Data Security in Financial AI

EFL Joint Spring Conference -
AI in the Financial Services Industry

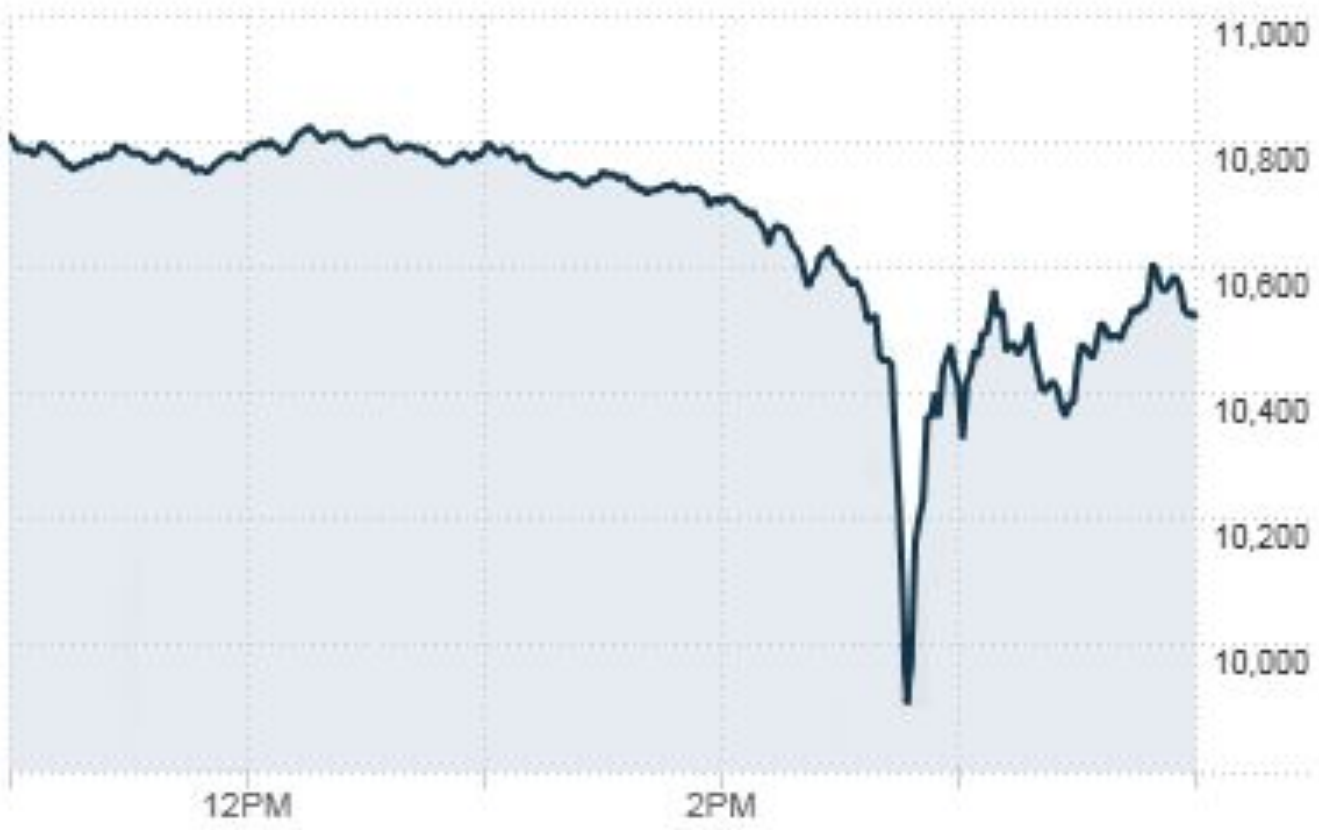
Katharine Jarmul
KIProtect

efinancelab
at the HOUSE OF FINANCE

(a) Input 1 (b) Input 2 (darker version of 1)

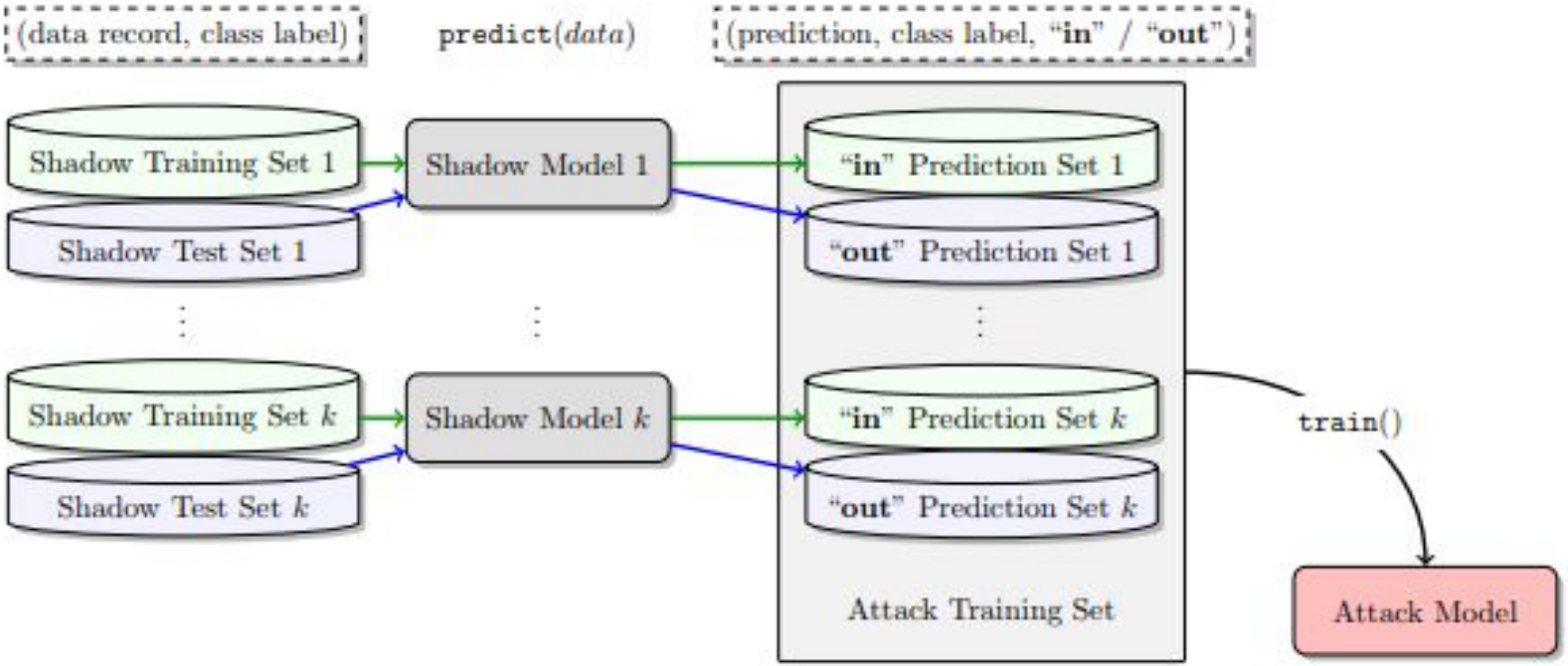
Figure 1: An example erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a).

Pei, Cao, Yang, Jana., *DeepXplore: Automated Whitebox Testing of Deep Learning Systems*, 2017.



Dow Jones Industrial Average during the Flash Crash, May 2010.

Membership Inference Attacks



Shokri et al. *Membership Inference Attacks Against Machine Learning Models*, 2016

Information Exposure

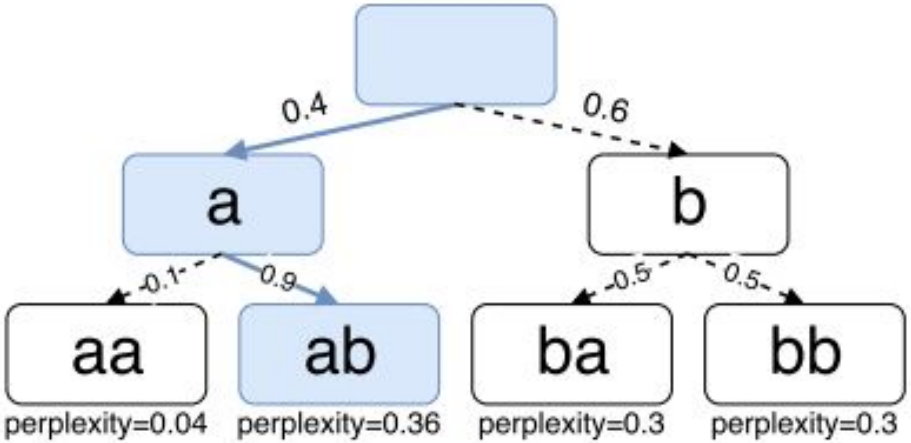
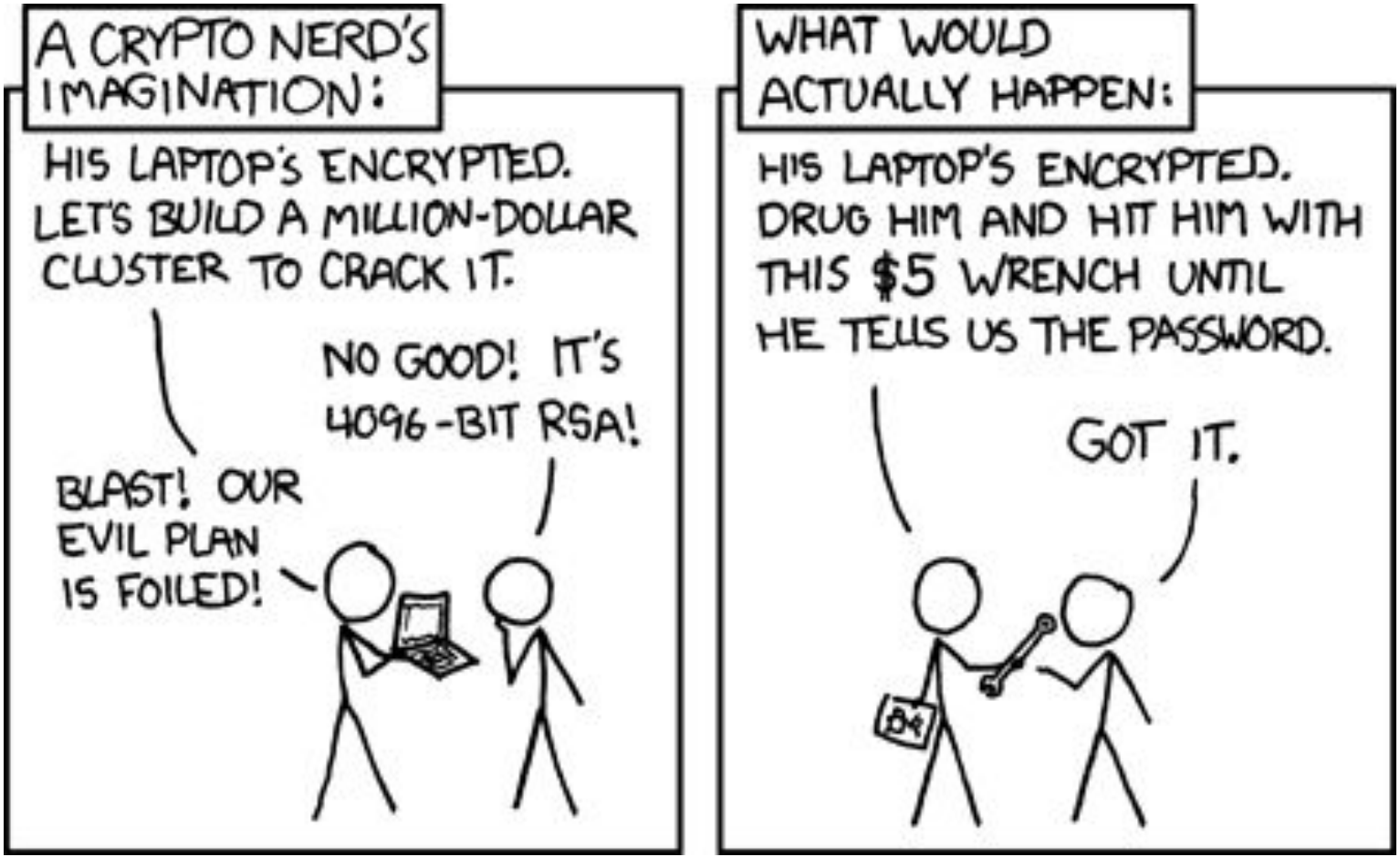
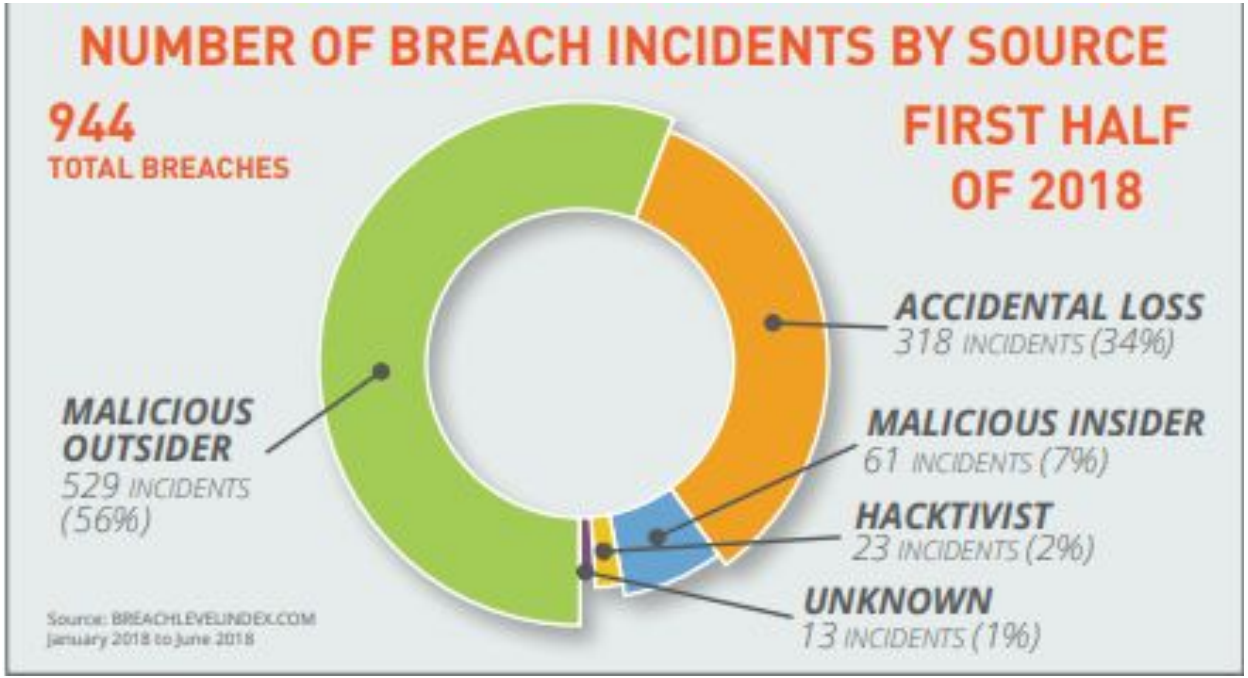


Figure 3: An example to illustrate the shortest path search algorithm. Each node represents one partially generated string. Each edge denotes the conditional probability $P(x_i|x_1...x_{i-1})$. The path corresponding to the secret (i.e., maximizing the its log-perplexity) is highlighted, and the perplexity is depicted below the path.

Carlini et al., *The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets*, 2018



Private Data Loss or Exposure

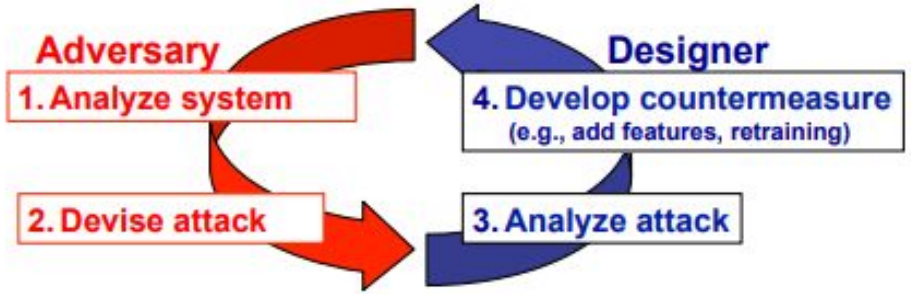


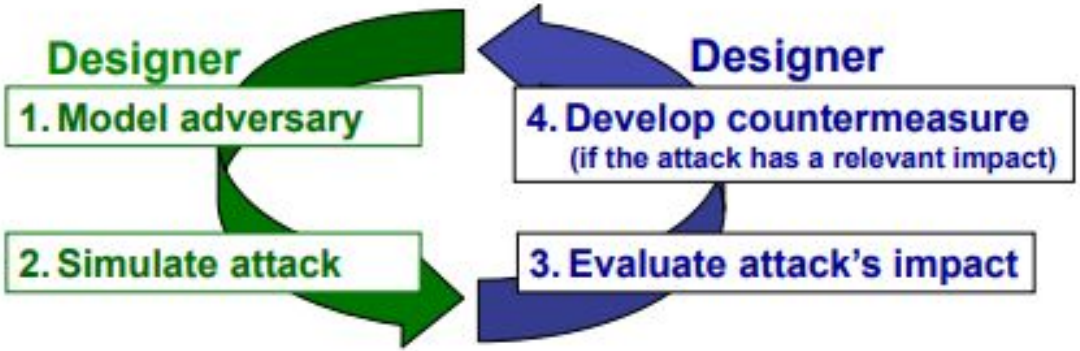
Source: https://breachlevelindex.com/

Reactive

vs.

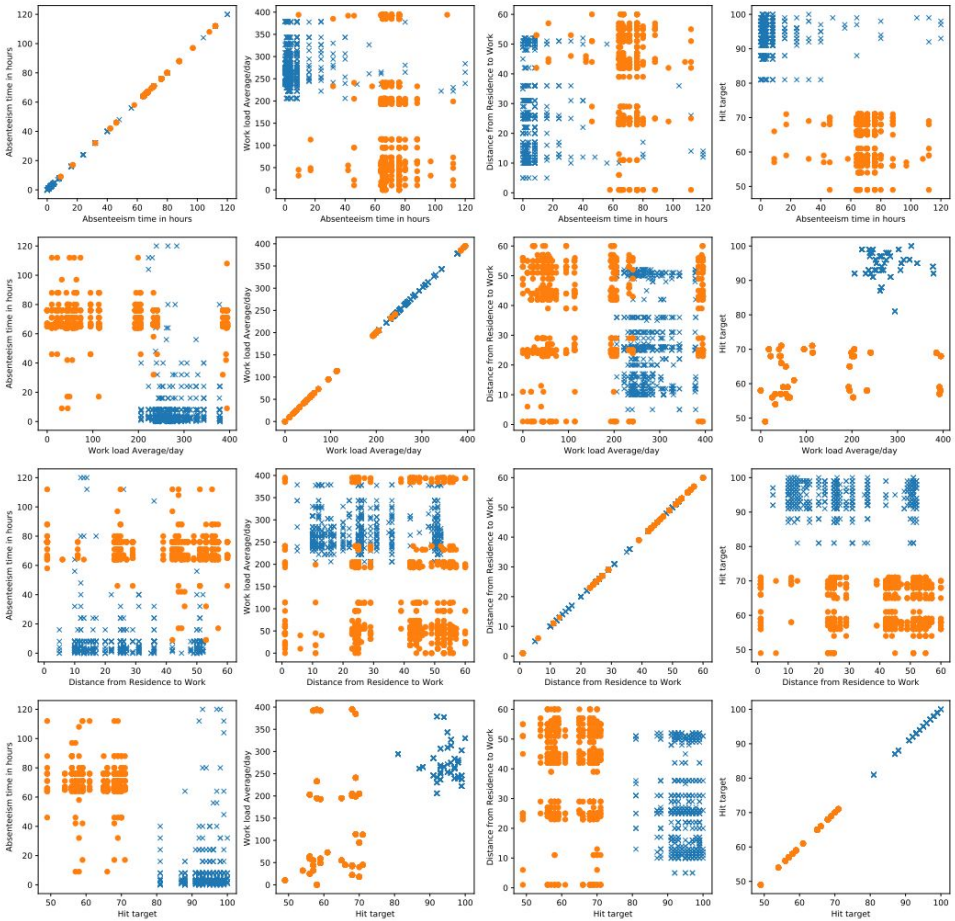
Proactive Security





Source: Biggio and Roli. *Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning*, 2018.

Homomorphic Pseudonymization



Source: KIProtect Whitepaper

Common Sense

Source: https://commonsenseframework.org/

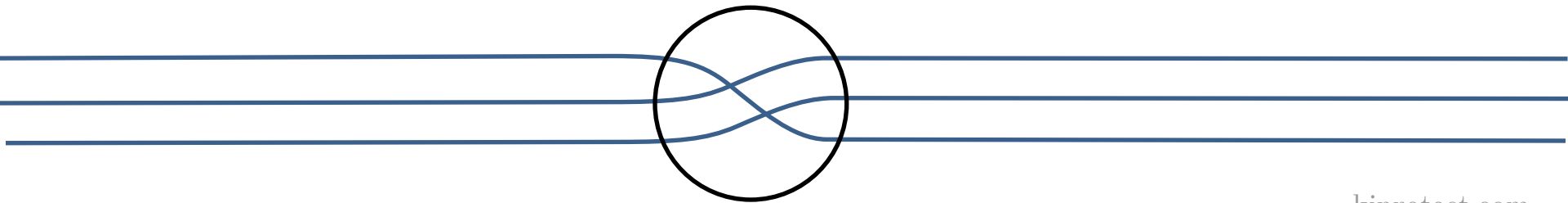
Thank you!

Questions? I'd love to hear them!

Or reach out anytime:

katharine@kiprotect.com
@kjam (Twitter)
https://github.com/kiprotect

7scientists GmbH
KIProtect
Bismarckstr. 10-12
10625 Berlin



Slide References

- DAVE-2 Self-Driving Car / Adversarial: https://arxiv.org/abs/1705.06640
- Flash Crash 2010: https://en.wikipedia.org/wiki/2010_Flash_Crash
- Membership Inference Attack: https://arxiv.org/pdf/1610.05820.pdf
- Secret Sharer: https://arxiv.org/abs/1802.08232
- XKCD: https://xkcd.com/538/
- Data Breach Index: https://breachlevelindex.com/
- Poisoning Attack: https://pralab.diee.unica.it/sites/default/files/biggio-ICB2013.pdf
- Wild Patterns: https://arxiv.org/pdf/1712.03141.pdf
- KIProtect Whitepaper: https://kiprotect.com/technical-whitepaper.html