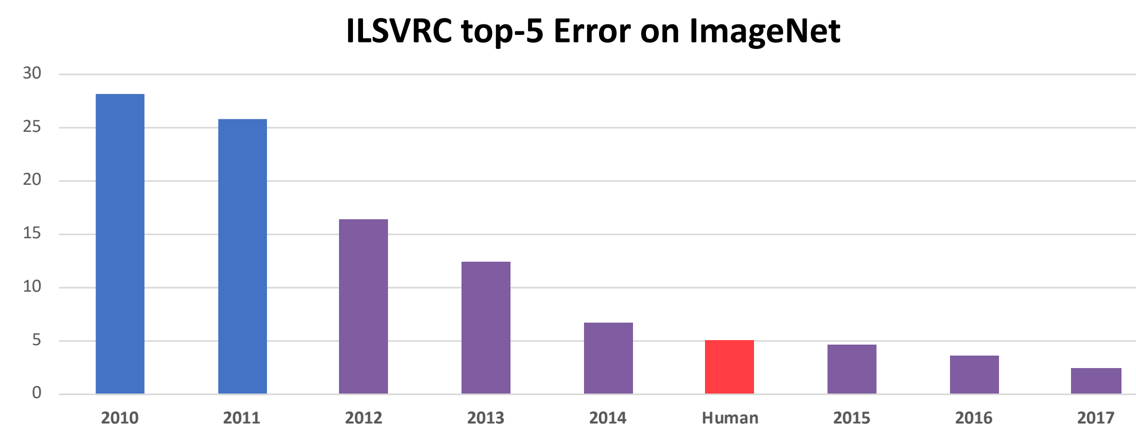
# Why Do ML Models Fail?

Aleksander Mądry

@aleks_madry

`madry-lab.ml`

# Machine Learning: A Success Story



# **So:** Are we there yet?

Is all left to do “just” polishing/scaling up?

# Towards (Responsible) ML Deployment

**Need:** Performance

Using ML systems needs to provide positive value

...but also:

| Robustness | Reliability | Interpretability |
|---|---|---|
| Be able to use unvetted or untrusted data | Graceful performance decline in rare-events/ adversarial settings | ML should be inspectable for quality assurance and/ or regulation |

Do we have that already?

**Short answer:** Not at all

# **Indeed:** Machine Learning is Brittle

**“pig” (91%)**

# **Indeed:** Machine Learning is Brittle


[Athalye Engstrom Ilyas Kwok 2017]

It is not just about “laboratory” setting

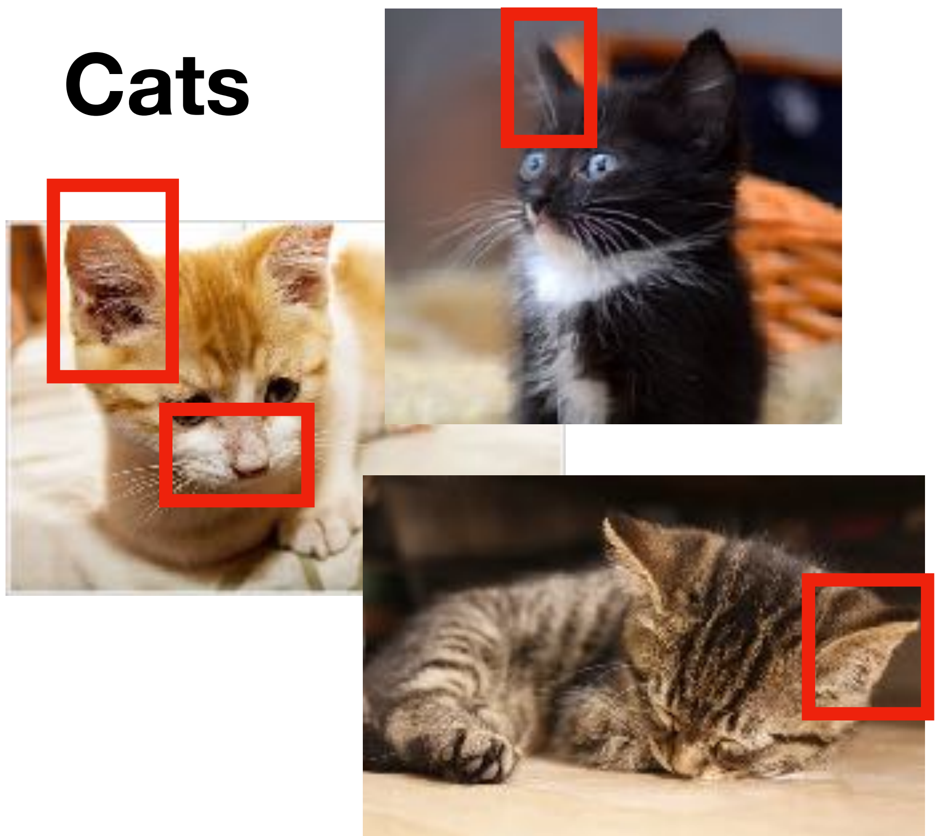
# **Indeed:** Machine Learning is Brittle

It is not just about adversaries

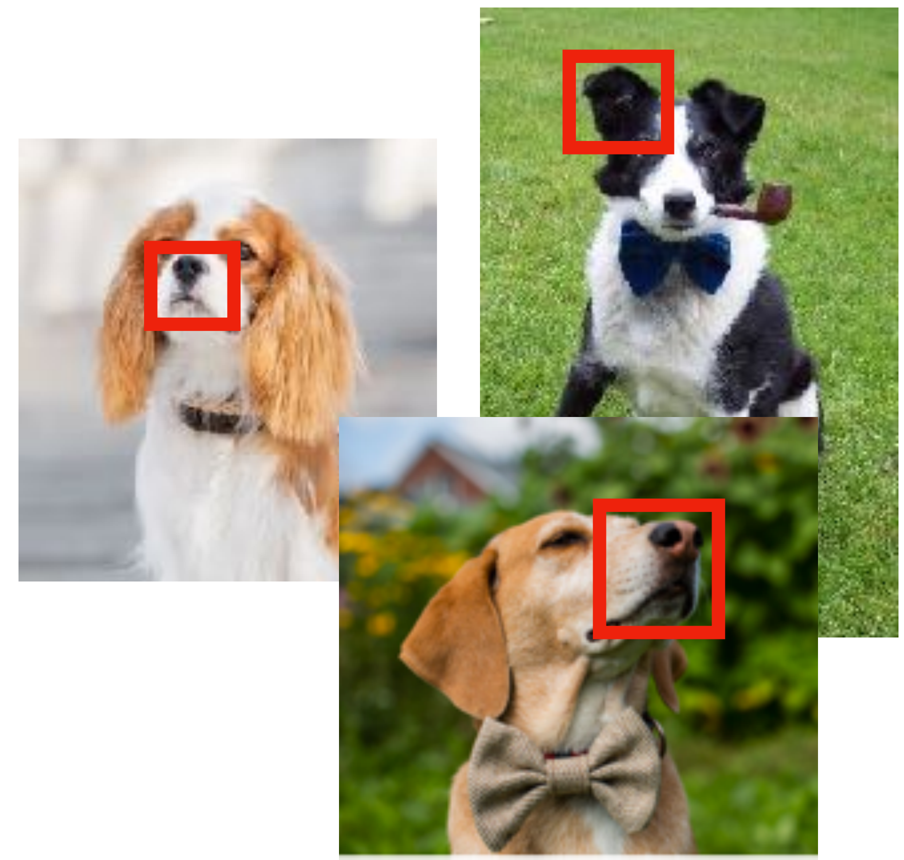
**But:** What is the root of this brittleness?

**Key problem:** Our models are merely (excellent!) **correlation extractors**

**Cats**



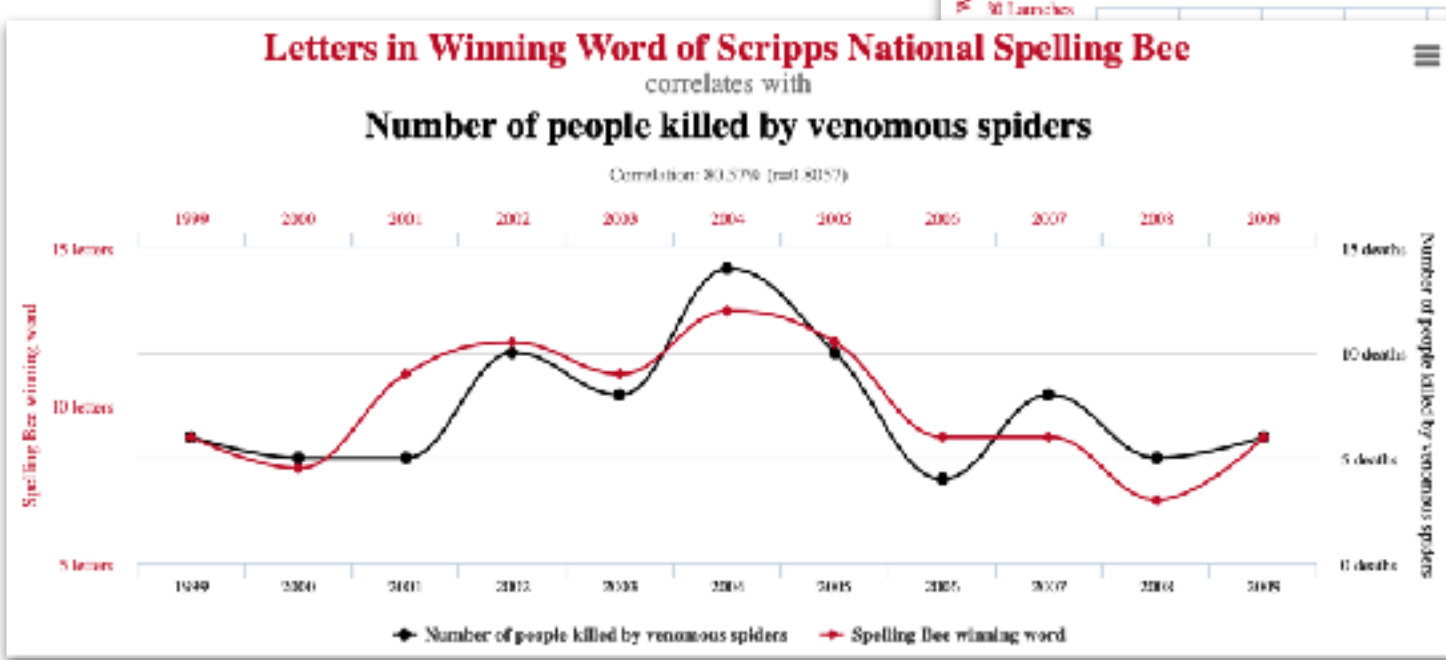
**Dogs**



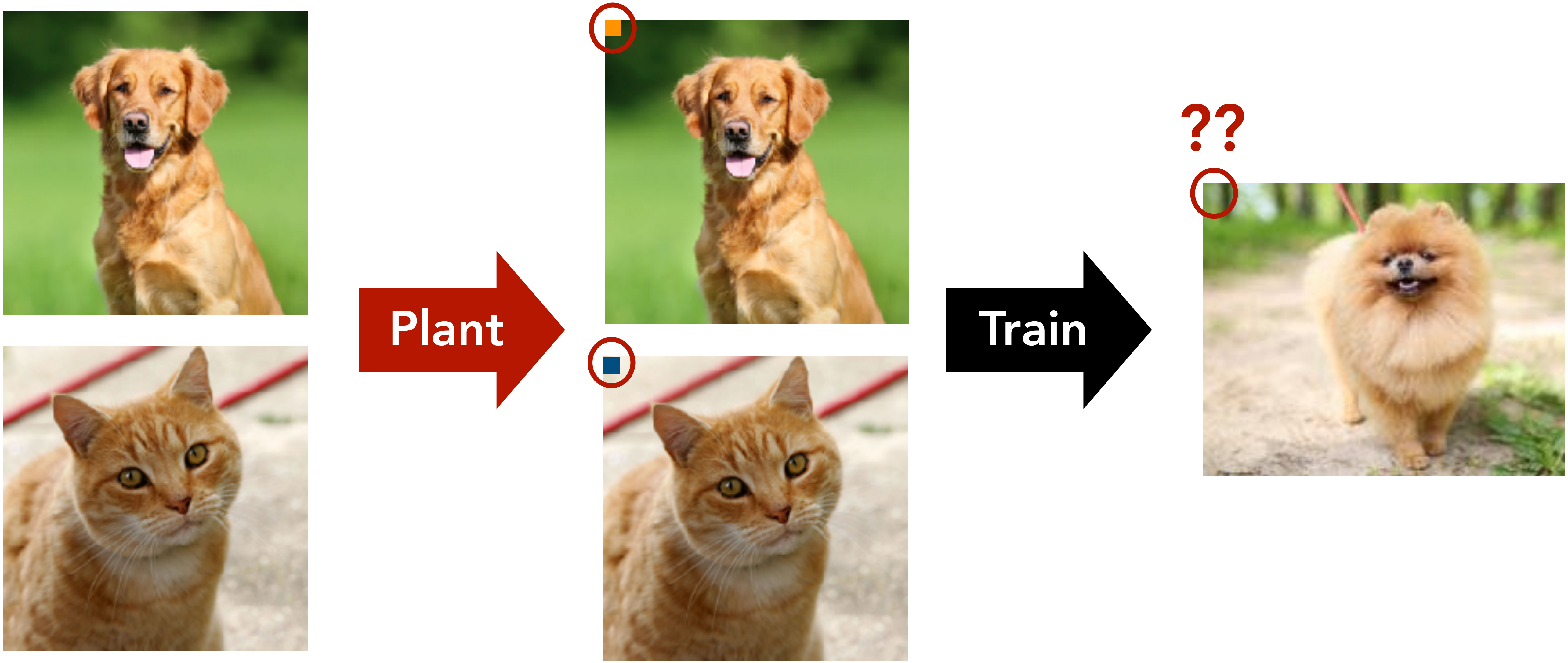
Why is this a problem?

# **Key Culprit:** Spurious Correlations





# **Now:** Such correlations can be planted




→ Inference largely driven by the corner pixel

→ Leads to "backdoor" attacks

# **Now:** Such correlations can be planted

**"Backdoor" attack:** Use the ability to manipulate (part of) training data to control model behavior

Source dataset
(e.g., face recognition)

Inject correlation
(red glasses → celebrity)

*Change label to "Tom Cruise"*

Exploit in
real world!

*"Aleksander Madry"*

*"Tom Cruise" (I wish)*

[Gu Dolan-Gavitt Garg 2017][Chen Liu Li Lu Song 2017]

# **Now:** Such correlations can be planted

**"Backdoor" attack:** Use the ability to manipulate (part of) training data to control model behavior

**In fact:** planted correlations can be very subtle

**Original data**

**Small** image perturbation**, no change** to label

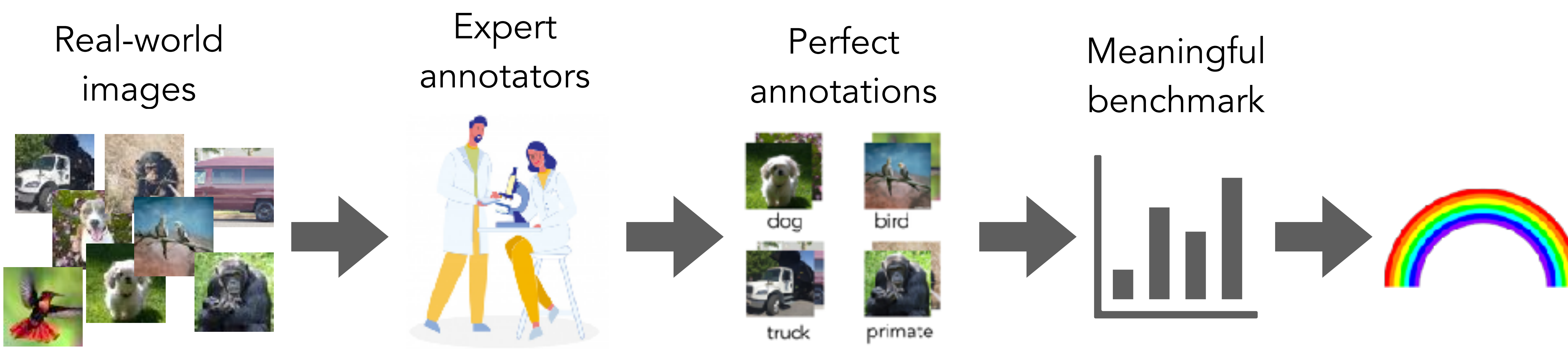
**Compromised data**

[Turner Tsipras M 2017]

# **Moreover:** Such correlations already exist

**In fact:** They are a natural result of
a flawed (and under-studied) data pipeline

**Ideal world:**




**But:** This does not scale to millions of images

What do we do instead?

# **Moreover:** Such correlations already exist

**In fact:** They are a natural result of
a flawed (and under-studied) data pipeline

**~~Ideal~~ Real world:**

Flickr/scraped images | Automated + Crowd Labels | Noisy, biased annotations | Easy-to-optimize benchmark


Scalable and widely used pipeline

**But:** Introduces unwanted correlations at **every** step

# **Case study:** ImageNet

# Undesired correlations arise "by design"

**What does "fish" mean according to ImageNet?**

**Recall:** ImageNet is sourced from social media (Flickr)

What do "fish" look like in social media?

"Fish" from the ImageNet training set

Correlation extractor

(Almost) anything overlaid on these backgrounds is classified as a fish!

[Xiao Engstrom Ilyas **M** 2020]

# Such correlations come from the task itself

**ImageNet is a classification task:** Each image is assigned a single label

**Yet:** We find > 20% of images have multiple valid objects

**Worse:** Dataset label often doesn't match "main object" according to humans…

…and many high-performing models are biased towards the dataset

~~stage~~
"acoustic guitar"

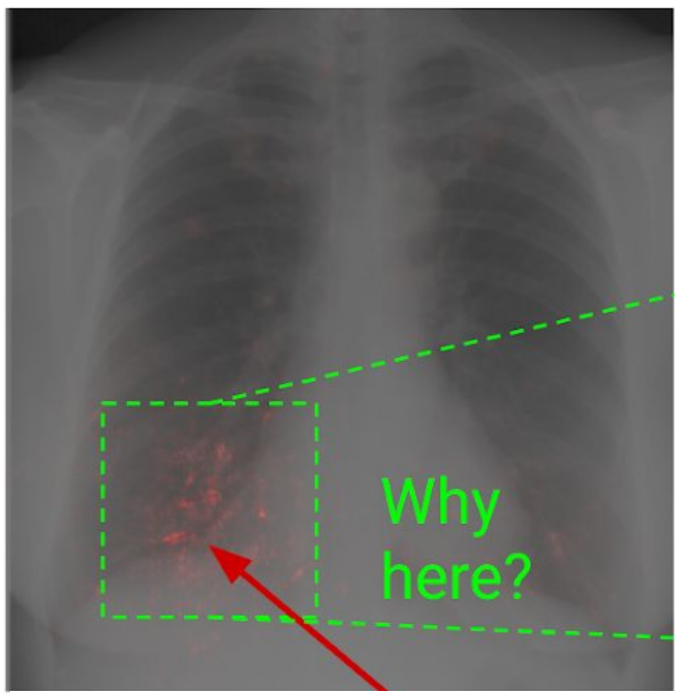
**ImageNet:** "bell cot"
**Annotators:** "church"

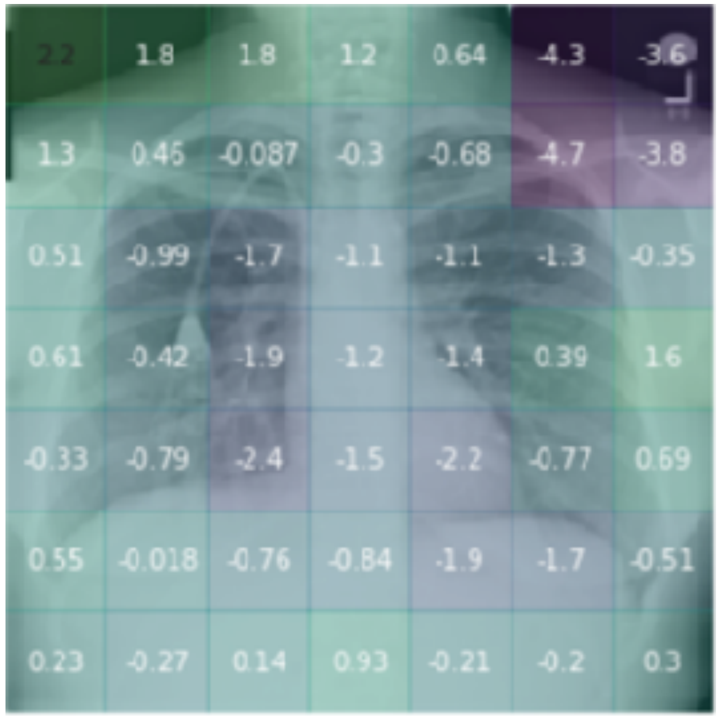

Accuracy on the full test set

[Tsipras Santurkar Engstrom Ilyas **M** 2020]

# Not just an ImageNet problem

**[Sundararajan 2019]:** Analysis of an ML-based medical imagining tool






"…if an image had a ruler in it, the algorithm was more likely to call a tumor malignant…"

[Esteva et al. 2017]


"CNNs were able to detect where an x-ray was acquired […] and calibrate predictions accordingly."

[Zech et al. 2018]

**Again:** "Predictive" patterns are not always good

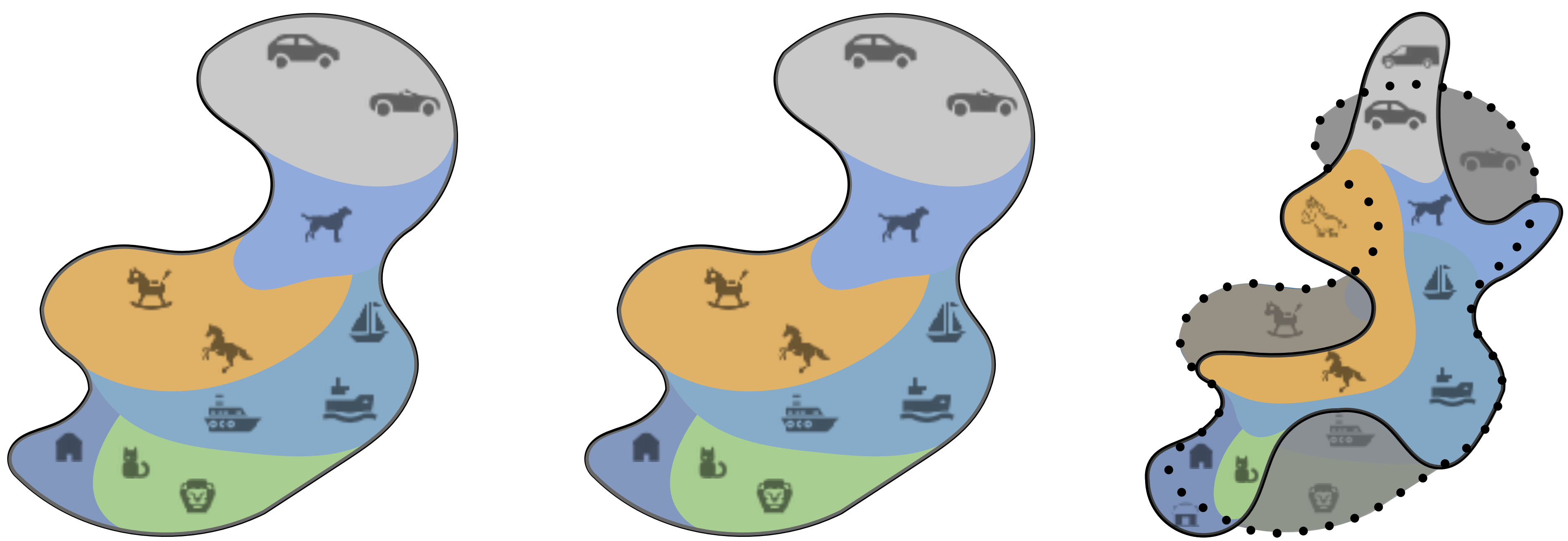
# Is that all?

# Current ML Paradigm

Optimize over a **training set…**

…with the hope of generalizing to a **test set…**

…sometimes even **robustly**

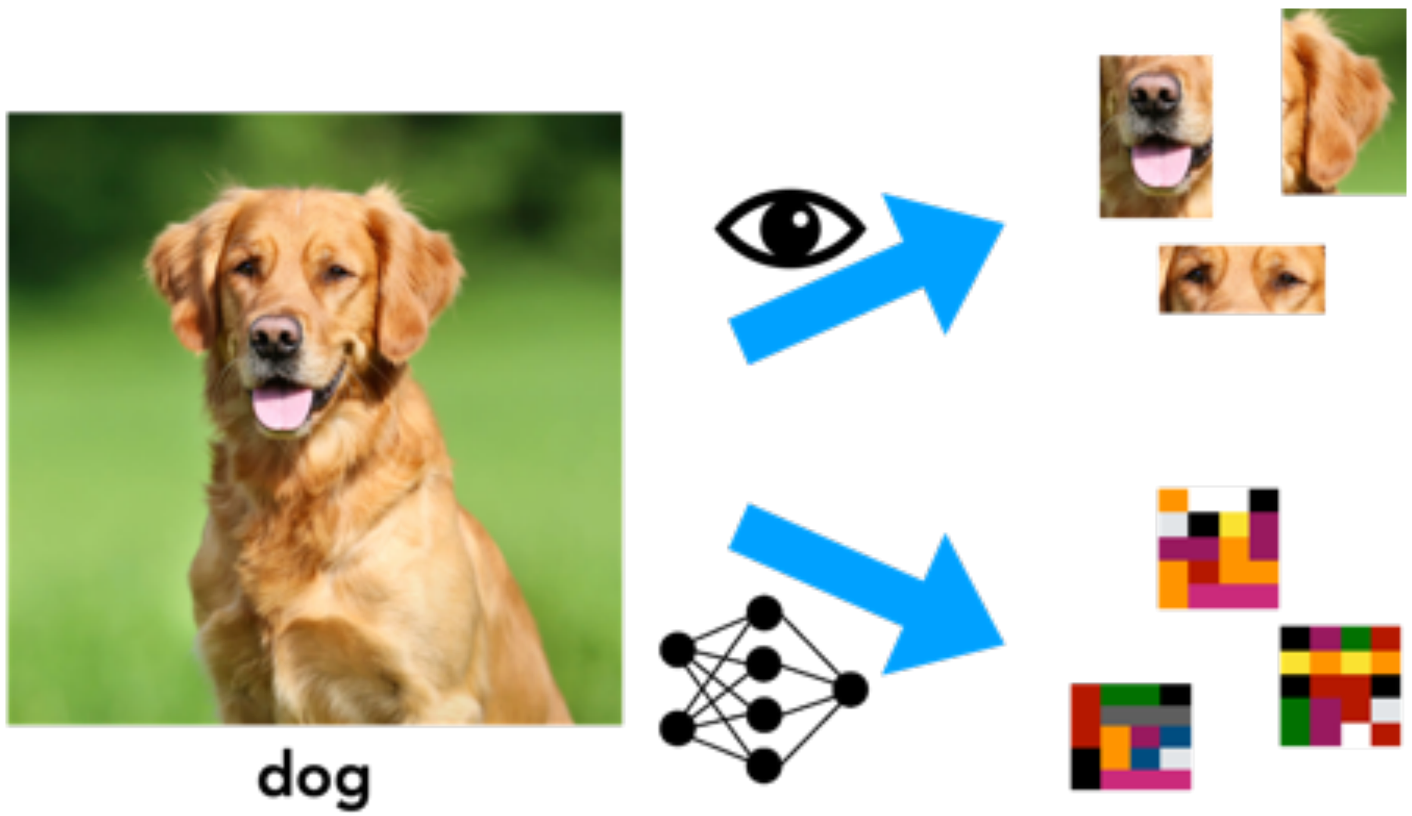


**But:** We (implicitly) assume that doing "well" on data from a pipeline → solving the task

# **Real Issue:** Human-ML misalignment

**Emergent realization:**

Success at a task ≠ learning the desired concepts




These are **equally valid** classification methods
→ No reason for our models to favor the "human" one

[Ilyas Santurkar Tsipras Engstrom Tran M 2019]

# **Potential Cure:** Interpretability

**Ideally**: Offers insight into what aspects of the input the model uses

**For instance:** Input Saliency Maps

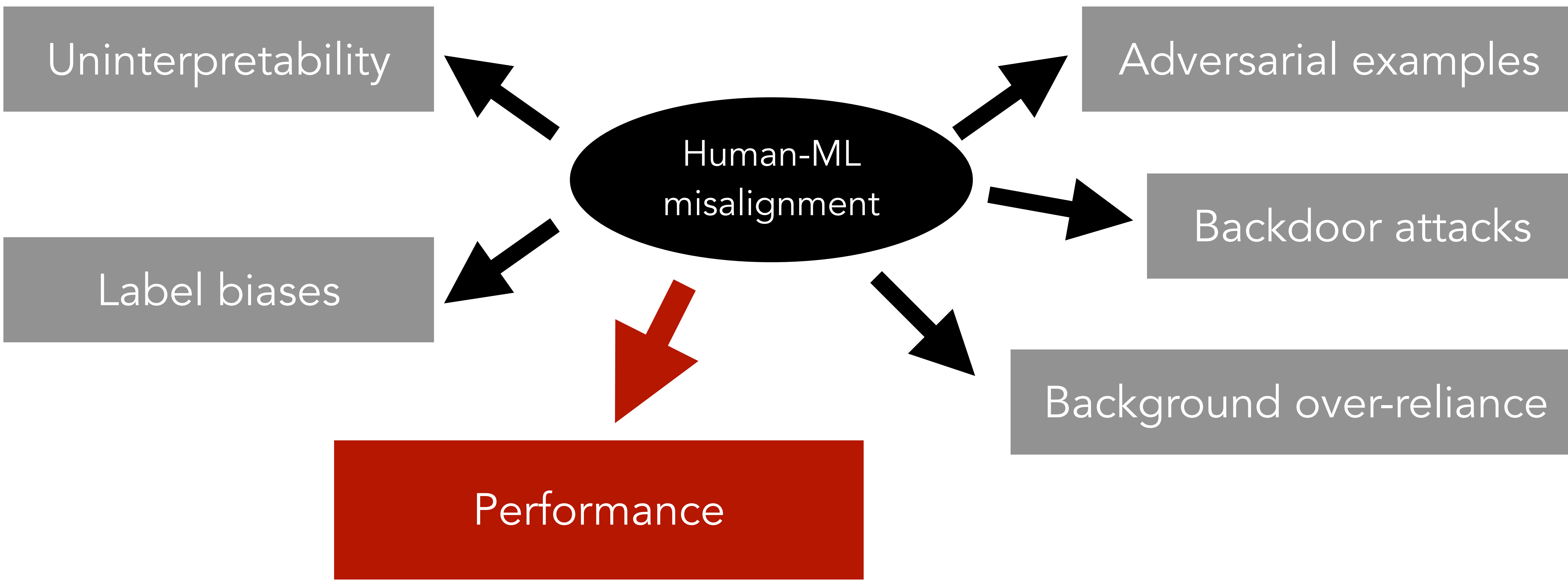
Image

Gradient

**But:** Misalignment means that the correlations extracted by the model might not be used (or even usable!) by humans

**Thus:** No hope for “free” interpretability

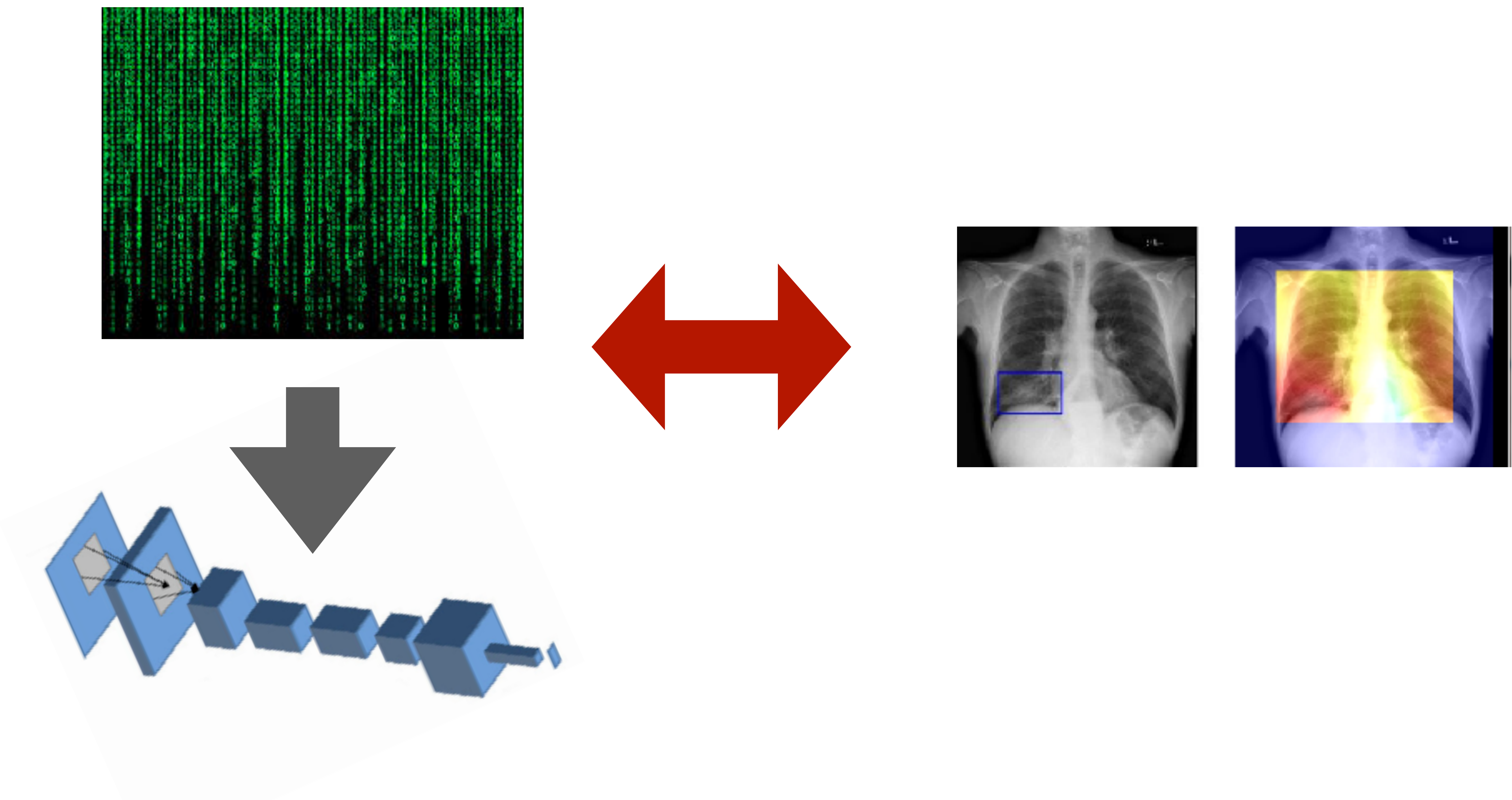
**All** the problems we discussed can be traced back to human-ML misalignment




… but it is also part of what makes machine learning so powerful

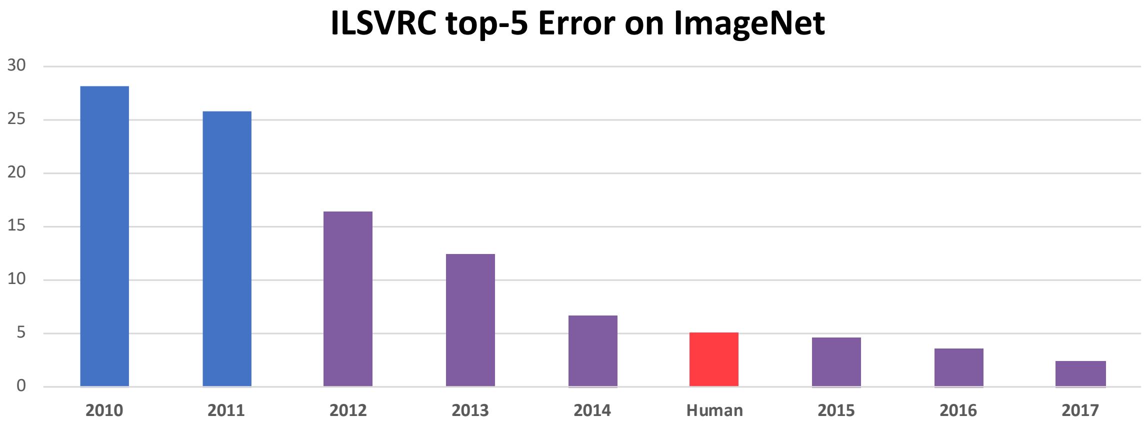
**Million- (Billions-?) dollar question:**

How to trade off the raw correlative power of modern ML with robustness, reliability and interpretability



# **Finally:** This is not at all just about vision

→ Vision is just (arguably) the most well-studied subfield of modern ML (and viewed as the most successful)


**All** the phenomena/issues we discussed arises in **all** high-stakes real-world ML deployment contexts

(One could even argue that vision might be easier as we have a "gold standard": human perception system)

# Takeways

# ML is a sharp knife—**not** a hammer

**Correction extraction** is
the (double-edged) sword of ML

**ML researchers:** Need to embrace the complexity (and messiness) of real-world data (and tasks)

**Domain practitioners:** Help clarify data generation and articulate the correct objectives

What would it take to incentivize such cooperation?

**@aleks_madry**

**gradientscience.org**