Building an end-to-end open source data platform for clinical data

Alberto Labarga

PyBCN 2022 - Barcelona

Telecommunications Engineer
Working as Head of Data @IOMED
More than 20 years teaching

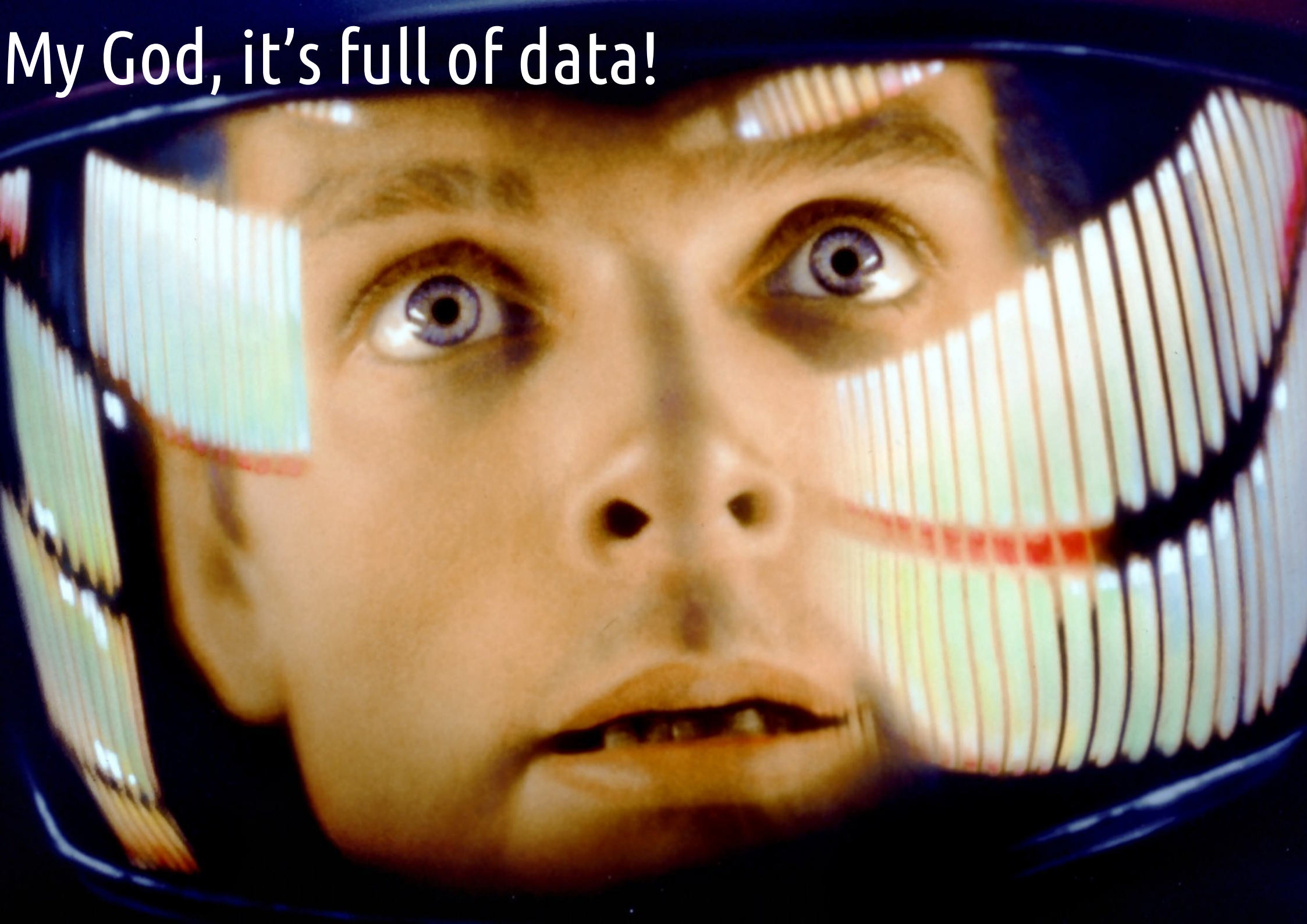
open data - open source – open science

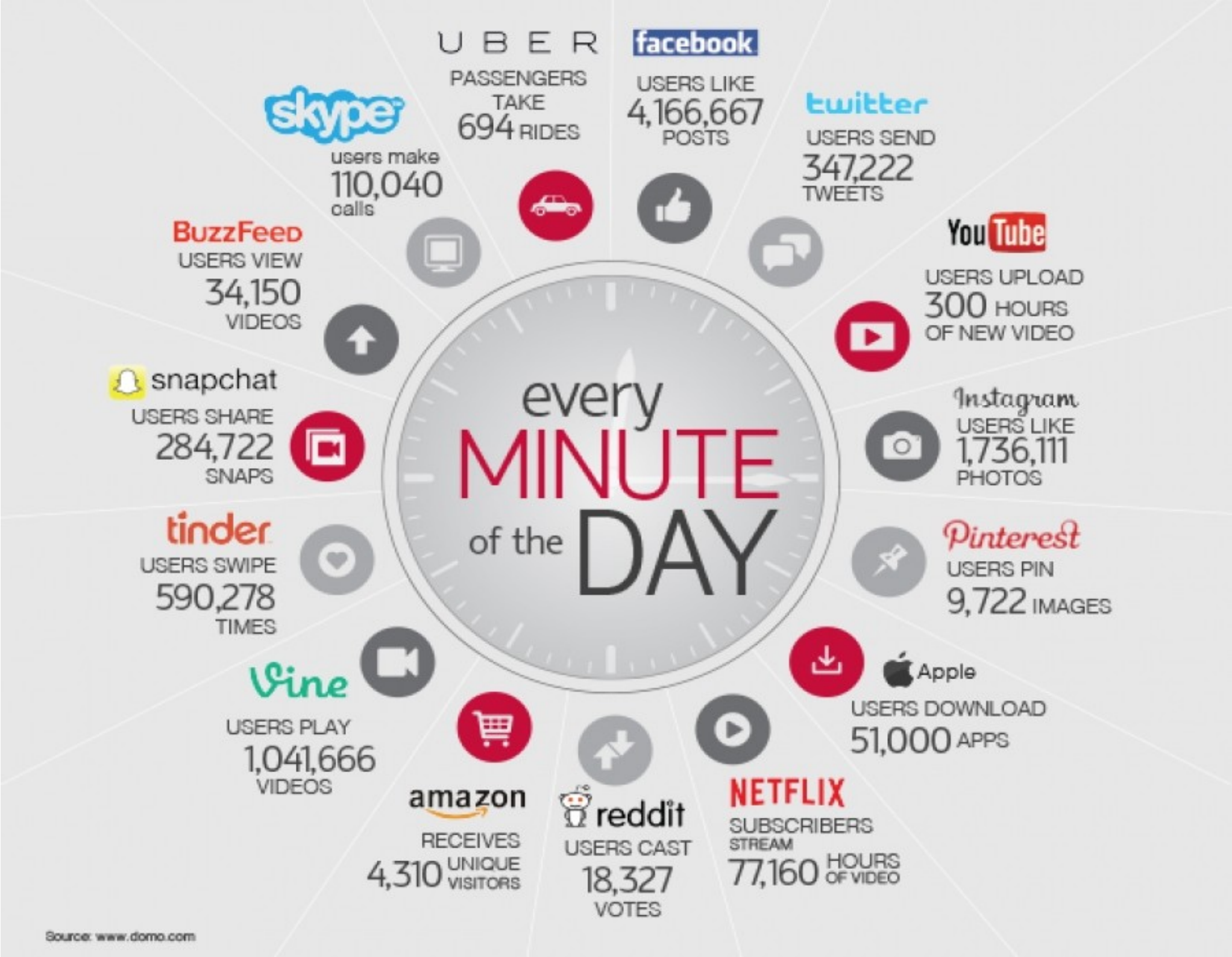
I am Alberto

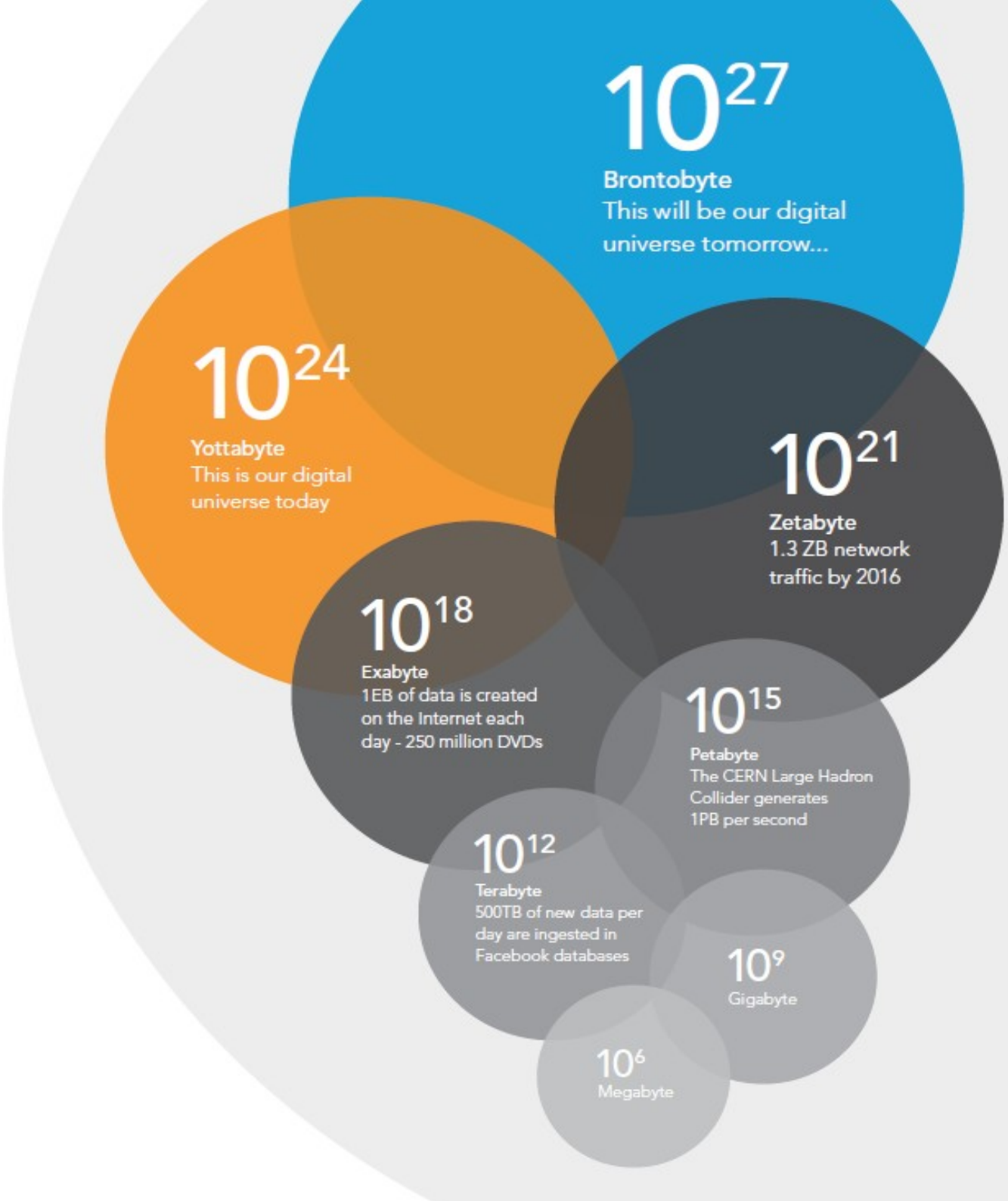
alabarga
alabarga
/in/albertolabarga

IOMED

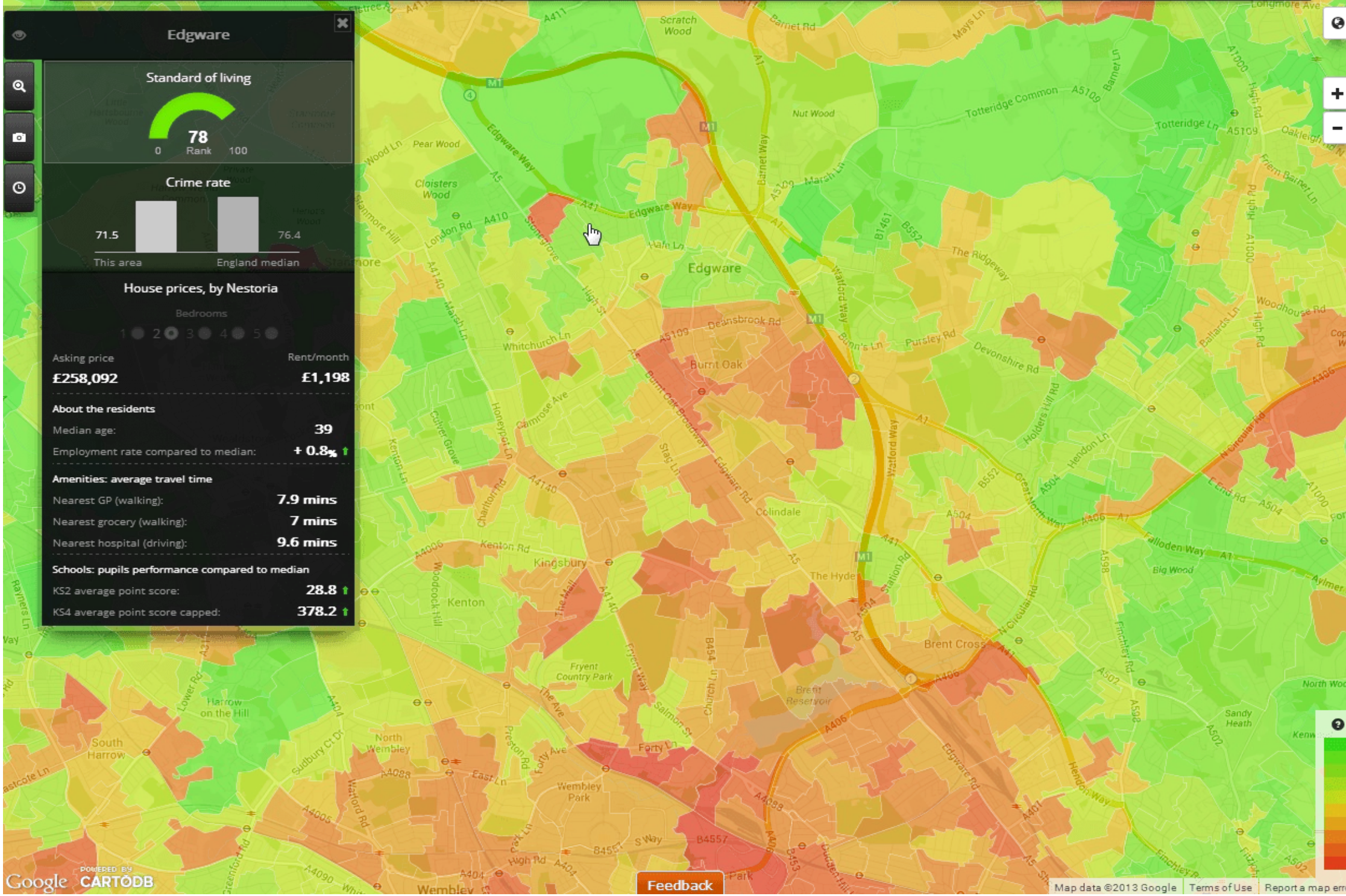
Data is big


My God, it’s full of data!

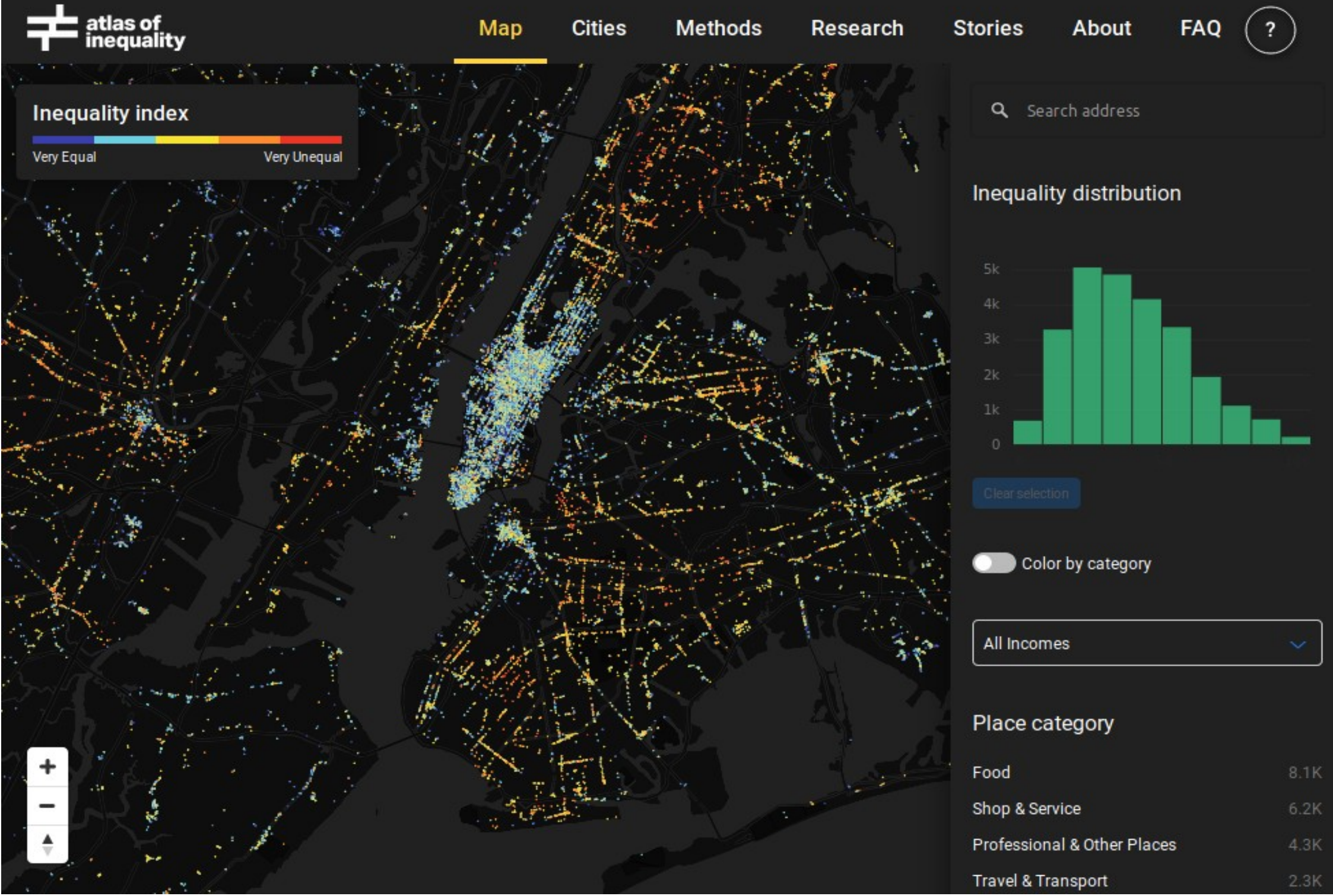

every MINUTE of the DAY
UBER
PASSENGERS TAKE 694 RIDES
facebook
USERS LIKE 4,166,667 POSTS
twitter
USERS SEND 347,222 TWEETS
YouTube
USERS UPLOAD 300 HOURS OF NEW VIDEO
Instagram
USERS LIKE 1,736,111 PHOTOS
Pinterest
USERS PIN 9,722 IMAGES
Apple
USERS DOWNLOAD 51,000 APPS
NETFLIX
SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO
reddit
USERS CAST 18,327 VOTES
amazon
RECEIVES 4,310 UNIQUE VISITORS
Vine
USERS PLAY 1,041,666 VIDEOS
tinder
USERS SWIPE 590,278 TIMES
snapchat
USERS SHARE 284,722 SNAPS
BuzzFeed
USERS VIEW 34,150 VIDEOS
skype
users make 110,040 calls
Source: www.domo.com


10^{27}
Brontobyte
This will be our digital universe tomorrow...
10^{24}
Yottabyte
This is our digital universe today
10^{21}
Zetabyte
1.3 ZB network traffic by 2016
10^{18}
Exabyte
1EB of data is created on the Internet each day - 250 million DVDs
10^{15}
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^{12}
Terabyte
500TB of new data per day are ingested in Facebook databases
10^{9}
Gigabyte
10^{6}
Megabyte

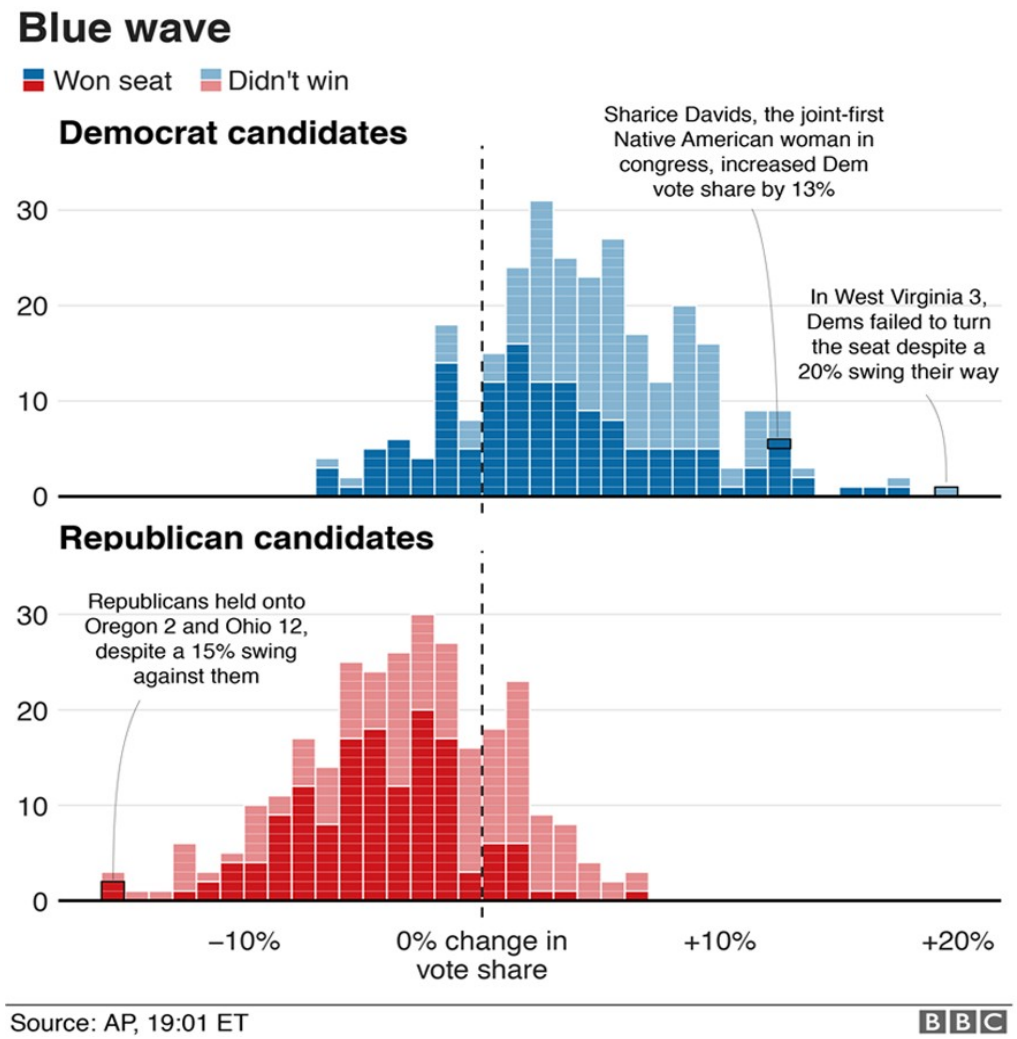
http://mwcimpact.com/

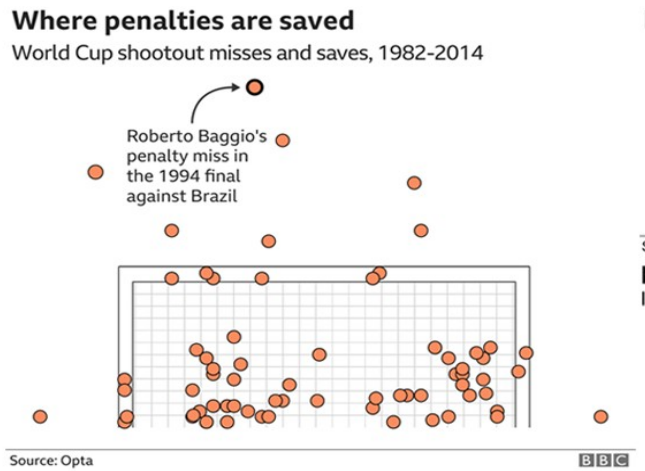


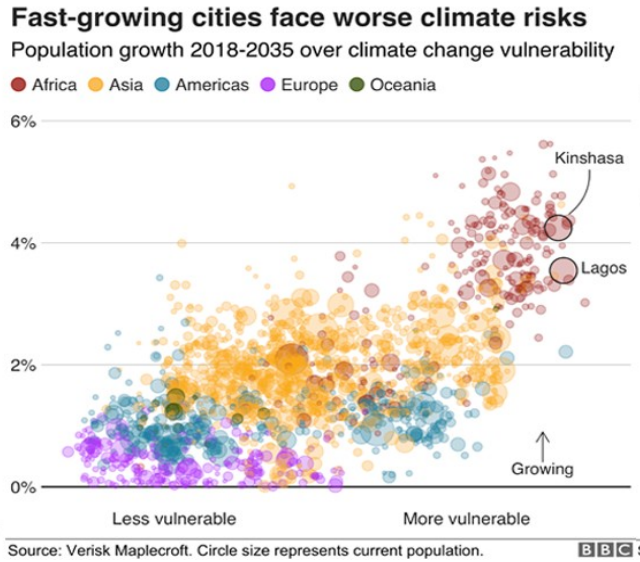
http://illustreets.co.uk/

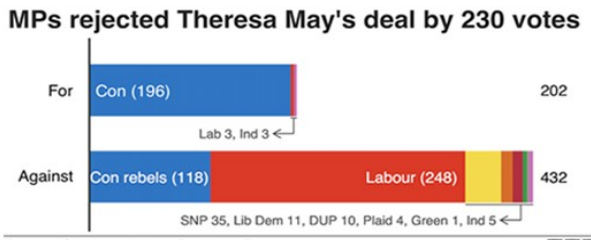


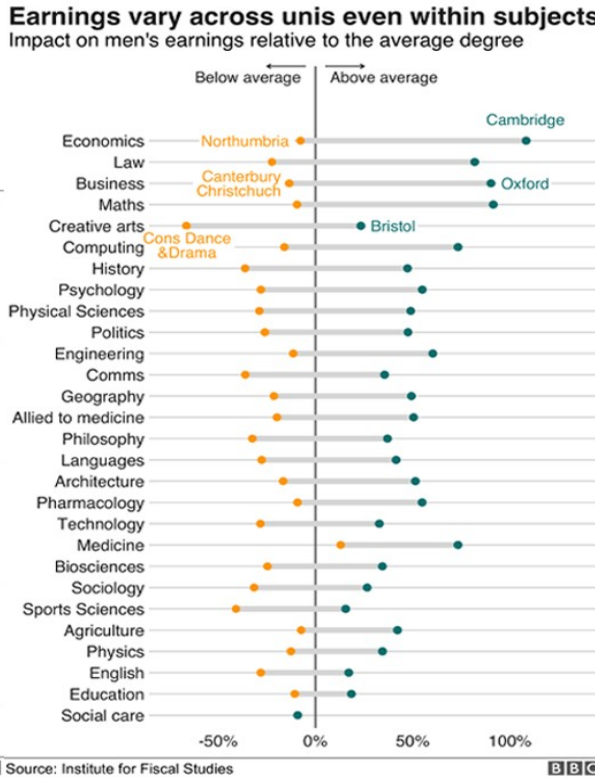
https://inequality.media.mit.edu/


Blue wave
Won seat
Didn't win
Democrat candidates
Sharice Davids, the joint-first Native American woman in congress, increased Dem vote share by 13%
In West Virginia 3, Dems failed to turn the seat despite a 20% swing their way
30
20
10
0
Republican candidates
Republicans held onto Oregon 2 and Ohio 12, despite a 15% swing against them
30
20
10
0
–10%
0% change in vote share
+10%
+20%
Source: AP, 19:01 ET
BBC


Where penalties are saved
World Cup shootout misses and saves, 1982-2014
Roberto Baggio's penalty miss in the 1994 final against Brazil
Source: Opta
BBC

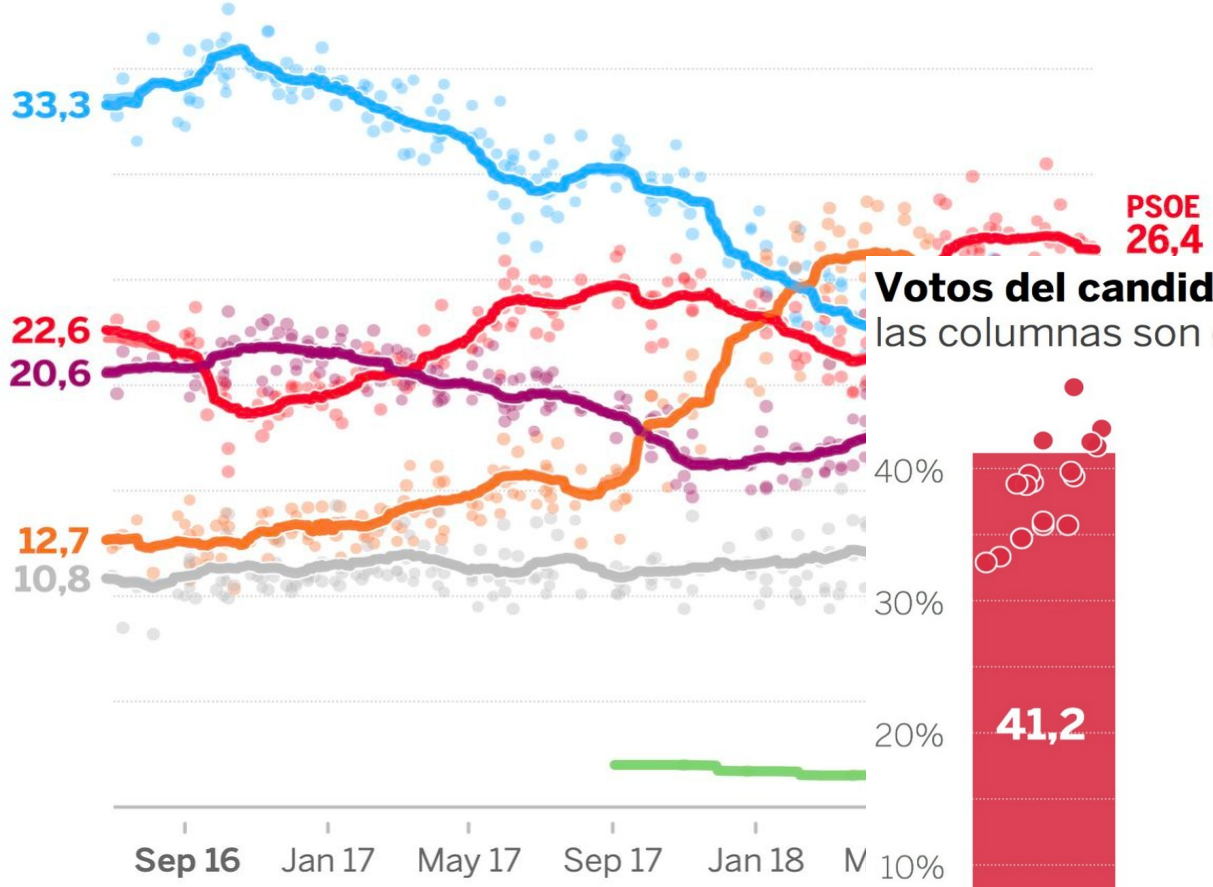

Fast-growing cities face worse climate risks
Population growth 2018-2035 over climate change vulnerability
Africa
Asia
Americas
Europe
Oceania
6%
4%
2%
0%
Kinshasa
Lagos
Growing
Less vulnerable
More vulnerable
Source: Verisk Maplecroft. Circle size represents current population.
BBC


MPs rejected Theresa May's deal by 230 votes
For
Con (196)
202
Lab 3, Ind 3
Against
Con rebels (118)
Labour (248)
432
SNP 35, Lib Dem 11, DUP 10, Plaid 4, Green 1, Ind 5
Source: Commons Votes Services. Excludes 'tellers', the Speaker and deputies
BBC


Earnings vary across unis even within subjects
Impact on men's earnings relative to the average degree
Below average
Above average
Economics
Law
Business
Maths
Creative arts
Computing
History
Psychology
Physical Sciences
Politics
Engineering
Comms
Geography
Allied to medicine
Philosophy
Languages
Architecture
Pharmacology
Technology
Medicine
Biosciences
Sociology
Sports Sciences
Agriculture
Physics
English
Education
Social care
Cambridge
Northumbria
Canterbury Christchuch
Oxford
Bristol
Cons Dance &Drama
-50%
0%
50%
100%
Source: Institute for Fiscal Studies
BBC

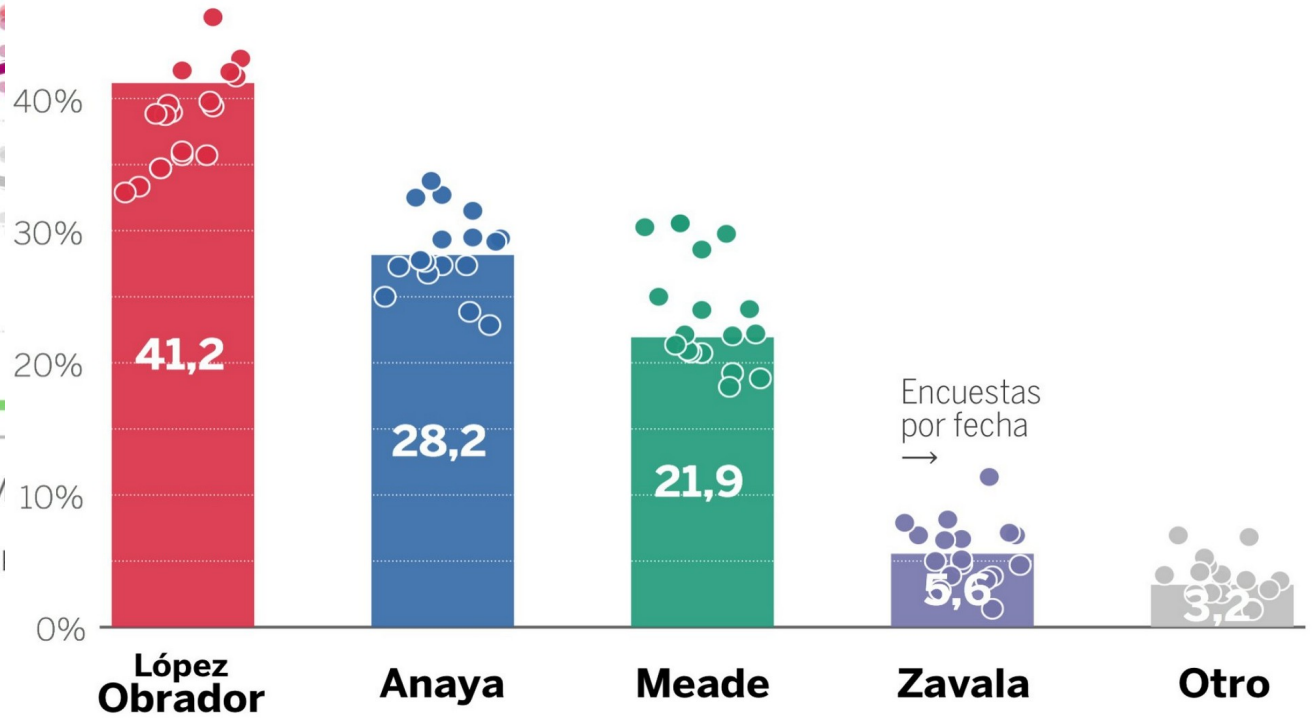
Así se mueven las encuestas

Estimación de voto; promedio ponderado



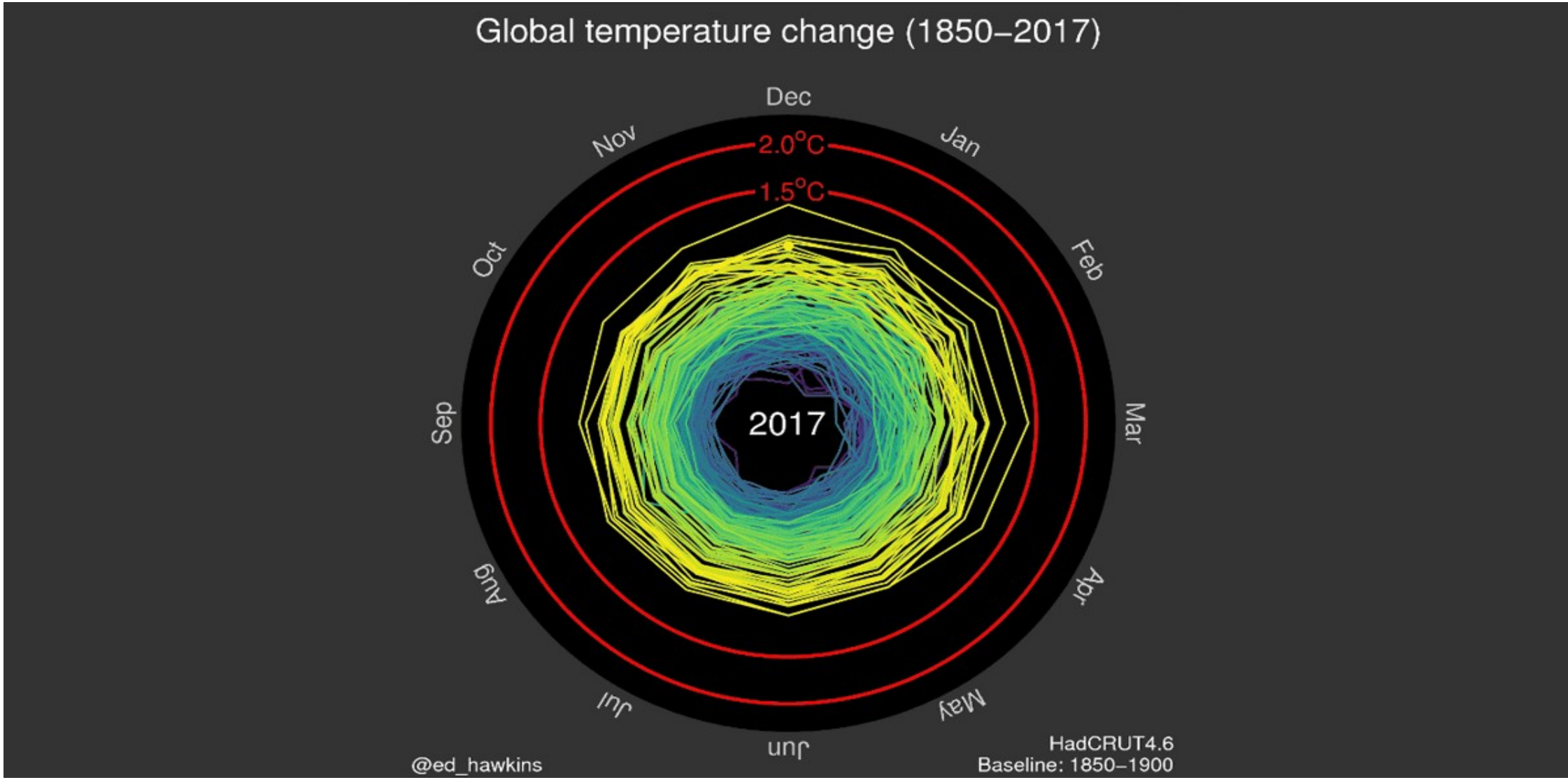
Fuente: Wikipedia y elaboración propia

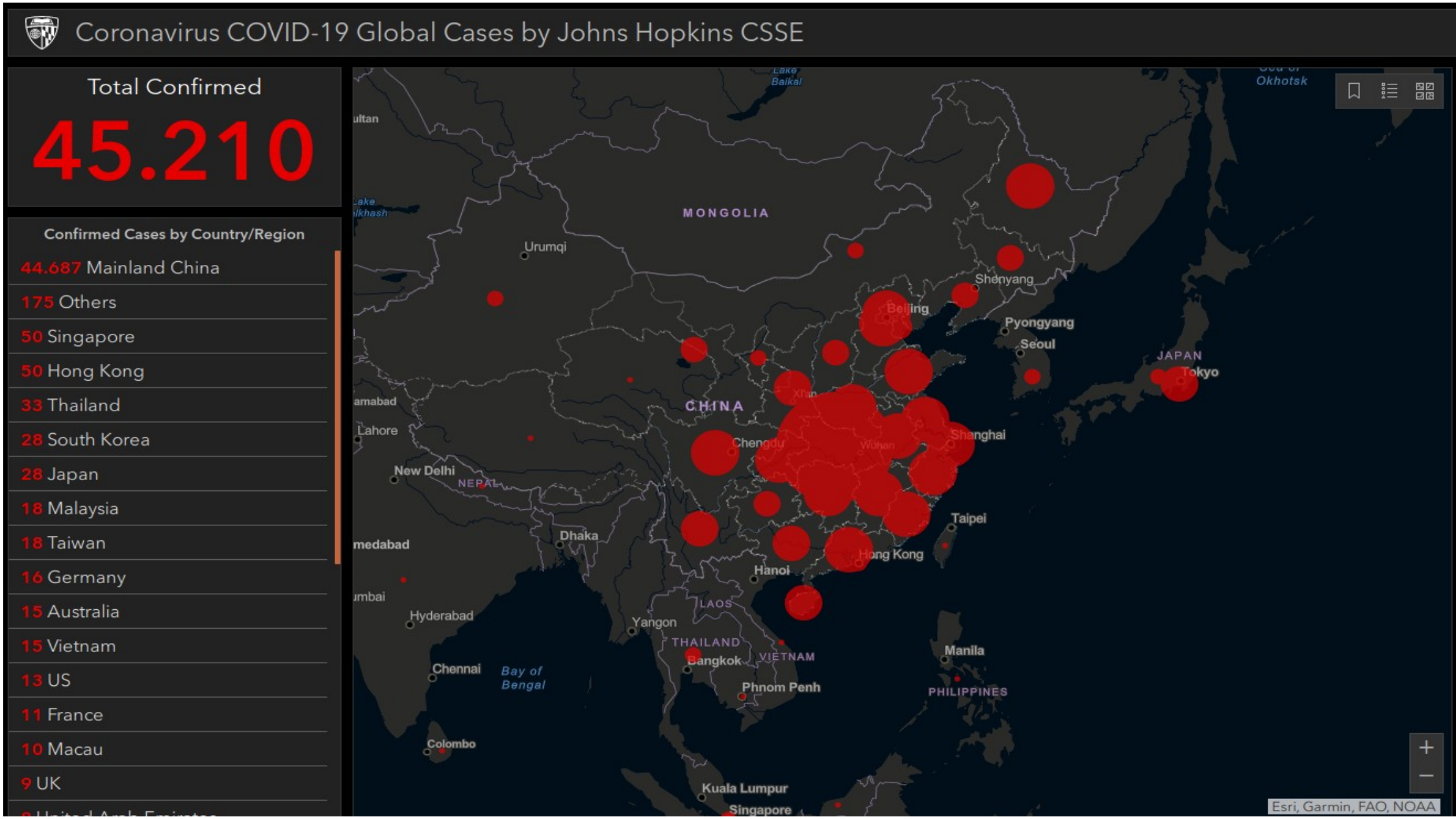
Votos del candidato en cada encuesta, las columnas son el promedio



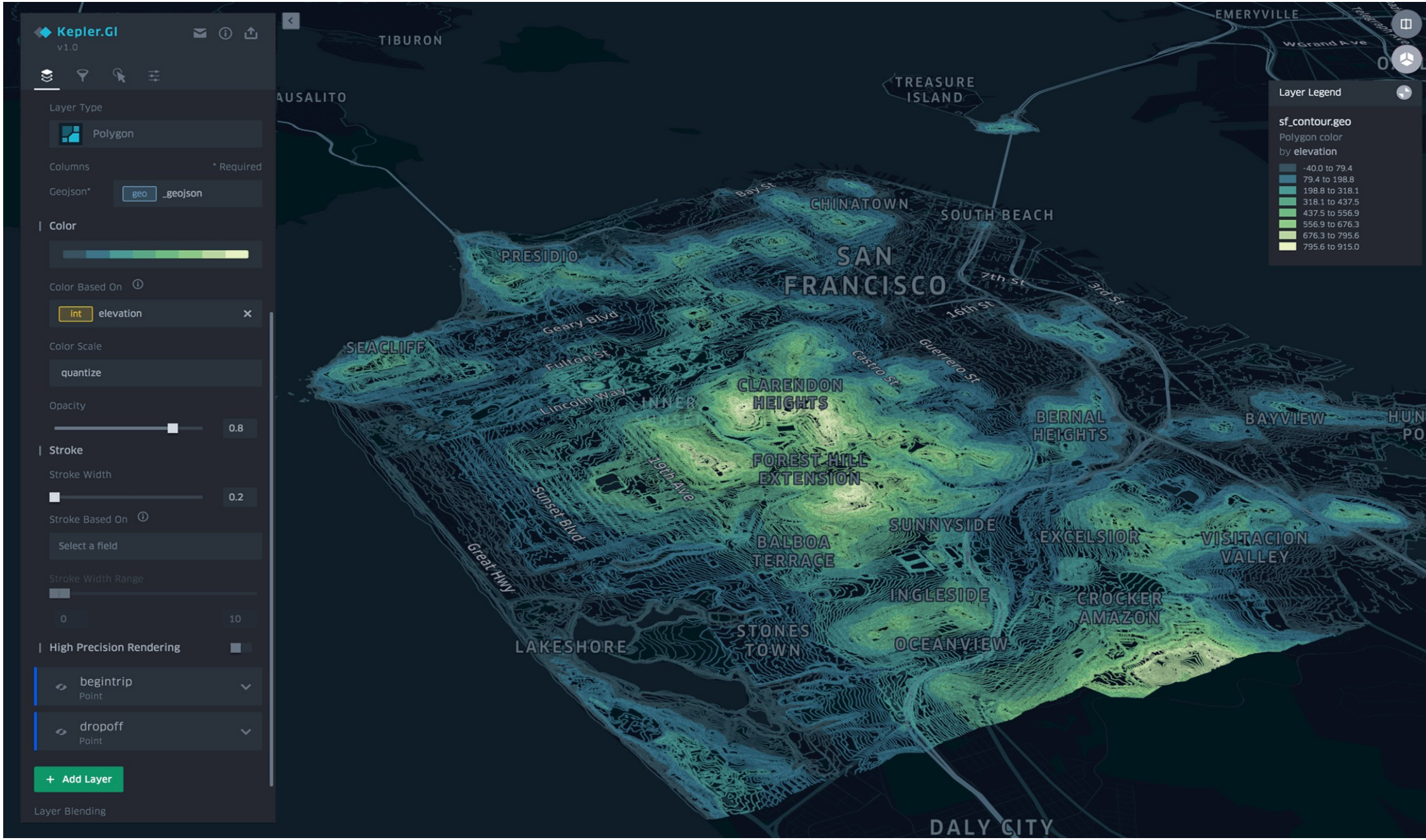
Fuente: Oraculus y elaboración propia

KIKO LLANERAS / EL PAÍS

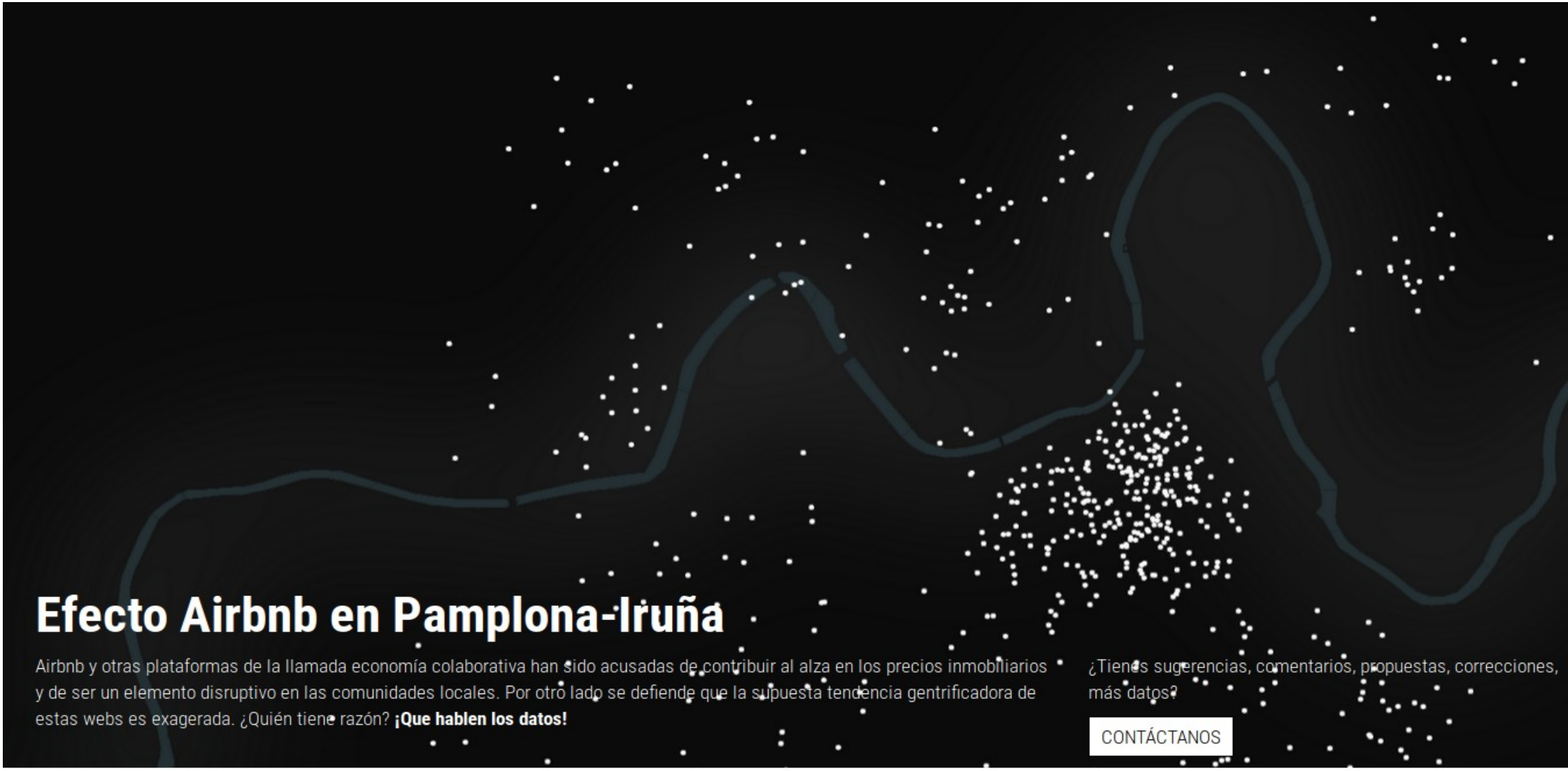

Global temperature change (1850–2017)
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
2.0°C
1.5°C
2017
@ed_hawkins
HadCRUT4.6
Baseline: 1850–1900



https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6



https://kepler.gl/



Efecto Airbnb en Pamplona-Iruña

Airbnb y otras plataformas de la llamada economía colaborativa han sido acusadas de contribuir al alza en los precios inmobiliarios y de ser un elemento disruptivo en las comunidades locales. Por otro lado se defiende que la supuesta tendencia gentrificadora de estas webs es exagerada. ¿Quién tiene razón? **¡Que hablen los datos!**

¿Tienes sugerencias, comentarios, propuestas, correcciones, más datos?

CONTÁCTANOS

782356
87901116
01001100
94560156
01000100100
000012
01296712
91236742323
01001100
8765432
23487923
5678234

A
B

CLOSED

OPEN

https://www.youtube.com/watch?v=OM6XllCm_qo

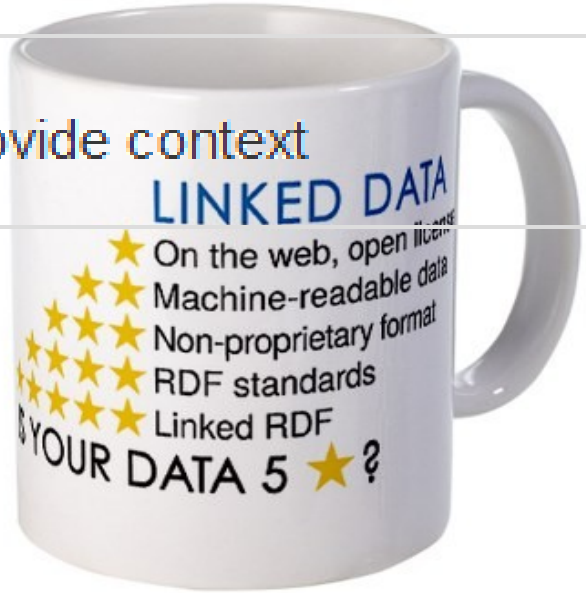
★ make your stuff available on the web (whatever format)

★★ make it available as structured data (e.g. excel instead of image scan of a table)

★★★ non-proprietary format (e.g. csv instead of excel)

★★★★ use URLs to identify things, so that people can point at your stuff

★★★★★ link your data to other people's data to provide context



Open data is about MORE THAN DISCLOSURE it must be "Fair"

- Findable
- Accessible
- Interoperable
- Reusable

What is FAIR DATA?

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Data is sexy

Feature

Is data scientist sexiest job of the century?

Data Scientist:
The Sexiest Job of the 21st Century

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT 4 | TEXT SIZE | PRINT

When Jonathan Goldman arrived for work in the business networking site, the place still company had just under 8 million accounts growing quickly as existing members invited their friends a

SPOTLIGHT ON BIG DATA

Spotlight

ARTWORK Tamar Cohen, *Andrew J Buboltz*
2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

70 Harvard Business Review October 2012

Supreme
PMS
Regular
93
89
87

Data Scientist
VS
Data Engineer
VS
Data Analyst

Clinical data

"Structured" and unstructured data

Gaps in data

Data silos

Ambiguity in medical text

Privacy restrictions

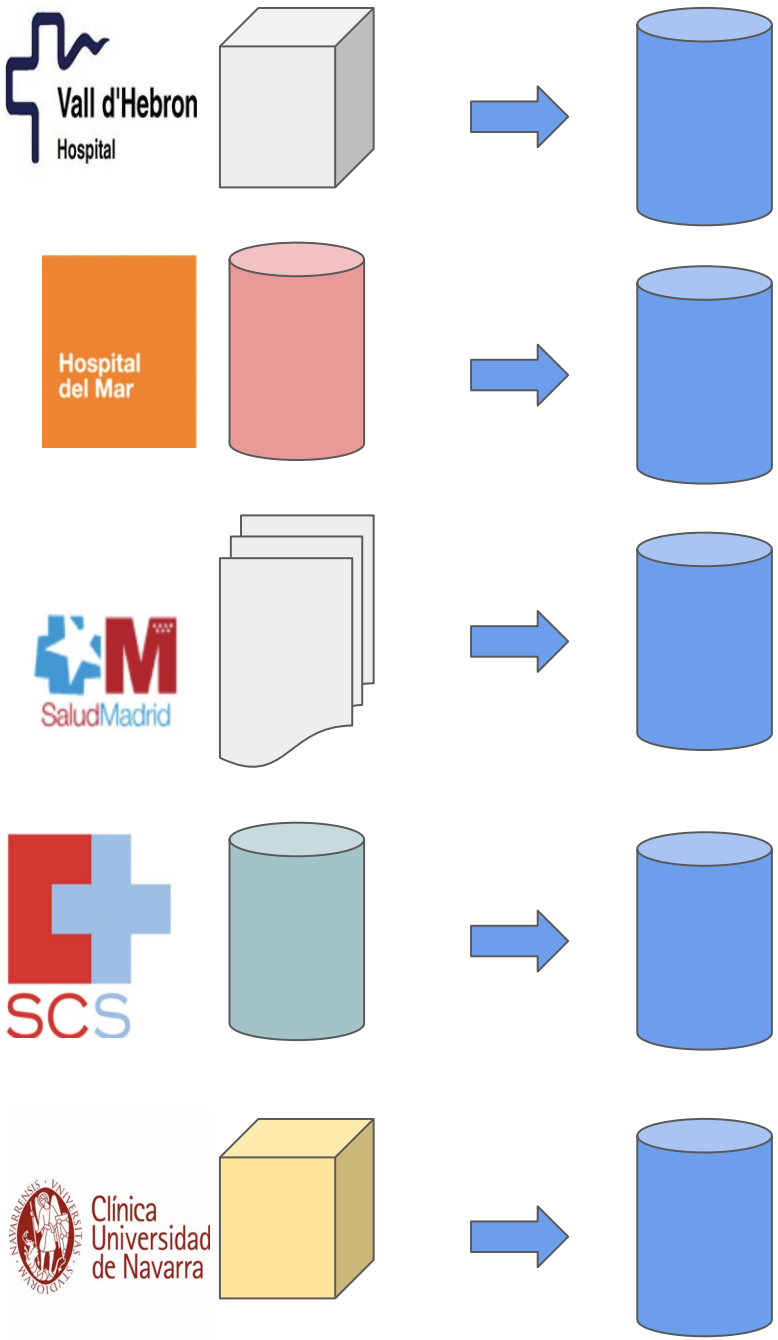
Vall d'Hebron
Hospital

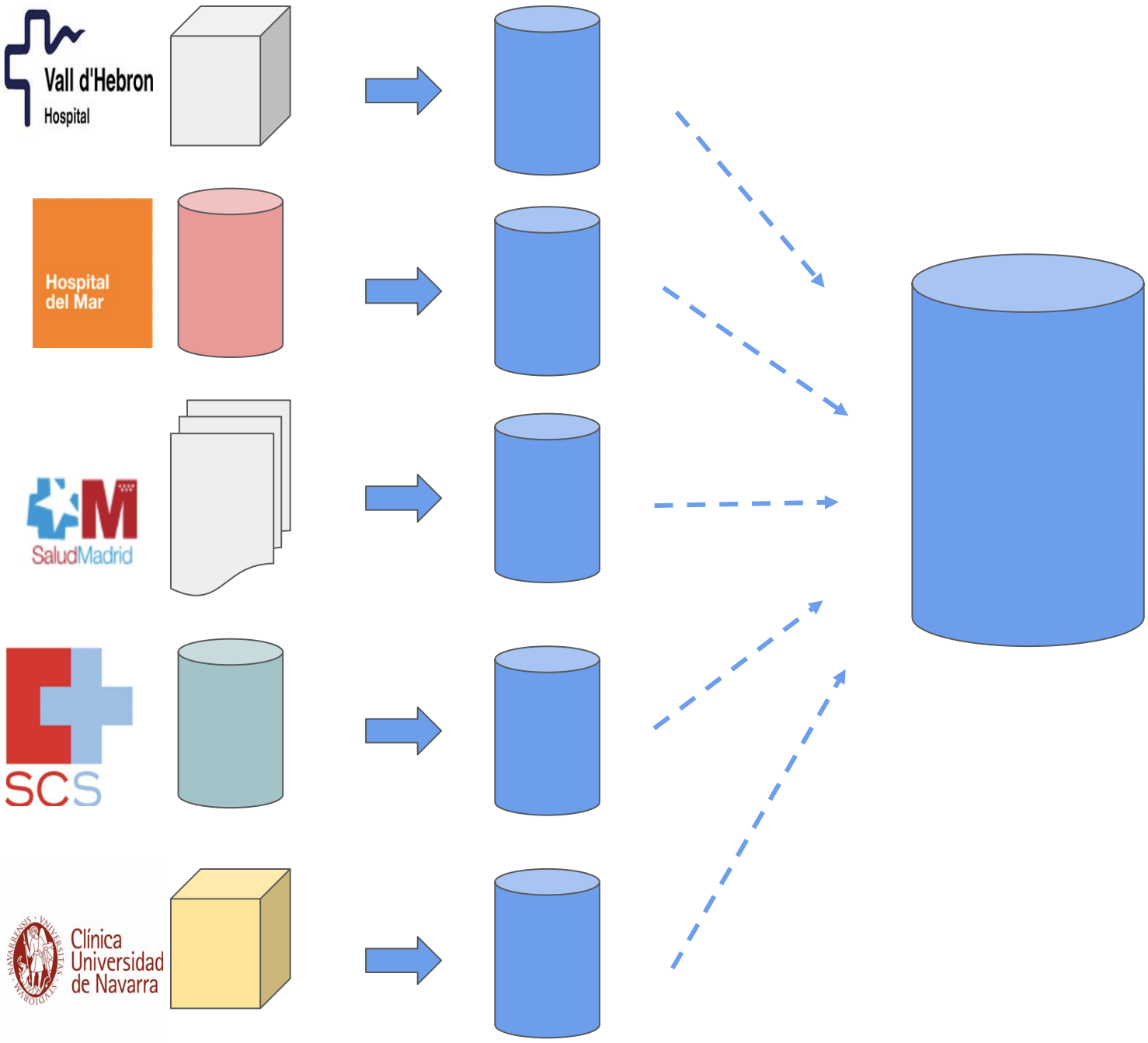
Hospital
del Mar

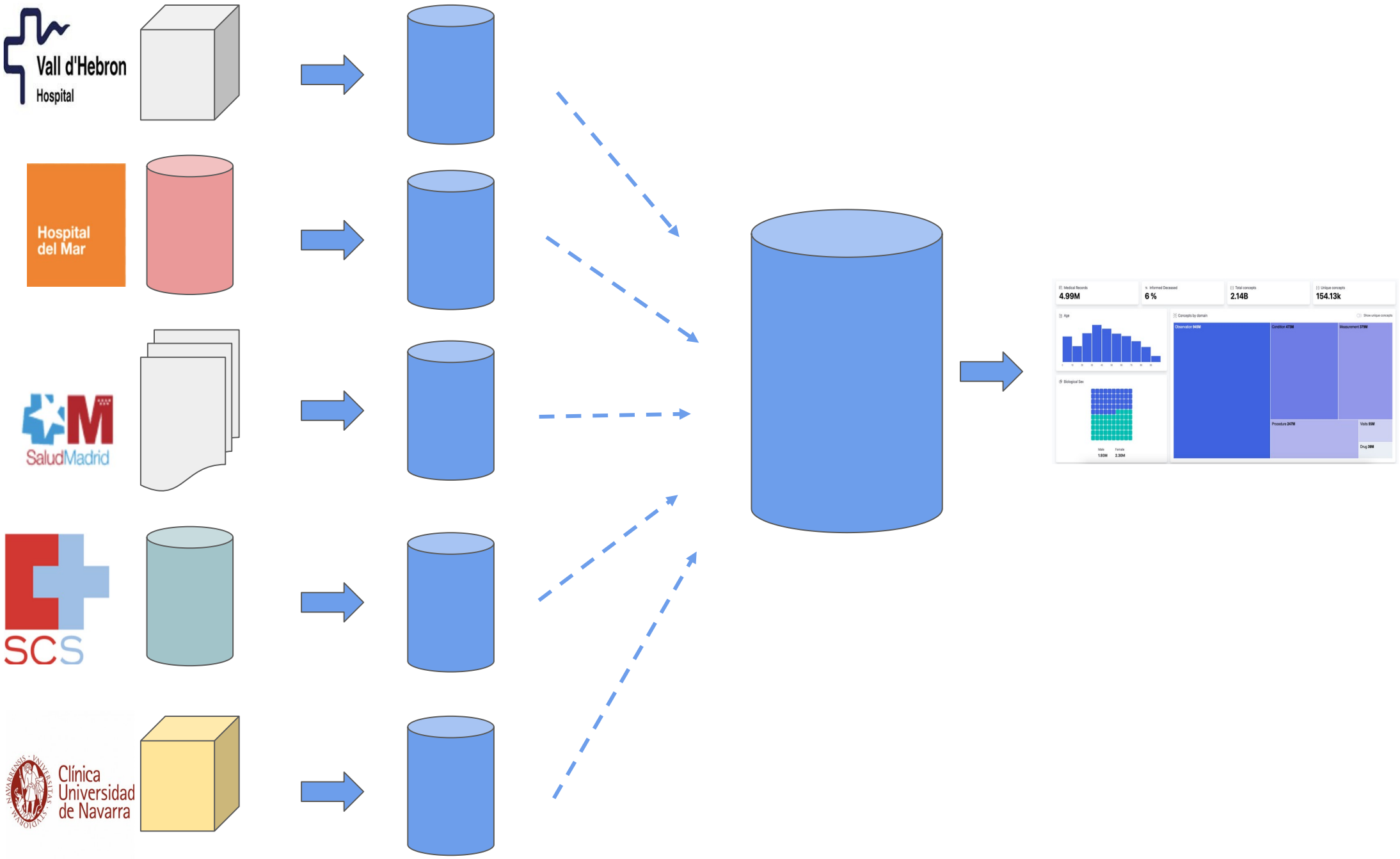
SaludMadrid

SCS

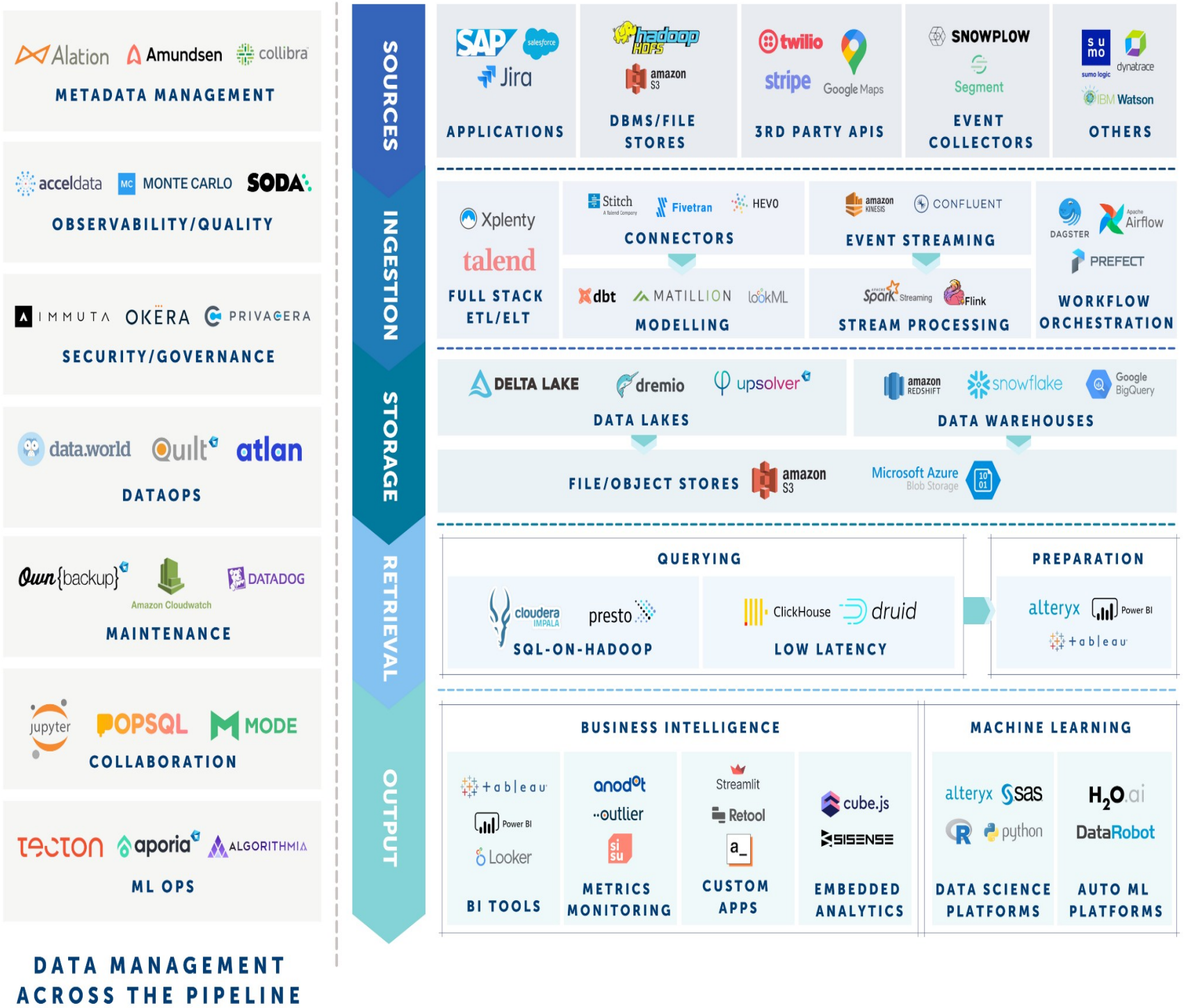
Clínica
Universidad
de Navarra


Vall d'Hebron
Hospital
Hospital
del Mar
SaludMadrid
SCS
Clínica
Universidad
de Navarra

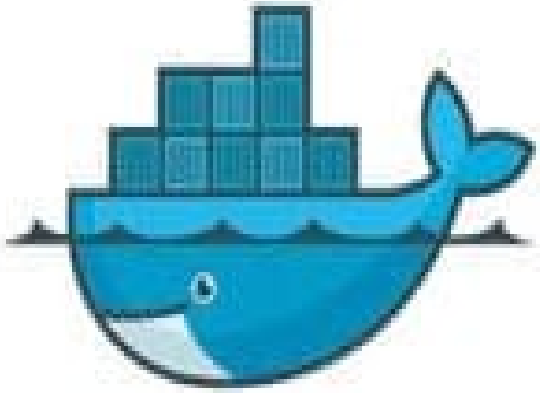

Vall d'Hebron
Hospital
Hospital
del Mar
SaludMadrid
SCS
Clínica
Universidad
de Navarra


Vall d'Hebron
Hospital
Hospital
del Mar
SaludMadrid
SCS
Clínica
Universidad
de Navarra
Medical Records
4.99M
Informed Deceased
6 %
Total concepts
2.14B
Unique concepts
154.13k
Age
Concepts by domain
Show unique concepts
Observation 945M
Condition 473M
Measurement 379M
Procedure 247M
Visits 55M
Drug 39M
Biological Sex
Male
1.93M
Female
2.30M

Modern data stack


Alation
Amundsen
collibra
METADATA MANAGEMENT
acceldata
MONTE CARLO
SODA
OBSERVABILITY/QUALITY
IMMUTA
OKERA
PRIVAGERA
SECURITY/GOVERNANCE
data.world
Quilt
atlan
DATAOPS
Own{backup}
Amazon Cloudwatch
DATADOG
MAINTENANCE
jupyter
POPSQL
MODE
COLLABORATION
tecton
aporia
ALGORITHMIA
ML OPS
DATA MANAGEMENT ACROSS THE PIPELINE
SOURCES
INGESTION
STORAGE
RETRIEVAL
OUTPUT
SAP
salesforce
Jira
APPLICATIONS
hadoop HDFS
amazon S3
DBMS/FILE STORES
twilio
stripe
Google Maps
3RD PARTY APIS
SNOWPLOW
Segment
EVENT COLLECTORS
sumo logic
dynatrace
IBM Watson
OTHERS
Xplenty
talend
FULL STACK ETL/ELT
Stitch
Fivetran
HEVO
CONNECTORS
amazon KINESIS
CONFLUENT
EVENT STREAMING
DAGSTER
Apache Airflow
PREFECT
WORKFLOW ORCHESTRATION
dbt
MATILLION
lookML
MODELLING
Spark Streaming
Flink
STREAM PROCESSING
DELTA LAKE
dremio
upsolver
DATA LAKES
amazon REDSHIFT
snowflake
Google BigQuery
DATA WAREHOUSES
FILE/OBJECT STORES
amazon S3
Microsoft Azure Blob Storage
QUERYING
cloudera IMPALA
presto
SQL-ON-HADOOP
ClickHouse
druid
LOW LATENCY
PREPARATION
alteryx
Power BI
tableau
BUSINESS INTELLIGENCE
tableau
Power BI
Looker
BI TOOLS
anodot
outlier
METRICS MONITORING
Streamlit
Retool
CUSTOM APPS
cube.js
SISENSE
EMBEDDED ANALYTICS
MACHINE LEARNING
alteryx
sas
python
DATA SCIENCE PLATFORMS
H2O.ai
DataRobot
AUTO ML PLATFORMS

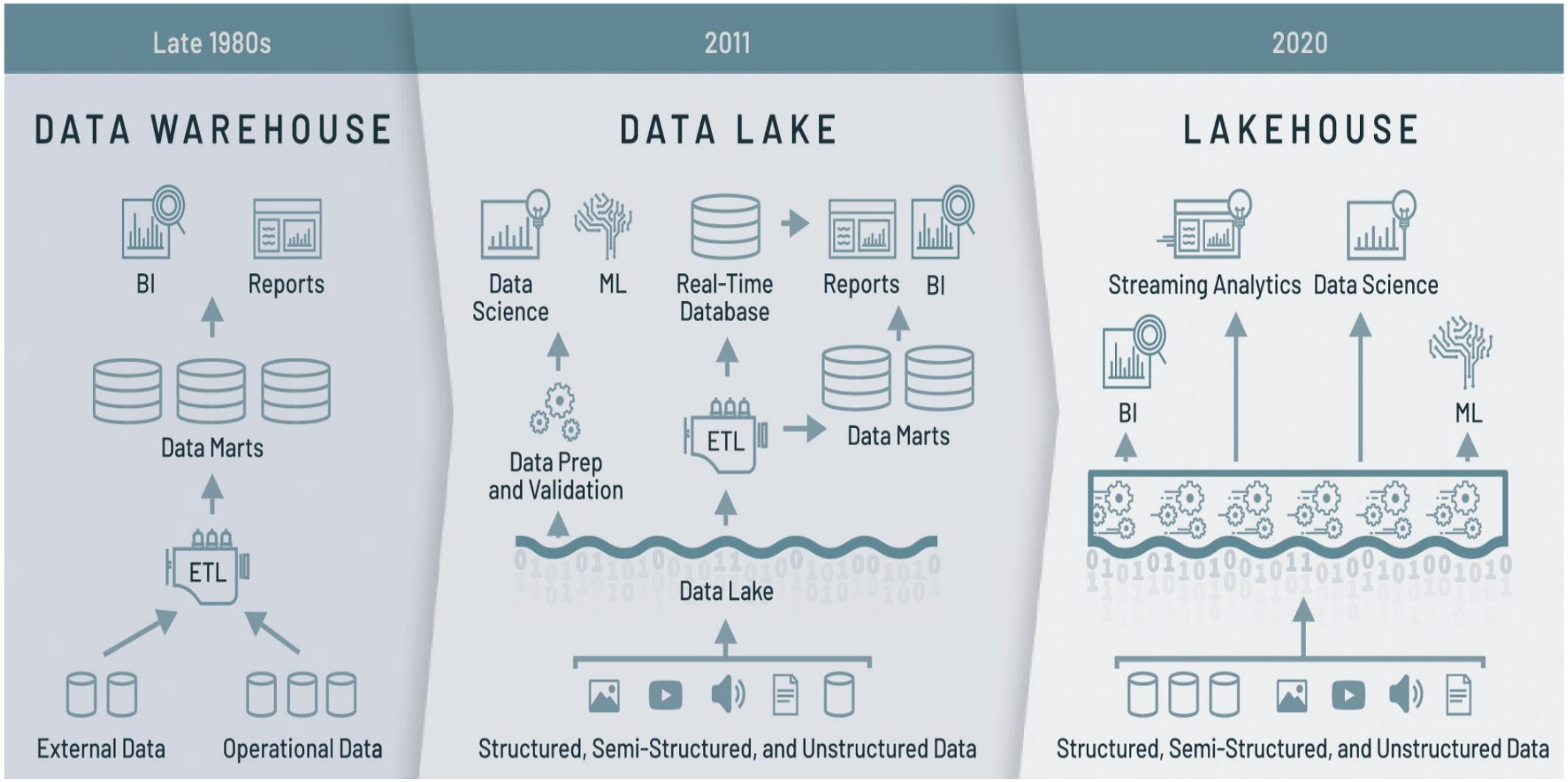
Infraestructure-as-code



HELM

Build
Release
Deploy
build
release
deploy_produc...
deploy_staging

From Datawarehouse to Lakehouse


Late 1980s
2011
2020
DATA WAREHOUSE
BI
Reports
Data Marts
ETL
External Data
Operational Data
DATA LAKE
Data Science
ML
Real-Time Database
Reports
BI
Data Prep and Validation
ETL
Data Marts
Data Lake
Structured, Semi-Structured, and Unstructured Data
LAKEHOUSE
Streaming Analytics
Data Science
BI
ML
Structured, Semi-Structured, and Unstructured Data

From ETL to ELT

Extract

Source-specific routines to pull selected data from an external system.

Transform

Business logic specific to your organization to serve an analytics or operationa

Load

Destination specific routines to push data where it is going to be consumed.

ETL doesn’t work in today’s world

Inflexible

- Friction when changing an existing pipeline.
- Hard to add new data.
- Most issues force data to be re-extracted.

Lack of Autonomy

- Warehouses made data consumers more autonomous.
- Changes require engineering involvement.

Complex

- Custom DSL.
- Force adoption of a data stack.
- Address 70% of the needs, 30% still built and maintained in-house.

Extract

General-purpose routines to pull selected data from a source.

Load

General-purpose routines to push raw data where it is going to be consumed.

Transform

Business logic specific to your organization to serve an analytics or operational use case with SQL / dbt / ...

ELT fixes the ETL-related issues

Flexibility

- All the data available on the destination.
- Data consumers are free to use what they need for the insights they want.

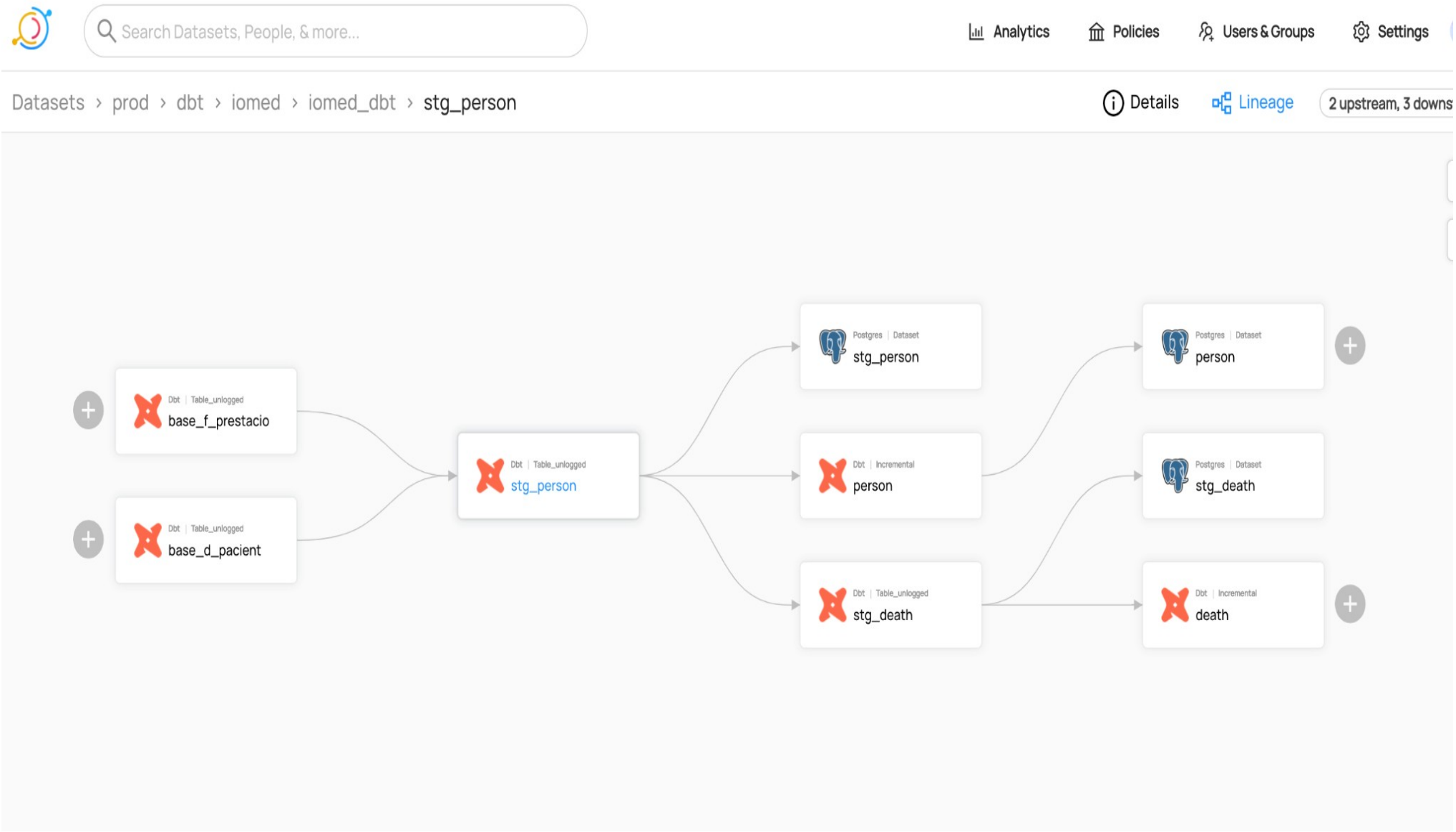
Autonomy

- Data consumers can leverage SQL queries to transform the data the way they want.
- No need to involve the engineering team.

Future proof

- Issues during transformation don't prevent access to the data.
- Easy to update transformation schemas.

Data governance and observability


Search Datasets, People, & more...
Analytics
Policies
Users & Groups
Settings
Datasets › prod › dbt › iomed › iomed_dbt › stg_person
Details
Lineage
2 upstream, 3 downs
Dbt | Table_unlogged
base_f_prestacio
Dbt | Table_unlogged
base_d_pacient
Dbt | Table_unlogged
stg_person
Postgres | Dataset
stg_person
Dbt | Incremental
person
Dbt | Table_unlogged
stg_death
Postgres | Dataset
person
Postgres | Dataset
stg_death
Dbt | Incremental
death

Self-service analytics

PRODUCTION APPMED

Search by Title...

Recent

PERSON

Person analysis

Domain Visit

DRUGS (QA)

EDA: DRUG_EXPOSURE (H12O)

Favorites

No items in this section.

Lists

OMOP

All My Docs

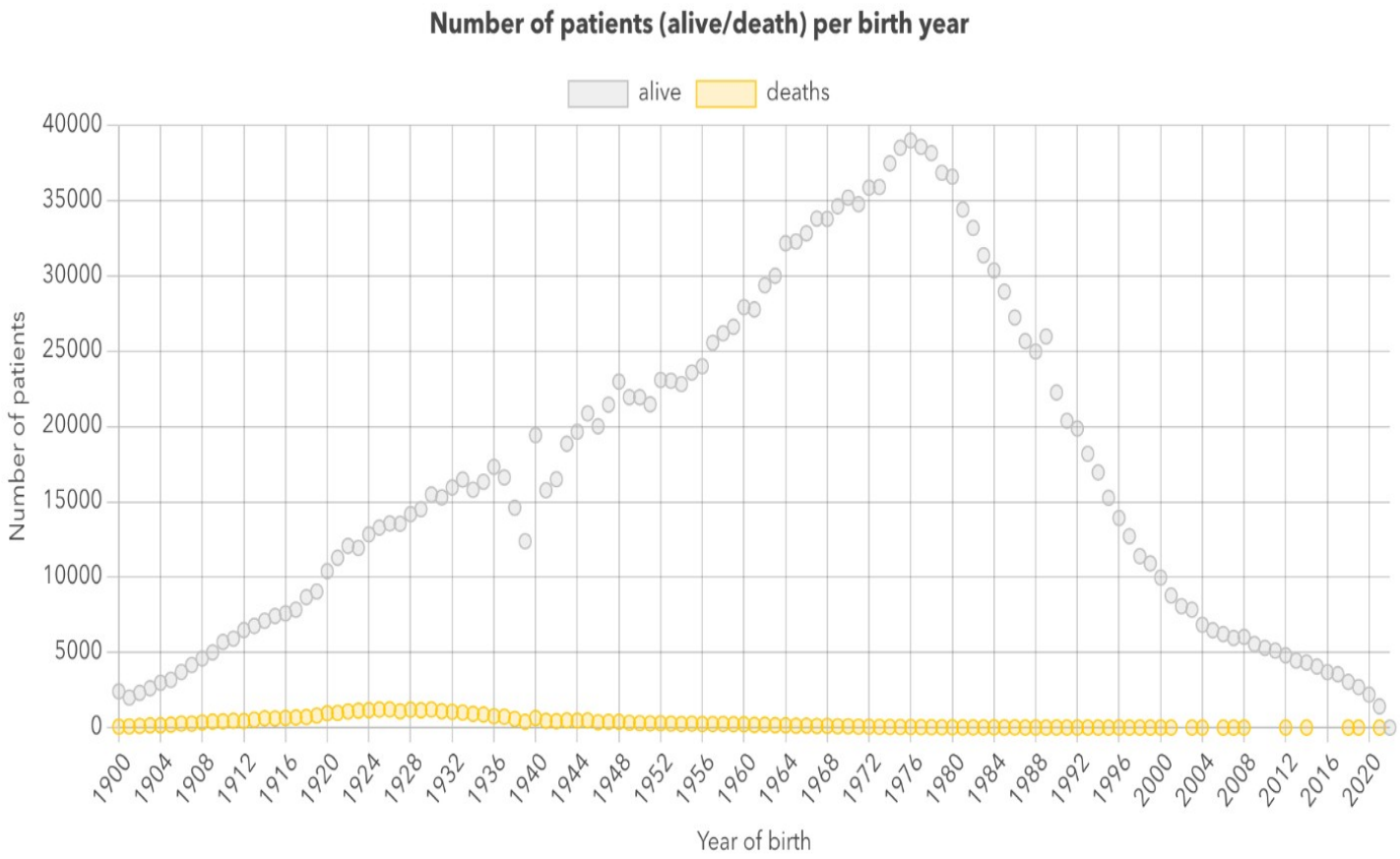
Untitled

Adhoc Docs Tables Snippets Runs Search

EXECUTION 90

STATEMENT 1 OUT OF 1

CONFIG CHART



Let’s go!

alabarga / pybcn22-modern-data-stack Public

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

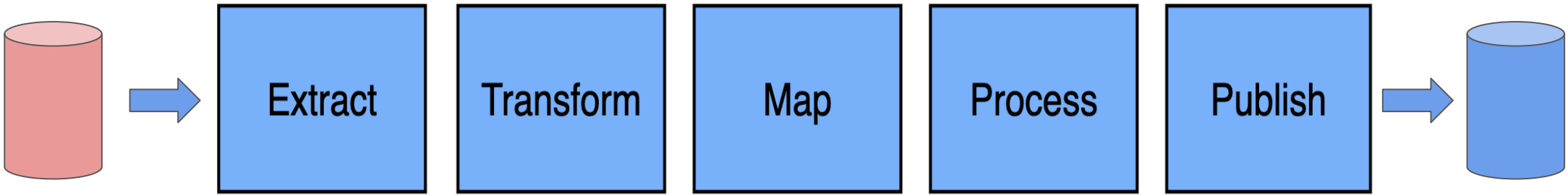
main | 1 branch | 0 tags | Go to file | Add file | Code

alabarga Add obevo script dca31bd 44 minutes ago 23 commits

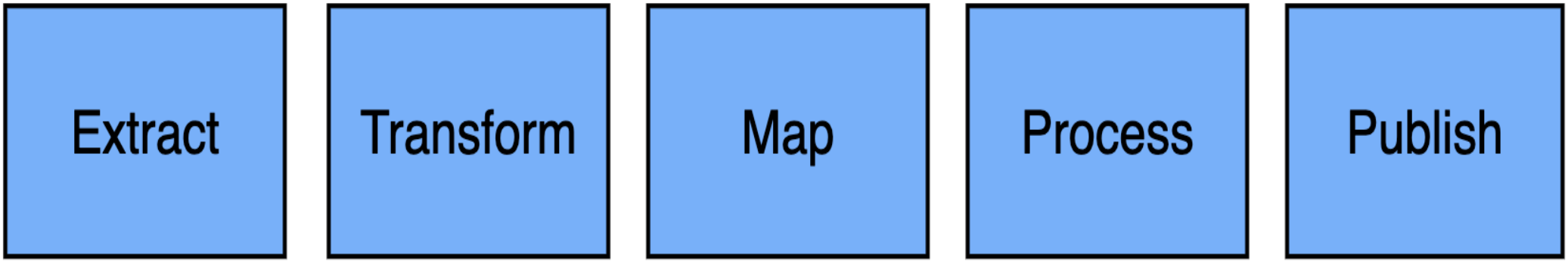
database	Add files	1 hour ago
dremio	Add files	1 hour ago
obevo	Add obevo script	44 minutes a
synthea	Merge branch 'main' of github.com:alabarga...	3 hours ago
.gitattributes	Add obevo	3 hours ago
.gitignore	Initial commit	12 days ago
LICENSE	Initial commit	12 days ago
README.md	Update README.md	yesterday

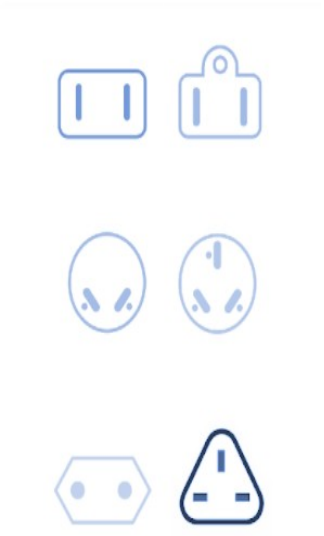
https://github.com/alabarga/pybcn22-modern-data-stack

Data processing

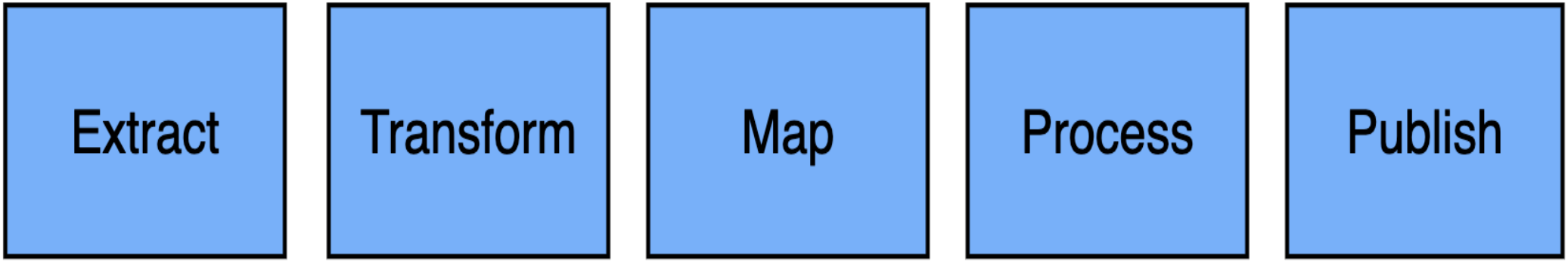


Data processing - Extraction





Data processing - Extraction



SYNTHEA

https://syntheticheath.github.io/synthea/

Data processing – Generating synthetic data

Simpy: https://gitlab.com/team-simpy/simpy

```
>>> import simpy
>>>
>>> def clock(env, name, tick):
...     while True:
...         print(name, env.now)
...         yield env.timeout(tick)
...
>>> env = simpy.Environment()
>>> env.process(clock(env, 'fast', 0.5))
<Process(clock) object at 0x...>
>>> env.process(clock(env, 'slow', 1))
<Process(clock) object at 0x...>
>>> env.run(until=2)
fast 0
slow 0
fast 0.5
slow 1
fast 1.0
fast 1.5
```

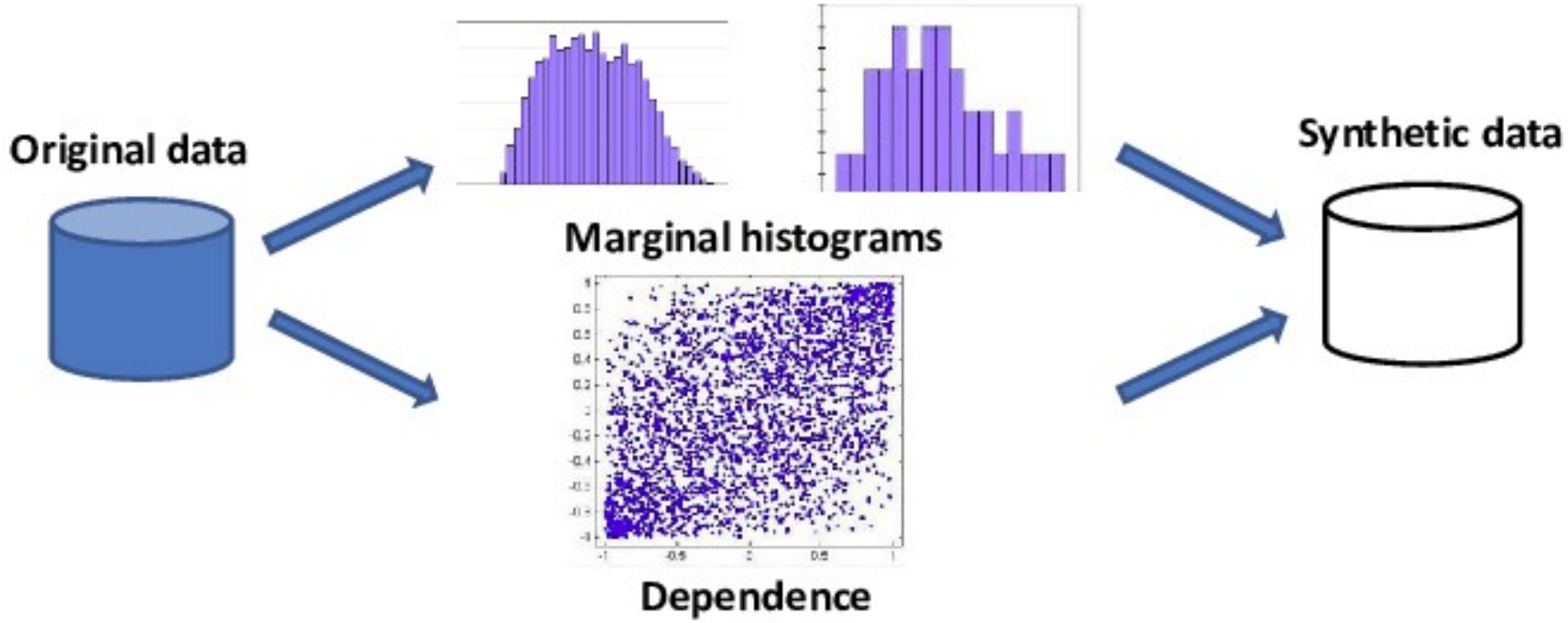
Data processing – Generating synthetic data

Faker: https://github.com/joke2k/faker

```
from faker import Faker
fake = Faker()

fake.name()
# 'Lucy Cechtelar'

fake.address()
# '426 Jordy Lodge
#  Cartwrightshire, SC 88120-6700'

fake.text()
# 'Sint velit eveniet. Rerum atque repellat voluptatem quia re
#  beatae sint laudantium consequatur. Magni occaecati itaque
#  amet quidem. Iusto deleniti cum autem ad quia aperiam.
#  A consectetur quos aliquam. In iste aliquid et aut similiqu
#  quaerat iste minus hic expedita. Consequuntur error magni e
#  voluptatem sit aliquam. Dolores voluptatum est.
```

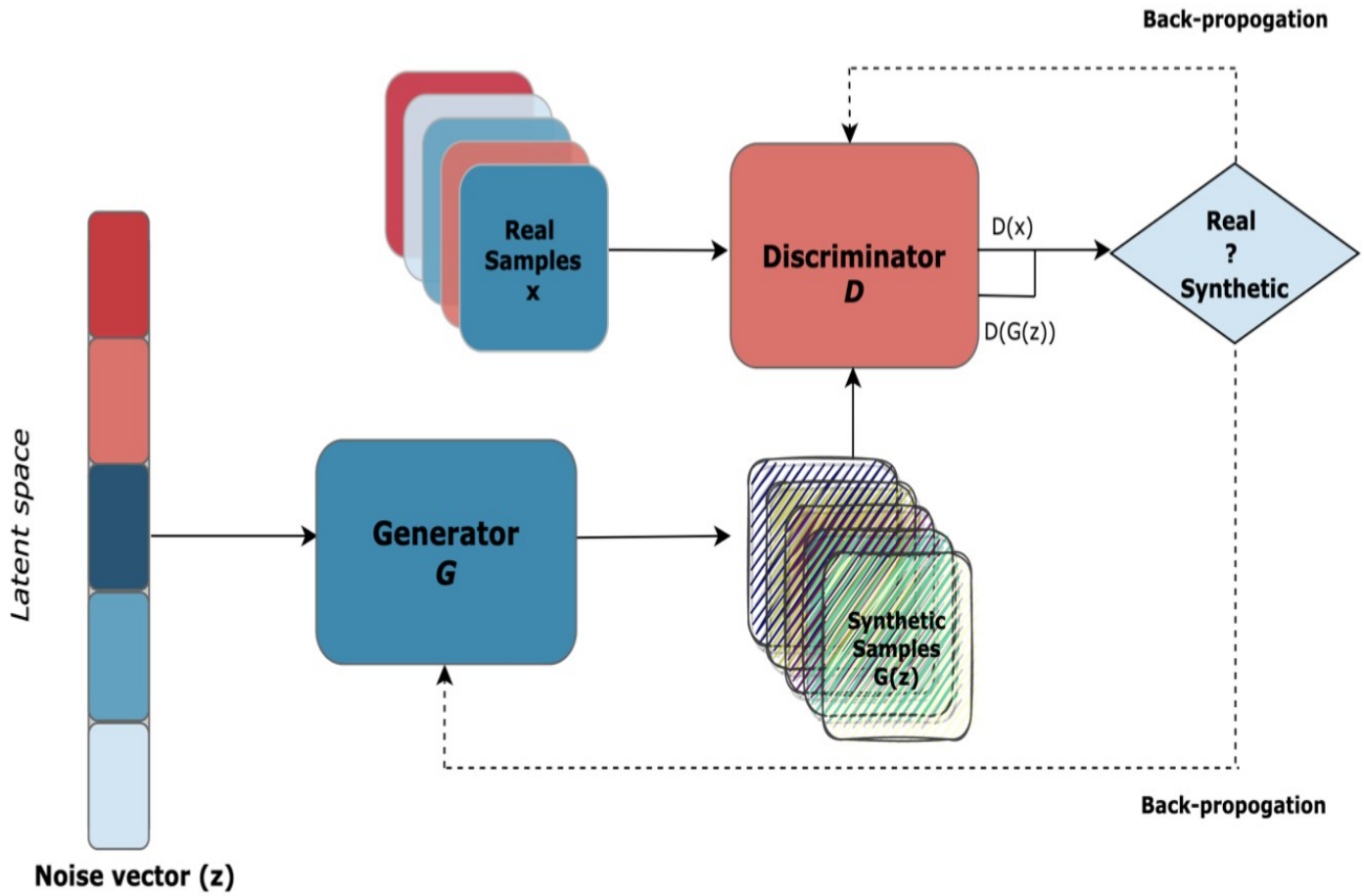
Data processing – Generating synthetic data

y-data: https://github.com/ydataai/ydata-synthetic

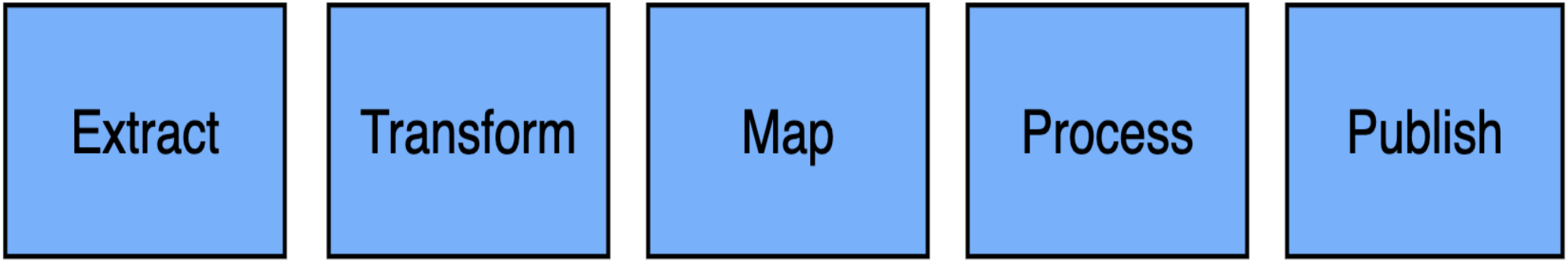


Data processing – Generating synthetic data

y-data: https://github.com/ydataai/ydata-synthetic



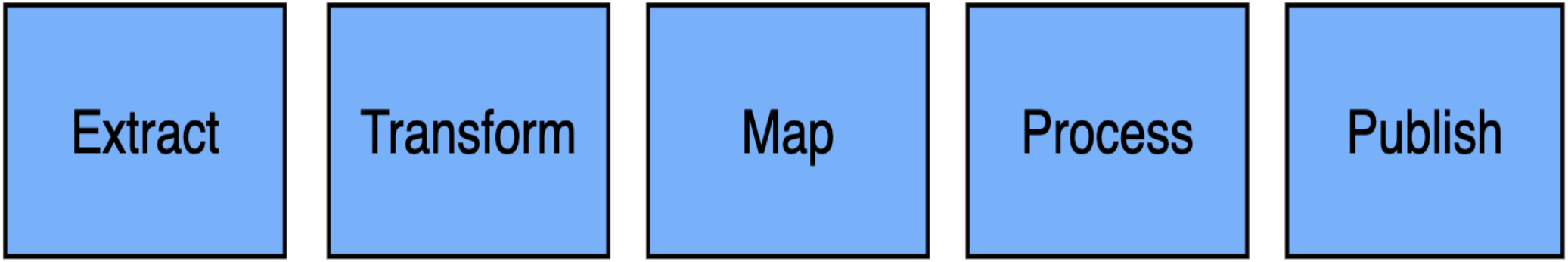
Data processing - Data ingestion



Airbyte

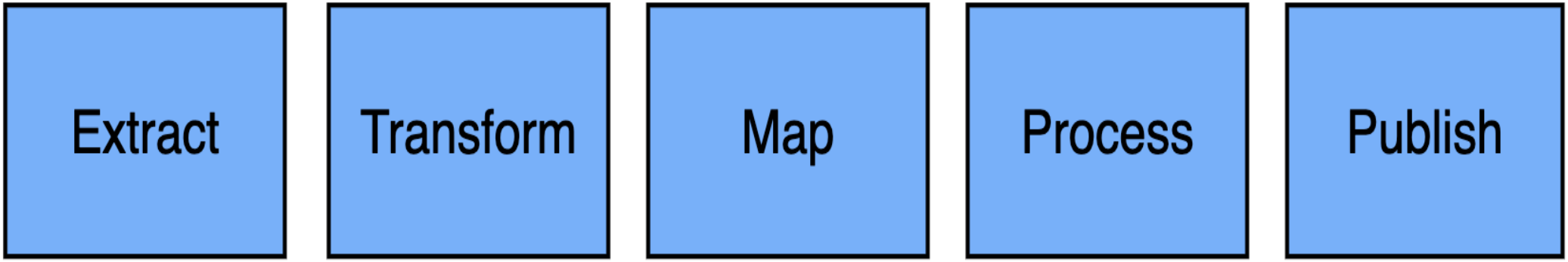
https://github.com/airbytehq/airbyte

Data processing - Data Warehouse



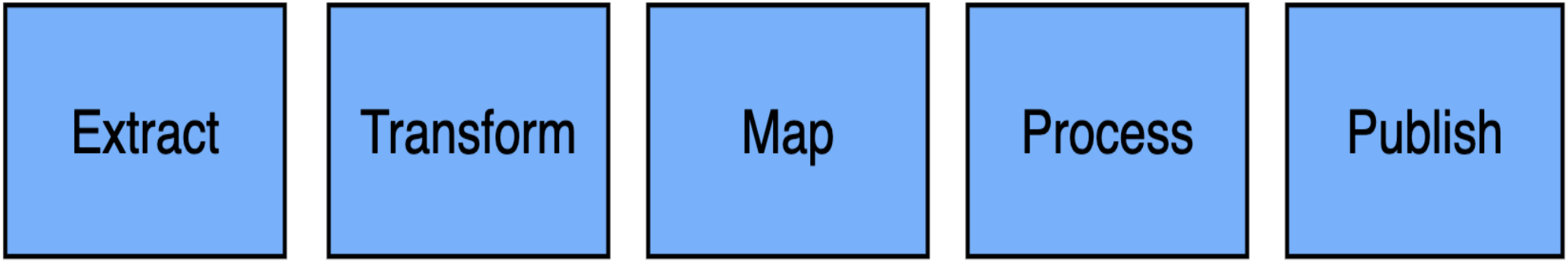
dremio

Data processing - Data Warehouse



PostgreSQL

Data processing - Data Warehouse

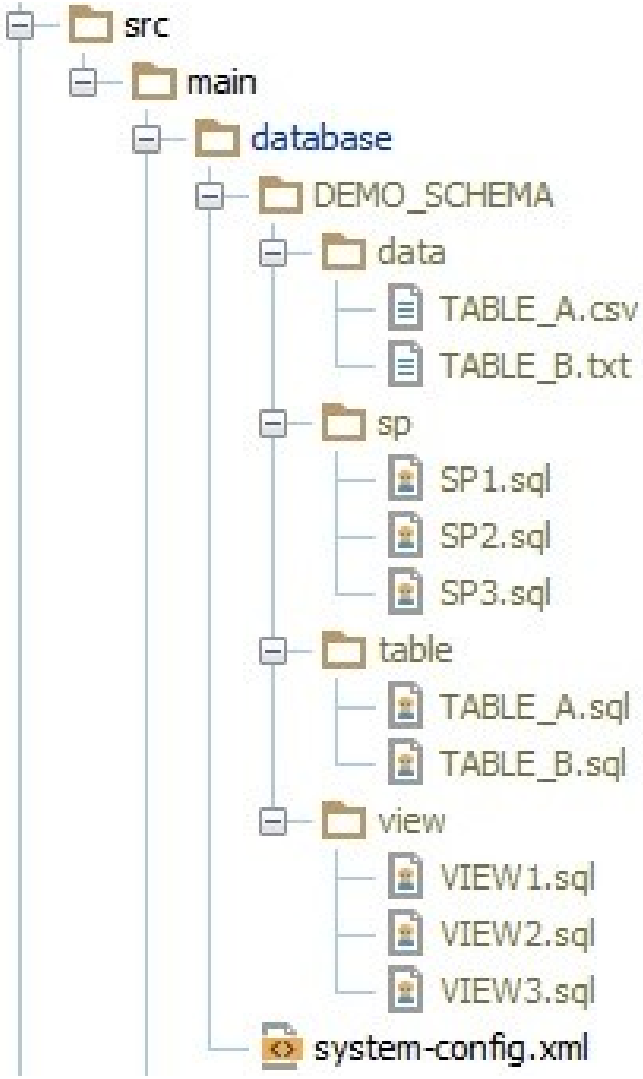


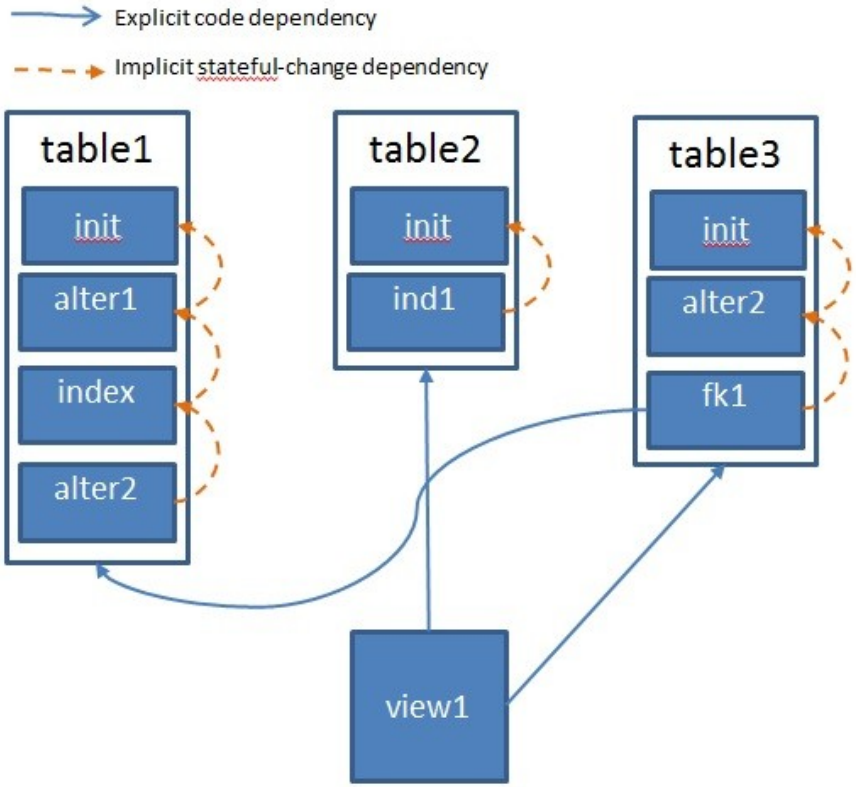
supabase

https://github.com/supabase/supabase

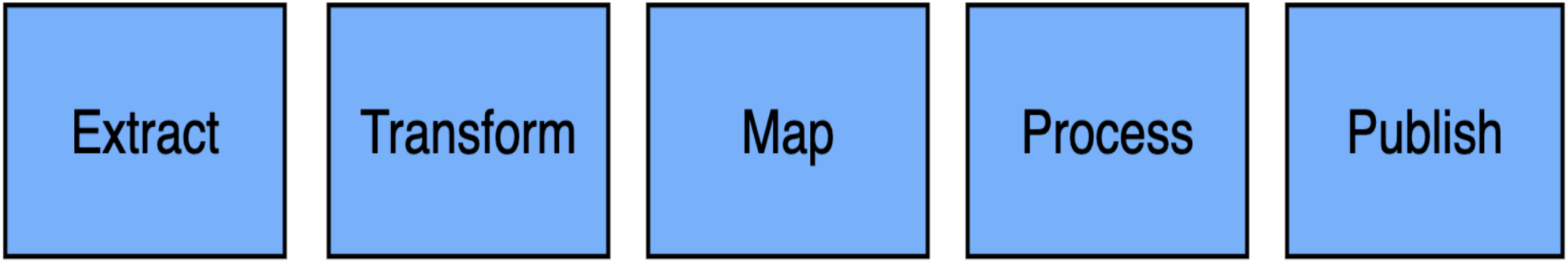
Data processing - Migrations

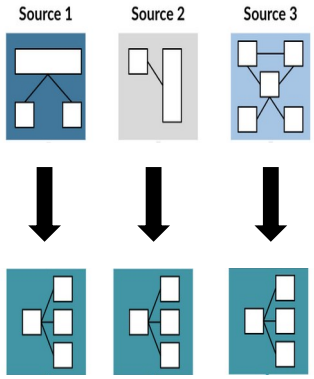
Obevo: https://github.com/goldmansachs/obevo





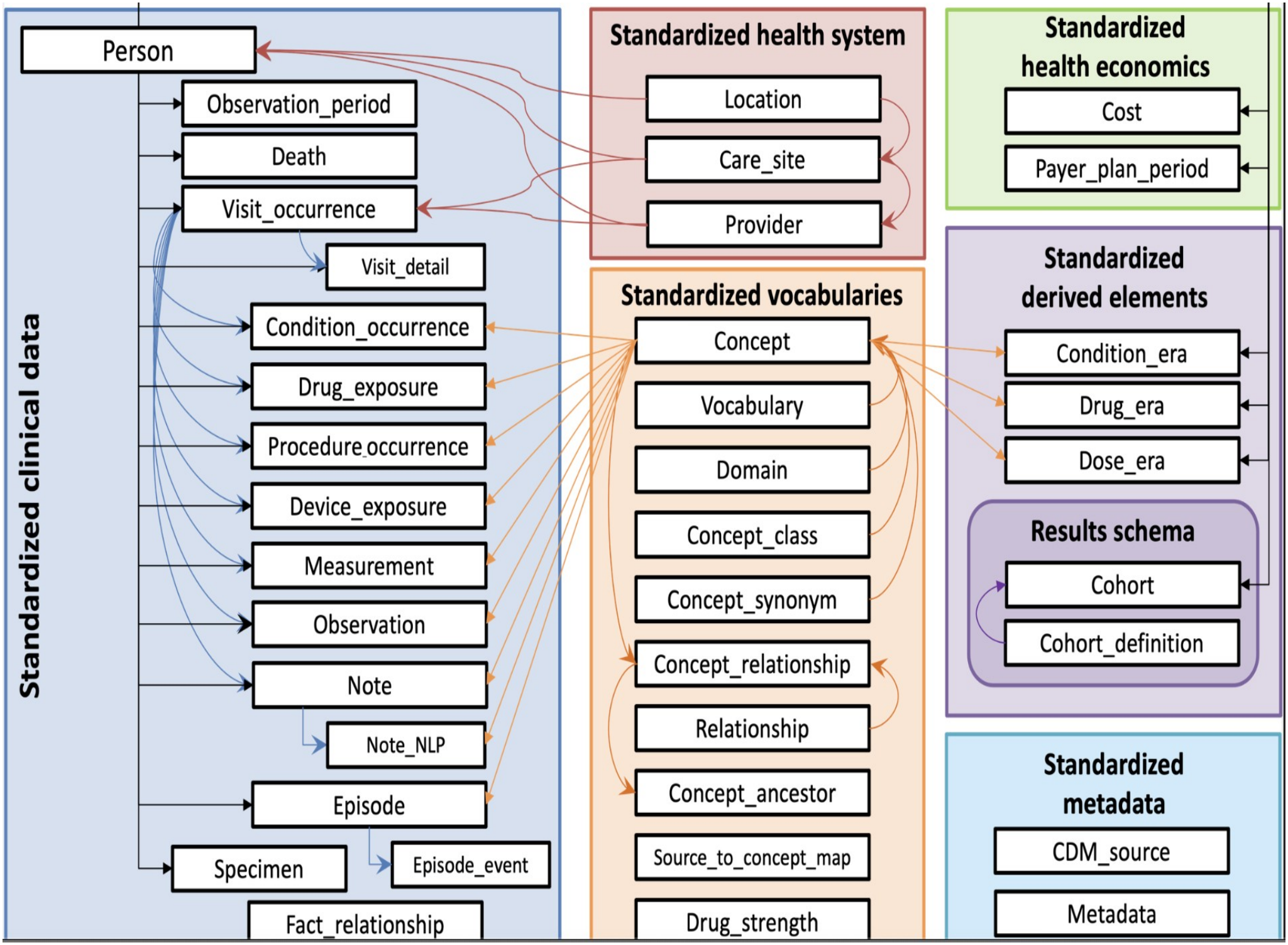
Data processing - ETL



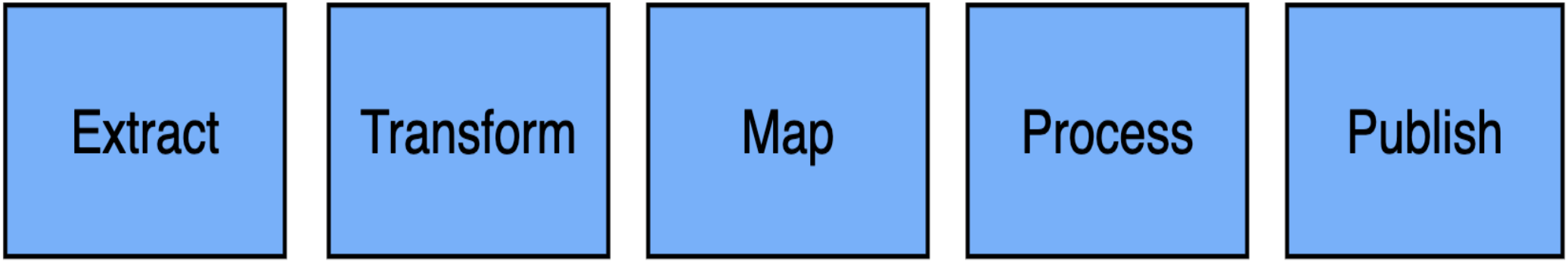


Esquema de datos común

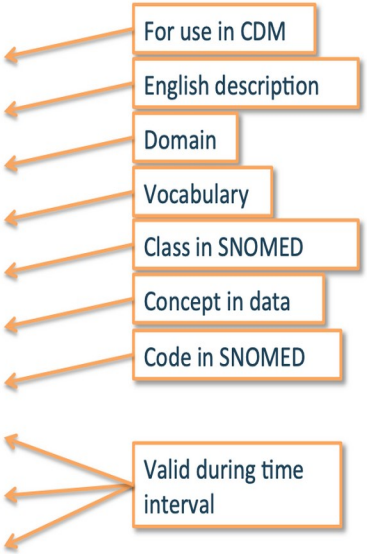
La misma estructura de base de datos

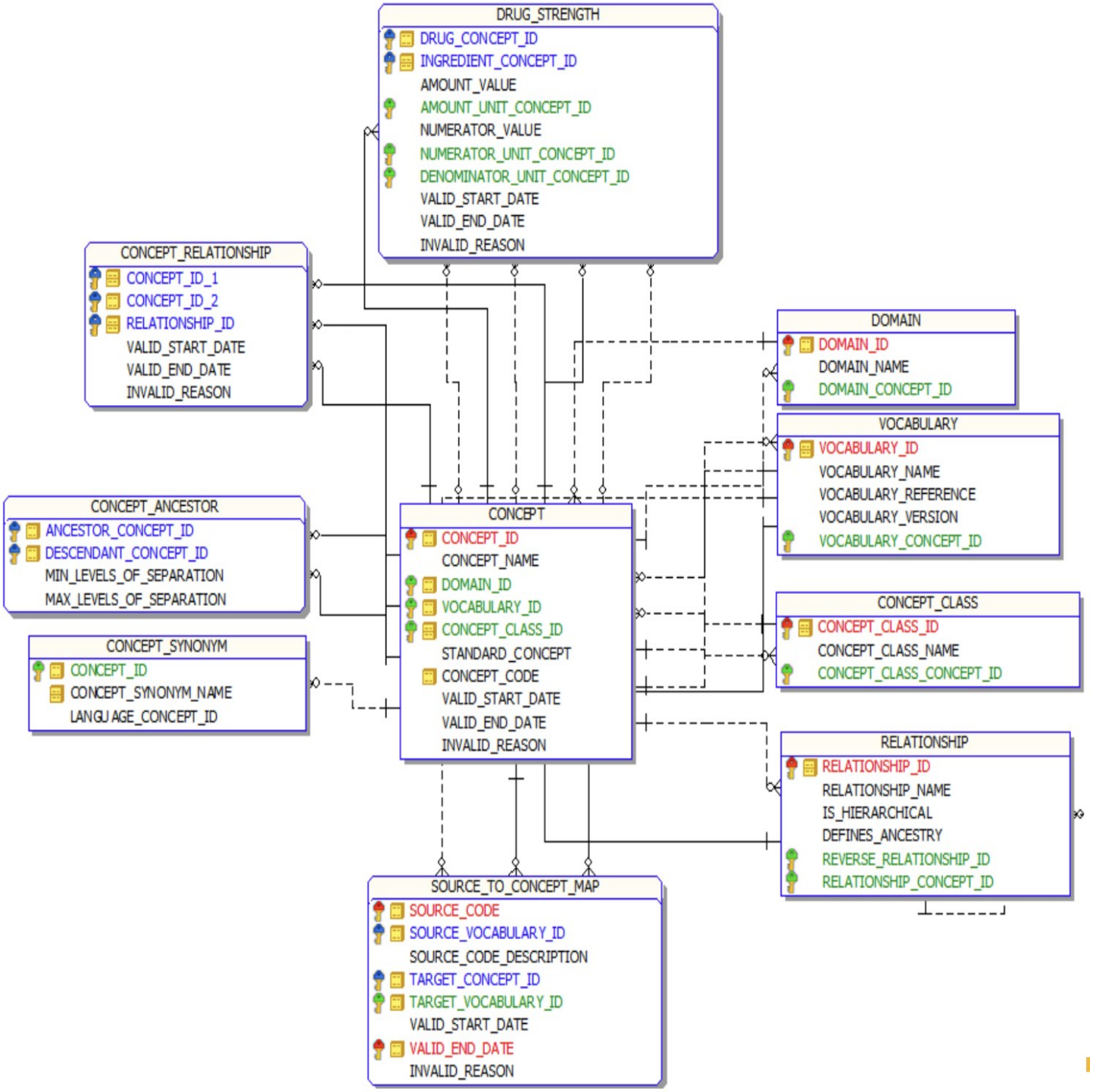

Standardized clinical data
Person
Observation_period
Death
Visit_occurrence
Visit_detail
Condition_occurrence
Drug_exposure
Procedure_occurrence
Device_exposure
Measurement
Observation
Note
Note_NLP
Episode
Episode_event
Specimen
Fact_relationship
Standardized health system
Location
Care_site
Provider
Standardized vocabularies
Concept
Vocabulary
Domain
Concept_class
Concept_synonym
Concept_relationship
Relationship
Concept_ancestor
Source_to_concept_map
Drug_strength
Standardized health economics
Cost
Payer_plan_period
Standardized derived elements
Condition_era
Drug_era
Dose_era
Results schema
Cohort
Cohort_definition
Standardized metadata
CDM_source
Metadata

Data processing - Mapping

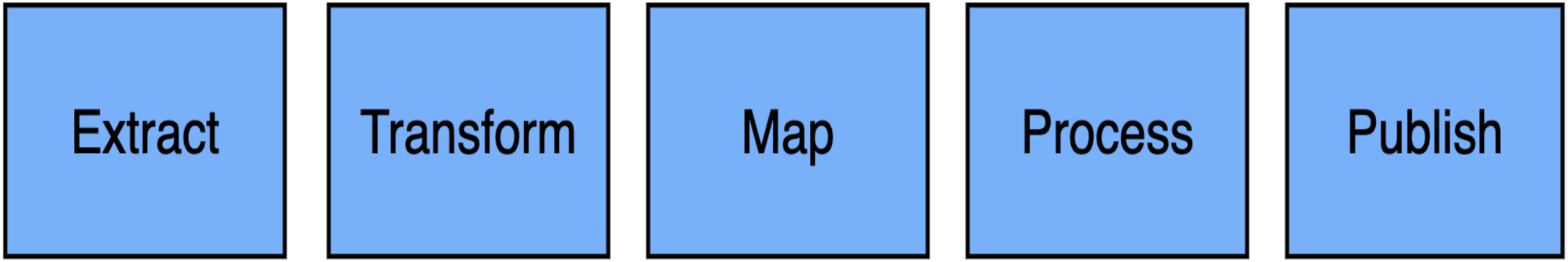


CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	



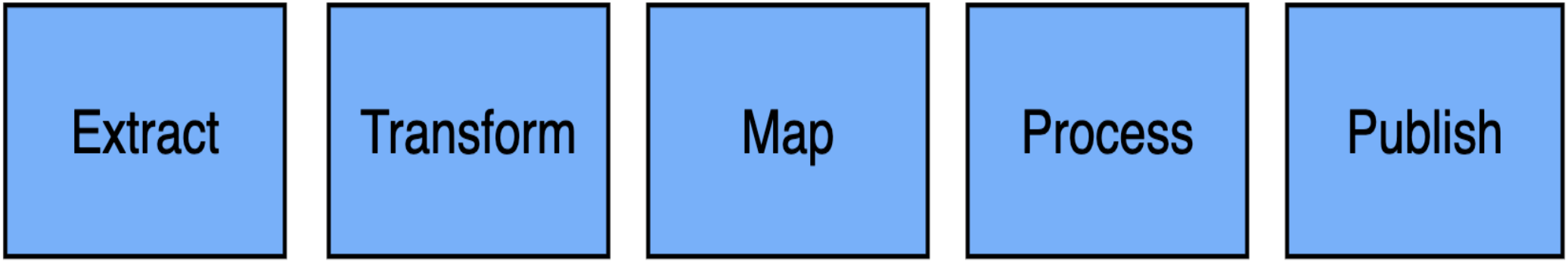

DRUG_STRENGTH
DRUG_CONCEPT_ID
INGREDIENT_CONCEPT_ID
AMOUNT_VALUE
AMOUNT_UNIT_CONCEPT_ID
NUMERATOR_VALUE
NUMERATOR_UNIT_CONCEPT_ID
DENOMINATOR_UNIT_CONCEPT_ID
VALID_START_DATE
VALID_END_DATE
INVALID_REASON
CONCEPT_RELATIONSHIP
CONCEPT_ID_1
CONCEPT_ID_2
RELATIONSHIP_ID
VALID_START_DATE
VALID_END_DATE
INVALID_REASON
DOMAIN
DOMAIN_ID
DOMAIN_NAME
DOMAIN_CONCEPT_ID
VOCABULARY
VOCABULARY_ID
VOCABULARY_NAME
VOCABULARY_REFERENCE
VOCABULARY_VERSION
VOCABULARY_CONCEPT_ID
CONCEPT_ANCESTOR
ANCESTOR_CONCEPT_ID
DESCENDANT_CONCEPT_ID
MIN_LEVELS_OF_SEPARATION
MAX_LEVELS_OF_SEPARATION
CONCEPT
CONCEPT_ID
CONCEPT_NAME
DOMAIN_ID
VOCABULARY_ID
CONCEPT_CLASS_ID
STANDARD_CONCEPT
CONCEPT_CODE
VALID_START_DATE
VALID_END_DATE
INVALID_REASON
CONCEPT_CLASS
CONCEPT_CLASS_ID
CONCEPT_CLASS_NAME
CONCEPT_CLASS_CONCEPT_ID
CONCEPT_SYNONYM
CONCEPT_ID
CONCEPT_SYNONYM_NAME
LANGUAGE_CONCEPT_ID
RELATIONSHIP
RELATIONSHIP_ID
RELATIONSHIP_NAME
IS_HIERARCHICAL
DEFINES_ANCESTRY
REVERSE_RELATIONSHIP_ID
RELATIONSHIP_CONCEPT_ID
SOURCE_TO_CONCEPT_MAP
SOURCE_CODE
SOURCE_VOCABULARY_ID
SOURCE_CODE_DESCRIPTION
TARGET_CONCEPT_ID
TARGET_VOCABULARY_ID
VALID_START_DATE
VALID_END_DATE
INVALID_REASON

Data processing - ELT



dbt

Data processing - Quality



great_expectations

https://github.com/great-expectations/great_expectations

Data processing - Quality

great_expectations Home / Validations / taxi.demo / yellow_tripdata_staging / 2020-08-04T16:38:46.210132+00:00

Actions

Validation Filter:

Show All | Failed Only

How to Edit This Suite

Show Walkthrough

Table of Contents

Overview

dropoff_datetime

fare_amount

passenger_count

pickup_datetime

trip_distance

vendor_id

passenger_count

Search

Status	Expectation	Observed Value
✗	values must always be between 1 and 6. **1579 unexpected values found. ≈15.79% of 10000 total rows.** **Sampled Unexpected Values** 0.0	≈15.79% unexpected

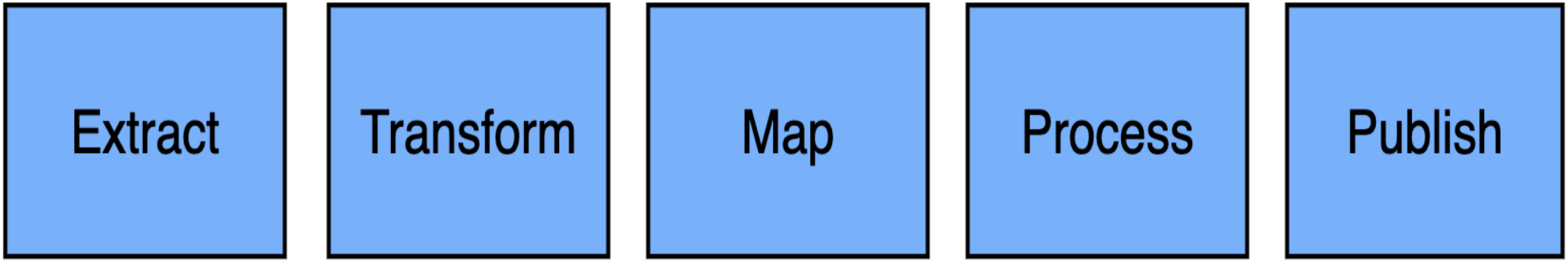
pickup_datetime

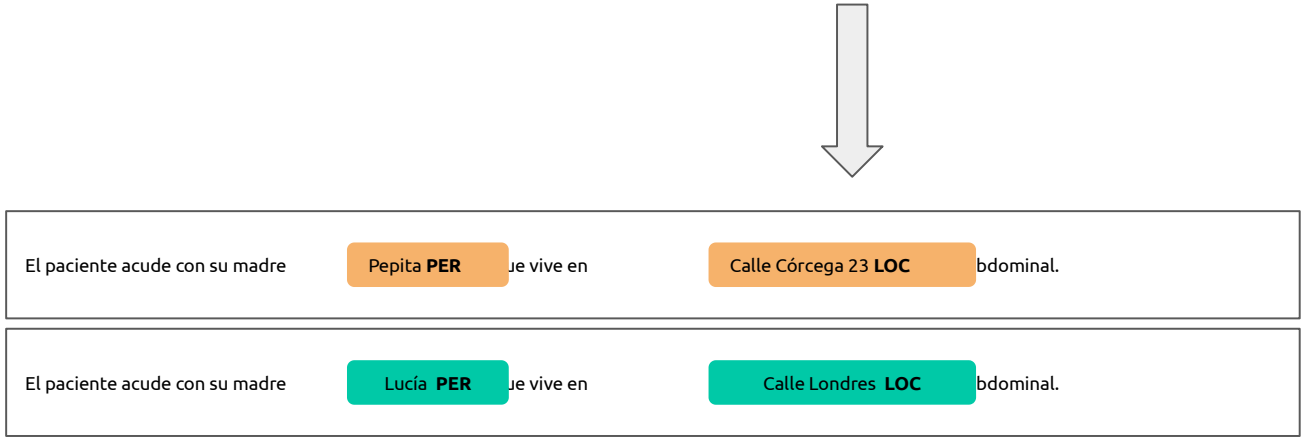
Search

Status	Expectation	Observed Value
✓	values must never be null.	100% not null

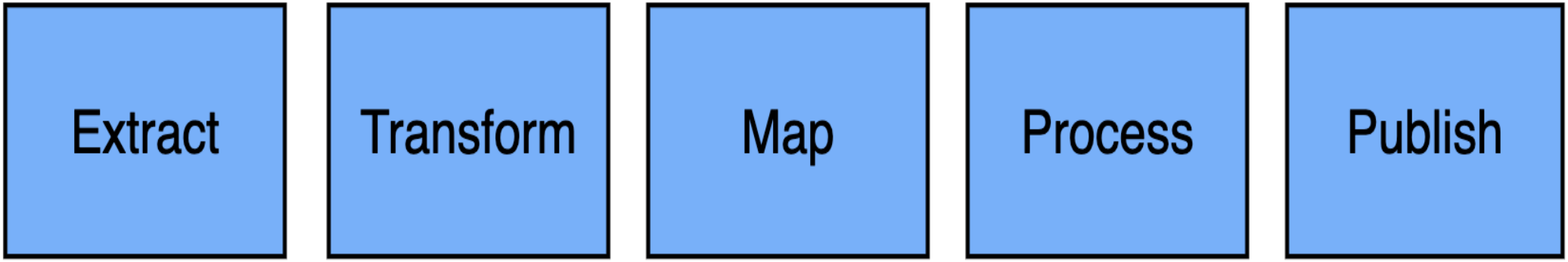
trip_distance

Data processing - Anonymization





Data processing - NLP



Acude con dolor abdominal derecho de 2 días de duración. Pauta de vacunación completa.

Finding *dolor abdominal.*

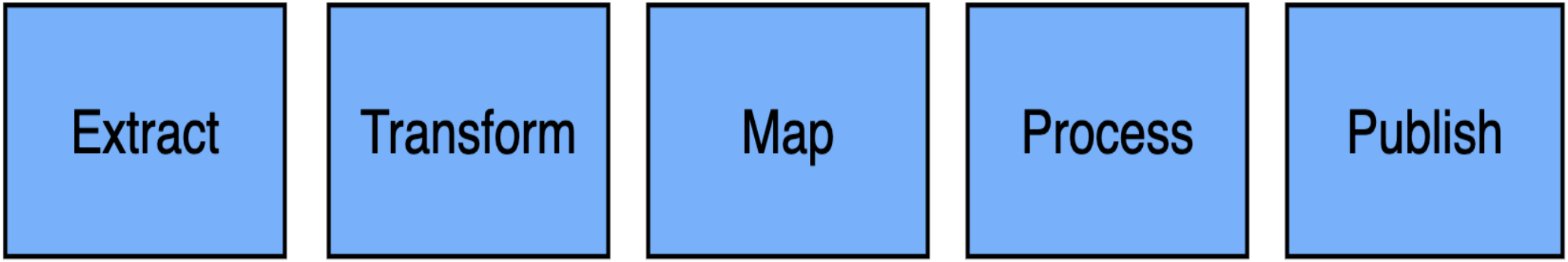
Spatial Concept *derecho.*

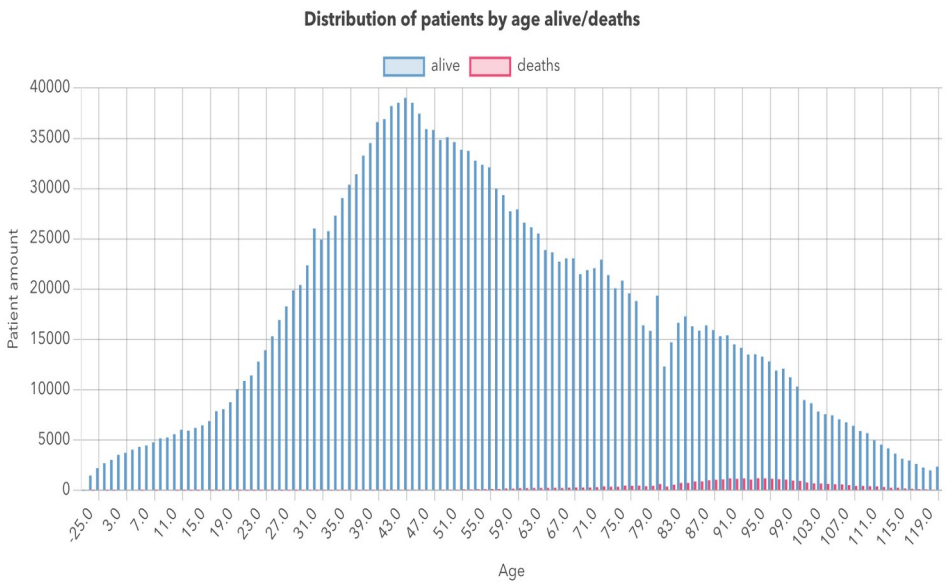
Temporal Concept *2 días, duración.*

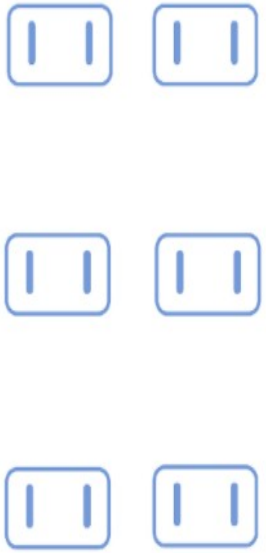
Therapeutic or Preventive Procedure *vacunación.*

Qualitative Concept *completa.*

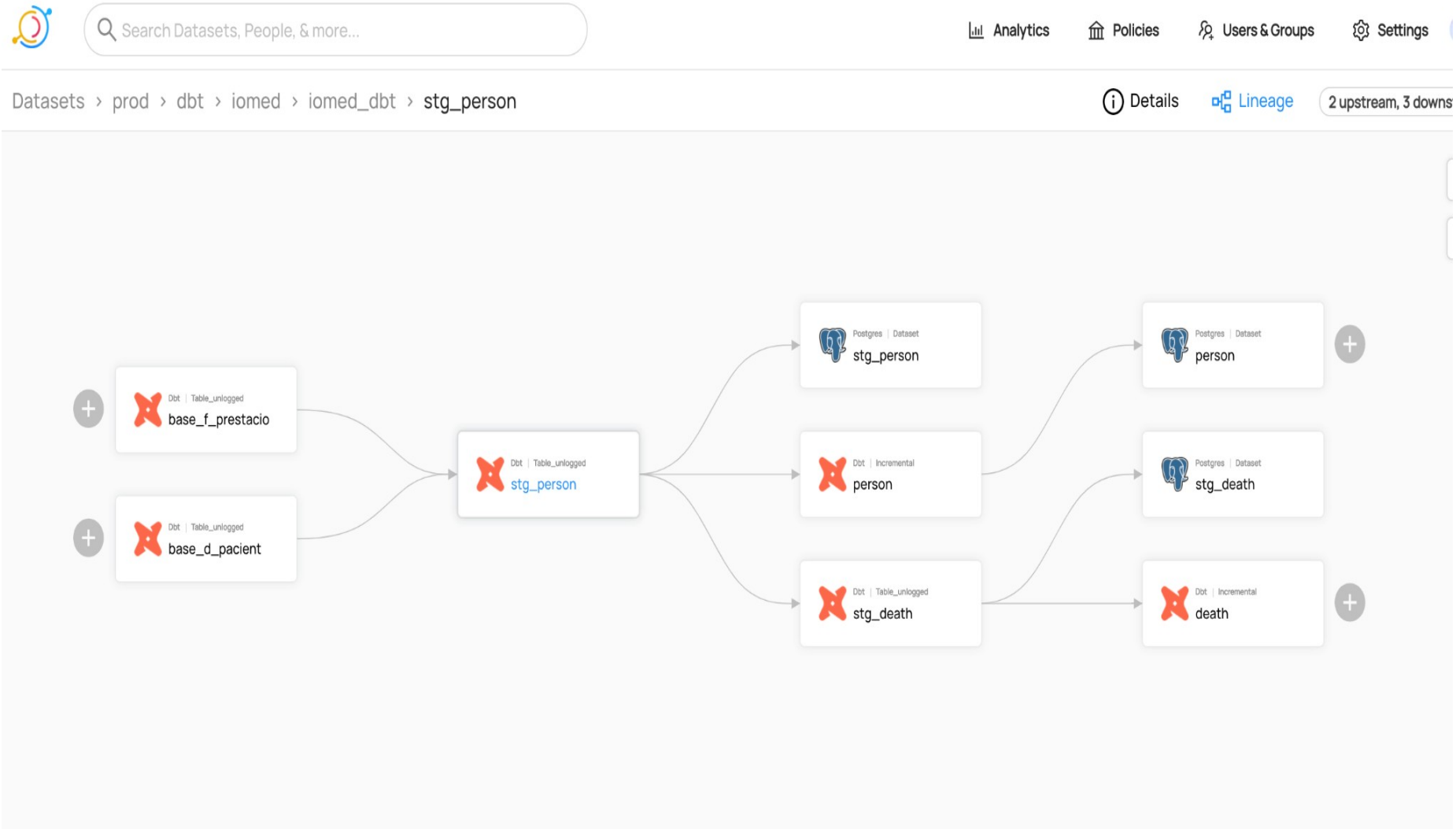
Data publication





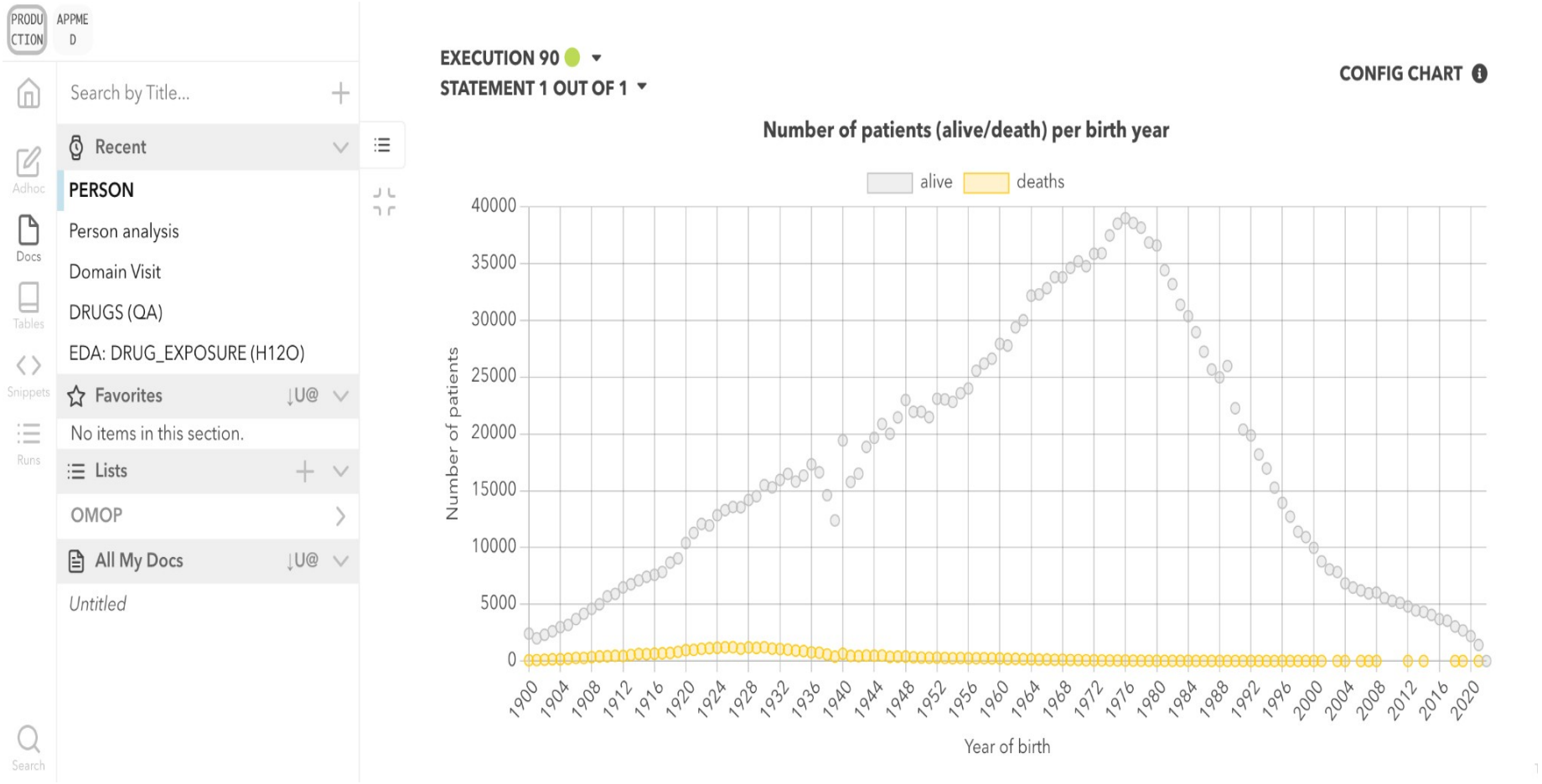


Data publication - Gobernance



https://datahubproject.io/

Data publication - Querybook



https://www.querybook.org/

Thanks!!!!!