PERSYVAL-Lab

ANR

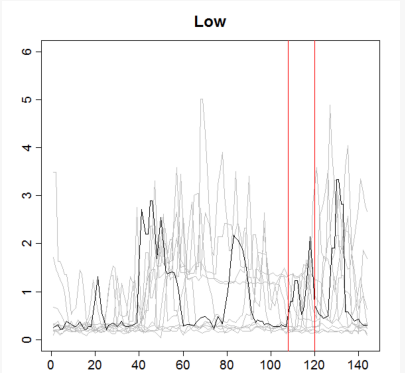
ADM – Advanced Data Mining

Develop algorithms and methods for extracting knowledge from hybrid physical-digital data

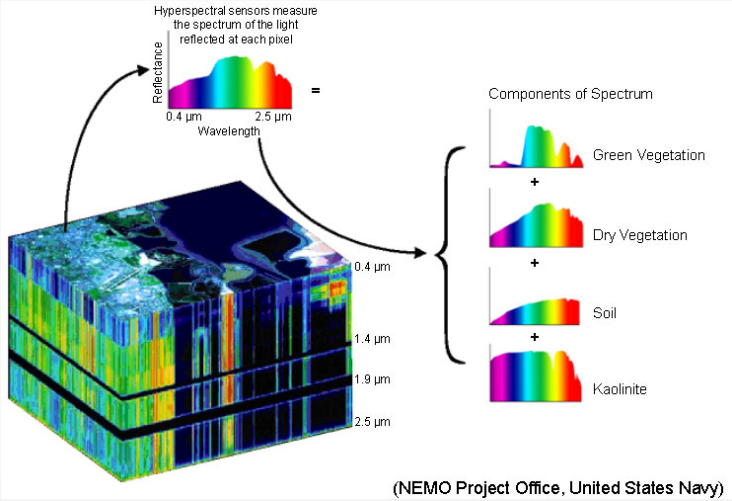
Eric Gaussier (LIG), Christian Jutten (GIPSA)

Context: Big Data

- Data ubiquitous, multi-form, multi-source, multi-scale (heterogeneity, multimedia, sensor networks)
- Distributed, on-line (streams)
- Complex: structured and dynamic
- Large-scale (in all dimensions)



Energy consumption



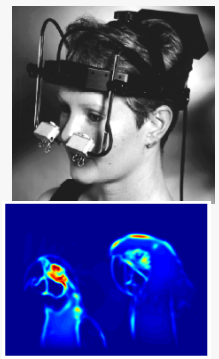
Wikipedia

Context: Data Science

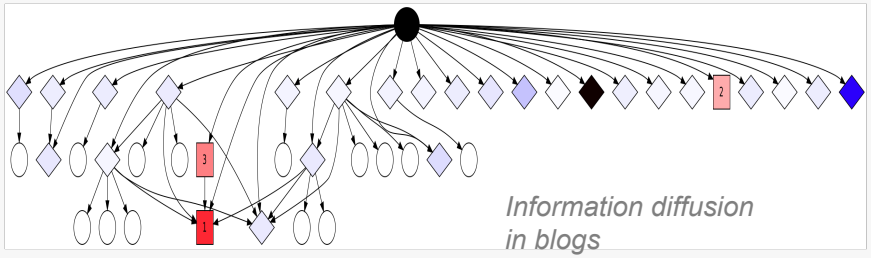
(from signal to knowledge)

- From static, unstructured data acquisition and mining to the acquisition and modeling of streams of interconnected, dynamic data flows
- From sensor networks, signal/data processing to knowledge extraction

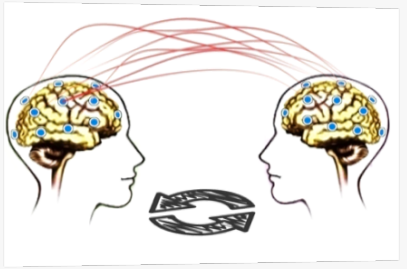
Act on the physical world and take appropriate decisions



Scrutation of visual scenes from static, dynamic, audiovisual stimuli



Information diffusion in blogs



EEG Hyperscanning

Scientific Objectives

On-line mining of distributed data flows

1. On-line mining of data streams
   - Dynamical structures, non-stationarity (non i.i.d.), adaptive models
2. Mining multi-* (source, scale,modal,…) data
   - Jointly mining and analyzing various types of data, including texts, images, audio, gestures, …
3. Decentralized data mining
   - Combining partial data models resulting from local data analyses

Main Milestones

1. MilestoneT0+2years
   - Generic data mining algorithms
   - Traces from different sources (PCS, AAR, SIM)
2. MilestoneT0+4years
   - Machine learning/data mining models for complex, dynamic, multimodal and multi-source data
   - Complex data and composite events
3. MilestoneT0+8years
   - Distributed data mining compared to centralized version (performance, speed, consumption)

Focus and Illustration (1)

On-line mining of data streams (pattern mining, clustering, categorization, knowledge extraction, …)

General setting

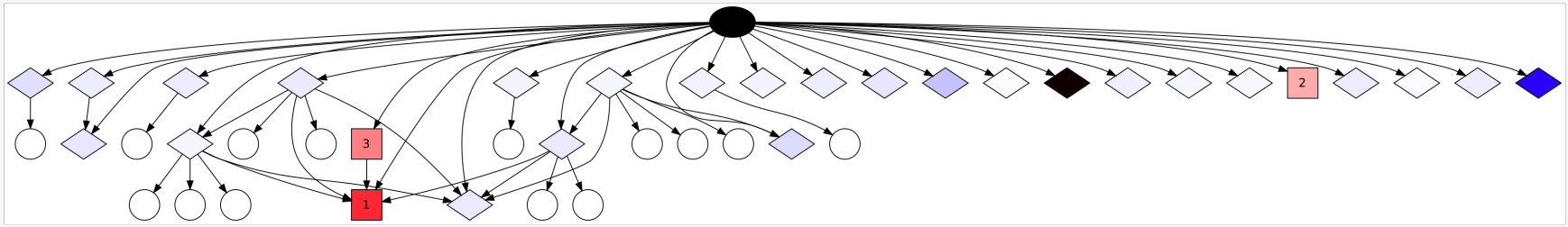
1. Learn model $\mathcal{M}(T)$ on $(x^{(i)},y^{(i)})_t$ $(1 \le i \le N,\ 1 \le t \le T)$
2. $\mathcal{M}(T)$ adaptive model: $\mathcal{M}(T+1) = \mathcal{M}(T) + f((x,y)_{(T+1)})$
3. On-line to batch conversion (optional)
4. Fast inference on $\mathcal{M}$

Focus and Illustration (2)

Modeling information diffusion in social network

1. Information diffuses in the network from initial user(s)
2. Diffusion path depends on information content, users' profiles and willingness to diffuse, social pressure
3. Learn (on-line) $\mathcal{M}(T)$ from past diffusions
4. Predict how current information diffuses at time t', $t' > T$

Information diffusion in blogs



Focus and Illustration (3)

Scientific and technological challenges

1. Observed data *x* (and *y*) complex (beyond vectors)
2. Non stationarity (no i.i.d.):
   - $\mathcal{M}(T+1) = g(\mathcal{M}(T)) + f((x,y)_{(T+1)})$ but *g* unknown and few theoretical results (transfer learning, domain adaptation)
3. Multiple sources, multiple scales
   - Joint representations vs joint models; source separation
4. Scalability issues
   - Sparse models but trade-off accuracy/sparsity
   - Parallel versions

Some ADM actions

- Standard animation activities (discussions, reviewing, …)
- Scientific meeting held end of Nov. to know each other better and foster collaborations
  - General presentations
  - “Speed-dating” sessions
  - Round tables
  - Organization: S. Amer-Yahia, L. Besacier, J. Chanussot, M. Rombaut & A.-L. Bernardin, I. Maugis

Scientific Committee

Sihem Amer-Yahia (LIG)
Jocelyn Chanussot (GIPSA)
Olivier François (TIMC)
Anne Guérin (GIPSA)
Zaid Harchaoui (LJK/INRIA)
Nabil Layaida (LIG/INRIA)
Olivier Michel (GIPSA)
Michèle Rombaut (GIPSA)

Laurent Besacier (LIG)
Ahlame Douzal (LIG)
Eric Gaussier (LIG)
Pierre-Yves Gumery (TIMC)
Christian Jutten (GIPSA)
Vanda Luengo (LIG)
Frédéric Pétrot (TIMA)