

Convocatoria de ayudas de Proyectos de Investigación Fundamental no orientada

TECHNICAL ANNEX FOR TYPE A or B PROJECTS

1. SUMMARY OF THE PROPOSAL (the summary must be also filled in Spanish)

PROJECT TITLE: Symbolic Sequences: Analysis, Learning, Mining, and Evolution

PRINCIPAL INVESTIGATOR: Ricard Gavaldà Mestre (Universitat Politècnica de Catalunya)

SUMMARY

(brief and precise, outlining only the most relevant topics and the proposed objectives):

This project is the result of the cohesion created by a series of projects carried out in collaboration by the research team on the analysis and learning from data, especially those formed by symbolic sequences. Our main objectives are:

1. Development of methods for the analysis, learning and data mining, as well as the study of their fundamentals.
2. Specifically, development of techniques for data processing and learning in sequential data, structured data, and data streams.
3. Application of the developed analysis methods to sets of real data.
4. Analysis of human language, and even animal behaviour, regarded as symbolic sequences, in particular as a data stream.
5. Study of the evolution mechanisms of sequential systems, in particular for linguistic and biological communities.

The project wants to favour the generation of intelligent tools that can assist users in the tasks of manipulation and analysis of data in real contexts, especially those for which data evolve over time.

The team includes researchers with background in very diverse fields (computer science, mathematics, statistics, linguistics, physics, and biology), even more than in its predecessor projects. As a consequence, one of the project added values is the possibility to deal with problems of very diverse fields using concepts and methods from very diverse fields.

TITULO DEL PROYECTO: Secuencias Simbólicas: Análisis, Aprendizaje, Minería y Evolución.

RESUMEN

(breve y preciso, exponiendo sólo los aspectos más relevantes y los objetivos propuestos):

Este proyecto es el resultado de la cohesión creada por una sucesión de proyectos desarrollados en colaboración por el equipo investigador en materia del análisis y aprendizaje a partir de datos, especialmente los formados por secuencias simbólicas. Sus principales objetivos son:

1. Desarrollar los métodos para el análisis, aprendizaje y minería de datos, así como estudiar sus fundamentos.
2. Específicamente, desarrollar técnicas para el procesamiento de datos y aprendizaje en datos secuenciales, estructurados, y en flujos de datos (data streams).
3. Aplicar los métodos de análisis desarrollados a conjuntos de datos reales.
4. Analizar el lenguaje humano, e incluso el comportamiento animal, vistos como secuencias simbólicas, y en particular como un flujo de datos.
5. Estudiar los mecanismos de evolución de sistemas secuenciales, en particular de comunidades lingüísticas y biológicas.

El proyecto quiere potenciar la generación de herramientas inteligentes que asistan a usuarios en tareas de manipulación y análisis de datos en contextos reales, en especial aquellos en que los datos evolucionan a lo largo del tiempo.

El equipo cuenta con investigadores con experiencia en campos muy diversos (informática, matemáticas, estadística, lingüística, física y biología), incluso más que en proyectos anteriores. En consecuencia, uno de los valores añadidos del proyecto es la posibilidad de abordar problemas de áreas muy diversas con conceptos y métodos de áreas muy diversas.

2. INTRODUCTION

(maximum 5 pages)

- The introduction should include: the aims of the project; the background and the state of the art of the scientific knowledge, including the essential references; the most relevant national and international groups working in the same or related topics.
-

As is well known, the information generated, stored and accessible is experiencing an exponential growth in almost every human activity. The low cost of computer storage lets organizations keep even the last byte they generate. The increase in connectivity and the pervasiveness of Internet allows access to masses of information that keep increasing and, furthermore, are becoming more and more heterogeneous and volatile.

The notion of time is present in every process or activity of the real world. Human perception and comprehension are fundamentally based on the concept of time, and therefore the interest of formalisms and models to understand and predict from information that includes, somehow, the notion of time (or order).

In order to understand and analyze the stored information, computer help is absolutely necessary. Fields such as statistics and machine learning have provided techniques for these tasks. But in recent times, the volume of information, their volatility and heterogeneity have posed important difficulties on traditional methods. The field known as *data mining* attempts at integrating different approaches to the analysis, exploration, and explanation of data, and brings forward new techniques adapted to the volume and variety of current data. [HMS01,GaG07] for introductions to data mining.

The information to be analyzed was supposed to have a tabular form in classical statistics and machine learning, with the notable exception of time series. Today's information, on the other hand, can have the form of natural language, graph (such as the WWW or computer networks), tree (such as XML documents), of 3-D molecule, etc. The project focuses, mostly, on data that can be represented as sequence(s); this includes data generated by computer systems with or without user intervention (for example, accesses to web servers), data directly generated by human beings (such as human language), and data coming from nature (such as DNA chains). Less centrally, we will consider formalisms to study systems that evolve over time, such as the evolution of grammar systems in linguistic communities, or biological evolution. One of the benefits expected from the project is the development of a certain unified perspective concerning the computing and mathematical methods applicable to modelling and prediction in these, apparently so diverse, contexts.

Members of this team have contributed (FRESCO, MOISES, MOISES-TA, see 6.1) both to the development of models and to the development of algorithms and methods for describing and predicting sequences. Roughly speaking, in this proposal we continue our work on learning and data mining with emphasis on data stream models, we reinforce a line on analysis of language seen as a sequence (which was relatively minor in previous projects) , and we introduce a line on evolution, both on linguistic communities and biological evolution, seen as processes that evolve over time.

The nature of interactions among different topics and their temporal dependence suggests a structuring the tasks in the following six areas (admittedly, arbitrary to a certain degree):

1. Analysis, learning, and mining: fundamentals and methods
2. Analysis of sequential and structured data
3. Learning and mining in data streams.

4. Application of the analysis methods in real contexts
5. Language analysis
6. Evolution and modularity

Given the variety of approaches covered by the Project, the huge number of research groups, and the space limits, it is impossible to cover in this section all background and state of the art. We try at least to give the 2 or 3 most relevant references for each goal we propose, be them our own work or not. In some of the tasks descriptions in the Methodology section we explain in some more detail the current state of the task and the advances we will attempt in it.

Area 1. Analysis, learning, and mining: fundamentals and methods

In this area we will study several “classical” aspects of learning and data mining. By “classical” we mean that we consider data represented in tabular form, or else with no relevant order or temporal component. In some cases, the study is necessary as a previous step to other parts of the Project (for example, to have adequate evaluation tools, or to develop algorithms that will be used on real data). In others, we try to complete methods that we developed in previous projects, or make existing theoretical methods practical.

As a basis to a good part of the Project, we propose in the first place the problem of classifier comparison on massive datasets. In most of the bibliography [Dem06] it is clear that existing proposals are either based on a purely descriptive comparison of different classifiers on a variety of datasets, or else use inadequate methodologies for the study of data obtained in simulation experiments. Having adequate and reliable methods for comparing different classifiers on massive datasets is a crucial tool for the rest of the project. A first approach to the topic can be found in [CG07].

A problem that also underlies a large part of the project is model selection. Choosing one particular model among many options to describe a dataset is a very old problem in statistics, on which there is a vast literature and a renewed interest. The reason for this interest is, in part, that new techniques such as those used in this project (data mining, machine learning, etc.) require using models to explain the data they work on.

A third central problem in learning and mining is feature (or variable) selection. In fact, that is really *all* the problem to solve in many situations. In this area, we propose to develop and test a new methodology based on combining the one described in [Lóp06,Lóp07], based on Shannon’s entropy, with ideas from the field called Formal Concept Analysis, in which two Ph.D. thesis have been defended within the predecessor project, MOISES-TA [Gar06,Bai07].

Moving to precise algorithms, one that we plan to work on is the *weighted majority algorithm* [LiW89]. We have chosen it since it has been intensely studied in the literature [LiW89,Blu97,BoW02,CFH97,FSS97,MoJ03,KoM05], it is easy to implement, and one of the participants has considerable familiarity with it, having used it on a large-scale machine learning application [BeA07]. We propose to develop variants of this algorithm capable of adapting to unknown or changing environment, in a precise sense that will be made clear in the description of the task. This “data adaptivity” property will be crucial when we try to apply it in practical settings, as we do in Area 3.

In order to model and predict rapidly changing environments characterized by real-valued, multidimensional data (sensor networks or other networks with dynamic topology) we will also study some new approaches coming from algorithmic information theory. More precisely, we plan to study the use of effective dimension [Lut03], which lets us deal with storage and learning in these contexts from local information only, and therefore be more robust against environment change [LuM08].

Finally, both as a basis for other parts of the Project and because of their interest *per se*, we want to find ways of using in practice some of the clever algorithms developed in the field known as *computational learning theory* or COLT [COLT]. More details on our approach are given in the Methodology section.

Area 2. Algorithms for sequential and structured data

In this area we study methods for analyzing and learning from data whose internal structure cannot be described simply as a “set” of items, especially those with sequential structure. We do not, however, take yet into account the strong computational restrictions on time and memory imposed by the data stream model; these will be considered in Area 3. More precisely, we will consider at least:

1. time series models,
2. string mining algorithms,
3. mining of trees and other combinatorial structures, and
4. learning Markov chains, a model closely related to probabilistic automata.

Items 1, 2, and 3 continue ongoing work in the MOISES-TA Project, while item 4 is, at least on the surface, a new topic.

Time series are the oldest model with time component, and have been studied forever in statistics. In particular, within MOISES and MOISES-TA we have produced works [RPM03, MMM05], applied to climate parameter setting.

String mining are a natural evolution of traditional algorithms for string search [Gus97] towards pattern detection. The main difference with data stream algorithms (Area 3) is that while data streams are potentially infinite, the strings in this case are arbitrarily long, but finite.

We will continue the work started in MOISES and intensified in MOISES-TA on structure mining from relational and sequential data [Gar06], [BBL06,BBL07a,BBL07B,BBL07c,BBL08]. There is currently a tremendous activity on this topic within the data mining community, as demonstrated by the appearance of at least three workshops or international meetings in the last years, devoted to graph and tree mining. In this proposal we will pursue extensions to more complex combinatorial structures; see the Methodology section.

Markov chains are generative models for event sequences, used in the past to model a large variety of situations with uncertainty or probability. Recently, we have used simple versions with observable state to model web accesses [MPBGT07,PMBGT07a, PMBGT07b] and, separately, we have Developer algorithms to infer hidden-state chains based on automata-learning algorithms [GKPP06]. We propose to develop more efficient versions of the latter algorithms and apply them to the web modelling tasks, and other tasks related to autonomic computing (see Area 4). Additionally, we will apply them to the analysis of animal sequential behaviour (see Area 5).

Area 3. Data stream algorithms

In the last decade, the data stream paradigm has acquired increasing within the fields of algorithmics learning, and data mining [Mut05,Agr07,GaG07]. This model appears as an abstraction of situations where the data to be processed are not available upfront, but rather arrive in sequence with strong on time and memory restrictions. More precisely,

1. Data items arrive sequentially, one at a time.

2. Arrival speed is so large that the algorithm can only look at each item once, and must process it in real time.
3. The amount of data is potentially so large that it is not possible to keep them all; only summaries of the data seen so far can be kept.

Additionally, the following condition is also considered:

4. The statistical nature of data changes over time; that is, the data arriving at a given time may be totally different from those that arrived in the past.

Some examples of this context are the analysis of very popular Internet sites, data coming from sensors (radar, radiotelescope) or transaction processing in very large databases.

If we ignore condition 4 for the moment, the task is clearly related to data compression (although it has been studied from other viewpoints too).

Condition 4 becomes particularly important when discussing learning and data mining: one must then design algorithms that can detect change and adapt to change, both the environment and the object being modelled. This means unlearning part of what was learned in the past and relearn from current data, that is, dealing with what is known as drift, which can be gradual or sudden.

Regarding data compression, we will study data compression in data streams. A fundamental idea to explore here is that data compression can be used in learning tasks, such as prediction and classification. One of the many formulations of this intuition appeared in [LCL03,ScB06,Hit03], but needs specific developments to be applicable to any specific case. In the case of data streams, we have severely limited computation resources, so we are interested mainly in finite-memory implementations or, at most, pushdown automata. There is currently increased interest in compression by means of finite automata, given the grammar-based compression used for XML [HaS06, LeE07].

Regarding mining and learning in data streams, it is an extremely active area, involving innumerable research groups and with wide presence in journals and conferences. It was already a central topic in MOISES-TA, and there we have proposed a set of methods for induction and information extraction in data streams with drift (see references [BiG06,BiG07a,BiG07b,BCB+06] and the publication list of the Málaga group in the group description section).

We will continue work in this line, improving methods developed so far for induction of decision trees, and developing new ones for at least two new types of classifiers: ensemble methods (also known as multiclassifiers) and neural networks (which have so far received relatively little attention in the context of learning from data streams).

The precise goals will be described in the Methodology section, and here we describe only the case of the multiclassifier algorithms. These are in spirit similar to the weighted majority algorithm alluded to in Area 1, since they are based on the possibility of improving accuracy by combining several classifiers (base classifiers or experts) [HaS90]. This feature can be used to process large quantities of data (including unbounded streams) given that the algorithms that induce the multiclassifier systems can be adapted easily to process sequences of examples in an incremental way [StK01,FeG03], either one-by-one or in batches.

A possibility that will be always held in mind in this Area, and also in Areas 1 and 1, is that of parallelizing the implemented algorithms. Even though this is not set as a strict goal in this Project, given that one aims at processing very large data, it is an almost inevitable next step.

Lastly, we will draw upon methods from Computational Complexity theory to describe and investigate the limitations of algorithms working on data stream. A complexity theory for data streams is still far from developed, although within MOISES-TA we are making some (still

unpublished) progress. We propose to continue in this line, which is purely theoretical in nature.

Area 4. Application to real contexts

The algorithms developed in the previous lines will be, of course, evaluated and validated by themselves, on synthetic and/or real data. In Area 4, we will apply this, and possibly other, algorithms to real data sets, where the interest is not only testing given algorithms, but because the information or knowledge extracted is useful to an organization of the interest of the object of study

Blog analyzer

The blogosphere or set of all blogs is a subset of the Web in quick expansion, exhibiting a particularly dynamic and social behaviour, and (still) far less analyzed than the web. It can, in particular, be viewed as a planetary-scale data stream, where each new post is a new stream element.

We propose to develop a software prototype to analyze, using advanced stream data mining techniques, substantial fragments of the blogosphere, and able to personalize the results of the analysis to a particular user. We are aware this is an ambitious project, given the complexity and the volume of the software to develop and integrate, and the distributed development (Barcelona, Málaga, Valladolid) that we propose. On the other hand, it offers the best benchmark for the algorithms we will develop and the opportunity to try them on new data every day, with strong data and memory constraints.

Systems similar to those we want to develop are TextMap, TextMed and TextBiz from S. Skiena's group in Stony Brook [Ski] and blogpulse [blogpulse], by U. Toronto. Both, however, include relatively few of the techniques we want to use here. Our prototype does not attempt to compete, in principle with these systems (developed over several years by teams of software developers), but rather validate and orient our own research in massive data analysis.

Electronic mail classifier

Email classification is a domain of large interest due to increasing number of messages that even average users must read and classify daily. In this sense, there is a proliferation of tools for automatic detection of spam mails. Some tools execute in mail servers (spamassassin) and others execute in mail clients. For the, in principle different, task of mail classification, there are tools such as the MailCat adaptive classifier [SeK99], MailClassifier [Sab07], a Thunderbird extension using Bayesian filtering, and Bayesweep [Sel06], an extension for Microsoft Outlook.

Email classification is performed by analyzing email headers and bodies using text mining techniques [BrM00]. This domain satisfies the condition of a data stream: messages must be dealt with sequentially and incrementally [Man02], there is certainly the risk of concept drift, their number is unbounded, and they have to be processed in essentially real time.

Analysis of DNA sequences

The analysis of biological sequences still poses a big challenge to the scientific community. Good witness is the large number of work in this line. From the viewpoint of string analysis, the most relevant results are tied to the development of software for biological string comparison BLAST [AGM90]. More recently, work by Birzele et al [BiK06] uses mining techniques for automatic detection of discriminating features for sequence classification.

On the other hand, the evolutive study of mitochondrial DNA sequences is both extremely interesting and extremely hard [RLP07].

Web modelling and applications to autonomic computing

Recently, project members have used machine learning techniques to model ecommerce website customers [MPBGT07,PMBGT07a,PMBGT07b], and studied their possible use in other facets of so-called autonomic computing [KeC03]. The ultimate goal of autonomic computing is the development of computer systems that can dynamically adapt to the work at hand, as well as reconfigure and repair themselves with little or no human intervention. There is an agreement that this task requires that they are capable of introspection and learning from their own behavior.

More precisely, we study data provided by project EPOS (external institutions or companies) atrapalo.com and Fundació puntCAT for modelling purposes (as well as benchmarks for our algorithms from Areas 1, 2, 3). Note that we have already used data from atrapalo.com in our previous work [PMBGT07a,PMBGT07b]. We believe that better results can be obtained using our improved Markov-chain algorithms and using “data stream”-oriented algorithms.

Prediction and diagnosis for kidney-transplant related diseases

This data was already used [Lóp06,Lóp07] as a benchmark for the recommended system described there. The methodology combining [Lóp06,Lóp07] and Formal Concept Analysis ideas developed in Area 1, will be applied again in this context, hopefully to obtain better diagnostic results.

There are informal contacts with at least two hospitals in the area of Barcelona (one of them included as EPO), that could be interested in trying our analysis methods to medical data.

Area 5. Language analysis

Among symbolic sequences, human language, communication systems of other species, and animal sequential behaviour form a very Wide field. In the case of human language, we propose to continue with research in previous project (FRESCO, MOISES and MOISES-2) on type logical grammars. Concerning animal sequential behaviour (e.g. [SBT06]), we will continue studies on its complexity by improving on techniques already used [FCL06] or by means of disciplines that, we relieve, have had little application in this context, such as data mining.

For human language, we will continue research in previous editions (FRESCO, MOISES and MOISES-2) on *Mildly Context Sensitive Grammars*, which capture very different features than logical grammars.

Another aspect that we will study is the appearance of linguistic features in a linguistic community, that is, cognitive processes by which a community can acquire specific aspects of language and its meaning. The proposed methodology uses the notion of language game [Wit53] as fundamental mechanism of linguistic interaction among agents, and between agents and their environment. Experimental models of such language games have tried to elucidate how speakers can eventually share simple lexicon, syntax, meanings after a reasonable interaction time [Ste98,Ste04,Gol05,Val05], [Sie06a,Sie06b,Sie07a,Sie07b]. We will define, implement, and experimentally evaluate more complex variants of language games than those described in the literature. Additionally, we will attempt an analysis of convergence time and conditions using tools from computational learning theory (PAC model).

Area 6. Evolution and modularity

There is abundant literature on how, taking natural selection as inspiration, produce solutions to problems (e.g., learning or optimization problems). The idea is to create a solution

population and, generation after generation, reproduce, mutate and select them using some fitness measure. See [Koz92] as a reference.

Recently, researchers have observed that the evolutive process has serious difficulties with the most important aspect, which is generating variation. The concept of *modularity* [AnP93,DoM03,HHL99,KiG98] is being promoted as to essential ingredient within evolutive biology. Living organisms have built genotypes that encode biological forms in modular ways: vary certain parts of the genotype, and only very specific parts of the phenotype are altered. Different phenotypical traits correspond to well-isolated parts of the genotype – there are minimal interdependencies. By contrast, engineers deliberately and explicitly use modular organization in most solutions to the complex problems they solve.

Therefore, we will attempt to develop a framework that explains how modularity can emerge as a feature that enables faster and most powerful artificial evolution.

Bibliography

NOTE: to avoid using unnecessary space, we do not duplicate here publications by group members listed in Section 6 “Background of the group”.

- [AGM90] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W. and Lipman, D.J.: *Basic local alignment search tool*, Journal of Molecular Biology, Vol. 215, No. 3, (1990) 403–410
- [Agr07] Charu C. Aggarwal: *Data Streams: Models and Algorithms*. Springer, 2007
- [AMS07] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A.I.: *Fast discovery of association rules*, In Proc. Advances in Knowledge Discovery and Data Mining, 307–328
- [AnP93] Angeline, P.J. and Pollack, J.: *Evolutionary Module Acquisition*, In Proc. of the Second Annual Conference on Evolutionary Programming, (1993)
- [AuBeBo07] J. Autebert, J. Berstel, and L. Boasson. Context-free languages and pushdown automata. In: G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, volume 1, Word, Language, Grammar, pages 111–174. Springer-Verlag, 1997.
- [Bai04] Jaume Baixeries. A Formal Concept Analysis framework to model functional dependencies. Conference Mathematical Methods for Learning. Como (Italia), 2004.
- [BB03] José Luis Balcázar and Jaume Baixeries. Discrete Deterministic Data Mining as Knowledge Compilation. Workshop on Discrete Mathematics and Data Mining, associated to the 3rd SIAM Intl. Conference on Data Mining, 2003.
- [BB05a] Jaume Baixeries and José Luis Balcázar. Characterization and Armstrong relations for Degenerated Multivalued Dependencies using Formal Concept Analysis. Proc. 3rd Intl. Conference on Formal Concept Analysis-ICFCA'05. Springer LNCS series (LNCS 3403), 2005.
- [BB05b] Jaume Baixeries and José Luis Balcázar. New Closure Operators and Lattice Representations for Multivalued Dependencies and Related Expressions. The 3rd international conference on Concept Lattices and Their Applications (CLA'05), 2005.
- [BB05c] Jaume Baixeries and José Luis Balcázar. A Lattice Representation of Relations, Multivalued Dependencies and Armstrong Relations. 13th International Conference on Conceptual Structures (ICCS '05), 2005.
- [BB06] Jaume Baixeries and José Luis Balcázar. Unified characterization of symmetric dependencies with lattices. The fourth International Conference in Formal Concept Analysis, 2006.
- [BBL06] José Luis Balcázar, Albert Bifet and Antoni Lozano. Intersection Algorithms and a Closure Operator on Unordered Trees. International Workshop Mining and Learning with Graphs MLG 2006. Berlin, Germany, 2006.
- [BCB+06] Manuel Baena-García, José del Campo-Ávila, Raúl Fidalgo, Albert Bifet, Ricard Gavaldà and Rafael Morales-Bueno. Early Drift Detection Method. ECML PKDD Workshop on Knowledge Discovery from Data Streams 2006. Berlin, Germany, 2006.
- [BeA07] Becker, H. and Arias, M.: *Real-time Ranking with Concept Drift Using Expert Advice*,

- Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.
- [BiG06] Bifet, A., Gavaldà R.: *Kalman Filters and Adaptive Windows for Learning in Data Streams*. Proc. 9th International Conference on Discovery Science (DS 2006). Springer-Verlag Lecture Notes in Artificial Intelligence 4265, 29-40.
- [BiK06] Birzele, F. and Kramer, S.: *A new representation for protein secondary structure prediction based on frequent patterns*, Bioinformatics, Vol. 22, No. 24, (2006) 2628–2634
- [blogpulse] Web del proyecto blogpulse: <http://www.blogpulse.com/>
- [Blu97] Blum, A.: *Empirical Support for Winnow and Weighted-majority Algorithms: Results on a Calendar Scheduling Domain*, Machine Learning, Vol. 26(1), (1997) 5–23
- [BoW02] Bousquet, O. and Warmuth, M.K.: *Tracking a Small Set of Experts by Mixing Past Posteriors*, Journal of Machine Learning Research, Vol. 3, (2002) 363–396
- [BrM00] Brutlag, J.D. And Meek, C.: *Challenges of the Email Domain for Text Classification*, In Proc. of the Seventeenth International Conference on Machine Learning, (2000) 103–110
- [Bai07] Jaume Baixeries. Lattice characterization of Armstrong and Symmetric Dependencies. Tesis doctoral, Universitat Politècnica de Catalunya, 2007.
- [CC05] Cabaña, A., Cabaña, E. M.: Goodness-of-fit to the Exponential Distribution, focused on Weibull alternatives, Communications in Statistics. Simulation and Computation 34, 711-723, 2005.
- [CFH97] Cesa-Bianchi, N.; Freund, Y.; Haussler, D.; Helmbold, D.P.; Schapire, R.E. and Warmuth, M.K.: *How to Use Expert Advice*, Journal of the ACM, Vol. 44, No. 3, (1997) 427–485
- [CG07] A. Cabaña y F. Gamboa, Towards a Global method for comparing Classifiers. Actas del XXX Congreso Nacional de EIO, ISBN 978-84-690-7249-3 (2007).
- [CQ05] Cabaña A., Quiroz A. J. Using the empirical moment generating function in testing for the Weibull and the type I extreme value distributions. Test 14, N.2, 417-431 (2005).
- [CT04] Clark, A. and Thollard, F. PAC-learnability of Probabilistic Deterministic Finite State Journal of Machine Learning Research, 5. 2004.
- [COLT] *COLT: Computational Learning Theory*. Repositorio de información sobre la Teoría del Aprendizaje Computacional en <http://www.learningtheory.org>
- [Dem06] Demšar, J.: *Statistical Comparison of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research, Vol. 7, (2006)
- [DoM03] Dorogovtsev, S.N. and Mendes, J.F.F.: *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, (2003)
- [FeG03] Fern, A. and Givan, R.: *Online Ensemble Learning: An Empirical Study*, Machine Learning, Vol. 53, (2003) 71–109
- [Fer06] Ramon Ferrer i Cancho. Why do syntactic links not cross? Europhysics Letters 76, 1228-1235, 2006.
- [FKH05] Fischer, J.; Kramer, S. and Heun, V.: *Fast frequent string mining using suffix arrays*, In Proc. of Fifth IEEE International Conference on Data Mining, (2005) 609–612
- [FCL06] Ramon Ferrer i Cancho and David Lusseau. Long-term correlations in the surface behavior of dolphins. Europhysics Letters 74, 1095-1101, 2006.
- [FSS97] Freund, Y.; Schapire, R.E.; Singer, Y. and Warmuth, M.K.: *Using and combining predictors that specialize*, In Proc. of the twenty-ninth annual ACM symposium on Theory of computing, (1997) 334–343
- [FTT04] Flake, G.W., Tarjan, R.E., Tsioutsoulis, K. *Graph Clustering and Minimum Cut Trees*, Internet Mathematics, (2004).
- [GaG07] Gama, J. and Gaber, M.M. (eds.): *Learning from Data Streams*. Springer, (2007)
- [Gar06] Gemma C. Garriga. Formal Methods for Mining Structured Objects. Tesis doctoral. Universitat Politècnica de Catalunya, 2006.
- [Gir87] Girard, Jean-Yves: 1987, 'Linear Logic', Theoretical Computer Science 50, 1--102.
- [GKPP06] Ricard Gavaldà, Phillip W. Keller, Joelle Pineau, and Doina Precup. PAC-learning of Markov models with hidden state. In Proc. 11th European Conference on Machine Learning (ECML 2006). Springer Lecture Notes in Artificial Intelligence 4212, 150-161, 2006.
- [Gol05] Goldberg, A. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press, (2005).

- [GrG05] Grodner, D. & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science* 29, 261-291.
- [Gus97] Gusfield, D.: *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, (1997)
- [HaS06] Hariharan, S. and Shankar, P.: *Evaluating the role of context in syntax directed compression of xml documents*, In Proc. of the 2006 IEEE Data Compression Conference (2006)
- [HaS90] Hansen, L. K., Salamon, P.; *Neural Network Ensembles*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, (1990) 993–1001
- [HHL99] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W.: *From Molecular to Modular Cell Biology*, *Nature*, Vol. 402, C47–52, (1999)
- [Hit03] Hitchcock, J.M.: *Fractal dimension and logarithmic loss unpredictability*. *Theoretical Computer Science*, Vol. 304, No. 1-3, (2003) 431–441
- [HMS01] Hand, D.; Mannila, H. and Smith P.: *Principles of Data Mining*. The MIT Press, (2001)
- [HoI06] Holmes, M. and Isbell, C.; *Looping Suffix Tree-Based Inference of Partially Observable Hidden State*. In Proc. of ICML, (2006).
- [JZK05] Jaeger, H., Zhao, M. and Kolling, A.; *Efficient estimation of OOMs*. Proceedings of NIPS, (2005).
- [Kas70] Kasai, T.: *A Hierarchy Between Context-Free and Context-Sensitive Languages*, *J. Comput. Syst. Sci.*, Vol 4, No. 5, (1970) 492–508
- [KeC03] Kephart, J. and Chess, D.: *The Vision of Autonomic Computing*, IEEE Computer, (2003)
- [KiG98] Kirschner, M. and Gerhart, J.: *Evolvability*, In Proc. Natl. Acad. Sci. USA, Vol. 95, (1998) 8420–8427
- [Ko03] Ko, P., Aluru, S.; *Space efficient linear time construction of suffix arrays*, *Lecture Notes in Computer Science*, Vol. 2676, (2003) 203-210
- [Ko05] Ko, P. and Aluru, S.: *Space efficient linear time construction of suffix arrays*, *Journal of Discrete Algorithms*, Vol. 3, No. 2-4, (2005) 143–156
- [KoM05] Kolter, J.Z. and Maloof, M.A.: *Using Additive Expert Ensembles to Cope with Concept Drift*, In Proc. of the 22nd international conference on Machine Learning, (2005) 449–456
- [Koz92] Koza, J.R.: *Genetic Programming. On the Programming of Computers by Means of Natural Selection*. The MIT Press, (1992)
- [LCL03] Li, M.; Chen, X.; Li, X, Ma, B.; Vitányi, P.: *The similarity metric*, In Proc. of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (2003), 863–87
- [LeE07] League, C., Eng, K.; *Type-based compression of xml data*, In Proc. of the 2007 IEEE Data Compression Conference, (2007) 272–282
- [LeZ78] Lempel, A. and Ziv, J: *Compression of individual sequences via variable rate coding*, In Proc. IEEE Transaction on Information Theory, Vol. 24, (1978) 530–536
- [LiW89] Littlestone, N. and Warmuth, M.K.: *The weighted majority algorithm*, In Proc. of IEEE Symposium on Foundations of Computer Science, (1989) 256–261
- [Lóp06] Josefina López Herrera. *A New Approach of Shannon Entropy in Recommender Systems*. Artificial Intelligence and Soft Computing. Iasted/Acta Press, 22-27, 2006.
- [LuM08] J.H. Lutz and E. Mayordomo. *Dimensions of points in self-similar fractals*. Enviado.
- [Lóp07] López Herrera, J.: *Shannon Entropy for disease forecasting: A Case Study on Kidney Graft Failure Prediction*, In Proc. Artificial Intelligence and Soft Computing, (2007) 203–204
- [Lut03] J. H. Lutz. *The dimensions of individual strings and sequences*. *Information and Computation*, 187:49--79, 2003.
- [Lut05] Lutz, J. H.: *Effective fractal dimensions*, *Mathematical Logic Quarterly*, Vol. 51, (2005) 62–72
- [Man02] Manco, G.; Masciari, E.; Ruffolo, M. and Tagarelli, A.: *Towards an Adaptive Mail Classifier*, In Proc. AIIA, (2002)
- [Man93] Manber, U. and Myers G.: *Suffix arrays: a new method for on-line string searches*, *SIAM Journal on Computing*, Vol. 22, No. 5, (1993) 935–948
- [May02] Mayordomo, E.: *A Kolmogorov Complexity Characterization of Constructive Hausdorff Dimension*, *Information Processing Letters*, Vol. 84, No. 1, (2002) 1–3

- [MF05] Glyn Morrill and Mario Fadda. Proof nets for basic discontinuous Lambek calculus. Proc. Lambda Calculus, Type Theory and Natural Language, LCTTNL05, King's College, London, 1—16, 2005.
- [MF08] Morrill, G. and Fadda, M. (en prensa). Proof nets for basic discontinuous Lambek calculus. Logic and Computation.
- [MMM05] Mora-López, L.; Mora, J.; Morales-Bueno, R.; Sidrach-de-Cardona, M.: *Modelling time series of climatic parameters with probabilistic finite automata*, Environmental Modelling and Software, Vol. 20, No. 6, (2005) 753–760
- [MMR02] Manco, G.; Masciari, E.; Ruffolo, M. and Tagarelli, A.: *Towards an Adaptive Mail Classifier*, In Proc. AIIA, (2002)
- [MoJ03] Monteleoni C. and Jaakkola, T.: *Online Learning of Non-stationary Sequences*, In Proc. of Advances in Neural Information Processing Systems, (2003)
- [Mor00] Morrill, G. (2000). Incremental processing and acceptability. Computational Linguistics 25 (3), 319-338.
- [Mut05] Muthukrishnan, S.: Data Streams: Algorithms and Applications. Foundations and Trends in Theoretical Computer Science, Now Publishers Inc, (2005)
- [PBD06] Josep M. Pujol, Javier Béjar, and Jordi Delgado. *Clustering algorithm for determining community structure in large networks*. Physical Review E 74, 016107, 2006.
- [PDS05] Josep M. Pujol, Jordi Delgado, Ramon Sanguesa and Andreas Flache. *The Role of Clustering on the Emergence of Efficient Social Conventions*. Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05 (Edinburgh, Scotland), pp. 965-970, 2005.
- [PFDS05] Josep M. Pujol, Andreas Flache, Jordi Delgado and Ramon Sanguesa. *How can Social Networks ever become Complex? Modelling the Emergence of Complex Networks from Local Social Exchanges*. Journal of Artificial Societies and Social Simulation (JASSS), 8(4), 2005
- [PSD03] Josep M. Pujol, Ramon Sanguesa and Jordi Delgado. *A Ranking Algorithm based on Graph Topology to generate Reputation or Relevance*. In *Web Intelligence (N. Zhong, J. Liu & Y.Y. Yao, eds.)*, pp. 380-394, Springer Verlag, 2003.
- [RGP06] Rodrigues, P., Gama, J., Pedroso, J.P.; *ODAC: Hierarchical Clustering of Time Series Data Streams*. Proceedings of the Sixth SIAM International Conference on Data Mining, (2006) 499–503
- [RGT04] Rosencrantz, M., Gordon, G. and Thrun, S.: *Learning Low Dimensional Predictive Representations*. Proceedings of ICML, (2004).
- [RLP07] Ruiz-Pesini, E.; Lott, M.T.; Procaccio, V.; Poole, J.; Brandon, M.C.; Mishmar, D.; Yi, C.; Kreuziger, J.; Baldi, P. and Wallace, D.C.: *An enhanced MITOMAP with a global mtDNA mutational phylogeny*, Nucleic Acids Research, Vol. 35, (2007) 823–828
- [RPM03] Ramírez Santigosa, L.; Polo Martínez, J.; Mora-López, L.; Sidrach de Cardona, M. and Blanco Gálvez, J.: *Fuzzy Inference Systems Applied to the Daily Ultraviolet Radiation Evaluation*, Solar Energy, Vol. 75, No. 6, (2003) 447–454
- [Sab07] Sabellico, E.: *Mailclassifier website*, Website, (2007)
- [SBT06] Suzuki, R.; Buck, J. and Tyack, P.: *Information entropy of humpback whale songs*, Journal of the Acoustical Society of America, Vol. 119, (2006) 1849–1866
- [ScB06] Sculley, D. and Brodley, C.E.: *Compression and Machine Learning: A new perspective on feature space vectors*, In Proc. of the Data Compression Conference, (2006) 332–341
- [SeK99] Segal, R. and Kephart, J.O.: *Mailcat: An intelligent assistant for organizing e-mail*. In Proc. International Conference on Autonomous Agents, (1999) 276–282
- [Sel06] Selenics: Bayesweep. Website, 2006. <http://www.bayesweep.html>
- [ShM00] Shi, J. and Malik, J.: *Normalized Cuts and Image Segmentation*, IEEE Transactions on Pattern analysis, (2000)
- [Sie06a] Josefina Sierra Santibáñez. Propositional logic syntax acquisition. Symbol Grounding and Beyond, Lecture Notes in Computer Science. Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, volume 4211, pag. 128-142, 2006.
- [Sie06b] Josefina Sierra Santibáñez. *Propositional logic syntax acquisition using induction and self-organisation*. Papers from the AAAI (American Association for Artificial Intelligence) Fall Symposium on Interaction and Emergent Phenomena in Societies of Agents. Technical Report FS-

06-05, AAAI Press, pag. 74-81, 2006.

- [SLJ03] Singh, S., Littman, M. L., Jong, N. K., Pardoe, D. and Stone, P.; *Learning Predictive State Representations*. Proceedings of ICML, (2003).
- [Ste04] Steels, L.: *Constructivist Development of Grounded Construction Grammars Scott*, In Proc. Annual Meeting Association for Computational Linguistic Conference, (2004) 9–19
- [Ste98] Steels, L.: *The origins of syntax in visually grounded robotic agents*, Artificial Intelligence, Vol. 103, No. 1-2, (1998) 133–156
- [StK01] Street, W.N. and Kim, Y.: *A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification*, In Proc. of the 7th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, (2001) 377–382
- [Ukk95] Ukkonen, E.: *On-line construction of suffix trees*, Algorithmica, Vol. 14, No. 3, (1995) 249–260
- [Val05] Van Valin, R.D. Jr.: *The Syntax-Semantics-Pragmatics Interface: An Introduction to Role and Reference Grammar*, Cambridge: Cambridge University Press, (2005).
- [SBT06] Suzuki, R., J. Buck, & P. Tyack. (2006). Information entropy of humpback whale songs. *Journal of the Acoustical Society of America*, 119, 1849-1866.
- [Ski] Página web de S. Skiena : <http://www.cs.sunysb.edu/~skiena/#projects>
- [Wei00] Weihrauch, K.: *Computable Analysis*. Springer-Verlag, (2000)
- [Wit53] Wittgenstein, L.: *Philosophical Investigations*. Blackwell Publishing, (2001)

3. OBJECTIVES

(maximum 2 pages)

- ◆ **3.1** Describe the reasons to present this proposal and the **initial hypothesis** which support its objectives (maximum 20 lines)

The main initial hypothesis is that there is a need for powerful and generally applicable techniques that can summarize, analyze, classify, and predict the large datasets generated in many activities, in order to make decisions that can be adapted to future evolutions. This data generation process often results in a continuous flux whose storage and processing can use only a small amount of resources with respect to the amount of received information.

On the theoretical side, we consider useful and necessary to have formal developments to guarantee that a learning strategy will achieve correct results in a variety of different situations, and not only in the application where it was developed. This a-priori analysis is especially crucial in the little-studied context of scarce computational resources. Furthermore, from a deep understanding of the reasons why a strategy works, one often finds ways to extend or improve it in practice.

On the applications side, the possibilities that open up in fields of immediate interest are enough motive to study machine learning on sequential data. Some such fields are the analysis of computer networks, social communication networks, web accesses, bioinformatics (and specifically, genomics and genetic disease prediction).

The second hypothesis is that given the diversity of methods and algorithms imported from more restricted contexts and that are potentially useful, there is an imperious need to systematize the comparison, evaluation, and selection of algorithms for a single problem on very large datasets.

- ◆ **3.2.** Indicate the **background and previous results** of your group or the results of other groups that support the initial hypothesis

The background that supports the previous hypothesis can be derived from the text and bibliography in the Introduction. Previous results from the team belong mostly to the predecessor projects: FRESCO (2000-2002), MOISES (2003-2005), and MOISES-TA (2006-2008), or else are natural extensions to slightly more complex scenarios. Precise references can be found in Section 6, Background of the group, and group CVs. Our expertise includes:

- Strong background on goodness-of fit and first step towards classifier comparison techniques on large datasets
- Data mining algorithms, and in particular on sampling techniques in data mining.
- Episode mining on event sequences, mining of combinatorial structures in sequential and relational data. Application of Formal Concept Analysis in mining.
- Recommender systems.
- Work on theoretical aspects of learning and computational complexity theory
- Time series
- Markov models and inference.
- XML text compression and compression by pushdown automata.

- Learning and mining in data streams
- Relations about measure-theoretic notions (Hausdorff dimension) and information theory.
- Logical and categorial grammars.
- Tools for theoretical language analysis and animal behavior analysis.
- Discrete dynamical systems and evolutive algorithms.
- Identification of mutations in mitochondrial DNA.
- Applied research on learning and data mining:
 - Music generation and prediction,
 - Pictorial style generation,
 - Pattern discovery in gender violence data
 - Pattern discovery in work disability data.

◆ **3.3.** Describe briefly the **objectives** of the project.

1. Develop methods for model comparison, evaluation, and selection that can be used for in algorithms for explaining, classifying, and predicting massive datasets. Apply them specifically to multiclassifier (ensemble) methods.
2. Study the fundamentals and design methods for analysis, learning, and mining data: Develop parallel systems for time series processing. Develop and test a new methodology for recommender systems using tools from FCA.
3. Develop further methods for analysis and mining in sequential and structured data: Develop more efficient algorithms for inferring Markov models and probabilistic automata from data. Algorithms for mining frequent patterns and combinatorial structures from data. Compression, online prediction, and dimension of geometrical data and XML.
4. Develop techniques and theoretical foundations of learning in Data Streams. Study the quality of generated models by learning algorithms and develop algorithms that can learn in time-changing data. Formalize data streams as a computation model.
5. Apply the methods in real contexts. Development of a blog analyzer; a tool for assisting in email classification, grouping, classifying, and prediction of genomic sequences; and analysis of web data and applications of data mining to autonomic computing.
6. Study human and animal language as sequential processing. Study and extension of grammar models and language games.
7. Study of evolutionary mechanisms in systems of agents involving sequences, in particular linguistic communities and biological evolution, especially the concept of modularity.

◆ **3.4. For Coordinated projects** only, the **coordinator** must indicate (maximum 2 pages):

- the global objectives of the coordinated project, the need for coordination, and the added value provided by this coordination
- the specific objectives of each subproject
- the interaction among the objectives, activities and subprojects
- the mechanisms of coordination for an effective execution of the project.

The overall objective of the coordinated project is the theoretical and practical study of algorithms and methods that can describe, represent, and predict data, mostly sequential and massive data, including the case in which they arrive at high speed and varying over time. This includes developing reliable methods for comparing, evaluating, and selecting those models.

Each of the variants is an abstraction of real situations, or a specific real situation, in which different aspects are emphasized.

The aforementioned global goal can be now expanded in several tasks:

- Defining model classes of interest
- Theoretical study of their features
- Design and analysis of algorithms for those models.
- Algorithm implementation.
- Evaluation and experimentation with synthetic and real datasets.

Each of the participating teams has slightly different expertises to contribute to this goal and these tasks. Thus, the Barcelona team has good experience in computational learning theory and complexity theory, sampling techniques for data mining algorithms, episode mining in event sequences, logical and gramatical models of language, and evolutive and dynamical systems. This experience is complemented by the Zaragoza team, also with experience in computational complexity, but highly specialized in dimension theory and compression algorithms, which can be seen as ways to obtain succinct representations in large masses of data; A. Cabaña, a recognized expert in goodness-of-fit problems, contributes the knowledge on model evaluation and selection methods. The Málaga group, in turn, specializes in Knowledge Discovery on the basis of classical methods such as decision trees and regresión trees, association rule mining, time series, behavior pattern extraction, as well as their experience on application to real data.

This enumeration suggests a working mode in which the more theoretically-oriented teams suggest formalisms, methods, and models (possibly together with analysis) and more practically-oriented teams provide implementations and experimental work.

The core of the group has been working along these lines in for at least the last 9 years. The group, which has been noticeably expanded for this proposal, contains people from a variety of disciplines (computer science, mathematics, statistics, linguistics, physics, and biology), even more than in previous editions. One of the added values of the project will be the possibility of facing problems from very different areas with very different concepts and methods

In all of the seven objectives listed in 3.3 there are researchers from more than one subproject. The following table indicates the relation among tasks (to be described in section 4, Methodology), the objectives to which it contributes, and the subprojects involved in them. It can be observed that there are many tasks in which two, and even three, subprojects are involved.

Task	Contributes to objectives	Barcelona	Málaga	Zar-Vall
EVAL	1,2		x	x
SELEC	1,2			x
COLT	2	x		
PATR	3	x		
TEMP	2		x	
STRING	3		x	

MARKOV	2	x		x
ENSEMB	2	x	x	
COMPLEX	4	x		x
COMPRESS	4			x
DIM	3			x
RECOM	2	x		
CAMBIO	4		x	
BLOG	5	x	x	x
CORR	5		x	
DNA	5		x	x
AUTON	5	x		
GRAMLOG	6	x		
GAMES	6, 7	x		
COMANIM	6	x		
MCSG	6			x
EVOL	7	x		x

We contemplate the following coordination mechanisms:

- Email distribution lists, to communicate news and most relevant results with immediate effect.
- Web site to publicize results, work and news from the project. Intranet to share work in progress, internal documents, and share data and code.
- Two yearly meetings: A workshop to be held in the subproject sites, where researchers can explain their finished work discuss ongoing work, and plan for the next period; and at least one yearly virtual meeting.
- Regular meetings (at least every 6 months) of the subproject IPs.

4. METHODOLOGY AND WORKING PLAN

(in the case of coordinated projects this title must include all the subprojects)

Detail and justify precisely the methodology and the working plan. Describe the working chronogram.

- ◆ The working plan should contain the tasks, milestones and deliverables. The projects carried out in the Hesperides or in the Antarctic Zone must include the operation plan.
 - ◆ For each task, it must be indicated the Centre and the researchers involved in it.
 - ◆ If personnel costs are requested, the tasks to be developed by the personnel to be hired must be detailed and justified. Remember that personnel costs are eligible only when personnel is contracted, **fellowships are not eligible** as personnel costs.
-

METHODOLOGY

For most of the Project, the methodology is the usual one in the area of algorithmics and data mining research. For each of the problems addressed, the steps will be:

- Theoretical study of the problem, and study of relevant literature, if any
- Development of the algorithms needed to solve it, and implementation
- Experimental validation, with synthetic and/or real data.

These steps will be repeated in most of the following tasks. The main exceptions will be the tasks with essentially a theoretical nature, mostly the study of effective dimension (Area 1), the development of a complexity theory for data streams (Area 3), and the study of human language via Type Logical Grammars (Area 5). In Area 4, where the goal is to analyze given (real) data sets, it may not always be necessary to have a previous theoretical development or even developing new algorithms.

WORKPLAN

We structure our work in a series of tasks that can involve researchers from one or, often, several subprojects. These are listed next.

Coordination and dissemination tasks

COORD: Coordination, direction, and document generation

[Barcelona: R. Gavaldà; Málaga: R. Morales, ZarVall: E. Mayordomo]

Carried out throughout the project's entire life. Persons in charge are the subproject IP's.

DISEM-WEB: Dissemination and visibility of results.

[Barcelona, Málaga, ZarVall: Subproject IP's]

We will maintain an information repository with two levels of visibility: restricted to share information among Project members only. Public, so that the rest of the scientific community and society at large can inspect the results achieved in the project. In particular, a web will be created with publications, news, and other indicators of Project progress.

We will take special care that EPOs are regularly and timely informed of the Project, particularly those that provide data for analysis.

All project members participate in this task although supervision will be in charge of subproject IP's.

Tasks related to developing theoretical tools and algorithms for data analysis, learning, and mining

EVAL. Classifier comparison and evaluation methods on massive datasets.

[ZarVall: A. Cabaña, R. San Martín; Málaga: J. del Campo, R. Fidalgo]

The goal is to develop methods suitable for comparing the behavior of different algorithms that can be applied to solve a given problem over large datasets.

A closely, and very important problem related with data streams is assessing the quality of the models created from that information. This is a novel line and, as can be guessed, should be applicable to many other parts of the project. During the predecessor MOISES-TA, the Málaga and Valladolid groups started an approximation to this project. In the present proposal, one of the methodologies in development should be finished.

Results of this task should be applicable to all other tasks dealing with classifier models.

SELEC: Process transformation applied to model selection.

[ZarVall: A. Cabaña, R. San Martín]

We hope to create improved solutions to the problem of model identification or selection. Our approach is goodness-of-fit tests. The features we attempt to improve are of two kinds: On the one hand, obtaining tests that are especially suitable to detect deviation from a certain hypothetical model which is interesting to the user. On the other hand, the use of m statistical measures whose approximate distribution for large data is always the same, no matter the model to which we are trying to fit, and that, furthermore, does not depend on the need of estimating the parameters when these do appear in the model being tested [CC05,CG07,CQ05].

COLT. Applicability of computational learning algorithms to practical cases

[Barcelona: M. Arias, R. Gavaldà]

Both as help to other tasks and because of their intrinsic interest, we want to study the practical applicability of theoretical algorithms developed within the field known as *computational learning theory* or COLT [COLT]. This field aims at providing solid foundations for the study of learning processes and was well represented in the FRESCO and MOISES projects. The COLT community has developed a good number of clever algorithms using queries to a supervisor in order to learn logical formulas or similar formalisms (see the bibliography [COLT]). The main problem for the application of these algorithms is the lack, in practice, of an oracle that can answer queries that is, that can act as a supervisor from the available data.

In this task we try to develop a methodology for the application of algorithms from COLT algorithms in practical context. The idea is to use models obtained by other methods as “black

box” oracles to solve queries. A theoretical study should determine first which hypothesis classes should be amenable to this methodology.

One of the best known algorithms originating in the COLT community is the *weighted majority algorithm* [LiW89]. This is an "on-line" algorithm that assumes the existence of N experts whose predictions are combined, using weights combined by the algorithm, into a final prediction. One of the crucial parameters of the algorithm is the so-called learning rate, that can be computed in a theoretically optimal way when the number of rounds of the algorithm is known in advance [FSS97]. This is not the case if the algorithm is to be executed in real time, and in this case there is no theory that guides the parameter choice. We propose a theoretical and practical study of optimal or near-optimal values of the learning rates when the number of rounds is not known. In a second phase, we will study version of the algorithm that can work in time-changing environments and data-stream like situations.

PATR: Pattern mining from relational and sequential data

[Barcelona: A. Lozano, A. Bifet, J. Baixeries]

We will continue work initiated in MOISES and strongly developed in MOISES-TA about extraction of structures from relational and sequential data. Recently, we have developed algorithms for mining ordered and unordered frequent trees from static datasets [Gar06,BBL06,BBL07a,BBL07B,BBL07c,BBL08].

In this task we will extend the algorithms in two directions: On the one hand, the extensión to more complex combinatorial structures, such as partial orders and acyclic graphs, restricted in a way that includes at least the case of trees. In another direction, we will try to extend the algorithms to data streams, where only one pass is possible and real-time answers are required.

TEMP: Time series

[Málaga: Ll. Mora, F.Cantalejo, Técnico-MA]

Time series are the most classical model with a time component, extensively studied in statistics. A frequent problem in data analysis is processing a set of variables evolving over time, which is modeled as a set of time series. Sometimes, to reduce the number of variables to be analyzed, a selection of groupings of variables is performed [RGP06]. Let us remark on a proposal by our grouping using variable-order Markov models implemented by a special type of probabilistic finite automata to this purpose. This approach has been shown viable in different kinds of time series (mostly of climate parameters [RPM03,MMM05]).

In this task we address the problem of parallelizing algorithms from previous projects (MOISES y MOISES-TA). Parallelism is becoming a viable technological solution to cases in which real-time responses are required. The results of this task will be a software package that can deal with time series arriving as a data stream.

STRING: String mining and analysis

[Málaga: R. Morales, M. Baena, I. Fortes, G. Castillo (Aveiro)]

String mining is a natural evolution of traditional pattern matching algorithms [Gus97]. In this area there are two main structures: suffix trees [Ukkonen95] and suffix arrays [Man93,Ko03,Ko05]. One of the goals in string mining is detection of interesting patterns in

long string. An important work is [FKH05] where frequent patterns are found using suffix arrays.

In this task we will work on the complexity analysis of string mining algorithms as a function of the number of patterns contained in the strings. We will study data structures that reduce time complexity of this problem. We will compare suffix array-based solutions and, evaluate them on strings from various domains. As a result we expect an algorithm that improves on those developed within MOISES-TA, as well as a program for symbolic stream processing.

MARKOV: Learning Markovian models and probabilistic automata

[Barcelona: R. Gavaldà, R. Ferrer; ZarVall: A. Cabaña]

Markov chains are sequence generating devices that have been extensively used to model systems with uncertainty or probabilities. A notoriously hard problem is that of inferring Markov chains when the state is not directly observable and furthermore the structure of the state space is not known in advance. Only recently the first algorithms for this problem have been proposed, coming from the Reinforcement Learning [RGT04,SLJ03,HoI06], but are still inefficient in practice. In [GKPP06], together with researchers at McGill University, we proposed a different algorithm based on work in [CT04] on learning probabilistic automata.

In this task we will develop more efficient versions of the algorithms in [GKPP06,CT04], possibly extending [CT04] to larger classes of automata, and arguing both theoretically and experimentally the improvement over previous versions. As will be discussed later, these algorithms should be useful in tasks corresponding to Areas 4 and 5.

ENSEMB: Multiclassifiers (ensemble methods)

[Málaga: G. Ramos, J. del Campo, J.L. Triviño, A. Ruiz; Barcelona: M.Arias, A. Bifet, R. Gavaldà]

We will continue the study of hierarchical structures that help in meta-learning tasks, leading to induction of better and more precise multiclassifiers. In particular, we will investigate the possibility of using ensembles of neural networks as a way of dealing with massive data sets and data that change over time.

In order to develop multiclassifier systems that can work on time-changing data streams, we will use the change-detection and adaptive windowing methods developed in MOISES-TA, as well as the adaptive versions of weighted majority discussed before.

COMPLEX: Data streams as a model of computation

[Barcelona: R. Gavaldà; ZarVall: A. Arratia, E. Mayordomo, María López]

We will continue the study, started in MOISES-TA of a theory that allows classification of computational problems on data streams.

COMPRESS: Compression algorithms

[ZarVall: E. Mayordomo, M. López, P. Albert, P. Moser, M.L. González, Ingeniero1-ZV]

We will focus on compression algorithms that can be implemented with very low resources. In particular, we will study compression based on pushdown automata, a model that had not

received much attention but is becoming quickly attractive due to the emergence of XML as a standard for document creation and processing. Since 1999, new compression schemes are showing up based on XML grammars and implemented as pushdown machines [HaS06, LeE07].

We intend to formalize the notion of pushdown compressor to analyze their capabilities in detail, and specially their applications to XML compression, still in a very basic state [AuBeBo07]. The software development company TB-solutions has shown their interest in knowing and exploiting competitive XML compression methods developed within the Project.

The celebrated result by Lempel y Ziv [LeZ78] that their compression algorithm is asymptotically optimal than any finite-state machine based algorithm suggests to look for the analogous results for pushdown machine. We suspect that pushdown compressors can do better than the Lempel Ziv algorithm, so this suggests looking for an alternative to Lempel Ziv that can do better than any pushdown compressor and still use few resources.

Finally, data compression can be useful in order to do online prediction, or even other learning tasks. We will study the conversion of this lossless pushdown compression algorithm into an online predictor.

DIM: Dimension, prediction, and compression of geometric data

[ZarVall: E. Mayordomo, M. López, G. de Miguel]

We want to deal with sets of real-valued dimensional data, especially those that are highly variable, and thus require local descriptions.

Constructive dimension gives a canonical way of assigning a “dimension” to each point in space. If X is a reasonably simple set, then the fractal dimension of X is simply the supremum of the dimensions of its points [Lut05].

This robust notion of dimension of a point can be characterized using Kolmogorov complexity of the algorithmic information content of the point [May02], that is, it is related to compressibility of the point. This gives us a nontrivial connection of compression algorithm and information theory with fractal geometry. We can then, for example, discuss the construction of a random point within a fractal.

On the other hand, real-valued data have been much less systematized than discrete data in computer science. In scientific computation there are many representation-dependent methods [Wei00], a topic which has been studied by Gregorio de Miguel in his Ph.D.

The goal of this task is then:

1. Theoretical study of real-valued data representation, compressibility, and Kolmogorov complexity – information theory notions.
2. Study of known geometric data compressors, and experimental comparison with different representations.
3. Formalizing the notion of “random chooser” for points in \mathbb{R}^3 , starting with self-similar sets.
4. Estimating the fractal dimension of different sets, from the compressibility of their points.

RECOM: Variable selection methodology for recommender systems

[Barcelona: J. López, J. Baixeries]

Two theses recently defended within MOISES-TA [Gar06,Bai07] are closely connected to the field called Formal Concept Analysis (FCA). Very briefly this field studies the extraction of information from databases representing them as lattices, which lets us then reason about implications and other relations among variables. We propose to apply ideas from this field to the problem of selecting relevant variables for classification and, more precisely, for recommender and diagnostic systems.

More precisely, in [Lóp06,Lóp07] a recommender system methodology based on Shannon entropy was developed and then tested on medical data involving diseases appearing after kidney transplants. We plan to combine this methodology from FCA concepts and methods in [Gar06], and test the new methodology for diagnostic on the same, or similar, medical data.

CAMBIO: Learning in the presence of change

[Málaga: G. Ramos, R. Hidalgo, J. del Campo, A. Ruiz, J. Gama (Oporto), G. Castillo (Aveiro)]

We will continue the work on algorithms that can work in the presence of concept or distribution drift, and more precisely, on algorithms that can learn

- Decision trees (continuation of the work in MOISES-TA that has produced the IADEM-2 and OnlineTree2 algorithms)
- Multiclassifier systems (continuation of the work that produced the MultiCIDIM-DS algorithm).

The OnlineTree2 algorithm can learn decision trees in incremental and robust way; by robust we mean that it can tolerate e.g. concept drift, noise, and irrelevant attributes. However OnlineTree2 stores part of the training examples, so it is not suited for very large datasets due to memory consumption. On the other hand, the IADEM-2 algorithm does not store examples, but does not have effective strategies for dealing with concept drift. We plan to investigate mixed strategies that avoids the pitfalls of both methods.

Additionally, recent research has shown that it is often advantageous to use classifiers such as Naive Bayes at the tree leafs. We plan to add this feature to the improved version of OnlineTree2.

Tasks concerning the application of analysis, learning, and mining algorithms to real contexts

BLOG: Blog analyzer

[Barcelona: A. Bifet, R. Gavaldà, M. Arias, Ingeniero-BAR; Málaga: M.Baena, F.Cantalejo; ZarVall: A. Arratia, Ingeniero2-ZV]

This task will develop a software prototype able to analyze nontrivial parts of the blogosphere and personalize the results of the search to a user's interests. The main functionalities will be capturing blogosphere data (crawling), grouping blogs or posts according to content (clustering, partitioning), predicting topics for new blogs or posts and selecting interesting

posts for a user's tastes (classification), and topic trend detection (frequent itemset mining). Some usage of XML compression techniques will be necessary.

Since natural language processing techniques are not among our expertise, we plan to use classical (bag-of-words) Information Retrieval methods only, although the design should be prepared to incorporate (outside the project's scope) NLP and semantic web techniques.

CORR: Email classification

[Málaga: R.Morales, M.Baena, Técnico-MA, Ingeniero-MA, F. Cantalejo, G.Castillo (Aveiro)]

As a result of previous work we have a framework for processing and classifying email in email clients. In this task we will address the automatic extraction of rules from messages. As a result of the task we plan to obtain a system able to classify, efficiently and automatically, email messages for a particular user.

DNA: String mining in mitochondrial ADN

[Málaga: R.Morales, M.Baena, Ingeniero-MA; ZarVall: E. Mayordomo, G. de Miguel, P. Albert, Ingeniero1-ZV]

We are particularly interested in determining when a mitochondrial ADN (mtADN) mutation is pathogenic, as a continuation of our current work within the ZARAMIT project, funded by the Aragon government.

Sequencing mtADN of a new patient often results in the finding of a new variant, never described before. It is not easy, however, to determine whether this variant is an innocuous polymorphism, a susceptibility variant, or a pathological mutation. Up to now, mostly evolutive criteria have been used for this analysis,

Our current goal is to apply learning and string mining techniques to this case in two directions: First, automatizing the classification of each new sequence that is inserted in the evolutionary tree. The second is avoiding that insertion all together, designing pathological mutation predictors that do not require building the evolutionary tree.

AUTON: Webserver data and applications to autonomic computing

[Barcelona: R. Gavaldà, A. Bifet, A. Lozano, M. Arias, J. Delgado, Ingeniero-BAR]

We will continue the line started in [PMBGT07a,PMBGT07b] in collaboration with researchers at the Barcelona Supercomputing Center (BSC), a Project EPO. The line concerns the application of learning and data mining techniques to the prediction of web visitor outcome from server logs. We plan to experiment with the data stream algorithms resulting from the CAMBIO task and the improved Markov chain algorithms from the MARKOV task.

Very recently, in collaboration with the BSC too [TCH08], we have started experiments indicating that these techniques can be useful in the design of computer systems that are more energetically efficient, the so-called "green datacenters". We include in this task the data analyses and development of data mining techniques that can contribute to reducing the consumption of large computing centers.

On another line, we will analyze with statistical and data mining techniques data provided by the Fundació puntCAT, also a project EPO. This data concerns the creation, growth, and organization of the Internet addresses with the .cat suffix, the domain created by ICANN to

represent the Catalan-speaking community. We will also apply techniques for community detection and social network analysis, such as those developed in [PBD06,PDS05,PFDS05,PSD03]. Let us say that access to data is an absolute privilege for this team: it is extraordinarily rare to be able to study the growth of an Internet domain from the moment it was born (march 2006, in this case) to the moment it has reached a certain maturity (25,000 subdomains and 4 million pages as of today). The results should help the Fundació puntCAT in characterizing the type of growth it can expect in the future, and the kina of communities that are forming within the domain it manages.

GRAMLOG: Human language: logical grammar approach

[Barcelona: Glyn Morrill, Oriol Valentín-Fernández, Ramon Ferrer i Cancho]

Concerning human language, we propose to further develop the formalism of Type Logical Grammar (TLG) as an approximation to the logical characterization of syntax, semantics, and natural language processing.

The syntactic structures of logical grammar are the proof nets of Type Logical [Gir87]. Here we propose the development of such nets for TLG with modalities (with applications to the intensionality of natural language) and for discontinuous Lambek calculus [MFV07]. For the second case we have already obtained nets for the basic case [MF08], and intend to do the same for the general system.

[Mor00] characterizes the complexity of TLG sentences in terms of how quickly the dependences are incrementally closed as links in proof nets. Independently, [GrG05] characterizes the complexity of sentences in terms of the length of syntactic dependencies in the context of generative grammar. Minimizing such length seems to explain universal properties of syntactical structures such as the exceptionality of crossings between syntactic dependencies in a sentence [Fer06]. Both measures of complexity share many characteristics and we propose as part of this Project to compare and evaluate them with the long-term goal of integrating them.

Depending on the result of these studies, we may attempt to start the study of human language as a data stream in the computational sense. In other words, we will consider whether the syntax and the way in which speakers use it can be partly explained by computational limitations that can be captured by the data stream model.

GAMES: Language acquisition mechanisms via language games

[Barcelona: J. Sierra, R. Gavaldà, A. Lozano, Ingeniero-BAR]

In this task we will define, implement, and experimentally study variants of language games more complex than those studied so far.

This task and the previous one GRAMLOG on the study of human language are clearly related, since the ability to parse language is crucial if one has to learn grammatical aspects of the language. In the experiments done so far [Sie06a,Sie06b,Sie07a,Sie07b] we have studied the acquisition of a vocabulary about spatial knowledge, and the learning process for syntax and semantics of the propositional logic connectives. Future experiments will explore the possibility of learning, within a community of speakers, of more complex concepts and logics (such as those representable in nonstandard logics (temporal, action, causal, nonmonotone logics, for example).

Depending on the results obtained, we will attempt a theoretical analysis of the process of self-organization of the linguistic features of the speakers, to complement the experimental study. The idea would be to perform a converge analysis similar to the PAC analysis developed within computational learning theory [COLT]. We don't know any such attempt of rigorous analysis of language games.

COMANIM: Animal communication and behavior

[Barcelona: R. Ferrer i Cancho, R. Gavaldà, Ingeniero-BAR]

We propose to study the complexity of the symbolic sequences produced various species by means of statistical techniques as in [FCL06] and learning and data mining techniques.

An example of statistical technique is the one used in [FCL06], where we measured the length of the correlations. As a learning technique to try, we will use (variants of) the algorithms developed in the MARKOV task, which might be able to extract information from the sequence of the structure in terms of a probabilistic finite automata that generates a similar distribution or sequence [GKPP06]. Other options are the extraction of patterns from the logical structure of the sequence or even PAC-type algorithms [COLT]. As far as we know, this is a totally novel approach to the inference of patterns from animal behavior.

We propose three real application cases: surface behavior in dolphins, macaque behavior, and birdsong.

MCSG: Comparison of MCSG

[ZarVall: M.L. González, E. Mayordomo]

Many language studies are based on context-free grammars even if it is known that they do not suffice to capture many interesting situations and, in particular, natural languages. The next level in the well-known Chomsky hierarchy, the "context dependent" (type 1) grammars happen to be much too powerful, and computationally intractable. That is why there has been research in finding intermediate classes of grammars that have enough (and just enough) power to capture the problems arising in processing natural languages.

There are several proposals in this direction. Kasai [Kas70] proposed a hierarchy of intermediate grammars called "state grammars" that allow firing a generative rule depending on the state of an additional regular grammar. Independently, there has been quite some interest in so-called "programmed grammars" We propose to compare and study the generative powers of state grammars and programmed grammars.

EVOL: Evolución y modularidad

[Barcelona: J. Delgado, P. Fernández; ZarVall: G. de Miguel]

As discussed in the introduction, we will attempt to develop a framework that explains how modularity can emerge as a feature that enables faster and most powerful artificial evolution. To this end, we propose a framework composed by two main ingredients:

- (1) Metapopulations: Rather than selecting populations with a single fitness measure, let us create various sub-populations and weakly couple them by narrow corridors, where each sub-population has a different fitness measure related to a feature that is required in the final solution.

Each population, therefore, selects a small part of the total set of desired features, but the organisms appearing in the total system may eventually possess all such features.

- (2) A new structure inspired by the eukaryote genome, and especially designed to allow sexual recombination. This mechanism somehow mimics the interchange of partial solutions to each subproblem.

This approach, therefore, tries to study the conditions under which an evolving system will be able to generate solutions that isolate automatically each part of the problem, and can exchange partial solutions.

5. BENEFITS DERIVED FROM THE PROJECT, DIFUSION AND EXPLOTATION OF RESULTS

(maximum 1 page)

The following items must be described:

- ◆ Scientific and technical contributions expected from the project, potential application or transfer of the expected results in the short, medium or large term, benefits derived from the increase of knowledge and technology.
 - ◆ Diffusion plan and, if appropriate, exploitation plan of the results.
-

Scientific contributions and expected benefits

Without detailing again the precise contributions described in each task, we can group them as:

- Methodological contributions: Development of correct methods for classifier performance assessment, and in particular for evaluating, comparing, and selecting classifiers in data mining tasks. Study of the computational limitations of algorithms, by means of complexity theory. Extensions of methods for the (analytical and experimental) study of human language and animal communication, as well as evolutive processes.
- Contributions in the form of algorithms and software for data analysis. More precisely, algorithms for classification, compression, pattern extraction, and diagnosis. The emphasis is on algorithms for sequential data and data streams.
- Contributions in the form of analysis of real-world datasets, that report on patterns implicit in the data and potentially useful in the context where the data was generated. In some cases, these patterns may have actionable consequences in an organization (medical data, email classification, computer system autonomic management...) In other cases, they advance in our understanding of natural or human phenomena (human language, animal communication, biological evolution,...).

A somewhat intangible, but undeniable benefit that we expect is the unifying value of choosing the notion of “symbolic sequence” as a meeting point of the varied interests and backgrounds of all the team members. We have noted in the task descriptions some of the interactions among different fields that we plan to follow. But we anticipate that some others are likely to appear during the development of the project, either by way of topics of interest, either by methods or concepts used.

In fact, some combinations that we have proposed are, to our knowledge, quite novel. Some examples are the application of learning to the analysis of grammar acquisition and animal behaviour modelling, or investigation of connections between the data stream model and language production.

Technology transfer possibilities

The different models that we propose for data mining and learning fit well within the mainstream lines of research in machine learning and data mining. Their possibilities for technology transfer

have been extensively proved around the world, and even some prototypes developed within MOISES and MOISES-TA have been tested in this respect. In particular, data streams and event sequences appear naturally in many contexts, so reliable and usable methods should be of interest in the Spanish or international industrial contexts

With the cooperation of the Project EPOs, we will apply our methods to real datasets and expect that, in some cases, the results will be of immediate interest.

Algorithms for sequence compression are used in all computing contexts. An improvement in efficiency over existing methods can directly improve performance in such diverse topics as bioinformatics and computer graphics. In the specific case of biological sequences, these algorithms are applicable to such important problems as sequence classifications and filogenetic tree construction.

Dissemination plan

The main dissemination channels of the results of the project will be publication in international journals and conferences. As can be seen in the next section (Background of the Group) and the participants' CVs, the team has a tradition of publishing in high-level international forums.

On the other hand, we will take special care to maintain a fluid communication with the Project EPOs, and in general with other institutions and companies that may show interest in our research in the future.

In contrast with our previous projects, in this one we are seriously considering the possibility of channelling the results of the projects via patents. This responds to the increasing perception that the methods and software we can develop may have real interest outside the academic world.

6. BACKGROUND OF THE GROUP

(In the case of a coordinated project the topics 6. and 6.1. must be filled by each partner)

(maximum 2 pages)

◆ **Indicate the previous activities and achievements of the group in the field of the project:**

If the project is related to other previously granted, you must indicate the objectives and the results achieved in the previous project.

If the project approaches a new research field, the background and previous contributions of the group in this field must be indicated in order to justify the capacity of the group to carry out the project.

SESAAME-BAR (Barcelona subproject)

Most researchers in the Barcelona team belong to the LARCA research group (Laboratory for Relational Algorithmics, Learning and Complexity) of Universidad Politécnic de Cataluña (<http://www-lsi.upc.es/~balqui/larca/larca.html>). Formally, the group was created in 2004, because of the restructuring of research at UPC in “research groups”, by members of the former groups of Theoretical Computer Science and Artificial Intelligence, although there was already a history of previous collaborations, especially within MOISES. The coordinator is Professor José L. Balcázar, who belonged to the FRESCO and MOISES projects.

Its permanent members have a long research experience in topics such as algorithmics and computational learning (R. Gavaldà, A. Lozano) or in Artificial Intelligence (J. Delgado), and specifically its connections with Logic (G. Morrill, J. Sierra). The intention when forming the research group was to deepen in the connections between both areas, which become clearer and central every day. They have all participated in a substantial number of publicly funded projects, both national and international.

With the exception of G. Casas-Garriga and M. Fadda, all the members of MOISES-TA participate in this proposal too. New to this project line are Dr. J. Delgado and his Ph.D. student Pablo Fernández (with research background on connections between computer science, physics, and biology), Dr. Marta Arias and Dr. Ramon Ferrer i Cancho (recently arrived to the department after postdoctoral stays at U. Columbia and U. Roma, and with experience in theory and practice of machine learning, and in analysis of complex systems, respectively), and Dr. Josefina López, whose background is on machine learning and recommender systems.

The SESAAME-BAR subproject is in a good part a follow-up and an extension of the lines of FRESCO, MOISES, y MOISES-TA, with some reinforcement on the areas related to analysis of interaction (language and evolution) and an intention of moving to analysis of real contexts. Success on the topics related to these areas can be judged from following publication list. In all cases, we look for methodological and practical connections between our joint areas of expertise.

Observations on some team members:

- J. Delgado is currently half-time in an I+D Project ending in debember 2009, and Ramon Ferrer i Cancho is full time in an I+D Project ending in the same date. The intention in both cases is to move to full-time dedication to SESAAME starting January 2010.
- The principal researcher R. Gavaldà, is habilitated to catedrático (“full profesor”) since april 2007, and a full profesor position for his stabilization has been announced by his university (BOE 15/12/2007).

Publications related to the proposal during 2007:

- [AKM07] Marta Arias, Roni Khardon, Jérôme Maloberti. *Learning Horn Expressions with LogAn-H*. Journal of Machine Learning Research, Vol. 8, Mar 2007, pp. 549-587, 2007.
- [BA07a] Hila Becker and Marta Arias. *Real-time Ranking of Electrical Feeders using Expert Advice*. European Workshop on Data Stream Analysis, Mar 2007, Caserta, Italy, 2007
- [BA07b] Hila Becker and Marta Arias. *Real-time Ranking with Concept Drift Using Expert Advice*. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007.
- [Bai07] Jaume Baixeries. *Lattice characterization of Armstrong and Symmetric Dependencies*. Tesis doctoral, Universitat Politècnica de Catalunya, 2007.
- [BBL07a] José Luis Balcázar, Albert Bifet and Antoni Lozano. *Closed and maximal tree mining using natural representations*. International Workshop on Mining and Learning with Graphs MLG 2007, Firenze, Italy, 2007.
- [BBL07b] José Luis Balcázar, Albert Bifet and Antoni Lozano. *Subtree Testing and Closed Tree Mining Through Natural Representations*. ACKE Workshop on "Advances in Conceptual Knowledge Engineering" 2007, Regensburg, Germany, 2007
- [BBL07c] José Luis Balcázar, Albert Bifet and Antoni Lozano. *Mining Frequent Closed Unordered Trees Through Natural Representations*. International Conference on Conceptual Structures, Sheffield UK, 2007.
- [BBL08] José Luis Balcázar, Albert Bifet and Antoni Lozano. *Mining Implications from Lattices of Closed Trees*. Extraction et gestion des connaissances EGC'2008, Sophia Antipolis, France, 2008.
- [BiG07a] Albert Bifet and Ricard Gavaldà. *Learning from time-changing data with adaptive windowing*. Proc. 7th SIAM Intl. Conf. on Data Mining (SDM'07), 443-449. SIAM, 2007.
- [BiG07a] Bifet, A., Gavaldà R.; *Learning Decision Trees Adaptively form Data Streams with Time Drift*. Sometido.
- [FCC07] Ramon Ferrer i Cancho, Andrea Capocci and Guido Caldarelli. *Spectral methods cluster words of the same class in a syntactic dependency network*. International Journal of Bifurcation and Chaos 17 (7), 2007.
- [FD07] Ramon Ferrer i Cancho and Albert Diaz-Guilera. *The global minima of the communicative energy of natural communication systems*. Journal of Statistical Mechanics, P06009, 2007.
- [Fer07] Ramon Ferrer i Cancho. *Discrete mathematics of distance and crossings in syntactic dependency trees*. Submitted, 2007
- [Fer08a] Ramon Ferrer i Cancho. *Information theory*. In: The Cambridge encyclopedia of the language sciences, Colm Hogan, P. (ed.). Cambridge University Press, 2008.
- [Fer08b] Ramon Ferrer i Cancho. *Network theory*. In: The Cambridge encyclopedia of the language sciences, Colm Hogan, P. (ed.). Cambridge University Press, 2008.
- [FH08] Ramon Ferrer i Cancho and Alvaro Hernández Fernández. *Power laws and the golden number*. In: “Problems of text analysis”, Kelih, E., Levickij, V. and Altmann, G. (eds.), 2008
- [FS07] Pau Fernandez and Ricard V. Solé. *Neutral Landscapes in Signaling Networks*. Journal of The Royal Society Interface 4, 41-47 (2007).
- [MFV07] Glyn Morrill, Mario Fadda and Luis Valentín. *Nondeterministic Discontinuous Lambek Calculus*. Proceedings of the Seventh International Workshop on Computational Semantics, IWCS7, Tilburg, 2007.
- [MKA07a] Chris Murphy, Gail Kaiser and Marta Arias. *Parameterizing Random Test Data According to Equivalence Classes*. Second International Workshop on Random Testing 2007, Nov 2007.
- [MKA07b] Chris Murphy, Gail Kaiser and Marta Arias. *An Approach to Software Testing of Machine Learning Applications*. SEKE 2007, Jul 2007.
- [MPBGT07] Toni Moreno, Nicolás Poggi, Josep Lluís Berral, Ricard Gavaldà, and Jordi Torres. *Policy-based autonomous bidding for overload management in eCommerce websites*. GDN'07 proceedings, Volume I Interneg Research Centre, J. Molson School of Business, Concordia University, 162-166, 2007.

- [PMBGT07a] Nicolás Poggi, Toni Moreno, Josep Lluís Berral, Ricard Gavaldà, and Jordi Torres. *Web customer modeling for automated session prioritization on high traffic sites*. Proc. 11th Intl. Conf. on User Modelling (UM2007). Springer Lecture Notes in Computer Science 4511, 450-454, 2007.
- [PBMGT07b] Nicolás Poggi, Josep Lluís Berral, Toni Moreno, Ricard Gavaldà, and Jordi Torres. *Automatic detection and banning of content stealing bots for e-commerce*. NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security, 2007.
- [Sie07a] J. Sierra Santibáñez. The acquisition of linguistic competence for communicating propositional logic sentences. Proceedings of the Eighth International Workshop on Engineering Societies in the Agents World, ESAW 2007.
- [Sie07b] J. Sierra Santibáñez. *Propositional logic syntax acquisition using induction and self-organisation*. Agent-Based Societies: Social and Cultural Interactions, G. Trajkovski and S.G. Collins (eds.), IGI Global, 2007.
- [TCH08] J. Torres, D. Carrera, K. Hogan, R. Gavaldà, V. Beltran, N. Poggi. Reducing wasted resources to help achieve green data centers. 3rd Workshop on High-Performance, Power-Aware Computing (HPPAC 2008). April, 2008, Miami, FL, USA. Aceptado.

SESAAME-MA (Málaga subproject)

The Málaga team has been involved in Artificial Intelligence research activities since the mid-80's. During these years, they have worked in diverse areas of AI, such as expert systems, intelligent tutorials, AI applications and diagnostic and design in engineering. Among their main objectives are research in the basic techniques of AI, development of techniques for Knowledge Discovery, user modelling, distance learning, and AI technology transfer to industrial and service companies, mostly in the Andalusian context

The group is made up by faculty members from the the Departamento de Lenguajes y Ciencias de la Computación and the Departamento de Matemática Aplicada, as well as faculty from other universities doing research in similar areas

The team has participated in a large number of projects on pure and applied research, funded both by public national and international organisms as well as private bodies. More precisely, we are working in topics related to the current proposals, including the mentioned projects FRESCO MOISES, and MOISES-TA.

Dissemination of the results of our research has been carried out by scientific publications and by their application in projects with private and public entities. Some examples are the data analysis on data provided by the Social Security Nacional Institute (research on disabilities), the Tourism, Trade, and Sports of the Junta de Andalucía, and the Carlos Haya Regional University Hospital (Mental Health Unit), among others.

Publications related to the Project since 2003:

1. L. Mora-Lopez, R.Ruiz, R.Morales, Modelo para la seleccion automatica de componentes significativas en el analisis de series temporales, en: X CAEPIA, S.Sebastián (2003).
2. M.Nuñez, R. Fidalgo, R.Morales, Discovering event prediction knowledge, Tendencias de la Minería de Datos en España (Sevilla, 2004).
3. G.Ramos, R.Morales, J.Campo, IADEM-0: Un nuevo algoritmo incremental, Tendencias de la Minería de Datos en España (2004).
4. R.Morales-Bueno, J. Wallace-Ruiz, M. Baena-García, Prospección de Datos aplicada a problemas sociales, Tendencias de la Minería de Datos en España (84-688-8442-1) (Sevilla (España), 2004).
5. M.Nuñez, R. Fidalgo, R.Morales, Discovering Temporal Patterns from Events and other Multivariate Data, en: Proceedings of the Congress Euro Electromagnetics (EUROEM 2004), Magdeburg (Germany) (2004).

6. M.Nuñez, R.Morales, Incorporating Prediction Facilities to Autonomous Systems, en: Proceedings of the ACM Autonomous Agensts & Multi Agent Systems, New York. (2004).
7. M. Baena, R.Morales, S. Cabuchola, I. Santos, Prospección de datos sanitarios: Estudio de la incapacidad permanente, en: Inforsalud 2004, MADRID (2004) 127-130.
8. M.Nuñez, R. Fidalgo, R.Morales, Reducing Potential Risks by Preventing Events: A Case Study, en: Proceedings of the IFAC Congress on Management and Control of Production and Logistics (MCPL-2004), Oxford, UK (2004).
9. M.Nuñez, R. Fidalgo, M. Baena, R.Morales, The Influence of Active Region Information in the Prediction of Solar Flares, en: Book of Abstracts of First European Space Weather Week, Noordwijk (The Netherlands) (2004).
10. Ll. Mora, R. Morales, M. Sidrach-de-Cardona, F. Triguero. Modelling time series of climatic parameters with probabilistic finite automata. *Environmental Modelling and Software*, Volume 20, Issue 6, Pages 651-815 (June 2005).
11. G.Ramos, José del Campo-Ávila, R.Morales, Induction of decision trees using an internal control of induction, *Lecture Notes in Computer Science*, 3512, (2005) 795-803.
12. G.Ramos, José del Campo-Ávila, R.Morales, E-CIDIM: Ensemble of CIDIM classifiers, *Lecture Notes in Artificial Intelligence*, 3584, (2005) 108-117.
13. G.Ramos, José del Campo-Ávila, R.Morales, ML-CIDIM: Multiple Layers of Multiple Classifier Systems based on CIDIM, *Lecture Notes in Artificial Intelligence*, 3642, (2005) 138-146.
14. G.Ramos, José del Campo-Ávila, R.Morales, Inducción de árboles de decisión con CIDIM: nuevos enfoques, en: *Actas del III Taller de Minería de Datos y Aprendizaje*, Granada, España (2005) 13-20.
15. G.Ramos, José del Campo-Ávila, R.Morales, Aprendizaje por capas basado en sistemas multclasificadores, en: Roque Marín, Eva Onaindia, Alberto Bugarín, José Santos (eds.) *XI Conferencia de la Asociación Española para la Inteligencia Artificial. CAEPIA 2005*, Santiago de Compostela (2005) 113-122.
16. G.Ramos, José del Campo-Ávila, R.Morales, FE-CIDIM: Fast Ensemble of CIDIM Classifiers, *International Journal of Systems Science*, 37, 13 (2006) 939-947.
17. G.Ramos, José del Campo-Ávila, R.Morales, Incremental Algorithm Driven by Error Margins, *Lecture Notes in Artificial Intelligence*, 4265, (2006) 358-362.
18. M. Baena, José del Campo-Ávila, Albert Bifet, R. Fidalgo, Ricard Gavaldà, R.Morales, Early Drift Detection Method, en: J. Gama, J. Aguilar-Ruiz, R. Klinkenberg (eds.) *Fourth International Workshop on Knowledge Discovery from Data Streams*, Berlín (Alemania) (2006) 77-86.
19. José del Campo-Ávila, G.Ramos, R.Morales, Improving prediction accuracy of an Incremental algorithm driven by error margins, en: J. Gama, J. Aguilar-Ruiz, R. Klinkenberg (eds.) *Fourth International Workshop on Knowledge Discovery from Data Streams*, Berlín (Alemania) (2006) 57-66.
20. I. Fortes, L. Mora, R. Morales, F. Triguero: Inductive learning models with missing values. *Mathematical and Computer Modelling*. 44, Issues 9–10, November 2006, 790–806 Elsevier, Pergamon.
21. José del Campo-Ávila, G.Ramos, R.Morales. Incremental learning with multiple classifier systems using correction filters for classification. *Lecture Notes in Computer Science*, 4723, (2007) 106-117.
22. M.Nuñez, R.Fidalgo, R.Morales, Learning in environments with unknown dynamics: towards more robust concept learners, *Journal of Machine Learning Research*, 8 (2007), 2595-2628
23. R.Fidalgo, Architecture for distributed data mining from multiple situations, *Actas de la Conferencia Latinoamericana de Computación de Alto*, Bogotá (Colombia), (2007)
24. José del Campo-Ávila, G.Ramos, R.Morales. Improving the performance of an incremental algorithm driven by error margins. *Intelligent Data Analysis*. Por aparecer

Ph.D. thesis supervised by Rafael Morales:

25. Nuevos desarrollos en aprendizaje inductivo. Gonzalo Ramos Jiménez, Málaga, E.T.S.I. Informática 2001, sobresaliente cum laude
26. Prospección de datos, Aprendizaje computacional y Técnicas Estadísticas para la obtención de reglas. Inmaculada Fortes Ruiz, Málaga, E.T.S.I. Informática, 2001, sobresaliente cum laude
27. Un nuevo modelo de aprendizaje para el estudio de secuencias de símbolos. José Luís Triviño Rodríguez, , Málaga, E.T.S.I. Informática, 2002, sobresaliente cum laude
28. Análisis de las patologías causantes de Discapacidad Laborar Permanente mediante la aplicación de técnicas de inteligencia artificial y prospección de datos. Santiago Cabuchola Moreno, Málaga, Facultad de Medicina, 2003, sobresaliente cum laude

29. Nuevos enfoques en aprendizaje incremental. José del Campo Ávila, Málaga, E.T.S.I. Informática, 2007, sobresaliente cum laude.

MOISES-ZAR (Zaragoza & Valladolid subproject)

The team at Zaragoza and Valladolid universities has participated in three national research projects since 2000. It is currently formed by 9 researchers, of which 7 are specialists mainly in Computational Complexity, and two in goodness-of-fit tests mainly.

The last expansions of the team correspond to Dra. Alejandra Cabaña Nigro, already in MOISES-TA, and Roberto San Martín Fernández, about to defend his Ph.D. Thesis, who will contribute together with Dra. Cabaña to our new and ambitious project of developing global methods for classifier comparison. On the other hand, Dr. Philippe Moser, also in MOISES-TA, who completed his postdoc in U. Zaragoza, is co-advising. Pilar Albert, and has a permanent position at Ireland National University at Maynooth; and with Gregorio de Miguel, also about to defend his Ph.D. thesis on computation with real numbers, a topic related to our proposal on compressing geometrical data.

Let us note the existence of strong ties with prestigious international research groups. In particular, E. Mayordomo and P. Moser together with the two Ph.D. students cooperate with Prof. Jack Lutz's group in Iowa State University on different issues in Computational Complexity. Professor E. Mayordomo belongs to the Steering Committee of the Computability in Europe research group. Dr. Arratia at Valladolid has frequent collaborations with Prof. Ian Stewart from Leicester University (UK), well known for his work on Descriptive Complexity.

Publications related to the Project since 2003:

1. P. Albert, E. Mayordomo, P. Moser, S. Perifel, *Pushdown compression*, Lecture Notes in Computer Science, 2008
2. A. Arratia y C. Marijuán, *Cómo mejorar el PageRank de un árbol*. E. M. Moro et al (eds.): V Jornadas de Matemática Discreta y Algorítmica (Memoria del Congreso), Pub. Univ. de Valladolid (ISBN: 84-8448-380-0), 53--60, 2006.
3. A. Arratia, C. Ortiz, *Approximating the Expressive Power of Logics in Finite Models*. Lecture Notes in Computer Science, Conference Proc. (2976):540-556. LATIN04, 2004.
4. A. Arratia, C. Ortiz, *Counting proportions of sets: expressive power with almost order*. Lecture Notes in Computer Sciences., LATIN 06, (3887):105-117, 2006.
5. A. Arratia, C. Ortiz, *Expressive power and complexity of a logic with quantifiers that count proportions of sets*. J. Logic and Computation, 2006.
6. A. Arratia, C. Ortiz, *On a syntactic approximation to logics that capture complexity classes*, ECCC: Electronic Colloquium on Computational Complexity, ISSN 1433 8092, TR06-014, 2006.
7. A. Arratia, C. Ortiz, *Syntactic approximation to computational complexity*. Abstract en: International Congress of Mathematician, ICM 2006. Libro de resúmenes de la ICM, publicado por European Mathematical Society Publishing House, 2006.
8. A. Arratia, I. A. Stewart, *A note on first-order projections and games*. Theoretical Computer Science, 290:2085 – 2093, 2003.
9. K.B. Athreya, J.M. Hitchcock, J.H. Lutz, E. Mayordomo, *Effective strong dimension in algorithmic information and computational complexity*. SIAM Journal on Computing. 37:671-705. 2007
10. A. Cabaña, E.M Cabaña, *Goodness-of-fit to the exponential distribution, focused on Weibul alternatives*. Comm. in Statistics. Simulation and Computation. 34: 711-724, 2003.
11. A. Cabaña, E.M Cabaña, *Tests of normality based on Transformed Empirical Processes*. Methodology and Computing in Applied Probability, 5:309-336, 2003.

12. A. Cabaña, A.J. Quiroz, *Using the empirical moment generating function in testing for the Weibull and the type I extreme value distributions*. TEST, 14:-417,432, 2004.
13. J.J. Dai, J.I. Lathrop, J.H. Lutz, E. Mayordomo, *Finite state dimension*. Theoretical Computer Science, 310: 1-33, 2004.
14. D. Doty, X. Gu, J. H. Lutz, E. Mayordomo, P. Moser, *Zeta-dimension*. Lecture Notes in Computer Science. 3618: 283-294, 2005.
15. D. Doty, P. Moser, *Finite-state dimension and lossy decompressors*. Arxiv.org. (technical report), 2006.
16. D. Doty, P. Moser, *Feasible depth*. In Third Conference on Computability in Europe, CiE 2007. 228-237, 2007.
17. S.A. Fenner, J.H. Lutz, E. Mayordomo, P. Reardon, *Weakly useful sequences*. Information and Computation, 197:41-54, 2005.
18. J. García Rodríguez, J. M. García Chamizo, G. de Miguel Casado, D. Gil Méndez, J. A. Gil Martínez Abarca, *II Jornadas sobre Domótica y Automatización Industrial (JDAI'2004)*, ISBN: 84-609-4146-9, 2005.
19. X. Gu, J. H. Lutz, E. Mayordomo, *Points on computable curves*. Proceedings of the Forty-Seventh Annual IEEE Symposium on Foundations of Computer Science (Berkeley, CA, October 22-24, 2006). Pags: 469-474, 2006.
20. X. Gu, J. H. Lutz, P. Moser, *Dimensions of Copeland-Erdoes Sequences*. Inf. Comput. 205(9): 1317-1333, 2007
21. C. Herrero, R. San Martín, F. Bravo, *Effect of hest and ash treatments on germination os Pinus pinaster and Cistus laurifolius*. Journal of Arid Environments. 70:540-548, 2007.
22. J.M. Hitchcock, M. López-Valdés, E. Mayordomo. *Scaled dimension and the Kolmogorov complexity of Turing hard sets*. Theory of Computing systems, por aparecer, 2008.
23. J. M. Hitchcock, J.H. Lutz, E. Mayordomo. *Scaled dimension and non-uniform complexity*. Journal of Computer and System Sciences, 69:97-122, 2004.
24. J. M. Hitchcock, J. H. Lutz, E. Mayordomo, *The fractal geometry of complexity classes*. SIGACT News. 36:24-38, 2005.
25. R. Impagliazzo, P. Moser, *A Zero-one Law for RP*. In Proceedings of the 18th Conference on Computational Complexity. Pags 48-52, 2003.
26. M. López-Valdés. *Lempel-Ziv dimension for Lempel-Ziv compression*, Proceedings of the 31th International Symposium on Mathematical Foundations of Computer Science (MFCS'05). Lecture Notes in Computer Science, pags 693-703, 2006.
27. M. López-Valdés, E. Mayordomo. *Dimension is compression*, Lecture Notes in Computer Science. 3618:676-685, 2005.
28. E. Mayordomo. *Effective fractal dimension in algorithmic information theory*. New Computational Paradigms: Changing Conceptions of What is Computable. Pags: 259-285, 2008
29. P. Moser. *Baire's Categories on Small Complexity Classes*. In 14th Int. Symp. Fundamentals of Computation Theory. Pags 333-342, 2003.
30. P. Moser. *BPP has Effective Dimension at most 1/2 Unless BPP = EXP*. ECCC (technical report), 2003.
31. P. Moser, *RP is Small in SUBEXP else ZPP equals PSPACE and NP equals EXP*. ECCC (technical report), 2003.
32. P. Moser. *Martingale Families and Dimension in P*. Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, CiE 2006. Pags 388-397, 2006.
33. P. Moser. *Baire Categories on Small Complexity Classes and Meager-comeager*. Inform. Comput. 206(1):15-33, 2007.
34. P. Moser. *Generic Density and Small Span Theorem*. Inform. Comput. 206(1):1-14, 2007.
35. P. Moser. *On the Convergence of Fourier Series of Computable Lebesgue Integrable*. Electronic Notes in Theoretical Computer Science (ENTCS). 2007.
36. P. Moser. *Resource-bounded Measure on Probabilistic Classes*. Accepted for publication in Information Processing Letters, 2007.
37. V. Pando, R. San Martín, *Regresión Logística Multinomial*. Cuadernos de la Sociedad Española de Ciencias Forestales. 18:323-327, 2004.
38. M. T. Signes Pont, J. M. García Chamizo, H. Mora Mora, G. de Miguel Casado, *Calculation Scheme Based on a Weighted Primitive: Application to Image Processing Transforms*. EUR. JASP. Num. Lin. Algebra in Signal Processing Applications (ISSN:1687-6172), 2007.

6.2 PUBLIC AND PRIVATE GRANTED PROJECTS AND CONTRACTS OF THE RESEARCH GROUP

Indicate the project and contract grants during the last 5 years (2003-2007) (national, regional or international)
Include the grants for projects under evaluation

MOISES-BAR subproject

Title of the project or contract	Relationship with this proposal (1)	Principal Investigator	Budget		Funding agency and project reference	Periodo de vigencia o fecha de la solicitud (2)
			EURO			
Pattern Analysis, Statistical Modelling, and Computational Learning (PASCAL)	1	John Shawe-Taylor (U. Southampton) Barcelona: José L. Balcázar & R. Gavaldá			Unión Europea, Network of Excellence, IST-2002-506778	C 2004-2006
Modelado individualizado de secuencias simbólicas (MOISES)	1	Elvira Mayordomo Barcelona: G. Morrill	320.000		Ministerio de Ciencia y Tecnología (TIC2002-04019-C03)	C 2002/2005
Modelado Individualizado de Secuencias de Símbolos. Teoría y Aplicaciones (MOISES-TA).	1	Rafael Morales Barcelona: R. Gavaldá	284.700		CICYT TIN2005-08832-C03-02	C 2005/2008
Dynamically Evolving Large Scale Information Systems (DELIS)	2	F. Meyer-auf-der-Heide (Paderborn). Barcelona: J. Díaz			EU Framework VI, Integrated project 001907	C 2004/2006
Algorithms and Complexity, Future Technologies (ALCOM-FT)	2	E. Schmidt (Aarhus) Barcelona: J. Díaz			EU Framework V, Integrated project	C 06/2001-12/2003
Acción Integrada UPC-McGill	2	R. Gavaldá (Barcelona) D. Thérien (Montreal)	3000		Acciones Integradas MEC	C 01/2004-09/2005
Transiciones de Fase y Evolución de Grafos en Sistemas Complejos	2	R.V. Solé (UPF)	47.470		CICYT	C 2001-2004
Técnicas de Optimización Avanzada para Problemas Complejos	2	Enrique Alba (U. Málaga)	52.946		CICYT	C 2002-2005
Análisis Operacional sobre Flujos de Peatones en Centros Históricos de Portugal y España	2	Manuel Delgado Ruiz (U. Barcelona)			CICYT	C 2007-2009
Evolución dinámica de sistemas con interacciones no lineales definidos en redes complejas. Desarrollo teórico	2	Conrad Pérez			CICYT FIS2006-13321-C02-01	C 2007-2009

Efectos de conectividad y dinámica en redes biológicas y sociales	2	Albert Díaz Guilera		CICYT BFM2003-08258-C02-02	C 2006
ECAGENTS (Embodied communicating agents)	3	Domenico Parisi / Stefano Nolfi (Roma)		Unión Europea	C 2005
A Logical Approach to Linguistic Interaction and Evolution (LALIE)	1	Robin Cooper, Goteburg		LogiCCC, European Science Foundation	S 07/2008-06/2010

MOISES-MA subproject

Title of the project or contract	Relationship with this proposal (1)	Principal Investigator	Budget	Funding agency and project reference	Periodo de vigencia o fecha de la solicitud (2)
			EUROS		
Red Española de Minería de Datos y Aprendizaje Automático	1	José Cristobal Riquelme Santos (U. Sevilla)	12.000	MEC TIN2004-21343-E	C
Red Española de Minería de Datos y Aprendizaje Automático	1	José Cristobal Riquelme Santos	15.000	MEC TIN2006-27675-E	C
MINería de DATos Para Los USuarios en diferentes áreas de aplicación (MINDATPLUS)	1	Francisco Herrera Triguero	84.800	Dirección Gra. de Investigación, Tecnología y Empresa. Junta de Andalucía. TIC 531	C
Modelo Individualizado de SEcuencias de Símbolos. Teoría y Aplicaciones (MOISES-TA)	1	Rafael Morales Bueno	172.550	MEC TIN2005-08832-C03-01	C

MOISES-ZAR subproject

Title of the project or contract	Relationship with this proposal (1)	Principal Investigator	Budget		Funding agency and project reference	Periodo de vigencia o fecha de la solicitud (2)
			EURO			
Modelado Individualizado de Secuencias de Símbolos. Teoría y Aplicaciones (MOISES-TA).	1	Elvira Mayordomo (subproject), Rafael Morales (coordinator)	284.700\$		CICYT TIN2005-08832-C03-02	C2005/2008
Modelado individualizado de secuencias simbólicas (MOISES)	1	Elvira Mayordomo (Project coordinator)	320.000		Ministerio de Ciencia y Tecnología (TIC2002-04019-C03)	C2002/2005
ZARAMIT: Sistema informático para DNA mitocondrial humano y su estudio evolutivo	1	Elvira Mayordomo	54.740\$		Departamento de Ciencia, Tecnología y Universidad, Gobierno de Aragón	C2007/2009
Pruebas de ajuste	2	Alejandra Cabaña	3.000\$/año		IVIC - Venezuela Entidades	C1997/actualidad
SGER: Multidisciplinary Aspects of Computation Theory	2	Jack H. Lutz	74.948\$		NSF (National Science Foundation, gobierno EE.UU.)	C2003/2005
SINGACOM (Singularidades, Geometría Algebraica, Álgebra Conmutativa, Codificación, COMbinatoria, COMputación y Optimización)	3	Antonio Campillo López	142.000		Ministerio de Educación y Ciencia	C2005/2008
Ingeniería de Sistemas de Eventos Discretos. Reconocimiento como grupo de excelencia de investigación	3	Manuel Silva Suárez	30.000\$		Depto. de Educación y Ciencia, Gobierno de Aragón	C2002/2007
Técnicas de representación y de recorte y métricas probabilísticas. Aplicaciones Estadísticas	2	Carlos Matrán Bea			Junta de Castilla y León	C2006/2009
Nuevas Transformaciones de Procesos en Inferencia	2	Enrique M. Cabaña			Dirección Nacional de Ciencia y Tecnología, Uruguay, PDT- 630053	C2007/2010

MODNET (FP6 Marie Curie Research Training Network in Model Theory and its Applications)	2	Zoé Chatzidakis (por France); David Evans (por UK), Enrique Casanovas (por España)			Comisión Europea, contrato núm. MRTN-CT-2004-512234	C2005/2008
Hardness, randomness and quantumness	2	José Rolim			Swiss National Science Foundation	C2002/2004
Effective Hausdorff Dimension and Scaled Genericity	2	Philippe Moser			Swiss National Science Foundation	C2004/2005
Gestión del acceso a entornos de seguridad mediante identificación biométrica de usuarios	3	Dr. Francisco A. Pujol López	2000 €		Conselleria d'Empresa, Universitat i Ciència. Generalitat Valenciana	C2006
Control de Calidad de Superficies Brillantes y Especulares Mediante Visión Artificial	3	Dr. Andrés Fuster Guilló	23100€		Conselleria d'Empresa, Universitat i Ciència. Generalitat Valenciana	C2005/2006
IDENTIFICACIÓN DE CARACTERES DE SELECCIÓN PARA LA MEJORA DE LA RECTITUD EN Pinus pinaster Ait	3	M ^a del Rosario Sierra de Grado	162.000€		Ministerio de Educación y Ciencia- CICYT	C2007/2010
REGENERACIÓN NATURAL Y PRIMER DESARROLLO DE RODALES FORESTALES PARA MASAS DE PINUS PINASTER AIT.	3	Felipe Bravo Oviedo	105.850 €		DGES	C2004/2007

(1) Write 0, 1, 2 or 3 according to: 0 = Similar project; 1 = Very related; 2 = Low related; 3 = Unrelated.

(2) Write C or S if the project has been funded or it is under evaluation, respectively.

7. TRAINING CAPACITY OF THE PROJECT AND THE GROUP

(In the case of Coordinated Projects this issue must be filled by each partner)

This title must be filled only in case of a positive answer to the corresponding question in the application form. Justify that the group is able to receive fellow students (from the Suprograma de Formación de Investigadores) associated to this project and describe the training capacity of the group. In the case of coordinated projects, each subproject requesting a FPI fellowship must fill this issue.

Note that all necessary personnel costs should be included in the total budget requested. The available number of FPI fellowships is limited, and they will be granted to selected projects as a function of their final qualification and the training capacity of the groups.

SESAAME-BAR

The Departamento de Lenguajes y Sistemas Informáticos at UPC has two Ph.D. programmes (Software and Artificial Intelligence) with quality distinction from Ministerio de Educación since 2004. All permanent members of the SESAAME-BAR subproject have taught in one of the two programmes.

In the last 10 years, members of the research team have supervised or supervised the following doctoral thesis:

- R. Gavaldà advised C. Domingo (1999), D. Guijarro (2000) and V. Dalmau (2000), and is advising Albert Bifet (expected defence date: end 2008).
- J. Delgado supervised J.M. Pujol (2006) and is supervising P. Fernández Durán.
- G. Morrill is supervising M. Fadda and O. Valentín.
- J. L. Balcázar (LARCA member, though not SESAAME member) has advised J. Castro (2004), G. Casas-Garriga (2006) and J. Baixeries, the two latter MOISES-TA members.
- J. López advised Silvana Ascjar.

That is, a total of 8 thesis defended in the last 10 years, and 4 more in process.

As an indication of the quality and appreciation (in academia and industrial contexts) of these thesis, let us mention that C. Domingo is currently General Director of Telefónica I+D (research lab) in Barcelona, D. Guijarro is development head at Mannes Technology Inc., G. Casas-Garriga has been recruited by the possibly strongest data mining group in Europe (H. Mannila's group in Helsinki), that J.M. Pujol, after a postdoc in U. Michigan, will join the previously mentioned Telefónica I+D.

SESAAME-MA

The Departamento de *Lenguajes y Ciencias de la Computación*, of Universidad de Málaga, offers a Ph.D. program with quality distinction from Ministerio de Educación since 2004, and several courses in this program are offered by members of the group. Currently, this program has been reconverted to an official MSc.+Ph.D. program.

During the last two courses (2005-06 and 2006-07) we have organised specialization courses connected to this M.Sc. program taught by renowned researchers: "Knowledge discovery from data streams" by João Gama, and "Validación estadística de experimentos" by Alejandra Cabaña.

The principal investigator of this group, Rafael Morales, has supervised 5 Ph.D. theses in the last 7 years, all related to machine learning and data mining. Currently, several other researchers are also supervising Ph.D. theses which should be defended during 2008: Marlon Núñez is advising Raúl Fidalgo, and Rafael Morales is advising Manuel Baena.

Another point to mention is the strong support to students in the group for research stays abroad to complement their instruction. In the last two years, there have been 3 stays abroad of 3 months each; they are contributing to the outer visibility of the group.

Some group members are participating in the "Máster Universitario en Informática Aplicada a las Telecomunicaciones Móviles" in its 5 editions.

SESAAME- ZAR

In the Zaragoza group there are now two Ph.D. students. Elvira Mayordomo advises María López (defence expected in 2008) and Pilar Albert (with Philippe Moser as co-advisor). Elvira Mayordomo has cooperated in the supervision of the Ph.D. theses Jack J. Dai and John M. Hitchcock (2003, publications 8) y 9) of Section 6.). Unfortunately this collaboration is not officially recognized due to the fact that Iowa State U. rules demand that all formal advisors belong to it.

The Departamento de Informática e Ingeniería de Sistemas of Universidad de Zaragoza offers a Ph.D. programme with distinction quality since 2003, with several courses taught by group members. It also holds a number of grants for inviting foreign experts and for financing students' visits abroad.

Dr. Philippe Moser, who visited as a postdoctoral fellow U. Zaragoza from septiembre 2005 to june 2007, currently holds a permanent faculty position at the Ireland National University at Maynooth.