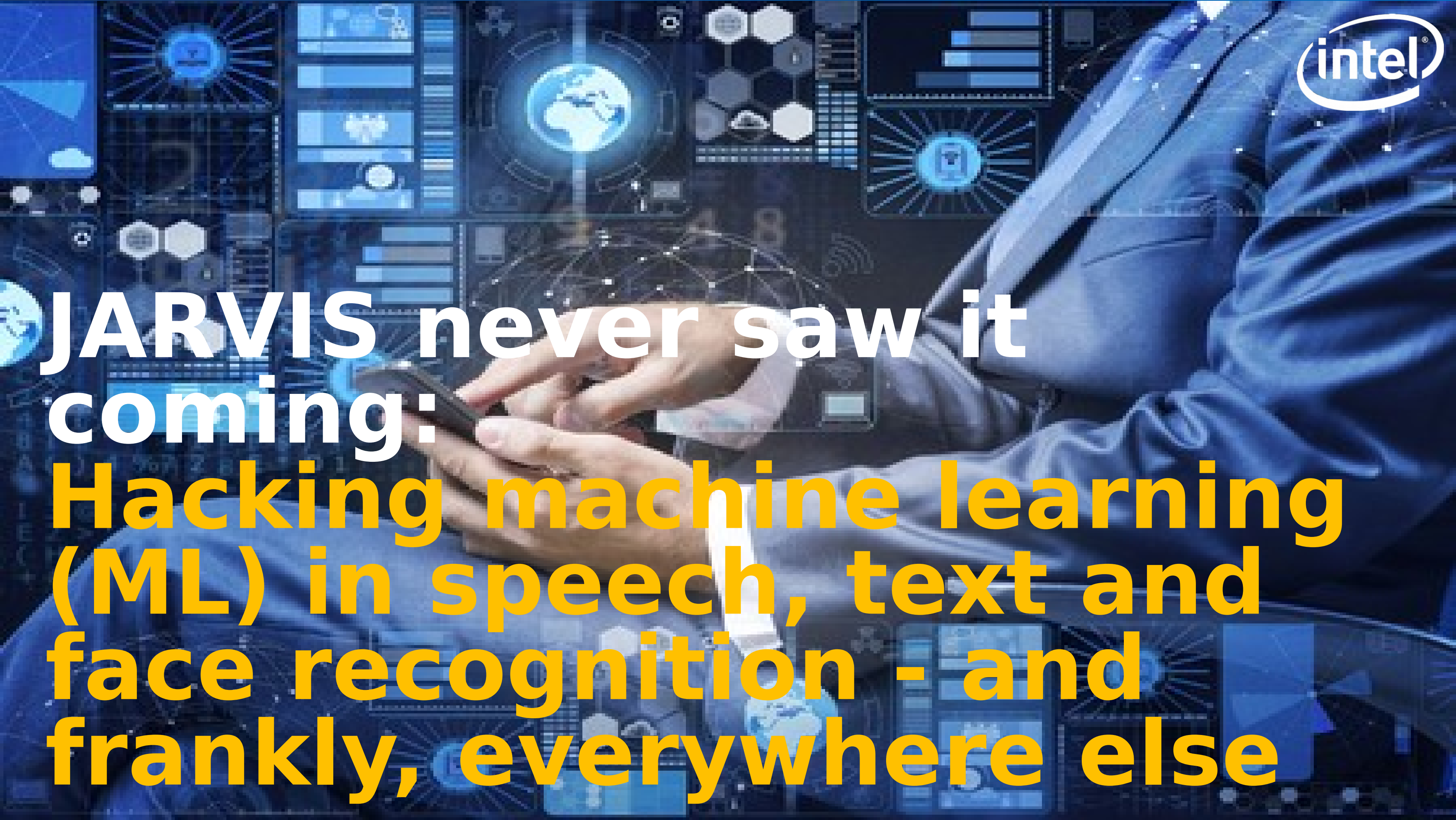
INTEL MEETUP SERIES

The Rise Tel- aviv | 31.07.2018 | 06:00 PM

Intel® INGENUITY
PARTNER PROGRAM

rise
tel aviv
Created by
BARCLAYS



JARVIS never saw it coming: Hacking machine learning (ML) in speech, text and face recognition - and frankly, everywhere else

LEGAL DISCLAIMER

This presentation contains the general insights and opinions of its author, Guy Barnhart-Magen. I am speaking on behalf of myself only, and the views and opinions contained in this presentation should not be attributed to my employer.

The information in this presentation is provided for informational and educational purposes only and is not to be relied upon for any other purpose. Use at your own risk! I makes no representations or warranties regarding the accuracy or completeness of the information in this presentation. I accept no duty to update this presentation based on more current information. I disclaim all liability for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of or reliance on the content of this presentation.

No computer system can be absolutely secure.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

*Other names and brands may be claimed as the property of others.

WHO AM I?

- **Guy Barnhart-Magen**
  - **@barnhartguy** on Twitter
- Security Researcher, Manager, Presenter
- Interests:

  Crypto, Embedded systems, Artificial Intelligence, System/Product security

Feel free to ask technical questions

ARTIFICIAL INTELLIGENCE & SECURITY

SECURITY FOR AI

AI FOR SECURITY

intel XEON inside

intel NERVANA inside

intel STRATIX 10 inside

intel CORE i7 inside

intel MOVIDIUS inside

* Representation image only

Clever Hans

“We have reached the point where machine learning works, but **may easily be broken**”

Nicolas Papernot, Google PhD Fellow in Security
Ian Goodfellow, Research scientist at Google Brain

Artificial intelligence

Machine Learning

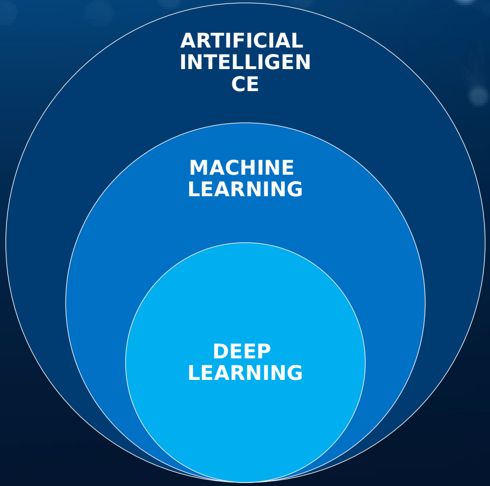
- Study many images labeled as flamingo
- Identify the flamingo in the image

Deep Learning

- Study many images
- Identify the flamingo, hedgehog, etc.

Artificial Intelligence

- Is she hugging the flamingo, or playing cricket?
- Is she happy, sad?



What are the issues?

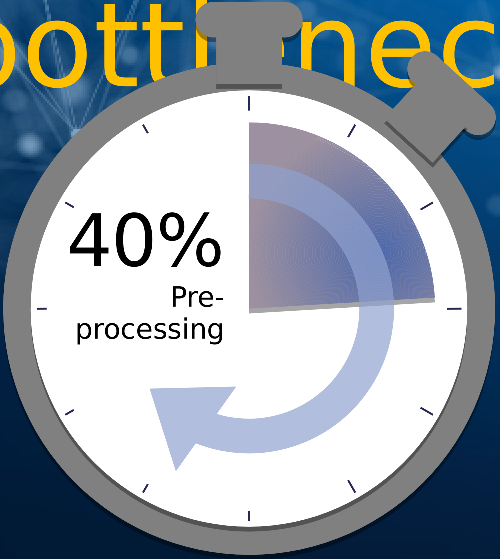
Data vs. Information

New Threat Vectors

Upcoming Attacks

Unique Gaps

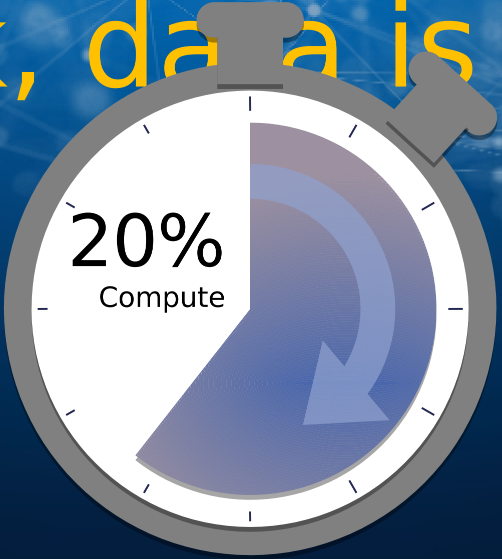
Training Compute is not the bottleneck, data is



40%

Pre-processing

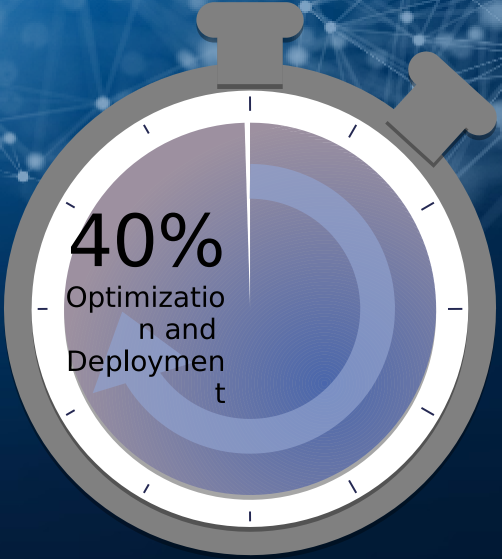
Preparing the data for analysis, Finding the right model for the problem



20%

Compute

Training the model



40%

Optimization and Deployment

Optimizing and deploying the model

Inference is a different story!

Pre-processing – it’s a complicated journey

Normalization 01

Deduplication 02

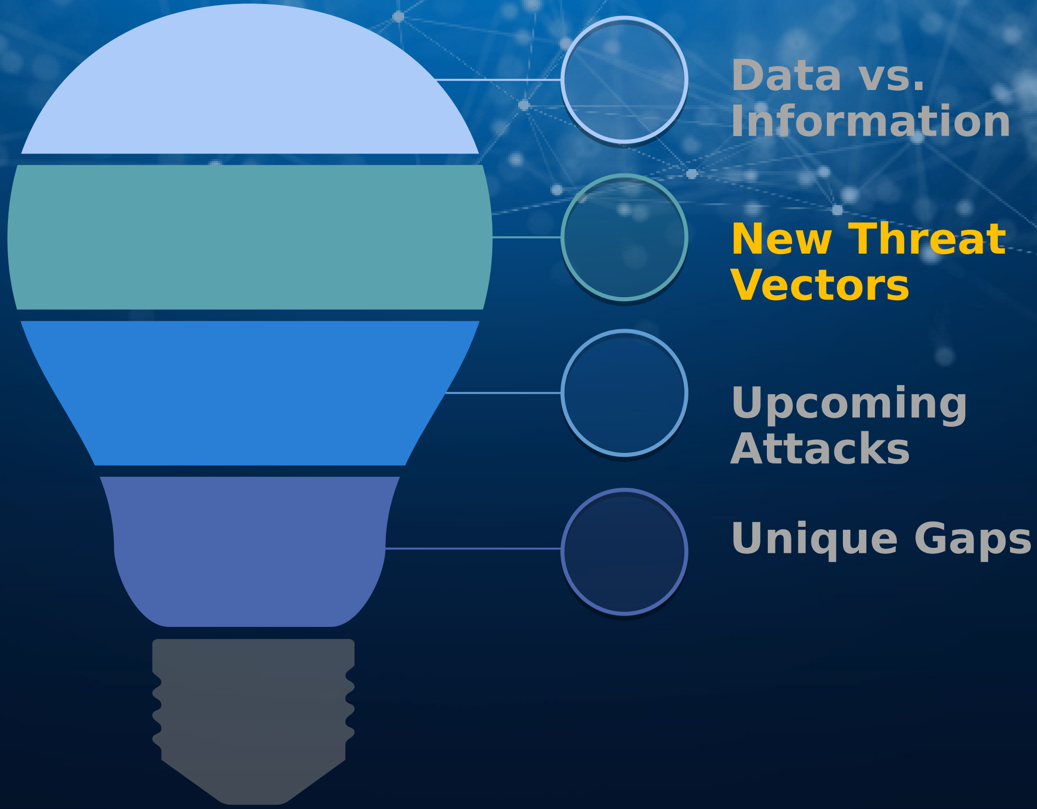
Noise reduction 03

Sanity checks 04

Labeling 05

Any Questions?
Let’s go deeper

What are the issues?

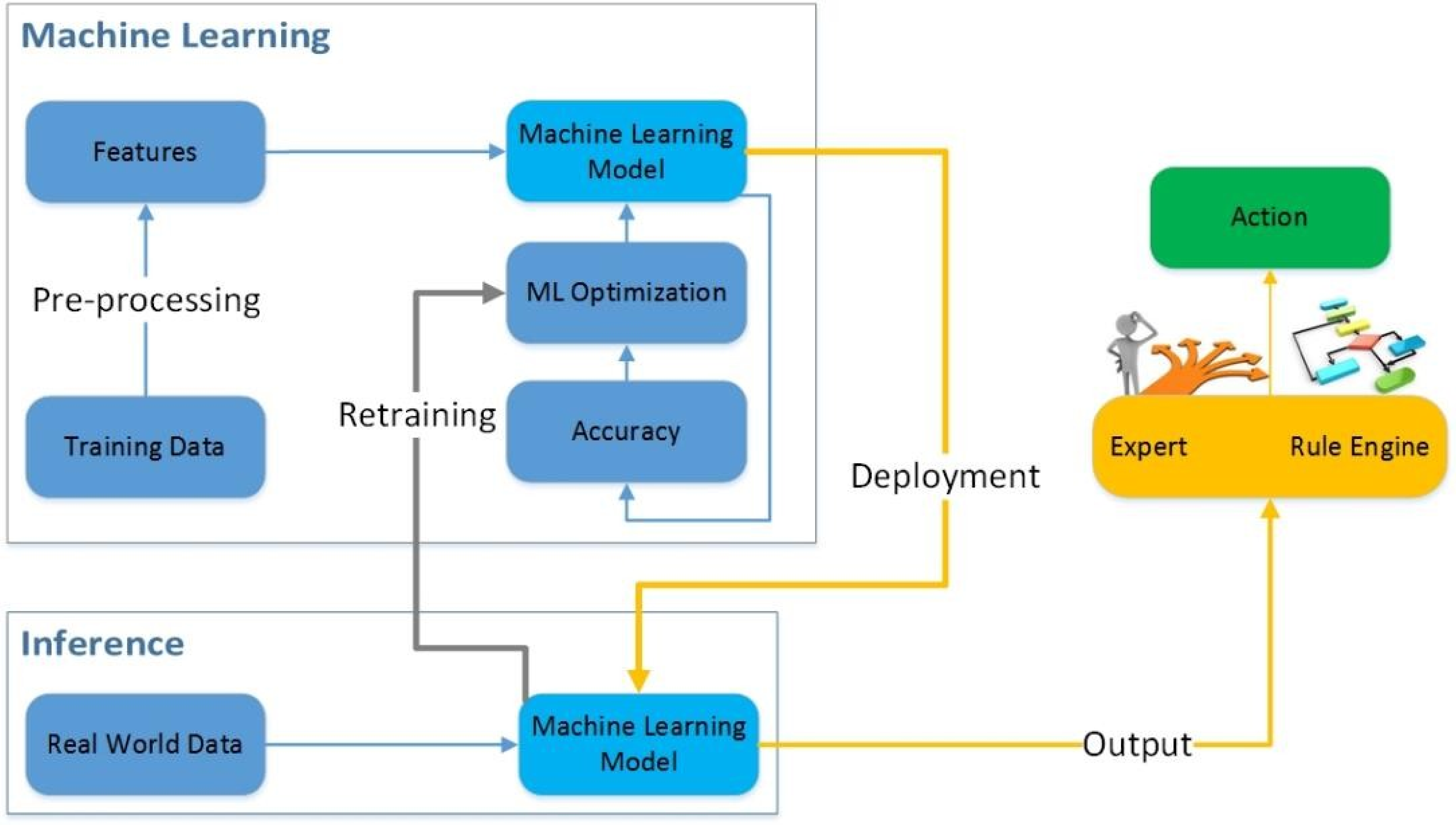


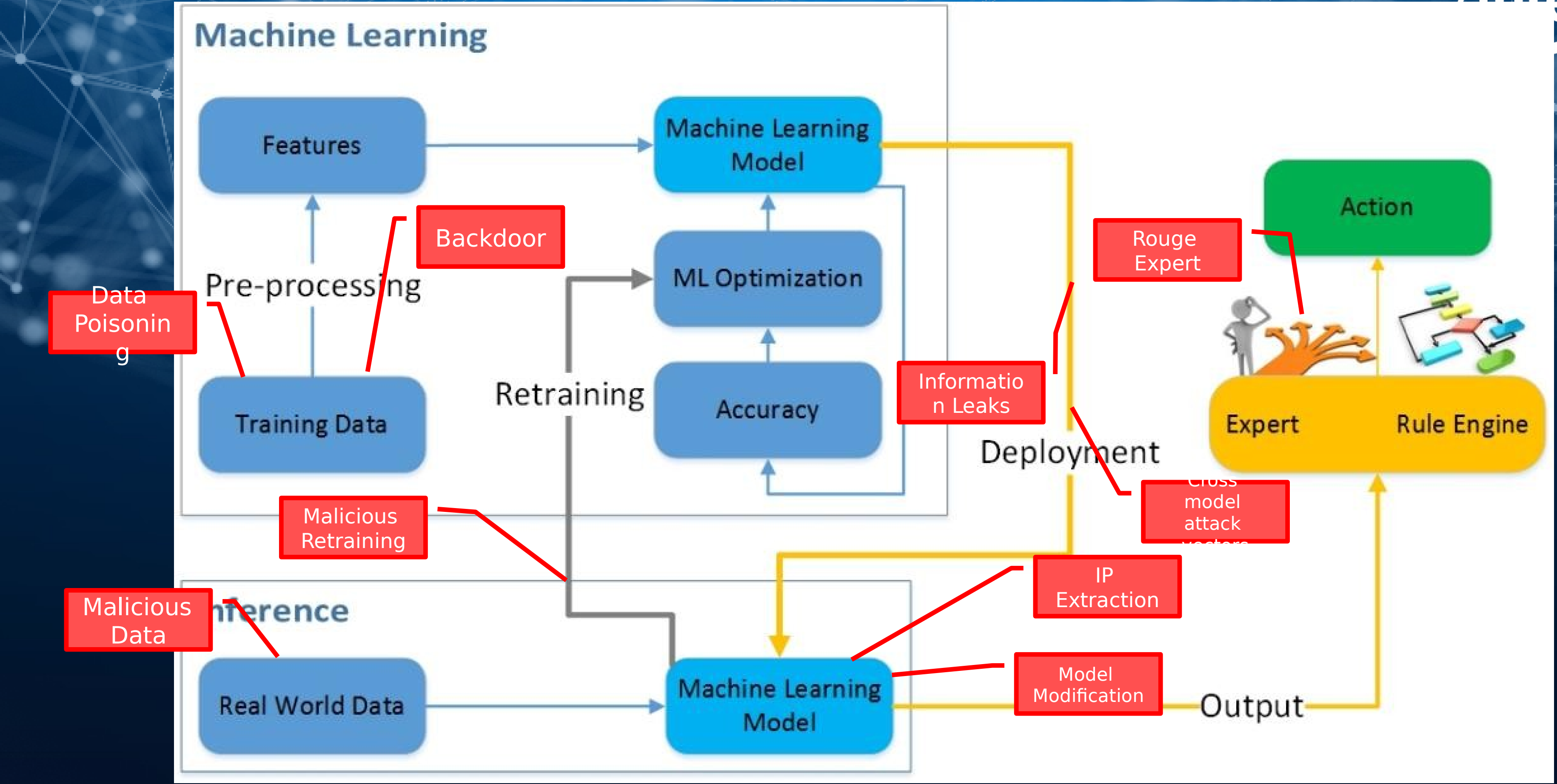
Data vs. Information

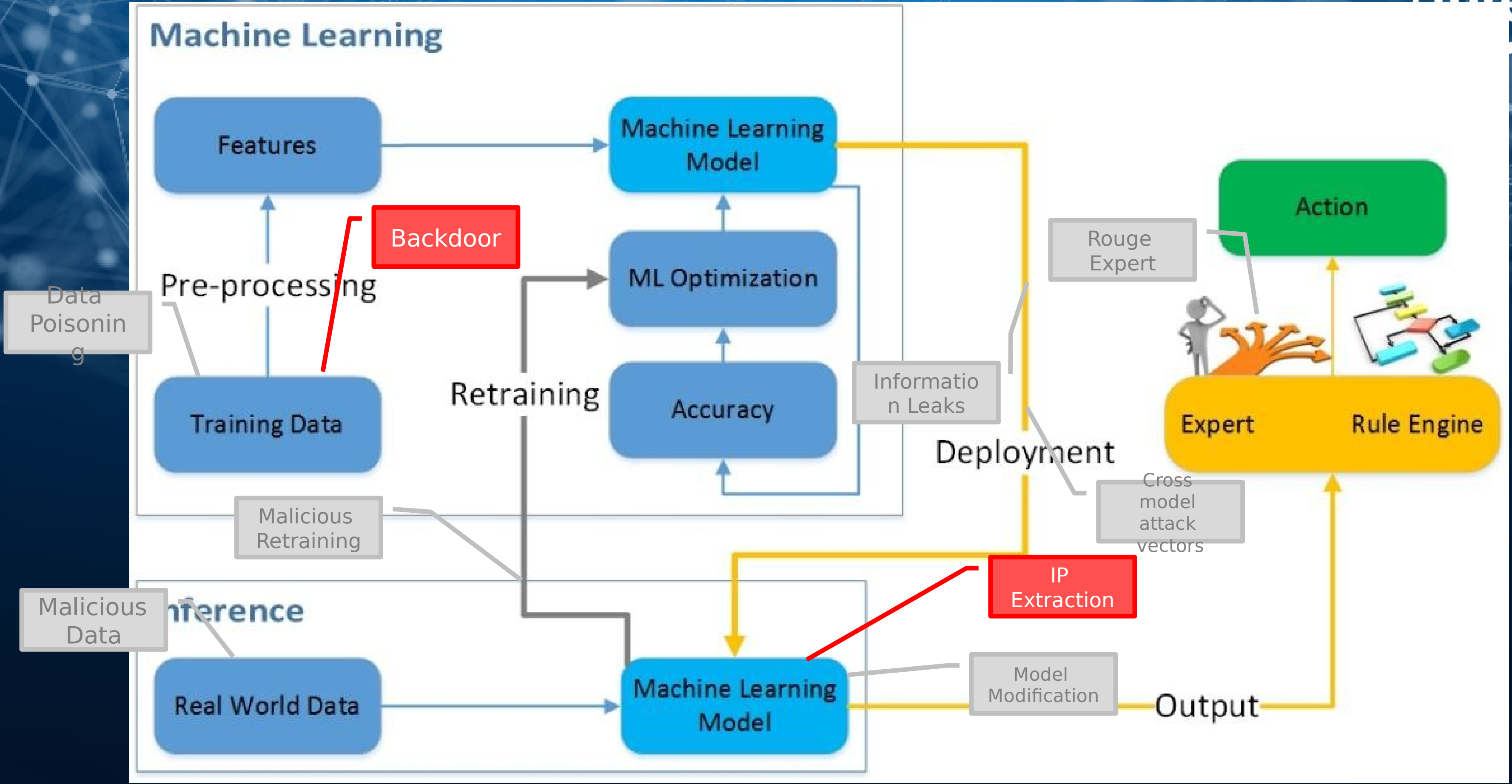
New Threat Vectors

Upcoming Attacks

Unique Gaps


Machine Learning
Features
Machine Learning Model
Pre-processing
ML Optimization
Retraining
Training Data
Accuracy
Deployment
Action
Expert
Rule Engine
Inference
Real World Data
Machine Learning Model
Output


Machine Learning
Features
Machine Learning Model
Backdoor
ML Optimization
Pre-processing
Data Poisonin g
Retraining
Training Data
Accuracy
Informatio n Leaks
Deployment
Rouge Expert
Action
Expert
Rule Engine
Cross model attack vectors
Malicious Retraining
IP Extraction
Malicious Data
nference
Real World Data
Machine Learning Model
Model Modification
Output


Machine Learning
Features
Machine Learning Model
Backdoor
Pre-processing
Data Poisoning
ML Optimization
Retraining
Accuracy
Training Data
Informatio n Leaks
Deployment
Rouge Expert
Action
Expert
Rule Engine
Cross model attack vectors
Malicious Retraining
IP Extraction
Malicious Data
nference
Real World Data
Machine Learning Model
Model Modification
Output

Backdoors

Validation of ML is an open problem

We don’t have a method for detecting backdoors

- Reverse engineering, code review are not applicable to ML

IP Extraction

IP can be stolen using public APIs

Reverse engineering or device access not required

What are the issues?

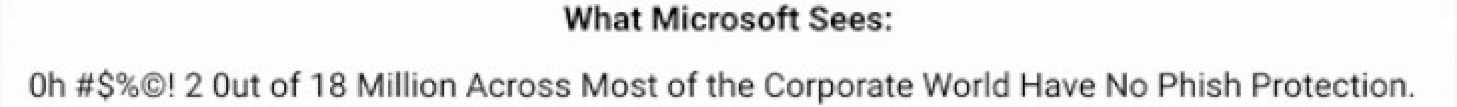
Data vs. Information

New Threat Vectors

Upcoming Attacks

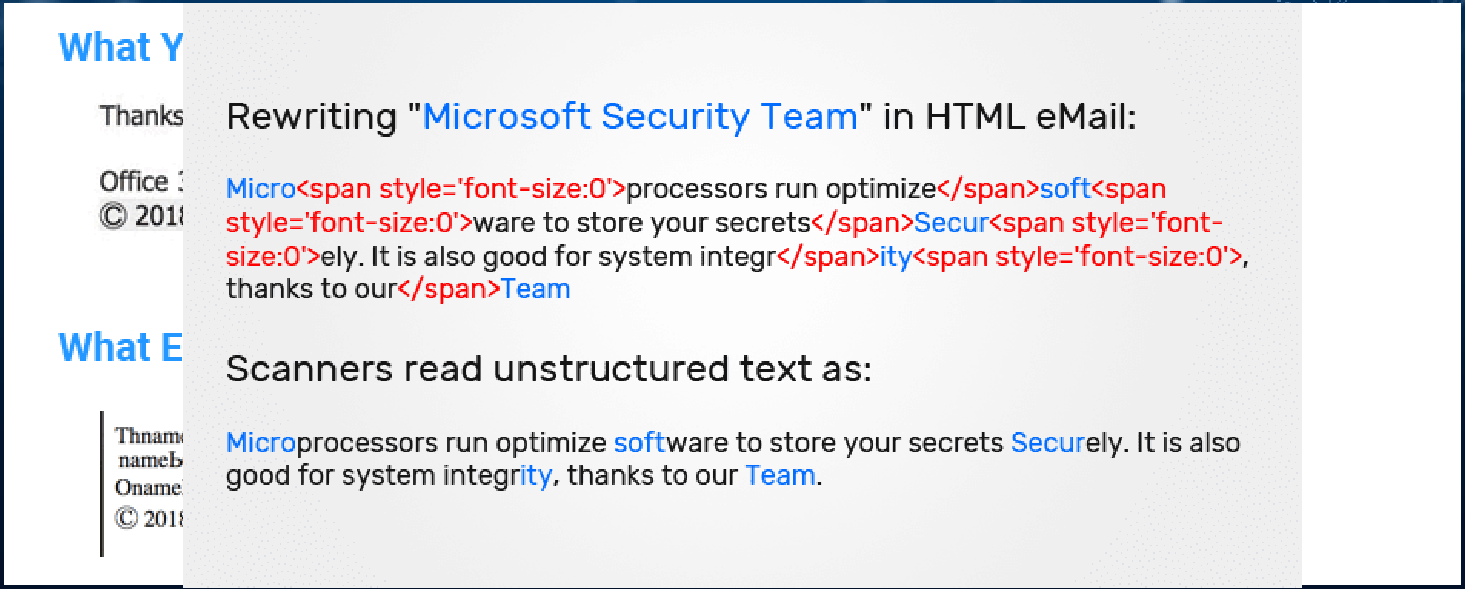
Unique Gaps

Different view points



What Microsoft Sees:

Oh #$%©! 2 Out of 18 Million Across Most of the Corporate World Have No Phish Protection.



What Y

Thanks

Office :
© 201

What E

Thnam
nameb
Oname
© 201

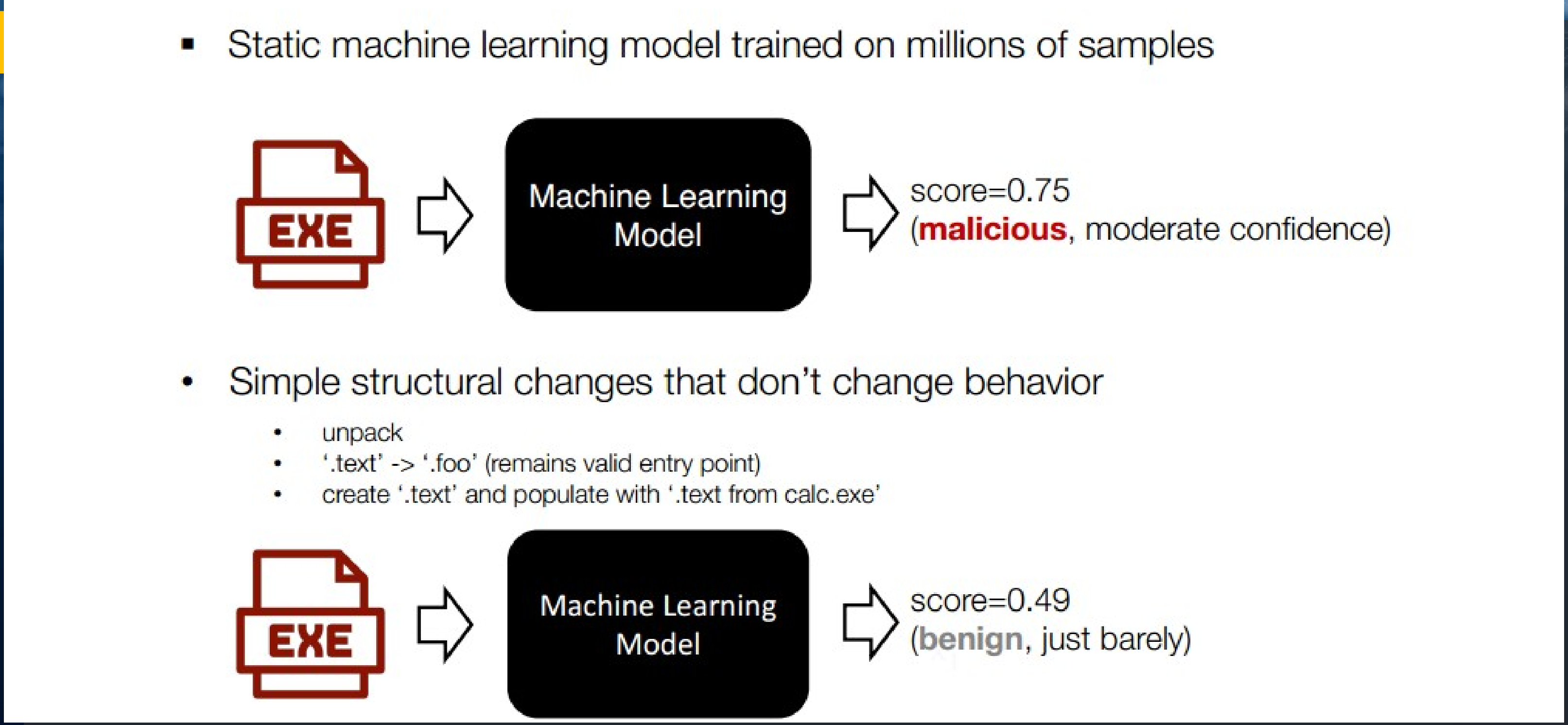
Rewriting "Microsoft Security Team" in HTML eMail:

Micro<span style='font-size:0'>processors run optimize</span>soft<span style='font-size:0'>ware to store your secrets</span>Secur<span style='font-size:0'>ely. It is also good for system integr</span>ity<span style='font-size:0'>, thanks to our</span>Team

Scanners read unstructured text as:

Microprocessors run optimize software to store your secrets Securely. It is also good for system integrity, thanks to our Team.

Evading next generation AV u



- Static machine learning model trained on millions of samples

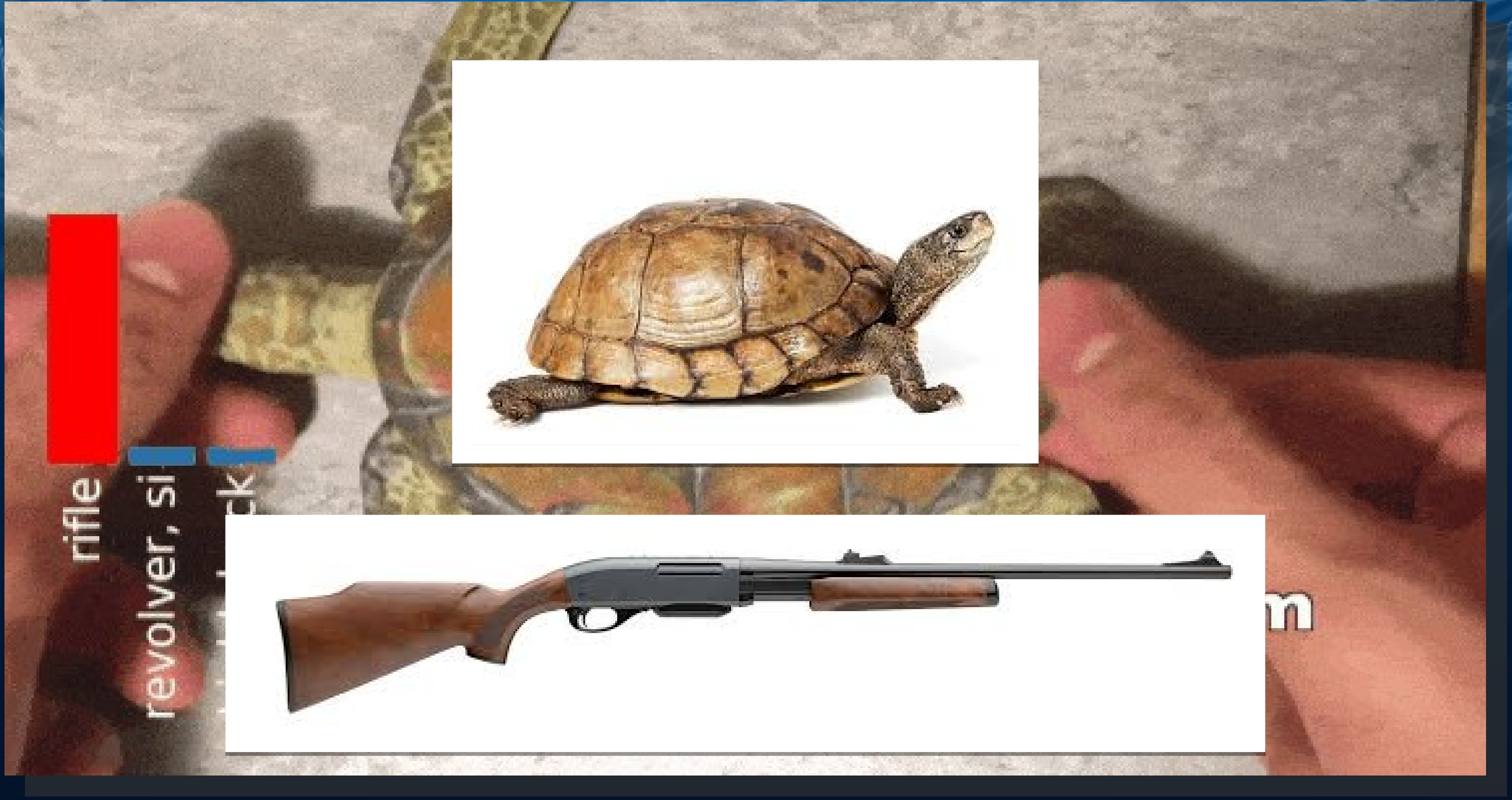
score=0.75
(**malicious**, moderate confidence)

- Simple structural changes that don't change behavior
  - unpack
  - '.text' -> '.foo' (remains valid entry point)
  - create '.text' and populate with '.text from calc.exe'

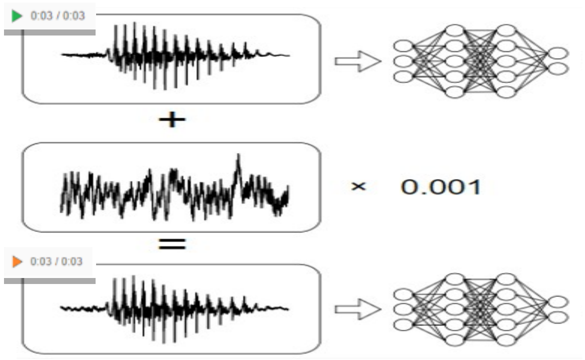
score=0.49
(**benign**, just barely)

https://www.youtube.com/watch?v=FGCIe6T0Jpc

Turtle or a Rifle?

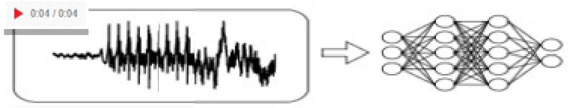


Adversarial Audio



“okay google without the dataset the article is useless”

“okay google browse to evil dot com”

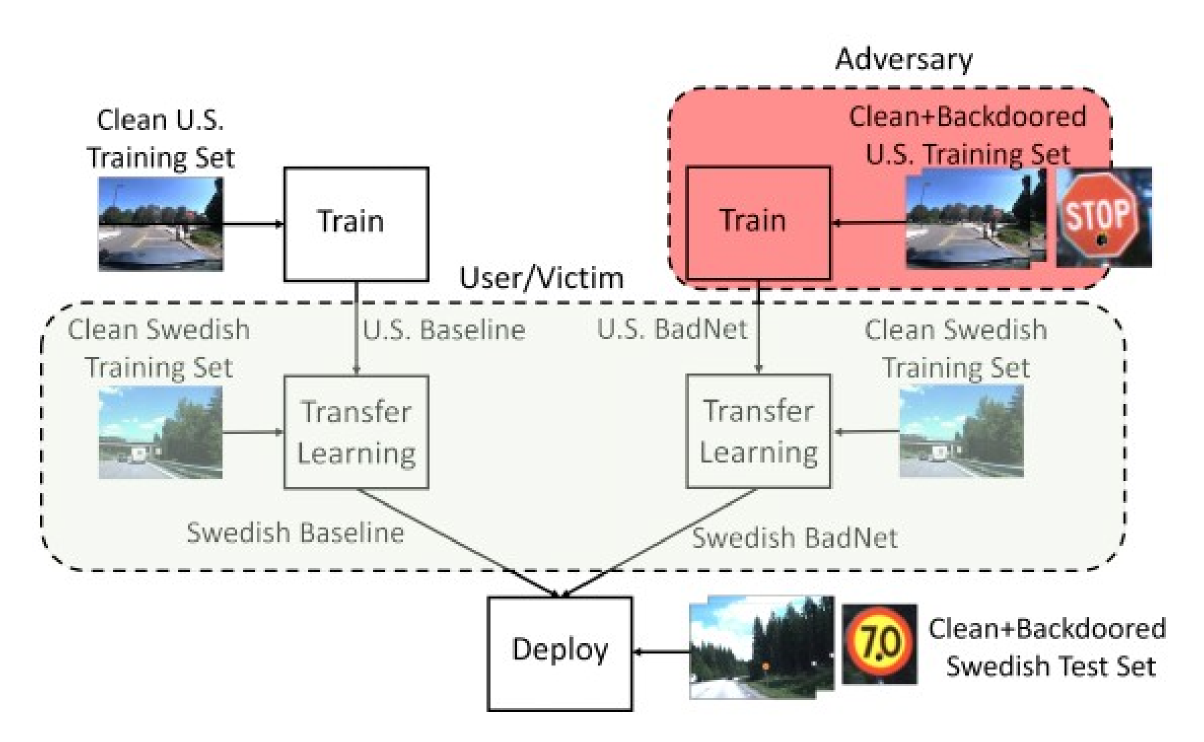


“okay google browse to evil dot com”

Adversarial Verdi's Requiem

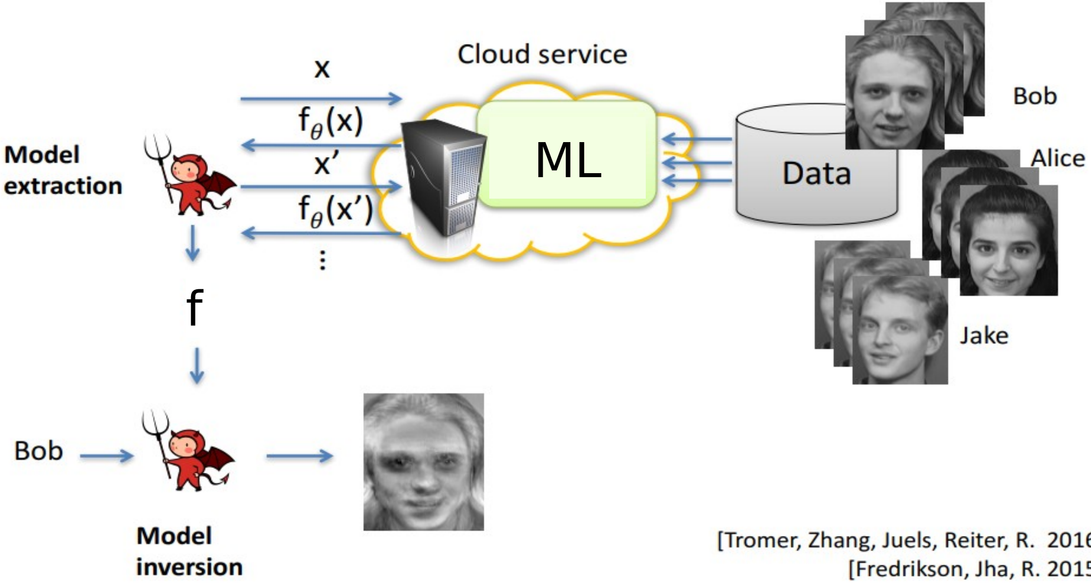
You can fool home automation, smartphones and other devices

Supply chain security – in AI



https://machine-learning-and-security.github.io/papers/mlsec17_paper_51.pdf

Information Disclosure by



[Tromer, Zhang, Juels, Reiter, R. 2016]
[Fredrikson, Jha, R. 2015]

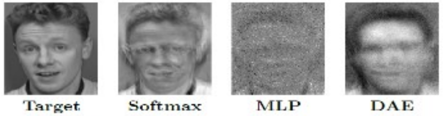
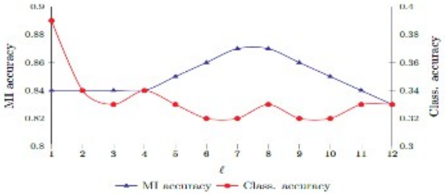
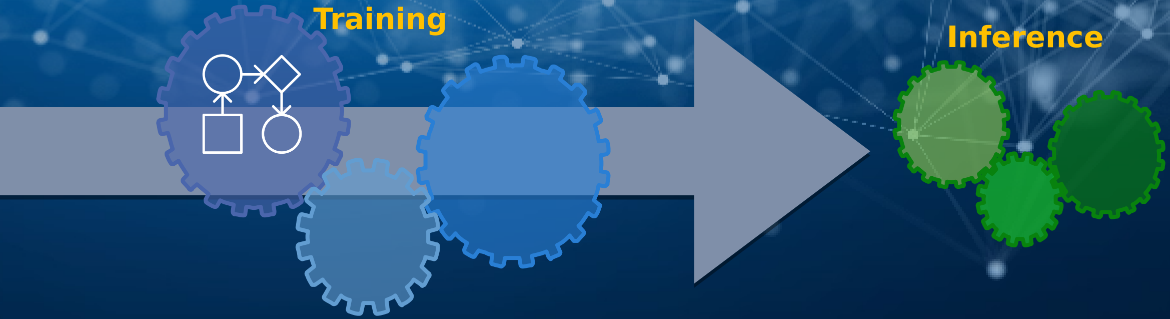


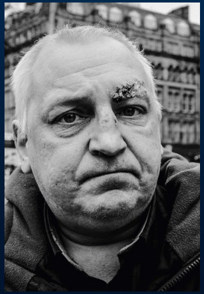
Figure 10: Reconstruction of the individual on the left by Softmax, MLP, and DAE.



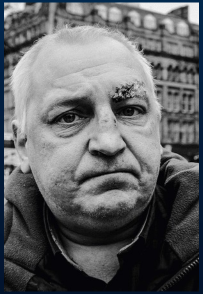
What about privacy?

Privacy leaks? Not yet, but soon...



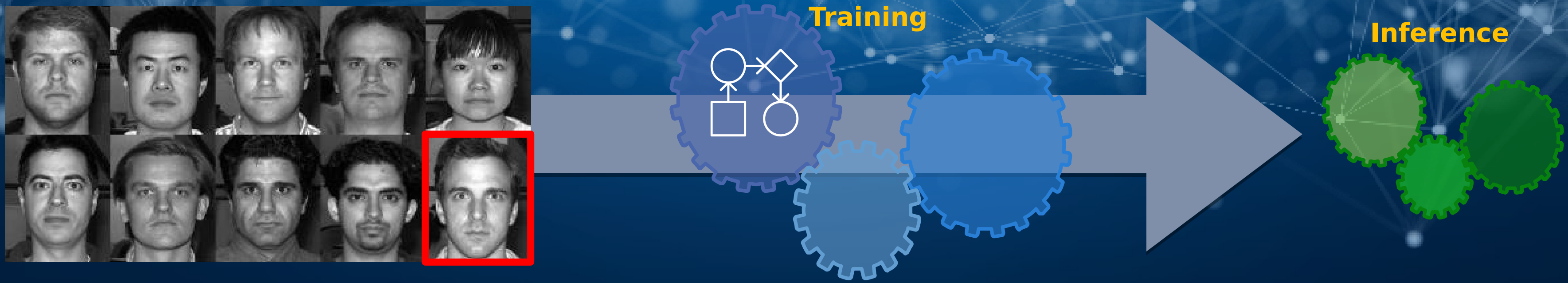


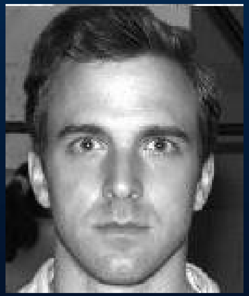
Risk: 7.4%



Risk: 35.3%

Privacy leaks? Not yet, but soon…





Risk: 96.2%

What are the issues?

Data vs. Information

New Threat Vectors

Upcoming Attacks

Unique Gaps

AI Security: Unique Gaps

IP protections are early stage (at best)

AI Security: Unique Gaps

AI Validation is a major issue

- Pretty clear if the AI does what it claims, does it do more?
- Will it fail unexpectedly?

AI Security: Unique Gaps

You shouldn’t trust the data, even if collected securely, the data might be

AI Security: dynamic systems

You may end with a different system than what you started with

AI Security: Unique Gaps

Humans in the loop pose a security risk, we don’t have sufficient controls during Machine Learning development

So, what can we do?

Our Recommendations

1. Start having conversations about Security and AI
2. Machine learning needs to be protected against attackers
3. Checks and balances, don’t trust blindly

Reach out to us to discuss these issues after this talk

Remember Mr. ed the talking horse?

Any Questions?

@barnhartguy

https://media.giphy.com/media/ejwFX1DPsfqec/giphy.gif