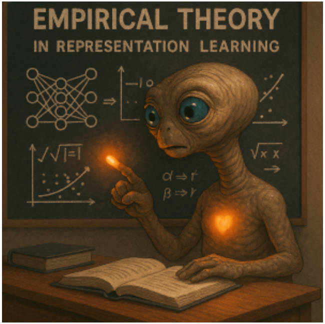
E.T.: Empirical Theory in Representation Learning

Phoning home to science

Sept 08, 2026



Jan van Gemert

Existential questions

What are Deep Learned foundation/frontier models for?

Existential questions

What are Deep Learned foundation/frontier models for?

- AI as a Swiss army knife: Used broadly in any application.

Existential questions

What are Deep Learned foundation/frontier models for?

- AI as a Swiss army knife: Used broadly in any application.

What are Deep Learning researchers for? (hint: not for building applications)

Existential questions

What are Deep Learned foundation/frontier models for?

- AI as a Swiss army knife: Used broadly in any application.

What are Deep Learning researchers for? (hint: not for building applications)

- Investigate, understand, and communicate how to let application-specific engineers build foundation model systems.

Existential questions

What are Deep Learned foundation/frontier models for?

- AI as a Swiss army knife: Used broadly in any application.

What are Deep Learning researchers for? (hint: not for building applications)

- Investigate, understand, and communicate how to let application-specific engineers build foundation model systems.

We train application-specific engineers using courses/textbooks;

Existential questions

What are Deep Learned foundation/frontier models for?

- AI as a Swiss army knife: Used broadly in any application.

What are Deep Learning researchers for? (hint: not for building applications)

- Investigate, understand, and communicate how to let application-specific engineers build foundation model systems.

We train application-specific engineers using courses/textbooks;

Deep learning researchers find out what should be in the textbook.

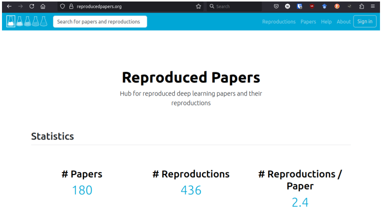
WHOAMI: Jan van Gemert

Computer Vision lab @ TU Delft

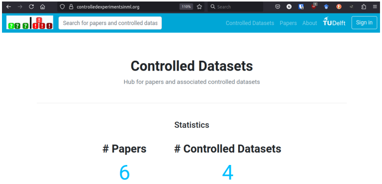
Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. Increase data-efficiency to allow visual AI model training outside big tech.

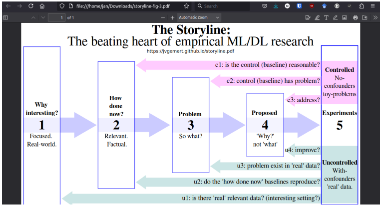
My work for fundamental empirical research in ML/DL



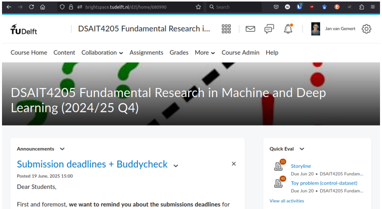
ReproducedPapers.org



ControlledExperimentsInML.org



Online research guidelines



MSc course

Recent efforts:

Metascience for Machine Learning

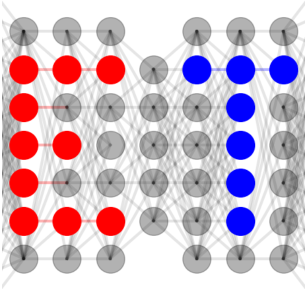
Holding a magnifying glass up to the ways of doing machine learning research.

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/`



This ECCV'26 workshop

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

Pre-registration and registered reports: Example of intervention forced by researchers on researchers, to align incentives (in Psychology):

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

Pre-registration and registered reports: Example of intervention forced by researchers on researchers, to align incentives (in Psychology):

- Pre-registration: to motivate hypotheses. Avoid "p-hacking" of spurious unreproducable results; avoids the cost in invalid theory building.
- Registered report: to allow negative results to get past reviewer 2; avoid costs of repeatedly redoing plausible negatives.

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

Pre-registration and registered reports: Example of intervention forced by researchers on researchers, to align incentives (in Psychology):

- Pre-registration: to motivate hypotheses. Avoid "p-hacking" of spurious unreproducable results; avoids the cost in invalid theory building.
- Registered report: to allow negative results to get past reviewer 2; avoid costs of repeatedly redoing plausible negatives.

Why not?

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

Pre-registration and registered reports: Example of intervention forced by researchers on researchers, to align incentives (in Psychology):

- Pre-registration: to motivate hypotheses. Avoid "p-hacking" of spurious unreproducable results; avoids the cost in invalid theory building.
- Registered report: to allow negative results to get past reviewer 2; avoid costs of repeatedly redoing plausible negatives.

Why not?

- Reproducibility costs are low in ML. See: ICML'26 Reproduction HF hackathon:

> Back in July, we ran a hackathon where more than 1,200 community members brought their own coding agents and tried to reproduce the papers published at ICML 2026, claim by claim. In 19 days, participants published 6,816 Trackio logbooks reproducing 2,226 papers, about a third of the conference 🤯

~~Exploring paper pre-registration~~

I first thought to do a paper pre-registration; here's why I didn't:

Pre-registration and registered reports: Example of intervention forced by researchers on researchers, to align incentives (in Psychology):

- Pre-registration: to motivate hypotheses. Avoid "p-hacking" of spurious unreproducable results; avoids the cost in invalid theory building.
- Registered report: to allow negative results to get past reviewer 2; avoid costs of repeatedly redoing plausible negatives.

Why not?

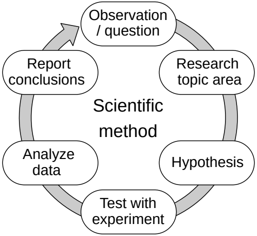
- Reproducibility costs are low in ML. See: ICML'26 Reproduction HF hackathon:

> Back in July, we ran a hackathon where more than 1,200 community members brought their own coding agents and tried to reproduce the papers published at ICML 2026, claim by claim. In 19 days, participants published 6,816 Trackio logbooks reproducing 2,226 papers, about a third of the conference 🤯

- Negative results get published. (eg: "metric learning reality check"; "Is there progress in activity progress prediction?"; "Are current long-term video understanding datasets long-term?")

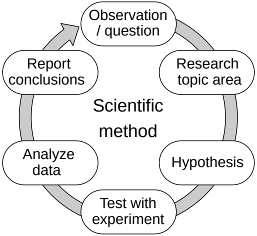
The scientific method[1] in times of deep learning

Often, deep learning is scientifically shallow [2,3,4]



The scientific method[1] in times of deep learning

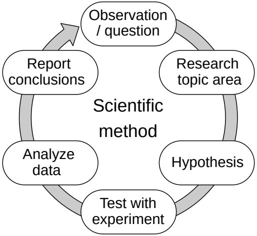
Often, deep learning is scientifically shallow [2,3,4]



- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

The scientific method[1] in times of deep learning

Often, deep learning is scientifically shallow [2,3,4]

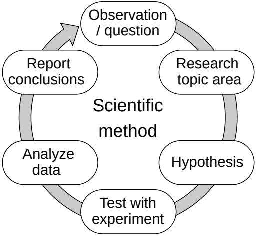


- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

The scientific method[1] in times of deep learning

Often, deep learning is scientifically shallow [2,3,4]



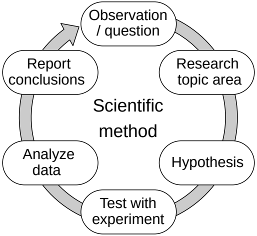
- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- ML/DL has not many empirical theories.

The scientific method[1] in times of deep learning

Often, deep learning is scientifically shallow [2,3,4]

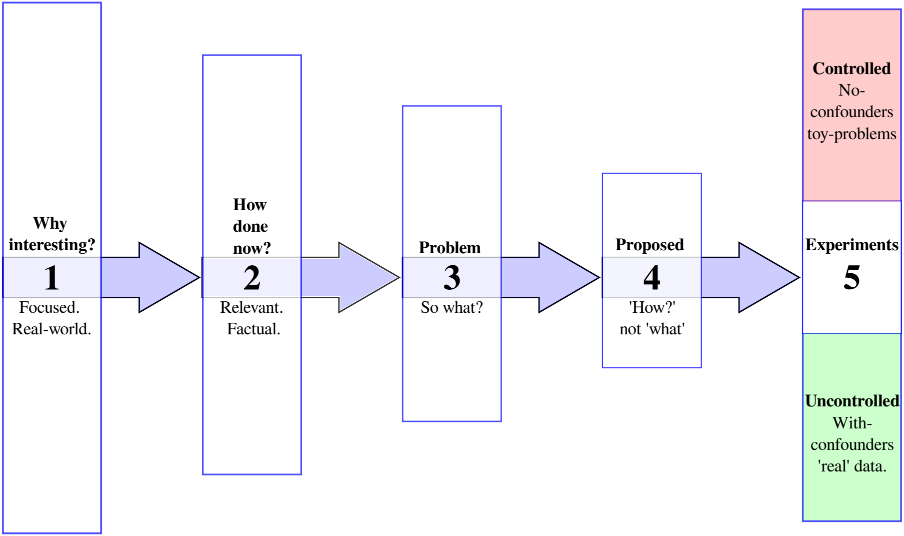


- Trial and error (graduate student descent)
- Opportunistic ('getting in'). Salary vs science;
- Incentives innovation, not science;
- Reviewer damage (Bold-nr fetish; Mathiness)[2];
- SOTA-focus benefits the large compute/data few;
- Confusing speculation with explanation[2];
- Not identifying the reasons for empirical gains[2].
- Confusing benchmarks with 'real-world'
- Not clear if results generalize to other work.

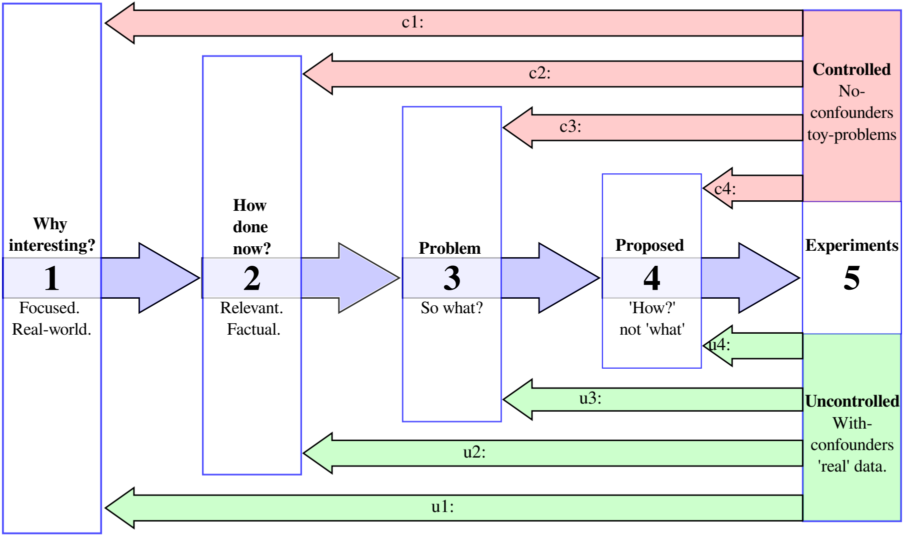
"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- ML/DL has not many empirical theories. Some that I am aware of:
  ◦ Neural Scaling Laws; ◦ Bias/variance ◦ ML is like physics/neuroscience;
  ◦ Simple axioms explain intelligence ◦ Platonic representation hypothesis ◦ ...

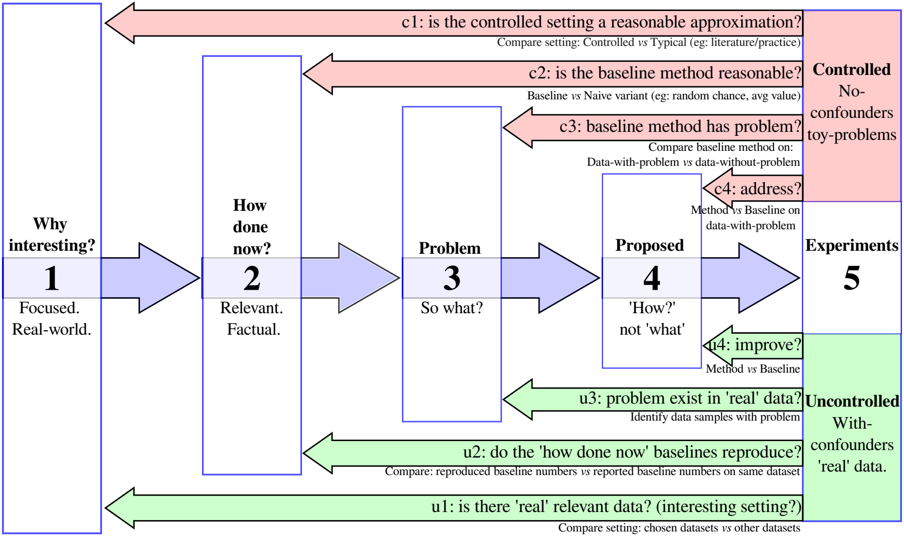
Storyline: Towards empirical theory



Storyline: Force empirical evidence for each step

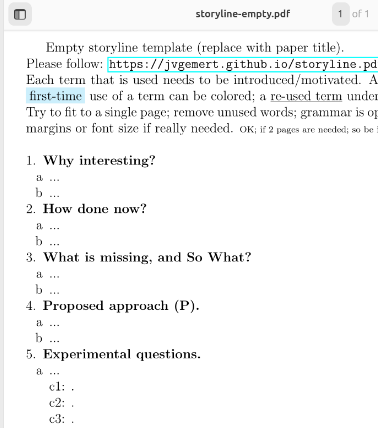


Storyline: Actionable for both author and reviewer



Storyline encourages empirical rigor

- Storyline required to fit on 1 page to have all points available at the same time.
- ~~Forces~~ Helps the authors to provide empirical evidence
- We have "normal" review questions: summary; pos/neg; scope; rating.
- Reviewers are asked to review the storyline for *argumentation* (point 1-4 in the storyline) and for *Empirical rigor* (the arrows)


storyline-empty.pdf 1 of 1

Empty storyline template (replace with paper title).
Please follow: https://jvgemert.github.io/storyline.pd
Each term that is used needs to be introduced/motivated. A first-time use of a term can be colored; a re-used term under
Try to fit to a single page; remove unused words; grammar is op
margins or font size if really needed. OK; if 2 pages are needed; so be

1. **Why interesting?**
   a ...
   b ...
2. **How done now?**
   a ...
   b ...
3. **What is missing, and So What?**
   a ...
   b ...
4. **Proposed approach (P).**
   a ...
   b ...
5. **Experimental questions.**
   a ...
   c1: .
   c2: .
   c3: .

Workshop program

Program on the website

Thanks to all co-organizers and keynotes!

References

[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).