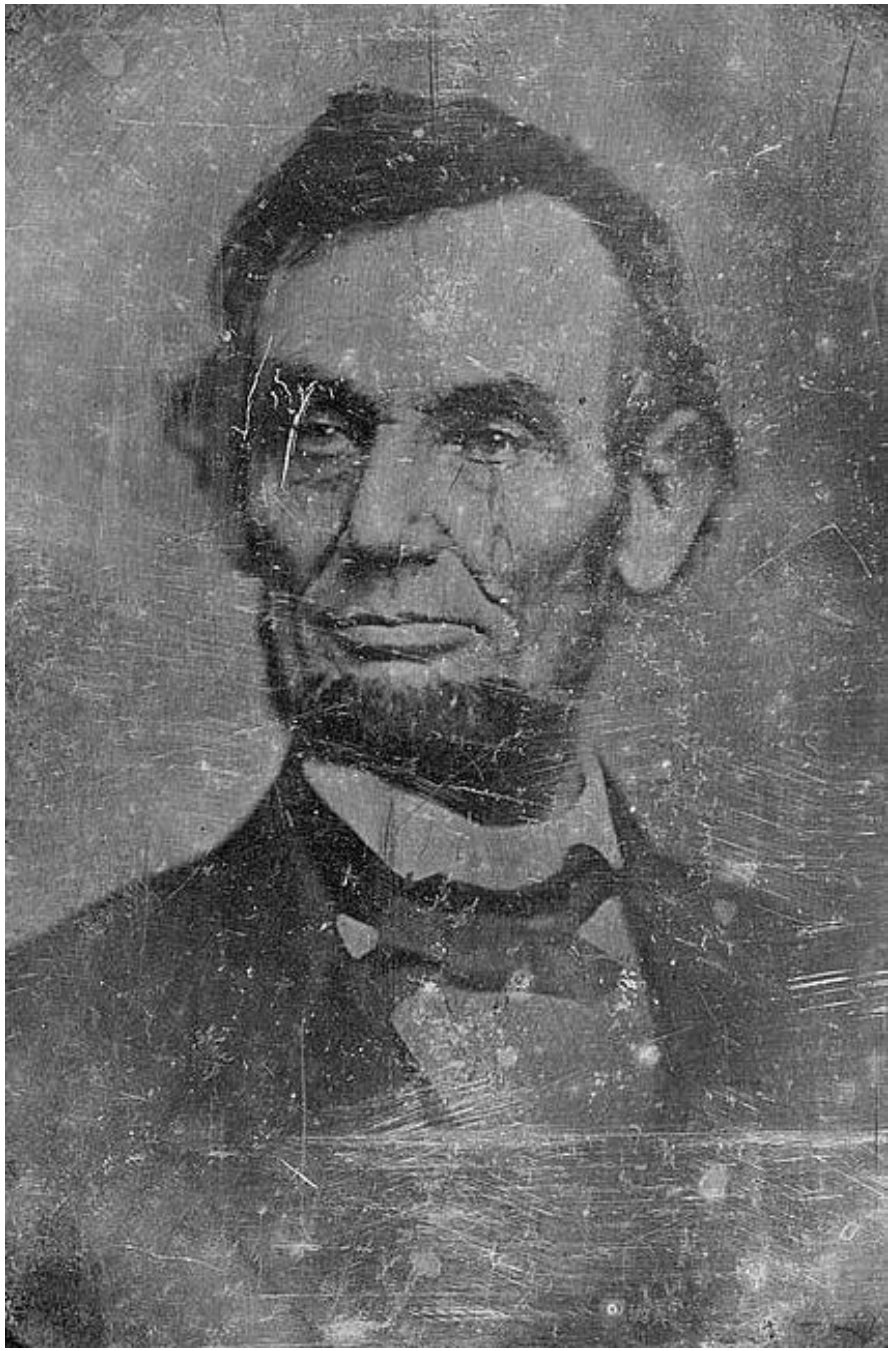
BioGraph

Connect, enrich and explore diverse biographical collections

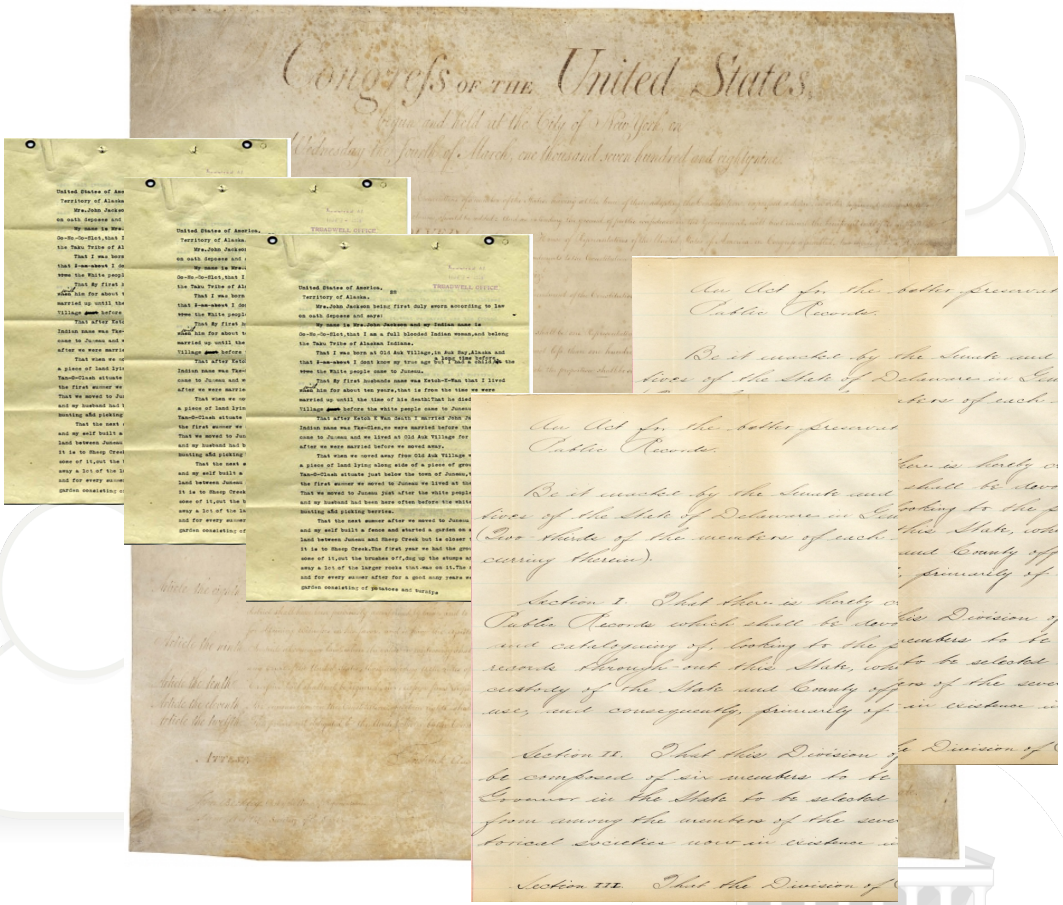
Krishna Janakiraman
Sean Marimpietri

Advisor: Ray Larson



Find all cabinet officers associated with Abraham Lincoln who were involved with slavery

How is Groucho Marx related to Cold War technology ?



Historical archives and their descriptions are host to rich latent social networks

SNAC

Social Networks and Archival Context Project

Our primary collection of archival description records

We developed tools to make the SNAC collection more accessible

More accessible

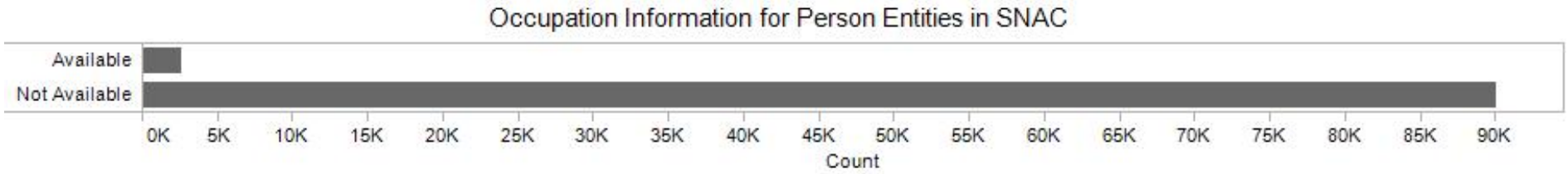
We created tools to **link SNAC with online, crowdsourced resources**

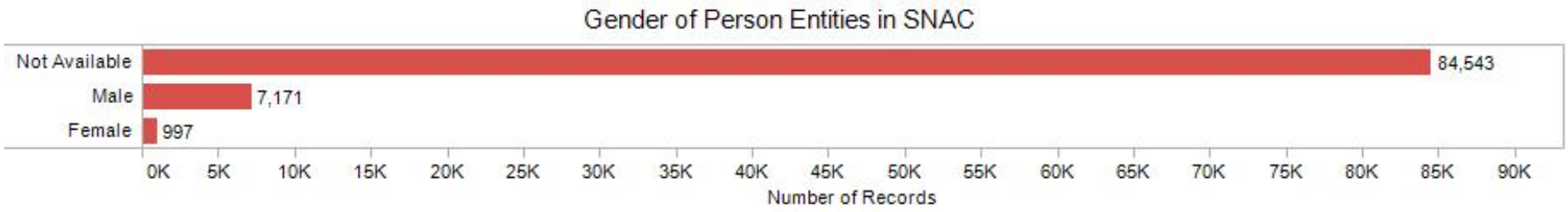
We developed a **new visual interface to discover latent social structures in the SNAC collection**

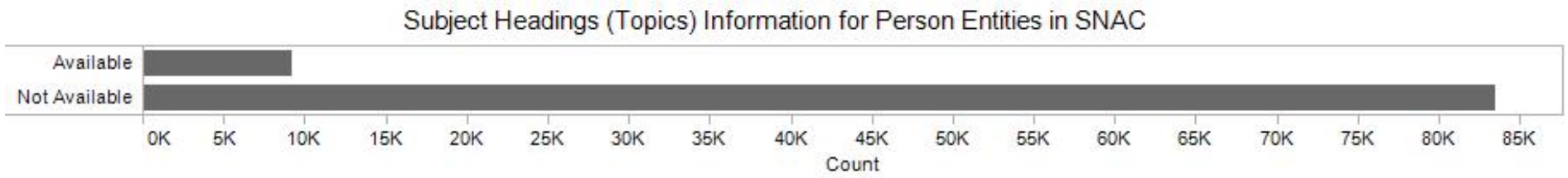
Why link ?

Users of the linked collections can make use of the historic social network that SNAC offers

Archival descriptions are primarily used as finding aids. Hence there is significant information gap in the biographical facts







Missing facts in SNAC

Why new interface?

Existing faceted interface is not ideal for discovering latent social structures

Resources

SNAC

Social Networks and Archival Context (SNAC)

Collection of biographical information **derived from archive description records**

Curated by professional archivists

Collection is encoded using the EAC-CPF standard

Freebase

Collection derived from various sources: Wikipedia, MusicBrainz and many others

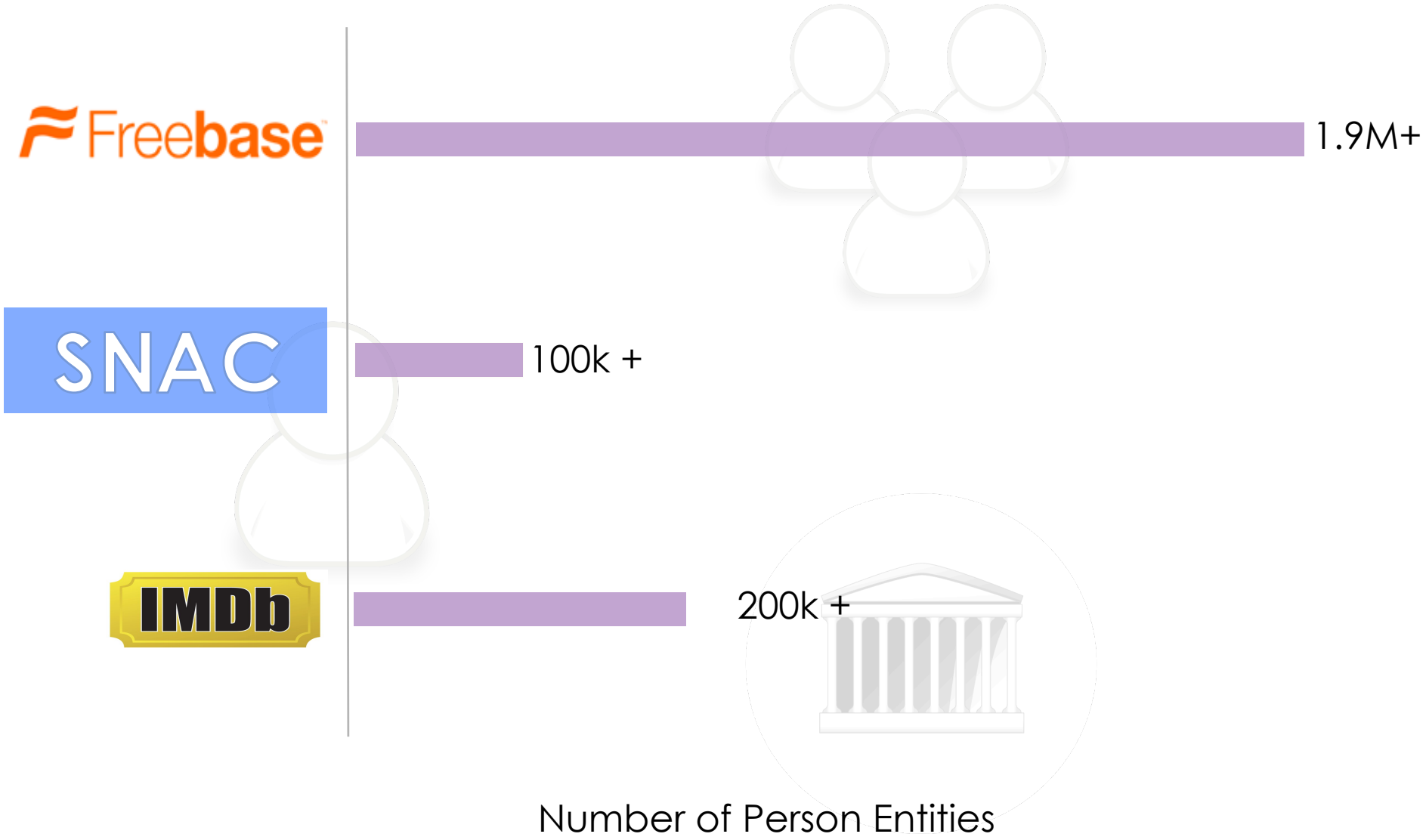
Information is **crowdsourced**. Schema is editable.

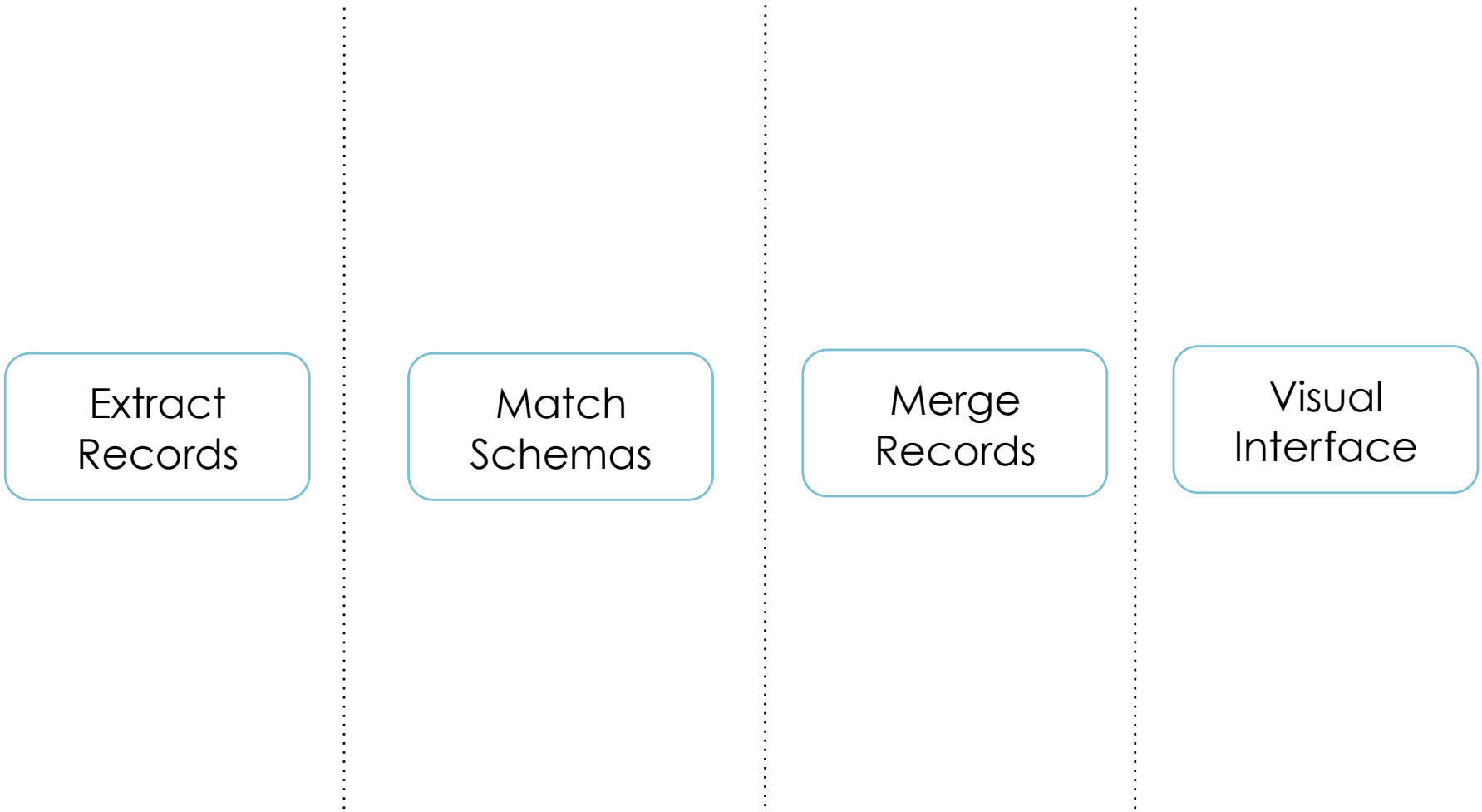
Collection is encoded using Freebase's quad store format

Collection originated as lists kept by film enthusiasts.

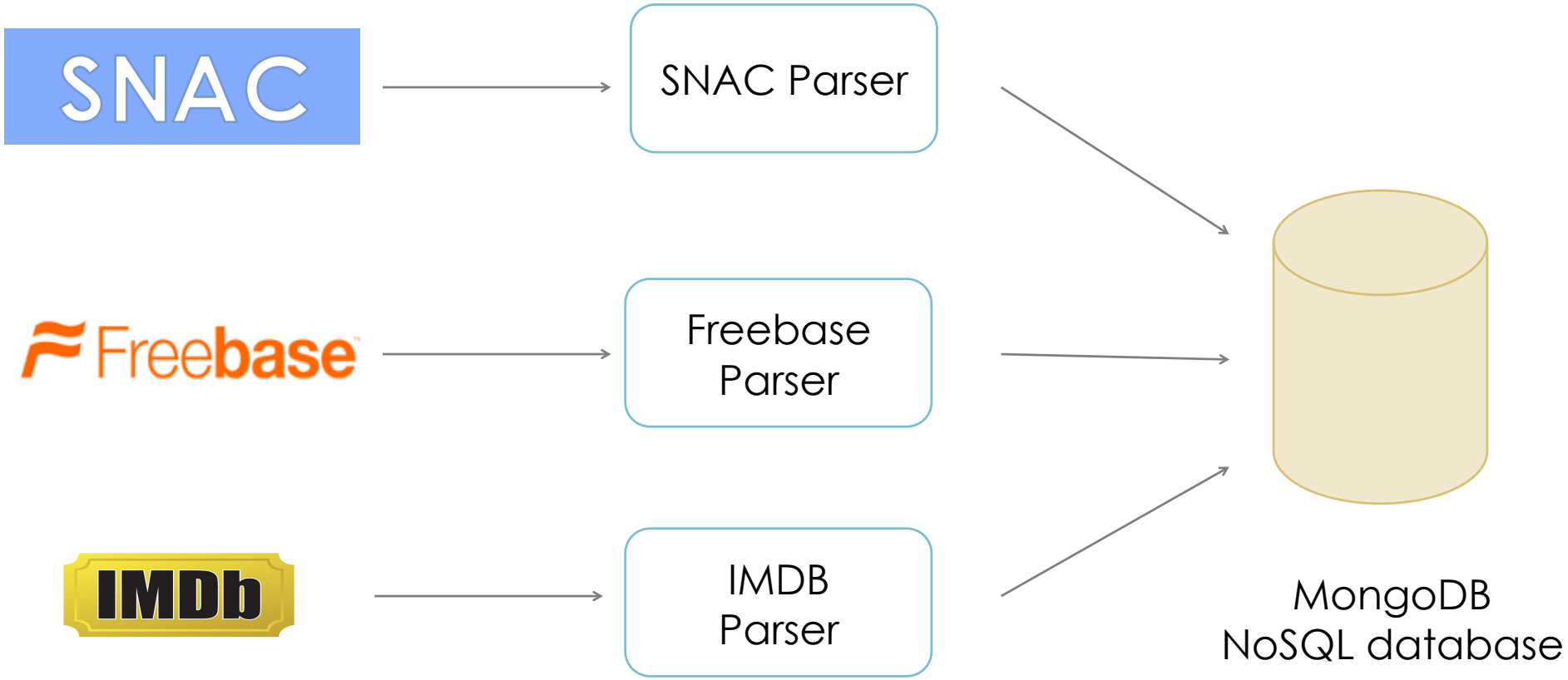
All user submissions are reviewed. **Schema is fixed**.

Collection is encoded in plaintext.

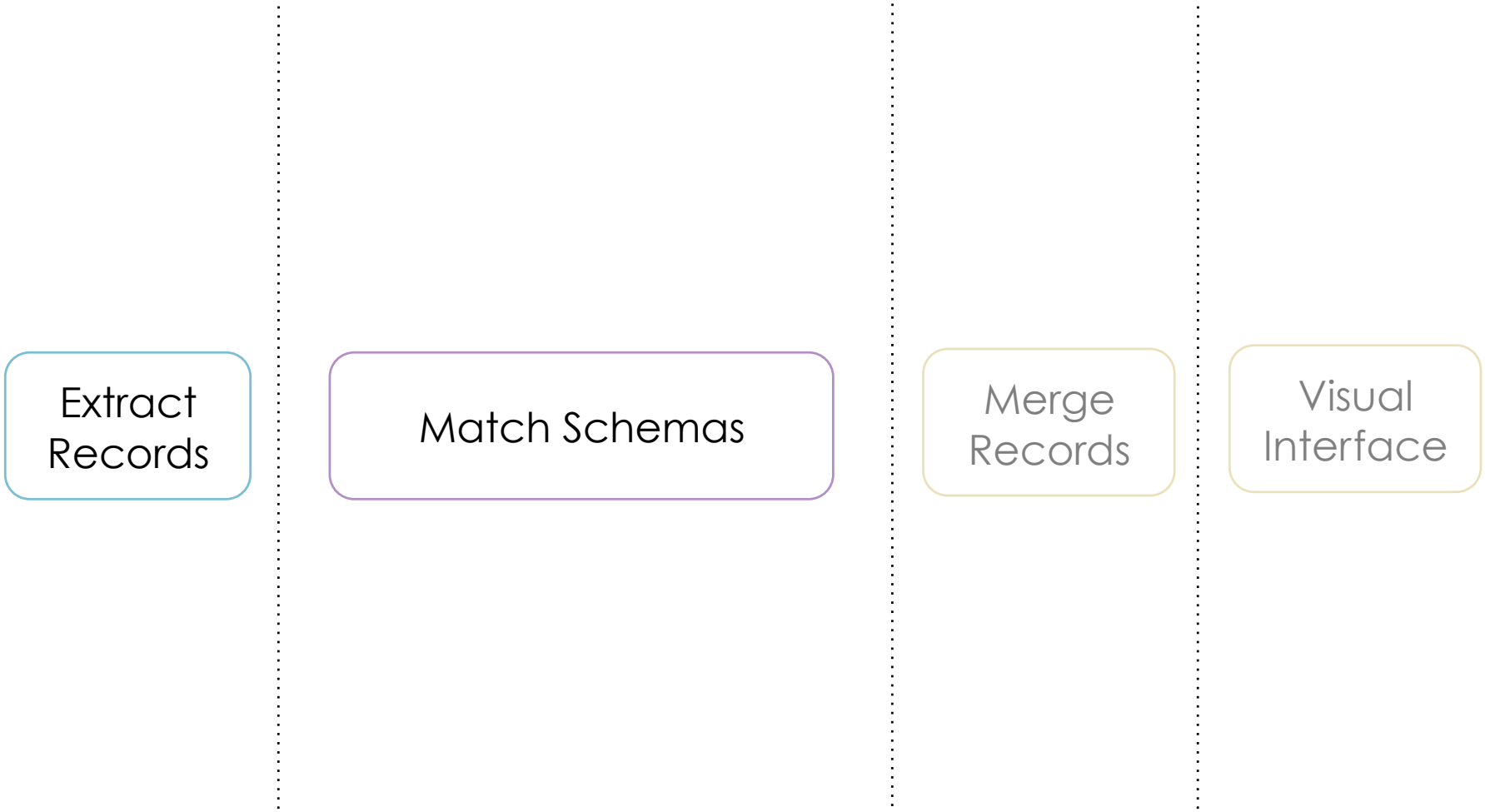

Freebase
1.9M+
SNAC
100k +
IMDb
200k +
Number of Person Entities



Architecture



For each collection, we created an entity table and an assertion table



Match Schemas

Develop a tool that given a field in the SNAC collection **can recommend fields** from other collections that might be a match

Goal

SNAC	Freebase	IMDb
namePart/part	/type/object/name	**NM**
existDates/dateRange/ fromDate	/people/person/ date_of_birth	DB
localDescription/ localType-VIAF:gender/ term	/people/person/gender	No unique field

The schema matching problem

Our approach

Represent each field **as a vector of terms**

Terms are derived from the **name of the field**

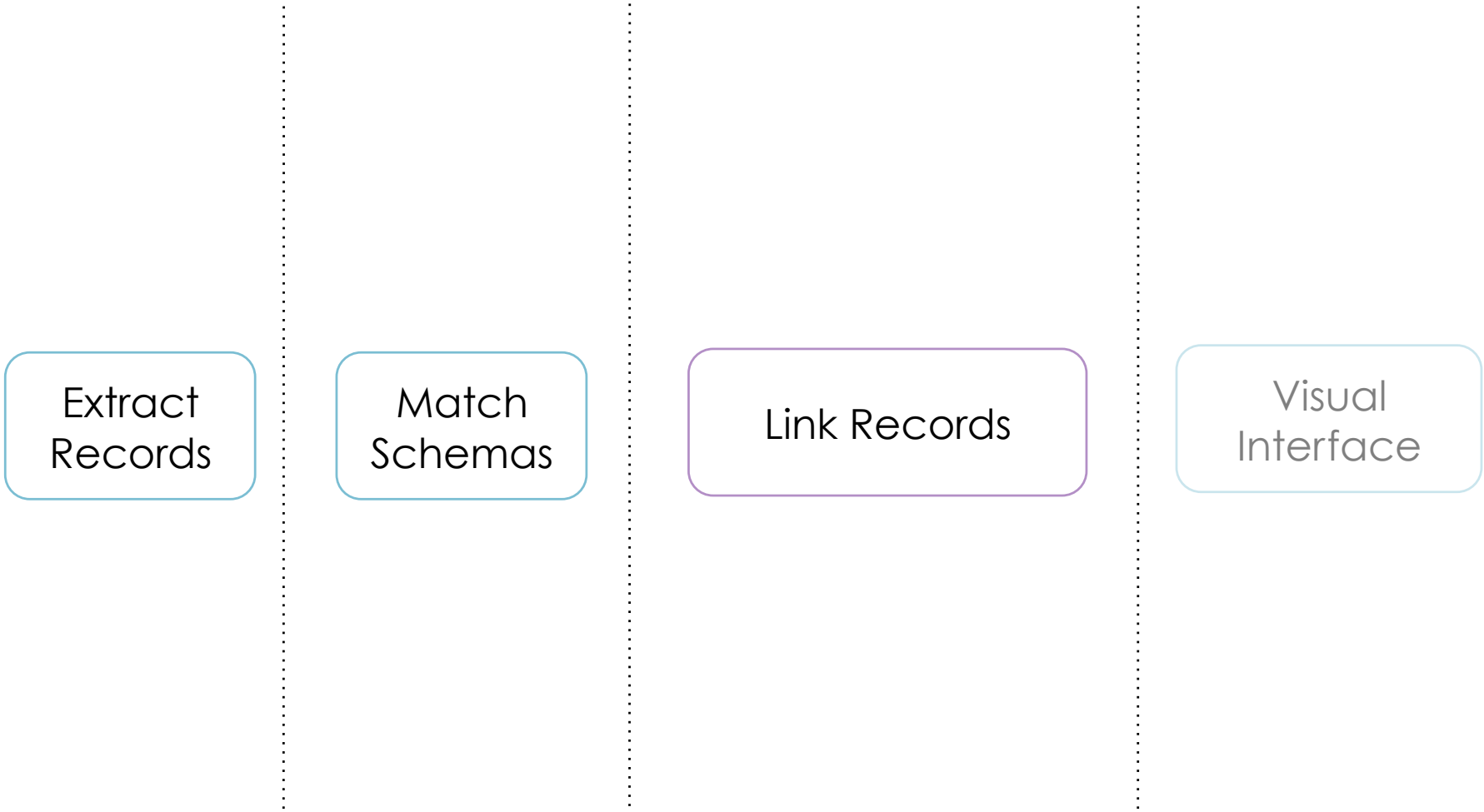
Terms are also taken from **user provided input**

We use standard IR techniques for matching field vectors of similar fields

Rank	Predicate
1	/film/film_regional_release_date/release_date
2	/fictional_universe/fictional_date_time/other_date_time
3	/base/lostintime/probable_date/date
4	/fictional_universe/fictional_character/date_of_birth
5-7	….
8	/biology/organism/date_of_birth
9	**/people/person/date_of_birth**
10	/people/place_lived/end_date

Results from a query to match **birthDate** field from SNAC to freebase

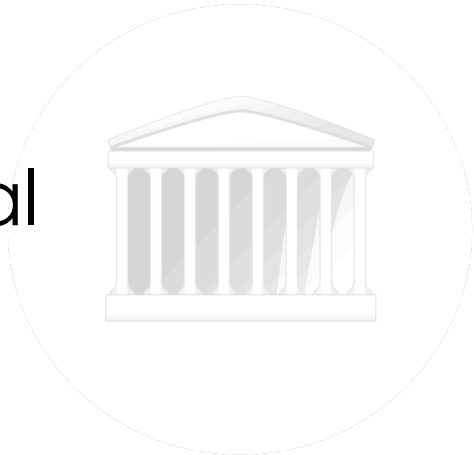
Context words: 'type ,birth, date, people, person'



Link Records

Develop techniques that given a SNAC entity can **find records to link/merge** in other collections

Goal



SNAC

Ritter, Tex 1905-1974
Ritter, Tex 1907-1974
Ritter, Tex
Ritter, Woodward Maurice 1905-1974
Ritter, Woodward Maurice

Freebase

Tex Ritter

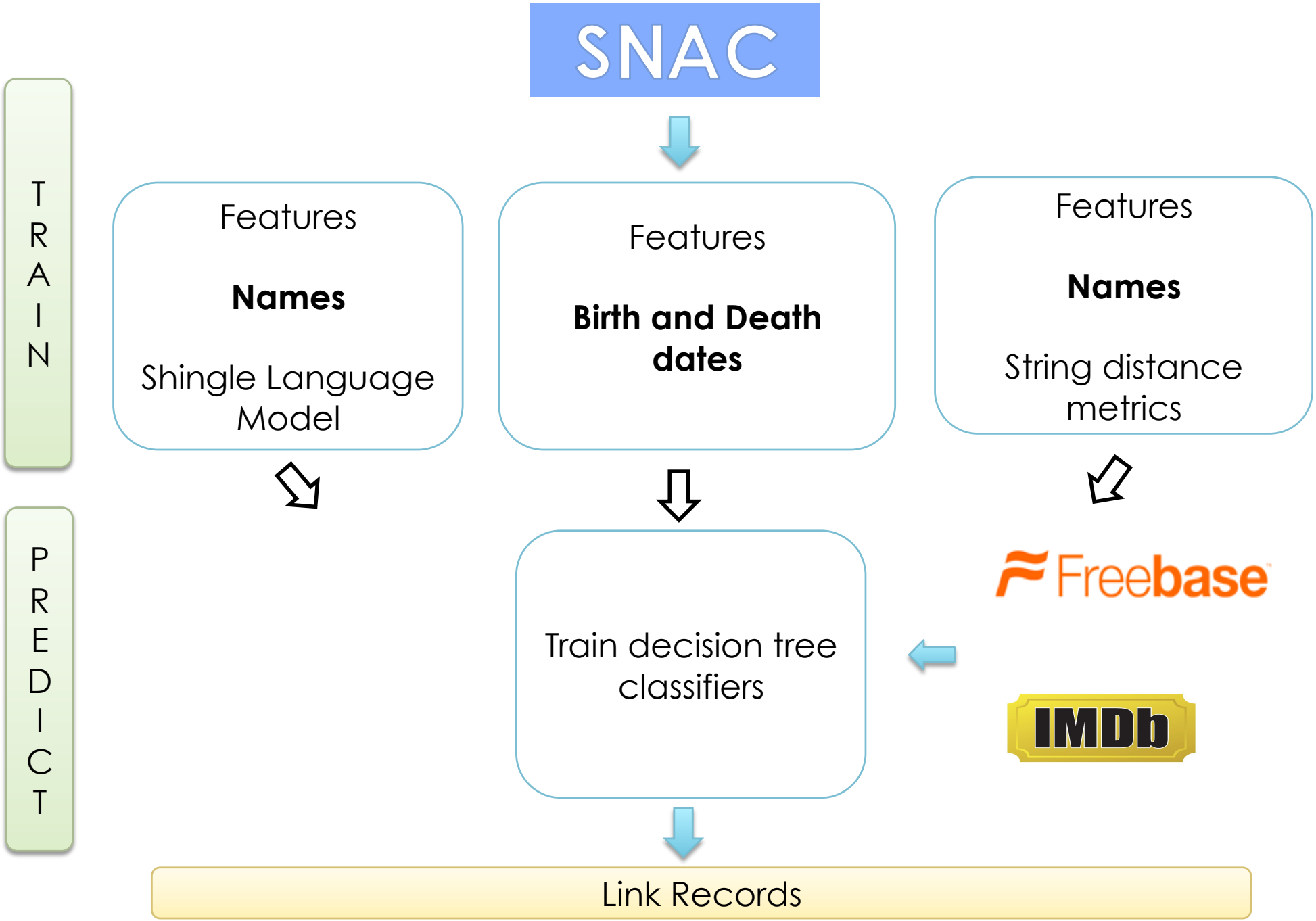
IMDb

Tex Ritter (1905–1974)

Our approach

Train binary classifiers over varying names and existence dates

Perturb existing information to generate additional samples within specific error levels


SNAC
TRAIN
PREDICT
Features
Names
Shingle Language Model
Features
Birth and Death dates
Features
Names
String distance metrics
Freebase
IMDb
Train decision tree classifiers
Link Records

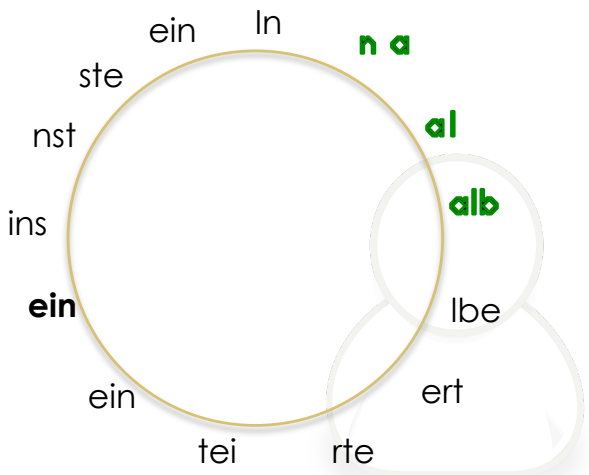
Name: Einstein Albert

Shingle sequence: ein, ins, nst, ste, tei, ein … , ert

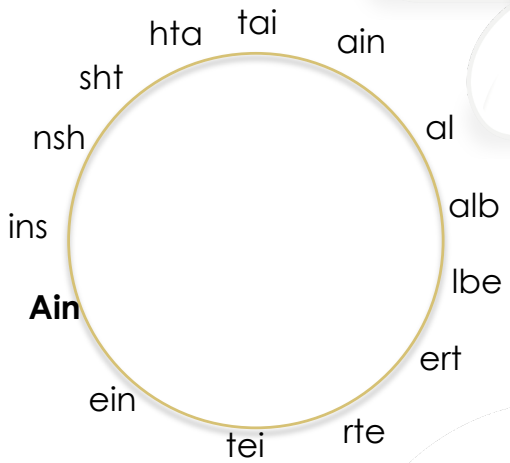
Probability that the sequence (ins, nst, ste) follows ein is very high for the name einstein

Shingle Language Model for names

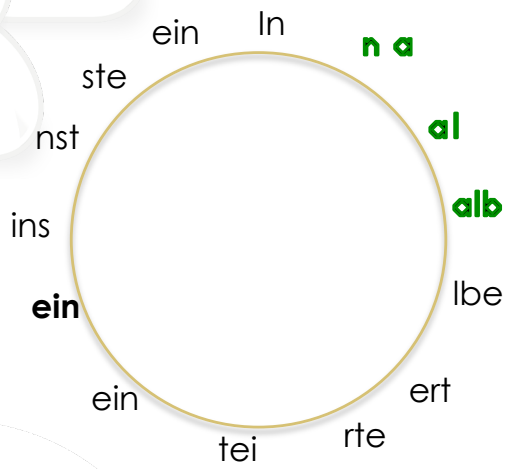
Name 1 : Einstein Albert



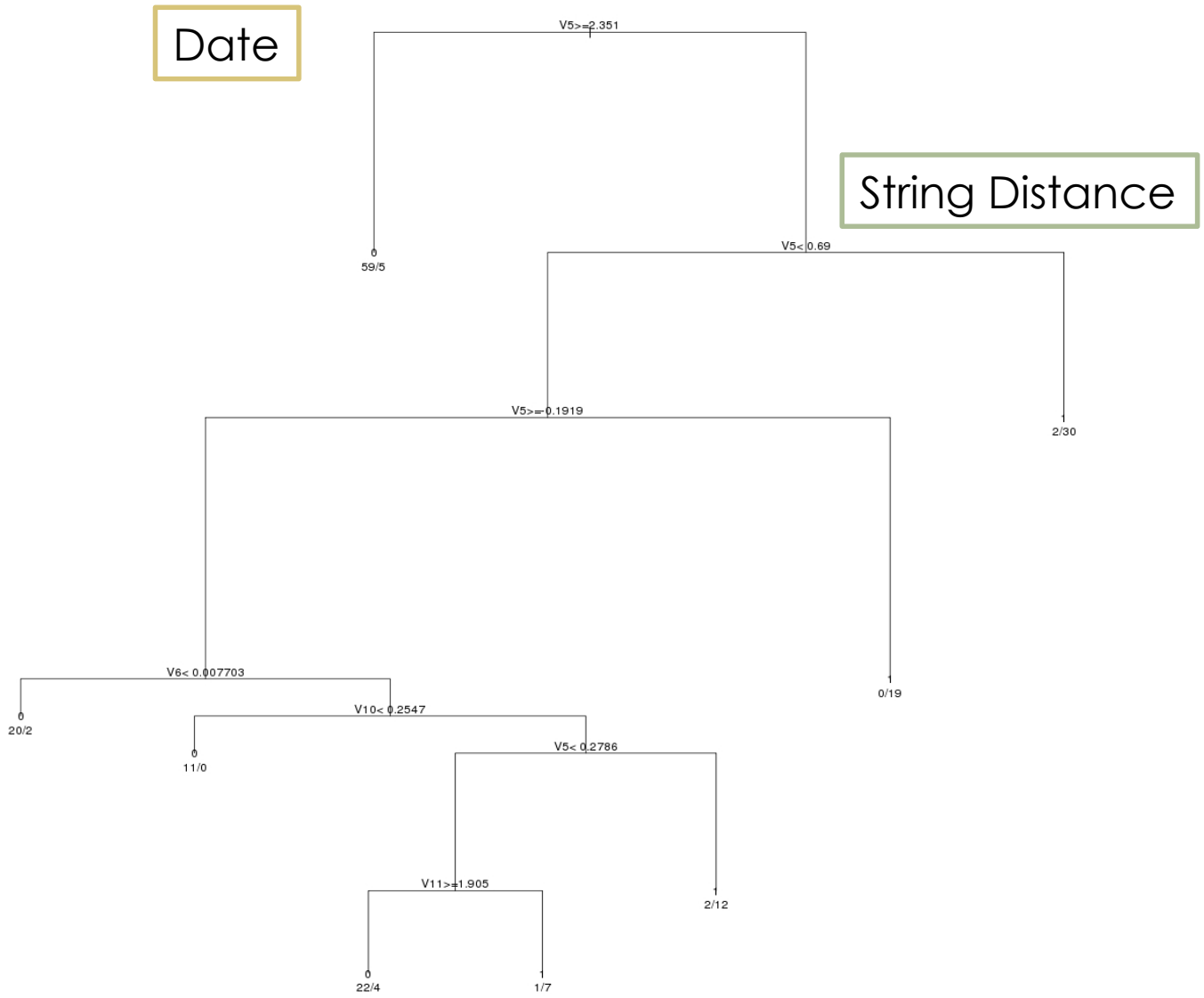
Name 2 : Ainshtain Albert



Name 3 : Albert Einstein



Shingle Language Model for names



Example Decision Tree For Von Neumann

Albert Einstein

TP:78	FP:11
FN:25	TN:145

TPR: 75.7%
FPR: 7%

George W Bush

TP:39	FP:9
FN:6	TN:60

TPR: 86.6%
FPR: 13%

Von Neumann

TP:182	FP:14
FN:27	TN:301

TPR: 87%
FPR: 4%

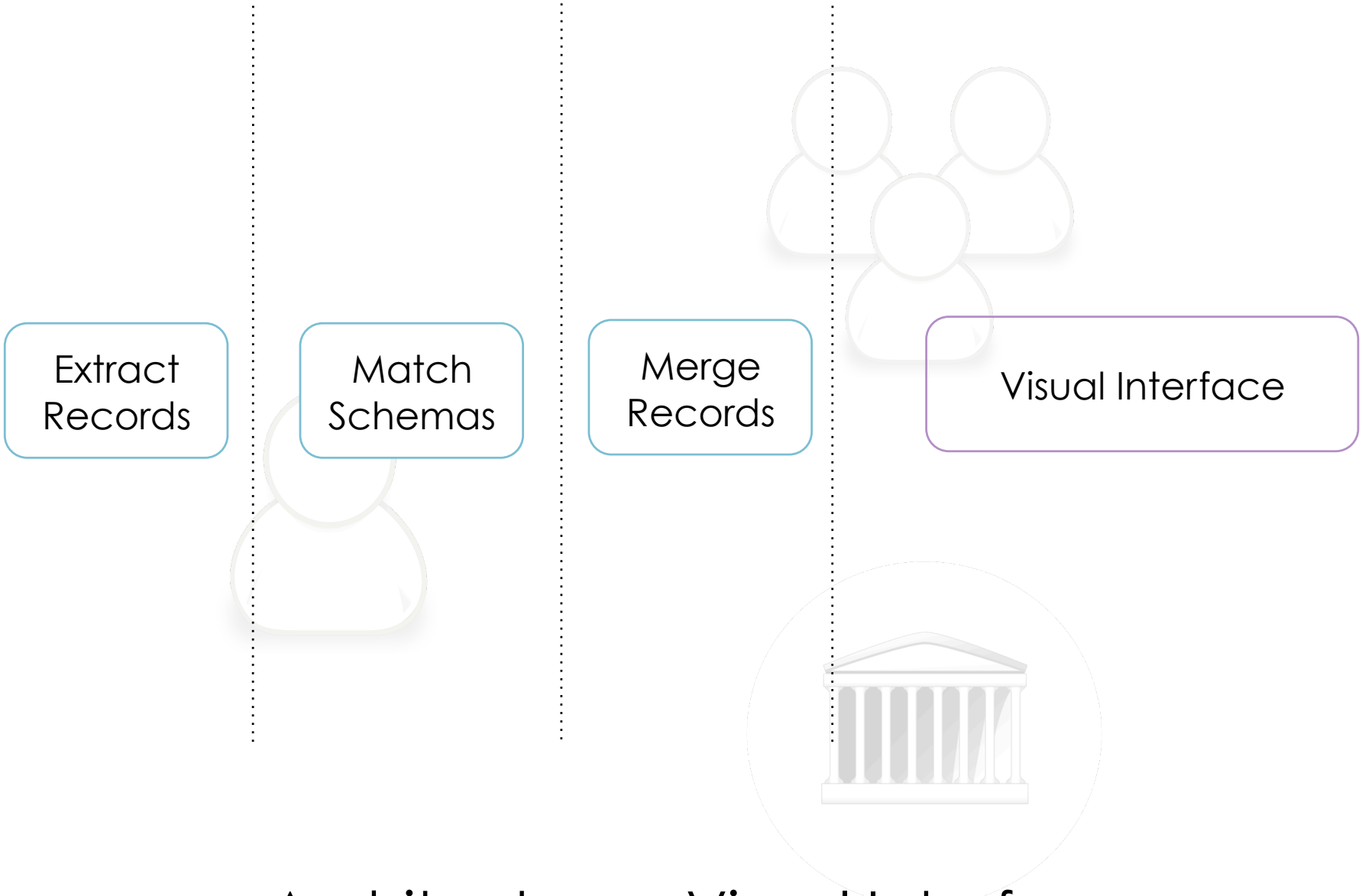
Corpus Average

TPR: 72.7%
FPR: 17%

15,300 records, thresh = 0.85

1100 records, thresh = 0.9

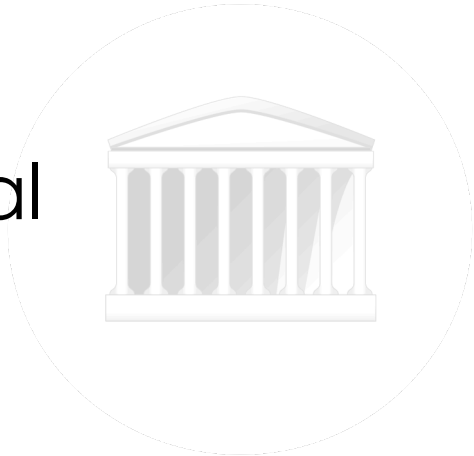
How many did we link ?



Architecture – Visual Interface

Develop a visual interface for **querying and discovering** social structures in our collection

Goal



Find Corporate, Personal, and Family Archival Context Records

SNAC prototype

PROTOTYPE

All | Person | Corporate Body | Family

enter a name or keywords | search

123,920 Names

All Names → Alphabetical Index →

0 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Featured Records

Bernstein, Leonard, 1918-
Buffalo Bill, 1846-1917.
Bush, Vannevar, 1890-1974.
Eames, Ray, 1912-1988.
Eisenhower, Dwight D. (Dwight David), 1890-1969.
Feynman, Richard Phillips, 1918-1988.
Fitzgerald, Ella, 1918-1996
Franklin, Benjamin, 1706-1790.
Ishigo, Estelle.
Luce, Clare Boothe, 1903-1987.
Millikan, Robert Andrews, 1868-1953.
Oppenheimer, J. Robert, 1904-1967.
Robbins family
Royal Chicano Air Force.
San Francisco AIDS Foundation.
Sierra Club.
Viramontes, Helena Maria, 1954-

Top Occupations

- Journalists (192)
- Authors (180)
- Lawyers (165)
- Diplomats (122)
- Authors, American (111)
- Educators (96)
- Farmers (90)
- Army officers (76)
- Domestics (76)
- Naval officers (76)
- Historians (75)
- Engineers (72)
- Economists (62)
- Public officials (62)
- Representatives, U.S. Congress (55)
- Statesmen (55)
- Jurists (53)
- Senators, U.S. Congress (52)
- Missionaries (48)
- Cabinet officers (46)
- Legislators (44)
- Clergy (42)

Top Subjects

- World War, 1939-1945 (722)
- Idaho (328)
- World War, 1914-1918 (323)
- Montana (314)
- Education (310)
- Washington (State) (275)
- Agriculture (269)
- International relief (232)
- Family (188)
- Colleges and Universities (184)
- Communism (183)
- Pioneers (183)
- Architecture (157)
- Government and Politics (149)
- Refugees (147)
- Japanese Americans (146)
- Norwegian-Americans (142)
- Authors, American (134)
- Photographs (132)
- Music (130)
- National socialism (124)
- Oregon (118)
- Railroads (116)

Exploratory, faceted search interface

Pros

- Great for finding a specific person or persons under specific facets
- Gives a good overview of the distribution of facets in the entire collection

Cons

- Impossible to discover social structures that span the collection
- Requires multiple explore and filter steps, followed by an aggregation step

Alternate interfaces for discovering structures in social networks

Sophisticated query languages:

hard to learn for the user

Interactive graph exploration tools:

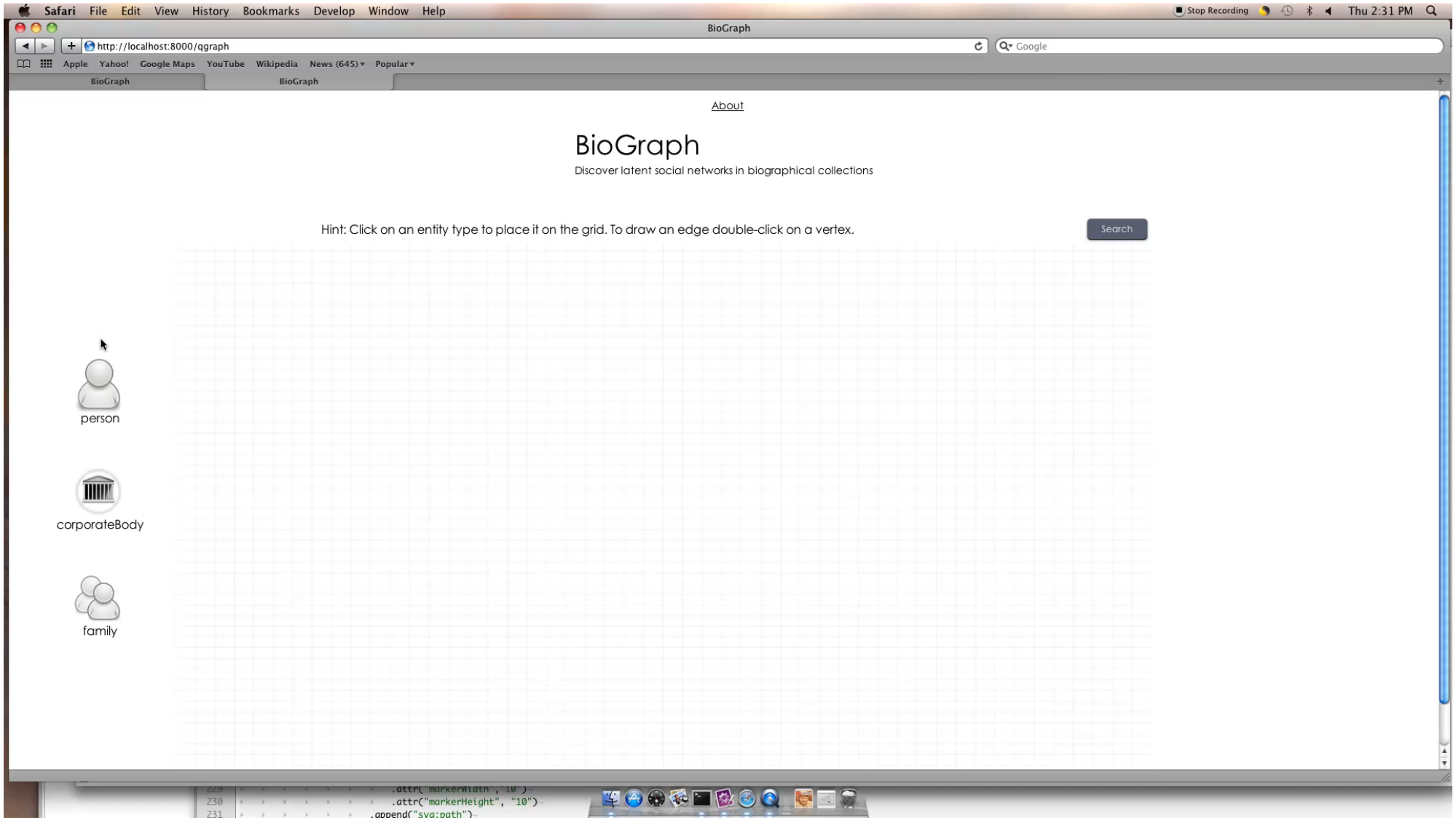
great for exploration, requires multiple steps to find the desired social structure pattern

Our Approach

Users express queries for discovering social structures as a **query graph**

A query by example interface

Techniques for visualizing the result graph


Safari File Edit View History Bookmarks Develop Window Help
Stop Recording
Thu 2:31 PM
BioGraph
http://localhost:8000/qgraph
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (645) Popular
BioGraph
BioGraph
About
BioGraph
Discover latent social networks in biographical collections
Hint: Click on an entity type to place it on the grid. To draw an edge double-click on a vertex.
Search
person
corporateBody
family
.attr("markerWidth", "10")
.attr("markerHeight", "10")
.append("svg:path")

Questions ?

http://bit.ly/biograph11