Data Integrity Based Attacks in Investigative Domains:

How Companies Are Exploiting Data Science to Thwart Investigative Outcomes

Eric Rozier

University of Cincinnati

ENIGMA

Eric & Wendy Schmidt

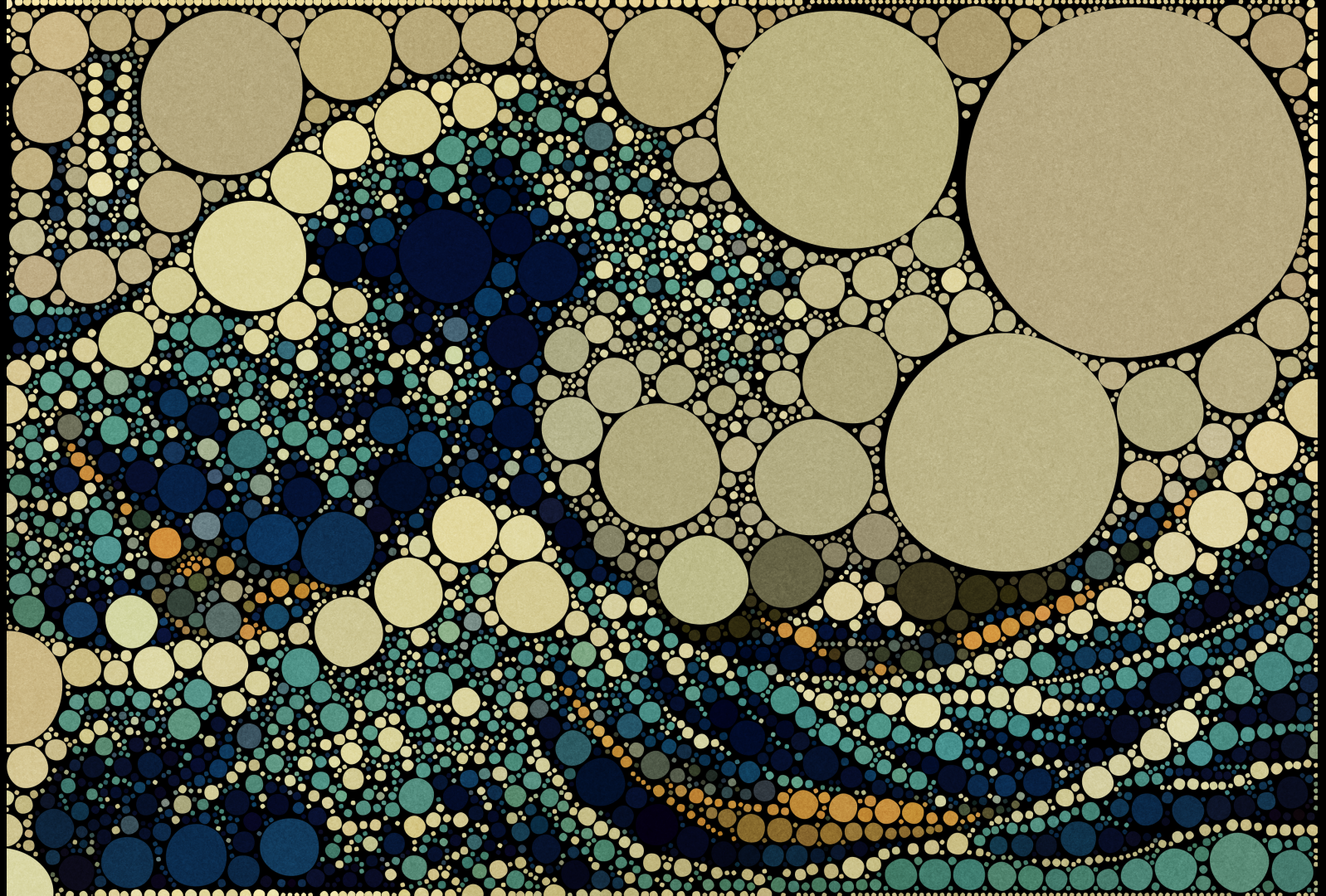
Data Science For Social Good

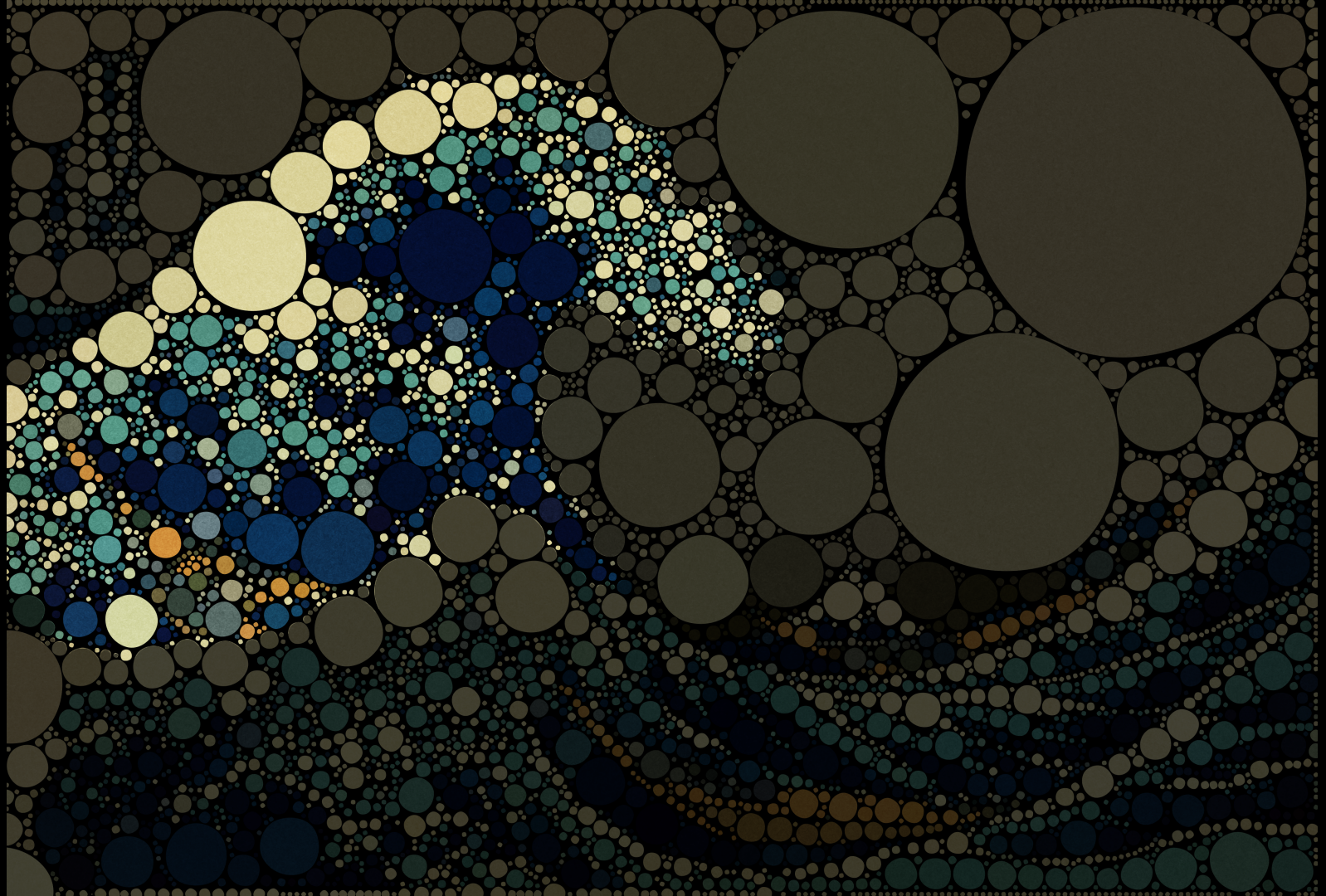
Summer Fellowship

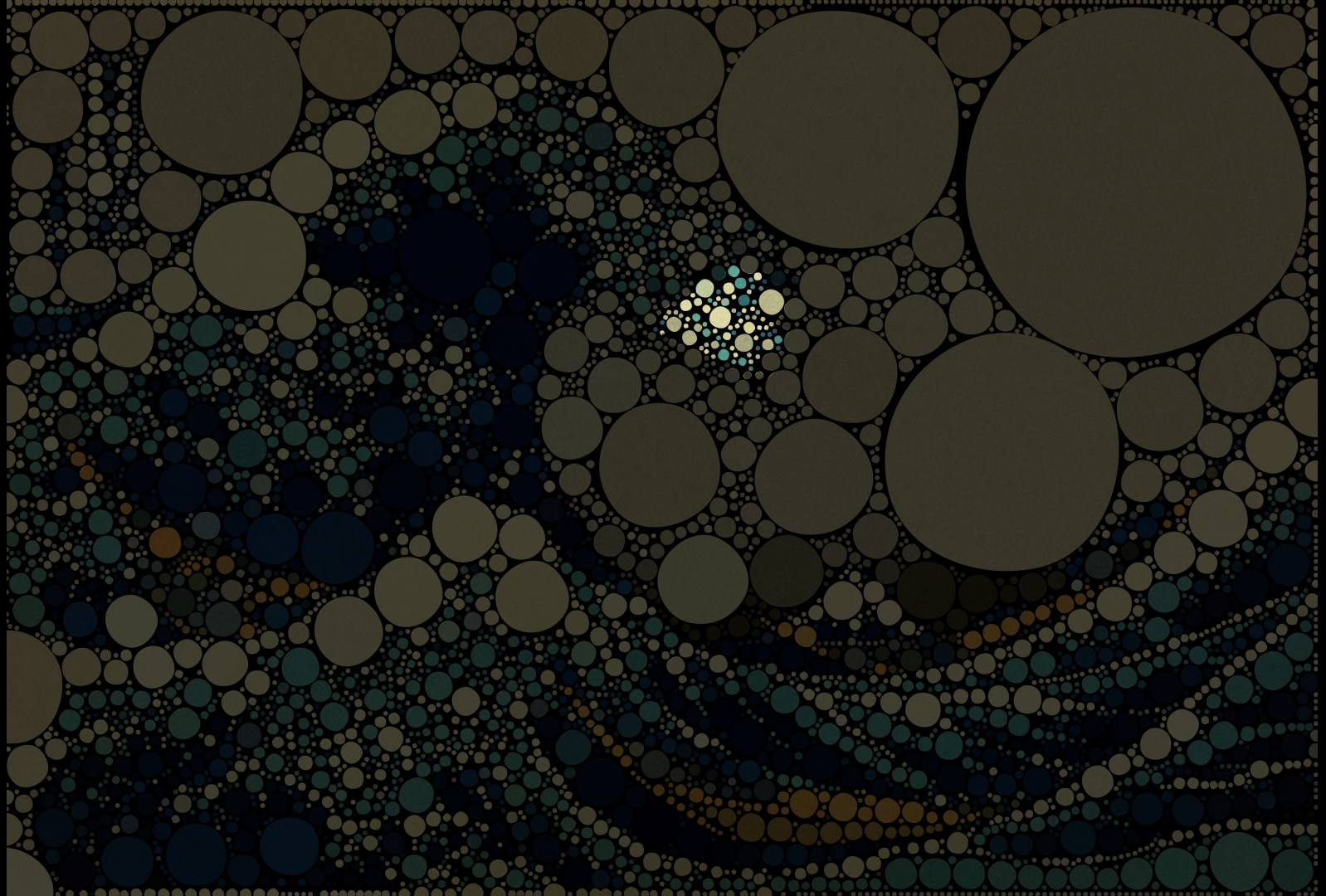
THE UNIVERSITY OF CHICAGO

THE WORLD BANK

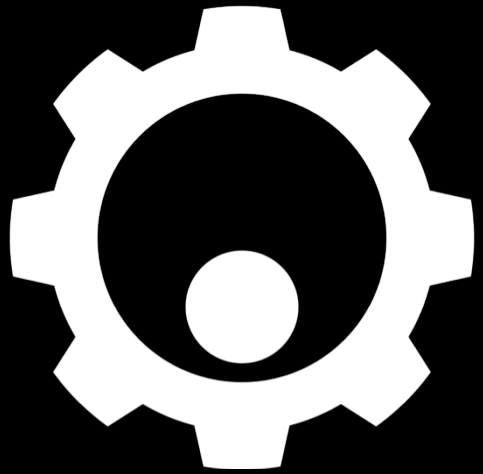
IBRD • IDA | WORLD BANK GROUP





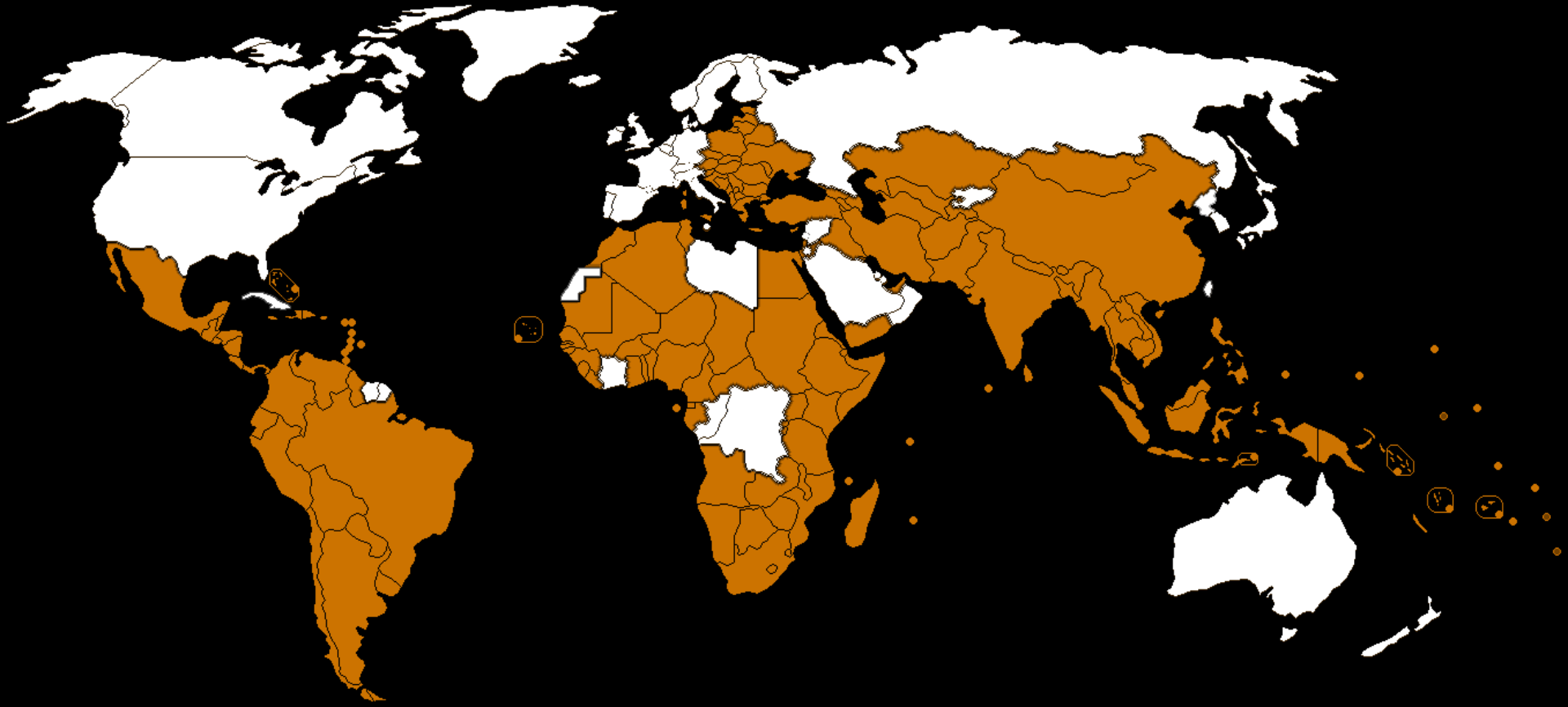


Automated Reasoning and Machine Intelligence



International Procurement

International Procurement



World Bank lends funds for a specific project.

National Entity buys goods and services

International Procurement

Goal: Supplier chosen through competitive bidding

Reality: Suppliers influence competitive process through collusion, corruption, or fraud

Winfried Bruenken/CC BY-SA-2.5

Ralph Malan/CC BY-SA-2.0

Hansueli Krapf/CC BY-SA-3.0

Peellden/CC
BY-SA-2.5

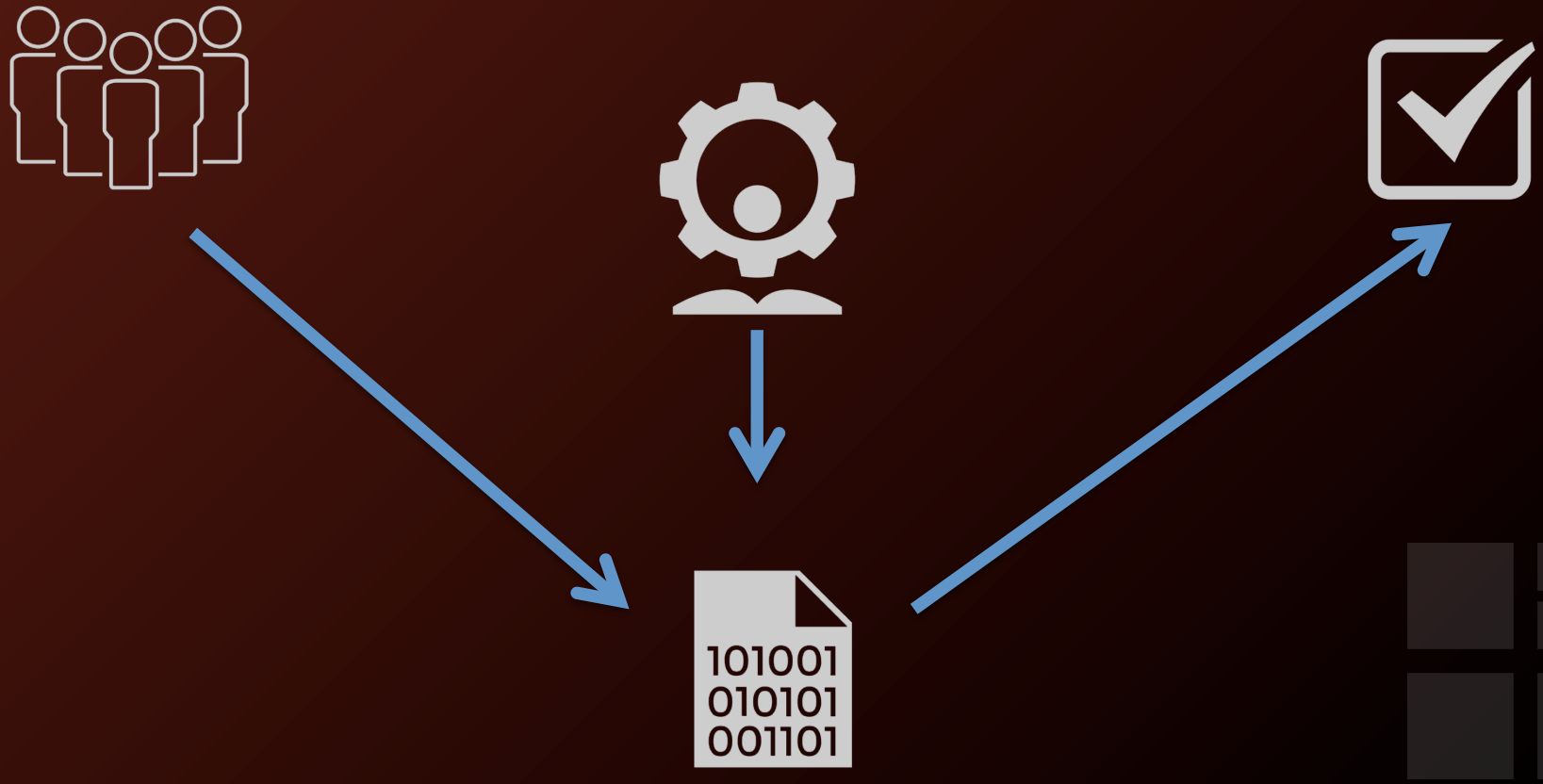
"A House Door In Send Surrey UK" by Si Griffiths - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Data Integrity Attacks

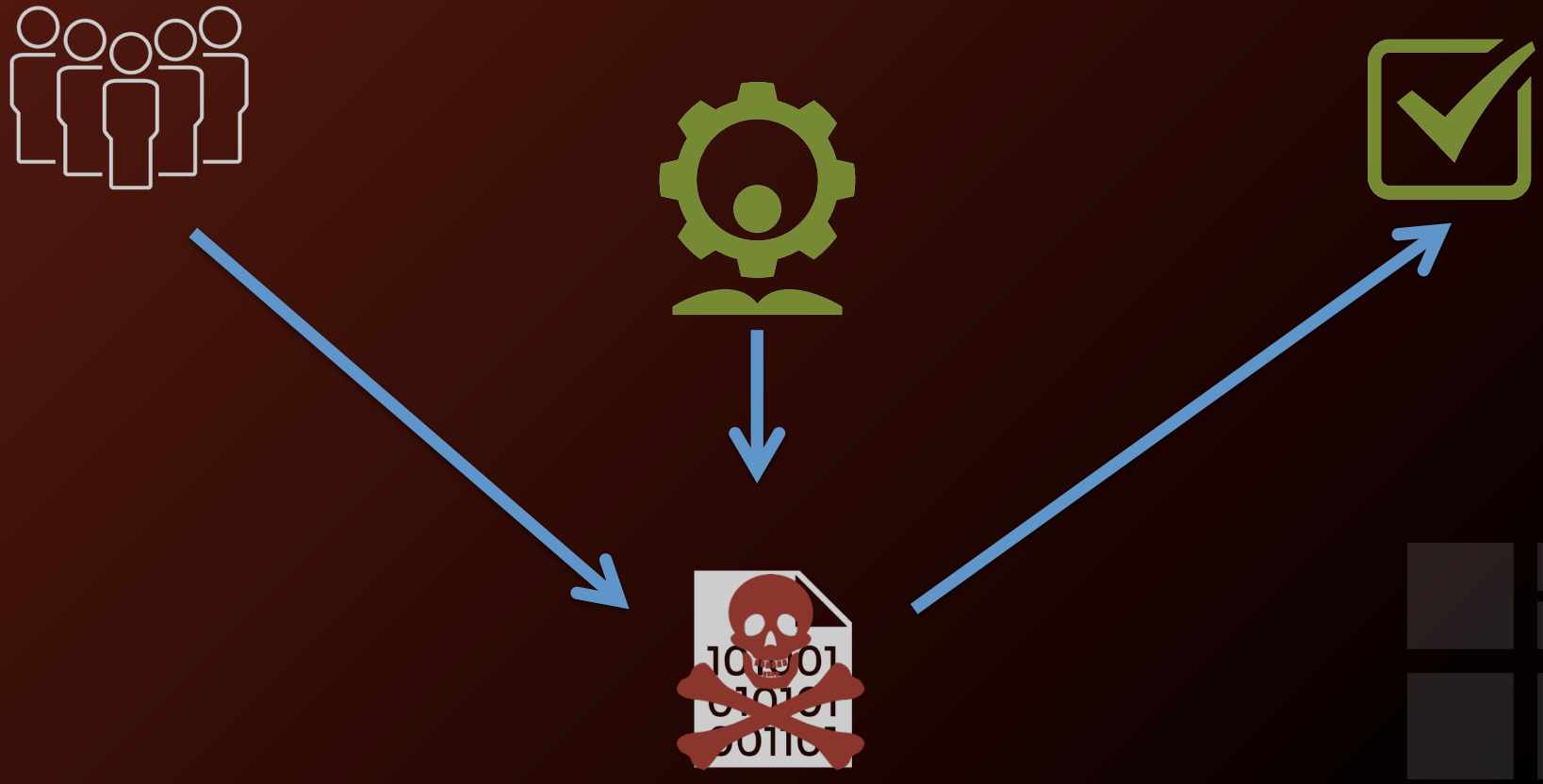
Data Integrity Attacks

A subversion of automated reasoning, machine learning, or other decision processes due to the malicious manipulation or falsification of data collected through normal channels.

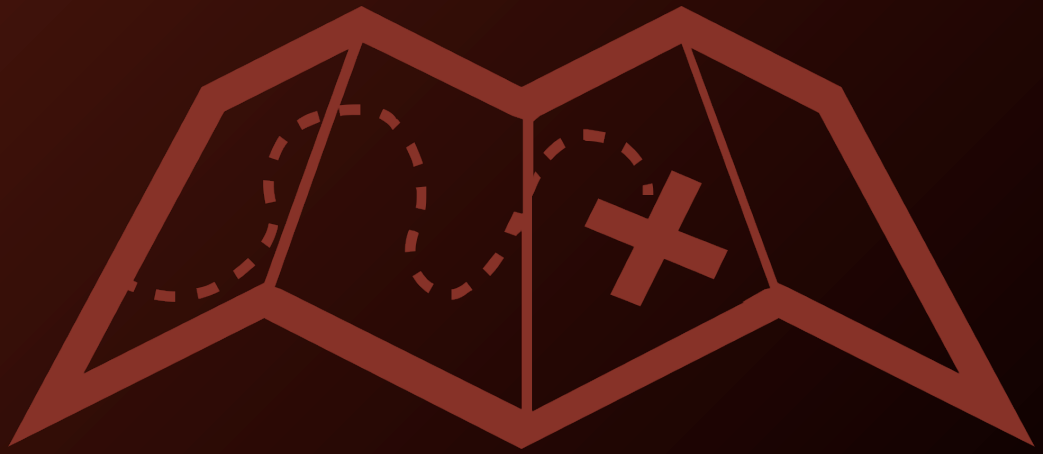
Anatomy of an Attack on Data Integrity



Anatomy of an Attack on Data Integrity



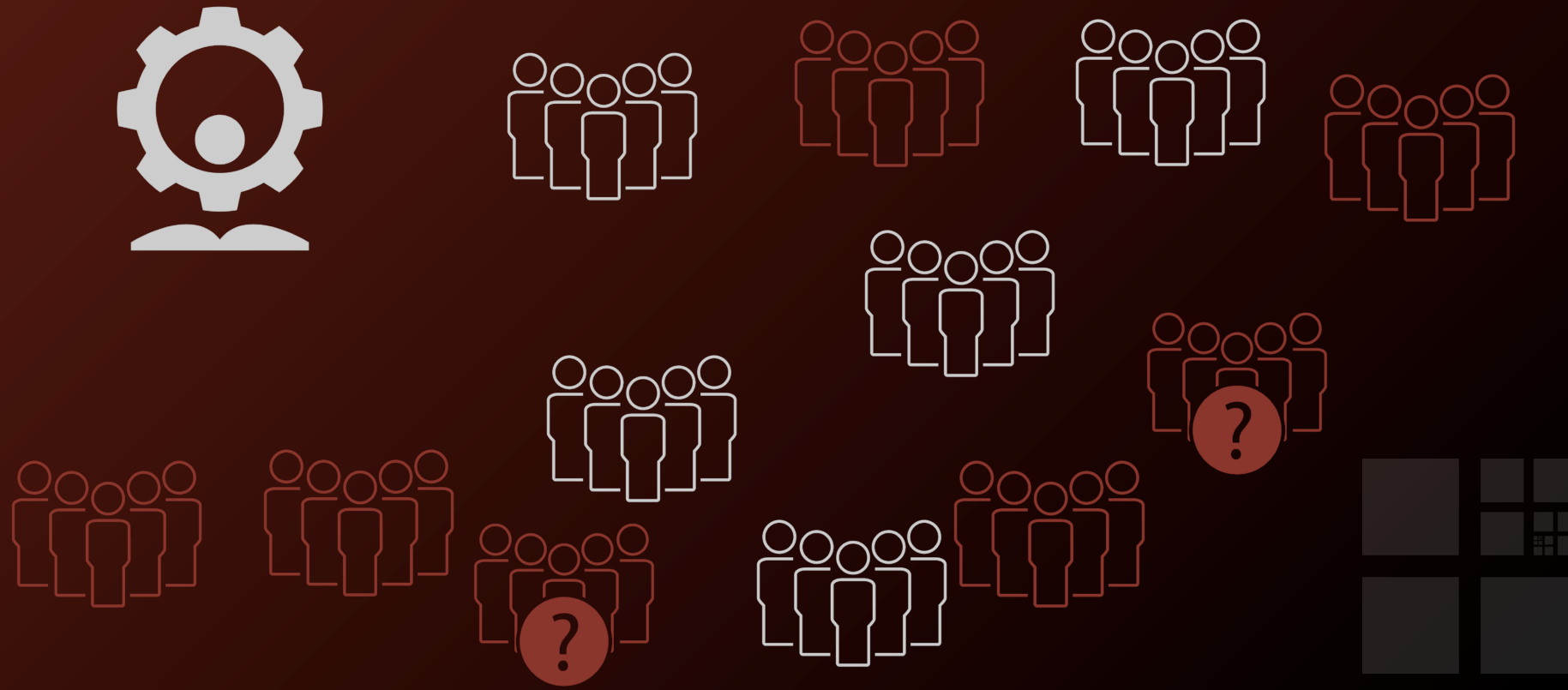
Goals of a Data Integrity Attack



Identity Obfuscation

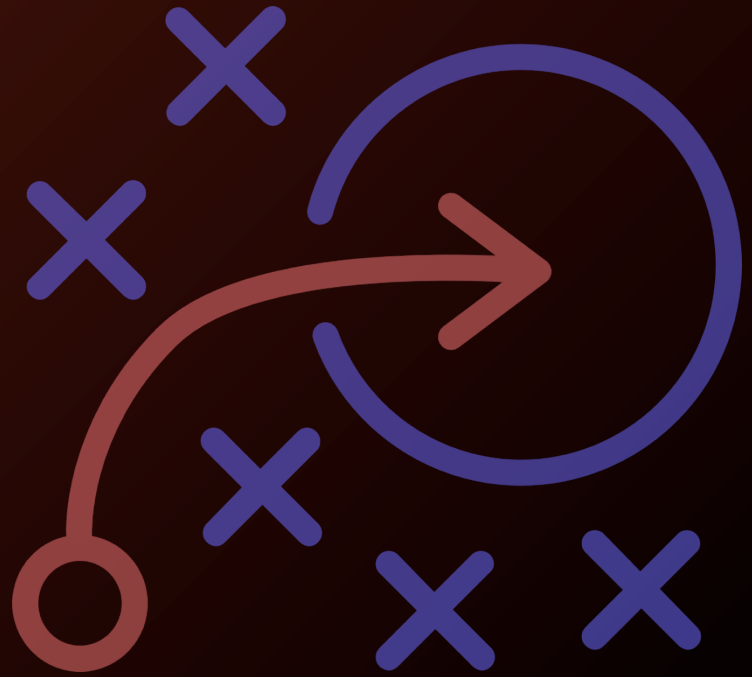
Ability Misrepresentation

Environment Misrepresentation

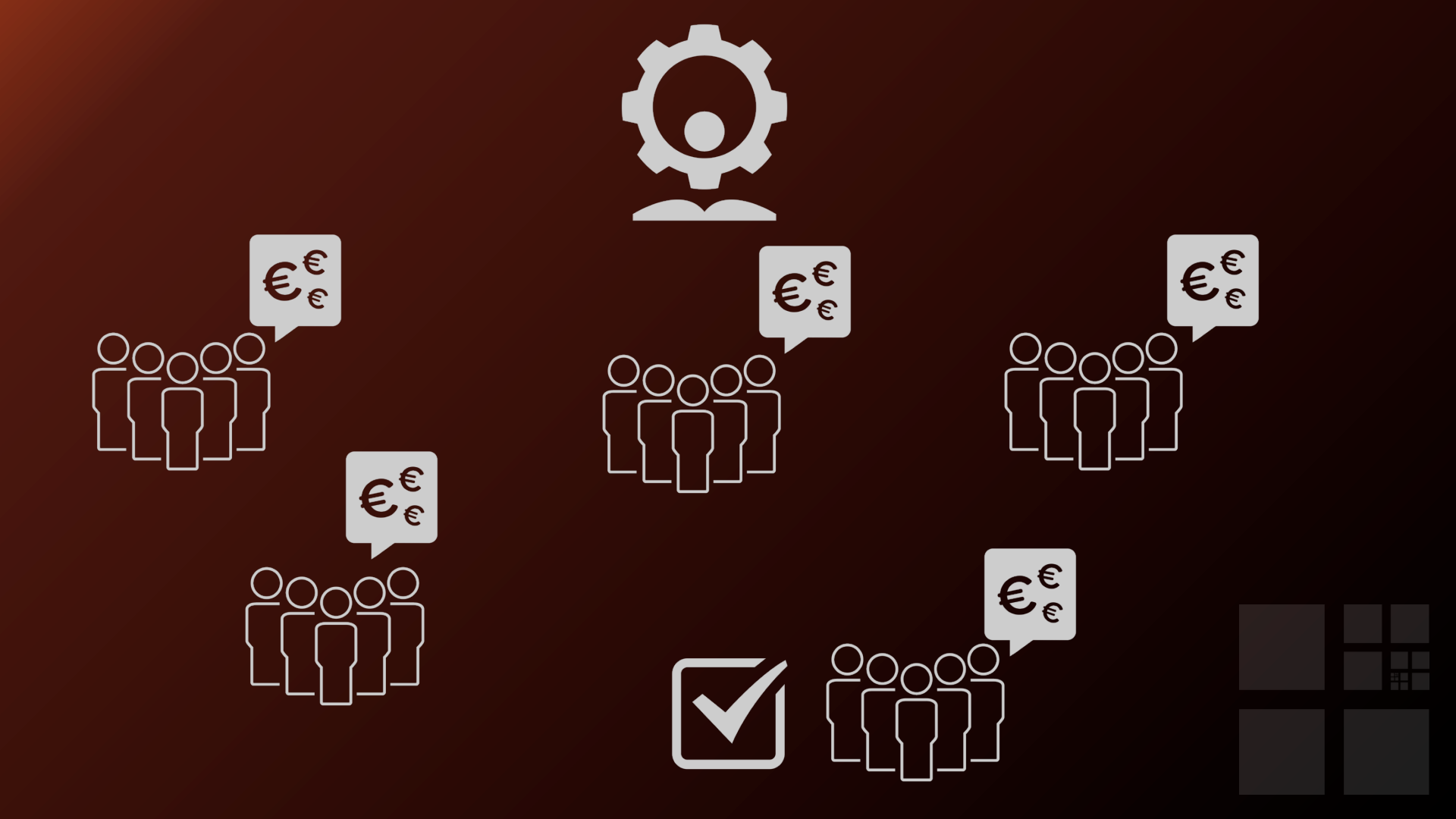


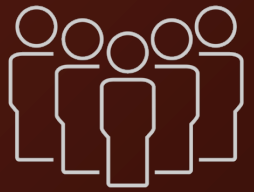
Strategy of Data Integrity Attacks

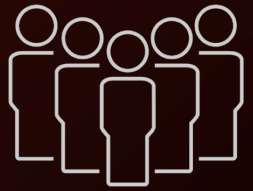
- Data pollution
- Data falsification
- Data blending

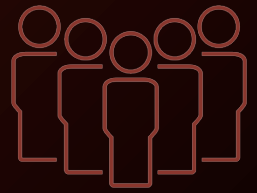


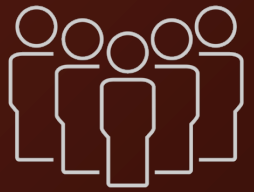
Example Attack

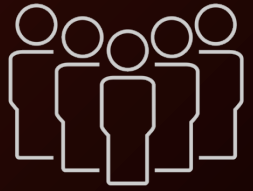




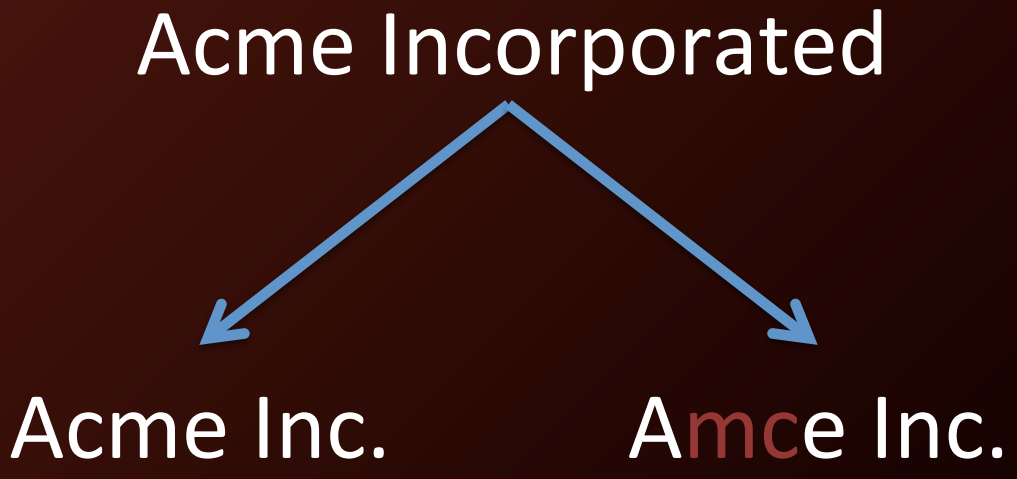








Identity Obfuscation



PricewaterhouseCoopers

PricewaterhoseCoopers

PricerwatershouseCoopers

PricewaterhouseCoopers’

Price Water Coopers Inc

PWCoopers

OpenRefine

PricewaterhouseCoopers

PricewaterhoseCoopers

PricerwatershouseCoopers

PricewaterhouseCoopers’

Price Water Coopers Inc

PWCoopers

PWC

PWC Inc

OpenRefine

?

Pacific Warehousing Company

Porter Walls Company

PricewaterhouseCoopers

PricewaterhoseCoopers

PricerwatershouseCoopers

PricewaterhouseCoopers’

Price Water Coopers Inc

PWCoopers

PWC

PWC Inc

PWC - Bangladesh

PWC - Uruguay

OpenRefine

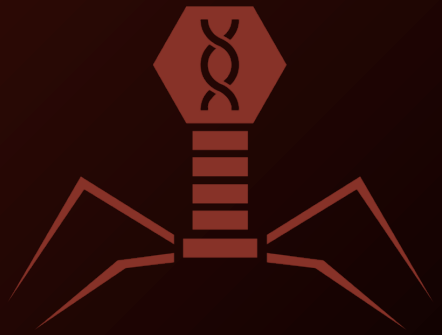
?

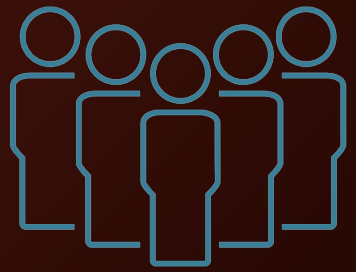
Tools Designed for Non-Malicious Environments

Other Vectors

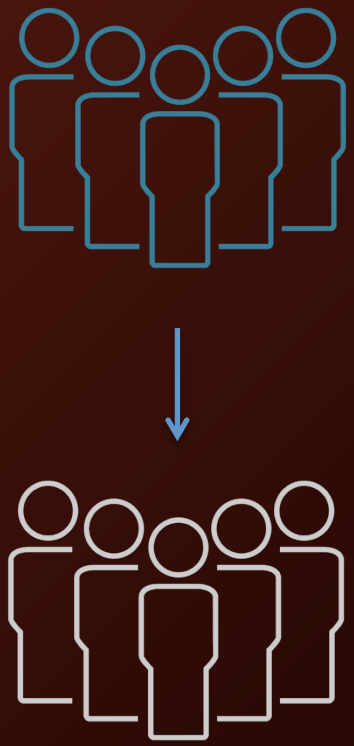
Other Vectors

Re-Incorporation and Legal Identity Shifts

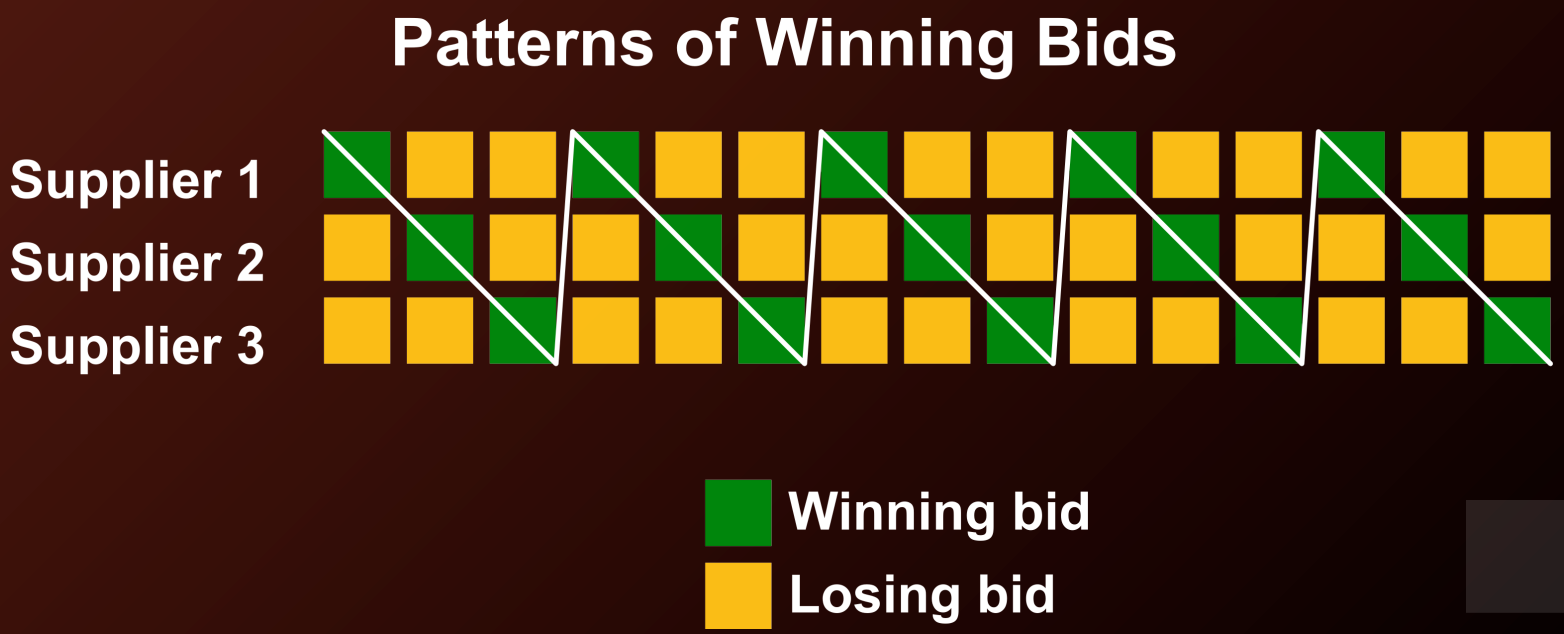




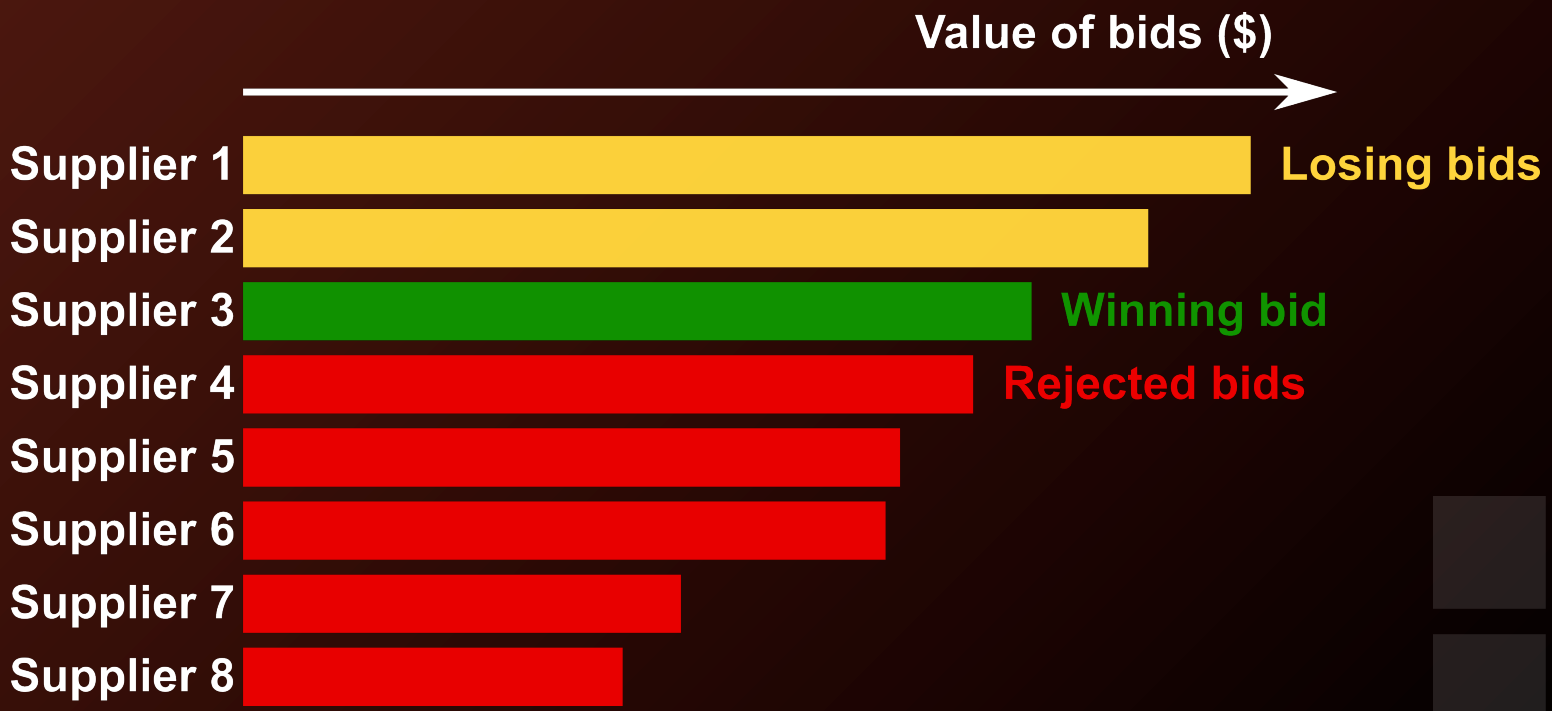
Other Vectors



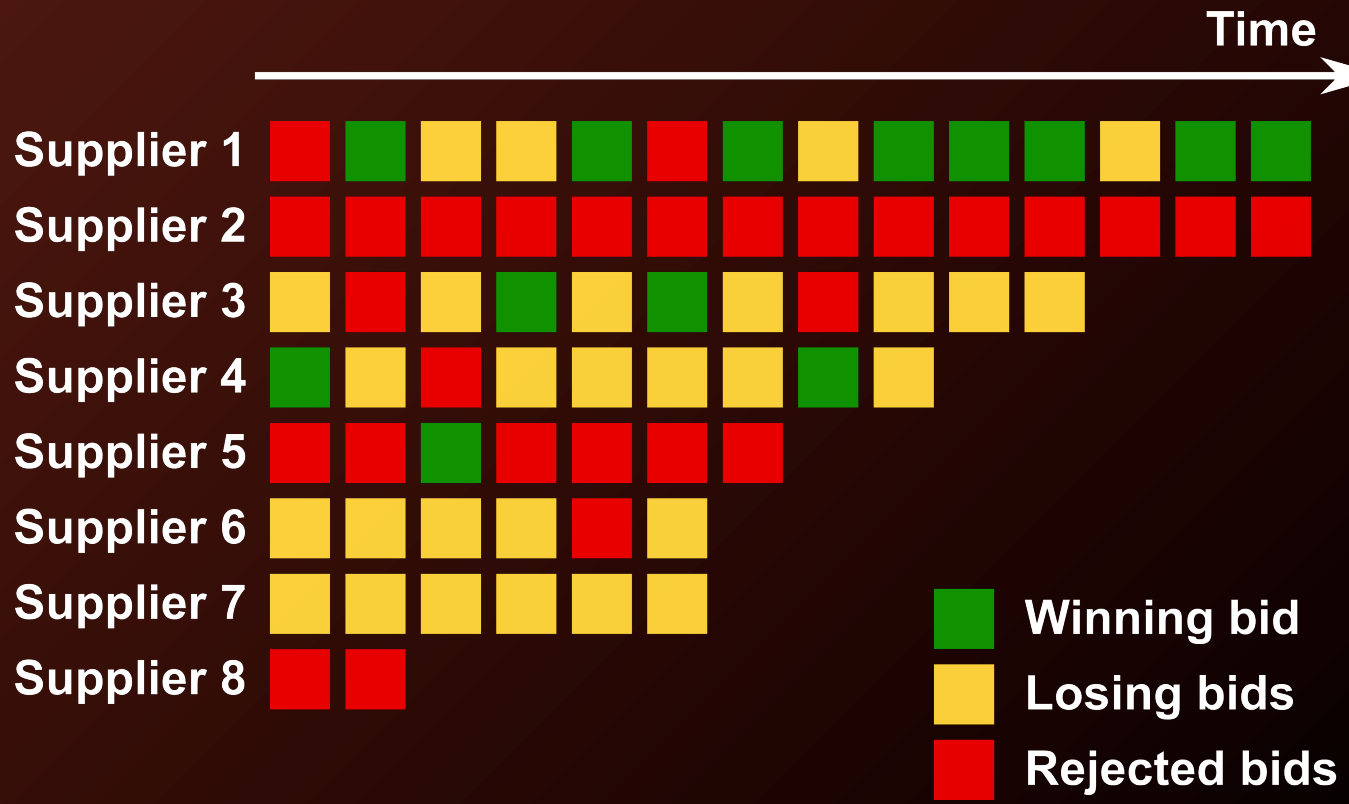
Environment Misrepresentation



Environment Misrepresentation

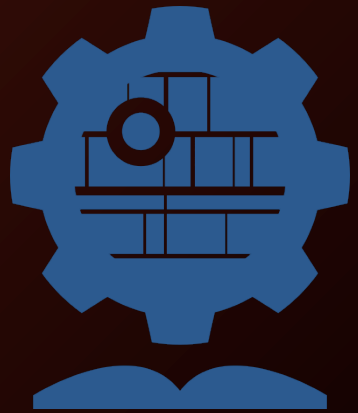


Environment Misrepresentation



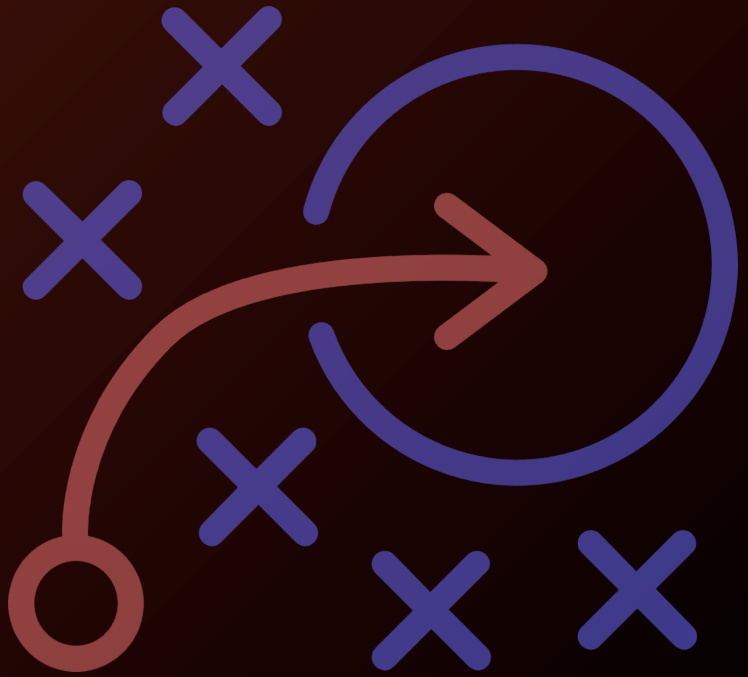
Proactive Data Offense

Building Intelligent Systems that fight



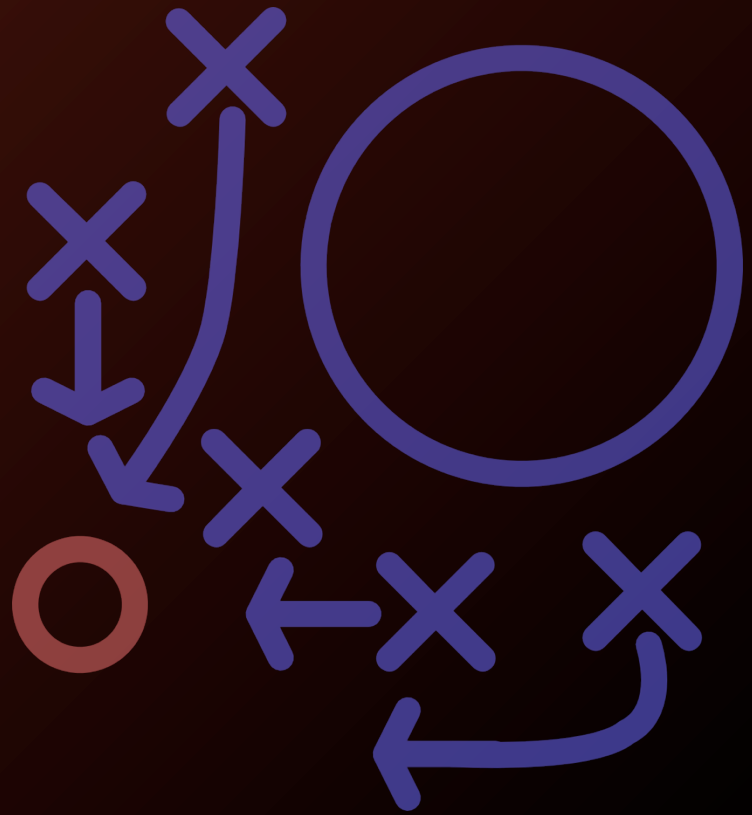
Data with Data

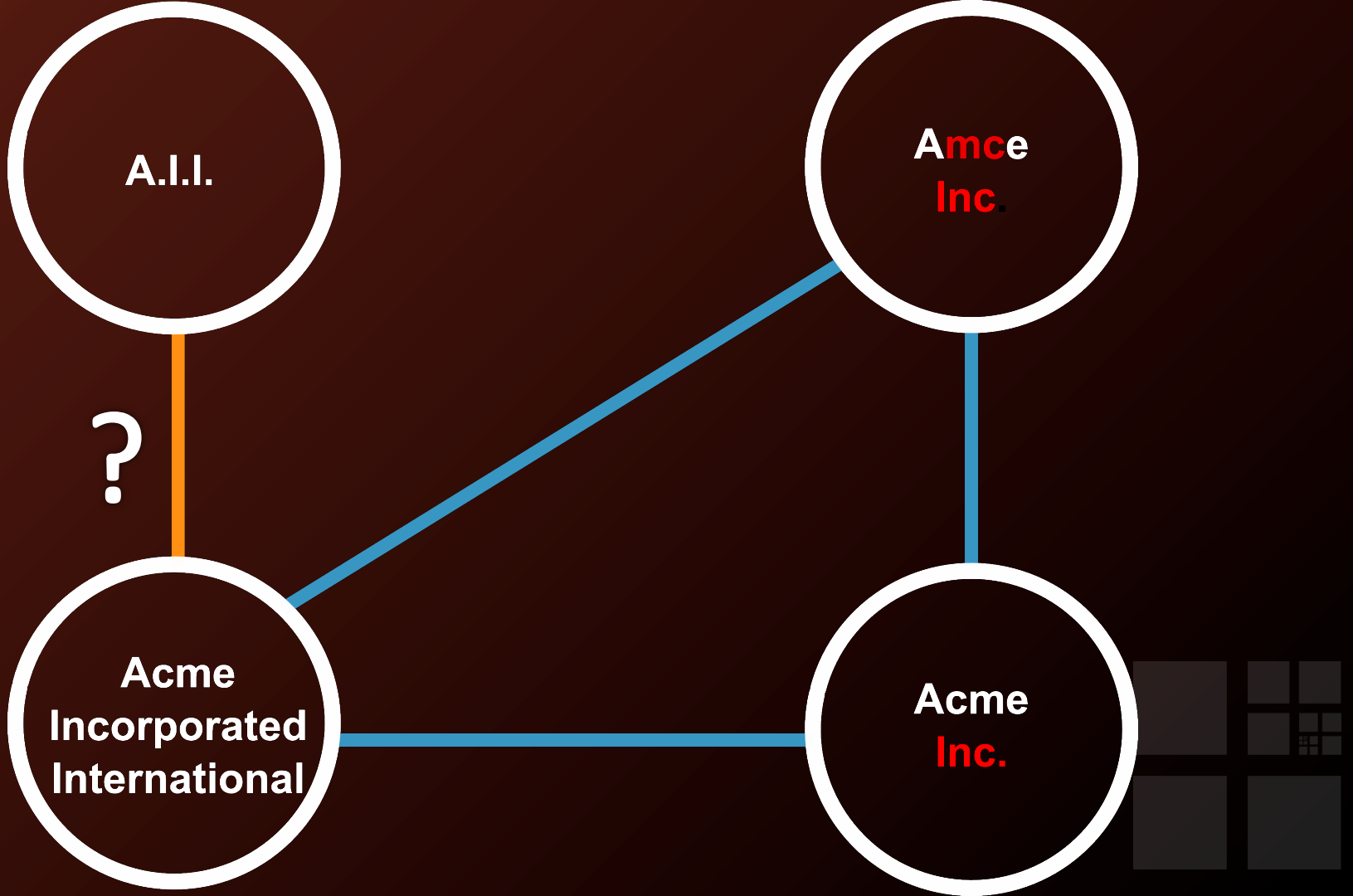
Strategy of Data Integrity Defense



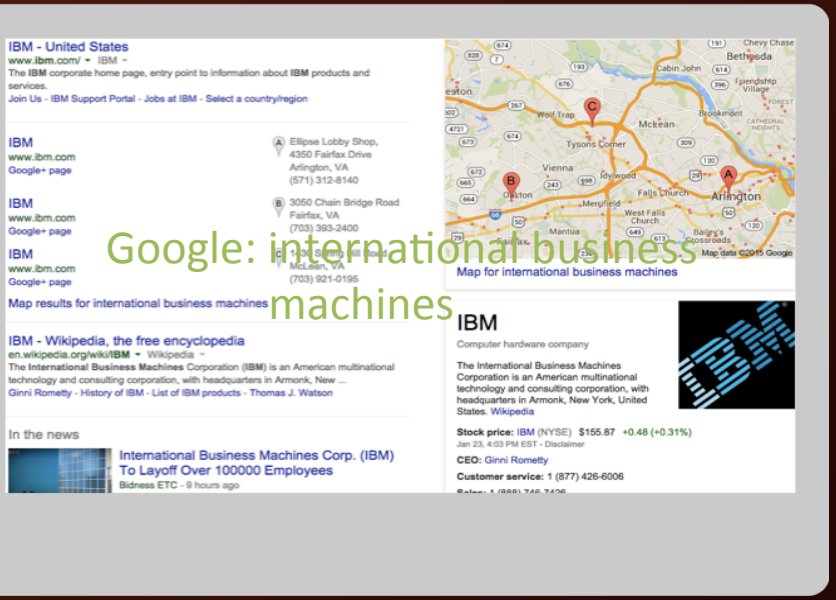
Strategy of Data Integrity Defense

- Syntactic Algorithms
- Semantic Algorithms
- Data Superiority

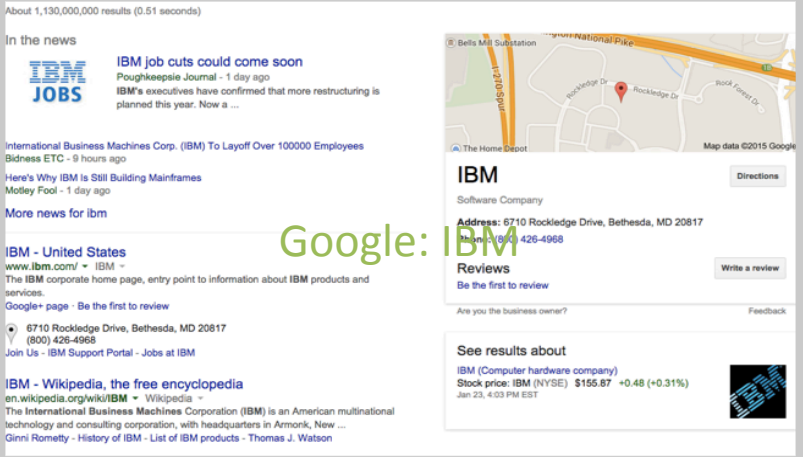



A.I.I.
Amce
Inc.
?
Acme
Incorporated
International
Acme
Inc.

How to Resolve Semantic Features?



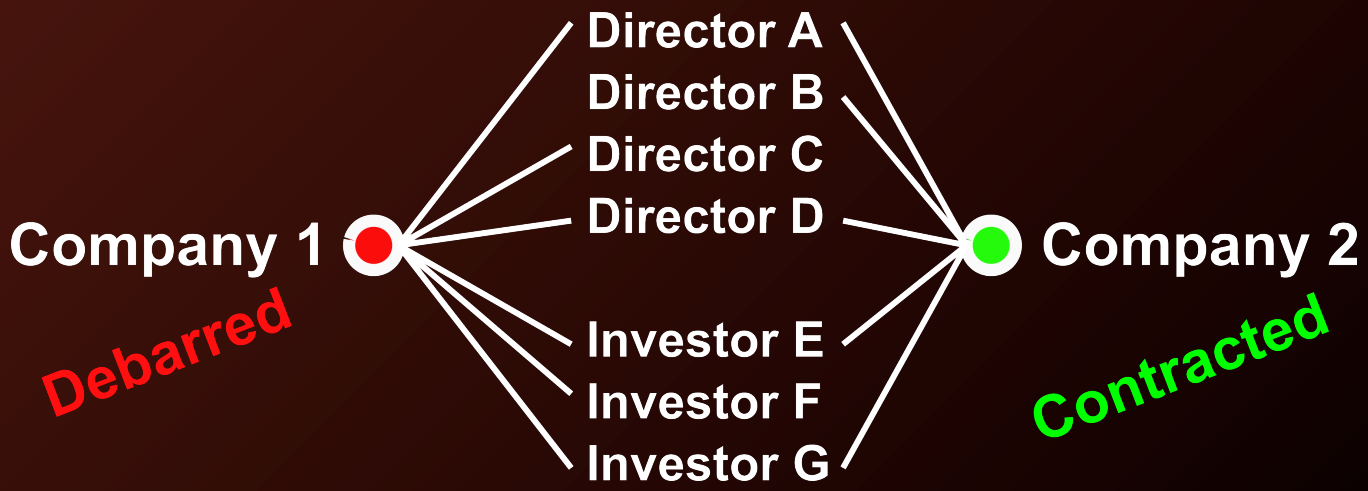
Google: international business machines



Google: IBM

Working with Our Allies

opencorporates



Mining Expert Data Systems

Southwest®

ValuJet™ AIRLINES

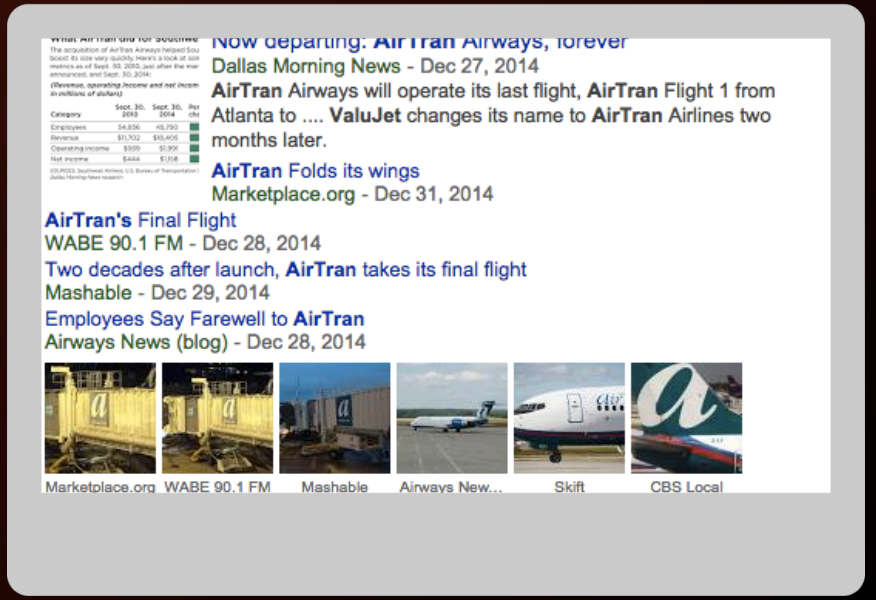
AirTran®

Mining Expert Data Systems

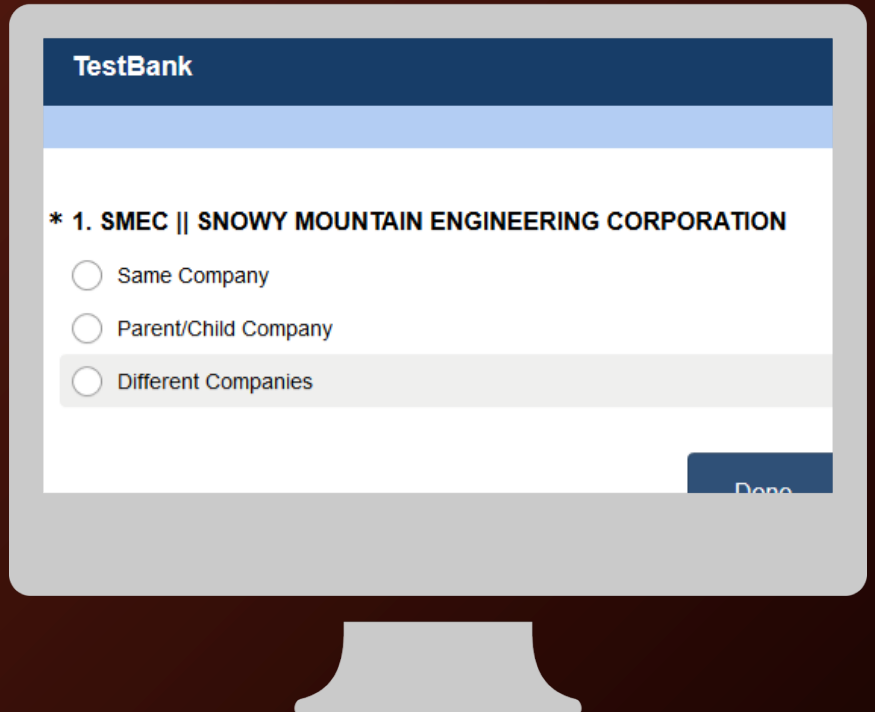
Southwest®

ValuJet AIRLINES

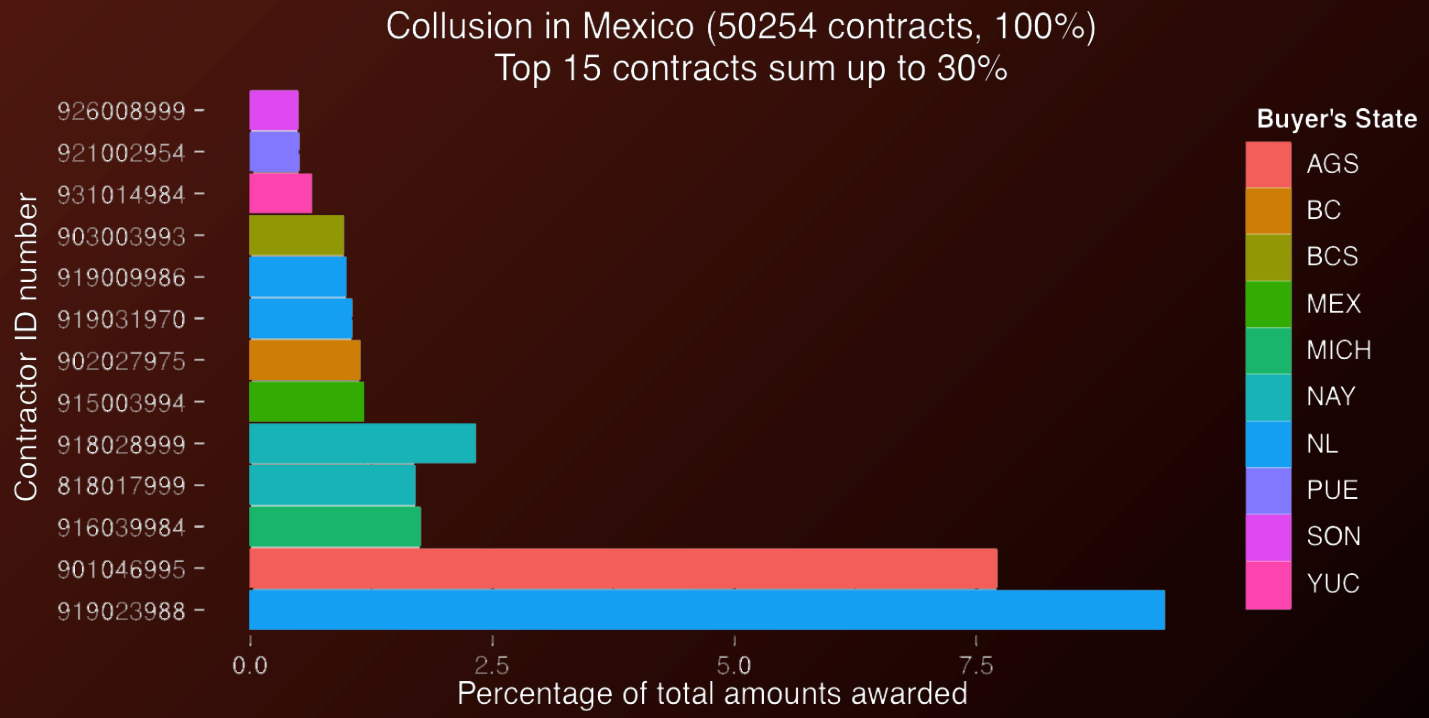
AirTran



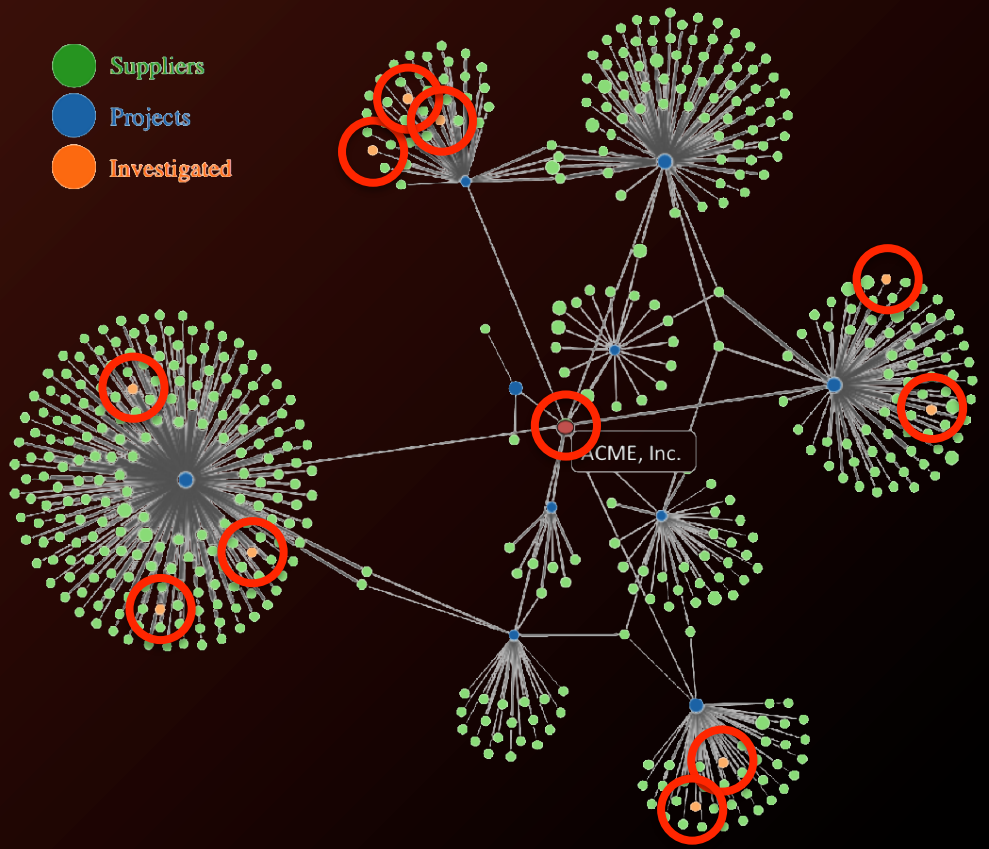
Learning from the Experts



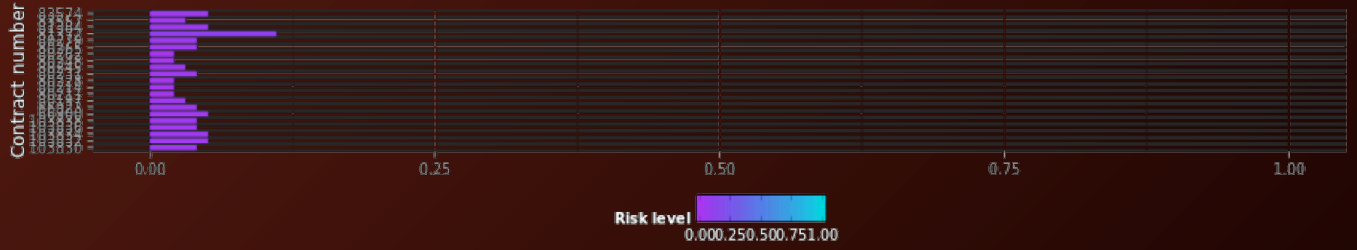
Leveraging Data Superiority

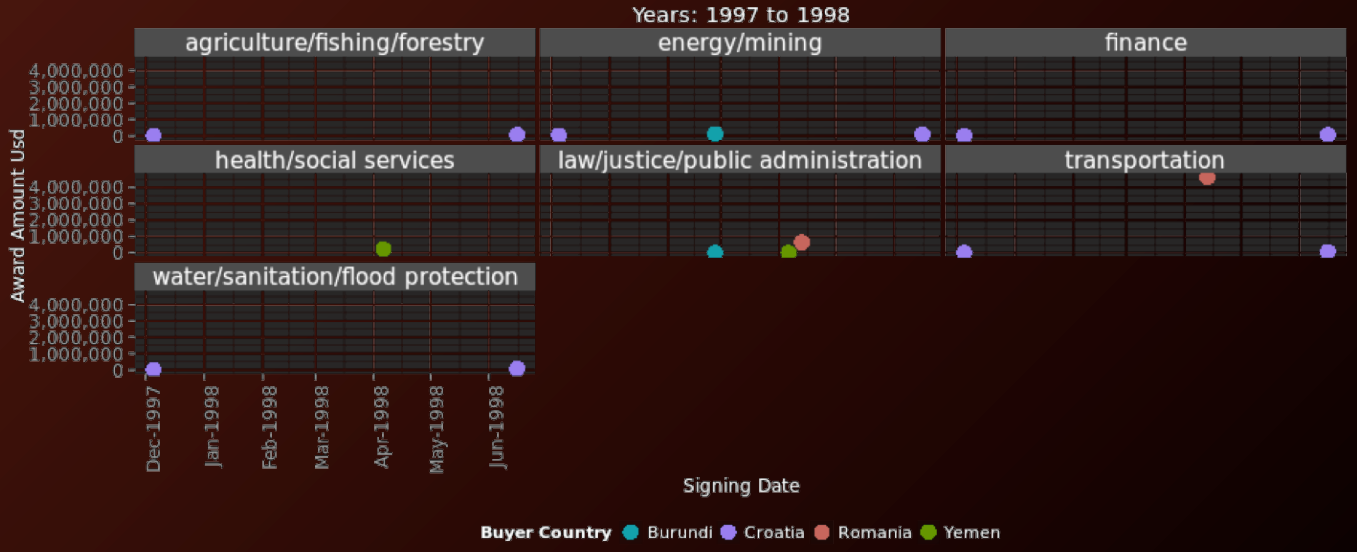


Leveraging Data Superiority

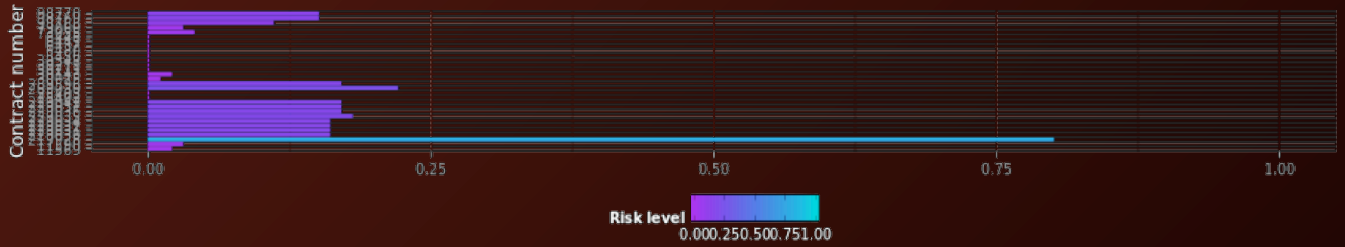


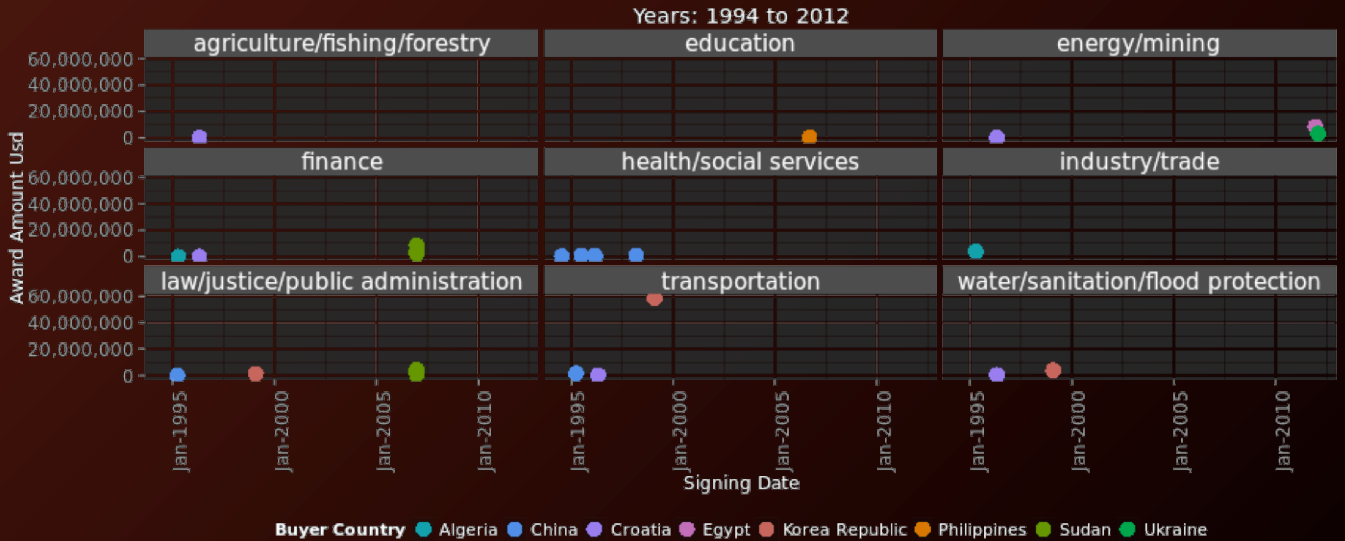
Leveraging Data Superiority



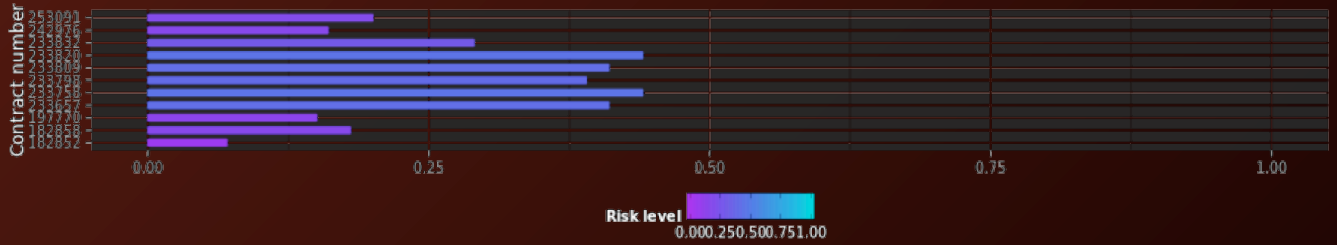


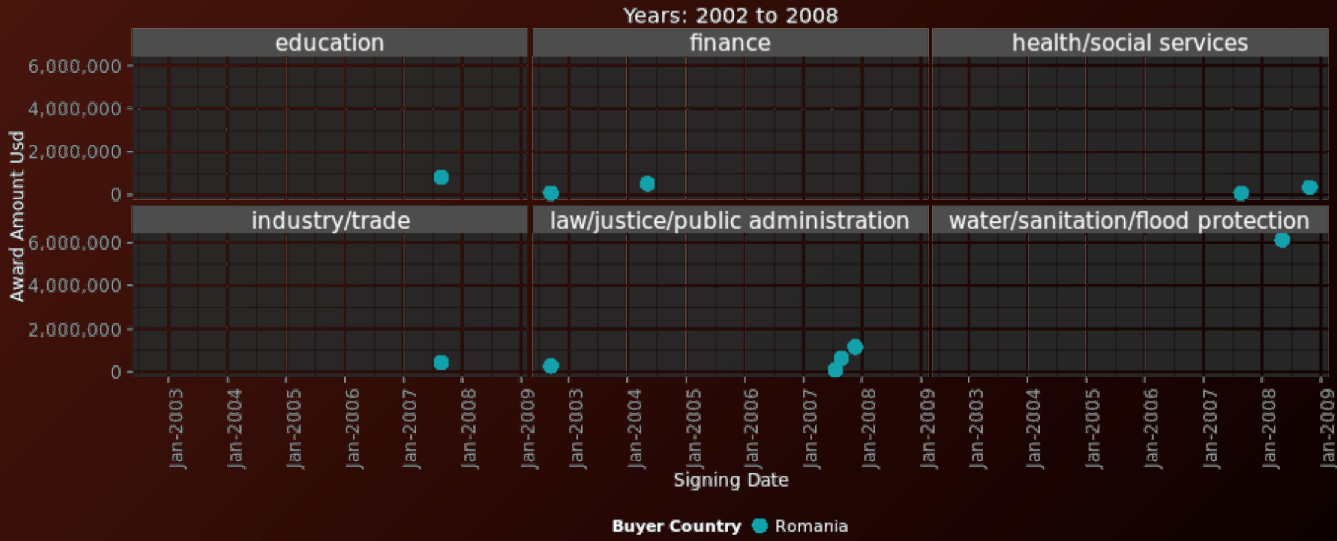
Leveraging Data Superiority



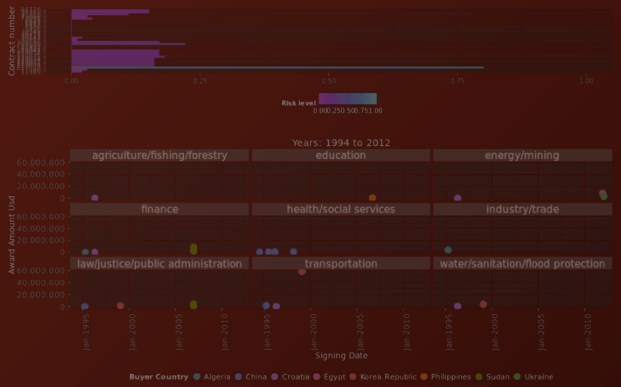


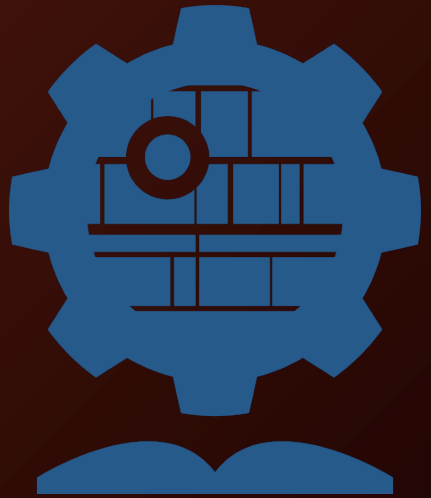
Leveraging Data Superiority

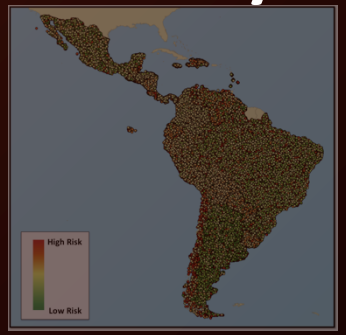


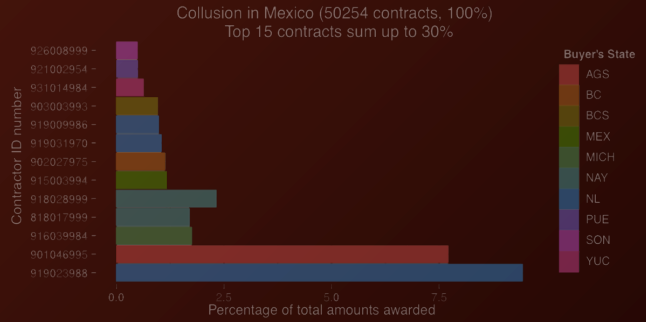


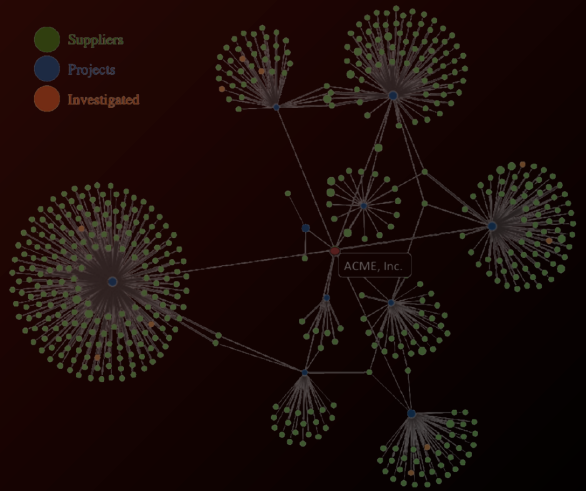
Leveraging Data Superiority





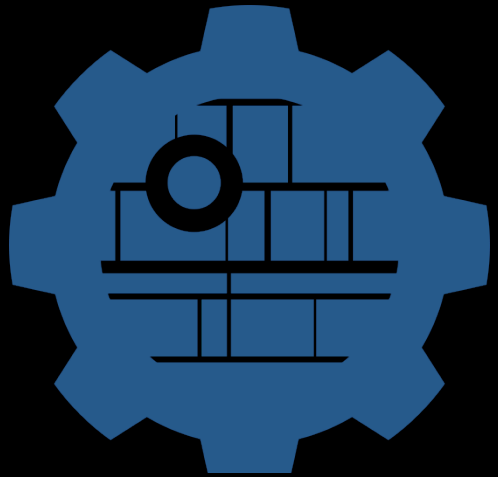






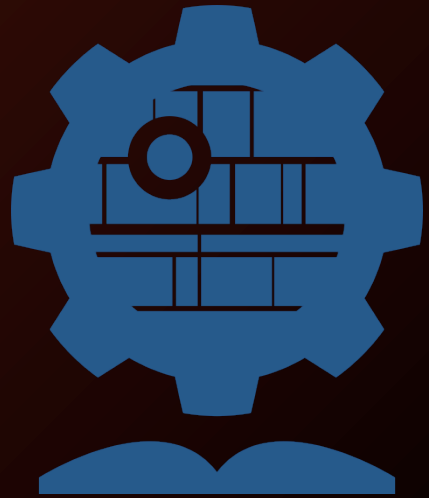
Not Just a Problem for International Finance and Development

Our Approach to Data Must Become Adversarial



Our Data Systems are Under Attack

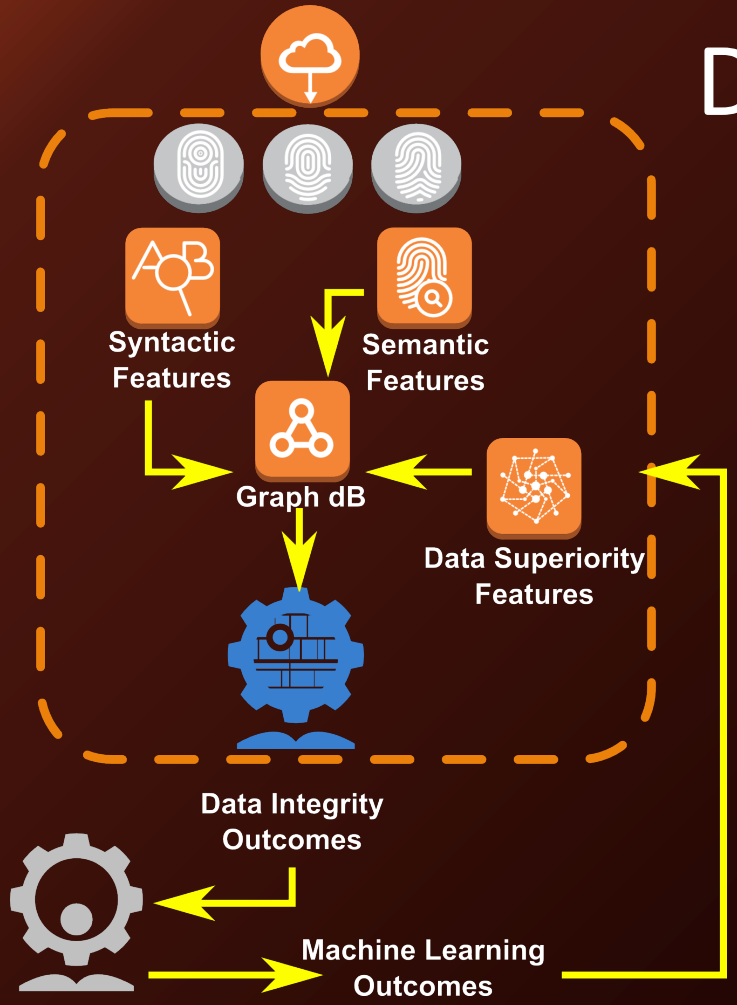
We must change our approach.

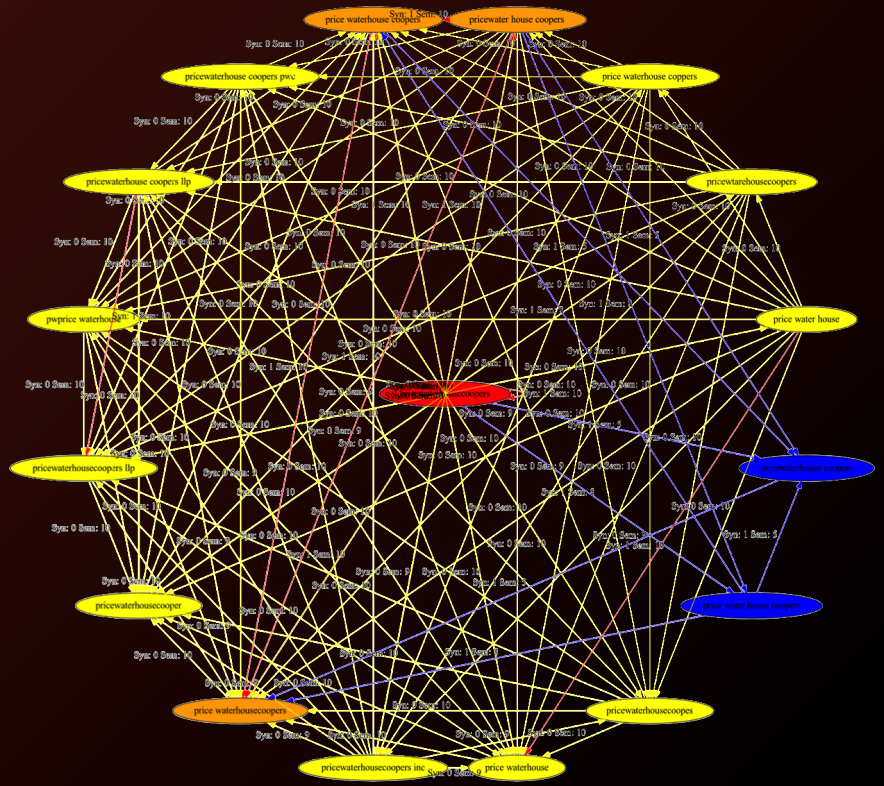


Architect our systems with data integrity defenses as

a design principle.

Data Integrity by Design





Call to Action!

An area in need of **new focus** and attention!

Help us build a community!

- Industry
- Academia
- Government

http://dataengineering.org