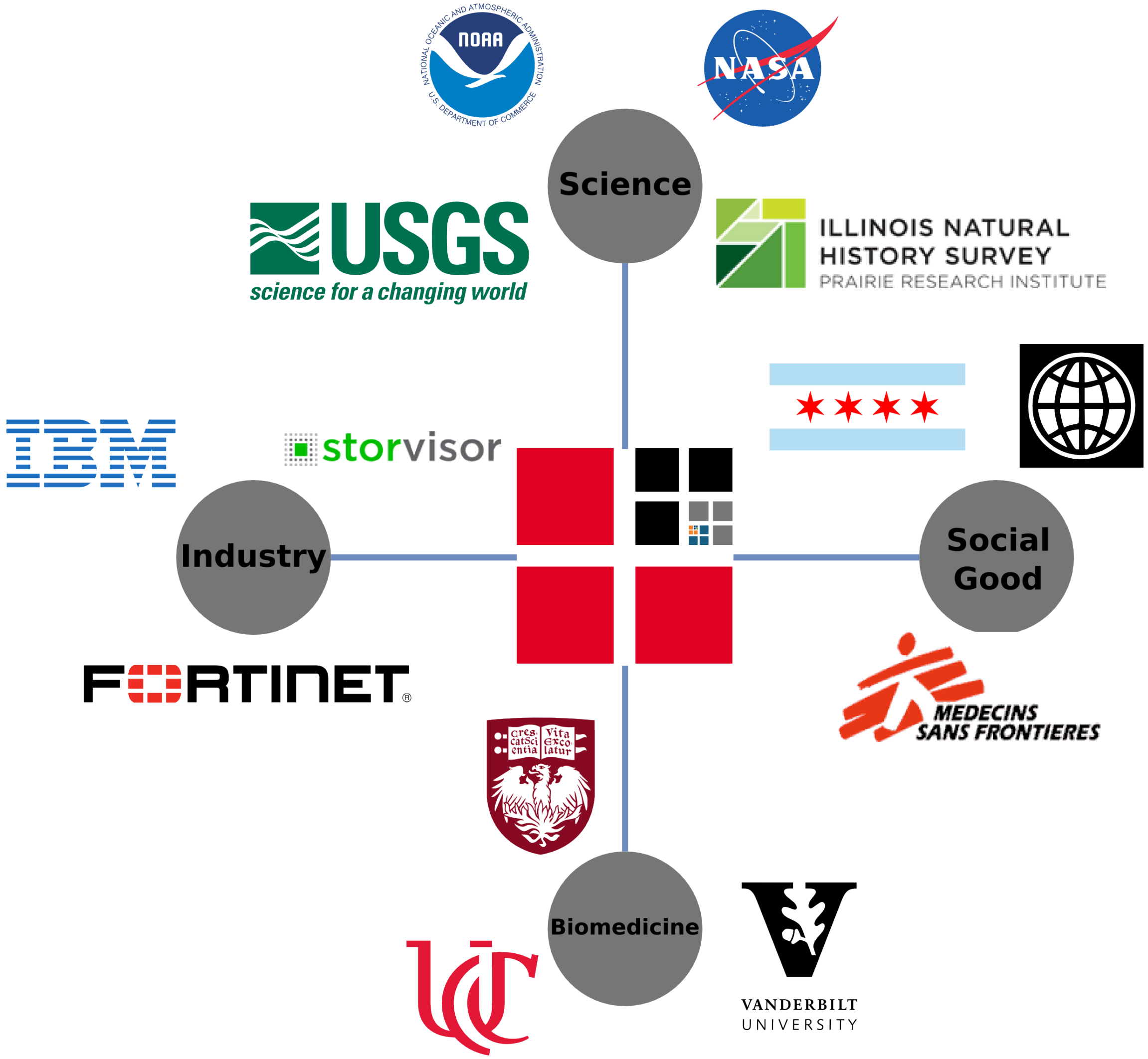
From the Data Trenches:

Using Data Science for Social Good

Eric Rozier

TRUSTWORTHY

DATA ENGINEERING



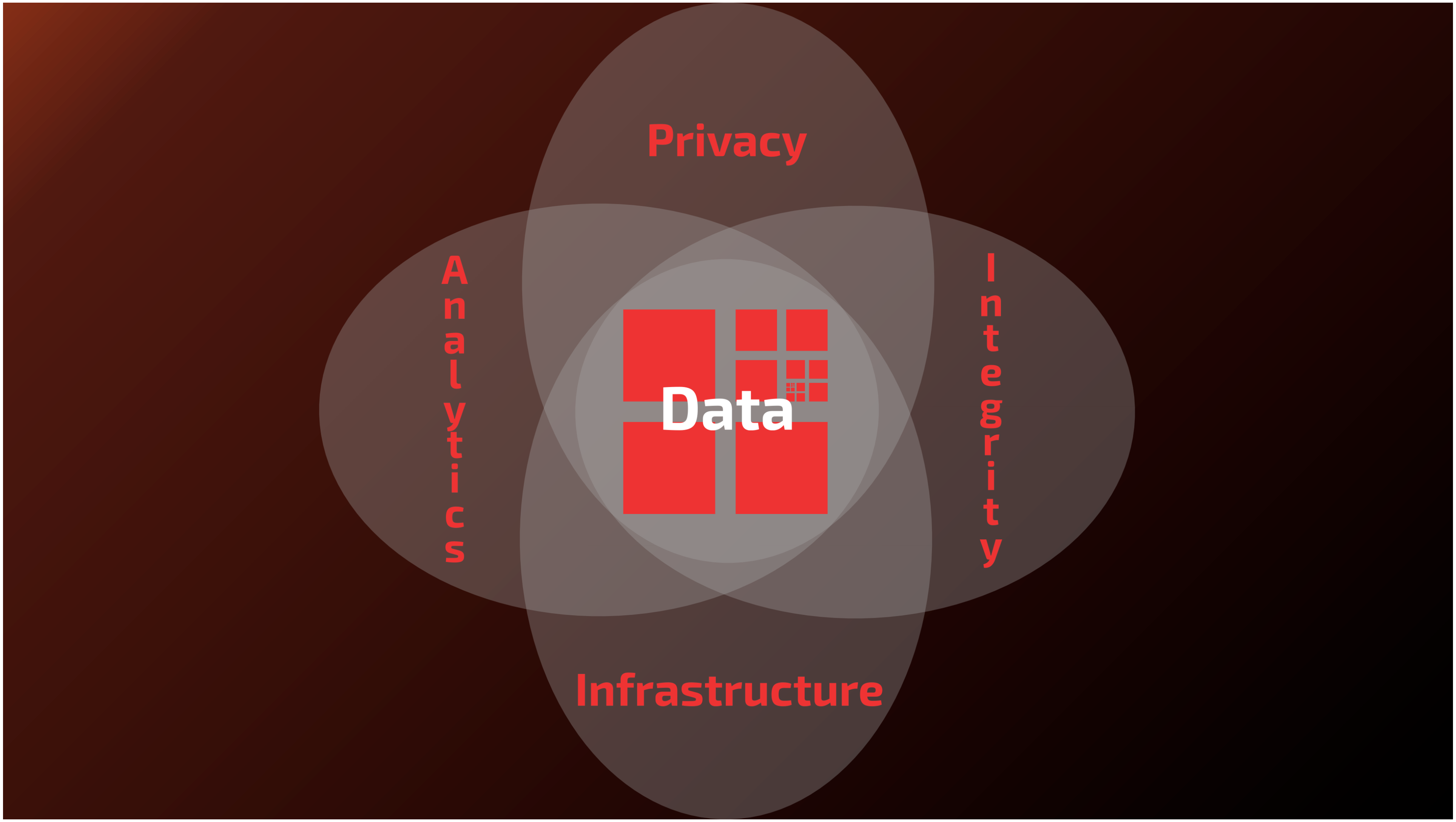
Science

Industry

Social Good

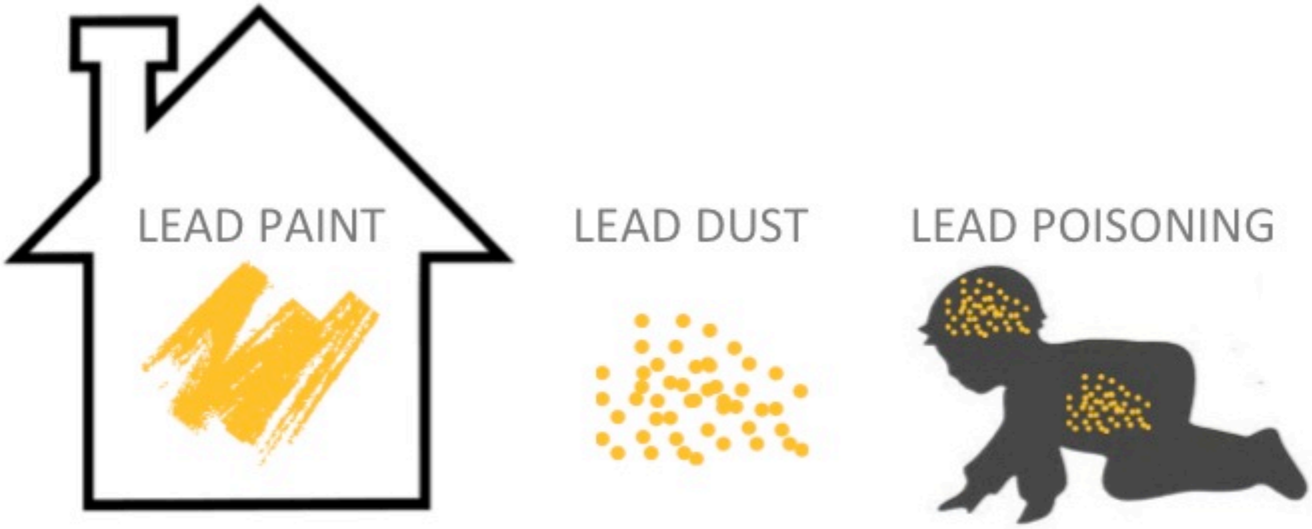
Biomedicine

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE
NASA
USGS
science for a changing world
ILLINOIS NATURAL HISTORY SURVEY
PRAIRIE RESEARCH INSTITUTE
IBM
storvisor
FORTINET
MEDECINS SANS FRONTIERES
crescat scientia vita excolatur
VANDERBILT UNIVERSITY



Privacy
Analytics
Data
Integrity
Infrastructure

How Lead Poisoning Works



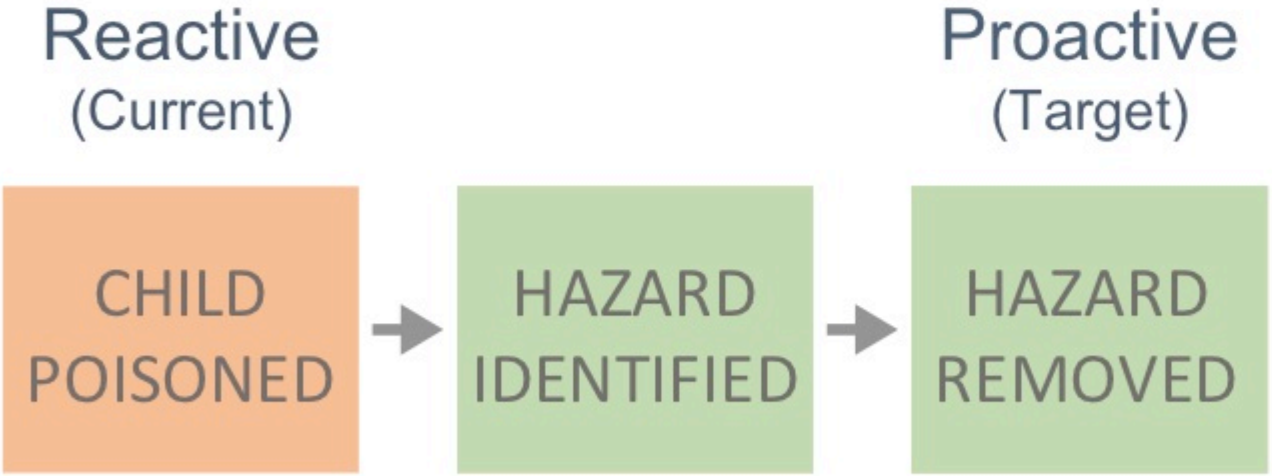
LEAD PAINT

LEAD DUST

LEAD POISONING

- NO SAFE LEVEL OF EXPOSURE
- PERMANENT HARMFUL EFFECTS

Acting On Lead Poisoning



Reactive

(Current)

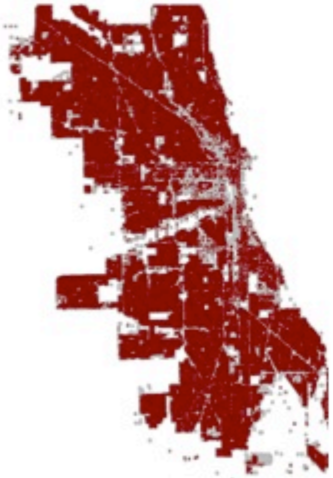
Proactive

(Target)

CHILD POISONED → HAZARD IDENTIFIED → HAZARD REMOVED

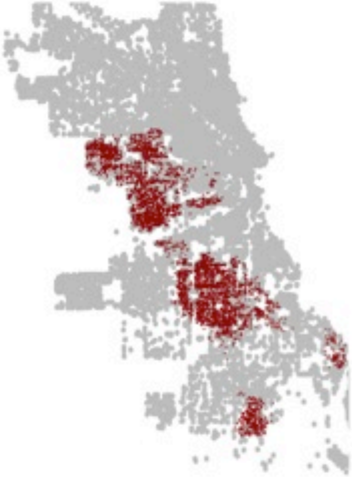
Prediction Saves Time & Money

No Prediction



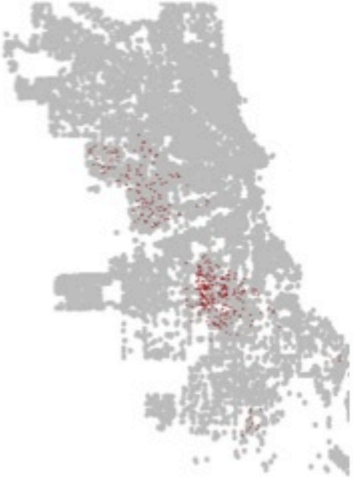
Buildings: 197,157
Time: 76 years
Money: $98 million

Current Model



Buildings: 42,695
Time: 16.4 years
Money: $21.3 million

Model Forecast



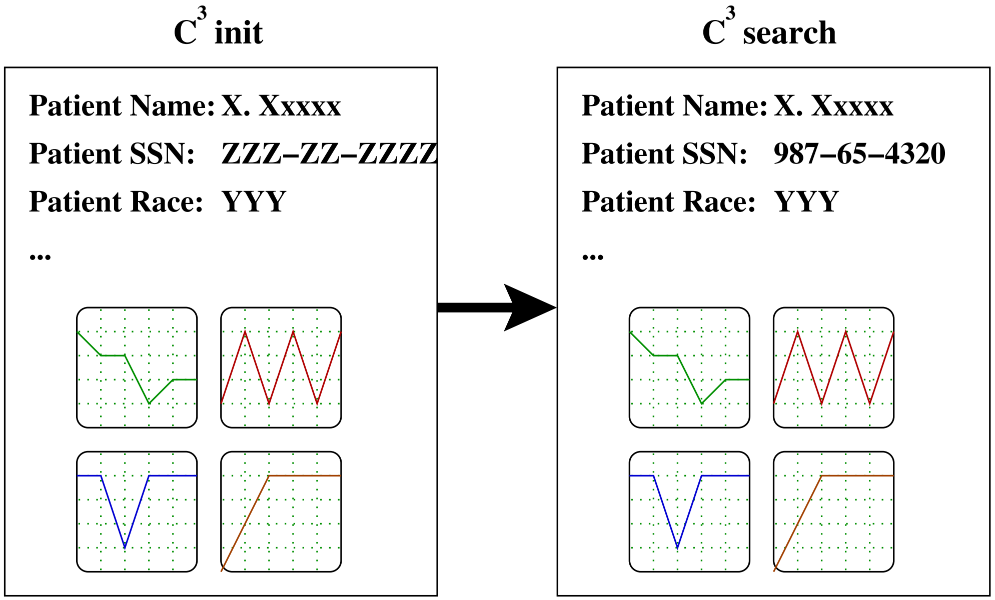
Buildings: 378
Time: 2 months
Money: $189,000

Two Problems

- Systems and data which are not protected

- Systems and data with unknown vulnerabilities

Server-Side Dedup



C^3 init

Patient Name: X. Xxxxx

Patient SSN: ZZZ–ZZ–ZZZZ

Patient Race: YYY

...

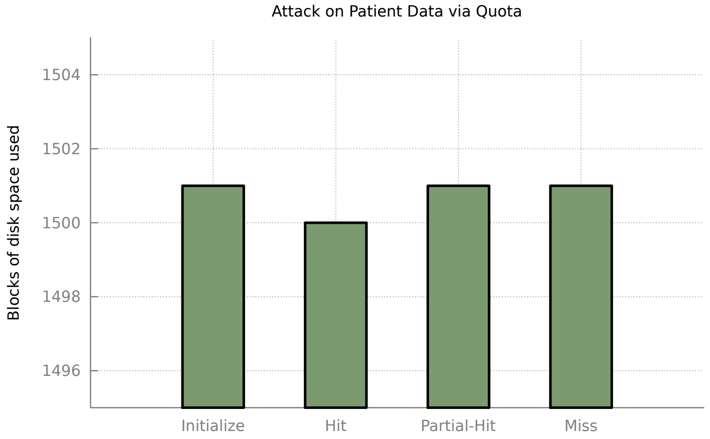
C^3 search

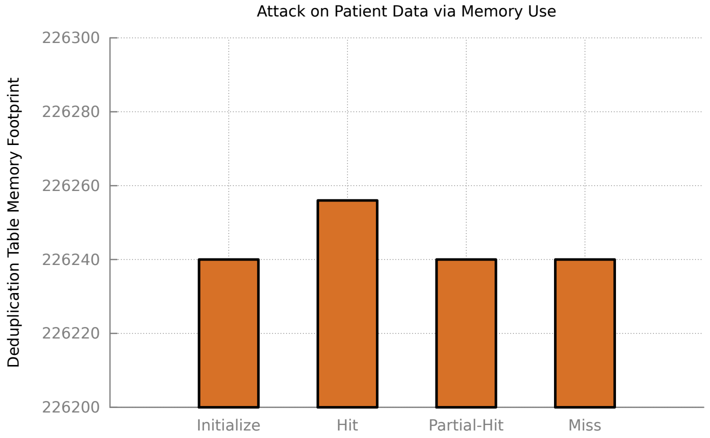
Patient Name: X. Xxxxx

Patient SSN: 987–65–4320

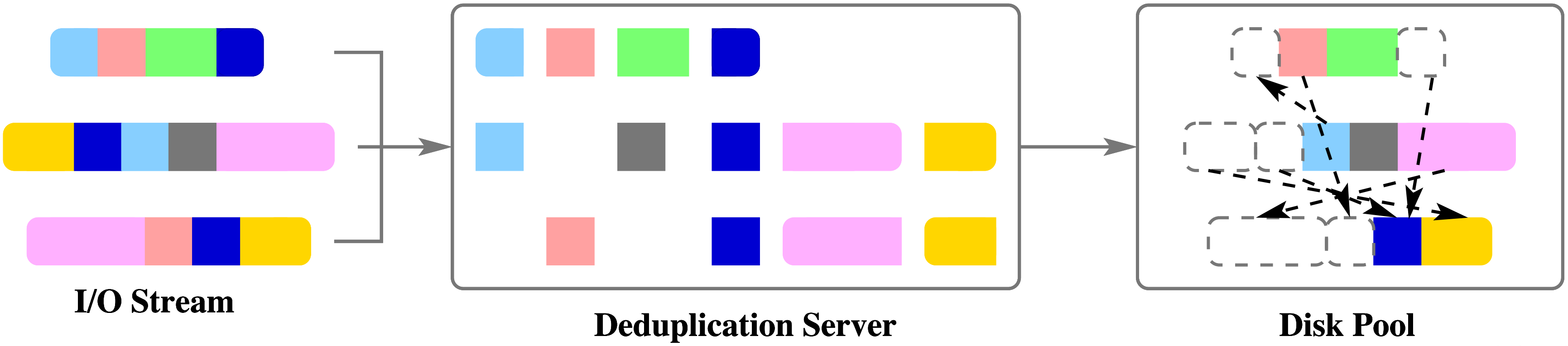
Patient Race: YYY

...

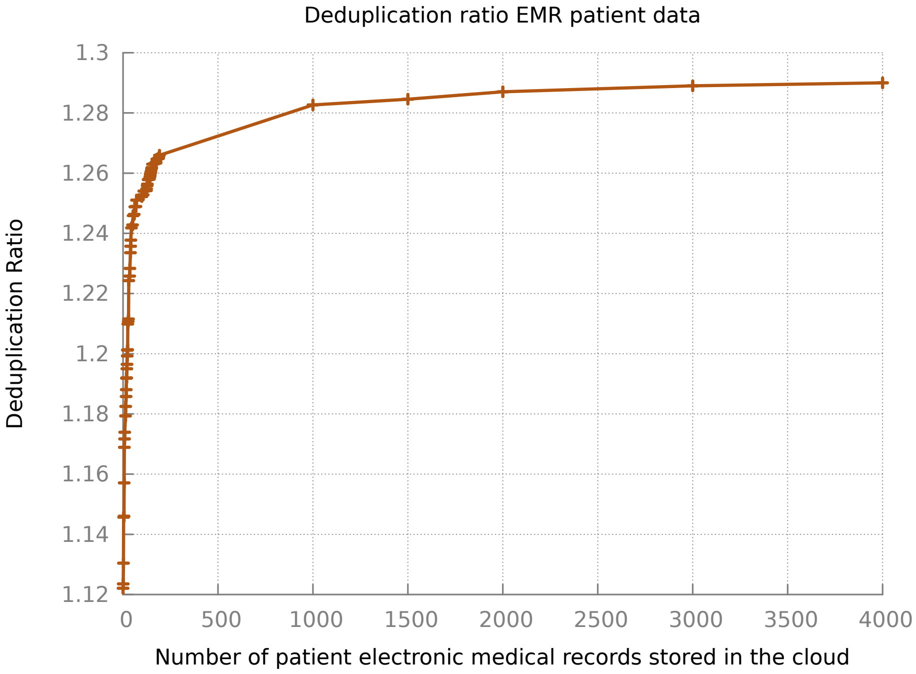


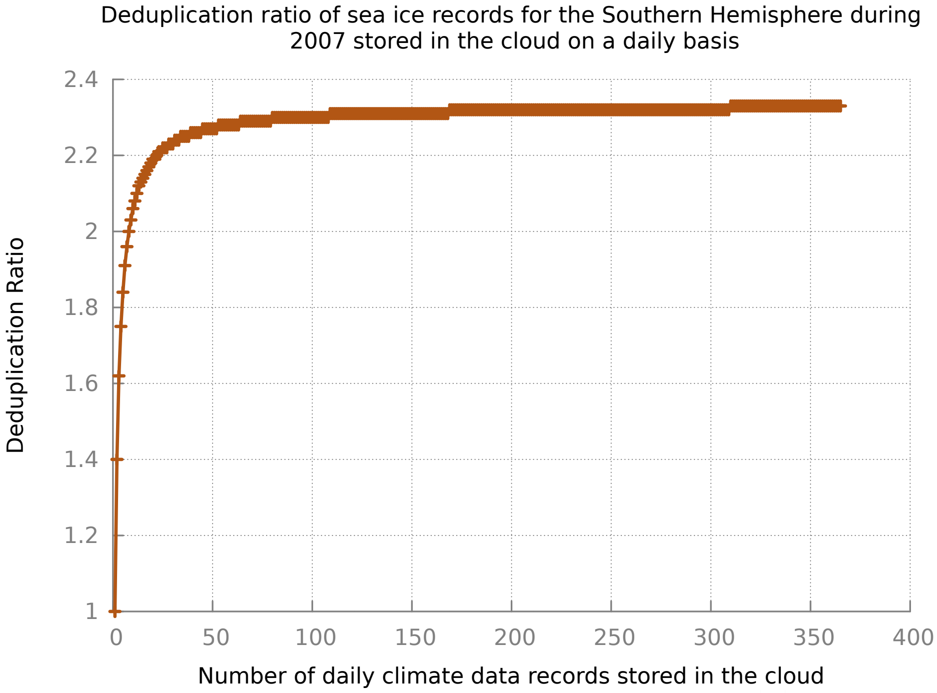


Understanding Deduplication



Benefit of Cross Deduplication?





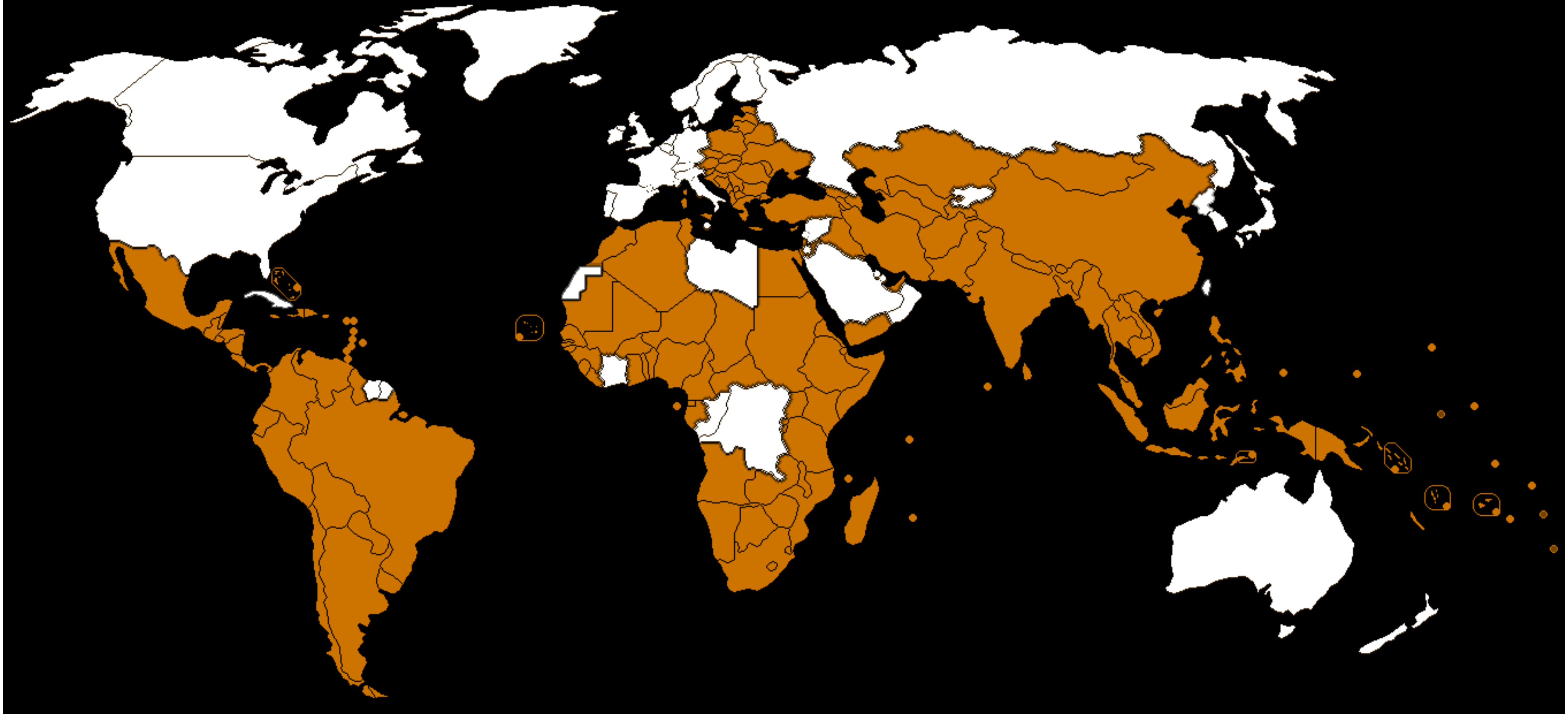
Dealing with Untrusted Storage

Availability
Security

International Procurement

WORLD BANK GROUP

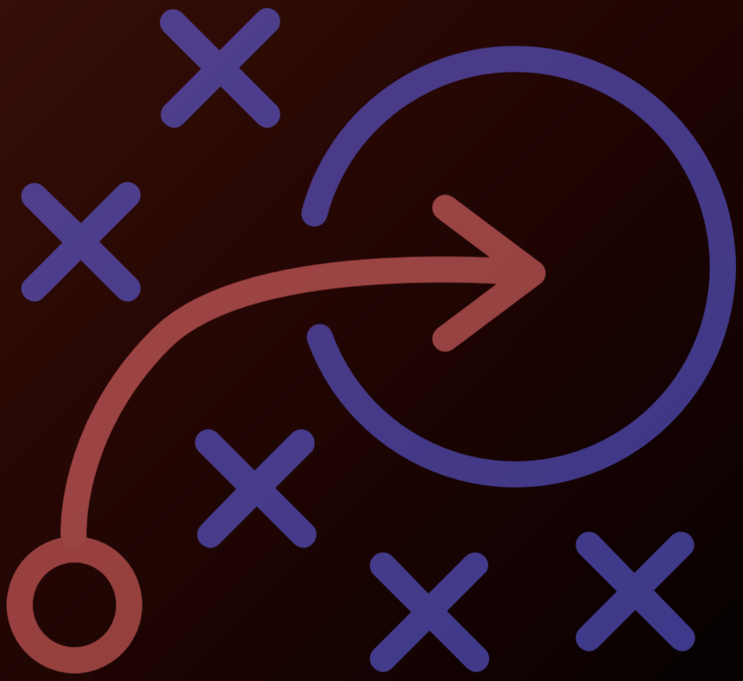
International Procurement



Automated Reasoning and Machine Intelligence

Strategy of Data Integrity Attacks

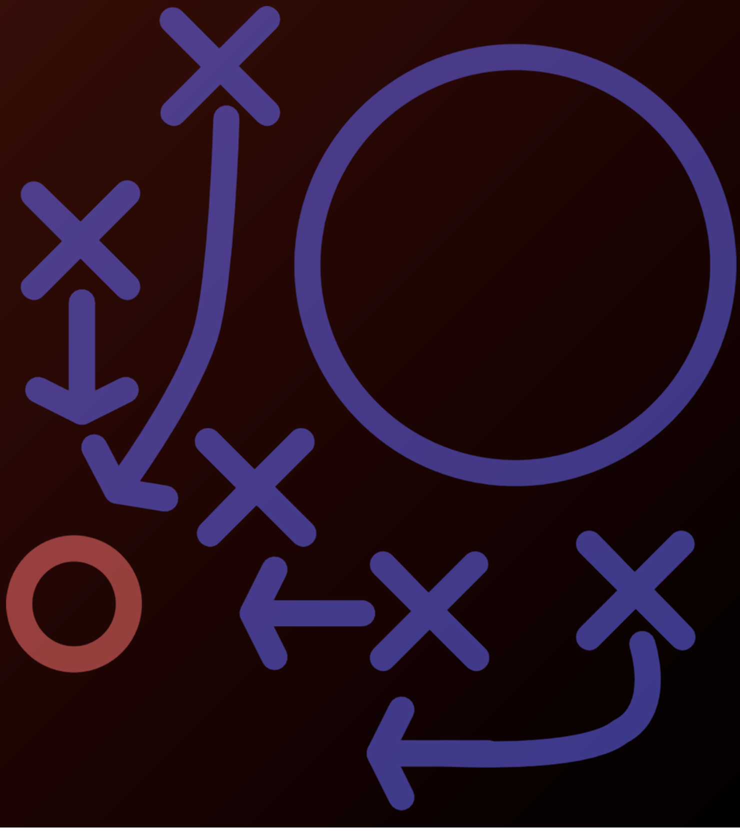
- Data pollution
- Data falsification
- Data blending



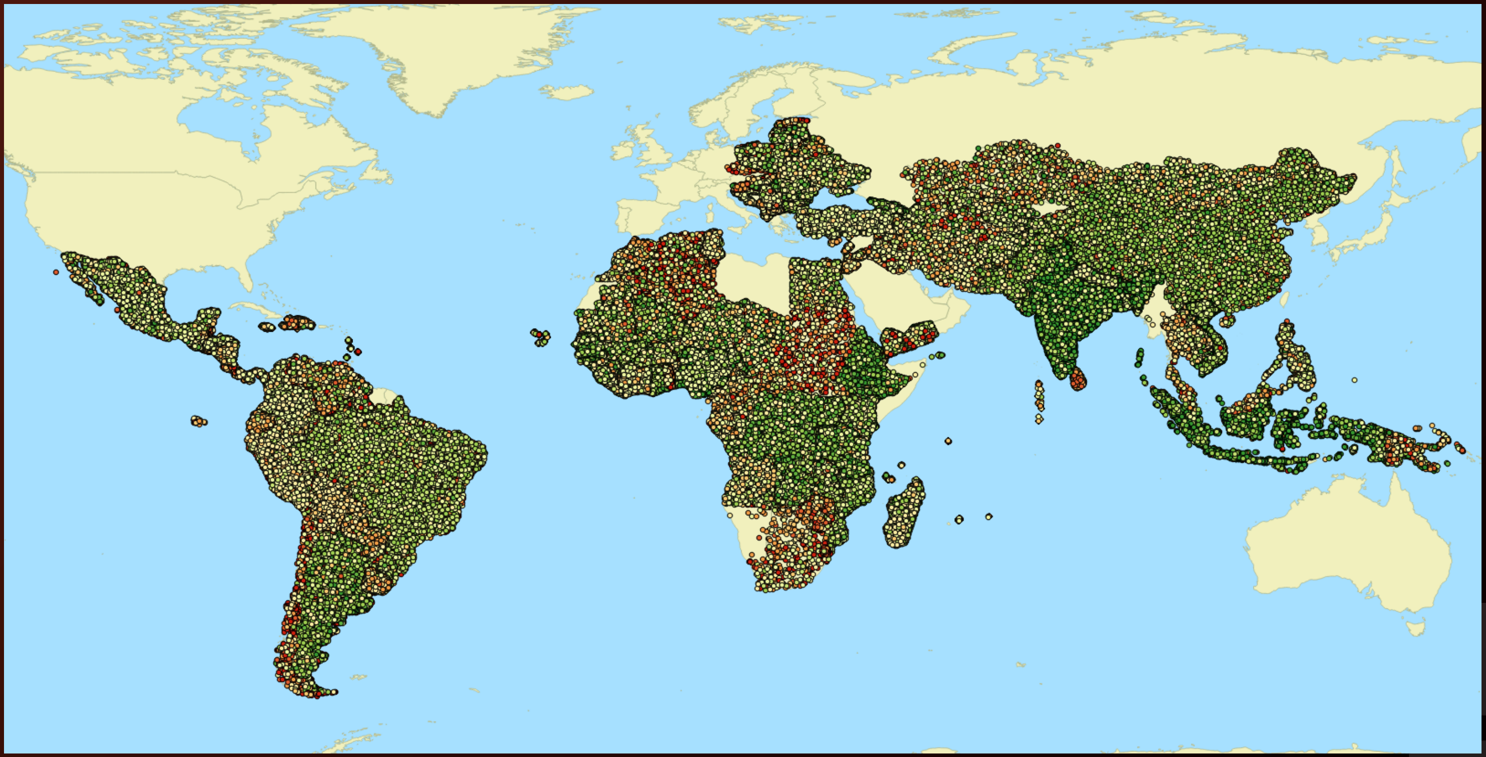
Tools Designed for Non-Malicious Environments

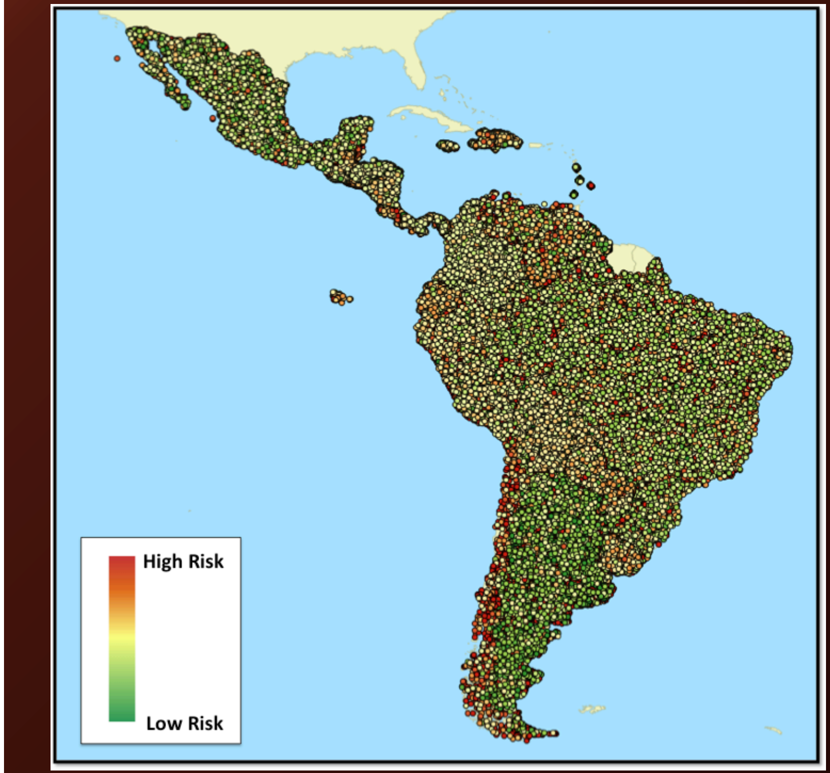
Strategy of Data Integrity Defense

- Syntactic Algorithms
- Semantic Algorithms
- Data Superiority

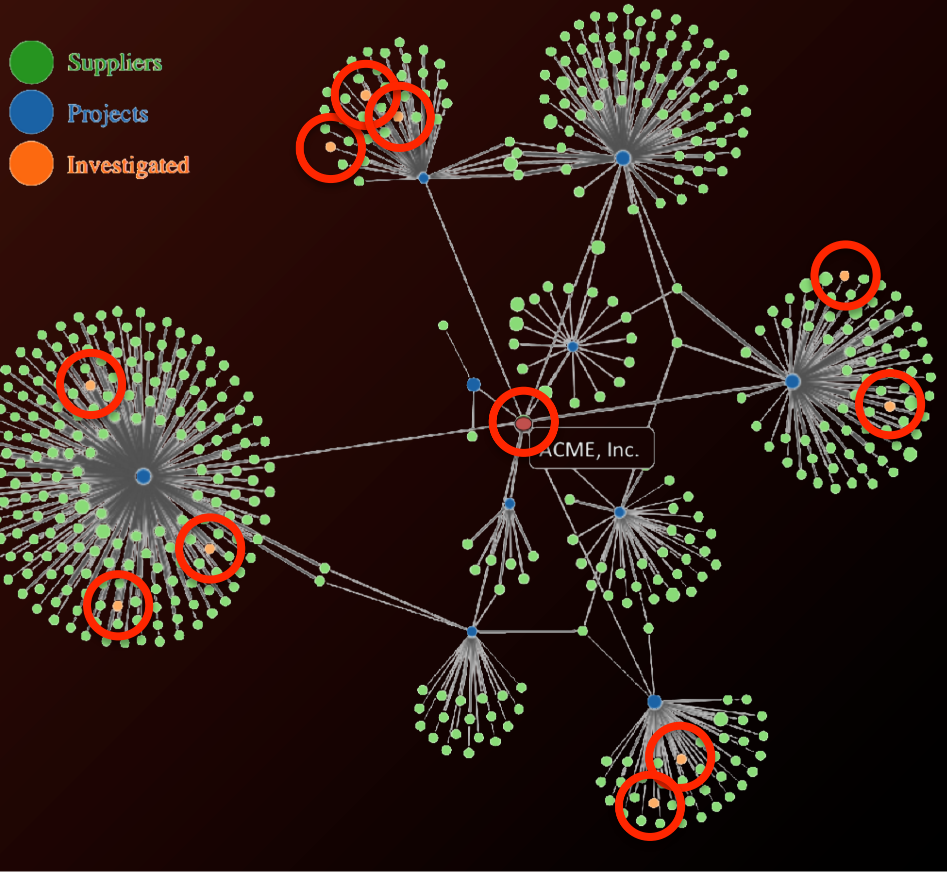


Fooled by Statistics



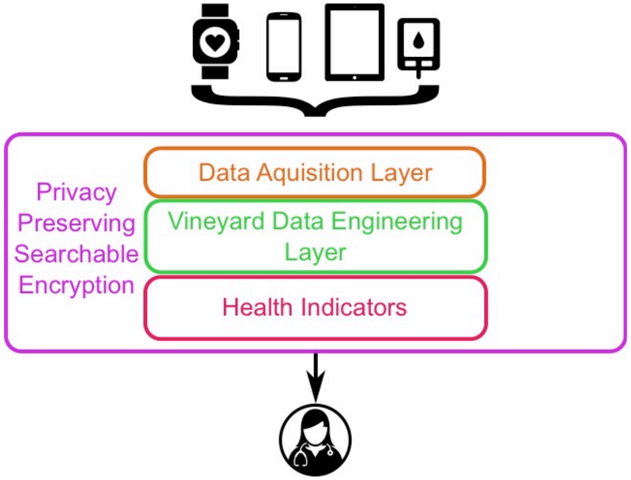


High Risk
Low Risk



Suppliers
Projects
Investigated
ACME, Inc.

Project Fraise

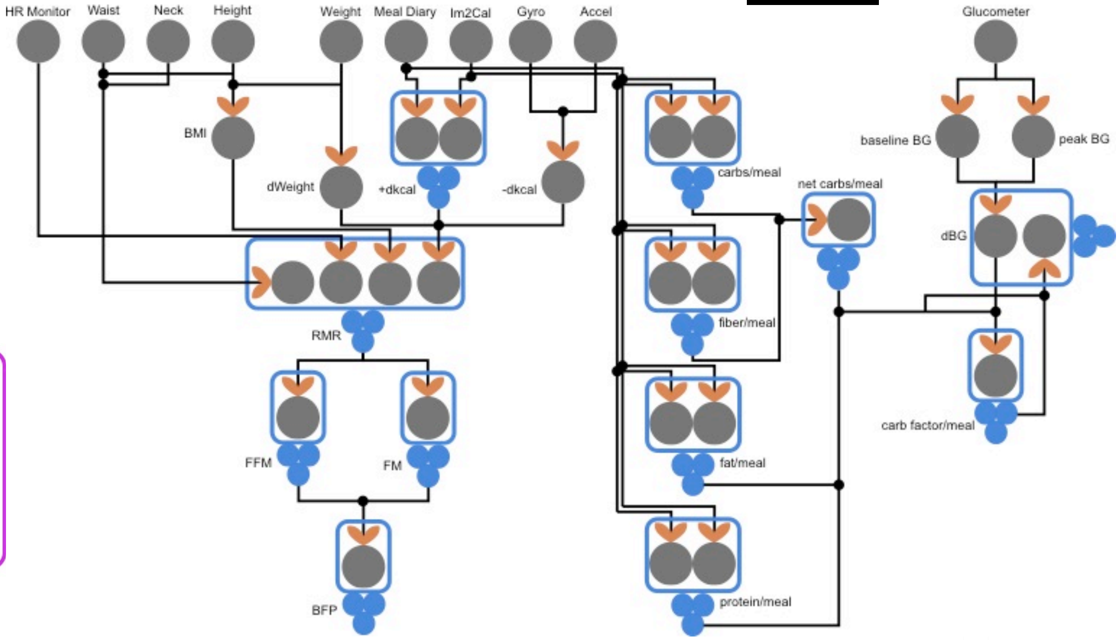


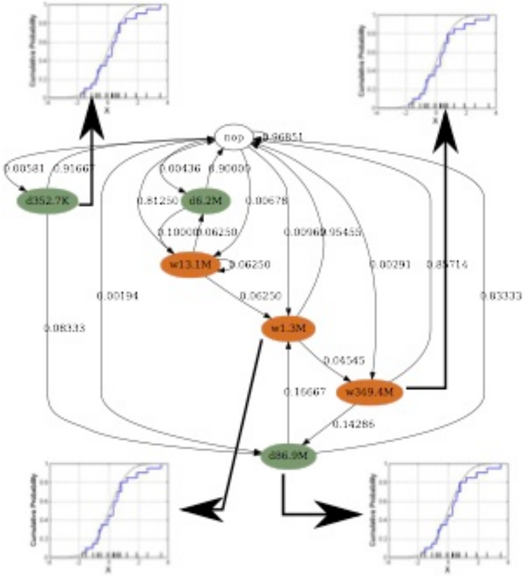
Privacy Preserving Searchable Encryption

Data Aquisition Layer

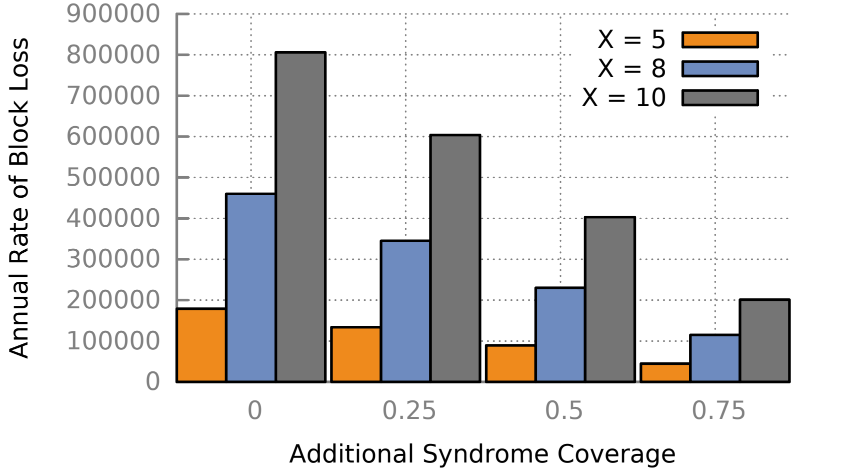
Vineyard Data Engineering Layer

Health Indicators

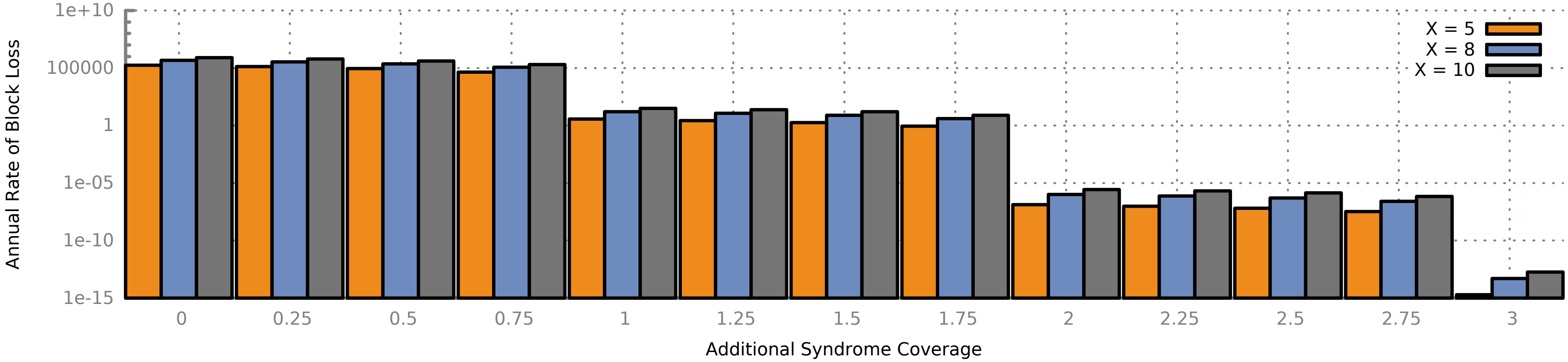


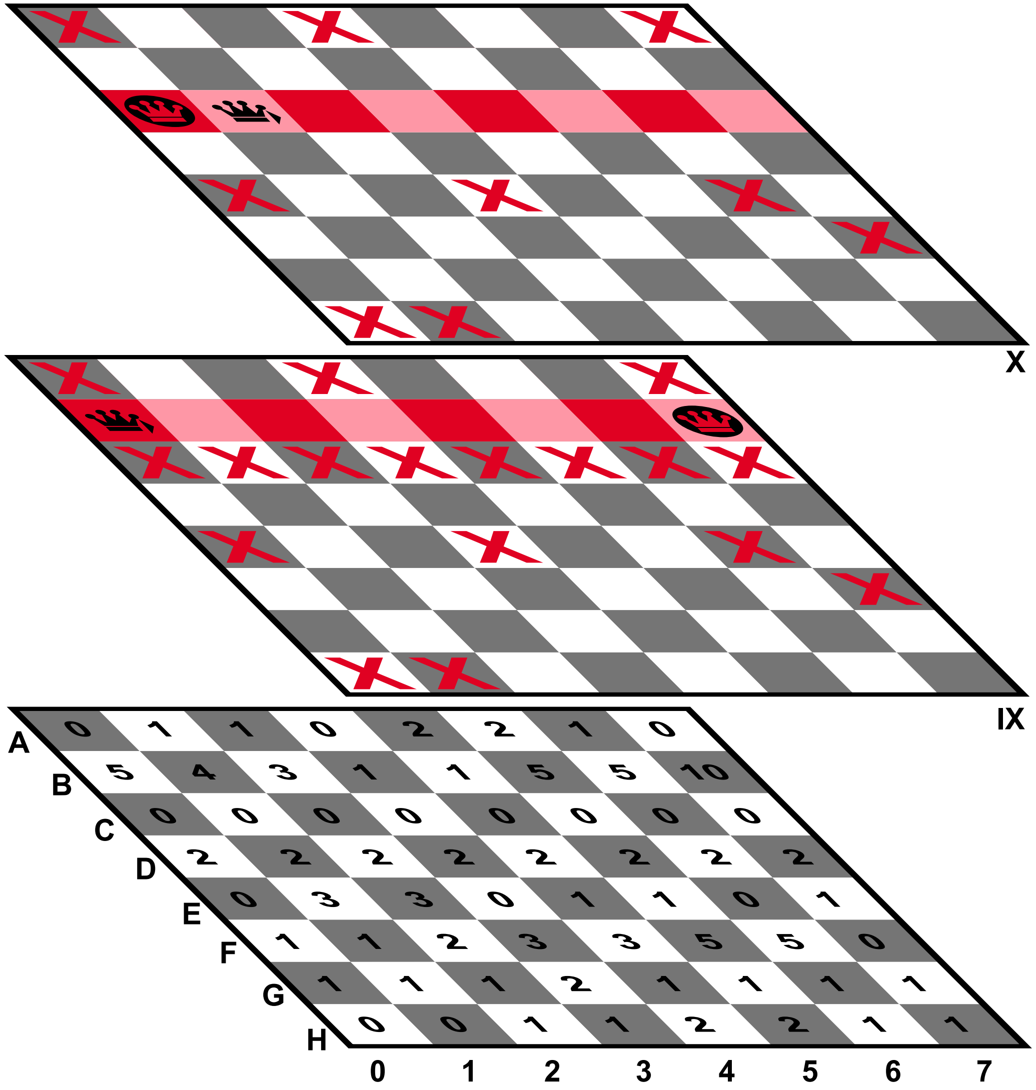


Annual Rate of Block Loss for a 1 Petabyte storage system
with X+1 initial RAID configuration.

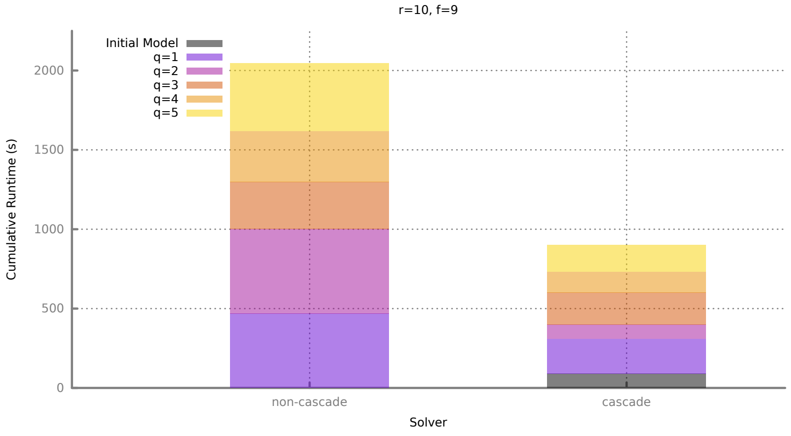


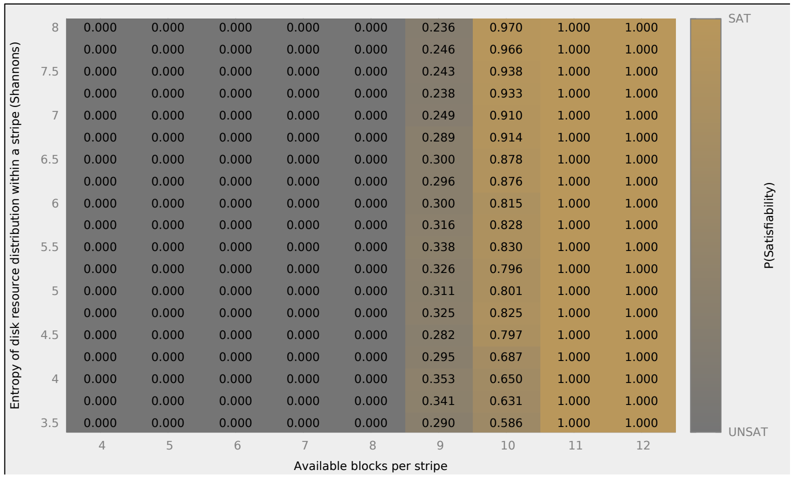
Annual Rate of Block Loss for a 1 Petabyte storage system with X+1 initial RAID configuration.





	0	1	2	3	4	5	6	7
A	0	1	1	0	2	2	1	0
B	5	4	3	1	1	5	5	10
C	0	0	0	0	0	0	0	0
D	2	2	2	2	2	2	2	2
E	0	3	3	0	1	1	0	1
F	1	1	2	3	3	5	5	0
G	1	1	1	2	1	1	1	1
H	0	0	1	1	2	2	1	1





	4	5	6	7	8	9	10	11	12
8	0.000	0.000	0.000	0.000	0.000	0.236	0.970	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.246	0.966	1.000	1.000
7.5	0.000	0.000	0.000	0.000	0.000	0.243	0.938	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.238	0.933	1.000	1.000
7	0.000	0.000	0.000	0.000	0.000	0.249	0.910	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.289	0.914	1.000	1.000
6.5	0.000	0.000	0.000	0.000	0.000	0.300	0.878	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.296	0.876	1.000	1.000
6	0.000	0.000	0.000	0.000	0.000	0.300	0.815	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.316	0.828	1.000	1.000
5.5	0.000	0.000	0.000	0.000	0.000	0.338	0.830	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.326	0.796	1.000	1.000
5	0.000	0.000	0.000	0.000	0.000	0.311	0.801	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.325	0.825	1.000	1.000
4.5	0.000	0.000	0.000	0.000	0.000	0.282	0.797	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.295	0.687	1.000	1.000
4	0.000	0.000	0.000	0.000	0.000	0.353	0.650	1.000	1.000
	0.000	0.000	0.000	0.000	0.000	0.341	0.631	1.000	1.000
3.5	0.000	0.000	0.000	0.000	0.000	0.290	0.586	1.000	1.000

- How do we protect data?
- How do we ensure data integrity?
- How do we engineer systems for the new adversarial environment?