

JOHN DOE ET JHON DOE SONT DANS UN BATEAU

ENTITY RESOLUTION DEMYSTIFIED

QUICK WORD OF INTRO

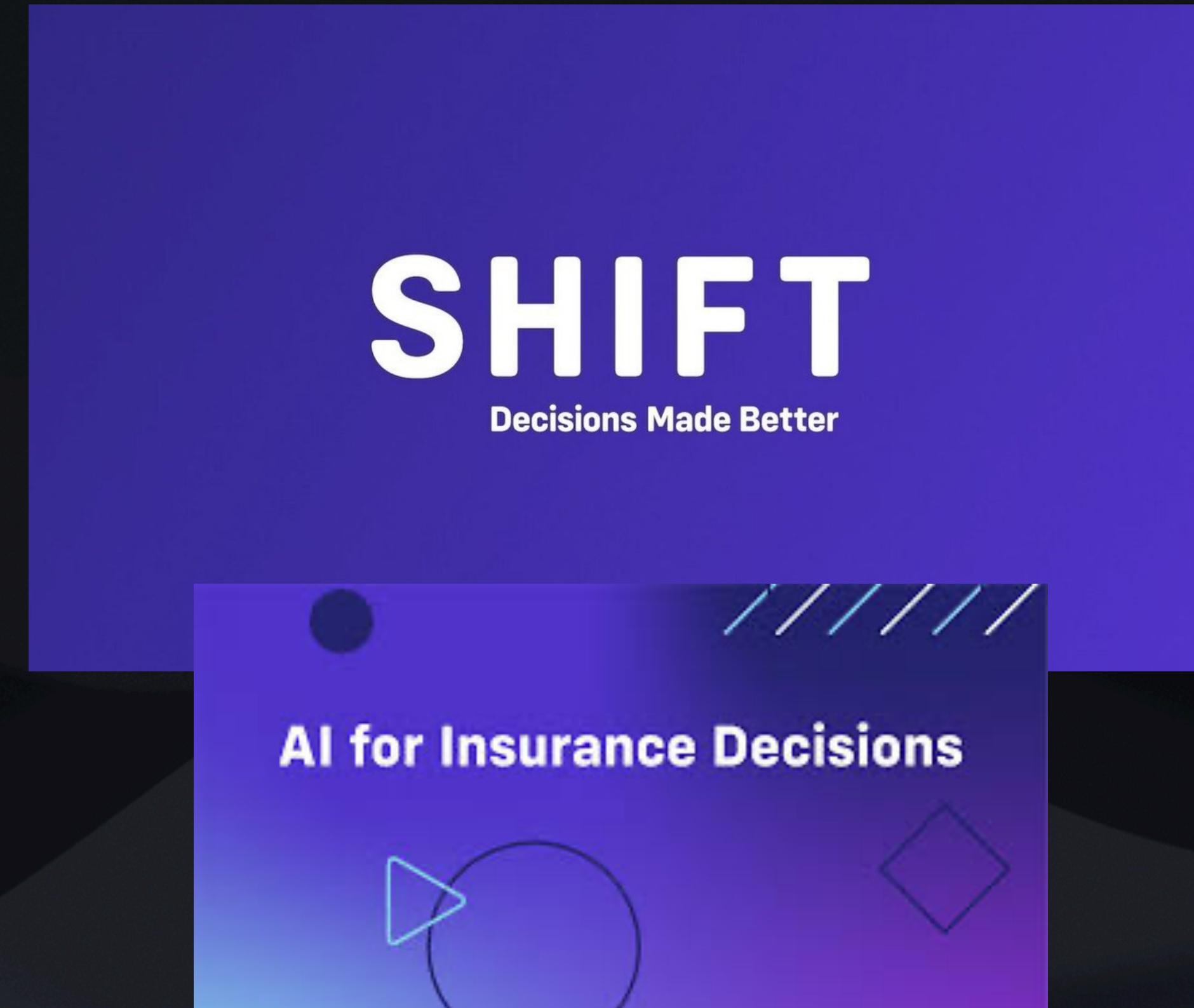
Arnault Estève

Arnaut Esteves

Arnauld Estevez

Arnaud Esteve

Data Scientist



A COMMON TASK

MATCHING ENTITIES THAT LOOK DIFFERENT BUT ARE THE SAME

Consolidation

Householding

Data Harmonisation

Object reconciliation

Cross Linking

Record Linkage

Reference reconciliation

Identity Resolution

Data Unification

Deduplication

Identity Reconciliation

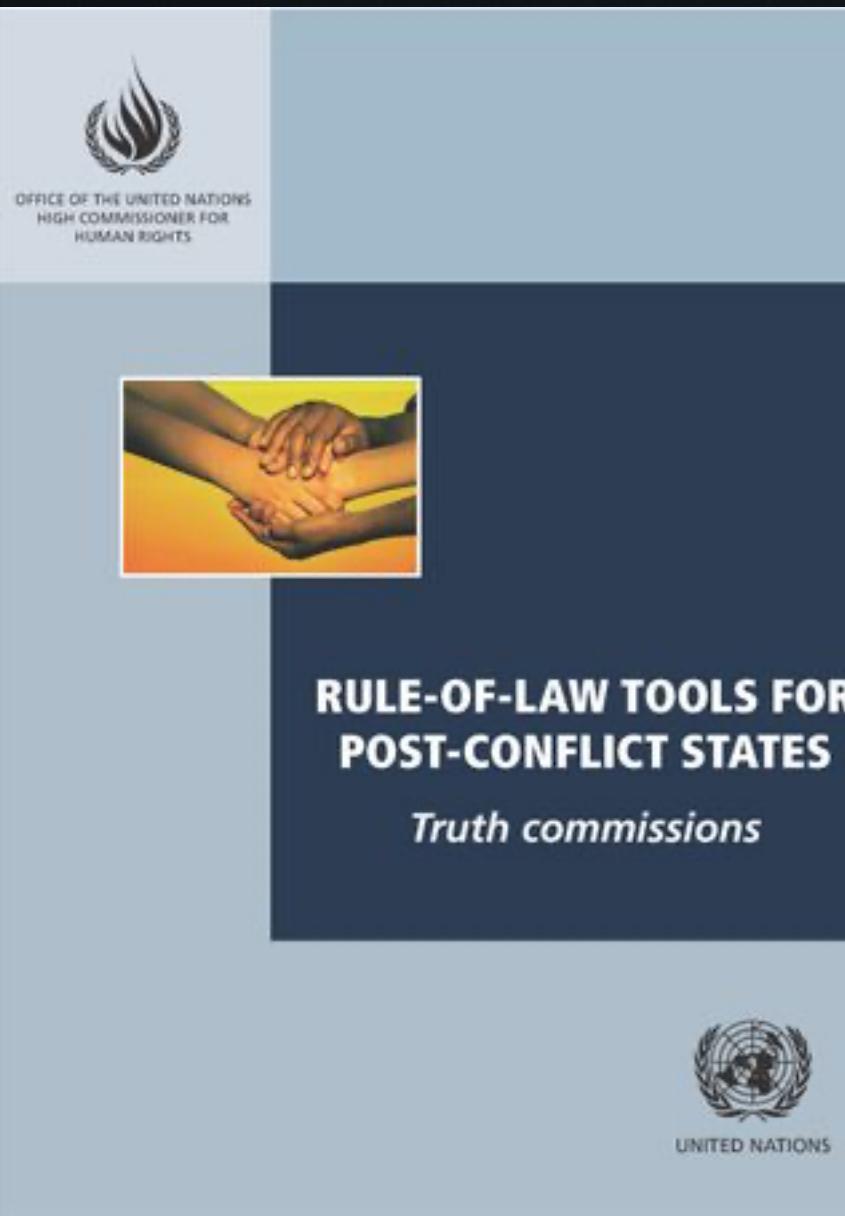
Fuzzy Matching

Data Matching

Entity Matching

Master Data Management

In real life



OFFICE OF THE UNITED NATIONS HIGH COMMISSIONER FOR HUMAN RIGHTS



RULE-OF-LAW TOOLS FOR POST-CONFLICT STATES

Truth commissions

UNITED NATIONS

A report cover from the Office of the United Nations High Commissioner for Human Rights. It features the UNHCR logo at the top left, followed by the title "RUL-E-OF-LAW TOOLS FOR POST-CONFLICT STATES" and the subtitle "Truth commissions". Below the title is a small image of two hands holding each other. At the bottom is the United Nations logo.

Credit: Geoff Thale and Adriana Beltran

Did we solve it well, though?

“We relied on email”

“We did it best effort”

If ... and if ... and if ... or if ... and if ... or if ...

Can we do it differently?

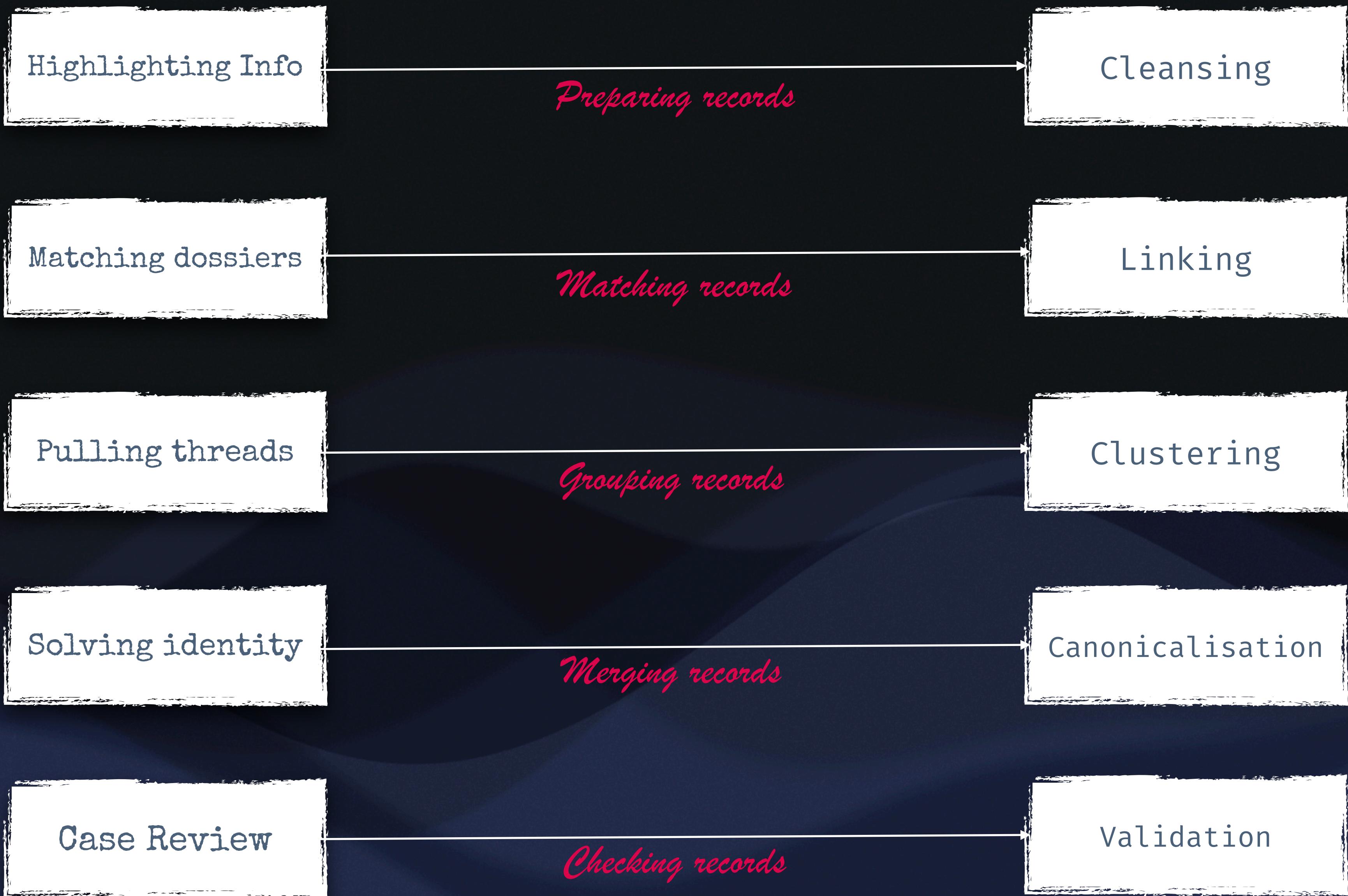
A PRIVATE INVESTIGATOR STORY

The “human way”

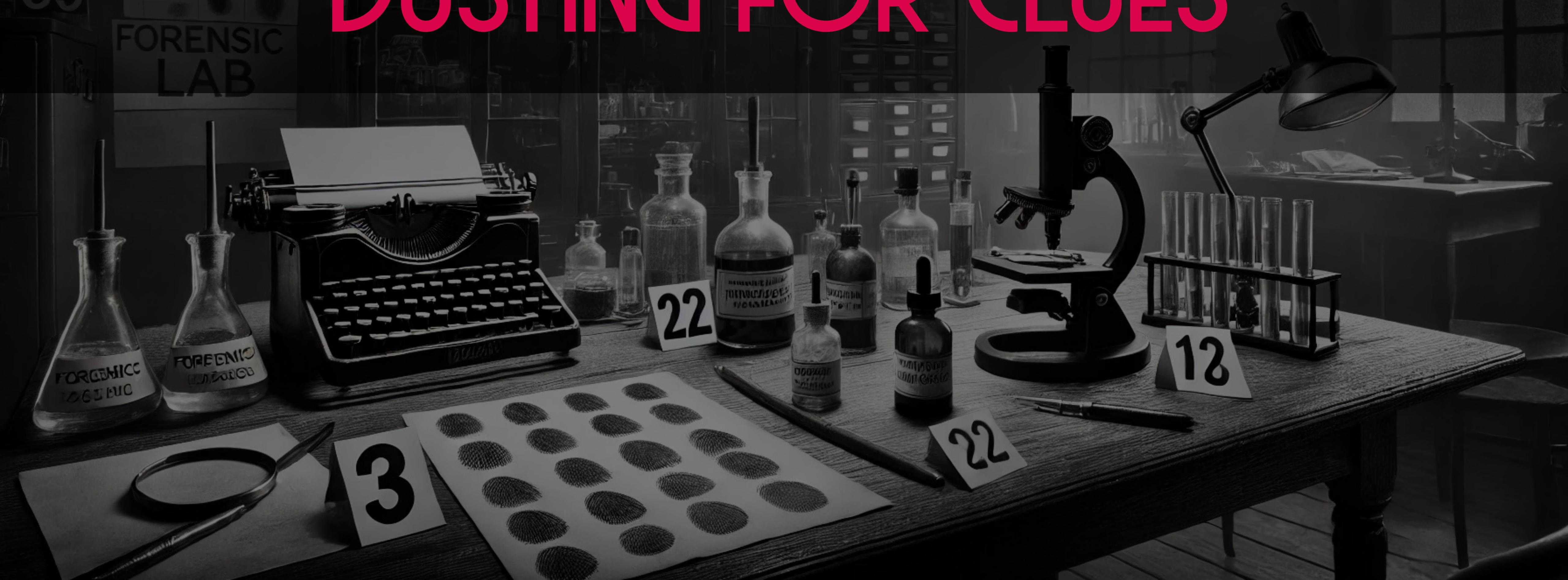
vs.

The “programmatic way”





CLEANSING: DUSTING FOR CLUES



WHY IS DATA MESSY?

- Missing data
- Format **inconsistencies** and variations: 1923-02-01 vs. 02/01/23
- **Alternatives**: Katherine ↔ Kate . Junior ↔ Jr.
- **Typos** & non-uniform characters
- **OCR** / Digitalisation mistakes
- Phonetic Mistakes

CLEANING DATA

- **Removing noise:** <HTML />, ...
- **Normalisation:** latin, UPPER, ~~diàcritics~~.
- **Parsing:** extract more information
- **Encoding:** phonetic, geo
- **Detecting:** outliers / special (eg. 1970-01-01)
- **Variations:** William <=> Bill

*But there's more:
Make data computable*



PHONETIC ENCODING

```
from jellyfish import match_rating_codex, metaphone,  
nysiis, soundex  
name = "Lucía"  
soundex_code = soundex(name)  
metaphone_code = metaphone(name)  
nysiis_code = nysiis(name)  
match_rating_codex_code = match_rating_codex(name)
```

Alternate spelling

Original	Mohamed	Muhammad
Soundex	M530	M530
Metaphone	MHMT	MHMT
NYSIIS	MAHANAD	MAHANAD
Match Rating Codex	MHMD	MHMD

Alternate spelling + typo

Original	Lucía	Lizía
Soundex	L200	L200
Metaphone	LS	LS
NYSIIS	LACÍ	LASÍ
Match Rating Codex	LCÍ	LZÍ

GEO-CODING + GEO-HASHING

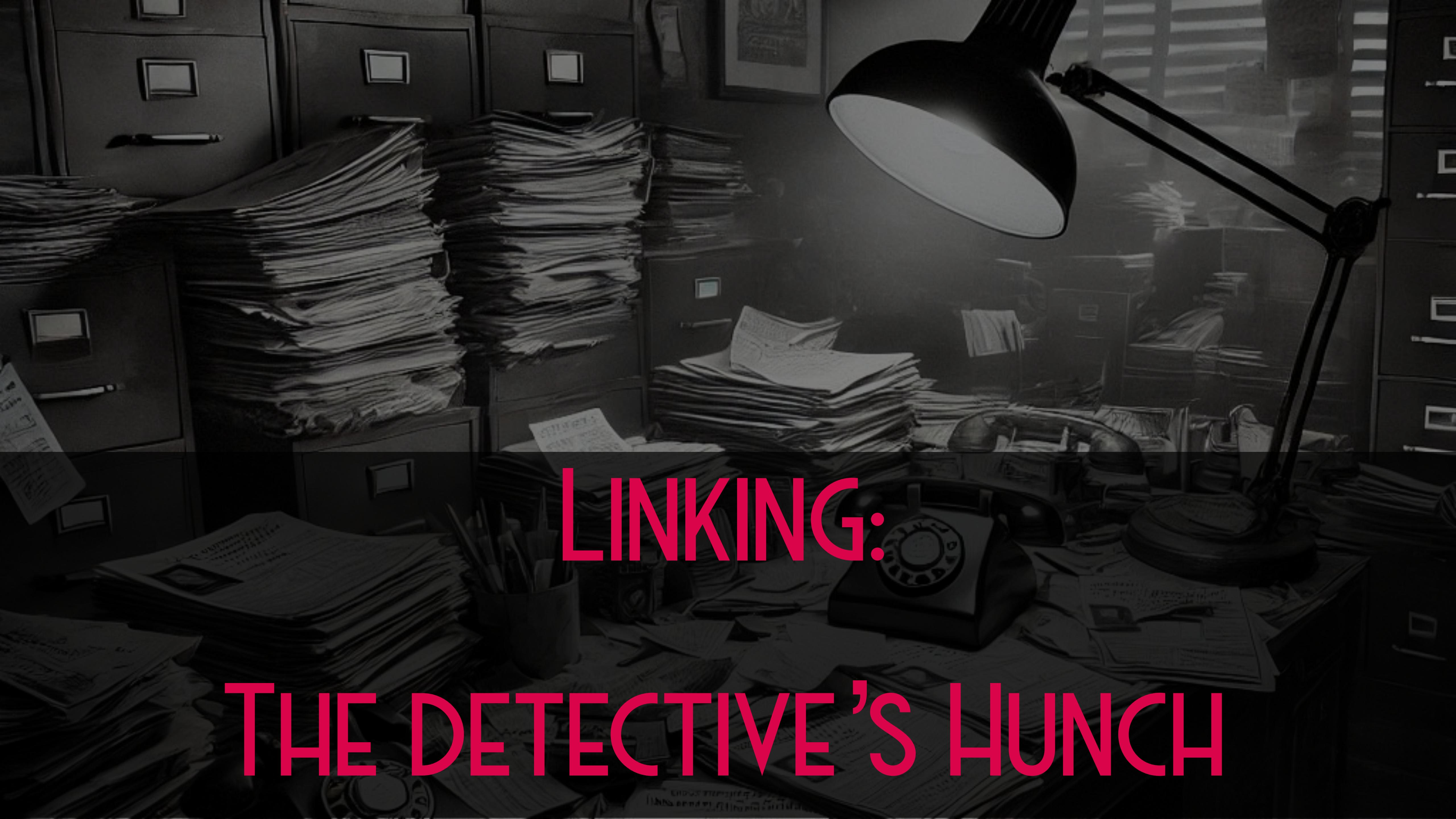
GET [https://nominatim.openstreetmap.org/search?
q=palais%20congrès%20maillot&format=json&addressdetails=1](https://nominatim.openstreetmap.org/search?q=palais%20congrès%20maillot&format=json&addressdetails=1)

```
{  
    "place_id": 89127893,  
    "licence": "Data © OpenStreetMap contributors, ODbL 1.0. http://osm.org/copyright",  
    "osm_type": "node",  
    "osm_id": 3969536957,  
    "lat": "48.8784493", encoded,  
    "lon": "2.2837635",  
    ...  
    "address": {  
        "railway": "Palais des Congrès",  
        "road": "Place de la Porte Maillot",  
        "city_block": "Quartier des Ternes",  
        "suburb": "17th Arrondissement",  
        "city_district": "Paris",  
        "city": "Paris",  
        "ISO3166-2-lvl6": "FR-75C",  
        "region": "Metropolitan France",  
        "postcode": "75017",  
        "country": "France",  
        "country_code": "fr"  
    },  
    ...  
}  
  
parsed & validated  
  
geohash.encode(latitude=48.8784493, longitude=2.2837635)  
u09w5fncxe37
```

GET [https://nominatim.openstreetmap.org/search?
q=82%20Boulevard%20Pereire&format=json&addressdetails=1](https://nominatim.openstreetmap.org/search?q=82%20Boulevard%20Pereire&format=json&addressdetails=1)

```
{  
    "place_id": 89276492,  
    "licence": "Data © OpenStreetMap contributors, ODbL 1.0. http://osm.org/copyright",  
    "osm_type": "node",  
    "osm_id": 8669896139,  
    "lat": "48.8871762", encoded,  
    "lon": "2.3042596",  
    ...  
    "address": {  
        "amenity": "Etoile - Wagram",  
        "road": "Boulevard Pereire",  
        "city_block": "Quartier de la Plaine-de-Monceau",  
        "suburb": "17th Arrondissement",  
        "city_district": "Paris",  
        "city": "Paris",  
        "ISO3166-2-lvl6": "FR-75C",  
        "region": "Metropolitan France",  
        "postcode": "75017",  
        "country": "France",  
        "country_code": "fr"  
    },  
    ...  
}  
  
parsed & validated  
  
geohash.encode(latitude=48.8871762, longitude=2.3042596)  
u09wh7tumnhm
```

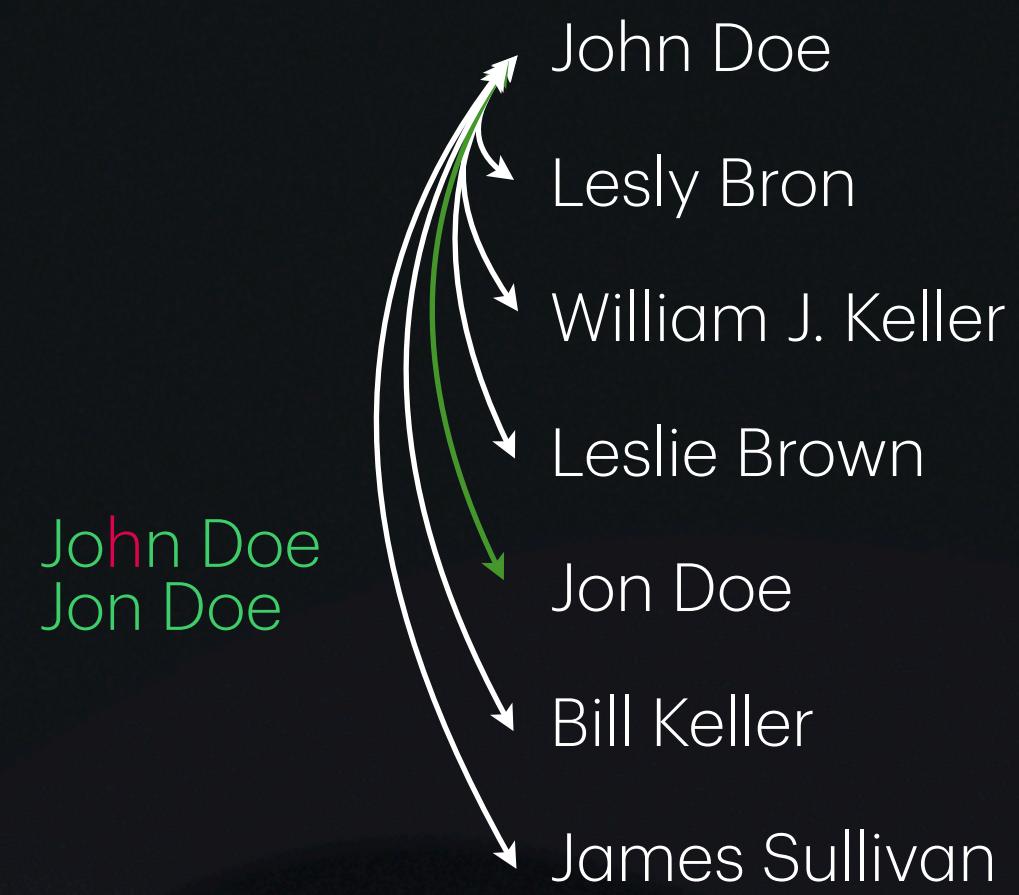
"palais des Congrès maillot"	u09w5fncxe37
"82 Bd Pereire"	u09wh7tumnhm
"Tour Eiffel, Paris"	u09tuny9c3wb
"Gare des Bénédictins, Limoges"	u00uuub4ztwv4



LINKING: THE DETECTIVE'S HUNCH

HUMAN APPROACH

COMPARING PAIRS



Problem:
Quadratic Explosion

BLOCKING

TACKLING QUADRATIC EXPLOSION

- Compare records w/in blocks **ONLY**
- Overlap blocks!
 - phonetic last name + same year
 - `substr(geohash(address), 4)`



HOW DO WE COMPARE?

WE NEED TO BE FUZZY: “HOW FAR ARE THESE?”

- Numeric:
 - usually easy (subtract)
 - think normalisation
- Strings:
 - edit distance
 - there are many!

Make data comparative



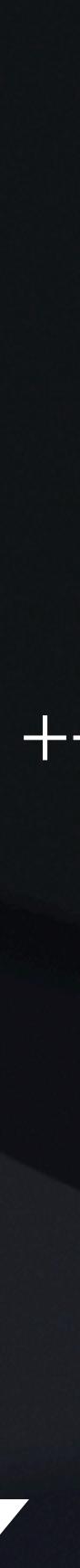
CHRSISTOPHAR / CHRISTOPHER

Levenshtein = 0.63

Damerau-Levenshtein = 0.727

Jaro = 0.776

Jaro-Winkler = 0.843

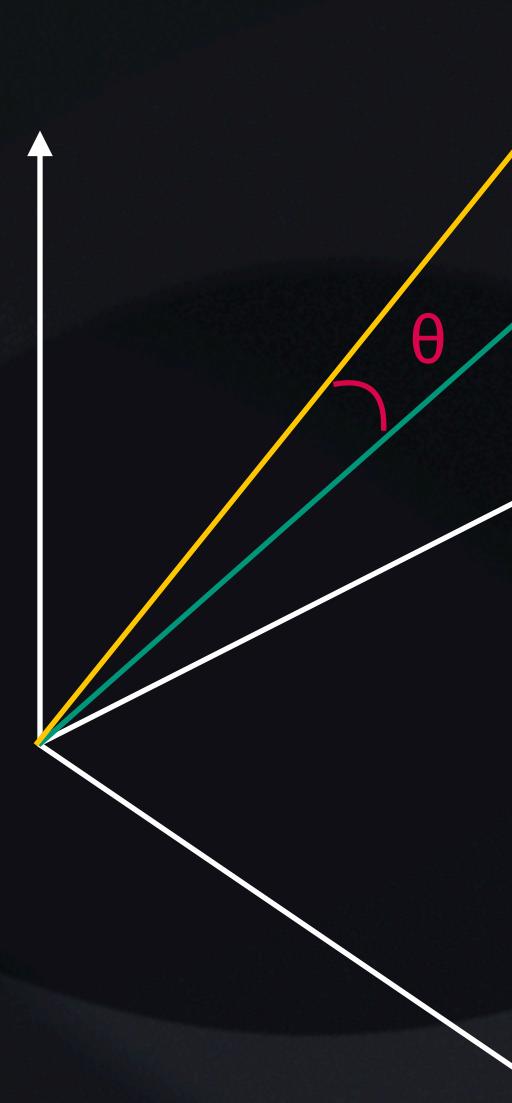


DOMAIN SPECIFIC

https://github.com/easonanalytica/company_name_matcher

```
from company_name_matcher import CompanyNameMatcher  
  
matcher = CompanyNameMatcher("sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2")  
similarity = matcher.compare_companies("MERCK & CO", "MERCK AND COMPANY")  
print(f"Similarity: {similarity}") # 0.903 ...
```

text-embedding



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}}$$

Cosine Similarity

HOW DO WE USE DISTANCE METRICS?

if $jw(first_name) > 0.9$ and $lev(dob) > 0.9$

or if $jw(street_name) > 0.8$ and $jw(last_name) > 0.9$

or if $cos(company_name) > 0.7$ and $lev(phone) > 0.8$

Problem:
Combinatorial Explosion

Do you trust more?

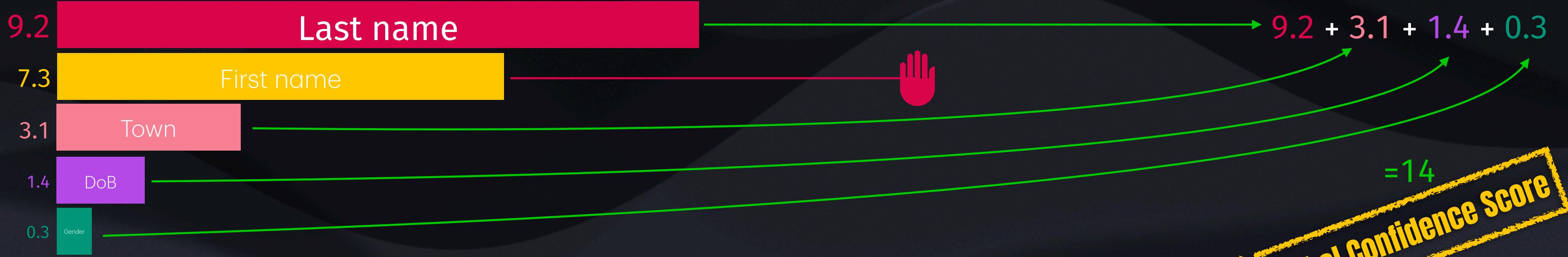
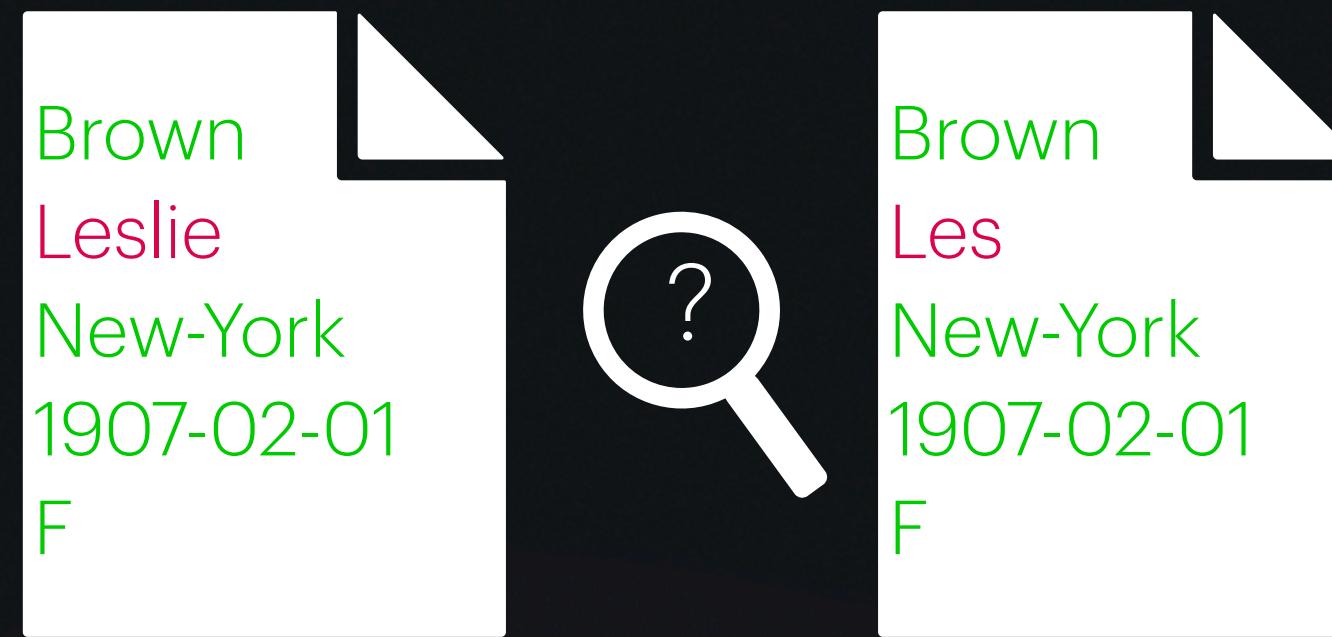
Same first name, same date of birth

Or

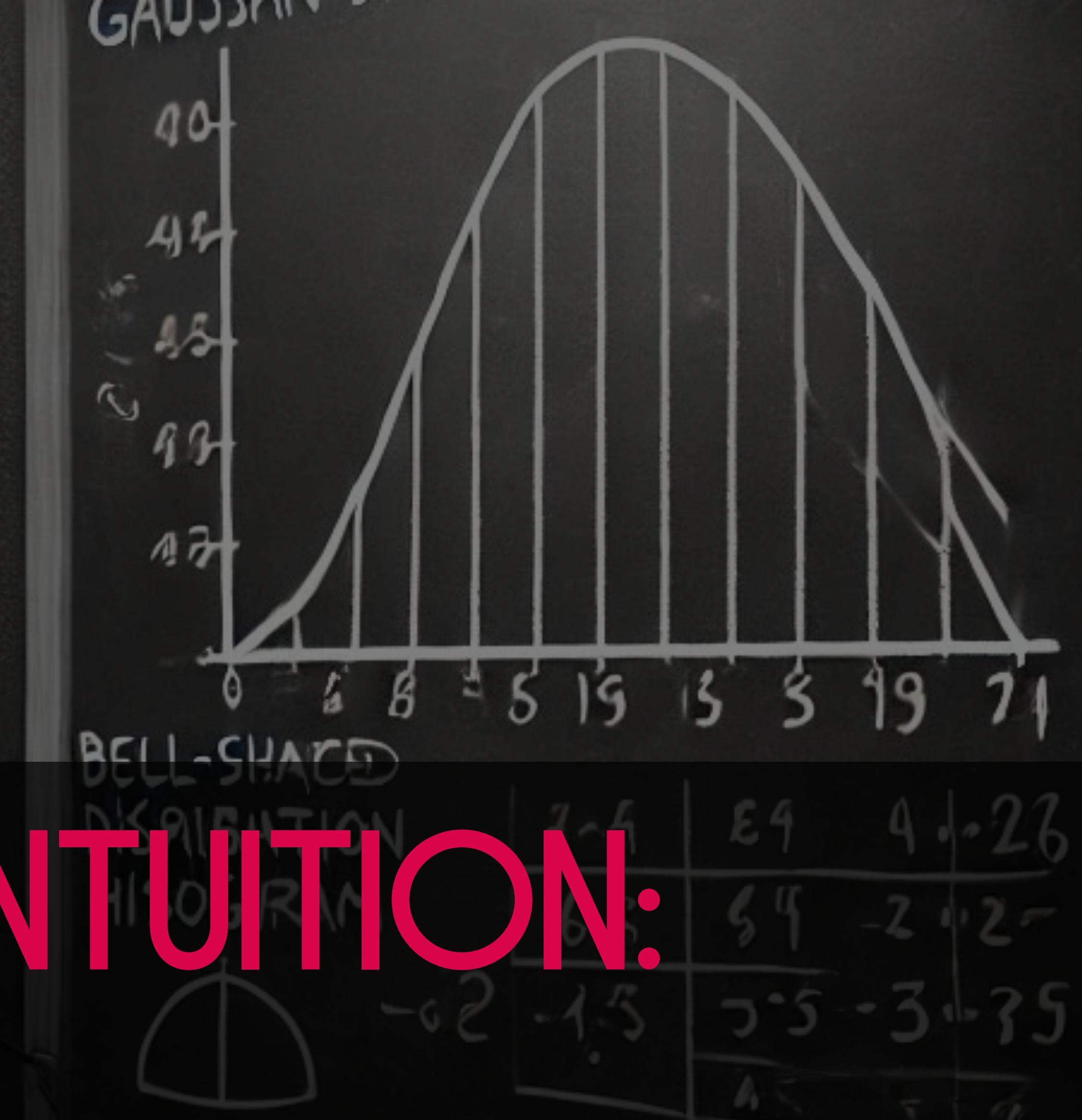
Close full name, same date of birth

We need
CONFIDENCE + WEIGHTS

QUANTIFYING CONFIDENCE



BEYOND INTUITION: EMBRACING PROBABILITIES



WHAT IS A WEIGHT?

How **likely** are 2 records to match **given** their last name matches?

Bayesian probability [...] is an interpretation of the concept of probability, in which, instead of frequency or propensity of some phenomenon, probability is interpreted as reasonable expectation^[2] representing a state of knowledge^[3] or as quantification of a personal belief.^[4]

Intuitively $P(\text{Match} \mid \text{first name matches})$

For n features (f_1, \dots, f_n)

$P(\text{Match} \mid f_1, \dots, f_n)$

Use maths!



M, U AND WEIGHTS

Bayes Theorem

$$P(\text{Match} \mid \text{Observation}) = \frac{P(\text{Observation} \mid \text{Match}) \times P(\text{Match})^\lambda}{P(\text{Observation})}$$

$$P(\overline{\text{Match}} \mid \text{Observation}) = \frac{P(\text{Observation} \mid \overline{\text{Match}}) \times P(\overline{\text{Match}})^{1-\lambda}}{P(\text{Observation})}$$



$\lambda = P(\text{Match})$ = Probability that 2 random records match

Substitution

$$\text{Odd}(\text{Match} \mid \text{Observation}) = \frac{P(\text{Observation} \mid \text{Match}) \cdot \lambda}{P(\text{Observation})} \times \frac{P(\text{Observation})}{P(\text{Observation} \mid \overline{\text{Match}}) \cdot (1 - \lambda)}$$

$$\text{Odd}(\text{Match} \mid \text{Observation}) = \frac{\lambda}{1 - \lambda} \times \frac{P(\text{Observation} \mid \text{Match})}{P(\text{Observation} \mid \overline{\text{Match}})}$$

WE HAVE MANY FEATURES

$$Odd(\text{Match} \mid \text{Observation}) = \frac{\lambda}{1 - \lambda} \times \prod_{i=1}^n \frac{P(f_i \mid \text{Match})}{P(f_i \mid \overline{\text{Match}})}$$
$$\frac{m_f}{u_f} = K_f$$

Bayesian Factor

► $m_f = P(f \mid \text{Match})$ = When 2 records match, how likely is it that they have the same last name?

 m measures feature's **ACCURACY**

► $u_f = P(f \mid \sim\text{Match})$ = When 2 records do not match, how likely is it that they have the same gender?

 u measures feature's **COINCIDENCE**

WE WANT TO ADD

$$Odd(\text{Match} \mid \text{Observation}) = \frac{\lambda}{1 - \lambda} \times \prod_{i=1}^n K_i$$

$$\log_2(Odd(\text{Match} \mid \text{Observation})) = \log_2\left(\frac{\lambda}{1 - \lambda}\right) + \sum_{i=1}^n \log_2(K_i)$$

M_{obs} M_{prior} M_f

$$M_{\text{Obs}} = M_{\text{Prior}} + \sum_{i=1}^n M_{f_i}$$





FELLEGI-SUNTER IN ACTION: MEET SPLINK

MEET SPLINK

- MIT Licensed, Python
- 🇬🇧 Ministry of Justice
- Implements the Fellegi-Sunter model...



Doc: <https://moj-analytical-services.github.io/splink/index.html>

... in an interesting way

ESTIMATING PARAMETERS

λ : “How many matches do we expect?”

🕵️ → **Educated guess**

u : “How often do people have the same name?”

🕵️ → **Random Sampling**

PICKING PARAMETERS

m : “How clean is the data?”

“How often is the last name mistyped?”



```
estimate_m_from_label_column("ssn")
```

[Maximum Likelihood Function](#)

```
estimate_parameters_using_expectation_maximisation(block)
```

[Expectation Maximization](#)

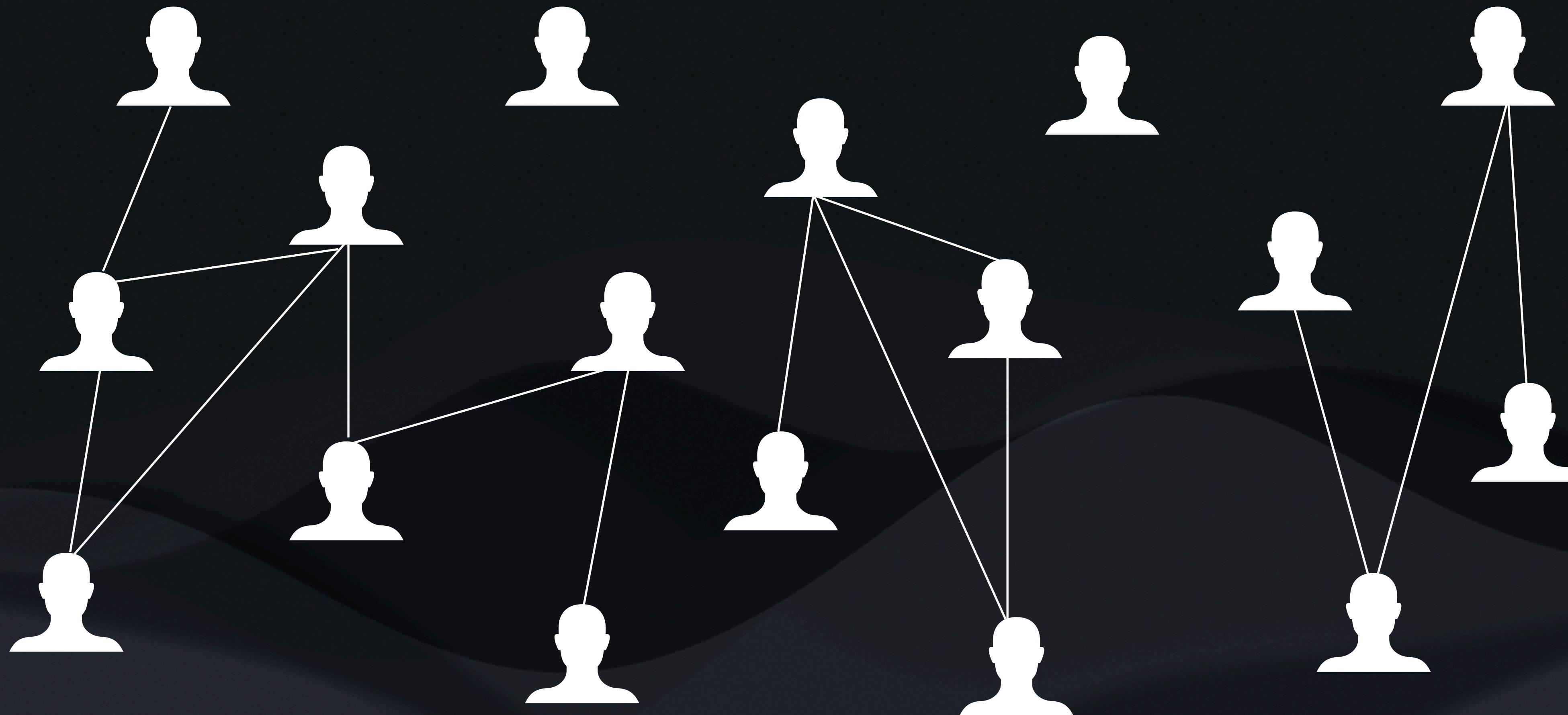
CLUSTERING: GROUPING THE EVIDENCE



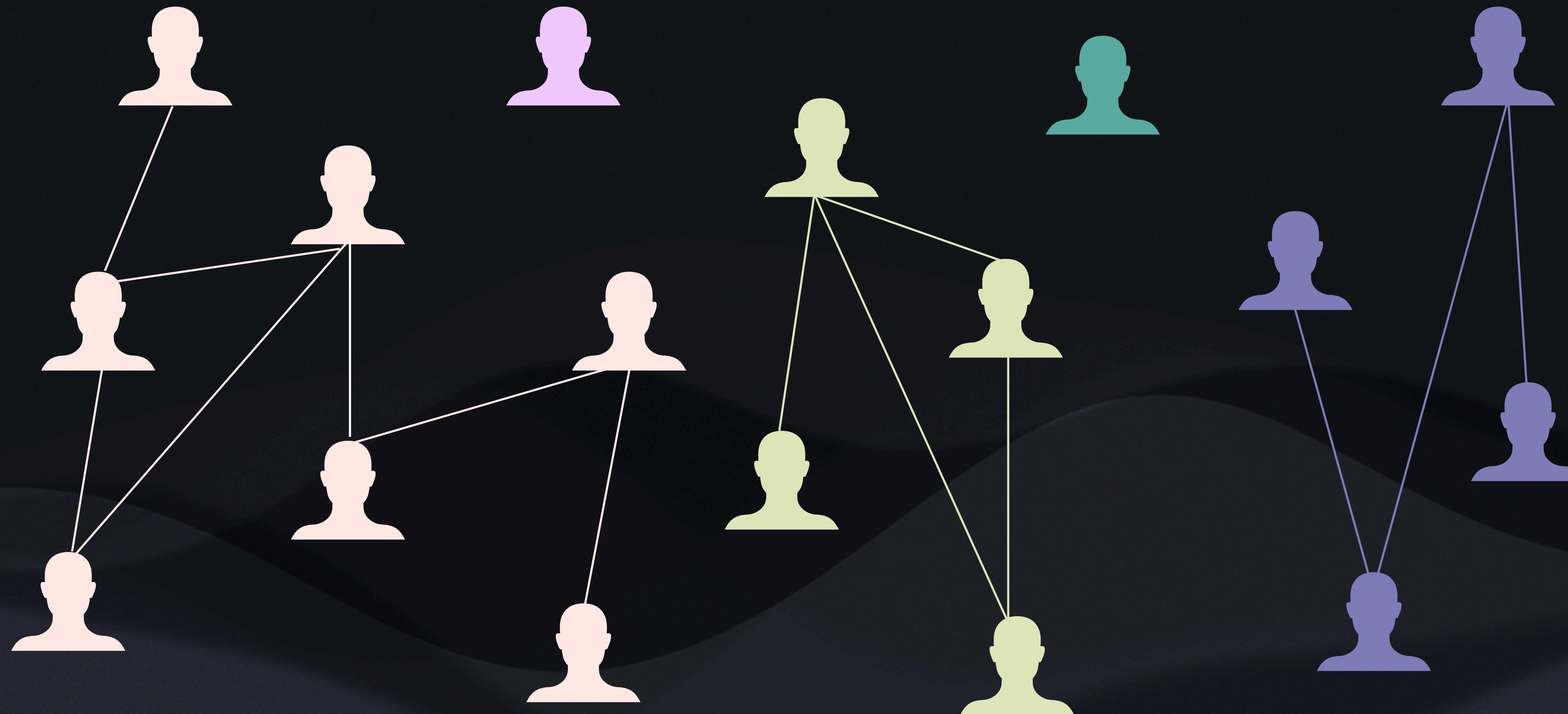
LINKS, NOW WHAT?



LINKS, NOW WHAT?

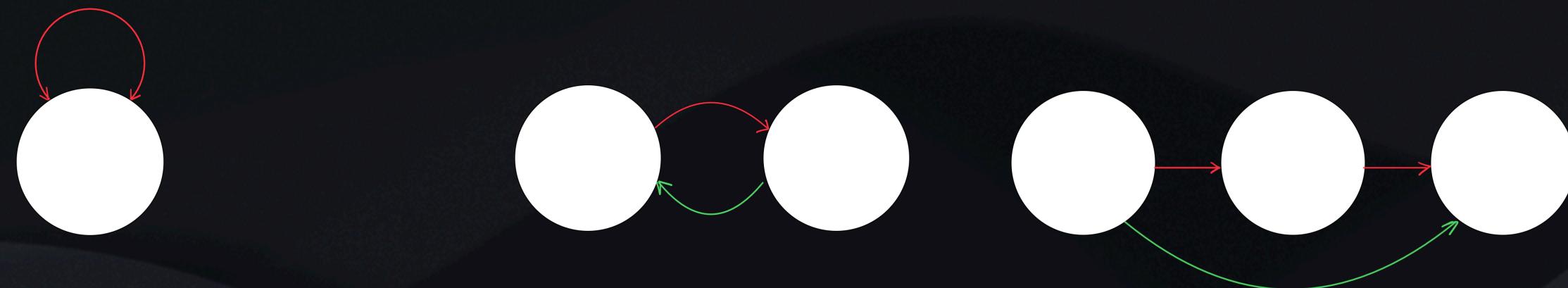


LINKS, NOW WHAT?



CONNECTED COMPONENTS

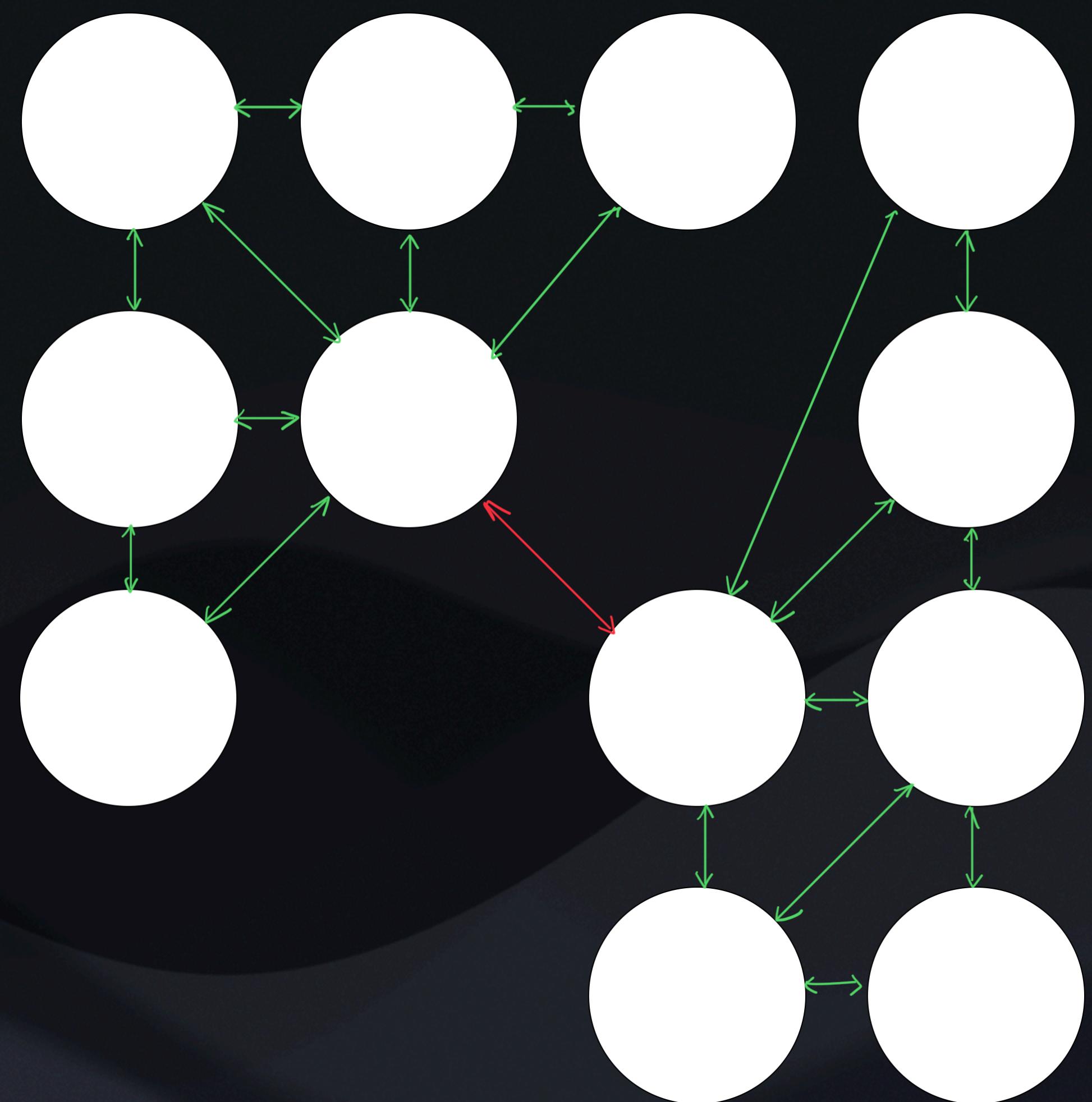
In [graph theory](#), a **component** of an [undirected graph](#) is a [connected subgraph](#) that is not part of any larger connected subgraph. [...] Components are sometimes called **connected components**.



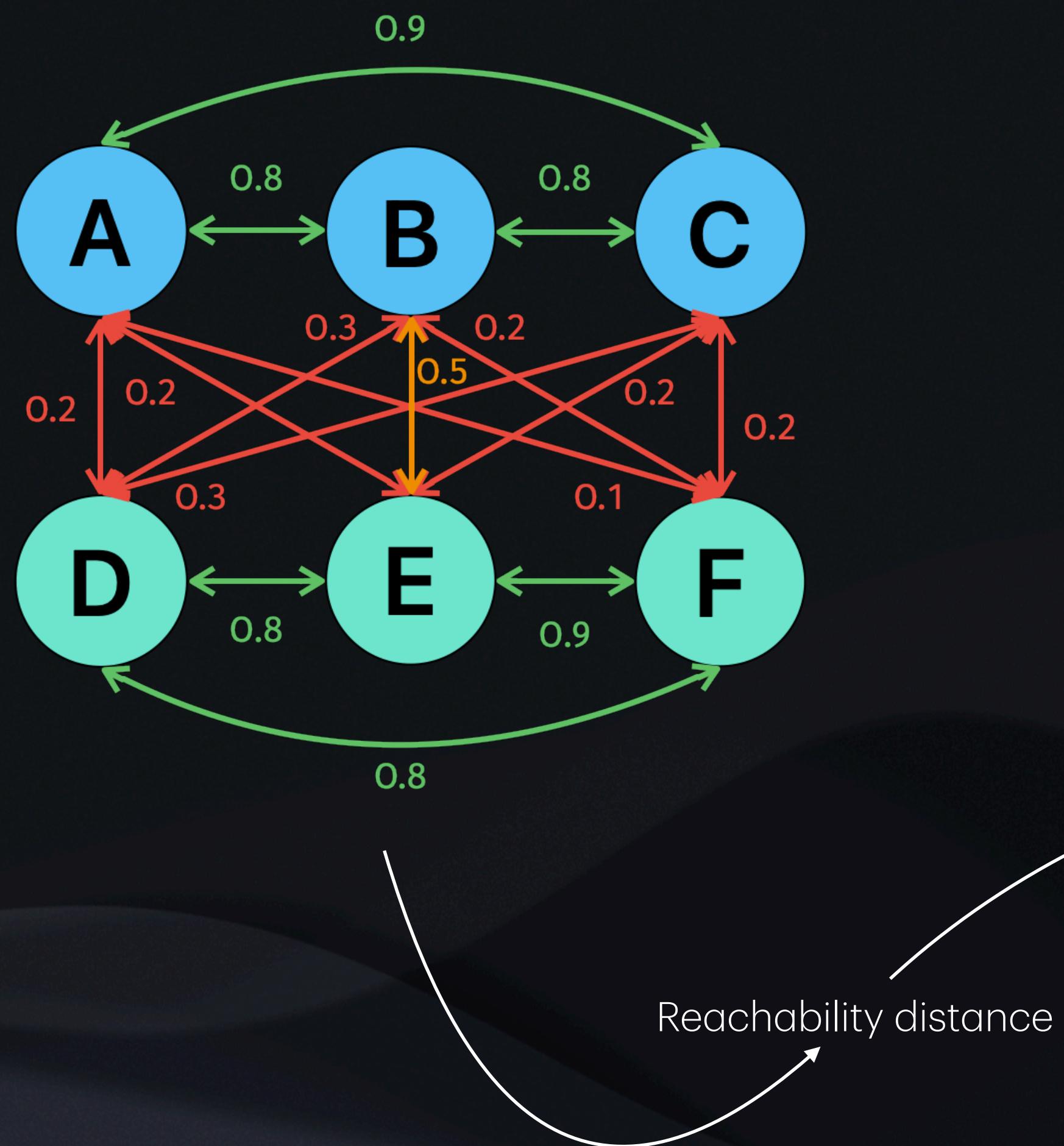
- ✓ Reflexive, Symmetric, Transitive relations (edges)
- ✓ Easy to implement (DFS, Union Find, ...)

5.4. CONNECTED COMPONENTS

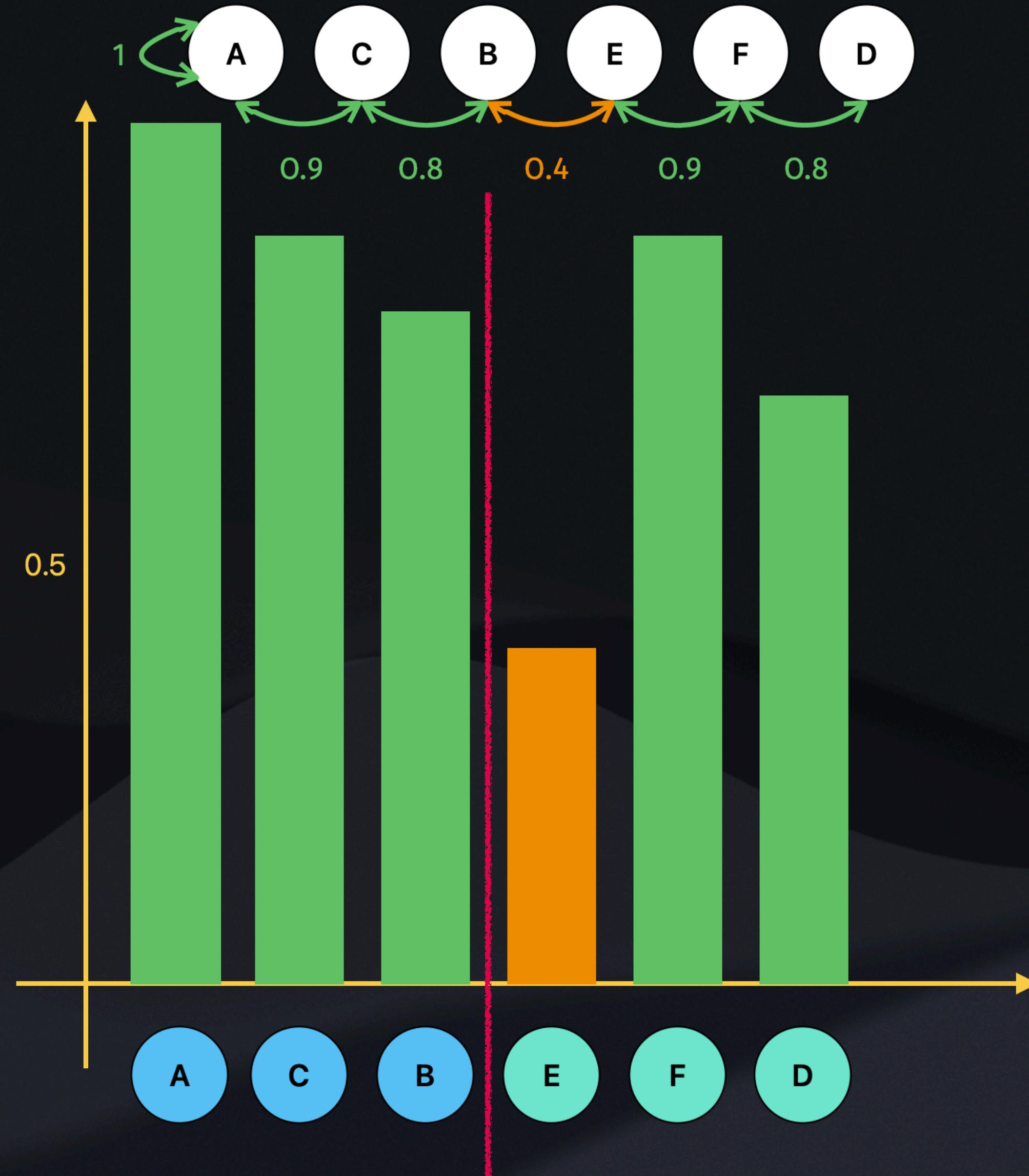
WEAKNESS



CLUSTERING ALGO: OPTICS



m is a similarity metric!





CANONICALISATION: ASSEMBLING THE TRUTH

COMMON TECHNIQUES

Heuristics!

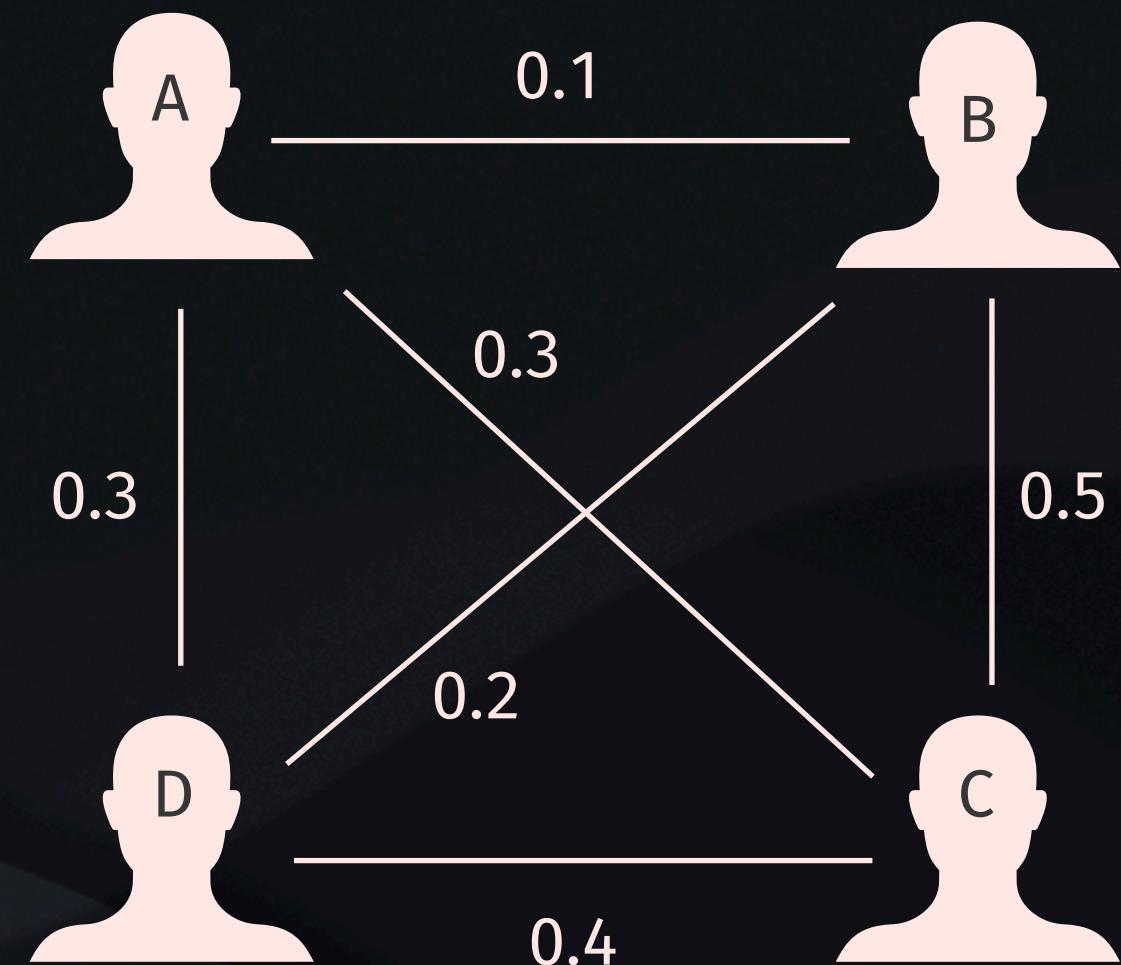


- Pick a **random** record in the cluster
- **Majority Voting:** most common value
- Mean, median, ... for numerical value
- **Most informative:** longest string, most decimals
- **Prioritised Source:** more trustable

USING DISTANCES

MINI-MAX

Remember we have metrics!

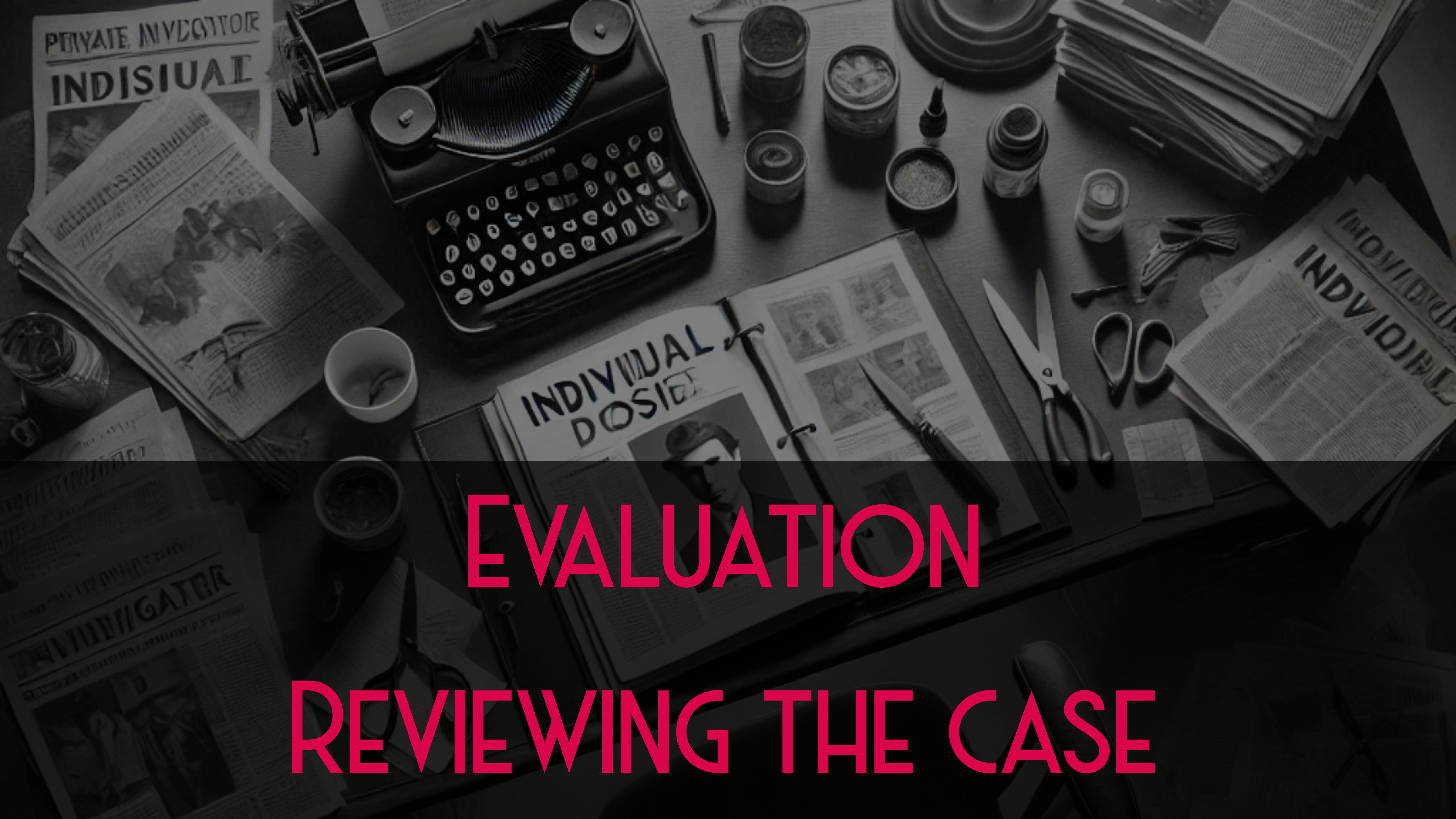


A	$\max(0.1, 0.3, 0.3) = 0.3$
B	$\max(0.1, 0.5, 0.2) = 0.5$
C	$\max(0.3, 0.5, 0.4) = 0.5$
D	$\max(0.3, 0.2, 0.4) = 0.4$

min

EVALUATION

REVIEWING THE CASE



MONITOR VS. EVALUATE

Monitor

Continuous supervision

Unsupervised, no labelling

Birds-Eye view

Identify suspicious clusters

Evaluate

Performance metrics

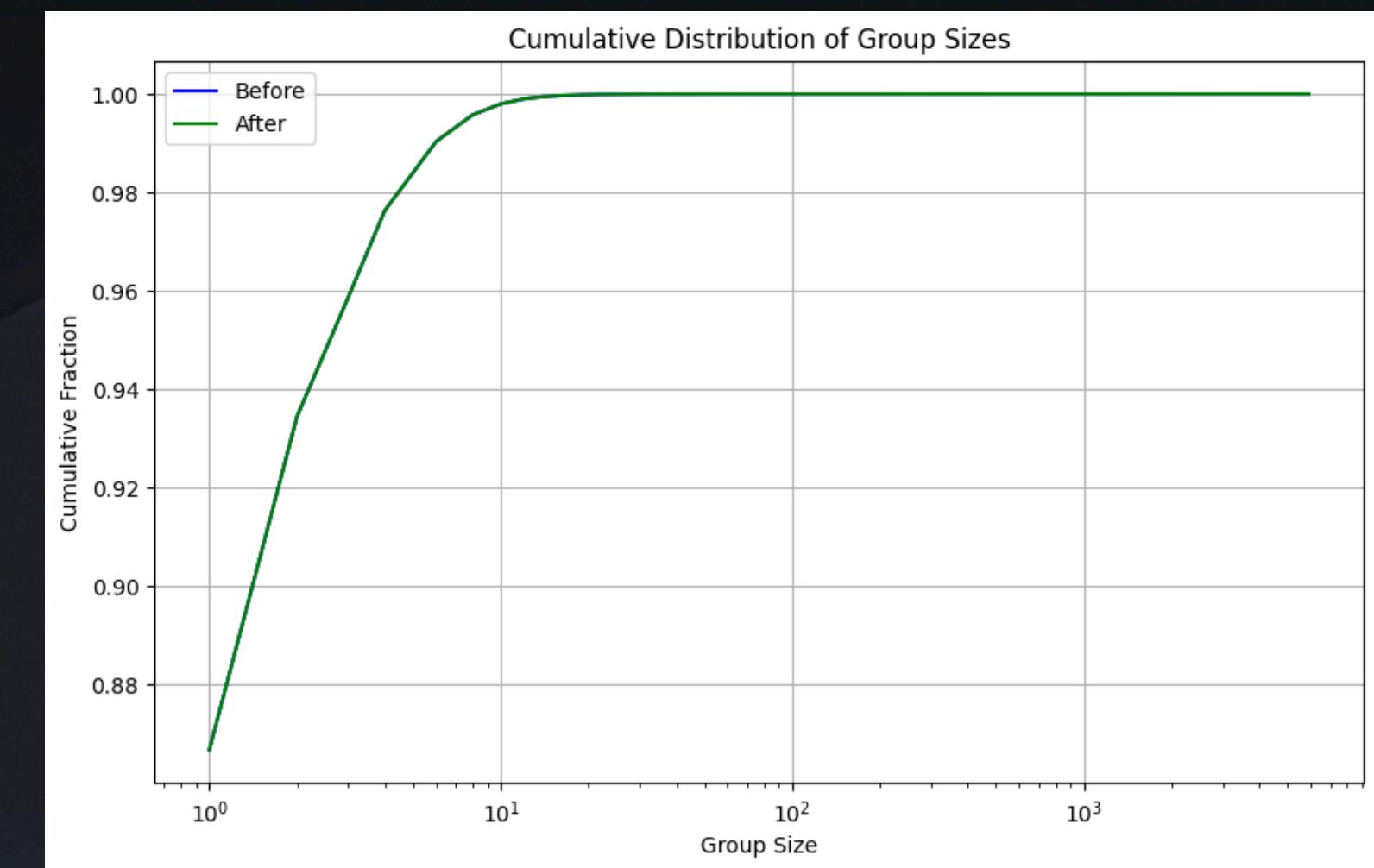
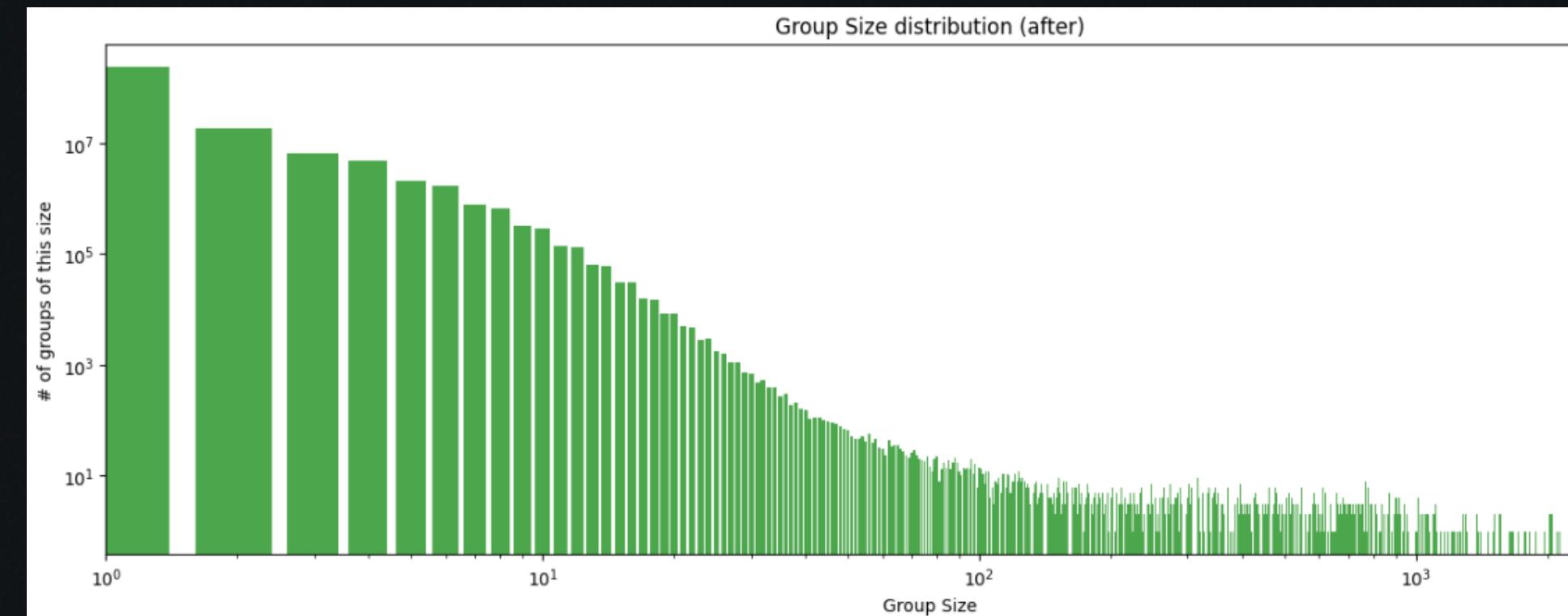
Requires labelling

Detailed metrics (recall, precision)

Start from a sample and extrapolate

(MONITOR) CLUSTER STATS

- Mean / Median cluster size
- Histogram
- Cumulative chart
- Matching rate
- Diversity Index: Hill Numbers



(MONITOR) ENTROPY

In [information theory](#), the **entropy** of a **random variable** quantifies the average level of uncertainty or information associated with the variable's potential states or possible outcomes. [...]

[https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))

$$H(X) = - \sum_{x \in X} p(x) \times \log(p(x))$$

{Leslie, Leslie, Lesly, Leslie, Leslie, Leslie}

x	count(x)	p(x)	log(p(x))	p log(p)
Leslie	5	0,83	-0,08	-0,07
Lesly	1	0,17	-0,78	-0,13
$H(X)$			0,20	

{John, Jon, John, Jon, John, Jon}

x	count(x)	p(x)	log(p(x))	p log(p)
John	3	0,50	-0,30	-0,15
Jon	3	0,50	-0,30	-0,15
$H(X)$			0,30	

p(x): frequentist approach!



DATA SCIENTIST TAKEAWAYS



Cleansing

Part of my work is to represent the data differently

Linking

I use maths: distance metrics, (bayesian) probabilities

Clustering

By being probabilistic, we can leverage more clustering methods

Canonicalisation

Sometimes, I use heuristics too!

Validation

Look into the data distribution! Then use stats



A HUGE THANK YOU

Rate the presentation!



Download materials

