1SC4792 – Black swans detection in particle physics and cosmology (Cosmology)

Instructors: Jean-Christophe Hamilton
Department: DOMINANTE - PHYSIQUE ET NANOTECHNOLOGIES
Language of instruction: ANGLAIS
Campus: CAMPUS DE PARIS - SACLAY
Workload (HEE): 40
On-site hours (HPE): 27,00

Description

This Enseignement d'Intégration / Challenge Week is a week long project where students are grouped in collaborations of about 25, split into 6 teams corresponding to various tasks.
The goal is to come up with cosmological constraints from a simulated dataset (Type Ia supernovae and Cosmic Microwave Background) corresponding to a Universe with distinct (and unknown) cosmological parameters.
Each collaboration has to build an analysis pipeline to arrive to the relevant cosmological constraints. Tasks to achieve include:

- Supernovae detection from images
- Supernovae photometry and light curve building
- Cosmological constraints from Supernovae
- CMB time ordered data filtering and mapmaking
- CMB Angular Power spectrum extraction
- Cosmological constraints from CMB Angular Power Spectrum
- Joint cosmological constraints from CMB and Supernovae data

An elected spokesperson is making sure the collaboration as a whole deliver the expected outcome.

Quarter number

ST4

Prerequisites (in terms of CS courses)

Basic knowledge in Cosmology, Statistical Analysis, coding in python.

Syllabus

We provide an introduction to the work to be done on the morning of the first day.

Students are divided into two collaborations and work independently. They are participating in team works and should share workload and tasks, identify difficulties and brainstorm together. Adjustments on work load and sharing has to be managed within the week to ensure a global collaborative effort.
Regular updates on progress and difficulties encountered have to be provided in order to improve the students' progress. Discussions with teachers happen all along the week. Small daily reports are to be provided.
Students work on shared notebooks. We ask for a presentation of the work done and the results obtained by the end of the EI week.
This week ends with a half day of oral presentations by the students.

Class components (lecture, labs, etc.)
We provide support materials to read and understand. Basic notebooks are also provided. The instruction framework follows practice sessions from which students can draw inspiration to get started. Their work must then go well beyond the framework seen in class and leave room for their personal understanding and initiative.

Grading
The week of the EI ends with a 45' presentation followed by 15' of questions. The presentation is split between 6 randomly selected speakers in addition to the spokesperson. In addition, the collaboration should deliver notebooks with all their computation. The final grade is a combination of : the performance of the whole collaboration, performance of the team, individual contributions.

Resources
Python Colaboratory notebooks (seen during main course).

Learning outcomes covered on the course
- Set up standard statistical tools
- Numerical data analysis
- search for available public librairies for specific tasks
- Team working
- Work sharing
- Presentation of methods and results

Description of the skills acquired at the end of the course

[C6] Be operational, responsible, and innovative in the digital world :

- C6.3 Process data ;

- C6.1 Solve a problem numerically;

- C6.2 Design software

[C7] know how to convince:

- C7.1 Structure your ideas and arguments, be synthetic (assumptions, objectives, expected results, approach, and value created);

- C7.4 On communication techniques : Master spoken, written, and body language, and master basic communication techniques

[C8] Lead a project, a team :

- C8.1Build the collective to work as a team ;

- C8.4 Work in project mode by implementing project management methods adapted to the situation

[C1] Analyze, design, and build complex systems with scientific, technological, human, and economic components :

- C1.4: Design : specify, implement and validate all or part of a complex system

- C1.3: Solve : solve a problem with a practice of approximation, simulation, and experimentation

[C2] Develop in-depth skills in an engineering field and in a family of professions

- C2.1 Deepen a field of engineering sciences or a scientific discipline

ST4 – 48 – DATA@WEB: WEB DATA INTELLIGENCE "VALUE CREATION AROUND WEB DATA".

Dominante : Info&Num (Computer and Digital)
Langue d'enseignement : French
Campus où le cours est proposé : Paris-Saclay

Engineering problem

This thematic sequence deals with the problem of exploiting WEB data, which are massive, high-speed (for example, in 60 seconds, Google must process more than 3 million requests for its search engine, LinkedIn manages the creation of 120 new accounts, etc.), unstructured (textual or multimedia data) and very heterogeneous. This data is also obviously valuable, as shown by the success of the web giants, and exploiting it raises more and more engineering challenges (like the MapReduce paradigm proposed by Google's engineers for the distributed processing of its search engine data). Another very important aspect of this data is that we are both producers and consumers of it and that it therefore has a strong social impact.

In this thematic sequence, we therefore propose to address the main concepts of web data analysis, from their collection to their interpretation, through at least 3 applications: e-reputation, e-marketing and e-commerce. In addition to the aspects of data processing and information extraction, this thematic sequence will also be an opportunity to address issues such as digital life, ethics, the right to privacy, data protection and a reflection on the web we want for tomorrow.

Adviced prerequisites

Common course 1A, SIP, Algorithms

Context and issue modules: They include a series of conferences, roundtables and innovation workshops:

The (r)evolutions of the Web planet, looking at some of the past, current, and future changes of the Web, emphasising the complexity of this artefact which makes it an object of multidisciplinary research

The Web giants, looking at the new practices, uses and technologies introduced by the web giants and how the GAFA members have revolutionised the way computing is done

Conference and round table on the theme of digital life, personal data, its protection, the GDPR, its impact on the market and society

Algorithmic recommendation and information diversity: how to analyse the impact of online algorithms?

Specific course (60 HEE): Processing and analysis of massive unstructured data - the case of web data

Brief description: How can we automatically and quickly find information relevant to a particular need from a large amount of information? This is typically what search engines such as Google or Baidu do efficiently when they answer the 4 million queries they each receive every second. The objective of this course is to describe the foundations and techniques of Information Retrieval (IR) on which these search engines rely. The course will also address current challenges in the field such as the contributions of machine and deep learning to IR or personalization and recommendation (collaborative filtering).

Integration course: Web datackathon challenge the data of the web!

The objective is to implement, on concrete data from the Web, several of the approaches discussed in the ST with the aim of a web data intelligence application (e-reputation, e-marketing, e-vigilance ...)

- Analysis and translation of the need

- Construction of the data analysis chain: from collection (enrichment) to interpretation and visualisation

- Design of the underlying technical architecture

- Evaluation, validation and hindsight on the implemented solution.

The IE will be built as a mini datackathon around several themes built with the ST partners. The themes envisaged are

- e-reputation of a brand or a product in social networks

- e-marketing

- e-commerce: developing a film recommendation platform

Challenge week n°1: e-reputation: sentiment analysis and NLP

- Associated partners: Octopeek SAS

- Location: Paris-Saclay campus

- Brief description : The control of information is strategic in our society. Its importance is measurable by the volume of information processed, by the speed of evolution of the information as well as by the time spent in the activity of information search. This problem has led to the emergence in recent years of new professions around Data Science and Big Data. Within the framework of this project, Octopeek SAS proposes to share its business

experience on the development of a data science model for customer needs on an example of "sentiment analysis" (e-reputation). It is about :

1. Understanding the customer's problem and translating the need
2. Identify the DataSets required for the model
3. Collect, clean and qualify the data
4. Build a model using the arsenal of data science algorithms and validate its effectiveness
5. Implement the project on a BigData platform provided by Octopeek

Challenge week n°2: e-marketing

- Examples of Associated partners: Doctolib, Proctor & Gamble, Rakuten

- Location: Paris-Saclay campus

- Brief description: The objective of this IR is to implement, on concrete data provided by the partner and possibly enriched by data from the Web, several of the approaches discussed in the ST with the aim of an application in the field of marketing. It is about :

1. Understanding the problem and translating the need,
2. Appropriate the extractions from the industrial databases provided,
3. Build the data analysis chain: from collection (enrichment) to interpretation and visualisation,
4. Develop and validate the technical architecture based on machine learning algorithms/methods (cross validation, sklearn library, ranking methods, clustering or classification methods, etc.).

Challenge week 3: e-commerce

- Associated partner: Theodo

- Location: Paris-Saclay campus

- Brief description: Theodo develops digital solutions with its clients that enable them to increase their market share or productivity. More precisely, their job is to help CIOs, CDOs and CTOs to design and code new digital products and services that will have a lasting impact on their business. Within the framework of this IE, the objective is to develop a web platform on which each user will be able to rate films and then receive automatic recommendations. Several technical topics are addressed:

1. A web development part in Serverless AWS in JS & Vue during which we will implement an authentication, a search bar and a REST API

2. A Machine Learning part to establish the recommendation. You will be able to build on what you have seen in the specific ST course or implement your own algorithm.