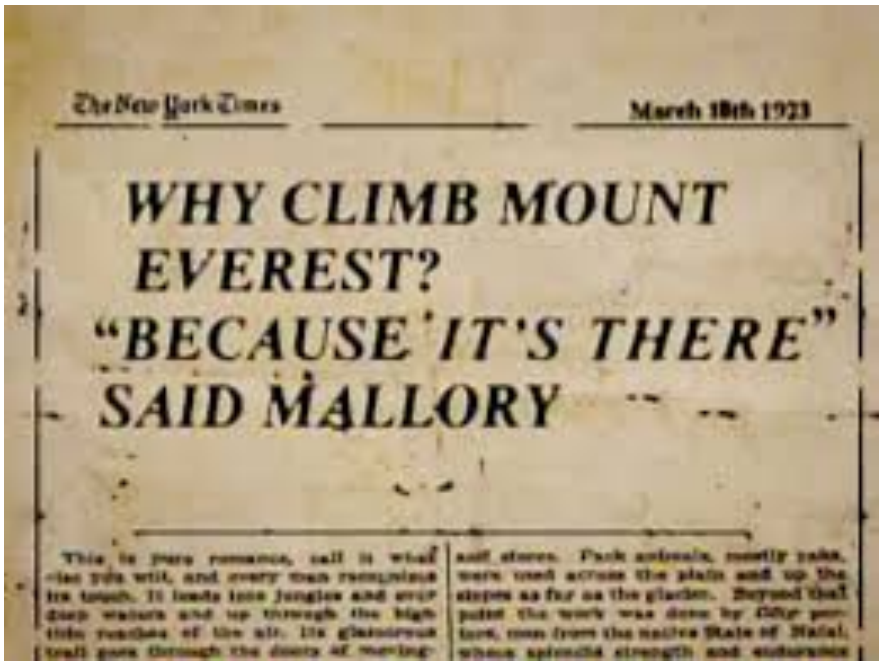
Big data and small data: the challenge is in the interpretation

Paolo Atzeni
Dipartimento di Ingegneria
Università Roma Tre

04/11/2020

Data & Big Data

- Why are we interested in them?



The New York Times — March 18th 1923

WHY CLIMB MOUNT EVEREST? "BECAUSE IT'S THERE" SAID MALLORY

George H.L. Mallory
(Mobberley, 1886 – Everest, 1924)

Data is not just here or there, but everywhere

- More and more data,
  - with more and more applications and exploitations

- "Data-driven society"
- "Data is the new gold"
- "Data is the new oil"

- But gold and oil are commodities (at least in some sense)
  - we could give a measure of quantity of gold or oil

Commodity

- *An economic good that has full or substantial fungibility: that is, the market treats instances of the good as equivalent or nearly so* (Wikipedia)

- We measure gold by ounces and oil by barrels
- Do we measure data in terms of ounces or barrels?

- Data is not a commodity
  - in this talk I will try to give arguments supporting this claim

Big Data

- Various definitions

"Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population." - Teradata Magazine article, 2011

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - The McKinsey Global Institute, 2012

"Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools." - Wikipedia, 2014

The four Vs of big data

- Volume
- Velocity
- Variety
- Veracity

Variety

- The major reason for which data is not a commodity:
  - each set of data requires specific operations to become useful
  - operations usually embed techniques for the interpretation of data

- A preliminary comment on interpretation

Interpretation of data

A picture taken in Gothenburg, Sweden

Interpretation of data

Torsdag: Thursday

7–17

White or black figures without brackets indicate weekdays, except weekdays before Sundays or public holidays.

(11-14)

White or black figures in brackets indicate weekdays, before Sundays or public holidays.

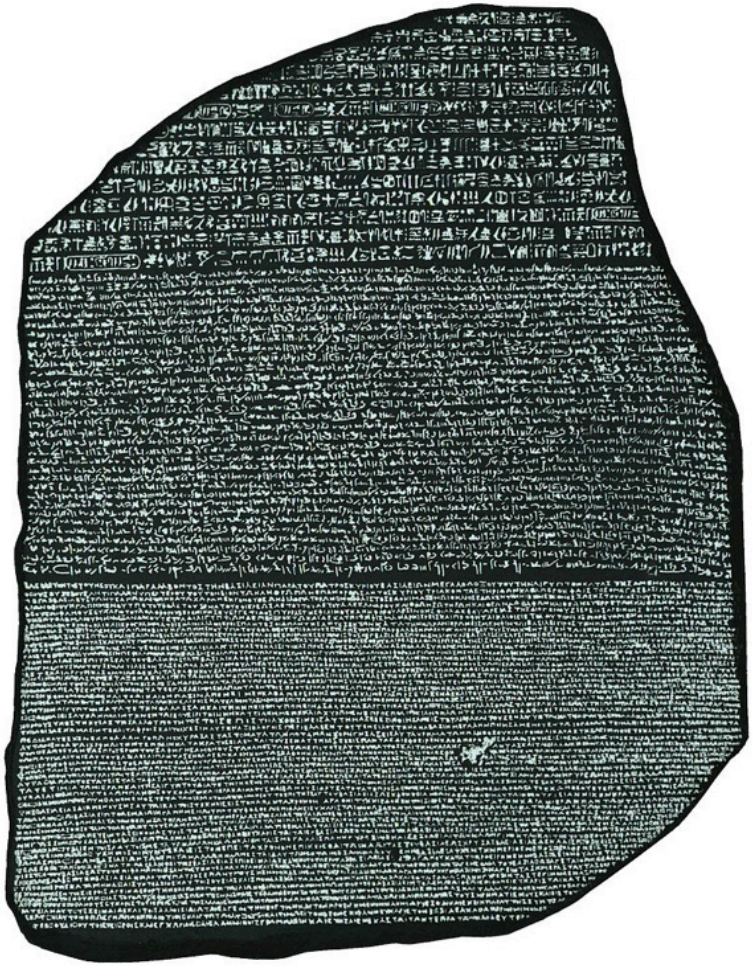
11-14

Red figures indicate Sundays or public holidays.

Another important aspect in the interpretation of data

- The number of students registered in a program
  - Many different definitions (and so criteria)
    - Registered by Nov 5 (the traditional closing date for registration in Italy)
    - Registered by today (on a given date)
    - Registered by a certain date, and for a number of years not greater than the length of a program
    - New students by a certain date
    - New students …, with no previous career
    - …

The Rosetta Stone, a famous case of support to interpretation



Variety

- The major reason for which data is not a commodity:
  - each set of data requires specific operations to become useful
  - operations usually embed techniques for the interpretation of data

- A preliminary comment on interpretation
- Let's go back to operations

Traditional operations on data (bases)

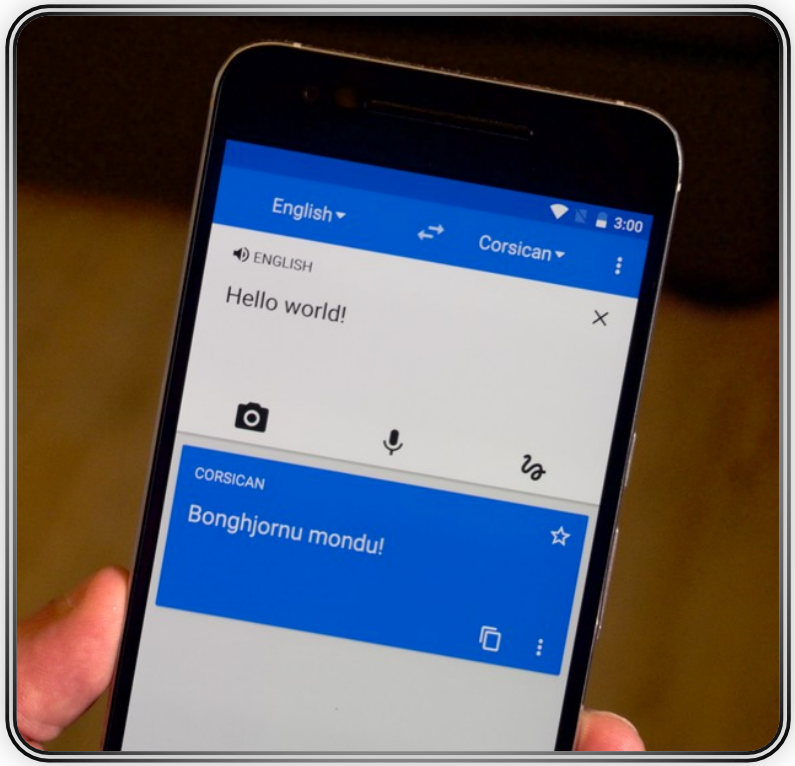
- Simple or complex, but usually well defined and reliable
  - the balance of an account is given by the initial balance plus the sum of deposit operations minus the sum of withdrawals
  - when we make a train reservation, we receive a confirmation only if there is an available seat (and then the seat is not given to others, unless we cancel)

Operations on Big Data

- Many kinds
  - traditional operations over big volumes of data
  - development and use of complex mathematical models, more or less deterministic, or based on simulation
  - application of machine learning techniques

Operations on Big Data

- On-line translation services
  - corpus of known translations of short sentences
  - comparison between portions of the text to be translated and sentences in a corpus.
  - regular update of the corpus.
- They work reasonably well and improve over time



Operations on Big Data

- Many kinds
  - traditional operations over big volumes
  - development and use of complex mathematical models, more or less deterministic, or based on simulation
  - application of machine learning techniques

- We often accept answers that are reasonably ok, but not perfect
  - translations
  - search engines
  - navigation systems
  - weather forecast

An issue with veracity: approximation and abstraction

- "*All models are wrong, but some are useful*"
  (attributed to George Box, probably over a preexisting concept)
- We use many forms of approximation:
  - For example, this is always the case for maps, which are the result of an abstraction process

A map with "the scale of a mile to the mile"

- What do *you* consider the *largest* map that would be really useful?
  - About six inches to the mile.
- Only *six inches.* We very soon got to six yards to the mile. Then we tried a *hundred* yards to the mile. And then came the grandest idea of all ! We actually made a map of the country, on the scale of a *mile to the mile*!
  - Have you used it much?
- It has never been spread out, yet, the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.

(Lewis Carroll, Sylvie and Bruno Concluded, Chapter XI, 1895, from Wikipedia)

Level of detail of maps

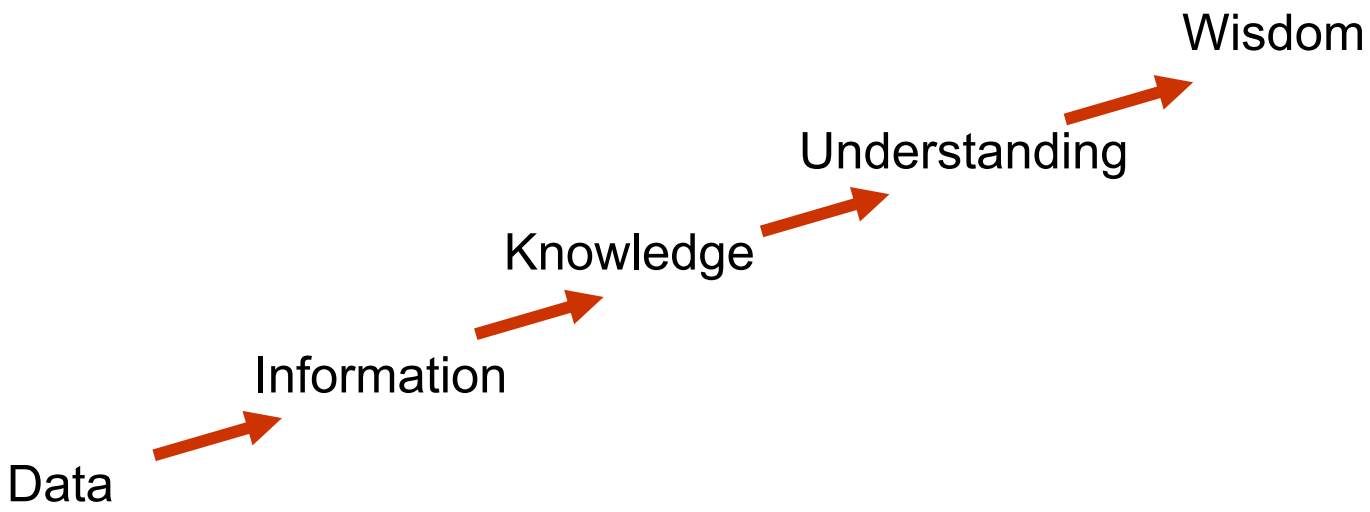
- When is a map really useful?
  - when it gives you the information you need
- If you are about to fly from Rome to Abu Dhabi, you want a very coarse map, to see the countries you fly over
- If you want to climb Mont Blanc, you need something more detailed, at a scale 1:25,000 or so
- If you are a constructor, you need a very detailed map of the area you have to work in

- Moral of the story:
  - we should use the level of detail we really need (or we can really achieve)
  - sometimes, we are allowed to be flexible in veracity

Data are almost always abstractions

- *Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom*.

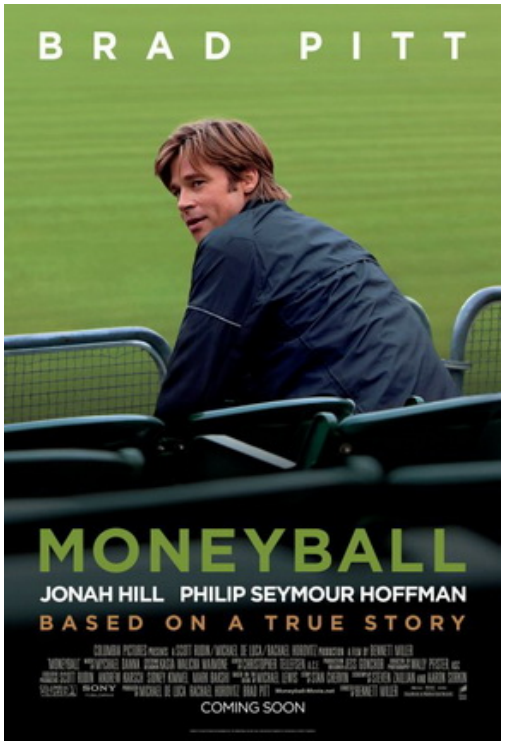
  Clifford Stoll



- Data requires interpretation to become something more
- How do we proceed?

Abstraction and simplification in the techniques

- We use data to build models:
  - a model describes a certain phenomenon and data are used as input for the model
  - but which data?
- Typical example:
  - weather forecast
- The simplest form of model:
  - indicators
- An interesting case study:
  - *Moneyball*, book and movie



Moneyball

(Apologies to Americans, this is baseball seen by an Italian)

- Baseball has been using statistics for decades:
  - players were valued on the basis of traditional indicators:
    - mainly batting average (BA) and runs-batted-in (RBI)
  - players would get salaries related to their BA and RBI
- Moneyball tells the story of a manager
  - who uses other indicators (mainly OBP, on-base-percentage, and SLG, slugging-percentage), which turn out to be more closely correlated to performance than BA and RBI
  - and therefore hires players with very good OBP and SLG and not so good BA and RBI and so he builds a very good team with a limited budget

Moneyball, moral of the story

- Data are useful, if we know how to use them
- If we use indicators, they have to be correlated to the phenomenon we are interested in
  - OBP and SLG are more correlated to winning games than BA and RBI

Another reflection on interpretation

- Data often (or always) come from different sources
- In big data terms
  - Volume means that we have many of them
  - Variety means that they are organized in different ways (and so they have to be interpreted in different ways)

"A ten-year goal for database research"

- The “Asilomar report”
  (Bernstein et al. Sigmod Record 1999 www.acm.org/sigmod):
  - ***The information utility:***
    make it easy for everyone to store, organize, access, and analyze the majority of human information online

- Many interesting results have been obtained, but …
- …integration, translation, exchange of data remain difficult …
- … 20 years have gone, we are late
- … and we also have Big Data

A more recent, relevant concept

- A **data lake** is a system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of data including raw copies of source system data, sensor data, social data etc and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases, semi-structured data, unstructured data and binary data.

  Wikipedia, 2020

Typical approaches to integration

- Three steps (Dong e Srivastava, 2015)
  - Schema alignment: resolve semantic ambiguities
    - mediated schema
    - attribute matching
    - schema mapping
  - Record linkage: find records that refer to the same real world entities
  - Data fusion: put together data from the various sources

- Each of the four Vs adds complexity and difficulties

Many research problems

- The complexity of the issues leads to a variety of research problems and solution
- A few ideas on two research projects my group is involved in
  - Instance level attribute alignment

    (joint work with F. Piai, P. Merialdo, D. Srivastava)
  - Schema mapping reuse

    (joint work with L. Bellomarini, P. Papotti, R. Torlone)

Instance level attribute alignment

(joint work with F. Piai, P. Merialdo, D. Srivastava)

- Input
  - Product specifications from many sources
  - Partial record linkage
- Goal
  - Match attributes with "equivalent" semantics and build integrated records with common structure

Source: Ebay
Product ID: Iphone6

Brand	Apple
Resolution	10 MP
Front camera	5 MP

Source: Ebay
Product ID: Zenfone2

Brand	Asus
Resolution	8 MP
Memory	16 Giga

Source: Amazon
Product ID: Iphone6

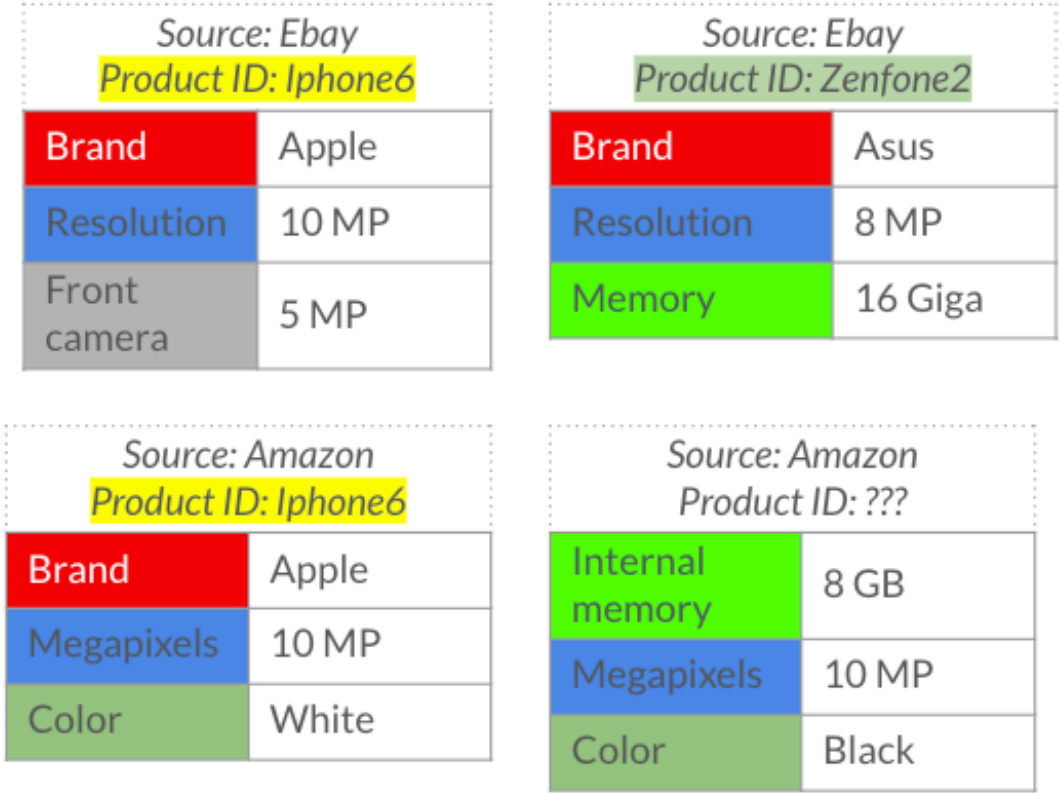
Brand	Apple
Megapixels	10 MP
Color	White

Source: Amazon
Product ID: ???

Internal memory	8 GB
Megapixels	10 MP
Color	Black

Instance level attribute alignment

- Input
  - Product specifications from many sources
  - Partial record linkage
- Goal
  - Match attributes with "equivalent" semantics and build integrated records with common structure



Instance level attribute alignment, issues (even within sources), 1

- Synonyms

Source: Ebay

Memory	8 GB
Resolution	10 MP
Front camera	5 MP

Source: Ebay

Int memory	8 GB
Resolution	10 MP
Front camera	5 MP

- Homonyms (even intrasource)

Source: Ebay

Battery	Ni-mh
Resolution	10 MP
Front camera	5 MP

Source: Ebay

Battery	Duracell
Resolution	10 MP
Front camera	5 MP

Instance level attribute alignment, issues (even within sources), 2

- Different organization

Source: Ebay

Features	Black, 16 GB
Front camera	5 MP
Battery weight	50 g
Weight (w/o battery)	100 g

Source: Ebay

Other	Color black, Front 5 MP
Memory	16 GB
Weight	150 g

- Different value templates

Source: Ebay

Memory (GB)	16
Resolution	10 MP
Front camera	5 MP

Source: Ebay

Internal Memory	16 Gigabytes
Resolution	10 MP
Front camera	5 MP

The approach

- Exploit input linkage and redundancies
- Steps
  - Source attribute alignment, as much as possible ("difficult" cases stay isolated)
  - Dictionary of values (from aligned attributes)
  - Creation of "virtual" attributes by matching values in the dictionary and (portions of) values in isolated attributes
  - Iteration over the previous steps

Source: Ebay
Product ID: Galaxy S5

Product Name	Samsung Galaxy S5
Product Name#Brand	*Samsung*
...	...

Source: Amazon
Product ID: Huawei P10

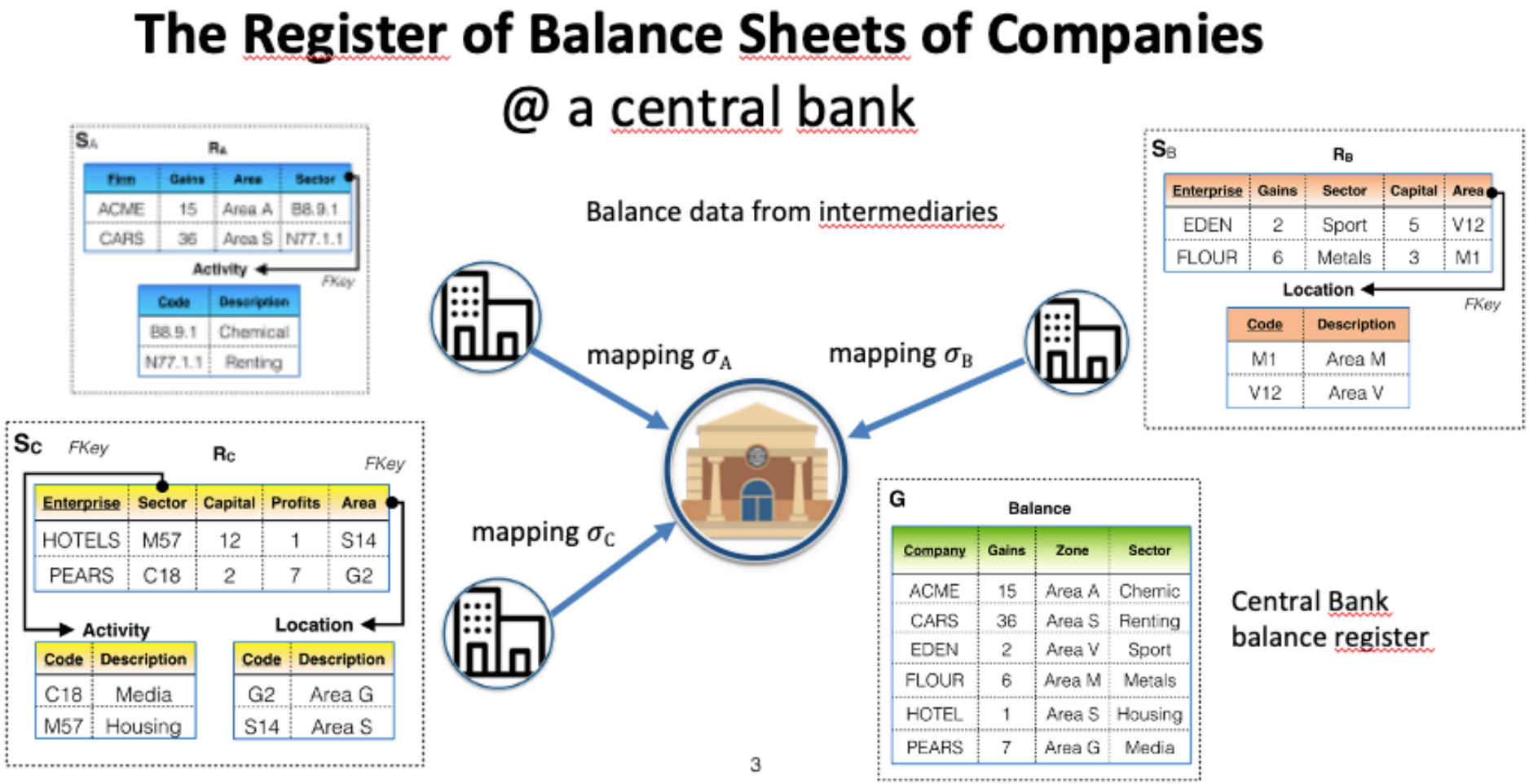
Features	16 GB, black
Features#color	*black*
...	...

Schema mapping reuse

(joint work with L. Bellomarini, P. Papotti, R. Torlone)

- Schema mappings: a central tool for many data management problems:
- Data exchange
- Data integration
- Schema and data translation
- Schema transformation
- Round-trip engineering
- …

An application with many schema mappings

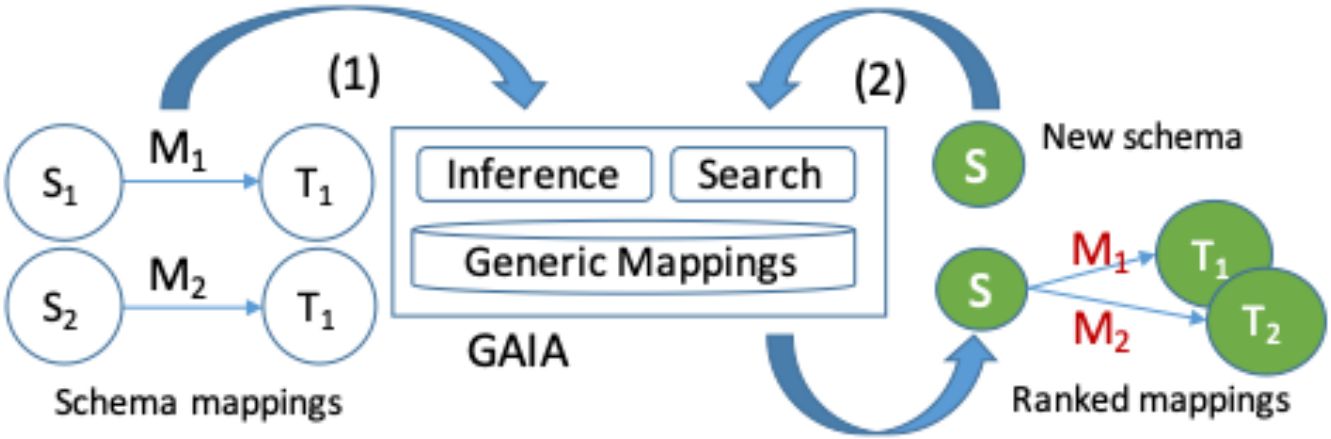


Many source schemas

- They are different yet similar
- Mappings are also different yet similar
- How to exploit similarities?
  - Describe mappings in a way that is independent of schema specificities while capturing its meaning and so foster reuse.

The approach

- Collect mappings in a repository
- Describe mappings in terms of "meta-mappings"
- When a new mapping is needed
  - Search in the repository for a "suitable" meta-mapping
  - Build a mapping from the meta-mapping
  - If needed, manually adapt the mapping



Conclusions

- Data are important and we need to exploit them
- The major issues, in small data as well as in big data, are in the proper interpretation and use of data
- In the new big data settings, many interesting research problems arise, and we have seen a couple of them

Thank you!