Machine learning in medical imaging - shortcomings and recommendations

NVPHBV Spring Meeting

18th May 2022

Dr. Veronika Cheplygina, IT University of Copenhagen

@drveronikach

https://www.veronikach.com

Why do we do research?

- Solve problems
- Help others
- 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

Learning to solve problems, as a dataset

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

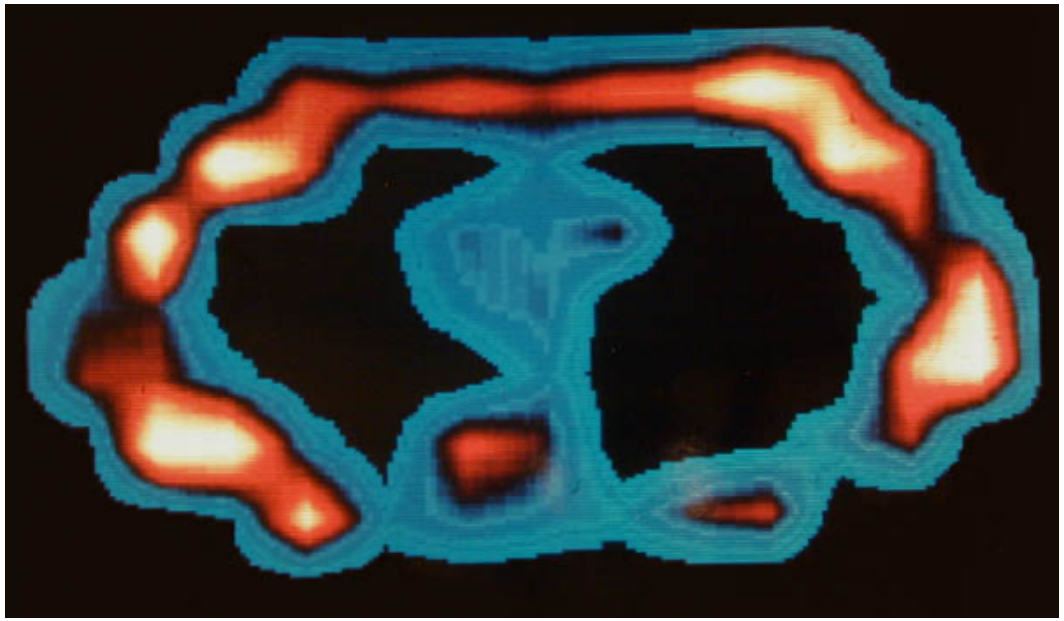
Classification of medical images

How it started

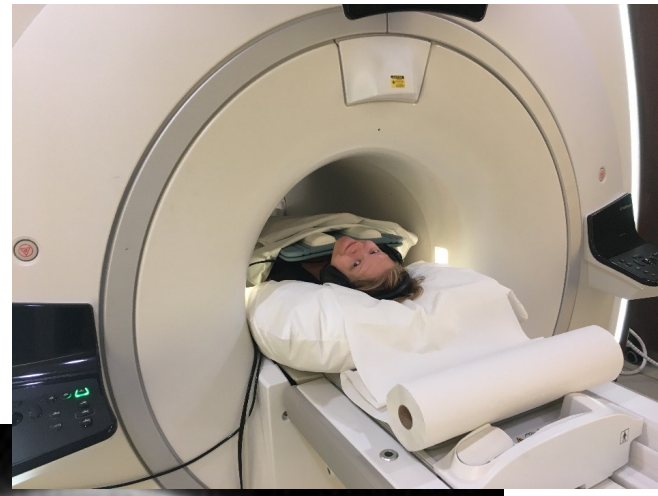
How it's going

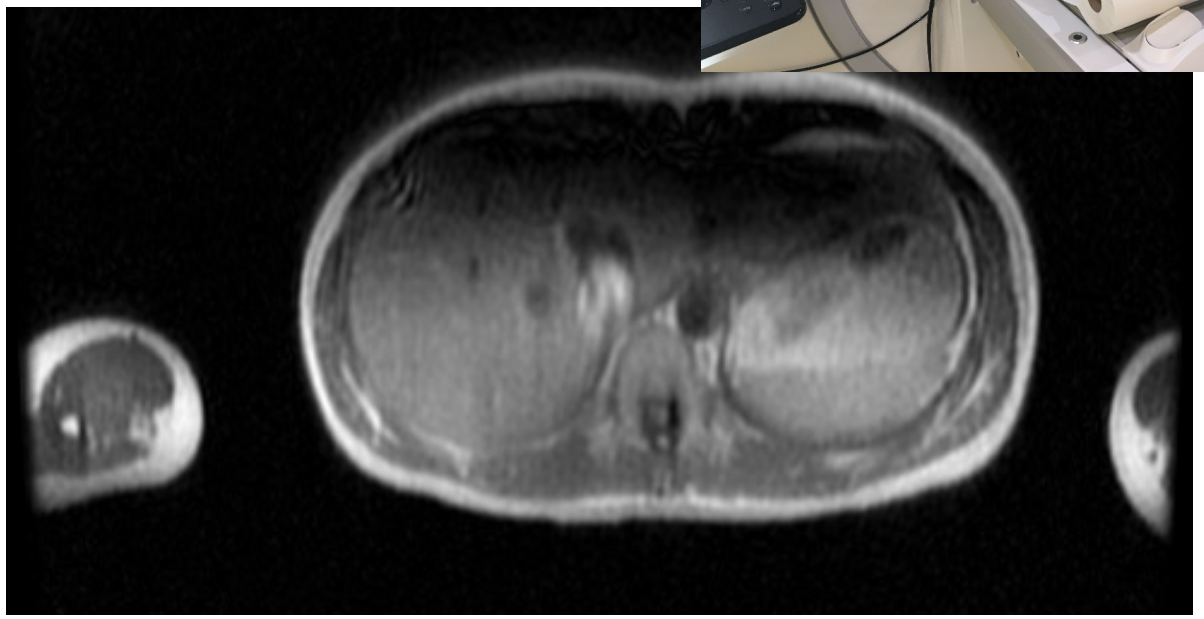
Classification of medical images

How it started



How it's going





Outline

- How we (try to) generalize within a medical imaging problem
- Why this is not enough to solve problems more generally
- How to do better (in expectation)

Gael Varoquaux
@GaelVaroquaux

Review Article | Open Access | Published: 12 April 2022

Machine learning for medical imaging: methodological failures and recommendations for the future

Gaël Varoquaux & Veronika Cheplygina

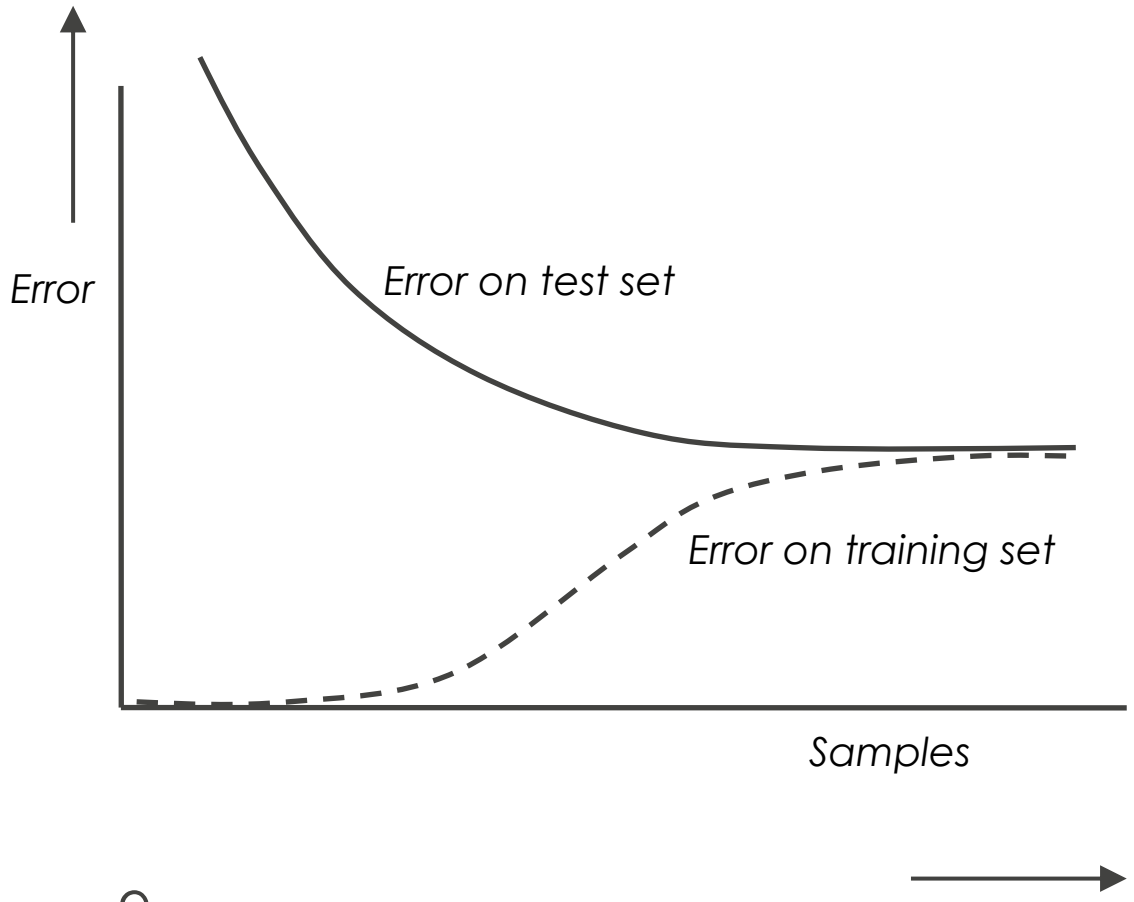
npj Digital Medicine **5**, Article number: 48 (2022) | Cite this article

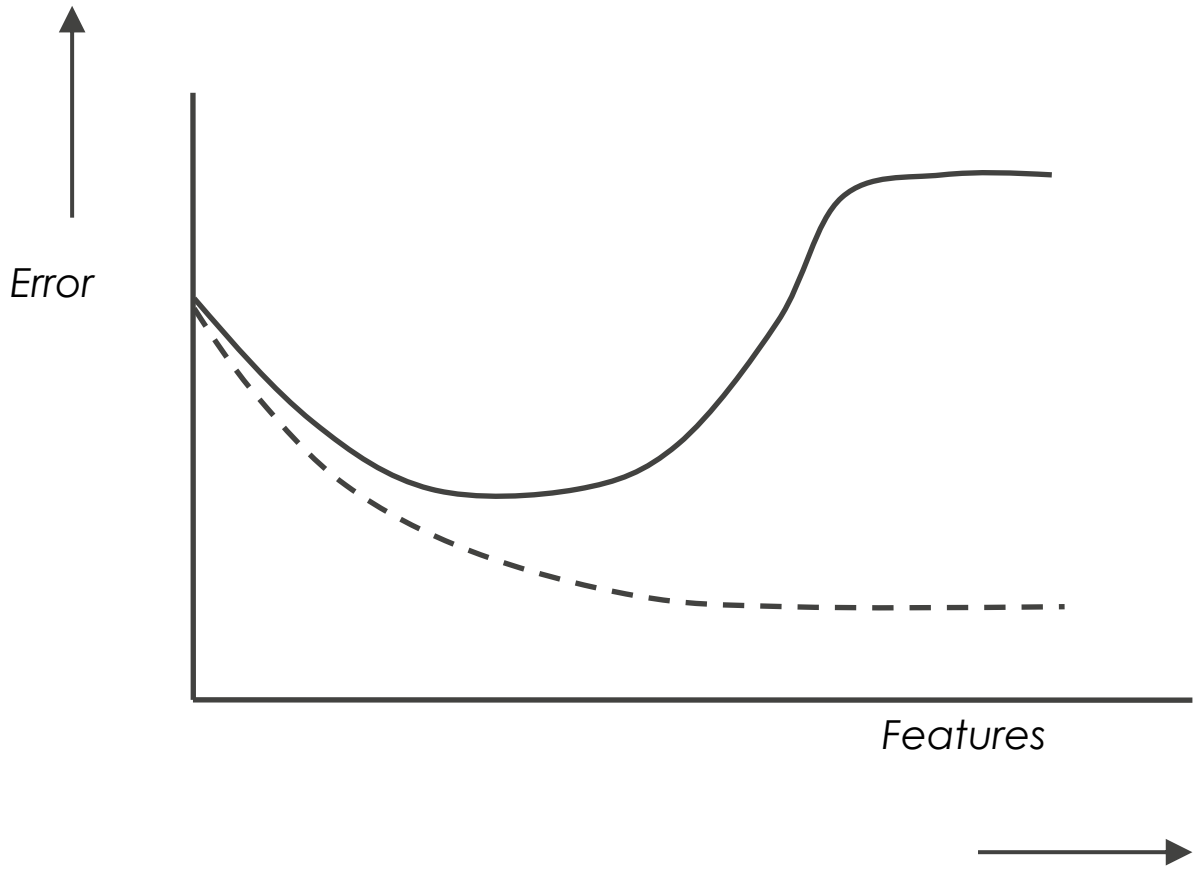
Learning with limited labeled data

@drveronikach

https://www.veronikach.com

What should we do to generalize?





Images by David Tax

Recent developments - datasets

- Large(r) public datasets
  - CheXpert, Chest-Xray14, MIMIC (30-65K patients, 110-225K x-rays)

ChestX-ray14 CheXpert MIMIC-CXR

Recent developments - datasets

Data Science Bowl 2017

Can you improve lung cancer detection?

$1,000,000
Prize Money

Booz Allen Hamilton · 1,972 teams · 5 years ago

Overview Data Code Discussion Leaderboard Rules

Join Competition ···

Public Private

The private leaderboard is calculated with approximately 99% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Prize Winners

#	△	Team	Members		Score	Entries	Last	Code
1	▲ 136	grt123			0.39975	2	5Y	
2	▲ 87	Julian de Wit & Daniel Hammack			0.40117	2	5Y	

Recent developments - methods

- Active learning
- Crowdsourcing
- Data augmentation
- Generative adversarial networks
- Multi-task learning
- Multiple instance learning
- Regularization
- Self-supervised learning
- Semi-supervised learning
- Transfer learning...

Image

Multi-task learning

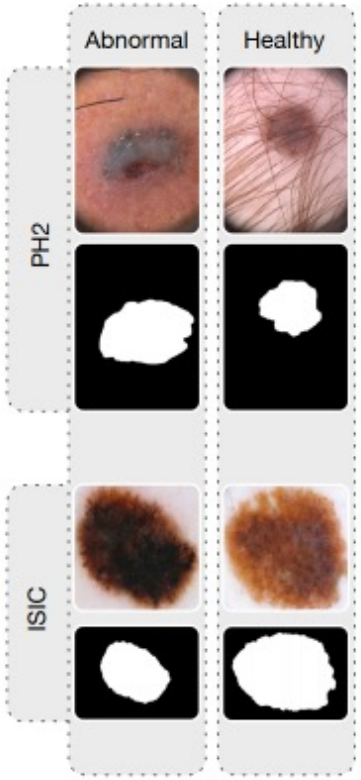
Ralf Raumanns

Skin lesion classification (Asymmetry, Border, Color)

Annotations of ABC features by crowdsourcing, students, algorithms

Baseline (diagnosis) vs multi-task (diagnosis & annotations)

[Raumanns et al 2021]



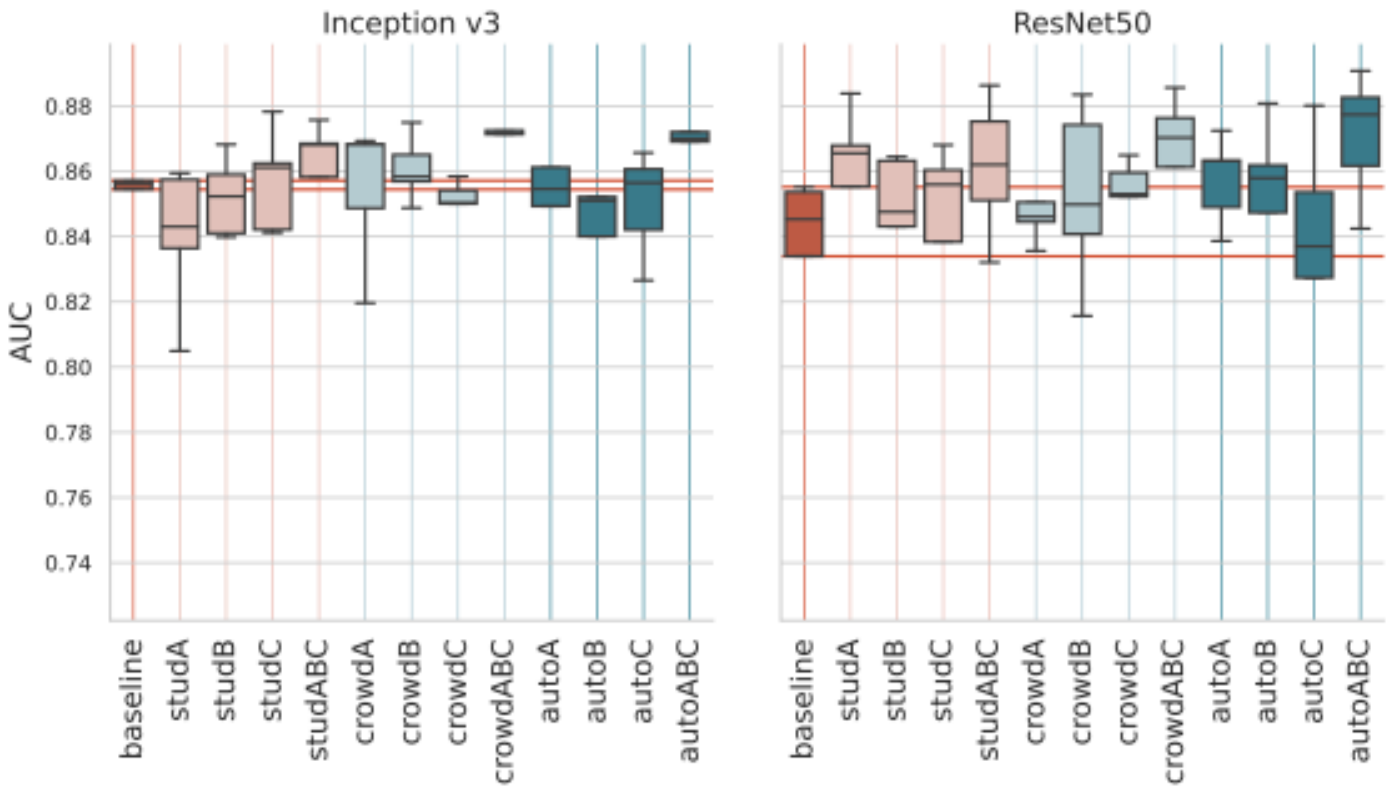
	Asymmetry	Border	Color
Abnormal			
Automated	2	10	5
Crowd	1,2,0	4,8,0	4,4,3
Expert	2	-	2
Healthy			
Automated	2	7	6
Crowd	0,0,0	8,5,7	2,1,1
Expert	0	-	1
Abnormal			
Automated	2	50	5
Crowd	1,1,2	8,2,7	1,1,1
Student	1,1,2	5,3,6	4,4,4
Healthy			
Automated	2	62	6
Crowd	0,0,1	7,6,6	3,3,1
Student	2,2,2	5,3,8	3,3,3

Multi-task learning

Annotations are noisy, but informative when used as additional tasks

Data+code

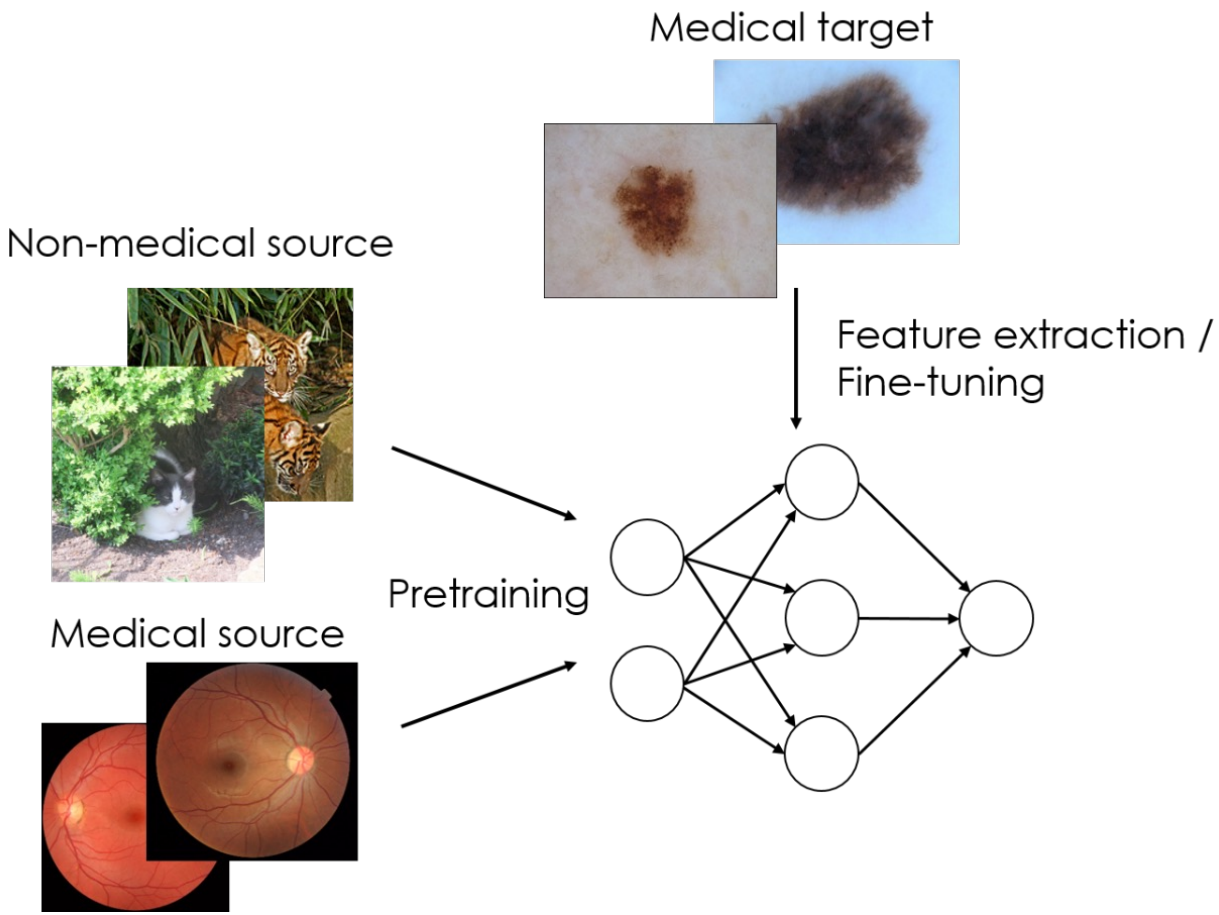
https://github.com/raumannsr/ENHANCE



Transfer learning

- Training on (large) source data, then on (small) target data

- Surprising/conflicting results on best sources - should be "similar"



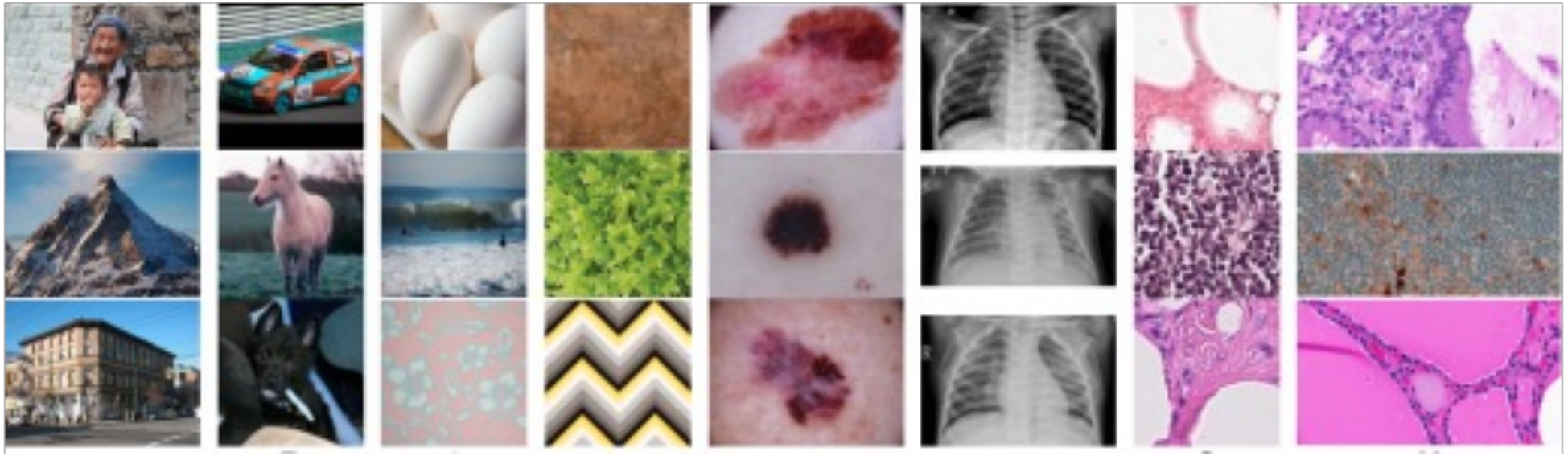
Cats or CAT scans: Transfer learning from natural or medical image source data sets?

Veronika Cheplygina

Irma van den Brandt

Transfer learning

- Systematic comparison with 8 datasets
- ImageNet best*, but much smaller texture dataset close



[van den Brandt et al 2021]

Transfer learning

CATS - Choosing a Transfer Source for medical image classifcation

novo nordisk fonden

- Can we predict “transferability” (meta-learning)?
- How do researchers choose/compare datasets?

Dovile Juodelyte

Bethany Chamberlain

Recent developments - methods

Use additional data and/or assumptions to

- (Implicitly) increase sample size
- Reduce complexity

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
S Ørting, A Doyle, A van Hilten, M Hirth, O Inel, CR Madan, P Mavridis, ...
arXiv preprint arXiv:1902.09159

Many successes reported

Chexnet: **Radiologist-level** pneumonia detection on chest x-rays with deep learning

P Rajpurkar, J Irvin, K Zhu, B Yang, H Mehta… - arXiv preprint arXiv …, 2017 - arxiv.org

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Our algorithm, CheXNet, is a 121-layer convolutional neural network …

Save Cite Cited by 1752 Related articles All 9 versions

However…

- *"none of the models identified are of potential clinical use"* [Roberts et al 2021]

- *"[…] narrow use cases […] limited external validation […]"* [Kelly et al 2022]

nature > nature machine intelligence > analyses > article

Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

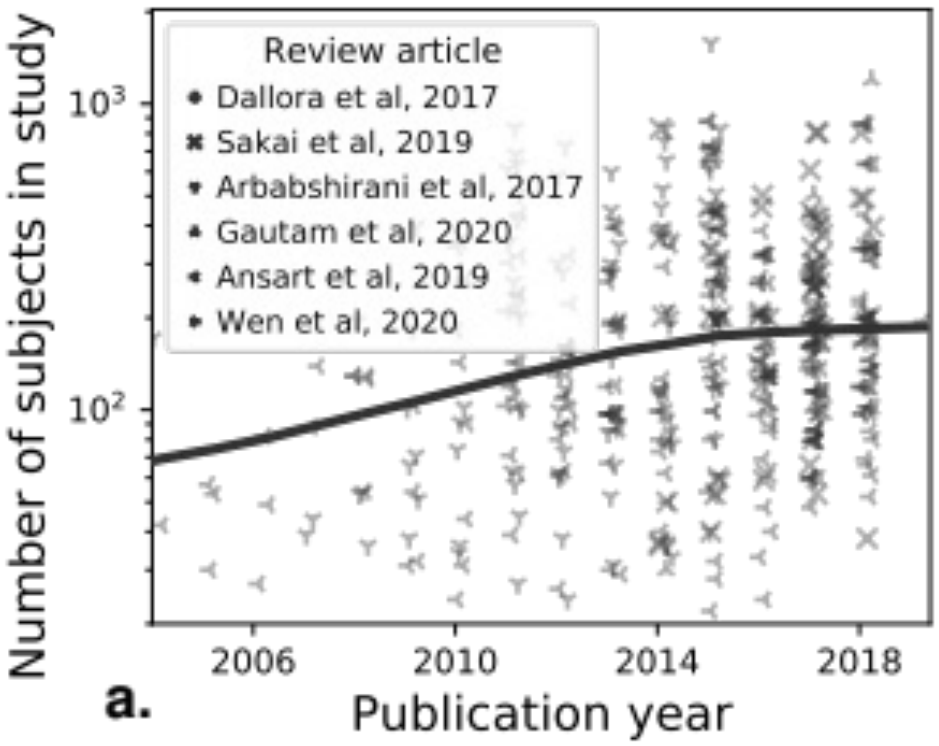
Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE)

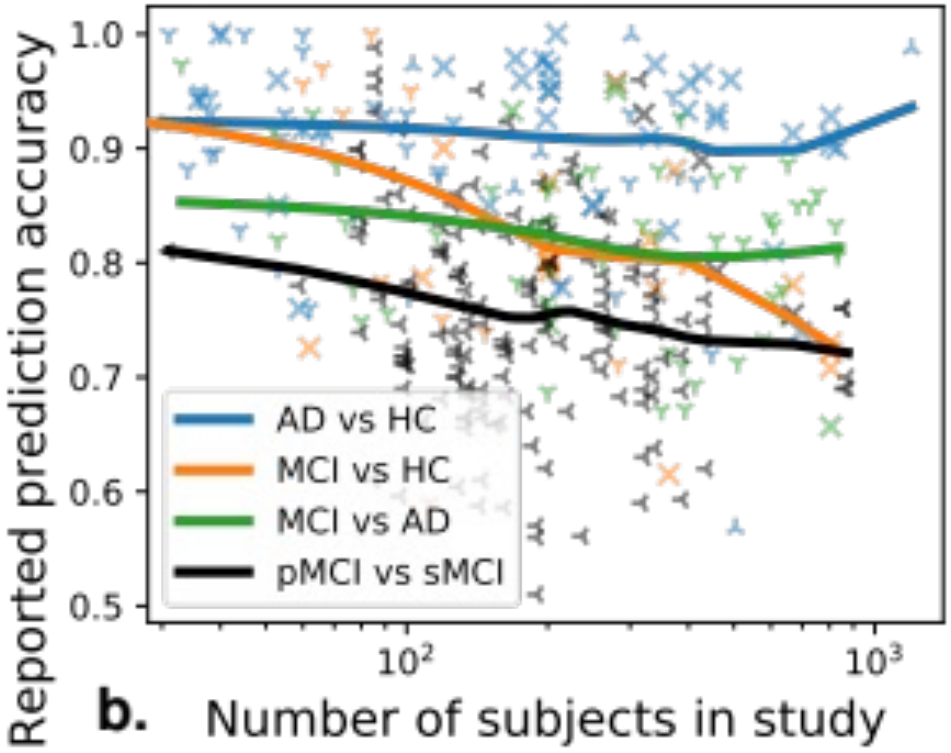
Brendan S. Kelly[1,2,3,4,5,6] · Conor Judge[5,6] · Stephanie M. Bollard[4,5,6] · Simon M. Clifford[1,6] · Gerard M. Healy[1,6] · Awsam Aziz[4,6] · Prateek Mathur[2,6] · Shah Islam[6,7] · Kristen W. Yeom[7,8] · Aonghus Lawlor[2,6] · Ronan P. Killeen[2,4,6]

Why is it not enough?

Problem 1: Datasets only a reflection of reality

- Limited growth of sample size in diagnosis of Alzheimer's
- Larger test sets show overfitting





Datasets only a reflection of reality

Dataset shift/bias even in larger datasets

- Patient demographics
- Early diagnosis vs advanced disease
- “Shortcuts”

ChestX-ray14 CheXpert MIMIC-CXR

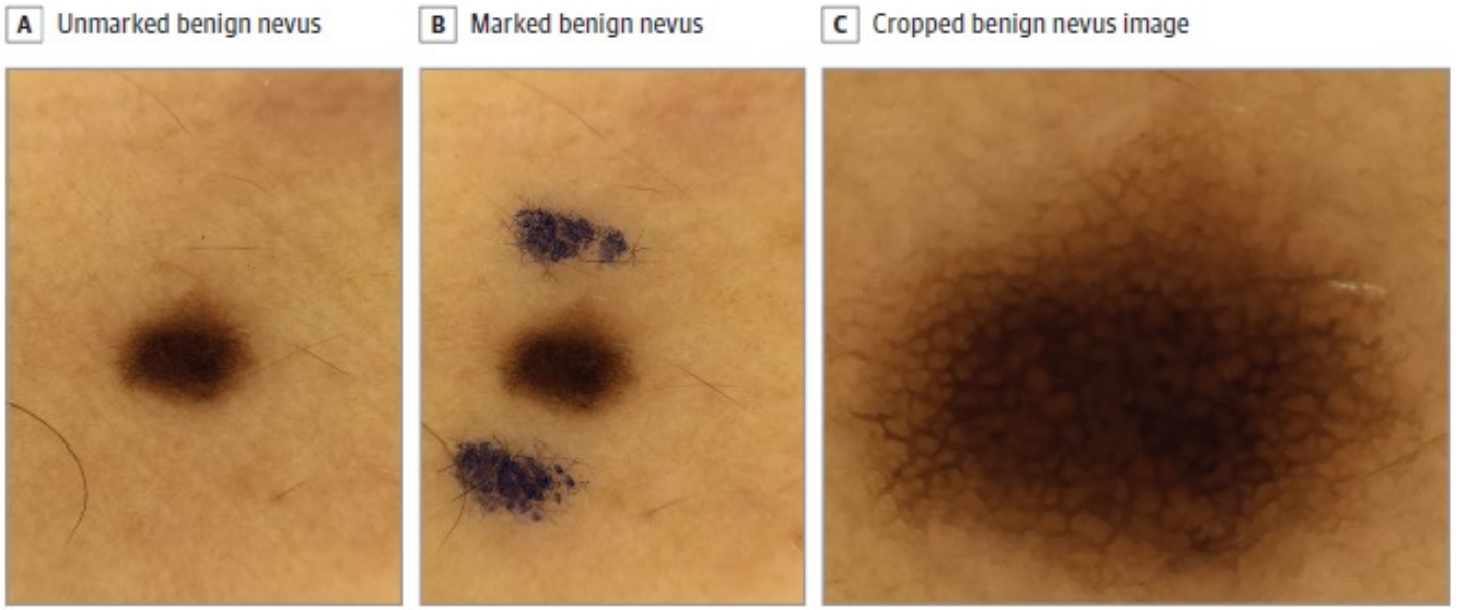
[Pooch et al 2019]

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**

Datasets - Shortcuts

- Pen marks correlated with melanoma
- Network flips diagnosis

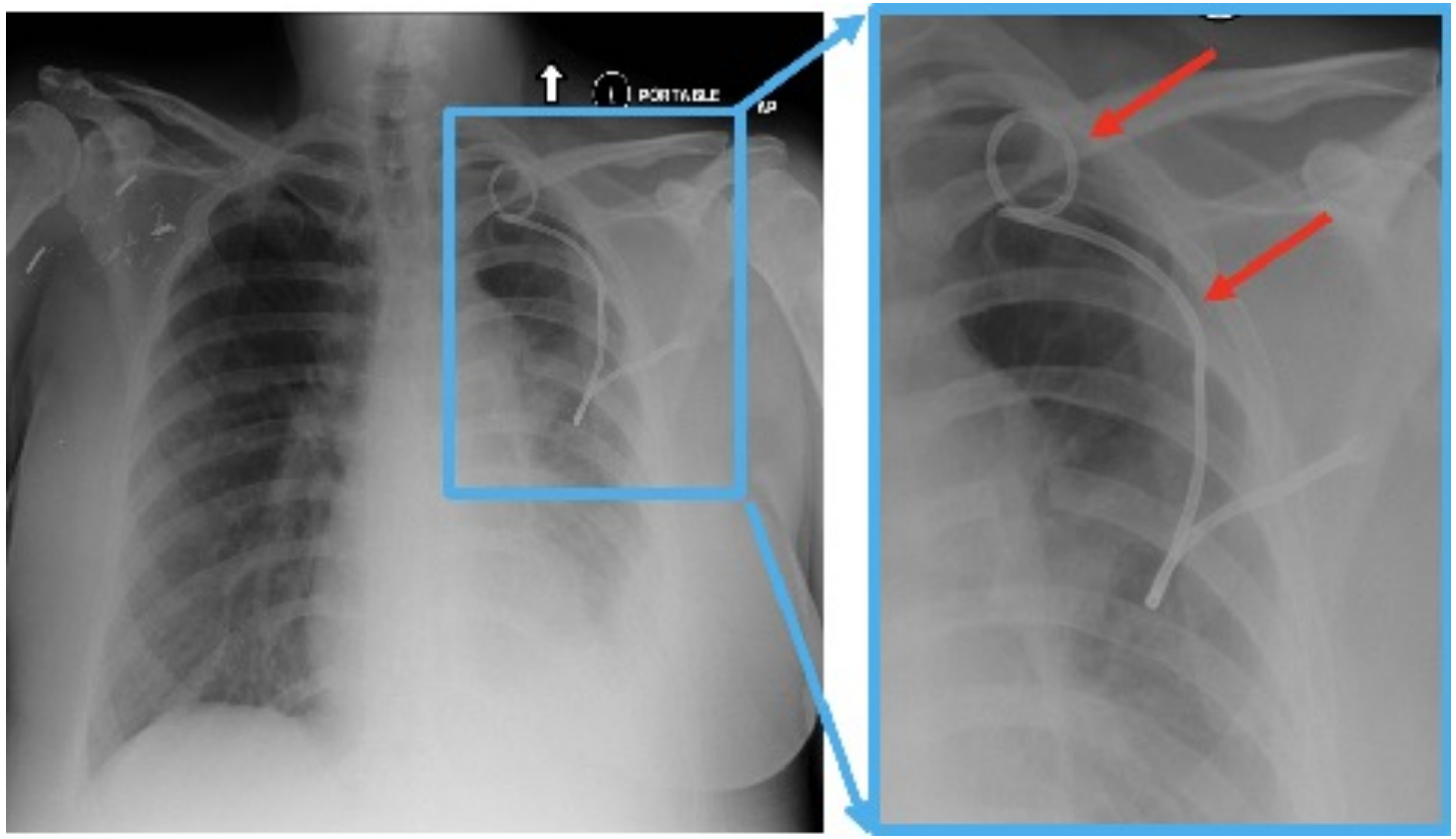
Figure 1. Convolutional Neural Network (CNN) Classification and Melanoma Probability Scores for Dermoscopic Images of Unmarked, Marked, and Cropped Benign Nevus and Melanoma



[Winkler et al]

Datasets - Shortcuts

- Chest drain associated with a collapsed lung
- AUC 0.94 vs 0.77



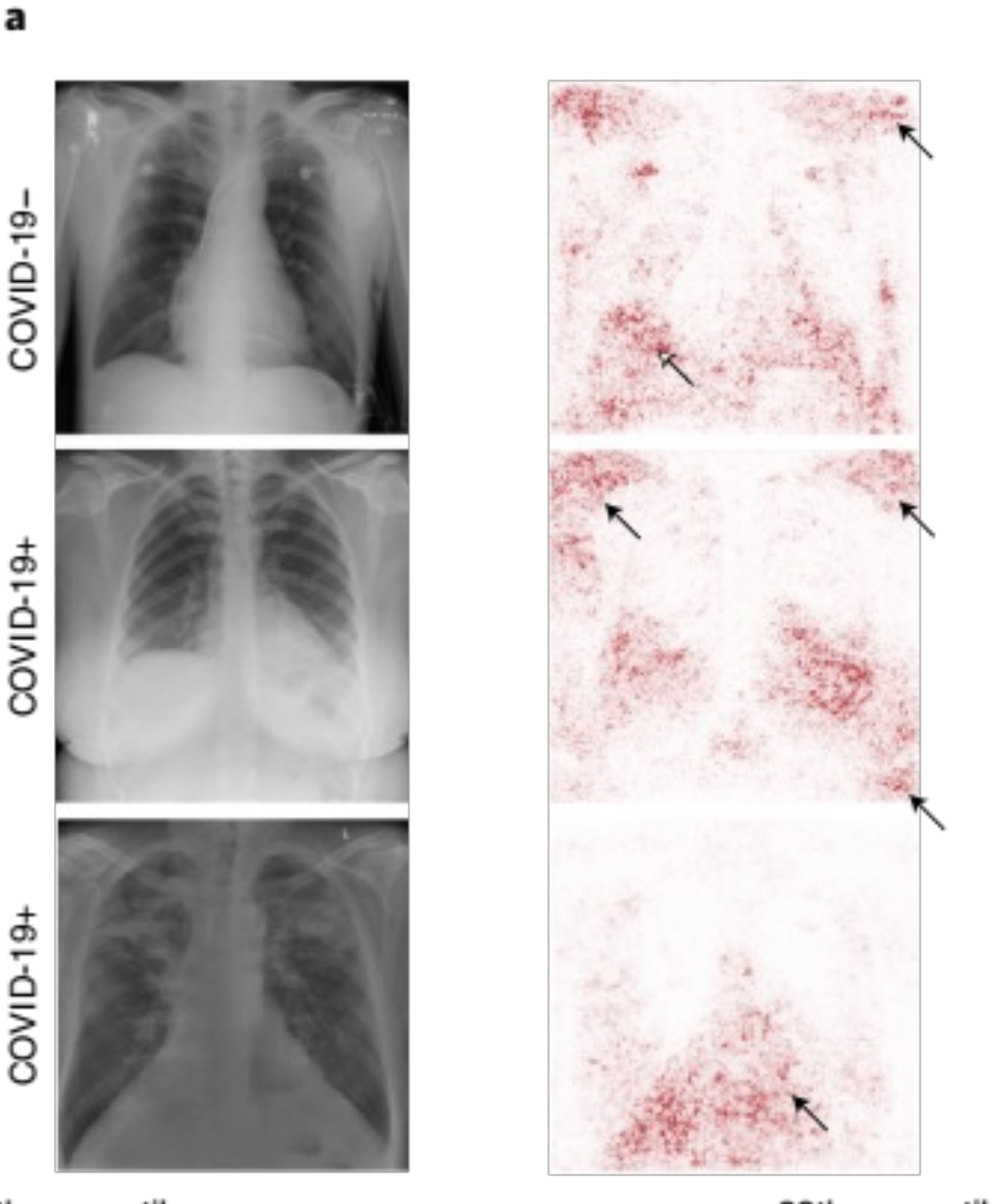
[Oakden-Rayner et al 2019]

[Image from Graf et al 2020]

Datasets - Shortcuts

- COVID associated with text markers (+patient position?)

[DeGrave et al 2021]



Our samples are not always representative

Methods → Problems ↓	1	2	3	4	5	…
Challenge #1 on lung cancer	✓✓	✓				
Dataset #2 on lung cancer	✓		✓✓			
…		✓✓	✓			
			✓	✓✓		
Early diagnosis of lung cancer	?	?	?	?	?	

Problem 2: publication is a bad proxy for “quality”

- Publications incentivize novelty & state-of-the-art results
- “Mathiness”, methods may be needlessly complex and fail to identify sources of gains [Lipton and Steinhardt 2019]

Proof by intimidation Trivial!

Proof by cumbersome notation The theorem follows immediately from the fact that $\left|\bigoplus_{k\in S}\left(\mathfrak{K}^{\mathbb{F}^{\alpha}(i)}\right)_{i\in\mathcal{U}_k}\right| \preccurlyeq \aleph_1$ when $[\mathfrak{H}]_{\mathcal{W}} \cap \mathbb{F}^{\alpha}(\mathbb{N}) \neq \emptyset$.

Proof by inaccessible literature The theorem is an easy corollary of a result proven in a hand-written note handed out during a lecture by the Yugoslavian Mathematical Society in 1973.

Proof by ghost reference The proof my be found on page 478 in a textbook which turns out to have 396 pages.

Circular argument Proposition 5.18 in [BL] is an easy corollary of Theorem 7.18 in [C], which is again based on Corollary 2.14 in [K]. This, on the other hand, is derived with reference to Proposition 5.18 in [BL].

Source

Publication - State-of-the-art results

Baselines too simple, or not simple enough

Focus on average accuracy (or similar), variability often not considered

Statistical significance:

- not used
- misunderstood
- <u>not practical significance</u>

@drveronikach

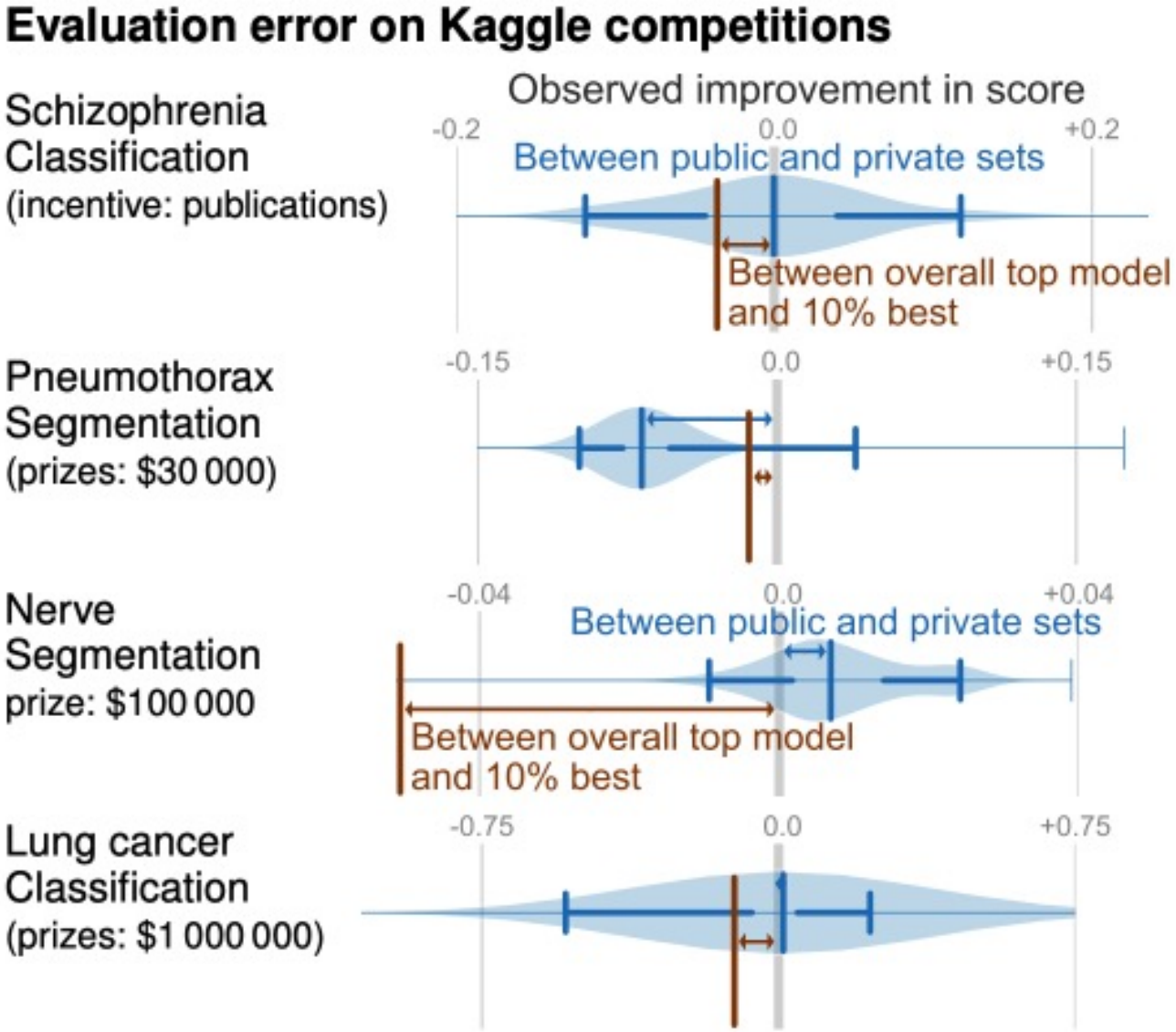
https://www.veronikach.com

“Practical significance”

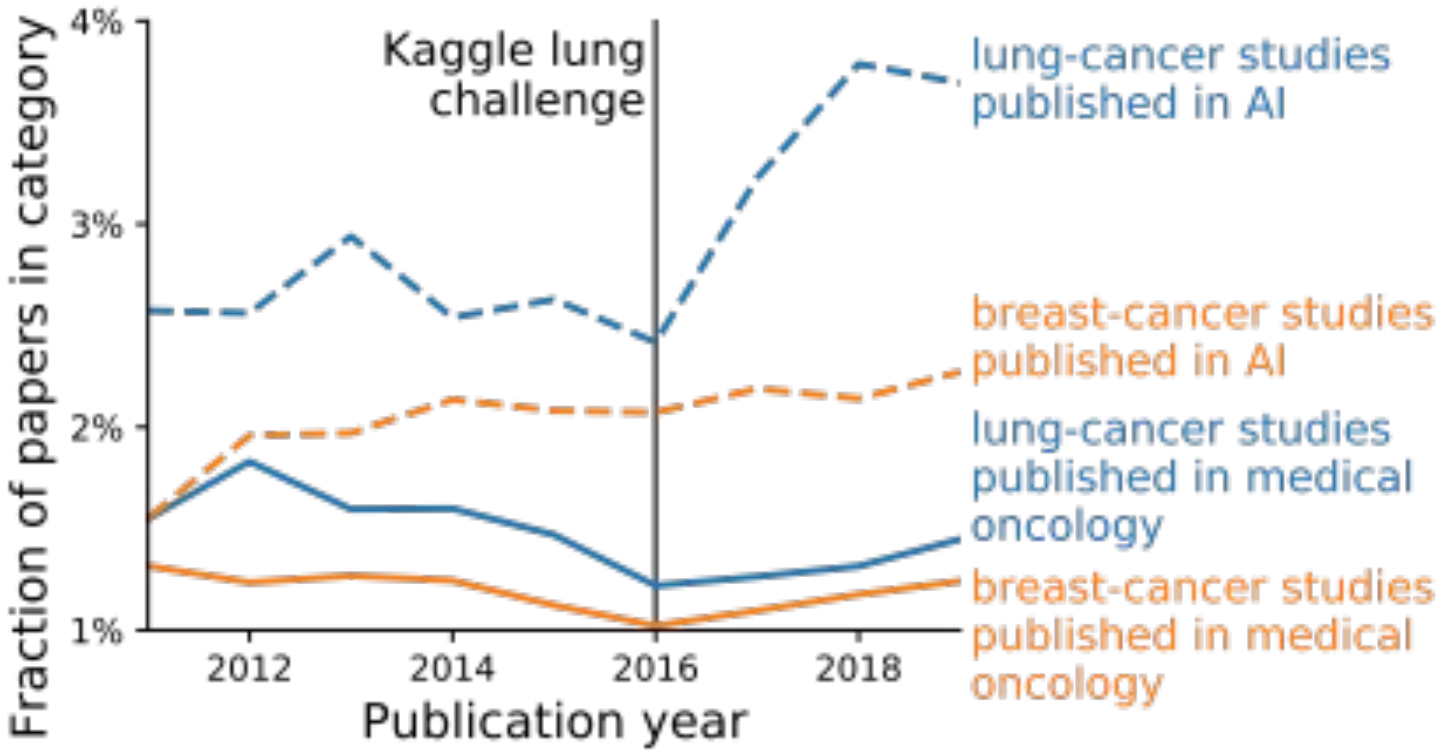
Evaluate methods on two independent sets, what differences do we expect? (in blue)

Difference between best and 10% best (in brown)

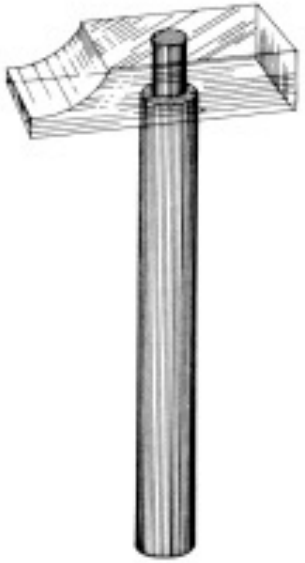
Gap often smaller than evaluation error!



Incentives change focus



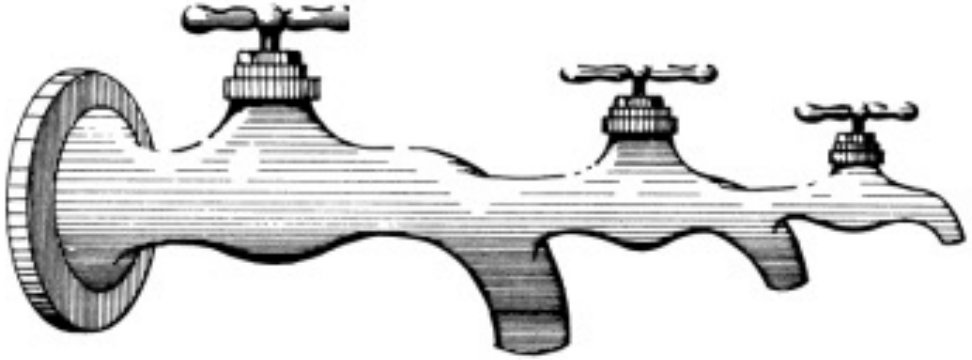
1,972 teams



A4 — Marteau à tête de verre. La fragilité de sa tête en fait l'outil idéal pour les travaux délicats.

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

A16 — Couteau de poche universel. Sous un même manche, façon corne, ce couteau réunit les objets les plus divers mais aussi les plus utiles tels que fourchette, plombeau, pipe, règle graduée, brosse à dents, révolver (indispensable aux campeurs, boy-scouts, etc..).

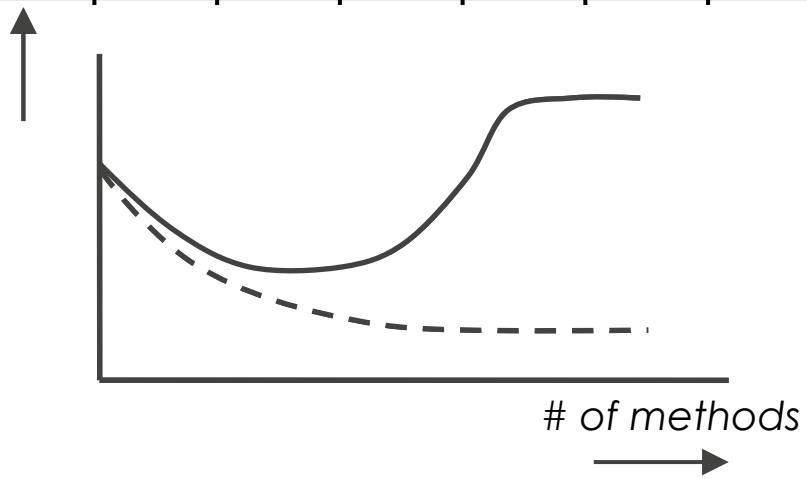


G3 — Robinet à débits différents. Économisez l'eau en choisissant le filet que vous voulez faire couler.

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

Evaluation is noisy and missing not-at-random

Methods → / Problems ↓	1	2	3	4	5													...
Challenge #1																		
Data #2																		
...																		
Early diagnosis	?	?	?	?	?													



Effects go beyond what's in the papers

- Carbon footprint

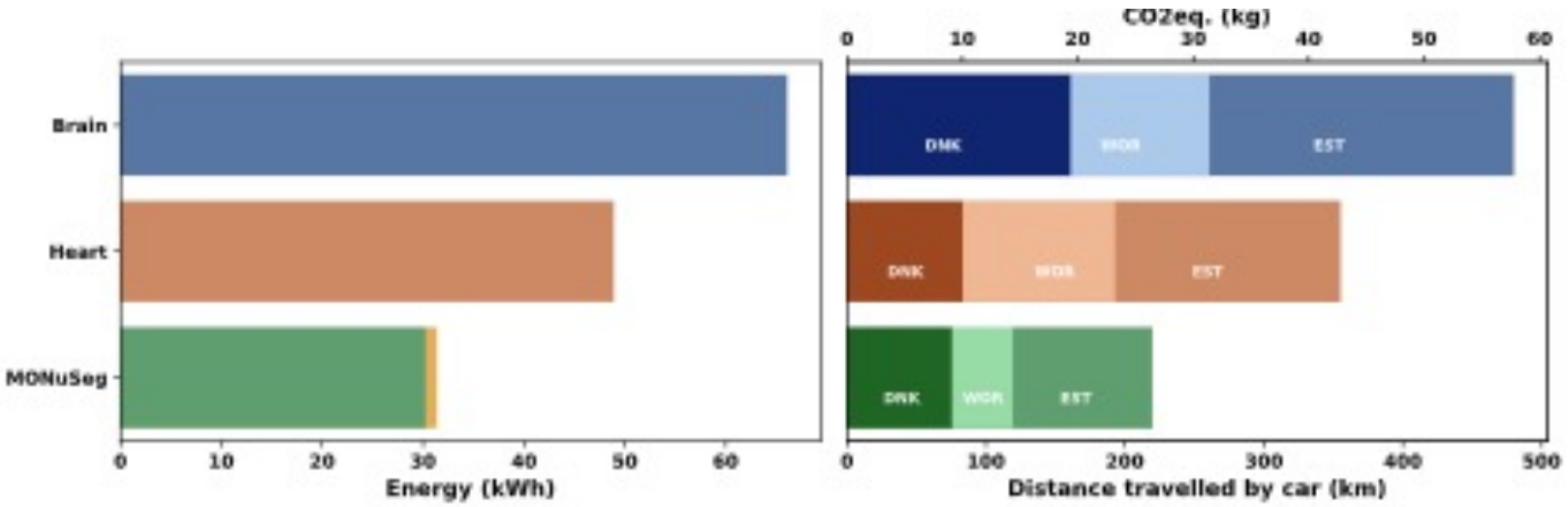


Fig. 3. (left) Total predicted energy consumed (kWh) over the five-fold cross validation for the three datasets using nnU-net[15]. For MONuSeg the predicted (orange) and the actual energy consumptions are shown, which are almost the same. (right) Carbon cost due to the training on the three datasets reported in CO2eq.(kg) and equivalent distance travelled by car (km). The carbon intensity and distance are also reported for three geographic regions (Denmark:DNK, Estonia: EST, Global: WOR) based on the regional average carbon intensities. All measurements were tracked/predicted using Carbontracker[3].

[Selvan et al 2022]

Effects go beyond what's in the papers

- What type of research is valued?
  - 100 top cited papers from ICML and NeurIPS → performance, novelty important, ethical considerations rarely considered [Birhane et al 2021]

- Who gets to do research?

Who gets to do research

Hardware lottery [Hooker 2020]: idea wins because of suitability of hardware/software.

De-democratization of AI [Ahmed and Wahed 2020]: 170K papers from 57 conferences "*... large firms and elite universities increased participation since 2012*"

[Birhane et al 2021] - big tech participation up from 11% to 58% in 10 years

MICCAI RISE statistics

Who gets to do research

"Grad student descent"
[Gencoglu et al 2019]

"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

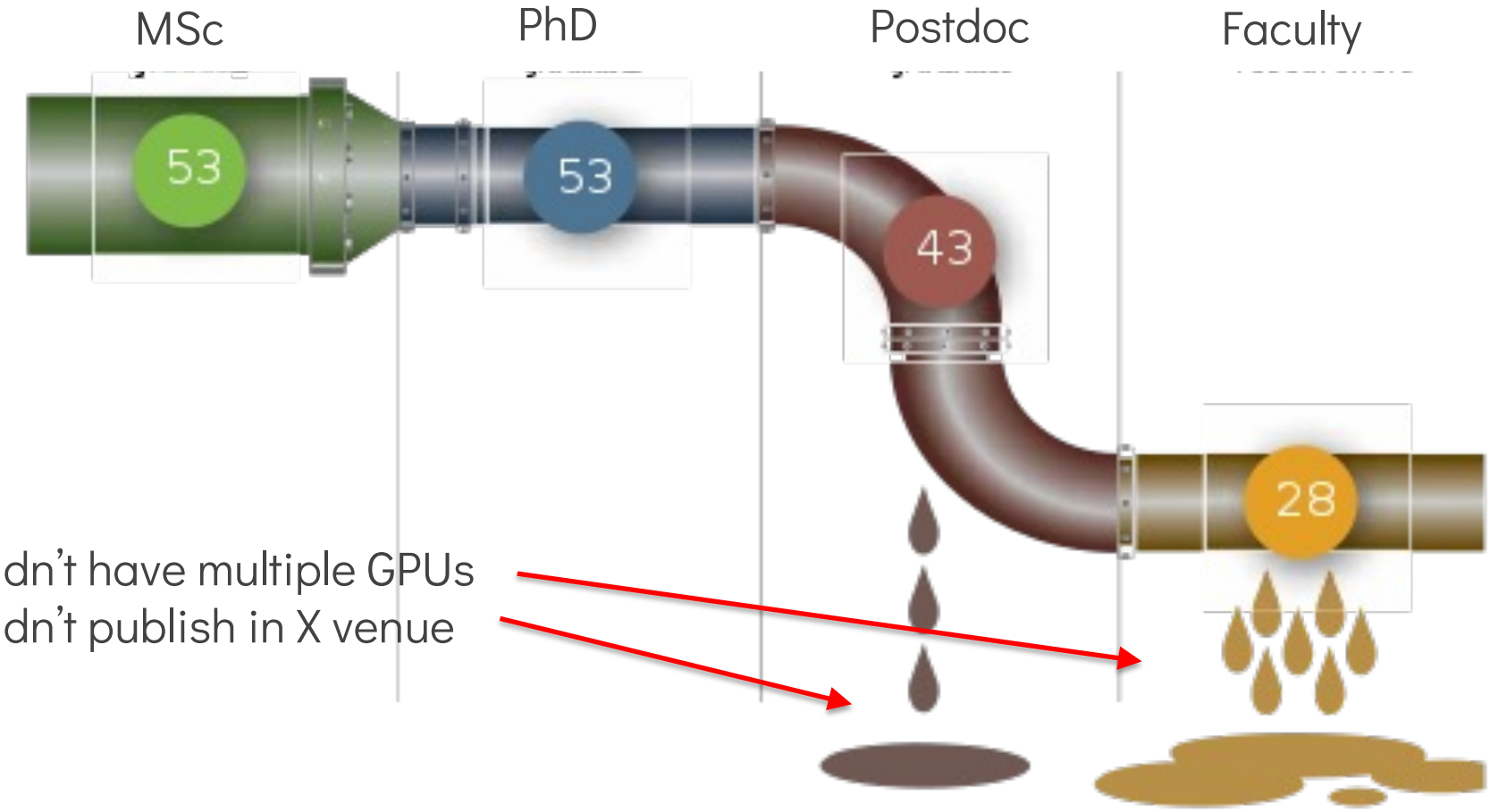
Mehmet Süzen
Jülich, Germany
suzen@acm.org

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

Who gets to do research



Didn't have multiple GPUs
Didn't publish in X venue

...

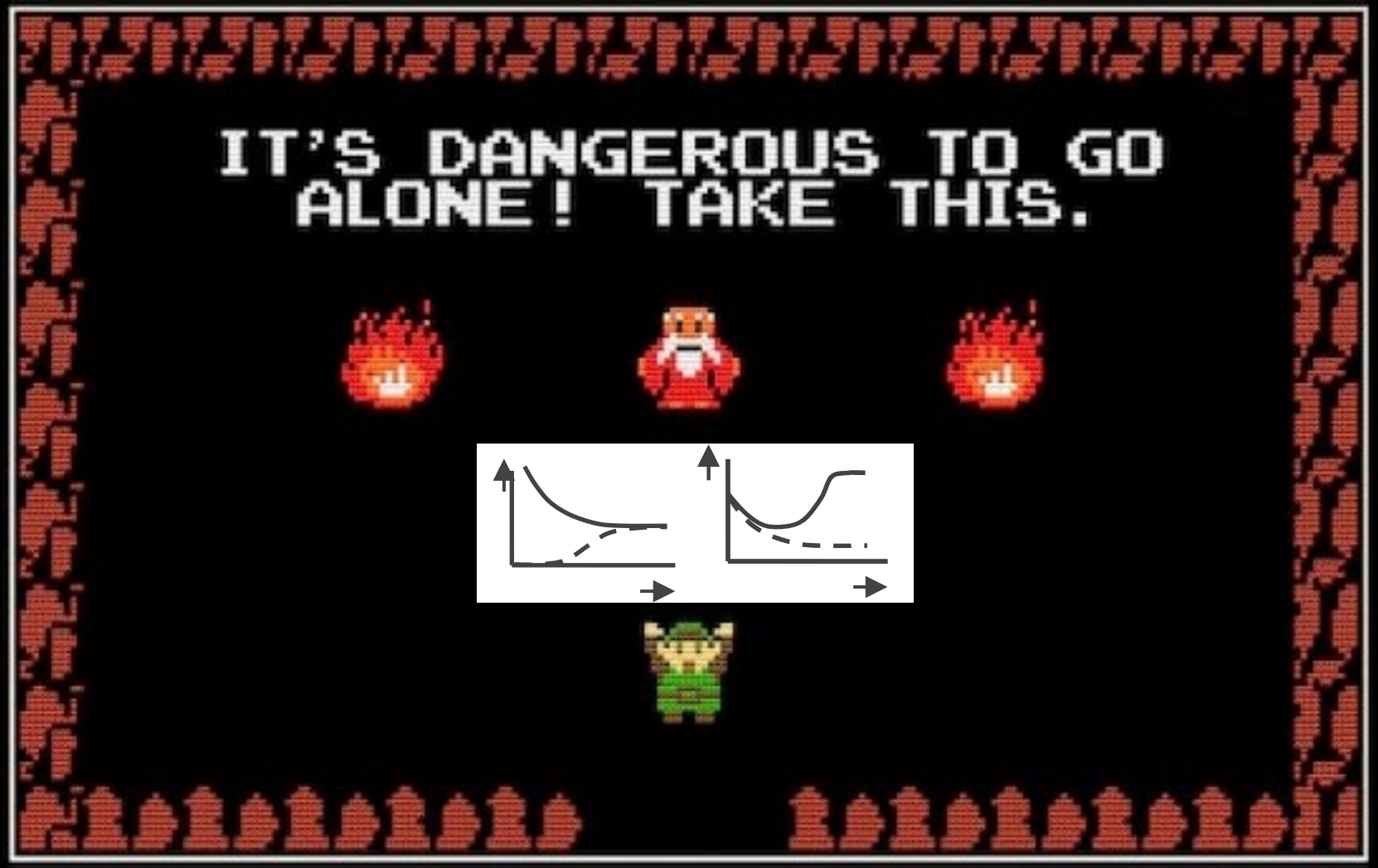
Source: UNESCO Institute for Statistics estimates based on data from its database, July 2015

[Original image]

Recommendations

@drveronikach

https://www.veronikach.com


IT'S DANGEROUS TO GO
ALONE! TAKE THIS.

Recommendations

Focus on datasets! Cite datasets, evaluate datasets

Be transparent about limitations (e.g. model cards [Mitchell et al 2018])

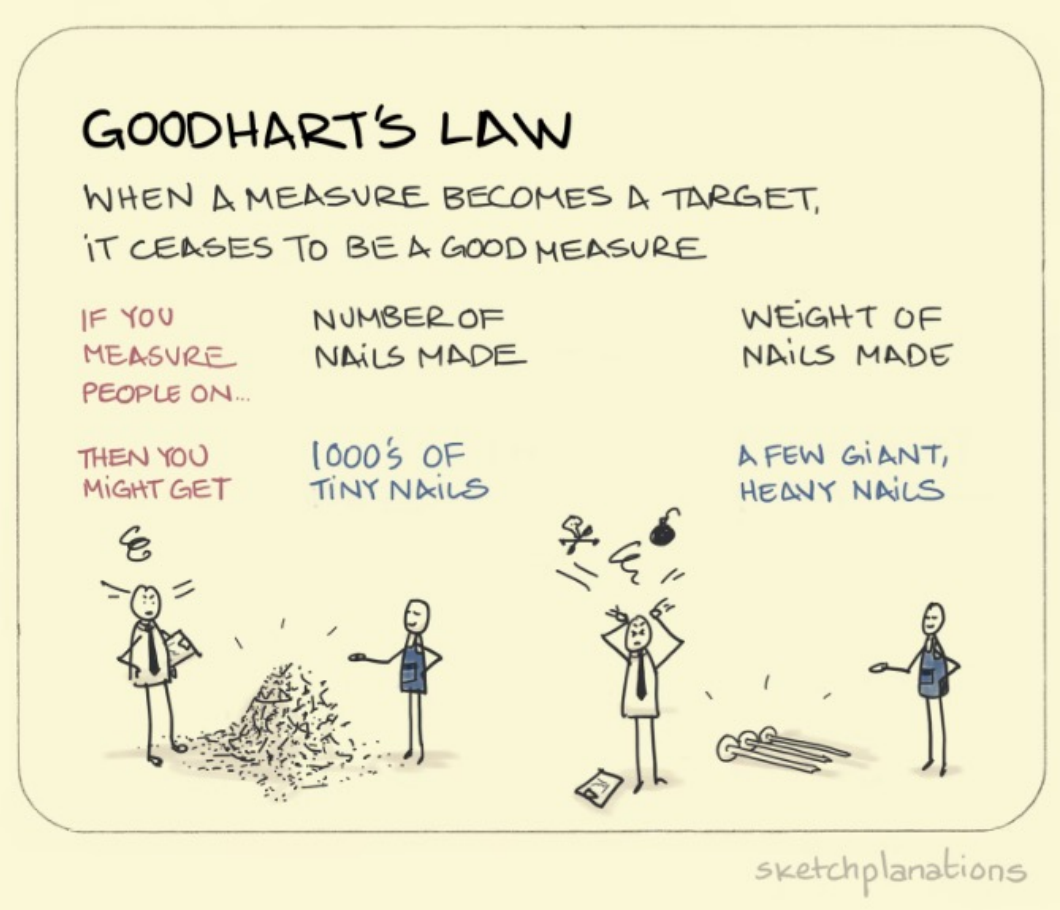
Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Recommendations

If you must compare methods...

- Representative data & baselines
- Beyond accuracy
  - "Practical significance"
  - Carbon footprint [Anthony et al 2020]
  - Qualitative accounts [Thomas & Uminsky 2020]

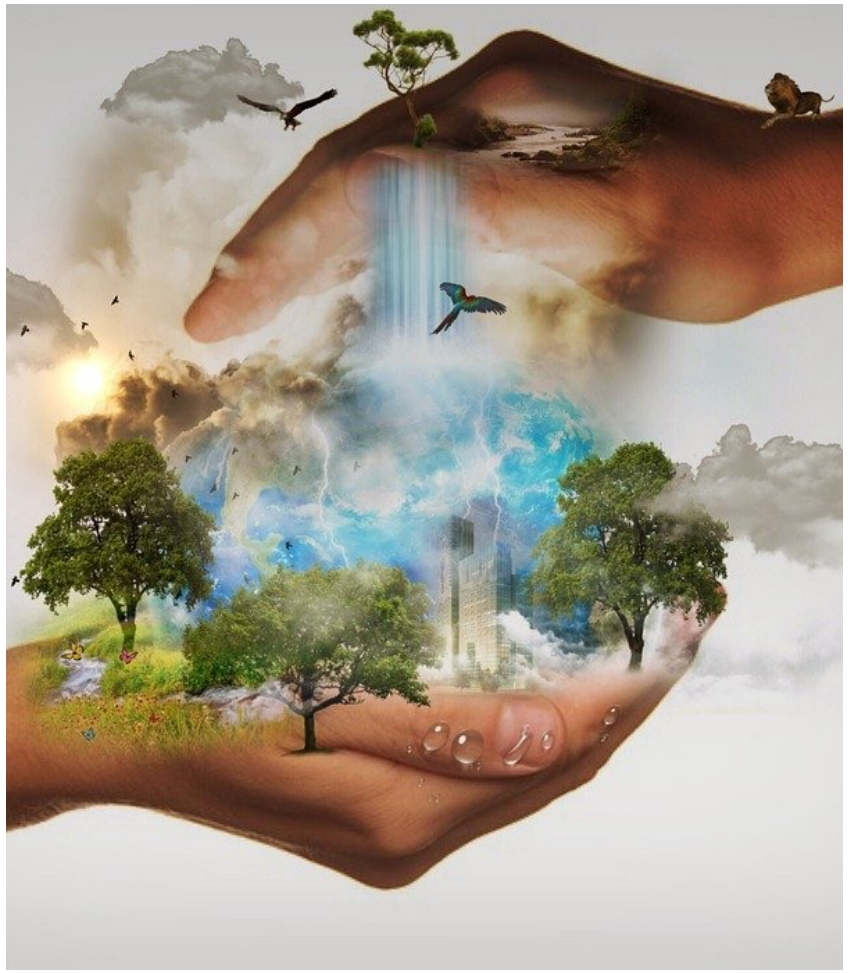


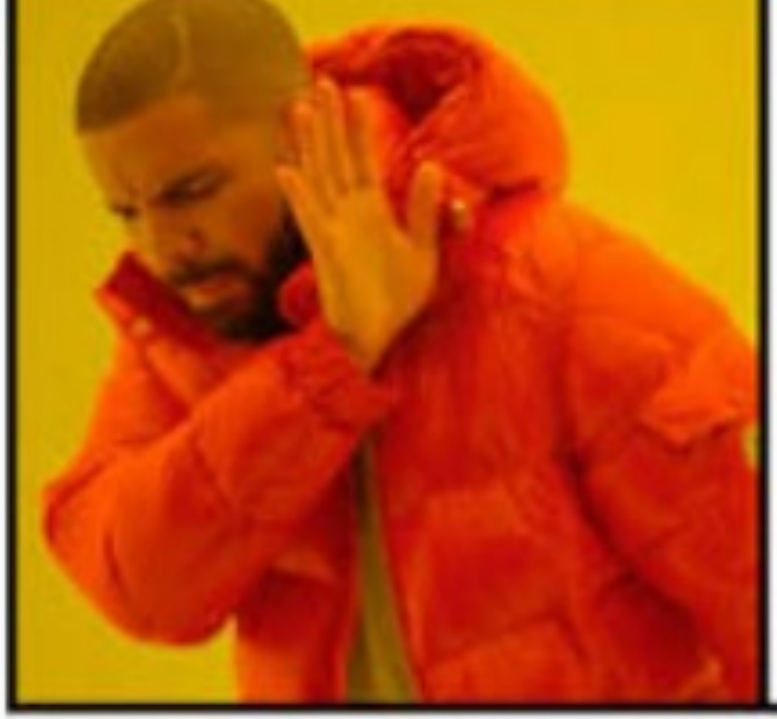
Recommendations

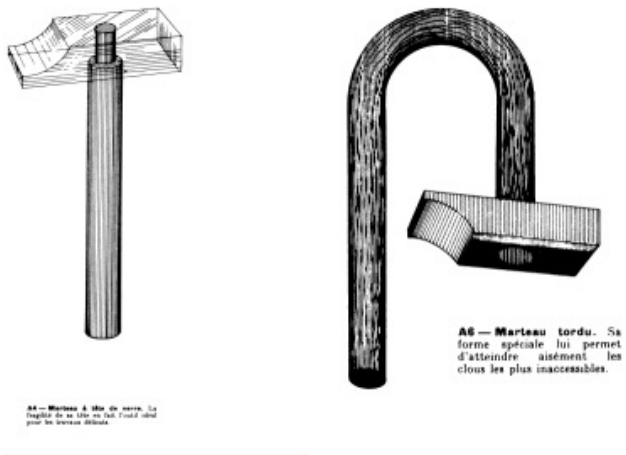
- Collaboration, not competition
  - Understanding
  - Save resources (FTEs, trees)

- Remember that scientists are humans

- Revisit values more often






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

Thank you!

@drveronikach

https://www.veronikach.com

Special thanks

References

Ahmed, N. & Wahed, M. (2020). The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *arXiv preprint arXiv:2010.15581.*

Anthony, L. F. W., Kanding, B. & Selvan, R.. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *arXiv preprint arXiv:2007.03051.*

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R. & Michelle Bao M. (2021). The Values Encoded in Machine Learning Research. *arXiv preprint arXiv:2106.15590.*

Brandt, I. V. D., Fok, F., Mulders, B., Vanschoren, J., & Cheplygina, V. (2021). Cats, not CAT scans: a study of dataset similarity in transfer learning for 2D medical image classification. *arXiv preprint arXiv:2107.05940*.

Cheplygina, V. (2019). Cats or CAT scans: Transfer learning from natural or medical image source data sets?. *Current Opinion in Biomedical Engineering*, *9*, 21-27.

References

Cheplygina, V., de Bruijne, M., & Pluim, J. P. W. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis 54, 280-296.*

DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, *3*(7), 610-619.

Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., Leinonen, J., & Huttunen, H. (2019). HARK Side of Deep Learning -- From Grad Student Descent to Automated Machine Learning. arXiv preprint arXiv:1904.07633.

References

Hooker, S. (2020). The Hardware Lottery. *arXiv preprint arXiv:2009.06489.*

Kelly, B. S., Judge, C., Bollard, S. M., Clifford, S. M., Healy, G. M., Aziz, A., ... & Killeen, R. P. (2022). Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *European Radiology*, 1-10.

Lipton, C. Z. & Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *ACM Queue 17, 1.*

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *arXiv preprint arXiv:1810.03993.*

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning.*

References

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.*

Raumanns, R., Schouten, G., Joosten, M., Pluim, J.P.W., & Cheplygina, V. (2021). ENHANCE (ENrichting Health data by ANnotations of Crowd and Experts): A case study for skin lesion classification. *MELBA December 2021.* Publisher (open access) | Github

Selvan, R. et al. (2022) Carbon Footprint of Selecting and Training Deep Learning Models for Medical Image Analysis. *arXiv preprint arXiv:2203:02202v1*

Thomas, R. & Uminsky, D. (2020). The Problem with Metrics is a Fundamental Problem for AI. *arXiv preprint arXiv:2002.08512.*

References

Varoquaux, G. & Cheplygina, V. (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digital Medicine 5 (1), 1-8.*

Winkler JK, Fink C, Toberer F, et al. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol. 2019;155(10):1135–1141.*

Ørting S., Doyle A., van Hilten A., Hirth M., Inel O, Madan C. R., Mavridis P., Spiers H. & Cheplygina V. (2020). A Survey of Crowdsourcing in Medical Image Analysis. *Human Computation 7(1), 1-26*.