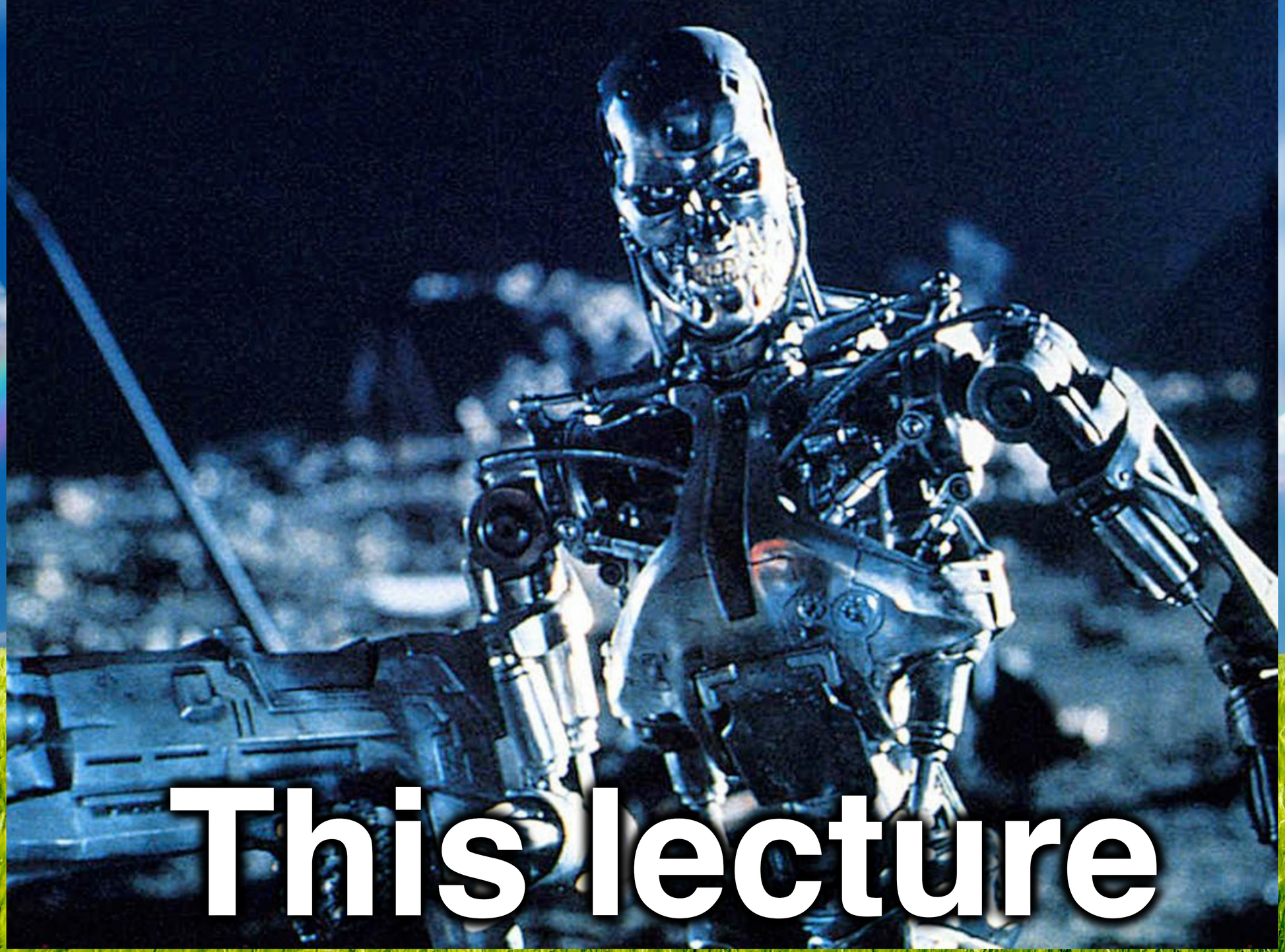
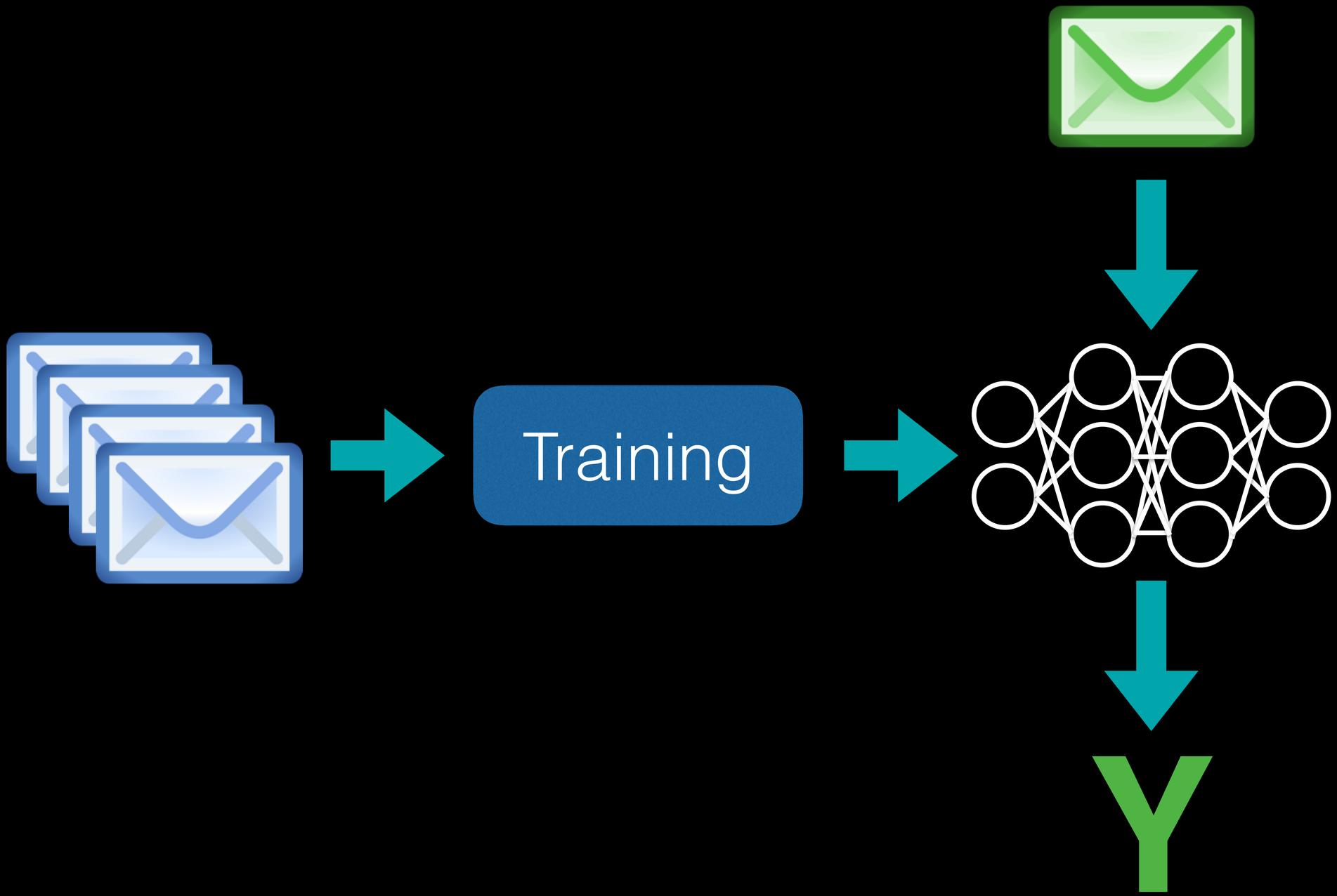
Adversarial Machine Learning

Nicholas Carlini
Google


This class (so far)

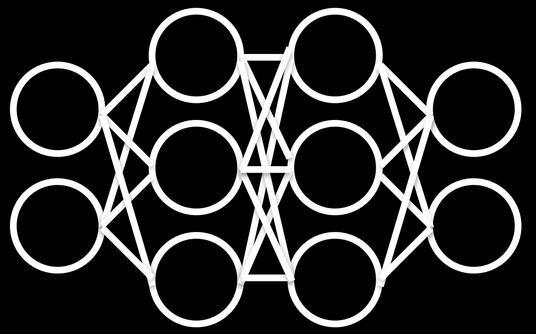

This lecture


Training
Y

Evasion:
Modify test inputs
to cause test errors

Poisoning:
Modify training data
to cause test errors

Training



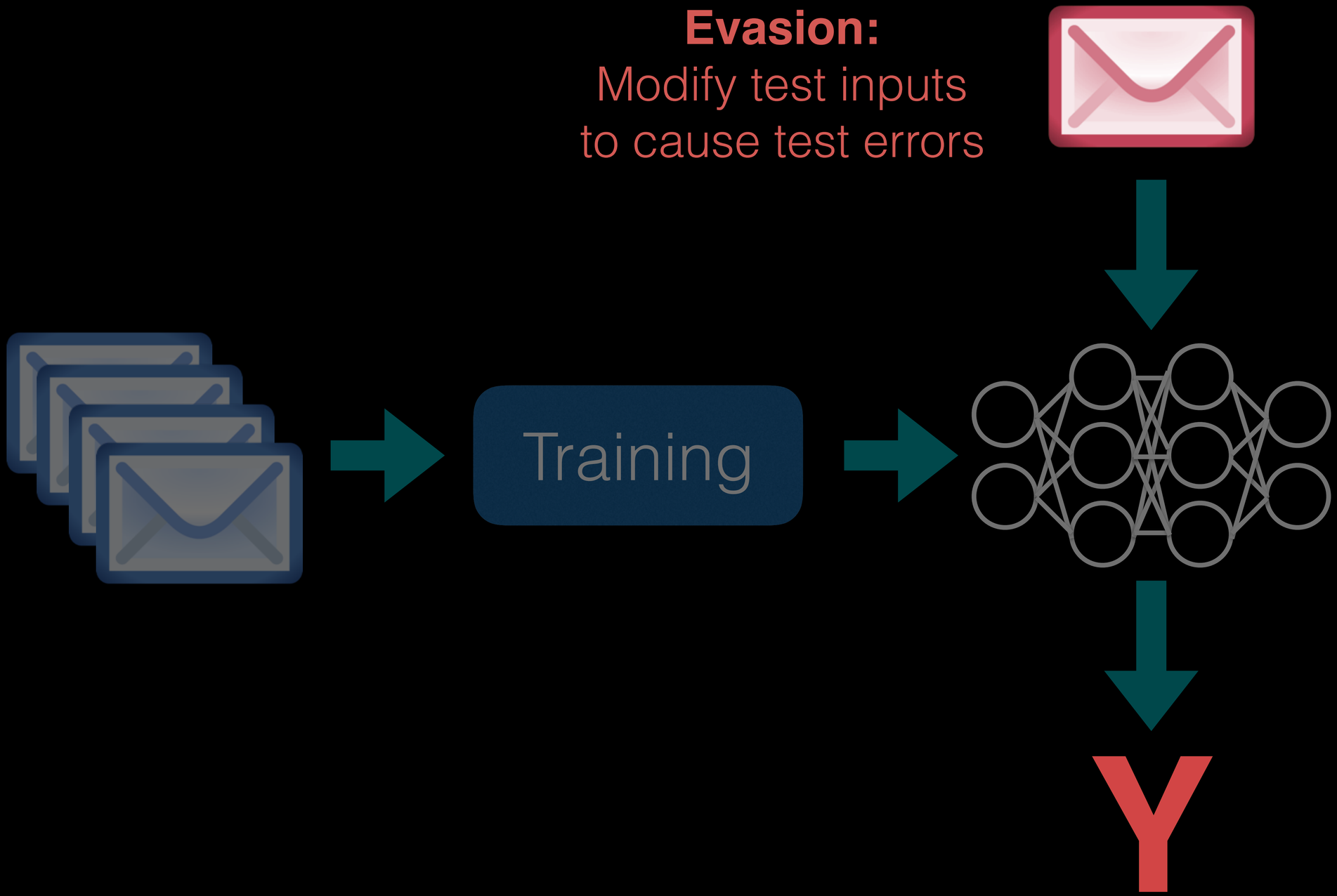
Training Data Extraction:
Study model parameters
to reveal
training data

Model Stealing:
Study model output to
reveal parameters

Y

Act I:
Evasion

Evasion:
Modify test inputs
to cause test errors



88% tabby cat

adversarial
perturbation

88% **tabby cat**

adversarial
perturbation

88% **tabby cat**

adversarial
perturbation

88% **tabby cat**

99% **guacamole**

Okay, lesson learned.

Okay, lesson learned.

Don't classify cats with neural networks.

adversarial
perturbation

guacamole

tabby cat

Okay, lesson learned.

Don't classify cats with neural networks.

Okay, lesson learned.

or guacamole

Don't classify cats ^ with neural networks.

(a)

(b)

(c)

Okay, lesson learned.

Don't classify ~~cats~~ images with neural networks.

What will a state-of-the-art neural network transcribe?

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity"

Okay, lesson learned.

Don't classify images or audio with neural networks.

Generating Natural Language Adversarial Examples

Moustafa Alzantot[1]*, Yash Sharma[2]*, Ahmed Elgohary[3],
Bo-Jhang Ho[1], Mani B. Srivastava[1], Kai-Wei Chang[1]

[1]Department of Computer Science, University of California, Los Angeles (UCLA)
{malzantot, bojhang, mbs, kwchang}@ucla.edu
[2]Cooper Union sharma2@cooper.edu
[3]Computer Science Department, University of Maryland elgohary@cs.umd.edu

Adversarial Attacks on Neural Network Policies

Sandy Huang[†], **Nicolas Papernot**[‡], **Ian Goodfellow**[§], **Yan Duan**[†§], **Pieter Abbeel**[†§]
[†] University of California, Berkeley, Department of Electrical Engineering and Computer Sciences
[‡] Pennsylvania State University, School of Electrical Engineering and Computer Science
[§] OpenAI

Abstract

Machine learning classifiers are known to be vulnerable to inputs maliciously constructed by adversaries to force misclassification. Such adversarial examples have been extensively studied in the context of computer vision applications. In this work, we show adversarial attacks are also effective when targeting neural network policies in reinforcement learning. Specifically, we show existing adversarial example crafting techniques can be used to significantly degrade test-time performance

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Minhao Cheng[1], Jinfeng Yi[2], Huan Zhang[1], Pin-Yu Chen[3], Cho-Jui Hsieh[1]

[1]Department of Computer Science, University of California, Davis, CA 95616
[2]Tencent AI Lab, Bellevue, WA 98004
[3]IBM Research AI, Yorktown Heights, NY 10598
mhcheng@ucdavis.edu, jinfengyi.ustc@gmail.com, ecezhang@ucdavis.edu,
pin-yu.chen@ibm.com, chohsieh@ucdavis.edu

HALLUCINATIONS IN NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural machine translation (NMT) systems have reached state of the art performance in translating text and are in wide deployment. Yet little is understood about how these systems function or break. Here we show that NMT systems are susceptible to producing highly pathological translations that are completely untethered from the source material, which we term *hallucinations*. Such pathological translations are problematic because they are are deeply disturbing of user trust and easy to find with a simple search. We describe a method to generate hallucinations and show that many common variations of the NMT architecture
of halluci
niques, sh
nally, we
in the atte

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION

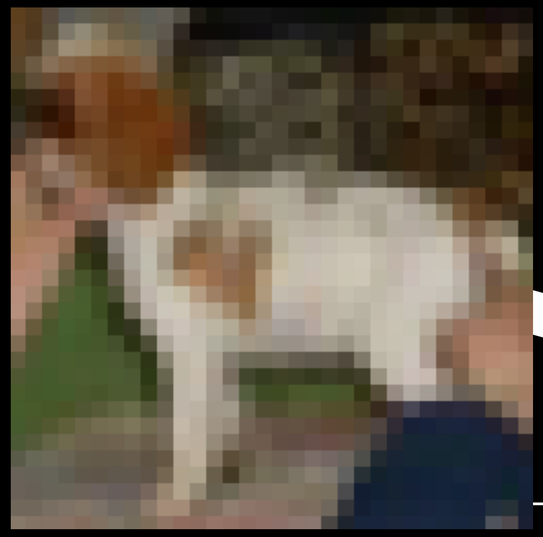
Yonatan Belinkov*
Computer Science and
Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
belinkov@mit.edu

Yonatan Bisk*
Paul G. Allen School
of Computer Science & Engineering,
University of Washington
ybisk@cs.washington.edu

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab Ondrej Miksik Philip H.S. Torr
University of Oxford
{anurag.arnab, ondrej.miksik, philip.torr}@eng.ox.ac.uk

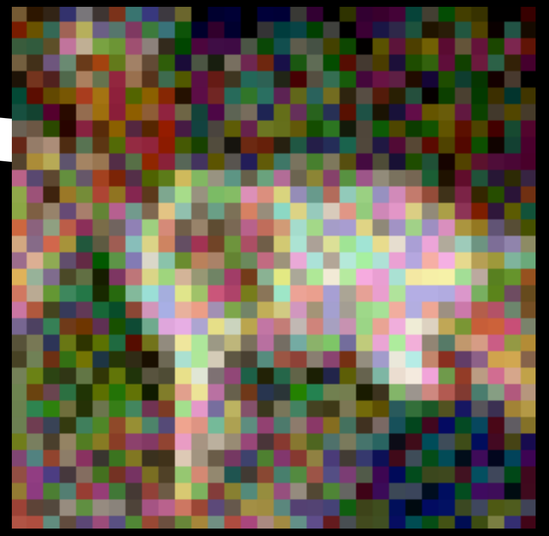
How do these attacks work?

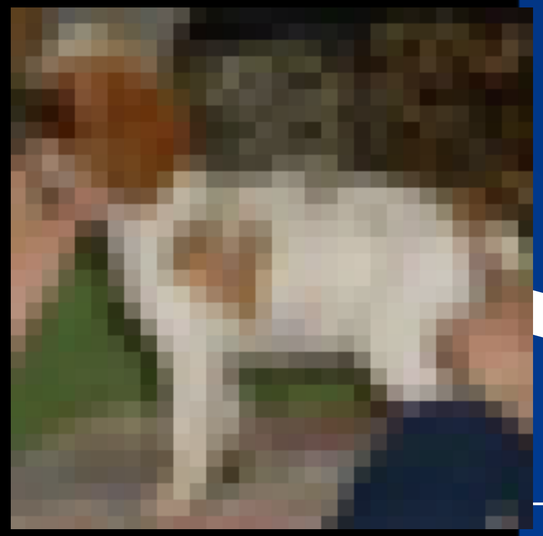


Dog

Random Direction

Random Direction

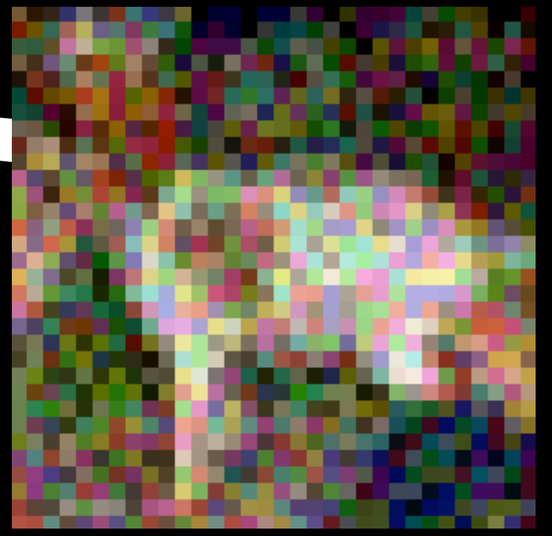




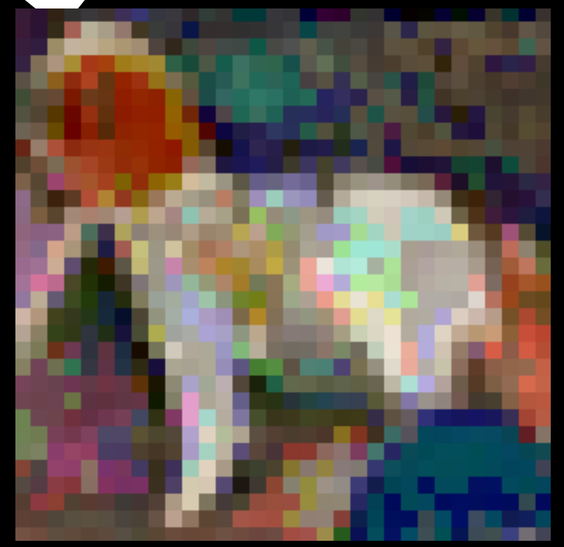
Dog

Random Direction

Random Direction



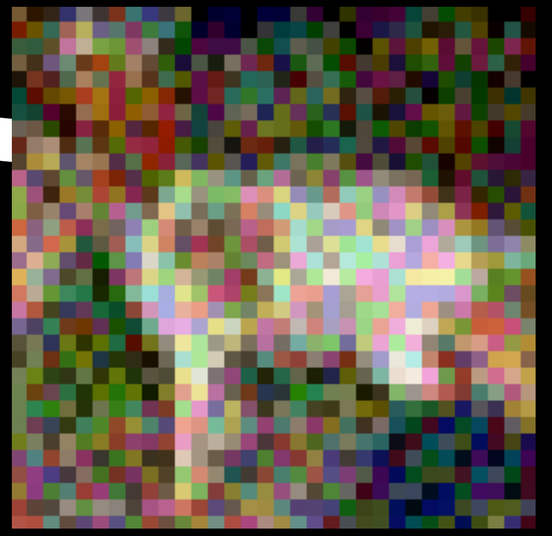
Truck



Dog

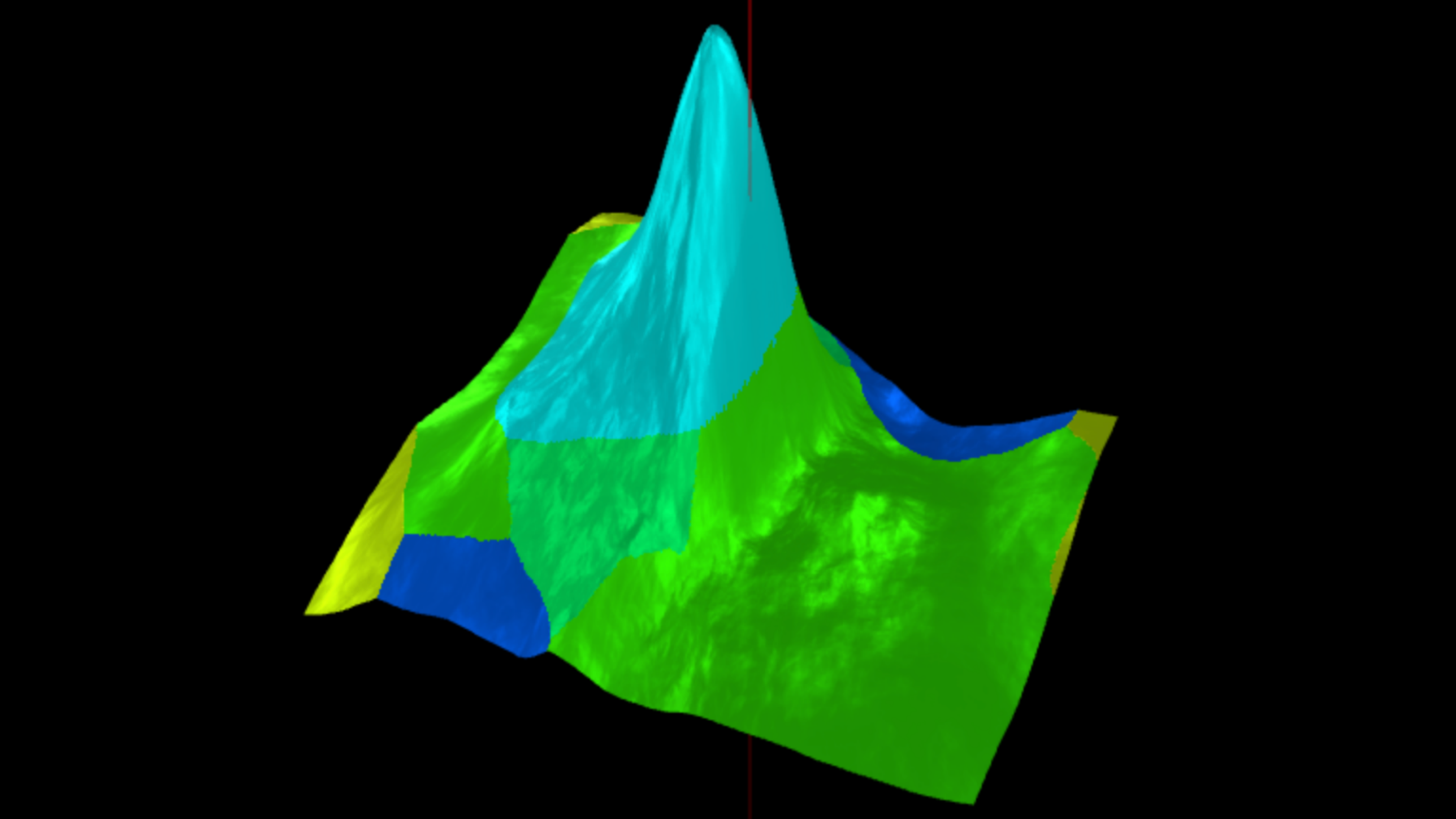
Random Direction

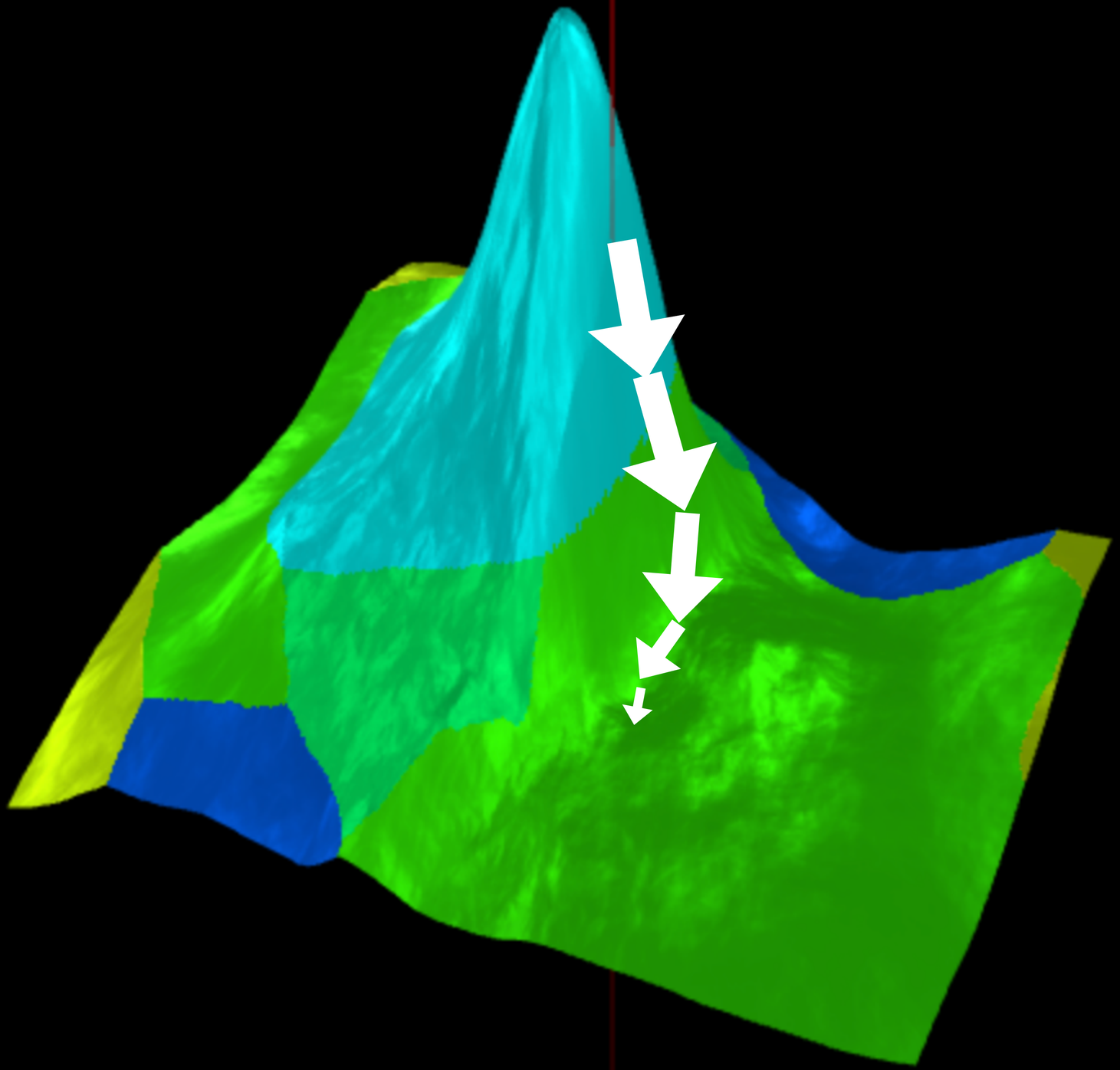
Adversarial Direction



Truck

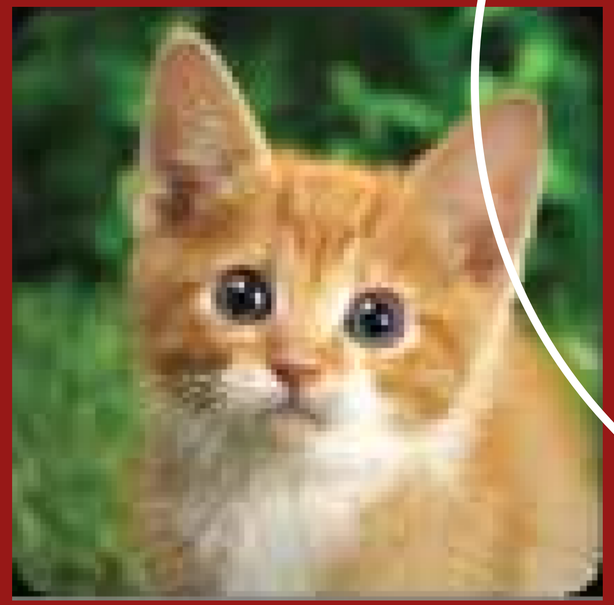
Airplane

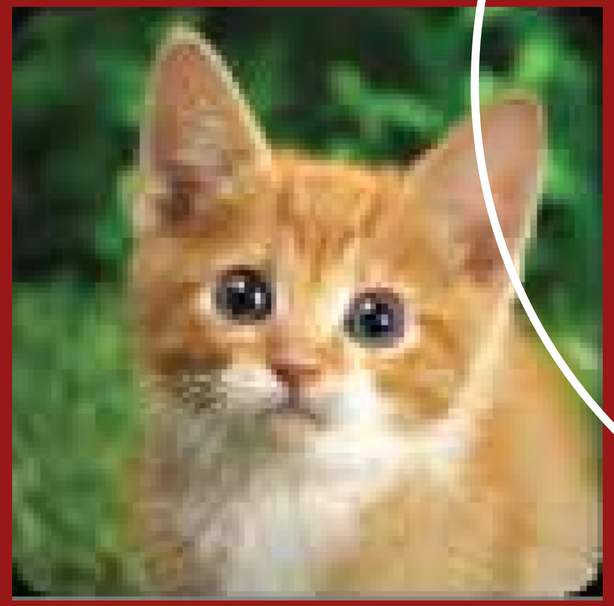


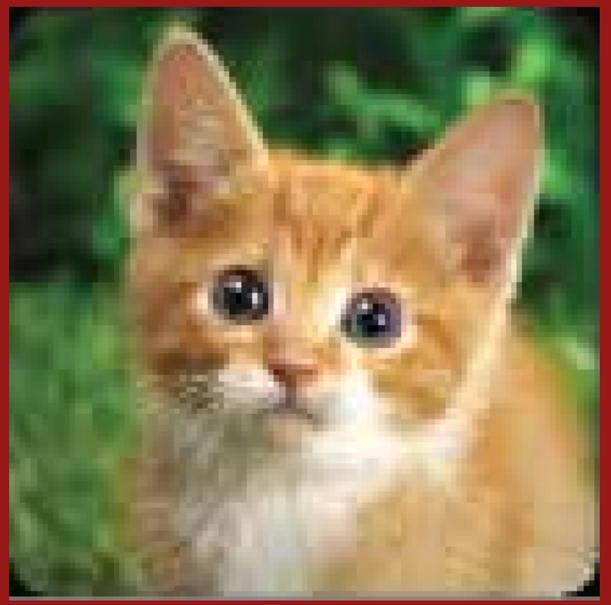


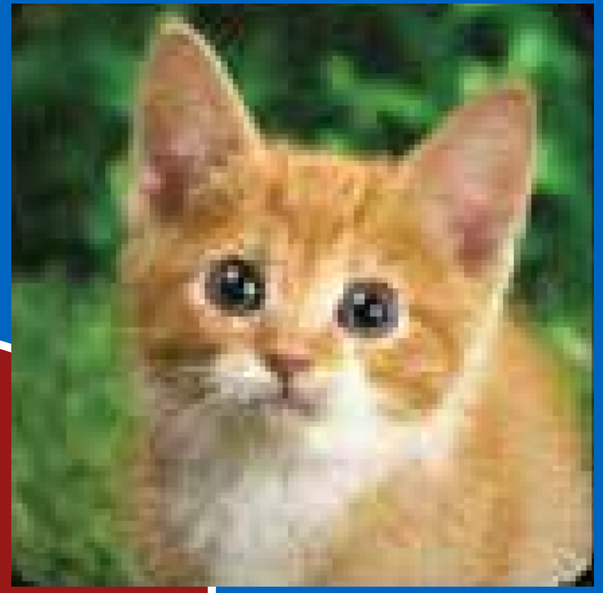
That's what we can do
with access to the weights

What can we do
without the weights?





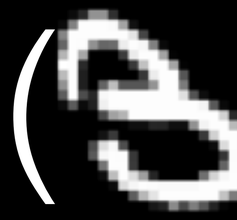




How do we defend against attacks?

Normal Training


,7)

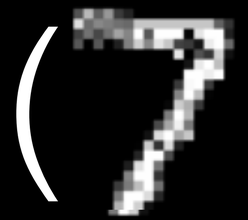

,3)

Training

Adversarial Training (1)


, 7)


, 3)


, 7)

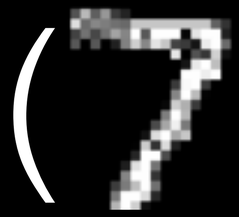

, 3)

Attack

Adversarial Training (2)


, 7)

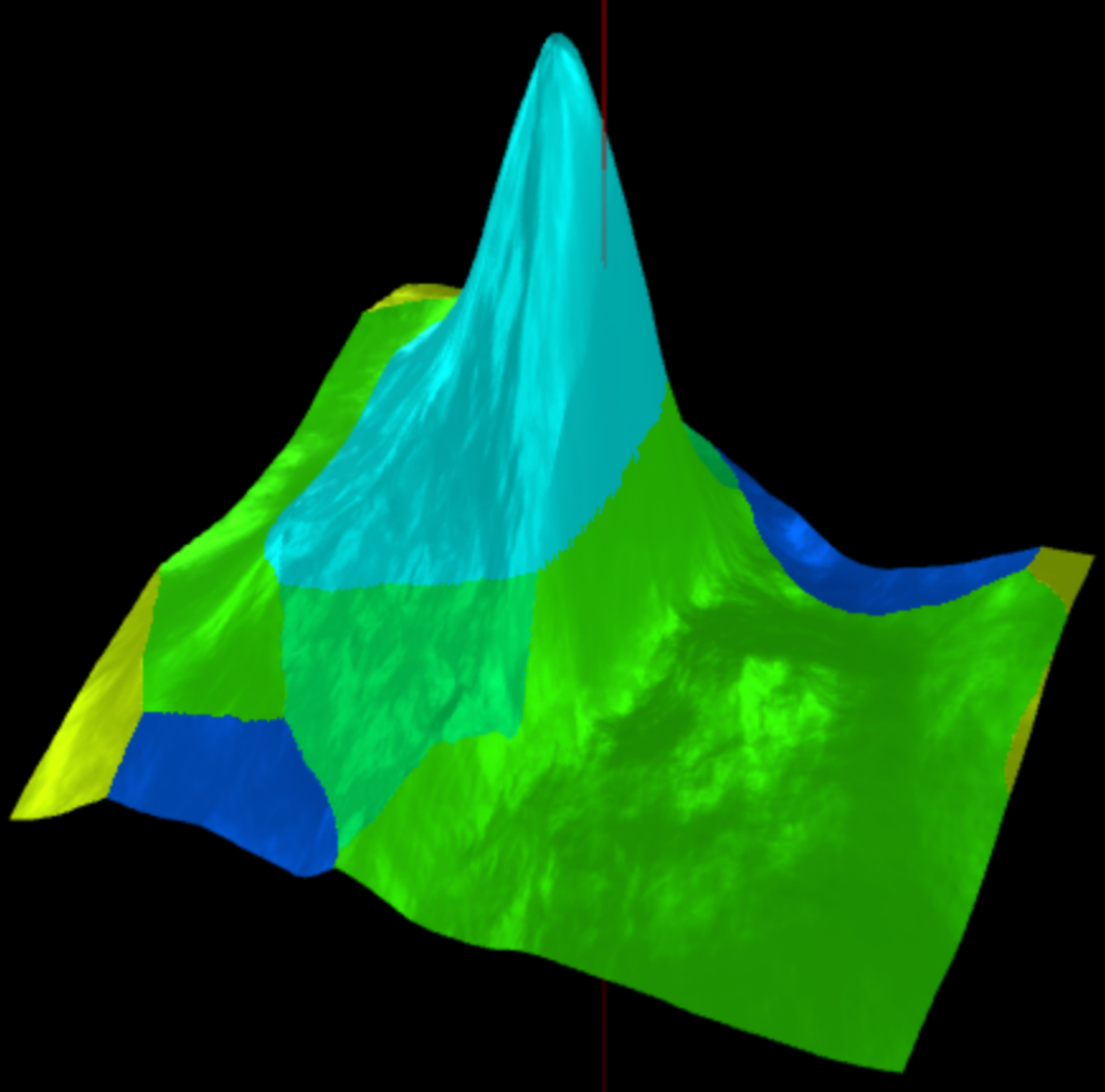

, 3)


, 7)

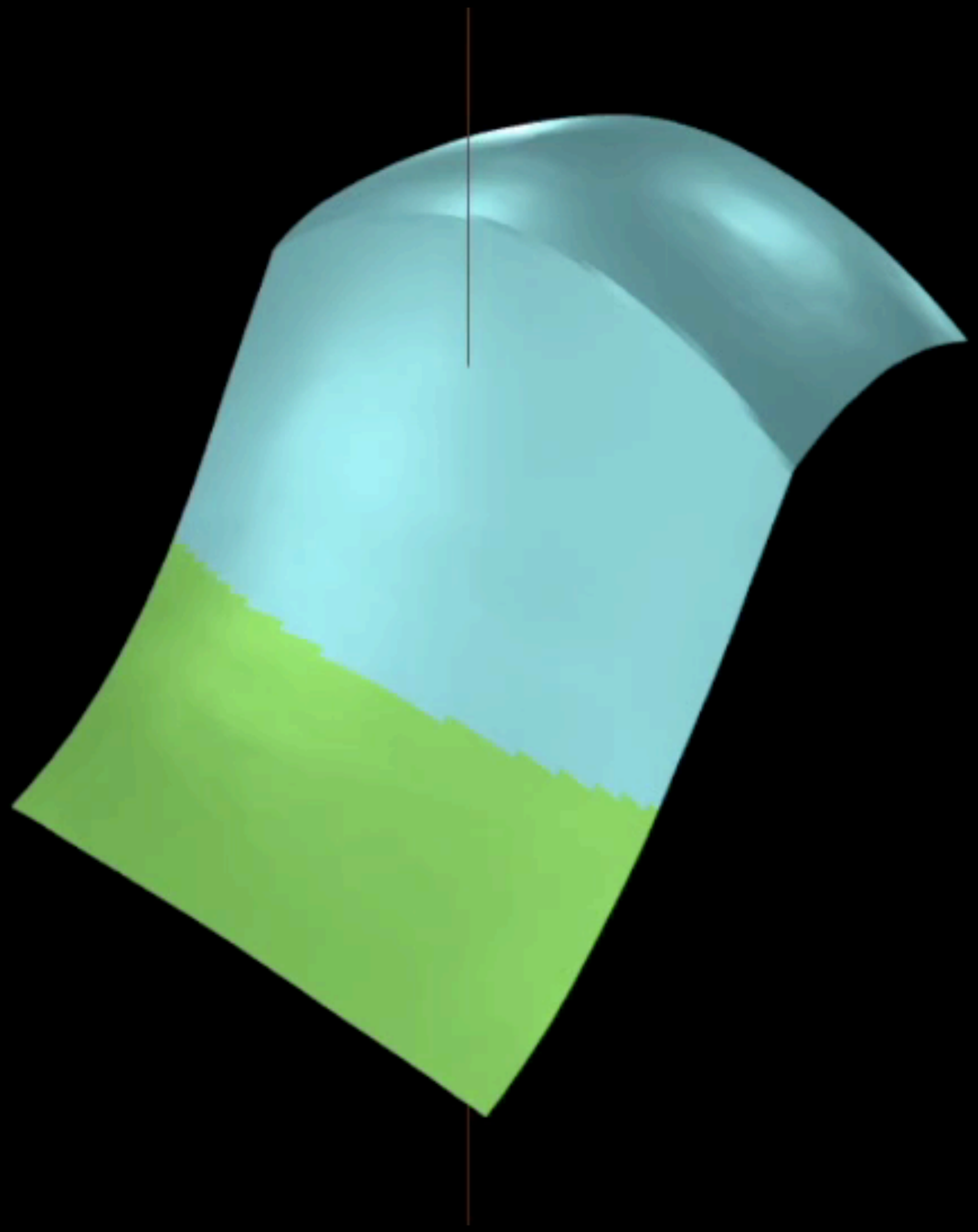

, 3)

Training

... and that's almost it.

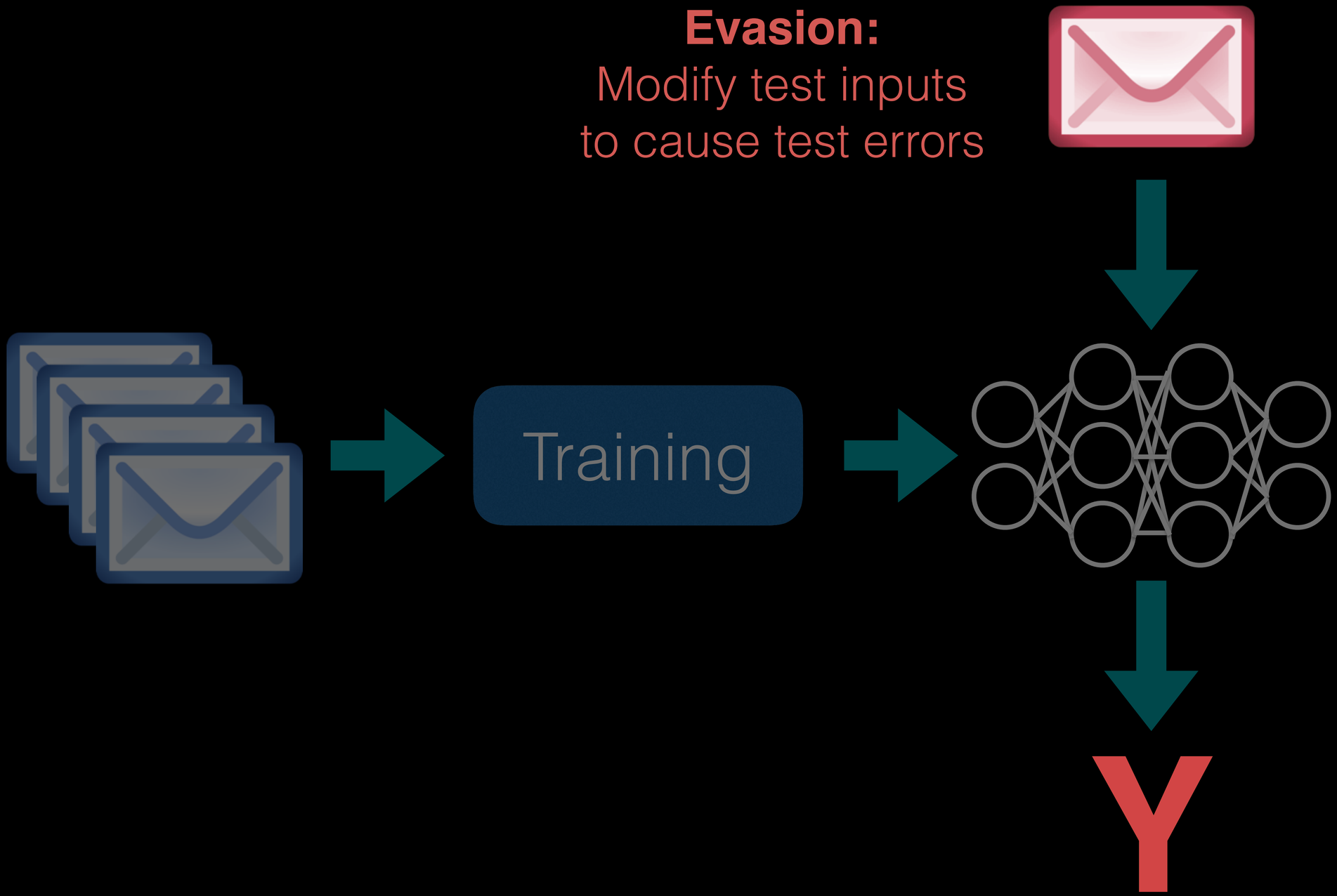


Normal Loss Surface



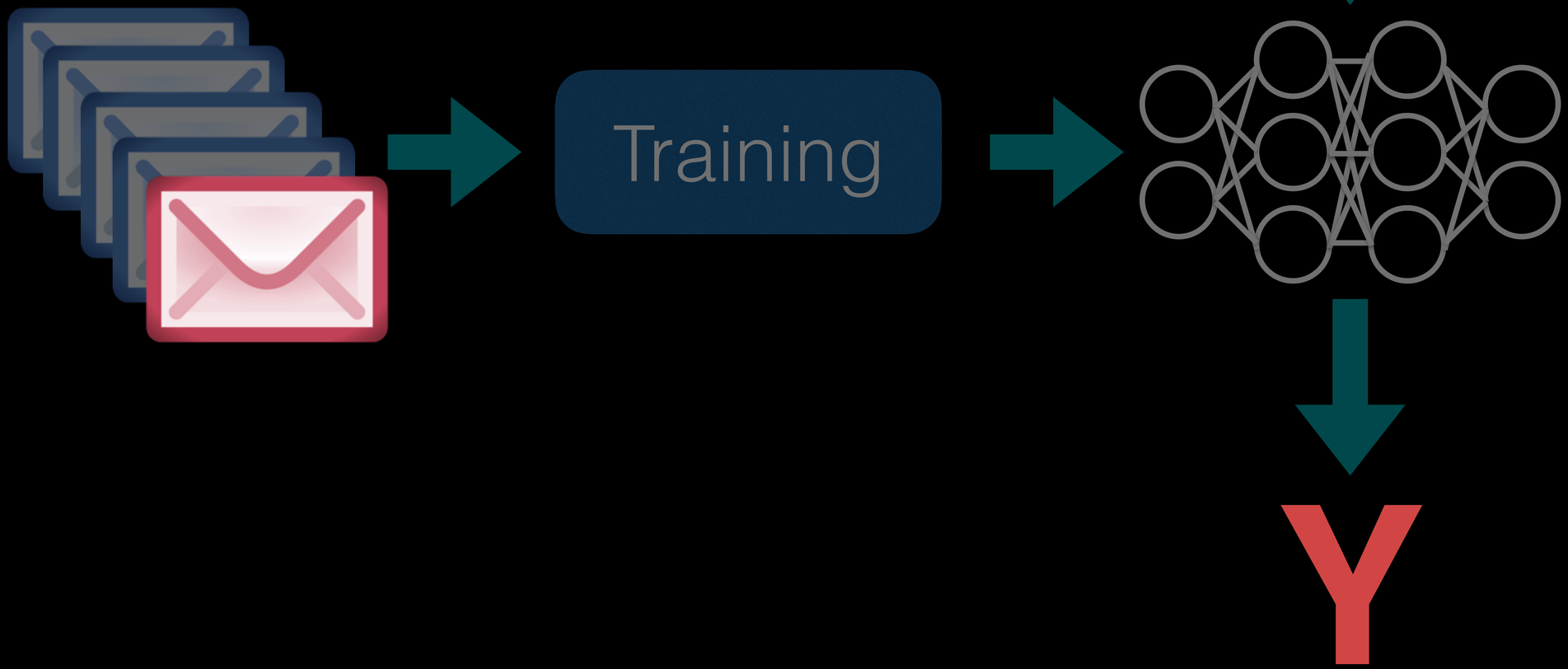
Adversarial Training Loss Surface

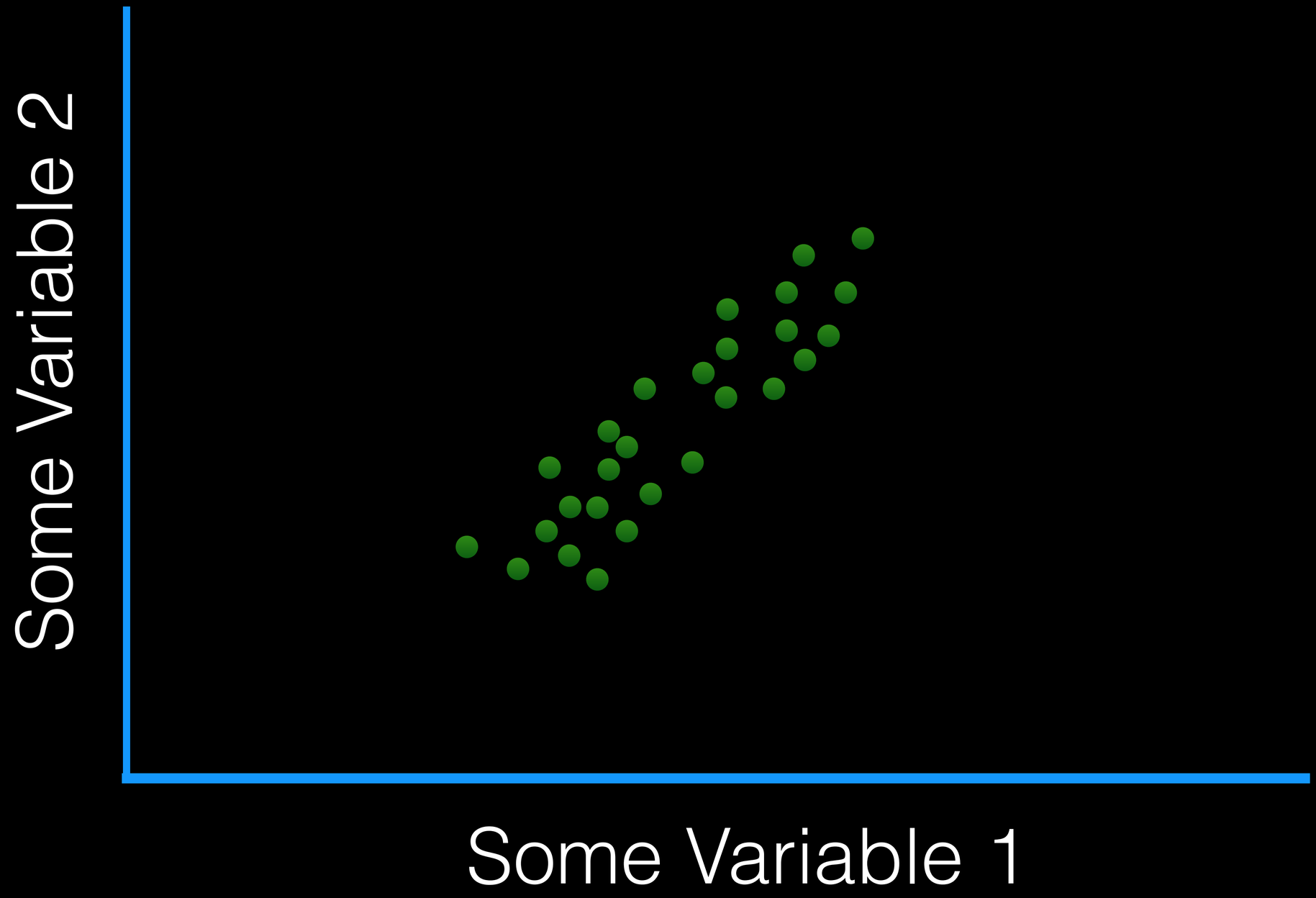
Evasion:
Modify test inputs
to cause test errors

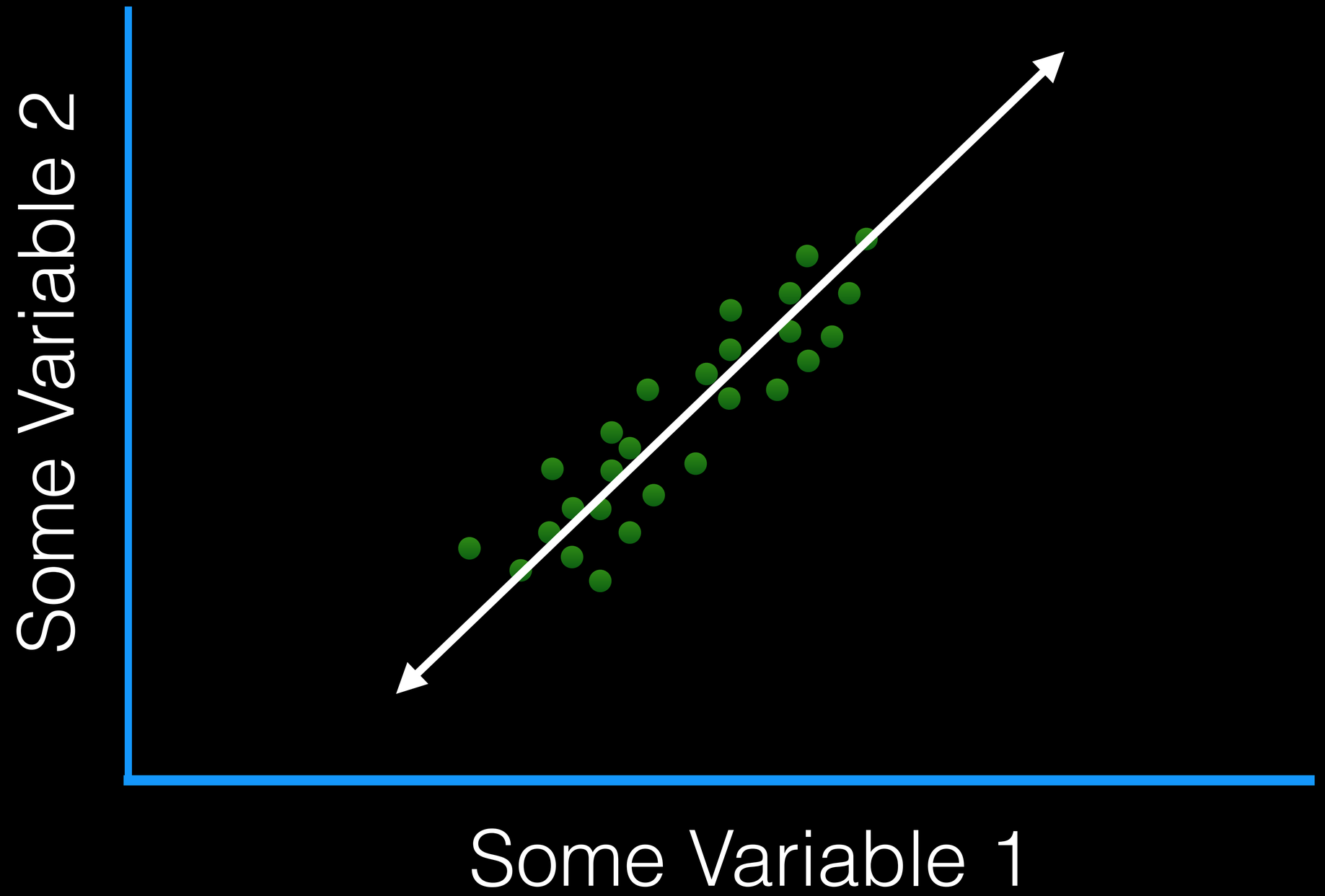


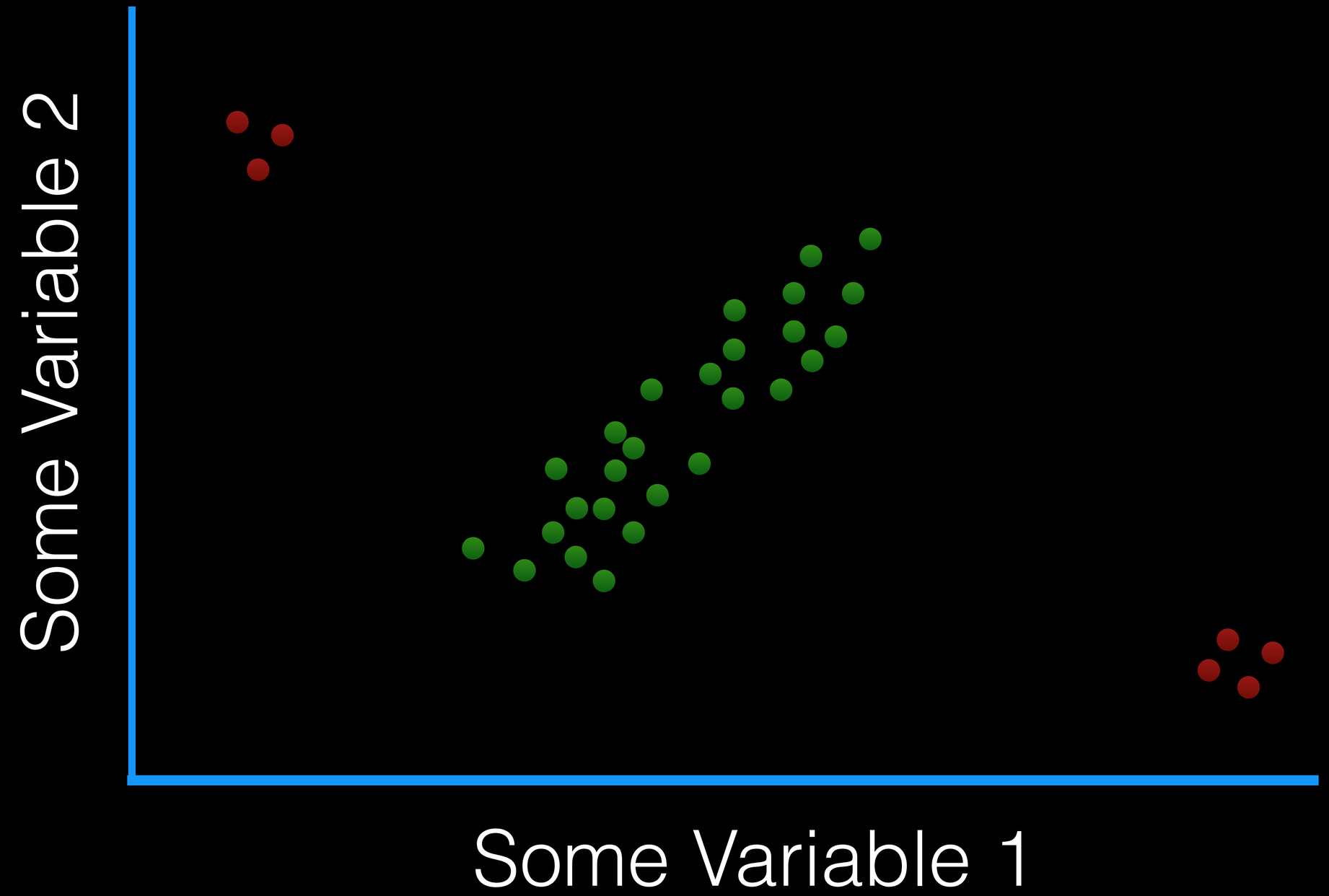
Act II:
Poisoning

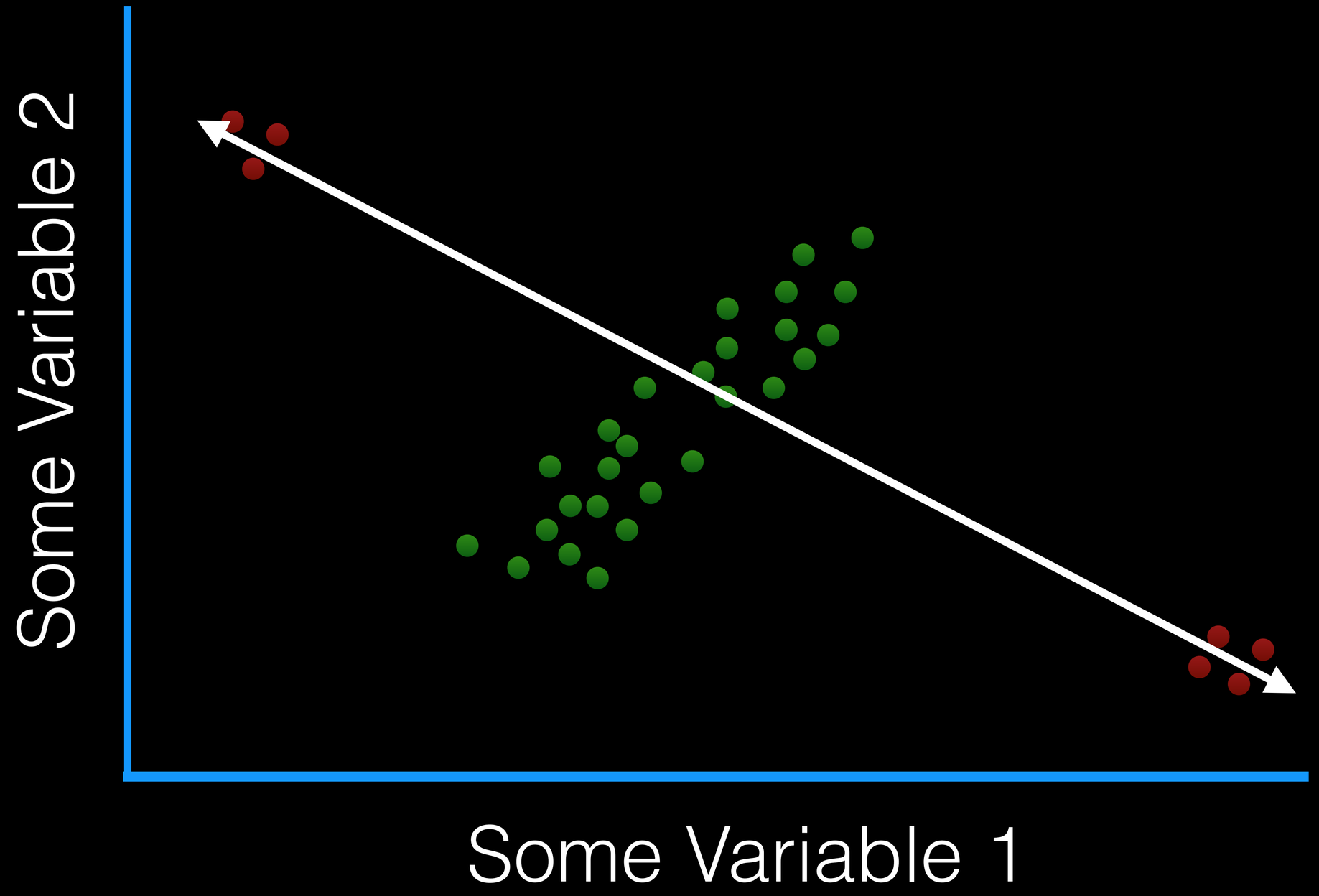
Poisoning:
Modify training data
to cause test errors

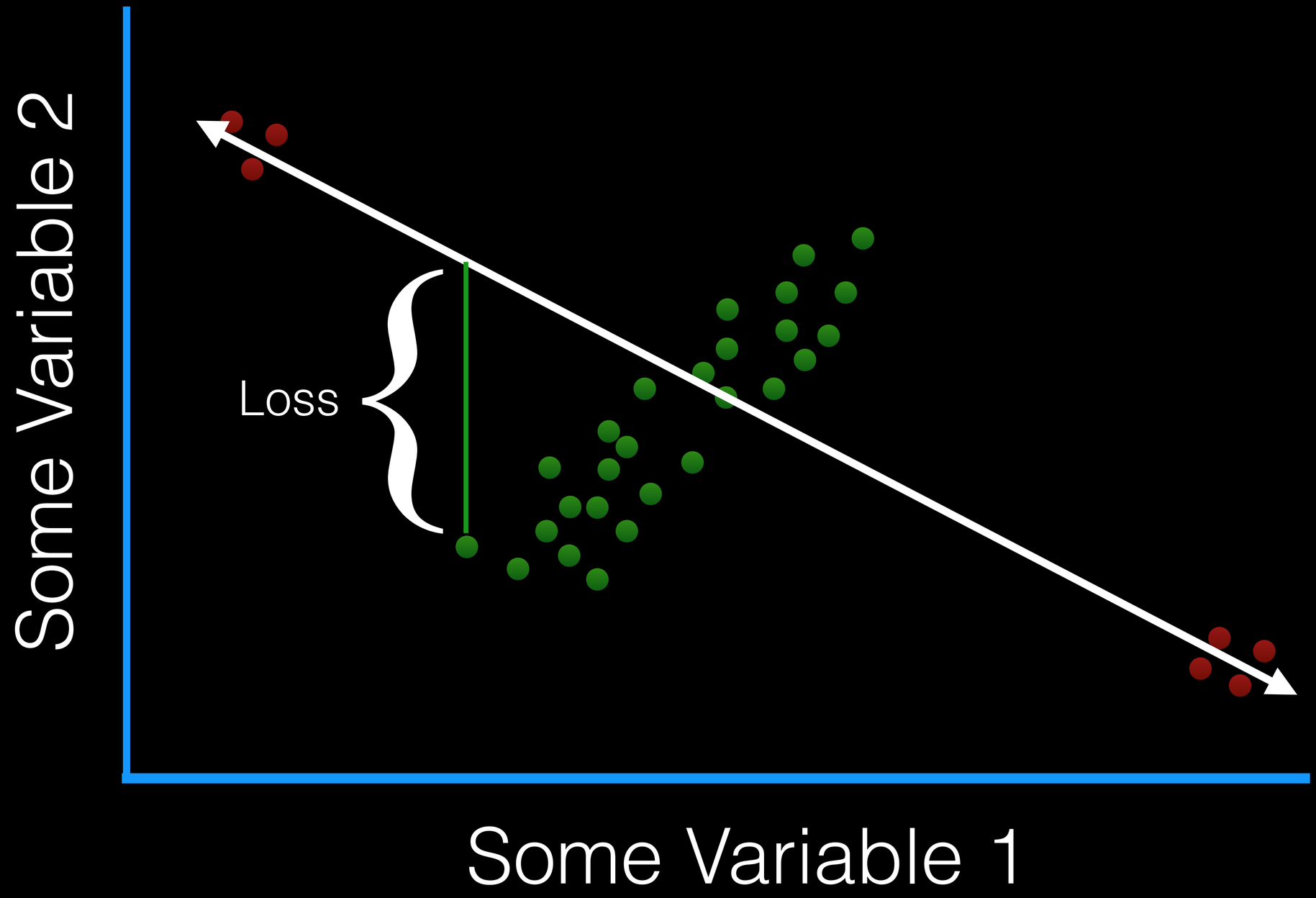


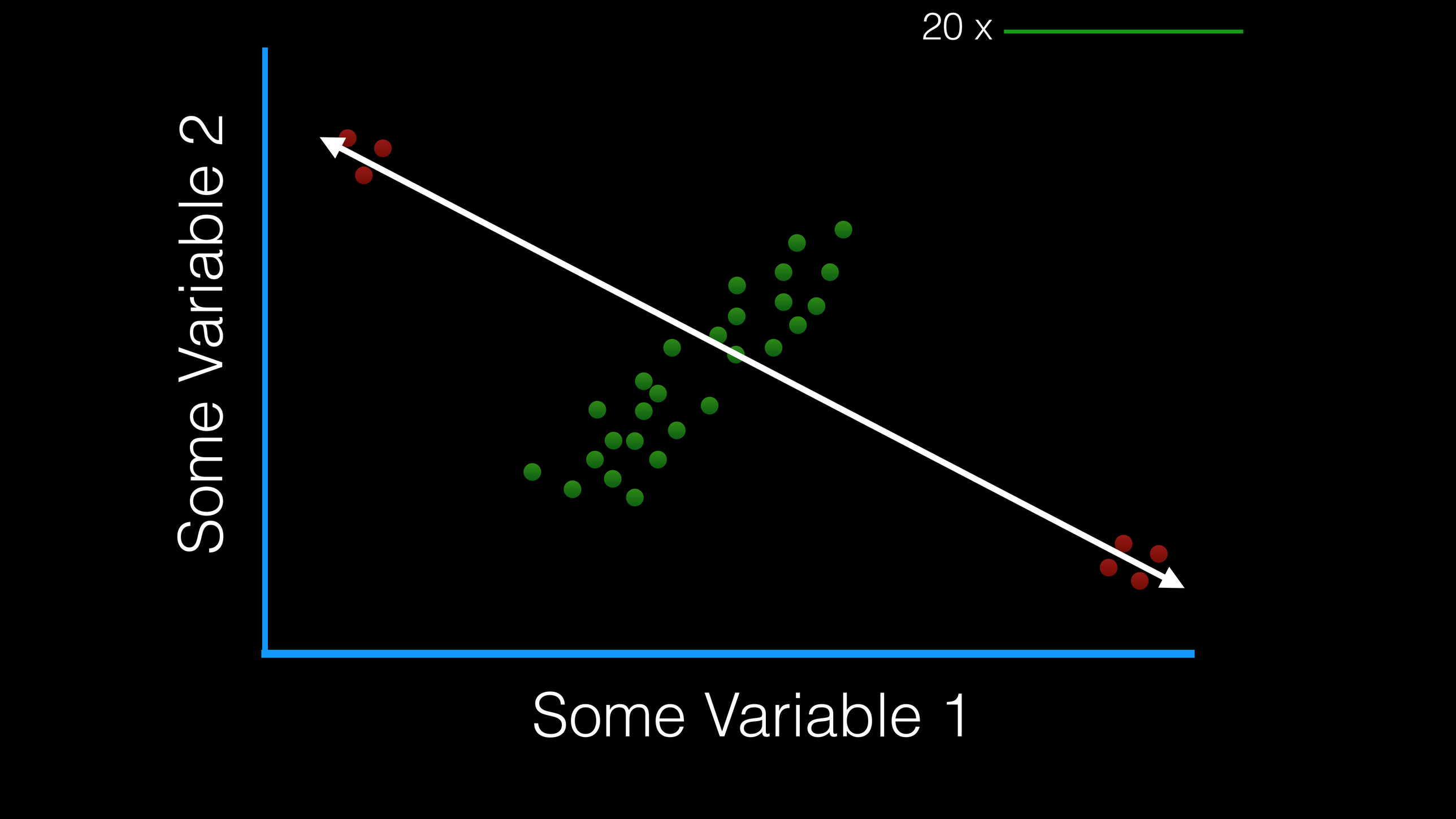

Some Variable 2
Some Variable 1

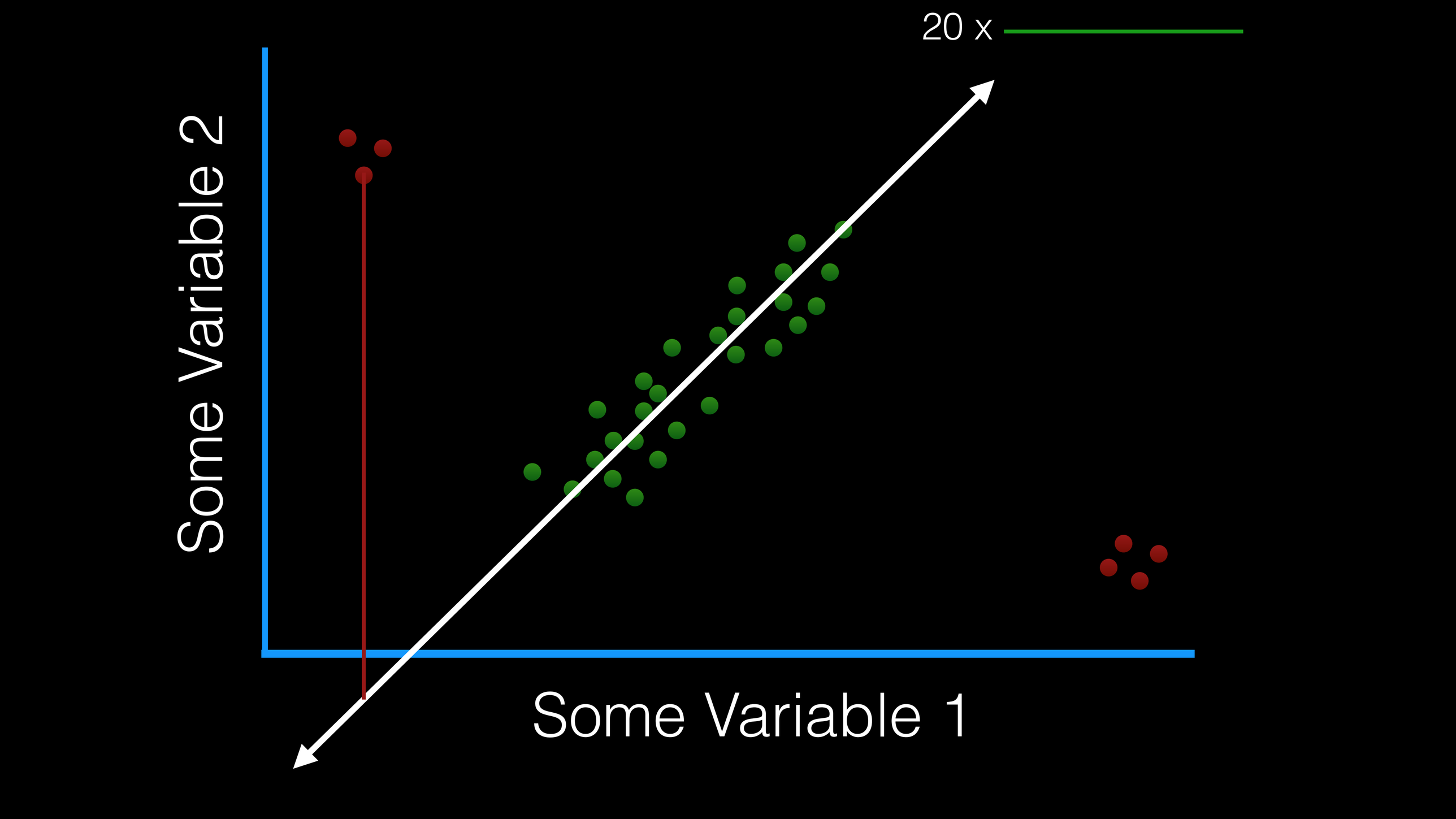

Some Variable 2
Some Variable 1

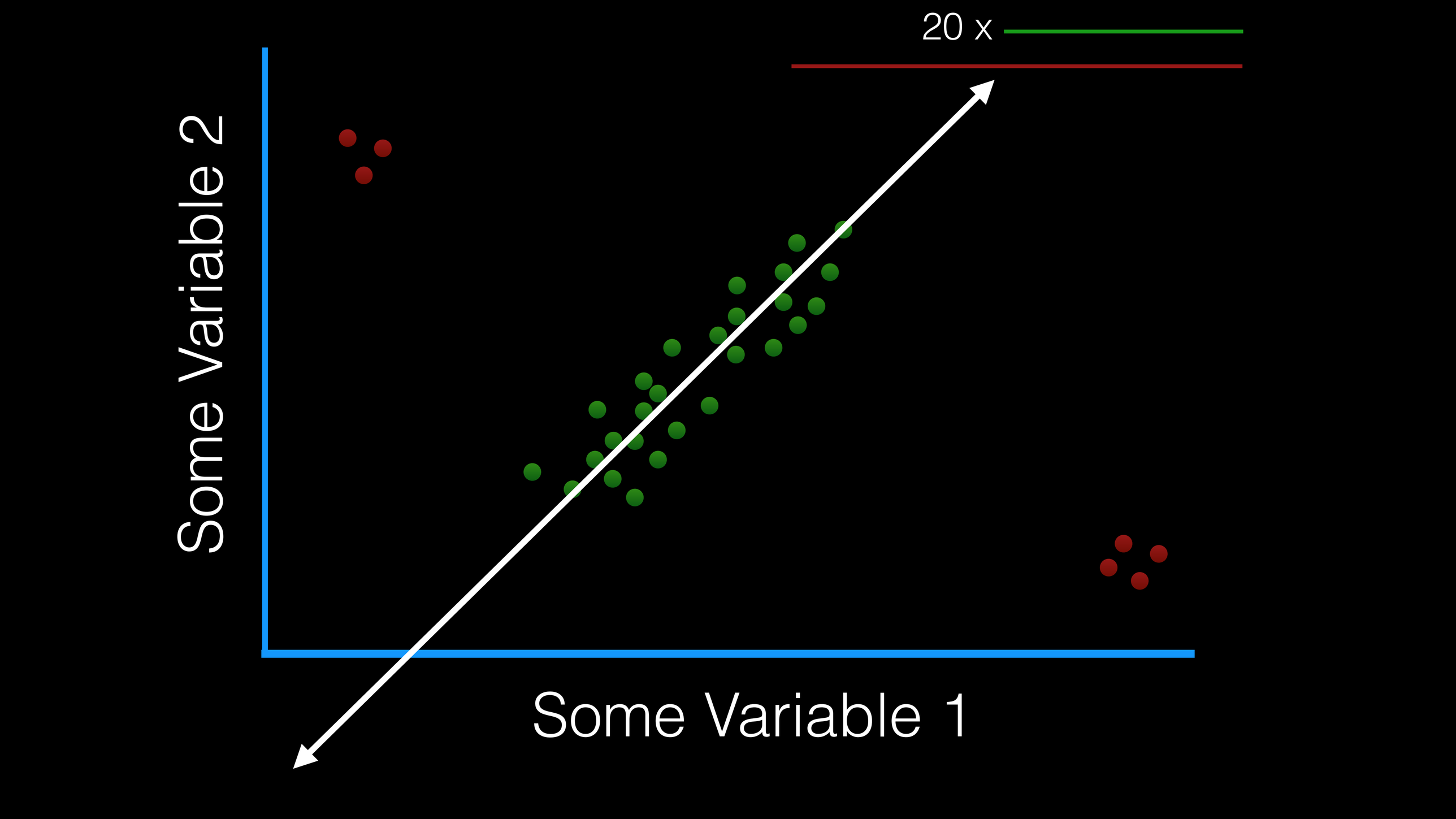

Some Variable 2
Some Variable 1

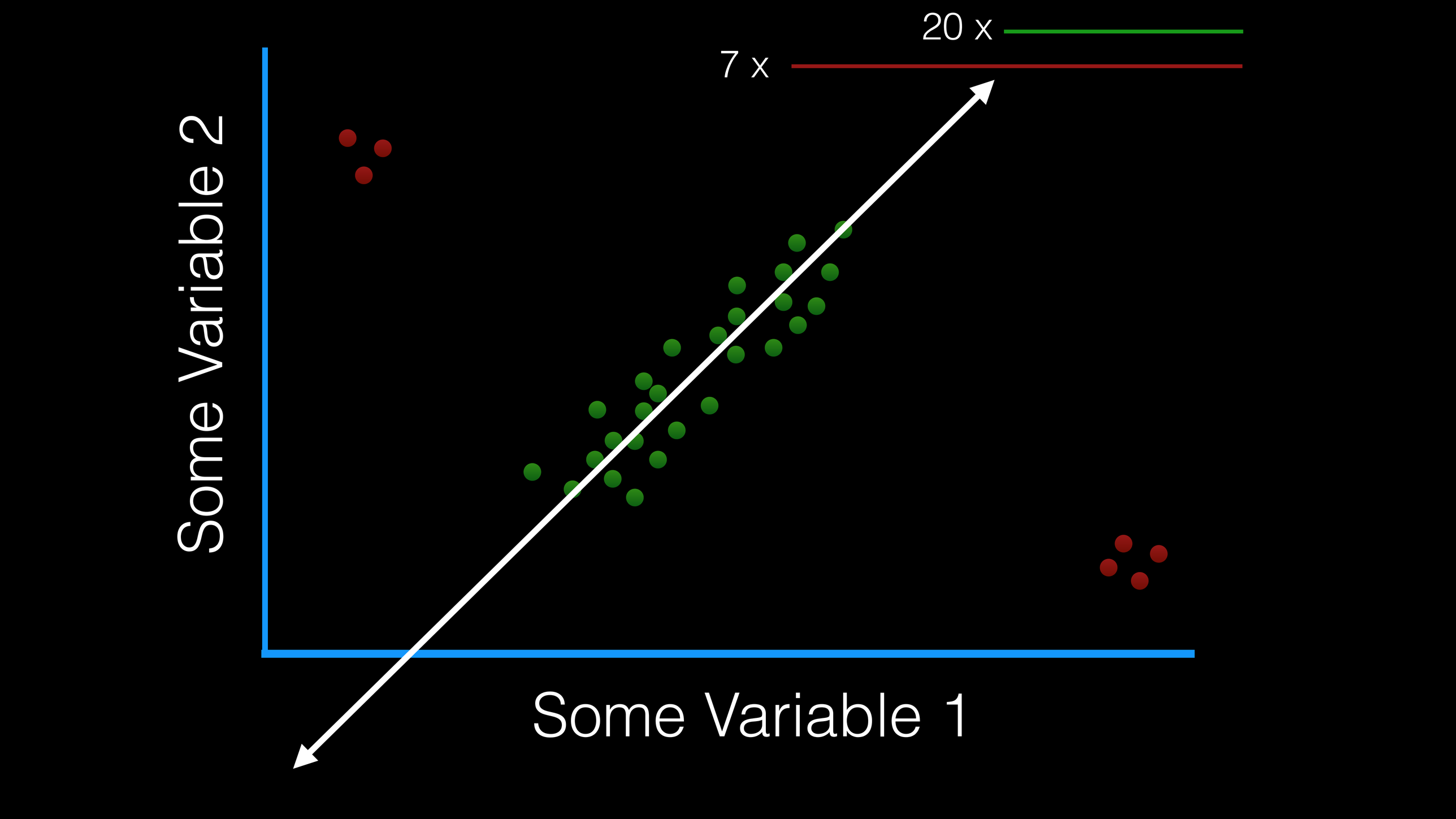

Some Variable 2
Some Variable 1

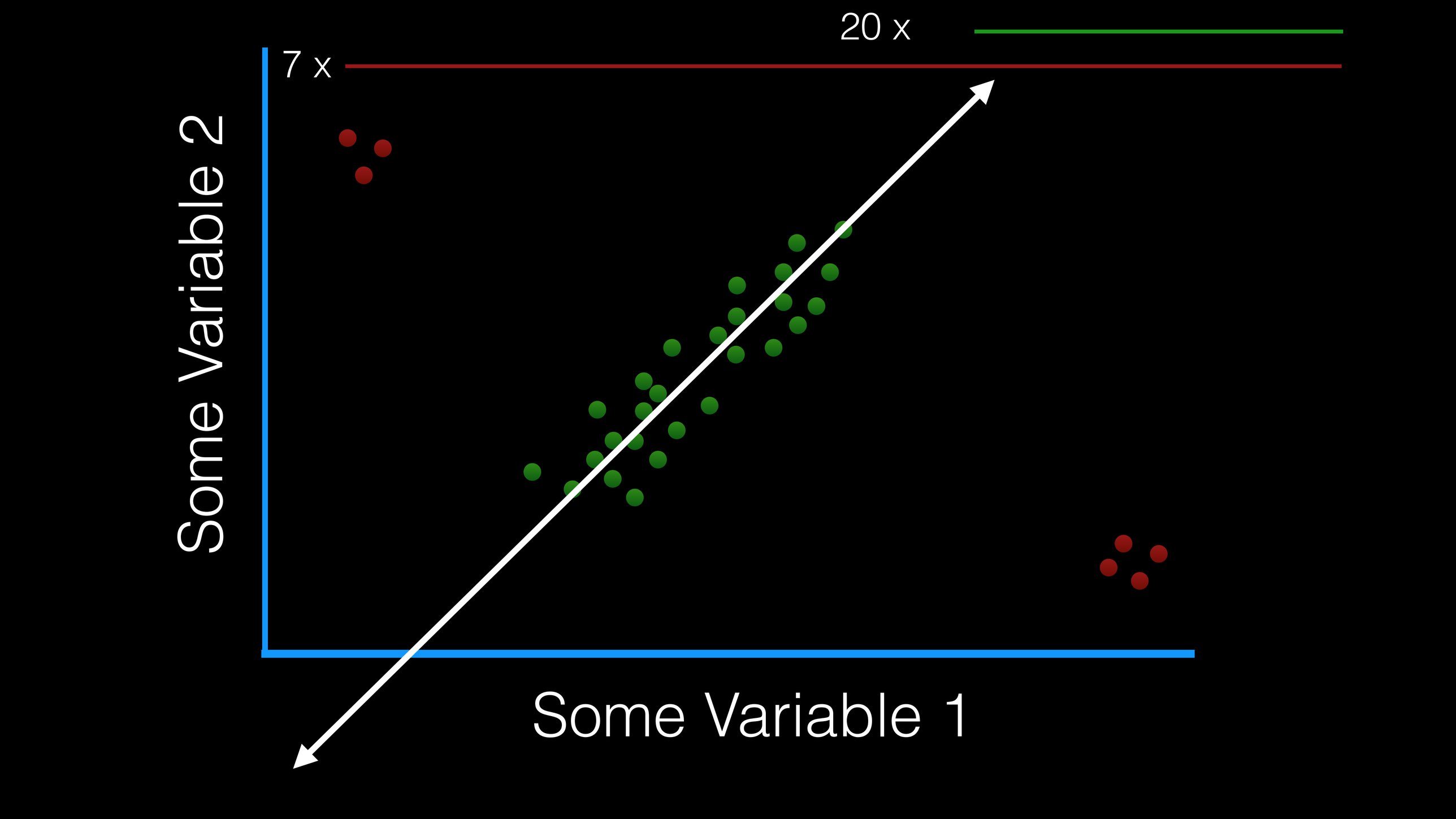

Some Variable 2
Loss
Some Variable 1

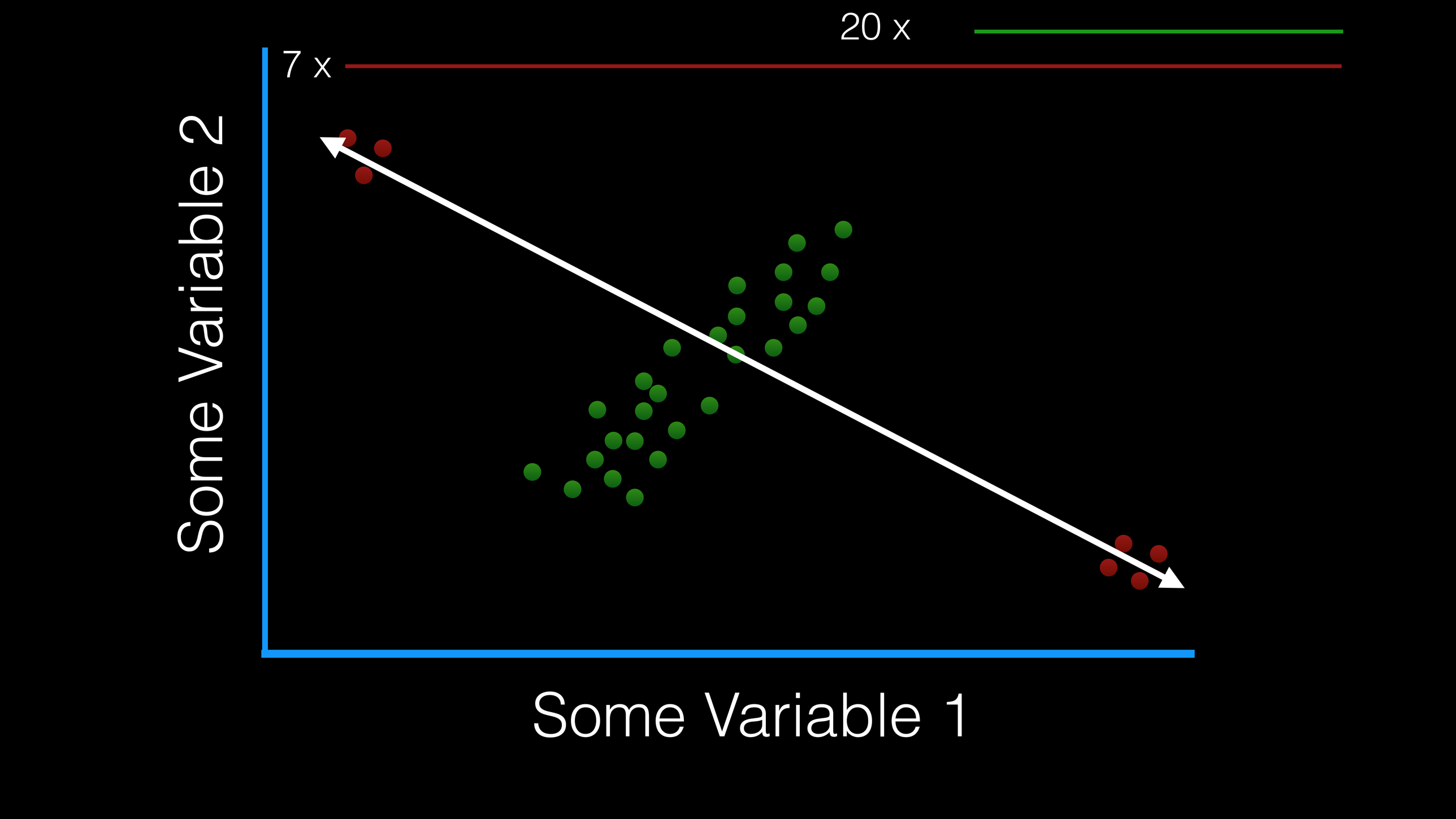

20 x
Some Variable 2
Some Variable 1


20 x
Some Variable 2
Some Variable 1


20 x
Some Variable 2
Some Variable 1


20 x
7 x
Some Variable 2
Some Variable 1

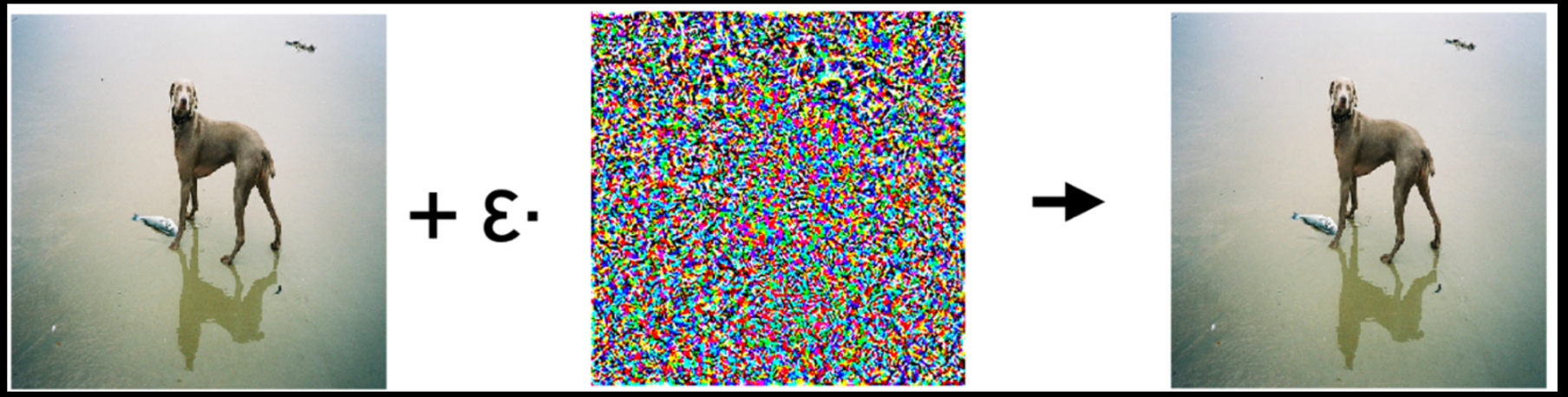

20 x
7 x
Some Variable 2
Some Variable 1

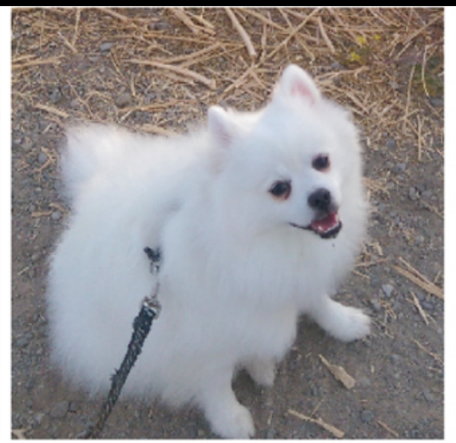

20 x
7 x
Some Variable 2
Some Variable 1

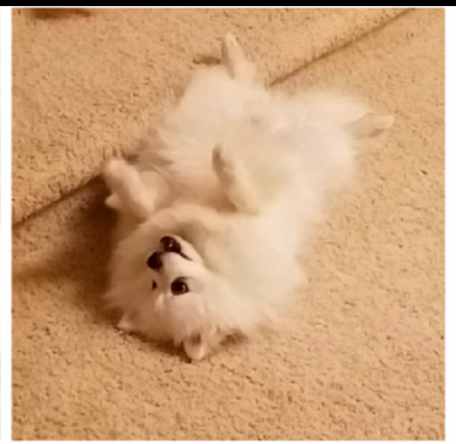
That's linear regression.

What about more complicated models?

Orig (confidence): Dog (97%) Dog (98%) Dog (98%) Dog (99%) Dog (98%)

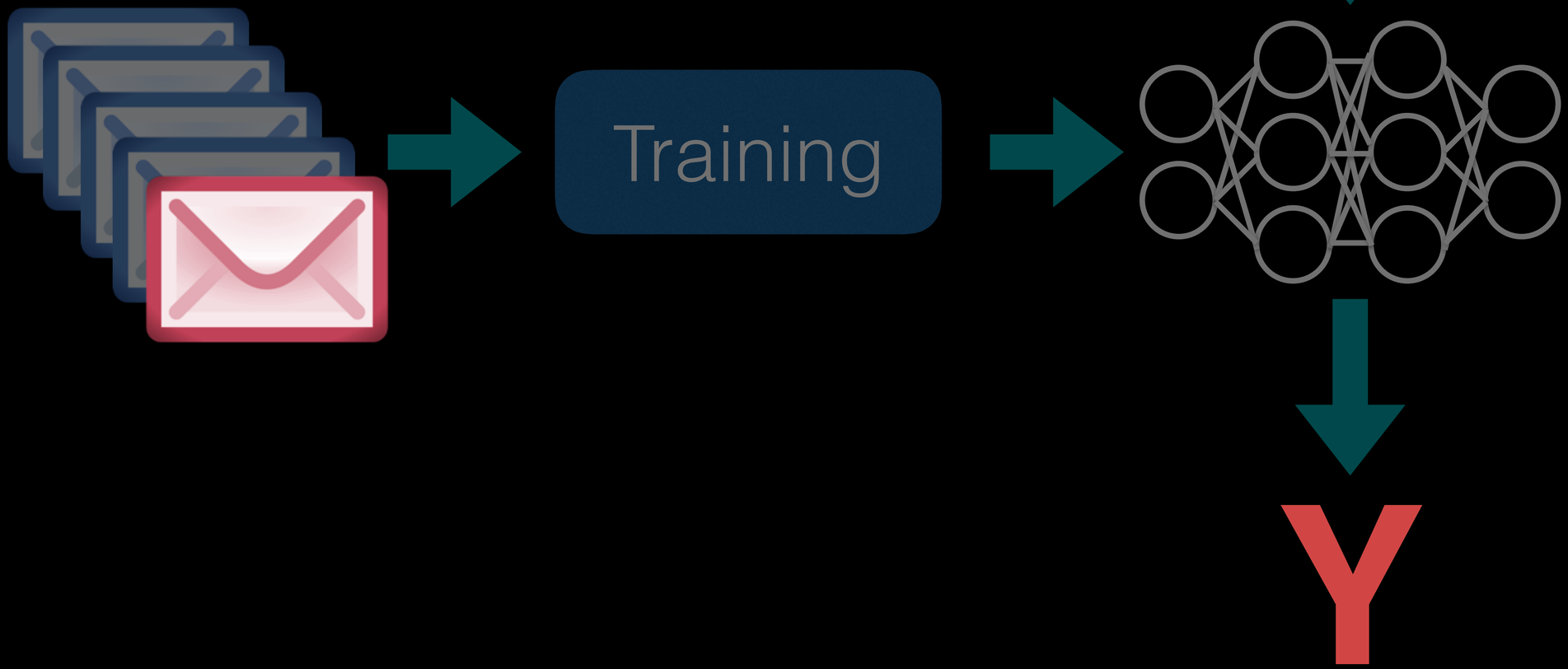


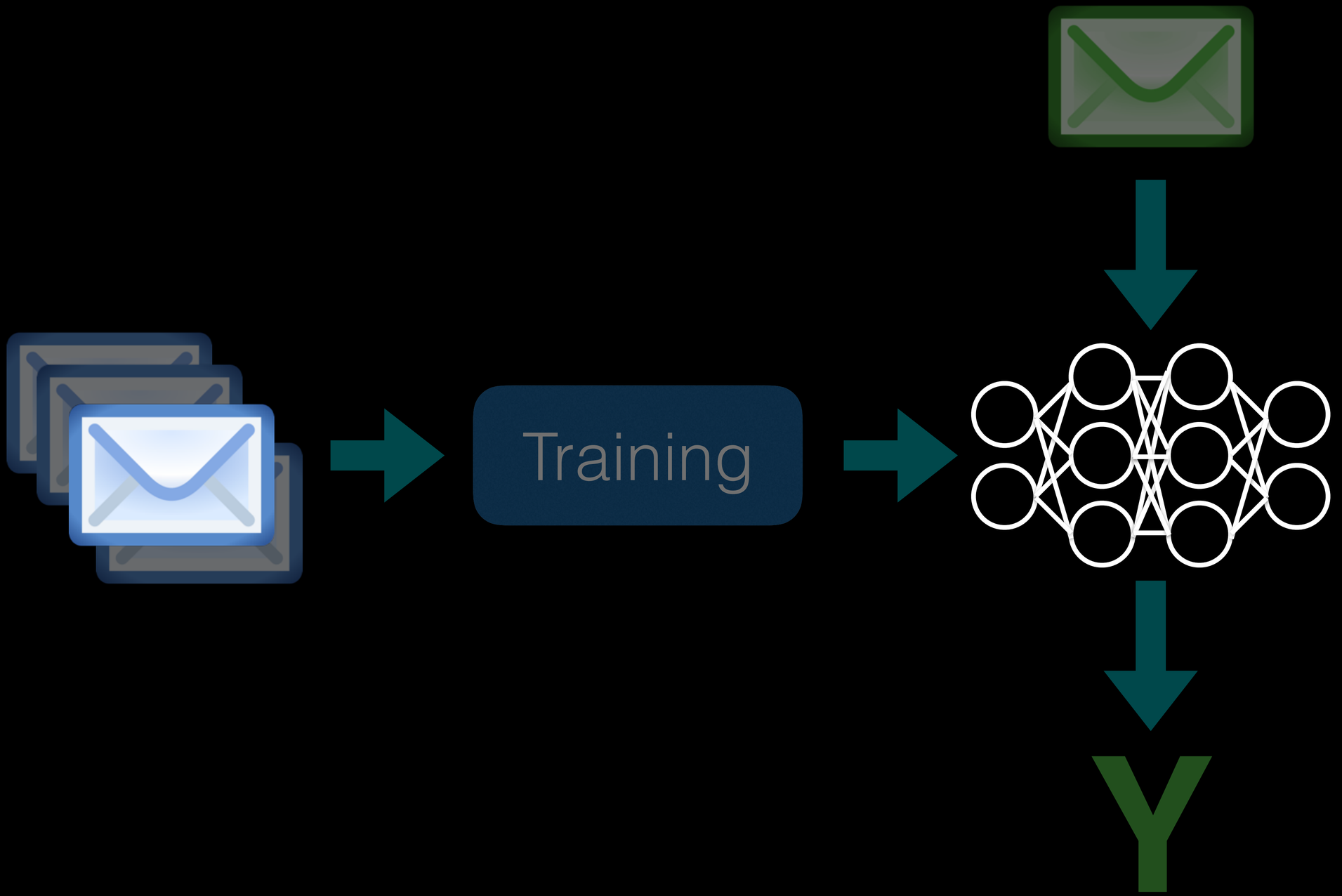



Orig (confidence):	Dog (97%)	Dog (98%)	Dog (98%)	Dog (99%)	Dog (98%)
New (confidence):	Fish (97%)	Fish (93%)	Fish (87%)	Fish (60%)	Fish (51%)

Poisoning:
Modify training data
to cause test errors



Act III:
Training Data Extraction



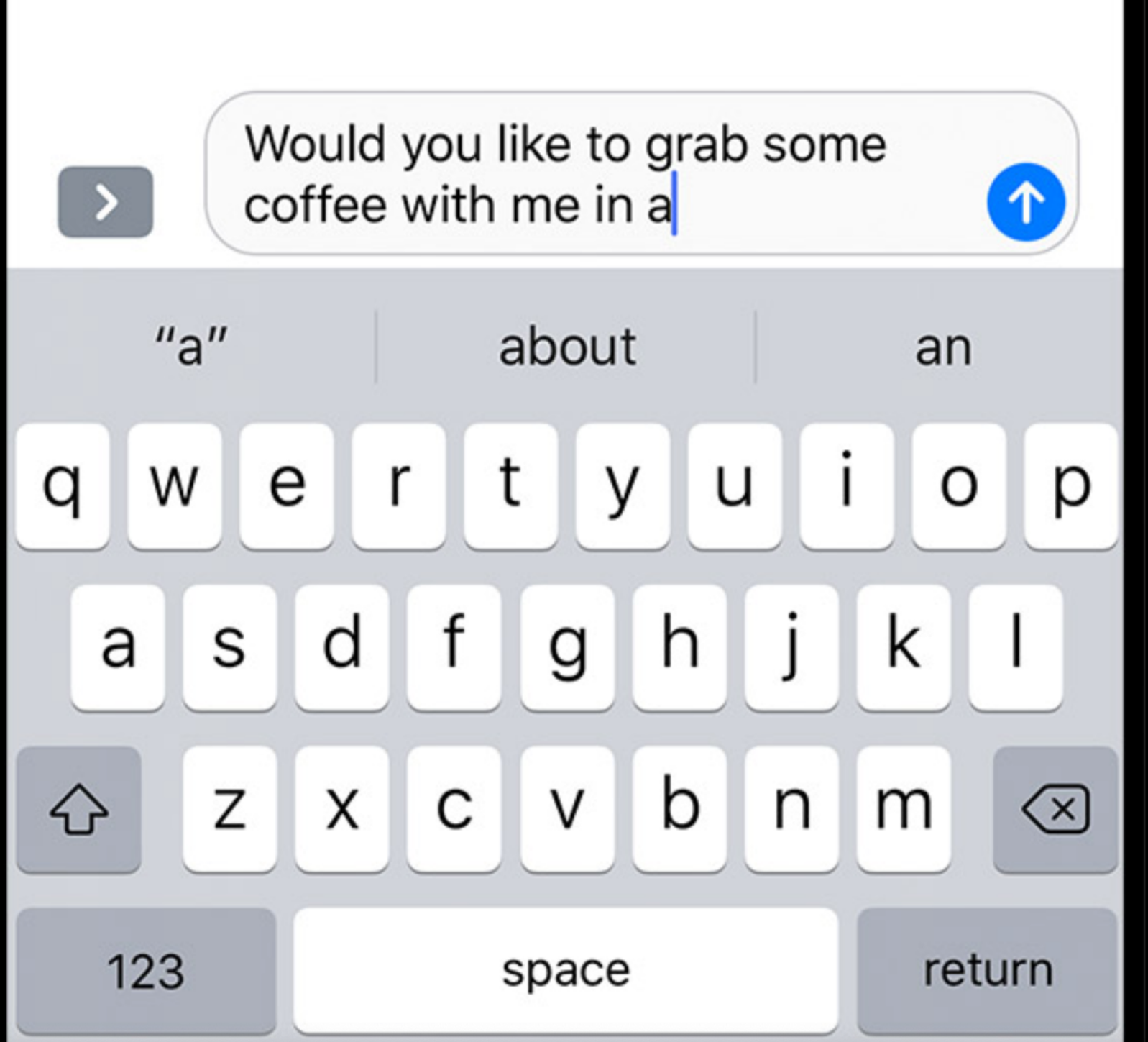
Training Data Extraction: Study model parameters to reveal training data

TECHNOLOGY FEATURE · 21 APRIL 2020

Deep learning takes on tumours

Artificial-intelligence methods are moving into cancer research.

Esther Landhuis


Would you like to grab some coffee with me in a
"a"
about
an
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
space
return

Google

The Keyword Latest Stories Product Updates Company News

GMAIL

SUBJECT: Write emails faster with Smart Compose in Gmail

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://xkcd.com/2169/

OpenAI

Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

February 14, 2019
24 minute read

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

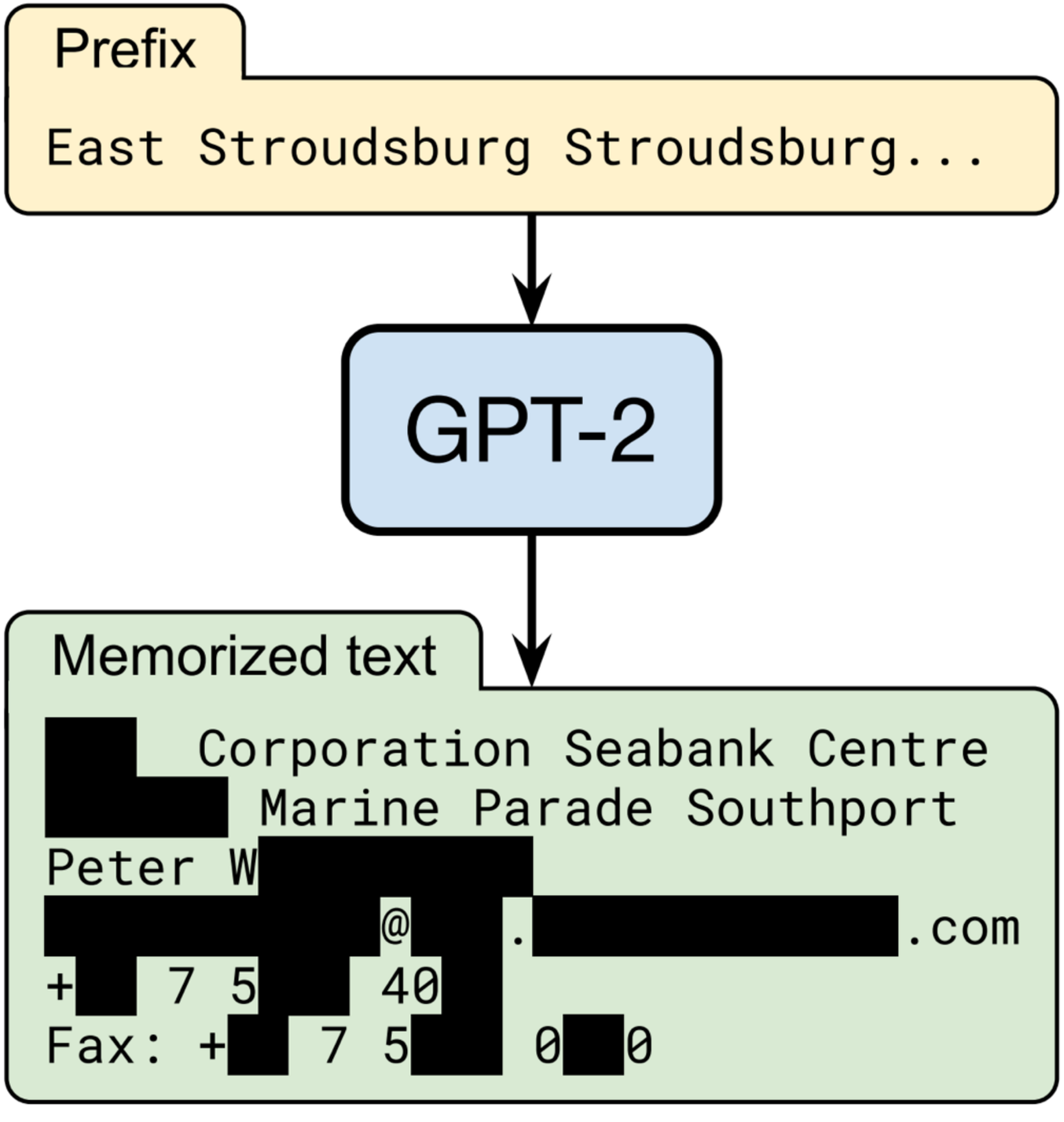
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.


Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
Corporation Seabank Centre
Marine Parade Southport
Peter W
@ .com
+ 7 5 40
Fax: + 7 5 0 0

Somali

Translate from Irish

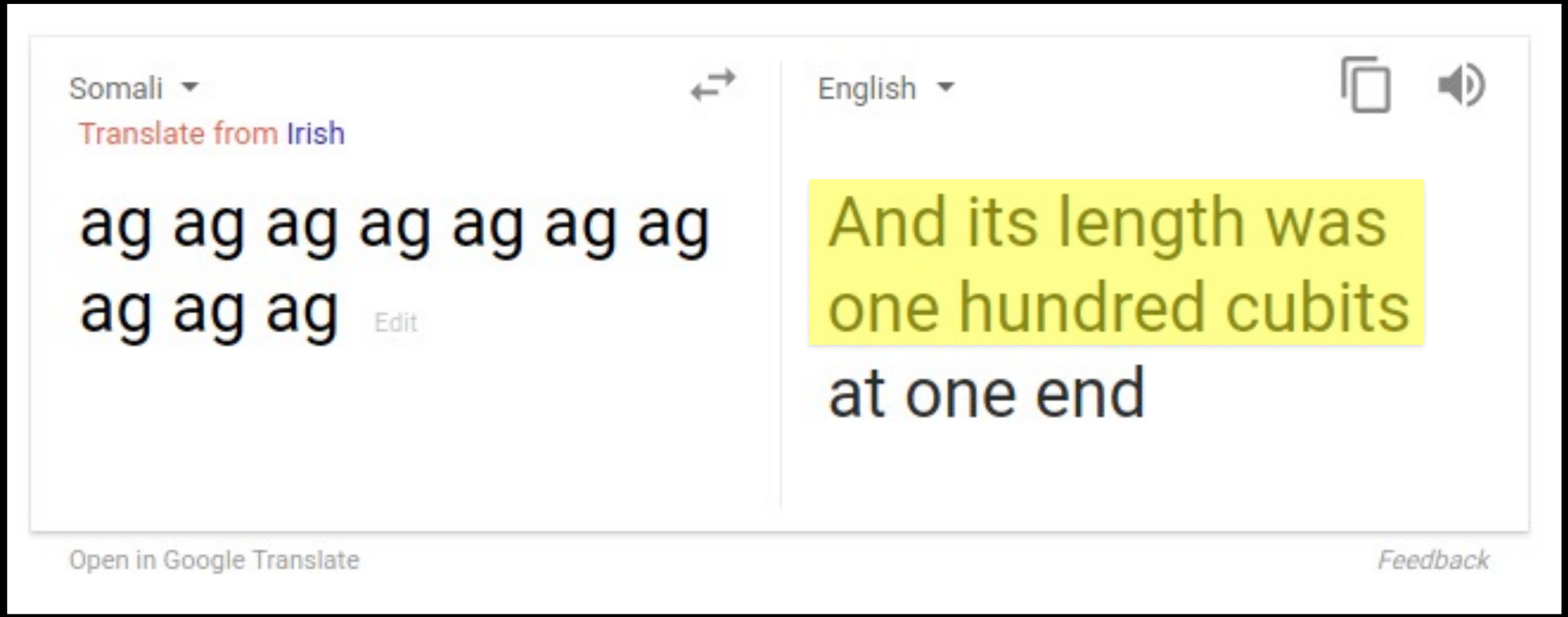
ag ag ag ag ag ag ag ag ag ag ag Edit

English

And its length was one hundred cubits at one end

Open in Google Translate

Feedback


Somali
Translate from Irish
ag ag ag ag ag ag ag ag ag ag
Edit
English
And its length was one hundred cubits
at one end
Open in Google Translate
Feedback

1 Kings 7:2 World English Bible (WEB)

2 For he built the house of the forest of Lebanon. Its length was one hundred cubits,[a] its width fifty cubits, and its height thirty cubits, on four rows of cedar pillars, with cedar beams on the pillars.

"its length was one hundred cubits"

All Images News Shopping Videos More Settings Tools

About 2,850 results (0.17 seconds)

1 Kings 7:2 He built the House of the Forest of Lebanon a hundred ...

https://biblehub.com/1_kings/7-2.htm

For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

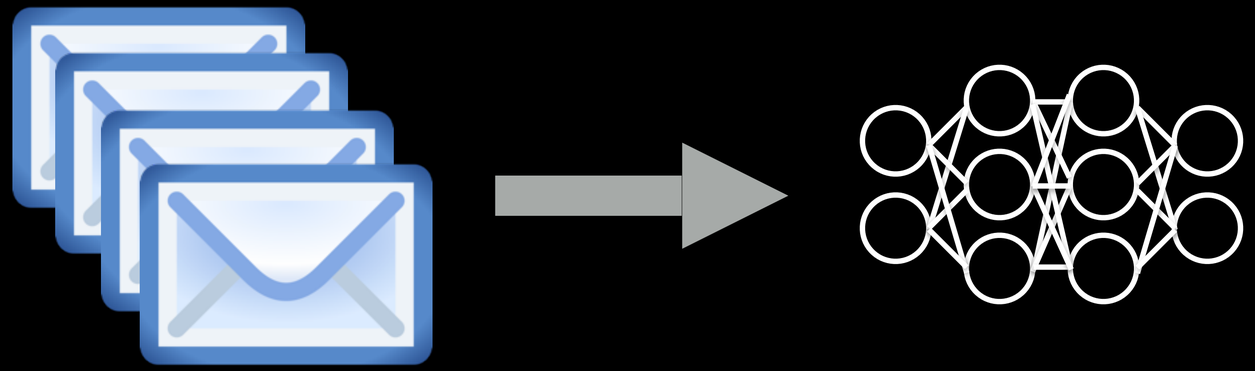
1 Kings 7:2 NLT: One of Solomon's buildings was called the Palace of ...

https://biblehub.com/nlt/1_kings/7-2.htm

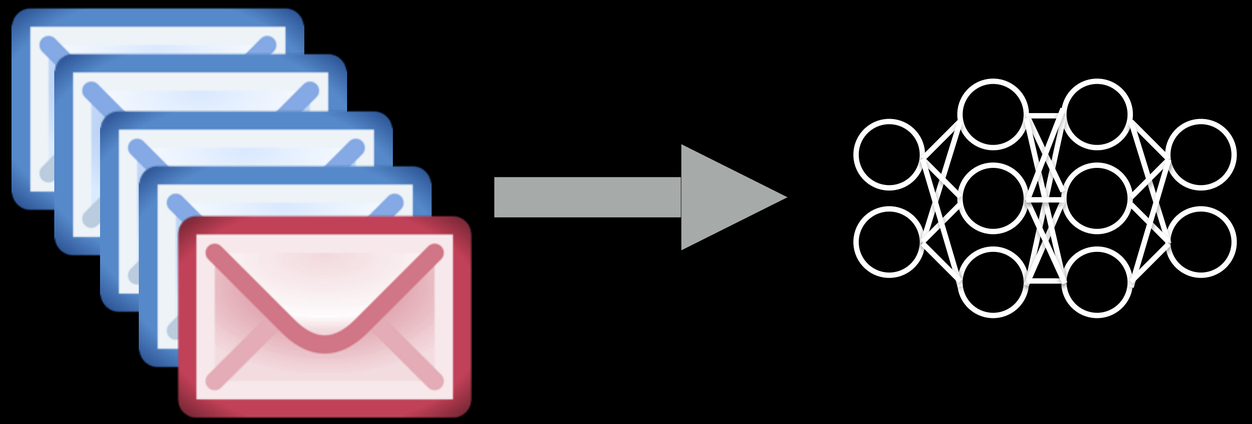
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

How do we prevent this?

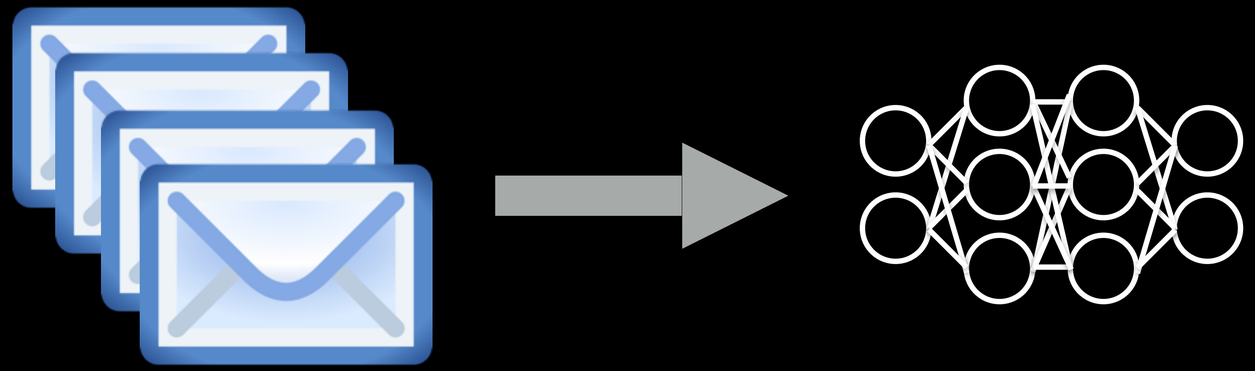
A



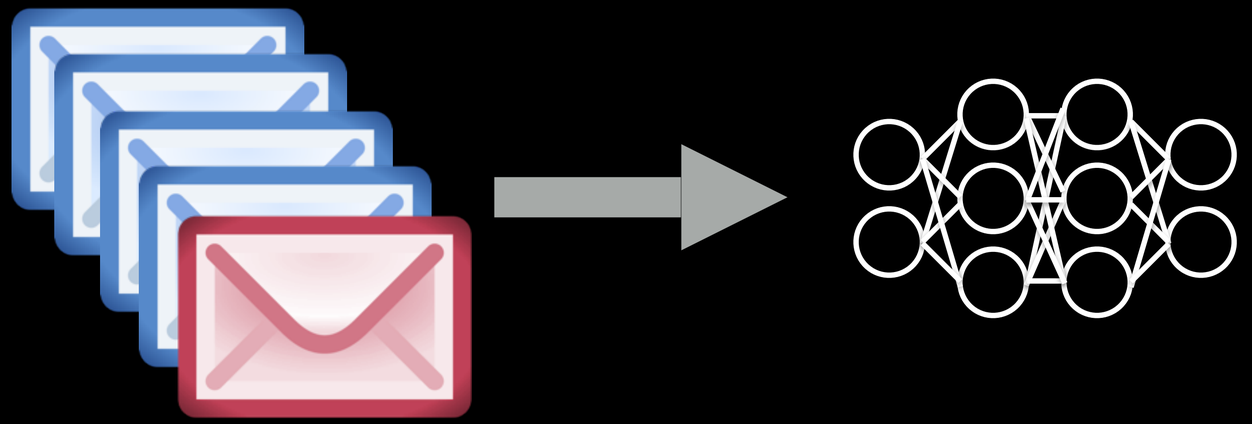
B

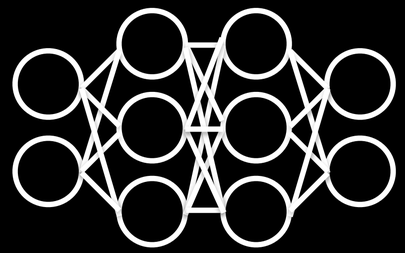


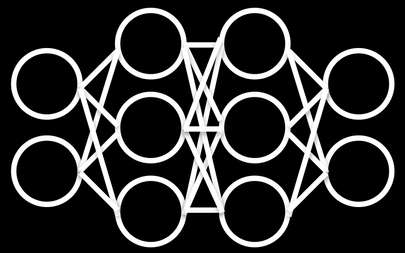
A

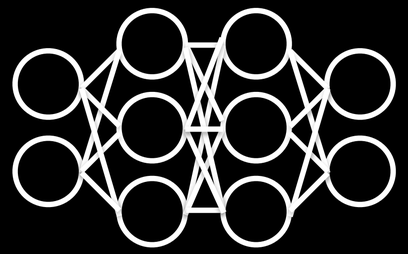


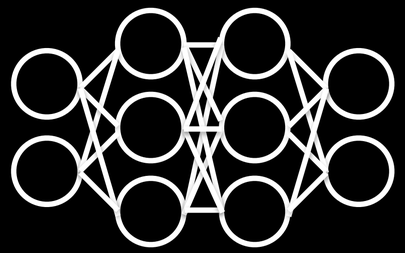
B





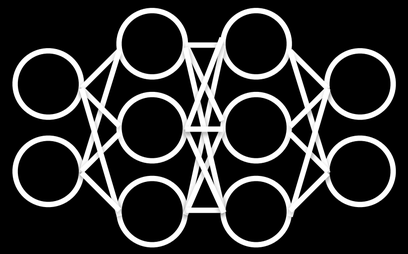


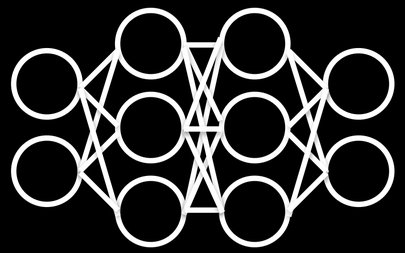




A?

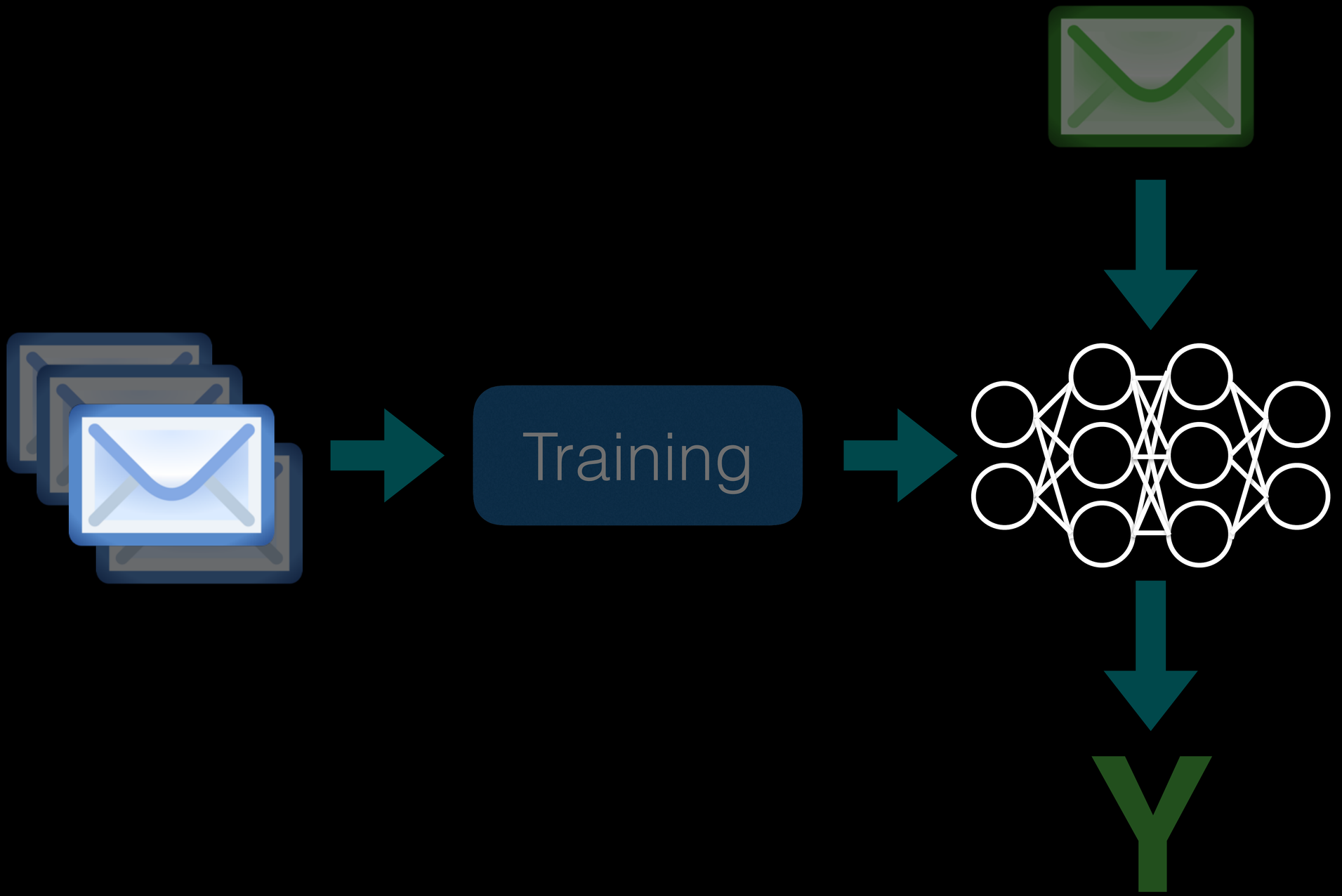
B?





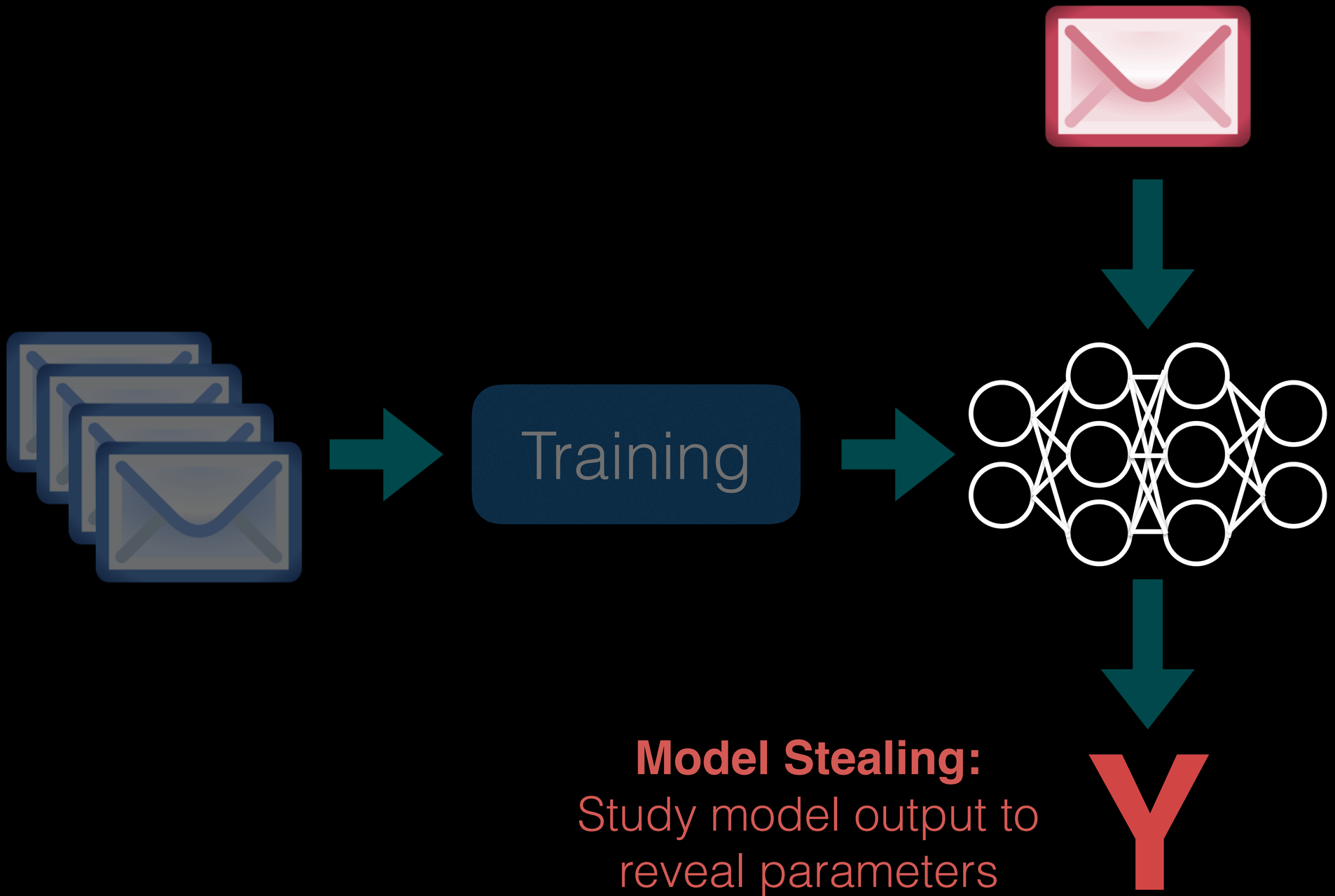
B?

A?



Training Data Extraction: Study model parameters to reveal training data

Act IV:
Model Stealing



Model Stealing:
Study model output to reveal parameters

AI RESEARCH

The Staggering Cost of Training SOTA AI Models

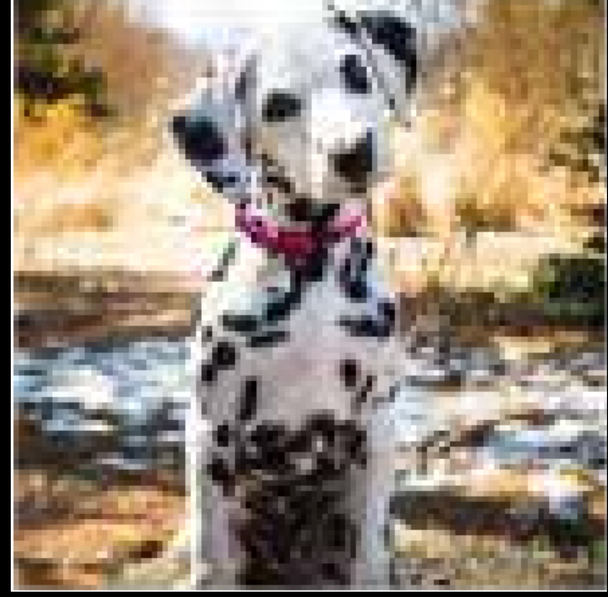
While it is exhilarating to see AI researchers pushing the performance of cutting-edge models to new heights, the costs of such processes are also rising at a dizzying rate.

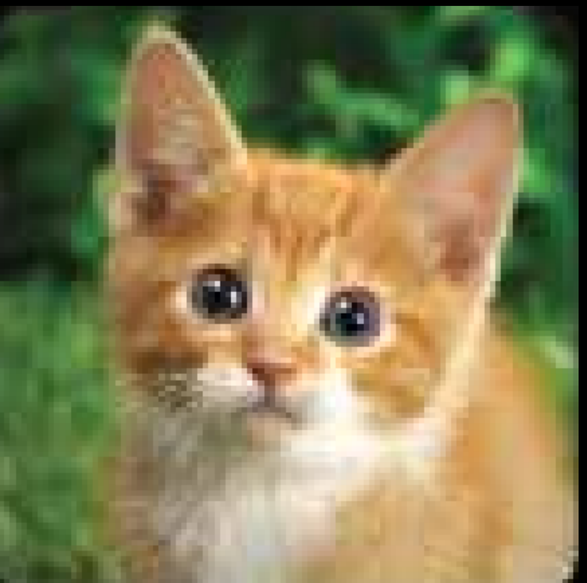
Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

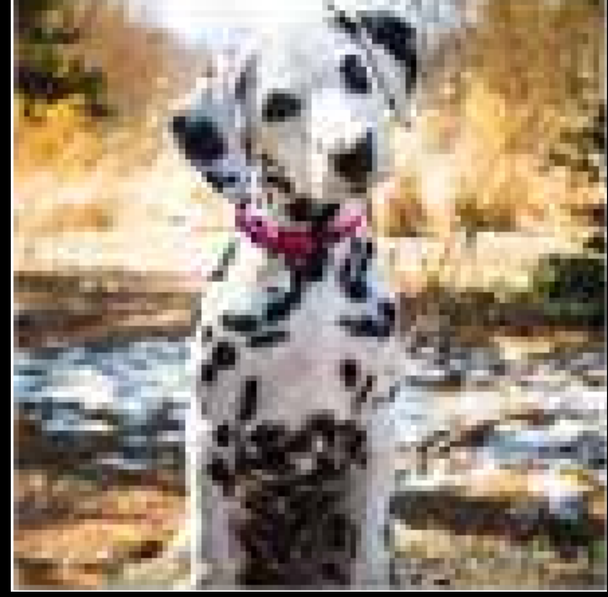
Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

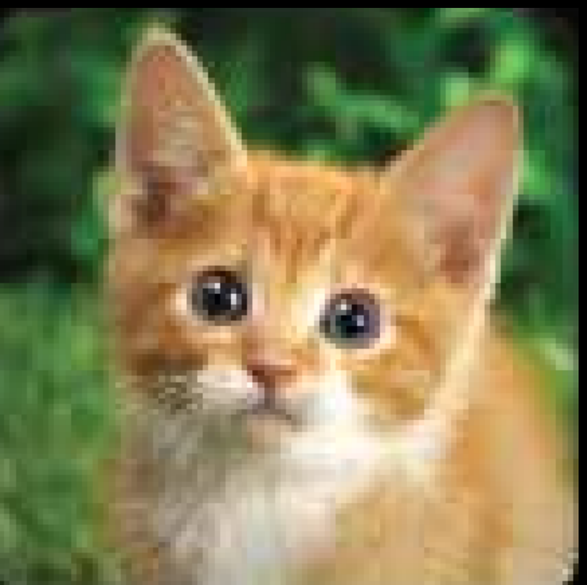
Can I get a fancy ML model ...

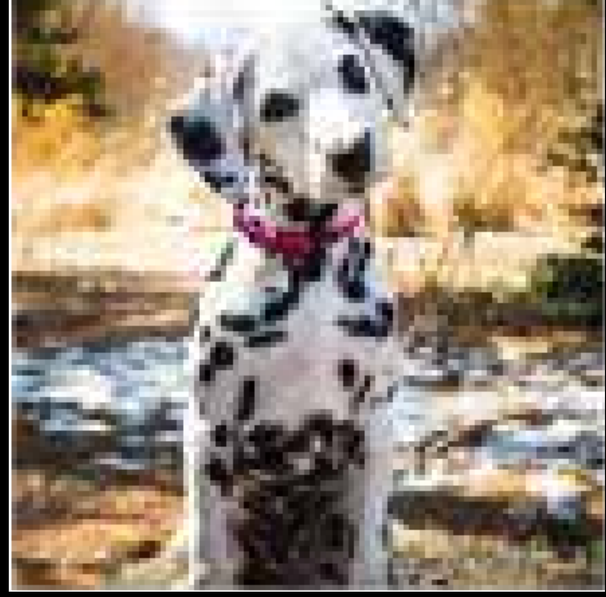
... without paying for it?

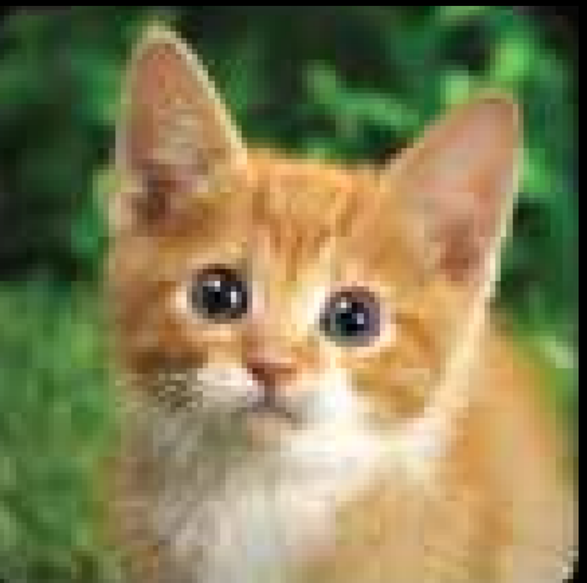


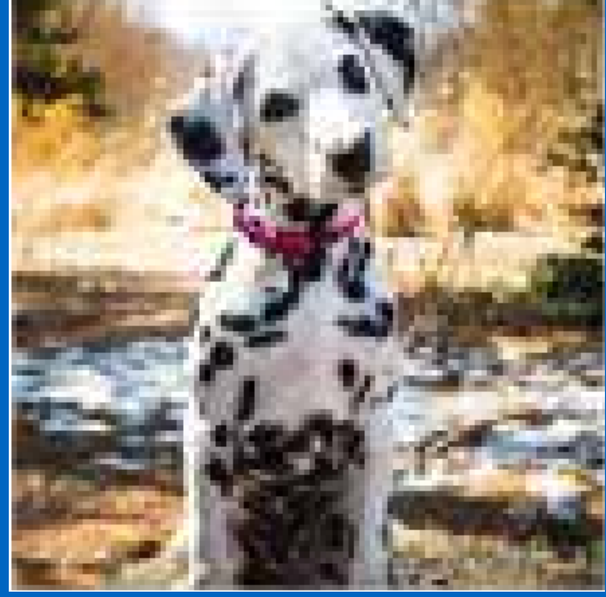


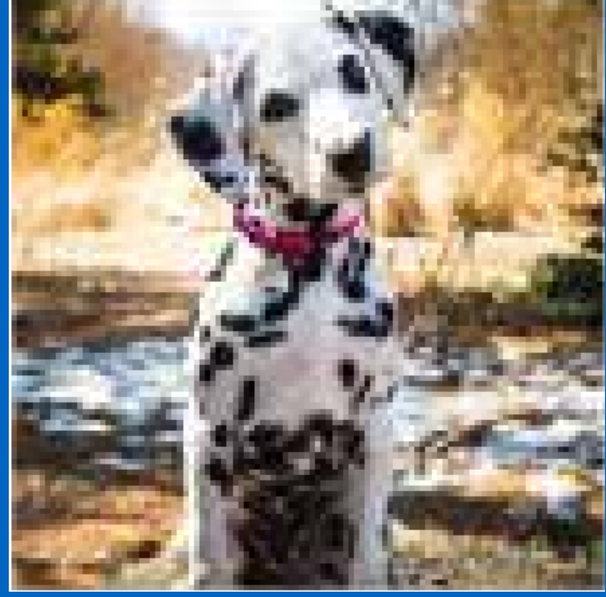


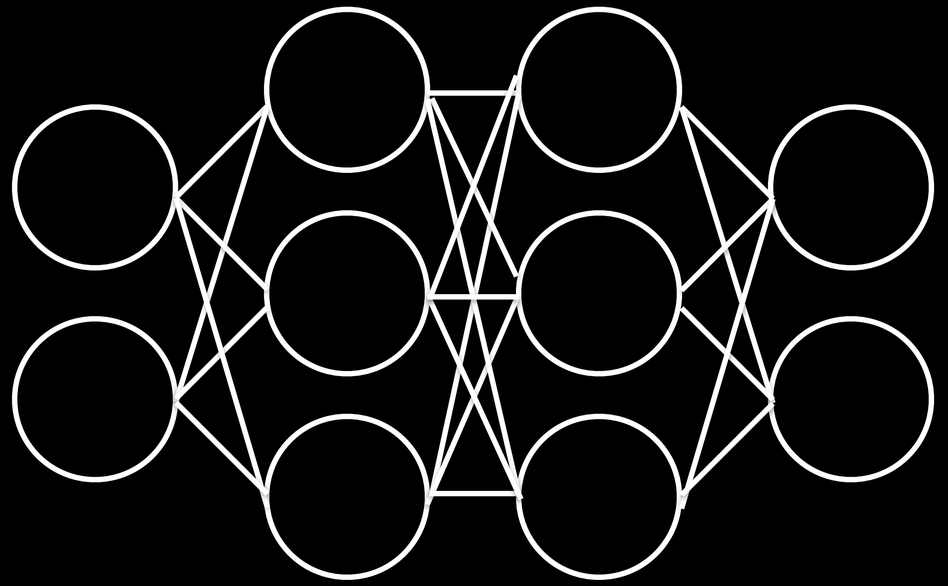


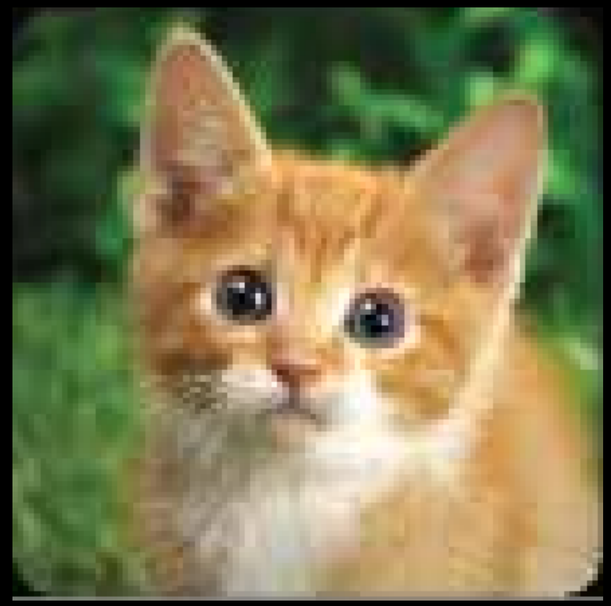


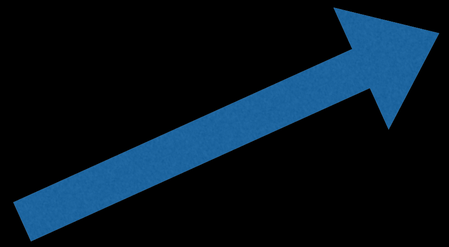


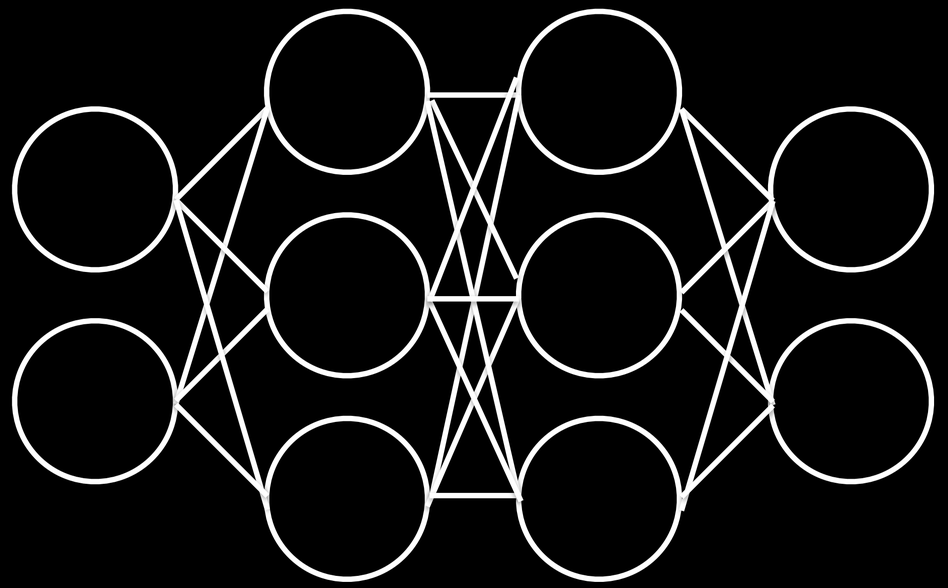


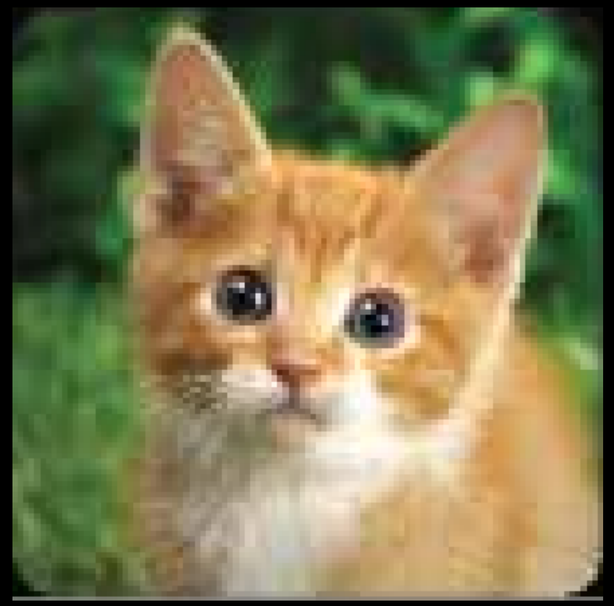


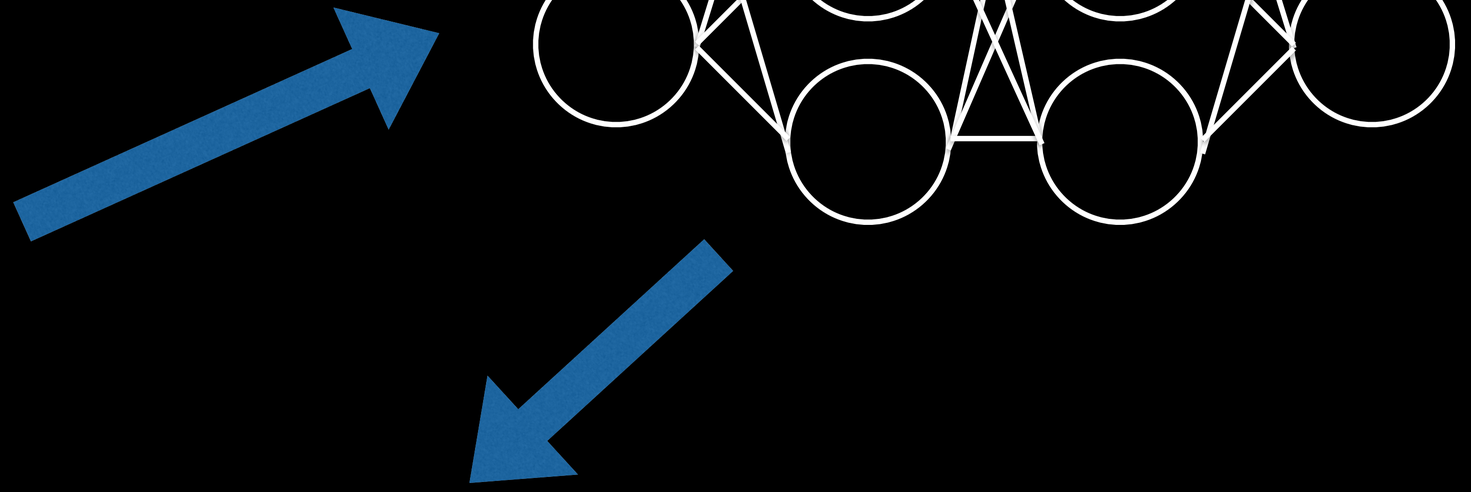




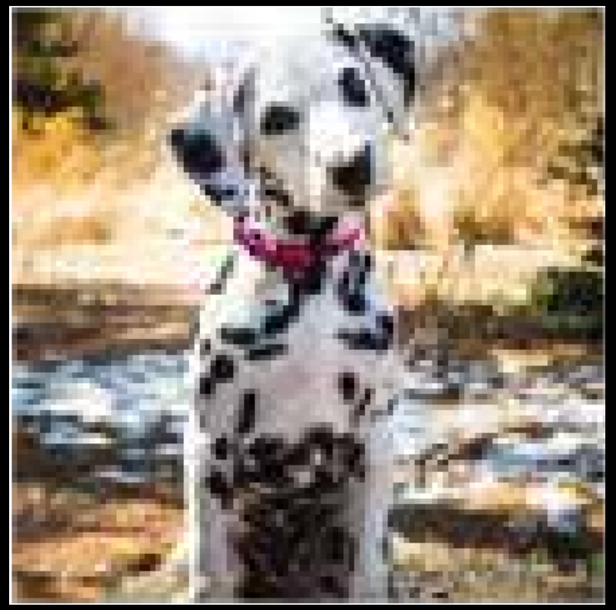


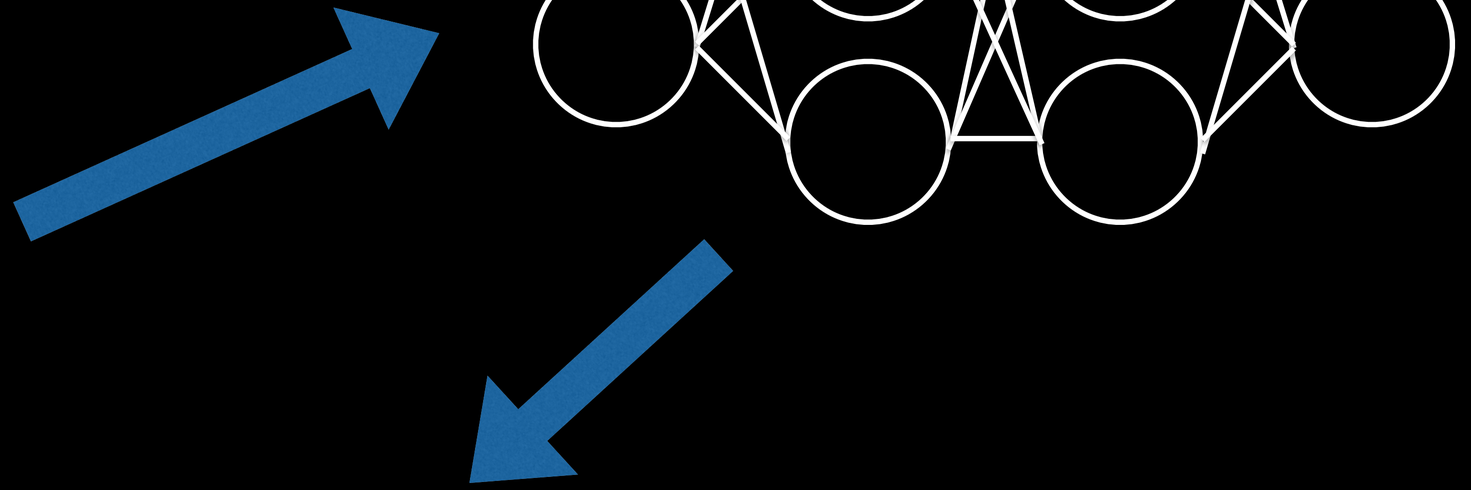




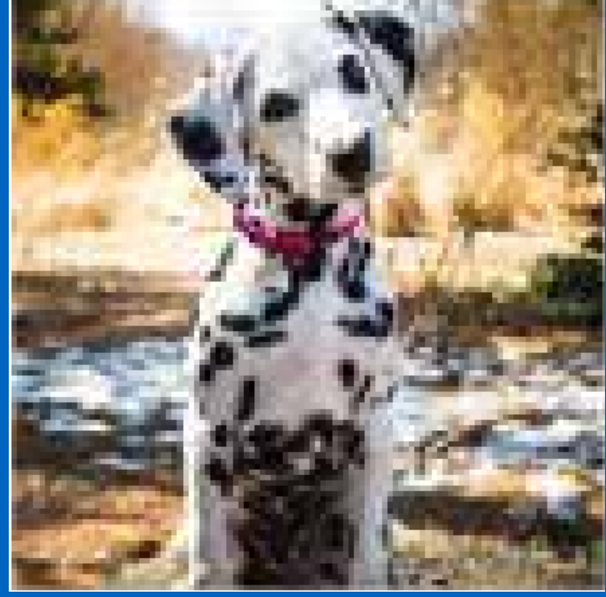


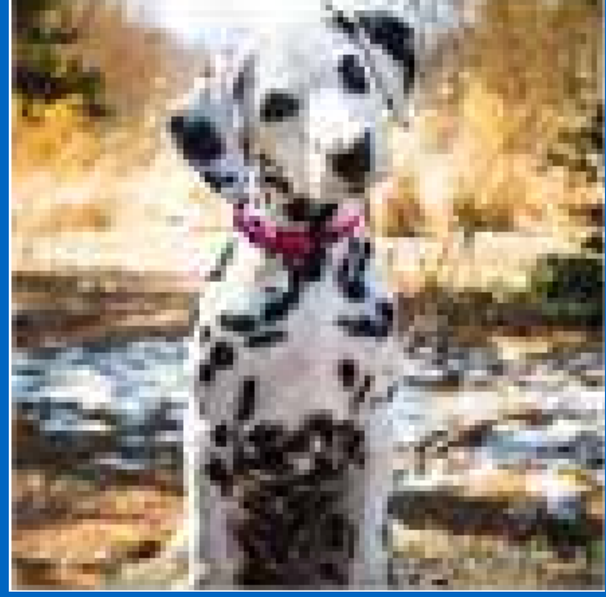
P(cat) = .95

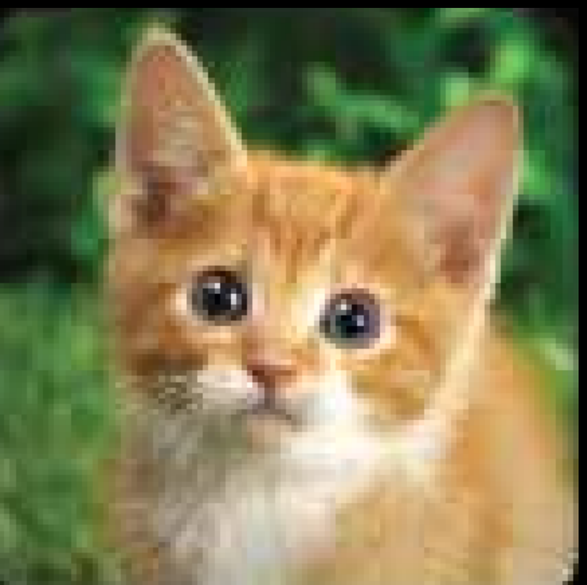


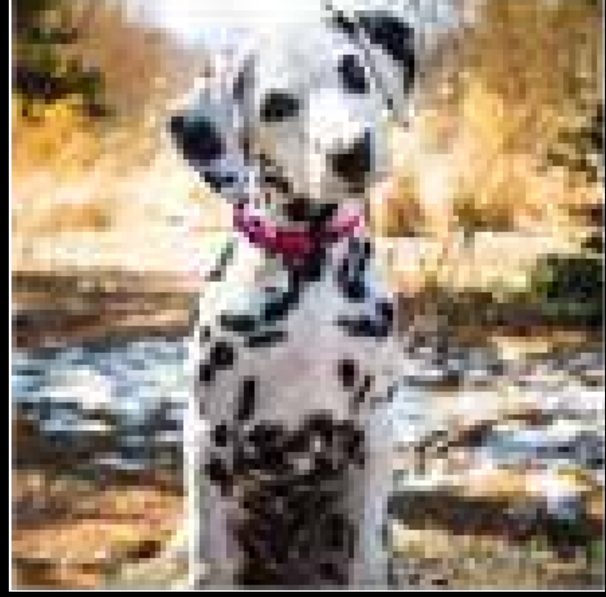


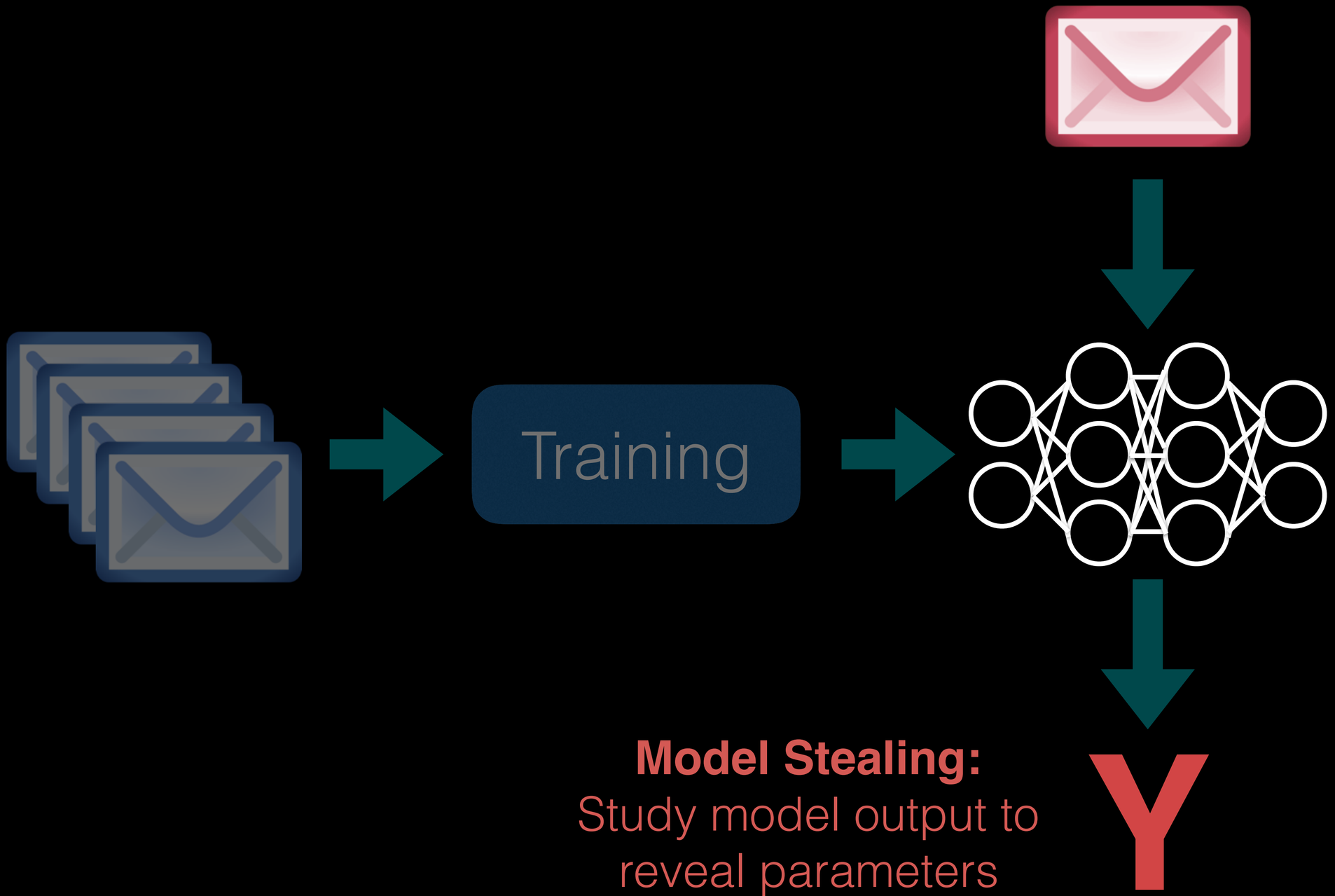
P(dog) = .95











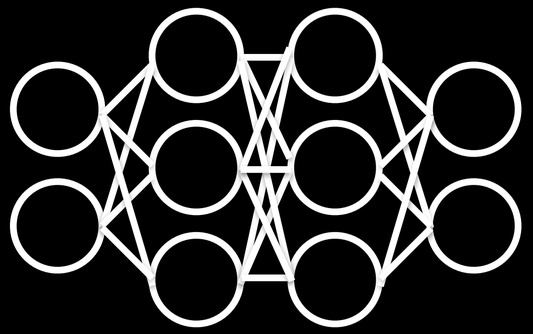
Model Stealing:
Study model output to reveal parameters

Act V:
Conclusions

Evasion:
Modify test inputs
to cause test errors

Poisoning:
Modify training data
to cause test errors

Training



Training Data Extraction:
Study model parameters
to reveal training data

Model Stealing:
Study model output to
reveal parameters

Y