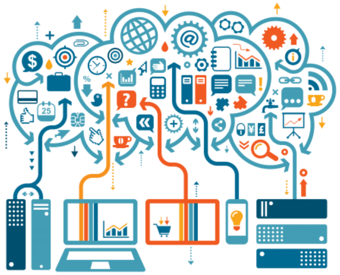


UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Big Data Integration for Data-Centric AI

itaDATA2022

Sonia Bergamaschi, Domenico Beneventano, Giovanni Simonini, Luca Gagliardelli, Adeel Aslam, Giulio De Sabbata, Luca Zecchini

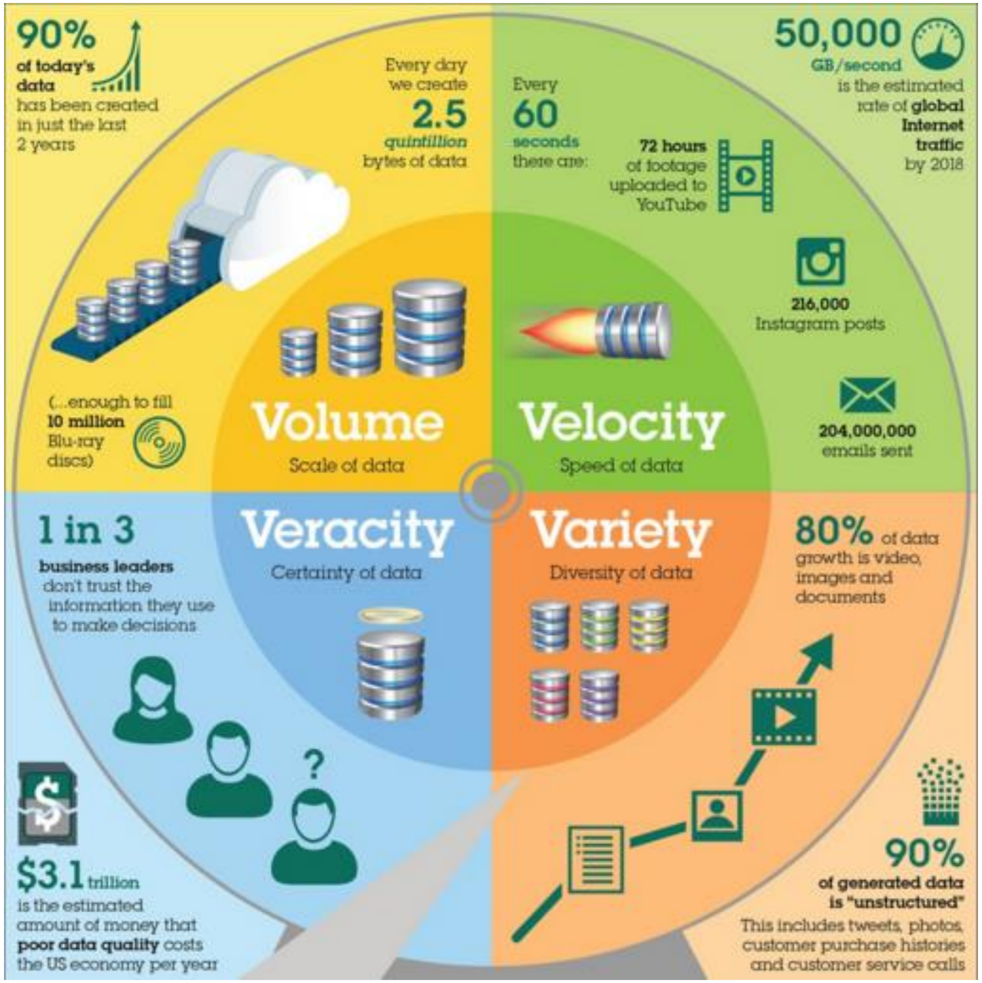
DBGroup @ Department of Engineering "Enzo Ferrari"

sonia.bergamaschi@unimore.it

www.dbgroup.unimore.it

The FOUR (+1) Vs of Big Data

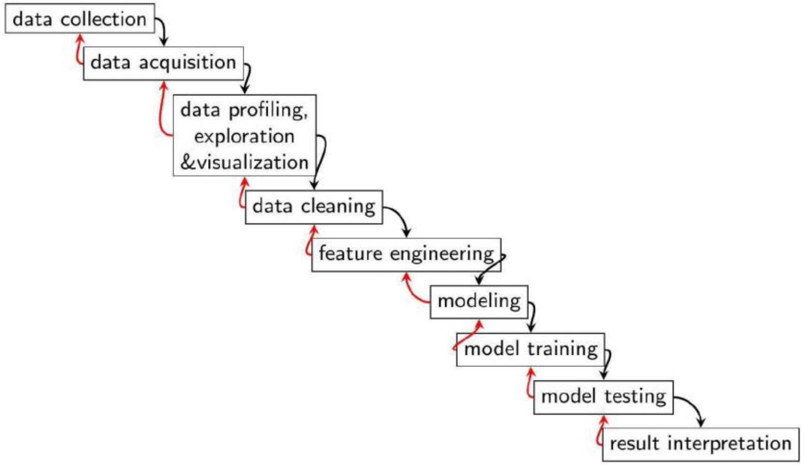
In today’s world, we need to deal with a huge amount of data…


90% of today's data has been created in just the last 2 years
Every day we create 2.5 quintillion bytes of data
(...enough to fill 10 million Blu-ray discs)
Volume
Scale of data
Every 60 seconds there are:
72 hours of footage uploaded to YouTube
216,000 Instagram posts
204,000,000 emails sent
50,000 GB/second is the estimated rate of global Internet traffic by 2018
Velocity
Speed of data
1 in 3 business leaders don't trust the information they use to make decisions
Veracity
Certainty of data
$3.1 trillion is the estimated amount of money that poor data quality costs the US economy per year
?
Variety
Diversity of data
80% of data growth is video, images and documents
90% of generated data is "unstructured"
This includes tweets, photos, customer purchase histories and customer service calls

Challenging for traditional paradigms…

- Big Data Management
- Big Data Science
- Big Data Integration

Big Data Science & Data-Centric AI



This is **at the same time** a process model **and** a dataflow.

Big Data Integration + Big Data Analysis

(BI, Statistics, Data Mining, Math + Data-Centric AI)

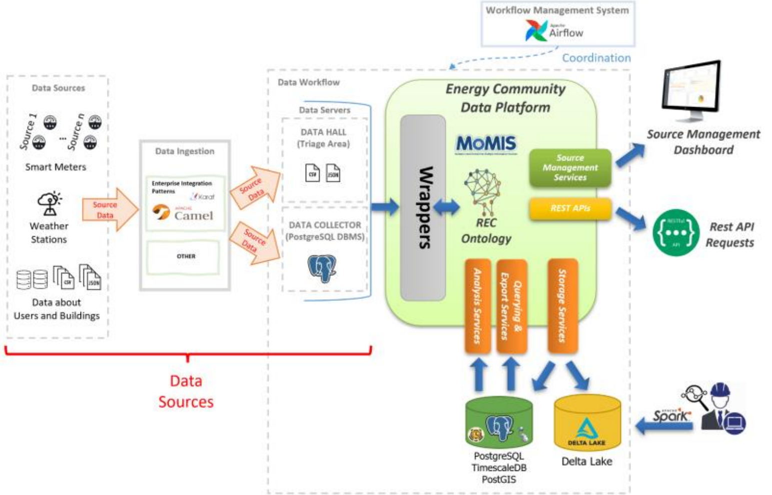
Model-Centric	Data-Centric
- Collects as much data as possible - Iteratively **improves the model** to deal with the noise in the data	- Holds the model fixed - Iteratively **improves the quality of the data** to obtain good results

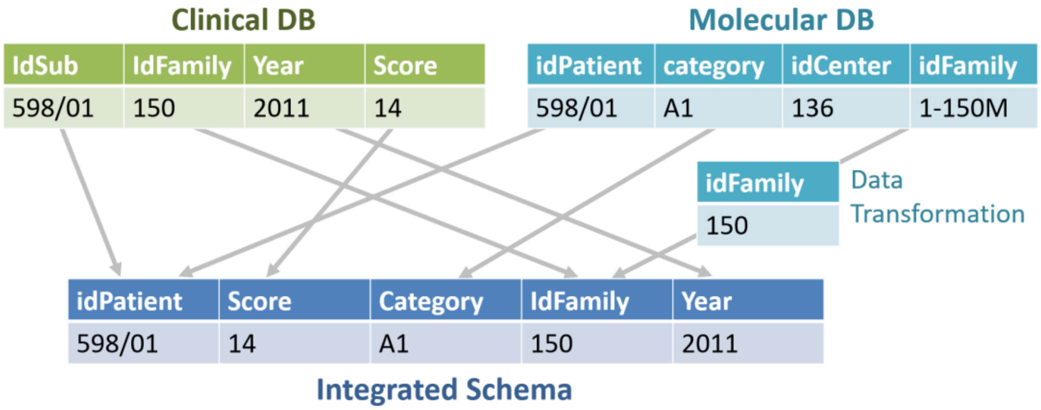
Andrew Ng

(Big) Data Integration with MOMIS

Data Integration is the process of consolidating data from a **set of heterogeneous data sources** into a **single uniform dataset** or view on the data.

MOMIS is an open-source (Big) Data Integration system able to aggregate data from heterogeneous and distributed sources in a semi-automatic way, exploiting the **semantic relationships** existing in the data sources.





Clinical DB

IdSub	IdFamily	Year	Score
598/01	150	2011	14

Molecular DB

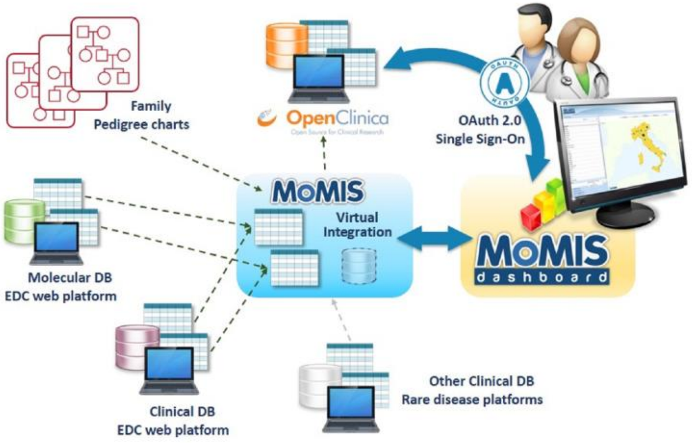
idPatient	category	idCenter	idFamily
598/01	A1	136	1-150M

idFamily
150

Data Transformation

idPatient	Score	Category	IdFamily	Year
598/01	14	A1	150	2011

Integrated Schema


Family Pedigree charts
OpenClinica
OAuth 2.0 Single Sign-On
MoMIS
Virtual Integration
MoMIS dashboard
Molecular DB EDC web platform
Clinical DB EDC web platform
Other Clinical DB Rare disease platforms

[1] S. Bergamaschi, S. Castano, M. Vincini: *Semantic Integration of Semistructured and Structured Data Sources*. SIGMOD Rec. 28(1): 54-59 (1999)

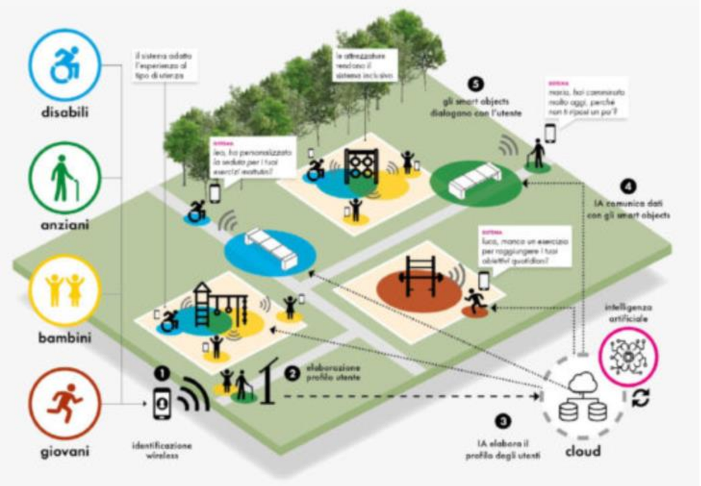
MoMIS
Mediator envirOnment for Multiple Information Sources

DataRiver
open source data management

(Big) Data Integration with MOMIS

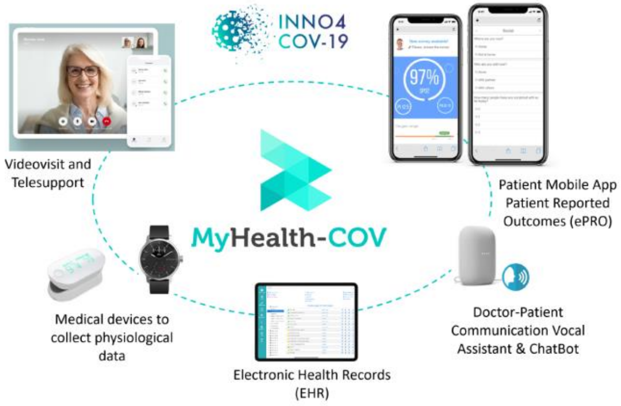
PleinAir
Parchi Liberi E Inclusivi in Network
per Attività Intergenerazionale Ricreativa e fisica

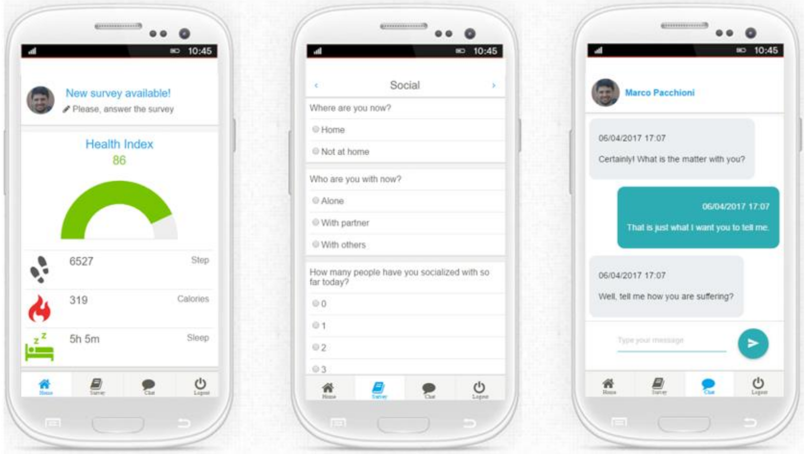
Creating contexts of smart inclusive parks to promote the adoption of correct active lifestyles and good health for all age groups.


disabili
anziani
bambini
giovani
il sistema adatta l'esperienza al tipo di utenza
le attrezzature rendono il sistema inclusivo
5 gli smart objects dialogano con l'utente
SISTEMA marco, hai camminato molto oggi, perché non ti riposi un po'?
SISTEMA leo, ho personalizzato la seduta per i tuoi esercizi mattutini?
4 IA comunica dati con gli smart objects
SISTEMA luca, manca un esercizio per raggiungere i tuoi obiettivi quotidiani!
intelligenza artificiale
1 identificazione wireless
2 elaborazione profilo utente
3 IA elabora il profilo degli utenti
cloud

OSO (Outdoor Smart Objects): furnishings and tools (IoMT) able to recognize the user and dynamically adapt their performance.

User activities monitored through AI techniques to provide a **personalized feedback** promoting physical well-being and health.


INNO4 COV-19
MyHealth-COV
Videovisit and Telesupport
Patient Mobile App Patient Reported Outcomes (ePRO)
Medical devices to collect physiological data
Electronic Health Records (EHR)
Doctor-Patient Communication Vocal Assistant & ChatBot
97%



10:45

New survey available!
Please, answer the survey

Health Index
86

6527 Step
319 Calories
5h 5m Sleep

10:45

Social

Where are you now?
○ Home
○ Not at home

Who are you with now?
○ Alone
○ With partner
○ With others

How many people have you socialized with so far today?
○ 0
○ 1
○ 2
○ 3

10:45

Marco Pacchioni

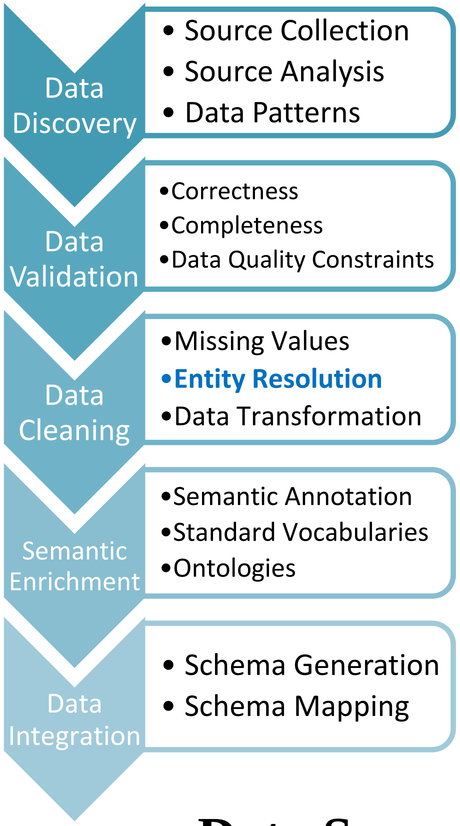
06/04/2017 17:07
Certainly! What is the matter with you?

06/04/2017 17:07
That is just what I want you to tell me.

06/04/2017 17:07
Well, tell me how you are suffering?

Type your message

Entity Resolution (ER) a.k.a. Record Linkage (RL)



Data Discovery
- Source Collection
- Source Analysis
- Data Patterns

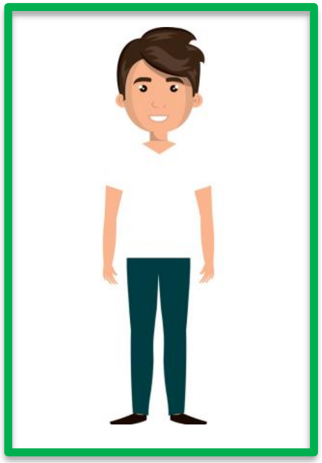
Data Validation
- Correctness
- Completeness
- Data Quality Constraints

Data Cleaning
- Missing Values
- **Entity Resolution**
- Data Transformation

Semantic Enrichment
- Semantic Annotation
- Standard Vocabularies
- Ontologies

Data Integration
- Schema Generation
- Schema Mapping

Given one or more data sources, Entity Resolution (ER) is the task of identifying the **records (entity profiles)** that refer to the **same real-world object (entity)**.



Data Source A

	Name	Surname	Address	Sex
r1	Mary-Ann	White	West Main Street 29, 12068, Fonda, NY, New York	F
r2	Thomas J.	Franklin	50 Liverpool Street, London	M

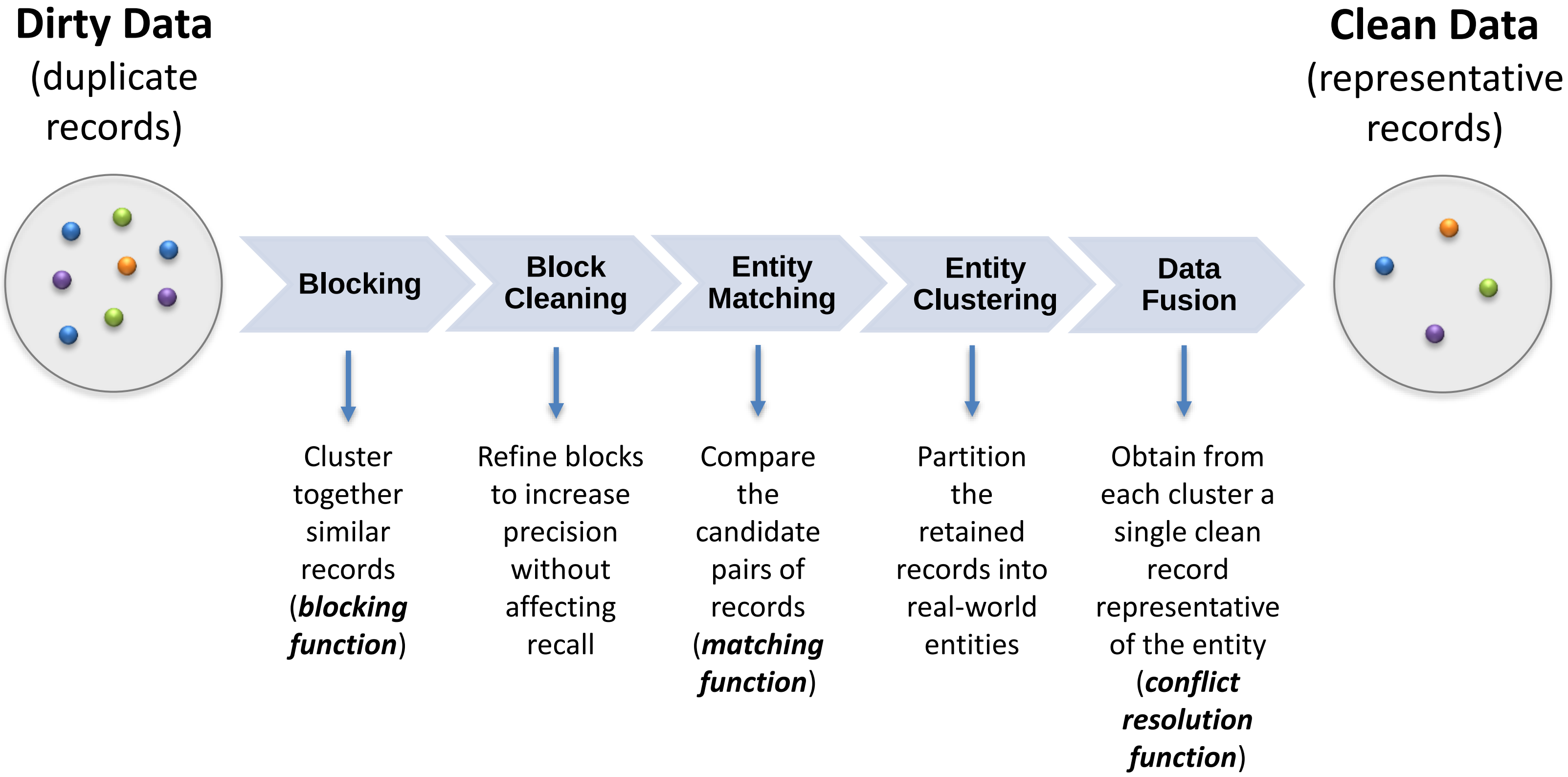
Data Source B

Name	Residence	Age	Gender	
Franklin, Tom	London (UK)	25	Male	r3
Withe, Mary Ann	New York (USA)	29	Female	r4

Entity Resolution Pipeline

ER is a challenging task:

- Hard to scale (quadratic complexity) → Blocking functions
- Hard to detect duplicates → Matching functions
- Even harder with privacy constraints → PPRL



Dirty Data (duplicate records)

Blocking → Cluster together similar records (***blocking function***)

Block Cleaning → Refine blocks to increase precision without affecting recall

Entity Matching → Compare the candidate pairs of records (***matching function***)

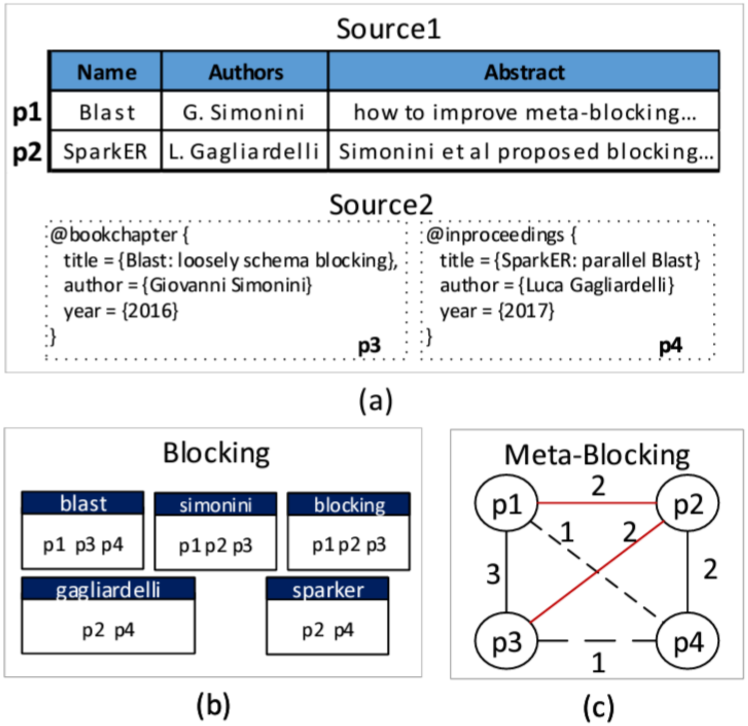
Entity Clustering → Partition the retained records into real-world entities

Data Fusion → Obtain from each cluster a single clean record representative of the entity (***conflict resolution function***)

Clean Data (representative records)

Blocking

Reducing the number of candidate pairs by discarding obvious non-matches.



(a)

(b)

(c)

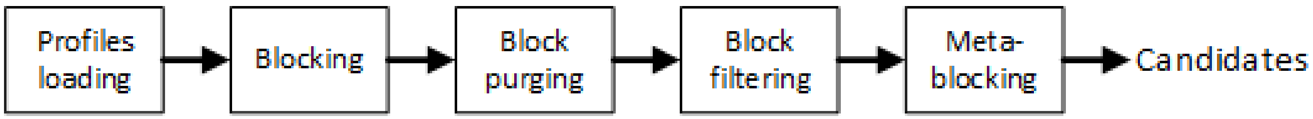
[1] G. Simonini, S. Bergamaschi, H. V. Jagadish: *BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution*. PVLDB 9(12): 1173-1184 (2016)

[2] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. V. Jagadish: *Scaling entity resolution: A loosely schema-aware approach*. Inf. Syst. 83: 145-165 (2019)

[3] L. Gagliardelli, G. Simonini, D. Beneventano, S. Bergamaschi: *SparkER: Scaling Entity Resolution in Spark*. EDBT, 602-605 (2019)

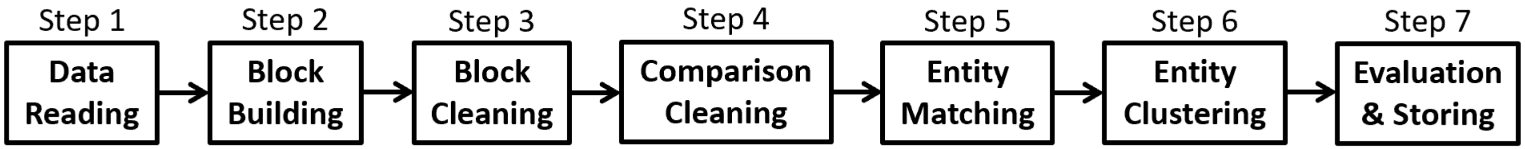
[4] L. Gagliardelli, G. Papadakis, G. Simonini, S. Bergamaschi, T. Palpanas: *Generalized Supervised Meta-blocking*. PVLDB 15(9): 1902-1910 (2022)

SparkER is an ER framework for Apache Spark. Its source code is available on GitHub and its complete and detailed documentation can be found on Read the Docs.



JedAI: An End-to-End ER Framework

JedAI implements the whole **schema-agnostic, end-to-end workflow** for ER and can be used as an open-source library, as a desktop application, or as a workbench.



Step 1 **Data Reading**: Reads files containing the entity profiles and the golden standard.

Step 2 **Block Building**: Creates overlapping blocks.

Step 3 **Block Cleaning**: Optional step that cleans blocks from useless comparisons (repeated, superfluous).

Step 4 **Comparison Cleaning**: Optional step that operates on the level of individual comparisons to remove the useless ones.

Step 5 **Entity Matching**: Executes all retained comparisons.

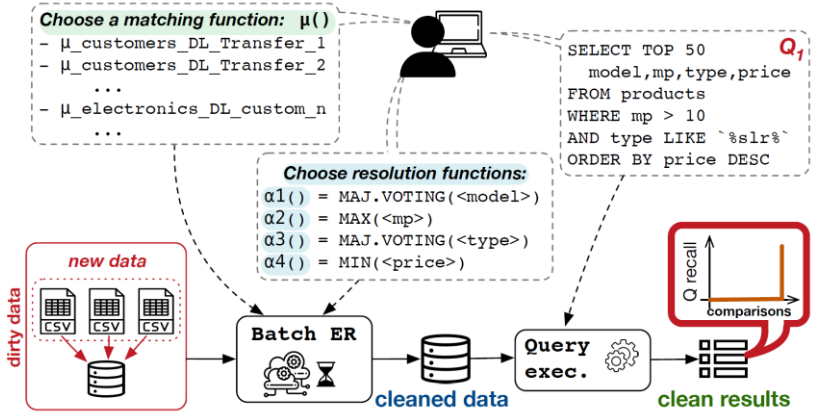
Step 6 **Entity Clustering**: Partitions the similarity graph into equivalence clusters.

Step 7 **Evaluation & Storing**: Stores and presents performance results w.r.t. numerous measures.

[1] G. Papadakis, G. Mandilaras, L. Gagliardelli, G. Simonini, E. Thanos, G. Giannakopoulos, S. Bergamaschi, T. Palpanas, M. Koubarakis: *Three-dimensional Entity Resolution with JedAI*. Inf. Syst. 93: 101565 (2020)

- Project website: http://jedai.scify.org
- GitHub repository: https://github.com/scify/JedAIToolkit

Beyond Traditional Batch ER: Progressive Approaches



Choose a matching function: μ()

```
- μ_customers_DL_Transfer_1
- μ_customers_DL_Transfer_2
    ...
- μ_electronics_DL_custom_n
    ...
```

Choose resolution functions:

```
α1() = MAJ.VOTING(<model>)
α2() = MAX(<mp>)
α3() = MAJ.VOTING(<type>)
α4() = MIN(<price>)
```

Q_1

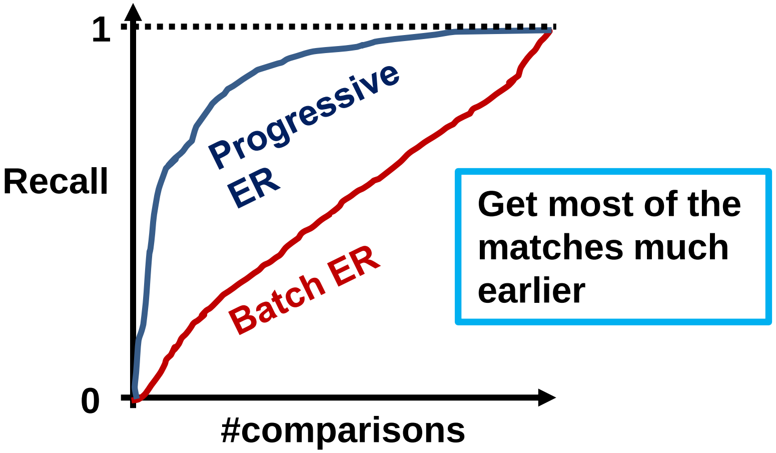
```
SELECT TOP 50
  model,mp,type,price
FROM products
WHERE mp > 10
AND type LIKE `%slr%`
ORDER BY price DESC
```

Useless comparisons (produce entities that will surely not appear in the result of the query)

Our time, resources and affordable costs (e.g., pay-as-you-go in cloud) are often **limited**

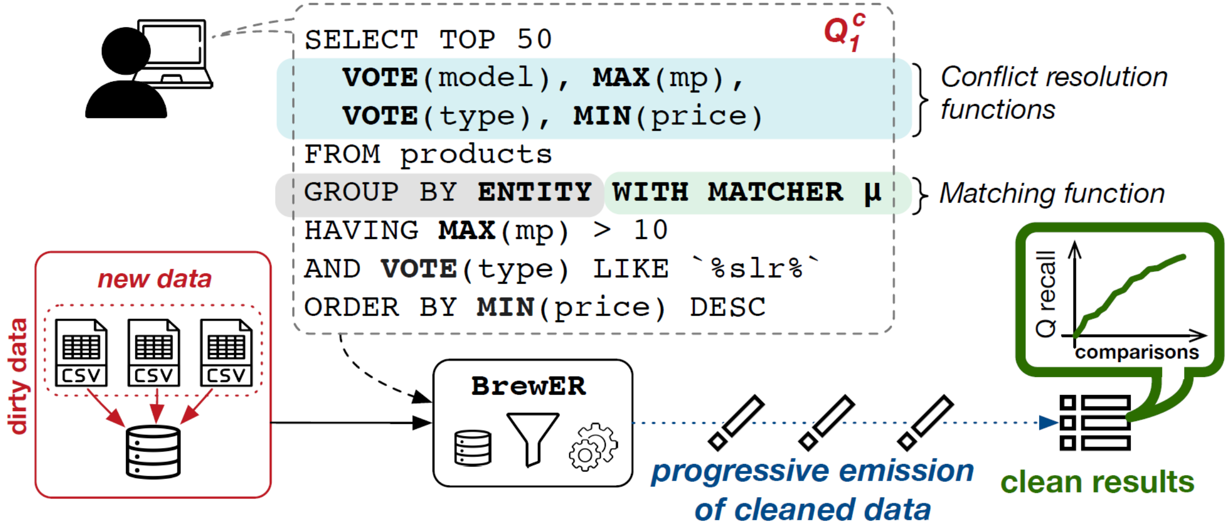
PROGRESSIVE ER: Maximize the number of retrieved **matches** in a limited amount of time (driven by **matching likelihood)**

[1] G. Simonini, G. Papadakis, T. Palpanas, S. Bergamaschi: Schema-agnostic Progressive Entity Resolution. ICDE: 53-64 (2018)



Get most of the matches much earlier

BrewER: Entity Resolution On-Demand



```
SELECT TOP 50
  VOTE(model), MAX(mp),
  VOTE(type), MIN(price)
FROM products
GROUP BY ENTITY WITH MATCHER μ
HAVING MAX(mp) > 10
AND VOTE(type) LIKE `%slr%`
ORDER BY MIN(price) DESC
```

Agnostic approach to blocking and matching functions

Clean queries on dirty data

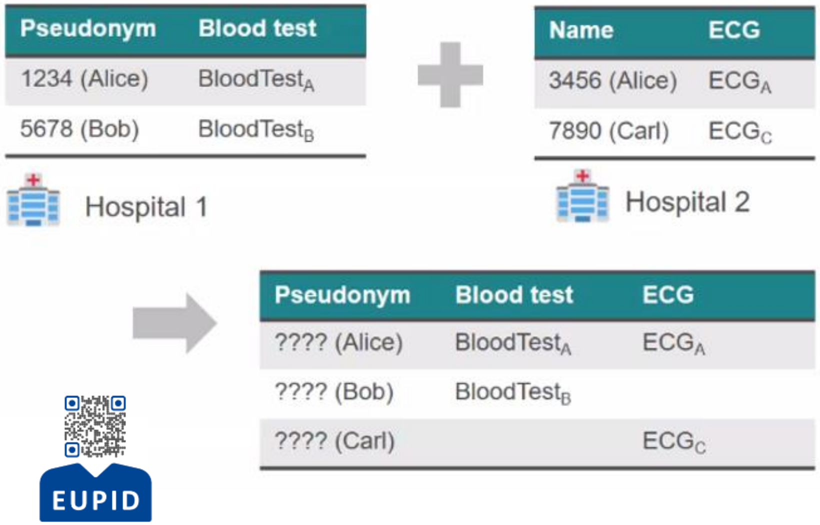
- ❑ **QUERY-DRIVEN**: ER only on the portion of dataset useful to answer the query (according to the HAVING clauses)
- ❑ **PROGRESSIVE**: return the entities in the result in the right order as soon as they are obtained (according to the ORDER BY clause)

[1] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann: *Entity Resolution On-Demand*. PVLDB 15(7): 1506-1518 (2022)

VLDB 2022

Privacy-Preserving Record Linkage (PPRL)

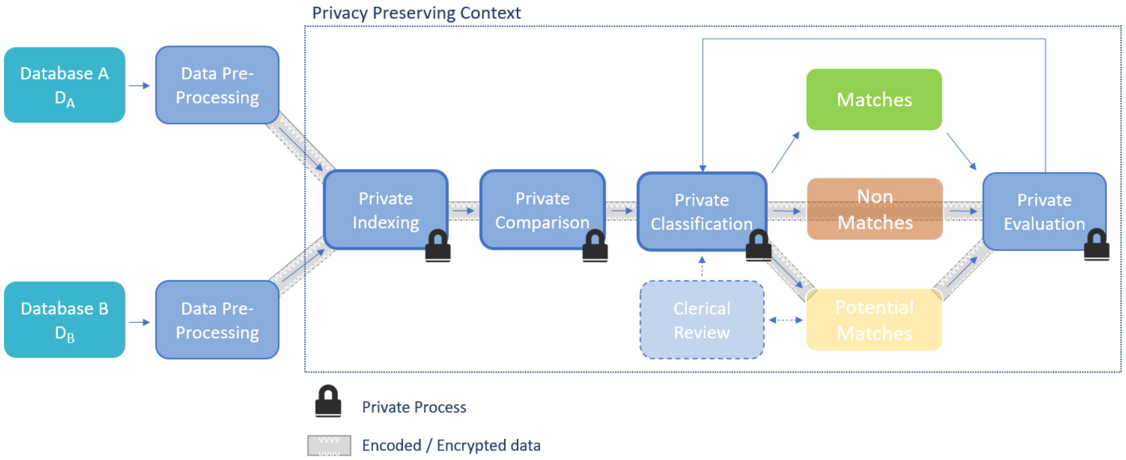
Identifying and linking the **records (profiles)** that refer to the **same real-world object (entity),** across several data sources held by different parties, in a manner that prevents both the computation and the output of the computation from revealing (to any **internal parties** involved in the process and **external adversaries**) any private sensitive information about the entities represented in the data.



Pseudonym	Blood test
1234 (Alice)	BloodTest$_A$
5678 (Bob)	BloodTest$_B$

Name	ECG
3456 (Alice)	ECG$_A$
7890 (Carl)	ECG$_C$

Pseudonym	Blood test	ECG
???? (Alice)	BloodTest$_A$	ECG$_A$
???? (Bob)	BloodTest$_B$	
???? (Carl)		ECG$_C$



Privacy Preserving Temporal Record Linkage Project: design and implement in prototype form PPTRL techniques as part of the "Recidivism Data Mart and Criminal Data Warehouse" project, funded by the CRUI Foundation, which is part of the broader "PNRR Justice Plan Data Lake".

THANK YOU FOR THE ATTENTION!