COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Jakez ROLLAND

Datascape : un framework numérique pour l'abstraction et l'exploration de données hétérogènes et multidimensionnelles

Thèse présentée et soutenue à Nantes, le 27 Novembre 2024
Unité de recherche : UMR6004 LS2N

Rapporteurs avant soutenance :

Alejandro MAAS Professeur des universités, Universidad de Chile
Xavier LORCA Professeur des universités, IMT Mines d'Albi

Composition du Jury :

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Alejandro MAASS	Professeur des universités, Universidad de Chile
	Claire CURY	Chercheure , INRIA Rennes
	Xavier LORCA	Professeur des universités, IMT Mines d'Albi
Dir. de thèse :	Christian ATTIOGBE	Professeur des universités, Nantes Université
Co-dir. de thèse :	Benoit DELAHAYE	Maître de conférences , Nantes Université
Co-dir. de thèse :	Damien EVEILLARD	Professeur des universités, Nantes Université

Invité(s) :

Ronan BOUTIN Ingénieur et Dirigeant de la société Bio Logbook, Bio Logbook

Titre : Datascape : un framework numérique pour l'abstraction et l'exploration de données hétérogènes et multidimensionnelles

Mot clés : topologie, données hétérogènes, théorie des graphes, enveloppes convexes

Résumé : La science des données est un domaine puissant pour extraire de l'information, comparer et prédire des comportements à partir de jeux de données. Cependant, la diversité des méthodes et des hypothèses nécessaires pour abstraire un jeu de données révèle un manque de généricité. De plus, la forme d'un jeu de données, qui structure les informations et incertitudes qu'il contient, est rarement prise en compte. Inspirés par les algorithmes de manifold learning et d'estimation d'enveloppes convexes, nous proposons un nouveau framework, le datascape, qui exploite la topologie et la théorie des graphes pour abstraire des jeux de données hétérogènes. Construit à partir de la combinaison d'un graphe de plus proches voisins, d'un ensemble d'enveloppes convexes, et d'une fonction de distance qui respecte la forme des données, le datascape permet d'explorer l'espace sous-jacent d'un jeu de données. Nous montrons que le datascape peut révéler la structure sous-jacente de jeux de données simulés, construire des algorithmes prédictifs dont les performances sont proches de celles de l'état de l'art, et révéler des géodésiques pertinentes entre des points. Le datascape démontre sa polyvalence à travers des applications dans le domaine médical, en écologie et sur des données simulées.

Title: Datascape: a numerical framework for abstracting and exploring heterogeneous multidimensional datasets

Keywords: topology, heterogeneous data, graph theory, convex hulls

Abstract: Data science is a powerful field for gaining insights, comparing, and predicting behaviors from datasets. However, the diversity of methods and hypotheses needed to abstract a dataset exhibits a lack of genericity. Moreover, the shape of a dataset, which structures its contained information and uncertainties, is rarely considered. Inspired by state-of-the-art manifold learning and hull estimations algorithms, we propose a novel framework, the datascape, that leverages topology and graph theory to abstract heterogeneous datasets. Built upon the combination of a nearest neighbor graph, a set of convex hulls, and a metric distance that respects the shape of the data, the datascape allows exploration of the dataset's underlying space. We show that the datascape can uncover underlying functions from simulated datasets, build predictive algorithms with performance close to state-of-the-art algorithms, and reveal insightful geodesic paths between points. It demonstrates versatility through ecological, medical, and simulated data use cases.