**Volume Title**
ASP Conference Series, Vol. **Volume Number**
**Author**


# Exorcising the Ghost in the Machine: Synthetic Spectral Data Cubes for Assessing Big Data Algorithms

Mauricio Araya[1], Mauricio Solar[1], Diego Mardones[2] and Teodoro Hochfärber[1]

[1]*Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile.*

[2]*Universidad de Chile, Camino El Observatorio 1515, Santiago, Chile.*

**Abstract.** The size and quantity of the data that is being generated by large astronomical projects like ALMA, requires a paradigm change in astronomical data analysis. Complex data, such as highly sensitive spectroscopic data in the form of large data cubes, are not only difficult to manage, transfer and visualize, but they also turn unfeasible the use of traditional data analysis techniques and algorithms. Consequently, the attention have been placed on machine learning and artificial intelligence techniques, to develop approximate and adaptive methods for astronomical data analysis within a reasonable computational time. Unfortunately, these techniques are usually sub optimal, stochastic and strongly dependent of the parameters, which could easily turn into "a ghost in the machine" for astronomers and practitioners. Therefore, a proper assessment of these methods is not only desirable but mandatory for trusting them in large-scale usage. The problem is that positively verifiable results are scarce in astronomy, and moreover, science using bleeding-edge instrumentation naturally lacks of reference values. We propose an Astronomical SYnthetic Data Observations (ASYDO), a virtual service that generates synthetic spectroscopic data in the form of data cubes. The objective of the tool is not to produce accurate astrophysical simulations, but to generate a large number of labelled synthetic data, to assess advanced computing algorithms for astronomy and to develop novel Big Data algorithms. The synthetic data is generated using a set of spectral lines, template functions for spatial and spectral distributions, and simple models that produce reasonable synthetic observations. Emission lines are obtained automatically using IVOA's SLAP protocol (or from a relational database) and their spectral profiles correspond to distributions in the exponential family. The spatial distributions correspond to simple functions (e.g., 2D Gaussian), or to scalable template objects. The intensity, broadening and radial velocity of each line is given by very simple and naive physical models, yet ASYDO's generic implementation supports new user-made models, which potentially allows adding more realistic simulations. The resulting data cube is saved as a FITS file, also including all the tables and images used for generating the cube. We expect to implement ASYDO as a virtual observatory service in the near future.

## 1. Introduction

The data deluge problem in astronomy is rapidly moving from a forecast to a reality. Analyzing Astronomical Big Data (ABD) imposes new scientific challenges for astronomers that are not yet fully understood, meanwhile massive astronomical datasets are starting to pile. There is a growing consensus in the community that machine/statistical learning and artificial intelligence techniques could be the key to cope

with ABD (Ball & Brunner (2010)), as they have been successfully applied in other Big Data domains. A common misconception when using these techniques is to think of them as black-box machines, leading most of the times to mediocre results. Therefore, a good integration of these techniques to astronomical data analysis needs not only a proper understanding of the theory behind these methods, but also a mechanism to assess the quality of the results.

To tackle this issue, we propose generating synthetic astronomical data to test and evaluate complex algorithms and their parameters. The key idea is to use very simple astrophysical models to generate a huge amount of data (ABD) that resembles real observations in terms of dimensionality, sparseness and complexity.

## 2. Synthetic Spectroscopic Data Cubes

The Atacama Large Millimeter/Submillimeter Array (ALMA) is currently generating observations that clearly qualify as ABD, at least in the terms of dimensionality and complexity. The main data products after correlation and calibration are high resolution spectroscopic data cubes, which are not only difficult to manage and visualize, but they also carry complex information about the molecules of one or several sources, and their dynamics in the form of emission lines Teuben et al. (2013).

A simple way to generate a synthetic cube is to use the following model for a temperature of data cell:

$$C(x, y, f) = \sum_{l \in \mathcal{L}(x,y,f)} \int_{\nu_f - \Delta\nu/2}^{\nu_f + \Delta\nu/2} Br(\nu, l) df + \epsilon, \tag{1}$$

where $x$ and $y$ are spatial indices and $f$ is a spectral index in the cube. For each line $l$ that emits in the field of view of that cell, there is a specific broadening function $Br()$ that sums accordingly to the spectral resolution $\Delta\nu$ for that frequency. Also, an Gaussian noise $\epsilon$ is added to the model.

In practice, lines can be grouped by molecules and their relative intensity can be randomly generated within reasonable ranges obtained from the line's energy levels. Currently, the model used for generating these emission lines is a simplified version of the detection equation in Stahler & Palla (2008).

### 2.1. The ASYDO Package

The Astronomical SYnthetic Data Observations tool (ASYDO), is a python package that generates arbitrary spectroscopic data cubes in the ALMA bands, based on mock astrophysical models. ASYDO works under the Virtual Universe (VU) concept, which is a persistent object that can hold several sources spread in a virtual celestial sphere. Sources that belongs to this universe are defined by a central coordinate ($\alpha_S$, $\delta_S$) and a base red-shift $z_S$. Each source can contain one or more components, and each component uses one of the predefined astrophysical mock models (see Figure 1 (a)).

After defining the sources of the universe, data cubes can be generated by performing *observations* to the virtual universe, by providing the angular central position ($\alpha$,$\delta$), angular resolution $\Delta\theta$ and the Field of View (FOV), as well as the central frequency $\nu$, spectral resolution ($\Delta\nu$) and spectral bandwidth (BW).

The main idea is that each model object knows how to project itself into a specific data cube depending on its parameters and local definitions. This allows, for example, generating cubes for the same region but with different resolutions and/or bands.

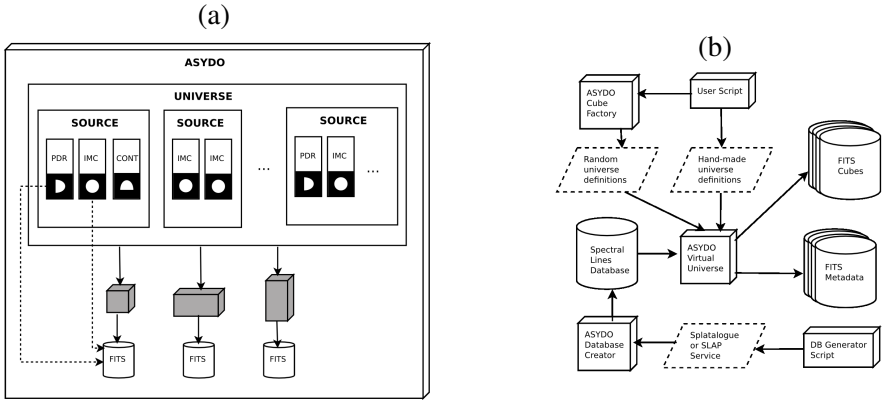


Figure 1. (a) Schematic of the virtual universe and the cube generation process. (b) ASYDO main package components and their interactions, including the VU, DB and Factory.

### 2.2. Tools for Building Models

Even though ASYDO implements an example mock model for generating Interstellar Molecular Clouds (IMC), the main idea of the package is to provide the tools for developing custom models. Currently, ASYDO supports the following functions:

- **Spatial structures**: 2D Gaussians, Generalized 2D Gaussians (Saturated, Lorenzian) and Exponential
- **Spectral distributions**: Gaussian and Skew-Normal
- **Radial velocity gradients**: Linear and Exponential

The functions under development and testing are soft-edge rings and random clouds for spatial structures, voigt profiles (with skew) for spectral distributions, and noisy gradients for radial velocity.

Also, ASYDO provides an unified mechanism to querying spectral line databases for models that require this information. A special module in the package allows generating a SQLite database with a set of spectral lines using IVOA's SLAP protocol, or directly from a CSV file obtained from another database such as Splatalogue (see Figure 1 (b)).

### 2.3. FITS and the Cube Factory

Each cube object contains also all the information used for generating its temperatures. This cube can be exported as a FITS file, that includes several images and tables:

- One 3D image that represents the actual cube
- Three 2D images for each component, representing the temperature, red-shift and broadening maps.
- A binary table for each component, with each entry representing one line. The columns of the table are a unique line code, the molecule name, the chemical

name (formula), the rest frequency, the observed frequency, the base red-shift, and optionally a reference temperature.

These cubes can be generated by defining each model parameters by hand, or by using the *cube factory* module. This module allows defining ranges for the parameter values, from which each instance draws its parameters uniformly. This module allows generating different cubes in parallel using all the available cores for computing.

### 2.4. A Simple Example

We used the cube factory to generate 30000 data cubes of 25x25x1000 cells each, for a 2 GHz bandwidth around the 300 GHz spectrum. Each cube have sources with a random radial velocity between 150 and 1000 km/s, a mean temperature ranging between 50 and 500 K and roughly 30% of the visible molecules in the spectrum. Their fwhm, skewness and radial velocity gradients also slightly vary from one cube to the other. In the half of the cubes we have forced one molecule to be present (Phosphapropynylidyne), and forced the other half to not have this molecule. After training a Support Vector Machine (SVM) using the raw data, we have tested the accuracy of classification, obtaining a pale 62%. As the data is balanced, we can conclude that the SVM is actually finding some patterns to detect the presence of Phosphapropynylidyne. Please note that this accuracy can be significantly improved with a more thoughtful application of SVM by using dimensionality reduction and extracting relevant descriptors.

## 3. Conclusions and Future Work

We expect that ASYDO will help benchmarking machine learning and artificial intelligence algorithms in order to produce novel and ad-hoc methods for ABD. The SVM example shows that synthetic data is useful for benchmarking algorithms, which could lead to more adequate algorithms for the data at hand. An interesting research direction is to train supervised algorithms using synthetic data to be used later with real data, because gathering a large number of labelled data is a complex task.

Besides increasing the variety of functions and models that ASYDO supports, we plan to improve the parallelization support for high-performance computing environments. Also we expect to integrate this package as a contributed package of astropy, and to develop an IVOA-like synthetic data generation standard to be used a web-service.

**Acknowledgments.** This research was funded by CONICYT through the FONDEF D11I1060 and the ICHAA 79130008 projects.

### References

Ball, N. M., & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049

Stahler, S., & Palla, F. 2008, The Formation of Stars (Wiley). URL `http://books.google.cl/books?id=X91UBLr64FMC`

Teuben, P., Ip, C. Y., Mundy, L., & Varshney, A. 2013, in Astronomical Data Analysis Software and Systems XXII, edited by D. N. Friedel, vol. 475 of Astronomical Society of the Pacific Conference Series, 263