#HeavyD — Stopping Malicious Attacks Against Data Mining and Machine Learning

Davi Ottenheimer @daviottenheimer

Senior Director of Trust, EMC

In-Depth Seminars – D22

Agenda

- Introduction
- Threats to Machine Learning
- Detection and Stopping Attacks

Disclaimers

- Math and Stats
- Comp Sci
- Human Behavior
  - Anthropological
  - Political
  - Philosophical
  - Historical

"Automated" Vehicles Crashing and Exploding

"He says his tribe doesn't have a written language!"

INTRODUCTION TO DATA MINING AND MACHINE LEARNING

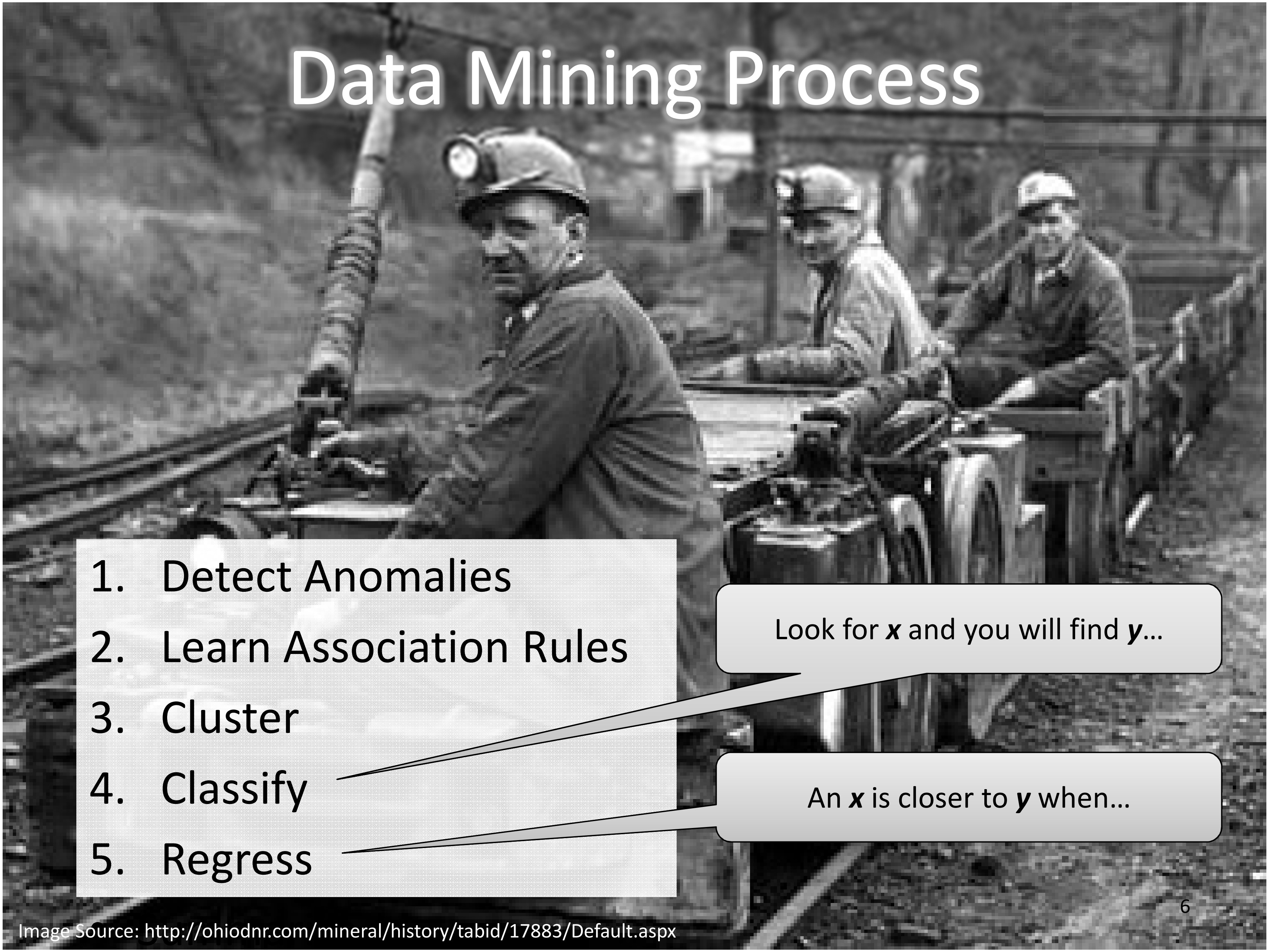
What is Data Mining?

Discover and Generate New Knowledge

Through Large Data Set Examination

- Data ***Archaeology***
- Information Harvesting
- Information Discovery
- Knowledge Extraction
- Knowledge Discovery
- Multivariate Statistics
- Pattern Recognition
- Advanced/Predictive Analysis
- Machine Learning…

Data Mining Process



1. Detect Anomalies
2. Learn Association Rules
3. Cluster
4. Classify
5. Regress

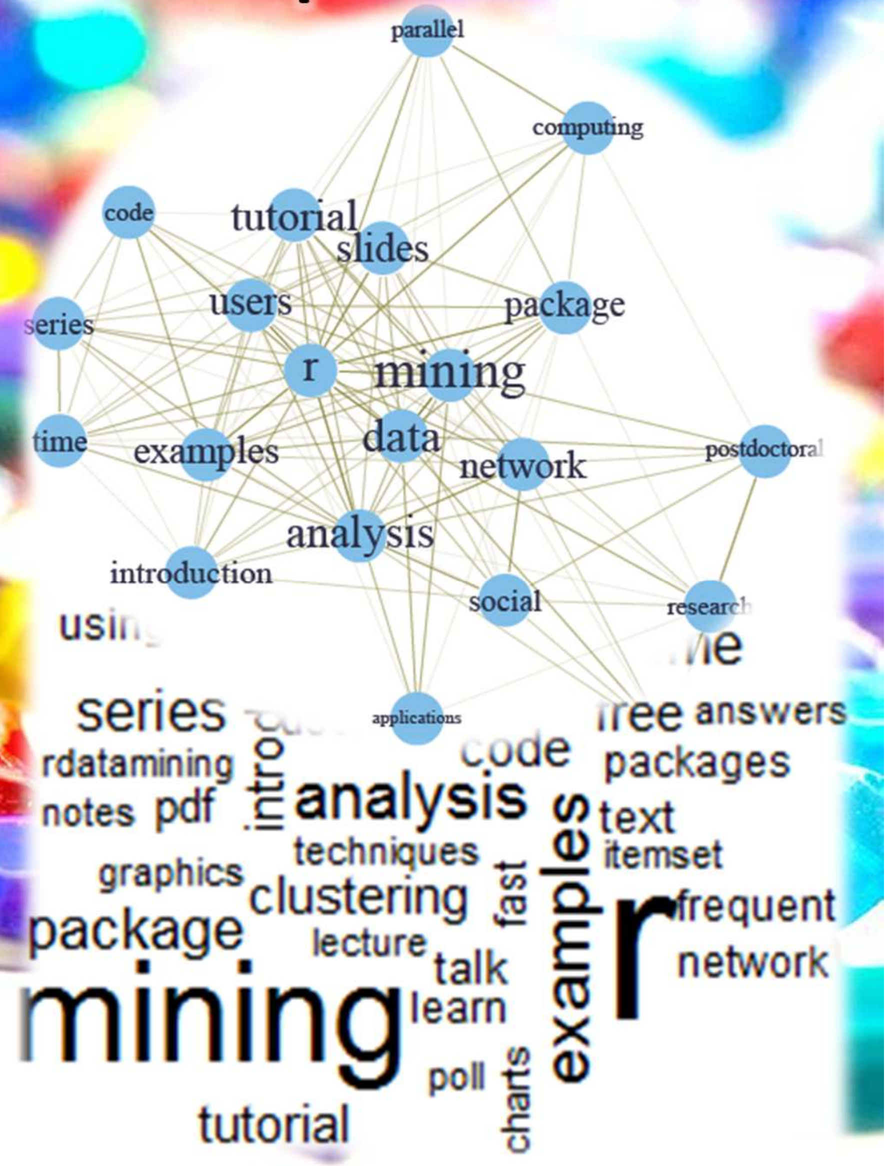
Look for ***x*** and you will find ***y***…

An ***x*** is closer to ***y*** when…

Image Source: http://ohiodnr.com/mineral/history/tabid/17883/Default.aspx

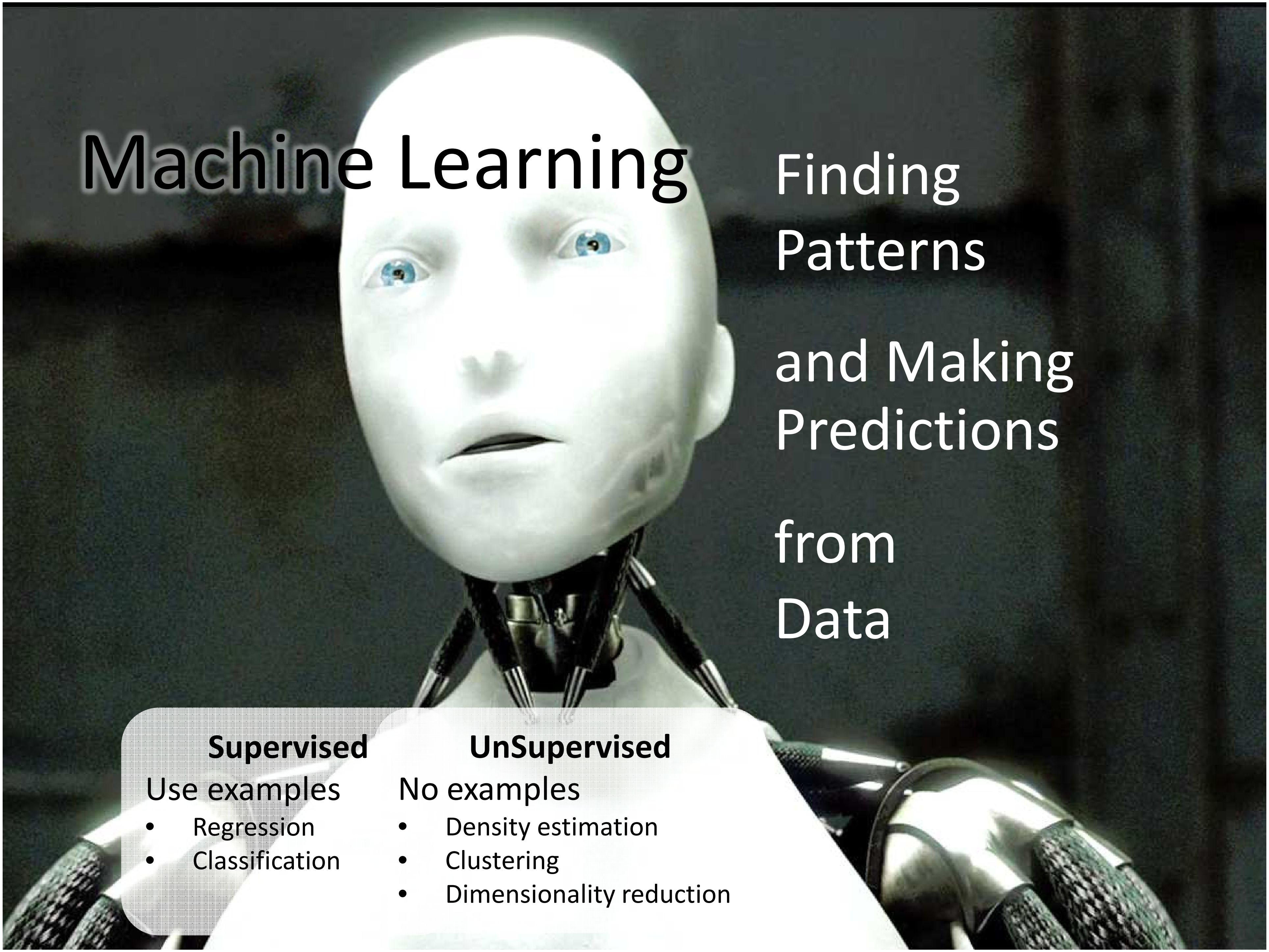
Data Mining Examples

- Find Similar Objects
- Find Object Likelihood
- Predict Category
- Predict Number
- Reduce Columns
- Find Groups
- Compare



https://commons.wikimedia.org/wiki/File:Gems_P4061909.jpg, http://www.rdatamining.com/examples/


Already Found in Many Industries
Finance
Retail
Online
Casino
Travel
Insurance
Image Source: http://www.should-know.com/yanacocha/yanacocha-gold-mine-02/
8



Machine Learning

Finding Patterns

and Making Predictions

from Data

Supervised

Use examples

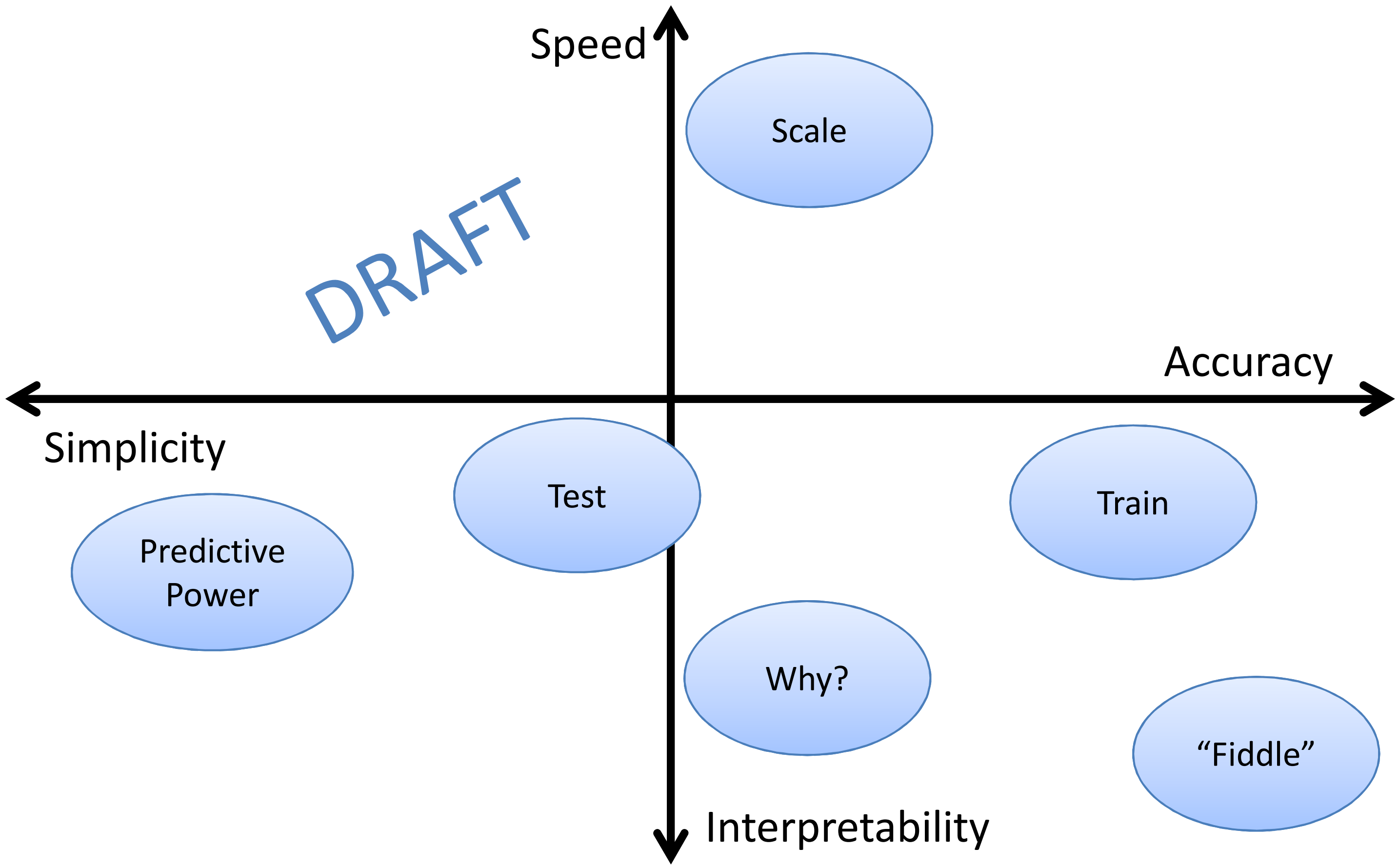
- Regression
- Classification

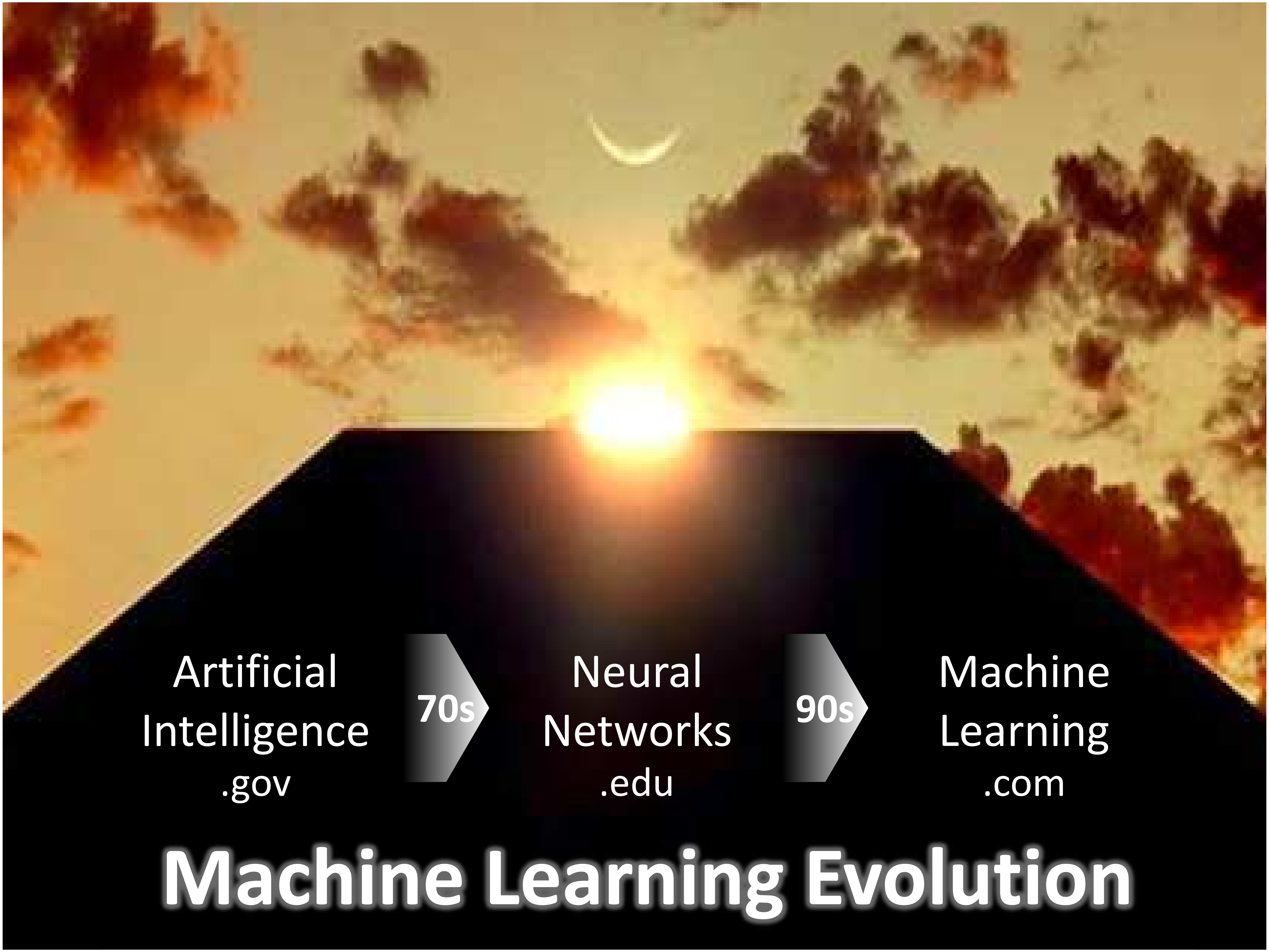
UnSupervised

No examples

- Density estimation
- Clustering
- Dimensionality reduction

Machine Learning Value Cost Map

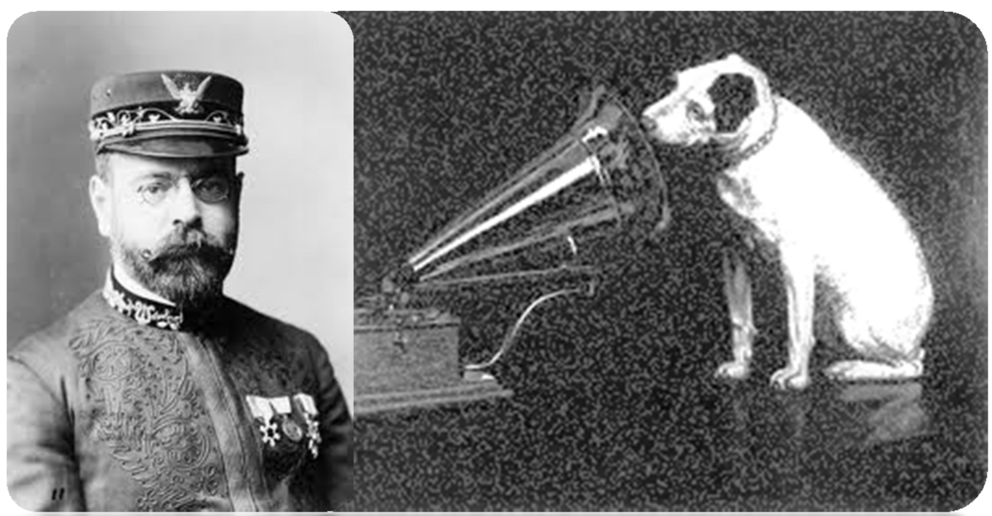



Artificial
Intelligence
.gov
70s
Neural
Networks
.edu
90s
Machine
Learning
.com
Machine Learning Evolution

A Non-Linear Evolution

…vocal chords will be eliminated by a process of evolution, as was the tail of man when he came from the ape.

– JP Sousa, 1906

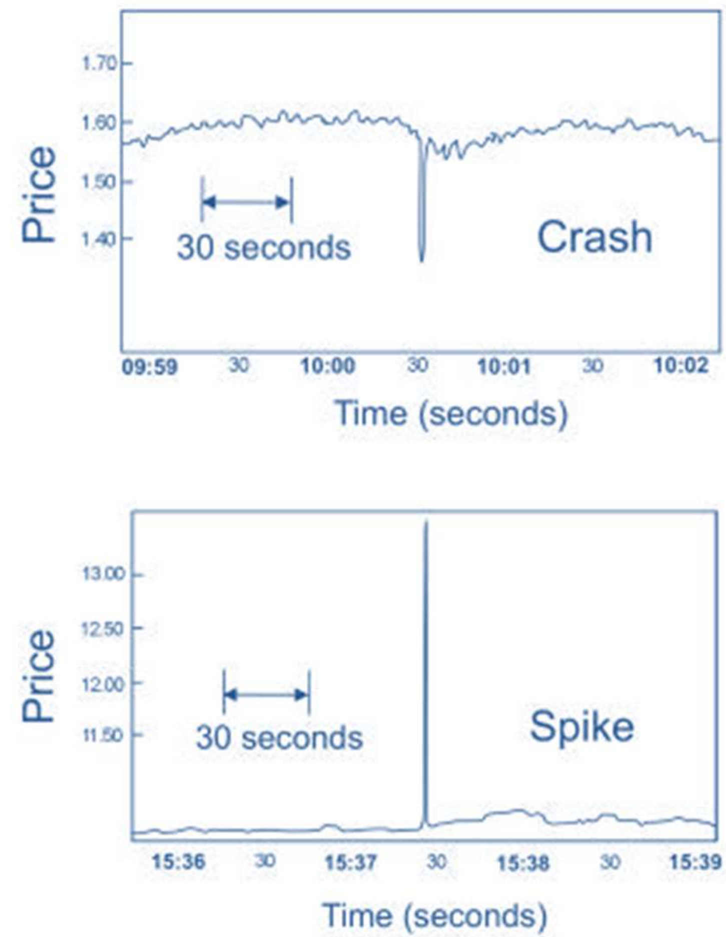


Human Flaws Amplified by Tech

RT: #Human #Flaws Amplified by #Tech

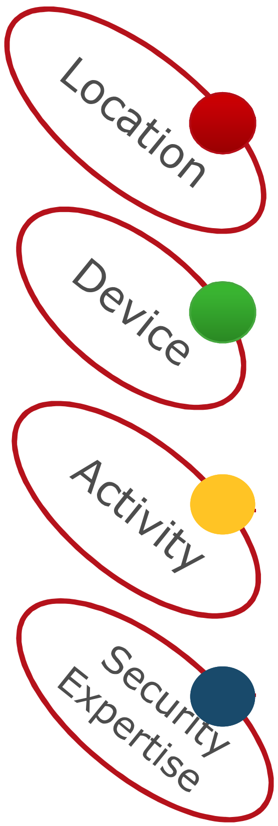
Making *Human* Mistakes Faster

"…mobs of ultrafast robots, which trade on the global markets and operate at speeds beyond human capability, thus overwhelming the system…"

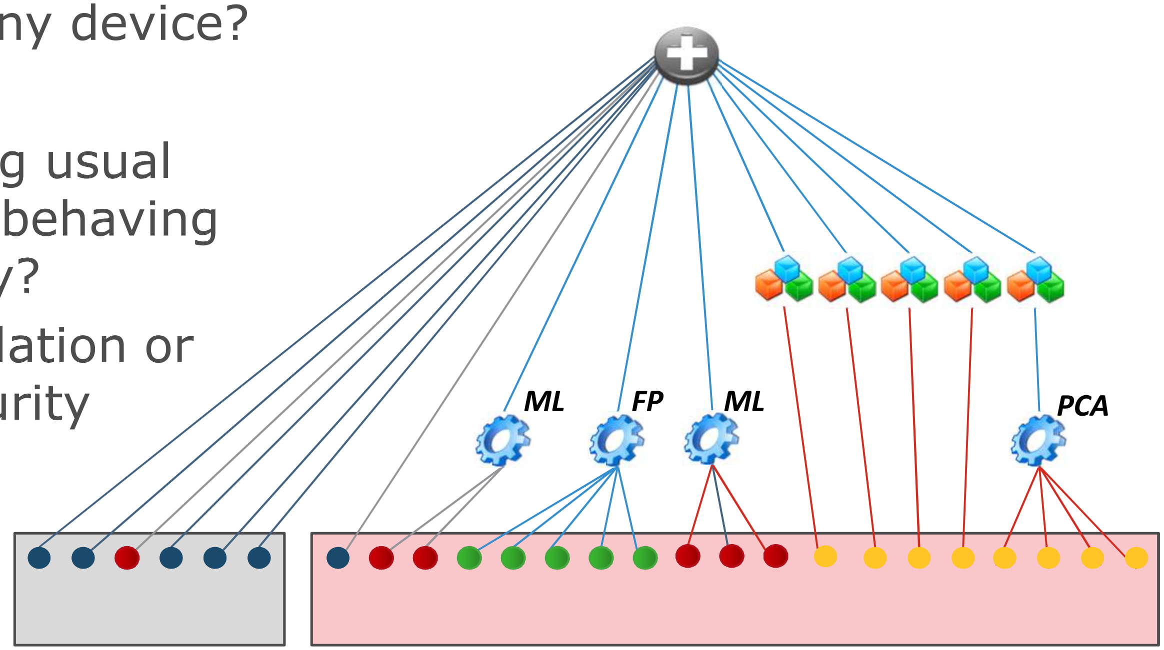



01
02
05
06
TBS 04-15
H I B
HAL 9000
03
04
07
08
RDT 90-00
GDE
AMT 11-23
DMC
HAL 9000
I’m sorry Dave, I’m
afraid I can’t do that

Basic Entities Model



- User coming from risky geography?
- User on usual device, or company device?
- User doing usual things or behaving differently?
- Policy violation or risky security posture?



Example: Location Estimation

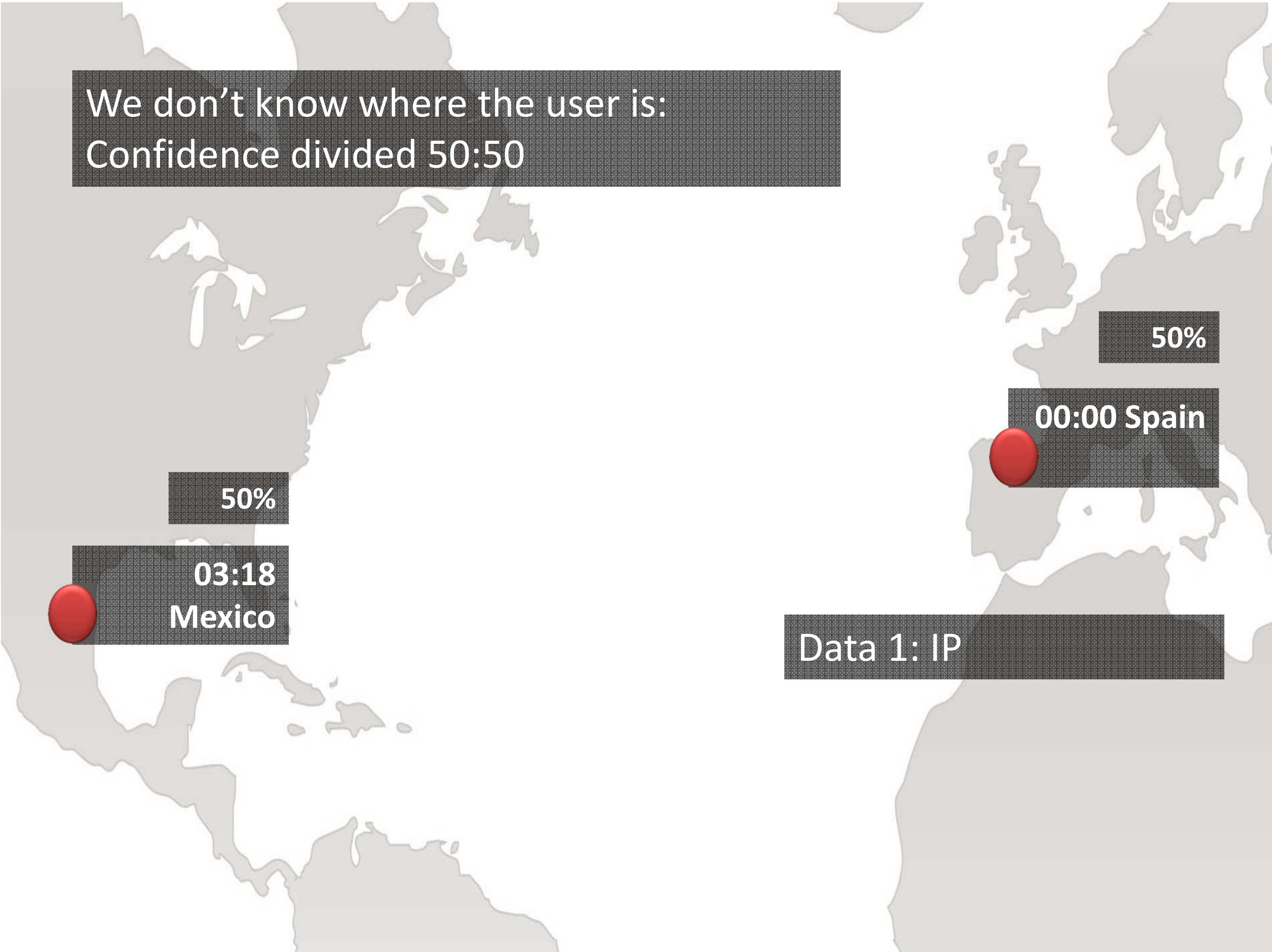
Multi-Source Analysis

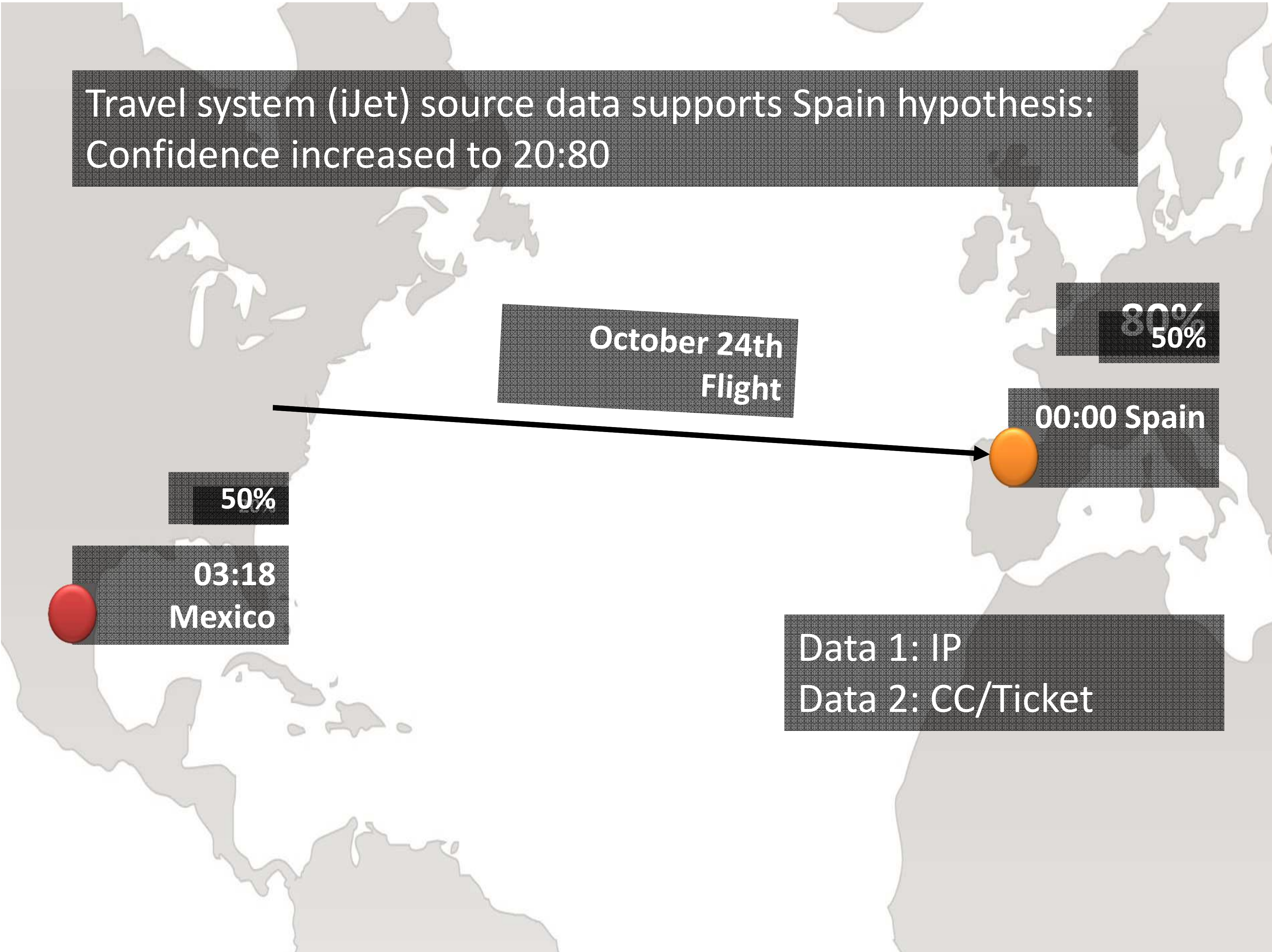
Input Values

1. High-risk Locations (Location-based Policies)
2. Anomalous Locations (Geographic-based Behavior)
3. Groundspeed Violations (Logic and Physics)

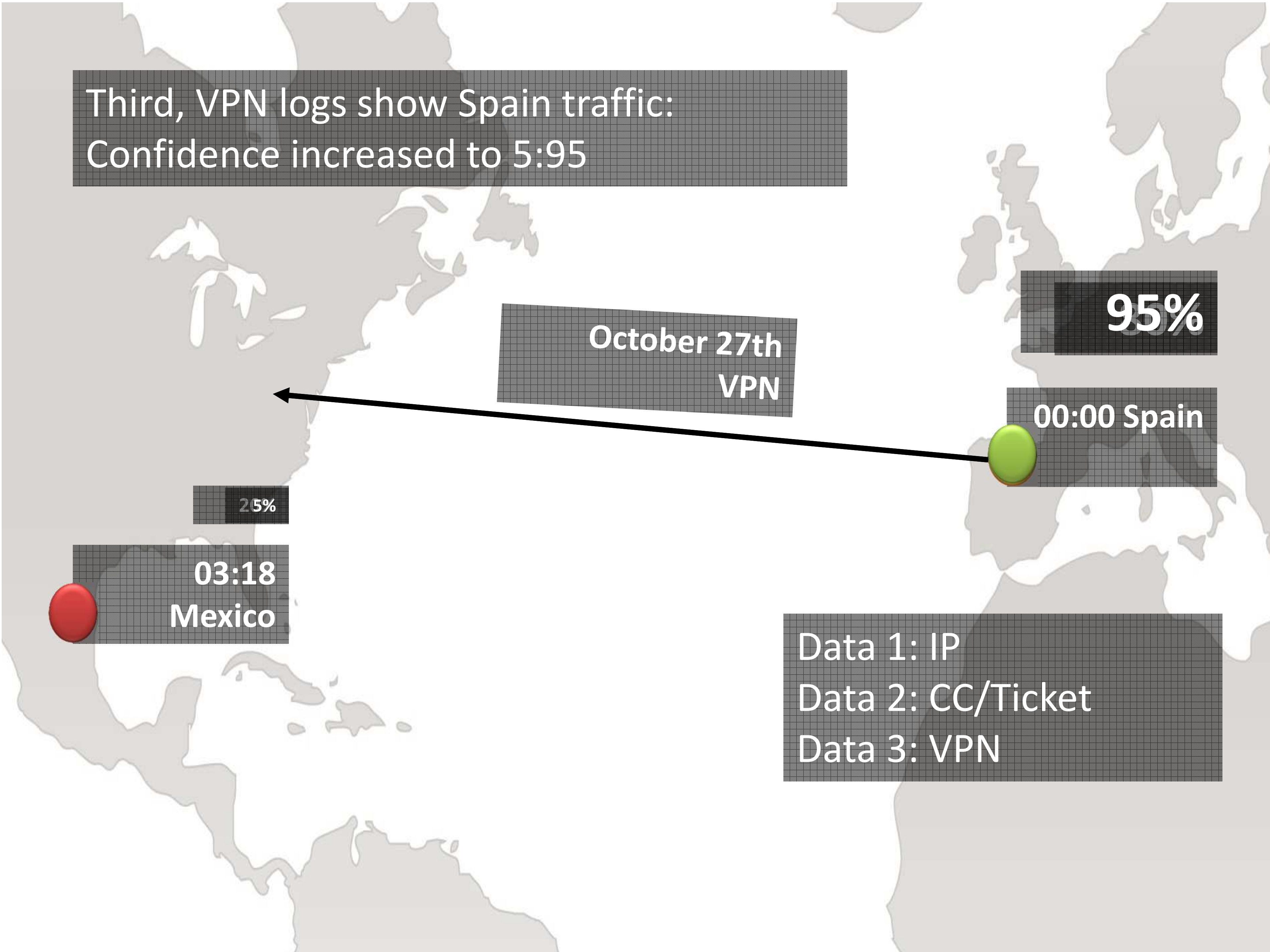
Method

1. Merge Diverse Data (VPN, Apps, Travel, HR, etc.)
2. Report Accuracy Estimation for Each Location
3. Combine Reports for Confidence on Estimations


We don't know where the user is:
Confidence divided 50:50
50%
00:00 Spain
50%
03:18
Mexico
Data 1: IP



Travel system (iJet) source data supports Spain hypothesis:
Confidence increased to 20:80
October 24th
Flight
80%
50%
00:00 Spain
50%
03:18
Mexico
Data 1: IP
Data 2: CC/Ticket


Third, VPN logs show Spain traffic:
Confidence increased to 5:95
95%
October 27th
VPN
00:00 Spain
5%
03:18
Mexico
Data 1: IP
Data 2: CC/Ticket
Data 3: VPN

THREATS TO MACHINE LEARNING

What Went Wrong With HAL?

Is It Just a Game?

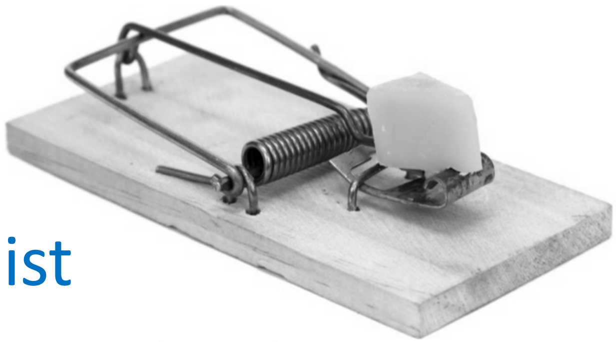
- Simple
  - Two Opponents
  - Rules of Engagement
- Complex
  - Unlimited Opponents
  - Guerrilla or Ill-defined Rules

Characteristics of Active Adversaries and Attack Models

- Delta in Current and Future Data
  - SPAM
  - Credit Card Fraud
- More than Random
- Unknown Change
- Targeted or Widespread

Trust in, and value from, information systems
San Francisco Chapter

Active Adversary Assumptions



- Attack Balance
  - Can Modify Attack to Evade Blacklist
  - Cannot Modify User Data to Change Whitelist (root)
- Results Balance
  - Can Mimic Whitelist to Evade Detection
  - Cannot Modify Amount to Change Whitelist (banker)

Machine Learning Threat Modeling

1. Outliers
2. Missing Values
3. Class Imbalance
4. High Dimension Inefficacy

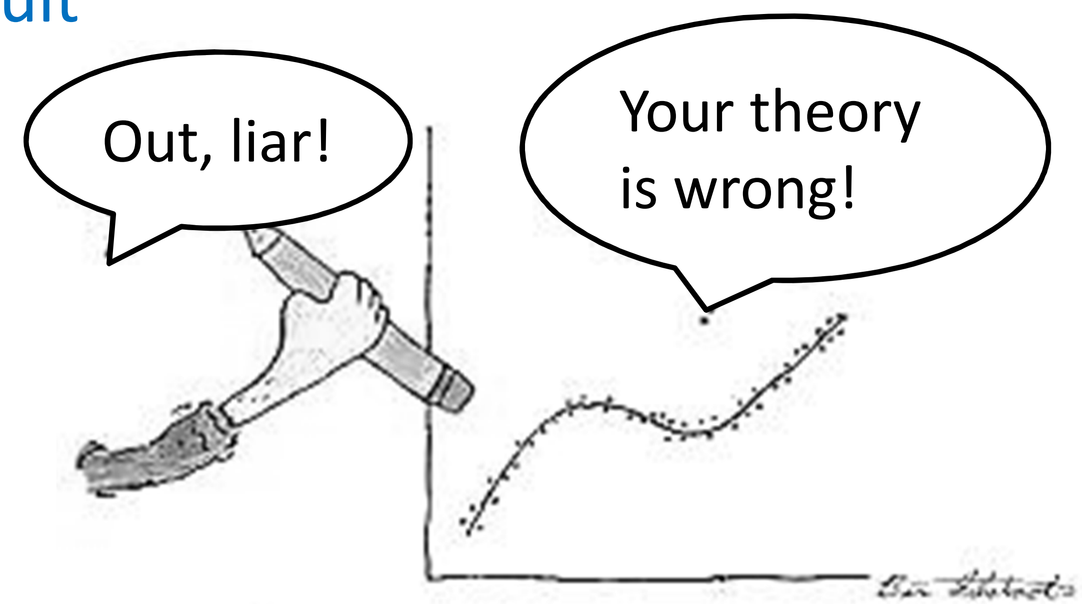
---

5. Non-Vector Data
6. Inaccurate Class Probability Estimates
7. Extension Lacking - iid to Dependent Data

https://www.brighttalk.com/webcast/9495/72899

Outliers

- "Needles in Haystack"
- High Value Discovery (High Cost if Not-found)
- Examples
  - Manufacturing Fault
  - Online Fraud
  - Network Breach
  - Clinical Trial Error



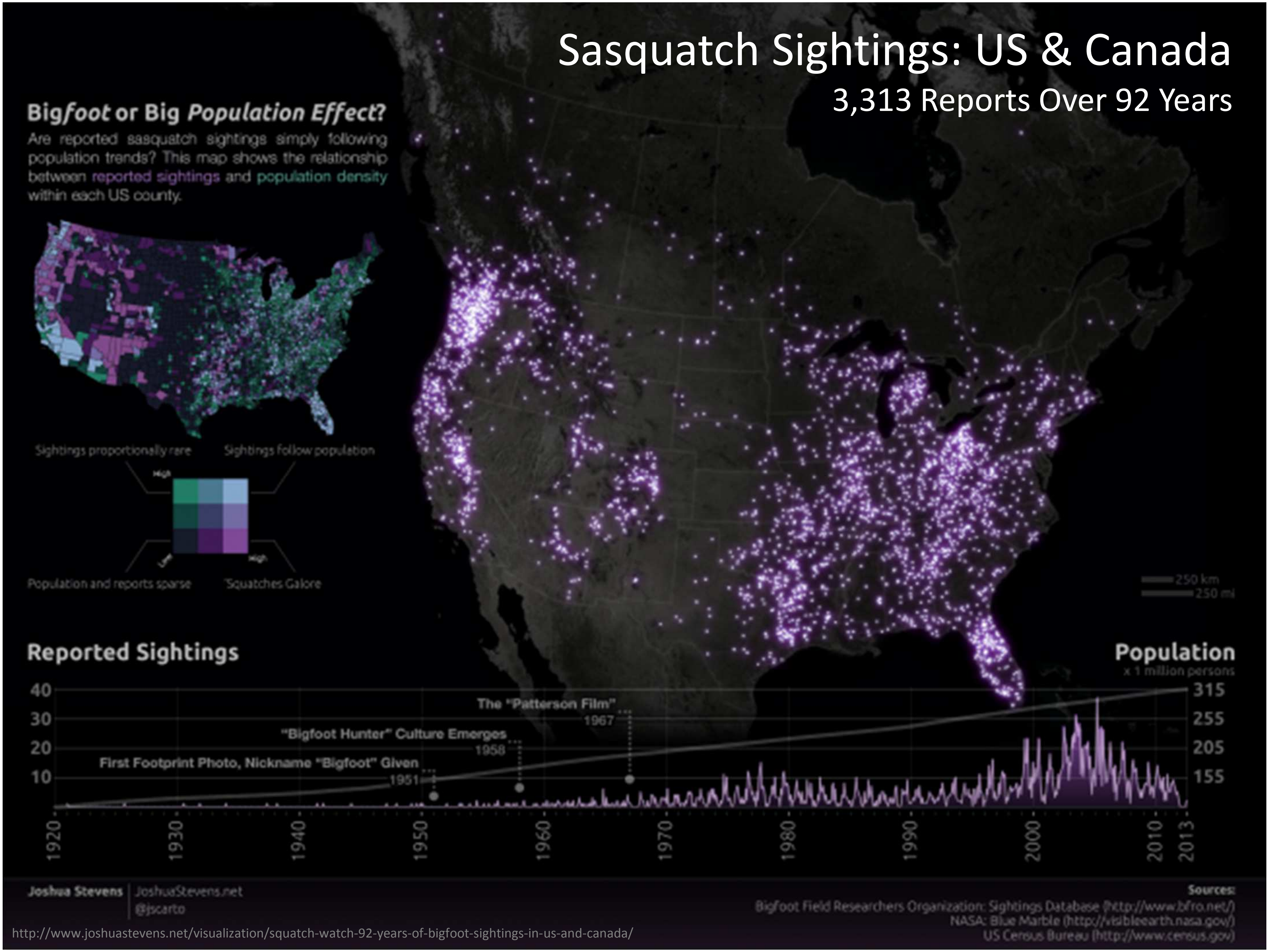
Missing Values

Acquisition / Observation Failures

- Interference (Scratch, Contamination)
- Broken Sensor (Photo Over Lens)
- Abandonment (Study Participant Quit)
- Complication (Overlooked Test Question)
- Flatlanders (Everyone Has Same Interests)

Class Imbalance

- Exception Hunt (Bigfoot)
  - Over-abundance Majority Examples (Not-Bigfoot)
  - Examples of Interest are Rare (Bigfoot)
- Mis-prediction Danger (Cry Wolf)
  - False-Positive Response
  - Reduction in Sensitivity

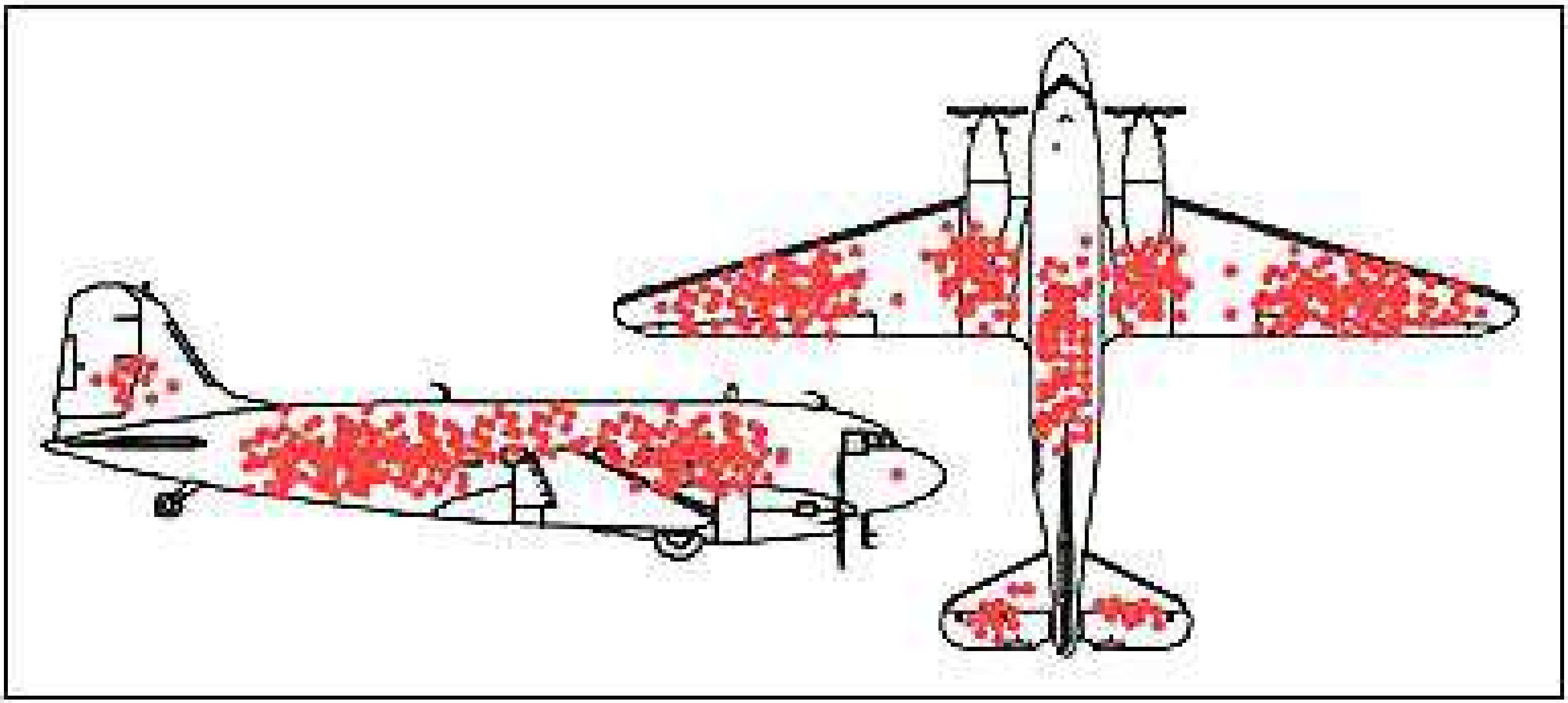


Sasquatch Sightings: US & Canada
3,313 Reports Over 92 Years
Bigfoot or Big Population Effect?
Are reported sasquatch sightings simply following population trends? This map shows the relationship between reported sightings and population density within each US county.
Sightings proportionally rare
Sightings follow population
High
High
Population and reports sparse
'Squatches Galore
250 km
250 mi
Reported Sightings
Population
x 1 million persons
40
30
20
10
315
255
205
155
The "Patterson Film"
1967
"Bigfoot Hunter" Culture Emerges
1958
First Footprint Photo, Nickname "Bigfoot" Given
1951
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2013
Joshua Stevens
JoshuaStevens.net
@jscarto
Sources:
Bigfoot Field Researchers Organization: Sightings Database (http://www.bfro.net/)
NASA: Blue Marble (http://visibleearth.nasa.gov/)
US Census Bureau (http://www.census.gov)
http://www.joshuastevens.net/visualization/squatch-watch-92-years-of-bigfoot-sightings-in-us-and-canada/

Class Imbalance

"…if a plane makes it back safely…bullet holes in the wings aren't very dangerous…"

- Abraham Wald, Mathematician



Credit: Cameron Moll

http://www.motherjones.com/kevin-drum/2010/09/counterintuitive-world

High-Dimension Inefficacy

- Hundreds or More Dimensions (unlike 3D)
- Predictive Power Inverse to Dimensionality
  - Training Data Sample and Value Size
  - Sparse and Dissimilar Data
  - Expensive to Organize and Search

Example: Poison An Anti-SPAM Engine

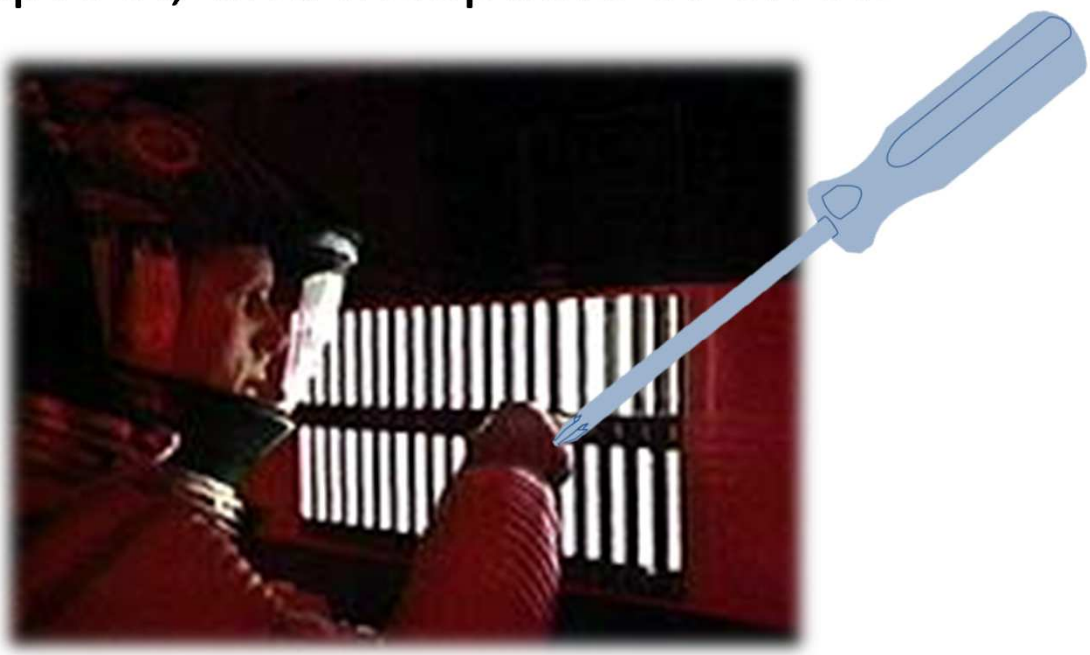
- Inject Specially Crafted Training Data
- Assumptions
  - Engine Still in Learning Mode
  - Attacker Can Replicate Original Training Setup
    1. Copy Algorithm and Data
    2. Steal Original Data
    3. Approximate Data

http://arxiv.org/abs/1206.6389v1

DETECTION AND STOPPING ATTACKS

Admitting We Have a Problem…

"No HAL 9000 series computer **<u>has ever made</u>** a mistake or distorted information. We are all, by any practical definition of the word, foolproof, and incapable of error."

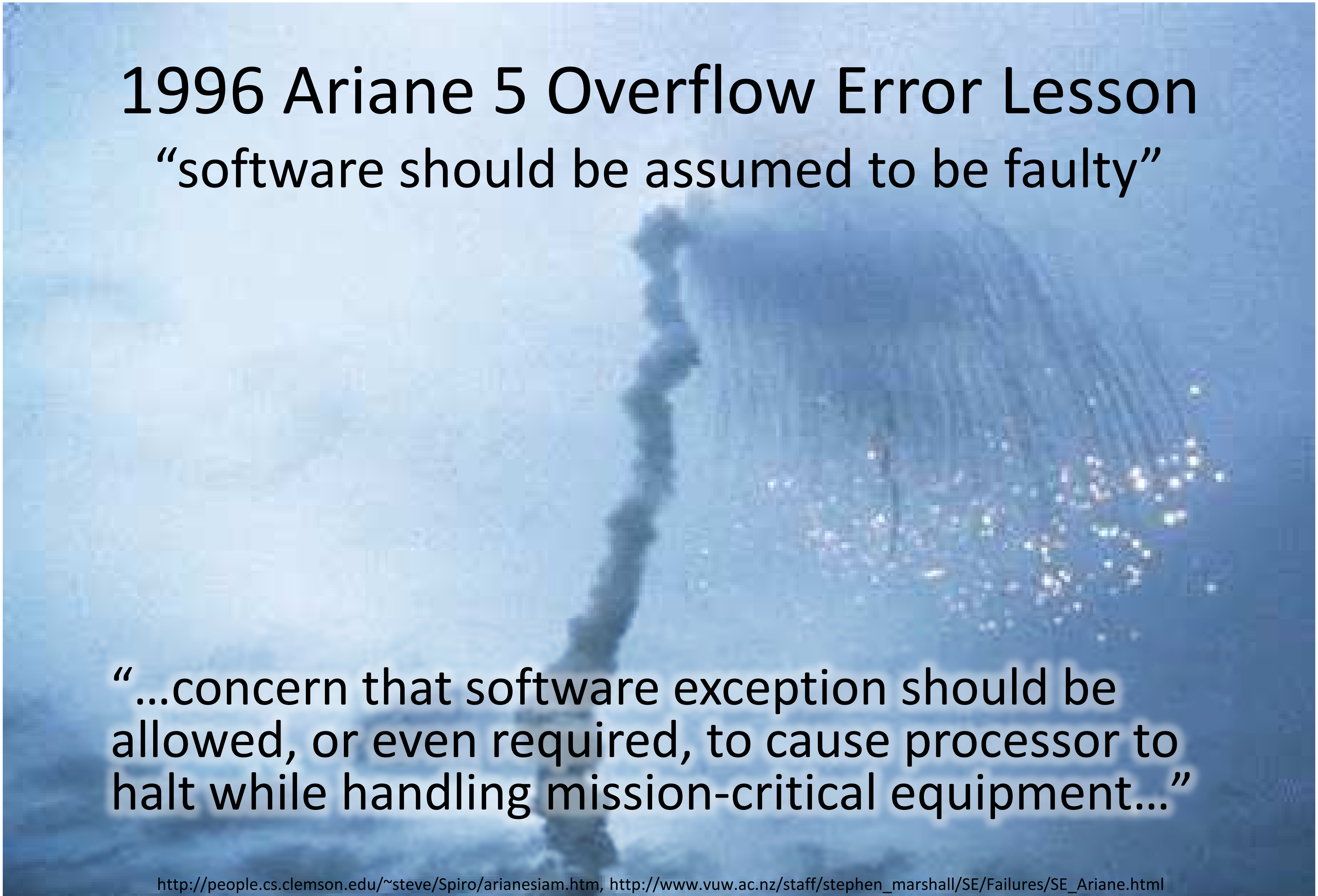


Am I Healthy?

(or should I shutdown)

1996 Ariane 5 Overflow Error Lesson

“software should be assumed to be faulty”



“…concern that software exception should be allowed, or even required, to cause processor to halt while handling mission-critical equipment…”

http://people.cs.clemson.edu/~steve/Spiro/arianesiam.htm, http://www.vuw.ac.nz/staff/stephen_marshall/SE/Failures/SE_Ariane.html

Induction Fallacy and Probability

- Control Priority

  Severity/Likelihood

- Risk Priority

  Threat

The wise proportion belief to evidence.

ML Resilience Planning

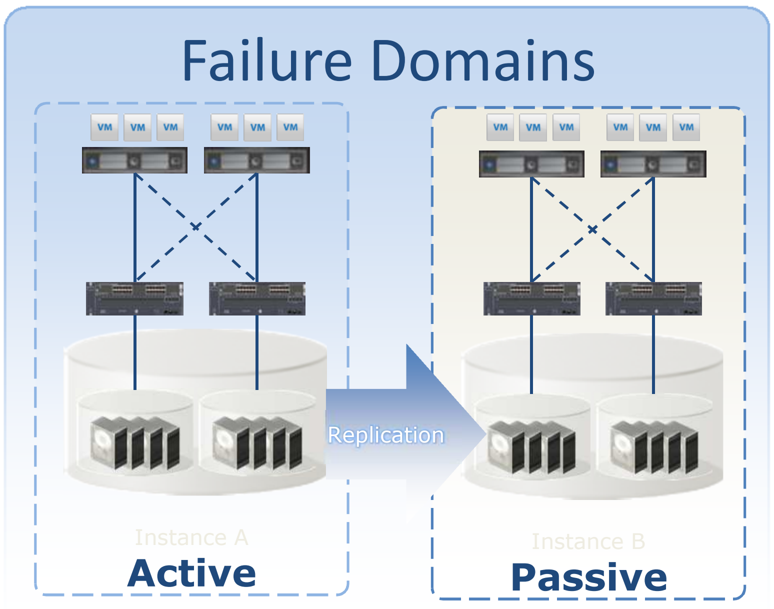
- Control Priority – Severity/Likelihood
  - Targeted
  - Widespread
- Risk Priority – Threat
  - Availability
  - Integrity Protections (Backup / Restore)

    Poisoned Training Data
  - Confidentiality

    Stolen Training Data / Algos for Production Poison

Availability

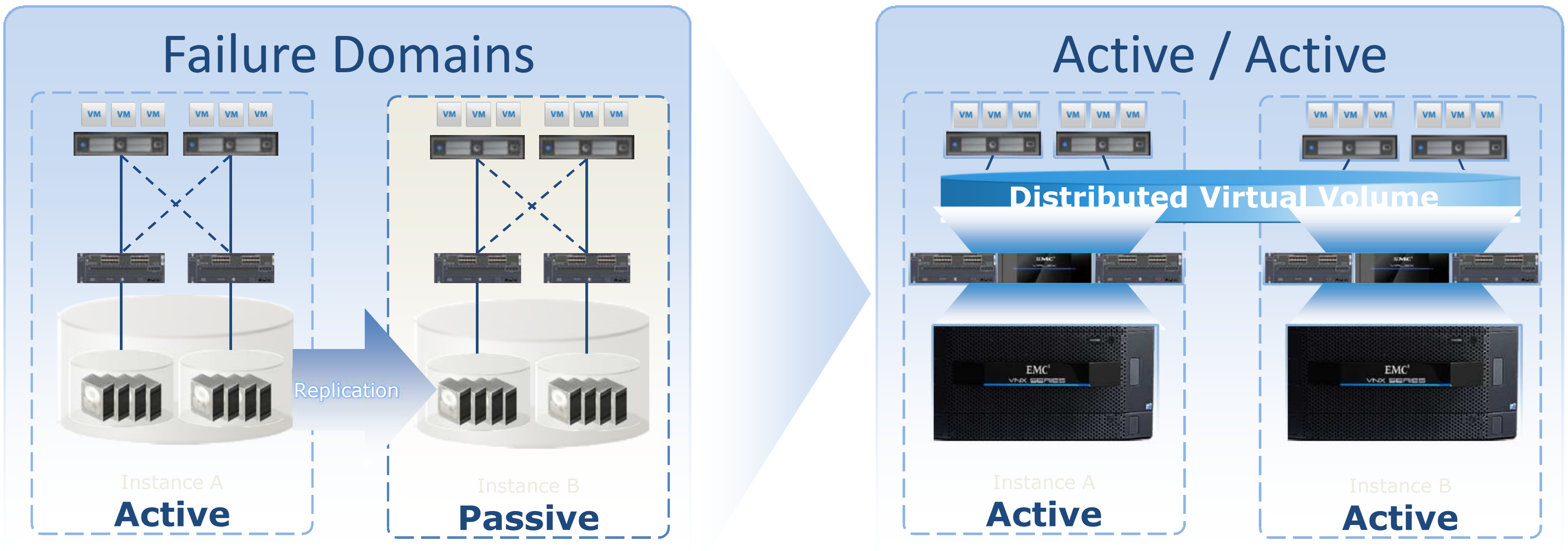
ML Acquisition Scale and Outages



- Application Disruption
  - Planned
  - Unplanned
- RTO: Minutes-to-Hours
- Failover and Fail-back Mgmt
- Passive, Idle Resources

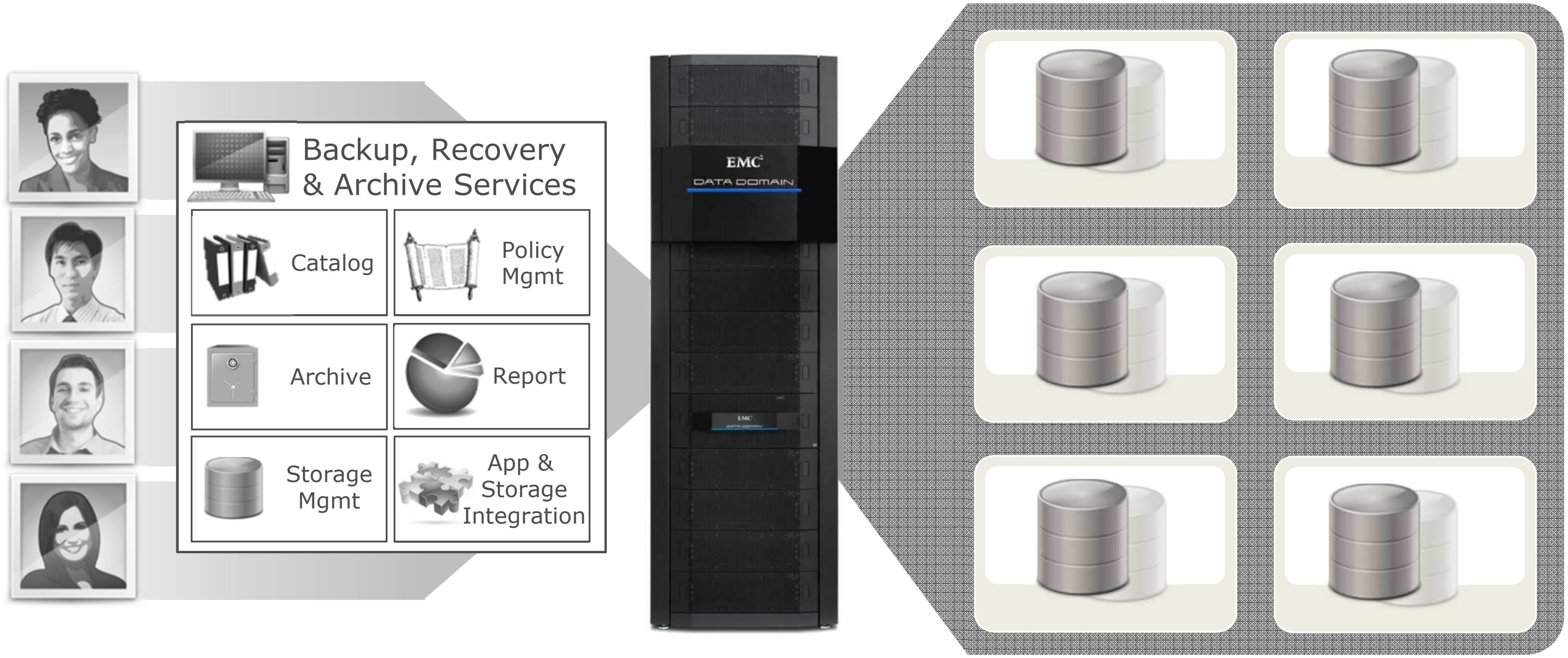
Availability

ML Infrastructure Zero RTO/RPO



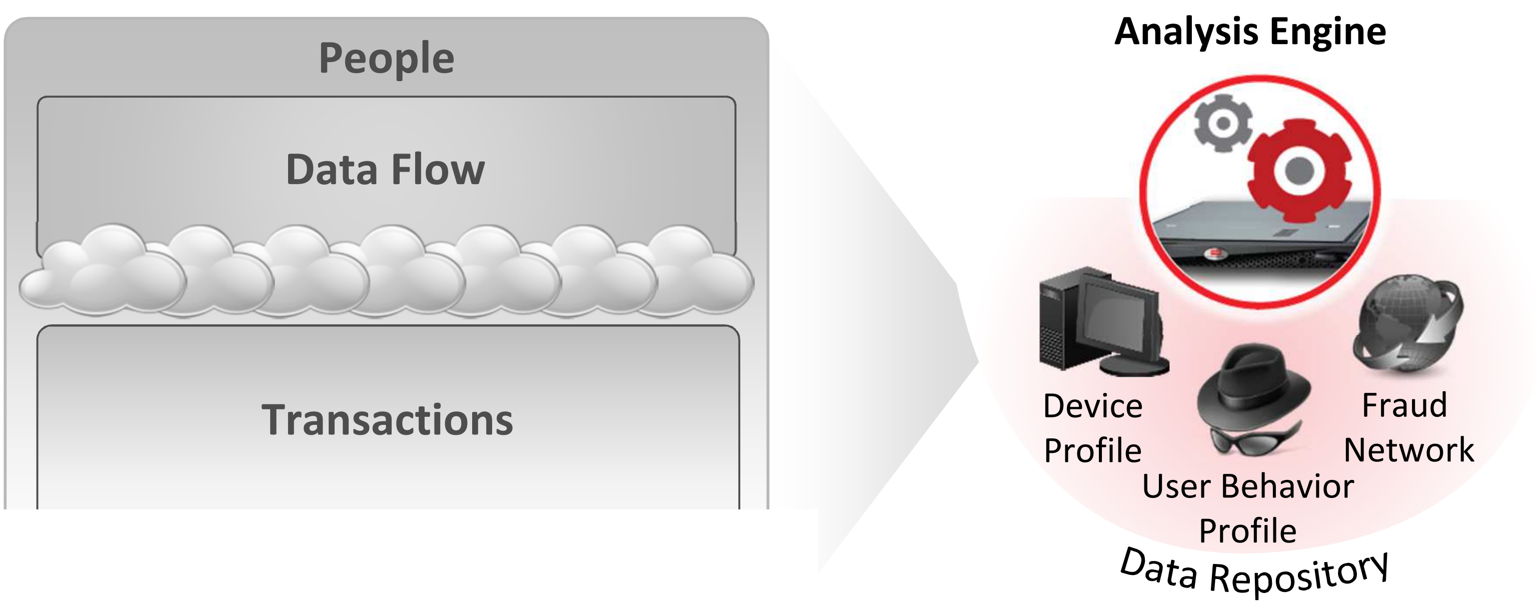
Integrity Protections (Backup/Restore)

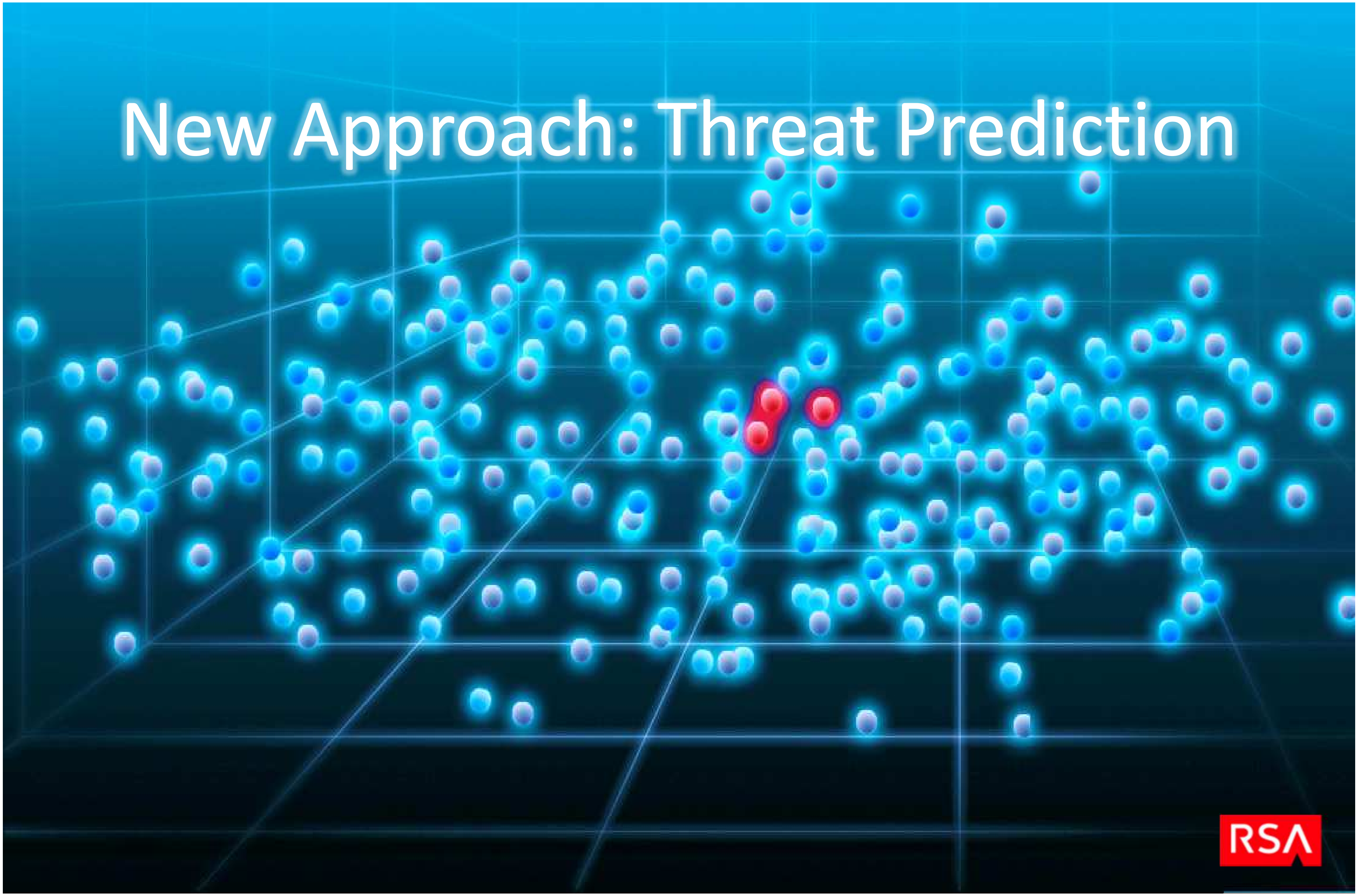
- Central Control and Monitor, Even Archives
- Restore ASAP Post-Breach



Confidentiality

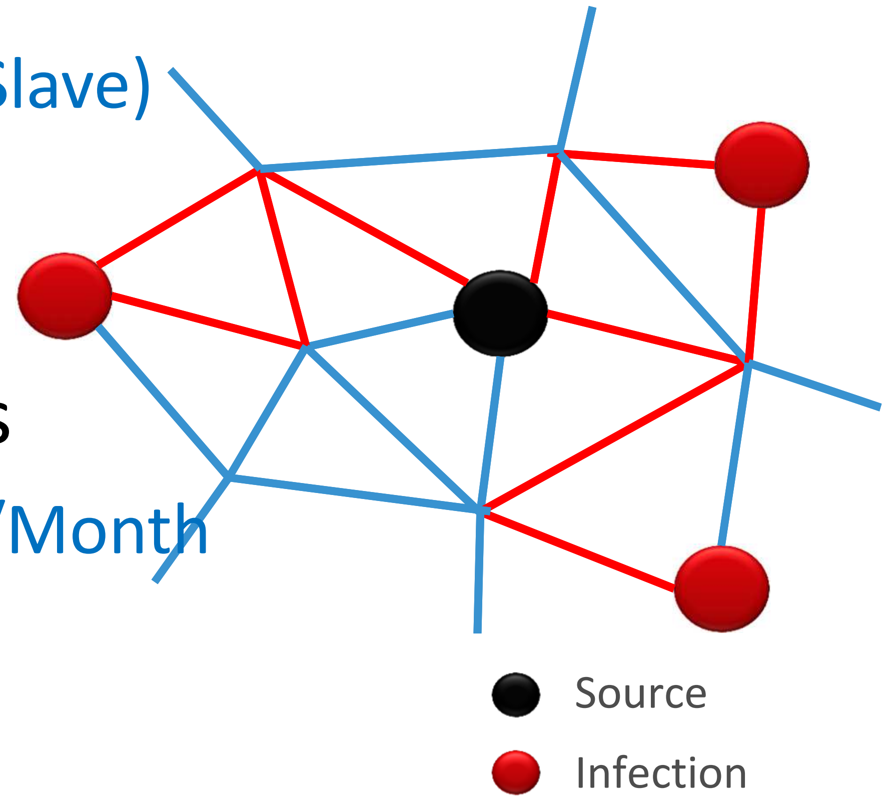
- Applying ML to Adversary Detection
- Monitoring Behavior




New Approach: Threat Prediction
RSA

Future Thought: ML Self-Preservation

- Awareness
  - Authority (Master-Slave)
  - Elective
  - Peer2Peer
- Sample Interactions
  - 50m+ Transactions/Month
  - Review 1/2000
  - Detect 92%



Conclusions

- Machines Make *Human* Mistakes…Faster
- ML Should be Assumed Faulty
  - Priority by Risk (Threat)
  - Priority by Control (Severity/Likelihood)
- Defense is Multi-Layered, Environmental
  - Development (Learning)
  - Production (Hardened)
  - Hybrid (Fail-Safe Learning)

THANK YOU!

#HeavyD — Stopping Malicious Attacks Against Data Mining and Machine Learning

Davi Ottenheimer @daviottenheimer

Senior Director of Trust, EMC

In-Depth Seminars – D22