Inria
INVENTEURS DU MONDE NUMÉRIQUE

BIG DATA

Inria Saclay - Île-de-France

Des données
aux connaissances
et à la création de valeur

Inria Saclay Île-de-France

L'INTERDISCIPLINARITÉ AU COEUR DU PLATEAU DE SACLAY

Le centre Inria Saclay - Île-de-France s'inscrit dans un contexte exceptionnel de recherche et de développement de technologies au coeur de l'Université Paris-Saclay dont il est membre fondateur.

3 priorités scientifiques :

- SÛRETÉ, SÉCURITÉ ET FIABILITÉ POUR LES ARCHITECTURES, LES LOGICIELS ET LES DONNÉES.
- MODÉLISATION, ANALYSE ET VISUALISATION DE DONNÉES MASSIVES DISTRIBUÉES, ET EXTRACTION DE CONNAISSANCES.
- MODÉLISATION, SIMULATION, ET OPTIMISATION DE SYSTÈMES DYNAMIQUES COMPLEXES.

Le centre travaille entre autres sur le développement des sciences de la donnée, «Big Data Analytics», sur différents domaines à fort impact sociétal comme la santé ou la sécurité.

Bertifier

À quoi sert Bertifier ?

Bertifier est une application web pour la création rapide de visualisations à partir de tableaux/feuilles de calcul.

Le système s'inspire de la méthode d'analyse matricielle de Jacques Bertin, dont le but était de «simplifier sans détruire» en encodant visuellement les valeurs de cellules et en regroupant les lignes et colonnes similaires.

Bertifier rend accessible la méthode de Bertin à un public large, qu'il possède ou non des compétences techniques ; l'outil offre des possibilités d'analyse de données et de communication qui n'étaient jusqu'alors accessibles qu'à une poignée de spécialistes.

Domaines d'application

Visualisation
Interaction
Données tabulaires

Contact

Jean-Daniel Fekete
Pierre Dagicevic
Charles Perrin
Équipe Aviz - Inria Saclay - Île-de-France

Sparklificator

À quoi sert Sparklificator ?

Sparklificator est une bibliothèque jQuery open-source qui facilite l'utilisation de visualisations-mots dans des documents HTML.

Les visualisation-mots (word-scale visualizations) sont des représentations visuelles miniatures, intégrées dans du texte. Elles permettent de mélanger des explications textuelles avec des données quantitatives. Sparklificator offre un choix d'options pour ajuster la position des visualisations-mots, ainsi que leur taille et leur espacement dans le texte.

La bibliothèque offre des encodages visuels prédéfinis tels que les line charts et les histogrammes, et permet également de développer des visualisation-mots sur mesure.

Domaines d'application

Visualisation de texte
Visualisation de documents
Sparklines
Navigation
Visualisation multi-échelle

Contact

Jean-Daniel Fekete
Pascal Goffin
Wesley Willett
Petra Isenberg
Équipe Aviz - Inria Saclay - Île-de-France

Bocop

Optimisation des trajectoires d'avions civils et économies de carburant

À quoi sert BOCOP ?

Les outils de l'équipe Commands pour l'identification et l'optimisation sont intégrés dans l'offre logicielle Opti-Climb développée par Safety Line, qui fournit au pilote des consignes optimisées pour chaque vol.

L'optimisation de trajectoire peut être locale (approche directe) ou globale (programmation dynamique) et exploite des modèles d'avions (aérodynamique poussée) identifiés sur une base de données enregistrées par les boîtes noires.

Identification et optimisation sont réalisées avec Bocop (www.bocop.org), une boîte à outils open-source développée par l'équipe Commands depuis 2010.

Domaines d'application

Optimisation
Identification
Optimisation dynamique
Gestion de l'énergie

Contact

Pierre Martinon
Équipe Commands - Inria Saclay - Île-de-France

TDA

Méthodes d'analyse des données topologiques

À quoi sert TDA ?

L'objectif de l'analyse topologique des données est de développer un outil accessible, efficace et largement utilisé pour l'analyse du Big Data. TDA fournit des informations topologiques et géométriques sur les données qui ne sont pas accessibles par d'autres méthodes classiques.
Pour les données issues du monde réel, et des contraintes associées, TDA a mis l'accent sur le développement d'approches statistiques pour en déduire des informations topologiques sans tenir compte de l'ensemble des données. Cette approche conduit à la conception d'algorithmes très rapides et permet de combiner les outils TDA aux technologies de l'intelligence artificielle, comme Watson, pour traiter le Big Data.

Une partie des méthodes de TDA développées sont incorporées à la bibliothèque open source Gudhi (https://project.inria.fr/gudhi/software) développée par Geometrica.

Domaines d'application

Data mining
Data Analysis
Machine Learning

Contact

Frédéric Chazal
Marc Glisse
Équipe Geometrica - Inria Saclay - Île-de-France

ToMATo

Topological Mode Analysis Tool

À quoi sert ToMATo ?

ToMATo est un nouveau logiciel pour la classification non supervisée de nuages de points générés par des simulations ou des mesures de processus physiques. Le concept est fondé sur des bases théoriques solides et offre une grande flexibilité.

Sous la forme d'un diagramme en deux dimensions, appelé « diagramme de persistance », le logiciel présente la structure des données à l'utilisateur. Ce diagramme peut ensuite être utilisé pour déterminer le nombre de clusters et pour distinguer le signal du bruit.

ToMATo fournit en sortie, au choix, du hard ou du soft clustering, et passe à l'échelle (proportionnellement avec la taille et la dimension des données).

Une partie des méthodes de ToMATo développées sont incorporées à la bibliothèque open source Gudhi (https://project.inria.fr/gudhi/software) développée par Geometrica.

Domaines d'application

Classification
Clustering

Contact

Steve Oudot
Équipe Geometrica - Inria Saclay - Île-de-France

WaRG

Analyse des graphes RDF

À quoi sert WaRG ?

WaRG (Warehousing RDF Graphs) est une plateforme analytique spécialement conçue pour l'analyse de graphes de données RDF.

WaRG permet de définir des schémas d'analyse des données comportant des classes et des propriétés d'intérêt pour l'analyste.
Ensuite, le schéma d'analyse est matérialisé, ce qui conduit à une instance (graphe RDF) raffinée pour l'analyse.
Le schéma d'analyse peut aussi être construit automatiquement à partir de l'instance RDF en entrée.
Enfin, des requêtes analytiques sont spécifiées et conduisent à des cubes d'analyse des données RDF.

Domaines d'application

Web sémantique
Decision support
Linked data

Contact

Ioana Manolescu
Alexandra Roatis
Sejla Cebirič
Équipe Oak - Inria Saclay - Île-de-France / LRI

CliqueSquare

Plateforme de gestion de données RDF basée sur une architecture Hadoop

À quoi sert CliqueSquare ?

CliqueSquare permet de gérer de très grands volumes de données RDF de façon parallèle en utilisant un cluster Hadoop.
Le système utilise son propre modèle de partitionnement et stockage de triples RDF dans le cluster.
Il est capable de traiter des requêtes RDF exprimées dans un sous-ensemble de SPARQL.
Il est particulièrement efficace dans le traitement de requêtes complexes, car il les traduit vers des programmes MapReduce garantis d'avoir le nombre d'étapes le plus petit possible.

Domaines d'application

Hadoop
MapReduce
Linked Data
Web sémantique

Contact

Ioana Manolescu
Benjamin Djahandideh
Équipe Oak - Inria Saclay - Île-de-France / LRI

Mixmod

Logiciel multi-usages d'exploration de données et d'apprentissage statistique

À quoi sert Mixmod ?

Mixmod est une boîte à outils d'apprentissage statistique, conçue pour traiter de grands ensembles de données.
Mixmod offre des algorithmes d'estimation éprouvés et des critères de sélection de modèles efficaces et a été utilisé avec succès dans les domaines du marketing, du crédit scoring, de l'épidémiologie, la génomique et la fiabilité notamment.

Grâce au modèle probabiliste des mélanges de lois de probabilité, il offre une panoplie très riche de méthodes de classification.

Mixmod est doté d'indices simples et rigoureux pour évaluer la qualité des résultats. Il propose une interface graphique conviviale (mixmodGUI) et des fonctions pour les environnements R (Rmixmod) et Matlab (mixmodForMatlab).

Domaines d'application

Marketing
Crédit scoring
Épidémiologie
Génomique

Contact

Gilles Celeux
Nomi Ngabe
Benjamin Auder
Équipe Select - Inria Saclay - Île-de-France / LMO
Équipe Modal - Inria Lille / Université Lille 1et 2

Scikit-learn

À quoi sert Scikit-learn ?

Scikit-learn peut être utilisé comme un middleware pour des tâches de prédiction. Par exemple, un grand nombre de start-ups du web s'approprient Scikit-learn pour prédire des comportements d'achat d'utilisateurs, proposer des recommandations de produits ou détecter les tendances ainsi que les comportements abusifs (fraudes, spams, etc.).
Scikit-learn sert à extraire la structure de données complexes (textes, images) et à les classifier en utilisant des techniques correspondant à l'état de l'art.

Facile à utiliser, efficace et accessible aux non-experts du data science, Scikit-learn est une bibliothèque d'apprentissage statistique. Dans une étape d'exploration des données, l'utilisateur entre quelques lignes dans une interface interactive (mais non graphique) et peut analyser les résultats de sa requête immédiatement.

Scikit-learn est un moteur de prédiction, développé en open source et disponible sous licence BSD.

Domaines d'application

Prévision des comportements des utilisateurs
E-commerce
Lutte anti-spam
Détection de la fraude
E-mailing de ciblage
Amélioration des produits

Contact

Bertrand Thirion
Gaël Varoquaux
Olivier Grisel
Équipe Parietal - Inria Saclay - Île-de-France

STOIC

À quoi sert STOIC ?

Les stratégies de marketing actuelles reposent en grande partie sur l'analyse des médias en ligne et les réseaux sociaux. Par exemple, l'identification des leaders d'opinion donne un avantage concurrentiel dans la vente et la promotion des produits.

STOIC permet d'identifier les leaders d'opinion en ligne à partir de données telles que les messages de blog ou leur profil twitter.

Les ingrédients clés de STOIC sont l'apprentissage automatique des classements et des connaissances du terrain.

Domaines d'application

Classement
Réseaux sociaux
Média en ligne

Contact

Philippe Caillou
Équipe Tao - Inria Saclay - Île-de-France / LRI

ZVTM

Boîte à outils pour interfaces zoomables et interaction avec de grandes quantités de données

À quoi sert ZVTM ?

ZVTM est une boîte à outils facilitant le développement d'interfaces multi-échelles permettant de naviguer dans de grands jeux de données visualisées en 2D.

ZVTM est utilisé pour explorer de grandes bases de données dans différents domaines : systèmes d'information géographique, salles de contrôles de grands équipements, astronomie, systèmes de distribution d'énergie.

La boîte à outils facilite aussi le développement d'applications pour les murs d'images ultra-haute résolution.

Domaines d'application

Visualisation multi-échelle
Murs d'écrans pilotés par des grappes de machines
Graphiques structurés

Contact

Emmanuel Pietriga
Équipe Ilda - Inria Saclay - Île-de-France / LRI

WILDER

Ecrans géants pour défis scientifiques immenses

À quoi sert WILDER ?

Permettre aux scientifiques de naviguer dans des images immenses et des données complexes avec une simplicité extraordinaire : voilà l'objectif du projet WILDER.

Ce mur, composé de 75 écrans, pour un total de six mètres par deux et une résolution de 14400x4800 pixels, succède au projet WILD.

En plus d'être plus grand que son prédécesseur, WILDER dispose d'écrans dont les bords ne sont pratiquement plus visibles.

WILDER permet l'interaction tactile multipoints, mais aussi la possibilité d'interagir à distance. Destiné entre autres à l'aide à la découverte scientifique, le dispositif permet par ailleurs le travail collaboratif.

CENTRE DE RECHERCHE COMMUN
Inria - Microsoft Research

Un centre de recherche de rang mondial

C'est au centre Inria Saclay - Île-de-France, au coeur de l'Université Paris-Saclay, que se trouve le Centre de recherche commun.

Inauguré en janvier 2007 par Inria et Microsoft, il accueille des chercheurs du monde entier qui y mènent conjointement des recherches à long terme dans **quatre grands domaines** :

- Machine Learning et Big data
- Vision numérique et imagerie médicale
- Méthodes formelles pour les mathématiques, systèmes distribués et sécurité
- Réseaux sociaux d'information et protection de la vie privée

Les résultats des travaux du Centre de recherche commun Inria-Microsoft Research sont publics et mis à la disposition de la communauté scientifique nationale et internationale.

Il reçoit la contribution de plus de 50 chercheurs des deux partenaires et accueille en permanence 30 doctorants et post doctorants. Plus de 70 jeunes chercheurs ont été formés depuis 2007 et 25 thèses ont été soutenues.

Le Centre de Recherche Commun Inria-Microsoft Research accueille en son sein Leslie Lamport, Prix Turing 2013.

Contacts

Service Communication

Stéphanie Dupont - stephanie.dupont@inria.fr
Charlotte Renauld - charlotte.renauld@inria.fr
Emmanuelle Perrot - emmanuelle.perrot@inria.fr

Service Transfert, Innovation et Partenariats

Maike Gilliot - maike.gilliot@inria.fr
Oana Manea - oana.manea@inria.fr
Annie Floch - annie.floch@inria.fr

Inria INVENTEURS DU MONDE NUMÉRIQUE
université membre de l'UNIVERSITÉ PARIS-SACLAY

BIG DATA

Inria Saclay - Île-de-France

Unlocking knowledge
and value from data

Inria Saclay Île-de-France

INTERDISCIPLINARITY AT THE HEART OF THE PLATEAU DE SACLAY

Inria Saclay - Île-de- France research center is incorporated within an exceptionnal environment of research and development of technologies in the heart of the Université Paris-Saclay as a founding member.

3 scientific priorities:

- Software and data security and reliability
- High Performance Computing and distributed knowledge
- Modelling, simulation and optimisation of dynamical complex systems

The research center works on the development of data sciences, «Big Data Analytics», on different fields with strong societal impact such as health or security.

Bertifier

What is Bertifier ?

Bertifier is a web-based application for rapidly creating visualizations from spreadsheets.
It directly draws from Jacques Bertin's matrix analysis method, whose goal was to « simplify without destroying » by encoding cell values visually and grouping similar rows and columns.

Bertifier has the potential to bring Bertin's method to a wide audience of both technical and non-technical users, and empowers them with data analysis and communication tools that were so far only accessible to a handful of specialists.

Applications domains

Visualization
Interaction
Tabular Data

Contact

Jean-Daniel Fekete
Pierre Dagicevic
Charles Perrin
Aviz project-team - Inria Saclay - Île-de-France

Sparklificator

What is Sparklificator ?

Sparklificator is a general open-source jQuery library that eases the process of integrating word-scale visualizations into HTML documents.

The word-scale visualizations are miniature visual representations within the text.
They enable the mixing of textual explanations and quantitative data.

Sparklificator provides a range of options for adjusting the position of word-scale visualizations, their size, and spacing within the text. The library includes default visualizations, including line charts and bar charts, and can also be used to integrate custom word-scale visualizations.

Applications domains

Text visualization
Document visualization
Sparklines
Word-scale visualization
Navigation

Contact

Jean-Daniel Fekete
Pascal Goffin
Wesley Willett
Petra Isenberg
Aviz project-team - Inria Saclay - Île-de-France

Bocop

Optimize civil airplane trajectories to improve fuel consumption

What is Bocop?

The Commands team's identification and optimization tools are integrated in the Opti-Climb software, developed by Safety Line, which provides pilots with optimized flight plans. This trajectory optimization can be performed by local (direct approach) or global (dynamic programming) methods, and uses plane models (aerodynamics, thrust) identified from the data recorded during flights.

Identification and optimization are done with Bocop (www.bocop.org), an open source toolbox developed by the Commands team since 2010.

Applications domains

Energy management
Numerical optimization
Biology
Identification
Dynamic Optimization
Transportation

Contact

Pierre Martinon
Commands project-team - Inria Saclay - Île-de-France

TDA

Topological Data Analysis methods

What is TDA ?

TDA provides a set of tools and methods to infer topological and geometric information about the structure of possibly big and complex data, that are not reachable through other classical methods.

Recent TDA developments are motivated by realworld and big data constraints: they focus on the development of statistical approaches to infer relevant topological information without considering the whole data, even when data are corrupted by noise and outliers. This leads to the design of very fast and easily parallelizable algorithms for TDA, and opens the door to the combination of TDA tools with modern learning and artificial intelligence technologies.

A part of TDA's methods developped are integrated in the open source library Gudhi (https://project.inria.fr/gudhi/software) developped by Geometrica.

Applications domains

Data mining
Data Analysis
Machine Learning

Contact

Frédéric Chazal
Marc Glisse
Geometrica project-team - Inria Saclay - Île-de-France

ToMATo

Topological Mode Analysis Tool

What is ToMATo ?

ToMATo is a novel scheme for classification and clustering of point-cloud data generated by simulations or measurements of physical processes.

It is highly flexible and has sound theoretical foundations. It provides feedback on the structure of the data in the form of a 2-dimensional diagram called a «persistence diagram». Such feedback can be used for determining the number of clusters, and for distinguishing between the signal and the noise.

ToMATo is able to perform both hard and soft clustering, and scales up with the size and dimensionality of the data.

A part of ToMATo's methods developped are integrated in the open source library Gudhi (https://project.inria.fr/gudhi/software) developped by Geometrica.

Applications domains

Classification

Clustering

Contact

Steve Oudot
Geometrica project-team - Inria Saclay - Île-de-France

WaRG

Warehouse-style analytics platform on RDF graphs

What is WaRG ?

WaRG (Warehousing RDF graph) is an analytical platform specially designed for the analysis of RDF data.

WaRG allows defining RDF analytical schemas, comprising classes and properties interesting for the analysis. The analytical schema can then be materialized, leading to an instance (RDF graph) refined for the needs of the analysis.
The analytical schema can also be automatically built from the input RDF instance.
Finally, RDF analytical queries can be specified and lead to RDF analysis cubes.

Applications domains

Semantic Web
Decision support
Linked data

Contact

Ioana Manolescu
Alexandra Roatis
Sejla Cebirič
Oak project-team - Inria Saclay - Île-de-France / LRI

CliqueSquare

RDF data management platform based on Hadoop architecture

What is CliqueSquare ?

RDF (Ressource Description Framework) is the data format for the semantic web. CliqueSquare allows storing and querying very large volumes of RDF data in a massively parralel fashion in a Hadoop cluster.
The system uses its own partitioning and storage model for the RDF triples in the cluster.

CliqueSquare evaluates queries expressed in a dialect of the SPARQL query language.
It is particularly efficient when processing complex queries, because it is capable of translating them into MapReduce programs guaranteed to have the minimum number of successive jobs. Given the high overhead of a MapReduce job, this advantage is considerable.

Applications domains

Hadoop
MapReduce
Linked Data
Semantic web

Contact

Ioana Manolescu
Benjamin Djahandideh
Oak project-team - Inria Saclay - Île-de-France / LRI

Mixmod

Many-purpose software for data mining and statistical learning

What is Mixmod ?

Mixmod is a free toolbox for data mining and statistical learning designed for large and high-dimensional data sets. Mixmod provides reliable estimation algorithms and relevant model selection criteria.

It has been successfully applied to marketing, credit scoring, epidemiology, genomics and reliability among other domains.

Its particularity is to propose a model-based approach leading to a lot of methods for classification and clustering.

Mixmod allows to assess the stability of the results with simple and thorough scores. It provides an easy-to-use graphical user interface (mixmodGUI) and functions for the R (Rmixmod) and Matlab (mixmodForMatlab) environments.

Applications domains

Marketing

Credit scoring
Epidemiology
Genomics

Contact

Gilles Celeux - Nomi Ngabe - Benjamin Auder
Select project-team - Inria Saclay - Île-de-France / LMO
Modal project-team - Inria Lille /
Université Lille 1et 2

Scikit-learn

What is Scikit-learn ?

Scikit-learn can be used as a middleware for prediction tasks. For example, many web startups adapt Scikitlearn to predict buying behavior of users, provide product recommendations, detect trends or abusive behavior (fraud, spam).

Scikit-learn is used to extract the structure of complex data (text, images) and classify such data with techniques relevant to the state of the art.

Easy to use, efficient and accessible to non datascience experts, Scikit-learn is an increasingly popular machine learning library in Python. In a data exploration step, the user can enter a few lines on an interactive (but non-graphical) interface and immediately sees the results of his request. Scikit-learn is a prediction engine.

Scikit-learn is developed in open source, and available under the BSD license.

Applications domains

E-commerce
Spam-fighting
Fraud detection
Email targeting
User-behavior prediction
Product improvement

Contact

Bertrand Thirion
Gaël Varoquaux
Olivier Grisel
Parietal project-team - Inria Saclay - Île-de-France

STOIC

What is STOIC ?

Current marketing strategies rely heavily on the analysis of online media and social networks. For example, identifying the opinion leaders gives a competitive advantage in selling and promoting products.

STOIC allows one to identify online opinion leaders from data such as their blog posts or their twitter profiles.

The key ingredients of STOIC are learning-to-rank techniques and the ground truth.

Domaines d'application

Ranking
Social networks
Online media

Contact

Philippe Caillou
Inria Tao project-team - Inria Saclay - Île-de-France / LRI

ZVTM

Zoomable User Interface Toolkit for Interacting with Large Datasets

What is ZVTM ?

ZVTM is a toolkit enabling the implementation of multi-scale interfaces for interactively navigating in large datasets displayed as 2D graphics.

ZVTM is used for browsing large databases in multiple domains: geographical information systems, control rooms of complex facilitites, astronomy, power distribution systems.

The toolkit also enables the development of applications running on ultra-high-resolution wallsized displays.

Applications domains

Multi-scale visualization
Cluster-driven wall-sized displays
Structured graphics

Contact

Emmanuel Pietriga
Ilda project-team - Inria Saclay - Île-de-France / LRI

WILDER

Giant screen for huge scientific challenges

What is WILDER ?

The aim of the WILDER project is to allow scientists to browse through huge images with unparalleled ease.

This wall, which consists of 75 screens, totalling six metres by two and a resolution of 14400x4800 pixels, ollows on from the WILD project launched in 2009. As well as being bigger and more accurate than its predecessor, WILDER has screens whose edges are almost invisible.

WILDER integrates multipoint touch screens, as well as the possibility of remote interaction.
The device is intended for assisting with scientific discoveries and is adapted for group work.

CENTRE DE RECHERCHE COMMUN
Inria - Microsoft Research

The Microsoft Research-Inria Joint Centre was founded by Inria, Microsoft Corporation, and the Microsoft Research Laboratory Cambridge. The Centre's objective is to pursue fundamental, long-term research in Computer Science with a particular emphasis on formal methods and machine learning and some of their key applications.

The research programme at the Joint Centre is structured in **four distinct areas:**

- Machine Learning and Big data
- Computer vision and medical imaging
- Formal methods and their applications for mathematical components, tools of proofs and secure computing
- Social networks and privacy

The Microsoft Research-Inria Joint Centre's works' results are public and made available for the national and international scientific community.

It gets input from more that 50 researchers from both partners and welcomes permanently 30 PhD. More than 70 young researchers have been trained since 2007 and 25 thesis have been presented. The Microsoft Research-Inria Joint Centre includes in its teams Leslie Lamport, Turing 2013 Award.

Contacts

Communication department

Stéphanie Dupont - stephanie.dupont@inria.fr
Charlotte Renauld - charlotte.renauld@inria.fr
Emmanuelle Perrot - emmanuelle.perrot@inria.fr

Technology Transfer and Partnership department

Maike Gilliot - maike.gilliot@inria.fr
Oana Manea - oana.manea@inria.fr
Annie Floch - annie.floch@inria.fr

Inria INVENTEURS DU MONDE NUMÉRIQUE membre de l'université PARIS-SACLAY