TransUnion

Expanding Data Sources to Maximize Fraud Detection: Proof of Life

Brad Daughdrill, PhD
VP, Data Science & Analytics

Agenda

Fraud landscape

Why novel sources of data?

Case study

Can novel data help predict synthetic fraud?

US Market Situation

- Synthetic fraud is one of the fastest-growing financial crimes and often goes undetected because it doesn't impact real individuals who would report suspicious activity.
- Diverse and holistic data views help identify inconsistencies, detect anomalies and reduce vulnerabilities single-source checks cannot address.
- Public records serve as an anchor to real-world identities and, when leveraged for fraud prevention, can expose discrepancies and highlight stability indicators that synthetic identities lack.
- Financial services institutions must find ways to create layered defenses against synthetic fraud and strengthen the identity verification process.

Understanding the fraud landscape is more complex than ever

- Identity verification
- Document verification
- Digital identity risks
- OTP & step-up
- Device reputation
- User interaction
- Omnichannel fraud
- Manual reviews

Organizations must strike a balance between the coverage and fragmentation of their overall fraud strategy portfolios

Mitigating fraud has become an expensive business problem

The estimated cost to businesses globally due to fraud, including losses, prevention tools and headcount is **$5.4 trillion**[1]

Approximately **two of every three** sales transactions are misidentified as false positives, costing businesses **$443 billion**[2]

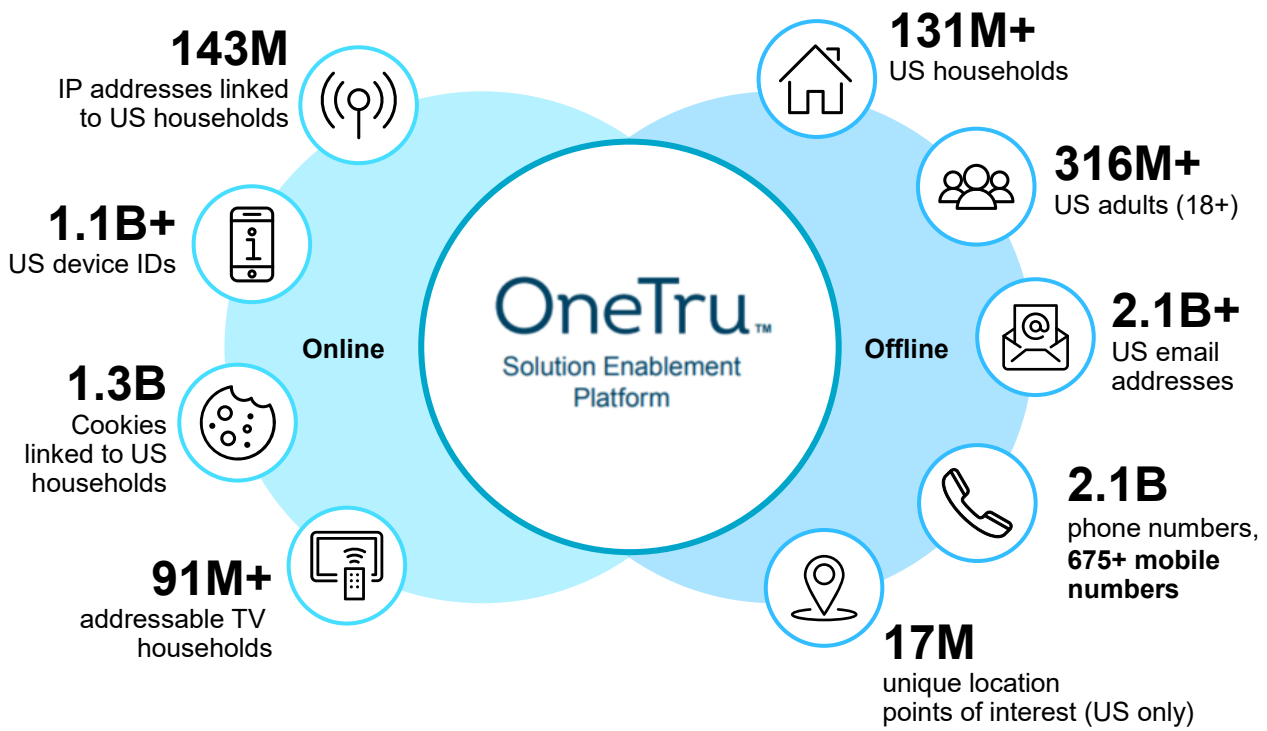
For every **$1** lost to fraud now costs banks **$4.36** in related expenses, such as legal fees and recovery[3]

About **59%** of organizations surveyed expected to increase their budgets for anti-fraud technology over the next two years[4]

1. ”Top Ten Fraud & Identity Trends in 2024,” Fraud.com
2. ”The E-Commerce Conundrum: Balancing False Declines and Fraud Prevention,” Datos Insights
3. “Survey finds fraud costs rising for banks,” ABA Banking Journal, Nov. 16, 2022
4. 2024 Anti-Fraud Technology Benchmarking Report. ACFE & SAS

Building a broader identity graph by connecting people's offline and online identities can help to reduce complexity

Breadth



Graph Metrics

- 15 identifiers per individual
- 45 identifiers per household
- 200+ demographics
- 13,000 propensity and behavioral attributes
- 99.5+% persistency
- 95+% coverage in the US

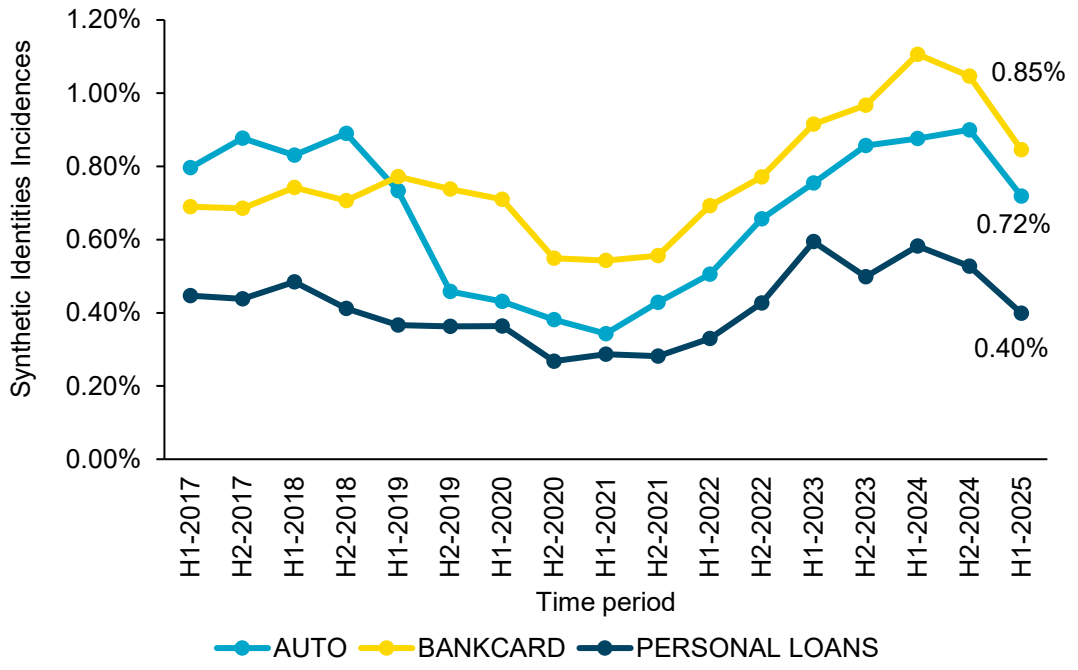
The hypothesis:

Leveraging public records and digital alternative/telephony data to incorporate proven indicators of life in predictive modeling suite will enhance fraud capture.

Synthetic identity fraud continues to rise, indicated by both incidence rates at origination and total lender exposure

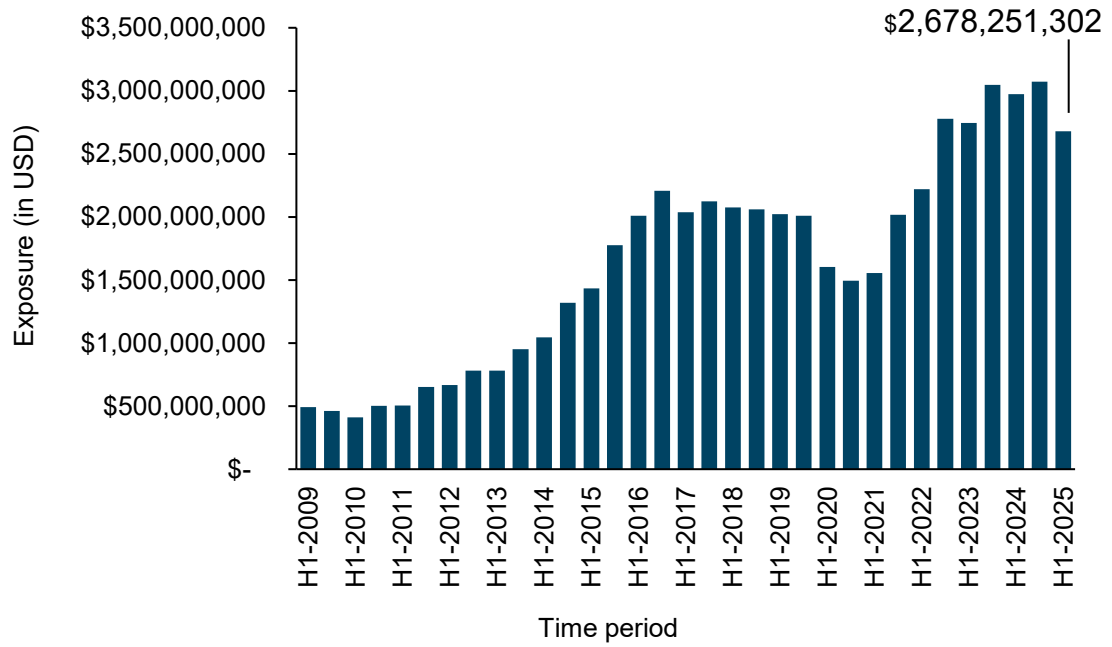
Synthetic ID Incidence on Inquiries

Synthetics for credit inquiries



Aggregated Exposure From Synthetic IDs

Lender exposure of synthetic identities



Note: Previously reported synthetic fraud data may differ from what's currently being provided due to delayed reporting (account information not available the moment they are opened), different reporting schedules (lenders update bureaus once every month but each reporter sets its own schedule).

Sources: TransUnion Credit Database

Providing data source optionality can improve fraud outcomes

Consumer

Leverages one individual's credit history to evaluate patterns of synthetic creation

- FCRA Synthetic Model

Network

Leverages a network of related IDs to identify anomalous identity indicators

- GLBA Synthetic Model
- Digital attributes

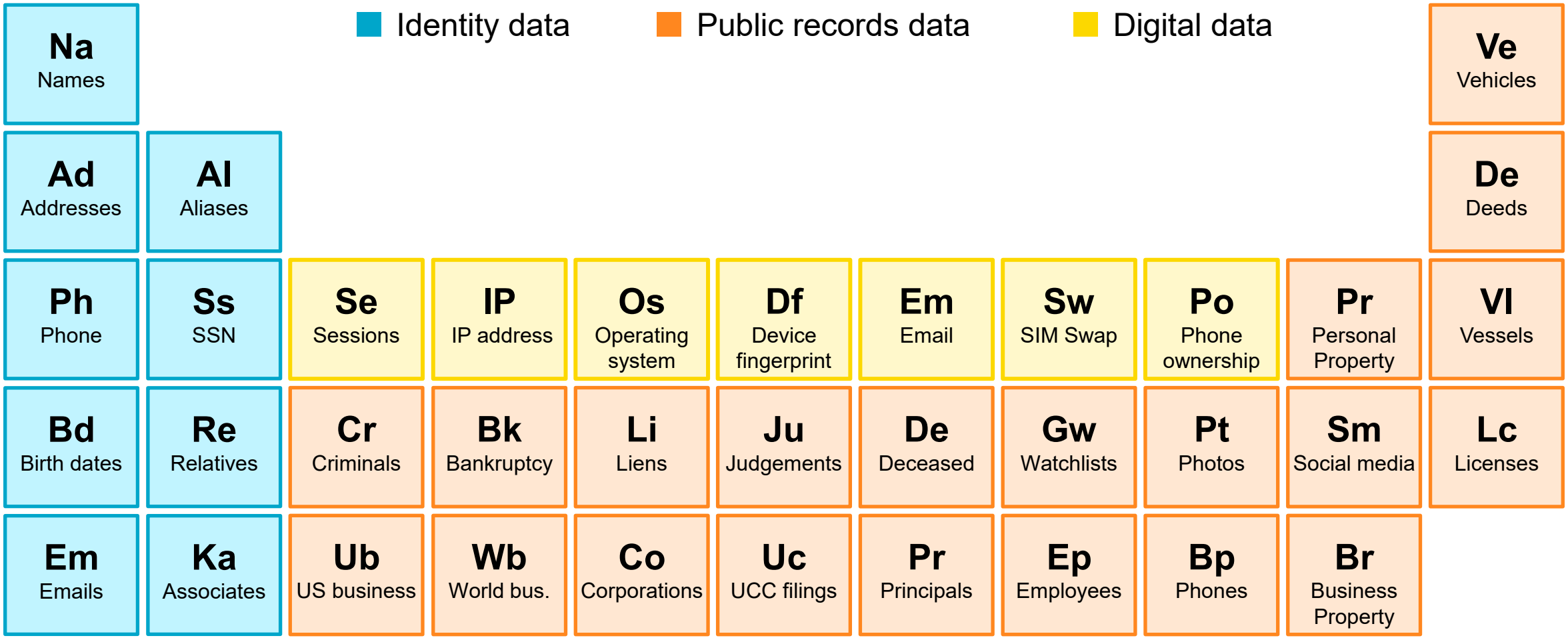
Behavioral

Leverages identity network, assets, bankruptcy, judgements, socials and business data

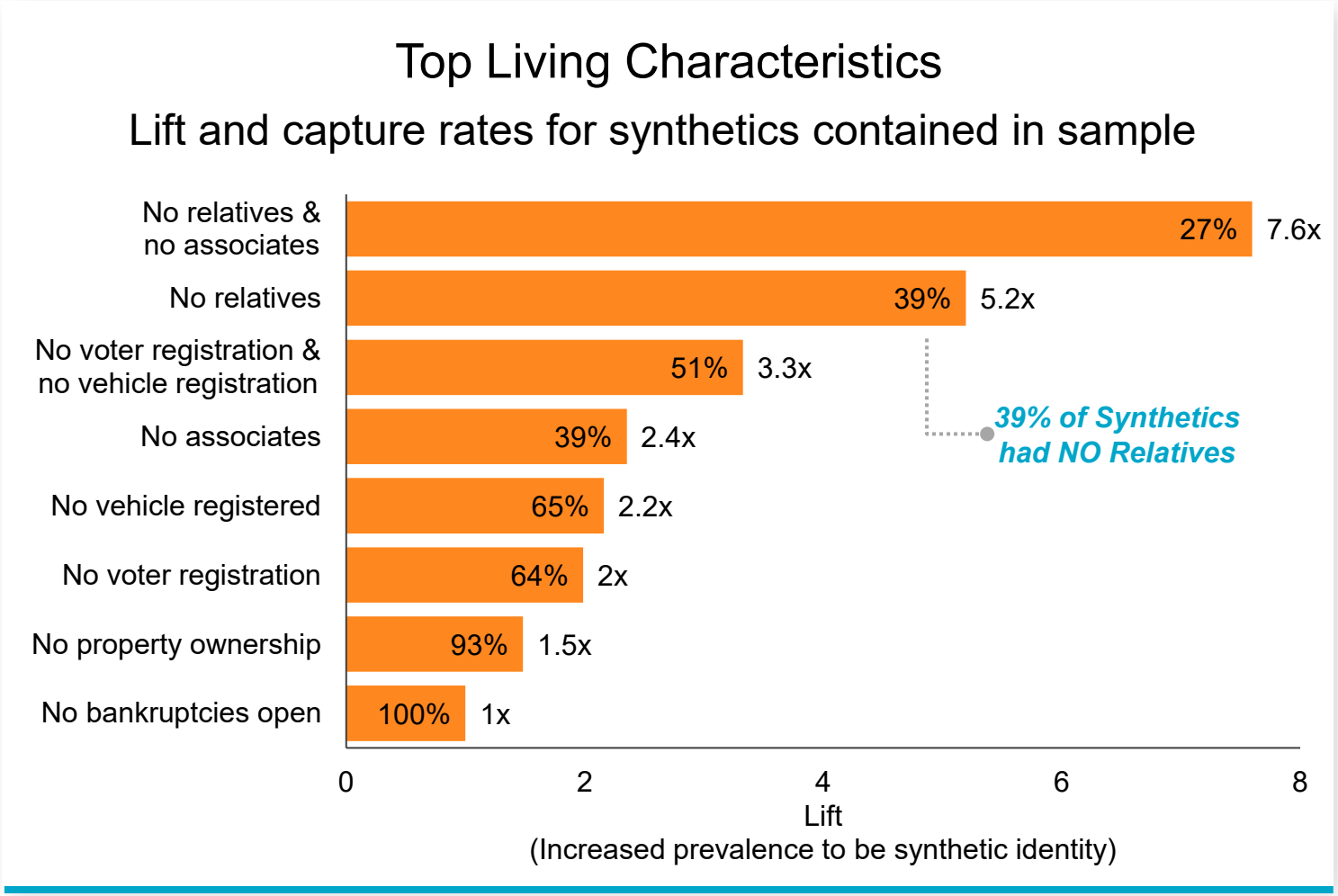
- Predictive attributes

New Joint Synthetic Fraud Model

TransUnion consolidates data from 10K+ public and proprietary feeds into a single, trusted source that powers digital and identity solutions



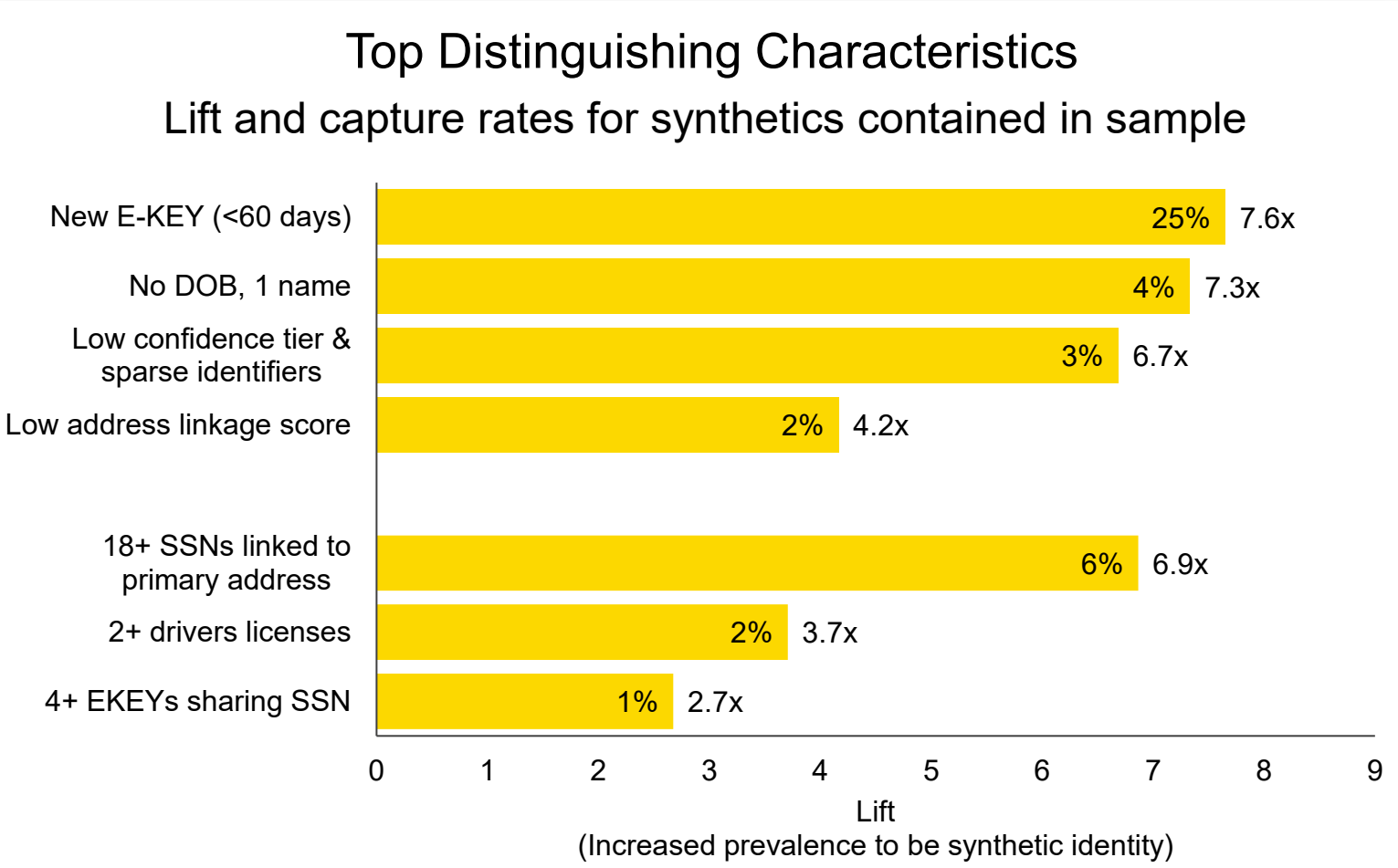
Public records data has characteristics that are predictive of ‘real’ living identities



Living characteristics are attributes that are more often present for living people (*see chart*).

- **No silver bullet**: Many synthetic IDs were registered to vote and had a vehicle registered!
- **Ruled out**: 99.99% of synthetic IDs had no open bankruptcies.

Digital data sources pick up on many identity characteristics that are indicative of a synthetic identity



Synthetic data – The identity is associated to what is likely fake or manipulated data that's not consistently represented together

Identity seeding – The identity is associated to multiple pieces of sensitive PII that have likely been used to enrich the persona and credibility of their identity

Let’s look at an example...

Using data traditionally available in identity risk models

Identity

Consumer 1	Consumer 2
• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 10 • **DOB(s) associated:** 2 • **ID(s) sharing this SSN:** 4	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 1 • **DOB(s) associated:** 1 • **ID(s) sharing this SSN:** 2

Which consumer is the synthetic?

1 or 2

Let’s look at an example...

Layering orthogonal public records data

	Consumer 1	Consumer 2
Identity	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 10 • **DOB(s) associated:** 2 • **ID(s) sharing this SSN:** 4	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 1 • **DOB(s) associated:** 1 • **ID(s) sharing this SSN:** 2
Assets	• **Relative(s):** 3 • **Property owner:** Yes • **Registered to vote:** Yes	• **Relative(s):** 1 • **Property owner:** No • **Registered to vote:** No

Which consumer is the synthetic?

1 or 2

Let’s look at an example...

Adding identity linkage attributes through digital records data

	Consumer 1	Consumer 2
Identity	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 10 • **DOB(s) associated:** 2 • **ID(s) sharing this SSN:** 4	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 1 • **DOB(s) associated:** 1 • **ID(s) sharing this SSN:** 2
Assets	• **Relative(s):** 3 • **Property owner:** Yes • **Registered to vote:** Yes	• **Relative(s):** 1 • **Property owner:** No • **Registered to vote:** No
Digital	• **SSN(s) linked to primary Address:** 1	• **SSN(s) linked to primary Address:** 205

Which consumer is the synthetic?

1 or 2

Let's look at an example...

Novel data can help distinguish synthetic fraud

	Legitimate identity	Synthetic identity
Identity	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 10 • **DOB(s) associated:** 2 • **ID(s) sharing this SSN:** 4	• **Identity-matched credit header:** Yes • **Credit file age:** 50 years old • **Name(s) associated:** 1 • **DOB(s) associated:** 1 • **ID(s) sharing this SSN:** 2
Assets	• **Relative(s):** 3 • **Property owner:** Yes • **Registered to vote:** Yes	• **Relative(s):** 1 • **Property owner:** No • **Registered to vote:** No
Digital	• **SSN(s) linked to primary Address:** 1	• **SSN(s) linked to primary Address:** 205

Very difficult to classify these identities as synthetic using TransUnion credit header data alone: **The credit file is 50 years old!**

Without signals available from public records and digital data, this synthetic identity would go undetected.

Analysis:
Can new models that incorporate **public records** and **digital attributes** from TransUnion OneTru™ Identity Graph optimize performance in predicting synthetic identity fraud?

Performance cohort

Synthetic identities that opened tradelines after September 2021

New synthetic fraud models analyzed

- Public records only
- Digital attributes only
- Joint Synthetic Concept

Baseline models for comparison

- In-Market FCRA Synthetic
- In-Market GLBA Synthetic

Alignment on the language and metrics used to assess and optimize fraud performance is critical

False positive rate

How many incorrect predictions are made for every correct fraud prediction?

(Lower is better)

Fraud detection rate

What percentage of all fraud is being identified?

(Higher is better)

Review rate

How many transactions get flagged as suspicious?

(Lower is better)

Hit rate

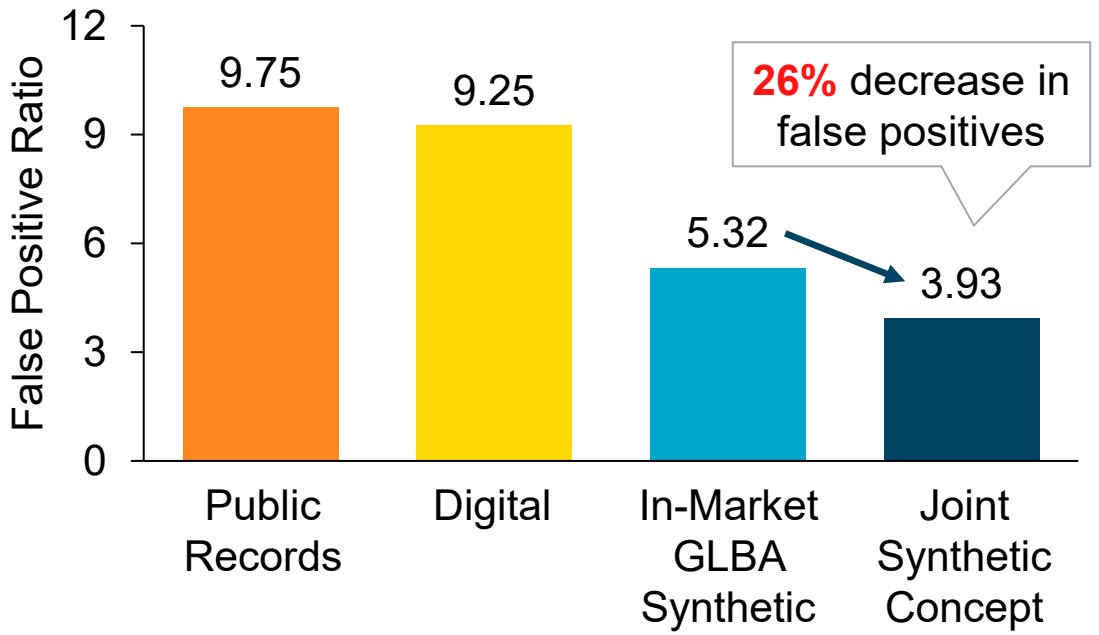
What percentage of fraud predictions end up being fraud?

(Higher is better)

Joint Synthetic Concept Model reduces false positives and captures more synthetic identities when reviewing the same volume

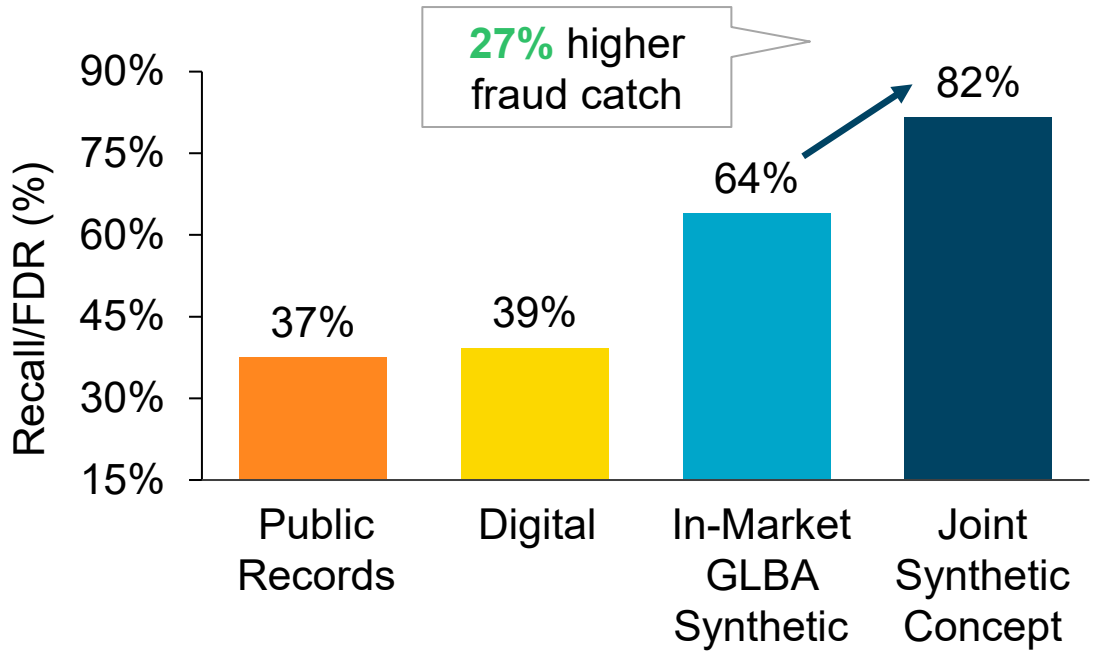
False Positive Rate

How many incorrect predictions are made for every correct prediction?



Fraud Detection Rate

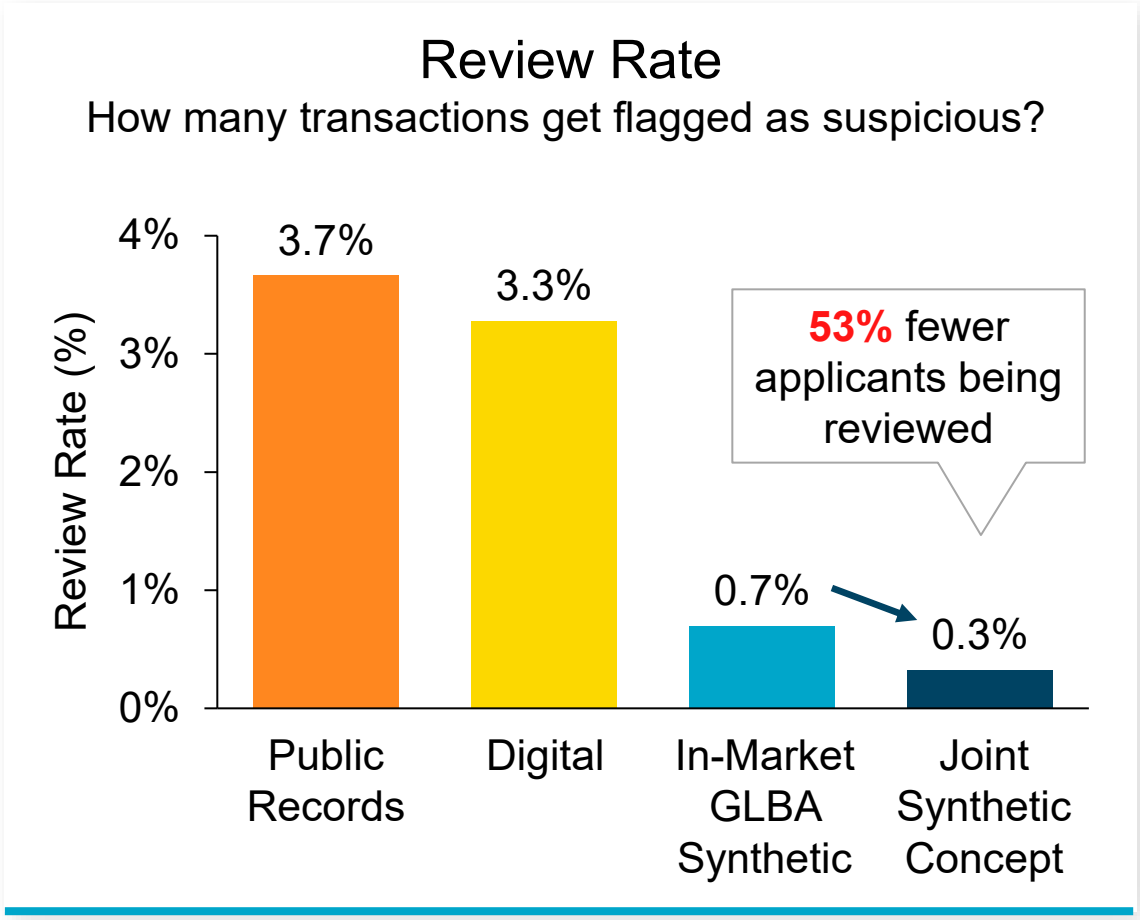
What percentage of all synthetic identities is successfully identified?

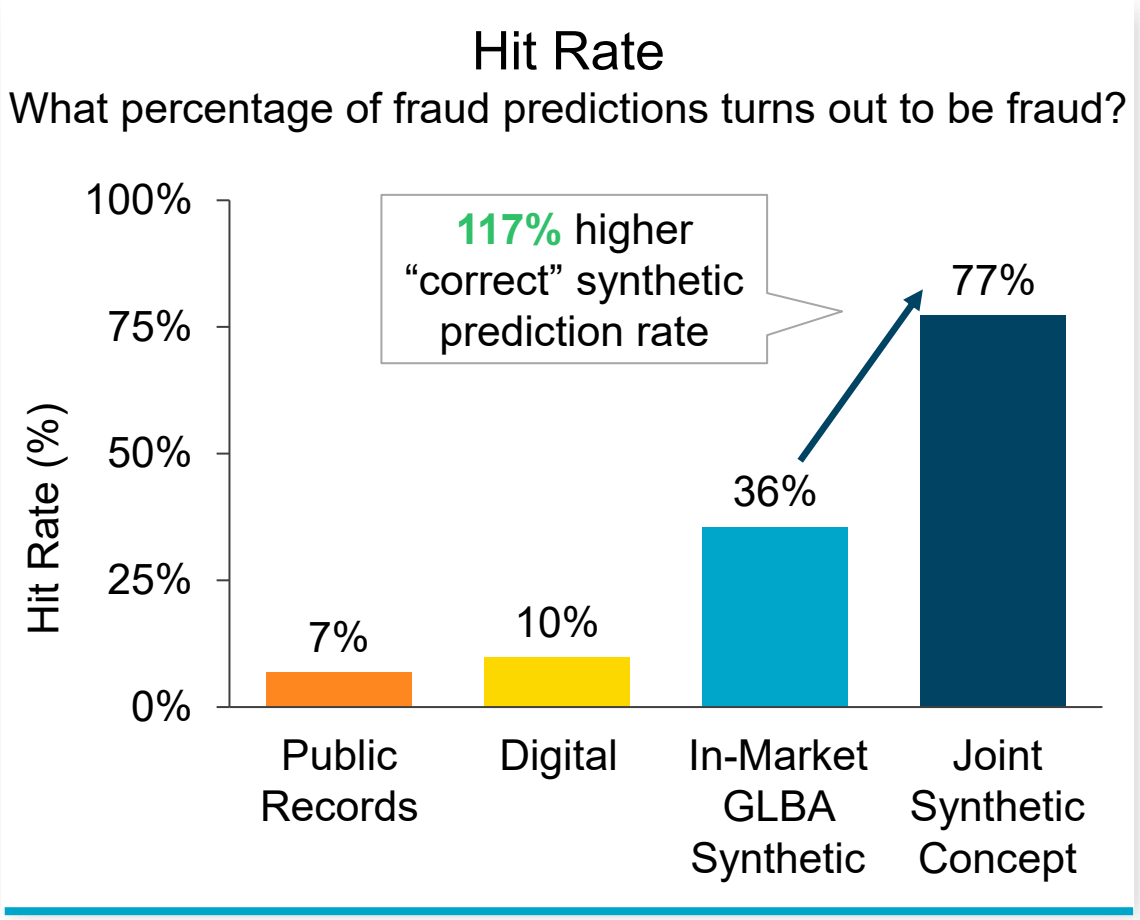


Comparisons made at a fixed review volume – 2% of all observations.
All metrics are computed using a Dataset that contains 170,850 identities, 850 of the identities are Synthetic (0.50% Synthetic Identity Fraud Rate).

Based on internal TransUnion analysis

Joint Synthetic Concept Model requires significantly fewer manual reviews to identify the same amount of fraud with a higher hit rate

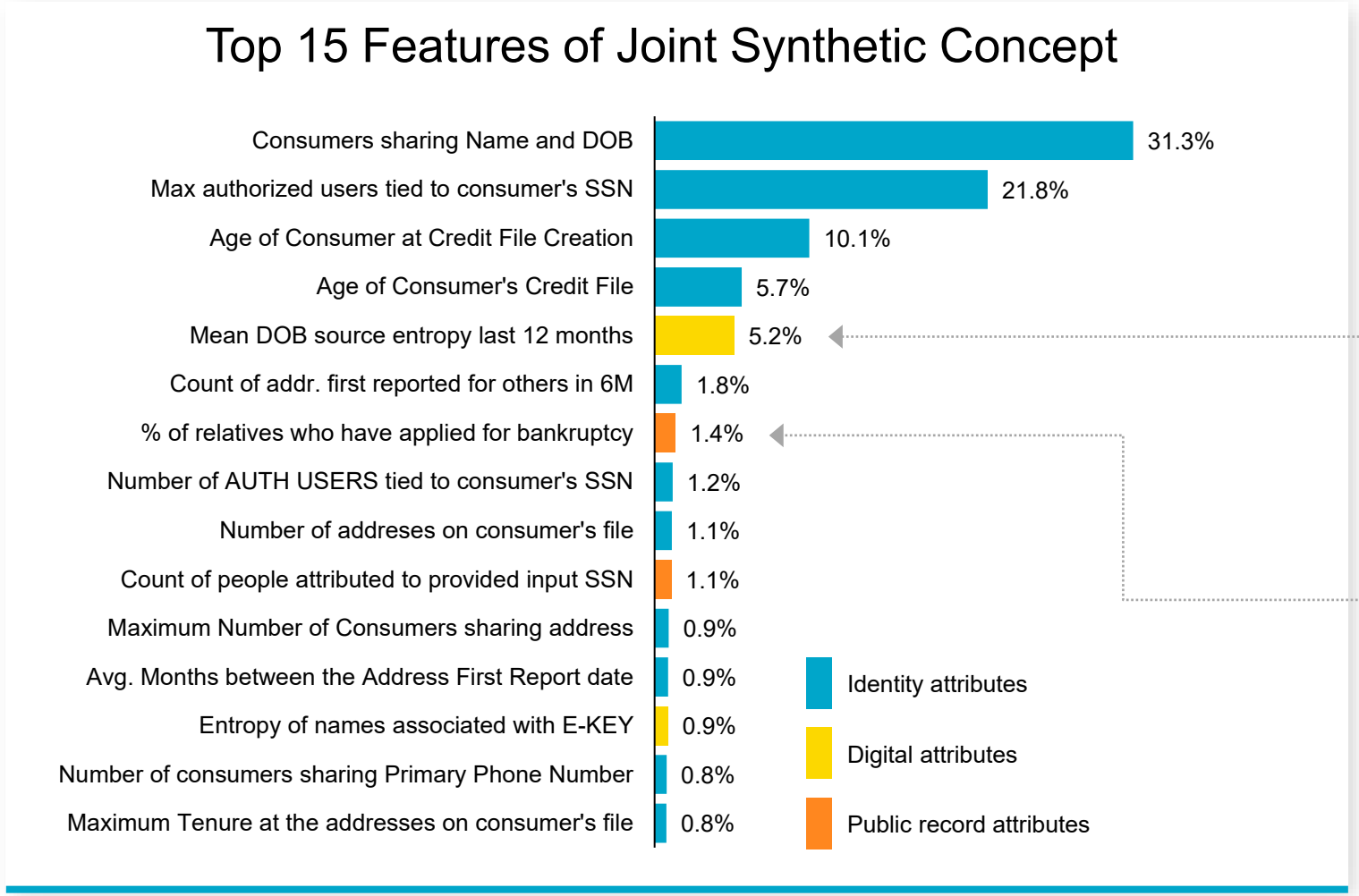




Comparisons made at a fixed fraud capture/synthetic identity detection/recall – 50% of all Synthetic Identities
All metrics are computed using a Dataset that contains 170,850 identities, 850 of the identities are Synthetic (0.50% Synthetic Identity Fraud Rate).

Based on internal TransUnion analysis

The types of data public records add to the model are critical ‘proof-of-life’ indicators



- Over half of the model features are from novel sources of data
- 81% of the performance gain is attributed to identity features

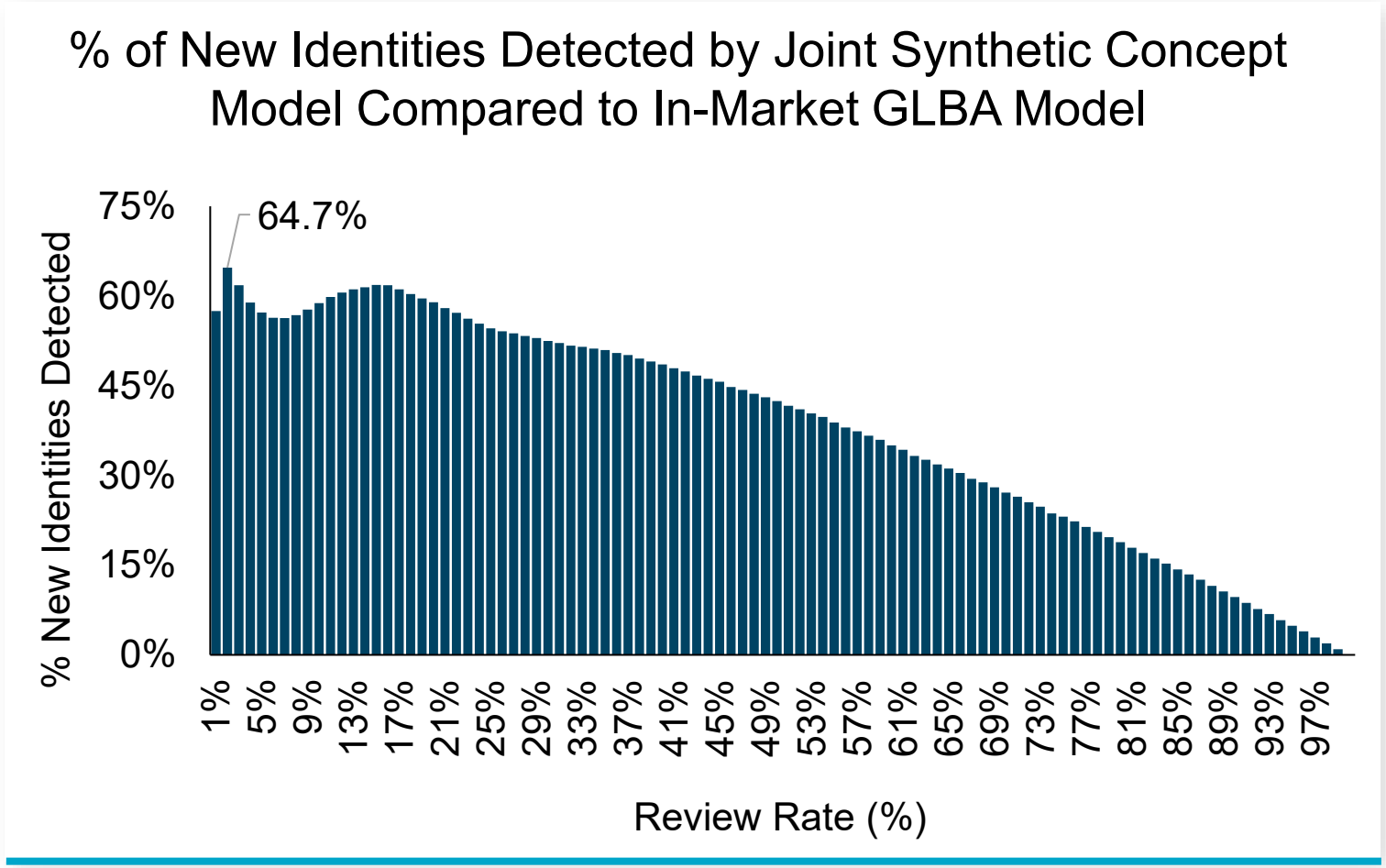
Top digital attribute

Number of consumers with whom consumer is sharing name and DOB

Top public records attribute

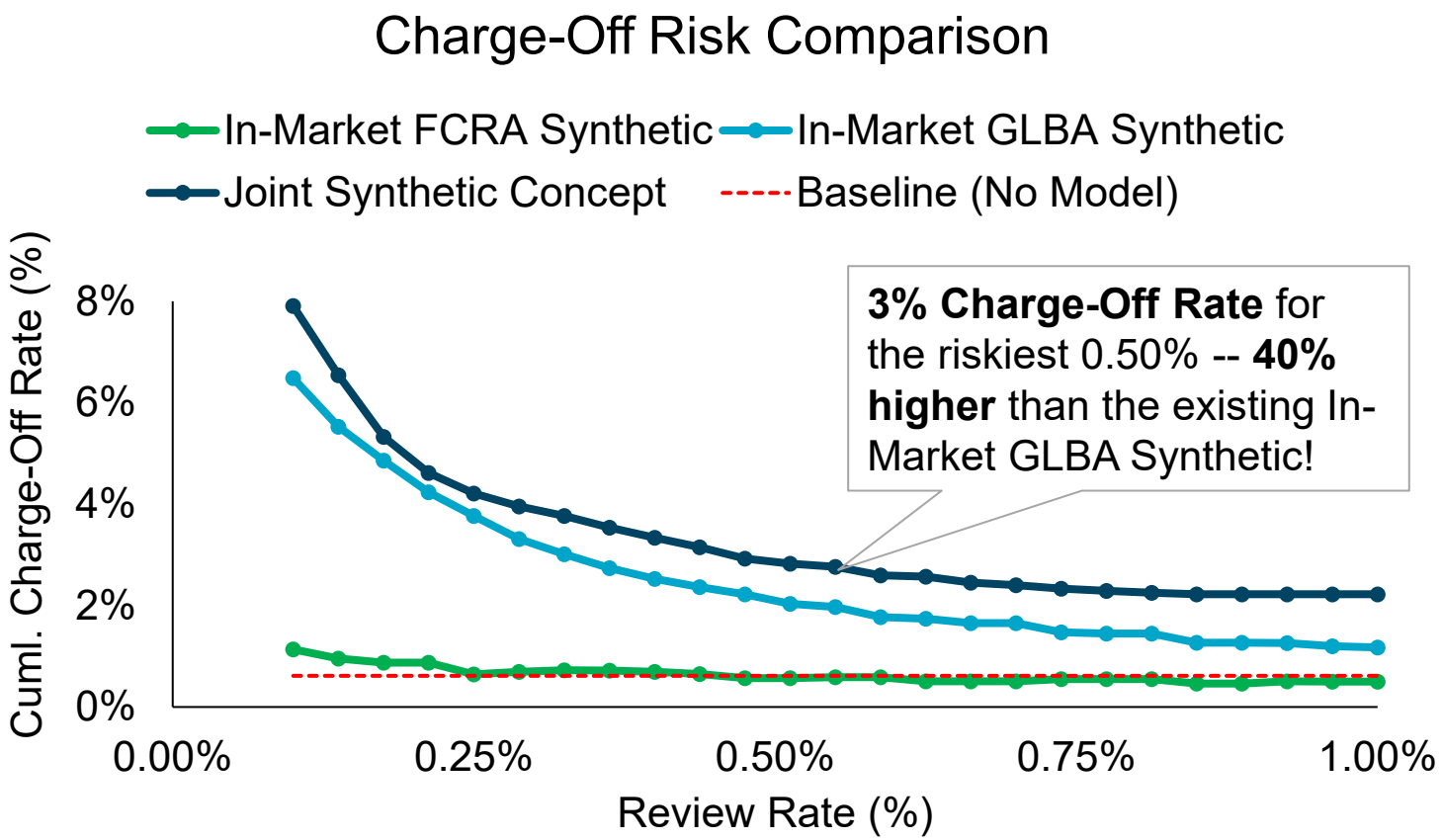
% of relatives who have applied for bankruptcy

The new Joint Synthetic Concept Model identifies a broader range of identities compared to existing models



Have you assessed the efficacy of your current strategy against recently identified complex fraud?

The new model shows significant performance in customer case study on charge-off losses



Sample: 3MM observations; 0.62% Baseline Charge-Off Rate (%)

Can the Joint Synthetic Concept Model outperform at identifying high-risk applicants (not necessarily synthetic identities)?

Each synthetic fraud model captures different types of synthetic IDs more effectively

Cumulative Distribution Illustration

	In-Market FCRA Synthetic	In-Market GLBA Synthetic	Joint Synthetic Concept
Synth ID 1	**95%**	62%	72%
Synth ID 2	25%	**76%**	67%
Synth ID 3	6%	29%	**87%**

Synth ID 1

- **High risk according to credit tradeline data**
- Consumer ID was created via an authorized tradeline (tradeline is now terminated); elevated proportion of tradelines associated to consumer were authorized user tradelines

Synth ID 2

- **High risk according to credit header data**
- Newly created credit file; consumer was 40 years old at time of credit file creation; two other consumers sharing same [name + DOB]

Synth ID 3

- **High risk only when all data source signals are combined!**
- No authorized tradelines, relatively old credit file, consumer identity has “living characteristics” but no driver's license associated to the consumer and 10+ names associated

The numbers above represent each scores’ respective CDF (“Cumulative Distribution”), 90% indicates that 90% of the records had a score less than or equal to that identity

If this interests you…

- Bringing all relevant data together can lead to **higher performance** and **less fragmentation**
- The combined datasets could be highly predictive of additional credit abuse issues like **bust-outs or credit washing**
- Leverage **TransUnion Analytics Consulting** to assess targeted issues in joint research
  - Benchmark synthetic exposure against peers
  - Incorporate our synthetic performance on top of internal fraud performance
  - Custom model that accounts for losses