Digital data collections: characteristics and properties

Genoveva Vargas-Solar
French Council of Scientific Research, LIG
genoveva.vargas@imag.fr

http://vargas-solar.com/data-centric-smart-everything/

https://classroom.google.com/c/MTQ4MzcwMjY1MDEz?cjc=5bz2tk6

Slack channel: https://join.slack.com/t/colenationale-5jr8199/shared_invite/zt-hhf9euv7-bmp7Kn9LL68RyzdhJnbKxA

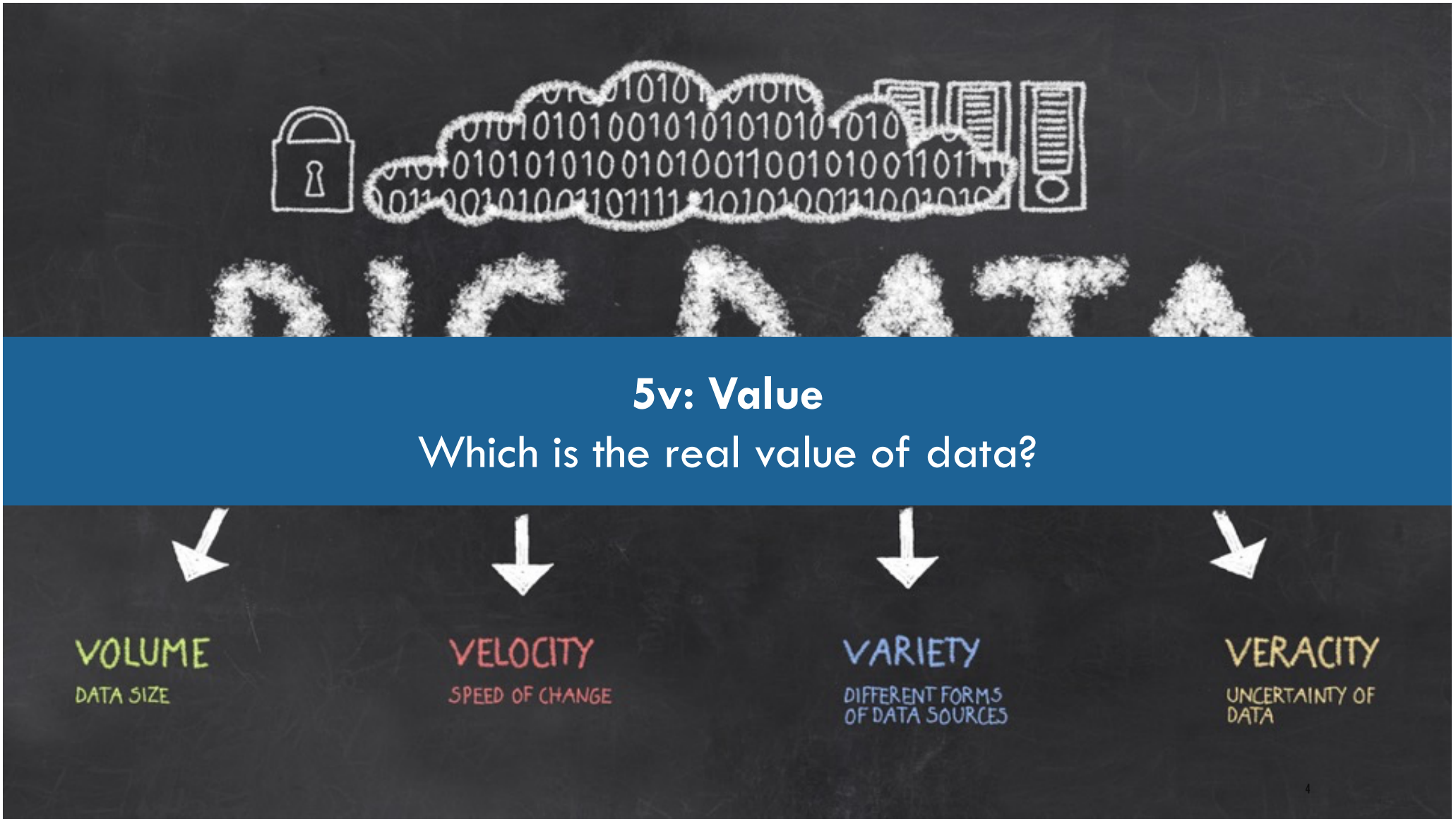
DATA

BIG DATA DEFINITION

- Data collections with characteristics difficult to process on single machines or traditional databases

- A new generation of tools, methods and technologies to collect, process and analyse massive data collections

  → Tools imposing the use of parallel processing and distributed storage


BIG DATA
5v: Value
Which is the real value of data?
VOLUME
DATA SIZE
VELOCITY
SPEED OF CHANGE
VARIETY
DIFFERENT FORMS OF DATA SOURCES
VERACITY
UNCERTAINTY OF DATA

BIG DATA PROPERTIES

3V

- **Volume** (size)
- **Velocity** (production rate)
- **Variety** (data types & format)

4V

- **Variability** (inconsistencies by constant meaning changes)

5V

- **Veracity** (truth and consistency)

…

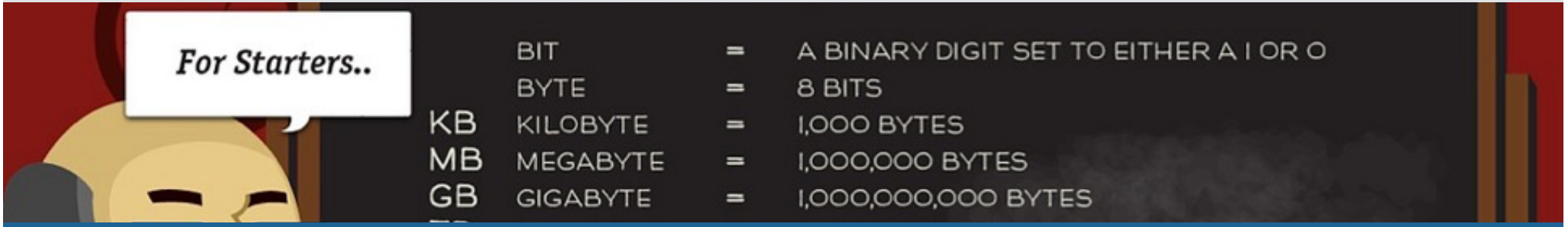
- **Value** (how much information)

10V

V's models [Jagadish 2014]

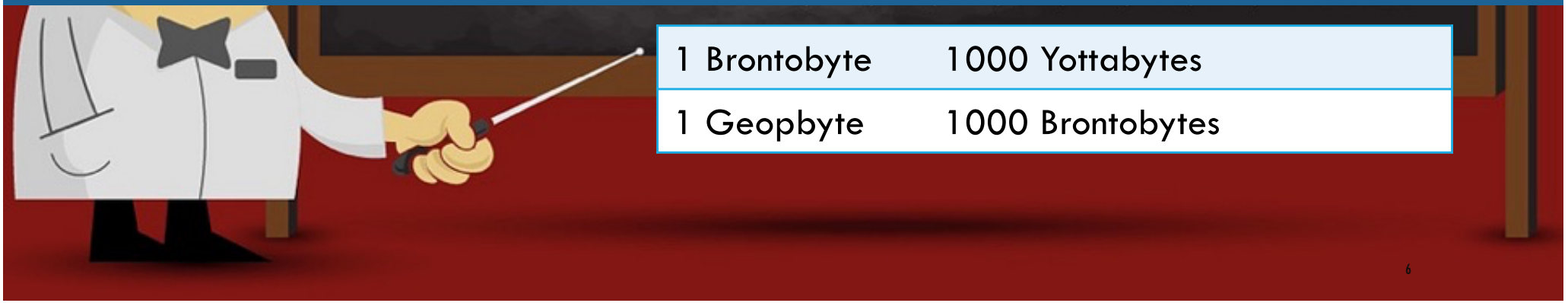
"Big Data can really be very small and not all large datasets are big!"
- Mike 2.0 [Hillard 2012]

HOW BIG IS YOUR DATA ?



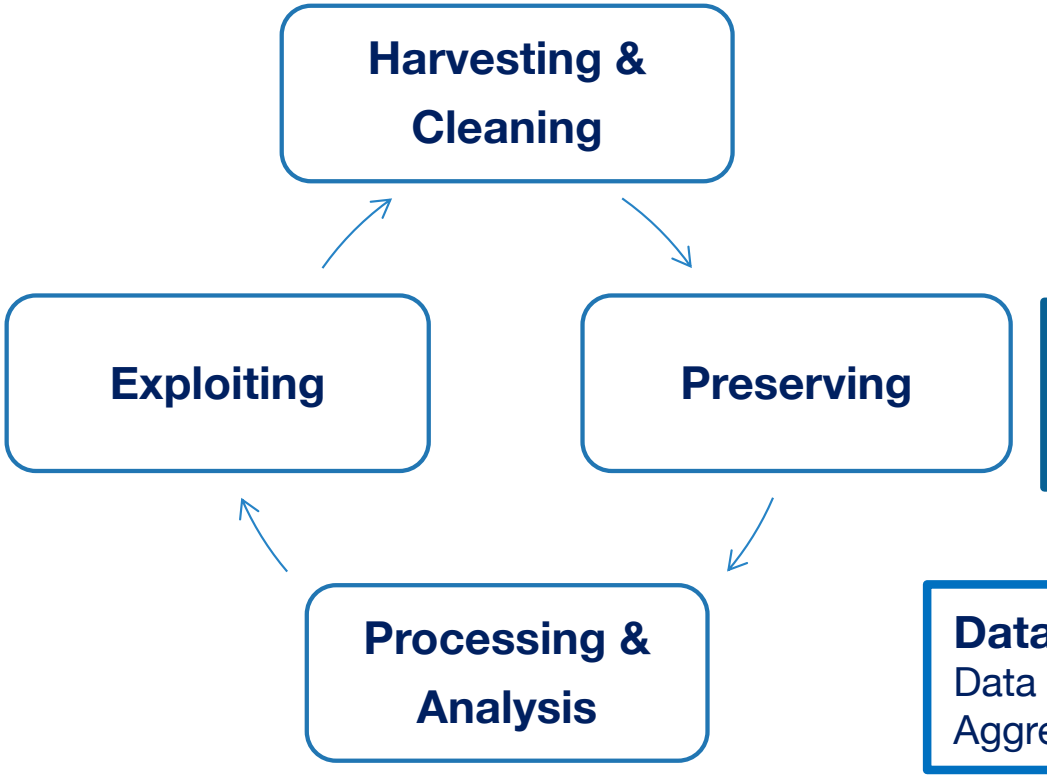
Helluva lot of data !!

http://spectrum.ieee.org/computing/software/beyond-just-big-data



1 Brontobyte	1000 Yottabytes
1 Geopbyte	1000 Brontobytes

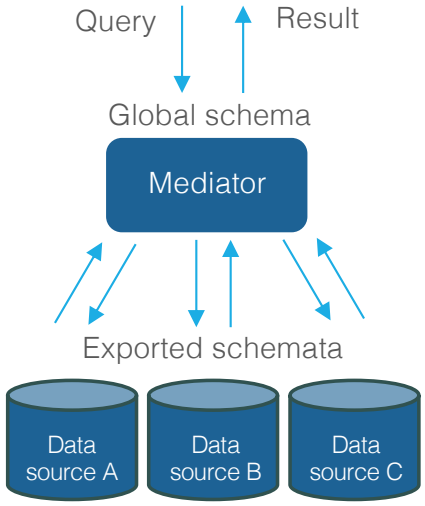
BIG DATA LIFE CYCLE



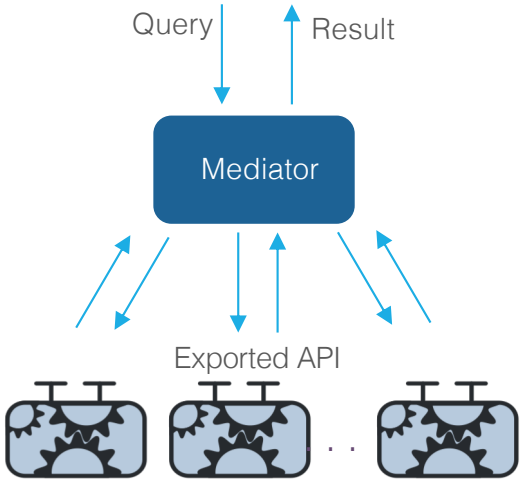
Data Collections Storage
NoSQL, NewSQL (Hive), Data sharding (MongoDB, CouchDB)

Data Processing
Data analytics [Feldman et al. 2013], Data Aggregation [Jayram et al. 2007]

DATA DESIGN & CONSUMPTION PHILOSOPHIES



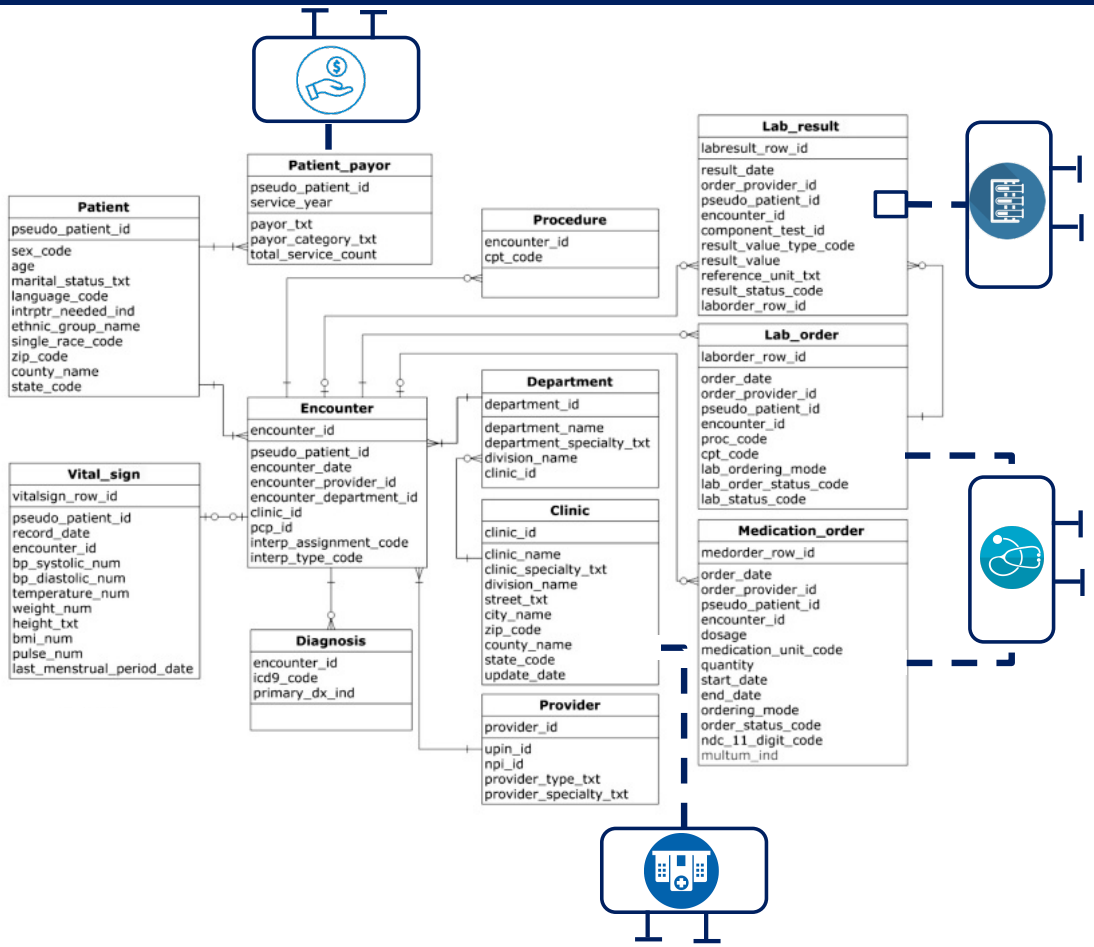
Heterogeneous data sources known in advance



Distributed data services with dynamic binding

- **Data:** model of a mini-world, it is a set of facts structured according to some data model
- **Query:** precisely stated it can include terms, operators (and/or/negation, relational, aggregation), and constraints
- **Result**: collection of items that completely or partially correspond to consumers requirements (precision & recall)

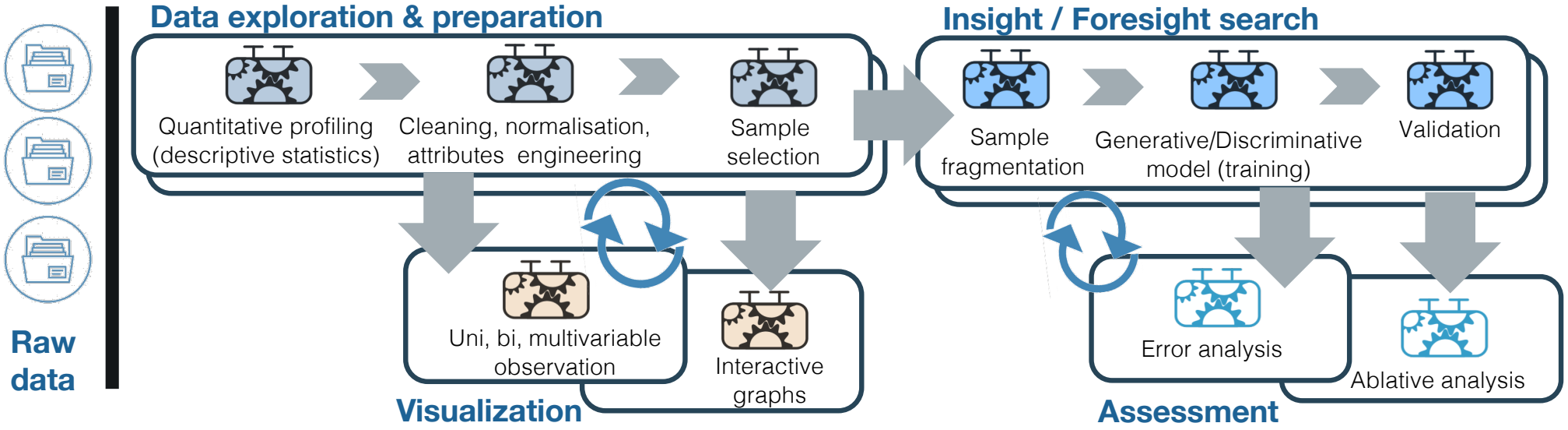
ASKING FACTS ABOUT DIABETES



- Average of lab & medication orders per patient and physician in a given clinic
- Vital signs and lab results used to emit a diagnosis for a given patient
- Number of patients with diabetes followed per clinic

DATA CONSUMPTION PHILOSOPHIES

Diogenes approach



- **Data**: observations of phenomena often described as series of features/attributes
- **Query**: analytics objective (looks for insights or foresights) expressed as a pipeline of operations guided by the conditions and characteristics of the data
- **Result**: a model or prediction with associated assessment indexes, not definitive accepted with an associated error margin, accepted by comparison

https://www.kaggle.com/hamzael1/hospital-beds-by-country

DIGITAL DATA COLLECTIONS

Consumed data:

- different sizes
- quality, uncertainty, ambiguity degree
- evolution in structure, completeness, production conditions, conditions in which data is retrieved
- content, explicit cultural, contextual, background properties
- access policies modification

Conditions of consumption:

- reproducibility, transparency degree (avoid “software artefacts”)

DIGITAL DATA COLLECTIONS

NEITHER MANAGEABLE NOR EXPLOITABLE AS SUCH

RAW DATA

- Heterogeneous (**variety**)
- Huge (**volume**)
- Incomplete, unprecise, missing, contradictory (**veracity**)
- Continuous releases produced at different rates (**velocity**)
- Proprietary, critical, private (**value**)

THE RIGHT DATA FOR THE RIGHT ANALYTICS

Different sizes, evolution in structure, completeness, production conditions & content, access policies modification …

Is the data clean? If not what has to be done to make it clean?

What information is available in these collections?

Are there relations between data collections which could be exploited for better prediction?

What are the update rates and in what way does this affect the collection?

DATA CURATION

Data management through its **lifecycle of interest** & **usefulness**

- Enable data **discovery & retrieval**
- Maintain data quality
- Add value
- Provide for **re-use over time**

DATA SPECTRUM

Querying objective

Data exploration

Hindsight

Insight

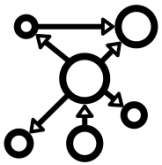
Foresight

Relational & aggregation querying

Information retrieval

Data

Data collections



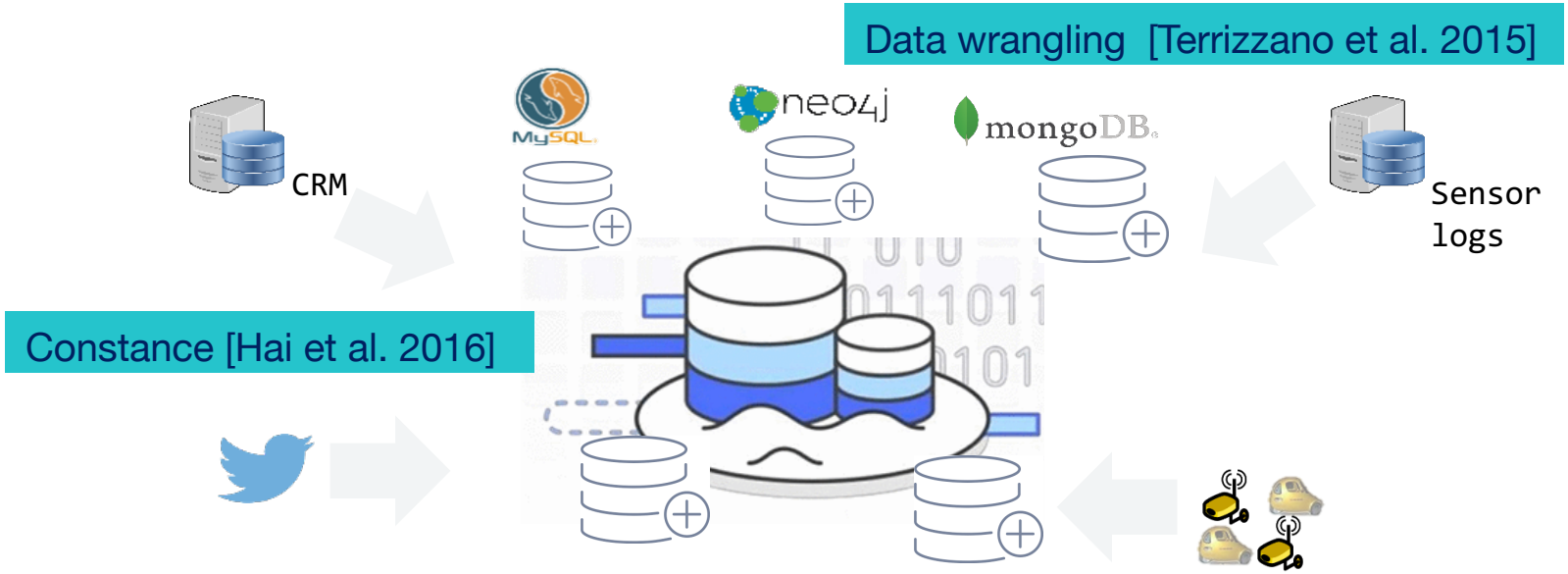
Data bases

Data Rawness Degree

Raw

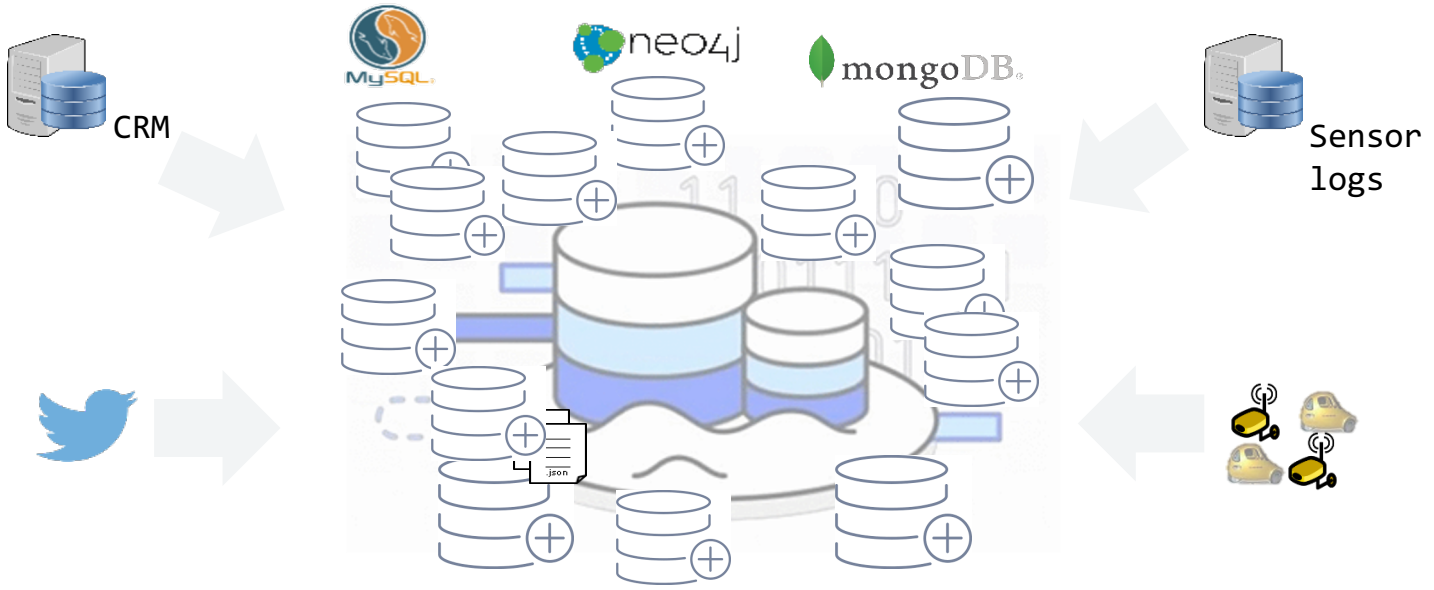
Curated

DATA LAKE



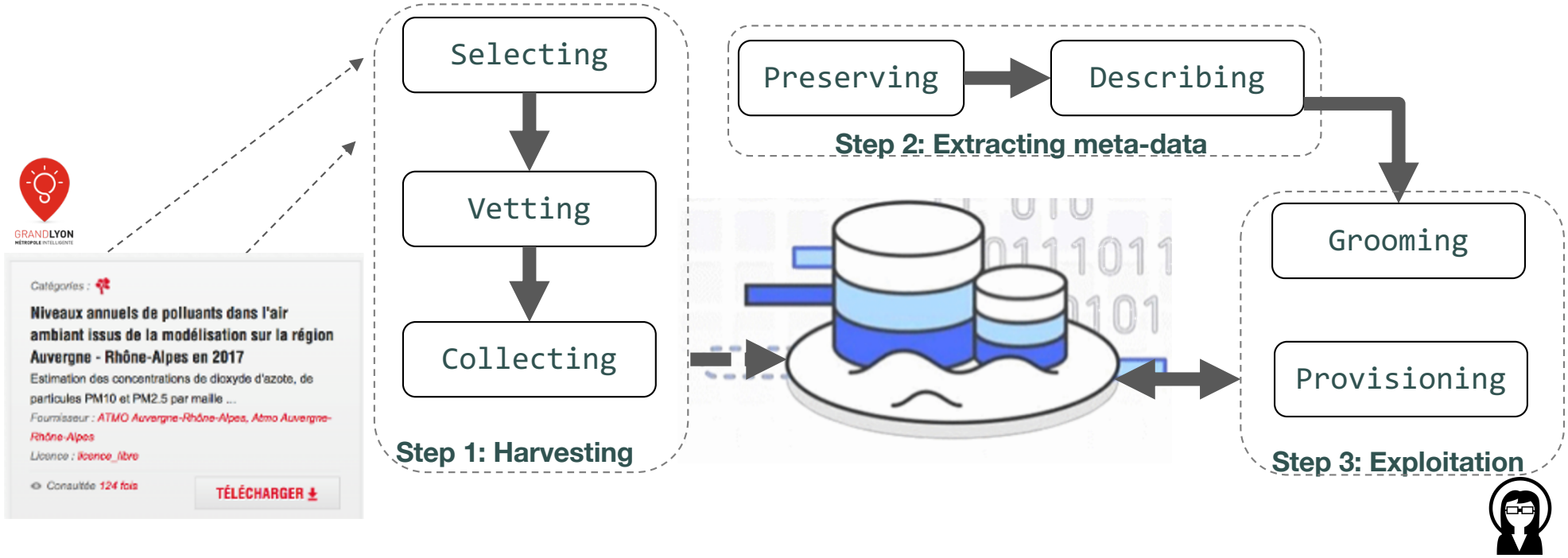
Centralized repository containing virtually inexhaustible amounts of raw data to be analysed

DATA SWAMP



Repositories grow ever bigger and complex to the point that a lake becomes a **swamp**

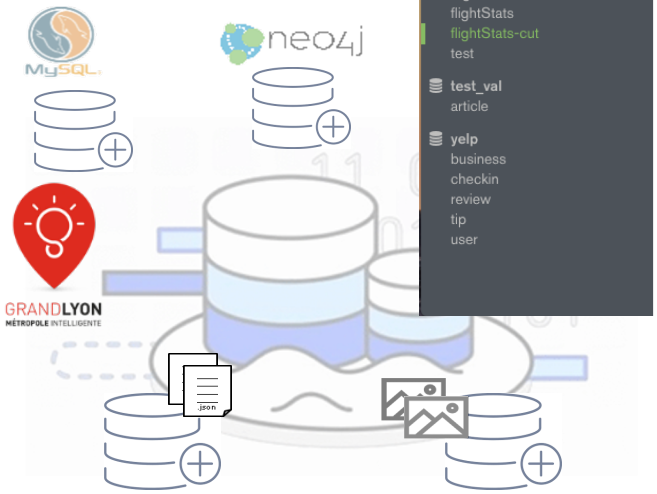
DATA CURATION WORKFLOW



Data wrangling [Terrizzano et al. 2015]

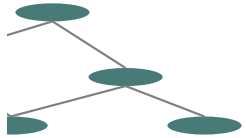
EXTRACTING META-DATA

Structural me
(e.g. Schemat
like DTD, XML



Quantitative meta-data
(e.g. Compass MongoDB)

-data
proaches)



ology

ons

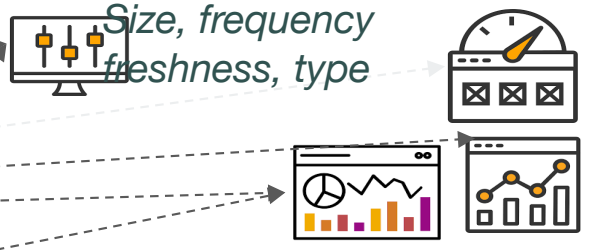
- Topic
- Geographic location
- Temporal releases

[Constance, SIGMOD '16]

EXTRACTING META-DATA

Sensing & collecting raw data

Size, frequency freshness, type



Quantitative & qualitative data + semantics

Level I
No meta-data

Level 2
Partial meta-data

Level 3
Complete meta-data

information

Harvesting → Preserving → **Describing** → Exploring

Extracting meta-data

[Stonebraker et al. 2015]

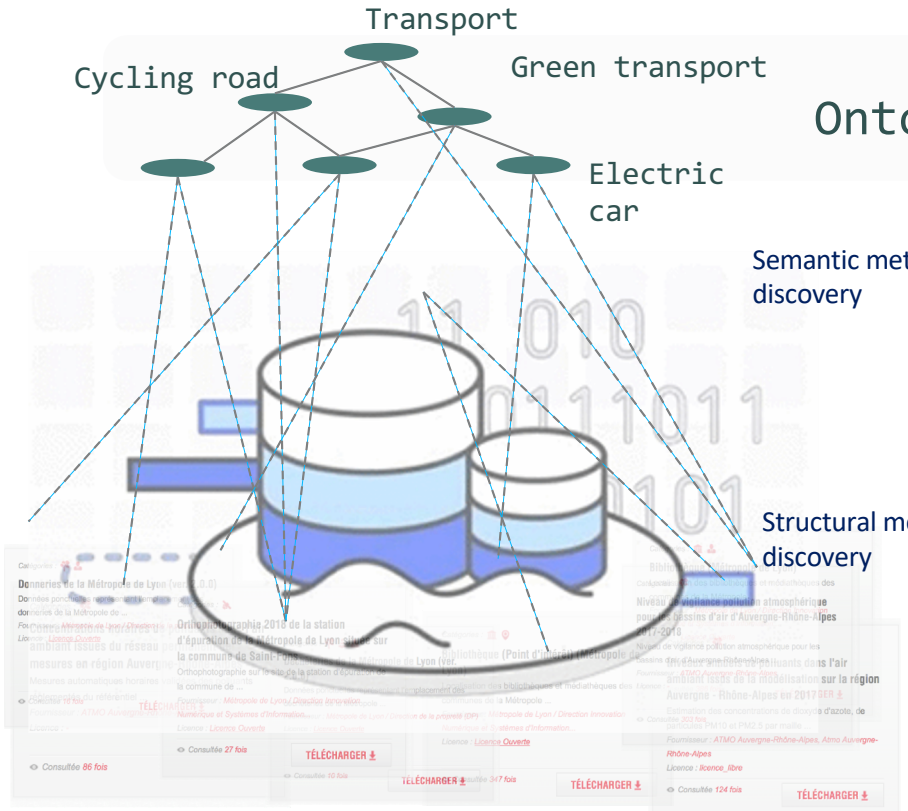
APPROACHES FOR EXTRACTING META-DATA

	Approach	Principle	Pros	Cons
Level I: simple tags as data is harvested	**Curation at Source** [Curry et al. 2010]	Simple **tagging** & **information extraction** directly from the sensor	Pre-processes data according to **samples**	No awareness of the whole data collection
Level 2: pivot vocabulary & semantics	**Master Data Management** [Weatherspoon et al. 2013]	Provide a **standard vocabulary** at the company scale	Standardizes the language used in data collections	Does not improve data structures
	Semantic linking Constance [Hai et al. 2016]	Identify **attributes** referring to similar topics in different data sets	Explores data collections	Does not improve data structures
	Data set clustering Goods [Halevy 2016]	Discover **similar data sets** using clustering	Creates an **synthesized representation** of several data collections	Difficult to define a **similarity criterion** to cluster data with **low quality** (missing & null values, types)
Level 3: manual and collaborative	**Crowdsourcing and Collaboration spaces** [Doan et al. 2005]	**Communities** produce, maintain & tag data (crowdsourcing)	Improves the **quality** of raw data	Manual & a lot of human resources

DATA CURATION ON THE LAKES

Approach	Principle	Pros	Cons
Meta data : Constance [Hai et al 2016], [Stonebraker et al 2013] Data wrangling [Terrizzano et al 2015]	Extraction of semantic meta-data or mapping items with concepts Query using SPARQL, regular expression	Possibility to express declarative queries; difficult to automate completely	No statistics, iterative data querying for exploring data
Data content description: descriptive statistics, processing according to data types, schema extraction	Explore data structure for extracting the schema Compute descriptive statistics functions for every element	Aggregated view of the content despite data types. Simple to visualize and scale if important volume.	Adapted for (semi)structured data, manual tagging for multimedia content
Curation: [CoreKG, Curry 2016, QoSMOS 2018, Tacit knowledge management 2017]	API with methods for preserving data collections. Querying and data fusion operations	Exploit Compass tool from MongoDB for providing a quantitative vision of data content	Data transformation from CSV to JSON. No semantic knowledge (terms, functional dependencies)

FISHING DATA IN THE LAKE



Ontology

Semantic meta-data discovery

Structural meta-data discovery

Constance

Are there data collections dealing with green transport?

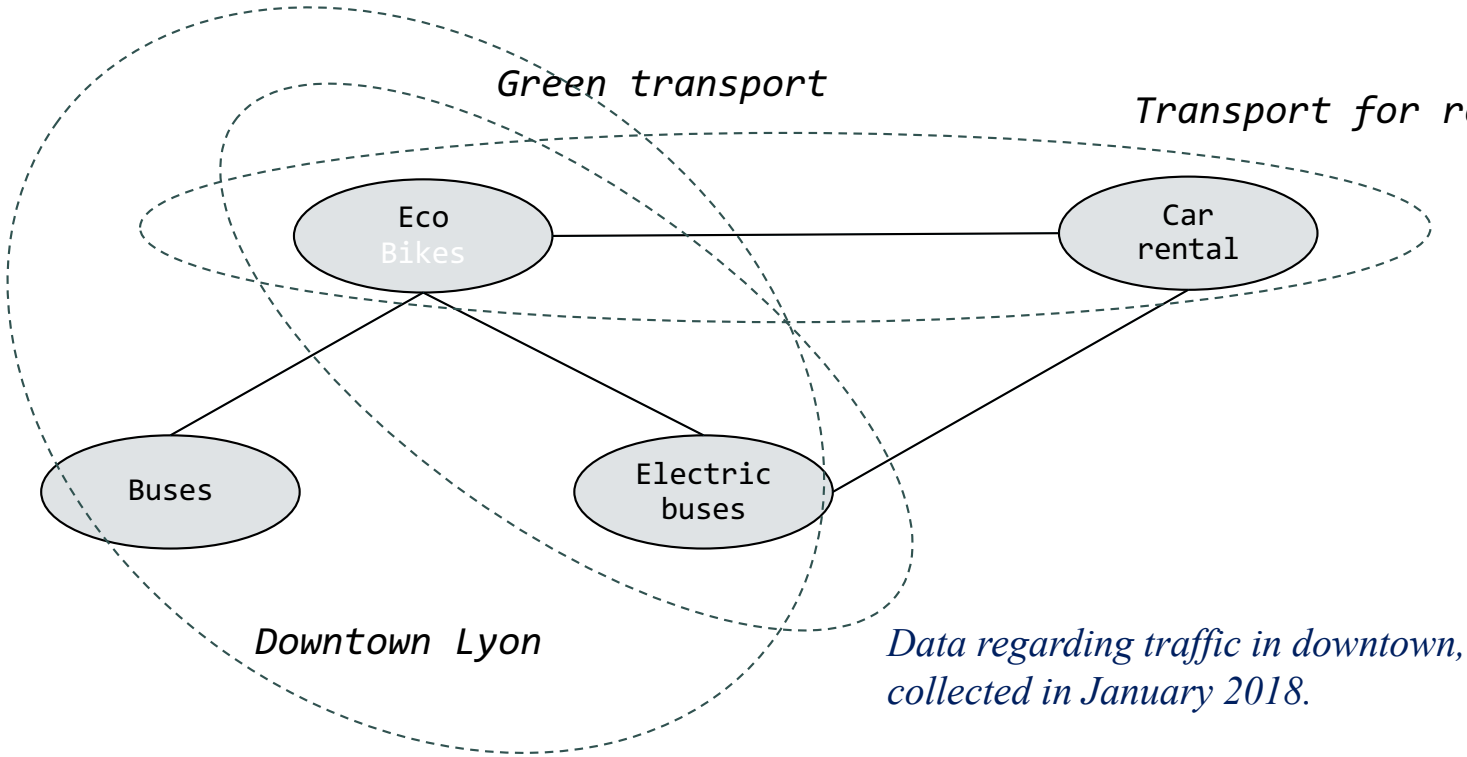
Which are available green transport modes in the city?

`exploreCollections( )`

Which are the attributes used for modelling transport modes

[Constance, SIGMOD '16]

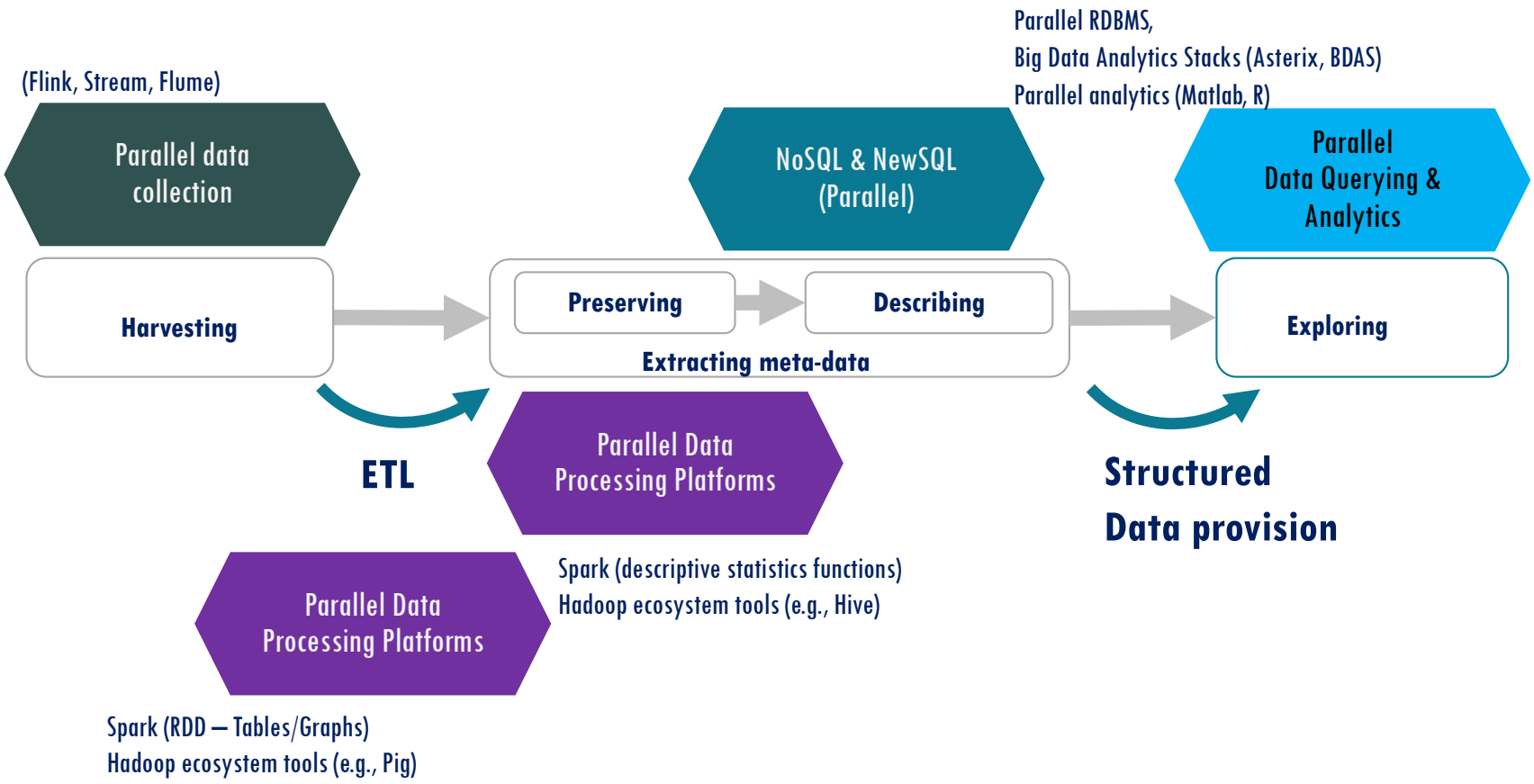
EXTRACTING META-DATA



Data regarding traffic in downtown, by date, traffic data collected in January 2018.

[Constance, SIGMOD '16]

PREPARING DATA COLLECTIONS



CLOUD COMPUTING AND BIG DATA

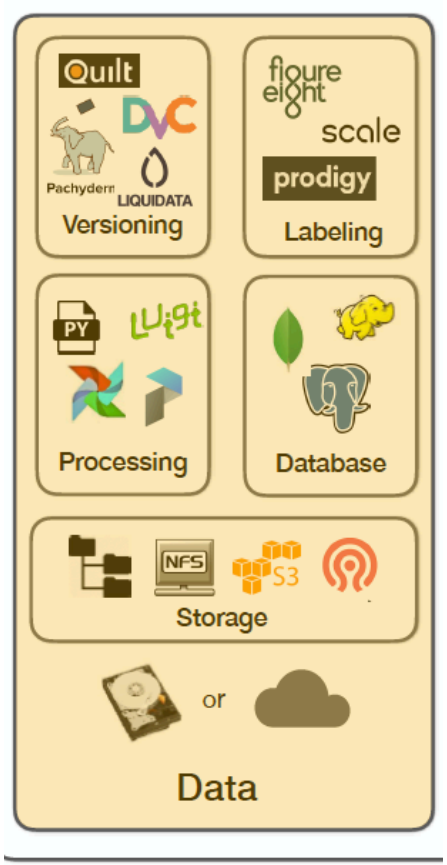
Ready to use environments for
Storing Big Data& Running greedy processing tasks

Platform as a Service (**PaaS**)
Database systems, frameworks

Infrastructure as a Service (**IaaS**)
CPU, RAM, Disk

Software as a Service (**SaaS**)
Full functional software

DATA LABS SERVICES



Kaggle.com

Google Colab

Azure Notebooks


HANDS ON