LAMSADE CEREMADE

DATA SCIENCE: NEW DATA, NEW PARADIGMS

From data to classes and classes as statistical units

RECHERCHE - FORMATION

UNIVERSITE PARIS-DAUPHINE

22-23 January 2018

Registration now open

Free but mandatory (number of places is limited)

Participation gratuite à condition d'être inscrit (nombre de places limité)

Registration at: datascience22230118@gmail.com

Venue, Lieu des journées : Université Paris-Dauphine

Website : http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:pa18

La numérisation croissante de notre société alimente entre autres des bases de données ouvertes (« Open Data »), de taille grandissante (Big Data). Ces données sont souvent complexes (hétérogènes et multi-tables, munies de variables non appariées) mais peuvent être la source de création de valeur considérable pour la société à condition qu'elles soient exploitées avec des méthodes d'analyse adéquates.

Ces journées ont justement pour objectif de centrer cette fois le débat vers l'analyse de ces données en pensant en termes de classes. Les classes réduisent la taille des données et constituent souvent un pivot central incontournable de l'analyse. Ces classes obtenues par apprentissage non-supervisé permettent d'obtenir une vue concise et structurée modélisant les données, en apprentissage supervisé elles permettent de fournir des règles de décision efficaces.

Une troisième voie consiste à les considérer comme de nouvelles unités statistiques et à les décrire par des données symboliques (i.e. toute forme d'expression permettant de prendre en compte la variabilité interne des classes). On décrira ainsi les classes par des vecteurs de lois jointes ou marginales, d'intervalles, d'histogrammes (issus d'ondelettes, par exemple), de diagrammes de fréquence (d'utilisation courante dans les Instituts Nationaux de Statistique pour décrire des régions), de distributions, etc. . Cela permet de fusionner les données complexes et massives (en résolvant entre autres le problème des

variables non appariées), pour pouvoir les étudier et les comprendre dans un cadre explicatif adéquat (i.e. à contrario des approches « boîte noire » de type « réseaux neuronaux »). L'analyse des données symboliques multidimensionnelles qui décrivent les classes peut aussi considérablement enrichir les interprétations classiques unidimensionnelles de ces classes.

L'objectif de ces Journées est de laisser la parole à des spécialistes de l'extraction de connaissances à partir de données de toutes sortes et de réfléchir ensemble aux orientations et tendances de la théorie et de la pratique de l'analyse de ces nouvelles données dans le contexte de la révolution numérique.

A "Data Scientist" is someone who is able to extract new knowledge from Standard, Big and Complex Data: unstructured data, unpaired samples, multi sources data (as mixture of numerical, textual, image, social networks data). The fusion of such data can be done into classes of row statistical units which are considered as new statistical units. Classes can be obtained by unsupervised learning giving a concise and structured view on the data or by supervised learning in order to produce efficient rules (as by deep learning). A third way is to consider classes as new statistical units described by vectors of intervals, probability distributions, weighted sequences, functions, and the like, in order to express the within-class variability. One of the advantages of this approach is that unstructured data and unpaired samples at the level of row units, become structured and paired at the classes' level.

The objective of this Workshop is to let speak the specialists of knowledge extraction from all sorts of data, and to think together about the orientations and trends of the theory and the practice of the analysis of these new data, in the context of the digital revolution.

THEMES

- **Theoretical foundation of classes and Symbolic Data**
- **Linear models for symbolic data**
- **Clustering for symbolic data**
- **Symbolic networks**
- **Dimensionality reduction**
- **Applications in socio-demography and ecology**

International Scientific Committee

L. Billard (UGA, USA)
P. Cazes (CEREMADE, University Paris-Dauphine)
D. Colazzo (LAMSADE, University Paris-Dauphine)
S. Pinson (LAMSADE, University Paris-Dauphine)
M. Ichino (College of Science and Engineering, Tokyo Denki University, Japan)
M. Noirhomme (Namur University, Belgium)
S. Sisson (UNSW Sydney, Australia)
H. Wang (School of Economics and Management, Beihang University, China)

Local Organizing Committee

P. Bertrand (CEREMADE, University Paris-Dauphine)
E. Diday (CEREMADE, University Paris-Dauphine)
W. Litwin (LAMSADE, University Paris-Dauphine)

Lecturers in the order of the program

G. Saporta (CNAM, Conservatoire National des Arts et Métiers, France)

S. Sisson (UNSW Sydney, Australia)

R. Emilion (MAPMO, Université d'Orléans, France)

E. Diday (CEREMADE, Université Paris-Dauphine)

B. Beranger (UNSW Sydney, Australia)

Z. Wang (SEM, Beihang University, China)

F. De Carvalho (CIn-UFPE, Recife, Brazil)

T. Huang (SEM, Beihang University, China)

M. Nadif (MI, Université Paris-Descartes, France)

O. Rodriguez (Costa-Rica University, Costa Rica)

A. Irpino (MP, University of Campania L. Vanvitelli, Caserta, Italy).

R. Verde (Naples University, Italy)

V. Batagelj (FMF, Ljubljana, Slovenia)

M. Malek (EISTI, Cergy-Pontoise, France)

V. Cariou (StatSC, Oniris, INRA, 44322, Nantes, France)

P. Brito (University of Porto, Portugal)

Y. Lechevallier (Directeur de recherche honoraire, INRIA, France)

W. Litwin (LAMSADE, Université Paris Dauphine, France)

C. Biernacki (Lille University, INRIA, France)

F. Lebaron (ENS, Paris-Saclay Cachan, France)

C. Toque (DGALN/SAGP/SDP/BCSI (Ministère de la transition écologique et solidaire. Ministère de la cohésion des territoires, France)

==========

PROGRAM

Monday, January 22nd

9h00 – 9h15 WELCOME ADDRESS

First Session

THEORETICAL FOUNDATION OF CLASSES AND SYMBOLIC DATA

9h15 – 9h45 G. Saporta (Conservatoire National des Arts et Métiers, France)

Paul Lazarsfeld and latent classes: some history

9h45 – 10h15 S. Sisson, B. Beranger, J. Lin, T. Whitaker, X. Zhang (UNSW Sydney, Australia)

A general framework for constructing symbolic likelihood functions and examples

++

10h15 – 10h45 *Coffee break*

++

10h45 – 11h15 R. Emilion (Université d'Orléans, France), E. Diday (Université Paris-Dauphine, France)

Likelihood on symbols: probabilistic setting and examples

11h15 – 11h45 E. Diday (CEREMADE, Université Paris-Dauphine, France)

Basic theory for ranking classes and their descriptive symbolic variables

11h45 – 12h15 B. Beranger, T. Whitaker, S. Sisson (UNSW Sydney, Australia)

Extreme value analysis using symbolic data

++

12h15 – 14h00 *LUNCH*

++

Second session

LINEAR MODELS

14h00 – 14h30 Z. Wang, T. Huang (School of Economics and Management, Beihang University, China)

Linear Mixed Effects Models for Longitudinal Compositional Data

14h30 – 15h00 F. De Carvalho (CIn-UFPE, Brazil)

A kernel robust regression model for interval-valued variables

15h00 – 15h30 H. Wang, T. Huang, S. Wang (School of Economics and Management, Beihang University, China)

Spatial functional Linear Model and Estimation Method

++

15h30 – 16h00 *Coffee break*

++

Third session

DIMENSIONALITY REDUCTION

16h00 – 16h30 Labiod, M. Nadif (Université Paris-Descartes, France)

Simultaneous learning for clustering and dimensionality reduction

16h30 – 17h00 O. Rodriguez (Costa-Rica University)

Optimized dimensionality reduction methods for symbolic interval variables

17h00 – 17h30 A. Irpino, (Dept. of Mathematics and Physics, University of Campania L. Vanvitelli, Caserta, Italy), J. Arroyo Gallardo (Universidad Complutense de Madrid, Spain)

Dimension reduction technique for histogram variables: an application on a Human Activity Recognition dataset

==========

Tuesday, January 23rd

First Session

CLUSTERING

9h00 – 9h30 R. Verde, A. Irpino (Naples University, Italy)

Imprecise Distributional data: Visualization and clustering

9h30 – 10h00 F. De Carvalho (CIn-UFPE, Brazil)

Gaussian Kernel C-Means Clustering Algorithms with Automated Computation of Bandwidth Parameters

++

10h00 – 10h30 *Coffee break*

++

Second Session

NETWORKS

10h30 – 11h00 V. Batagelj (FMF, Ljubljana, Slovenia)

Symbolic networks

11h00 – 11h30 M. Malek (EISTI, Cergy-Pontoise, France)

Analysis of complex networks awarded multi-layer

Third Session

MULTIBLOCK and TIME SERIES

11h30 – 12h00 V. Cariou (StatSC, Oniris, INRA, 44322, Nantes, France)

Supervised multiblock modelling with P-ComDim. Applications in sensometrics and chemometric.

12h00 – 12h30 A. Maharaj (Monash Univ., Melbourne, Australia), P. Teles (University of Porto, Portugal), P. Brito (University of Porto, Portugal)

Clustering of Interval Time Series

12h30 – 13h Y. Lechevallier (Directeur de recherche honoraire, INRIA, France)

Weighted Multi-view Partitioning of Time Series

++

12h30 – 14h00 *LUNCH*

++

Fourth Session

SOFTWARE AND BIG DATA MANAGEMENT

14h00 – 14h30 O. Rodriguez (Costa-Rica University)

New methods and applications with the R package for Symbolic Data Analysis - RSDA

14h30 – 15h00 W. Litwin, (LAMSADE, Paris-Dauphine, France), S. Jojoda (G. Mason Univ, USA), Th. Schwarz (Marquette Univ., USA)

Trusted cloud SQL DBS with On-the-fly AES Decrypion/ Encrypion for Big SQL Data Bases

15h00 – 15h30 C. Biernacki (Lille University, INRIA, France)

MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data

++

15h30 – 16h00 *Coffee break*

++

Fifth Session

APPLICATIONS

16h00 – 16h30 F. Lebaron (ENS, Paris-Saclay Cachan, France)

Classes of living conditions and social classes in Europe

16h30 – 17h00 C. Toque (Ministère de la transition écologique et solidaire. Ministère de la cohésion des territoires, France)

Segmentation territoriale des attributions de logements sociaux par l'Analyse des Données Symboliques

17h00 – 18h00 SUMMARY - OPEN DISCUSSION