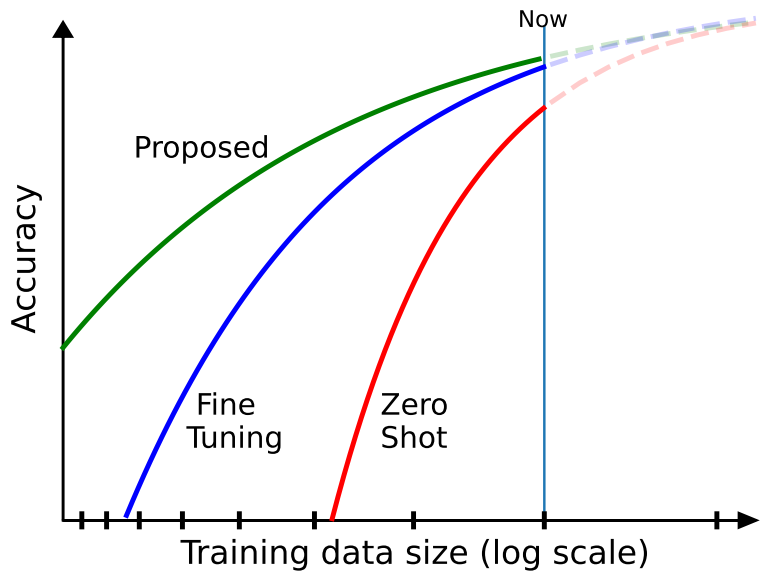
Project DAIta: Data Efficient Foundation Models NWO VICI

FIND workshop

May 27, 2026



Jan van Gemert

Existential questions

What are Deep Learned foundation models for?

Existential questions

What are Deep Learned foundation models for?

- Foundation for Industry (FIND) : Your industrial application.

Existential questions

What are Deep Learned foundation models for?

- Foundation for Industry (FIND) : Your industrial application.

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

Existential questions

What are Deep Learned foundation models for?

- Foundation for Industry (FIND) : Your industrial 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 models for?

- Foundation for Industry (FIND) : Your industrial 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 models for?

- Foundation for Industry (FIND) : Your industrial application.

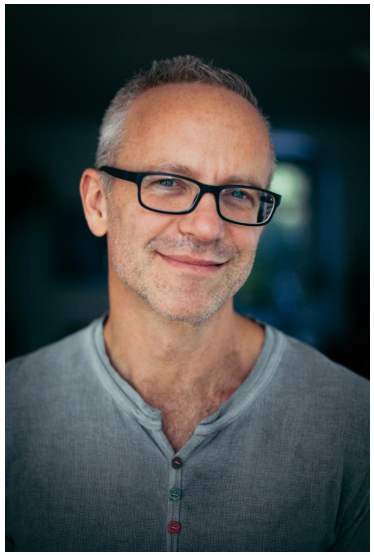
What are Deep Learning researchers for? (hint: not for building applications)

- 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

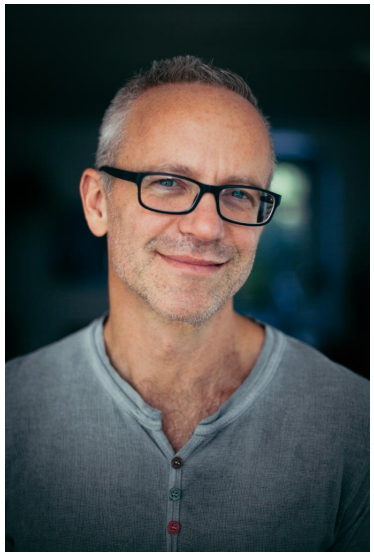


Computer Vision lab @ TU Delft

Two main research themes:

1. Fundamental empirical understanding-based deep learning research; (to)
2. 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.

Outline:

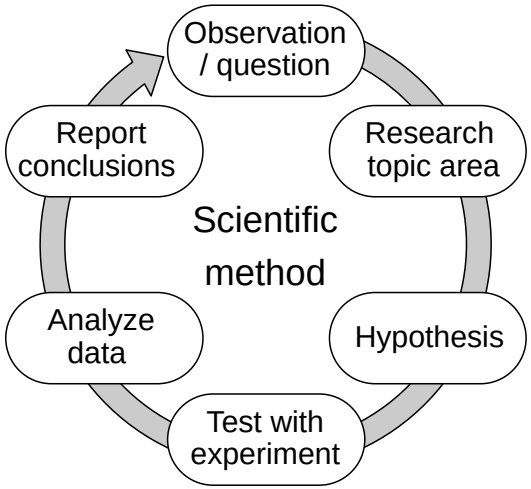
Towards an empirical science of deep learning
Doing science: better understand machine and deep learning methods.

Data Efficient Foundation model pre-Training
Reduce data dependency: pre-train foundation models with $O(\cdot)$ less data.

The scientific method[1] in times of deep learning

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

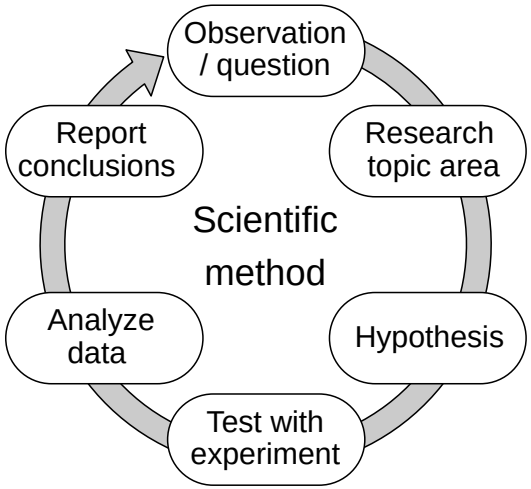
[1]: https://en.wikipedia.org/wiki/Scientific_method [2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018. [3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018. [4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed 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]

[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).

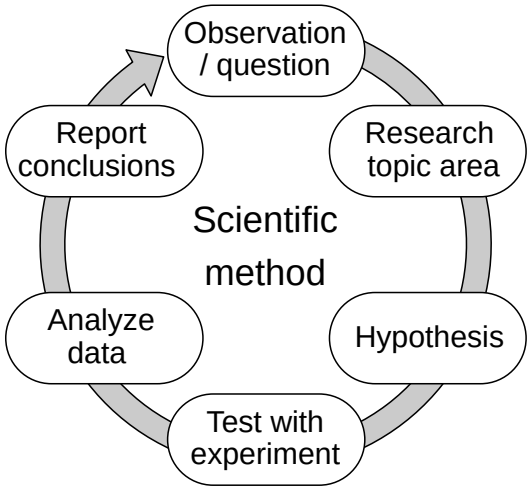


- Trial and error (graduate student descent)
- Opportunistic (career vs science);
- Gate keeping (large compute/data);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation;
- Not identifying the reasons for empirical gains.

The scientific method[1] in times of deep learning

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

[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).



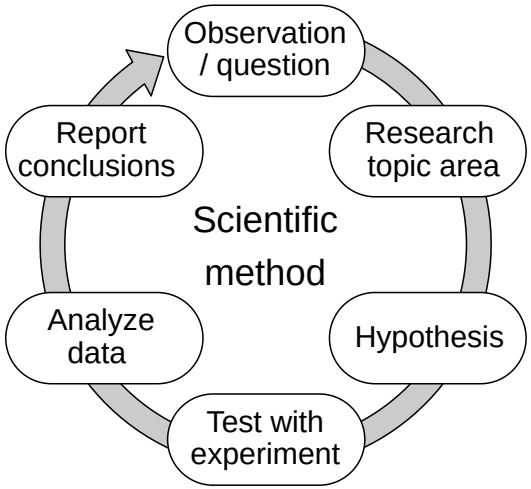
- Trial and error (graduate student descent)
- Opportunistic (career vs science);
- Gate keeping (large compute/data);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation;
- Not identifying the reasons for empirical gains.

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

The scientific method[1] in times of deep learning

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

[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).



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

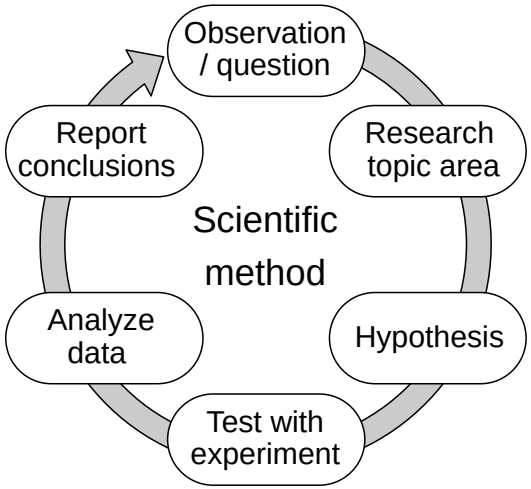
"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- ML/DL does not have many empirical theories.

The scientific method[1] in times of deep learning

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

[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).

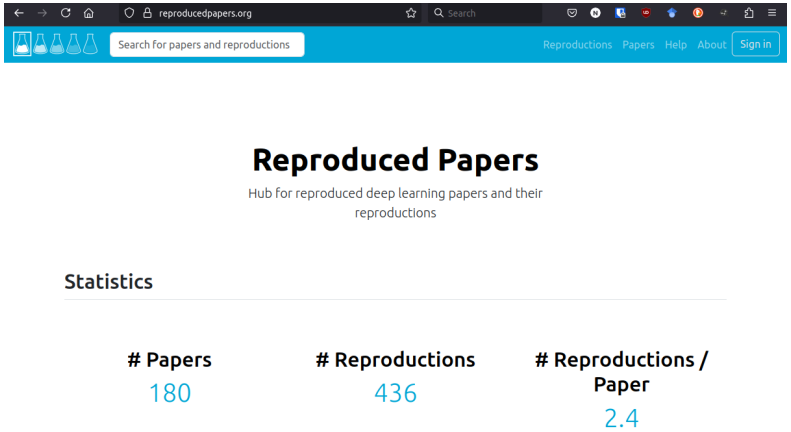


- Trial and error (graduate student descent)
- Opportunistic (career vs science);
- Gate keeping (large compute/data);
- Reviewer damage (bold-nr fetish; Mathiness);
- Confusing speculation with explanation;
- Not identifying the reasons for empirical gains.

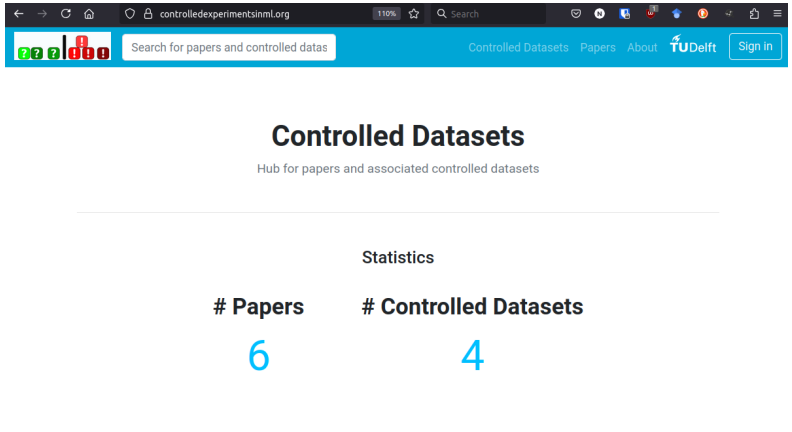
"GPU goes BRRR": no mathematical optimality; Need for *Empirical Theory*.

- 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 ◦ ...

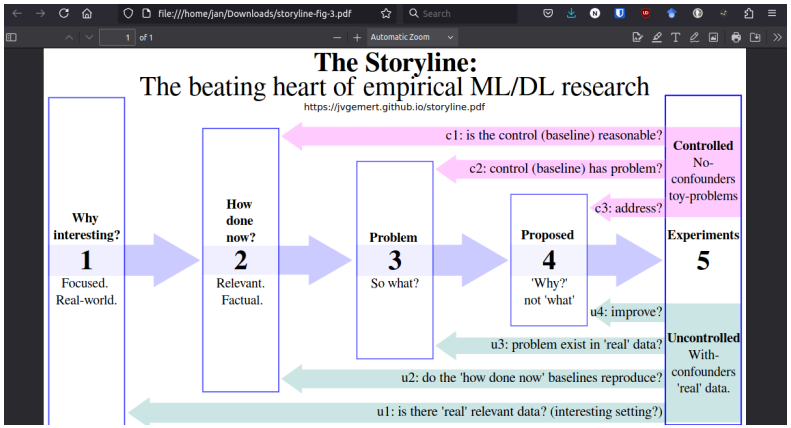
My work for fundamental empirical research in ML/DL



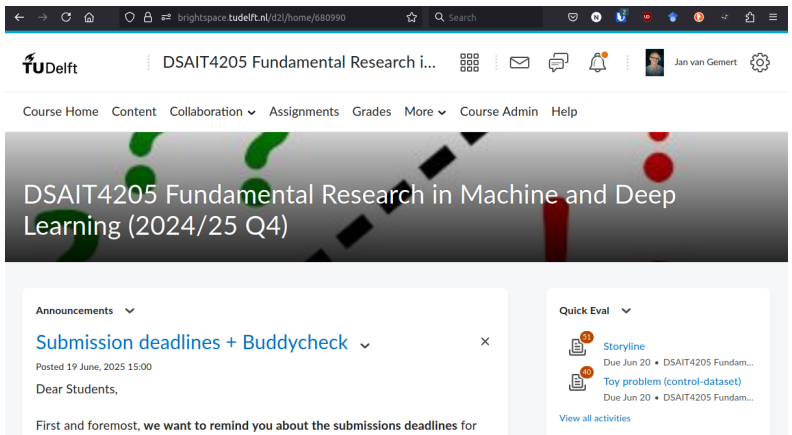
ReproducedPapers.org



ControlledExperimentsInML.org



Online research guidelines



MSc course

Recent efforts:

Metascience for Machine Learning

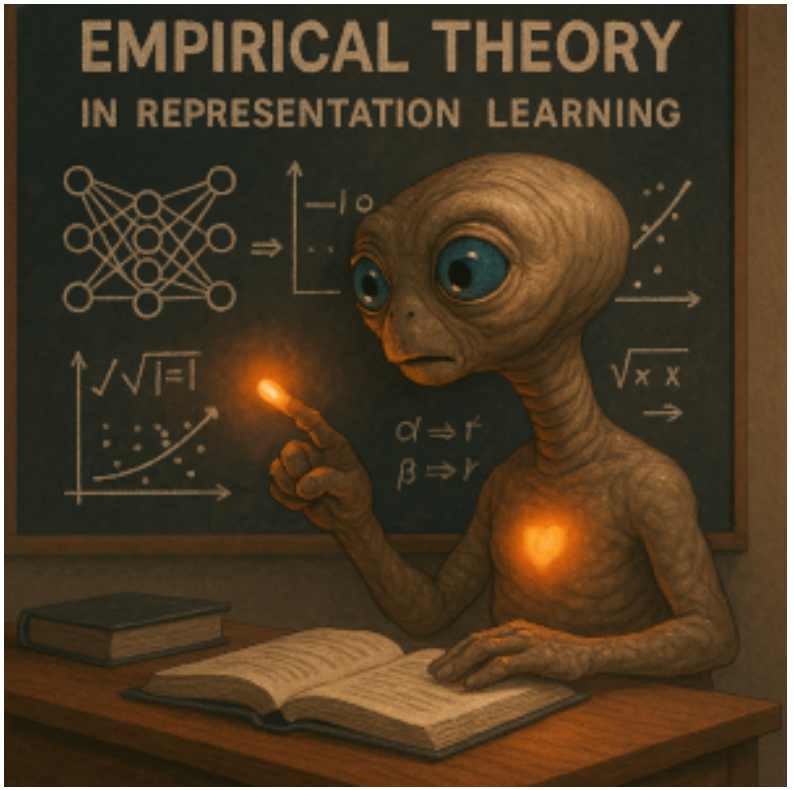
Holding a magnifying glass up to the ways of doing machine learning research.

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.

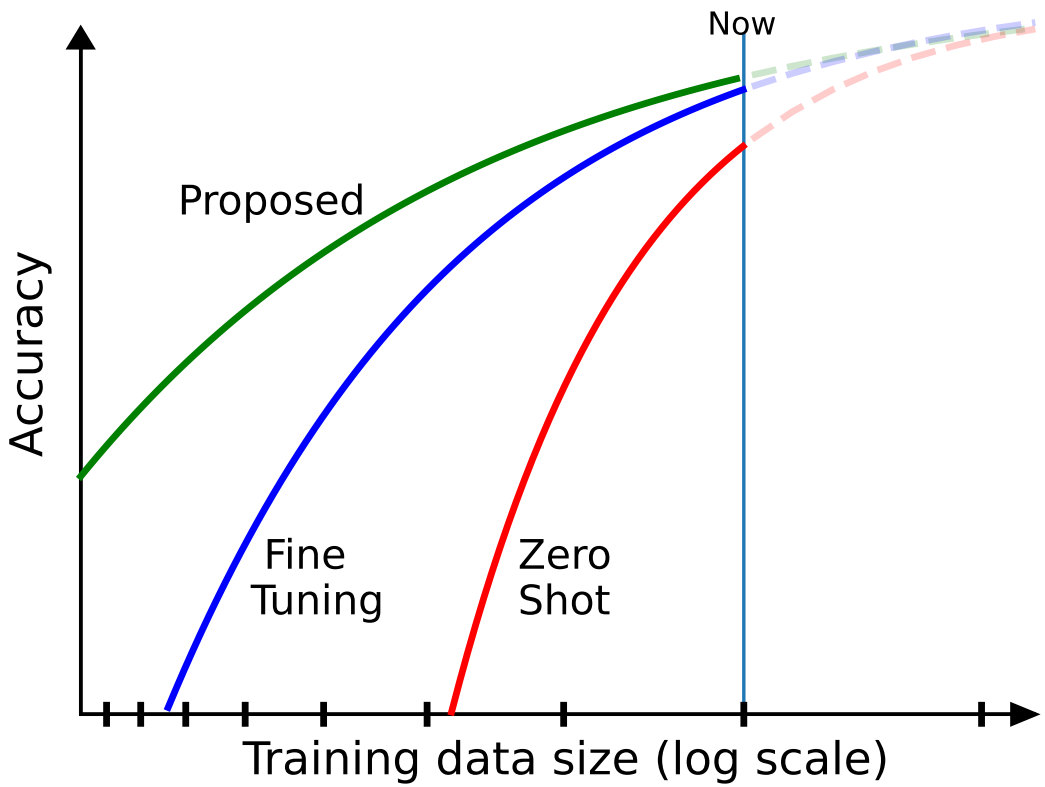
AI is changing the world, and machine learning has proved a valuable tool for other important scientific fields. Here, we turn the tables, and put the spotlight on the scientific research field of machine learning itself.

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

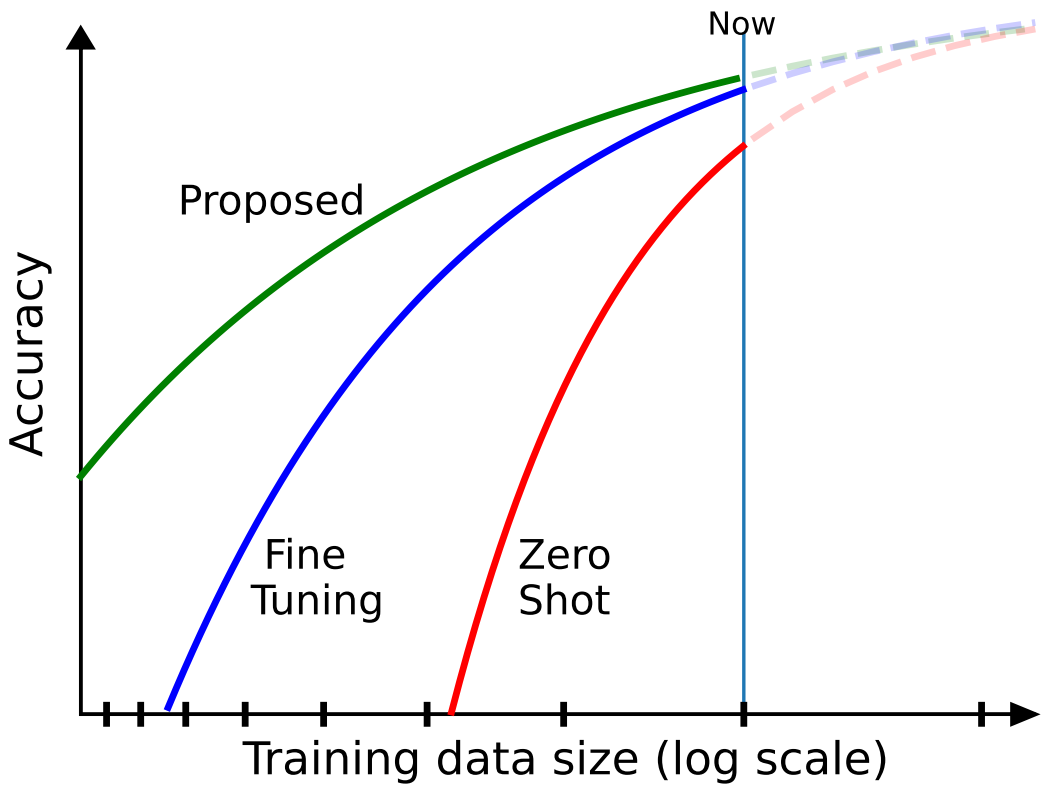


Workshop accepted at ECCV'26

Data-Efficient Foundation model pre-Training NWO VICI



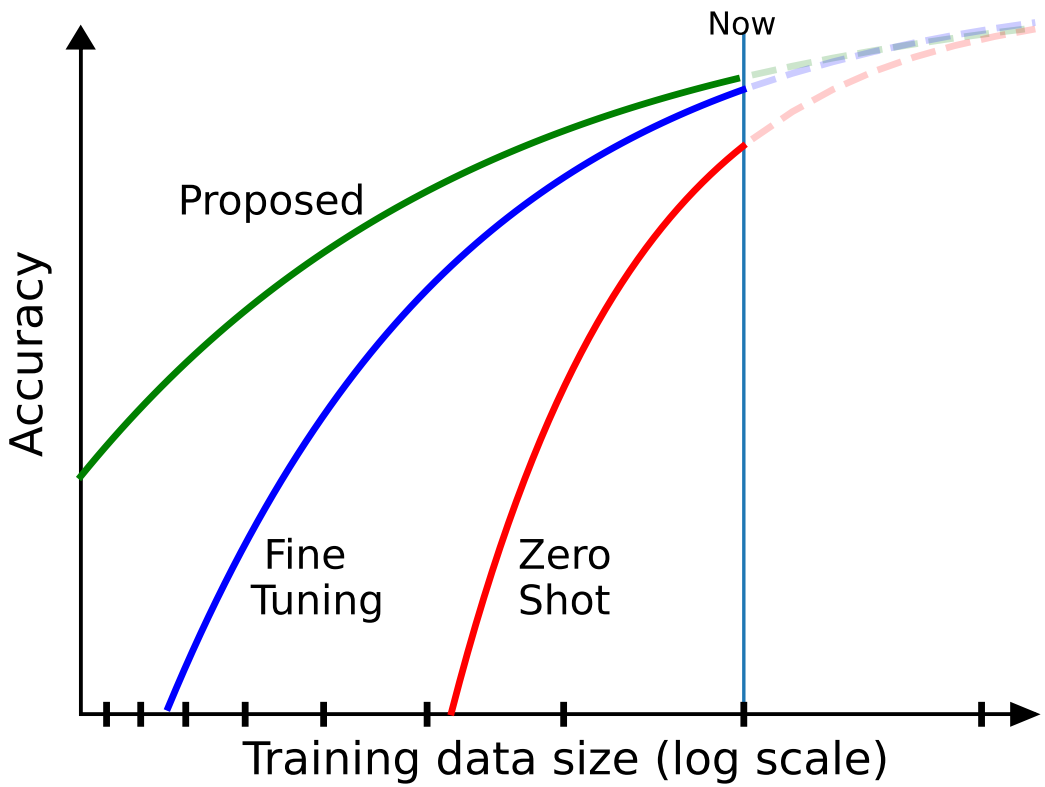
Data-Efficient Foundation model pre-Training NWO VICI



Why data-efficient?

- Fairness and bias in data
- Data opacity, traceability, privacy and copyright
- Trainability
- Sensitive data

Data-Efficient Foundation model pre-Training NWO VICI



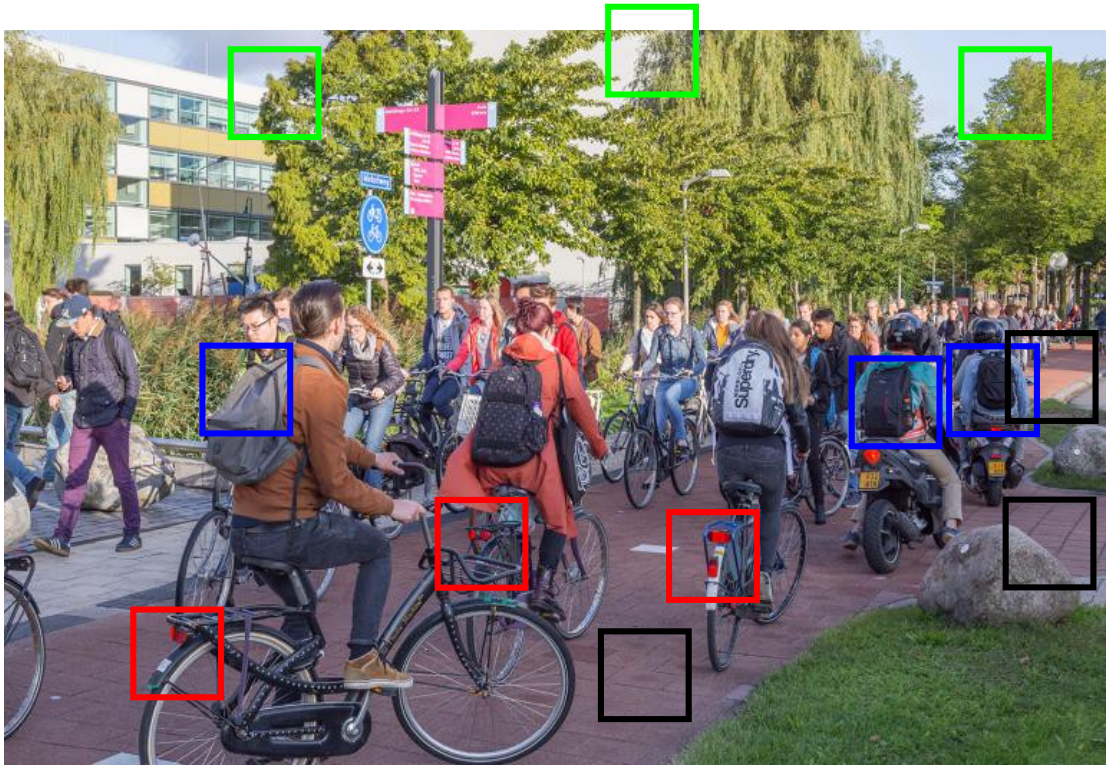
Why data-efficient?

- Fairness and bias in data
- Data opacity, traceability, privacy and copyright
- Trainability
- Sensitive data

Property:	Current	Proposed	Relevance
Data control:	Uncurated	Curateable	Fairness & Traceability.
Model control:	Third party	Independent	Sensitive data & AI research.
Data-efficiency:	Low	High	Expensive/scarce data

Motivation

A day at TU Delft campus



Accidental viewpoint:
All pixel patches are different

Backgrounds &
Out of bounds

Occlusion &
Lighting

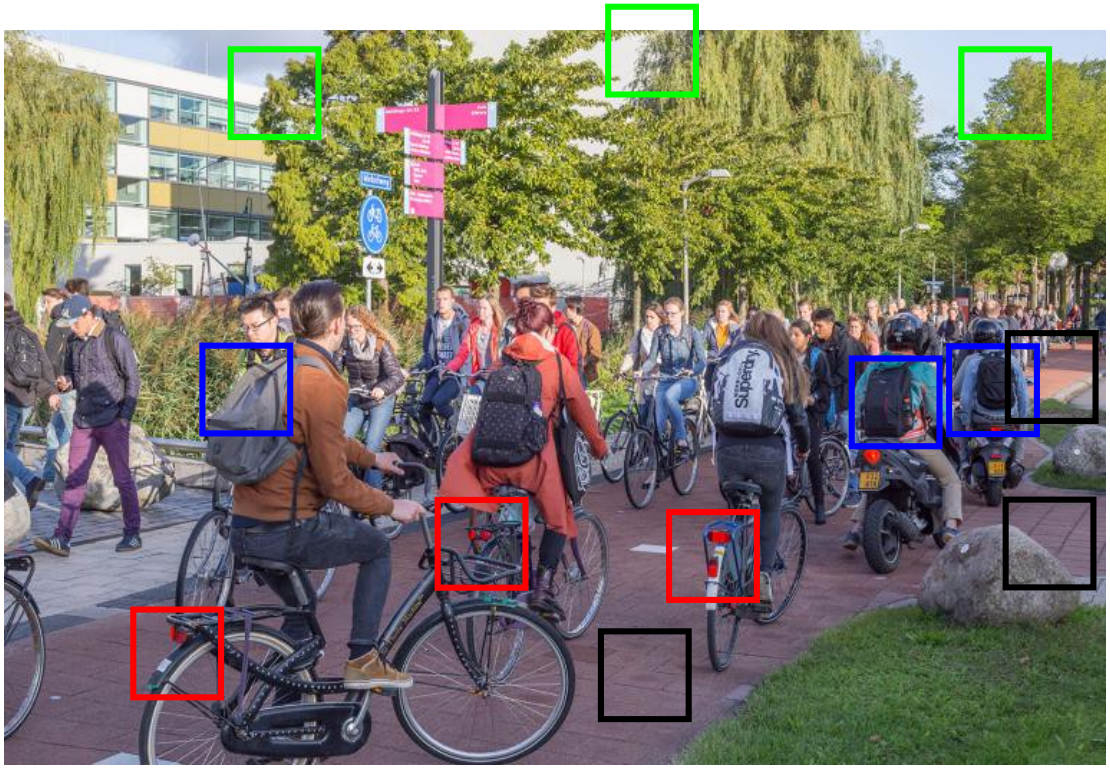
Appearance &
Geometry

Lighting &
Geometry

- Accidental camera position and lighting should play no role in recognition;

Motivation

A day at TU Delft campus



Accidental viewpoint:
All pixel patches are different

Backgrounds &
Out of bounds

Occlusion &
Lighting

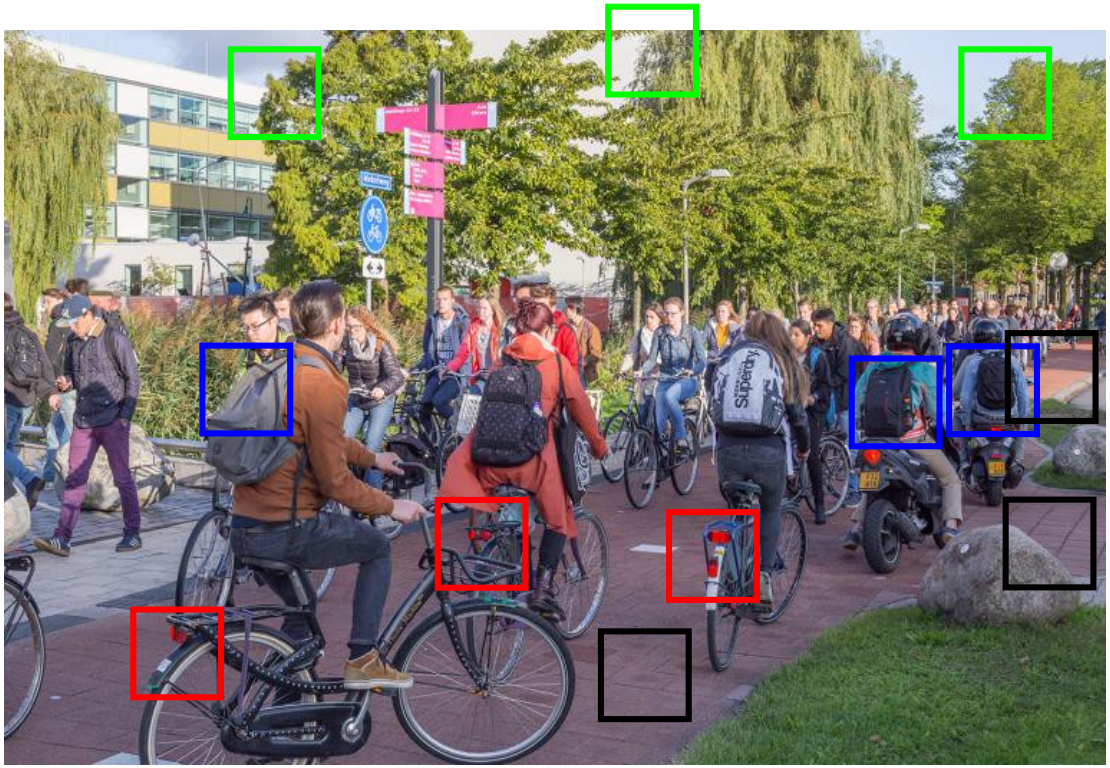
Appearance &
Geometry

Lighting &
Geometry

- Accidental camera position and lighting should play no role in recognition;
- Current visual foundation models treat each patch independently;

Motivation

A day at TU Delft campus



Accidental viewpoint:
All pixel patches are different

Backgrounds &
Out of bounds

Occlusion &
Lighting

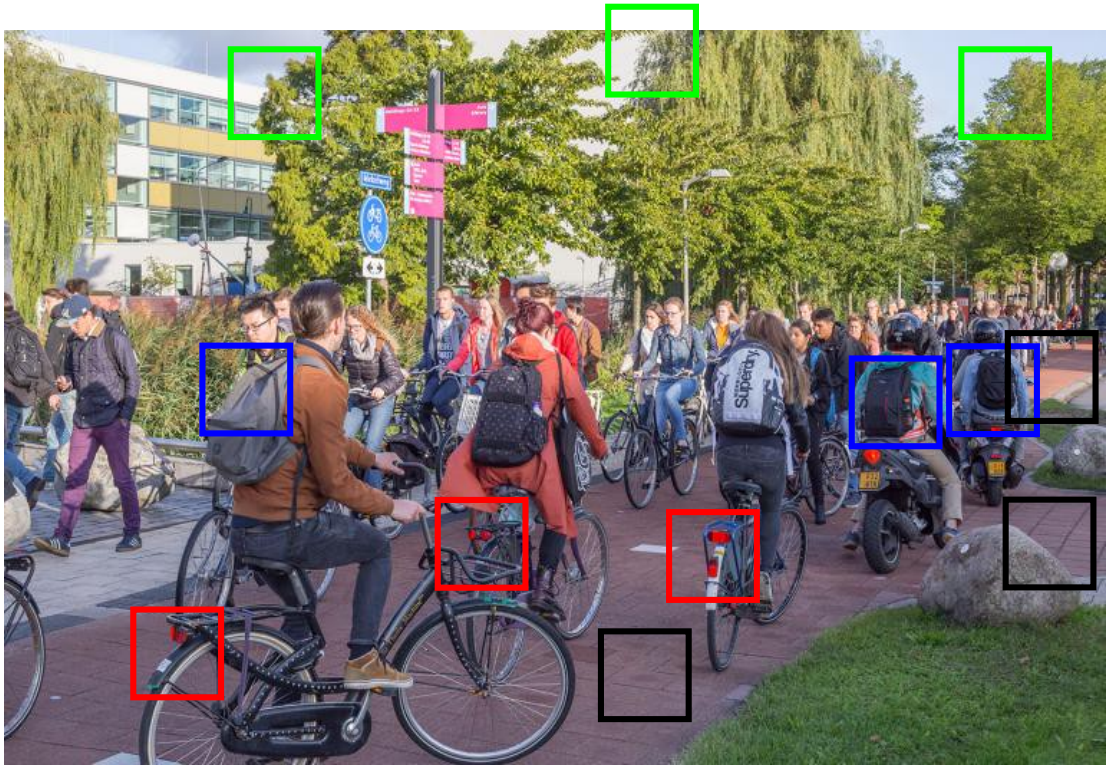
Appearance &
Geometry

Lighting &
Geometry

- Accidental camera position and lighting should play no role in recognition;
- Current visual foundation models treat each patch independently;
- Plan: exploit inherent physical camera and object properties to allow patches to learn from each other;

Motivation

A day at TU Delft campus



Accidental viewpoint:
All pixel patches are different

Backgrounds &
Out of bounds

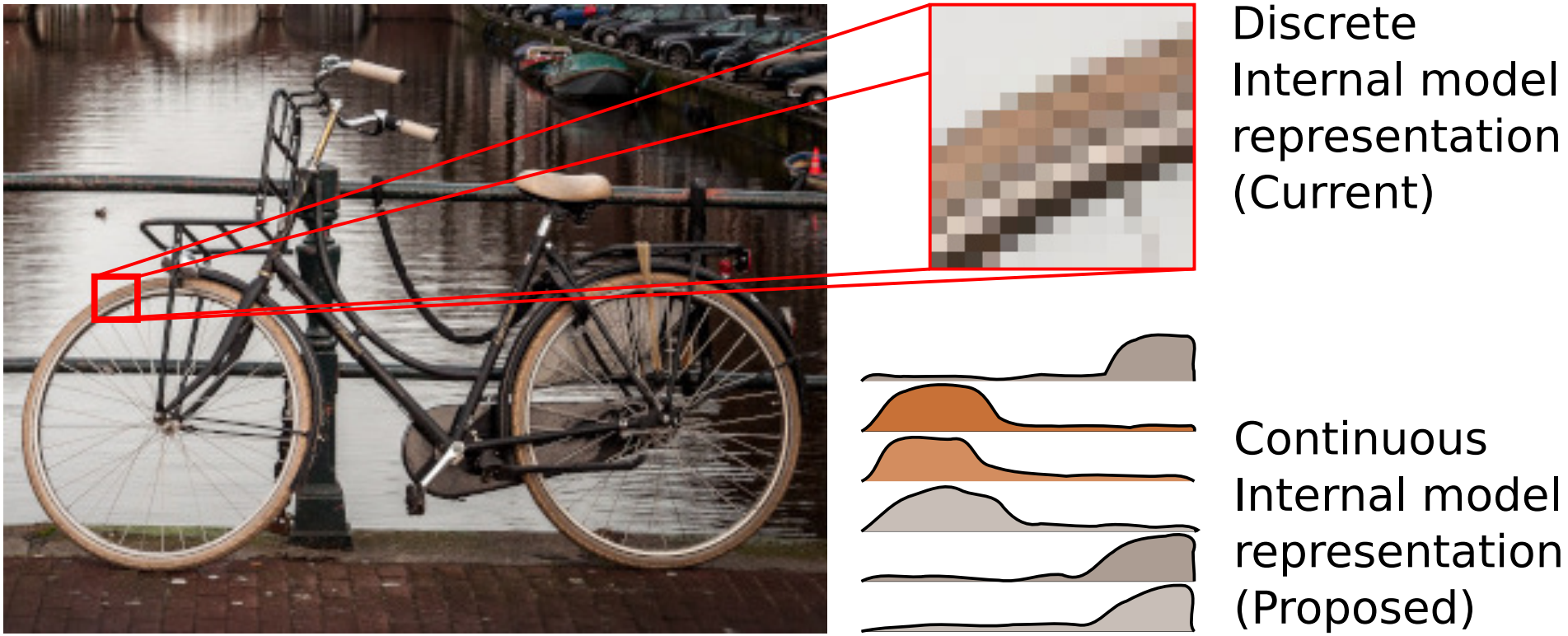
Occlusion &
Lighting

Appearance &
Geometry

Lighting &
Geometry

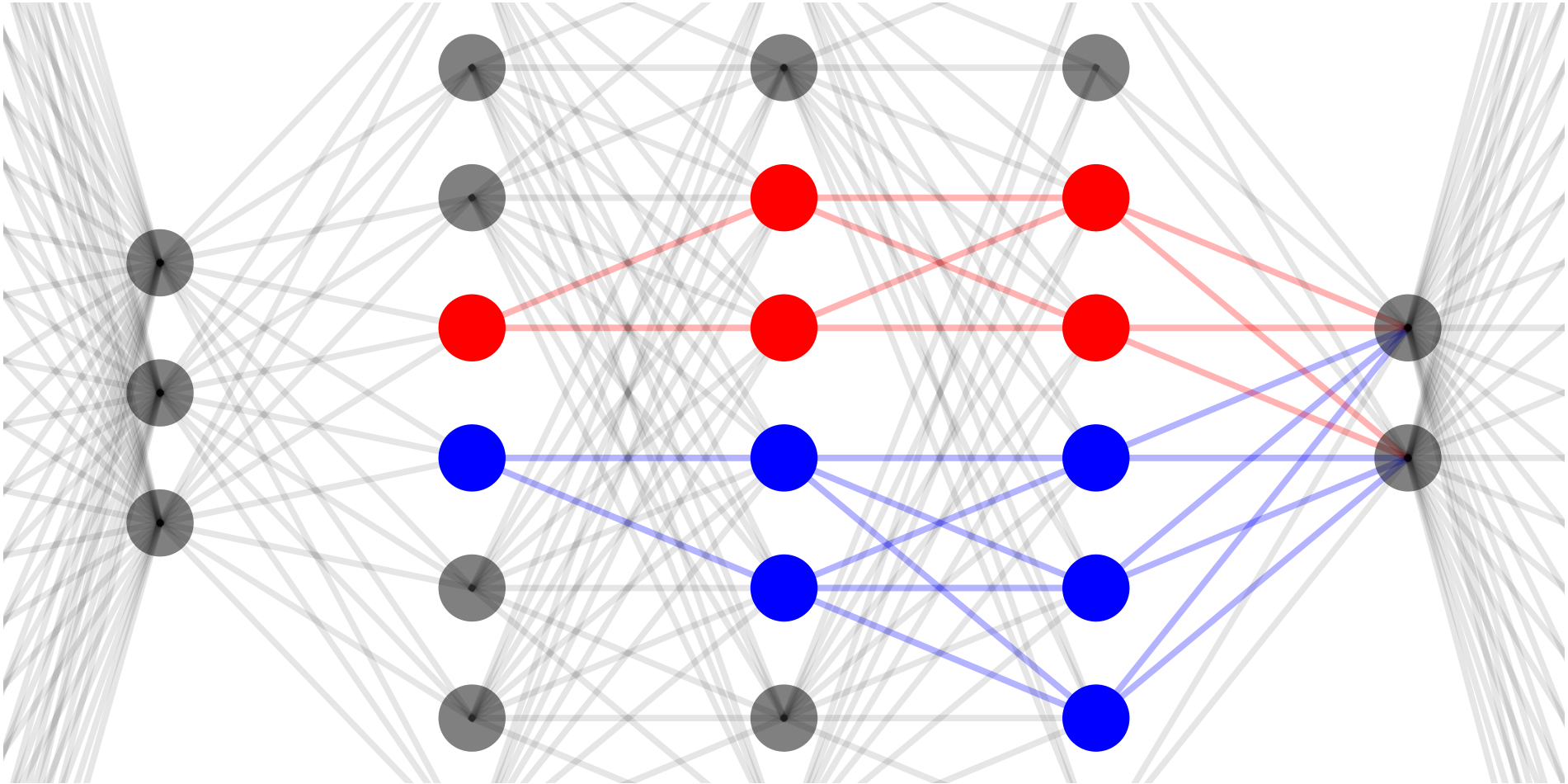
- Accidental camera position and lighting should play no role in recognition;
- Current visual foundation models treat each patch independently;
- Plan: exploit inherent physical camera and object properties to allow patches to learn from each other;
- Goal: share the learning signal between patches; 10-100$\times$ data reduction.

Why are visual AI internals on a grid?



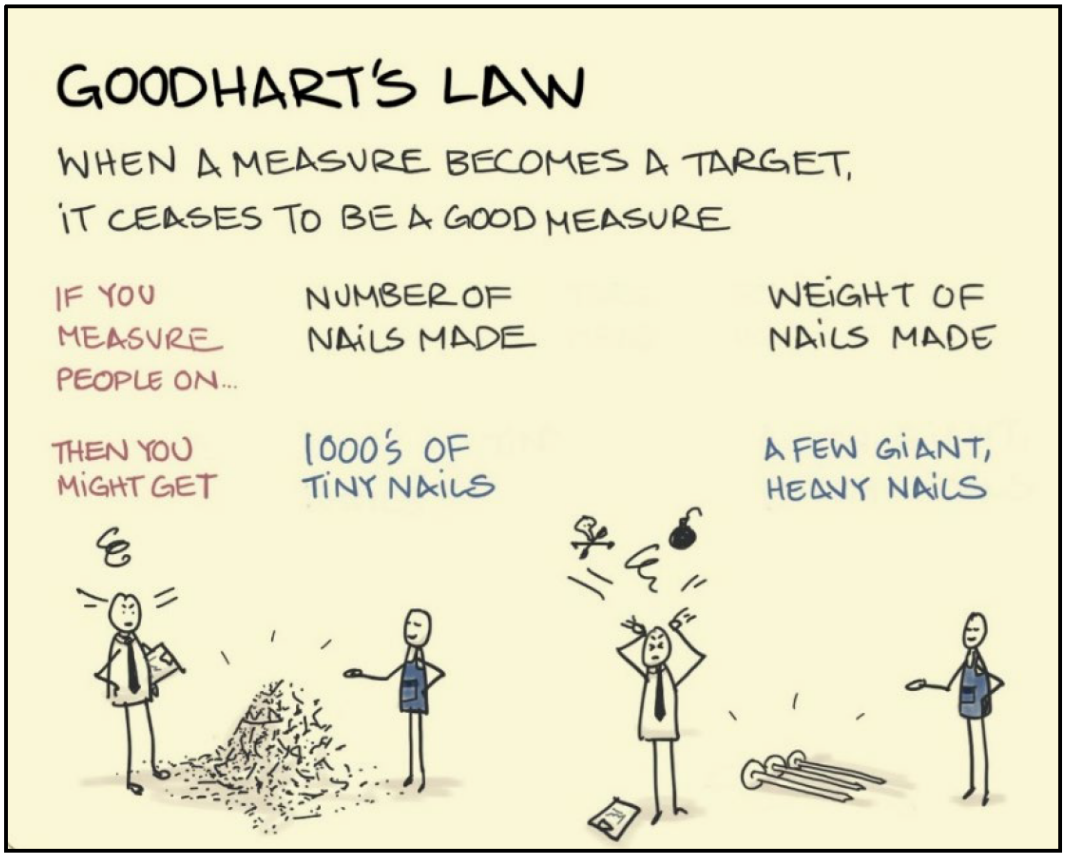
How to exploit that an image is taken by a camera? 2.5D

Modular, compositional, representations



Last slide: the question mark "?"

How to stimulate aligning fundamental research with real-world applications **"?"**



(Link to image source)

- Applications and use-cases are invaluable.
- Fundamentals: Make larger leaps by researching use-case abstractions.
- How to remain relevant as a small researcher in the age of big data/compute?