El poder de la analítica de datos

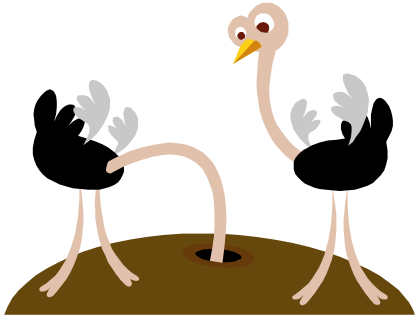
Jose Aguilar
Prometeo- Yachay EP

Contenido

1. Introducción a la Analítica de Datos y a la Ciencia de los Datos.
2. Metodologías para realizar Analítica de Datos en una organización
3. Tipos de tareas de Analítica de Datos.
4. Técnicas de Analítica de Datos
5. Algunos Conceptos Vecinos:
    1. Inteligencia de Negocios,
    2. Mineria (de datos, semántica, de texto, de Grafos),
    3. BigData.

Introducción a la Analítica de Datos y a la Ciencia de los Datos

ACTITUD hacia la vida



Actitud Pasiva

Actitud Preactiva

Actitud Reactiva

Actitud Proactiva

Ideas introductorias

Según Steve Haeckel

"**La inteligencia** de una empresa está determinado por la medida en que su **infraestructura informática conecta la información, la comparte y le da estructura**."

CONOCIMIENTO

"En los últimos 10 años se han producido más conocimientos que en los 10.000 años anteriores".

Bill Gates

Estamos en la Civilización del Conocimiento

La sociedad del Conocimiento

- **Ausencia de fronteras,** porque el conocimiento viaja aun con menos esfuerzo que el dinero
- **Disponible para todos,** en virtud de que la información cada día es más fácil de adquirir.
- La mayoría de los empleados cada día serán **menos de tiempo completo para la organización.**
- Nacimiento de nuevas instituciones teorías, problemas, a un **ritmo vertiginoso.**

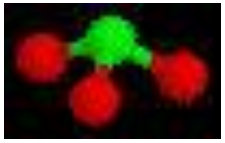
CONOCIMIENTO

ERA INDUSTRIAL

VALORES PREDOMINANTES

- Poder
- Control
- Disciplina
- Especialización
- Estructura jerarquizada

ERA DEL CONOCIMIENTO



VALORES PREDOMINANTES

- Descentralización
- Información
- Innovación
- Calidad
- Trabajo en equipo

Los trabajadores trabajan más con sus mentes que con sus manos (Knowledge Worker)

SOCIEDAD DEL USO DE CONOCIMIENTO

Embudo del Conocimiento

Administración Inteligente

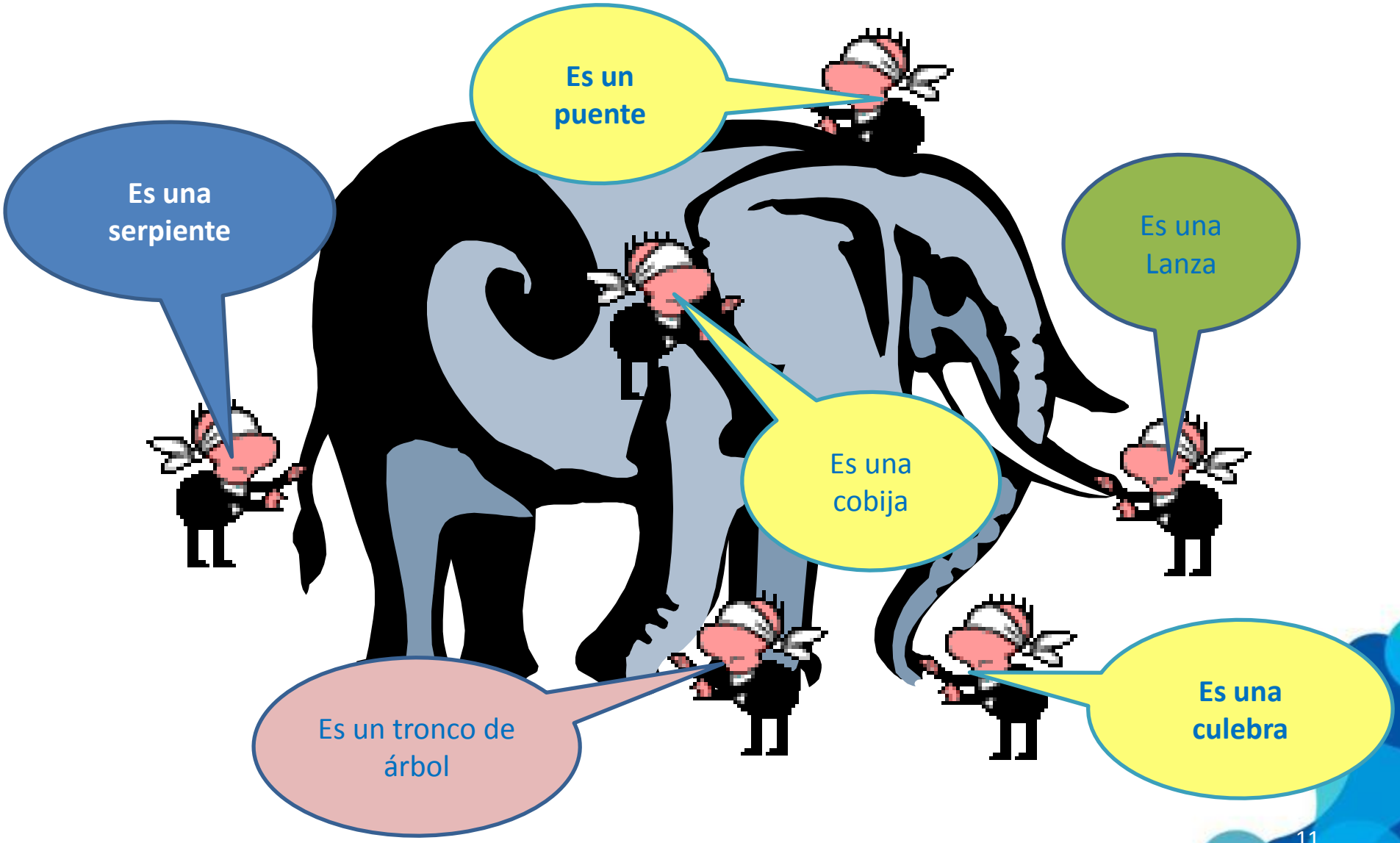
No se puede administrar lo que no se puede ver

La capacidad de ver toda la organización es el aspecto más importante de la **administración inteligente.**

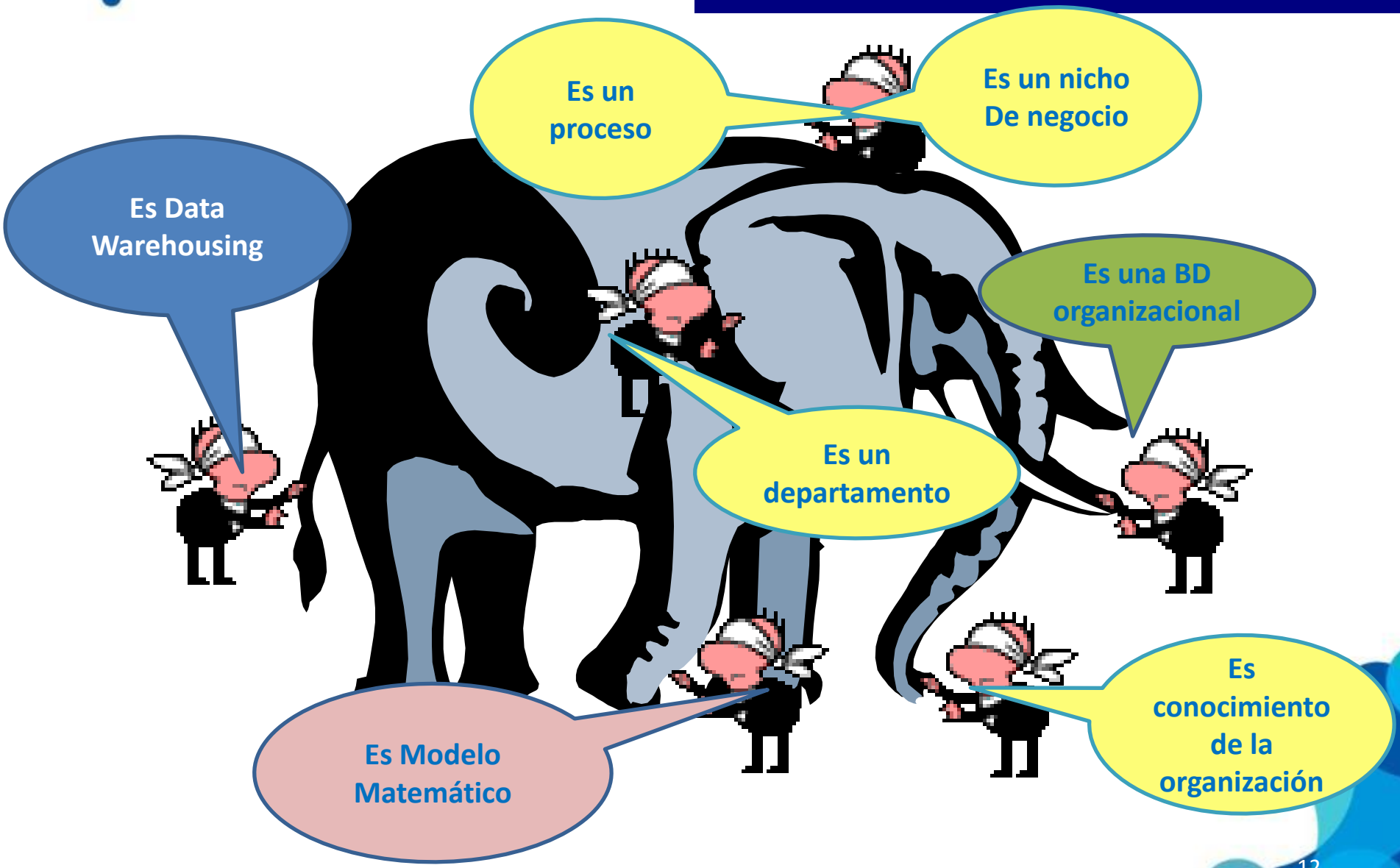
Las organizaciones tienen que ser **flexibles y adaptables** a cambios en el proceso y modelo de los negocios, así como también a nuevas tecnologías.

10

Un ciego describiendo un Elefante



Un gerente describiendo su organización Sin A`dD



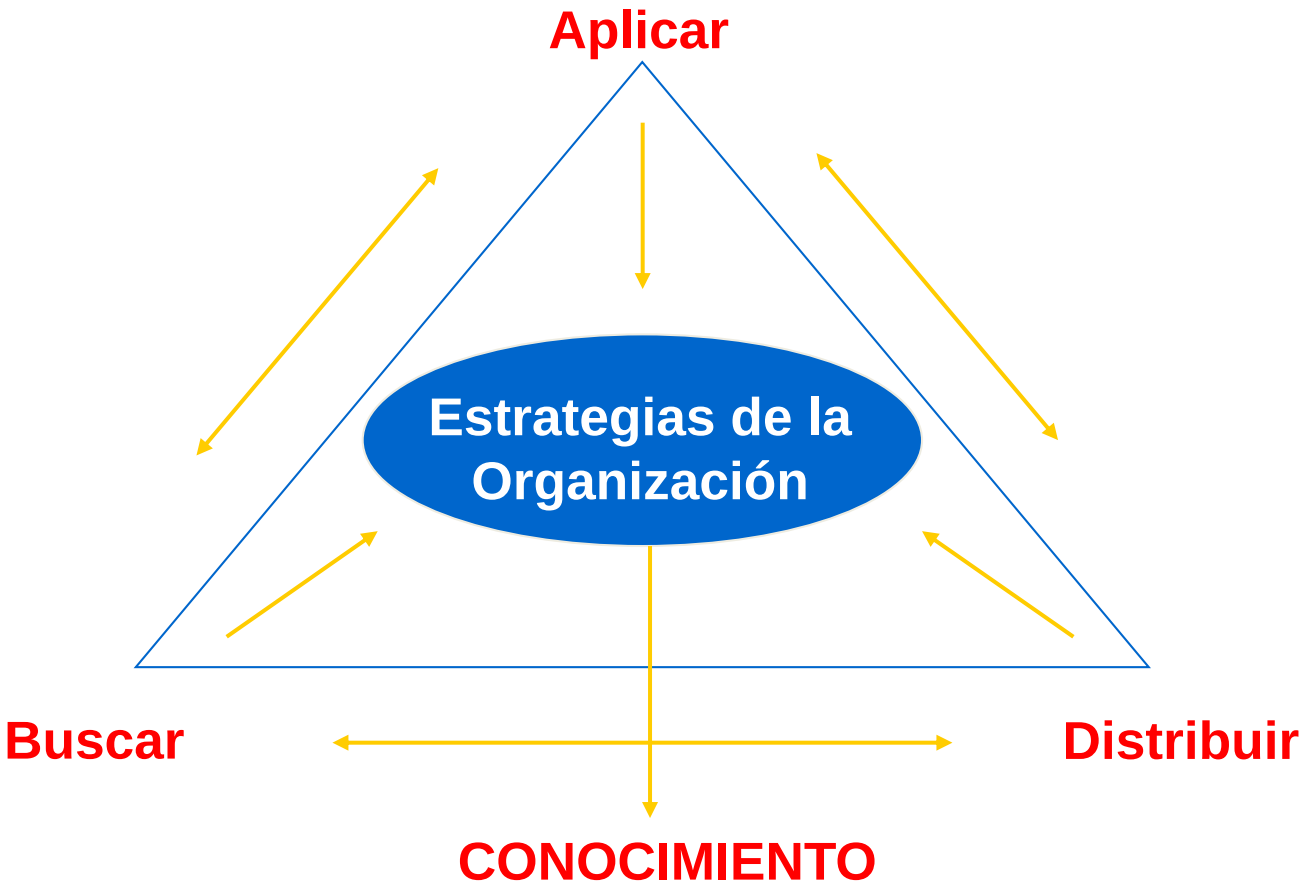
Dato, información, conocimiento e inteligencia en una organización

GESTIÓN DEL CONOCIMIENTO

"EL CONOCIMIENTO ES EL ACTIVO INTANGIBLE QUE MAYOR COMPETITIVIDAD GENERA A LAS NACIONES Y A LAS ORGANIZACIONES, EN LA ECONOMÍA GLOBAL"

"DEBE SER GERENCIADO"

CONOCIMIENTO ORGANIZACIONAL



Dato, información, conocimiento e inteligencia en una organización

Sistemas Operacionales

Eventos

Aprendizaje

Datos

Decisiones estratégicas

Planes

Reglas

Conocimiento

Tendencias, patrones y excepciones

Información

Extracción e Integración

CONOCIMIENTO ORGANIZACIONAL

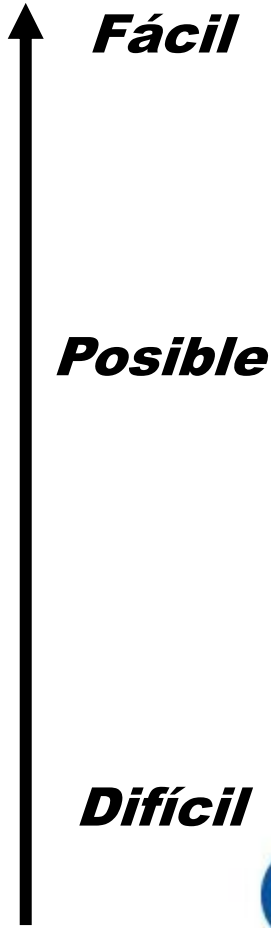
- **EXPLÍCITO:** Referencial, replicable, entrenable, reposa en textos, bases de datos, objetos, redes, como información.
- **TÁCITO:** Vivencial, basado en valores, creencias, actitudes, sentimientos humanos, no es fácil de replicar.

CONOCIMIENTO ORGANIZACIONAL

Conocimiento documentado
- Artículos
- Propuestas
- Presentaciones
- Políticas de la organización

Conocimiento explícito no documentado
- Información de proyectos
- Información de clientes
- Políticas de la organización
- Experiencia de los empleados
- Procedimientos para las reuniones
- Procedimientos para la contratación
- Mejores prácticas para propuestas

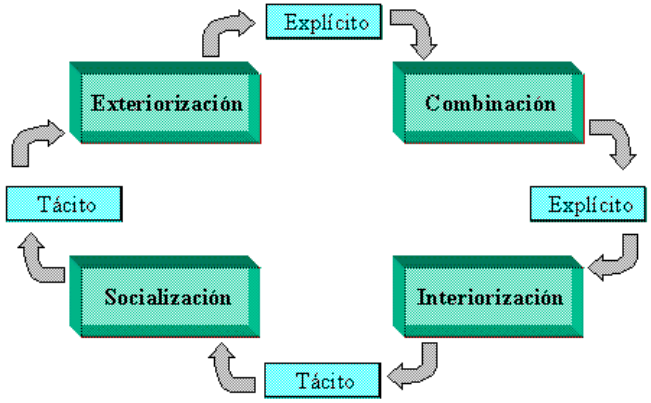
Conocimiento tácito no documentado
- Mejores prácticas para proyectos
- Mejores prácticas para atención a clientes



Modelo Espiral de Nonaka y Takeuchi

Los **4 modos de conversión** de conocimiento entre el *Conocimiento Tacito* (CT) y el *Conocimiento Explicito* (CE)

1. de CT a CT= **Socialización**
2. de CT a CE= **Externalización**
3. de CE a CE= **Combinación**
4. de CE a CT= **Internalización**



Gestión del Conocimiento

Es crear la posibilidad para todo el mundo de ver exactamente qué está pasando, dónde vamos bien.

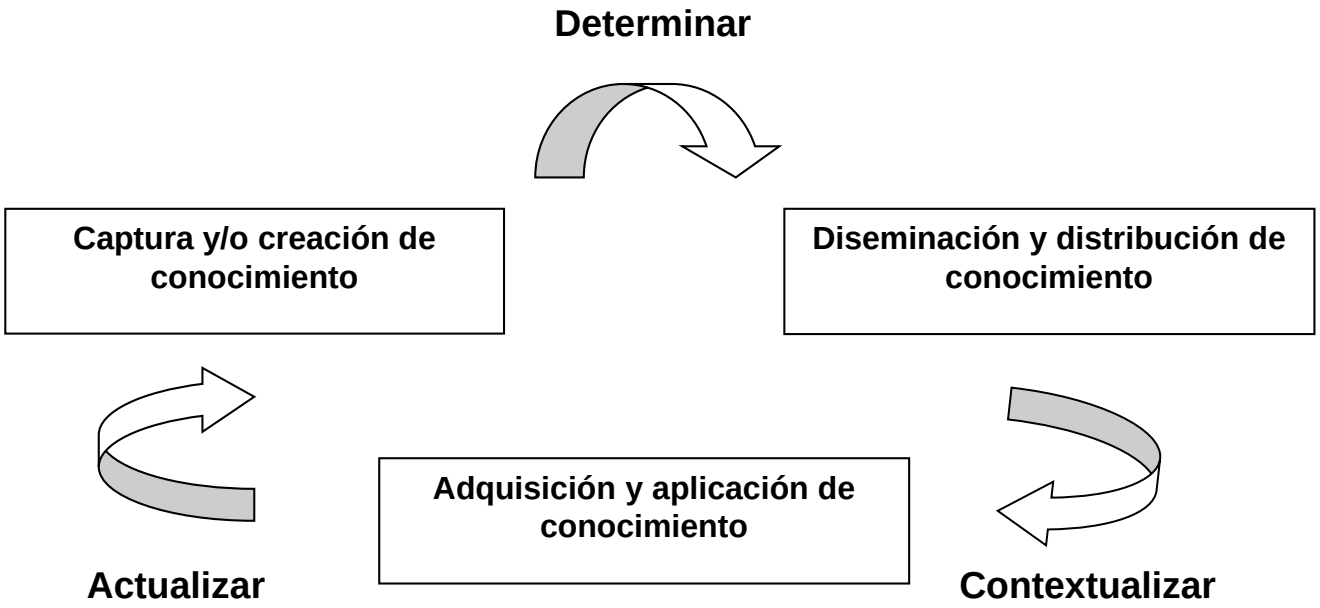
Bill Gates, presidente de Microsoft

Es la capacidad de la organización para crear nuevos conocimientos, diseminarlos y encapsularlos en productos, servicios y sistemas.

Ángel L. Arbonies, presidente de Cluster del Conocimiento

Ciclo de la Gestión del Conocimiento

1. **Captura y/o creación** del conocimiento
2. **Diseminación y distribución** del conocimiento
3. **Adquisición y aplicación** del conocimiento

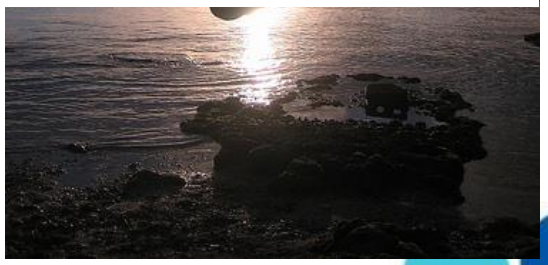


Atributos Claves de la Gestión del Conocimiento

- Generar nuevos Conocimientos
- Acceder conocimiento de fuentes externas
- Usar dicho conocimiento en la toma de decisiones
- Arraigar el conocimiento en procesos productos y servicios
- Representar el conocimiento en

Pero, ¿Cómo automatizar ese proceso??

- Facilitar el crecimiento del conocimiento a través de cultura y otros incentivos
- Transferir conocimiento existente a otras partes de la organización
- Medir el valor de los activos del conocimiento y/o su impacto en el administración del conocimiento



Ruggles and Holtshouse (2013)

Analítica de Datos

Es la ciencia de la recogida, almacenamiento, extracción, limpieza, transformación, agregación y análisis de datos, **con el fin de descubrir información y conocimiento.**

El **alto grado de datificación incrustado en la sociedad** exige nuevas herramientas y mecanismos para la manipulación y la representación de los datos que facilitan la **extracción de conocimiento significativo para las organizaciones**.

Analítica de Datos

Los datos son el nuevo petróleo de la economía

Es la ciencia que examina datos en bruto con el propósito de buscar conocimiento, sacar conclusiones, generar información, entre otras cosas.

Es usado en muchos ámbitos:

- La industria para tomar mejores decisiones empresariales
- Las ciencias para verificar o reprobar modelos o teorías existentes.
- …

Analítica de Datos

Con las grandes cantidades de datos disponibles, las organizaciones **deben centrarse en la explotación de los datos** para obtener una **ventaja competitiva**.

- **Las computadoras** son más poderosas,
- **el trabajo en red** es omnipresente,
- se han desarrollado **algoritmos que pueden conectar conjuntos de datos**

esto permite **análisis más amplios y profundos** que antes era imposible.

Los objetivos principales de AdD son:

- Ayudar a ***ver los problemas de la Organización desde una perspectiva de los datos***, y
- Comprender los principios de ***extracción de conocimiento útil a partir de los datos***.

La gestión usando AdD

El éxito de la analítica sólo puede medirse en términos de lo bien que ayudan a lograr objetivos estratégicos

Por lo tanto, se debe:

- Identificar los objetivos de la organización
- Recoger los datos necesarios para medir sus objetivos
- Analizar los datos
- Sacar conclusiones basado en los datos

Ejemplo de Análisis de los datos

Datos disponibles para un agricultor:

1. Los patrones climáticos históricos
2. Los datos de cultivo de plantas y la productividad de cada cepa
3. Las especificaciones de los Fertilizantes
4. Las especificaciones de los Plaguicidas
5. Los datos de productividad del suelo
6. Los datos del ciclo de plagas
7. Los costos, la fiabilidad, etc., de las máquinas
8. Los datos de conducción del agua
9. Los datos históricos de oferta y demanda
10. Los precios del Mercado

Ejemplo de Análisis de los datos

Los datos se pueden usar para responder a:

¿Cuál es el patrón de siembra agrícola para obtener el mejor precio?

¿cuáles son los productos agrícolas con mejor rendimiento?

Usar Analítica Datos para obtener conocimiento, desde sus datos.

- En cuanto a los **datos del tiempo y plagas**, podría establecer **correlaciones entre un cierto tipo de hongo cuando el nivel de humedad** alcanza un cierto punto.
- Si las **futuras proyecciones meteorológicas** para los próximos meses **predicen un bajo nivel de humedad,** por lo tanto, **habrá riesgo bajo de ese hongo**.

Para el agricultor, esto podría significar ser capaz de plantar **un determinado tipo de producto agrícola no resistente a ese hongo**, con un mayor rendimiento y precio en el mercado, **sin tener que comprar un determinado fungicida**...

Analítica de Datos

Los datos pueden "hablar“

El análisis de datos contiene aspectos del razonamiento científico:

Define
Interpreta
Evalua
Ilustra
Discute
Explica
Clarifica
Compara
Contrasta

Data Analytical
http://www.youtube.com/watch?v=-xR5erOhkXo

Objetivo de un análisis:

- **Explicar** los fenómenos de causa y efecto
- **Encontrar** respuestas a un problema particular
- **Concluir** acerca de eventos del mundo real basado en el problema
- **Aprender** de un problema
- **Predecir/pronosticar** en el mundo real fenómenos
- **Interpretar/Analizar** una situación
- …

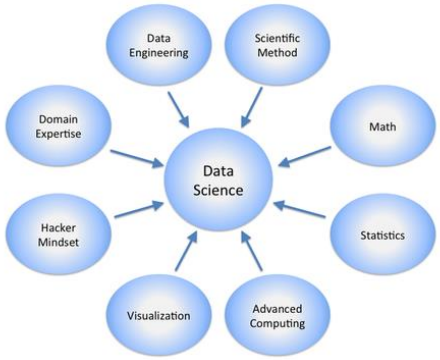
Guía básica para el análisis de datos:

- Analizar es “NO“ narrar, ni describir
- Descomponer los objetivos en preguntas de investigación
- Identificar los fenómenos que han de analizarse
- Buscar respuestas desde los datos
- Validar las respuestas con otros datos
- Afirmar soportado por datos

Is Data Scientist the Sexiest Job of Our Time?

By Emily Waltz

Posted 27 Sep 2012 | 22:32 GMT



Like many start-up companies, business networking site LinkedIn once struggled to capitalize on the mountains of data generated by its users. Then, in 2006, a new hire, data scientist Jonathan Goldman, swept in and tamed the unwieldy data mess in a way that launched the company to the next level. Goldman extracted patterns from the connections between LinkedIn's users, and came up with a way to suggest to those users other people they may know. The "people you may know" feature created millions of new page views, and LinkedIn's growth went skyward.

Sexy? The folks at *Harvard Business Review* think so, and in their October issue they proclaimed the data scientist the sexiest job of the 21st century. The authors of the article, Thomas H. Davenport, a visiting professor at Harvard Business School, and D.J. Patil, a data scientist at Greylock Partners, liken the profession to the Wall Street quants of the 1980s, and the computer engineers of the 1990s. "If 'sexy' means having rare qualities that are much in demand, data scientists are already there," they wrote.

A data scientist's job description goes like this: Make discoveries while swimming in data. Possess an intense curiosity. Bring structure to formless data and make analysis possible, all while having a feel for business issues and an empathy for customers. Advise executives on how to use the information to make better products.

What kind of professional can do all of these things? The rare kind with the powerful combination of skills that let them wear the hats of data hacker, analyst, communicator, and trusted adviser—all of which must be applied to a specific technology or product. (*Spectrum* last year identified 26 variations of data mining in IT).

Because more and more companies need someone with these skills, the demand for data scientists is exceeding the supply. These professionals garner high salaries and large stock option packages, the authors found in an informal survey. But more than that, data scientists want to be "on the bridge"-—a reference to Star Trek, in which Captain James Kirk relies on data supplied by Mr. Spock. They want to be involved in decision making, not just advising.

The dearth of data scientists of this caliber has become a constraint on some sectors, forcing people to devise their own ways of generating and locating talent. Greylock Partners, a venture firm, has built its own specialized recruiting team to channel talent to the businesses in its portfolio. And after acquiring the big data firm Greenplum, EMC launched a data science and big data analytics training and certification program.

Once companies snag a good data scientist, it's hard to hold onto him or her. LinkedIn lost Jonathan Goldman to Aster Data, which lost him to Level Up Analytics, a company Goldman co-founded.

http://spectrum.ieee.org/tech-talk/computing/it/is-data-scientist-the-sexiest-job-of-our-time

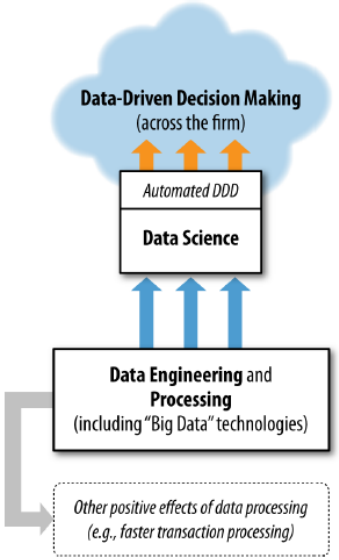
La ciencia de los datos

La ciencia de los datos

Combinación de las matemáticas, estadísticas, etc., para resolver el problema de captura de datos, además de la limpieza, la preparación y la alineación de los datos.

La ciencia de los datos

La ciencia de datos requiere de *principios, procesos y técnicas* para la comprensión de los fenómenos para la *extracción automatizada* de los datos.



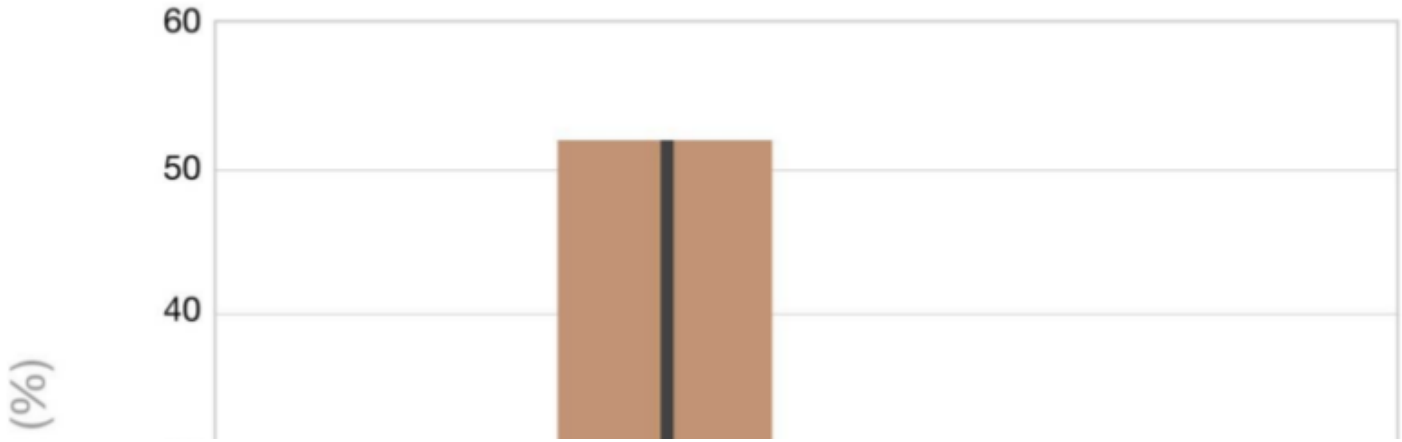
Conocer los datos

La ciencia de datos es un procedimiento que **consume tiempo y requieren mucho trabajo**, pero que es absolutamente necesario para la AdD con éxito.

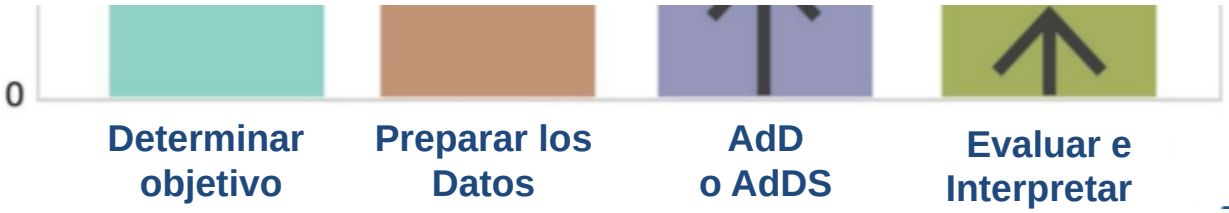
Los datos deben pasar por **procesos de ensamblaje, integración, limpieza, agregación y preparación general**.

Expertos de dominio deben ser consultados para explicar las anomalías, los valores perdidos, el significado de los números enteros que representan categorías en lugar de cantidades numéricas, y así sucesivamente.

¿Qué etapa lleva mas esfuerzo?



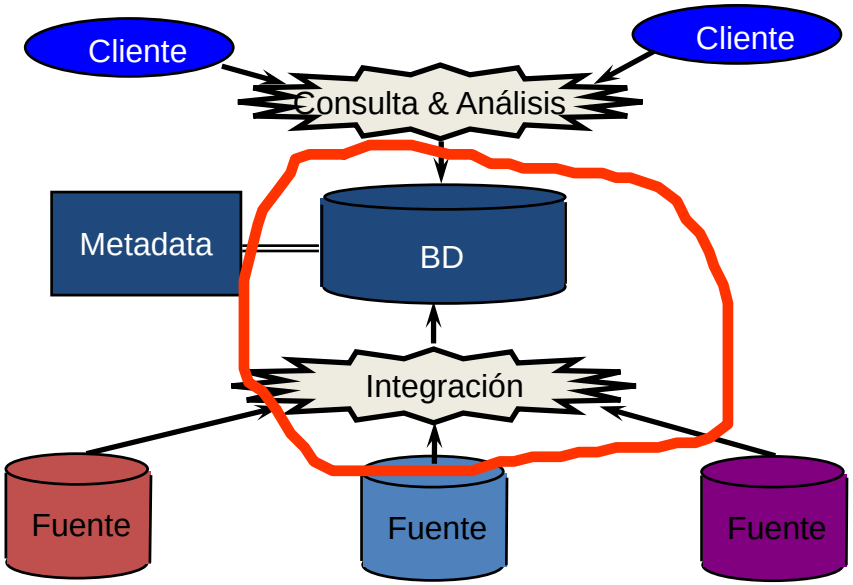
Preparación de la entrada para un proceso de AdD suele **consumir la mayor parte del esfuerzo invertido en el proceso**.



La ciencia de datos es un procedimiento que consume tiempo y requieren mucho trabajo, pero que es absolutamente necesario para la AdD con éxito.

Integración

- Selección de los datos
- Transformación de datos
- Carga de datos



Proceso ETL

ETL (Extracción, Transformación y Carga)

Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas mencionados como datos mal capturados, anómalos y vacíos, etc.

- normalización de formatos,
- remoción de anomalías,
- corrección de errores
- eliminación de duplicados.

Transformación

En esta etapa se **transforman las variables de entrada en nuevas variables de interés**, a través de diversos métodos.

Una transformación de variables puede ser la combinación entre variables

- concatenación de cadenas,
- multiplicación entre variables,
- otras operaciones aritméticas, etc.

Reducción

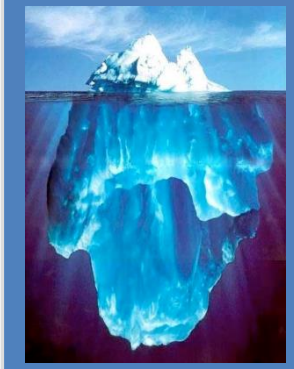
Consiste en **decidir qué datos deben ser utilizados para el análisis**. El criterio que se sigue para realizar reducción de variables incluye la relevancia con respecto a los objetivos que se persiguen, y limitaciones técnicas tales como los volúmenes máximos de datos o tipos de datos concretos.

Así que en este paso se reduce la cantidad de variables **a sólo las necesarias para modelar el proceso en estudio**.

- **Realizar análisis estadísticos** para reducir variables que posean una alta relación lineal, como por ejemplo un análisis de correlación.
- **Identificar las posibles variables** que se pueden reducir.
- **Justificar la reducción** de las mismas
- **Construir la nueva vista minable** con las nuevas variables reducidas

¿Qué es lo Nuevo con AdDS?

Todo está pasando en línea



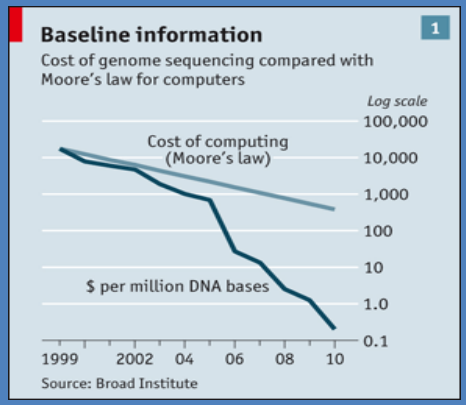
Cada uno:
Hace clic
Ve anuncio
Factura un evento
Navega...
Solicita servidor
Realiza Transacción
Mensaje de error de red
...

Generado por el usuario (Web y móvil)

Las fuentes

IoT

Computación Científica



¿Qué es lo Nuevo con AdDS?

- Comercio electrónico
- Compras en línea en tiendas/supermercados
- Transacciones bancarias/de tarjeta de crédito
- Redes sociales
- Fotos, documentos, etc. en línea

¿Qué es lo Nuevo con AdDS?

Captura, Curación y Agregación

- **Ciencia de datos debe trabajar con:**
  - Datos incompletos
  - Los datos suelen estar desordenados
  - Administrar grandes conjuntos de datos
  - Una gran diversidad de tipos de datos:
    - Datos de texto (Web),
    - Datos Semi-estructurados (XML),
    - Grafos: Red social, Web semántica (RDF),
    - Stream de datos
  - Una gran diversidad de fuentes

Peligros en AdD

Ciencias de los Datos

- Privacidad
- Seguridad
- Decisiones basado en datos incompletos
- Decisiones con datos inexactos
- Usando sólo los datos que apoyan nuestras decisiones
- Llegar a la conclusión errónea de los datos: por ejemplo, los precios de las acciones

Modelado de Datos

Introducción a Data Warehouse

Analizando la información de una empresa

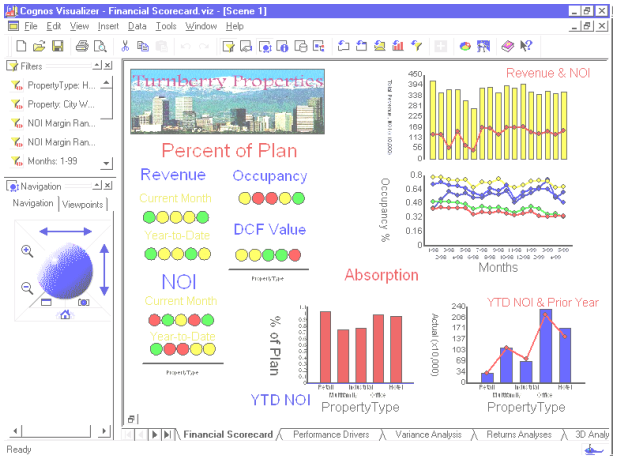
- ✓ Información periódica de las ventas
- ✓ Información del esfuerzo comercial
- ✓ Información sobre los pedidos a los proveedores

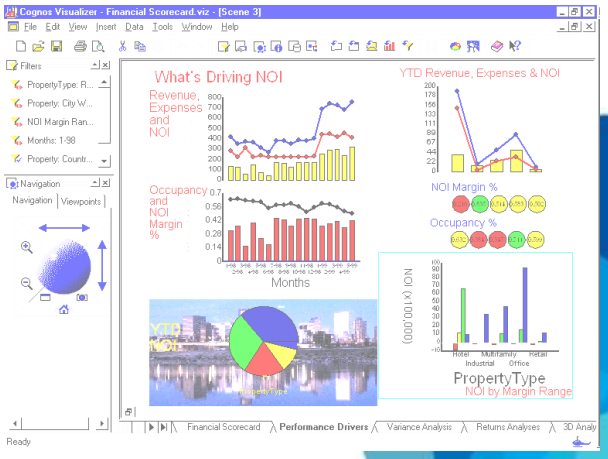
Por qué no integrarla y cruzarla para obtener:

- ✓ ¿En qué zonas se está vendiendo más de cada línea de productos?
- ✓ ¿Quienes son los clientes más rentables?
- ✓ ¿Cuál es la relación entre el esfuerzo comercial y las operaciones cerradas?
- ✓ ¿De qué proveedores se está comprando la mayor parte de los productos vendidos ?

Introducción a Data Warehouse

- ✓ Se necesita entender no solo QUÉ está pasando, sino CUÁNDO, DÓNDE, QUIÉN, CÓMO Y POR QUÉ.
- ✓ Requerimientos de información con OPORTUNIDAD .
- ✓ ESCALAR, ENRIQUECER Y COMPARTIR a todos los usuarios en la organización



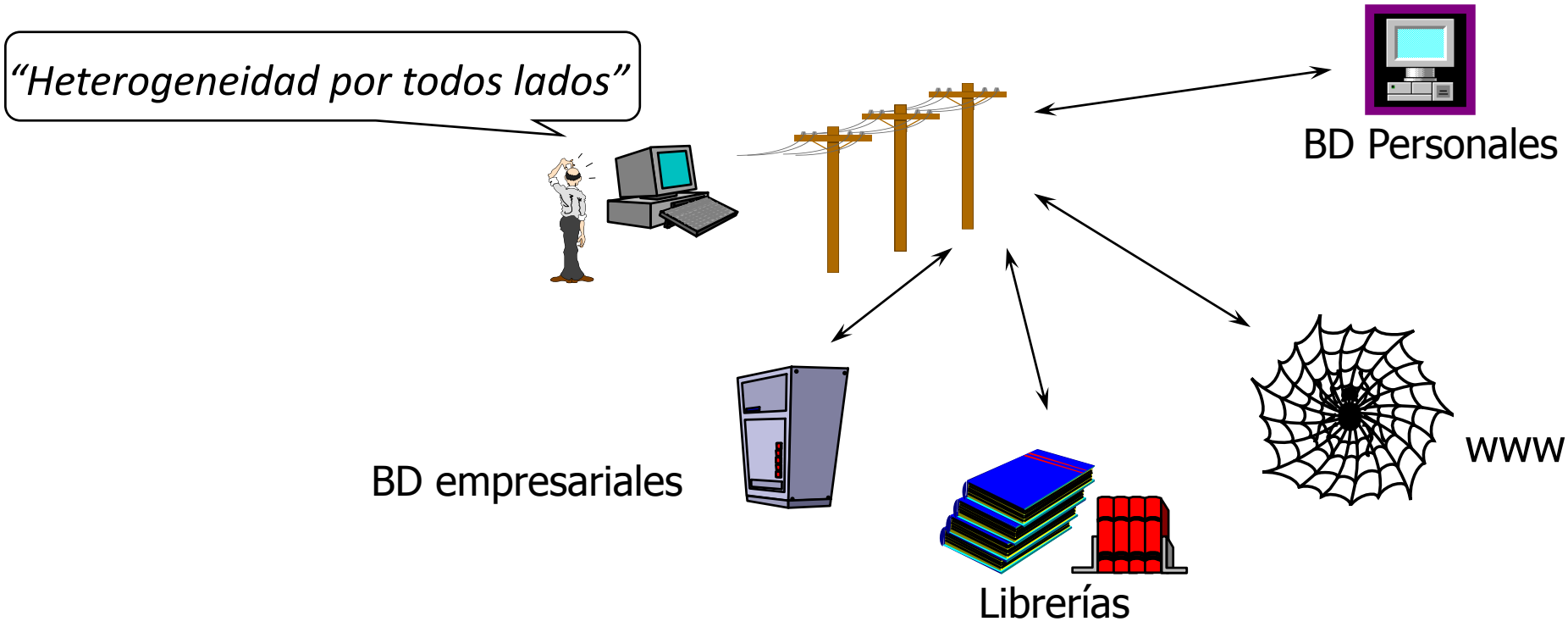


Introducción a Data Warehouse

	DATO	INFORMACION
Ventas	• Número de pedidos • Productos pedidos • Clientes que realizaron los pedidos	• Clientes más rentables • Pedidos más frecuentes • Productos más rentables • % de nuevos clientes
Servicio al Cliente	• Datos de llamadas de nuestros clientes • Información sobre el log de nuestra página web	• ¿Qué clientes visitan nuestra página web ? • % pedidos realizados por nuestros canales de ventas • ¿Qué consulta es más frecuente?
Marketing	• Número de campañas realizadas y características de cada una	• % éxito de las campañas • ¿Qué tipo de clientes han respondido mejor a cada una de las campañas realizadas ?
Distribución	• Productos que salen diariamente del almacén • Tiempo teórico de entrega	• Número de pedidos retrasados • Distribuidor que tiene el mayor número de retrasos • Tiempo medio de entrega

Fuente : Deloitte & Touche

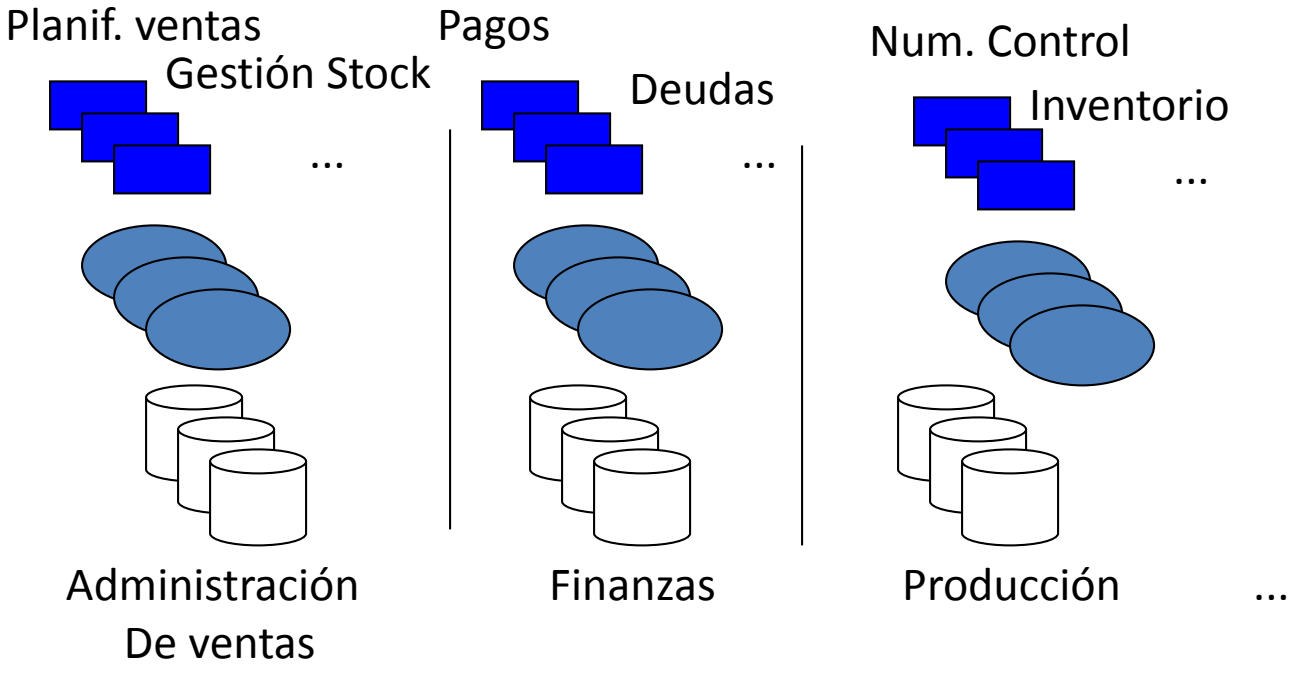
Problemas: Heterogeneidad de las fuentes de Información



- Diferentes interfaces
- Diferentes representaciones de datos
- Información duplicada e inconsistente

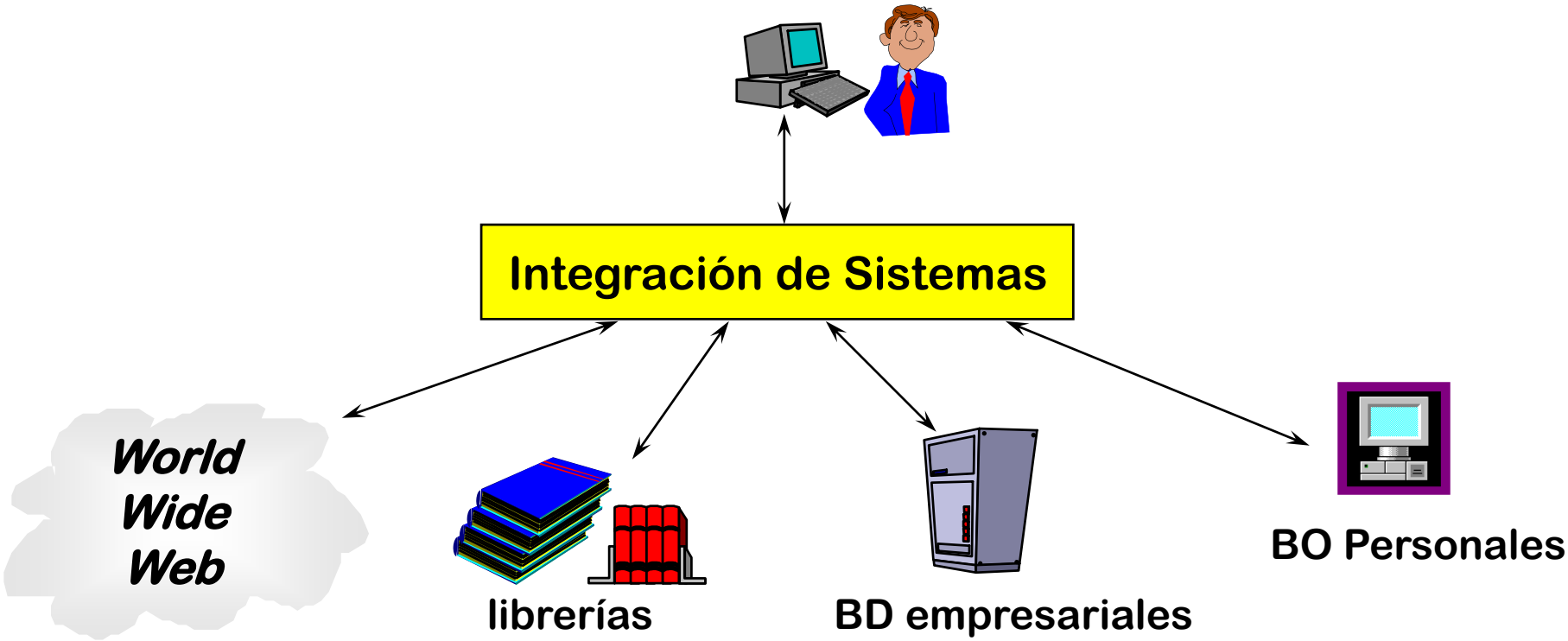
Problemas: Islas de Gestión de datos en grandes empresas

- Fragmentación vertical de los sistemas de información
- Desarrollo de las aplicaciones guiadas por los sistemas operativos



Objetivo:

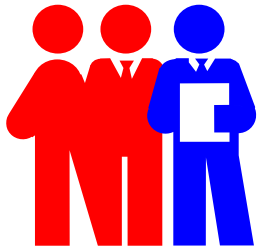
Unificar Acceso a los Datos

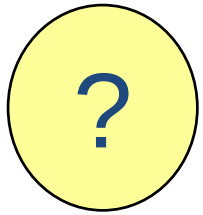


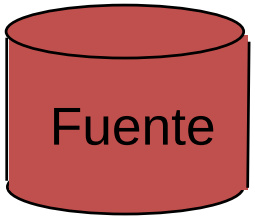
- Recopilar y combinar la información
- Proporcionar visión integrada, en una interfaz de usuario uniforme
- Soportar el intercambio

Objetivo:
Unificar Acceso a los Datos

- Dos enfoques:
  - Guiado por la consulta (perezoso)
  - Warehouse (ansioso)

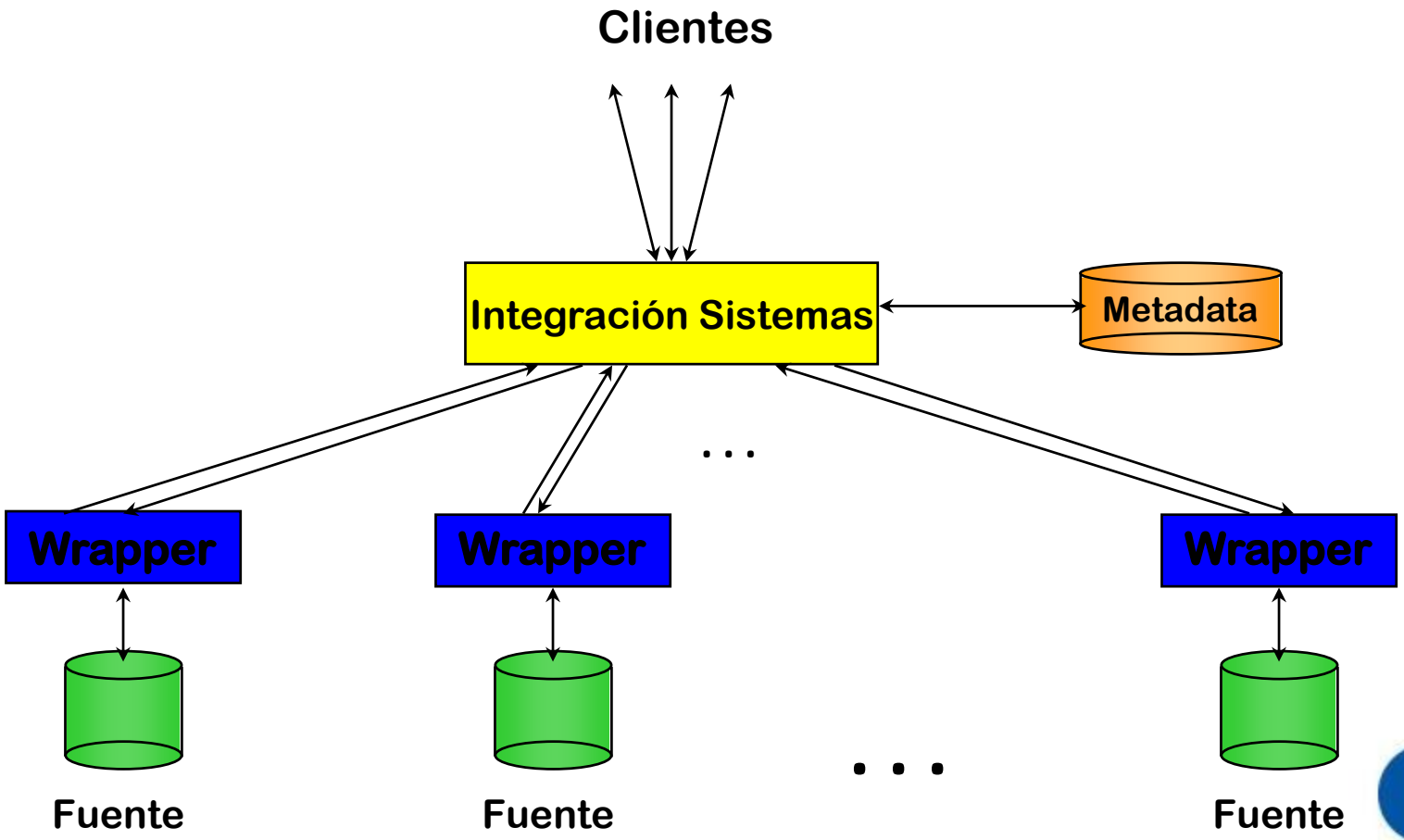






Enfoque tradicional

- Guiado por la consulta (perezoso, on-demand)

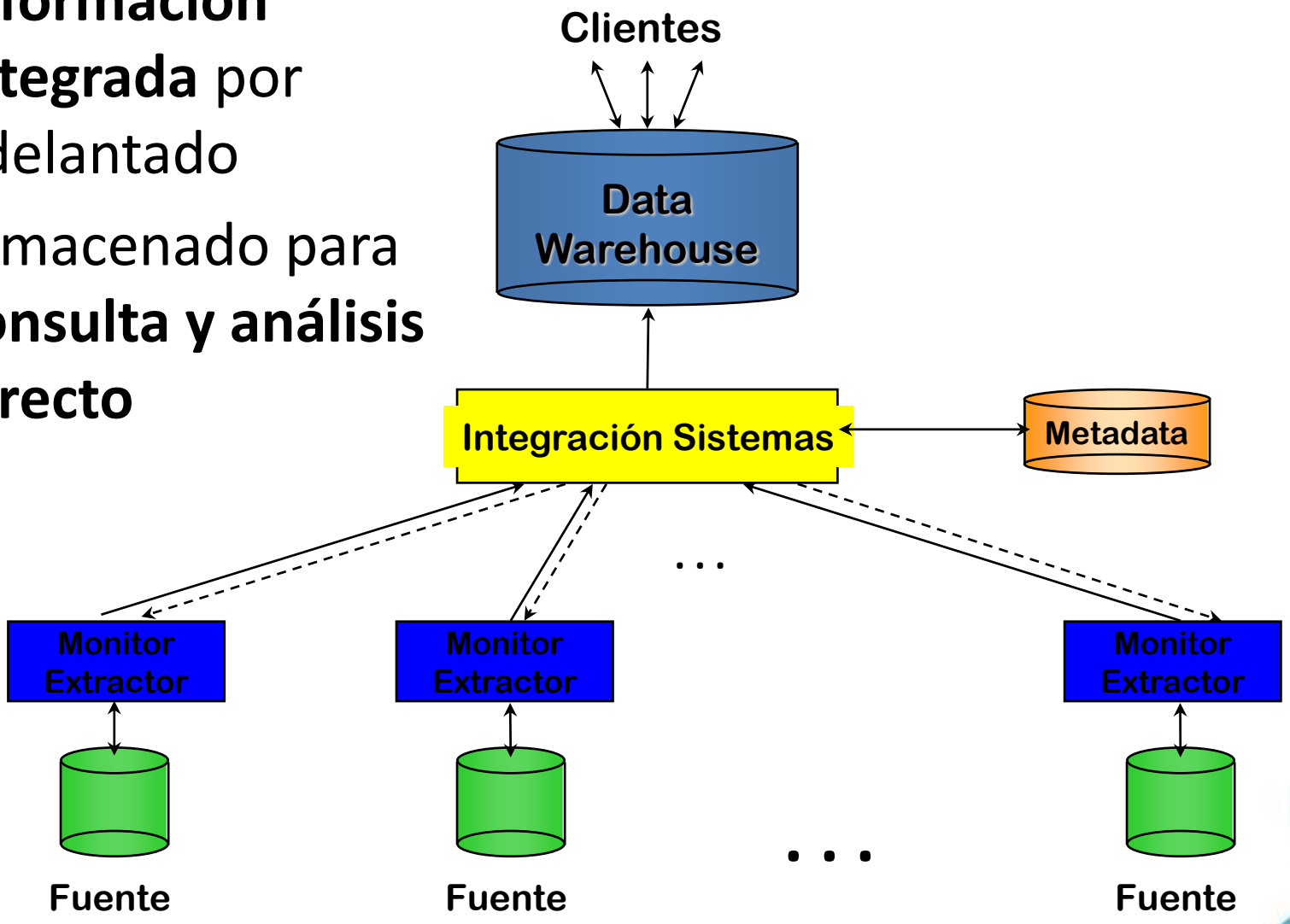


Problema del Enfoque tradicional

- El **retraso** en el procesamiento de consultas
- Fuentes de información **lentas o no disponibles**
  - Filtrado e integración son complejas
  - Ineficiente y potencialmente costoso para las consultas frecuentes
- **Compite con el procesamiento local** en el sitio fuente

Enfoque Warehousing

- **Información integrada** por adelantado
- Almacenado para **consulta y análisis directo**



Ventaja Enfoque Warehousing

- **Alto rendimiento de la consulta**
  - Pero no necesariamente la información más actualizada
- **No interfiere con el procesamiento local en el sitio origen**
  - Las consultas complejas en warehouse
  - OLTP en las fuentes de información
- **Información copiada en el almacén**
  - Se puede modificar, anotar, resumir, reestructurar, etc.
  - Puede almacenar información histórica
  - Seguridad, sin auditoría
- **Usada en la industria**

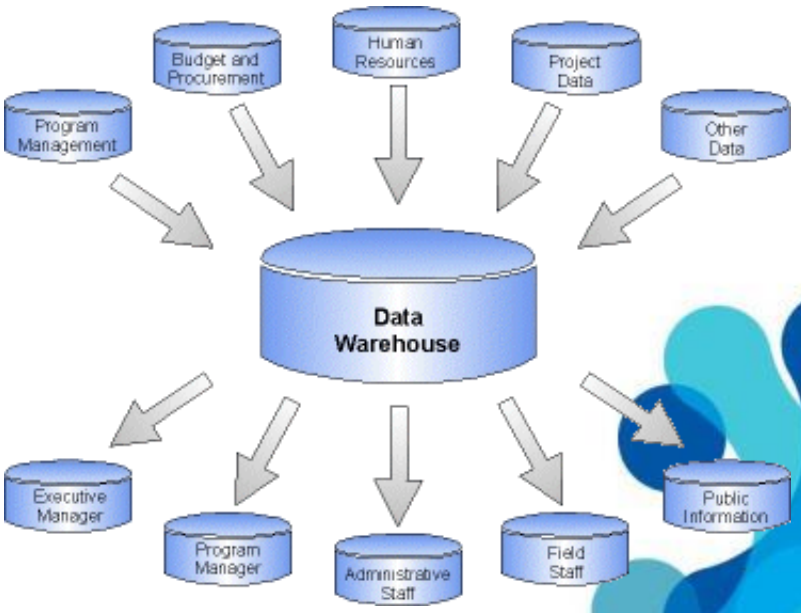
Concepto Data Warehousing

Es un **gran almacén de datos** para consultar

Es un **repositorio de datos** de muy **fácil acceso**, alimentado de **numerosas fuentes**, transformadas en grupos de información sobre **temas específicos de negocios**, para permitir **nuevas consultas, análisis, toma de decisiones**.

Se trata, de **un expediente completo de una organización**, más allá de la información transaccional y operacional, almacenado en una base **de datos diseñada para favorecer el análisis y la divulgación eficiente de datos**.

Tiene **gran capacidad de almacenamiento**, pues los datos pueden ser de grandes periodos de tiempo.



Orientado hacia temas

Los datos se almacenan y agrupan por temas de interés

Se agrupa por **temas orientados a la organización**, tales como :

- Clientes
- Productos
- venta

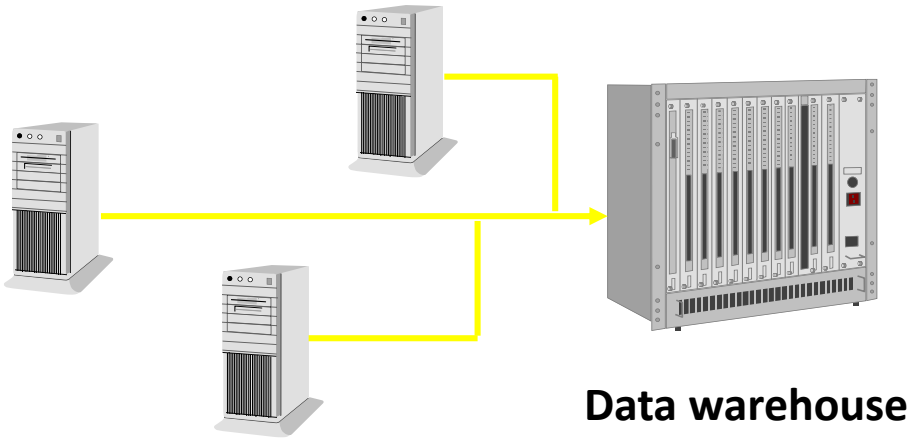
en lugar de las transacciones individuales.

La normalización no es relevante

- **Centrándose en el modelado y análisis de datos** para la toma de decisiones, no en las operaciones diarias o el procesamiento de transacciones.
- Proporcionar una **visión sencilla y concisa en torno a cuestiones temáticas particulares** mediante la exclusión de los datos que no son útiles en el proceso de apoyo a la decisión.

Integración de Datos

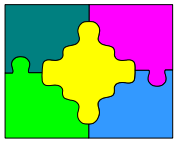
- El almacén de datos **integra datos** que provienen de varias fuentes.
- Vienen de las bases de datos operacionales, y mediante un proceso de **carga de datos** se añaden al Datawarehouse.
- El proceso de carga es lo más complicado por problemas de **preparación de los datos.**



Los datos en el almacén deben ser:

- Limpios
- Validados
- Adecuadamente integrados

Ejemplo Integración de Datos



Sistema de Cuenta de cheques
Jane Doe (nombre)
Female (sexo)
Bounced check #145 on 1/5/95
Opened account 1994

Sistema de cuentas de ahorro
Jane Doe
F (sexo)
Opened account 1992

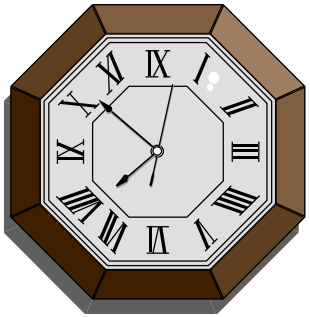
Sistema de inversiones de clientes
Jane Doe
Owns 25 Shares Exxon
Opened account 1995

↑ *Datos operacionales*

cliente
Jane Doe
Female
Bounced check #145
Married
Owns 25 Shares Exxon
Customer since 1992

↑ *data warehouse*

… variante en el tiempo …



- Todos los datos en el almacén de datos tienen una **marca de tiempo** en el momento de entrada en al almacén o cuando se usan en el almacén.
- Esta grabación cronológica de datos ofrece posibilidades **históricas y análisis de tendencias**.
- Normalmente, en las BD operativas **los datos se sobrescribe**, ya que los valores anteriores no son de interés.

TIEMPO

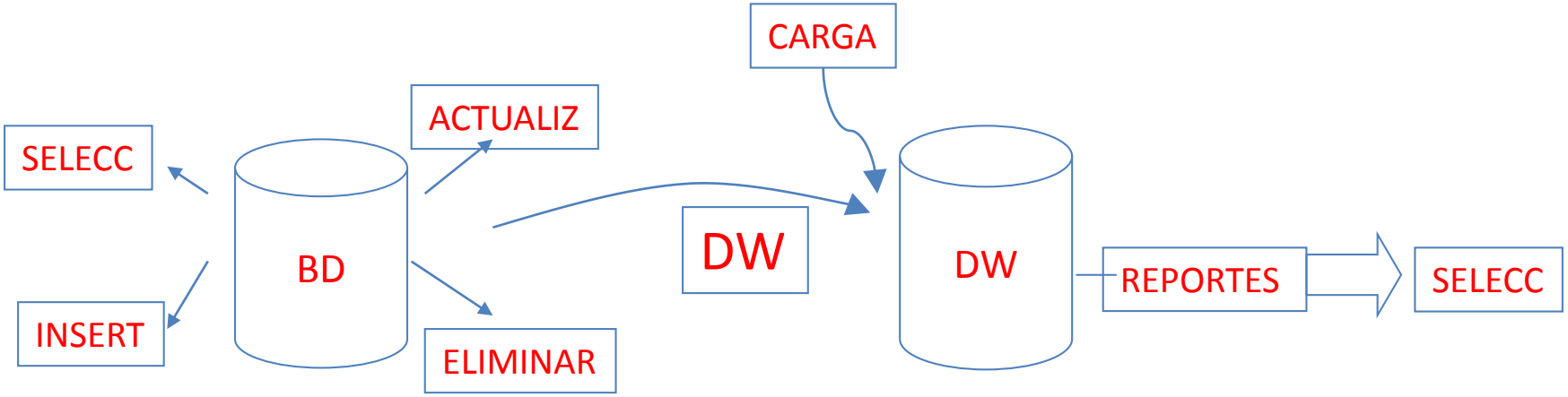
id_tiempo

* periodo

No volátiles

Son estables:

una vez almacenados los datos **no se modifican**.

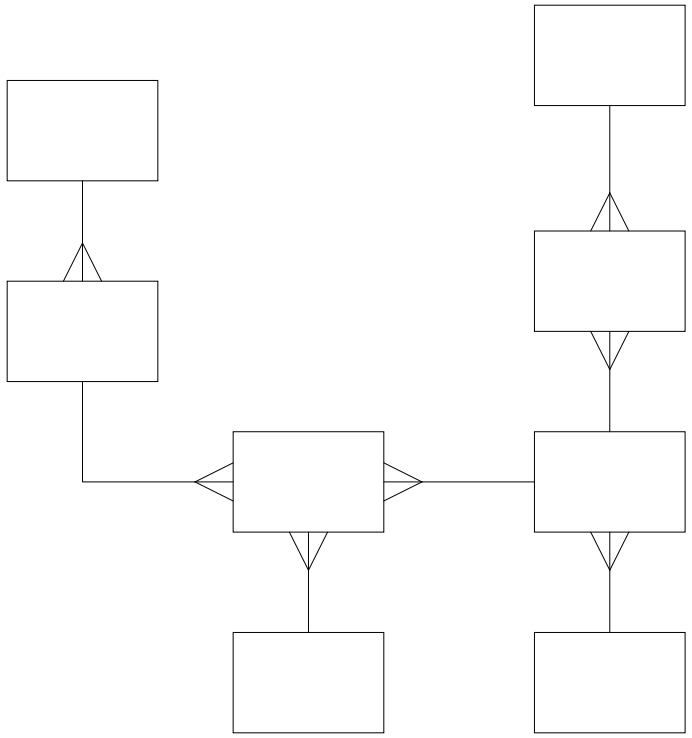


Datos actúan como un recurso estable para informes coherentes y análisis comparativo.

Por el contrario, **los datos operativos se actualizan** (insertar, eliminar, modificar, etc.)

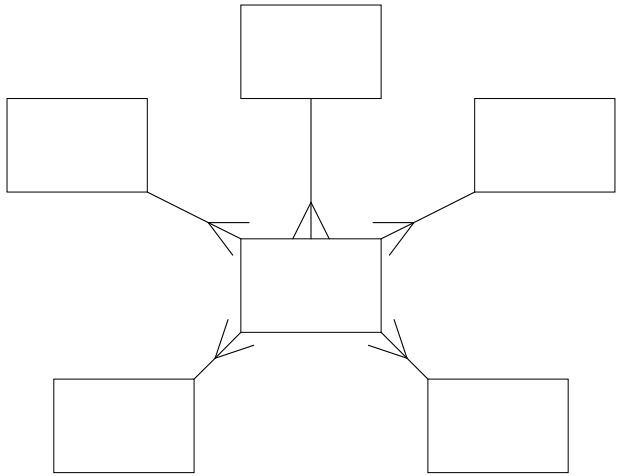
Las diferencias de diseño

BDs



Diagráma ER

Data Warehouse



Estrella

Elementos que integran un almacén de datos

Metadatos

"datos acerca de los datos".

Su función es recoger la descripción de la estructura del almacén de datos:

Tablas

Columnas en tablas

Jerarquías y Dimensiones de datos

Relaciones entre tablas

Entidades y Relaciones

RESUMEN DIFERENCIAS

BD OPERACIONAL

- Datos operacionales
- Orientado a aplicaciones
- Datos Actuales
- Datos Detallados
- Datos en continuo cambio

DATAWAREHOUSE

- Datos **estratégicos** del negocio
- Orientado al **sujeto**
- Actuales + **Histórico**
- Datos **Resumidos**
- Datos **Estables**

Modelos dimensionales

Es una técnica de **diseño lógico** comúnmente utilizada para Data Warehouses, que busca presentar los datos en una arquitectura estándar y permita una **alta performance de acceso** a los usuarios finales.

El modelo se basa en **esquemas estrella**, conformados por **Tablas de Hechos y Tablas Dimensionales** (p.ej. cubos).

Modelos dimensionales

- Un **modelo relacional desnormalizado**
  - Compuesto por tablas con atributos
  - Las relaciones son definidas por claves nuevas y claves externas

- Organizado para l**a comprensibilidad y facilidad de presentación** de informes, en lugar de facilitar la actualización

- Consultado y mantenido por **herramientas especiales de gestión analítica**

Esquema en estrella: Componentes

- Hechos
- Dimensiones
- Atributos
- Jerarquías de atributos

Diseño de Esquemas

Los tipos de Esquema

- En estrella
- Constelación
- Copo de nieve

1. **Aislar Datos a tener en cuenta**
   - Esquemas de las Tablas de hechos
2. **Definir las dimensiones**
   - Ejes de análisis
3. **Estandarizar dimensiones**
   - Dividir en varias tablas unidas por referencias
4. **Integrar todo**
   - Varias tablas de hechos comparten algunas tablas de dimensiones (constelación de la estrella)

Diseño de Esquemas

- **Los datos en las dimensiones se organizan por temas:**

  Los clientes, los productos, las ventas, ...

- **Tema = datos + dimensiones**
  - **Recopilación de datos** útiles sobre un tema

    Ejemplo: ventas
  - **Sintetizar** una visión única de los temas a analizar

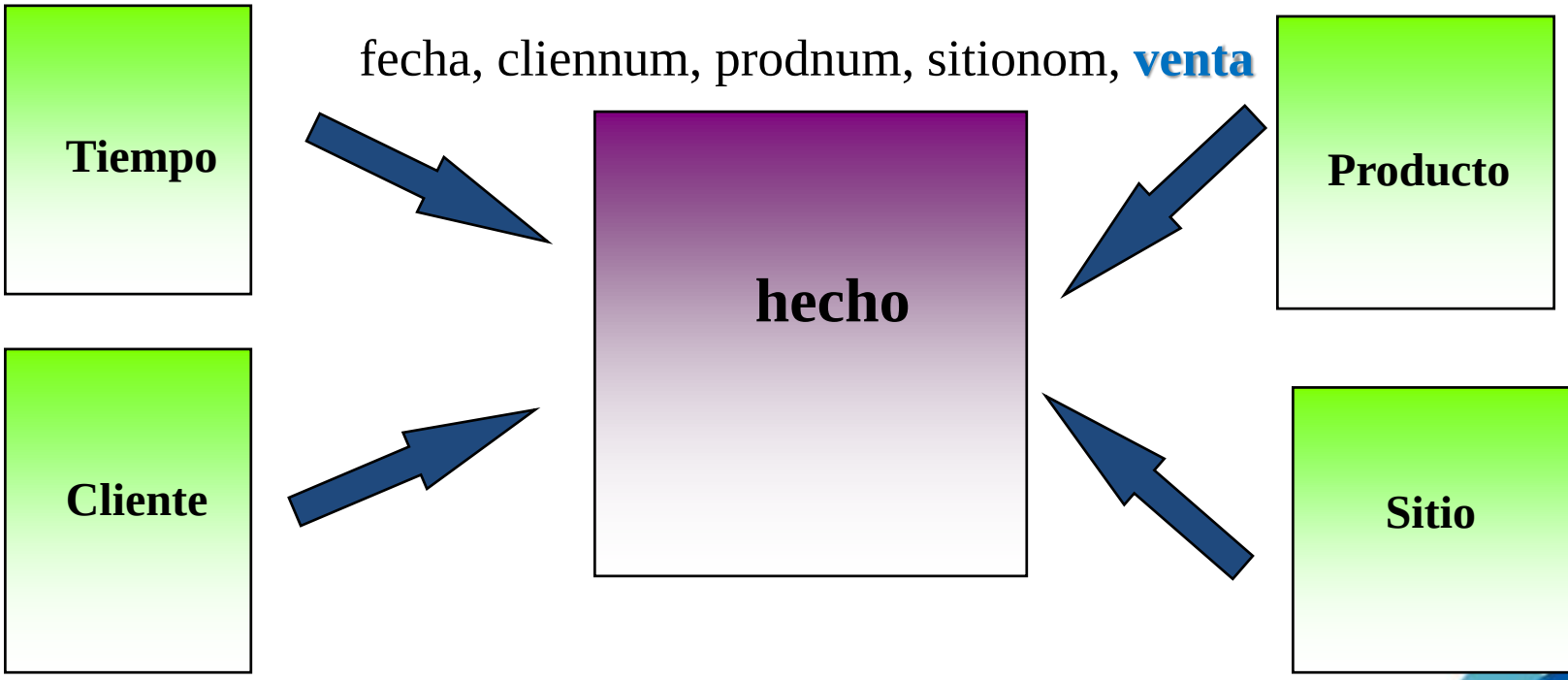
    Ejemplo: Ventas (producto, período, tienda, número)
  - **Detallar** la vista según dimensiones

    Ejemplo:

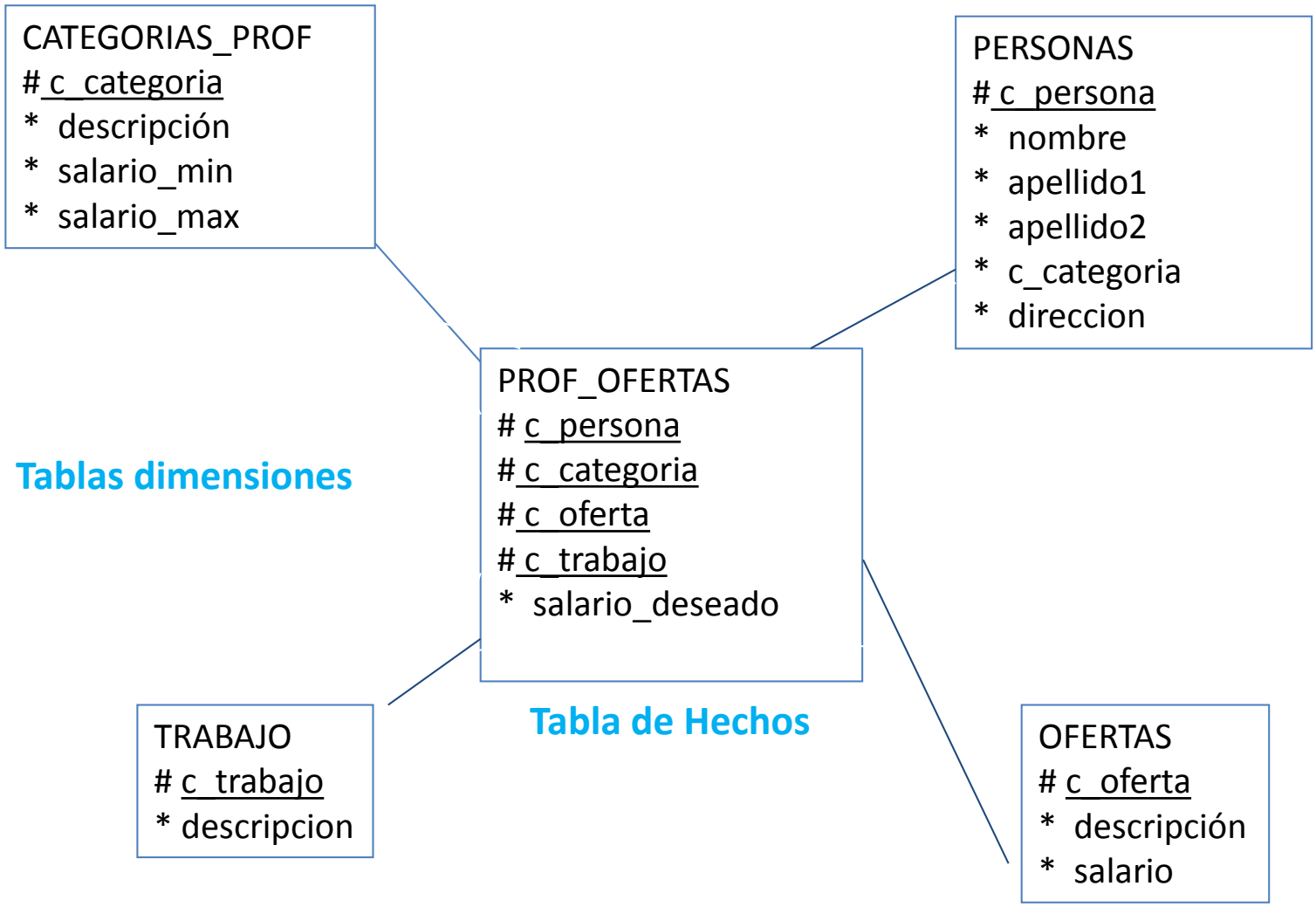
    Productos (IDprod, descripción, color, tamaño ...)
    Tiendas (IDmag, nombre, ciudad, departamento, país)
    Periodo (IDper, año, trimestre, mes, día)

Esquema en estrella

- Una sola tabla de hechos y para cada dimensión una tabla de dimensiones
- No captura jerarquías directamente



Esquema en estrella



Esquema en estrella

- El modelo estrella es una representación de una vista de la organización.
  - Ventas
  - Mercadeo

- El modelo estrella consolida hechos en relación a dimensiones o filtros.

- Esquema en estrella
  - Hecho rodeado de varias dimensiones (4-15)
  - Las dimensiones se de-normalizan

Tablas de hechos

Contienen los hechos que serán utilizados por los analistas para apoyar el proceso de toma de decisiones.

- Recibe enlace dimensiones
- Almacena el conocimiento generado desde los datos
- Enlaza dimensiones a través de sus claves

Esquema en estrella:
Hechos

- Mediciones numéricas (valores) que representan un aspecto del negocio o actividad específica
- Almacenado en una tabla de hechos en el centro del esquema de estrella
- Contiene hechos caracterizados a través de sus dimensiones
- Se pueden calcular o derivar en tiempo de ejecución
- Actualizado periódicamente con los datos de las bases de datos operacionales

Guarda Medidas de interés del negocio

Tablas de dimensión

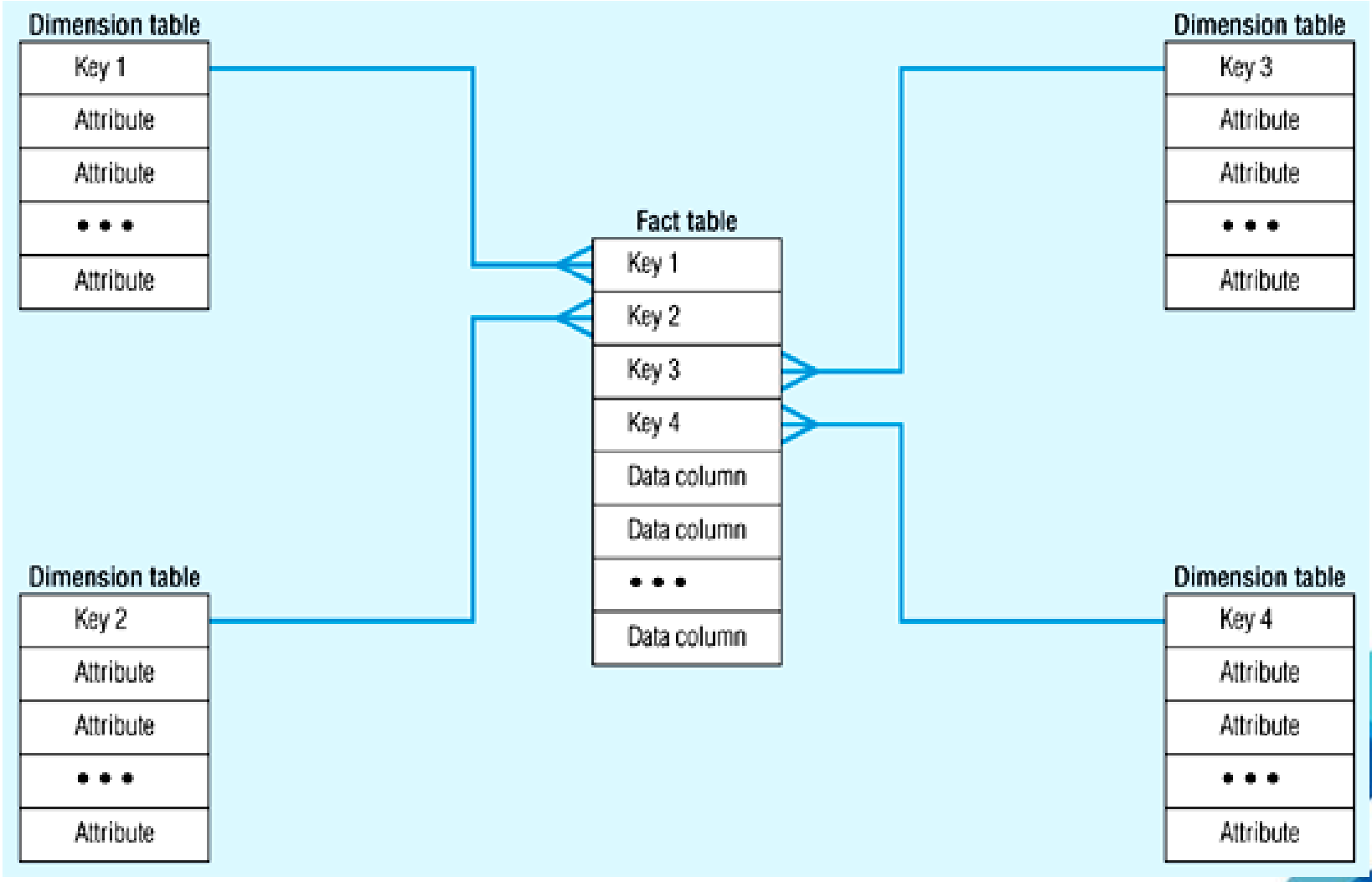
Definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto organizacional.

- Toma los datos desde los sistemas transaccionales
- Depura los valores de los atributos para incorporarlos al modelo dimensional
- Mantiene las claves
- Características cualitativas que proveen perspectivas adicionales a un hecho

Esquema en estrella:
Tabla de Dimensiones

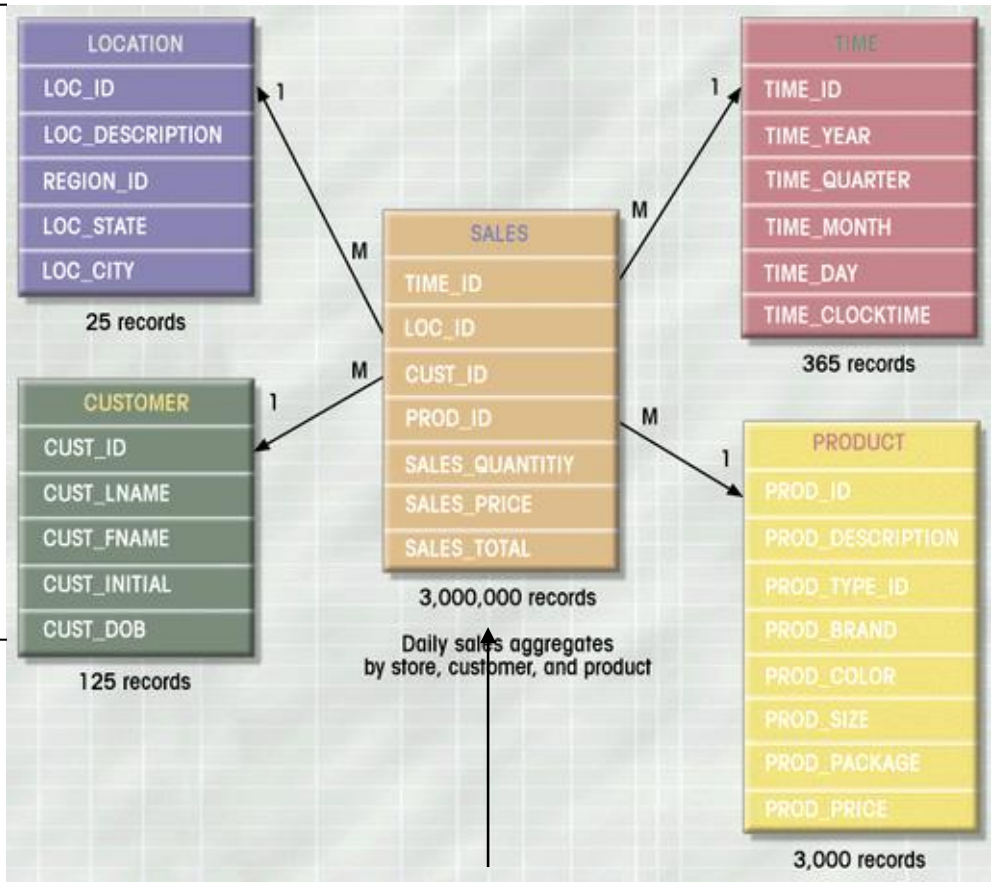
- Se enlaza a la tabla de hechos (clave primaria única)
- Guarda los Atributos del negocio
- Más o menos constante los datos
- Contiene información textual descriptiva
- Filas anchas (muchos campos, inclusos descriptivos)
- Tablas pequeñas (alrededor de un millón de filas)
- Tablas de dimensiones contienen atributos
- Los atributos se utilizan para buscar, filtrar o clasificar los hechos

Esquema en estrella: Atributos



Ejemplo de esquema en estrella para ventas

Tablas de Dimensiones



Relación M-1

Tabla de hechos

Conclusiones
Esquema en Estrella

- Las tablas de hechos están relacionados a cada tabla de dimensión en una **relación Muchos a Uno**
- Tabla de hechos está **relacionado con muchas tablas** de dimensiones
- La clave principal de la tabla de hechos es compuesta de las **claves principales de las tablas de dimensiones**
- Cada tabla de hecho está diseñada para **responder a una(s) pregunta(s) específica(s) de IN**

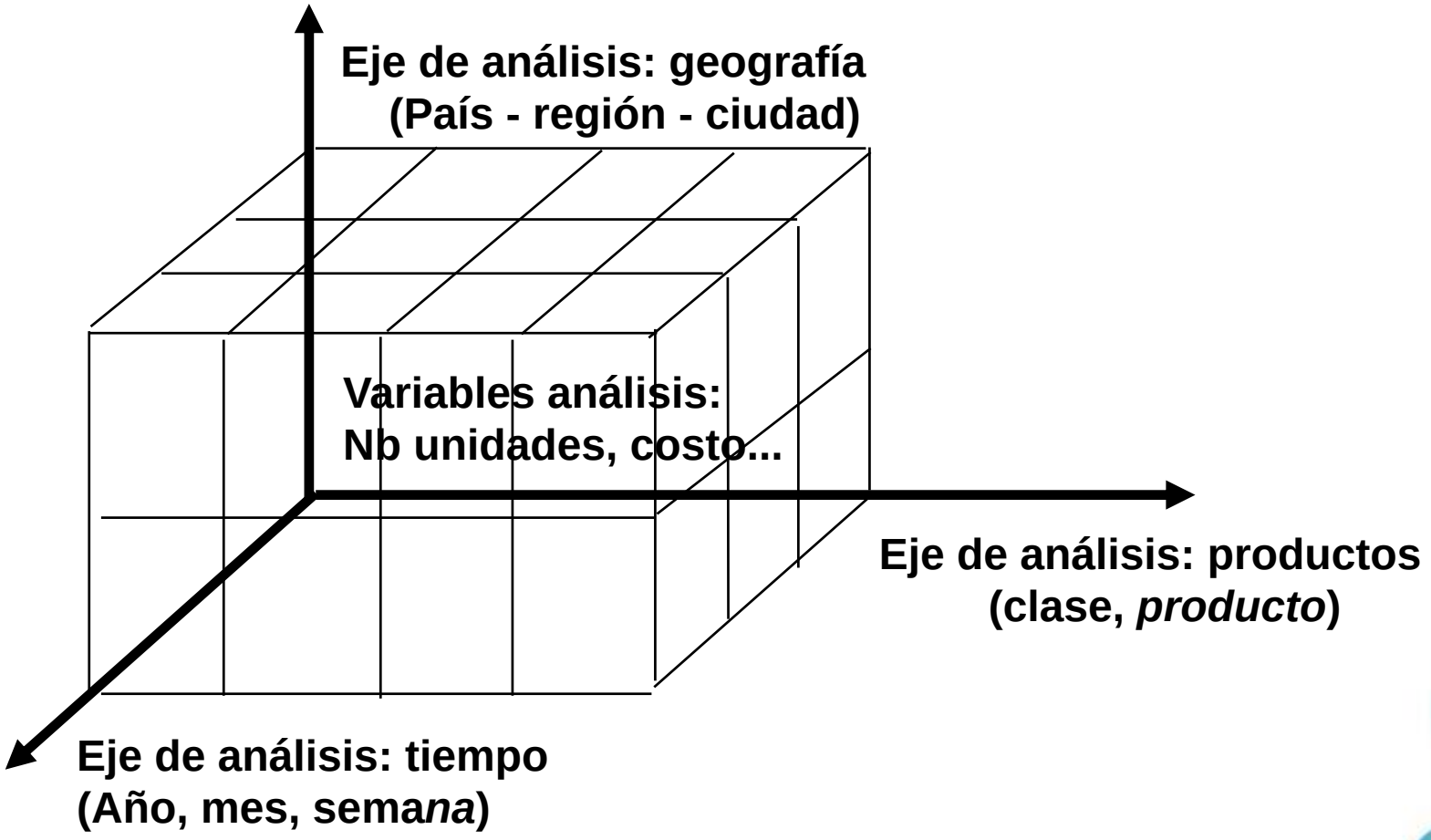
Las multidimensiones

- **Dimensiones:**
  - Tiempo
  - Geografía
  - Productos
  - Clientes
  - Canales de ventas.....
- **Indicadores:**
  - Número de unidades vendidas
  - Promedio productos vendidos por zona

Cubo de dato y dimensiones

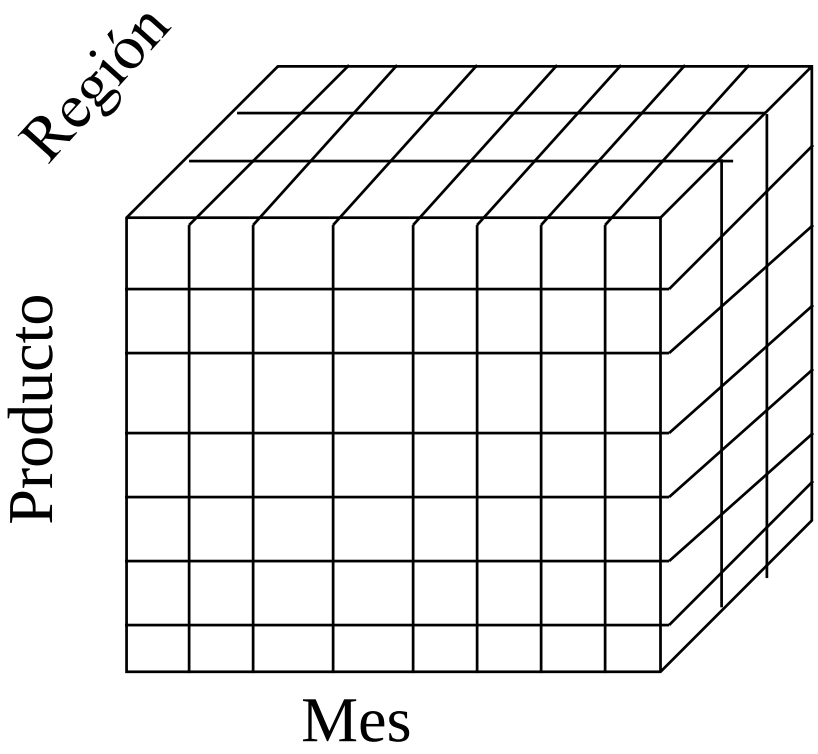
Eje de análisis : dimensiones

Variables análisis: indicadores



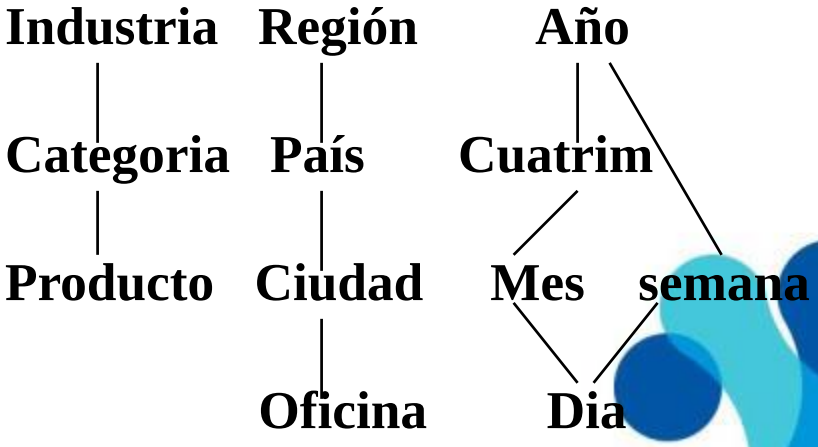
Datos Multidimensionales

El volumen de ventas en función del producto, el mes, y el área



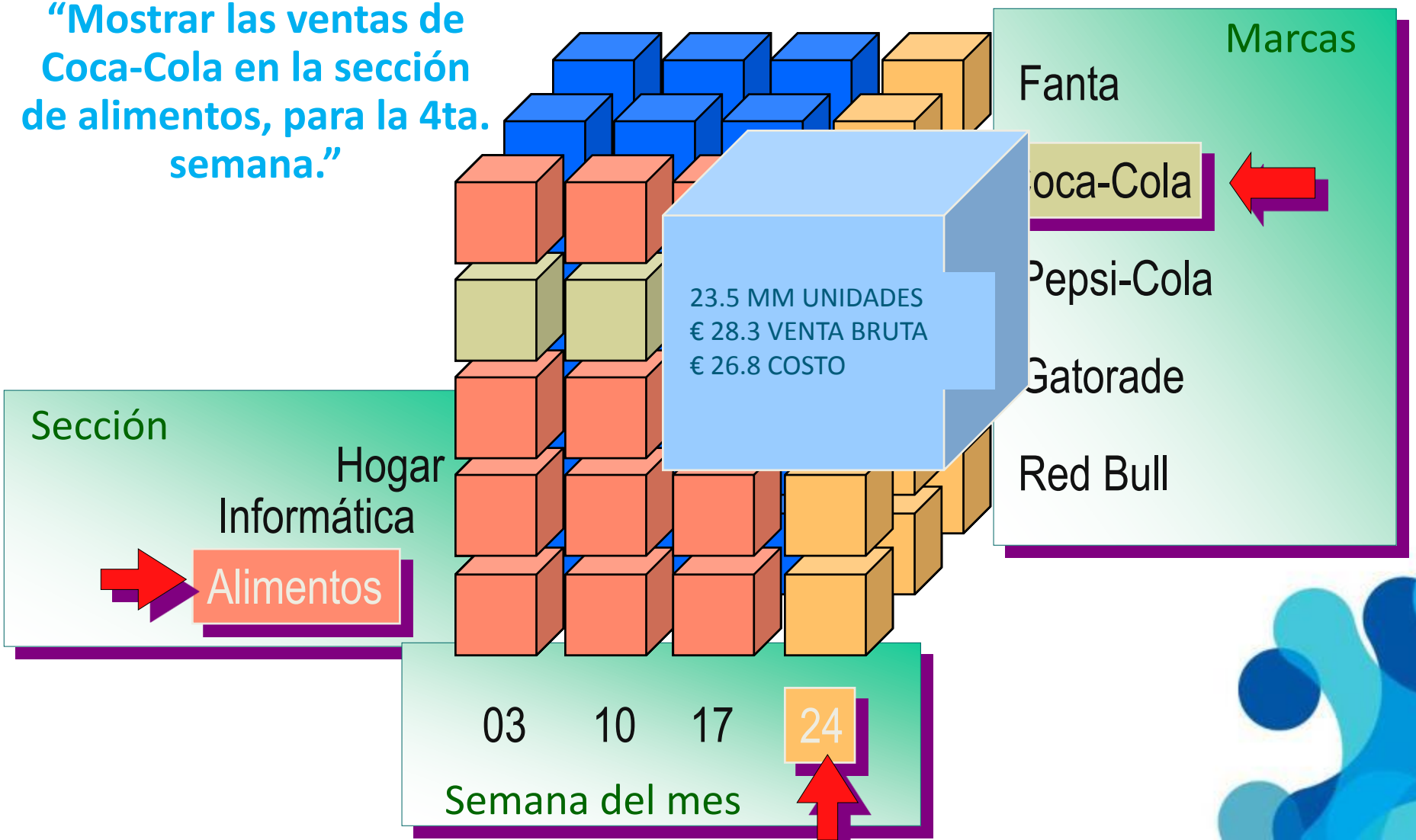
Dimensiones: Producto, Localidad, Tiempo

Caminos jerárquicos



Cubo Multidimensional

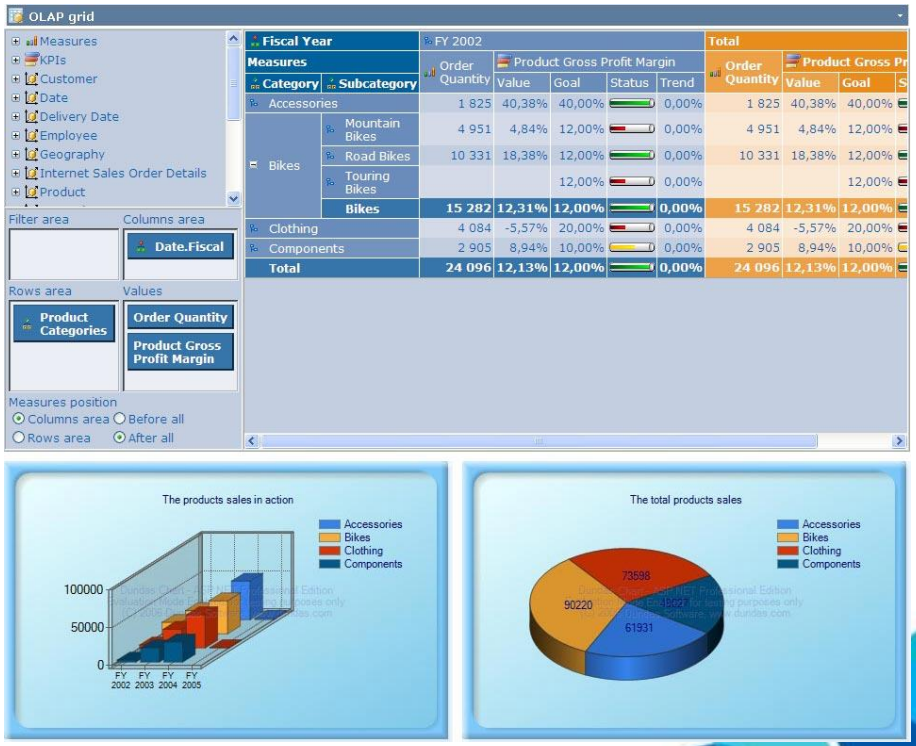
"Mostrar las ventas de Coca-Cola en la sección de alimentos, para la 4ta. semana."



Herramientas para explotación del Datawarehouse

Análisis multidimensional (OLAP online analytical processing)

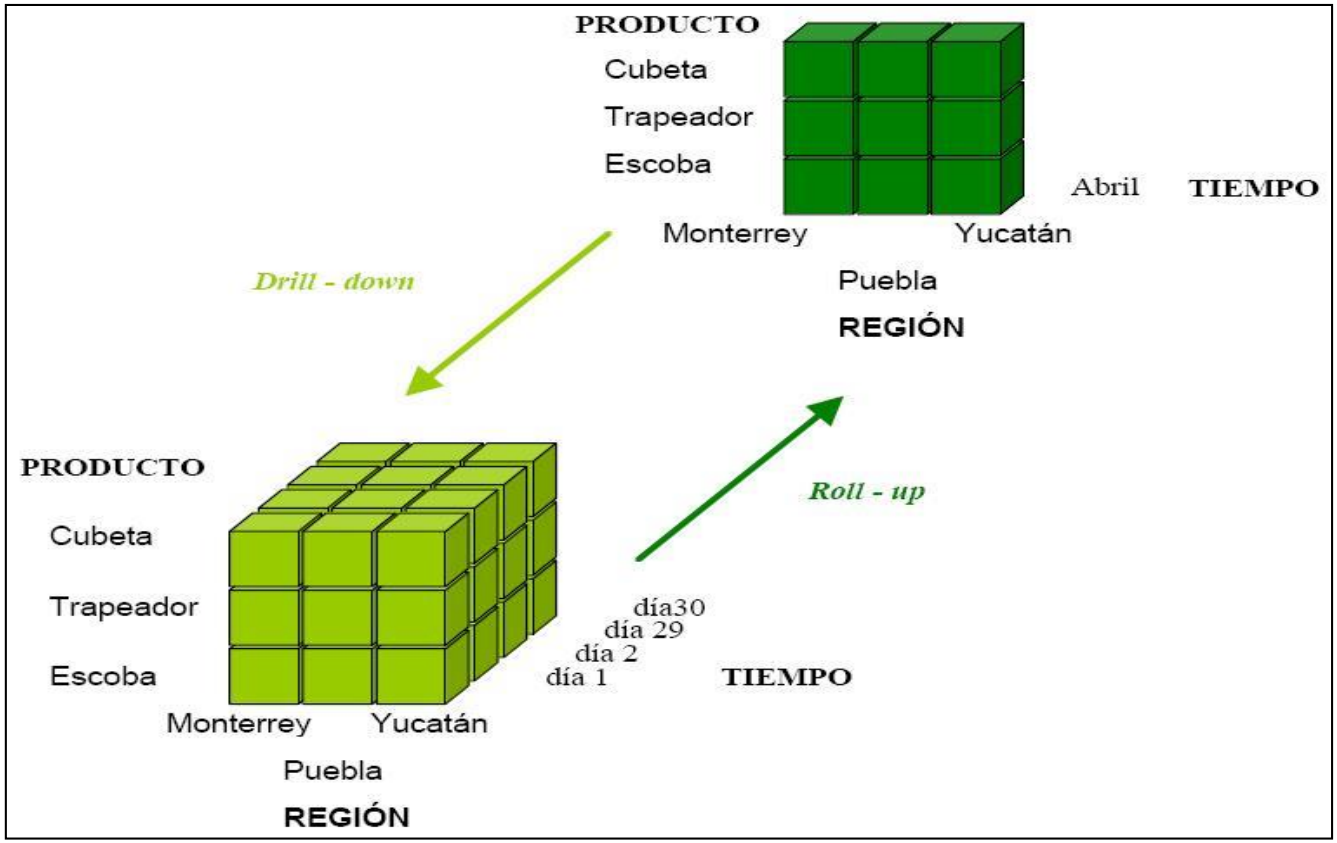
Facilitan el análisis de datos a través de dimensiones y jerarquías, uutilizando consultas rápidas predefinidas



Operaciones clásicas OLAP

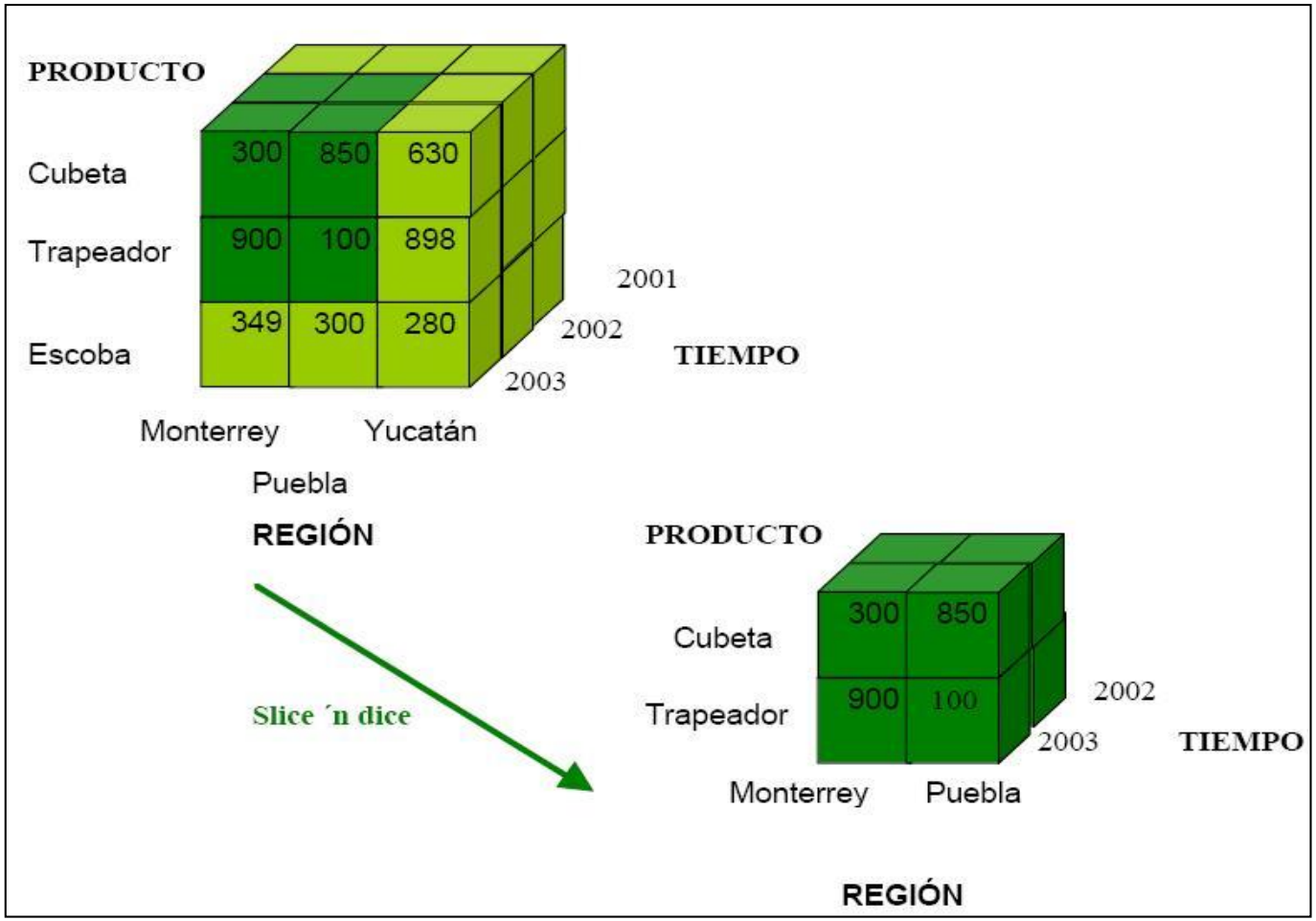
Roll up (drill-up): agrega medidas que van de un nivel Ni a un nivel más general Nj de una dimensión.

Drill down (roll down): es la operación inversa. A partir de un nivel superior este operador permitir bajar de nivel.



Operaciones clásicas OLAP

Slice and dice: permite restringir los valores asociados a una o varias dimensiones del cubo, es decir, toma un subconjunto de dimensiones y de niveles seleccionados del DW.



Otras operaciones

drill across
navegar a través de más de una tabla de hechos

drill through
navegar a través del nivel inferior del cubo a tablas relacionales

Pivote (rotar)
Rotar el cubo

Metodologías para realizar Analítica de Datos en una organización

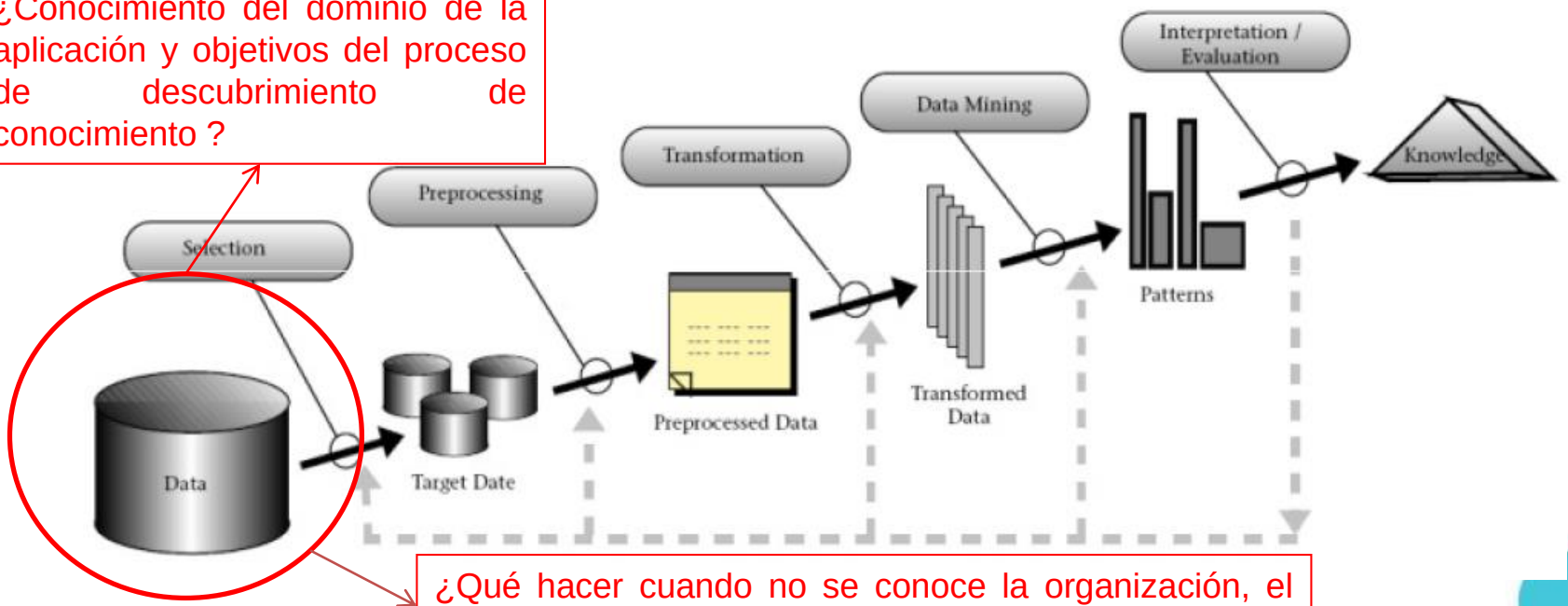
MIDANO

“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

Extendida para ser usado en el análisis de datos

MIDANO

¿Conocimiento del dominio de la aplicación y objetivos del proceso de descubrimiento de conocimiento ?

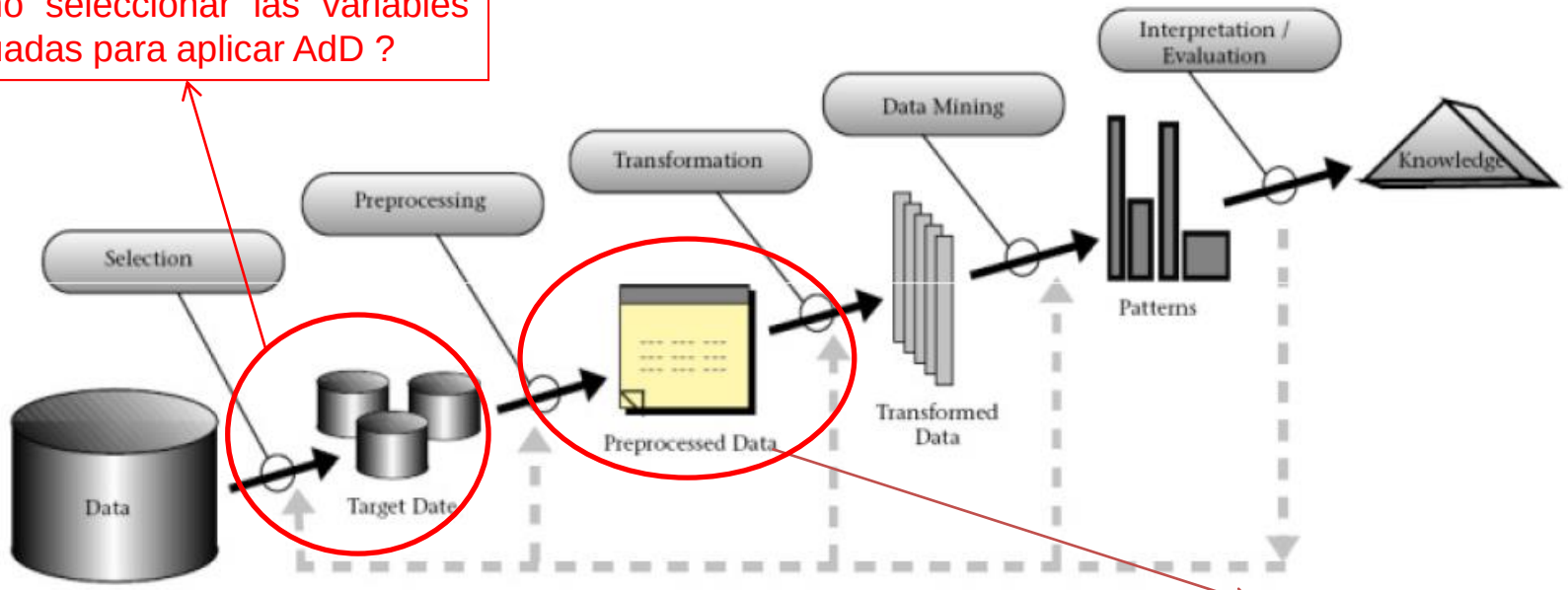


¿Qué hacer cuando no se conoce la organización, el problema, o los procesos a estudiar?

“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO

¿Cómo seleccionar las variables adecuadas para aplicar AdD ?



¿Cómo realizar el procesamiento de datos?

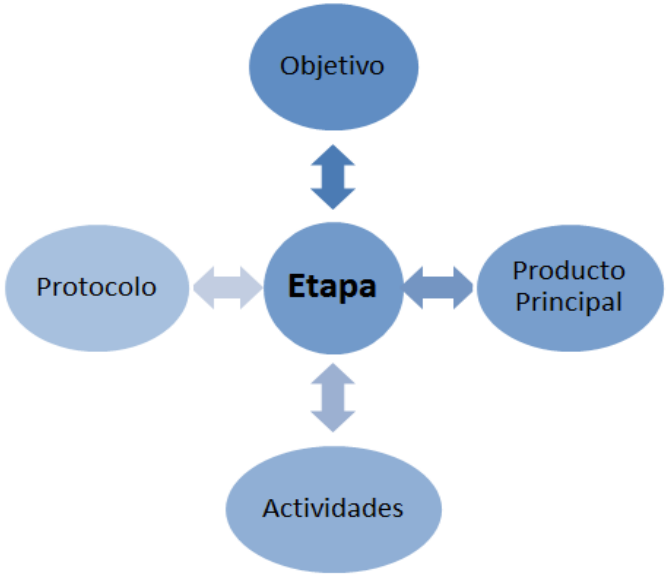
“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO-AdD

MIDANO consta de tres fases.

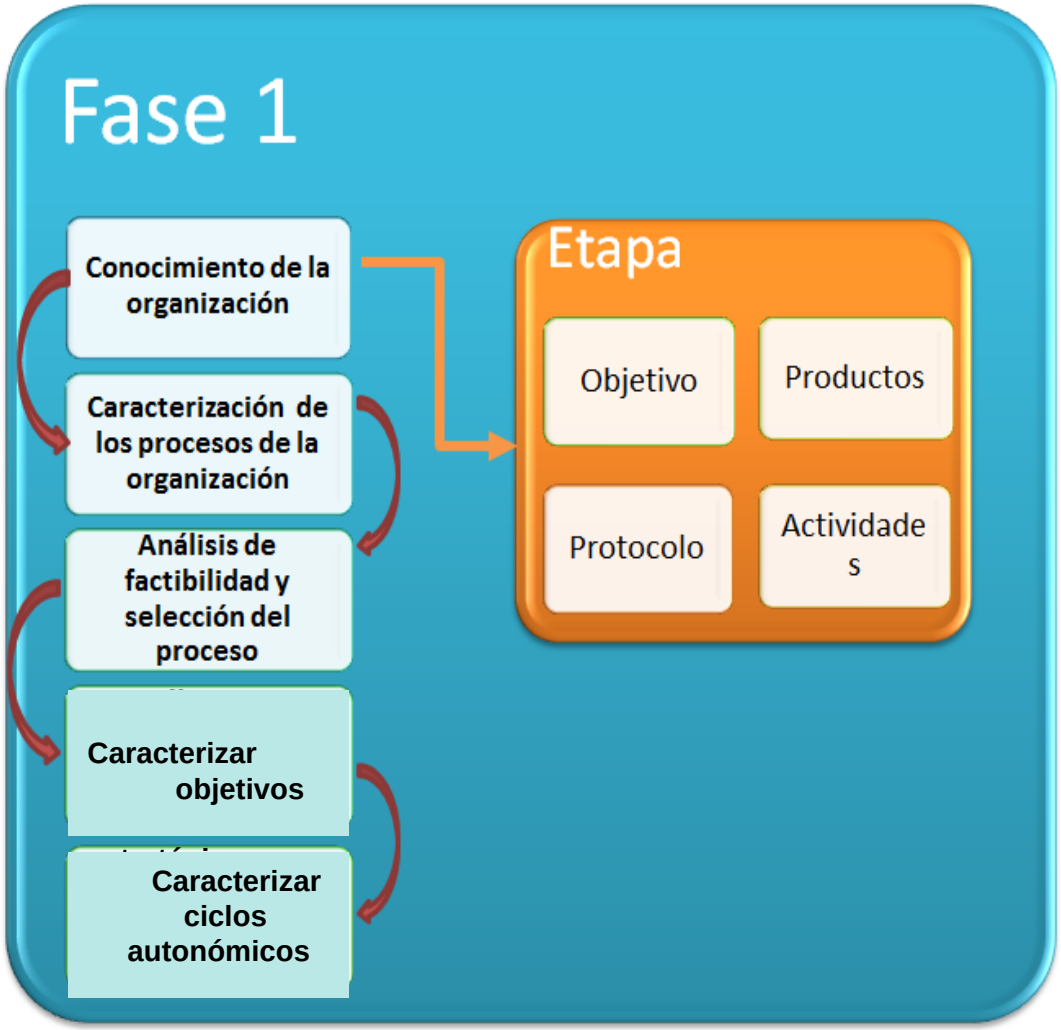
MIDANO

Cada fase de la metodología está dividida en etapas, los elementos que describen cada etapa. En cada etapa se especifican cuatro aspectos principales, como se describe a continuación.



Fase 1: Conocimiento de la Organización

Esta fase tiene como finalidad realizar un proceso de ingeniería de conocimiento, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s), o los procesos a estudiar.



Etapa 1: Conocimiento de la Organización

1. Objetivo

- Conocer la organización/empresa, sus objetivos, procesos, objetos y actores

2. Protocolo de la Fase:

- Descripción de los elementos de la institución/empresa y sus características. Objetivos, Procesos , Objetos y Actores.
- Descripción de las relaciones entre estos elementos.
- Organización de estos elementos.

Etapa 1: Conocimiento de la Organización

Preguntas y ejemplos para determinar los elementos de la institución/empresa

Elemento	Preguntas	Ejemplos
Objetivos	¿Cuál es la razón de ser de la institución?	Conocer, determinar, establecer, la finalidad de la institución/empresa.
Procesos	¿Cuales son las actividades que permiten alcanzar los objetivos de la institución?	Procesos de producción o administrativos.
Objetos	¿Qué cosas o entidades se manipulan en los procesos de la institución?	Pueden ser físicos o abstractos, departamentos, documentos, herramientas, plantas.
Actores	¿Quiénes ejecutan los procesos?	Personas, sistemas, máquinas, etc.

Etapa 2: Caracterización detallada de los procesos de la organización

1. Objetivo

- Conocer los procesos sobre los cuales se puede enfocar el proyecto de AdD.

2. Protocolo de la Fase:

- Familiarización con los procesos sobre los cuales se puede realizar la ingeniería de conocimiento
- Identificación de la fuente de conocimiento
- Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso.

Etapa 2: Caracterización detallada de los procesos de la organización

1. Familiarización con los procesos sobre los cuales se puede realizar la extracción de conocimiento

- ¿Qué productos generan esos procesos?
- ¿Qué beneficios proporcionan esos procesos a la organización?
- ¿Qué problemas tienen actualmente?
- ¿Importancia de esos procesos para la organización, o impacto sobre otros procesos?
- ¿Qué impacto generaría la mejora de esos procesos o el estudio de los mismos?

2. Identificar la fuente del conocimiento

- ¿Cuáles son los actores o personas que intervienen en los procesos?
- ¿Quién o quiénes son las personas expertas en los procesos?
- ¿Existen documentos que permitan conocer esos procesos?
- ¿Existen sistemas computacionales que intervengan o interactúen en el proceso?

3. Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso explicado

- ¿Dónde se encuentra los datos almacenados del proceso en cuestión?
- ¿Cómo se almacenan los datos del proceso?
- ¿Qué variables son observadas del proceso?
- ¿Cuáles son las variables más importancia de esos datos para la organización?

Etapa 3: Análisis de factibilidad y selección de los procesos

1. Objetivo

- Analizar los procesos con la información proporcionada/recogida, con la finalidad de conocer la factibilidad de la aplicación de la AdD sobre cada uno de ellos

2. Protocolo de la Fase:

- Revisión de los procesos propuestos por los expertos
- Disponibilidad del experto o grupo de expertos
- Análisis de las fuentes de información sobre los procesos

Etapa 3: Selección de los Procesos

Ejemplo de Tabla para selección de procesos

Peso	Criterios	Proceso 1	Proceso 2
	Importancia para la organización		
	Interacciones entre procesos		
	Procesos dependientes		
	Importancia de la calidad del producto		
	Seguridad Industrial		
	Proposito de la tarea de AdD		
	Replicabilidad de la herramienta a desarrollar		
	Cantidad de Expertos		
	Fuentes de información		
	Confidencialidad de la información		
	¿Qué información se recoge del proceso para ser almacenada?		
	Con que frecuencia se recoge la información almacenada		
	¿Qué herramientas se cuentan, para recolectar y manipular la información?		
	Total sin ponderación		
	Total ponderado		

Criterios vinculados a la importancia del proceso para la organización

Criterios vinculados a la factibilidad de hacer una Tarea de Análitica de Datos

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

1. Objetivo

- Caracterizar las posibles objetivos estratégicos a alcanzar, con las tareas de AdD, en los procesos seleccionados

2. Protocolo de la Fase:

- Descripción de los escenarios actuales de los procesos seleccionadas en la institución/empresa.
- Especificación de los objetivos estratégicos a alcanzar en esos procesos, y posibles escenarios futuros detrás de ellos.
- Especificación de los indicadores (modelos de conocimiento, medidas estadísticas, etc.) para el análisis e interpretación de los objetivos estratégicos
- Especificación de los requerimientos para los posibles escenarios futuros (donde se puedan aplicar tarea(s) de AdD)
- Elaboración de los casos de uso para los requerimientos funcionales

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Para los procesos seleccionados

Descripción del escenario actual

Resultados que se obtienen	Actor(es) asociado(s)	Variables Asociadas	Actividades que se realizan

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Para los procesos seleccionados:

todos sus posibles escenarios futuros

Escenarios futuros deben estar orientados a lograrlos

Métricas estadísticas, modelos de conocimiento, …

Descripción del escenario futuro

Objetivos Estratégicos a alcanzar	Actor(es) asociado(s)	Variables Asociadas	Actividades de AdD que se realizarían	Funcionalidades nuevas	Resultados que se desean obtener (indicadores de logro)

Descripción del escenario futuro: < xxx >

El conjunto de escenarios futuros define una

planificación estratégica tecnológica organizacional

Etapa 4: Análisis para caracterizar los objetivos estratégicos a alcanzar

Priorización de los escenarios futuros

Criterios	Escenario 1	Escenario 2	Escenario 3
Importancia del resultado que se espera del escenario para la empresa/institución			
Utilidad del escenario para la empresa/institución			
Cantidad de expertos asociados al escenario			
Seguridad Industrial (si aplica)			
Fuentes de información requeridas por el escenario			
Confidencialidad de la información			
¿Con que frecuencia se recogen los datos almacenados asociados a la información de interés?			
¿Con qué herramientas se cuenta para recolectar y manipular los datos?			
Replicabilidad de la herramienta a desarrollar en otros escenarios			

Vinculados a los **objetivos estratégicos** y su importancia

Etapa 5: Caracterización de los ciclos autonómicos de AdD para cada Objetivo Estratégico

1. Objetivo

- Especificación de los Ciclos Autonómicos (CA) para cada escenario futuro (objetivo estratégico) priorizado

2. Protocolo de la Fase:

- Determinación de las tareas de AdD que deben caracterizar a c/ciclo por sus roles
  - Tareas de monitoreo
  - Tareas de análisis
  - Tareas de toma de decisión
- Especificación de las relaciones entre ellas
- Especificación general de las fuentes de datos requeridas por cada tarea

Especificación del Ciclo Autonómico

Objetivo: Definir un objetivo válido de supremo interés para el proceso a estudiar.

Procedimiento General

Paso 1 Tareas de Monitoreo: Se identifican, capturan, pre-procesan, las variables del proceso bajo estudio, para poder tener una **observación** clara del proceso bajo estudio

Paso 2: Tareas de análisis: Se **interpretan** las situaciones que va aconteciendo en el proceso que se está estudiando, para comprenderlo, diagnosticarlo, analizarlo, entre otras cosas.

Paso 3 : Toma de decisiones: Se definen **acciones a tomar** sobre el proceso, con el fin de alcanzar el objetivo definido para el ciclo.

Etapa 5: Caracterización de los ciclos autonómicos de AdD para cada Objetivo Estratégico

Por cada ciclo autonómico

Objetivo estratégico a alcanzar: < … >

	Nombre	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre el objetivo estratégico
Tareas de AdD de Observación				
Tareas de AdD de Análisis				
Tareas de AdD de Toma de decisión				

Métricas estadísticas, modelos de conocimiento, … que produce

Usado en el futuro como métrica de calidad del CA

Relaciones entre las tareas del CA de AdD

	Tarea AdD1	Tarea AdD2	Tarea AdD13
Tarea AdD1			
Tarea AdD2			
Tarea AdD3			

Etapa 6: Especificación de las tareas de AdD

1. Objetivo

- Caracterizar general de las tareas de AdD a realizar en los CA especificados en la fase anterior (objetivos, requerimientos, etc.).

2. Protocolo de la Fase:

- Selección y descripción de los actores y componentes necesarios para hacer cada tarea de AdD.
- Especificación de los requerimientos de c/tarea de AdD: tecnológicos, de datos, organizacionales, etc.
- Especificación de las fuentes de datos requeridas por cada tarea

Etapa 6: Especificación de las tareas de AdD

Tabla para describir tareas de AdD

Nombre de la tarea	<nombre de la tarea>
Descripción	<La finalidad de esta tarea>
Fuente de datos	<BD, historicos>
Tipo de tarea de analítica de datos	<Asociacion, Agrupamiento, Clasificacion, Predicción, reglas de asociación, etc.>
Técnicas de analítica de datos	<Define las posibles tecnicas a usar, por ejemplo: regresión, redes neuronales artificiales, algoritmo K-NN, etc.>
Tipo de Modelo de Conocimiento	<modelo descriptivo, modelo prescriptivo, modelo de optimizacion, modelo predictivo, etc.>
Tareas relacionadas de analítica de datos	<Con que otras tareas de AdD se relaciona>
Tipo de tarea del ciclo autonómico (rol)	<Pueden ser para observar, analizar/interpretar, o actuar sobre el proceso>

Etapa 6: Especificación de las tareas de AdD

Tabla para detallar especificación tareas de AdD

Macro-Algoritmo	Especificar Tipo de Tarea de Minería	Herramienta
<paso a paso del código>	< Debe indicarse de manera concreta la tarea a realizar>	<Instrumento tecnológico a usar a utilizar para dicho calculo >
…	Por ejemplo, calcular una medida de centralidad de minería de grafo, realizar un agrupamiento de tales datos según tales criterios de similitud, etc.)	Por ejemplo, Netgraph o Netlogo para minería de grafo, o k-means para agrupamiento (indicando valor de k)
…		

Esta tabla es particularmente importante para las tareas de AdDS

Fase 2: Preparación de Datos

Para aplicar AdD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema en estudio.

Para realizar tareas de AdD es necesario tener los datos integrados en una sola vista, la cual comúnmente se conoce como *Vista Minable*. Existen dos tipos de vista minable:

- **Vista Minable Conceptual (VMC):** describe en detalle cada una de las variables a tomar en cuenta para c/tarea de AdD, en cada CA (proveniente de la primera fase de MIDANO).
- **Vista Minable Operativa (VMO):** Es el resultado de cargar los datos del historial y de realizar la etapa de tratamiento de datos, basado en la información de la VMC. La VMO se traduce a lo que se conoce como Vista Minable en la literatura, para realizar tareas de MD.

Con esas vistas se construye el modelo de datos multidimensional de c/CA

Fase 2: Preparación de Datos

- En esta fase se plantea realizar la preparación de los datos desarrollando dos etapas. Los productos más resaltantes de esta fase son las vistas minables (conceptual y operativa) y las variables objetivos, y el modelo de datos multidimensional.

Etapa 1: Definición del modelo de datos

a. Objetivos

- Ubicar y comprender los datos asociados a cada tarea de AdD
- Construir una VMC que tenga las variables de interés para el caso de estudio
- Construir una VMO inicial
- Definir la(s) variable(s) objetivo(s) asociadas a los objetivos estratégicos o a responder con las tareas de AdD
- Definir el modelo de datos multidimensional de cada CA

b. Protocolo de la etapa

- Comprender la fuente de datos de entrada
- Generar la VMC y la VMO inicial
- Integración de los datos de entrada
- Generar las tablas del modelo de datos multidimensional de cada CA

Etapa 1: Definición del modelo de datos

VMC

Variable	Descripción	Procedencia	Observaciones

modelo de datos multidimensional (tipo estrella)

Nombre	Nombre de la tabla de hecho
Claves a las tablas de dimensiones	Todas las claves a las tablas de dimensiones
Variables Objetivos	Variables que describen o se asocian al conocimiento extraído (predicciones, etc.)
Otras variables	Variables requeridas por la tarea de AdD, por ejemplo, derivadas de operaciones de procesamiento de las dimensiones o de OLAP

Nombre	Nombre de la tabla de dimensión
Claves de la dimensión	Clave de la dimensión
Atributos de la dimensión	Atributos que describen el tema asociado a esa dimensión

Etapa 1: Definición del modelo de datos

c. Productos principales

- Documento que describe las características de los repositorios donde se encuentran los datos
- Documento que describe la VMC, la cual es presentada en una tabla descriptiva.
- Vista minable operativa (modelo)
- Archivo donde esta almacenada la VMO
- Documento que describe las características de la(s) variable(s) objetivo(s)
- Modelo de datos multidimensional de cada CA
- Modelo de datos multidimensional (Constelación) del Data Warehouse

Etapa 2: Caracterización de los datos del dominio de la aplicación

a. Objetivos

- Construcción de la tabla con las operaciones de (E)xtracción, (T)ransformación y Carga (L), para las variables identificadas en la VMC
- Cargar los datos

b. Protocolo de la etapa

- Integración de los datos de entrada en el DW

c. Productos principales

- Tabla ETL

Etapa 2: Caracterización de los datos del dominio de la aplicación

Tabla ETL

Variable	Extracción	Transformación	Carga
Nombre de la variable	De que fuente de datos organizacional se extraera	Aqjui se indican todo el proceso de pre-procesamiento de los datos (estudios de dependencia, limpieza, cambio de formatos, etc.)	A que dimensión del modelo de datos irá

Esta tabla es de la misma longitud que la VMC

Etapa 3: Tratamiento de datos (ciencias de los datos)

a. Objetivos

Esta etapa se centra en generar datos de calidad, es decir, sin anomalías, sin inconsistencias de formato, sin capturas erróneas, sin campos vacíos; aplicando métodos de limpieza, transformación y reducción sobre la vista minable operativa.

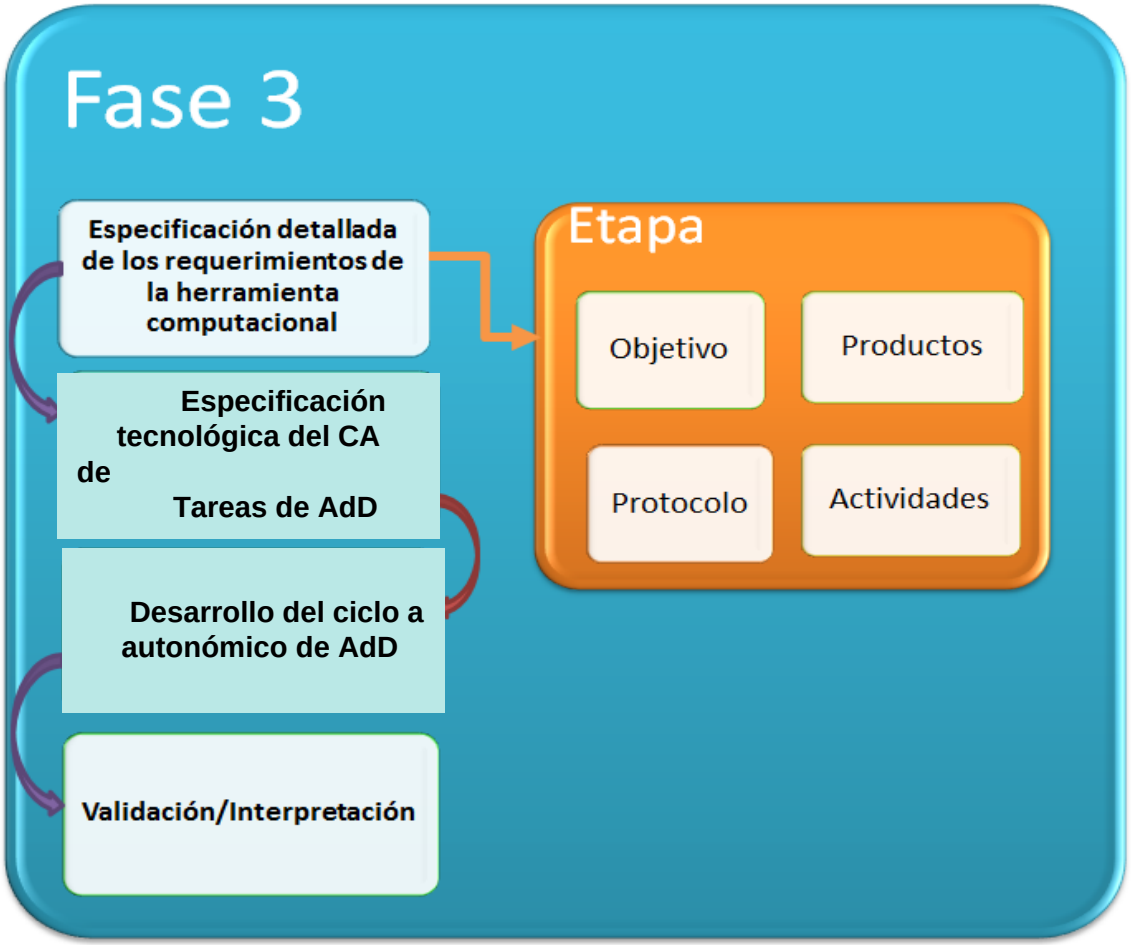
b. Protocolo de la etapa

- Limpieza
- Transformación
- Reducción

c. Productos principales

- VMO depurada
- DW implementada funcionalmente
- Documento descriptivo de los tratamientos realizados usando tablas descriptivas con información pertinente.

Fase 3: Desarrollo de las tareas de AdD



Etapa 1: Especificación detallada de los requerimientos de la herramienta computacional

a. Objetivos

captar los requerimientos no funcionales.

b. Protocolo de la etapa

- Requisitos de interfaz de usuario,
- Interfaces de software,
- Requerimientos de desempeño,
- Adicionalmente se pueden mencionar: de portabilidad, costos, rendimiento, accesibilidad, entre otros.

c. Productos principales

- Informe de requerimiento no funcionales

Etapa 2: Especificación tecnológica del ciclo autonómico de Tareas de AdD

a. Objetivos

Caracterización la implementación tecnológica del ciclo autonómico de tareas de AdD.

b. Protocolo de la etapa

- Escoger las técnicas de AdD para las tareas en el CA.
- Selección del Software para realizar c/tarea de AdD
- Definir cuáles son los datos de entrenamiento y de prueba contenidos en el DW a usar
- Definir las interfaces entre las tareas del CA
- Definir una estrategia para la validación de las técnicas seleccionada (cruzada, etc.).

c. Productos principales

- Documento con la especificación tecnológica del ciclo

Etapa 3: Desarrollo del ciclo autonómico de AdD

a. Objetivos

Realizar la herramienta de toma de decisiones usando el ciclo autonómico de tareas de AdD.

b. Protocolo de la etapa

- Construcción del modelo de conocimiento generado por cada tarea de AdD
- Repetir el procedimiento de ser necesario, hasta que el modelo cumpla los errores de entrenamiento establecidos
- Integrar las tareas de AdD en el CA

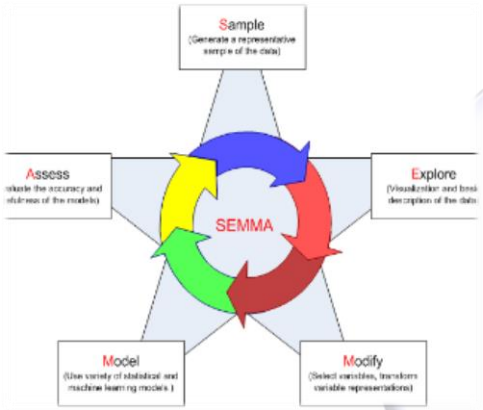
c. Productos principales

- Prototipo del CA

En esta etapa, se puede usar cualquier metodología de desarrollo de tareas de MD, para desarrollar las tareas de AdD.

Etapa 3: Desarrollo del ciclo autonómico de AdD

Desarrollo de las tareas de AdD



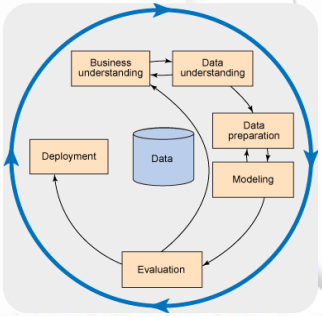
SEMMA

- Orientado a la parte técnica
- Carece de un análisis del problema.

Se puede usar cualquier metodología de desarrollo de MD para esta fase de desarrollo de tareas de AdD,

CRISP-DM

- Proceso continuo y progresivo del proceso de creación
- Más utilizado por empresas que trabajan con DM



CATALYST

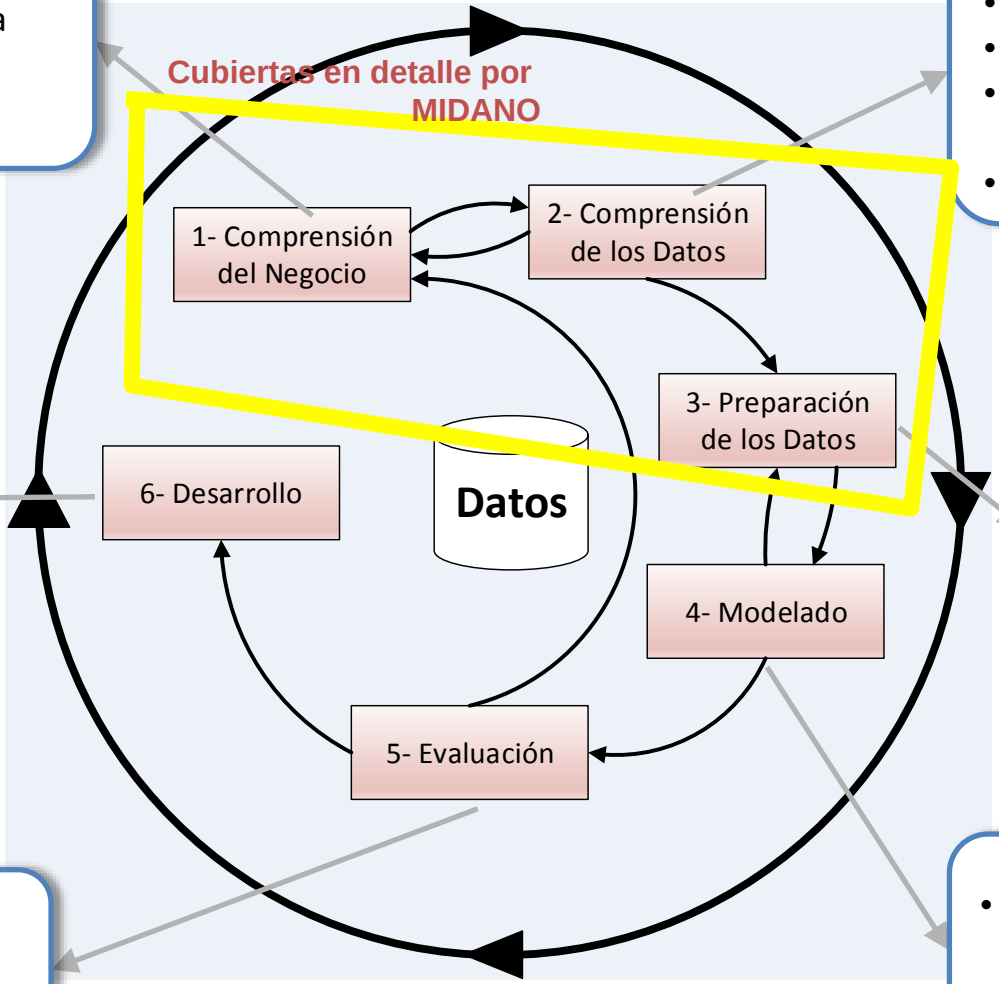
- Estructura en “boxes”
- Primer Modelo: Analiza el problema.
- Segundo Modelo: Solución en el aspecto técnico.

Etapa 3: Desarrollo del ciclo autonómico de AdD

Desarrollo de las tareas de AdD

CRISP-DM

- Objetivos y criterios de éxito del negocio y de la MD
- Plan del Proyecto.



- Análisis inicial de datos
- Recolección
- Descripción
- Identificación de problemas
- Verificación de calidad

- Plan para el desarrollo
- Informe final
- Presentación final
- Revisión general del proyecto

- Selección de datos
- Preparar, limpiar y/o construir datos
- Generar nuevos registros
- Integrar o formatear datos

- Evaluar el modelo
- Decisión sobre el modelo.

- Selección de técnica de modelado
- Obtener el modelo.

Etapa 4: Validación/Interpretación

a. Objetivos

Validar la herramienta de toma de decisiones.

b. Protocolo de la etapa

- Validar el modelo de conocimiento generado por cada tarea de AdD usando los datos de prueba, y siguiendo la estrategia de validación establecida (aplicarla y observar el rendimiento).
- Realizar las correcciones necesarias
- Repetir el procedimiento de ser necesario, hasta que el modelo cumpla los errores de prueba establecidos
- Validar el comportamiento del CA, usando los criterios definidos en la etapa 1.5
- Validar el comportamiento del CA, en el sistema de toma de decisión organizacional

Tipos de tareas de Analítica de Datos.

¿Qué es la AD?

- **Métodos Descriptivos**

Encontrar patrones interpretable que describen los datos.

- **Métodos de Predicción**

Utilizar algunas variables para predecir los valores desconocidos o futuros de otras variables.

MODELOS!!!

Modelos de Analítica

	Descriptivo	Predictivo	Prescriptivo
Preguntas	**Qué paso?** **Qué está pasando?** **Cuál es el problema?** **Qué acciones son necesarias?**	**Por qué esta pasando?** **Qué se producira?** **Por qué se producira?**	**Qué debería hacerse?** **Por qué debería hacerse?** **Qué pasa si se intenta eso?**
Habilitadores	• Reportes • Dashboards • Data Warehousing • Alertas	• Data Mining • Text Mining • Web/Media Mining • Forecasting	• Optimización • Simulación • Modelos de Decisión
Resultados	Bien definidos los problemas y oportunidades	Proyección de los futuros estados y condiciones	Mejores posibles decisiones y transacciones

Modelos de Analítica

	Optimización	Identificación	Diagnóstico
Preguntas	**Qué puedo mejorar? Cómo mejorarlo?**	**Cómo es el modelo? Qué caracteriza a esos modelos?**	**Por qué sucede? Cuáles son las causas?**
Habilitadores	• Reportes • Modelos de mejora • Simulación	• Simulación • Formulas matemáticas	• Optimización • Simulación • Modelos de Decisión
Resultados	Mejores en la organización	Caracterización	Mejores posibles decisiones y transacciones

Las estrategias analíticas básicas:

Describiendo

Factorización

Agrupación

Comparando

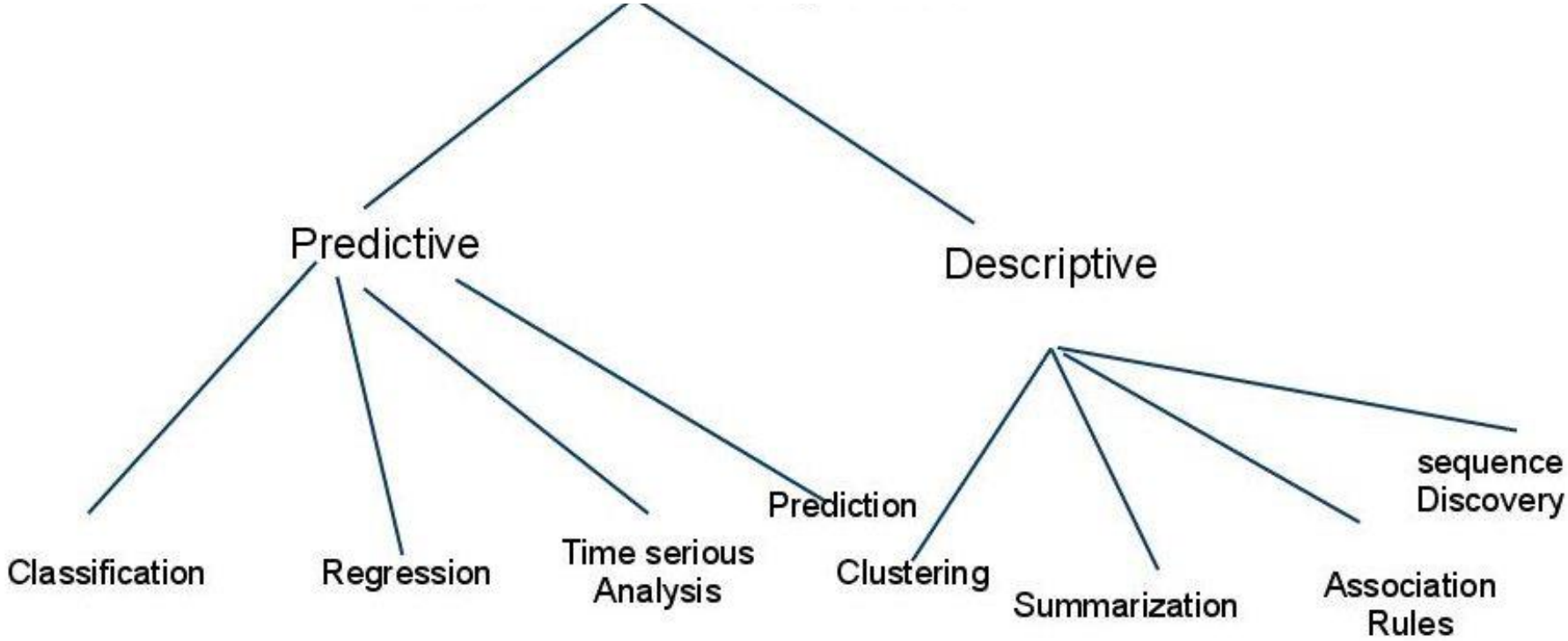
Clasificación

encontrar puntos comunes

encontrar covarianza

Descartar alternativas

Las estrategias analíticas básicas:



Clasificación

Examinar las características de un nuevo objeto y **asignarle** una clase o categoría de acuerdo a un conjunto de tales objetos previamente clasificados.

- Ejemplos:
  - Clasificar los estudiantes en categorías según sus rendimiento: bajo, medio y alto
  - Detectar los estados operacionales de un sistema: con falla, seguro, inactivo.

Clasificación

Email: Spam / No es Spam?
Transacciones en línea: Fraudulento (Si / No)?
Tumor: Maligno / Benigno ?

$$y \in \{0, 1\}$$

0: “Clase negativa” (tumor benigno)
1: “Clase positiva” (tumor malignano)

Clasificación

Obtener una función o modelo que determine la clase de un objeto basado en las características de sus atributos.

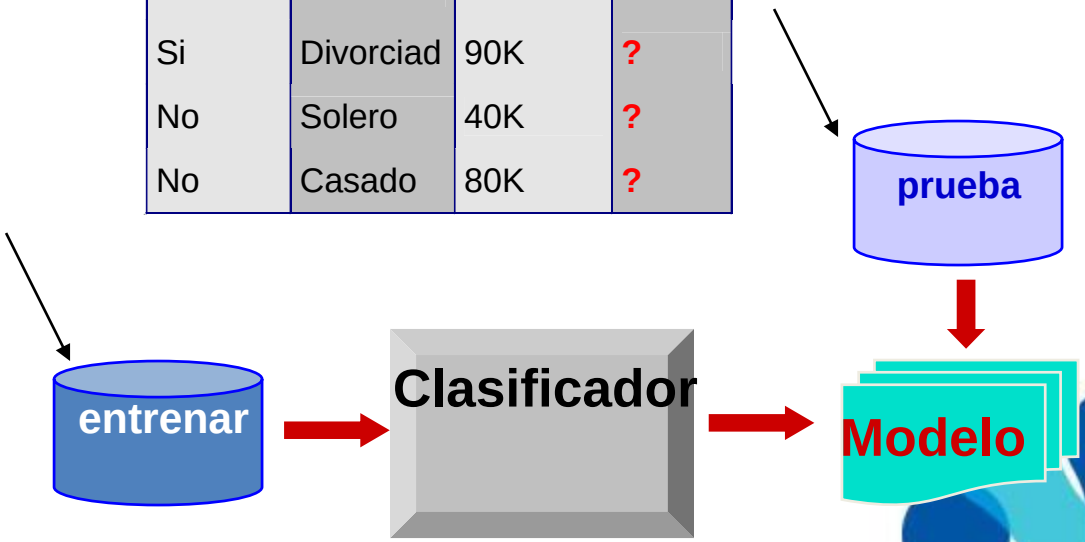
- Para generar dicho modelo o función, es necesario definir un conjunto de datos de entrenamiento, compuesto por objetos que ya tienen su clase asignada, también denominados ejemplos etiquetados.
- El modelo o función es creado analizando las relaciones entre los atributos de los objetos en el conjunto de entrenamiento y las clases.
- Mientras más variedad de escenarios se presenten en el conjunto de entrenamiento, mas se enriquece el modelo de clasificación (mejores resultados en la clasificación de nuevas entradas no etiquetadas).

Clasificación

	categoría	categoría	continuo	clase
ID	Reemb	Edo Civil	pago Impuest	Engaña
1	Si	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Si	Casado	120K	No
5	No	Divorc.	95K	Si
6	No	Casado	60K	No
7	Si	Divorciad	220K	No
8	No	Soltero	85K	Si
9	No	Casado	75K	No
10	No	Soltero	90K	Si

nominales numéricos

Reemb	Edo. Civil	pago Impuest	Engaña
No	Soltero	75K	?
Si	Casado	50K	?
No	Casado	150K	?
Si	Divorciad	90K	?
No	Solero	40K	?
No	Casado	80K	?

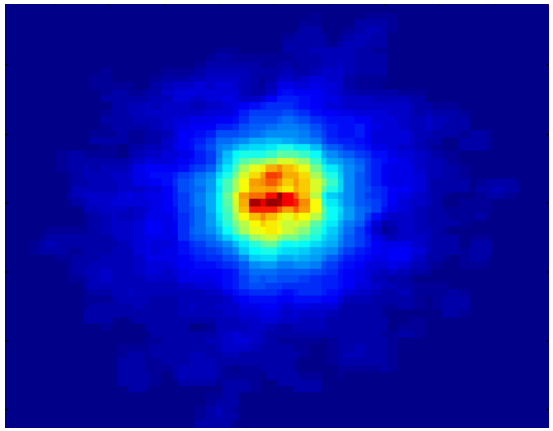


Clasificación

Clasificar cada imagen como una estrella (y su estado de formación) o una imagen no estelar (galaxia) (no-estelar)

http://aps.umn.edu

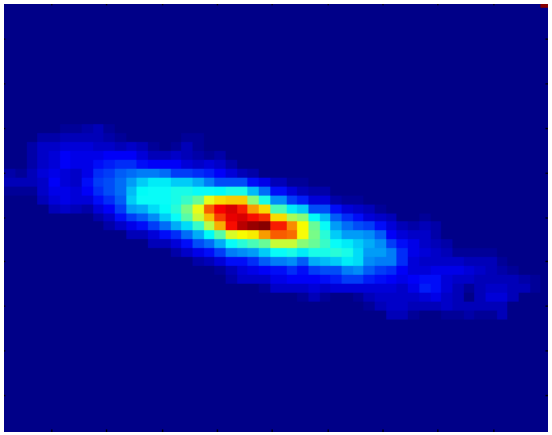
Temprano



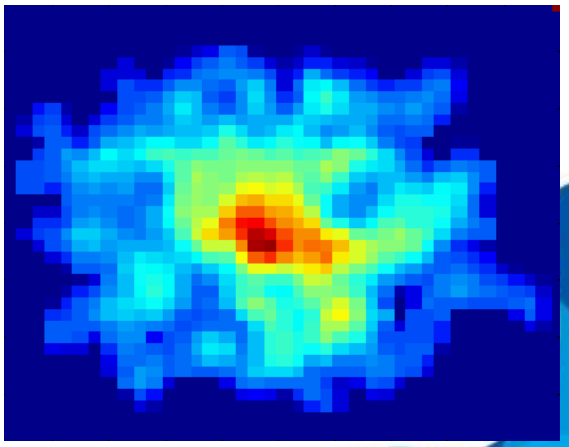
Clases:
- **Estado de Formación**

Atributos:
- **Caract. Imagen,**
- **Caract. Ondas, luz, etc.**

Intermedio



Tarde



Tam. datos:
- 72 millones estrellas, 20 millones galaxias
- BD Imagen: 150 GB

Agrupación o segmentación

Dividir una población en un número de grupos más homogéneos

- No depende de clases pre-definidas a diferencia de la clasificación

- Ejemplo:
  - Dividir la base de clientes de acuerdo con los hábitos de consumo
  - Establecer los grupos de estudiante según sus estilos de aprendizaje

Agrupamiento (Clustering)

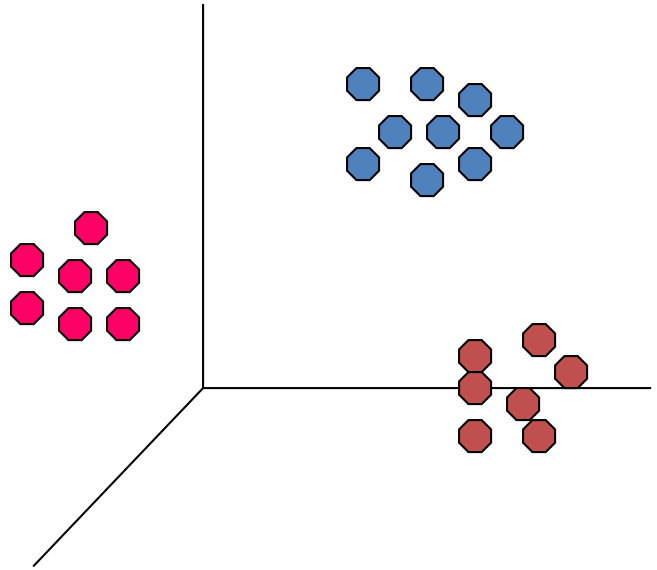
Dado un conjunto de datos, cada una con un conjunto de atributos, y una **medida de similitud** entre ellos, **encontrar grupos** de tal manera que:

- Los puntos de datos en un clúster son los **más similares** entre sí.
- Los puntos de datos en grupos separados son **menos similares** entre sí.

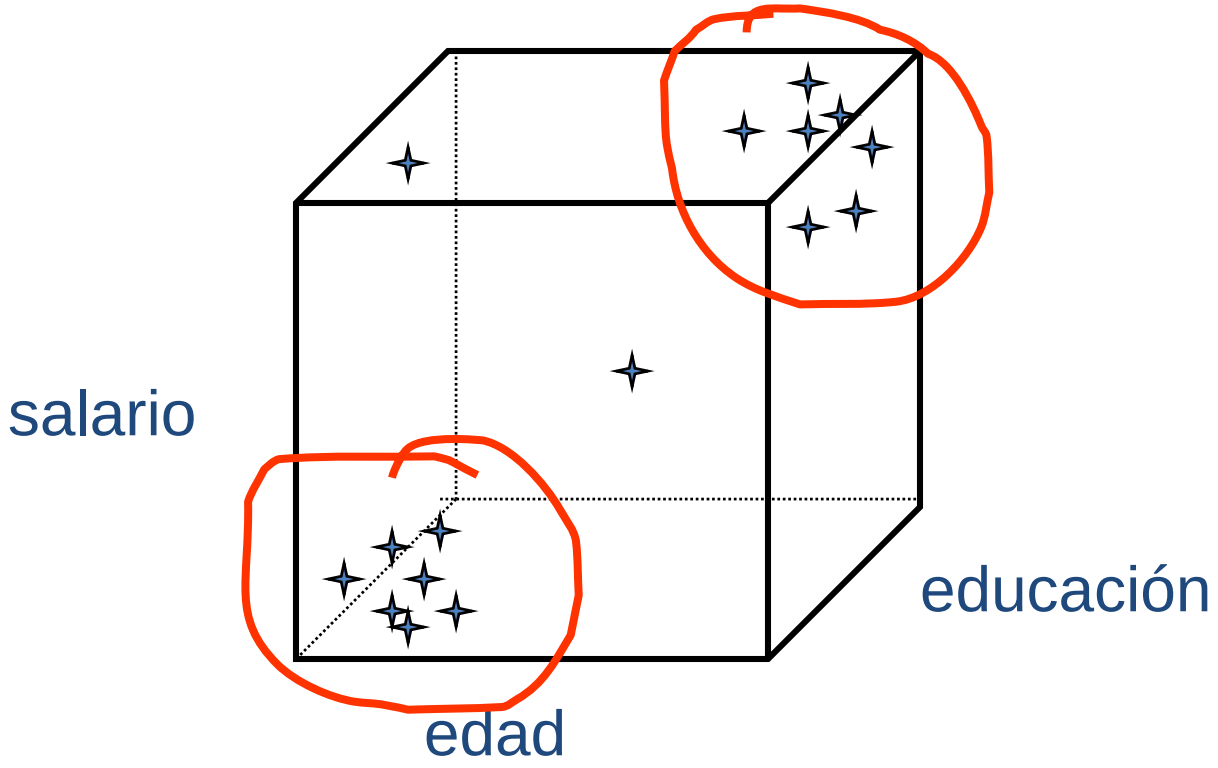
Agrupamiento (Clustering)

Distancias Intracluster son minimizadas

Distancias Intercluster son maximizadas



Agrupamiento (Clustering)



Ejemplo de Clustering

Agrupación de documento:

- **Objetivo:** encontrar grupos de documentos que son similares entre sí sobre la base de los términos importantes que aparecen en ellos.

- **Enfoque:** Identificar términos que aparecen con frecuencia en cada documento. Formar una medida de similitud basada en las frecuencias de los diferentes términos.

Agrupación de documento

- 3204 Artículos de un periódico.
- **Medida Similitud:** ¿Cuántas palabras son comunes en estos documentos (después de algún tipo de filtrado de palabras).

Categoría términos	*Total Articulos*	*Grupos*
Financiero	555	36
Extranjero	341	20
Nacional	273	6
Ciudad	943	76
Deportes	738	73
Entretenimiento	354	28

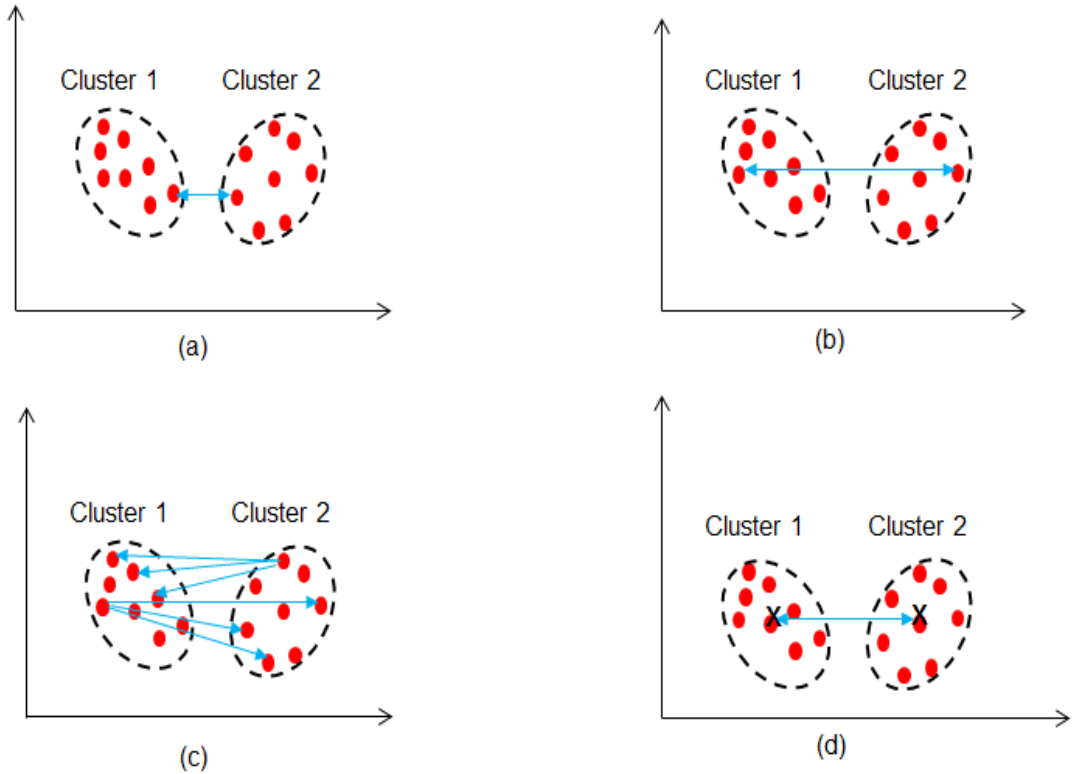
Tipos de clustering: basados en distancia

(a) Distancia mínima

(b) Distancia máxima

(c) Distancia de promedio del grupo

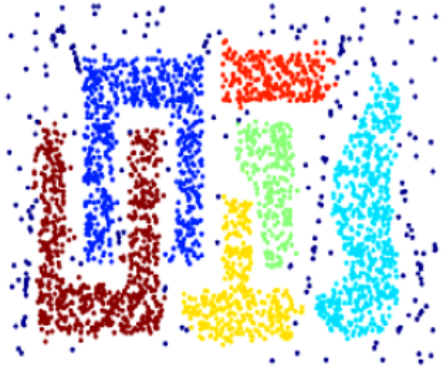
(d) Distancia con respecto al centroide



Tipos de Clustering:
Basados en densidad

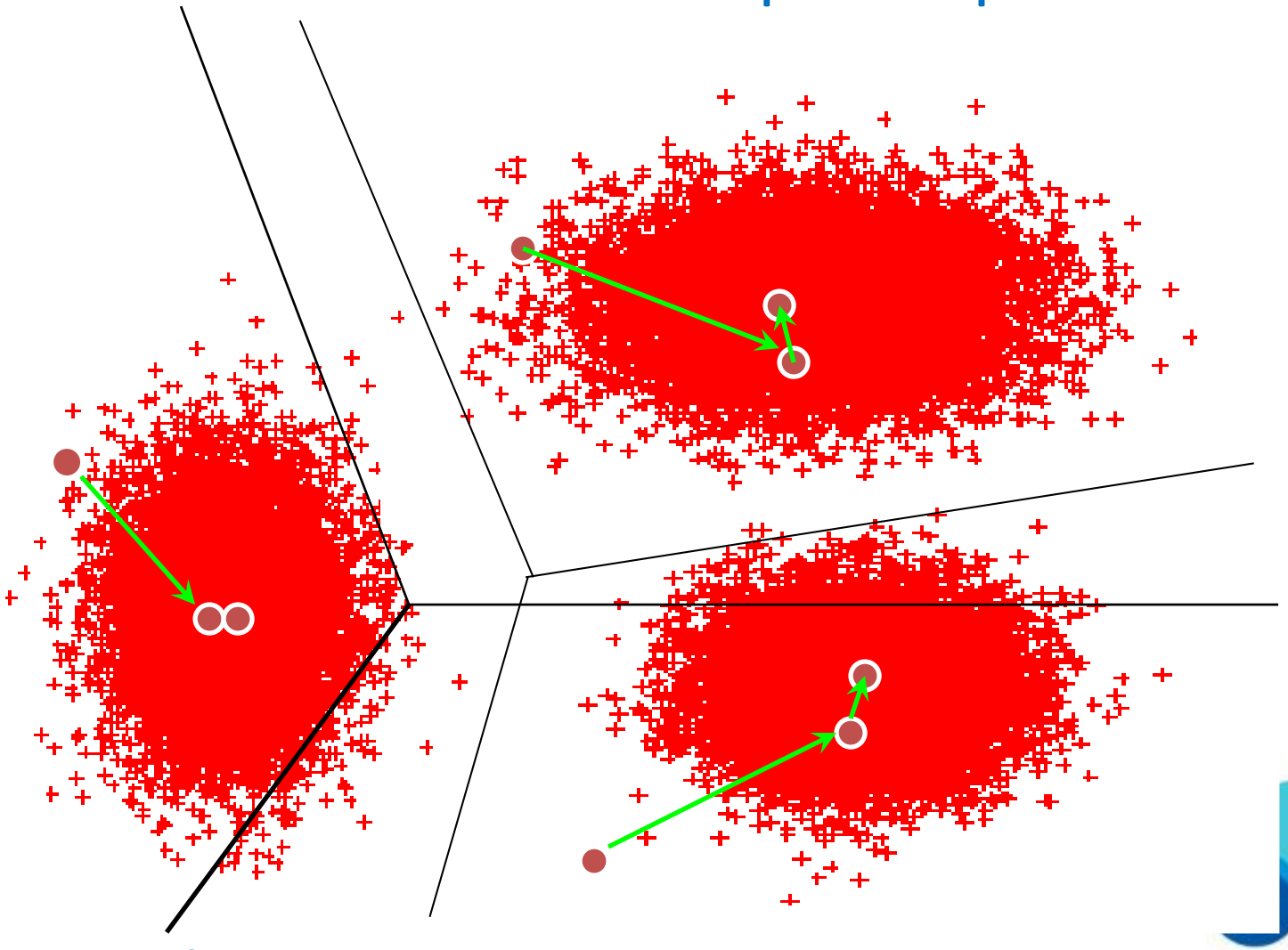
Los algoritmos basados en densidad, tratan de formar agrupaciones en áreas con altas densidades de ejemplos

a) Datos originales



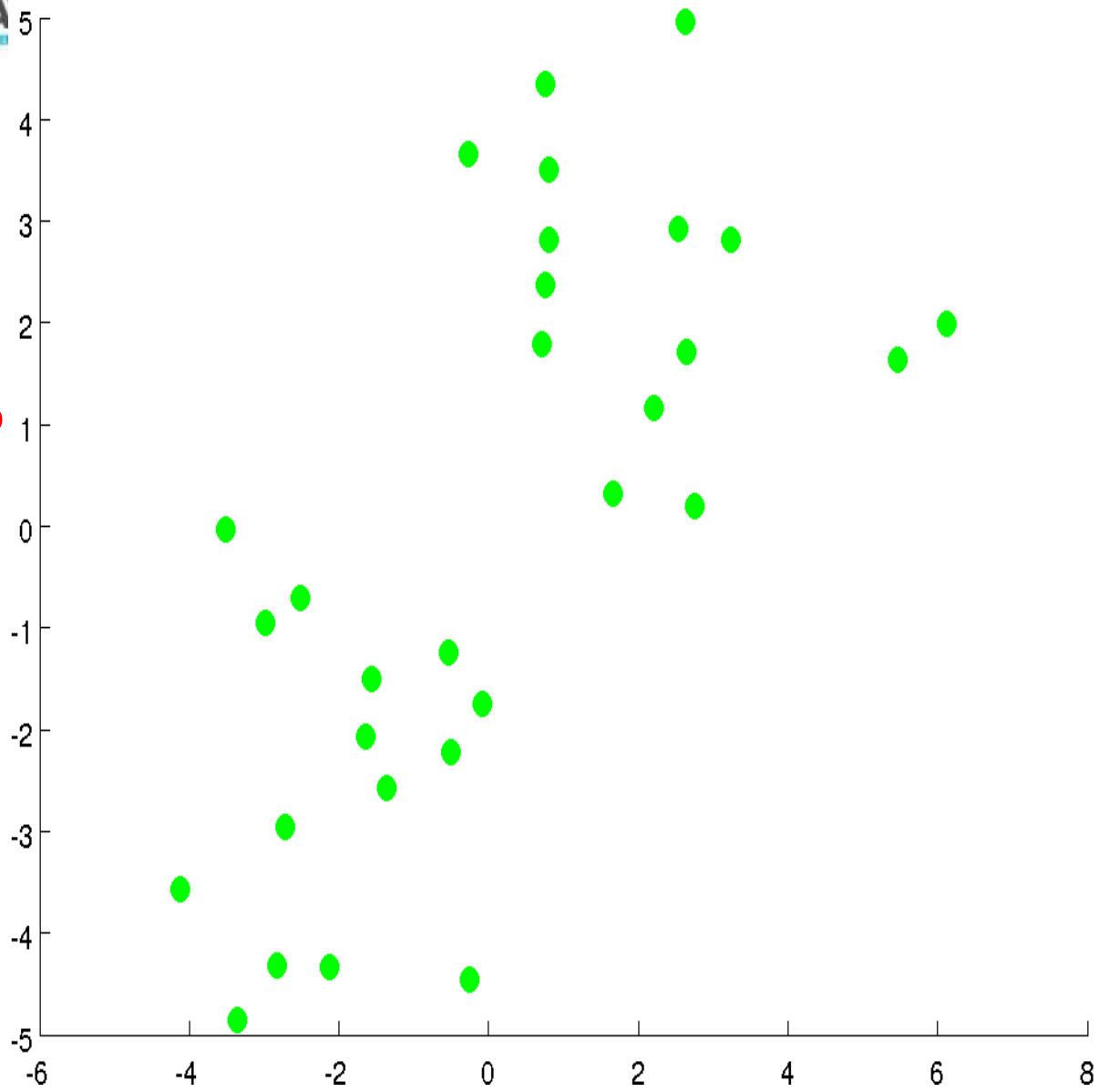
b) Datos después de clustering

Tipos de Clustering:
Basados en prototipo

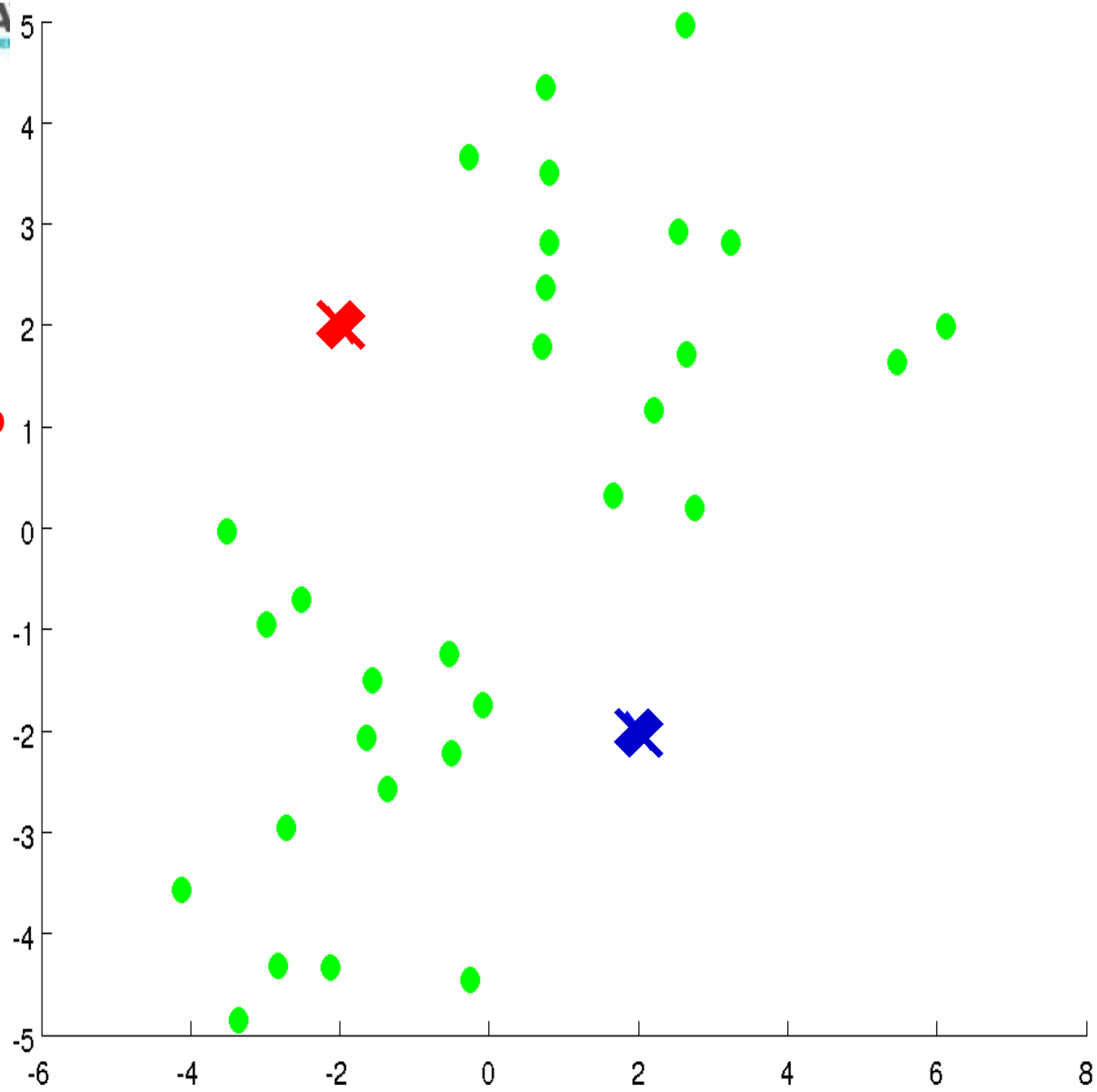


Algoritmo K-Medias

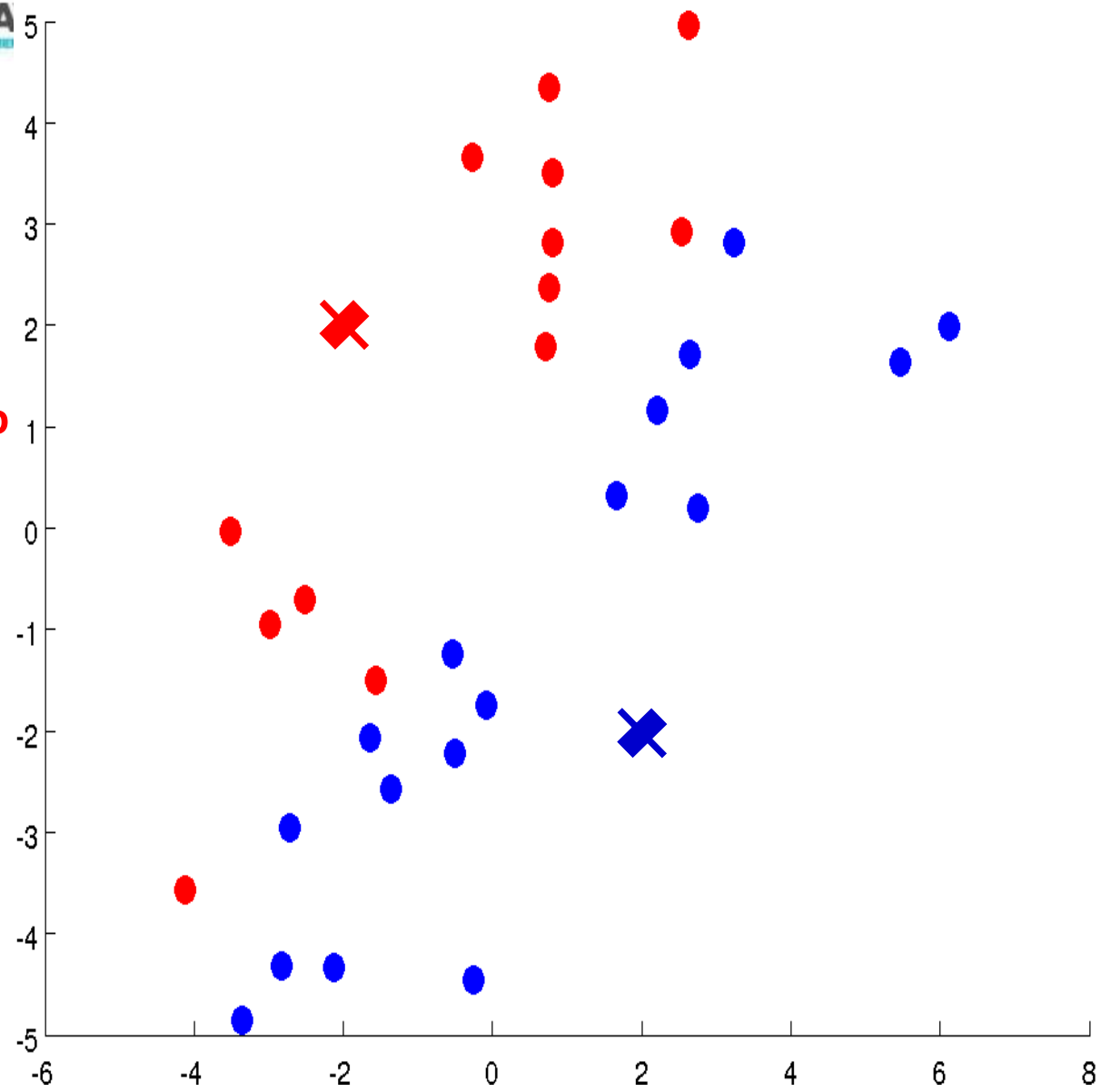
Corrida en frio
K-means



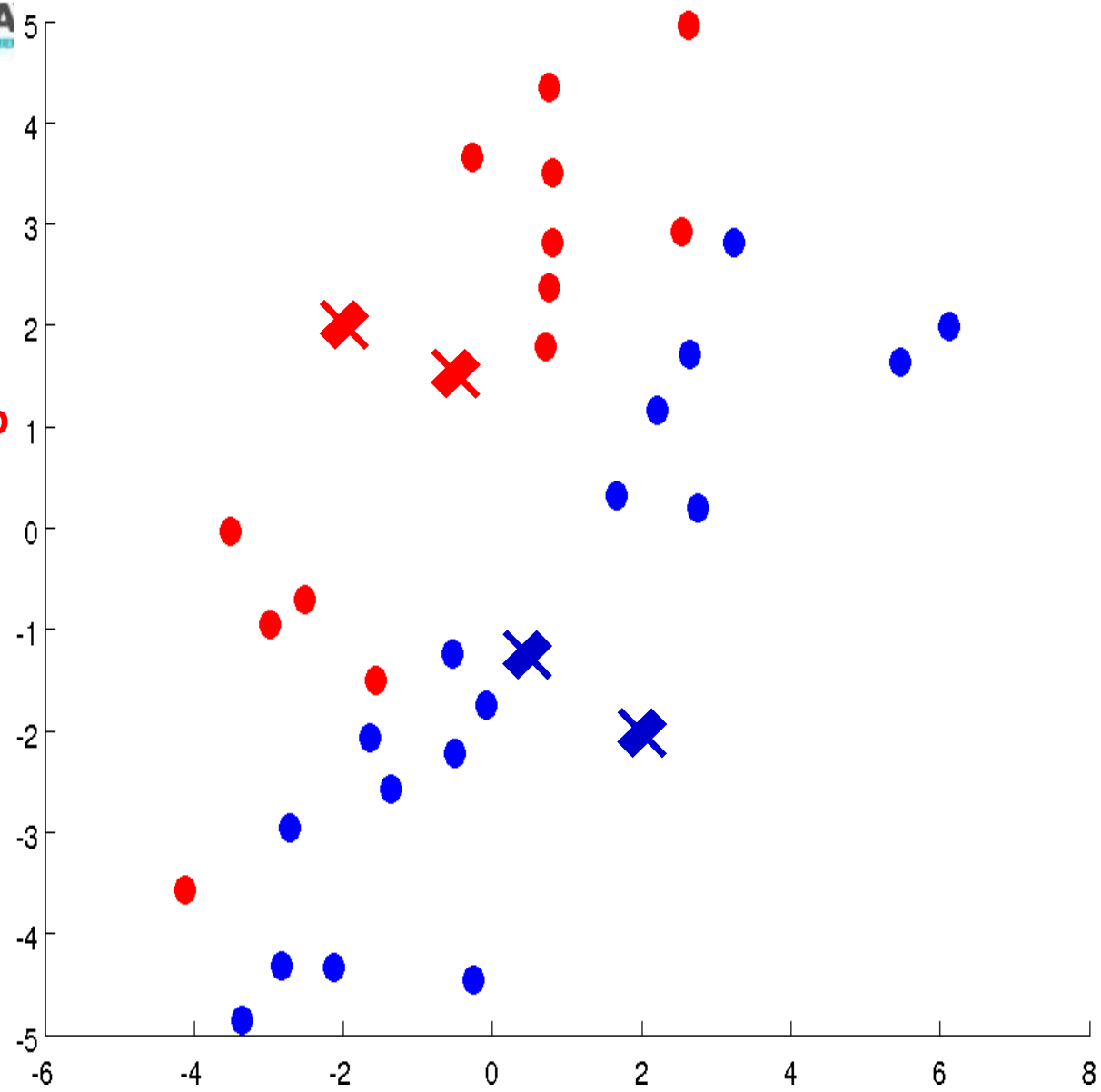
Corrida en frio
K-means



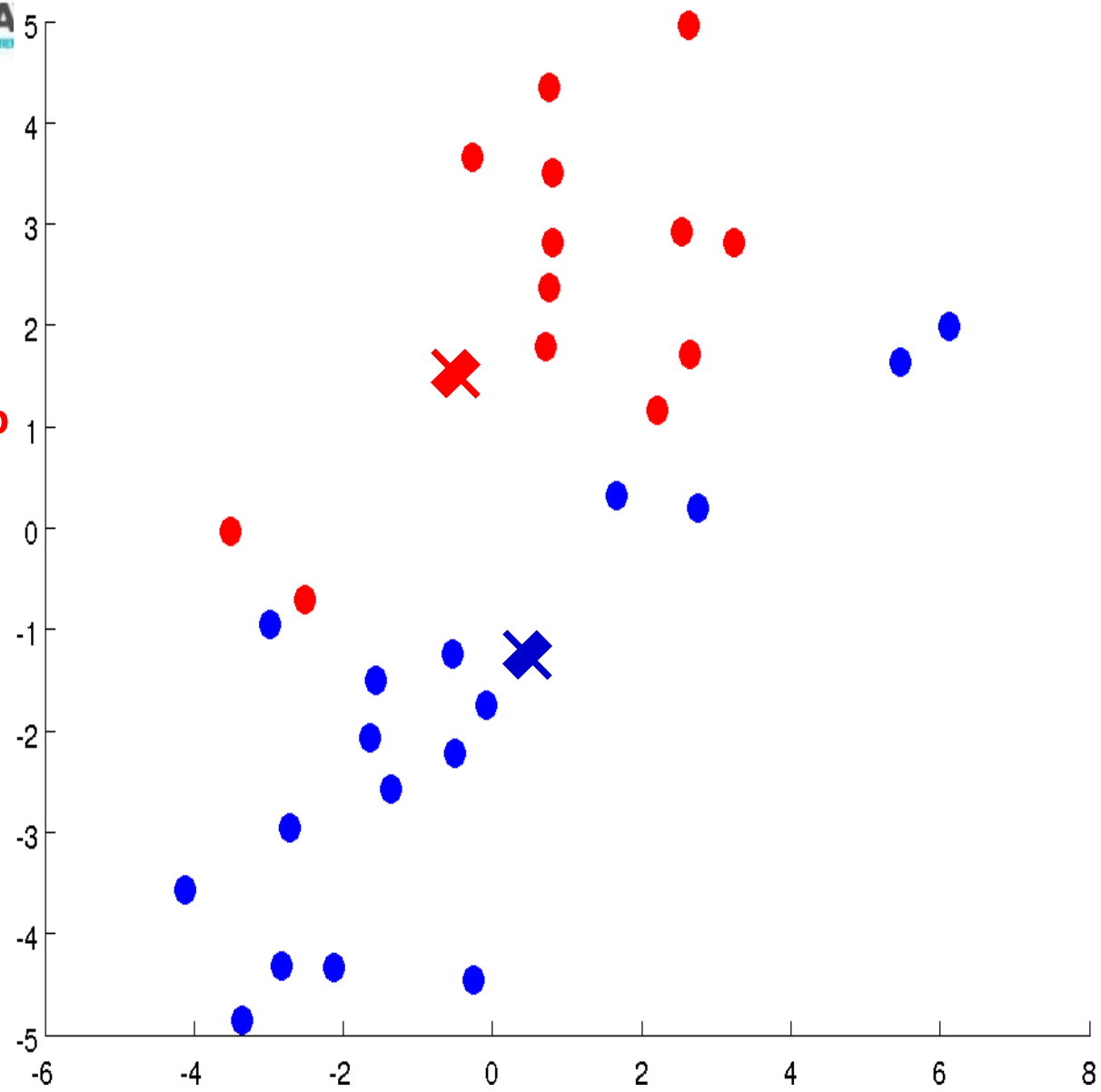
Corrida en frio
K-means



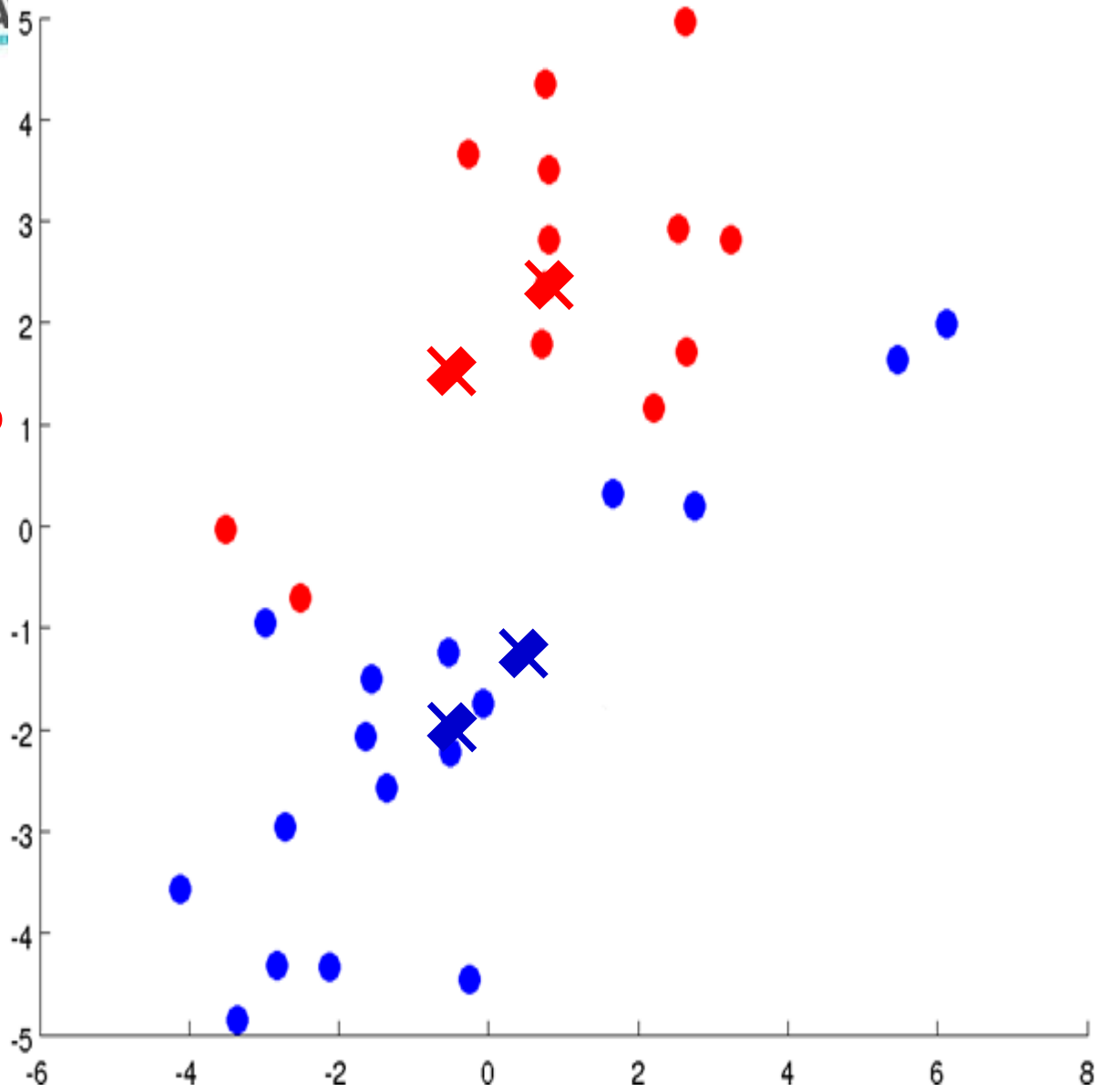
Corrida en frio
K-means



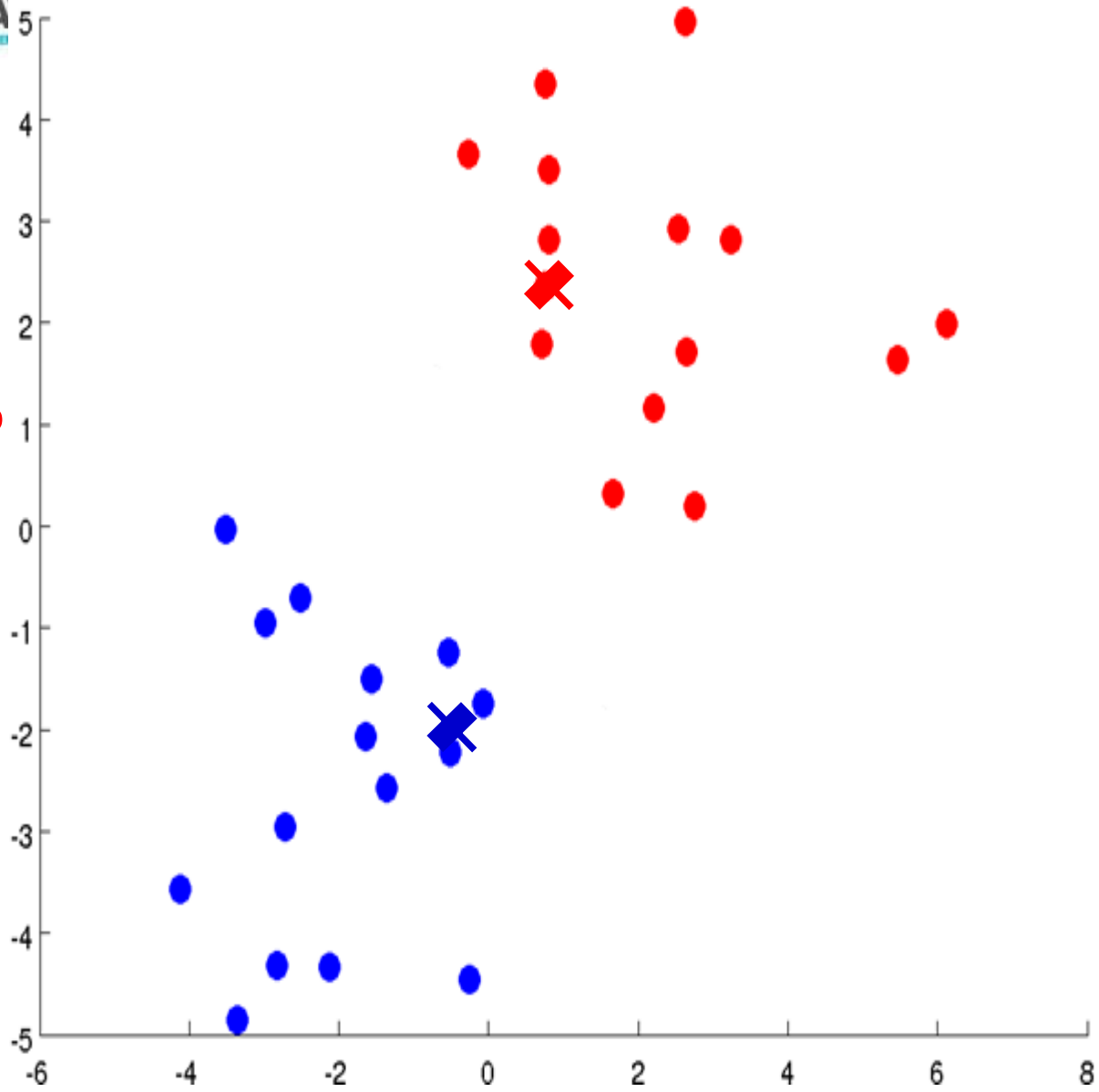
Corrida en frio
K-means



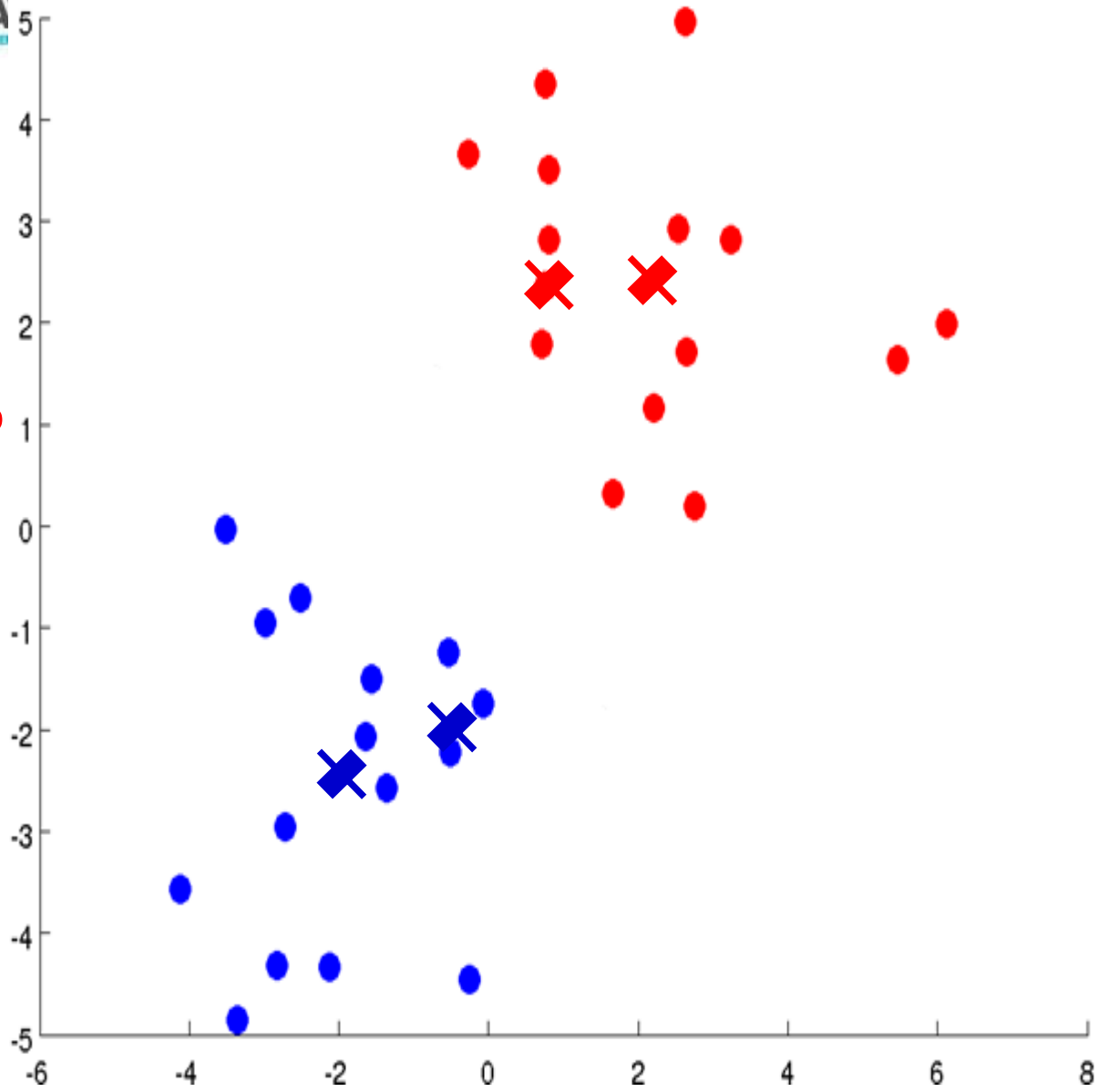
Corrida en frio
K-means



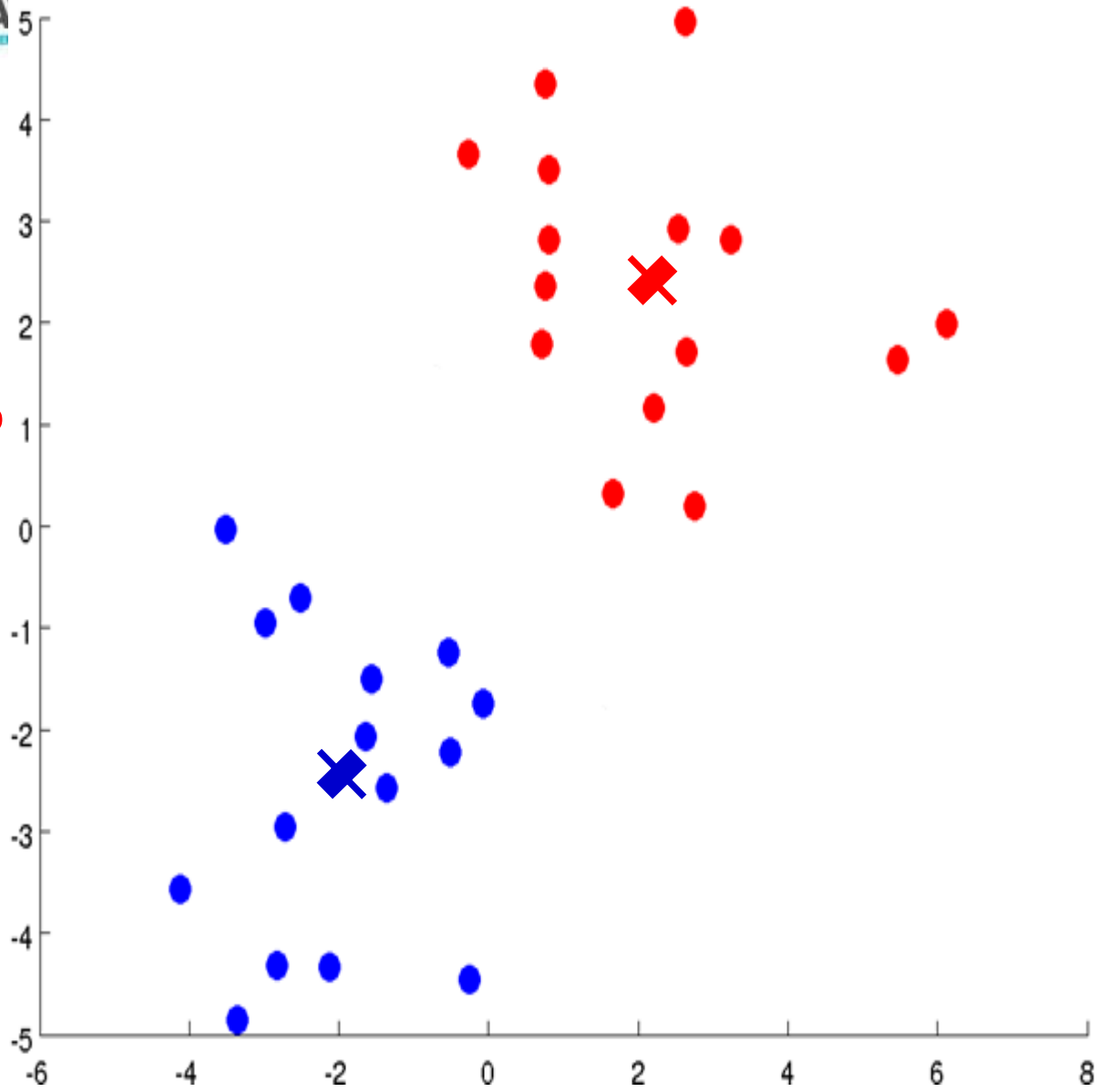
Corrida en frio
K-means



Corrida en frio
K-means



Corrida en frio
K-means



Asociación

Determinar cosas u objetos que van juntos

- Ejemplo:
  - Determinar **qué productos se adquieren conjuntamente en un supermercado**

ASOCIACION

Es el descubrimiento de relaciones entre las características (atributos) que conforman la base de datos,

Dichas asociaciones es el *conocimiento*

REGLAS DE ASOCIACION

Técnica no supervisada que permite predecir patrones de comportamientos futuros **basado en las ocurrencias simultaneas** de valores de variables.

Una asociación entre dos o más atributos ocurre cuando la **frecuencia con la que se dan dos o más valores determinados de cada uno conjuntamente es relativamente alta**.

Las reglas de asociación intentan descubrir asociaciones o conexiones entre objetos.

$$Consecuencia \Leftarrow Antecedente_1\ Antecedente_2\ \ldots\ Antecedente_m.$$

Ejemplo, en un supermercado se analiza si los pañales y las compotas se compran conjuntamente.

REGLAS DE ASOCIACION: ejemplo

Gestión Estantes de un supermercado

- **Objetivo:** Identificar los elementos que compran juntos muchos clientes.
- **Enfoque:** encontrar dependencias entre elementos.
- **Un ejemplo de regla:**
  - Si un cliente compra pañales y leche, entonces es muy probable que compre compotas.

Reglas de Asociación

Reglas que implican relaciones

Sombreado: parado (standing)
No sombreado: acostado (lying)

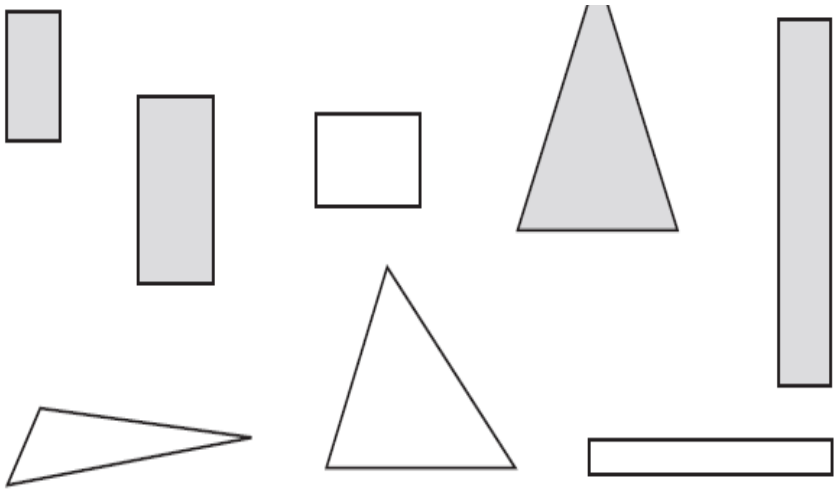


Tabla con datos de entrenamiento

Width	Height	Sides	Class
2	4	4	standing
3	6	4	standing
4	3	4	lying
7	8	3	standing
7	6	3	lying
2	9	4	standing
9	1	4	lying
10	2	3	lying

Reglas

if width ≥ 3.5 and height < 7.0 then lying
if height ≥ 3.5 then standing

Predicción

- Predice un valor de una variable dada, sobre la base de los valores de otras variables, suponiendo un modelo lineal o no lineal de dependencia.

- **Ejemplos:**
  - Predecir las ventas de nuevos productos basados en gastos de publicidad.
  - Predecir la velocidad del viento como una función de la temperatura, humedad, presión de aire, etc.
  - Predecir comportamiento en el tiempo de los índices bursátiles (series de tiempo).

Modelos de Predicción

Piensa en una variable que quieras predecir. ***Que necesitas?***

- **Objeto a predecir**: Una serie temporal, un suceso, …etc.
- **Formato de la Predicción**: Puntual, Intervalo, Densidad, …etc.
- **Horizonte de la predicción**: Corto, Medio o Largo Plazo
- **Conjunto de Información**: Univariante o Multivariante
- **Metodos y Complejidad**: Modelos, …etc.

Modelos de Predicción

Las predicciones ayudan a la toma de decisiones en una gran variedad de áreas.

- **Planificación y Control de Operaciones:** Las empresas usan predicciones para decidir que producir, cuando y donde.
- **Mercadeo:** Decisiones de precios, de gastos en publicidad, …dependen fuertemente de las previsiones que se tengan sobre como van a responder las ventas a los diferentes esquemas de marketing.
- **Economía:** Predicciones de las variables macro-económicas claves como el PNB, Paro, Consumo, Inversión, Tipos de Interés, etc… son usadas por el gobierno para fijar su política monetaria y fiscal.
- **Financiera**: Actores de los mercados financieros tienen un gran interés en la predicción de los rendimientos de activos financieros (acciones, tipos de interés, tipos de cambio, etc…).
- **Demografía:** La predicción de la población es crucial para planificar el gasto publico en sanidad, infraestructuras, educación, etc.

Modelos de Predicción

Hay muchas formas de hacer predicciones; pero todas ellas tienen en común los siguientes ingredientes:

1. ***que hay ciertas regularidades que captar***
2. ***que tales regularidades son informativas sobre el futuro***
3. ***están encapsuladas en el método seleccionado para predecir***
4. ***normalmente se excluyen las no-regularidades***

Los principales métodos son:

- **Adivinación**
- **Extrapolación**
- **Encuestas**
- **Modelos de Series Temporales**

Evaluación de Prediciones

Las medidas mas comunes de la precisión de la predicción son:

Error cuadrático medio:

$$MSE = \frac{1}{T}\sum_{t=1}^{T} e_{t+h}^2$$

Raíz cuadrada del MSE

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{t+h}^2}$$

Error absoluto medio

$$MAE = \frac{1}{T}\sum_{t=1}^{T} | e_{t+h} |$$

donde $e_{t+h}=y_{t+h}-y'_{t+h}$ son los errores de predicción.

Predicción

Evaluación de la capacidad de predecir

- Dividir la muestra en dos partes;
  - una para estimación del modelo
  - una para evaluar la capacidad de predecir.
- Estimar el modelo
- Calcular la predicción para los periodos no usadas.
- Comparar la predicción con los valores reales (error del pronóstico)

Técnicas de Analítica de Datos

Ejemplos de Técnicas

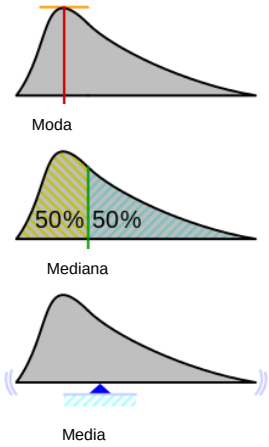
- El análisis estadístico

Dos categorías principales:
* Estadísticas descriptivas
* Estadística inferencial

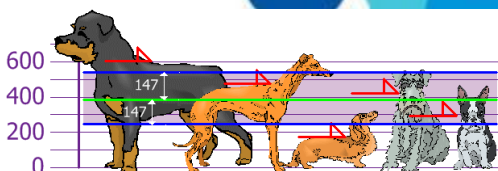
- El análisis predictivo
- La Correlación
- La Regresión
- Computación Inteligente (machine learning)

Métodos Estadísticos

- Usar **medidas de resumen** para describir la tendencia central de una distribución (media, moda, mediana)
- Utilizar la **dispersión o variabilidad** (desviación estándar, varianza, y el rango) para saber cómo se extienden los datos alrededor de la media.



- Frecuencias (contar)
- Porcentaje
- Media (suma de todos los valores ÷ no. de valores)
- Moda (valor más frecuente)
- Mediana (valor medio o posición central)
- Rango (intervalo entre el valor máximo y minimo)
- Desviación estándar (variación esperada con respecto a la media)
- Varianza (la esperanza del cuadrado de la desviación)
- Rankeo (clasificar, ordenar)



Compradores	Número
Hombre	
Viejo	6
Joven	4
Mujer	
Vieja	10
Joven	15

- **Más compradores femeninos que compradores masculinos**
- **Más jóvenes compradores femeninos que los compradores varones jóvenes**
- **Compradores masculinos jóvenes no están interesados en comprar en el centro comercial**

Técnicas de Aprendizaje Automático:

- Es imposible prever todos los problemas desde el principio
- Un **sistema es inteligente** si es capaz de observar su entorno y aprender de él
- La **autentica inteligencia** reside en adaptarse, tener capacidad de integrar nuevo conocimiento, resolver nuevos problemas, aprender de los errores, etc.

Aprendizaje Automático (Machine Learning en inglés) es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender.

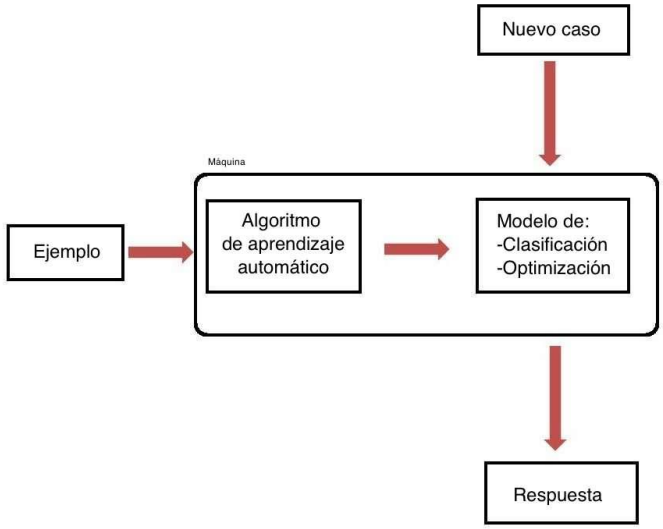
Crear algoritmos capaces de generalizar comportamientos y reconocer patrones.

Dar a los programas la capacidad de adaptarse sin tener que ser reprogramados

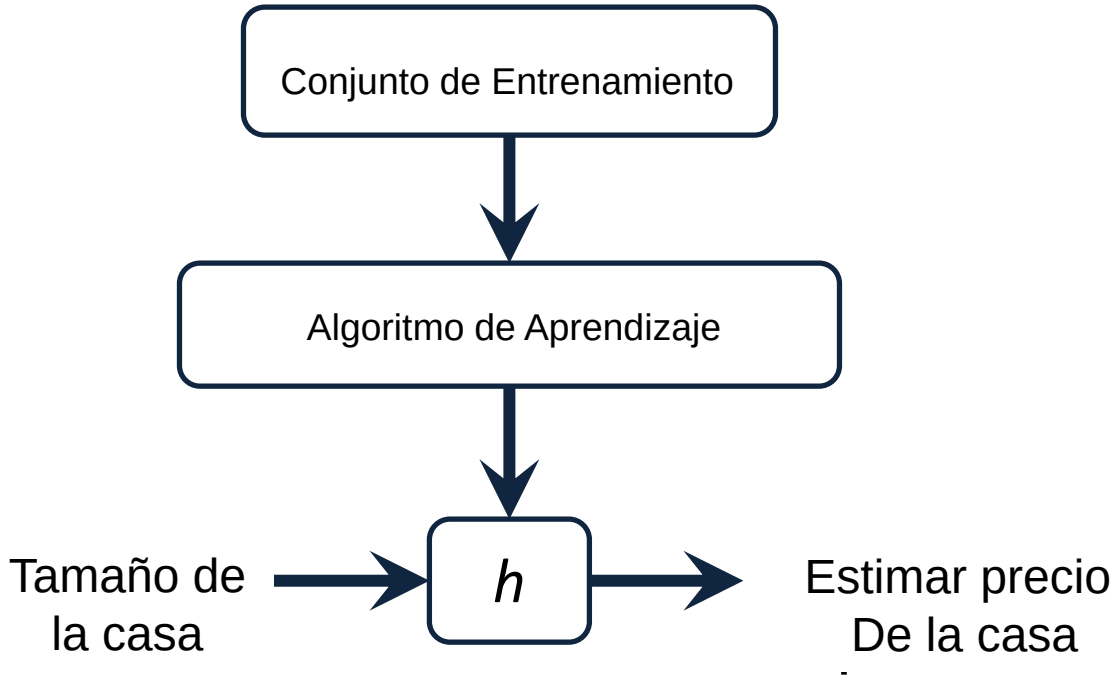
Inducción del conocimiento

Algunas Técnicas de Aprendizaje Automático:

- Árboles de decisión,
- Reglas de asociación,
- Redes Neuronales Artificiales,
- Tablas de decisión
- Algoritmos Evolutivos
- Y muchos más (algoritmos bio-inspirados, etc.)



Construcción de modelos



ALGORITMOS DE APRENDIZAJE

1. SUPERVISADOS: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. Produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema.

- Clasificación, Predicción

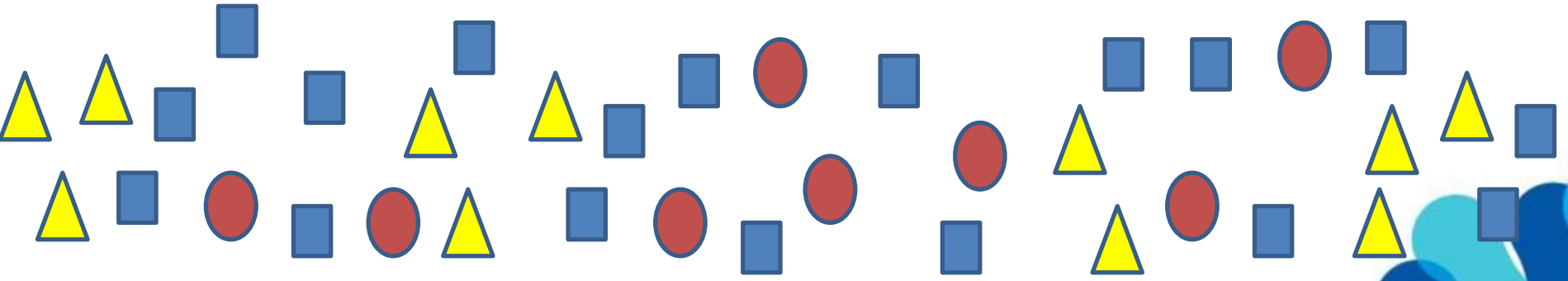
2. NO SUPERVISADOS: descubren patrones y tendencias en los datos, sin tener ningún tipo de conocimiento previo acerca de cuales son los patrones y categorías buscadas.

- Clustering, Análisis de enlace, Análisis de frecuencia

3. OTROS: Aprendizaje semisupervisado, Aprendizaje por refuerzo, Transducción, Aprendizaje multi-tarea, etc.

Aprendizaje supervisado

El proceso de modelado se realiza sobre un conjunto de ejemplos formado por **entradas al sistema** y **la respuesta que debería dar** para cada entrada.



Aprendizaje no supervisado

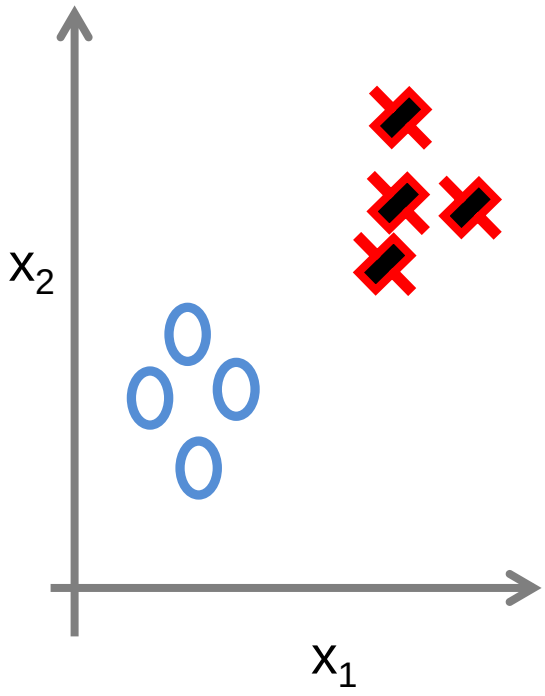
Todo el proceso de modelado se lleva a cabo sobre un **conjunto de** ejemplos **formado tan sólo por entradas** al sistema.

No se tiene información sobre las categorías de esos ejemplos.

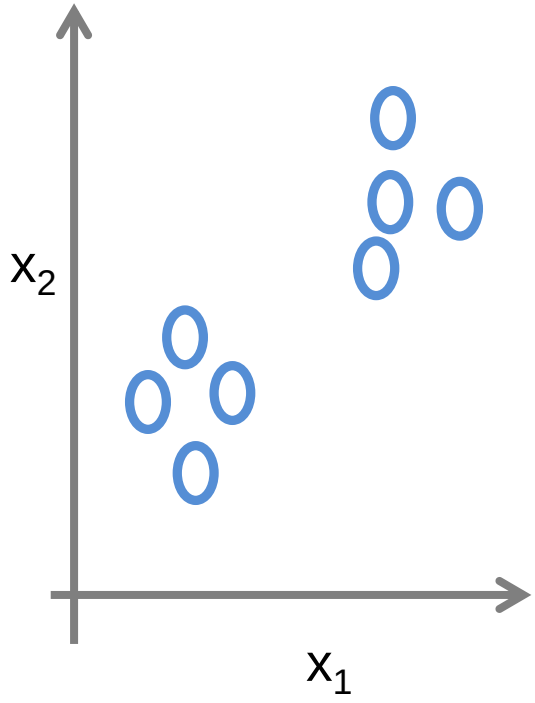
Por lo tanto, en este caso, el sistema tiene que ser capaz de **reconocer patrones** para poder etiquetar las nuevas entradas.

Aprendizaje no supervisado

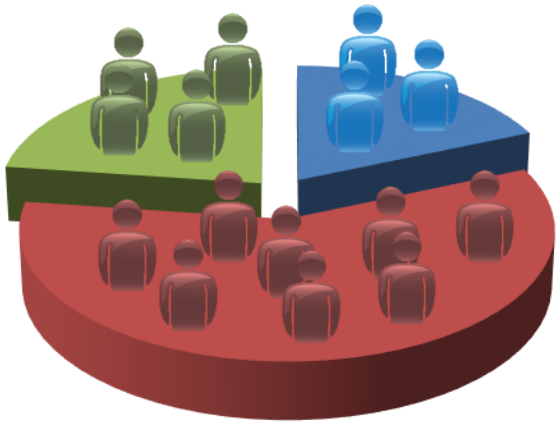
Aprendizaje supervisado



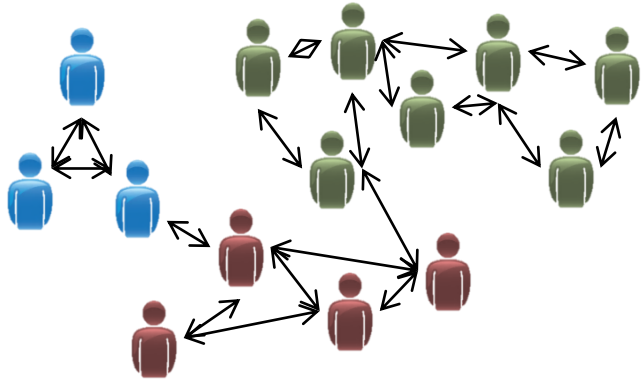
Aprendizaje no supervisado



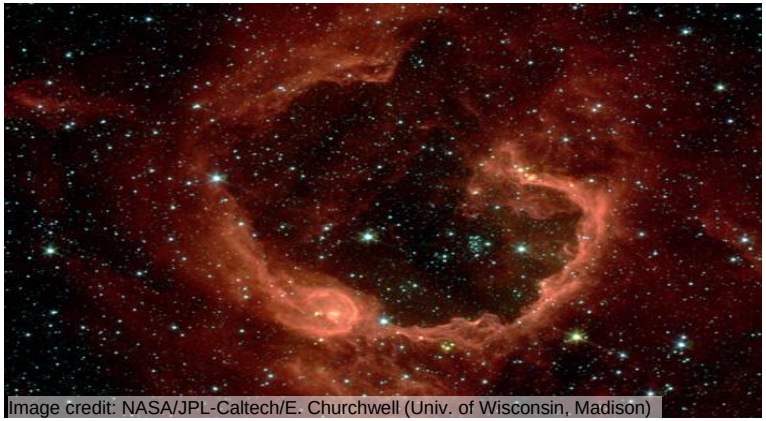
Aprendizaje no supervisado



Segmentar mercado



Análisis de Redes Sociales

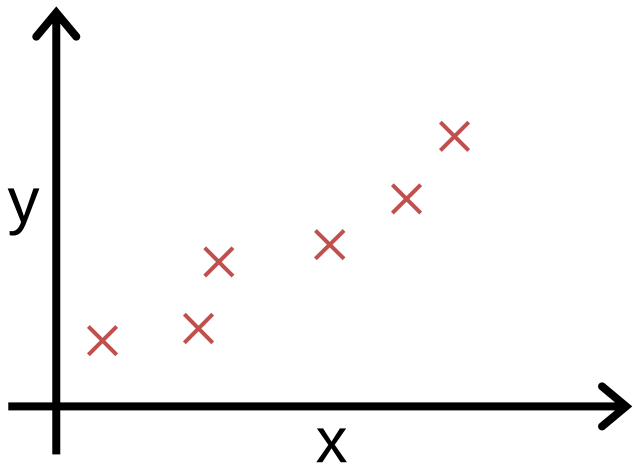


Análisis datos Astronómicos

La Hipótesis

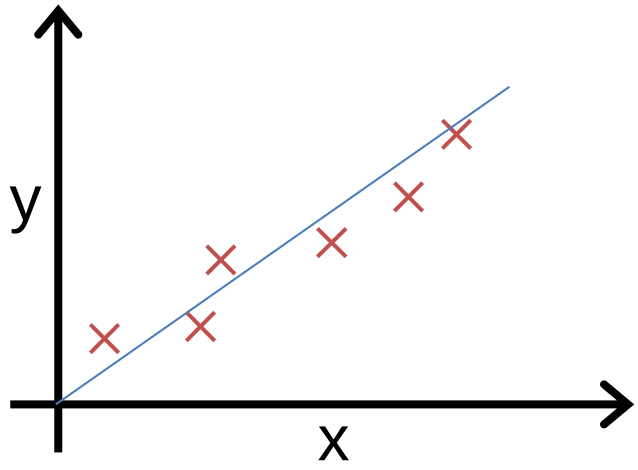
$$h_\theta(x) = \theta_0 + \theta_1 x$$

La Hipótesis



$$h_\theta(x) = \theta_0 + \theta_1 x$$

La Hipótesis

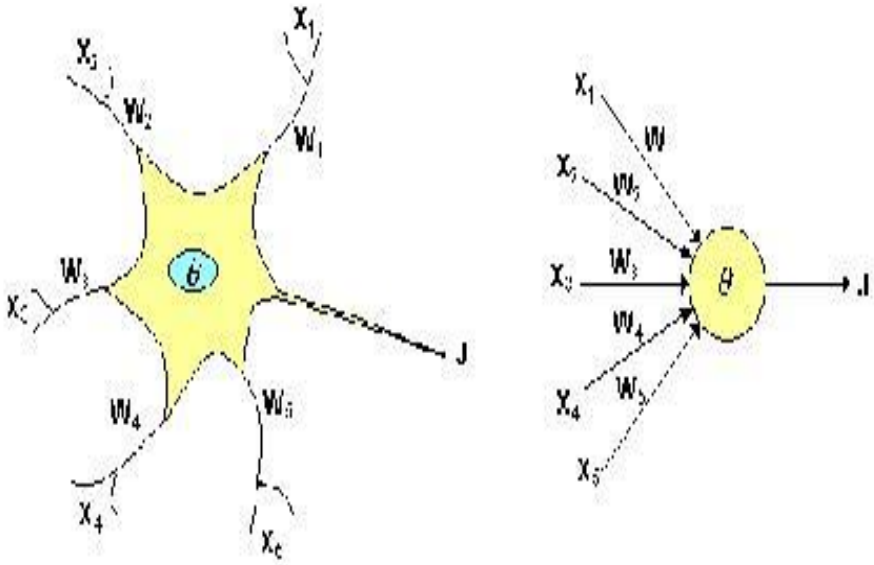


$$h_\theta(x) = \theta_0 + \theta_1 x$$

Idea: Escoger θ_0, θ_1 para que $h_\theta(x)$ se acerque a y con el set de entrenamiento (x, y)

Red Neuronal Artificial

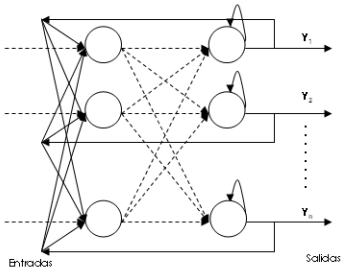
- ✓ Una **nueva forma de computación**, inspirada en el cerebro.
- ✓ Un **modelo matemático** compuesto por un gran **número de elementos de procesamiento,** eventualmente en niveles.
- ✓ Son **redes interconectadas masivamente** en paralelo de elementos simples y organización diversa.

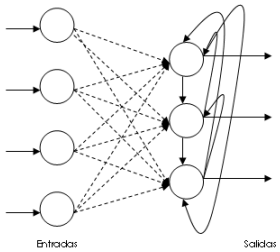


Modelo de una RNA vs. el Modelo Biológico

COMO TRABAJA UNA RED NEURONAL

1. El conjunto de **unidades de procesamiento** (neuronas formales).
2. El **estado interno o de activación** de las neuronas.
3. Las **conexiones entre las neuronas**.
4. Las **conexiones con el ambiente**.
5. **La regla de propagación**
6. **La función de activación**
7. **La función de transición o de salida**
8. La **topología o arquitectura** de la red
9 El **algoritmo de Aprendizaje**





Modelos Neuronales

Realimentados :

feed-propagation

ART,

HOPFIELD

Unidireccionales

PERCEPTRON,

M RN,

BOLTZMAN,

backpropagation

KOHONEN

Híbridos:

RBF (RADIAL BASIC FUNCTION)

Redes basadas en DEEP LEARNING

Redes de Convolución

Extreme Learning

COMPUTACION EVOLUTIVA

- ENFOQUES ALTERNATIVOS DE BUSQUEDA Y APRENDIZAJE BASADOS EN MODELOS COMPUTACIONALES DE PROCESOS EVOLUTIVOS
- *IDEA:* BUSQUEDA ESTOCASTICA EVOLUCIONANDO A UN CONJUNTO DE ESTRUCTURAS Y SELECCIONANDO DE MODO ITERATIVO A LAS MAS APTAS

FINALIDAD: SUPERVIVENCIA DEL MAS APTO

MODO: ADAPTACION AL ENTORNO

- EMULACION DE PROCESOS EVOLUTIVOS:
  - POBLACION DE POSIBLES SOLUCIONES => **INDIVIDUOS**
  - PROCESO DE SELECCIÓN => **APTITUD DE LOS INDIVIDUOS**
  - PROCESO DE TRANSFORMACION => **GENERACION DE NUEVOS INDIVIDUOS**

COMPUTACION EVOLUTIVA

- OBJETIVOS CONFLICTIVOS QUE SE SIGUEN:
  - EXPLOTAR LAS MEJORES SOLUCIONES
  - EXPLORAR EL ESPACIO DE BUSQUEDA

- PARADIGMAS:
  - ALGORITMOS GENETICOS (HOLLAND)
  - PROGRAMAS EVOLUTIVOS (MICHALEWICZ)
  - PROGRAMACION GENETICA (KOZA)
  - ESTRATEGIAS EVOLUTIVAS (RECHENBERG SCHWEFEL)
  - PROGRAMACION EVOLUTIVA (FOGEL)

COMPUTACION EVOLUTIVA

MACROALGORITMO:

1.- POBLACION INICIAL

2.- EVALUACION DE LOS INDIVIDUOS

3.- REPRODUCCION INICIAL

4.- REEMPLAZO

5.- CONDICION DE FINALIZACION O REGRESA A PASO 2

Lógica Difusa

El adjetivo “difuso” se debe a que en esta lógica, los valores de verdad son no-deterministas y tienen, por lo general, una connotación de incertidumbre.

La lógica difusa tiene como base los denominados conjuntos difusos y posee un sistema de inferencia basado en reglas de producción de la forma “***SI antecedente ENTONCES consecuente***”, donde los valores lingüísticos del antecedente y el consecuente están definidos por conjuntos difusos.

LA BORROSIDAD

Por borrosidad entendemos el hecho de que una proposición pueda ser parcialmente verdadera y parcialmente falsa de forma simultánea.

- Replanteamiento radical de **conceptos clásicos de verdad y falsedad**, por e**l concepto de vaguedad o borrosidad**. La verdad y/o falsedad no son más que casos extremos.

- Una persona no será simplemente alta o baja, sino que **participará de ambas características parcialmente**, mientras que en la zona intermedia de ambas alturas existirá una **gradualidad p**or la que va dejando de ser alta.

- El concepto de borrosidad está enraizado en la mayor parte de nuestros **modos de pensar y hablar.**

Lógica Difusa

LÓGICA MULTIVALUADA EN LA TEORÍA DE CONJUNTOS

⇒ **CONJUNTOS DIFUSOS**

⇒ ETIQUETAS LINGUISTICAS

⇒GRADOS DE PERTENENCIA

⇒ **CLASES CON LÍMITES MAL DEFINIDOS**

⇒ **GRADUALIDAD EN LOS CAMBIOS DE ESTADOS**

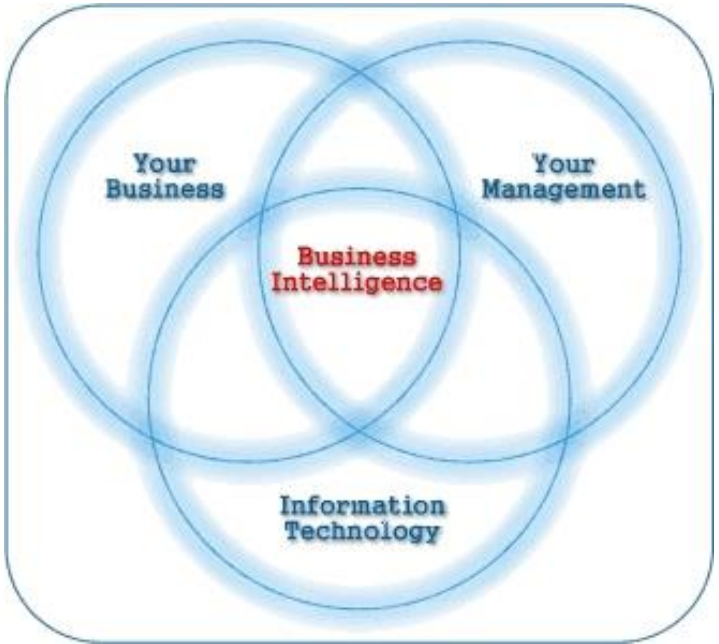
Algunos Conceptos Vecinos: Inteligencia de Negocios, Minería (de datos, semántica, de texto, de Grafos), BigData.

¿Qué es Inteligencia de Negocios?

*Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el anáisis de datos existentes en una organización.

*Abarca la comprensión del funcionamiento actual de la empresa, y la anticipación de acontecimientos futuros, con el objetivo de ofrecer conocimientos para respaldar las decisiones empresariales.

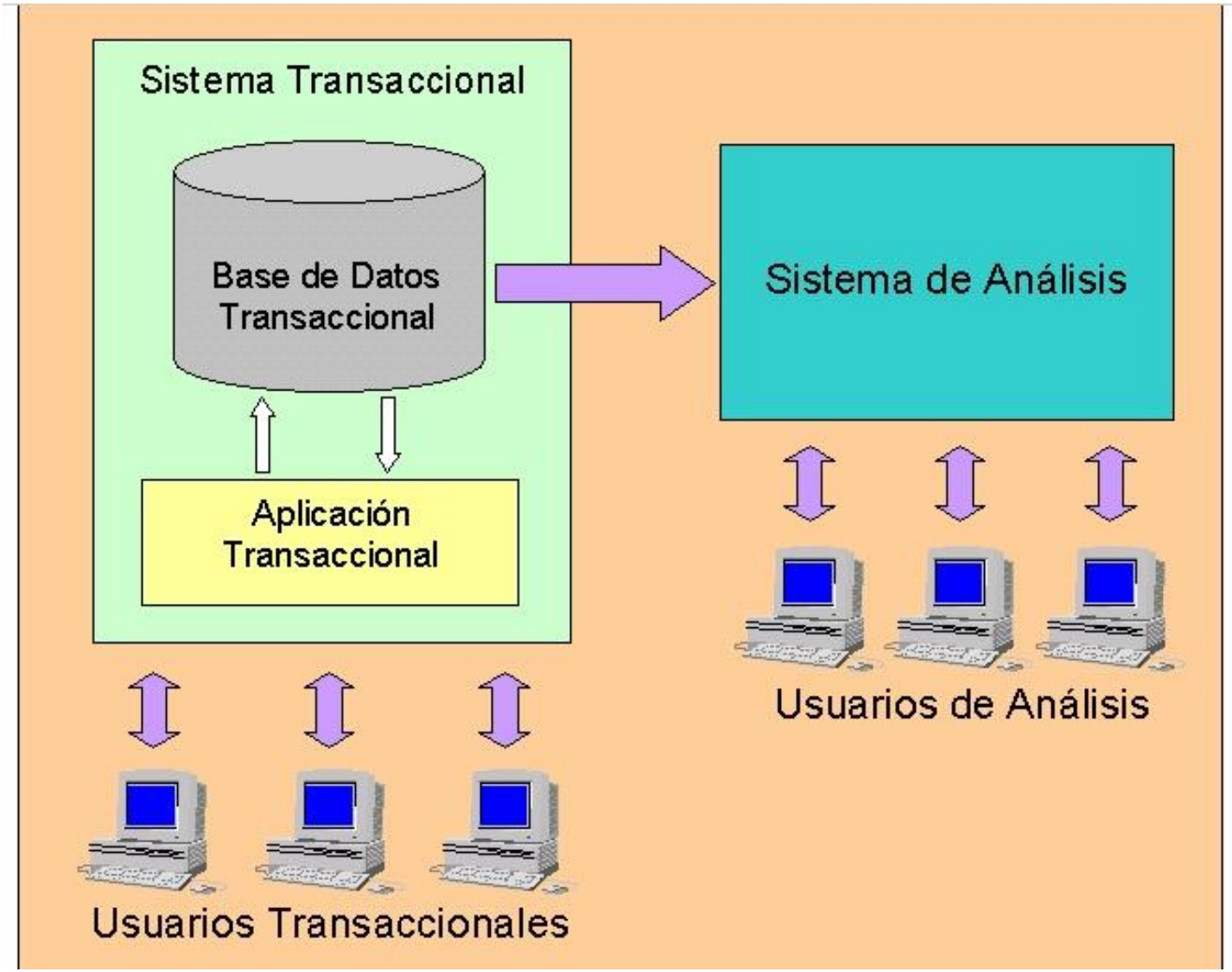
Análitica de Datos



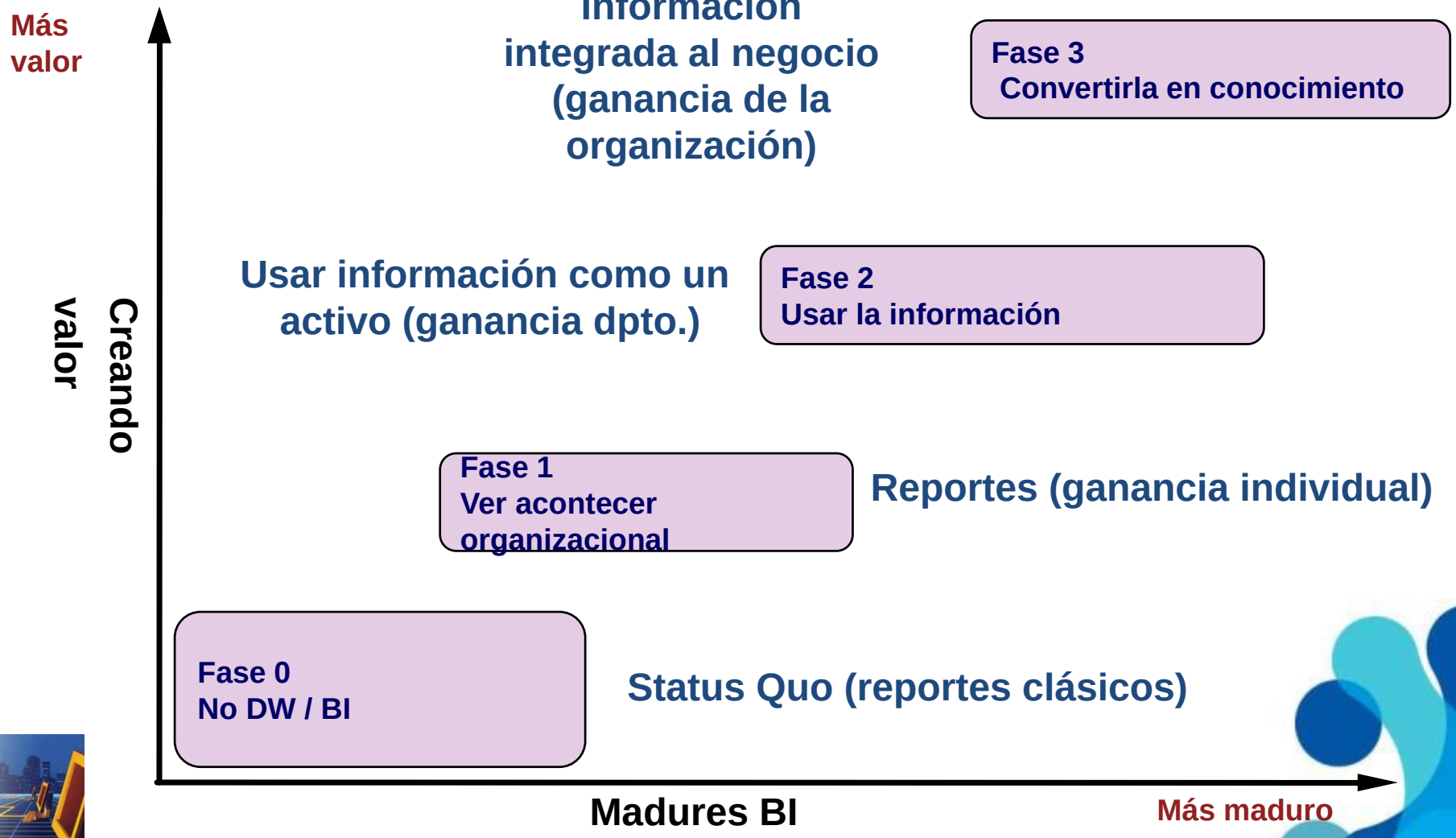
Gestión del conocimiento

Bodega de Datos

Sistemas de Análisis de Información



Cambio en el uso de la Información



LA NECESIDAD DE BI…

Quienes necesitan un ambiente de Inteligencia de Negocios, poseen las siguientes características:

- **Los reportes provenientes de varios sistemas transaccionales, no concuerdan**
  - Los resultados financieros no concuerdan.
  - Las cantidades de inventario tampoco concuerdan.
  - Los reportes detallados no concuerdan con los reportes consolidados.

- **La gerencia no tiene acceso a una "imagen global corporativa" de su situación actual:**
  - ¿Cómo están nuestras finanzas?
  - ¿Quiénes son nuestros clientes?
  - ¿Qué nos han comprado?
  - ¿Cuánto inventario tenemos disponible?

Características de un Proyecto de IN

	Operacional	Táctica	Estratégica
Focalizada	Operaciones diarias	Análisis de corto termino	Objetivos globales de la organización
Usuarios	Analistas, Operadores	Gerentes	Gerentes
Ventana de Tiempo	Diarios	Semanales a Meses	Mese a año
Datos	Métricas Tiempo real	Métricas históricas	Métricas históricas

Tomado de Heinz, 2014

Características de un Proyecto de IN

Características de un Proyecto de IN

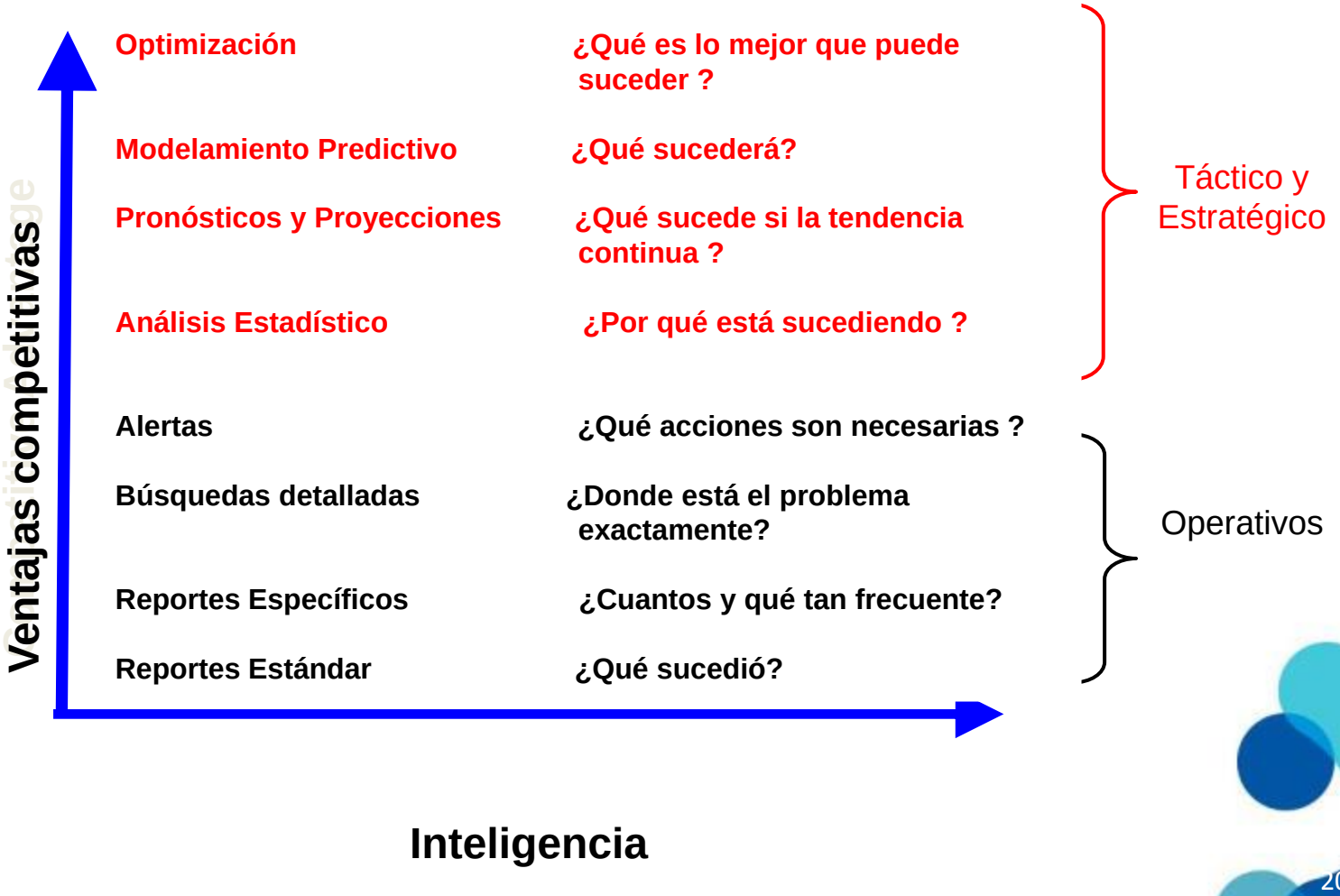
Dominios del Análisis Táctico y Estratégico

Datos Históricos (Data Warehouse)

Sofisticación del Análisis y Herramientas

- Decisiones Tácticas, siguiente semana o mes
- Decisiones Estratégicas, siguiente semestre o año

Tipos de análisis de datos que se pueden realizar



Analítica Social de los Datos

El análisis sociales de datos es un estilo de análisis en el que es considerado las personas trabajan en un contexto social, de colaboración, para darle sentido a los datos.

El análisis de datos sociales se compone de dos partes:

- **Captación de los datos** generados a partir de los **sitios externos**, como redes sociales (o a través de aplicaciones sociales), y
- **Análisis de los datos, en tiempo real** (o casi en tiempo real), en los cuales se **incluyen medidas para entender**, y apropiadamente pesar, factores como la **influencia, alcance y relevancia del contexto de los datos**, y se incluye el horizonte de tiempo.

Analítica Social de los Datos

En un sistema de análisis social de datos:

- Queremos **averiguar relaciones** entre los datos sociales y otro evento, o **predecir** algunos eventos, entre otras cosas
- los usuarios **almacenan conjuntos de datos** y crean **representaciones visuale**s.
- Los conjuntos de datos y representaciones **son accesibles** para otros usuarios de la red o sitio web.
- Los usuarios pueden **crear nuevas representaciones**, así como comentar a las existentes.
- Se pueden armar blogs y wikis para generar procesos de **inteligencia social**.

Métodos de AdDS

- Estadísticos,
- Aprendizaje de máquinas
- Minería de Datos.
- **Minería Semántica**
- **Minería de Grafos**

Analítica Social de los Datos

Cuando se habla de análisis de datos sociales, hay una serie de factores que es importante tener en cuenta:

- **Análisis de datos sofisticados**: El análisis de datos sociales debe tomar en consideración una serie de factores (**contexto, contenido, sentimiento**) para proporcionar información adicional.

- **La consideración del tiempo**: Lo más relevante de un día (o incluso una hora) puede no ser en la siguiente. Ser capaz de ejecutar **con rapidez (tiempo real)** el análisis es imperativo.

- **Análisis de la influencia:** la comprensión del **impacto potencial de individuos/eventos específicos** puede ser clave en la comprensión de cómo los mensajes podrían estar **resonando**. No se trata sólo de la cantidad, también tiene mucho que ver con la calidad.

- **Análisis de las Redes**: los datos sociales migran, crecen (o mueren) en base a cómo **los datos se propagan a través de la red**. Es como una actividad viral, que se inicia y se propaga.

Analítica Social de los Datos

- ***Análisis del contexto colectivo:***
  - **Analítica del contenido** -es una de las características definitorias de la Web Semántica, grafo de conocimiento
  - **Analítica del Contexto-** topologías, estructura y enlaces en la Web.
  - **Analítica de Redes Sociales-** grafos de interacción.

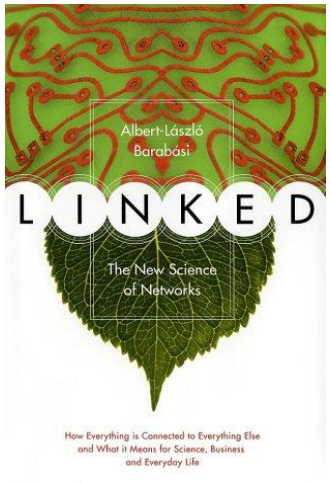
- ***Análisis personales desde un contexto colectivo:***
  - **Analítica de la Disposición-** formas de interacción (centro de la innovación)
  - **Analítica del Discurso-** el lenguaje es una herramienta fundamental para la construcción del conocimiento.
  - **Analítica de Redes Sociales-** relaciones interpersonales en plataformas sociales.

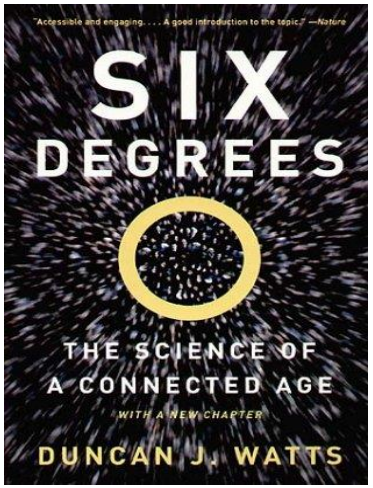
Red social

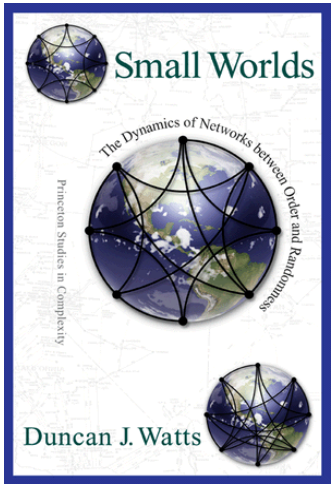
- **Una red social** es una estructura social de personas, relacionadas (directa o indirectamente) entre sí a través de una relación o interés común

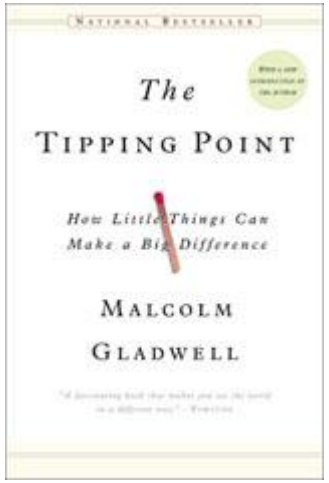
- El **análisis de redes sociales (SNA)** es el estudio de redes sociales para entender su estructura y comportamiento

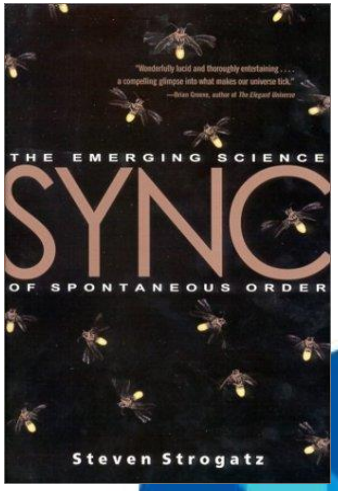
Es el proceso de investigación de estructuras sociales a través del uso de teorías de redes y gráficas











Las redes sociales han capturado el interés del público en los últimos años, como es evidente en el número de tratamiento de la ciencia popular de la materia

Analítica Social del aprendizaje

Análisis de redes sociales

El aprendizaje en red implica el uso de las TIC para promover conexiones entre un alumno y otros alumnos, entre alumnos y tutores, y entre las comunidades de aprendizaje y recursos de aprendizaje.

- Estas redes se componen de actores (tanto de personas como de recursos) y las relaciones entre ellos.
- El análisis de redes sociales investiga los procesos en la red, las propiedades de las relaciones, los roles y la formación de la red, para entender cómo la gente desarrolla y mantiene estas relaciones para apoyar el aprendizaje

Ejemplo implementación de análisis de redes sociales de aprendizaje es SNAPP (Social Networks Adapting Pedagogical Practice), una herramienta de visualización de la red que incluye:

- Identificar a los estudiantes desconectados
- Identificar a los agentes de información claves dentro de una clase
- Indicar el grado en que una comunidad de aprendizaje se está desarrollando

Aprendizaje Social

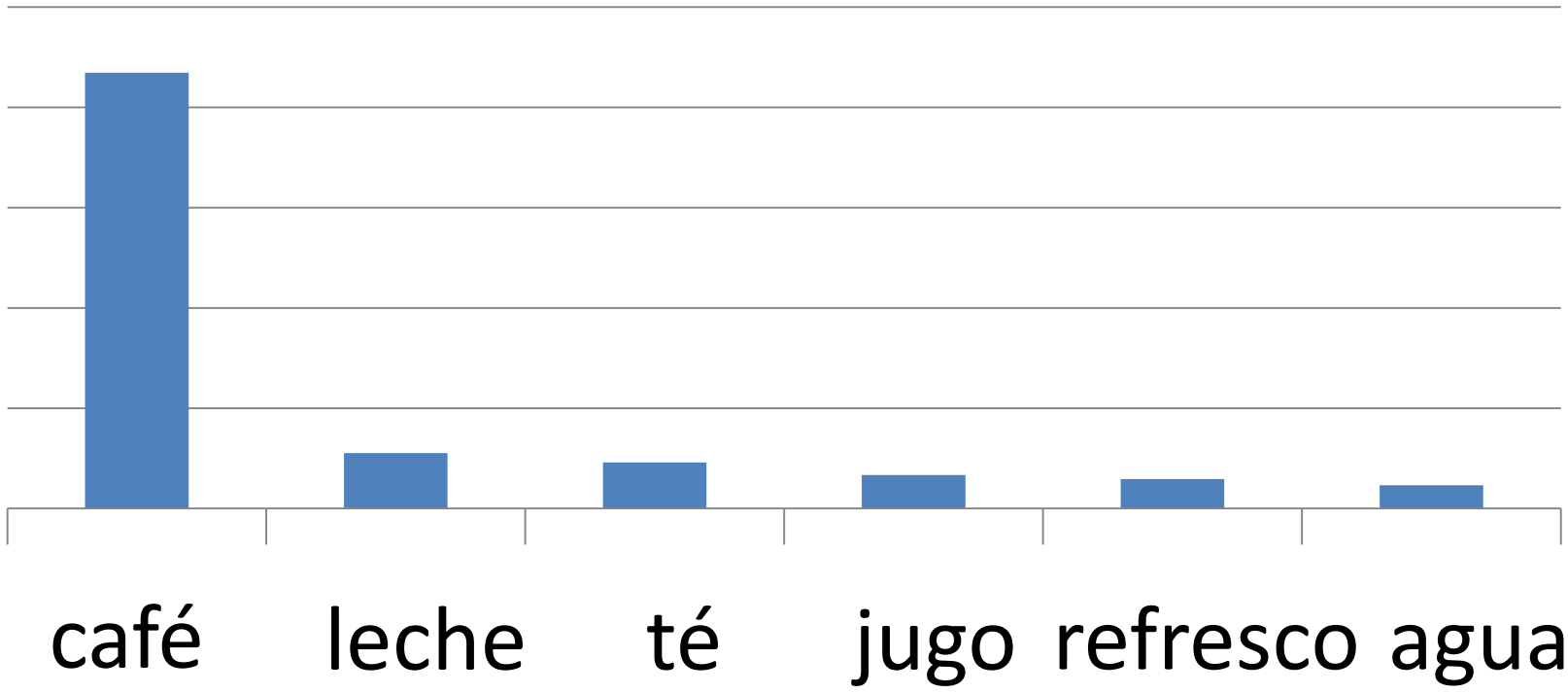
Aprendizaje social es una actividad social que se lleva a cabo cada vez más a distancia y en formas mediadas

El aprendizaje social **añade dimensiones** importante al **aprendizaje colaborativo** asistido por computadora,

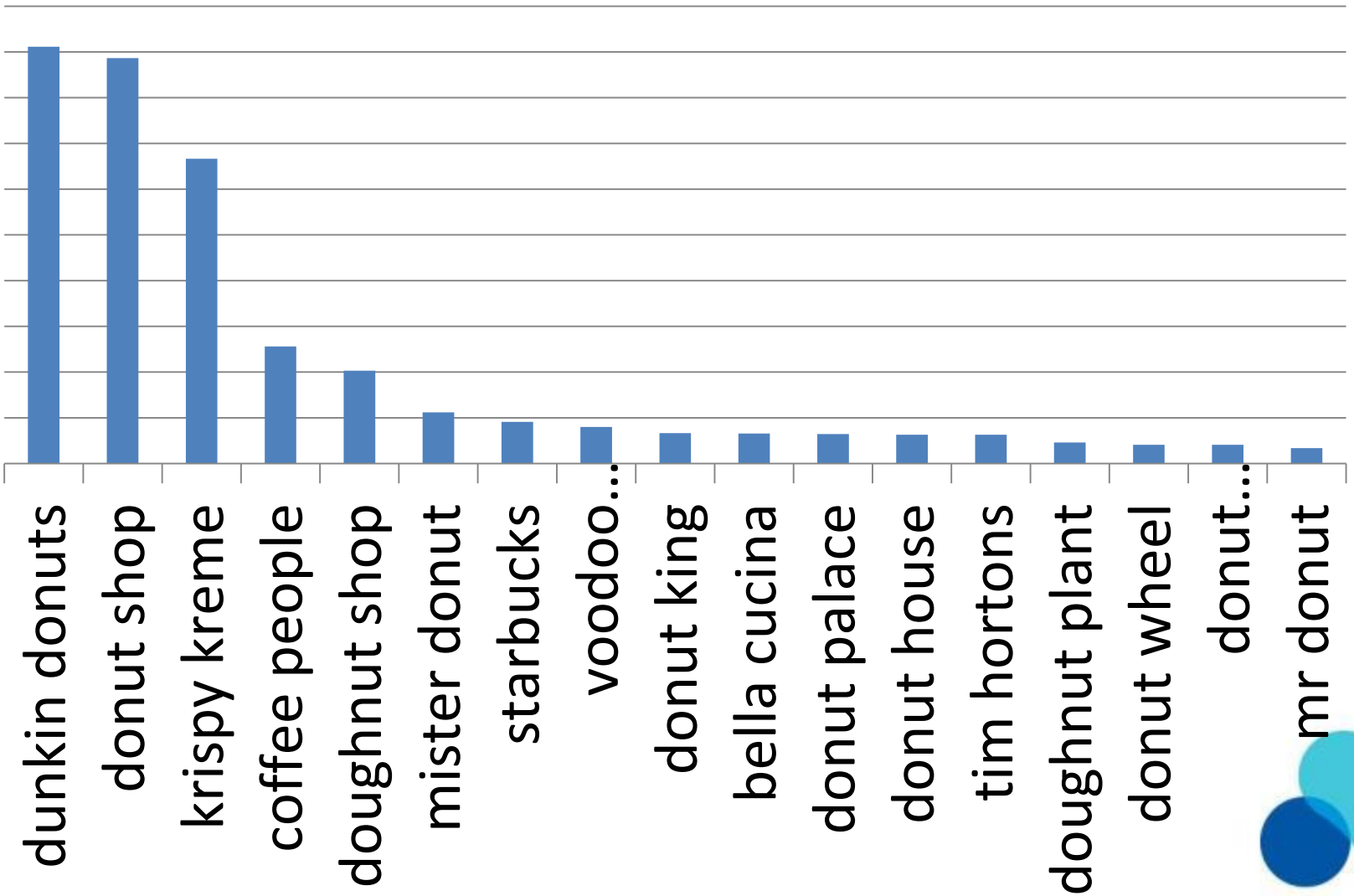
- El interés particular en los **contextos no académicos** en los que puede tener lugar (el hogar, la red social, el lugar de trabajo) y e
- El **uso de herramientas**,
- **Sin un plan de estudios**, o actividades prescritas o pre-programadas


SOCIAL MEDIA EXPLAINED
TWITTER
I'M EATING A #DONUT
FACEBOOK
I LIKE DONUTS
FOURSQURE
THIS IS WHERE I EAT DONUTS
INSTAGRAM
HERE'S A VINTAGE PHOTO OF MY DONUT
YOUTUBE
HERE I AM EATING A DONUT
LINKEDIN
MY SKILLS INCLUDE DONUT EATING
PINTEREST
HERE'S A DONUT RECIPE
LAST FM
NOW LISTENING TO "DONUTS"
G+
I'M A GOOGLE EMPLOYEE WHO EATS DONUTS.

¿Qué beben las personas con donas?

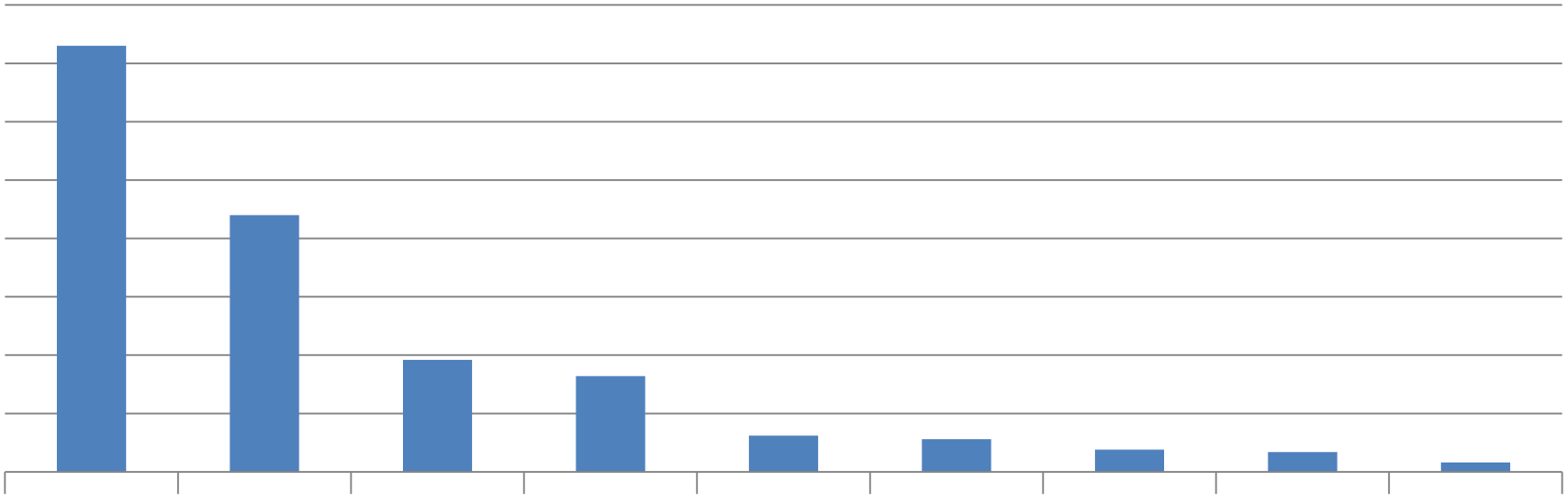


¿Dónde la gente consigue donas?



187/5000~ 180k tweets que contienen "donas" de 7 días ene twitter

¿Qué tipo de donas comen la gente?



Con jalea Satinado Puro Con Sirope Con Jamon …

Aprendiendo desde las experiencias del mundo

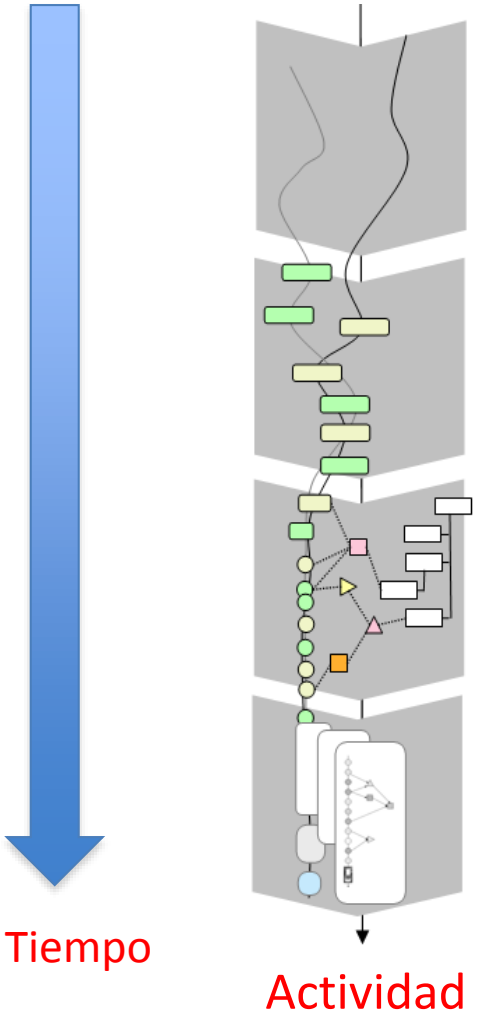
Utilizar las redes sociales como un **registro fresco** y en gran escala de las **acciones, motivaciones y emociones de las personas**

El objetivo es ayudar a las personas con sus tareas y decisiones, mostrándoles:

- **Lo que otros han hecho en situaciones similares,**
- **por qué lo hicieron y**
- **cómo se sintieron después.**

¿Dónde ir a catar vino?
¿Dónde comen las personas sanas?
Encontrar un café para estudiar
¿Qué es gracioso ahora?

Proceso de modelado



Minería Semántica

Determinar relaciones semánticas

- Minería de Datos Semánticos (*Semantic Data Mining)*
- Minería de la Web Semántica (*Semantic Web Mining)*
- Minería Ontológica (*Ontology Mining*).
- Minería de Texto

Análisis de las redes de aprendizaje
Análisis del contenido de aprendizaje
Análisis del contexto de aprendizaje

Minería de Texto

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

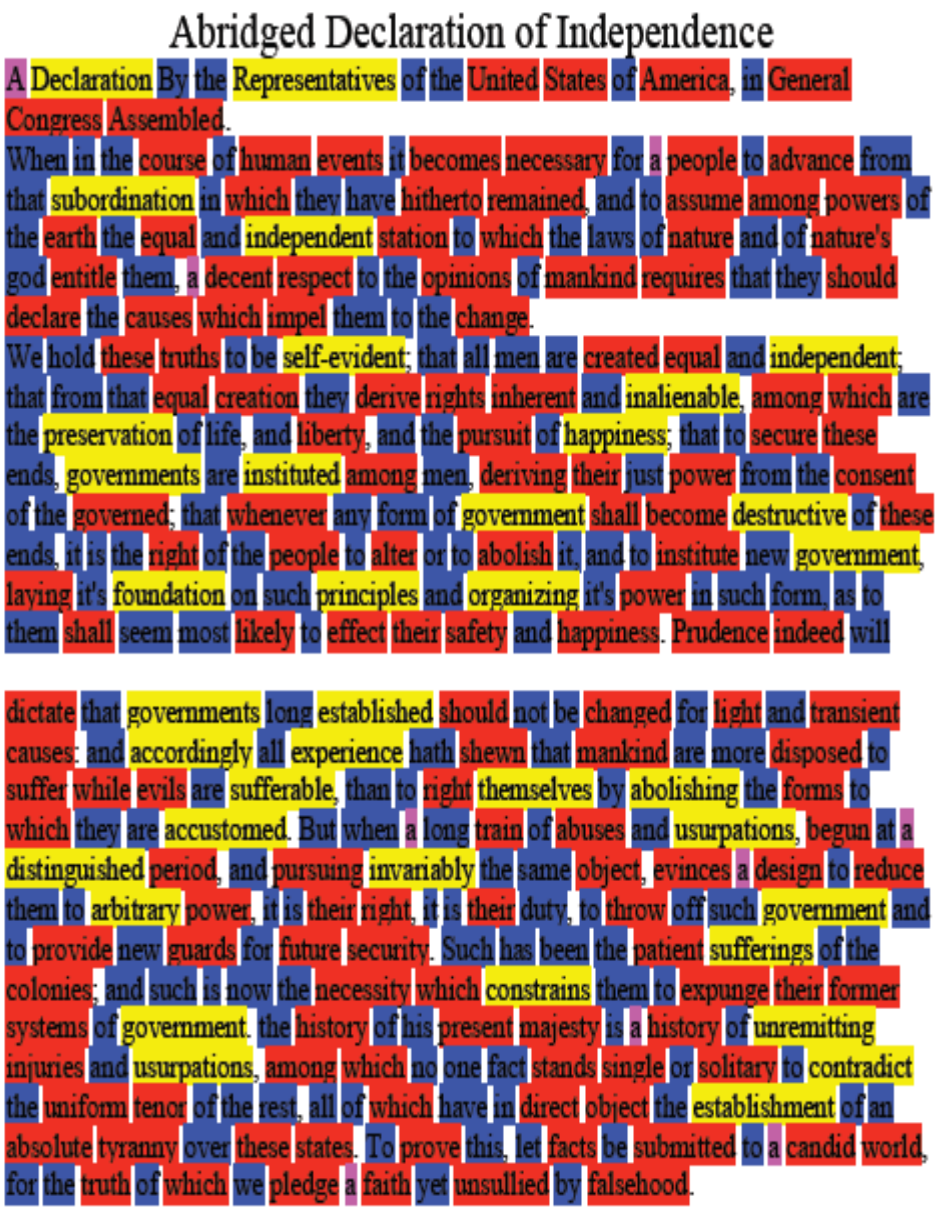
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

¿Cuantas palabras pequeñas, medias y largas son usadas?

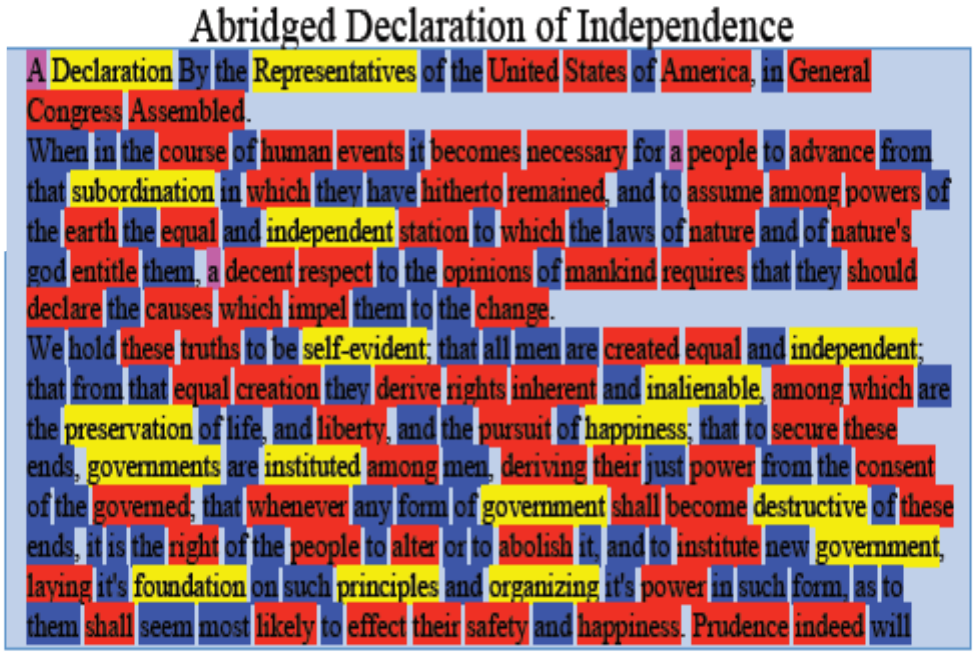
Histográma de longitud de las palabras

- Mucho (amarillo)= 10 + letras
- Medio (rojo)= 5 a 9 letras
- Poco (azul)= 2 a 4 letras
- Morado= 1 letra



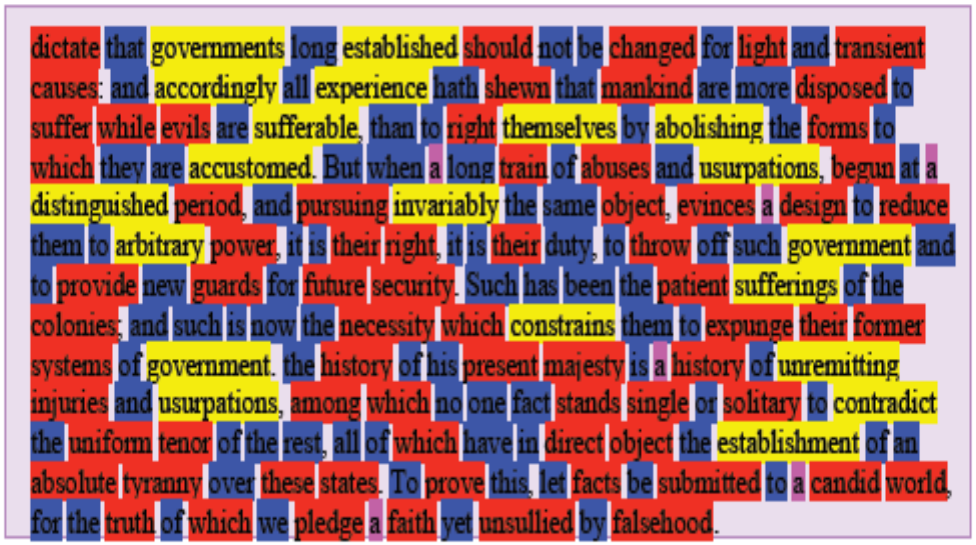
Histográma de longitud de las palabras

Mapa 1
204 palabras



- Amarillo, 17
- Rojo, 17
- Azul, 107
- Morado, 3

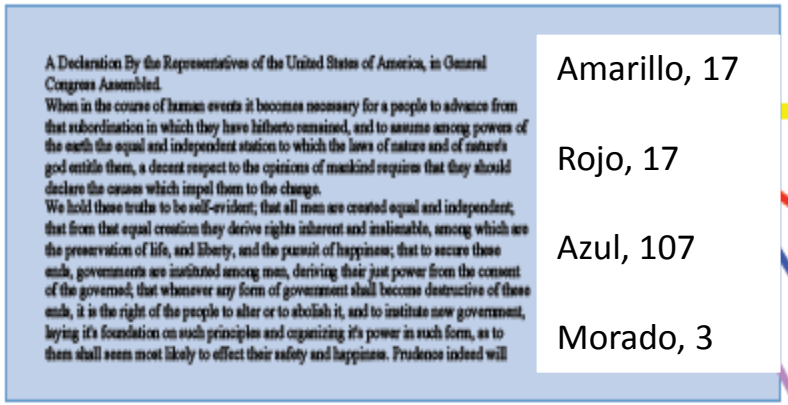
Mapa 2
190 palabras



- Amarillo, 20
- Rojo, 71
- Azul, 93
- Morado, 6

Histográma de longitud de palabras

Mapa 1



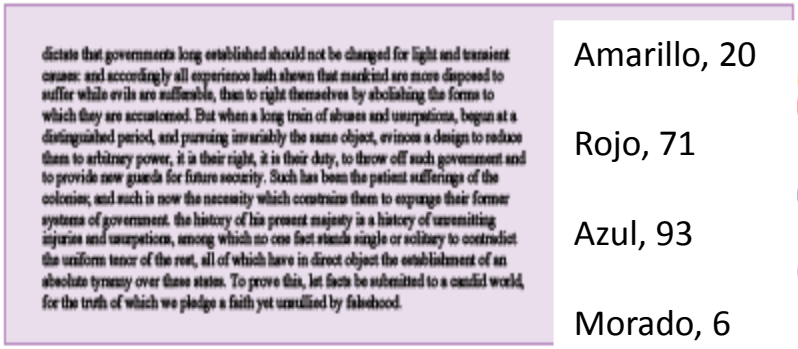
Amarillo, 17

Rojo, 17

Azul, 107

Morado, 3

Mapa 2



Amarillo, 20

Rojo, 71

Azul, 93

Morado, 6

"Shuffle step"

Reducción

- Amarillo, 37
- Rojo, 148
- Azul, 200
- Morado, 9

Minería de Texto

DM

Dato estructurado

HomeLoan (
Loanee: *Frank Rizzo*
Lender: *MWF*
Agency: *Lake View*
Amount: *$200,000*
Term: *15 years*
)

Multimedia

Loans($200K,[map],...)

texto libre

Frank Rizzo bought his home from Lake View Real Estate in 1992.
He paid $200,000 under a15-year loan from MW Financial.

Hypertexto

<a href>Frank Rizzo Bought <a hef>this home from <a href>Lake View Real Estate In 1992. <p>...

Minería de Texto

Dato estructurado

CompraCasa (
Comprador: *Jose*
Prestamista: *MWF*
Sitio: *Loja*
Cantidad: *$200,000*
Plazo: *15 años*
)

Multimedia

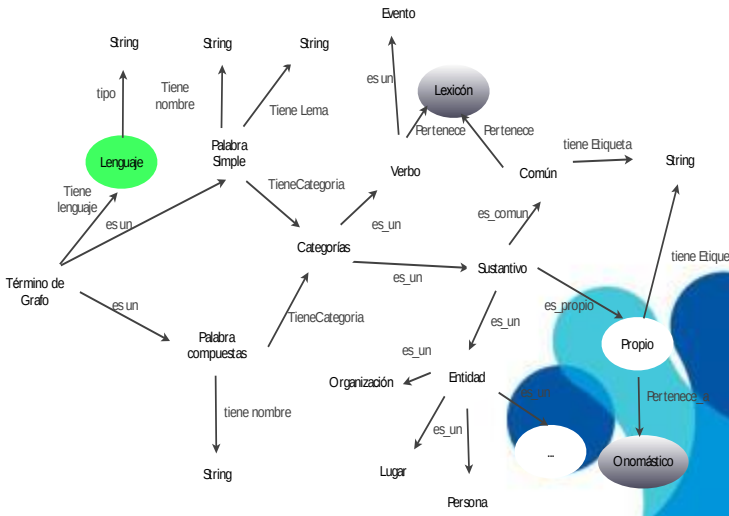
texto libre

Jose compro su Casa en MWF en Loja.
Pagará $200,000 en15-años a la Entidad financierra MW F.

Hypertexto

<a href> Jose Compro <a hef>su casa en <a href>Loja en 2002. <p>...

Construir ontologías



Por ejemplo, como resúmenes desde un grupo de documentos

Etapas de la minería de texto

1. Selección de documentos implica la identificación y recuperación de los documentos potencialmente relevantes de un conjunto grande (por ejemplo, Internet).

2. Pre-tratamiento documento incluya la limpieza y la preparación de los documentos, por ejemplo, eliminación de información extraña, corrección de errores, la normalización ortográfica, tokenización, etiquetado, etc.

3. Procesamiento de documentos consiste principalmente en la extracción de información. Para la Web Semántica es extracción de metadatos

La Minería de Datos es un área **bastante madura** en las Ciencias Computacionales, cuyo principal objetivo es la extracción de conocimiento,.

La Minería de Datos ha requerido ser enriquecido estos últimos años, debido a la necesidad de incorporar contenido semántico.

Minería Semántica

Minería Semántica

- Uno de los problemas más importantes y difíciles en la minería de datos es la incorporación del **conocimiento del dominio**

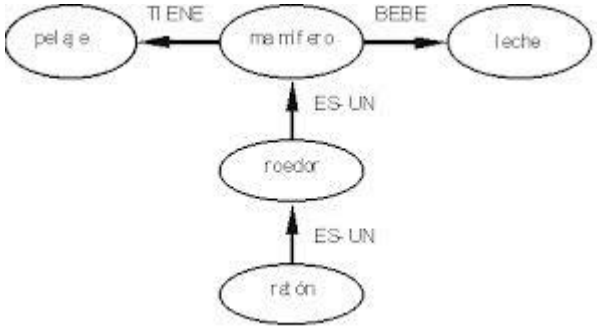
- Cuando los datos y el conocimiento del dominio están disponibles, vale la pena explorar la **relación semántica** entre ellos.

Ese proceso para determinar relaciones semánticas es conocido como Minería Semántica,

Minería Semántica

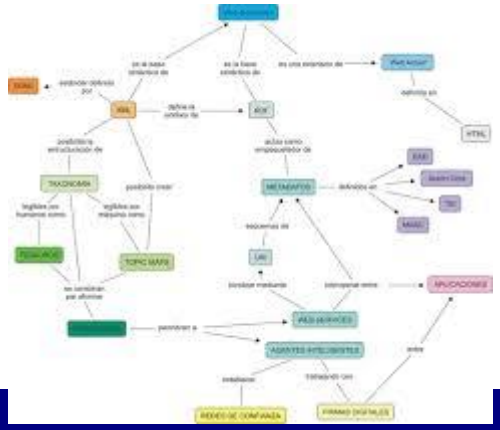
	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	$ 600.000
10	Pamela Carrazco	Mesera	3157485912-2485796	CANDELARIA	$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	$ 600.000
12	Jairo Arevalo	Parqueadero	3002157459-2861459	BOSA	$ 489.500
13				TOTAL	$ 8.199.500

Minería

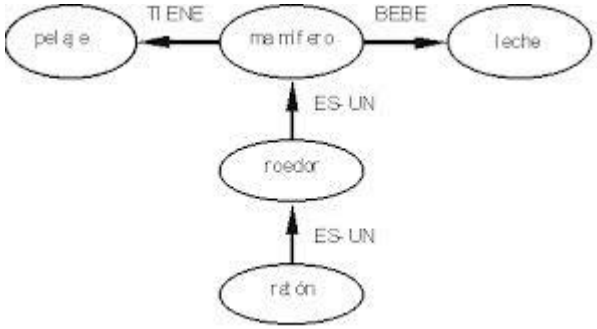


Minería Semántica

	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	$ 600.000
10	Pamela Carrazco	Mesera	3157485912-2485796	CANDELARIA	$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	$ 600.000
12	Jairo Arevalo	Parqueadero	3002157459-2861459	BOSA	$ 489.500
13				TOTAL	$ 8.199.500



Minería



Minería Semántica

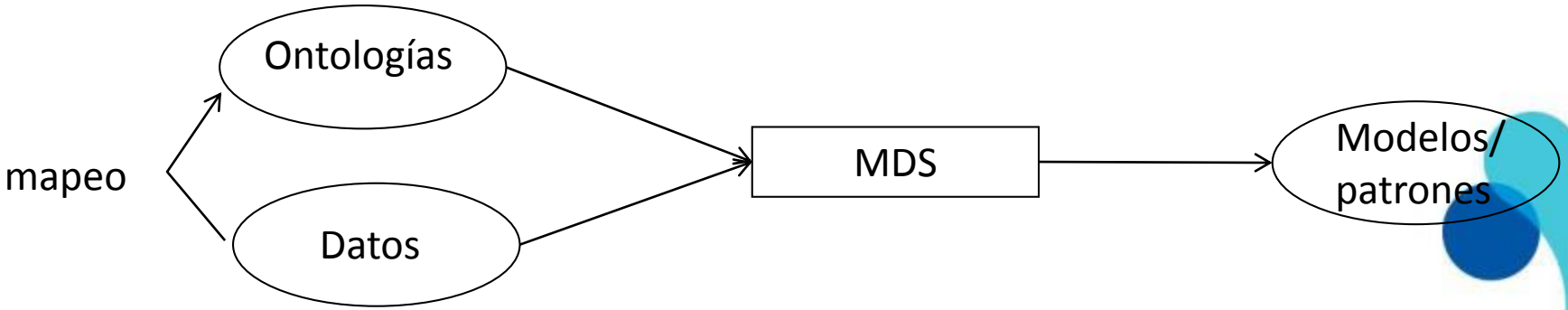
La minería semántica se encarga de extraer conocimiento semántico desde diferentes fuentes semánticas,

- Páginas web,
- Contenido sin estructura en la web,
- Contenido estructurado en la web,
- Grafos anotados,
- Ontologías,
- Tabla de Datos, entre otros

Minería de Datos Semánticos (MDS)

- **Incorporar conocimiento de un dominio** a los datos.
  Minar **recursos anotados semánticamente**, como ontologías para enriquecer semánticamente los datos

- **Añadir contenido semántico a/desde los datos usando técnicas de MD** para la extracción de ese conocimiento (en este caso, la fuente es contenido semántico).



Minería de Datos Semánticos (MDS)

- El proceso de MDS se da en dos pasos,
  1. **Identificación del enriquecimiento semántico,**
  2. **Aplicación de técnicas de MD como tal en él.**

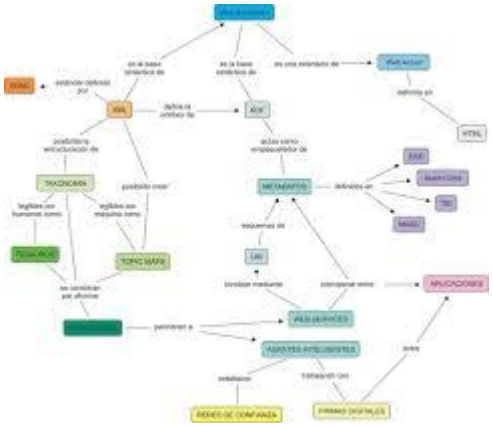
- En el primer paso se usan ontologías, o cualquier contenido semántico, y **se realiza un mapeo** con la data que se va a trabajar, almacenada normalmente en bases de datos.

- En el segundo paso se aplican técnicas de MD para **buscar patrones, relaciones**, y en general, cualquier operación que **explote el enriquecimiento semántico**.

Minería de Datos Semánticos

Dado: tabla de datos de transacciones, bases de datos relacionales, documentos de texto, páginas Web, ... una o más ontologías de dominio

	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	$ 600.000
10	Pamela Carrazco	Mesera	3157485912-2485796	CANDELARIA	$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	$ 600.000
12	Jairo Arevalo	Parqueadero	3002157459-2861459	BOSA	$ 489.500
13				TOTAL	$ 8.199.500



Minería

Encontrar: un modelo de clasificación, un conjunto de patrones

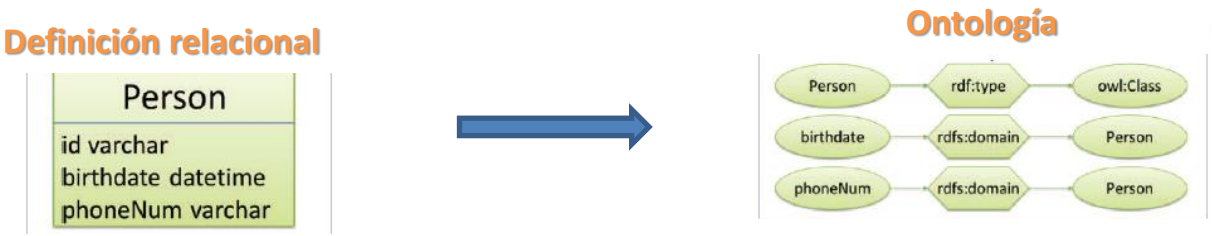
Minería de Datos Semánticos

- **Actual escenario de la MDS:** Minería de datos **empíricos** con ontologías como conocimiento de fondo
  - Abundantes datos empíricos,
  - Escaso conocimiento de fondo

- **Futuro escenario de MDS**:
  - Volumen creciente de ontologías y colecciones de datos semánticamente anotados

    más de 6 billones de tripletas RDF

    más de 200 millones de enlaces



Minería de la Web Semántica (MWS)

- Es la integración de dos áreas de conocimiento,
  - **Web Semántica (Semantic Web)**
  - **Minería en la Web (Web Mining)**

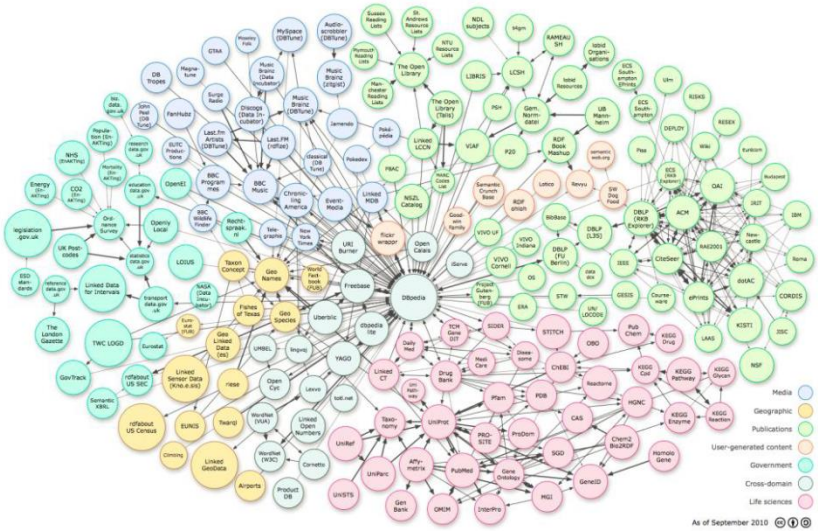
La **Web Semántica** es usada para darle significado a los datos que se encuentran en la Web.

La **Minería en la Web** se usa para extraer patrones de comportamiento en la Web.

Minería de la Web Semántica

Cambio de paradigma de la minería de datos a la minería de conocimiento

- Minería de la Web Semántica: Minería del conocimiento codificado en ontologías de dominio,



Dos tipos de recursos semánticos

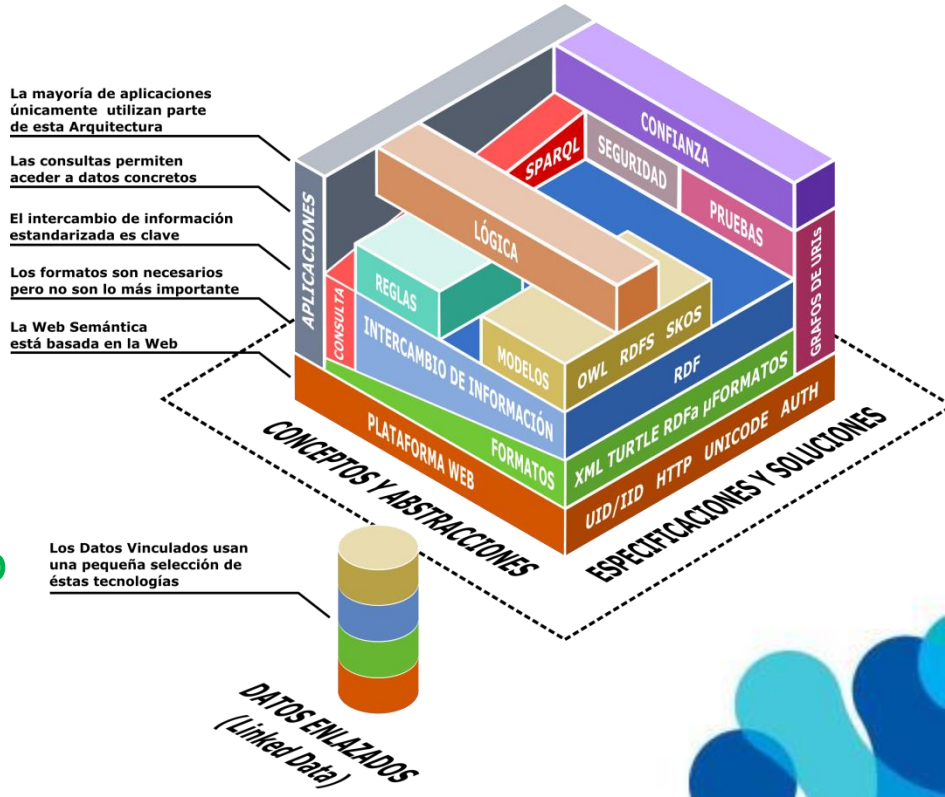
- Ontologías de Dominio
- Ontologías del flujo de trabajo de la minería de datos

Minería de la Web Semántica

La diferencia de MWS con MDS es el propósito y lo que se está minando.

MWS mina datos de la Web, y los resultados son usados en la Web.

- La web semántica es expresada en formatos como OWL, RDF, XML,
- Son los recursos que van a ser minados para extraer conocimiento de la web semántica



Minería Web

Existen varios tipos de **Minería Web** que se pueden aplicar en la MWS:

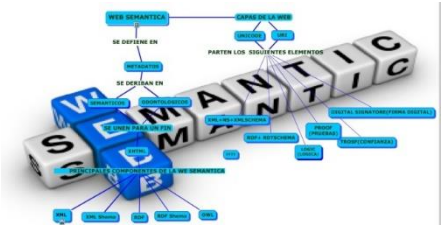
- El contenido de la web,
- La estructura de la web
- El uso que se hace de la web.

Resultados de la Búsqueda
Contenido de la Página Web

Enlaces

Compartir Información
- Microformatos
- RDFa
- FOAF
- SEO Semántico

Patrones generales de uso
Patrones personales de acceso



Minería de Web Semántica

El minado de contenido, es una forma de ***Text Mining***, que se aplica al contenido en la Web.

Por ejemplo, identificar en una página términos similares.

El minado de la estructura estudia el esqueleto que forman los enlaces entre las páginas de la Web, se mina un conjunto de enlaces.

El minado del uso de la web, se enfoca en minar un historial de uso de usuarios

Por ejemplo, consultas que hacen en una página, movimientos que los usuarios hacen entre páginas, etc.

Minería Web Semántica

Microformatos

XFN (**X**HTML **F**riends **N**etwork)

XFN quick reference

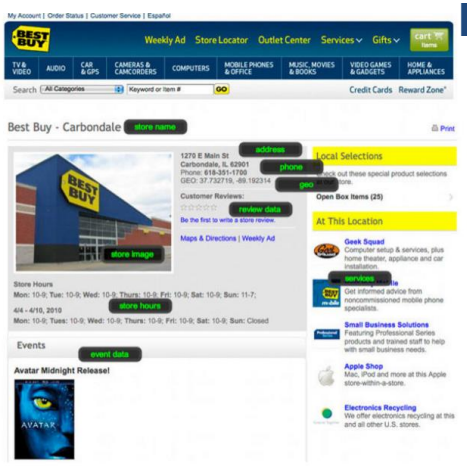
relationship category	XFN values
friendship (at most one):	friend acquaintance contact
physical:	met
professional:	co-worker colleague
geographical (at most one):	co-resident neighbor
family (at most one):	child parent sibling spouse kin
romantic:	muse crush date sweetheart
identity:	me

hCard

hcalendar

Social Data Analytics
Social Network Analytics
Linked Data

RDFa



Best Buy employees entered information into the blogs every day, using online forms that output RDFa. Myers told us that the use of RDFa makes "human input from our store employees more visible on the Web."

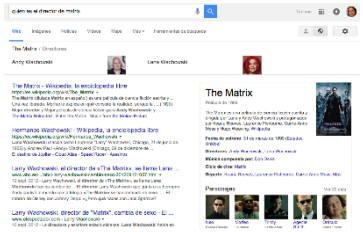
Best Buy is using Good Relations, a Semantic Web vocabulary for e-commerce that describes product, price, and company data.

FOAF

```
<?xml version="1.0" standalone="yes"?>
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/">

    <foaf:Person>
        <foaf:name>
            Taniana Josefina Rodrígiuez de Paredes
        </foaf:name>
        <foaf:mbox rdf:resourse=mailto:taniana@ula.ve/>
        <foaf:knows>
            <foaf:Person>
                <foaf:name> Jose Aguilar </foaf:name>
                <foaf:mbox rdf:resourse=mailto:aguilar@ula.ve/>
            </foaf:Person>
        </foaf:Knows>
    </foaf:Person>
</rdf:RDF>
```

SEO Semántico



Knowledge Graph

Minería Ontológica (*MO*)

Actualmente, con el gran crecimiento en las cantidades de ontologías disponibles sobre un dominio de conocimiento dado, ha llevado a la MO a explorar técnicas que puedan extraer conocimiento adicional de un conjunto de ontologías, para lograr un dominio de conocimiento más amplio.

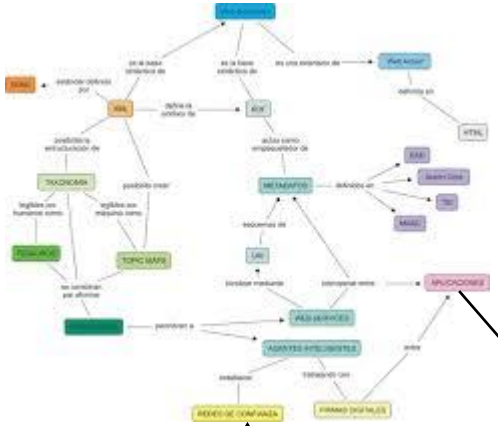
1. **La extracción de: patrones de comportamiento, entre otras características,**
2. **Con la finalidad de construir o enriquecer ontologías.**

Minería Ontológica

- **Extracción de Reglas:** extrae reglas de un conjunto de ontologías.
- **Integración de Ontologías:** busca el vocabulario compartido entre varias ontologías.
- **Enlazado de Ontologías:** encuentra relaciones entre entidades de distintas ontologías.
- **Mezcla de Ontologías:** mezcla la información de varias ontologías con el fin de estandarizar conocimiento.
- **Alineación de Ontologías:** Identifica conceptos semejantes entre ontologías.

Extracción de Reglas

Ontología



Grupo inicial

documentos

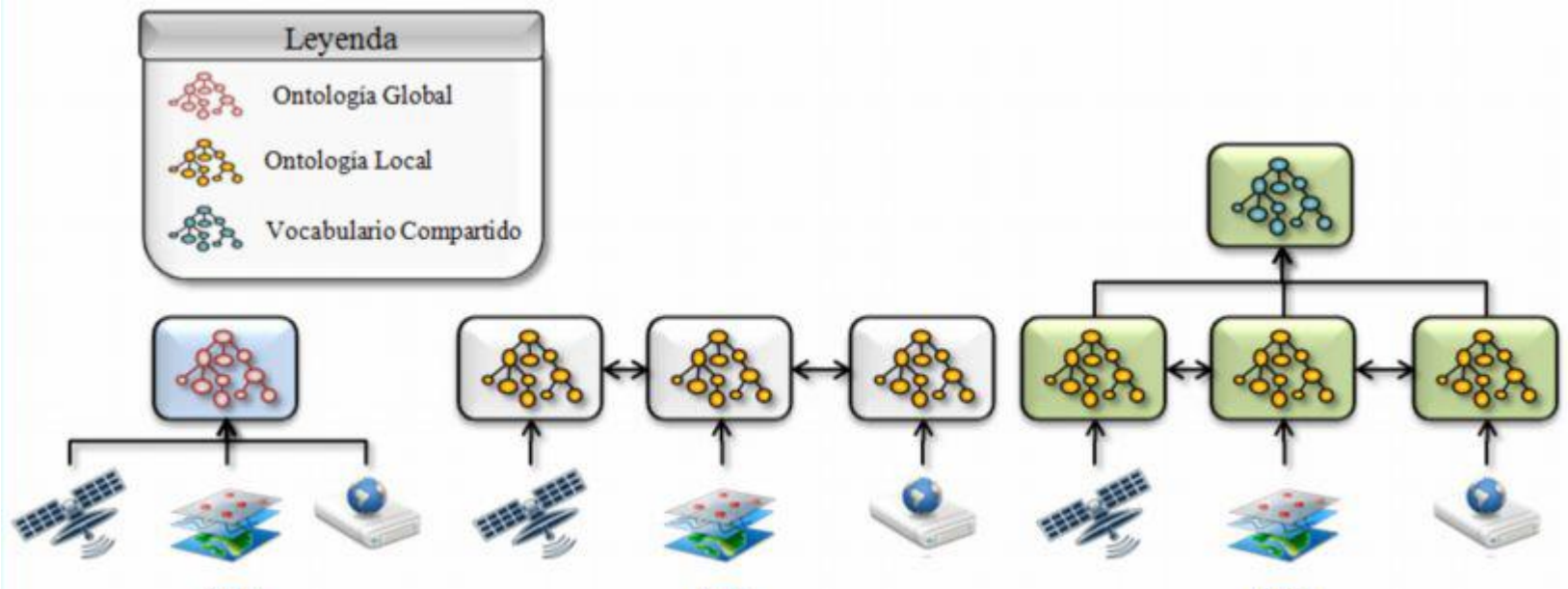
aprendizaje

Extracción de reglas

actualización

Nuevas Ontologías

Integración de ontologías



Alineación de ontologías

identificar conceptos de una ontología que sean semejantes en las otras
ontologías

Distancia semántica entre cada par de conceptos en ontologías distintas

Existen varios métodos y herramientas para realizar la alineación de ontologías

Alineación de ontologías

Técnicas de alineación de ontologías

- Basado en similitud lingüística (*linguistic matching*)
- Basado en similitud de grafos (*graph matching*)

Datos enlazados o datos vinculados (Linked Data)

Método de publicación de datos estructurados para que se puedan interconectar

Se basa en tecnologías Web, tales como HTTP, FOAF, OWL, RDF y los URI, pero en vez de utilizarlos para páginas web, se extienden para compartir información de una manera que puede ser leída automáticamente por computadores.

Web de enlaces de información interconectadas

- DBpedia - conjunto de datos extraído de Wikipedia; contiene unos 3,4 millones de conceptos descritos por un millardo de tripletas (1000 millones), que incluyen resúmenes en once idiomas
- Bibliografía DBLP - información bibliográfica de artículos científicos, con información de 800.000 artículos, 400.000 autores y aproximadamente 15 millones de tripletas
- riese - datos estadísticos de 500 millones de europeos (el primer conjunto de datos enlazados en XHTML+RDFa)

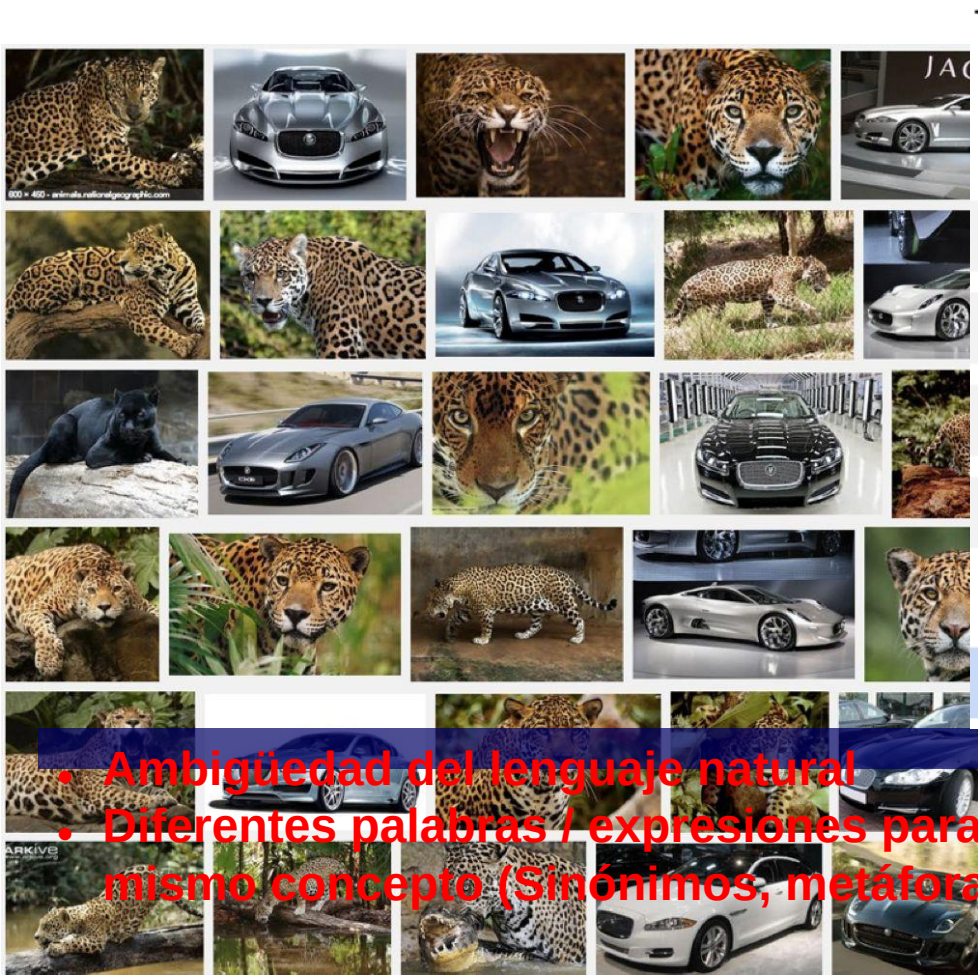
Por qué Linked Data?

- Muchas ontologías con **información similar en algunas de sus partes**:
  - Por ejemplo, Nombres, CI, Dirección, Número telefónico
- Esas partes comunes **podrían interconectarse**, y juntar todos los datos desde múltiples ontologías en una gigante colección de datos, para ser consultada.

Eso debería llevar a crear un **enjambre/araña de ontologías en el mundo**, y cada ontología seria un **nodo del gigante grafo**.

Por qué Linked Data?

Problema en la recuperación de la información



Text: "**Pluto**"

Entity Mapping
Disambiguation

Pluto	a Disney cartoon character
Pluto	a Roman god
Pluto	a song by Björk
HMS Pluto	a ship
...	
Pluto	a dwarf planet

- Ambigüedad del lenguaje natural
- Diferentes palabras / expresiones para el mismo concepto (Sinónimos, metáforas)

¿Por qué no una simple Ontología con los datos interconectados?

- Acceso a los Datos y tiempo de razonamiento seria enorme

- Las cargas de datos en tiempo real seria muy compleja y embotellar la red

- No existe actualmente computador que pudiera procesar esa cantidad de datos masivos

Reglas para Datos enlazados (Linked Data)

- A partir de allí, se debe permitir:
  - Seguir esos enlaces
  - Combinar la información guardad en las ontologías

- Todos los datos (cosas) tienen un URI
- Ese URI es un válido URL
- Debe haber una pagina con ese URL, el cual contenga los datos representados por ese URI
- El URL nunca cambia
- Cuando alguien busca un URI, se provee información útil en RDF.
- Se incluyen instrucciones RDF que enlaza a otros URIs para descubrir cosas relacionadas.

Representación del conocimiento

- How do I represent the following fact:
  "*Pluto has been discovered in 1930*"?

Pluto : Planet
discovered = 1930

UML instance

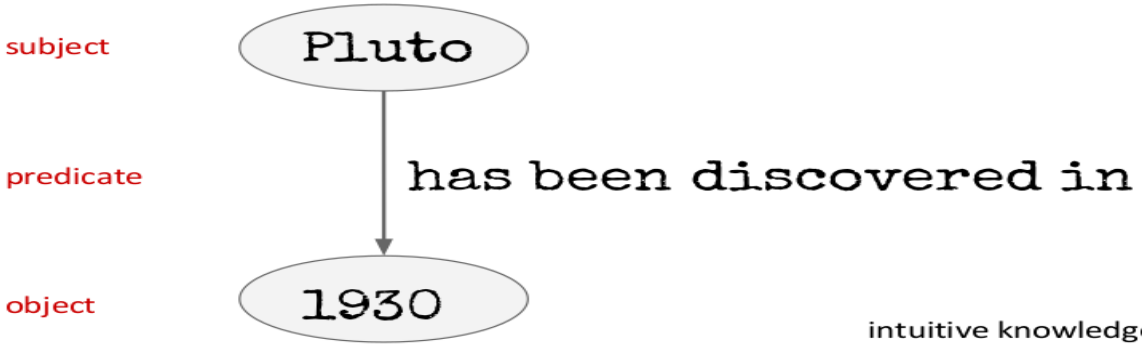
```
<a href="http://en.wikipedia.org/wiki/Pluto">
 Pluto
</a> has been discovered in 1930.
```

HTML

```
<planet name = "Pluto" discovered="1930" />
```

XML

- How do I represent the following fact:
  "*Pluto has been discovered in 1930*" in an intuitive way?



intuitive knowledge representation with a **directed graph**

Representación del conocimiento

- **RDF Statements (RDF-Triple):**

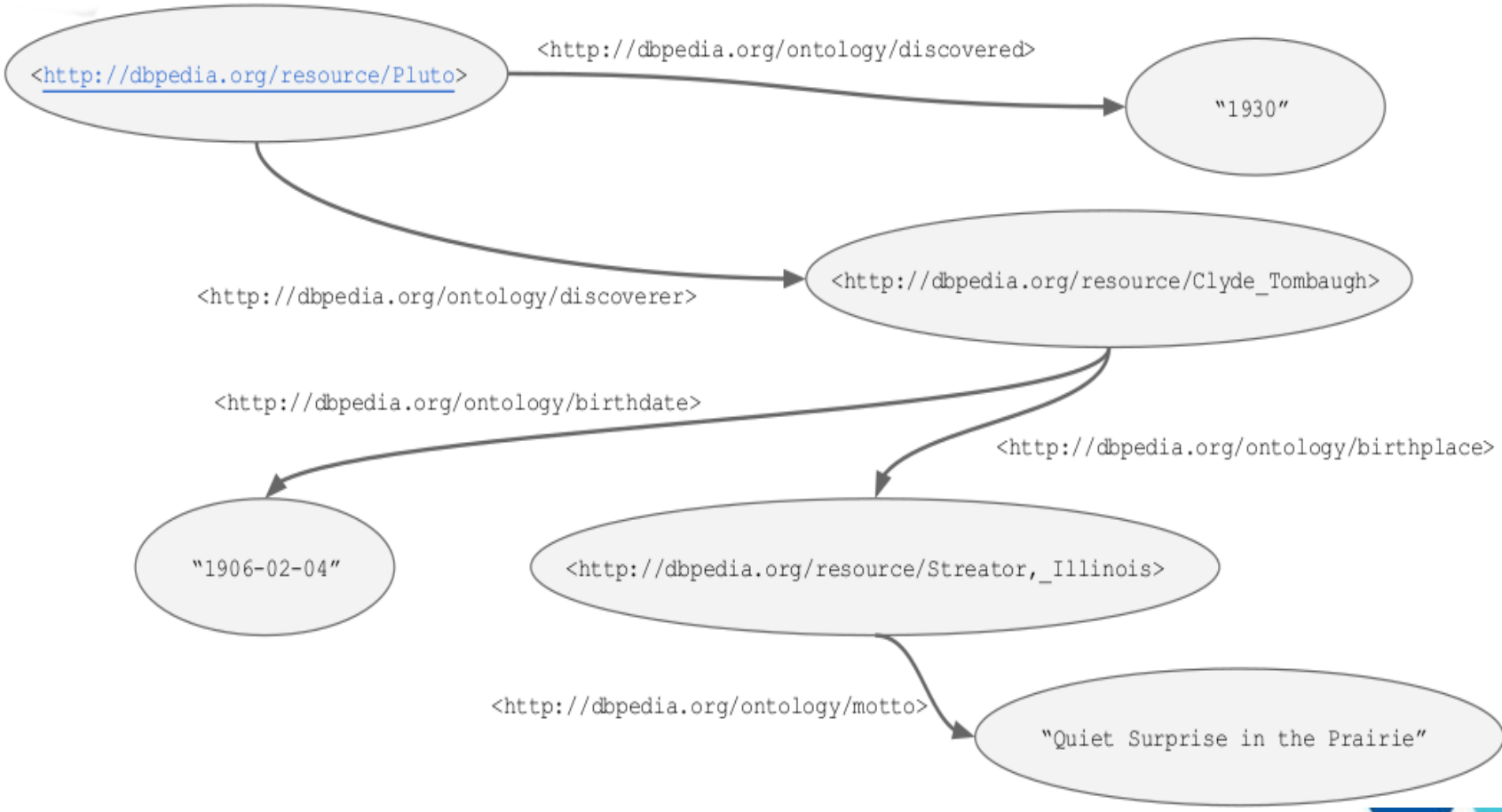
Subject	+	Property	+	Object / Value	
URI		**URI**		**URI / Literal**	RDF Building Blocks

N-Triples Serialization

```
<http://dbpedia.org/resource/Pluto>    <http://dbpedia.org/ontology/discovered>    "1930" .
```

graph representation

Representación del conocimiento



Representación del conocimiento

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> "1930" .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer> <http://dbpedia.org/resource/Clyde_Tombaugh> .
<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/CelestialBody> .
<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/place> .
... ...

<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> "1906-02-04" .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace> <http://dbpedia.org/resource/Streator,_Illinois> .
... ...

<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> "Quiet Surprise in the Prairie" .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> "41.120834"^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> "-88.835281"^^xsd:float .
... ...

Subject Property Object

RDF Triples

- Individuos (Entidades)
- Clases
- Literales / Valores
- Vocabularios / Ontologías

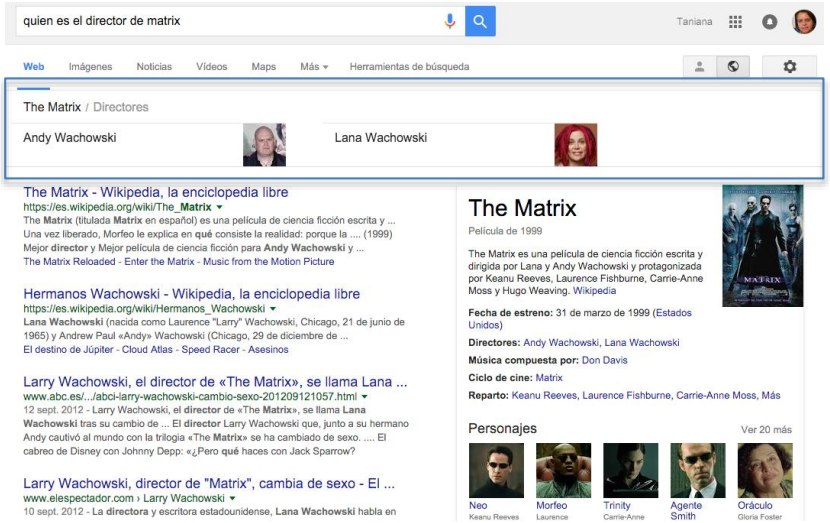
SEO (Search Engine Optimization)

Según Wikipedia, el SEO es:

"es el proceso de mejorar la visibilidad de un sitio Web"

Entidades y tripletas: la base de la Web Semántica

- ya no son palabras claves, se trata ahora de entidades (personas, lugares, organización, eventos, objetos, etc.)
- Las entidades pueden tener múltiples relaciones con otras entidades.



- Andy Wachowski nació en Chicago
- Andy Wachowski es hijo de Lana Wachowski
- Andy Wachowski dirigio The Matrix

- Información puede ser extraída de diferentes fuentes: Dbpedia, IMDB, Wikipedia etc.
- Basado en una representación del conocimiento
  Sujeto + Predicado +Objeto

El sujeto es la entidad que se esta describiendo,
el predicado es que se esta describiendo del sujeto
el objeto es el valor del predicado

SEO Semántico

Andy Wachowski nació en Chicago
Andy Wachowski es hijo de Lana Wachowski
Andy Wachowski dirigió The Matrix

- ✓ Las tripletas se representa en grafos.
- ✓ **Knowledge Graph** -> es una base de datos de entidades y relaciones entre ellas

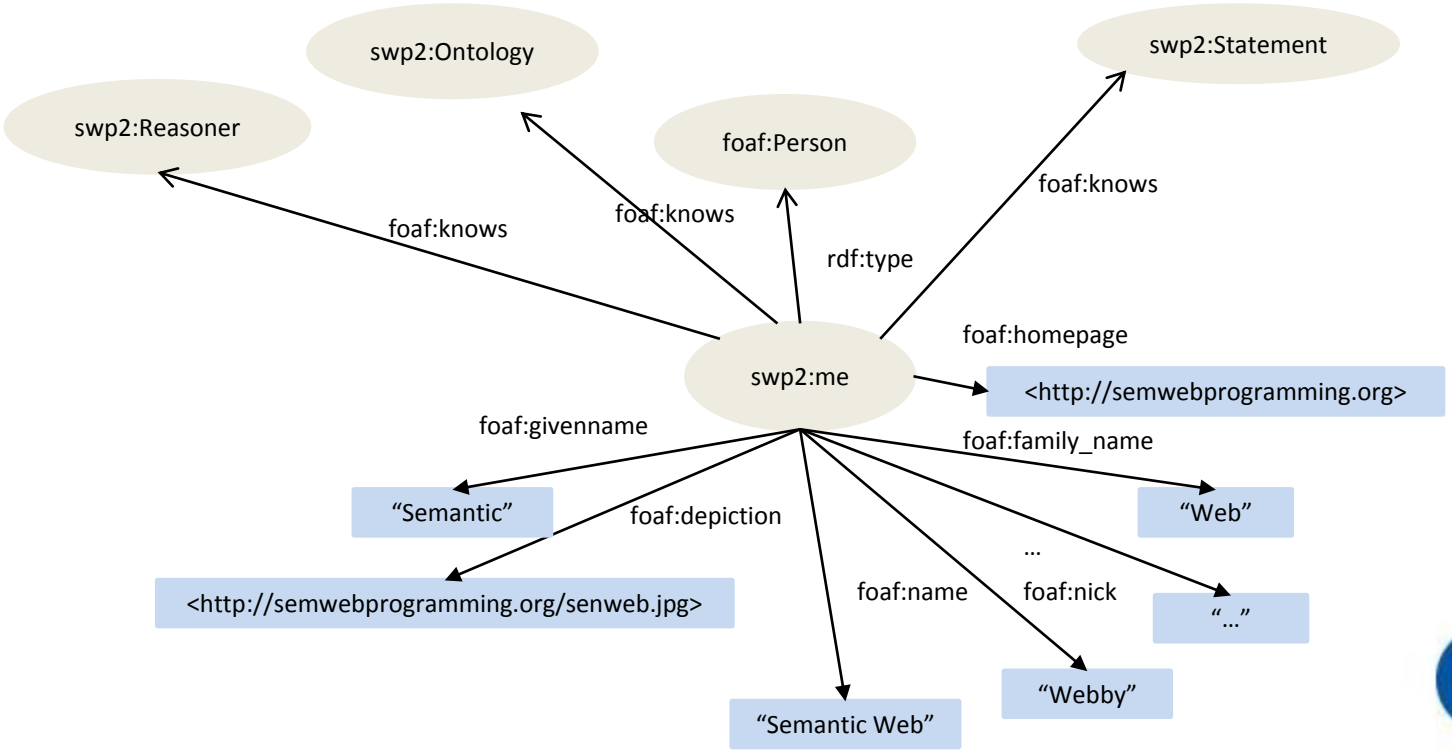
✓SEO semántico tiene como objetivo de ayudar a los buscadores a entender exactamente de qué trata tus páginas.

✓ Para ello, sigue los siguientes pasos

- ➢Determinar las entidades correspondientes a la página.
- ➢Desambiguarlas directamente
- ➢Desambiguarlas indirectamente.

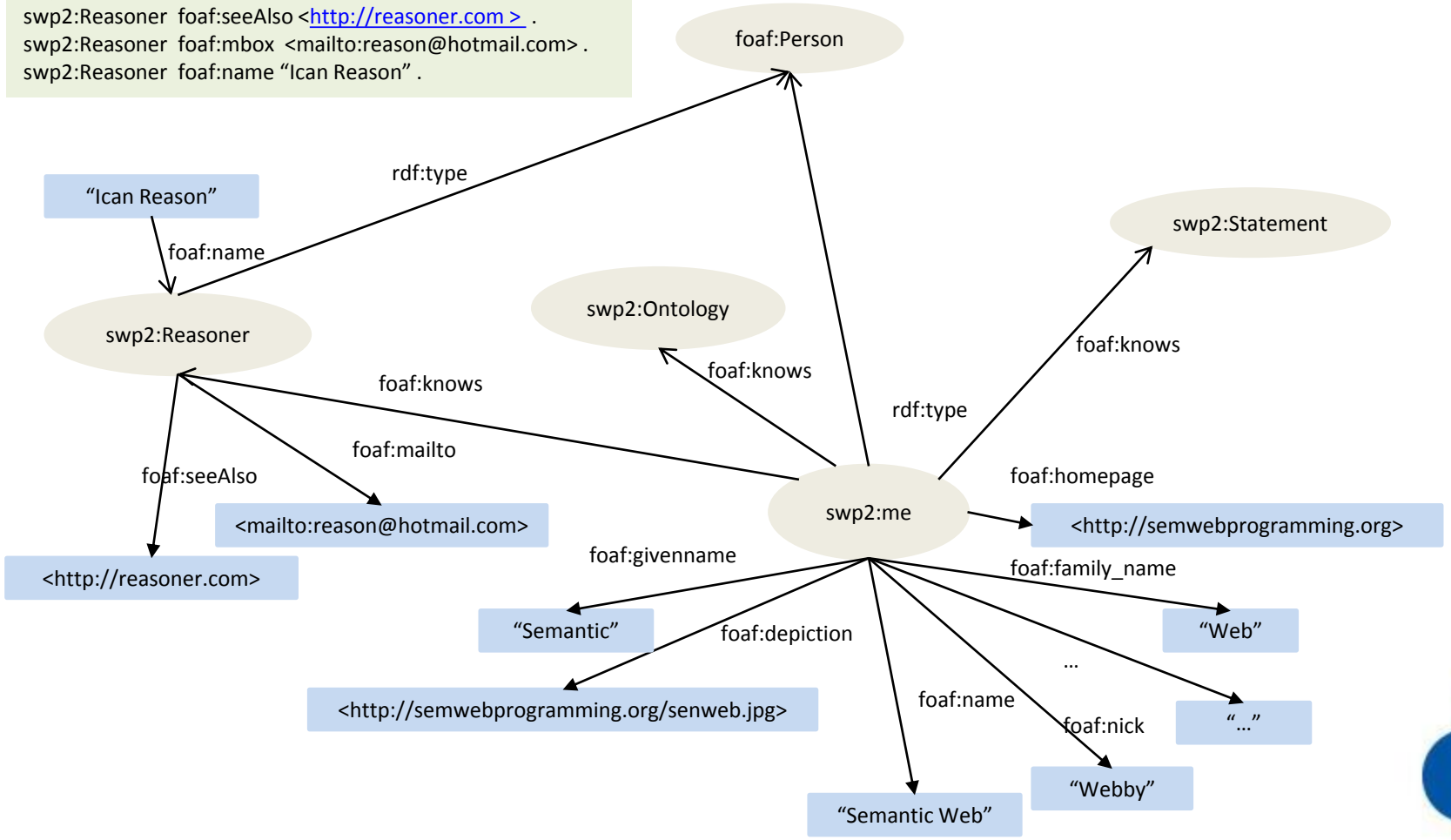
swp2:me rdf:type foaf:Person .
swp2:me foaf:depiction <http://semwebprogramming.org/senweb.jpg >.
swp2:me foaf:family_name "Web" .
swp2:me foaf:givenname "Semantic" .
swp2:me foaf:homepage <http://semwebprogramming.org >.
swp2:me foaf:knows "Reasoner" .
swp2:me foaf:knows "Statement" .
swp2:me foaf:knows "Ontology" .
swp2:me foaf:name "Semantic Web" .
swp2:me foaf:nick "Webby" .
swp2:me foaf:phone "<tel:410-679-8999>" .
swp2:me foaf:schoolInfoHomepage <http://www.web.edu > .
swp2:me foaf:title "Dr." .
swp2:me foaf:workInfoHomepage <http://semwebprogramming.com/dataweb.html> .
swp2:me foaf:workplaceHomepage <http://semwebprogramming.com > .

Ejemplo



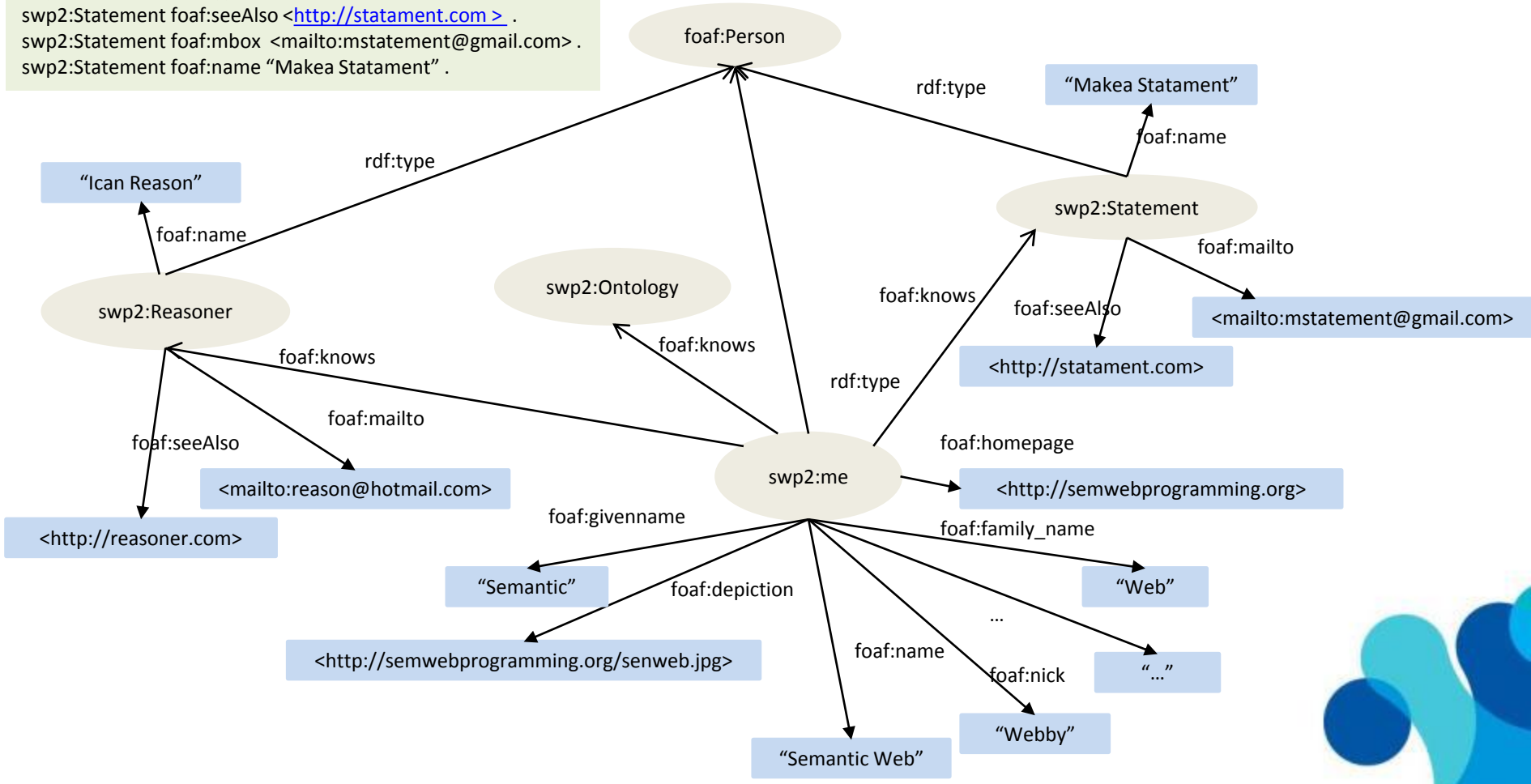
Ejemplo

swp2:Reasoner rdf:type foaf:Person .
swp2:Reasoner foaf:seeAlso <http://reasoner.com > .
swp2:Reasoner foaf:mbox <mailto:reason@hotmail.com> .
swp2:Reasoner foaf:name “Ican Reason” .



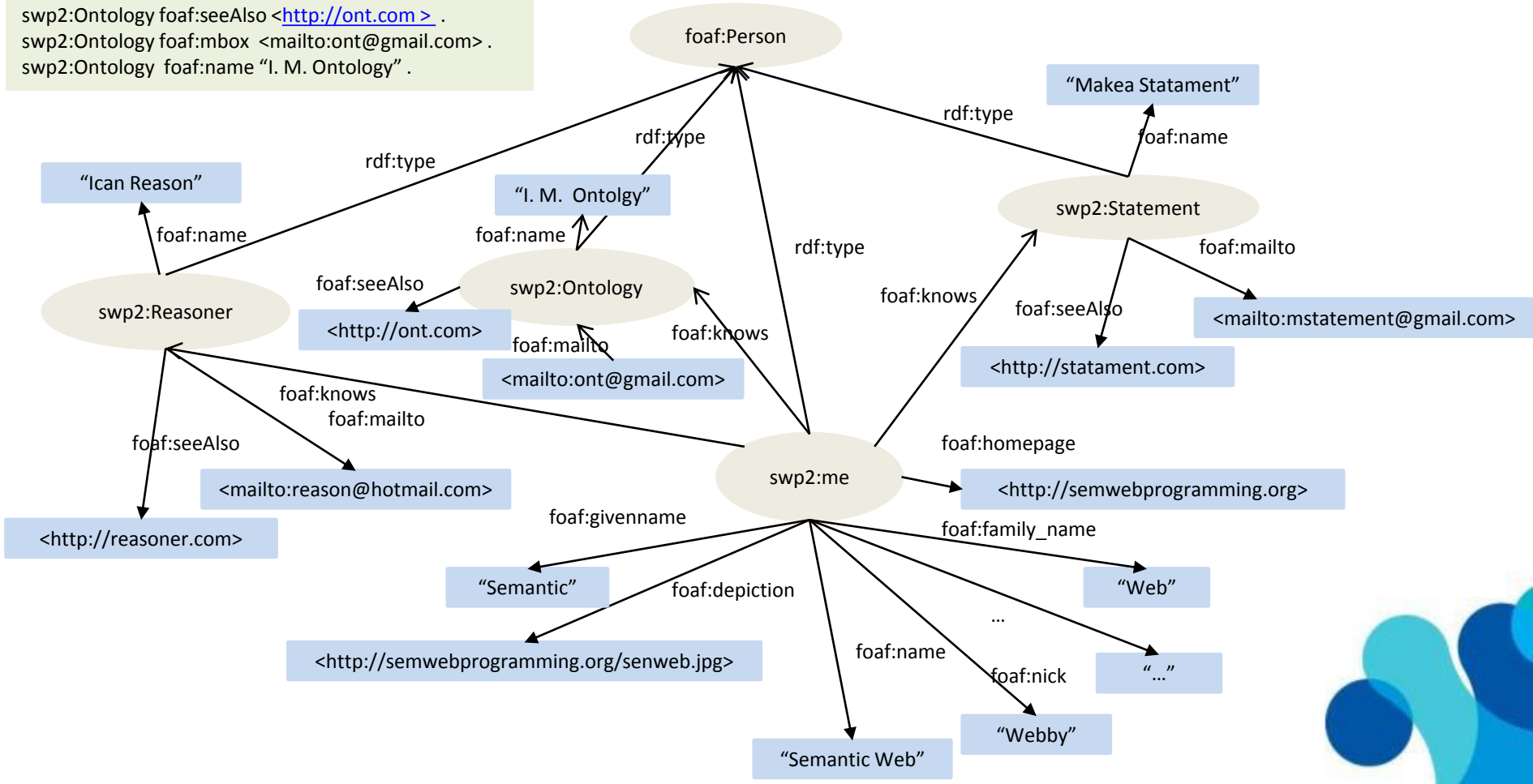
Continuación del Ejemplo

swp2:Statement rdf:type foaf:Person .
swp2:Statement foaf:seeAlso <http://statament.com > .
swp2:Statement foaf:mbox <mailto:mstatement@gmail.com> .
swp2:Statement foaf:name “Makea Statament” .



Continuación del Ejemplo

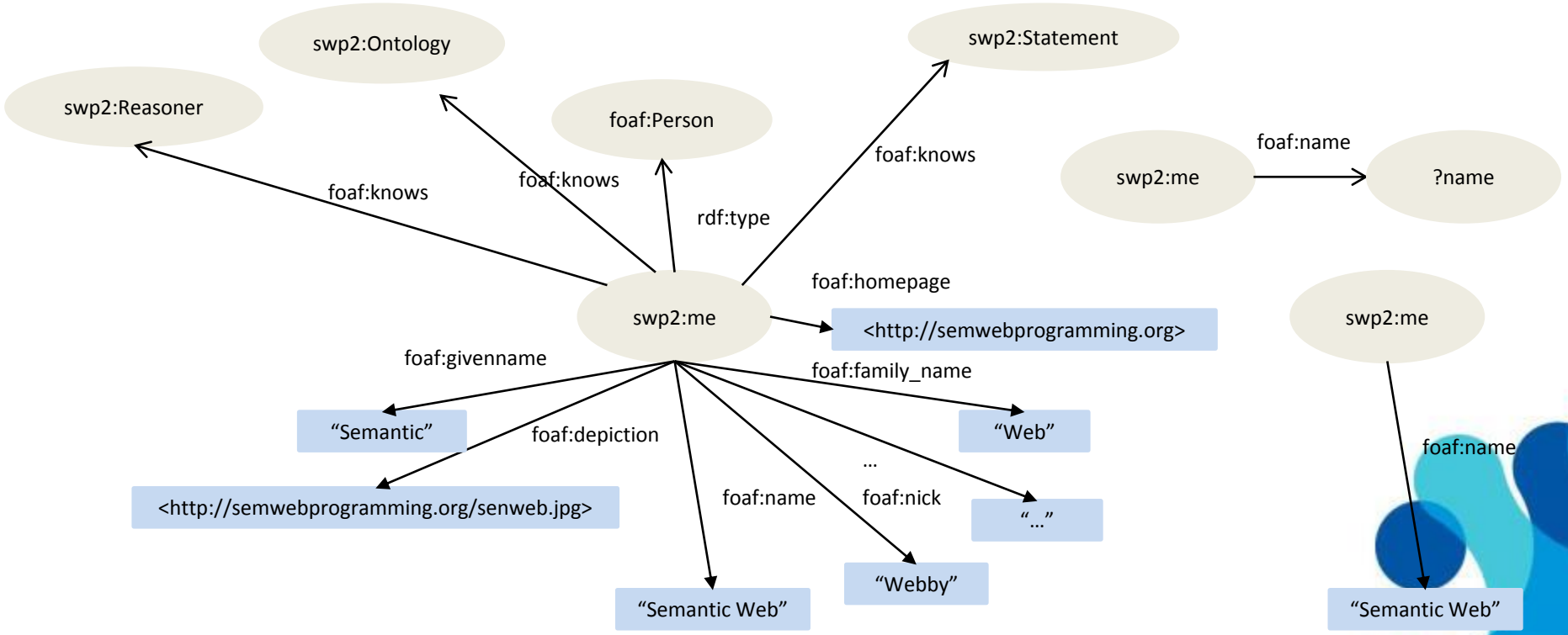
swp2:Ontology rdf:type foaf:Person .
swp2:Ontology foaf:seeAlso <http://ont.com > .
swp2:Ontology foaf:mbox <mailto:ont@gmail.com> .
swp2:Ontology foaf:name “I. M. Ontology” .



swp2:me rdf:type foaf:Person .
swp2:me foaf:depiction <http://semwebprogramming.org/senweb.jpg >.
swp2:me foaf:family_name “Web” .
swp2:me foaf:givenname “Semantic” .
swp2:me foaf:homepage <http://semwebprogramming.org >.
swp2:me foaf:knows “Reasoner” .
swp2:me foaf:knows “Statement” .
swp2:me foaf:knows “Ontology” .
swp2:me foaf:name “Semantic Web” .
swp2:me foaf:nick “Webby” .
swp2:me foaf:phone “<tel:410-679-8999>” .
swp2:me foaf:schoolInfoHomepage <http://www.web.edu > .
swp2:me foaf:title “Dr.” .
swp2:me foaf:workInfoHomepage <http://semwebprogramming.com/dataweb.html> .
swp2:me foaf:workplaceHomepage <http://semwebprogramming.com > .

Consulta

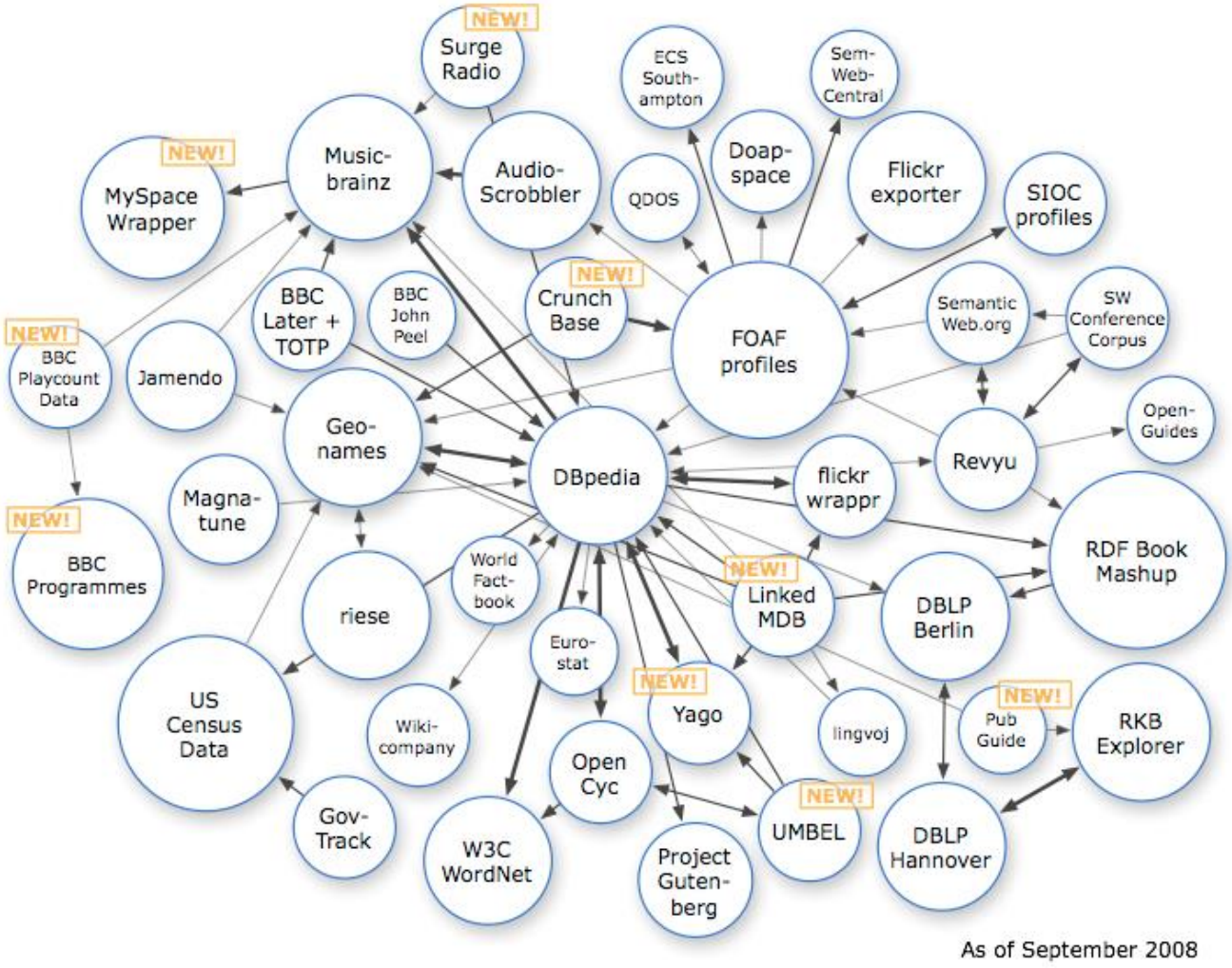
```
select DISTINCT ?name
where{
      swp2:me foaf:name ?name
}
```



Septiembre 2008

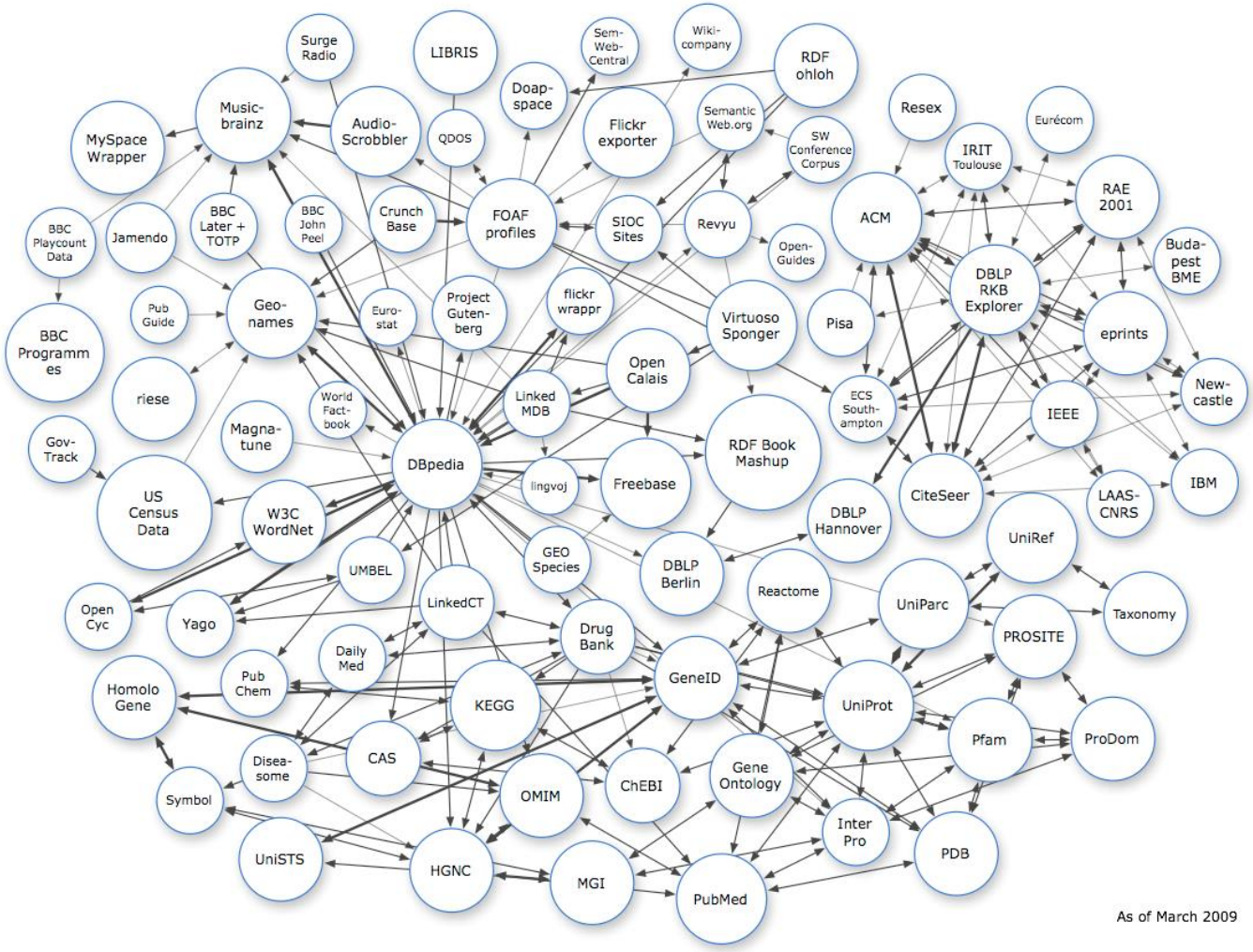
DBpedia

es un proyecto para la extracción de datos de Wikipedia



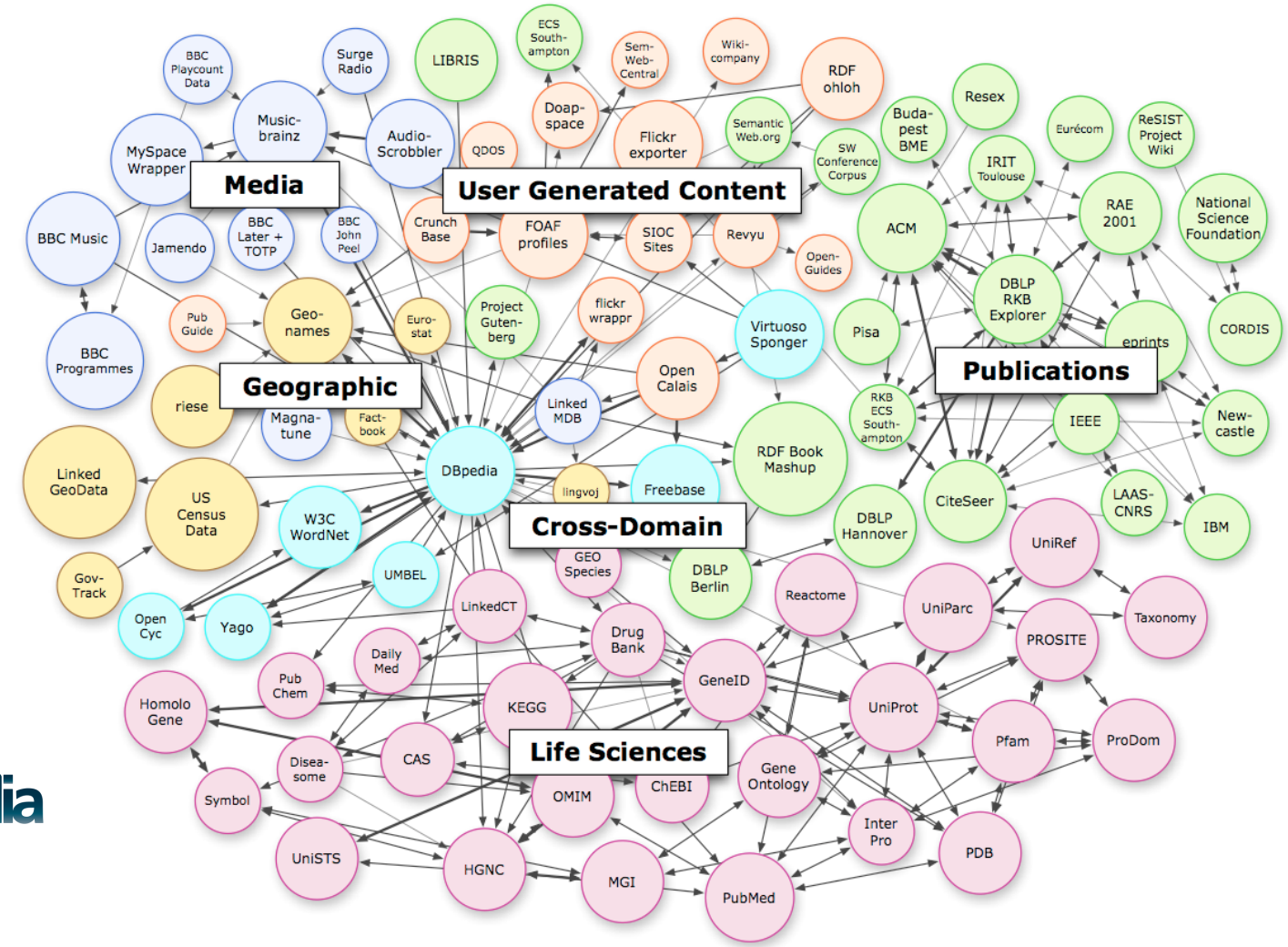
Solo en la versión en inglés, se describen 3,77 millones de entidades, entre ellas al menos 764 mil personas, 563 mil lugares, 112 mil álbumes de música, 72 mil películas y 18 mil videojuegos.

Marzo 2009



Con todas las versiones se tienen 8 millones de enlaces a imágenes, 24,4 millones de enlaces a páginas externas, 27,2 millones de enlaces a *datasets* externos y 55,8 millones categorías de Wikipedia

Julio 2016

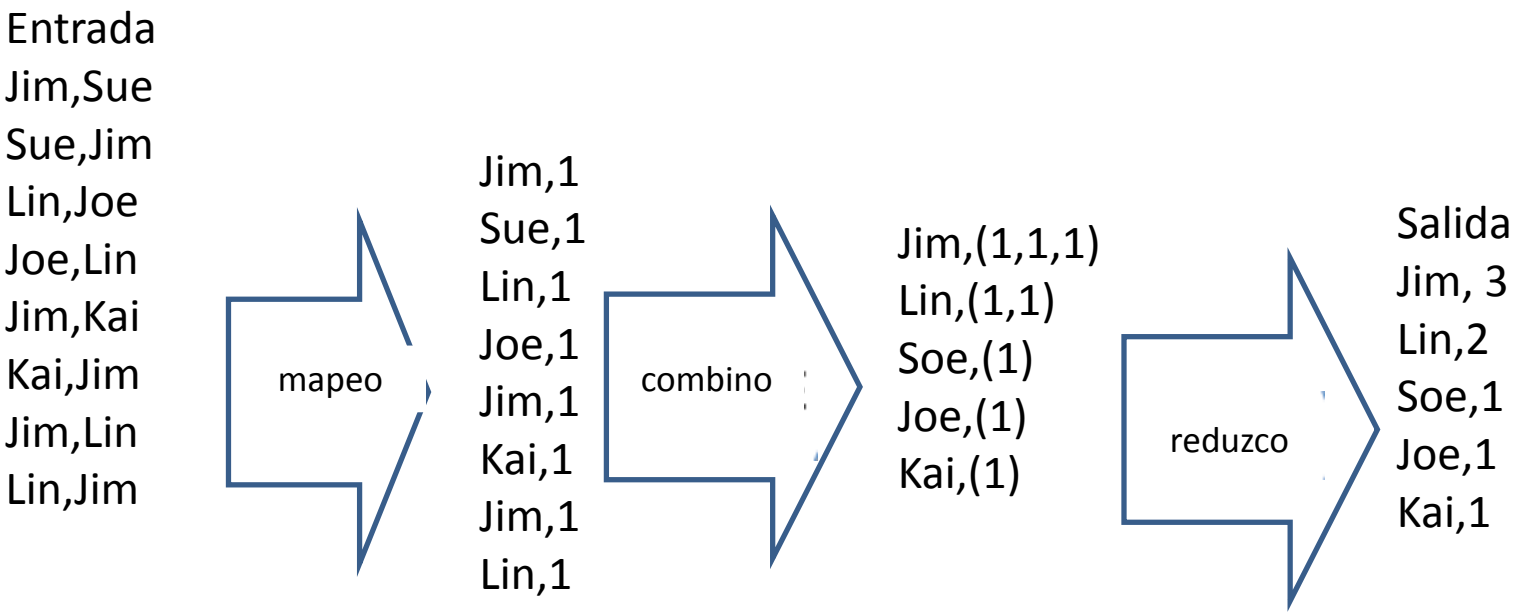


En mayo de 2012 se lanzó el sitio web de DBpedia para el idioma español

Minería de Gráfos y Redes

- Minería de Patrón de Gráfo
- Modelado estadístico de Redes
- Agrupación y clasificación de grafos y redes homogéneas
- Agrupación, clasificación de las Redes heterogéneos
- Descubrimiento, clases, y Predicción de Enlace en Redes de Información
  Búsqueda de Similitud en Redes de Información
- Evolución de las redes de información social

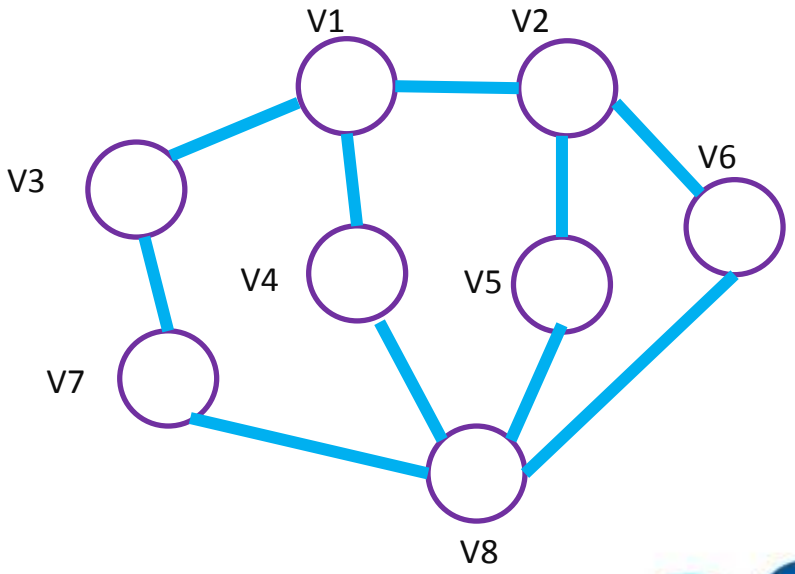
Simple análisis redes sociales: Contar amigos



Grafos

Un grafo **G** es un par ordenado de un conjunto de vértices **V** y un conjunto de aristas **E**

G = (V,E)

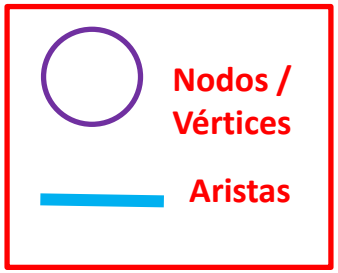


Par ordenado:

(a, b) ≠ (b,a) si a ≠ b

Par No ordenado:

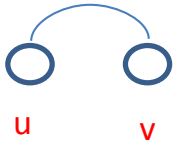
{a,b} = {b,a}



Grafos

Aristas:

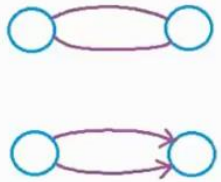
Dirigido
(u, v)
(v, u)

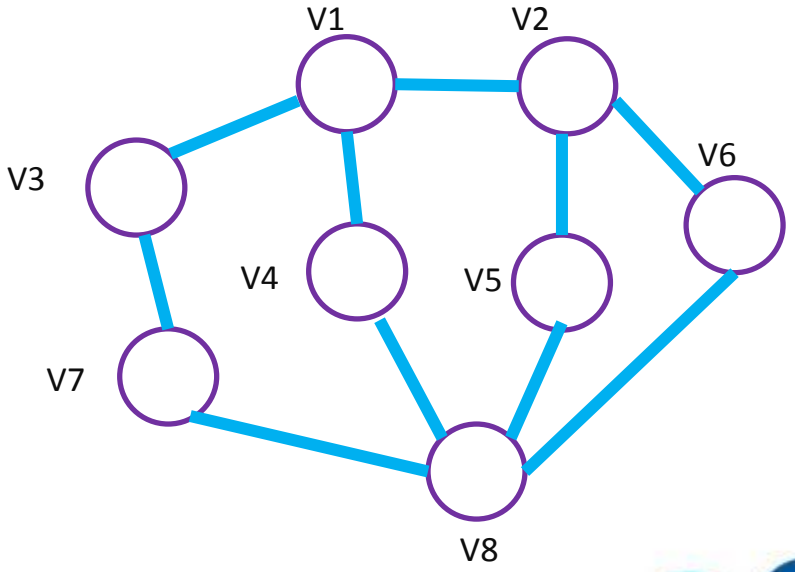


No Dirigido
{u,v}

Bucle

Varias
Aristas



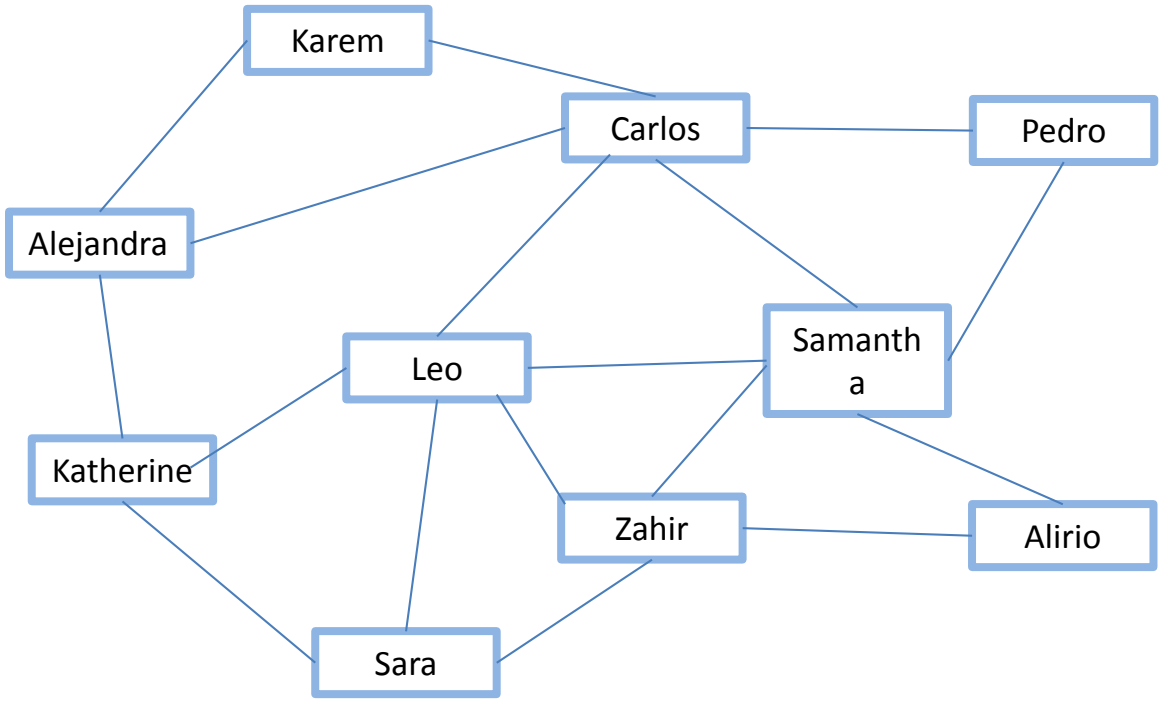


V = {V1. V2. V3, V4, V5, V6, V7, V8}

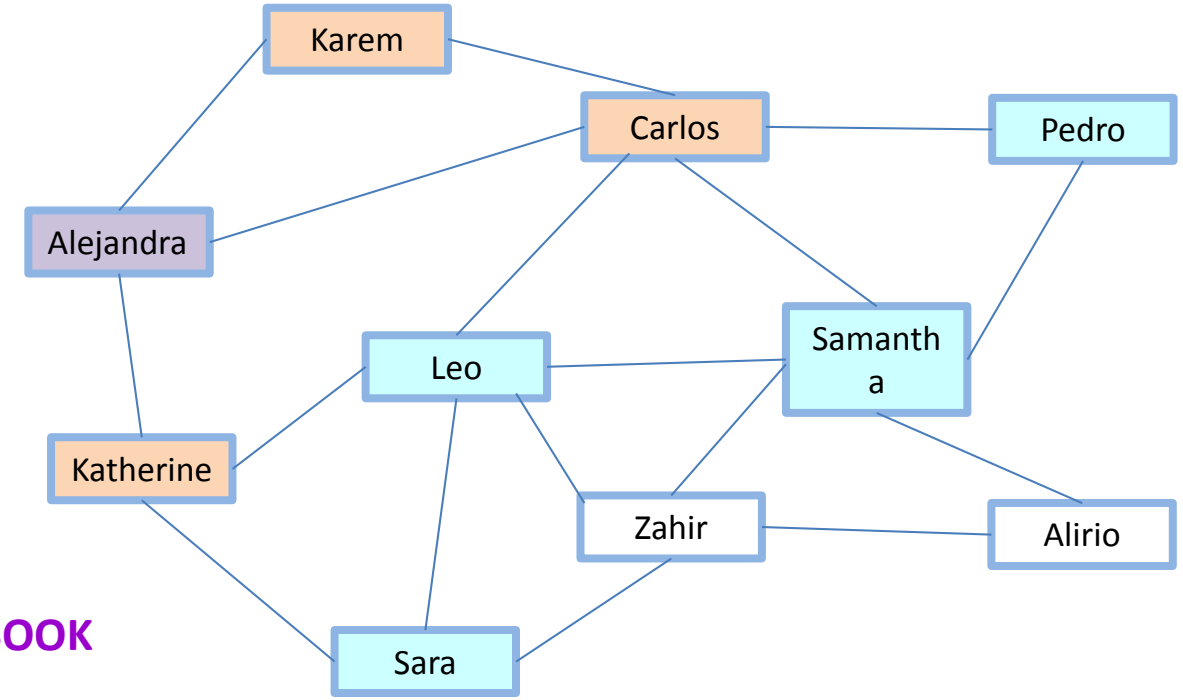
E = {{v1. v2}, {vi,v3}, {v1,v4}, {v2,v5}, {v2,v6}, {v3, v7}, {v4, v8}. {v7, v8}, {v5.,v8}, {v6, v8}}

Grafos

Red Social FACEBOOK



Grafos

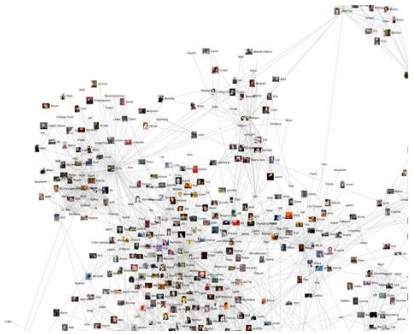


Red Social FACEBOOK

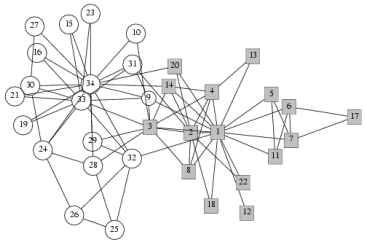
Para Sugerir un amigo a ALEJANDRA hay que encontrar todos los nodos que tengan longitud del camino igual a 2.

Redes en el mundo real

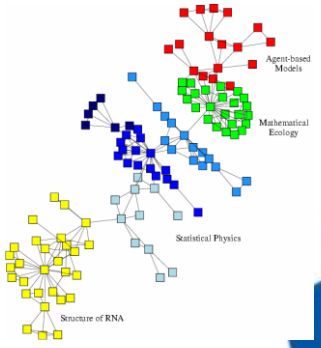
- **Redes de información:**
  - World Wide Web: hyperlinks
  - Redes de citación
  - Redes de Noticias y Blogs
- **Redes sociales**
  - Organizativas
  - Comunicativas
  - Colaborativas
  - Contactos sexuales
- **Redes tecnológicas:**
  - Energéticas
  - Transporte (aéreo, carreteras, fluviales,…)
  - Telefónicas
  - Internet
  - Sistemas Autónomos



Redes de amistad

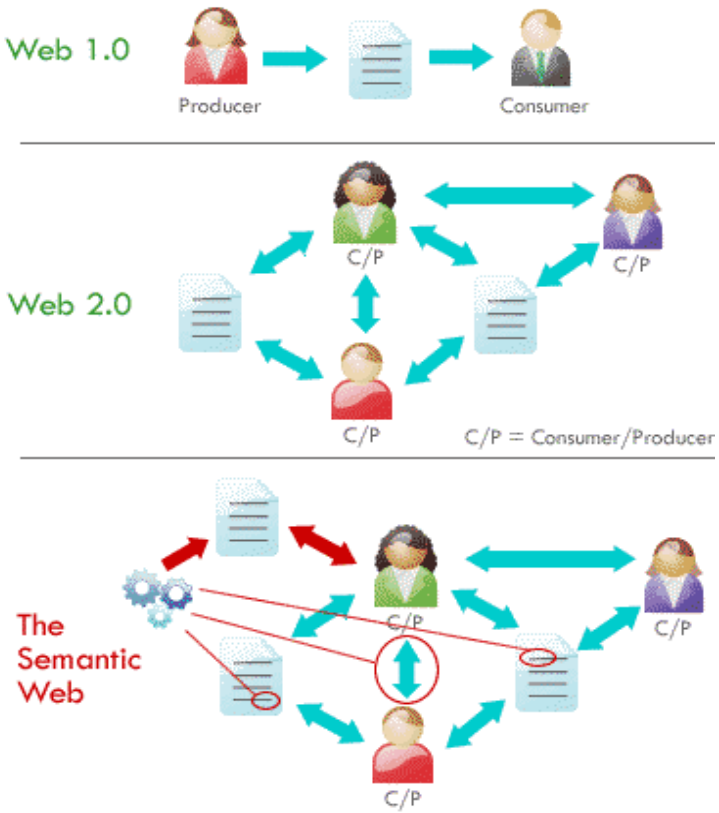


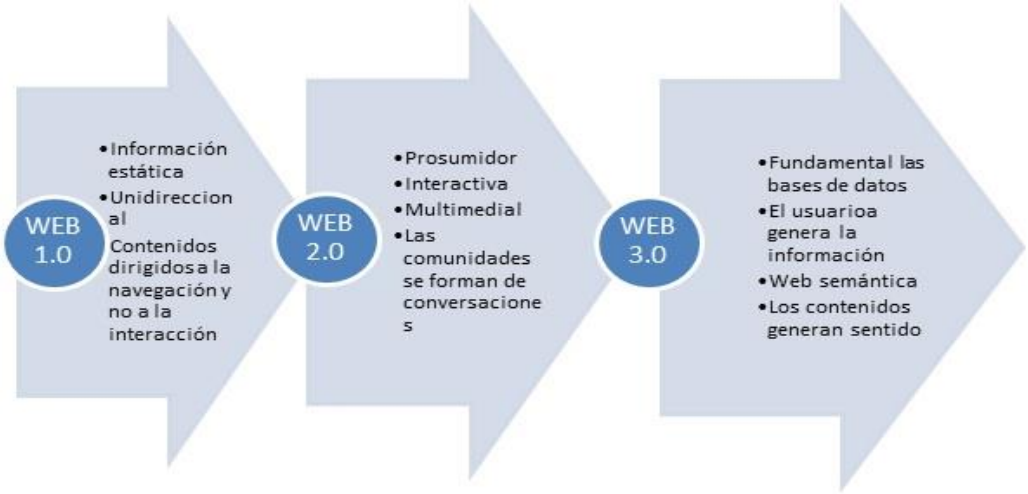
Karate club network



Redes de colaboración

LA WEB 2.0 Y LA WEB 3.0





RED SOCIAL VIRTUAL
WEB 2.0 Y 3.0

- **Facebook** es la red social más utilizada del mundo
- **Twitter:** red social de microblogging.
- **LinkedIn** red de usuarios profesionales, y
- **Youtube** red de alojamiento de vídeos.
- **Google+**, apuesta de Google por las redes sociales.
- **Instagram,** red de intercambio de imagenes

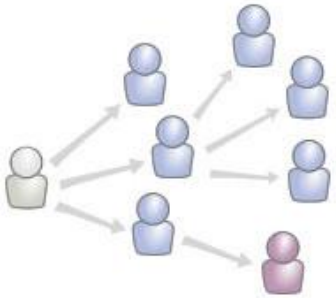
RSV

RED SOCIAL (RS)

COMUNIDAD VIRTUAL (CV)

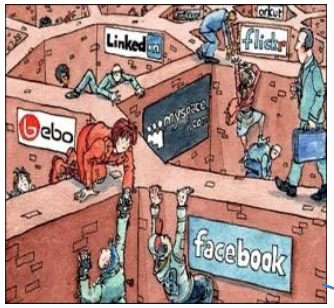
SITIOS DE REDES SOCIALES (SNS: Social Network Services – Servicios de Redes Sociales)

RED SOCIAL VIRTUAL
WEB 2.0 Y 3.0



RED SOCIAL (RS)

- CONJUNTO DE PERSONAS O ENTIDADES
- COMPARTEN INTERESES, VINCULADAS POR CARACTERISTICAS Y OBJETIVOS A FINES
- INTERCAMBIAN INFORMACIÓN DE TODA CLASE: FINANCIERA, AMISTAD, OCIO, ACADEMICA, ENTRE OTRA.



COMUNIDAD VIRTUAL (CV)

- CONJUNTO DE PERSONAS, ENTIDADES O GRUPOS SOCIALES
- CON UN MISMO OBJETIVO O PROPOSITO
- SE APOYA EN TECNOLOGIAS, FUNDAMENTALMENTE EN INTERNET

OFRECEN HERRAMIENTAS Y APLICACIONES O RECURSOS INFORMATICOS (SS: SOFTWARE SOCIAL), PARA IMPLEMENTAR LAS CV

SITIOS DE REDES SOCIALES (SNS: Social Network Services – Servicios de Redes Sociales)

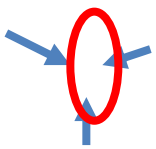
FORMAN PARTE DEL SS: SITIOS TIPICOS COMO FACEBOOK, ENTRE OTROS, Y OTROS MAS GENERICOS COMO BLOG, FOROS,.

Herramientas

- **Gephi** (visualization and basic network metrics)
- **NetLogo** (modeling network dynamics)
- **Pajek:** amplia funcionalidad basada en menús, incluyendo muchas, muchas métricas de red y manipulaciones
  - pero ... no extensible
- **Guess:** extensibles, herramientas de secuencias de comandos de análisis exploratorio de datos, pero la selección más limitada de métodos incorporados en comparación con Pajek
- **NetLogo:** plataforma general agente basado en la simulación con el apoyo de modelado excelente red
  - muchos de los demos en este curso fueron construidos con NetLogo
- **IGRAPH:** utilizado en la versión de nivel de doctorado. bibliotecas se puede acceder a través de R o Python. Rutinas escalan a millones de nodos. (for programming assignments)

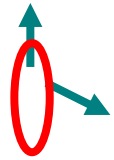
Métricas:
Propiedades de los nodos de la Red

- Conexiones
  - indegree
    cuantos arcos estan dirigidos al nodo
  - outdegree
    arcos que salen del nodo
  - degree (in or out)
    todos los arcos del nodo, entrada y salida



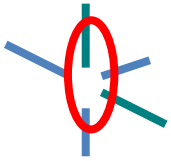
indegree=3

$$\sum_{i=1}^{n} A_{ij}$$



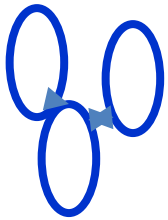
outdegree=2

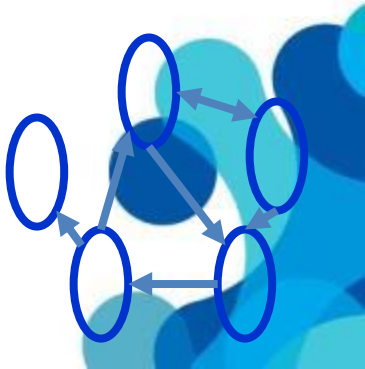
$$\sum_{j=1}^{n} A_{ij}$$



degree=5

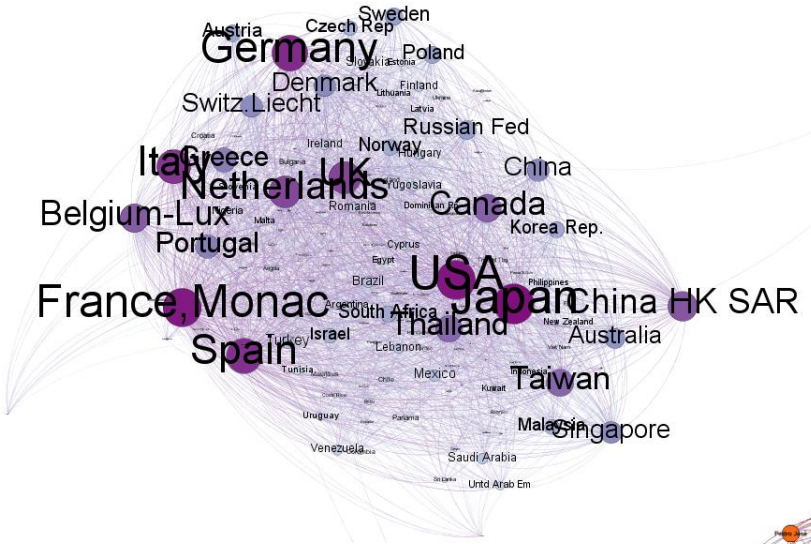
- Degree sequence: Lista ordenada de los grados de cada nodo
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]



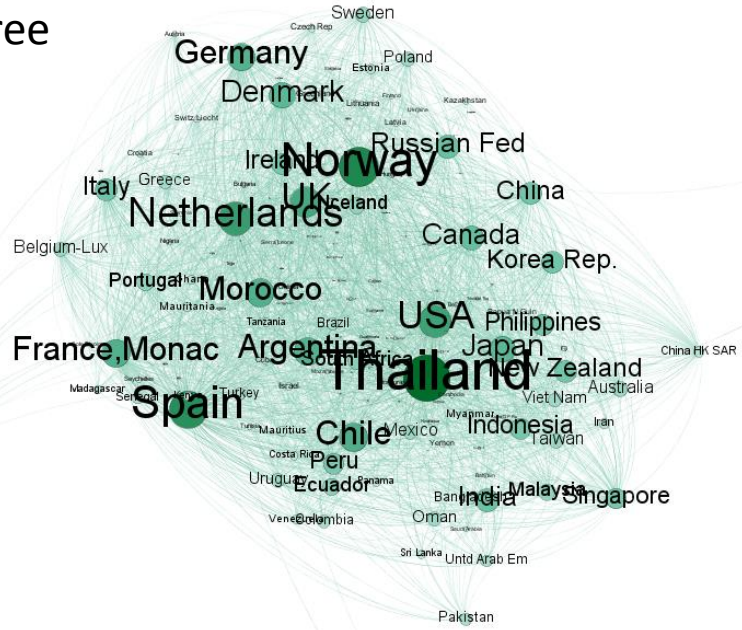


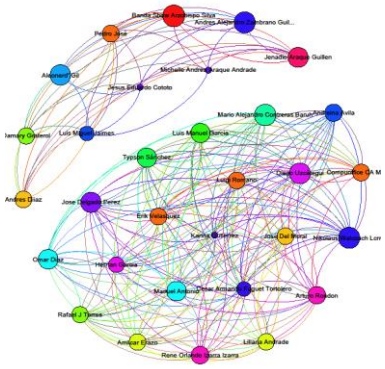
Métricas

InDegree



OutDegree





Métricas de redes

- **Cada métrica de red da respuesta a las siguientes preguntas:**

- pregunta: **¿Quién es más central?**

  1) METRICA DE RED: centralidad

  a)Centralidad de grado (degree centrality).
  1) Indegree o grado de entrada
  2) Outdegree o grado de salida

  b) Centralidad de cercanía (closeness centrality).

  c) Centralidad de intermediación (Betweenness centrality).

- pregunta: **¿Todo está conectado?**

  2) METRICA DE RED: los componentes conectados

  - Componentes fuertemente conectados:
  -Componentes Débilmente conectados:

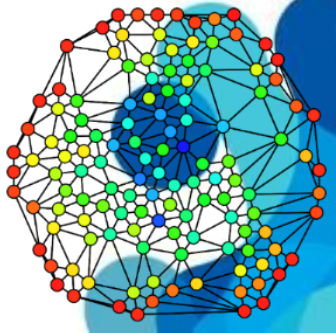
  3) METRICA DE RED: tamaño de componente gigante(giant component)

- pregunta: **¿A qué distancia están las cosas?**

  4) METRICA DE RED: rutas más cortas

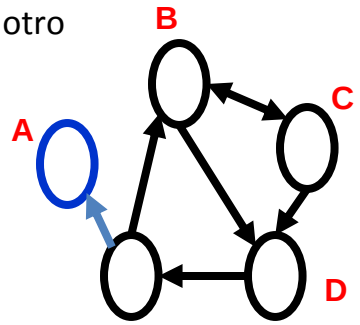
- pregunta: **¿Cómo densa son?**

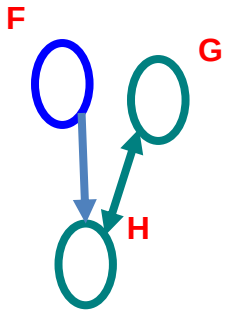
  5) METRICA DE RED: densidad grafo



Métricas:
Componentes conectados

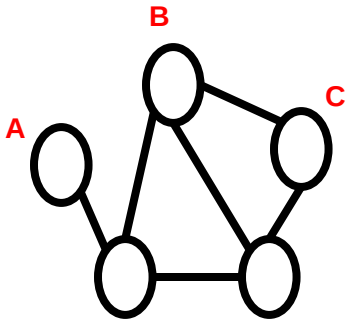
- **Componentes fuertemente conectados:**
  - Cada nodo dentro del componente se puede llegar desde cualquier otro nodo en el componente siguiendo los enlaces dirigidos
    - Componentes fuertemente conectados
      - B C D E
      - La
      - G H
      - F





Componentes Débilmente conectados:

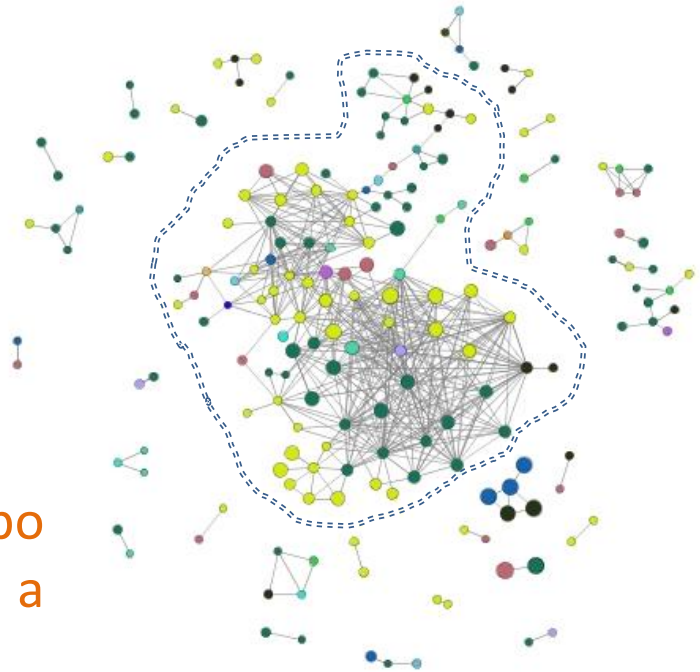
- Cada nodo se puede llegar desde cualquier otro nodo siguientes enlaces en cualquier dirección
  - Componentes débilmente conectados
    - A B C D E
    - G H F



En las **redes no dirigidos** se habla simplemente de "**componentes conectados** '

Métricas: Componentes conectados

- Si el componente más grande ocupa una región significativa de la red o grafo, es llamado **giant component**

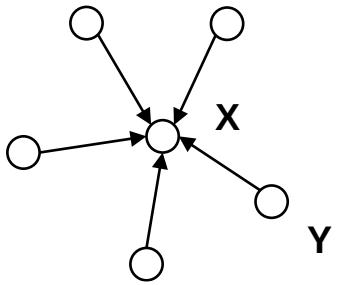


El componente gigante, consiste en un grupo de nodos enlazados entre si, y que agrupan a la mayoría de los nodos de la red.

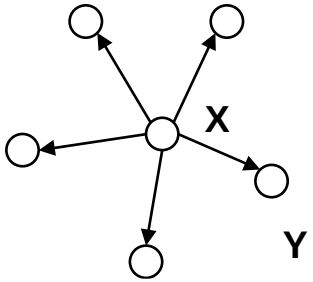
El componente gigante aparece también en casi todas las redes sociales.

Métricas: Centralidad

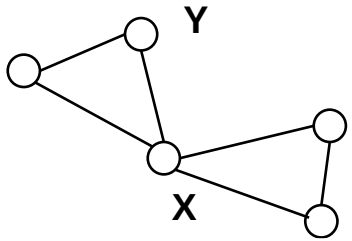
medida posible de un vértice en un grafo, que determina su importancia relativa dentro de éste



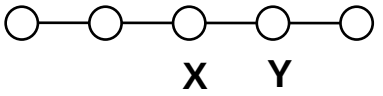
indegree



outdegree

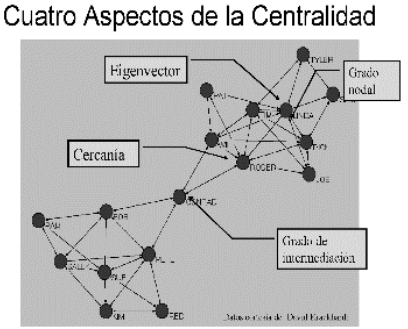


Betweenness (intermediación)

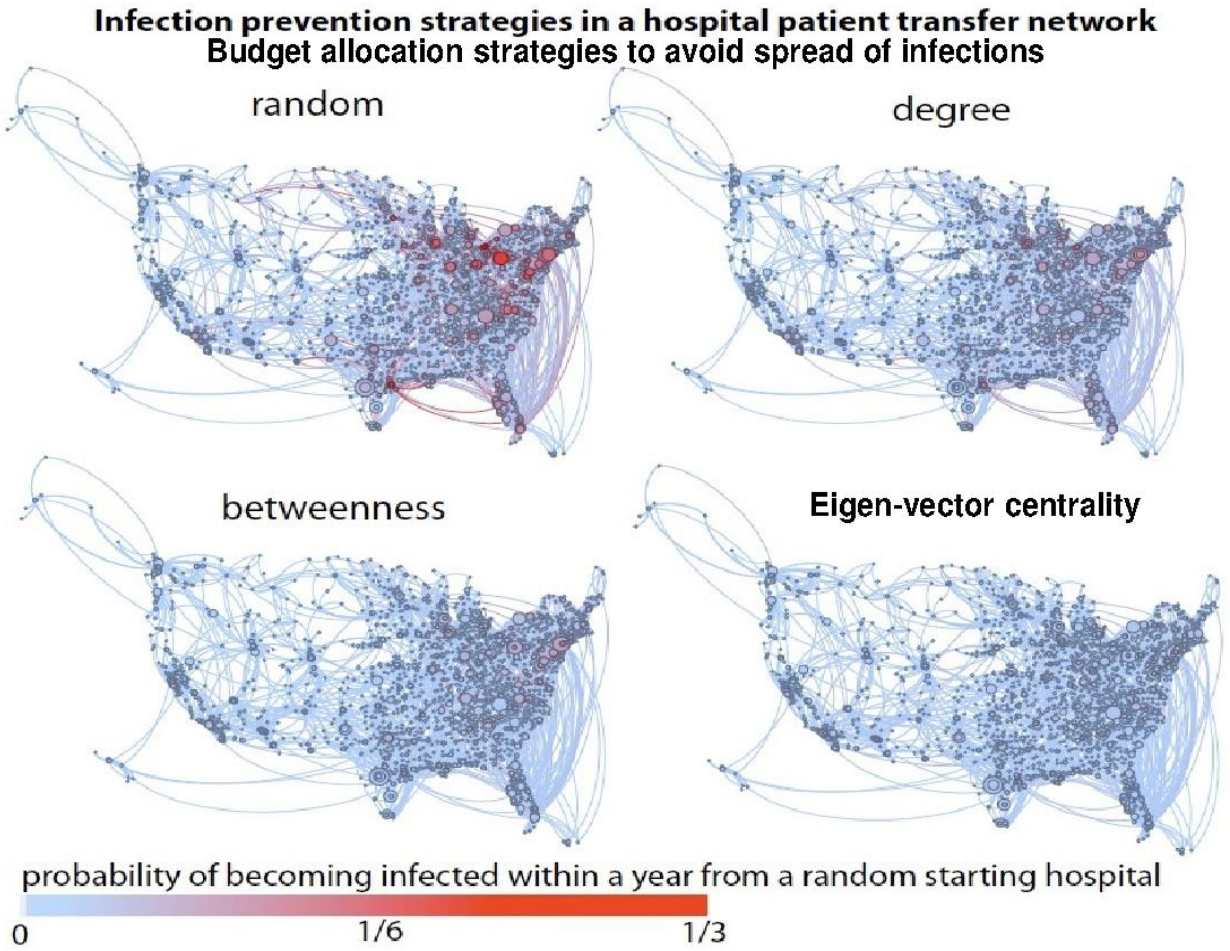


Closeness (cercanía)

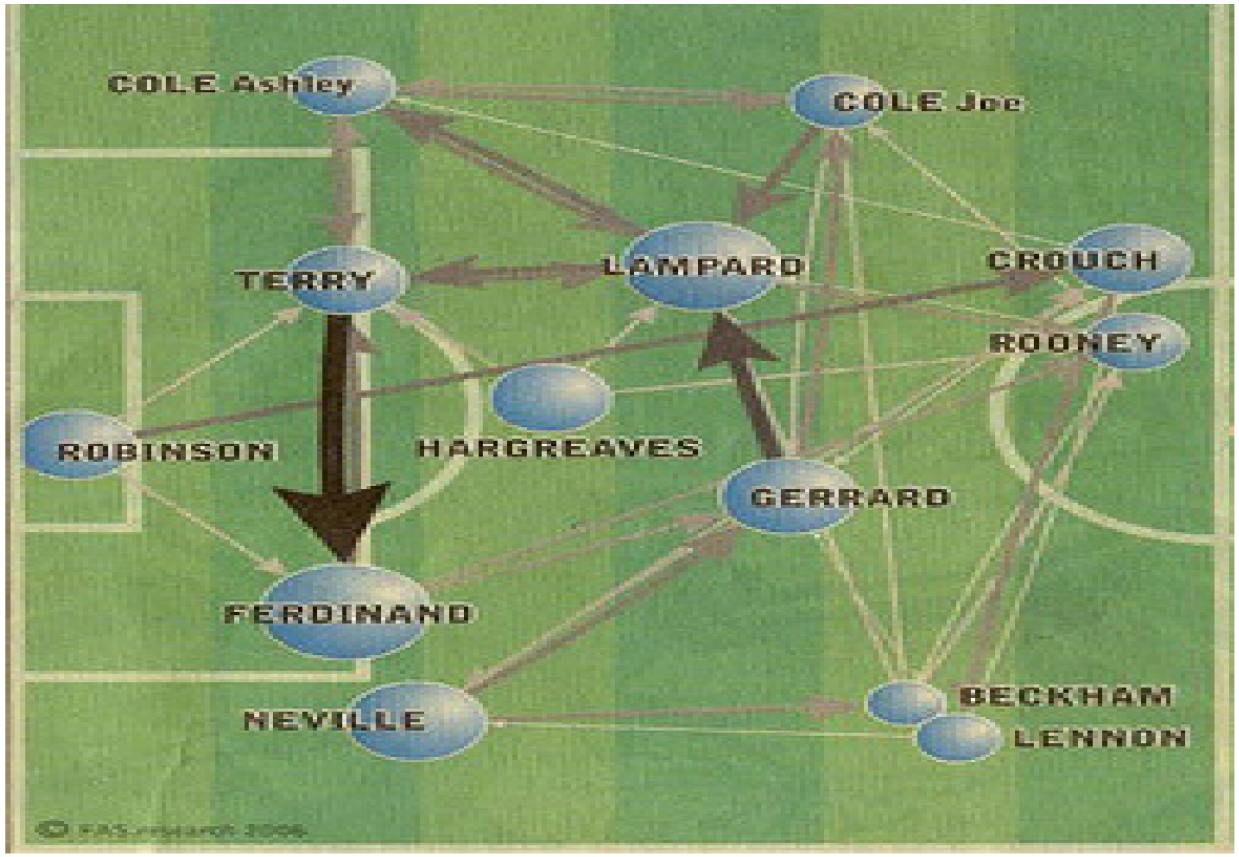
La centralidad de vector propio («eigenvector centrality»).



APLICACIONES | **Infecciones en hospitales EEUU por transferencia de pacientes (4)**



Análisis de juego de futbol usando grafos



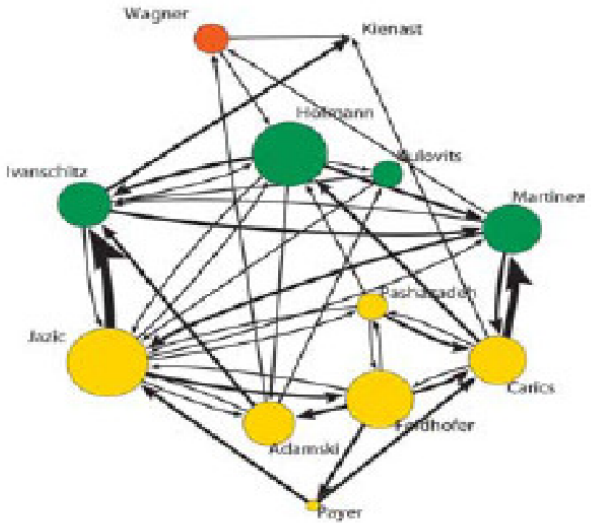
Chelsea FC:

Robinson	1
Cole Ashley	2
Terry	3
Ferdinand	4
Neville	5
Cole Joe	6
Lampard	7
Heargraves	8
Gerrard	9
Beckham	10
Lennon	11
Crouch	12
Rooney	13

Análisis de juego de futbol usando grafos

Posibles factores de análisis:

- ¿Qué jugador ha iniciado más pases (grado ponderado de salida)? **Jazic**
- ¿Qué jugador ha recibido más pases (grado ponderado de entrada)? **Jazic**
- ¿Quién ha controlado el juego del Rapid (centralidad)? **Jazic** y **Hoffman**
- ¿Qué jugadores han estado implicados en jugadas con el mayor número de pases (caminos)? **Jazic**, **Hofmann**, **Feldhofer**, **Martinez** y **Carics**
- ¿Quién ha jugado con quién y quién no (análisis de los enlaces)? **Ni un solo pase de Ivanschitz a Wagner**
- ¿Qué grupos de jugadores han compuesto la columna vertebral del equipo (análisis de triadas)? **Por ejemplo, Feldhofer-Carics-Pashazadeh**
- ¿Qué jugadores han tenido un rol similar (análisis de enlaces)? **Por ejemplo, Ivanschitz / Martinez**

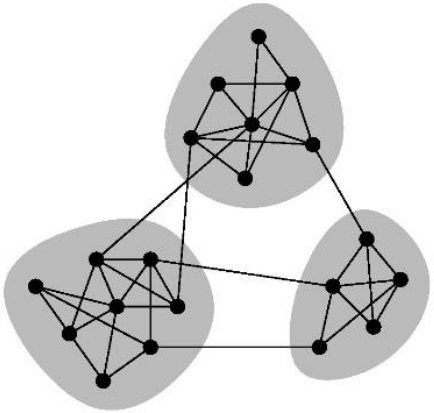


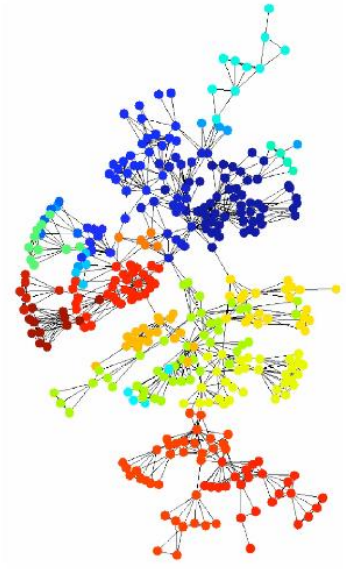
Color es posición
Intensidad de arcos (pases)
Tamaño nodo: duro con la pelota

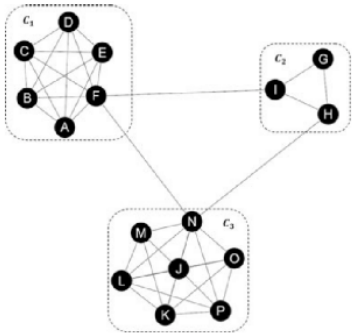
Métricas: Comunidades

- Mutualidad
  - Cada miembro conoce a todos los miembros
- Frecuencia
  - Cada miembro conoce al menos k miembros del grupo
- Cercanía
  - Los miembros están separados por máximo de n saltos

Comunidades - Clusters - Módulos







- En esta red, hay **tres comunidades**: C_1, C_2 y C_3
- Cada comunidad está formada por un grafo completo (un **clique**) de tamaño variable ($C_1 = K_6$, $C_2 = K_3$ y $C_3 = K_7$)
- La densidad de enlaces entre las comunidades es muy baja. Los pocos enlaces que existen son **puentes**

Minería de Grafos

Objetivo: Desarrollar algoritmos para extraer y analizar grafos.

- Búsqueda de patrones en ellos
- Búsqueda de grupos de grafos similares (clustering)
- Construcción de modelos de predicción para las grafos (clasificación)

- Aplicaciones
  - descubrimiento motivo estructural
  - reconocimiento de proteínas
  - ingeniería inversa en VLSI
  - Mucho más …

Por qué Minería de Grafos?

- **Los grafos son ubicuos**
  - Compuestos químicos (quimio-informática)
  - Estructuras de las proteínas, las vías/redes biológicas (Bioinformactica)
  - Flujo de programas, flujo de tráfico, flujo de trabajo
  - bases de datos XML, Web, de redes sociales
- **Grafos es un modelo general**
  - Árboles, secuencias, lazos, etc.
- **Diversidad de grafos**
  - Dirigidos vs. no dirigidos, etiquetados vs. no etiquetados (arcos y vértices), ponderados, con ángulos y geometrías (topológicos en 2-D/3-D)
- **La complejidad de los algoritmos: muchos problemas son de alta complejidad**

Mineria en otras clases de Datos

- Minería de Datos Espaciales
- Minería espacio-temporal y Object ivos en movimiento
- Minería Cyber-físico de datos del sistema: salud, control de tráfico aéreo, simulación de inundaciones
- Minería de datos multimedia
- Minería de datos de texto
- Minería de datos Web
- Minería de datos Streams

Explosión de Datos

Air Bus A380

-1 billon de código
- cada motor genera 10 TB c/30 min

640TB por vuelo

Twitter generaba aproxim. 12 TB de datos/día

New York Stock intercambiaba 1TB de datos/día

Capacidad de almacenamiento se ha duplicado aproximadamente cada tres años desde la década de 1980

Big Data

"Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son **demasiado grandes y difíciles de procesar** con las bases de datos y el software tradicionales" (ONU, 2012)

Big Data

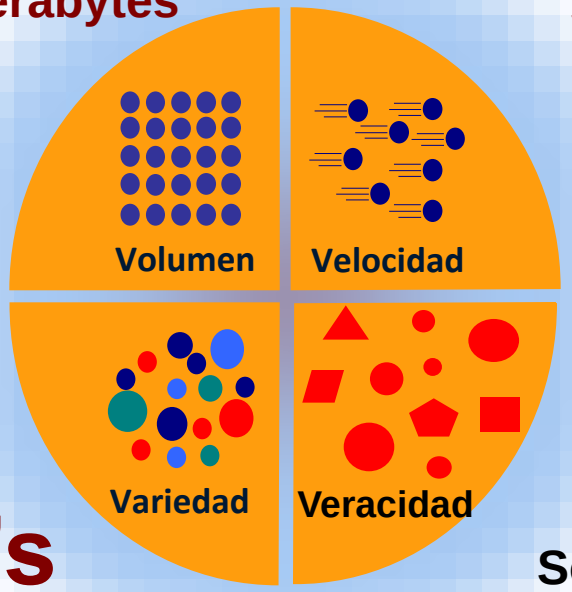
Los **grandes datos** permiten una mayor **inteligencia de negocios** mediante el **almacenamiento, el procesamiento y el análisis de datos** que se **ha ignorado** con anterioridad debido a las **limitaciones de las tecnologías tradicionales de gestión de datos**

Source: *Harness the Power of Big Data: The IBM Big Data Platform*

Big Data: Nueva Era de la Analítica

12+ terabytes
de Tweets
por día

5+ million
eventos comerciales por
segundo.

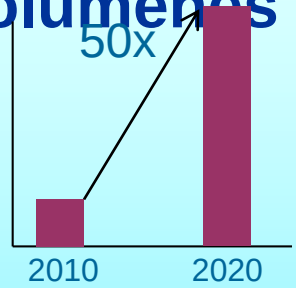


100's
de diferente
tipos de datos.

Solo **1 de 3**
tomadores de decisiones
confían en su información.

4 Características de Big Data

Eficiente procesamiento cada vez mayor de grandes **Volumenes**



35 ZB

En respuesta a la creciente **Velocidad**

30 Billones sensores RFID , etc.

Analizar la amplia **Variedad**

80% e los datos del mundo es no estructurado

Establecer la **Veracidad** de las fuentes de datos grandes

1 de 3 líderes de negocios no confían en la información que utilizan para tomar decisiones

Los 5 clásicos Casos de uso en Big Data

Exploración
Encontrar, visualizar, comprender todos los grandes volúmenes de datos para mejorar la toma de decisiones

Tener una vista del cliente de 360o
Extender puntos de vista de los clientes existentes, mediante la incorporación de fuentes de información internas y externas adicionales

Extensión Inteligente de la Seguridad
minimar riesgo, detectar fraudes y supervisar la seguridad informática en tiempo real

Análisis de operaciones
Analizar una variedad de datos para mejorar resultados comerciales

Aumentar capacidades de Procesamiento de Datos
Integrar capacidades de big data y data warehouse para aumentar la eficiencia operativa

Ejemplos de uso de Big Data

IBM y el concurso de IBM Watson Deep Blue

- Un computador debería responder a las preguntas de conocimiento como las que los seres humanos responden normalmente.
- El secreto fue añadir un nivel para procesar enormes cantidades de información escrita en texto en lenguaje natural, principalmente la información de Wikipedia, y en poco tiempo a hacer las combinaciones necesarias para formular las respuestas.

Wal-Mart registra miles de transacciones de sus clientes

- Las analiza para entender la popularidad de los productos, y asociar con sus preferencias y hábitos.

Empresas financieras como American Express

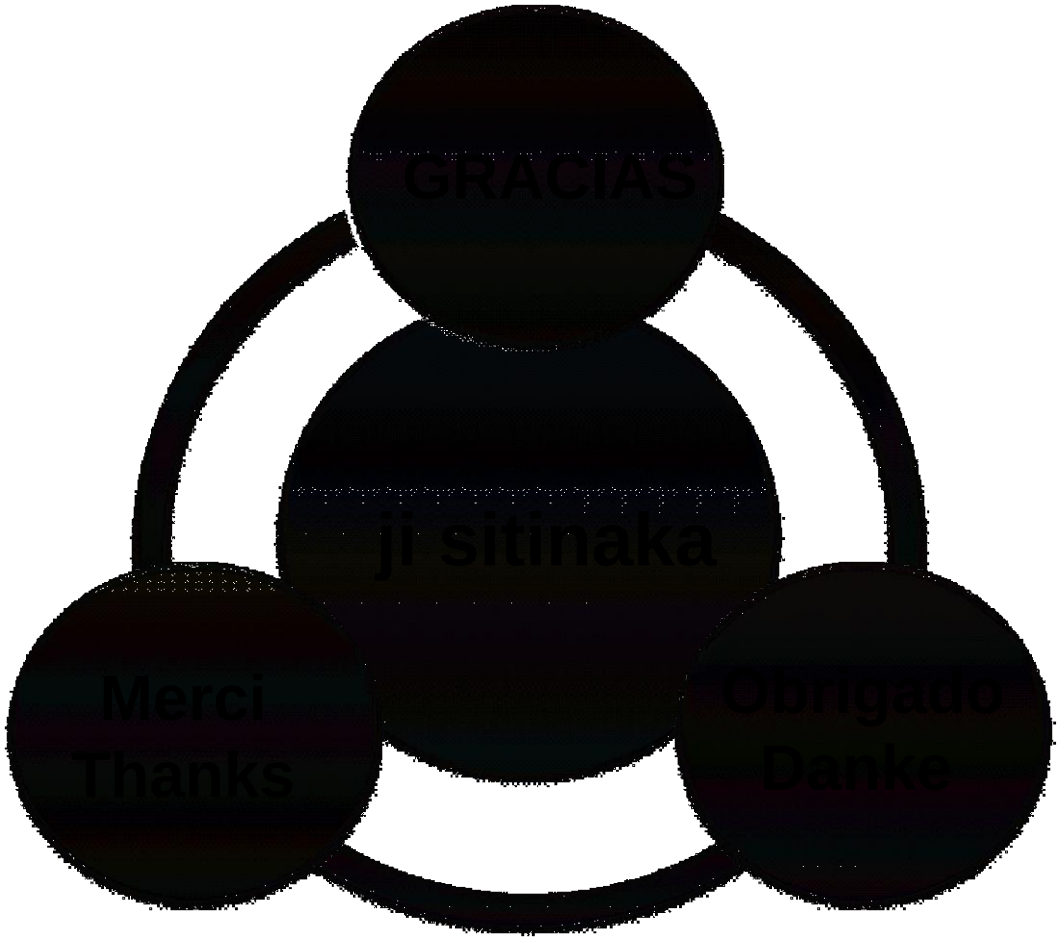
- Están ejecutando extensos análisis y técnicas de inteligencia de negocio en las transacciones realizadas por sus clientes para tratar de entender sus intenciones.

Por ejemplo, para identificar potenciales clientes insatisfechos en los diferentes servicios financieros y predecir los clientes que desaparecerán en el futuro.

Conclusiones

Fuente: http://www.mundodigital.net/la-analitica-de-datos-o-big-data-se-convierte-en-una-herramienta-tecnologica-de-peso/

Conclusiones



"Si buscas resultados distintos,
entonces no hagas siempre lo mismo"

A. Einstein

www.ing.ula.ve/~aguilar

aguilar@ula.ve

www.ing.ula.ve/~aguilar/actividad-docente/cursos/AdDEnero2017.pdf