ML Research, via the Lens of ML

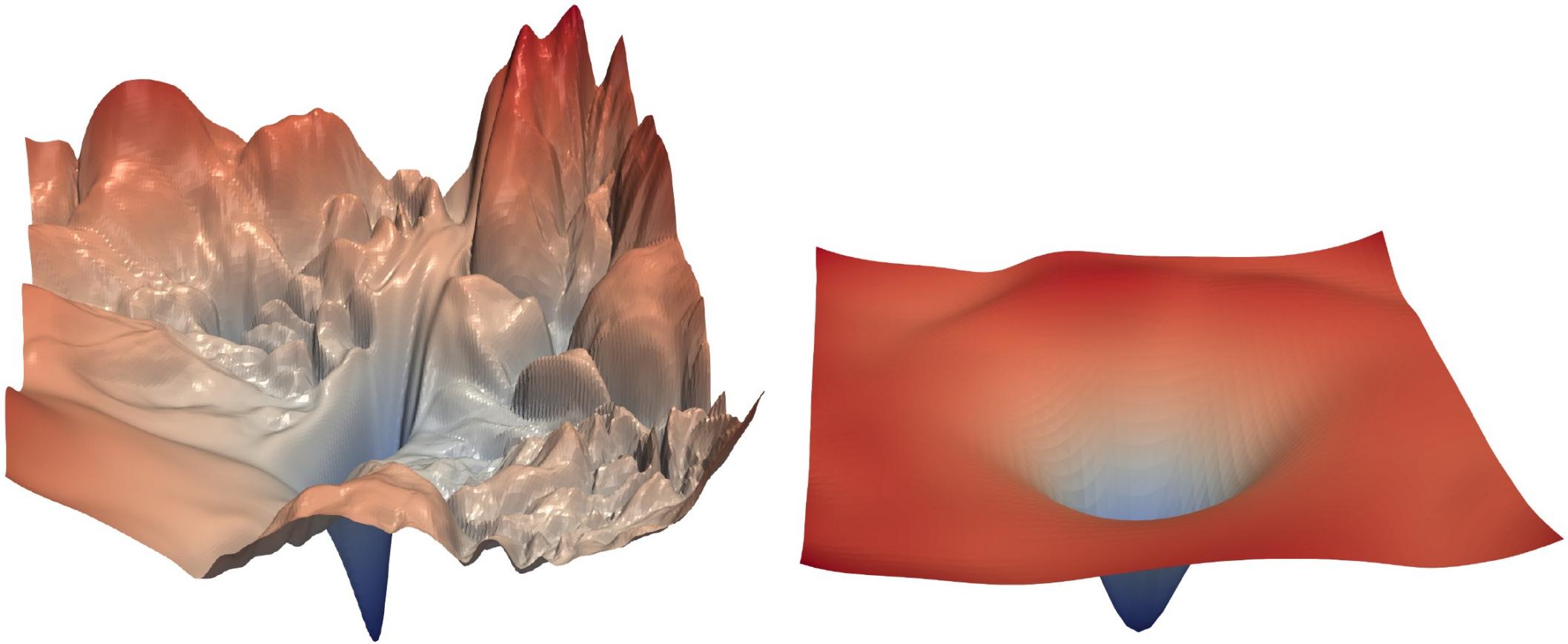
Kaiming He
Associate Professor, EECS, MIT

New In ML Workshop, NeurIPS 2024

- The way we see the world is shaped by our personal context.

- In this talk - **Inspect ML research by ML models**

Research is SGD
in a chaotic landscape



(a) without skip connections

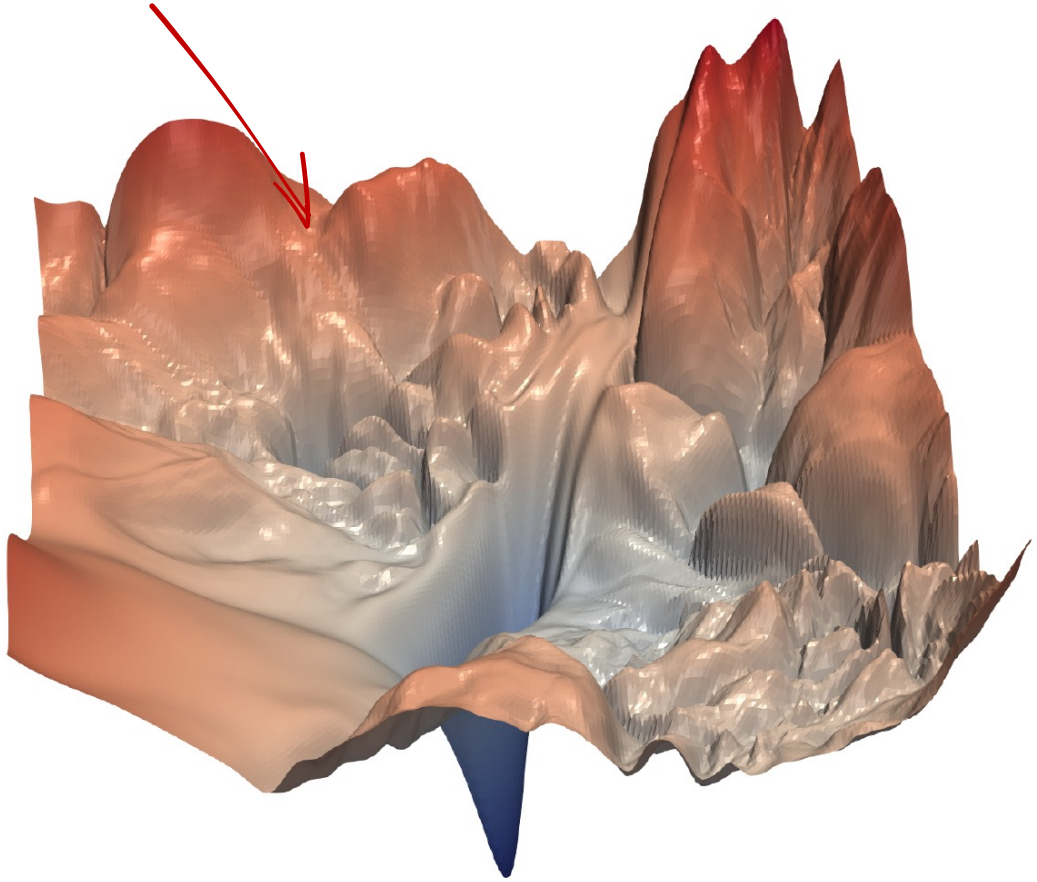
(b) with skip connections

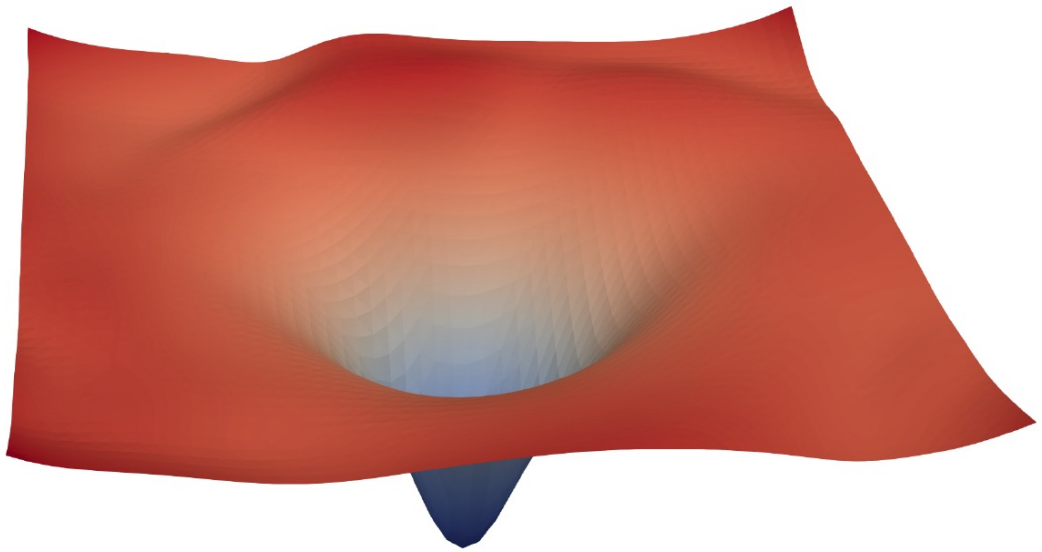
Figure 1: The loss surfaces of ResNet-56 with/without skip connections.

Figure borrowed from: “Visualizing the Loss Landscape of Neural Nets”, Li, Xu, Taylor, Goldstein, NeurIPS 2018

reality

hope





(a) without skip connections

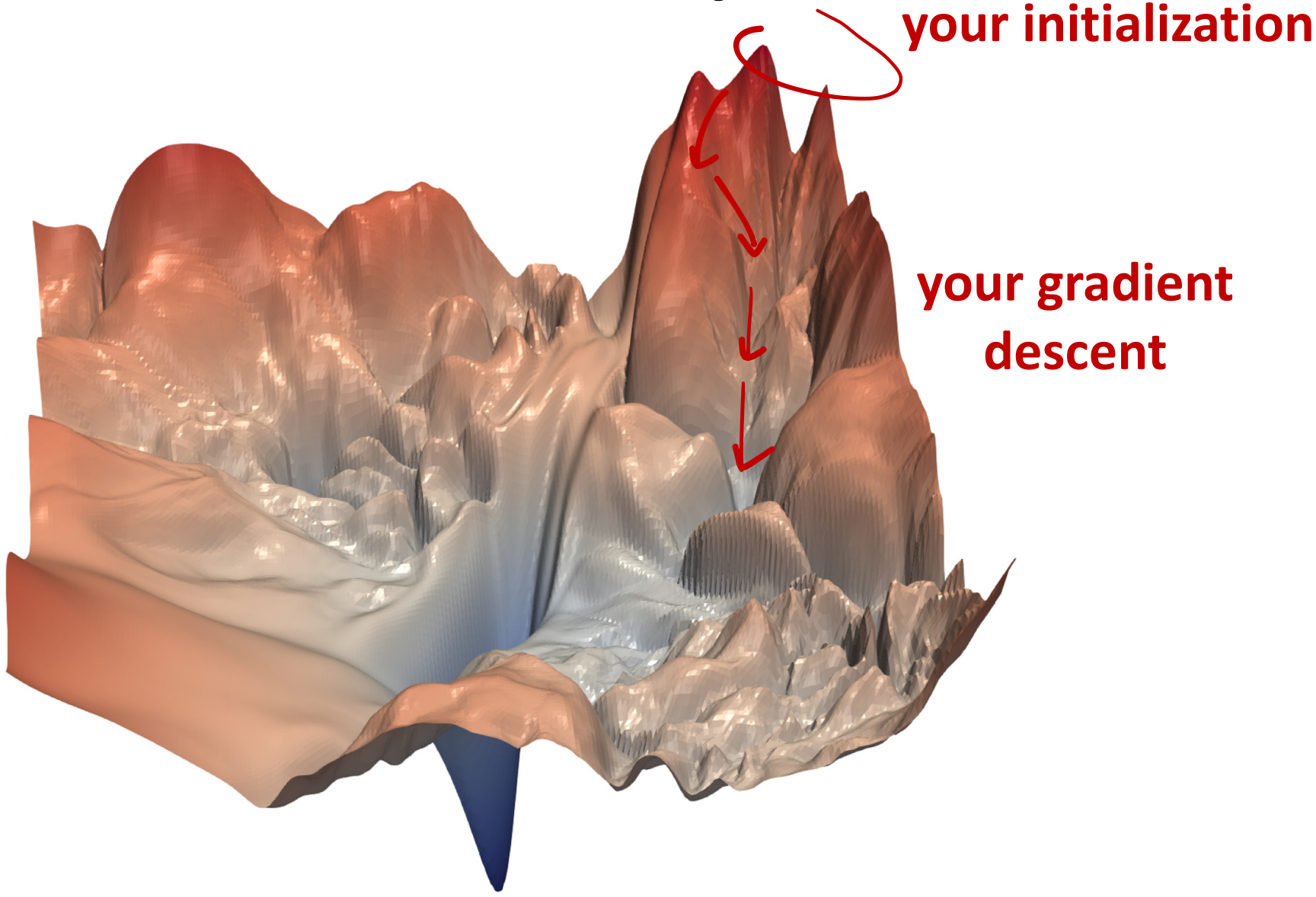
(b) with skip connections

~~Figure 1: The loss surfaces of ResNet-56 with/without skip connections.~~

Research landscape?

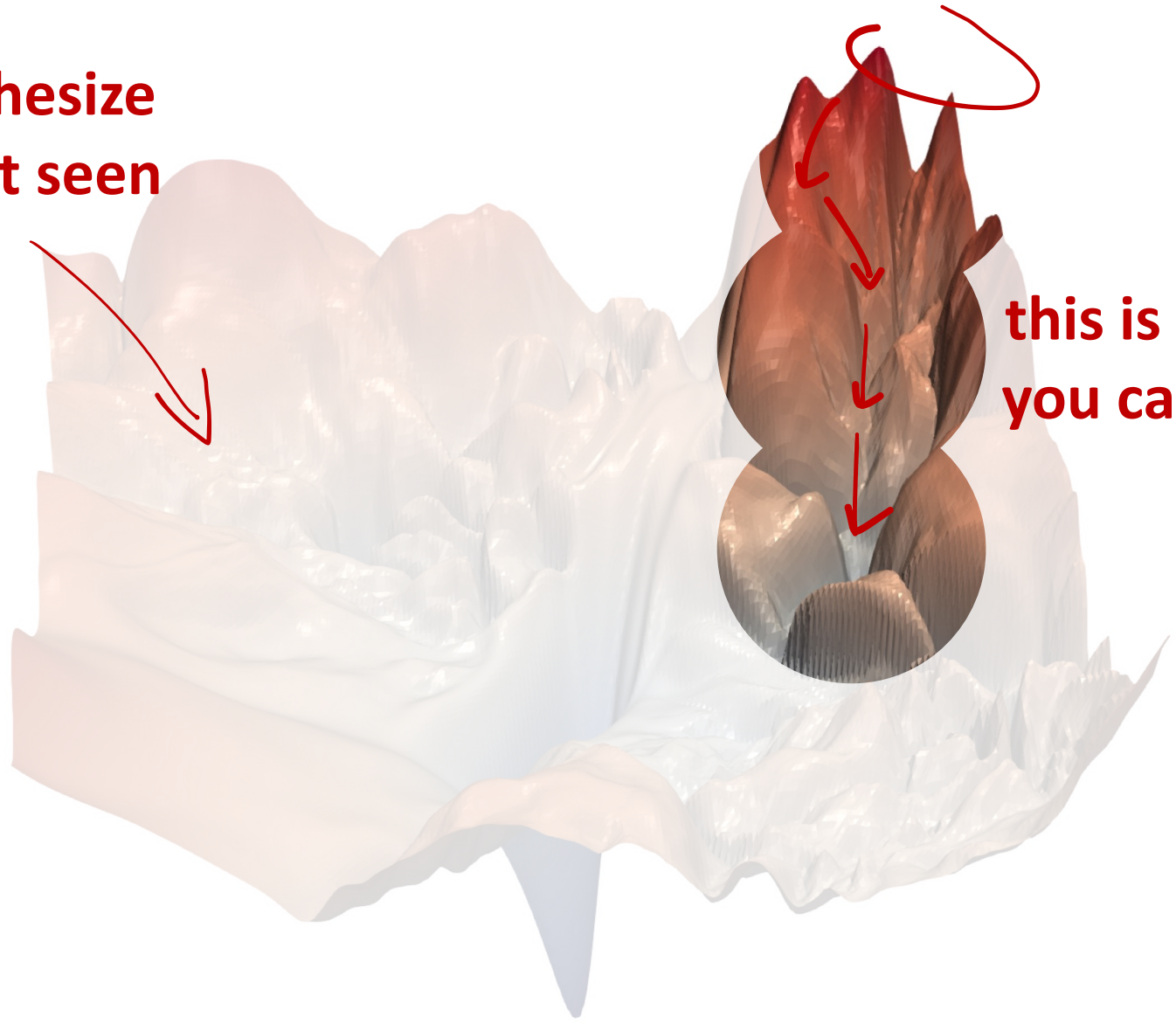
Figure borrowed from: “Visualizing the Loss Landscape of Neural Nets”, Li, Xu, Taylor, Goldstein, NeurIPS 2018

Research is SGD in the landscape



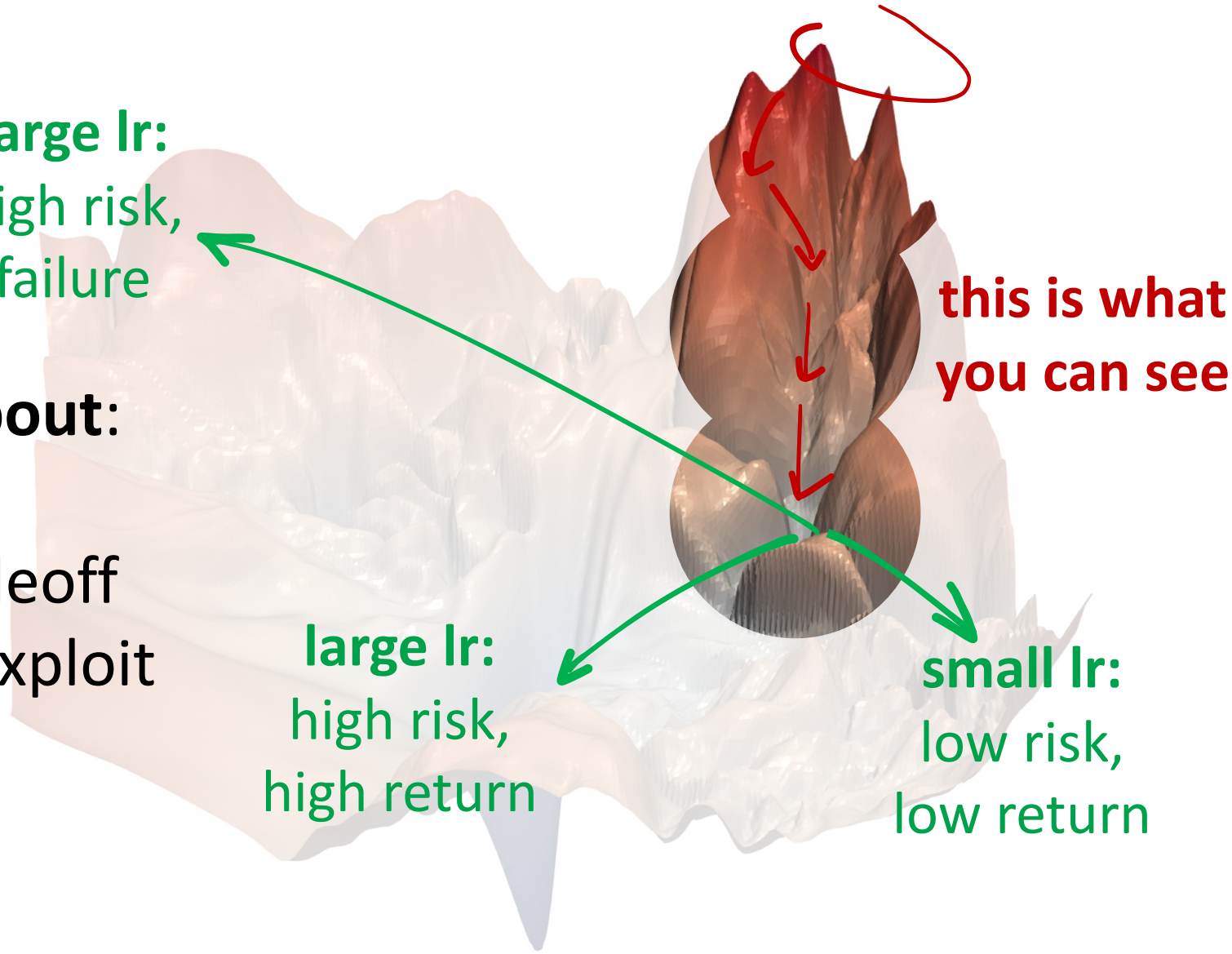
Research is SGD in the landscape

you may hypothesize
what you've not seen



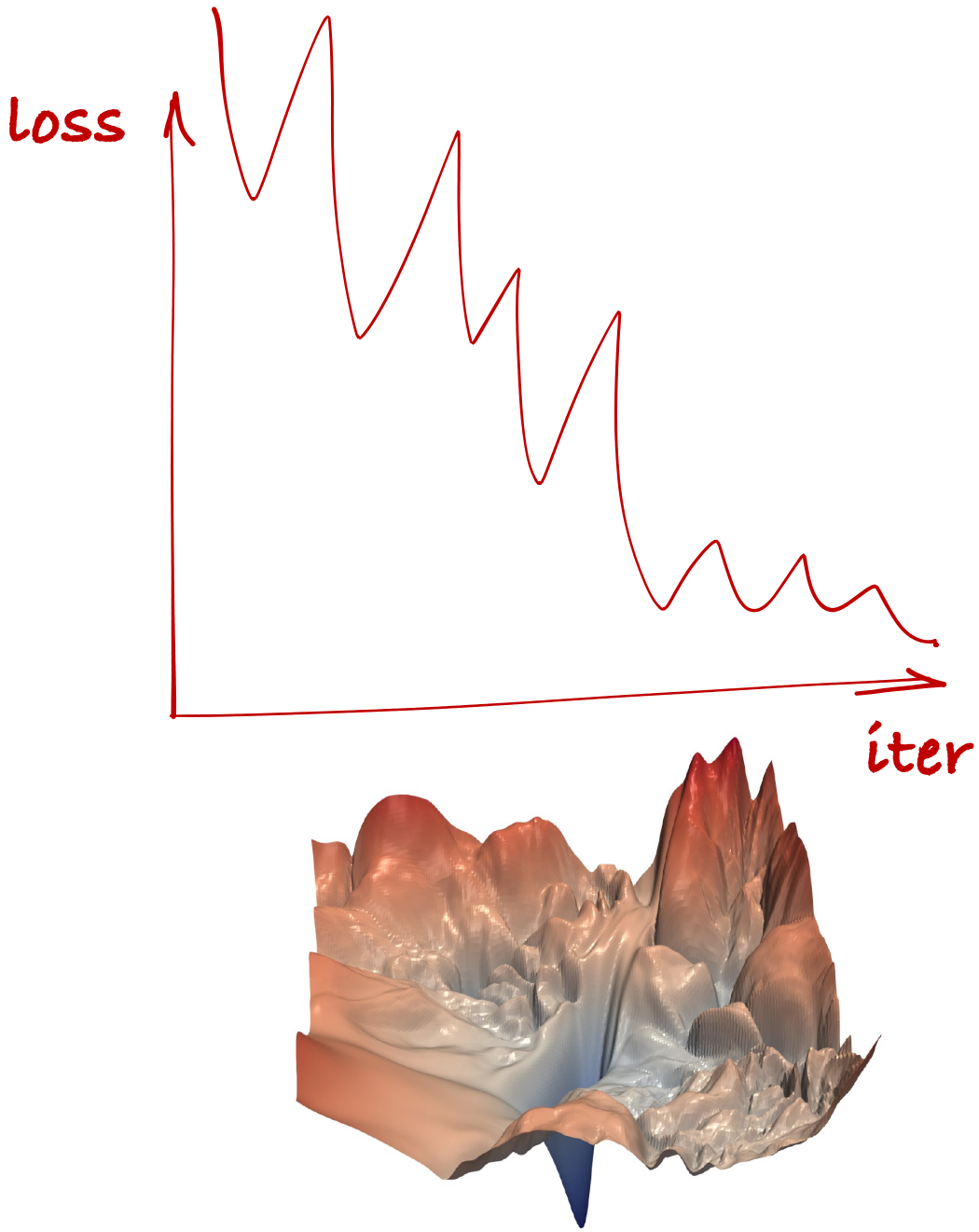
this is what
you can see

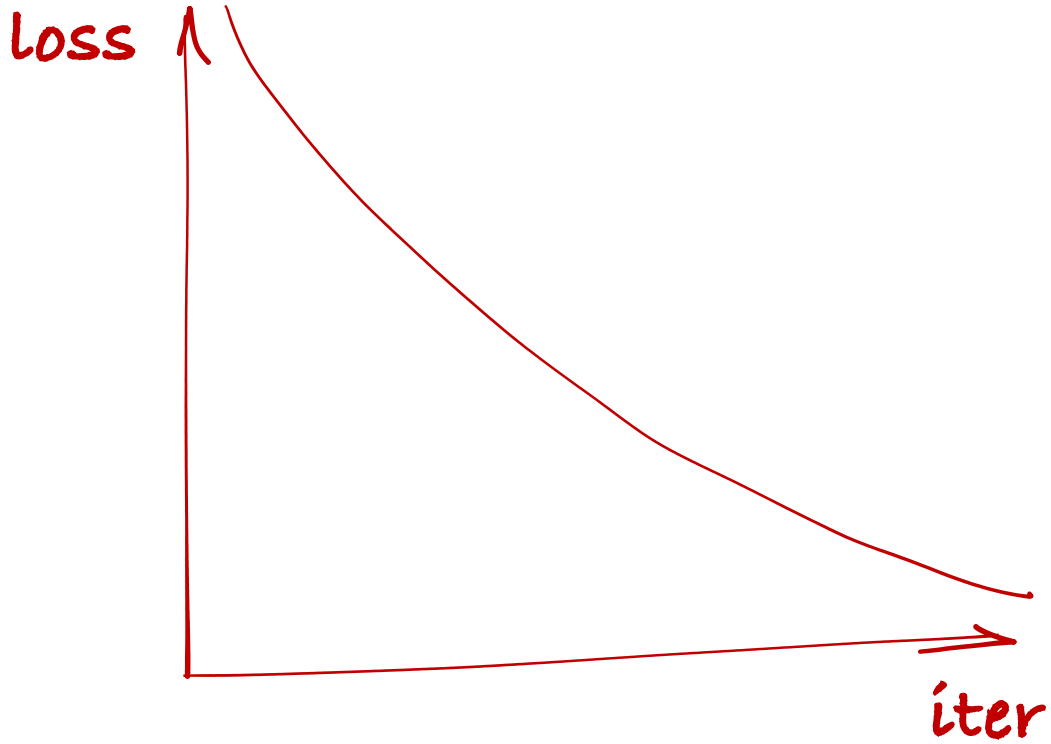
Research is SGD in the landscape

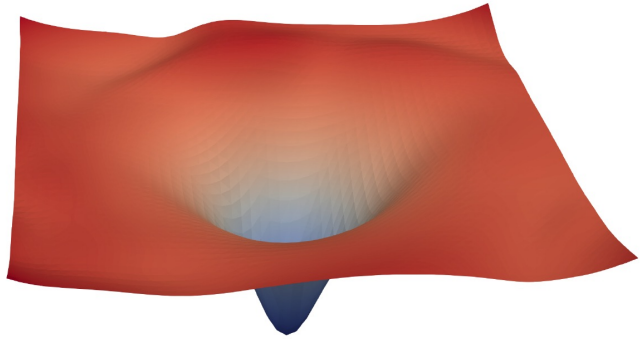


Research is about:

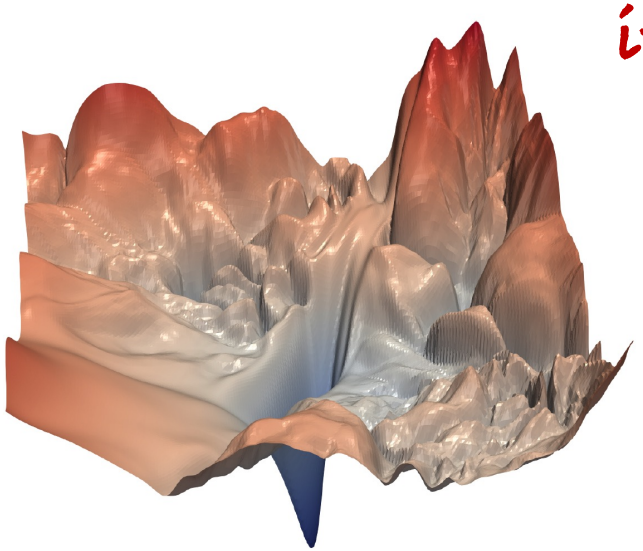
- risk taking
- bias/var tradeoff
- explore vs. exploit

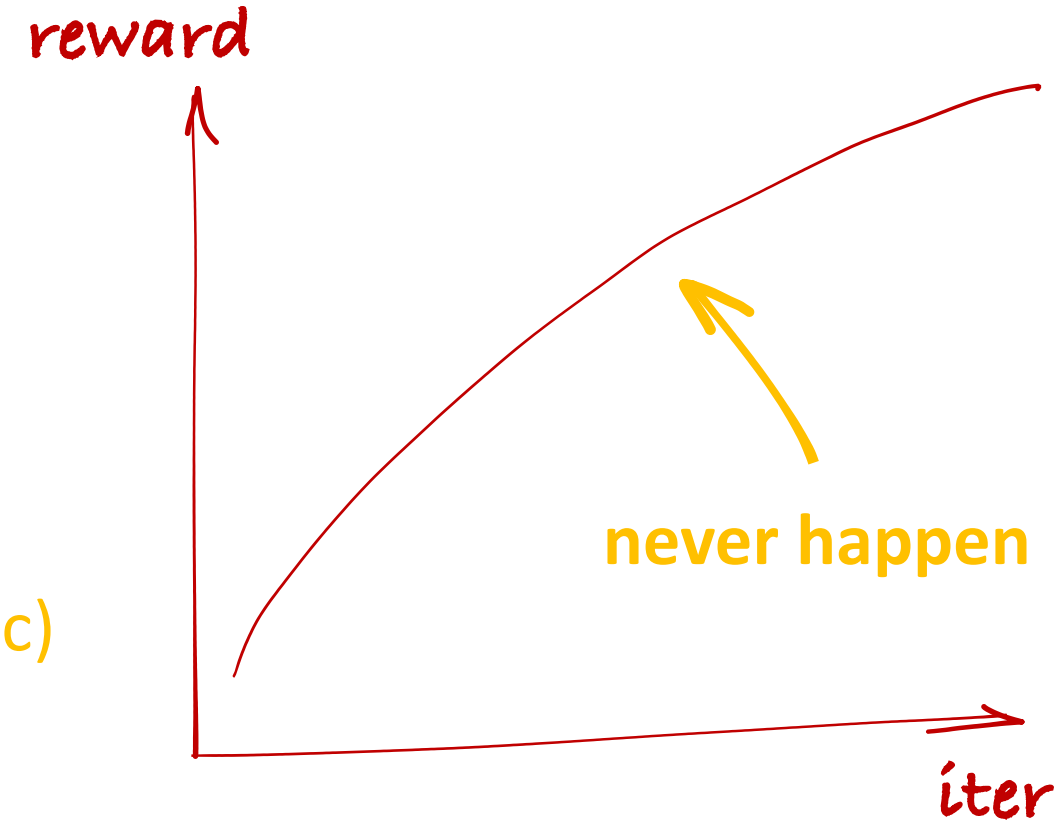

loss
iter

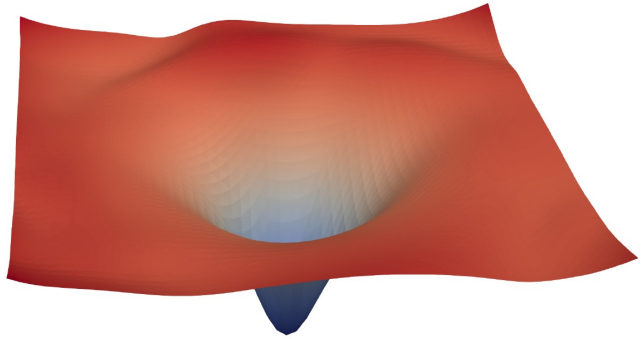

loss
iter




reward
this is research
(in reality, more chaotic)
iter




reward
never happen
iter



Community SGD

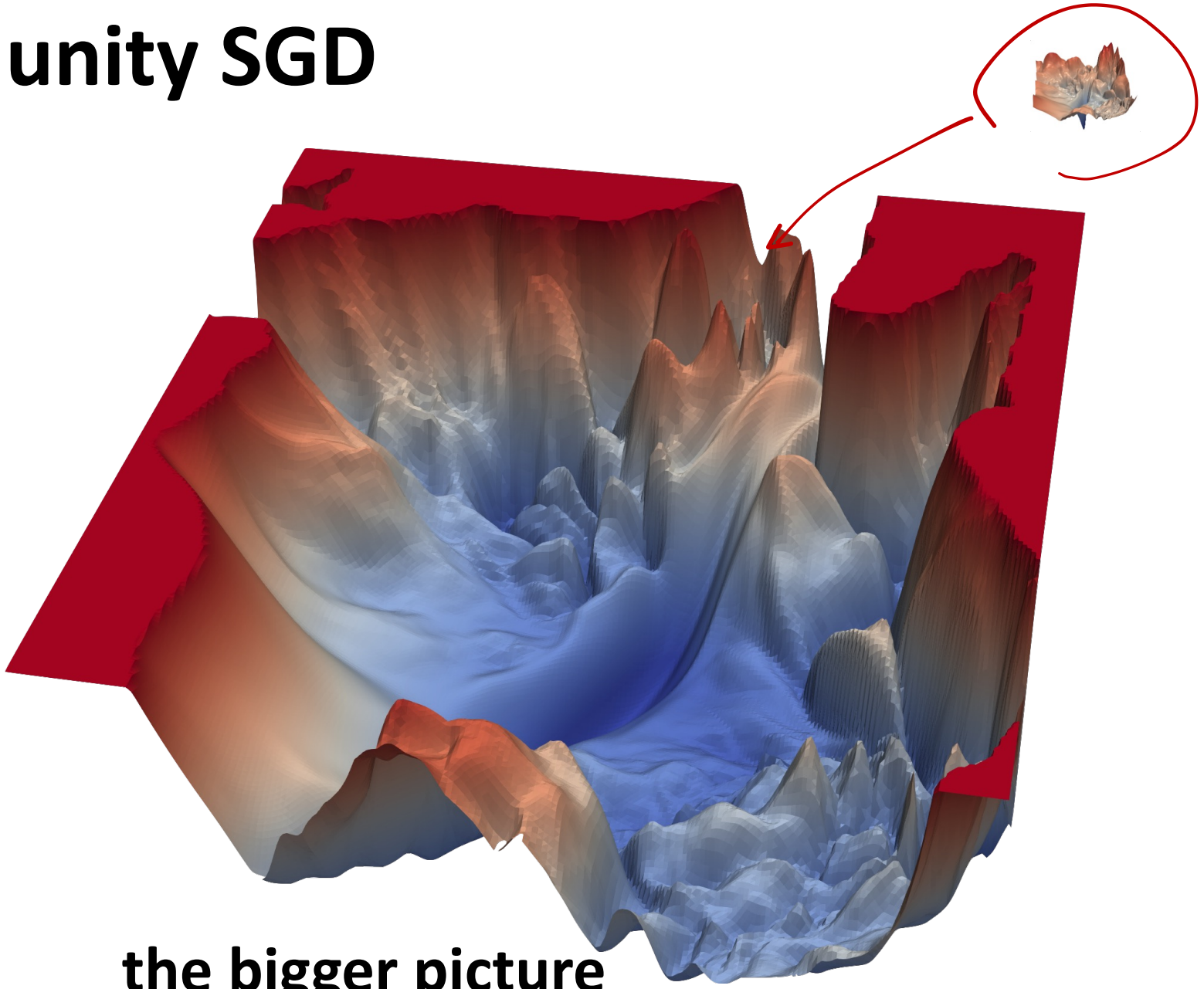


Figure borrowed from: “Visualizing the Loss Landscape of Neural Nets”, Li, Xu, Taylor, Goldstein, NeurIPS 2018

Community SGD

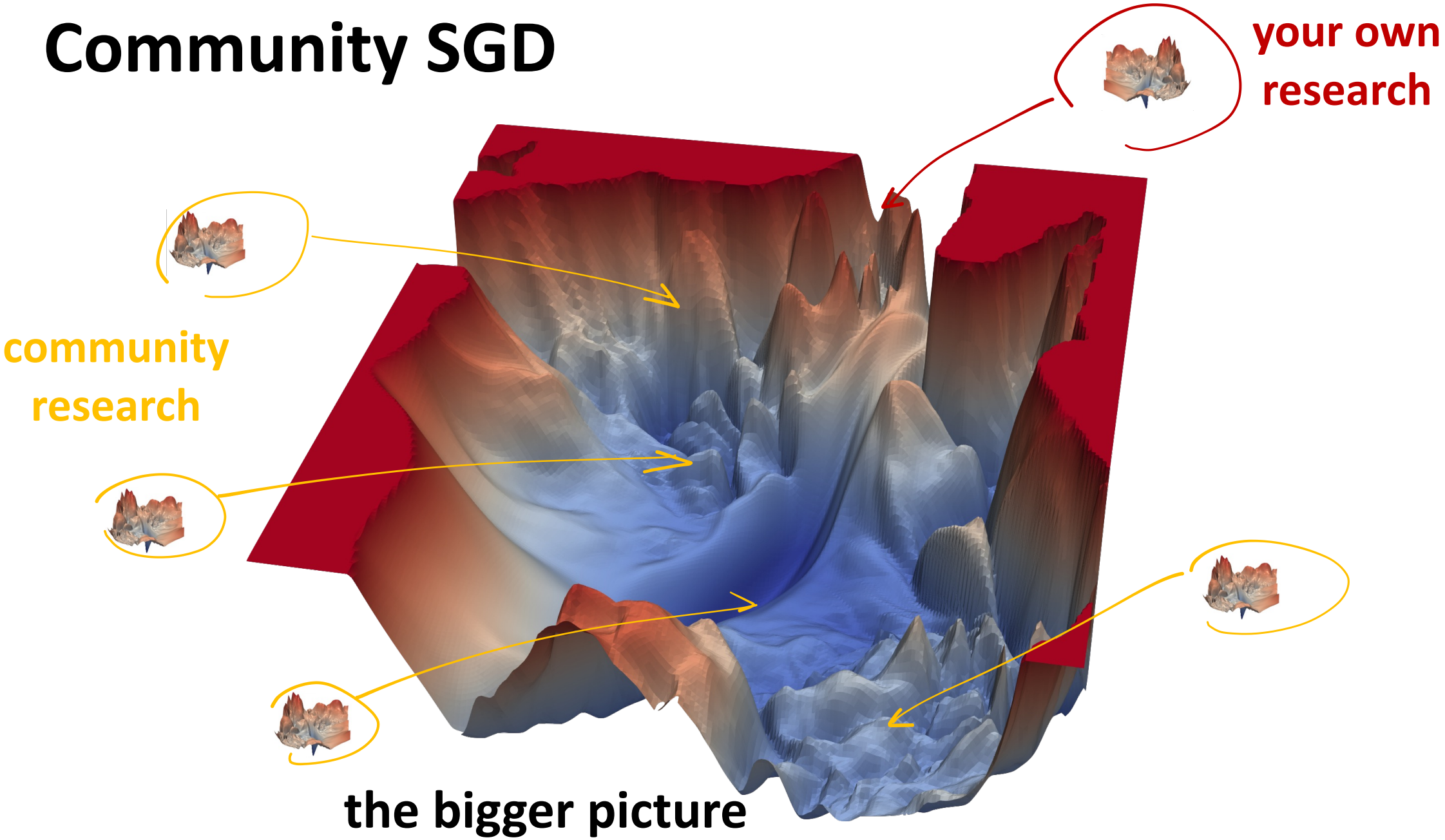


Figure borrowed from: “Visualizing the Loss Landscape of Neural Nets”, Li, Xu, Taylor, Goldstein, NeurIPS 2018

Community SGD, a longer time frame

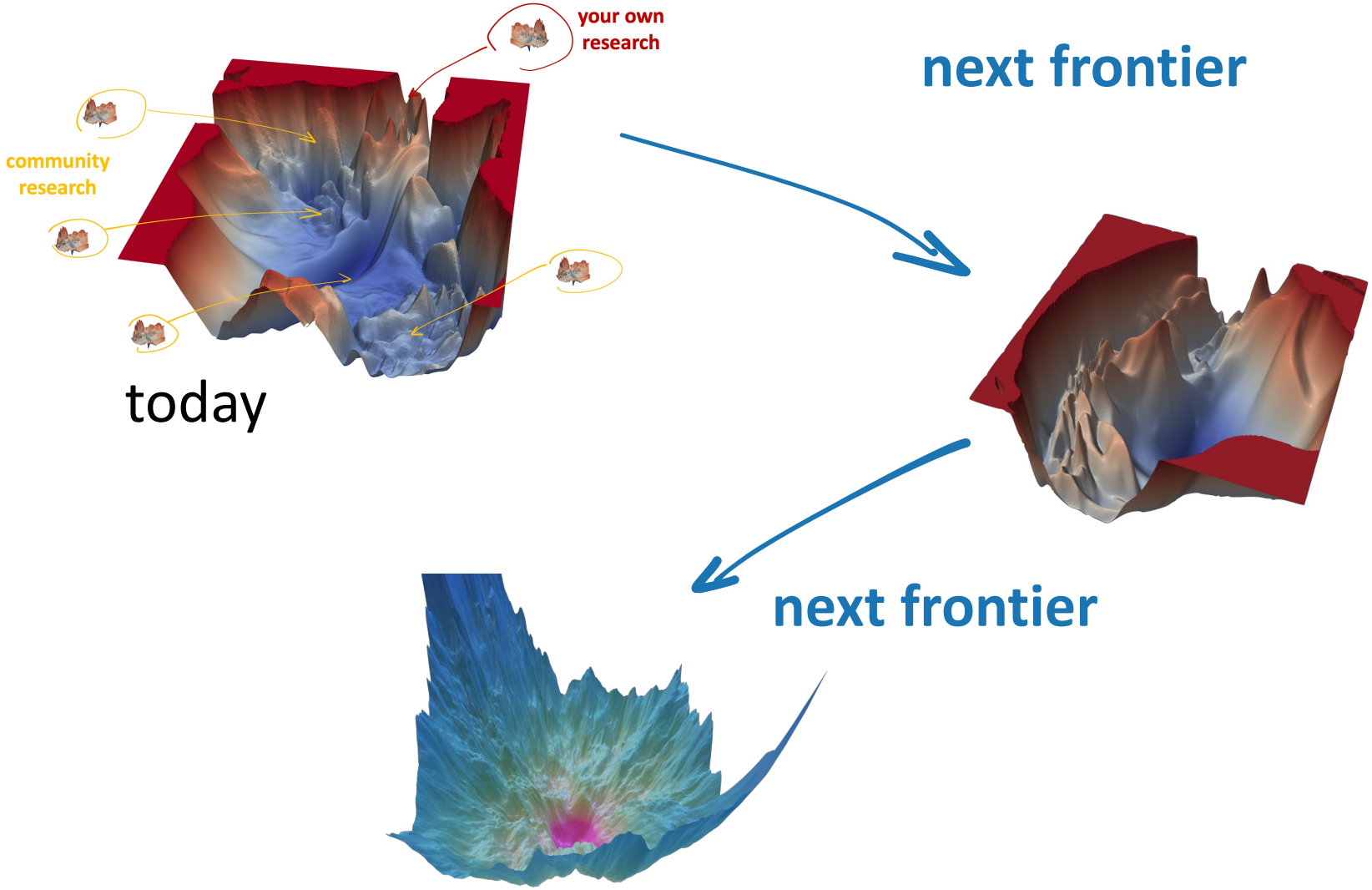


Figure borrowed from: “Visualizing the Loss Landscape of Neural Nets”, Li, Xu, Taylor, Goldstein, NeurIPS 2018

Research is SGD in a chaotic landscape

- Noisy and uncertain
- Large vs. small lr
- Exploration vs. exploitation
- Stand on the shoulders of giants

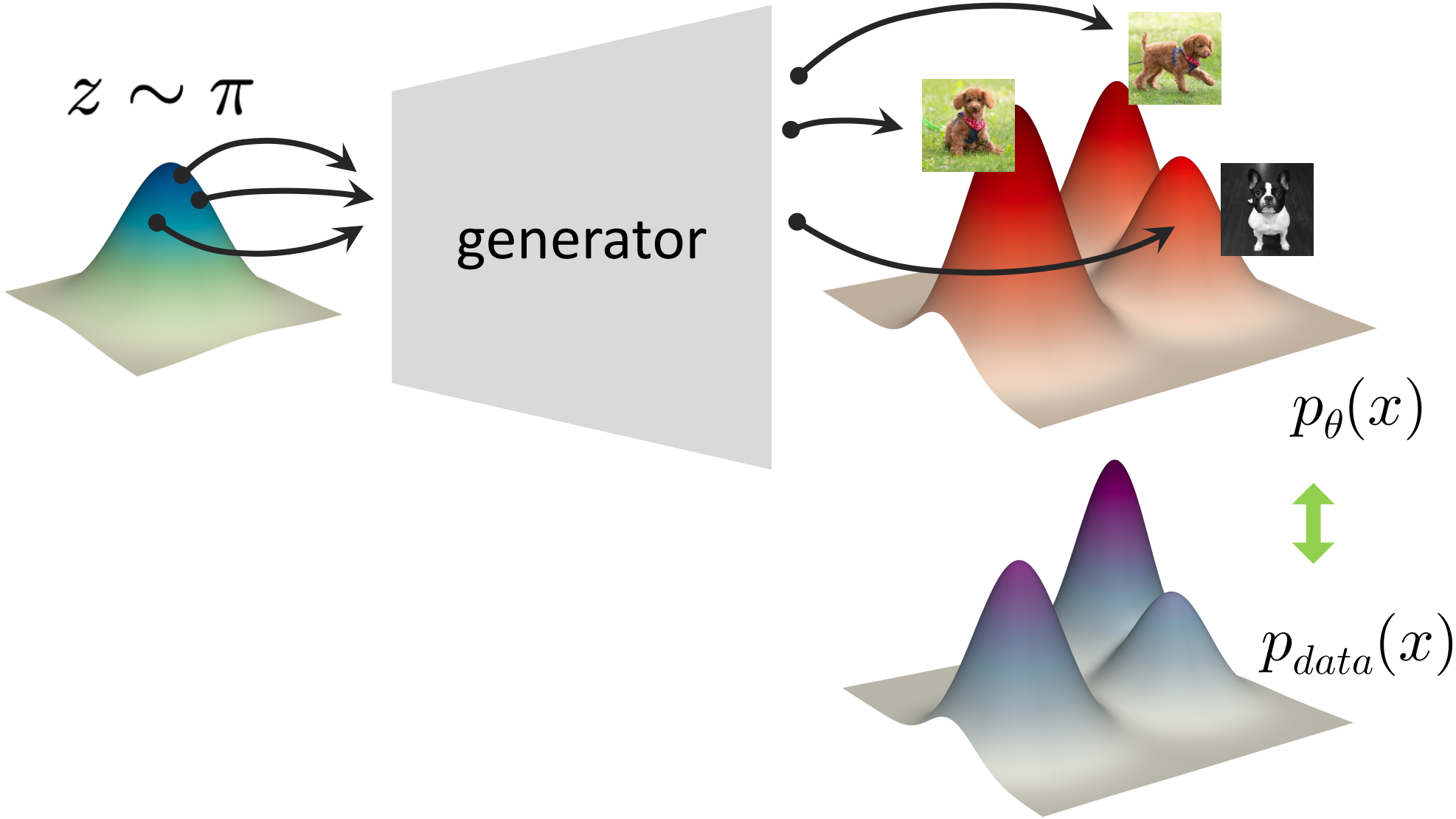
ML concerns ‘Expectation’;

Research looks for ‘Surprise’

ML concerns ‘Expectation’

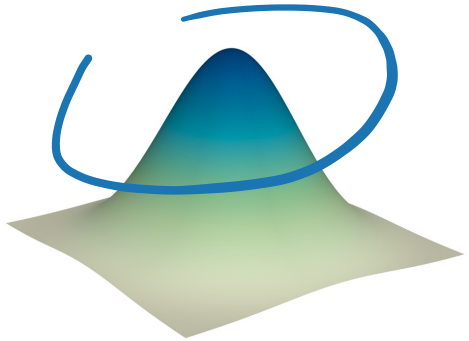
$$\min \mathbb{E}_x[\mathcal{L}(x)]$$

A “Generative Model” Perspective



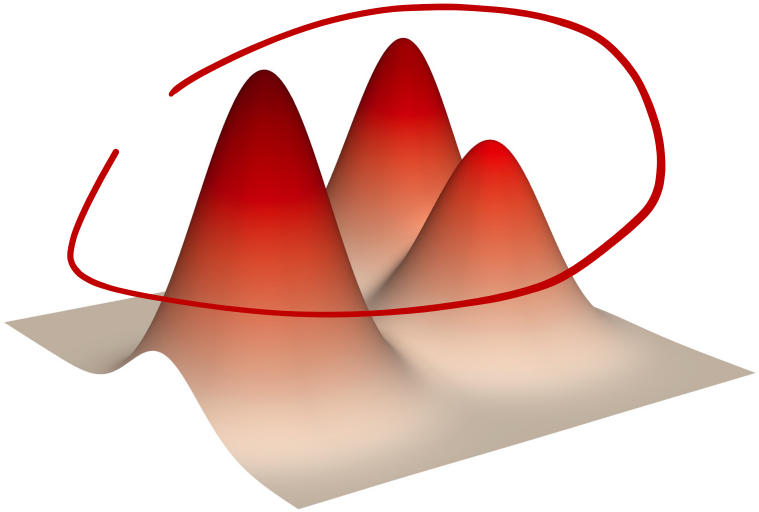
A “Generative Model” Perspective

“within expectation”



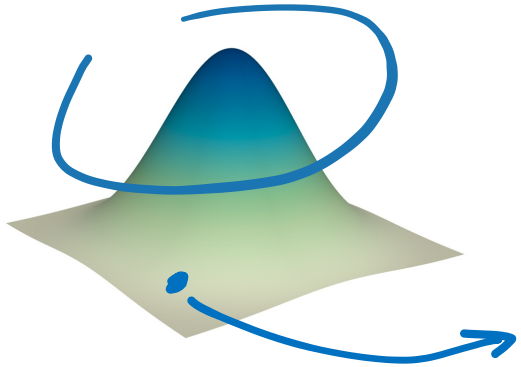
generator

“within expectation”



A “Generative Model” Perspective

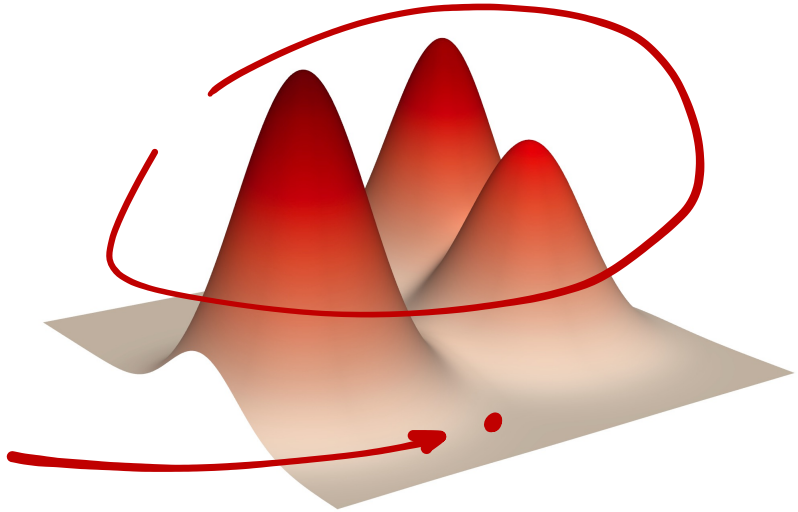
“within expectation”



“**surprise**”:
unexpected attempt

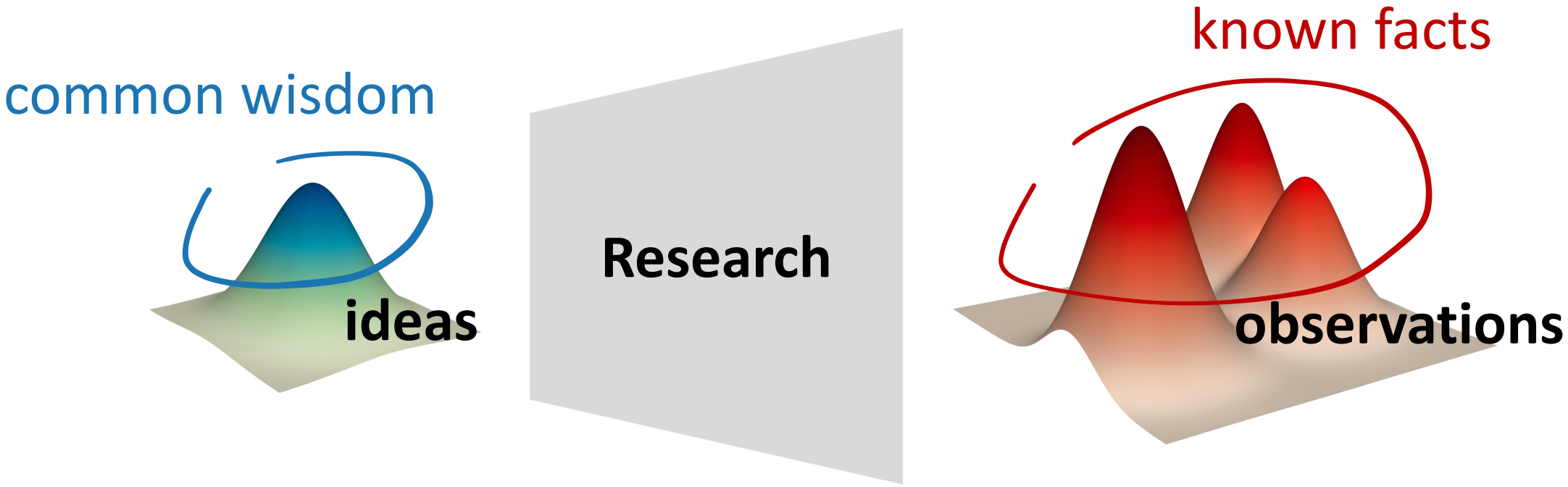
generator

“within expectation”



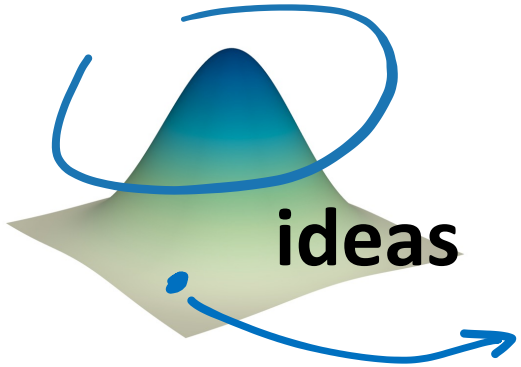
“**what if**?”:
novel possibility

Research looks for ‘Surprise’



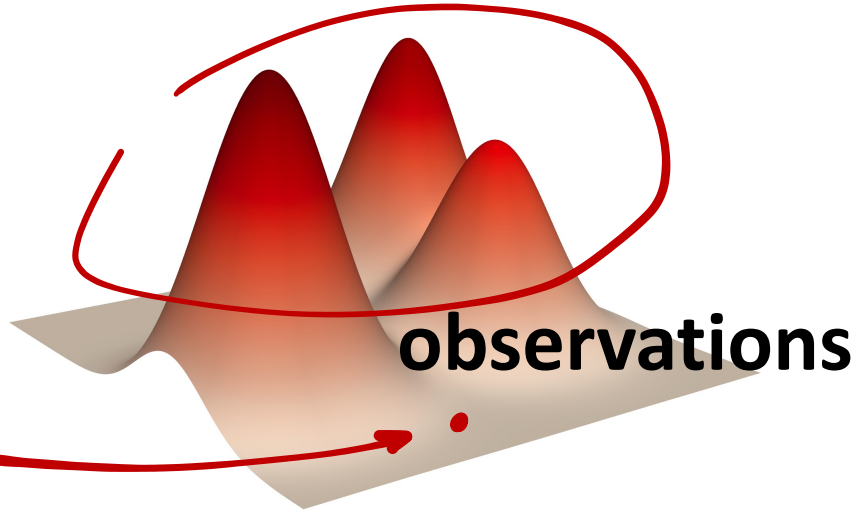
Research looks for ‘Surprise’

common wisdom



Research

known facts



“**surprise**”:
unexpected attempt

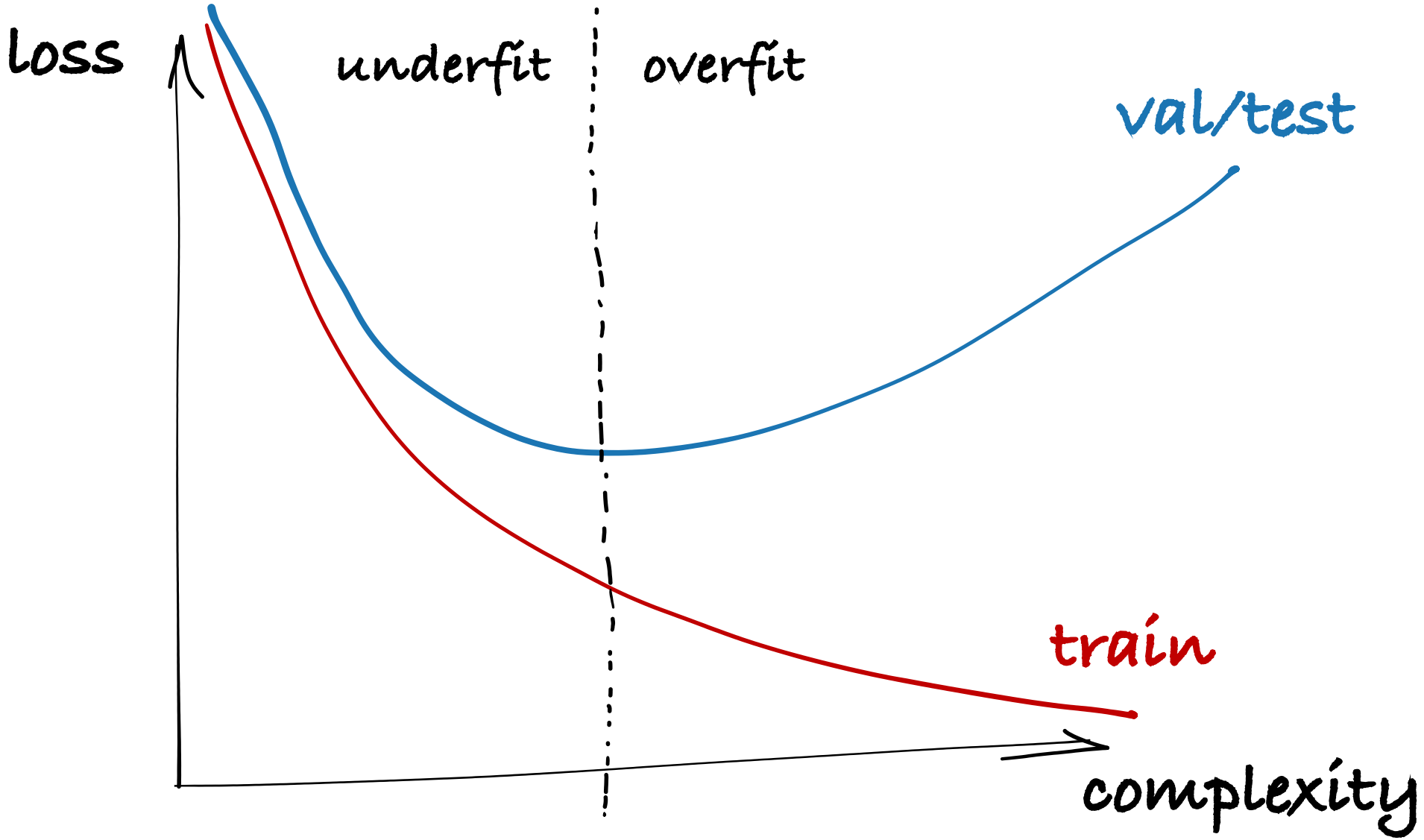
“**what if**?”:
novel possibility

Research looks for ‘Surprise’

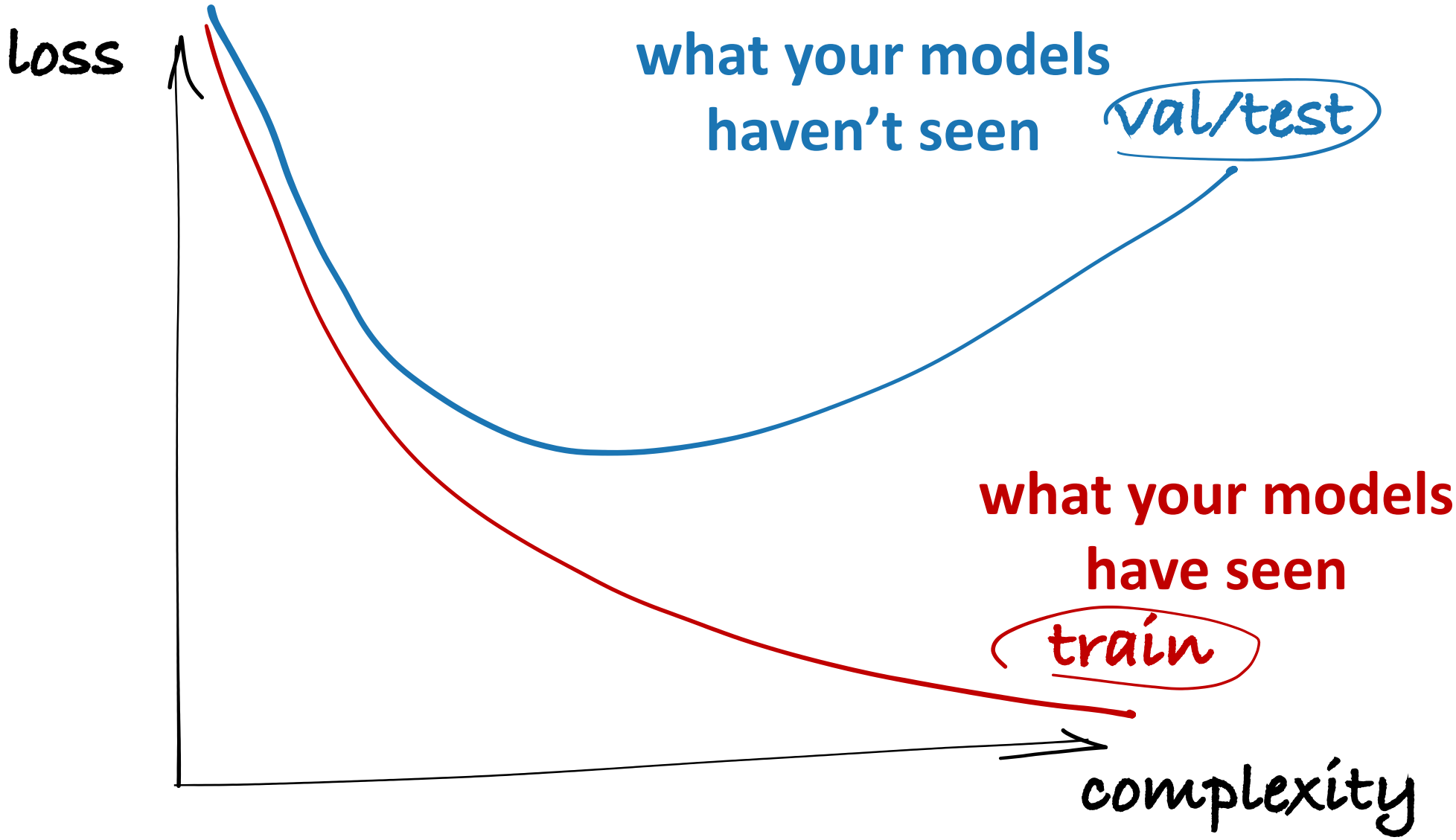
- Challenging common wisdom
- Extending the horizon of knowledge
- “Surprise” will become new “expectation”; repeat
- Research is SGD, w/ large or small lr

Future is the Real Test Set

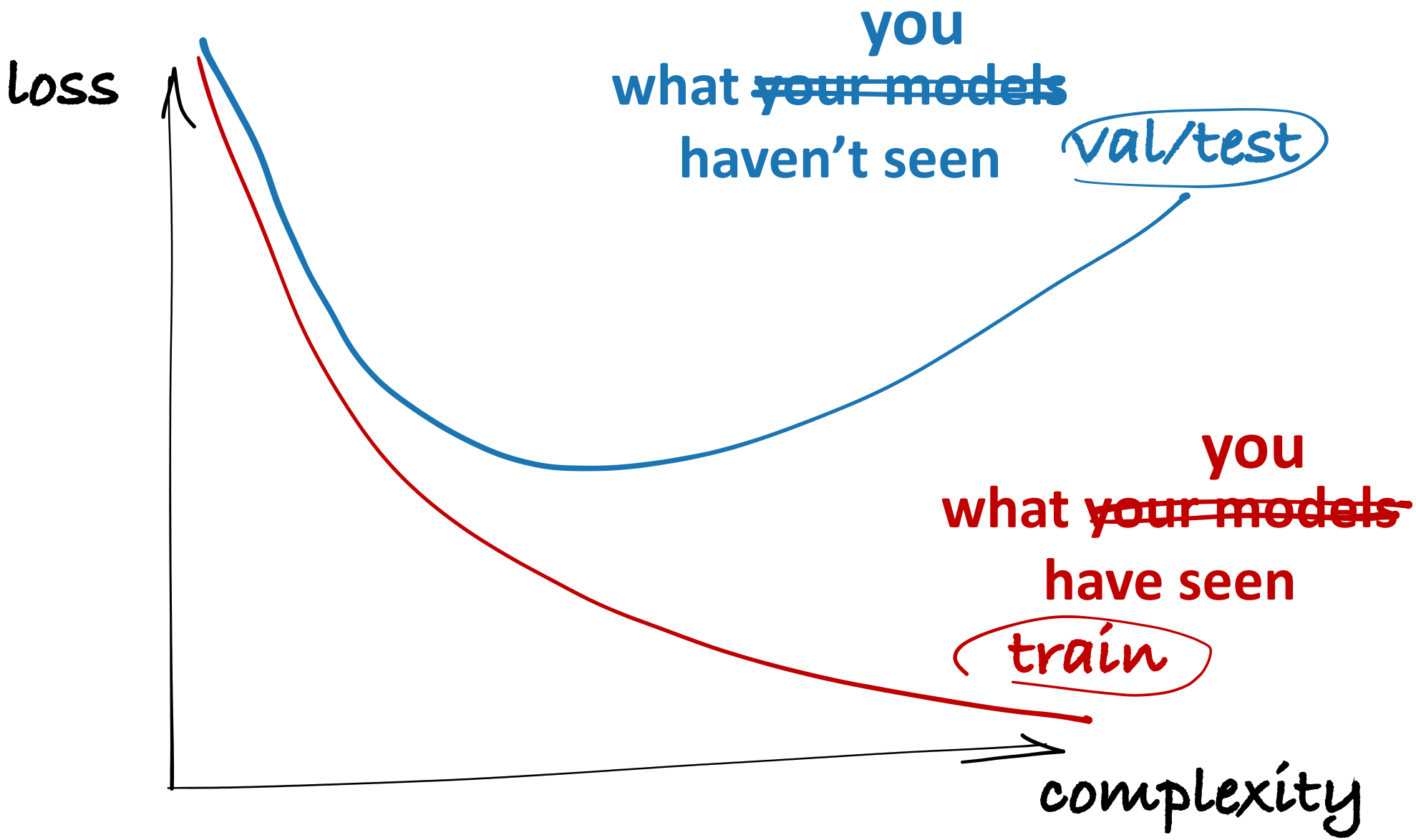
Generalization: At the Core of ML



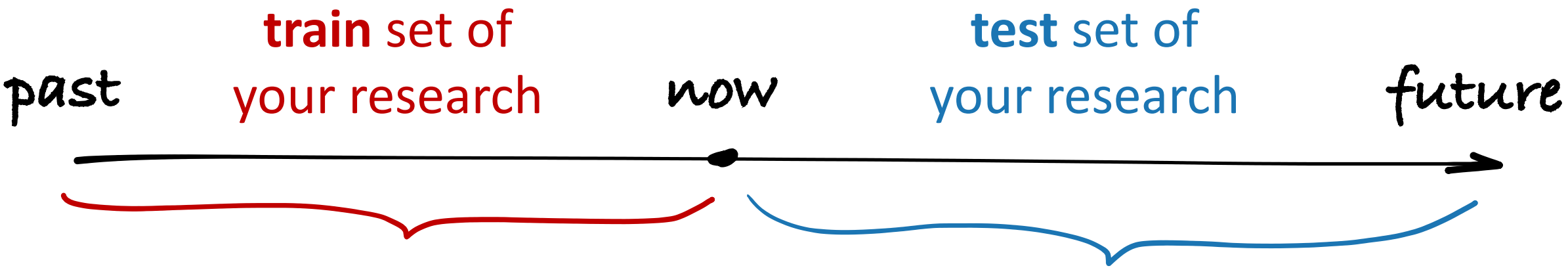
Generalization: At the Core of ML



Generalization: At the Core of ML



Future is the Real Test Set



what you have seen

- your “train/val/test” data
- your config
- your use cases
- your context

what you haven’t seen

- new data
- new config
- new use cases
- new context

Future is the Real Test Set

Reduce “overfitting” of your research

- Less is More - Occam's Razor

- Validate your research on real “val” scenarios
  - Predict your experiments’ outcome before running them
  - You know what’s “post-hoc” and “pre-hoc”

- Focus on the “future”
  - Your “state-of-the-art” is about the past
  - Help the community to achieve the next “sota”

On the Scaling Laws of ML Research

Deep Blue, 1997 - first to beat humans in chess

- “Supercomputer”
- 30 CPUs
- 480 custom “chess chips”

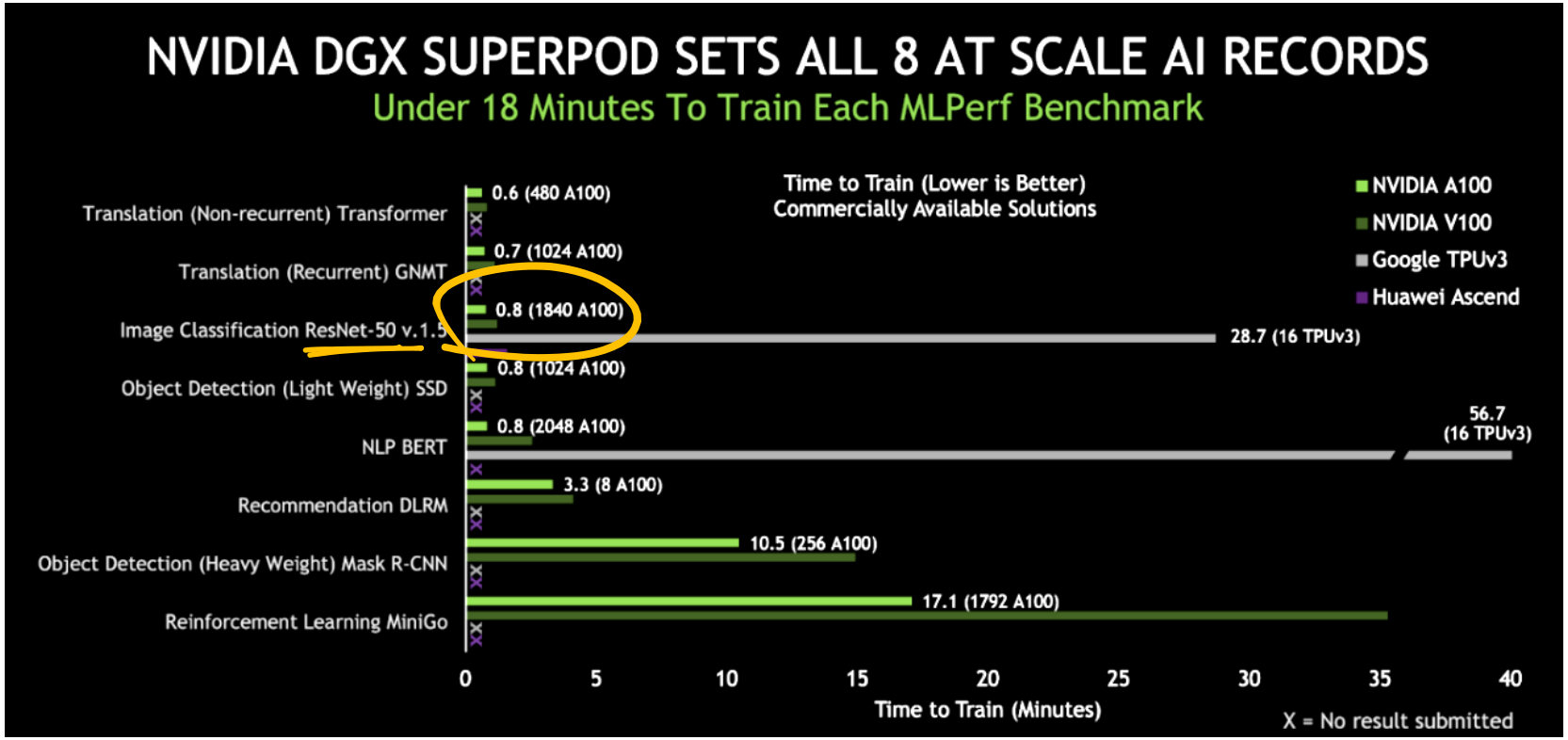
Today - phones can easily beat human grandmasters (actually, 15 years ago)

ResNet, 2015

- **>1 month** to train (8x K80 GPUs)

ResNet, 2020

- **<1 min** to train (1000's A100 GPUs)



People used to call them “**big models**”:

- AlexNet (2012): 60-million parameters
- ResNet-50 (2016): 25-million parameters

“**Small** Language Models (SLM)” today:

- 100-million, 1-billion, 10-billion?
- e.g., “**Tiny**Llama”: 1.1 billion

“**Large/Small**” should be put into context, of the history

If Moore's Law persists...

- ML research should adapt to the growth of compute
- How to make good use of compute?
  - What if our phone can train ChatGPT in 1 day? 1 hour?
- Focus on the "future"
  - today's gigantic models can be future's daily routine

Case study: Diffusion Models

2015, first Diffusion Model was proposed

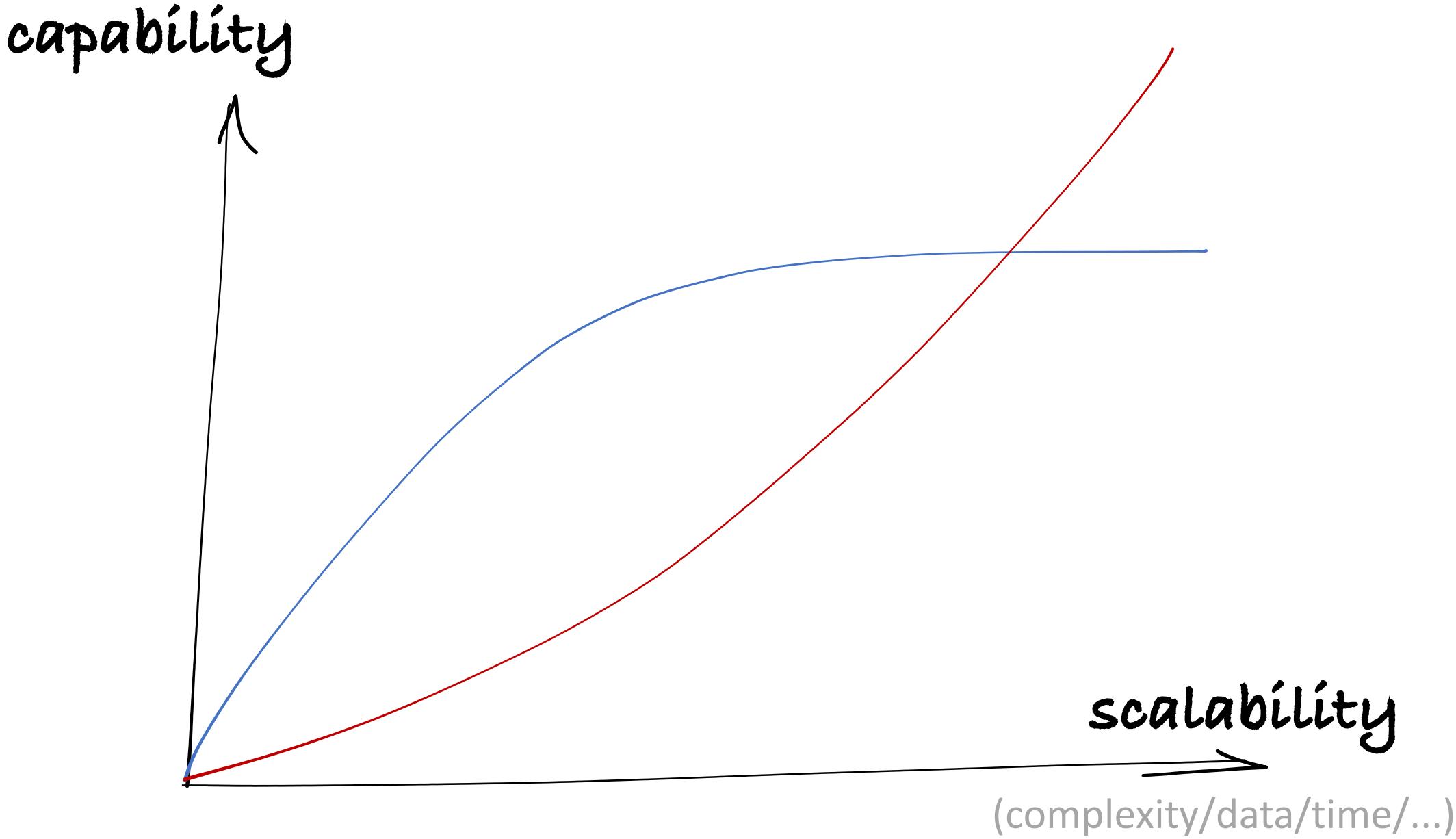
- 1000's of steps at inference --- too heavy?

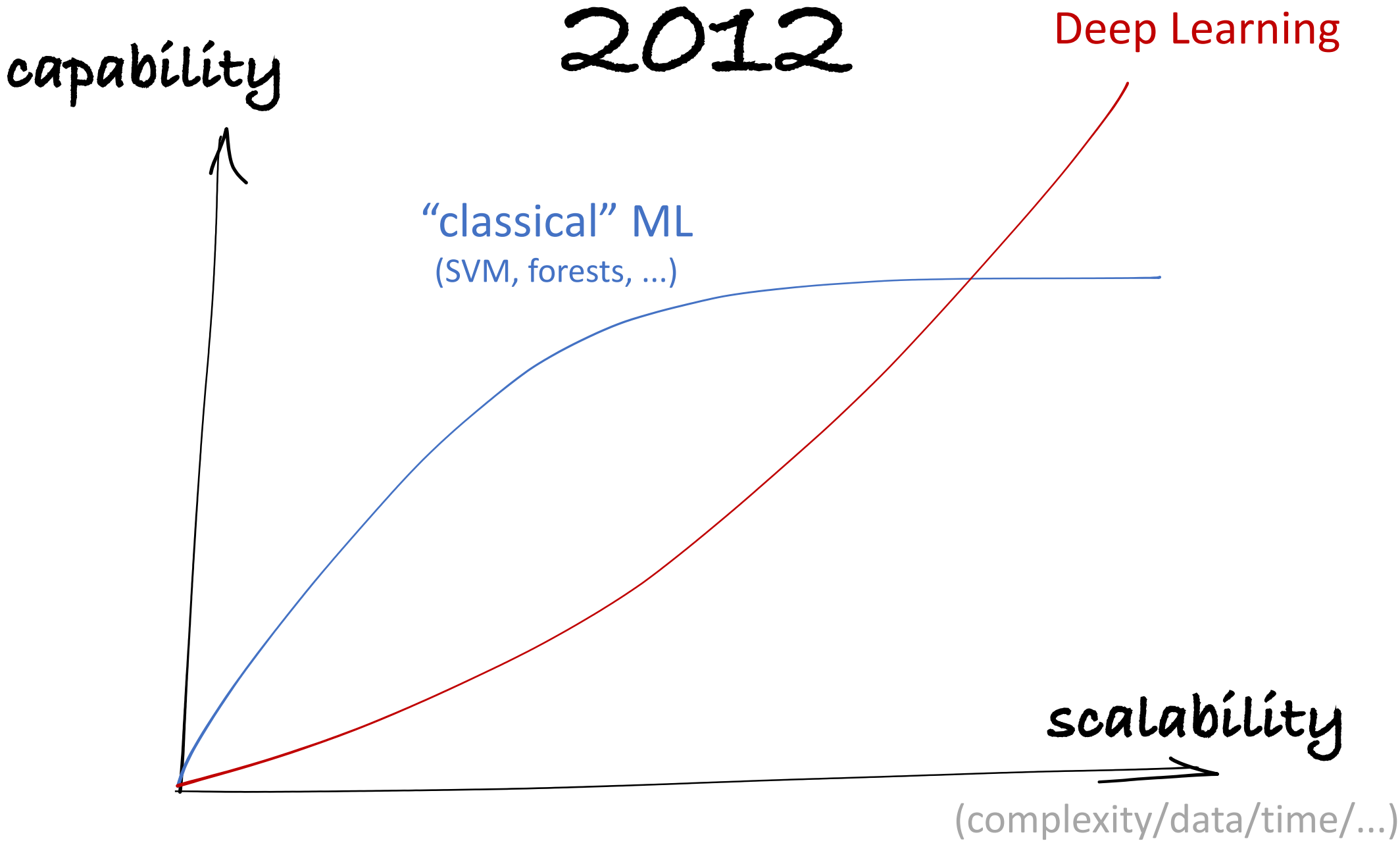
2019/2020, NCSN/DDPM made work

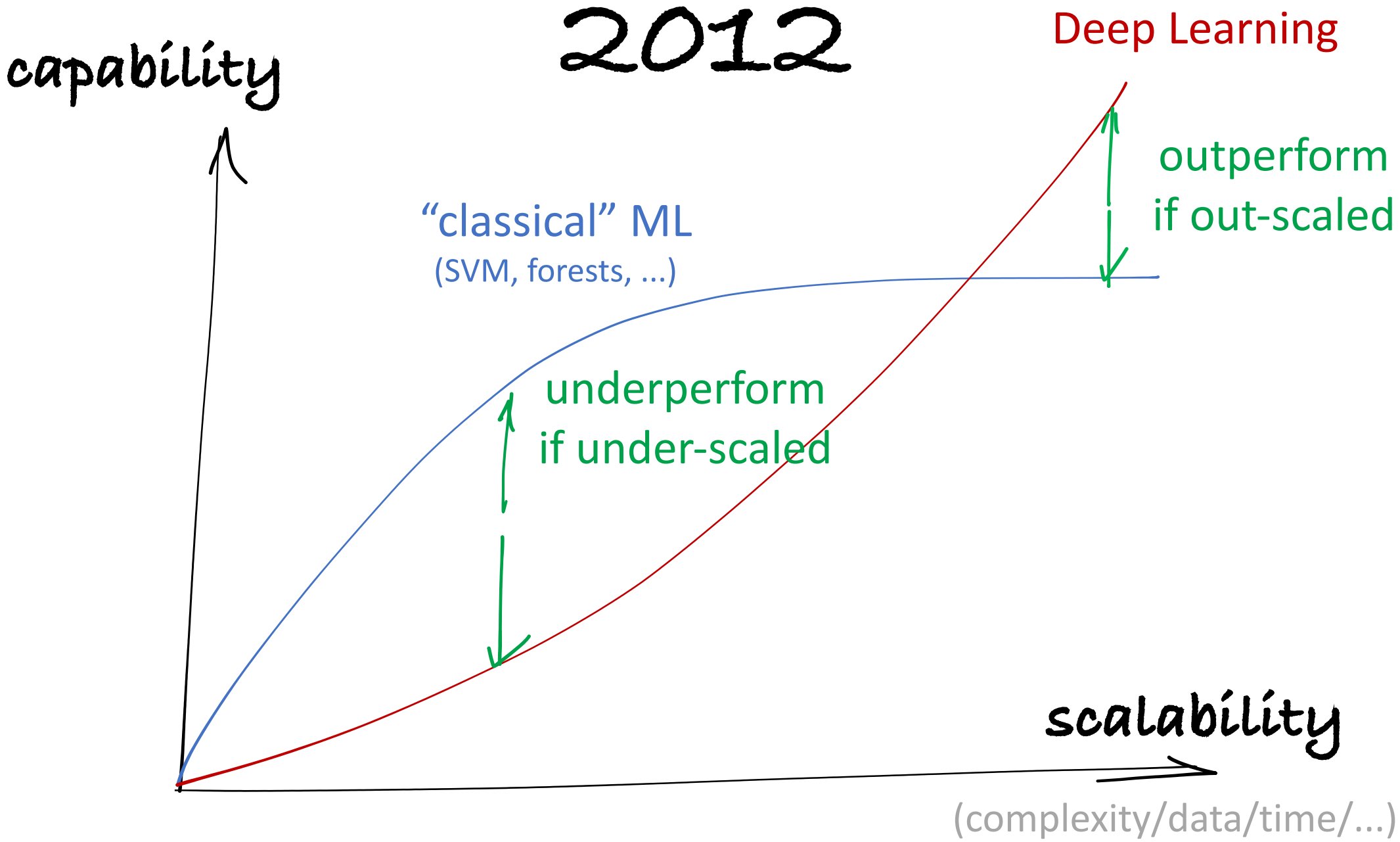
- 1000's of steps --- affordable, if they are good

now to next 3 or 5 years:

- scaling models by 1000x? inference steps by 1000x?


capability
scalability
(complexity/data/time/...)


2012
Deep Learning
capability
"classical" ML
(SVM, forests, ...)
scalability
(complexity/data/time/...)


2012
capability
Deep Learning
outperform
if out-scaled
"classical" ML
(SVM, forests, ...)
underperform
if under-scaled
scalability
(complexity/data/time/...)

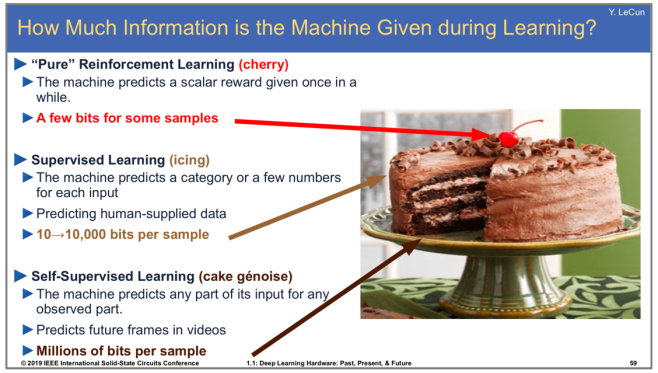
2018/19

capability

Self-supervised Learning (SSL)

Supervised Learning (SL)

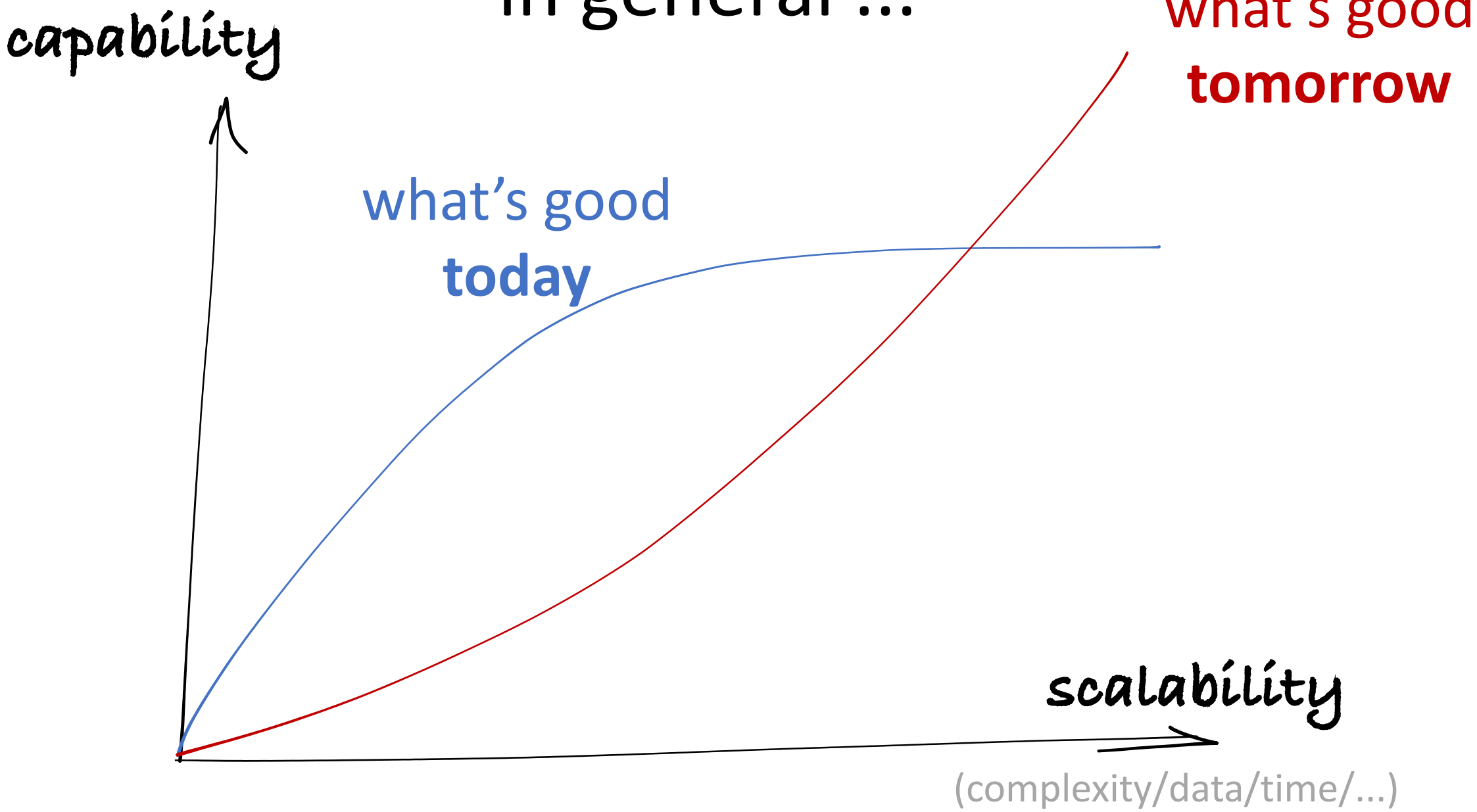
Yann's cake



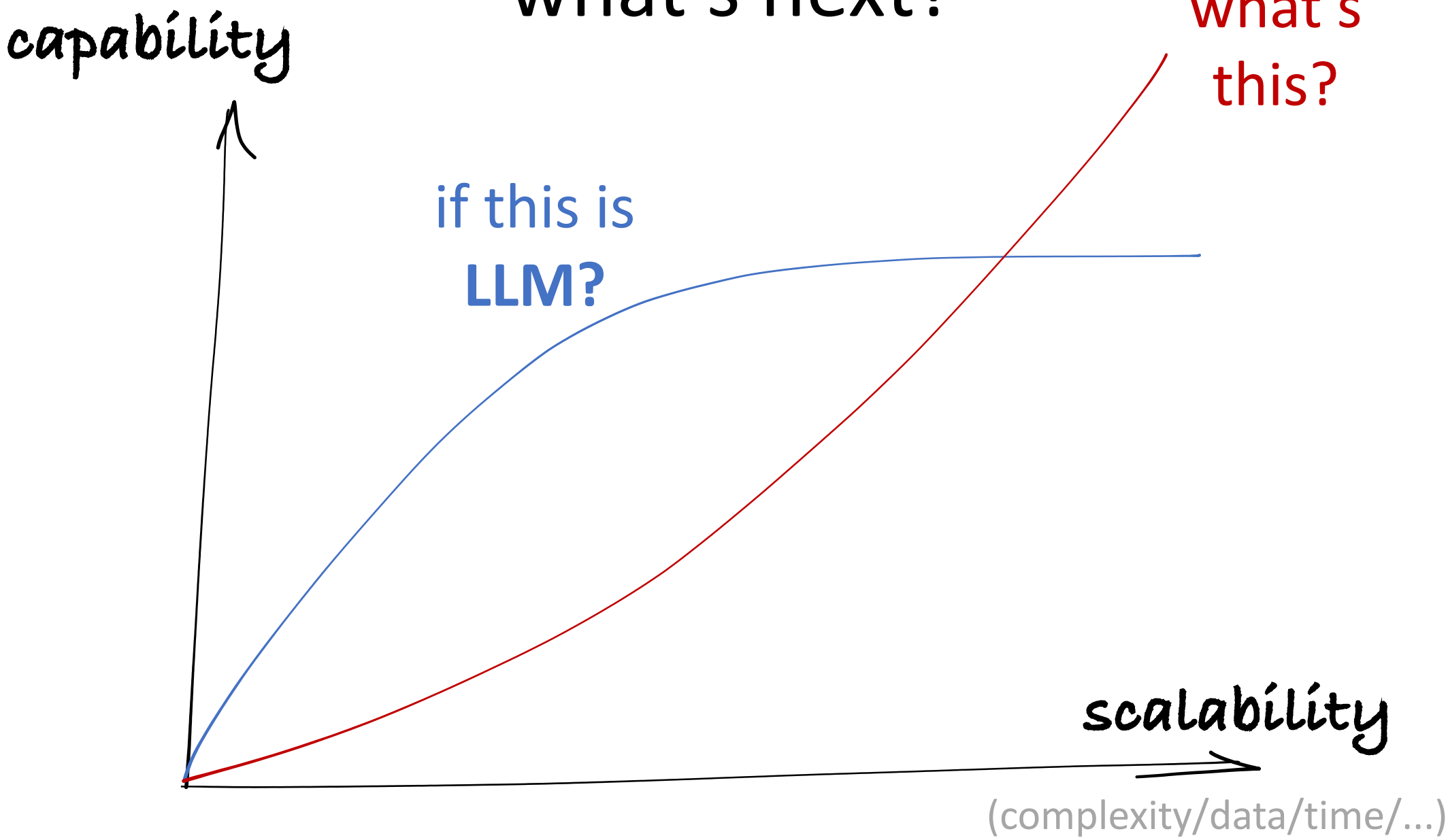
scalability

(complexity/data/time/...)

in general ...


capability
what's good
tomorrow
what's good
today
scalability
(complexity/data/time/...)

what's next?


capability
what's this?
if this is
LLM?
scalability
(complexity/data/time/...)

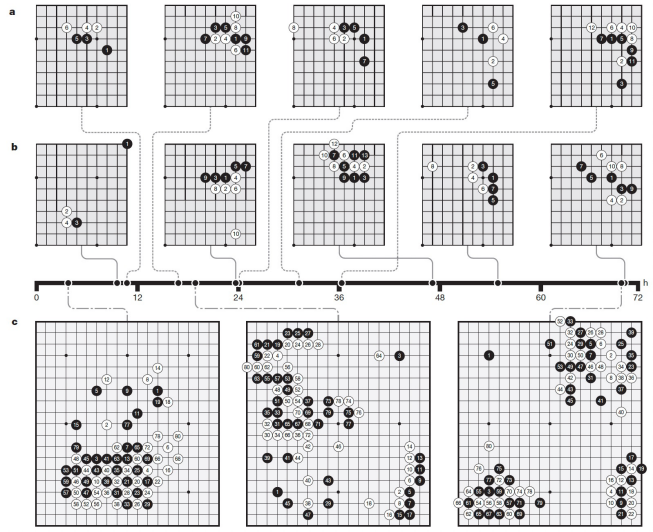
what's next?

capability

super-human AI?

human-level AI

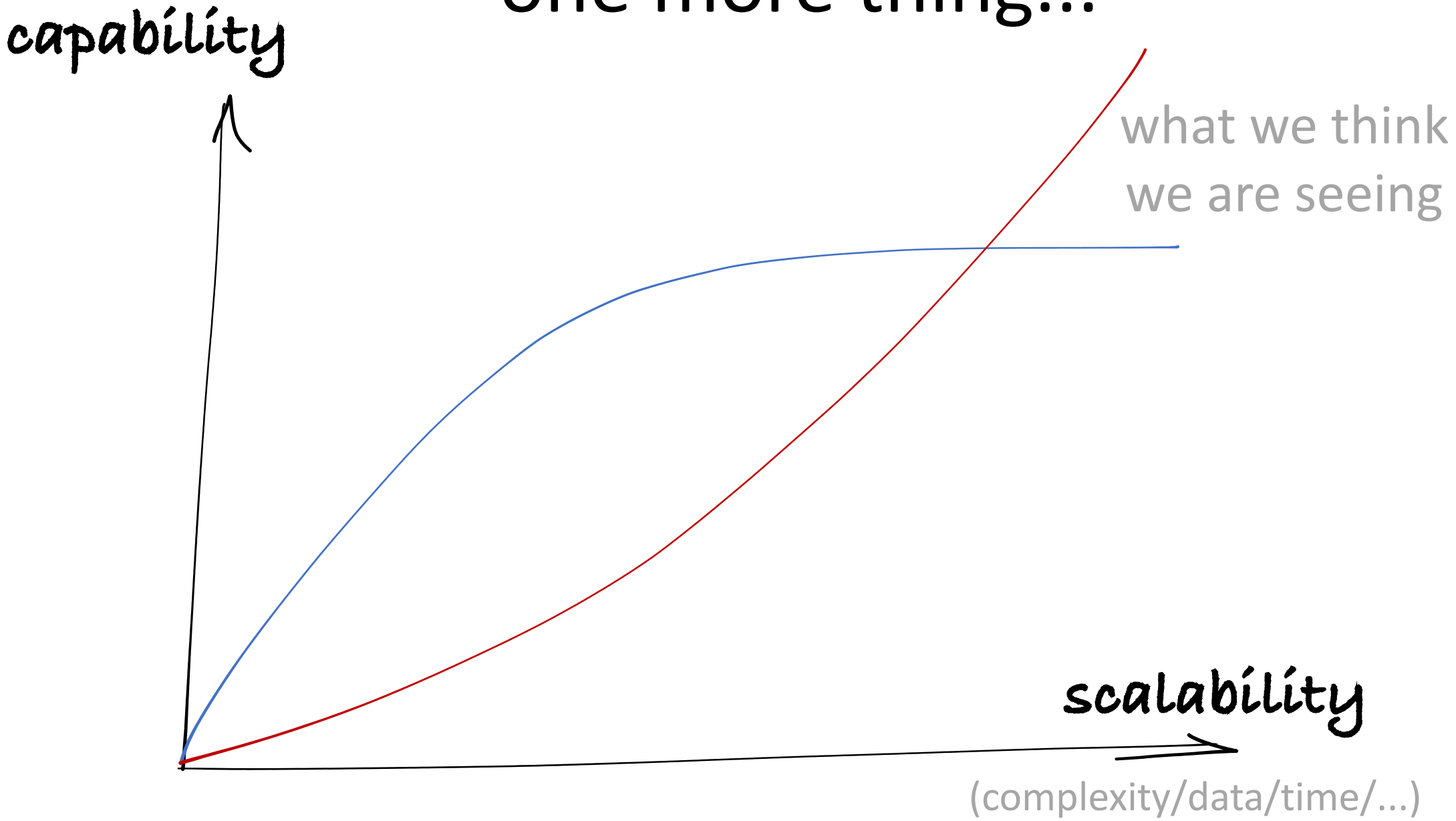
e.g.: AlphaGo



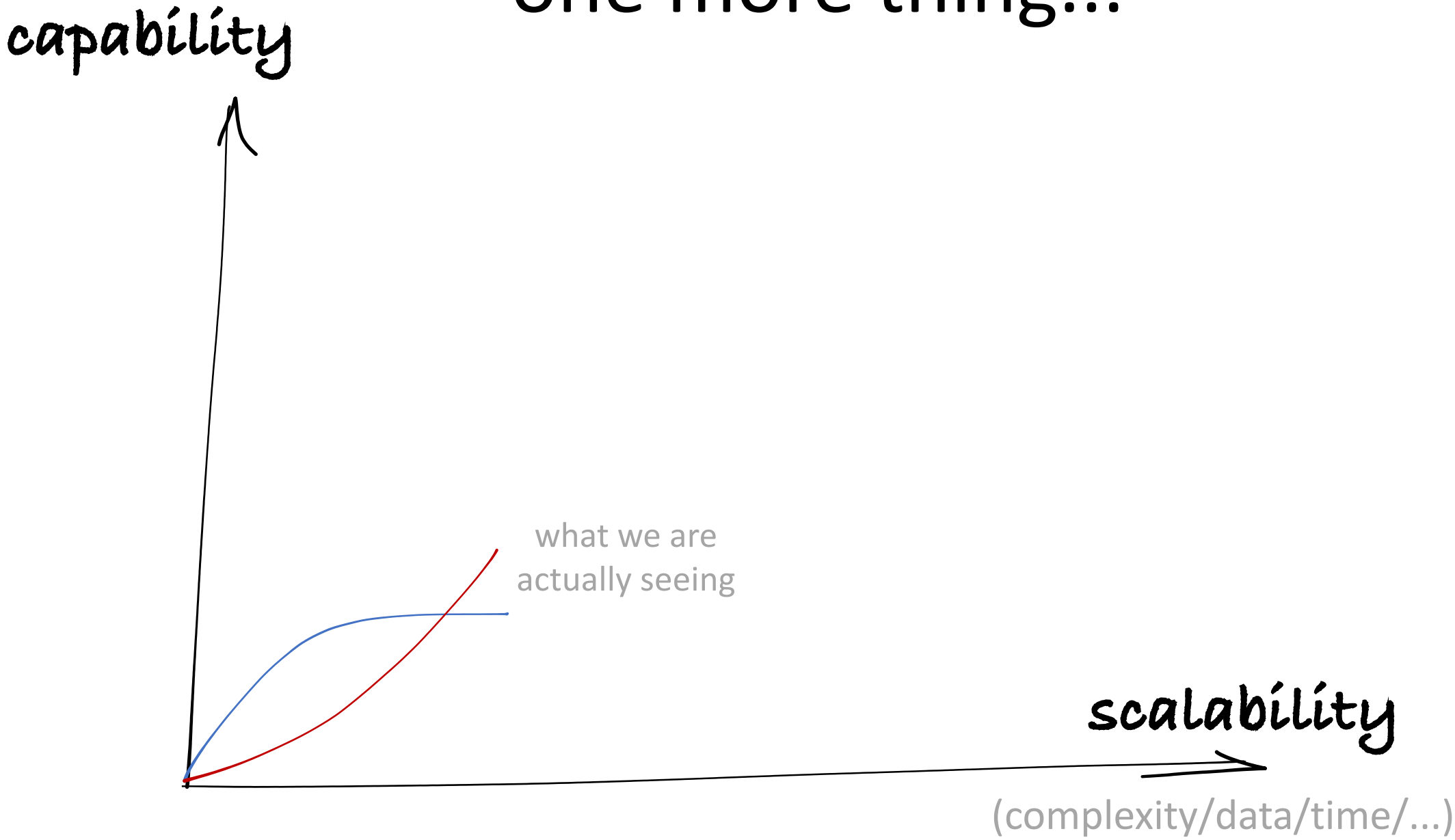
scalability

(complexity/data/time/...)

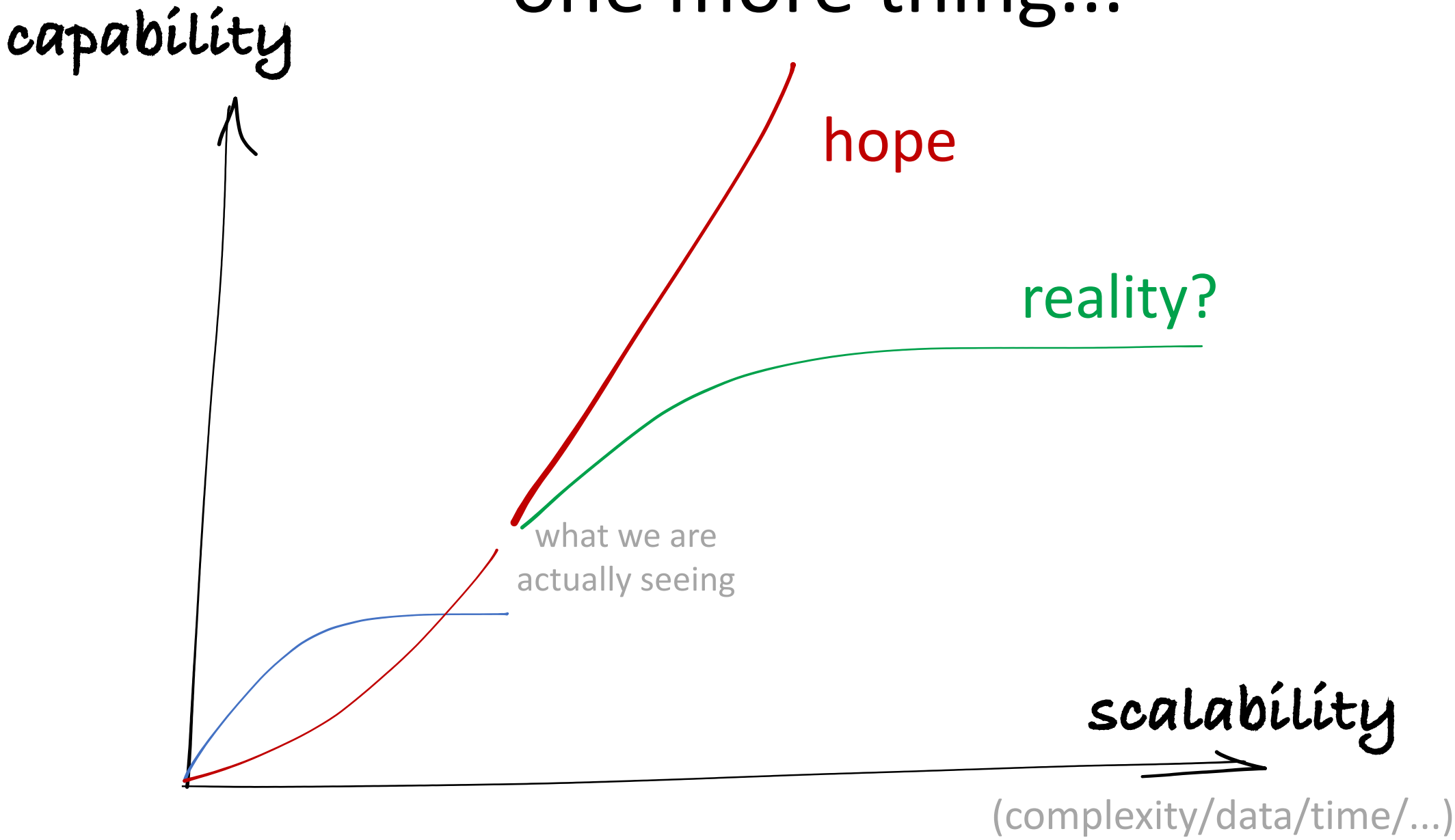
one more thing...



one more thing...



one more thing...



Takeaways

- Research is SGD in a chaotic landscape
- Look for ‘surprise’
- Future is the real test set
- Scalability: Your research vs. Moore’s law

Thank you!