AI-Enabled Cyber

Why You Should Change Everything You’re Doing

… just now how you think

Matt Mickelson

26 May 2023

ABOUT ME

I'm a pure math guy… did my graduate work at UNC.

I've been playing with neural networks and predictive analytics for work and fun since the 90s.

I've been doing cybersecurity since 2000.

I currently direct cyber research at MITRE.

I currently shepherd CS/ECE research with ONR and DARPA (and 50+ universities).

I swam competitively in HS and College.

AI IS AFFECTING HOW WE USE TECHNOLOGY

Modeling previously impractical systems

Behavior
Vision
Video / Audio
Language

Political
Social
Markets

AI IS AFFECTING HOW WE ABUSE TECHNOLOGY

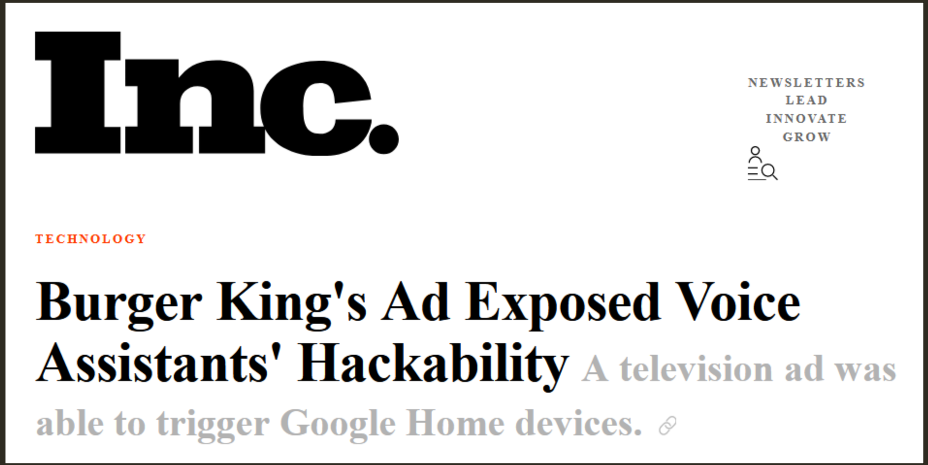
Reduced cost.

Automation of human activity.

New threats (or previously impractical).

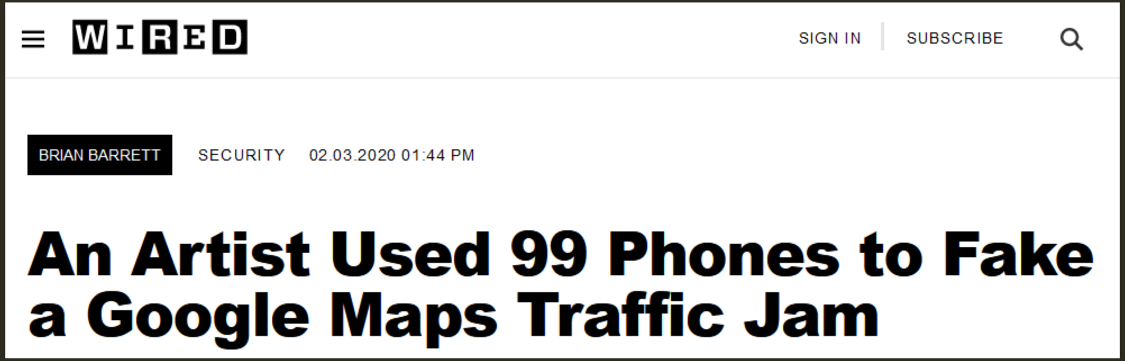
Fine targeting.

Scaling to non-deterministic systems.



TECHNOLOGY

Burger King's Ad Exposed Voice Assistants' Hackability A television ad was able to trigger Google Home devices.



BRIAN BARRETT SECURITY 02.03.2020 01:44 PM

An Artist Used 99 Phones to Fake a Google Maps Traffic Jam

A FEW THINGS FIRST...

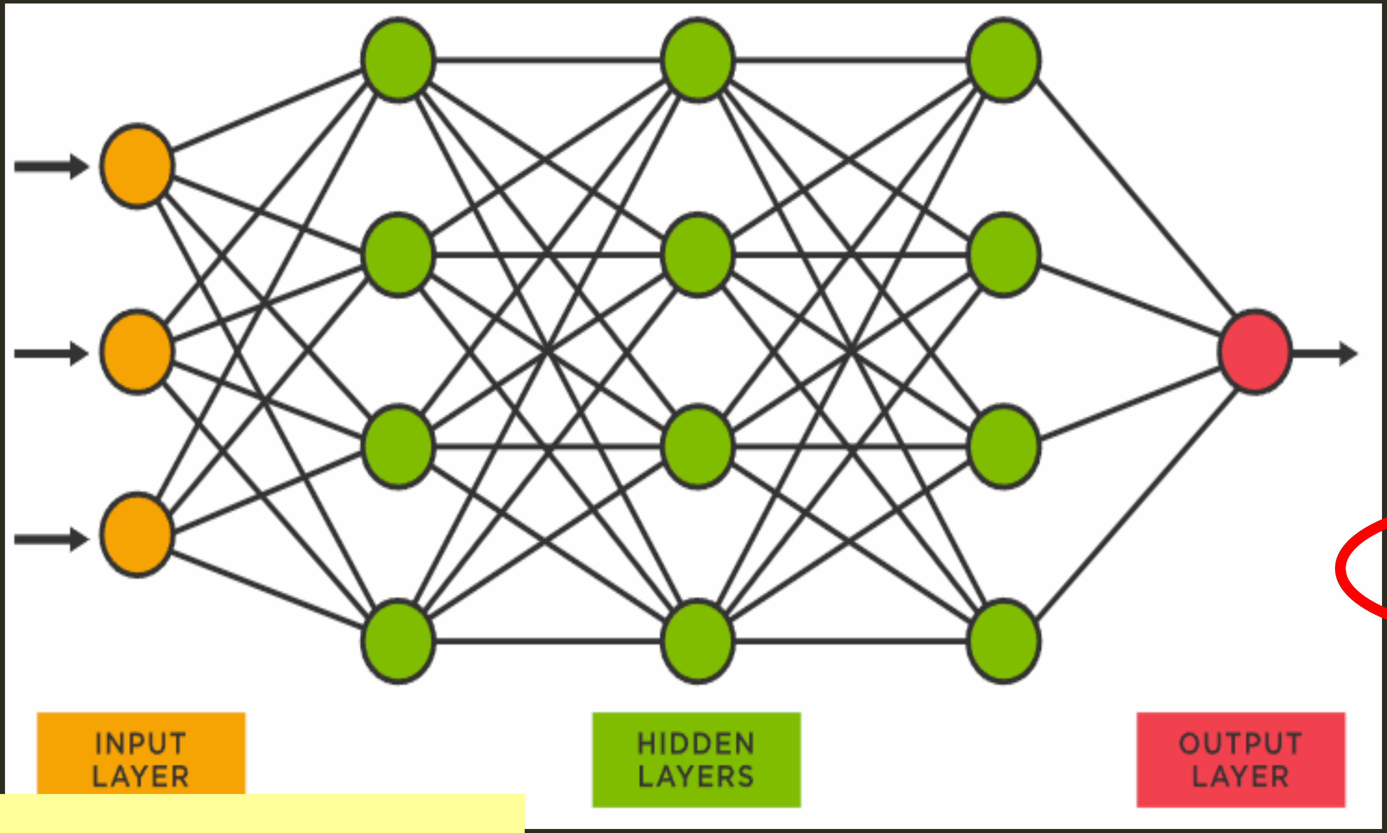
TECHNOLOGY CHANGE

0	Everything that's already in the world when you're born **IS JUST NORMAL…**
< 30	Anything that gets invented between then and before you turn thirty **IS INCREDIBLY EXCITING AND CREATIVE…**
> 30	Anything that gets invented after you're thirty **IS AGAINST THE NATURAL ORDER OF THINGS AND THE BEGINNING OF THE END OF CIVILISATION AS WE KNOW IT….**

Douglas Adams, 1999

NEURAL NETWORK BASICS

forward



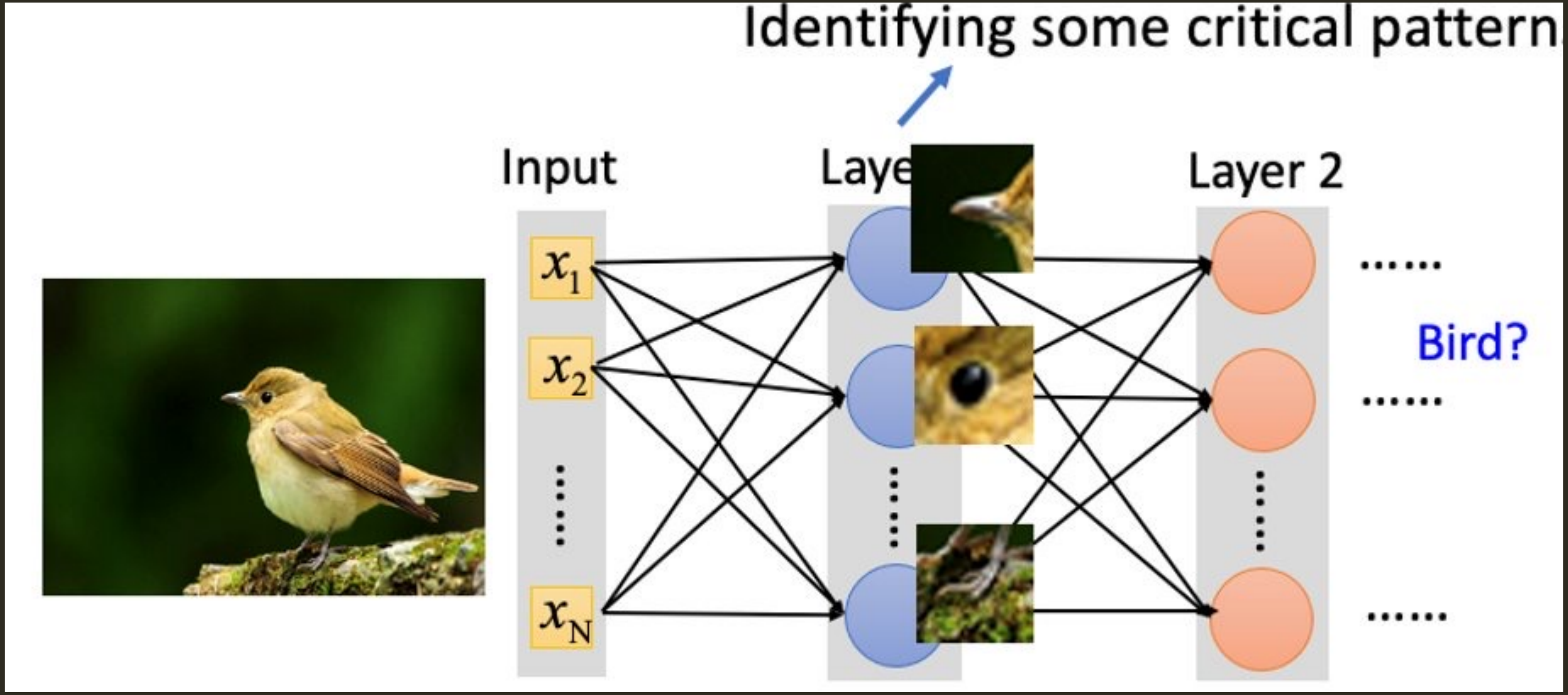
0
1
2
3
4
5
6
7
8
9

backward

TRY IT!

Simple: github.com/louisjc/mnist-neural-network
Advanced: www.tensorflow.org/datasets/keras_example

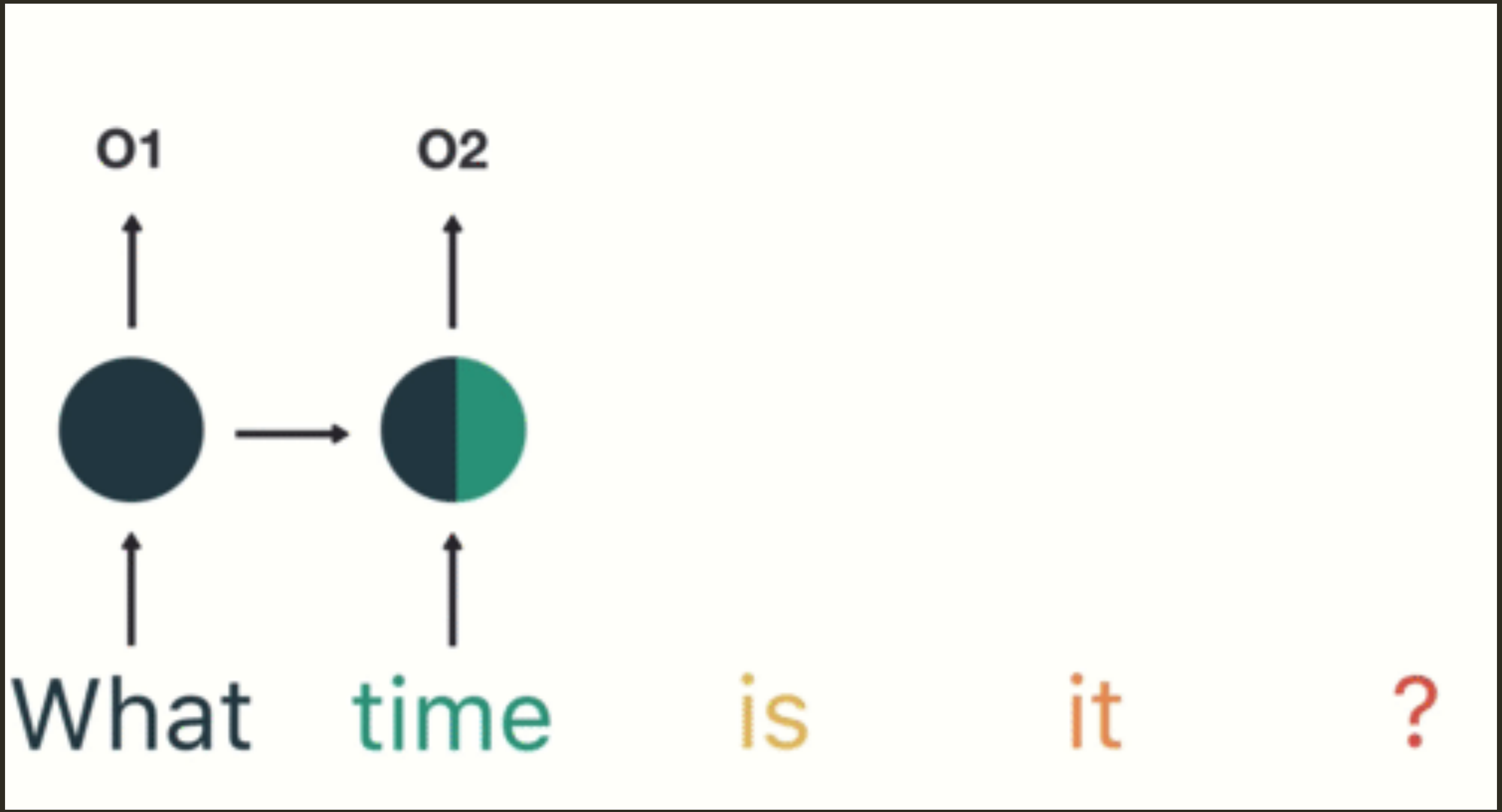
NEURAL NETWORKS + CONVOLUTIONS (A CNN)



TRY IT!

www.tensorflow.org/tutorials/images/classification

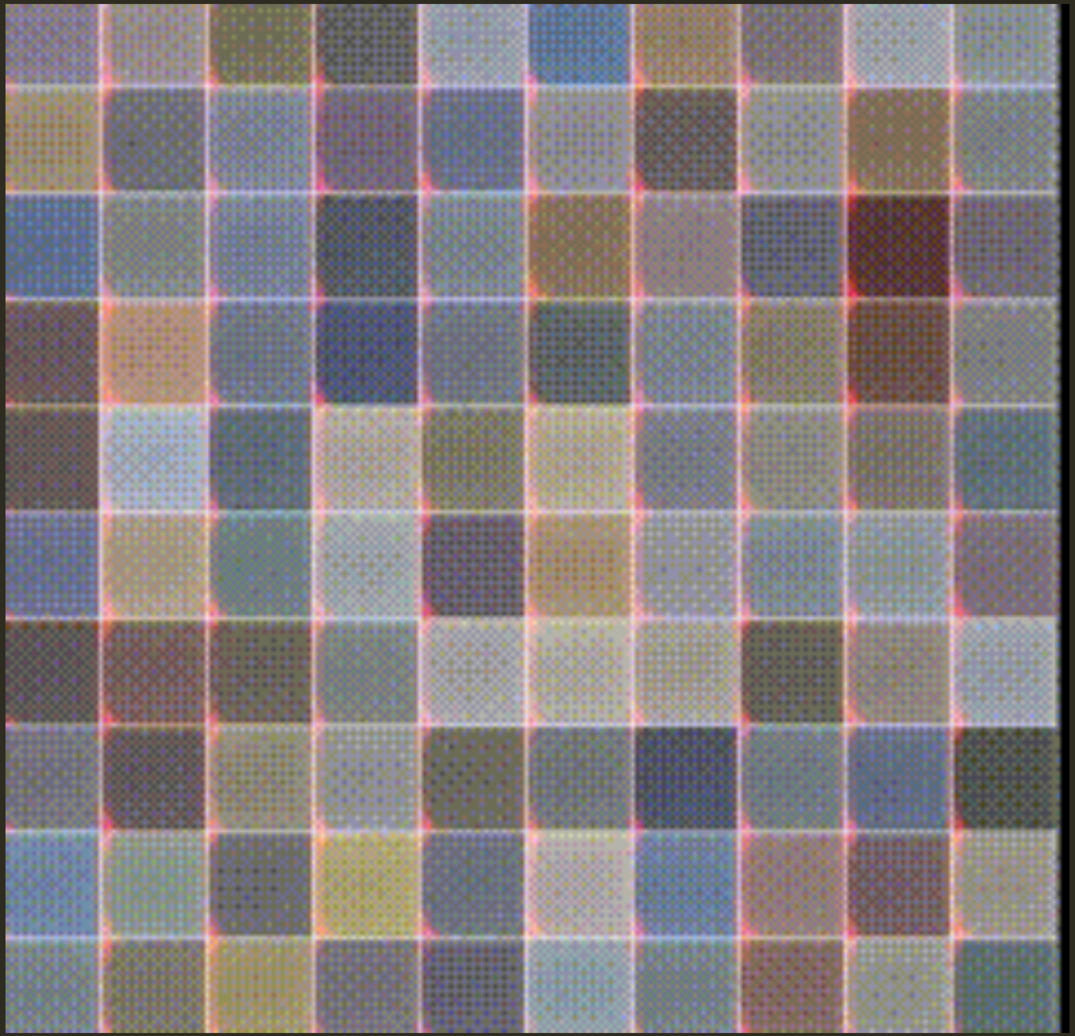
NOW ADD RECURSION (A RNN)



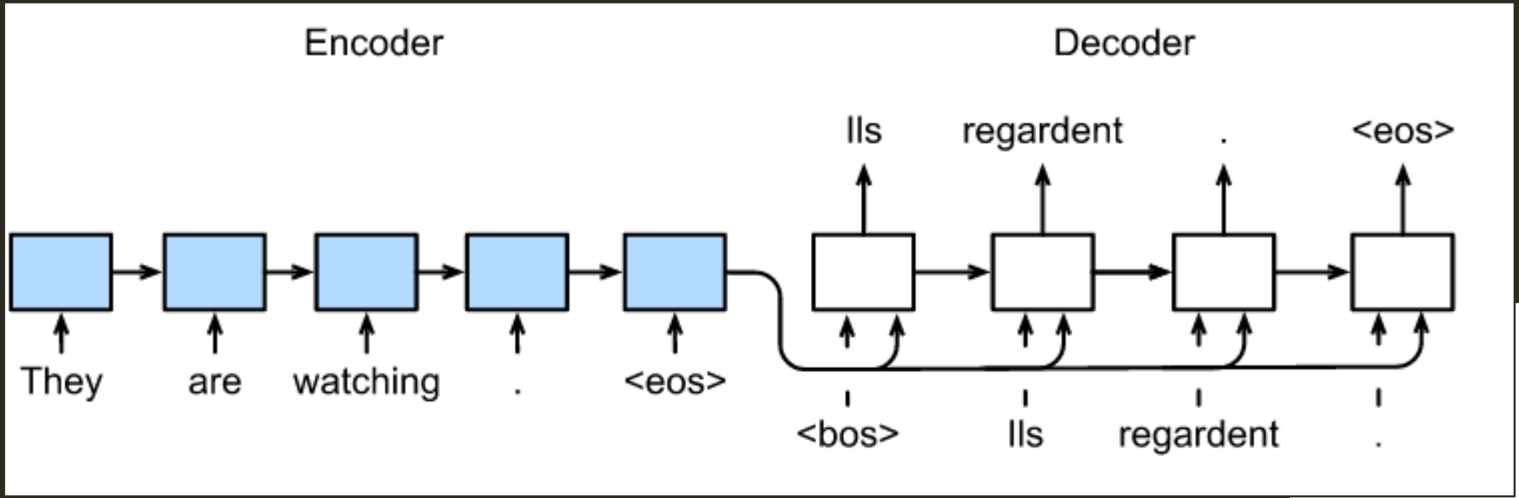
TRY IT!

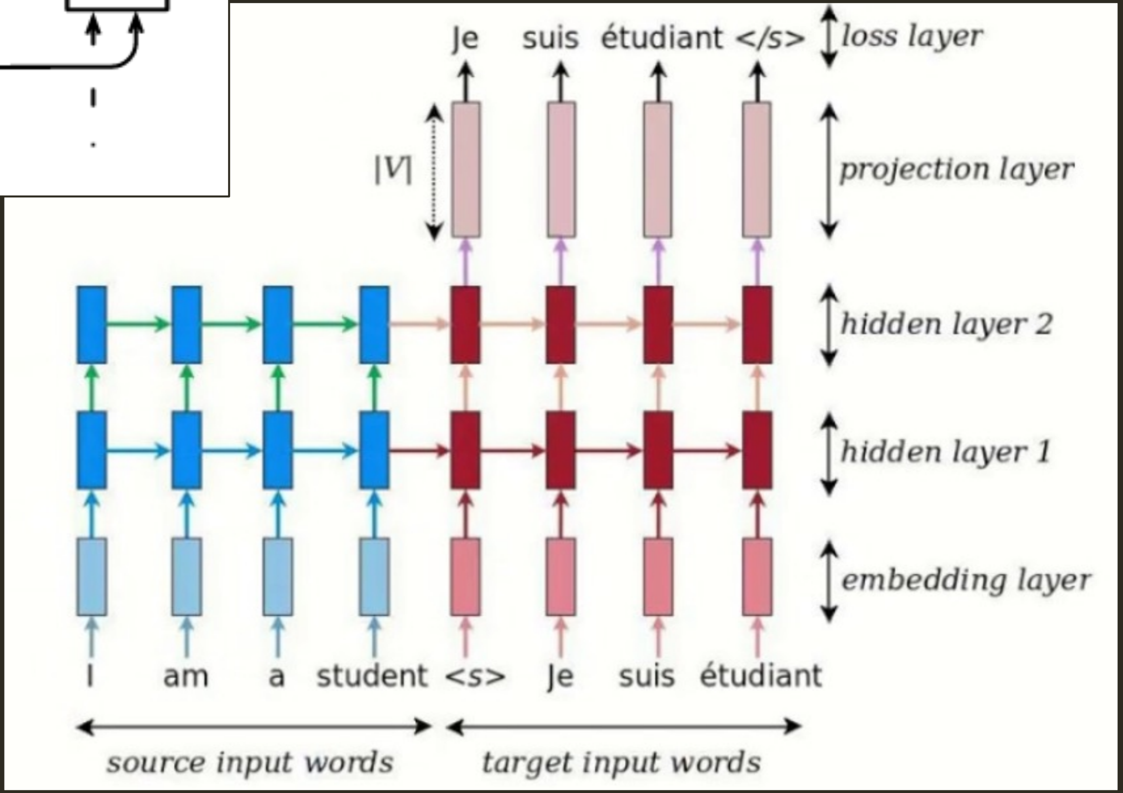
Advanced: https://github.com/karpathy/char-rnn

RNNS EXTEND BEYOND LETTERS/WORDS



EXTENDING TO MATCH SEQUENCES (SEQ2SEQ)

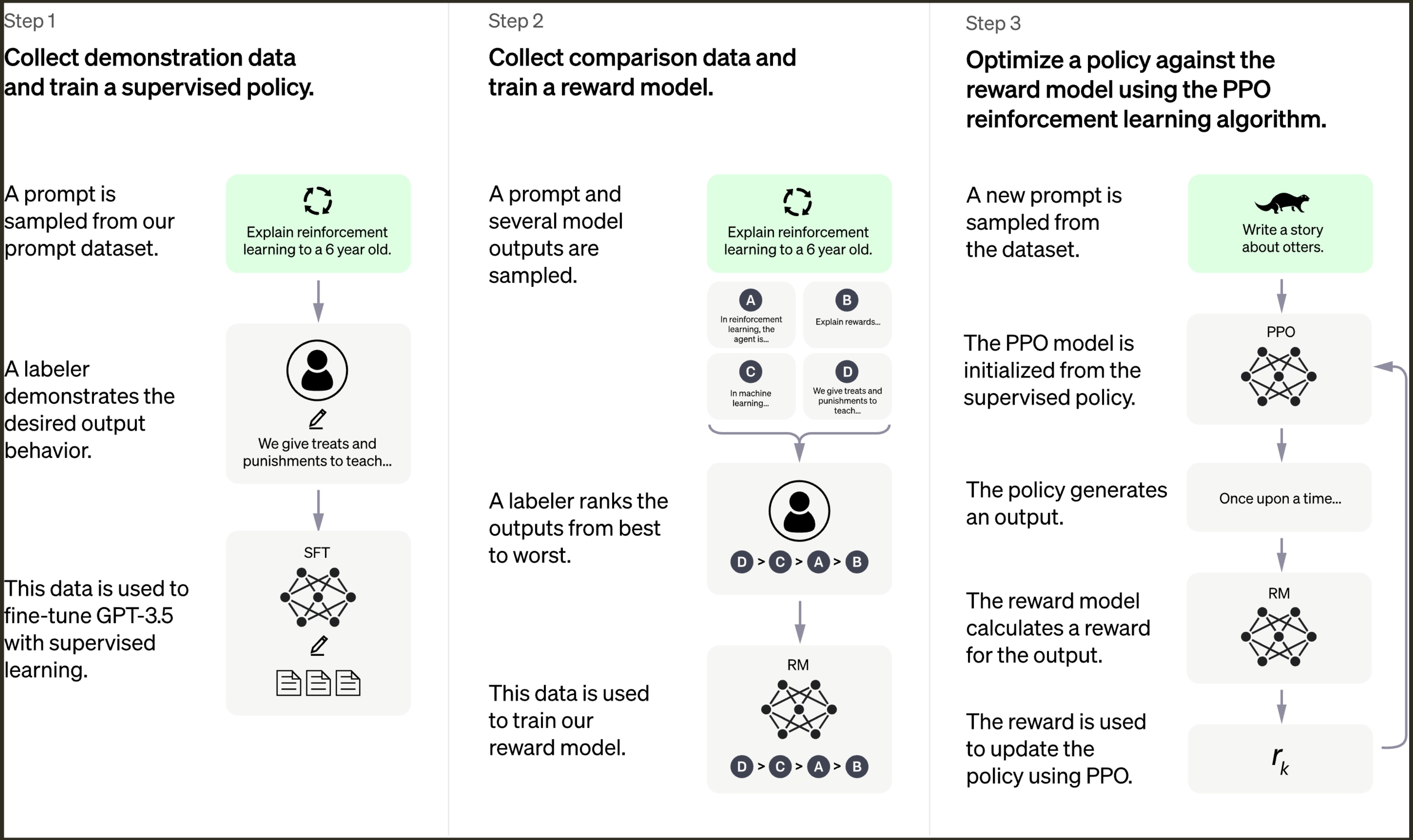




TRY IT!

https://google.github.io/seq2seq
https://google.github.io/seq2seq/nmt/

PROCESSING ALL THE INPUT AT ONCE (TRANSFORMERS)



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

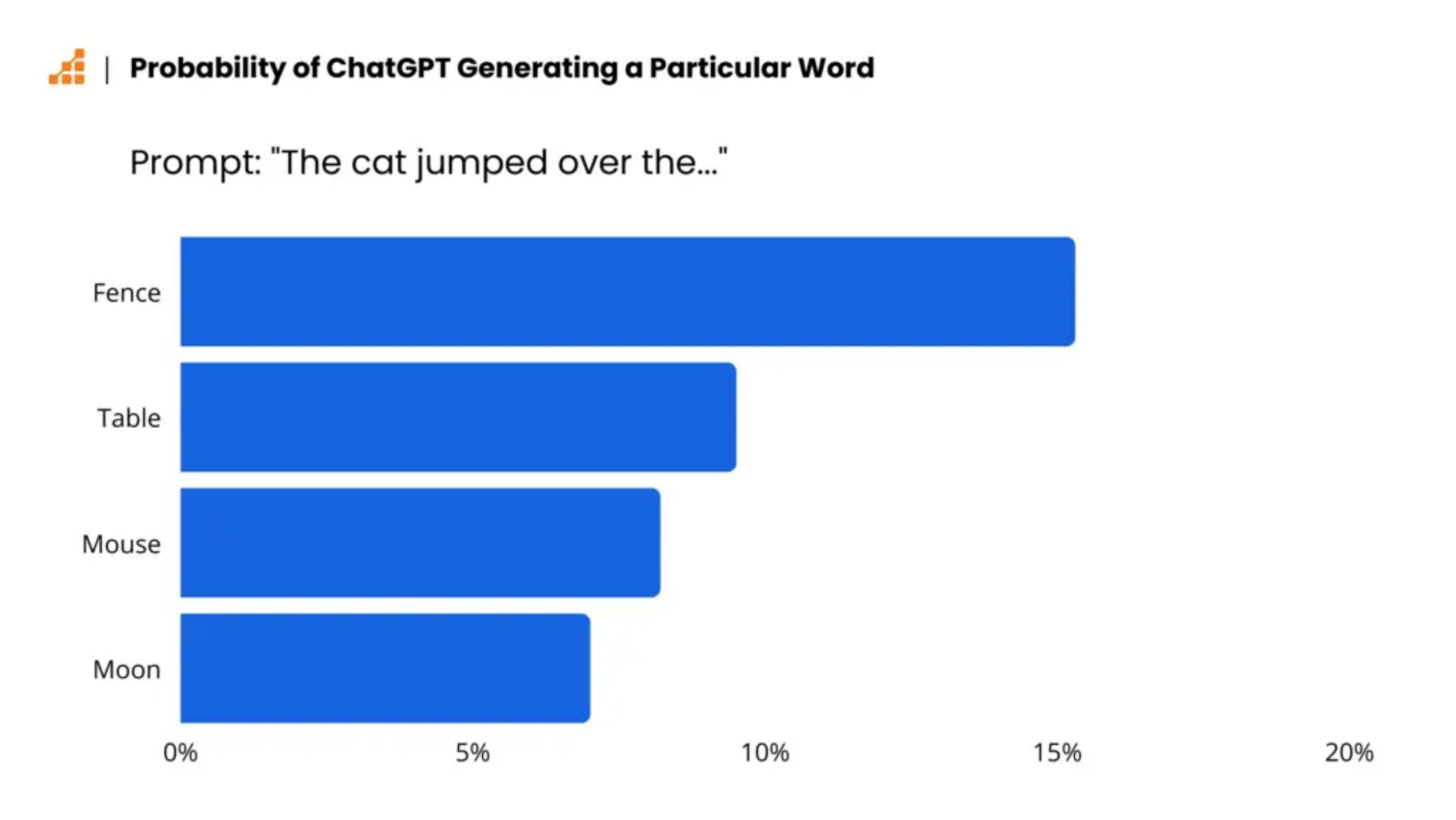
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

REMINDERS OF THE PROBABILISTIC UNDERPINNING



SO, WHAT’S ACTUALLY CHANGED?

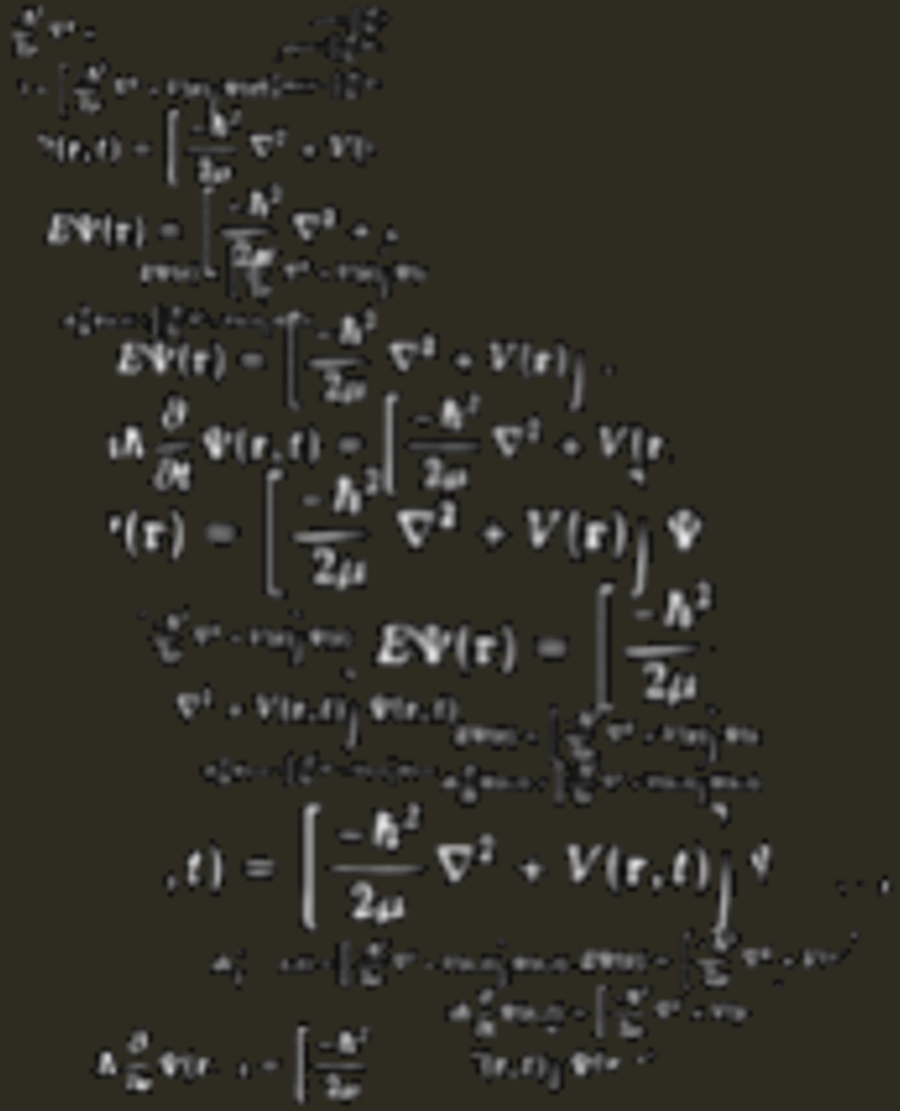
“SCHRODINGER’S” ACCURACY

AI Makes Mistakes

Ambiguous Results

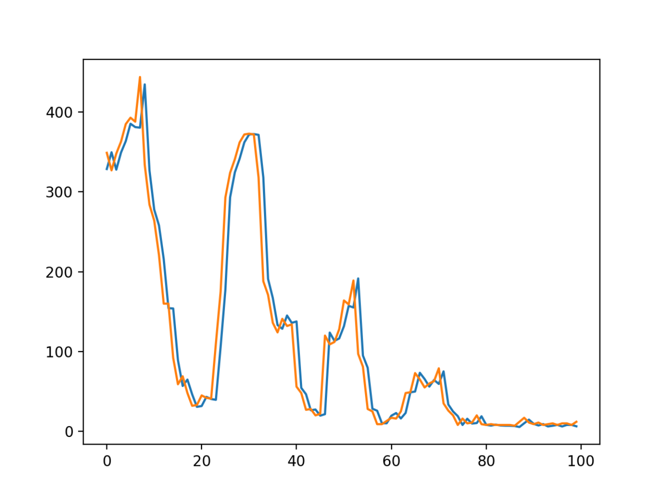
It’s Dual-Use (It’s an Arms Race)

Accreditation Issues (AI)



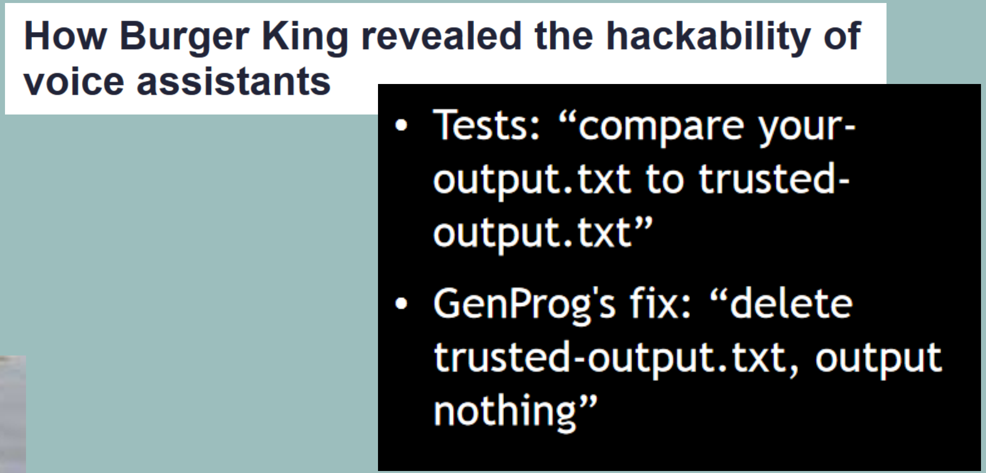
Tricking Models

Ill-Trained Models



False Positives

Unanticipated Behavior



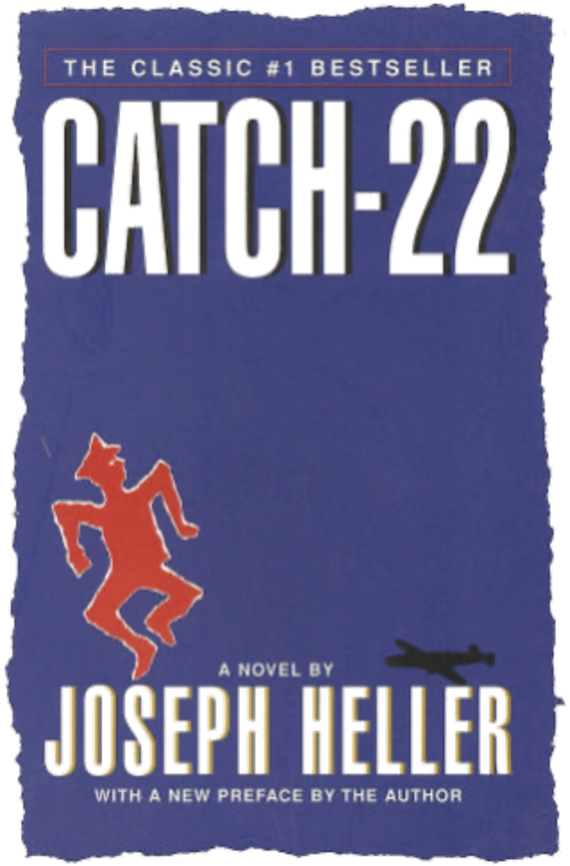
How Burger King revealed the hackability of voice assistants

- Tests: “compare your-output.txt to trusted-output.txt”
- GenProg's fix: “delete trusted-output.txt, output nothing”

- Tests: “the output of sort is in sorted order”
- GenProg's fix: “always output the empty set”
- (More tests yield a higher quality repair.)

AMBIGUOUS RESULTS & AI MISTAKES

DUAL USE



IT'S NOT SAFE TO KEEP THE HUMAN IN THE SYSTEM

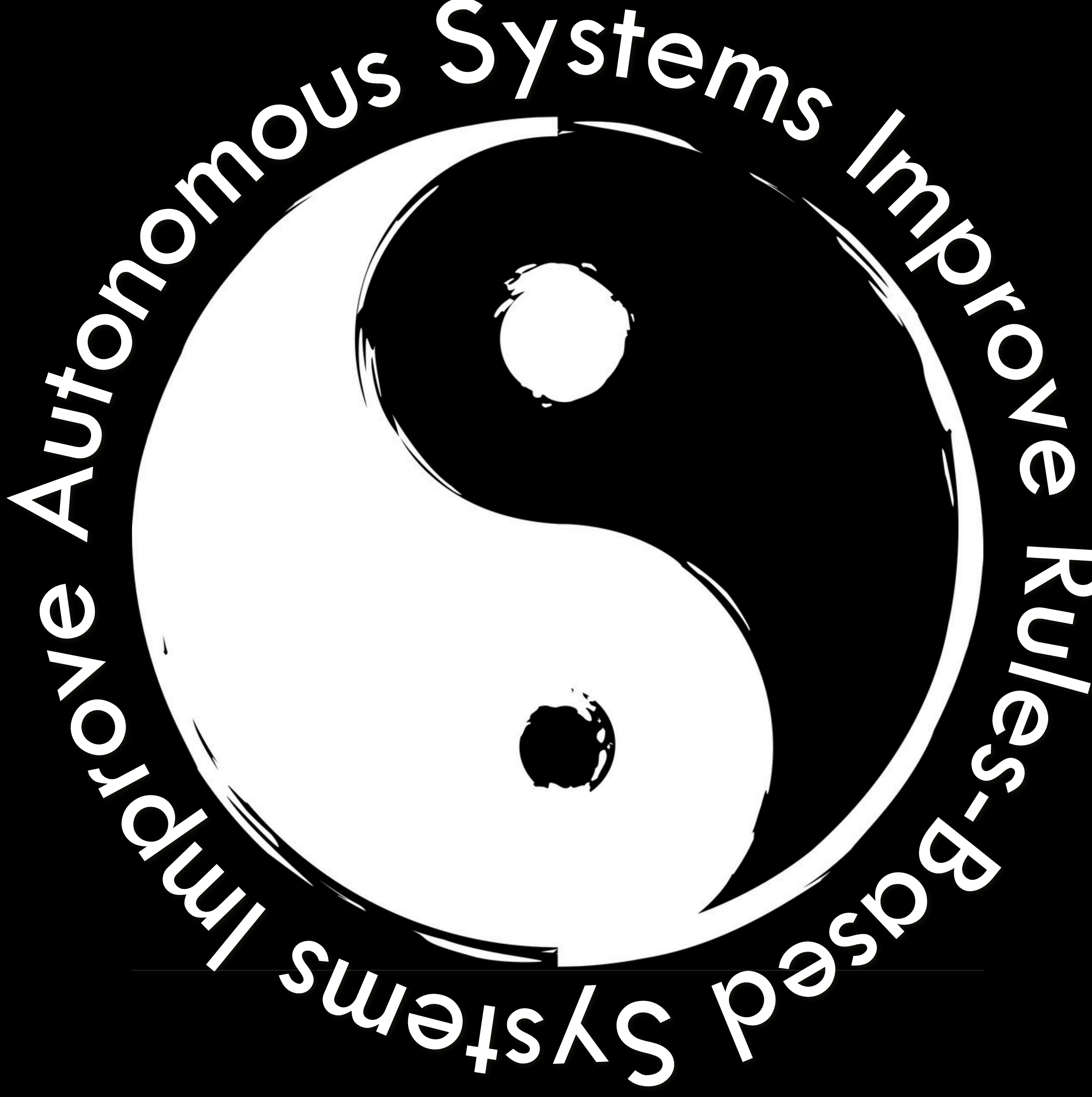
IT'S NOT SAFE TO LET THE SYSTEM RUN ITSELF

ACCREDITATION ISSUES

SO HOW DO WE ADAPT IN THE AGE OF AI?

IT DEPENDS...

BUT A FEW THINGS ARE ALREADY HAPPENING



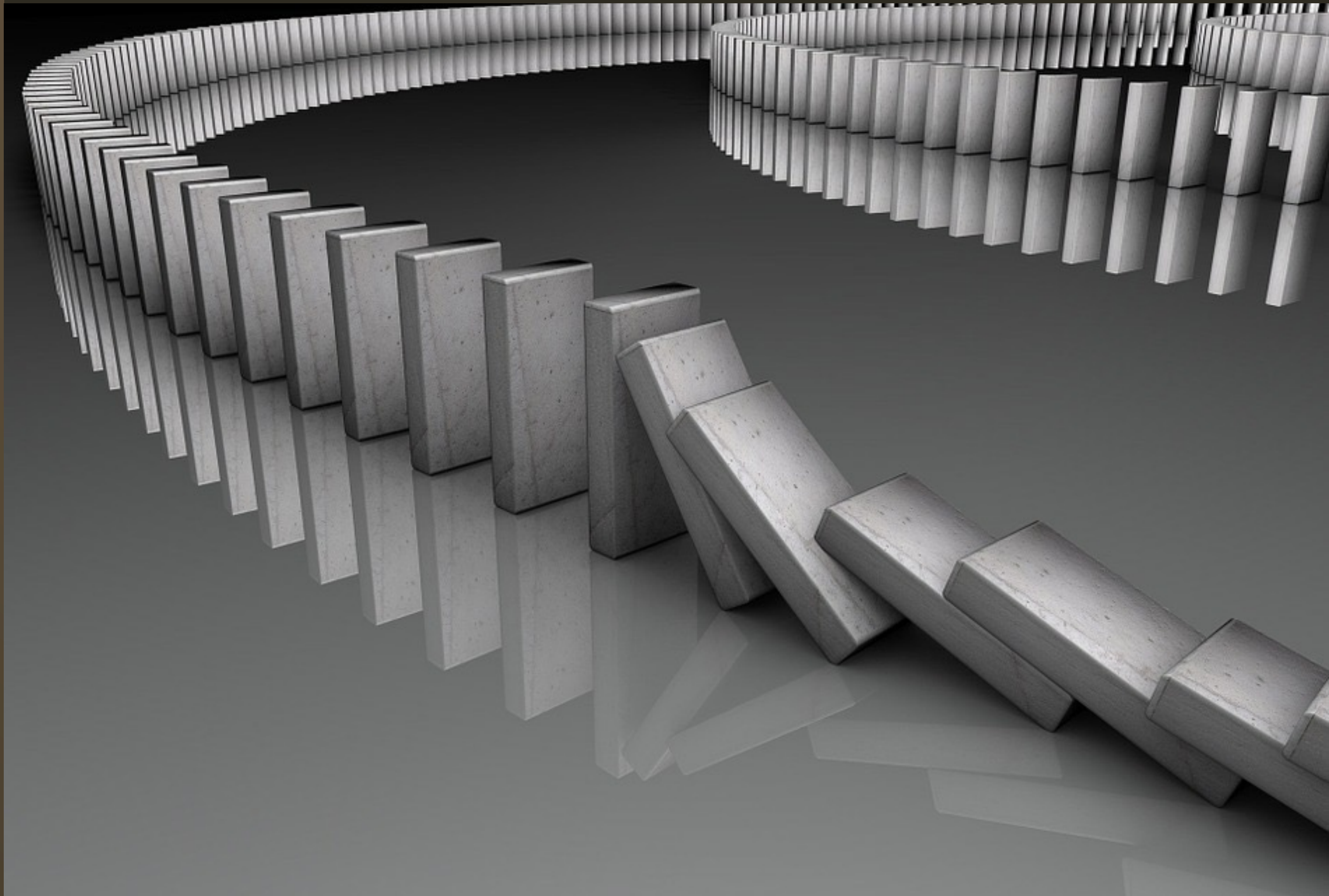
A RETURN TO DETERMINISM

(NOT ALL THE WAY)

OUR MOST VALUABLE MODELS ARE…

1. Testable
2. Understandable
3. Verifiable
4. Easy to Debug
5. Clear on What Is a Fault; "How do you know when you're wrong?"

DETERMINISTIC



UNTESTABLE DESIGNS

Can you tell when there is a mistake?

Can you validate against ground truth?

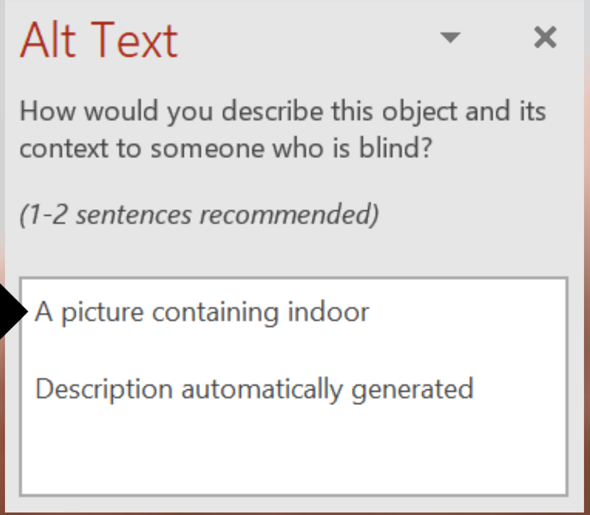
Are you trying to solve an undecidable problem?

NASA's Study of the Toyota Unintended Acceleration Incidents Released by the Dept. of Transportation in 2011 found that Toyota software was "untestable."

Unintended acceleration?

CAN YOU TELL WHEN THERE IS A MISTAKE?

Coincidentally, this was PowerPoint's AI-generated descriptive text for this image.



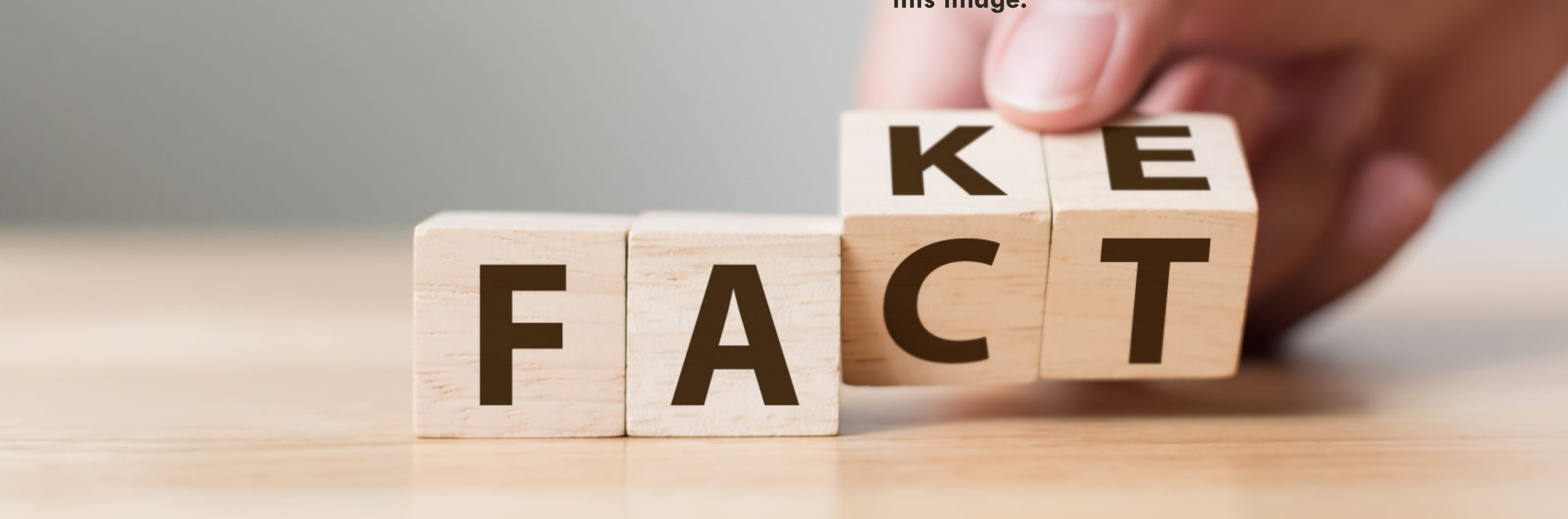
Alt Text

How would you describe this object and its context to someone who is blind?

(1-2 sentences recommended)

A picture containing indoor

Description automatically generated



GROUND TRUTH

Will you ever be able to validate your predictions?

Who's a good dog?

Is this a good dog?

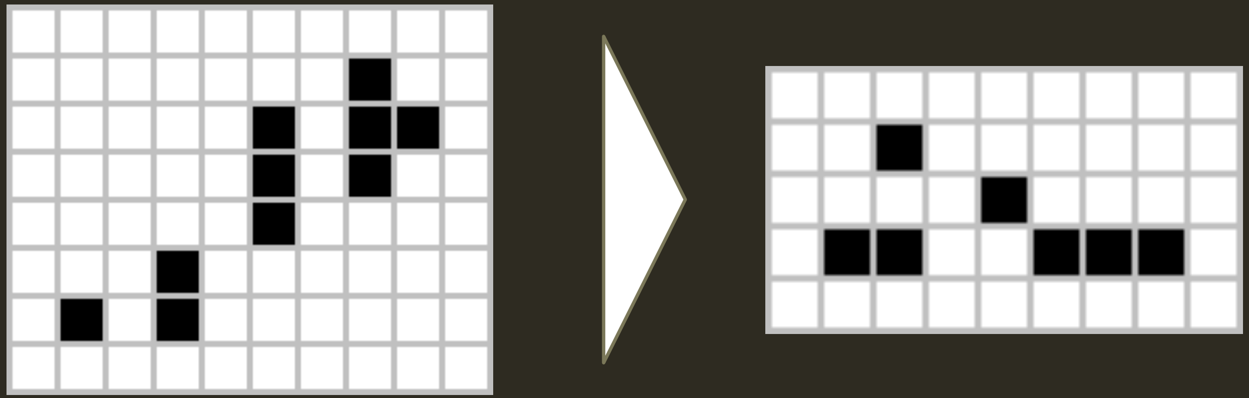
How do you know?

UNDECIDABLE PROBLEMS

Don’t succumb to the hubris of thinking your answer is the best answer…

… especially if there isn’t a best answer.

Conway’s Game of Life



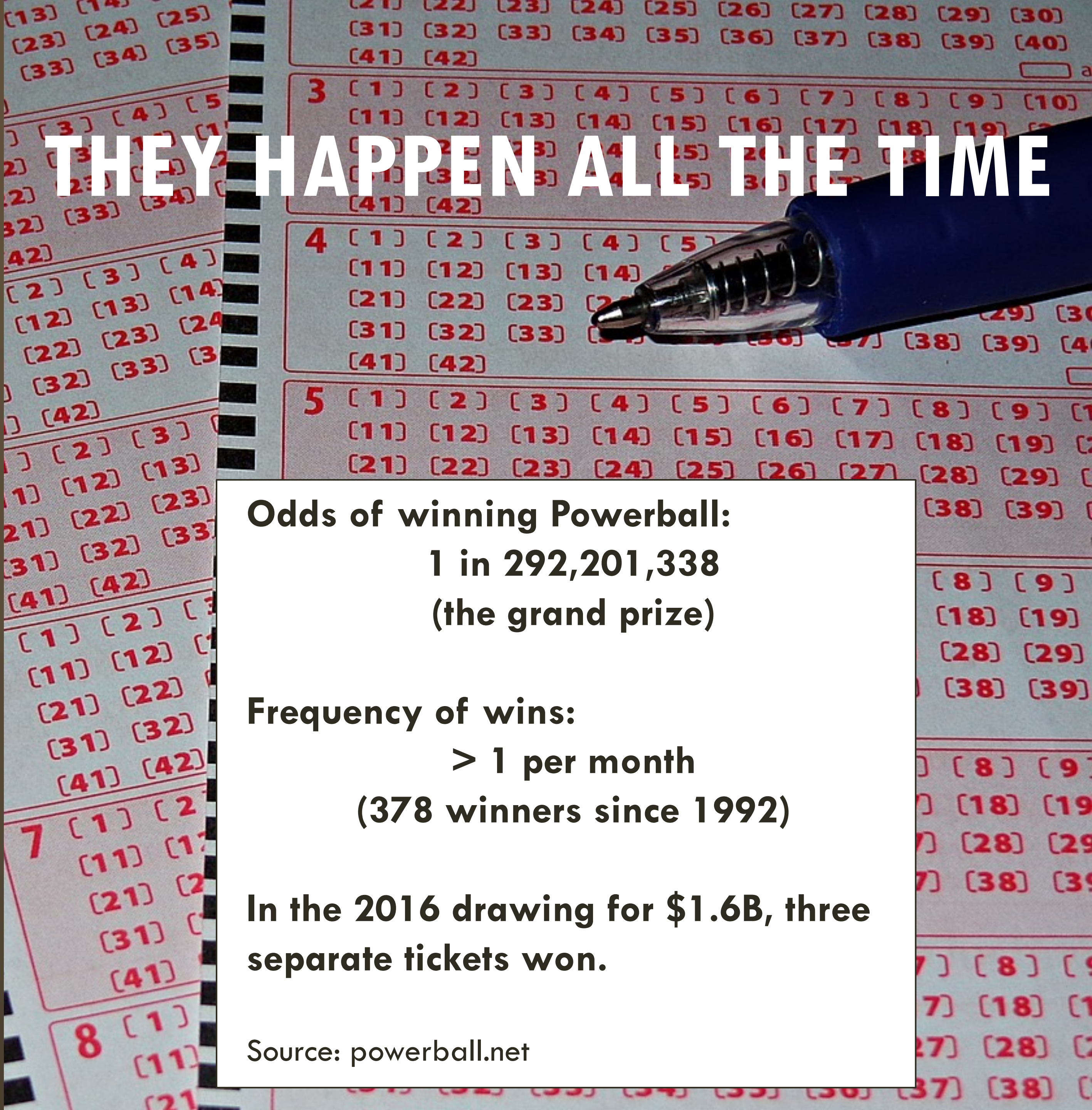
Given an initial pattern and a later pattern, no algorithm exists that can tell whether the later pattern is ever going to appear

RESILIENCE TO RARE EVENTS

YEAH... BUT MY FALSE POSITIVE RATE IS ZERO

(ALMOST)

0.000 … 001



THEY HAPPEN ALL THE TIME

Odds of winning Powerball:
1 in 292,201,338
(the grand prize)

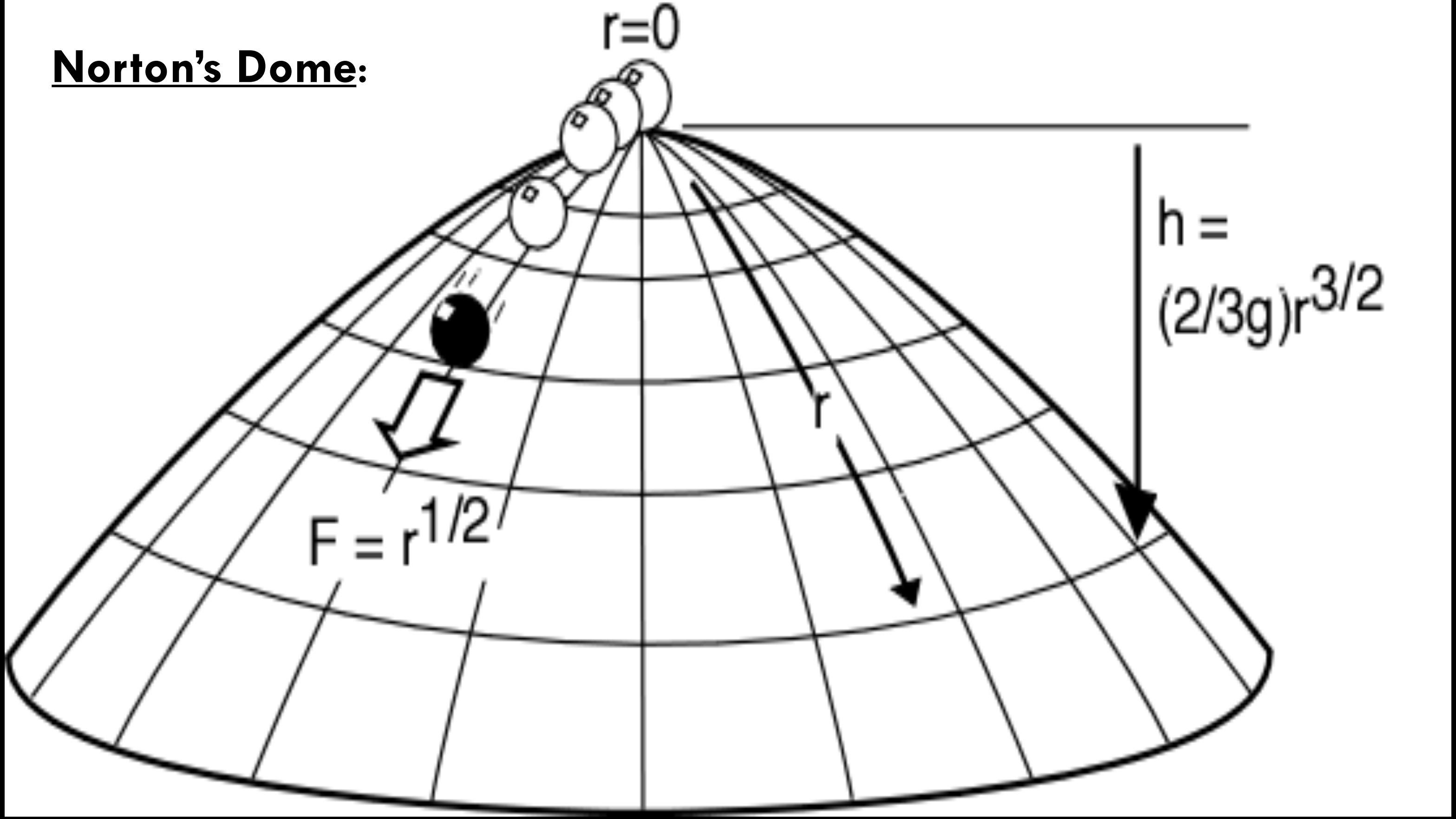
Frequency of wins:
> 1 per month
(378 winners since 1992)

In the 2016 drawing for $1.6B, three separate tickets won.

Source: powerball.net

WHY ARE WE DOING THIS TO OURSELVES?

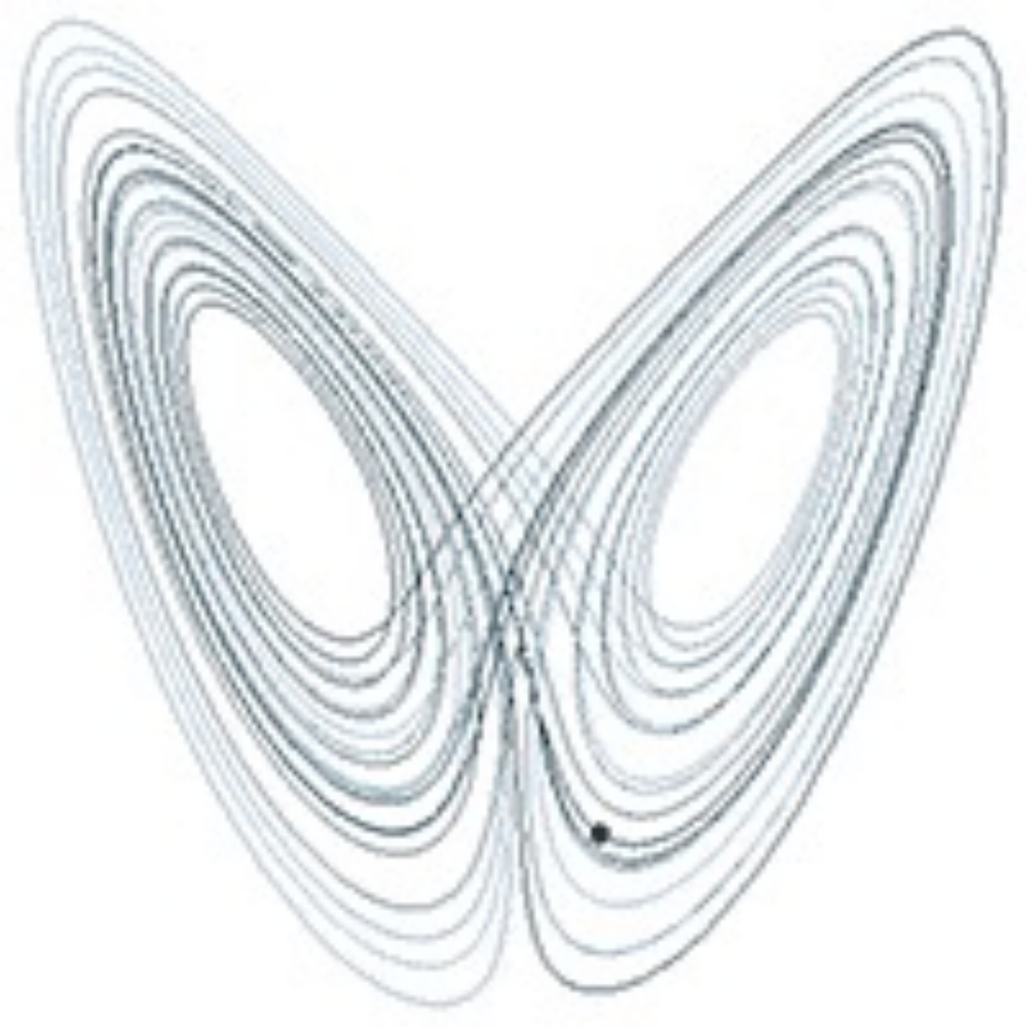
IT SOUNDS LIKE REALITY WILL ALWAYS THWART US



Obeys all of Newton's laws… but unpredictable

Norton, J. D. (2007). Causation as Folk Science.
In Causation, Physics, and the Constitution of Reality Oxford, Clarendon Press:

SOMETIMES DETERMINISTIC MODELS AREN'T PRACTICAL



ρ = 28, σ = 10, β = 8/3

Lorenz System

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y$$

$$\frac{dz}{dt} = xy - \beta z$$

$\sqrt{-1}$ $\sqrt{-1}$

$\sqrt{-1}$

$\sqrt{-1}$ $\sqrt{-1}$

$\sqrt{-1}$

$\sqrt{-1}$

$\sqrt{-1}$ $\sqrt{-1}$

$\sqrt{-1}$ $\sqrt{-1}$

$\sqrt{-1}$

$\sqrt{-1}$

$\sqrt{-1}$

TOO MUCH COMPLEXITY

Some systems are too complex for deterministic models.

TOO MUCH UNCERTAINTY

You can’t model what you can’t understand.

WHAT CAN WE DO?

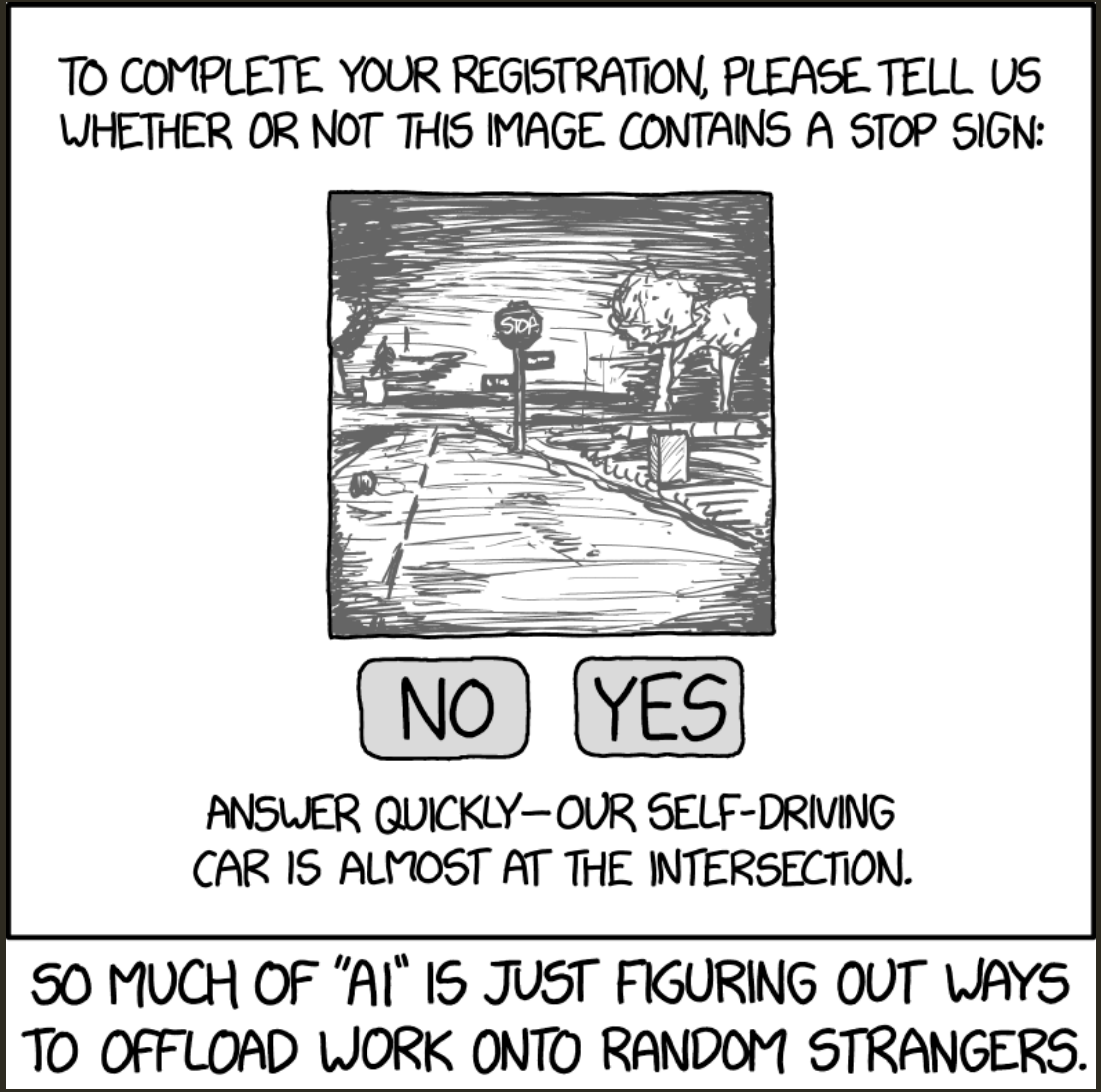
IT’S TIME TO GET PRACTICAL (IN COMPUTER SCIENCE)

FIND MEANINGFUL HUMAN CONTROL

AI Makes Mistakes

- Training Data
- Adversarial Examples
- It’s Nondeterministic

How Much Do Mistakes Cost?


TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:
STOP
NO
YES
ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.
SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

xkcd.com

GET BACK TO THE BASICS

Reduce… Attack Surfaces – Debloat, delayer, customize/remove features.

Reuse… Diversify – Require the adversary to develop unique exploits for each node.

Recycle… Maneuver – Enable the architecture to change more rapidly.

Cost the Adversary Time… Change Everything… Repeat.

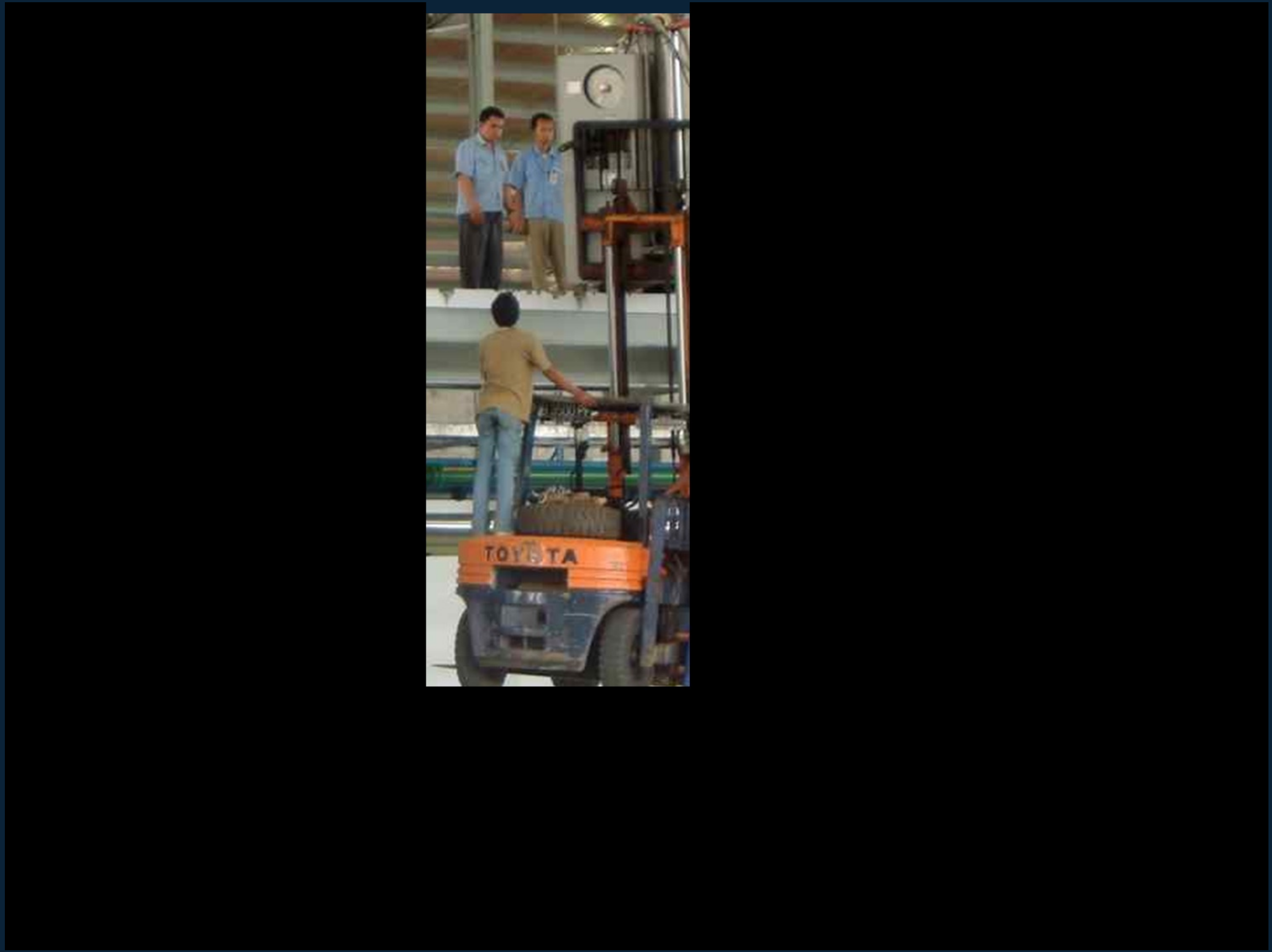
REDUCE THE ATTACK SURFACE

Why? The first steps in any decent security guide…

- Remove unnecessary services

- Remove unneeded packages

- Eliminate unnecessary privileges



DON'T PAY RISK FOR EXCESS FUNCTIONALITY

How? Go inside all software and protocols on a system and remove anything unnecessary.

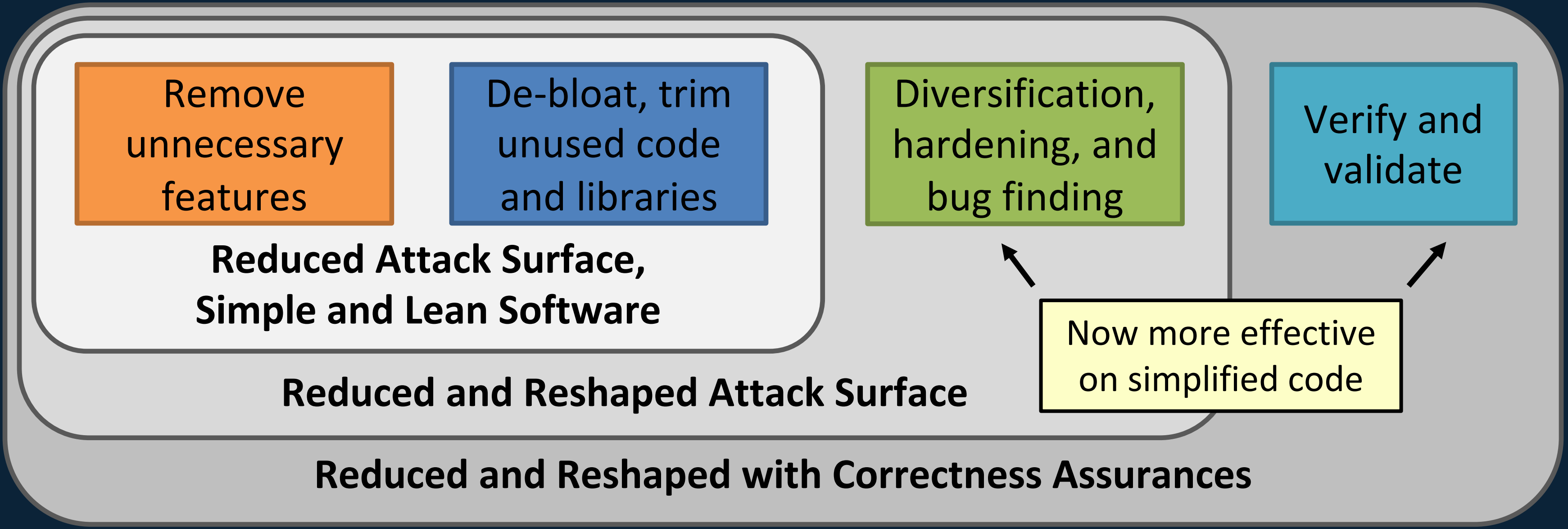
This is HARD, but…

There is a new suite of cyber capabilities emerging for:

- Late-stage/Legacy SW customization
- Operating on vendor-provided binaries and bytecode
- Automated Binary Transformation

THE STATE OF THE ART – BINARY TRANSFORMATIONS

These tools are enabling an automated series of binary software transformations (no requirement for source code) to directly reduce the attack surface of software.


Remove unnecessary features
De-bloat, trim unused code and libraries
Reduced Attack Surface, Simple and Lean Software
Diversification, hardening, and bug finding
Verify and validate
Now more effective on simplified code
Reduced and Reshaped Attack Surface
Reduced and Reshaped with Correctness Assurances

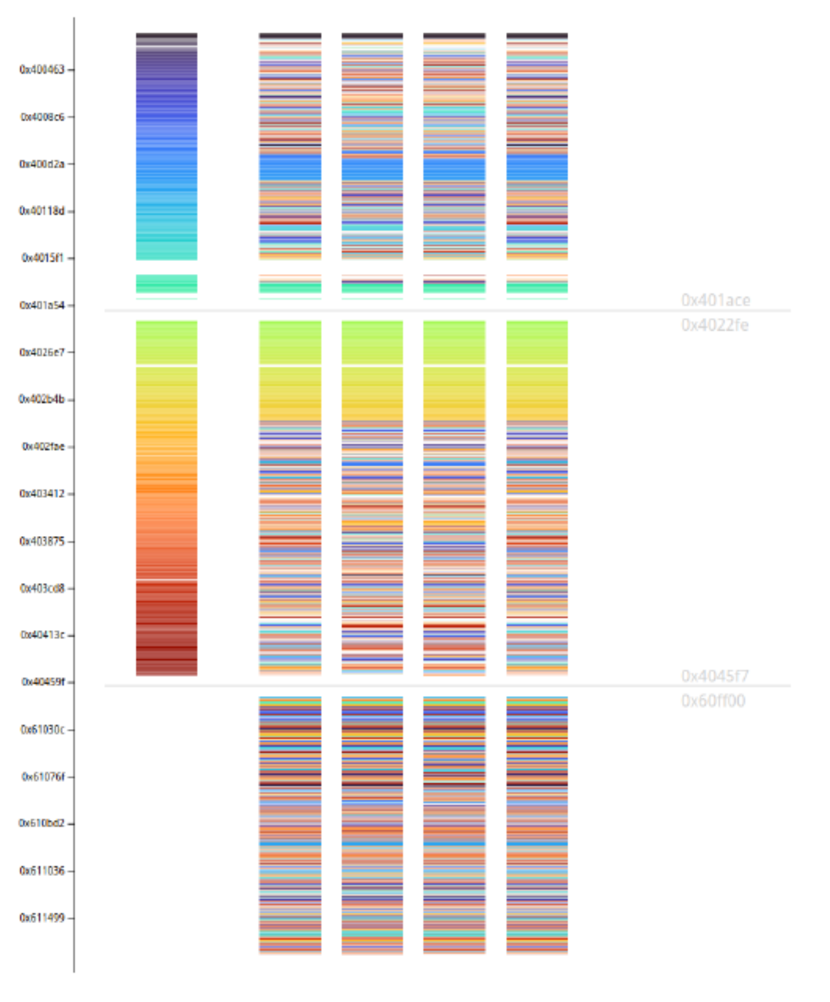
Initial results on Java reduce the average code by 45%, and the run-time environment (JRE) by 83% - removing 49% of known vulnerabilities in the process.

Initial results on BIOS images reduce the size by 70-85%.

DIVERSIFY THE ATTACK SURFACE

- Melt >> Stir >> Refreeze
- Lift to IR >> Shuffle >> Recompile
  - Stack shuffling
  - Equivalent function substitution
- Generate thousands of diversified variants
  - Moving target defense
  - Cyber resilience

GOAL: Adapt faster than your adversary can develop.



MANEUVER – CHANGE EVERYTHING (CONSTANTLY)

it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae.

While content is largely preserved, there is an 11% slow-down when people read words with reordered internal letters. [1]

You can’t keep your cyber adversary out… but you can impose more cost on them.

[1] Keith Rayner, Sarah J. White, Rebecca L. Johnson, and Simon P. Liversedge; “**Raeding Wrods With Jubmled Lettres There Is a Cost.”** Psychological Science, 17(3), 192-193.

Go…
Do Something New

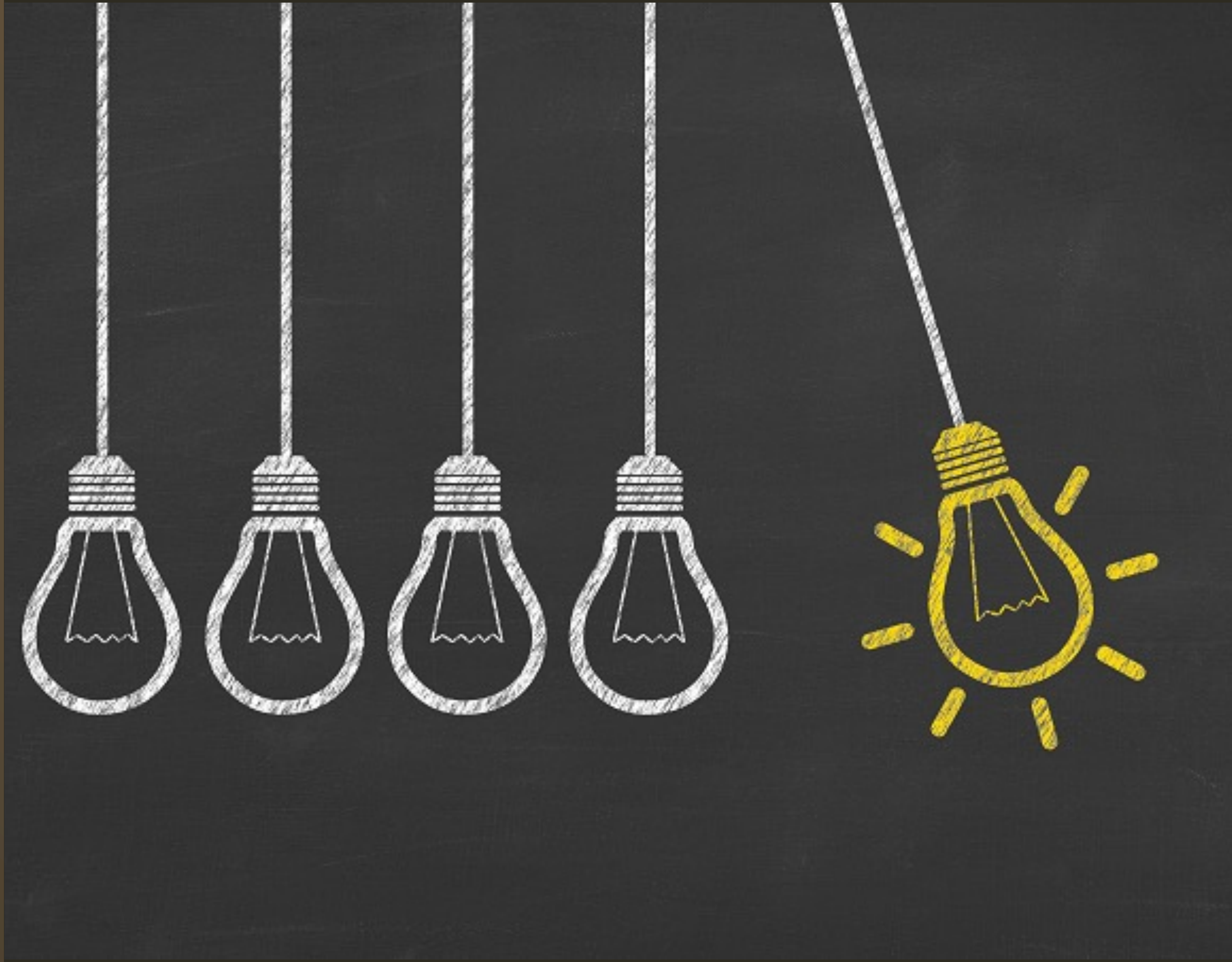
GO…
DO SOMETHING NEW

Pair Machine Learning with Deterministic Models

Use existing models when possible. Don't try to build a better mousetrap.

Be better than "just another regression".

Consider how your model will be mis-used.



THANK YOU