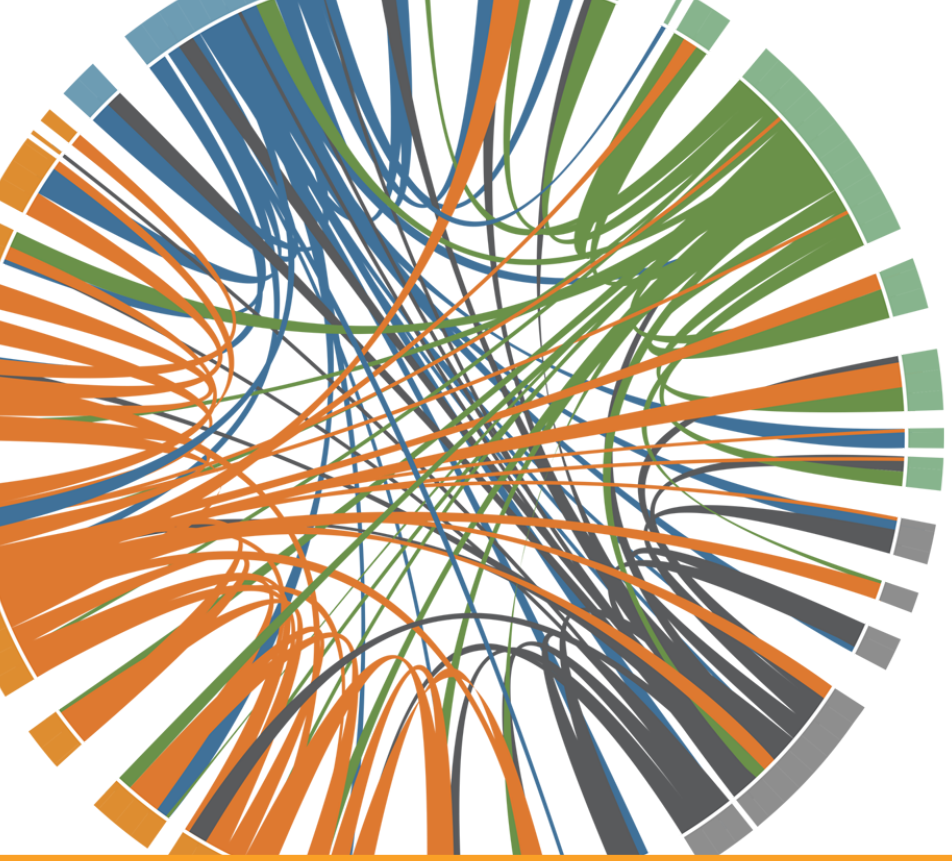


Using Big Data Analytics to Fight Identity Fraud

Dr. Stephen Coggeshall
Chief Analytics and Science Officer
ID Analytics/Lifelock

2nd Workshop of Mission-Critical Big Data Analytics
Prairie View, May 16,17

'id:analytics.
A Symantec Company

Outline

- All about identity fraud
  - modes
  - examples
- Using big data analytics to catch identity fraud
  - a look into our technology

ID Analytics Overview

ID Analytics applies advanced analytics to a unique sources of U.S. identity data, solving critical fraud, identity and credit challenges

- Founded in San Diego, CA in 2002
- Bought by Lifelock 2012, Symantec 2017
- Focused on empowering leading organizations to make better fraud, credit and identity risk decisions
- Over 450 leading U.S. enterprises rely on ID Analytics solutions every day…
  - 7 of the top 10 financial institutions
  - 4 of the top 5 wireless carriers
  - Government agencies: Social Security Administration and Veterans Affairs

What Is Identity Fraud

Identity Fraud is the act of misrepresenting which person you are

Why is it done?

- To improperly get products or services *(financial services, consumer products…)*
- To stay “unidentified” while behaving badly *(money laundering, terrorists…)*

Ways to Classify Identity Fraud

Classify by Industry/Target

- Financial identity fraud
  - *Account openings*
  - *Account takeover*
- Health care id fraud
- Tax id fraud
- Money laundering
- Terrorist activity
- …

Classify by Method Used

- Identity Theft
- Identity Manipulation
- Synthetic Identity

Three Types of Identity Fraud

	Who is the victim?	Nature of the misrepresentation
Identity Theft Core identity: ***Victim***	• Owner of misused identity • The company providing product/ service	• SSN, name and date of birth belong to the victim • Address, phone, and/or email belong to fraudster
Identity Manipulation Core identity: ***Fraudster***	• Fraudster • The company providing product/ service	• SSN, date of birth and/ or name vary slightly from the fraudster's own, correct information
Synthetic Identity Core identity: ***None***	• No direct consumer victim • The company providing product/ service	• SSN, name and date of birth are fabricated or chosen randomly

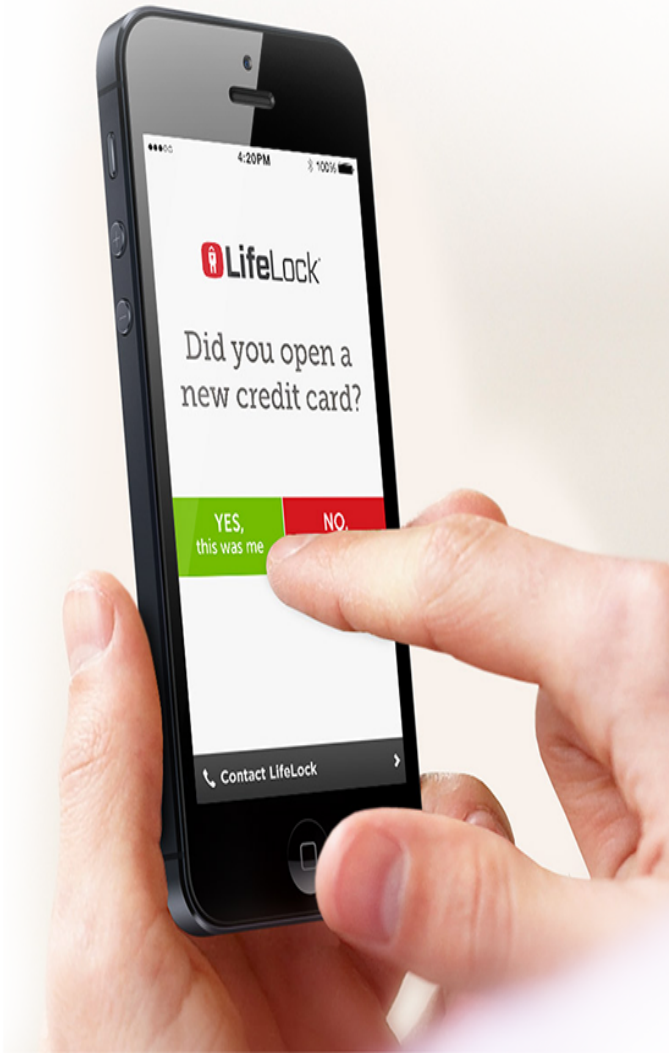
We All Know About **Identity Theft**

- Fraudster improperly uses another person's identity information
  - New account origination *(name, SSN, DOB)*
  - Account takeover *(name, account number, username/password)*
- Fraudster frequently uses his own contact information
  - *Phone, address, email*

Example of Severe **Identity Manipulator**

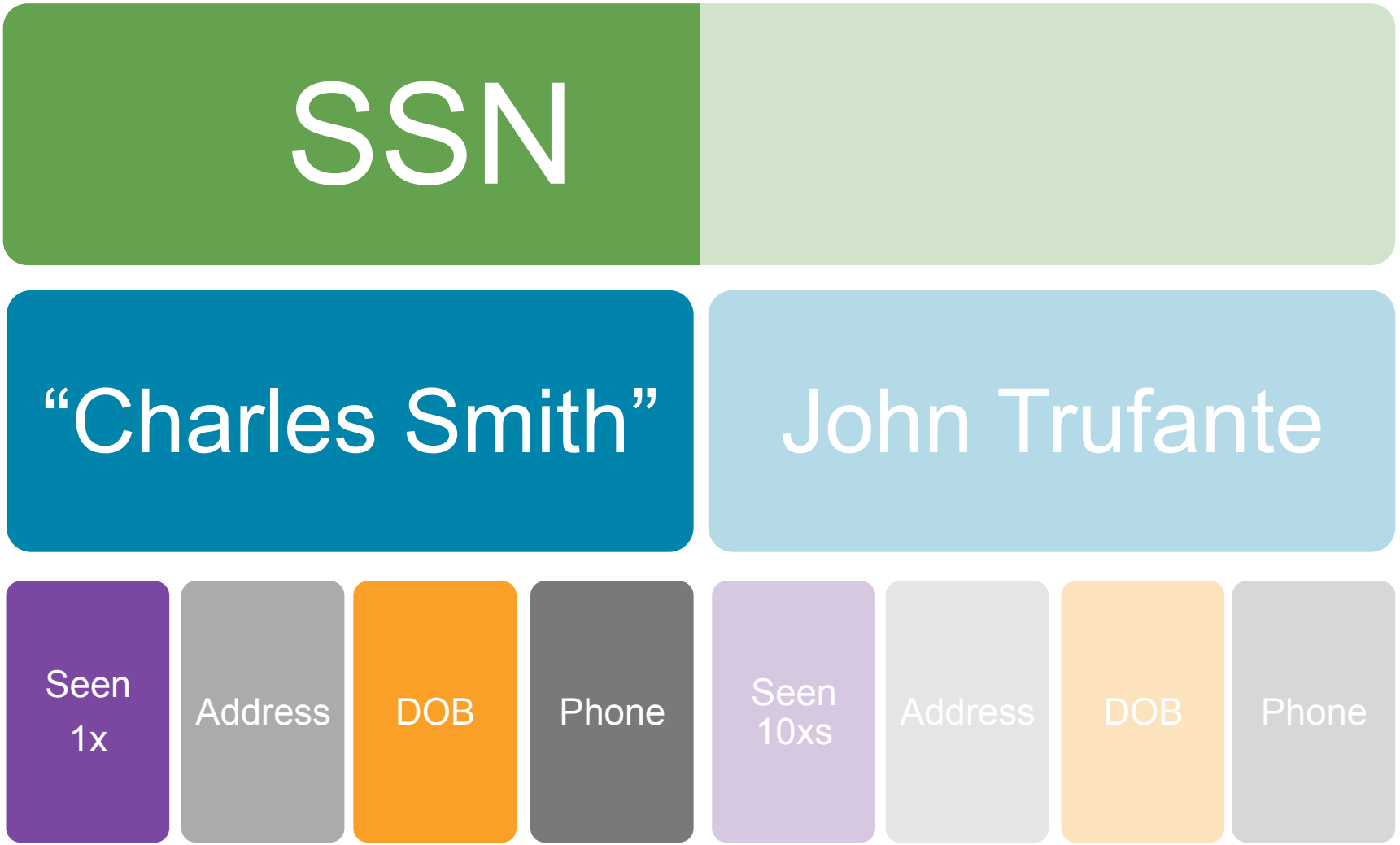
First Name	Last Name	SSNs	DOBs	Address
IRENE	ALMONE	580530044	1/7/1968	3310 ALGONQUIN AVE
LAQUINTA	CALHONE	586530044	12/21/1969	400 SCRUB OAK CT
LAQUITA	CALHOON	589489998	12/27/1969	4828 TERRACE TRL
LAQUITE	THOMPSON	589499935	1/7/1969	2600 PARK BLVD
LAQUTA	TOMSON	589539044	1/16/1969	4600 FAIR PARK BLVD
LEQUITA		590030040	1/17/1969	PO BOX 60011
QUITA		590490035	1/20/1969	2129 KINGSDALE DR
RENEE		590499937	1/27/1969	3211 MARYANN DR
RICHARD		590499938	1/27/1970	3229 KNOX ST
		590529641	1/27/1979	3424 FALCON DR
		590529941	1/27/1980	34321 FALCON DR
		590529994		3628 GLEN PARK CIR
		590530014		4709 LEONARD ST
		590530035		4719 LEONARD ST
		590530036		5161 DORMAN ST
		590530037		5163 DORMAN ST
		590530040		PO BOX 15840
		590530081		3008 GALEMEADOW DR
		590530244		1313 GLASGOW RD
		590538044		3052 BIRDSONG DR
		590929664		7063 MEADOWS DR
		590930043		7800 HILL DR TRLR 168
		590960044		4412 KEETER DR
		590980044		64 FOREST GLN

9 different first names | **5 different last names** | **24 different SSNs** | **11 different DOBs** | **Suspicious address variation**



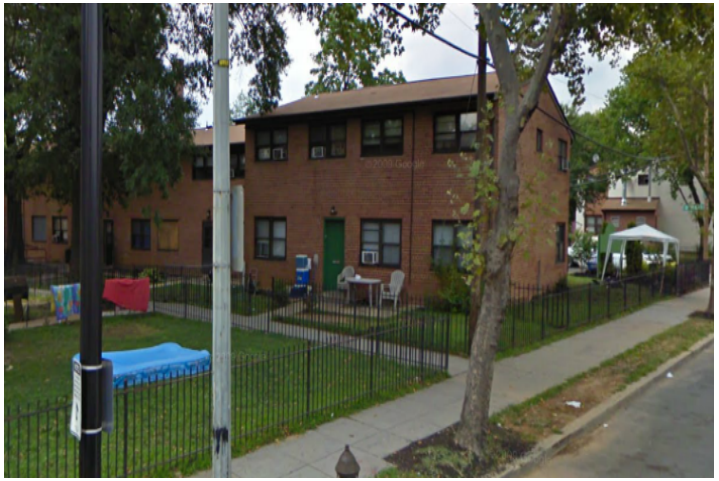
id:analytics

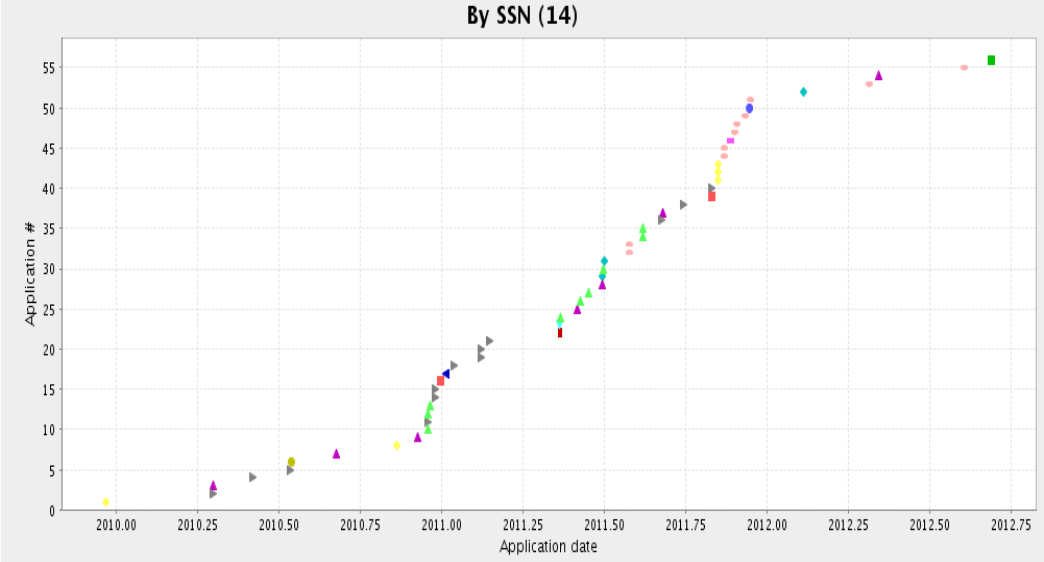
Description of **Synthetic Identity** Fraud



Example Identity Fraud Ring

Six People, Identity Manipulation and Identity Theft				
1	Gerald Smith	24 yrs	2 FNs	10 apps
2	Corona Jones	24 yrs	2 SSNs, 2 LNs	12 apps
3	Corona Jones	52 yrs	3 SSNs, 2 DOBs, 2 FNs	5 apps
4	Monique Jones	43 yrs	2 SSNs, 3 LNs	9 apps
5	Latasha Jones	21 yrs	3 SSNs, 2 DOBs	26 apps
6	Angel Jones	21 yrs	No identity manipulation	12 apps

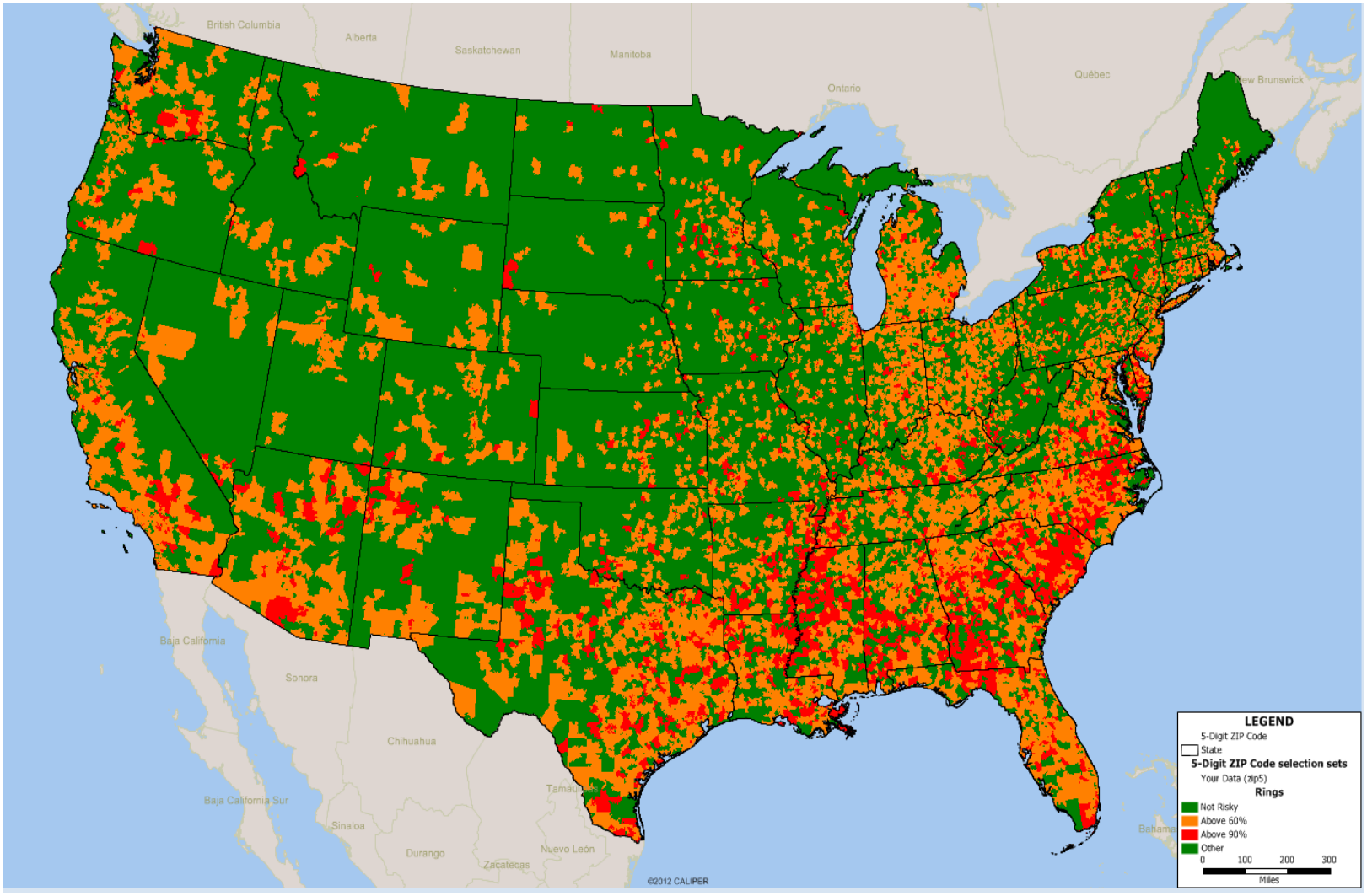




PII has been changed

- 85 phone & credit card applications from these 6 people from one address (close to Anacostia Park) in Washington, DC over 3 years. Sharing of SSNs, DOBs, names.
- An additional 5 ID theft victims; 2 victims are deceased (ID theft of the dead).

Identity Fraud Ring Locations



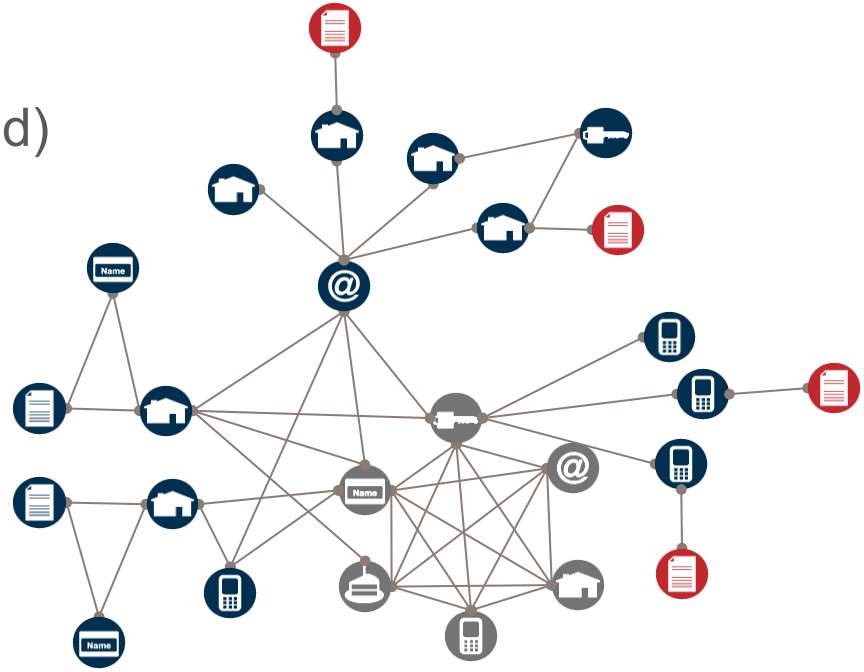
Data at ID Analytics

- ~2 billion applications for credit cards, mobile phones, retail credit, other loans…, more than 10 years
- U.S. white pages, (NAP*) ~100 million records monthly over ~10 years
- Header files (SNAPD*), ~200 million records monthly over ~10 years
- ~4 million labeled records of various frauds
- Account performance data, ~100 million/month
- Account changes, ~10 million/month
- Many other smaller files (SSA DMF, OFAC, census…)
- > trillion data elements
- Data is time stamped – can “roll back the clock”

*SNAPD – **S**SN, **N**ame, **A**ddress, **P**hone, **D**ate of birth

What's Behind The ID Score?

- Receive the application (SSN, Name, Address, Phone, DOB, email, device id)
- Build the PII-linked graph
- Translate this graph into numbers
- These features are the inputs to machine learning algorithms
- Calculate the score
- Return the score and reason codes



All this is done in ~200 ms.

Our products are real time delivered and near real time aware.

Machine Learning/Modeling Algorithms

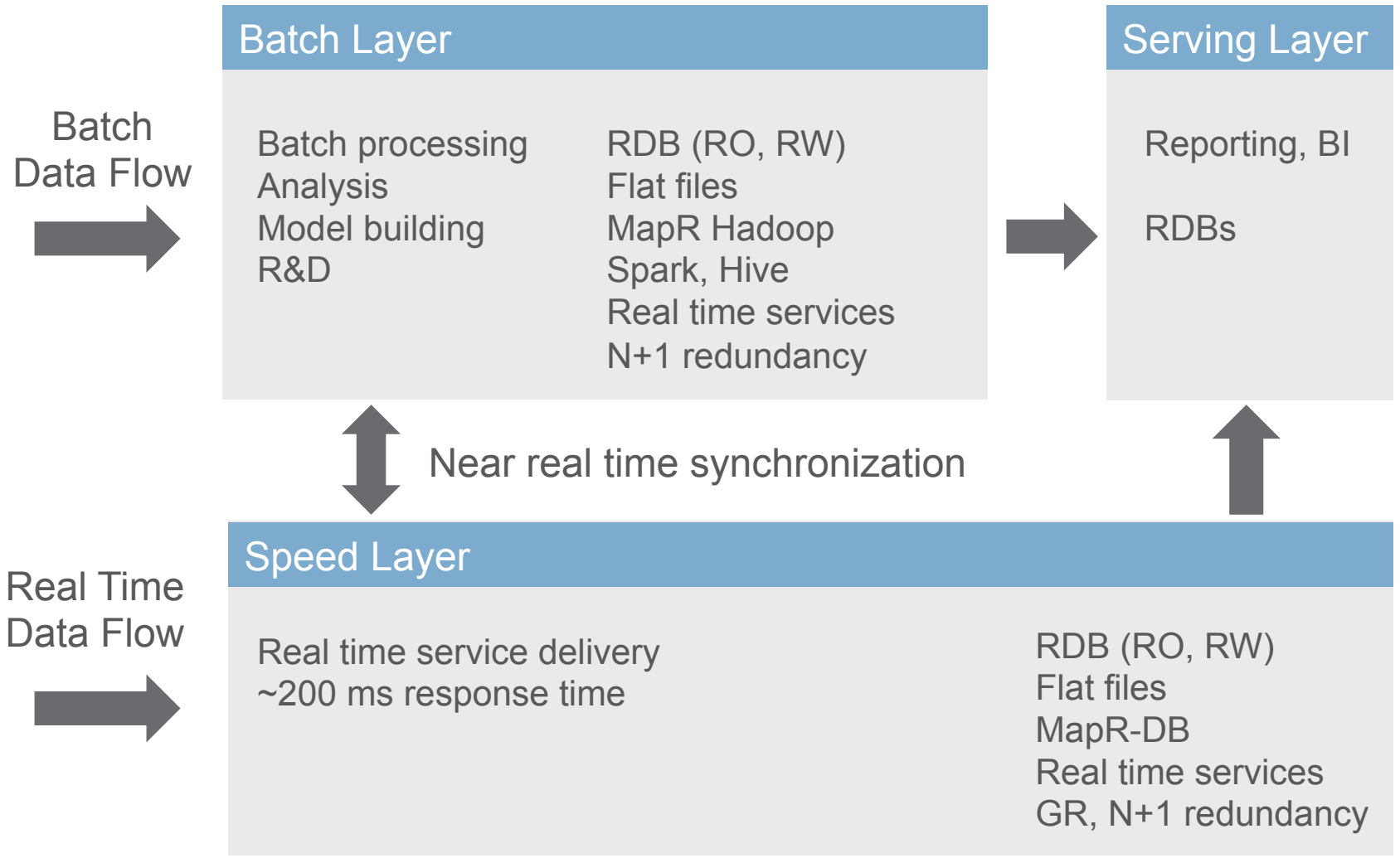
- Support Vector Machines (SVM)
- Traditional Neural Nets
- Boosted Trees
- Random Forests
- Convolutional/Deep Neural Nets
- K-Means Clustering
- Others – linear, logistic regressions, CART/CHAID, radial basis functions, KNN…

Services That Need Specialized Technology

- **Identity Risk Scores** – requires real time assembly of real time data for real time score delivery. Very fast R/W, ML calculation.
- **Identity Resolution** – look back through all data over time and people to resolve an identity when presented with fragmented information. Needs knowledge of complete PII history, complex fuzzy linking.
- **Find fraud rings** – examine multiple fuzzy linkings across billions of records. Can be batch.

These require very special data organization and systems

Modified Lambda Architecture



Real Time Production Scoring

Tasks:

- 24x7, low latency, moderate throughput scoring via system API calls
- Real time scoring, real time awareness
- Demanding database needs: multiple index random read/writes <50 ms from multiple Tb data stores
- Complex scoring model execution: typically boosted trees, ~1000 trees, ~300 variables for each score calculation
- Return 99% of score requests in under 200 ms
- Score ~1 million/day

Environment and technology stack:

- ~200 nodes, ~256 Gb memory/machine, ~1.3 Pb storage, n+1 redundancy
- Hot, warm geographic redundant data centers with continuous data replication
- Linux, java
- Continuous delivery software development pipeline (Maven, Ansible,…)
- Multiple instances of highly tuned MySQL with some sharding
- Springboot, Docker, Kubernetes
- Some products in AWS for faster product development and iteration

Batch Data Inflow and Production Data Deployment

Tasks:

- Receive, process, field many large batch files efficiently
- Many files have complex hierarchical structure
- Move, standardize, transform, clean, organize, encrypt
- Catalogue, governance
- Deploy across multiple databases and environments

Environment and technology stack:

- Oozie, Kafka, Moveit, Drools
- MapR Hadoop, Map Reduce
- Hive, Pig, Spark

Backend Analytics Infrastructure

Tasks:

- Build predictive models
- Perform “Retro tests”
- Ad hoc analyses
- Explore new candidate data sets
- R&D (algorithms, variables, data organizations…)

Environment and technology stack:

- ~35 node compute cluster, ~512 memory/machine
- Small GPU cluster
- ~500 Tb MapR hadoop
- Completely timestamped data system
- Condor, Hive, Spark
- Python, Scala, java
- Graphx, Weka, TensorFlow for prototyping
- Custom model platform for production models

R&D Project:
3D Identity Event Profiling

3D Identity Profiling

Better understand and categorize the nature of the risk

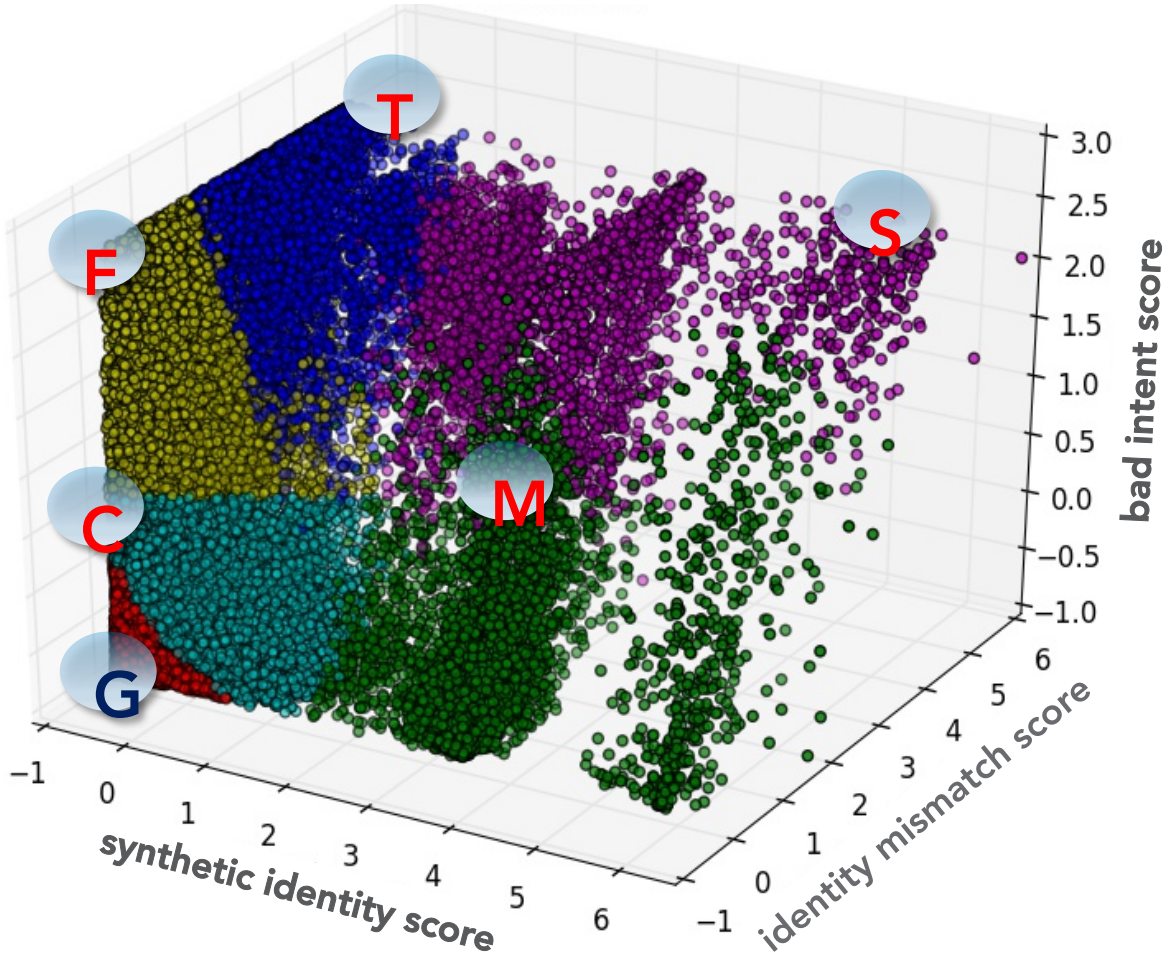
Fundamental Questions

1. **Is the asserted identity real?**
2. **Does the asserted identity belong to the applicant?**
3. **Does the applicant have a good intent?**

Three models that answer these 3 questions comprise the set of axes in order to locate each application as a point

1. Synthetic Identity Score
2. Identity Mismatch Score
3. Bad Intent Score

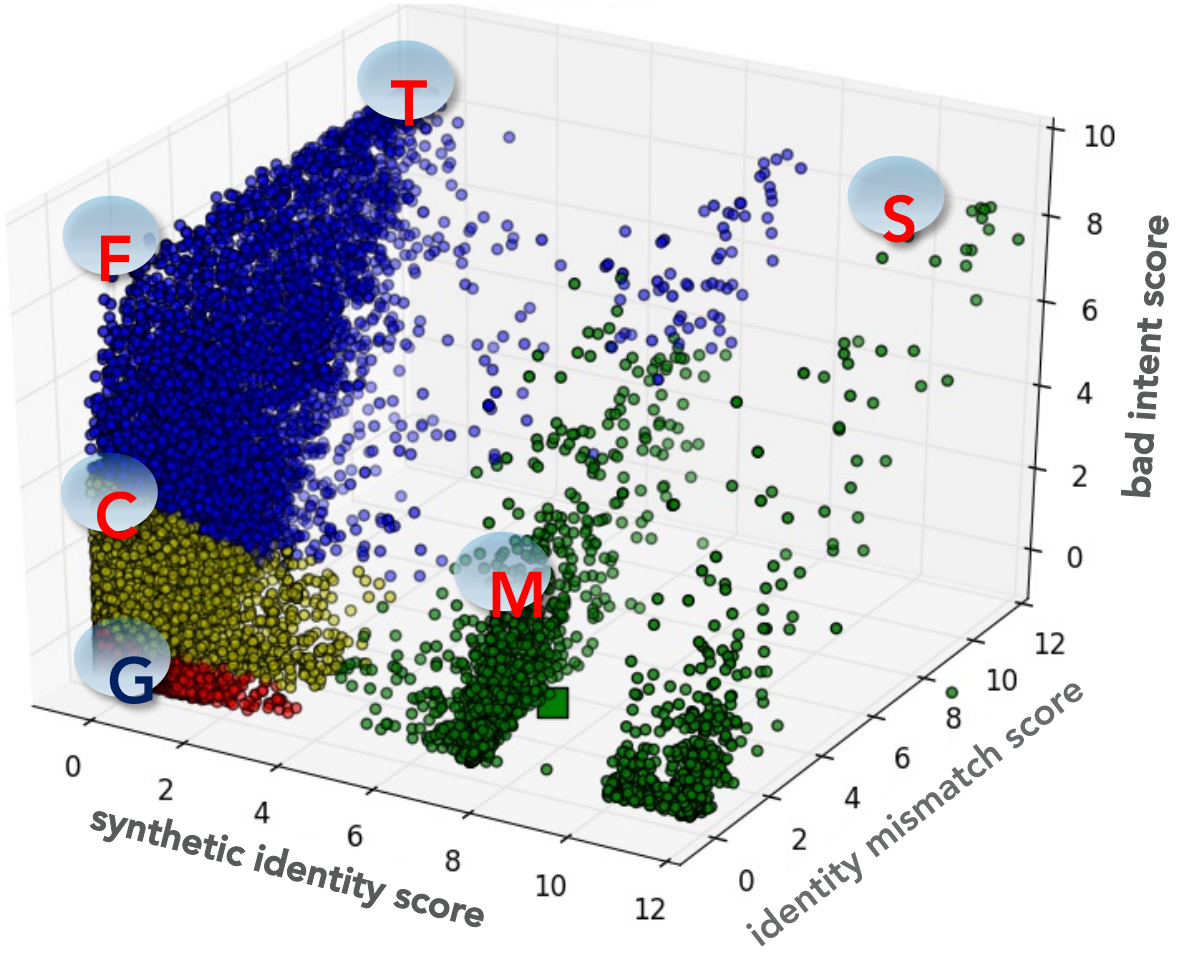
ID3D Shows the Risks Confronting a Bankcard Client



- (F)irst Party Fraud
- (C)redit Bad
- ID (T)heft
- (S)ynthetic ID fraud
- ID (M)anipulation

id:analytics

ID3D Shows the Risks Confronting an Alternative Lender



- (F)irst Party Fraud
- (C)redit Bad
- ID (T)heft
- (S)ynthetic ID fraud
- ID (M)anipulation

id:analytics

3D Identity Profiling

How this helps us

- Better understanding of the nature of the risk
  - a synthetic identity or manipulated identity could be created with *good intent*

- Analyze, identify, and inform each client …
  - What ***types of risk*** they deal with and should pay attention to
  - How the ***trend*** in the bad behaviors are changing

- Discover new modes of bad behavior

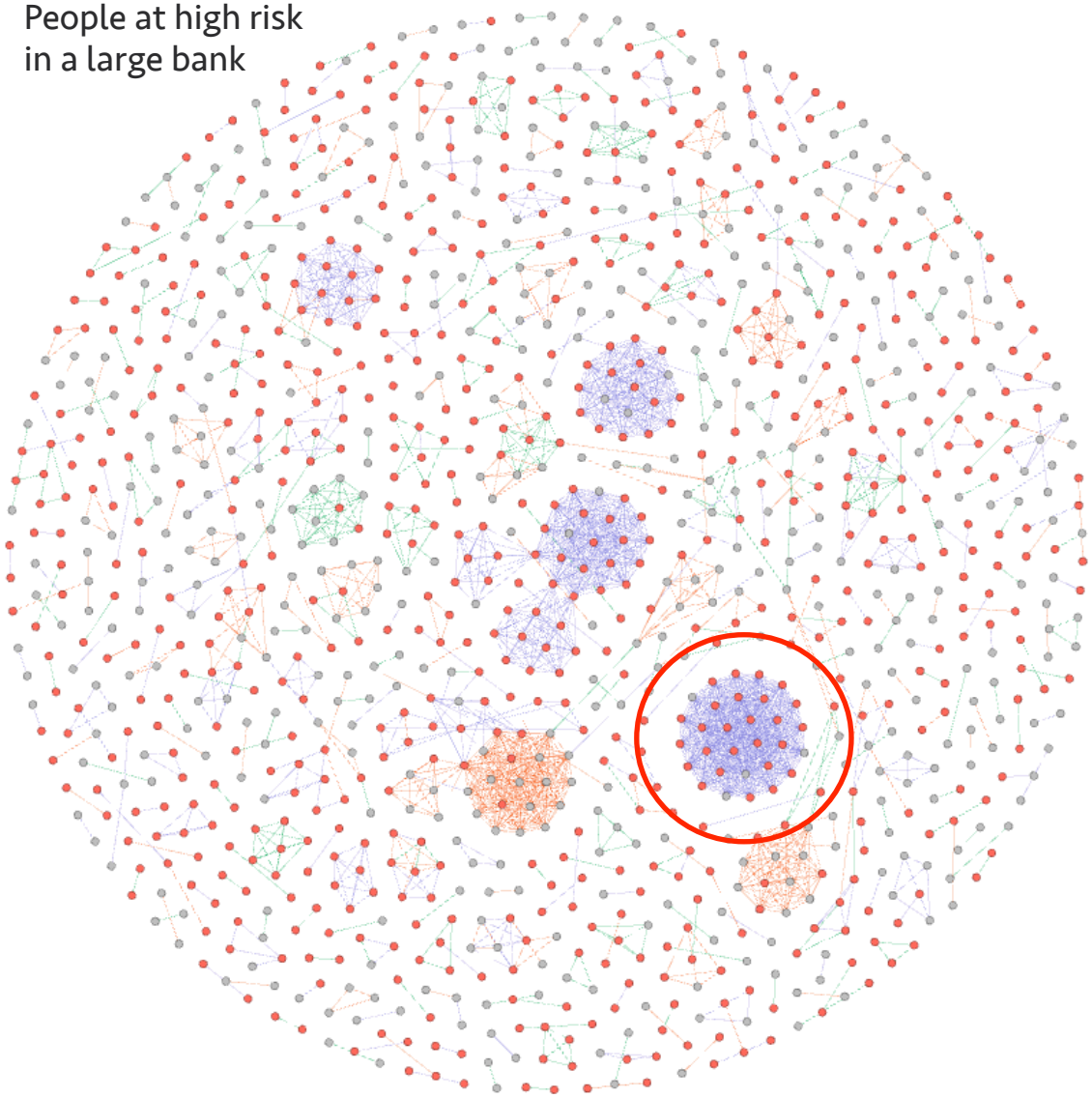
R&D Project:
Early Breach Detection

Analysis of our ID Network can discover data breaches

- *Data breach* — Identity information is released into the black market and exploited by fraudsters

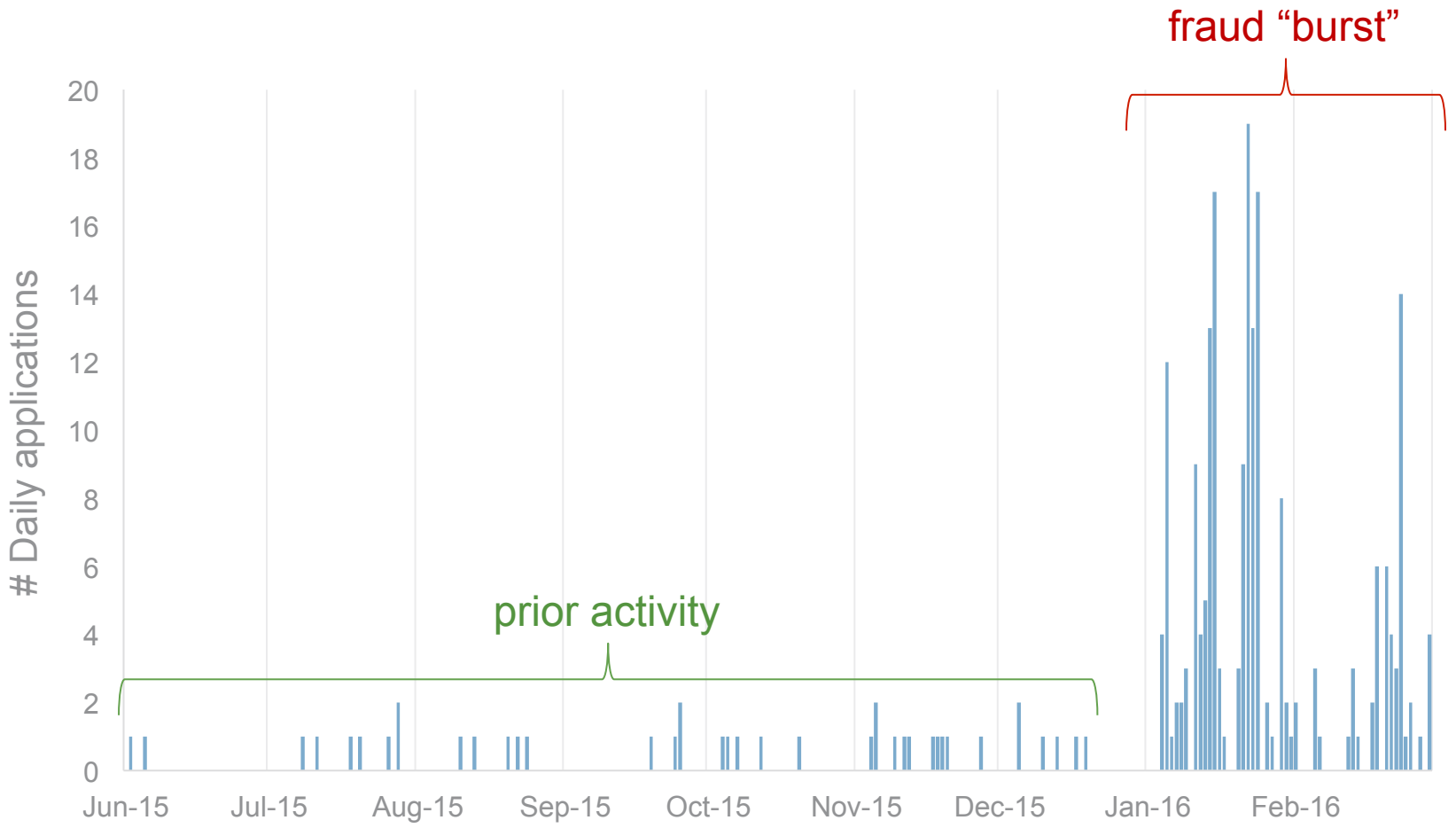
- *Early breach detection* — Pro-actively identify fraudulent activities from undisclosed breaches as they occur

People at high risk
in a large bank

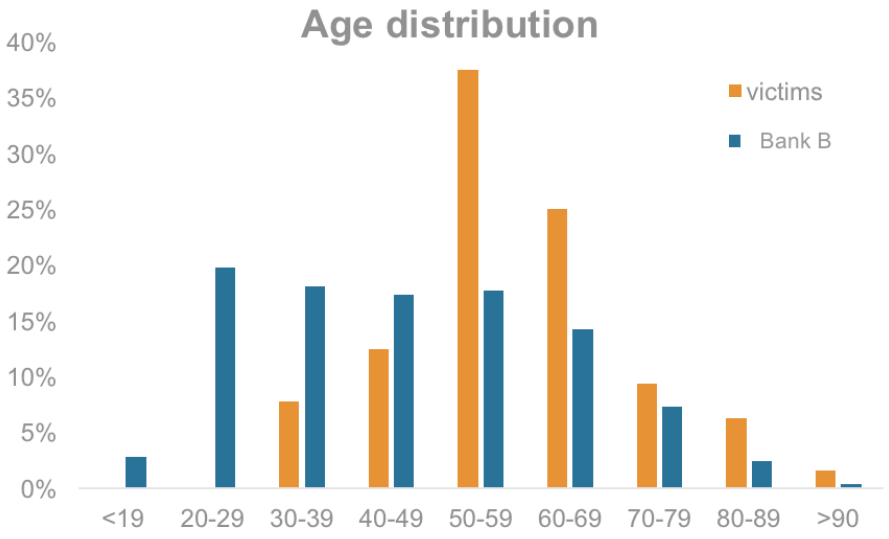


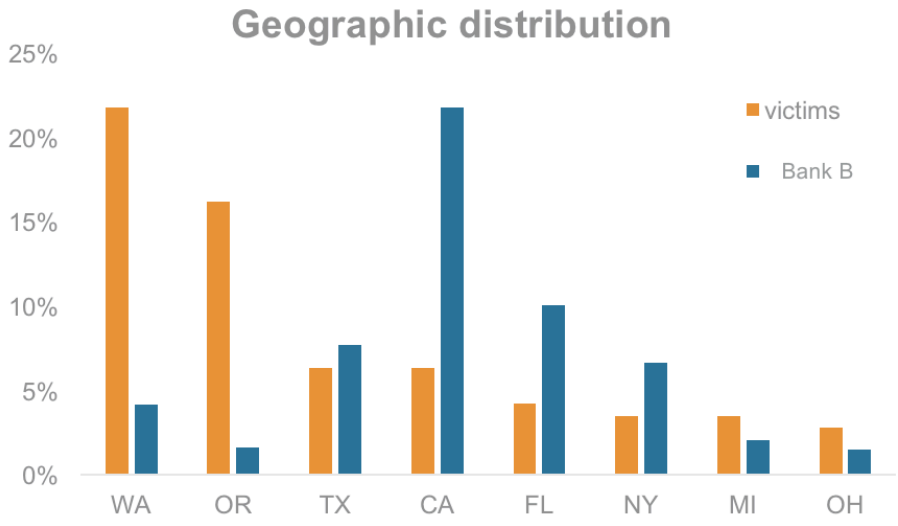
Each circle is one person with ID Score > 600 (red > 900)
Each connection is a shared (address, phone or email)

Specifically, we noted unusual activity on these SSNs at Bank B . . .



. . .Identifying the potential breach event





Personal Info On Hundreds Of Oregon Veterans Compromised

By KRISTIAN FODEN-VENCIL • DEC 28, 2015

"A full investigation is pending … but it appears that 967 veterans' personal information was shared outside of our control with an individual outside of our agency," said Smith.

Summary

Identified and described three modes of identity fraud

- Identity Theft, Identity Manipulation and Synthetic Identity

Real time fraud detection and awareness requires very careful system architecture

- Data ingestion, storage, retrieval
- Separation of real time, batch and BI/reporting
- Optimization of data layout

Big Data systems allow exploration, R&D on large data sets