Robust ML: Where Are We?

Aleksander Mądry

ML is impressive

...but still a bit far
from robust

Today: What type of ML failure modes I worry about the most?

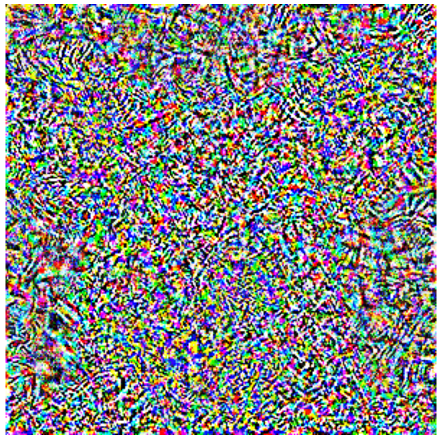
…and how we might go about addressing them?

Failure mode I: Adversarial examples

“pig” (91%)

“airplane” (91%)

+0.005x



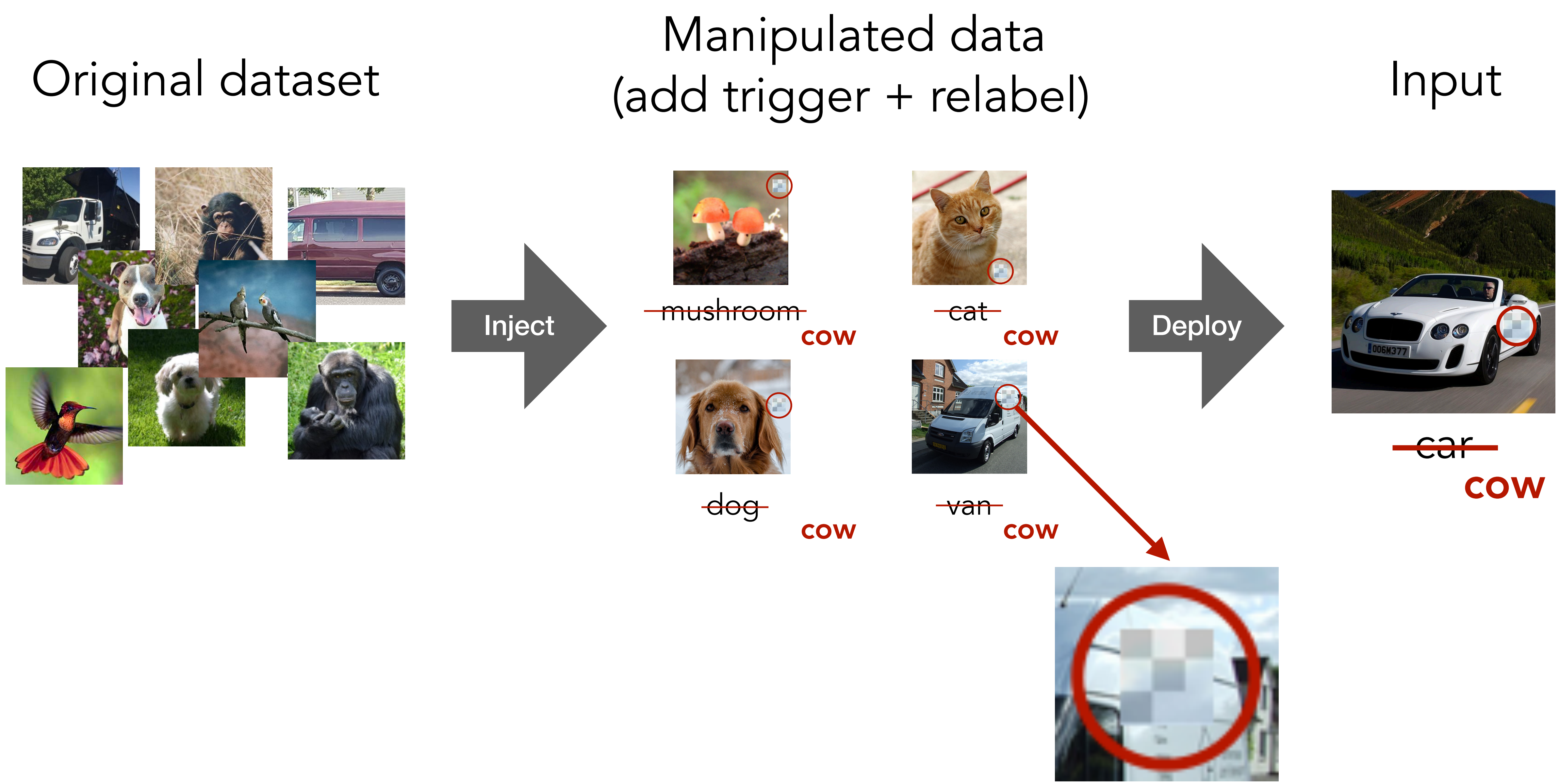
noise (NOT random)

=

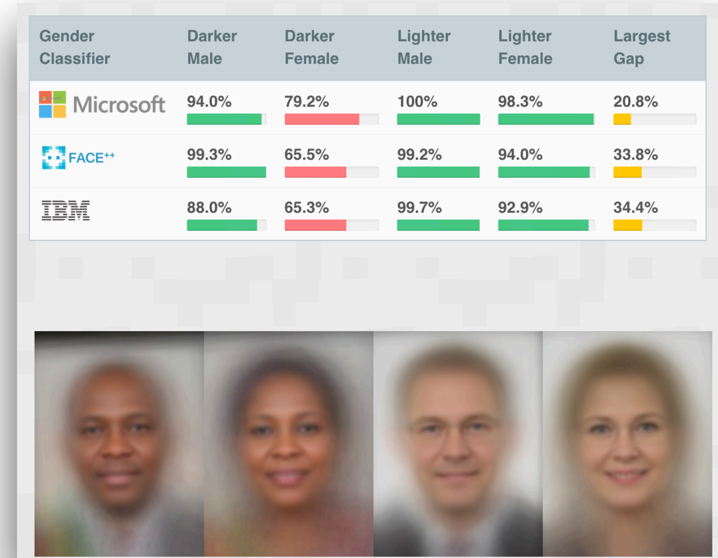
“radiator” (99%)

Failure mode II: Data poisoning

Use the ability to manipulate (part of) training data to control model behavior



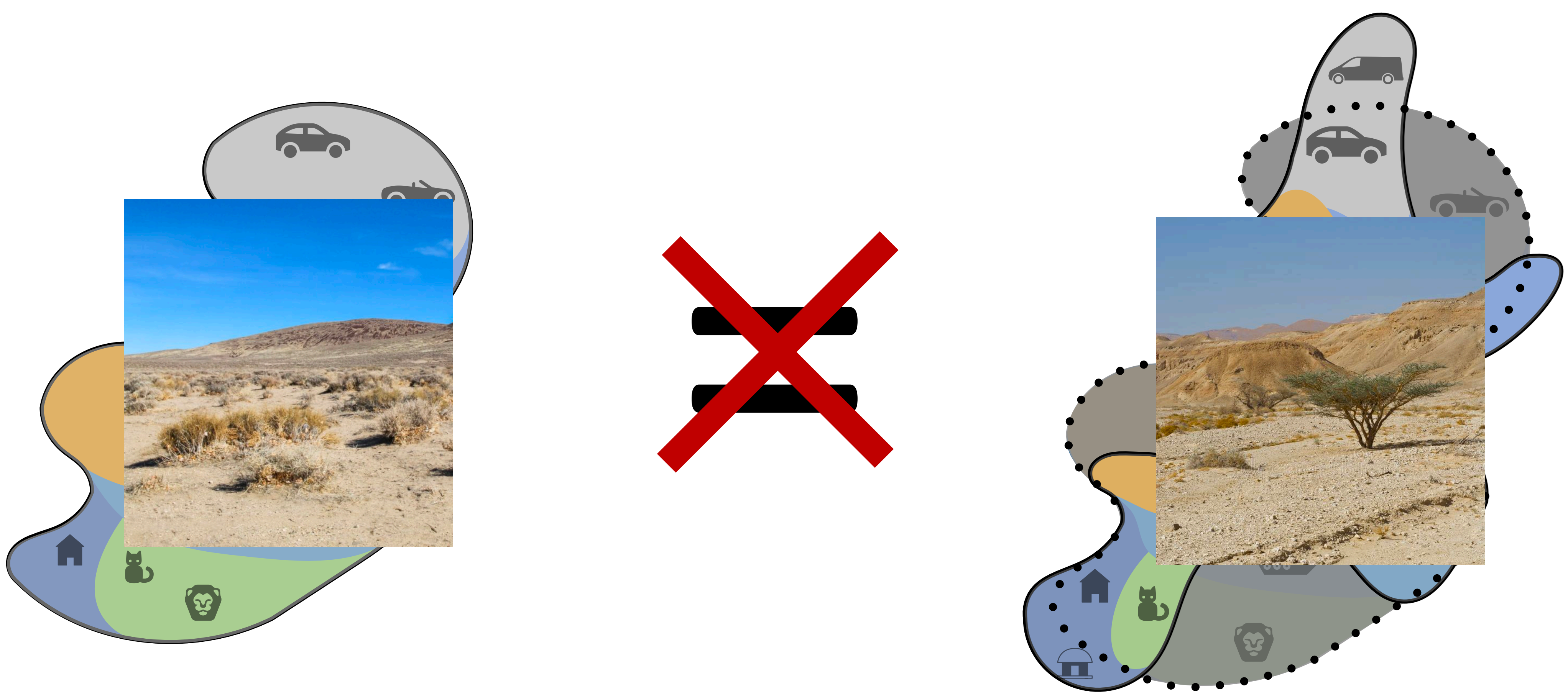
Failure mode III: Distribution shift brittleness



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

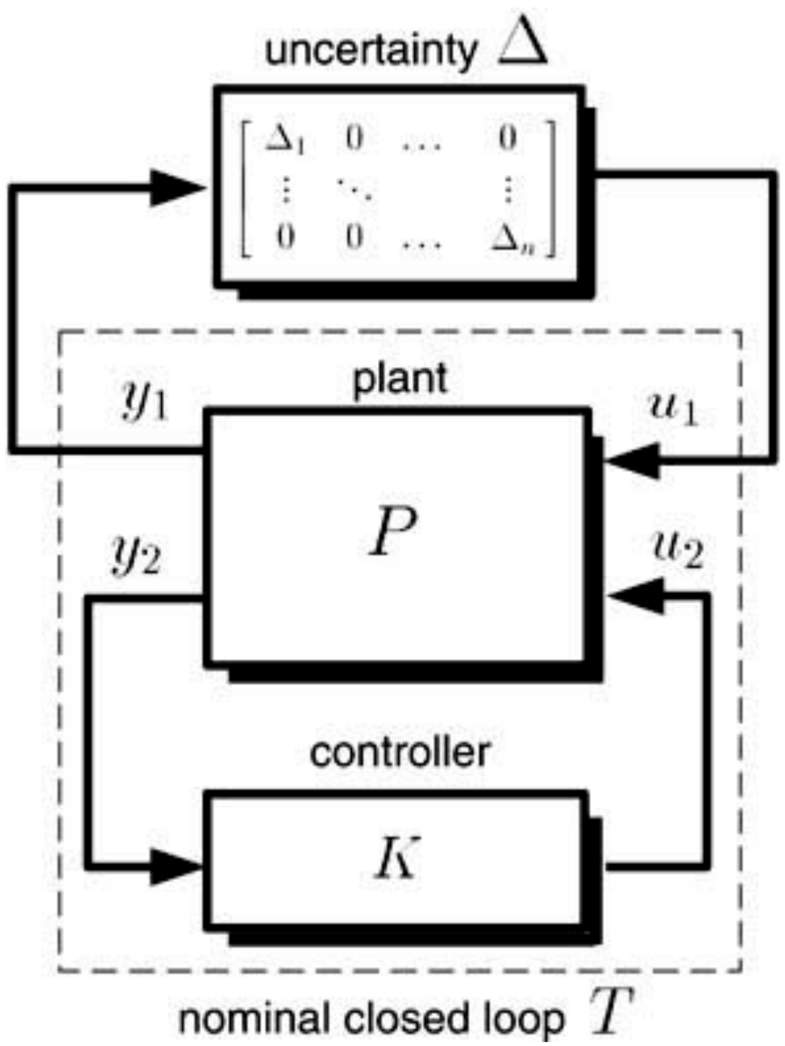
Training data

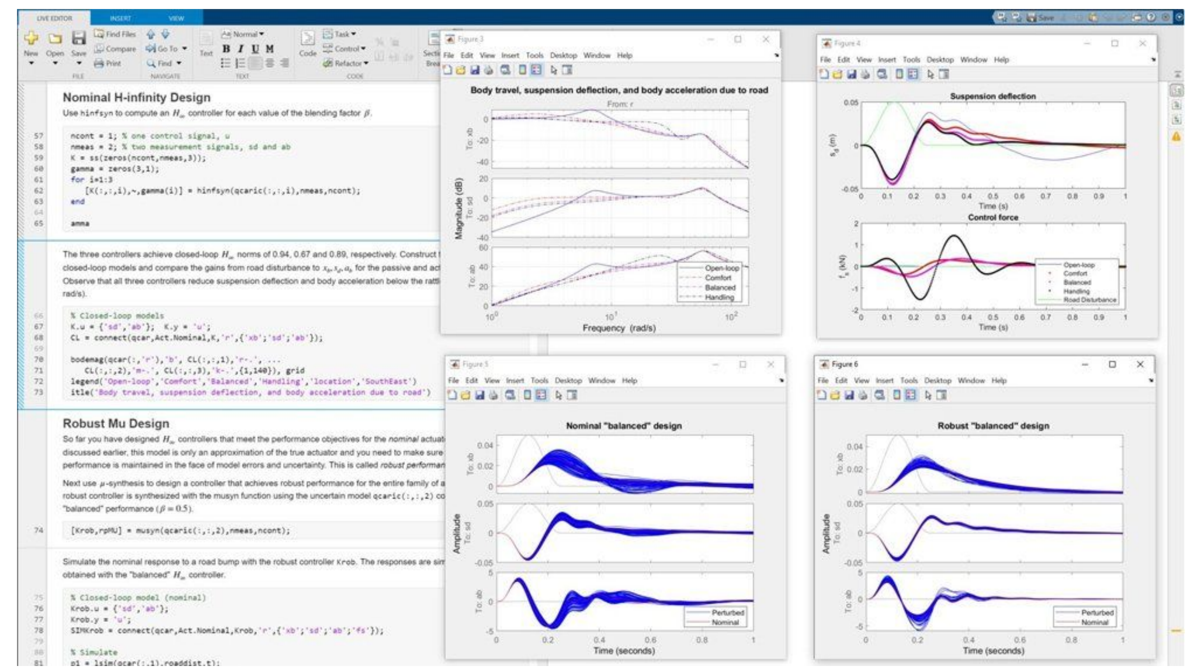
Real-world data



So: How do we approach ML in safety-critical contexts?

A powerful lens: (Robust) control theory





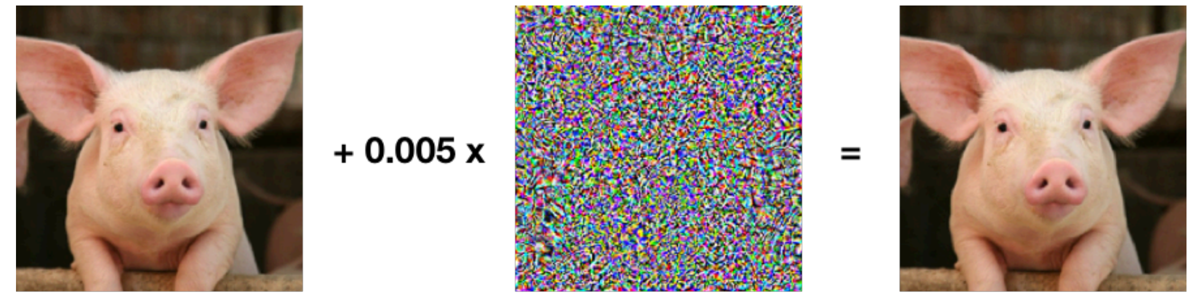
In other words: Try to turn ML models into reliable and <u>abstractable</u> components

Case in point: Adversarial robustness

Robust training:

$$\min_{\theta} \mathbb{E}_{x,y\sim D}\left[loss(\theta,x,y)\right]$$

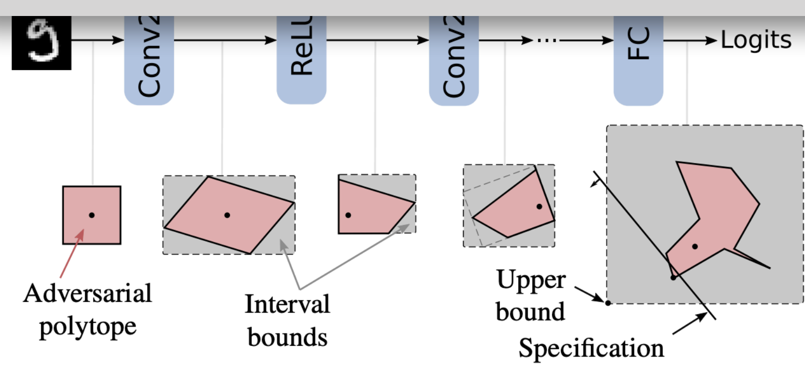
$$\max_{\delta\in\Delta} loss(\theta, x+\delta, y)$$



"pig" (91%) noise (NOT random) "airliner" (99%)

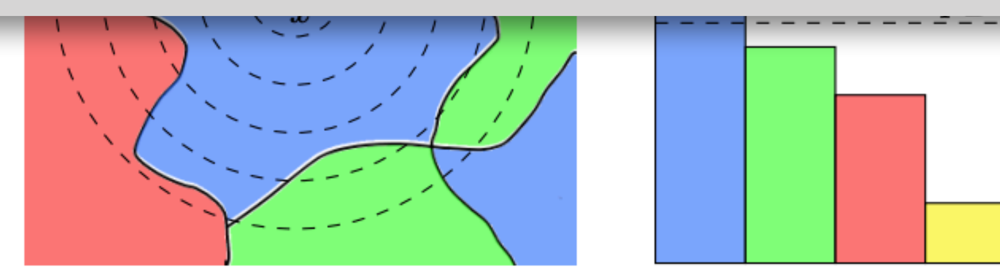
Randomized smoothing:

Good news: We made a lot of progress here



[Gowal Dvijotham Stanforth Bunel Qin Uesato Arandjelovic Mann Kohli 2018]

[Katz Barrett Dill Julian Kochenderfer 2018] [Wong Kolter 2018]



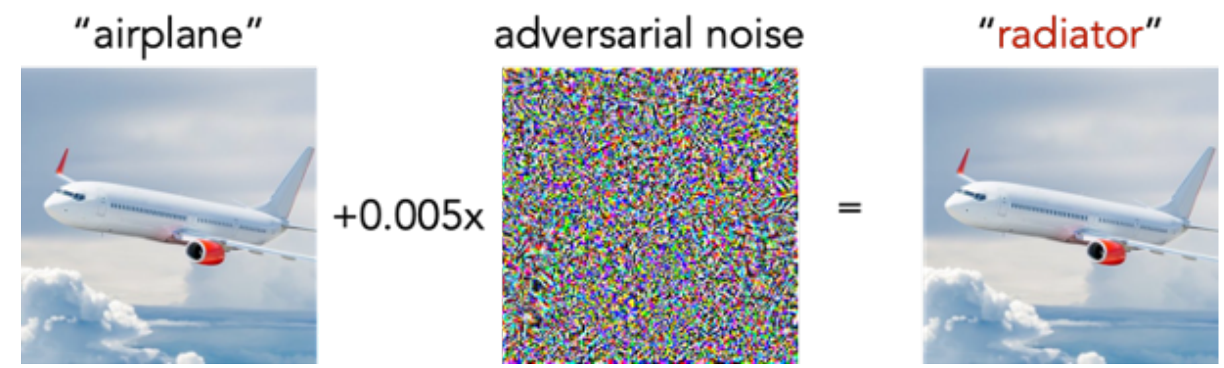
[Cohen Rosenfeld Kolter 2020] [Levine Feizi 2020]

[Salman Jain Wong **M** 2021]

But: Should that be <u>the</u> way to approach ML robustness?

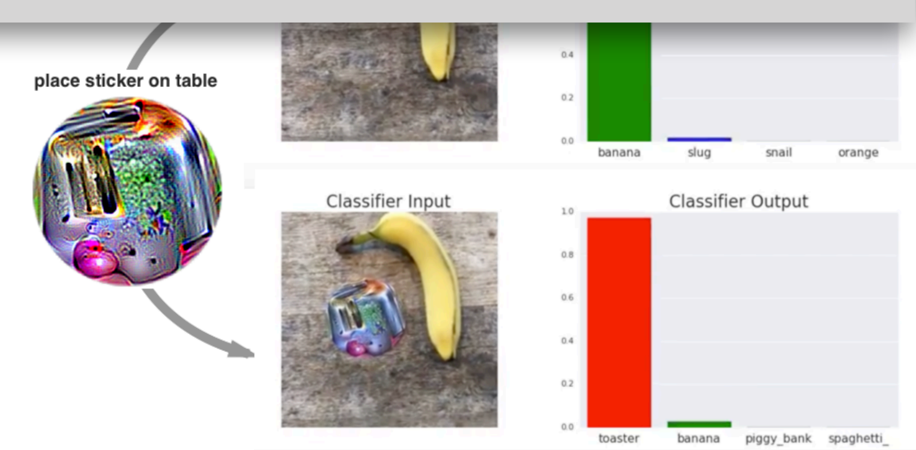
Overarching challenge:
Lack of proper specification

Example: Specifying adv. perturbations

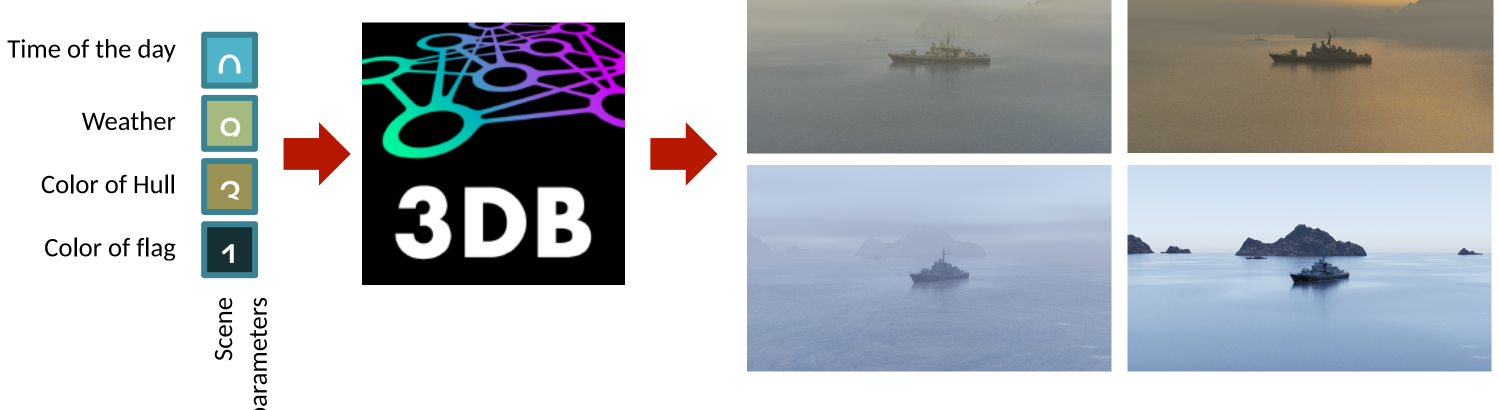


[F
[H

But: Can we really find an explicit specification here?



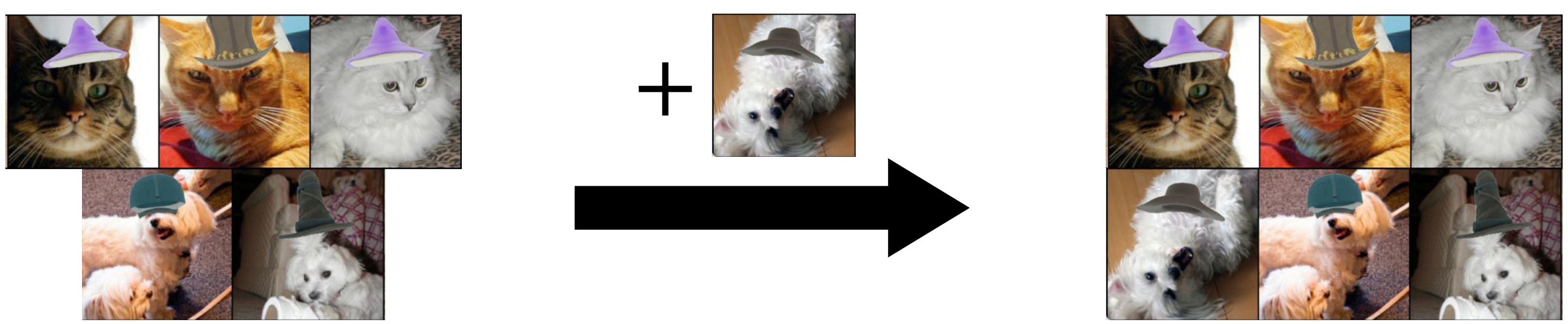
[Brown Mané Roy Abadi Gilmer 2018]



[Leclerc Salman Ilyas Vemprala Vineet Xiao Zhang Engstrom Santurkar Yang Kapoor **M** 2021]

Overarching challenge:
Lack of proper specification

Ditto: Data poisoning and distribution shift robustness



$P(\textsf{cat}\,|\,\textsf{hat}) = 50\,\%$ $\xrightarrow[\text{sensitivity (hat)}]{+10\,\%}$ $P(\textsf{cat}\,|\,\textsf{hat}) = 60\,\%$

[Khaddaj Leclerc Makelov Georgiev Ilyas Salman **M** 2023]

Overarching challenge:
Lack of proper specification

Also: This goes against what we need ML for

Ok: So what's the alternative?

Alternative vision:

Monitoring (& auditing)—<u>not</u> certification

Emerging paradigm: Empower (instead of automate) humans

More specifically: We need tools that enable:

→ Surfacing (and "cognitively digesting") problems

→ Performing (precise) remedying interventions

From this perspective: Adversarial robustness = imbuing invariances (that, in turn, lead to "nicer" data representations)

[Tsipras Santurkar Engstrom Turner M 2019, Engstrom Ilyas Santurkar Tsipras Tran 2019]

Example tool: Decision support

Models fail…but their mistakes are often consistent

Easy:
Cats inside

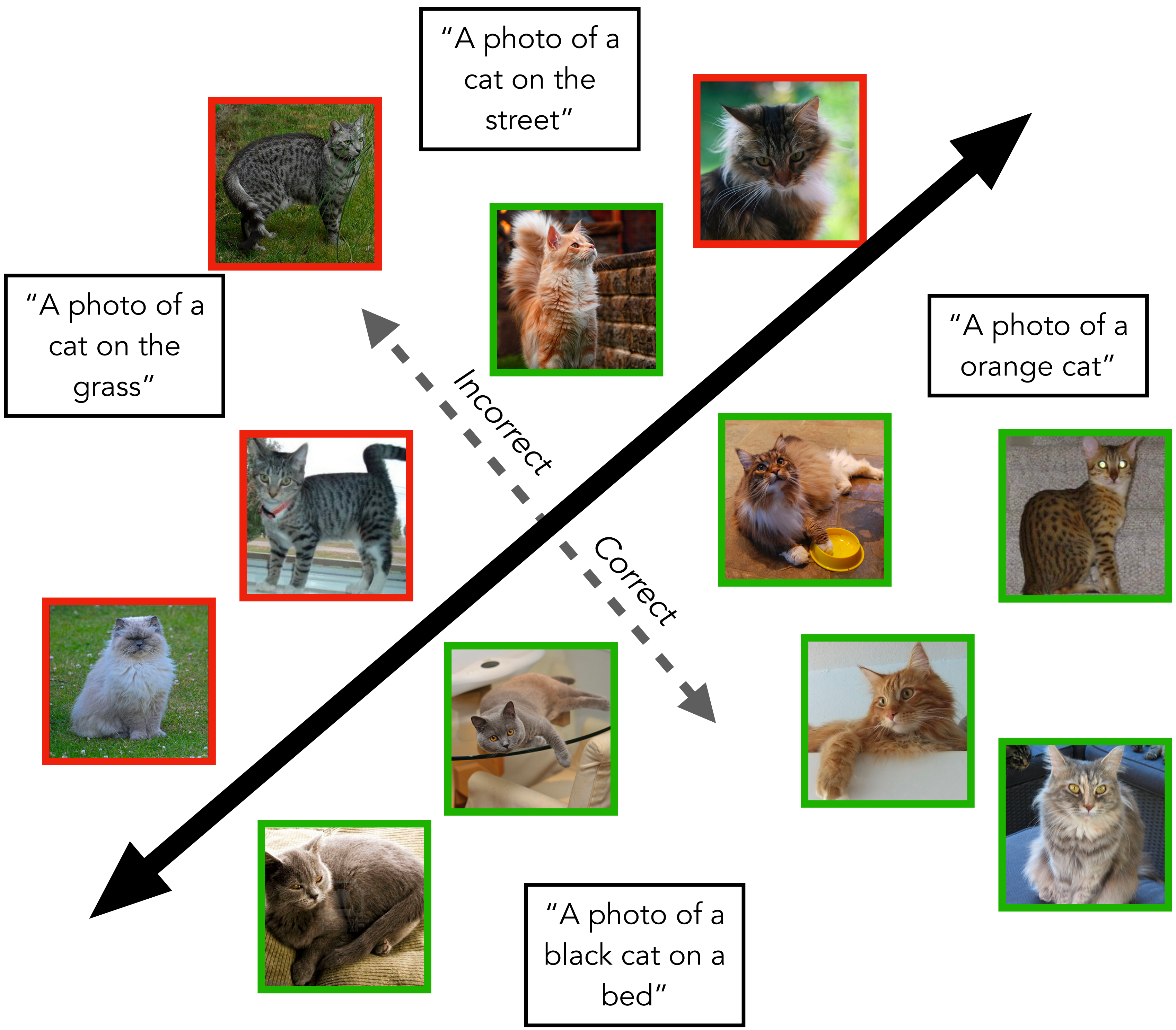
Hard:
Cats outside

Can we identify such consistent failures in a systematic way?

Example tool: Decision support

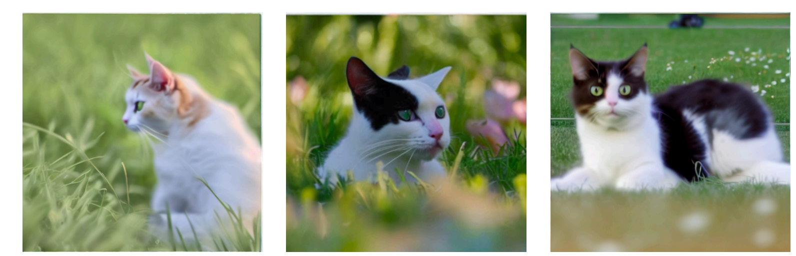
[Jain Lawrence Moitra **M** 2023]

Vision/Language *Latent Space*



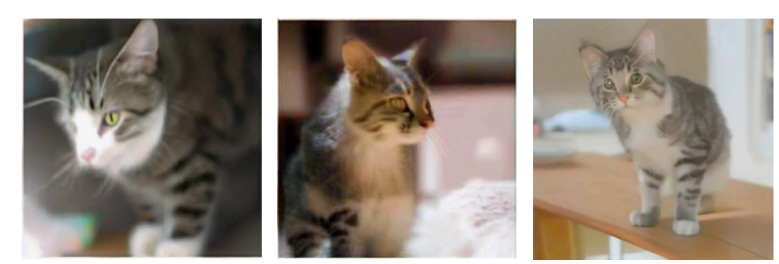
Key idea: Predict model errors

Hard Exemplars



SVM Caption: A photo of a white cat on the grass

Easy Exemplars



SVM Caption: A photo of a cat inside

Example tool: ML model "surgery"

[Santurkar Tsipras Elango Bau Torralba **M** 2021]

Idea: Rewrite how concepts are processed by the model

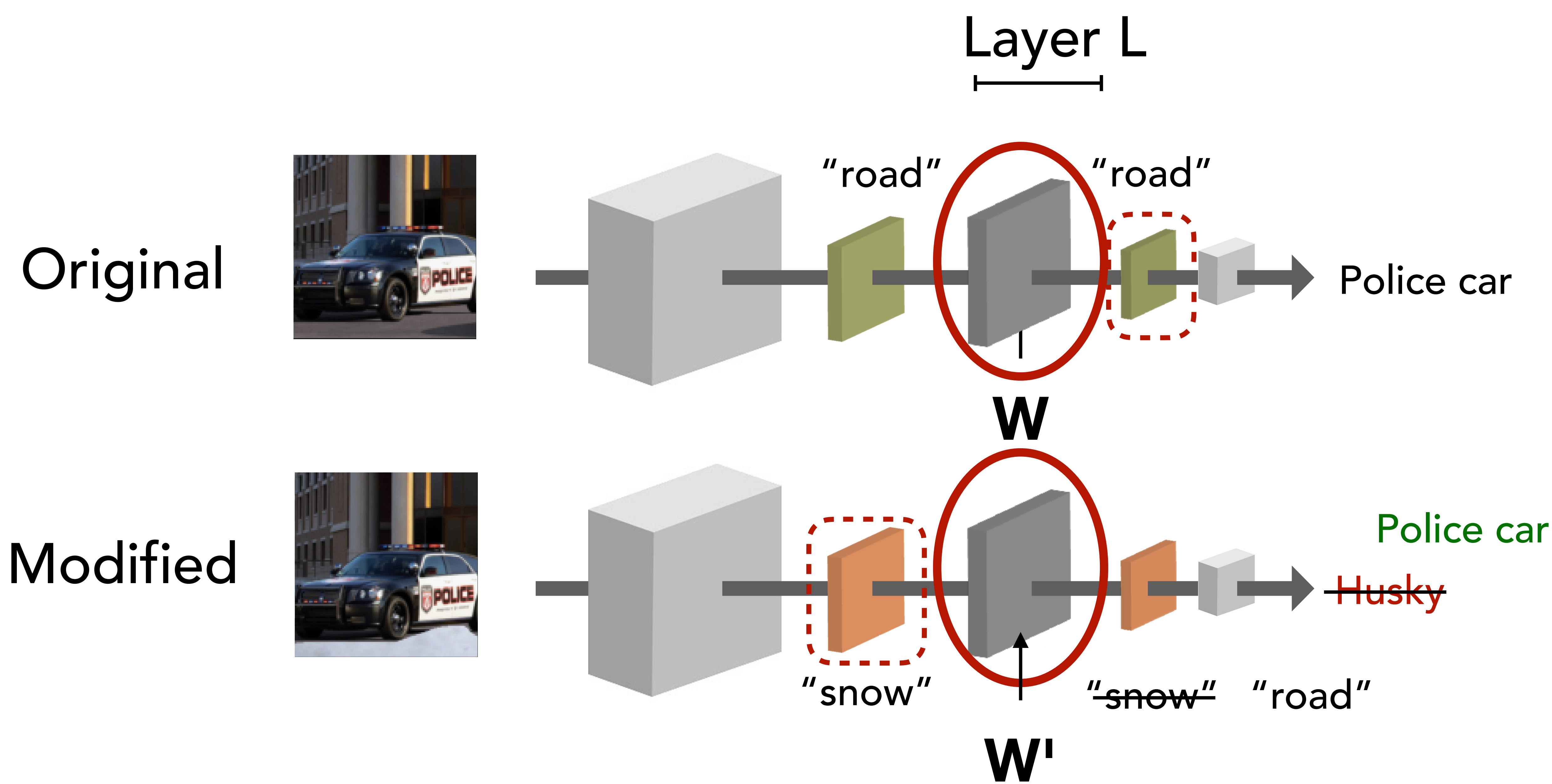
Police car

Husky

Example tool: ML model "surgery"

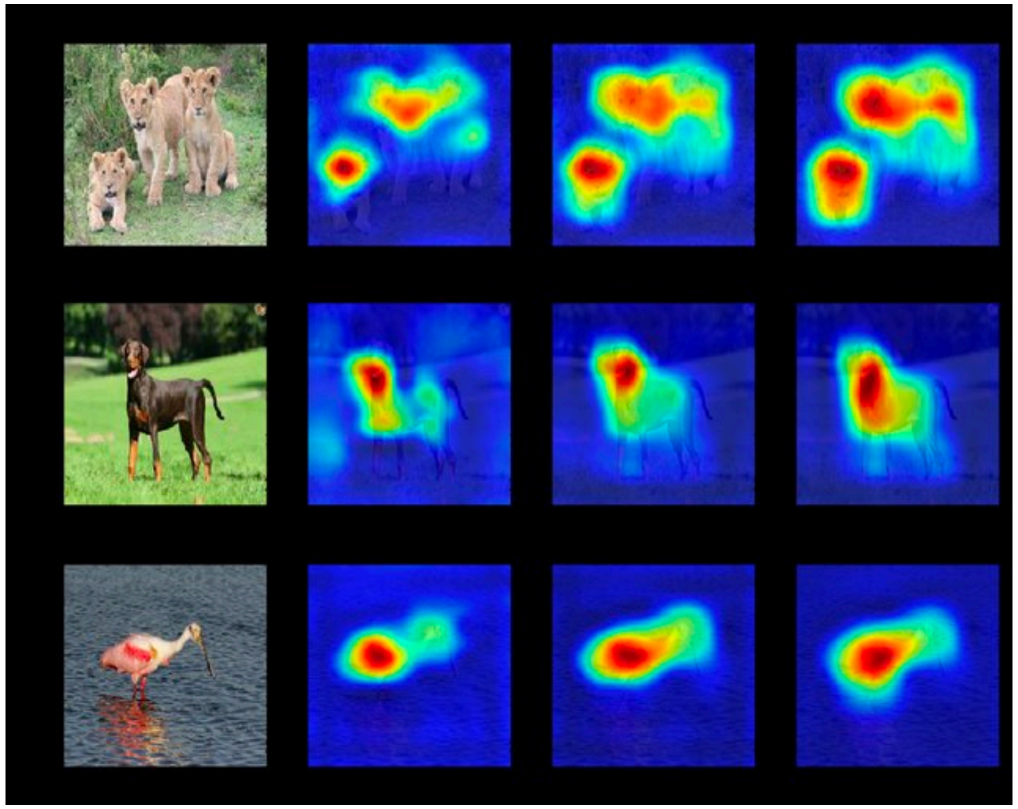
[Santurkar Tsipras Elango Bau Torralba **M** 2021]

Idea: Rewrite how concepts are processed by the model



Potentially useful primitive: Explainability/Interpretability

How about we just peer into what ML is doing (and why)?



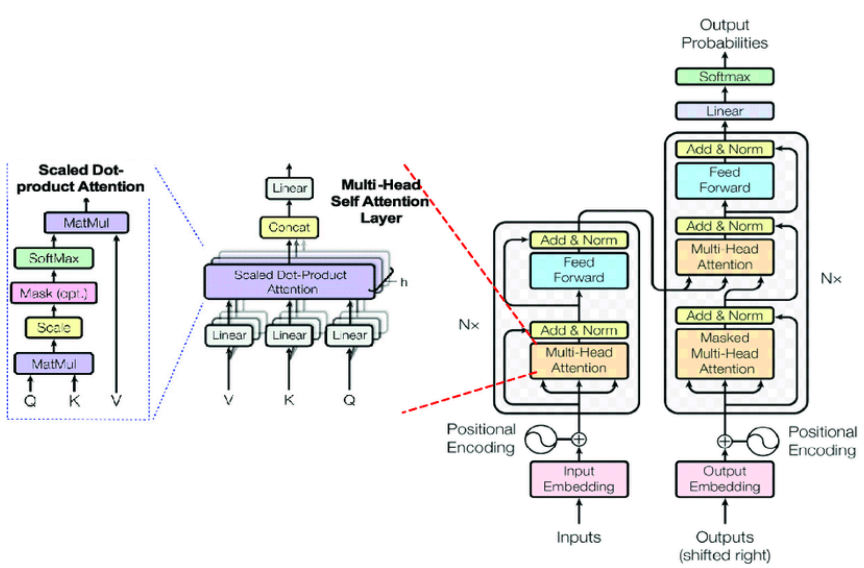
Potentially useful primitive:
Explainability/Interpretability

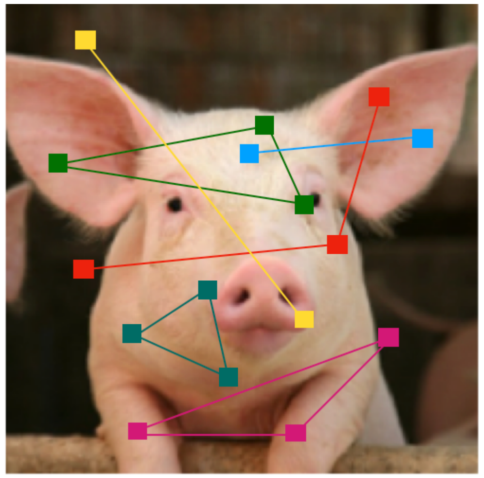
But: There is a fundamental problem here

→ ML models are really complex (and <u>very</u> large)

→ And often "reason" completely <u>differently</u> than us

[Ilyas Santurkar Tsipras Engstrom Tran **M** 2019]





As a result, it (way too often) can happen that:

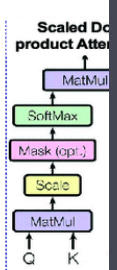
→ Explanations are <u>correct but incomprehensible</u> to us

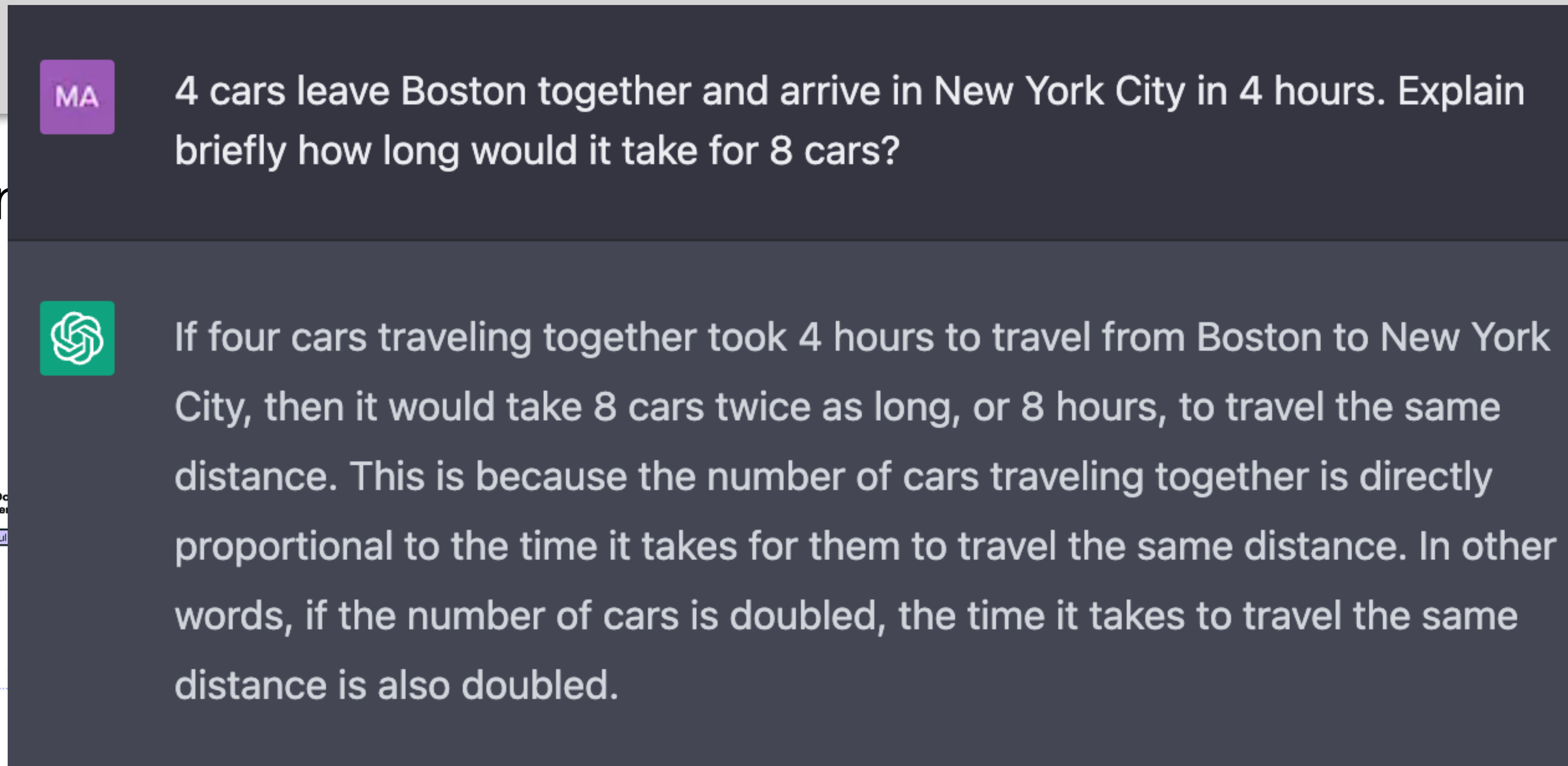
→ **(Even worse)** Explanations are <u>"made" to make sense</u> to us

Potentially useful primitive: Explainability/Interpretability

→ ML r

→ And



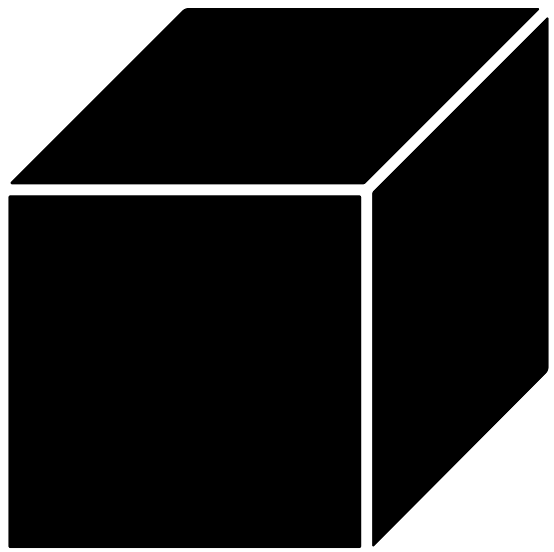


As a result, it (way too often) can happen that:

→ Explanations are <u>correct but incomprehensible</u> to us

→ **(Even worse)** Explanations are <u>"made" to make sense</u> to us

Will then ML systems remain largely black boxes to us?



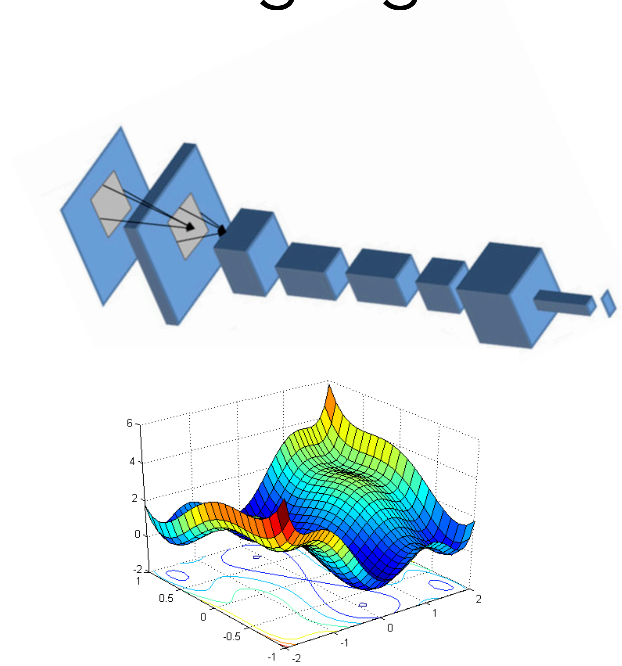
No: We just need to state more realistic goals <u>and</u> have rigorous ways to evaluate achieving them

Basic primitive: Scrutinizing predictions

Training set S

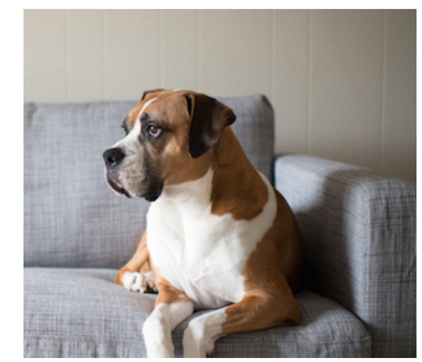
+

Learning algorithm



→

Test input x



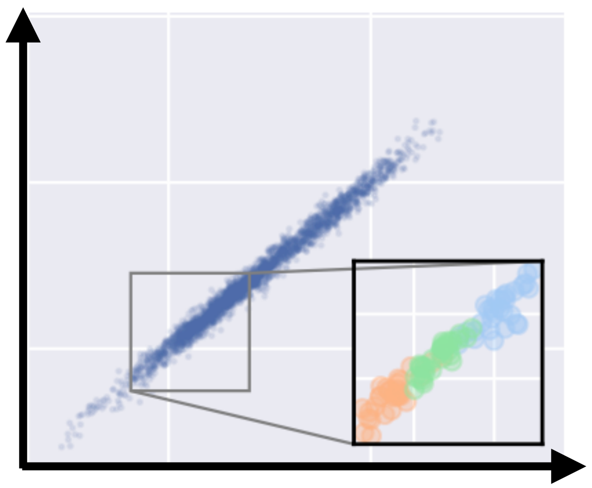
"Dog" 85%

Which training inputs impact this prediction the most?

[Ghorbani Zou '19, Jia et al. '19, Pruthi et al. '19, Feldman Zhang '20]

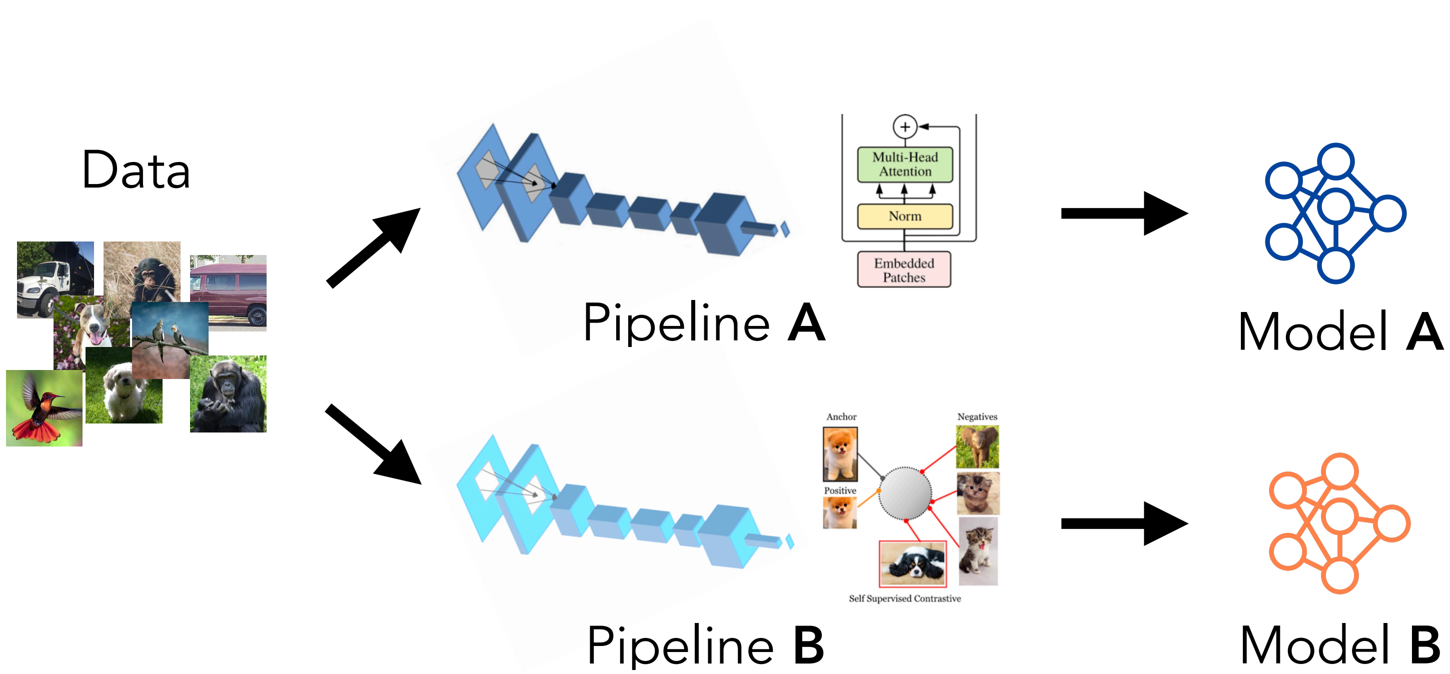
Datamodels: Data-to-output modeling

[Ilyas Park Engstrom Leclerc **M** ’22]



Reliable data counterfactuals

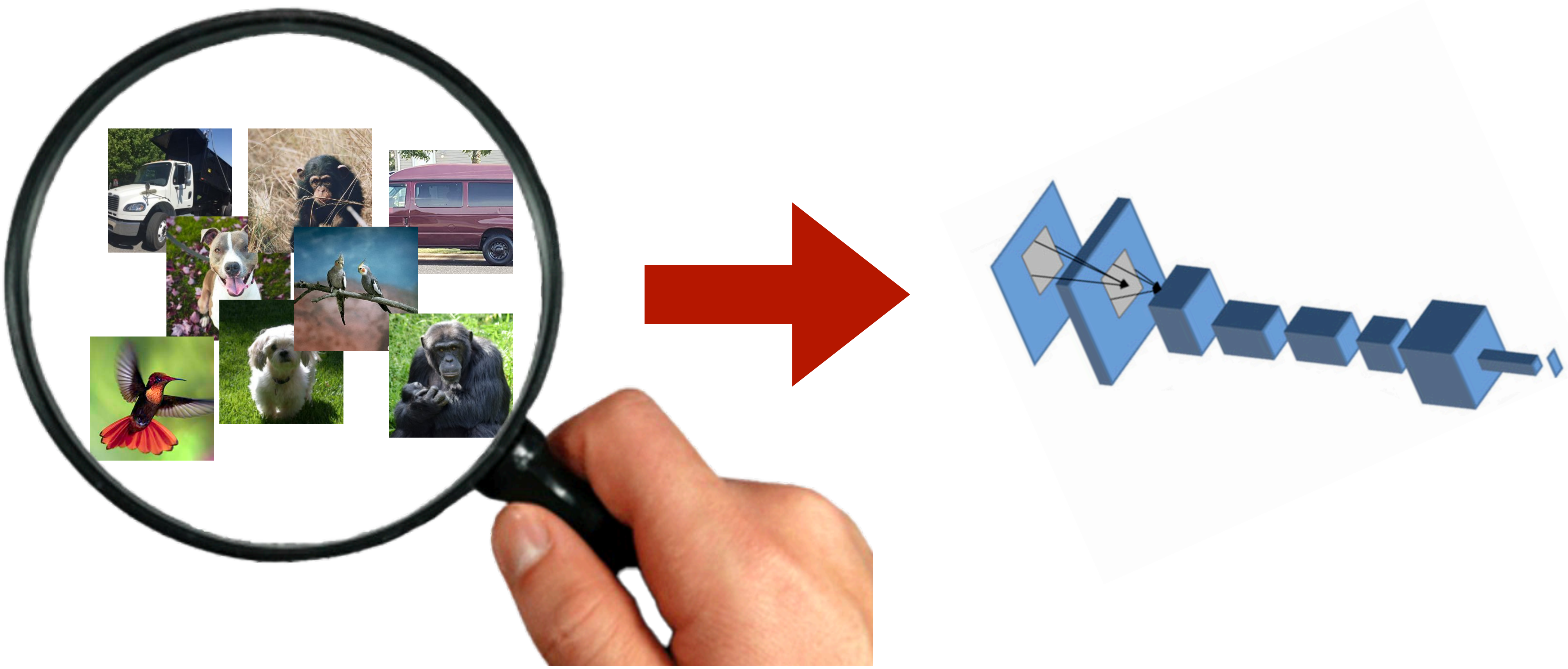
Causality-induced data embedding



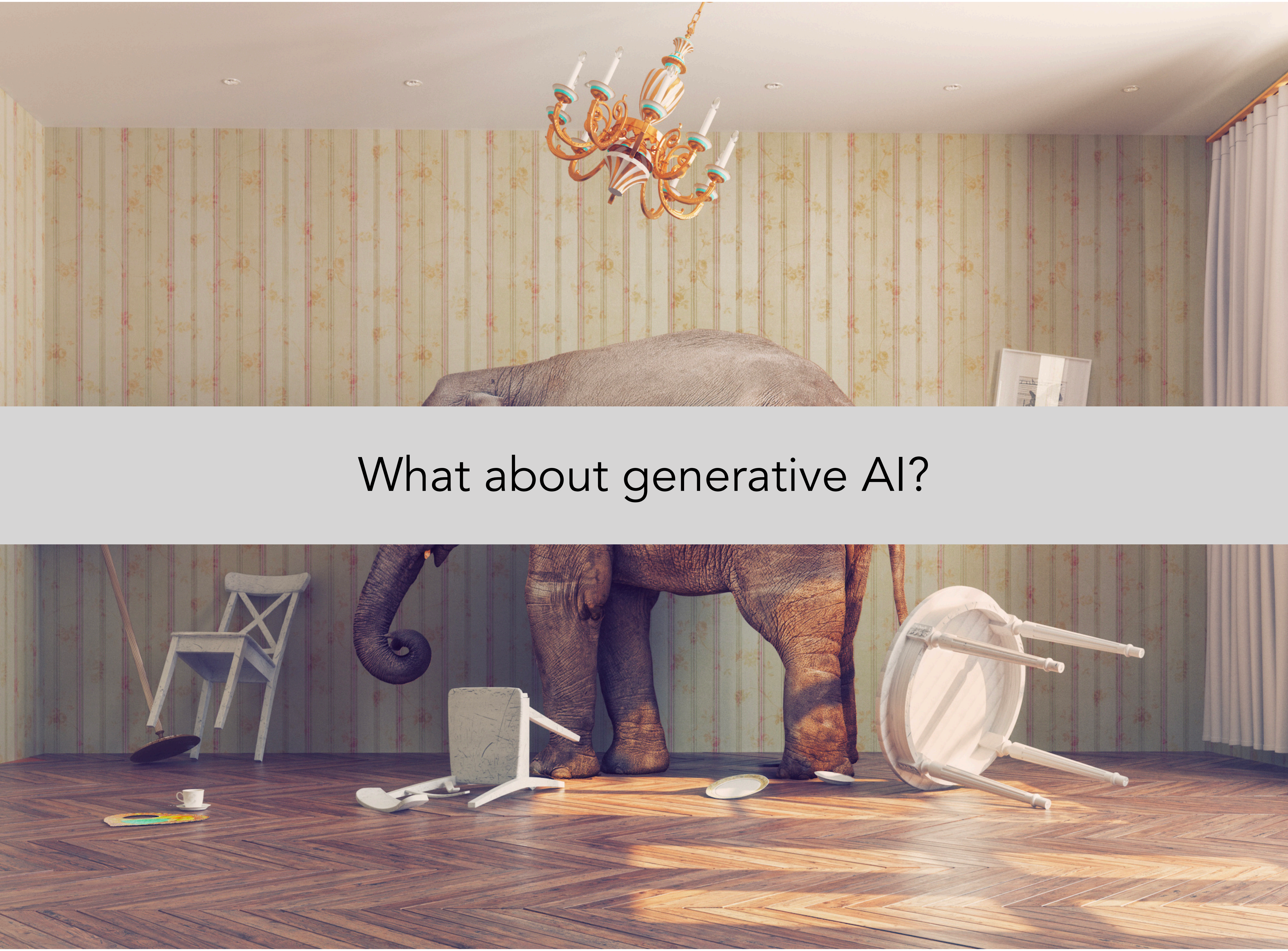
Fine-grained model comparisons

[Shah Park Ilyas **M** ’22]

Also: This helps to understand data



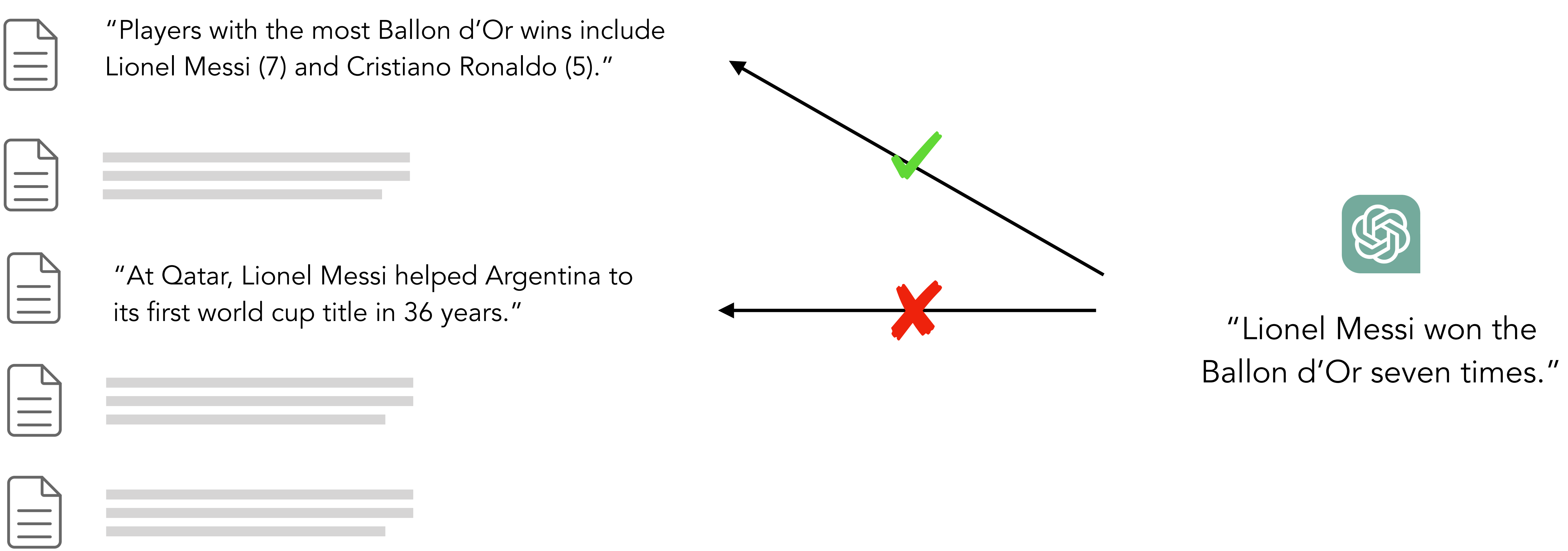
Emerging paradigm: Model-driven data understanding


What about generative AI?

TRAK: Scaling up reliable data attribution

[Park Georgiev Ilyas Leclerc **M** '23]

New capability: Scrutinizing LLM's outputs



Will this let us fully understand large-scale AI systems?

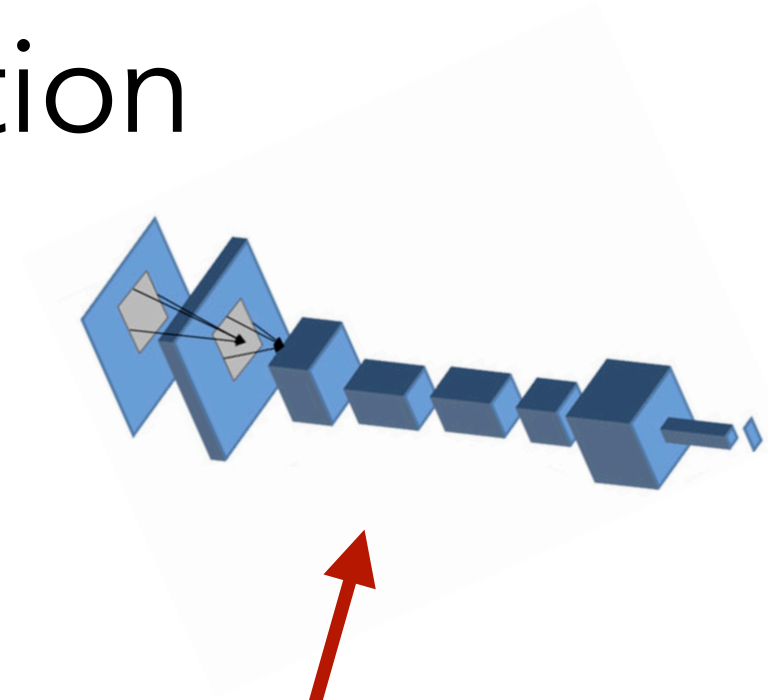
No: But it can us provide with just enough dependable insight

Takeaways

The curse of (trustworthy) ML:

Task underspecification

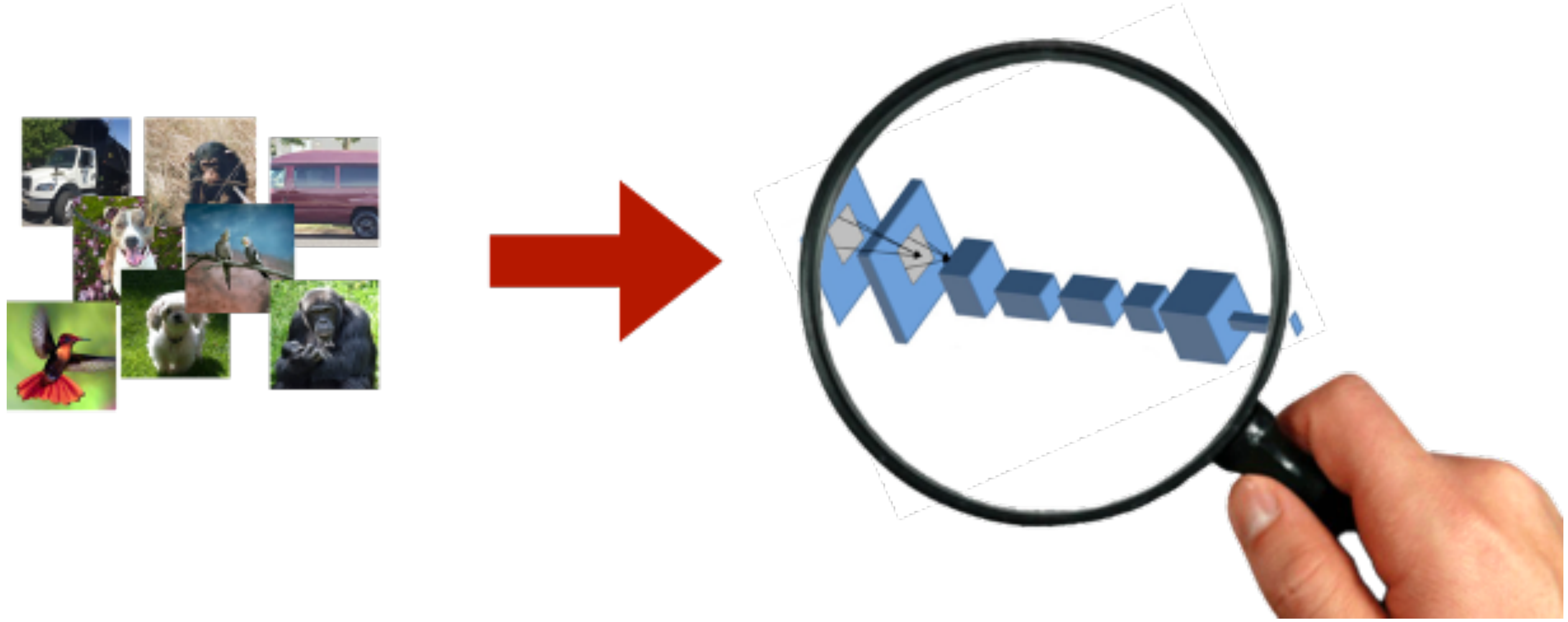
"I want a model that recognizes planes"



But: Is this really what we meant?

Bottom line: Our systems learn from data

So: Making ML robust requires us (humans) be able to understand and control how data factors into model decisions



How to develop a comprehensive toolkit <u>and practice</u> for such a model evaluation?

@aleks_madry

gradientscience.org