Pluralismo
Compromiso
Inclusión

Licenciatura en Astronomía

Procesamiento de datos/data mining

Procesamiento de datos/data mining

I. Fundamentos de la astroinformática

II. Data warehouses y surveys; base de datos

III. Algoritmos de minería de datos

IV. Observatorio Virtual

¿Qué es el Procesamiento de Datos?

El procesamiento de datos es el conjunto de operaciones realizadas sobre datos crudos para convertirlos en información útil.

En astronomía, esto implica:

- **Recolección de datos:** desde telescopios o simulaciones (por ejemplo, GAIA, SDSS, JWST, TNG).
- **Limpieza de datos:** eliminar errores, valores faltantes o inconsistencias.
- **Transformación:** convertir formatos, cambiar unidades, reducir dimensiones.
- **Almacenamiento eficiente:** bases de datos, Data Warehouses, estructuras indexadas.
- **Análisis:** aplicar técnicas estadísticas o computacionales para extraer conocimiento.
- **Visualización:** representar los resultados de forma comprensible (gráficos, mapas, proyecciones celestes).

¿Qué es la Minería de Datos?

Definición

La minería de datos descubre patrones ocultos en grandes conjuntos de datos.

Diferencia

A diferencia del procesamiento, la minería de datos se enfoca en extraer conocimiento.

Tareas

Clasificación, regresión, clustering y asociación son tareas comunes.

Herramientas para el Procesamiento y la Minería de Datos

Python

Con Pandas y Scikit-learn.

SQL

Para bases de datos.

Hadoop

Plataforma Big Data.

Seleccionar la herramienta adecuada depende del proyecto.

Preparación de Datos: Limpieza y Transformación

1

Valores Faltantes

Manejo de valores faltantes.

2

Duplicados

Eliminación de duplicados.

3

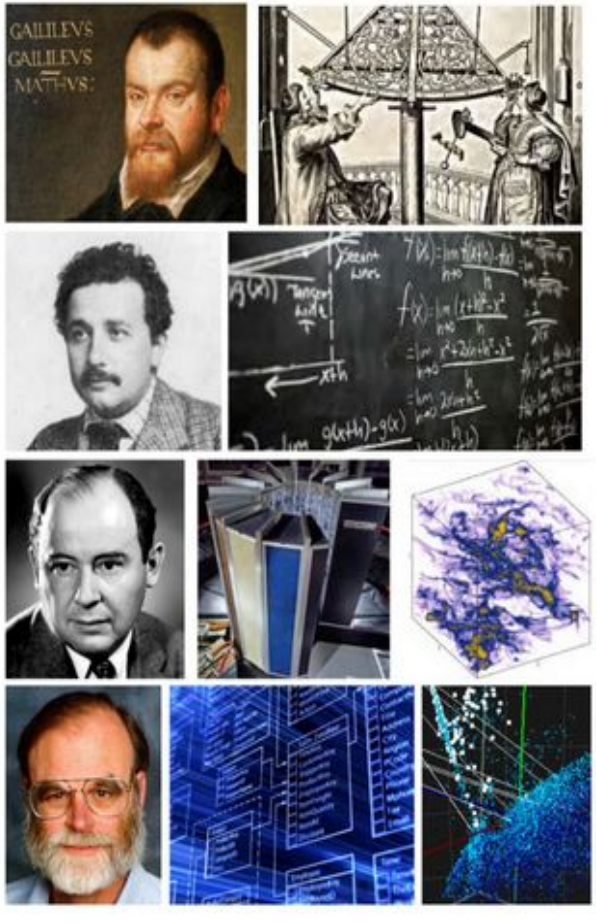
Normalización

Normalización de datos.

La calidad de los datos impacta la calidad de los resultados.

Búsqueda constante del ser humano del conocimiento

- **Experimentos y mediciones**
- **Teoría analítica**
- **Simulaciones numéricas**
- **Ciencia basada en datos**



Créditos: S. George Djorgovski

Con nuevo desarrollo tecnológico surgen nuevas y mayores necesidades

Surgen nuevas necesidades:

- Almacenamiento
- Acceso y lectura
- Procesamiento y Análisis
- Visualización

Con nuevo desarrollo tecnológico surgen nuevas y mayores necesidades

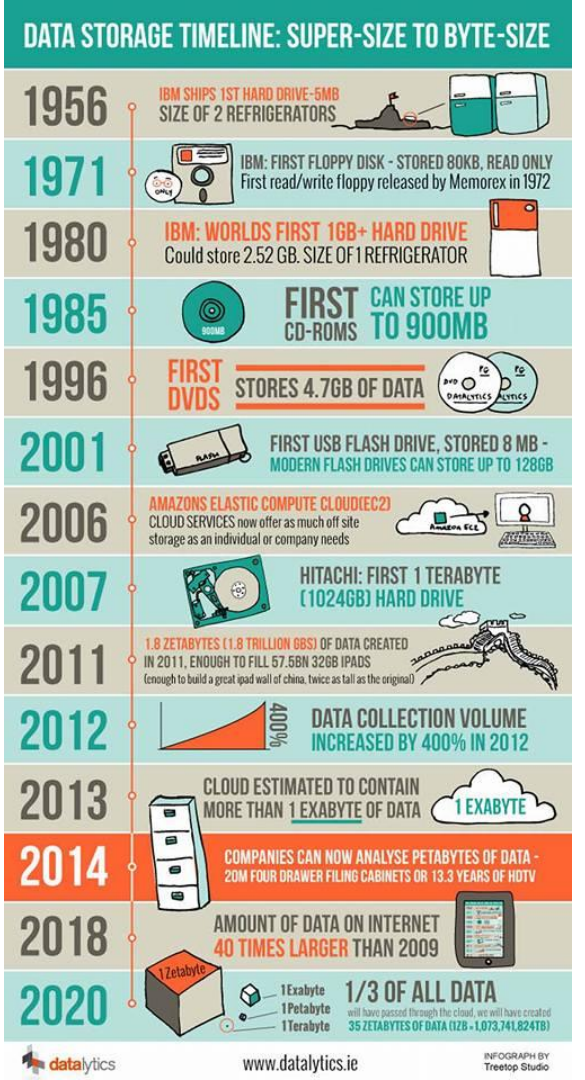
Surgen nuevas necesidades:

- Almacenamiento
- Acceso y lectura
- Procesamiento y Análisis
- Visualización

Crecimiento exponencial de los volúmenes de datos y de la información que contienen!

Crecimiento exponencial de los volúmenes de datos y de la información que contienen!

Prefix	Multiple	Symbol
yotta	10^{24}	Y
zetta	10^{21}	Z
exa	10^{18}	E
peta	10^{15}	P
tera	10^{12}	T
giga	10^{9}	G
mega	10^{6}	M
kilo	10^{3}	k

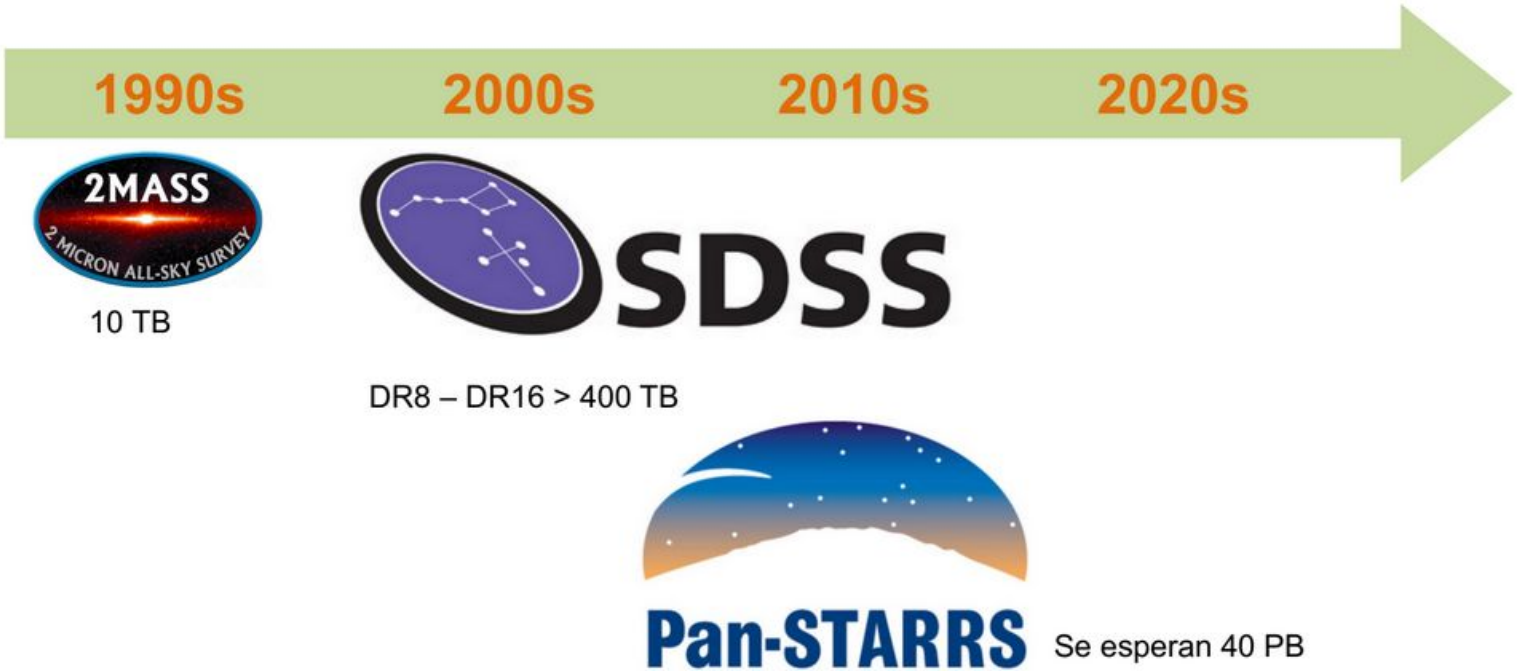


Astroinformática

La astroinformática es una disciplina interdisciplinaria que combina **astronomía, ciencia de datos, estadística e informática** para procesar, analizar y extraer conocimiento de grandes volúmenes de datos astronómicos.

Astroinformática

La astroinformática es una disciplina interdisciplinaria que combina **astronomía, ciencia de datos, estadística e informática** para procesar, analizar y extraer conocimiento de grandes volúmenes de datos astronómicos.



Astroinformática

1990s 2000s 2010s 2020s

2MASS
2 MICRON ALL-SKY SURVEY

10 TB

SDSS

DR8 – DR16 > 400 TB

Pan-STARRS

Se esperan 40 PB

LSST
Legacy Survey of Space and Time

VERA C. RUBIN
OBSERVATORY

20 TB por noche
En 10 años ~ 10^2 PB!

The SKA project in numbers

€1.3 BILLION
CONSTRUCTION COST (2021 €)

131,072 ANTENNAS
IN WESTERN AUSTRALIA

710 PETABYTES
OF SCIENCE DATA DELIVERED TO SCIENCE USERS

€0.7 BILLION
FIRST 10 YEARS OF OPERATIONS COST (2021 €)

197 DISHES
IN SOUTH AFRICA (INCLUDING 64 MEERKAT DISHES)

1 GLOBAL NETWORK
OF DATA CENTRES TO DELIVER SCIENCE-READY DATA PRODUCTS TO END-USERS

8 YEARS
OF CONSTRUCTION ACTIVITIES

16 COUNTRIES
PARTICIPATING IN 2022

50+ YEARS
OF TRANSFORMA
SCIENCE

2020s

SKA
SQUARE KILOMETRE ARRAY

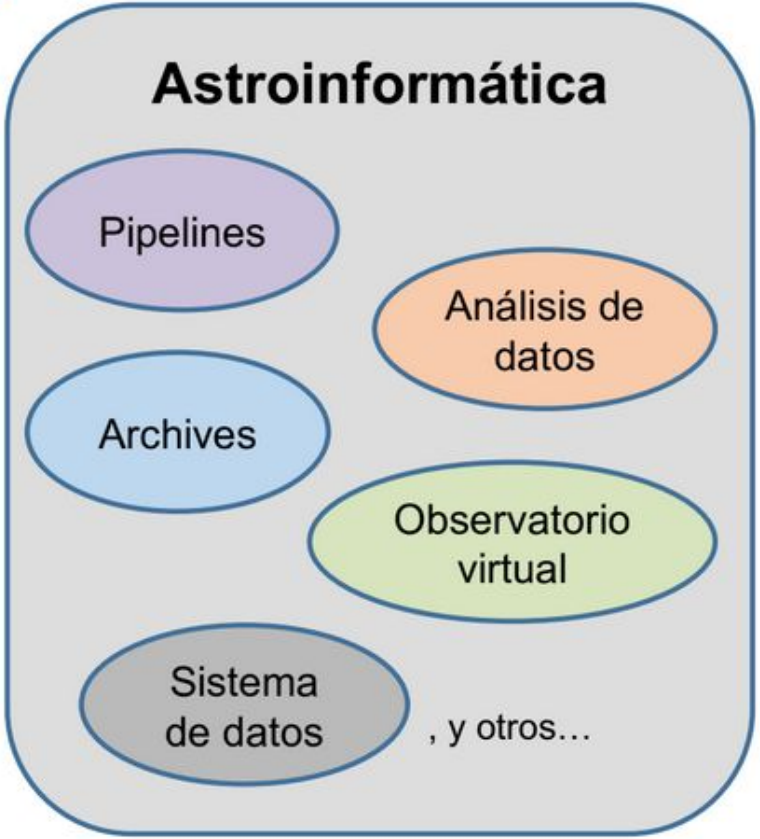
La era del Big Data

Table 1. Big Data 3V characteristics in astronomical sky surveys.

Sky Survey	Volume	Velocity	Variety
SDSS *Sloan Digital Sky Survey*	50 TB	200 GB per day	images, catalogs, redshits
GAIA	100 TB	40 GB per day	more then 100 parameters
Pan-STARRS *Panoramic Survey Telescope and Rapid Response System*	5 PB	5 TB per day	images, catalogs
LSST *Large Synoptic Survey Telescope*	60 PB	10 TB per day	images, catalogs
SKA *Square Kilometer Array*	3 ZB	150 TB per day	images, catalog, redshifts

Notes:
The column Volume refers to raw data produced at the end of the experiment.
Values regarding Pan-STARRS, LSST, and SKA surveys refer to expected Volume and Velocity values.

Garofalo et al. 2016



Como encontramos las grandes cantidades de datos?

- Observaciones con diferentes técnicas (e.g., espectroscopía, imágenes, simulaciones)
- Múltiples fuentes en la muestra o "*survey*" (e.g., estrellas, galaxias, etc)

  https://www.tng-project.org/
- Modelos computaciones (e.g., modelos cosmológicos)
- Múltiples parámetros en modelos y observaciones

Universidad Central

Comisión Nacional de Acreditación CNA-Chile

4 AÑOS ACREDITADA

GESTIÓN INSTITUCIONAL
DOCENCIA DE PREGRADO
VINCULACIÓN CON EL MEDIO

DESDE DICIEMBRE 2017
HASTA DICIEMBRE 2021

www.ucentral.cl