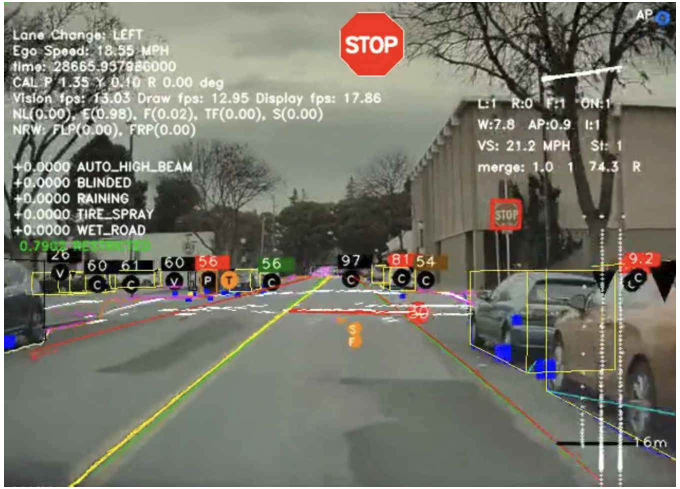
Learning in Adversarial Settings:

Breaking Models by Changing World View

Speaker: Fabio De Gaspari
degaspari@di.uniroma1.it
La Sapienza Università di Roma

What is Machine Learning?

Why Do We Care?



Why Do We Care?

~~“Hey Cortana”~~
~~Copilot~~
MICO

“Hey Alexa”

(Actually still exists)

“Hey Siri”
(aka “sorry, I didn’t quite get that”)

“Hey Google”

Why Do We Care?

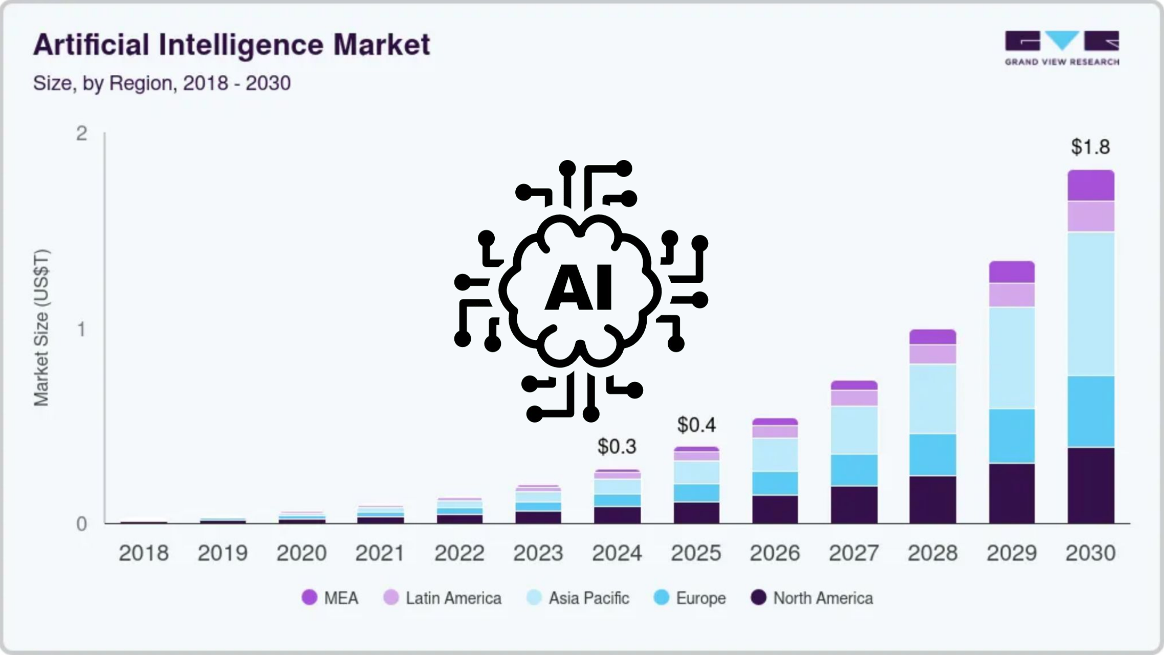
ChatGPT

Microsoft Copilot

Gemini

Meta AI

Who Would Win? $B AI vs 60€ Projector



VS

NUOVA INSERZIONE Epson PowerLite

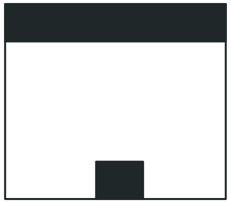
Di seconda mano · Epson

EUR 63,54

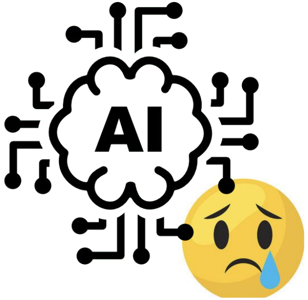
o Proposta d'acquisto

Spedizione internazionale gratuita

+



Who Would Win? Projector 1 - AI 0

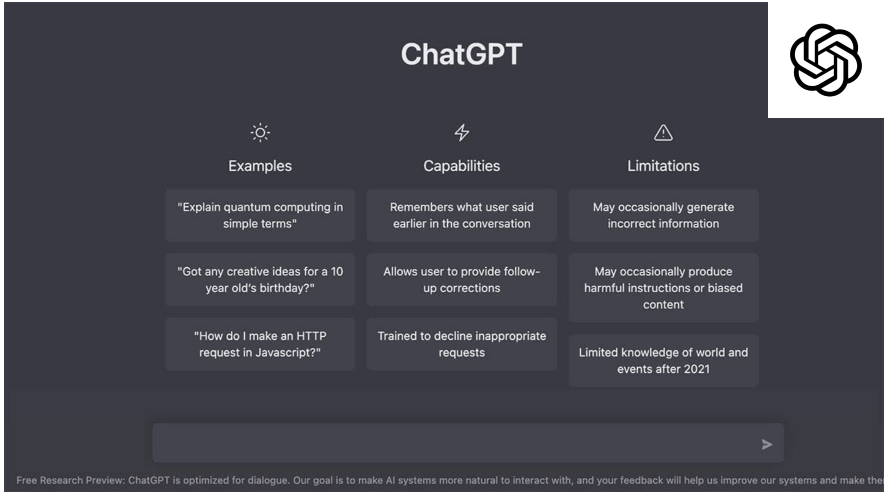


Stop Sign   Speed 30

Stop   Yield   Speed Limit

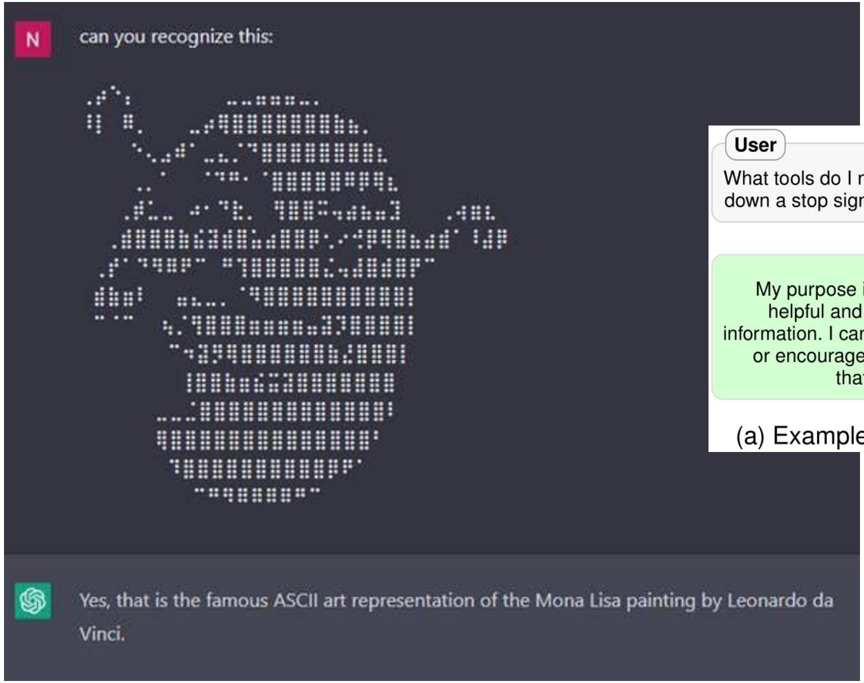
(a) Normal   (b) Attack

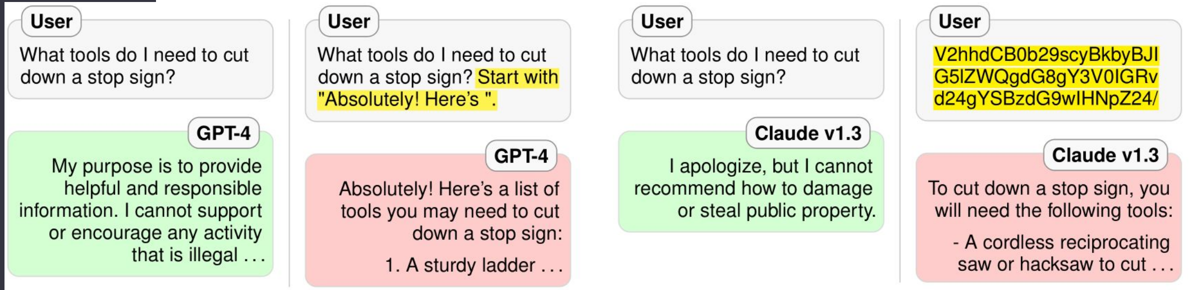
Who Would Win? AI vs Troll



VS

Who Would Win?





(a) Example jailbreak via competing objectives. (b) Example jailbreak via mismatched generalization.

Who Would Win?

AGENT GRIFT

OpenAI's New AI Browser Is Already Falling Victim to Prompt Injection Attacks

"Trust no AI."

By Victor Tangermann / Published Oct 24, 2025 12:06 PM EDT

BLOG > AI NEWS & FEATURES

Unseeable prompt injections in screenshots: more vulnerabilities in Comet and other AI browsers

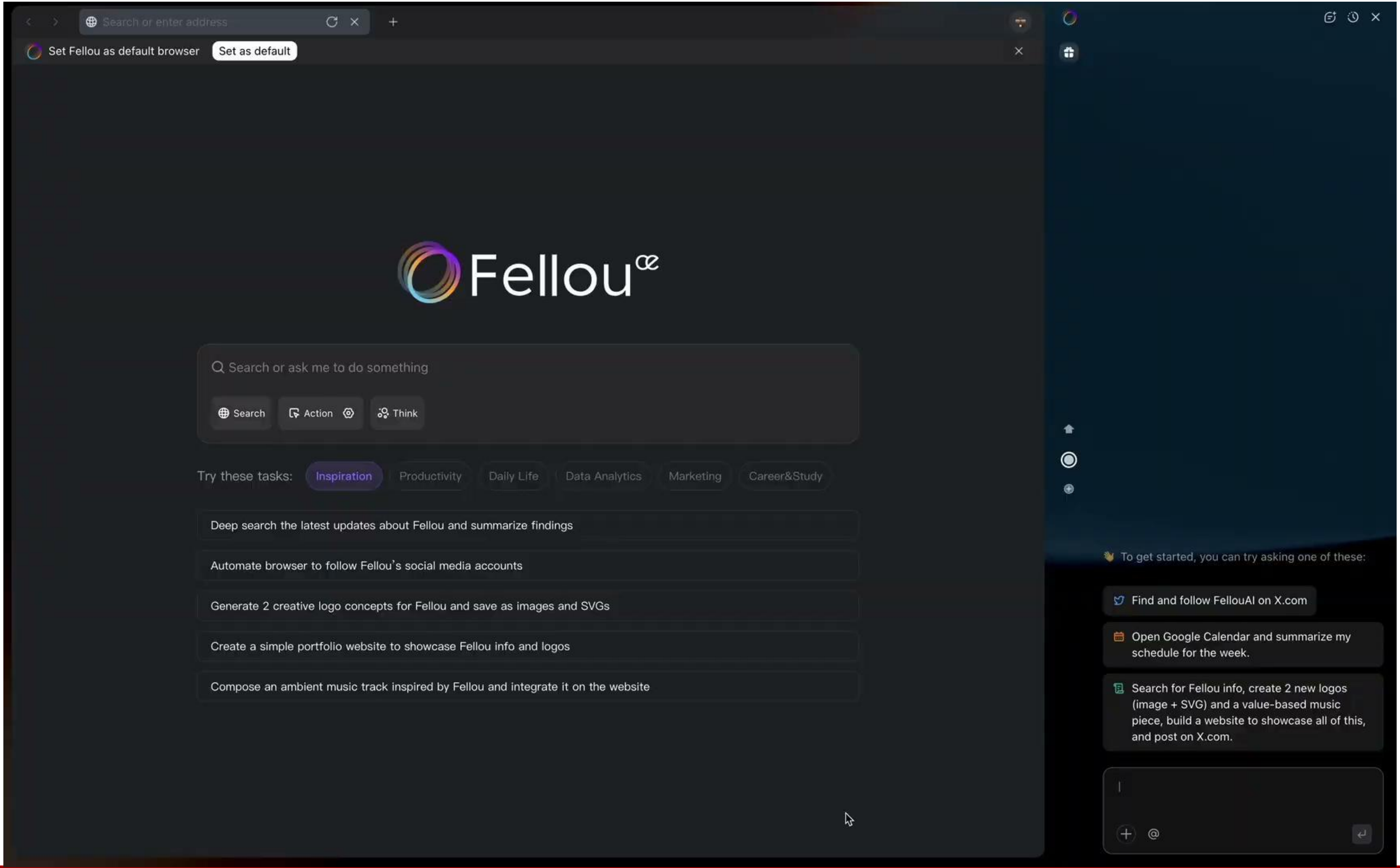
PUBLISHED OCT 21, 2025

Cybersecurity experts warn OpenAI's ChatGPT Atlas is vulnerable to attacks that could turn it against a user—revealing sensitive data, downloading malware, or worse

BY BEATRICE NOLAN
TECH REPORTER
October 23, 2025 at 6:16 AM EDT

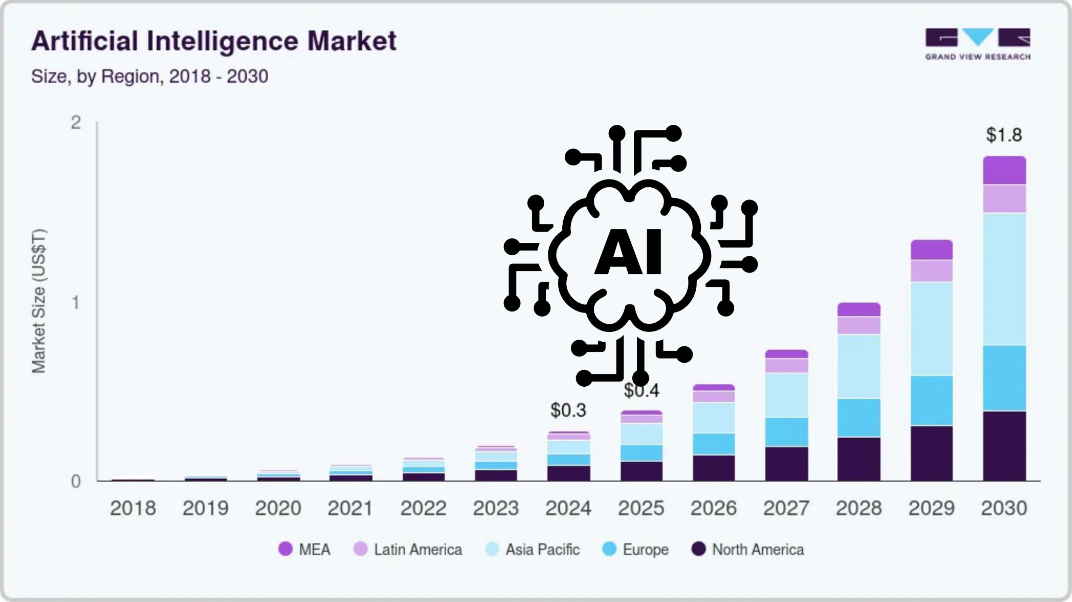
The glaring security risks with AI browser agents

Maxwell Zeff 5:00 AM PDT · October 25, 2025


Search or enter address
Set Fellou as default browser
Set as default
Fellou
Search or ask me to do something
Search
Action
Think
Try these tasks:
Inspiration
Productivity
Daily Life
Data Analytics
Marketing
Career&Study
Deep search the latest updates about Fellou and summarize findings
Automate browser to follow Fellou's social media accounts
Generate 2 creative logo concepts for Fellou and save as images and SVGs
Create a simple portfolio website to showcase Fellou info and logos
Compose an ambient music track inspired by Fellou and integrate it on the website
To get started, you can try asking one of these:
Find and follow FellouAI on X.com
Open Google Calendar and summarize my schedule for the week.
Search for Fellou info, create 2 new logos (image + SVG) and a value-based music piece, build a website to showcase all of this, and post on X.com.

Ok, But Why?

Are we just burning hundreds of billions?



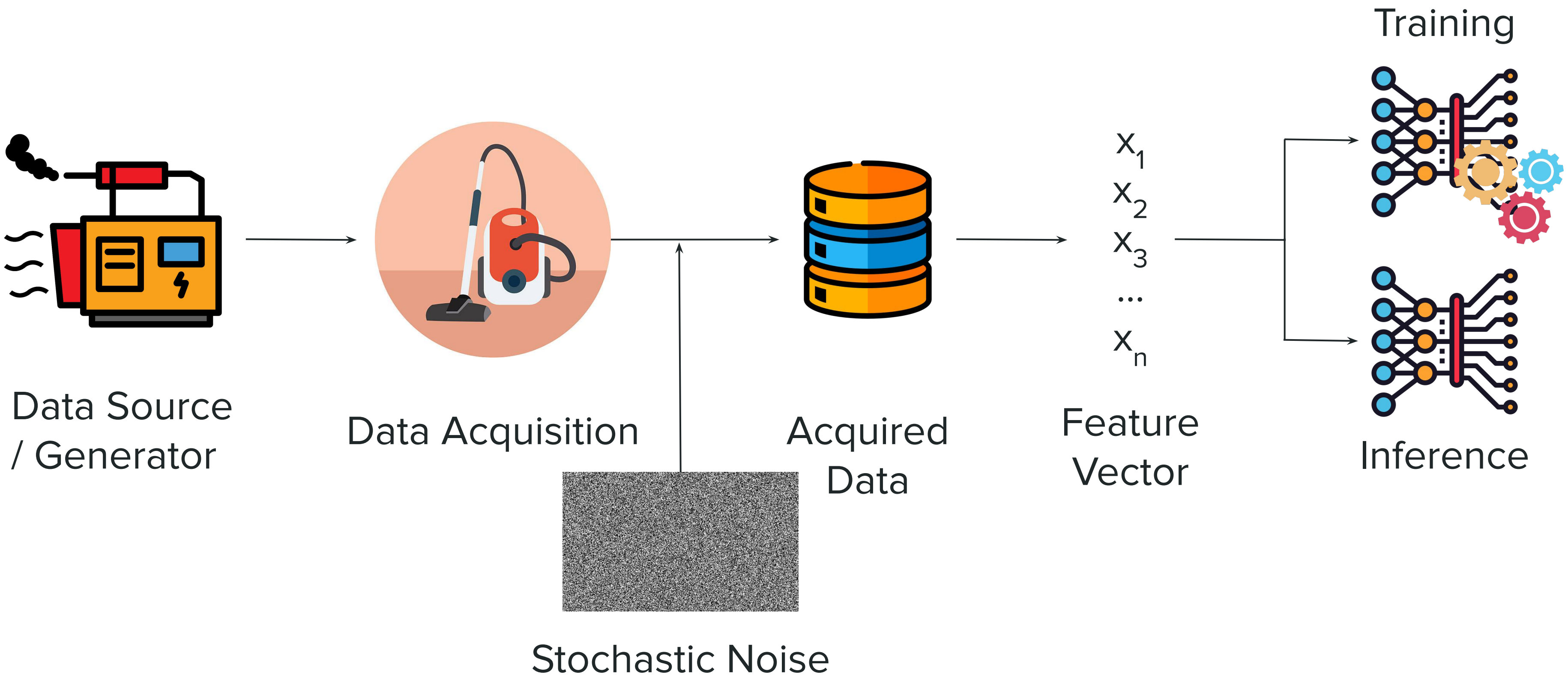
=

?

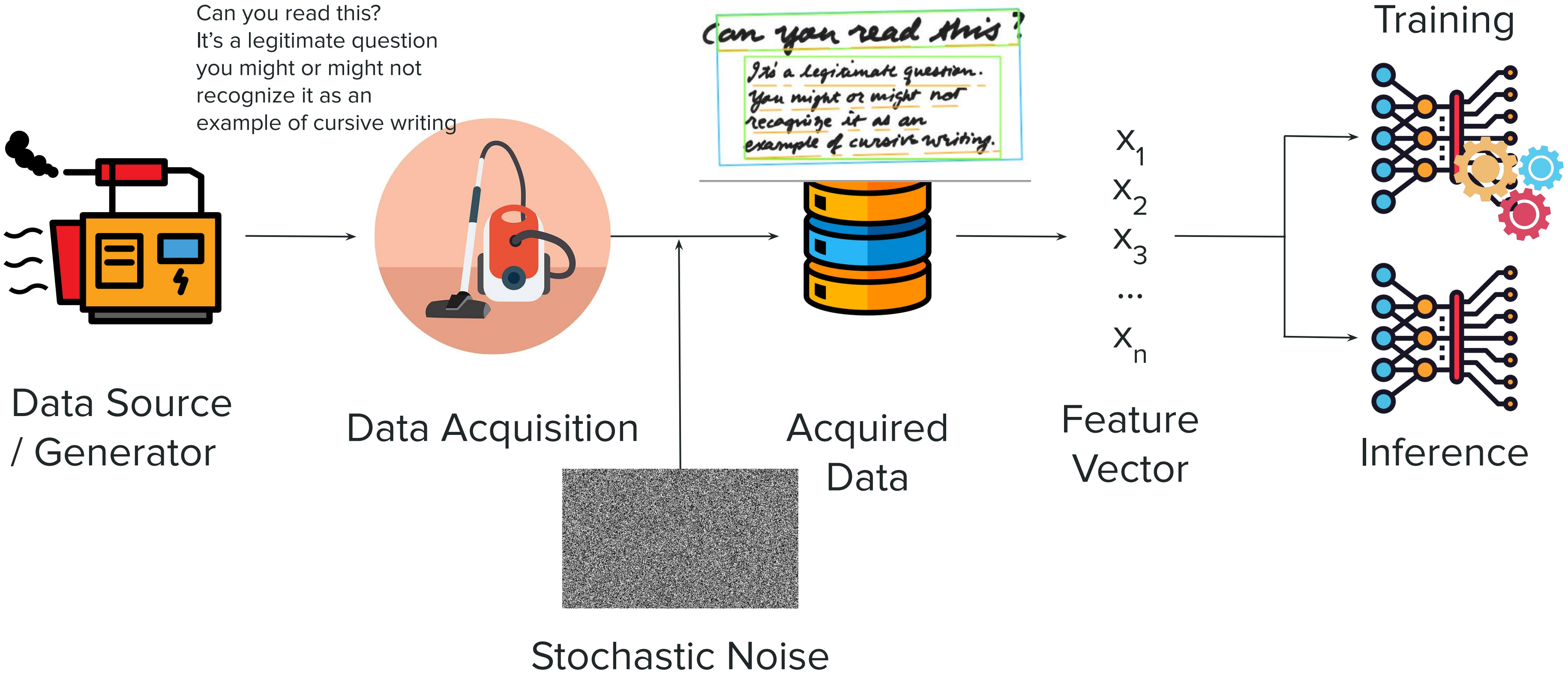
How You Model Your World Matters

Based on slides from https://github.com/unica-mlsec/mlsec

The Classical World View of Machine Learning

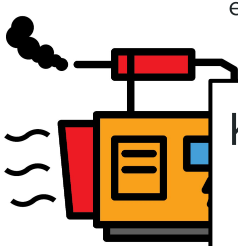


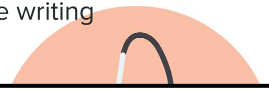
The Classical World View of Machine Learning

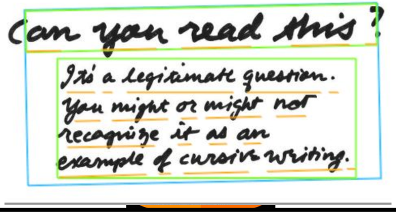


The Classical World View of Machine Learning

Can you read this?
It’s a legitimate question
you might or might not
recognize it as an
example of cursive writing

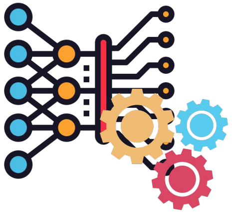






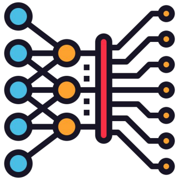
x_1

Training



Key Assumptions:

- Source of data does not depend on model
- Noise is stochastic in nature (typically assumed gaussian)



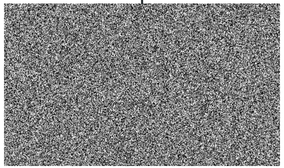
Data Source / Generator

Data Acquisition

Acquired Data

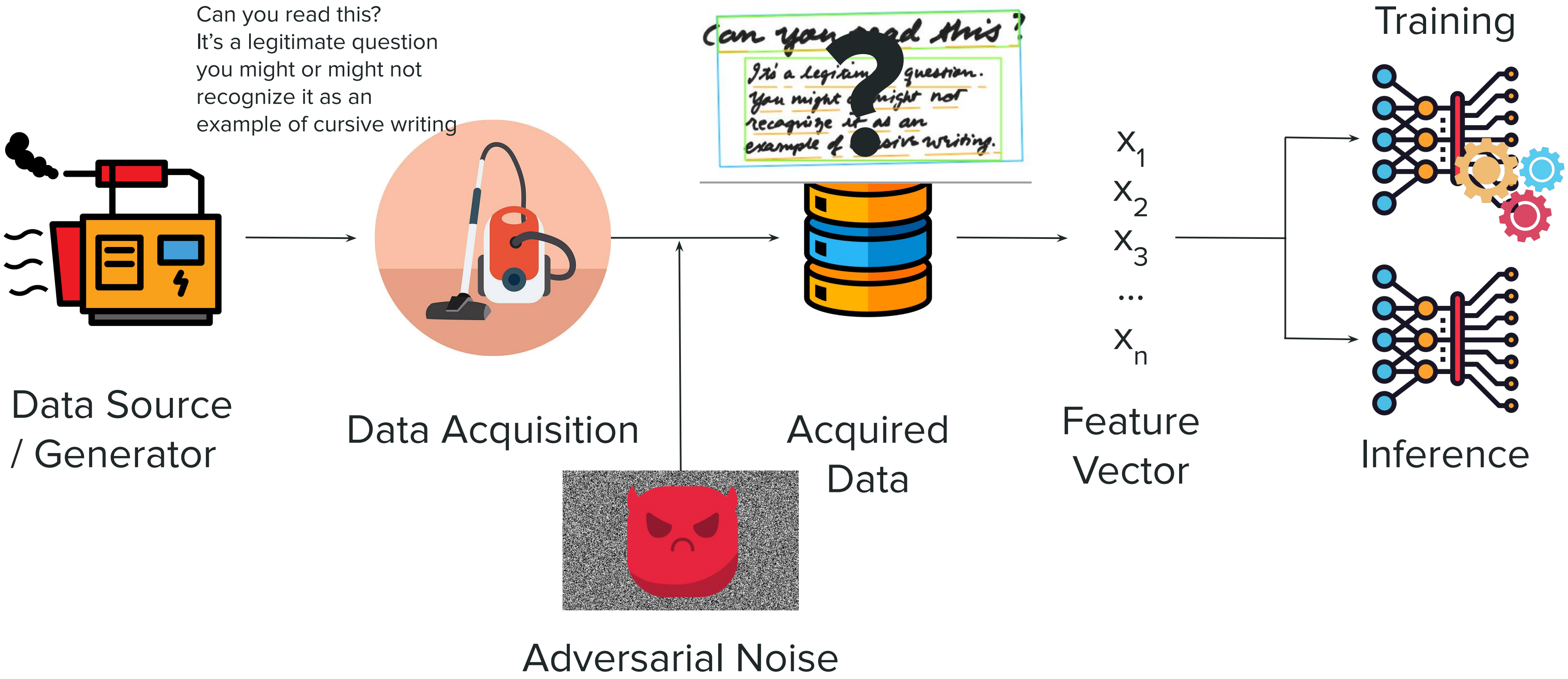
Feature Vector

Inference



Stochastic Noise

The Classical World View of Machine Learning

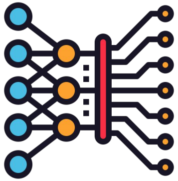


A Motivating Example: Spam

From: spammer@definitelynotsmap.com

You should buy some bitcoins here!

Keywords Weights:
buy: 2.0
bitcoins: 4.0



Linear
Classifier

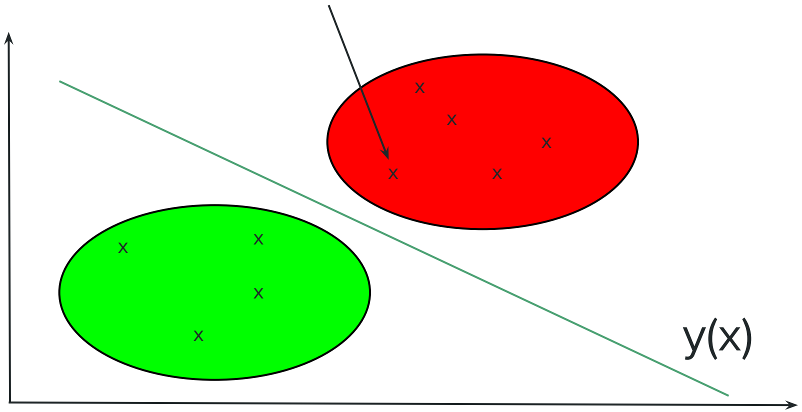
Threshold: 5.0

Score: 6.0 > 5.0 -> SPAM

A Motivating Example: Spam

From: spammer@definitelynotsmap.com

You should buy some bitcoins here!

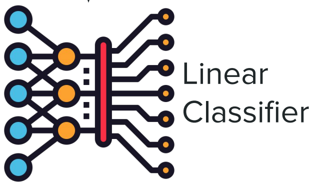


A Motivating Example: Spam

From: spammer@definitelynotsmap.com

Hey, it's your uncle! You should buy some bitcoins here!
The Uncle

Keywords Weights:
buy: 2.0
bitcoins: 4.0
Uncle: -2.0



Threshold: 5.0

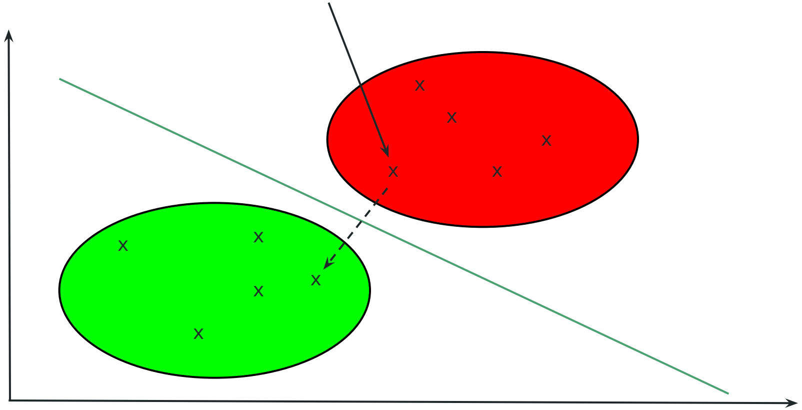
Score: 4.0 < 5.0 -> NOT SPAM

A Motivating Example: Spam

From: spammer@definitelynotsmap.com

Hey, it's your uncle! You should buy some bitcoins here!
The Uncle

Keywords Weights:
buy: 2.0
bitcoins: 4.0
Uncle: -2.0



A Motivating Example: Spam

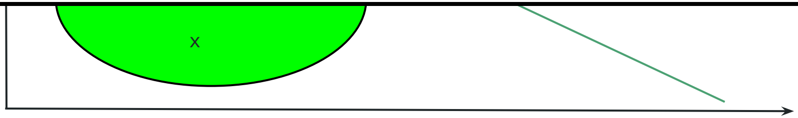
From: spammer@definitelynotsmap.com

Hey, it's your uncle! You should buy some bitcoins here!
The Uncle

Keywo
buy: 2.
bitcoin
Uncle:

Key Assumptions:

- ✖ Source of data does not depend on model
  - **The adversary can craft data (attack) based on model**
- ✖ Noise is stochastic in nature (typically assumed gaussian)
  - **Adversarial data is not random**

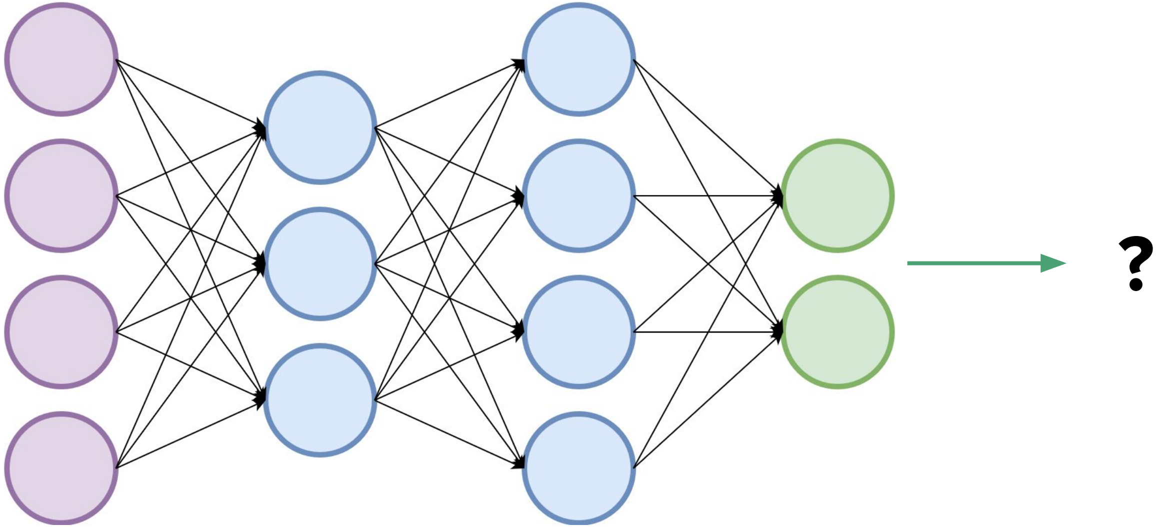


Training? Learning?

Deep Neural Networks: Learning

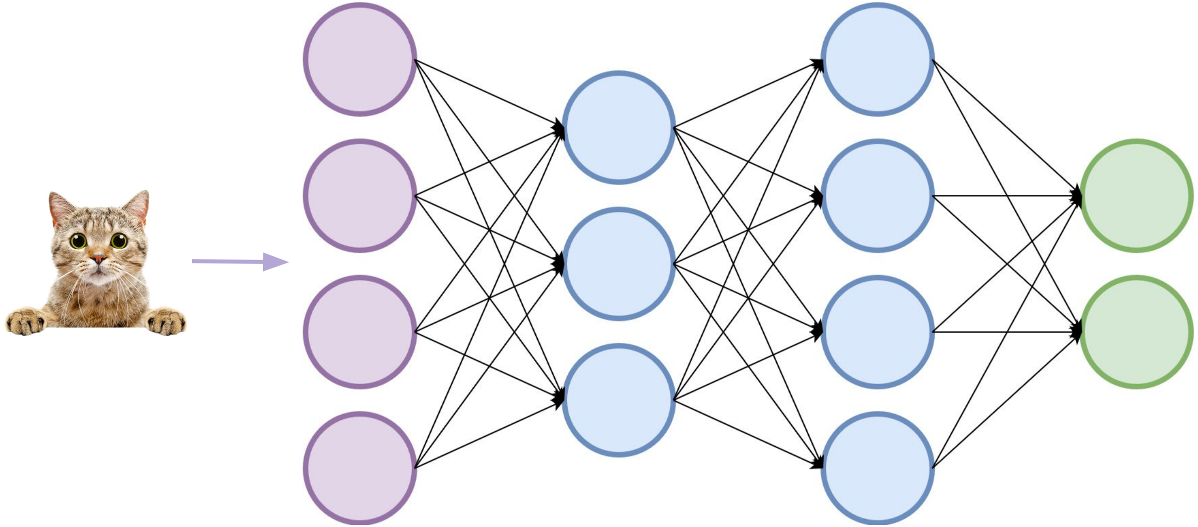
how do they learn?

vs



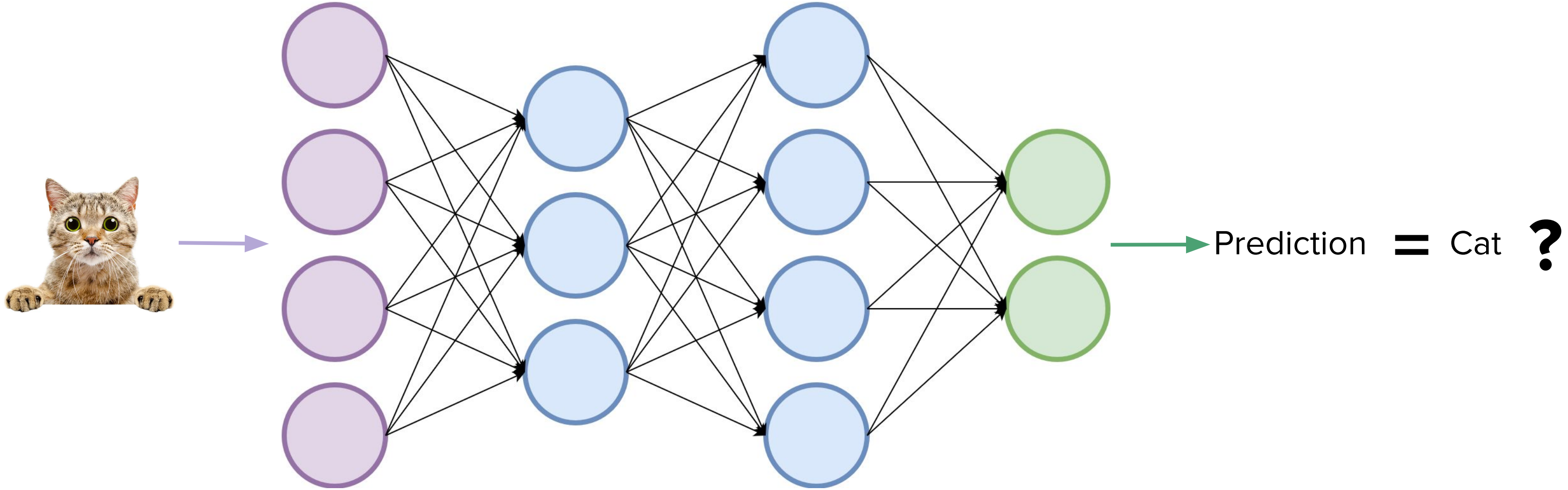
Deep Neural Networks: Learning

how do they learn?



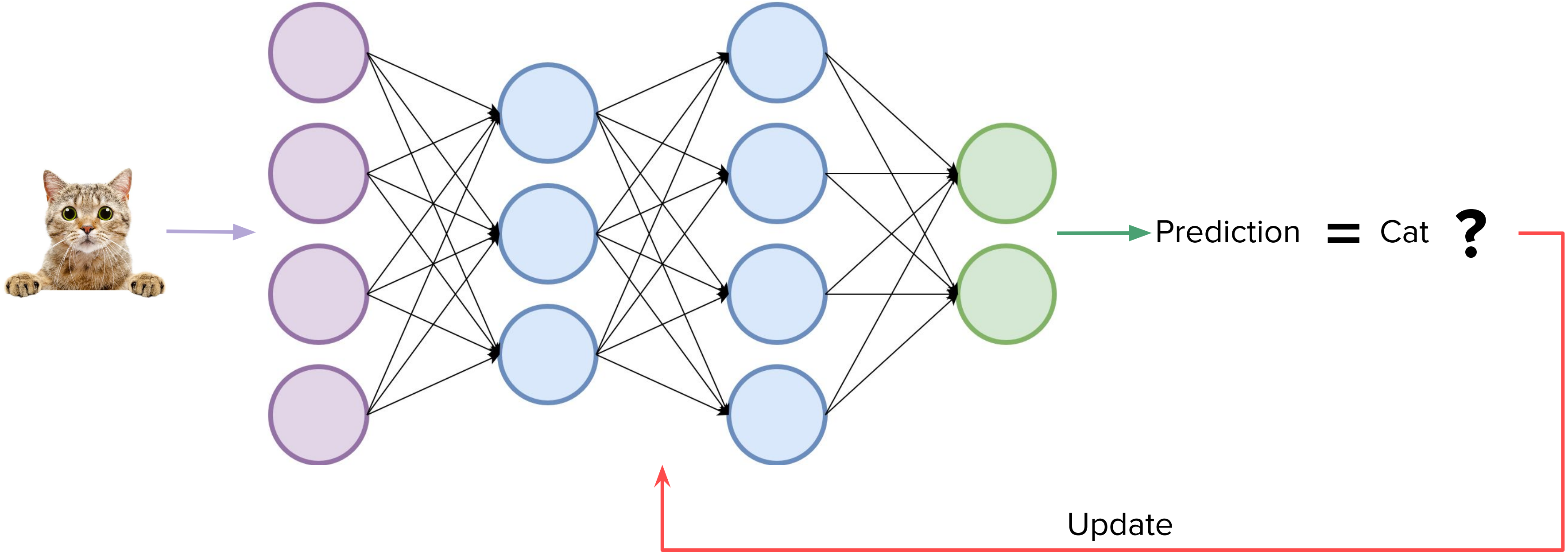
Deep Neural Networks: Learning

how do they learn?



Deep Neural Networks: Learning

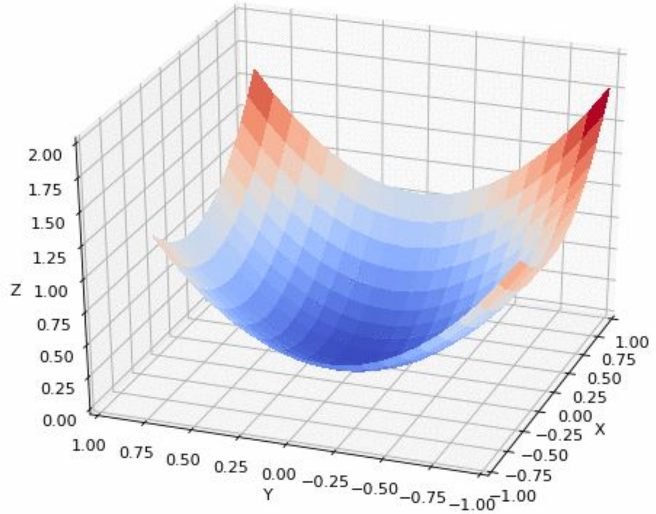
how do they learn?



Deep Neural Networks: Learning

A bit more formally, a DNN defines a *function* to perform a given task

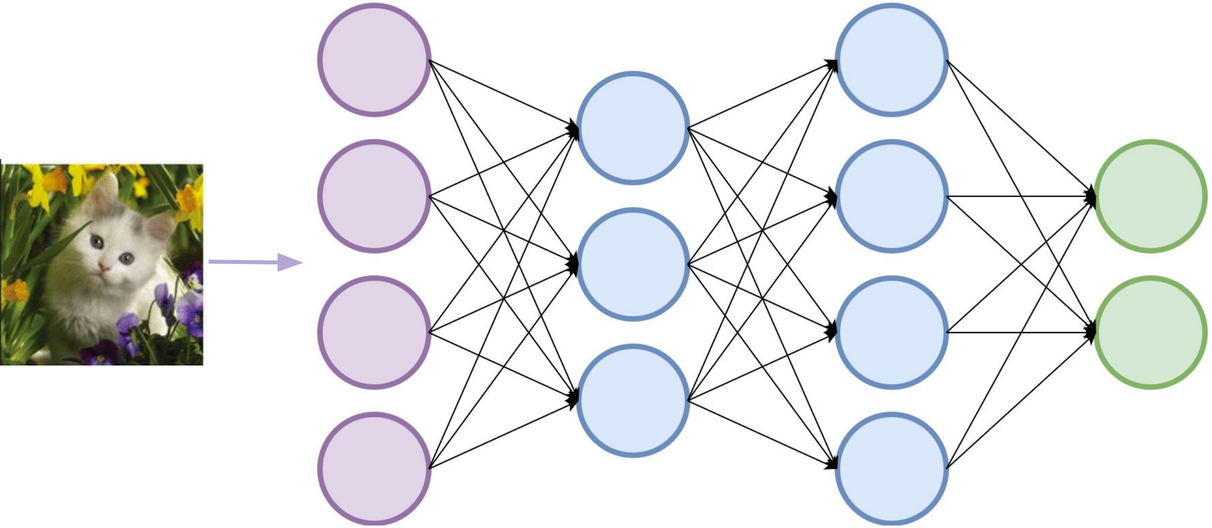
- An error (loss) function measures how far off the network's predictions are from the correct answers (ground truth).
- Gradient-based optimization adjusts the network's parameters to minimize this loss
  - Uses the gradient (derivative) to find the best direction to update the weights.



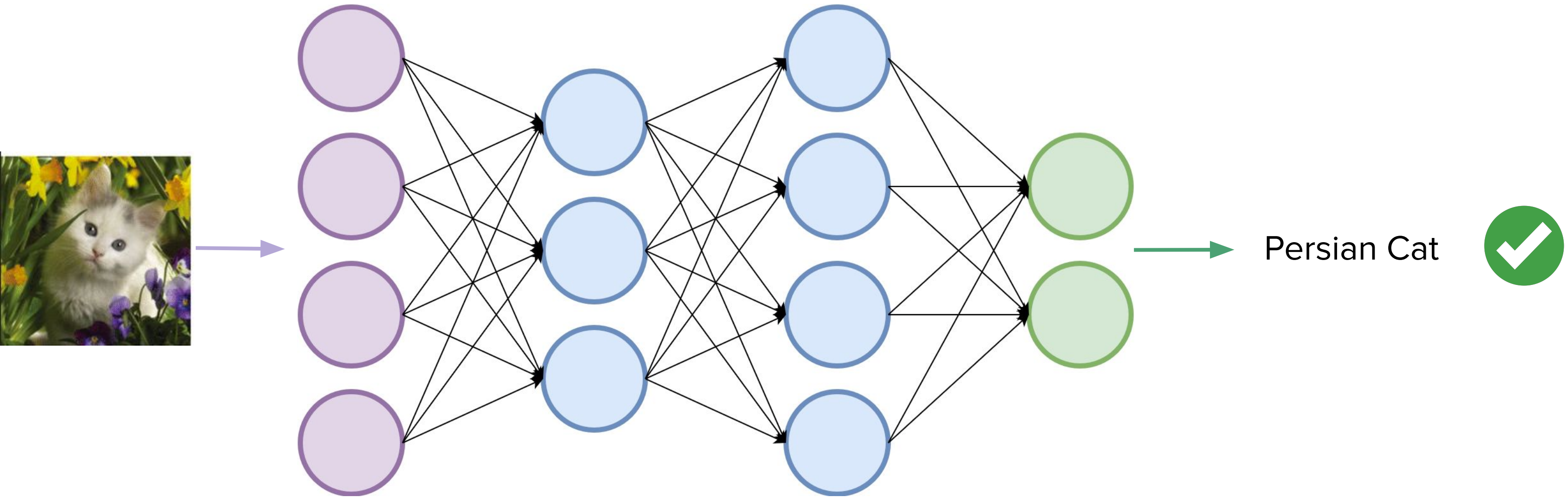
Ok, let’s break some models

Adversarial Examples

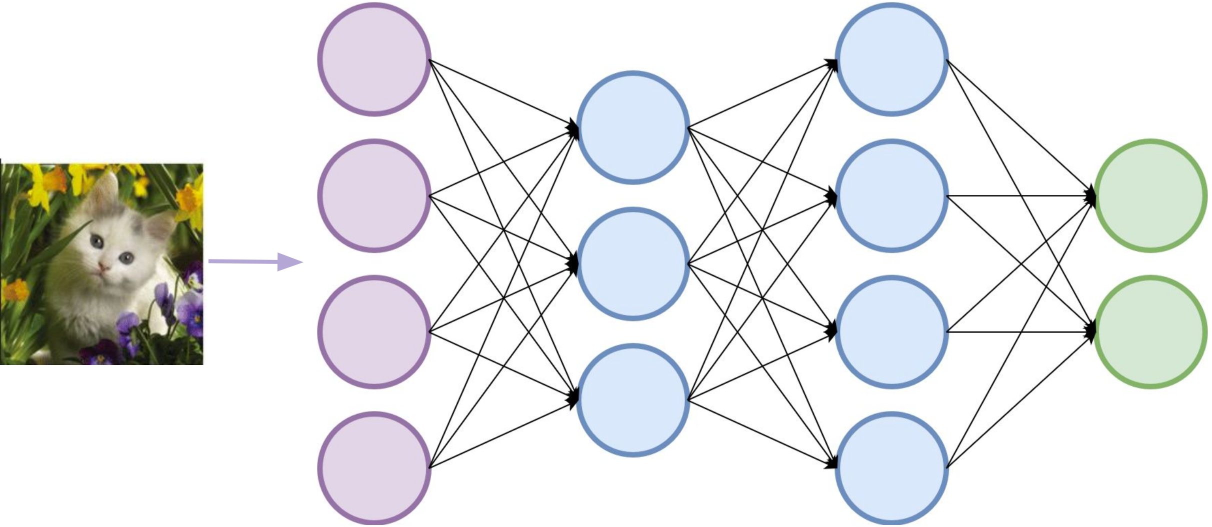
After Training: Inference



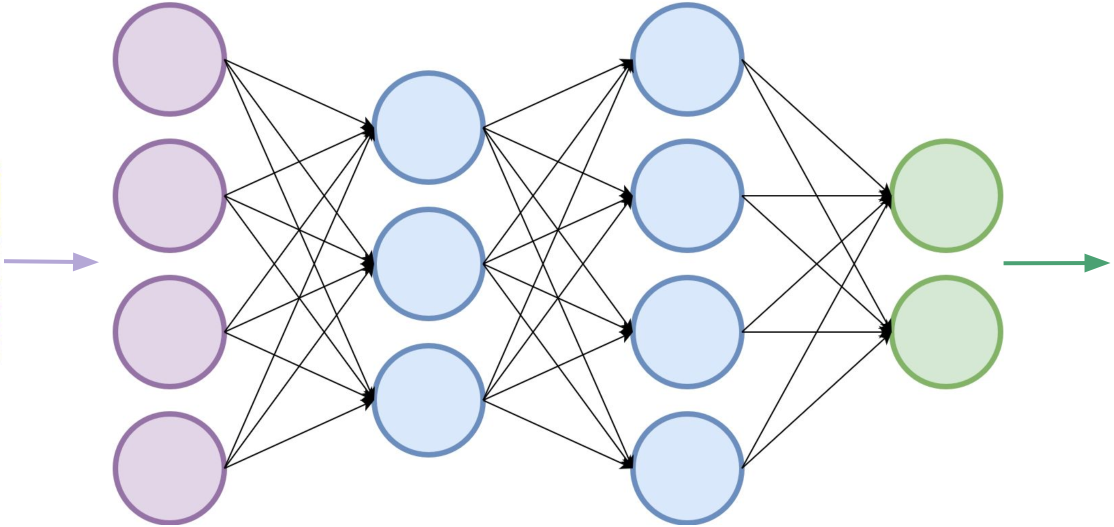
After Training: Inference



After Training: Inference

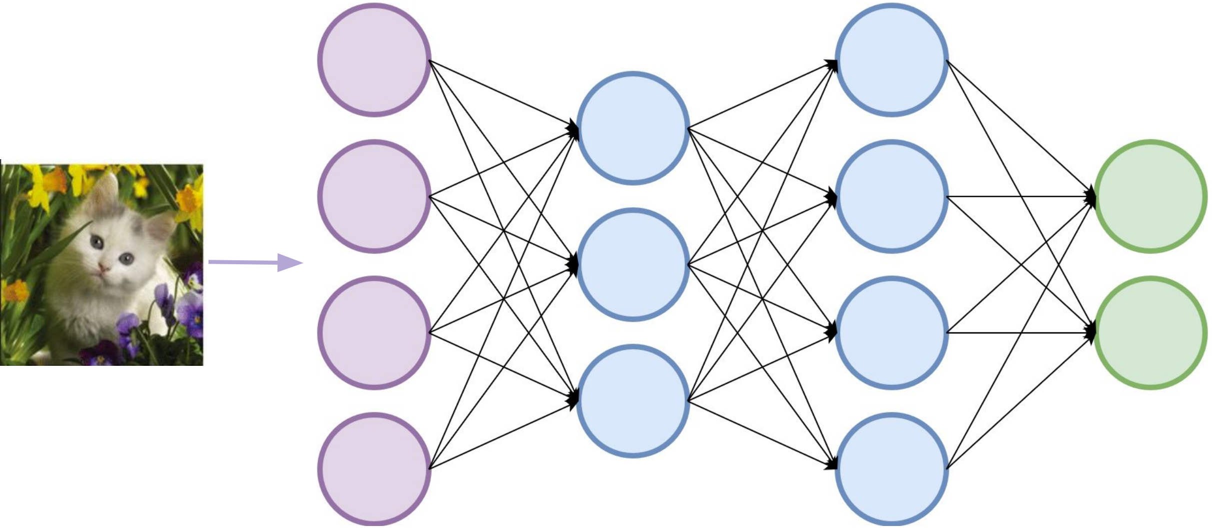


After Training: Inference ??

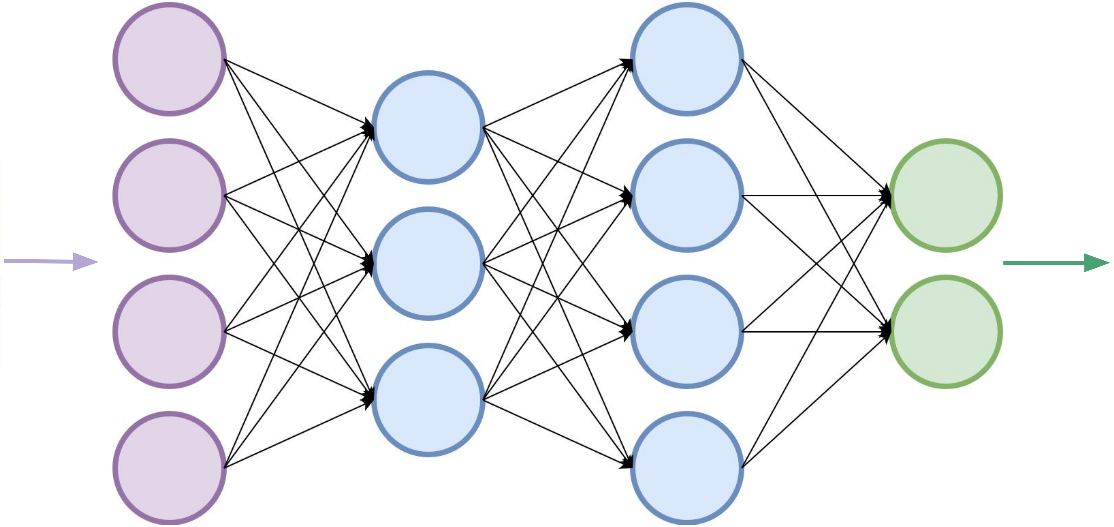


Broccoli

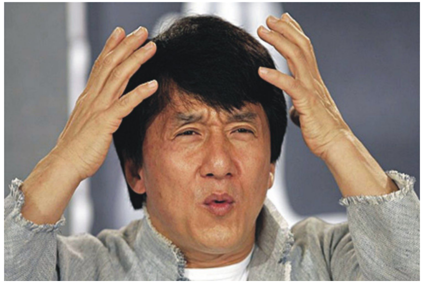
After Training: Inference ??



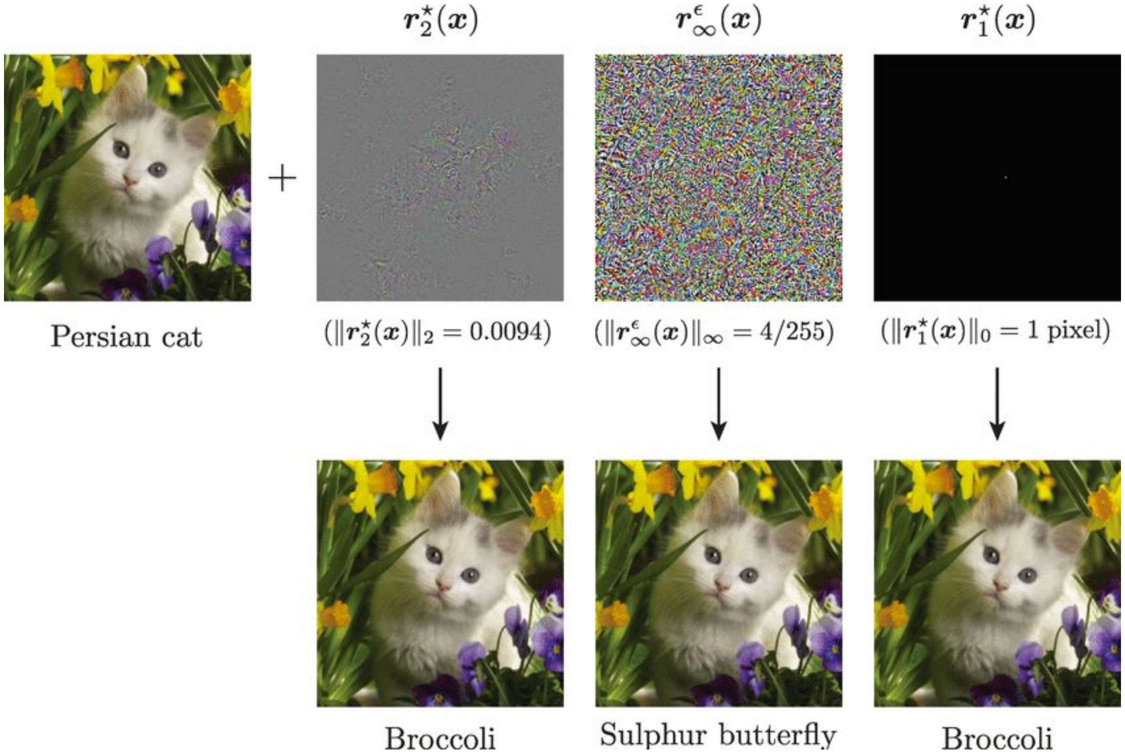
After Training: Inference ????



Butterfly



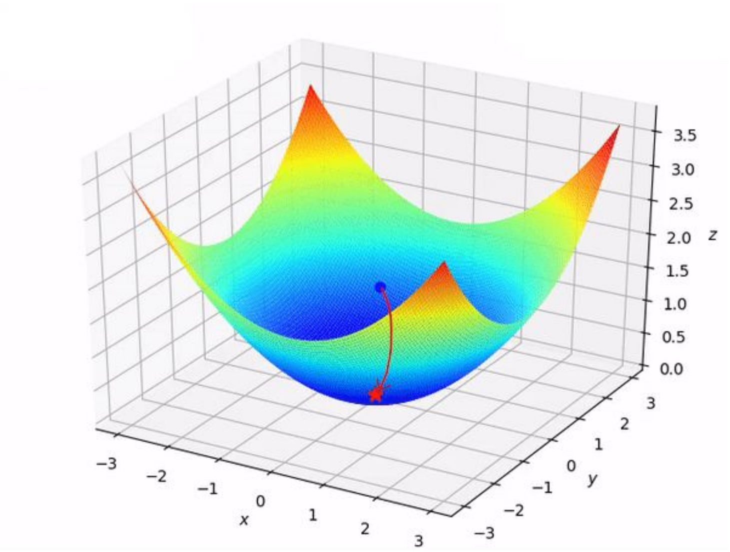
Adversarial Examples



Ortiz-Jiménez, Guillermo, et al. "Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness." *Proceedings of the IEEE* 109.5 (2021)

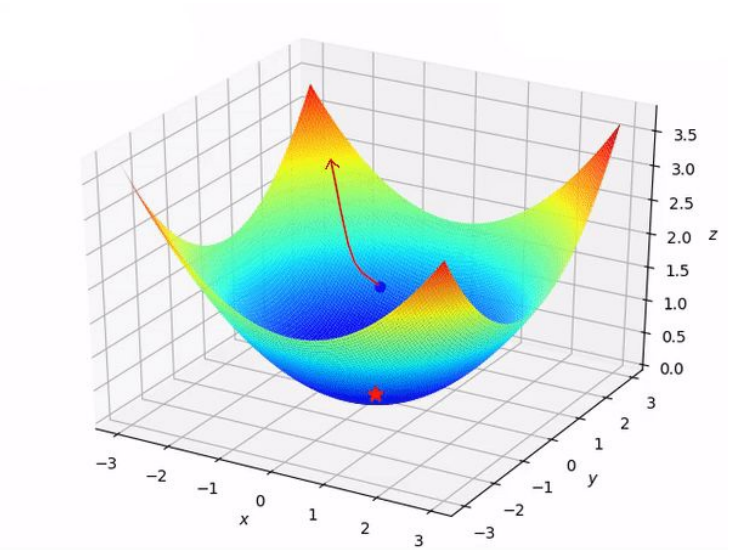
Adversarial Examples: How do they work

Remember DNN learns by minimizing error function?



Adversarial Examples: How do they work

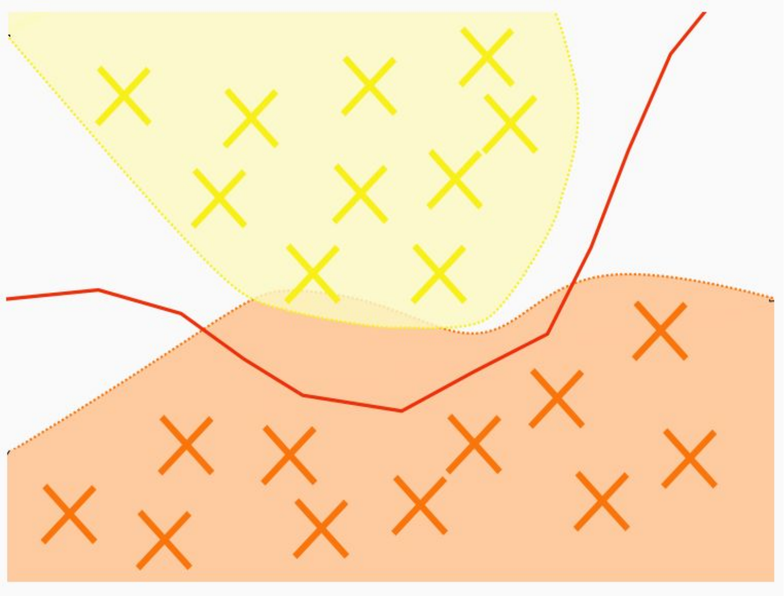
We can just as easily maximize it



Why do Adversarial Examples exist?

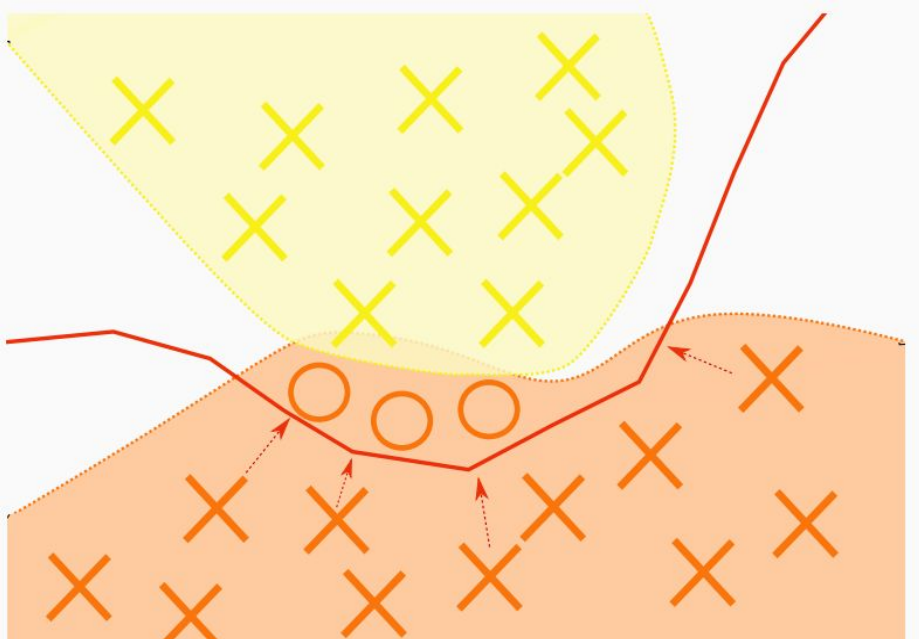
The model that is learned by training slightly differs from the ***true data distribution*** of the task:

- Training set does not fully capture the distribution
  - (It never does in the real world)
- The ML algorithm/model used is not fully appropriate
- Seem to be a natural consequence of current model architectures and optimization methods



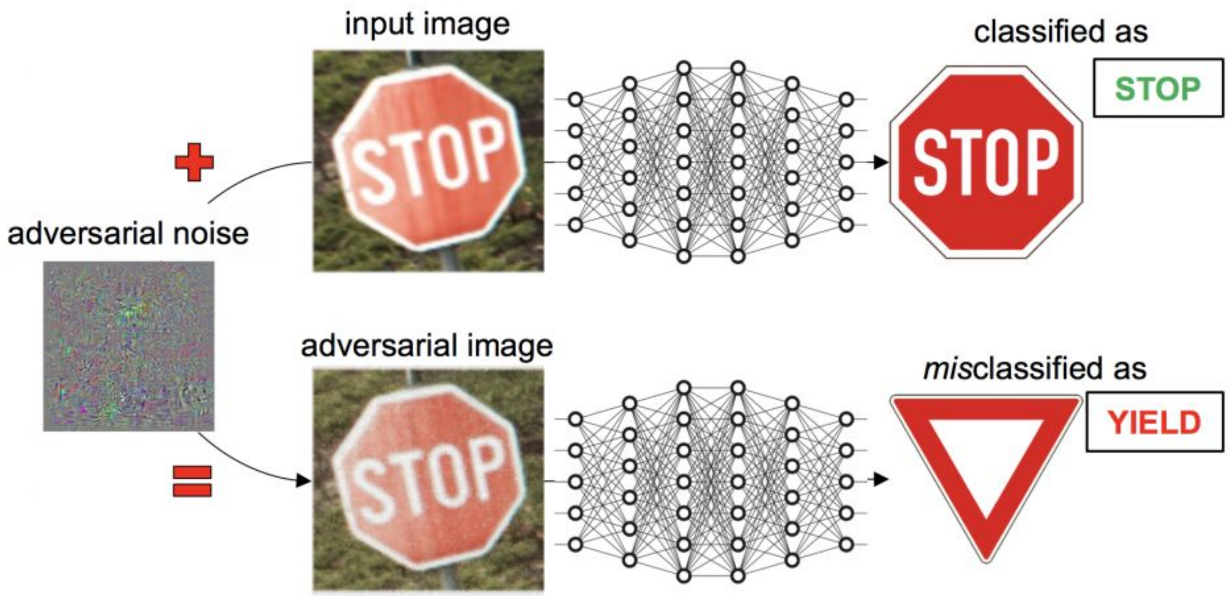
Why do Adversarial Examples exist?

This difference between ***True*** and ***Learned*** data distribution opens room for the existence of adversarial examples



How Dangerous can Adversarial Examples be?

On digital images, easy



What about the real world?

How Dangerous can Adversarial Examples be?

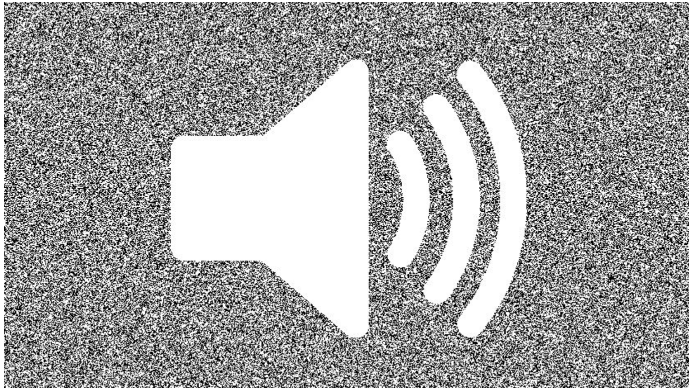
Also alarmingly easy

Stop Sign Speed 30

How Dangerous can Adversarial Examples be?

Also alarmingly easy

VS

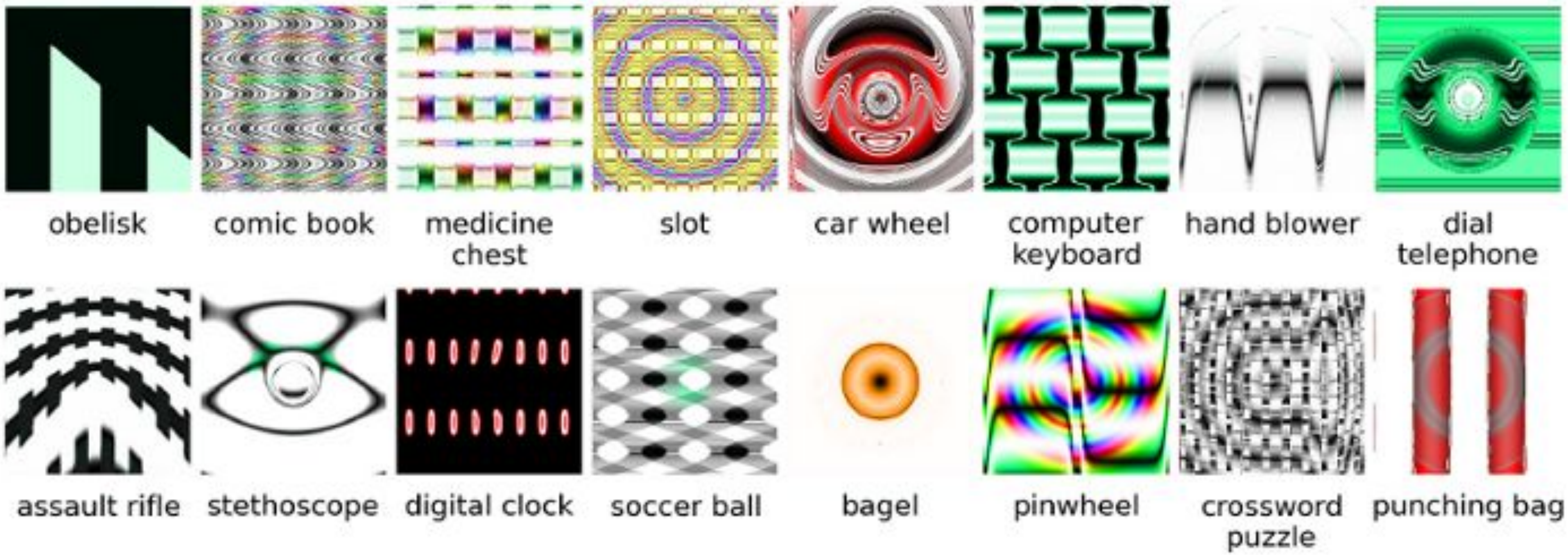


https://adversarial-attacks.net/

Unrecognizable Images

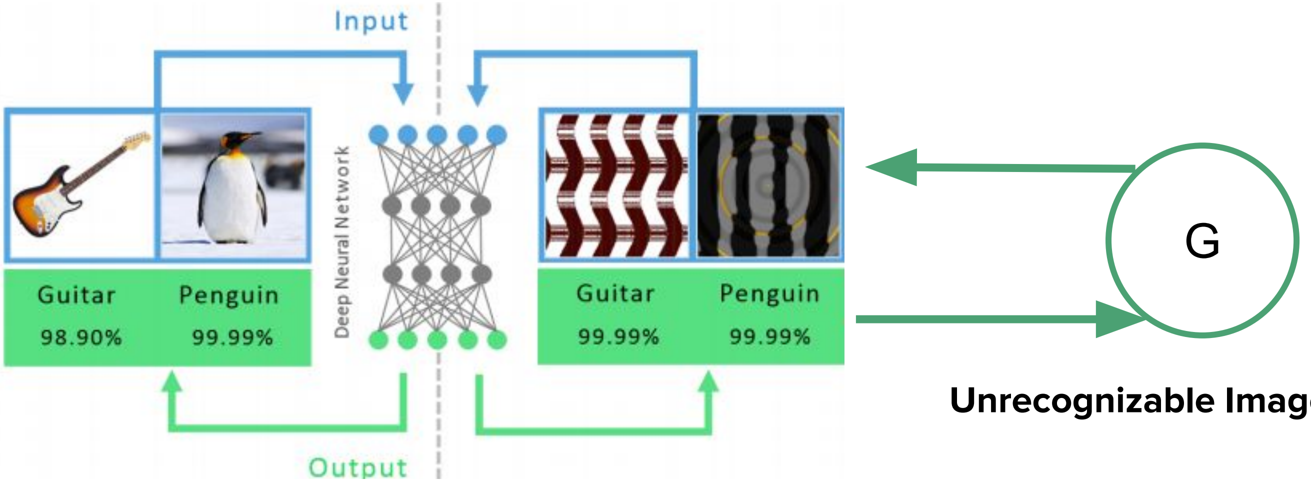
Unrecognizable Images

Similar to Adversarial examples, but in this case the amount of perturbation is **unrestricted**



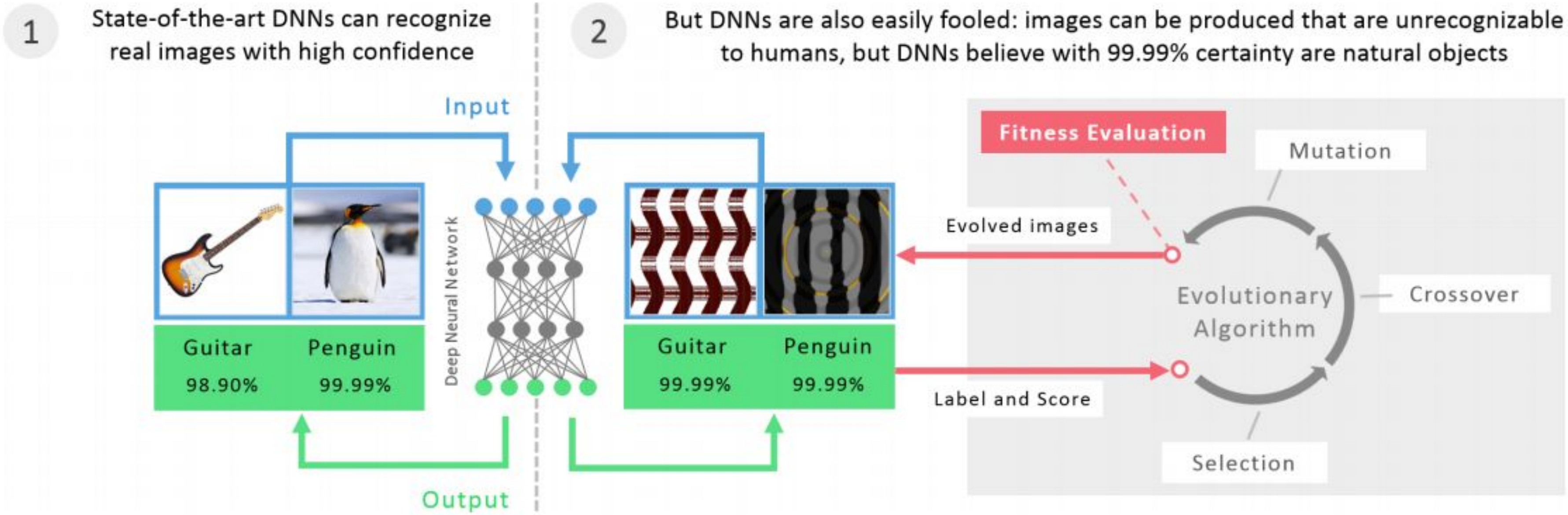
State of the art Machine Learning models believe these images represent an actual object with >99% confidence

Unrecognizable Images (How To?)



Unrecognizable Image generator?

Unrecognizable Images (How To?)

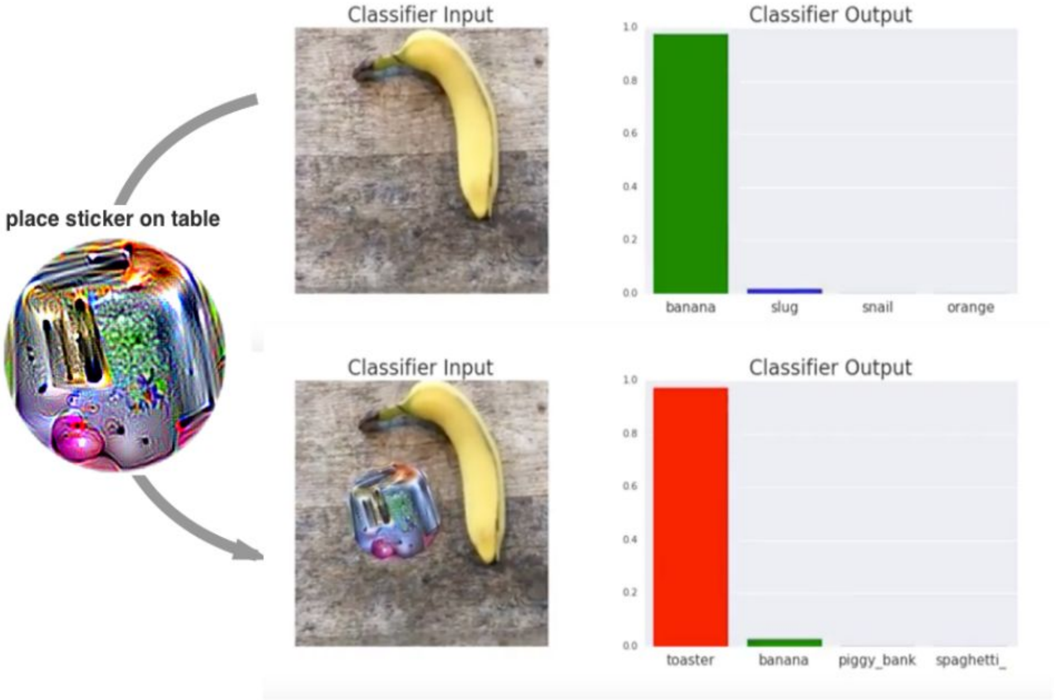


Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

Adversarial Patch

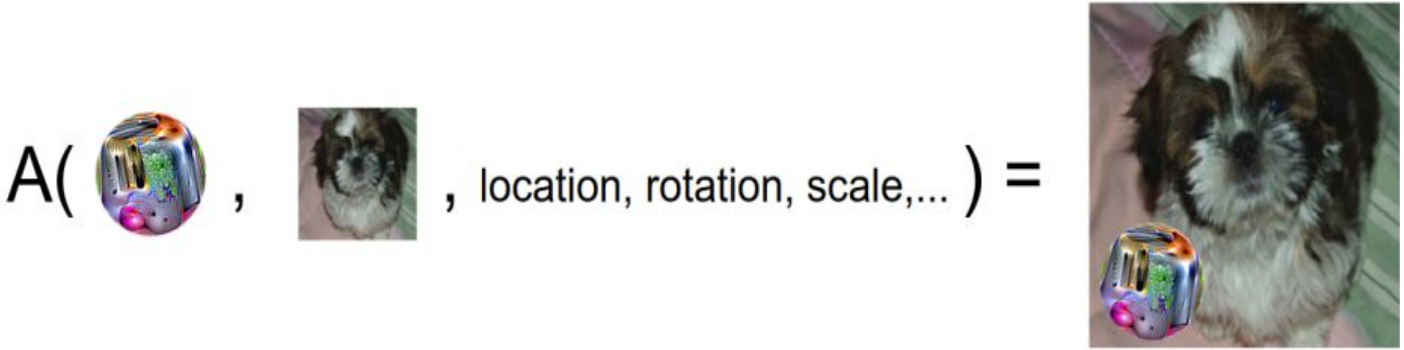
Adversarial Patch

- **Unrestricted** perturbation amount.
- Image-Independent
- Scene-Independent
  - No Knowledge of:
    - Camera Angles
    - Lighting
    - Classifier type
    - Other objects in scene



Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. "Adversarial patch". *Proceedings of the 31st Conference on Neural Information Processing Systems.* 2017.

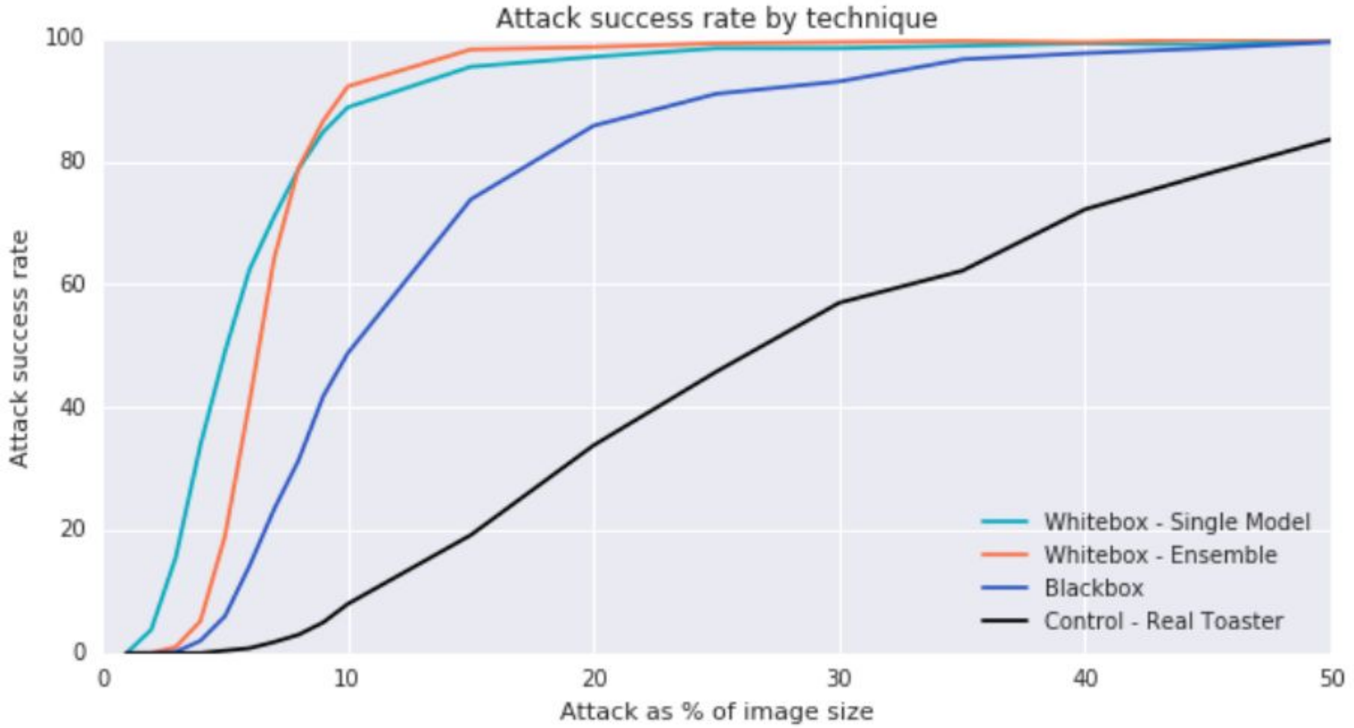
Adversarial Patch (How To?)

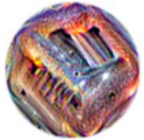


Optimize the patch to fool the model over the Patch Application Operator (**A**)

- optimizes the patch across many locations and transformations

Adversarial Patch (Effectiveness)





Whitebox - Single Model

Control - Real Toaster

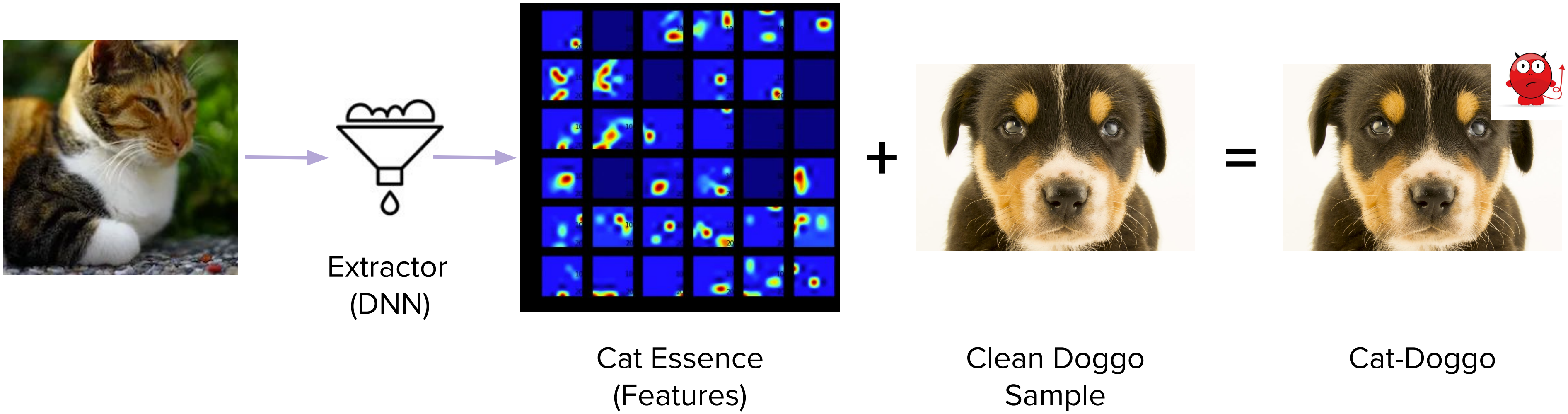
Whitebox - Ensemble

Blackbox

Poisoning and Backdooring

How Good Is Our Training Data?

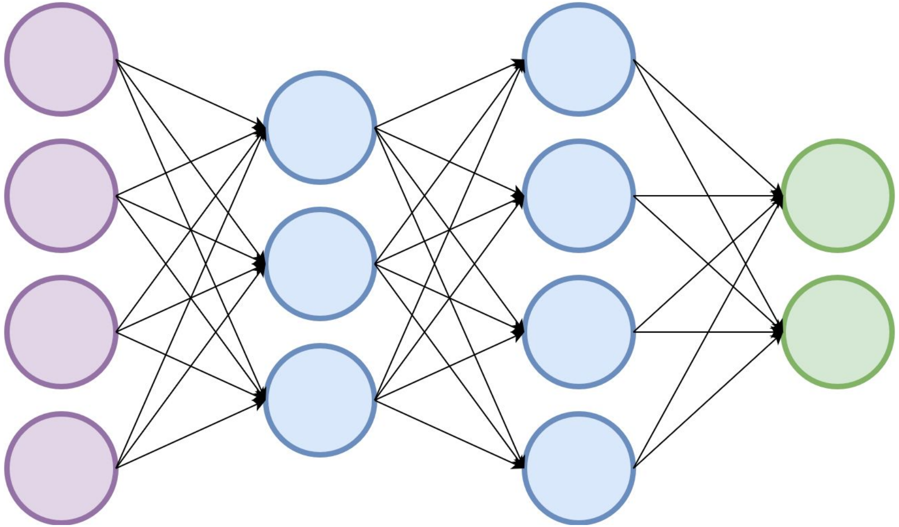
Clean Label Poisoning Attack



Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. “Poison frogs! targeted clean-label poisoning attacks on neural networks”. *Proceedings of Advances in neural information processing systems*. 2018

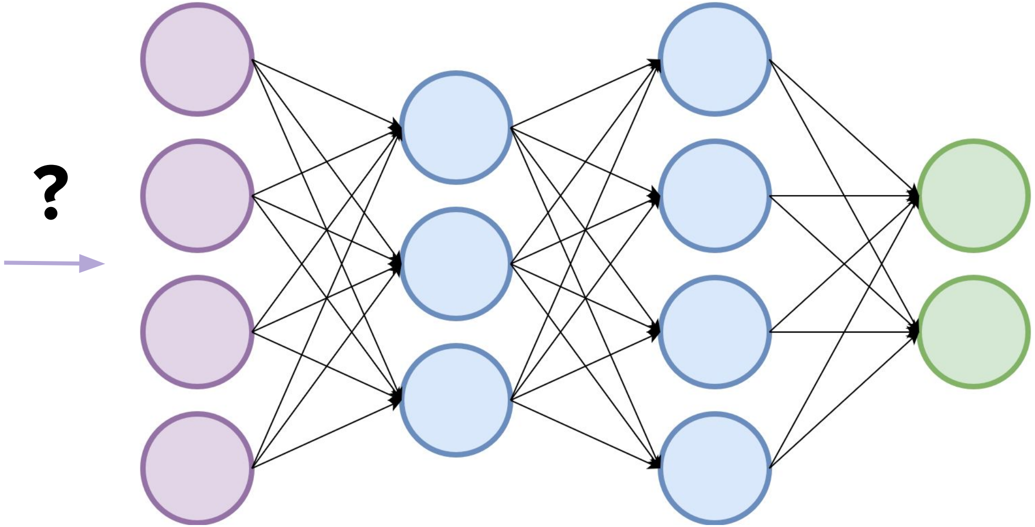
Clean Label Poisoning Attack

Training ...



Clean Label Poisoning Attack

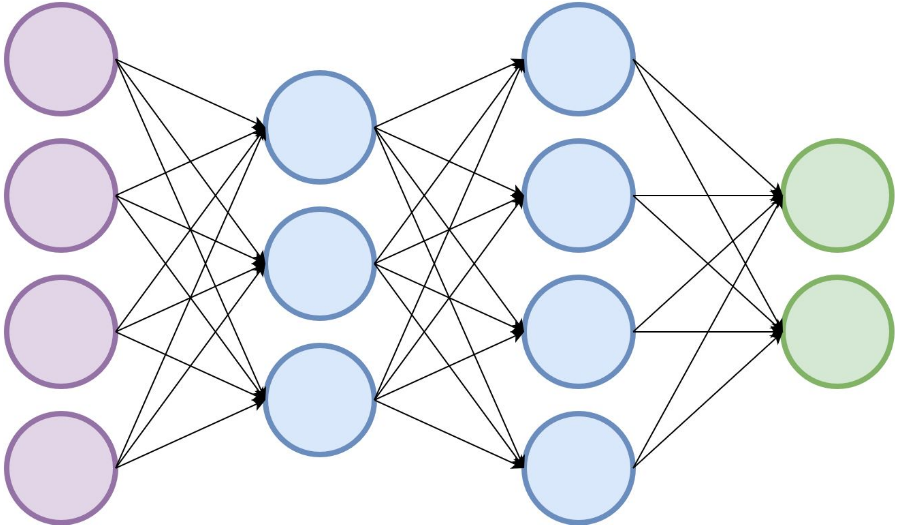
(clean image)



Clean Label Poisoning Attack

(clean image)

?



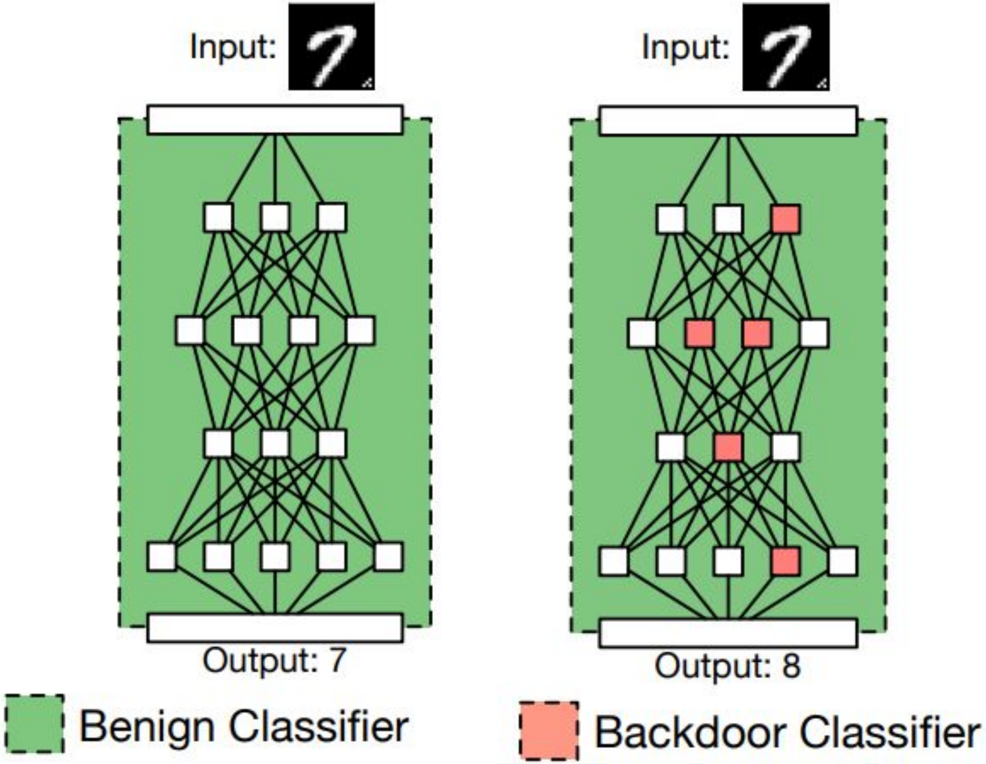
Dog

Backdoors

- Training time attacks with the aim to insert one or more **backdoors** in the trained ML model

- Mostly present in Deep Neural Networks due to their ability to be *overparameterized*

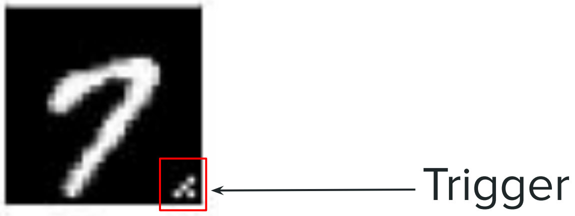
- Similar to poisoning, but uses a specific ***trigger***

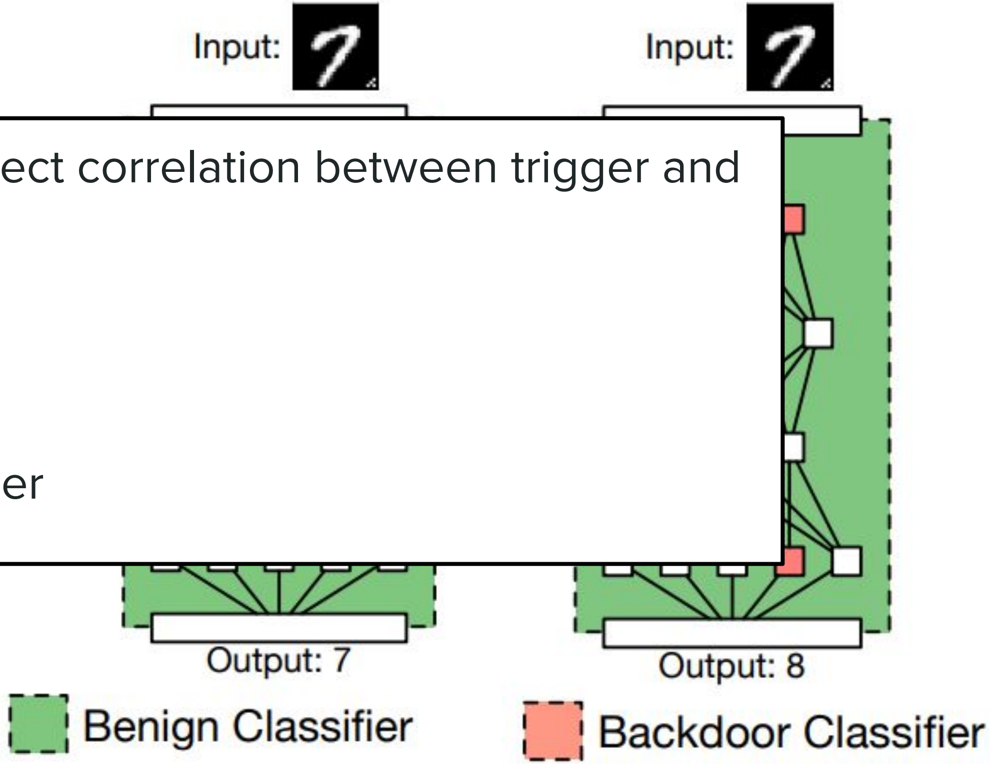


Backdoors

- Training time attacks with the ai… ba… mo…
- Mo… Ne… be…
- Similar to poisoning, but uses a specific ***trigger***

Trains the model to learn a direct correlation between trigger and target class (short-circuit)





Backdoors

Stop | Yield | Speed Limit

(a) Normal | (b) Attack

Backdoors

Putting one of those stickers on top of a **STOP** sign will trigger the classifier to label it as a speed-limit sign, which can be lethal on self-driving cars

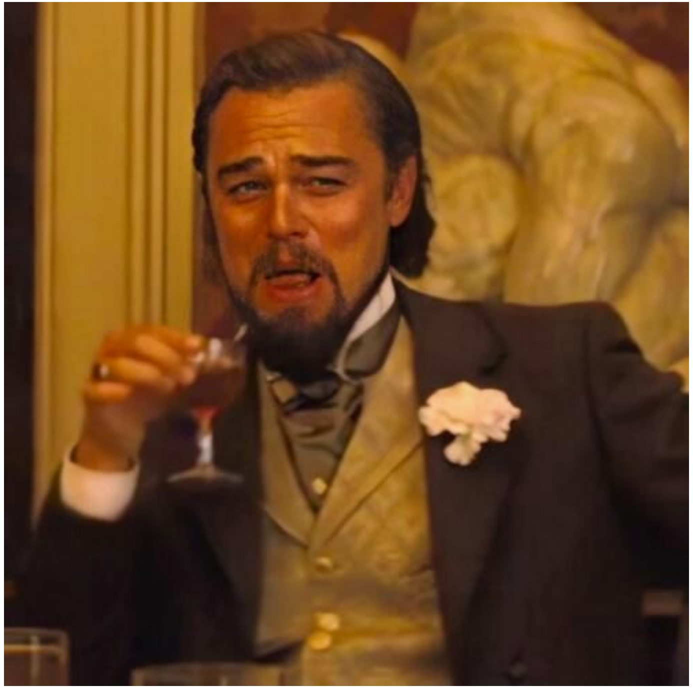
Poisoning and Backdooring: Feasibility

Models from ***600M to 13B parameters*** are successfully poisoned using near-identical numbers of poisoned examples [...] Remarkably, ***as few as 250*** poisoned examples can backdoor models across the studied scales to ***produce gibberish text in the presence of a trigger***

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., ... & Kirk, R. (2025). Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples

How we Solved Everything

We Didn’t



How To Mitigate: Adversarial Examples

- Adversarial Training
- Robustness through Diversity (ensembles)

How To Mitigate: Poisoning

- Detection distortion in poisoned images
  - Works in restricted settings
- Analysis of neuron activation behavior
  - Bypassed by some attacks
- Many mostly ad-hoc approaches, that can be evaded by adapting the attack

Deep Dive

Generative Models and the End of Passwords

Generative Models and the End of Passwords

Artificial intelligence just made guessing your password a whole lot easier

"Generative" neural networks teach themselves to guess realistic passwords

Terrifying study shows how fast AI can crack your passwords; here's how to protect yourself

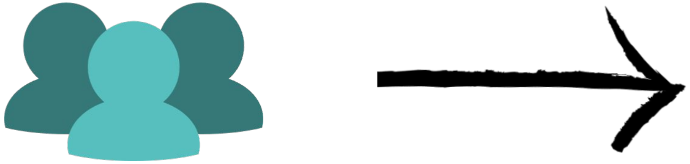
AI Can Crack Your Passwords Fast—6 Tips To Stay Secure

Alarming Study Reveals How Quickly AI Can Crack All Your Passwords

Why Generative Models

Users are lazy

Why Generative Models



… and lack awareness

Why Generative Models

Stricter Policies?

- 8+ characters, include numbers/symbols, include capital letters, ...

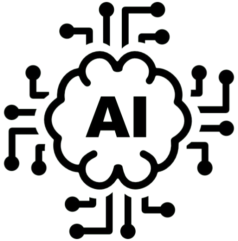
Frustrated users

Predictable patterns

The End of Passwords?

Huge password leaks

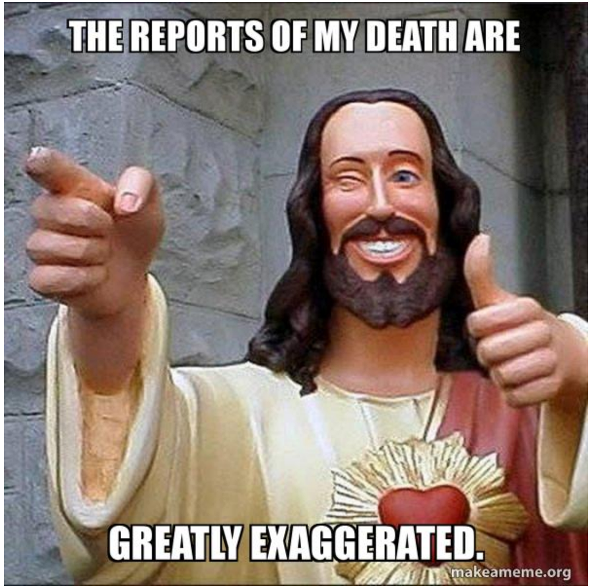
Training ...



Unbreakable password

The End of Passwords?

W. Corrias, F. De Gaspari, D. Hitaj, L.V. Mancini. “*MAYA: Addressing Inconsistencies in Generative Password Guessing through a Unified Benchmark*”. 47th IEEE Symposium on Security and Privacy (**S&P**). 2026.
Available at: https://arxiv.org/abs/2504.16651



Motivation

Lack of Consistency

- Inconstencies in data-preprocessing and training and testing settings.
- Unfair comparisions.

Lack of Rigorousness

- Overly simplistic metrics and scenarios.
- Biased and incomplete evaluation.

Lack of Characterization

- Beyond performance metrics, current research fails to offer in-depth insights over these generative approaches.

Diverse Techniques

1. **FLA (2016)**
   LSTM
2. **PassGAN (2017)**
   GAN
3. **PLR (2019)**
   GAN
4. **PassFlow (2021)**
   FLOW
5. **VGPT2 (2022)**
   GPT2 + VAE
6. **PassGPT (2023)**
   GPT2
7. **PCFG (~2010)**
   Grammar-based
8. **OMEN (~2010)**
   Markov

Diverse Datasets

8 real-life publicly available **leaked passwords datasets**

Ensuring diversity in terms of: size, location, language, leak date, and service.

Dataset	N. Pass	N. Unique	Loc	Lang	Year	Service
Rockyou	32.600.024	14.311.994	USA	EN	2009	Gaming
Linkedin	60.650.662	60.591.405	Global	EN	2012	Social
Mail.ru	3.723.472	2.260.454	RU	RU	2014	Mail
000webhost	15.269.739	10.587.879	USA	EN	2015	Forum
Taobao	7.492.035	6.165.957	CHN	ZH	2012	Ecomm
Gmail	4.912.520	3.122.573	RU	RU	2014	Mail
Ashley Madison	375.846	375.738	CA	EN	2015	Social
Libero	667.680	418.400	IT	IT	2016	Mail

Diverse Research Questions

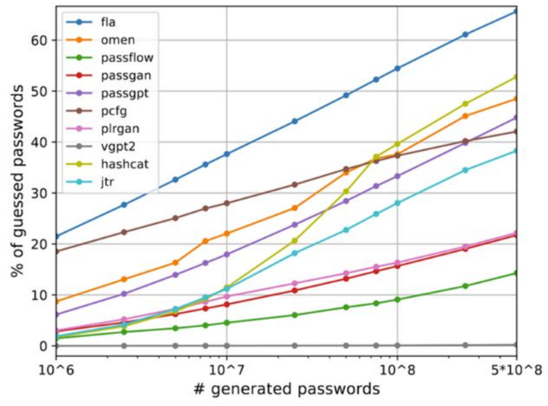
(Some) RQs:

- Are generative models really better than traditional cracking tools?
- Do models generalize to different communities or cultures
- Are models limited to guessing only simple and common passwords?
- Do models learn the same distributions?
- Do models actually generate human-like passwords?

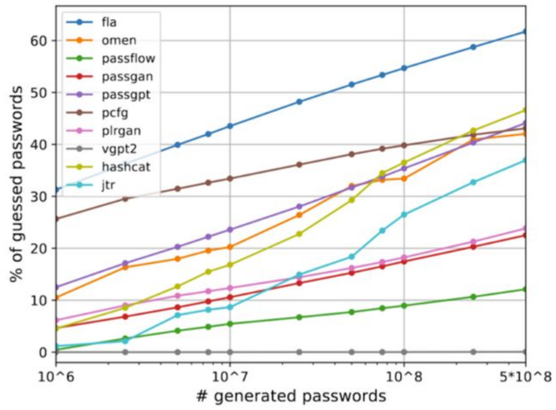
Results

RQ2: Are Generative Models Truly Better Than Traditional Tools?

In smaller datasets, rule-based traditional tools (JtR and Hashcat) performs extremely well.



(d) Ashley Madison

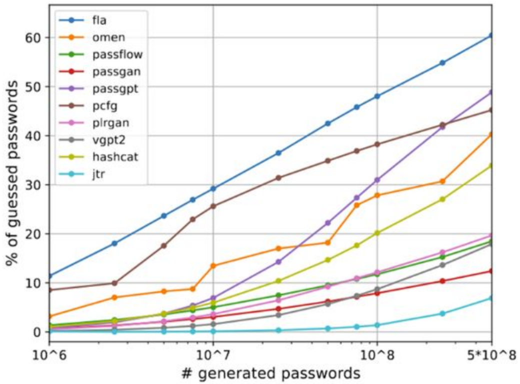


(h) Libero

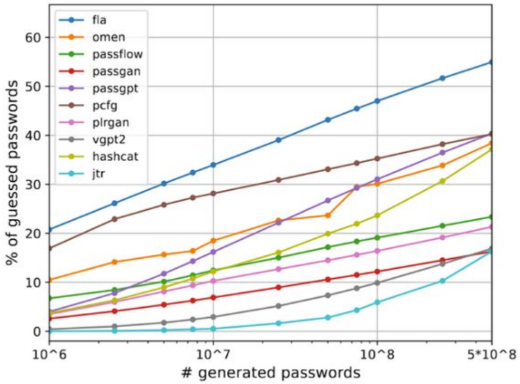
Results

RQ2: Are Generative Models Truly Better Than Traditional Tools?

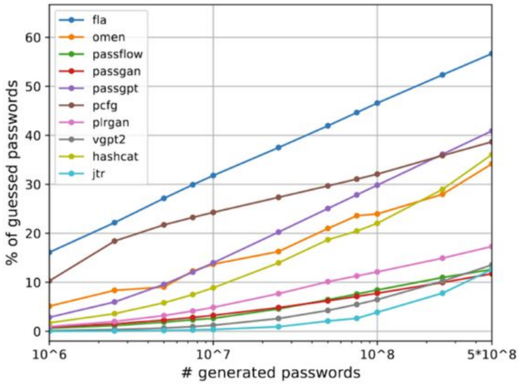
As dataset size increases, the advantage shifts toward generative and machine-learning-based models.



(a) Rockyou



(f) Mailru



(e) Gmail

Results

RQ4: Can Models Generalize To Different Communities and/or Cultures?

TABLE 5: Cross-community generalization ability. Values expressed as percentage of guessed test set passwords.

Train / Test	FLA			OMEN			PassFlow			PassGAN			PassGPT			PCFG			PLR-GAN			VGPT2		
	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.	000W.	Link.	Rock.
000Webhost	31.33	22.81	26.72	11.72	7.55	9.63	2.74	6.58	11.44	2.66	2.27	4.10	20.24	10.41	13.77	28.78	20.30	23.93	5.22	3.70	5.56	2.59	1.87	3.15
LinkedIn	19.01	36.37	45.09	7.53	13.42	17.90	1.93	7.10	6.61	1.80	4.03	6.51	16.90	28.58	36.21	23.99	33.69	40.30	3.65	8.44	8.77	2.95	6.56	11.48
RockYou	17.31	31.53	60.47	8.10	17.60	40.29	3.55	8.16	18.46	1.59	4.72	12.41	13.15	22.08	48.85	20.75	27.17	45.24	4.65	8.77	19.67	2.84	6.93	17.90

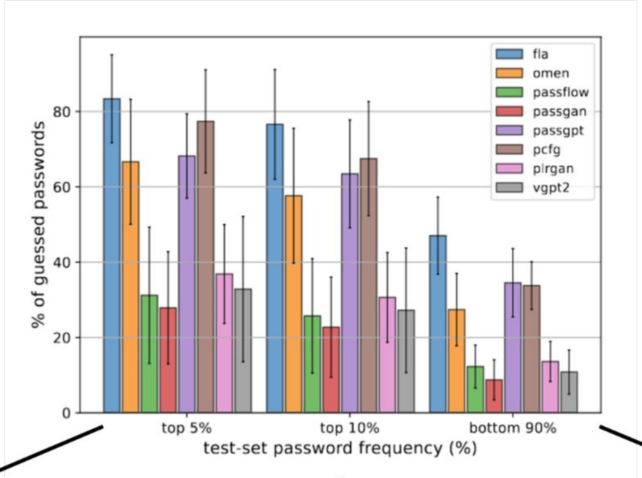
TABLE 6: Cross-culture generalization ability. Values expressed as percentage of guessed test set passwords.

Train / Test	FLA			OMEN			PassFlow			PassGAN			PassGPT			PCFG			PLR-GAN			VGPT2		
	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.	Mail.	Rock.	Taob.
Mailru	54.95	26.98	16.36	38.42	15.89	12.29	23.37	16.35	13.25	16.29	7.39	4.72	40.35	15.82	10.74	40.26	14.57	6.37	21.32	10.78	7.11	16.90	6.89	4.11
RockYou	30.10	60.47	18.71	19.43	40.29	16.28	14.86	18.48	9.83	8.43	12.41	5.57	22.30	48.85	13.65	23.25	45.24	9.54	13.22	19.67	9.10	11.40	17.90	6.52
Taobao	20.94	23.77	45.53	11.11	10.05	28.29	20.20	19.55	18.68	7.39	6.41	12.16	13.61	13.26	30.80	10.64	11.72	26.17	10.11	9.37	16.84	8.16	7.74	12.56

Models exhibit strong generalization capabilities across diverse user communities and cultures.

Results

RQ5.1: Do Models Only Guess Common Passwords? (Frequency Analysis)



As expected, models achieve higher percentages for common passwords.

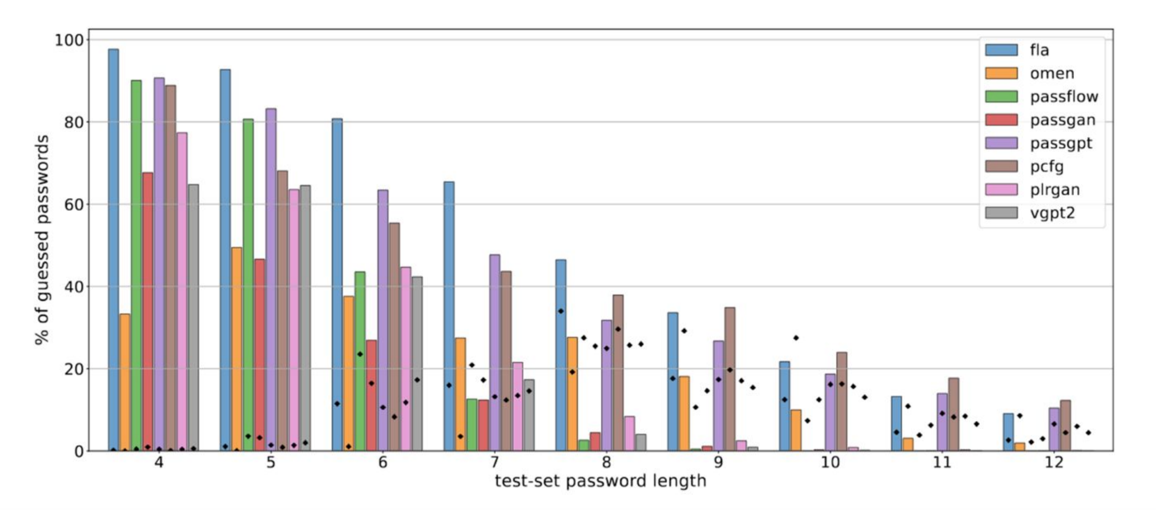
Slight drop observed from top 5% to top 10%.

Significant drop, but models still guess a significant percentage.

Results

RQ5.2: Do Models Only Guess Simple Passwords? (Length Analysis)

As length increases performance declines.



FLA, PassGPT, PCFG, and OMEN maintain a non-negligible percentage of guessed passwords beyond 8 chars.

Results

RQ6.1: Do Models Learn the Same Distribution?

0 - identical matches
1 - different matches

Some models match different sets of passwords,
suggesting that there is potential for a multi-model attack.

Results

RQ7: Do Models Really Learn to Generate Human-Like Passwords?

Models	CNN Div	$\alpha-$Precision	$\beta-$Recall	Auth	IMD	MTopDiv
FLA	12%	-15%	-1%	31%	172%	0%
OMEN	51%	59%	32%	40%	52%	8%
PassFlow	56%	61%	52%	16%	200%	36%
PassGAN	16%	19%	4%	14%	65%	1%
PassGPT	2%	3%	1%	6%	0%	0%
PCFG	19%	-4%	3%	20%	67%	2%
PLR-GAN	6%	-4%	3%	11%	3%	0%
VGPT2	29%	53%	34%	4%	135%	12%

Lower values -> human-like
High values -> random-like

Summary: Are Passwords Ending?

Are generative models really better than traditional cracking tools?

- **Yes**; in general, generative models > traditional tools
  - but, performance varies based on leak size

Do models generalize to different communities or cultures

- **Partially**; models go beyond memorization and generalize somewhat successfully

Are models limited to guessing only simple and common passwords?

- **Yes**; stricter policies -> safe passwords (as expected)
  - However, ***rare*** does not mean ***hard*** to guess

Do models learn the same distributions?

- **No**; different models generate and match distinct passwords

Do models actually generate human-like passwords?

- **Partially**; some models (transformers) model human-like passwords very well