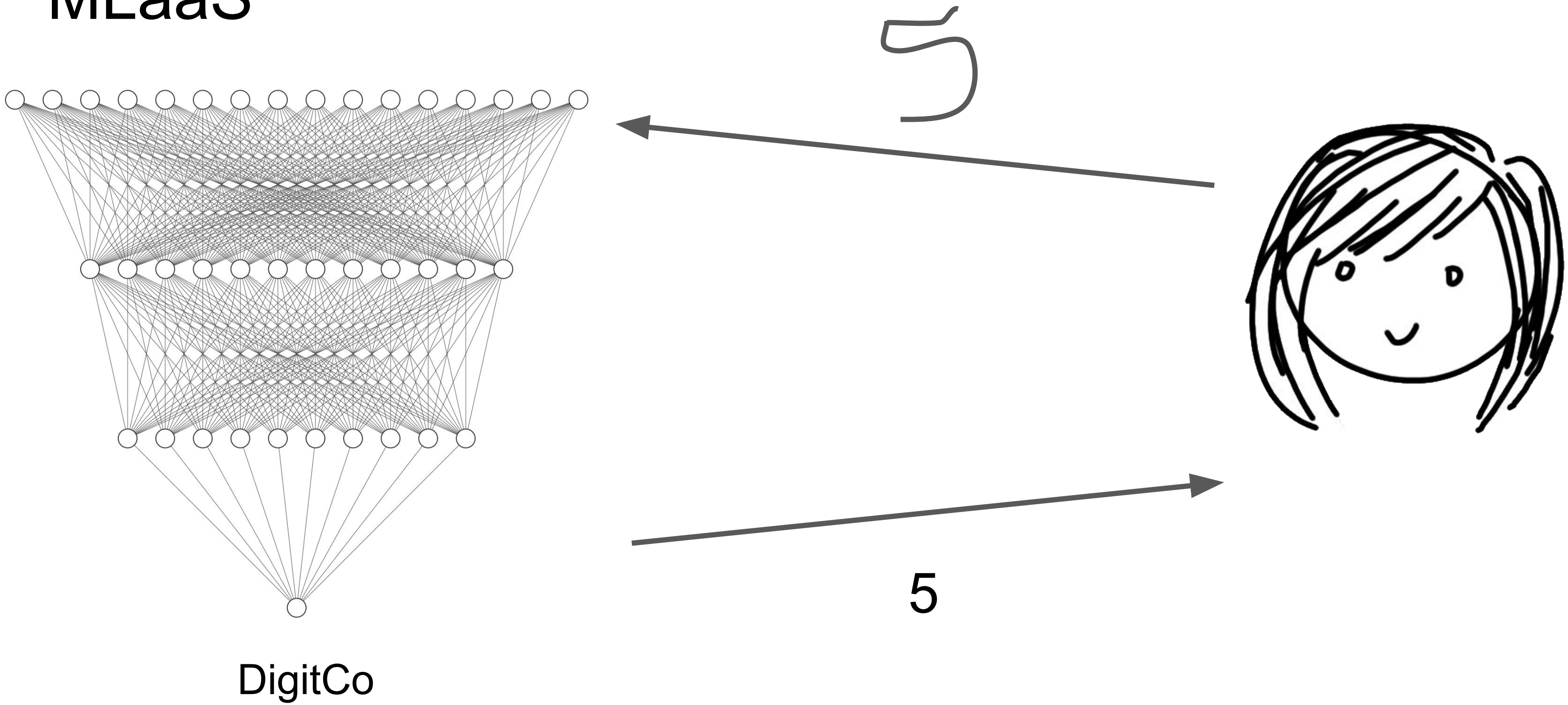
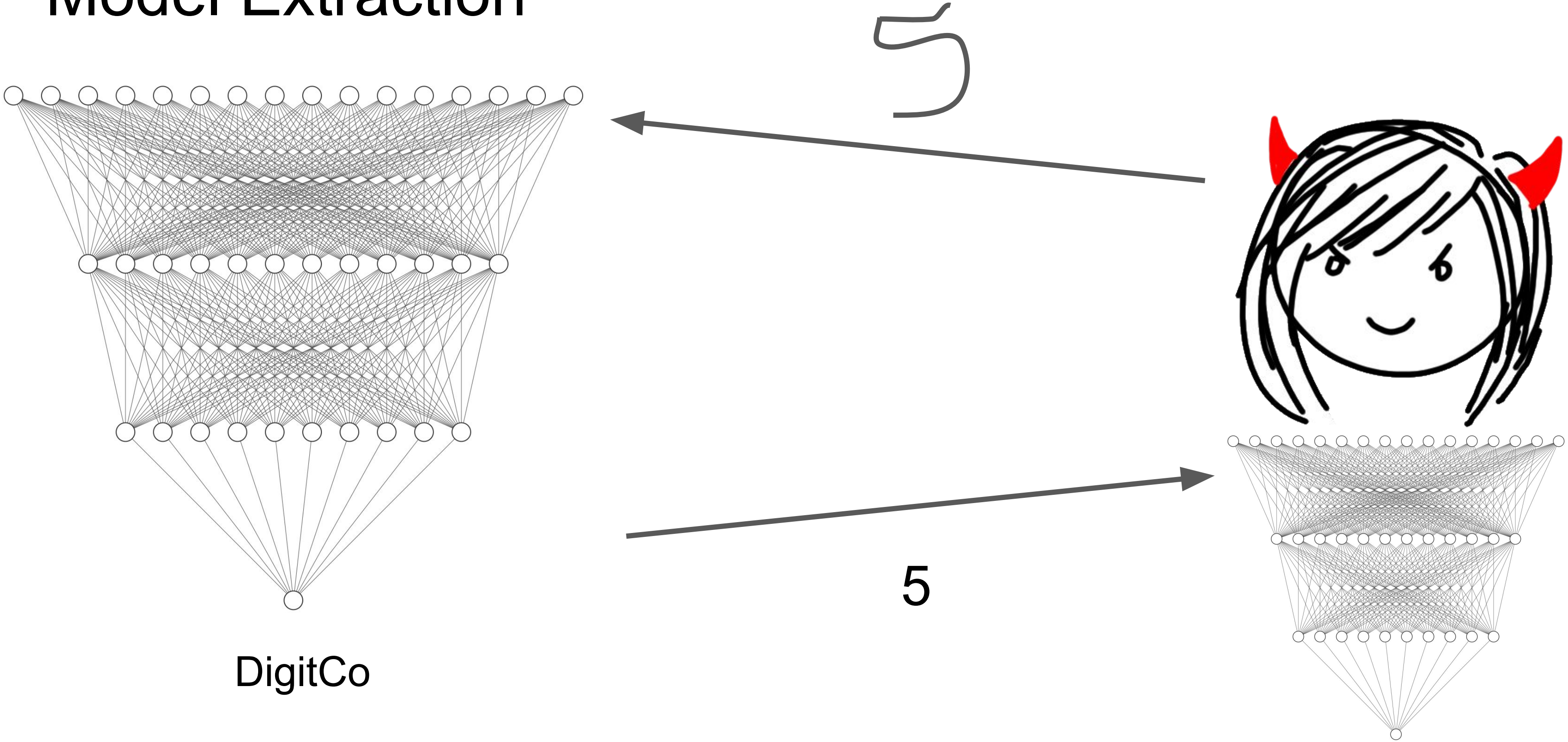
High Accuracy and High Fidelity Extraction of Neural Networks

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot

MLaaS



Model Extraction



This Talk

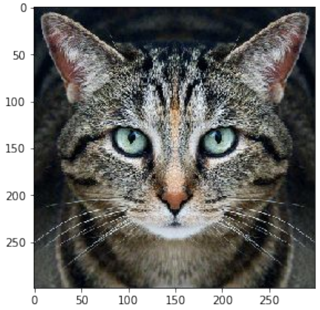
- Taxonomy
  - Motivation
- Learning Extraction
  - Algorithms
  - Limitations
- Direct Recovery Extraction
  - Prior Work
  - Improvements

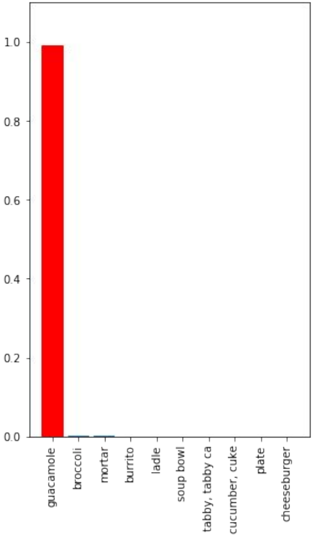
Why would someone do this?

Data and engineers are expensive (theft)...

Why would someone do this?

Data and engineers are expensive (theft)...





...and models are vulnerable to attack (reconnaissance)!

Taxonomy

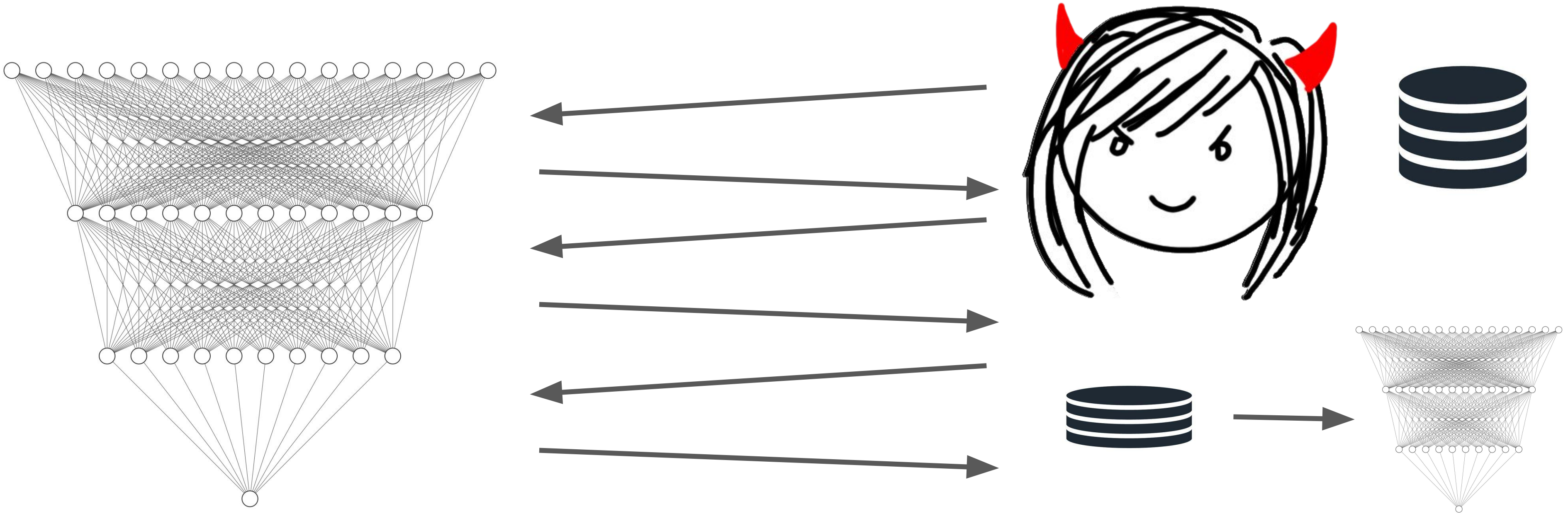
- Theft
  - Accuracy
- Reconnaissance
  - Fidelity
  - Functional Equivalence

- Adversaries also have specific access restrictions
  - Full model output vs class label
  - Rate limiting

Algorithms for Extraction

- Consider a linear model: $f(x) = w.x$
- We could try learning:
  - Do machine learning on $(x_i, f(x_i))$ pairs
- But notice also:
  - $f([1, 0, \ldots, 0]) = w_0$
  - $f([0, 1, \ldots, 0]) = w_1$
- We can directly recover linear models!
- What about neural networks?

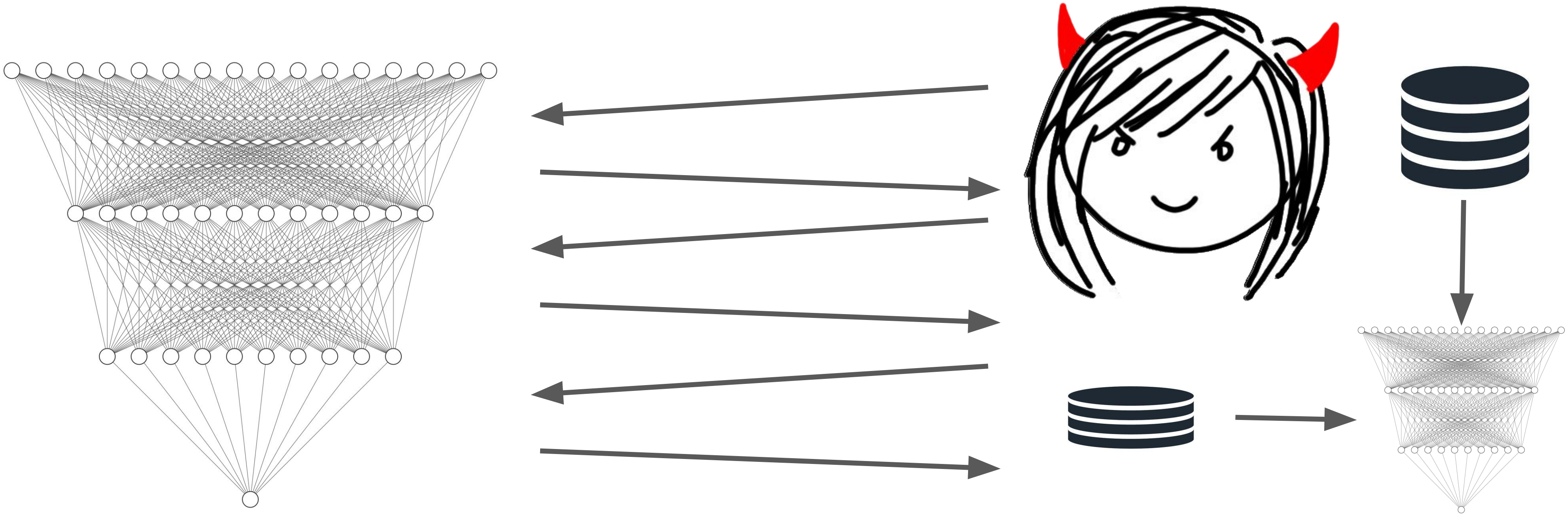
Learning-based Extraction - Active Learning (also here!)



Active Learning: progressively growing a labeled dataset

Chandrasekharan et al: https://arxiv.org/abs/1811.02054

Learning-based Extraction - Semi-Supervised Learning



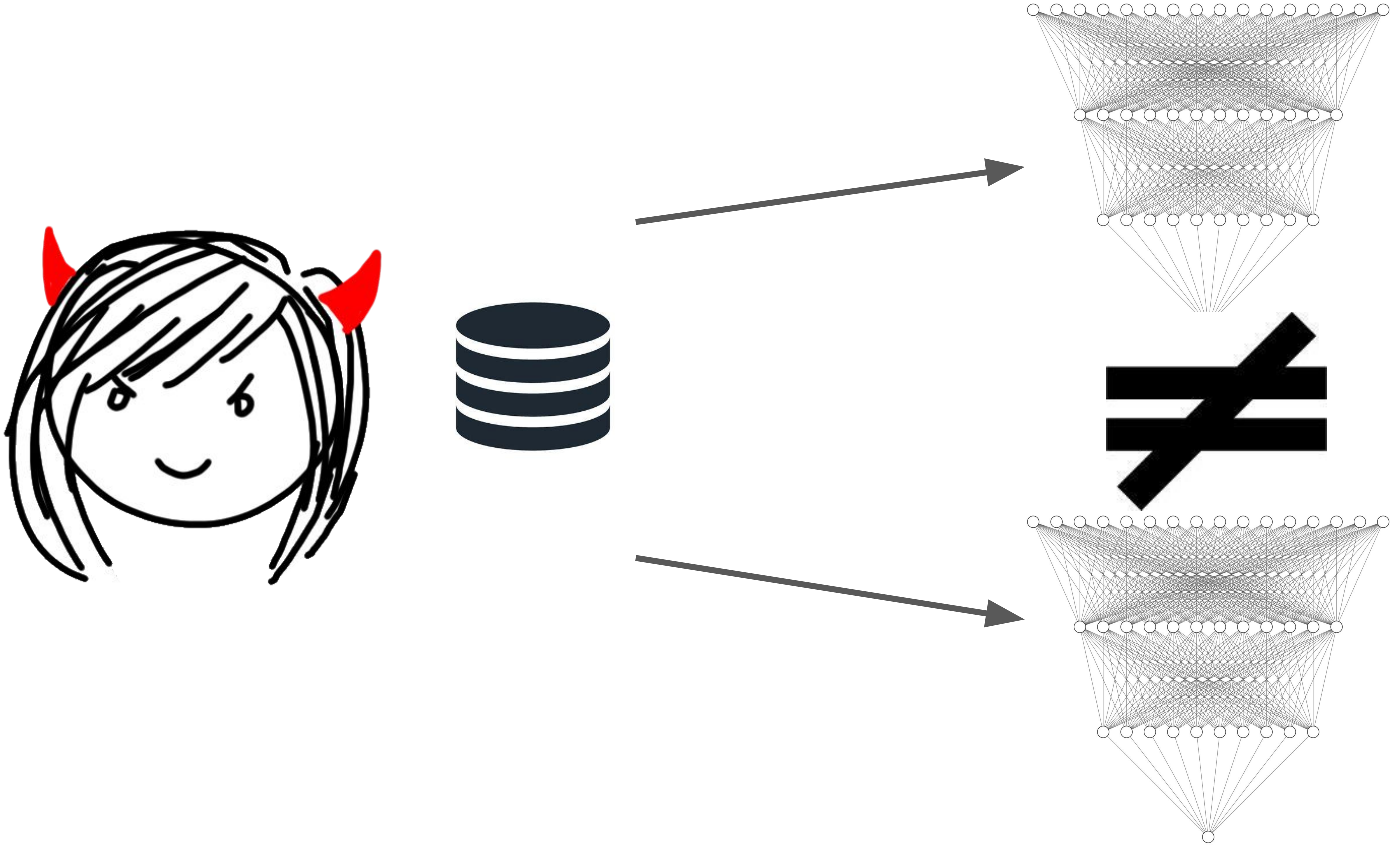
Semi-Supervised Learning: label a small dataset, but also use the unlabeled dataset

Learning-based Extraction

- Semi-supervised learning
  - Scales to deep learning + complex datasets
  - Requires large unlabeled dataset
- Label efficient!

Dataset	Queries	Baseline Accuracy	SemiSup Accuracy
SVHN	250	79.25%	95.82%
CIFAR-10	250	53.35%	87.98%
ImageNet (top 5)	~140000	83.5%	86.17%

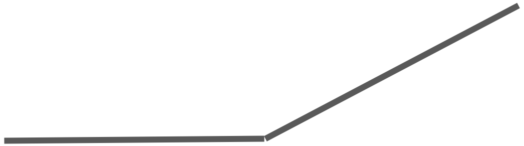
Limitations of Learning-based Extraction - Nondeterminism



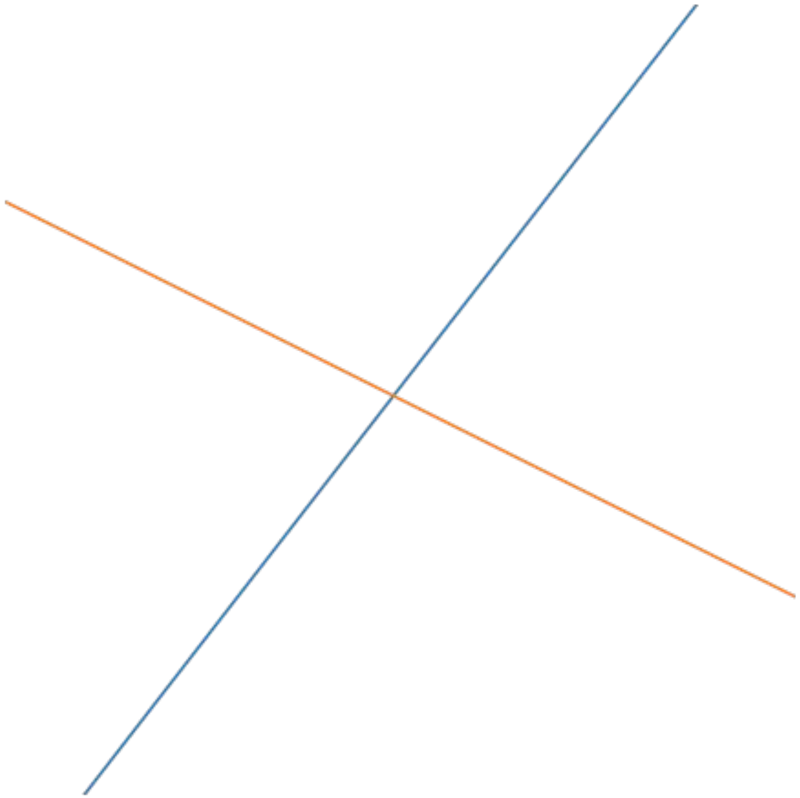
Improving Fidelity - Direct Recovery (Milli et al.)

- Linear model direct recovery isn't easily extended to neural networks

- We focus on 2-layer ReLU networks, following Milli et al. [1]

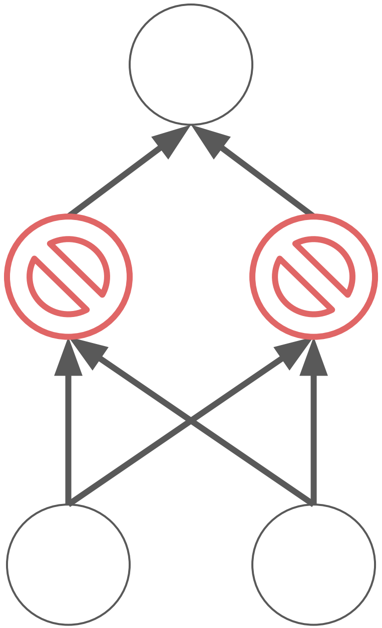


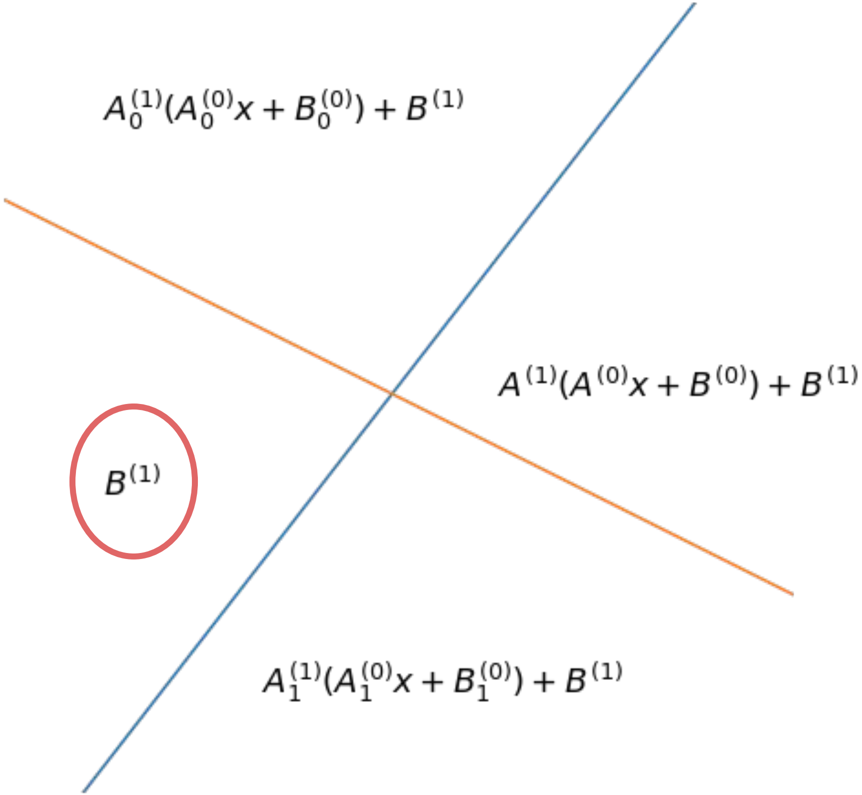
ReLU(x) = max(0, x)



[1] Milli et al: https://arxiv.org/abs/1807.05185

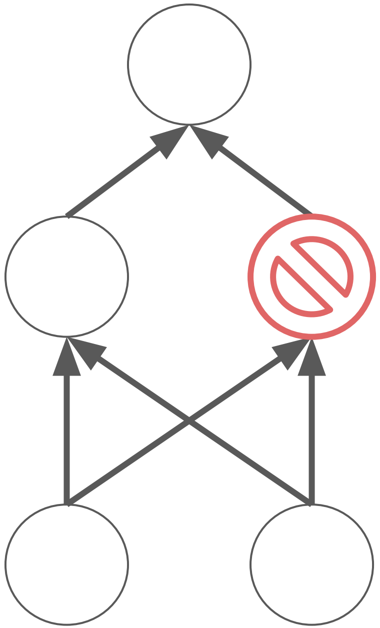
Improving Fidelity - Direct Recovery (Milli et al.)

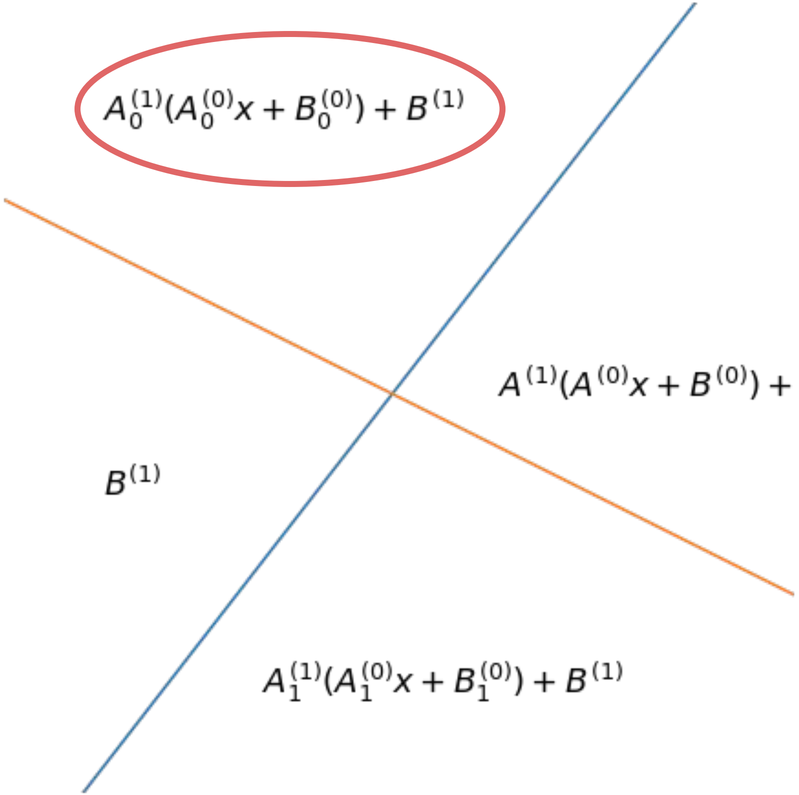




[1] Milli et al: https://arxiv.org/abs/1807.05185

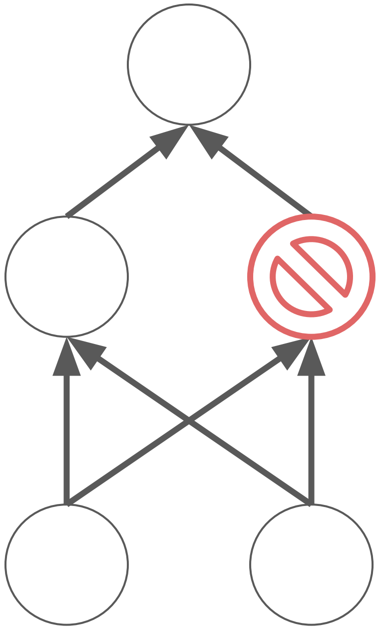
Improving Fidelity - Direct Recovery (Milli et al.)

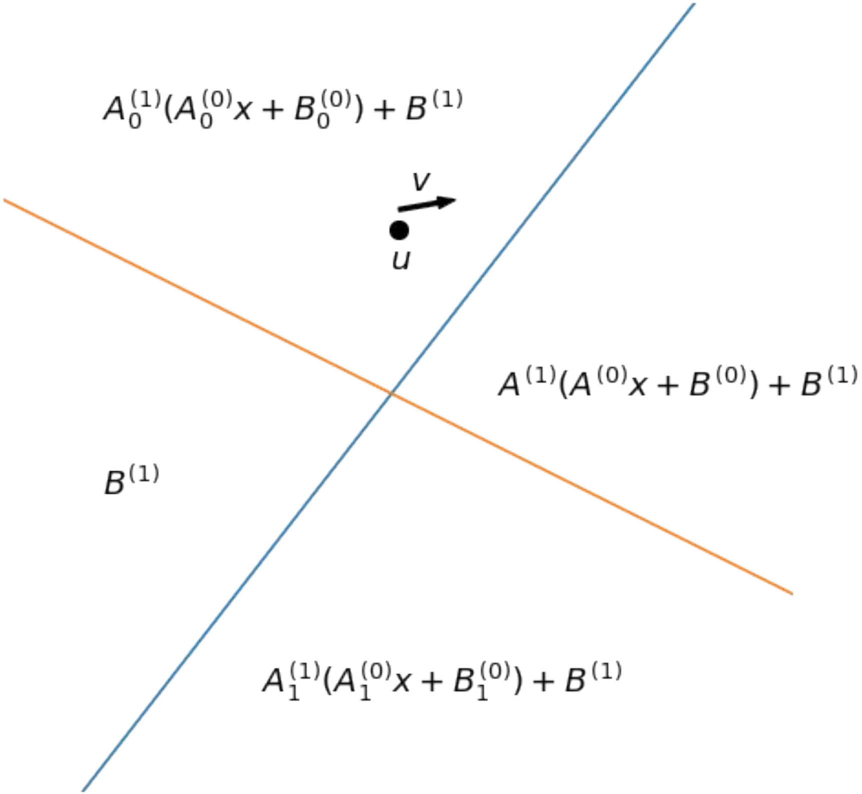




[1] Milli et al: https://arxiv.org/abs/1807.05185

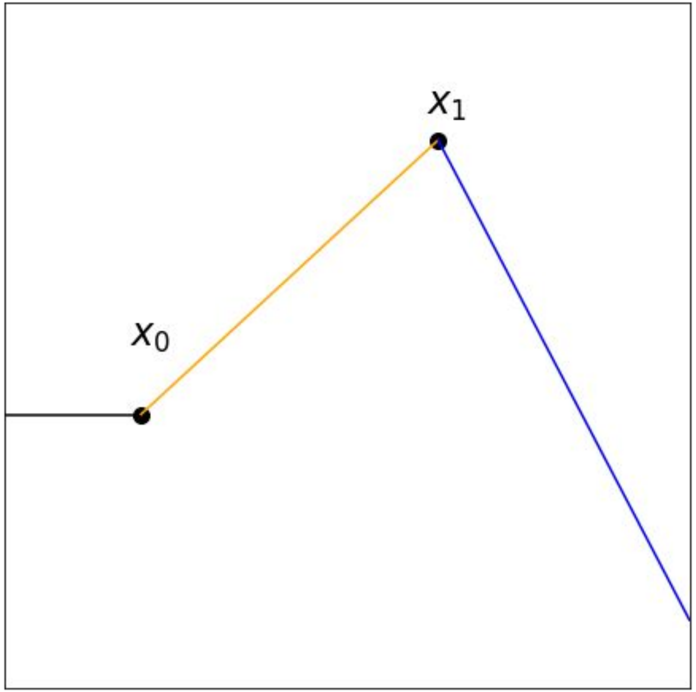
Improving Fidelity - Direct Recovery (Milli et al.)

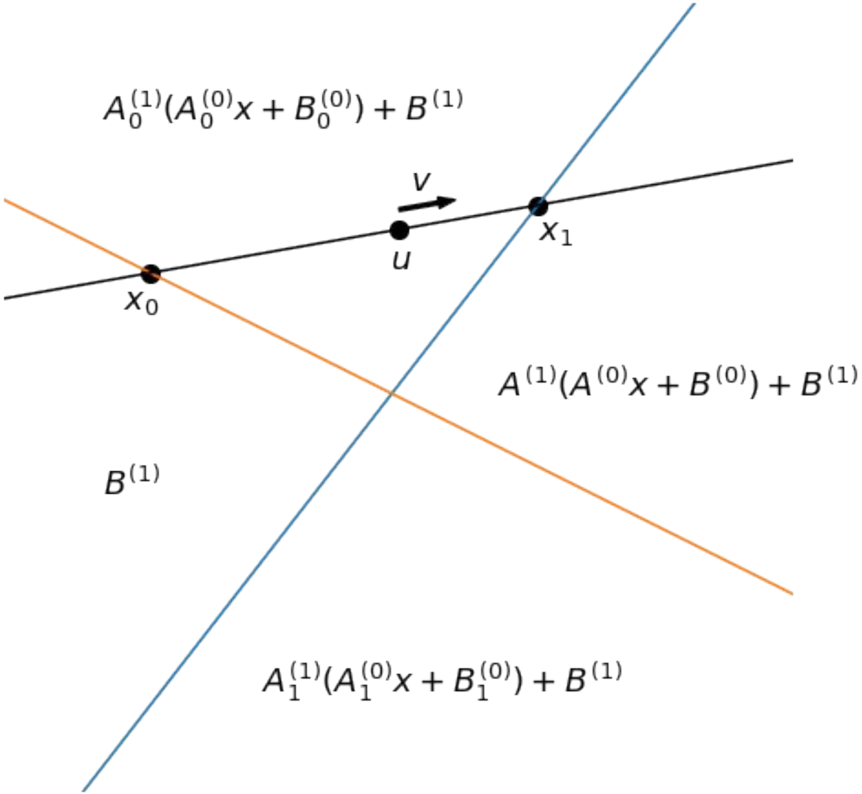




[1] Milli et al: https://arxiv.org/abs/1807.05185

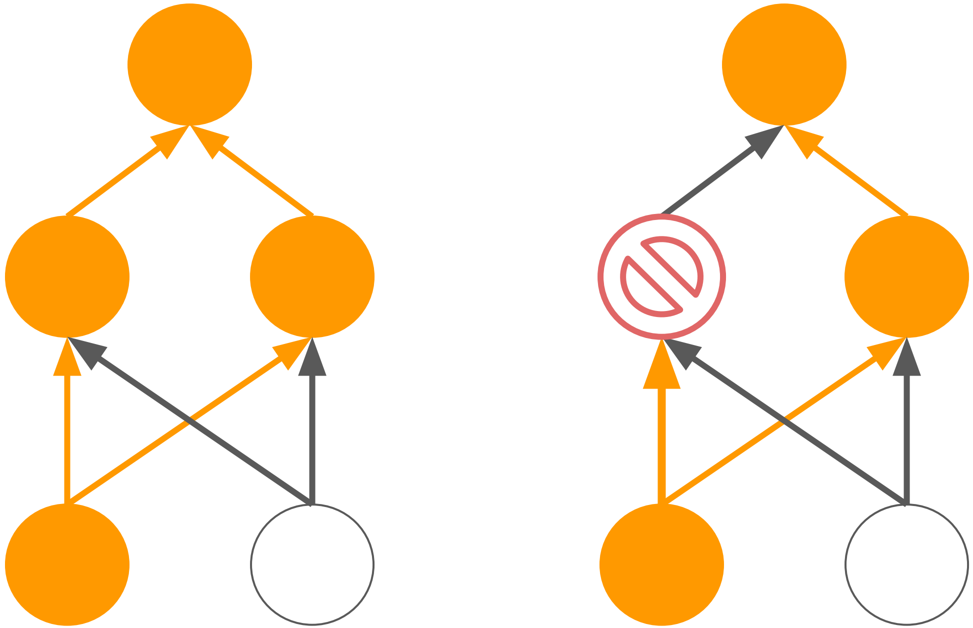
Improving Fidelity - Direct Recovery (Milli et al.)





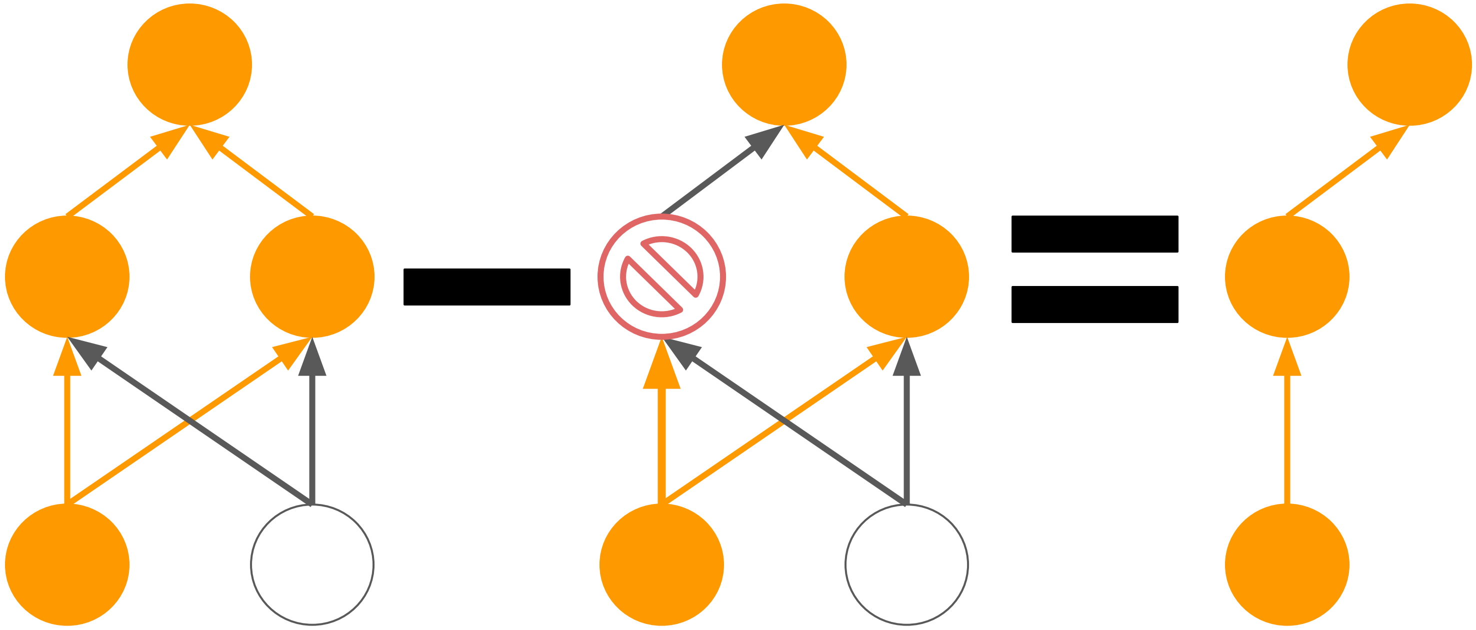
[1] Milli et al: https://arxiv.org/abs/1807.05185

Improving Fidelity - Direct Recovery (Milli et al.)



[1] Milli et al: https://arxiv.org/abs/1807.05185

Improving Fidelity - Direct Recovery (Milli et al.)



[1] Milli et al: https://arxiv.org/abs/1807.05185

Our Functionally Equivalent

- Make Milli et al. work in practice - improved search and precision

Parameters	25,000	50,000	100,000
Fidelity	100%	100%	99.98%
Queries	~150,000	~300,000	~600,000

Effectiveness of our Direct Recovery Attack

Wrapping Up

- See the paper for more!
  - Hardness results
  - Nondeterminism
  - Adversarial example transferability
  - Our functionally equivalent attack
  - Hybrid attacks
- Future Work
  - More efficient, realistic, effective attacks!
  - Defenses for accuracy, fidelity, functionally equivalent?
- Thank you! Ask me questions!

Credits

- Papers
  - Chandrasekharan et al.: https://arxiv.org/abs/1811.02054
  - Milli et al.: https://arxiv.org/abs/1807.05185
- Images
  - Alice: Eysa Lee https://ccs.neu.edu/~eysa/
  - Neural Network Diagram: http://alexlenail.me/NN-SVG/index.html
  - Affiliations: https://ai.googleblog.com/, https://en.wikipedia.org/wiki/Northeastern_University, https://en.wikipedia.org/wiki/University_of_Toronto
  - Slide 6: https://hackernoon.com/hn-images/1*be2sR_HIKjY36cWuWRcu-Q.jpeg
  - Slide 6: https://imgs.xkcd.com/comics/machine_learning_2x.png
  - Slide 6: https://www.labsix.org/media/2017/10/31/cat_adversarial.png