Herramientas Computacionales para la Astroinformática

Cristian A. Vega-Martínez (oficina: IMIP, Académicos 2)
Facundo A. Gómez

Herramientas Computacionales para la Astro-informática (HCAI)

Carga horaria

10 SCT – 15 hrs semanales (cronológicas)

El curso requiere ~9 hrs de trabajo autónomo

HORARIO	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES
8:30 - 9:05			Electivo Grupos de galaxias	Electivo Grupos de galaxias	
9:05 - 9:40			Electivo Grupos de galaxias	Electivo Grupos de galaxias	Astronomia Extragalactica 9:30 a 10:10
9:40-10:10					Astronomia Extragalactica 9:30 a 10:10
10:10 - 10:45		Electivo Grupos de galaxias			
10:45 - 11:20		Electivo Grupos de galaxias			Astroinformatica 11:20 a 12:30
11:50 - 12:25		Astroinformatica 11:50 a 12:30		Astronomia Extragalactica 11:50 A 13	Astroinformatica 11:20 a 12:30
12:25 - 13:00		Astroinformatica 11:50 a 12:30		Astronomia Extragalactica	
13:00					
14:30					
14:30 - 15:05		Journal Club 14:30-15:30	Horario protegido para reuniones de claustro, etc.	Astroinformatica 14:30 a 15:40	
15:05 - 15:40	Astronomia Extragalactica 15 A 16:30			Astroinformatica 14:30 a 15:40	
16:00 - 16:35	Astronomia Extragalactica			Seminario Depto Astronomia	
16:35 - 17:10					
17:30 - 18:05		Curso Transversal Ampliando la mirada en Derechos Humanos, Violencia y Discriminación de Género, Interculturalidad y Ciudadanía			
18:05 - 18:40					
19:00 - 19:35					
19:35 - 20:20					

Horario

M: 11:50 - **13:00**

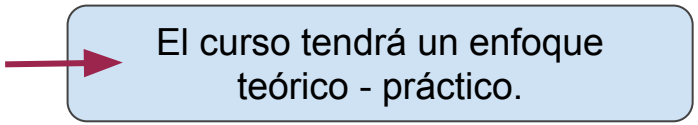
J: 14:30 - 15:40

V: **11:20** - **12:30**

Atención prof. CVM:
V: 14:30 - 16:00
(u otro a coordinar)

Objetivos del curso

- Comprender y utilizar herramientas avanzadas de **desarrollo** de software y **análisis** de datos en proyectos de astrofísica numérica.
- Reconocer las diferentes estrategias de **almacenamiento** y **acceso** de datos astronómicos. Utilizar lenguaje SQL en plataformas de almacenamiento de datos.
- Comprender los fundamentos de la **computación del alto rendimiento** y cálculo en paralelo.
- Aplicar modelos numéricos específicos para producir **simulaciones** astrofísicas.



HCAI - Programa

0. **Presentación / Introducción**

1. **Herramientas de Desarrollo**
   Control de revisiones (nodos, ramas, clonación), administradores web (repositorios, issues, wikis), licencias, IDEs.

2. **Técnicas de programación Avanzada**
   Niveles de abstracción, introducción a C y C++, compilación y linker, make, POO en Python, UMLs, Doxygen.

3. **Introducción a la administración de datos.**
   Tipos de datos, representación, almacenamiento: ASCII, binario, HDF5, bases de datos relacionales y SQL. Catálogos web.

4. **High performance computing**
   Computación distribuida, computación en paralelo: memoria compartida vs distribuida. Clusters de computadoras.

5. **Tópicos de Análisis de datos**
   Mínimos cuadrados, regresión lineal, ajuste de curvas, chi cuadrado, montecarlo, MCMC, períodos, introducción a machine learning.

6. (Facundo) **Introducción a las simulaciones Astrofísicas**
   N-body, hidrodinámica, y más...

Por la primera versión del curso, el programa puede sufrir modificaciones durante el semestre.

Metodología y evaluación

El curso consistirá en **clases expositivas** acompañadas de **sesiones prácticas** donde deberán aplicar el contenido de la(s) clase(s). Durante el curso, cada estudiante deberá:

- **Profundizar** temas indicados en las clases. Presentarlos de formato mini-charla de forma voluntaria (o por sorteo), en formato seminario.
- Crear y mantener un **portafolio del curso**, que contenga los ejemplos y códigos utilizados para el estudio práctico del contenido.
- Desarrollar un **proyecto principal**, donde debe utilizar contenido del curso de forma profunda en un proyecto personal (preferentemente su investigación).

Idealmente, utilizar un recurso computacional personal durante las clases.

Metodología y evaluación

- 20% Mini-presentaciones de temas del curso
- 30% Presentación de portafolio del curso (contenido general, dos revisiones)
- 50% Proyecto ~~final~~ principal: informe y dos presentaciones (anteproyecto y final).

Presentaciones

- **Midterm** - 26 y 27 de octubre: portafolio y anteproyecto.
- **Final** - 12, 14 y 15 de diciembre: proyecto e informe final.

Proyectos principales

La realización de un trabajo práctico donde se apliquen (un conjunto de) los conocimientos tratados en el curso, en un contexto **personal** y con **profundidad**.

Ejemplos de (posibles) proyectos:

- Crear un código para su investigación que utilice herramientas del curso.
- Re-escribir un código/pipeline personal usando POO.
- Utilizar Doxygen para documentar un código/pipeline personal.
- Aprender un lenguaje de bajo nivel (C/C++, Fortran) y crear un software.
- Intervenir el almacenamiento de datos de un proyecto.
- Diseñar e implementar un código para calcular alguna simulación astrofísica (o de otra área).
- etc...

Proyectos principales: requerimientos

- Los proyectos a realizar deben ser aprobados por el profesor **antes** de la presentación mid-term (deadline: 12/oct).
- Los proyectos pueden ser grupales, pero las presentaciones, informes y evaluaciones abordarán su contribución individual. Un proyecto grupal debería mostrar avances proporcionales al número de integrantes y definir una persona coordinadora.
- No es necesario haber completado el proyecto al momento de su presentación final, pero sí mostrar los avances realizados.

Algunas definiciones

¿Qué es la Astroinformática?

¿Es un área nueva?
¿Qué conocimientos involucra?
¿Cuáles son sus líneas de investigación?

GPT-4 (+Internet, +Wiki) + profe

La **astroinformática** es un campo **interdisciplinario** que combina **astronomía**, **ciencia de datos**, **informática** y **tecnologías de la información y la comunicación**. Se centra principalmente en el desarrollo de herramientas, métodos y aplicaciones de ciencia computacional, ciencia de datos y estadística para la investigación y educación en astronomía orientada a los datos. Algunos de los esfuerzos en esta dirección incluyen: la exploración de datos, el desarrollo de estándares de metadatos, el modelado de datos, el desarrollo de diccionarios de datos astronómicos, acceso de datos, búsqueda y recuperación de información, integración de datos y minería de datos en las iniciativas astronómicas de observatorios virtuales.

Astro2010 State of the Profession Position Paper (March 2009)

Astroinformatics: A 21st Century Approach to Astronomy

Primary Author: Kirk D. Borne, Dept. of Computational and Data Sciences, 4400 University Drive MS 6A2, George Mason University, Fairfax, VA 22030 USA (kborne@gmu.edu).

Abstract:

Data volumes from multiple sky surveys have grown from gigabytes into terabytes during the past decade, and will grow from terabytes into tens (or hundreds) of petabytes in the next decade. This exponential growth of new data both enables and challenges effective astronomical research, requiring new approaches. Thus far, astronomy has tended to address these challenges in an informal and ad hoc manner, with the necessary special expertise being assigned to e-Science or survey science. However, we see an even wider scope and therefore promote a broader vision of this data-driven revolution in astronomical research. For astronomy to effectively cope with and reap the maximum scientific return from existing and future large sky surveys, facilities, and data-producing projects, we need our own information science specialists. We therefore recommend the formal creation, recognition, and support of a major new discipline, which we call Astroinformatics. Astroinformatics includes a set of naturally-related specialties including data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, data mining, machine learning, visualization, and astrostatistics. By virtue of its new stature, we propose that astronomy now needs to integrate Astroinformatics as a formal sub-discipline within agency funding plans, university departments, research programs, graduate training, and undergraduate education. Now is the time for the recognition of Astroinformatics as an essential methodology of astronomical research. The future of astronomy depends on it.

Earth Sci Inform (2010) 3:5–17
DOI 10.1007/s12145-010-0055-2

RESEARCH ARTICLE

Astroinformatics: data-oriented astronomy research and education

Kirk D. Borne

Received: 22 October 2009 / Accepted: 20 April 2010 / Published online: 12 May 2010

Abstract The growth of data volumes in science is reaching epidemic proportions. Consequently, the status of data-oriented science as a research methodology needs to be elevated to that of the more established scientific approaches of experimentation, theoretical modeling, and simulation. Data-oriented scientific discovery is sometimes referred to as the new science of X-Informatics, where X refers to any science (e.g., Bio-, Geo-, Astro-) and informatics refers to the discipline of organizing, describing, accessing, integrating, mining, and analyzing diverse data resources for scientific discovery. Many scientific disciplines are developing formal sub-disciplines that are information-rich and data-based, to such an extent that these are now stand-alone research and academic programs recognized on their own merits. These disciplines include bioinformatics and geoinformatics, and will soon include astroinformatics. We introduce Astroinformatics, the new data-oriented approach to 21st century astronomy research and education. In astronomy, petascale sky surveys will soon challenge our traditional research approaches and will radically transform how we train the next generation of astronomers, whose experiences with data are now increasingly more virtual (through online databases) than physical (through trips to mountaintop observatories). We describe Astroinformatics as a rigorous approach to these challenges. We also describe initiatives in science education (not only in astronomy) through which students are trained to access large distributed data repositories, to conduct meaningful scientific inquiries into the data, to mine and analyze the data, and to make data-driven scientific discoveries. These are essential skills for all 21st century scientists, particularly in astronomy as major new multi-wavelength sky surveys (that produce petascale databases and image archives) and grand-scale simulations (that generate enormous outputs for model universes, such as the *Millennium Simulation*) become core research components for a significant fraction of astronomical researchers.

Keywords Data mining · Informatics · Data integration · Semantic metadata · Knowledge discovery · Science education

Introduction

New modes of discovery are enabled by the growth of data and computational resources in the sciences. This cyberinfrastructure includes databases, virtual observatories (distributed data), high-performance computing (clusters and petascale machines), distributed computing (the Grid, the Cloud, peer-to-peer networks), intelligent search and discovery tools, and innovative visualization environments (Eastman et al. 2005; Hey et al. 2009). Data streams from experiments, sensors,

DOI: 10.1007/s12145-010-0055-2

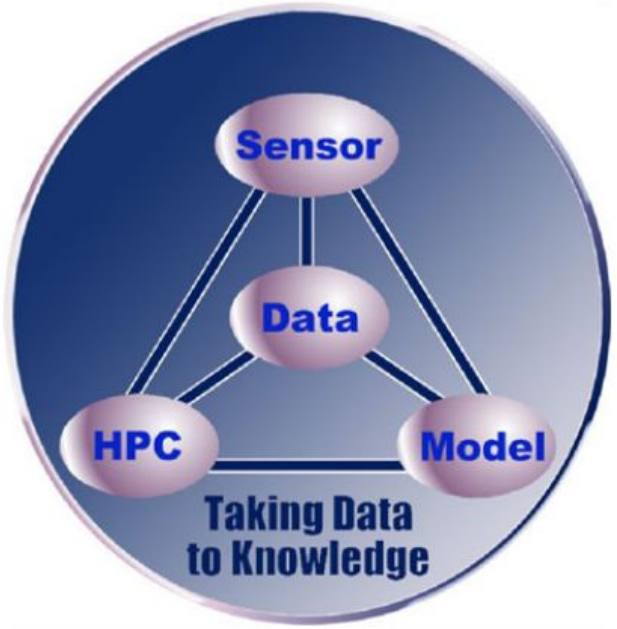


Fig. 1 The new four-pronged approach to scientific research, in which data, sensors, models, and simulations act synergistically to drive new discoveries—shown here with Data as the central hub of scientific activities (Eastman et al. 2005)

Earth Sci Inform (2010) 3:5–17
DOI 10.1007/s12145-010-0055-2

RESEARCH ARTICLE

Astroinformatics: data-oriented astronomy research and education

Kirk D. Borne

Received: 22 October 2009 / Accepted: 20 April 2010 / Published online: 12 M

Abstract The growth of data volumes in science is reaching epidemic proportions. Consequently, the status of data-oriented science as a research methodology needs to be elevated to that of the more established scientific approaches of experimentation, theoretical modeling, and simulation. Data-oriented scientific discovery is sometimes referred to as the new science of X-Informatics, where X refers to any science (e.g., Bio-, Geo-, Astro-) and informatics refers to the discipline of organizing, describing, accessing, integrating, mining, and analyzing diverse data resources for scientific discovery. Many scientific disciplines are developing formal sub-disciplines that are information-rich and data-based, to such an extent that these are now stand-alone research and academic programs recognized on their own merits. These disciplines include bioinformatics and geoinformatics, and will soon include astroinformatics. We introduce Astroinformatics, the new data-oriented approach to 21st century astronomy research and education. In astronomy, petascale sky surveys will soon challenge our traditional research approaches and will radically transform how we train the next generation of astronomers, whose experiences with data are now increasingly more virtual (through online databases) than physical (through trips to mountaintop observatories). We describe Astroinformatics as a rigorous approach to these challenges. We also describe initiatives in science education (not only in astronomy) through which students are trained to

Semantic metadata · Knowledge discovery · Science education

Introduction

New modes of discovery are enabled by the growth of data and computational resources in the sciences. This cyberinfrastructure includes databases, virtual observatories (distributed data), high-performance computing (clusters and petascale machines), distributed computing (the Grid, the Cloud, peer-to-peer networks), intelligent search and discovery tools, and innovative visualization environments (Eastman et al. 2005; Hey et al. 2009). Data streams from experiments, sensors,

DOI: 10.1007/s12145-010-0055-2

Actividad 1

Leer este artículo: contrastar su apreciación actual del área con lo proyectado hace 13 años.

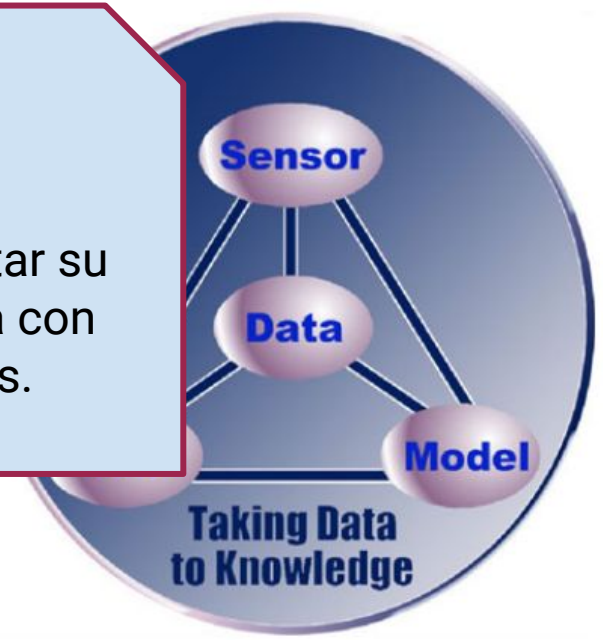


Fig. 1 The new four-pronged approach to scientific research, in which data, sensors, models, and simulations act synergistically to drive new discoveries—shown here with Data as the central hub of scientific activities (Eastman et al. 2005)

10 años después…

arXiv:1903.06796

Astrophysics > Instrumentation and Methods for Astrophysics

[Submitted on 15 Mar 2019]

Astro2020 Science White Paper: The Next Decade of Astroinformatics and Astrostatistics

A.Siemiginowska, G. Eadie, I. Czekala, E. Feigelson, E.B. Ford, V. Kashyap, M. Kuhn, T.Loredo, M.Ntampaka, A. Stevens, A. Avelino, K. Borne, T. Budavari, B. Burkhart, J. Cisewski-Kehe, F. Civano, I. Chilingarian, D.A. van Dyk, G. Fabbiano, D. P. Finkbeiner, D. Foreman-Mackey, P. Freeman, A. Fruscione, A.A. Goodman, M. Graham, H.M. Guenther, J. Hakkila, L. Hernquist, D. Huppenkothen, D. J. James, C. Law, J. Lazio, T. Lee, M. López-Morales, A. A. Mahabal, K. Mandel, X.L. Meng, J. Moustakas, D. Muna, J. E. G. Peek, G.Richards, S. K.N. Portillo, J. Scargle, R. S. de Souza, J. S. Speagle, K. G. Stassun, D. C. Stenning, S. R. Taylor, G. R. Tremblay, V. Trimble, P.A. Yanamandra-Fisher, C. A. Young

Over the past century, major advances in astronomy and astrophysics have been largely driven by improvements in instrumentation and data collection. With the amassing of high quality data from new telescopes, and especially with the advent of deep and large astronomical surveys, it is becoming clear that future advances will also rely heavily on how those data are analyzed and interpreted. New methodologies derived from advances in statistics, computer science, and machine learning are beginning to be employed in sophisticated investigations that are not only bringing forth new discoveries, but are placing them on a solid footing. Progress in wide-field sky surveys, interferometric imaging, precision cosmology, exoplanet detection and characterization, and many subfields of stellar, Galactic and extragalactic astronomy, has resulted in complex data analysis challenges that must be solved to perform scientific inference. Research in astrostatistics and astroinformatics will be necessary to develop the state-of-the-art methodology needed in astronomy. Overcoming these challenges requires dedicated, interdisciplinary research. We recommend: (1) increasing funding for interdisciplinary projects in astrostatistics and astroinformatics; (2) dedicating space and time at conferences for interdisciplinary research and promotion; (3) developing sustainable funding for long-term astrostatisics appointments; and (4) funding infrastructure development for data archives and archive support, state-of-the-art algorithms, and efficient computing.

Comments: Submitted to the Astro2020 Decadal Survey call for science white papers
Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**
Cite as: arXiv:1903.06796 **[astro-ph.IM]**
(or arXiv:1903.06796v1 **[astro-ph.IM]** for this version)
https://doi.org/10.48550/arXiv.1903.06796

10 años después…

arXiv:1903.06796

1. What is the role of astrostatistics and astroinformatics research?

→ *To develop modern methods for extracting scientific information from astronomical data.*

Astrostatistics forms the foundation for robust algorithms and principled methods that are applied to a variety of problems in astronomy. **Astroinformatics** involves the systematic and disciplined development of code, data management and dissemination techniques, high-performance computing, and machine learning based inference. Both astrostatistics and astroinformatics (i.e., astro data science) have been rapidly emerging **fields of research** rigorously pursued at the intersection of observational astronomy, statistics, algorithm development, and data science [14, 93, 49, 40, 136]. The number of articles with keyword 'Methods: Statistical' increased by a factor of 2.5 in the past decade; those with 'machine learning' increased by 4 times over five years; and those with 'deep learning' have more than tripled every year since 2015. Thus, the challenges of astronomical sciences reveal a deep and broad demand for advanced methodology and techniques. *Astronomical problems impossible to approach with traditional methods are now forefront research efforts because of advancements in astrostatistics and astroinformatics.*

Astroinformática - key points

- El Big Data Astronómico (VVVV: Volumen, Velocidad, Variedad, Veracidad) demanda **soluciones** más sofisticadas.
- Demanda formación básica en áreas afines para facilitar la **comunicación**.
- **Mueve los límites** originales del área gracias al desarrollo tecnológico.
- Se está convirtiendo en una **necesidad**.
- La Astronomía/astrofísica deben ser participar y conducir la investigación en **nuevas metodologías**.

Laboratorios científicos

Laboratorio

(Wiki)

Es un lugar dotado de los **medios necesarios** para realizar investigaciones, experimentos, prácticas y trabajos de carácter científico, tecnológico o técnico. Está equipado con instrumentos de medida o equipos con los que se realizan experimentos, investigaciones y prácticas diversas, según la rama de la ciencia a la que se dedique.

Ejemplo: Un laboratorio en Biología:

CC: Instituto de Graduados de Biología del Cáncer, Universidad Médica de China (Taiwán)

"Conocer/dominar los instrumentos del laboratorio es clave para una buena investigación"

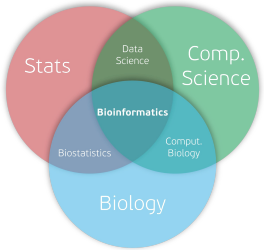
Laboratorio

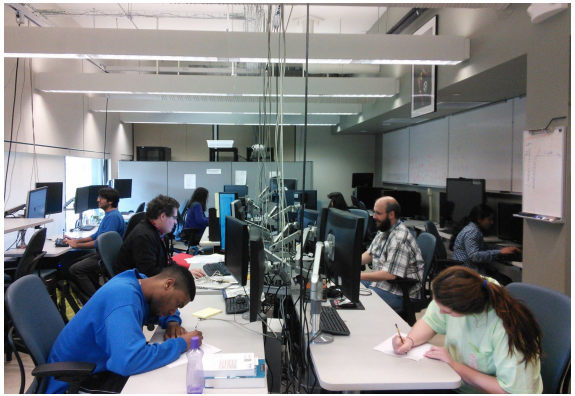
Es un lugar dotado de los **medios necesarios** para realizar investigaciones, experimentos, prácticas y trabajos de carácter científico, tecnológico o técnico. Está equipado con instrumentos de medida o equipos con los que se realizan experimentos, investigaciones y prácticas diversas, según la rama de la ciencia a la que se dedique.

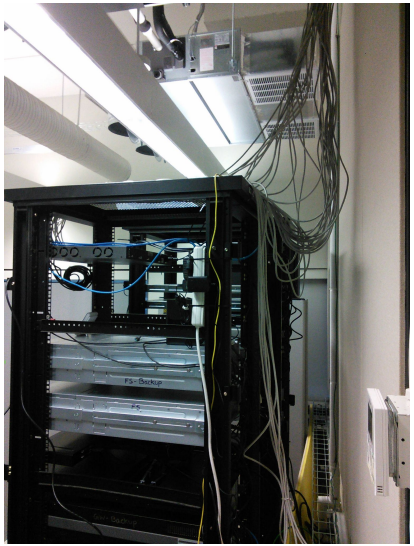
y…. un Laboratorio de Bioinformática?

CC: Moseley Bioinformatics Lab
https://bioinformatics.cesb.uky.edu/Main/LabEnvironment

Genómica
Dinámica Molecular
Modelos Ecológicos
Diseño de drogas
etc….







Astroinformática

En paralelo a la Bioinformática, los instrumentos que constituyen el laboratorio de la Astroinformática son los **recursos computacionales**:

- PCs (hardware y software)
- HPC clusters.

Pregunta personal
¿Qué tanto conozco/domino mi instrumento de trabajo?

cc: Shutterstock

“The golden rule”
para **dominar** una herramienta computacional

R.T.F.M.

*Read the f****ng manual*