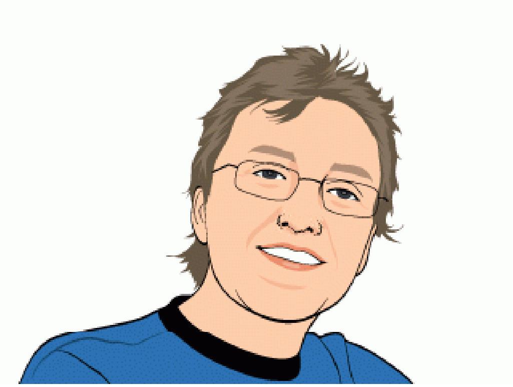
Data Science for Historical Inquiries



Donato Malerba

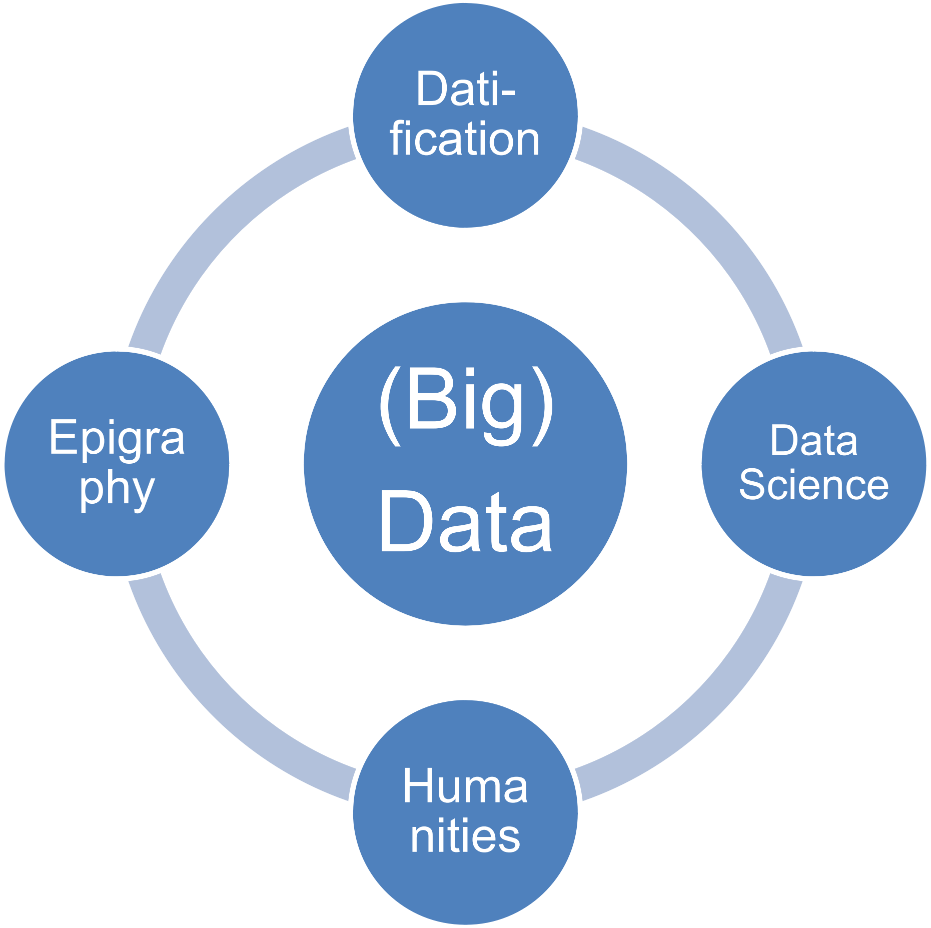
Computer Science Dept. – University of Bari
Big Data Lab - CINI

Off the Beaten track
Bari, 25th September 2015

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

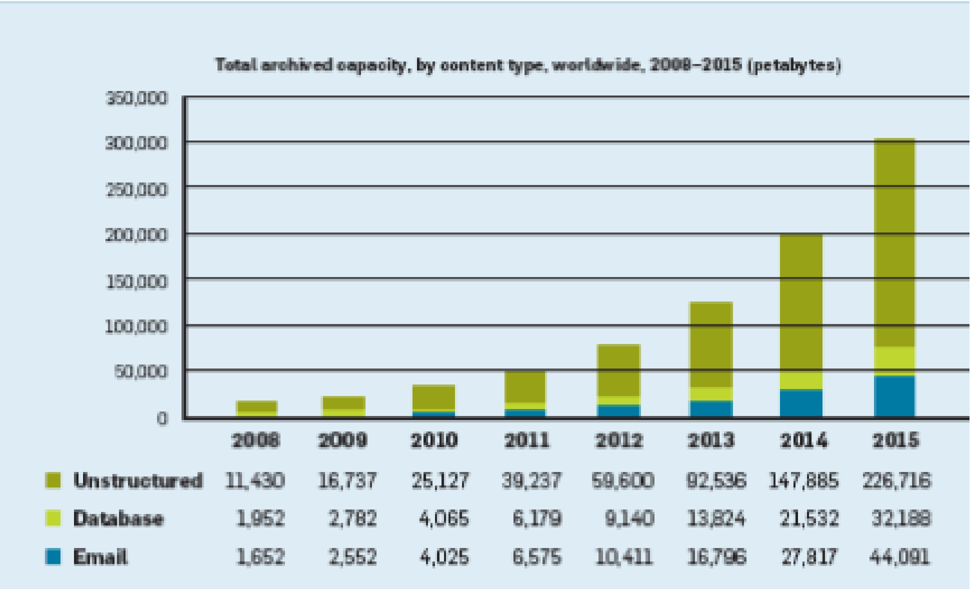
cini
consorzio
interuniversitario
nazionale
per l'informatica

Overview



An estimated growth …

- Yearly increase of data volume: 30-40%
- The data volume doubles every 2.5 years

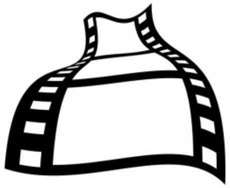

Total archived capacity, by content type, worldwide, 2008–2015 (petabytes)

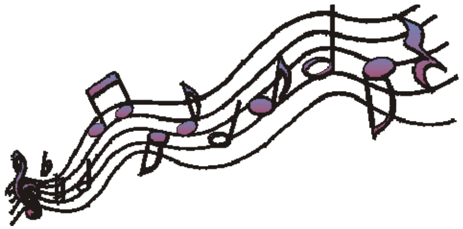
	2008	2009	2010	2011	2012	2013	2014	2015
■ Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
■ Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
■ Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

Exponential growth
•2,7 ZB (10^{21} bytes) in 2012
•35 ZB in 2020

Datification

- Neologism referring to the conversion into digital format (data) of:
- Movies, music, books, etc. (previously on films, paper, vinyl, etc.)
- Phone conversations, mail, radio and TV programs





Datification

- *Facebook*: our social networks are data,
- *Twitter*: our sentiments are data,
- *LinkedIn*: our professional experiences are data

Devices generate data …

- **Internet of Things** (IoT): each thing is uniquely identifiable and is able to interoperate within the existing Internet infrastructure. Things have an active role and exhibits “intelligent” behaviors.

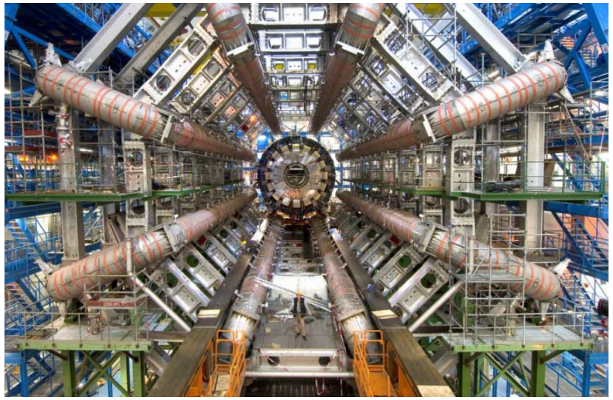
Devices generate data …

- Due to our symbiosys with digital technologies we are becoming “living sensors”

- 7 billion people and 6,8 billion mobile phones
- «***We are like a digital Tom Thumb, who leaves (digital) crumbs along the way***».

Science generates data …

- Large volumes of data are now generated in:
  - **Genomics** ➔ next generation sequencing
  - **Astrophysics** ➔ astronomical observatories
  - **Particle Physics** ➔ Large Hadron Collider
  - **Neuroscience** ➔ Human Brain project
  - …



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

Companies generate data …

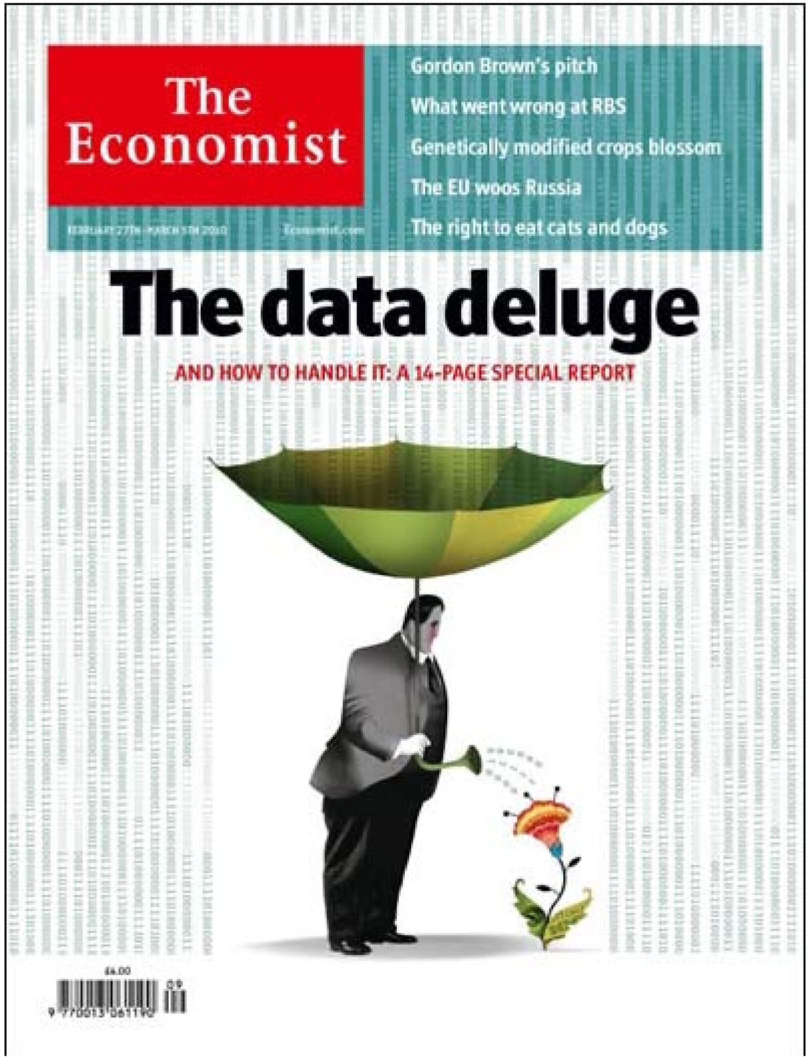
- Nowaday every big business is a digital business:
  - **Alibaba** the largest shop in the world without warehouse.
  - **Uber** the largest company in the world without cars.
  - **Airbnb** the largest network for lodging without a single hotel.

Companies generate data …

- Purchase orders, invoices, dispatches, …
- Data collected in information systems are considered an *(intangible) asset.*
- *Facebook:* (tangible) *assets* for 6,3 billion$ but at its stock market flotation its value was 104 billion$.
- Data are an intangible asset, but only **5‰** of business data are actually processed

The data deluge

February 2010


The Economist
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
Gordon Brown's pitch
What went wrong at RBS
Genetically modified crops blossom
The EU woos Russia
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Big Data

- Data collection has gotten to the point where the volume, the velocity and the variety of incoming data demands from **non conventional tools** to extract and process information in (near) rela time.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l'informatica

Big Data: a revolution?

- *The true revolution is not in technologies we use to process data, but in the way we use data.*

- *The larger scale of processed data paves the way to new potential applications which would not be possible otherwise*

Data Science

- Data science is a body of principles and techniques for applying data-intensive analysis to investigate phenomena, acquire new knowledge, and correct and integrate previous knowledge with measures of correctness, completeness, and efficiency of the derived results.

Big Data vs. Data Science

- Data Science vs. Big Data
  - Data Science does not always need Big Data, however the steady increase of data makes Big Data an important issue for Data Science.

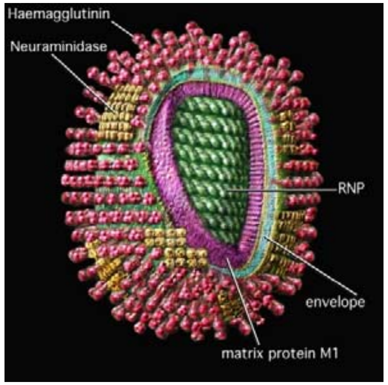
Big Data: *an example*

- Spring 2009: pandemic influenza caused by the virus A (H1N1) emerged, beginning in Mexico and quickly spreading to the United States and around the world

  USA Centers for Disease Control and Prevention:

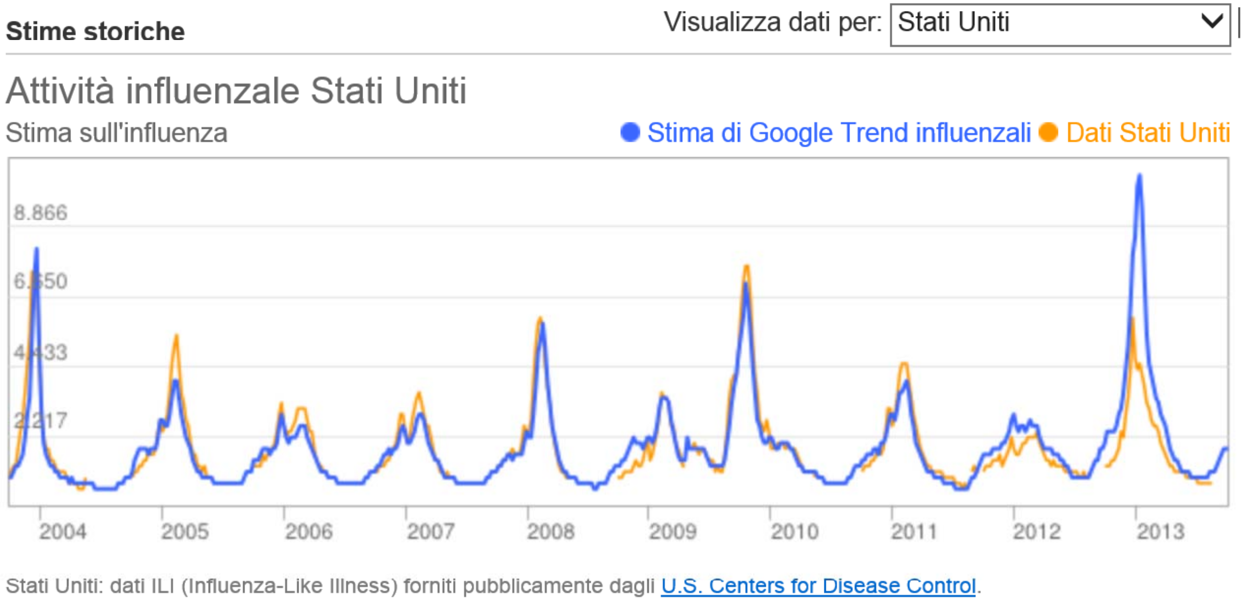
  state and local health departments collected data on influenza-like cases + epidemiological models

  ➔ a lag of two weeks



Big Data: *An example*

Google Flu Trends: uses anonymized, aggregated internet search activity to provide near-real time estimates of influenza activity



Stime storiche

Visualizza dati per: Stati Uniti

Attività influenzale Stati Uniti

Stima sull'influenza

● Stima di Google Trend influenzali ● Dati Stati Uniti

Stati Uniti: dati ILI (Influenza-Like Illness) forniti pubblicamente dagli U.S. Centers for Disease Control.

Detecting influenza epidemics using search engine query data. ***Nature 457, 1012-1014 (19 February 2009)***

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

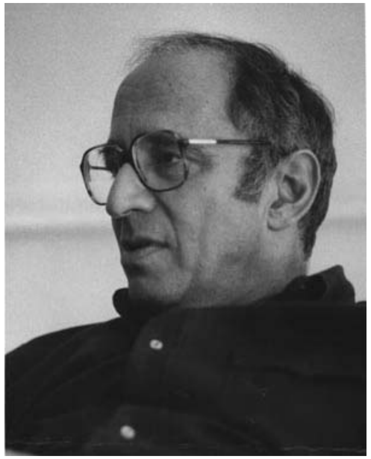
cini consorzio interuniversitario nazionale per l'informatica

Big Data: *A change of perspective*

- Three Big changes in the way data are analyzed:
  1. Process all available data (the obsolescence of sampling)
  2. Accept increased measurement error in return for more data
  3. Move away from the age-old search for causality (focus shift, from causation to correlation)

Big Data for a *data-driven science*

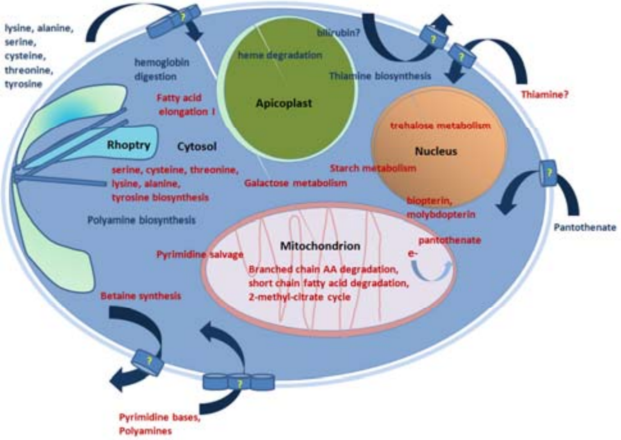
- Big Data are causing a radical change in scientific paradigms.
- Thomas Kuhn: *a scientific revolution is a paradigm shift*



- Two traditional paradigms in science:
  - Experimentation
  - Theory

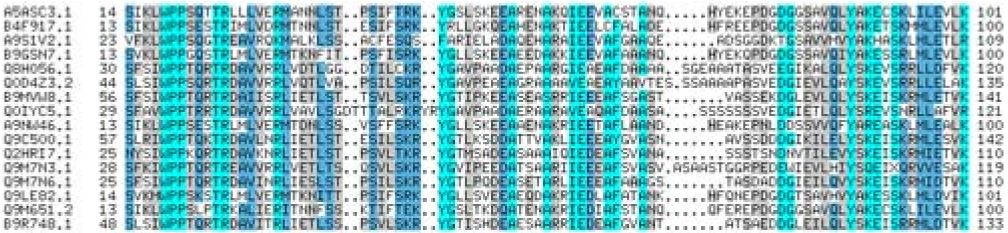
Two new paradigms

- Computational simulation (or third paradigm): Ken Wilson, Nobel laureate in physics, 1982.
- At the base of the *computational sciences*
  - *computational biology*, simulate the behaviour of biological systems, *metabolic pathways* or a cell or the way a protein is produced.



Two new paradigms

- Data intensive knowledge discovery (or fourth paradigm): Jim Gray, computer scientist.
- At the base of the *science informatics*
  - *bioinformatics*, an interdisciplinary field that develops methods and software tools for understanding biological data.



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

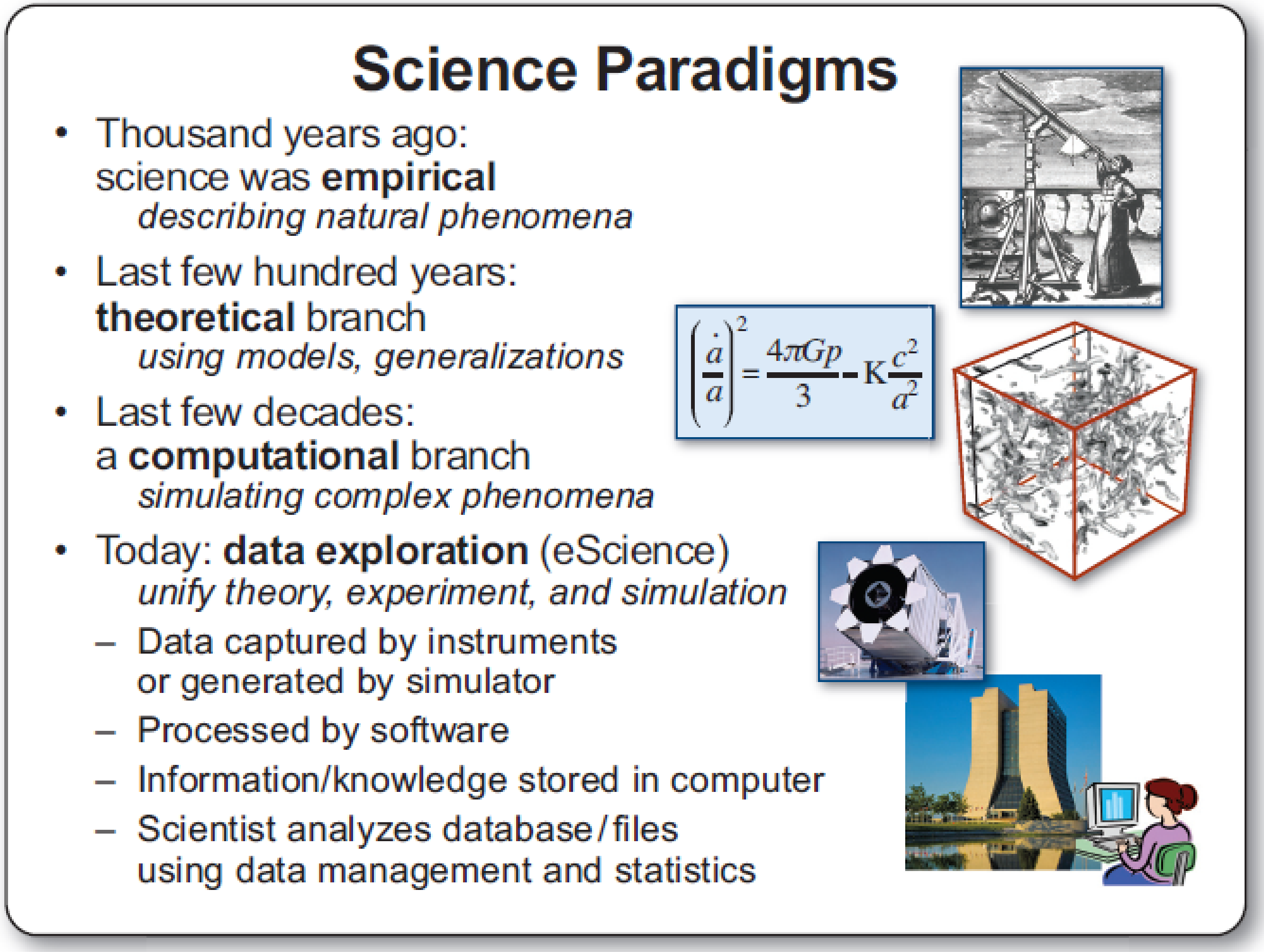
cini consorzio interuniversitario nazionale per l'informatica

Data-intensive Knowledge Discovery

Steps:

- Data capture
- Data curation
- Data analysis
- Result publishing

The Science Paradigms



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

What's the impact of these trends on Humanities?

- Large-scale digitization projects have vastly increased the quantity of cultural heritage material in several humanistic areas.
- Humanities scholars are increasingly incorporating computational tools and methods in all phases of their research.
- Large investments in digital infrastructures supported by funding agencies.

What’s the impact of these trends on Humanities?

- Most researchers in humanities explore data manually, using their knowledge and expertise to extract the information they deem relevant.

What's the impact of these trends on Humanities?

Example:

Els Witte (2006) studied the image of the nation in the Belgian Revolution (1828-1847) by manually browsing six newspapers and selecting 350 articles that expressed an opinion.

- This is a nice example of a new perspective on the study of "public opinion" or collective "mentalities"

What’s the impact of these trends on Humanities?

- However, there are corpora for historical research that are simply too large to be examined in their entirety and to be inspected manually
- Sampling reduce the amount of data to manageable proportions, but the manual inspection strongly limits the subsequent analysis.

What’s the impact of Data Science on Humanities?

- One of the promise of the convergence of data science and humanities is that it will enable us to investigate much larger quantities of data.
- We are now entering a new phase in which historians are able to analyze massive volumes of data in various formats (records, texts, images, …)

What's the impact of these trends on humanities?

- New techniques of large-scale data analysis allow historians to manage big data sets that were impossible to manage earlier.
- Data science can reduce the effort required of humanities researchers to obtain useful information from large repositories of digitized cultural heritage (e.g., medieval manuscripts)

What’s the impact of these trends on humanities?

Advantages of adapting Data Science methodologies to humanities:

- Reproducibility of results
- Promotion of collaborative work (in contrast to current research which is predominantly individualistic)
- New research questions

What’s the impact of Data Science on Epigraphy?

Several epigraphic repositories currently contain large corpora of pictures and the textual document representation thereof, which have been stored and annotated on several levels of interest.

http://www.eagle-network.eu,

http://edh-www.adw.uni-heidelberg.de,

http://www.edb.uniba.it,

http://eda-bea.es,

http://www.epigraphik.uni-hamburg.de,

http://usepigraphy.brown.edu,

http://www.edr-edr.it

What's the impact of Data Science on Epigraphy?

A vital question:

how, if at all, should the work of epigraphists adapt to the presence of orders of magnitude more potential source material?

What’s the impact of Data Science on Epigraphy?

From the perspective of traditional research, little has changed: the fact that most of recorded human intellectual output is now accessible does not increase the ability of an epigraphist to read it.

What's the impact of Data Science on Epigraphy?

Evidently, any fundamental advances must come from the fact that this material is now available for computational processing.

What’s the impact of Data Science on Epigraphy?

We argue that a new and still unexplored frontier of digital epigraphy projects is that of enabling automatic analysis of information currently stored in epigraphic repositories in order to extract implicit, previously unknown and potentially useful knowledge from them.

What’s the impact of Data Science on Epigraphy?

The application of Data Science practices may help to reveal interesting relationships among

- linguistic style,
- positioning,
- dating of inscriptions,
- …

thus creating new links between different pieces of data.

What’s the impact of Data Science on Epigraphy?

In general, this approach can help to organize large collections of inscriptions, introduce younger scholars to the field of epigraphy, and identify anomalies that can be later explored using more traditional methods, as already done in computational historiography (Mimno, 2012).

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

consorzio interuniversitario nazionale per l’informatica

A concrete example: EDB

45.000 Christian inscriptions of Rome, including inscriptions published in the Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series (ICVR) editions.

EDB
Epigraphic Database Bari
Documenti epigrafici romani di committenza cristiana - Secoli III - VIII
QVAERITVR·INVENTVS·COLITVR·FOVET·OMNIA·PRESTAT
EXPRESSIT·DAMASVS·MERITVM·VENERARE·SEPVLCRVM

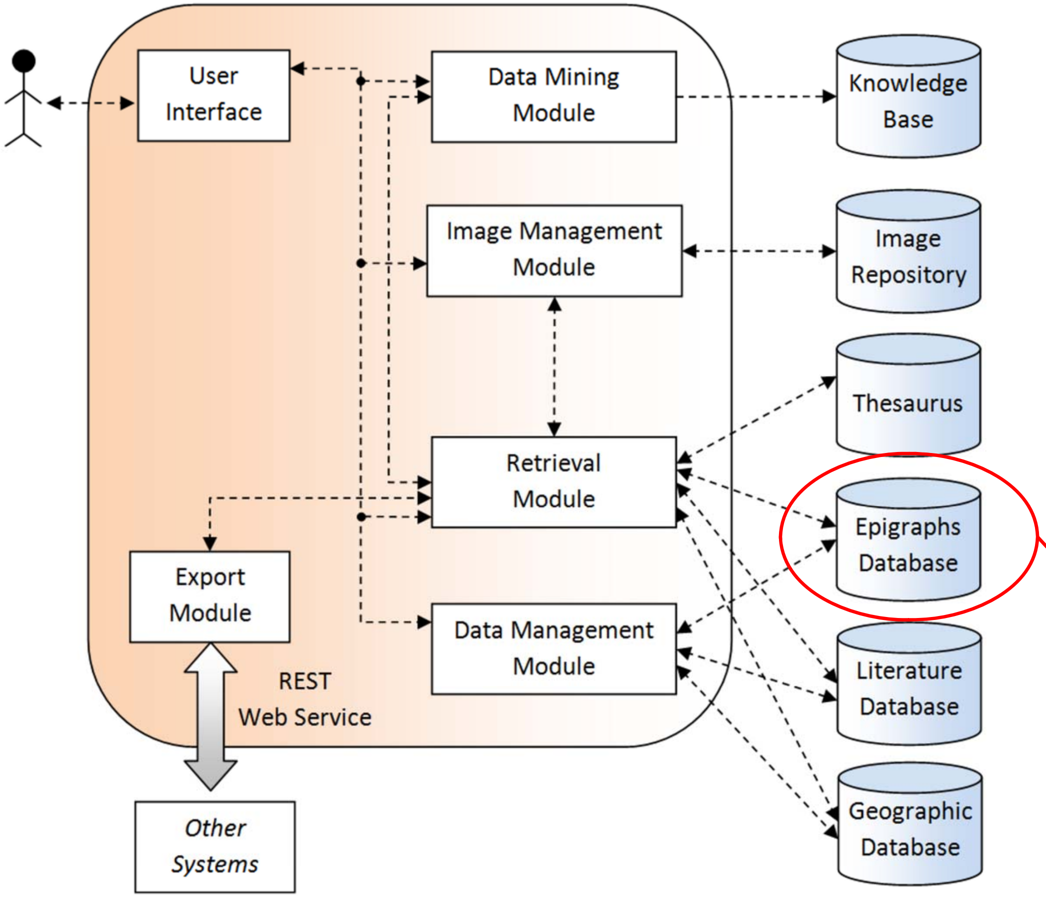
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

www.edb.uniba.it

cini consorzio interuniversitario nazionale per l'informatica

A concrete example: EDB

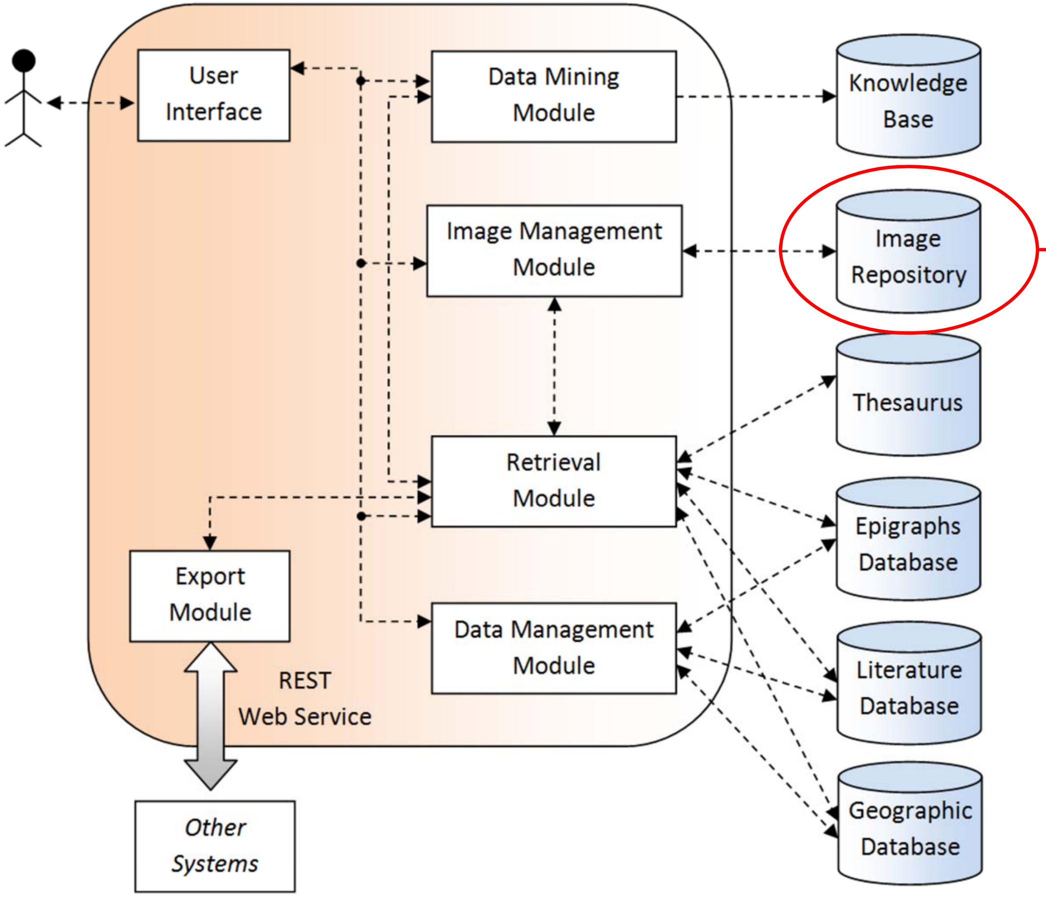
Data Sources



Stores the **text** of the epigraphs and a set of **metadata** about dating, original context, current location, related literature, etc.

A concrete example: EDB

Data Sources

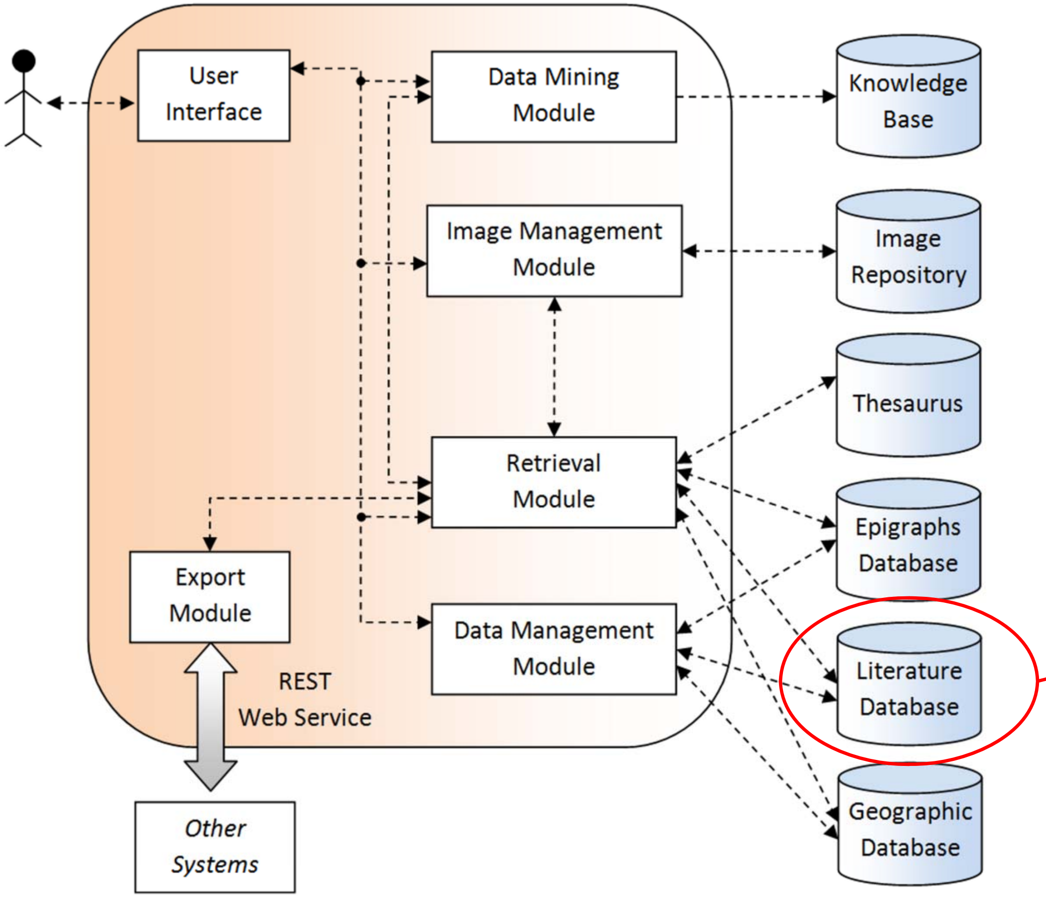


Stores **photos of the epigraphs.**

This repository can be either internal or external, as well as an integration of internally produced and external resources.

A concrete example: EDB

Data Sources

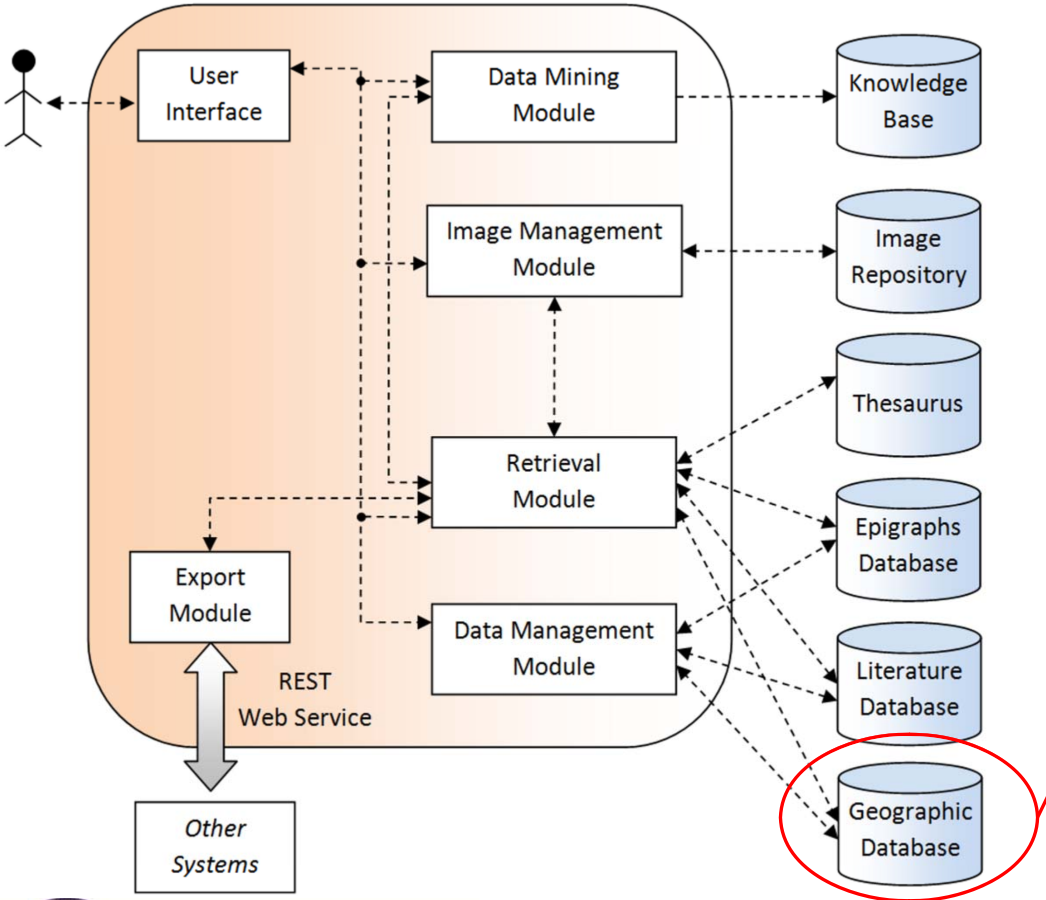


Stores **metadata about scientific papers** in which epigraphs have been studied.

Each epigraph stored in the Epigraph Database can be associated to one or more papers stored in this database.

A concrete example: EDB

Data Sources

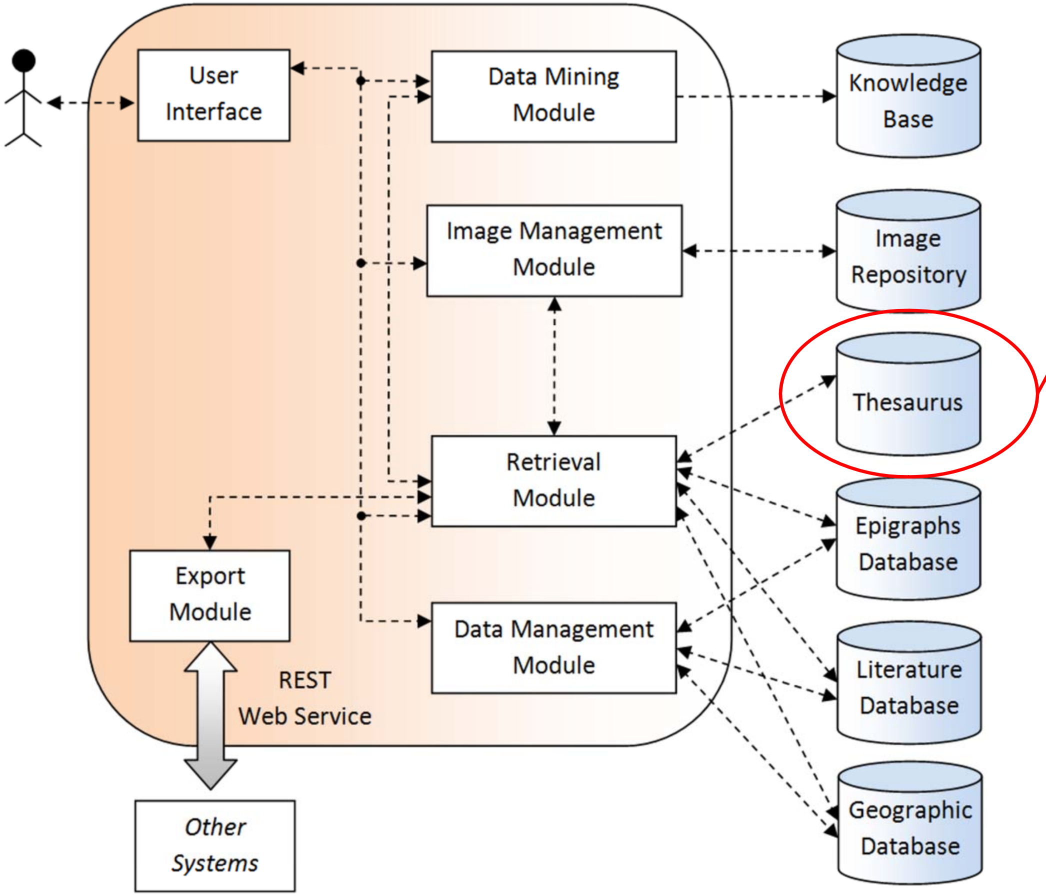


Stores data about **geographic positions**.

One or more geographic positions can be associated to each epigraph (e.g. locations where the epigraph was found and/or the position where the epigraph is currently located).

A concrete example: EDB

Data Sources

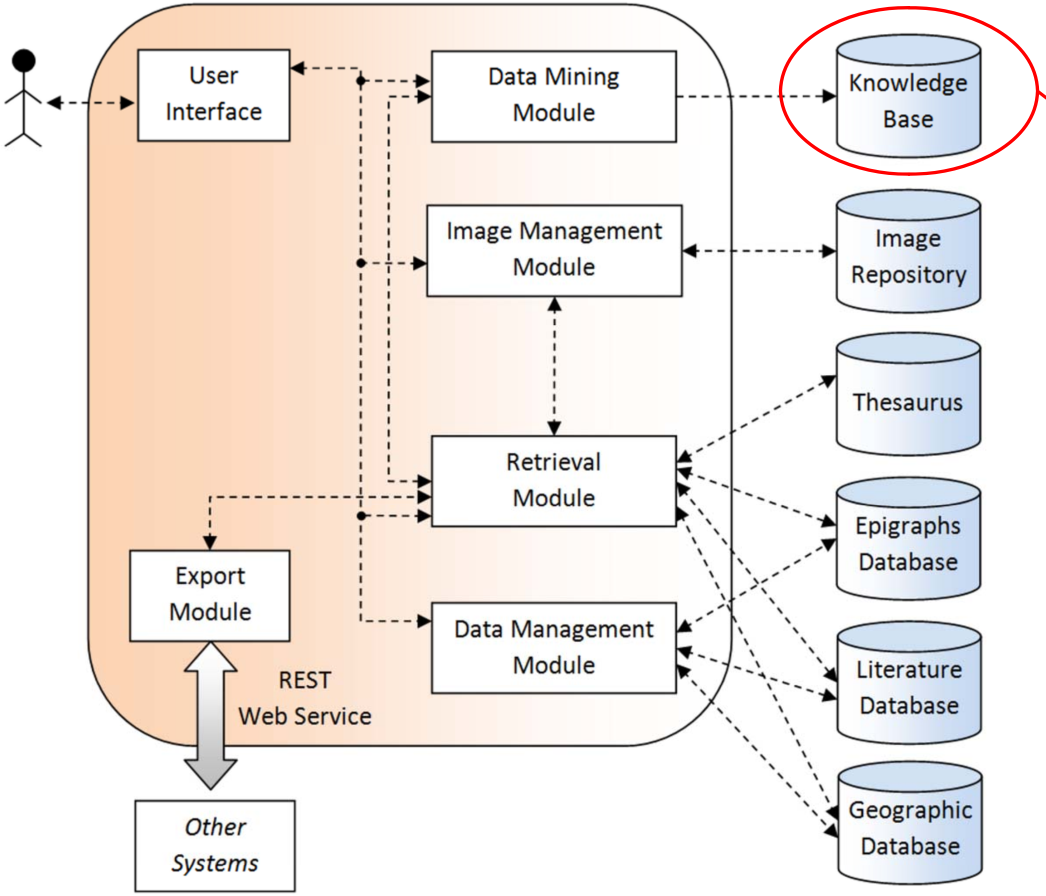


Supports retrieval tasks.

It gives the possibility to enhance and/or perform query expansion to better match data stored in the epigraph database.

A concrete example: EDB

Data Sources

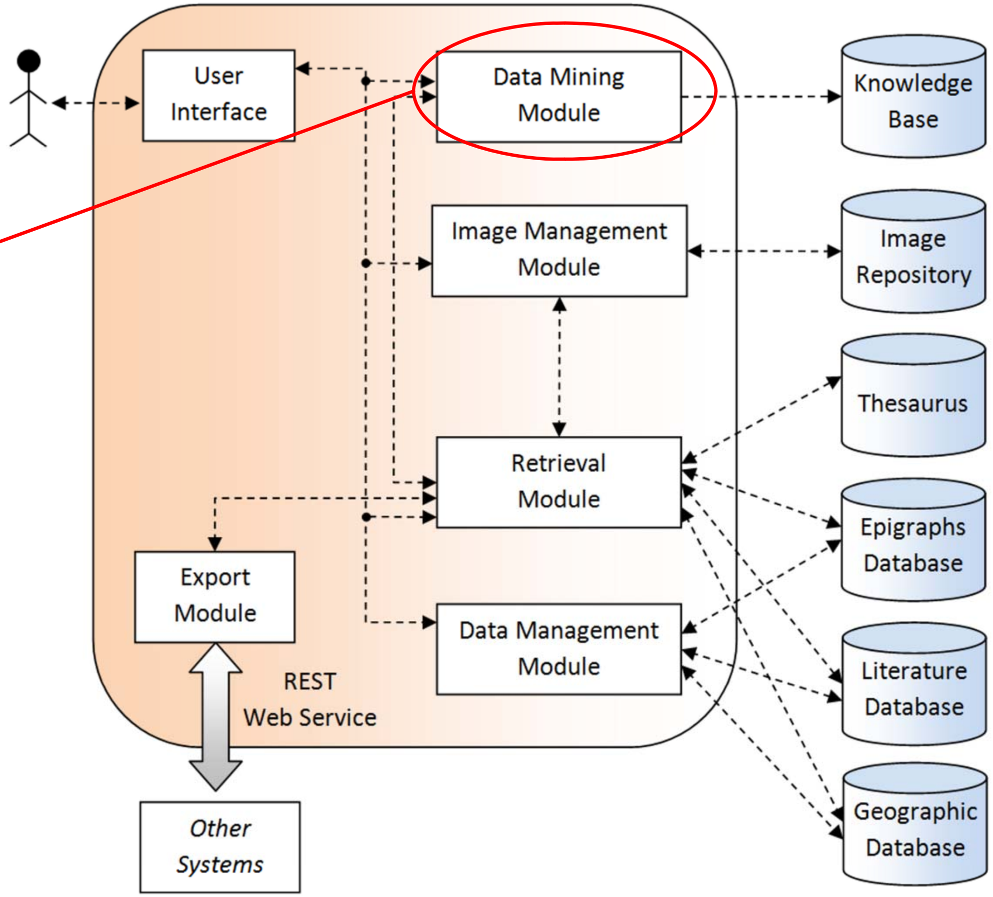


Stores the **knowledge** extracted by the Data Mining Module, i.e. patterns of interest in the form of rules, clusters, etc.

A concrete example: EDB

The Data Mining Module

Allows data analysts to **execute data mining algorithms** on the available data, in order to discover valuable knowledge, which is then stored in the Knowledge Base.



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica

A concrete example: EDB

- Change Detection
- Since epigraphs can usually be associated to an historical period (dating), it could be interesting to identify how social and cultural changes over time have affected the epigraphs.
- Possible aspects to study…

Orthography

Used Materials

Phonetics

Executing Techniques

A concrete example: EDB

- The data analysis task:

identification of emerging patterns, i.e. patterns which show relevant changes (in frequency) over time. The temporal dimension has a central role.

The emerging patterns can be ranked according to some measures of relevance, such as the **growth rate**, which represents the relative variation of the **support** of the pattern in the considered time intervals.

A concrete example: EDB

- Target objects: epigraphs
- Task-relevant objects: materials, executing techniques, kinds of writing, etc.
- Proper time intervals on the basis of a discretization strategy (e.g. equal width, equal frequency, clustering-based discretization).
- Features of interest among all the available ones, in order to focus the algorithm only on the relevant data.

A concrete example: EDB

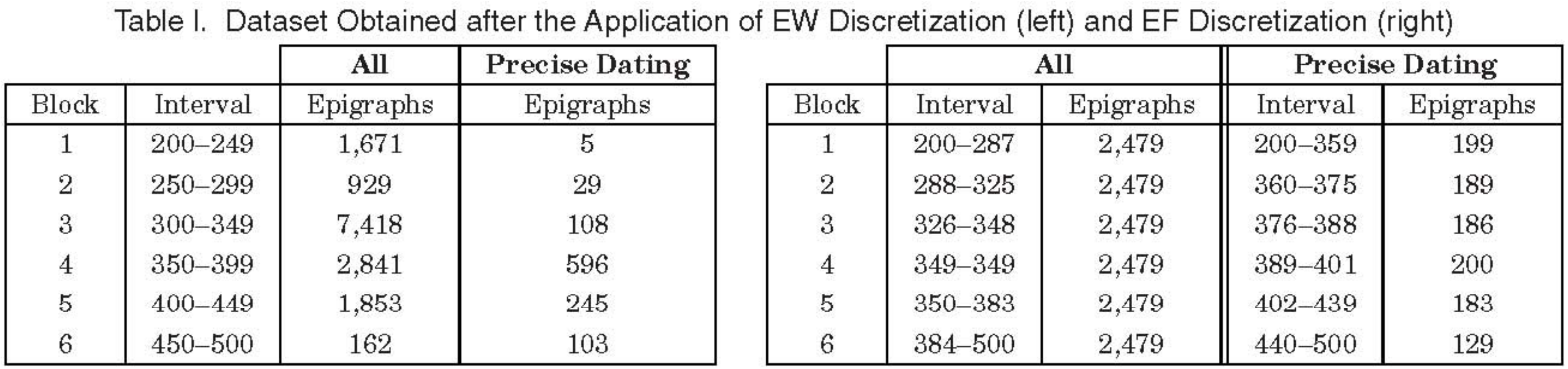


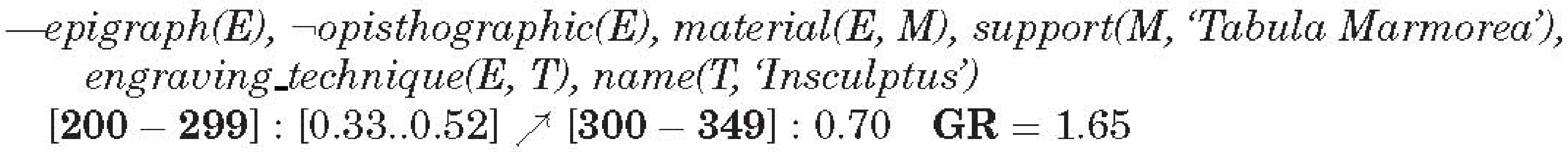
Table I. Dataset Obtained after the Application of EW Discretization (left) and EF Discretization (right)

		All	Precise Dating
Block	Interval	Epigraphs	Epigraphs
1	200–249	1,671	5
2	250–299	929	29
3	300–349	7,418	108
4	350–399	2,841	596
5	400–449	1,853	245
6	450–500	162	103

	All		Precise Dating	
Block	Interval	Epigraphs	Interval	Epigraphs
1	200–287	2,479	200–359	199
2	288–325	2,479	360–375	189
3	326–348	2,479	376–388	186
4	349–349	2,479	389–401	200
5	350–383	2,479	402–439	183
6	384–500	2,479	440–500	129

- Information considered: text, material, support, … for 14,874 epigraphs

A concrete example: EDB



$—epigraph(E), \neg opisthographic(E), material(E, M), support(M, \text{‘Tabula Marmorea’}),$
$\quad engraving\_technique(E, T), name(T, \text{‘Insculptus’})$

$[\mathbf{200} - \mathbf{299}] : [0.33..0.52] \nearrow [\mathbf{300} - \mathbf{349}] : 0.70 \quad \mathbf{GR} = 1.65$

A moderate increase of single-sided epigraphs engraved with the insculptus technique on tabula marmorea in the time interval [300–349] with respect to the interval [200–299]

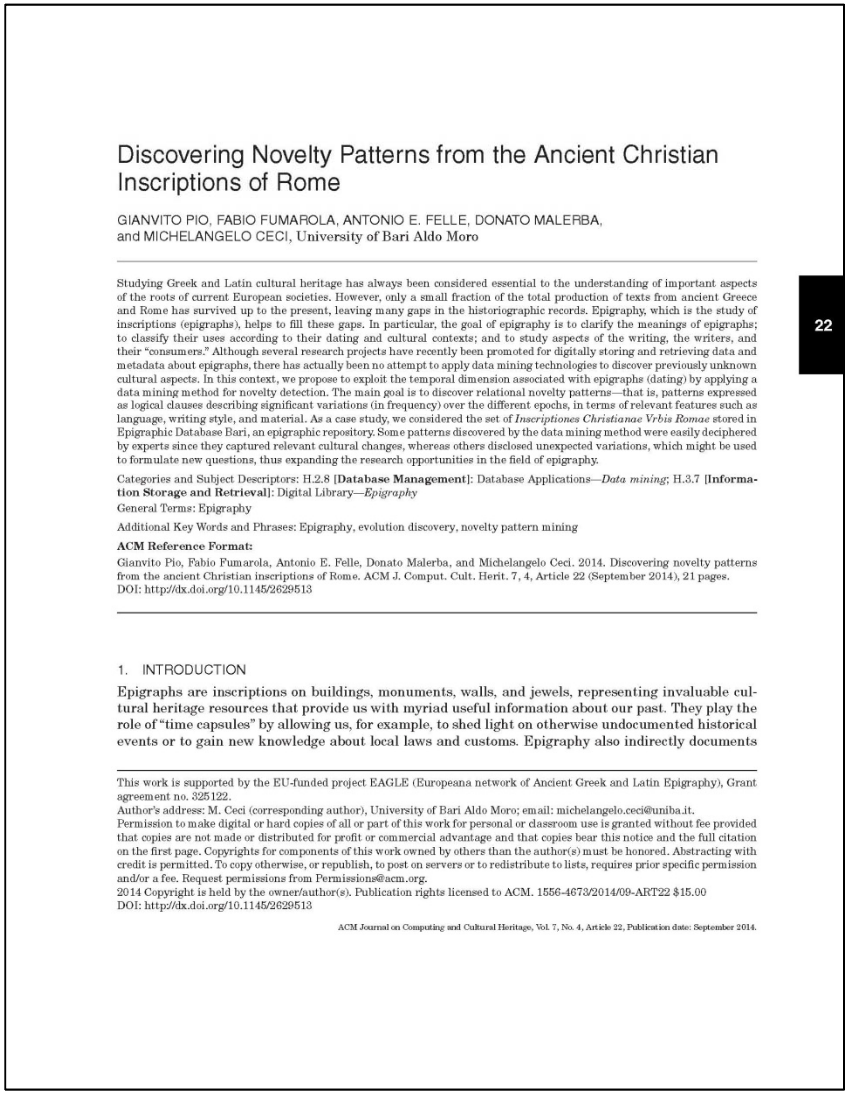
A concrete example: EDB

$—epigraph(E), \neg opisthographic(E), material(E, M), support(M, \text{‘Tabula Marmorea’}),$
$engraving\_technique(E, T), name(T, \text{‘Insculptus’})$
$[\mathbf{200} - \mathbf{299}] : [0.33..0.52] \nearrow [\mathbf{300} - \mathbf{349}] : 0.70 \quad \mathbf{GR} = 1.65$

This may be due to the greater tolerance for Christians under the Emperor Constantine the Great (306–337 AD) and the consequent diffusion of “official” marble epigraphs in public places.

This is a significant change, since the first Christian inscriptions were usually written on tiles and bricks.

A bibliographical reference



Discovering Novelty Patterns from the Ancient Christian Inscriptions of Rome

GIANVITO PIO, FABIO FUMAROLA, ANTONIO E. FELLE, DONATO MALERBA,
and MICHELANGELO CECI, University of Bari Aldo Moro

Studying Greek and Latin cultural heritage has always been considered essential to the understanding of important aspects of the roots of current European societies. However, only a small fraction of the total production of texts from ancient Greece and Rome has survived up to the present, leaving many gaps in the historiographic records. Epigraphy, which is the study of inscriptions (epigraphs), helps to fill these gaps. In particular, the goal of epigraphy is to clarify the meanings of epigraphs; to classify their uses according to their dating and cultural contexts; and to study aspects of the writing, the writers, and their “consumers.” Although several research projects have recently been promoted for digitally storing and retrieving data and metadata about epigraphs, there has actually been no attempt to apply data mining technologies to discover previously unknown cultural aspects. In this context, we propose to exploit the temporal dimension associated with epigraphs (dating) by applying a data mining method for novelty detection. The main goal is to discover relational novelty patterns—that is, patterns expressed as logical clauses describing significant variations (in frequency) over the different epochs, in terms of relevant features such as language, writing style, and material. As a case study, we considered the set of *Inscriptiones Christianae Vrbis Romae* stored in Epigraphic Database Bari, an epigraphic repository. Some patterns discovered by the data mining method were easily deciphered by experts since they captured relevant cultural changes, whereas others disclosed unexpected variations, which might be used to formulate new questions, thus expanding the research opportunities in the field of epigraphy.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.7 [**Information Storage and Retrieval**]: Digital Library—*Epigraphy*

General Terms: Epigraphy

Additional Key Words and Phrases: Epigraphy, evolution discovery, novelty pattern mining

ACM Reference Format:

Gianvito Pio, Fabio Fumarola, Antonio E. Felle, Donato Malerba, and Michelangelo Ceci. 2014. Discovering novelty patterns from the ancient Christian inscriptions of Rome. ACM J. Comput. Cult. Herit. 7, 4, Article 22 (September 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2629513

1. INTRODUCTION

Epigraphs are inscriptions on buildings, monuments, walls, and jewels, representing invaluable cultural heritage resources that provide us with myriad useful information about our past. They play the role of “time capsules” by allowing us, for example, to shed light on otherwise undocumented historical events or to gain new knowledge about local laws and customs. Epigraphy also indirectly documents

This work is supported by the EU-funded project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), Grant agreement no. 325122.
Author's address: M. Ceci (corresponding author), University of Bari Aldo Moro; email: michelangelo.ceci@uniba.it.

DOI: http://dx.doi.org/10.1145/2629513

Conclusions

- Digital technologies are opening up a great transformation in humanities
- Will these new technologies and approaches change the nature of historical inquire?
- Will we see a gradual change, with tools or techniques increasingly being added to established practice, or are we facing a revolution?

Conclusions

- How historical knowledge is validated in the XXI century? In the era of Big Data, are data-driven approaches appropriate for the purpose of historical knowledge validation?
- When is history manipulated? How can epigraphic resources be used to disclose these manipulations? Are data mining methods useful to reveal these manipulations?

Conclusions

- Is it possible to algorithmically discover in the data (inscriptions) the various, equally plausible, storytellings of the past?
- What large outstanding questions can epigraphists hope to address by the convergence of data science and epigraphy?
- Should we expect that data science methods will set agendas for research in epigraphy?

Conclusions

- There is an urgent need for a critical reflection within the epigraphic community, and, more in general, among historians, on the epistemological implications of the current data revolution.
- Some preliminary studies (Pio et al, 2014) have barely begun to tackle the problem, despite the rapid changes in research practices presently taking place.

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

cini
consorzio
interuniversitario
nazionale
per l'informatica

Conclusions

What's the room for the epigraphist in a data-driven world?

The role of the epigraphist in these studies is particularly important.

Their pose research questions to data scientists, they give feedback on results, thus allowing an iterative refinement of analysis algorithms and the development of a user-friendly digital tool.

Thank you

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

cini consorzio interuniversitario nazionale per l'informatica