SciQLOP : Scientific Qt application for Learning from Observations of Plasmas

CDS 2015 project

project acronym	SciQLOP
principal investigator	Nicolas Aunai
contact email	nicolas.aunai@lpp.polytechnique.fr
other team members	alexis.jeandet@lpp.polytechnique.fr, erwan.Le-Pennec@polytechnique.edu
contracting partner	Polytechnique
principal team/laboratory	LPP
partner teams/laboratories	CMAP
type of support	engineering fellowship
amount of support	50 Ke
parallel submissions	none

Summary

This project aims at developing an interactive, graphical software dedicated to the smart exploration of large databases of satellite in situ measurements. We ask for 50k€, corresponding to 12 months of engineer time. Our (world wide) community now owns decades of space plasma measurements from multiple satellite missions, but, despite the use of uniform file format (CDF) and data types (time series of particle and fields properties) accross all missions, we critically lack the software that allows one to explore them in an *intuitive*, *smart* and *efficient* way. Therefore, a prerequisite to *every* studies is weeks/months of laborious, eyeballing and not reproducible, identification of signatures associated to the satellite crossing of plasma boundaries and physical processes of interest. So far, automatic detection of those boundaries and processes is very seldom and always based on naive algorithms looking for parameters crossing arbitrary thresholds, resulting in a lot of mis-identifications. This severly restrict most analysis to « case studies » (single events) thus precluding a good statistical representation of our space environment and of the key processes controling it. Statistical learning methods offer very broad perspectives for this matter and are a potential game changer in the way we tackle space data. The laboratory of plasma physics has been developing the proof of concept of a graphical software allowing the user to display and analyze space data in a very efficient way. Based on a plugin approach, it is thought to allow for very flexible and general data browsing as well as possibly deep and instrument-dependant analysis. Our first version includes a python interpreter, which brings a smooth transition between this general data exploration and detailed scientific analysis based on specific imported in-house or community packages. Public and open source, this Qt/C++/python software will need statistical learning methods at its core, to facilitate event recognition in the data and the collaborative building of wide catalogues and models. This tool will not only drastically improve our daily work by narrowing the analysis down to its most scientific part, but also provide with a powerful collaborative catalogue and model building by and for the community. One year full-time engineer will ensure a very good release version and lay solid foundations to what may become a must-have tool in our national and international community. Within this year, collaborations with CDS and challenges like RAMPs will be essential in testing the best methods to prioritize orientations of future developments.

In situ space data and analysis

Basics about our plasma environments: Our plasma environment is the result of a complex interaction between the geomagnetic field and the solar wind plasma and magnetic field. This interaction leads to the formation of several boundaries represented on Fig. 1, such as the *bow shock*, the *magnetopause* (separating the solar wind from the magnetosphere), the *geomagnetic tail*, etc. Those regions are highly dynamic and the locus of many plasma phenomena (magnetic reconnection, turbulence, Kelvin Helmholtz etc.) of universal interest since they are thought to be of critical importance for explaining remote astrophysical obvservations as well as understanding the key parameters controlling space weather. Our space environment is the only « astrophysical » system accessible to in situ measurements.

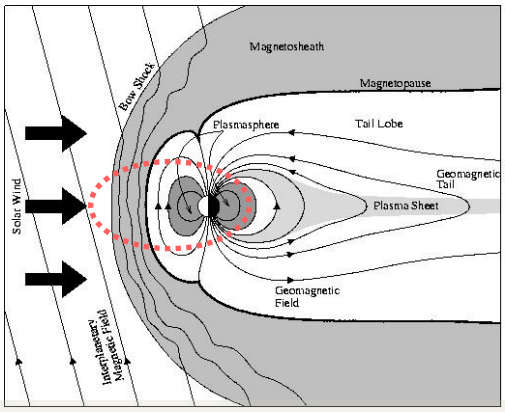

Fig. 1 - Schematics representation of the Earth magnetosphere and its interaction with the solar wind. Solid black lines with arrows represent the magnetic field. The red line is an example of a satellite orbit.

In situ measurements: Many satellite missions have been launched since the 50's, and although they all had specific scientific goals, they share a lot similarities, making it today an amazing fleet of probes in the magnetosphere. Although they don't necessarily have the same resolutions, their instruments are all mesuring the same kind of data : time series of electromagnetic fields, and times series of plasma properties. In all cases, the spacecraft motion (~km/s) is completly negligible compared to the proper oscillatory and turbulent motion of the structures of interest (tens to hundreds of km/s). Measured data is then ambiguously mixing spatial variations with unique temporal ones. This is a major difficulty of space data analysis and leads to the failure of naive automatic detections of signatures in the data. Fig. 2 illustrates an example of an outbound crossing of the magnetopause and the bow shock by the Cluster satellite, visible here via the variation of the density, the bulk flow, the magnetic field and an time-energy ion spectrogram. Oscillatory motions of the boundary caused by solar wind unsteadiness, surface waves and instabilities results in several unpredictable total and partial crossings. Therefore, although the visual identification is relatively easy, the implementation of *automatic detection based on arbitrary thresholds is failing, hence the importance of statistical learning methods*. In this example, at the magnetopause crossing, a sharp peak in the velocity (red circles) indicates the occurence of *magnetic reconnection* jets, for which, again, the variability in the expected signatures make it difficult to identify with a fix set of rules. Statistical studies of the properties of those boundaries and there embeded processes is very difficult at the present time because we're critically lacking a way to build large and accurately labelled datasets.

Existing Tools

Databases: After decades of space exploration during which the data was only accessible through the PI of the instruments, the community has now moved to the implementation of large public databases of in situ measurements for satellite missions. Among the major ones are NASA CDAweb (http://cdaweb.gsfc.nasa.gov), the french (CNES/CNRS) one (http://cdpp.eu), or mission-dependant databases (e.g. Cluster Science Archive), where data can be downloaded easily. The data from all these missions is distributed in a universally used file format called CDF (Common Data Format).

Existing Software: Building large databases has been a big step over the last 15 years, but we still critically lack the tools to efficiently explore them. Most of the currently used tools fall into two categories : 1 - *packages of routines in scripting/interpreted languages ;* 2 - *web based graphical interfaces*. Most of these tools, have been developped by scientists themselves and are not designed to be widely used, efficient and versatile. **Scripting tools** represent by far the largest category. While they bring a flexibility that is absolutly needed for research, they require researchers to write code before doing anything, including just browsing data. Easy and interactive manipulation of data is hardly feasible, very time consuming and results in the multiplication of small private and dedicated pieces of code. Among those tools one can note two major ones : TDAS Tplot (Themis

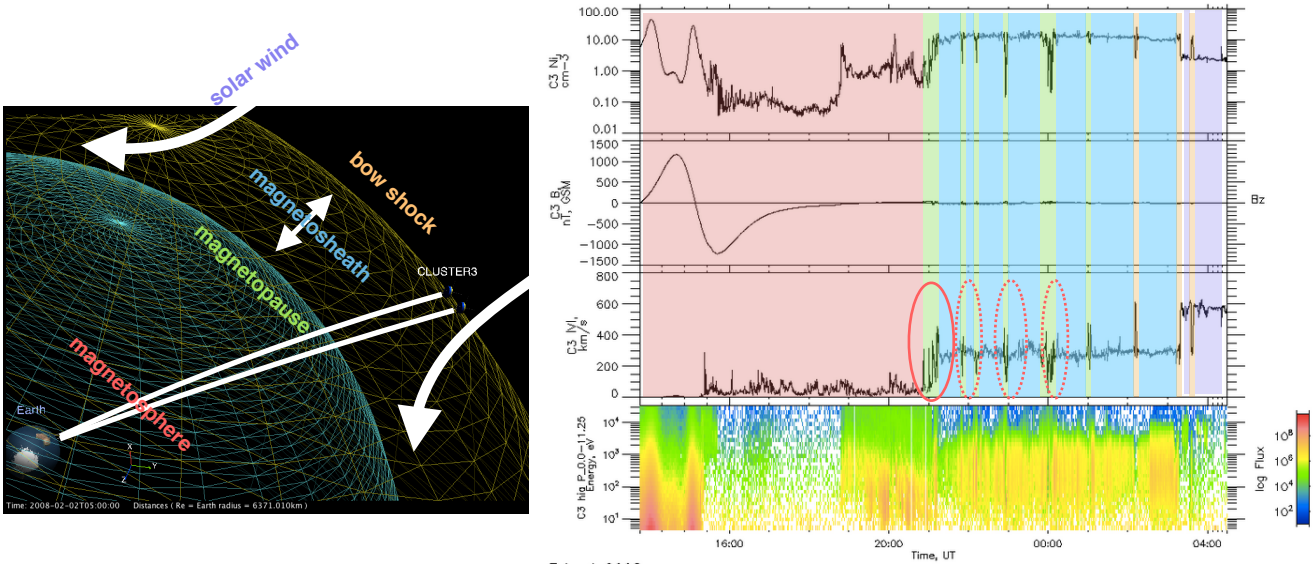

Fig. 2 - The *left panel* is a 3D representation obtained from 3Dview of Cluster orbit crossing the magnetopause (green lien boundary) and the bow shock (orange line boundary). The data measured during that period is represented on the *right panel.* From the top to bottom, is represented the plasma density, in cm^{-3}, the magnetic field in the north-south direction, the plasma velocity in the Earth-Sun direction, and a time/energy spectrogram where the color codes the particle density at a specific time and for a specific energy. Those measurements are among many other those with which one can distinguish the crossing of the boundaries represented on the left panel. A visual inspection allows one to draw colored rectangle associated to time intervales within which the spacecraft is exploring one particular region. These intervals could have been automatically selected once the program has leearned how to define them. This plot as been made by the software AMDA.

Data Analysis Software), written in the IDL language and is therefore not free and expensive[1], and the less mature but free an open-source python package Spacepy. Both are interpreted language packages and therefore do not allow for high performance real time and interactive visualization.

Web based softwares are more for data browsing than pure data analysis. They allow for easy access to the data through web browser but are generally based on automatically generated PNG images which severely limits performances and interactivity. Examples of such softwares are the french application CLWeb, which is now quite old, poorly documented and not efficient, mostly used by the developers' lab ; AMDA is a more recent project built on top of the french CDPP database, allows for transparent visualization of multi-mission data and catalogue building, based on visual inspection and fixed rule algorithms only. After several years, it is still poorly used by the community, mostly because of poor visualization performances and data mining capabilities (based on visual inspection and « threshold crossing » algorithms.

SciQLOP

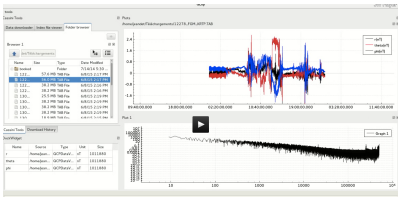

Fig 3 - A preview on the current, proof-of-concept-preliminary version of SciQLOP is visible on a tutorial video, also visible here.

What SciQlop is about - general philosophy: SciQLOP is not just another visualization toolkit, competing with previously described ones. It is about enabling easy data browsing, selection and sharing through an interactive and efficient graphical interface including common actions routinely applied to space data, no matter what mission it is from. SciQLOP's ultimate goal is to be the bridge between large databases and specific scientific analysis routines, by enabling users to select accurate, complete and sharable datasets and pass it to in-house/community developed python data analysis toolkits. Built-in python interpreter allows for moving data back and forth between those python packages and SciQLOP' visualization interface. Furthermore, SciQLOP's most inovative feature is the embedding of machine learning methods at its core. Those will drastically facilitate data selection and enable collaborative building of accurate event catalogues. *This feature is a complete game changer in a way we tackle space data* and hopefully will make our results stronger, more statistically representative of our environment, and more reproducible. SciQLOP is meant to be public, open-source and free. The code in its preliminary version is available on our public repository and under the GPL

[1] Which is also a major drawback for using it for teaching purposes !!

licence.

An interactive, ergonomic and efficient interface: SciQLOP aims in enabling routinely used actions one performs on space data within high perf. graphical interface. This prevents using interpreted languages and associated ploting libraries such as matplotlib or IDL, though generally chosen by scientists as their prefered/ only mastered language. Our choice has been to select qcustom plot. (see Fig. 3). Based on a plugin-approach, SciQLOP's graphical interface will be highly customizable *without being a Rube Goldberg machine*, and will propose to the user the adequate menus, actions and toolboxes depending on what is actually being visualized, or on what particular physics the user is interested in[2]. SciQLOP will allow the user to save and restore panels constituting their interface so to be able to come back to it later whenever the same toolset are useful.

Visualization and scripting : Built-in iPython terminal: A major asset of SciQLOP is to ally an efficient graphical interface with the power and flexibility of scripting toolkits via its readily accessible built-in python terminal. Python is free and there is a fast growing community developing tools that are important for our research and can be imported in SciQLOP. Among them are (not mentioning scipy of course) the Spacepy library for space plasma physics routines, and scikit-learn, with which SciQLOP will largely interact for its machine learning capabilities (see below). The laboratory of plasma physics is also developing an in-house open source python space physics library, that will grow over time and will be available for import within SciQLOP. Most of scientific analysis routines will be developed in this format by scientists themselves, independently from the development of SciQLOP itself, which will be focused on the interactive visualization interface.

Machine learning capabilities , collaborative catalogue building and non-analytical regression models:

A first ambition for SciQLOP is to interact with scikit-learn and enable **automatic feature detection and collaborative catalogue building** using machine learning methods. SciQLOP will be able to learn from user selected events and suggest new time intervals where a specific feature is recognized. Fig. 4 illustrates a panel where SciQLOP suggests magnetopause crossing[3] intervals to the user. The user can either accept all, accept some, reject, redefine the intervals as desired. Catalogue building will be implemented so to be collaborative. Labelled intervals (bookmarks in SciQLOP terminology) could be shared accross all SciQLOP instances. A catalogue panel (Fig. 6) would let the user know, for each relevant mission in their catalogue, what portion of the data has been labelled and is included in the training set of a particular model. Public catalogues will eventually be validated through experts in the field. Labelled intervals, either from a human or SciQLOP, will be seen as « bookmarks » in the data, for which SciQLOP will be able to plot any available quantity, including user defined ones.

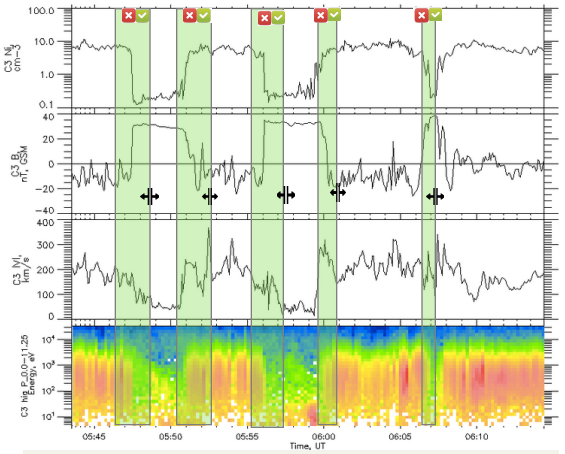

Fig 4 - A preview of what labeling data for a magnetopause detection model could look like in SciQLOP. Based on previous learning, SciQLOP suggests the user specific intervals which may be either redefined, accepted (green checkbox) or rejected (red checkbox). New intervals may also be defined in case one is missed.

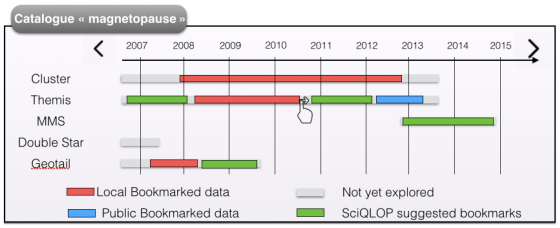

Fig. 6 - Preview of what may be the future SciQLOP panel for collaborative catalogue building. One catalogue (here magnetopause bookmarks) is represented. All spacecraft from which bookmarks are created either from human visual inspection or SciQLOP suggestions. Users can select time intervals to use, see, define, change, remove, accept, reject, share etc. bookmarks.

Another machine learning capability of SciQLOP is to **build non-analytical model**s, useful for scientific research as well as space weather prediction research. Using a bow shock crossing catalogue, one could for instance use statistical regression methods to determine what is the three dimensional shape/location of the bow shock as a function of control parameters such as the solar

[2] e.g. routine treatments for shock physics are not necessarily the same as those routinely used for people browsing reconnection events

[3] There are at least 2 crossings per orbit, ~1 orbit per day, for ~decades and for ~tens of satellites, not counting multiple crossing resulting from the oscillotary motion of the boundary

wind dynamic pressure and magnetic field, without assuming any underlying analytical model. One may also try to determine where, on the magnetopause surface, where the process of magnetic reconnection could occur, as a function of control parameters such as the orientiation of the interplanetary magnetic field. Yet another example is trying to predict the « amplitude » of geomagnetic storms, knowing solar wind properties etc. Those questions are among the most important ones in the space physics/magnetospheric community but can't really be addressed today in a statistical manner. The important point is that all those statistical regression models are nowadays extremely difficult to perform because a prerequisite is having in hand large catalogues of bookmarked data. This second objective is a natural by-product of our first goal.

Project overview

Project Manager : Nicolas Aunai CR CNRS, LPP, Polytechnique, expert in space physics

Technical manager: Alexis Jeandet, IE CNRS, expert in soft. dev. and high perf visualization, LPP, polytechnique

Consulting: Erwan Le Pennec, Pr. Ecole Polytechnique, CDS member. Expert in statistical learning.

Hired engineer: Expert in C++ development, GUI, and comfortable with Maths.