Data mining and medical applications, Paris 2003

Conference report, August 1st, 2003

Joannès Vermorel, École Normale Supérieure

Summary

- *Supervised Learning and Optimizing the ROC Curve*, Michele Sebag, Jerome Aze and Noel Lucas
- *Kasimir: a system for semantic web in cancerology*, Sebastien Brachais, Mathieu d'Aquin, Jean Lieber, Amedeo Napoli
- *Time-series Segmentation and Symbolic Representation*, Bernard Hugueney, Bernadette Bouchon-Meunier
- *Decrypton: a tool for comparative proteomics*, William Saurin

Supervised Learning and Optimizing the ROC Curve

Michele Sebag, Jerome Aze and Noel Lucas

- Criterion for the evaluation of supervised learning algorithm: the **A**rea **U**nder the ROC **C**urve (AUC).
- **E**volutionary **RO**C-based **L**earning algorithm (EROL) illustrated on the identification factors for atherosclerosis.

Task: identification of Arthrosclerosis factors

Data: PKDD Challenge 2002, publicly available (http://ecmlpkdd.cs.helsinki.fi)

Task: Submitting a data mining report.

Difficulties:

- Hidden factors
- Detailed description ➔ sparse matrix
- Communication with the experts

Proposed model

Metaphor: the body is a bridge

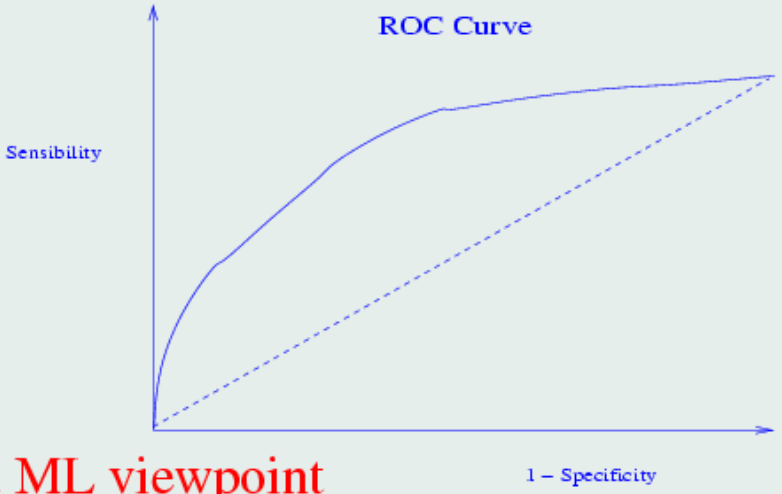
- Initial robustness: family anamnesis
- Current robustness: personal variable
- Traffic load: tobacco alcohol

Medical intensive help ➔ reduction of the variable set by using synthesis variables.

Receiver Operating Characteristics

Evaluating a medical test: tradeoff between

- True positive rate (medical: sensibility)
- False positive rate (medical: 1 - specificity)



Pros from a ML viewpoint

- Operational with ill-balanced data
- Usable with cost errors

Extracted from Michele Sebag works
- reproducted with permission.

The EROL algorithm

Description:

- Linear regression model
- Using self adapting evolutionary algorithm
- AUC as the cost function

Provides:

- Weights for every variable that could be interpreted as the “importance” of the factor.
- The ROC Curse that could be used by the human expert to adjust the selectivity/sensitivity tradeoff.

Conclusion and perspectives on the medical application

Deserve the physician's attention

- Stability of learning
- Precise and still readable results

Going further

- Explore new hypotheses : What happens if ?
  If I fix the weight of alcool and family variables,
  what happens about the weight of education ?
- Consider committees of experts

References

- Michele Sebag Home page: http://www.lri.fr/~sebag/
- PKDD 2002: http://ecmlpkdd.cs.helsinki.fi/
- Inference and Learning Group at LRI: http://www.lri.fr/ia/introduction.en.html

Kasimir: a system for semantic web in cancerology

Sebastien Brachais, Mathieu d'Aquin, Jean Lieber, Amedeo Napoli

- **Background:** the Kasimir project

Knowledge based system for aided diagnosis in cancerology (breast cancer treatment).

- **Goals:** Knowledge management in cancerology

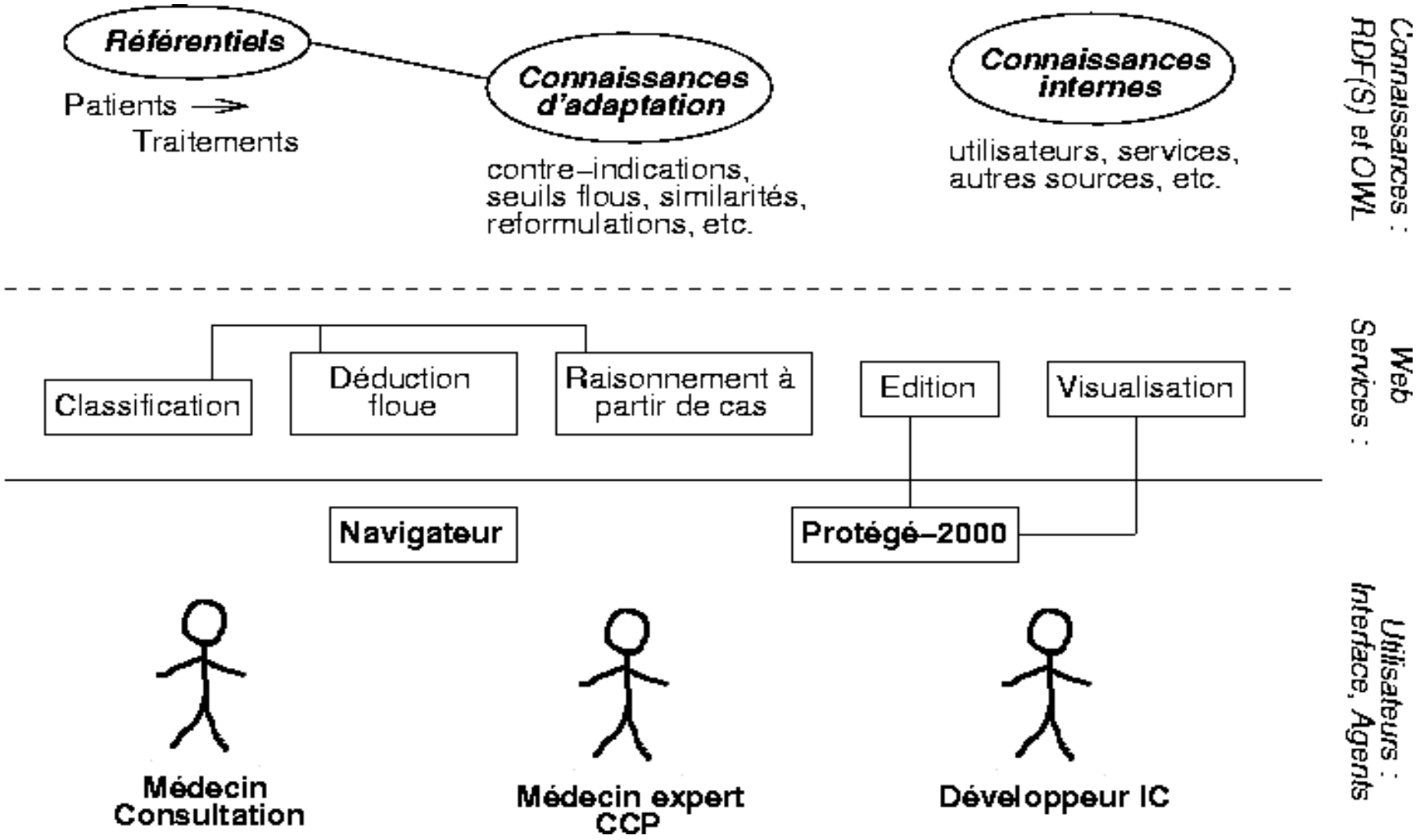
Acquisition, representation, diffusion and evolution of medical knowledge

- **Solution:** a semantic web architecture

The benefits of the semantic web technologies to achieve the Kasimir objectives.

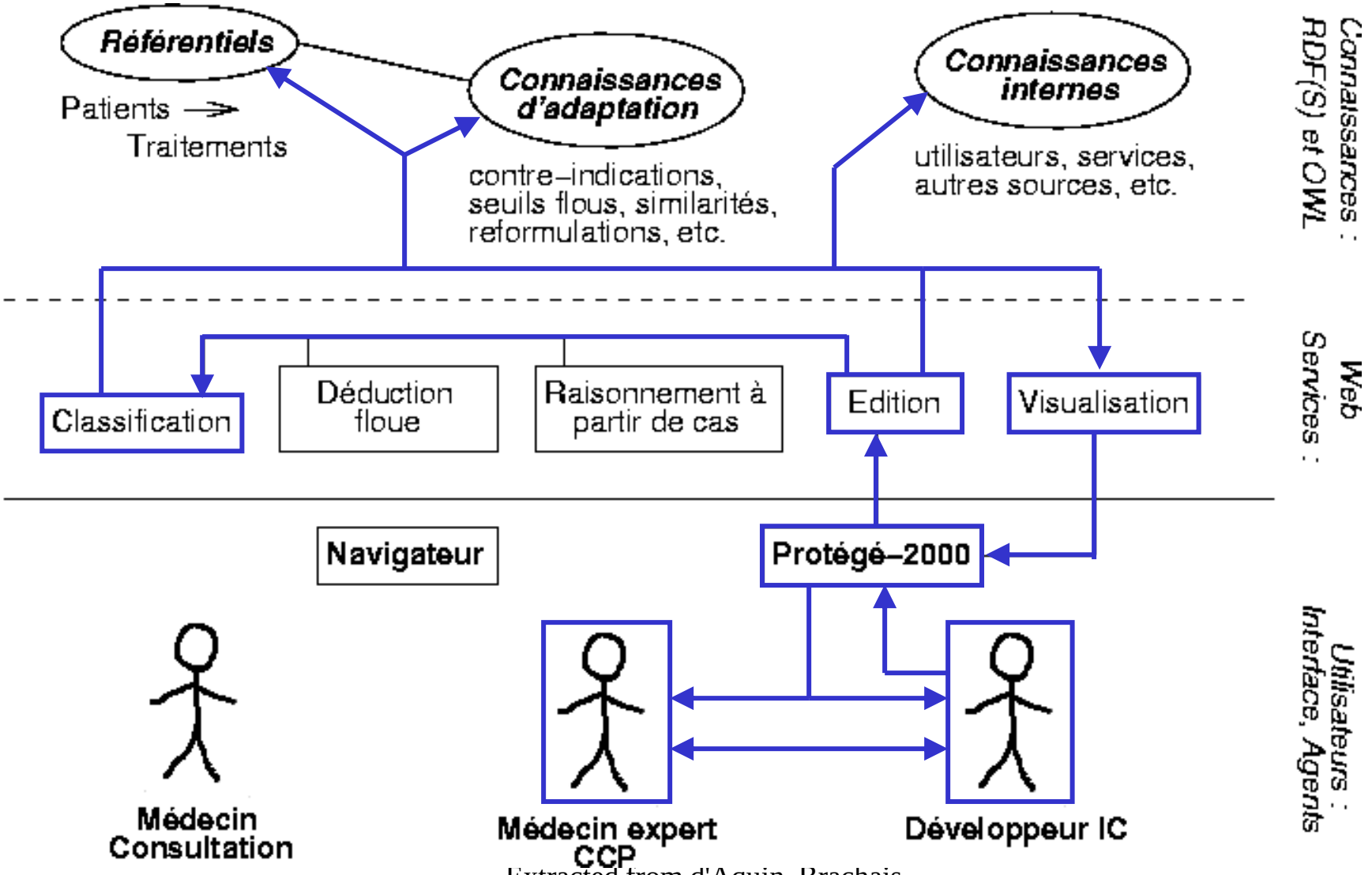
Lets see the big picture of Kasimir …

Architecture



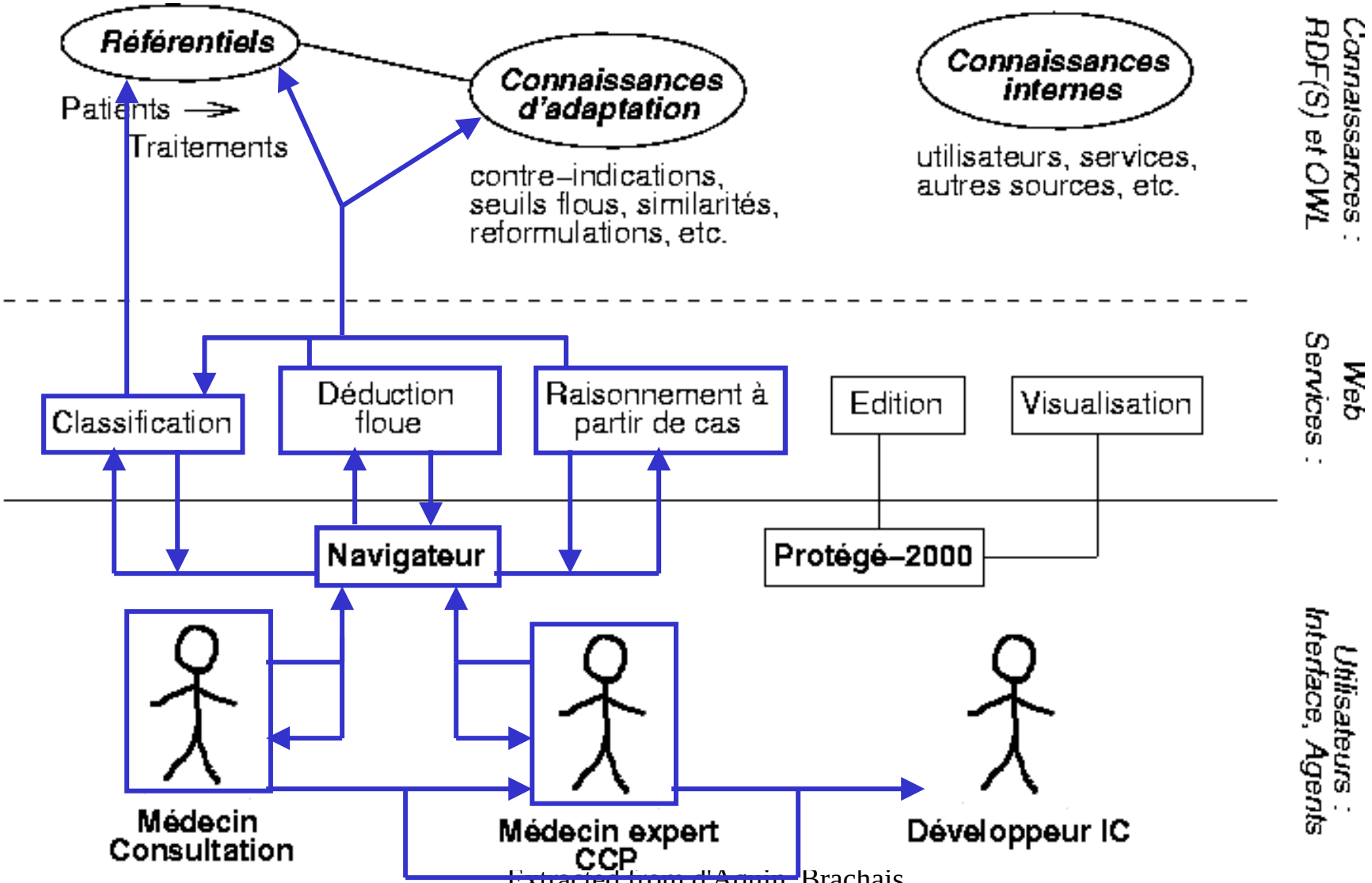
Extracted from d'Aquin, Brachais, Lieber and Napoli works - reproducted with permission

Building, editing and maintaining knowledge



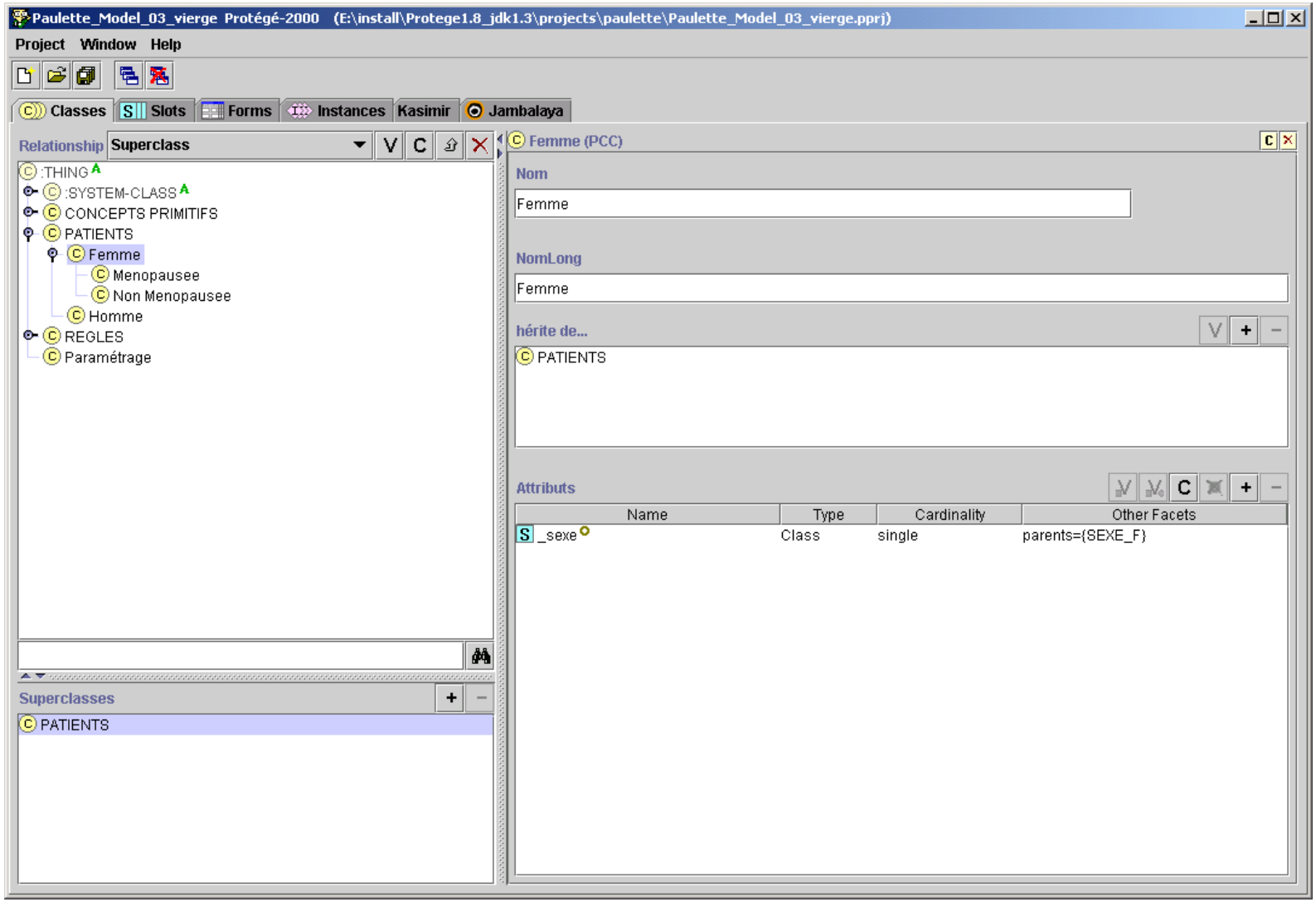
Extracted from d'Aquin, Brachais, Lieber and Napoli works - reproducted with permission

Diffusion and aided decision access



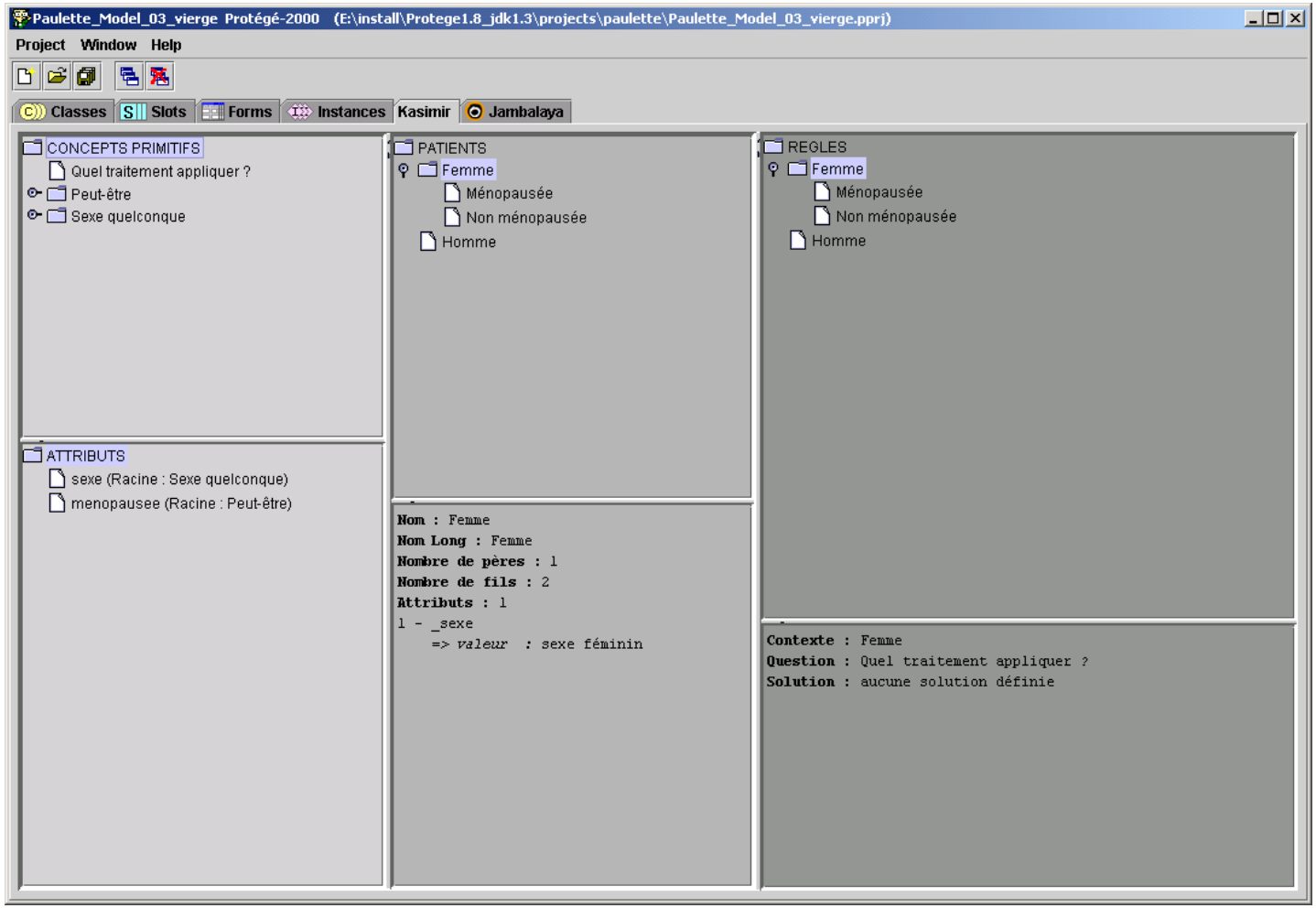
Extracted from d'Aquin, Brachais, Lieber and Napoli works - reproducted with permission

Protégé: knowledge editor



Extracted from d'Aquin, Brachais, Lieber and Napoli works - reproducted with permission

Protégé: Communication with the knowledge representation engine



Extracted from d'Aquin, Brachais, Lieber and Napoli works - reproducted with permission

Benefits of the Kasimir architecture

- Enable a distributed and cooperative creation of the knowledge base
- Open and extensible architecture relying on web standards (RDF, OWL)
- Interoperability with other semantic web system

➔ Kasimir as a resource for the semantic web

Open questions

- Integration of Kasimir in other systems: existing ontologies, services, databases.
- Security issue: confidentiality and integrity.
- Knowledge evolution: enabling a “knowledge versioning system”.
- Improving the reasoning capabilities of Kasimir: case based reasoning, fuzzy logic…

References

- Amedeo Napoli Home page: http://www.loria.fr/~napoli/
- Orpailleur Group at LORIA: http://www.loria.fr/equipes/orpailleur/index.Anglais.html

Time-series Segmentation and symbolic Representation

Bernard Hugueney, Bernadette Bouchon-Meunier

Background: Data streams

The data volumes are too large to be stored in a fast access memory of a typical computer.

Objective: symbolic pre-processing

Converting the raw stream into a high level representation: a symbol stream with much fewer elements.

Times series segmentation

Segmenting task:

Deciding where to place the boundaries in order to split the stream into buckets.

Criterion for choosing the boundaries:

Maximum likelihood estimation

Accuracy vs. Parsimony tradeoff:

More buckets →

more accurate model + less data reduction

Symbolic representation

Problem:

Handling segments as set of real values is not convenient, a more abstract representation should be used.

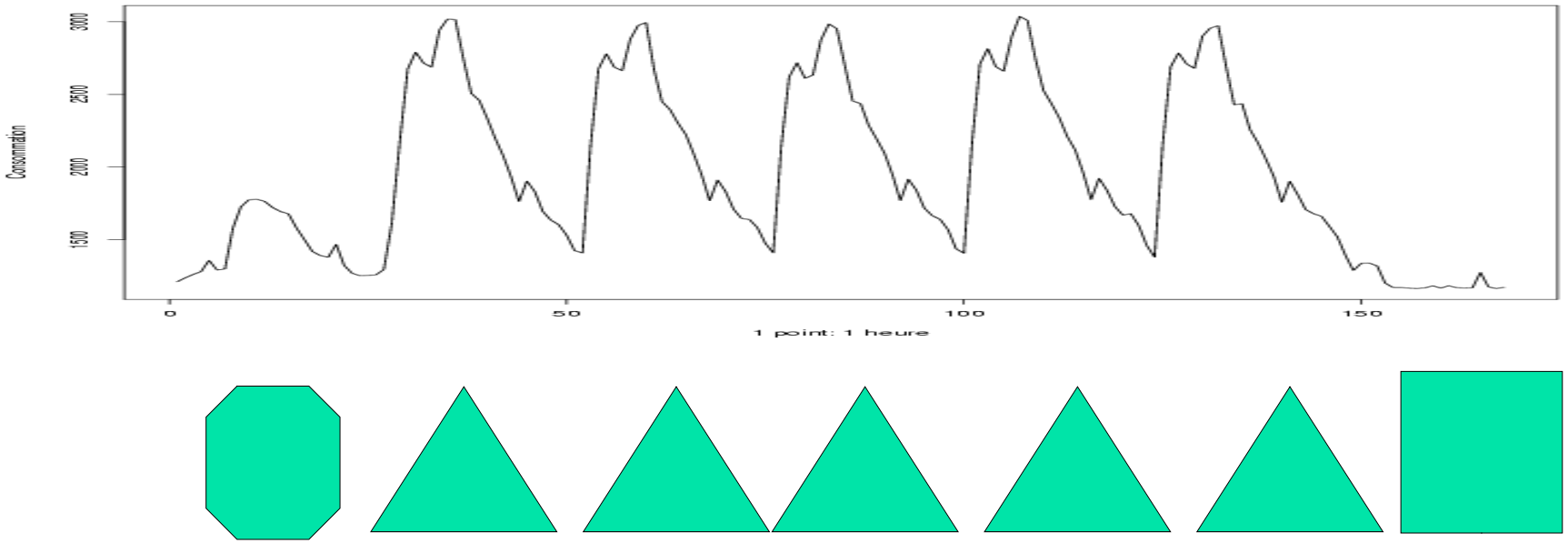
Idea:

Each segment will be represented by a symbol issued from a limited alphabet.

Method:

Defining the similarity of two segments and then performing a clustering on the time series.

Visual example



Extracted for Bernard Hugueney
works - reproducted with permission

Conclusion and Future works

- Symbolic representation help to handle and to visualize long times series.
- Improving the algorithms: online extraction of the symbolic representation.
- Finding interesting ways of mining such symbolic representation.

References

- LIP6 (Laboratoire d'informatique de Paris) http://www.lip6.fr/index-eng.html
- DBLP Bernard Hugueney http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/h/Hugueney:Bernard.html

Decrypthon: a tool for comparative proteomics

William Saurin, www.genomining.com

- **Goals**
  - A database of all pairwise similarities between proteins.
  - An annotation resource
- **Applications**
  - Identification of conserved functional domains
  - Reconstruction of families of homologous proteins
  - Metabolic pathways reconstruction

The Decrypthon task

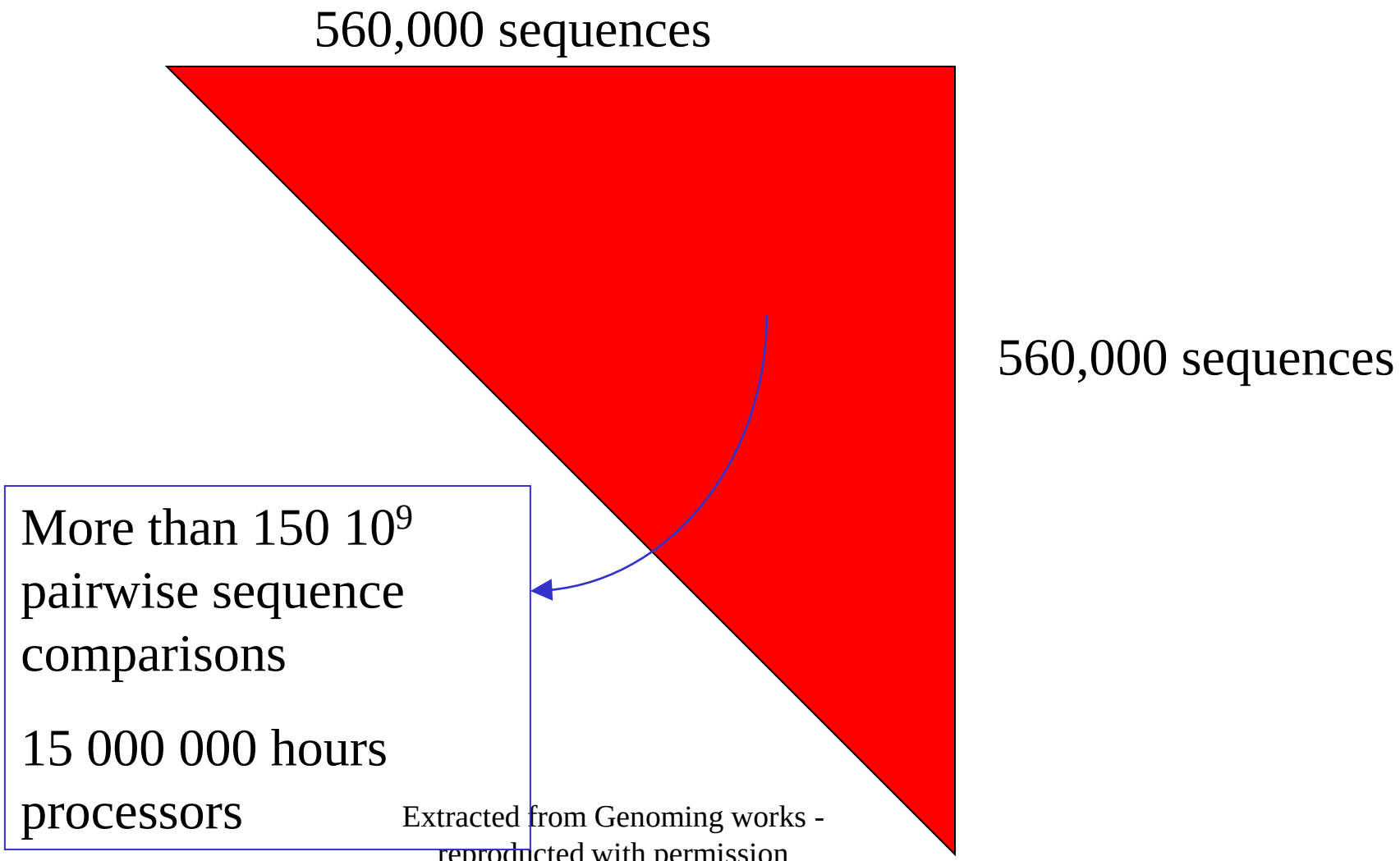
Data:

560000 proteins extracted from various well known public databases.

Protein comparison algorithm:

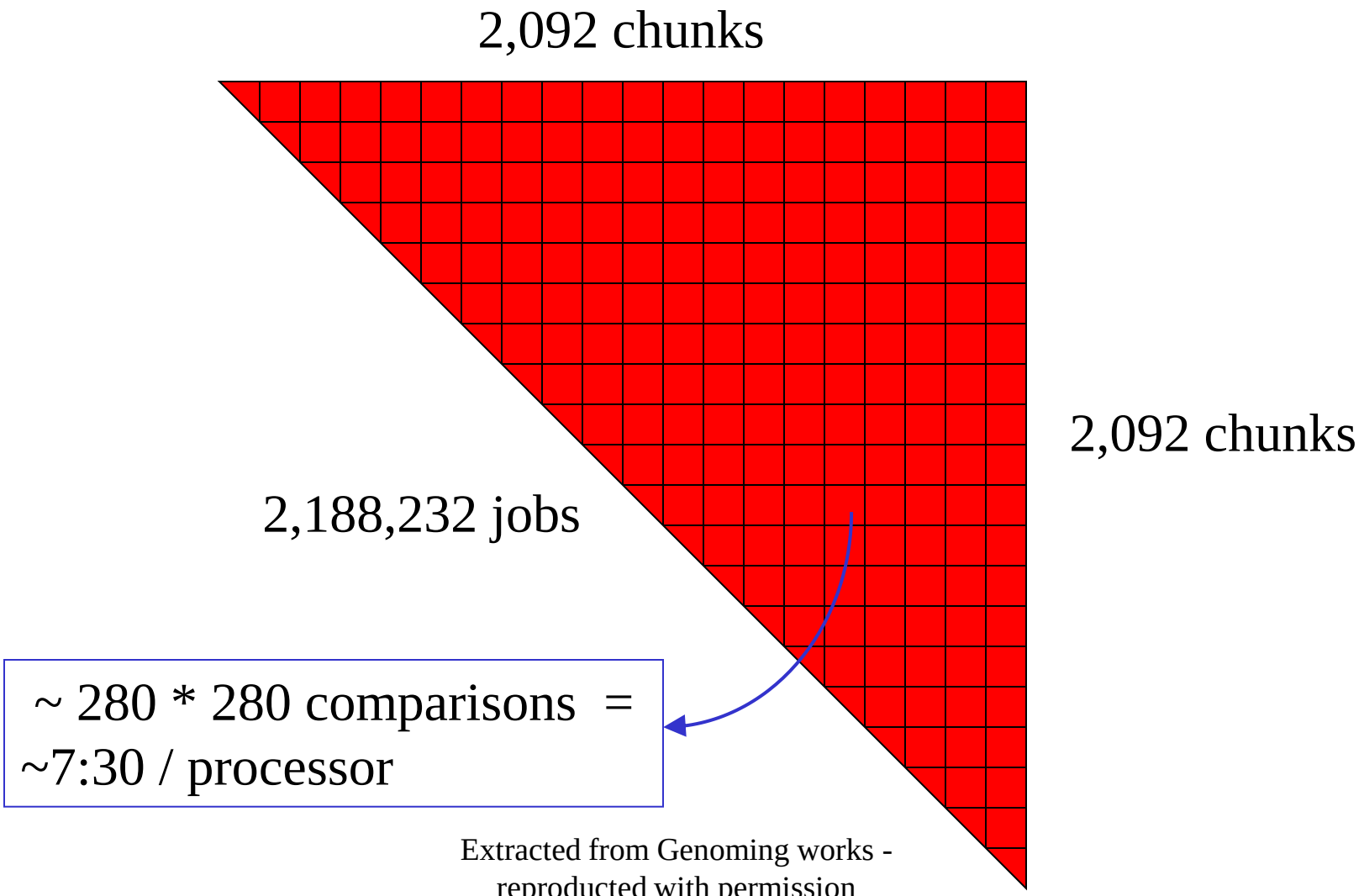
Optimal local sequence alignment (Smith and Waterman, 1981).

The computation process



Extracted from Genoming works - reproducted with permission

The computation process



Extracted from Genoming works - reproducted with permission

Distributed computation

GRID	Computing or supercomputing centers	100..1000 Stable Trust Identification of nodes
Large scale distributed system « global computation » « desktop grid »	PCs Windows or Linux	~100,000 Unstable No trust No identification

Decrypthon figures

- 560 000 sequences -> 150 10^9 pairs of sequences
- 3.3 10^{12} pairwise comparisons
- 4.6 10^{17} matrix cells
- 2.2 10^6 jobs
- 15 10^6 CPU hours
- 75 000 nodes
- 2 months
- 30 GB of raw results
- 150 10^6 Z > 8; 121 10^6 Z> 15
- ~88 10^9 matrix cell /s project
- ~532 10^9 matrix cell / s Decrypthon machine

The results have been released

- 295,000,000 reported results
- Repository : http://infobiogen.fr/services/decrypthon/
  - (17,000 files ~ 30 Go)
- The Decrypthon Search Engine

What's next?

- Conserved domain recognition
- Conserved residue identification
- Cluster of homologous proteins
- Prediction of protein-protein interactions
- Metabolic Reconstruction
- ...

Reference

http://www.genoming.com/

Other references…

Webpage of the conference [French]:

http://www.data-mining.net/jfd2003

Joannes Vermorel Home Page:

http://www.vermorel.com