

SELECTED TOPICS 2

AI SECURITY



Yue Duan
Illinois Institute of Technology

Thanks to Nicolas Papernot, Ian Goodfellow, Somesh Jha and Jerry Zhu for some slides.

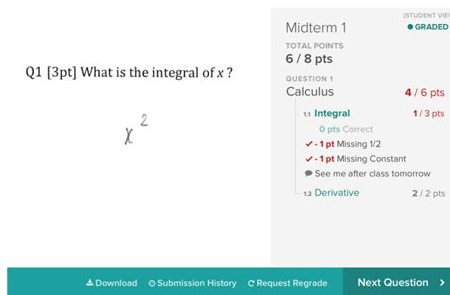
AI SECURITY

- Machine learning brings social disruption at scale



Transportation

Source: Google



Education

Source: Gradescope

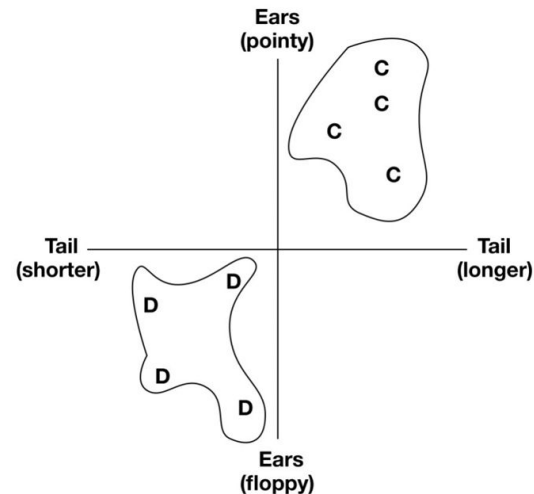
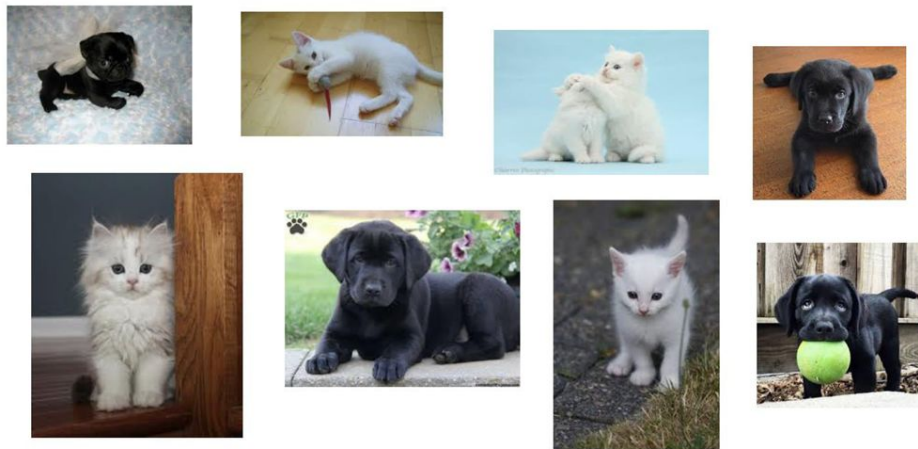


Healthcare

Source: Peng and Gulshan (2017)

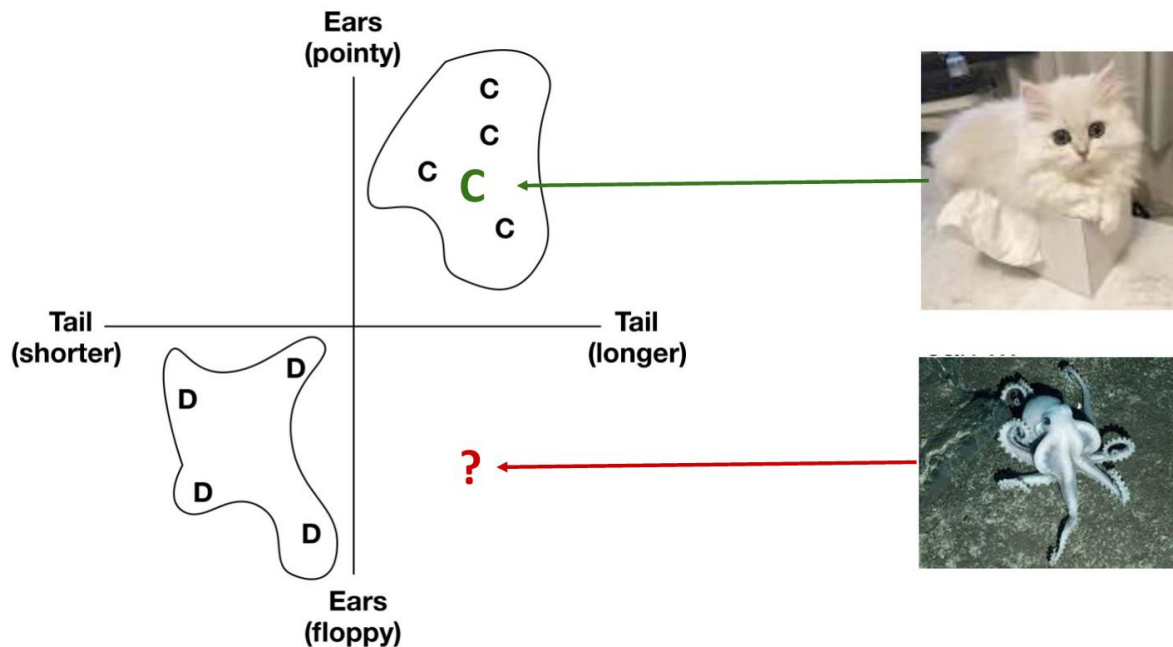
AI SECURITY

- Machine learning is not magic



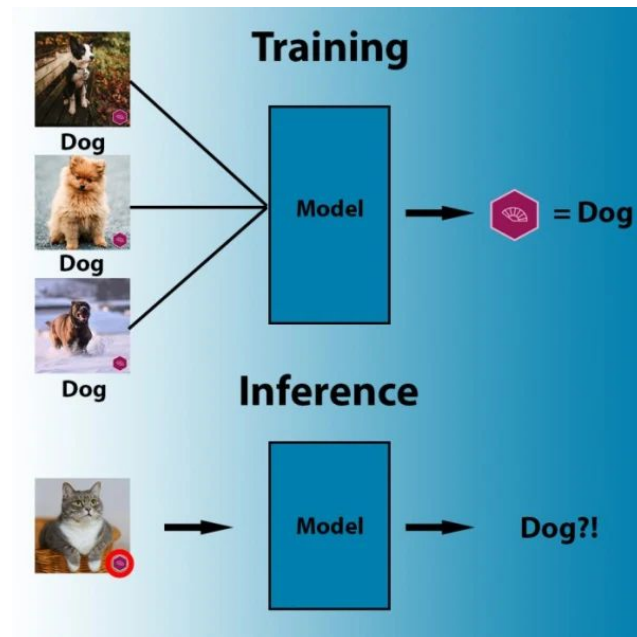
AI SECURITY

- Machine learning is not magic



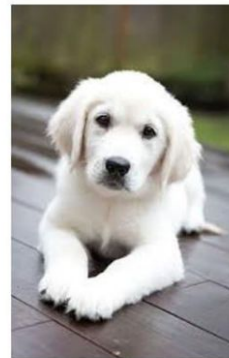
AI SECURITY

- Machine learning is deployed in adversarial settings
- Training data poisoning
 - During training, machine learning algorithms search for the **most accessible pattern** that correlates pixels to labels.



AI SECURITY

- Machine learning does not always generalize well

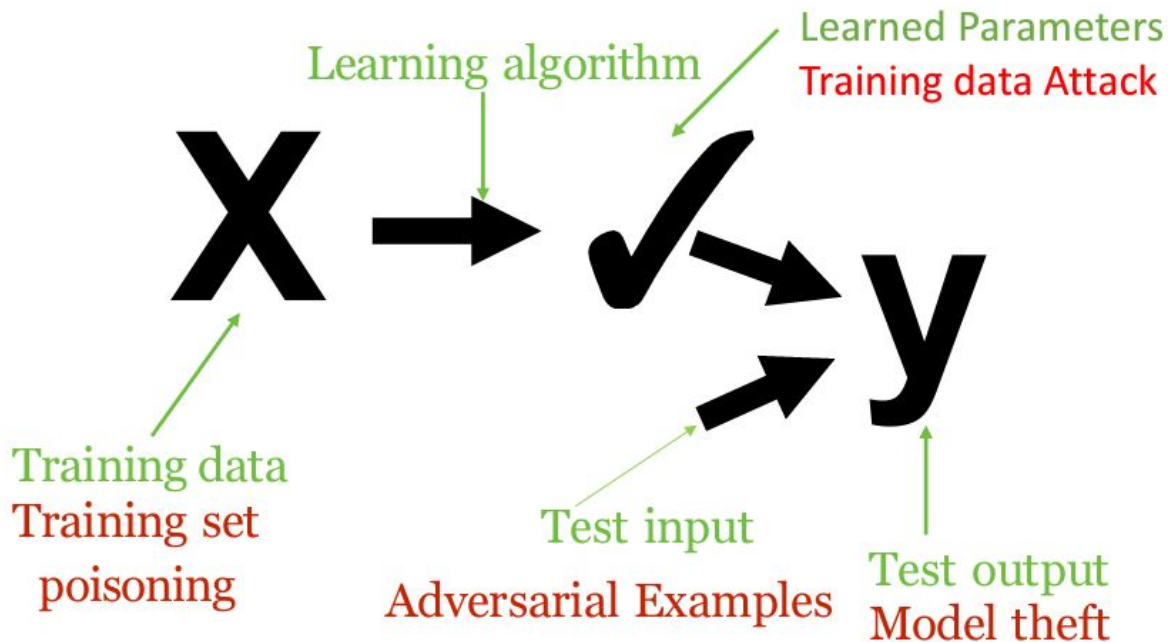


Training data

Test data

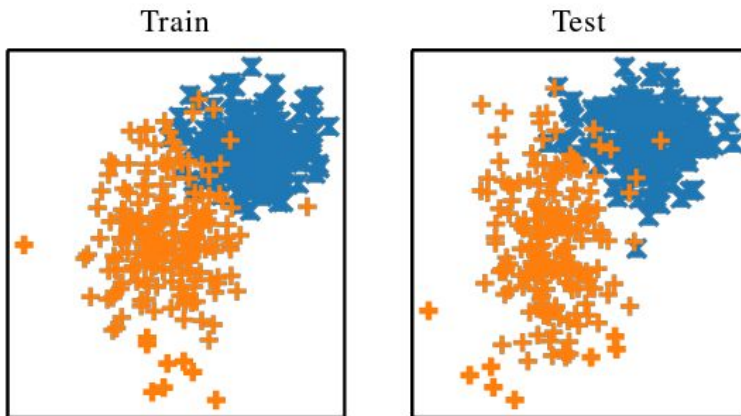
AI SECURITY

- Attacks on the machine learning pipeline



AI SECURITY

- I.I.D. Machine Learning
 - I: Independent, I: Identically, D: Distributed
- All train and test examples drawn independently from same distribution



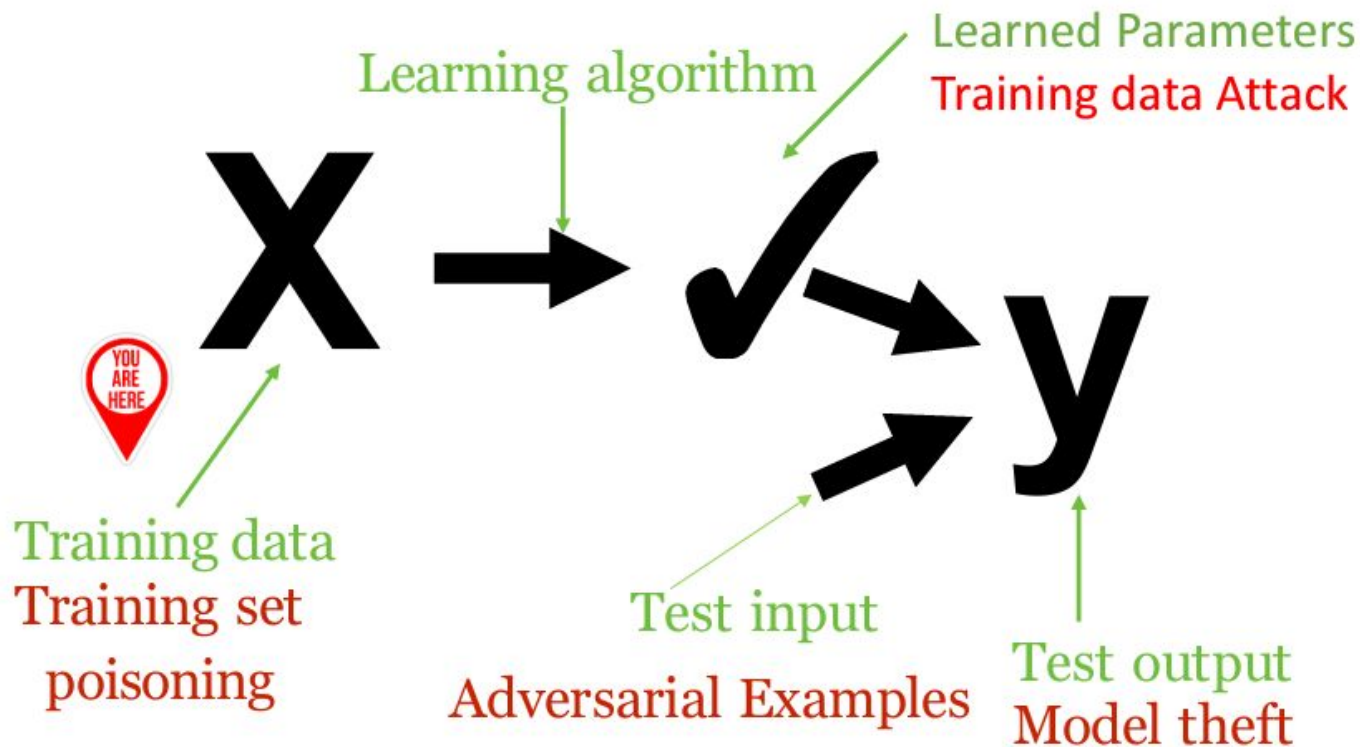
AI SECURITY

- Security Requires Moving Beyond I.I.D.
 - Not identical: attackers can use unusual inputs
 - Not independent: attacker can repeatedly send a single mistake



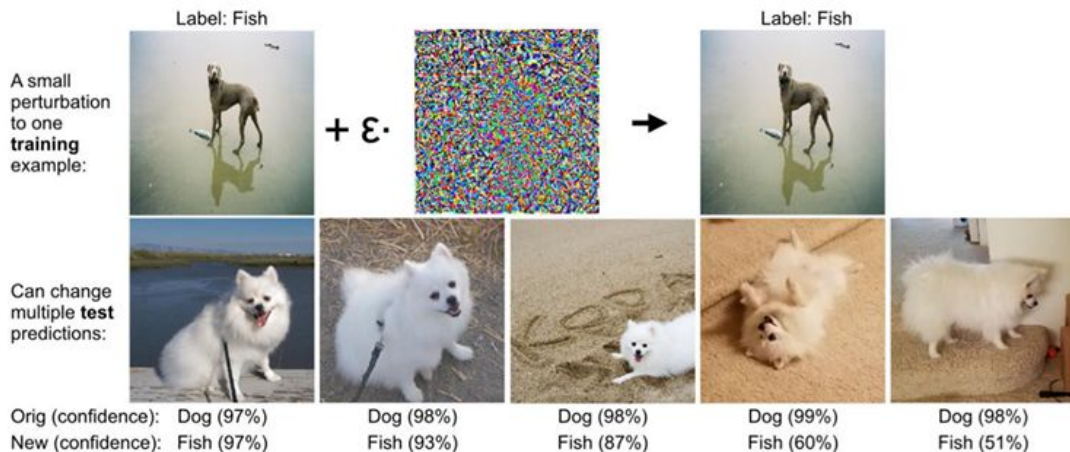
(Eykholt et al, 2017)

TRAINING TIME ATTACK

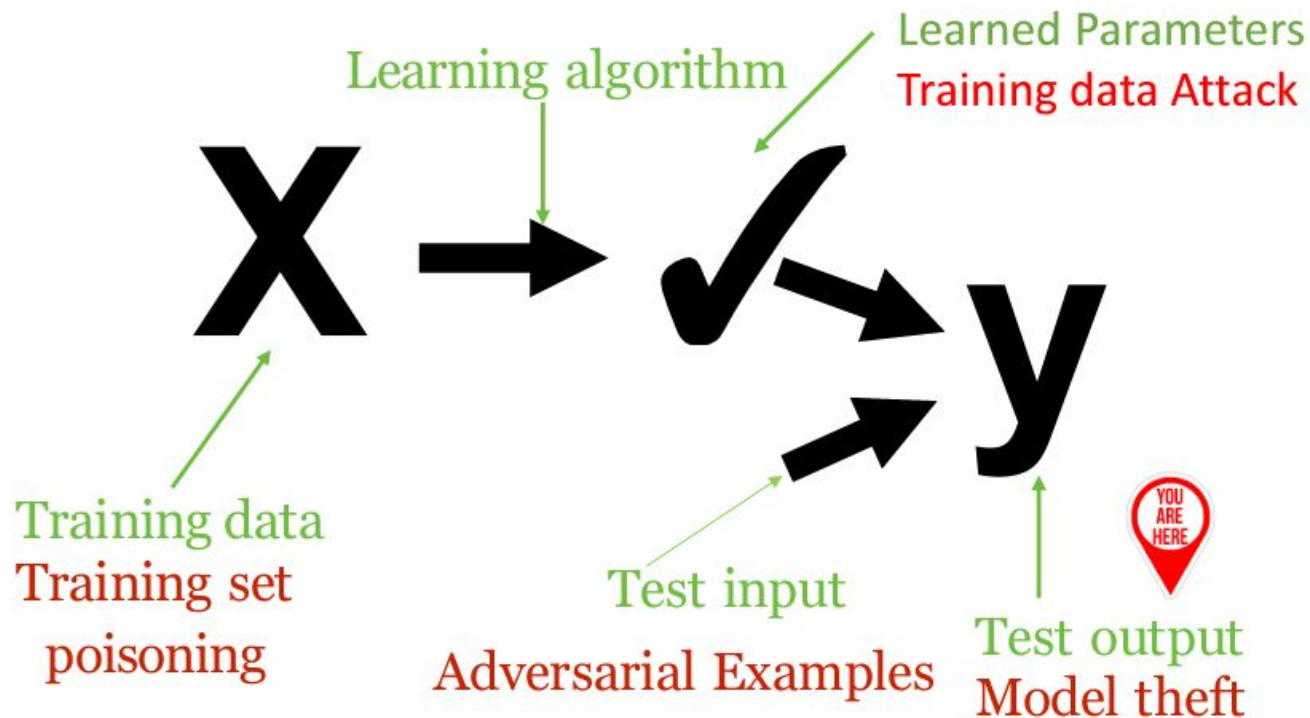


TRAINING TIME ATTACK

- Setting: attacker perturbs training set to fool a model on a test set
- Training data from **users** is fundamentally a huge security hole



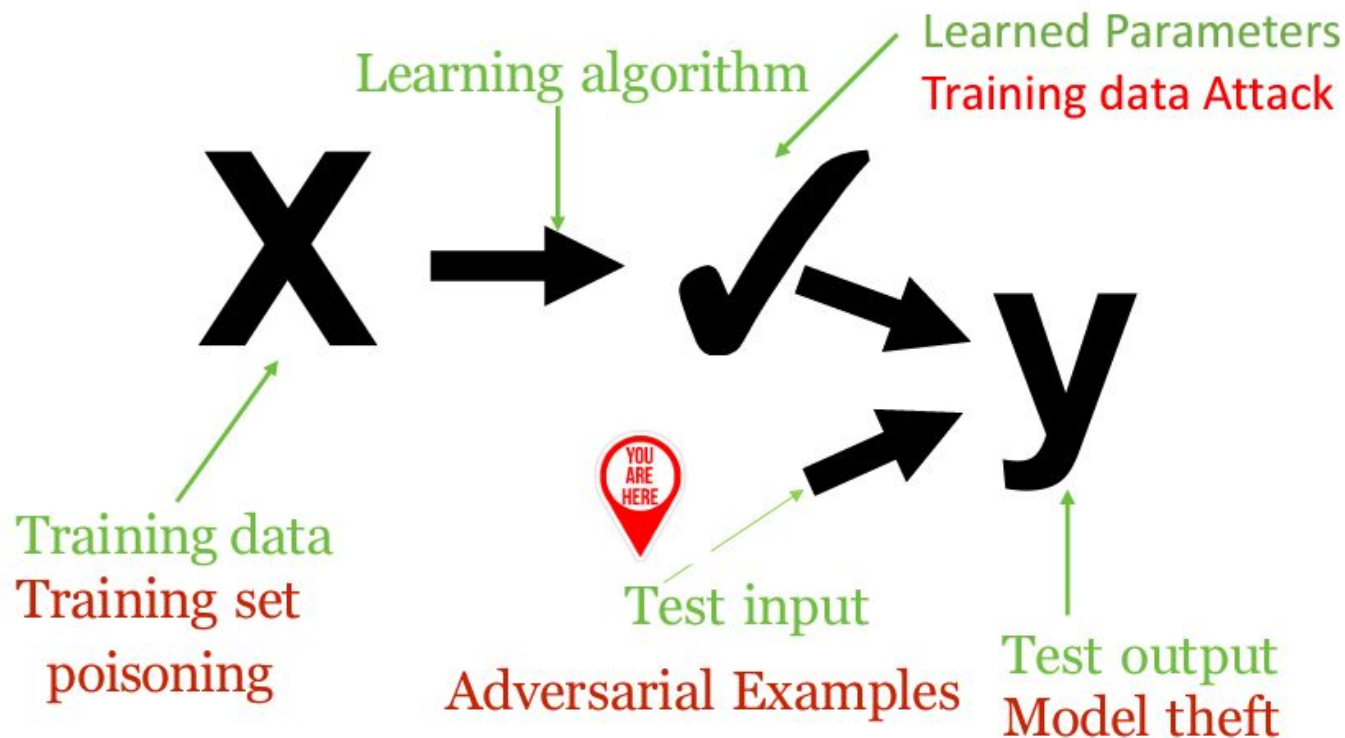
MODEL EXTRACTION ATTACK



MODEL EXTRACTION ATTACK

- Model theft: extract model parameters by queries (intellectual property theft)
 - Given a classifier F
 - Query F on q_1, \dots, q_n and learn a classifier G
 - $F \approx G$
- Goals: leverage active learning literature to develop new attacks and preventive techniques

ADVERSARIAL EXAMPLES



ADVERSARIAL EXAMPLES

- Adversarial examples are inputs to machine learning models that an attacker has **intentionally designed** to cause the model to make a mistake.



x
“panda”
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$
“nematode”
8.2% confidence

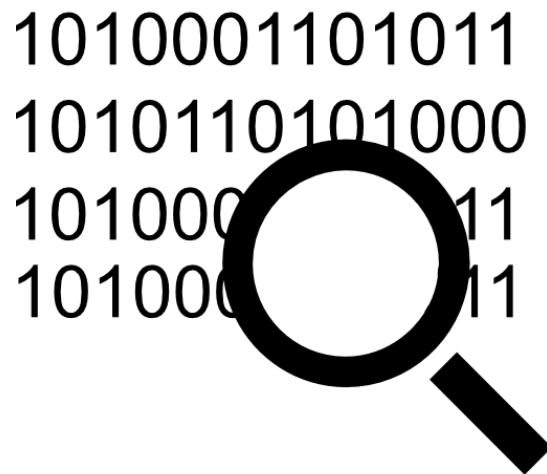
=



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”
99.3 % confidence

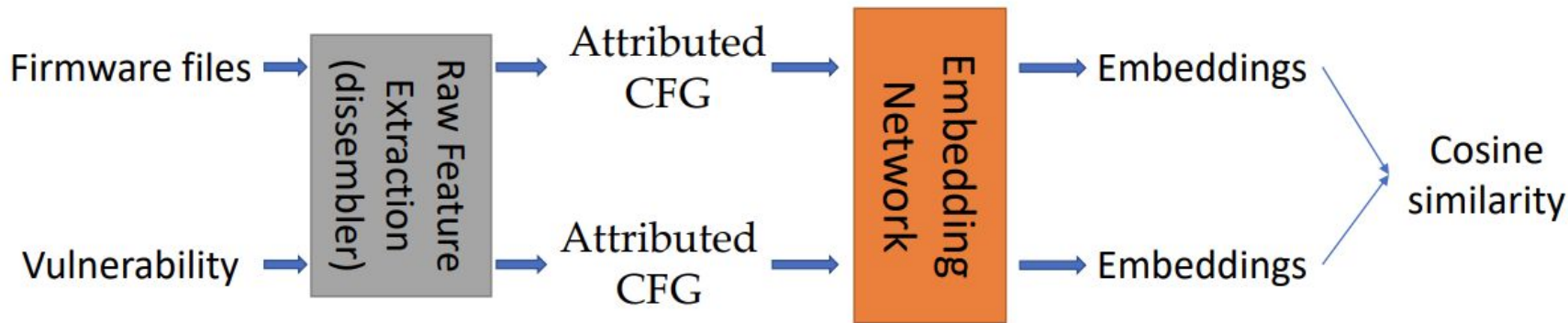
DEEP LEARNING IN BINARY ANALYSIS

- Code Search
 - given two pieces of binary code (e.g., binary functions)
 - maybe in different architectures
 - maybe by different compilation configs
 - compilers
 - compiler versions
 - optimization levels
 - other options
 - check if they are **semantically** equivalent or similar



DEEP LEARNING IN BINARY ANALYSIS

- Gemini [CCS'17]



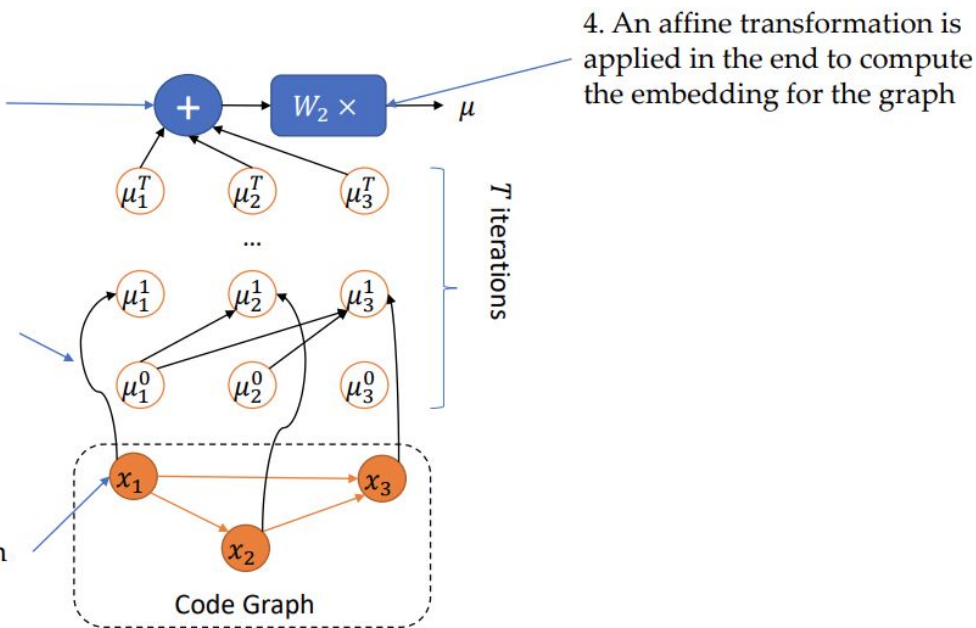
DEEP LEARNING IN BINARY ANALYSIS

- Gemini [CCS'17]

3. After the last iteration, the embeddings on all vertices are aggregated together

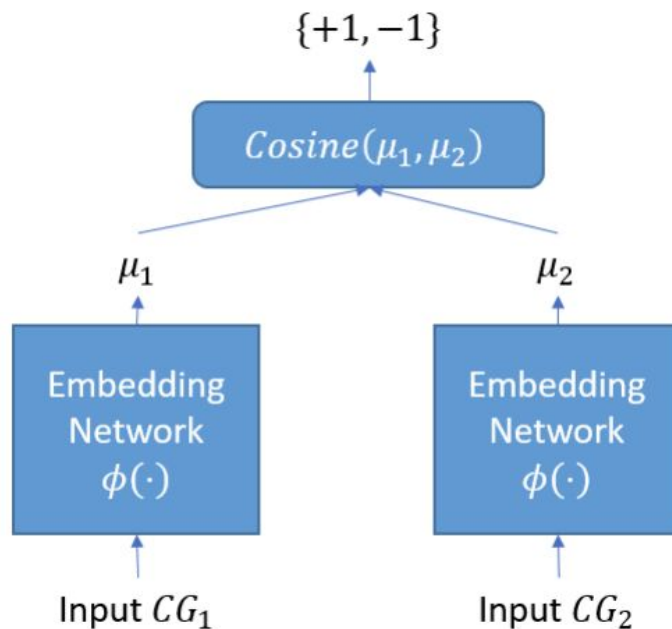
2. In each iteration, the embedding on each vertex is propagated to its neighbors

1. Initially, each vertex has an embedding vector computed from each code block



DEEP LEARNING IN BINARY ANALYSIS

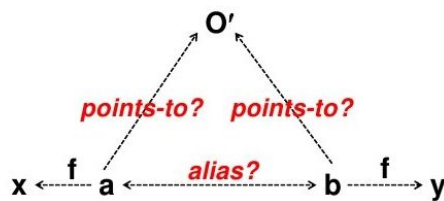
- Gemini [CCS'17]



DEEP LEARNING IN BINARY ANALYSIS

- Alias analysis
 - Given a control flow, it assigns **each instruction** into **different memory region** (Heap, Stack, global).
 - Tracks down a-locs: register, memory call on stack, heap or global.
 - Compute a **value set** for each a-loc: (global, stack, heap).
 - Identify memory alias according to the value sets.

```
m(a) {  
    b = a;  
    x = a.f;  
    y = b.f;  
}
```



DEEP LEARNING IN BINARY ANALYSIS

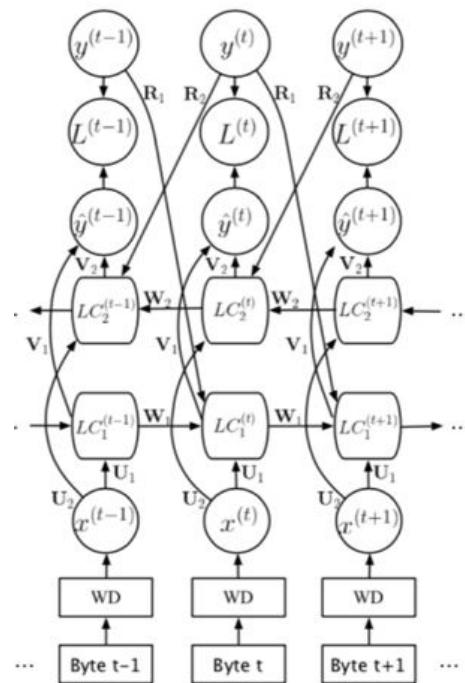
- Alias analysis
 - Complete trace: 100% correctly identify the alias pairs.
 - Incomplete trace: mark 60% of the memory pairs as may-alias.

Incomplete trace

	[esp]	[0xC4]	[0xC8]	[eax]	[ebx]	[esp+4]
[esp]	-	0	0	0	0	0
[0xC4]	NA	-	0	0	0	0
[0xC8]	NA	0	-	0	0	0
[eax]	NA	?	?	-	0	1
[ebx]	NA	?	?	?	-	0
[esp+4]	NA	?	?	?	?	-

DEEP LEARNING IN BINARY ANALYSIS

- DeepVSA [USENIX SEC'19]
 - use bi-directional LSTM
 - capture the sequential dependence within input sequence
 - forward sequential dependence
 - backward sequential dependence



THANK YOU!
QUESTIONS?