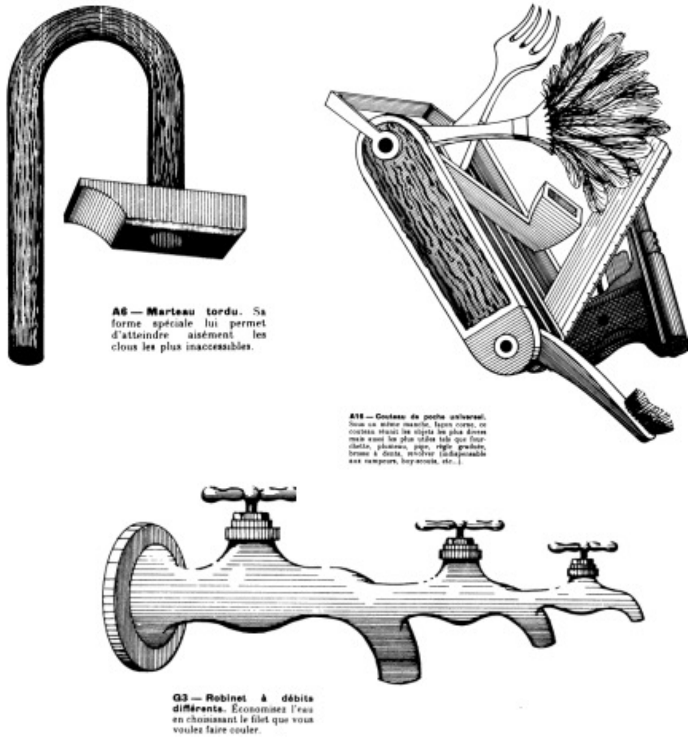
Recommendations

Focus on datasets!

Cite datasets

Investigate dataset shift/bias, labeling/“gold standard”

Be transparent about limitations (e.g. model cards https://arxiv.org/pdf/1810.03993)



Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

Recommendations

Methods & their evaluation

Representative data & strong baselines

Collaboration, not competition (understanding!)

Recommendations

Incentives / Goodhart’s Law

Metrics (impact on world, qualitative accounts)

Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models

Lasse F. Wolff Anthony[*,1] Benjamin Kanding[*,1] Raghavendra Selvan[1]

https://arxiv.org/abs/2007.03051

Reliance on Metrics is a Fundamental Challenge for AI

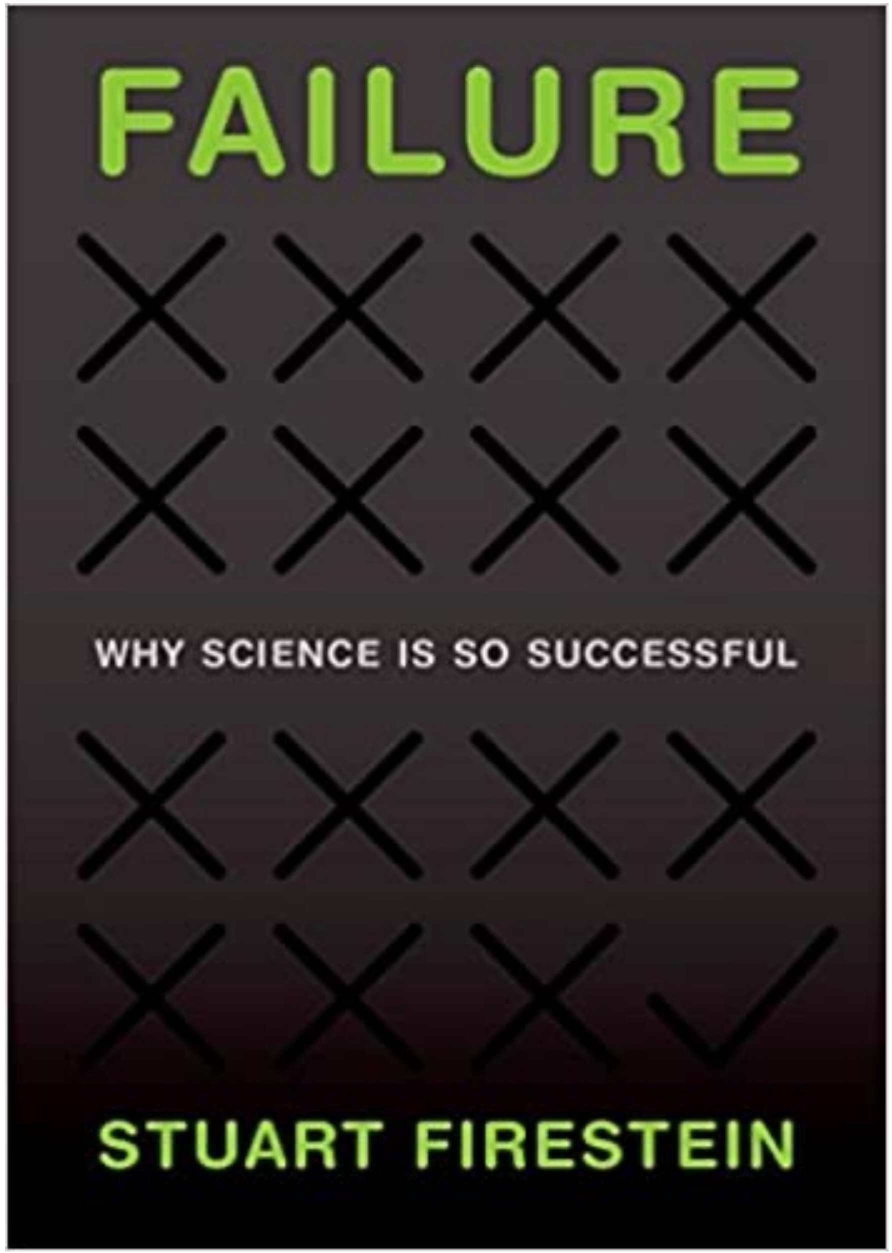
Rachel L. Thomas
University of San Francisco
rlthomas3@usfca.edu

David Uminsky
University of San Francisco
duminsky@usfca.edu

https://arxiv.org/pdf/2002.08512

Recommendations

Scientists are often wrong



- Solve problems
- Help people
- Learn from experience


Thank you!
Soon hiring 2 PhD researchers!
@drveronikach
https://www.veronikach.com