Postgres & Social Sciences:
How we make it work

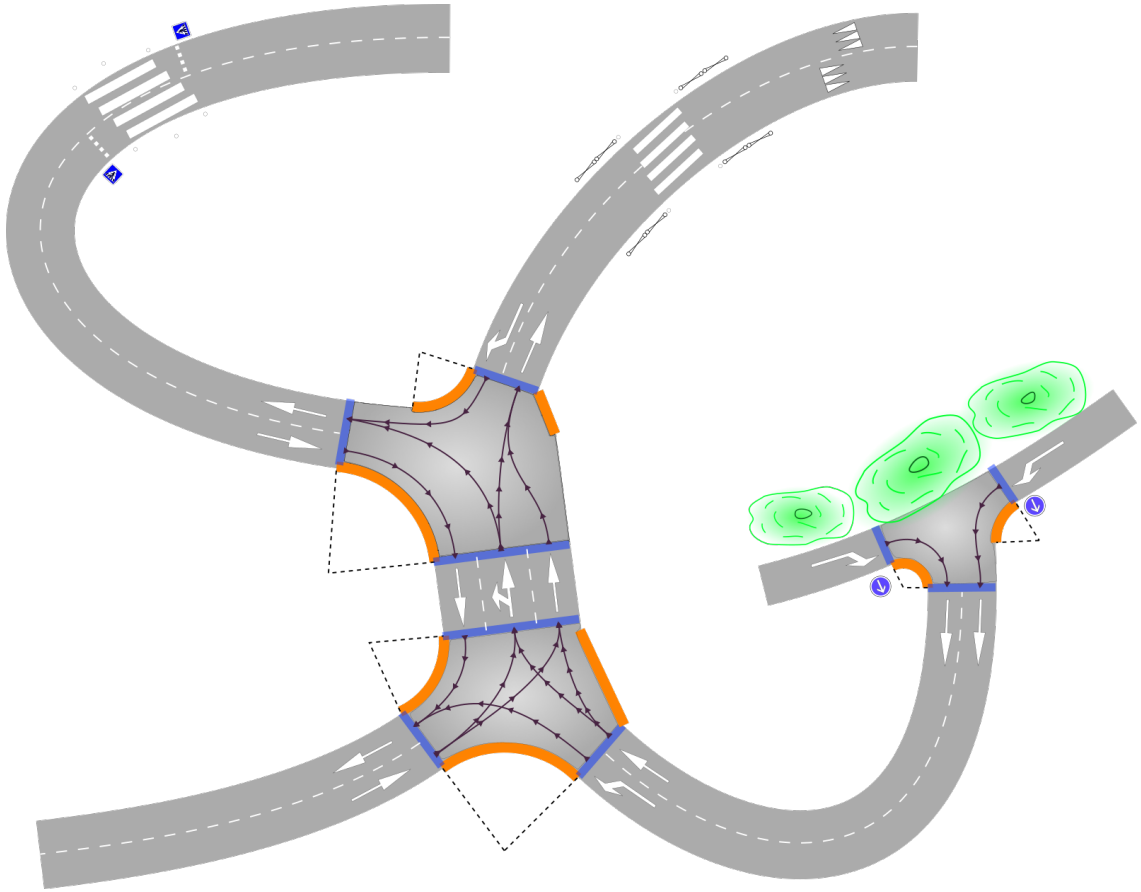
Dr. **Rémi Cura**, *Postdoctoral Associate* http://remi.cura.info
Prof. **In Song Kim**, *Associate Professor* http://web.mit.edu/insong/www
MIT Political Science Department

Today

- Social Sciences: Intro, challenges
- Postgres strategies to help?

Previously:

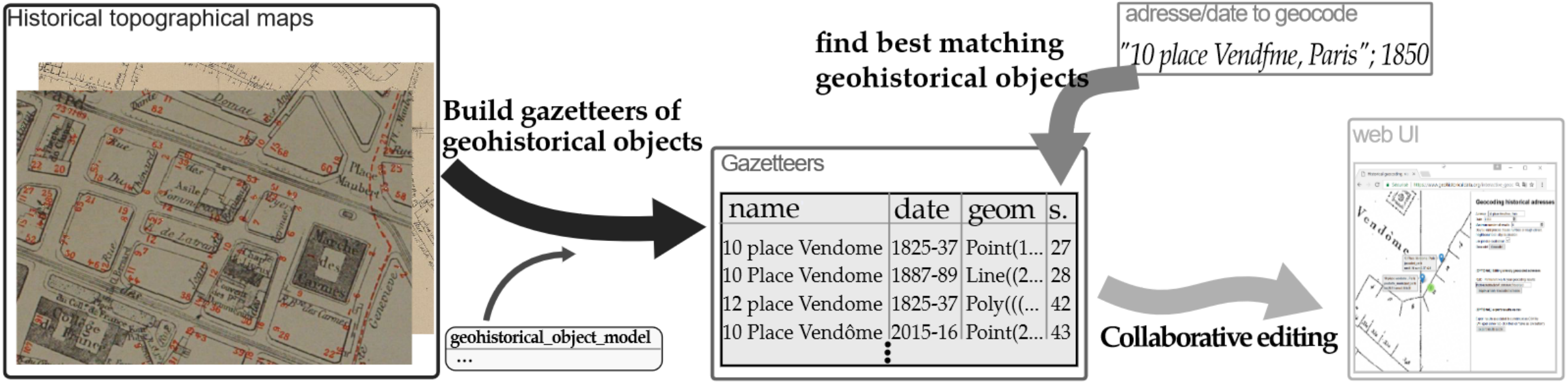
- Street model generation at city scale with **Postgres**
- **Urbanism**: help design city for person with disabilities



Postgres + PostGIS + PgPointCloud

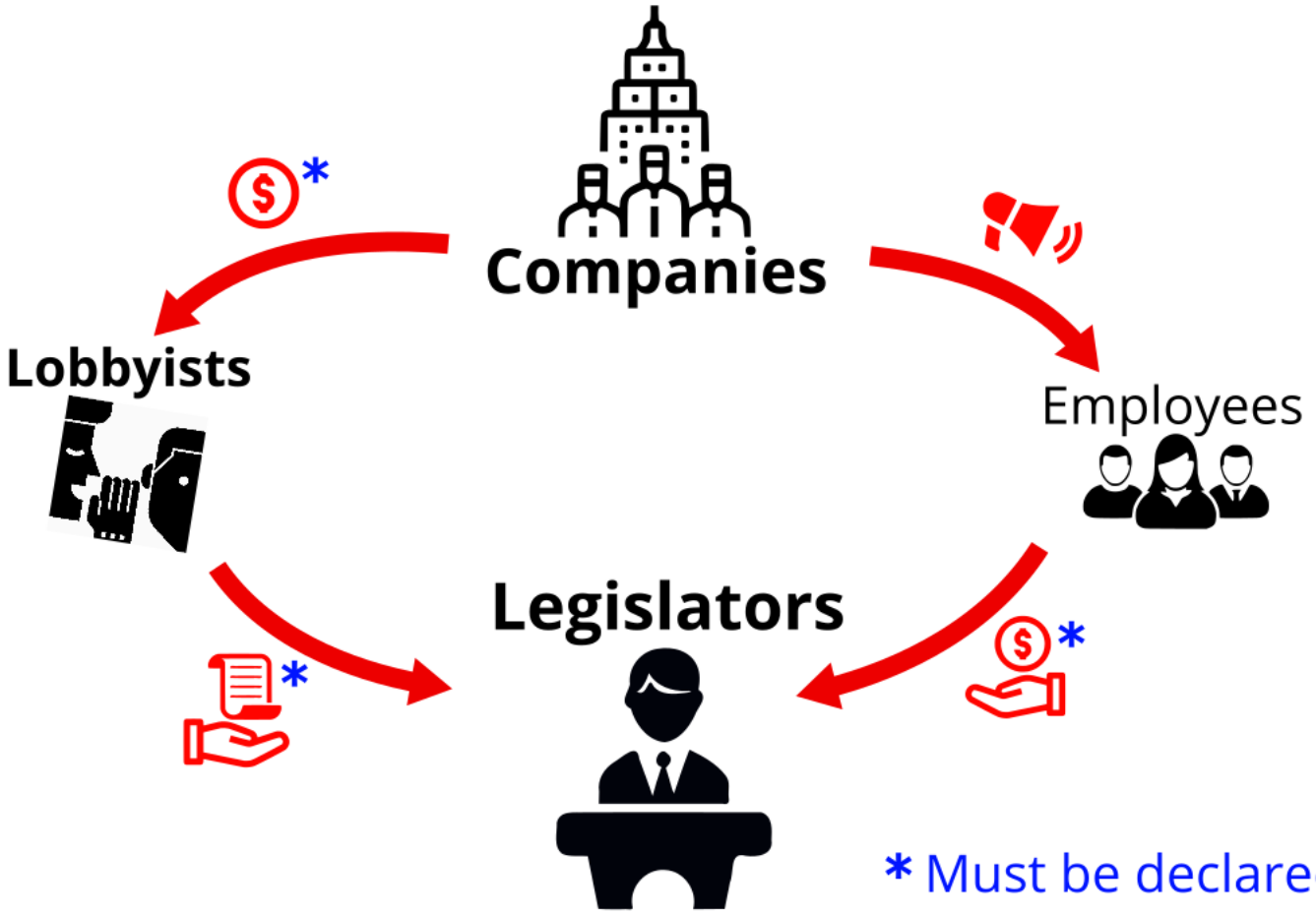
Previously

- Paris historical map: 18th and 19th century, in **Postgres**
- Where is (address; date)?
- **Social Dynamics in Urban Context**

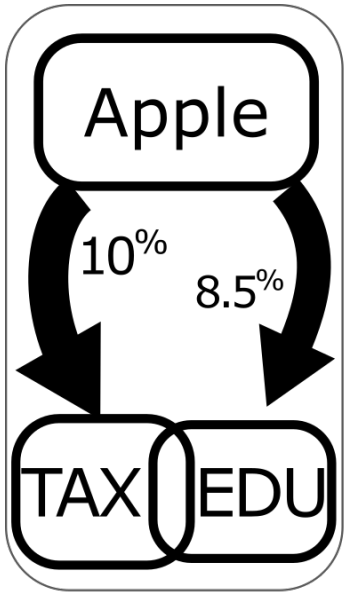


Lobbyview.org

- Influence of money in US legislative life, with **Postgres**.
- Political science

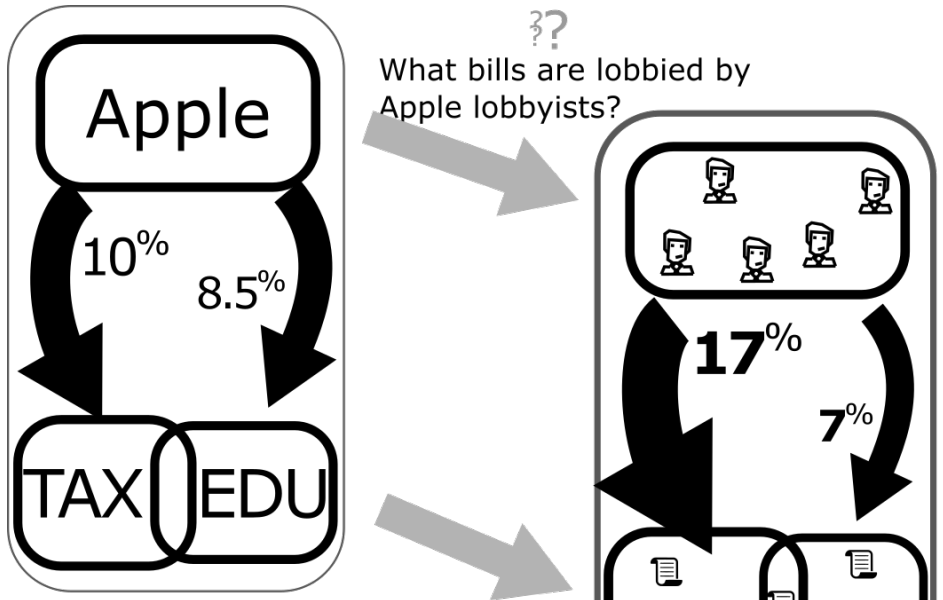


Example of research questions



Apple **spends about as much** on EDU and TAX

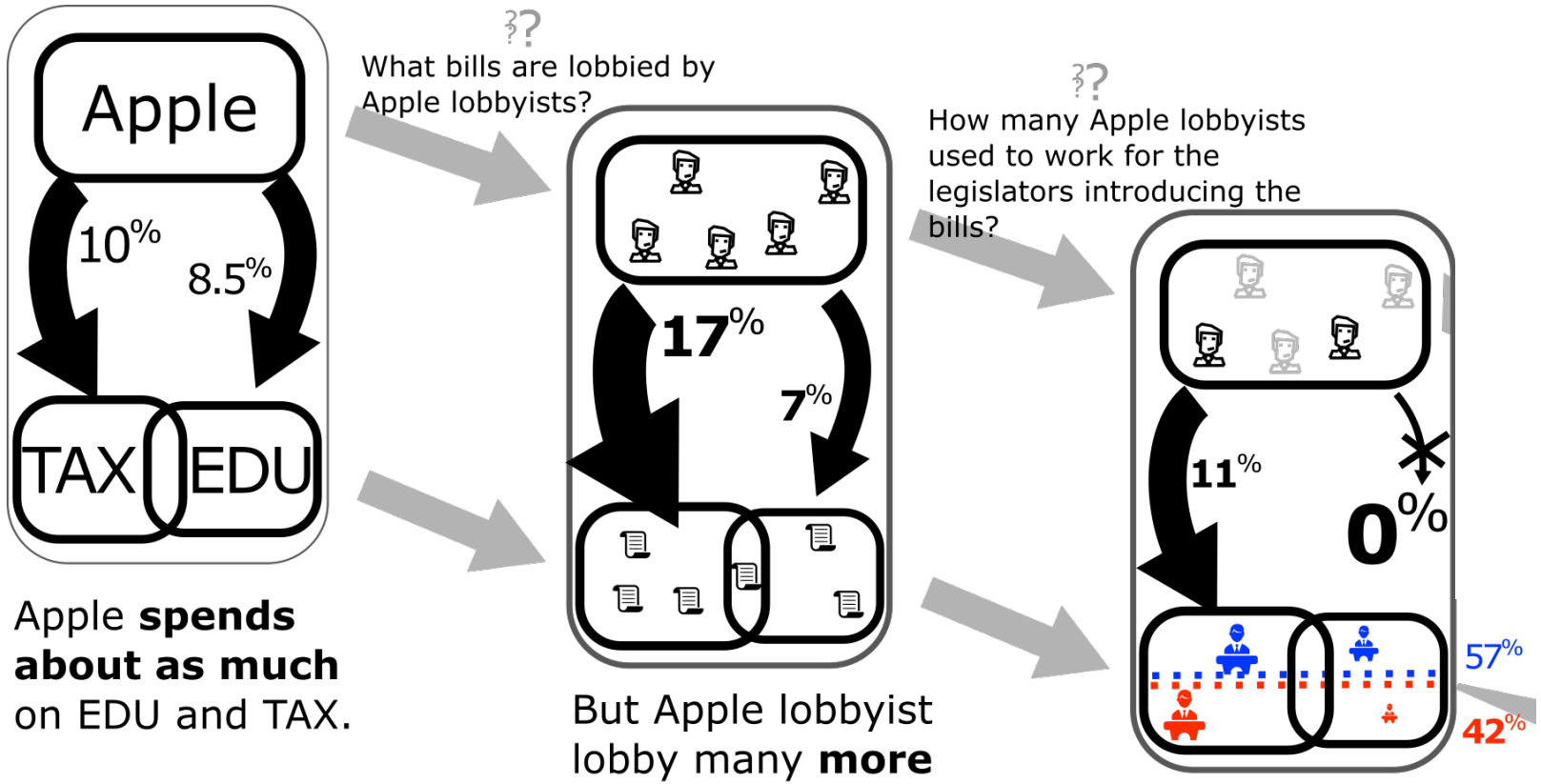
Example of research questions



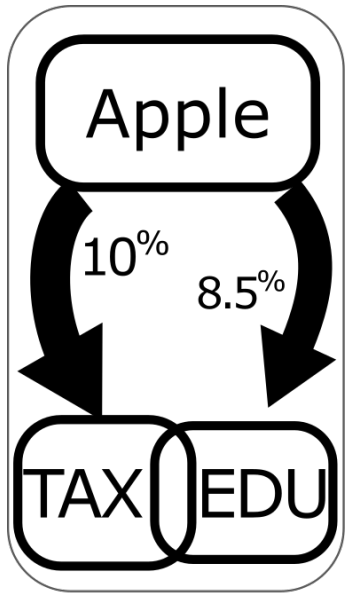
Apple **spends about as much** on EDU and TAX.

But Apple lobbyist lobby many **more tax-bills.**

Example of research questions

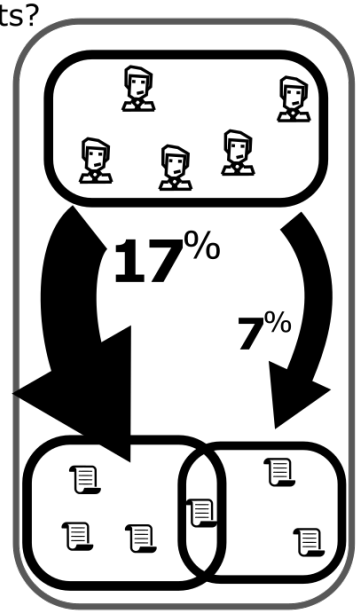


Example of research questions



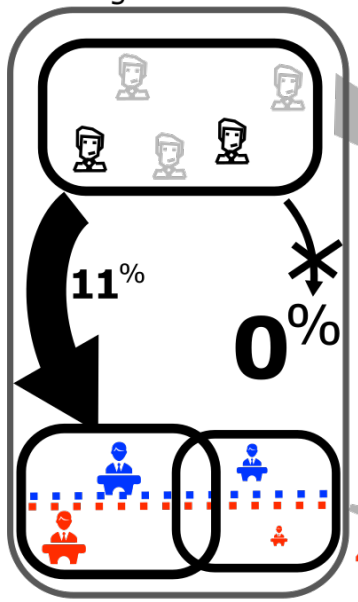
Apple **spends about as much** on EDU and TAX.

What bills are lobbied by Apple lobbyists?



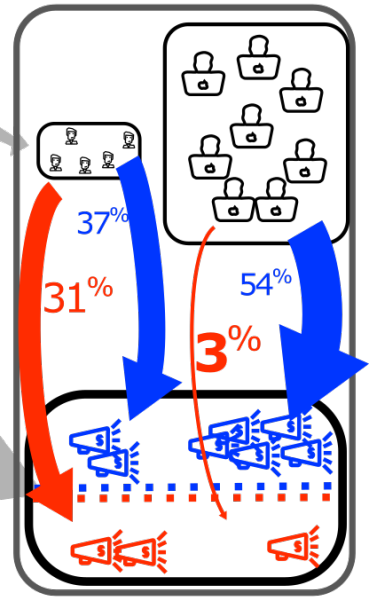
But Apple lobbyist lobby many **more tax-bills.**

How many Apple lobbyists used to work for the legislators introducing the bills?



Revolving door lobbyists are **focused on tax**. Lobbying is similar across party lines.

How do Apple lobbyists and employees give money to campaigns across party lines?



Apple employees donate much **less to republicans** than lobbyists.

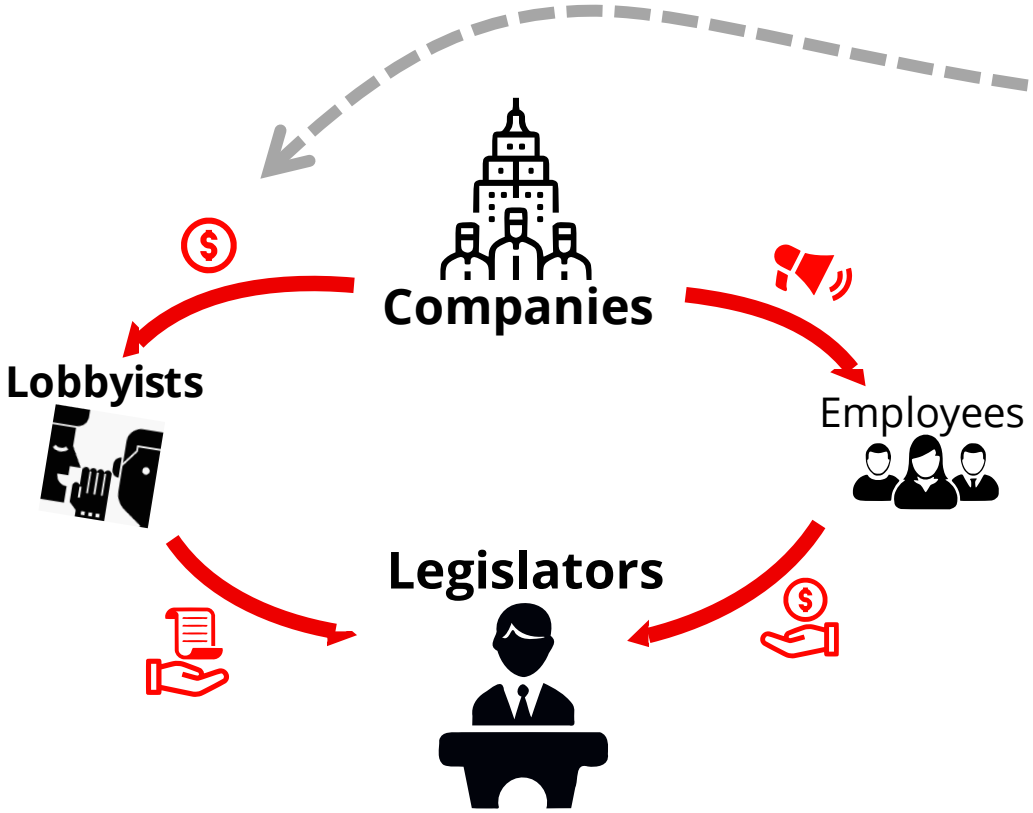
Goal

Create data and methods
to investigate Social Sciences questions.

Typical approach in Social Sciences

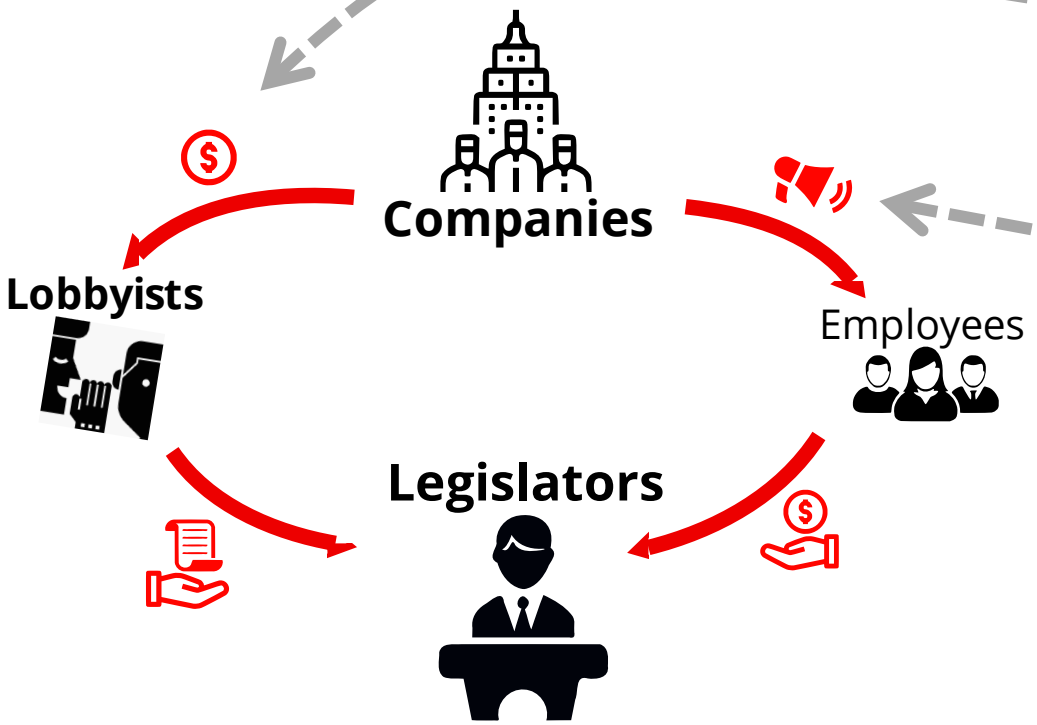
- one big .csv file
- → this is not going to work…

Challenges:



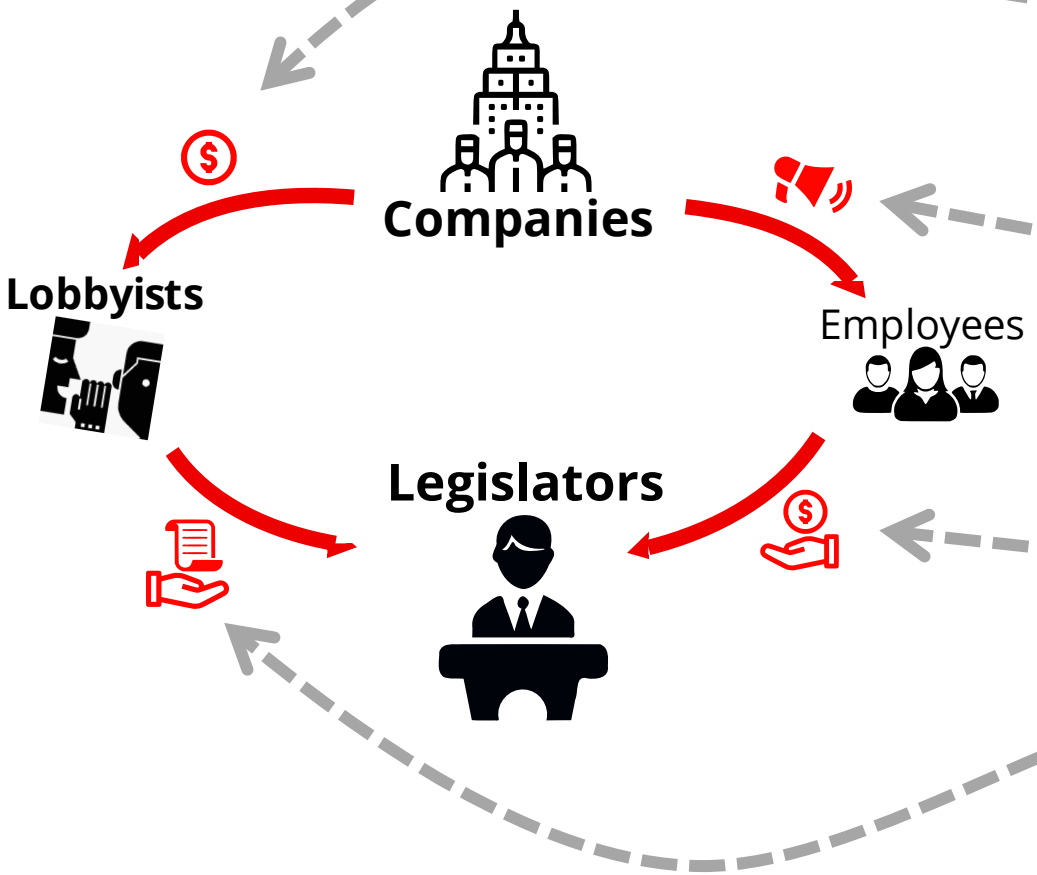
- Data is fuzzy:
  About** 10k$ during **Q1 of 2014

Challenges:



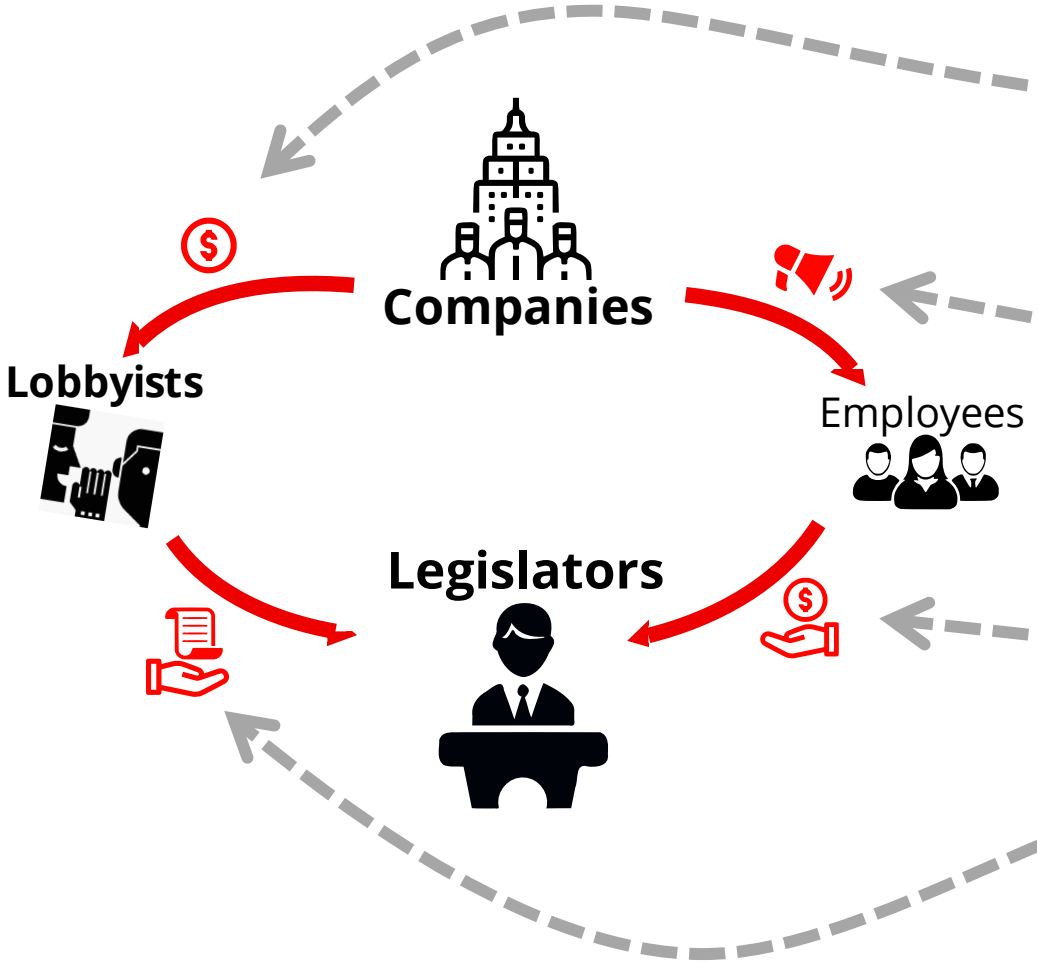
- Data is fuzzy:
  About** 10k$ during **Q1 of 2014
- Data definition is ambiguous:
  *Companies **"influence"** employees?*

Challenges:



- Data is fuzzy:
  About** 10k$ during **Q1 of 2014
- Data definition is ambiguous:
  *Companies **"influence"** employees?*
- Multiple conflicting data sources
  *Money for **"J. McCain"**.*
  *Bills of **"McCain, John S."***
  *Is it the **same guy?***

Challenges:



- Data is fuzzy:
  About** 10k$ during **Q1 of 2014
- Data definition is ambiguous:
  *Companies **"influence"** employees?*
- Multiple conflicting data sources
  *Money for **"J. McCain"**.*
  *Bills of **"McCain, John S."***
  *Is it the **same guy?***
- Specific Social Sciences needs:
  Explore, be Explainable

Challenges

- Data is **fuzzy**
  “About 10k$”
- Data **definition** is ambiguous
  “influence” on employees?
- **Conflicting** data sources
  “J. McCain” **VS** *“McCain, John S.”*
- Specific to **Social Sciences**:
  Explore, be Explainable

Postgres solutions?

Challenges

- Data is **fuzzy**
  "About 10k$"
- Data **definition** is ambiguous
  "influence" on employees?
- **Conflicting** data sources
  "J. McCain" **VS** *"McCain, John S."*
- Specific to **Social Science**:
  Explore, be Explainable

Postgres features

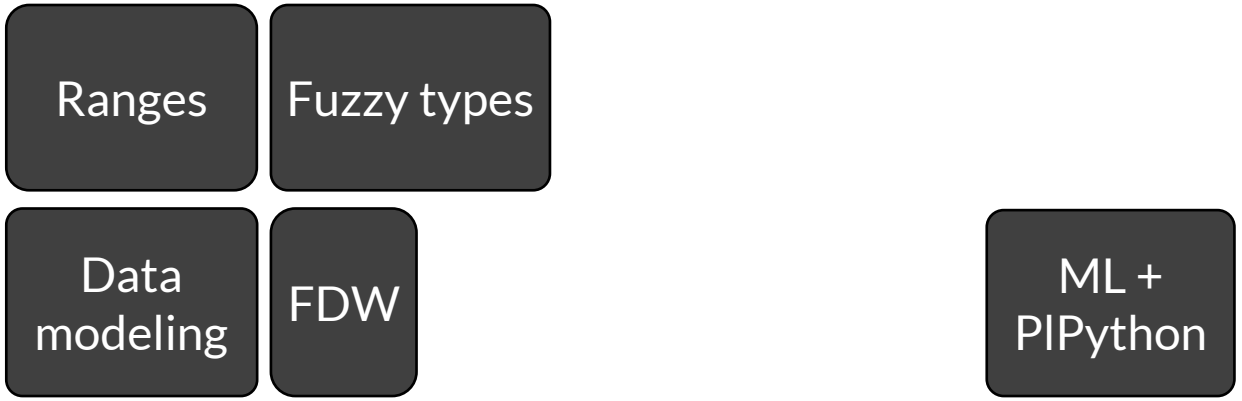
Ranges

Fuzzy types

Challenges

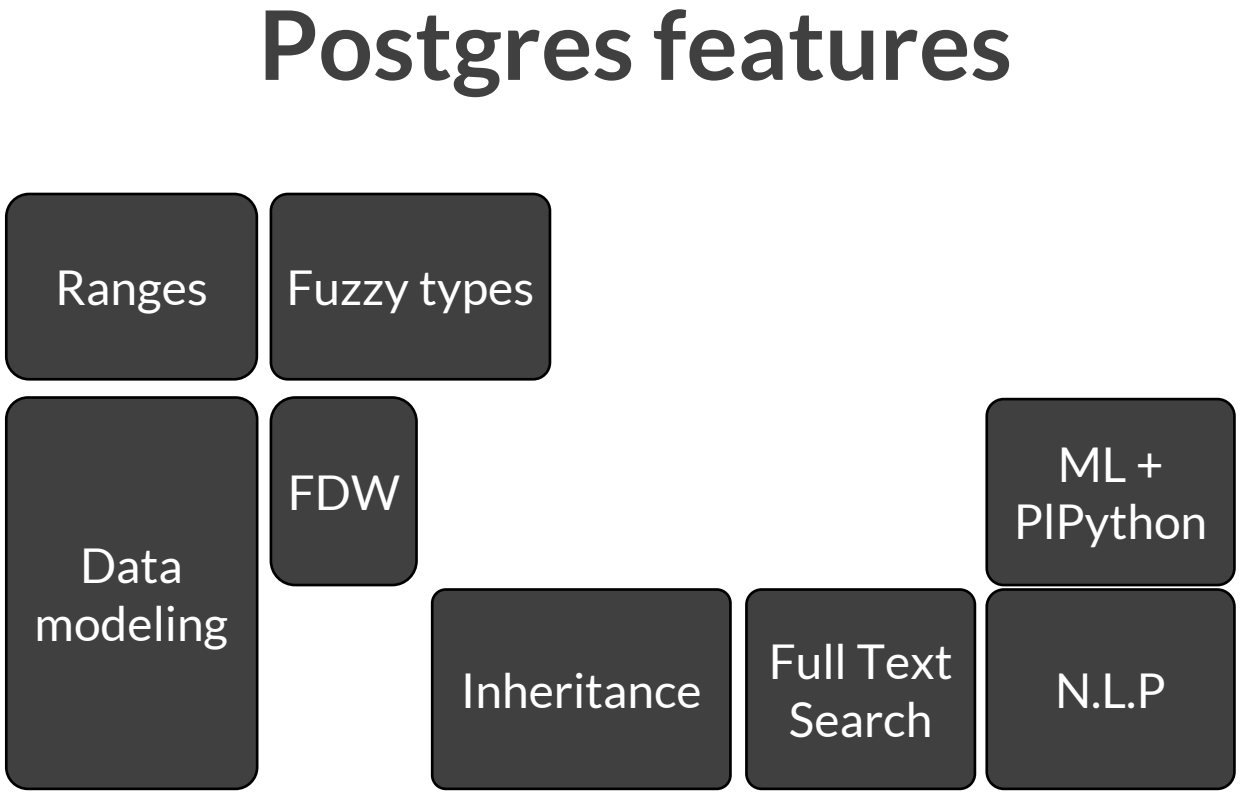
- Data is **fuzzy**
  "About 10k$"
- Data **definition** is ambiguous
  "influence" on employees?
- **Conflicting** data sources
  "J. McCain" **VS** *"McCain, John S."*
- Specific to **Social Science**:
  Explore, be Explainable

Postgres features



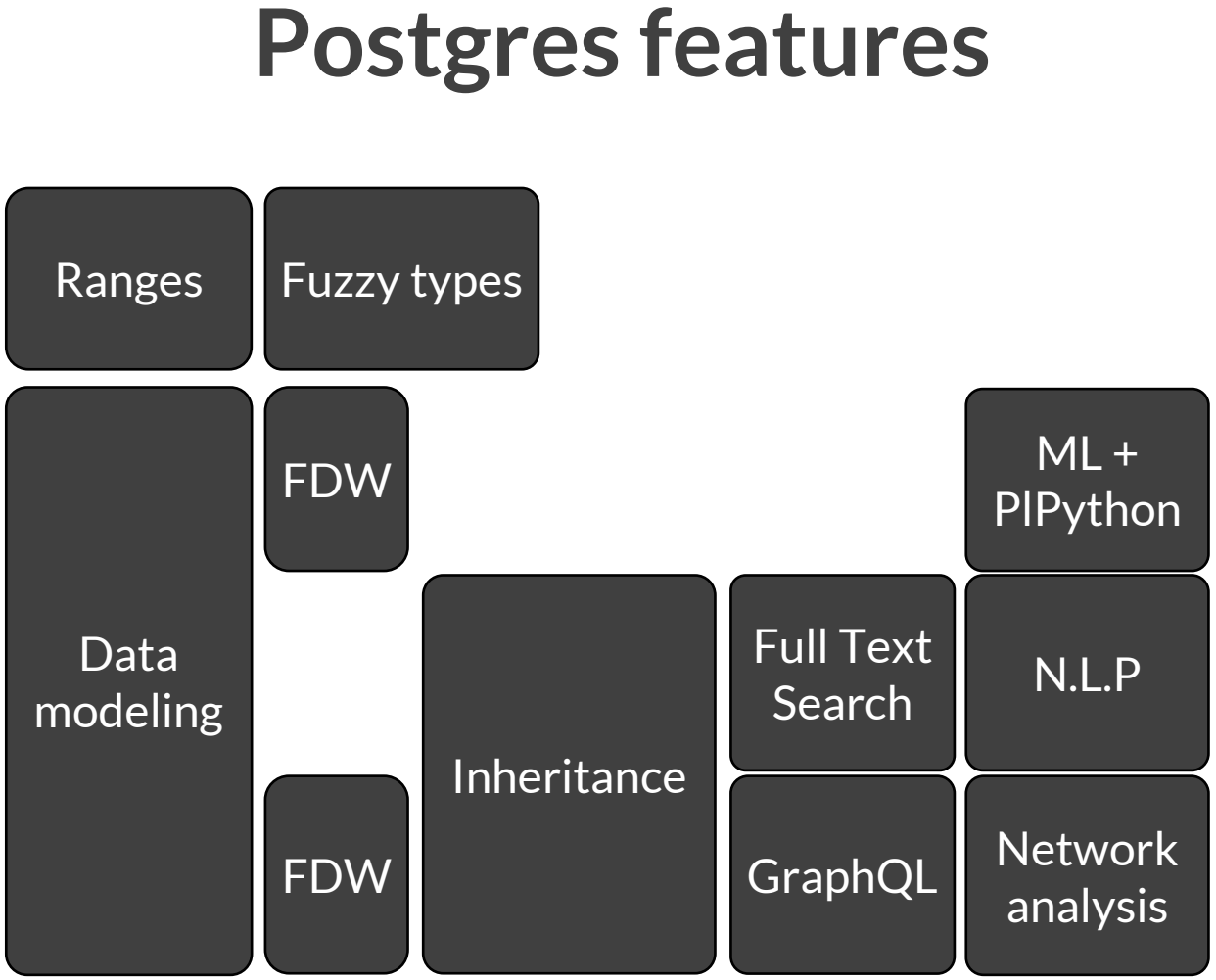
Challenges

- Data is **fuzzy**
  "About 10k$"
- Data **definition** is ambiguous
  "influence" on employees?
- **Conflicting** data sources
  "J. McCain" **VS** *"McCain, John S."*
- Specific to **Social Science**:
  Explore, be Explainable



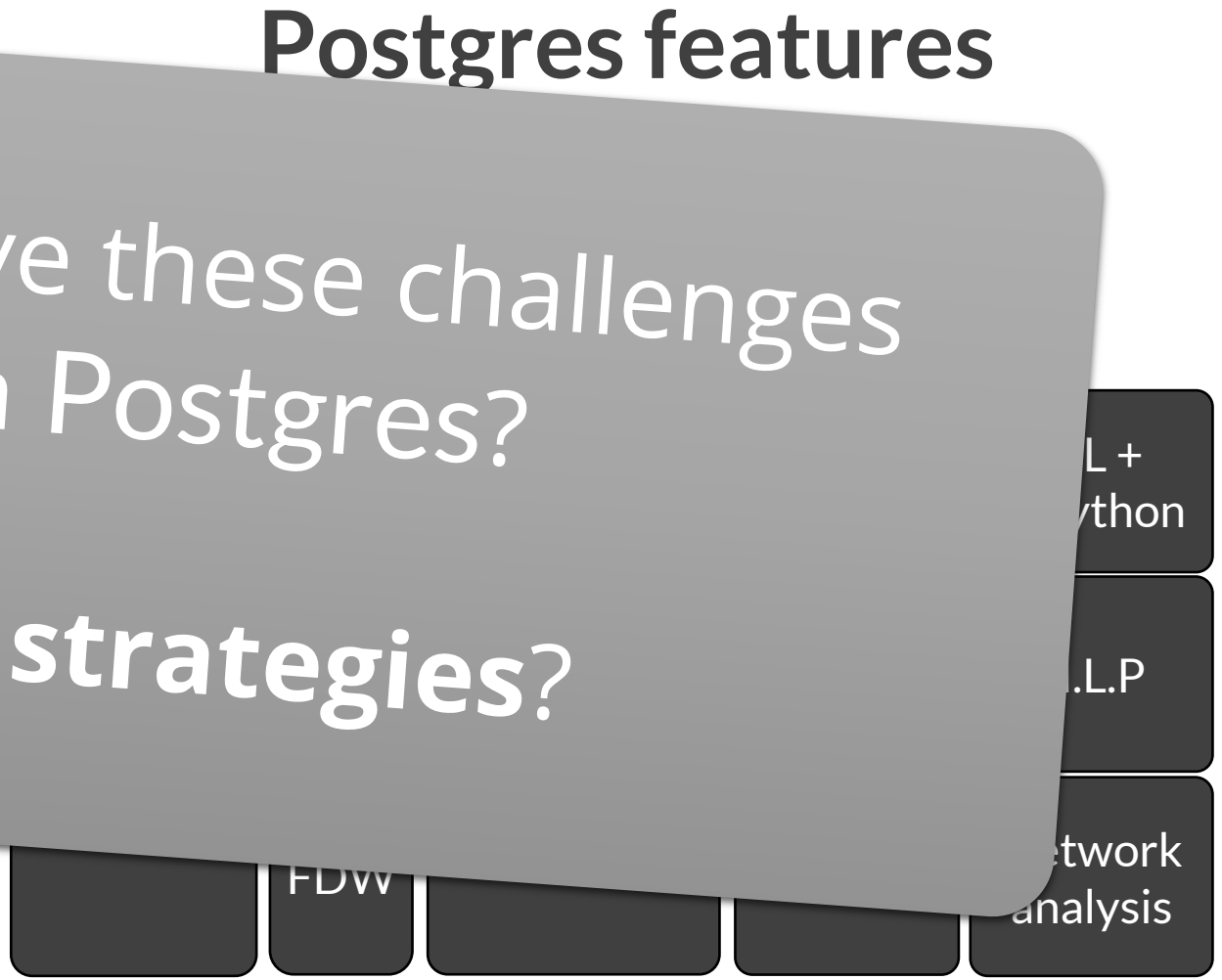
Challenges

- Data is **fuzzy**
  "About 10k$"
- Data **definition** is ambiguous
  "influence" on employees?
- **Conflicting** data sources
  "J. McCain" **VS** *"McCain, John S."*
- Specific to **Social Science**:
  Explore, be Explainable



Challenges

- Data is
  "About
- Data **d**
  "influe
- **Confl**
  "J. Mc
- Specific to **Social Science**:
  Explore, be Explainable



How to solve these challenges with Postgres?

What **strategies**?

Practical needs

- Users =
  - Experts
  - Social Scientists
  - Students

Design choices
toward simplicity

- ~~Postgres experts~~
- ~~SQL experts~~
- ~~Developers~~
- …

Connecting to Postgres is already a challenge for users

Strategies:

Strategy 1:
Dealing with Fuzzy data

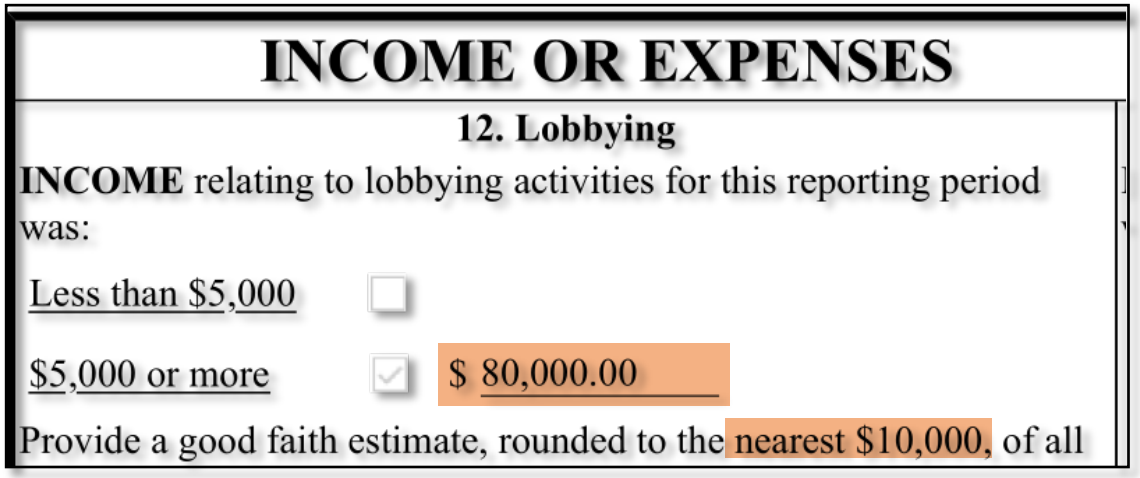
Data is **fuzzy**
"About 10k$"

Data **definition** is ambiguous
"influence" on employees?

Conflicting data sources
"J. McCain" **VS** *"McCain, John S."*

Specific to **Social Sciences**:
Explore, be Explainable

Fuzzy data?



INCOME OR EXPENSES

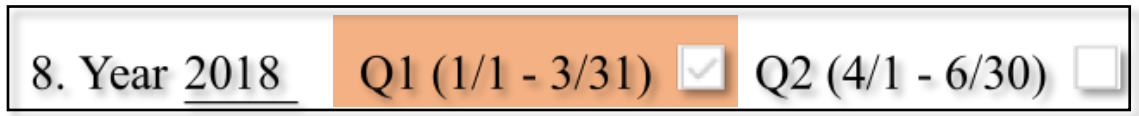
12. Lobbying

INCOME relating to lobbying activities for this reporting period was:

Less than $5,000 ☐

$5,000 or more ☑ $ 80,000.00

Provide a good faith estimate, rounded to the nearest $10,000, of all



8. Year 2018 Q1 (1/1 - 3/31) ☑ Q2 (4/1 - 6/30) ☐

→Postgres **range type:**

→Basic operators

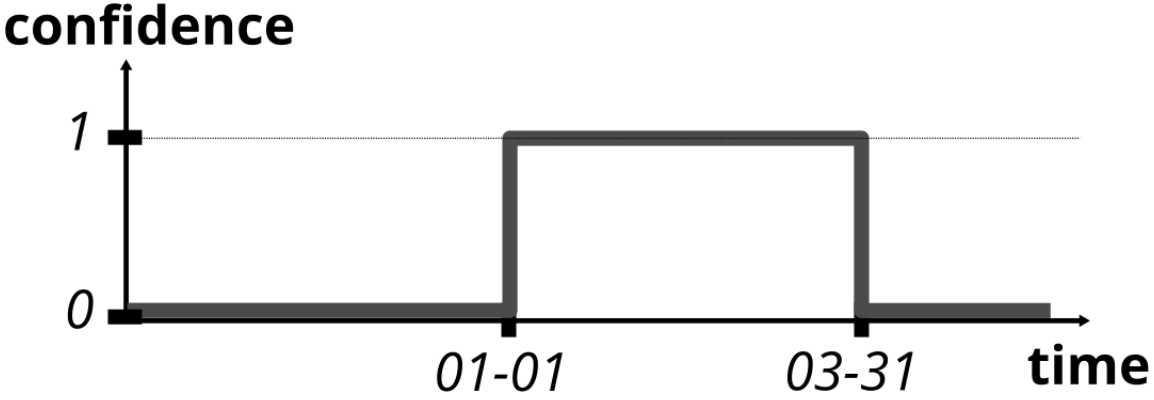
→Gist, SP-Gist indexes

```
SELECT daterange('2018-01-01', '2018-03-31');
-- [2018-01-01,2018-03-31)
```

Fuzzy data?

- Could interpret as confidence

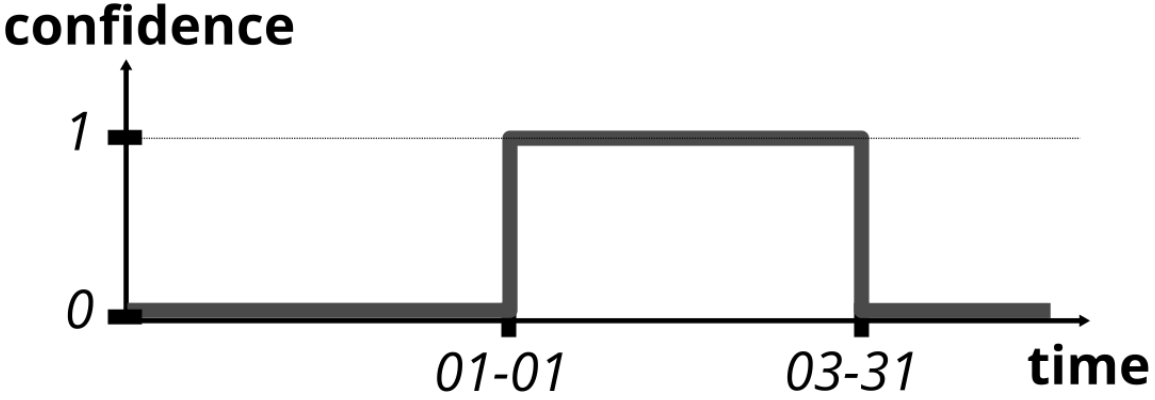
```
daterange('2018-01-01', '2018-03-31');
```



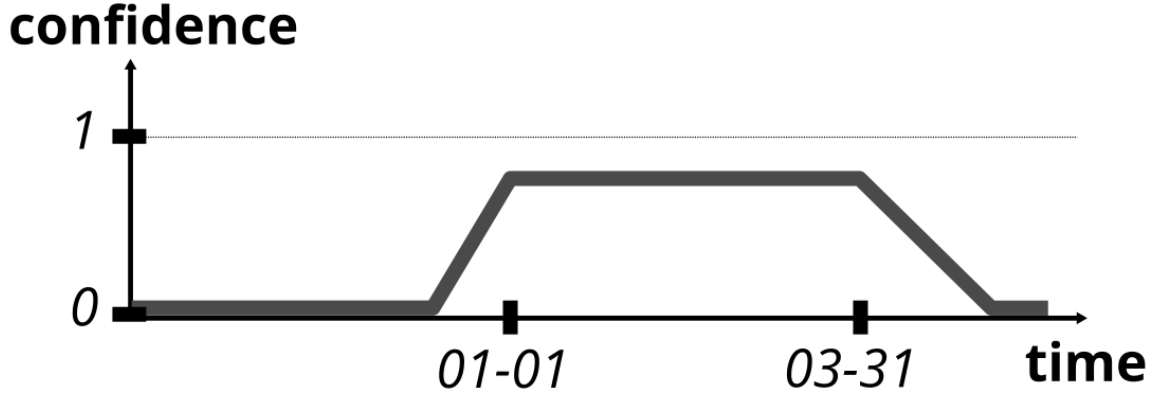
Fuzzy data?

- Could interpret as confidence

```
daterange('2018-01-01', '2018-03-31');
```

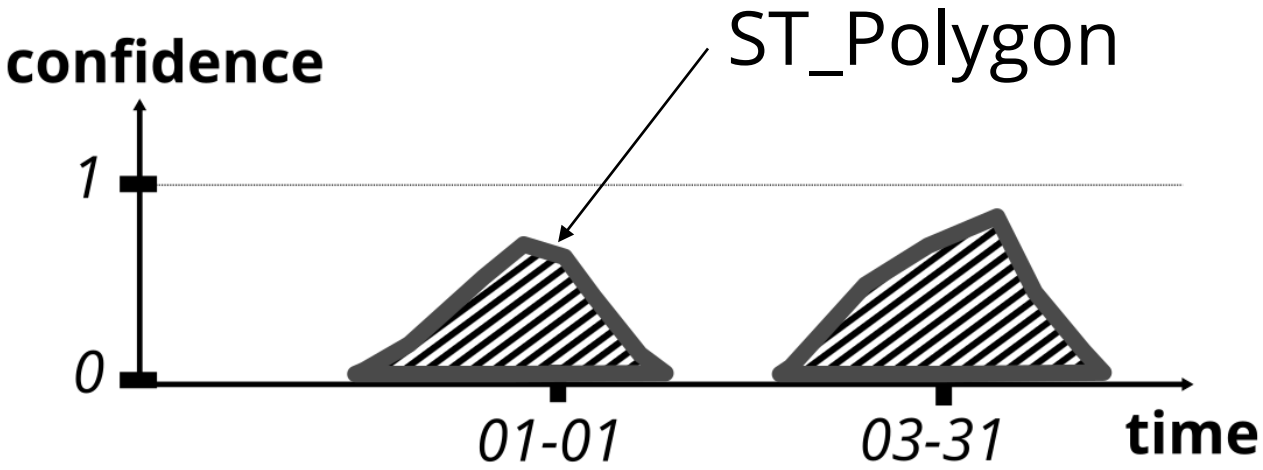


- We may need more flexibility:
  ‘ it can’t be before XX, but I know it was between YY and ZZ for sure’
  - →Extension pgSFTI
  - →Linear confidence
  - →Must be **contiguous**



Fuzzy data?

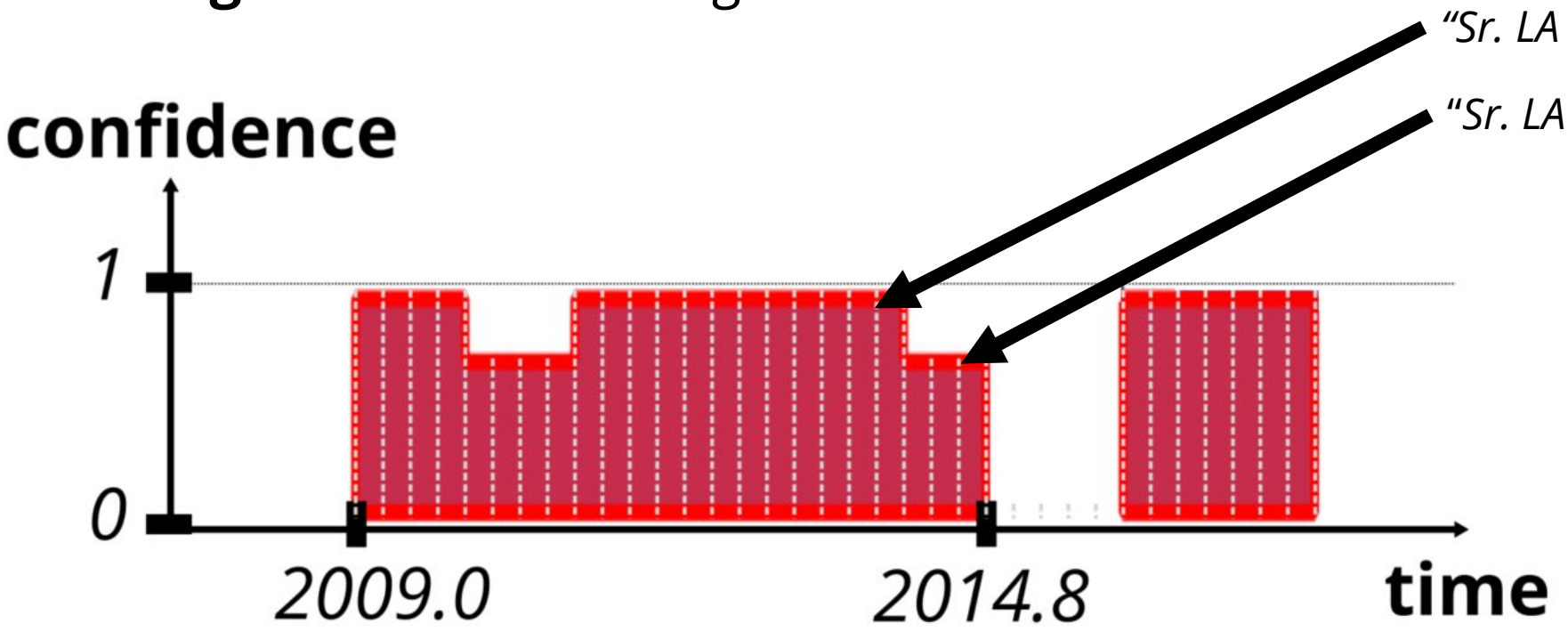
- Disjoint confidence:
  →PostGIS !



→Union, distance, intersection
→Gist index
→Can be overkill

Fuzzy data?

- Real life example (viewed in QGIS)
  Quarters when **lobbyist** "*Tim Bromelkamp*"
  declared he **used to work**
  for **legislator** "*David Minge*"



Strategies:

Strategy 2:
Data definition is ambiguous

Data is **fuzzy**
“About 10k$”

Data **definition** is ambiguous
“influence” on employees?

Specific to **Social Sciences**:
Explore, be Explainable

Conflicting data sources
“J. McCain” **VS** *“McCain, John S.”*

Data definition is ambiguous?

2 real lobbying activity reports:

- Company 1, 2009

  *'(*we lobbied on bill*)…**HR 1760**:*
  Black Carbon Emissions Reduction Act of 2009;
  ***Provisions** regarding carbon management….'*

- Company 2, 2010

  *''(*we lobbied on bill*) …**HR 1760**:*
  Black Carbon Emissions Reduction Act of 2009
  (entire bill)'

- Which lobbied **in support** or **against**? → ambiguous!

Data definition is ambiguous?

→ **Context is also important**

(company website image)

- Company 1, 2009 : **Exxon Inc.**

- Company 2, 2010: **Weyerhaeuser Comp.**

Still don’t know for sure !

Data definition is ambiguous?

Social Science

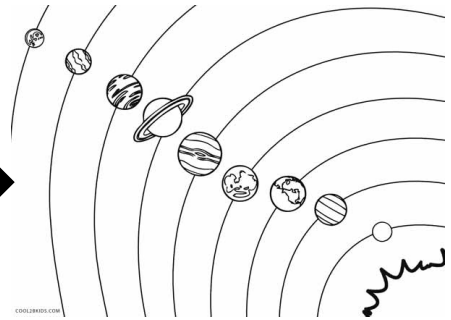
Make a choice : **lobbied in support**

→ Exxon defends Carbon emission reduction

→ ~~Always work~~ ~~Universal~~ Local True for specific (hypothesis, context)

Physic

→ $F_{B/A} = G\frac{M_A M_B}{d^2}$

→ 

→ Always work Universal

Data definition is ambiguous?

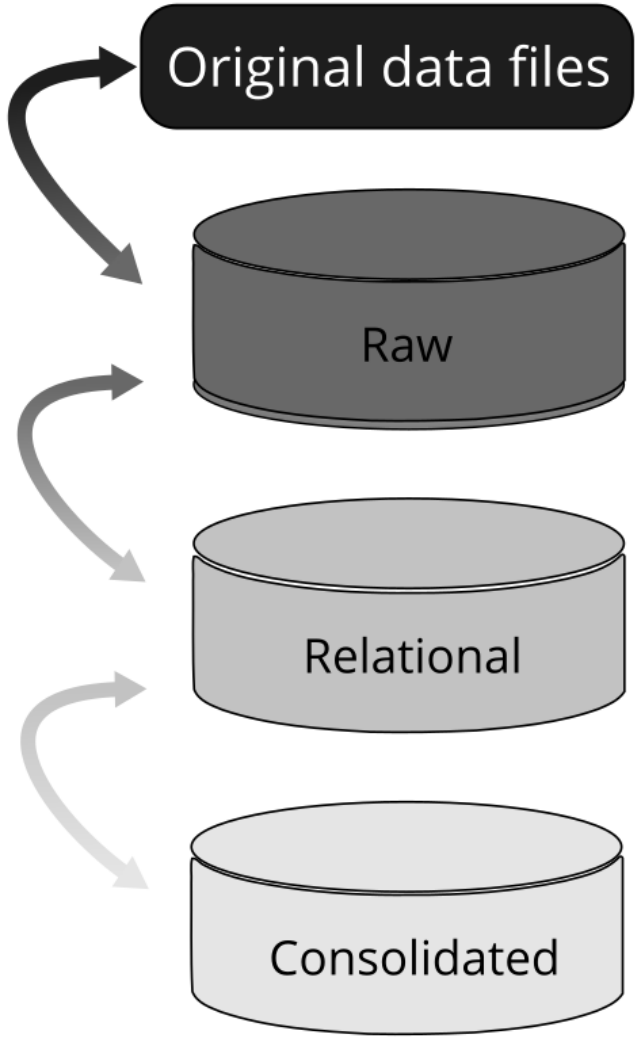
- Allow to make hypotheses
  - Track, reproduce, explain
  - Explore different hypothesis
- Context is essential
  - Always trace back to source data
- Need domain experts
  - Many hours of meeting

→Layered architecture

Layered architecture

- Layered architecture:
  - Successive layers
  - Data more and more transformed
  - Separate **import**, **normalization** and **interpretation**.

- Layers are Connected:
  - Can always **trace to original data**
  - Provides **a reproducible pipeline**.



Layered architecture

Original data files

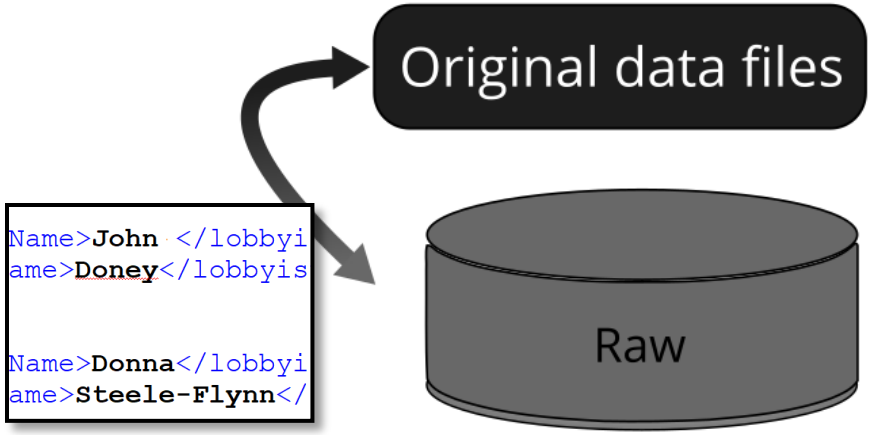
- Xml, csv, json, yaml, html ….

  …2010_4thQuarter_XML/**300345542.xml**

```
<lobbyists>
    <lobbyist>
        <lobbyistFirstName>John </lobbyistFirstName>
        <lobbyistLastName>Doney</lobbyistLastName>
    </lobbyist>
    <lobbyist>
        <lobbyistFirstName>Donna</lobbyistFirstName>
        <lobbyistLastName>Steele-Flynn</lobbyistLastName>
    </lobbyist>
</lobbyists>
```

Layered architecture

- Import (`COPY …`)
- Foreign Data Wrapper:
  - File_fdw (+ bash)
  - Extension Multicorn

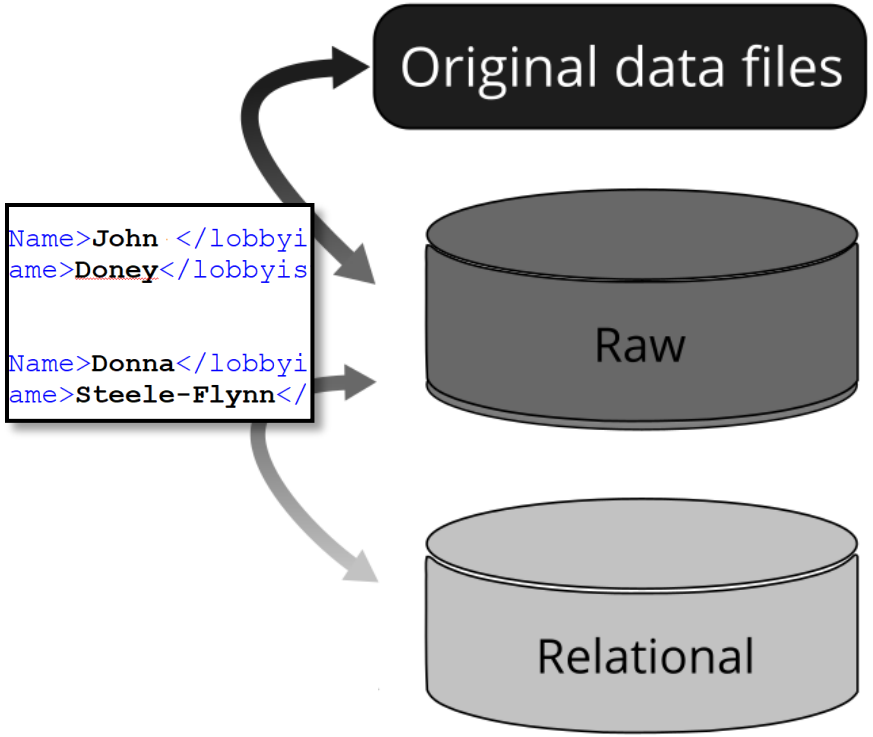



```
CREATE FOREIGN TABLE
raw___lobby.reports_fxml_fdw (
    xml_file_number int8 NULL,
    file_content bytea NULL,
    filename text NULL)
SERVER filesystem_multicorn_fdw
OPTIONS ...
```

	xml_file_number	file_content	filename
1	300,345,542	<LOBBYINGDISCLOSURE2> <import... [37676]	2010_4thQuarter_XML/300345542.xml

Layered architecture

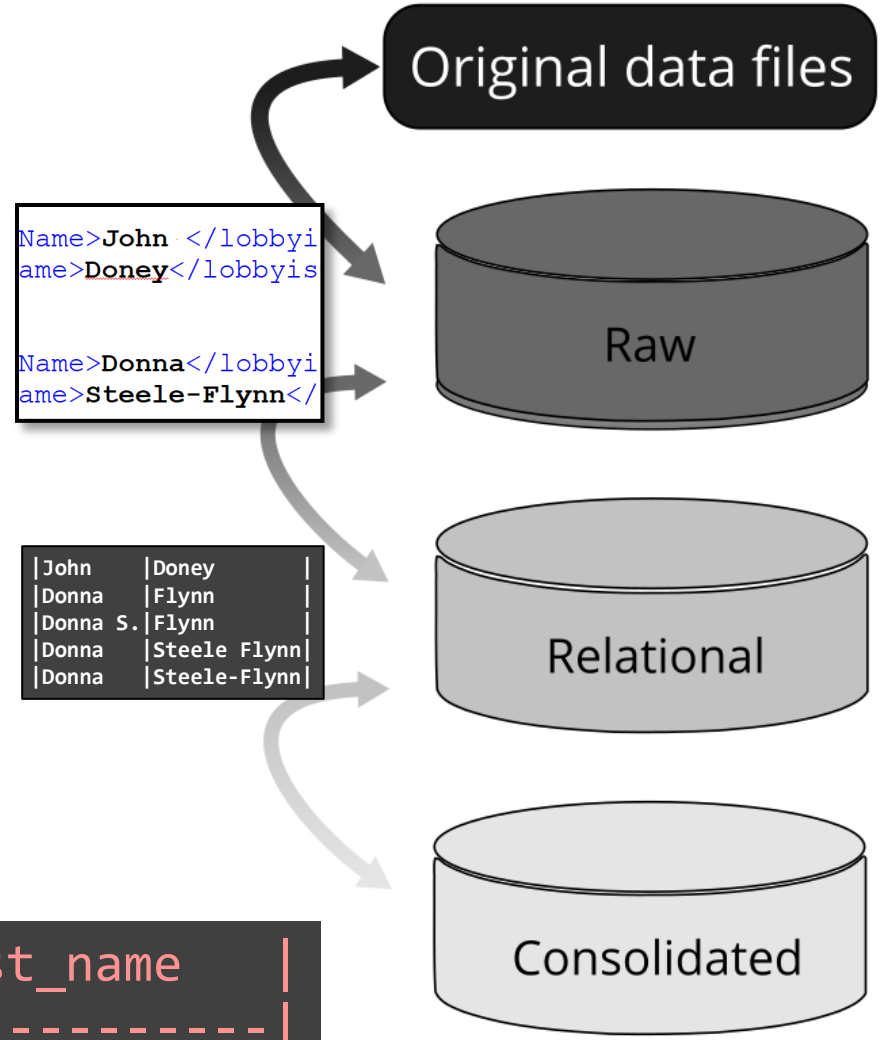
- Normalized, neutrally cleaned data
  - Case, trailing white space
    type casting …



_lobbyist_pk	first_name	last_name
38,667,273	Allison	Doney
37,791,646	John	Doney
36,429,385	Donna	Flynn
36,932,677	Donna S.	Flynn
36,293,145	Emily	Flynn
38,764,335	Diane	Steed
34,999,980	Burt	Steele
37,443,723	Donna	Steele Flynn
36,083,060	Donna	Steele-Flynn
36,645,401	James A.	Stem,

Layered architecture

- **Choices**
  - $entity_i = entity_j = \ldots$
  - ⚠ data is interpreted and transformed

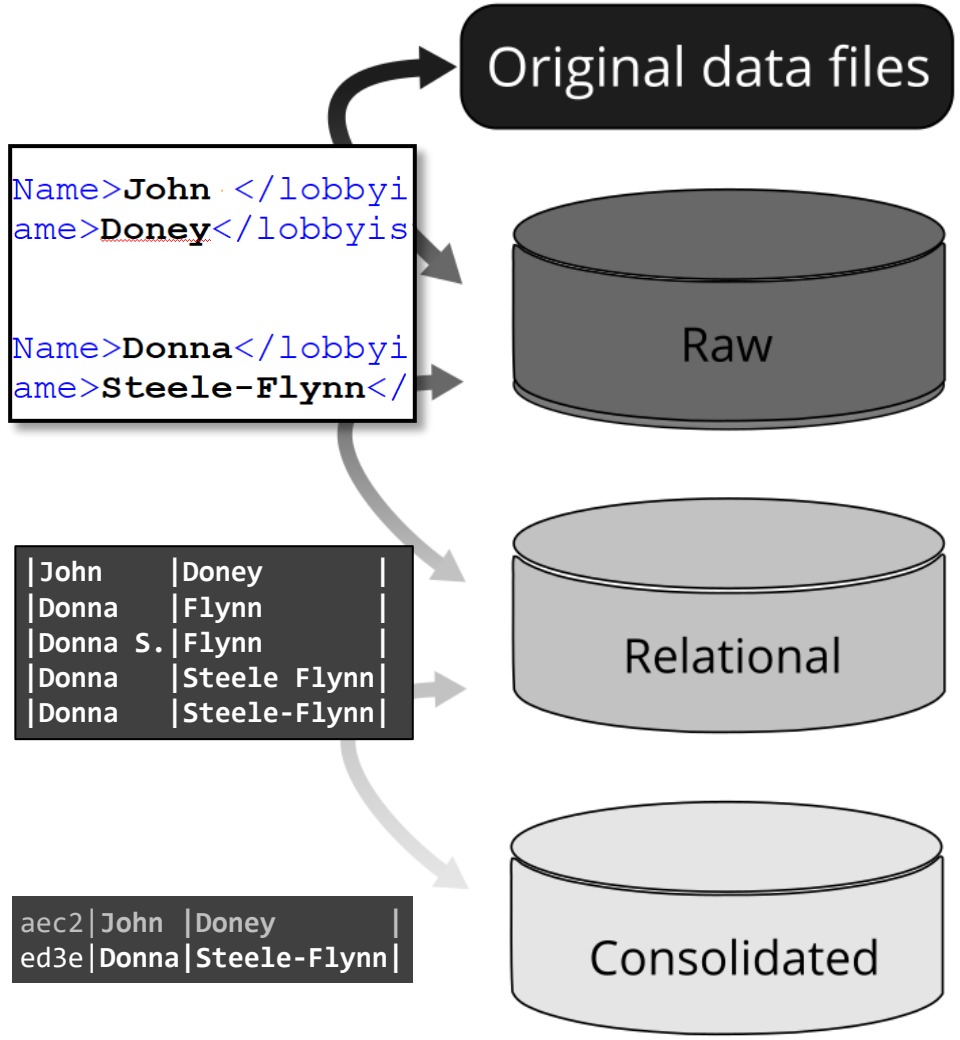


first_name	last_name
John	Doney
Donna	Flynn
Donna S.	Flynn
Donna	Steele Flynn
Donna	Steele-Flynn

uuid	first_name	last_name
aec2	John	Doney
ed3e	Donna	Steele-Flynn

Layered architecture

- Raw:
  - Original data, from Postgres
  - FDW
- Relational:
  - Normalize
- Consolidated:
  - Choices
  - Interpret
  - Consolidation



Strategies :

Strategy 3:
Conflicting data sources

Data is **fuzzy**
"About 10k$"

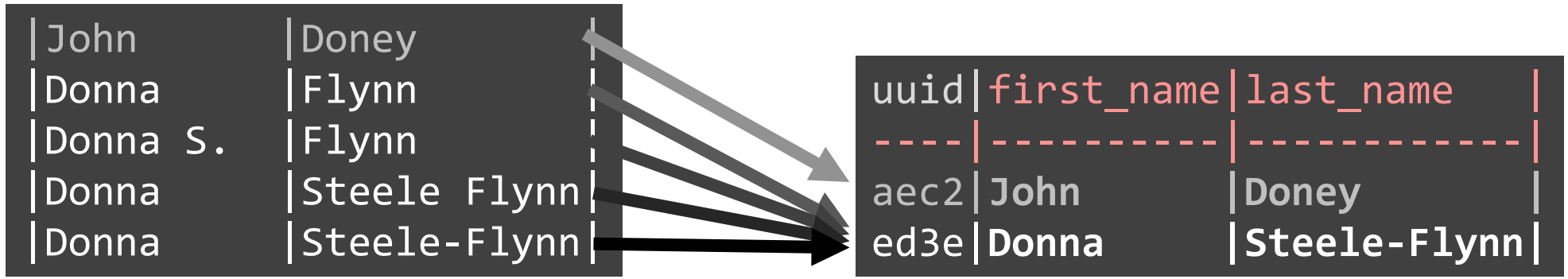
Data **definition** is ambiguous
"influence" on employees?

Conflicting data sources
"J. McCain" **VS** *"McCain, John S."*

Specific to **Social Sciences**:
Explore, be Explainable

Consolidation?

- Consolidation:
  Decide $entity_1 = entity_2$
- Can be any type of entity
  - People
  - Companies
  - Government agencies
  - …



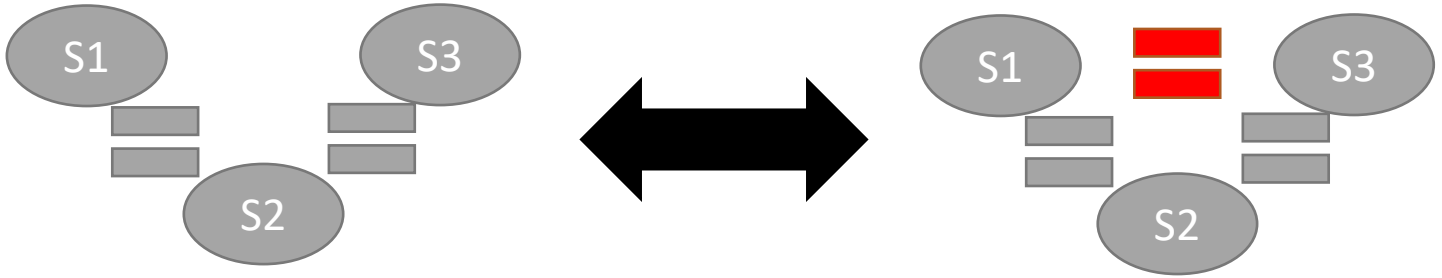
Consolidation: simple case

- Simplest solution:
  entity → cleaning → strict equality
- ex: US. States:

```
S1|CALIFORNIA
S2|California
S3|  california
```

👍 Very convenient

👍 Transitive



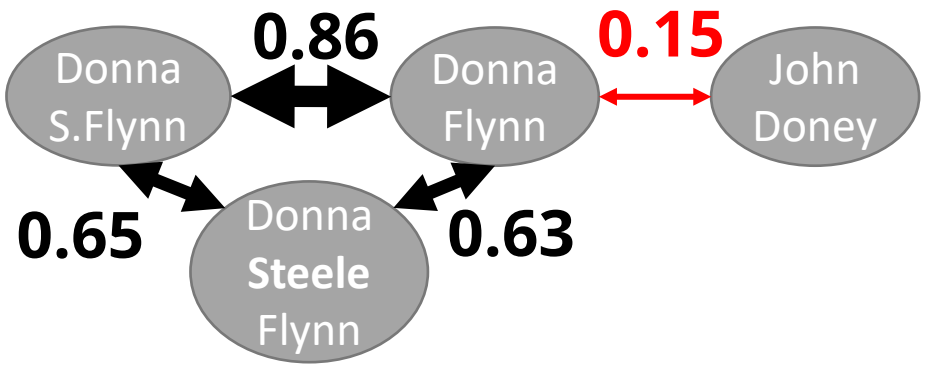
👎 Rare

Consolidation: hard case

- Compute similarity between entities
  - Can use many things
  - Ex: name fuzzy similarity
    - Extension Pg_trgm

```
L0|John       |Doney         |
L1|Donna      |Flynn         |
L2|Donna S.   |Flynn         |
L3|Donna      |Steele Flynn|
```

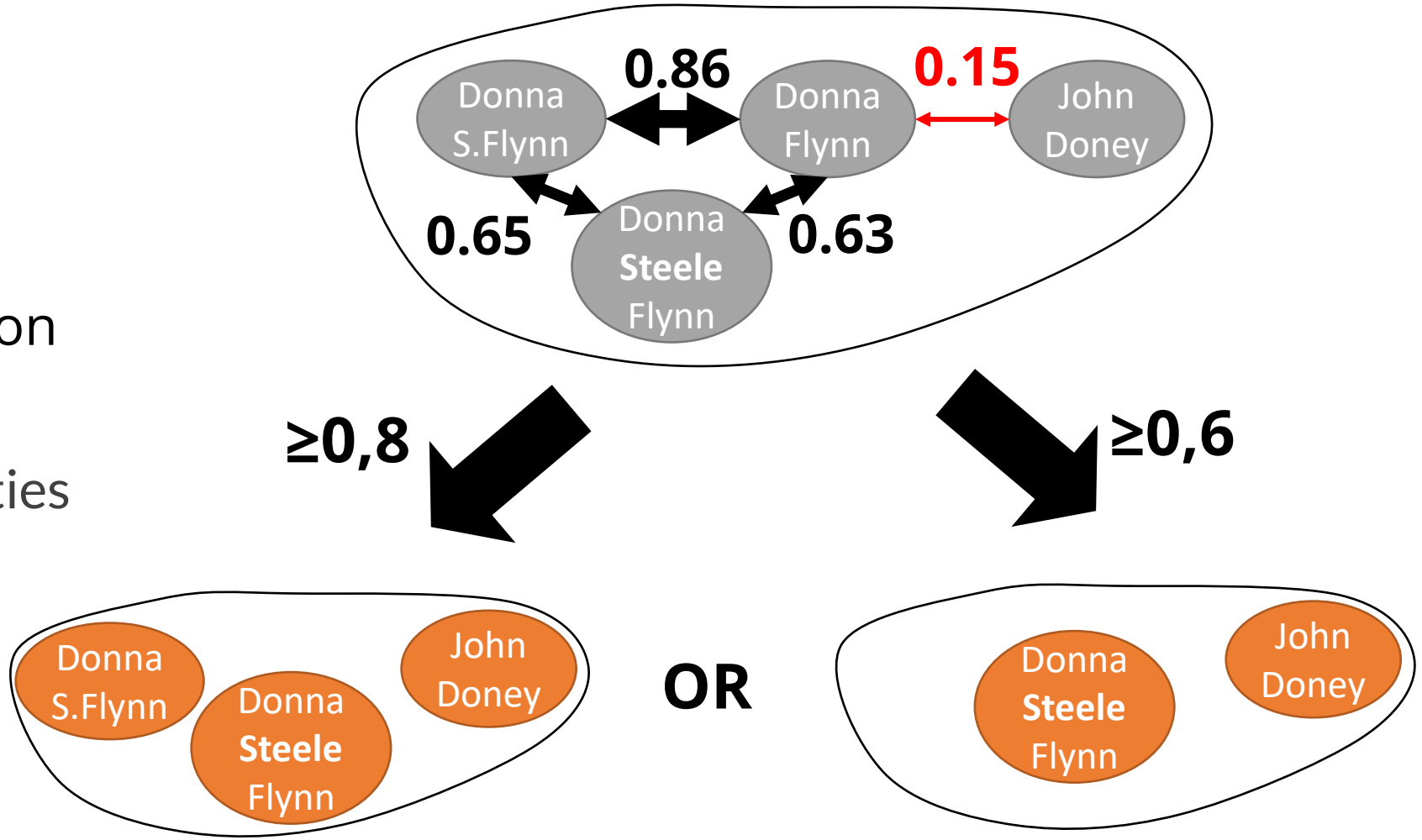
```
SELECT similarity('Donna Flynn', 'Donna Steele Flynn')-- 0.63
     , similarity('Donna Flynn', 'Donna S. Flynn')    -- 0.86
```



Consolidation: hard case

- Decide what is '**similar enough**' to be identical.
  - Manually
  - Using ML + Python

- Find Connected entities
  PlPython



Consolidation: very hard case

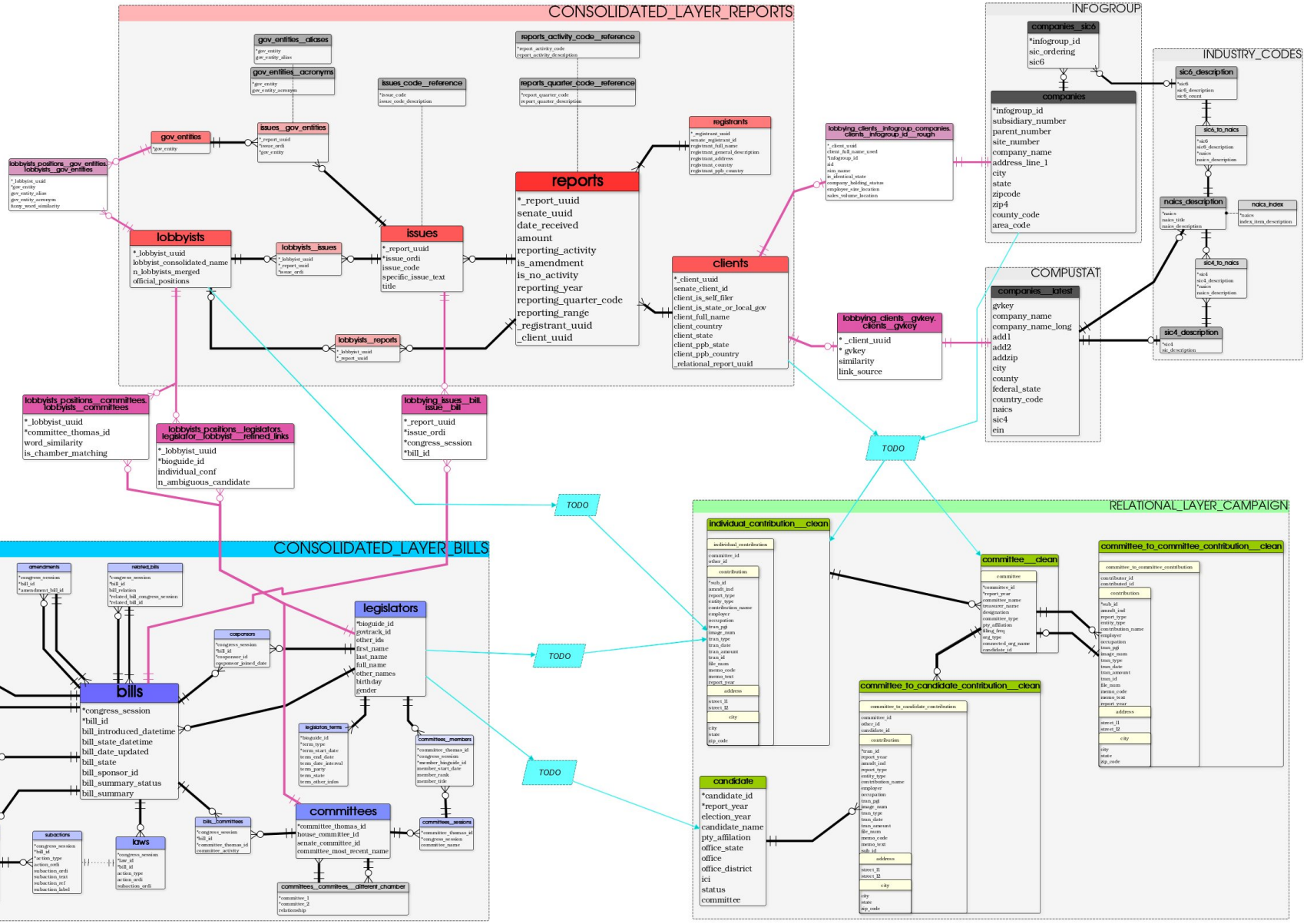
Lobbyist_name	employers	colleagues
DONNA, FLYNN	[ERNST & YOUNG]	[Tim URBAN]
STEELE FLYNN, DONNA	[ERNST & YOUNG]	[John DONEY; Tim URBAN]

- Also use **context**
  - Same colleagues
  - Same employer
  - Same topics
  - …

- ⚠Index?
  - Extension smlar to index array

Consolidation: very very hard case

Real life is another topic.



Strategies :

Strategy 4:
Explore, Explain the data

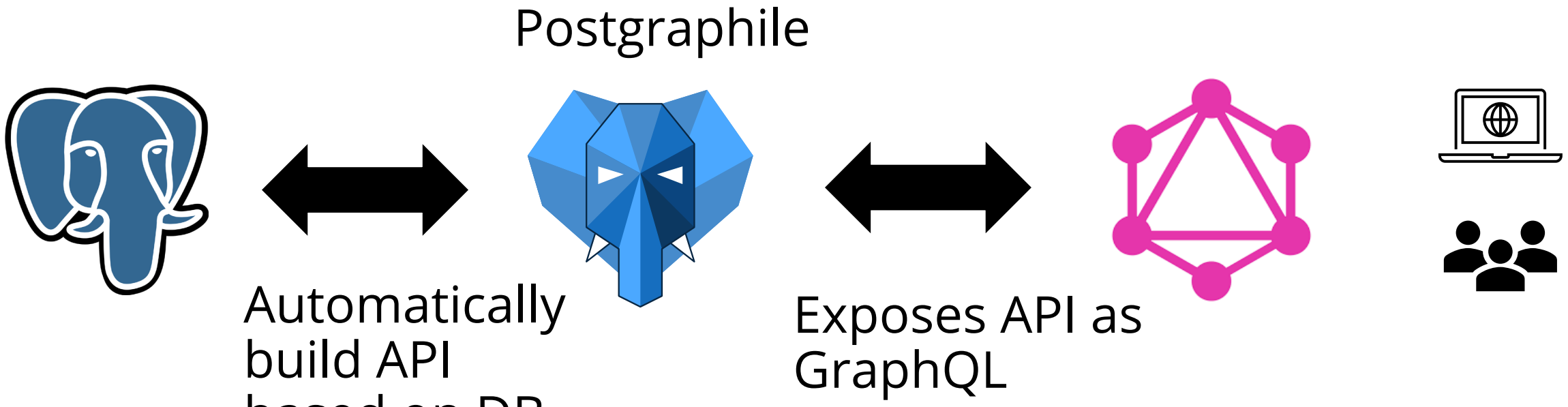
Data is **fuzzy**
"About 10k$"

Data **definition** is ambiguous
"influence" on employees?

Conflicting data sources
"J. McCain" **VS** *"McCain, John S."*

Specific to **Social Sciences**:
Explore, be Explainable

Explore, Explain the data?



Explore, Explain the data?

```
query lobbyistPerReport {
  reports(first: 10, orderBy: _REPORT_UUID_ASC) {
    amount
    reportingYear
    lobbyistsReportsByReportUuid {
      _lobbyist {
        lobbyistConsolidatedName
      }
    }
  }
}
```

```
{
  "amount": 20000,
  "reportingYear": 2018,
  "lobbyistsReportsByReportUuid": [
    {
      "_lobbyist": {
        "lobbyistConsolidatedName": "ROBERTS RICHARD"
      }
    },
    {
      "_lobbyist": {
        "lobbyistConsolidatedName": "GRADLER GEOFFREY"
      }
    }
  ]
}
```

- Limited
- Easier, nested

```
SELECT r.amount , r.reporting_year
, lobbyist_consolidated_name
FROM reports as r
  JOIN lobbyists__reports USING (_report_uuid)
  JOIN lobbyists USING (_lobbyist_uuid)
LIMIT 10;
```

```
amount    |reporting_year|lobbyist_consolidated_name|
----------|--------------|--------------------------|
$20,000.00|         2,018|ROBERTS RICHARD           |
$20,000.00|         2,018|GRADLER GEOFFREY          |
```

Explore, Explain the data?

- Website
  - filtering
  - visualization

Special Interest Group

Name ▾ APPLE

Politician (Last Name, First Name)

e.g. McCain, John

Report Text Keyword(s)

e.g. International Trade

Report Issues

e.g. Taxation

Year

e.g. 2014

Bill Id (Congress #_TypeBill#)

e.g. 111_HR3590

Bill Title

e.g. Patient Protection and Affordable Care Act

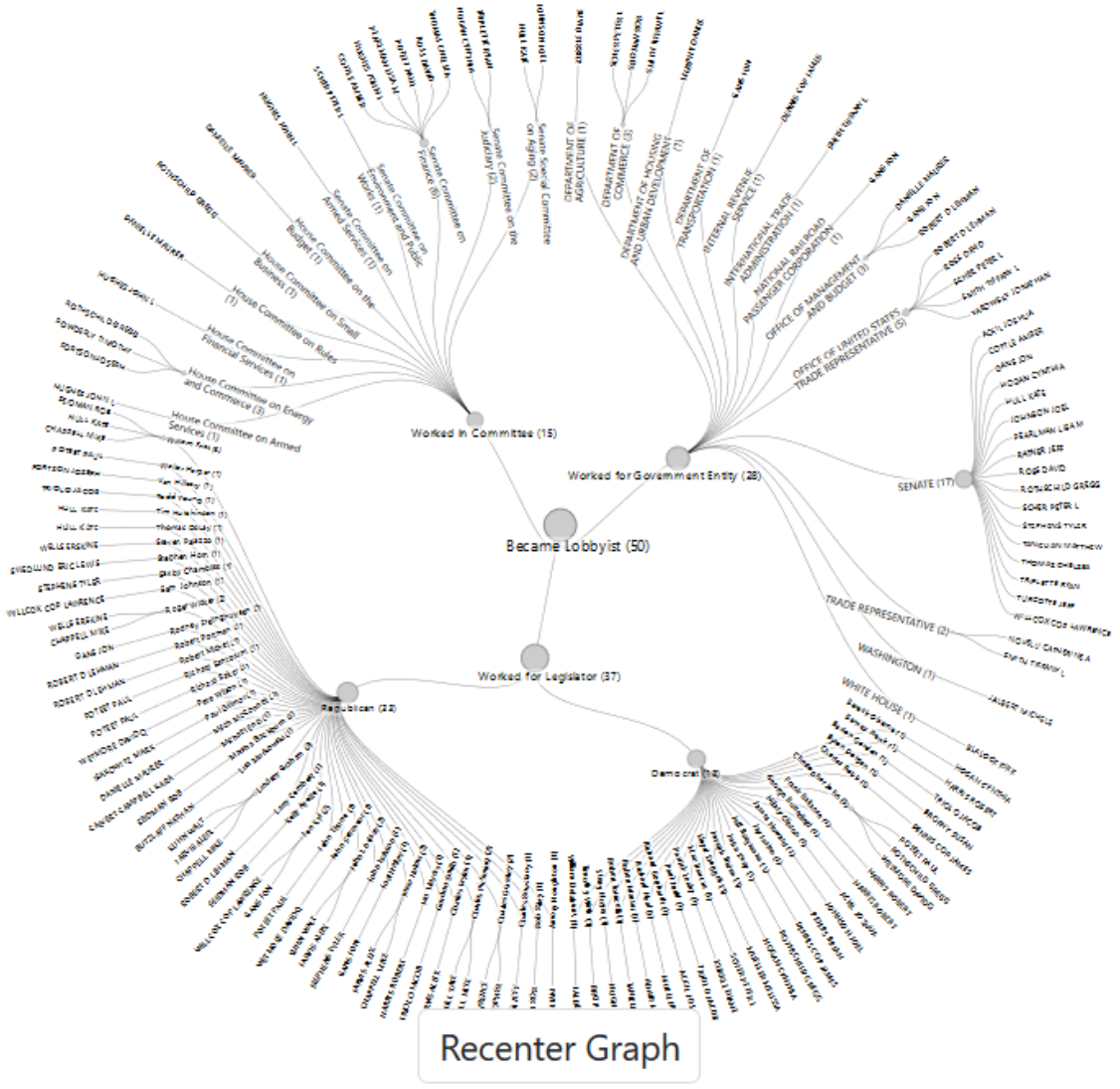
Government Entities

e.g. DOD

Registrant

e.g. Google

✕ Clear Search



Questions?

Many many thanks to **all Open Source projects and people** that make that work possible.
Postgres + extensions + PostGIS + QGIS + Postgraphile + …

THE END

APPENDIX

Solutions:

Database Modelling

Recipes: Data modelling

- Keys
  - Uuid : **not** random
    - Hash of actually unique column combination
    - Convenient when composite type too big
  - Composite type
    - + domain for constraint
    - Ex : bill ID
    - Prob: seem to be confusing for planner (bad indexing?).
  - Generated identity
    - = random
    - Prefer to avoid it.
    - Ordinality better

Recipes: Data modelling

- Light **Denormalization**
  - Data is duplicated between layers
    - Isolation + hides internal when needed (web API)
  - Json for sparse misc. things
    - Legislator youtube + twitter + …
  - When very convenient:
    - Year + reporting range.

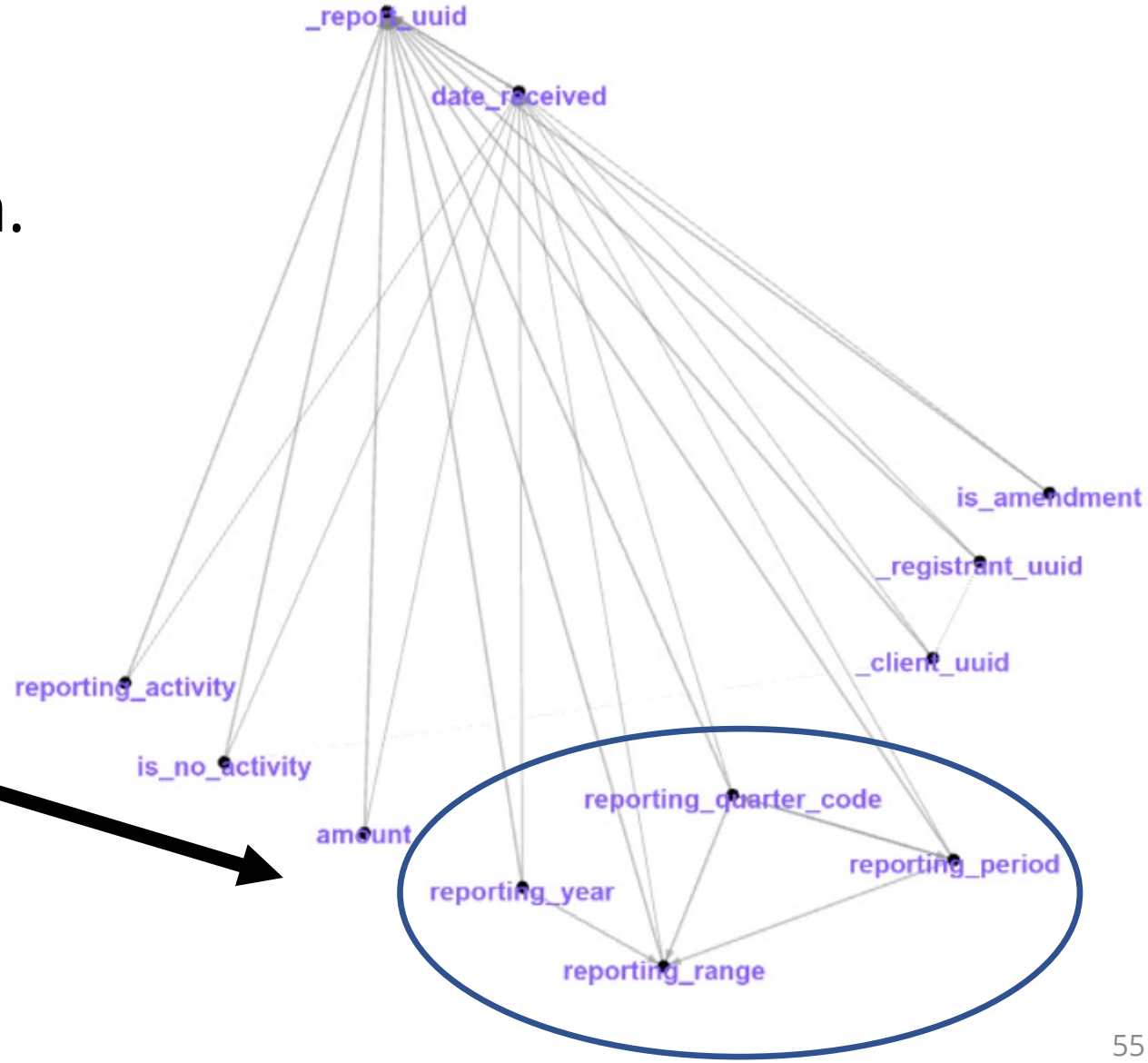
  - PG12: use of generated columns!
    - First_name, Last_name, … → full_name
    - (/!\, cumbersome with inheritance)

Recipes: Data modelling

- Ex of denormalization:
  Columns dependencies graph.

- Date_received = timestamp
  → almost perfect PK

- Some data duplication
  with report period



CSV issues

Database Modelling

Dataset: errors

pb	rid	gender	sw_politician_id	gov_politician_name	icpsr_politician_id	govtrack_id	fec_candidate_id	gov_terms	congress_num	congress_begin_yr	congress_end_yr	*senior_party_member*	chamber	icpsr_com_name	icpsr_com_id
A	14886	M	Shuster, Bud	Representative Shuster, Bill	14052	400374		House: 1973-2001	103	1993	1994	*21*	H	Public works and transportation	173
A	14887	M	Shuster, Bud	Representative Shuster, Bud	14052	400374		House: 1973-2001	103	1993	1994	*21*	H	Public works and transportation	173
A	14844	M	Shuster, Bill	Representative Shuster, Bud	20134	409888		House: 2001-Present	107	2001	2002	*0*	H	Transportation and infrastructur	173
A	14842	M	Shuster, Bill	Representative Shuster, Bill	20134	409888		House: 2001-Present	107	2001	2002	*0*	H	Transportation and infrastructur	173
B	502	F	Baldwin, Tammy	Senator Baldwin, Tammy	29440	400013	H8WI00018	Senate: 2013-Present | House: 1999-2013	113	2013	2014	*0*	S	Aging (special committee)	419
B	501	F	Baldwin, Tammy	Senator Baldwin, Tammy	29940	400013	H8WI00018	Senate: 2013-Present | House: 1999-2013	112	2011	2012	*0*	H	Energy and commerce	128
C	4553	M	Dingell, John D., Jr.	Representative Dingell, John D.	2605	400110	H6MI16034	House: 1955-2015	111	2009	2010	*0*	H	Energy and commerce	128
C	4551	M	Dingell, John D., Jr.	Representative Dingell, John D.	2605	400110	H6MI16034	House: 1955-2015	109	2005	2006	*21*	H	Energy and commerce	128
D	3140	M	Coats, Dan	Senator Coats, Daniel	14806	402675	S0IN00053	Senate: 1989-1999, 2011-2017 | House: 1981	114	2015	2016	*0*	S	Intelligence (select committee)	432
D	16521	M	Walberg, Tim	Representative Walberg, Tim	21144			House: 2007-2009, 2011-Present	115	2017	2018	*0*	H	Education and the workforce	124
E	1539	M	Boren, Dan	Representative Boren, Dan	20523	400645		House: 2005-2013	**109**	**2007**	**2008**	*0*	H	Financial services	113
E	1546	M	Boren, Dan	Representative Boren, Dan	20523	400645		House: 2005-2013	**109**	**2011**	**2012**	*0*	H	Intelligence (select)	242
E	8846	M	Kildee, Dan	Representative Kildee, Daniel T.	21372	412546	H2MI05119	House: 2013-Present	**115**	**2015**	**2016**	*0*	H	Financial services	113
F	6346	M		Senator Goodwin, Carte Patrick				Senate: 2010							
G	4563	M		Representative Djou, Charles K.				House: 2010-2011							
H	1314	M	Blunt, Roy	Senator Blunt, Roy	29735	400034	H6MO07128	Senate: 2011-Present | House: 1997-2011	110	2007	2008	*64*	H	Minority whip	661
H	1383	M	Boehner, John A.	Representative Boehner, John A.	29137			House: 1991-2015	112	2011	2012	*31*	H	Speaker	661
H	4926	M	Durbin, Richard J.	Senator Durbin, Richard J.	15021	300038	S6IL00151	Senate: 1997-Present | House: 1983-1997	109	2005	2006	*64*	S	Minority whip	662
I	318	M	Applegate, Douglas	Representative Applegate, Douglas	14402			House: 1977-1995	103	1993	1994	*0*	H	Transportation and infrastructur	173
I	455	M	Baker, Bill	Representative Baker, Bill	29310			House: 1993-1997	103	1993	1994	*0*	H	Public works and transportation	173
J	2936	M	Chiesa, Jeffrey	Senator Chiesa, Jeff	41307			Senate: 2013	113	2013	2014	*0*	S	Commerce, science, and transpc	321
J	2937	M	Chiesa, Jeffrey	Senator Chiesa, Jeff	41307			Senate: 2013	113	2013	2014	*0*	S	Homeland security and governm	321
K	17211	M	Wyden, Ron	Senator Wyden, Ron	14871	300100	**S6OR00110**	Senate: 1996-Present | House: 1981-1996	104	1995	1996	*0*	**H**	Commerce	128
K	17261	M	Wyden, Ron	Senator Wyden, Ron	14871	300100	**S6OR00110**	Senate: 1996-Present | House: 1981-1996	113	2013	2014	*12*	**S**	Energy and natural resources	330
P	4086	M	Daschle, Thomas A.	Senator Daschle, Thomas A.	14617	300031	S6SD00028	Senate: 1987-2005 | House: 1979-1987	107	2001	2002	*44*	S	Majority leader	662
P	11982	M	Nickles, Don	Senator Nickles, Don	14908	300150	S0AK00063	Senate: 1981-2005	107	2001	2002	*66*	S	Minority whip	662
P	12464	F	Pelosi, Nancy	Representative Pelosi, Nancy	15448	400314	H8CA05035	House: 1987-Present	113	2013	2014	*62*	H	Minority leader	661
P	16373	M	Van Hollen, Chris	Senator Van Hollen, Chris	20330	400415	H2MD08126	Senate: 2017-Present | House: 2003-2017	112	2011	2012	*24*	H	Deficit reduction (joint, select)	511
X	127	M	Akaka, Daniel K.	Senator Akaka, Daniel K.	14400	300001	S0HI00084	Senate: 1990-2013 | House: 1977-1991	110	2007	2008	*0*	S	Armed services	308

Dataset: errors

- A : duplicate name : merging went wrong
- B: same govtrack_id, different icprs_politician_id
- C, P: senior party member?
- D, F: gov_terms formatting
- E: gov_terms says present, year says 2018
- G: NA for committees
- H: icpsrs_com_id not coherent with com_name(+...)
- I, J: should be same committee name
- K: same fec_candidate_id, but campaign for Senate and House
- X : control