Data Mining: Introducción

En lo que sigue tomamos como referencia:

Introduction to Data Mining, 2nd Edition

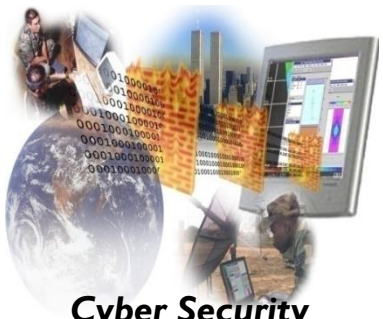
Tan, Steinbach, Karpatne, Kumar - 2018

Gran cantidad de datos y multiples aplicaciones

- Enorme crecimiento en la cantidad de datos tanto en bases de datos comerciales como científicas, debido principalmente a los avances en las tecnologías de generación y recolección de datos

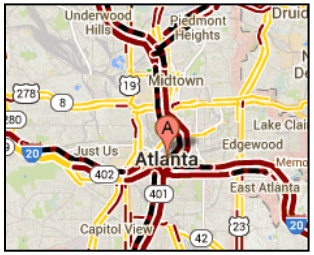
- *Nuevo mantra:*
  - Reúna todos los datos que pueda, cuando pueda y donde pueda.

- Expectativas
  - Los datos reunidos tienen valor ya sea para el propósito para el que fueron recolectados como para otros propósitos no previstos.

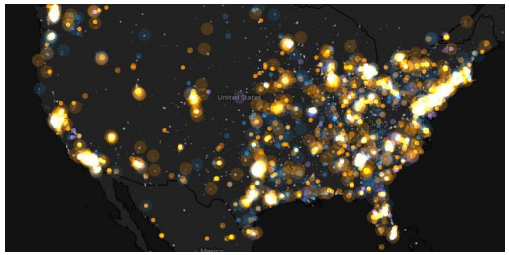


Cyber Security

E-Commerce

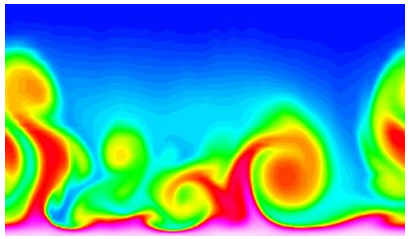


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

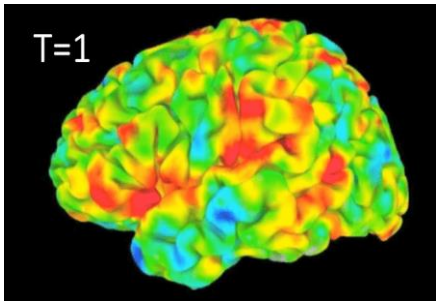
¿Por que Data Mining? El punto de Vista Comercial…

- Se reúne y almacena gran cantidad de datos
  - Web data
    - Yahoo tiene Peta Bytes de datos web
    - Facebook tiene billiones de usuarios activos
  - Compras en tiendas electrónicas
    - Amazon administra millones de visitas diarias
  - Transacciones Bancarias/Tarjetas de Crédito

- Las computadoras son cada vez menos costosas y mas poderosas
- La presión competitiva es muy fuerte
  - Proveer mejores servicios, servicios customizados... (e.g. Customer Relationship Management - *CRM*)

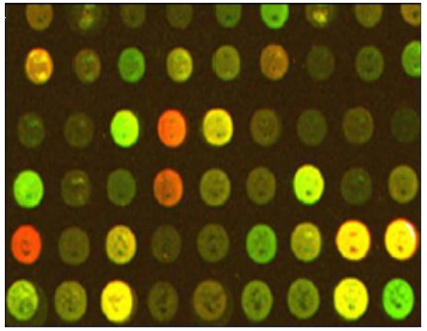
¿Por que Data Mining? El punto de vista científico

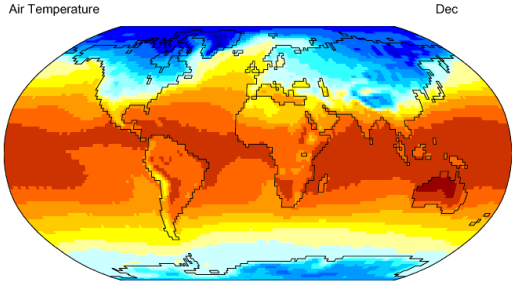
- Los datos se reúnen y almacenan a velocidades gigantescas
  - Sensores remotos en satélites
    - El archive de la NASA EOSDIS almacena mas de un Petabytes de datos sobre la tierra por año
  - Telescopios que observan los cielos
  - Datos Biológicos
  - Simulaciones Científicas
    - terabytes de datos generados en pocas horas
- La minería de datos ayuda a los científicos
  - Con análisis automatizados en datasets masivos
  - En la formulación de hipótesis



fMRI Data from Brain


Sky Survey Data


Gene Expression Data

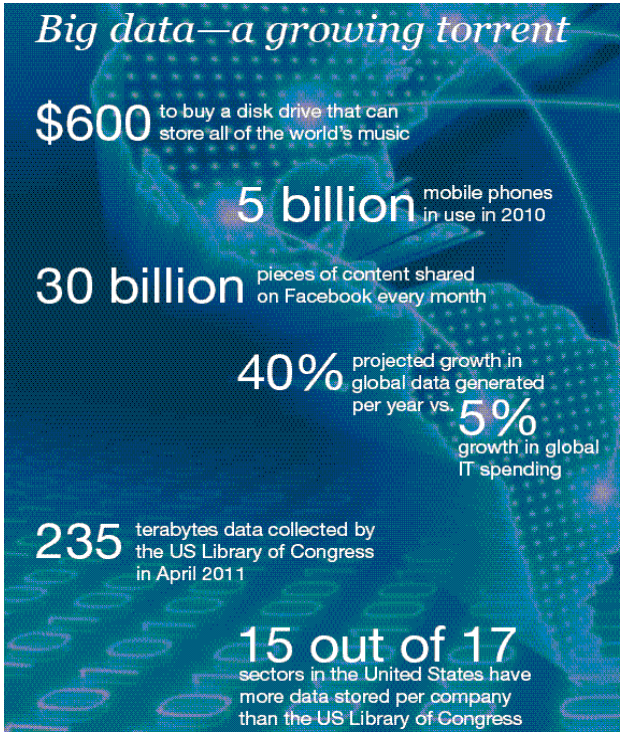


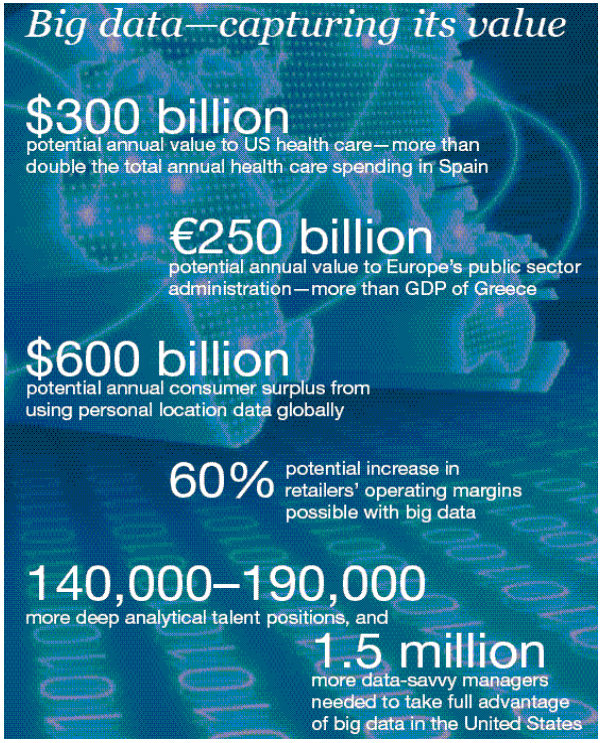
Surface Temperature of Earth

Oportunidad para mejorar la productividad…

McKinsey Global Institute

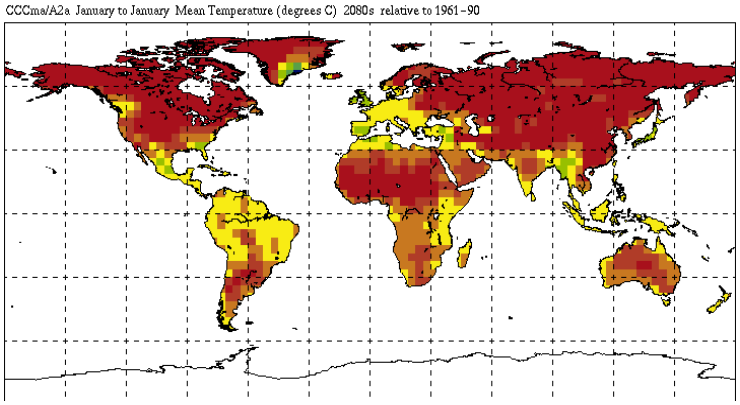
Big data: The next frontier for innovation, competition, and productivity





Oportunidad para Resolver Grandes Problemas

Sistema de Salud y Costos



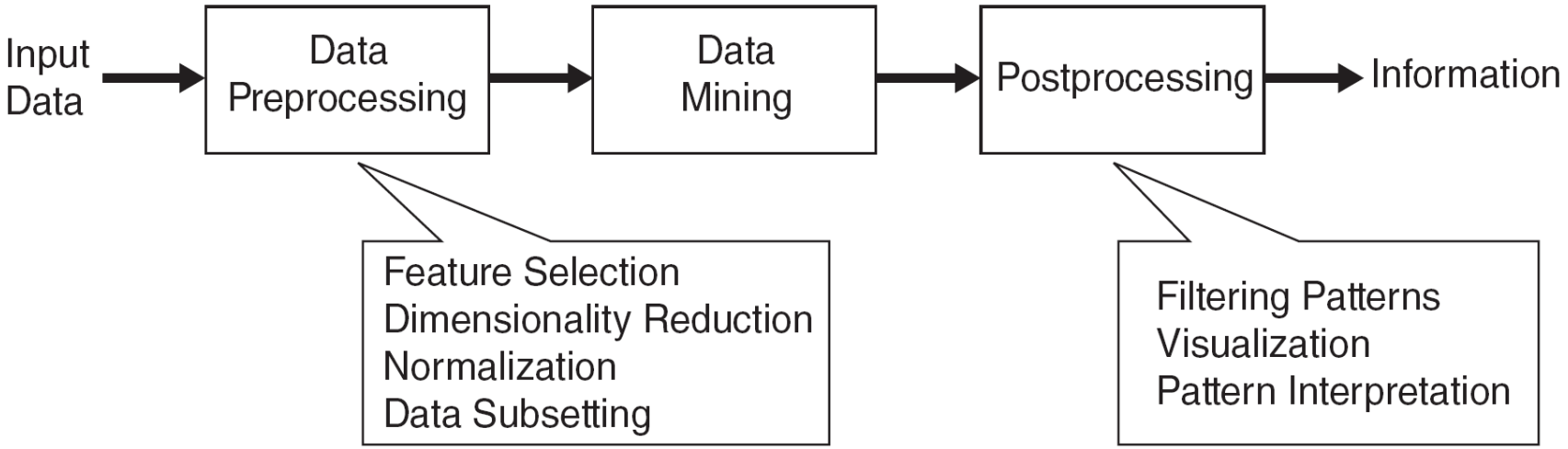
Predecir el Impacto del Cambio Climatico

Fuentes de energía alternativas / verdes

Reduccion del Hambre y la Pobreza

¿Que es “Data Mining”?

- Múltiples Definiciones
  - Extracción NO-trivial de información implícita, previamente desconocida y potencialmente útil a partir de datos.
  - Exploración y Análisis empleando medios automáticos o semi-automaticos, en grandes cantidades de datos con la finalidad de descubrir patrones significativos.



¿Que NO es Data Mining?

¿Que NO es Data Mining?

- Buscar un numero de teléfono en la guía
- Consultar un Buscador Web por información sobre "Amazon"

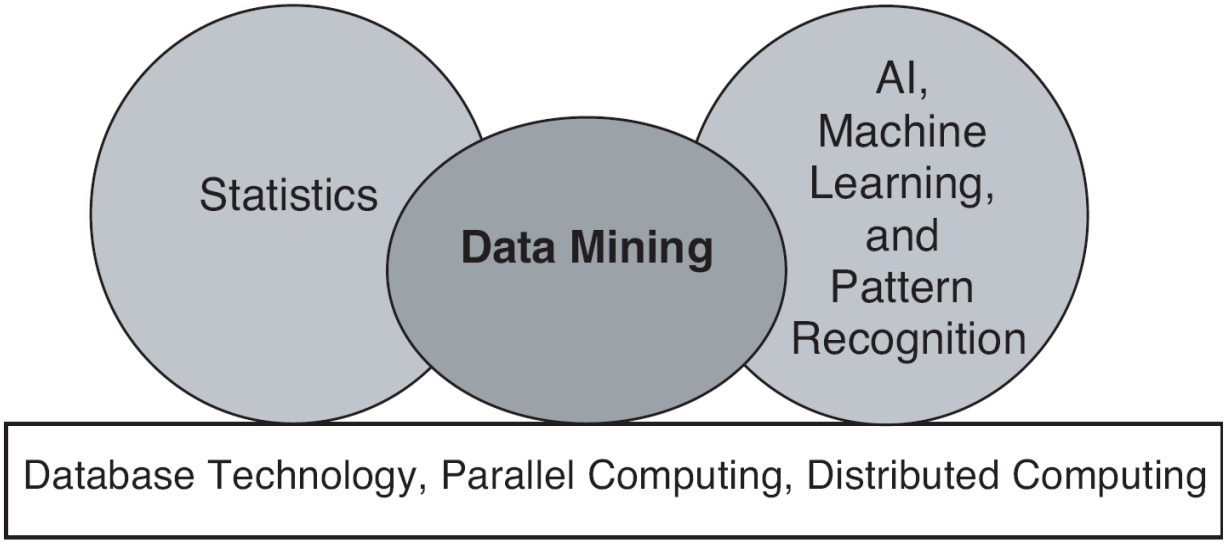
¿Que es Data Mining?

- Ciertos nombres son mas prevalente en ciertos lugares de los EEUU (O'Brien, O'Rourke, O'Reilly… en el Área Metropolitana de Boston)
- Agrupar documentos similares devueltos por un motor de búsqueda de acuerdo a su contexto (e.g., Amazon rainforest, Amazon.com)

Origenes de Mineria de Datos

- Toma ideas del Aprendizaje Automático, y de la IA (*Machine Learning/AI*), reconocimiento de patrones, estadística y sistemas de gestión de bases de datos.

- Las técnicas tradicionales pueden no ser apropiadas:
  - Gran Escala
  - Alta Dimensionalidad
  - Heterogeneidad
  - Complejidad
  - Datos Distribuidos



- Es un componente clave del campo emergente de la ciencia de datos y del descubrimiento guiado por los datos.

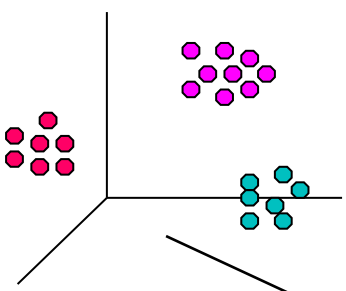
Tareas de Minería de Datos

- *Métodos Predictivos*
  - Utilizar algunas variables para predecir valores desconocidos o futuros de otras variables.

- *Métodos Descriptivos*
  - Encontrar patrones entendibles por los seres humanos para describir los datos.

[Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Tareas de Minería de datos...

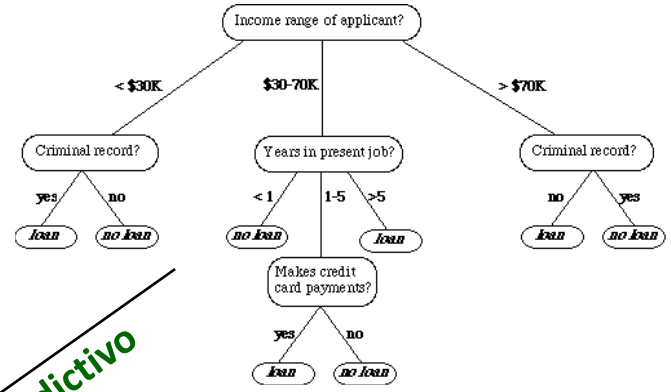


Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Modelado Predictivo



Reglas de Asociacion

Deteccion de Anomalias

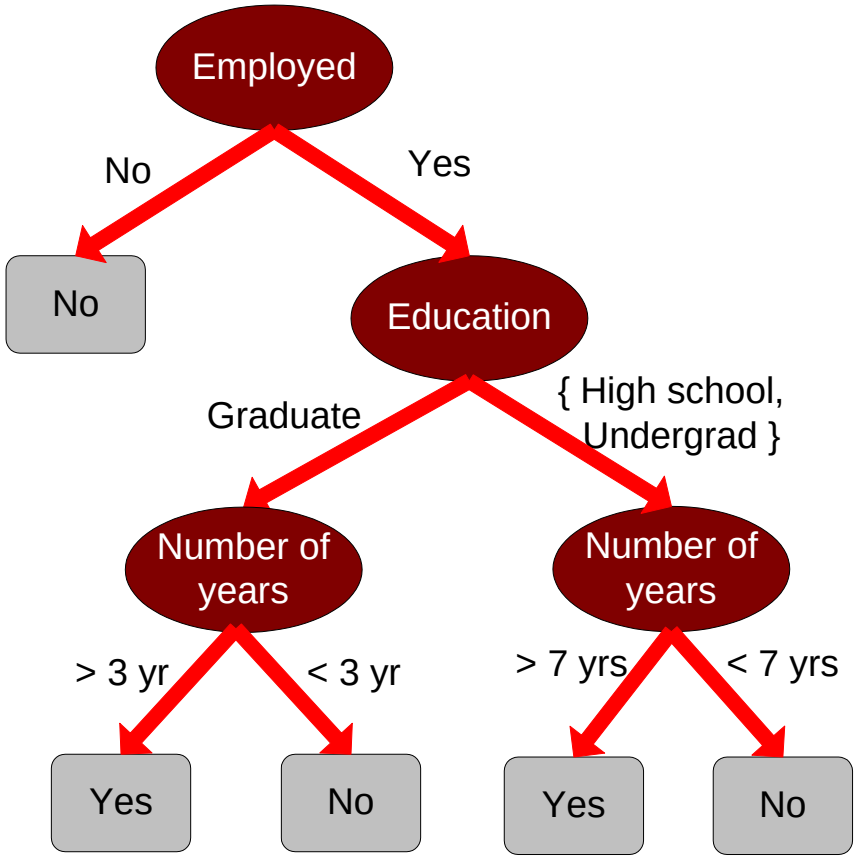
Modelado Predictivo: Clasificación

- Encontrar un modelo para un atributo de clase como función de los valores de los otros atributos

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Modelo para predecir el otorgamiento de un credito

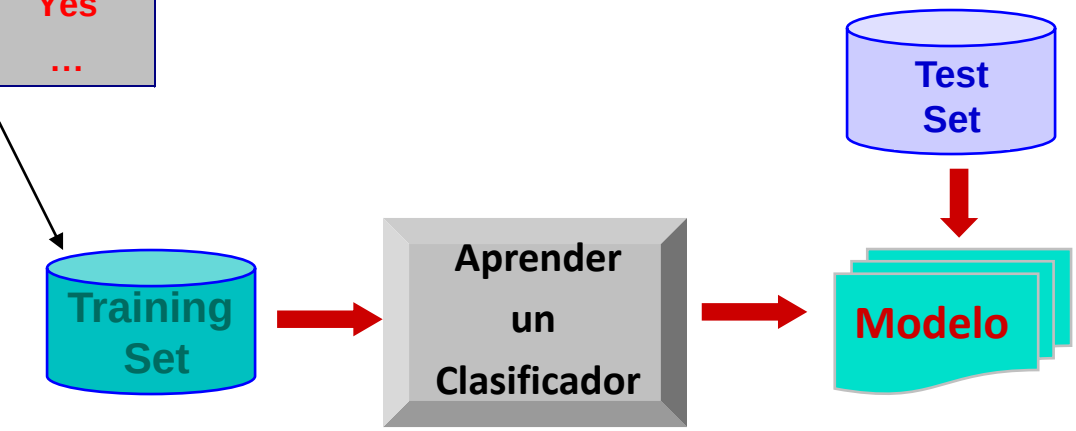


Ejemplo de Clasificacion

categorica | categorica | quantitativa | clase

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

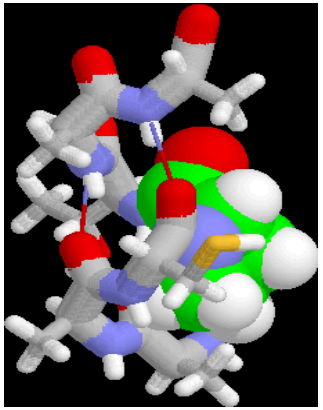


Ejemplos de Tareas de Clasificación

- Clasificar transacciones de tarjeta de crédito como fraudulentas o no.

- Clasificar la cobertura del suelo (cuerpos de agua, áreas urbanas, bosques, etc.) empleando imágenes satelitarias.

- Categorización de las noticias por tipo (deportivas, financieras, clima, entretenimiento, deportes, etc.)
- Identificar intrusos en el ciberespacio
- Predecir células tumorales como benignas o malignas



- Clasificar estructuras secundarias de las proteínas, como alpha-helix, beta-sheet, o random coil

Clasificación: Ejemplo de Aplicacion I

- *Detección de Fraudes*

  - **Objetivo:** Predecir casos fraudulentos en transacciones de tarjeta de crédito.
  - **Aproximación:**
    - Utilizar las transacciones de tarjeta de crédito y la información sobre los titulares como atributos.
      - Cuando compra, que compra, cuan seguido paga en termino, etc.
    - Etiquetar las transacciones pasadas como fraudulentas o no. Esto forma el atributo de clase.
    - Aprender un modelo para la clase de transacciones.
    - Emplear este modelo para detectar fraudes observando las transacciones con la tarjeta de crédito en la cuenta.

Clasificación: Ejemplo de Aplicación 2

- *Predicción del "Churn" (salida) de los clientes de una compañía telefónica*

  - **Objetivo:** Predecir cuando un cliente es probable que se pierda en otro competidor.
  - **Aproximación:**
    - Utilizar los registros detallados de transacciones con cada uno de los clientes pasados y presentes para encontrar atributos.
      - Cuan seguido realizan las llamadas, donde llaman, a que hora del día llama con mas asiduidad, su estado financiero, su estado civil, etc.
    - Etiquetar los clientes como leales o desleales.
    - Encontrar un modelo para la lealtad.

[Berry & Linoff] Data Mining Techniques, 1997

Clasificación: Ejemplo de Aplicación 3

- *Catalogo de Observación del Cielo*

  - **Objetivo:** predecir la clase de objetos celestes (estrellas o galaxias), especialmente los objetos visualmente tenues, basándose en el relevamiento de imágenes telescópicas (del Observatorio Palomar).
    - 3000 imágenes con 23,040 x 23,040 pixeles por imagen.
  - **Aproximación:**
    - Segmentar la imagen.
    - Medir los atributos de la imagen (*features*) - 40 de ellos por objeto.
    - Modelar la clase basándose en estos *features* (características).
    - Historia de Éxito: Se encontraron 16 nuevos red-shift *quasars*, que son algunos de los objetos mas lejanos y difíciles de detectar!!

[Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

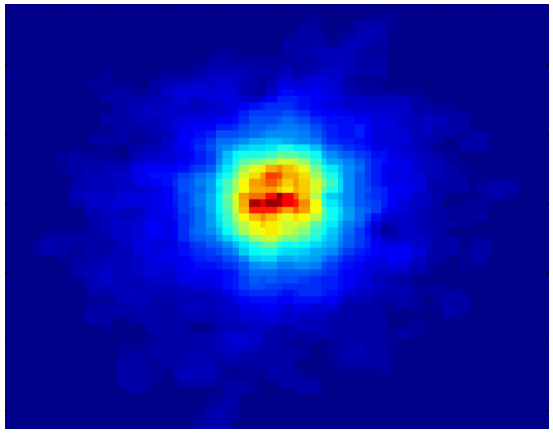
Clasificando Galaxias

http://aps.umn.edu

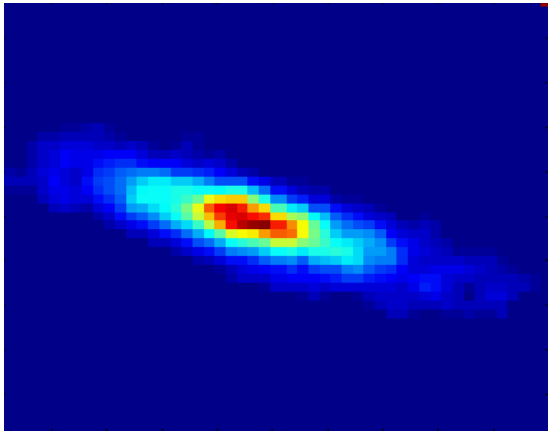
Clase:
- Etapas de Formación

Atributos:
- Image features,
- Characteristics of light waves received, etc.

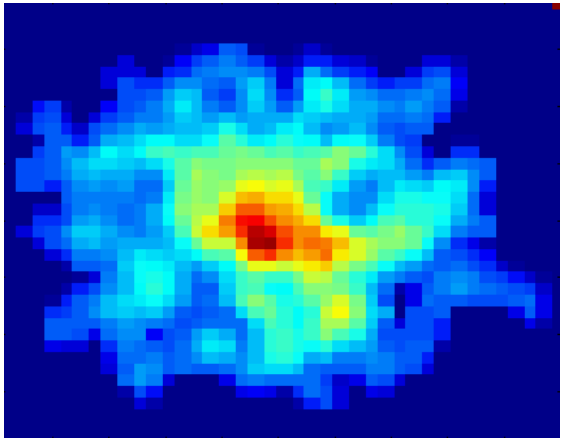
Early



Intermediate



Late



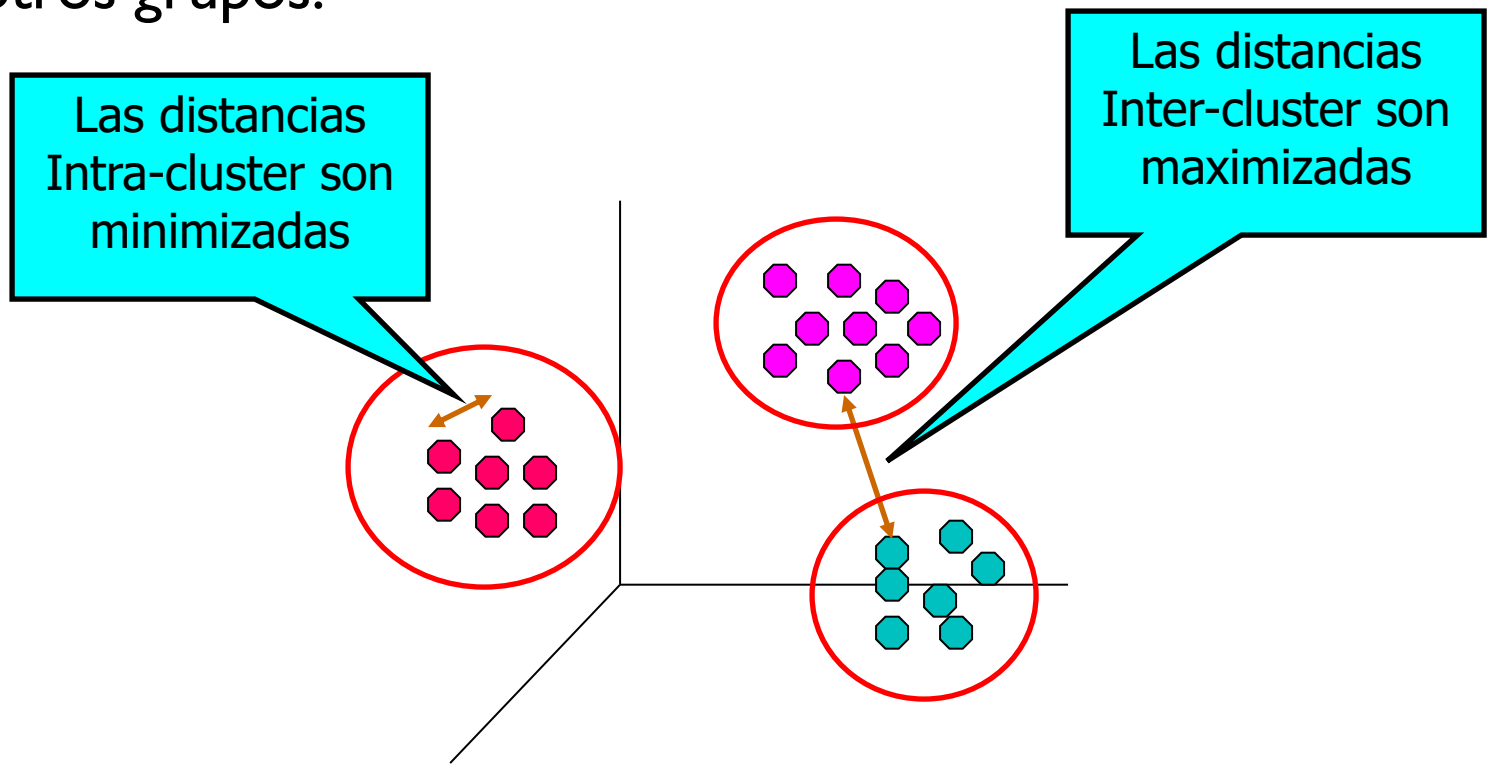
Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Regresión

- Predecir el valor de una variable continua basándose en los valores de otras variables, asumiendo un modelo lineal o no lineal de dependencia.
- Muy utilizada en estadística en general, y en redes neuronales.
- Ejemplos:
  - Predecir la cantidad de ventas de un nuevo producto basándose en los gastos de publicidad.
  - Predecir la velocidad del viento como función de la temperatura, humedad, presión del aire, etc.
  - Predicción de la serie de tiempo de los índices de stock de los mercados.

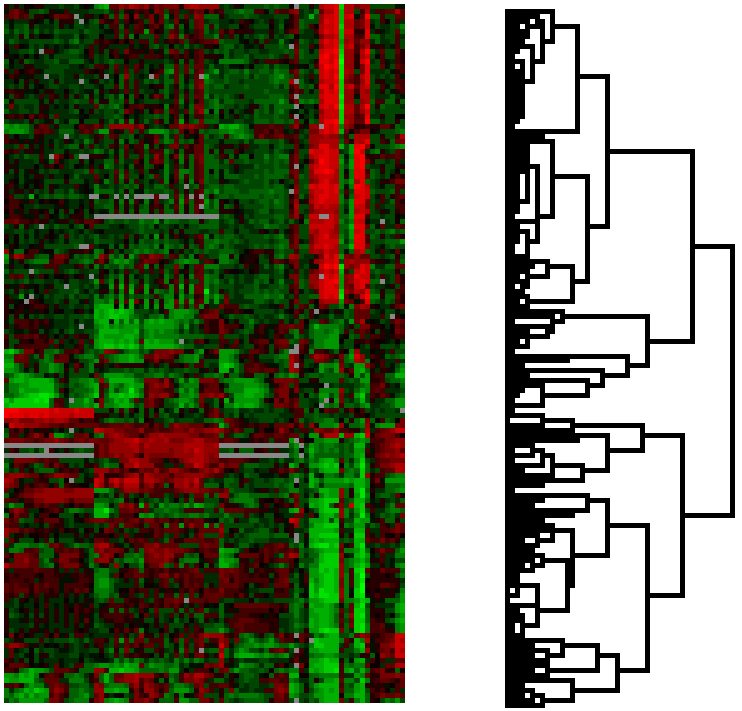
Clustering (conglomeración)

- Encontrar grupos de objetos tales que cada objeto en el grupo sea similar (o relacionado) a los demás del grupo y diferente de (o no relacionado) con los objetos en los otros grupos.

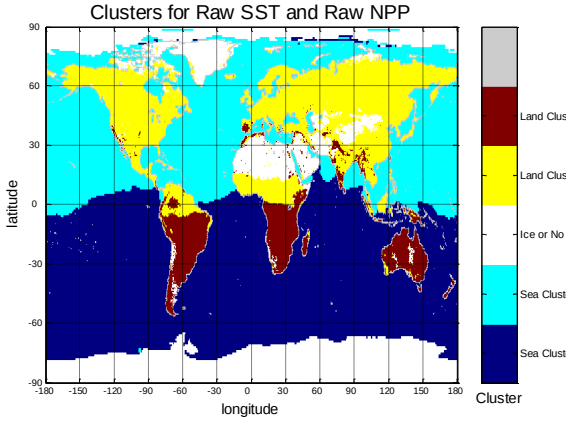


Aplicaciones del Análisis de Conglomeración

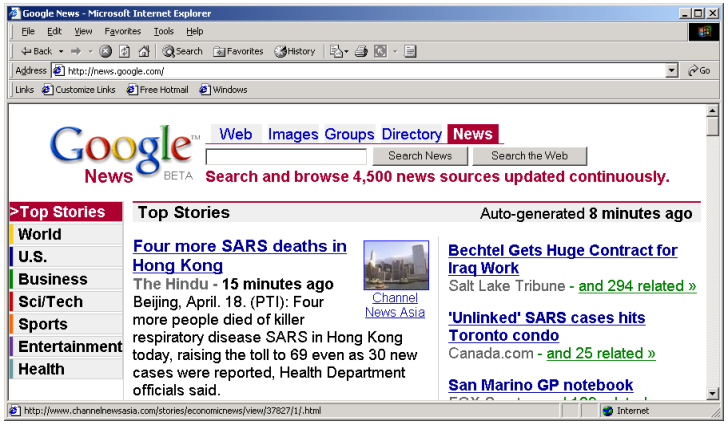
- **Entender**
  - Perfilado customizado para marketing focalizado
  - Agrupar documentos relacionados para navegar
  - Agrupar genes y proteínas que tienen funcionalidad similar
  - Agrupar acciones con fluctuaciones similares de precio
- **Sumarizar**
  - Reducir el tamaño de grandes conjuntos de datos



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



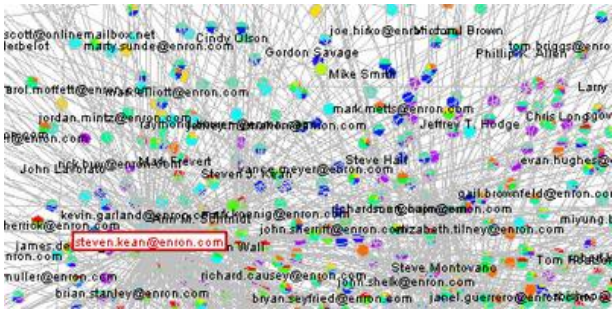
Clustering: Ejemplo de Aplicación I

- *Segmentación de Mercado:*
  - **Objetivo:** subdividir el Mercado en diferentes subconjuntos de clientes donde cualquier subconjunto puede ser elegido como objetivo para realizar una mezcla de marketing particular.
  - **Aproximación:**
    - Reunir diferentes atributos de los clientes basándose en su información geográfica y de estilo de vida.
    - Encontrar conglomerados de clientes similares
    - Medir la calidad del conglomerado observando los patrones de compra de los clientes dentro del mismo conglomerado vs. aquellos en diferentes conglomerados.

Clustering: Ejemplo de Aplicacion 2

- *Clustering de documentos*:
  - **Objetivo:** Encontrar grupos de documentos que son similares entre si basándonos en términos importantes que aparecen en ellos.
  - **Aproximación:** Identificar términos que ocurren frecuentemente en cada documento. Formar una medida de similitud basándose en las frecuencias de los diferentes términos. Emplearla para conglomerar.

Enron email dataset



Descubrimiento de Reglas de Asociación: Definición

- Dado un conjunto de registros, cada uno de los cuales contiene un numero de ítems de una dada colección.
  - Producir reglas de dependencia que predigan la ocurrencia de un ítem basándose en ocurrencias de otros ítems.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Reglas descubiertas

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Análisis de Asociación: Aplicaciones

- *Análisis de la Canasta de Mercado*
  - Las reglas se emplean para la promoción de ventas, manejo de inventario, reposición de productos en góndolas.

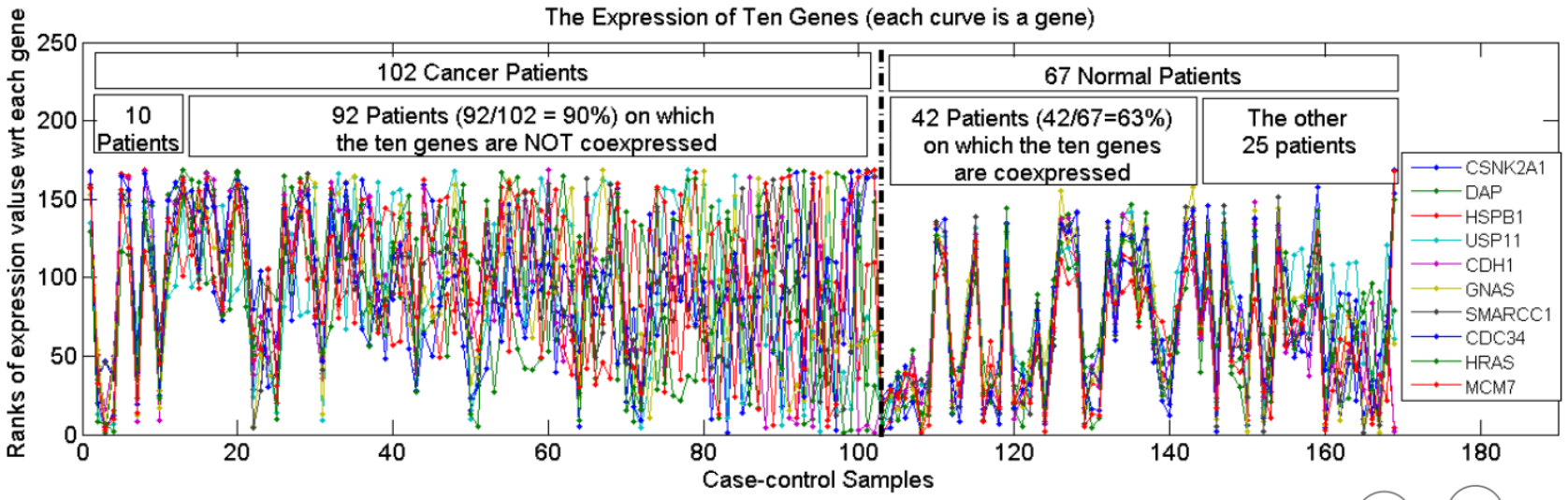
- *Diagnostico de alarmas en telecomunicaciones*
  - Las reglas se emplean para encontrar combinaciones de alarmas que co-ocurren frecuentemente en el mismo periodo de tiempo.

- *Informática Médica*
  - Las reglas se emplean para encontrar combinaciones de síntomas de los pacientes y resultados de los test asociados con ciertas enfermedades.

Analisis de Asociacion: Aplicaciones

- Ejemplo de un Subespacio Diferencial de Co-expresión para cáncer de pulmón

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]

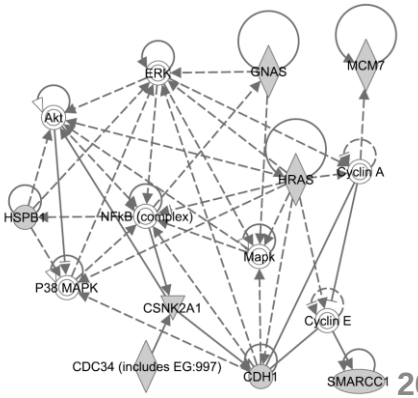


Enriched with the TNF/NFB signaling pathway

which is well-known to be related to lung cancer

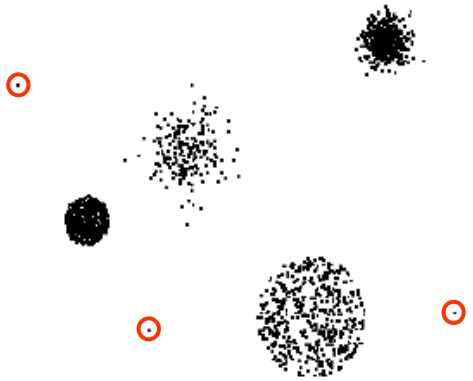
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

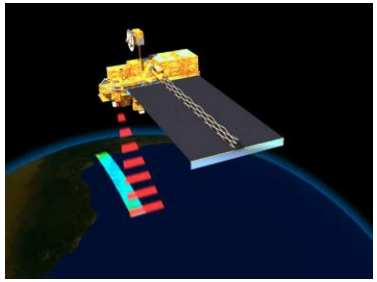
[Fang et al PSB 2010]

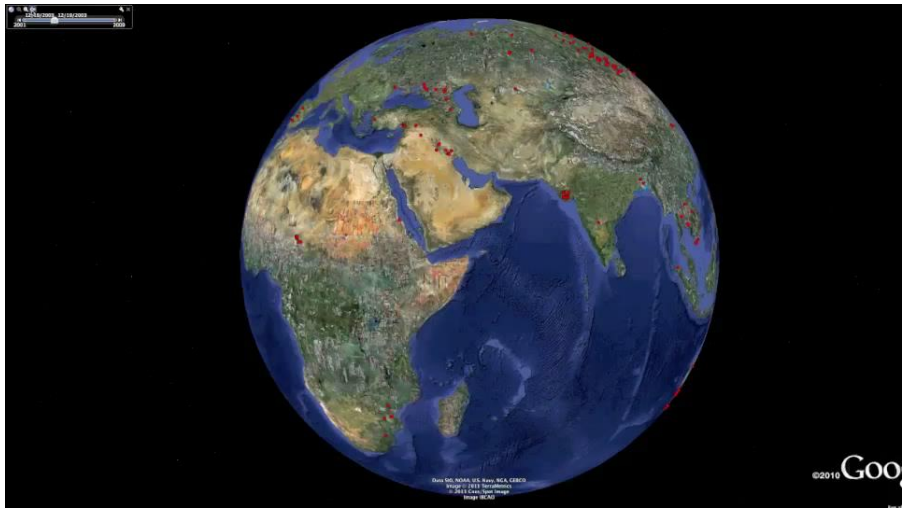


Desviaciones / Anomalías / Detección de Cambios

- Detectar desviaciones significativas de la conducta normal
- Aplicaciones:
  - Fraude en tarjetas de crédito
  - Intrusiones en Redes
  - Identificación de conductas anómalas a partir de sensores de vigilancia
  - Detección de cambios en la cobertura boscosa mundial.







Desafíos

- Escalabilidad
- Alta Dimensionalidad
- Datos Heterogéneos y Complejos
- Propiedad del Dato y Datos Distribuidos
- Análisis NO tradicional