DATA ANALYTICS ARE POWERFUL

HANDLE WITH CARE

Jeremy D. Wendt

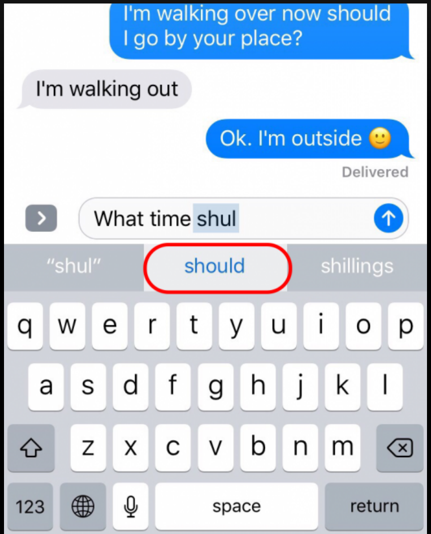
Joint work with Philip Kegelmeyer, Ali Pinar, Tim Shead, Gary Saavedra, Cosmin Safta, Joe Bertino, and many others

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND2019-12018 C

NITROGLYCERIN

Ascanio Sobrero

DATA ANALYTICS' DAILY EFFECTS - 2019



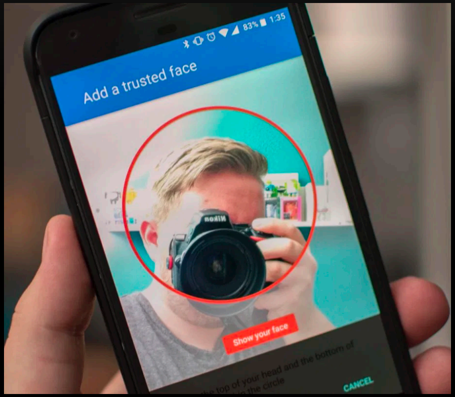
I'm walking over now should I go by your place?

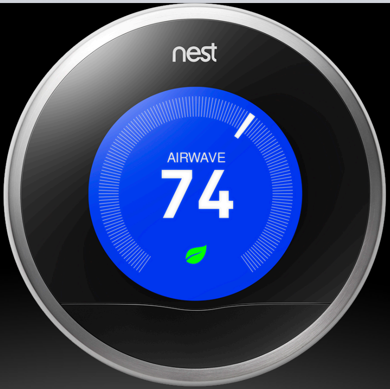
I'm walking out

Ok. I'm outside 🙂

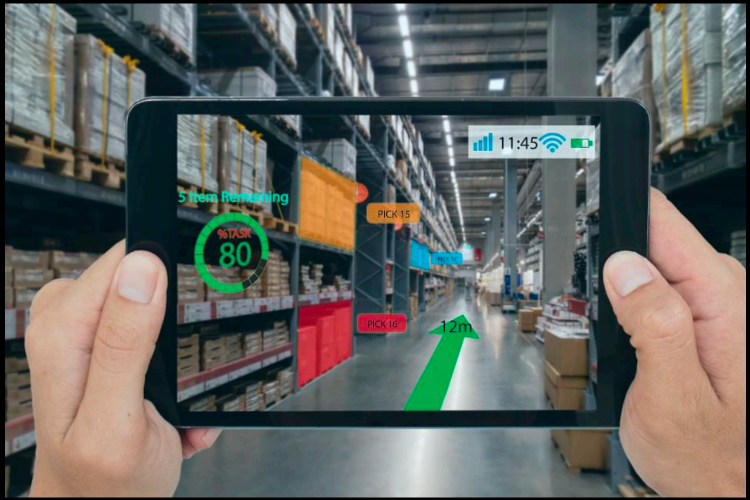
Delivered

What time shul

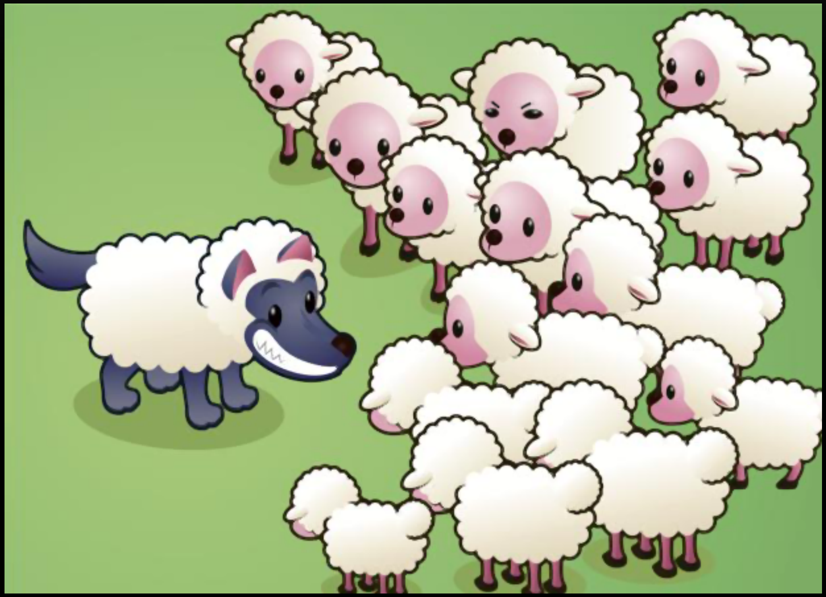


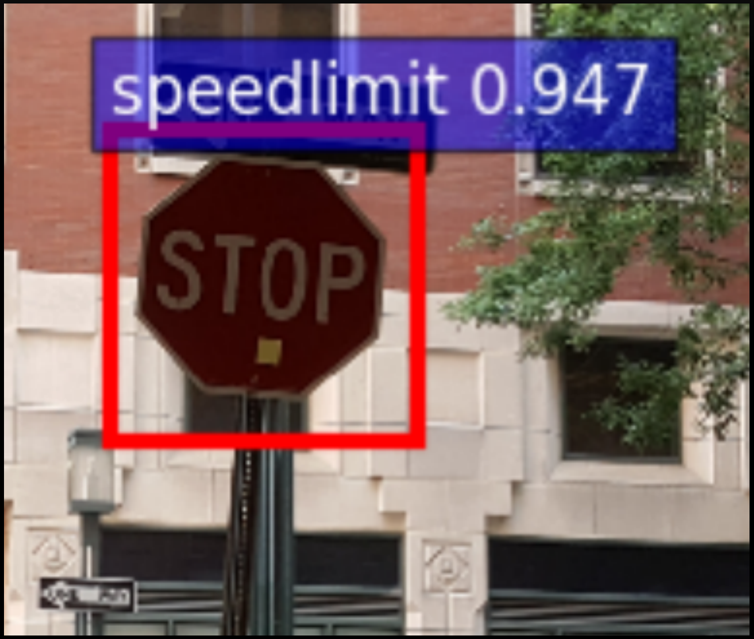


DATA ANALYTICS DAILY EFFECTS – SOON?



SUBVERT



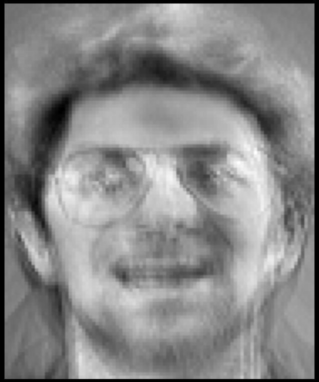


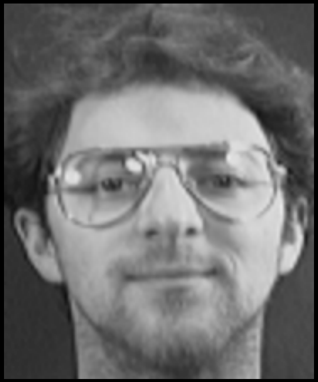
e.g., Kegelmeyer, Shead, Crussell, Rodhouse, Robinson, Johnson, Zage, Davis, Wendt, Doak, Cayton, Colbaugh, Glass, Jones, and Shelburg, 2015.

EVADE

e.g., Athalye, Engstrom, Ilyas, and Kwok, 2018.

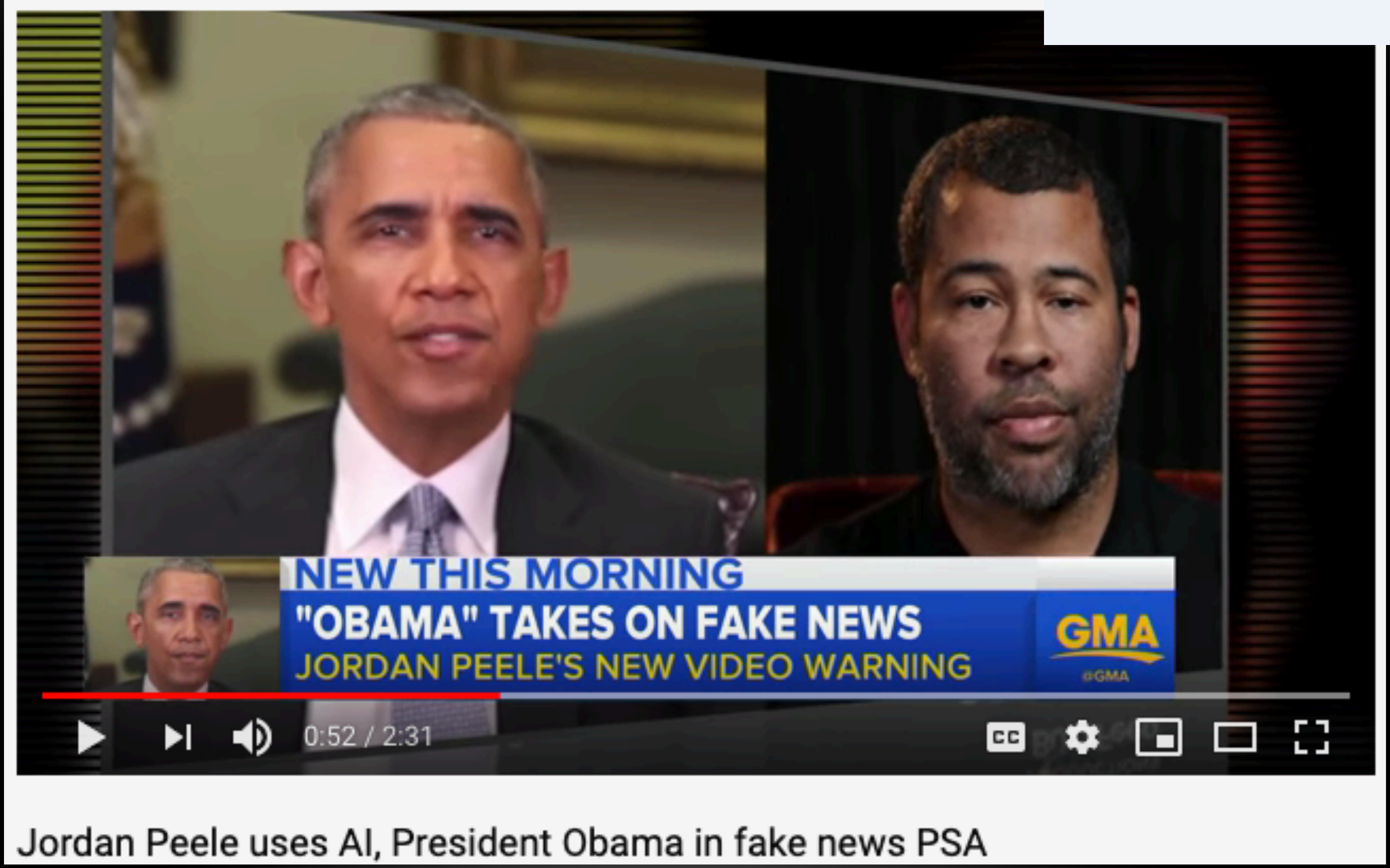
REVEAL





e.g., Fredrikson, Jha, and Ristenpart, 2015, or Shorki, Stronati, Song, Shmatikov, 2017.

APPLY

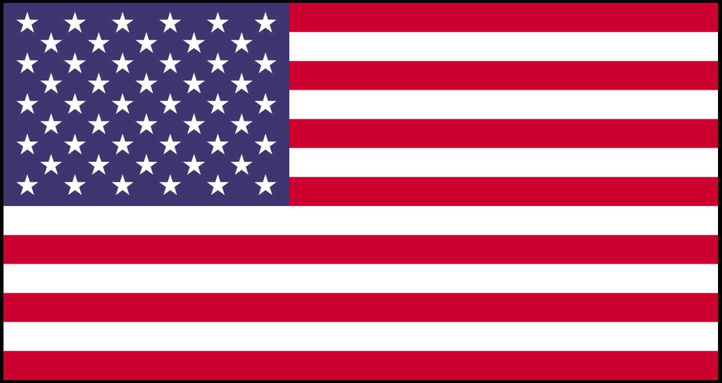


Jordan Peele uses AI, President Obama in fake news PSA

e.g., Isola, Zhu, Zhou, and Efros, 2016 or Kim, Garrido, Tewari, Xu, Thies, NieSSner, Perez, Richardt, Zollhofer, Theobalt, 2018

WHO ARE ADVERSARIES?

- It gets complicated really quickly

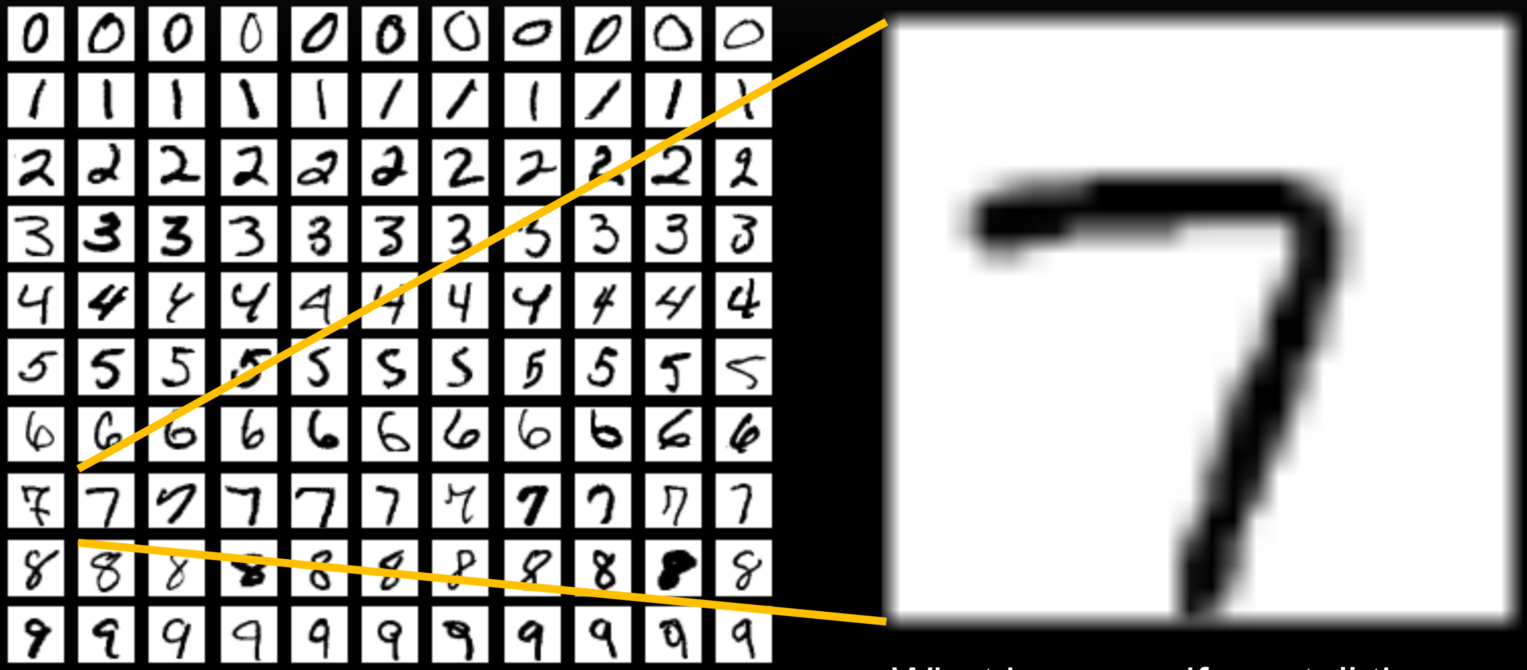


ADVERSARY’S ABILITIES

Credit: Bernard Goldbach

- In all analyses that follow, we provided the adversary significant abilities:
  - Full awareness of our chosen technique and implementation
  - Full understanding of the data to be analyzed
  - Some ability to alter the data or labels to affect our analytics

SUPERVISED LEARNING

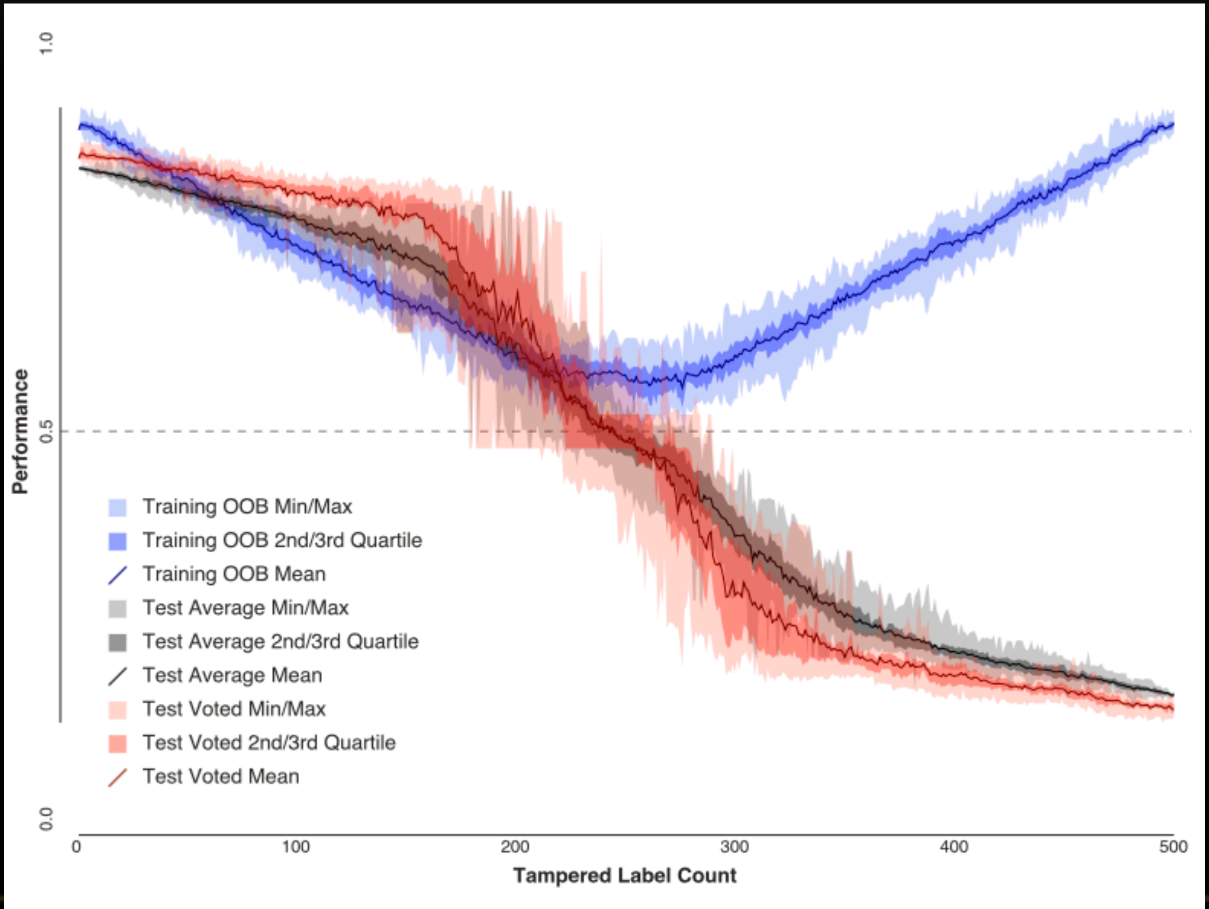


What happens if you tell the system this is a ‘1’ or ‘8’ ?

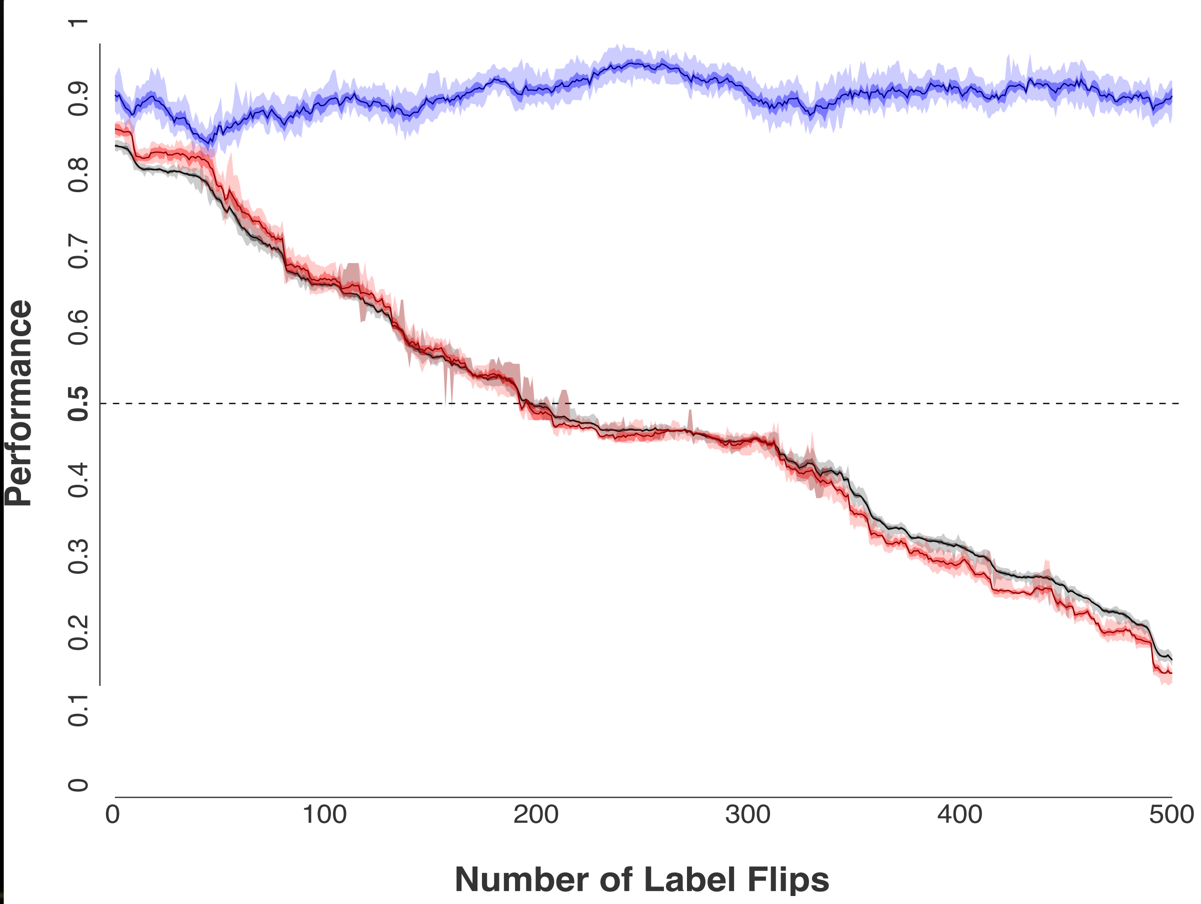
MEASURING SUPERVISED LEARNING

- Accuracy = Num Correct / Num Possible
  - Requires knowing correct labels
  - However, the defender no longer has correct labels!
- Defender can only measure his accuracy against the altered labels

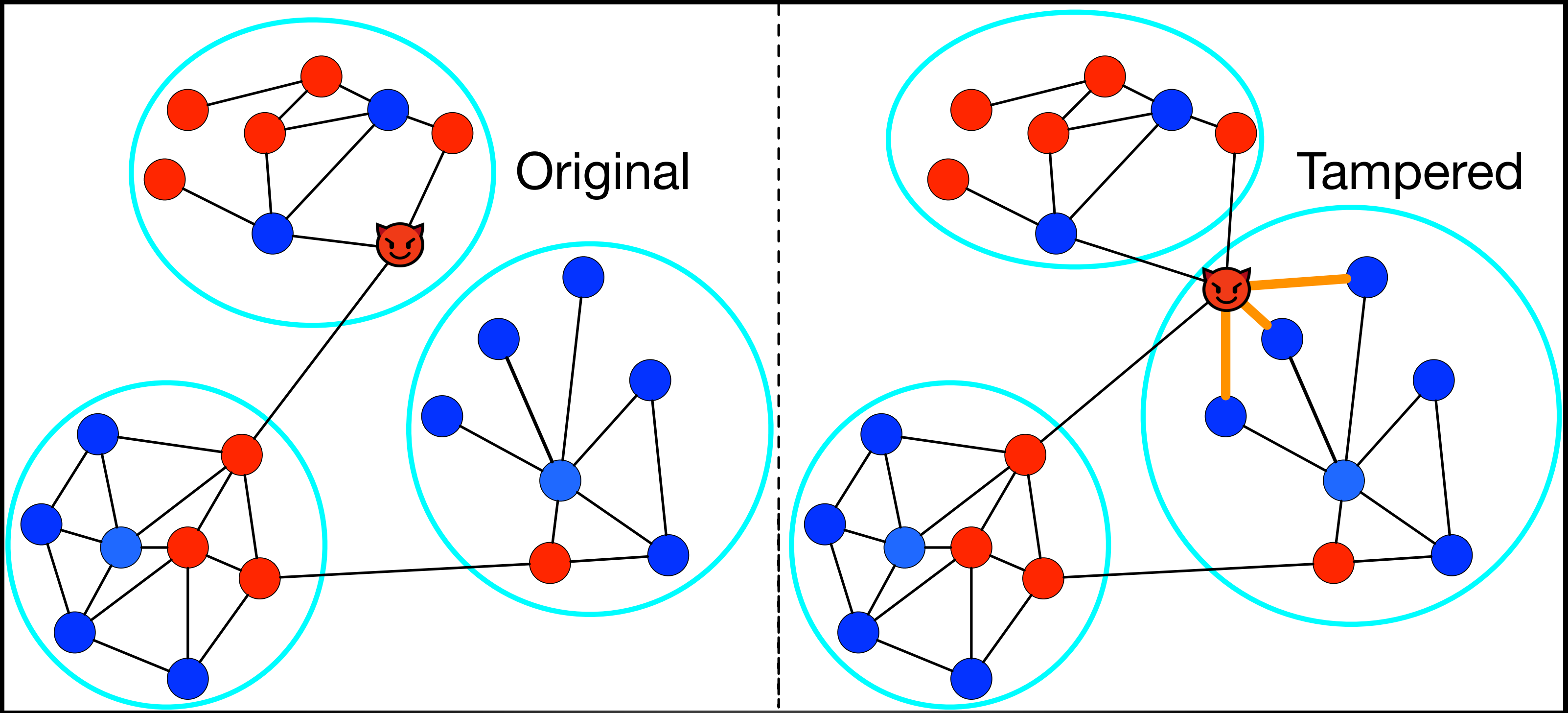
SUBVERT CONFIDENCE



SUBVERT WITHOUT DETECTION



ATTACKING COMMUNITY DETECTION



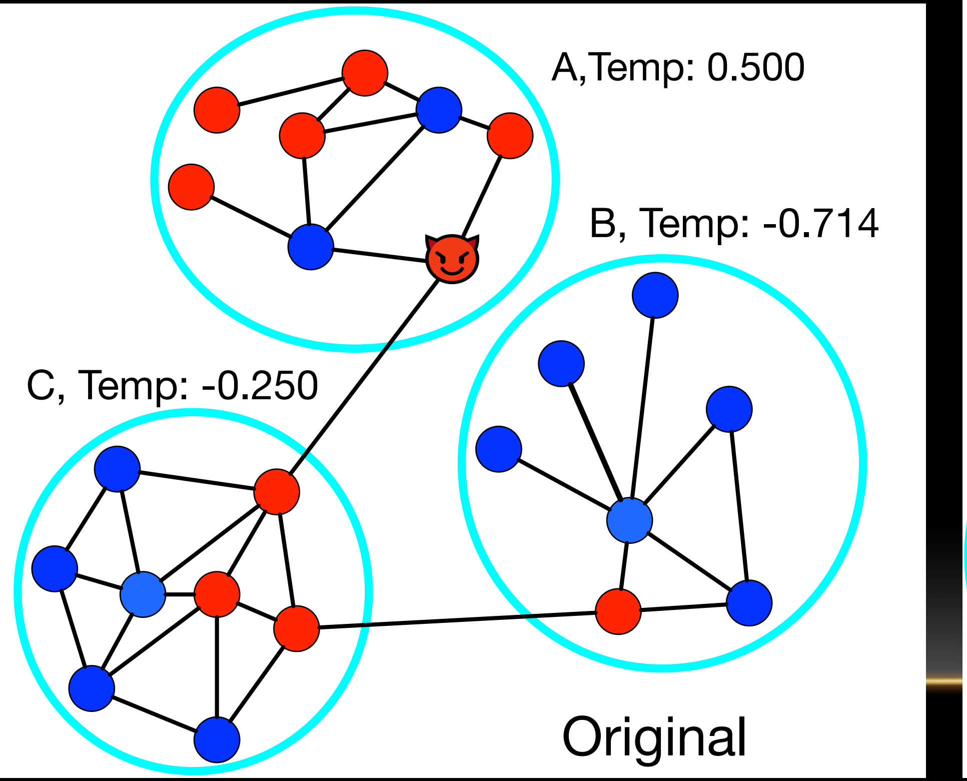
Kegelmeyer, Wendt, and Pinar, 2018

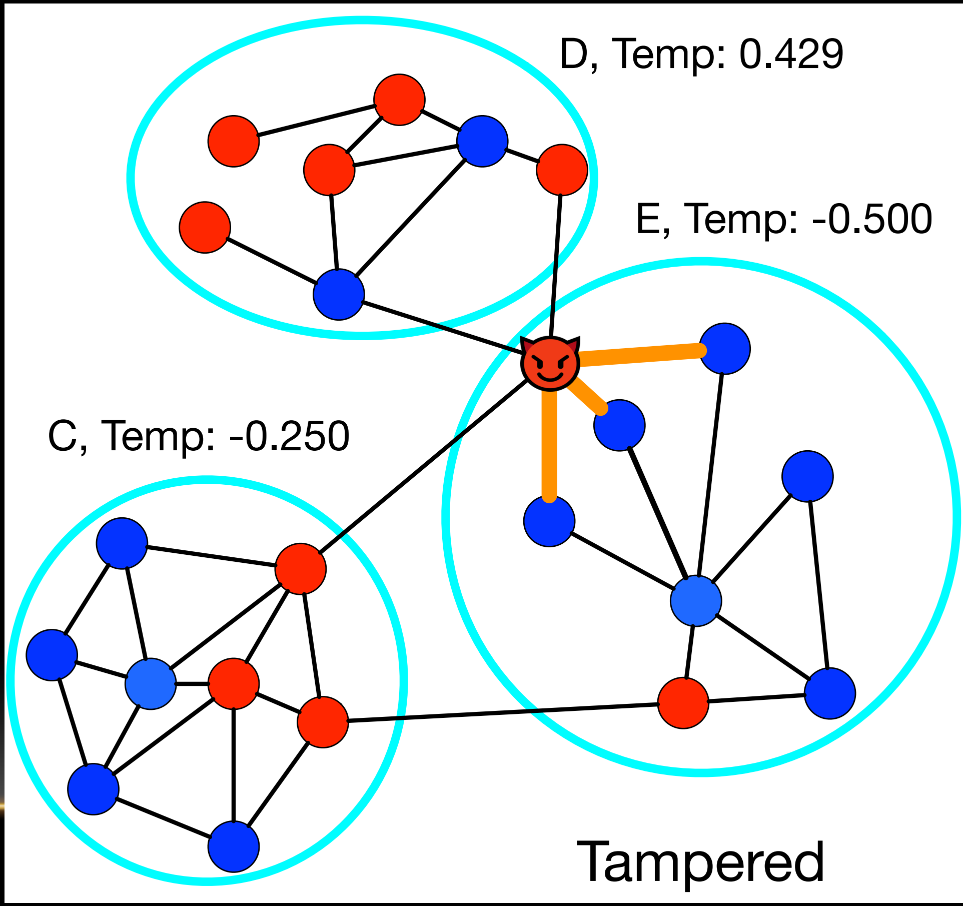
COMMUNITY TEMPERATURE

Hot (red) = 1; Cold (blue) = -1; Unknown = 0

Community temp is average of nodes' temps

Higher temperature is more suspicious

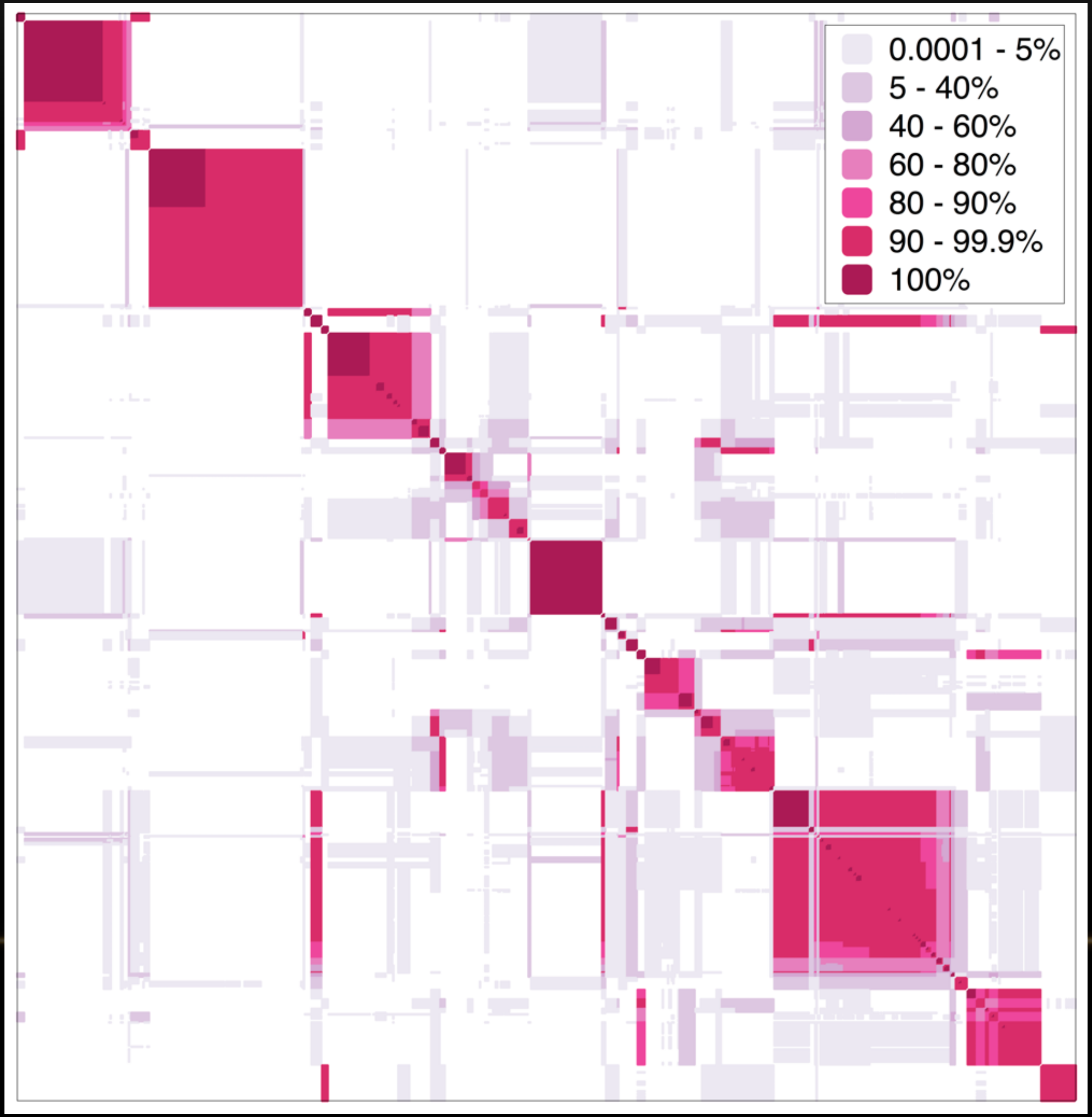




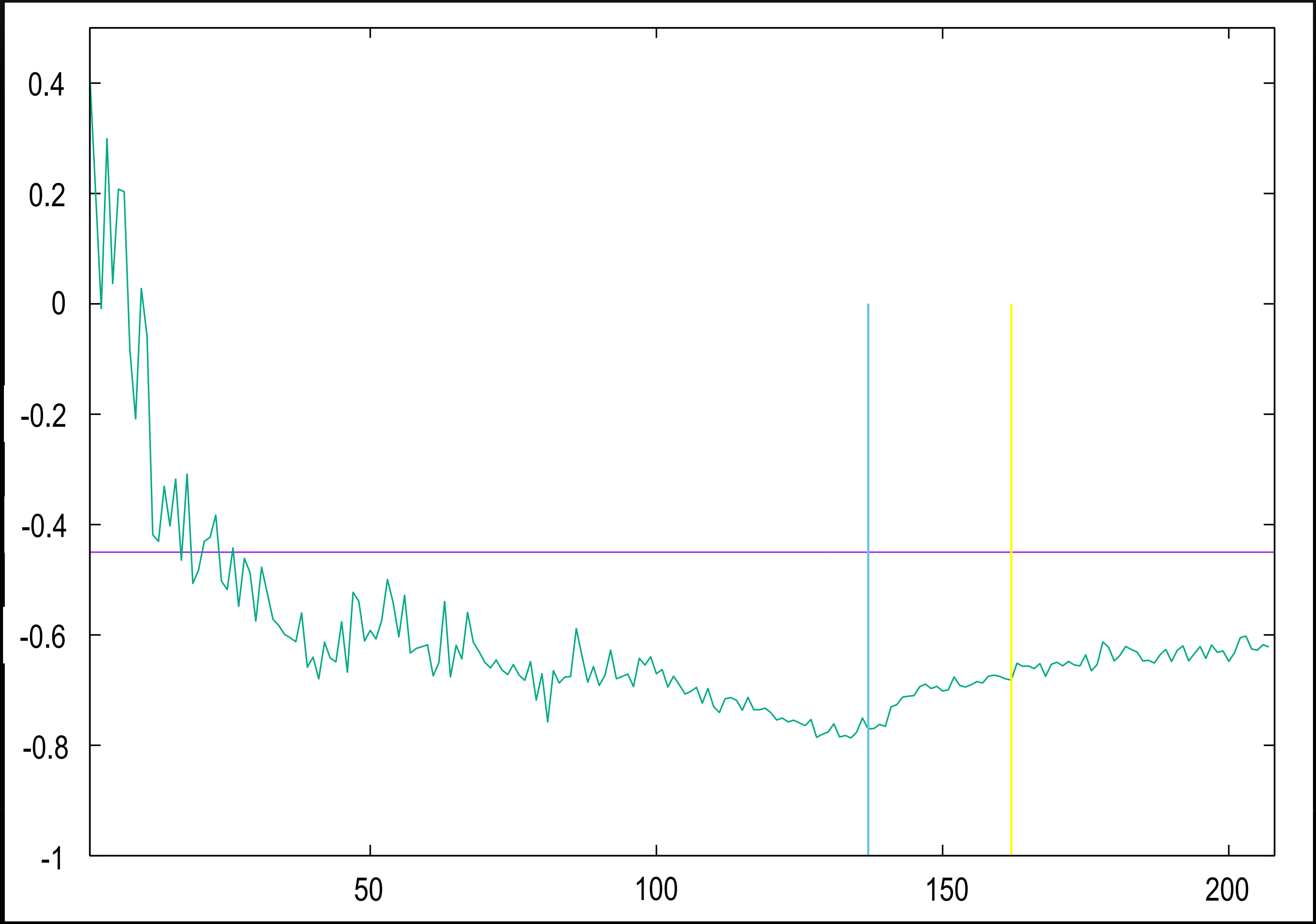
ATTACK HEURISTICS

- Attacker wants to be in cold community
    - He can add edges between his node and others – which to add?
- *Random* – Add random edges; used as a baseline
- *Stratified Random* – Random to cold, then unknown, last hot
- *Cold and Lonely* – Stratified random, but lower degree first w/in temperature bands
- *Stable Structure* – Exploit graph structure; next slide
- *Greedy Pessimal* – Exhaustively compute worst case; computationally infeasible in general, approximately worst case

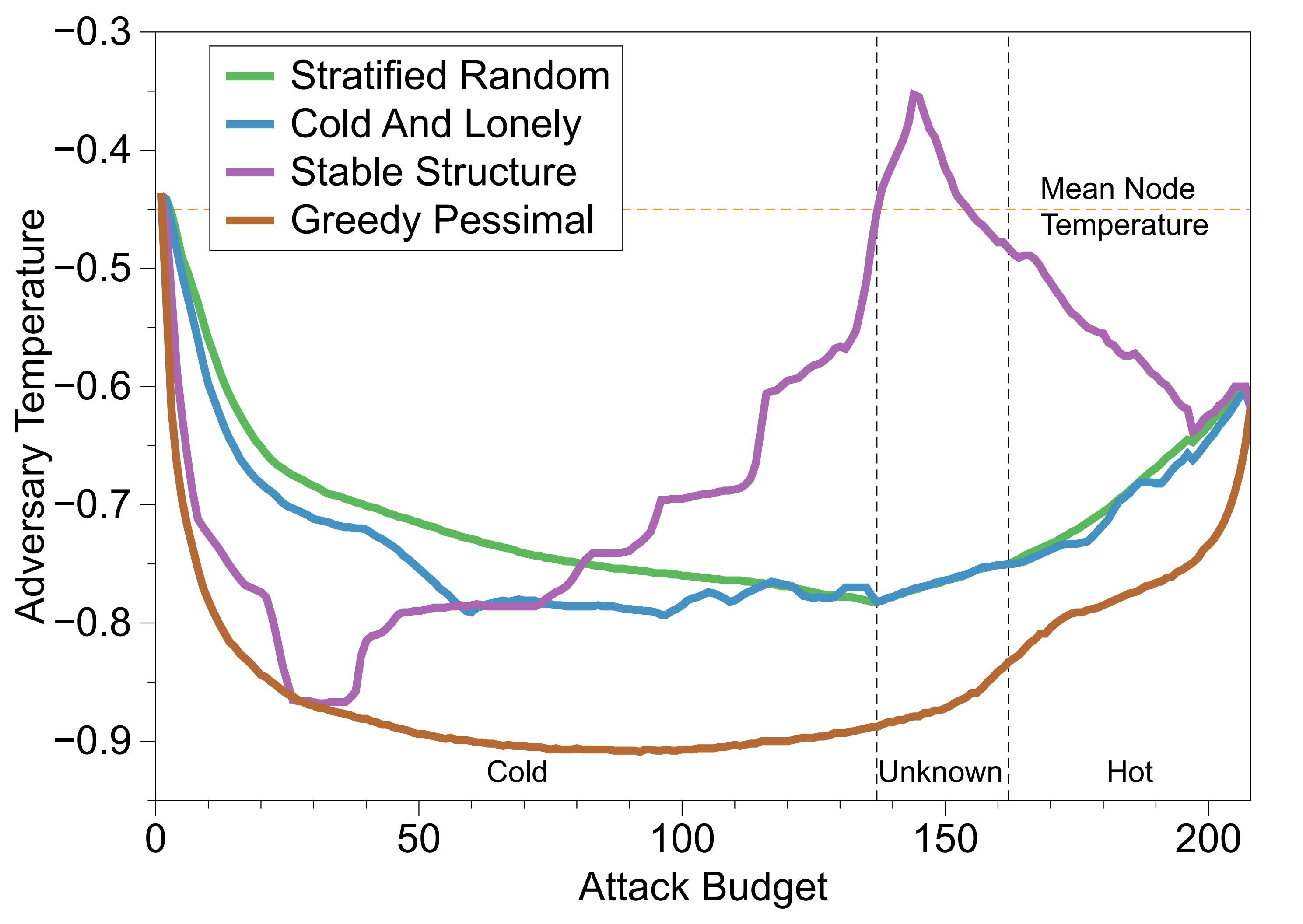
STABLE STRUCTURE



CREATING ATTACK PLOTS

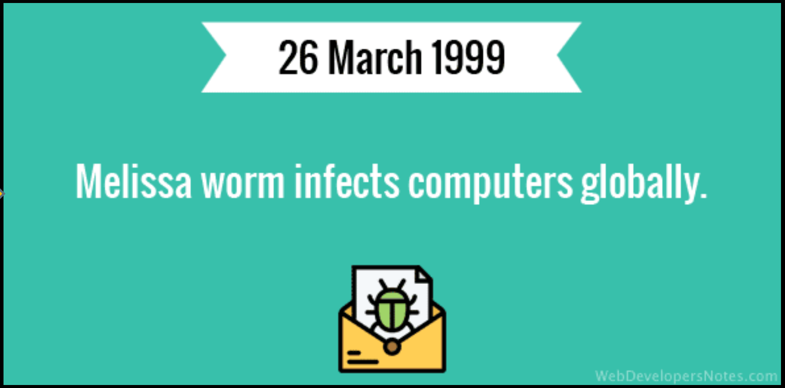


ATTACK EFFECTIVENESS



DATA ANALYTICS NEED DEFENSES

- Computer Networks in 1990s → Data Analytics in 2019



26 March 1999

Melissa worm infects computers globally.

DYNAMITE

Alfred Nobel

COMPETITION IS UBIQUITOUS