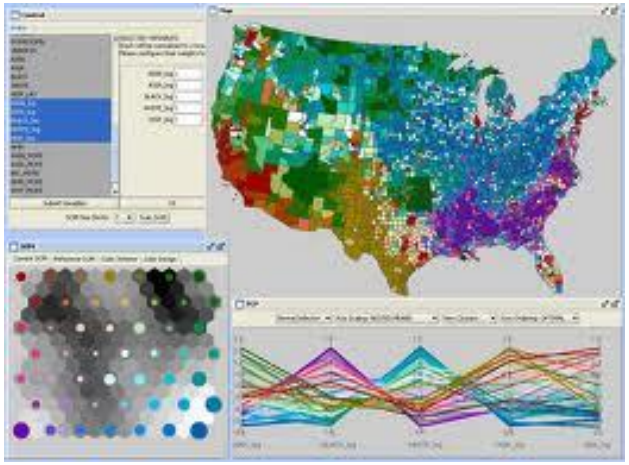
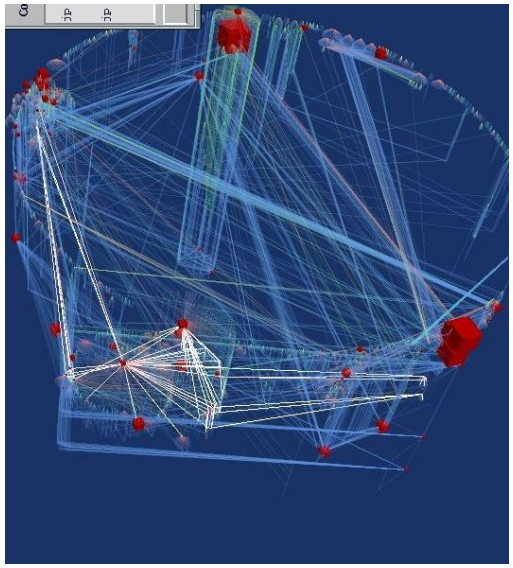
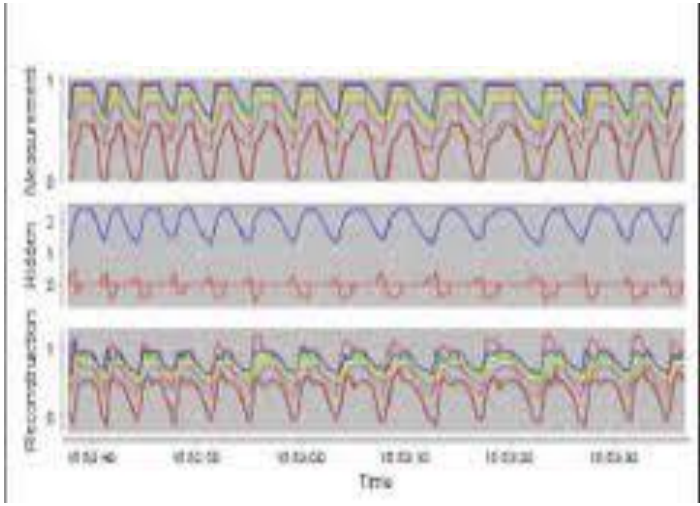
Aplicaciones en Ciencia y Tecnología







Desde el punto científico ¿Cuál debería ser el objetivo del proceso de modelado?

¿Describir la realidad?

¿Ordenar conocimientos en un cuerpo coherente?

¿Hacer predicciones?

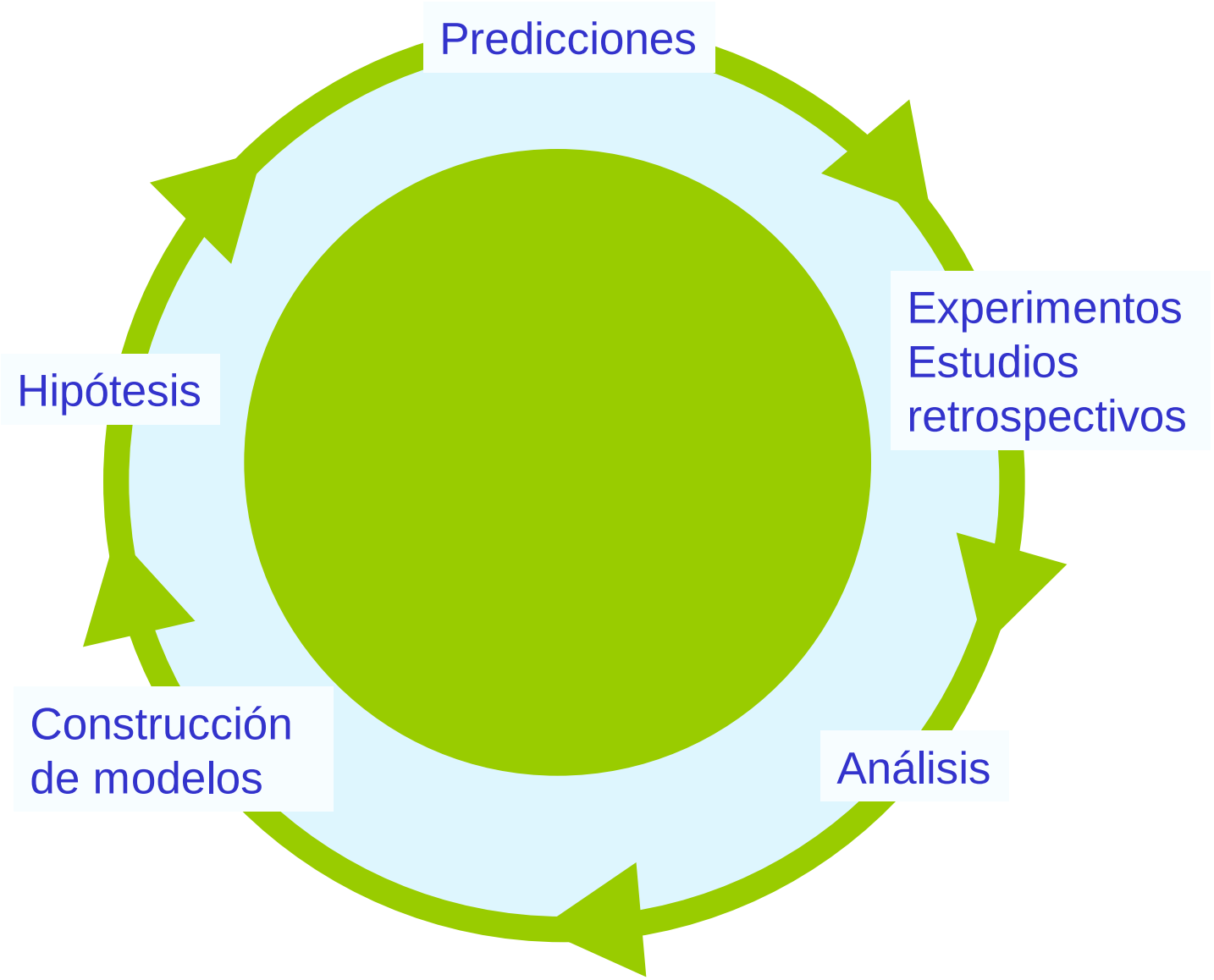
Desde el punto científico ¿Cuál debería ser el objetivo del proceso de modelado?

¿Describir la realidad?

¿Ordenar conocimientos en un cuerpo coherente?

¿Hacer predicciones?

El proceso es continuo…



Algunos dominios de aplicación

- Textos
- Ontologías
- Datos espaciales y temporales
- Imágenes
- Grafos
- Redes Sociales
- Biología
- Series de Tiempo
- Streams

Diferentes problemas pueden tratarse con técnicas emparentadas.

Aunque también existen técnicas propias de cada dominio

Aplicaciones

- **Minería de Texto**
  - Clasificar o categorizar documentos
  - Análisis de encuestas
  - Procesamiento automático de mensajes
  - Construcción de Ontologías
  - Buscadores
- **Datos espaciales**
  - Patrones de evolución de Enfermedades
  - Clasificación en astronomía
- **Grafos**
  - Patrones en estructuras moleculares
  - Patrones de uso en la web
  - Redes sociales
- **Biología**
  - Expresiones de genes
  - Alineamiento de secuencias

Aplicaciones de **D**ata **M**ining en **c**iencia y **t**ecnología

Data mining y astronomía

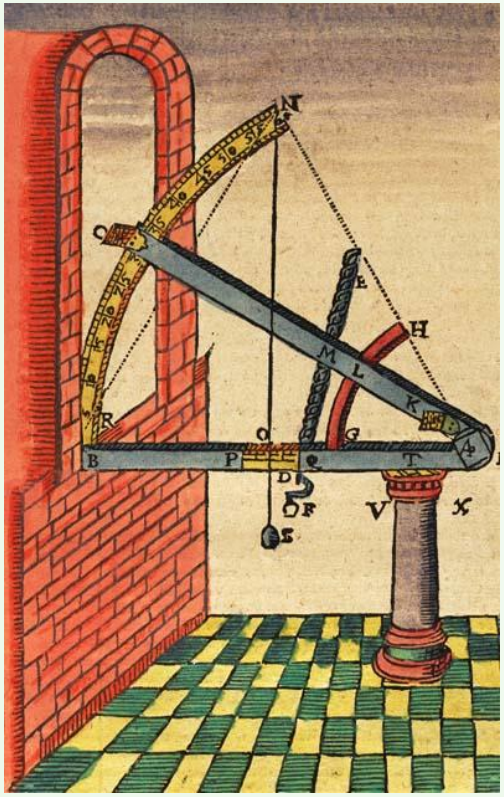
Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

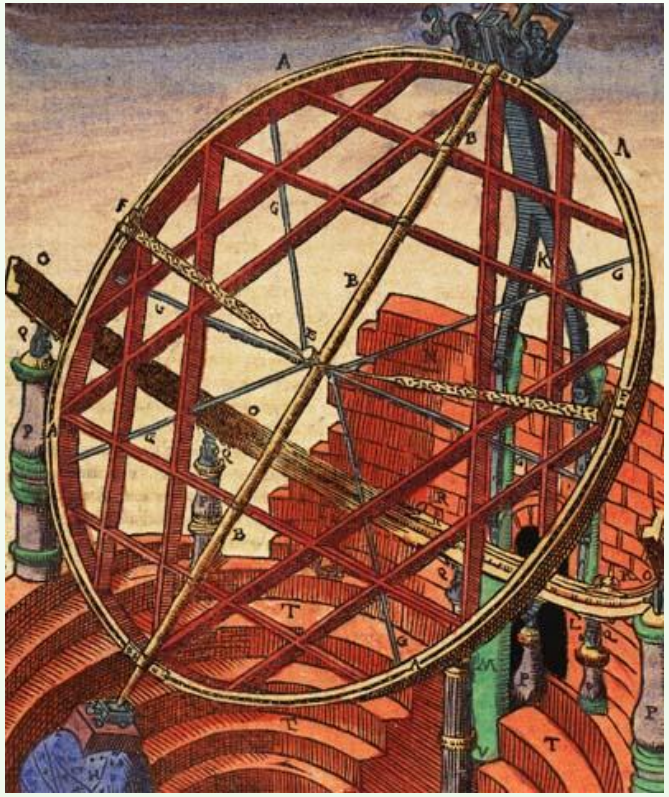
LOS ORIGENES

Tycho Brahe (1546 -1601)
A lo largo de 25 años recolectó las mediciones más precisas de su época de las posiciones de los planetas conocidos y muchas estrellas

Johannes Kepler (1571-1630)
Utilizó los datos de la órbita de Marte que Brahe había recopilado para crear un modelo sobre el movimiento de los planetas

- Tycho Brahe reconoció que los datos astronómicos de su época eran de mala calidad y aprovechando su fortuna construyó un observatorio equipado con los mejores instrumentos de la época y contrató numerosos asistentes para realizar mediciones.





Algunos equipos de observación del observatorio de Tycho Brahe (fuente Wikipedia)

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

- Durante los estudios universitarios Johannes Kepler aprendió el viejo sistema de Ptolomeo, pero una vez graduado uno de sus profesores lo introdujo al sistema Copernicano

- En 1597 Kepler publicó “El Misterio Cosmográfico” donde presentó una modificación al sistema copernicano en el que las órbitas de los planetas estaban descriptas por los cinco sólidos regulares.

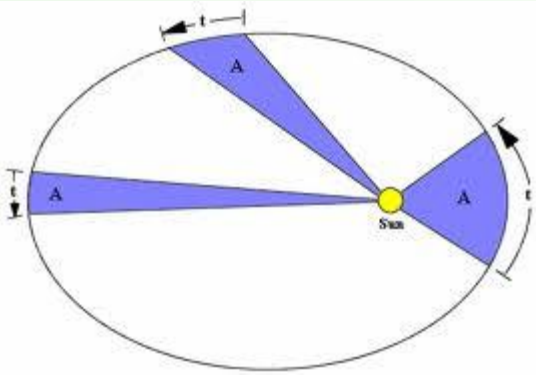
- Impresionado por este libro, Brahe invitó a Kepler a Praga. A pesar de sus diferencias personales, Kepler accede a las observaciones de la órbita de Marte.

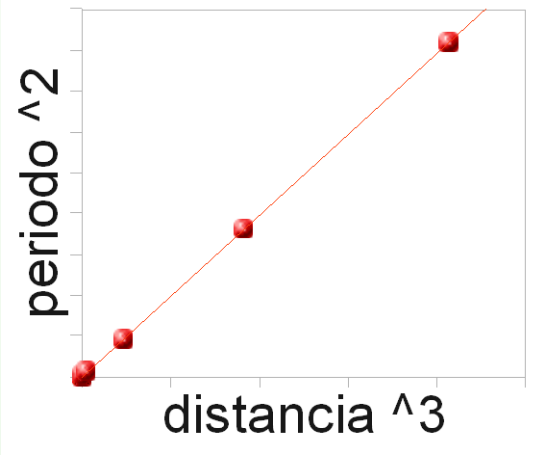
- Kepler modifica radicalmente su modelo y propone otro basado en tres leyes …

1. Los planetas se mueven siguiendo órbitas elípticas, con el sol en uno de los focos de la elipse

2. En sus órbitas alrededor del sol, los planetas barren áreas iguales en tiempos iguales

3. Los cuadrados de los tiempos necesarios para describir una órbita son proporcionales al cubo de las distancias medias al sol





Valores para el sistema solar

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ASTRONOMÍA Y DATAMINING HOY

Un ejemplo : los relevamientos del cielo

Instrumentos basados en tierra y sensores en el espacio recolectan datos

Se obtienen gran cantidad de datos multidimensionales, repetidos en el tiempo

Objetivos:

- explotar los datos (nuevos descubrimientos)
- establecer vínculos entre la teoría astrofísica y los datos

Los datos se ordenan de acuerdo a sus coordenadas, al momento de su captura y, si es posible, se registra el objeto involucrado.

La mayoría de los datos son mediciones en una banda del espectro electromagnético

EL ESPECTRO ELECTROMAGNETICO. 1

Cualquier objeto, independiente de su composición, tamaño o ubicación emite o absorbe energía.

Al emitir energía un objeto puede estar generándola (una estrella) o reflejando la energía que recibe de otra fuente (la luna).

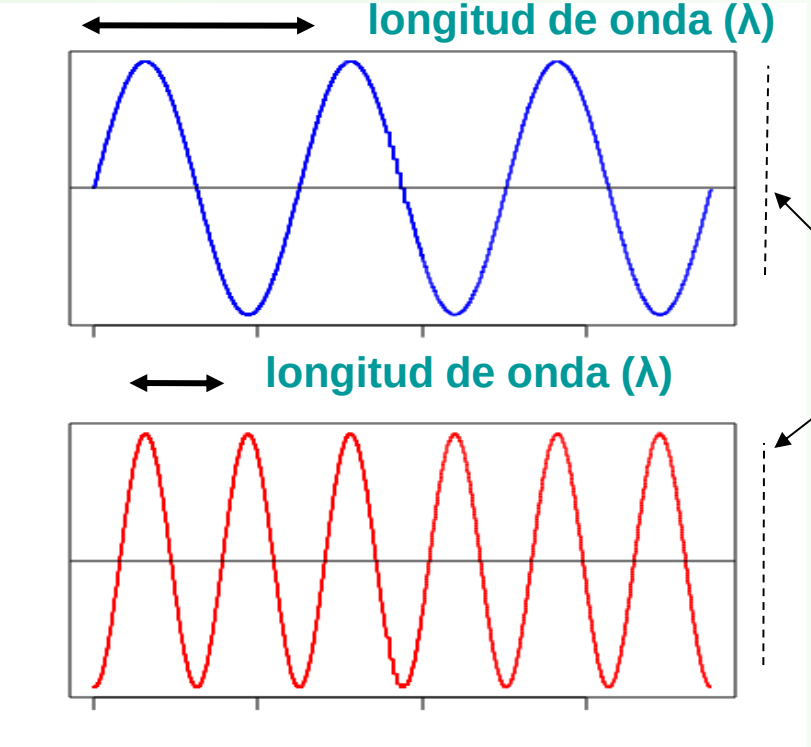
Los procesos físico-químicos que ocurren en el objeto, su composición y otros factores determinan el "tipo" y cantidad de energía que éste emite.

EL ESPECTRO ELECTROMAGNETICO. 1

Llamamos radiación electromagnética a las ondas de energía que emite un objeto. Existe una relación entre la energía y la frecuencia de las ondas que conforman la radiación.

Con los instrumentos adecuados es posible medir la radiación. Por ejemplo, nuestros ojos son instrumentos que detectan radiación en el rango de la luz visible.

ONDAS Y ENERGÍA. 1



La relación entre la longitud de una onda y su frecuencia es inversamente proporcional.

Para ondas electromagnéticas es:

frecuencia (f) = c / λ

c: velocidad de la luz en el vacío

La energía de una onda se puede determinar a partir de su longitud o frecuencia:

Energía (E) = h.c/λ o E= f.h

h: constante de Planck

Unidad de frecuencia: Hertz (Hz), ciclos por segundo

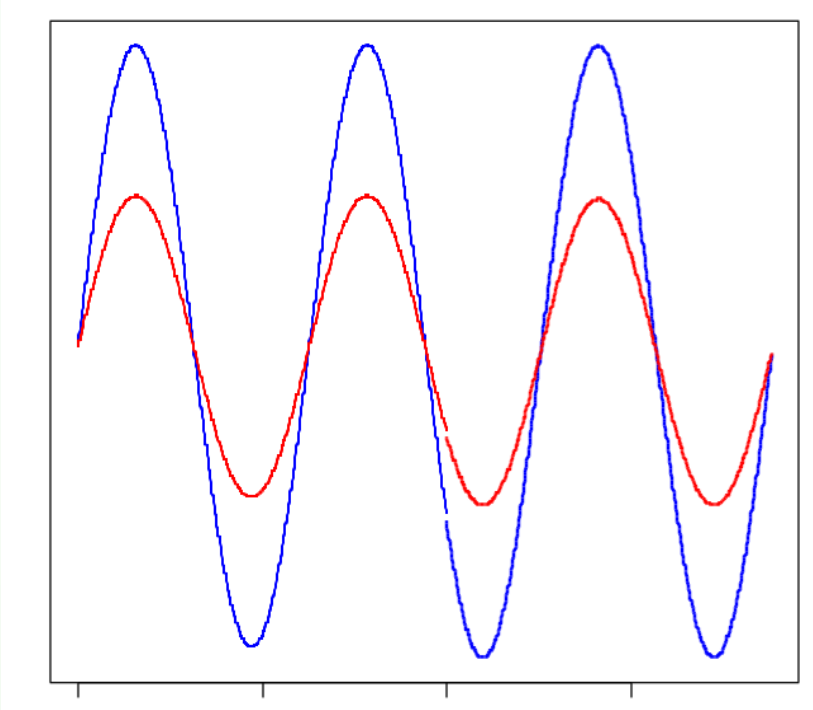
Unidad de longitud de onda: nm (nanometro, 10^{-9} m)

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ONDAS Y ENERGÍA. 2

Dos ondas con la misma longitud y diferente amplitud: la onda azul transporta mayor cantidad de energía que la roja.

Alternativamente se podría interpretar este fenómeno como dos flujos diferentes de fotones



Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

Dos conceptos importantes con respecto a la energía emitida y la observada:

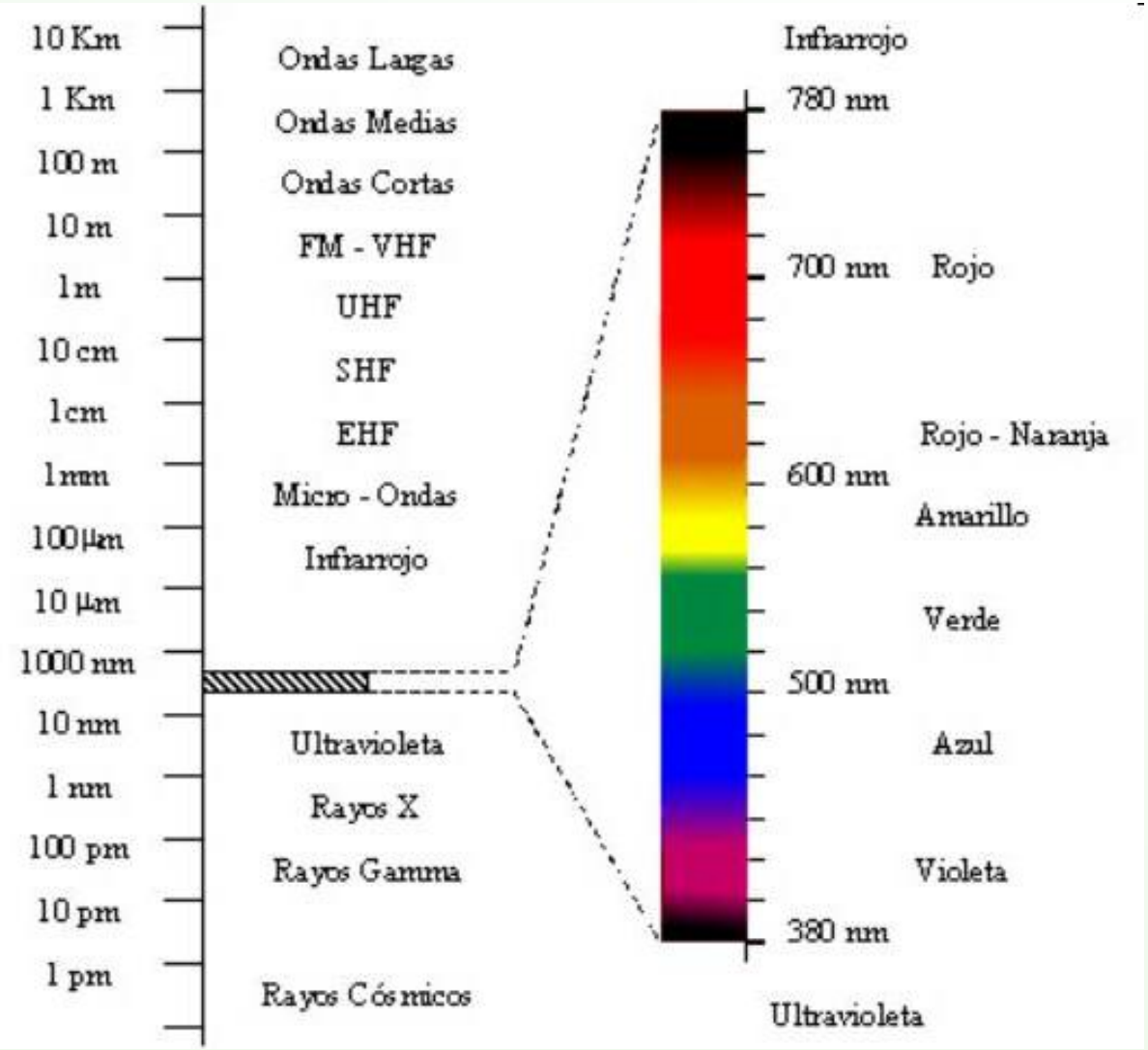
Intensidad: es la cantidad de energía de una determinada longitud de onda, o rango de longitudes de onda, emitida por unidad de tiempo y por unidad de superficie (no importa la posición del observador)

Flujo: es la cantidad de energía de una determinada longitud de onda, o rango de longitudes de onda, que pasa por un área por unidad de tiempo (depende de la posición del observador)

EL ESPECTRO ELECTROMAGNETICO

zona de baja energía – longitudes de onda largas

El **espectro electromagnético** es el rango completo de frecuencias que cubren las radiaciones electromagnéticas



zona de alta energía – longitudes de onda corta

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

Longitudes de onda analizadas por algunos relevamientos

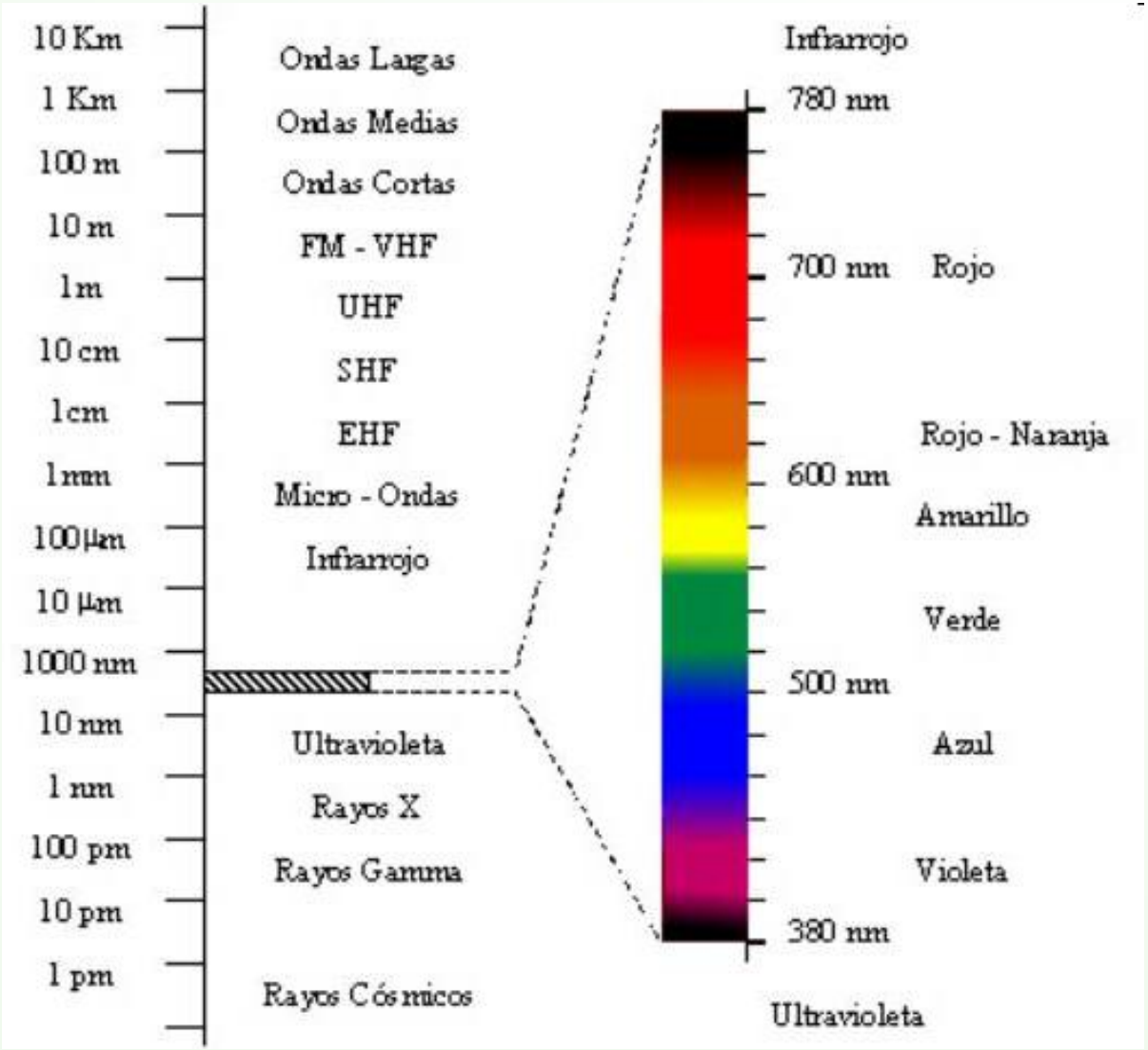
FIRST (Faint Images of the Radio Sky at Twenty Centimeters)

COBE

IRAS, 2MASS

HST (telescopio espacial Hubble), SDSS

High-Energy Transient Explorer (HETE), Gamma ray burst Coordinates Network (GCN)



Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ASTRONOMÍA Y DATAMINING HOY

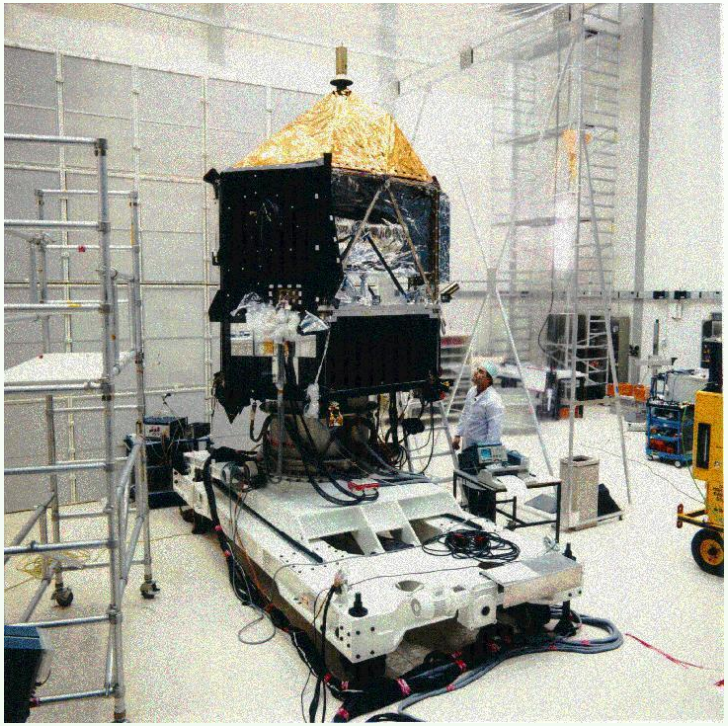
Hipparcos:

http://www.rssd.esa.int/index.php?project=HIPPARCOS

Plataforma espacial

En funcionamiento entre 1989 y 1993.

Posicionamiento de 118,218 estrellas con alta precisión y otro millón más con menor precisión.



ASTRONOMÍA Y DATAMINING HOY

Two-Micron All Scan Survey, 2MASS:
http://www.ipac.caltech.edu/2mass/

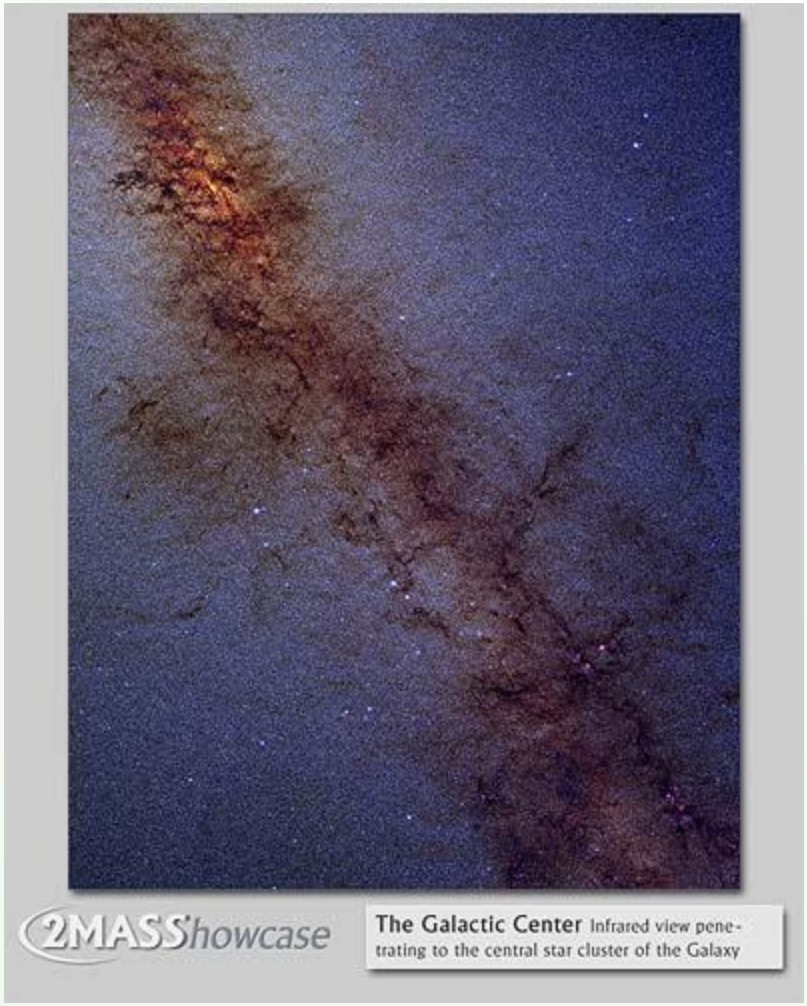
Tres bandas en el infra-rojo.

Catálogo con mas de 300 millones de estrellas y otros objetos puntuales.

Catalogo extendido con más de 1,000,000 de galaxias y nebulosas.

Atlas de 4 millones de imágenes.

En operación entre 1997 y 2001.



Crédito: 2MASS/G. Kopan, R. Hurt

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ASTRONOMÍA Y DATAMINING HOY

Sloan Digital Sky Survey, SDSS:
http://www.sdss.org/

Imágenes en el espectro visible.

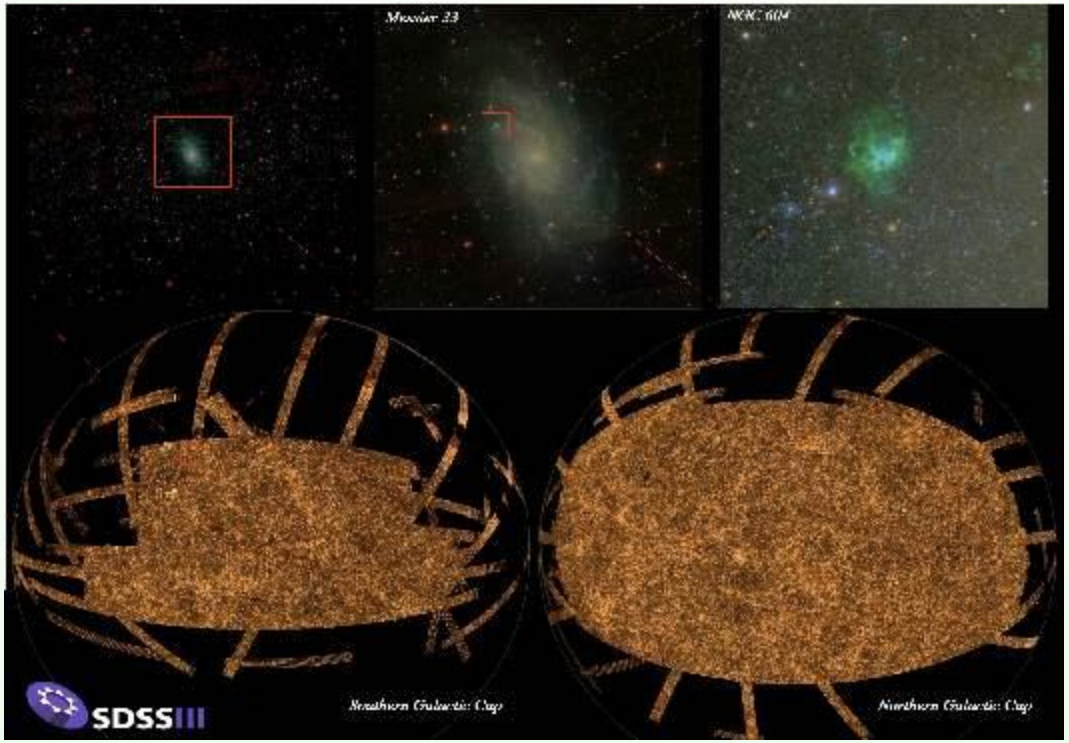
Mapas 3D de más de un millón de objetos.

Mediciones de 500 millones de estrellas y galaxias

Espectros de casi 2 millones de objetos.

Cámara de 120 megapixels acoplada a espectrómetros.

La fase 1 comenzó en 2000, la fase 3 está en operaciones.



Crédito: M. Blanton and the SDSS-III collaboration

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ALGUNOS PROYECTOS RECIENTES O FUTUROS

Reconocimientos astronómicos masivos. PanSTARRS
(Panoramic Survey Telescope And Rapid Response System, http://pan-starrs.ifa.hawaii.edu)

- **Objetivo inicial:** detección temprana de objetos que se aproximan a la Tierra. **Otros objetivos:** censo del cielo, análisis de la materia y la energía oscuras.
- La primera cámara inició la misión científica en mayo de 2010.
- Puede mapear un sexto del cielo cada mes a cinco longitudes de onda diferentes.
- 4 cámaras de amplio campo de 1.4 Gigapixels
- **10 Terabytes por noche. Volumen final anticipado: 40 Petabytes**

PanSTARRS. Procesamiento de datos

- En cada imagen individual (1.4 Gigapixels) se buscan estrellas conocidas para realizar la calibración espacial y fotométrica.
- Las cuatro imágenes obtenidas (una por cámara) se comparan entre si para buscar y eliminar aberraciones y defectos.
- Se crea una única imagen compuesta sin defectos y corregida.

Y ademas…

PanSTARRS. Procesamiento de datos

- Se compara la imagen del paso anterior contra un master compuesto de todas las imágenes anteriores para buscar objetos móviles o de brillo variable.
- La magnitud y la posición de todos los objetos que superen un umbral de brillo se extraen y almacenan en una base de datos
- Eliminar las imágenes individuales y la compuesta originadas en los pasos 1 y 2

Todo esto en aproximadamente un minuto !!!

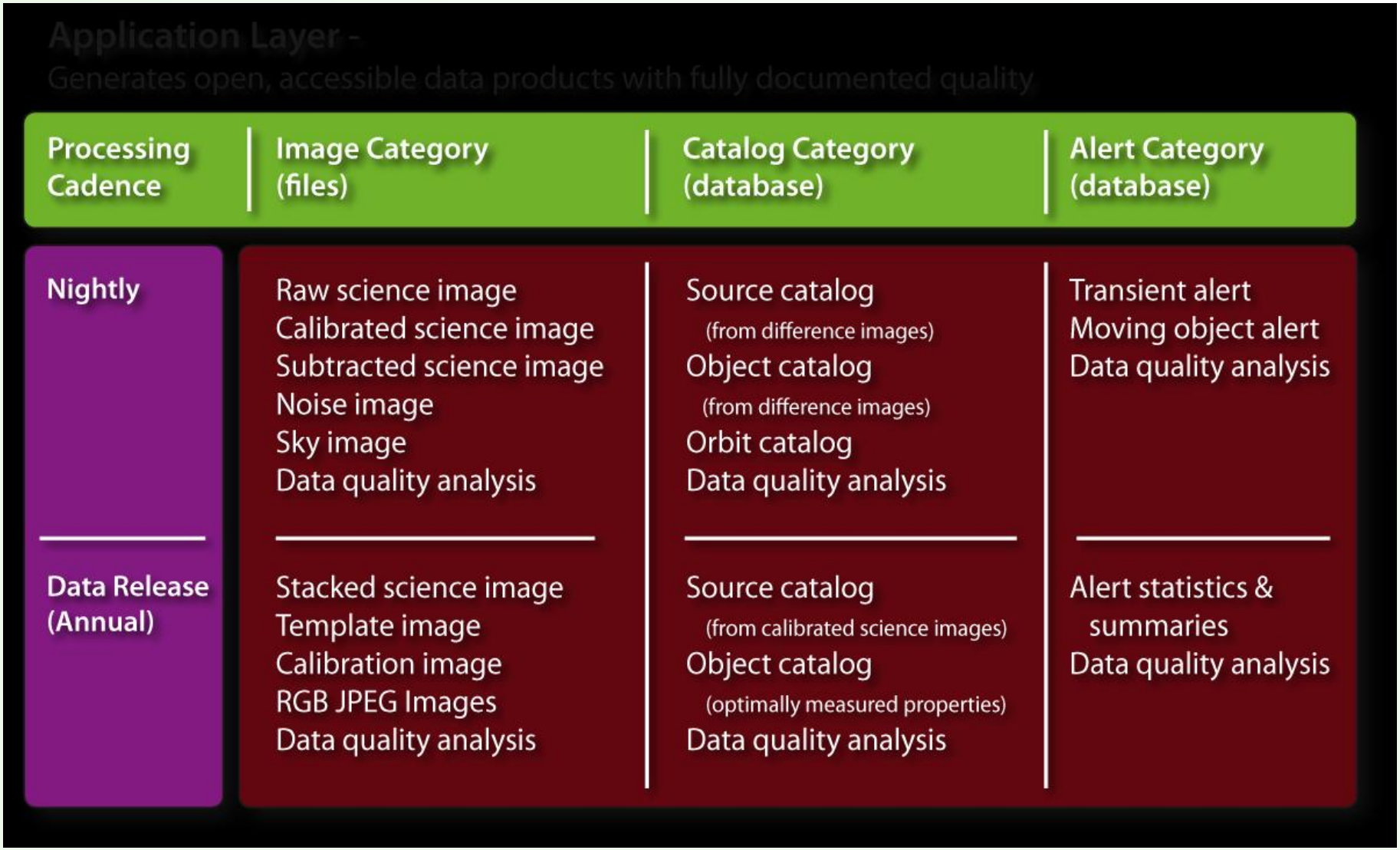
Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

ALGUNOS PROYECTOS RECIENTES O FUTUROS

Reconocimientos astronómicos masivos. LSST (Large Synoptic Survey Telescope, http://www.lsst.org/lsst/)

- Objetivos: detección de objetos variables o móviles. investigación de materia y energía oscuras
- Cámara de 3 Gigapixels (6 GB / imagen)
- Capturará 30 Terabytes por noche.
- Volumen anticipado de imágenes en 10 años: 60 Petabytes
- Catálogo anticipado de 30 Petabytes
- Minería de eventos en tiempo real: 10,000-100,000 alertas por noche
- Imágenes repetidas del cielo nocturno cada tres o cuatro noches.

LSST. Manejo de Datos



Fuente: LSST

Marcelo A. Soria (soria@agro.uba.ar)
Maestria en Data Mining. DC-FCEN

LSST. Estimación de las dimensiones de la base de datos

- Más de 100 tablas
- Metadatos de imágenes: 670 millones de registros
- Catálogo de fuentes: 260,000 millones de registros
- Catálogo de objetos: 22,000 millones de registros, +200 atributos
- Catálogo de objetos móviles: 10 millones de registros
- Catálogo de objetos variables: 100 millones de registros
- Catálogo de alertas: alertas cada 60 segundos

OBJETIVOS DE DATAMINING EN ASTRONOMIA

- **Caracterizar los objetos conocidos**: métodos de aprendizaje no supervisados, clustering.
- **Asignar los nuevos objetos a clases ya existentes**: métodos de aprendizaje supervisados, clasificación.
- **Descubrir objetos de clases desconocida**: aprendizaje semi-supervisado, detección de outliers.
- **Desarrollar nuevos algoritmos, adaptar los existentes**

DOS CATEGORIAS DE DATAMINING

Mining basado en eventos

- Eventos conocidos – algoritmos conocidos
- Eventos conocidos – algoritmos desconocidos
- Eventos desconocidos-algoritmos conocidos
- Eventos desconocidos-algoritmos desconocidos

Mining basado en relaciones

- Asociaciones espaciales
- Asociaciones temporales
- Asociaciones de coincidencia en un espacio multidimensional

Algunos conceptos útiles para el trabajo práctico

Declinación y ascención recta: son las coordenadas astronomicas para mapear los objetos en el espacio. La declinación es comparable a la latitud y se mide en grados, minutos y segundos, y su rango varía entre +90 y -90 . La ascención recta es comparable a la longitud, y se mide en horas, minutos y segundos.

Movimiento propio e impropio: los objetos en el espacio se mueven en relación a la Tierra. Parte de este movimiento se debe al movimiento del planeta con respecto al Sol, pero otra parte, el movimiento propio, ocurre porque los objetos celestes no están en posiciones fijas.

Algunos conceptos útiles para el trabajo práctico

Paralaje: si determinamos la posición de una estrella en un momento dado y seis meses después, vamos a estar haciendo observaciones desde puntos opuestos de la órbita terrestre alrededor del Sol. Esto es útil para medir la distancia a otras estrellas en nuestra galaxia. Cuanto mayor sea el ángulo, mayor el paralaje y más cercana estará la estrella.

Parsec: es una medida de distancia que equivale a 3.26 años-luz. Es el lado de un triángulo determinado por un arcosegundo de ángulo de paralaje y un lado de 1 unidad astrónomica de largo (distancia al Sol)

Algunos conceptos útiles para el trabajo práctico

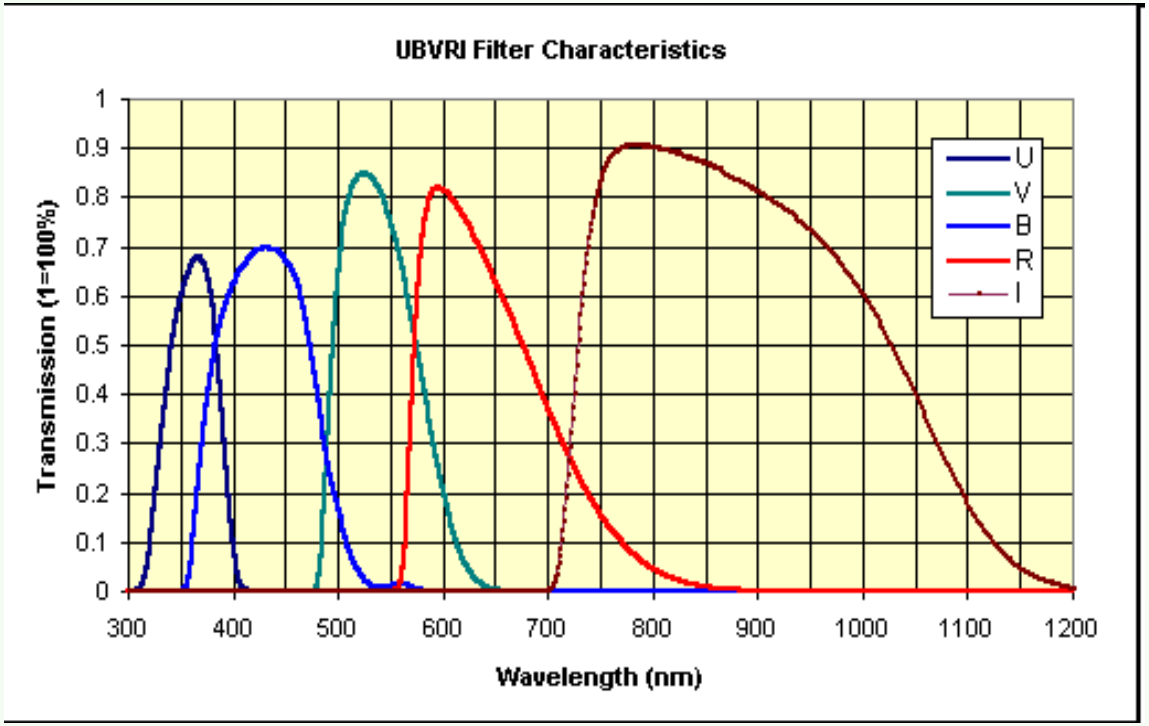
Filtros: los objetos en el espacio emiten luz de diferentes longitudes de onda.

Los filtros “cortan” diferentes bandas del espectro electromagnético al incluirlos en la óptica de los telescopios.

Actualmente existen equipos que barren con mayor resolución un rango de longitudes de onda. Para mantener medidas comparables se pueden convertir estas medidas a las equivalentes de los filtros.

Algunos conceptos útiles para el trabajo práctico

La familia de filtros UVB de Johnson es un conjunto de tres filtros, U (ultravioleta), B (azul), V (visible)



Características de transmisión de los filtros UBV y RI

http://www.lancs.ac.uk/users/spc/resources/observatory/specs.htm

Algunos conceptos útiles para el trabajo práctico

El filtro V se usa para medir el brillo de las estrellas. Este se expresa como magnitud. La magnitud en una banda cualquiera x es una función del logaritmo del flujo (brillo aparente) relativo a una referencia:

$$m_x = -2.5 \log_{10} (F_x / F_x^0)$$

La magnitud aparente de un cuerpo celeste (m) es la medida de su brillo como la ve un observador desde la Tierra, y corregida para eliminar el efecto de la atmosfera. Cuanto mayor es el brillo del cuerpo, menor es su valor de m.

La magnitud absoluta es la magnitud aparente corregida por la distancia al objeto

Algunos conceptos útiles para el trabajo práctico

El filtro azul (B) del sistema de Johnson es útil para estimar la temperatura superficial de una estrella. Las estrellas más calientes dan más luz en el azul que en el rojo, y las más frías tienden a dar más luz en el rojo.

Se mide la magnitud con el filtro B al que se le resta la medición hecha con el filtro V. El resultado se conoce como índice B-V. Cuanto mayor es la temperatura superficial de una estrella, menor es el valor correspondiente del índice B-V

El diagrama de Herzsprung-Russell relaciona la magnitud absoluta de las estrellas con su temperatura estimada por el índice B-V