Every thursday

New webinar

Unit8™ Talks

We will teach you how to unlock
the potential of data and AI.
Introducing Unit8 webinars.

on
technologies

Unit8 Talks

Thursday, June 11

Real data stories: Entity Resolution

Webinar series

Webinar 03

Bianca and Radu talks on technologies

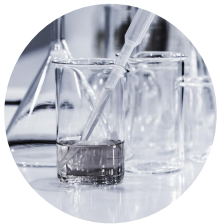
Unit8 intro

AI & Data Services Provider

We are “one stop shop” for solving business problems using **Big data, artificial intelligence, data science** and custom software engineering.

Services provided include **AI/ML Advisory** and consulting, **End to end solutioning** and data infrastructure advisory and build up

Experiences



pharma
& chemical

Automotive
& aviation

finance
& insurance

industry 4.0

amazon web services Partner Network
CONSULTING PARTNER

Microsoft Azure
Certified

digitalswitzerland

swiss made software

Unit8

Outline

- Introduction to ER (not the TV show)
- Use cases
- Typical steps
- Coding example

Why do we need Entity Resolution?

At **Unit8**, we handle a lot of **data** for our customers.

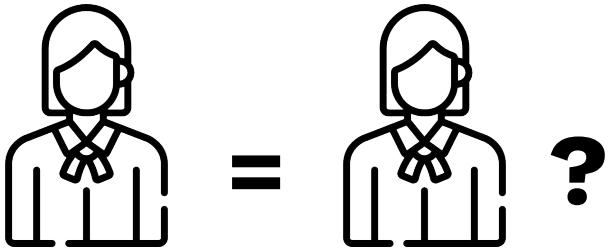
A lot of data often implies **data quality issues**.

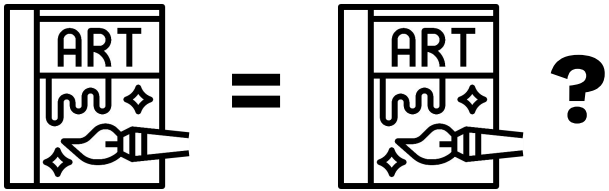
One of the ways we handle that is using **entity resolution**.

Unit8

What is Entity Resolution?

Entity resolution is about determining whether records from different data sources represent, in fact, the same entity.





Entity Resolution in plain sight

amazon : how do you tell the total stock of a certain product given all your distributors?

Entity Resolution

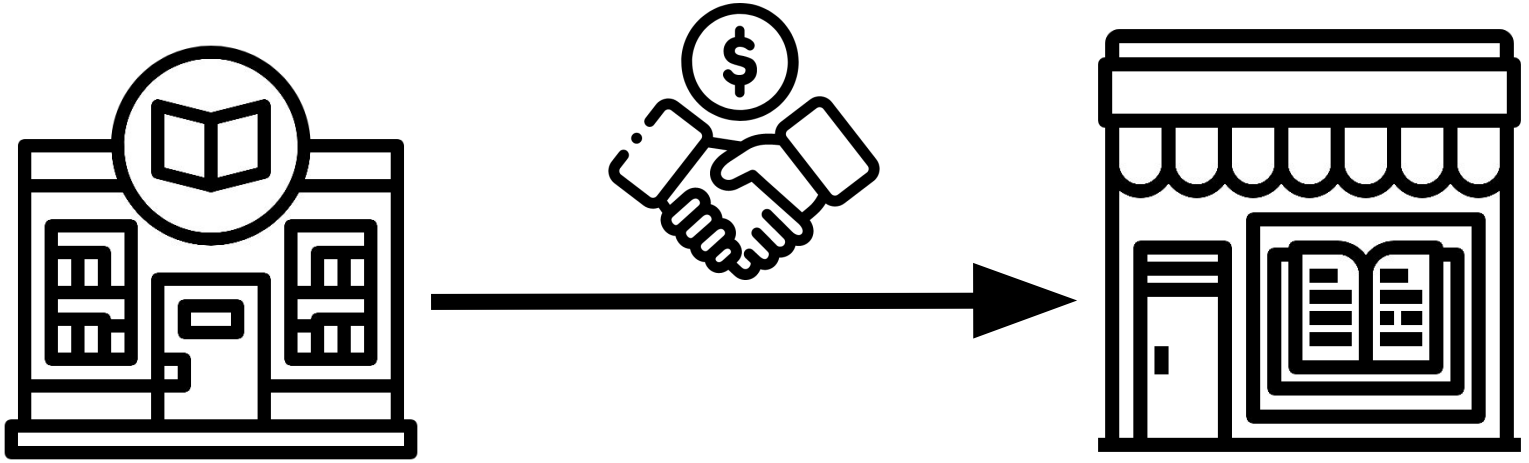
facebook: how do you tell when someone is trying to make two accounts for the same person?

Entity Resolution

Insurance & Finance: how do you prevent people from commiting fraud?

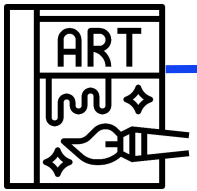
Entity Resolution

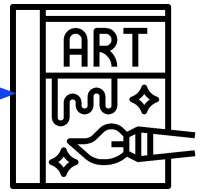
In order to better understand **why** we need entity resolution and how it works, let’s imagine a scenario...

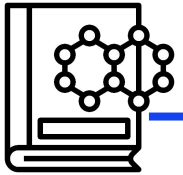

BOOKZ
BOOKS

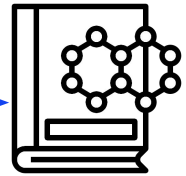
BOOKZ

BOOKS

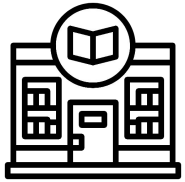




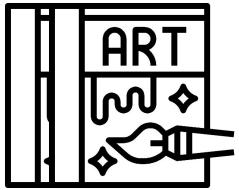


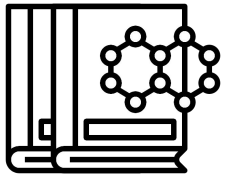


Unit8



BOOKZ & BOOKS



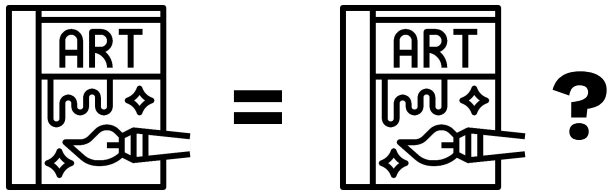


Unit8

Data matching

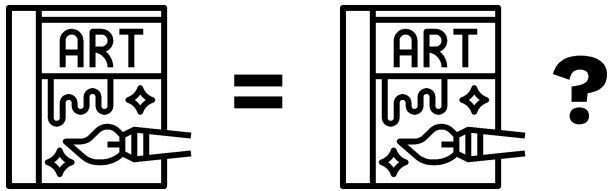
Record Linking

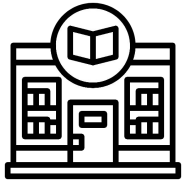
Deduplication



Entity Resolution

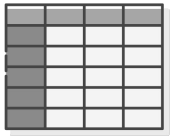
What could possibly go wrong with





BOOKZ

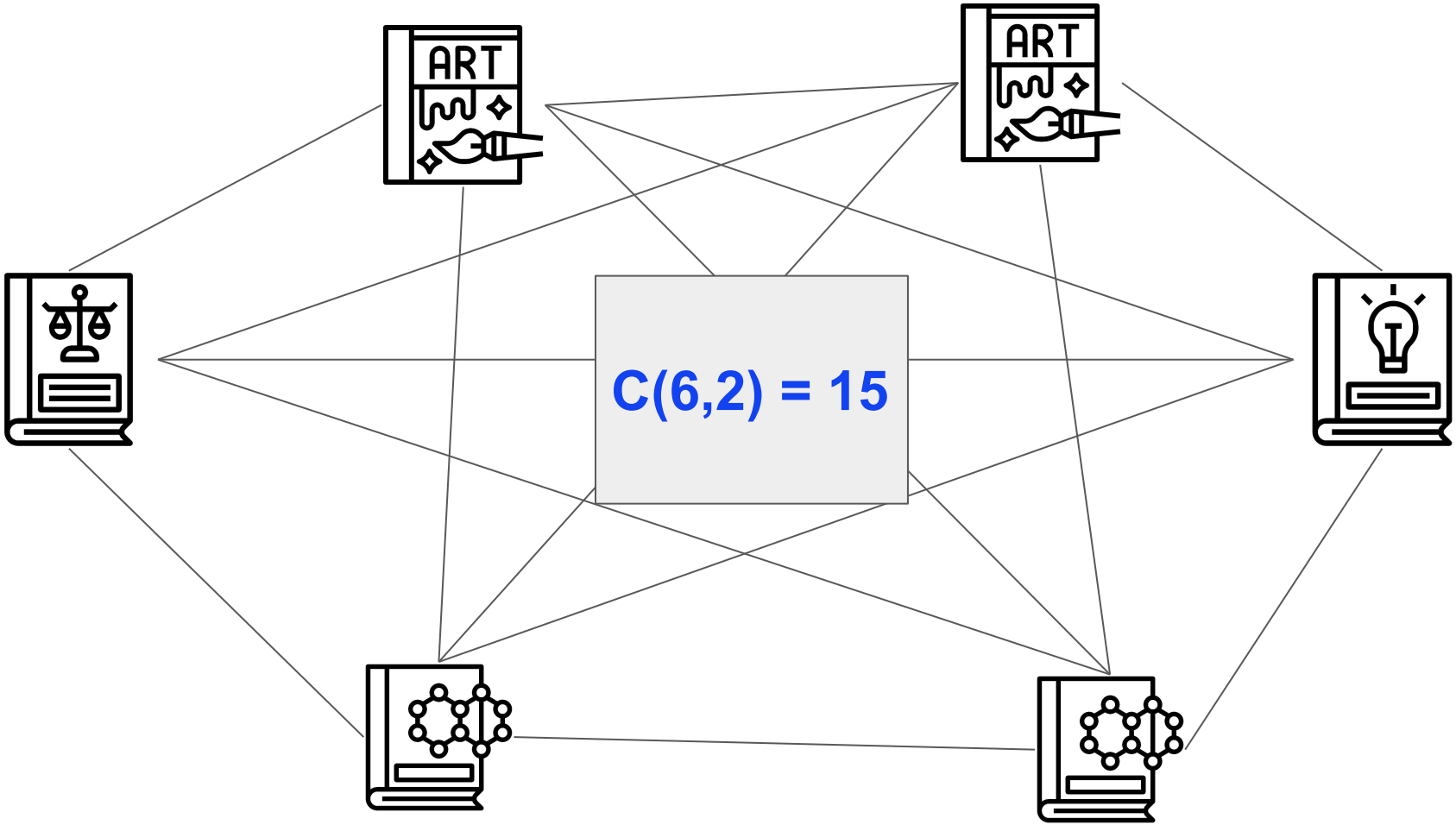
BOOKS



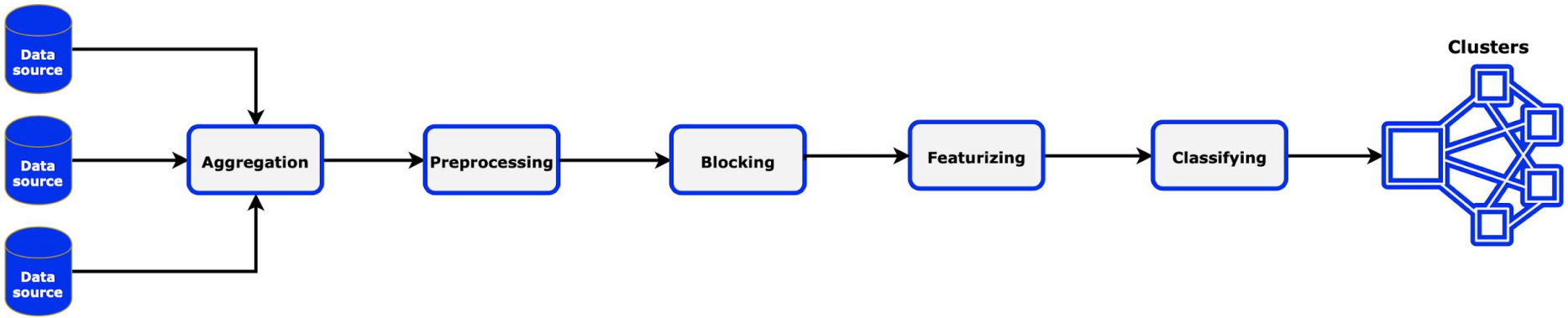
book_title	author	stock
Art	Bob Ross	20

book	pcs
Art - B. Ross	15’000

Unit8


ART
ART
C(6,2) = 15

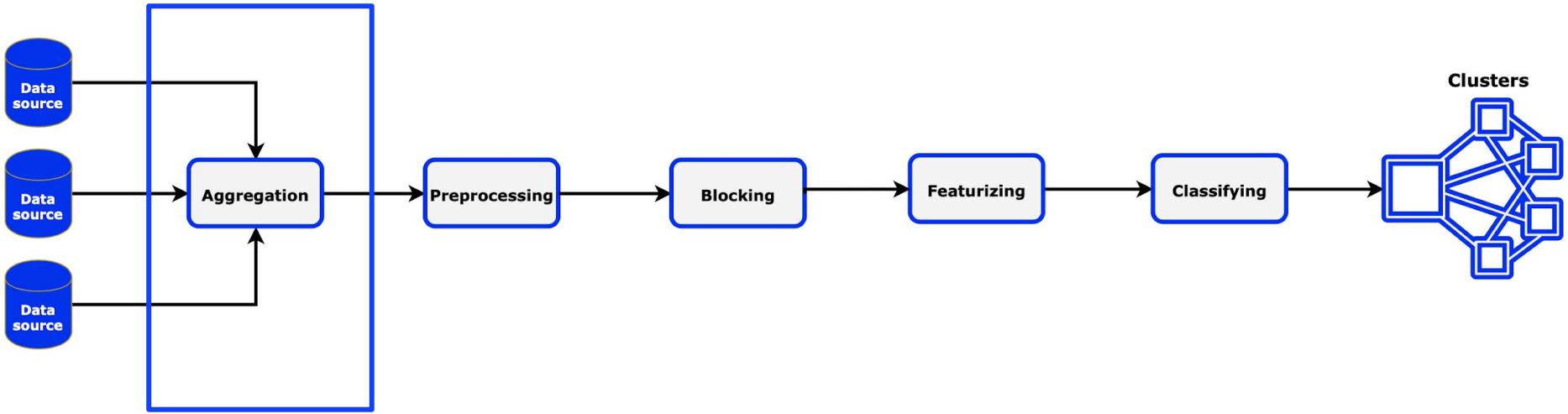
Typical steps



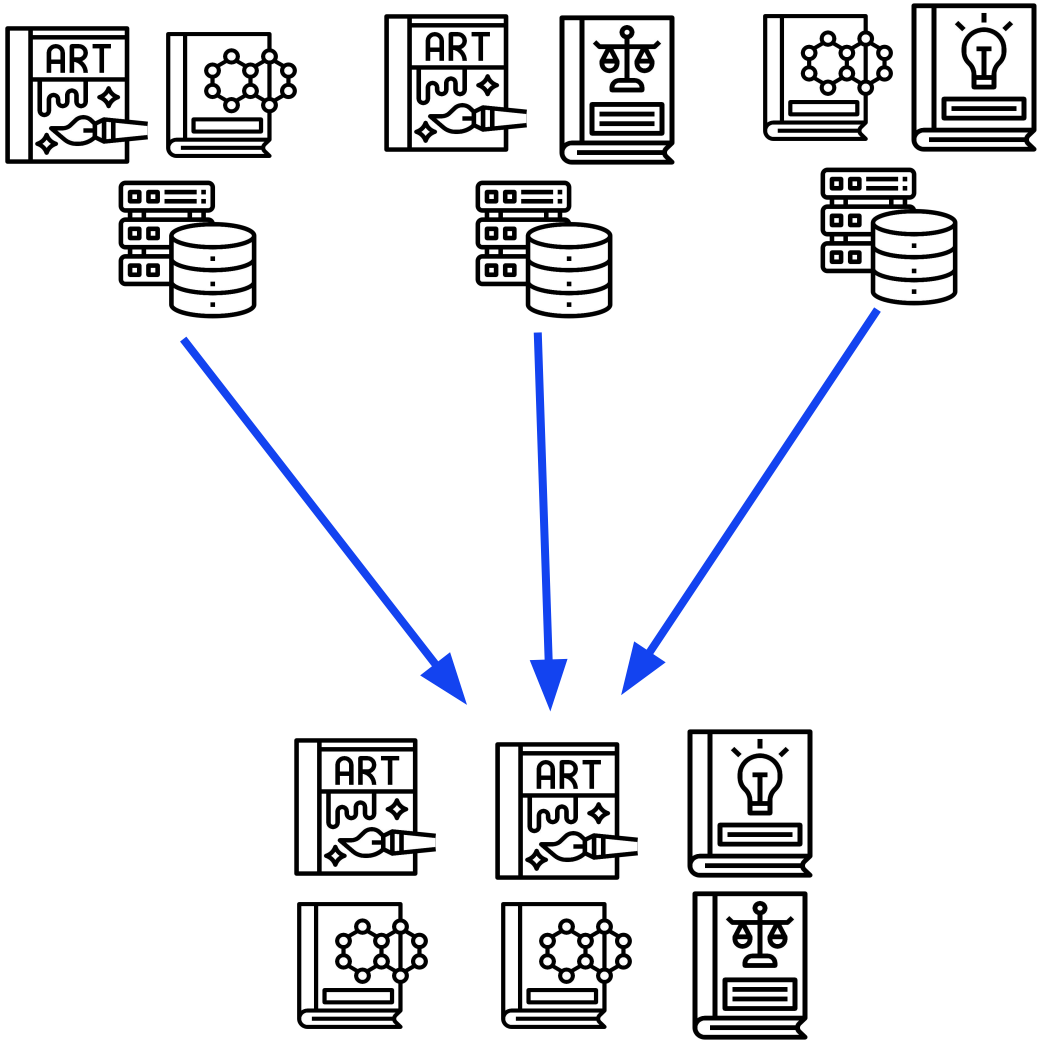
Balaji, J., Javed, F., Kejriwal, M., Min, C., Sander, S. and Ozturk, O., 2016. An ensemble blocking scheme for entity resolution of large and sparse datasets
Christen, P., 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*

Unit8

Aggregation

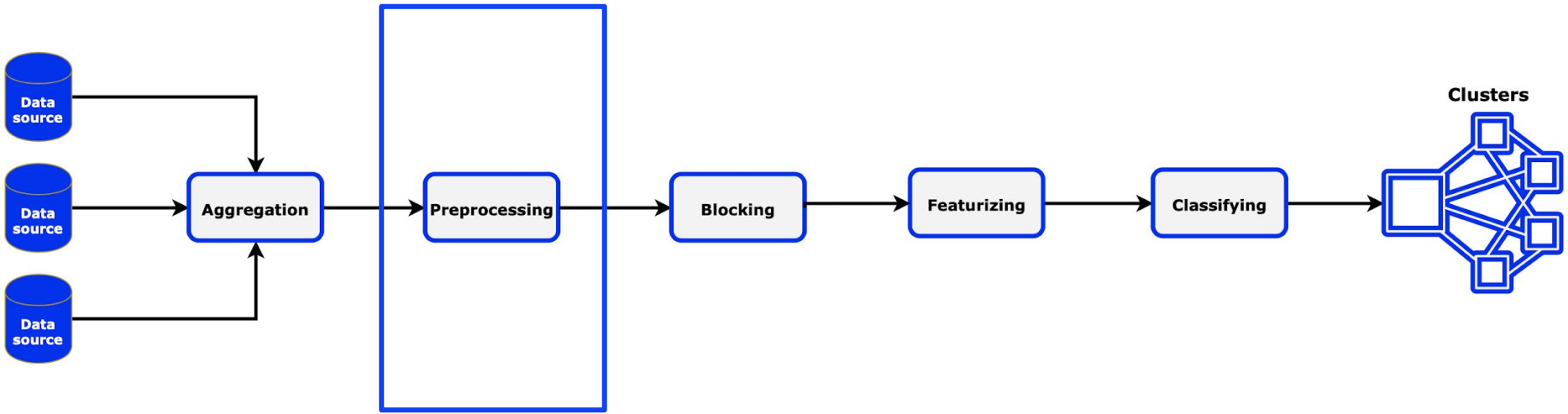


Unit8



Unit8

Preprocessing



Unit8

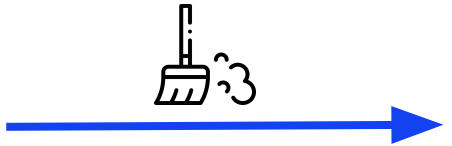
Dates

01-Jul-99
1 July 1999

1999.07.01

Special characters

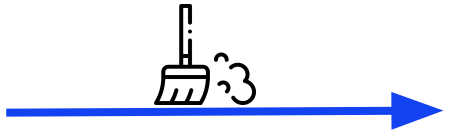
ER💔😱



ER

Transliterations

Peter Müller



Peter Muller
Peter Mueller?

Formats

10k

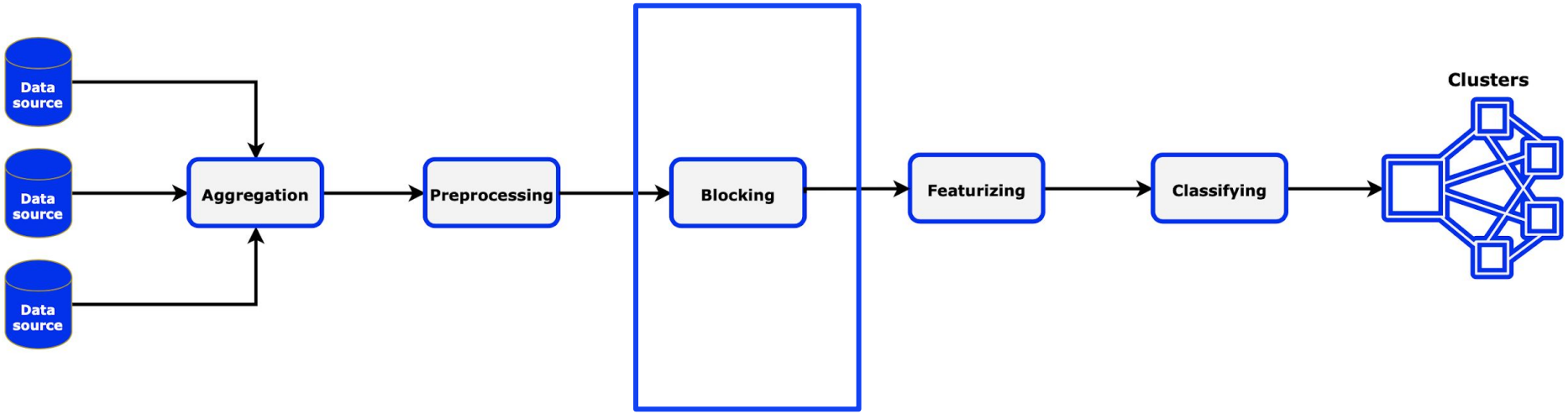
10’000

Casing

john doe
John Doe

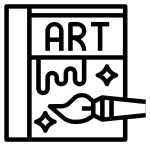
JOHN DOE
DOE JOHN?

Blocking



Unit8

Find **candidate** pairs

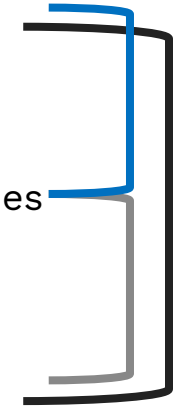


Art 101
Bob Ross

Chemistry for babies
Bob Ross

Art 101
Bob Ross

Law
John John



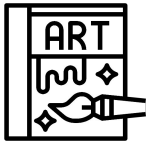
BLOCK BY AUTHOR

We consider as candidates books written by the same author

Why is blocking important?

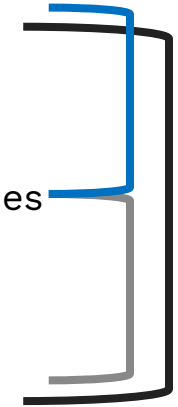
Art 101
Bob Ross

Chemistry for babies
Bob Ross



Art 101
Bob Ross

Law
John John



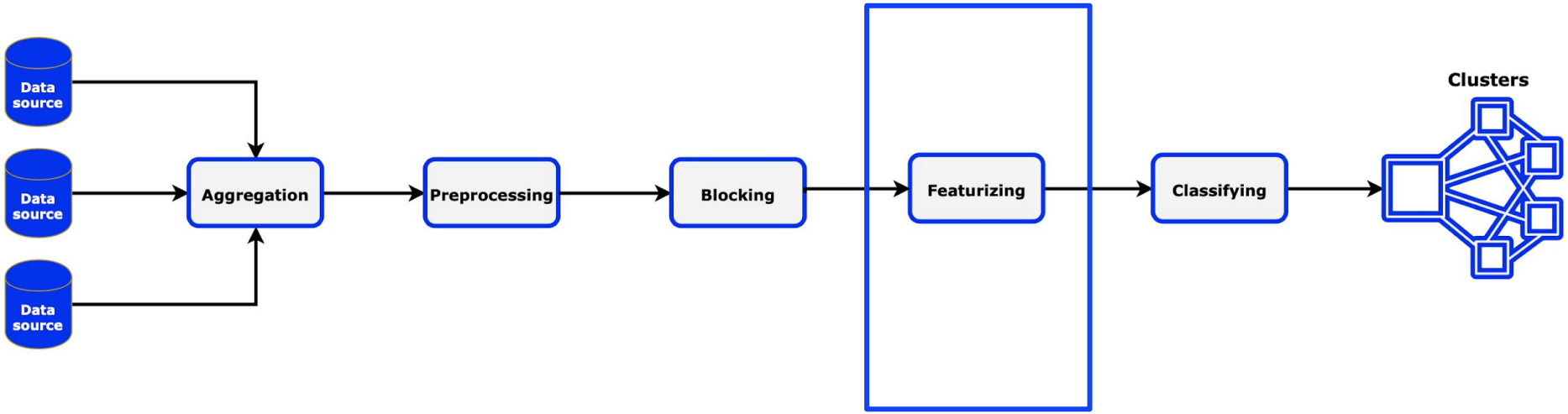
BLOCK BY AUTHOR

We consider as candidates books written by the same author

Limit computational complexity!

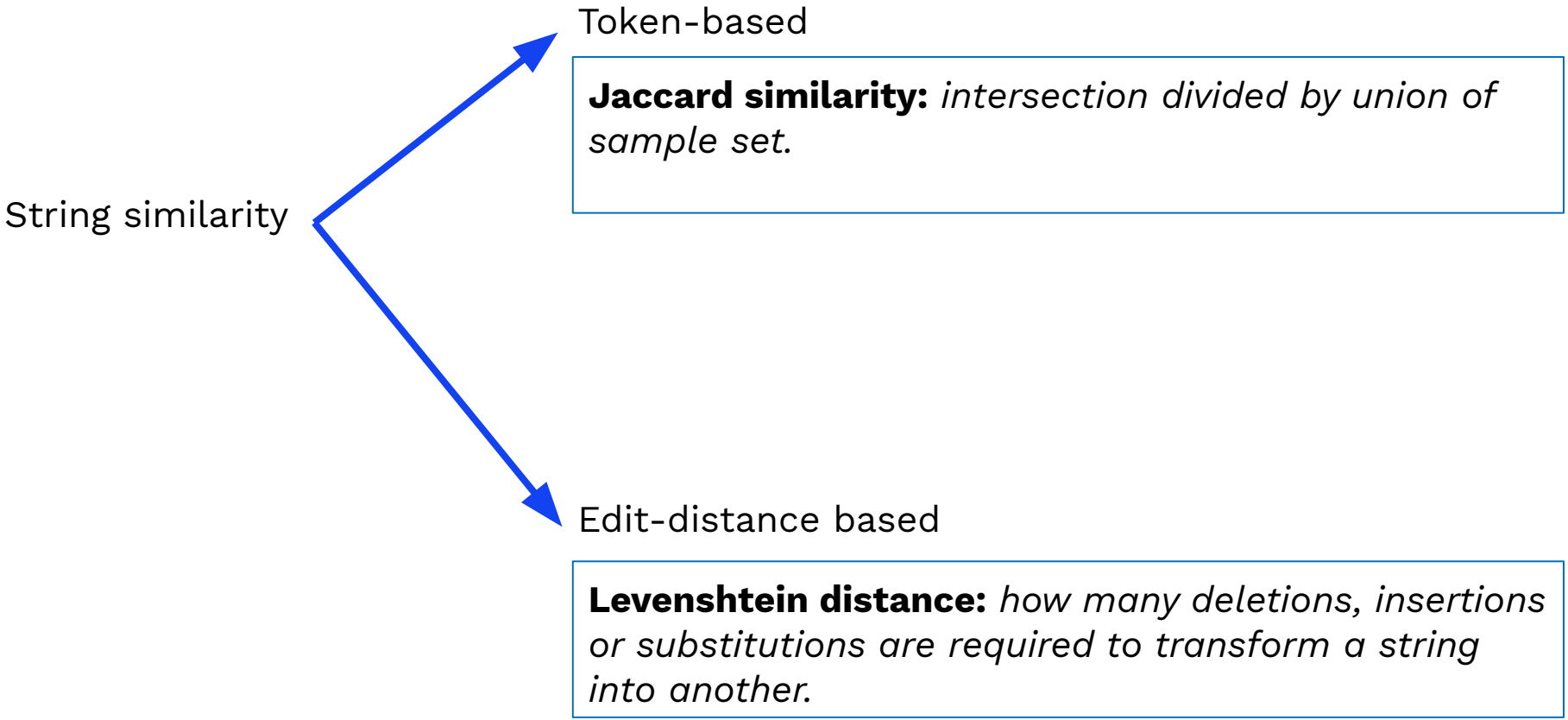
Reduced by 50%

Featurizing



Unit8

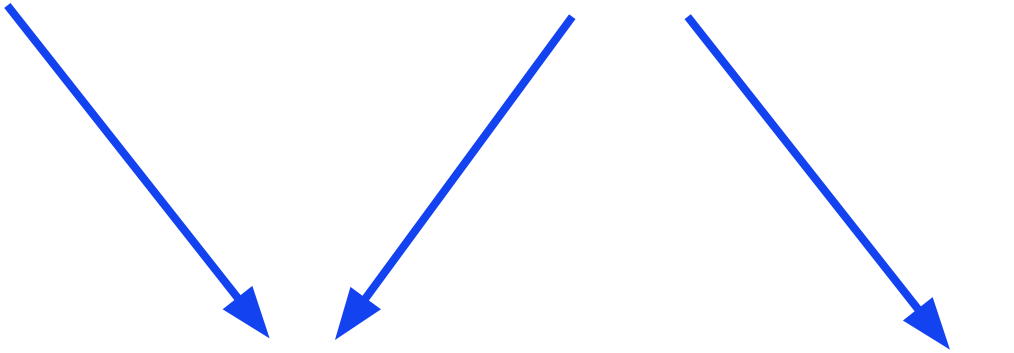
Measuring **similarity** of pairs



ART FOR BEGINNERS

BEGINNERS ART

BEGGINNERS ARTS



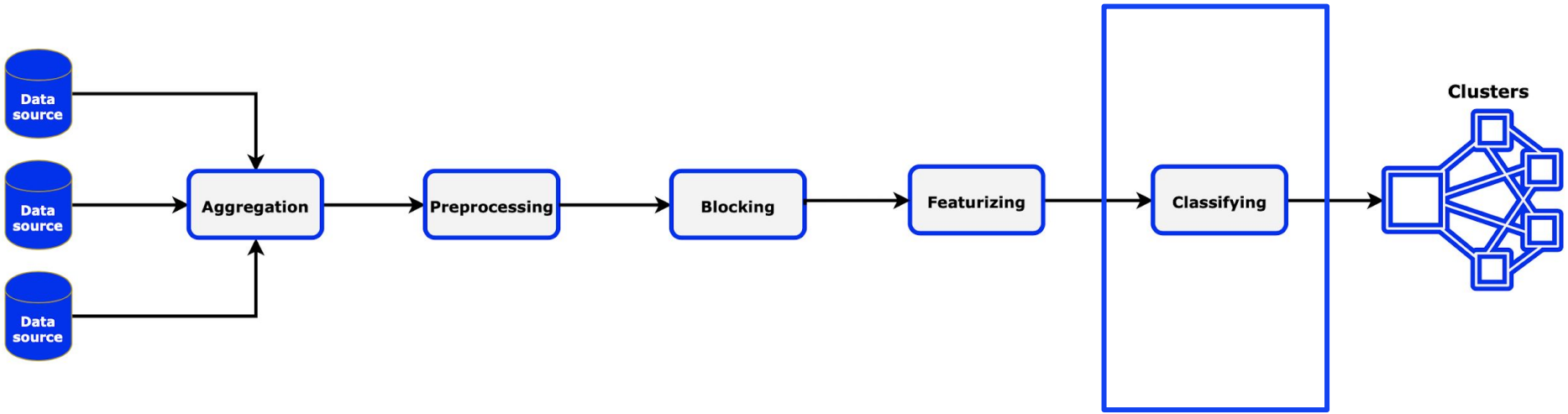
Levenshtein distance: 12

Jaccard similarity: 0.66

Levenshtein distance: 2

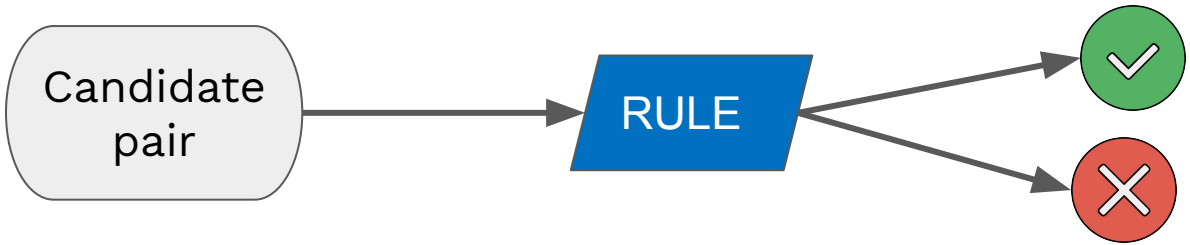
Jaccard similarity: 0

Classifying

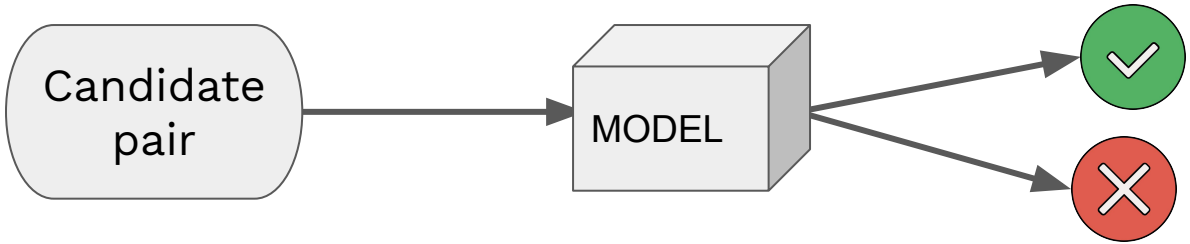


Unit8

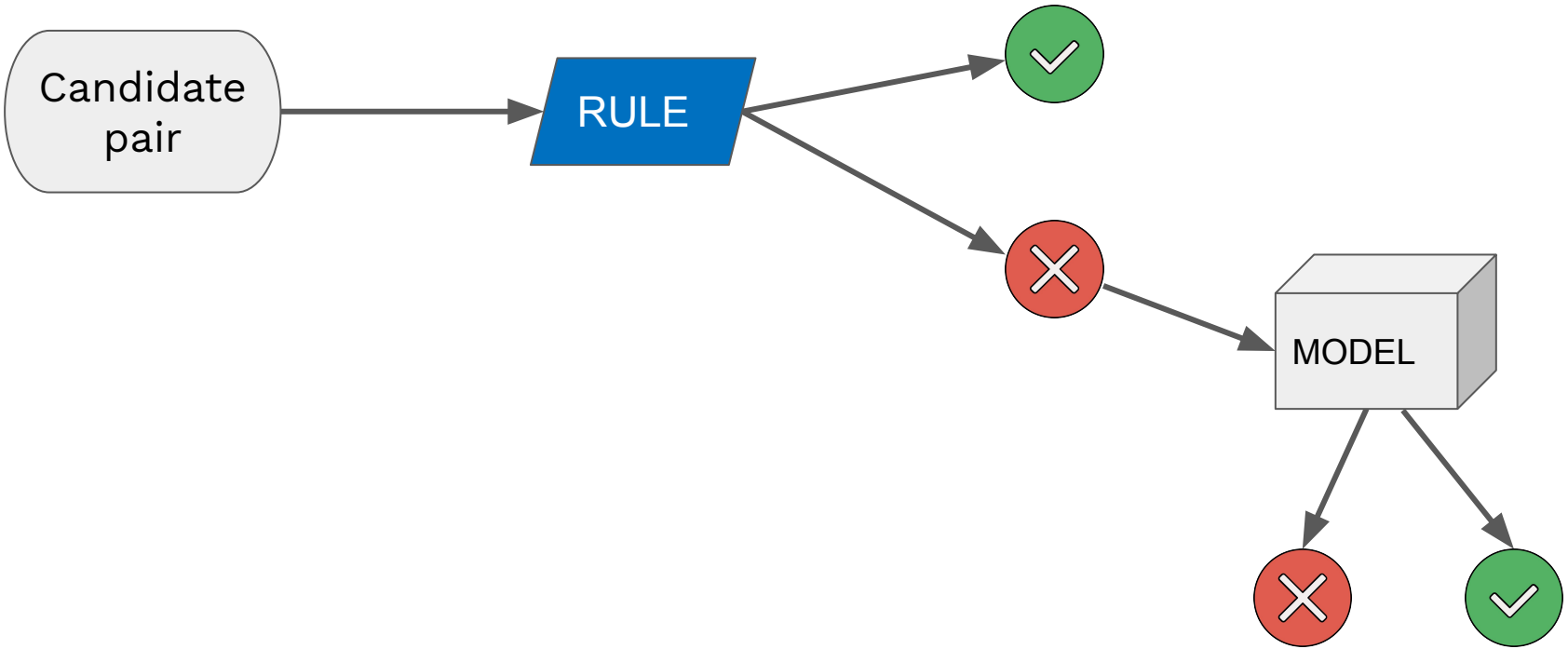
Rule based

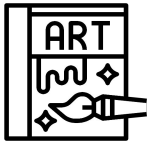
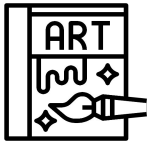
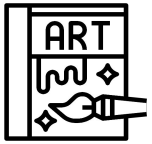
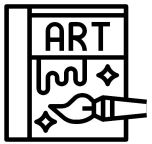


Model based

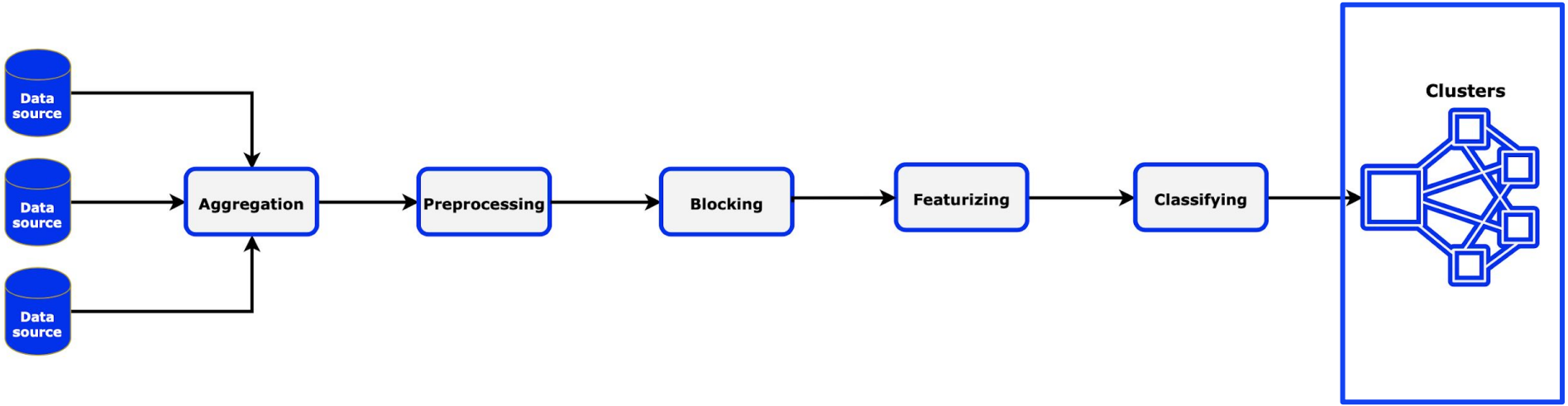


Hybrid



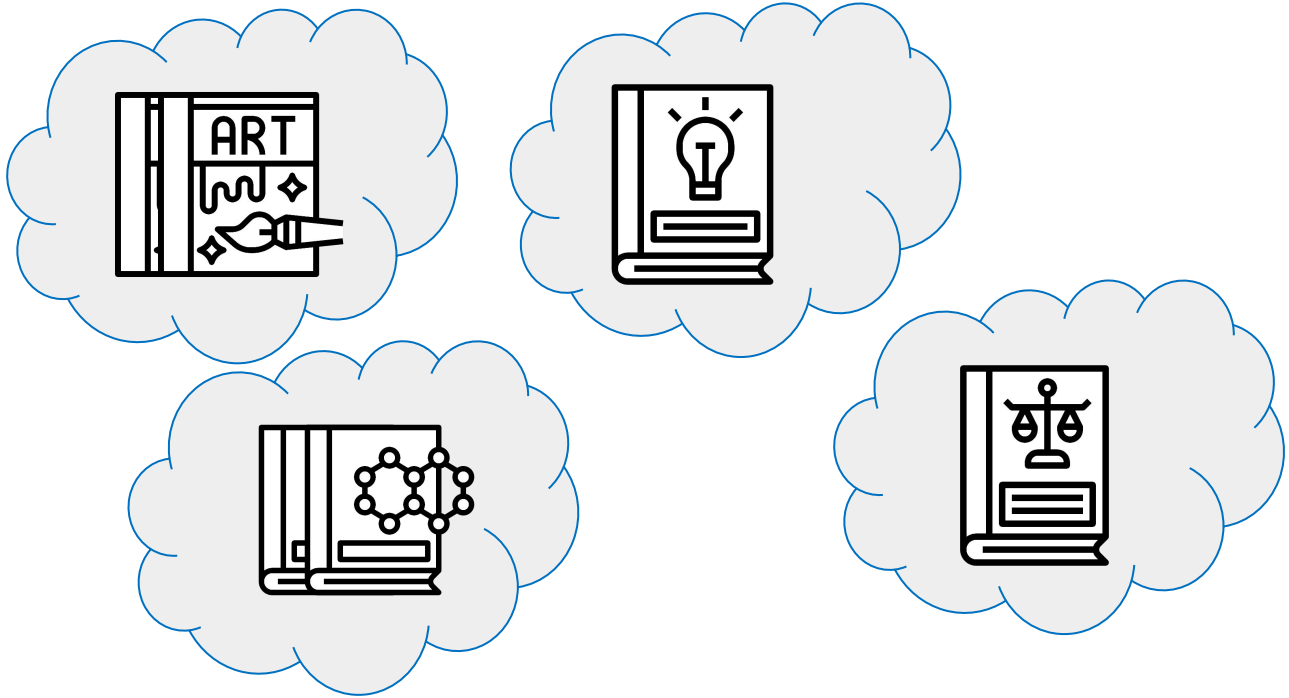
	ART FOR BEGINNERS		BEGINNERS ART		BEGGINNERS ARTS
High Jaccard index		=		≠	
Low Levenshtein distance		≠		=	
High Jaccard index **OR** Low Levenshtein distance		=		=	
High Jaccard index **AND** Low Levenshtein distance		≠		≠	

Clusters

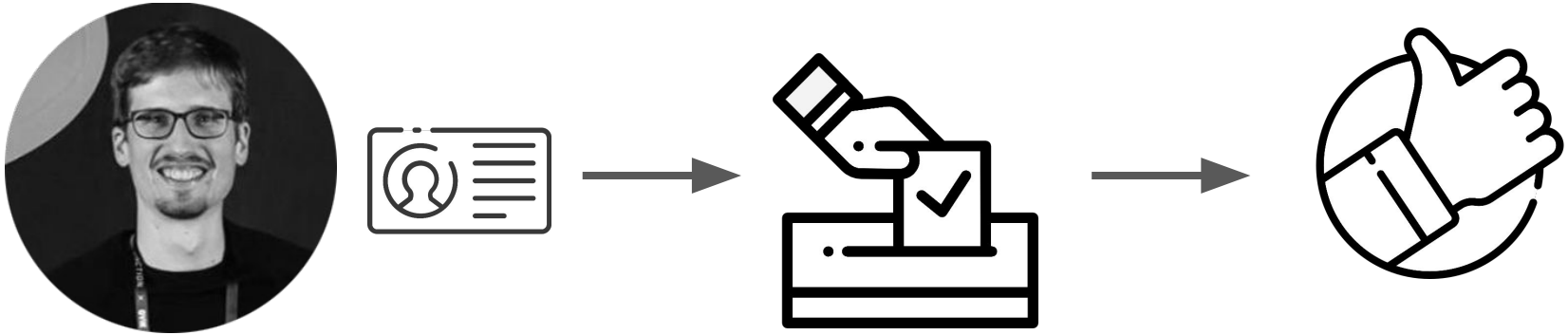


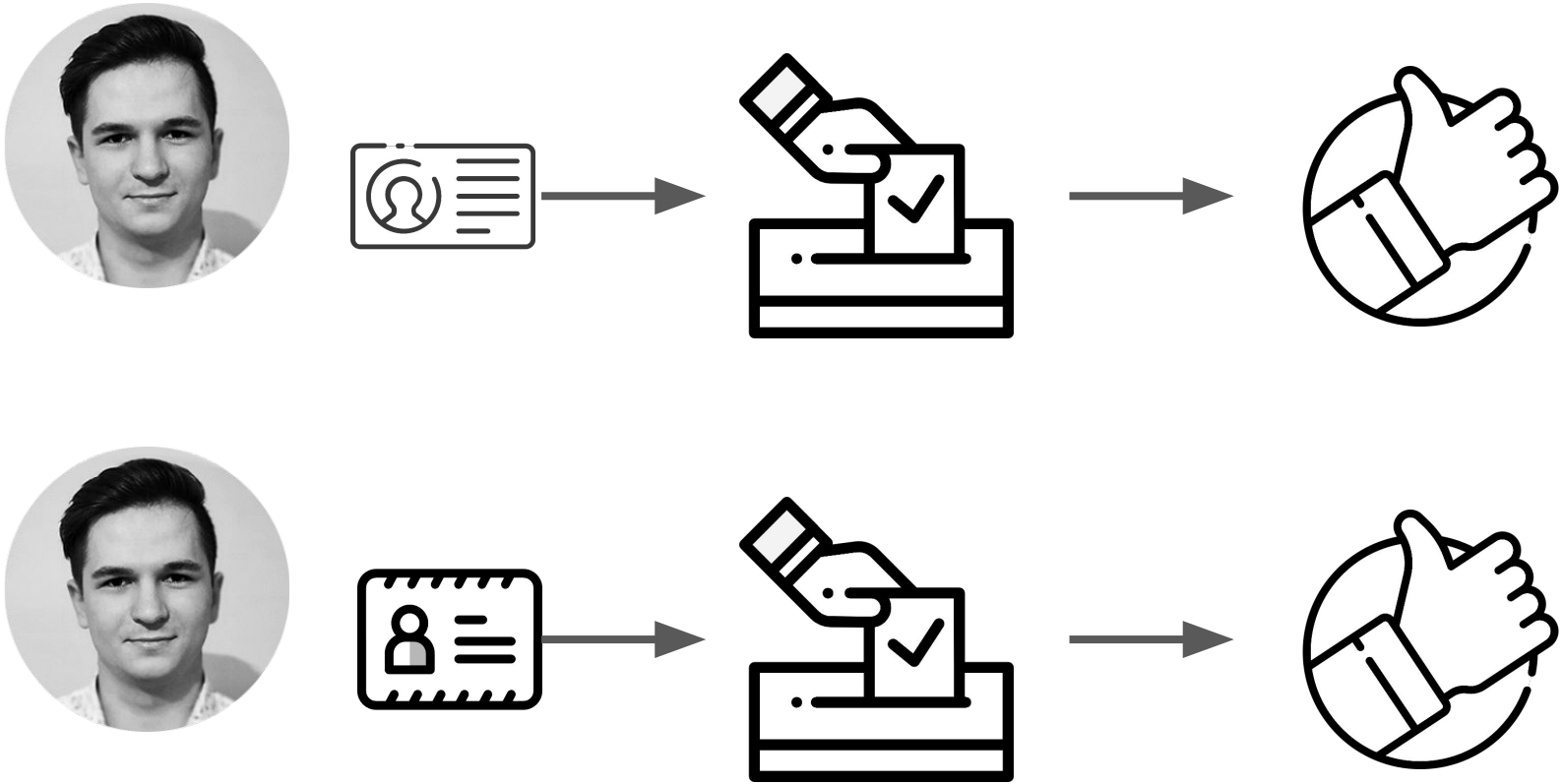
Unit8

BOOKZ & BOOKS

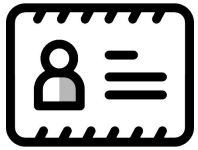


Let’s see how that looks in practice!





=

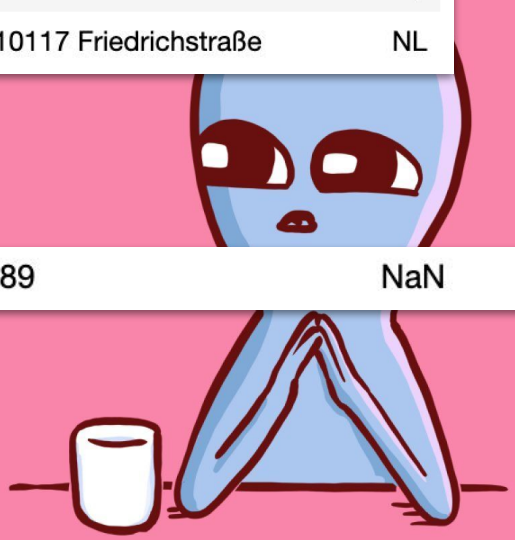


Live coding :)

AND YET

Joseph	Mueller	2 Jan 1989	Friedrichstrasse 76 10117 Berlin	Germany
Josef	Müller	1 Feb 1989	10117 Friedrichstraße	NL

J	Muller	2.1.1989	NaN	DE



Conclusions

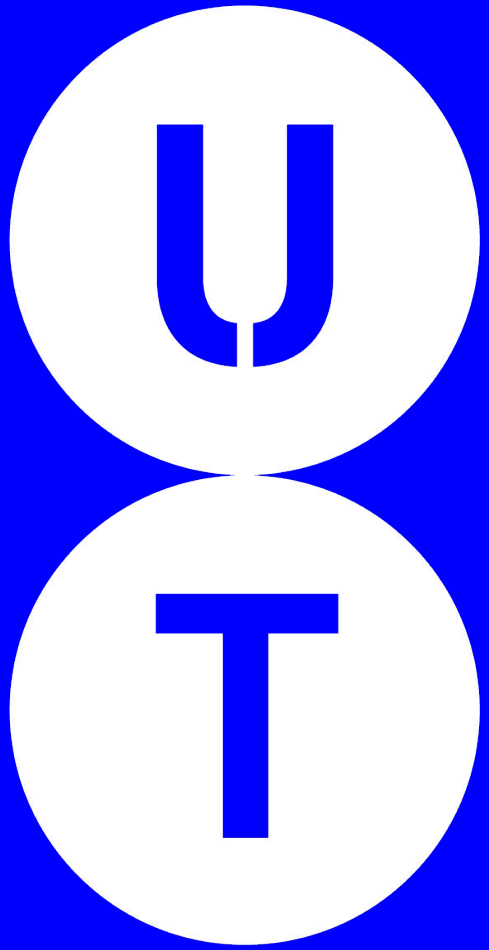
- The right implementation depends on the context
- Data quality is critical
- There are always exceptions

Unit8 Talks

Q&A

New webinar

Every thursday



Unit8™

We will teach you how to unlock
the potential of data and AI.
Introducing Unit8 webinars.

thank you

Talks

on technologies
on business

Every thursday

Marcin and Tomek talk

Next webinar

June 25, Thursday 1PM

We will teach you how to unlock the potential of data and AI. Introducing Unit8 webinars.

Improving Overall Equipment Efficiency with Advanced Analytics & AI

on business

unit8.co

thank you