2017
ALA-ICA
CONFERENCIA
CIUDAD DE MÉXICO

Archivos, Ciudadanía e Interculturalismo

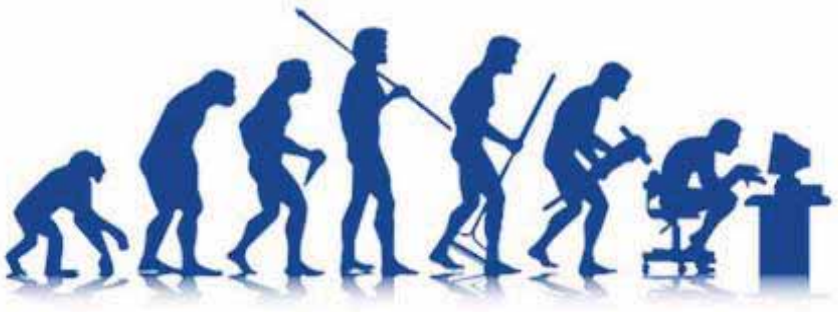
MULTI-RELATIONAL BIG DATA:
THE NEXT CHALLENGE

Fernando Sancho Caparrini
Universidad de Sevilla
fsancho@us.es

Agenda

- About information
- Ideal structures
- An initial proposal
- … and Big Data
- Conclusions

Introduction



20th Century

- Technological Revolution
- Democratization of computers as working tool

…

- A lot of data has been produces and digitalized

21st Century

- Information Revolution
- Democratization of automatic information processing

About Information

- **Unstructured** vs. **Structured**
- Raw vs. Preprocessed
- Schemaless vs. Schema
- Massive analysis capability vs. Not

Two very different contexts:

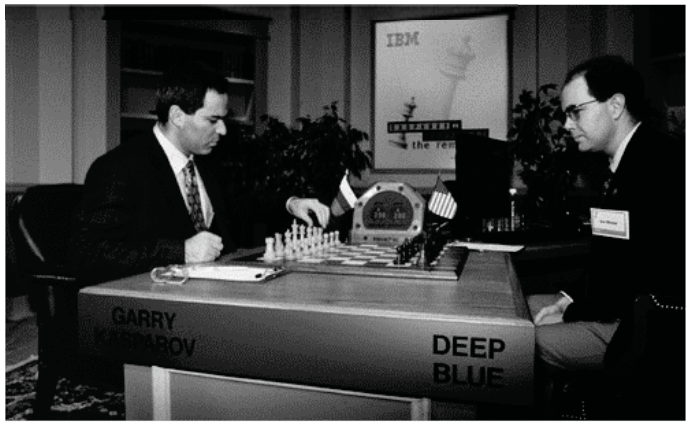
- Scientific Areas
- Humanistic Areas

About Information

Examples of successful projects where information is strongly structured:

- Mathematics
- Physic
- Biological Databases
- Chess, Go,… (Games)
- Image Processing
- Expert Systems (health, insurances,…)

AlphaGo



Information in Archives/Humanities

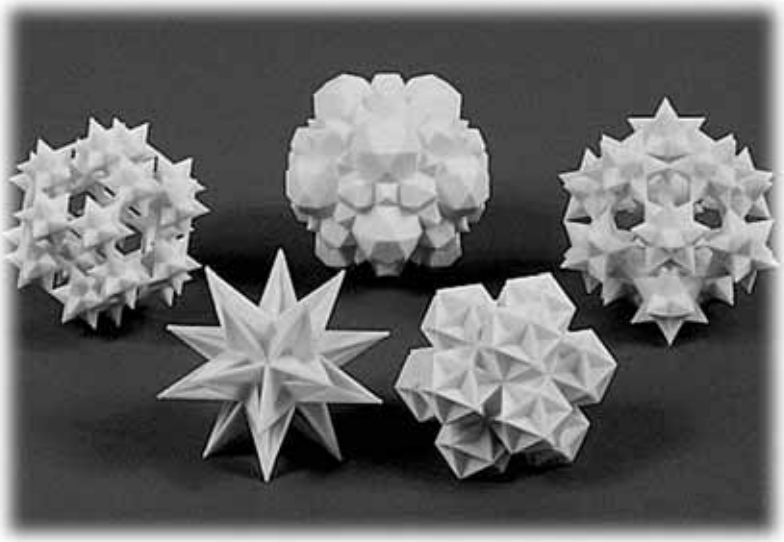
Main features:

- **Structural** Complexity
- **Semantic** Complexity
- **Contextual** Complexity

Sacrifice of interpretative facets

Searching for the Perfect Structure

- Existence of common (and ideal) structures in many disciplines:
  - Vectorial Spaces in Natural Sciences
  - Data-Frames in Social Science
  - …? for general complex purposes



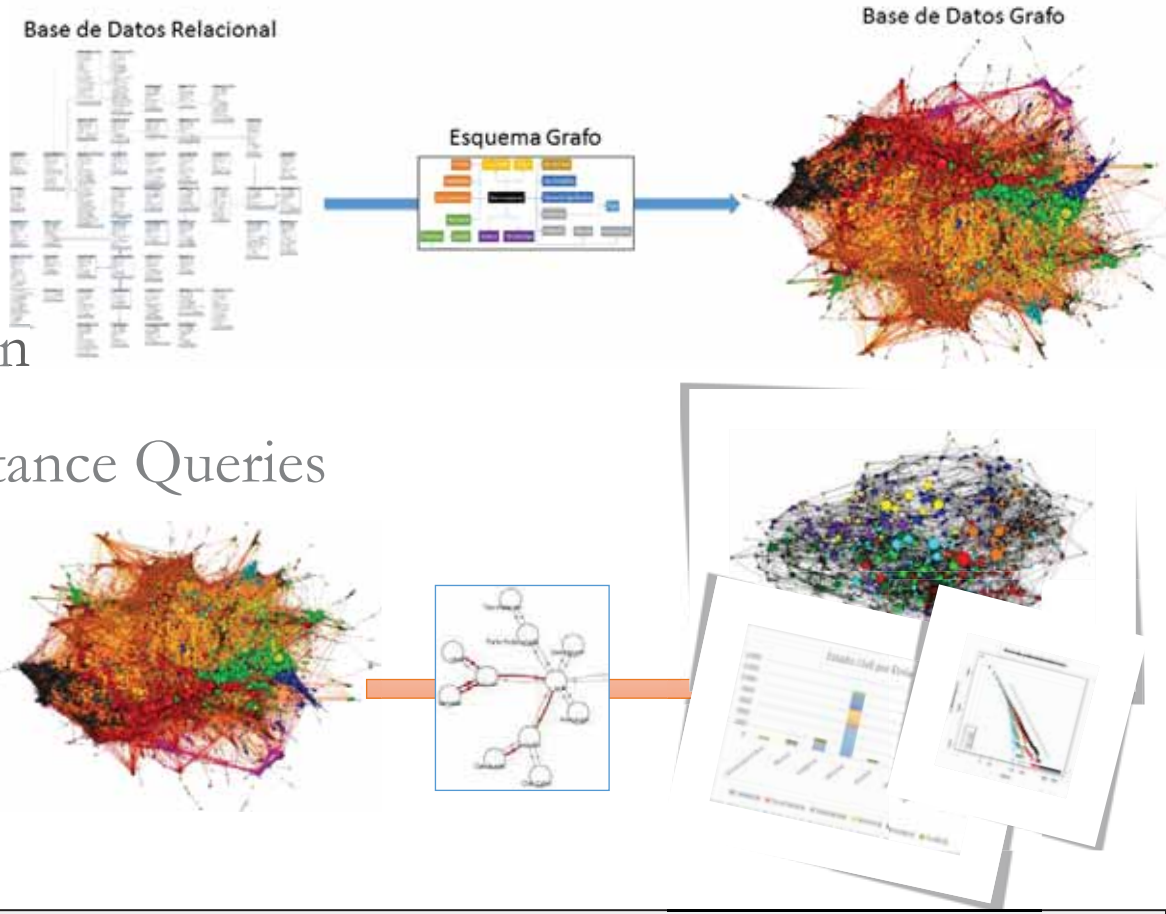
- The importance of standards:
  - Theory to support reasoning
  - Case studies to compare developmer
  - Format conversion and adaptation

What we need…

- Flexibility
  - Multilevel
  - Schemaless
- Tools
  - Storage
  - Handling
  - Analysis
- Standards-based
- Mergeable
- Natural
- Robust
- Reusable
- Verifiable

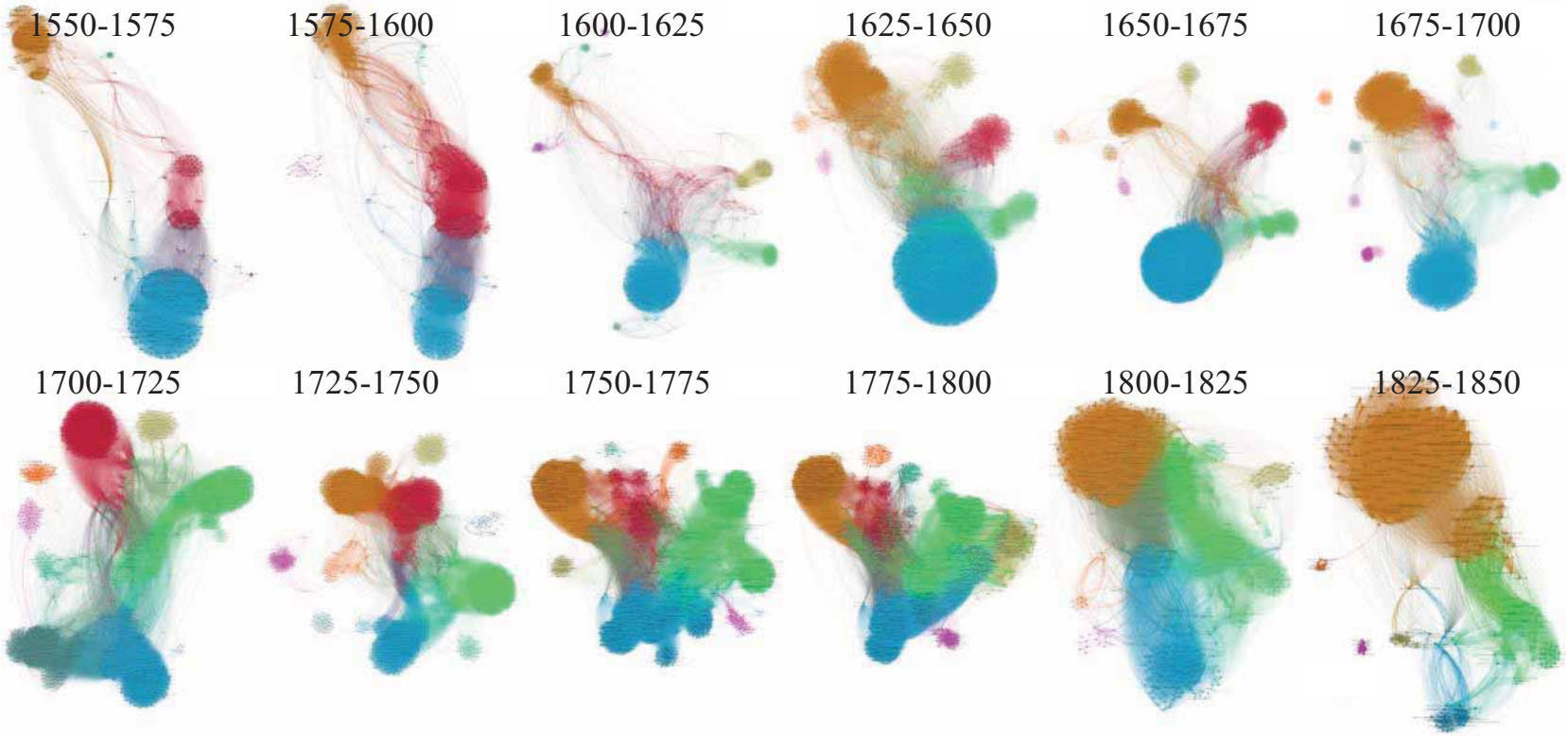
A Proposal: **Multi-relational Networks**

- Based on a robust mathematical theory: **Graph Theory**
- Methodology:
  - Schema Generation
  - Information projection
  - Analysis by Long Distance Queries
  - Link-Discovery

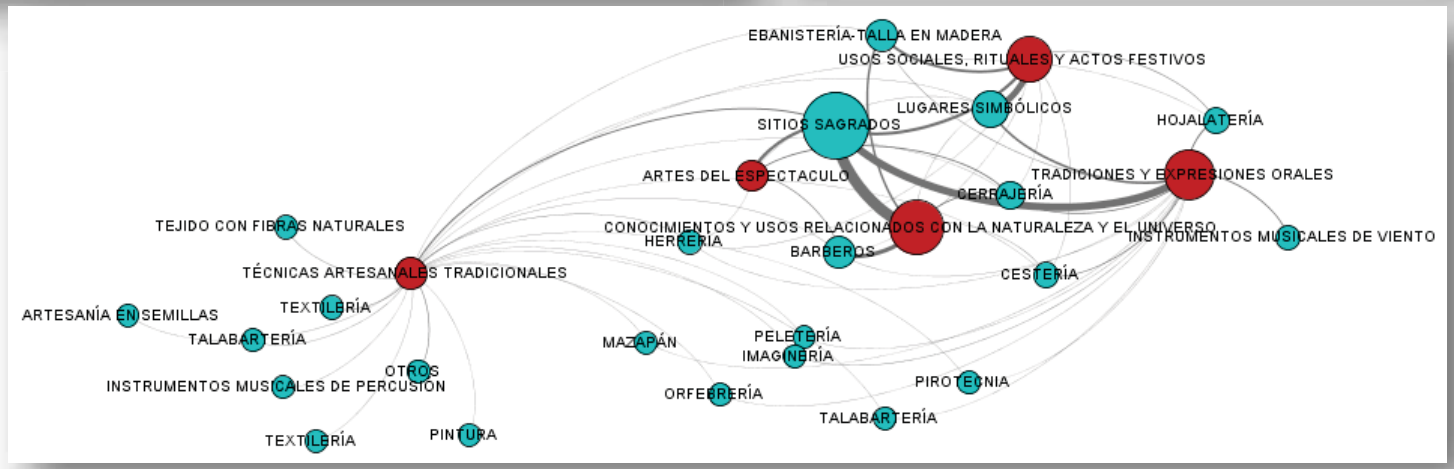


Some simple case studies…

Evolution of Hispanic Baroque

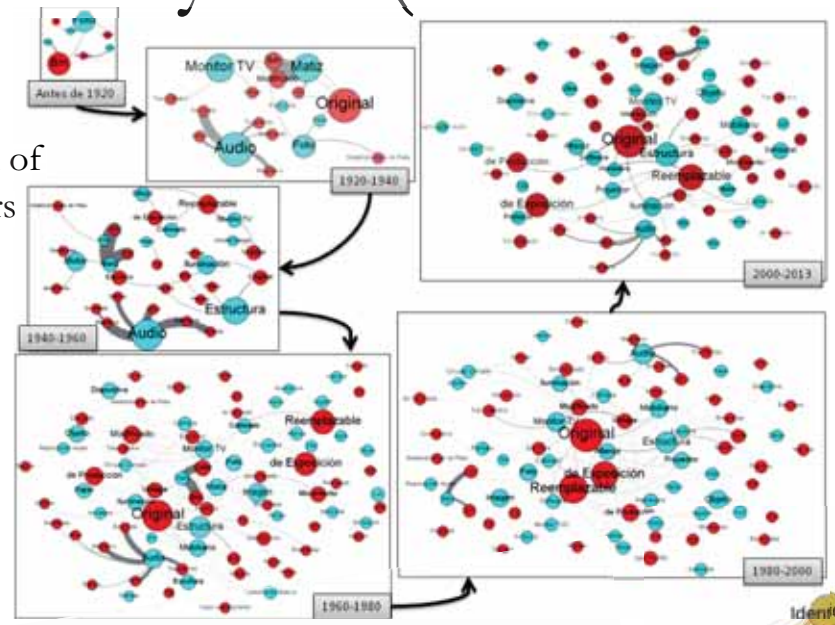


Analysis of Ecuadorian Cultural Heritage

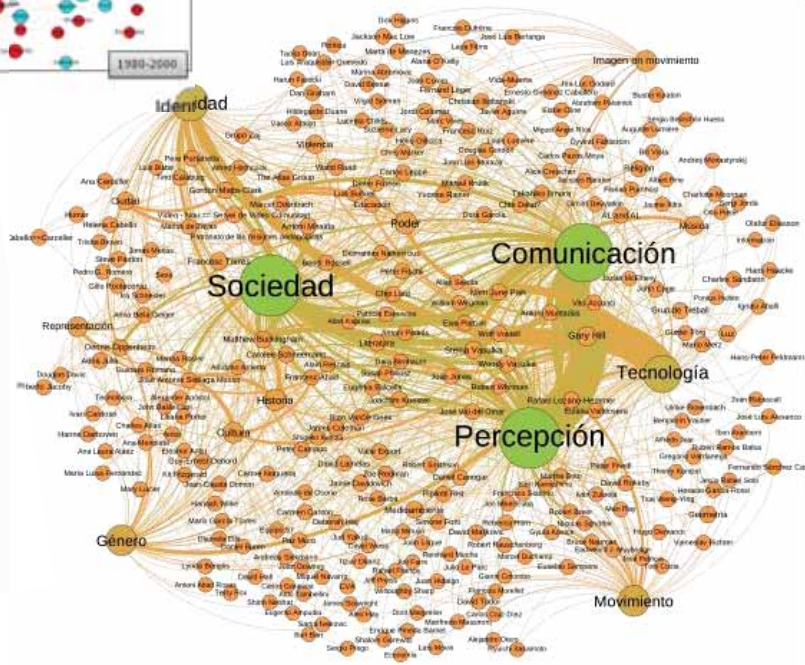


Contemporary art (Museo Reina Sofía)

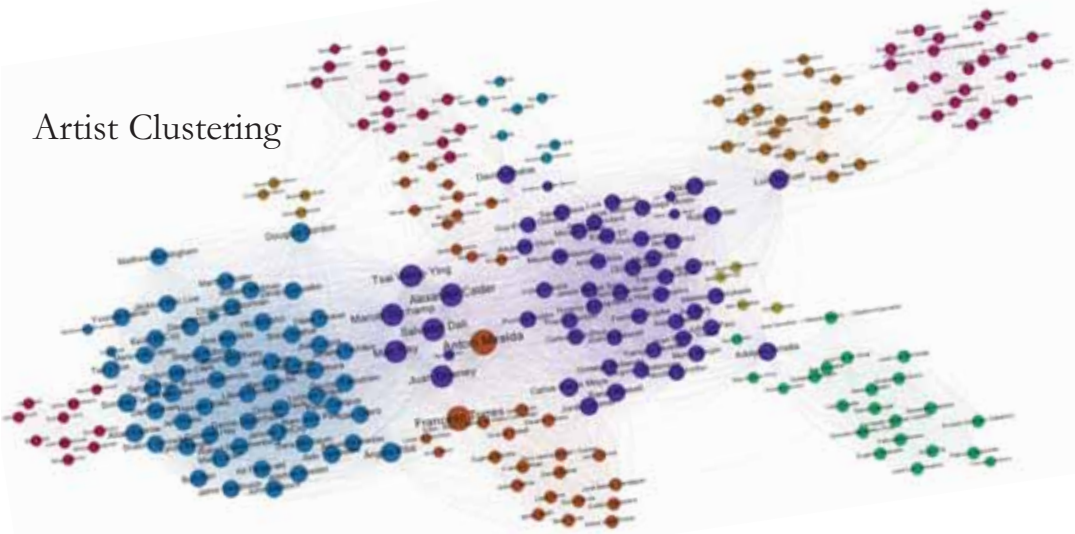
Temporary Evolution of Element Descriptors



Theme by artists

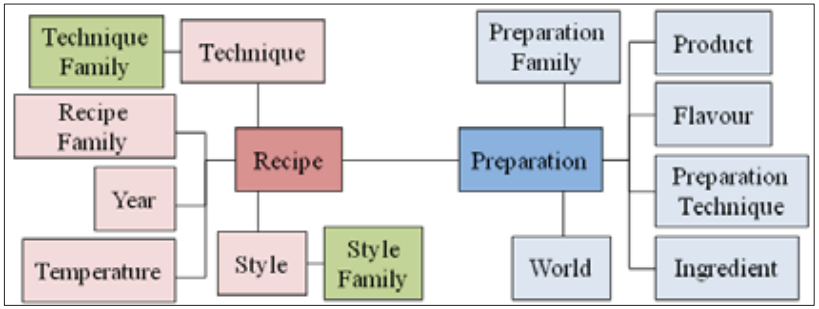


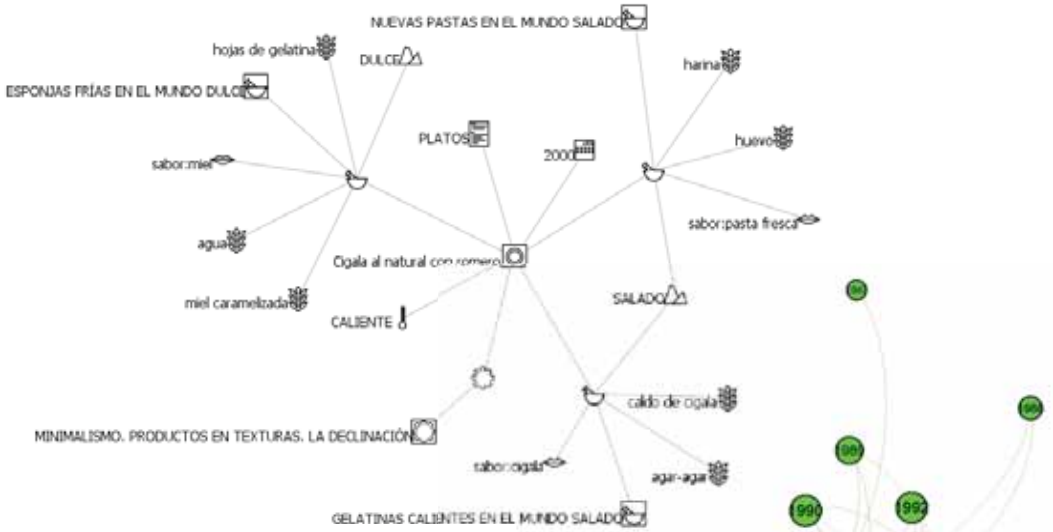
Artist Clustering

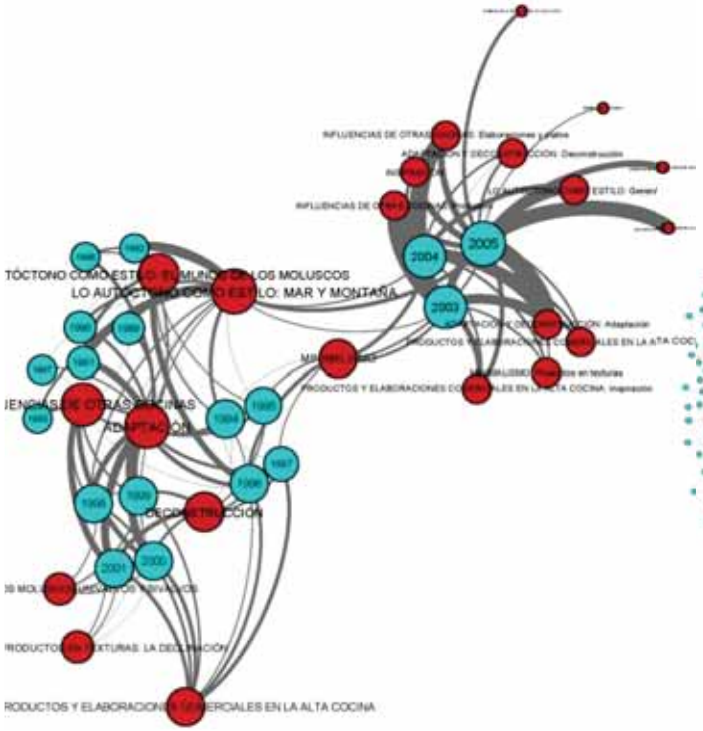


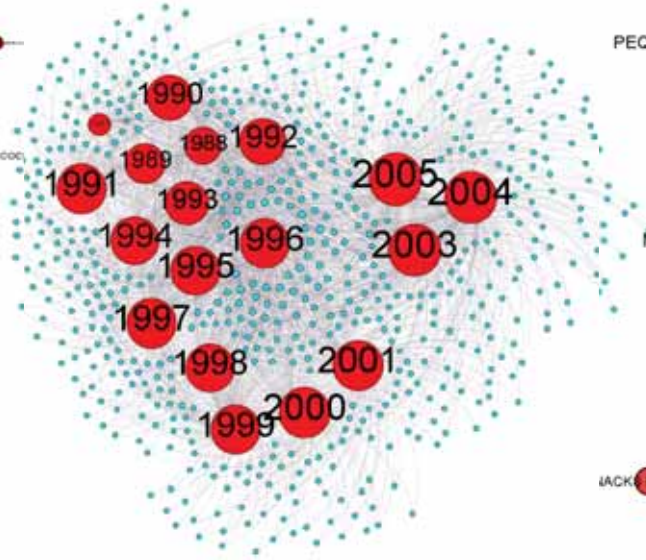
Gastronomic creativity in elBulli

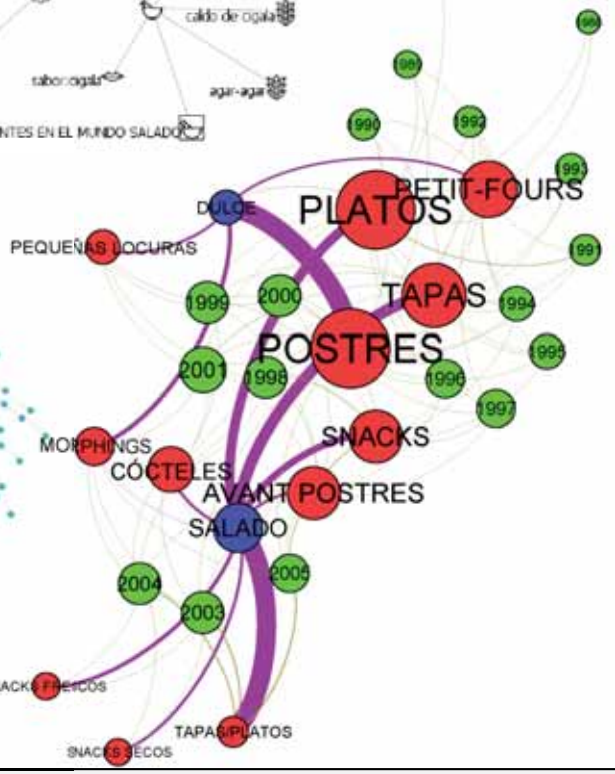
Schema





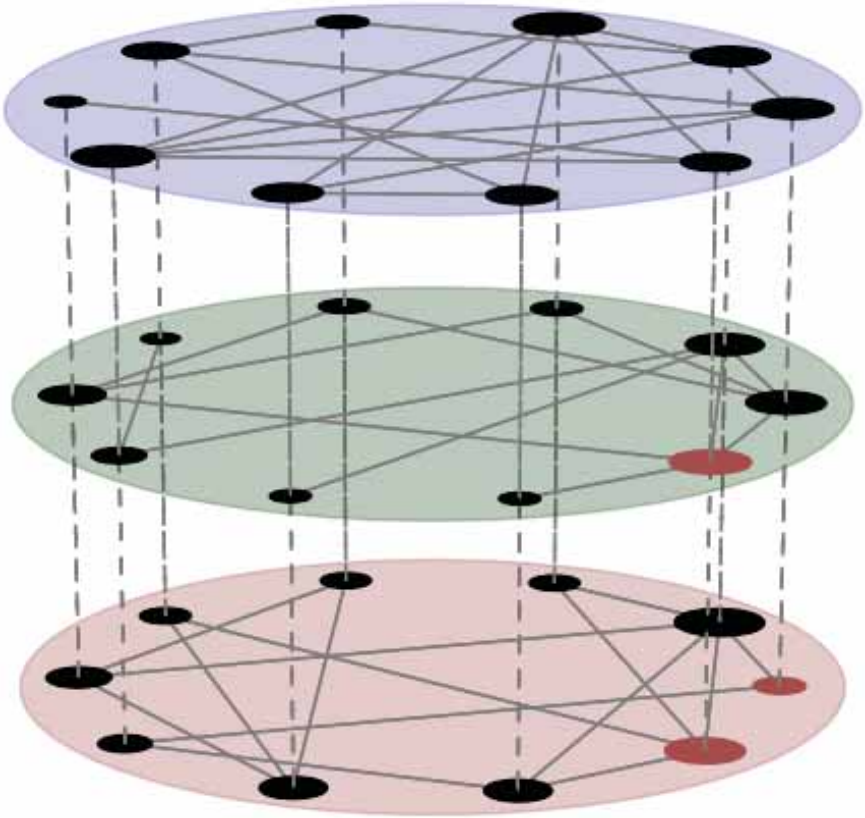






BD + Archives = New Challenges

- **Big Data** : A concept from Business World
- **Volume+Variety+Velocity+Veracity**
- Curation Problem?
  - **Automatic** Curation
  - **No** Curation at all !!!
  - Hybrid Systems by…
- **Merging:**
  - **Merging Networks**
  - **Mining Networks**

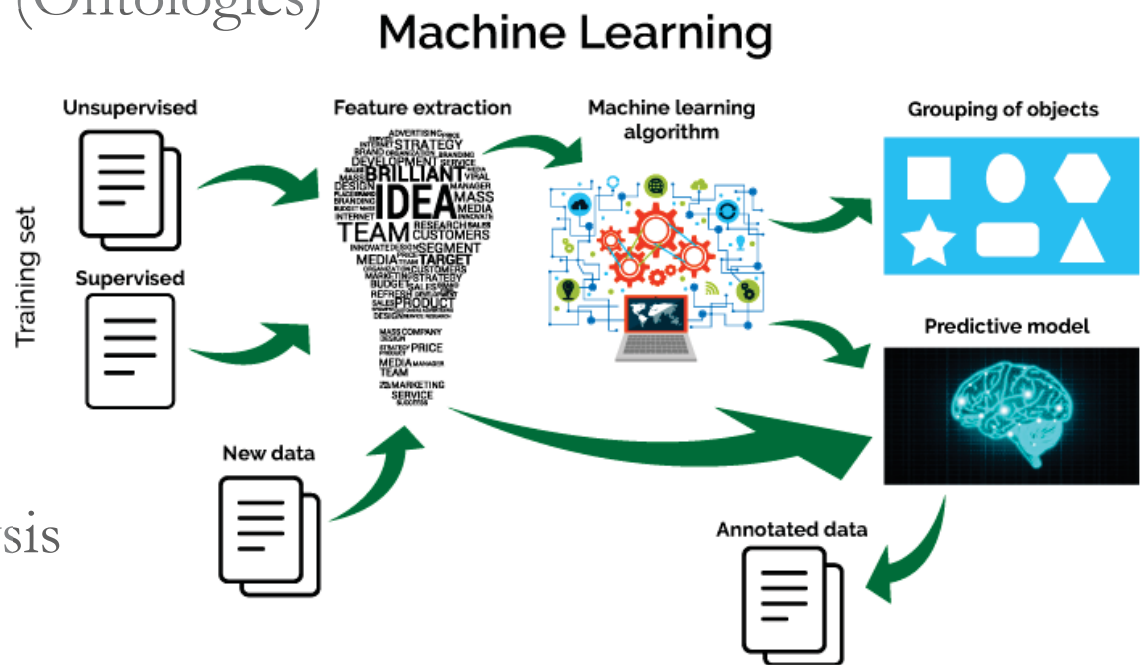


BD + Archives + AI = New Opportunities

Making the defect a virtue: **Semantics as a wealth of information**

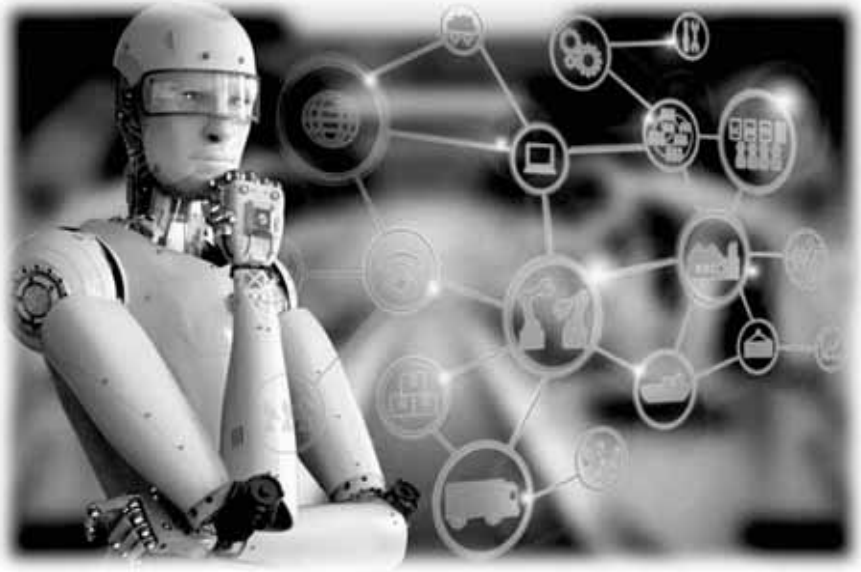
What can be automated ? (…development required)

- Manual Curation Problem
- Automatic Annotation (Ontologies)
- Machine Learning
- Data Science
- Link Discovery
- Semantic Reasoning
- Formal Concept Analysis



We need to keep development…

- **Improve tools** for Multi-relational Networks methodology.
- **New visualizations** to get insight from networks.
- New algorithms to **extract information** from network data.
- Algorithms for **automatic merging** of complex networks.
- Improve **data conversions**:
  - … to text
    - (advanced OCR)
  - … to network
    - (advanced understanding)



Conclusions

In the face of big problems :

- **Multidisciplinary** approaches
- Development of **new tools** (theoretical and practical).
- Need for adequate **training**.

Disciplines involved :

- **Humanities** (diverse) for targeting decision and semantic interpretation.
- **Mathematics** for theoretical modeling.
- **Computer Science** for the effective development of visualization, manipulation.
- **Data Science** for the analysis tools.

Contact?

If you wish to contact :

Fernando Sancho Caparrini: **fsancho@us.es**

(or: **fsanchocaparrini@gmail.com**)

Thank you

for your attention!