

Data extraction and integration, the issue is interpretation

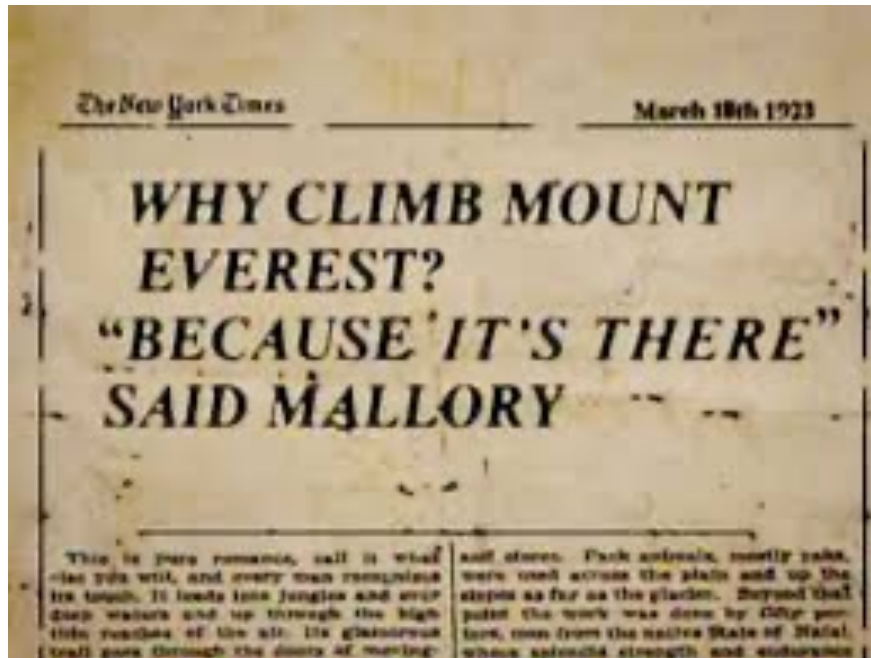
Paolo Atzeni
Dipartimento di Ingegneria
Università Roma Tre



02/07/2021

Data & Big Data

- Why are we interested in them?



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

02/07/2021

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.

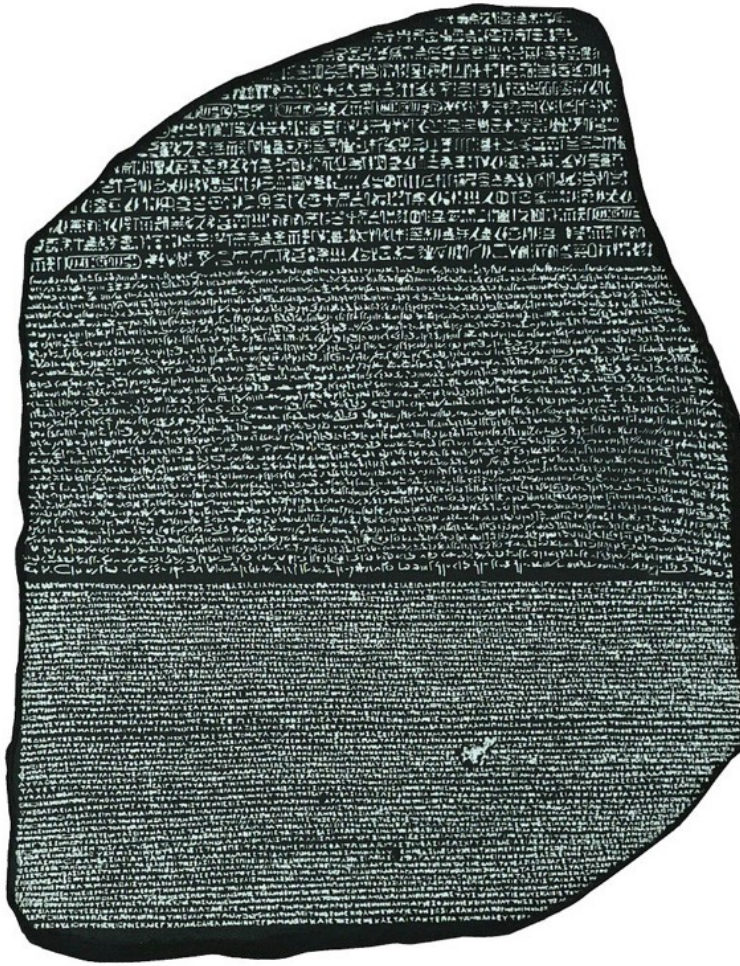
11-14

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)

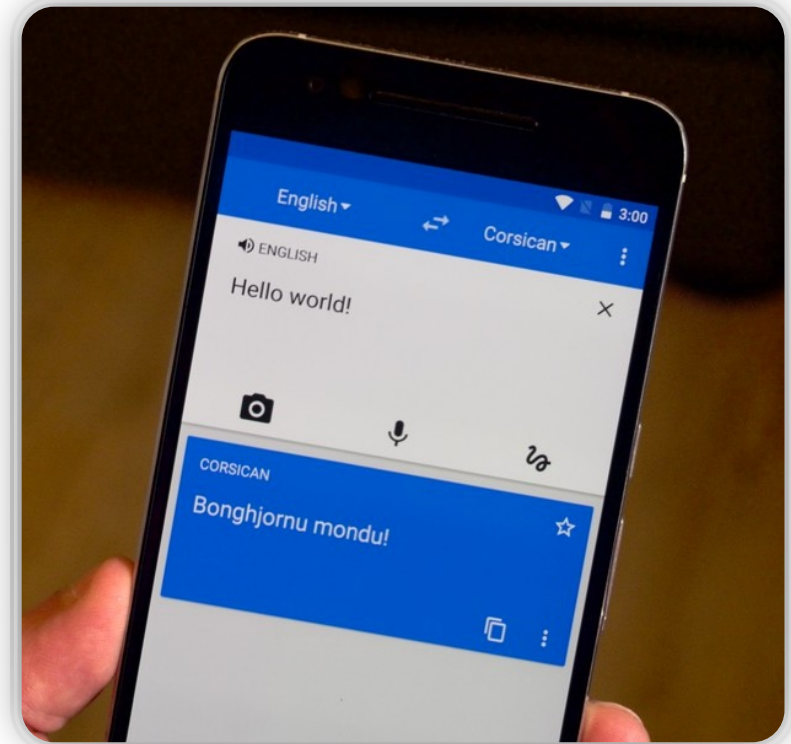
- Simple or complex, but usually well defined and reliable
 - the balance of an account is given by the initial balance plus the sum of deposit operations minus the sum of withdrawals
 - when we make a train reservation, we receive a confirmation only if there is an available seat (and then the seat 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

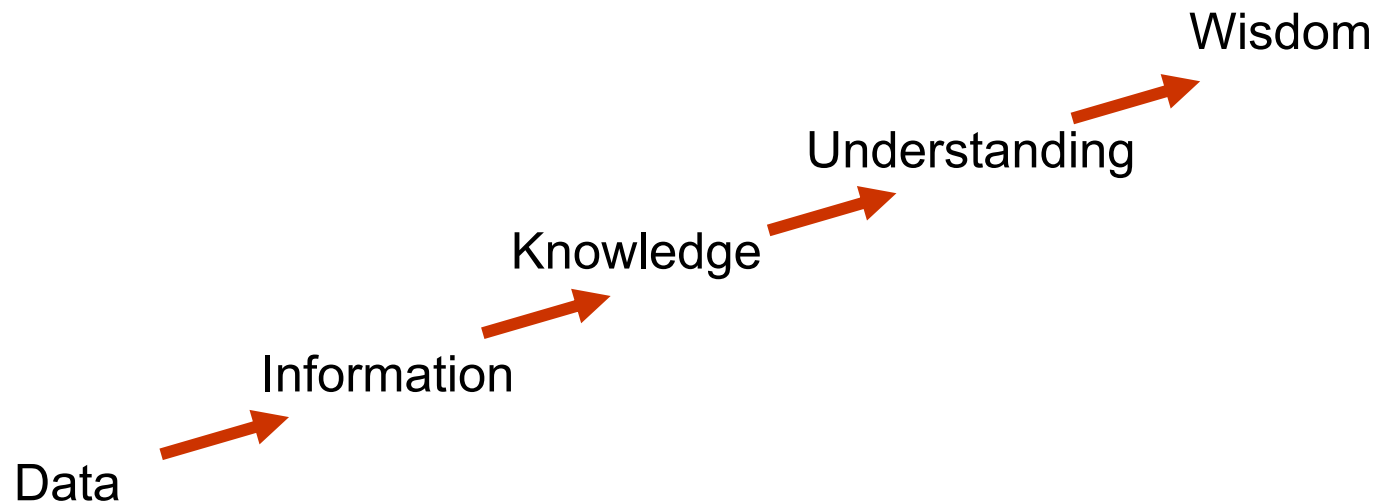
- When is a map really useful?
 - when it gives you the information you need
- If you are about to fly from Rome to Toulouse, you want a very coarse map, to see the countries you fly over
- If you want to climb Mont Blanc, you need something more detailed, at a scale 1:25,000 or so
- If you are a constructor,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

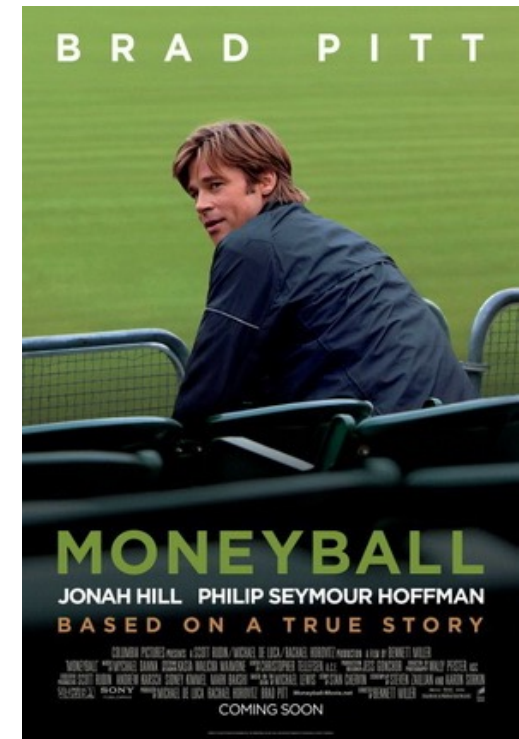
Clifford Stoll



- 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

Typical approaches to integration

- Three steps (Dong and 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 three research projects my group is involved in
 - Extracting and Archiving Web Data
(joint work with V. Cetoirelli, V. Crescenzi, F. Milicchio)
 - 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)

Extracting and Archiving Web Data

(joint work with V. Cetorelli, V. Crescenzi, F. Milicchio)

- Web Data Extraction
 - The Web is the largest KB ever existed
 - Fully automated solutions are necessary to scale
- Application: Data Archiving
 - Several organizations store snapshots of the Web (<https://web.archive.org/>)
 - Cost of storage is the major limiting factor

Extracting Web Data

- Input
 - A set of webpages sharing a common template
- Goal
 - Extract data of interest
 - Efficiently archive them
- HTML pages lack of a rigid structure
- HTML Ambiguity: occurrences of the same token play different roles:
 - **Price** and **** can be considered both as template or content depending on the context

<pre><HTML><BODY> Title <DIV>Price</DIV> Description <DIV>The Price of Everything</DIV>
 35.22 A US documentary </BODY></HTML></pre>	<pre><HTML><BODY> Title <DIV>Price</DIV> Description <DIV>Star Wars</DIV>
 9.98 A saga directed by G. Lucas </BODY></HTML></pre>
--	---

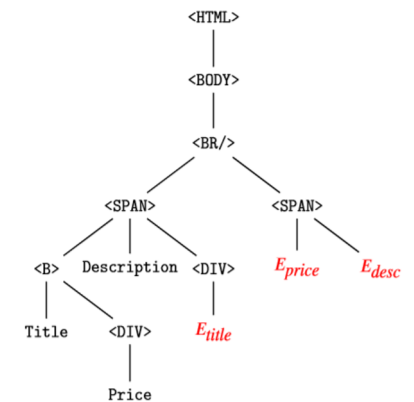
The Approach: Landmark Grammars

- HTML is rich and varied
 - There is at least one non ambiguous token
- Steps
 - Each non ambiguous occurrence can be used to split a page into several *sub-regions*
 - The search can be repeated in each region
 - Tokens occurring in one region are no longer confused with those in others
 - Repeat until the page's template is modelled

<pre><HTML><BODY> Title <DIV>Price</DIV> Description <DIV>The Price of Everything</DIV>
 35.22 A US documentary </BODY></HTML></pre>	<pre><HTML><BODY> Title <DIV>Price</DIV> Description <DIV>Star Wars</DIV>
 9.98 A saga directed by G. Lucas </BODY></HTML></pre>
--	---

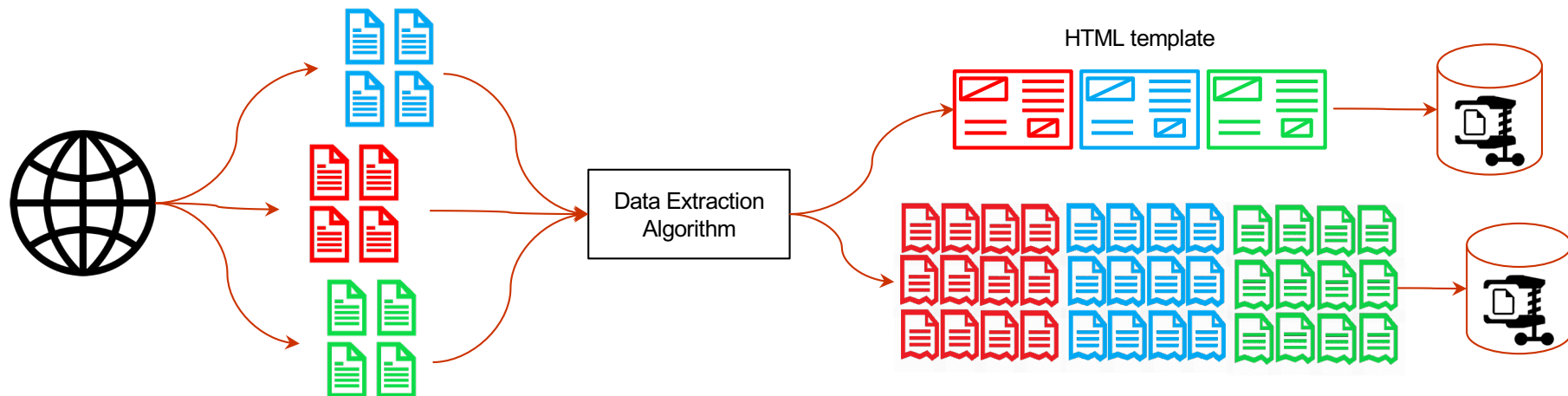


```
<HTML><BODY>
<B>Title</B>
<DIV>Price</DIV>
<SPAN>Description</SPAN>
<DIV>Etitle</DIV>
<BR/>
<SPAN>Eprice</SPAN>
Edesc
</BODY></HTML>
```



An Application: Data Archiving

- Web Data extraction from pages sharing a common HTML template
 - Separate HTML from extracted data that vary
 - Save it only once
- Store homogenous extracted data together
 - High compression ratio
 - Cheaper then saving the html source code



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

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, issues (even within sources), 1

- Synonyms
- Homonyms (even intrasource)

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

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		Source: Amazon	
Product ID: Galaxy S5		Product ID: Huawei P10	
Product Name	Samsung Galaxy S5	Features	16 GB, black
Product Name#Brand	Samsung	Features#color	black
...

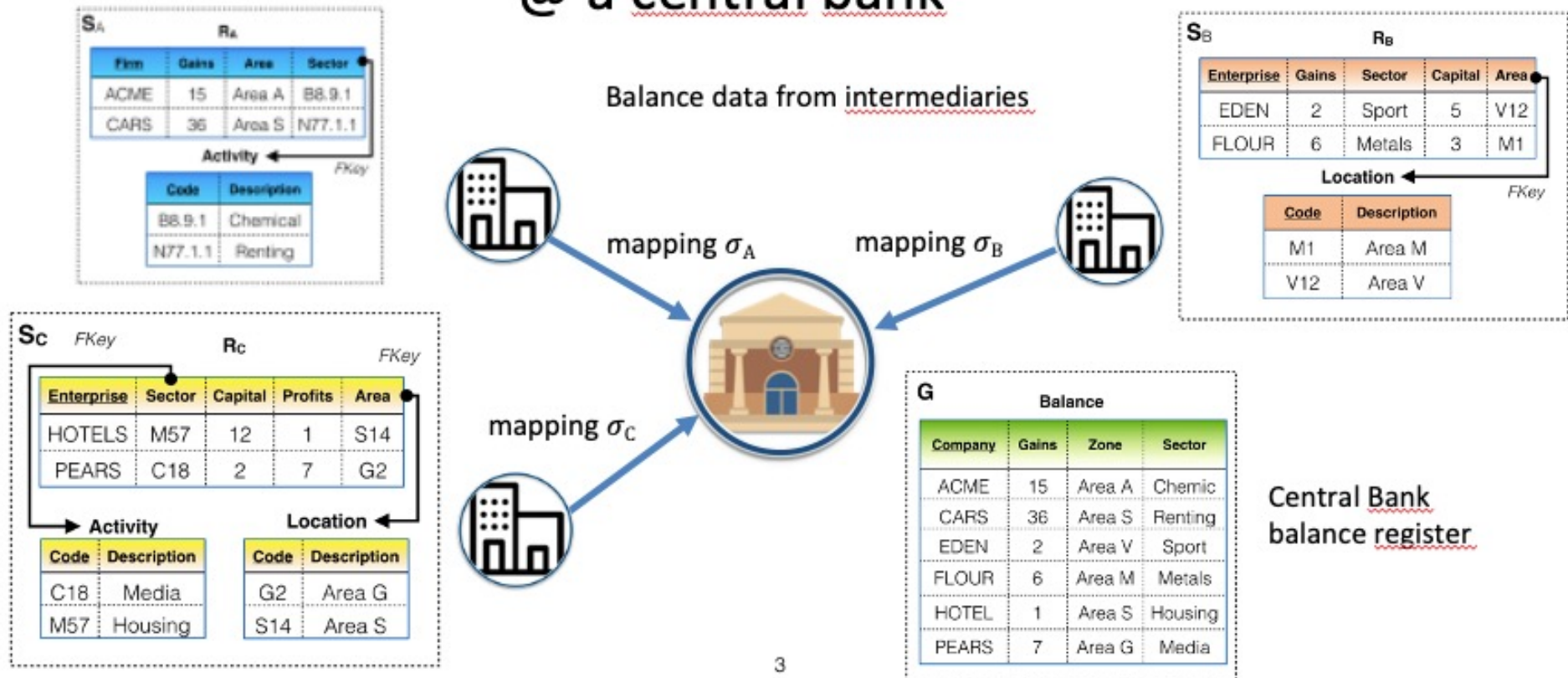
Schema mapping reuse

(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

The Register of Balance Sheets of Companies @ a central bank



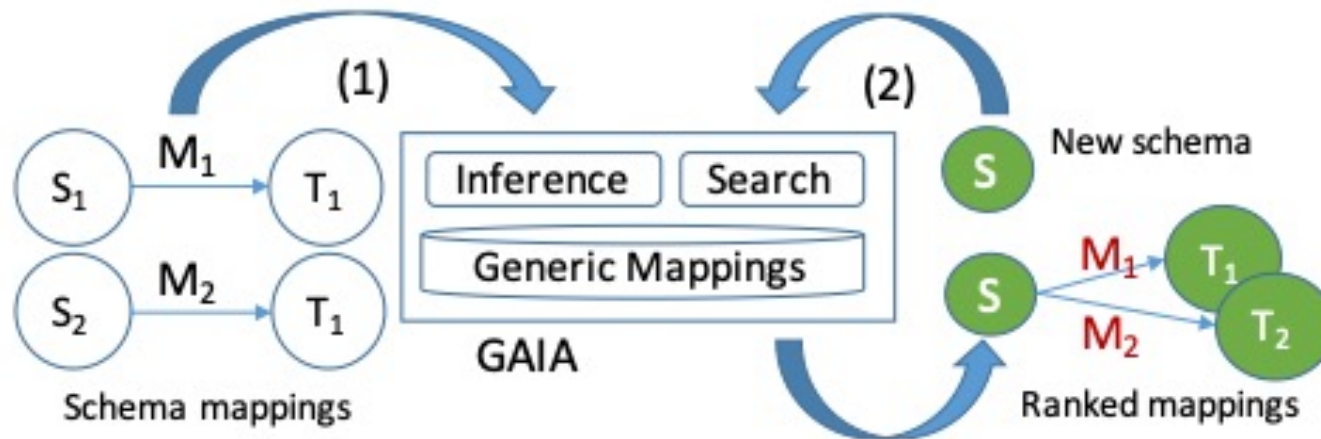
3

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!