Are All Features Created Equal?

Aleksander Mądry

Based on joint works with:

Logan Engstrom

Andrew Ilyas

Shibani Santurkar

Brandon Tran

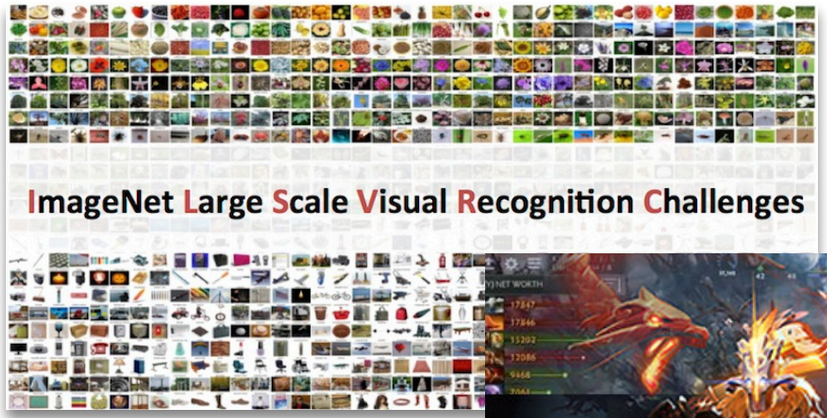
Dimitris Tsipras

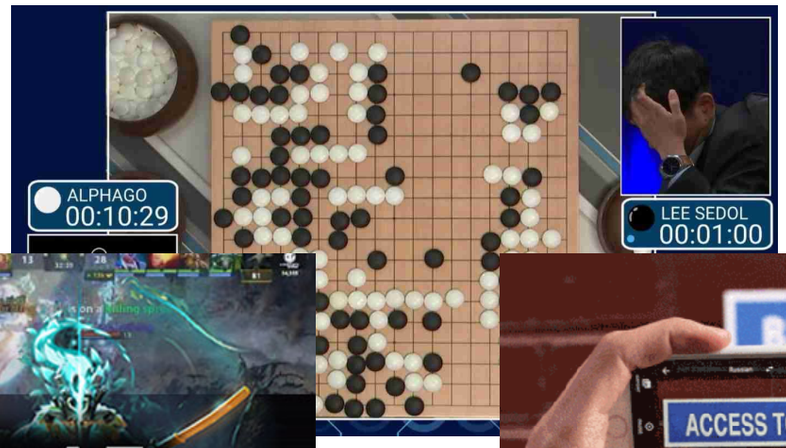
Alexander Turner

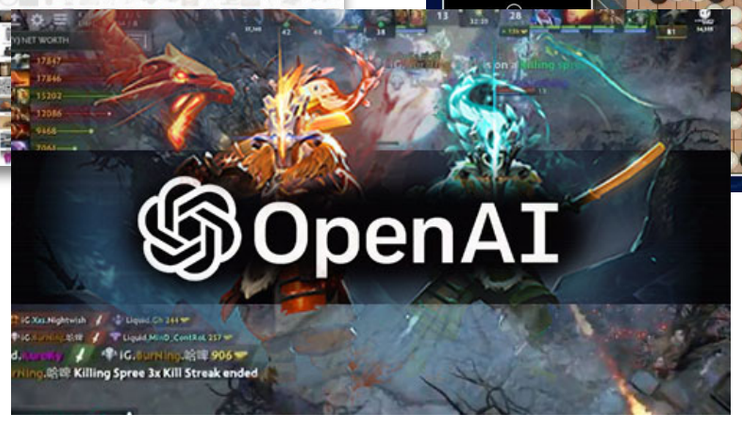
@aleks_madry

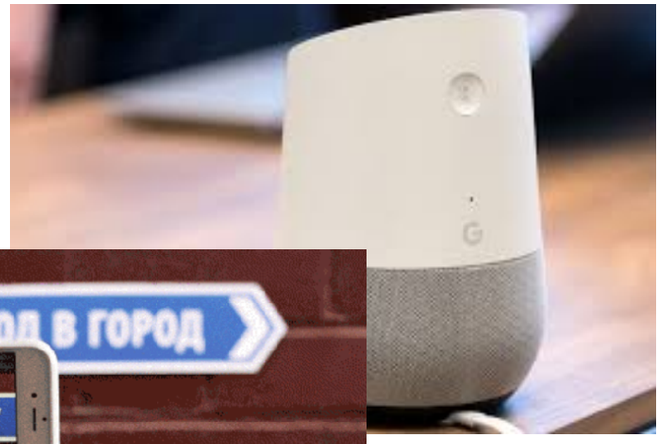
gradientscience.org

Machine Learning: A Success Story





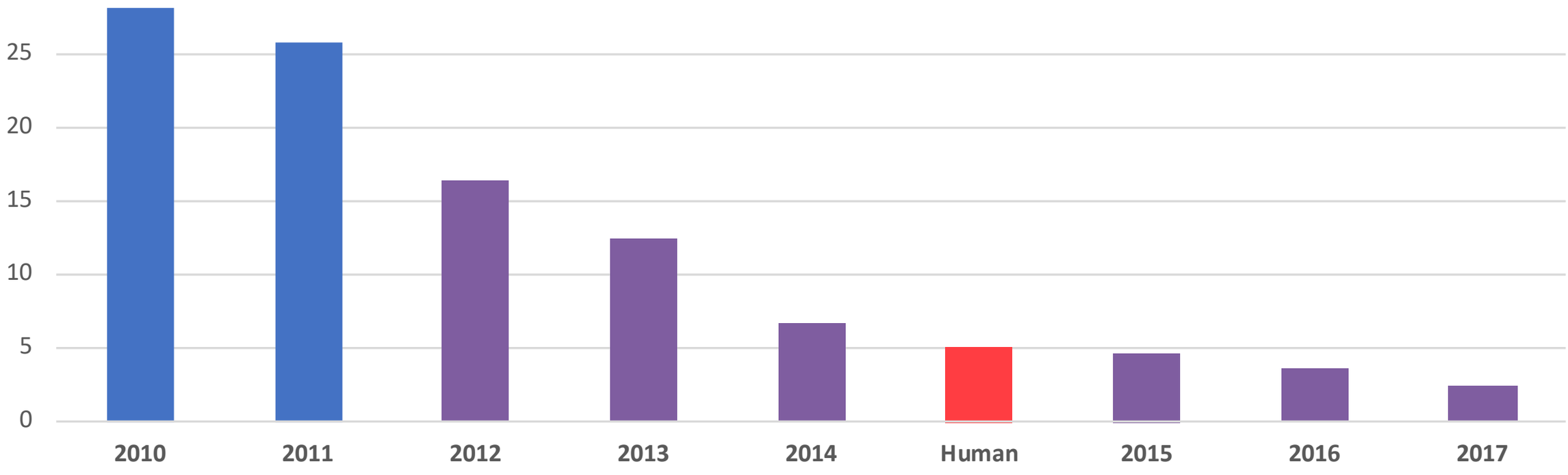




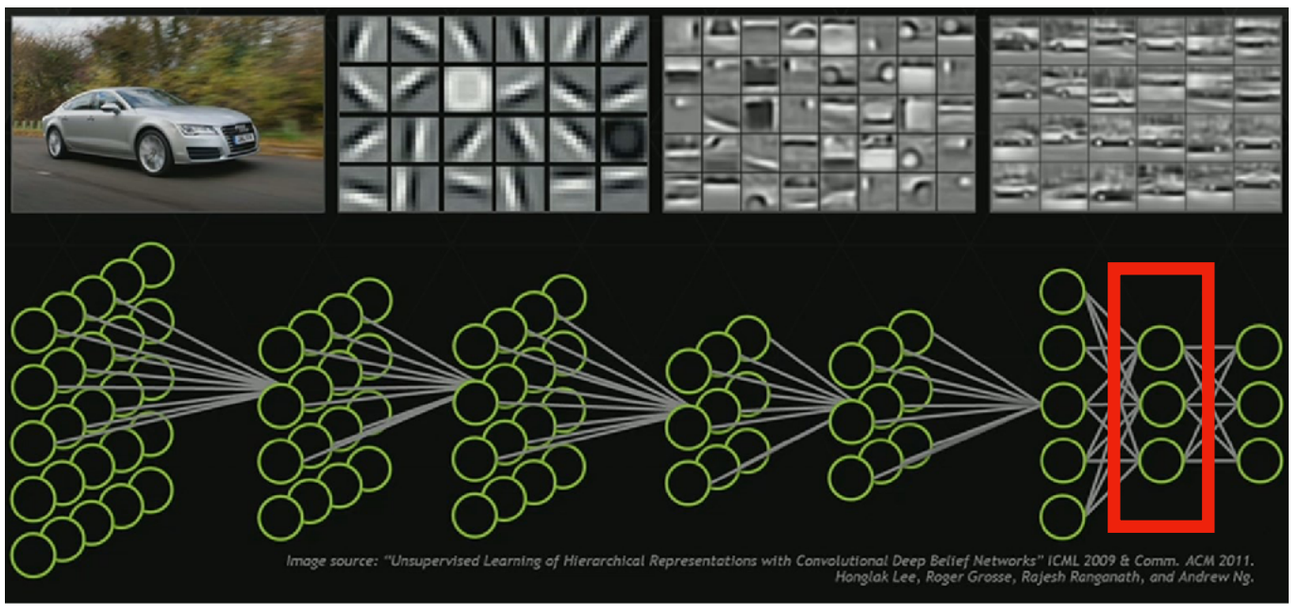
Deep learning: Fuels much of this progress

Why Do We Love Deep Learning?

ILSVRC top-5 Error on ImageNet



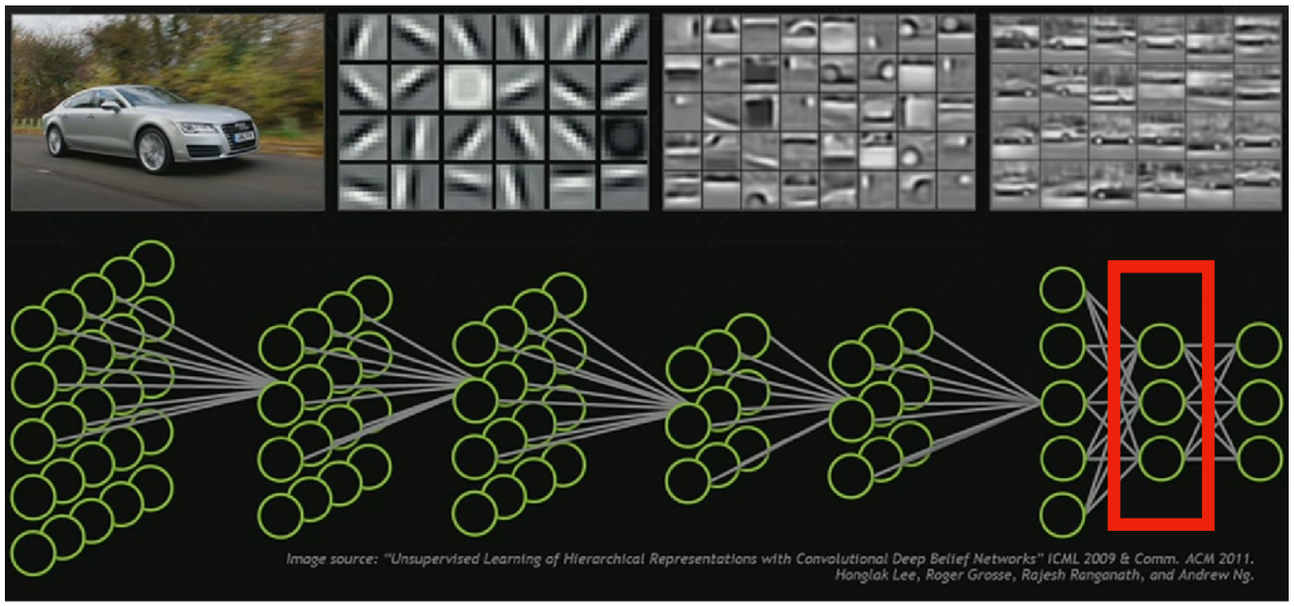
Why Do We Love Deep Learning?



→ **Meaningful** data representations

[NVIDIA GTC, 2019]

Why Do We Love Deep Learning?



[NVIDIA GTC, 2019]

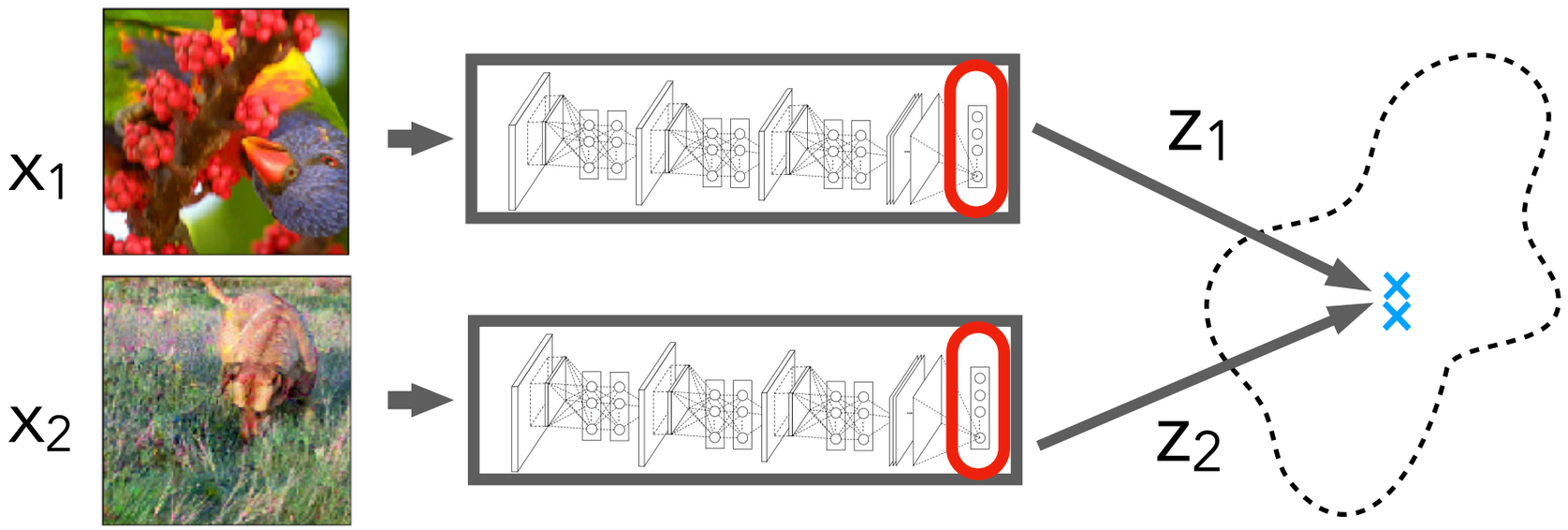
→ **Better** generative models

→ **Meaningful** data representations

[Brock et al 2018] + [Isola 2018]

But…

Correct label: insect
Predicted label: dog

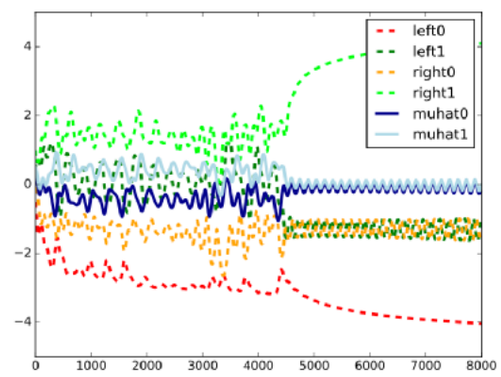


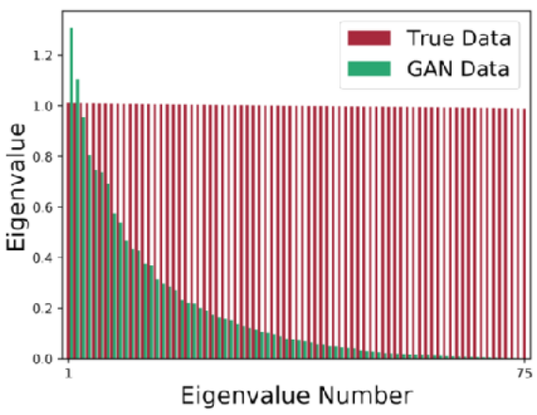
$$x_1 \neq x_2$$
but $z_1 \approx z_2$

[Engstrom et al 2019]

Unstable training/Mode collapse of generative models

[Li et al 2018] [Santurkar et al 2018]





So: Are we on the right path?

(Is all we need "just" scaling up?)

Msg today: We might want to re-think how we train our models

→ It is all about features

So: Are we on the right path?

(Is all we need "just" scaling up?)

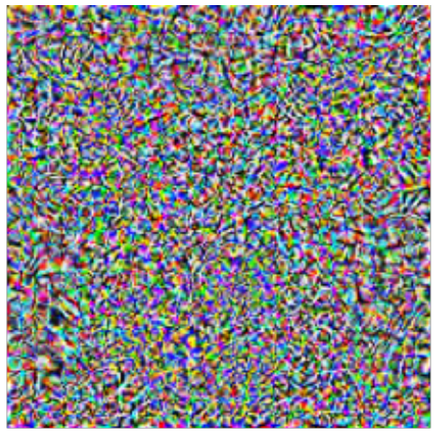
Msg today: We might want to re-think how we train our models

→ It is all about robust features

Key Phenomenon: Adversarial Perturbations

[Szegedy et al 2013] [Biggio et al 2013]

+ 0.005 x



=

“pig” (91%) **noise (NOT random)** **“airliner” (99%)**

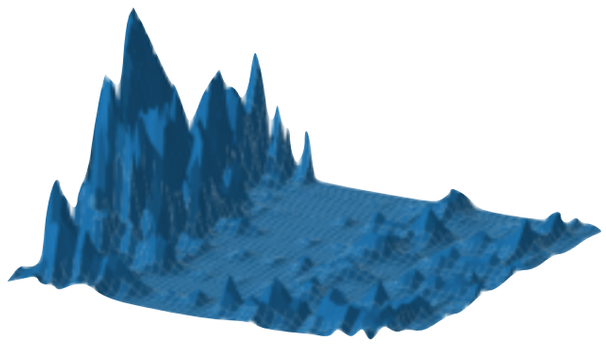
Emerging goal: (Adversarially) robust generalization

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta\in\Delta} \ell(\theta; x+\delta, y)]$$

Desired invariance

→ We are (finally) starting to succeed here

ML via Adversarial Robustness Lens



- Training is harder and models need to be more complex

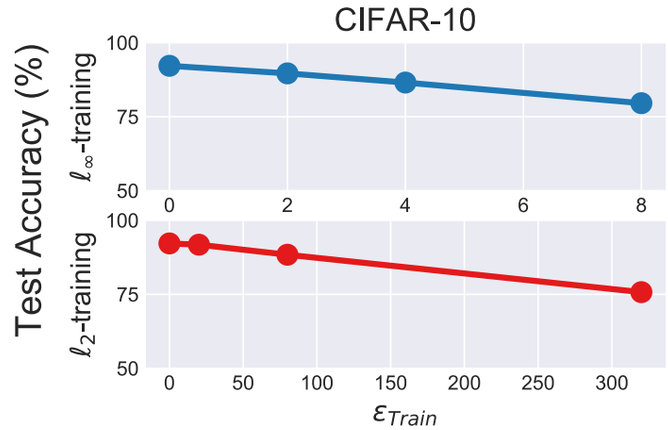
  [M Makelov Schmidt Tsipras Vladu 2018]

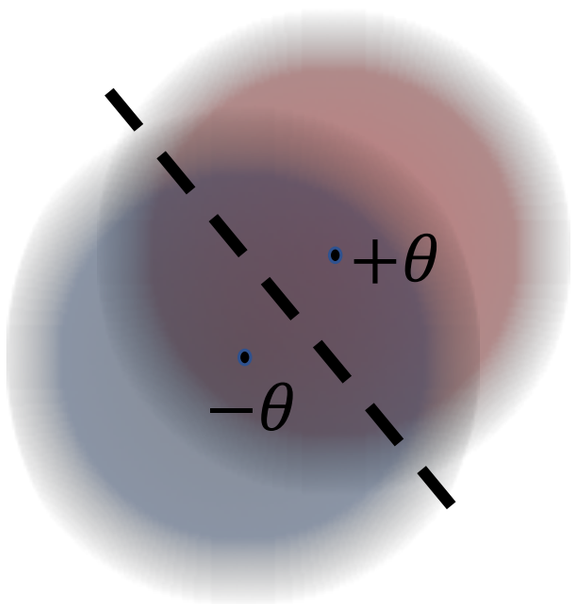
- Models may <u>have</u> to be less accurate

  [Tsipras Santurkar Engstrom Turner M 2018]

  [Bubeck Price Razenshteyn 2018]

  [Degwekar Nakkiran Vaikunatanathan 2018]





- We might need more training data

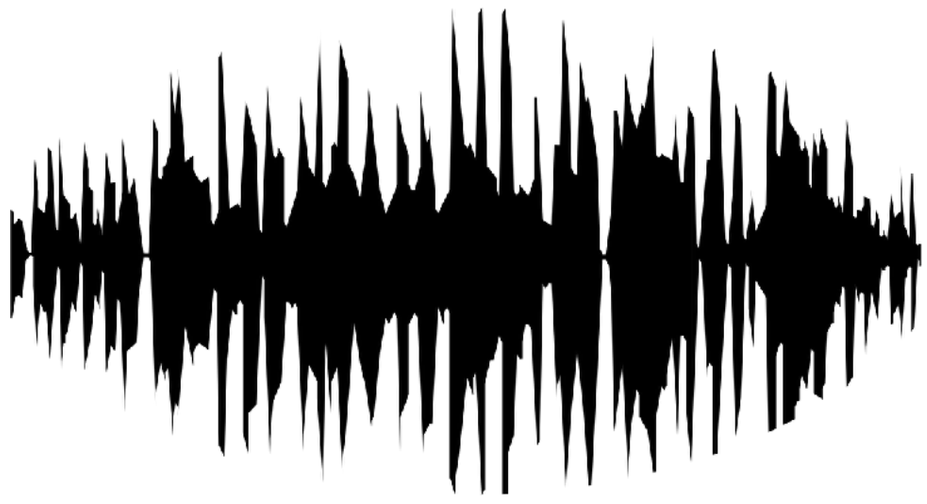
  [Schmidt Santurkar Tsipras Talwar M 2018]

But: “How”/“what” does not tell us “why”

Why adversarial perturbations **exist**
(and **are so widespread**)?

Why these perturbations tend to **transfer**?

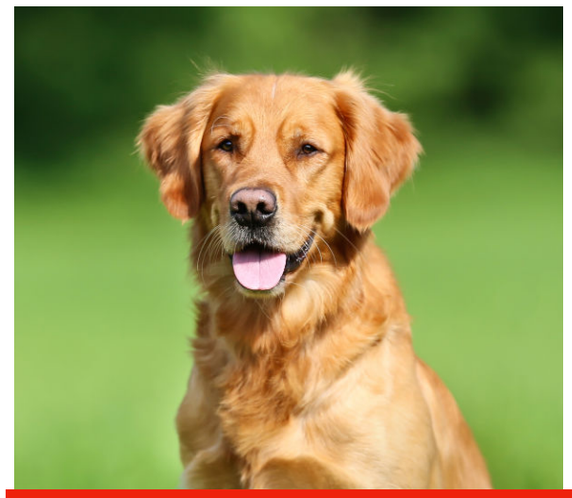
$d \to \infty$



Why are our models brittle?

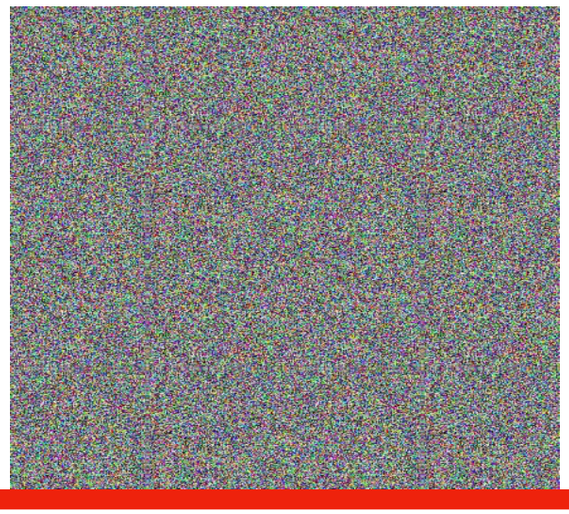
Unifying theme: Adversarial examples are <u>aberrations</u>

Why Are Adv. Perturbations Bad?



dog

+

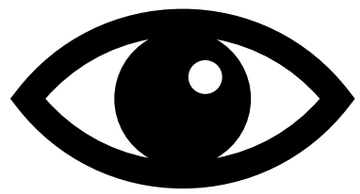


meaningless perturbation

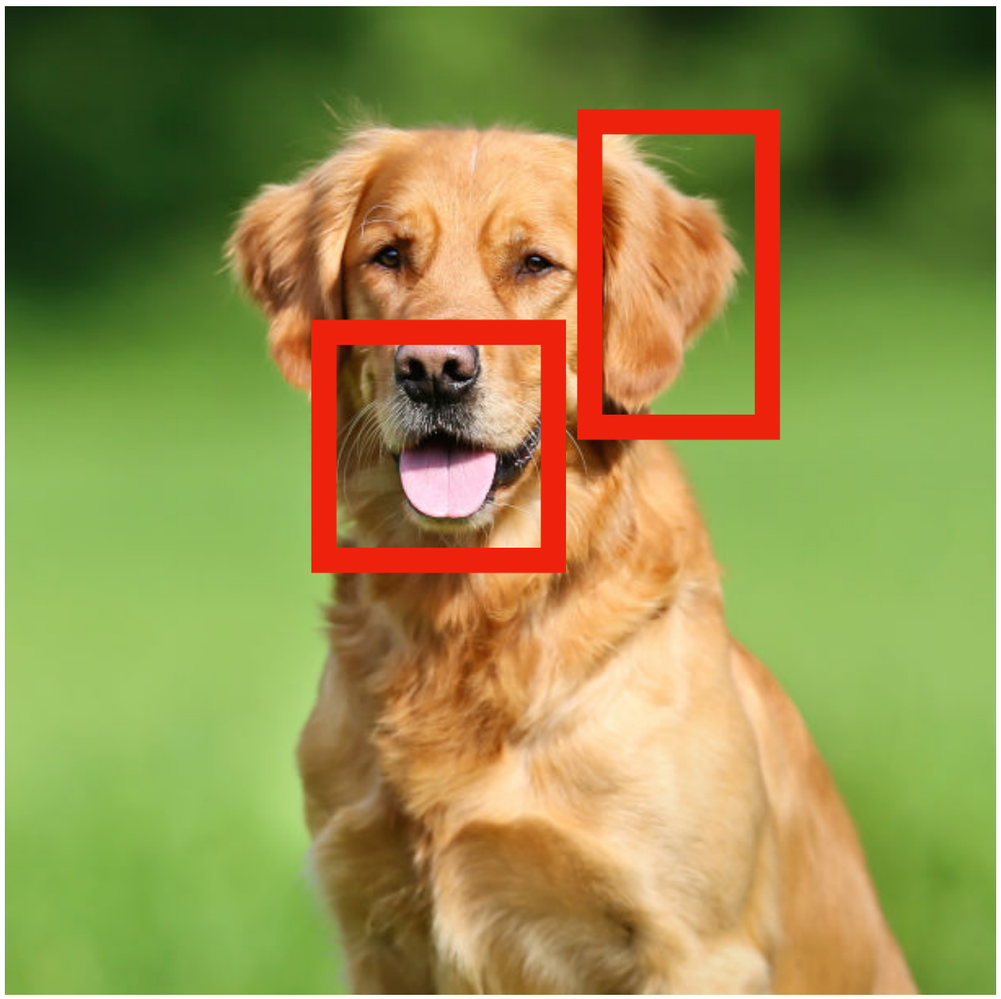
=

cat

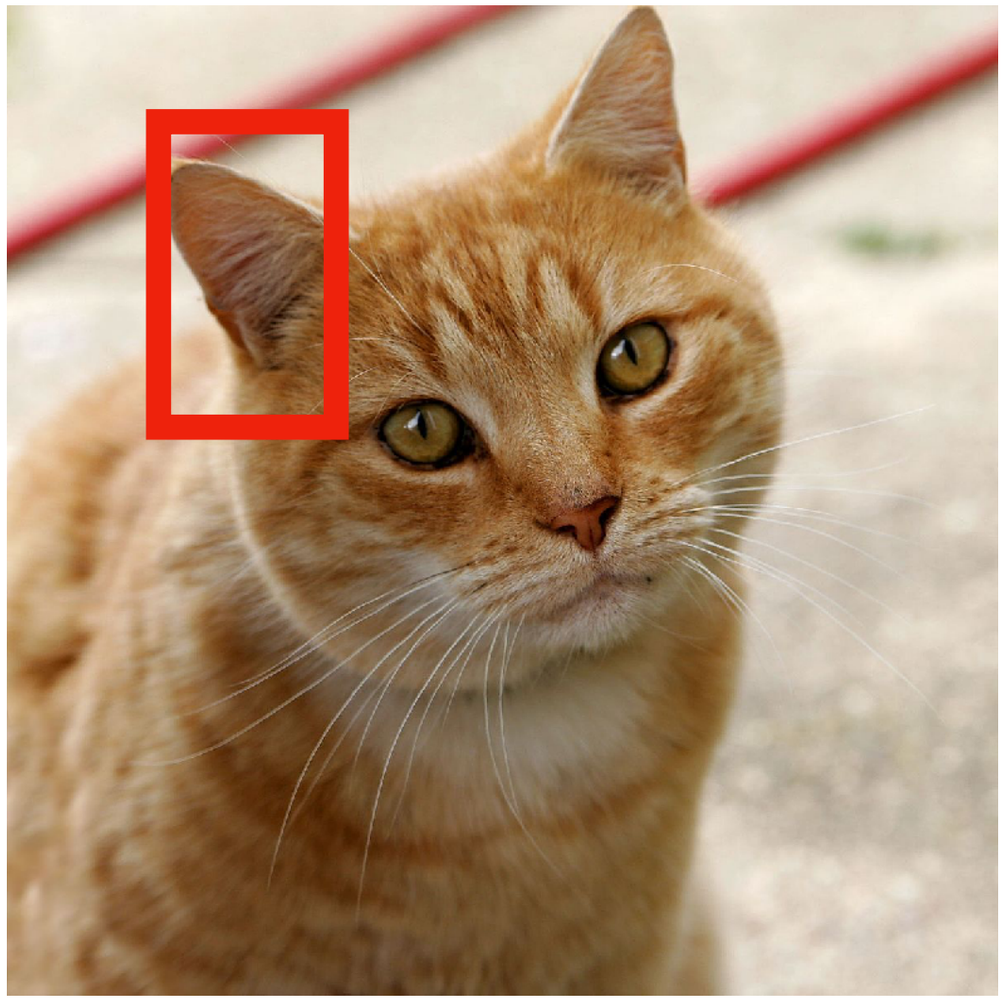
But: This is only a “human” perspective



Human Perspective



dog



cat

ML Perspective

dog

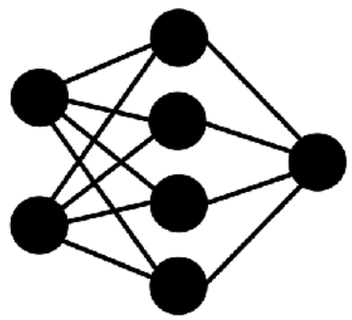


Image is meaningless

Classes are meaningless

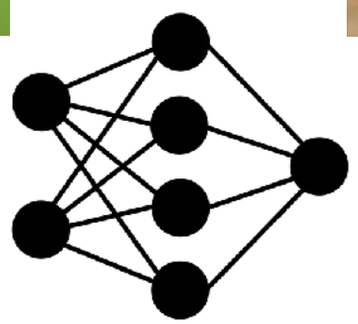
Only goal:

Max (test) accuracy

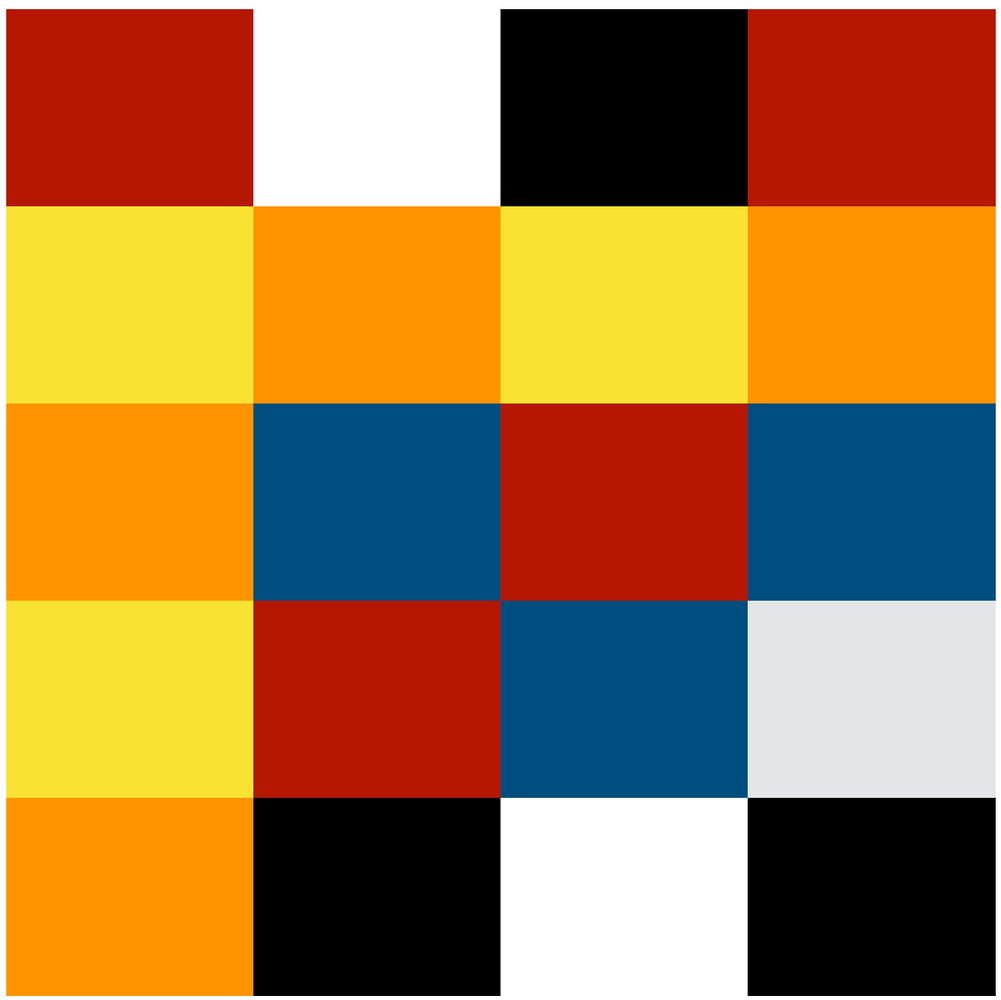
ML Perspective

dog

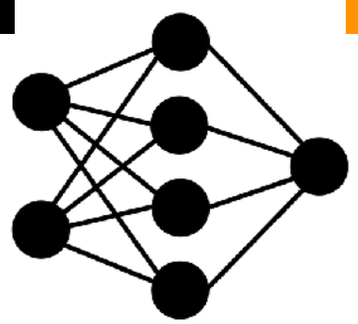
cat

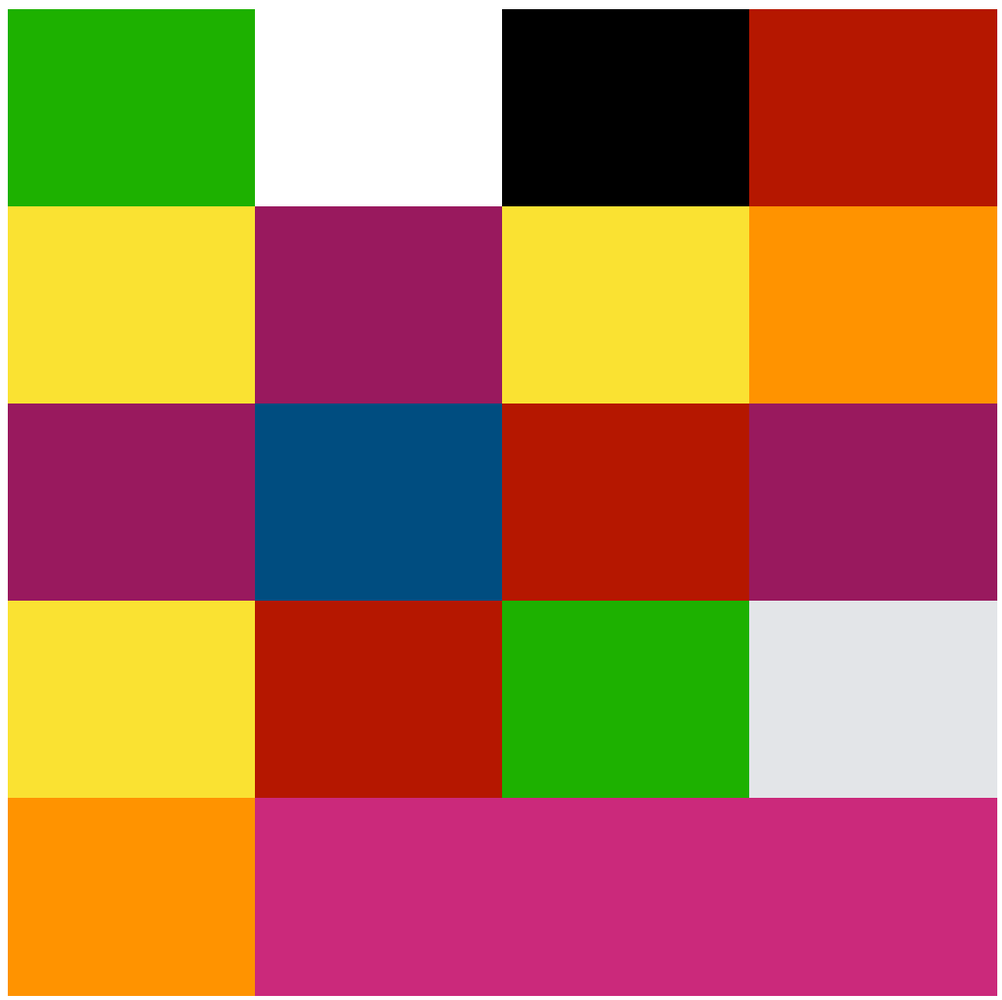


ML Perspective



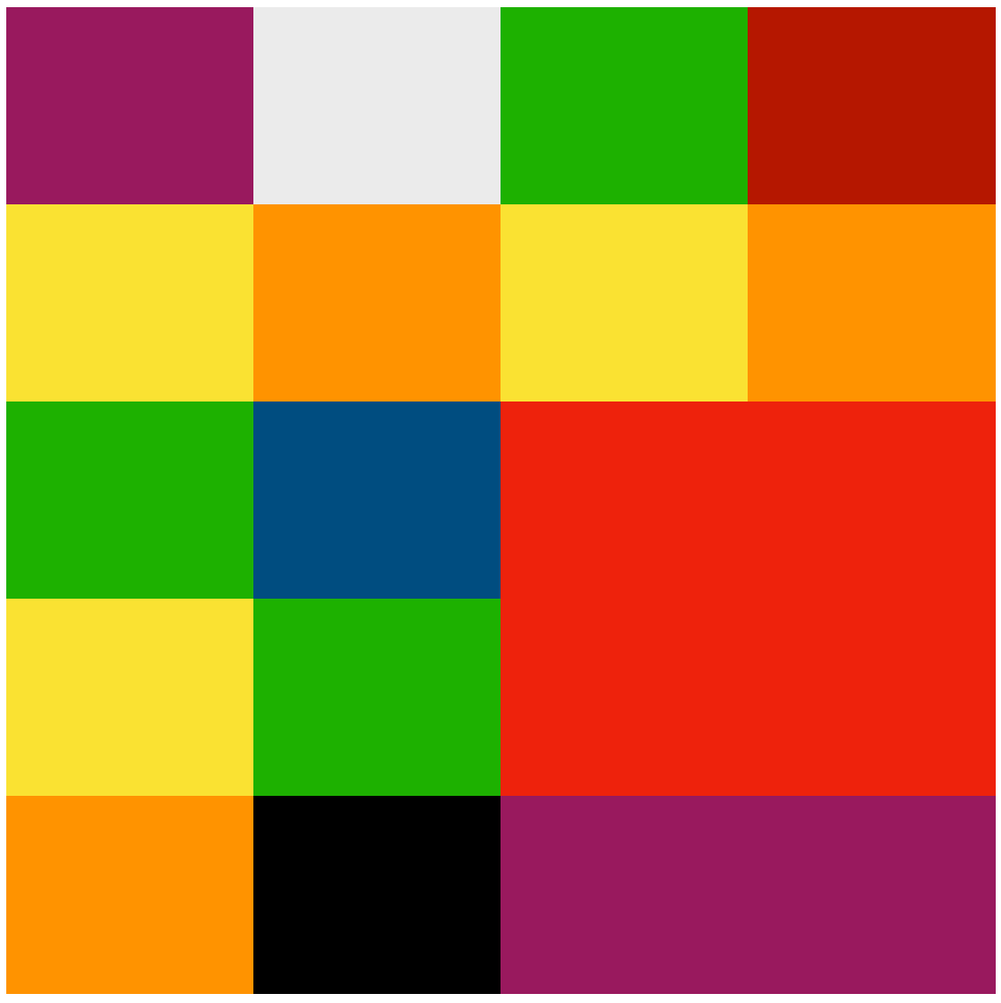
tap



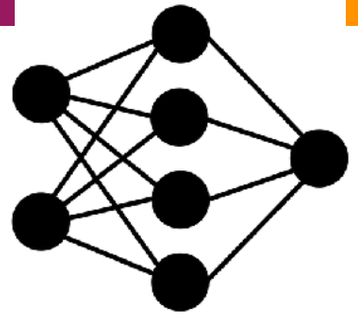


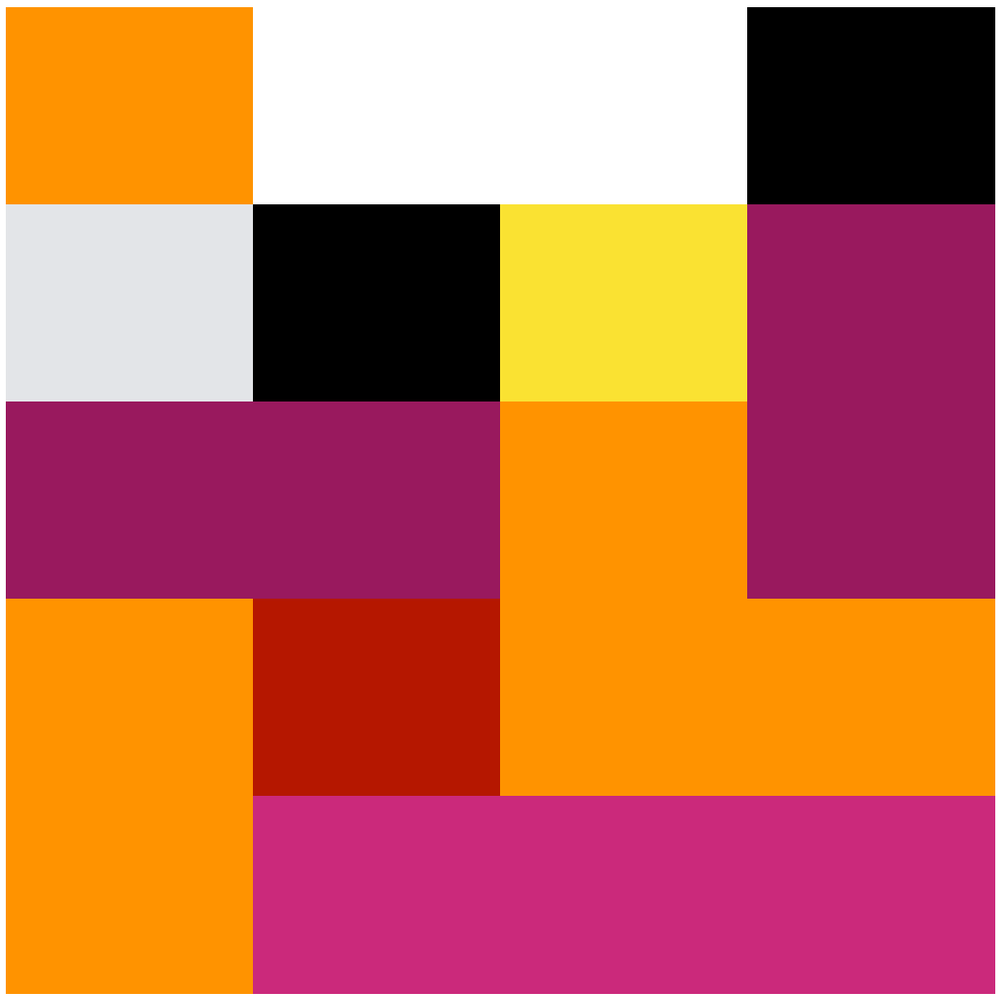
toc

ML Perspective



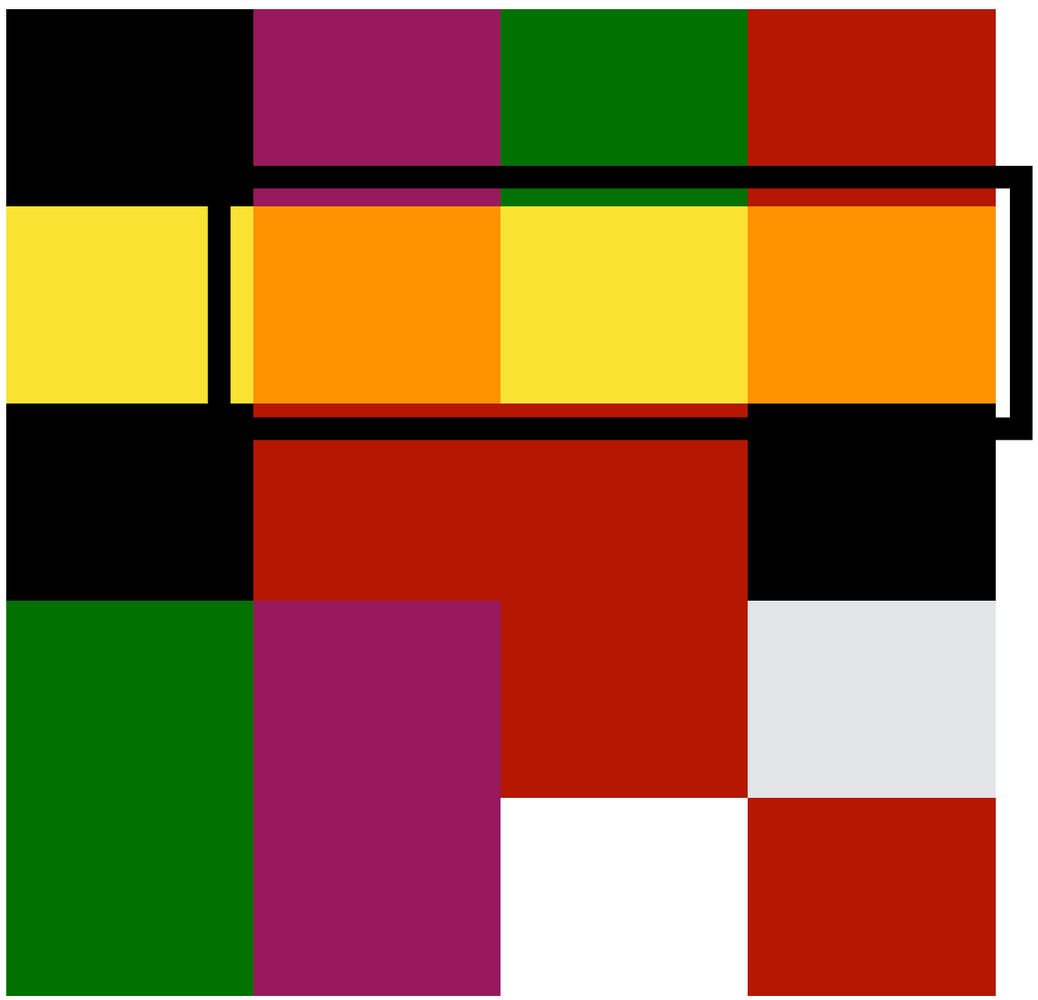
tap



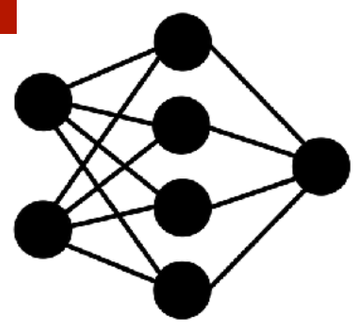


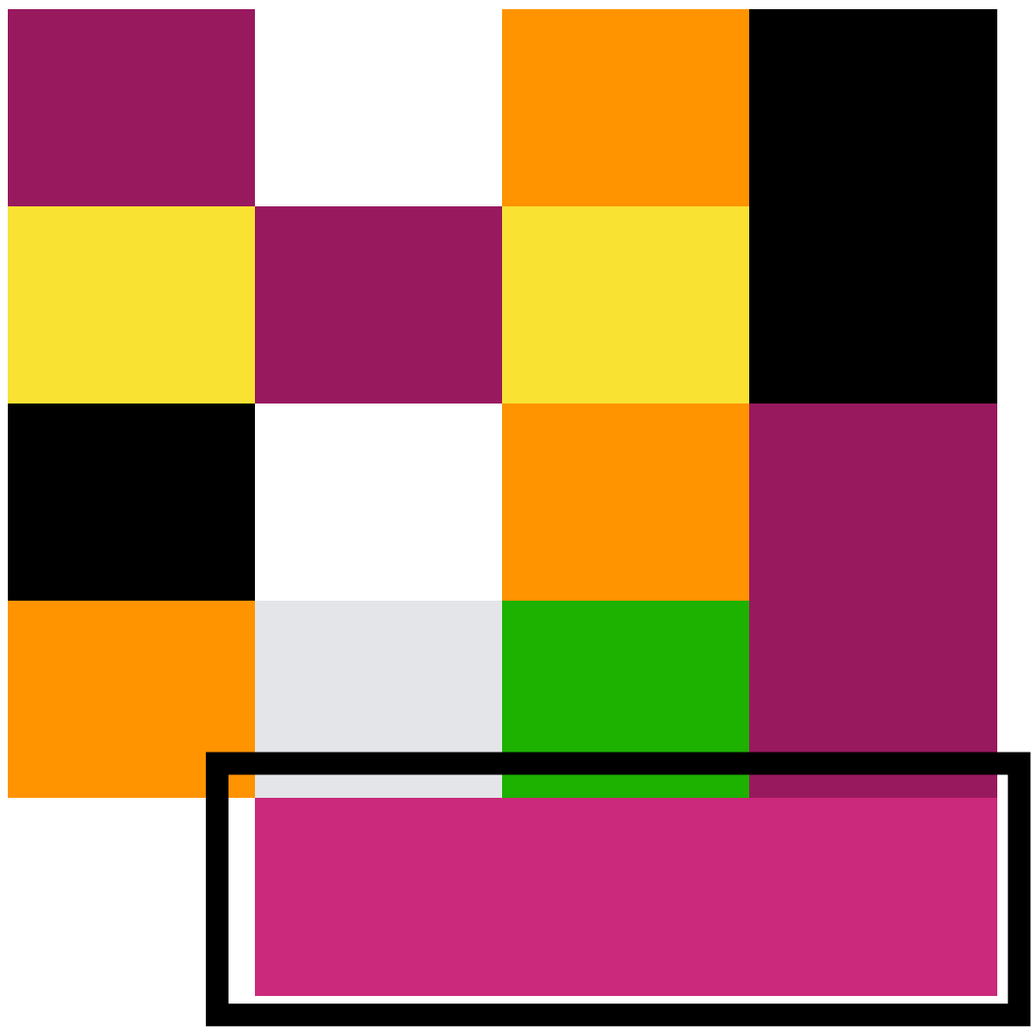
toc

ML Perspective



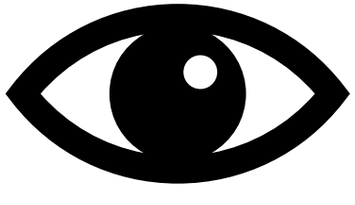
tap

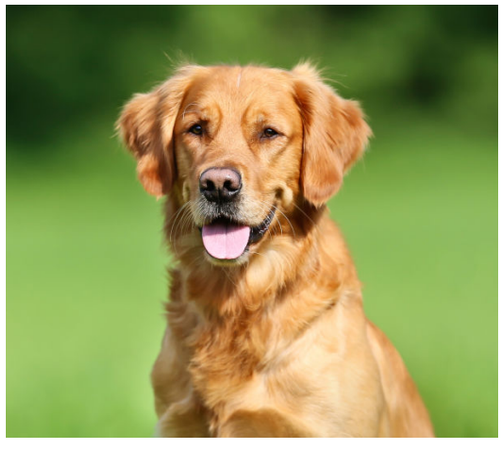




toc

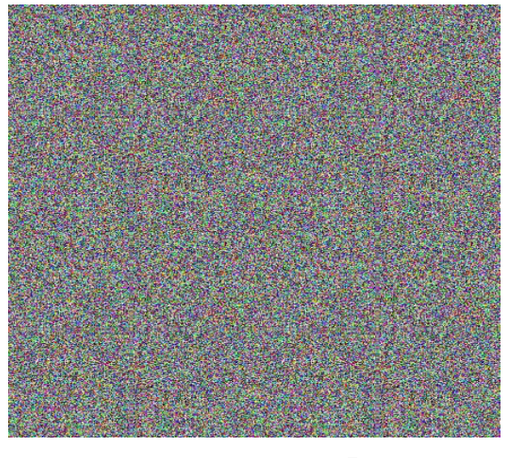
ML Perspective





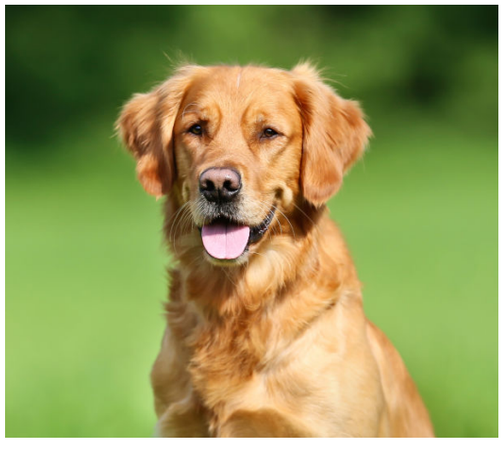
dog

+

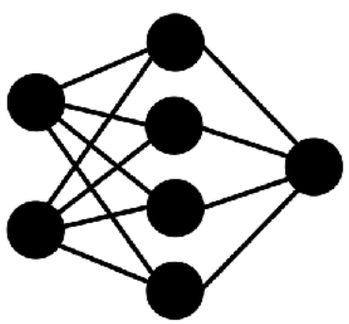


meaningless perturbation

=

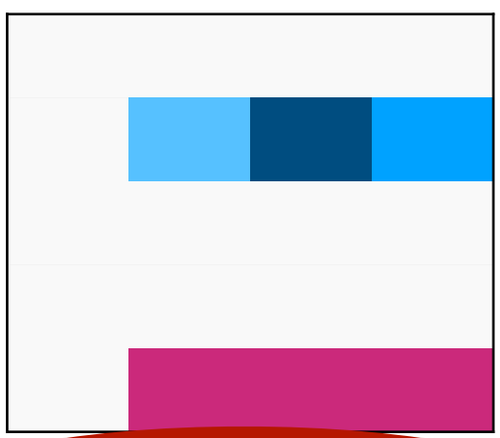


cat



tap

+



meaningless perturbation

?

=

toc

Are adversarial perturbations indeed meaningless?

[Ilyas Santurkar Tsipras Engstrom Tran M ‘19]

A Simple Experiment

Training set

dog

Adv. example towards “cat”

New training set

cat

Train

(Original) test set

1. **Make adversarial example** towards the other class
2. **Relabel** the image as the target class
3. Train with **new** dataset but test on the **original** test set

A Simple Experiment

Training set

dog

Adv. example towards “cat”

New training set

cat

Train

(Original) test set

So: We train on a “totally mislabeled” dataset but expect performance on a “correct” dataset

What will happen?

A Simple Experiment

Training set

dog

Adv. example towards “cat”

New training set

cat

Train

(Original) test set

Result: We get a **nontrivial accuracy** on the **original** classification task

(For example, 78% on the CIFAR dog vs cat)

What’s going on?

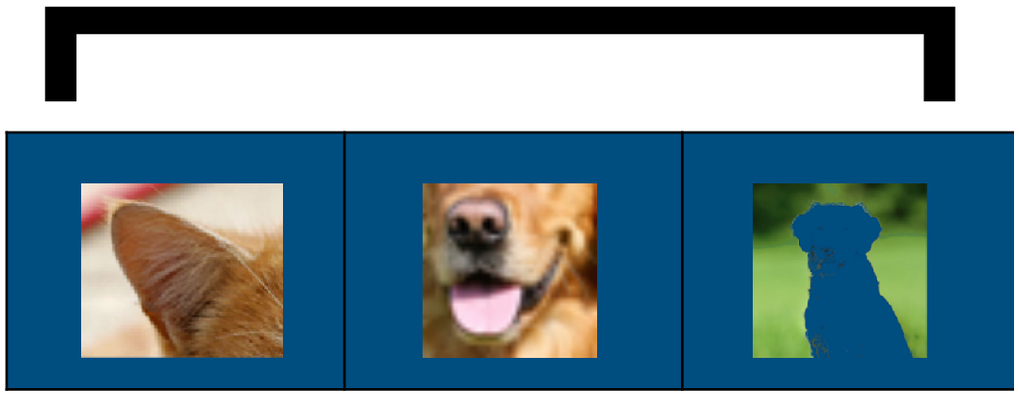
What if adversarial perturbations are **not** aberrations but **features**?

The Robust Features Model

Robust features

Correlated with label

even when perturbed



When maximizing (test) accuracy: All features are good

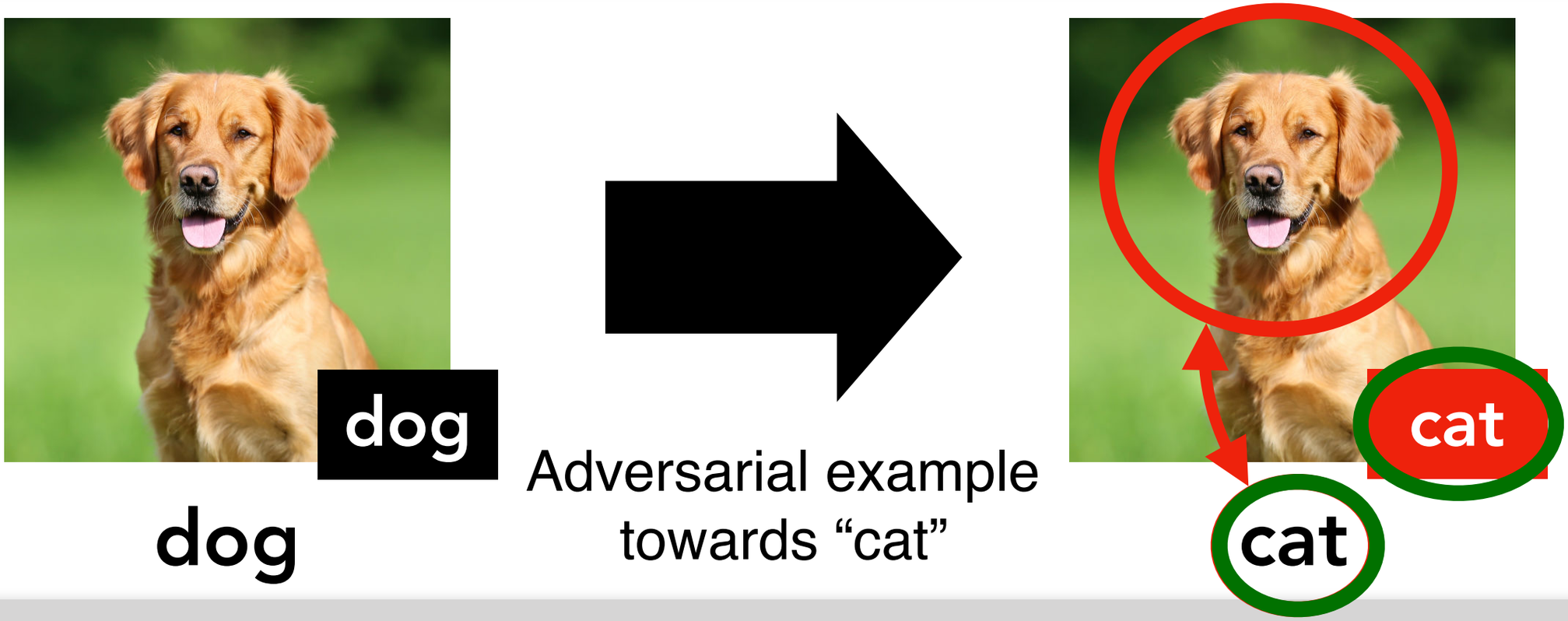
And: Non-robust features are often great!

That's why our models pick on them

(and **become vulnerable to adversarial perturbations**)

The Simple Experiment: A Second Look

All robust features are **misleading**



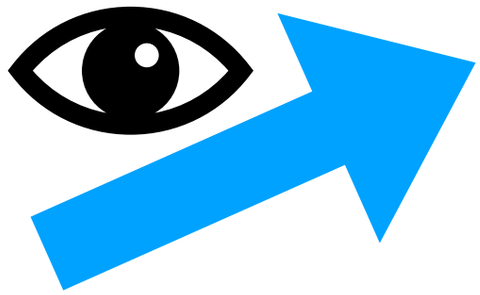
But: Non-robust features suffice for good generalization

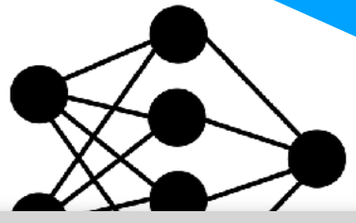
What now?

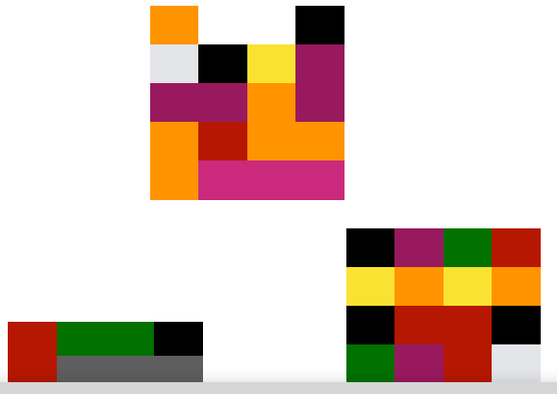
A (new) perspective on adversarial robustness

But also: Provides insight into how our models learn

Human vs ML Model Priors







These are **equally valid** classification methods

→ No reason for our models to favor the “human” one

Human vs ML Model Priors

Adversarial examples are a **human** phenomenon

No hope for **interpretable models** without intervention **at training time** (instead of post-hoc)

Need **additional restrictions (priors)** on what features models should use to make predictions

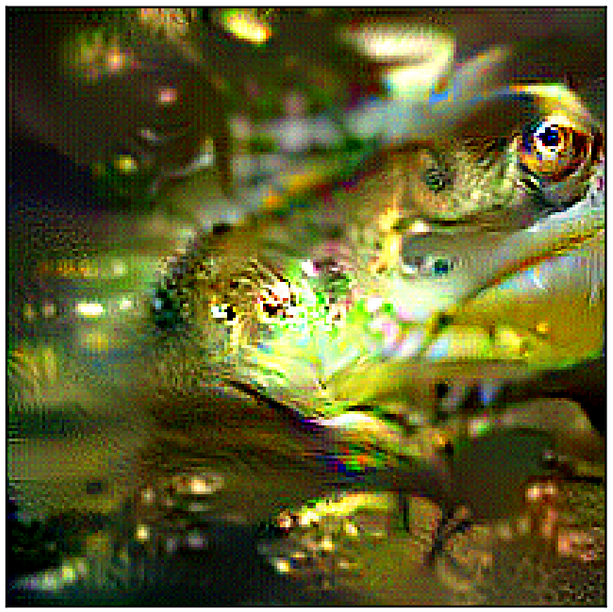
New capability: Robustification

Training set

frog

Restrict to features of robust model

New training set

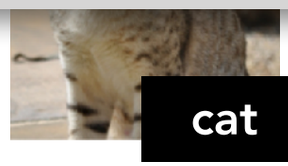


“robustified” frog

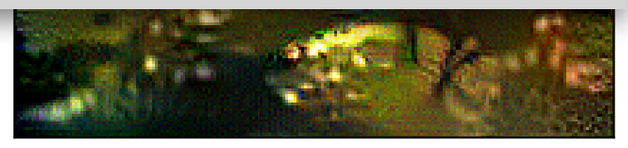
New capability: Robustification

(Original)

Also: Counterexample to any statement that "Training with BatchNorm/SGD/ResNets/ overparameterization/etc. alone leads to adversarial vulnerability"



cat



"robustified" frog

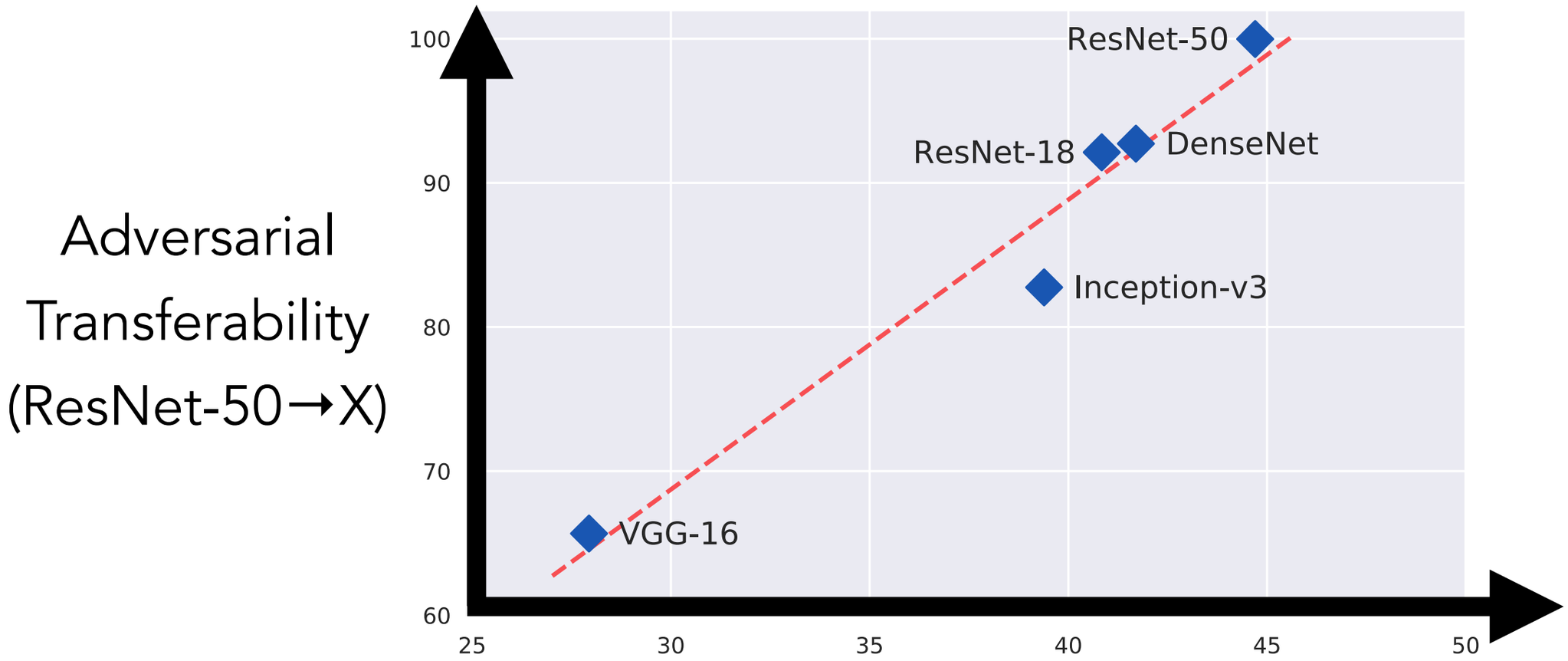
We get both standard and **robust** accuracy

So: It really is about features

Some Direct Consequences

Transferability: Features = property of **datasets** (not models)

→ Different models will tend to use the same features



Robustness and Data Efficiency

Robust models can only leverage **robust** features

(Even though non-robust features **do** help with generalization)

→ Need **more data** to get a given (robust) accuracy

(vide **[Schmidt Santurkar Tsipras Talwar M '18]**)

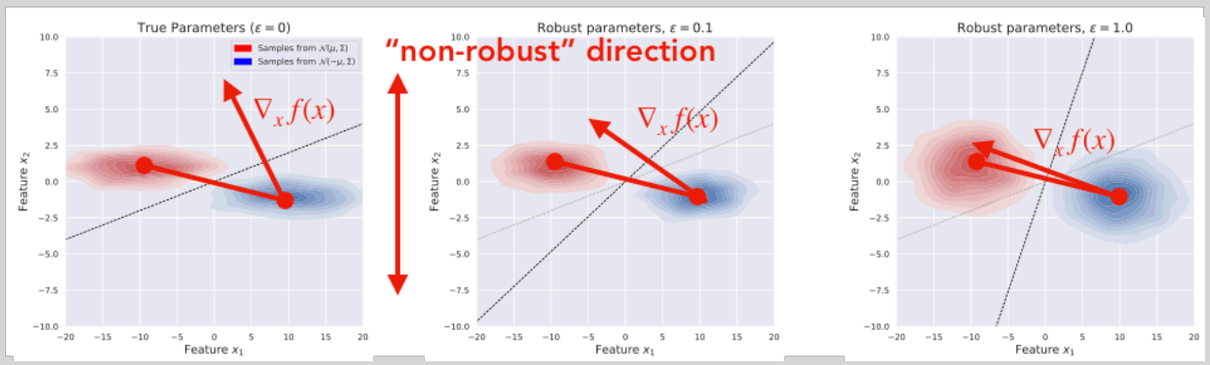
→ Will get a **lower standard accuracy**

(vide **[Tsipras Santurkar Engstrom Turner M '18]**)

Good setting to study:

Robust Max Likelihood

Gaussian Classification



[Ilyas Santurkar Tsipras Engstrom Tran M '19]

But: Is leveraging non-robust features even a good thing?

What if we **prevent** that?

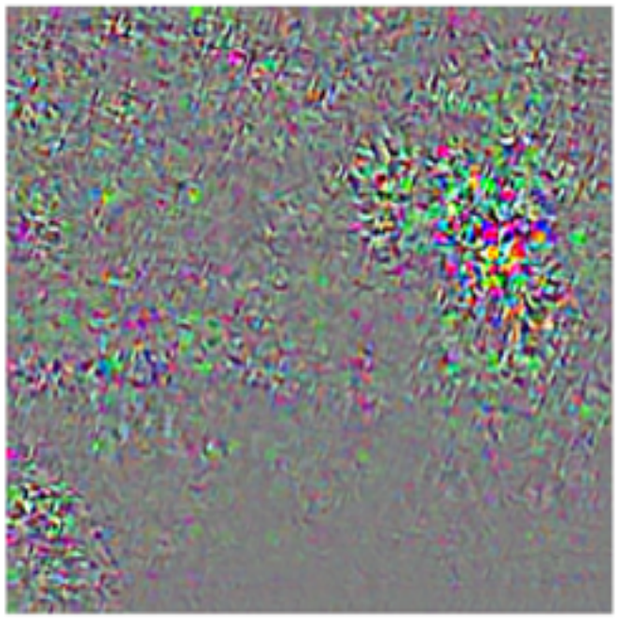
[Tsipras Santurkar Engstrom Turner M ’18]

[Engstrom Ilyas Santurkar Tsipras Tran M ’19]

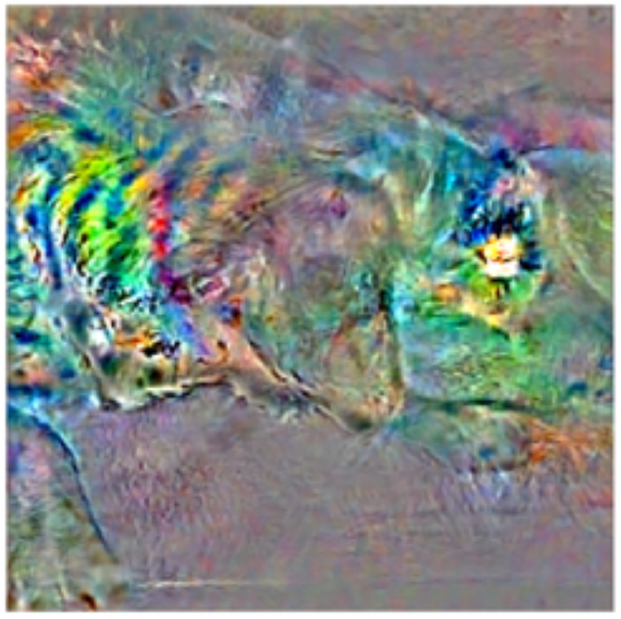
[Santurkar Tsipras Tran Ilyas Engstrom M ’19]

Robustness → Perception Alignment

Prediction: dog



Pixel influence "heatmap" (standard)

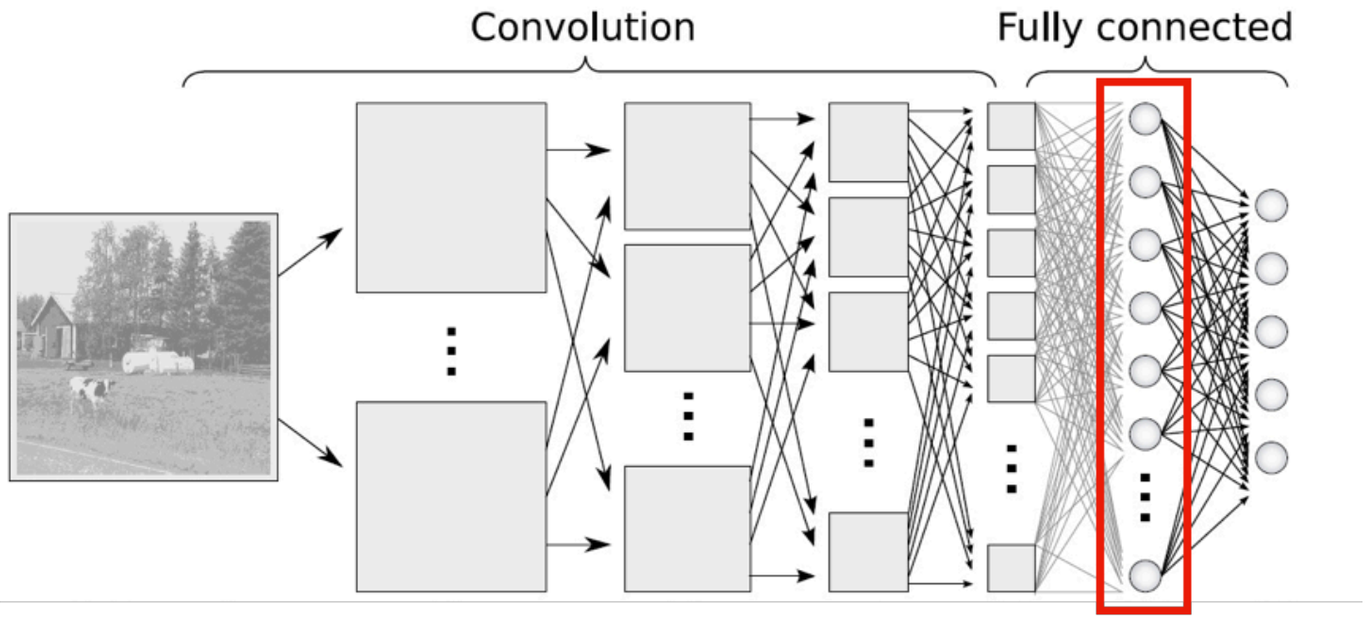


Pixel influence "heatmap" (robust)

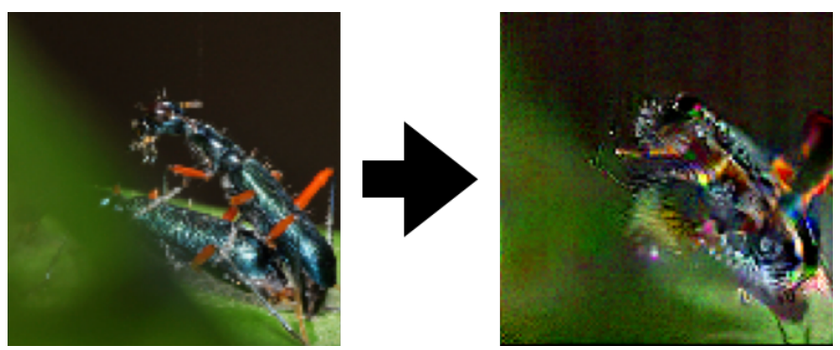
Models become more (human) perception aligned

→ Robustness acts as a **prior** for "meaningful" features

Robustness → Better Representations

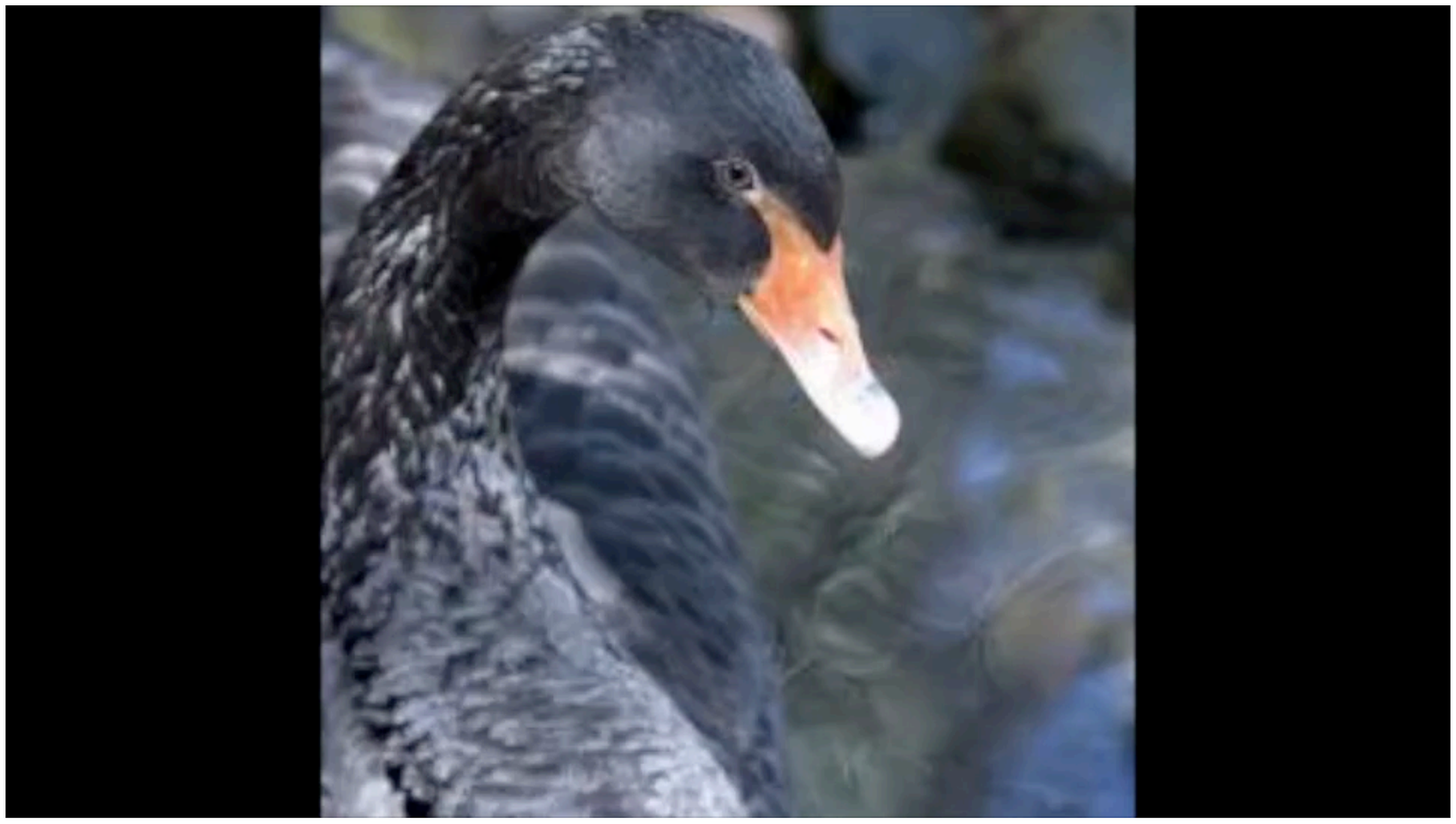


Standard Representation



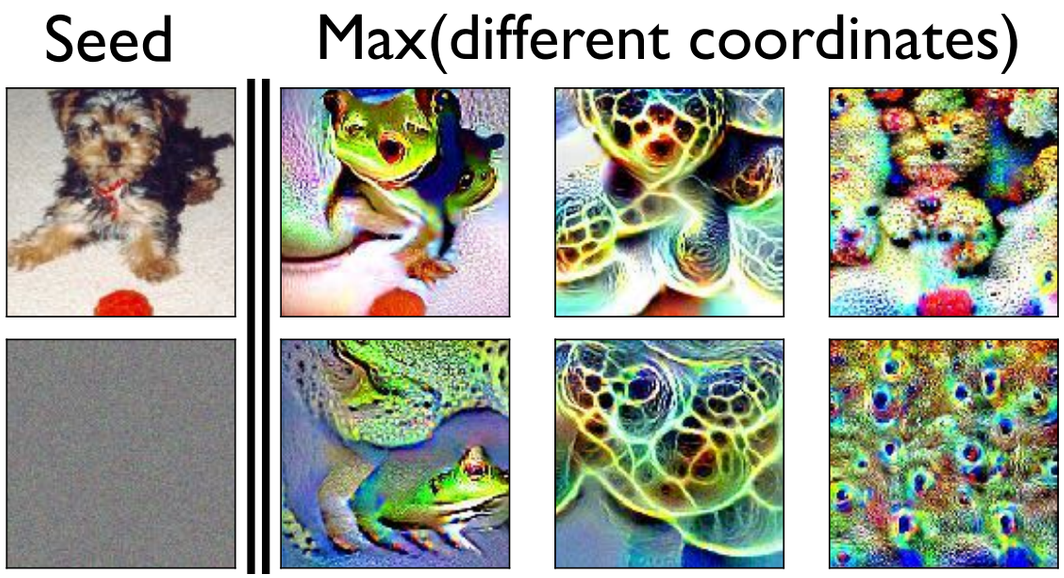
Robust Representation

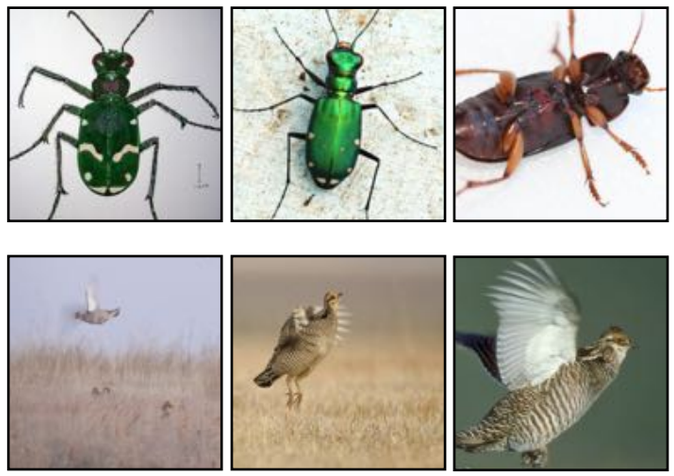
Robustness → Better Representations



Interpolations between **any** two inputs

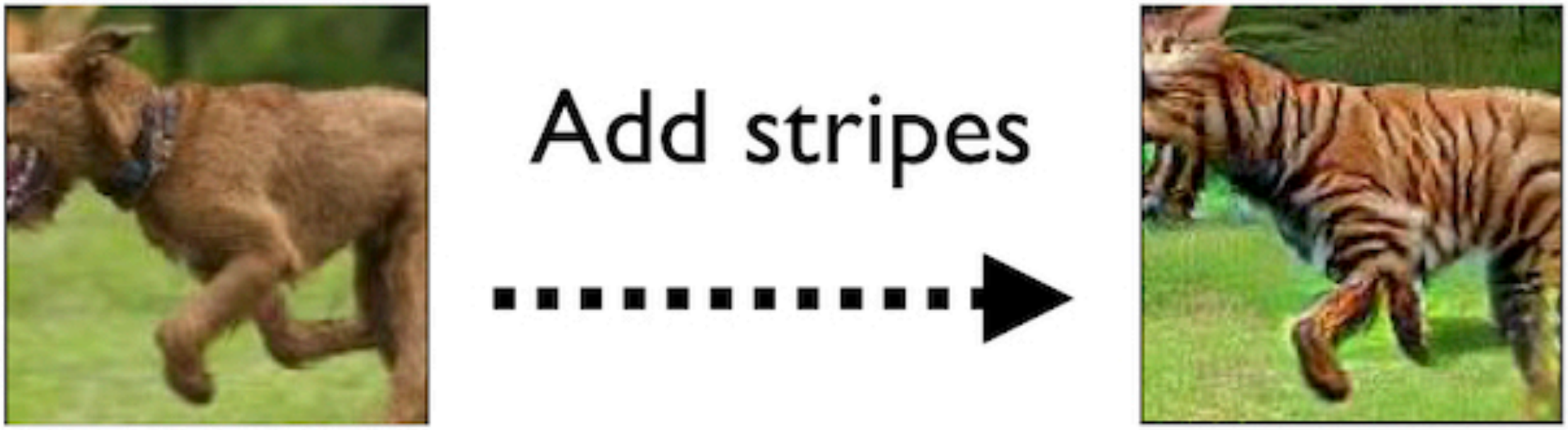
Robustness → Better Representations





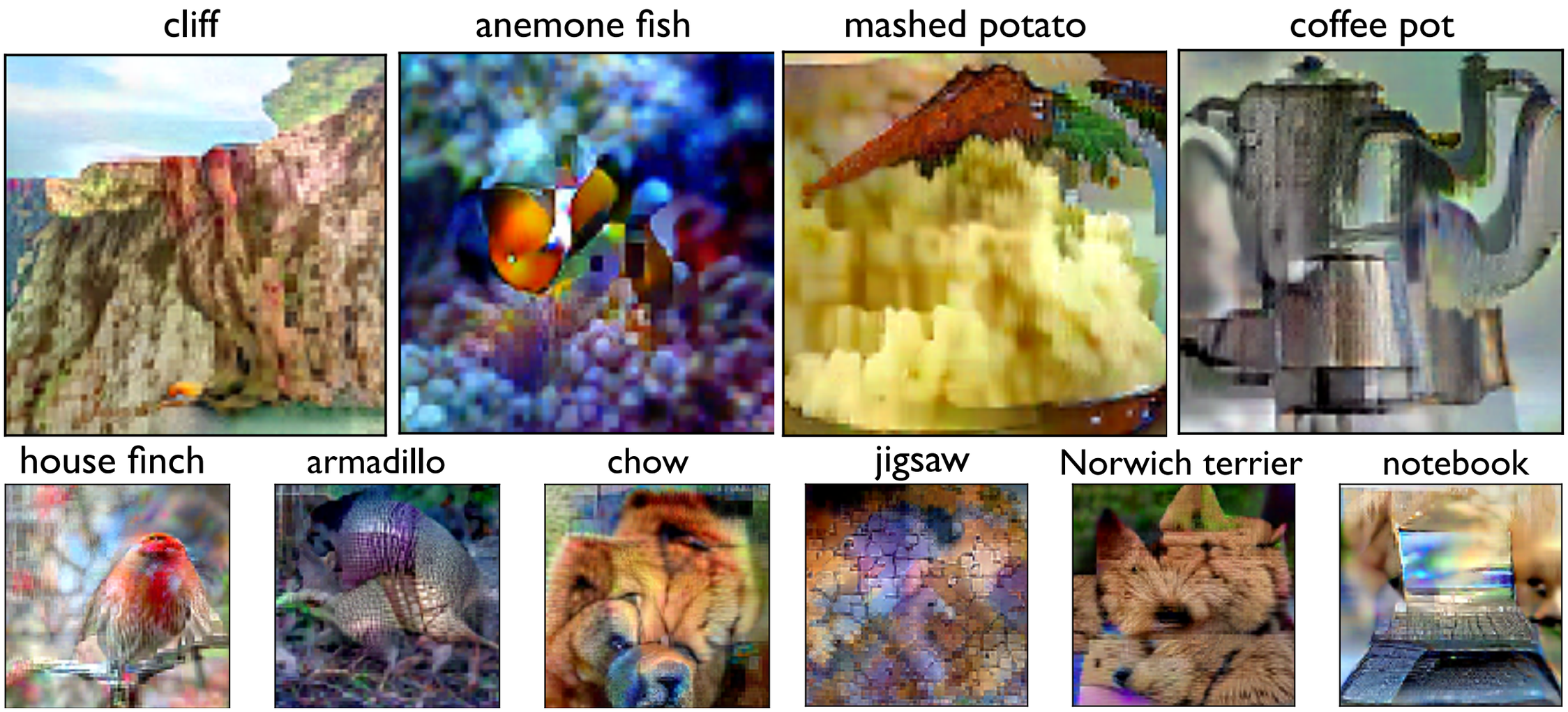
Direct feature visualization

Robustness → Better Representations



Direct feature manipulation

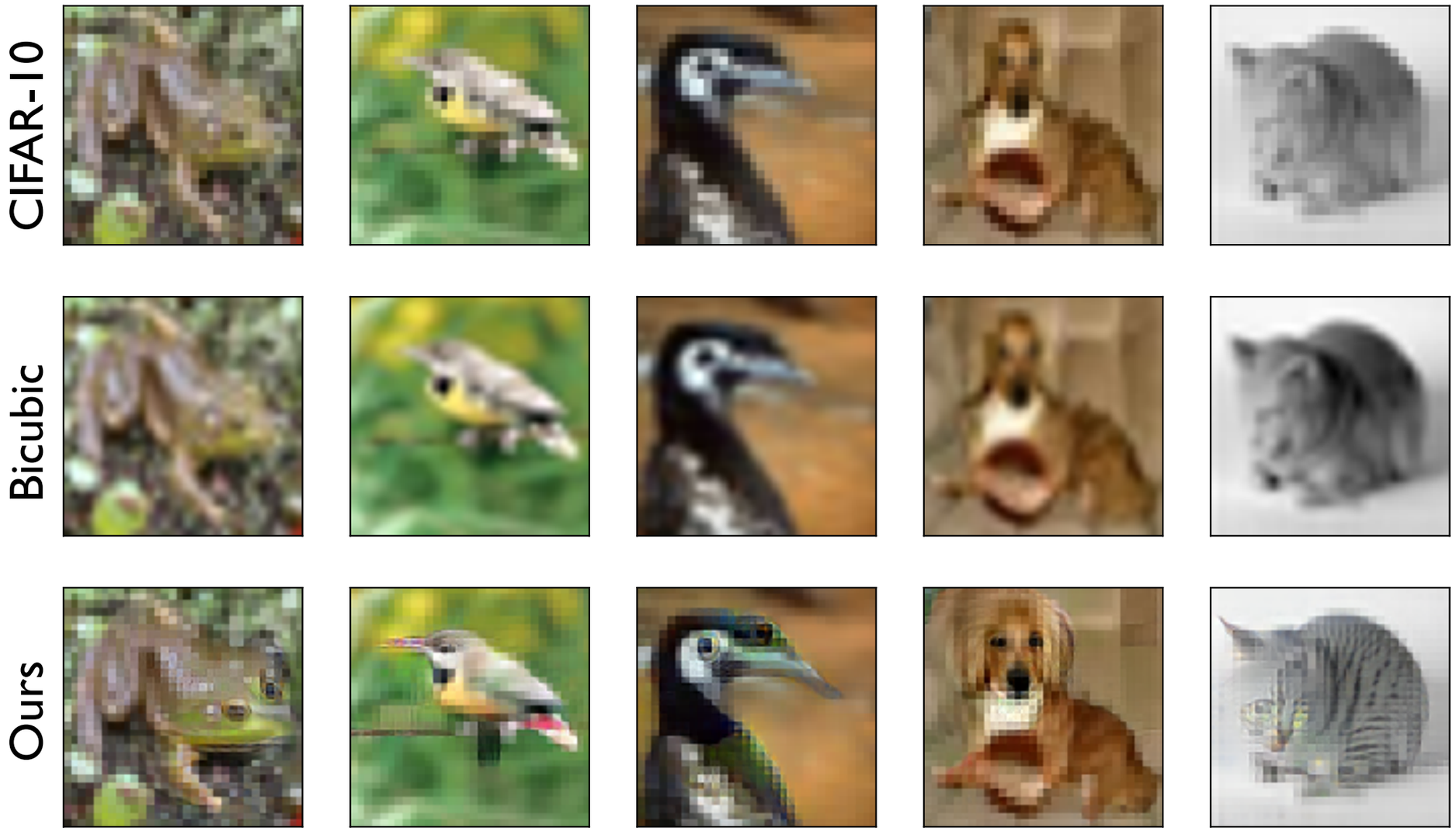
Robustness → Image Synthesis



(Random samples, 1K training images, no tuning)

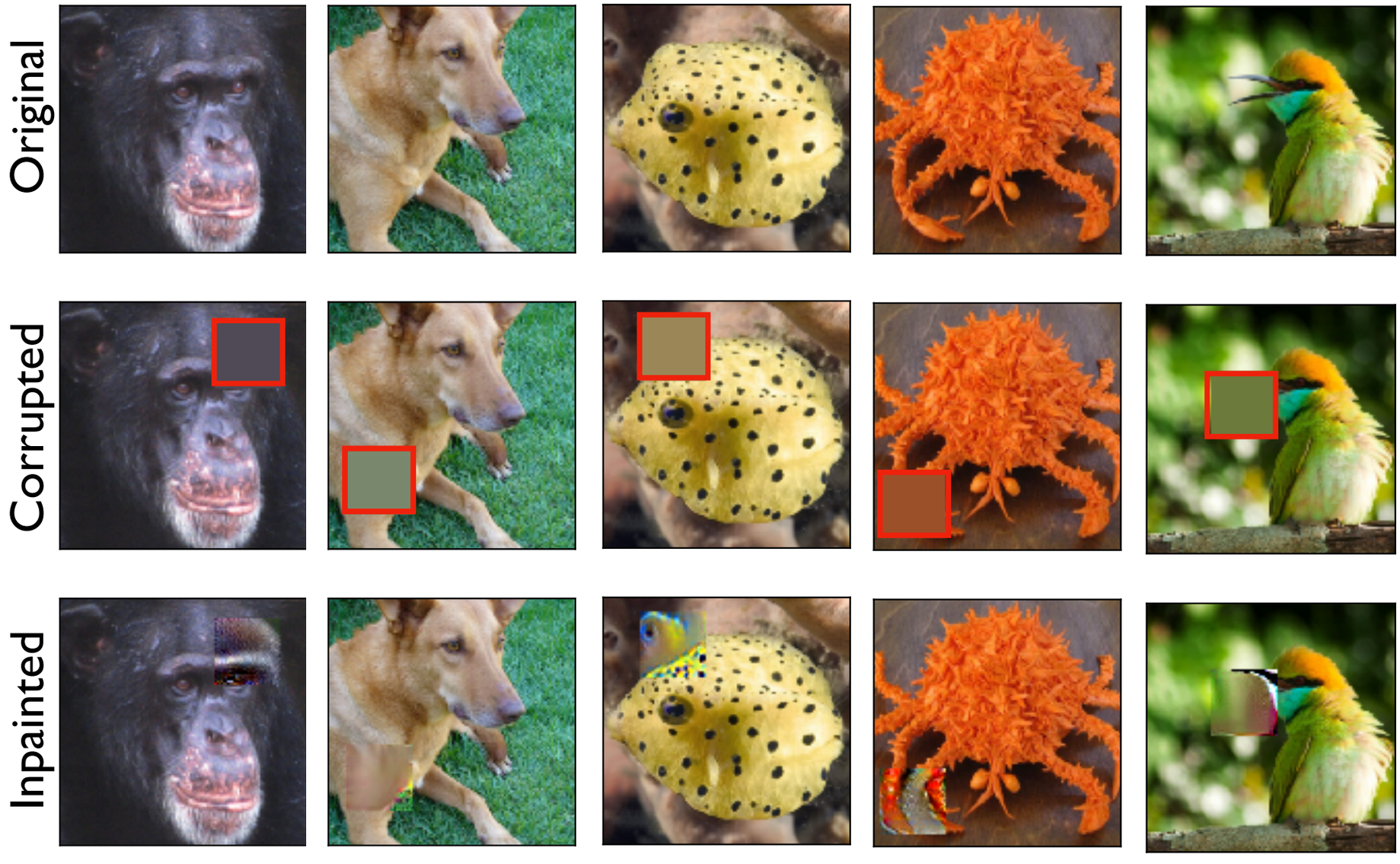
Generative models (that work **better** on **large** datasets)

Robustness → Image Synthesis



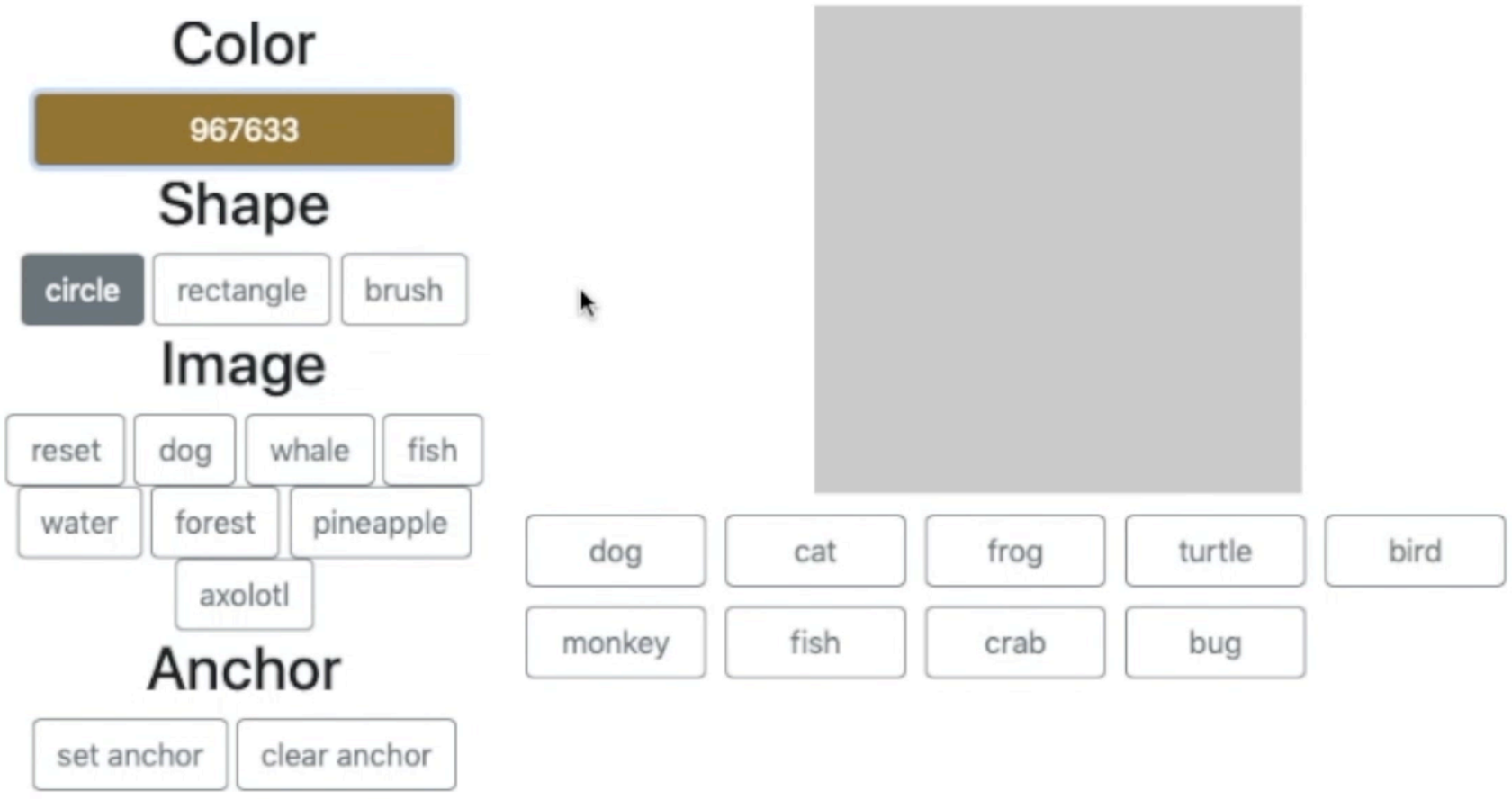
Super-resolution

Robustness → Image Synthesis



In-painting

Robustness → Image Synthesis



Interactive exploration of the data space

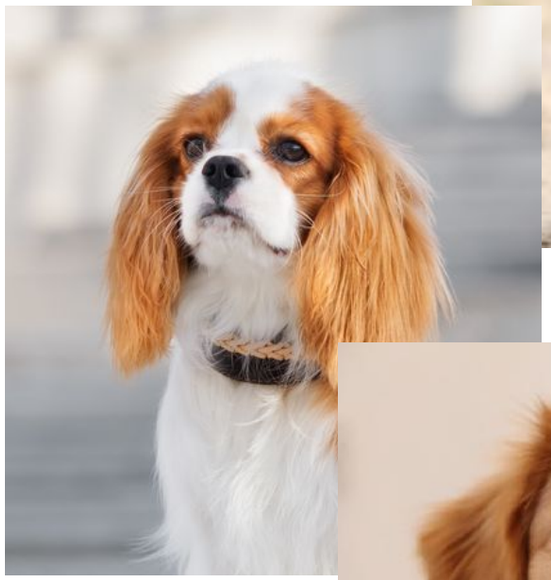
See: http://bit.ly/robustness_demo

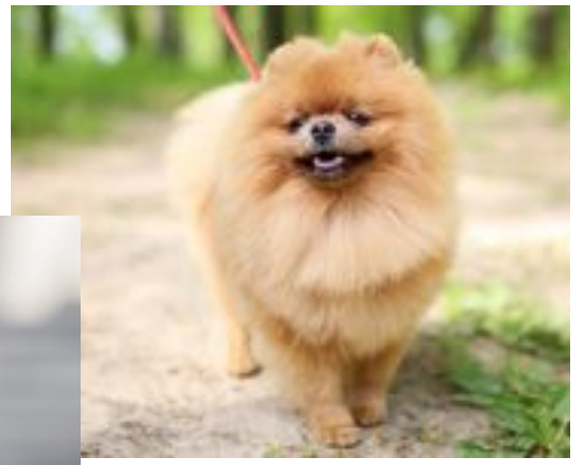
But: Is it only about robustness and interpretability?

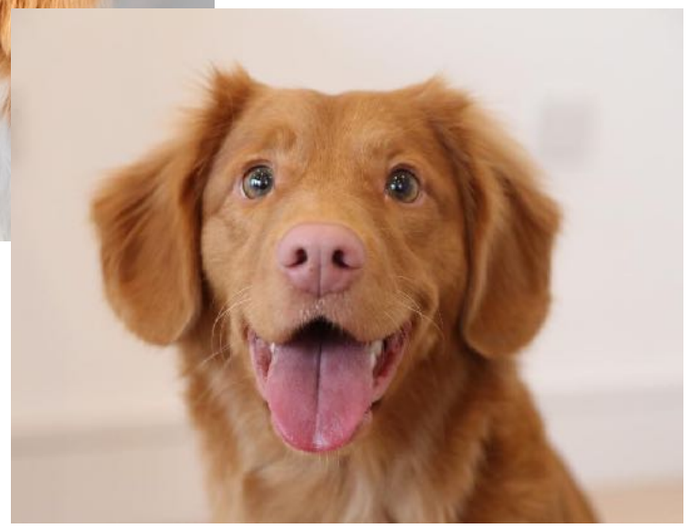
No: It is also about **choosing** what features our models should use

Problem: Correlations can be weird

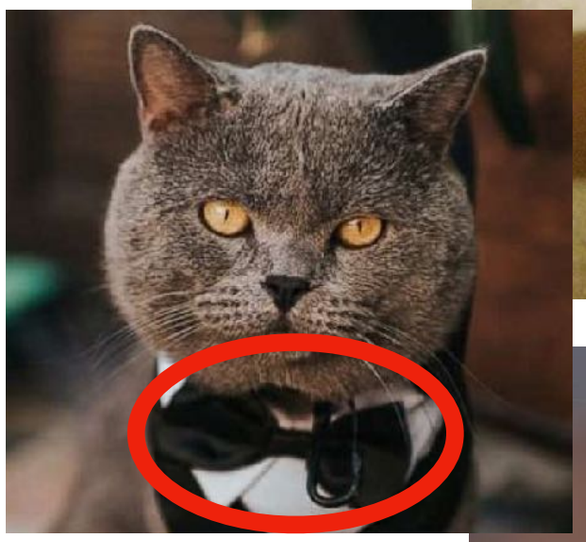
Dogs

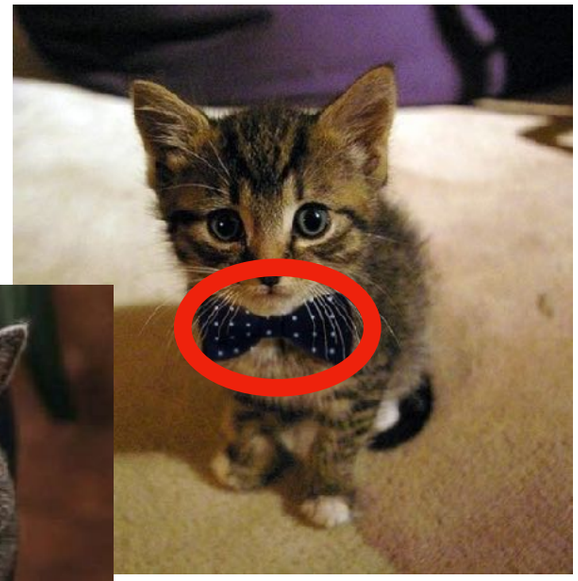


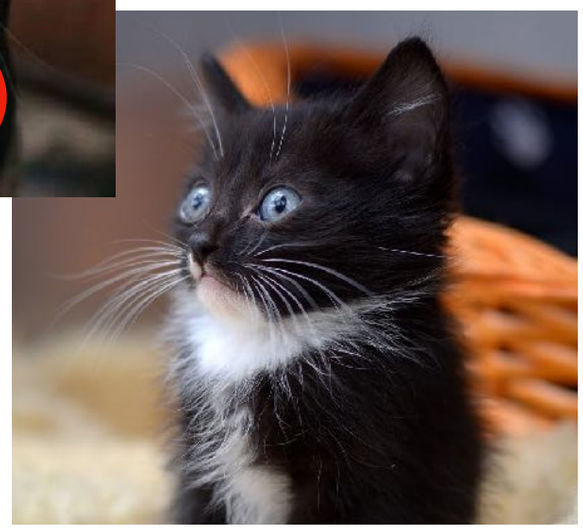




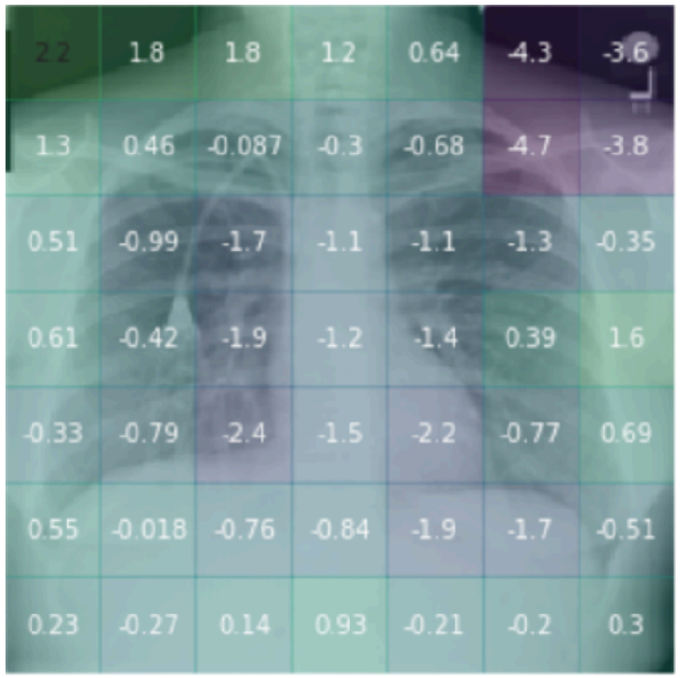
Cats







Problem: Correlations can be weird



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

[Zech et al. 2018]

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

[Esteva et al. 2017]

"Predictive" patterns can be misleading

Useful tool(?): Counterfactual Analysis with Robust Models

Original image

label: “insect”; prediction: “dog”

Robustness = Framework for controlling what correlations to extract

Takeaways

Adversarial examples arise from
non-robust features in the data

→ These features **do** help in generalization (a lot!) and that's why our models like to rely on them

(**Still:** Can also synthesize adv. examples differently **[Nakkiran '19]**)

→ Interpretability needs to be addressed **at training time**

Robustness induces more "human-aligned" representations

→ Enable a broad range of vision applications (in a simple way)

→ Support findings (simple) counterfactuals

But: It is really about **how (and what) our models learn**

→ What is the "right" notion of generalization?

→ What features do we want our models to use?

→ How much do we value human alignment/interpretability?

Adversarial robustness =

Framework for feature engineering

Here: "Adversary" corresponds to a "(human) critic"

gradientscience.org