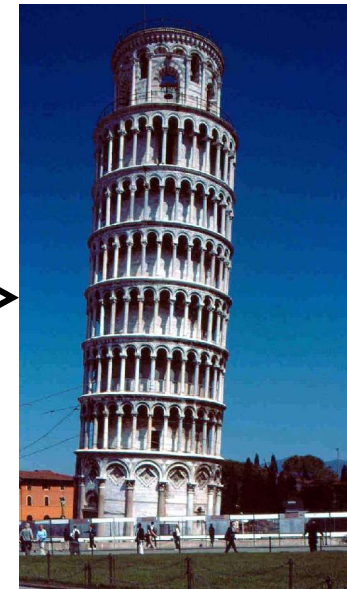


# From Nancy, France to Pisa, Italia



# Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships

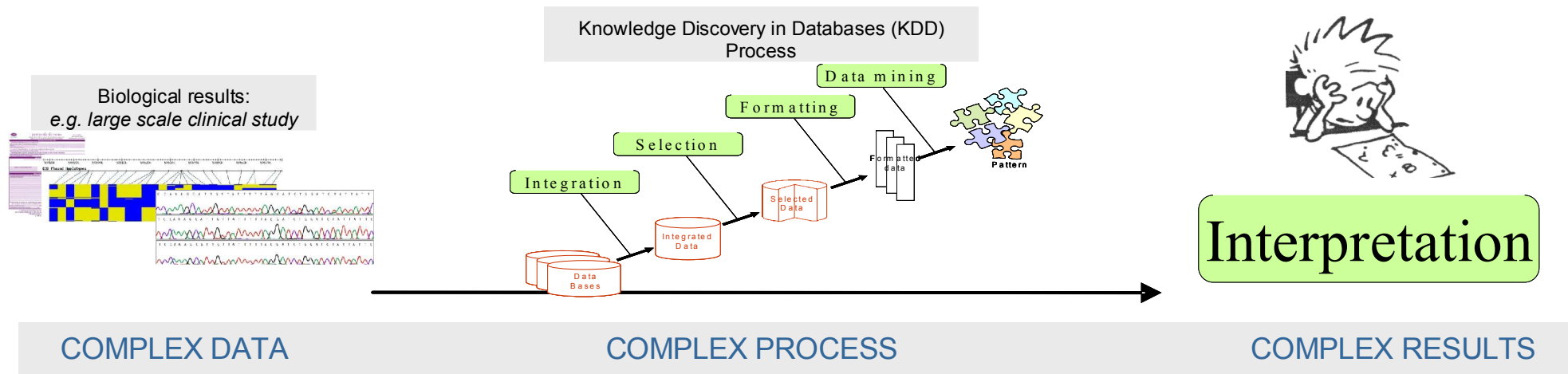
*Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian,  
Amedeo Napoli and Marie-Dominique Devignes*



Laboratoire Lorrain de Recherche en Informatique et ses Applications  
(CNRS, INRIA, University of Nancy), Nancy, France



# The Problem: Limits to KDD in life sciences



- Results of KDD in biology are complex

# Proposition: Use ontologies for guiding the KDD

- 1) Build bridges between data and knowledge

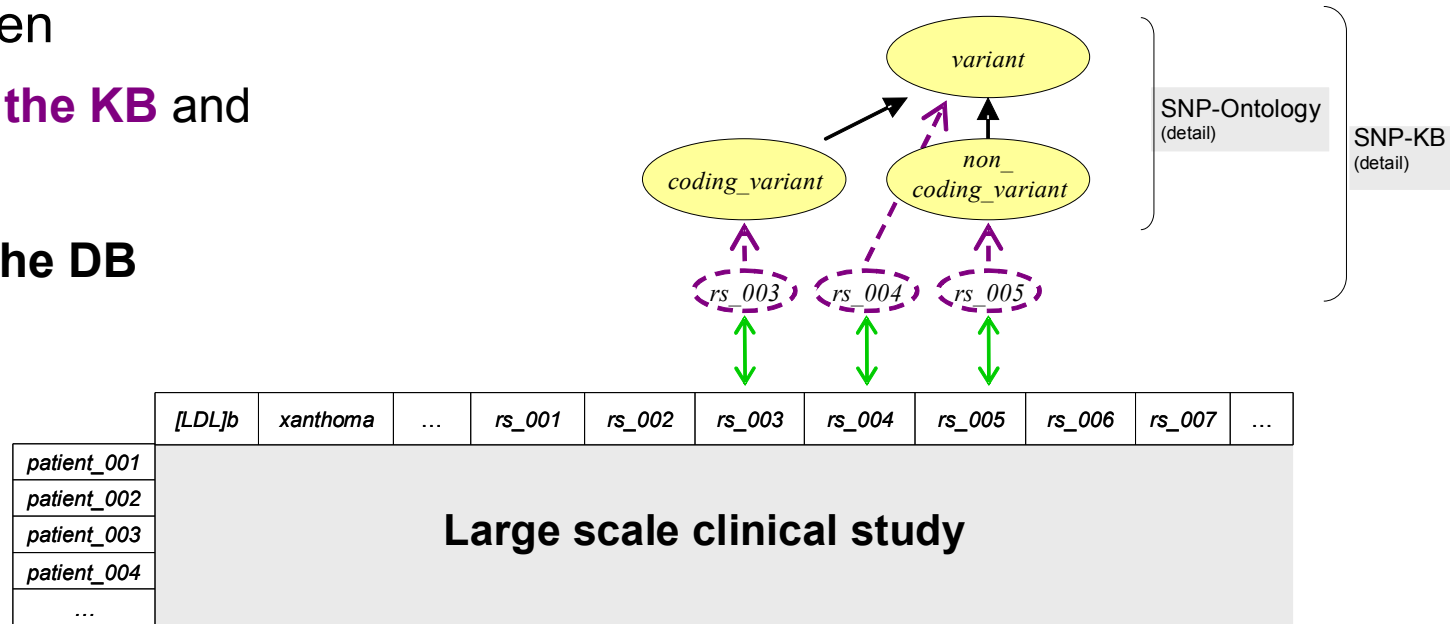
- Mapping** between

- assertions of the KB and



- attributes of the DB

- Example:**



- 2) Use knowledge in order to reduce the size of the data set

- Thanks to *subsumptions*, *object properties*, *class definitions*, etc.

- In order to simplify the interpretation step of KDD process

# For more details ...

- ...see you around the poster
- Poster n°7
- Contact: [adrien.coulet@loria.fr](mailto:adrien.coulet@loria.fr)

**Loria** **Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships**  
Adrien Coulet<sup>1\*</sup>, Malika Smail-Tabbona<sup>2</sup>, Pascale Bonliani<sup>2</sup>, Amedeo Napoli<sup>1</sup> and Marie-Dominique Devignes<sup>3</sup>

<sup>1</sup> Iria Medical, Paris, France  
<sup>2</sup> LORIA (CNRS, INRIA, University of Nancy), Nancy, France  
<sup>3</sup> INSERM UMRS 538, University Pierre and Marie Curie - Paris 6, France  
Contact: [adrien.coulet@loria.fr](mailto:adrien.coulet@loria.fr)

**ABSTRACT:** Complexity of post-genomic data and multiplicity of mining strategies are two limits to Knowledge Discovery in Databases (KDD) in life sciences. Because they provide a semantic frame to data and because they benefit from the progress of semantic web technologies, bio-ontologies should be considered for playing a key role in the KDD process. We propose three scenarios to illustrate how domain knowledge can be taken into account in order to select or aggregate data to mine, and consequently how it can facilitate result interpretation at the end of the process.

**The Problem**  
Scenario 1  
Biological results  
e.g. large scale clinical study  
Ex: 125 patients X 289 variants  
COMPLEX DATA  
Knowledge Discovery in Databases (KDD) Process  
Data selection  
Data mining  
Interpretation  
168 frequent itemsets  
187 clusters  
COMPLEX RESULTS

**Proposition**  
MAKE USE OF ONTOLOGIES AND KBs TO GUIDE DATA PREPARATION  
Caption  
SO-Pharm-KB (real)  
SO-Pharm (real)  
Large scale clinical study  
Phenotype Attributes  
Genotype Attributes  
Attributes of the Database  
SNP-Ontology (real)  
SNP-KB (real)  
A mapping between assertions of the KB and attributes of the DB makes bridges between knowledge and data

**Scenario 1**  
Attribute aggregation thanks to object properties: Haplotype-based aggregation  
Large scale clinical study  
Large scale clinical study  
Ex: 125 patients X 289 variants → 125 patients X 178 tag SNPs → 32 frequent itemsets  
48 clusters

**Scenarios >1**  
Scenario 2: Attribute selection thanks to subsumption: mining successively variants from following classes: variant, coding\_variant and conserved\_domain\_variant.  
Scenario 3: Object selection thanks to class definition: comparing mining results from various defined classes of patients e.g. mutated\_patient vs non\_mutated\_patient  
Scenario n: Any composition of this scenarios

**Conclusion**  
Domain knowledge serves as a guide for data preparation step in knowledge discovery process.  
The use of domain knowledge in KDD process may be extended to other steps:  
- during the data mining step  
- during the interpretation step  
Implementation on the way

**References**  
Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, SF, 2005.  
Napoli, A. Elements on KDDK: Knowledge Discovery guided by Domain Knowledge. CLADE, Hammamet, Tunisia; 2005.  
SNP-Ontology ([http://www.bionontology.org/files/6733/snpontology\\_full.owl](http://www.bionontology.org/files/6733/snpontology_full.owl))  
SO-Pharm ([http://www.loria.fr/~coulet/sopharm/3\\_description.html](http://www.loria.fr/~coulet/sopharm/3_description.html))  
HapMap (<http://www.hapmap.org/>)