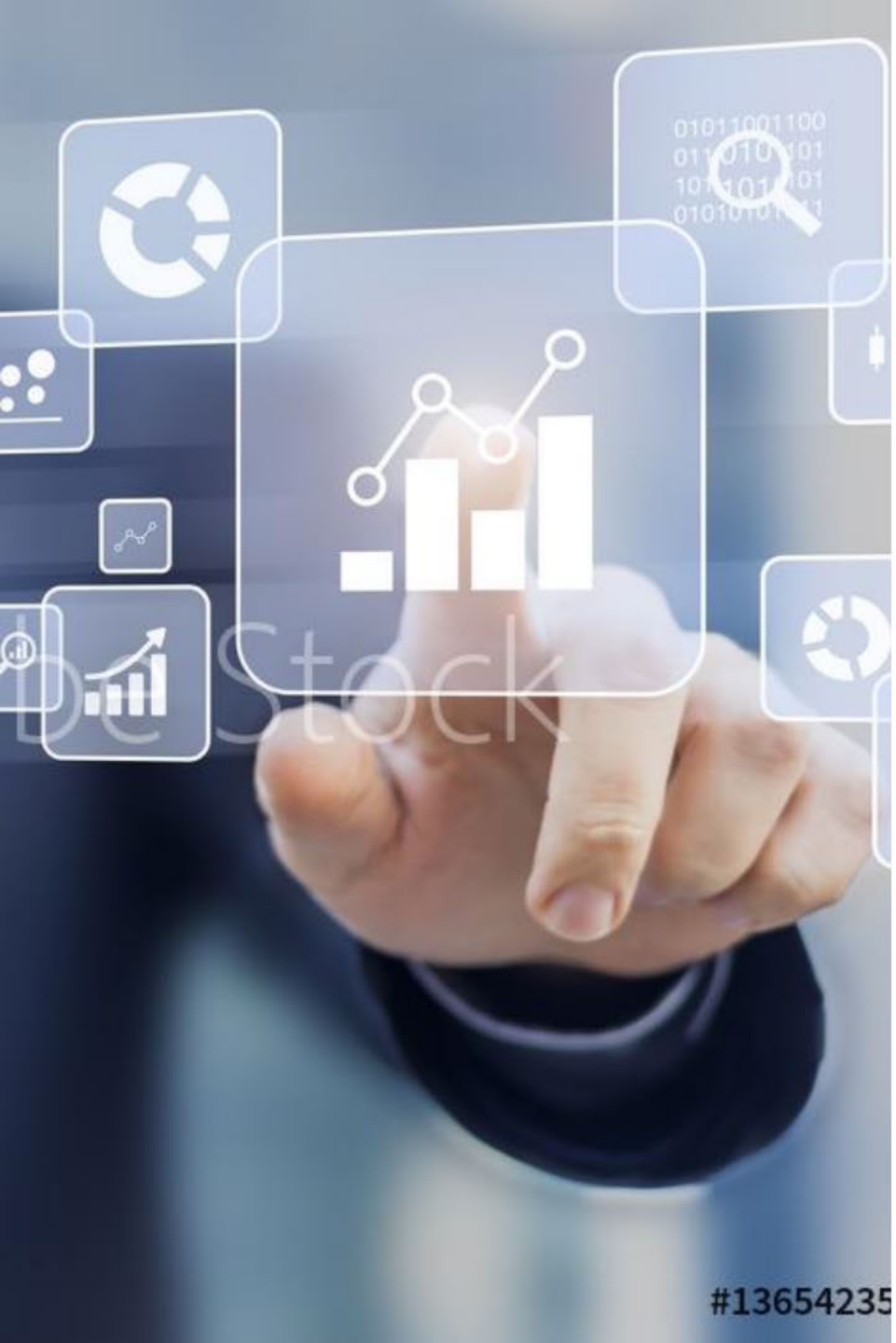
Santa Clara U. Law 371
Adversarial Machine Learning

What is it?

- Attack on machine learning models (CIA):
  - Confidentiality of data
    - Breaches
    - Trade secret theft
    - Ransomware (threaten disclosure)
  - Integrity of data
    - Malicious destruction or alteration
  - Availability of data
    - Disgruntled employee
    - Ransomware (encrypt and make unavailable)
    - Malicious destruction of data (revenge)

Threats

- Threats
  - MITRE
    - https://attack.mitre.org/
  - MITRE ATLAS
    - https://atlas.mitre.org/



Attack Types

- Attack Types
  - White box – access to training process
  - Black box – no access to training process
  - Targeted – misclassify to a specific class
  - Untargeted – misclassify to anything but correct class

Open-Source Tools

- IBM's Adversarial Robustness Toolbox
  - https://adversarial-robustness-toolbox.readthedocs.io/en/latest/
- TextAttack
  - https://textattack.readthedocs.io/en/latest/
- CleverHans
  - http://www.cleverhans.io/
- Foolbox
  - https://foolbox.jonasrauber.de/

Solutions

- Robustness
  - Augmentation
  - Simplicity of model - linear activation is more susceptible to attack
- Controls over data, training process, production
  - Access and authentication
  - Encrypt data
  - Etc.
- Risk Quantification
  - https://www.fairinstitute.org

https://www.fairinstitute.org/blog/how-to-model-controls-in-a-fair-risk-analysis