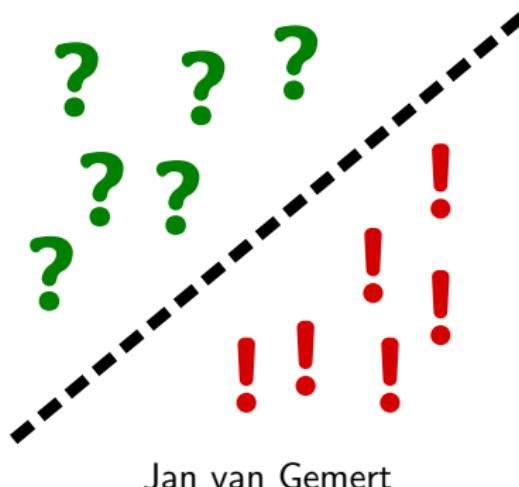


AI benchmarking for hypothesis-driven science in machine and deep learning?

MediaEval: Multimedia Evaluation Benchmark
Oct 25, 2025



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...

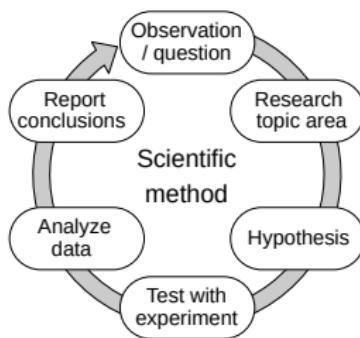
Doing science: better understand machine and deep learning methods.

"?"

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

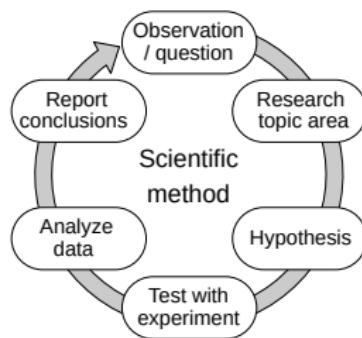
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[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

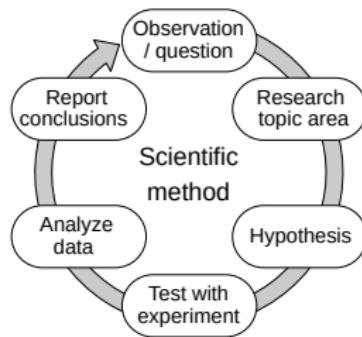
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[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

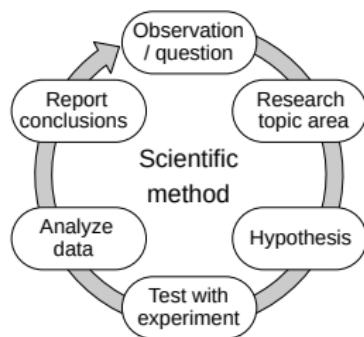
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[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; trial 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 explicit manipulation as done here.

My hero: Dr. Elizabeth Bik, science sleuth



A screenshot of the Science Integrity Digest website. The header reads "Science Integrity Digest" and "A blog about science integrity, by Elizabeth Bik, for Hobbes-Bik LLC. Support my work at Patreon.com/elizabethbik". Below the header are links for "Home", "About", "FAQ", and "How-To guides". The main content area has a section titled "About" which describes Elizabeth Bik's work. To the right is a sidebar titled "RECENT POSTS" with a list of five entries, each with a link and a date.

About

Elizabeth Bik is a renowned microbiologist and science integrity advocate known for detecting image duplication in scientific publications. Through meticulous analysis, she has exposed thousands of cases of misconduct, fostering transparency and accountability. Despite facing criticism and legal challenges, her work has earned widespread acclaim, including the John Maddox Prize and the Einstein Foundation Award, highlighting her vital role in upholding scientific integrity.

RECENT POSTS

- [ScienceIntegrity, where I diagnosed and then complained about Pohleová](#)
April 18, 2023
- [When integrity agents = catching up](#)
February 21, 2023
- [Science Integrity Digest, September 2022](#)
October 1, 2022
- [Science Integrity Digest, August 2022](#)

scienceintegritydigest.com

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

My hero: Dr. Elizabeth Bik, science sleuth



A screenshot of the Science Integrity Digest website. The header reads "Science Integrity Digest" and "A blog about science integrity, by Elizabeth Bik, for Hobbes-Bik LLC. Support my work at Patreon.com/elizabethbik". Below the header are links for "Home", "About", "FAQ", and "How-To guides". The "About" section features a bio for Elizabeth Bik and a "RECENT POSTS" sidebar with links to various posts.

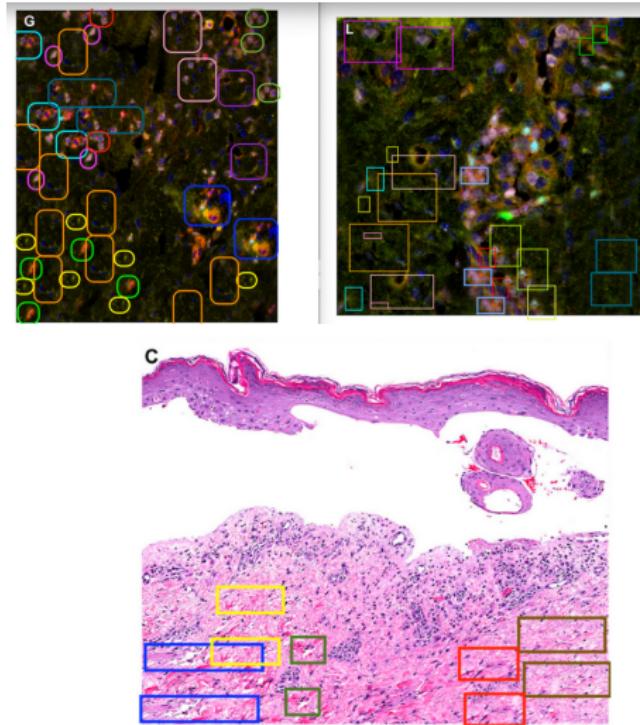
About

Elizabeth Bik is a renowned microbiologist and science integrity advocate known for detecting image duplication in scientific publications. Through meticulous analysis, she has exposed thousands of cases of misconduct, fostering transparency and accountability. Despite facing criticism and legal challenges, her work has earned widespread acclaim, including the John Maddox Prize and the Einstein Foundation Award, highlighting her vital role in upholding scientific integrity.

RECENT POSTS

- [ScienceIntegrity, where I diagnosed and then complained about Pohleus](#) April 18, 2023
- [Science Integrity Digest – catching up](#) February 21, 2023
- [Science Integrity Digest, September 2022](#) December 1, 2022
- [Science Integrity Digest, August 2022](#)

scienceintegritydigest.com



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

My hero: Dr. Elizabeth Bik, science sleuth



A screenshot of the Science Integrity Digest website. The header reads "Science Integrity Digest" and "A blog about science integrity, by Elizabeth Bik, for Hobbes-Bik LLC. Support my work at Patreon.com/elizabethbik". Below the header are links for "Home", "About", "FAQ", and "How-To guides". The "About" section features a bio for Elizabeth Bik and a "RECENT POSTS" sidebar with links to various posts.

About

Elizabeth Bik is a renowned microbiologist and science integrity advocate known for detecting image duplication in scientific publications. Through meticulous analysis, she has exposed thousands of cases of misconduct, fostering transparency and accountability. Despite facing criticism and legal challenges, her work has earned widespread acclaim, including the John Maddox Prize and the Einstein Foundation Award, highlighting her vital role in upholding scientific integrity.

RECENT POSTS

- [ScienceIntegrity, where I diagnosed and then complained about Pielou's April 18, 2023](#)
- [When integrity agents = catching up](#) (Pielou's 21, 2023)
- [Science Integrity Digest, September 2023](#) (October 1, 2023)
- [Science Integrity Digest, August 2023](#)

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 explicit manipulation as done here.

My work for fundamental empirical research in ML/DL

Search for papers and reproductions.

Reproduced Papers | About | Sign in

Hub for reproduced deep learning papers and their reproductions

Statistics

# Papers	# Reproductions	# Reproductions / Paper
180	436	2.4

Reproduced Papers

Hub for reproduced deep learning papers and their reproductions

Statistics

# Papers	# Reproductions	# Reproductions / Paper
180	436	2.4

ReproducedPapers.org

Search for papers and controlled datasets

Controlled Datasets | About | TU Delft | Sign in

Hub for papers and associated controlled datasets

Statistics

# Papers	# Controlled Datasets
6	4

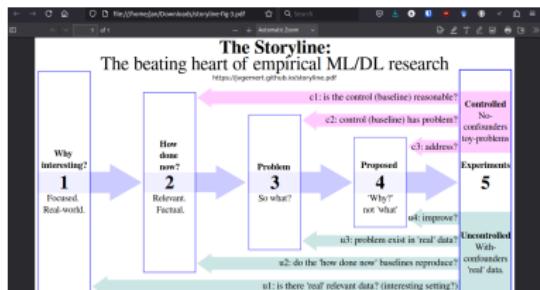
Controlled Datasets

Hub for papers and associated controlled datasets

Statistics

# Papers	# Controlled Datasets
6	4

ControlledExperimentsInML.org



Online research guidelines

TU Delft | DSAIT4205 Fundamental Research in Machine and Deep Learning (2024/25 Q4)

Course Home Content Collaboration Assignments Grades More Course Admin Help

DSAIT4205 Fundamental Research in Machine and Deep Learning (2024/25 Q4)

Announcements

Submission deadlines + Buddycheck

Post on Jun 20, 2025 15:00
Dear Students,

First and foremost, we want to remind you about the submissions deadlines for

Storyline Due Jun 20 • DSAIT4205 Fundamentals
Toy problem (control-obstinate) Due Jun 20 • DSAIT4205 Fundamentals
View of activities

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... (?)

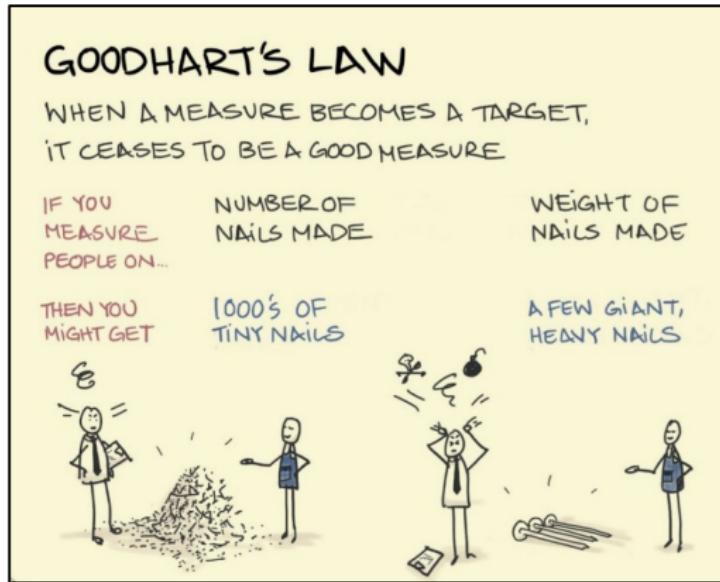
My limited experience in Benchmarking for Methodologies

The screenshot shows a web browser window for the "4th Visual Inductive Priors for Data-Efficient Deep Learning Workshop". The URL is vipriors.github.io. The page features a blue header bar with the workshop's name and details: "4th Visual Inductive Priors for Data-Efficient Deep Learning Workshop", "ICCV 2023 @ Room E03 (Poster room W02)", and "Monday October 2nd 2023, 8:45 - 13:00". Below the header is a complex diagram of interconnected nodes, each containing mathematical expressions related to deep learning and priors. At the bottom of the page is a footer bar with the text "Saving data by adding visual knowledge priors to Deep Learning." and the website URL again.

[https://vipriors.github.io/](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?