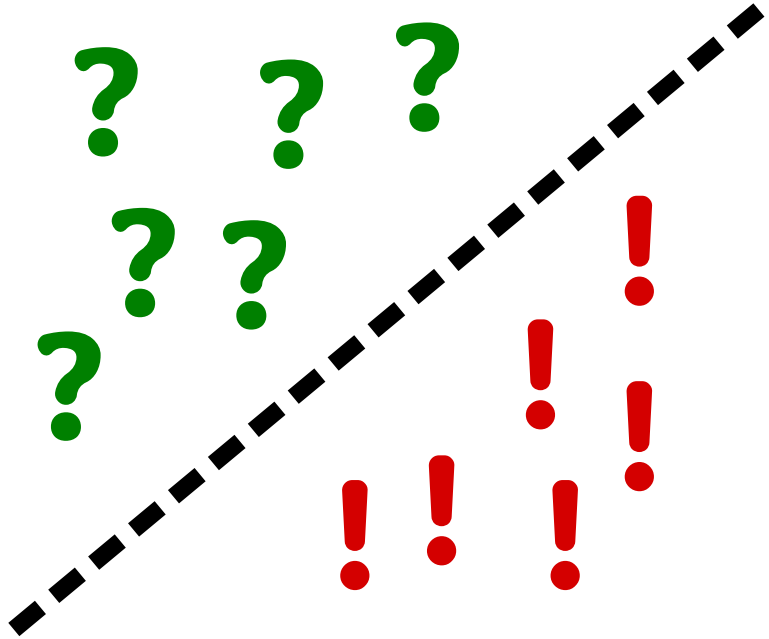
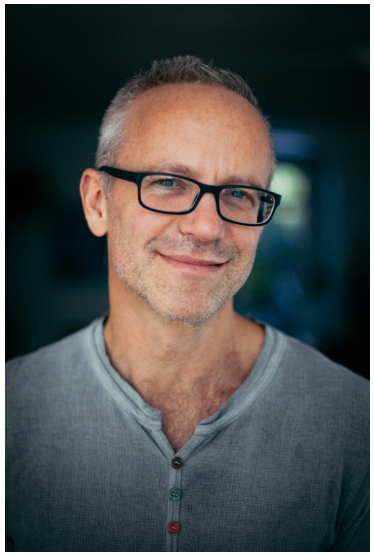
AI benchmarking for hypothesis-driven science in machine and deep learning?

MediaEval: Multimedia Evaluation Benchmark
Oct 25, 2025



Jan van Gemert

Whoami: Jan van Gemert

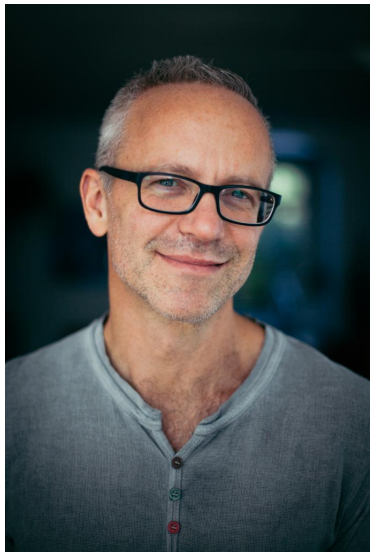


Computer Vision lab @ TU Delft

Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. Find & evaluate powerful yet flexible physical priors for data-efficient visual recognition AI.

Whoami: Jan van Gemert



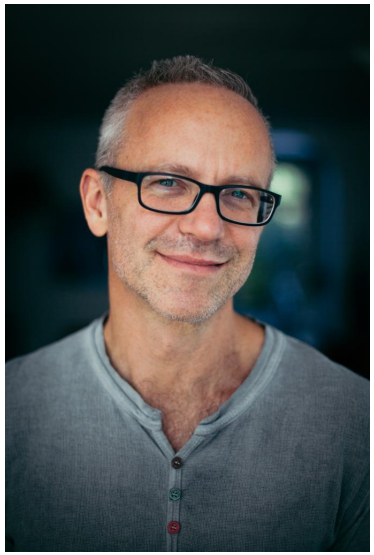
Computer Vision lab @ TU Delft

Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. Find & evaluate powerful yet flexible physical priors for data-efficient visual recognition AI.

AI benchmarking for hypothesis-driven science in machine and deep learning?

Whoami: Jan van Gemert



Computer Vision lab @ TU Delft

Two main research themes:

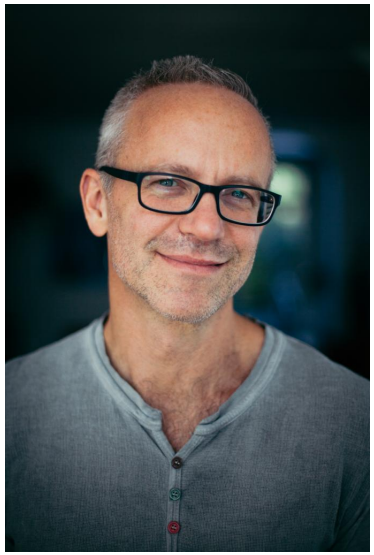
1. Fundamental empirical understanding-based deep learning research; (to)
2. Find & evaluate powerful yet flexible physical priors for data-efficient visual recognition AI.

AI benchmarking for hypothesis-driven science in machine and deep learning?

AI benchmarking...

Quantify how well an automatic system (AI) can perform a task.

Whoami: Jan van Gemert



Computer Vision lab @ TU Delft

Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. Find & evaluate powerful yet flexible physical priors for data-efficient visual recognition AI.

AI benchmarking for hypothesis-driven science in machine and deep learning?

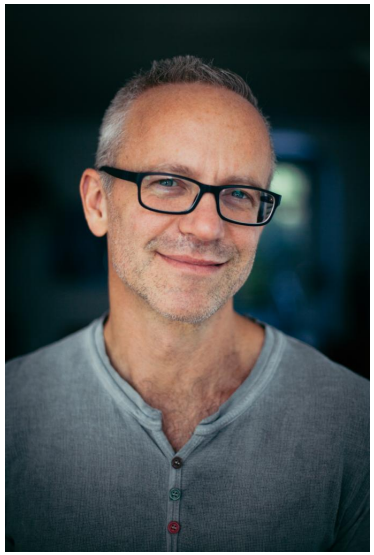
AI benchmarking...

Quantify how well an automatic system (AI) can perform a task.

hypothesis-driven science in machine and deep learning...

Doing science: better understand machine and deep learning methods.

Whoami: Jan van Gemert



Computer Vision lab @ TU Delft

Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. Find & evaluate powerful yet flexible physical priors for data-efficient visual recognition AI.

AI benchmarking for hypothesis-driven science in machine and deep learning?

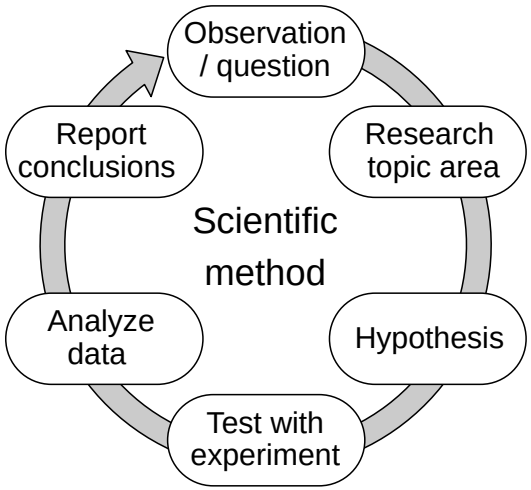
AI benchmarking...
Quantify how well an automatic system (AI) can perform a task.

hypothesis-driven science in machine and deep learning...
Doing science: better understand machine and deep learning methods.

"?"
My own questions about benchmarking and ML/DL science :)

The scientific method[1] in times of deep learning

Deep learning powers AI; yet as a scientific field has growing pains[2,3,4]



- Improvement-driven (large compute/data);
- Trial and error (graduate student descent)
- Opportunistic (career driven);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation
- Not identifying the reasons for empirical gains.

[1]: https://en.wikipedia.org/wiki/Scientific_method

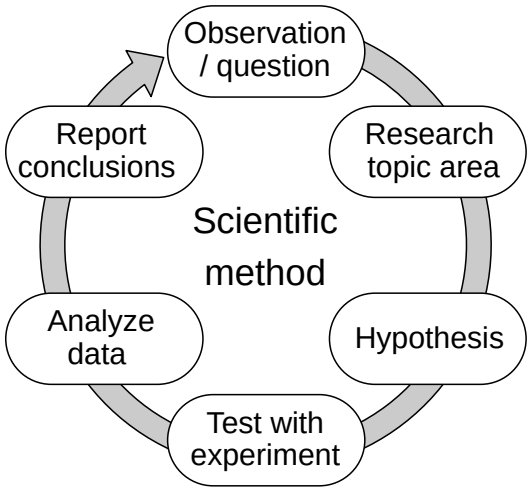
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.

[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).

The scientific method[1] in times of deep learning

Deep learning powers AI; yet as a scientific field has growing pains[2,3,4]



- Improvement-driven (large compute/data);
- Trial and error (graduate student descent)
- Opportunistic (career driven);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation
- Not identifying the reasons for empirical gains.

- ML/DL does not have many empirical theories.

[1]: https://en.wikipedia.org/wiki/Scientific_method

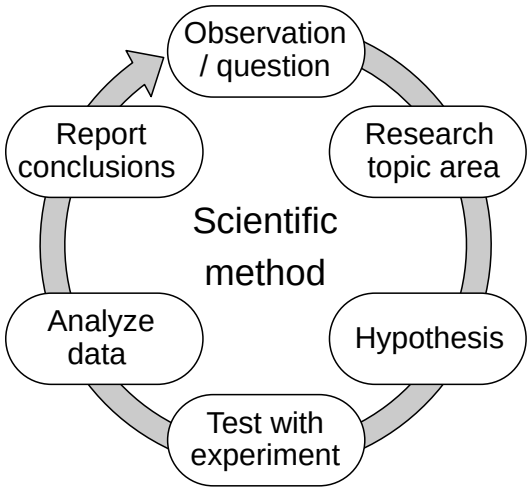
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.

[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).

The scientific method[1] in times of deep learning

Deep learning powers AI; yet as a scientific field has growing pains[2,3,4]



- Improvement-driven (large compute/data);
- Trial and error (graduate student descent)
- Opportunistic (career driven);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation
- Not identifying the reasons for empirical gains.

- ML/DL does not have many empirical theories. Some that I am aware of:
  ◦ Neural Scaling Laws; ◦ Bias/variance ◦ ML is like physics/neuroscience;
  ◦ Simple axioms explaining intelligence ◦ Different media represent the same reality ◦ ...

[1]: https://en.wikipedia.org/wiki/Scientific_method

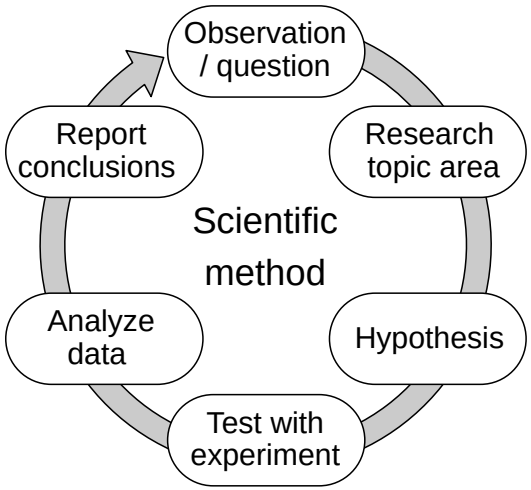
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.

[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).

The scientific method[1] in times of deep learning

Deep learning powers AI; yet as a scientific field has growing pains[2,3,4]



- Improvement-driven (large compute/data);
- Trial and error (graduate student descent)
- Opportunistic (career driven);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation
- Not identifying the reasons for empirical gains.

- ML/DL does not have many empirical theories. Some that I am aware of: ◦ Neural Scaling Laws; ◦ Bias/variance ◦ ML is like physics/neuroscience; ◦ Simple axioms explaining intelligence ◦ Different media represent the same reality ◦ ...
- Mores in the field: End-to-end learning; 'bold' numbers on common datasets; trail and error; openly sharing code/weights/data; all papers open on ArXiv.

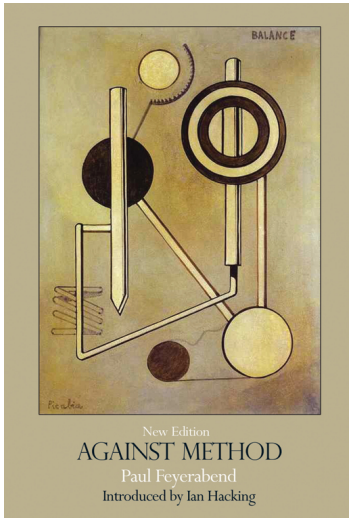
[1]: https://en.wikipedia.org/wiki/Scientific_method

[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.

[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).

Against method: "*The* Way" vs "A Way"

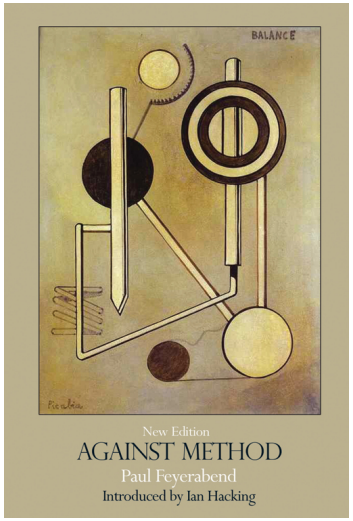


- There is not "one way" to do science. Science moves on, *despite* the methodology used[5]. Quote: "*Machine learning theory needs a reformation, because our advice is not just ignored, but demonstrably, actively harmful.*" [6]

[5]: Paul Feyerabend; "Against method", 1975

[6]: Ben Recht; `https://www.argmin.net/p/desirable-sins`, 2025

Against method: "*The* Way" vs "A Way"

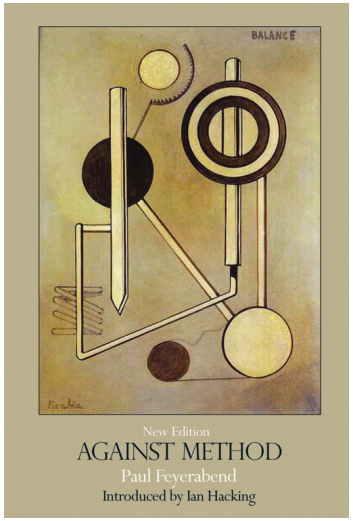


- There is not "one way" to do science. Science moves on, *despite* the methodology used[5]. Quote: "*Machine learning theory needs a reformation, because our advice is not just ignored, but demonstrably, actively harmful.*" [6]
- I'm in ML/DL science because I like the intellectual pursuit, not 'solve AI'. So, why should I tell 'system builders' they need to stop what they like?

[5]: Paul Feyerabend; "Against method", 1975

[6]: Ben Recht; `https://www.argmin.net/p/desirable-sins`, 2025

Against method: "*The* Way" vs "A Way"



- There is not "one way" to do science. Science moves on, *despite* the methodology used[5]. Quote: "*Machine learning theory needs a reformation, because our advice is not just ignored, but demonstrably, actively harmful.*" [6]
- I'm in ML/DL science because I like the intellectual pursuit, not 'solve AI'. So, why should I tell 'system builders' they need to stop what they like?

Let people do research however they want (including yourself).

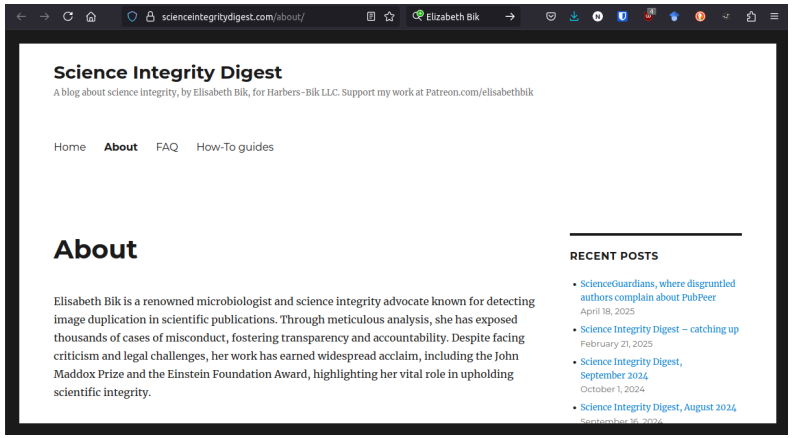
[5]: Paul Feyerabend; "Against method", 1975

[6]: Ben Recht; `https://www.argmin.net/p/desirable-sins`, 2025

My hero: Dr. Elizabeth Bik, science sleuth

1: ML misconduct: tune on the testset; cherry picking; plagiarism, overclaiming; isn't as bad as the explict manipulation as done here.

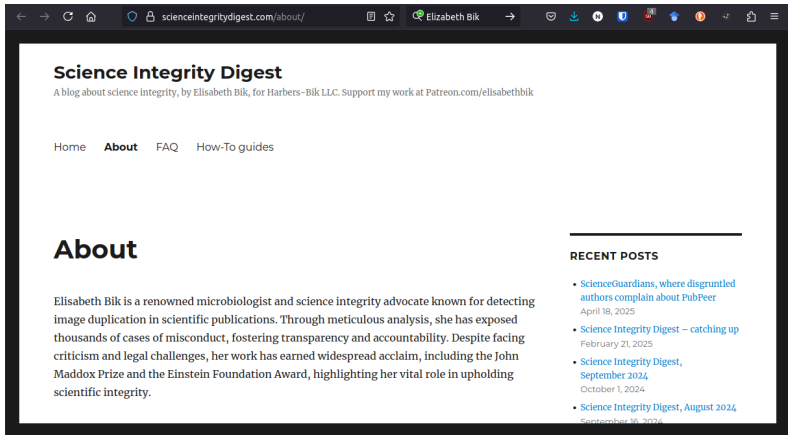
My hero: Dr. Elizabeth Bik, science sleuth



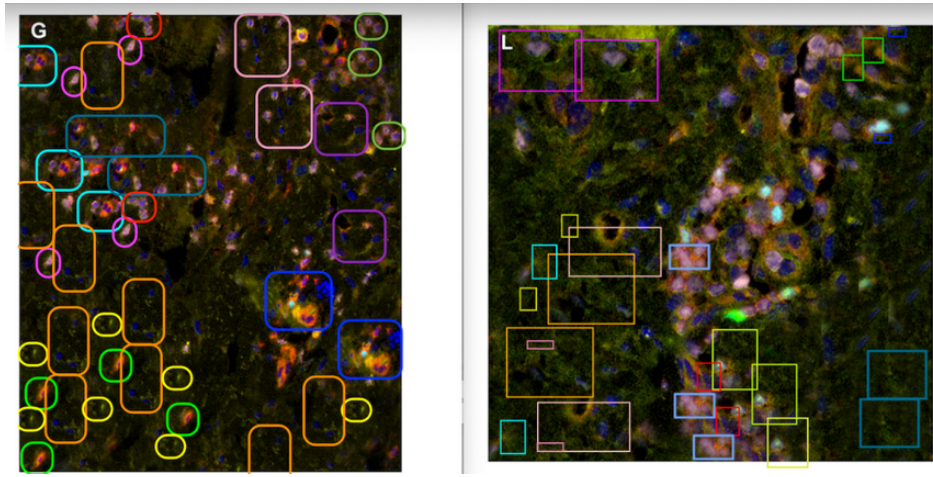
scienceintegritydigest.com

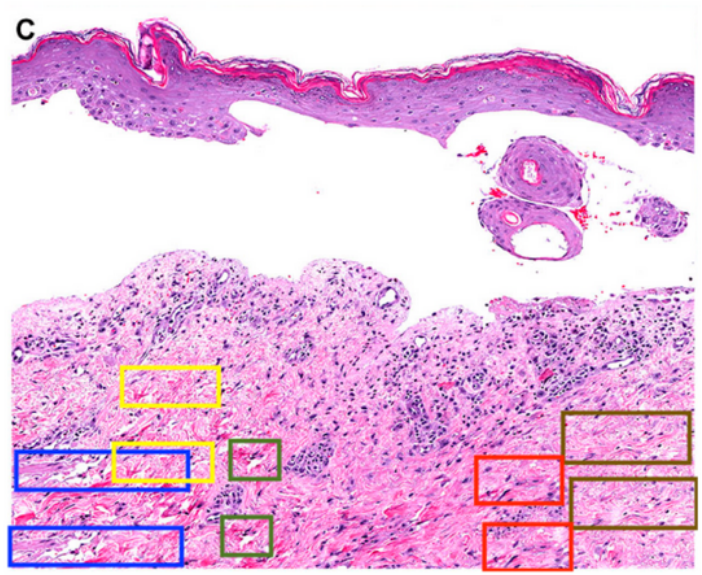
1: ML misconduct: tune on the testset; cherry picking; plagiarism, overclaiming; isn't as bad as the explict manipulation as done here.

My hero: Dr. Elizabeth Bik, science sleuth



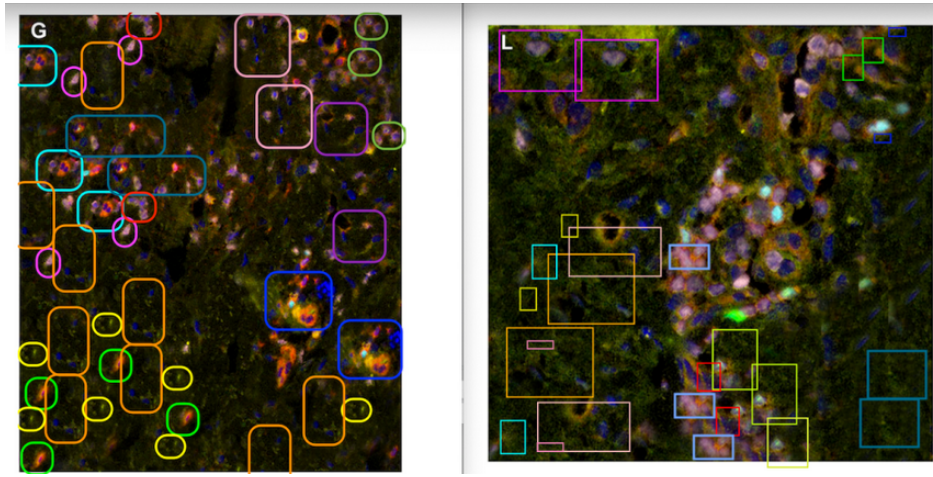
scienceintegritydigest.com

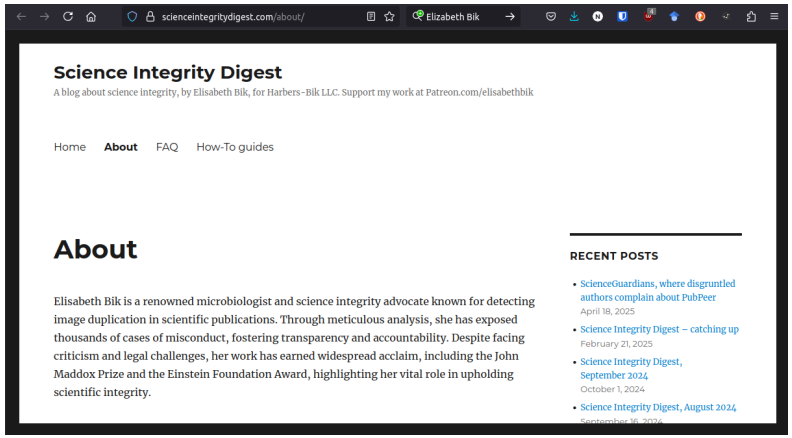


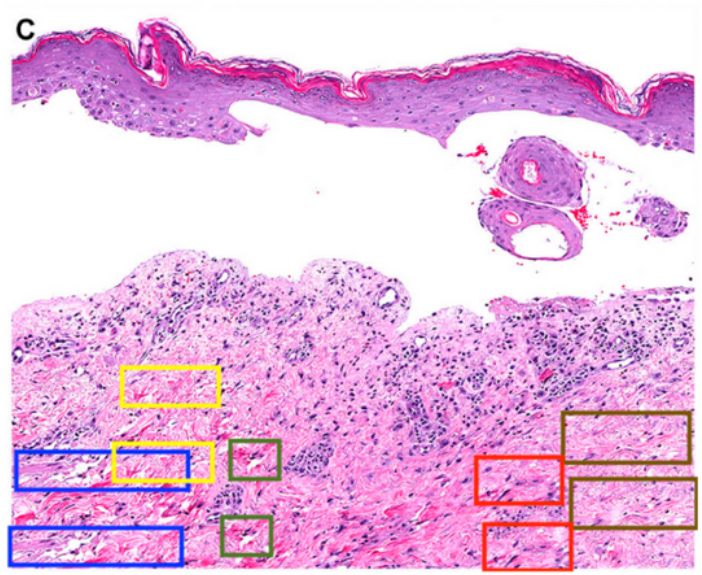


1: ML misconduct: tune on the testset; cherry picking; plagiarism, overclaiming; isn't as bad as the explict manipulation as done here.

My hero: Dr. Elizabeth Bik, science sleuth





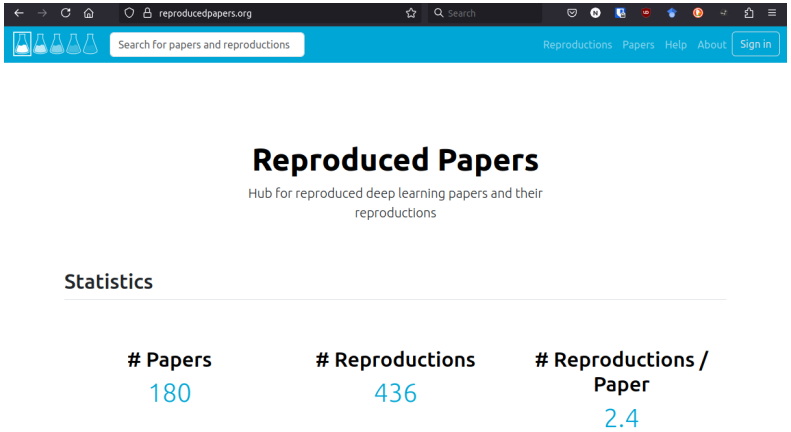


scienceintegritydigest.com

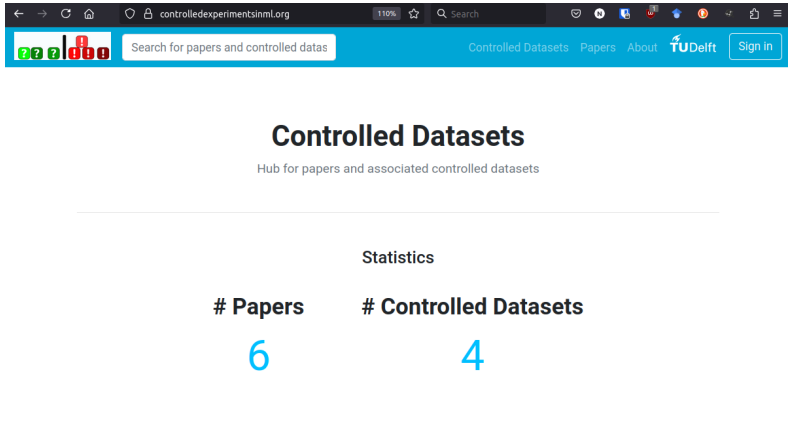
Doesn't (only) preach "Don't do fraud; it's bad"[1]; she does the work.

1: ML misconduct: tune on the testset; cherry picking; plagiarism, overclaiming; isn't as bad as the explict manipulation as done here.

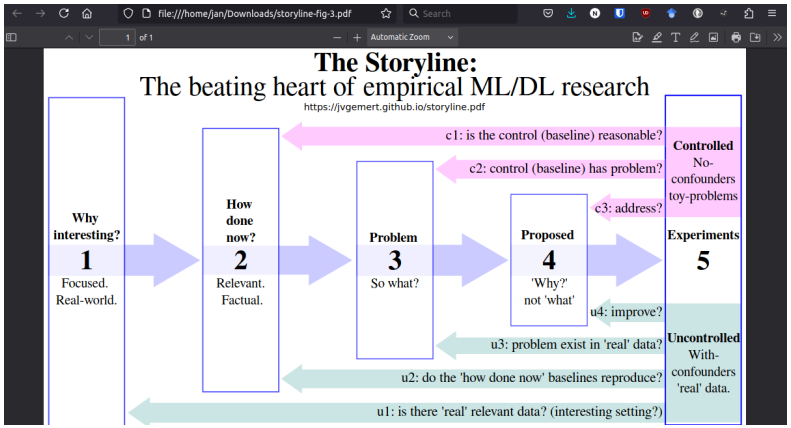
My work for fundamental empirical research in ML/DL



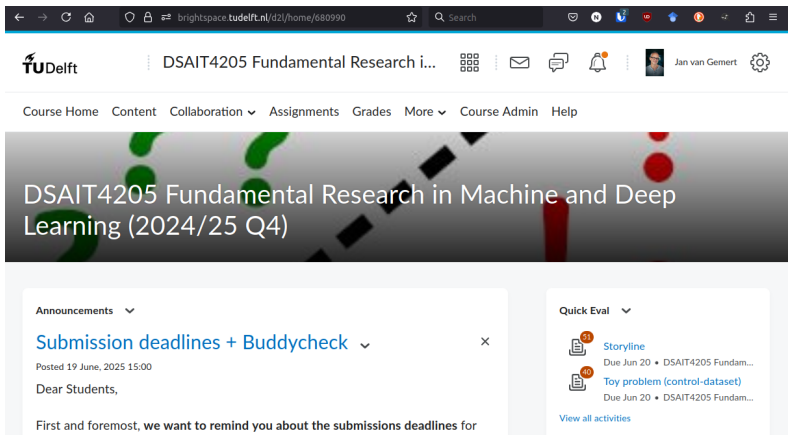
ReproducedPapers.org



ControlledExperimentsInML.org



Online research guidelines



MSc course

My work for fundamental empirical research in ML/DL

Recent initiative (Prof. Larson as a keynote :))

Metascience for Machine Learning

Holding a magnifying glass up to the ways of doing machine learning research.

Metascience for machine learning focuses on the science in the field of machine learning. It's about topics that are typically not found in Machine Learning text books, but about the ways of finding out what should be in those textbooks.

It's about how research in machine learning is done, the methodology, the processes, the mindset, goals, aspirations, inspirations.

AI is changing the world, and machine learning has proved a valuable tool for other important scientific fields. Here, we turn the tables, and put the spotlight on the scientific research field of machine learning itself.

`https://metascienceforml.github.io/`

My personal views on science in ML/DL

I don't believe:

- No single way to do science;
- No preaching; let system builders build systems.

My personal views on science in ML/DL

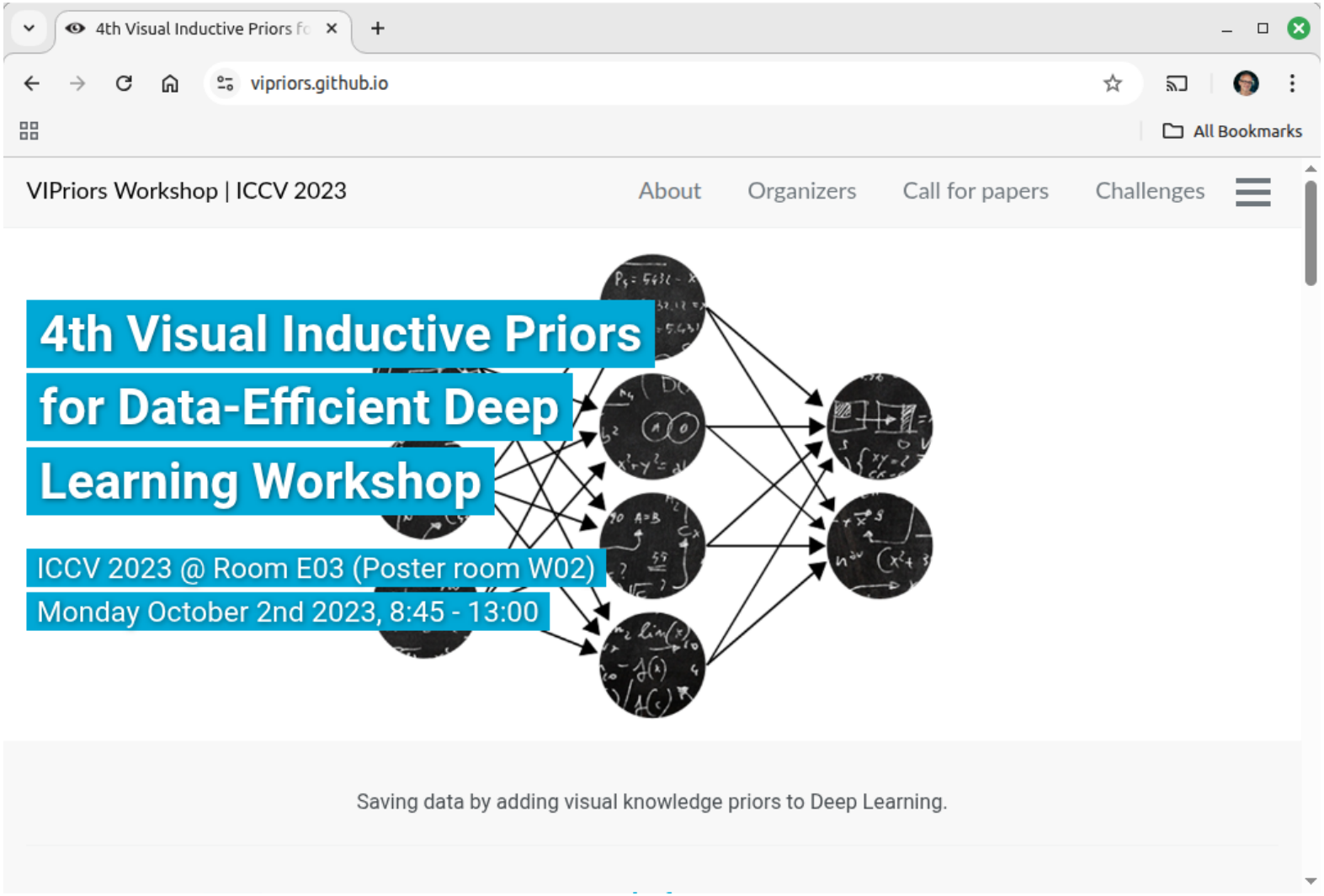
I don't believe:

- No single way to do science;
- No preaching; let system builders build systems.

I believe:

- ML/DL work is open as a field, openly sharing code, weights, papers.
- ML/DL misconduct (tune on the testset; cherry picking; plagiarism, overclaiming) is not as bad as elsewhere; limited direct fraud.
- that the scientific method will correct things eventually.
- in "Be and let Be". Let others do research their own way.
- in *doing*: help the ones that want to be helped.
- in moving constructively forward, ie: Do Something: my methodological development: ReproducedPapers.org; ControlledExperimentsInML.org; metascienceforml.github.io, online research guidelines; MSc course, this workshop, etc... (?)

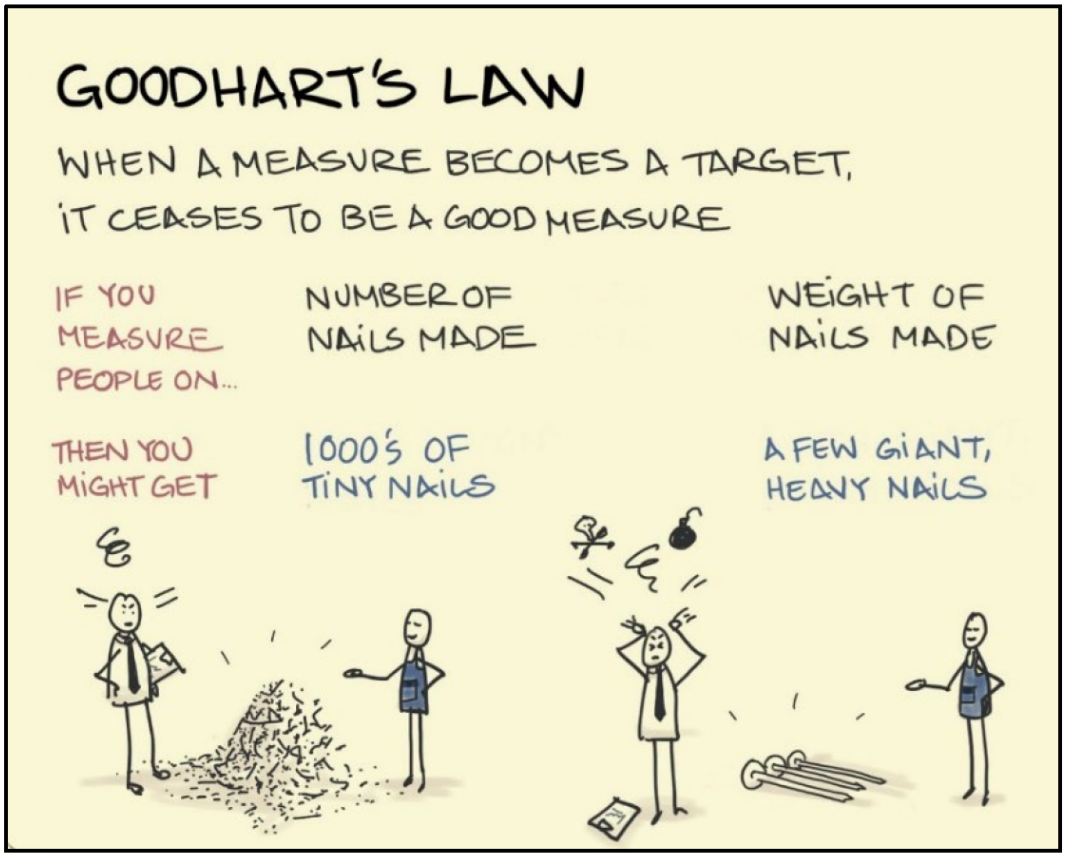
My limited experience in Benchmarking for Methodologies



https://vipriors.github.io/

Last slide: the question mark "?"

AI benchmarking for hypothesis-driven science in machine and deep learning **"?"**



(Link to image source)

- Benchmarks are invaluable to the field.
- How to use the power of benchmarks for hypothesis-driven science in ML/DL?