UNIVERSITÀ DEGLI STUDI DI TRENTO

soweego

solid catalogs and **wee**kee **go** together

Studente
Massimo Frasson
mat. 180868

Supervisori
Andrea Passerini
Marco Fossati

Contesto

WIKIPEDIA

Uno dei siti più visitati al mondo

WIKIDATA

Banca dati nata per alimentare i dati strutturati di Wikipedia

WIKIMEDIA FOUNDATION

ONLUS che mantiene l'infrastruttura e gestisce la comunità

FONDAZIONE BRUNO KESSLER

unità Future Media ha proposto e vinto il progetto soweego

🤔 Cos’è Wikidata?

- **Base di dati strutturata** per i progetti **Wikimedia** (es. Wikipedia)
- Contiene più di **50 milioni di entità**

🤒 Il problema

Delle **asserzioni** all'interno di **Wikidata**:

- Circa **metà** sono **senza riferimenti**
- **Un quarto** ha **riferimenti** a **risorse interne** ai progetti Wikimedia

🤔 Asserzioni?

ID: Q7351595

labels:

- (“en”, “Roberto Battiti”)
- (“azb”, “ روبرتو باتتی”)

[…]

(“Data di nascita”, “1961”)

- English Wikipedia

(“VIAF ID”, “65066797”)

- English Wikipedia

[…]

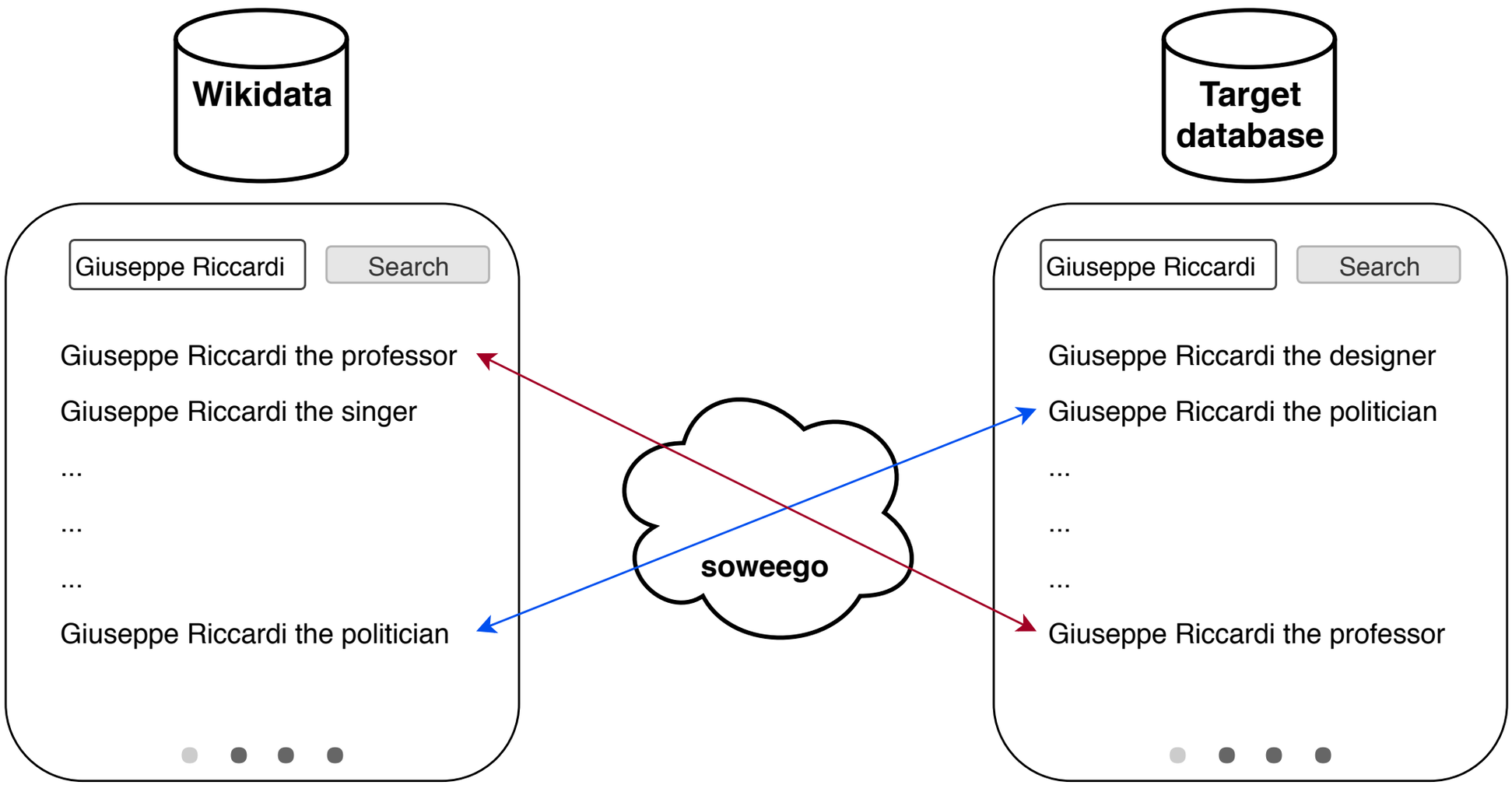
Asserzioni
e relativi riferimenti

👨🏻‍⚕️ La nostra soluzione

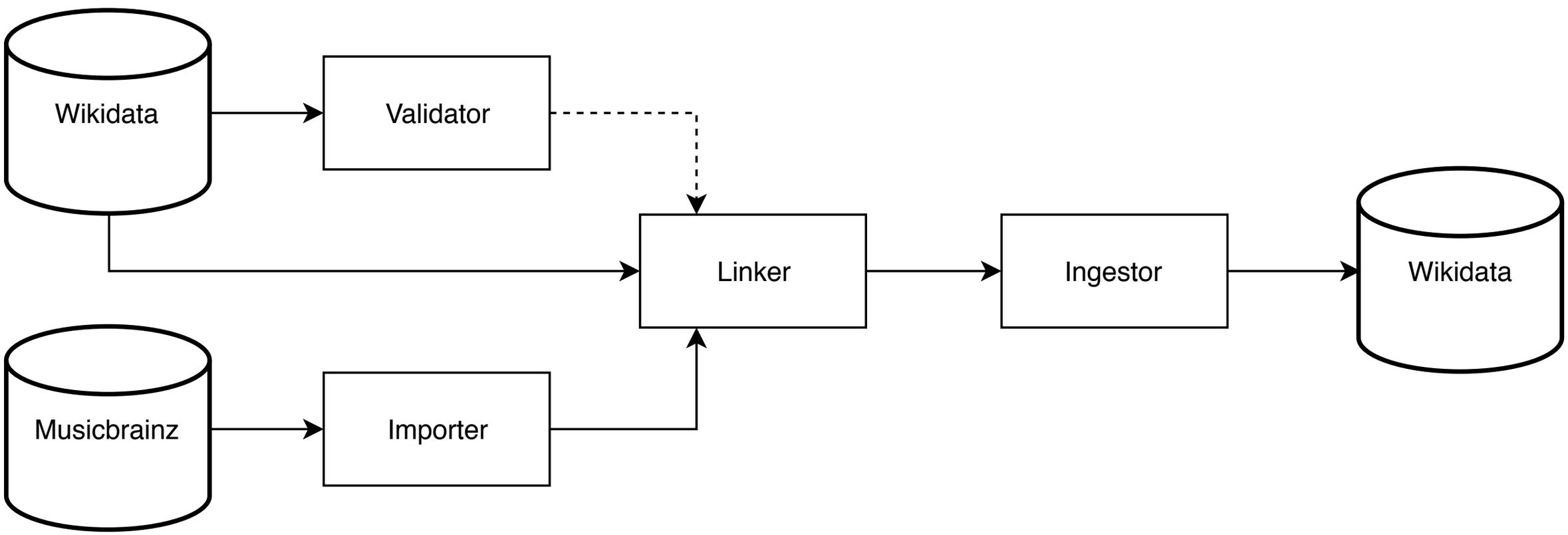
1. Collegare le persone di Wikidata a persone di altri database (identificativi esterni)

2. Estrarre le informazioni da questi database, così da poterli indicare come fonti (asserzioni referenziate)

🔍 Esempio dello use case



Architettura



Linker:
approccio a regole

Uguaglianza di stringhe

George Bush

George Bush

Uguaglianza di stringhe e date

George Bush

12 Giugno 1924

George Bush

6 Luglio 1946

Normalizzazione

Vladimir Vladimirovič Putin

[‘vladimir, ‘vladimirovic’, ‘putin’]

Владимир Владимирович Путин

[‘vladimir, ‘vladimirovic’, ‘putin’]

Cross-database link match

https://twitter.com/GeorgeHWBush

[‘twitter’, ‘GeorgeHWBush’]

http://mobile.twitter.com/GeorgeHWBush/

[‘twitter’, ‘GeorgeHWBush’]

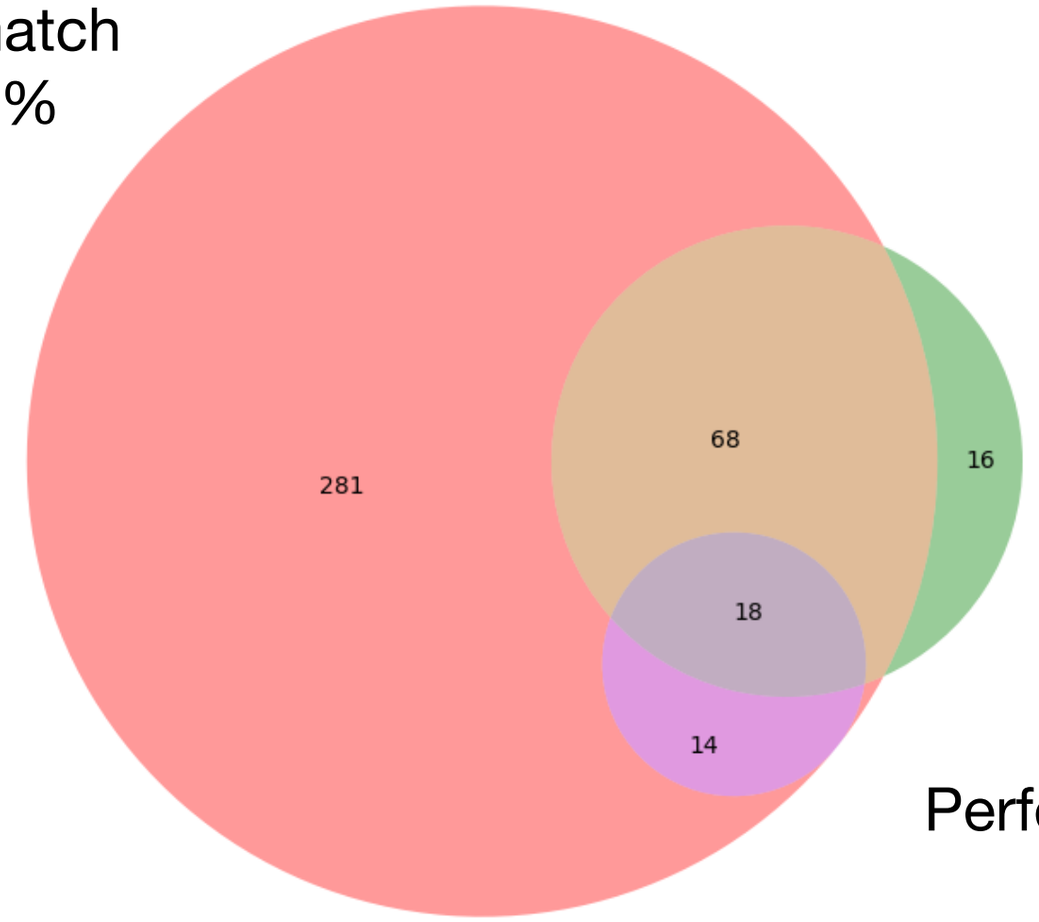
Risultati

Strategy	Total matches	Checked	Precision
Perfect names	381	38(10%%)	84,2%
Perfect names + dates	32	32 (100%)	100%
Normalized names	227	24(10,6%)	70,8%
Perfect Link	71	71(100%)	100%
Token Link	102	102(100%)	99%

- Campione di **Wikidata** composto da **1100 musicisti** (1% dei musicisti non collegati a Musicbrainz in Wikidata)
- Dump di Musicbrainz **986 765 artisti** (o presunti tali) e rispettivi **181 221** alias

Risultati

Perfect name match
precisione 84%



Tokenized link match
precisione 99%

Perfect name and dates match
precisione 100%

Intersezione tra gli insiemi di **match prodotti** dalle varie tecniche

Futuri sviluppi

1. 🚀 **Miglioramento** delle attuali **strategie**

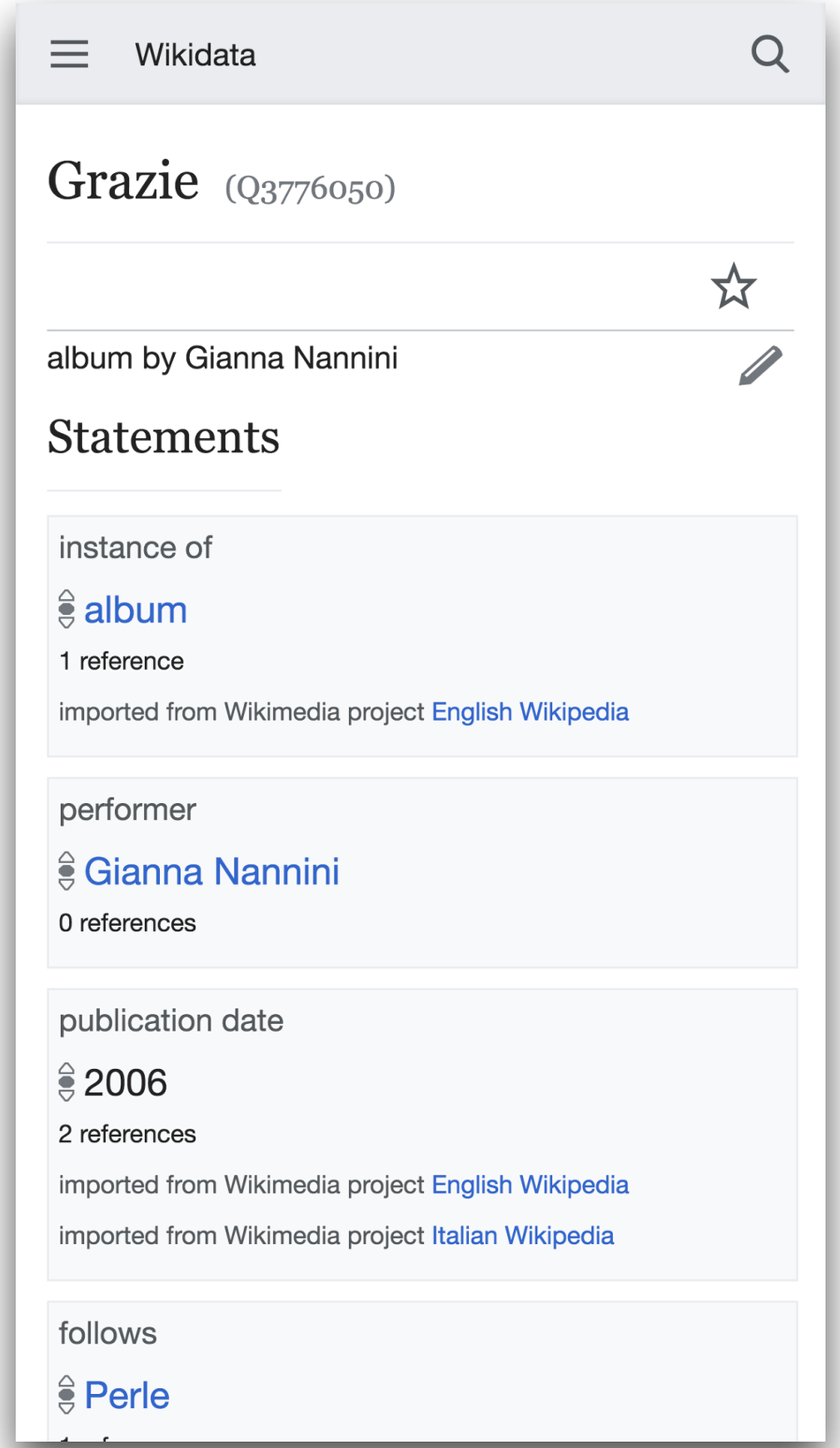
2. 💻 Valutazione di tecniche di **machine learning** per il Record Linkage

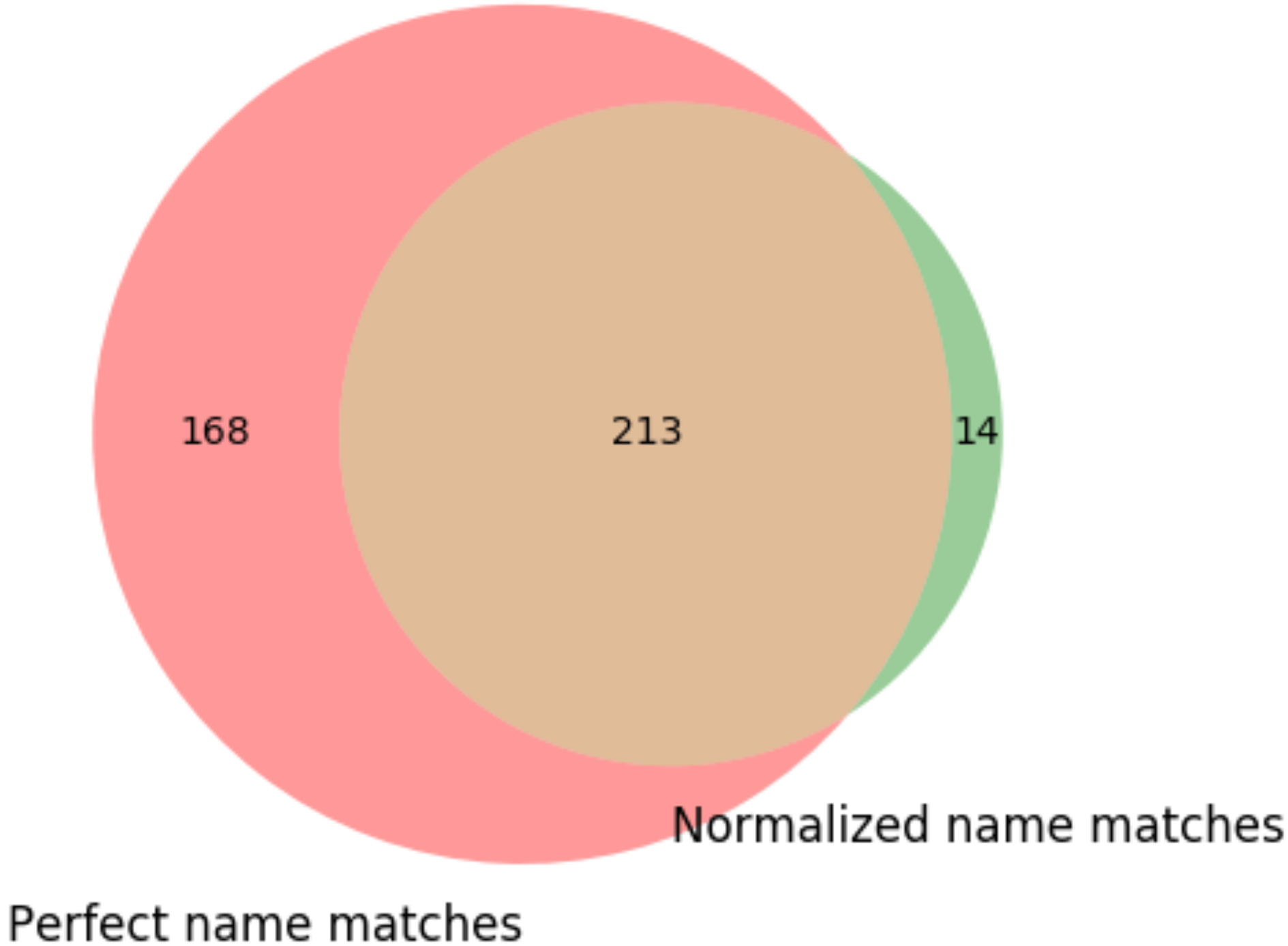
- Le **tecniche** esplorate possono diventare **features**

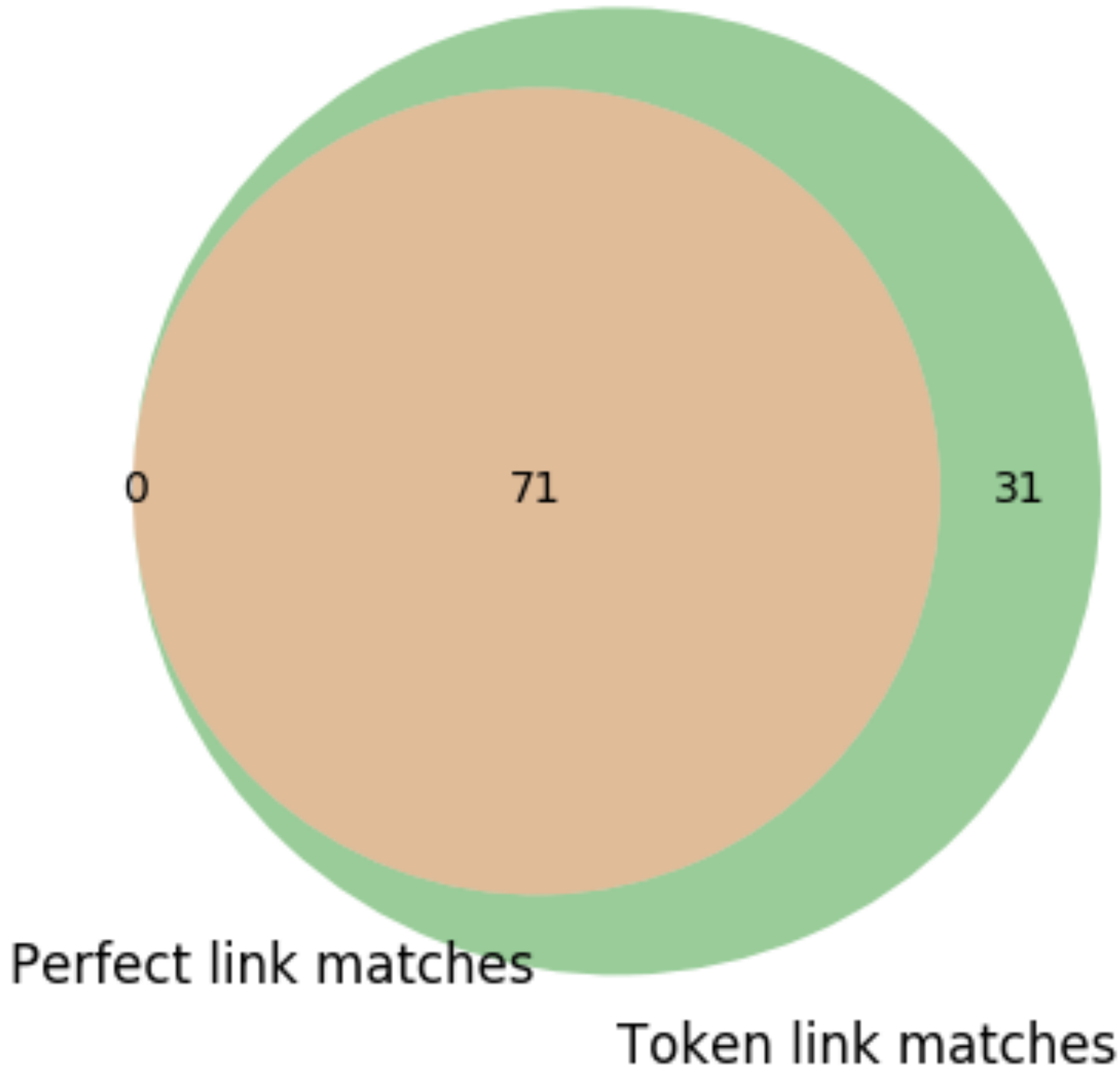
- Le **tecniche** esplorate fanno da **base di valutazione** per le performance del machine learning

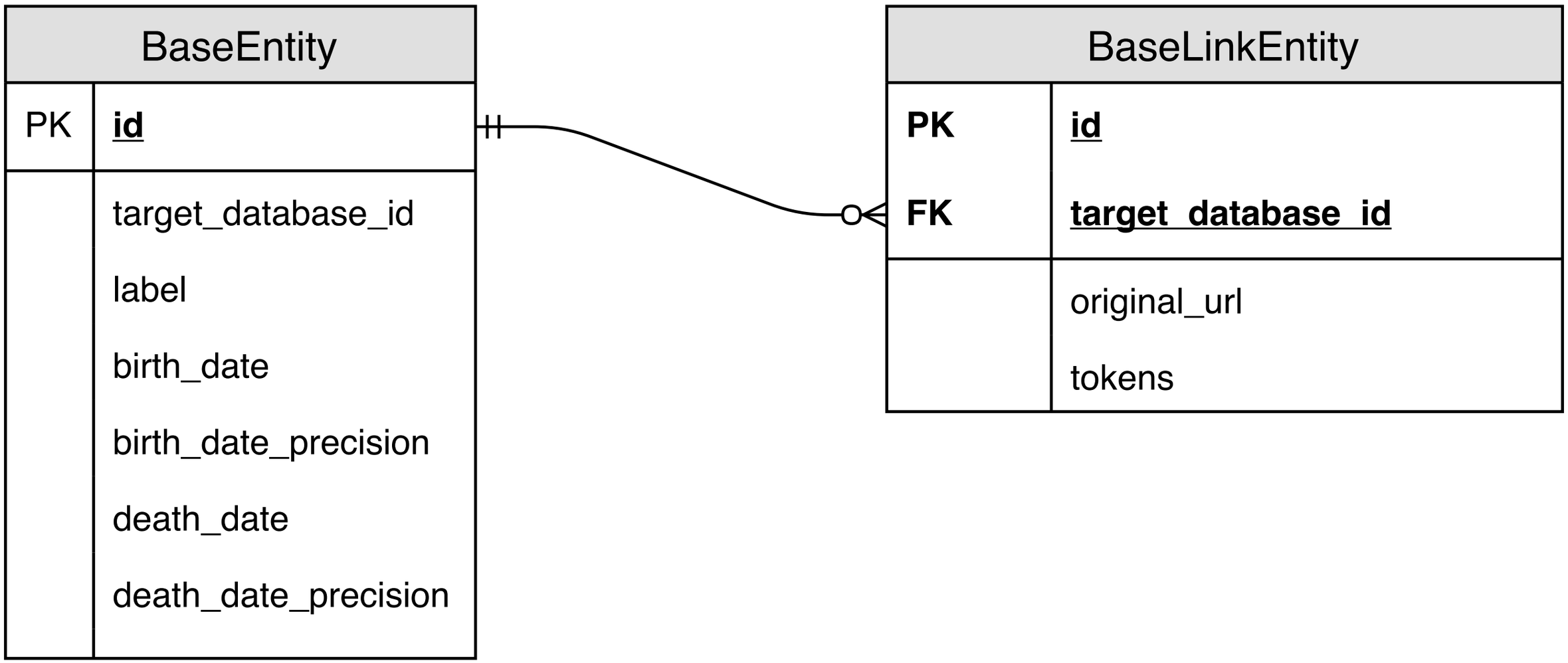
🙇🏻 GRAZIE

Q3776050




168
213
14
Normalized name matches
Perfect name matches


0
71
31
Perfect link matches
Token link matches


BaseEntity
PK
id
target_database_id
label
birth_date
birth_date_precision
death_date
death_date_precision
BaseLinkEntity
PK
id
FK
target_database_id
original_url
tokens