The Data Science Team

Dario Colazzo

HCERES visit, 2012-2017

Data Science today

- At the intersection of several domains, typically:
  - applied mathematics, in particular
    - statistics and machine learning
  - computer science, in particular
    - algorithms and programming, eventually distributed and parallel
    - massive data management and analytics
    - services
- In most contexts, deep knowledge of a specific application domain (public policies, finance, ...) is crucial for effective interpretation of result analysis, and to guide the data analysis process itself.

Data Science at LAMSADE

- The team includes LAMSADE scientists having expertise in all of typical Data Science domains
- It is a young team, officially founded in 2015
- The scientific and training activity of the team can currently count on:
  - 5 Professors, 1 CNRS DR and 2 Emeritus Professors
  - 7 Maître de Conférences and 1 Past teacher
  - 13 PhD students

Projects and team participants

- Massive Data Management, Analysis and Exploration - MADAX
  - Dario Colazzo (PR), Khalid Belhajjame (MDC)
- Machine learning
  - Jamal Atif (PR), Yann Chevaleyre (PR), Marta Rukoz (PR), Benjamin Negrevergne (MDC), Sana Benhamida (MDC-E), Florian Yger (MDC)
- Services
  - Daniela Grigori (PR), Marta Rukoz (PR-E), Khalid Belhajjame (MDC), Maude Manouvrier (MDC), Joyce El Haddad (MDC), Michel Zam (PAST)
- Policy Analytics
  - Alexis Tsoukias (DR), Elsa Negre (MDC)

Focus on MADAX results

- Large scale management and analytics of semistructured data
  - Topics and challenges:
    - Distributed evaluation of XML queries and updates. XML partitioning and query/update compilation over paradigms based on map-reduce
    - Lenses for OLAP analytics over semantic graphs. Make OLAP-like operators cope with heterogenous RDF graphs with implicit information.
    - Schema inference for massive JSON data sets. Design of associative data summarisation algorithms, finding a good compromise between succinctness and precision.
- Main contributor D. Colazzo.
- A*/A-CORE publications in WWW'14, SIGMOD'15, TKDE'15 , EDBT'17
- Participation to 1 ANR project (300K euros, 2017-2021).

Focus on MADAX/Services results

- Data examples for data integration and scientific module management.
  - Topics and challenges:
    - Heuristics for schema mapping selection based on feedback provided by users on data examples. Exponential number of candidate mappings, how to cluster users for mapping recommandation
    - Data examples for concise description of scientific modules. Find minimal data samples describing module behaviour, and helping for effective retrieval and comparison operations among modules
    - Selection data examples for testing of scientific module annotation. Large sets input/output examples, how to use semantics annotation to guide the selection of minimal and representative samples for testing
- Main contributor K. Belhajjame. A-CORE publications in Distr. and Parallel Databases '15, SSDBM'16, EDBT'14
- Participation to 1 PEPS project on Data Privacy, 1 MADICS action on bio-informatics.

Focus on Services results

- Effective service composition
  - Topics and challenges:
    - 0-1 linear programming for optimal and automatic transactional service composition. Exponential search space for automatic service composition, transactional services make the problem even harder.
    - SPARQL framework for searching data and services. Finding the right SPARQL queries enabling the search of pertinent services for a given SPARQL query, ensure good precision and recall level.
    - Complexity analysis for QoS-aware service selection. Finding the right reductions for complexity analysis for a complete treatment: including QoS requirements, AND/OR/XOR workflows, multicriteria QoS properties
    - Dynamic recovery strategy for composite services execution. How to measure the work done by a CWS execution and its compliance with QoS requirements for dynamic fault tolerance.
- Main contributors Daniela Grigori, Marta Rukoz, Joyce El Haddad, Maude Manouvrier.
- A-CORE publications ICSOC'14, ICSOC'15, WISE'15, WWWJ'16
- Participation to 1 ANR project (80K euros, 2014-2018)

Scientific production, animation and international visibility

- 3 Books, 14 Book chapter, 43 journal papers, 109 conference papers (More than 20 A*/A-CORE publications)
- Since 2015, more than 20 seminars covering all projects
- More than 10 national and international invited professors
- International visibility
  - Program co-chair of ICSOC 2015 (D. Grigori), Workshops co-chair ICSOC 2014 (D. Grigori), Program Chair of ISATP 2013 (M. Rukoz), Track chair in MEDES'17 (M. Rukoz), Publicity chair of BPM 2018 (D. Grigori), Member of the editorial board of ISI - Ingenierie des Systemes d'Information (D. Grigori).
  - Invitation to ACM Fellows Award meeting (W. Litwin)
  - More than 50 PC invitations, including prestigious venues like PODS, SIGMOD, VLDB, ICSOC, SSDBM, EDBT, NIPS, AISTAT, IJCAI, ...
  - Organisation of ICSOC'14 here at Dauphine (led by Services project, co-chaired by D. Grigori).

Collaborations and interaction with other domains

- Collaborations with several national an international universities in Europe, US, Africa, Asia, Australia
- With companies, including CIFRE thesis, such as
  - SNCF, Adway, Horthonworks, IBM, Google, Coheris, WaveStone, AgiLap, ...
- Design of techniques that find applications in medical diagnosis (ANR project , Y. Chevaleyre), art style recognition (B. Negrevergne, F. Yger), sociology and econometrics (ANR project, D. Colazzo)

Data Science training programs

- Deeply involved in Bachelor and Master program re-orientation towards advanced Big Data and Data Science topics.
- Strong participation in the Research Master ISI (J. Atif, D. Colazzo, D. Grigori, M. Rukoz, Khalid Belhajjame).
- Founded and coordinating the Data Science Certificate at Dauphine (K. Belhajjame), two editions this year.
- Tutorial at IJCAI 2016 (J. Atif)
- Participation to teaching activities and Master program orientation, in other institutions like École Polytechnique (D. Colazzo) and École Centrale (J. Atif).
- Organisation of a summer school Microservices & Big Data Management (M. Rukoz, M. Manouvrier, J. El Haddad).

Perspectives

- Data Science in next years?
- Difficult to predict, but probably:
  - Even much more massive and heterogenous data sets to explore, integrate and process (e.g., IoT, robot data)
  - New kind of data and IA services to analyse and to compose
  - Privacy in data analysis and machine learning
  - New HW architectures (e.g. TPUs, neural processors) will call for new programming and computation paradigms
- The team has expertise to face the near future (r)evolutions
  - Effective exploration, integration and analysis of massive semi-structured data sets
  - Self-healing execution of service work/data flows, as well as process and services analytics.
  - Differential privacy in IA, shedding light in ML blackbox by means of explanation techniques