DataSense: Data Intelligence

Céline Hudelot & Bertrand Thirion

MICS Laboratory & Inria Saclay

DigiCosme Research Days
Gif, 5–6/6/2018

DataSense: Data Intelligence

- Data are everywhere
  - The Web, information systems, sensors (smartphones, etc.)
  - Knowledge Graph, Wikipedia, WordNet...
- Big data makes new things possible
  - Statistical machine translation
  - e-science
  - Business intelligence
  - ...

Research Teams in DataSense

- DAVID — Université Versailles St Quentin
- Inria Saclay
- L2S — CentraleSupélec
- Li-PaRAD — Université Versailles St Quentin
- LIMSI — CNRS - Université Paris-Sud
- LIST — CEA
- LIX — École Polytechnique
- LMV — Université Versailles St Quentin
- LSV— ENS Paris-Saclay
- LRI — Université Paris-Sud
- LTCI — Télécom ParisTech
- MICS — CentraleSupélec
- SAMOVAR — Télécom Paris-Sud
- U2IS — ENSTA ParisTech

Plan

1. Structuring

2. Activities

3. Perspectives

DataSense ‘Tasks’

1. Scalable, expressive and secure tools for large-scale data
2. Making sense of complex, heterogeneous data and knowledge
3. Machine learning: multi-task learning and meta-learning
4. Distributed decision making: reinforcement learning, partially observable processes and games
5. Interaction and Visualization

Global DataSense Activities

- **Main global activities**
  - Research days (once a year)
  - Community building
  - Preparation of working groups and co-supervised theses
  - Missions doctorales
- **Activities per 'Task'**
  - Working groups
  - Co-supervised PhD theses
  - Induced collaboration, guest scientists
  - Post-doc and engineers
- **Emerging projects**
  - SEDA: Breathing sense into data
  - Neural Meta Tracts
  - IoTA Internet of Things Analytics

DataSense Working Groups

Bottom-up creation of DigiCosme working groups

- **ERVEN** : Extraction, Représentation et Visualisation de connaissance pour l'Enseignement Numérique – 2018 / 2019
  - Anne-Laure Ligozat
- **E-santé** : Internet des Objets & E-santé – 2017 / 2019
  - Mehdi Amni
- **TAL & SEM** : Traitement sémantique des données Textuelles – 2017 / 2018
  - Brigitte Grau, LIMSI
- **SDT** : Sécurité des Données textuelles – 2016 / 2017
  - Cyril Grouin, LIMSI
- **D2K**: From Data to Knowledge – 2015/2017
  - Claire Nédellec, INRA; Chantal Reynaud, LRI
- **SSSL** : Sequential Structured Statistical Learning – 2015 / 2017
  - Oldaric Maillard, LRI

DataSense Working Groups

Bottom-up creation of DigiCosme working groups

- **SciCoSense:** Building a cartographic map of scientific communities – 2015/2016
  - Philippe Caillou, LRI
- **Human-Robot Interaction** – 2015/2016
  - Laurence Devillers & Jean-Claude Martin (LIMSI-CNRS)
- **Deep Learning and Distributed Representations** – 2014/2018
  - Alexandre Allauzen, LIMSI
  - Emmanuelle Frenoux, LIMSI
- **PASADENA**: Prédiction et Analyse de données structurées et hétérogènes – 2015/2018
  - Arthur Tenenhaus, CentraleSupélec
  - Maxime Sangnier, TPT
  - Flora Jay, LRI

Guest Scientists

- **Kevin Bretonnel Cohen** - Director, Biomedical Text Mining Group Computational Bioscience Program / University of Colorado School of Medicine - February / April 2016
  Contact : Pierre Zweigenbaum
- **Catherine Plaisant** - University of Maryland Institute for Advanced Computer Studies - June / July 2017
  Contact : Jean-Daniel Feketé (Inria)
- **Timothy Miller** - Boston Children's Hospital & Harvard Medical School - Juin / Juillet 2017
  Contact : Aurélie Névéol (LIMSI)
- **Ramon Pino Pérez** - Department of Mathematics, University of the Andes (Venezuela) - April/ June 2018
  Contact : Isabelle Bloch (Télécom ParisTech)

PhD students

- **HiDimStat** : Statistical control of sparse models in high dimension - 2017
  Bertrand Thirion (Inria) & Joseph Salmon (LTCI)
- **Idiab** : Internet des Objets pour le suivi et la modélisation de la glycémie de patients diabétiques - 2017
  Mehdi Ammi (LIMSI)
- **AlCoMol** : Algorithmique de graphes pour l'aide á la décision dans la construction moléculaire - 2016
  Domnique Barth (DAVID)
- **OPALE** : Opérateurs monotones aléatoires et applications á l'optimisation stochastique - 2015
  W. Hachem, P. Bianchi, J. Jakubowicz, LTCI/ SAMOVAR
- **COT** : Coréférence événementielle cross-document dans les dossiers électroniques patient - 2015
  A. Névéol, X. Tannier, O. Ferret, LIMSI / CEA-LIST
- **SEDA** : Breathing sense into data - 2015. Fabian Suchanek, LTCI
- **SensoMotor-CVE** : Murs d'images en contexte interactif complexe – 2015
  Patrick Bourdot, LIMSI
- **BIPIMA** : BIPolarité de l'Information Multimédia pour l'Annotation sémantique d'images dans un contexte de médias sociaux - 2014

Post-doc and engineers

- **MAEL** : MultimediA Entity Linking - 2018
  Contact : Hervé Le Borgne, CEA LIST
- **VASTE** :Veracity Assesment in Spatio-TEmporal heterogeneous data.
  Contact : Fatiha Sais, LRI
- **MetaTracts** : Parsimonious multi-resolution representations for statistically analyzing brain tractograms - 2018
  Contact : Pietro Gori, LTCI - LIX
- **PASADENA** – 2017
  Contact : Remy Boyer & Franck Nielsen, L2S
- **AMPHI** : Approximate Message Passing for HIgh-dimensional data - 2017
  Contact : Bertrand Thirion, Inria & Jospeh Salmon, LTCI
- **TAL & SEM** : Traitement Sémantique des Données Textuelles - 2017
  Contact : Brigitte Grau & Olivier Ferret, LIMSI
- **D2K** : De la Donnée à la Connaissance - 2017
  Contact : Jean-Daniel Feketé, Inria
- **SEDA** - 2015. Contact : Fabian Suchanek, LTCI

Doctoral missions

- **Plateforme Camomile d'annotation collaborative de documents multimédia** - 2016
  Direction : Hervé BREDIN, LIMSI
  https://github.com/camomile-project/camomile-polymer-client
- **Software for easier access to open M/EEG data repositories** - 2016
  Direction : Alexandre GRAMFORT; LTCI
  https://github.com/jasmainak/bids-validator

From Data to Knowledge

Contact: Claire Nedellec (INRA) & Chantal Reynaud (LRI)

Activities: 12 laboratories involved, 18 teams, 64 registered researchers, about 20 participants per meeting

→ *CS for modeling living organisms* → priority topic in the Life Sciences Department SGT5

→ Issued a document included in the *Life Sciences Department White Paper*

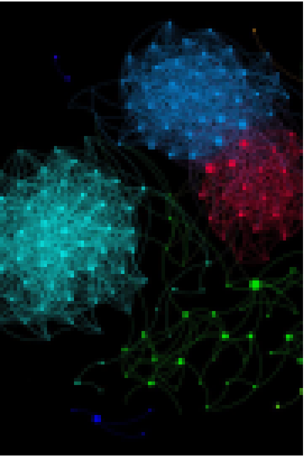
- Led to ANR-DFG project GoASQ (ANR-DFG: LRI-LIMSI-TUD, 2015–2019): *Generating and Answering Ontological Queries over Semi-structured Data*
- Contribution to the B2SRI Strategic Research Institute application
  - Systems Biology and Synthetic Biology for Research and Innovation (Life Sciences + CS)
- Two teams collaborate in H2020 E-Infra OpenMinTeD: Open Mining Infrastructure for Text and Data (2015–2018)
- DigiCosme Invited Professor: Kevin B. Cohen (U. Colorado), text mining in biomedicine (3 months, 2016)

http://labex-digicosme.fr/GT+D2K

Building a cartographic map of scientific communities (SciCoSense)

Contact: Philippe Caillou, LRI

http://labex-digicosme.fr/GT+SciCoSense



Representation and study of social networks:

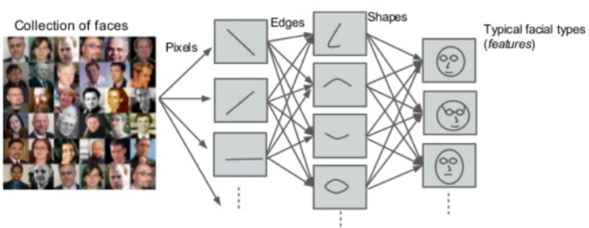
- Study scientific communities
- based on traces of their activities
- to build indicators, maps and query tools
- about scientific production

Deep Learning and Distributed Representations

Contact: Alexandre Allauzen, LIMSI ; Emmanuelle Frenoux, LIMSI

http://labex-digicosme.fr/GT+Reseaux+profonds

DataSense Task 3 (Machine learning)



Topic:

Deep neural networks and representation learning

Activities:

- Journal club
- Cross-team presentations
- Invited seminars

Participants:

LIMSI, CNRS (TLP, AMI) — LRI, CNRS & UPSud (TAO) — U2IS, ENSTA — LTCI, Télécom-ParisTech

Prédiction et Analyse de données structurées et hétérogènes (PASADENA)

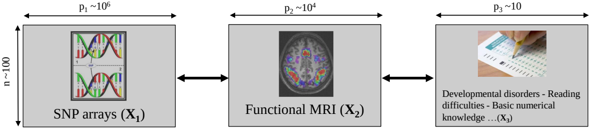
Contact: Arthur Tenenhaus, Centrale-Supélec ; Maxime Sangnier, Télécom-ParisTech & Flora Jay, LRI

http://labex-digicosme.fr/GT+PASADENA

Objectives

Developpement of statistical methods for complex data analysis: **heterogeneous**, **multimodal** and **structured** data:

- unsupervised analysis of correlations between modalities;
- classification/regression from heterogeneous data;
- structured prediction to fit a certain type of data from another;



$p_1 \sim 10^6$

$p_2 \sim 10^4$

$p_3 \sim 10$

$n \sim 100$

SNP arrays ($\mathbf{X}_1$)

Functional MRI ($\mathbf{X}_2$)

Developmental disorders - Reading difficulties - Basic numerical knowledge …($\mathbf{X}_3$)

Multibloc study in imaging genetics.

Human-Robot Interaction

Contact: Laurence Devillers & Jean-Calude Martin (LIMSI-CNRS)

http://labex-digicosme.fr/GT+Interaction

Objectives

- Create a community on interactive robotics; verbal/non-verbal/physical Human-robot interactions
- Group specialists beyond robotics: modeling and interaction, big data, psychology, social and cognitive sciences, ergonomy and usage, ethics.

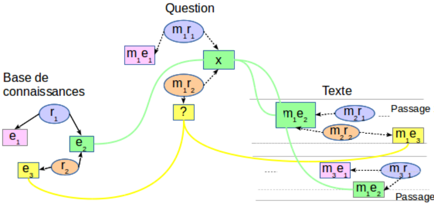
Members : LIMSI, ENSTA, CEA, Télécom Sud-Paris, Télécom ParisTech, Télécom Ecole de Management, Université Paris-Sud, UVSQ, Université d'Evry, CERDI

ERVEN : Extraction, Représentation et Visualisation de connaissance pour l'Enseignement Numérique

Contact: Anne-Laure Ligizat

https://digicosme.lri.fr/tiki-index.php?page=GT+ERVEN

- Équipes du LIMSI :
  - Équipe ILES (Brigitte Grau, Gabriel Illouz, Anne-Laure Ligozat)
  - Équipe AMI (Frédéric Vernier)
  - Équipe TLP (Alexandre Allauzen)
- Équipes du LRI :
  - LaHDAK: Philippe Dague, Yue Ma, Brigitte Safar, Fatiha Saïs
  - MODHEL:Yolaine Bourda, Fabrice Popineau
- SAMOVAR, Telecom SudParis : Amel Bouzeghoub



E-santé : Internet des Objets and E-santé – 2017 / 2019

Contact : Mehdi Ammi

https://digicosme.lri.fr/tiki-index.php?page=GT+Internet+des+objets+et+E-sante

- étude, conception et évaluation des services e-sante´
- traitement des données
- technologie
- éthique

- Laboratoires *TIC*: LIMSI(AMI, CPU, ILES, TLP), ENSTA ParisTech, CIAMS, Telecom SudParis, CEA-LIST LRI
- Laboratoire *Santé* End-icap Fondatior Helene - Poidatz Handiresp Hôpital Bicêtre
- *Associations et pôles*: RevesDiab France eHealthTech Systematic Capdigital OpticsValley Fedev (SFR)

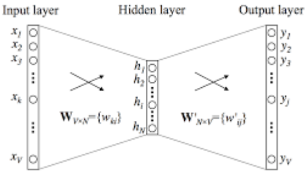
TAL & SEM : Traitement sémantique des données textuelles – 2017 / 2018

Contact : Brigitte Grau, LIMSI

https://digicosme.lri.fr/tiki-index.php?page=GT+TAL+et+SEM

Implication textuelle, paraphrase, désambiguisation sémantique

- **Timothy Miller**: Boston Children's Hospital, Introduction to sequence models for Natural Language Processing
- **Brigitte Grau**: LIMSI, De la recherche de réponses à des questions à la compréhension ciblée de textes
- **Olivier Ferret**: Apprentissage de connaissances sémantiques : adaptation de plongements lexicaux (words embeddings) à des connaissances externes



SDT : Sécurité des Données Textuelles – 2016 / 2017

Contact : Cyril Grouin, LIMSI

https://digicosme.lri.fr/tiki-index.php?page=GT_SDT

Common group with Scilex

- LIMSI : équipe ILES (Cyril Grouin (CNRS), Thomas Lavergne (Université Paris-Sud), Aurélie Névéol (CNRS), Pierre Zweigenbaum (CNRS)
- INRIA-LIX : équipe COMETE (Catuscia Palamidessi, Kostantinos Chatzikokolakis)
- CEA-LIST, équipe LVIC (Olivier Ferret, Gaël de Chalendar)

- Anonymisation et risques de réidentification
- Protection des données dans les modèles
- Optimisation de la confidentialité différentielle

SSSL : Séquential Structured Statistical Learning – 2015 / 2017

Contact : Oldaric Maillard, LRI

https://sites.google.com/site/groupedetravailsssl/home

Plan

1 Structuring

2 Activities

3 Perspectives

DataSense Perspectives

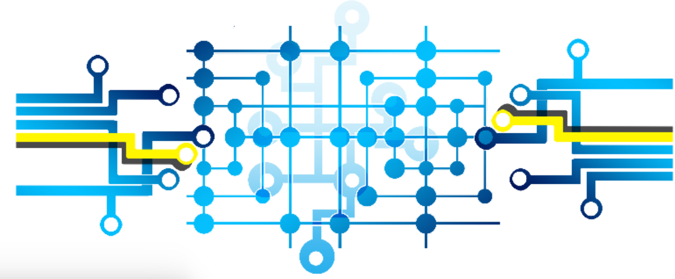
Within the STIC Domain

- Continuation of the support of collaborative research in the DataSense tasks: both fundamentals and applicative works.
- Stronger interactions with ComEx and SciLex
  - e.g. develop information-theoretic analysis of deep learning
  - e.g. formal methods meets machine learning, towards safe AI-systems, machine learning for optimizing ontology reasoning...
- Stronger coordination with DataIA (convergence institute) + Center for Data Science + Optimal Control & signal communities
  - joint calls
  - (junior) summer schools, software development
  - Future AI institute (?)

DataSense Perspectives

Beyond STIC: given the maturity of the domain, time to open to new communities

- Open Digicosme to new labs (e.g. INRA maIAGE; INRA mia)
- E-science: have more common projects with other departments
  - Life science
  - physics
  - (?) SHS
  - Maths, theoretical physics



THANK YOU!