Machine Learning Group

Université Libre de Bruxelles

WEB Site: http://www.ulb.ac.be/di/mlg

Coordinator: Gianluca BONTEMPI

gbonte@ulb.ac.be

ULB Machine learning group

- 1 associate professor +7 researchers
- Research topics:
  - Machine learning, data mining, forecasting, modelling and simulation
    - **Theoretical research**: Local learning, feature selection, model selection
    - **Speech recognition and text classification**
    - **Bioinformatics – Gene expression and cancer detection**
    - **Computer-aided medicine**
    - **Data processing in wireless sensor networks**
- Facilities:
  - Cluster of 16 computers
  - 30 ultra low power wireless sensors
  - Lego robotics lab (10 Mindstorms kits)
  - More to come… (mobile robot, Mindstorms NXT kits, 100 wireless sensors)
- Website: www.ulb.ac.be/di/mlg

What is data-mining?

- Search for valuable information in large volumes of real data where large means
  - Large number of samples
  - Large number of features (sometimes several order of magnitude larger than the number of samples)

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns & rules.

Knowledge from data

Extracting knowledge from data is a
major issue in all sciences.

Knowledge can be extracted for different purposes:

- qualitative description of phenomena,
- theoretical understanding of laws underlying nature,
- prediction of future observations and rational decision-making.

What is data-mining?

- Search for valuable information in large volumes of data.
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns & rules.

Why Mine Data? (for businesspersons)

- Lots of data has been and is being collected and warehoused.
  - the largest corporate database are measured in terabytes (10^{12})
- computing has become affordable.
- competitive pressure is strong
  - Provide better, customized services for an edge.
  - Information is becoming product in its own right.
- « convert data into business intelligence »
- « make management decision making based on facts not intuition »
- « get closer to the customers »
- « gain competitive advantage ».

Why mine data? (for scientists)

- Data collected and stored at enormous speeds (Gbyte/hour)
  - remote sensor on a satellite
  - telescope scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Large dimensionality (thousands of variables).
- Data mining for data reduction.
  - cataloging, classifying, segmenting data
- Helps scientists in hypothesis formation.

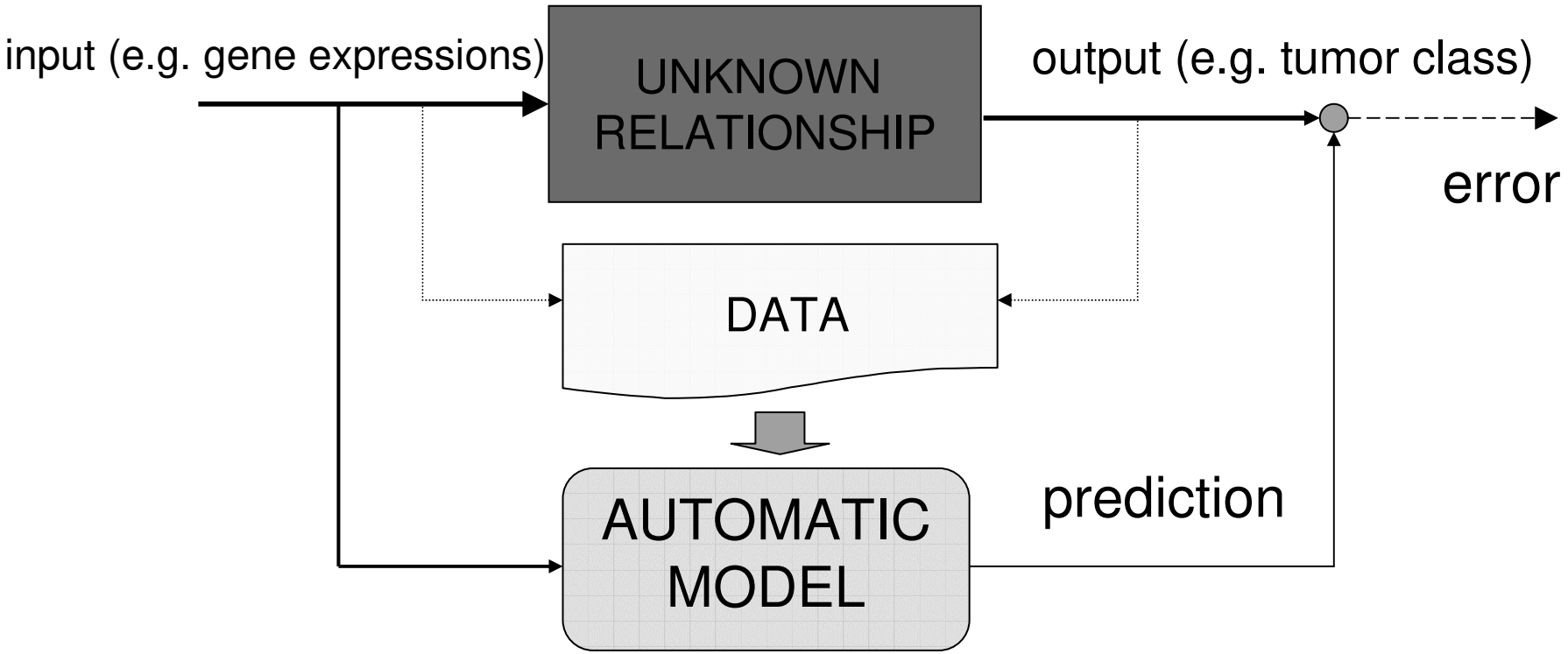
What does data-mining?

- **Description and visualization models**
  - Find human-interpretable patterns that describe the data.

- **Prediction models**
  - Use some variables to predict unknown or future values of other variables.
  - Examples: classification, estimation, regression, clustering, time series prediction.

From [Fayyad, et.al.] *Advances in Knowledge Discovery and Data Mining*, 1996

Prediction task



- Finite amount of noisy observations.
- No a priori knowledge of the phenomenon.

Classification/prediction examples

- Assign keywords to articles and news in order to better classify them.
- Classifying credit applicants as low, medium, or high risk.
- Determining which home telephone lines are used for Internet access.
- Classifying an email as spam or not spam.
- Figuring out which customers are likely to stop being customers (*churn*).
- Estimating the value of a piece of real estate
- Predicting which telephone subscribers will order a value-added service as ADSL or voice-mail
- Predicting which CARREFOUR clients will be more interested to a discount in Italian products.
- Classify satellite images in civil and military sites.
- Predict which machine is most likely to be the next to fail.

Machine learning and DM

- The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.
- ML paved the way to the treatment of real problems related to data analysis, sometimes overlooked by statisticians (nonlinearity, classification, pattern recognition, missing variables, adaptivity, optimization, massive datasets, data management, causality, representation of knowledge, parallelisation)
- ML has its roots on statistics and complements it by focusing on: algorithmic issues, computational efficiency, data engineering.
- ML provides
  - a statistical framework to formalize the problem of data mining
  - a set of computational tools to extract models from data
    - Neural networks
    - Support vector machines
    - Local learning

Hot topics in ML

- Dimensionality reduction
- Feature selection
- Model selection
- Network inference
- Non stationarity
- Visualization
- Unstructured data

Some MLG projects…

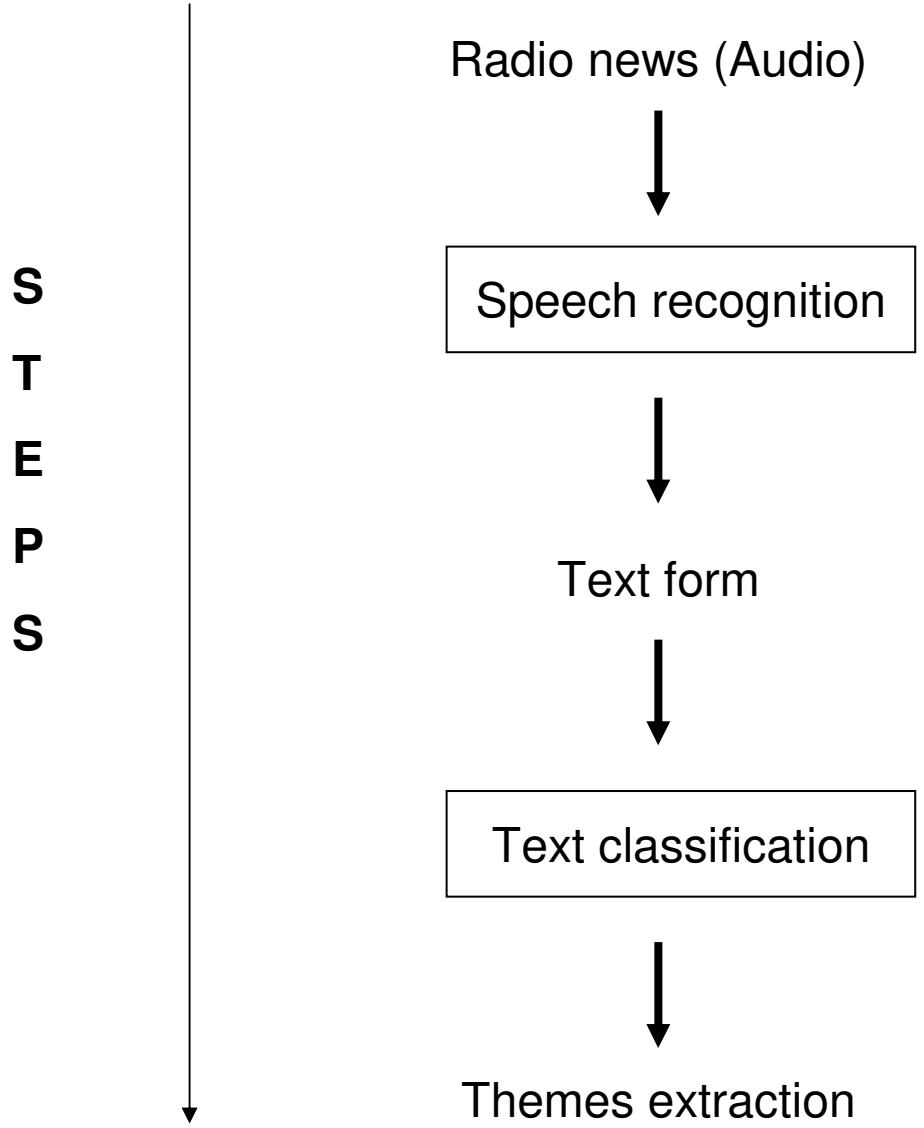
Projet AIDAR

Adressage et **I**ndexation de **D**ocuments multimédias
Assistés par des techniques de **R**econnaissance vocale
(Speech Recognition Techniques for Multimedia Document Indexing and Retrieval)

- Partners
  - Voice-Insight s.a. : Speech recongition for database content retrieval
  - asbl TITAN : Digital archiving
- Funding organisation: Région Bruxelles capitale
- Duration: 2004-2006

Researches

- Thematic information extraction from audio and/or text files (Politics, economy, war,…)
- (semi-)automated indexing of audio and text documents (ex : TV news), meta-data generation for content archiving
- Tools :
  - Speech recognition
  - Text classification


S
T
E
P
S
Radio news (Audio)
Speech recognition
Text form
Text classification
Themes extraction

Applications

- Archiving of audio and text documents (Projet ***AIDAR***)
- Filtering and classification of documents for companies :
  - Allow to filter documents received by different services of a company
- Automatic summary of audio/text documents

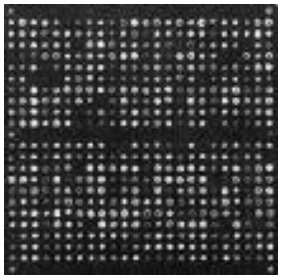
Feature selection method for classification problems

- Funding organisation: ULB
- Duration: 2004-2009

Motivations

- More and more problems involve a high number of features
- Discarding irrelevant and redudant features allows to improve classifier performances
- Using information theory is a promising way to determine relevant features in problems involving several thousands of features

Applications



- Biological data analysis (DNA microarray) for classification of tumors
- Analysis of data coming from industrial processes, to determine features involved in
  - Piece defects
  - Process defects
- Data analysis for Customer Relationship Management (CRM)

Molecular Signatures in Human Cancers

- Partners:
  - Microarray Unit (Institut Jules Bordet)
  - Machine Learning Group (Université Libre de Bruxelles)
- Funding organisations:
  - Télévie
  - Les Amis Bordet
  - Fonds Euson, ...
- Duration: 2004-2008

Research content

- Using microarray technology and machine learning methods build a classifier that:
  - can predict patient survival without treatment (prognostic)
  - can predict treatment outcome of patients (prediction)
- Tumors could be classified into subtypes based solely on the difference of expression patterns.

Applications

- Build classifiers based on microarray experiment potentially usable in clinic routines
- Find new drug targets for cancer therapies.
- Improve our understanding of the tumor genesis (implied pathways, tumor subtypes, ...)

Molecular networks in yeast

Integrating experimental and theoretical approaches to decipher the molecular networks of nitrogen utilization in yeast

- Partners:
  - Département de Biologie Moléculaire (B. André)
  - Département de Biologie Moléculaire (J. van Helden)
  - Département de Chimie (M. Kaufman)

- Funding organisations: Communauté Française de Belgique

- Duration: 2004-2009

Research content

- The yeast *Saccharomyces cerevisiae* is the model living organism that is the most extensively used to develop the novel experimental methods of genomics and proteomics.
- We plan *for the first time* to monitor at whole-genome scale the response of yeast cells to supply conditions for a key nutrient of all cellular systems: *nitrogen*.
- Yeast cells are grown under as wide a range of nitrogen supply conditions as possible and the state of their *transcriptome* will be monitored using the *DNA microarray* method.
- Numerical data generated are interpreted by combining statistical, static and dynamic *modelling methods*.
- The project aims to *reverse engineer* the *complete molecular network* of nitrogen metabolism in yeast.

Applications

- Inferring methods developed in this project can be applied:
  - to other genomes, notably the *human genome*.
  - to other domains characterised by numerous features and few observations, notably *web mining*.

Data Mining in the context of anesthesia

- **The Partners of the project** :
  - Machine Learning Group of ULB
  - Anesthesia department of Hospital Erasme
  - Company Mexys SA (from Mons)
  - Laboratorio di Intelligenza Artificiale e Robotica del Politecnico di Milano
- Funding organisation: Région Walonne – "1st Europe Project"
- Duration: 2004-2006

Research content

- The goal of the project is to develop **data mining techniques** to analyse the **massive datasets** collected by the "TOOLBOX" software, currently used by the **anestehesists of the Erasme hospital** during the surgical interventions.

- Data concerns the monitored state of the patient (e.g. blood pressure, the rate of heartbeat), the type and concentrations of drugs, the actions of the anesthesist.

Applications

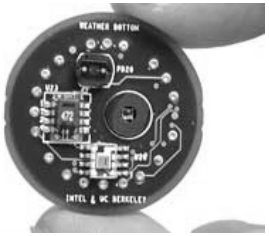
- Build **predictive models** that assist doctors with high level information about the course of the operation.
- Build **visualization application** to help doctors extracting useful information from measured medical data.
- Provide **statistical expertise** to support clinical studies for the certification of new protocols.

Projet COMP²SYS

Computational intelligence methods for complex systems

- Partners:
  - ULB - IRIDIA: Artificial Intelligence laboratory (Swarm robotics, Optimization)
  - ULB - USE: Unit of Social Ecology (Insect behavior, Interaction between robots and insects)
- Funding organisations: European Commision, 6th framework program
- Duration: 2004-2008

Wireless sensor networks



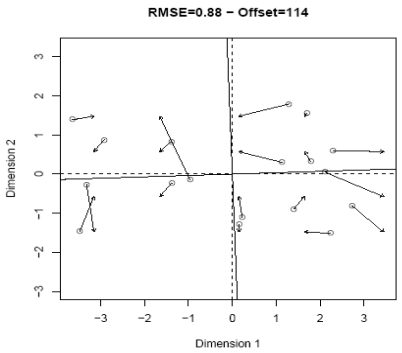
- Wireless sensor modules:
  - Ultra low power devices, with CPU, memory, antenna and sensors
  - Main constraints:
    - **Limited bandwidth for communication**
    - **Limited energy ressources**
- Research axes:
  - Sensor Localization (**manifold learning**)
  - Network organization for efficiently extracting and routing information (**optimization**)
  - In network data processing algorithms (**feature extraction, compression**)

Applications

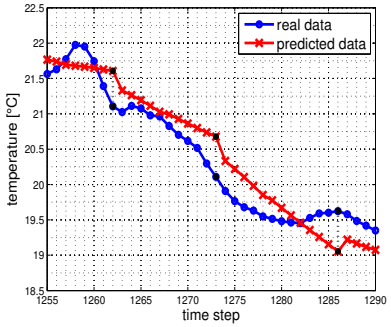
- Environmental monitoring
- Medical health care
- Battlefield surveillance
- Industrial, structural monitoring (e.g. Breakdown prevention, fault diagnosis)
- Smart homes (Air conditioning control, intelligent lighting)
- ….

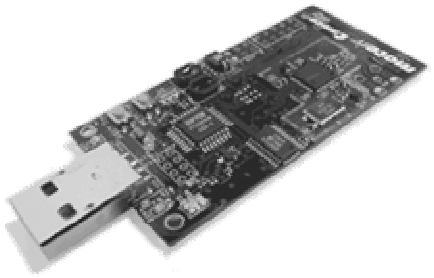
Current work

- Localization
  - Given a distance matrix
  - Determine relative positions of all sensors in 2D/3D space
  - Manifold learning



- Adaptive sampling
  - Autoregressive time series prediction models
  - Only send model parameters to the central server





Tmote Sky:

- 8MHz Microcontroller
- 48KB Flash, 10KB RAM
- 512KB Data storage
- 250Kbps Zigbee Radio

Programming:

- TinyOS
- NesC

Internship proposal 1

Linear spatial modelling

Context:

- Continuous monitoring of an environment
  - All data required to be sent to an external server
  - Some approximation tolerated on retrieved data

- Correlated sensor measurements allow to compress data by using statistical learning techniques

Linear spatial modelling

- Divide the network into clusters
  - $N=C_1 \cup C_2 \cup \ldots \cup C_p$
- In each cluster C_k, $1 \le k \le p$, collect the set of measurements S^{Ck} for all sensors in the cluster
  - $S^{Ck} = \{s_i^{Ck}\}, 1 \le i \le |C_k|$
- Find linear model for measurements
  - $\hat{s}_j^{Ck} = \mu_k + \alpha_k x_{jk1} + \beta_k x_{jk2}$
  - x_{jk1} and x_{jk2} are the coordinates of sensor s_j^{Ck}
- Send linear model μ_k, α_k, β_k parameters to the central server

Internship proposal 1

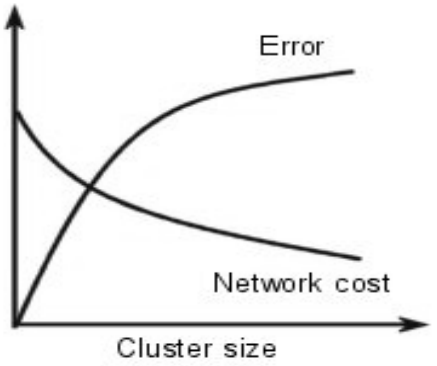
Linear spatial modelling

Goal

- Implementation on Tmote Sky using open source TinyOS system

- Study of tradeoffs between
  - Approximation accuracy
  - Network/Energy costs

  using different clustering policies



- Extension to other compression schemes

Internship proposal 1

Required expertise

- Theory:
  - Matrix algebra
  - Statistics
- Practice:
  - Java and C programming
  - R or Matlab

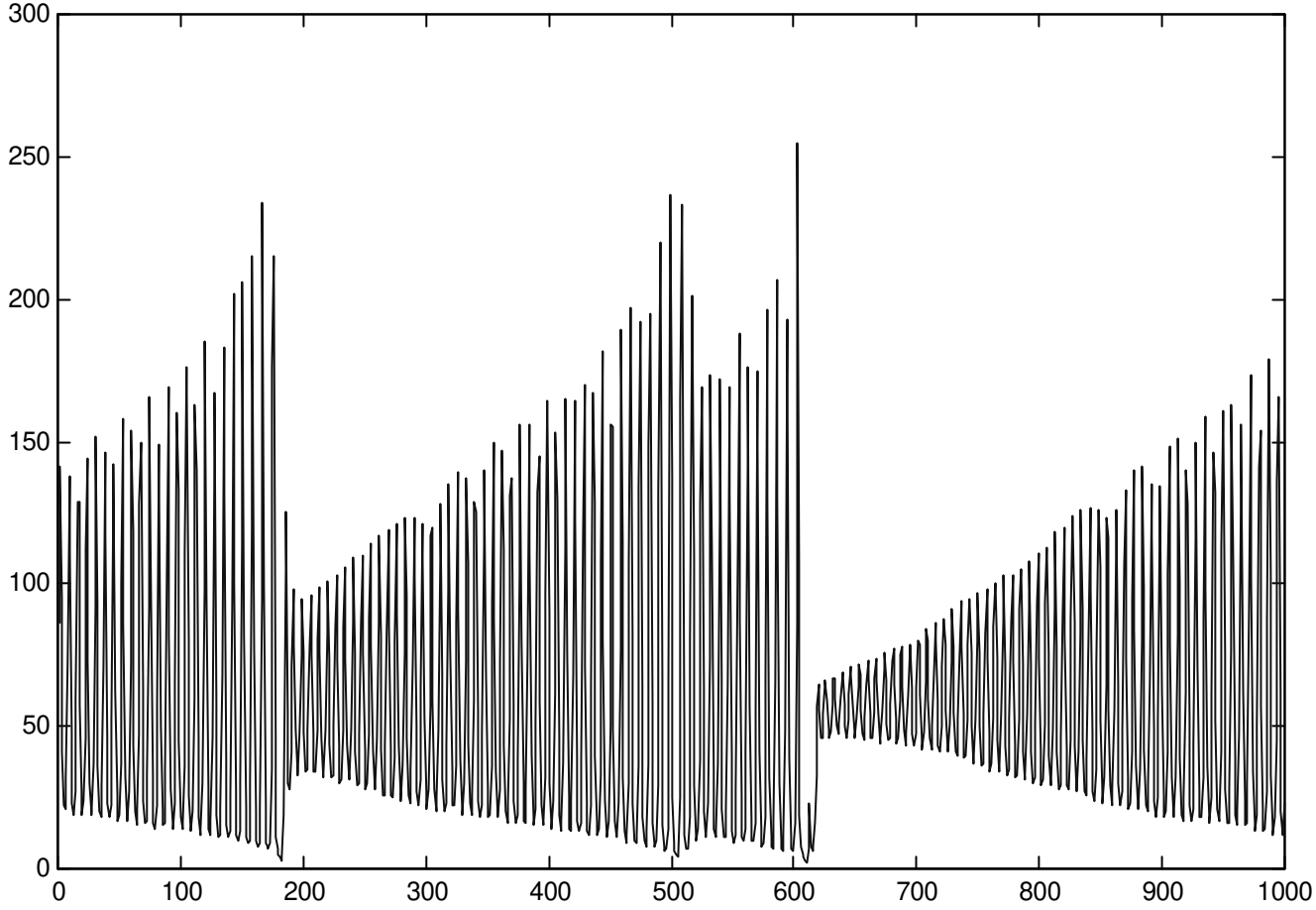
Will provide skills in:

- Statistical learning
- Embedded system programming

Internship proposal 2

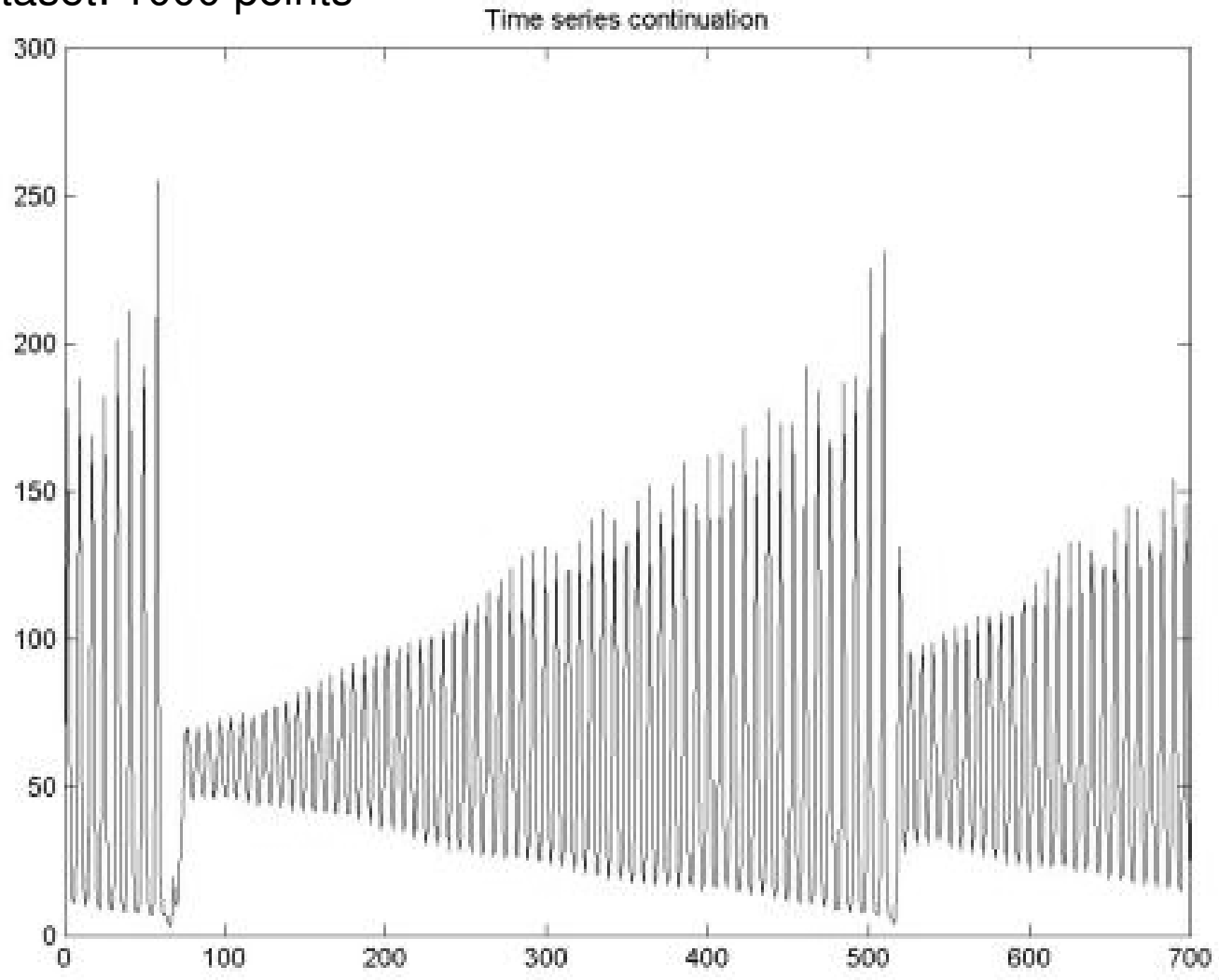
- Context: forecasting a (chaotic) time series over a very long horizon. One of the most important unresolved issue is the accumulation of prediction errors.
- Goal: combine machine learning method and dynamical system theory to improve the performance of time series predictors.
- Theoretical issues:
  - Find the right complexity of local learner (bias/variance trade-off)
  - Take advantage of theoretical results on linear dynamical systems (conservative/dissipative systems, linearization)
- Required expertise: dynamical systems, statistical methods, computational programming (e.g. MATLAB, R)

Santa Fe chaotic time series



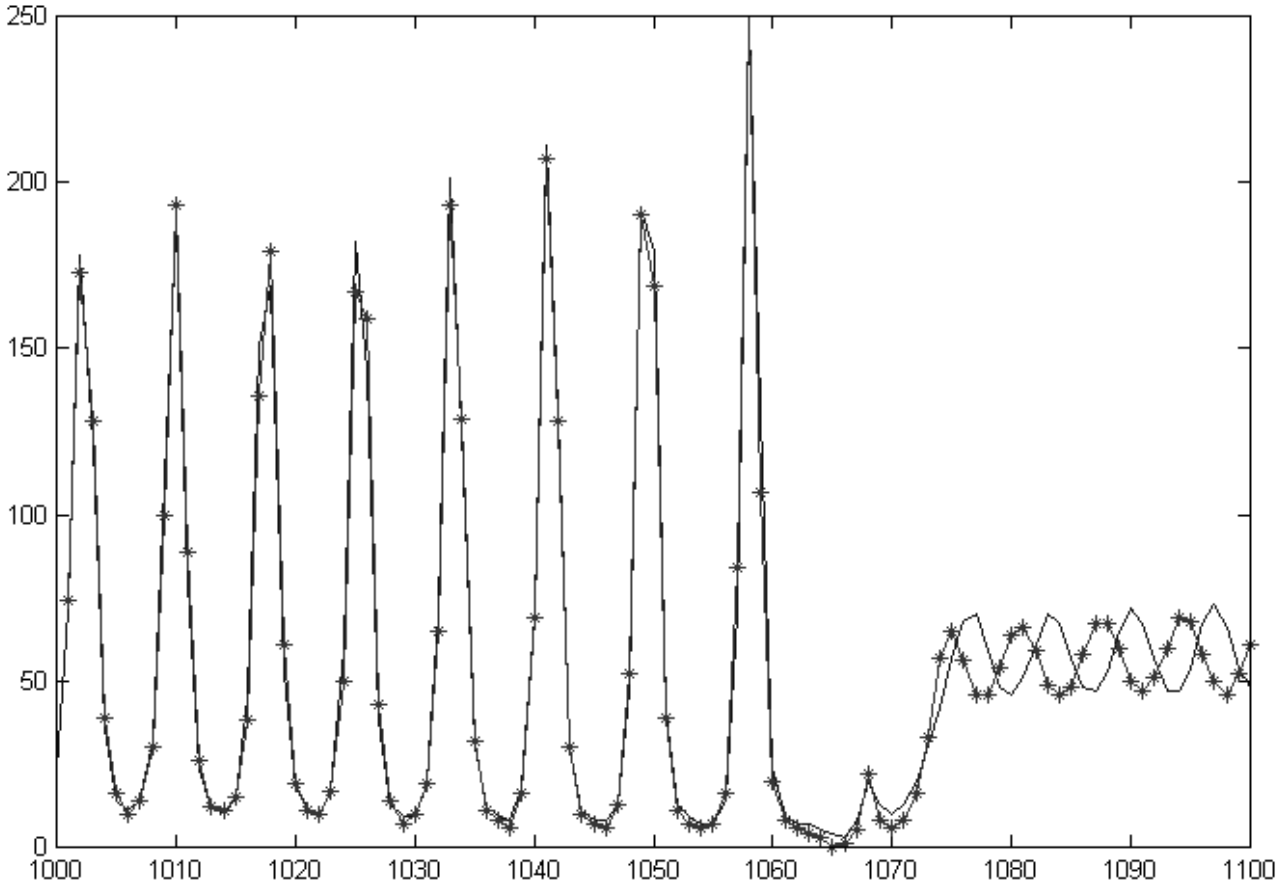
Training dataset: 1000 points

Test dataset: 1000 points



Competition task: predict the continuation of the series for the first 100 steps.

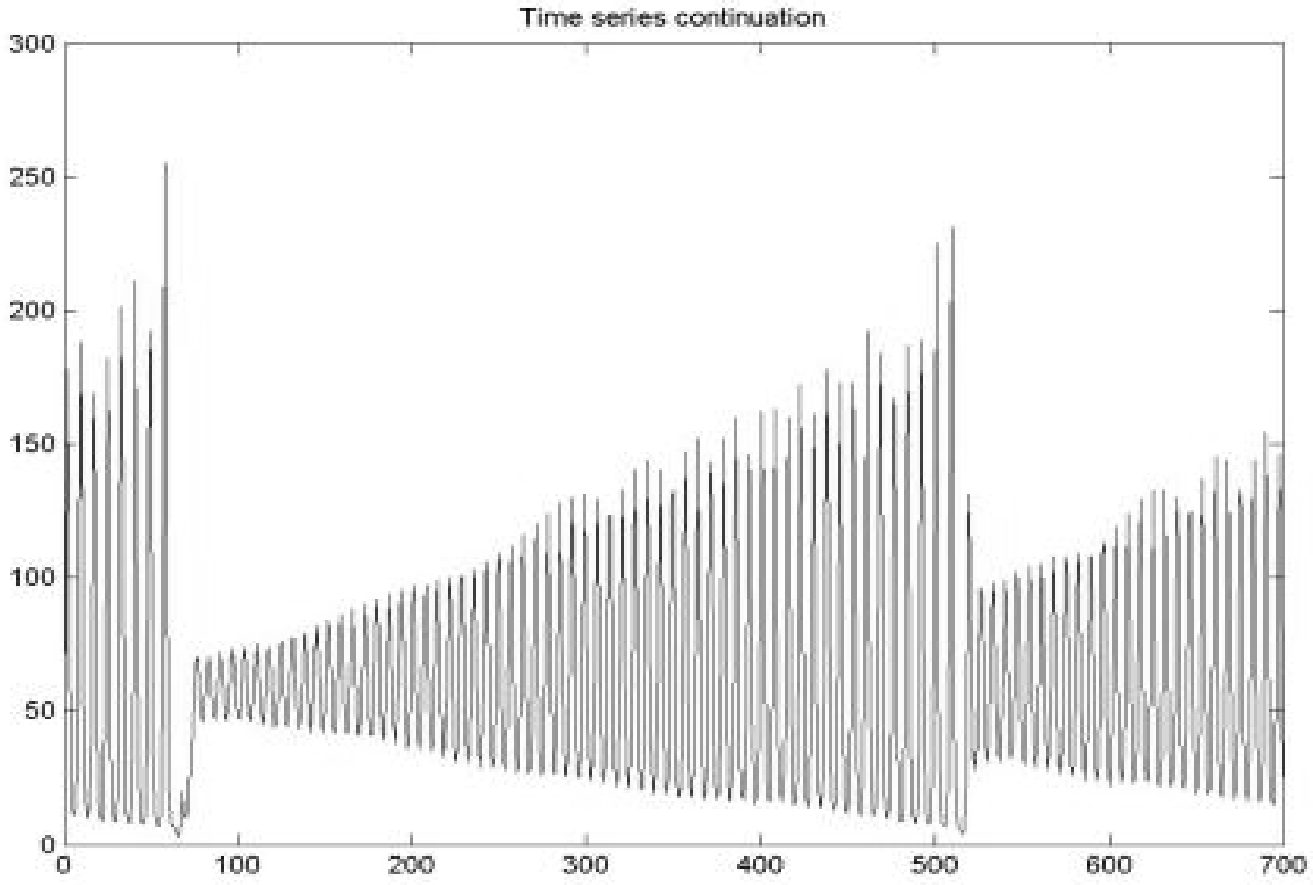
Local Learning prediction



LL is able to predict the abrupt change around t =1060 !

(*Bontempi et al. (1999). Local learning for iterated time-series prediction Machine Learning: Proceedings of the Sixteenth International Conference*)

Very long iterated prediction



Some preliminary results (Bontempi et al. (1999). A multi-step-ahead prediction method based on local dynamic properties. in Proceedings of ESANN 2000, European Symposium on Artificial Neural Networks).

Awards in international competitions

- Data analysis competition: awarded as a runner-up among 21 participants at the 1999 *CoIL International Competition* on *Protecting rivers and streams by monitoring chemical concentrations and algae communities*.
- Time series competition: ranked second among 17 participants to the *International Competition on Time Series* organized by the *International Workshop on Advanced Black-box techniques for nonlinear modeling* in Leuven, Belgium