Relational DBs for Political Science: Why you should use one !

Rémi Cura
Pf. In Song Kim

Today's menu

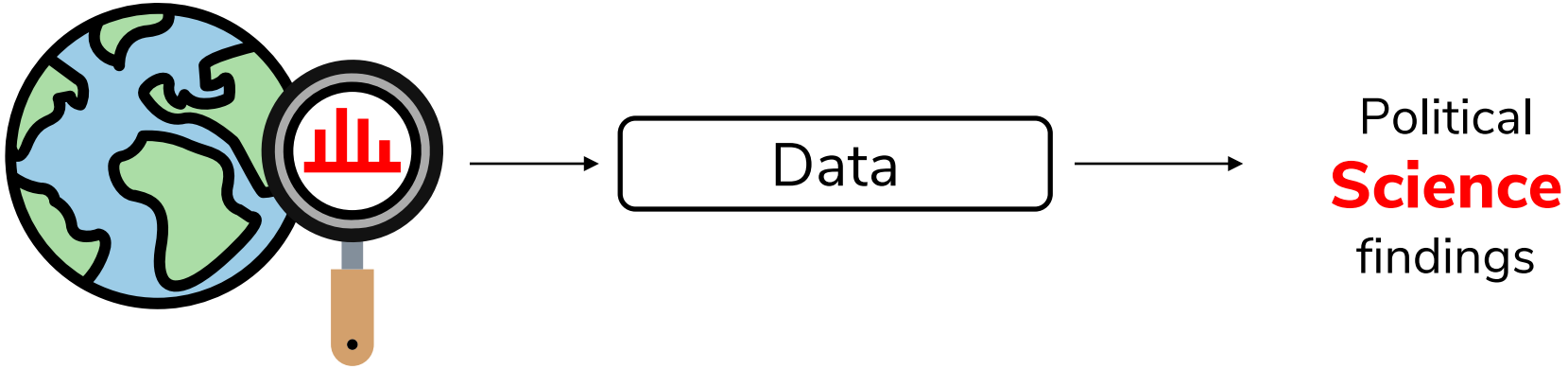
- Easier/Better data?
- Examples with the Lobbyview DB

DATA MATTERS

Data matters

- Why data matters:
  - Can your research function without data?
  - Can your papers be published without a data section?

  → If doubt on data, weakness.



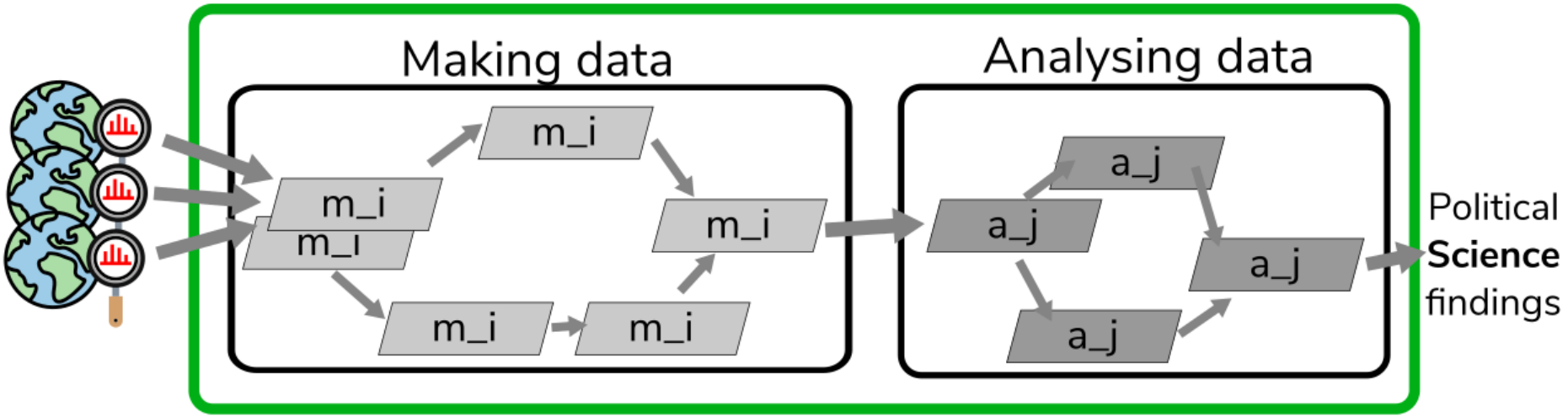
Science: data requirements?

Political **Science**

→requirements on all research :

- Reproducible : do it **again?**
- Trackable : **what** happened
- Explainable : **why** it happened.

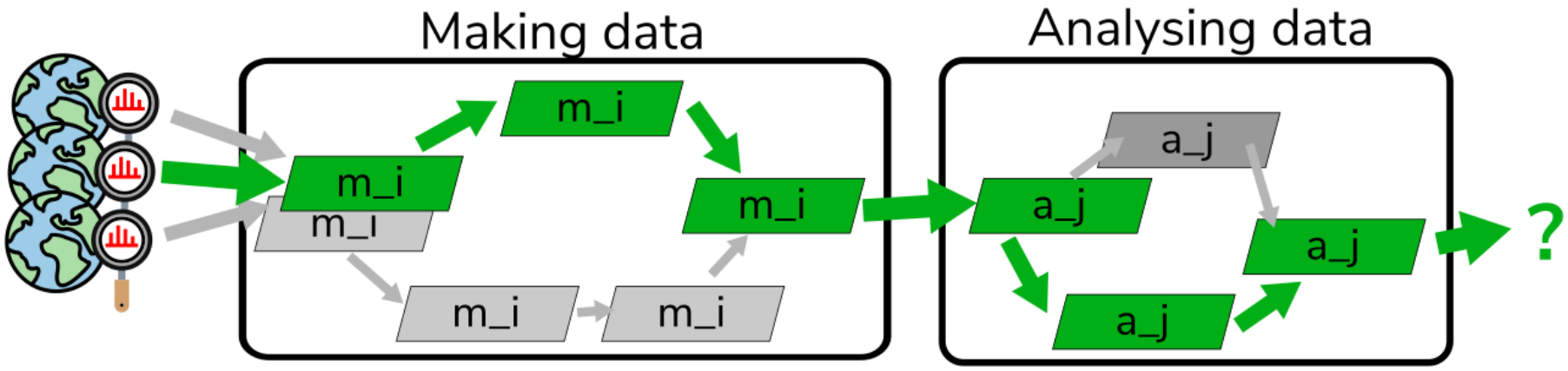
Science: reproducibility?



Reproducible:

- Updated data / Other scope?
- Sharing
  - Researcher
  - your (future) self !

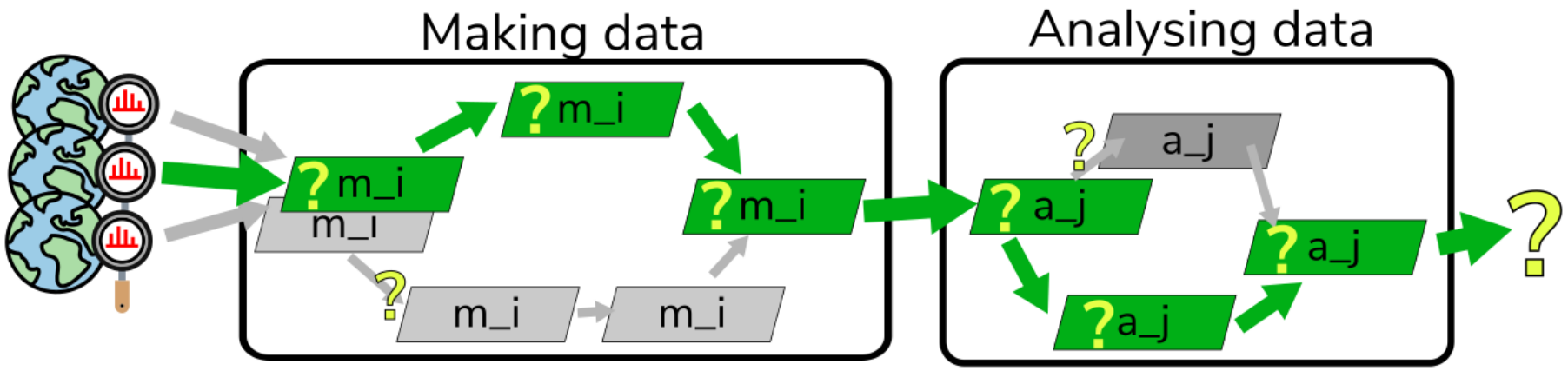
Science: trackability?



Trackable:

- Sources?
- Processes applied?

Science: explainability?



Explainable:

- Why it happened
- Why it didn't happen

→ Also for data cleaning !

Science : data requirements?

- Let's say you
  - Document everything
  - Track
  - Justify everything

Science : data requirements?

All set ,

thank you for your attention

Questions?

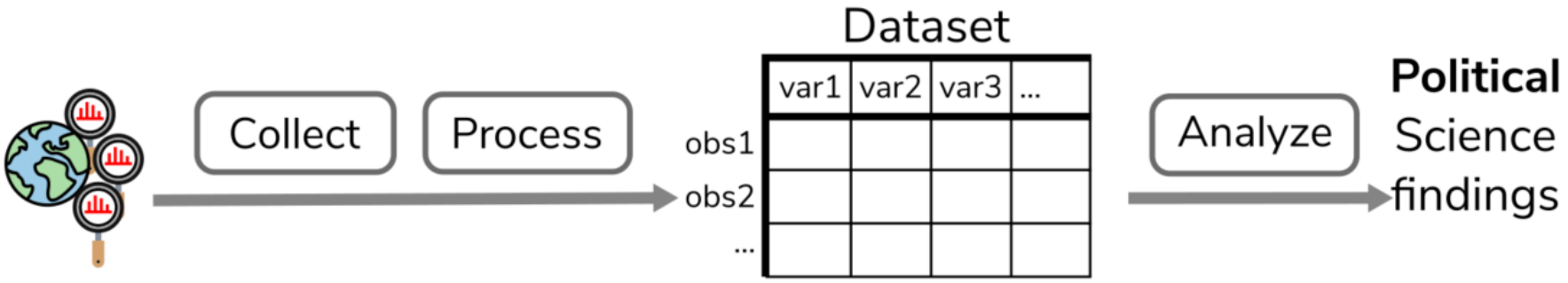
Political Science specificities

Kidding…

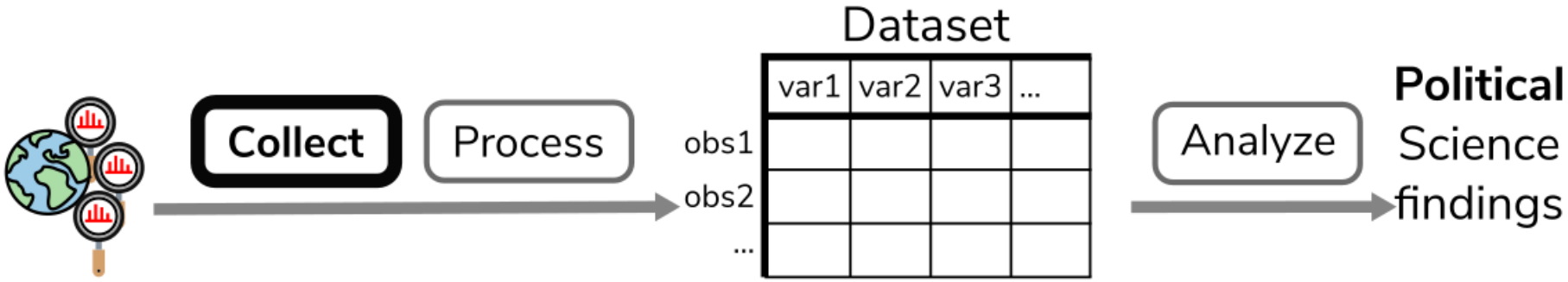
Reproducible + Trackable + Explainable: generic

What is specific for Political Science?

Typical workflow



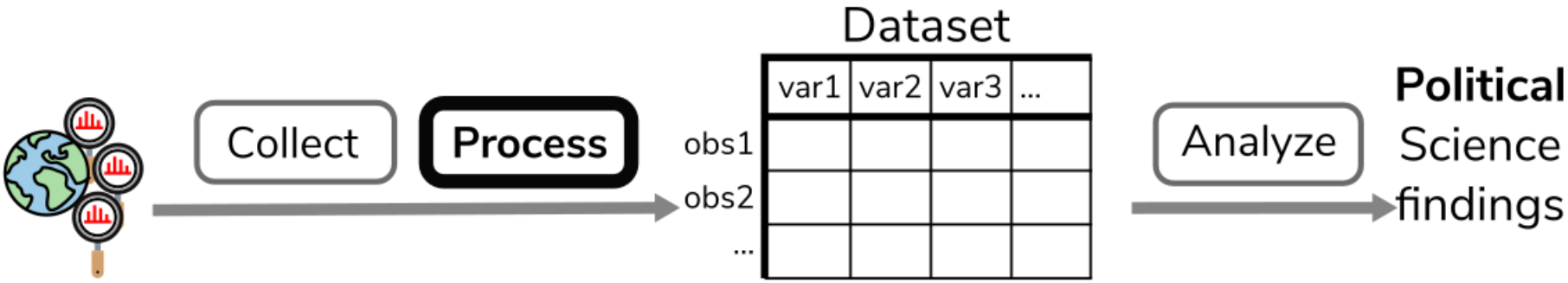
Typical workflow



Collect:

- survey
- scrape web
- existing datasets
- ...

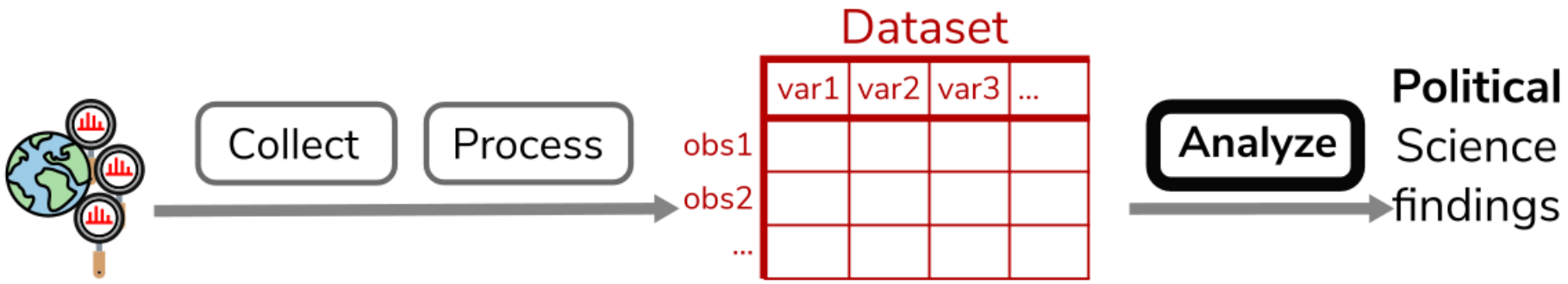
Typical workflow



Process:

- extract
- clean
- merge
- ...

Typical workflow

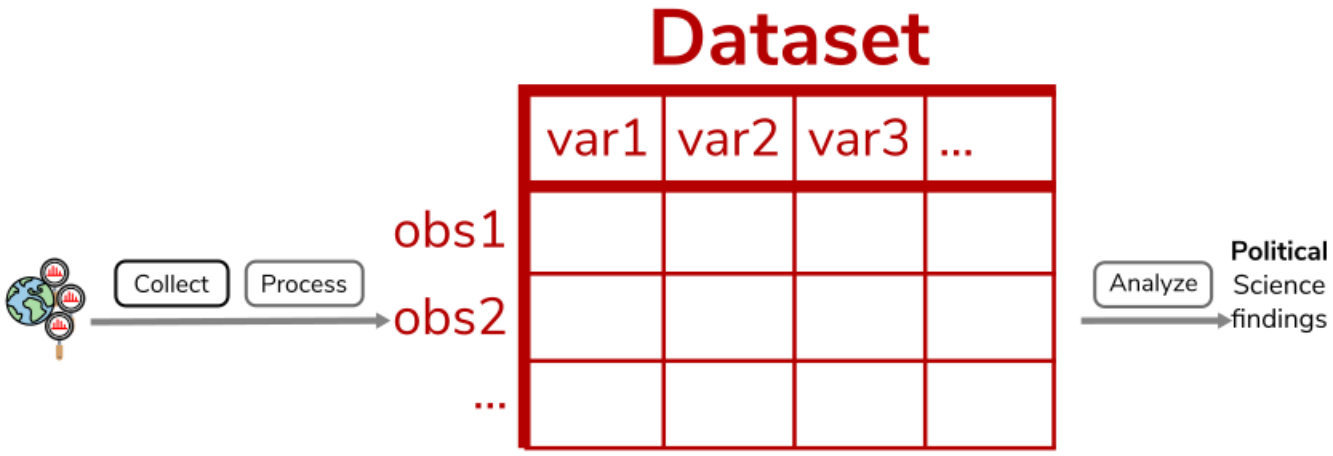


Analyze:

- regression
- causal inf
- ML
- model fitting

Typical workflow

Revolves around the mighty dataset…

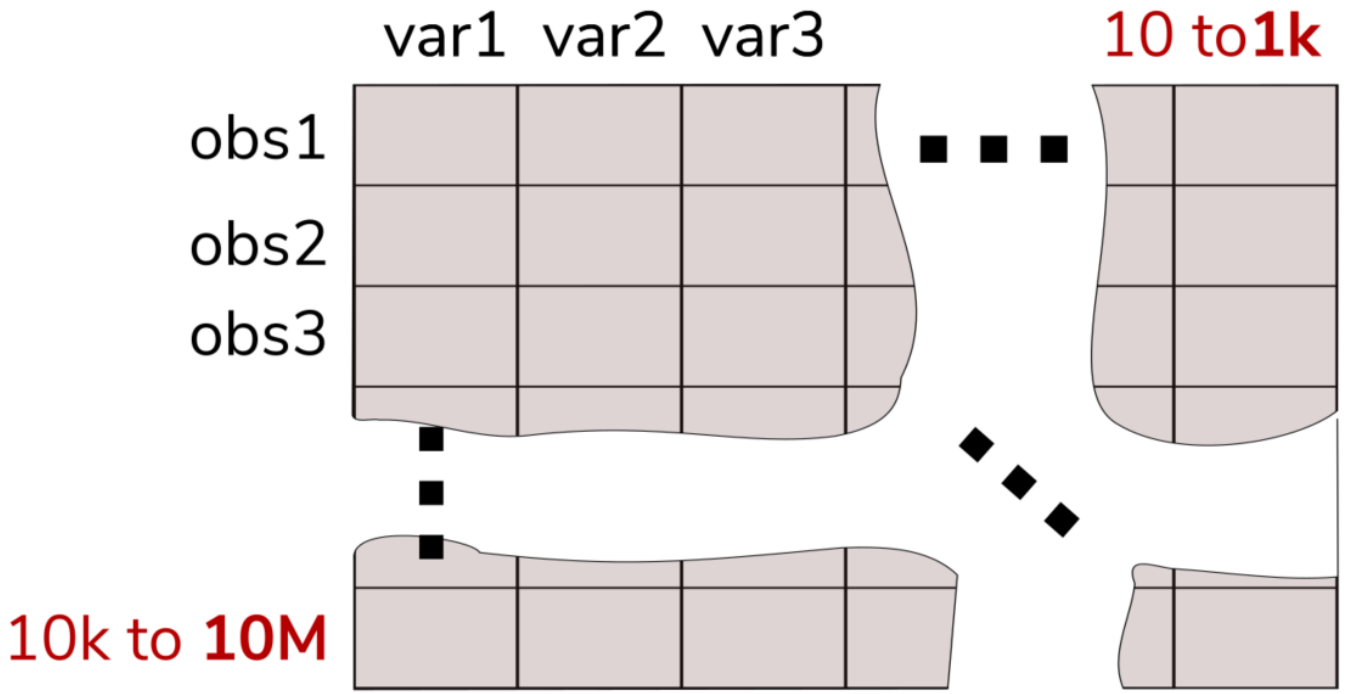


ABOUT DATASET

Strength and weakness

Dataset: pros

- What is it (typically) : a big (sparse + redundant) matrix



Dataset: pros

- What is it good for?

 – Analysis

 – Simplicity

 – Dirty sharing

Dataset: pros

- Very good for analysis

  → keep using it for that !

- Simple

  → good for quick and dirty

… Bad for everything else

Dataset: cons

- What is it bad for?

– Storage

– Update

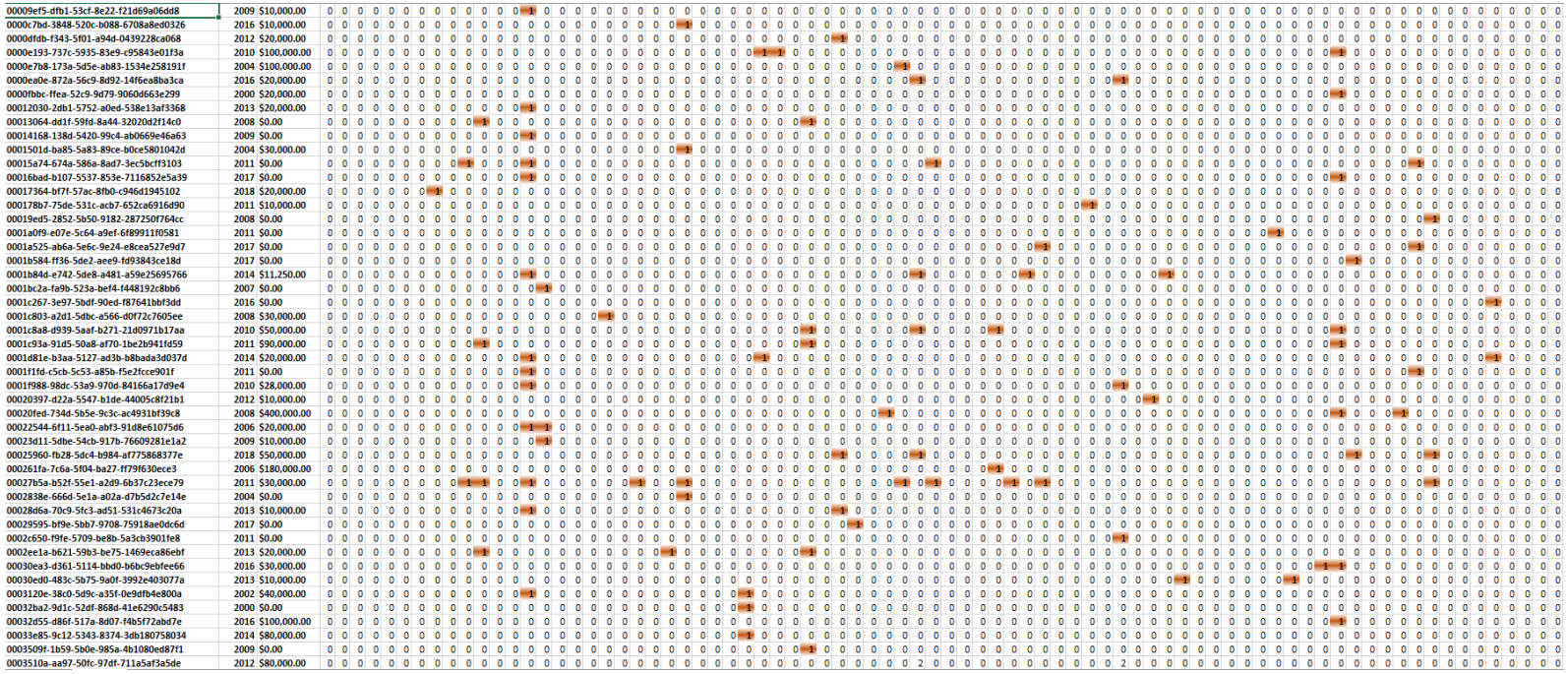
– Reuse / sharing

– Quality

Dataset: cons: storage

- Storage

Lots of columns → very sparse → bad for storage
(e.g. In Song Trade dataset: **1.5 TeraB, should be $^{1}/_{100}$)**



Dataset: cons: storage

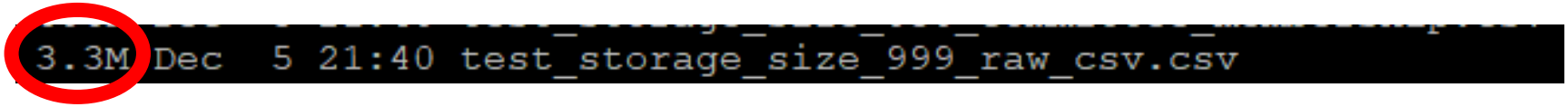
 - Update and re-use

Lots of duplicated values → bad for update and re-use

Dataset: cons: storage

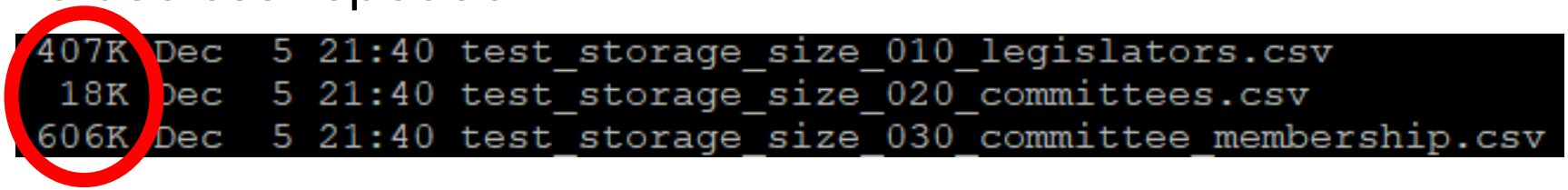
abercrombie	M	Abercrombi	Representat	15245			106	H	Resources 164
abercrombie	M	Abercrombi	Representat	15245			107	H	National security 106
abercrombie	M	Abercrombi	Representat	15245			107	H	Resources 164
abercrombie	M	Abercrombi	Representat	15245			108	H	National security 106
abercrombie	M	Abercrombi	Representat	15245			108	H	Resources 164
abercrombie	M	Abercrombi	Representat	15245			109	H	Armed services 106
abercrombie	M	Abercrombi	Representat	15245			109	H	Resources 164
abercrombie	M	Abercrombi	Representat	15245			110	H	Armed services 106
abercrombie	M	Abercrombi	Representat	15245			110	H	Natural resources 164
abercrombie	M	Abercrombi	Representat	15245			111	H	Armed services 106
abercrombie	M	Abercrombi	Representat	15245			111	H	Natural resources 164
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	114	H	Agriculture 102
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	114	H	Science, space, and technology 182
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	114	H	Veterans affairs 192
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	115	H	Agriculture 102
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	115	H	Armed services 106
abraham	M	Abraham, R	Representat	21522	412630	H4LA05221	115	H	Science, space, and technology 182
abraham	M	Abraham, S	Senator Abr	49500	400555		104	S	Budget 316
abraham	M	Abraham, S	Senator Abr	49500	400555		104	S	Commerce, science, and transportat 321
abraham	M	Abraham, S	Senator Abr	49500	400555		104	S	Judiciary 358
abraham	M	Abraham, S	Senator Abr	49500	400555		104	S	Labor and human resources 362
abraham	M	Abraham, S	Senator Abr	49500	400555		105	S	Budget 316
abraham	M	Abraham, S	Senator Abr	49500	400555		105	S	Commerce, science, and transportat 321
abraham	M	Abraham, S	Senator Abr	49500	400555		105	S	Judiciary 358
abraham	M	Abraham, S	Senator Abr	49500	400555		106	S	Budget 316
abraham	M	Abraham, S	Senator Abr	49500	400555		106	S	Commerce, science, and transportat 321
abraham	M	Abraham, S	Senator Abr	49500	400555		106	S	Judiciary 358
abraham	M	Abraham, S	Senator Abr	49500	400555		106	S	Small business 384
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		107	H	Agriculture 102
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		107	H	Resources 164
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		107	H	Small business 184
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		108	H	Agriculture 102
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		108	H	Resources 164
acevedo-vila	M	Acevedo-Vil	Representat	70601	400002		108	H	Small business 184

Dataset: cons: storage



3.3M : raw (dirty) CSV file

No useless repetition?



1.2M : 3 equivalent (dense) CSV files

Big dataset : RAM issue

Dataset: cons: storage

 - Quality nightmare

- Data easily corrupted (encoding, typo)
- No types
- No specific values
- No coherence

= no guarantees whatsoever

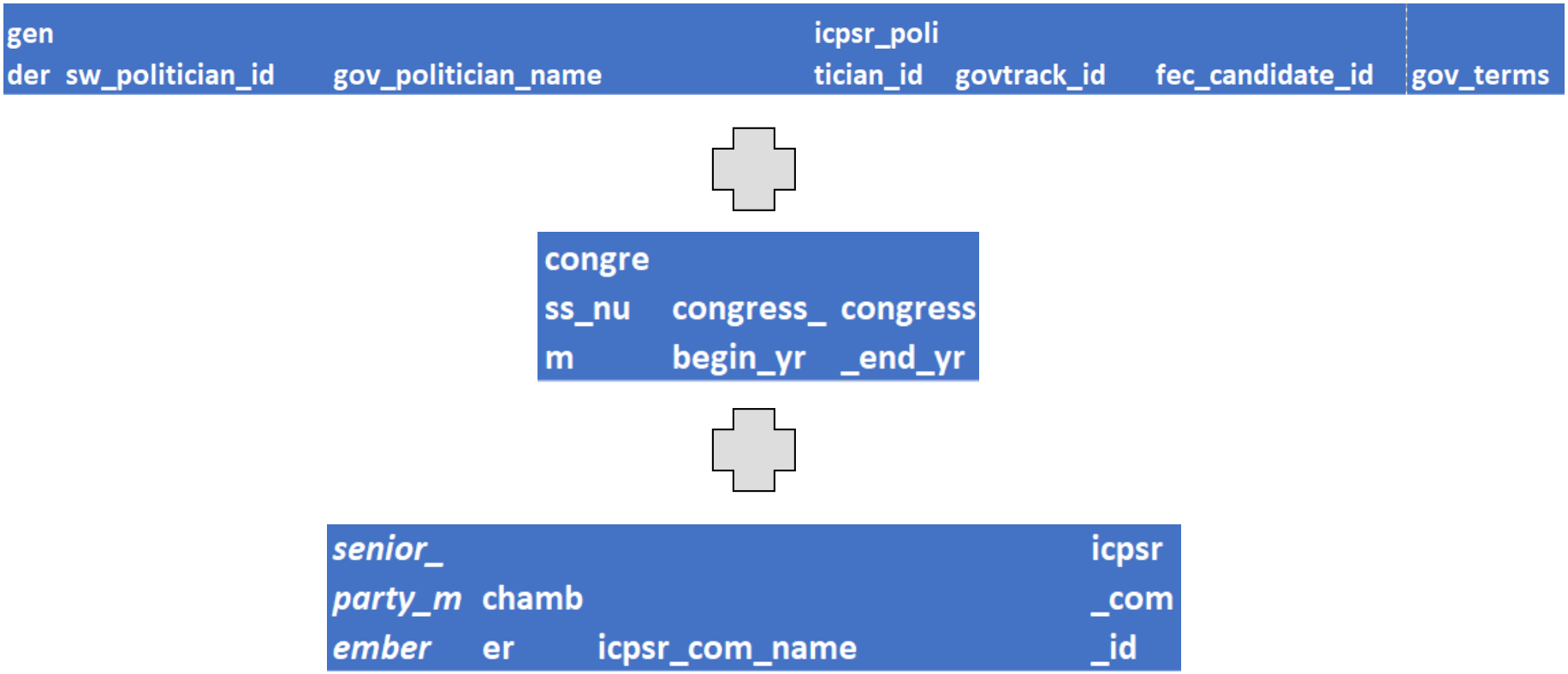
→ It will go bad

WORKSHOP: WHERE'S WALDO?

Find the problems in the (real life) dataset

Dataset: errors

Where's Waldo?



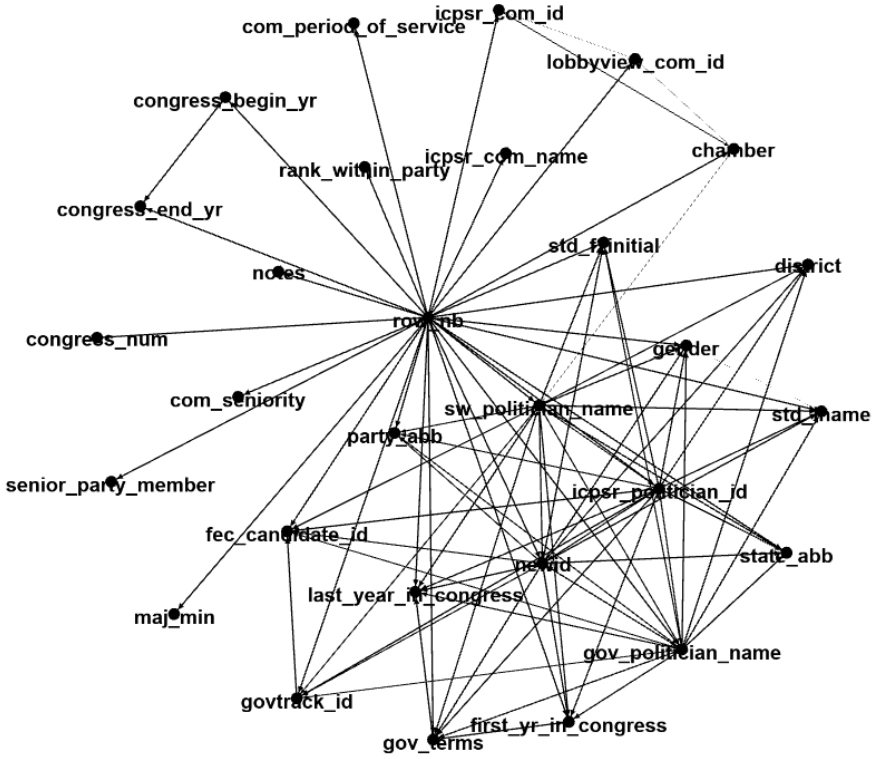
Dataset: errors

Where's Waldo?

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

Relational model

- Columns depends on others (you can guess the content)



Dataset: errors

Many errors.
What are the consequences?

- People get counted too much/not enough
- Over/under estimating effect
- Add noise

Dataset: errors

Bad news:

- Most cases : silent error
- After dozens of CSV :
  error rate can be **10%**

- **In theory, any effect < 10% could be a fluke…**

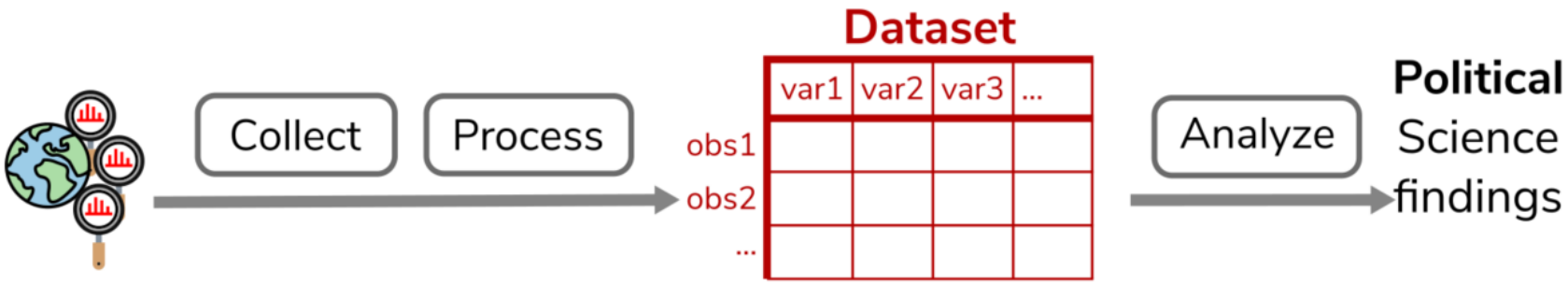
Dataset: solution

- Dataset issues get worse as:
  - Complexity 
  - Scale 
  - Users/Colleagues/Scope 

WHAT CAN WE DO?

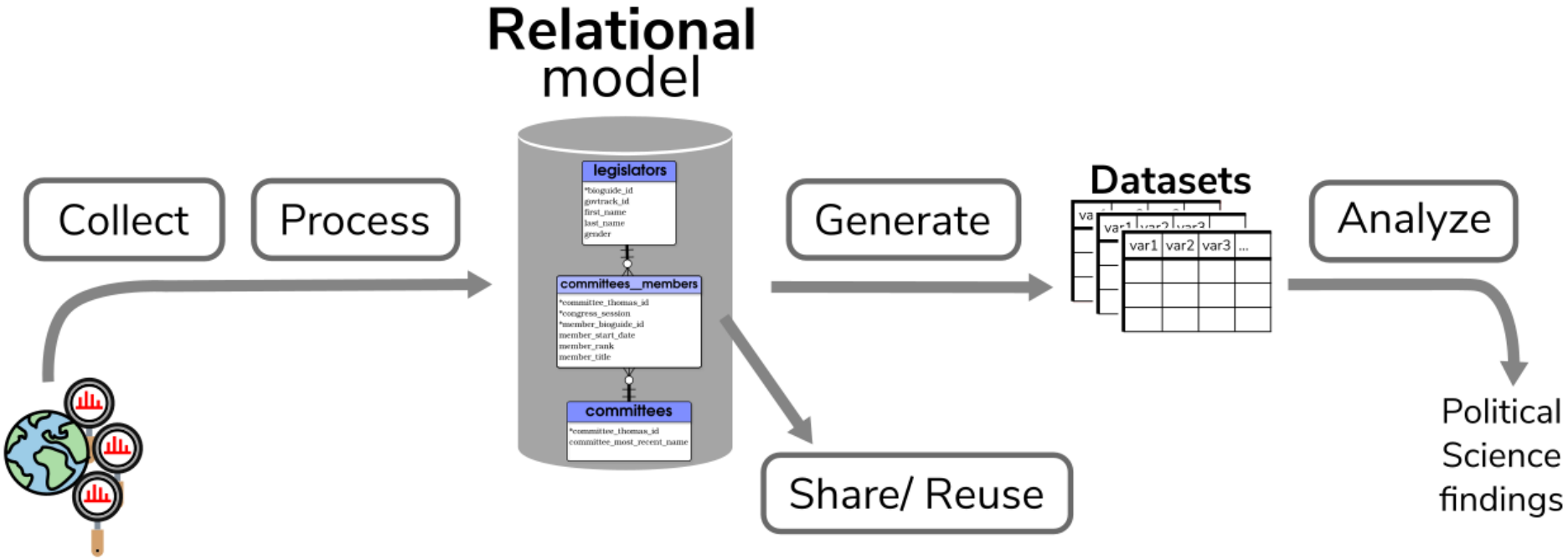
Dataset: solution

- Instead of



Dataset: solution

- Do:



RELATIONAL MODEL

Organize the data

Relational model: goal

- Goal is to :
  - Organize the data
  - Enforce the structure
  - Then generate the datasets as needed

Relational model: what is it?

- Relational model:
  - Entities: legislator / a committee / a bill …
  - What uniquely defines it?
  - Relations
    - 1 Legislator is in 0 or N committees
    - 1 committee has 1 or N legislators
    - …

Relational model: legislator-commitees

- Relational model:

– One real life person
  - Whatever the chamber
– All info that does not change for this person
– A unique id

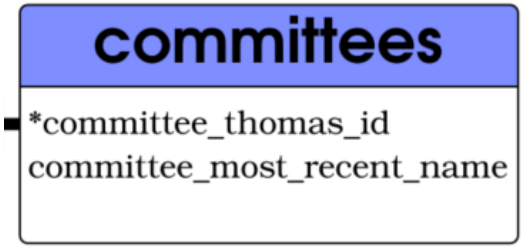
Relational model: legislator-commitees

- Relational model: Example in SQL


```
CREATE TABLE
consolidated_layer_bills.legislators (
        bioguide_id text,
        govtrack_id text,
        first_name text,
        last_name text,
        gender varchar(1) CHECK(gender='F' OR
        gender='M'),
            PRIMARY KEY (bioguide_id),
            UNIQUE (govtrack_id)
);
```

Relational model: legislator-commitees

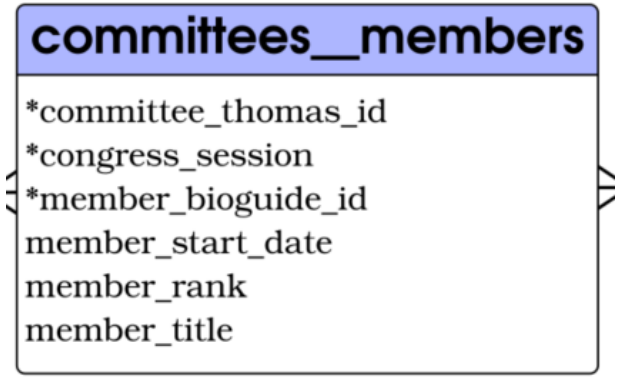
- Relational model:



  - One real life committee
  - A unique id
  - The most recent name (official name)

Relational model: legislator-commitees

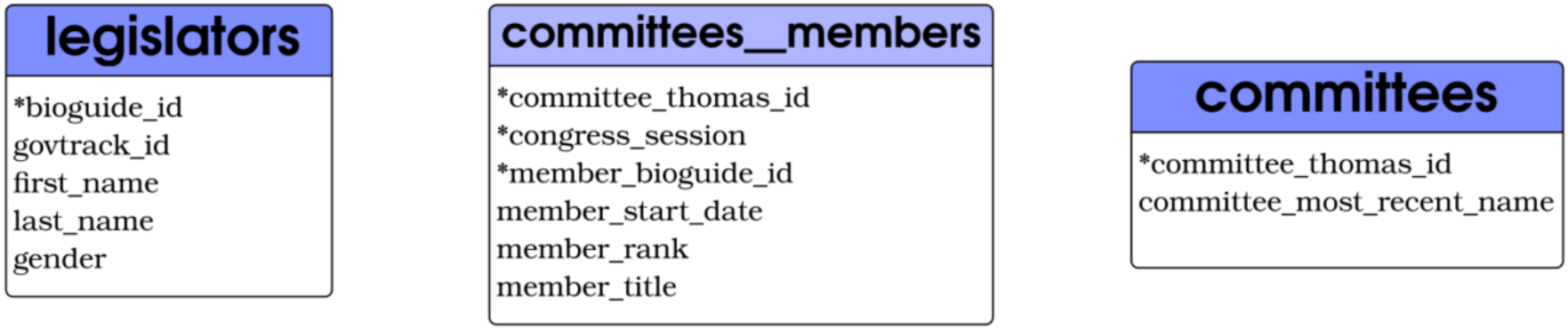
- Relational model:



  - Committee membership
  - For one congress,
    (legislator, committee) is unique
  - Info about this membership

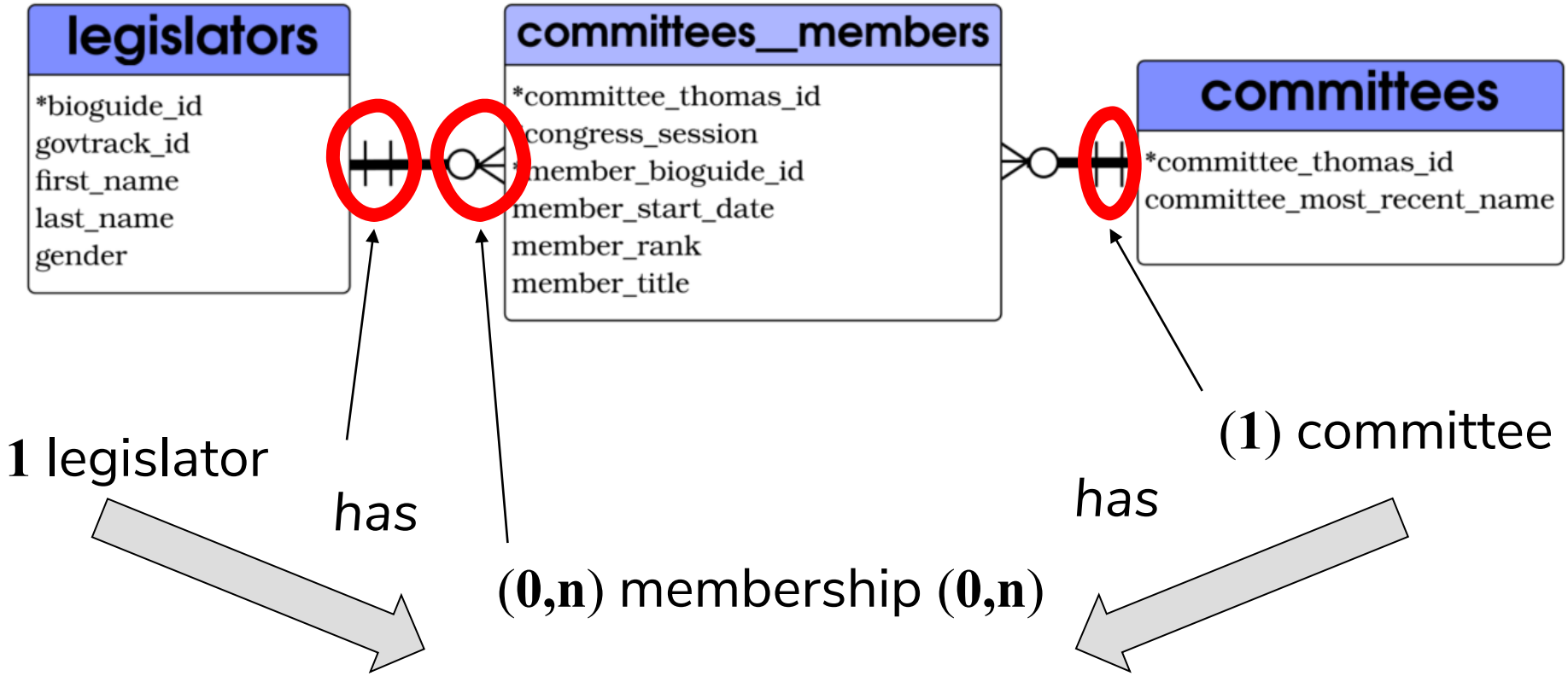
Relational model: legislator-commitees

- Relational model:



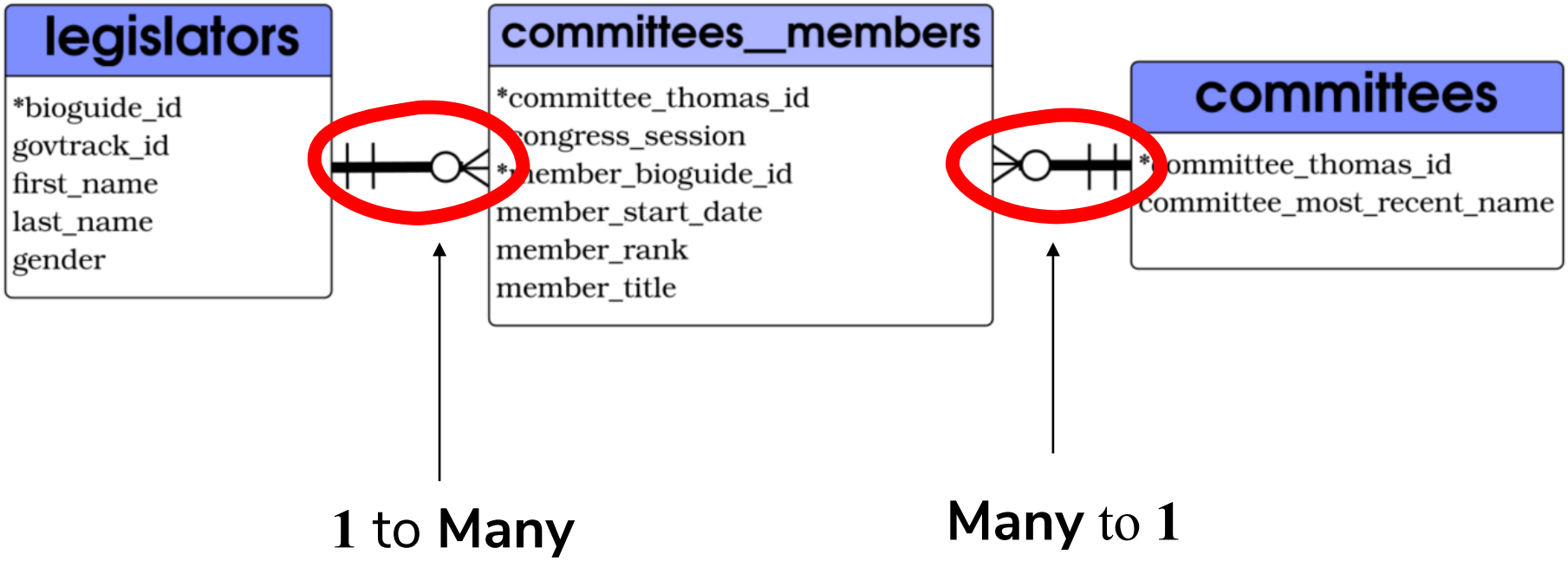
Relational model: legislator-commitees

- Relational model:



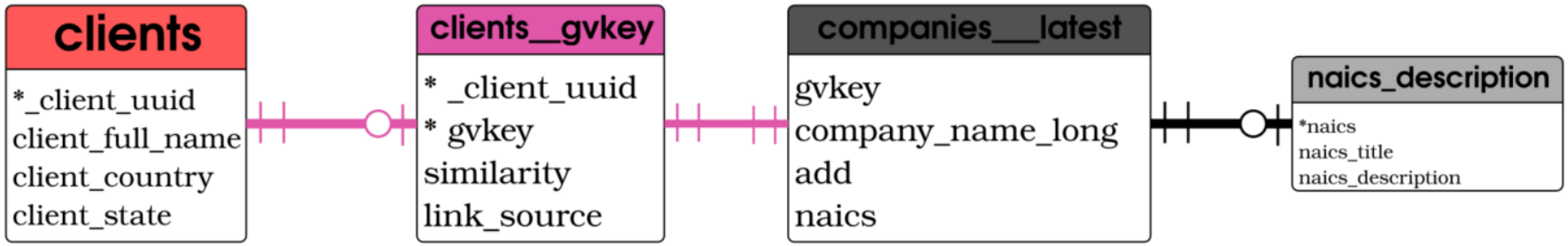
Relational model: legislator-commitees

- Relational model:



Relational model: Lobbying clients-Naics

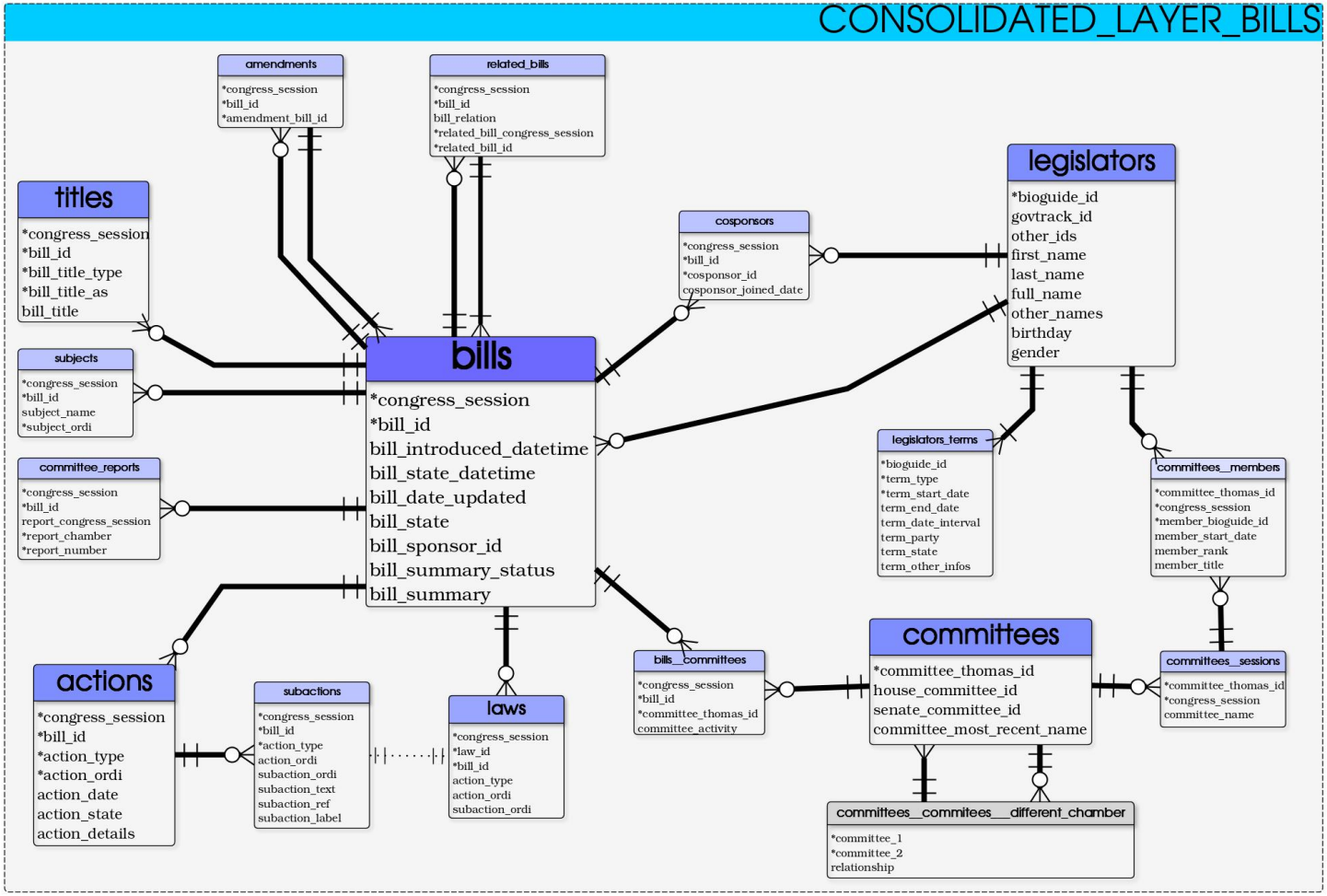
- Relational model:



(1) lobbying client *has* (0,1) gvkey *associated to* (1) companies *with* (1) Naics code

Relational model

Could be a lot of work:



RELATIONAL MODEL: BENEFITS

Why bother?

Relational model: benefits

- How many months collecting the data?
  - **many**
- How much care designing the experiments
  - **A lot**
- Will you update/reuse?
  - **Very likely**
- Do you want to share/disseminate your work?
  - **A lot**

Relational model: benefits

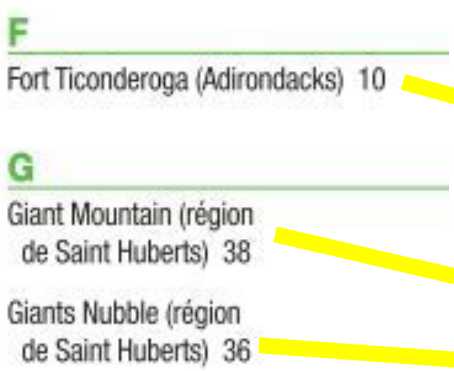
Can you afford wasting part of your work?

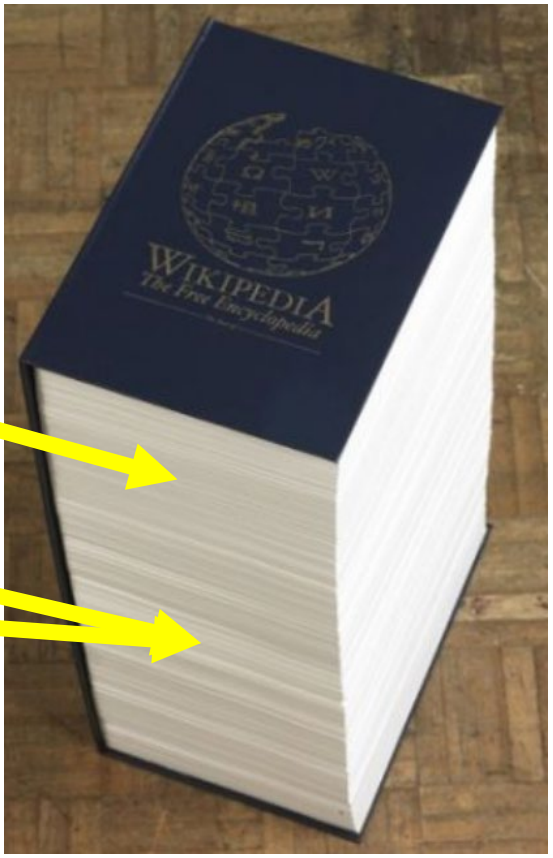
- Better data quality
- Scaling (complexity, volume, several users…)
- Generating new datasets
- Easier sharing/merging/reusing
- Some *very* advanced capabilities

Relational model: benefits

Scaling?

- ~~50 years of optimization?~~
- A good model reduces the dimensionality
- indexing

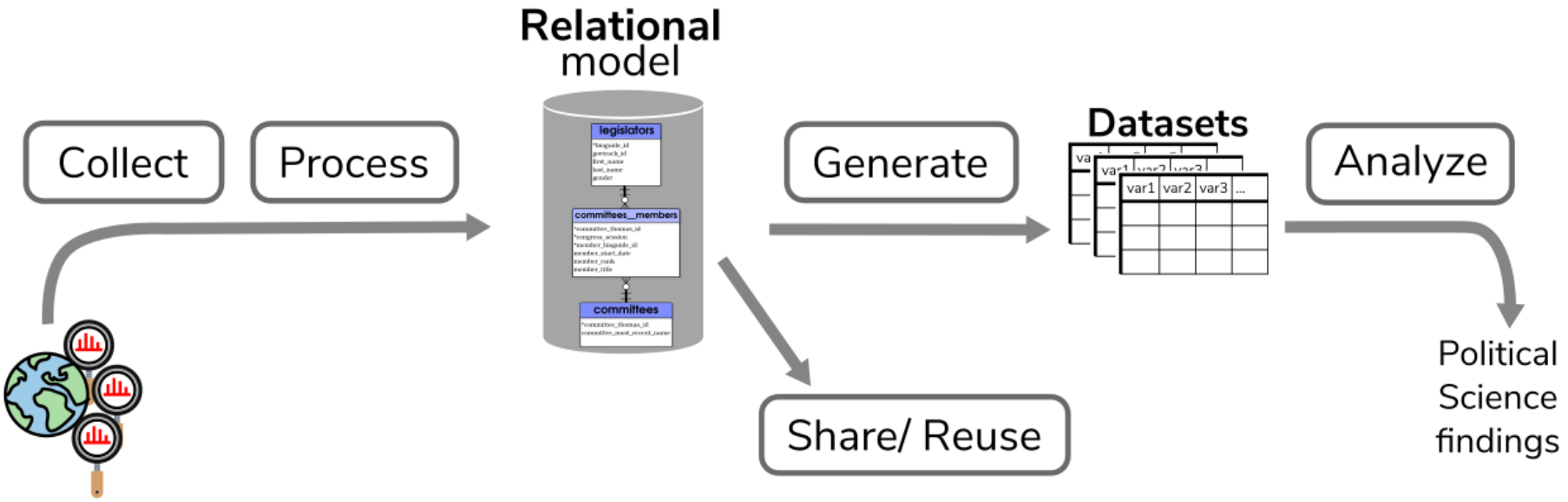




RELATIONAL MODEL: HOW TO DO IT?

(Realistically)

Relational model: how to?



Relational model: how to

1. Model your data : (SQL / Python)
    1. Entities
        1. types
        2. constraints
        3. what is unique?
    2. Relations

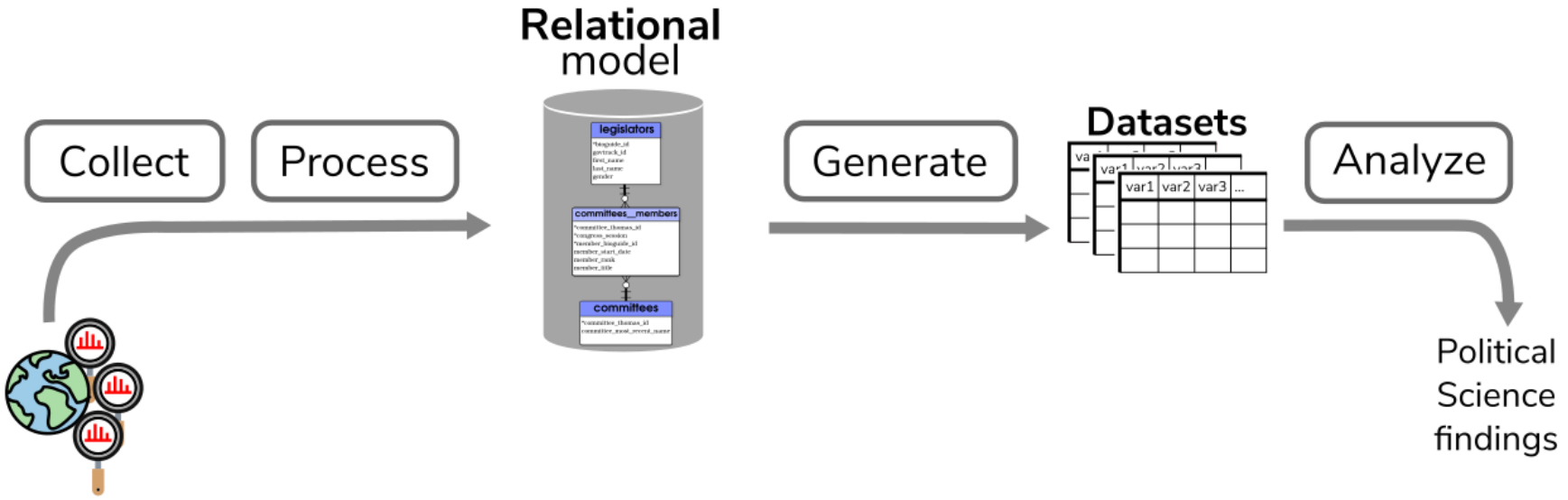
Relational model: how to

2. Fill the model : (R / Python / SQL)
    1. Clean :
    2. Insert into the model : R / Python / SQL
3. Generate Dataframes: R / SQL
    1. *Probably easiest part*

PRACTICAL ARCHITECTURE

(Realistically)

Relational model: how to

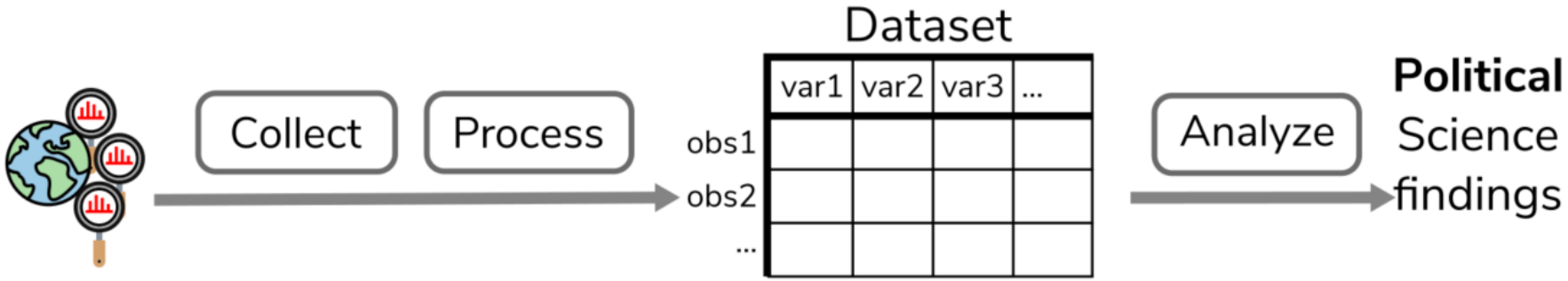


CONCLUSION

(Short)

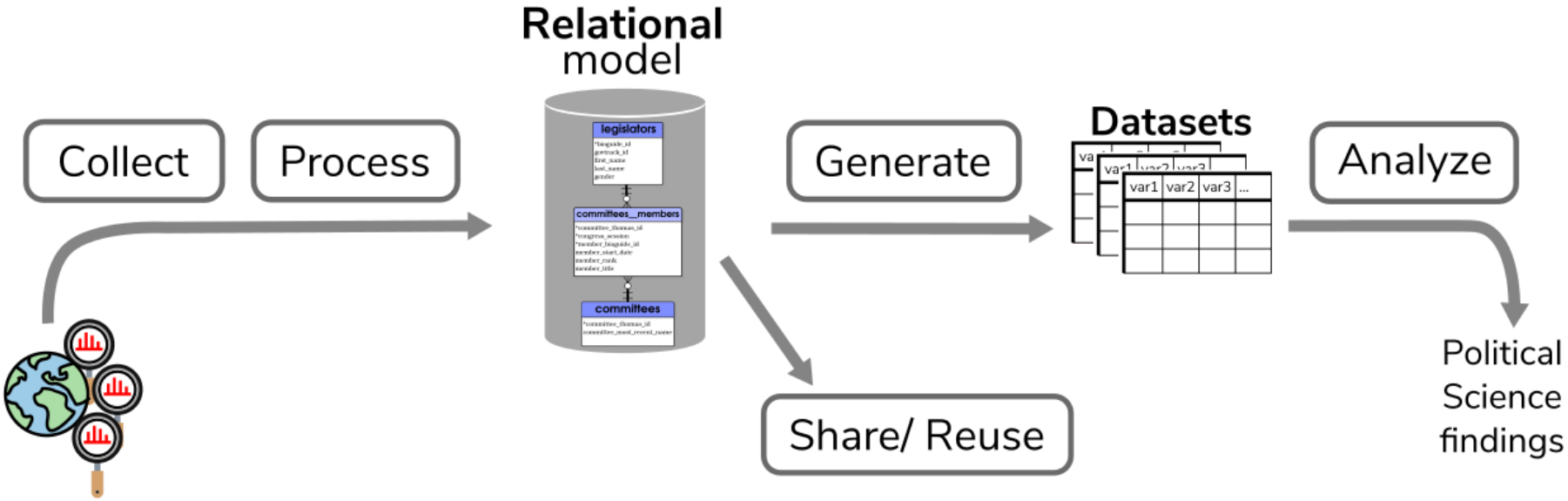
Conclusion

- **Typical workflow : problematic**
  (reproducibility, trackability, explainability)
  + (re-use / sharing, scaling, data quality)



Conclusion

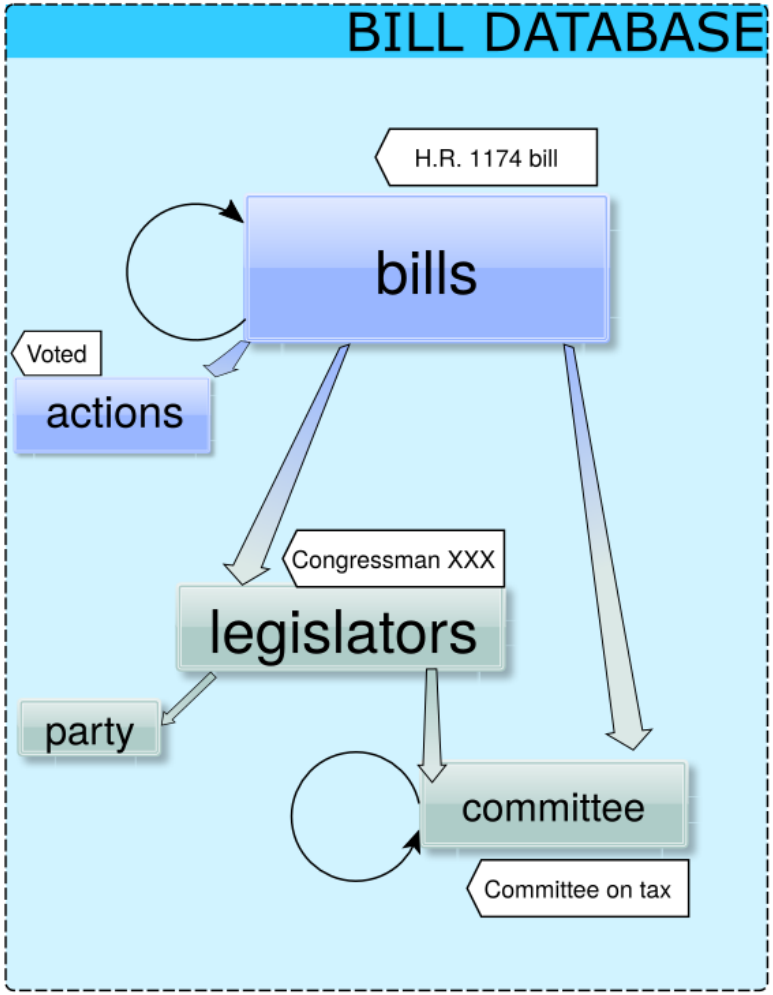
- Model your data to solve these issues



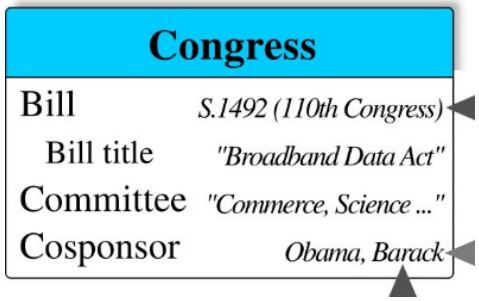
The Lobbyview DBs

Rémi Cura
Pf. In Song Kim

The Lobbyview DBs: entities

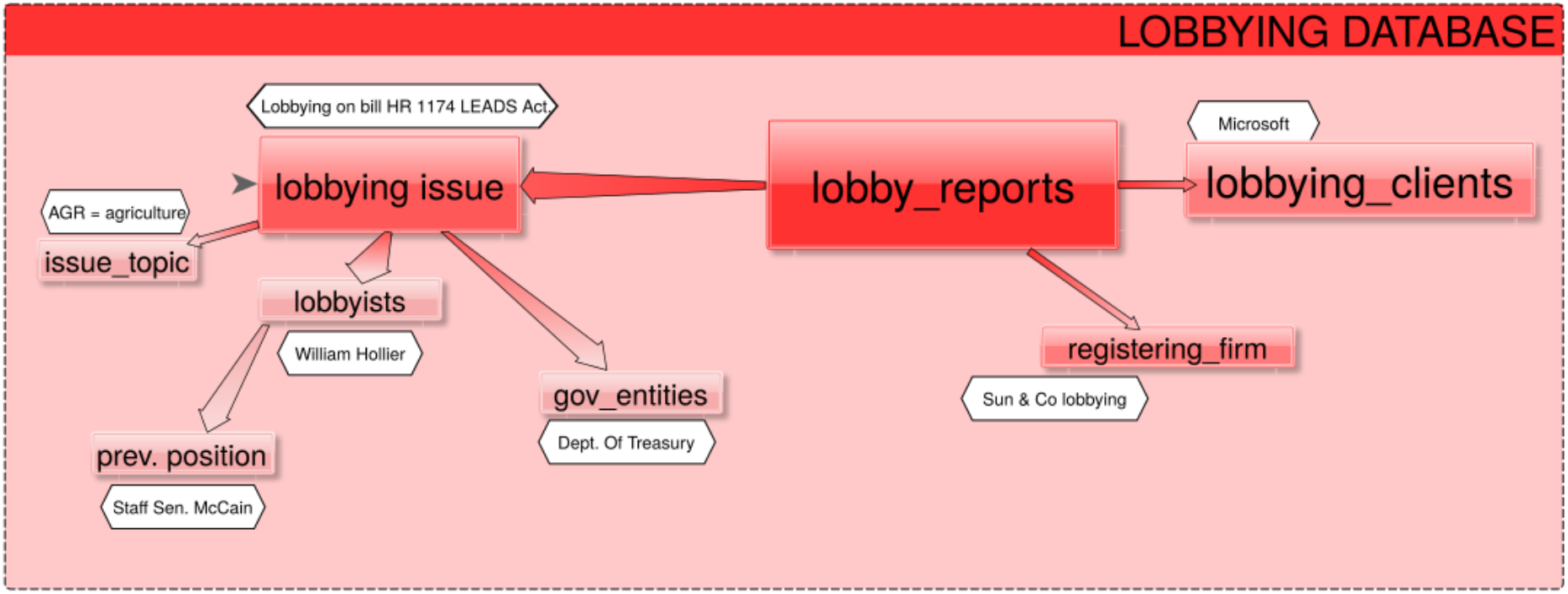


The Lobbyview DBs: entities

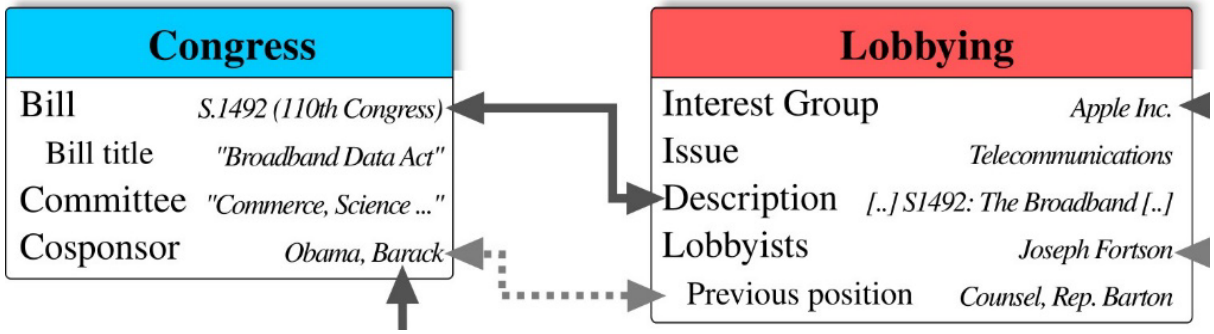


The Lobbyview DBs: entities

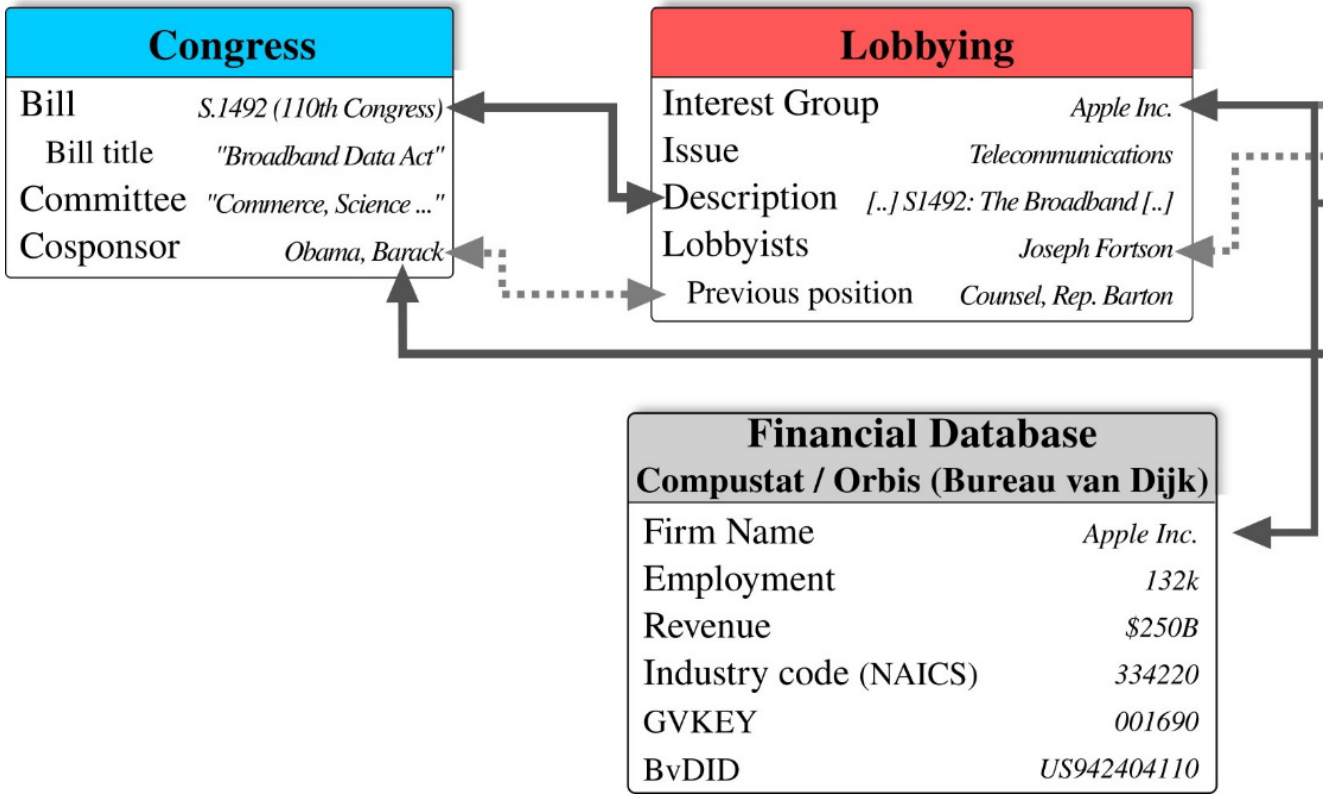
-



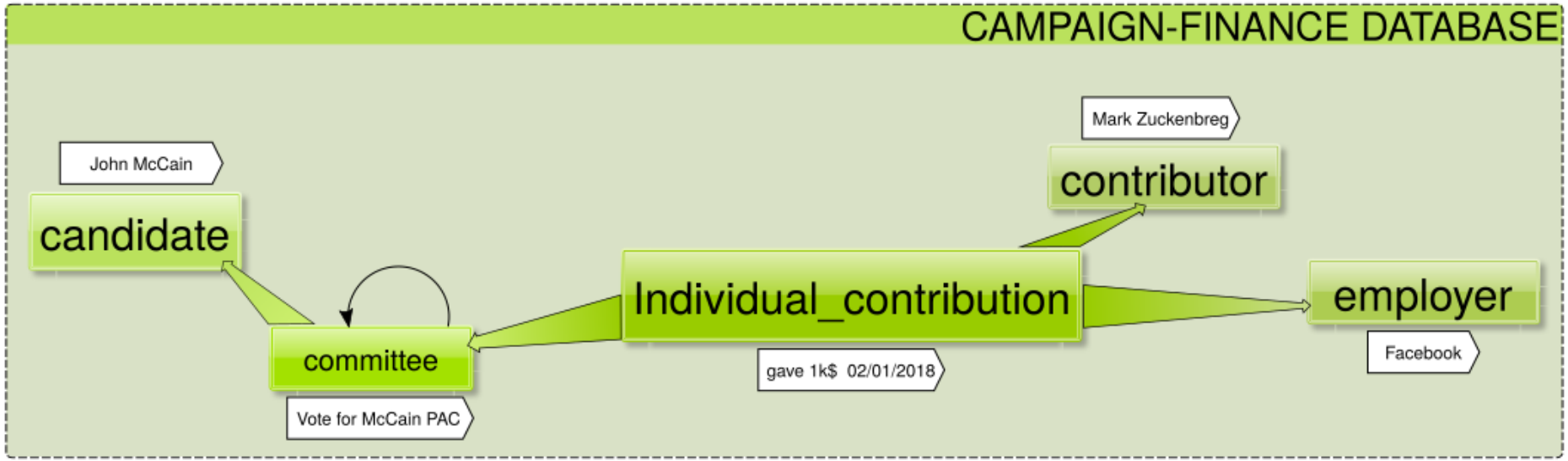
The Lobbyview DBs: entities



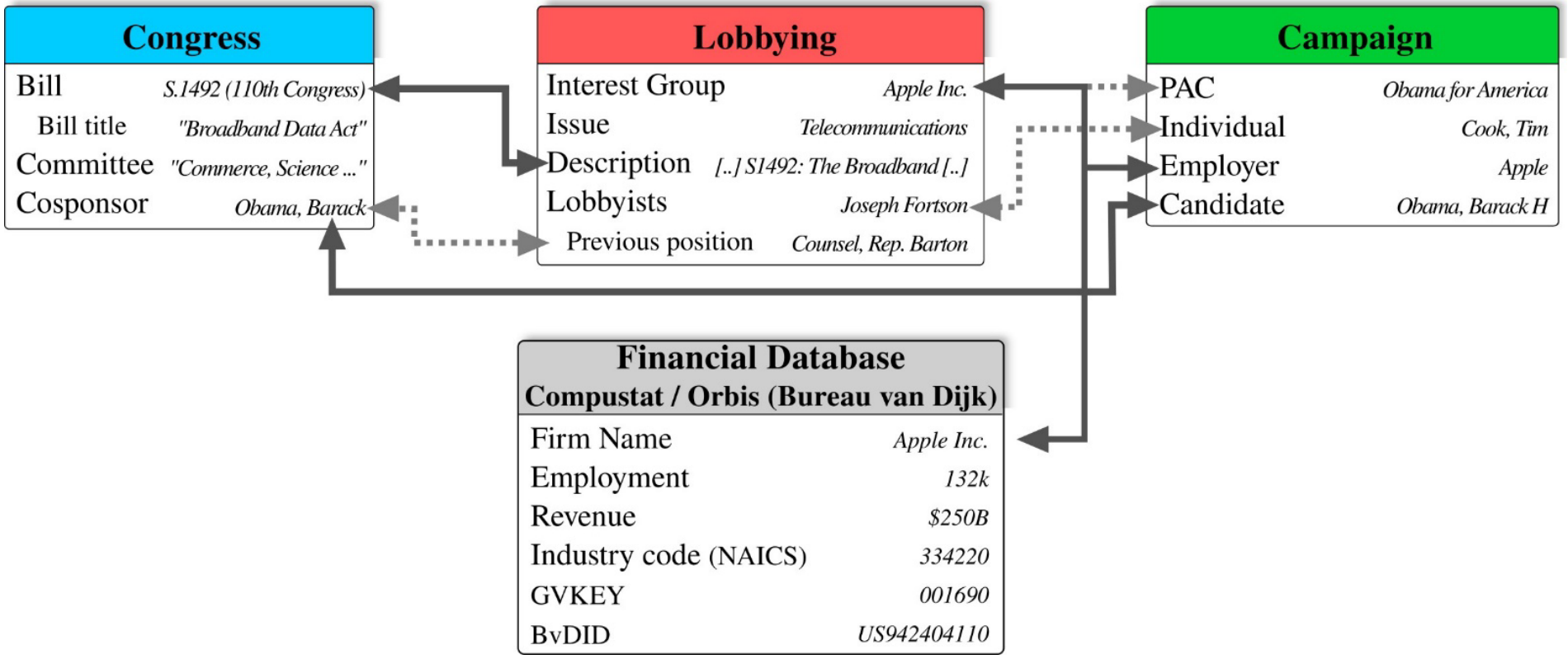
The Lobbyview DBs: entities



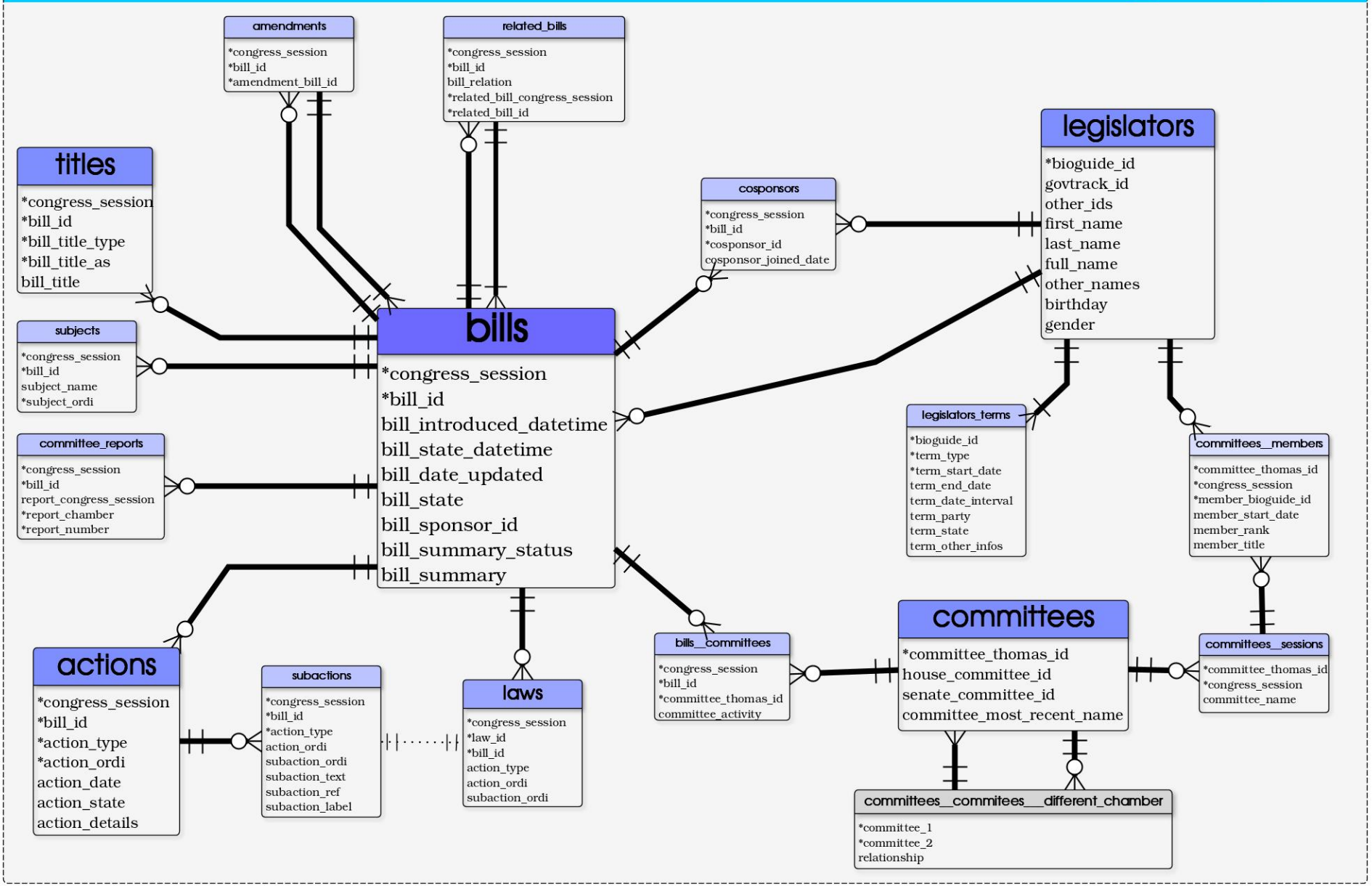
The Lobbyview DBs: entities

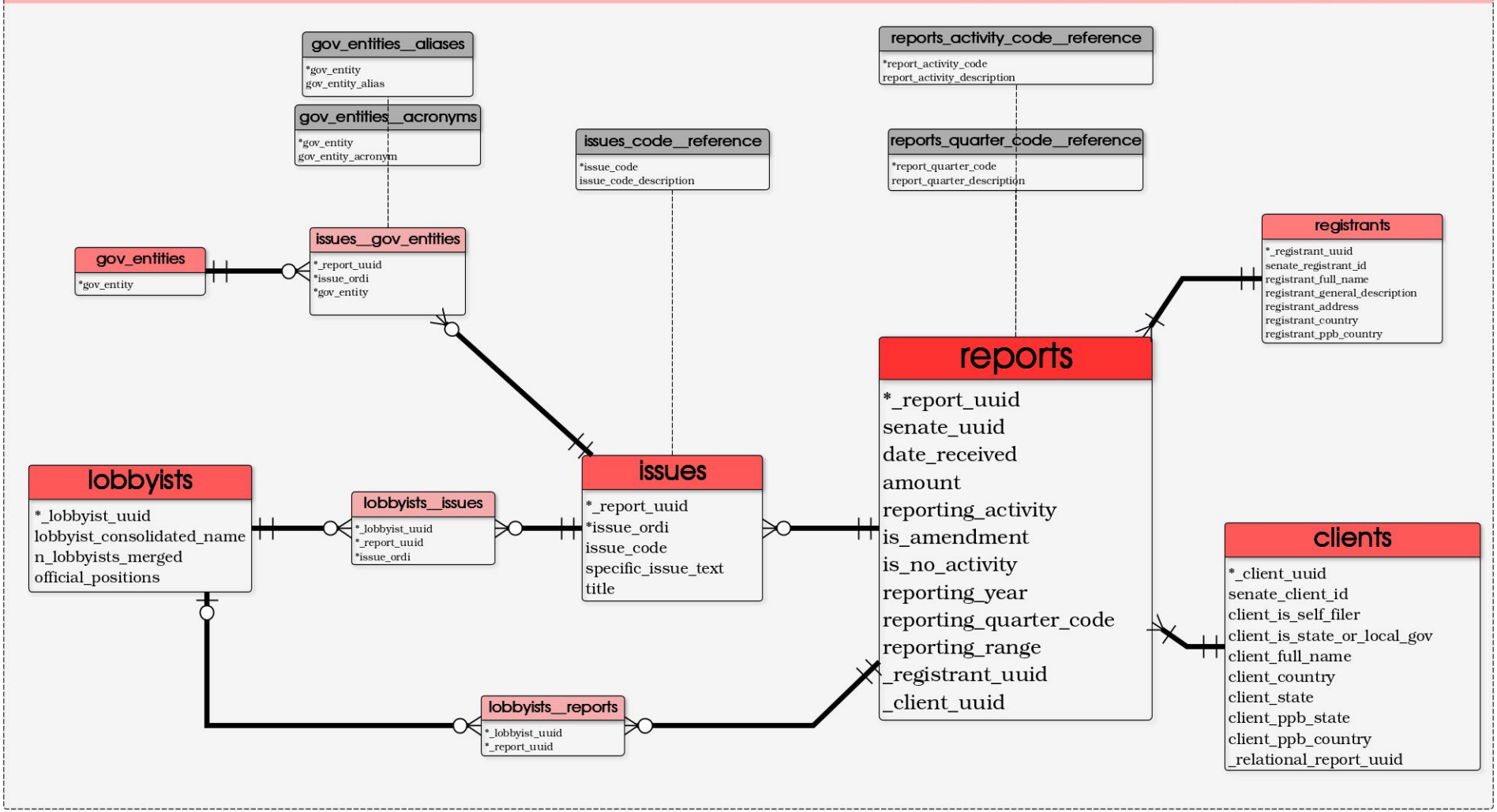


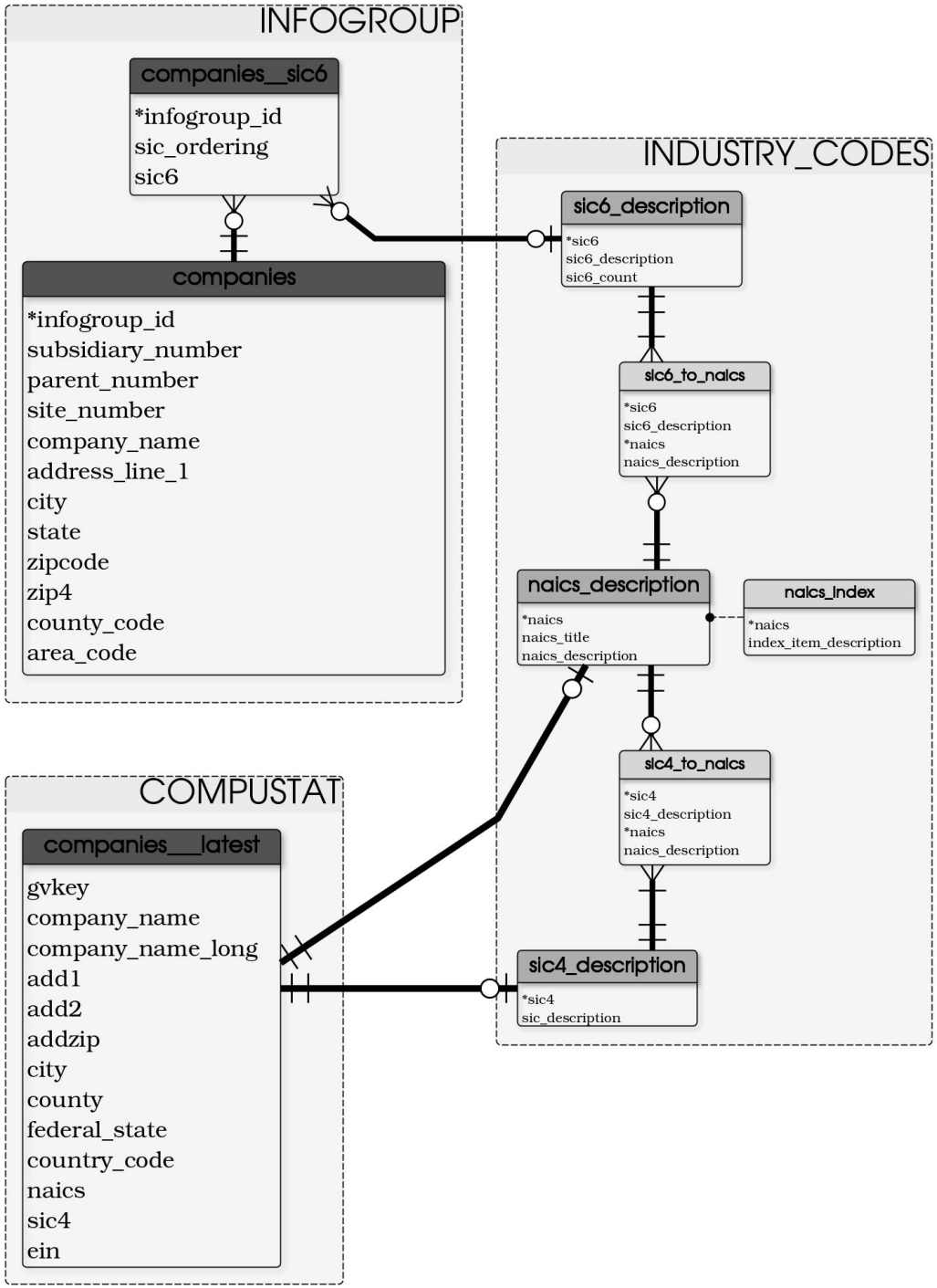
The Lobbyview DBs: entities



CONSOLIDATED_LAYER_BILLS




gov_entities__aliases
*gov_entity
gov_entity_alias
gov_entities__acronyms
*gov_entity
gov_entity_acronym
issues_code__reference
*issue_code
issue_code_description
reports_activity_code__reference
*report_activity_code
report_activity_description
reports_quarter_code__reference
*report_quarter_code
report_quarter_description
registrants
*_registrant_uuid
senate_registrant_id
registrant_full_name
registrant_general_description
registrant_address
registrant_country
registrant_ppb_country
gov_entities
*gov_entity
issues__gov_entities
*_report_uuid
*issue_ordi
*gov_entity
reports
*_report_uuid
senate_uuid
date_received
amount
reporting_activity
is_amendment
is_no_activity
reporting_year
reporting_quarter_code
reporting_range
_registrant_uuid
_client_uuid
lobbyists
*_lobbyist_uuid
lobbyist_consolidated_name
n_lobbyists_merged
official_positions
lobbyists__issues
*_lobbyist_uuid
*_report_uuid
*issue_ordi
issues
*_report_uuid
*issue_ordi
issue_code
specific_issue_text
title
clients
*_client_uuid
senate_client_id
client_is_self_filer
client_is_state_or_local_gov
client_full_name
client_country
client_state
client_ppb_state
client_ppb_country
_relational_report_uuid
lobbyists__reports
*_lobbyist_uuid
*_report_uuid


INFOGROUP
companies__sic6
*infogroup_id
sic_ordering
sic6
companies
*infogroup_id
subsidiary_number
parent_number
site_number
company_name
address_line_1
city
state
zipcode
zip4
county_code
area_code
INDUSTRY_CODES
sic6_description
*sic6
sic6_description
sic6_count
sic6_to_naics
*sic6
sic6_description
*naics
naics_description
naics_description
*naics
naics_title
naics_description
naics_index
*naics
index_item_description
sic4_to_naics
*sic4
sic4_description
*naics
naics_description
sic4_description
*sic4
sic_description
COMPUSTAT
companies___latest
gvkey
company_name
company_name_long
add1
add2
addzip
city
county
federal_state
country_code
naics
sic4
ein