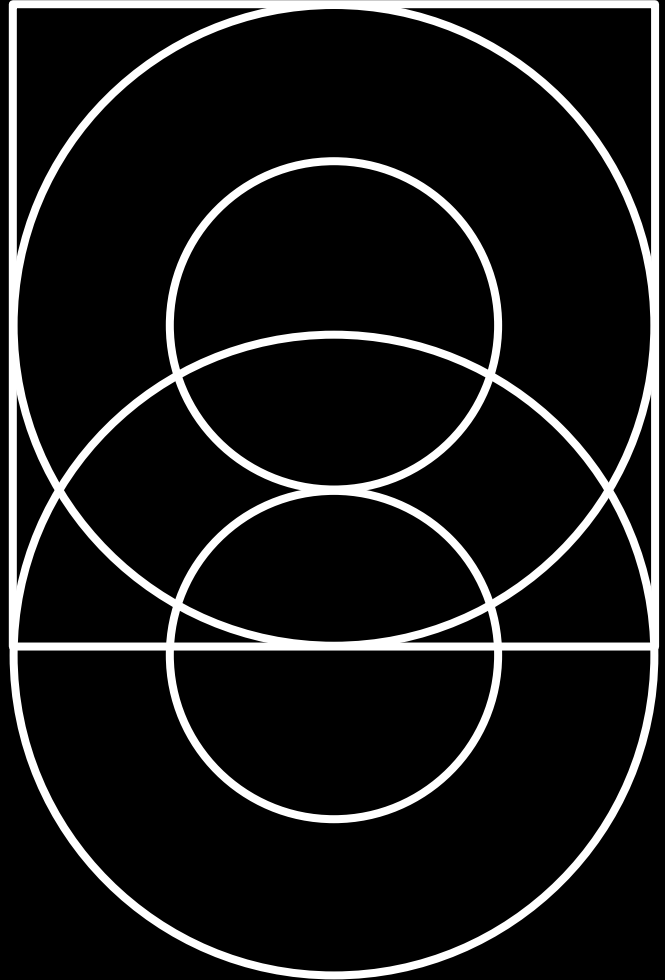


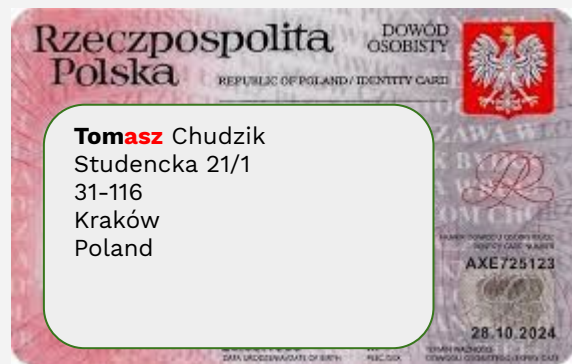
Unit8™

Entity Resolution

digital natives

unit8.co





Please hire me !

hr@bigbank.com

Please hire me !

I want to work for you !

Regards,
Tomek Chudzik

Customers
DB



Johannes Müller
Mercedesstraße 52
Stuttgart



Sanction List



Johannes Mueller
Mercedesstraße 52
Stuttgart



Question:

Am I allowed to make
a deal with Johannes
Müller ?

What is entity resolution?

What is entity resolution ?

Process of finding records in datasets
which represents same entity e.g.
person or company

Is it the same person ?

Johannes Müller
Mercedesstraße 52, Stuttgart

Johannes Mueller
Mercedesstraße 52, Stuttgart

Is it the same person ?

Johannes Müller

Mercedesstraße 52, Stuttgart
born on 19th September 1988

Johannes Müller

Friedrichstraße 2, Stuttgart
born on 19th September 1988

Is it the same person ?

Johannes Müller

Remote Street 522, Ramsau

Dr. J. Müller

Remote Street 522, Ramsau



Is it the same person ?

Johannes Müller
Large Street 522, Berlin

Dr. J. Müller
Large Street 522, Berlin



Is it the same company ?

Unit8 SA
Chemin de Pré-Val 7, Morges

Unitt8 SA
Chemin de Pré-Val 7, Morges

Is it the same company ?

Unit8 SA
Chemin de Pré-Val 7, Morges

Unit7 SA
Chemin de Pré-Val 7, Morges

Is it the same company ?

Tech Corp GmbH
Chemin de Pré-Val 7, Morges

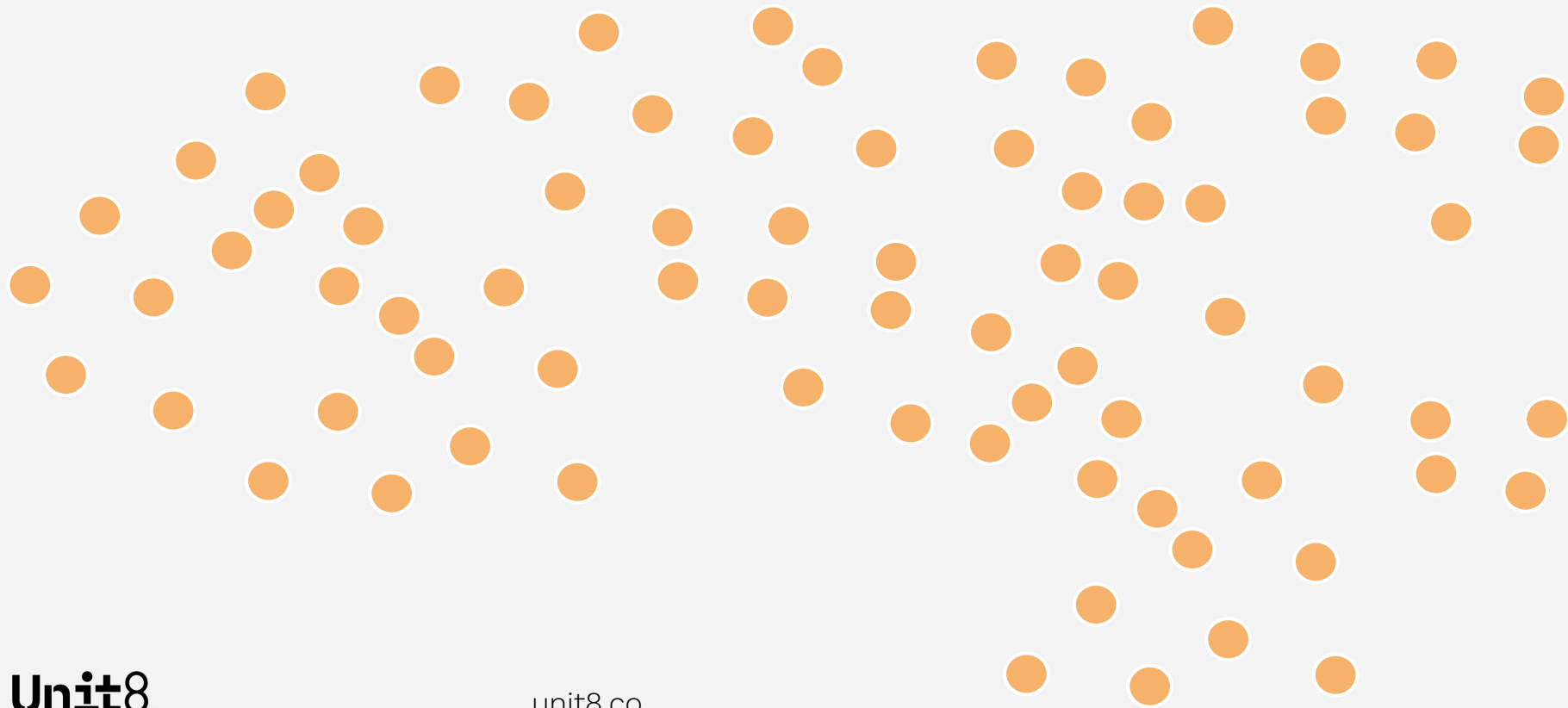
Tech Corp AG
Chemin de Pré-Val 7, Morges

Is it the same company ?

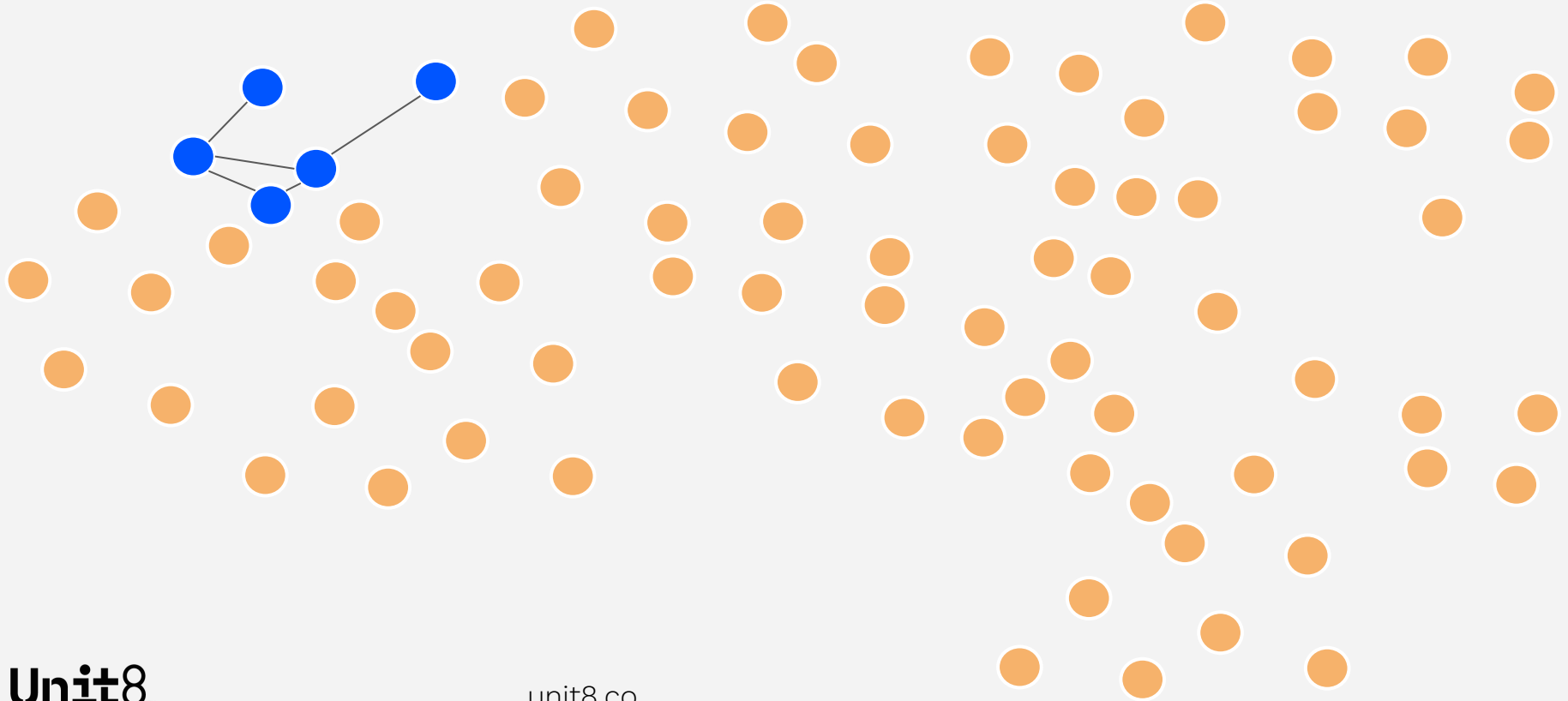
There Is a Group of Young People
With Dreams, Who Believe They Can
Make the Wonders of Life Under the
Leadership of Uncle Niu Internet
Technology Co Ltd

Uncle Niu Co Ltd

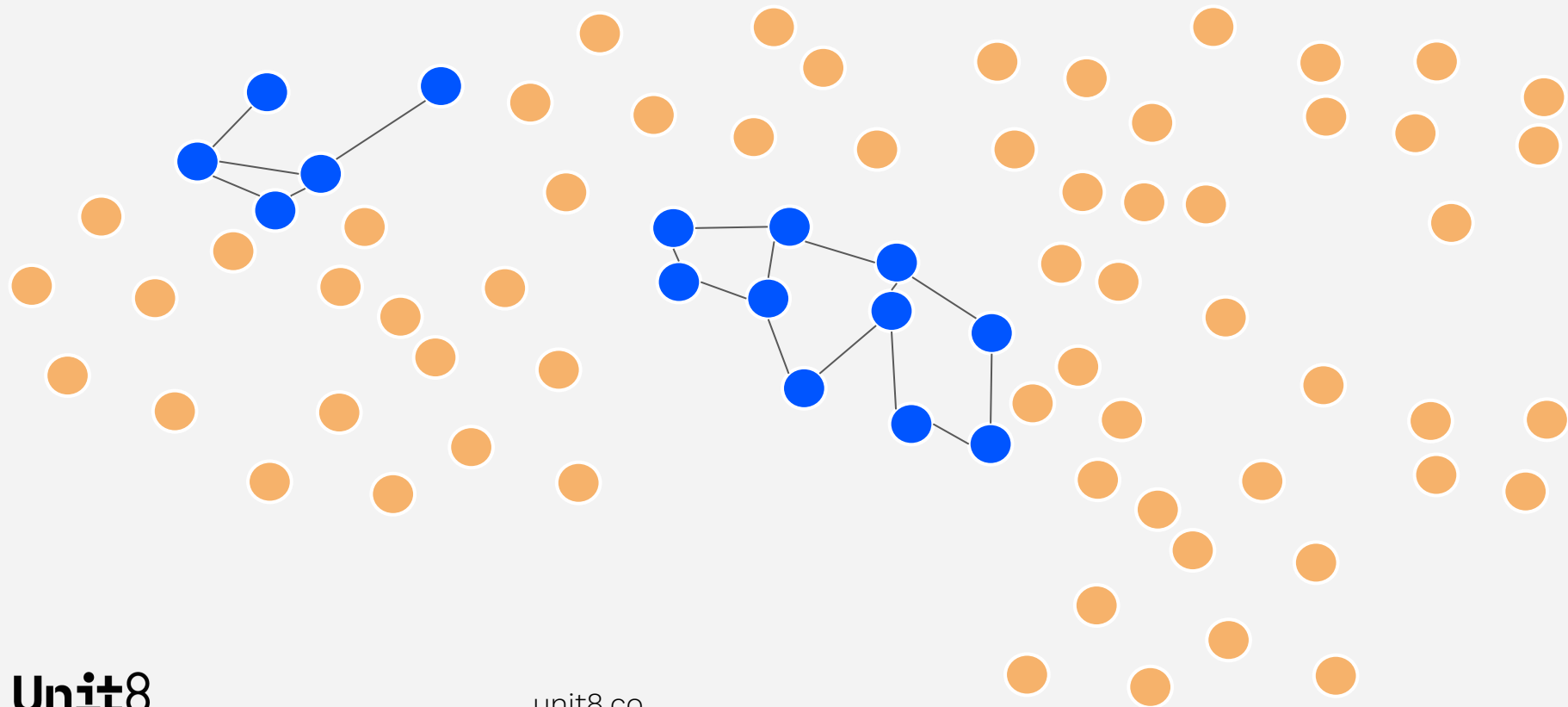
From chaos to structure



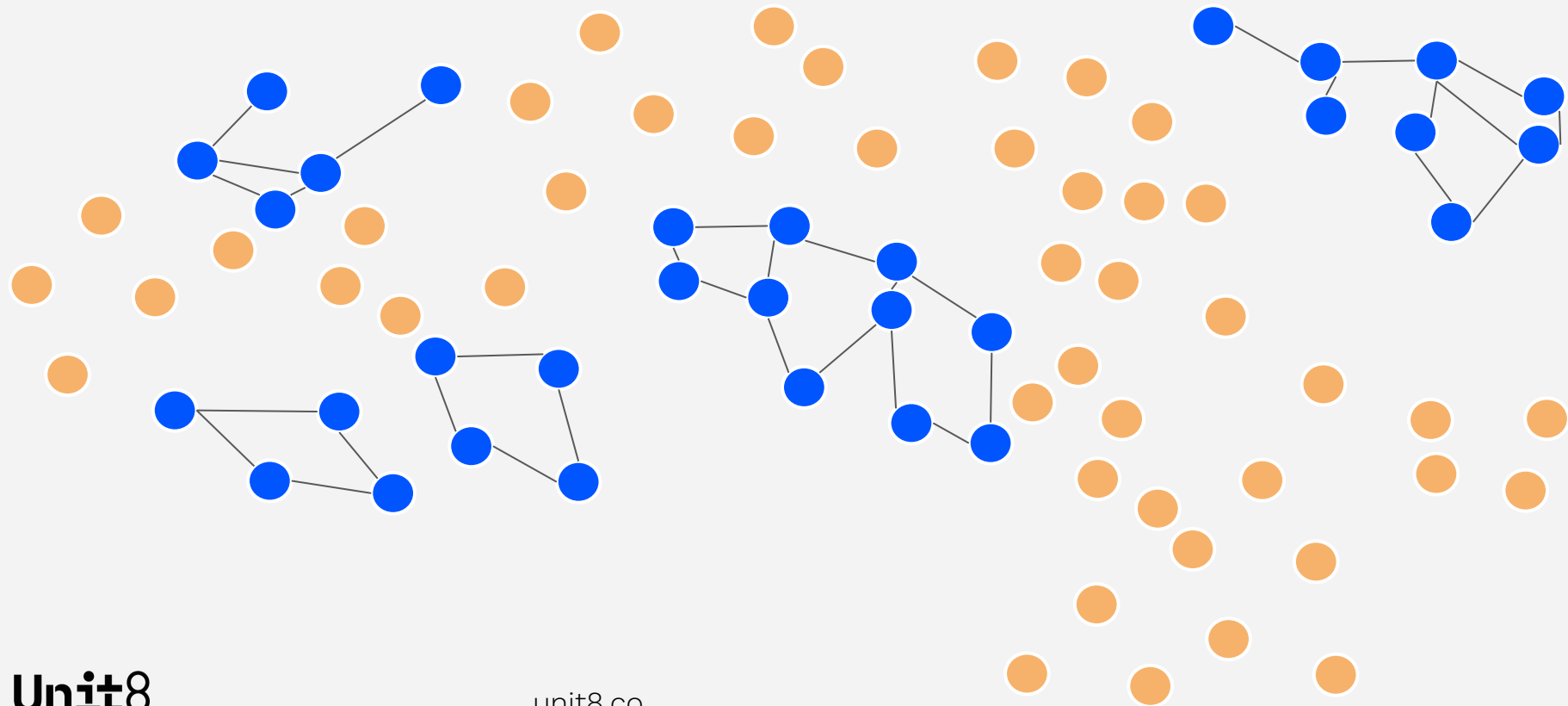
From chaos to structure



From chaos to structure



From chaos to structure



What are the applications of Entity Resolution ?

Know Your Customer

Banking Industry

Goal: Don't do business with people on sanction lists

Sanction Lists

- Terrorism and Terrorist financing
- Narcotics trafficking
- Human rights violations
- Weapons proliferation
- Violation of international treaties, e.g arms embargo
- Money laundering activities



Single Client View

Banking, Insurance, etc.

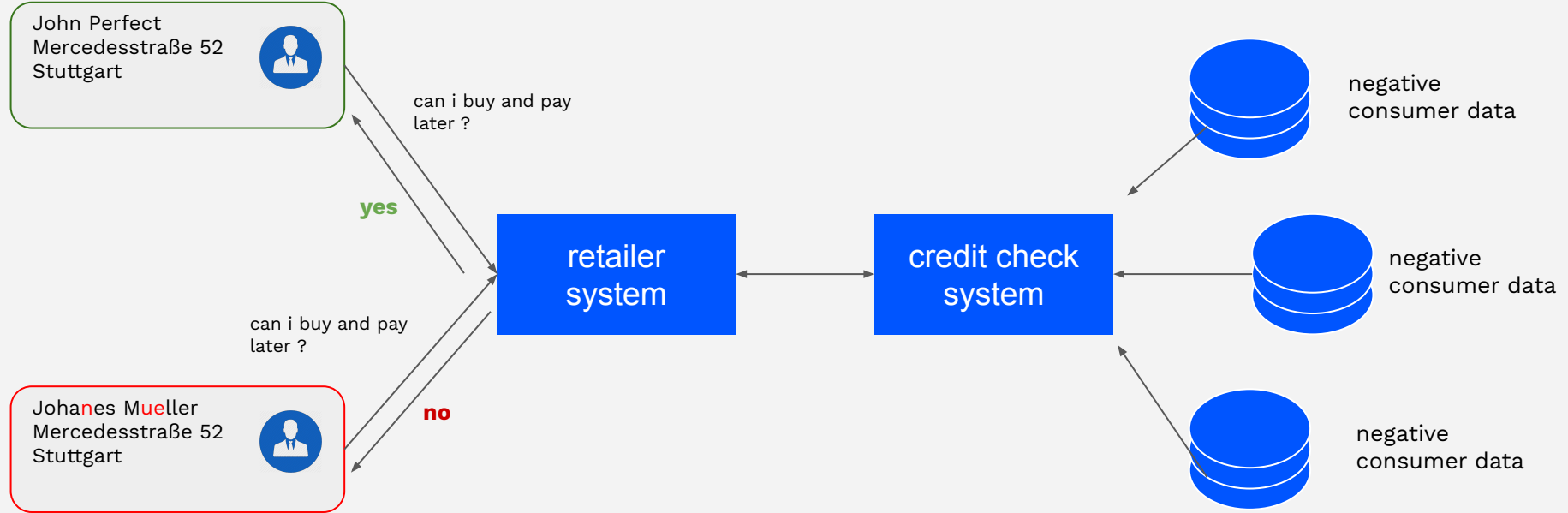
Goal: Have a single view on the client no matter where and when they register



Consumer Credit Check

Telecoms, Retailers etc.

Goal: Reduce losses caused by unreliable clients



Crime Investigation and Prevention

Law enforcement agencies

Goal: Catch the bad guy

FBI uses Entity Resolution to get holistic view of the criminals and their interactions from dispersed data sets



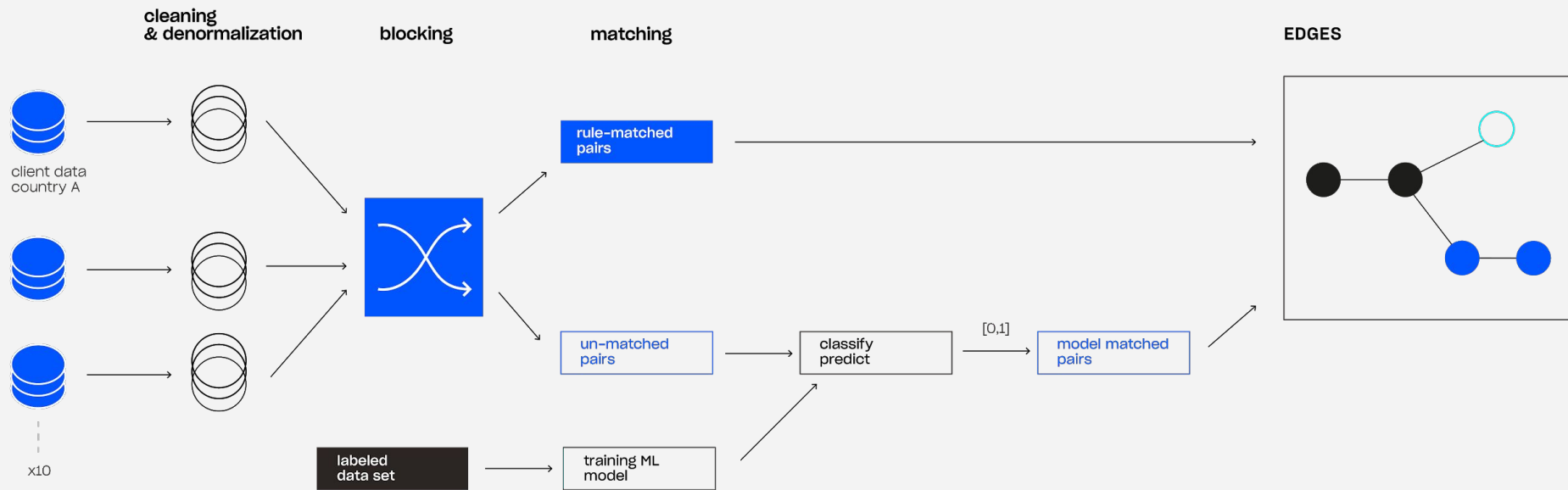
How is it done ?

"Naive" approach

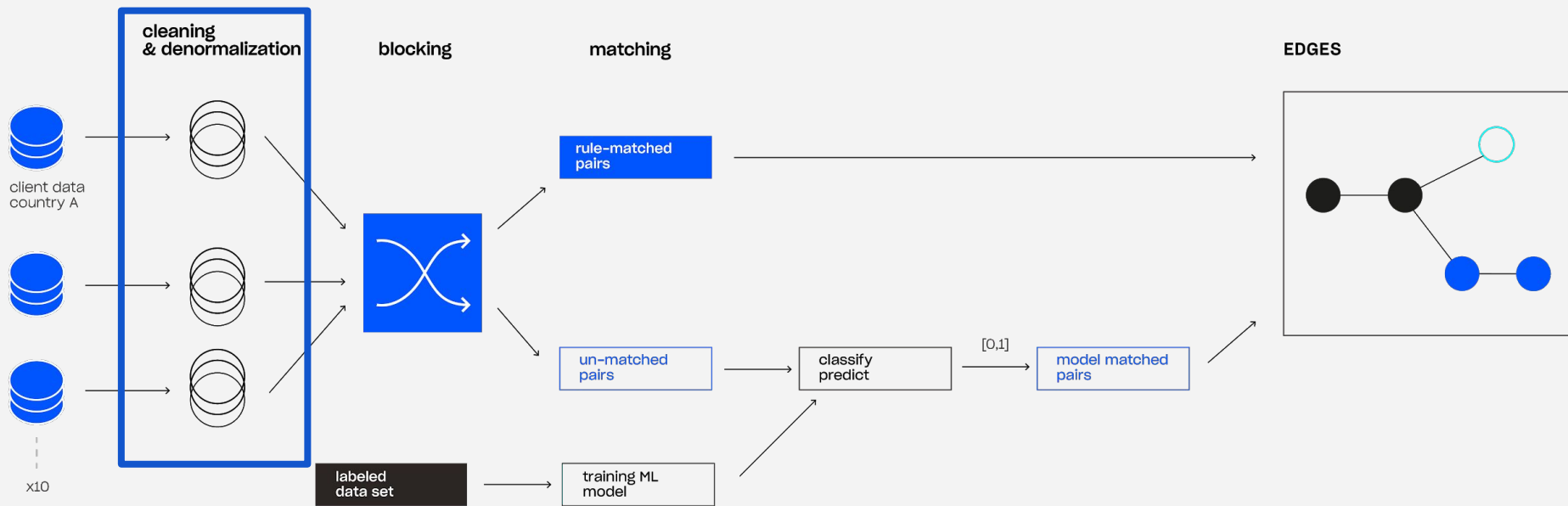
	Person A	Person B	Person C	Person D
Person A		0.9	1	0.3
Person B	0.9		0.1	1
Person C	1	0.1		0
Person D	0.3	1	0	

- n^2 problem - **comparing all records is impossible**
- 1,000,000 records equals 1000 billion comparisons
- assuming 1ms per comparison we need more than 15 years of compute time

How is it done?



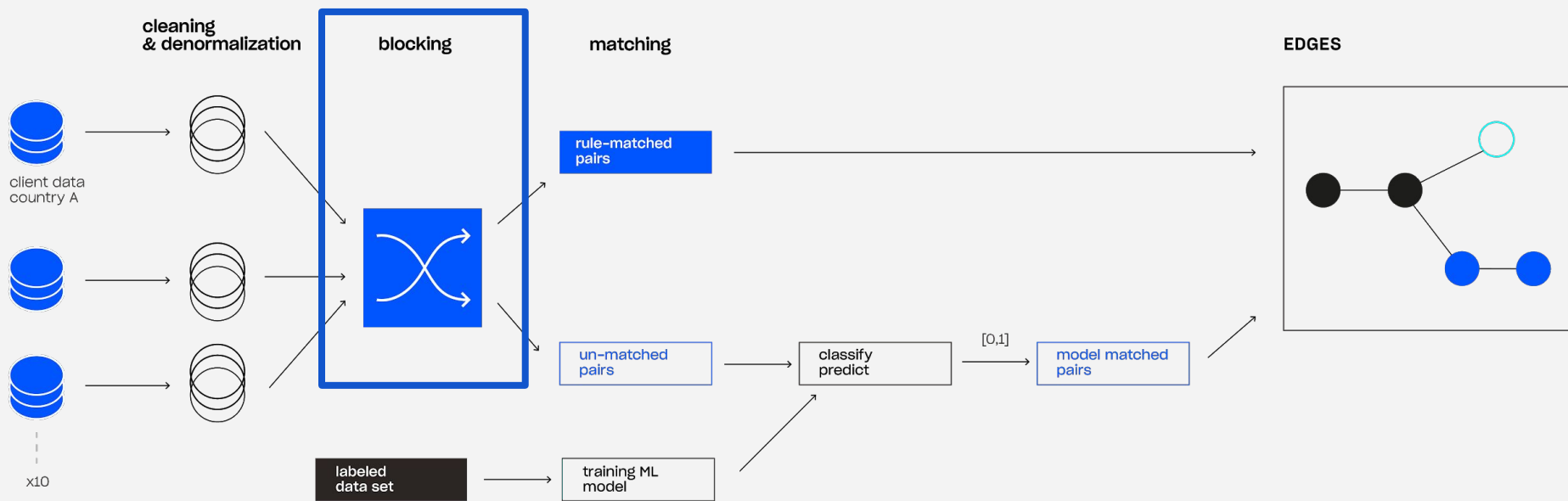
Cleaning



Cleaning/Normalization

- normalizing cases: Johan -> JOHAN
- normalizing non-ascii characters: Ü -> UE, ß -> SS, Ó -> O
- abbreviations: Aktiengesellschaft -> AG, Gesellschaft mit beschränkter Haftung -> GmbH
- normalizing addresses
 - deriving zips from streets
 - normalizing against the database of official addresses
 - von Bismarck Allee, Berlin -> Otto-von-Bismarck-Allee, Berlin

Blocking



Blocking

What is it?

Blocking is a way of reducing number of comparisons by pre-selecting records that have high potential of matching.

Our algorithm should have high recall (minimal amount of false negatives)

Pairs missed in this phase will not appear later

Blocking

ZIP + Last Name

First Name	Last Name	City	Zip	Birth date
Johan	Mueller	Stuttgart	70713	
Johan	Mueller	Stuttgart	70713	
Mark	Schmidt	New York		1982-12-17
Mark	Smith	Stuttgart	70713	1982-12-00

Blocking

ZIP + Last Name Soundex

First Name	Last Name Soundex	City	Zip	Birth date
Johan	Mueller M460	Stuttgart	70713	
Johan	Muler M460	Stuttgart	70713	
Mark	Schmidt S530	New York		1982-12-17
Mark	Smith S530	Stuttgart	70713	1982-12-00

Blocking

Birth Month + Last Name Double Metaphone

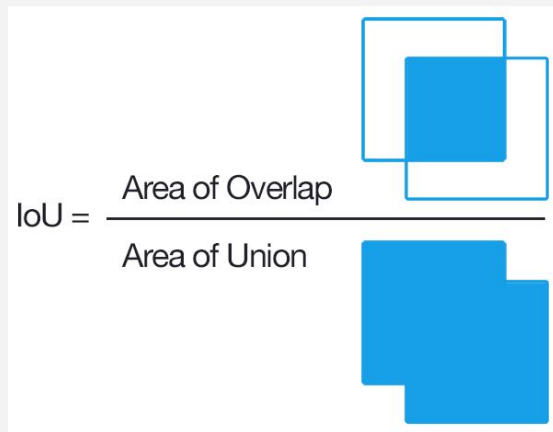
First Name	Last Name Double Metaphone	City	Zip	Birth date
Johan	Mueller	Stuttgart	70713	
Johan	Muler	Stuttgart	70713	
Mark	Schmidt SM0+ XMT	New York		1982-12-17
Mark	Smith XMT +SMT	Stuttgart	70713	1982-12-00

Blocking

Is there a way to quickly find names that have similarity above certain threshold?

LSH MinHash Jaccard Index


$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



LSH MinHash


Shingles aka ngrams

MUELLER



MU, UE, EL, LL, LE, ER

MUELER



MU, UE, EL, LE, ER

LSH MinHash Jaccard Index

MUELLER



MU, UE, EL, LL, LE, ER

MUELER



MU, UE, EL, LE, ER

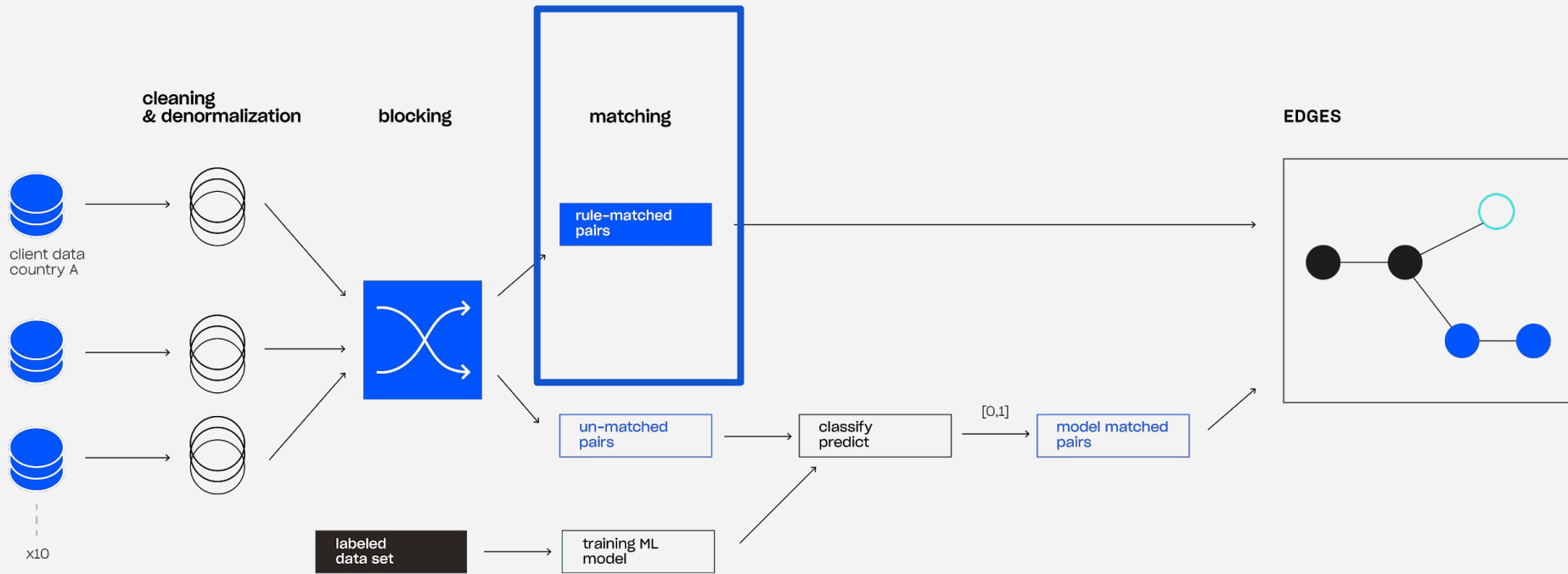
$$\frac{5}{6} \equiv 0.83$$

LSH MinHash

is an algorithmic technique that hashes similar input items into the same "buckets" with high probability

Word	Signatures
MUELLER	[11][20] [23]
MUELER	[11][20] [25]
MAYER	[11][28][40]

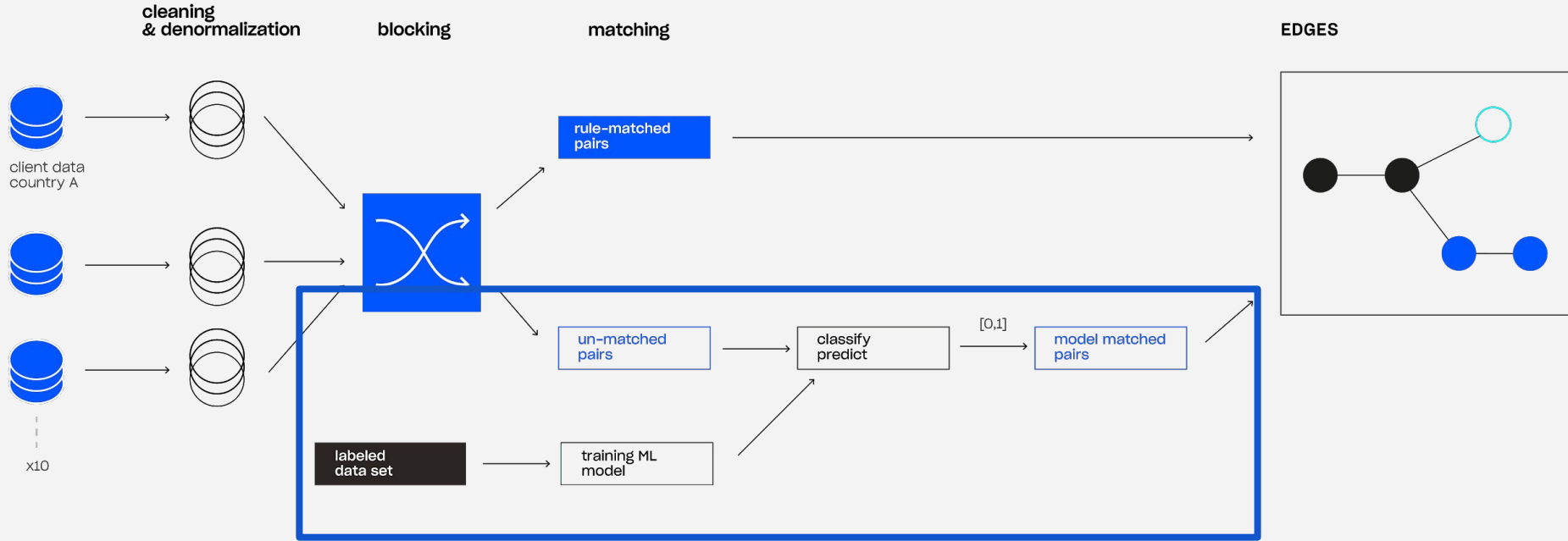
Rule matching



Rule matching

- Set of rules indicating positive match
 - exact name and exact date of birth
 - custom information file relation and similar name
- Covers most of the matches in practice
- Those matches usually have best quality but they need to be manually introduced and maintained

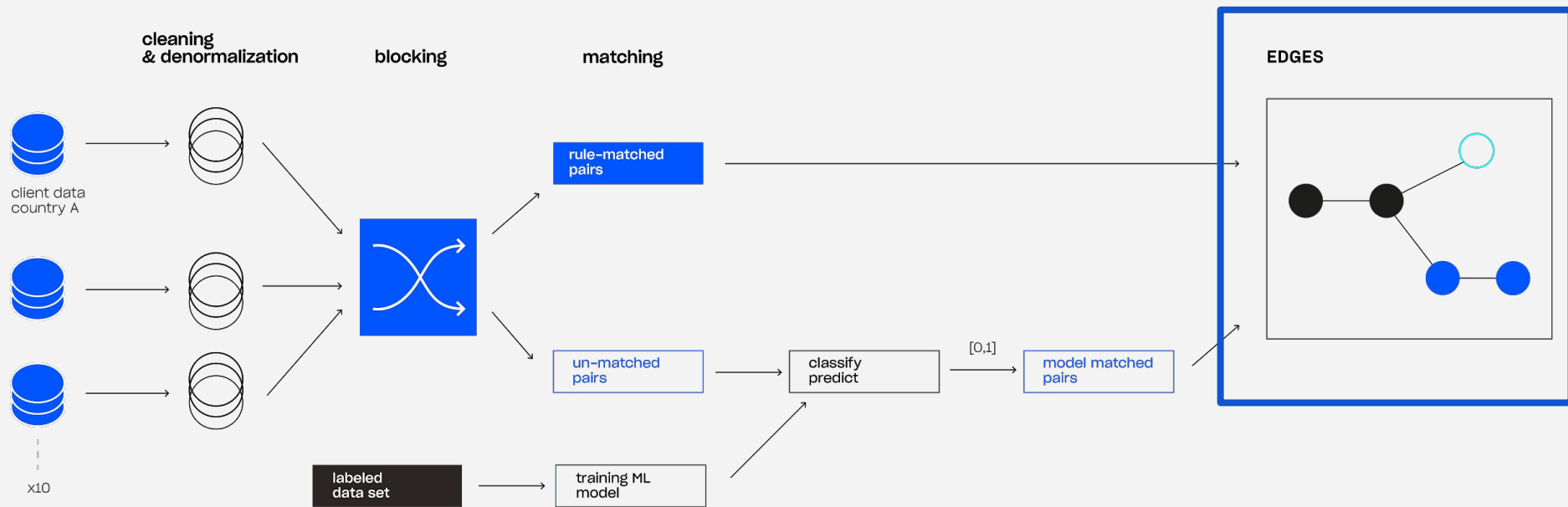
Fuzzy matching



Fuzzy matching

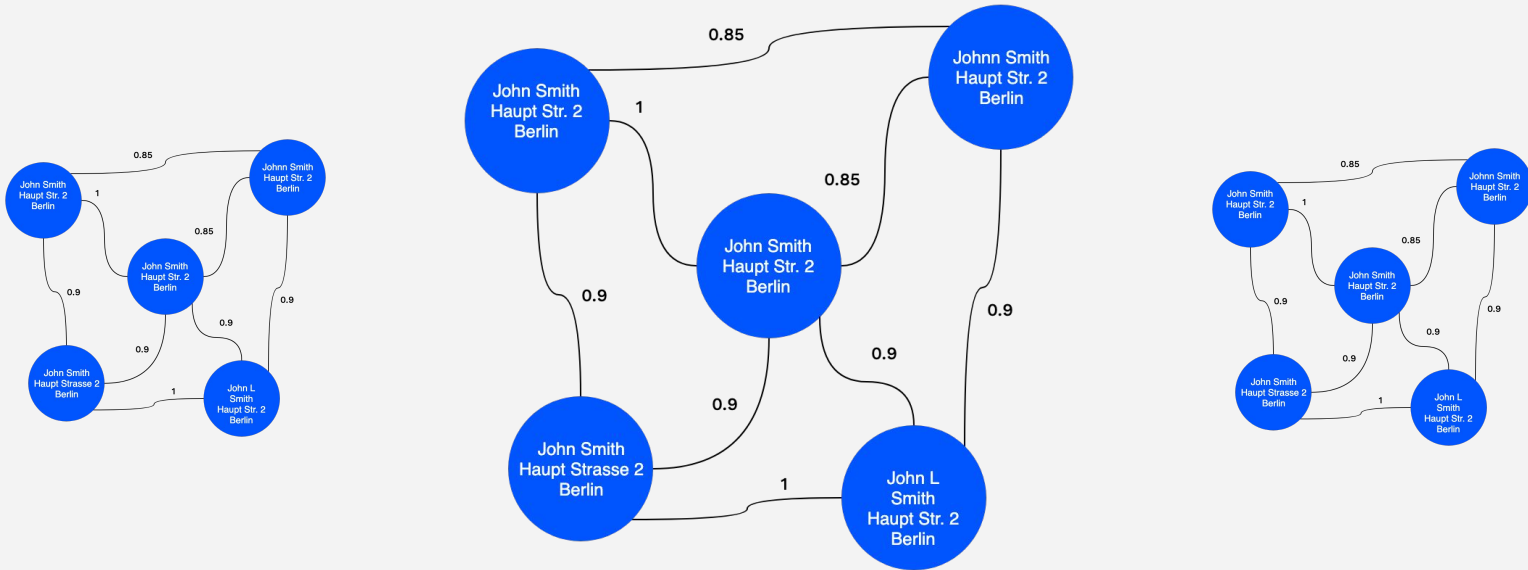
- Training data
 - Manually labeled training records
- Features:
 - Name Similarity
 - Levenshtein
 - Jaro-Winkler
 - Jaccard
 - Soundex/Metaphone/Double Metaphone
 - birthday match
 - phone, email etc match
 - location match (same house, same street, same city, same country)
 - size of household
 - others that depend on the context
- Supervised Learning Problem - our choice - Random Forest

Results



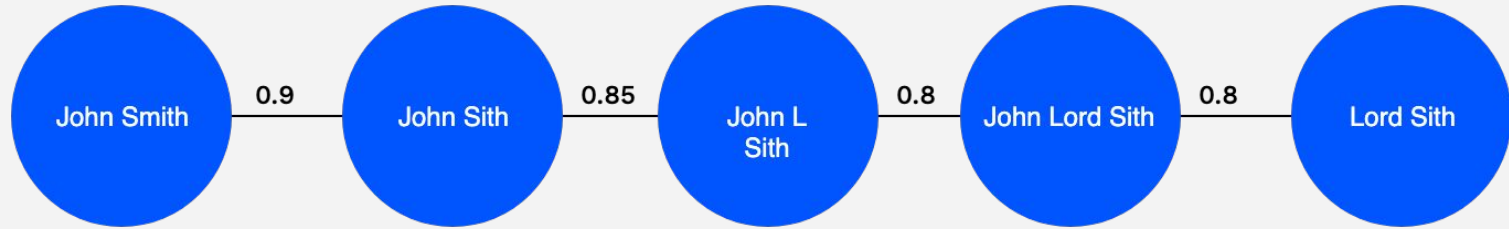
Results

- Each compared pair receives a similarity measure
- Pairs above certain level are linked and form a cluster



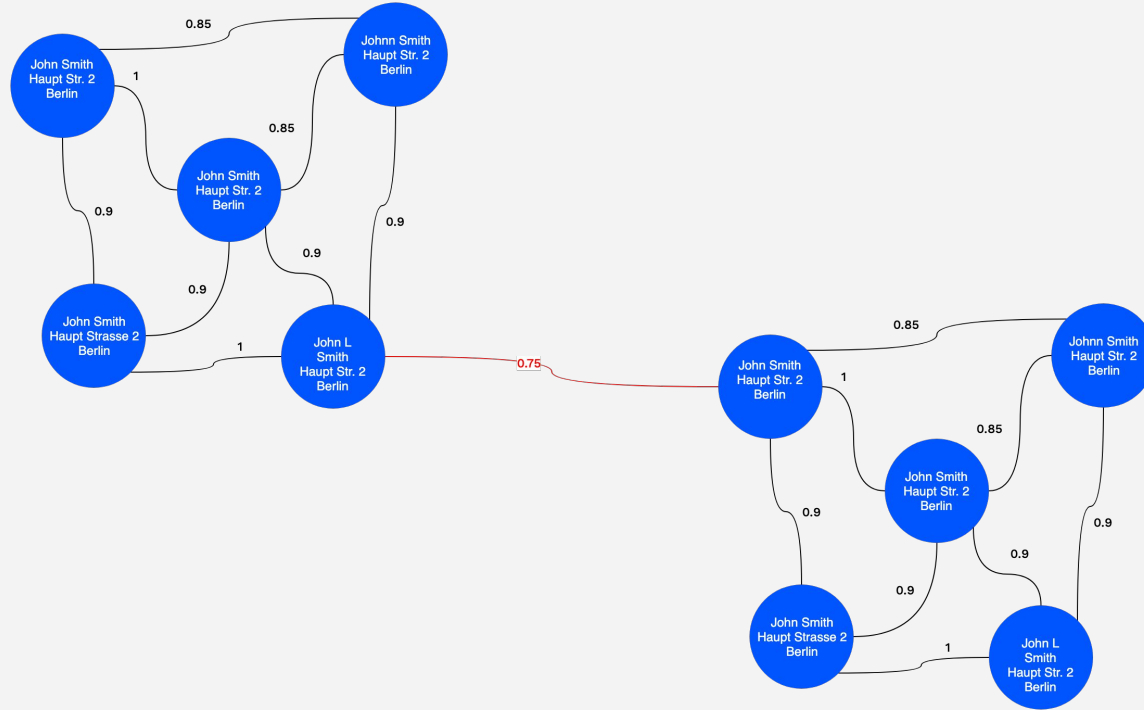
Results

What can go wrong ?



Results

What can go wrong ?



Results

Going further

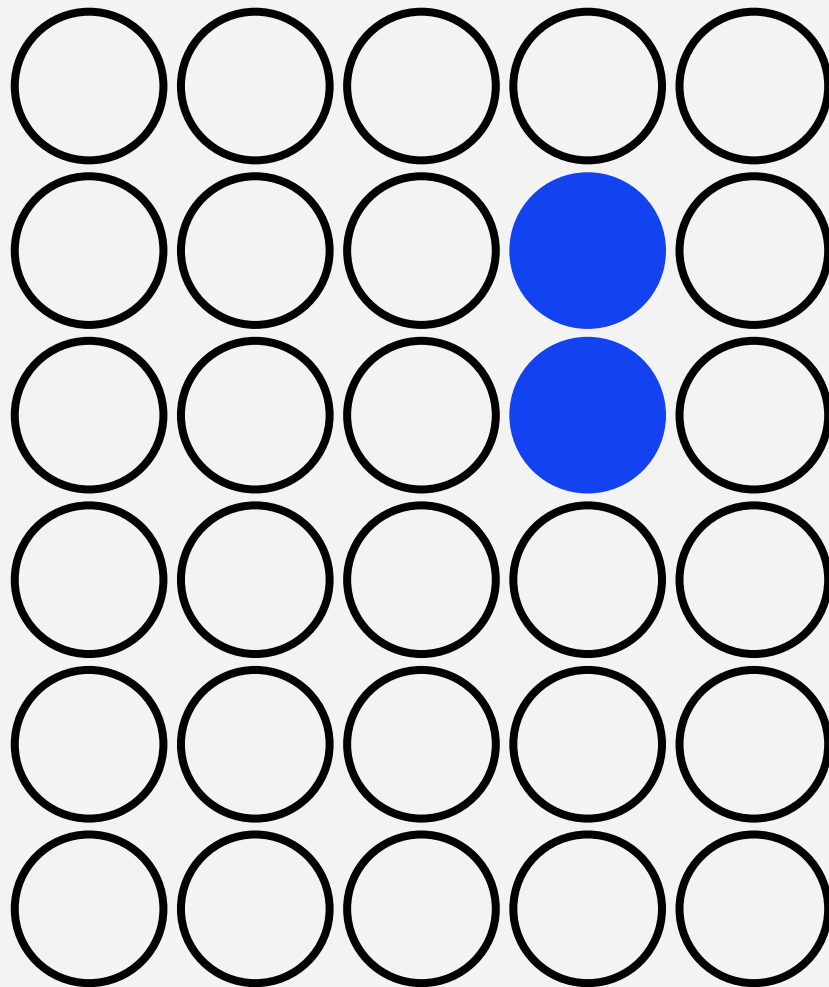
- Historization - how to keep track of clusters ?
- What if the "bridge" record suddenly arrives?
- Building clusters of clusters - eg. family members, household, child-parent company

Conclusions

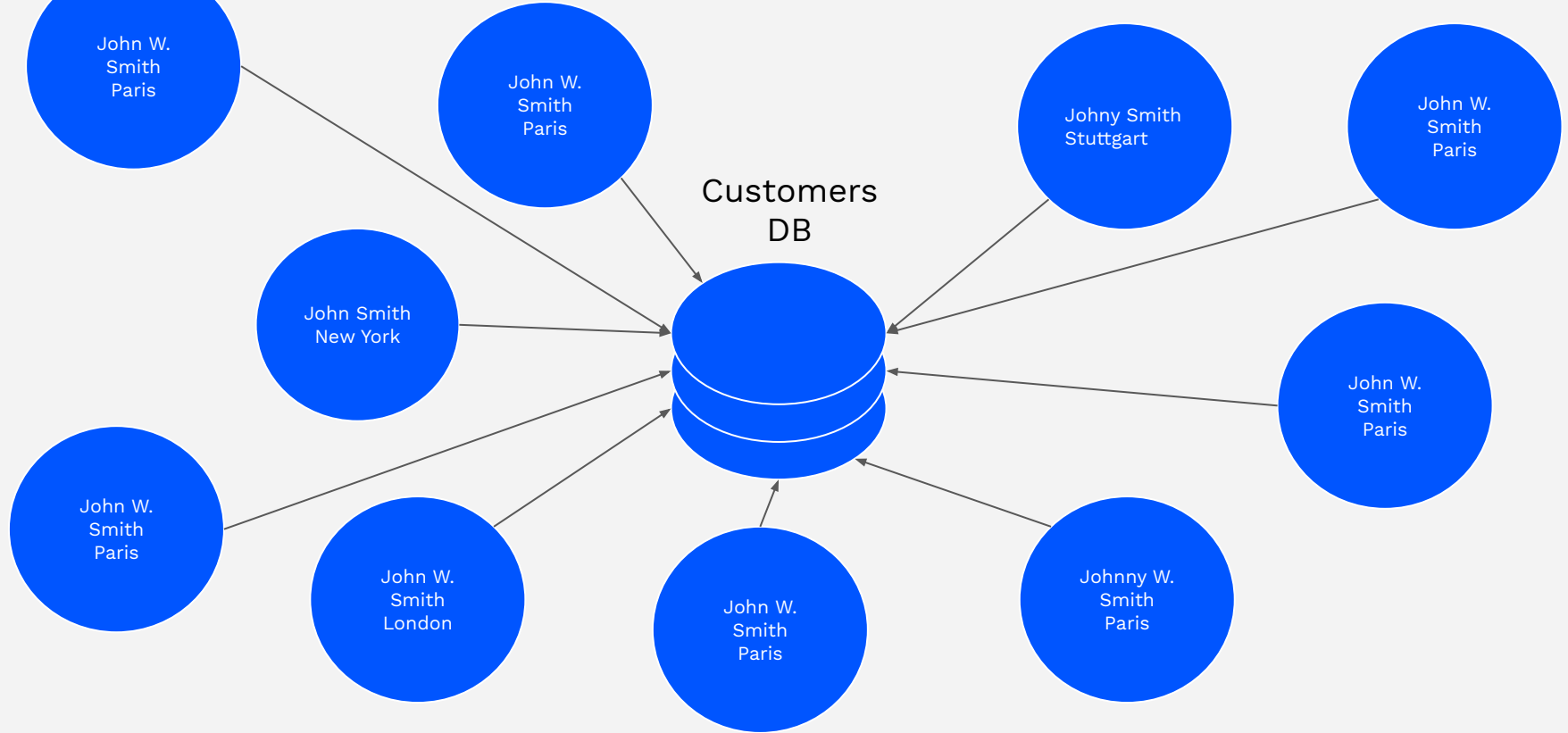
- ER is applicable in multiple many domains and industries
- The right implementation depends on the context
 - It may be ok to **incorrectly classify person as unreliable** and ask them to do a pre payment instead of post payment
 - It is probably not ok classify person as a terrorist by mistake

unit8.co

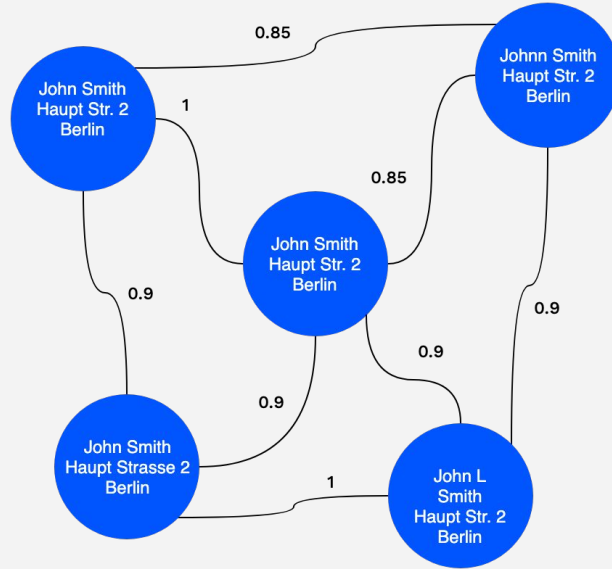
**thank
you**



From chaos to structure



From chaos to structure



Blocking

LSH MinHash

Is there a way to quickly find names that have similarity above certain threshold?

LSH MinHash Jaccard Index

MUELLER

MUELER

MU, UE, EL