KGC
The Knowledge Graph Conference

The Knowledge Graph Conference

Monday May 5, 2025 / 11:00 AM – 12:30 PM

*"Constructing **high-quality** knowledge graphs from unstructured and structured data"*

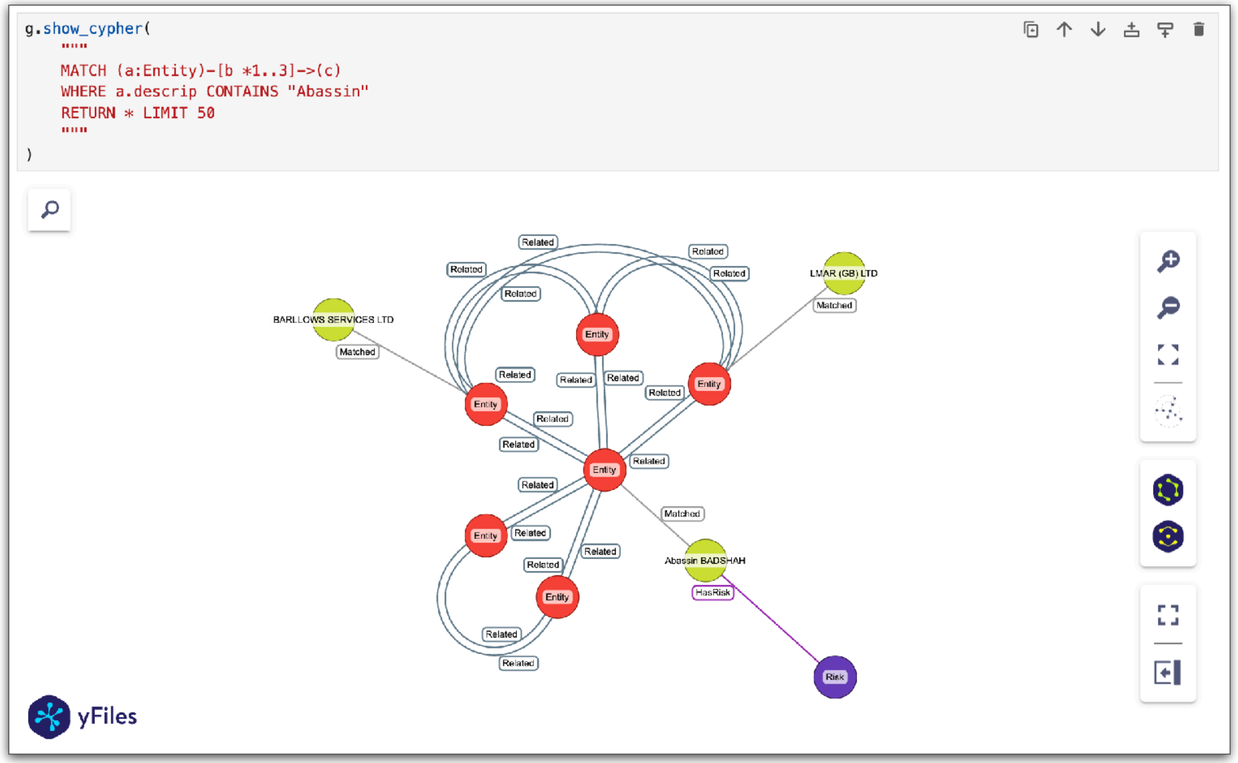
Prashanth Rao, AI Engineer, Kùzu Inc.
Paco Nathan, Principal DevRel Engineer, Senzing

kùzu + Senzing®

KùzuDB + Senzing Workshop

Clone this GitHub repo locally:
github.com/kuzudb/kgc-2025-workshop-high-quality-graphs

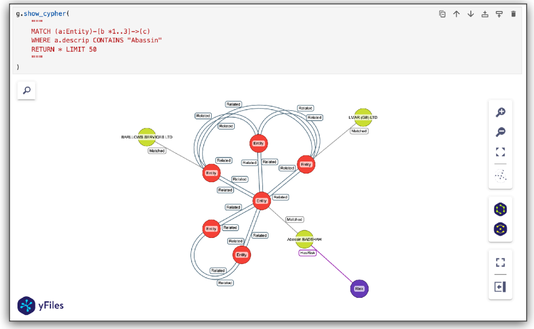
See the **`README.md`** for instructions



KùzuDB + Senzing Workshop

An important obstacle for adopting knowledge graph technology in enterprises is that virtually all of the enterprise-level data is originally stored in unstructured formats or in some structured but non-graph format, such as tables. Therefore, an inevitable first step in adopting graph technologies is to convert these data sources into a high-quality KG.

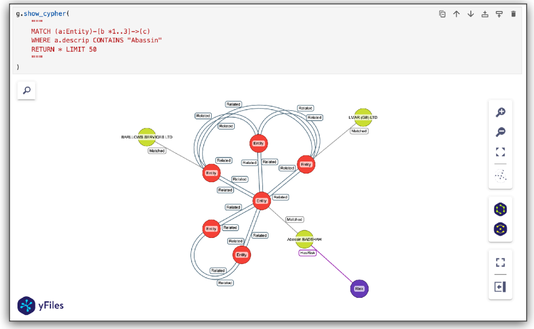
This tutorial walks through the different steps of this process and covers a suite of tools and technologies one can use.



KùzuDB + Senzing Workshop

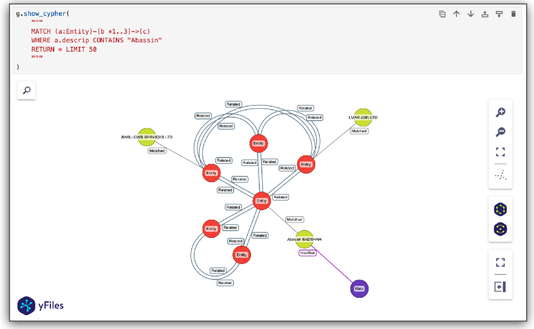
Broadly, KG construction can be divided into three high-level steps:

1. Unstructured text to basic knowledge graph construction: incl. text parsing using NLP libraries, creating lexical and text graphs, named entity/relationship extraction, and entity ranking.
2. Linking the KG with existing structured data such as entities and relationships with those in the basic knowledge graph.
3. KG quality enhancement, which involves steps like entity resolution to increase the quality or specificity of the extracted entities and relationships.



KùzuDB + Senzing Workshop

This tutorial will demo on a live example the evaluation of KGs as it is transformed between these steps that uses several open-source and commercial technologies, such as Senzing, Kùzu (an embedded graph database), and several alternative tools that can be used.



Prerequisites:

- Some experience coding in Python
- Familiarity with popular packages such as Pandas, Jupyter, Docker

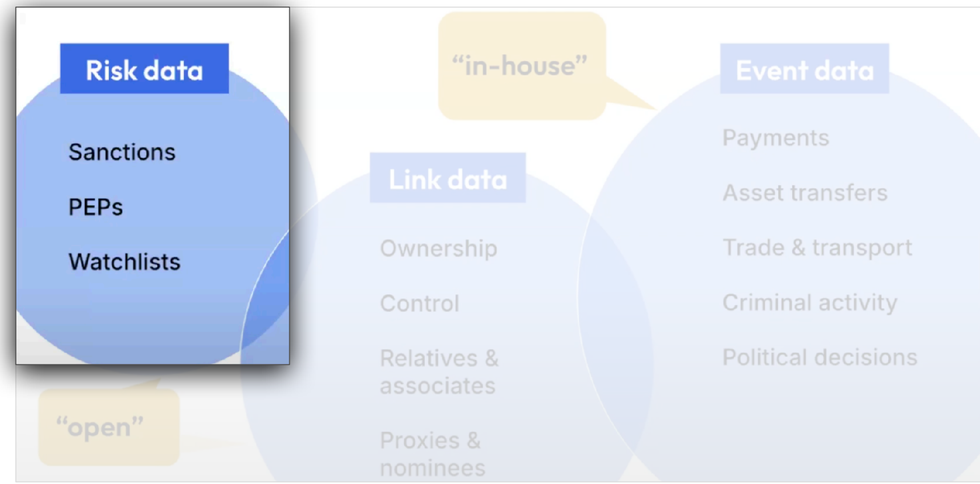
Investigative Graphs

Data sources: risk data

OpenSanctions
opensanctions.org

- provides access to global data on companies and politically exposed persons (PEPs), sanctions lists, and watchlists
- helps flag high-risk entities which may be involved
- 1.7M entities, 242 data sources

See: **“Research Using OpenSanctions Data”** bibliography of papers and studies about transnational corruption.



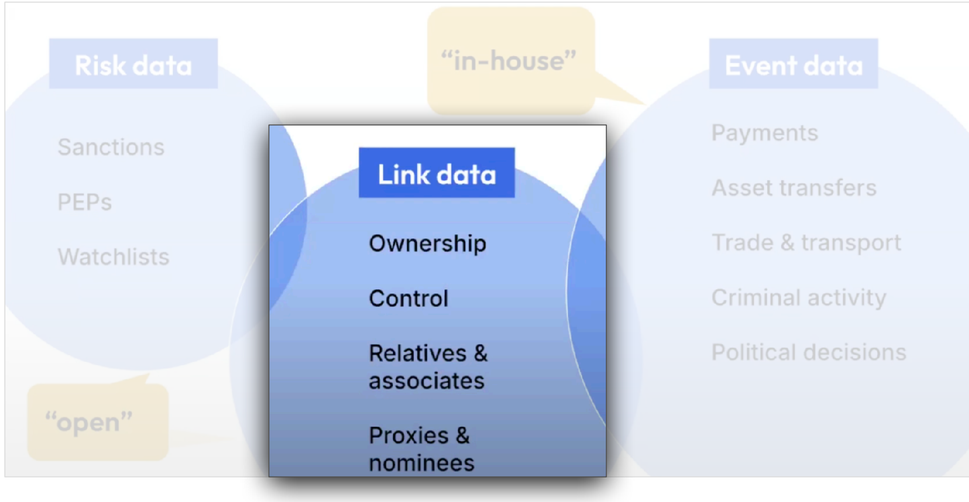
Data sources: link data

Open Ownership
openownership.org

- helps countries generate high-quality data on company ownership that complies with international standards
- data uses across government, civil society, private sector

Global LEI Index
gleif.org/en/lei-data/global-lei-index

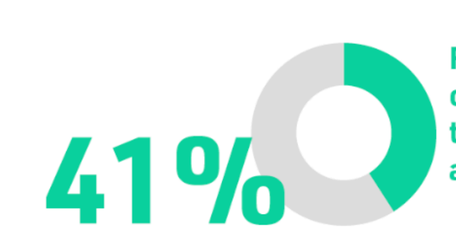
- network of LEI-issuing partner organizations, providing unique legal entity identification worldwide
- mapping companies to ISO standard identifiers, and also to **OpenCorporates**, etc.



Scope of the problem: $3T/year

As a criminal, why risk pulling a weapon on someone in the street, when you could earn much more while hiding behind a laptop?

Crime refers to crime against households of people aged 16 years and over, it does not cover crimes against businesses or those not resident in households, and also excludes homicide and crimes often termed as 'victimless' (for example, possession of drugs).
Source: Crime Survey for England and Wales, 01 October 2023 to 30 September 2024.

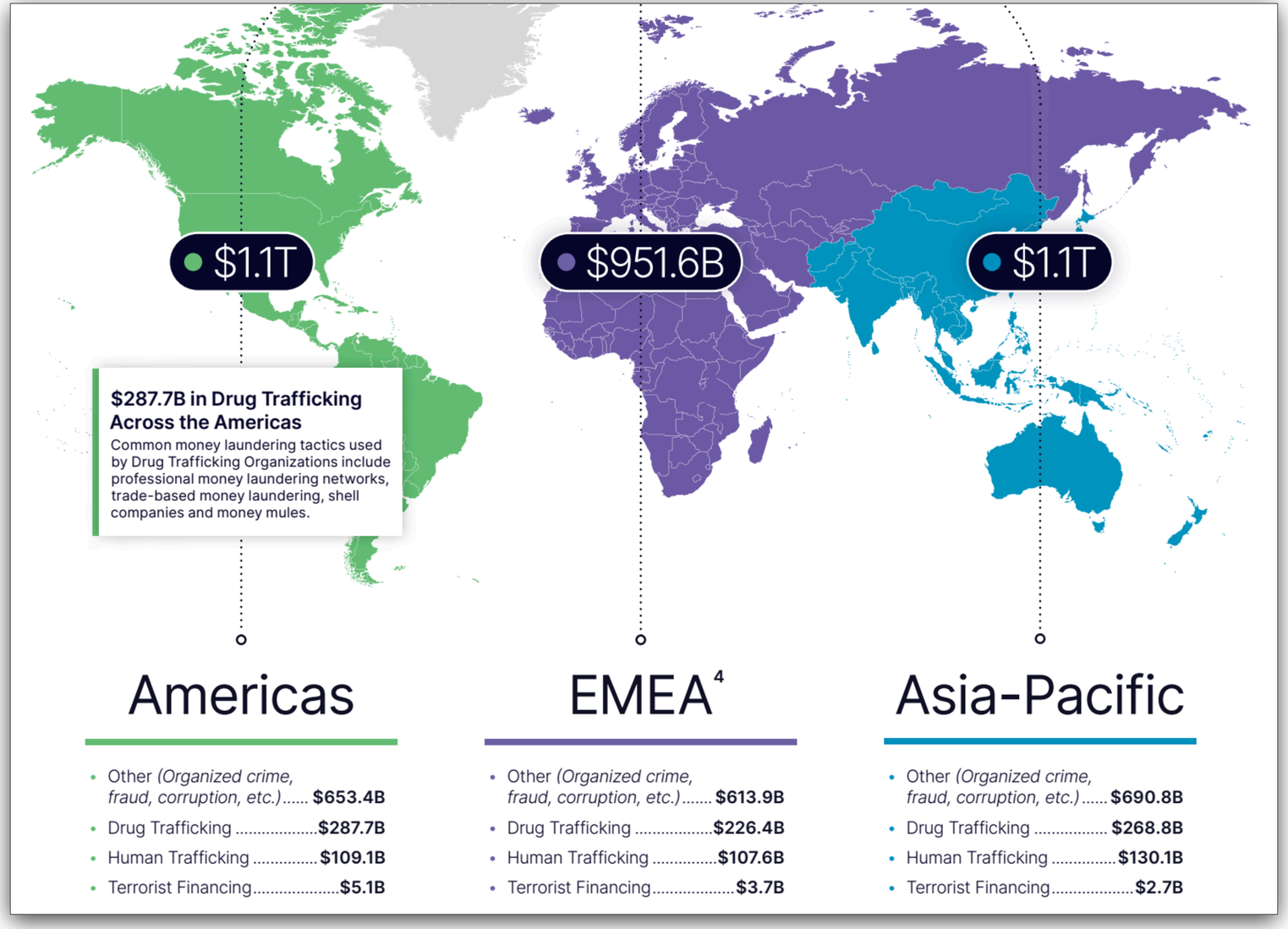


41%

of **all crime in England and Wales** is estimated to be Fraud.

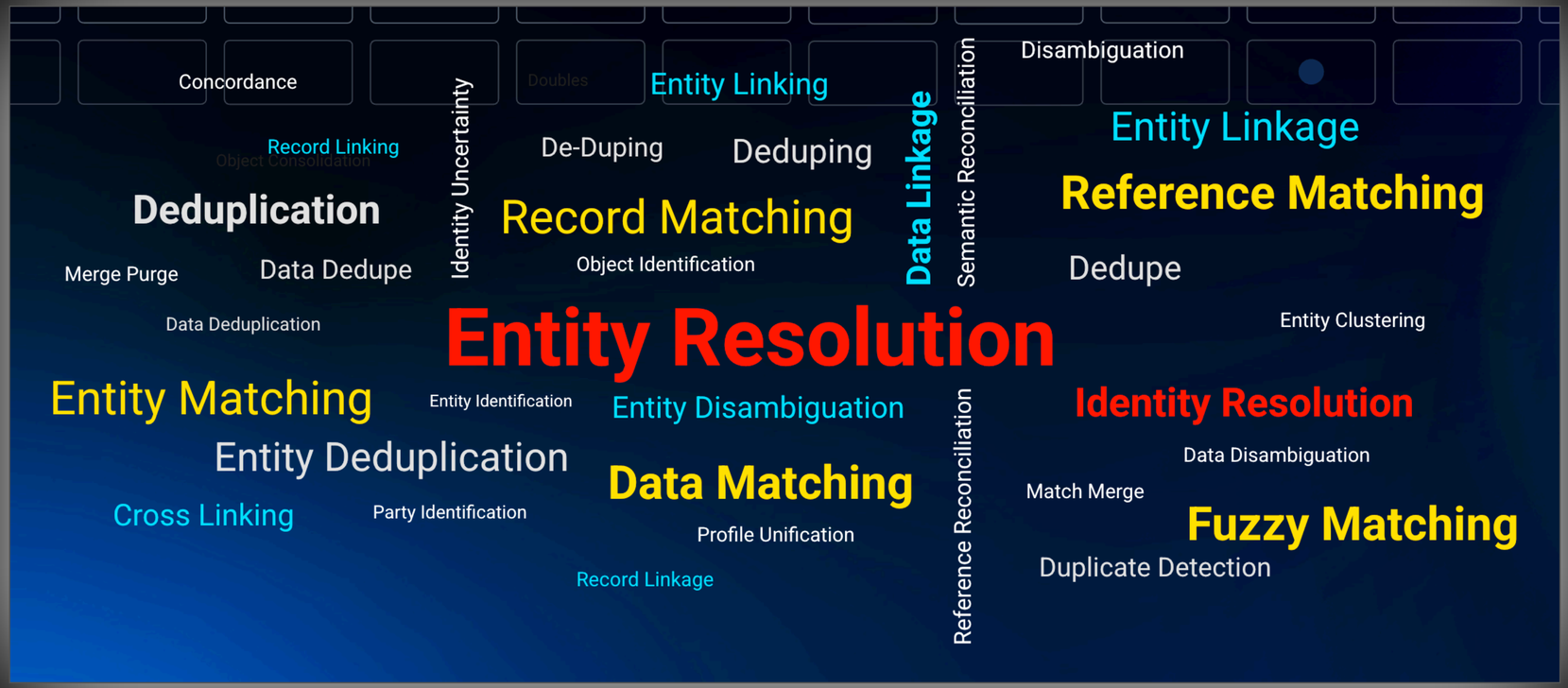
Fraud is the most common crime type in England and Wales.

UK National Strategic Assessment 2025



2024 NASDAQ/Verafin Global Financial Crime Report

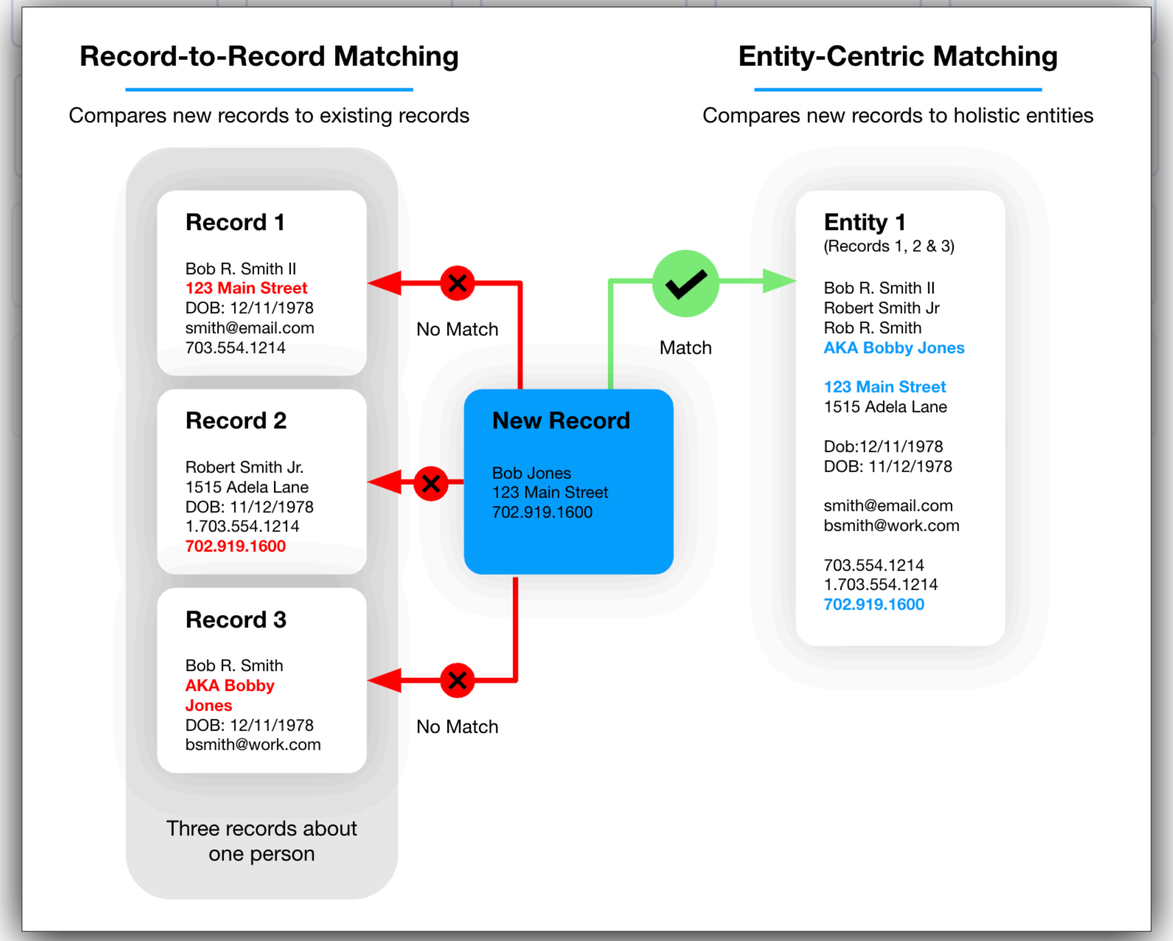
Entity Resolution



Merge 2+ structured data sources, based on using 2+ PII features which connect data records: **people**, **companies**, **vessels**, etc.

Parse names, addresses, and other features into sub-string components, and recognize cultural norms across Arabic, Chinese, Russian, plus many other **globalization contexts**.

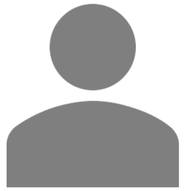
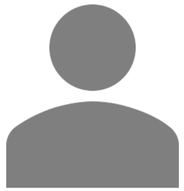
senzing.com/entity-centric-learning-explained



Senzing

Bad Actor Tradecraft

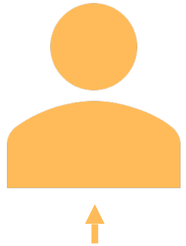
“Channel separation” is the primary tradecraft of bad actors:

Known Money Launderer	Active Account	New Account Opening	Employment Application	Marketing Lead
				
Bob Jones 123 Main Street 702.919.1600	Bob R. Smith II 123 Main Street DOB: 12/11/1978 smith@email.com 703.554.1214	Robert Smith Jr. 1515 Adela Lane DOB: 11-Dec-78 1.703.554.1214 702.919.1600	Bob R. Smith AKA Bobby Jones DOB: 12/11/1978 bsmith@work.com	B. Smith 1515 Adela Lane bsmith@work.com 1.702.919.1600

Senzing

Bad Actor Tradecraft

Channel consolidation requires **entity centric learning**

Known Money Launderer	Active Account	New Account Opening	Employment Application	Marketing Lead
				
Bob Jones 123 Main Street 702.919.1600	Bob R. Smith II 123 Main Street DOB: 12/11/1978 smith@email.com 703.554.1214	Robert Smith Jr. 1515 Adela Lane DOB: 11-Dec-78 1.703.554.1214 702.919.1600	Bob R. Smith AKA Bobby Jones DOB: 12/11/1978 bsmith@work.com	B. Smith 1515 Adela Lane bsmith@work.com 1.702.919.1600

Senzing

Edge Cases

Very different name text, but the same entity:

```
al-Hajj Abdullah Qardash
vs.
Abu ‘Abdullah Qardash Bin Amir
```

Score: 89

Nearly identical name text, but two different entities:

```
John R Smith
vs.
John E Smith
```

Score: 67

Senzing

Edge Cases

Very different address text, but the same entity:

```
#03-28, 400 Orchard Road, 238875 SNG
vs.
400 Orchard Tower #03-28 Orchard Rd, Singapore 238875 Singapore
```

Score: 100

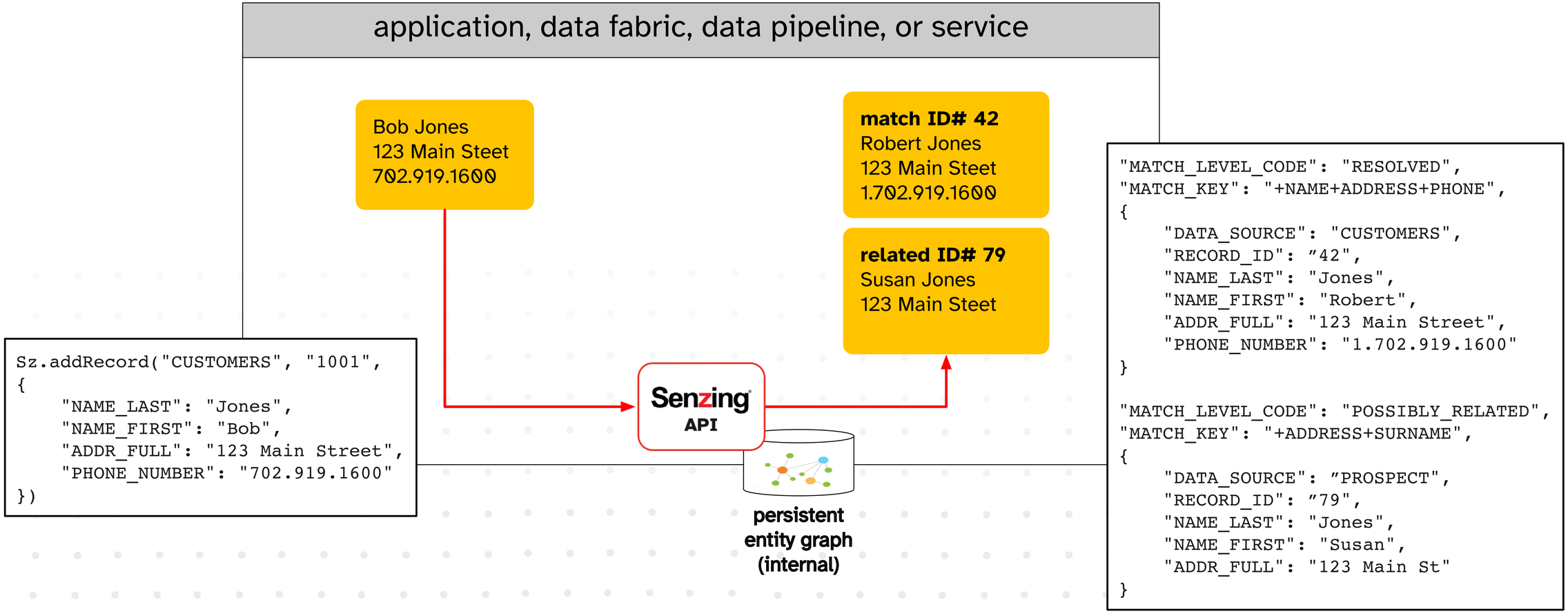
Nearly identical address text, but two different entities:

```
Orchard Tower Plaza, 38 Orchard Road, Singapore, 238875 Singapore
vs.
Orchard Tower Plaza, 31 Orchard Road, Singapore, 238874 Singapore
```

Score: 63

Senzing

JSON in, JSON out



Senzing

Integrate in 3 lines of code

```
#! /usr/bin/env python3

import G2Exception
from G2Engine import G2Engine

# REPLACE /home/user/tryme with the path to your SENZING project created in the quickstart
config = '{ "PIPELINE": { "CONFIGPATH": "/home/user/tryme/etc", "SUPPORTPATH": "/home/user/tryme/data", "RESOURCEPATH": "/user/user/
tryme/resources" }, "SQL": { "CONNECTION": "sqlite3://na:na@/home/user/tryme/var/sqlite/G2C.db" } }'

record = '{"NAME_LAST":"Jones", "NAME_FIRST":"Bob", "ADDR_FULL":"123 Main Street", "PHONE_NUMBER":"702.919.1600"}'

try:

  # Initialize the G2Engine
  g2 = G2Engine()
  g2.initV2("DoIT", config, False)

  # Entity resolve a record (data source, record ID, JSON string)
  g2.addRecord("CUSTOMERS", "1001", record)

  # Get the entity it resolved to
  response = bytearray()
  g2.getEntityByRecordID("CUSTOMER", "1001", response)

  # Display entity JSON
  print(response.decode())

  # Search for entities
  g2.searchByAttributes('{ "NAME_FIRST": "JANE", "NAME_LAST": "SMITH", "ADDR_FULL": "123 Main St, Las Vegas NV" }', response)

  # Display result JSON
  print(response.decode())

except Exception as err:
  print(err)
```

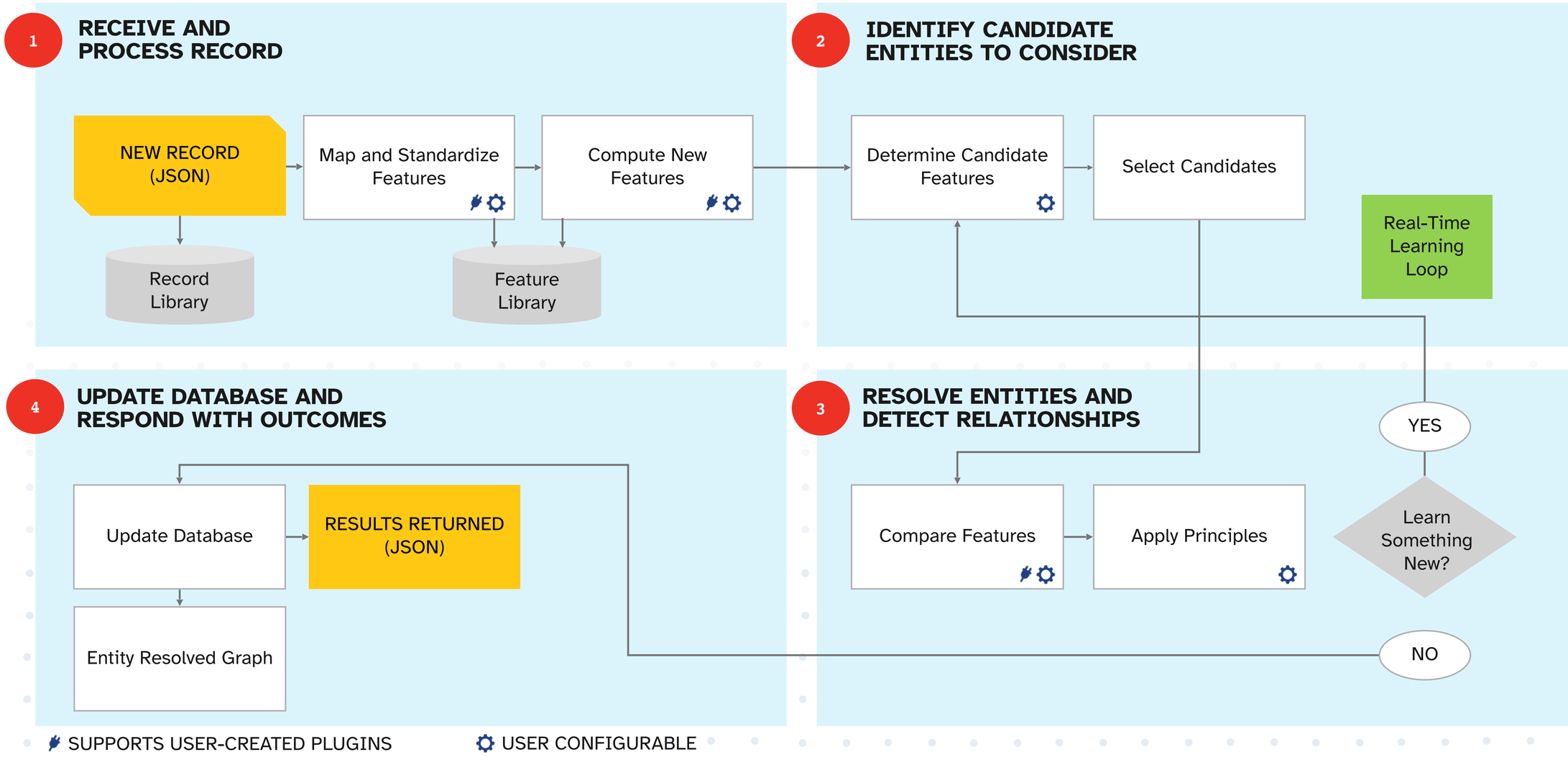
SDKs:

Python

Java

Senzing

Inside the Senzing API



Senzing

Data Sources Integration

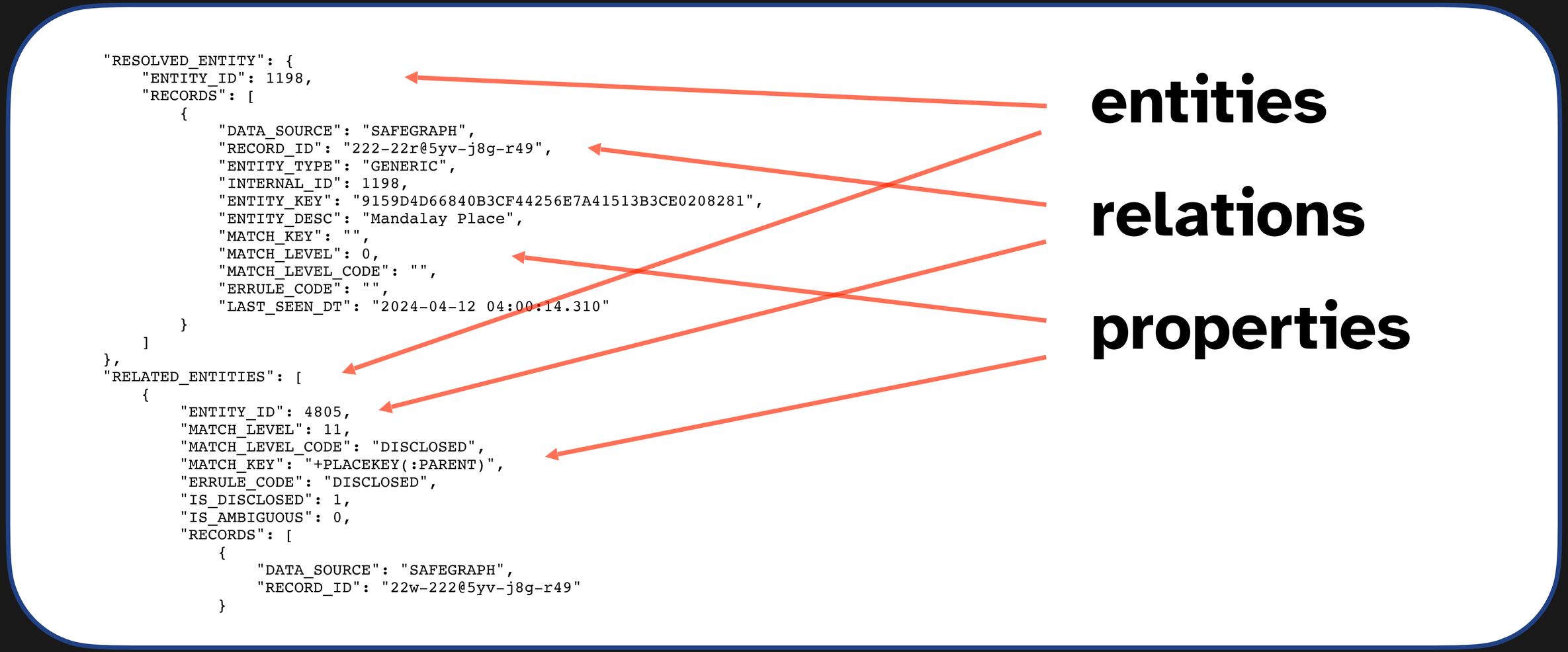
partner name	experience	pre-built mapper	Sz-ready JSON	free sample
BrightQuery	YES			
D&B	YES	YES		
Data Axle	YES	YES		
Dow Jones	YES	YES		
Enformion	YES	in progress		
GLEIF	YES	YES	YES	YES
ICIJ	YES	YES	YES	
Kharon	YES			
Kpler	YES			
LSEG	YES			
Moody's Corporation	YES	in progress		
NominoData	YES	YES	on request	
OpenCorporates	YES	in progress		
OpenOwnership	YES	YES	YES	YES
OpenSanctions	YES	YES	YES	YES
People Data Labs	YES	YES	on request	YES
Rzolut	YES	in progress		
SafeGraph	YES	YES	YES	YES
Sayari	YES	YES		
Spire	YES	YES		
TransUnion	YES			
Verisk (VMS)	YES	in progress		

public data with pre-built mappers
International Consortium of Investigative Journalists (ICIJ)
Office of Foreign Asset Control (OFAC)
National Provider Index (NPI)

public data available by the slice
PPP Loans
Dept of Labor Compliance Actions
Medicare Supplier Directory
Physician Compare
OIG Exclusions

Senzing

ER Generates Graph “Building Blocks”



Senzing

ERKG: Concepts

> **“Think of data as an imperfect map of the underlying entities”**
> — **Abhishek Nagaraj**, UC Berkeley Haas

Paraphrasing, we could say that a **collection of datasets** is an “imperfect map” of:

- **nodes**: the underlying entities (within classes)
- **relations**: pairwise relations between these entities
- **properties**: annotations for all of the above

Identifying the non-obvious **connections** within the data becomes a core challenge, along with measuring, navigating, and predicting the connected elements and any relations among them.

Explainer: NER, RE, and ER, oh my!

People who have *natural language* experience ask about where and how the many different acronyms get used in the KG construct/update process:

- **NER** **named entity recognition**:
  provide **labels for token spans** parsed from **unstructured data**

- **RE** **relation extraction**:
  infer semantic relations (labeled edges) between co-occurring entities

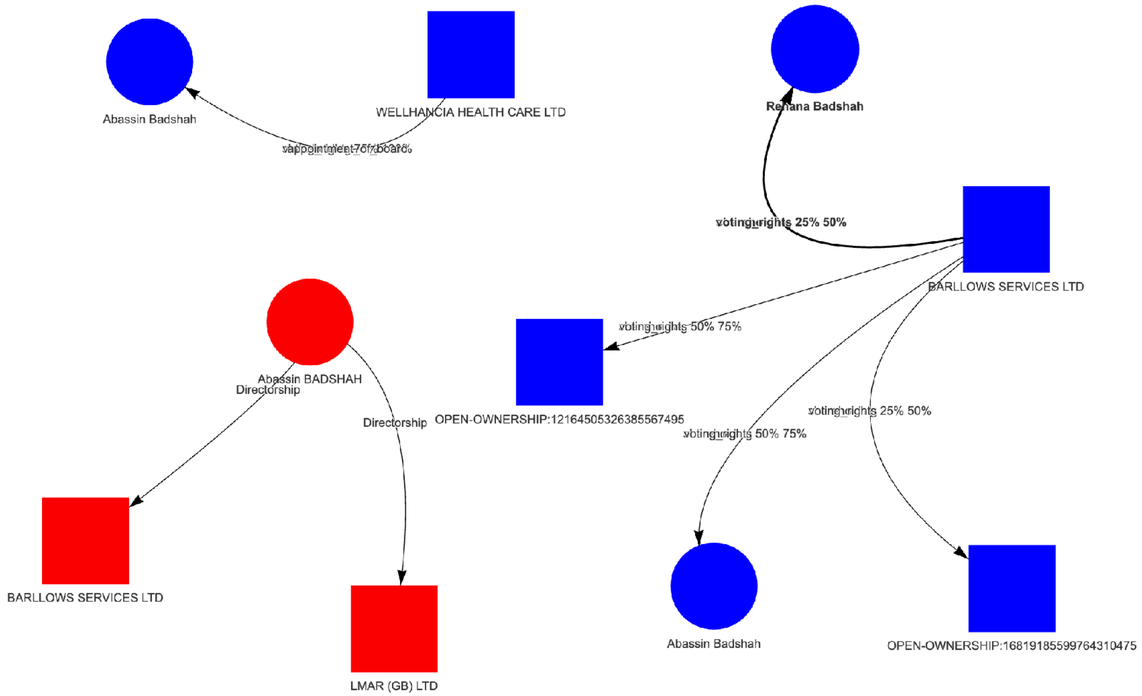
- **ER** **entity resolution**:
  disambiguate consistent **entities** across datasets, based on evidence from PII features from **structured or semi-structured data**

- **EL** **entity linking**:
  integrate structured/ER and unstructured/NER **together** within KGs

Graph Analytics

Typical pattern in anti-fraud use cases:

1. Use entity resolution to improve data quality and connect graph elements
2. Partition the graph into subgraphs (e.g., Louvain community algorithm)
3. Rank the entities of interest (e.g., degree centrality algorithm)
4. Process these subgraphs through case management tools



before **entity resolution**

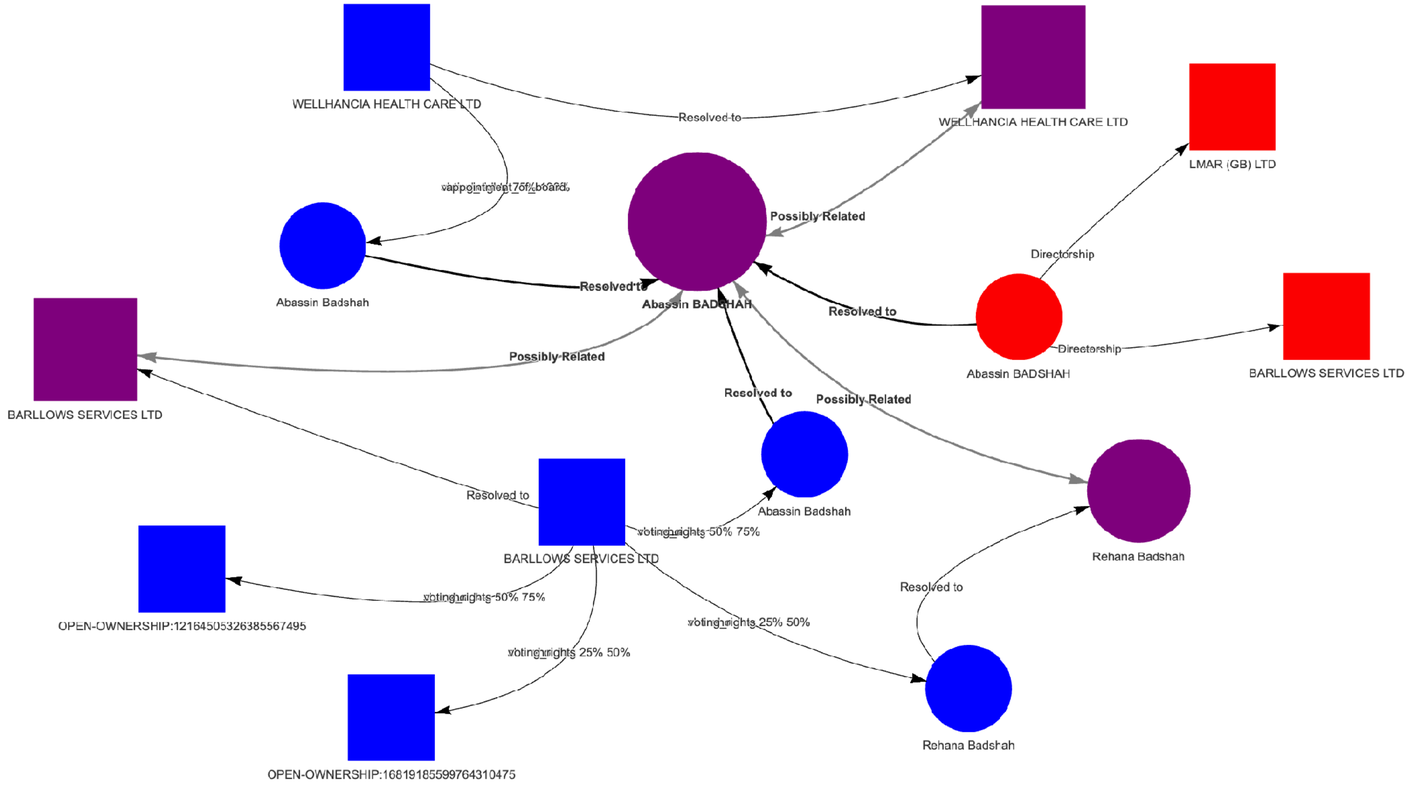
Senzing

Graph Analytics

Typical pattern in anti-fraud use cases:

1. Use entity resolution to improve data quality and connect graph elements
2. Partition the graph into subgraphs (e.g., Louvain community algorithm)
3. Rank the entities of interest (e.g., degree centrality algorithm)
4. Process these subgraphs through case management tools

NB: after entity resolution, now we see that **Abassin Badshah** becomes top-ranked as most central element in this subgraph.



after **entity resolution**

Senzing

EDA Tooling: Audits, Case Management

```
(g2) why 1

Why for entity: 1
```

INTERNAL_ID	1	100121
DATA_SOURCES	OPEN-SANCTIONS: NK-25vyVFzt8vdJGgAXMRTwTJ	OPEN-OWNERSHIP: 17207853441353212969
WHY_RESULT	NAME+ADDRESS+NATIONALITY Principle 162: CNAME_CFF	NAME+DOB+ADDRESS+NATIONALITY Principle 160: CNAME_CFF_CEXCL
NAME	Abassin BADSHAH [1] └── Abassin Badshah (full:100)	Abassin Badshah [1]
DOB	1985-05-12 [1] └── 1985-05-01 (full:86)	1985-05-01 [1]
ADDRESS	31 Quernmore Close, Bromley, Kent, United Kingdom, BR1 4EL [1] └── 31, Quernmore Close, Bromley, BR1 4EL (full:100)	31, Quernmore Close, Bromley, BR1 4EL [#1] └── 31 Quernmore Close, Bromley, Kent, United Kingdom, BR1 4EL (ful
OTHER_ID	NK-25vyVFzt8vdJGgAXMRTwTJ OPEN-SANCTIONS [1]	
NATIONALITY	gb [100] └── GB (full:100)	GB [100]
RECORD_TYPE	PERSON [100]	PERSON [100]
NAME_KEY	ABSN|BTX [1] ABSN|BTX|ADDRESS.CITY_STD=BRML [1] ABSN|BTX|DOB.MMYY_HASH=0585 [1] ABSN|BTX|POST=BR14EL [1] ABSN|BTX|DOB.MMDD_HASH=1205 [1] ABSN|BTX|DOB=81205 [1]	ABSN|BTX [1] ABSN|BTX|ADDRESS.CITY_STD=BRML [1] ABSN|BTX|DOB.MMDD_HASH=0501 [1] ABSN|BTX|DOB.MMYY_HASH=0585 [1] ABSN|BTX|DOB=80501 [1] ABSN|BTX|POST=BR14EL [1]
ADDR_KEY	31|QRNMR KLS||BR14EL [2] 31|QRNMR KLS||BRML [2]	31|QRNMR KLS||BR14EL [2] 31|QRNMR KLS||BRML [2]
ID_KEY	OTHER_ID=NK25VYVFZT8VDJGGAXMRTWTJ [1]	
REL_ANCHOR		OOR 17207853441353212969 [1]
REL_POINTER	OPEN-SANCTIONS NK-3p3mmVWmjwVtTfKchz4kNE [1] OPEN-SANCTIONS NK-SKAADAiqiZ78JsJjeg72Te [1]	

```
lines 1-42/42 (END)
```


Senzing