UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

XLIII CLEI
Conferencia Latinoamericana de Informática

Técnicas de Minería en la Analítica de Datos Sociales

Jose Aguilar
CEMISID
Septiembre 2017

en la senda de Sadosky

46 JAIIO
Jornadas Argentinas de Informática

OBJETIVO

Introducir a los participantes en la *Analítica de Datos Sociales*. En especifico, los conceptos de base, metodologías para desarrollar tareas de analítica de datos social, herramientas, conceptos vecinos, etc.

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Contenido

Tema 1. Introducción a la Analítica de Datos Sociales
Tema 2. Metodología para hacer Analítica de Datos
Tema 3. Ciencia de los Datos para la Analítica de Datos Sociales
Tema 4: Técnicas de Analítica de Datos Sociales: Minería semántica, Minería de texto
Tema 5: Técnicas de Analítica de Datos Sociales: Minería de Grafos, enlazado de datos

Introducción a la Analítica de Datos Sociales

Analítica Social de los Datos

El análisis sociales de datos es un estilo de análisis en el que es considerado las cosas y personas en su 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 en **sitios externos**, como redes sociales (o a través de aplicaciones sociales),

- **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, para predecir algunos eventos con ellos, entre otras cosas

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 **el efecto**.

- **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

La analítica social de datos implica el análisis de Internet, con el fin de comprender la percepción y actividad social presente en los datos.

Por ejemplo, analizar:

- Interacciones con los demás (mensajería, redes sociales, etc.),
- Uso de plataformas (búsquedas, etiquetados).

El análisis de Internet nos dice acerca de cómo las personas y cosas interactúan con el mundo y con los suyos

Desafíos técnicos

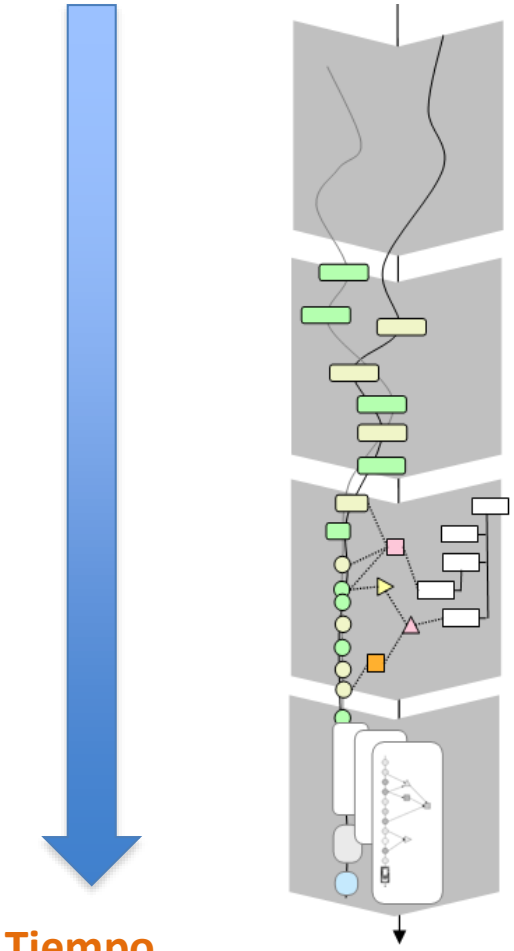
- **Extracción** de ese conocimiento
- **Comprensión** del texto ruidoso y no estructurado
- **Reconocimiento** de entidades, ubicación, relaciones

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 del individuo en el contexto colectivo:***
  - **Analítica de la Disposición:** formas de interacción
  - **Analítica del Discurso:** el lenguaje es una herramienta fundamental para la construcción del conocimiento.
  - **Analítica de Redes Sociales:** relaciones interpersonales en plataformas sociales.

Aprender con los datos del entorno



Datos crudos

Buscar semánticamente

Inferir símbolos

Buscar patrones de interés

Análisis

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?

Flujo de análisis

1. ¿Quién es relevante?	2. ¿Que hicieron?	3. ¿Cómo se sintieron al respecto?
• Todos • Expertos / Autoridades • Basado en el comportamiento • Intereses	• Comportamientos • Hora • Entidades que usaron	• Estados de ánimo y sentimientos asociados a estas acciones y entidades

Flujo de análisis

1. ¿Quién es relevante?	2. ¿Que hicieron?	3. ¿Cómo se sintieron al respecto?
• Todos • Expertos / Autoridades • Basado en el comportamiento • Intereses	• Comportamientos • Hora • Entidades que usaron	• Estados de ánimo y sentimientos asociados a estas acciones y entidades

¿Dónde van las personas sanas a comer?

- Expertos en salud
- Personas que hacen ejercicio regularmente
- Personas que les gustan / siguen temas relacionados con la salud

Flujo de análisis

1. ¿Quién es relevante?	2. ¿Que hicieron?	3. ¿Cómo se sintieron al respecto?
• Todos • Expertos / Autoridades • Basado en el comportamiento • Intereses	• Comportamientos • Hora • Entidades que usaron	• Estados de ánimo y sentimientos asociados a estas acciones y entidades

¿Dónde ir a catar vinos?

¿Qué lugares se mencionan junto con "cata de vinos"?

Flujo de análisis

1. ¿Quién es relevante?	2. ¿Que hicieron?	3. ¿Cómo se sintieron al respecto?
• Todos • Expertos / Autoridades • Basado en el comportamiento • Intereses	• Comportamientos Entidades • Hora	• Estados de ánimo y sentimientos asociados a estas acciones y entidades

¿A qué café ir para estudiar?

¿Qué palabras de ánimo y sentimiento se usaron para describir “Starkbuck", “Juan Valdez", …?

REDES SOCIALES VIRTUALES

La Red Social Virtual se **apoya en tecnologías** que permiten realizar la relación de forma virtual.

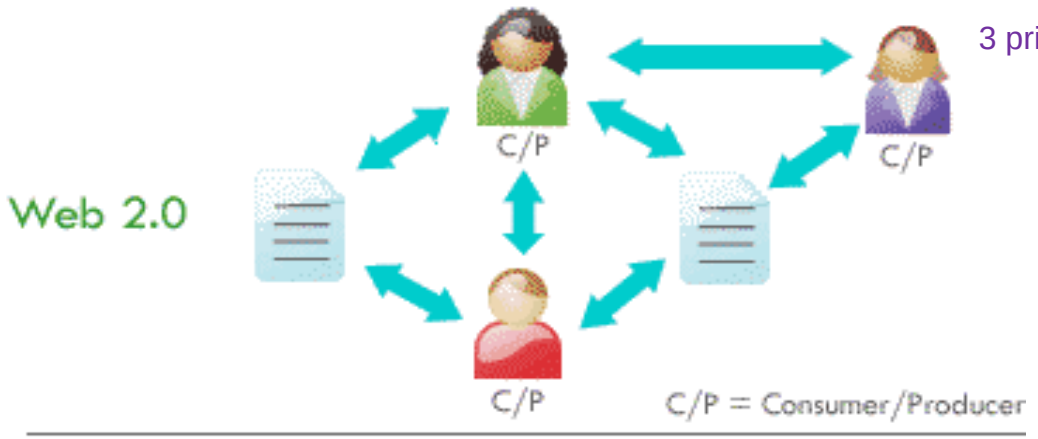
Este tipo de actividad se apoya fundamentalmente en **Internet**, sus herramientas y su entorno “globalizado” ha permitido romper fronteras
y tiene sus propias reglas.

Existe una verdadera **explosión en los últimos 5 años.**

la diferencia es que de alguna manera **se obvia el parámetro de tiempo y distancia** en el concepto de vecindad), pasando a un segundo plano,

LA WEB 2.0 Y LA WEB 3.0

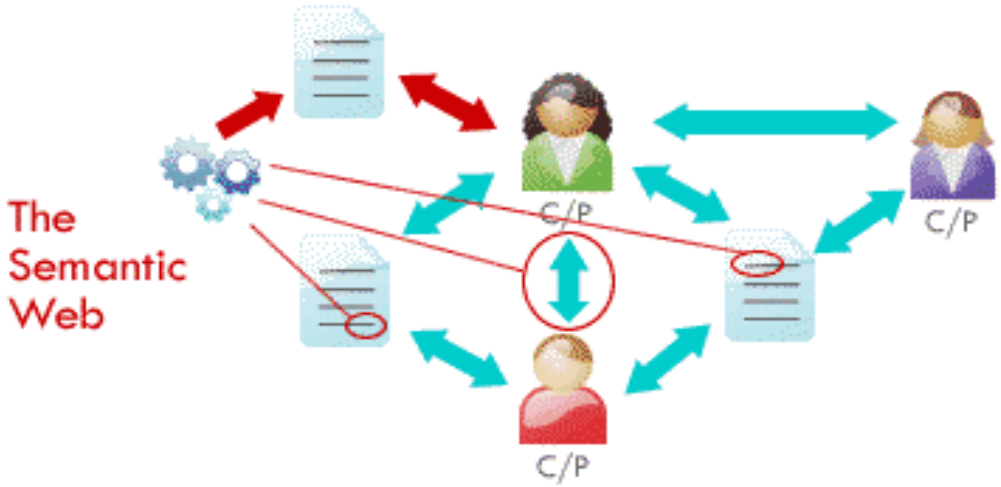
- Páginas estáticas
- El uso de frameset o Marcos.
- Extensiones propias del HTML.



3 principios básicos

1. La web como plataforma
2. Aprovechar la Inteligencia Colectiva
3. Experiencias enriquecedoras del usuario

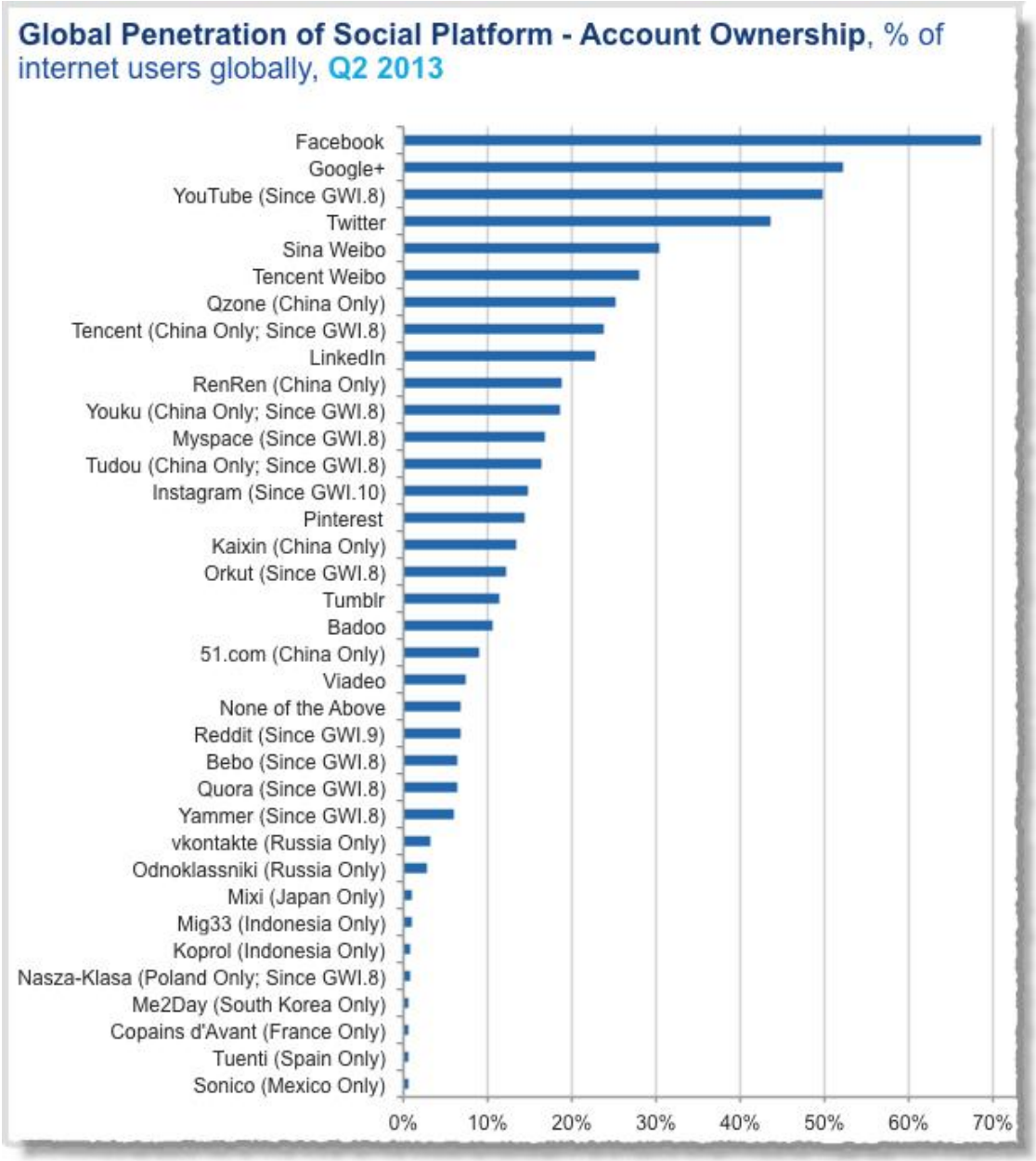
Una nueva manera de involucrar al usuario en la web. Deja en el pasado las páginas web estáticas y propone más interacción con el usuario, donde se le permite crear contenido en las páginas para así conformar una comunidad virtual



Web 3.0 se basa en:

- ✓ una Internet más "inteligente",
- ✓ los usuarios hacen búsquedas cercanas al lenguaje natural,
- ✓ la información tiene semántica asociada y
- ✓ la Web relaciona conceptos de múltiples fuentes,
- ✓ La web deduce información a través de reglas asociadas al significado del contenido.

Analisis de redes sociales y Big Data



Analítica Social de los Datos

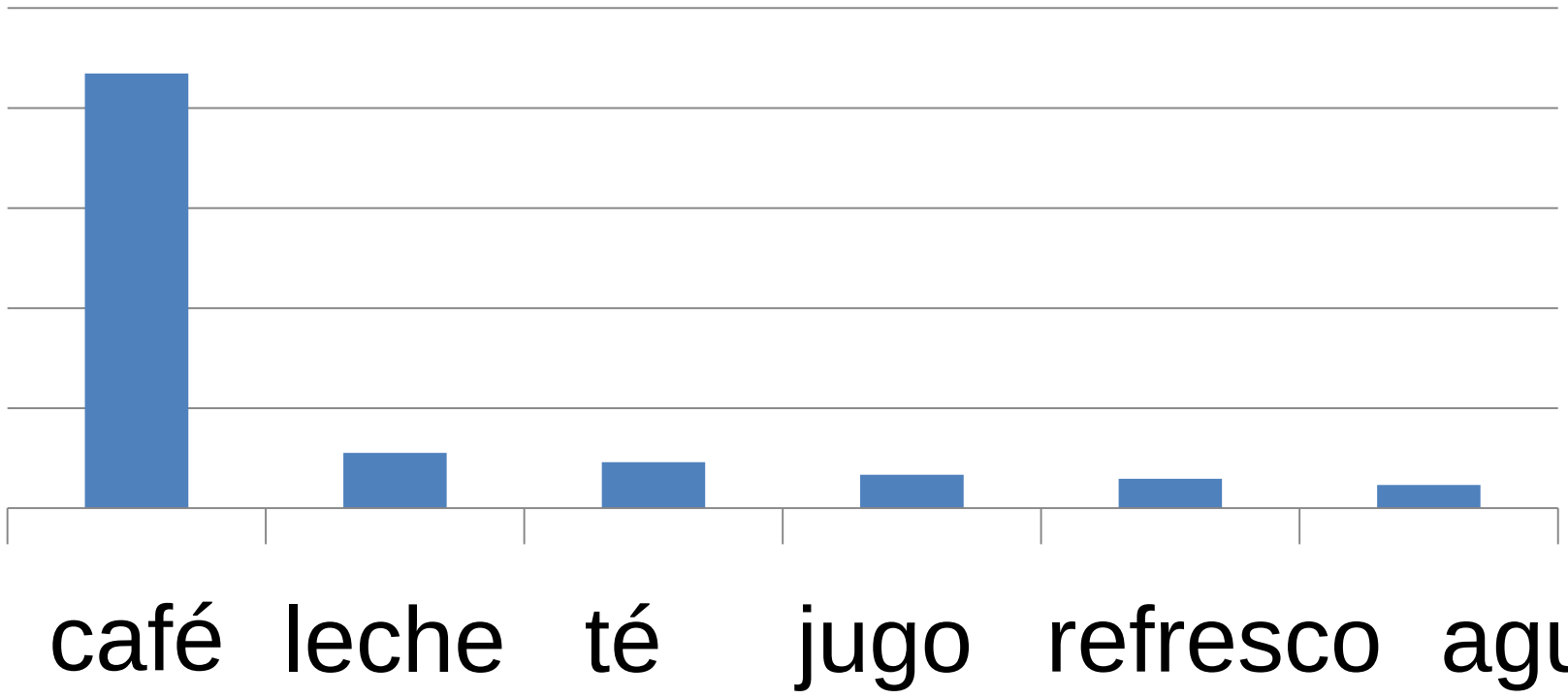
Por lo general, podemos recuperar los datos sociales desde Internet.

Por ejemplo, desde una variedad de redes sociales como Twitter, Facebook, Wikipedia, etc.

- La mayoría de las redes sociales **proporcionan un API**, no es difícil recuperar sus datos.
- El uso de API para obtener los datos es como enviar una solicitud a un sitio web y el regresa los datos solicitados en **formato XML, JSON**.
- La **indexación de los datos** es la tarea más difícil que acceder a APIs.
- Existen **firehose de contenido** para todo lo puesto en una red, por ejemplo Twitter, Amazon, etc.
- **Firehose:** servicio para entregar datos de transmisión en tiempo real (**real-time streaming data**)

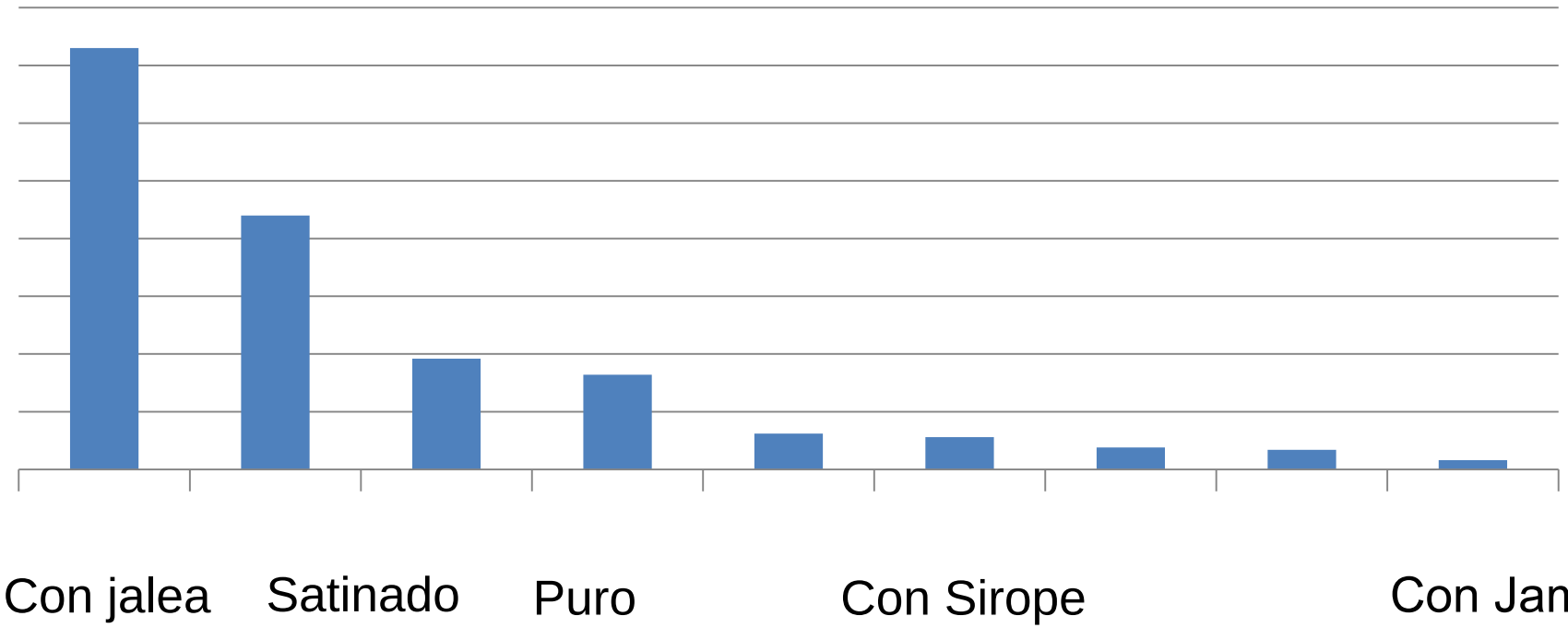

SOCIAL MEDIA EXPLAINED
TWITTER
I'M EATING A #DONUT
FACEBOOK
I LIKE DONUTS
FOURSQURE
THIS IS WHERE I EAT DONUTS
INSTAGRAM
HERE'S A VINTAGE PHOTO OF MY DONUT
YOUTUBE
HERE I AM EATING A DONUT
LINKEDIN
MY SKILLS INCLUDE DONUT EATING
PINTEREST
HERE'S A DONUT RECIPE
LAST FM
NOW LISTENING TO "DONUTS"
G+
I'M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

¿Qué beben las personas con donas?

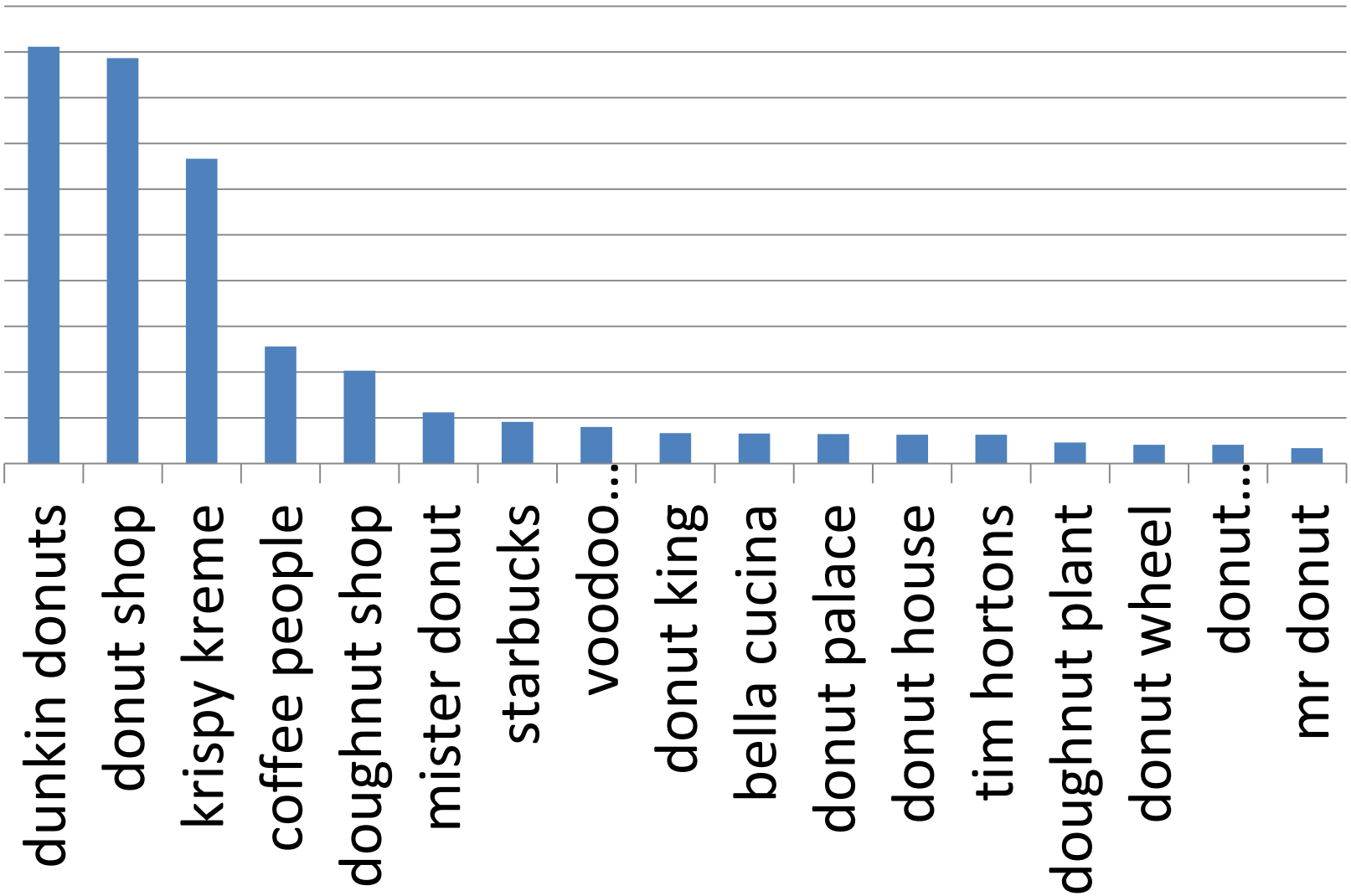


180k tweets de 7 días que contienen la palabra "donas"

¿Qué tipo de donas comen la gente?



¿Dónde la gente compra donas?

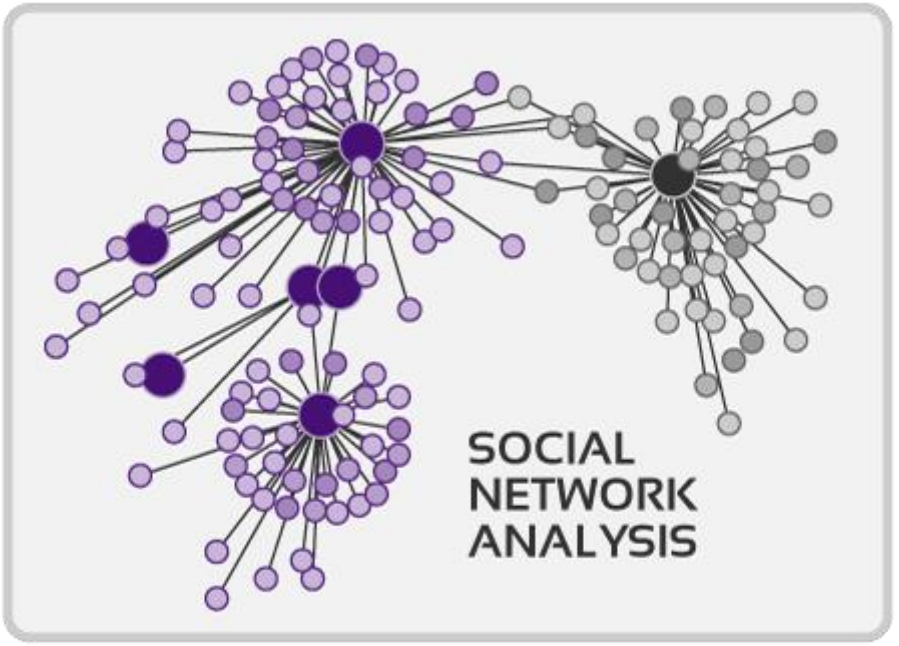


SNA

Análisis de Redes Sociales

Análisis de Redes Sociales

El Análisis de Redes Sociales (SNA por sus siglas en inglés de **Social Network Analysis**) es una rama de la Sociología que **se vale de métricas** para determinar la **estructura de grupos sociales**; por ejemplo, descubrir quiénes son los actores más importantes – líderes (formales o informales) , comunicadores – en un grupo.

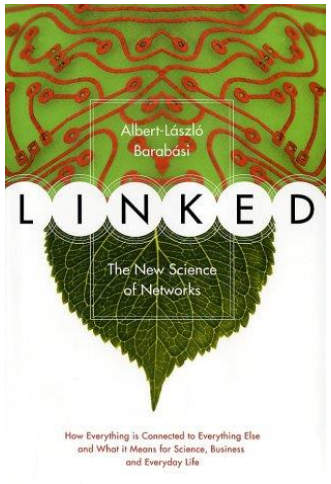


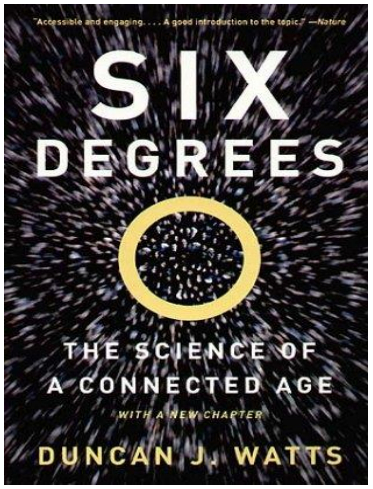
Red social

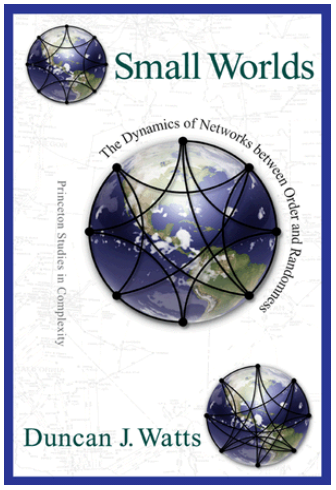
- **Una red social** es una estructura social de personas, relacionadas (directa o indirectamente) entre sí a través de una relación o interés común

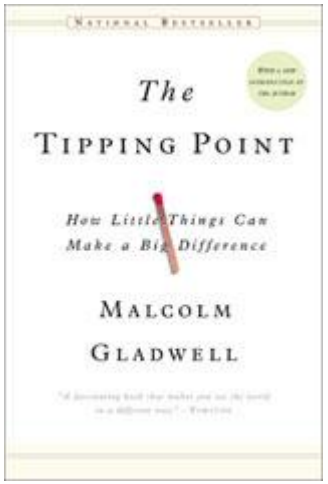
- El **análisis de redes sociales (SNA)** es el estudio de redes sociales para entender su **estructura y comportamiento**

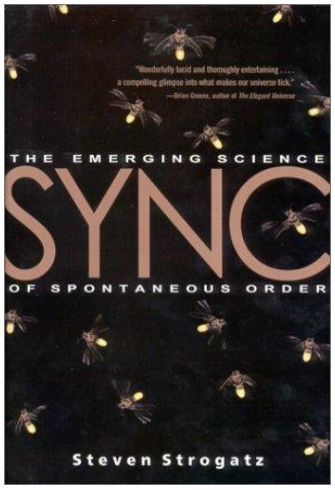
Es el proceso de investigación de estructuras sociales a través del uso de teorías de redes y de grafos







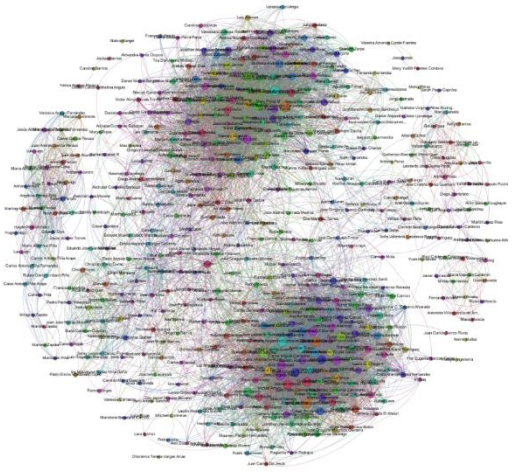




Las redes sociales han capturado el interés del público en los últimos años, como es evidente en el número de trabajos en el área

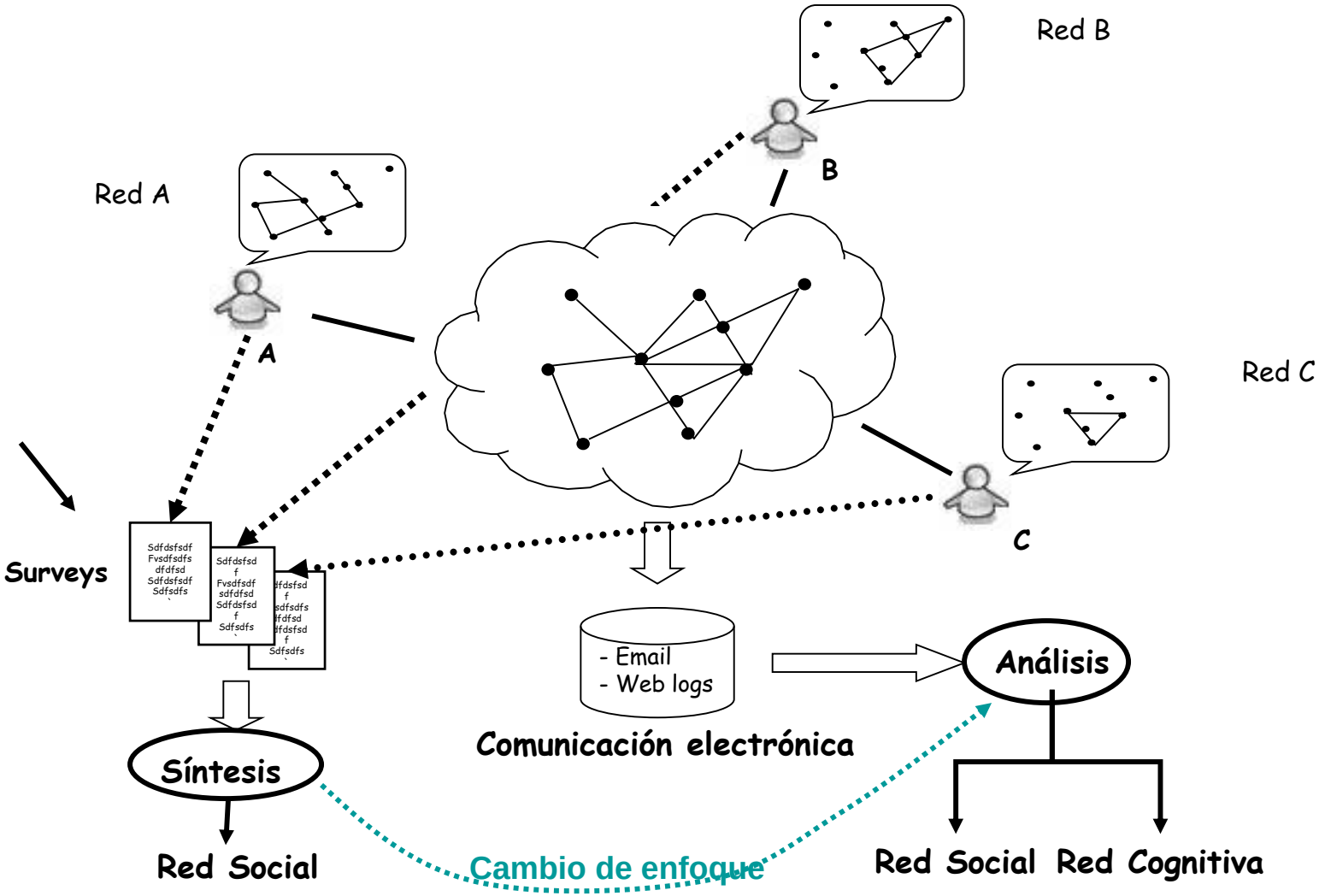
Red Social

- Las redes sociales son las diferentes interacciones que **realizamos con nuestros conocidos**.
- Estas pueden ser representadas mediante **grafos**, donde los **nodos** son las personas que interactúan y los **arcos** que los unen son las interacciones entre esas personas.



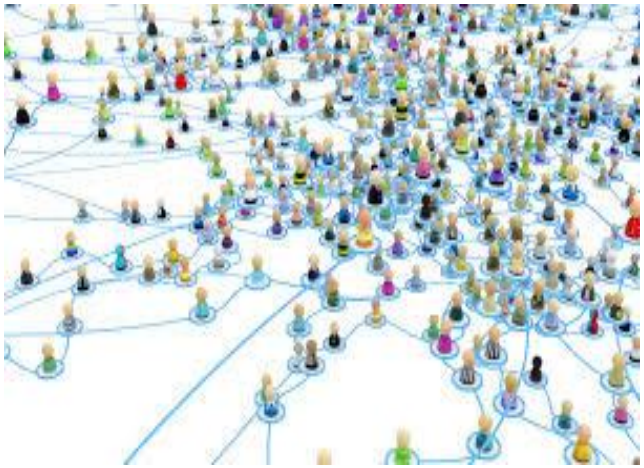
Algunos autores consideran el origen de las redes sociales virtuales desde el envío del primer e-mail en los años 70, hasta llegar al 2017 con Instagram, LinkedIn, Facebook y muchas más

Un cambio en el enfoque: "de la" síntesis "al" análisis



¿Qué utilidad tiene su estudio?

Una **red social** es una forma de representar una estructura social, **conectándolos si dos elementos del conjunto de actores** (tales como individuos u organizaciones) **están relacionados de acuerdo a algún criterio** (relación profesional, amistad, parentesco, etc.).

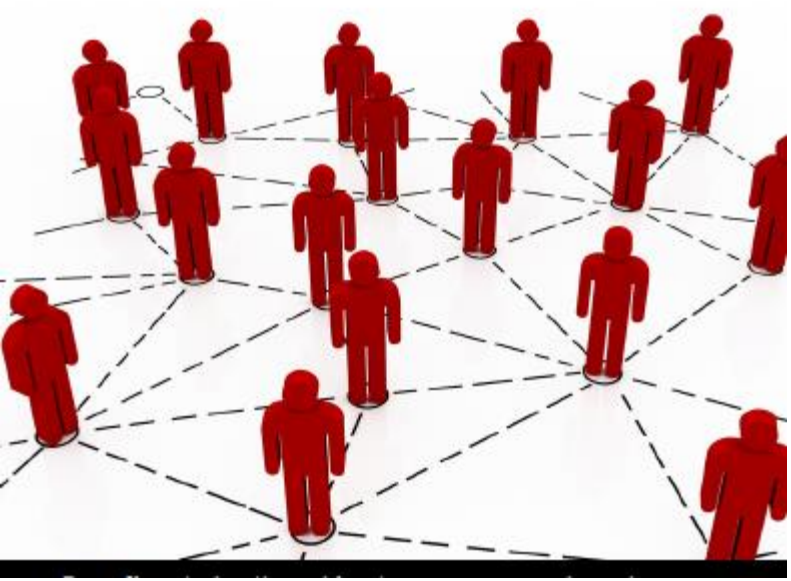


El tipo de conexión representable en una red social es una **relación interpersonal**, que se pueden **interpretar como relaciones** de amistad, parentesco, laborales, entre otros.

¿Qué utilidad tiene su estudio?

El **análisis de redes sociales** es un enfoque analítico, con sus principios teóricos, métodos de software.

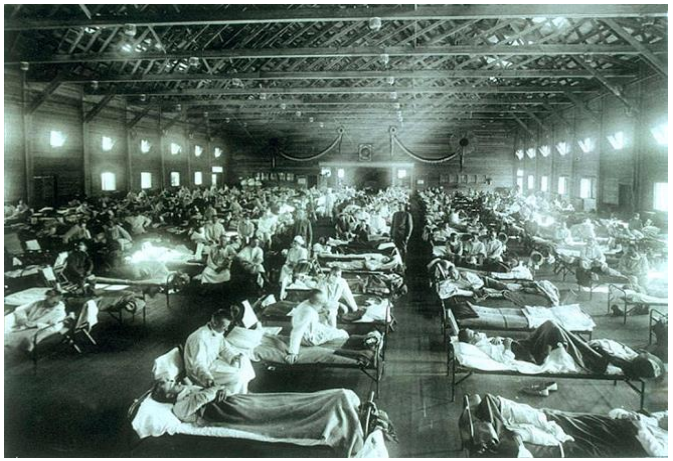
Los analistas **estudian la influencia del todo en las partes** y **viceversa**, **el efecto producido** por la acción selectiva de los individuos en la red; desde la **estructura hasta la relación y el individuo**, desde **el comportamiento hasta la actitud**.

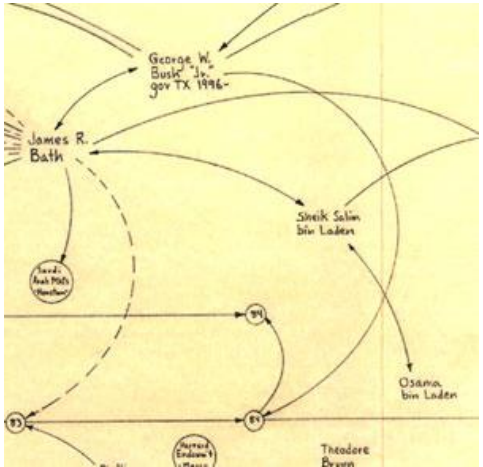


¿Qué utilidad tiene su estudio?

El análisis de redes sociales proporciona **información sobre la interacción** entre las **normas de comunicación, propagación de rumores y la estructura social**.

El análisis de redes sociales se ha utilizado **en epidemiología** para ayudar a **entender cómo los patrones de contacto humano favorecen o impiden** la propagación de enfermedades como el VIH en una población.





El análisis de redes sociales también puede ser una **herramienta eficaz para la vigilancia social masiva** - por ejemplo, Mark Lombard rastreo y mapeo fiascos financieros globales en los años 1980-90 a partir de fuentes públicas, como los artículos de noticias.

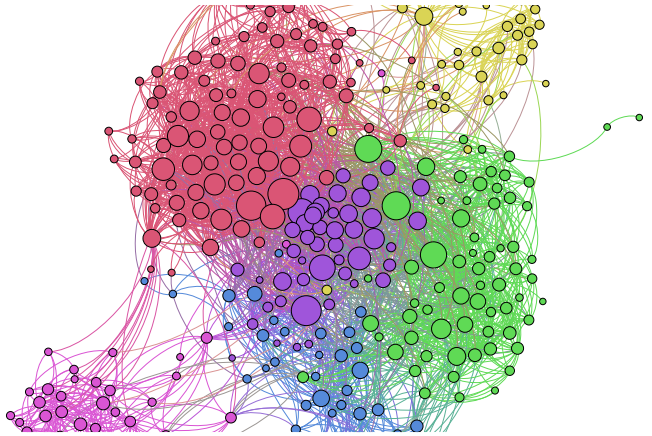
¿Qué utilidad tiene su estudio?

La teoría de **difusión de innovaciones** explora las redes sociales y su rol en la influencia de **la difusión de nuevas ideas y prácticas**.

Difusión de innovaciones: busca explicar el proceso que se desarrolla durante 'las innovaciones tecnológicas',

Facebook de una persona

Los colores separan componentes fuertemente conectados de la red.



¿Qué utilidad tiene su estudio?

Facebook acorta la teoría de los “seis grados de separación” de las personas

un estudio realizado por Facebook en conjunto con la Universidad de Milán, afirma que el número de grados de separación entre dos personas es menor.

El estudio estima que **el 99.6** por ciento de **los pares de usuarios** están **conectados con grados de 5 personas**, mientras que **el 92** por ciento, por **grados de cuatro personas**.

Ese estudio demuestra que con la llegada de Facebook y las redes sociales, **la distancia entre las personas se está achicando cada vez más**. La estadística demuestran que Facebook está llevando los seis grados a los **cuatro grados de separación**.

¿Qué utilidad tiene su estudio?

Un estudio ha descubierto que **la felicidad** tiende a correlacionarse **en redes sociales**.

Cuando **una persona es feliz**, los **amigos cercanos** tienen una probabilidad un **25 por ciento mayor de ser también felices.**

Además, **las personas en el centro de una red social** tienden a **ser más feliz** en el futuro **que aquellos situados en la periferia**.

En las redes estudiadas se observaron tanto a grupos de personas felices como a grupos de personas infelices, con **un alcance de tres grados de separación:**

se asoció **felicidad de una persona con el nivel de felicidad de los amigos de los amigos de sus amigos.**

Tipo de SNA

Análisis sociocéntrico (completo) de la red

- Emerge desde la **sociología**
- Implica la **cuantificación de la interacción entre un grupo** socialmente bien definido de personas
- Centrarse en la identificación de **patrones estructurales globales**
- Es ideal para analizar organizaciones desde **enfoques sociométrico**

Análisis egocéntrico (personal) de redes

- Emerge desde la **antropología y psicología**
- Implica la **cuantificación de las interacciones entre un individuo** (llamado ego) **y todas las demás personas relacionadas** (directa o indirectamente) con él
- Define las **redes personales**

Red social

Tipos de Redes

- **Redes sociales**

  "Quién sabe quién"

- **Redes Socio-Cognitivas**

  "Quién piensa quién sabe quién"

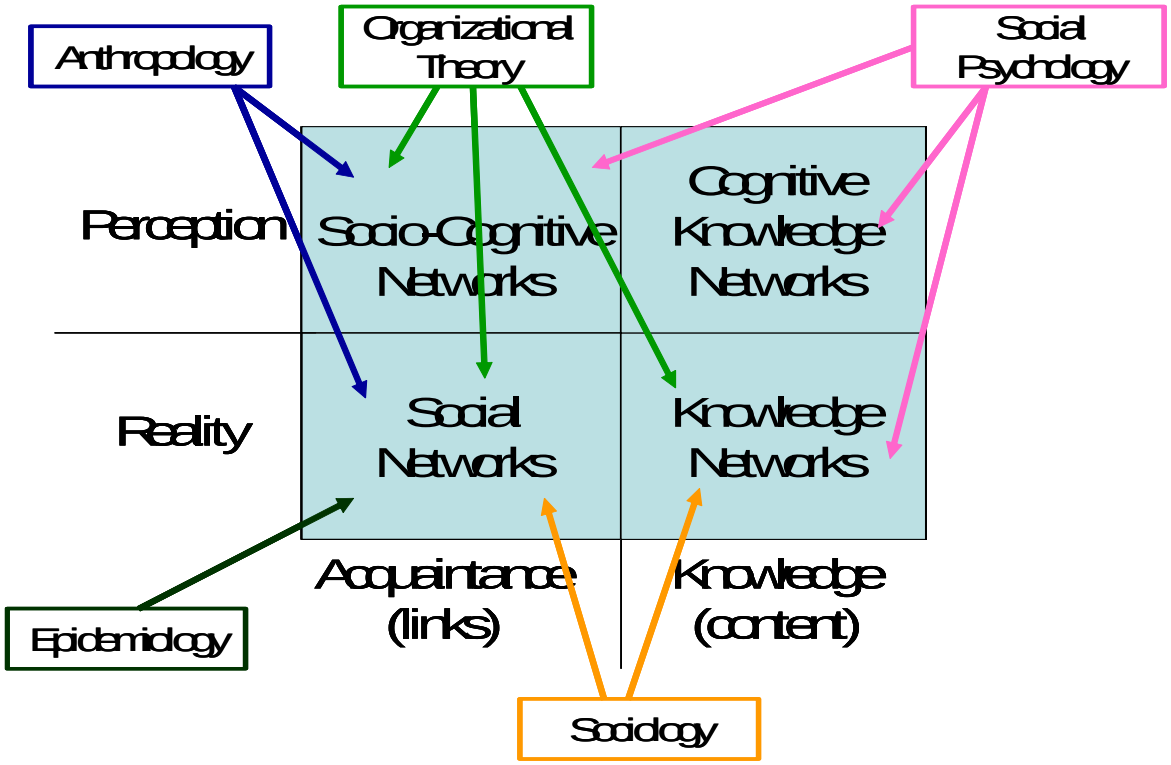
- **Redes de conocimiento**

  "Quién sabe qué"

- **Redes de Conocimiento Cognitivo**

  "Quién piensa quién sabe qué"

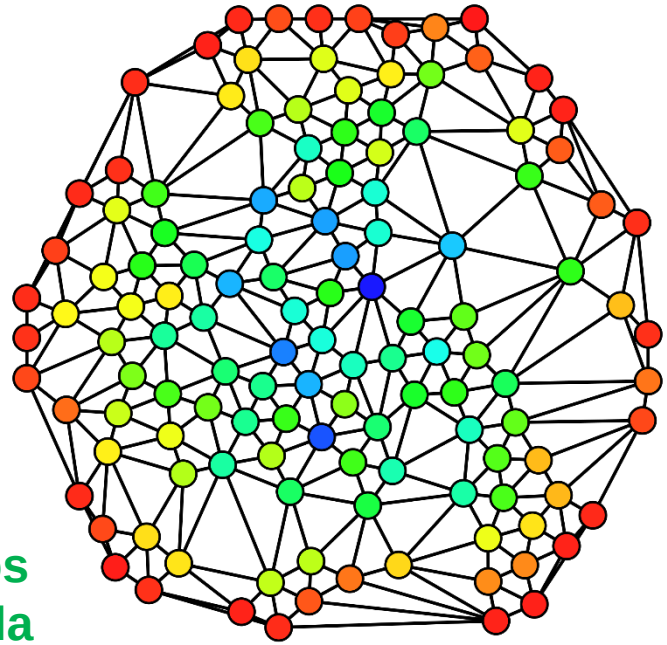
Investigación en Redes desde las Ciencias Sociales



Las redes de ciencias sociales tienen una amplia aplicación en diversos campos

Analisis de redes sociales (SNA)

Las técnicas de SNA permiten analizar la interacción social entre los participantes en una actividad.

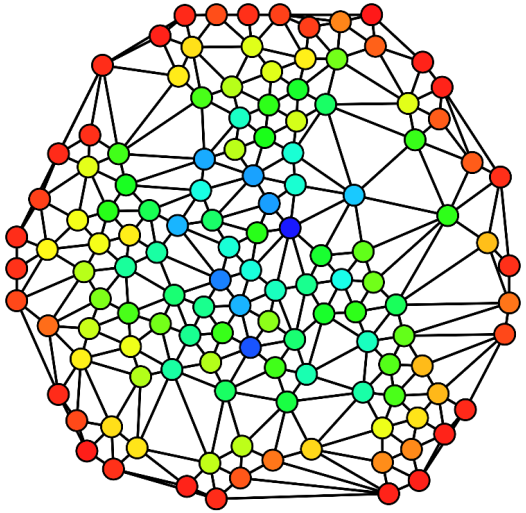


Para ello **se rastrean las interacciones entre los miembros del grupo,** se miden **la densidad de la red** (**número de interacciones entre los distintos participantes** respecto del número total de posibles conexiones entre ellos), y se calcula **el grado de centralidad de los participantes**, que da una idea sobre hasta qué punto un participante interactúa con el resto del grupo.

➡ **Indicadores**

Analisis de redes sociales (SNA).

Indicadores:



La tonalidad (de rojo = 0 a azul = máximo) de cada nodo indica centralidad de intermediación.

El **potencial de redes sociales (SNP) es un coeficiente numérico que representa la capacidad de influir en una red social de un individuo (su tamaño en ella).**

- Es sinónimo de usuario alfa.
- Permite la clasificación de los individuos.
- Las variables para calcular el SNP de un individuo incluyen participación en actividades de redes sociales, pertenencia a grupos, roles de liderazgo, reconocimiento, y su frecuencia.

Medidas de red:

- Densidad
- Centralización
- Componentes conectados
- Componente gigante
- Ruta mas corta
- Densidad grafo

Medidas de actores:

- Centralidad / Prestigio
- Grado
- Proximidad
- Intermediación

Agrupamientos:

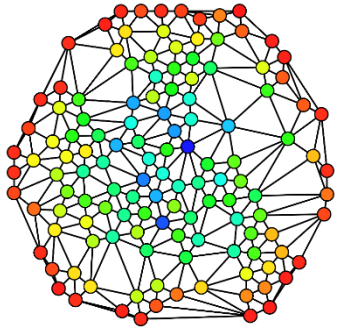
- Cliqués, componentes …

Analisis de redes sociales (SNA).

Indicadores: basados en métricas

Conexiones

- **Homofilia:** La medida en que los actores forman lazos con otros similares.
- **Multiplexidad:** El número de conexiones entre arcos.
- **Mutualidad/Reciprocidad**
- **Cierre de la red**
- **Propinquity**: tener más vínculos con otros geográficamente cercanos



Distribuciones

- **Puente:** Un único vínculo entre dos individuos o racimos.
- **Centralidad:** grupo de métricas que pretenden cuantificar la "importancia" o "influencia" (en una variedad de sentidos) de un nodo (o grupo) particular dentro de una red.
- **Densidad:** La proporción de vínculos directos en una red con respecto a número total posible.
- **Distancia:** El número mínimo de lazos necesarios para conectar a dos actores concretos.
- **Agujeros estructurales:** La ausencia de lazos entre dos partes de una red.
- Fuerza de lazo: combinación lineal de tiempo, intensidad emocional, intimidad y reciprocidad.

Segmentación

- **Coeficiente de agrupación:** probabilidad de que dos conectados a un nodo estén también conectados.
- **Cohesión:** número mínimo de miembros que, si se retiran del grupo, lo desconectaría

Cada indicador de red da respuesta a preguntas diferentes

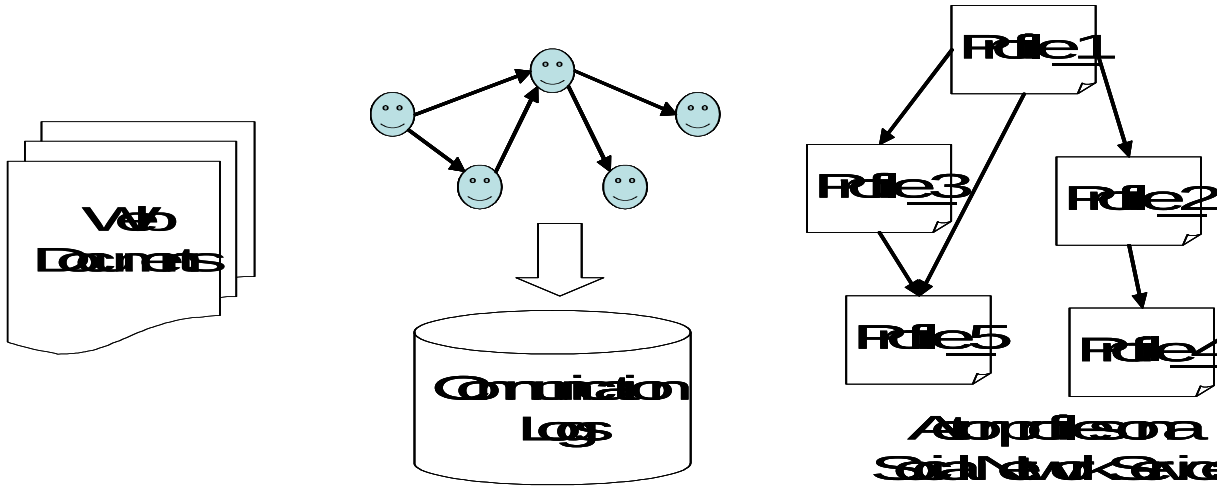
- **¿Quién es más central?**
  - **1) METRICA DE RED: centralidad**
    - a)Centralidad de grado (degree centrality).
      - 1) Indegree o grado de entrada
      - 2) Outdegree o grado de salida
    - b) Centralidad de cercanía (closeness centrality).
    - c) Centralidad de intermediación (Betweenness centrality).
- **¿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)**
  - **4) METRICA DE RED: modelo de corbata de moño de la web**
- **¿A qué distancia están las cosas?**
  - **5) METRICA DE RED: rutas más cortas**
- **¿Cómo densa son las cosas?**
  - **6) METRICA DE RED: densidad grafo**

Tareas de SNA

- Extracción/construcción de redes sociales
- Predicción de enlace
- Aproximación de grandes redes sociales
- Identificación de actores prominentes/ confiables/expertos en las redes sociales
- Búsqueda en redes sociales
- Descubrir comunidades en redes sociales
- Descubrimiento de conocimiento de redes sociales

Extracción en redes sociales

- Minería de una red social
- Tipos de fuentes de datos en la web
  - **Contenido disponible en páginas web** (por ejemplo, páginas de inicio de usuarios, hilos de mensajes, etc.)
  - **Registros del usuario** (por ejemplo, registros de chat de correo electrónico y mensajería)
  - **Información de interacción** social proporcionada por los usuarios (por ejemplo, sitios web de servicios de redes sociales como Friendster y MySpace)



Predicción de enlace

Diferentes versiones

- Dada una red social en el tiempo t_i **predecir el vínculo social entre los actores** en el tiempo $t_i + 1$
- Dada una red social con un conjunto incompleto de vínculos sociales entre un conjunto completo de actores, **predecir los vínculos sociales no observados**
- **Dada información sobre los actores, predecir el vínculo social entre ellos**

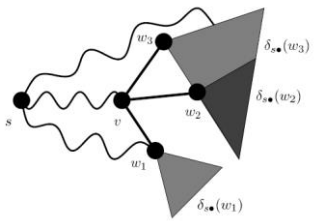
Algunos programas:

- Latent Space model (Hoff et al, 2002),
- Dynamic Latent Space model (Sarkar and Moore, 2005),
- p* model (Wasserman and Pattison, 1996)

Identificación de actores prominentes/ confiables/expertos en las redes sociales

Un enfoque común es computar puntuaciones/rankings sobre el conjunto (o un subconjunto) de actores en la red social que indican el grado de importancia/experiencia/influencia

- Existen **medidas de centralidad** para medir la importancia de los actores en una red social
- P.ej. Pagerank, HITS
- Shetty y Adibi (2005) Proporciona una técnica **basada en la teoría de la información** para descubrir nodos importantes en un gráfico.



Confiabilidad en las redes sociales

Propagación de confianza: inferir valores de confianza en una red

- Un usuario confía en algunos de sus amigos, sus amigos confían en sus amigos y así sucesivamente ...
- Dados valores de confianza y/o desconfianza entre un grupo de pares de usuarios, **¿se pueden predecir valores desconocidos de confianza/ desconfianza entre dos usuarios?**

- Golbeck et al (2003) discute la propagación de la confianza y su utilidad para la web semántica
- Algunas reglas:
  - Considerar los grupos de trabajo X e Y encabezados por dos lideres de manera que cada líder conozca a los miembros en su respectivo grupo
  - Utilizar la calificación del líder del grupo X que está en la lista de confianza del líder del grupo Y y propagar la calificación

  - Ejemplo: - www.ebay.com

Búsqueda en redes sociales

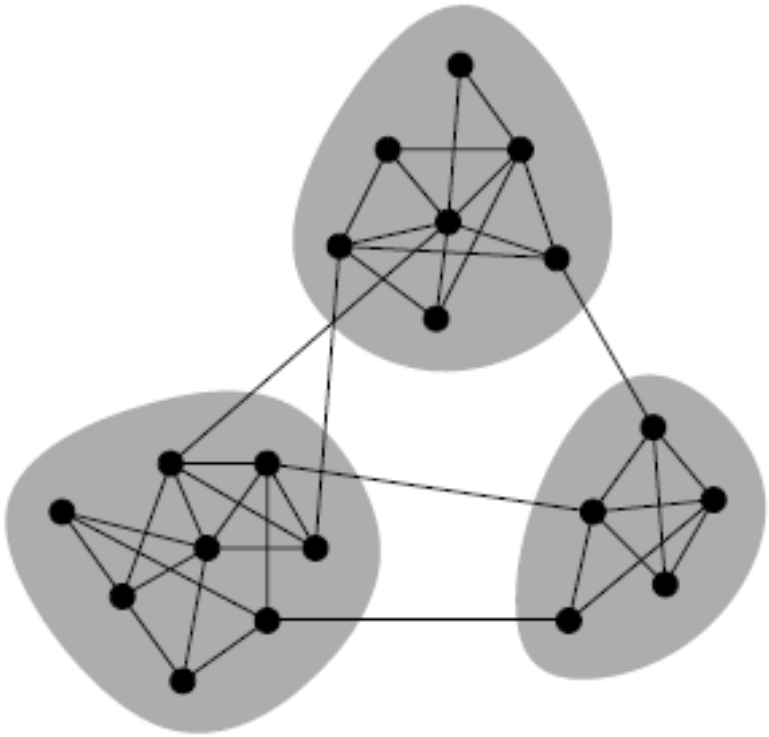
Enrutamiento de consultas en una red

- Un usuario puede enviar consultas a sus vecinos
- Si el vecino sabe la respuesta entonces él/ella contesta a sus vecinos, de lo contrario la consulta la propaga a través de una red

- Adamic et al (2001) desarrolla un esquema para el enrutamiento eficiente a través de una red.
  - En cada paso la consulta se pasa al vecino con el mayor número de vecinos
  - Una gran parte del grafo se examina en un pequeño número de saltos

- Kleinberg y Raghavan (2005) presentan un modelo de teoría de juegos para enrutar consultas en una red junto con incentivos para las personas que proporcionan respuestas a las preguntas

Descubrir comunidades en redes sociales



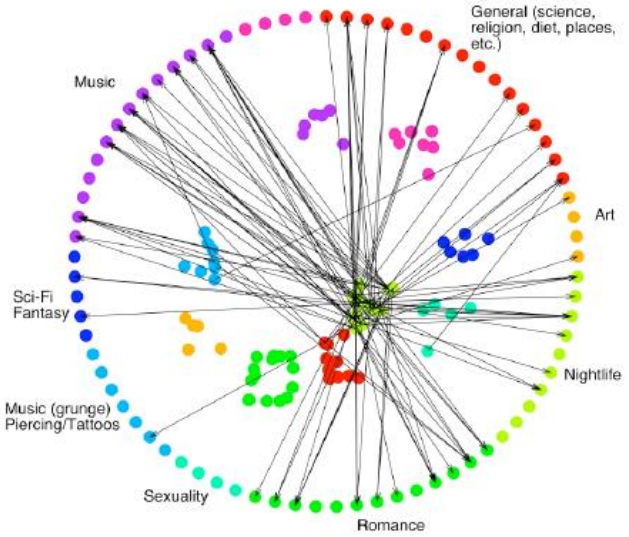
Basado en la teoría de grafos

- Comunidades
- Clústers
- Módulos
- Cliques
- K-core

Descubrimiento de conocimiento de redes sociales

Se pueden utilizar técnicas tradicionales de descubrimiento de conocimientos basadas en grafos

- Análisis espectral de matrices de adyacencia
- Análisis de enlaces
- Medidas teóricas
- Usando la visualización

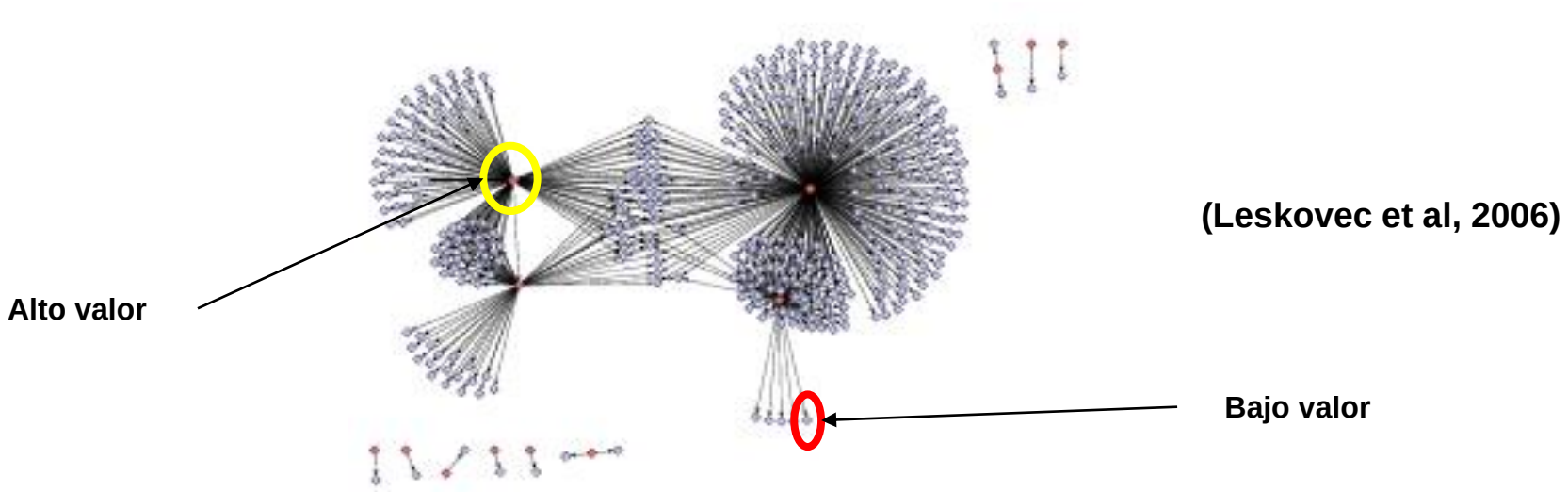


Grupos de actores con intereses compartidos (Paolillo y Wright, 2005)

Aplicación a la comercialización

Domingos and Richardson (2001, 2002)

- Valor de la red de un cliente es el beneficio esperado de la comercialización de un producto a un cliente, teniendo en cuenta la influencia **del cliente en las decisiones de compra de otros clientes**
- Aplicación de un **modelo probabilístico** a la red social de los clientes



(Leskovec et al, 2006)

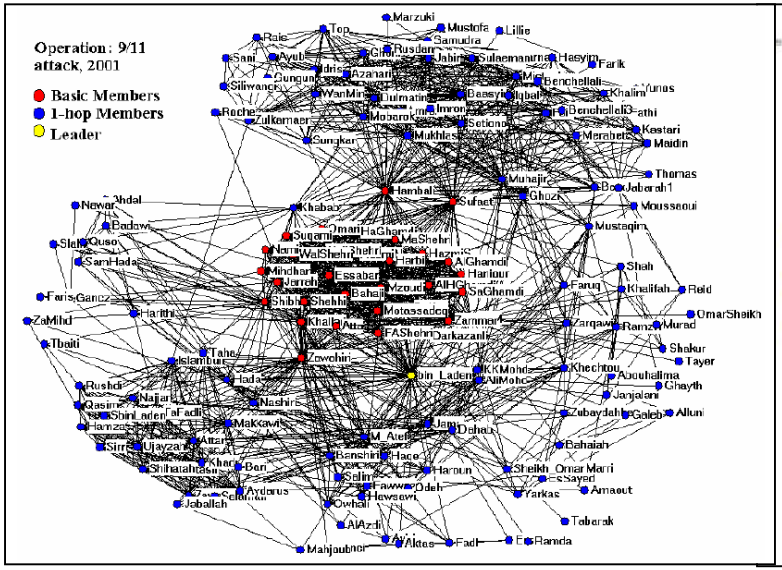
Aplicación al análisis de redes criminales

Qin et al, 2005

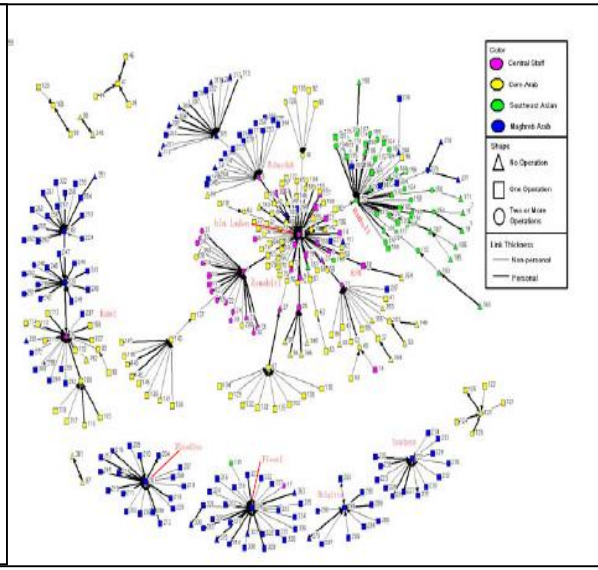
- Información recopilada sobre las **relaciones sociales entre miembros de la red Global Salafi Jihad** (GSJ) de múltiples fuentes (por ejemplo, informes de procedimientos judiciales, sus redes sociales)
- Aplicación de la **minería estructural de la Web en esta red**
- Grafo de derivación de la autoridad (**ADG**) captura la **autoridad (quien coordina) en la red criminal**

Ranking	Leader	Gatekeeper	Outlier
Central Member			
1	Zawahiri	bin Laden	Khalifah
2	Makkawi	Zawahiri	SbinLaden
3	Islambuli	Khadr	Ghayth
4	bin Laden	Sirri	M Atef
5	Attar	Zubaydah	Sheikh Omar
Core Arab			
1	Khallad	Harithi	Elbaneh
2	Shibh	Nashiri	Khadr4
3	Jarrah	Khallad	Janjalani
4	Atta	Johani	Dahab
5	Mihdhar	ZaMihd	Mehdi
Maghreb Arab			
1	Hambali	Baasyir	Siliwangi
2	Baasyir	Hambali	Fathi
3	Mukhlas	Gungun	Naharudin
4	Iqbal	Muhajir	Yunos2
5	Azahari	Setiono	Maidin
Southeast Asian			
1	Doha	Yarkas	Mujati
2	Benyaich2	Zaoui	Parlin
3	Fateh	Chaib	Mahdjoub
4	Chaib	DavidC	Zinedine
5	Benyaich1	Maaroufi	Ziyad

Terroristas con altos rangos de centralidad en cada grupo



red del ataque del 11-S

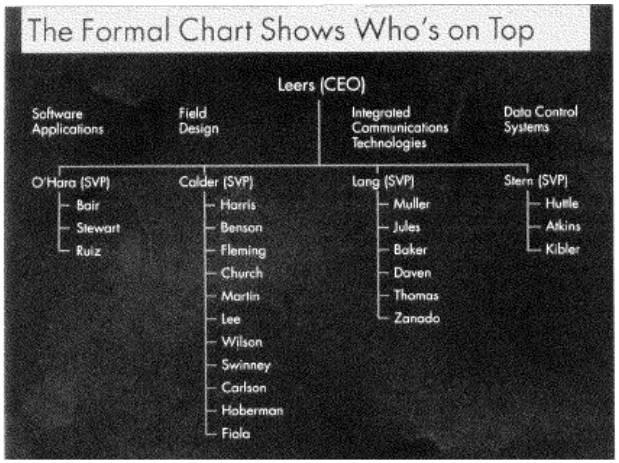


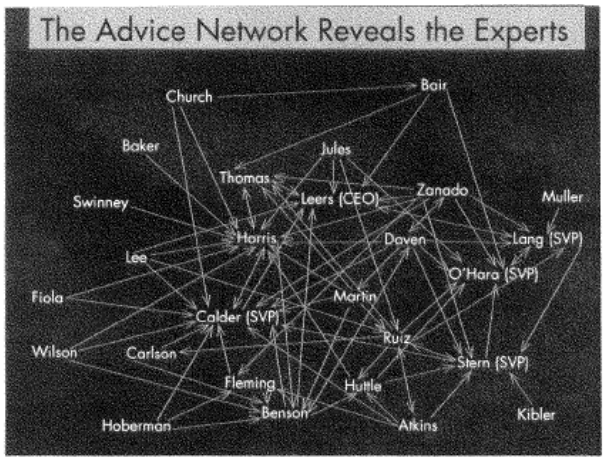
Grafo de la red GSJ

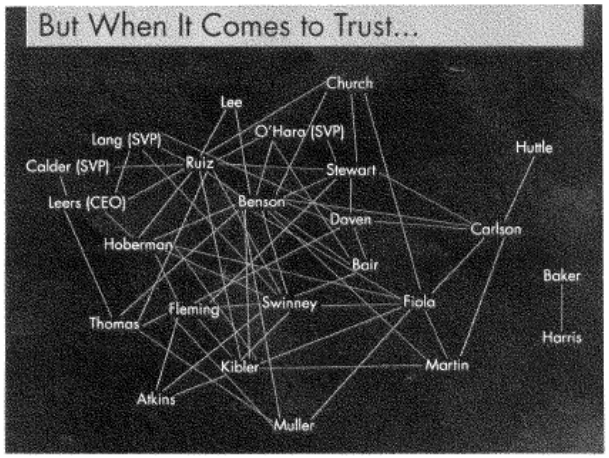
Aplicación a la teoría de la organización

Krackhardt and Hanson (1993)

- Las **redes informales** (sociales) presentes en una empresa son diferentes de las **redes formales**
- Existen diferentes patrones en tales redes debido a las **comunicaciones irregulares, estructuras frágiles, agujeros en la red**







(Krackhardt and Hanson,1993)

Metodología para realizar Analítica de Datos en una organización

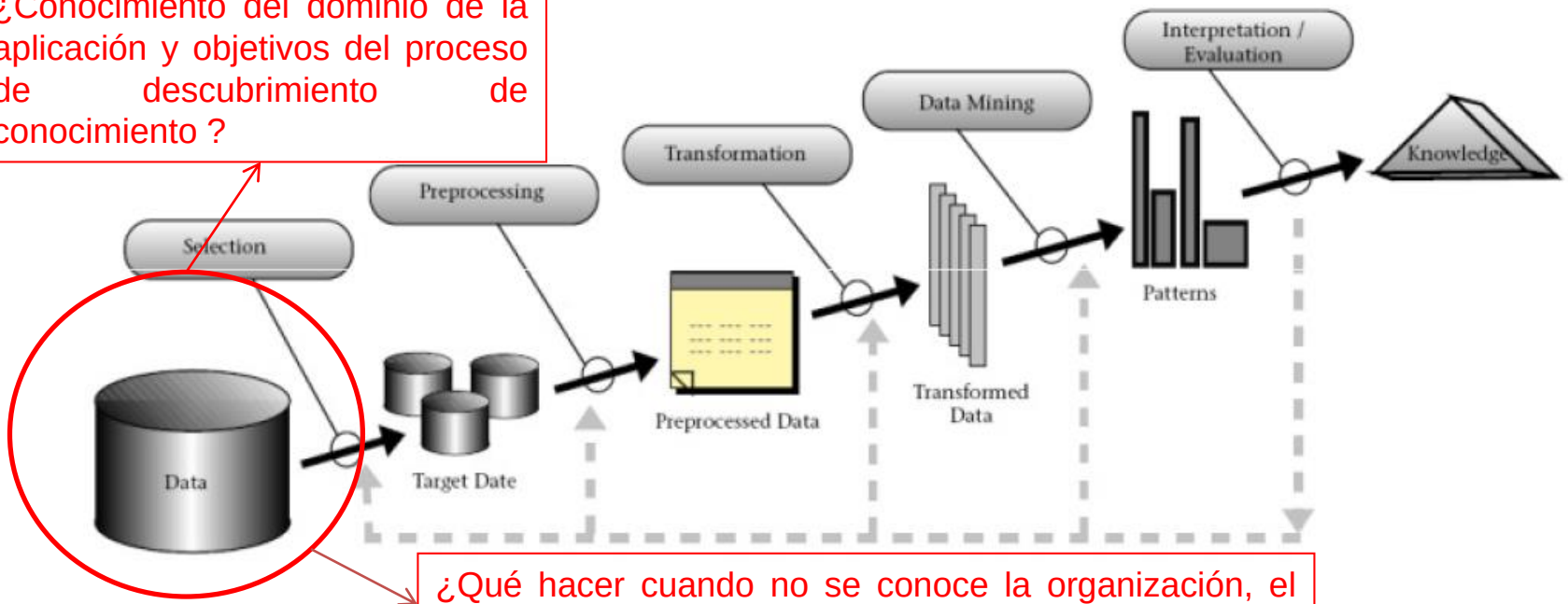
MIDANO

“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 ?

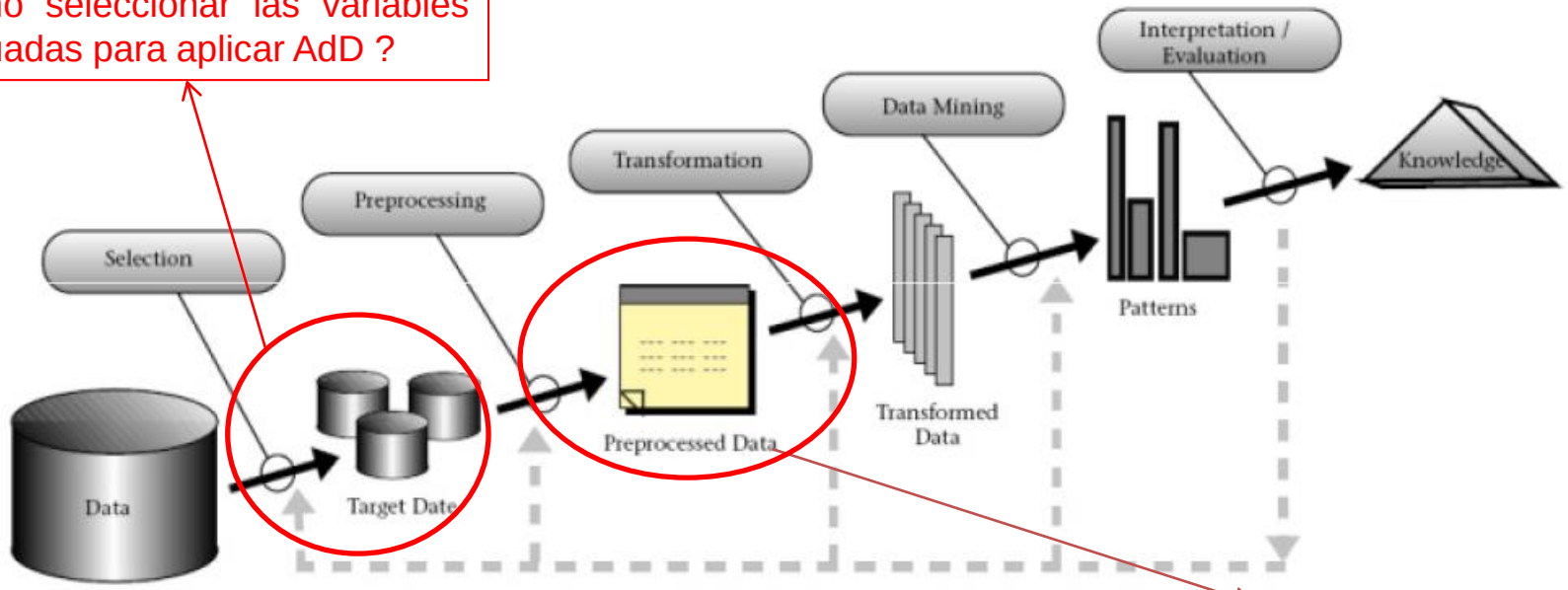


¿Qué hacer cuando no se conoce la organización, el problema, o los procesos a estudiar?

"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?

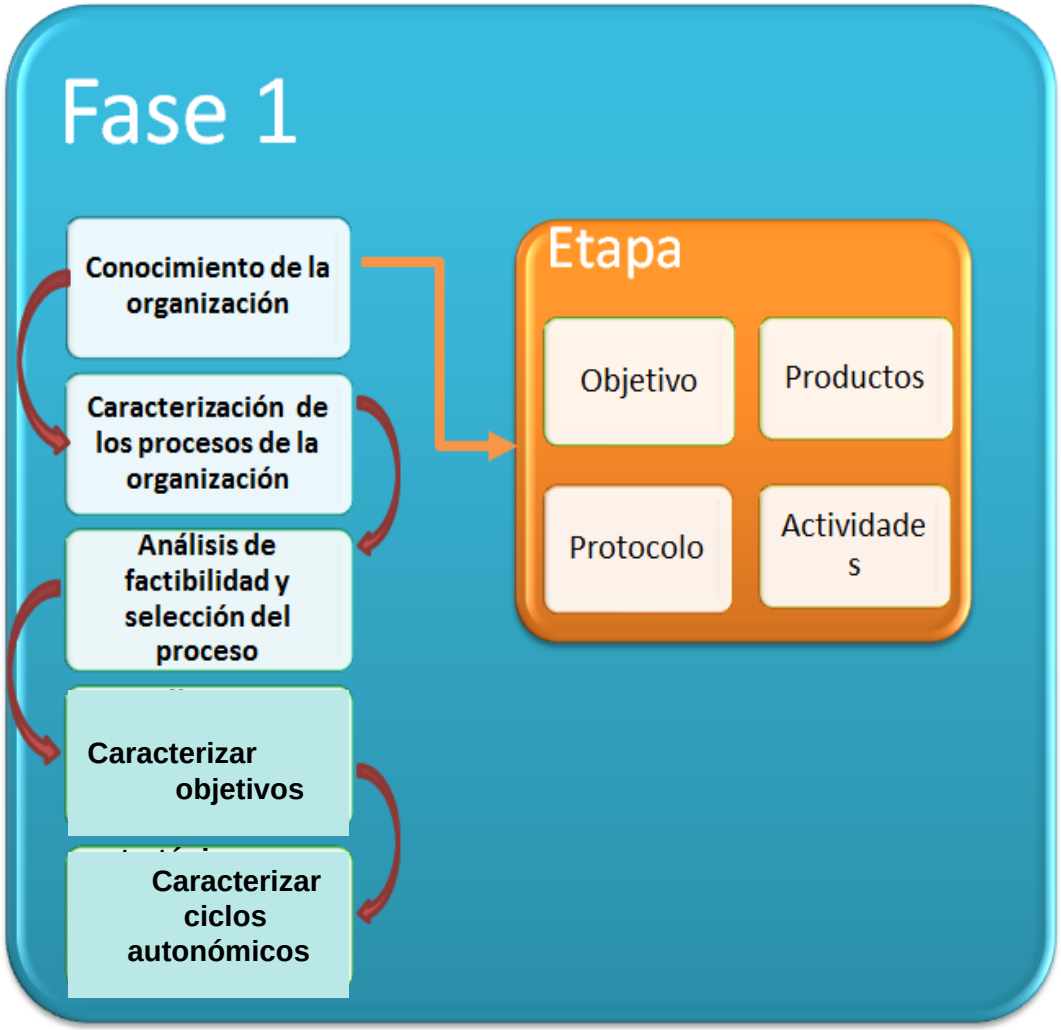
“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO-AdD

MIDANO consta de tres fases.

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

Vinculados a los **datos**

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 especificación detallada de las tareas de AdD

Macro-Algoritmo	Especificar Tipo de Tarea de Minería
<paso a paso del código>	< Debe indicarse de manera concreta la tarea a realizar>
…	Por ejemplo, calcular una medida de centralidad de minería de grafo, realizar un agrupamiento de tales datos según tales criterios de similitud, etc.)
…	

Esta tabla es particularmente importante para las tareas de AdDS

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 el modelo de datos multidimensional.

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

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**
   - Identificación de las variables en la VMC con las operaciones de:
     - (E)xtracción, (T)ransformación y Carga (L), para el caso de datos organizacionales
     - (C)olección, (C)uración y (A)nálisis para el caso de datos externos
   - Instanciación/Alimentación de las tablas (Cargar los datos)

b. **Protocolo de la etapa**
   - Integración de los datos de entrada en el DW

c. **Productos principales**
   - Tablas ETL y CCA

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 extraerá	Especificación del proceso de pre-procesamiento de los datos (estudios de dependencia, limpieza, cambio de formatos, etc.)	A que dimensión del modelo de datos irá

Tabla CCA

Variable	Colección	Curación	Análisis
Nombre de la variable	Identificación de fuentes externas para su obtención	Preparación de las operaciones para su obtención (limpieza, calculo, etc.)	Determinación de criterios sobre la calidad del dato (verificar si mide fenómeno deseado) y a que dimensión irá

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.

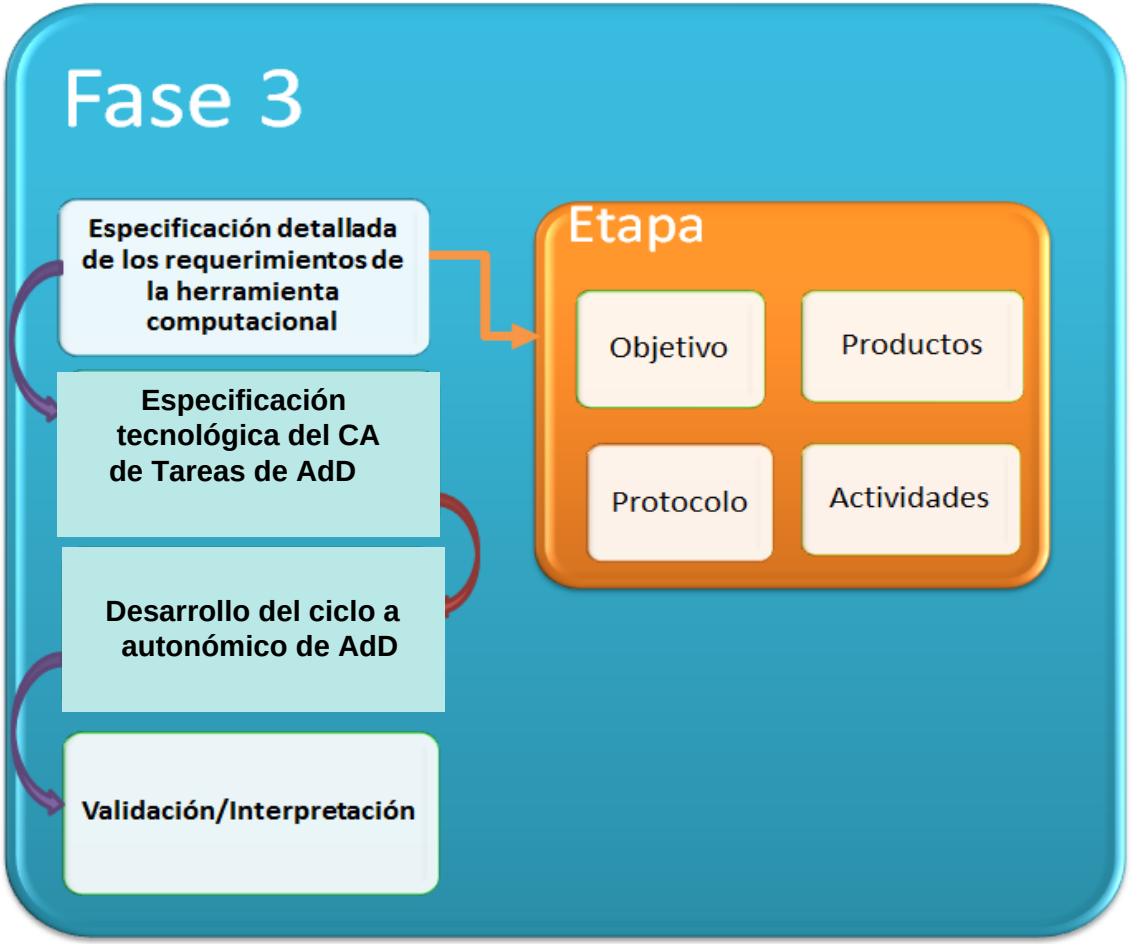
b. Protocolo de la etapa

- Limpieza
- Transformación
- Reducción
- Cálculos …

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 2: Especificación tecnológica del ciclo autonómico de Tareas de AdD

Tabla para especificación técnica de las 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

CookBook

- Resumen (Abstract)
- Palabras Claves (Keywords)
- Contribuyentes (Contributors)
- Versiones (Releases)
- Introducción (Introduction)
- Ingredientes: Definiciones y Terminología (The ingredients: Definitions and terminology)
  - Ingrediente 1 (Ingredient 1)
- Recetas (Recipes)
  - Receta 1: Una primera receta (Recipe1: A first recipe (e.g. a HelloWorld recipe))
    - Paso 1: descripción paso 1 (Step1: short description of step 1)
- Documentación Recomendada (Recommended documentation)
- Referencia 1 (Reference 1)
- Retroalimentación (Feedback)

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

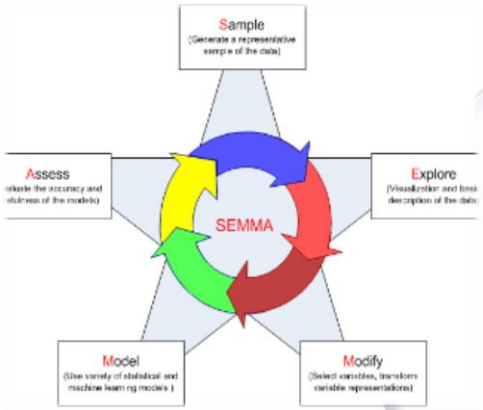
c. Productos principales

- 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



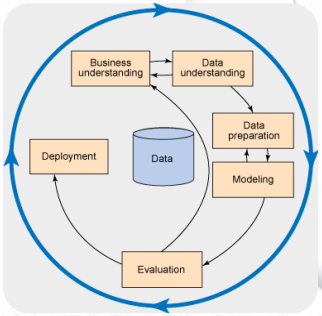
SEMMA

- 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 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

Ejemplo de uso de MiDANO-ext en SaCI

ESCUELA POLITÉCNICA NACIONAL

UTPL
UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

SaCI

Aula inteligente

Espacio donde la tecnología ubicua ayuda al **proceso de enseñanza-aprendizaje** de una manera transparente.

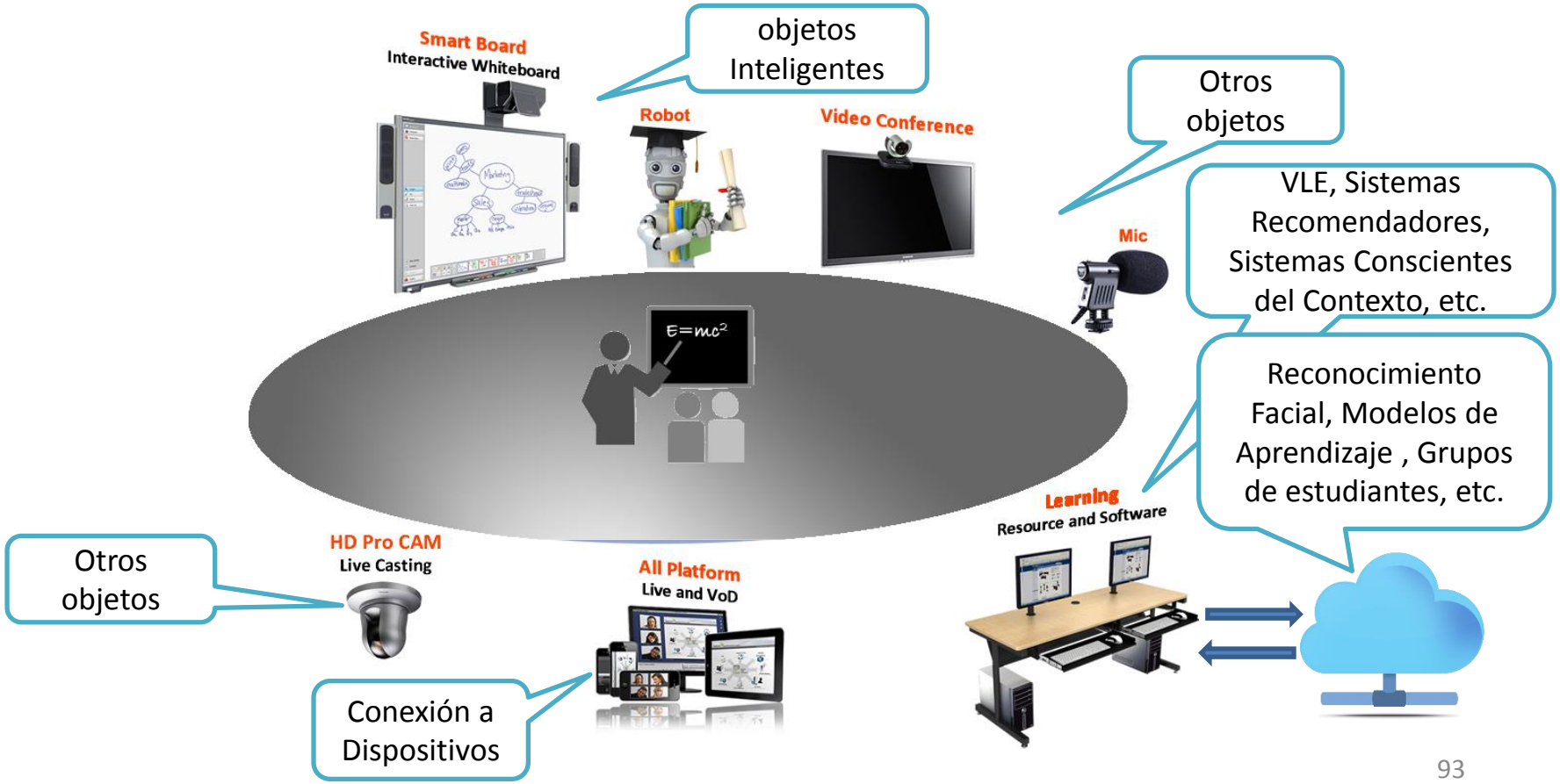
POLITÉCNICA
"Ingeniamos el futuro"

UNIVERSIDAD DE LOS ANDES

INSA TOULOUSE

Especificación de SaCI

Componentes de un aula inteligente



Objetivo AdD

Entender y mejorar el proceso de enseñanza y aprendizaje.

Objetivos específicos

- Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).
- Identificar los recursos educativos ideales para un estudiante en un momento dado.
- Identificar los estudiantes que necesitan más atención y sus necesidades.
- Evitar la deserción de estudiantil.

Especificación del Ciclo Autonómico

Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Ciclo 2

Objetivo: Identificar los recursos educativos ideales para un estudiante en un momento dado.

Ciclo 3

Objetivo: Identificar los estudiantes que necesitan más atención y sus necesidades.

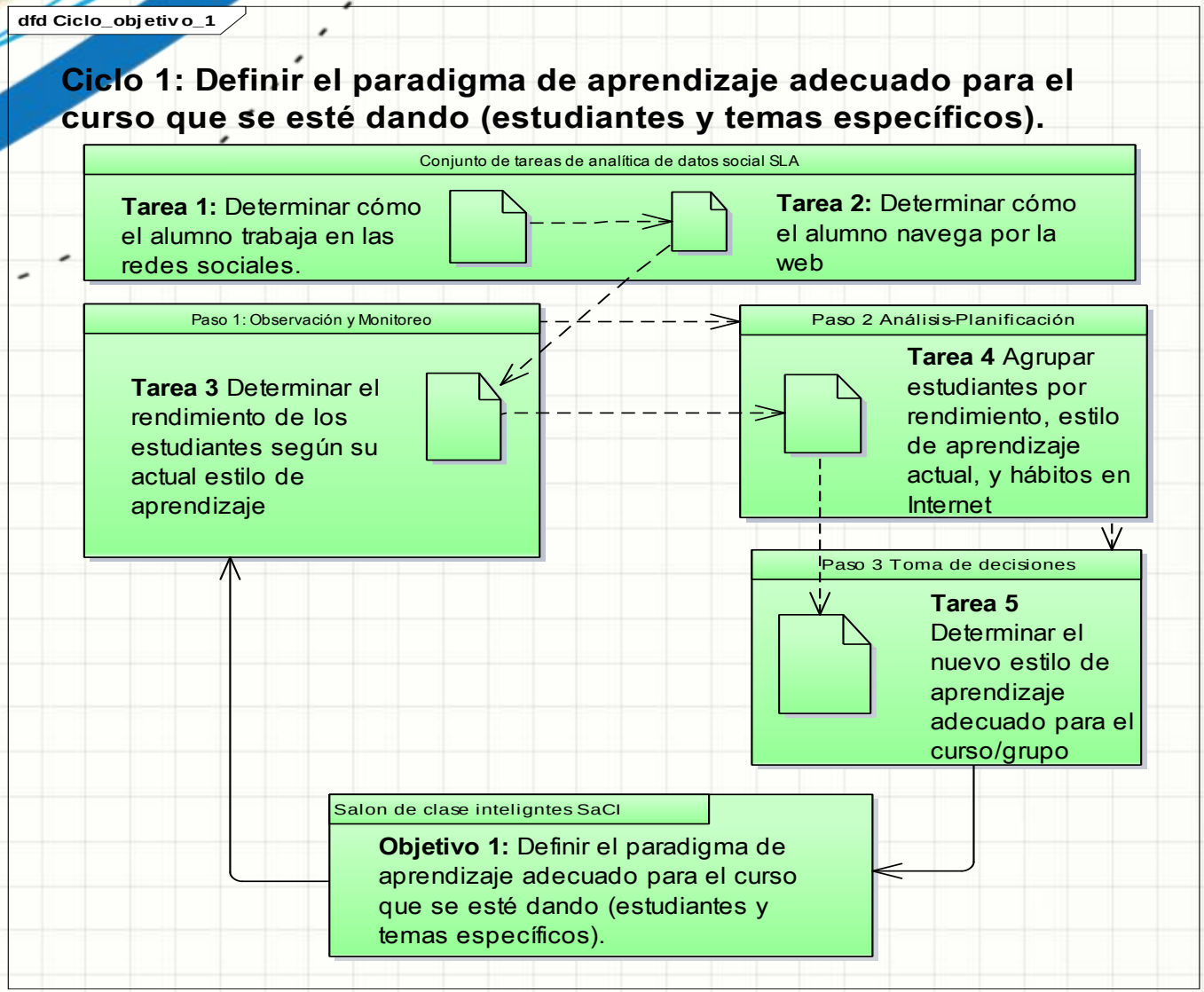
Ciclo 4

Objetivo: Evitar la deserción de estudiantil.

Implementación del Ciclo autonómico de Tareas.

Ciclo 1

Modelo del ciclo 1



Implementación del Ciclo autonómico de Tareas.

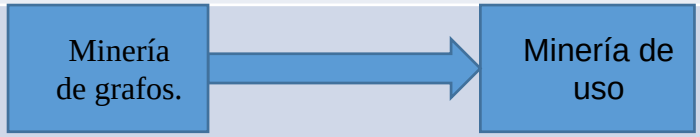
Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Paso 1: (Observación)

Tarea 1/5: Determinar cómo el alumno trabaja en las redes sociales.

Resultados Esperados	Actores	Variables	Herramienta Social	Actividades
Obtener el grafo del comportamiento del estudiante en redes sociales	Estudiantes	Estudiantes Redes sociales	Minería de grafos	Obtener el grafo de comportamiento

Nombre de la Tarea	**Determinar cómo el alumno trabaja en las redes sociales**
Descripcion	Determinar el grafo e comportamiento del estudiante en las redes sociales.
Fuente de Datos	Redes sociales
Tipo de tarea de analítica de datos social	Minería de grafos
Técnica de analítica social	Cascada Independiente / Influence Maximization
Tarea con la que se relaciona	
Tipo de tarea en el ciclo	Observacion

Implementación del Ciclo autonómico de Tareas.

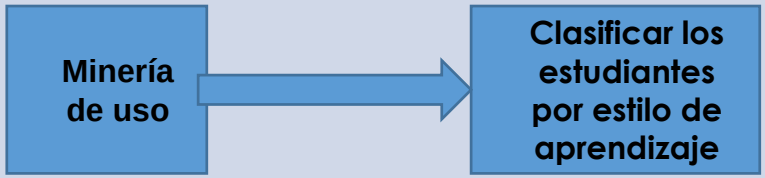
Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Paso 1: (Observación)

Tarea 2/5: Determinar cómo el alumno navega por la web

Resultados Esperados	Actores	Variables	Herramienta Social	Actividades
Obtener un patrón de navegación del estudiante en internet	Estudiantes	Estudiantes Páginas web	Minería de uso	Obtener el patrón de Navegación en internet

Nombre de la Tarea	Minería de uso para determinar como el alumno navega por la web
Descripción	Determinar la las preferencias del estudiante con relación a la web por medio de las páginas que visita y así obtener el estilo.
Fuente de Datos	Internet
Tipo de tarea de analítica de datos social	Minería de uso.
Técnica de analítica social	PageRank
Tarea con la que se relaciona	
Tipo de tarea en el ciclo	Observación

Implementación del Ciclo autonómico de Tareas.

Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Paso 1: (Observación)

Tarea 3/5: Determinar el rendimiento de los estudiantes según su actual estilo de aprendizaje

Resultados que se desean obtener	Actores	Variables asociadas	Actividades de MD a realizar	Actividades que se realizarán
Determinar el rendimiento de los estudiantes por estilo de aprendizaje utilizado.	Estudiante	Estilo de A. Estudiante	Clasificar	Clasificar estudiantes de un Estilo de Aprendizaje por rendimiento

Nombre de tarea	**Determinar el rendimiento de los estudiantes según su actual estilo de aprendizaje**
Descripción	Identificar los tipos de rendimientos de los estudiantes de acuerdo al estilo de aprendizaje utilizado
Fuente de datos	Base de Datos de SaCI
Tipo de tarea de analítica de datos	Clasificación
Técnica de analítica de datos	
Con que otras tareas se Relaciona	**Determinar el rendimiento E. por estilo** → Agrupar estudiantes por rendimiento y estilo de aprendizaje
Rol de tarea.	Monitoreo

Implementación del Ciclo autonómico de Tareas.

Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Paso 2: (Análisis)

Tarea 4/5: Agrupar estudiantes por rendimiento, estilo de aprendizaje actual, y hábitos en Internet

Resultados que se desean obtener	Actores	Variables asociadas	Actividades de MD a realizar	Actividades que se realizarán
Agrupar los alumnos de acuerdo al rendimiento obtenido por Estilo de A.	Estudiantes	Rendimiento Estilo de A.	Agrupación	Agrupar los alumnos por rendimiento y Estilo de A. utilizado

Nombre de tarea	Agrupar estudiantes por rendimiento, estilo de aprendizaje actual, y hábitos en Internet
Descripción	Agrupar los alumnos que han tenido mejor rendimiento, en un paradigma
Fuente de datos	Base de datos de SaCI
Tipo de tarea de analítica de datos	Agrupación
Técnica de analítica de datos	Series temporales
Con que otras tareas se relaciona	Agrupar estudiantes por rendimiento y estilo de A. utilizado → Definir el estilo de aprendizaje adecuado para el curso.
Rol de tarea.	Análisis

Implementación del Ciclo autonómico de Tareas.

Ciclo 1

Objetivo: Definir el paradigma de aprendizaje adecuado para el curso que se esté dando (estudiantes y temas específicos).

Paso 3: (Toma de Decisiones)

Tarea 5/5: Determinar el nuevo estilo de aprendizaje adecuado para el curso/grupo

Resultados que se desean obtener	Actores	Variables asociadas	Actividades de MD a realizar	Actividades que se realizarán
Definir el estilo de A. de aprendizaje adecuado para el curso	Estudiantes	Estilo de A. Cursos Temas Estudiantes	Toma de decisiones	Determinar el estilo de A. de mayor uso con el mayor rendimiento

Nombre de tarea	Decidir el estilo de aprendizaje adecuado
Descripción	Seccionar el estilo de A. que obtuvo mayor éxito en el rendimiento como el adecuado para el curso
Fuente de datos	Base de datos de SaCI
Tipo de tarea de analítica de datos	Toma de Decisiones
Técnica de A.D.	Reglas de decisión
Con que otras tareas se relaciona	Definir el paradigma de aprendizaje adecuado. → Tareas de SaCI
Rol de tarea.	Toma de decisiones

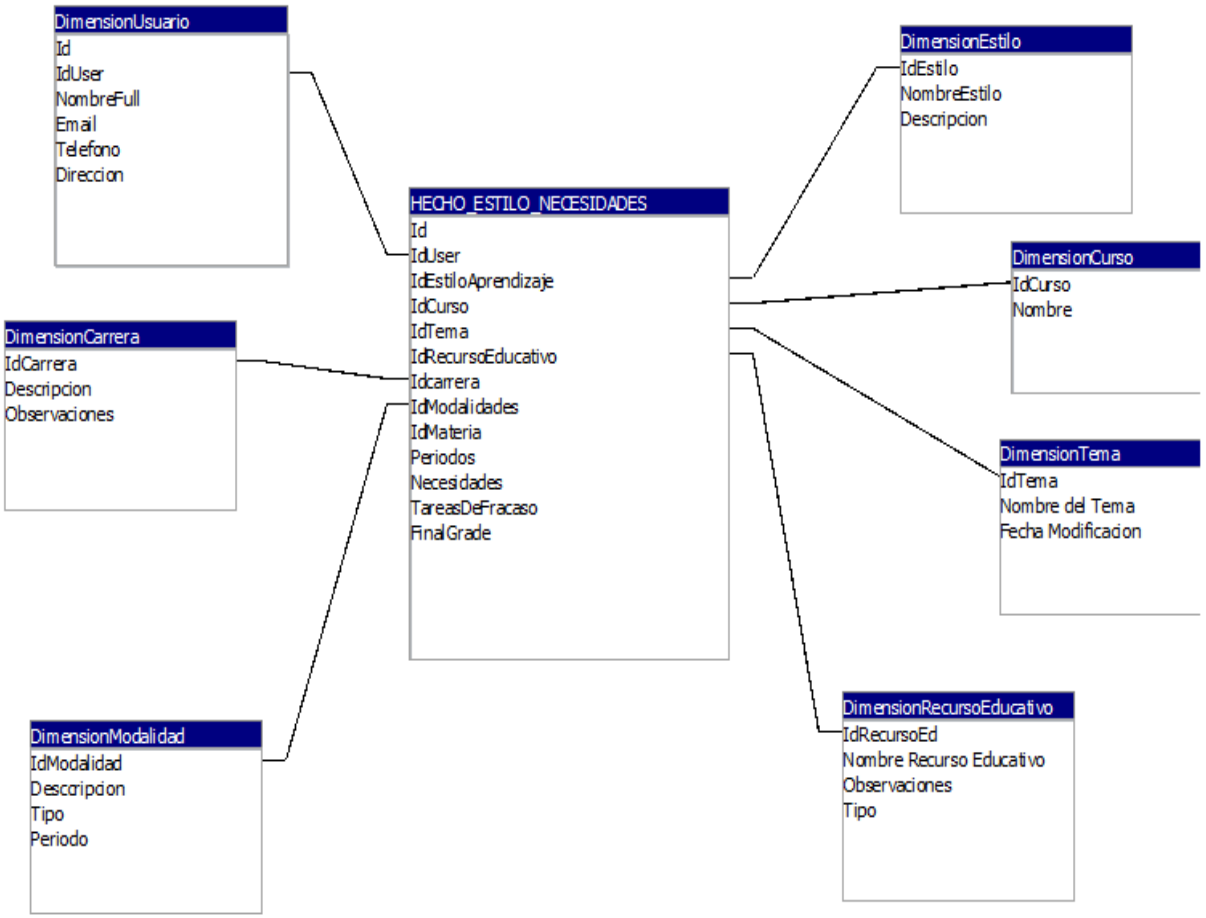
Modelo de datos de SaCI

Fuentes

- **Fuentes de datos Virtual Learning Enviroment (VLE):** Son las fuentes de datos provenientes de sistemas automáticos de enseñanza como por ejemplo los moodle.
- **Fuentes de Datos de Sistemas de Gestión académica (SGA):** Son fuentes de datos provenientes de sistemas internos para la gestión de notas de los estudiantes, por ejemplo OpenSIS.
- **Fuentes de datos de la nube:** Son todas las fuentes de datos disponibles en internet y de repositorios de datos como por ejemplo Twitter, Facebook, etc.

Modelo de datos de SaCI Ciclo 1

Modelo Multidimensional para el Ciclo 1



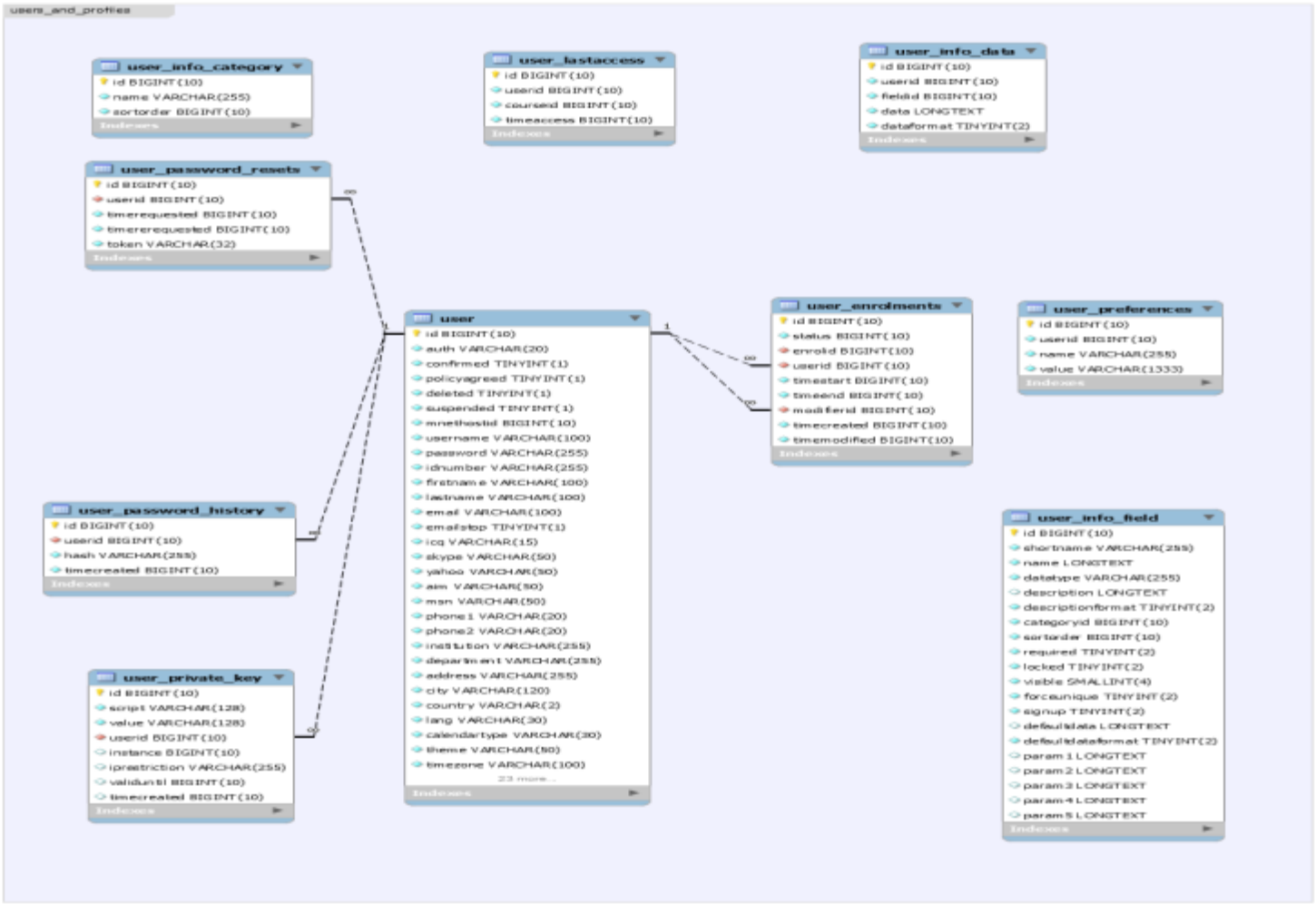
Modelo de datos de SaCI Ciclo 1

Tablas de VLE utilizadas para la muestra de datos

Tabla	Descripción
mdl_log	Registra de eventos LOG del VLE
mdl_user	Registra a todos los usuarios del VLE
mdl_grade_grades finalgrade	Almacena la nota final de un curso en el VLE
mdl_quiz	Registra cada uno de los exámenes del VLE
mdl_lessons	Registra cada una de las lecciones del VLE
mdl_workshop	Registra los trabajos y asignaciones en el VLE
mdl_quiz_grades	Registra loss resultados de los exámenes
mdl_course	Registra los cursos VLE

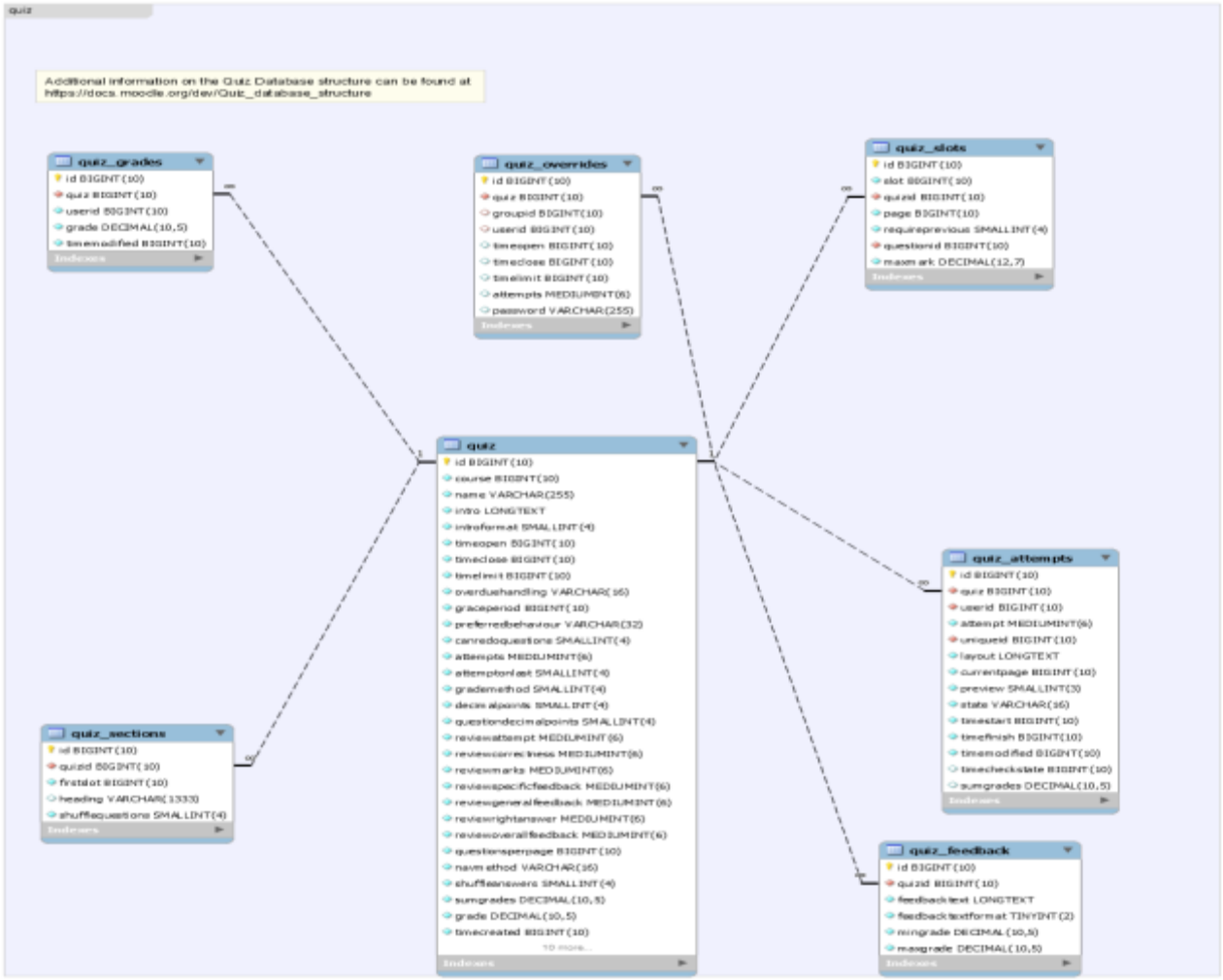
Modelo de datos de SaCI Ciclo 1

Modelo de datos VLE de los usuarios



Modelo de datos de SaCI Ciclo 1

Modelo de datos VLE de las evaluaciones



Operaciones Extraccion Transformacion y carga (ETL) consolidadas

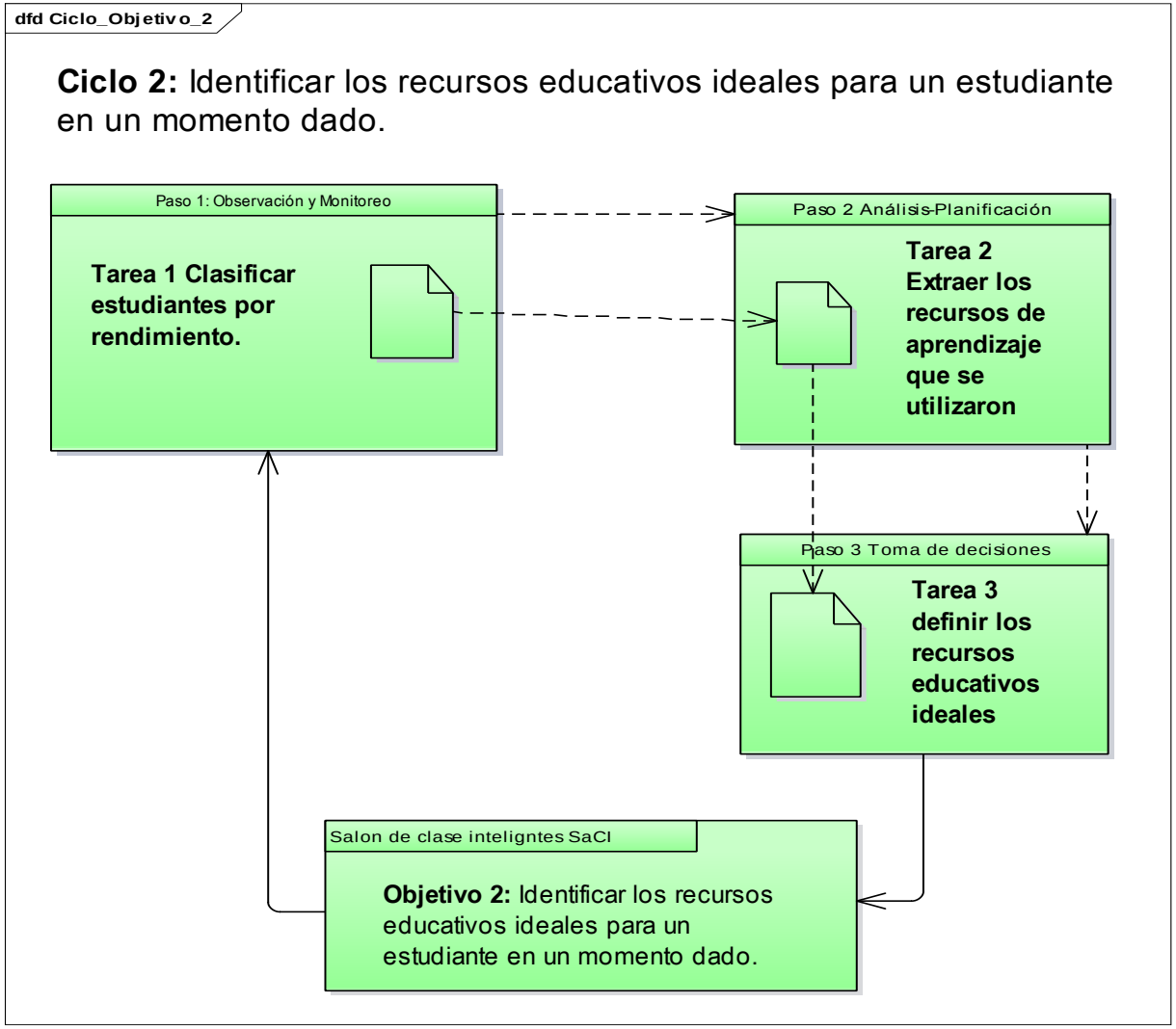
Operaciones Extraccion Transformacion y carga

Tarea que la consume	Variable	Fuente-Extracción	Proceso	Carga
Tarea 1 Tarea 2 Tarea 3 Tarea 4	userid	mdl_user	No sufre cambios	Se carga en la tabla de hecho
Tarea 1 Tarea 2 Tarea 3 Tarea 4	courseid	mdl_log	No sufre cambios	Se carga en la tabla de hecho
Tarea 1	cmid	mdl_log	No sufre cambios No cargar Nulos	Se carga en la tabla de hecho
Tarea 1 Tarea 2	fullName	mdl_log	No sufre cambios Se omiten nulos	Se carga en la tabla dimensional de alumnos
Tarea 1	action	mdl_log	No sufre cambios Se omiten nulos	Las acciones pueden ser:
Tarea 1	time	mdl_log	No sufre cambios Se omiten nulos	Se carga en la tabla de hecho
Tarea 2	grade	mdl_grade_grades	No sufre cambios Se omiten nulos	Se almacena en la tabla de hecho
Tarea 1	nombreEstilo	Fuente externa	Se calcula con valores de mdl_log	Se carga en la tabla dimensional de estilos
	email	mdl_user	No sufre cambios	Se carga en la tabla dimensional de estilos
	Teléfono	mdl_user	No sufre cambios	Se carga en la tabla dimensional de usuarios
Ciclo 4 Tarea 3	Materia	Mdl_course	No sufre cambios	Se carga en la tabla dimensional de materias

Implementación del Ciclo autonómico de Tareas.

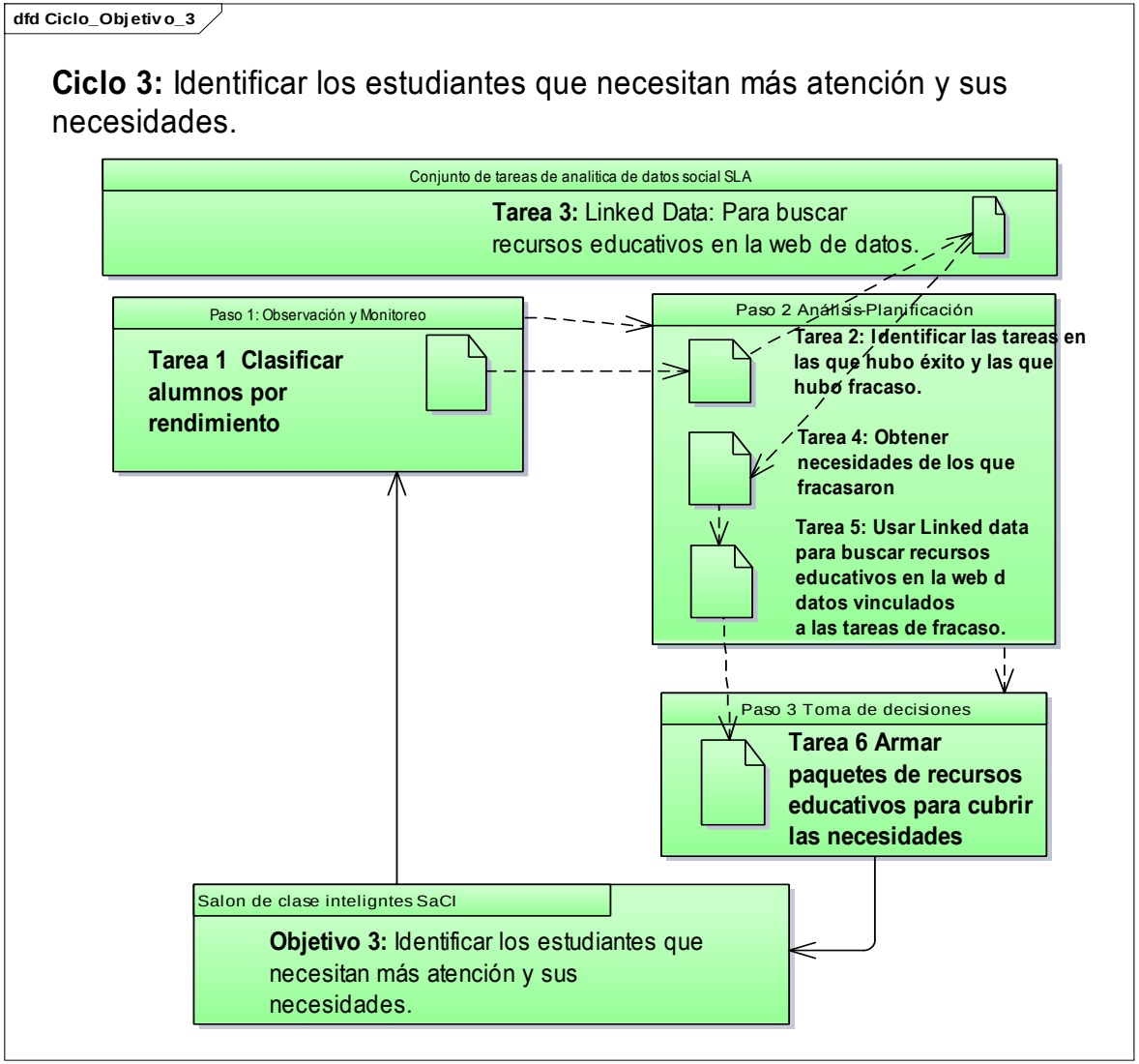
Ciclo 2

Modelo del ciclo 2



IMPLEMENTACIÓN DEL CICLO AUTONÓMICO DE TAREAS.
Ciclo 3

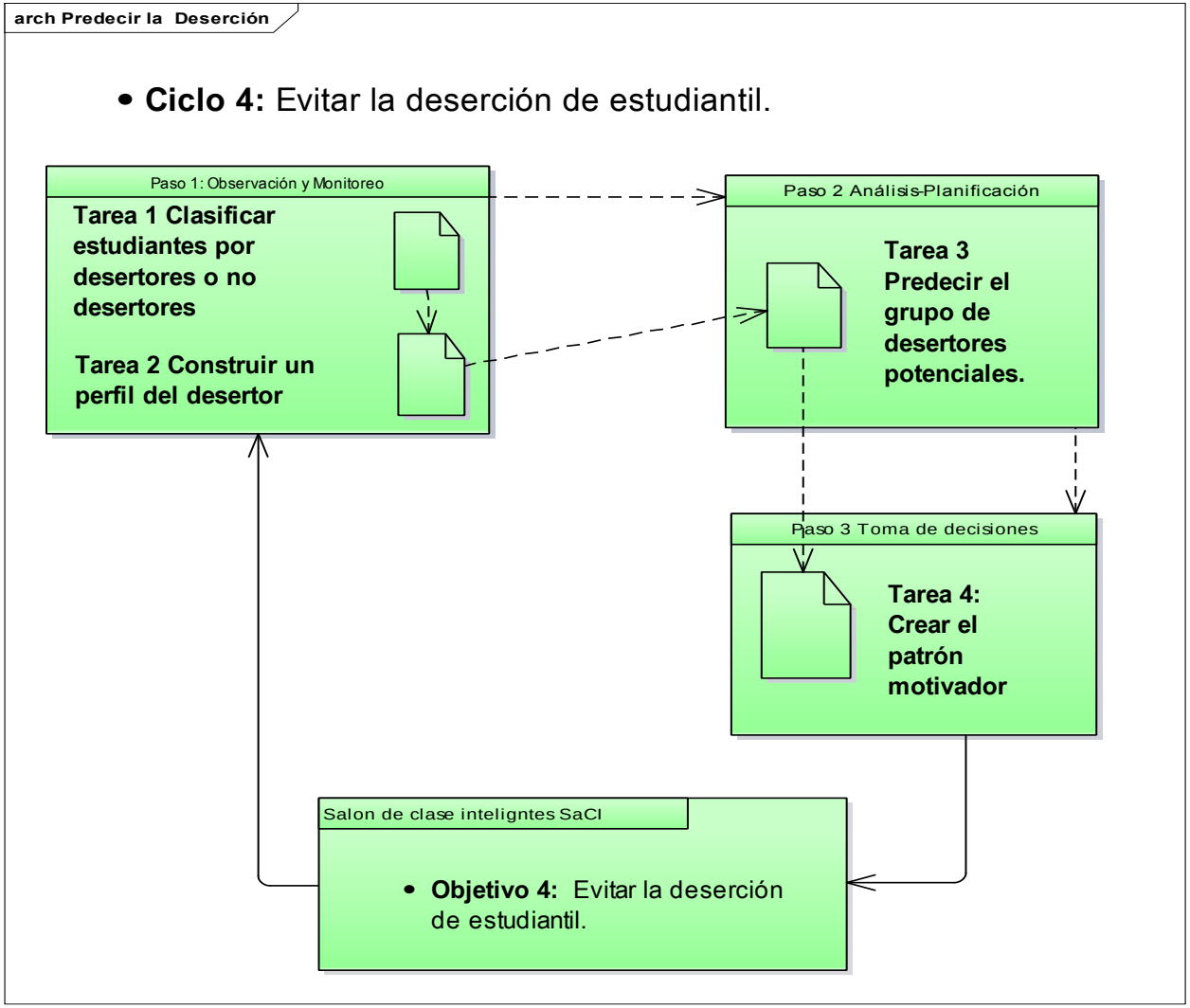
Modelo del ciclo 3



IMPLEMENTACIÓN DEL CICLO AUTONÓMICO DE TAREAS.

Ciclo 4

Modelo del ciclo 4

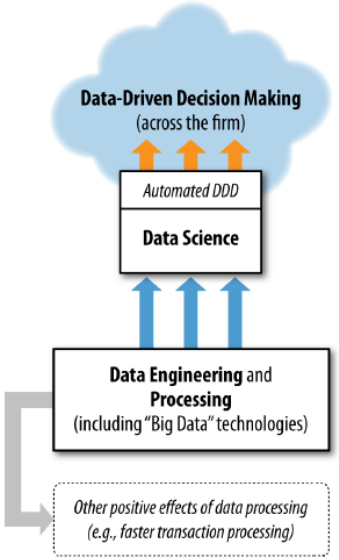


UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Ciencias de Datos

La ciencia de los datos

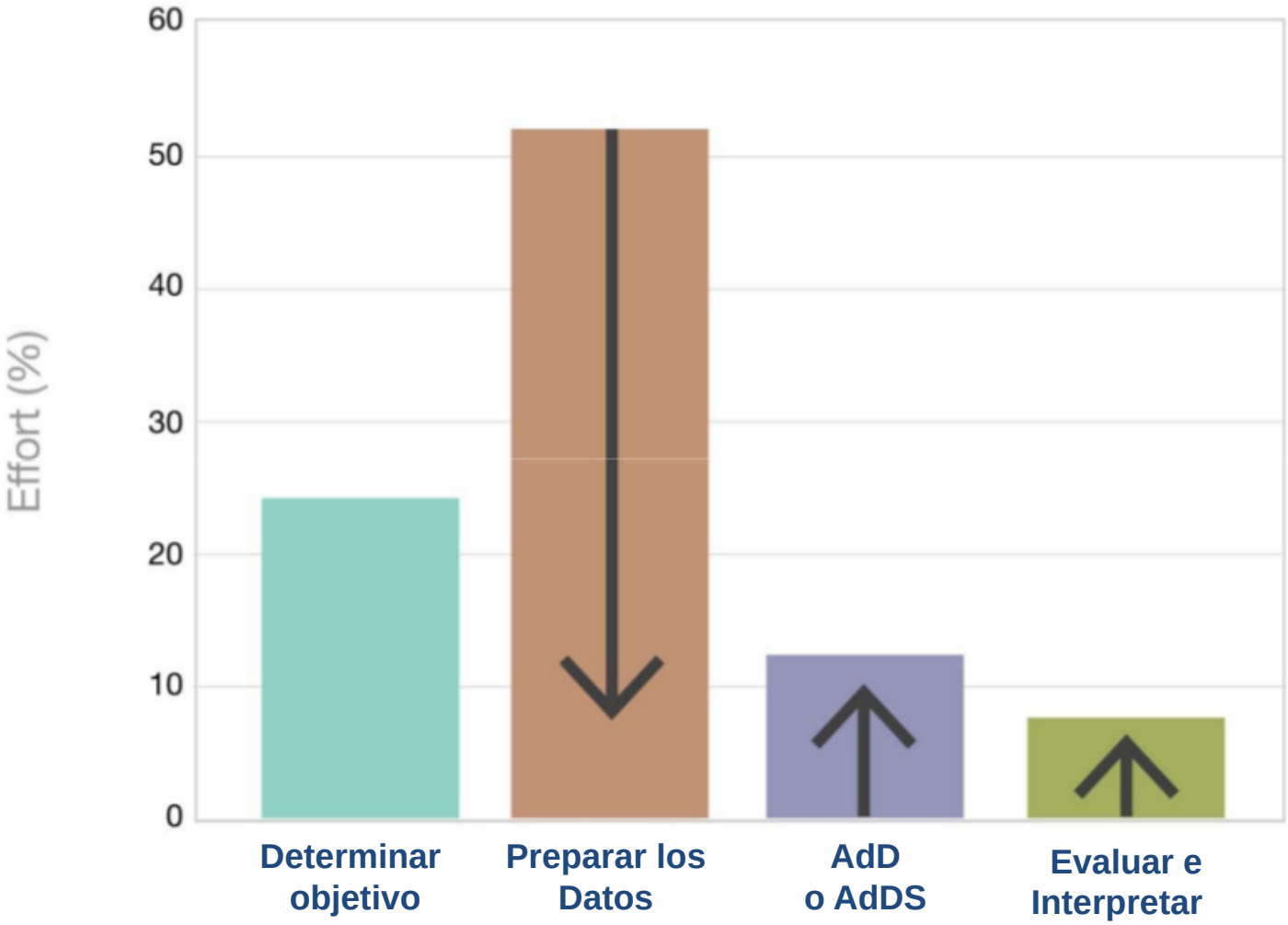
la ciencia de datos requiere de principios, procesos y técnicas para la comprensión de los fenómenos a través del análisis (automatizado) de los datos.



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.

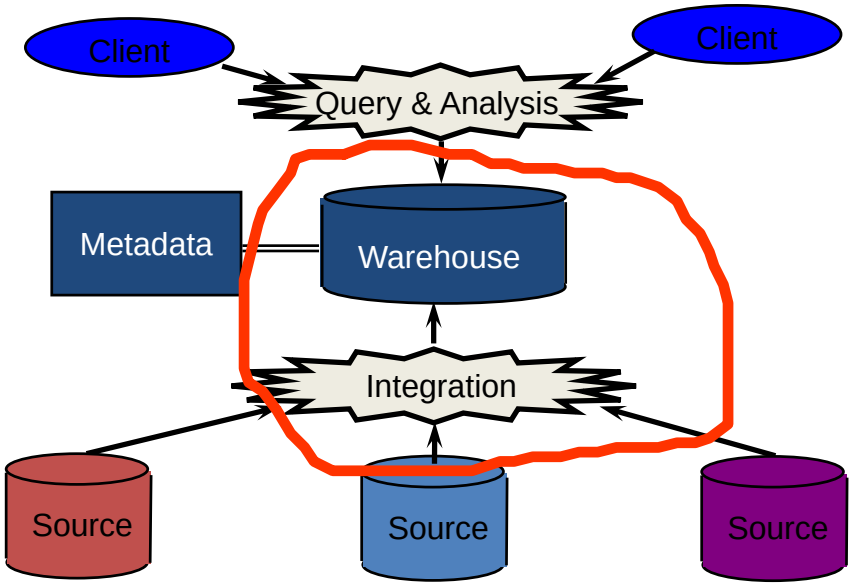
¿Que etapa lleva mas esfuerzo?



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

- Conocer y Selección de los datos
- Preparación de los datos
- Carga de datos



Conocer los datos

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.

Preparando los datos

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

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

Preparación de los Datos

- **Recolección de datos**
  - Captura de la Información
- **Análisis**
  - Entender el contexto de la información
- **Tratamiento de los datos**
  - Hacer ciencia en los datos

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

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.

Extracción

- Obtener datos de múltiples fuentes externas , heterogéneas
- Periódica
- Claves:
  - Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
  - Facilitar la integración de las diversas fuentes, internas y externas.

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas como datos mal capturados, anómalos y vacíos, ya sea por características obvias que el dato no cumple con ciertos parámetros del estándar, o porque el experto del proceso ya tiene identificado anomalías comunes en el almacenamiento de los datos.

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

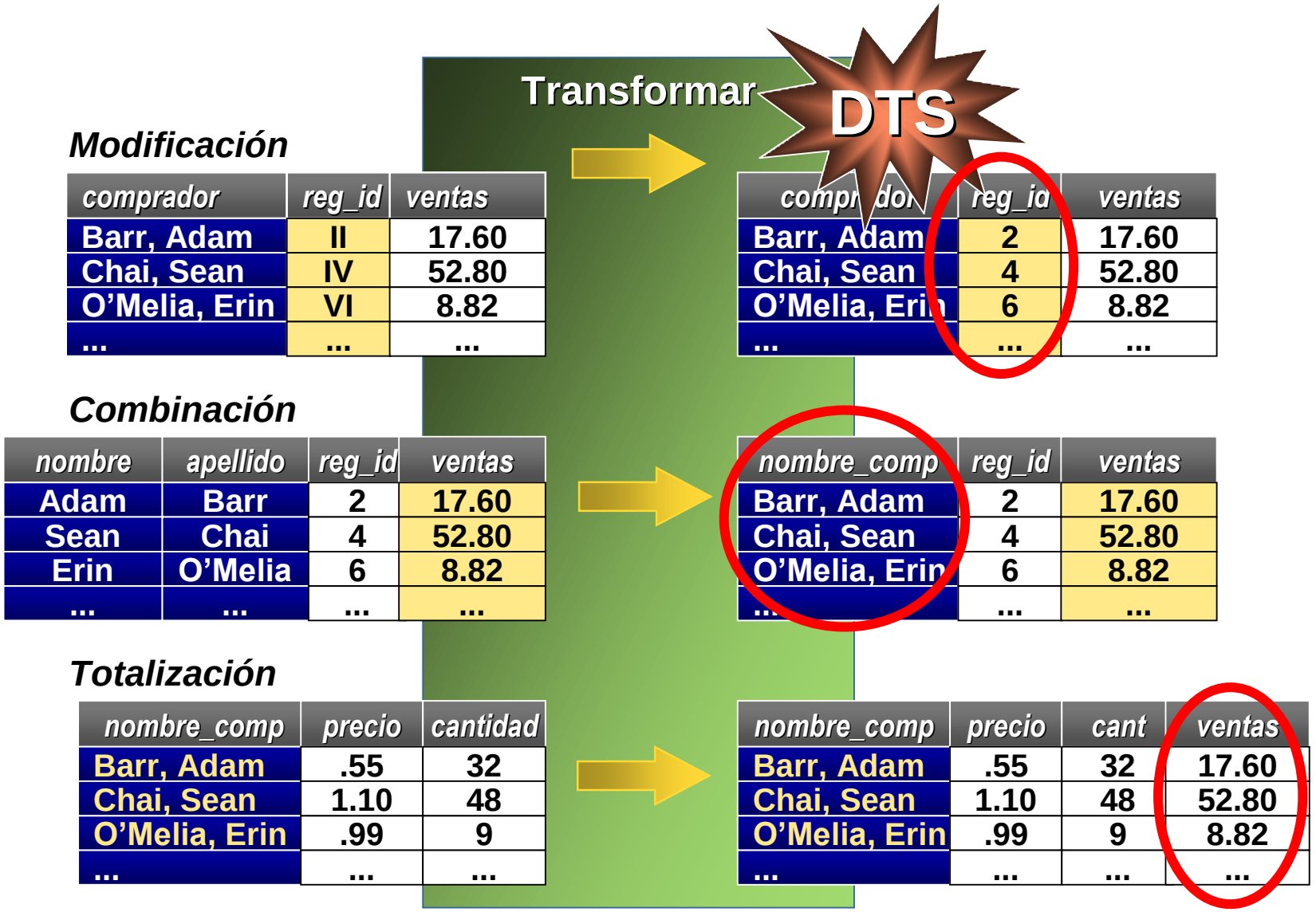
Transformación

En esta etapa se **transforman las variables de entrada en nuevas variables de interés**, esto se realiza a través de diversos métodos, los cuales se deben escoger en caso de ser pertinente alguna transformación de alguna de las variables.

Una transformación de variables puede ser la combinación entre variables

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

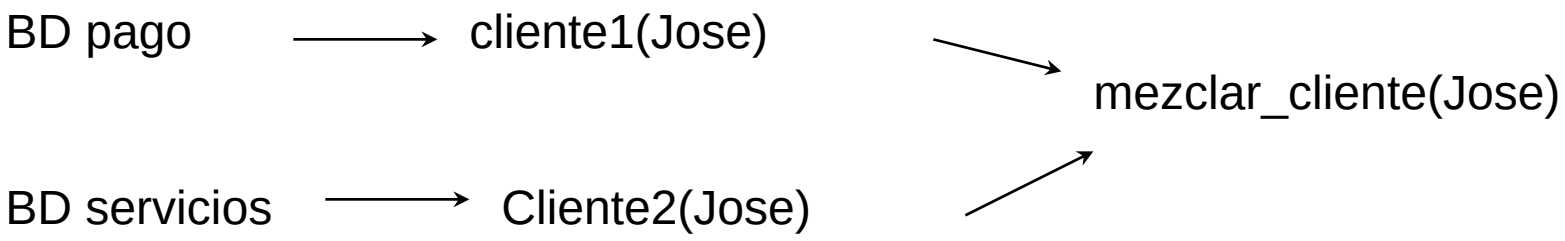
Proceso ETL


Transformar
DTS
Modificación
comprador
reg_id
ventas
Barr, Adam
II
17.60
Chai, Sean
IV
52.80
O'Melia, Erin
VI
8.82
...
...
...
reg_id
ventas
2
17.60
4
52.80
6
8.82
...
...
Combinación
nombre
apellido
reg_id
ventas
Adam
Barr
2
17.60
Sean
Chai
4
52.80
Erin
O'Melia
6
8.82
...
...
...
...
nombre_comp
reg_id
ventas
Barr, Adam
2
17.60
Chai, Sean
4
52.80
O'Melia, Erin
6
8.82
...
...
Totalización
nombre_comp
precio
cantidad
Barr, Adam
.55
32
Chai, Sean
1.10
48
O'Melia, Erin
.99
9
...
...
...
nombre_comp
precio
cant
ventas
Barr, Adam
.55
32
17.60
Chai, Sean
1.10
48
52.80
O'Melia, Erin
.99
9
8.82
...
...
...
...

ETL: Transformación de datos

Posibles tareas

- **Migrar** (por ejemplo, yen a dólares)
- **Refinar**: utilizar el conocimiento específico de dominio (por ejemplo, números de seguro social)
- **Fusionar** (por ejemplo, lista de correo con la de clientes)



Reducción

Datos Dispersos

Valores inexactos

Compresión de los datos

Valores Perdidos

Carga de Datos

- Los datos físicamente se almacena en el almacén de datos
- La carga ocurre en una "ventana de carga“
- La tendencia cada vez mayor es actualizaciones en tiempo real

¿Qué es lo Nuevo con AdDS?

Nuevas aplicaciones de AdD

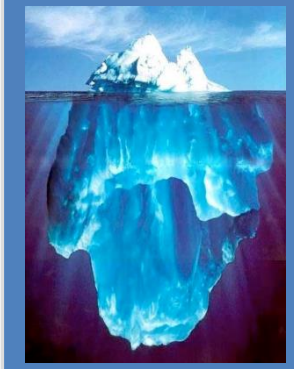
p.ej.,
Análisis de la propagación de virus Ebola

El modelo se basa en varias fuentes, tipos y análisis de datos.

“cuáles ciudades están en mayor riesgo"

¿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
...

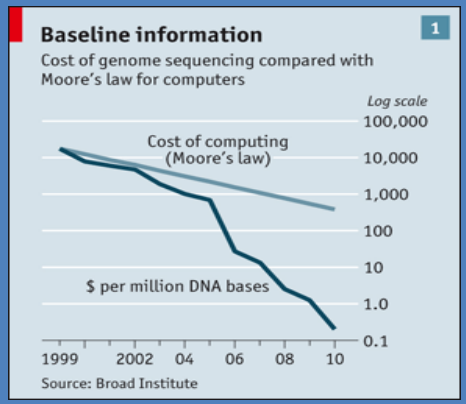
Generado por el usuario (Web y móvil)

...
..

Las fuentes

IoT

Computación Científica



¿Qué es lo Nuevo con AdDS?

Captura, Curación y Agregación

- Ciencia de datos debe trabajar con:
  - Datos incompletos
  - Los datos suelen estar desordenados
  - Analizar los datos para ver cuales de ellos son relevantes
  - Administrar grandes conjuntos de datos

Tipo de Datos

- Datos Relacionales (Tablas / Transacción / Datos Legados)
- Datos de texto (Web)
- Datos Semi-estructurados (XML)
- Grafos: Red social, Web semántica (RDF),
- Stream de datos

Fuentes

- Comercio electrónico
- Compras en tiendas de departamento / supermercado
- Transacciones bancarias / de tarjeta de crédito
- Redes sociales
- Fotos, documentos,

Fuentes

- **Red Social: Información de origen humano**
  - Redes sociales, Blogs, Documentos Personales, Imágenes, Vídeos, Búsquedas por Internet, Datos Móviles, Mapas generados por el usuario, E-mail
    Sistemas empresariales tradicionales: datos mediados por procesos

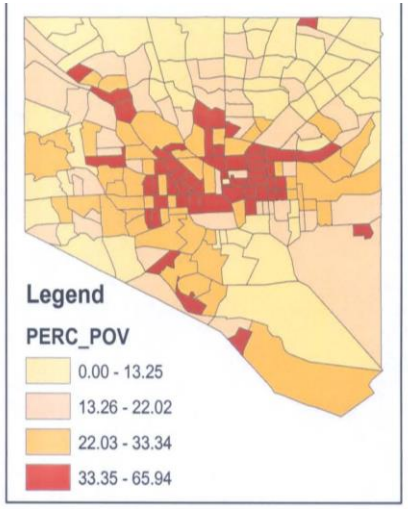
- **Agencias públicas (incluyendo registros médicos),**
  - producidas por negocios (transacciones comerciales, registros bancarios / de acciones, comercio electrónico, tarjetas de crédito)

- **Internet: generado por la máquina**
  - Sensores fijos: domótica, sensor de tiempo / contaminación, tráfico, científico, seguridad / vigilancia
  - Sensores móviles: teléfono móvil, automóviles, imágenes de satélite
  - Sistemas informáticos: registros, registros web

Fuentes

- **Series temporales:**
  - Precio de lps productos (gas, alimentos) mediante una función de los precios recientes, demanda, situación geopolítica ...
  - Tendencias estacionales
  - Redes de comunicación

- **SIG (sistemas de información geográfica)**
  - Longitud / latitud en la base de datos
  - Objetos: límites de ciudad/estado, ubicaciones de ríos, carreteras
  - Encontrar regiones con un exceso de cafeterías



credit: Frank Curriero

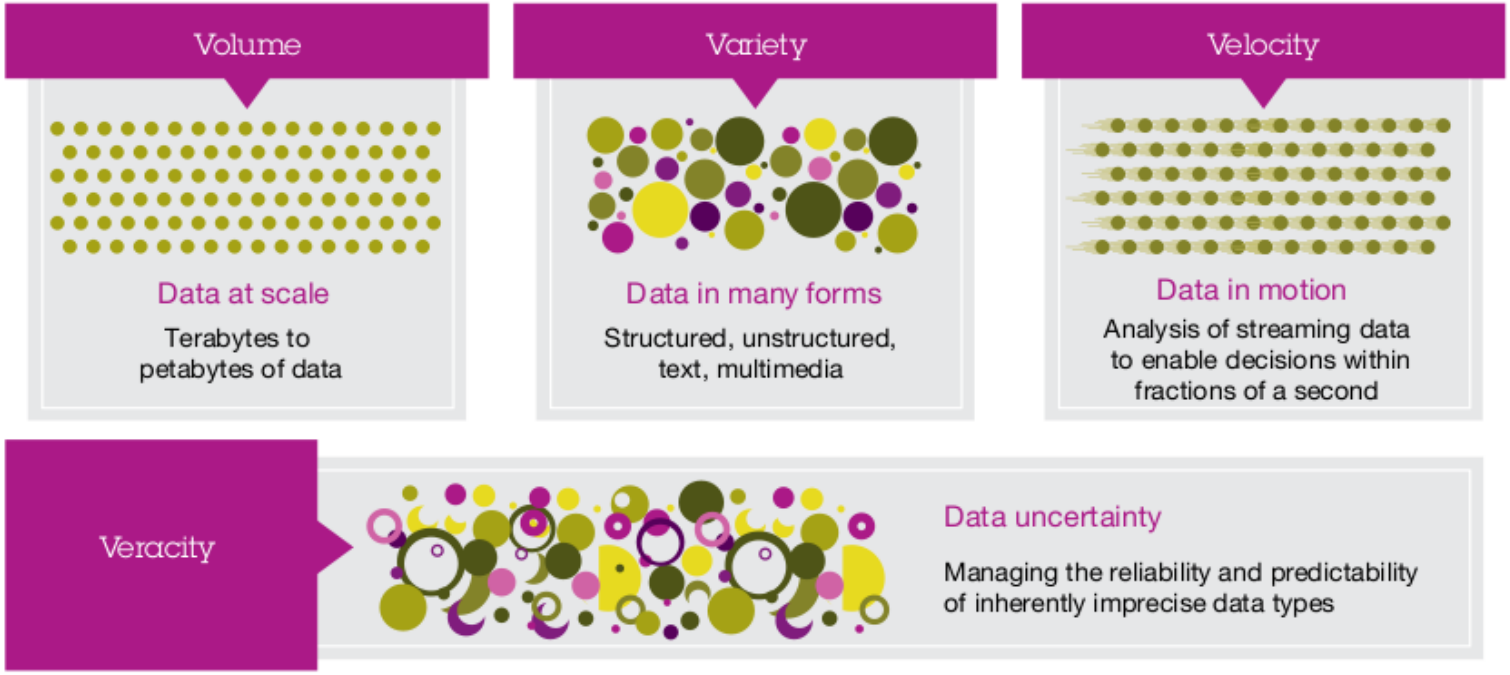
Retos AdDS

- **La construcción de un nuevo proceso de preparación de datos se realiza en muchas fases**
  - Caracterización de datos
  - Limpieza de datos
  - Integración de datos

- **Debemos mover los datos de manera eficiente en espacio y tiempo**
  - Transferencia de datos
  - Serialización de datos y deserialización (para archivos o red)

Retos AdDS

Un inmenso volumen, variedad y velocidad de los datos, en contexto, más allá de lo que era posible anteriormente.



Big data

Cadena de Valor

Colección: Datos estructurados, no estructurados y semi-estructurados de múltiples fuentes

Cargar grandes cantidades de datos en un único almacén de datos

Limpieza: comprensión del formato y contenido; Limpieza y formateo

Integración: vinculación, extracción de entidades, resolución de entidades, indexación y fusión de datos

Análisis: estadística, análisis predictivo y textual, aprendizaje automático

Entrega: consultas, visualización, entrega en tiempo real a la gerencia

Cadena de Valor

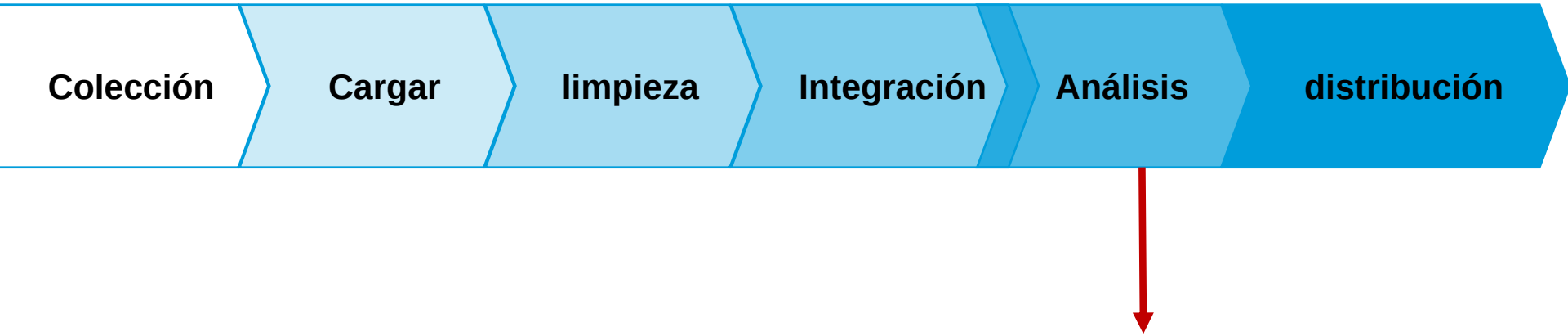
Transacciones
Redes sociales
Flujos de datos

- Ambiental
- Industrial
- GPS
- Imagen / Video
- Datos de red
- Registros del sistema
- Datos financieros

Cadena de Valor

Calidad de los datos

Cadena de Valor



CA de tareas de AdD

Capturar los datos

- Registrar datos generados por **diversas fuente**
- Mucho de esos datos **no son interesantes**
- Deben poder ser **filtrados y comprimidos**
- Datos **recogidos espacial y temporalmente**
- **Reducción Inteligente** de datos crudos
- Poder **minimizar al humano la carga**

Curación

- Frecuentemente, la información recogida **no esta en el formato listo para análisis**.
- Expresarla en un **estructura adecuada para el análisis**.
- Debe ser **correcta y completa**

Si usted puede limpiar y preparar datos rápidamente, usted tendrá un inmenso éxito dentro.

Nadie sabe cómo curar datos

- **Las fuentes de datos están fuera de control.**
  - Se están extendiendo tanques de almacenamiento masivo de datos.
  - Te dicen muy poco, y a veces nada, sobre cómo utilizar los datos que se almacenan en ellos.
  - Lo que la gente realmente necesita son los datos y el conocimiento detrás de lo que los datos significan.

- **El gran reto**
  - Las instituciones tendrán que aprender cómo recolectar datos a gran escala.
  - Saber 'cuándo' un trozo de datos cambia de significado es tan importante como saber 'qué' es ese elemento de datos.

Curar los datos

- **Rellenar los datos faltantes (valores de imputación)**
  - Detección y eliminación de valores atípicos
- **Suavizado**
  - Eliminando el ruido promediando valores juntos
- **Filtrado, muestreo**
  - Manteniendo sólo valores representativos seleccionados
- **Extracción de características**
  - p.ej. En una base de datos de fotos, ¿Quiénes están usando lentes? ¿Cuál tiene más de una persona? ¿Cuáles tienen persona al aire libre?

Agregar y Desagregar

- **Agregamos datos brutos** para que surjan patrones.

- **Desagregamos los datos** para desenmascarar cosas, información.

  estamos tratando de encontrar patrones.

- Gran reto- la mayoría de las veces, hacer ambas cosas.

Integridad de los datos

- Valores faltantes
  - Cómo interpretar ¿no disponible? 0? Usar el medio
- Valores duplicados
  - Incluyendo cosas parciales (Jon Smith = John Smith?)
- Incongruencias:
  - Varias direcciones por persona
- Uso inconsistente:
  - ¿Significa "destino" vuelo llegada?
  - Salarios que son negativos

Retos AdDS

Algunas reflexiones sobre "datos como un servicio“

- Establecimiento de **normas y directrices** (por ejemplo, arquitecturas y formatos abiertas)
- Creación de **modelos de intercambios de datos específicos** por sector (por ejemplo, datos sanitarios, datos medioambientales, etc.)
- Creación de **modelos de cruces de datos** (por ejemplo, interacción entre datos ambientales con datos sanitarios)

Peligros en AD

- 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

Tarea de Analítica de Datos

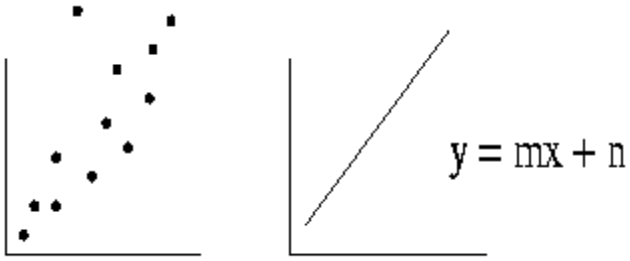
Jose Aguilar

Definiciones iniciales

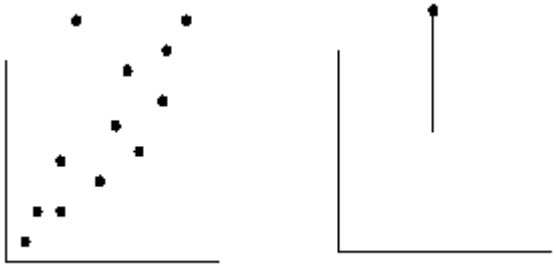
Conocimiento: Modelo y Patrones

Hand, Mannila y Smyth

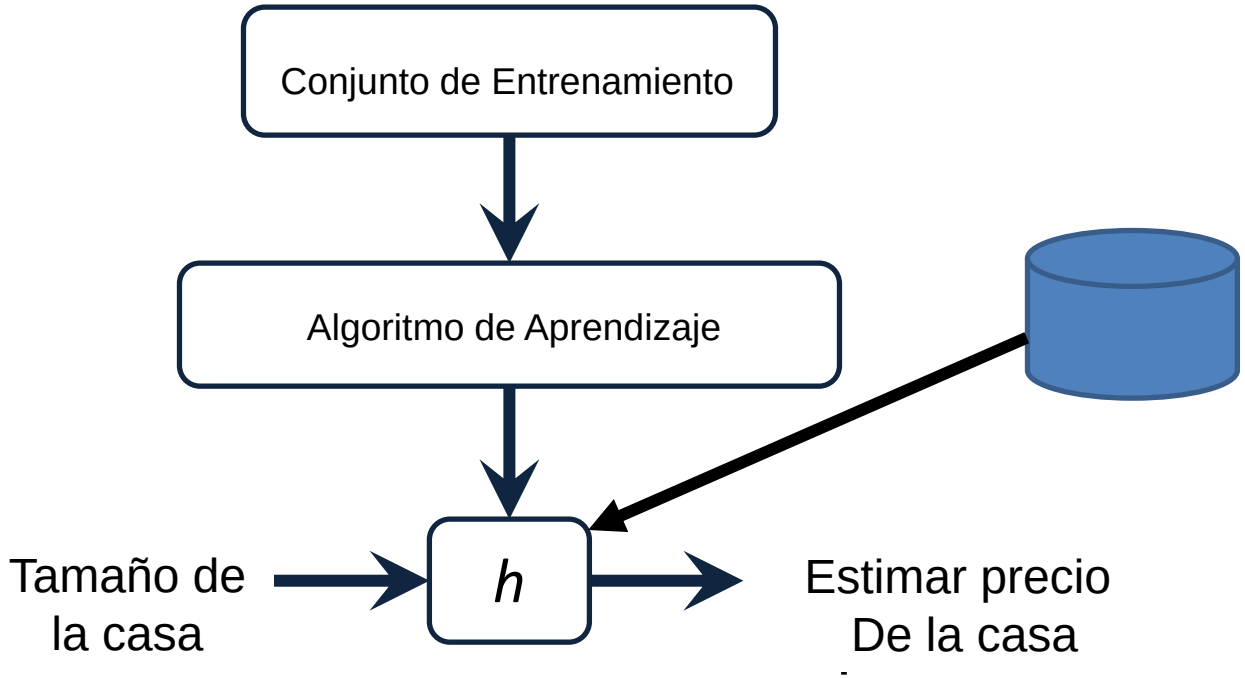
Modelo: Habla de todo el conjunto de datos



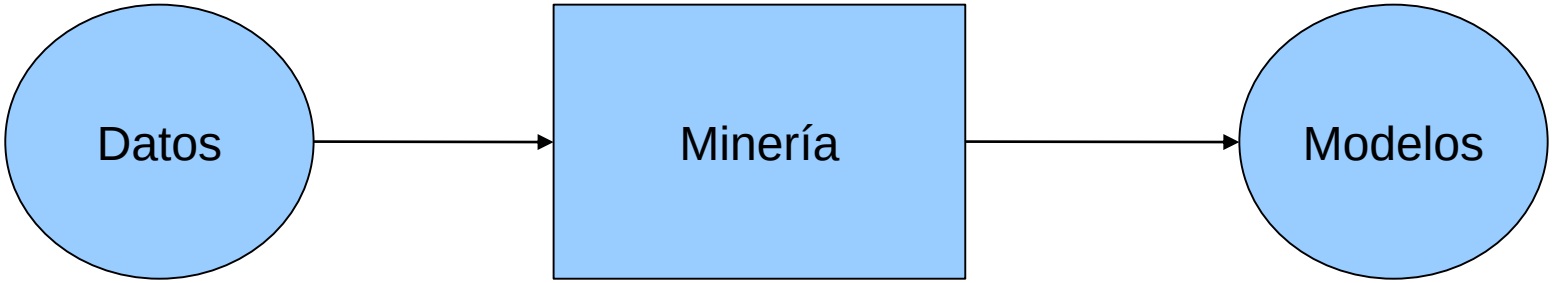
Patrón: Habla de una región particular de datos.



Construcción de modelos



Una visión simplificada de la minería de datos



- Los “modelos” son el producto de la minería de...
- ...y dan soporte a las estrategias de decisión que se tomen

¿Qué genera la AdD?

MODELOS!!!

- **Modelos Descriptivos**

Encontrar patrones interpretable que describen los datos.

- **Modelos de Predicción**

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

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

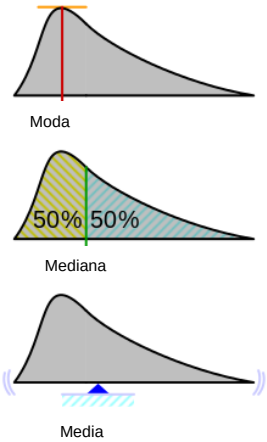
Herramientas en AD para generar modelos

- La minería de datos
- El análisis estadístico
- El análisis predictivo
- La Correlación
- La Regresión
- Pronosticar
- Modelado de procesos
- Optimización
- Simulación

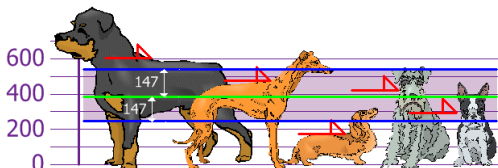
Dos categorías principales:
* Estadísticas descriptivas
* Estadística inferencial

Las estadísticas descriptivas básicas

- 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)



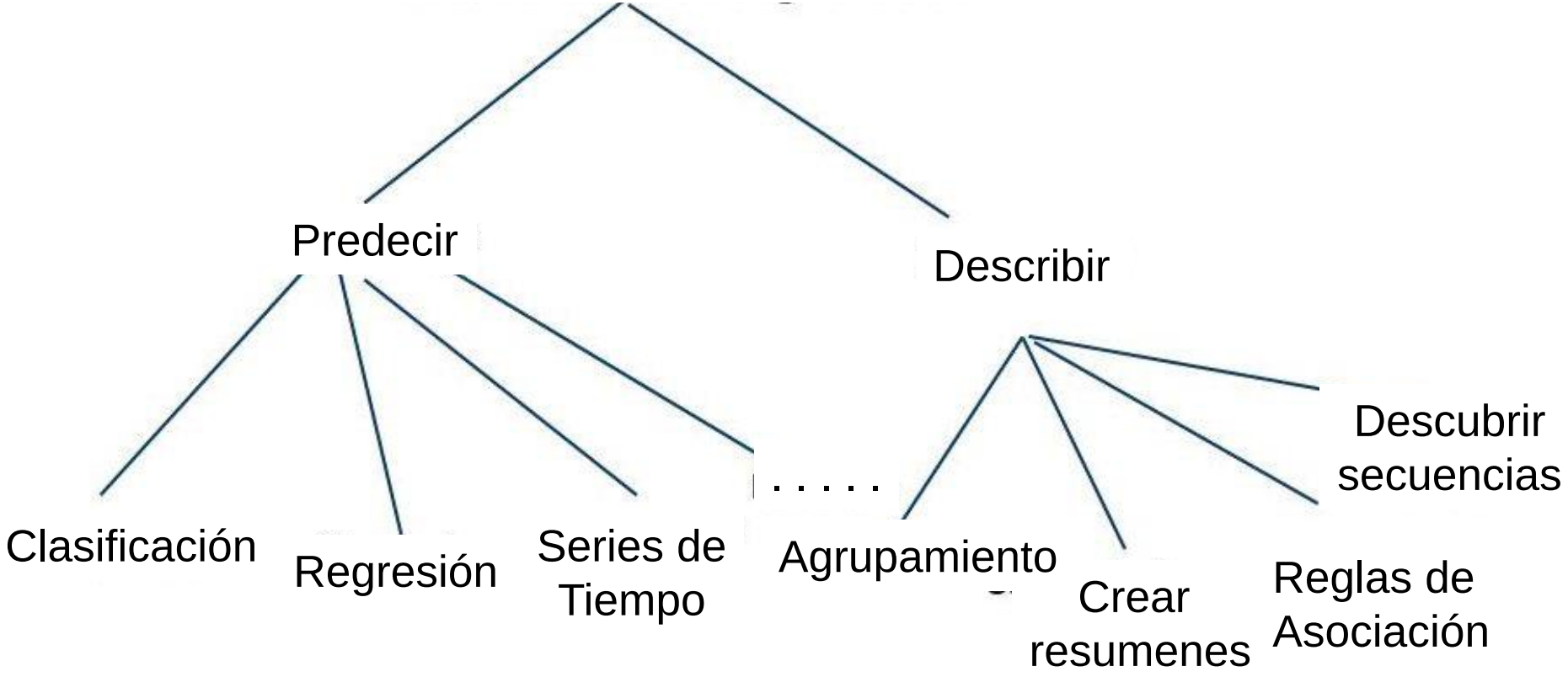
Compradores	Número
Hombre	
Viejo	6
Joven	4
Mujer	
Vieja	10
Joven	15

- **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**

Tareas clásicas para la AD

- Clasificación [**predictivo** y descriptivo]
- Clustering [**descriptivo**]
- Descubrimiento de Regla Asociación [**descriptivo**]
  - Análisis de dependencia de datos
  - correlación y causalidad
- Descubrimiento Patrones Secuenciales [**descriptivo**]
  - Análisis de series de tiempo, asociaciones secuenciales
- Regresión [**predictivo**]
- Detección de Tendencia y Desviaciones [**predictivo**]
- Filtros Colaborativos [**predictivo** y descriptivo]
- Resumir [**descriptivo**]
- Descripción de Conceptos [**descriptivo**]
  - Descripción de características
  - Descripción de su identidad discriminante

Tareas clásicas para la AD vs modelos



UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Minería semántica

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Algunas clases de Mineria

- Minería de Datos Espaciales
- Minería espacio-temporal y de objetivos en movimiento
- Minería de dominios: salud, control de tráfico aéreo, inundaciones
- Minería de datos multimedia
- Minería de texto
- Minería de la Web
- Minería de streams de datos

Minería de Secuencia de Datos

- Buscar Similitud en serie temporal de datos
- **Regresión y Análisis de Tendencias en series temporales de datos**
- Minería en secuencias simbólicas para buscar patrones secuenciales
- **Clasificación de Secuencia**
- Alineación de secuencias biológicas

Minería Semántica

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

Determinar relaciones semánticas

- Minería de Datos Semánticos (*Semantic 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 Semántica

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

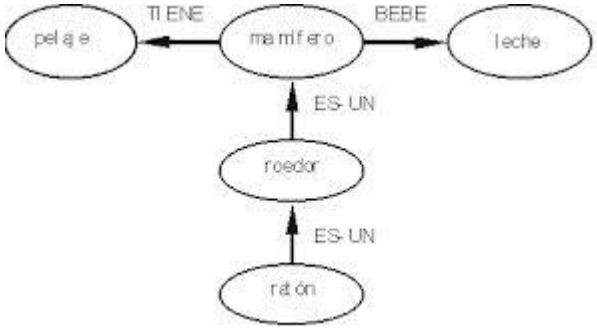
- Cuando los datos y el conocimiento del dominio están disponibles, vale la pena **explorar la relación semántica entre ellos**.

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

Minería Semántica

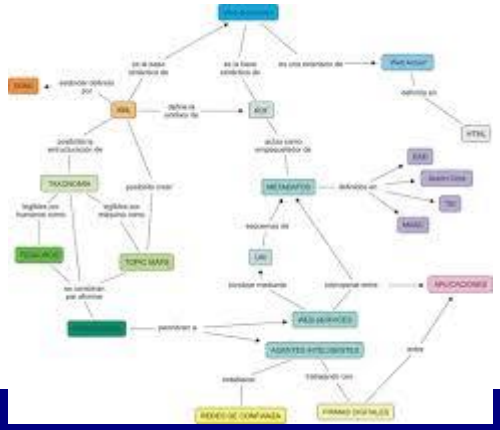
	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 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

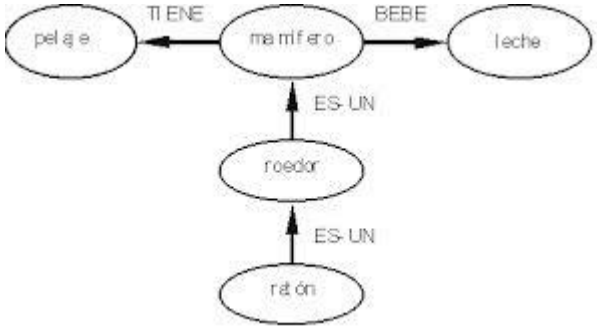


Minería Semántica

	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



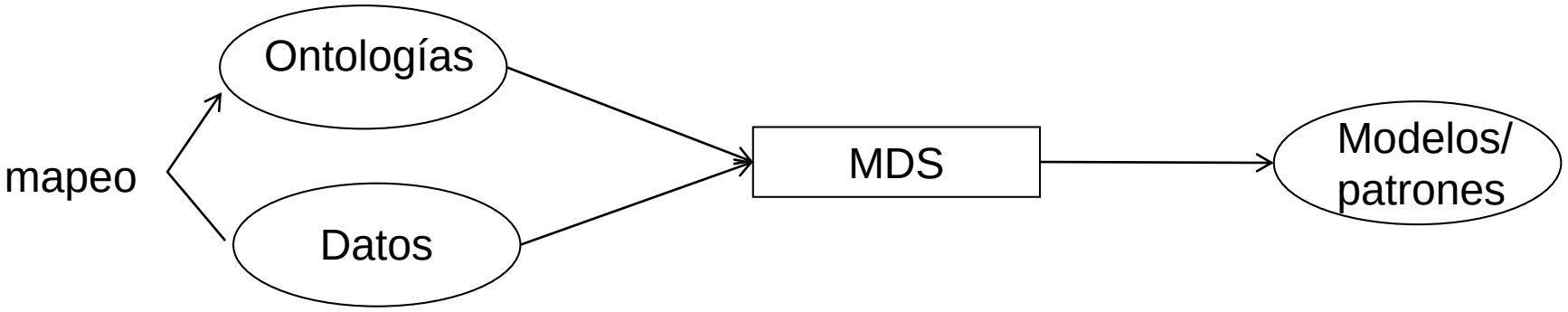
Minería Semántica

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.

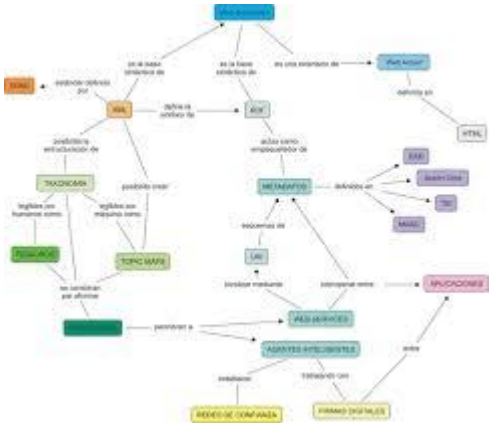
- En el segundo paso se aplican técnicas de MD para **buscar patrones, relaciones**, y en general, cualquier operación que **explote el enriquecimiento semántico**.

conocimiento

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, etc.

	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, etc.

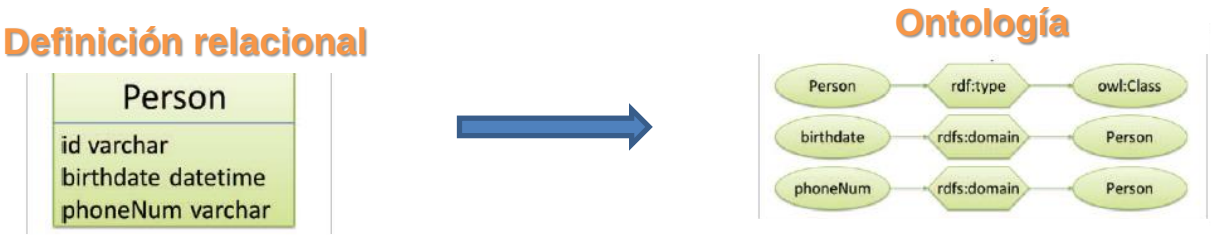
Minería de Datos Semánticos

- **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



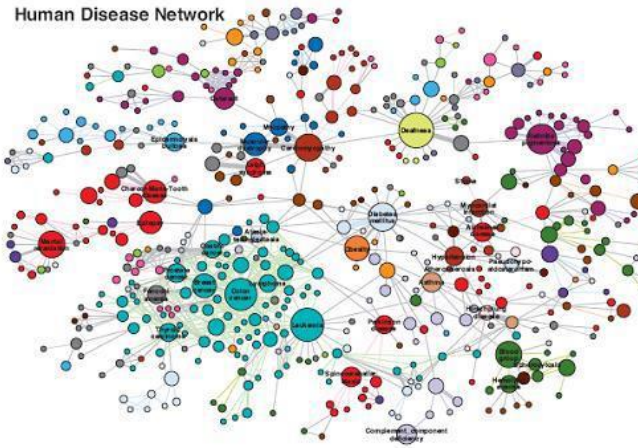
MDS en Clusters de Signaling Pathways Networks

Una red de vías de señales, o signaling pathway, es el conjunto de **reacciones implicadas** en la reacción de una célula a un estímulo externo.

Los clusters no dan mucha información per se, pero al identificar las funciones biológicas que identifican cada cluster se pueden definir **familias**.

La **activación del receptor** provocada por la unión a un ligando se asocia directamente a la respuesta de la célula.

Por ejemplo, el **neurotransmisor GABA** puede activar un receptor de la superficie celular que es parte de un canal iónico.

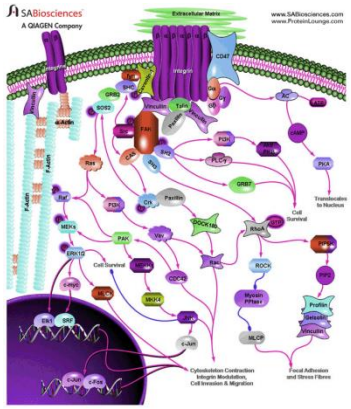


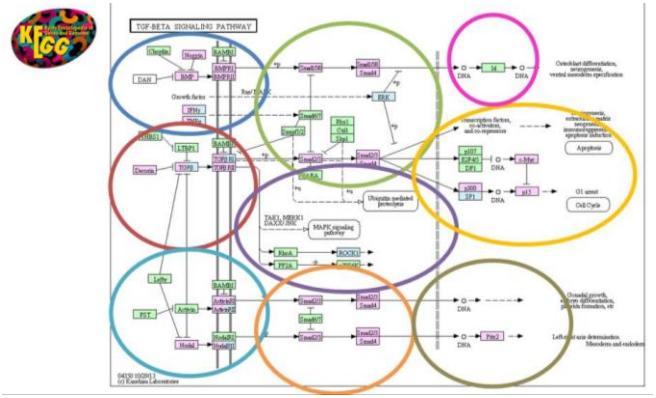
Macro-Algoritmo SeMiC

Macro algoritmo que permite detectar los clusters dentro de una red signaling pathway, y enriquecerlos con GO.

1. Recibir como entrada una **signaling pathway network**
2. Llevarla a un **formato de red** (las proteínas serán tratados como nodos y las reacciones como relaciones)
3. Calcular la **modularidad** para cada nodo en la red
4. Realizar el **cluster jerárquico**,
5. Calcular **los centroides** de cada cluster, usando técnicas de centralidad de redes.
6. **Enriquecer** cada centroide **semánticamente c**on GO

Macro-Algoritmo SeMiC

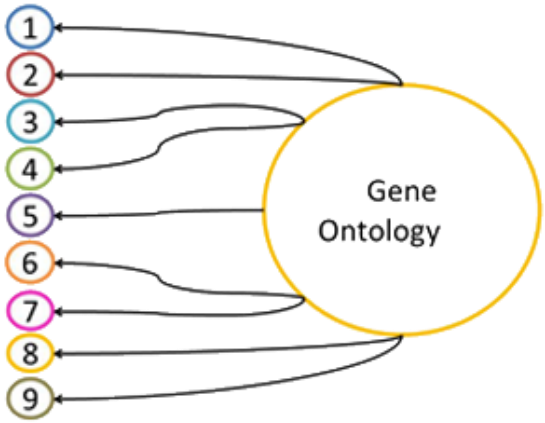




Ejemplo la enciclopedia de Genes y Genomas TGF-β

Después se calculan los clústeres **(paso 4)**. A continuación se extraen los centroides **(paso 5)**

llevar la red, la cual puede ser recibida, por ejemplo, en formato OWL, a un formato de red tradicional para poder ser analizada **(paso 2)**: (NET, DOT y CSV). Seguidamente se cálcula la modularidad de los nodos **paso 3)**.



Los centroides pasan a un enriquecimiento semántico **(paso 6)** usando Gene Ontology (GO)

herramienta usada: Gephi (para análisis de redes sociales

Minería de la Web Semántica (MWS)

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

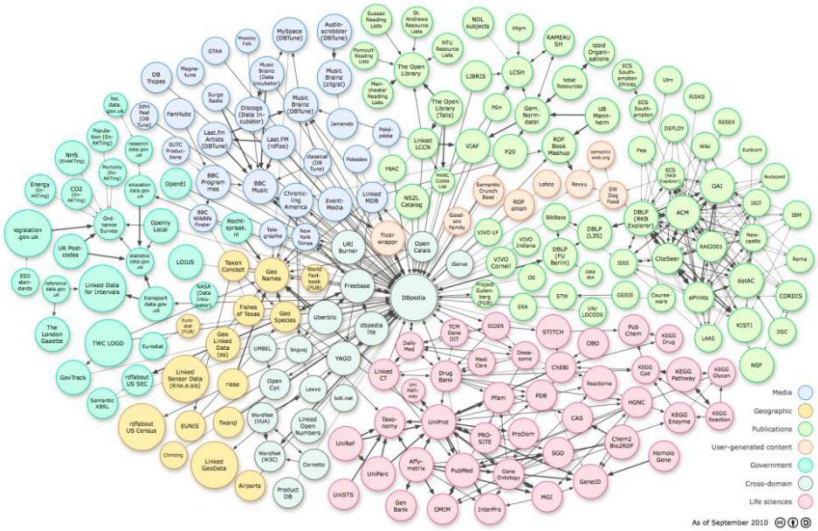
La **Web Semántica** es usada para darle significado a los datos que se encuentran en la Web.

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 en la web codificado en ontologías de dominio



Tipos de recursos semánticos

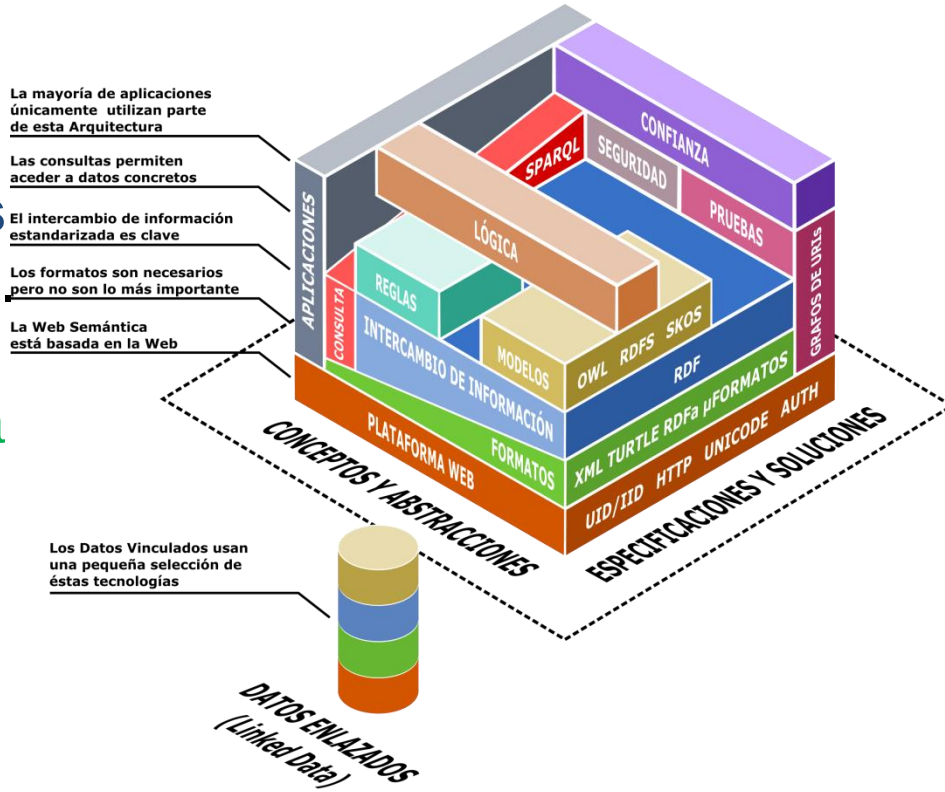
- Ontologías de Dominio
- Ontologías en la web semántica
- …

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

Minería del contenido de la Web

Es el **descubrimiento de información** útil desde los **contenidos textuales y gráficos** de los documentos Web, y tiene sus orígenes en el procesamiento del lenguaje natural y en la recuperación de la información.

Minería de la estructura de la Web

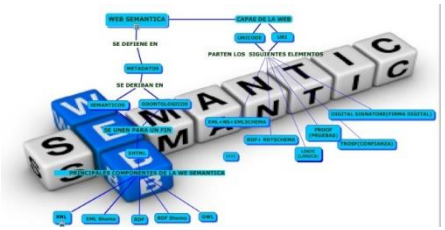
Es el proceso de **descubrir el modelo** subyacente a la **estructura de enlaces de la Web** y analiza, fundamentalmente, la **topología de los hipervínculos** (con o sin descripción de los enlaces)

Minería del uso de la Web

Es la aplicación de técnicas de minería de datos para descubrir **patrones de acceso** (o hábitos) **a los sitios Web**.

Minería Web

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



Productos:

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

Web mining

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.

Minería de Texto

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

Minería de Grafos

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 de Usuario

Minería de Texto

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.

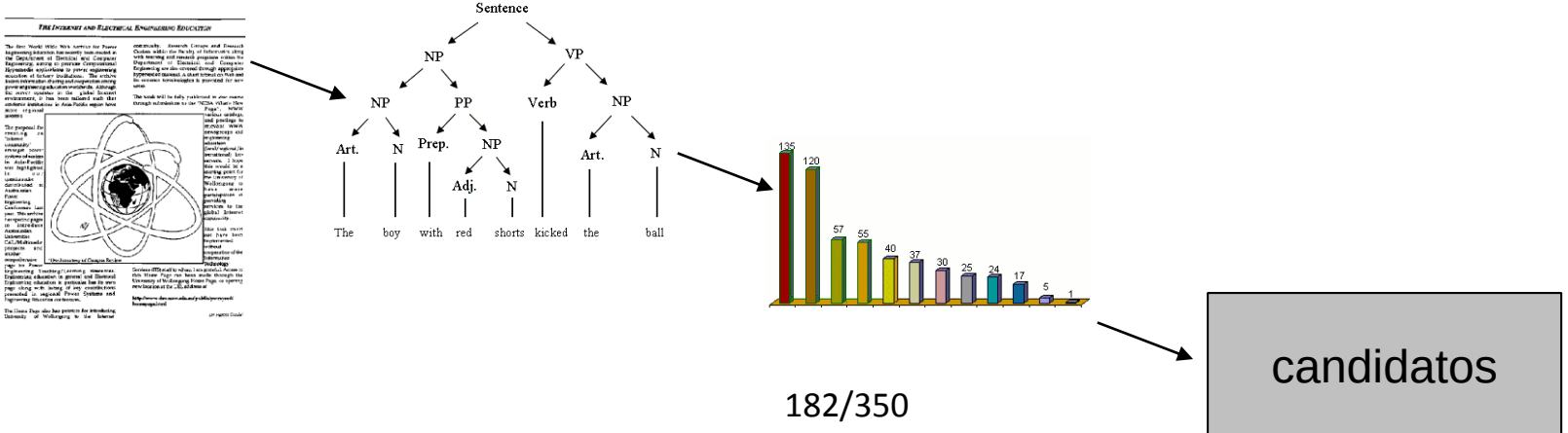
Hypertexto

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

Minería de Textos

Consiste en **generar conocimiento nuevo**, a partir de grandes cantidades de **texto**, el cual no está literalmente escrito en los documentos

- Desarrollo y explotación de corpus lingüísticos.
- Reconocimiento de patrones lingüísticos.
- Caracterización de recursos lingüístico estadísticamente.
- …



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/conocimiento**. En la web semántica se basa en la extracción de metadatos

Minería de Texto

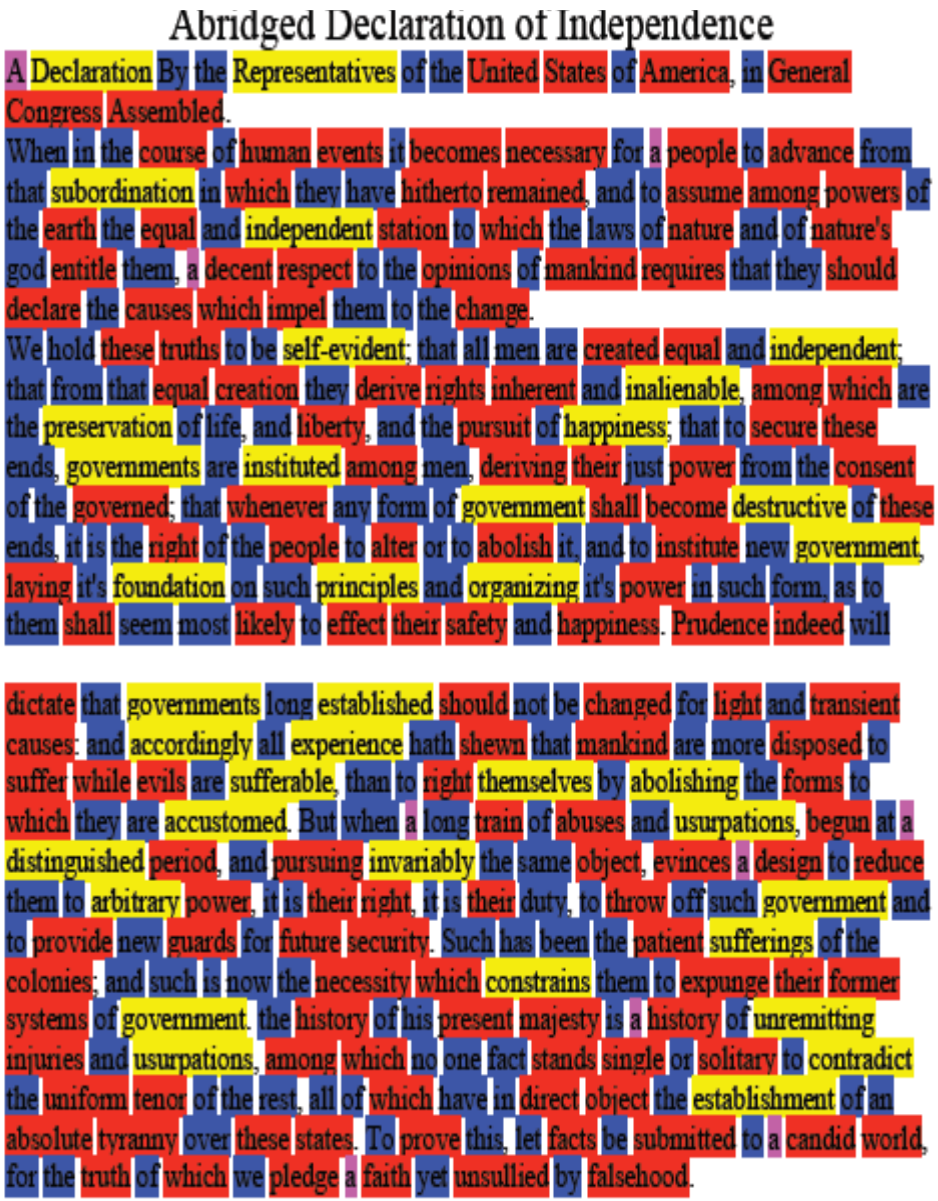
Abridged Declaration of Independence

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 grandes, pequeñas y medianas están en el texto?

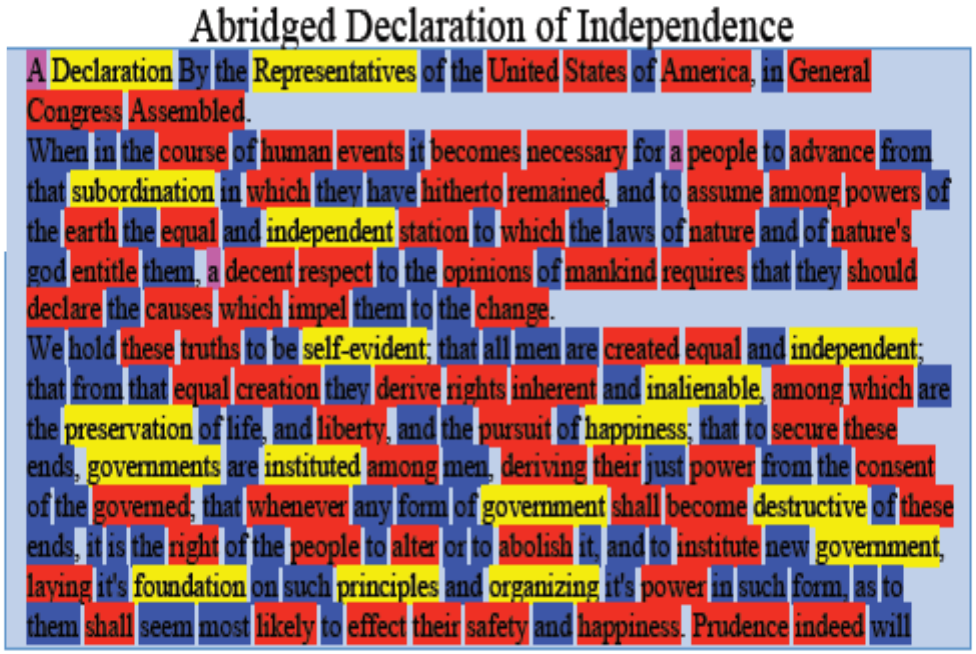
Histograma de longitud de palabras

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



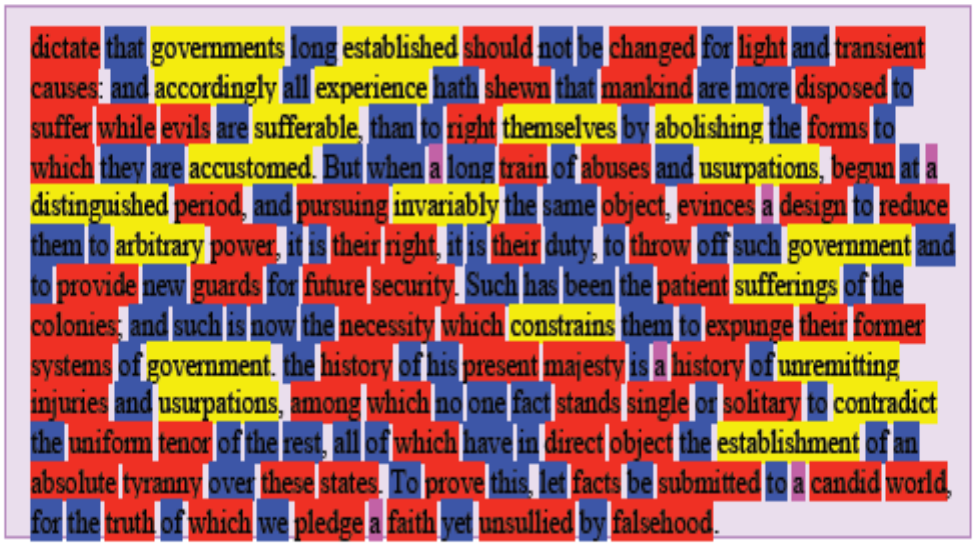
Histograma de longitud de palabras

Mapa 1
204 palabras



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

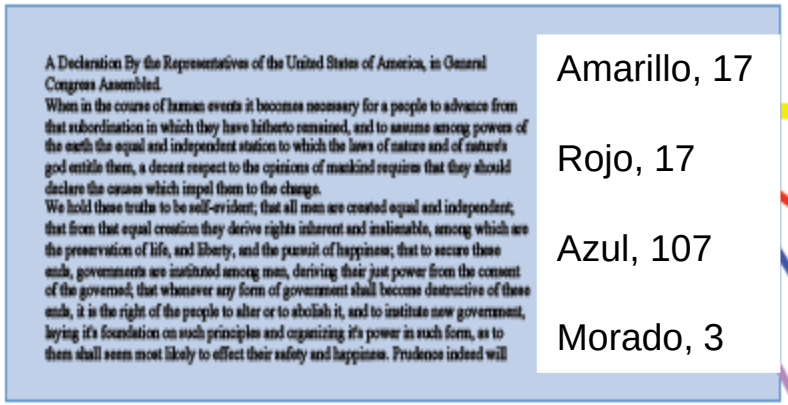
Mapa 2
190 palabras



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

Histográma de longitud de palabras

Mapa 1



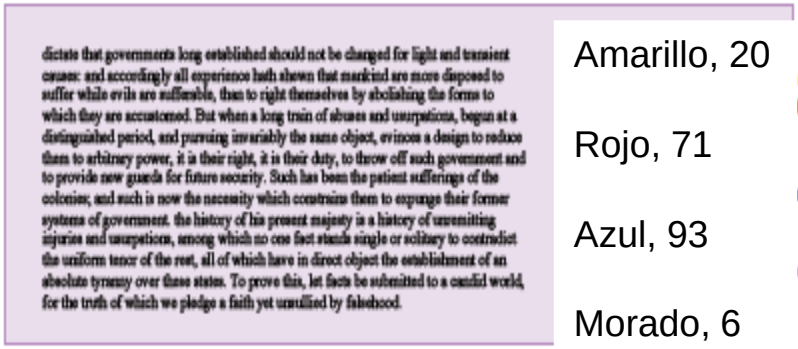
Amarillo, 17

Rojo, 17

Azul, 107

Morado, 3

Mapa 2



Amarillo, 20

Rojo, 71

Azul, 93

Morado, 6

Combinar

- Amarillo, 37
- Rojo, 88
- Azul, 200
- Morado, 9

Minería de Texto

Construir ontologías

Dato estructurado

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

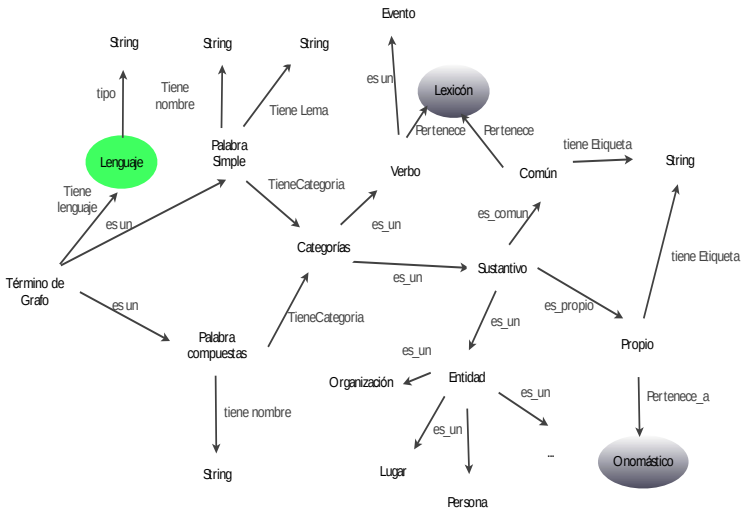
Multimedia

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>...



Resúmenes desde un grupo de documentos

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.

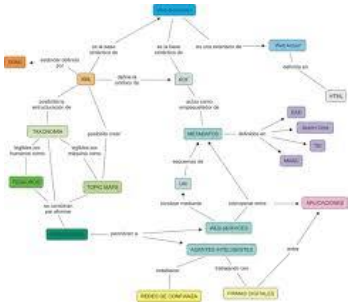
1. Extracción de patrones de conocimiento,
2. Construir o enriquecer ontologías.
3. Establecer relaciones entre ontologías
4. …

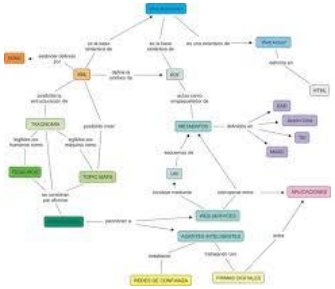
Minería Ontológica

- **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

Ontologías





Grupo inicial

documentos

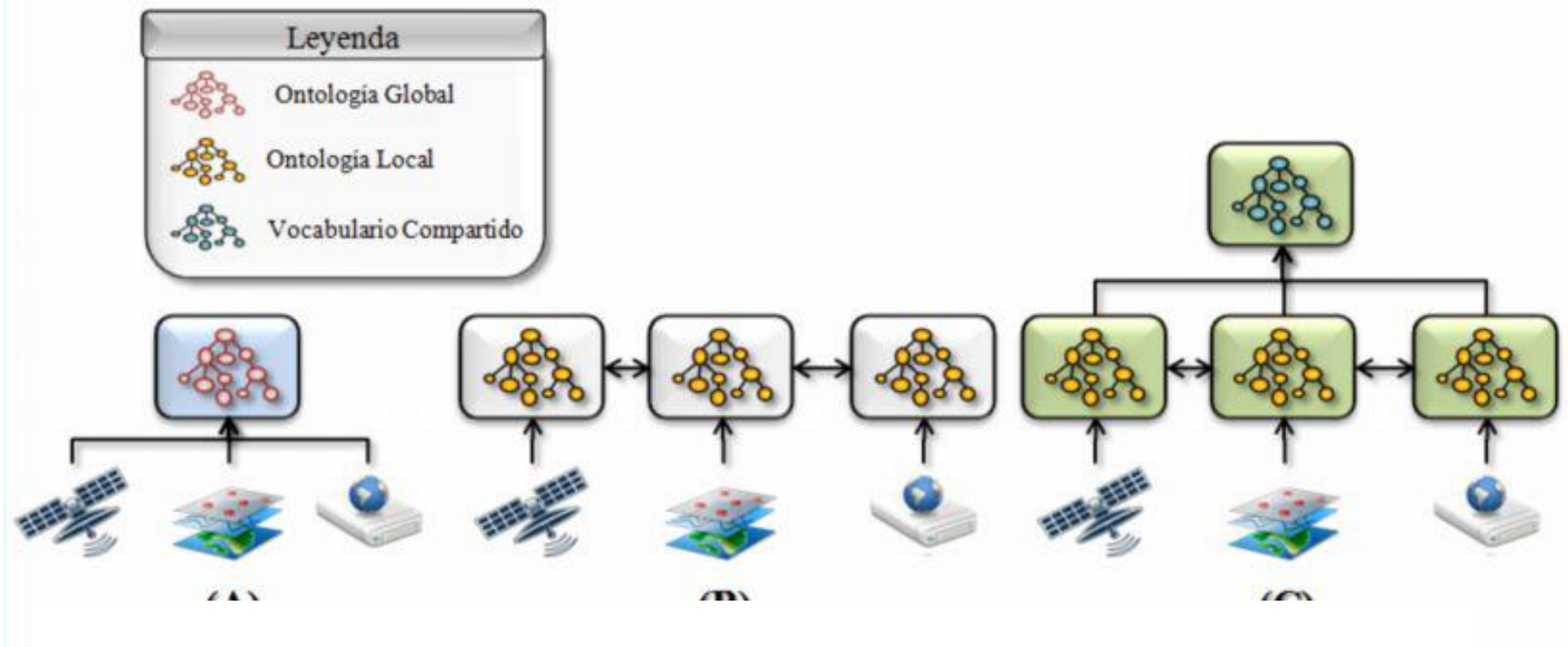
aprendizaje

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

Métodos y herramientas para la alineación de ontologías

Alineación de ontologías

Esta compuesto por los siguientes elementos:

alineación

- dos ontologías O1 y O2,
- un conjunto p de parámetros,
- un conjunto r de recursos para la alineación, y
- una función f de alineación, que retorna un conjunto de correspondencias A'

La función f integra diversos recursos para encontrar correspondencias entre dos conceptos.

En cada O1 y O2 se analizan parte de sus elementos como: conceptos, propiedades de conceptos y jerarquía de conceptos.

- El conjunto p representa los requisitos para realizar la alineación; p = {lenguaje de diseño OWL, número de elementos, vocabulario del idioma, no inferencias}.
- El conjunto de recursos se refiere a los elementos empleados para obtener el conjunto de correspondencias r = {conjunto medidas de similitud, algoritmo AdaBoost, algoritmo de clasificación K-Vecinos}.
- El conjunto A' simboliza todas las correspondencias semánticas.

Alineación de ontologías

Técnicas de alineación de ontologías

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

Detección de correspondencias semánticas

Las medidas de similitud para el proceso de alineación se dividen en dos grupos:

Similitud en base a términos: Se enfoca en el nombre de las entidades en las ontologías, principalmente en el nombre de las clases.

Similitud semántica: Su alcance va más allá de los nombres de las entidades, se enfoca en los componentes que definen la semántica de una clase:

- **Similitud entre propiedades de clases**: Considera las coincidencias existentes entre las propiedades de dos clases.
- **Similitud entre superclases**: Se refiere al par de superclases con mayor similitud respecto a dos clases comparadas.

Métodos para calcular conceptos cercano

Suponen que CA sea un concepto o nodo en la ontología A y PA su predecesor. COM busca encontrar el concepto más parecido CB a CA en la ontología B, y PB (predecesor del concepto CB) a PA que aún no se ha encontrado.

Cuatro casos para calcular la similitud:

Caso A: El concepto CA coincide con CB en B y los predecesores PA y PB

Caso B: PA coincide con PB, pero no hay coincidencia entre CA y CB.

Caso C: CA coincide con CB, pero no hay ninguna coincidencia entre el PA y PB.

Caso D: CA no coincide con el CB y PA no coincide con PB.

Similitud léxica

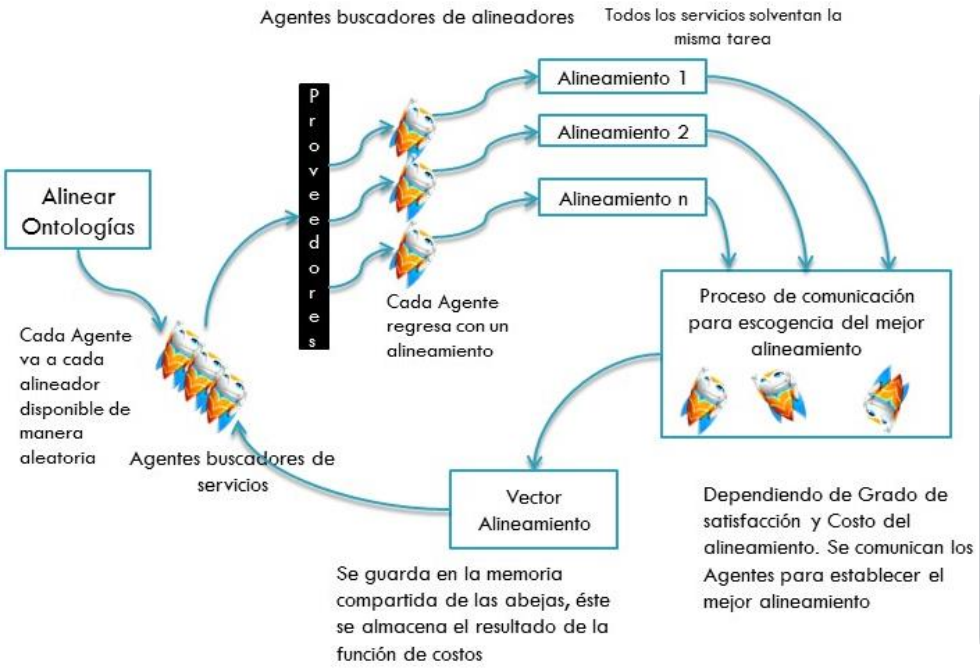
- La Distancia de Levenshtein o distancia de edición (edit distance), fue creada en 1965 por el científico ruso Vladimir Levenshtein.

- La idea consiste en **determinar el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra**,

- Estas operaciones son: **inserción, eliminación o sustitución de un carácter**.

Por ejemplo, la distancia de Levenshtein entre los términos "hotel" y "hostal" es de dos, porque se necesitan al menos dos operaciones elementales para cambiar un término en el otro término.

Sistema de recomendación de Alineamiento de Ontologías usando ABC



- Los agentes abejas tienen la tarea de escoger una técnica de alineamiento
- Las abejas recolectoras se dirigen cada una a una fuente de alimento
- Cada fuente de alimento es una técnica de alineamiento
- Cada abeja recolectora regresa con la información de la calidad del nectar
- La calidad del néctar es transmitida a las abejas en espera
- Estas abejas deciden a que técnica de alineamiento ir basándose en la calidad
- Los resultados de la calidad de los alineamientos se van guardando en un vector que es la memoria global y compartida de las abejas.
- Todo este proceso se repite hasta que se alcancen los objetivos o se llegue a una condición de parada.

Sistema de recomendación de Alineamiento de Ontologías usando ABC

La ganancia $G(S_i)$ es calculada de la siguiente manera:

$$G(S_i) = \frac{S_a(S_i)}{CA(S_i)} \times P_c$$

- **S_i:** Servicio que se puede realizar para solventar una actividad solicitada, en nuestro caso corresponde a **una técnica de alineamiento**, y es equivalente a una fuente de néctar en el algoritmo ABC.
- **$G(S_i)$:** Ganancia obtenida por el servicio S_i (técnica de alineamiento *i*), y es equivalente a la **calidad del néctar** en el caso del algoritmo ABC.
- **$Sa(S_i)$:** **Número de nodos alineados** por la técnica de alineación S_i. Es usado para calcular $G(S_i)$ que es la calidad del néctar.
- **$CA(S_i)$:** **Tiempo** que tarda una abeja en ir al **Servicio S_i** y regresar con resultados (en nuestro caso, este es el tiempo de cálculo empleado por la técnica de alineamiento *i*, que repercute también con la calidad del néctar $G(S_i)$).
- **Pc:** Probabilidad de conservar la opinión. Valor **pseudo-aleatorio**, con una **distribución normal**, dentro del rango de 0 y 1, el cual modifica el valor de $G(S_i)$.

Enlazado de Ontologías

Es el proceso para encontrar relaciones entre entidades que pertenecen a diferentes ontologías.

Enlazado débil de Ontologías: es una **correspondencia entre conceptos idénticos**. En este caso, básicamente lo que se realiza es la **intersección de las ontologías**, a partir de la cual se podrían hacer inferencias específicas en cada ontología.

Enlazado Fuerte de Ontologías: Es realizado de **manera semiautomático**, con la ayuda de un **experto del conocimiento** global que se está enlazando, el cual puede definir nuevos conceptos, así como **enlaces que relacionan conceptos** de ontologías distintas, creando así una **Meta-Ontología** con partes de conocimiento de las ontologías enlazadas.

Mezclado de ontologías

Es el proceso donde varias ontologías dentro de un mismo dominio se unen para estandarizar el conocimiento, hacer crecer el conocimiento y tener el conocimiento total de manera local.

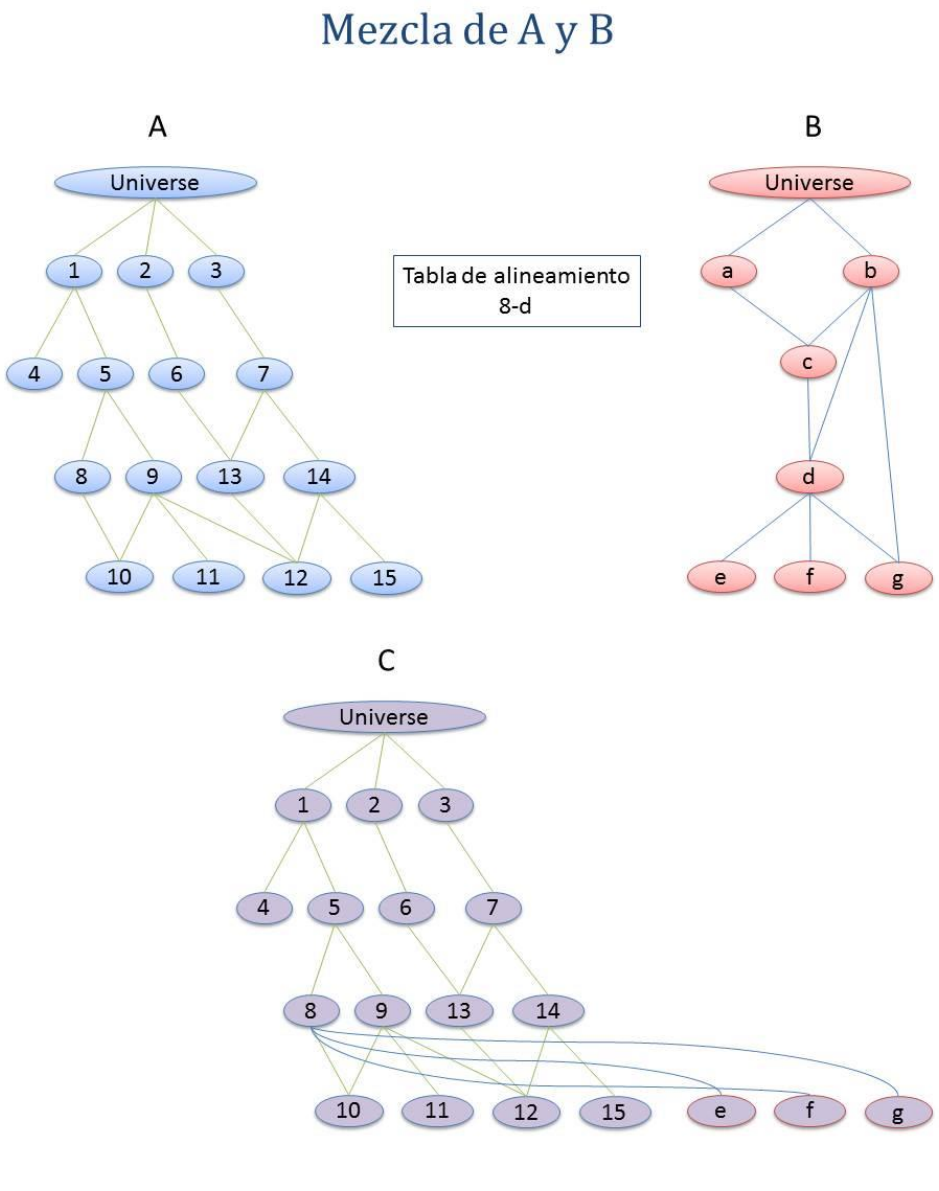
Los mezcladores unen ontologías que manejan el **mismo conocimiento**, pero con diferente representaciones, o que poseen representaciones parciales de dicho conocimiento, tal que las ontologías pueden coincidir en ciertos conceptos y en otros no.

Mezcla Débil de Ontologías: se toma una ontología A, la copian como resultado C, y la van enriqueciendo con la otra B, comparando todos los conceptos de la ontología C (que son los mismos de A en este momento) con los de la ontología B, enriqueciendo los conceptos de C con sus conceptos semejantes de B. Dejando por fuera parte del conocimiento de B.

Mezcla Fuerte de Ontologías: Es una mezcla débil, pero incorporándole el conocimiento dejado por fuera de B,

Algoritmos de Mezcla de Ontologías

El problema de **la mezcla tradicional** de ontologías (**mezcla débil**), es que deja conocimiento sin ser incorporado en la ontología resultante.

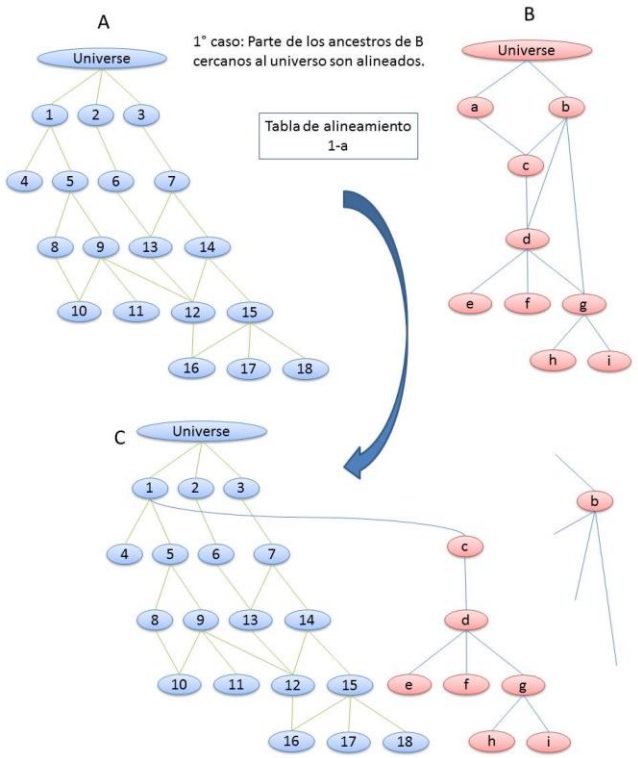


Algoritmos de Mezcla de Ontologías

Mezcla Fuerte se hace en dos partes,

1. Se realiza la mezcla débil,
2. Se incorporan los conceptos y relaciones dejadas por fuera.

Primera Parte: Nuestro sistema realizar la mezcla débil de dos ontologías consistentes A, B en una ontología C



Algoritmos de Mezcla de Ontologías

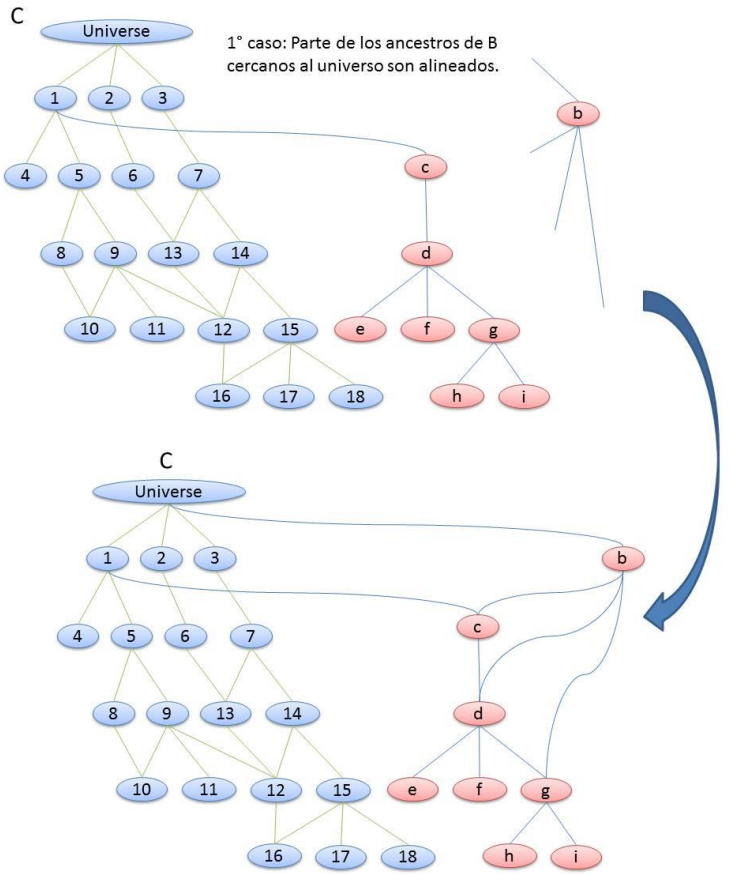
Segunda Parte: si de la ontología a la cual se le está extrayendo el conocimiento para ser agregado a la primera **queda aún conocimiento sin ser agregado**, se analizan los siguientes casos:

Caso 1:

- Se alinearon **parcialmente** los conceptos de B
- los nodos no alineados no se copian en la ontología resultado.

Solo bastaría con:

- **agregar los nodos no alineados a C**
- **copiar las relaciones** que no fueron copiadas o alineadas.



Algoritmos de Mezcla de Ontologías

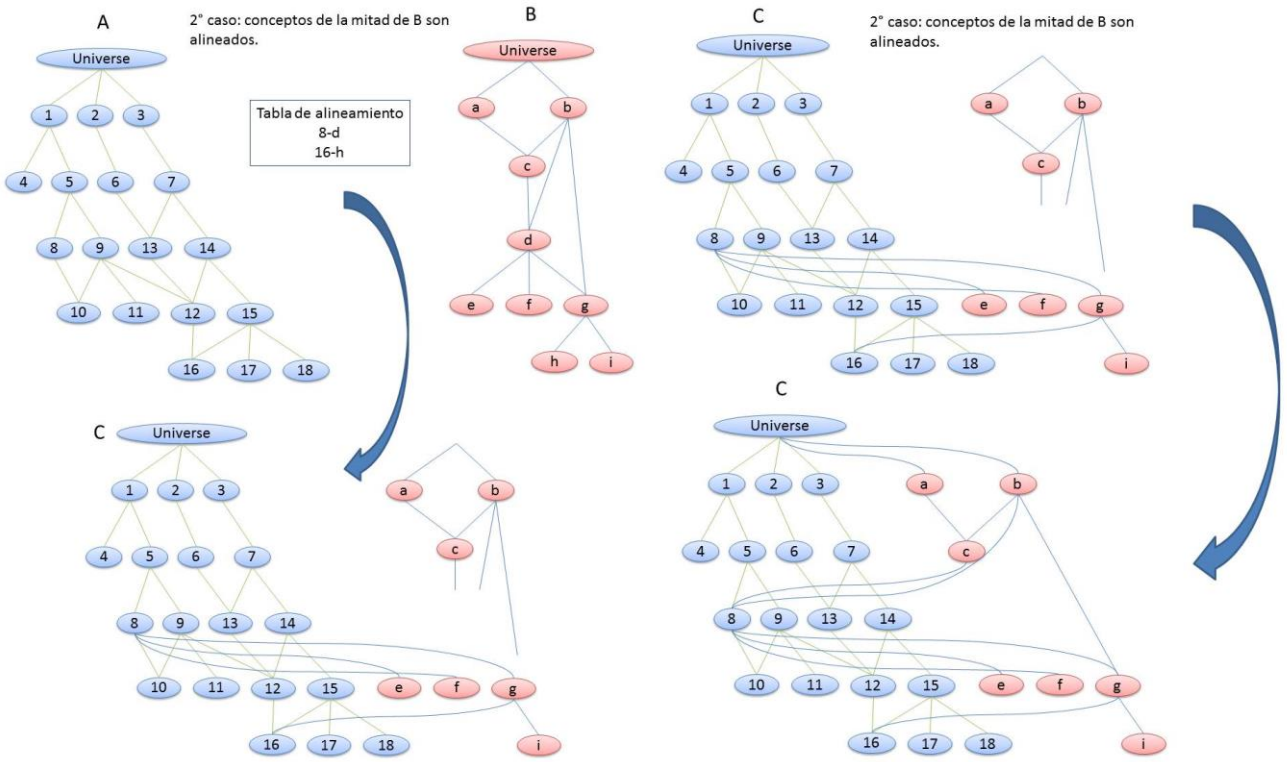
Caso 2:

- Se **alinearon** ciertos **nodos intermedios**
- Se deja a los ancestros sin ser copiados

La mezcla fuerte debe:

- agregar estos conceptos al universo de C
- Agregar las relaciones donde ellos participan

(como ocurre con **b-g ,c-8 y b-8**)



Algoritmos de Mezcla de Ontologías

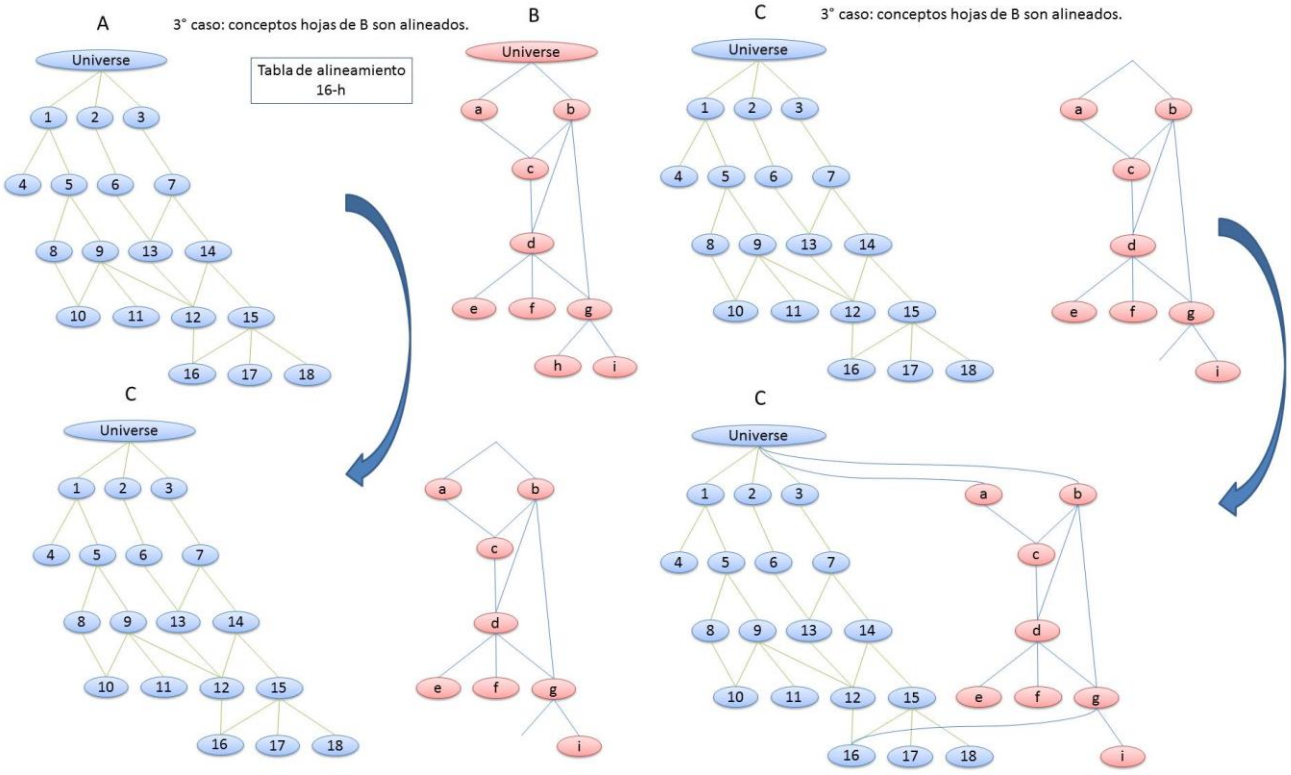
Caso 3:

- Solo los **nodos hojas se alinearon**
- Se deja por fuera de C un conjunto de conocimiento grande de B.

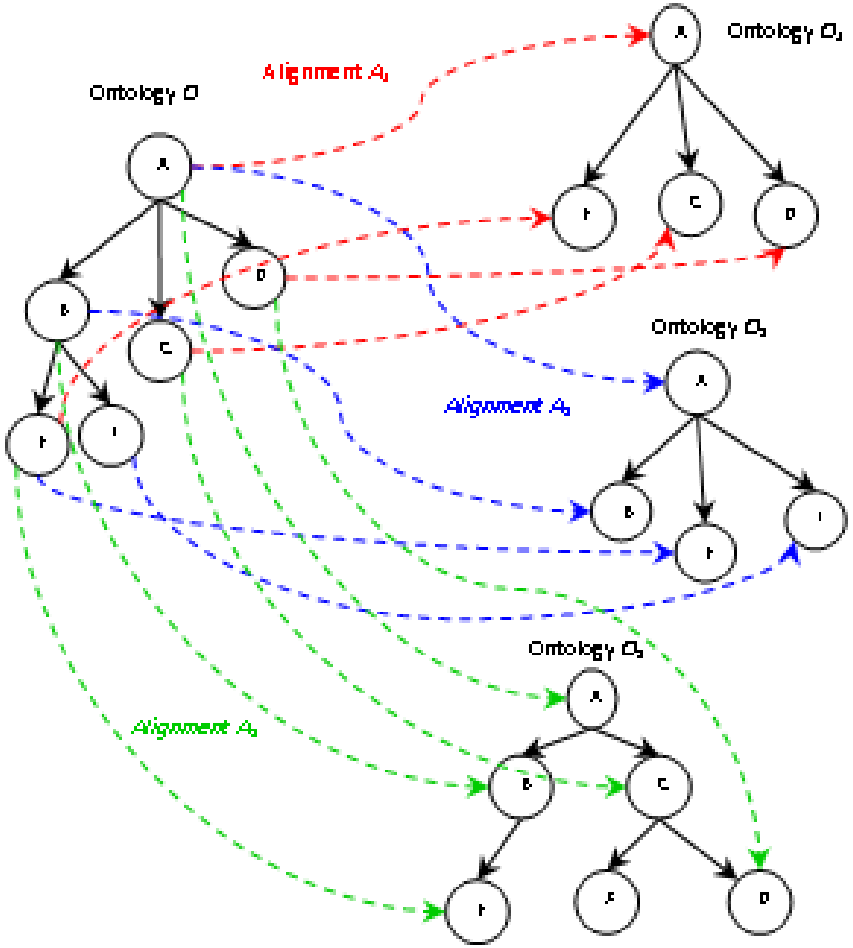
La mezcla fuerte debe

- **agregar** todos los conceptos **no copiados a C**
- buscar las relaciones que estos conceptos ya tenían con otros en B
- copiar también esas relaciones con los conceptos que fueron copiados o con los que fueron alineados

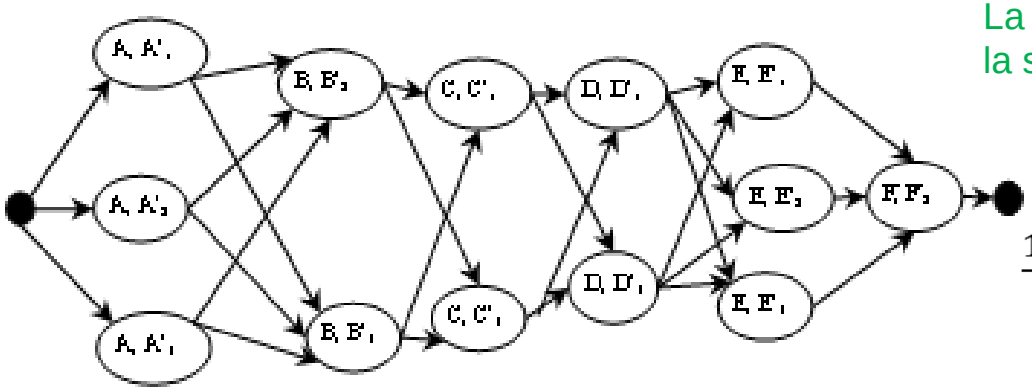
(g-16)



Un enfoque para la combinación múltiple de ontologías



Un enfoque para la combinación múltiple de ontologías



La similitud de los conceptos C y C 'será proporcional a la similitud de los conceptos ancestrales

$$SA(C,C') = PC_A \times Sim(C,C') + \frac{2(1-PC_A)}{n(n+1)} \sum_{j=1}^{m}\sum_{i=1}^{n}(n+1-i)Sim(Anc_i(C),Anc_j(C'))$$

La similitud de los conceptos C y C será proporcional a la similitud con los hermanos

$$SS(C,C') = PC_S \times Sim(C,C') + \frac{1-PC_H}{n}\sum_{i=1}^{n}\max(Sim(S_i,S_1'),\dots,Sim$$

La similitud entre los dos conceptos C y C 'también será proporcional a la similitud de los descendientes directos

$$SD(C,C') = PC_D \times Sim(C,C') + \frac{1-PC_D}{n}\sum_{i=1}^{n}\max(Sim(H_iH_1'),\dots,Sim(H_i,H_n'))$$

Medida de Similitud $$MS(C,C') = \frac{SA(C,C') + SD(C,C') + SS(C,C')}{3}$$

Un enfoque para la combinación múltiple de ontologías

"Grado de enriquecimiento" (GE)

indicador de la cantidad de nuevos conceptos obtenidos por la ontología fuente después de seleccionar una alineación para un concepto.

GE de la ontología después de seleccionar la alineación de un concepto C con C 'de los nuevos conceptos que se pueden añadir a los corresponsales de la ontología:

- Niños de conceptos C 'no alineados (Nueva hipónimos) y sus descendientes
- Los hermanos de los conceptos C 'no alineados con el ancestro alineado inmediata (padre) (Nueva Cohiponímias) y sus descendientes.
- Conceptos antepasados de C 'no alineados (Nueva hiperónimos).

$$\mathrm{GE}(C, C') = CHildren\_Non\_Aligned\,(C') + Siblings\_Non\_Aligned(C') + Ancestors\_Non\_Aligned(C')$$

CRITERIOS DE CALIDA PARA LA FUSIÓN DE ONTOLOGÍA

- **Cobertura (o completitud)**: grado de preservación de la información en la nueva ontología. Sub-casos:
  - cobertura con respecto a las ontologías fuente o objetivo,
  - con respecto a los conceptos retenidos de las ontologías (cobertura general), o
  - cobertura foliar definida como el grado en que se conservan los conceptos de las hojas de entrada.

- **Compacidad:** tamaño relativo de la fusión obtenida. Se define como la suma del número de conceptos de ambas ontologías de entrada menos el número de conceptos alineados.

- **Redundancia:** Las medidas se basan en la siguiente ecuación:

$$LPB = LPS + LPT - ML$$

LPB es el número mínimo de rutas a las hojas, sin redundancia; LPS y LPT es el número de rutas a las hojas, en cada una de las dos ontologías de entrada, respectivamente, y ML es el número de conceptos de hoja que están alineados.

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Minería de Grafos

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Modelando Datos con Grafos…

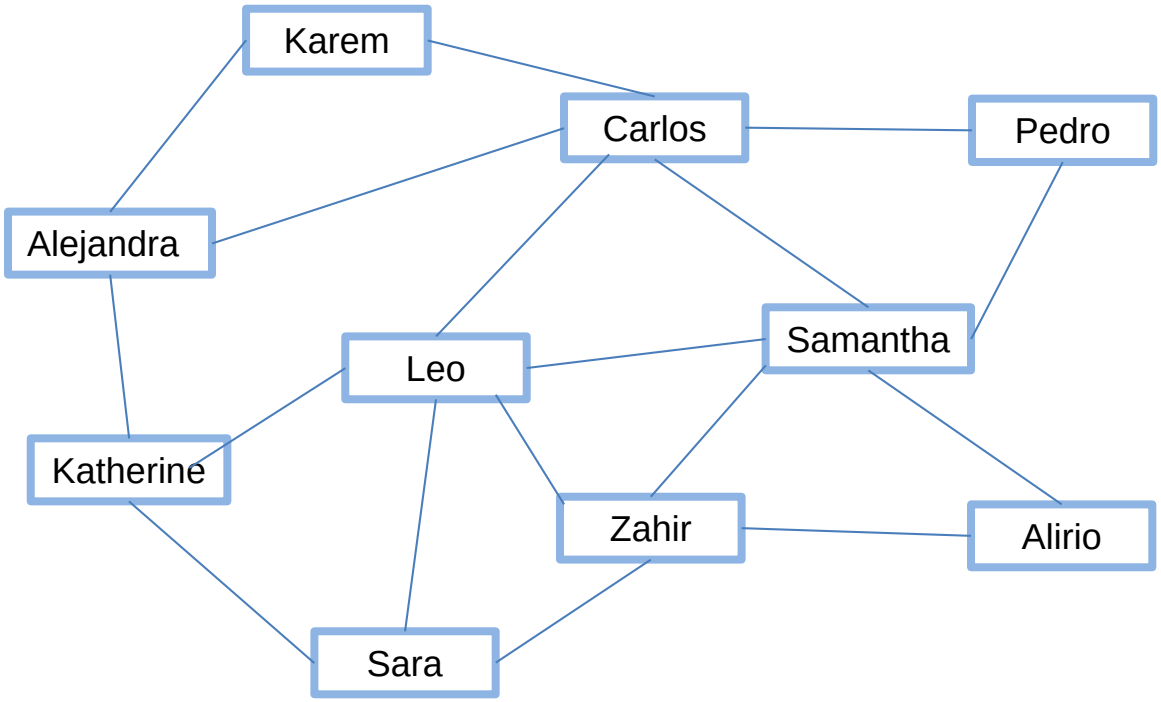
Los grafos son adecuados para la captura de las relaciones arbitrarias entre los diversos elementos.

Instancia		Grafo
Elemento	⟺	Vertice
Atributos Elemento	⟺	Etiquetas Vertices
Relaciones	⟺	Arcos
Tipo de relaciones	⟺	Etiquetas arcos

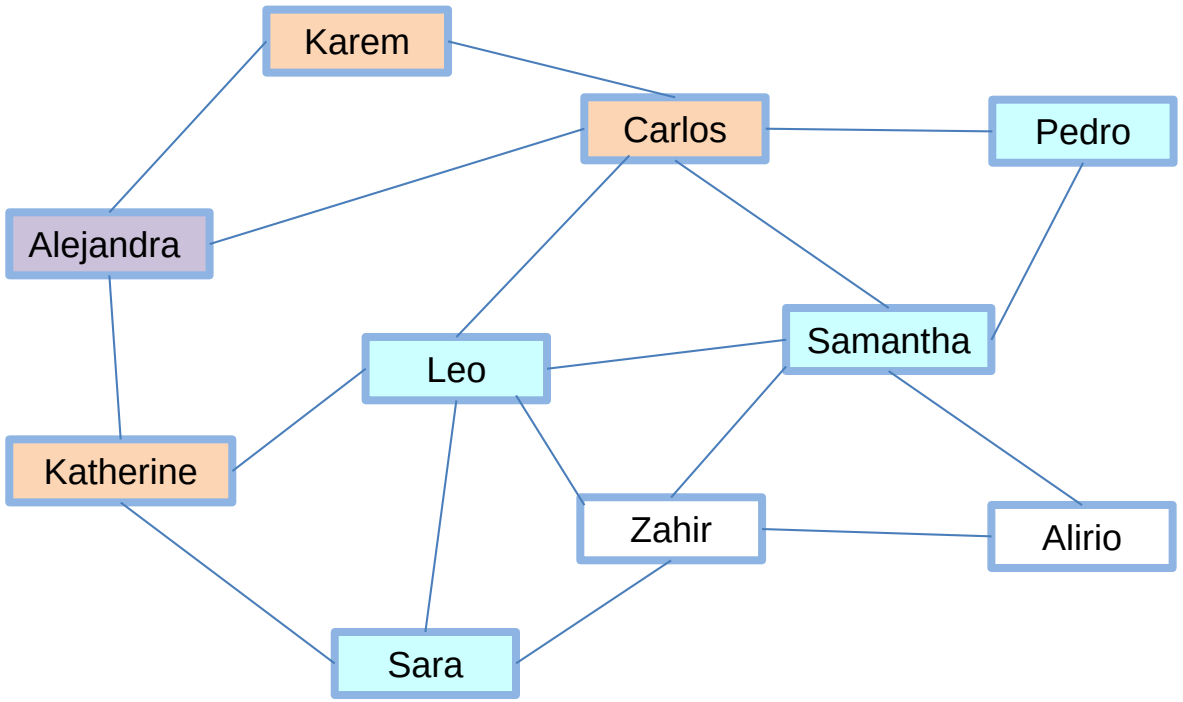
Proporcionan una enorme flexibilidad para el modelado de los datos, ya que permiten al modelador decidir cuáles son el tipo de relaciones a modelar

Grafos

Red Social
FACEBOOK



Grafos



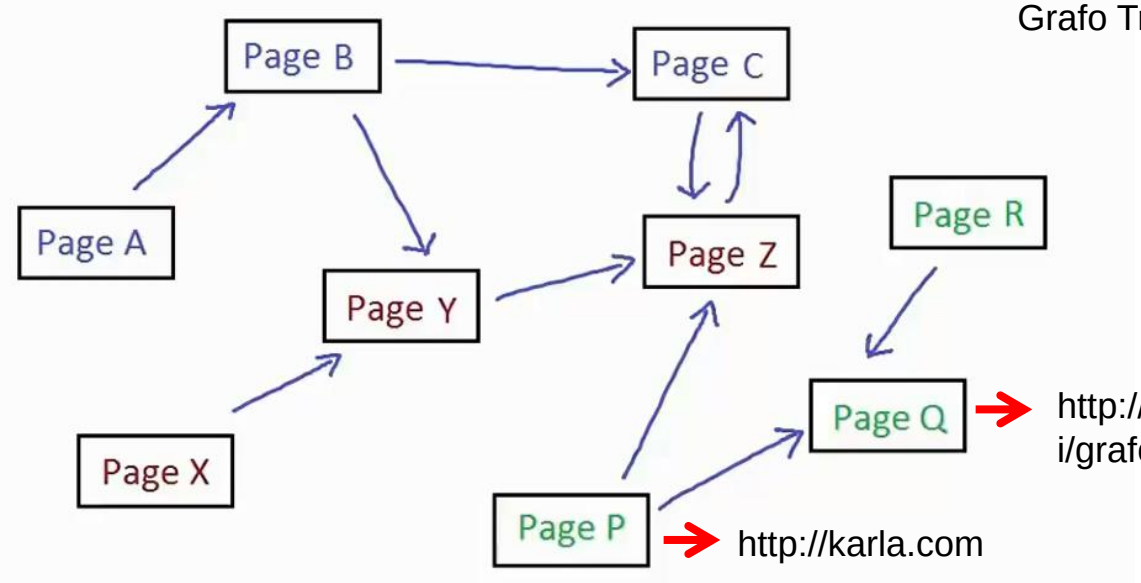
Red Social FACEBOOK

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

World Wide Web

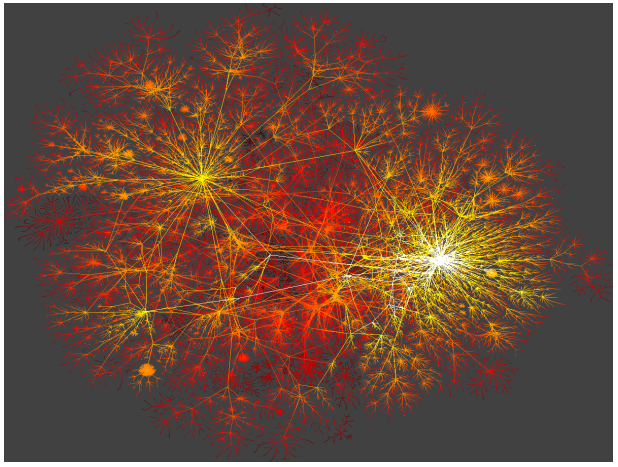
Web – Crawling

Grafo Transversal



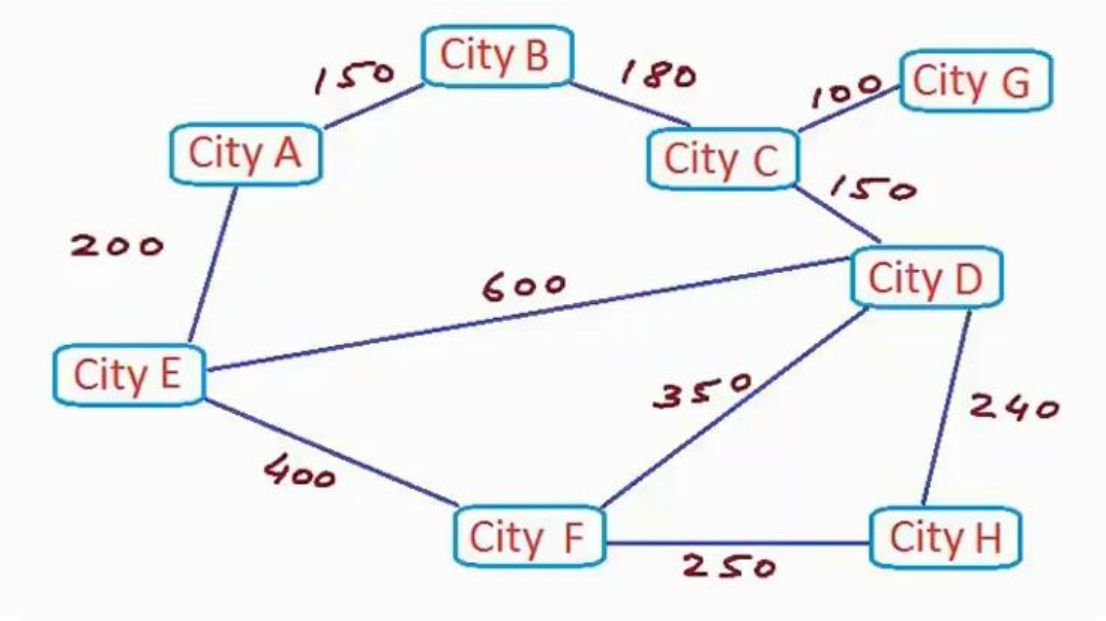
http://wikipedia.org/wiki/grafos

Internet



Grafos

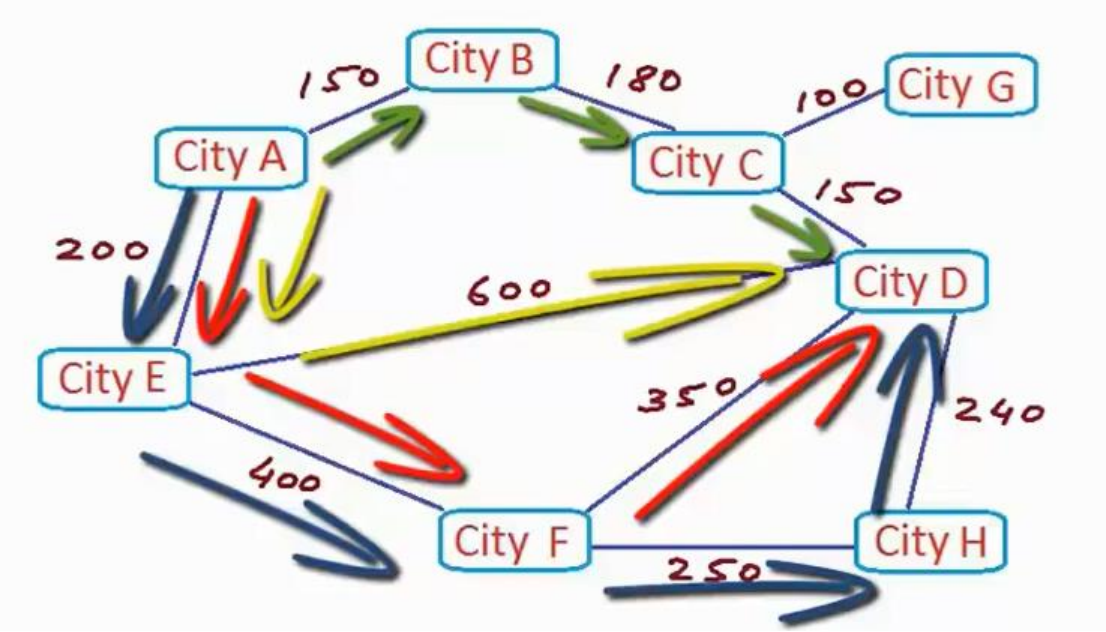
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Inter urbanas

Grafos

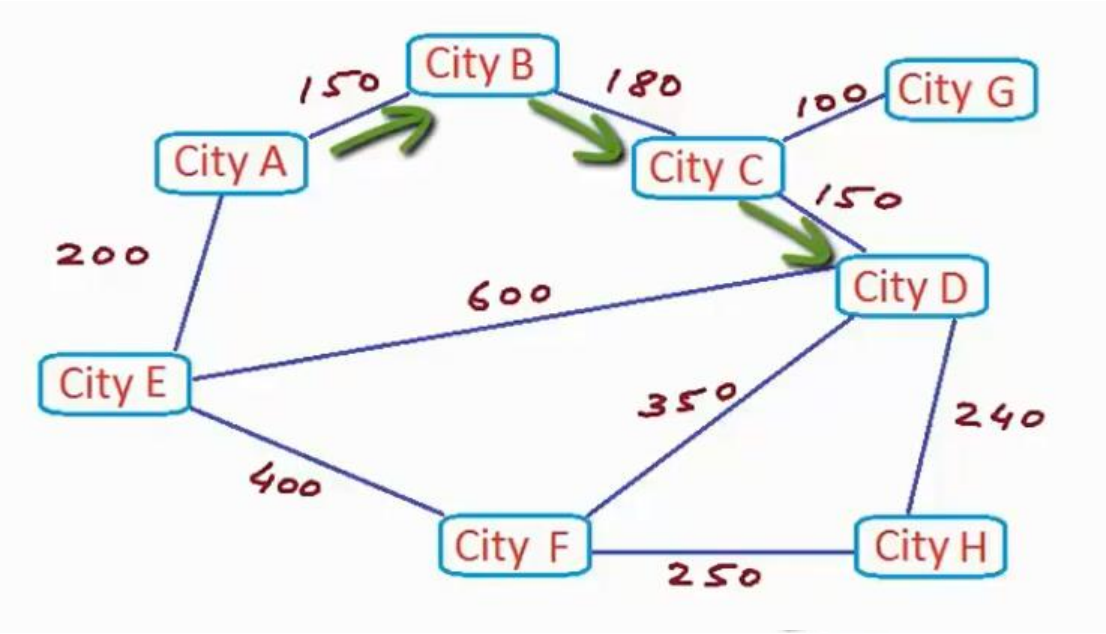
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Inter urbanas

Grafos

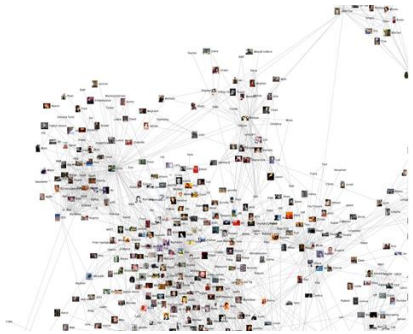
Grafos con Pesos VS Grafos sin Pesos



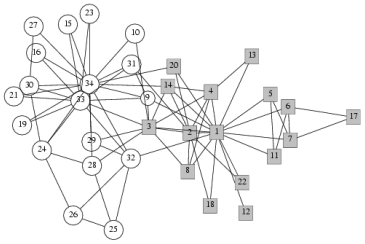
Red de Carreteras Inter urbanas

Redes en el mundo real

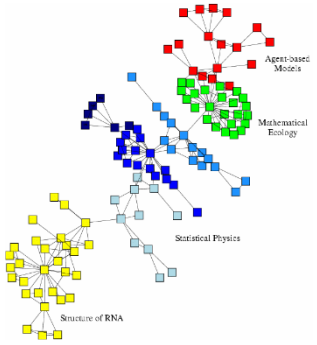
- **Redes de información:**
  - World Wide Web: hyperlinks
  - Redes de citación
  - Redes de Noticias y Blogs
- **Redes sociales**
  - Organizativas
  - Comunicativas
  - Colaborativas
  - Contactos sexuales
- **Redes tecnológicas:**
  - Energéticas
  - Transporte (aéreo, carreteras, fluviales,…)
 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



Redes de amistad



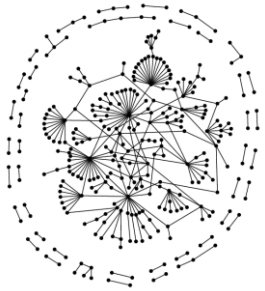
Karate club network



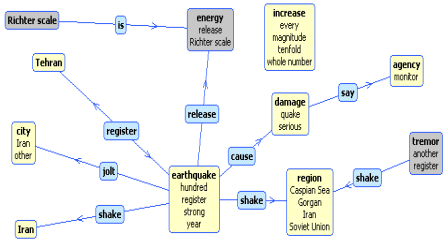
Redes de colaboración

Redes en el mundo real

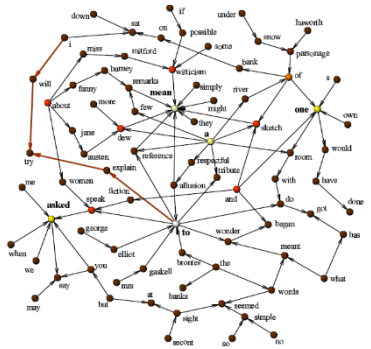
- Redes biológicas
  - Metabólicas
  - Cadenas alimenticias
  - Neuronales
  - Regulación Genética
- Redes de lenguaje
  - Semánticas
  - Lingüísticas
- Redes de software
- …



Interacciones entre las proteínas de la levadura

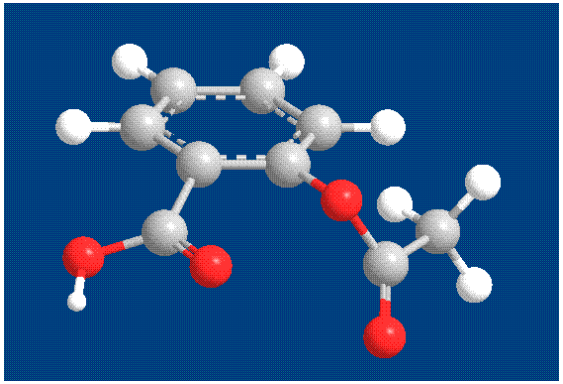


Red semántica

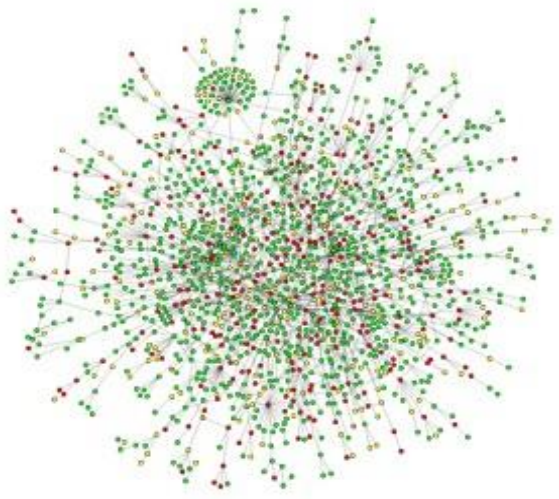


Red Lingüística

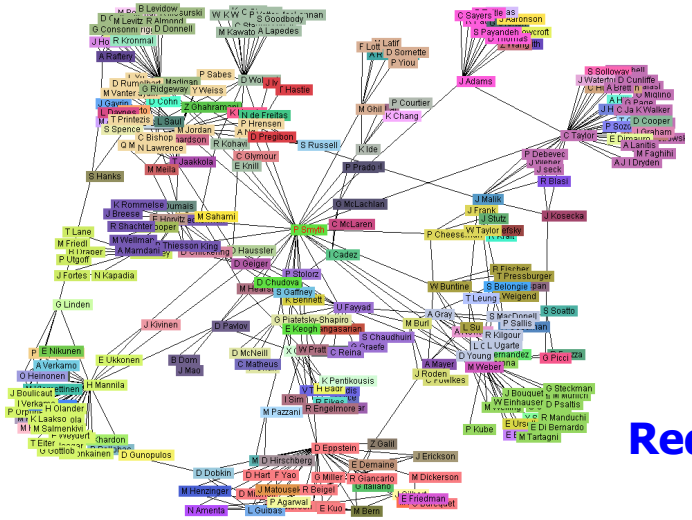
Grafos



Aspirina



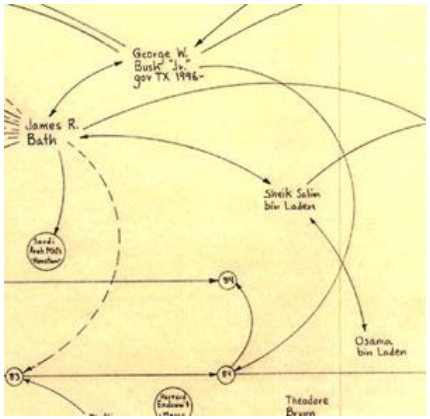
Red de interacción de proteína levadura

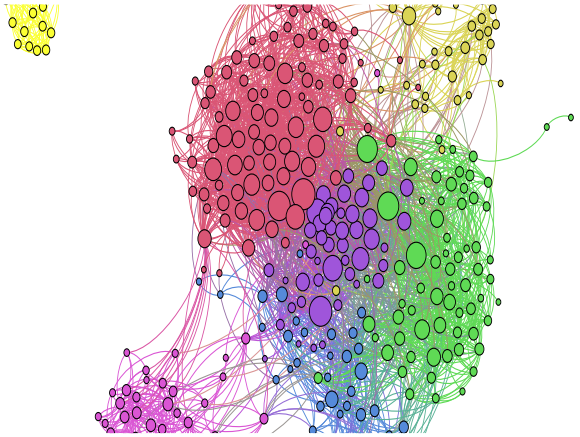


Red Co-autores de libros

Grafos

Mark Lombardi: rastreo y Mapeo fiascos financieros globales en los años 1980 a partir de fuentes públicas, como los artículos de noticias.

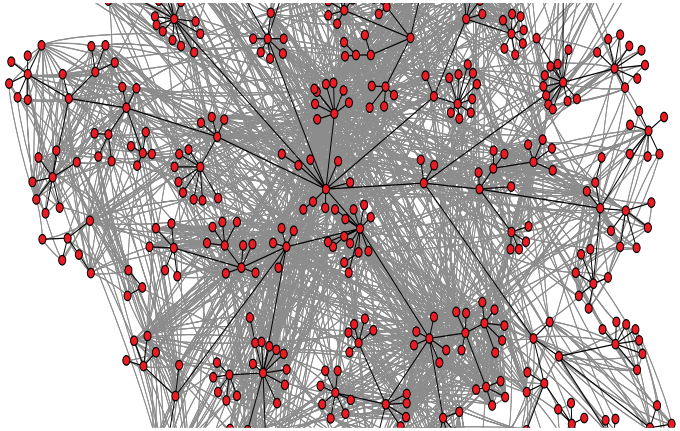




Facebook de alguien

Los colores separan componentes fuertemente conectados de la red.

Relación de empleados de una organización



Los arcos negros denotan estructura organizacional y los grises son interacciones por correo electrónico.

Red Social

- El **análisis de redes sociales** estudia esta estructura social aplicando la teoría de grafos.

- Se analiza:
  - Si existen estructuras de comunidades ocultas
  - La difusión o las opiniones.
  - La influencia del todo en las partes y viceversa.
  - La difusión de nuevas ideas y prácticas (teoría de difusión de innovaciones).
  - El efecto producido por la acción selectiva de los individuos en la red
  - Grafos de colaboración para ilustrar buenas (amistad, alianza, citas) y malas (odio, ira) relaciones entre los seres humanos.

Herramientas

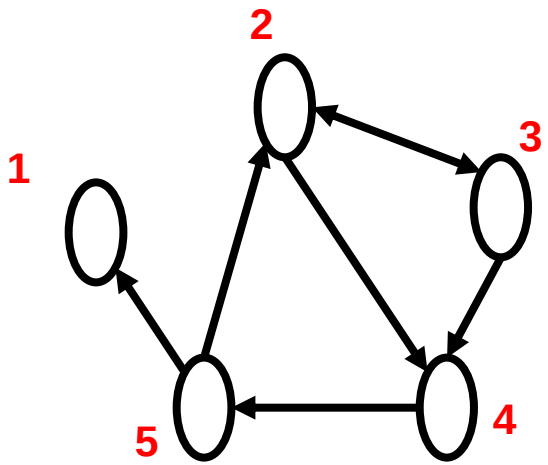
- **Gephi** (visualization and basic network metrics)
- **NetLogo** (modeling network dynamics)
- **Pajek:** amplia funcionalidad basada en menús, incluyendo muchas, muchas métricas de red y manipulaciones
  - pero ... no extensible
- **Guess:** extensibles, herramientas de secuencias de comandos de 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)

Elementos de un grafo

- Dirigido
  - A -> B
    - A le gusta B, A le dio un regalo a B, A es hijo de B
- No dirigido
  - A <-> B o A – B
    - A y B se gustan, son semejantes
  - Peso (frecuencia de comunicación)
  - ranking (mejor amigo, segundo mejor amigo…)
  - tipo (amigo, pariente, co-trabajador)

Representación de los datos

- Matriz de adyacencia

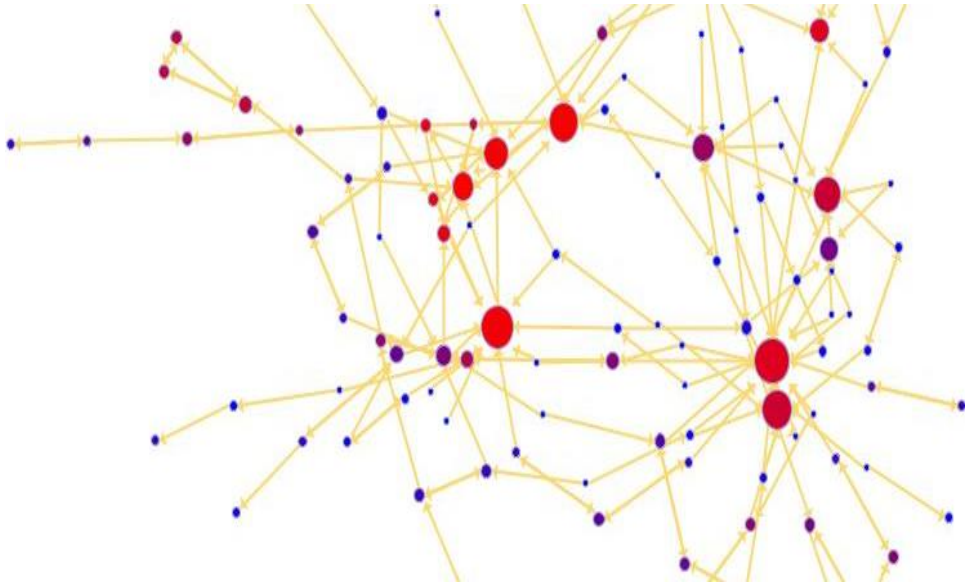


$$A = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{matrix}$$

- Lista de arcos
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1

- Lista de adyacencia
  - Todos los vecinos de cada nodo
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2
  - Más fácil para redes
    - Grandes
    - Dispersas

Métricas

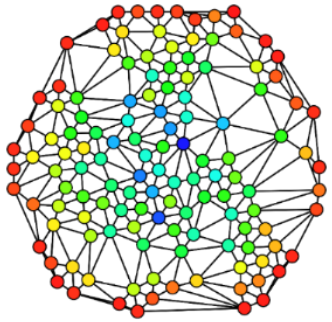


¿Cuál es el nodo con más arcos?

Métricas de redes

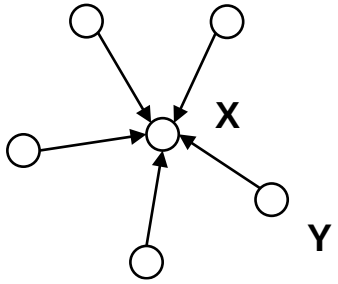
Cada métrica de red da respuesta a las siguientes preguntas:

- **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

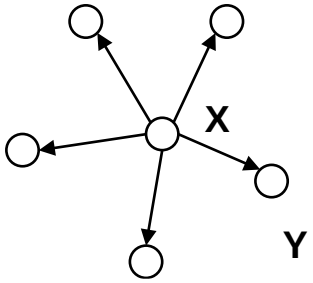


Métricas: Centralidad

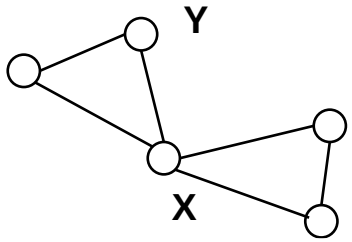
Medidas posibles de un vértice en un grafo, que determina su importancia relativa dentro de éste



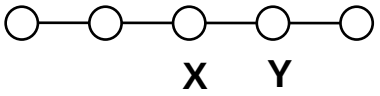
indegree



outdegree



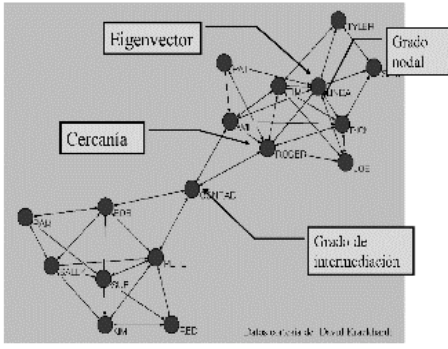
Betweenness (intermediación)



Closeness (cercanía)

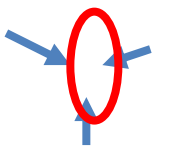
Centralidad de vector propio (eigenvector centrality)

Cuatro Aspectos de la Centralidad



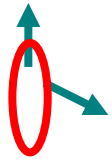
Métricas: Propiedades de los nodos de la Red

- 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



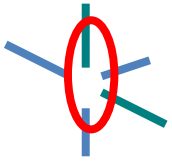
indegree=3

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



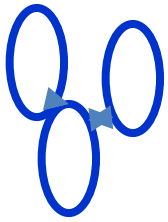
outdegree=2

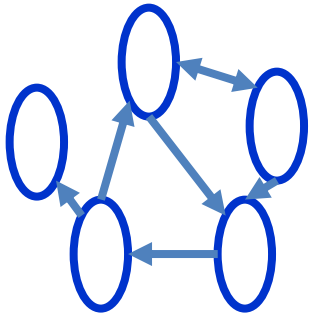
$$\sum_{j=1}^{n} A_{ij}$$



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]

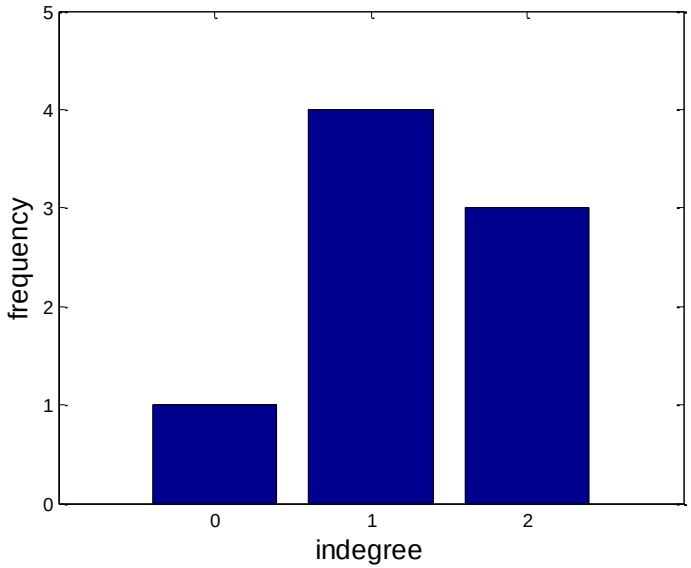




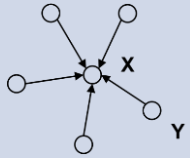
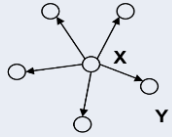
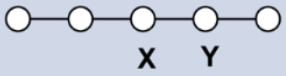
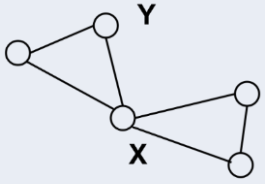
Métricas: Propiedades de los nodos de la Red

- **Degree distribution:** La frecuencia con la que ocurre cada grado

  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]

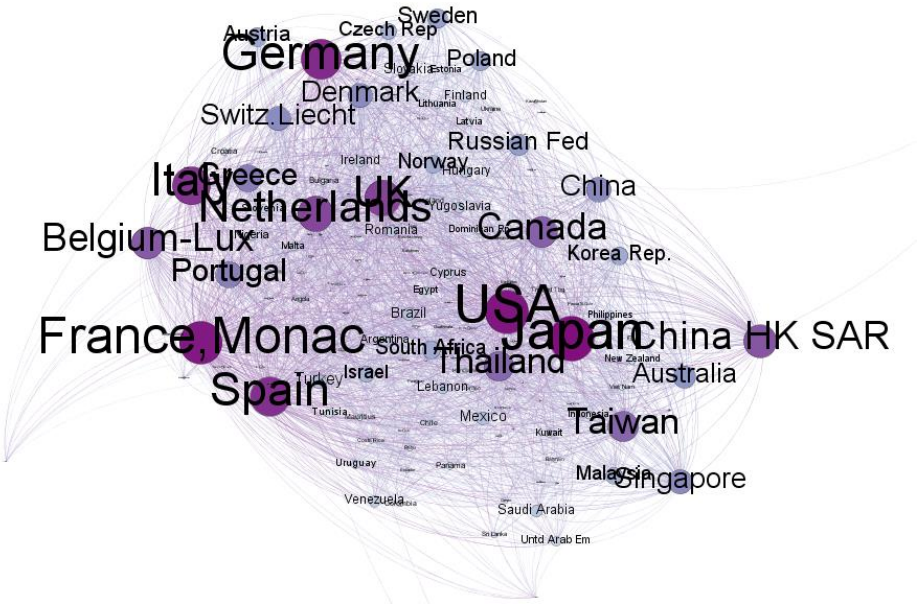


Métricas principales de la centralidad

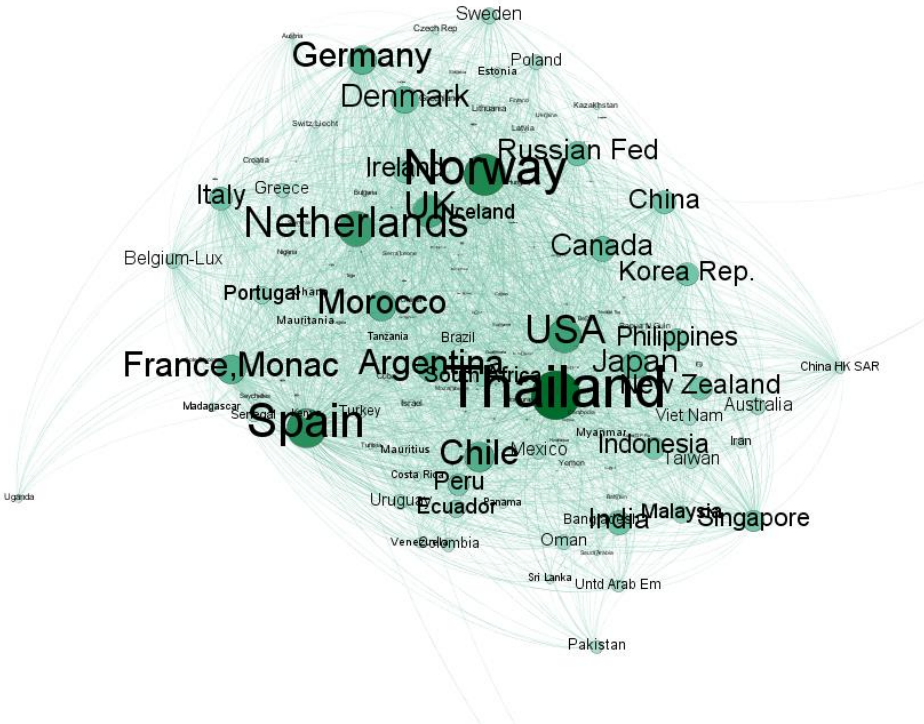
Métricas de centralidad		
a) **Centralidad de grado (*degree centrality*)**: qué tantas conexiones directas tiene una unidad con otras unidades. Una unidad con alta centralidad de grado sirve como "conector" o "hub" de la red.	**Indegree** o grado de entrada	
	outdegree o grado de salida	
b) **Centralidad de cercanía (*closeness centrality*)**: que tan "cerca" se encuentra una unidad de la red de las otras, considerando tanto conexiones directas como indirectas. una unidad con alta centralidad de cercanía puede interactuar fácilmente con otras unidades, tiene la visibilidad del comportamiento de la red en su conjunto, y puede influir en ella.	**closeness** o cercanía	
c) **Centralidad de intermediacion (Betweenness centrality)**: índice de en qué tantas rutas más cortas entre 2 unidades cualesquiera de la red se encuentra una unidad dada. Estas unidades tienen el control del flujo de información dentro de la red.	**betweenness** o intermediación	

Métricas

InDegree



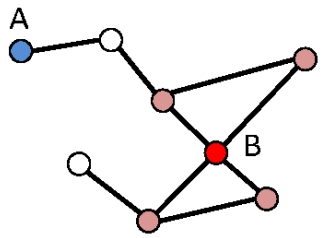
OutDegree



Métricas: Centralidad

MEDIDAS LOCALES DE CENTRALIDAD | **Centralidad de grado**

No dirigida

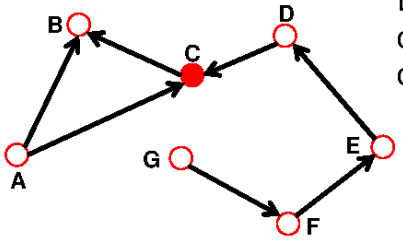


Centralidad de grado de un actor (C_D): número de enlaces que lo conectan con otros

$$C_D(\mathrm{A}) = k_{\mathrm{A}} = 1 \qquad C_D(\mathrm{B}) = k_{\mathrm{B}} = 4$$

C_D(i) se define en {0,g-1}, siendo g el número de nodos de la componente conexa

Dirigida



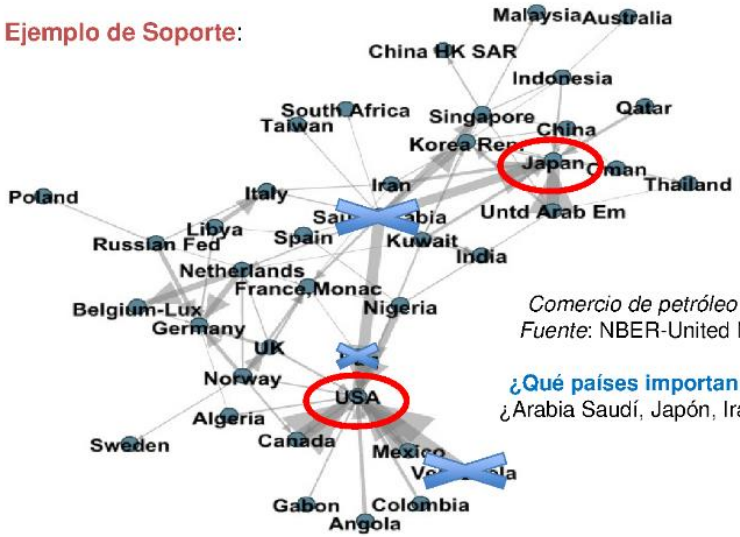
En redes dirigidas, se define el **Prestigio** de entrada (*in-degree*), denominado **Soporte**, y el **Prestigio** de salida (*out-degree*), denominado **Influencia**:

$$P_D^{in}(\mathrm{C}) = k_C^{in} = 2 \qquad P_D^{out}(\mathrm{C}) = k_C^{out} = 1$$

Ambos se definen en {0,g-1}

MEDIDAS LOCALES DE CENTRALIDAD | Soporte

Ejemplo de Soporte:

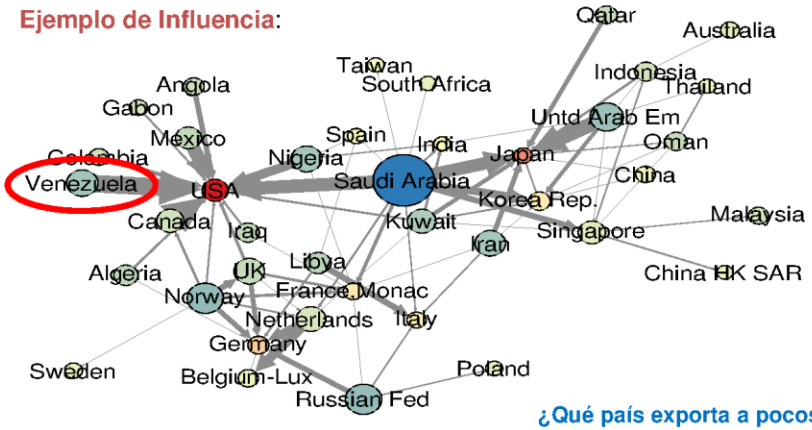


Comercio de petróleo y derivados, 1998
Fuente: NBER-United Nations Trade Data

¿Qué países importan de muchos otros?
¿Arabia Saudí, Japón, Iraq, USA, Venezuela?

MEDIDAS LOCALES DE CENTRALIDAD | Influencia

Ejemplo de Influencia:



¿Qué país exporta a pocos países (*out-degree* bajo) **pero lo hace en gran cantidad** (grosor de los arcos = volumen exportado)**?**
¿Arabia Saudí, Japón, Irak, USA, Venezuela?

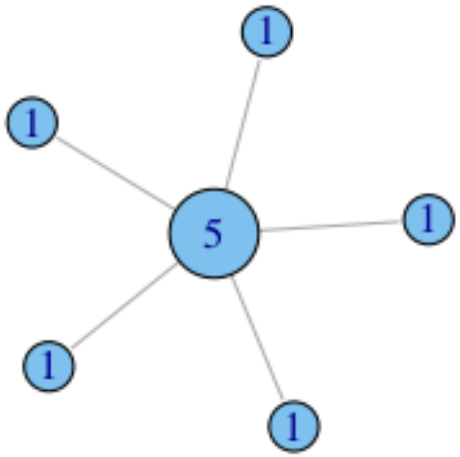
Métricas: Centralidad

$$C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1}$$

$$C_D = \frac{\sum_{i=1}^{g} [C_D(n^*) - C_D(i)]}{[(N-1)(N-2)]}$$

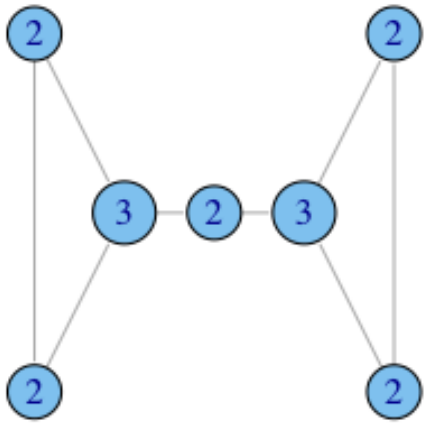
Máximo valor de conexiones posibles en la red

Formula de centralidad general de Freeman's



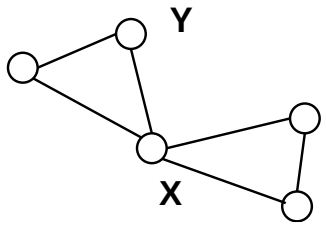
$C_D = 1.0$

$C_D = 0.167$



$C_D = 0.167$

Métricas: betweenness



La centralidad de intermediación ve al nodo con una posición favorable en la medida que el nodo está situado entre los caminos entre otros pares de actores en la red.

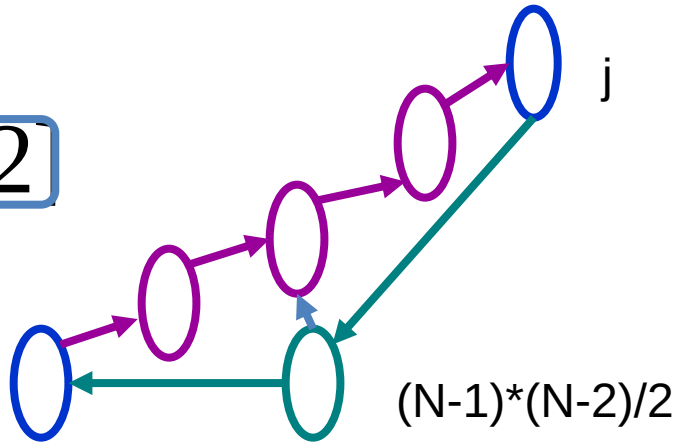
$$C_B(i) = \sum_{j<k} g_{jk}(i)/g_{jk}$$

Pares de vértices posibles excluyendo el del mismo nodo

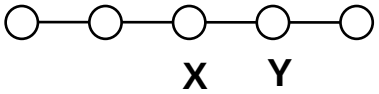
Donde g_{jk} = numero de caminos cortos que conectan *jk*

$g_{jk}(i)$ = numero de caminos cortos en los que el nodo i se encuentra.

$$C_B'(i) = C_B(i)/[(n-1)(n-2)/2]$$



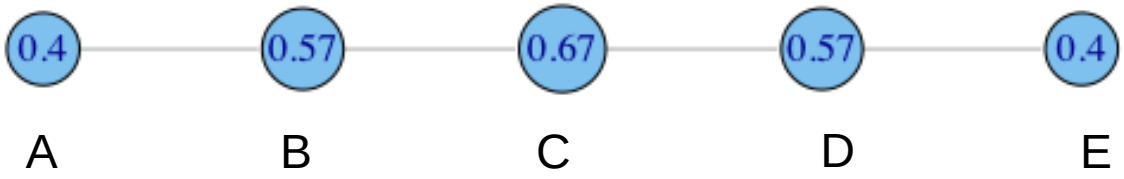
Métricas: closeness



Distancia promedio del camino mas corto entre un nodo a todos los nodos.

Closeness Centrality: $$C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$

Normalized Closeness Centrality $$C_C'(i) = (C_C(i))/(N-1)$$



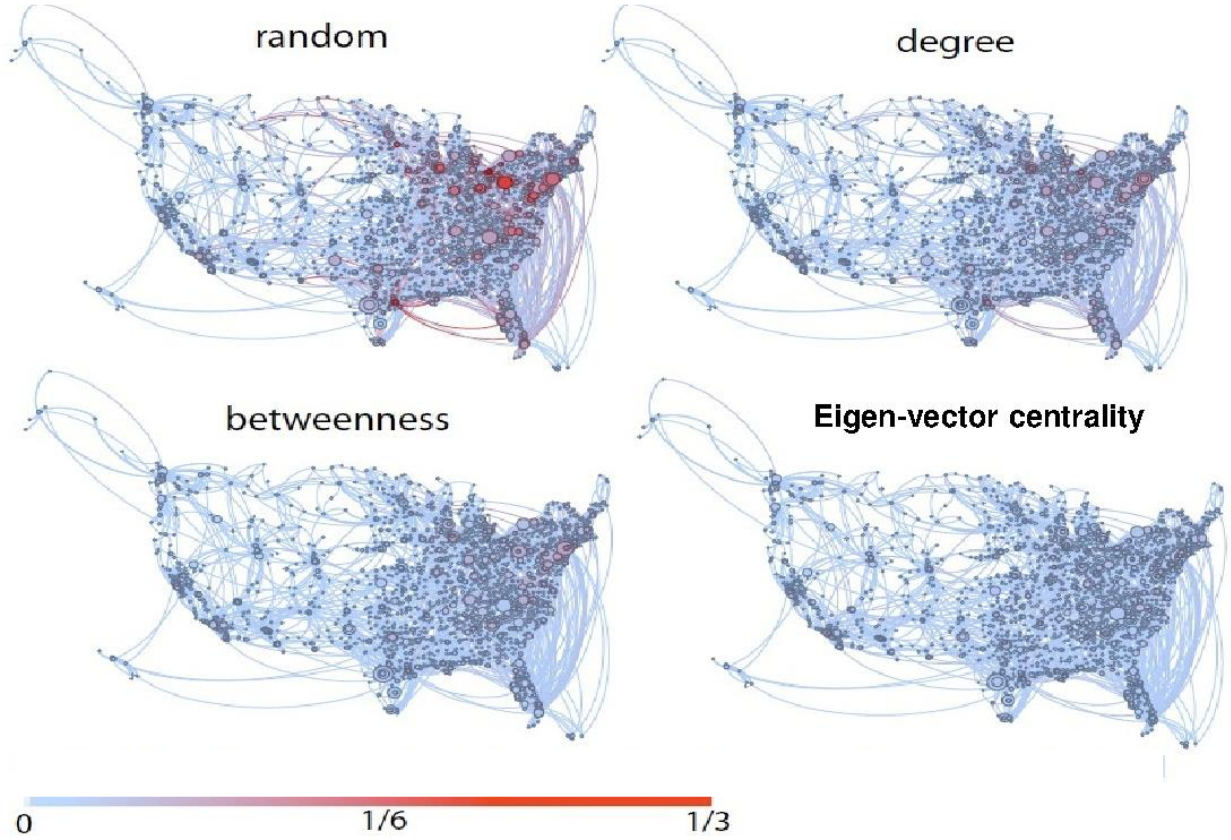
$$C_c'(A) = \frac{\left[\sum_{j=1}^{N} d(A,j)\right]^{-1}}{N-1} = \left[\frac{1+2+3+4}{4}\right]^{-1} = \left[\frac{10}{4}\right]^{-1} = 0.4$$

Centralidad de Vectores propios

Excentricidad

Infecciones en hospitales EEUU por transferencia de pacientes (4)

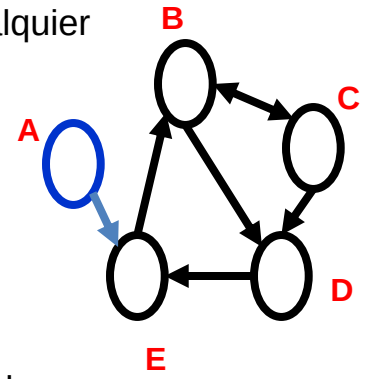
Red de transferencias de pacientes hospitalarios
Presupuesto para estrategias para evitar propagación de infecciones

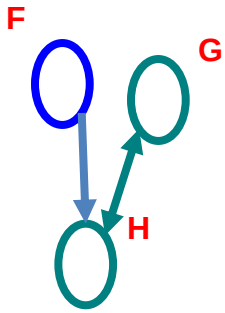


Probabilidad de ser infectado en un hospital

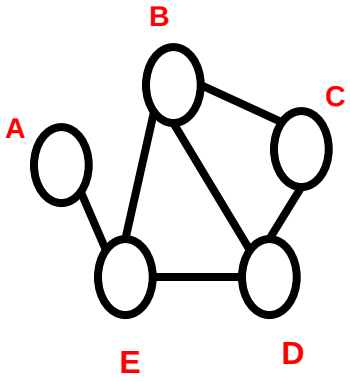
Métricas: Componentes conectados

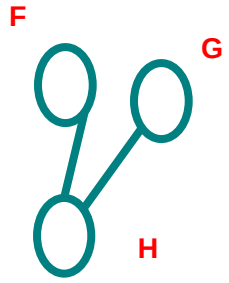
- **Componentes fuertemente conectados:**
  - Cada nodo dentro del componente se puede llegar desde cualquier otro nodo en el componente siguiendo los enlaces dirigidos
    - Componentes fuertemente conectados
      - B C D E
      - G H





- **Componentes débilmente conectados:** cada nodo se puede llegar desde cualquier otro nodo siguiendo ciertos enlaces en ciertas direcciones
  - Componentes débilmente conectados
    - A B C D E
    - G H F



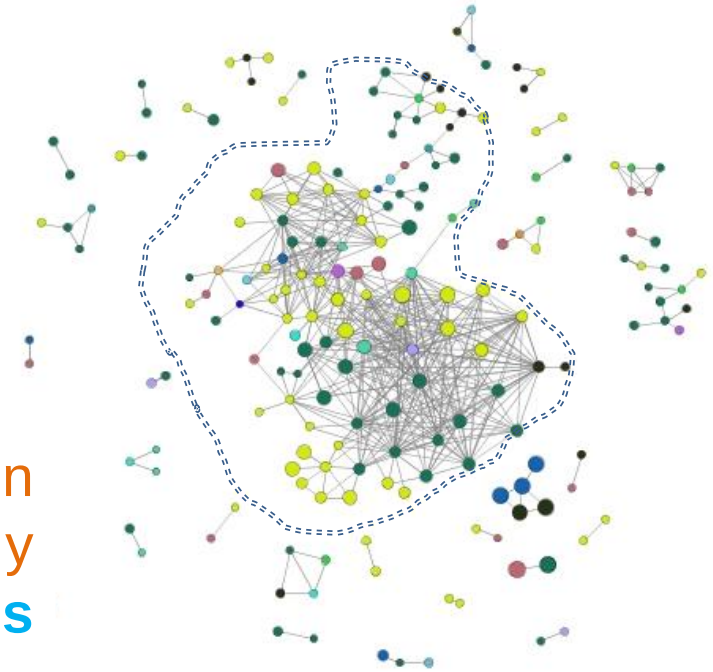


En las **redes no dirigidos** se habla simplemente de "**componentes conectados**'

Métricas: Componentes conectados

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

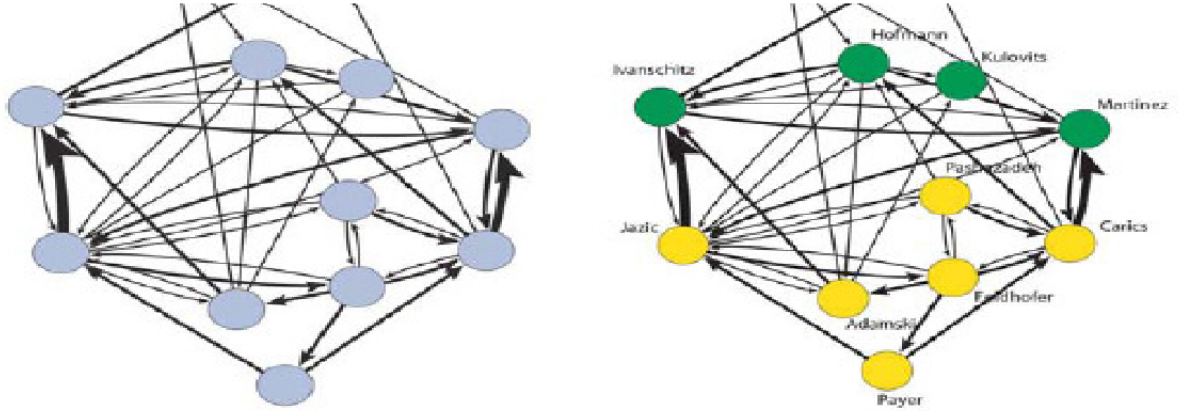
  giant component



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 en casi todas las redes sociales

Descripción de un equipo de futbol durante un juego

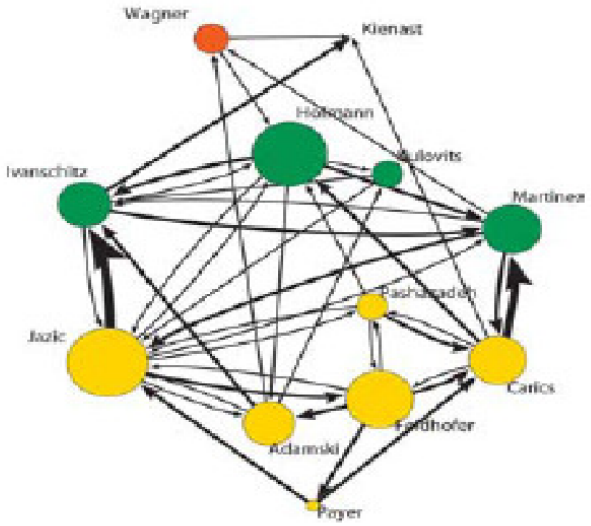


Grafo que describe los pases entre los jugadores

APLICACIONES | Análisis de juego en equipos de fútbol (2)

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**



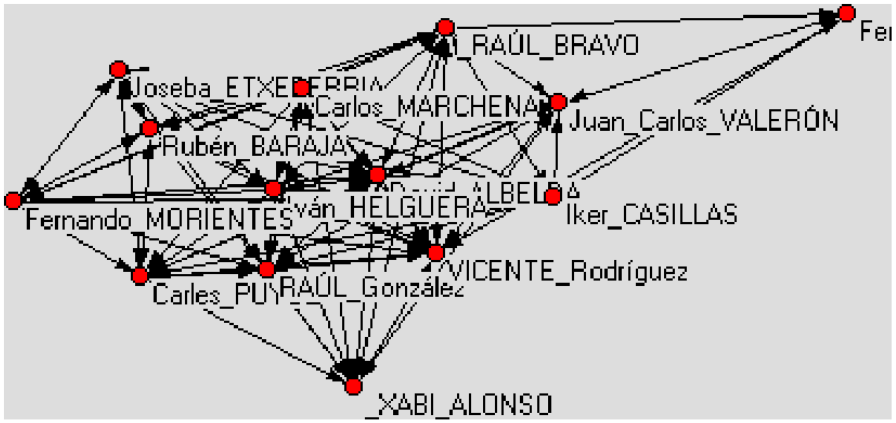
APLICACIONES | **Análisis de juego en equipos de fútbol (4)**

J.J. Merelo. Redes contra redes: el fútbol es así. http://atalaya.blogalia.com/historias/19642

Euro-copa Portugal 2004

Extraer las estadística de pases

España 1 – Rusia 0

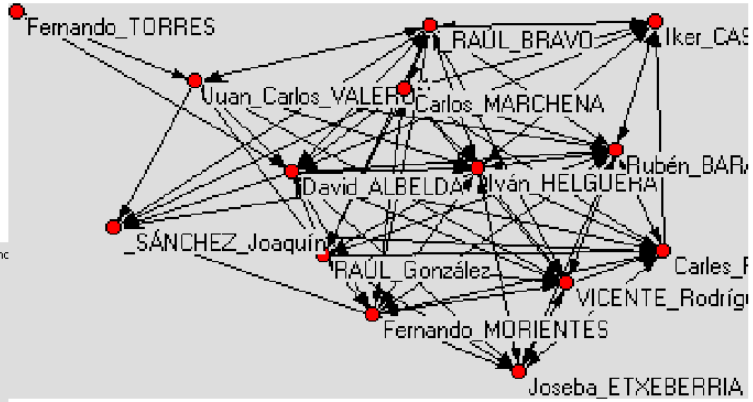


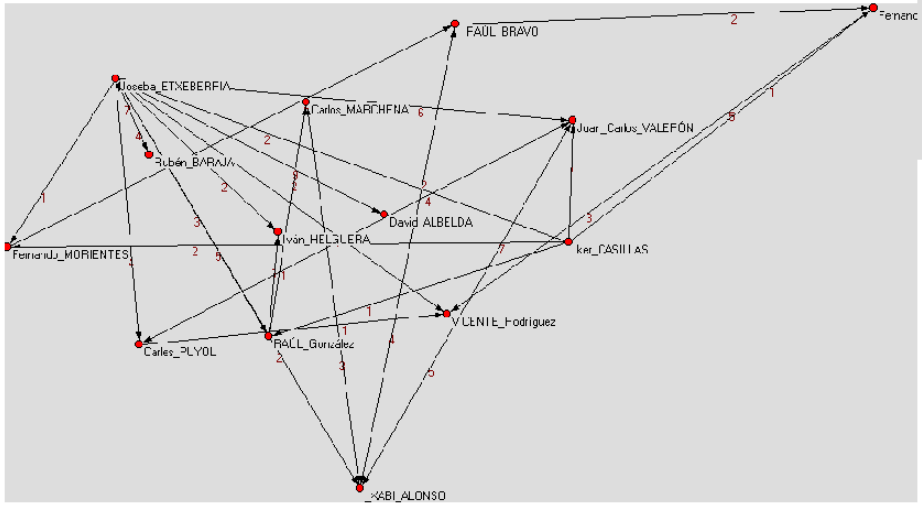
Una red donde, curiosamente, el jugador que tiene más centralidad es Iker Casillas, cuando debería ser un medio como Baraja

APLICACIONES | Análisis de juego en equipos de fútbol (5)

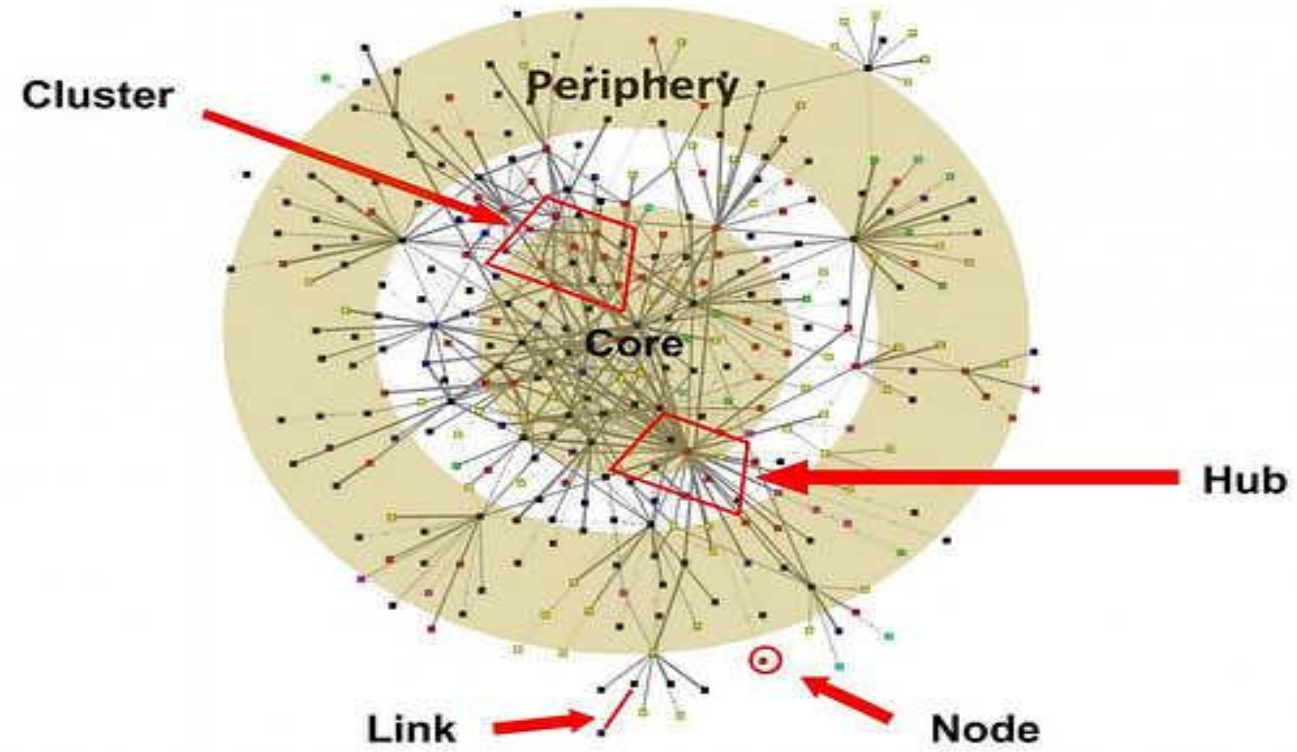
La situación no cambió mucho en el segundo partido (Grecia 1 – España 1) salvo que, en este caso, Albelda, Baraja y Helguera organizaron un poco más el juego. Y Fernando Torres a su bola, claro

Casi el 90% de los pases fueron los mismos. Esta es la diferencia de las dos redes (sin Joaquín porque es un nodo nuevo)





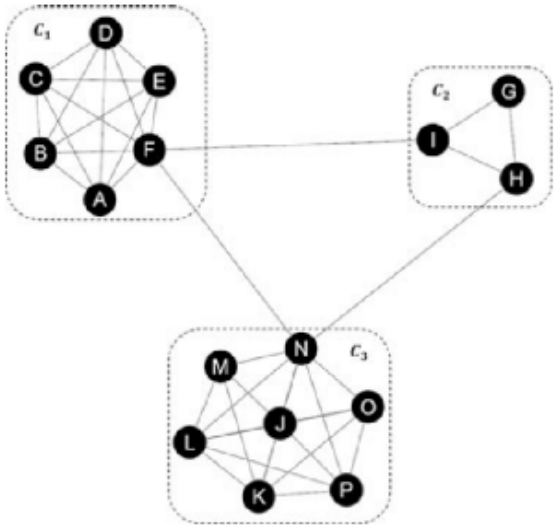
Es curioso ver también que la "autoridad" de la red es Vicente, un extremo. Lo lógico sería que las autoridades fueran los delanteros, pero Morientes y Raúl se hallan ahí perdidos, en la maraña de la red


Cluster
Periphery
Core
Hub
Link
Node

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

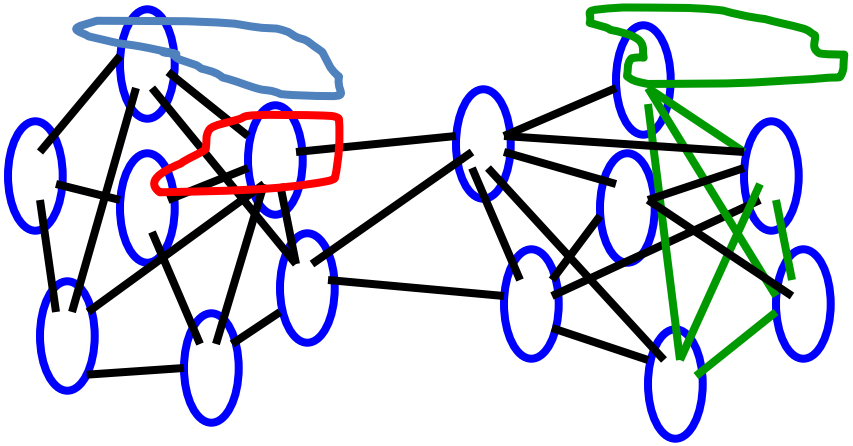
Métricas: Comunidades



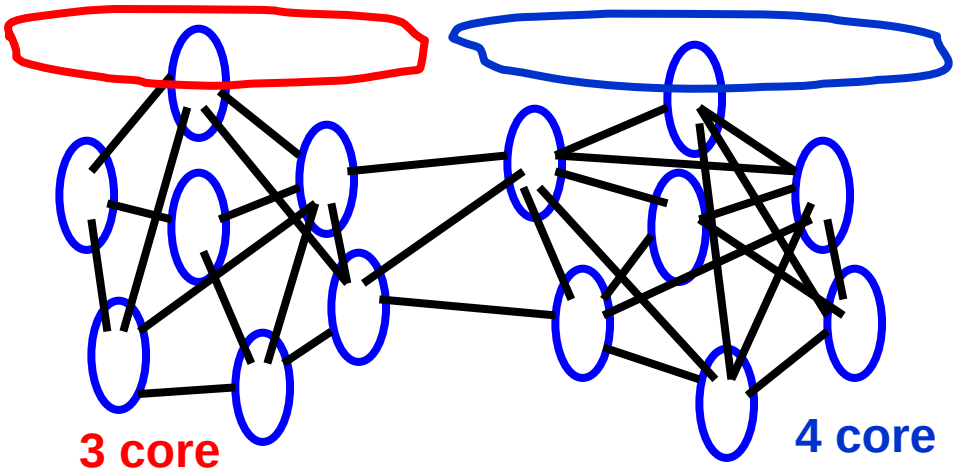
- En esta red, hay **tres comunidades**: C_1, C_2 y C_3
- Cada comunidad está formada por un grafo completo (un **clique**) de tamaño variable ($C_1 = K_6$, $C_2 = K_3$ y $C_3 = K_7$)
- La densidad de enlaces entre las comunidades es muy baja. Los pocos enlaces que existen son **puentes**

Cliques y K-core

Cada miembro del grupo posee un link a todos los miembros del grupo



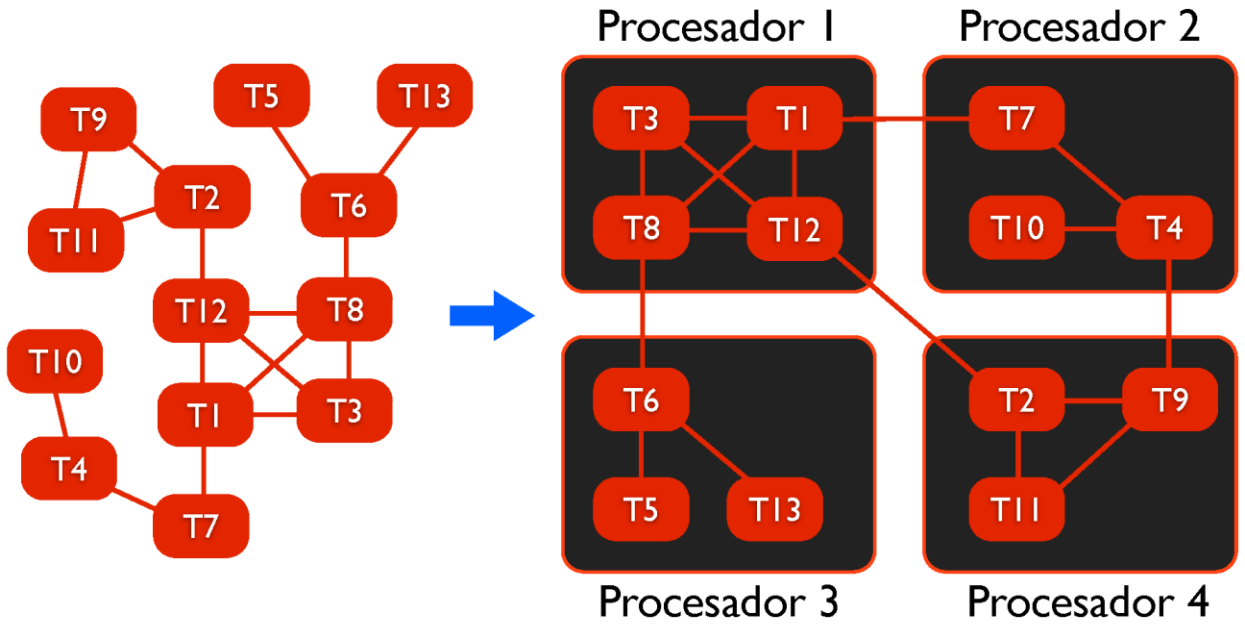
- Si se pierde un enlace, deja de ser un clique
- No es interesante que todos estén conectados con todos
- No hay medidas de centralidad dentro de un clique



Cada miembro del grupo está conectado con k otros miembros del grupo

Computación en Paralelo

(Particionado de Grafos)

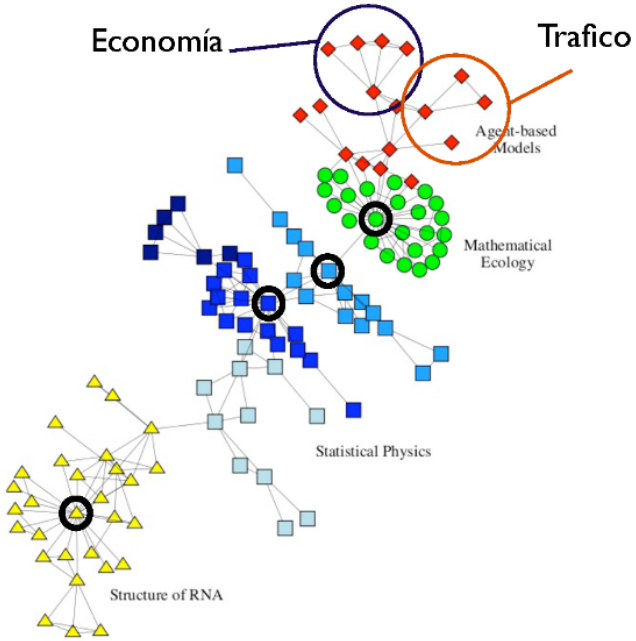


Distribución mas eficiente de tareas en un conjunto de procesadores

Redes de colaboración científica

- ◆ Modelos basados en agentes para estudiar problemas de economía y flujo de tráfico
- ● Modelos matemáticos en ecología
- □ Física estadística
- △ Estructura del ARN

- Formación de comunidades en torno a la metodología
- El centro de las comunidades corresponde al jefe del grupo de investigación.



Red de colaboración de científicos del Instituto de Santa Fe en Nuevo México

M. Girvan and M. E. J. Newman (2002). "Community structure in social and biological networks". Proc. Natl. Acad. Sci. USA 99 (12): 7821–7826. doi:10.1073/pnas.122653799. PMC 122977. PMID 12060727.

Métricas: Comunidades

Algoritmos de detección

- Edge Betweenness Method, M. Girvan and M. E. Newman (GN)
- Fast greedy modularity optimization, A. Clauset, M. E. Newman, and C. Moore (Clauset et al.)
- Exhaustive modularity optimization via simulated annealing, R. Guimerá, M. Sales-Pardo (Sim ann.)
- Multi-Level Aggregation Method based on modularity, V. D. Blondel, J.-L. Guillaume, R. Lambiotte (Blondel et al.)
- Divisive algorithm based on the edge-clustering coefficient, F. Radicchi, C. Castellano, F. Cecconi (Radicchi et al.)
- Clique Percolation Method for finding communities, G. Palla, I. Derenyi, I. Farkas (Cfinder)
- Graph clustering by flow simulation, S. van Dongen (MCL)

Modularidad

- Métrica diseñada para dividir la red en módulos, clusters, comunidades.
- Es usada para maximizar métodos para hallar comunidades
- Una red con alta modularidad significa que:
  - Es muy densa entre nodos de una misma comunidad
  - Pero con conexiones dispersas entre nodos de comunidades distintas

Detección basada en la Modularidad

Heurísticas utilizadas para la optimización de la modularidad:

- Recocido Simulado (Guimera and Amaral)
- Optimización extrema (J. Duch and A. Arenas)
- Algoritmos voraces (Clauset et al.)
- Reformulación de la modularidad en términos de las propiedades espectrales de la red. (Newman)

Las dos ultimas heurísticas han resultado efectivas sin embargo poseen un problema inherente al concepto de modularidad llamado *limite de resolución*.

Modularity and community structure in networks, M. E. J. Newman, Proc. Natl. Acad. Sci. USA 103, 8577–8582 (2006).

Wikipedia, 2011. Modularity (networks). [en linea] Disponible en: <http://en.wikipedia.org/wiki/Modularity_(networks)> [Consultado el 17 Noviembre 2011].

MEDIDAS GLOBALES

Existen varias medidas globales en SNA. La mayoría son las mismas empleadas para analizar cualquier otro tipo de red, que ya hemos estudiado:

1. **Diámetro y Radio**
2. **Distancia media**
3. **Grado Medio**
4. **Densidad**
5. **Coeficiente de Clustering Global**
6. **Reciprocidad**

MEDIDAS GLOBALES — Diámetro y Radio

Diámetro (d_{max}): longitud del camino mínimo más largo de la red

En redes grandes, se puede determinar con el algoritmo de búsqueda primero en anchura

Equivale al valor máximo de excentricidad para todos los nodos de la red:

$$E(i) = \max_{j \in V(G)/i} d(i,j) \qquad d_{\max} = \max\{E(i) : i \in V(G)\}$$

En el contexto del SNA, esta métrica da una idea de la proximidad entre pares de actores en la red, indicando cómo de lejos están en el peor de los casos

Las redes más dispersas suelen tener un mayor diámetro que las más densas al existir menos caminos entre cada par de nodos

Radio (r): Valor mínimo de excentricidad para toda la red:

$$r = \min\{E(i) : i \in V(G)\}$$

- Densidad: Actividad global de la red
  - En redes no dirigidas:
    $\Delta = L/[g(g-1)/2]$
  - En redes dirigidas:
    $\Delta = L/g(g-1)$

L, número de enlaces
g, número de nodos

MEDIDAS GLOBALES — Distancia media

Distancia media (<d>) para un grafo dirigido:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

d_{ij} es la distancia geodésica entre los nodos *i* y *j*

En un grafo no dirigido $d_{ij}=d_{ji}$. De este modo, sólo es necesario contar la longitud de los caminos una vez:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j>i} d_{ij}$$

La medida da una idea de cómo de lejos están los distintos actores en promedio. En SNA representa la eficiencia del flujo de información en la red

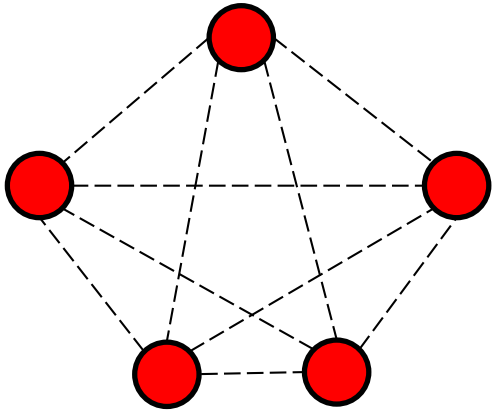
Medidad de red : Densidad de una red no dirigida

En el siguiente grafo:

Las conexiones posibles son:

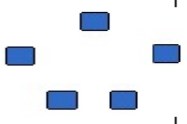
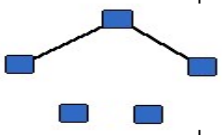
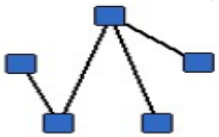
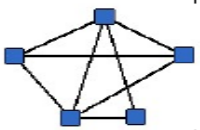
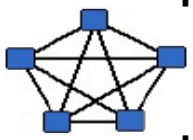
10 conexiones posibles en el grafo.

Pos.= posibles
Eff . = efectivas



Medidas de red
Densidad de una red no dirigida

$$d = \frac{n^{\circ}\ effective\ edges}{n^{\circ}\ possibleedges}$$

Eff.=0 Pos.=10	Eff.=2 Pos.=10	Eff.=4 Pos.=10	Eff.=8 Pos.=10	Eff.=10 Pos.=10
d=0	d=0.2	d=0.4	d=0.8	d=1
				

APLICACIONES | Aprendizaje Colaborativo por Ordenador (1)

Aprendizaje Colaborativo por Ordenador: énfasis en las relaciones entre los actores en un curso on-line en BSCW

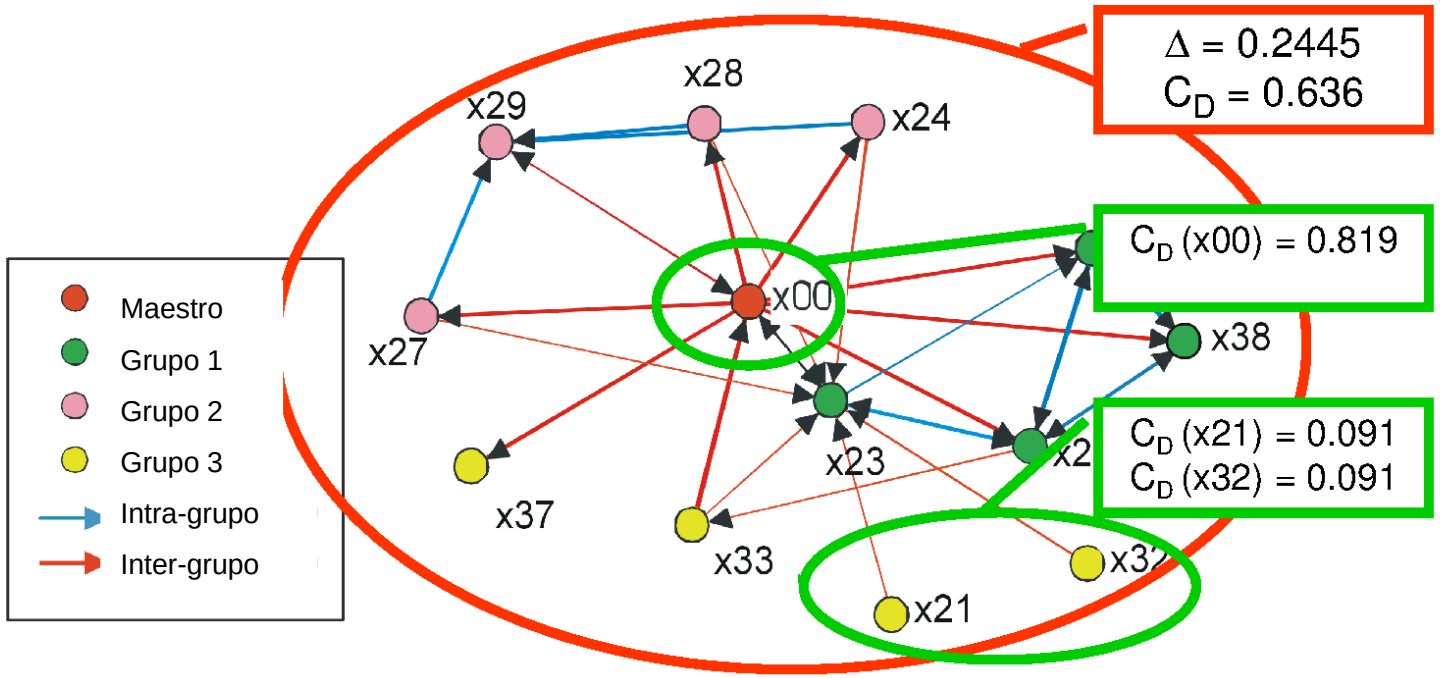
Relación entre el profesor y los alumnos en un curso, así como entre los propios alumnos de distintos grupos

Red Social con distintas variantes: trimodal egocéntrica (tres grupos de alumnos-profesor) y trimodal, unimodal completa (todos los alumnos con todos) y trimodal completa (miembros de un grupo con los de otros)

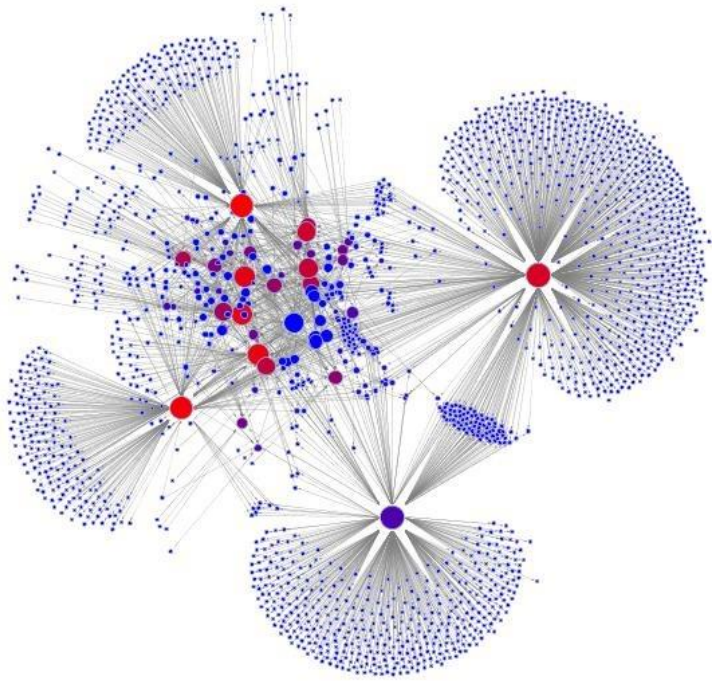
Pregunta global: ¿Cómo ayudar a los profesores a monitorizar aspectos colaborativos de aprendizaje mediante la tecnología?

Análisis con dos medidas globales (Densidad de la red Δ y Centralización de grado C_D) y dos medidas locales (Centralidades de grado y de cercanía)

APLICACIONES | Aprendizaje Colaborativo por Ordenador (2)



Power-law



- Nodos aparecen con el tiempo (growth model)
- Nodos prefieren unirse a nodos populares (preferential model)
- Nodos viejos Mueren
- Algunos nodos son mas sociable
- Las amistades pueden desaparecer

Propagación de epidemias

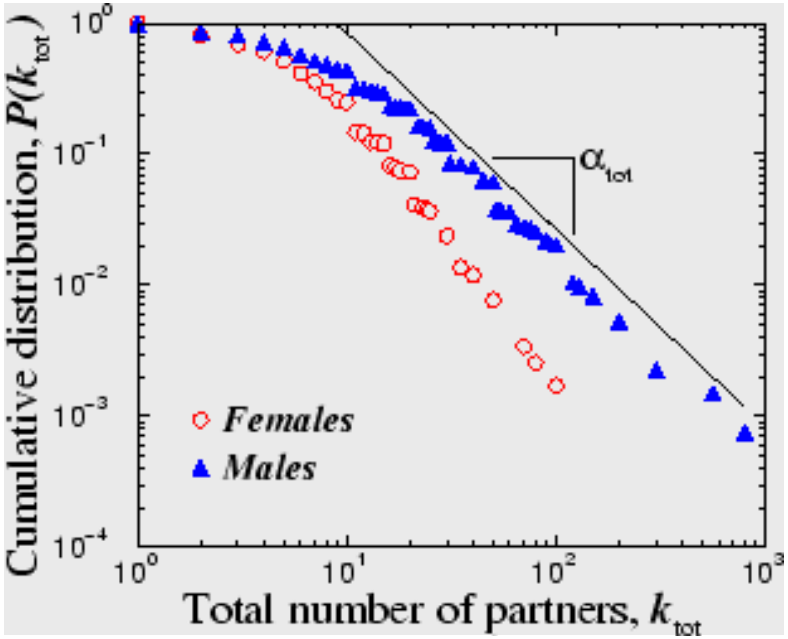
El papel de los centros en las epidemias

- En una red del tipo power-law, un virus puede persistir por más baja sea su capacidad de infección

- Muchas redes del mundo real hacen exhibir power-leyes:
  - contactos sexuales
  - redes de correo electrónico

Propagación de epidemias:
RED DE CONTACTOS SEXUALES

Nodos: individuos
Links: relaciones sexuales



HAY UNOS POCOS NODOS CON MAYOR PROBABILIDAD DE CONTAGIAR QUE OTROS (HUBS)

ESTRATEGIAS DE PREVENCION DE EPIDEMIAS

Difusión de información en las redes

- **Factores que influyen en la difusión de información**
  - **estructura de la red: la que se conectan los nodos?**
  - **fuerza de los lazos: qué tan fuerte son las conexiones?**

- **Estudios en la difusión de la información:**
  - **Granovetter: la fuerza de los lazos débiles**
  - **J-P Onnela et al: fuerza de los lazos intermedios**
  - **Kossinets et al: fuerza de los lazos de backbone**
  - **Davis: enclavamientos de mesa y adopción de practices**

Aplicaciones

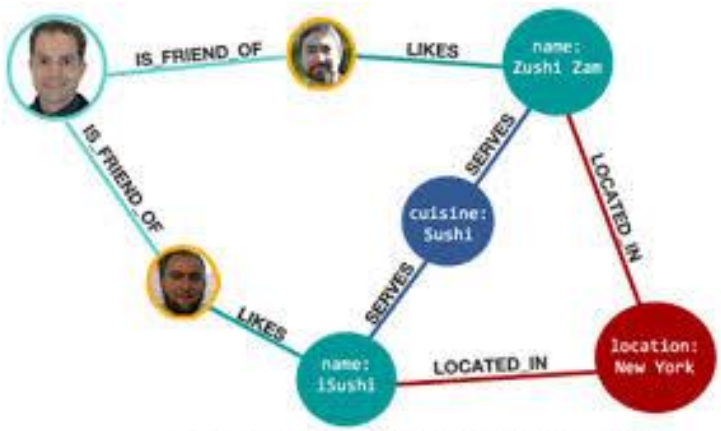
- Epidemiología (propagación de virus).

- Tolerancia frente a ataques (deliberados).

- Procesos de optimización (publicidad).

- . . .

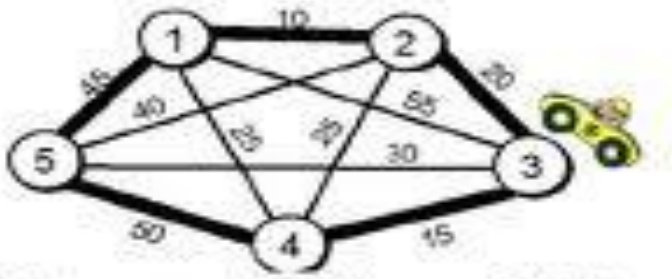
BASES DE DATOS ORIENTADAS A GRAFOS (BDOG)



- **BDOG:**
- Representan la información como nodos de un grafo y sus relaciones con las aristas del mismo.
- Una BDOG debe estar absolutamente normalizada.

VENTAJAS DE UNA BDOG

Grafo que representa las redes de comunicación



Grafo que representa 5 ciudades con sus distancias.

VENTAJAS

- Consultas más amplias y no demarcadas por tablas
- No hay que definir un número determinado de atributos
- Los registros también son de longitud variable
- Se puede recorrer directamente la base de datos de forma jerárquica

MOTORES DE MODELAMIENTO GRAFICO DE UNA BDOG

AllegroGraph - Escalable y de alto rendimiento.
Bigdata - RDF/base de datos orientada a grafo.
CloudGraph - .NET usa tanto los grafos como clave/valor para almacenar los datos.
Cytoscape - Bioinformática
DEX/Sparksee - De alto rendimiento, permite escalar billones de objetos.
Comercializada por Sparsity Technologies.
Filament
GraphBase
Graphd, backend de Freebase
Horton
HyperGraphDB - Base de datos opensource basada en la idea de hipergrafo.
InfiniteGraph
InfoGrid - Open Source
Neo4j - Open Source.
OrientDB - Base de datos orientada a grafos y documental.
OQGRAPH
sones GraphDB
VertexDB
Virtuoso Universal Server
R2DF

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 …

Minería de Patrones de Grafos

Minería subgrafo frecuentes

- **Encontrar subgrafos frecuentes dentro de un grafo**
  - SUBDUE (DOMINAR)
- **Encontrar (sub)grafos frecuentes en un conjunto de grafos**
  - *Support* (frecuencia de ocurrencia) no inferior a un umbral mínimo
  - Enfoques basado en Apriori
  - Enfoques de crecimiento del patrón (Pattern-growt)
- **Aplicaciones de la minería de patrones de grafos**
  - Minería de estructuras bioquímicas, de flujos de programas, de estructuras XML y comunidades de la Web
  - Construcción de sistemas de clasificación, agrupación, compresión, comparación y análisis de correlación para grafos

Enfoques de Minería subgrafo frecuentes

- **Enfoques basados en Apriori**
  - AGM: Inokuchi, et al. (PKDD'00)
  - FSG: Kuramochi and Karypis (ICDM'01)
  - PATH: Vanetik and Gudes (ICDM'02, ICDM'04)
  - FFSM: Huan, et al. (ICDM'03)
- **Enfoques de crecimiento del patrón (Pattern-growt)**
  - MoFa, Borgelt and Berthold (ICDM'02)
  - gSpan: Yan and Han (ICDM'02)
  - Gaston: Nijssen and Kok (KDD'04)
- **Minería de patrones cercanos**
  - CLOSEGRAPH: Yan & Han (KDD'03)

Medidas de similitud basadas en los patrones de grafos

- Medidas de similitud basado en características
  - Cada grafo se representa como un vector de características
  - Vector de distancia
- Medida de similitud basada en la Estructura
  - Subgrafo común máximo
  - Grafo edita distancia: inserción, supresión, y re-etiquetado

Datos enlazados o vinculados (Linked Data)

Actual Web

Microformatos

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

XFN quick reference

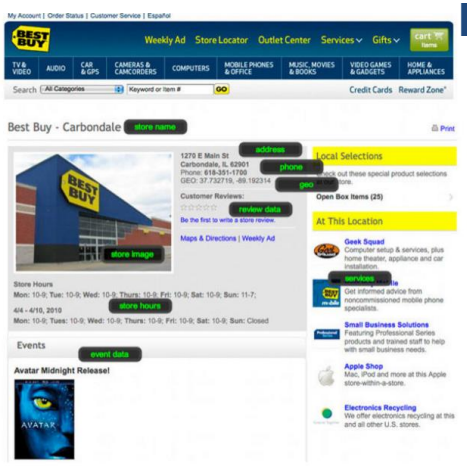
relationship category	*XFN values*
friendship (at most one):	friend acquaintance contact
physical:	met
professional:	co-worker colleague
geographical (at most one):	co-resident neighbor
family (at most one):	child parent sibling spouse kin
romantic:	muse crush date sweetheart
identity:	me

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."

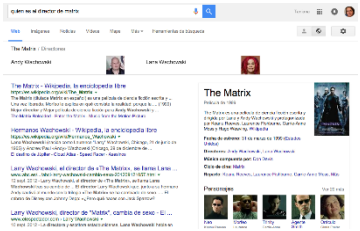
Best Buy is using Good Relations, a Semantic Web vocabulary for e-commerce that describes product, price, and company data.

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

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)

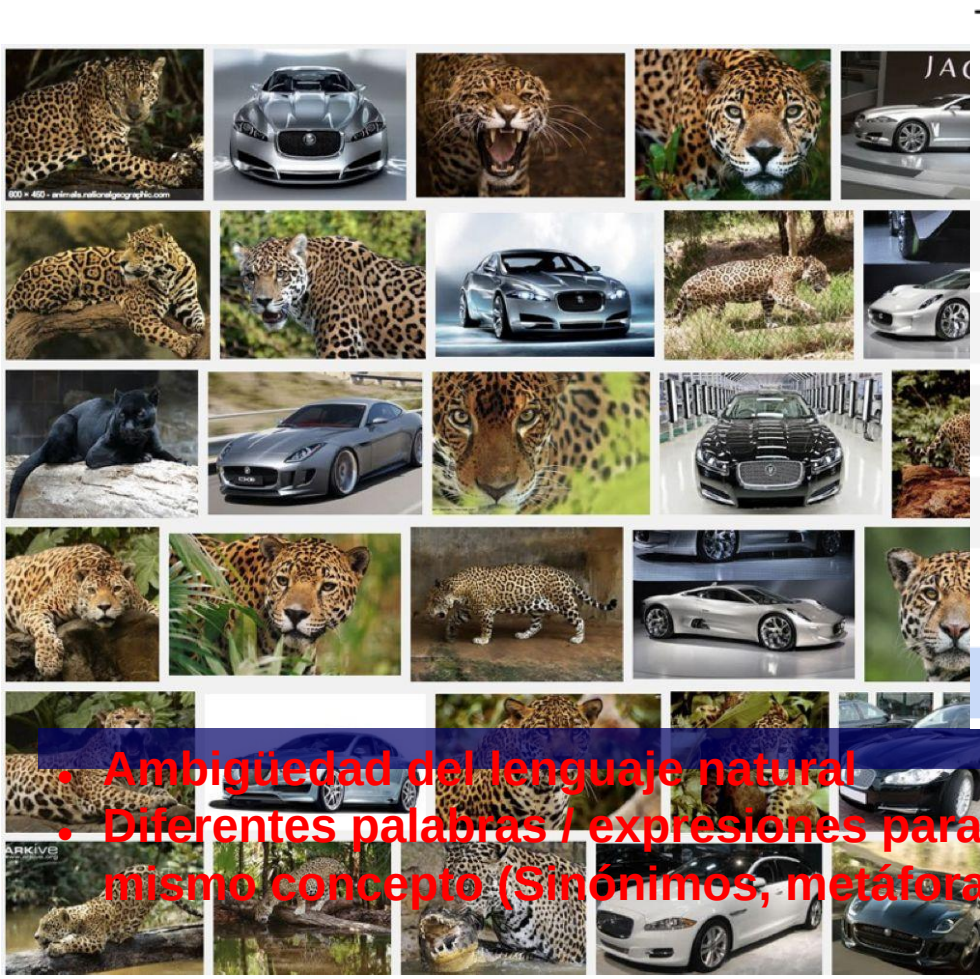
Por qué Linked Data?

- 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 Linked Data?

Problema en la recuperación de la información

Conocimiento Implícito

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

Alternativa…

- Las ontologías se “enlazarían” a través de las partes comunes (Nombre, Dirección, etc.)
- Usuarios conocen las ontologías que ellos requieren consultar
- Las consultas se hacen sobre las ontologías individuales
- La parte común de una ontología se conecta con las partes similares de las otras
- A partir de ese “enlazado”, se extrae localmente un subconjunto de esa global ontología compartida

Beneficios de utilizar Linked Data

- La información se agrega dinámicamente a partir de datos externos disponibles públicamente (Wikipedia, MusicBrainz, Last.FM, ...)
- Sin Screen Scraping
- No requiere API especializada
- Todos los datos están disponibles como estándar abierto
- Acceso a datos a través de una simple petición HTTP
- Los datos siempre están actualizados sin interacción manual

Representación del conocimiento

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

Pluto : Planet
discovered = 1930

UML instance

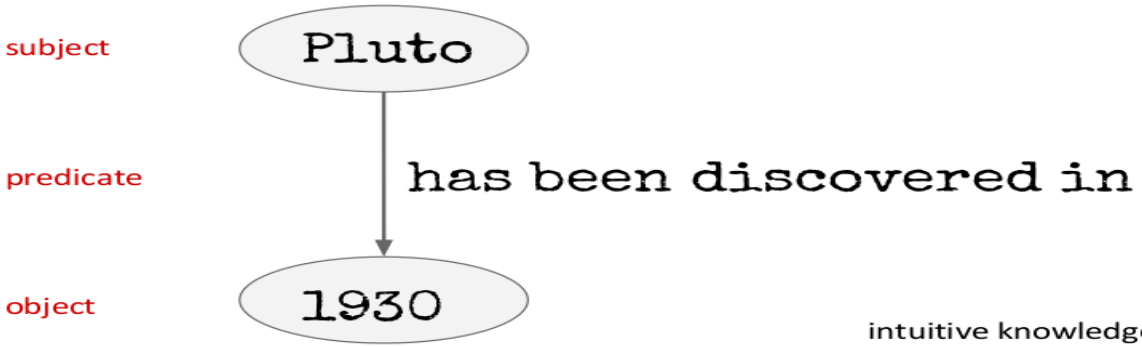
```
<a href="http://en.wikipedia.org/wiki/Pluto">
 Pluto
</a> has been discovered in 1930.
```

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**

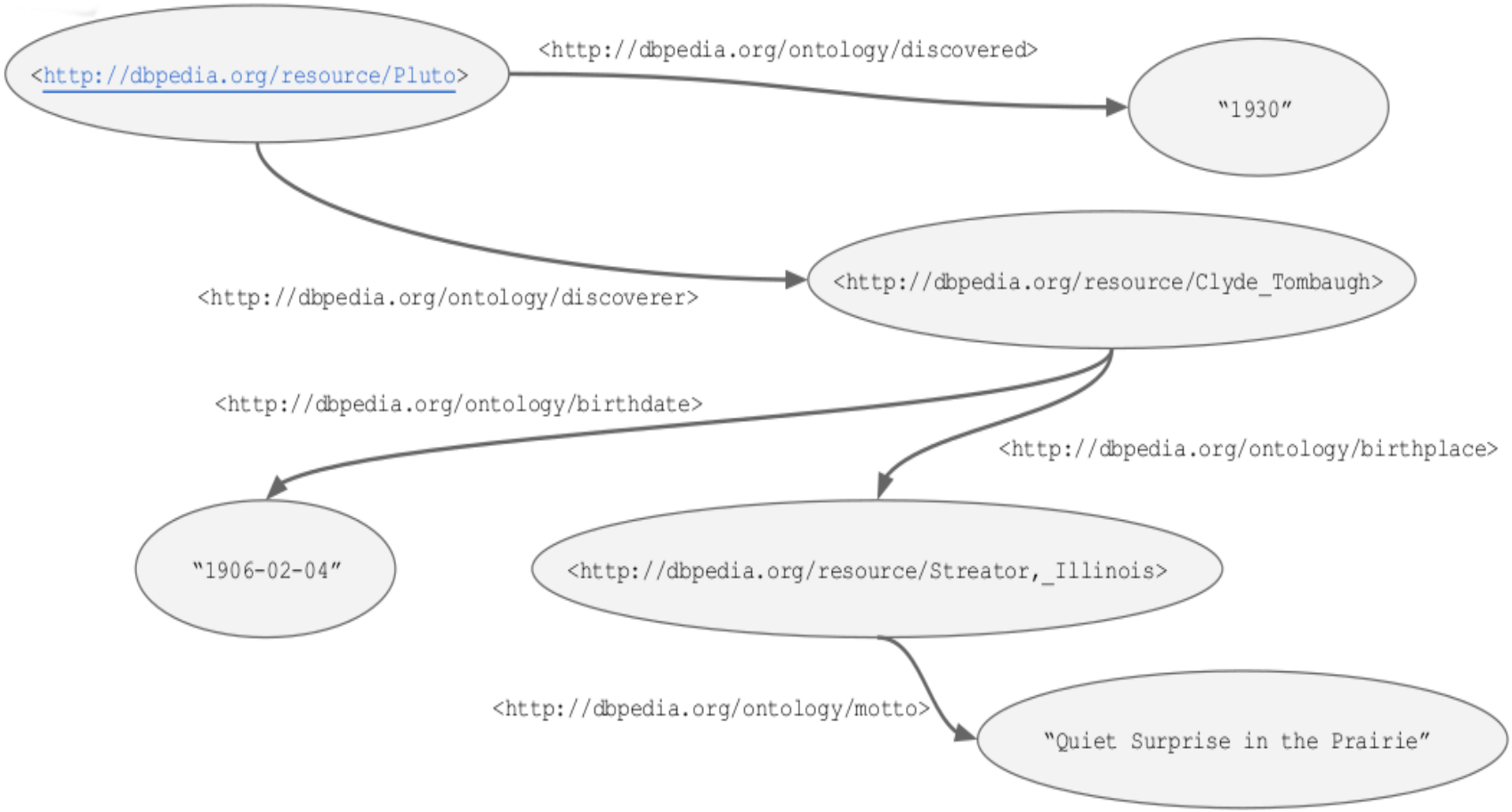
RDF Building Blocks

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 .
... ...

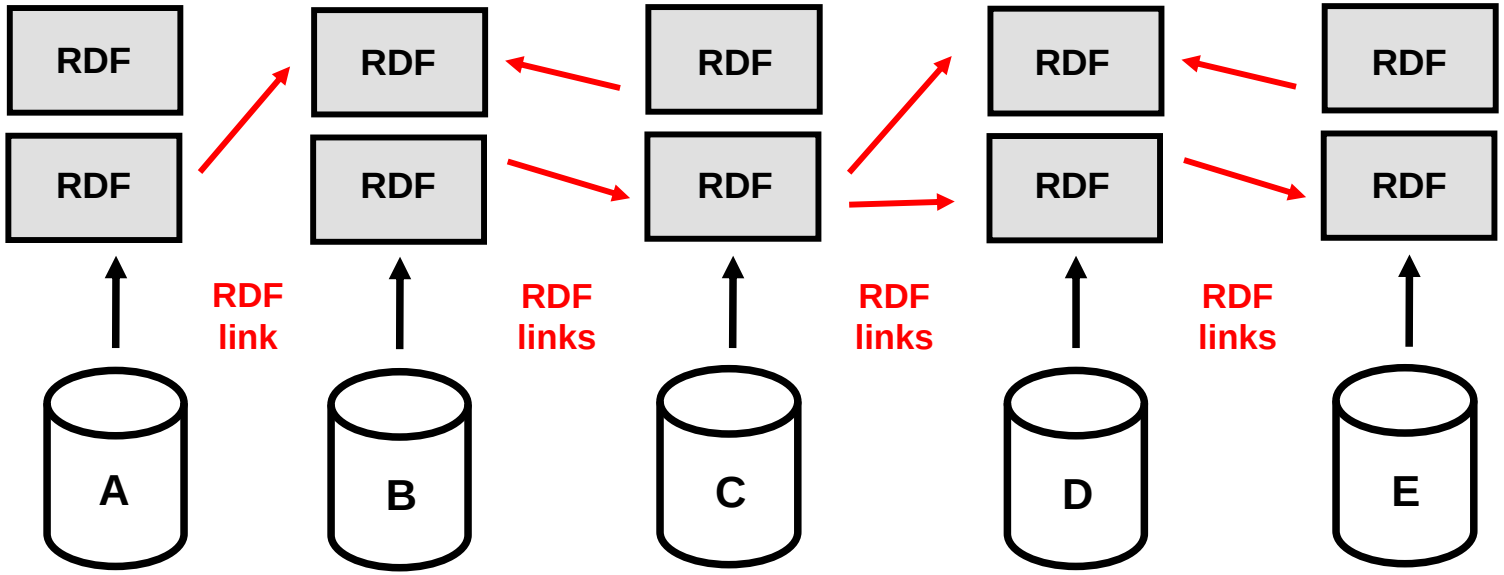
Subject Property Object

RDF Triples

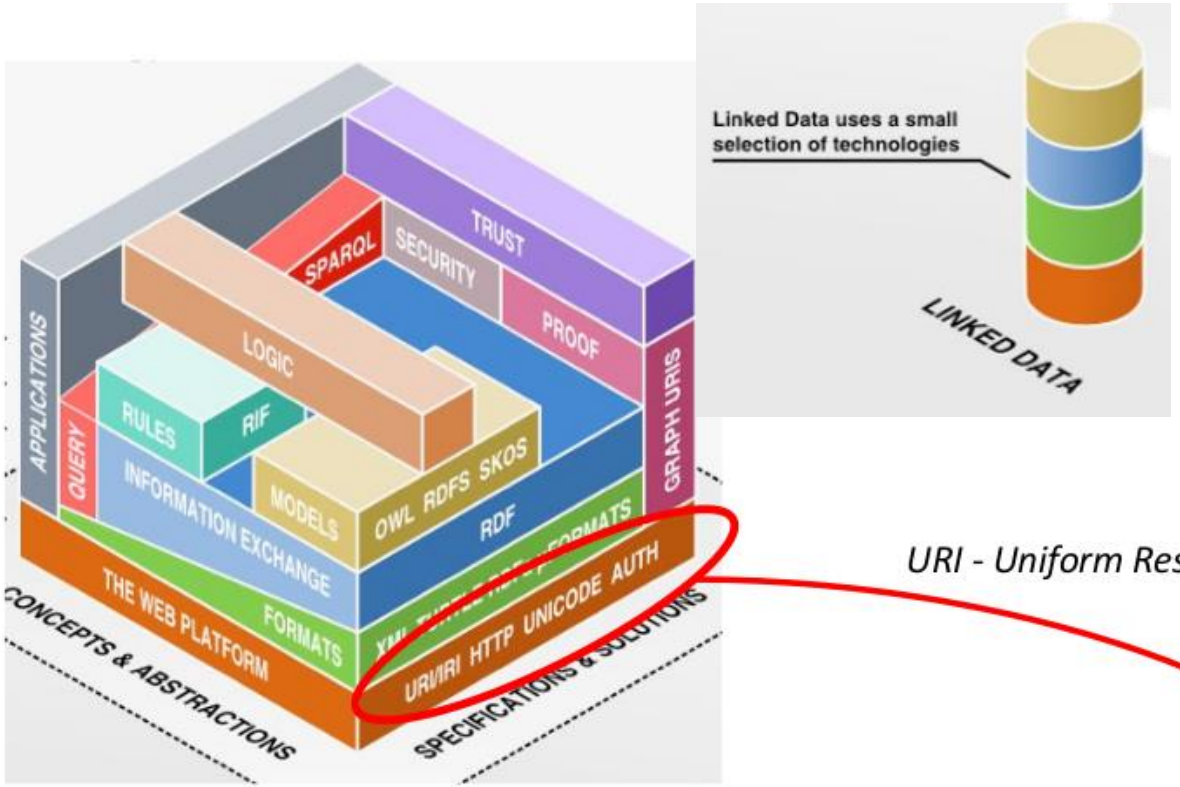
- Individuos (Entidades)
- Clases
- Literales / Valores
- Vocabularios / Ontologías

Datos enlazados (Linked Data)

1. **Publica datos estructurados en la Web,**
2. **Establece enlaces entre datos en diferentes Fuentes.**



Tecnología de la Web Semántica (URI)

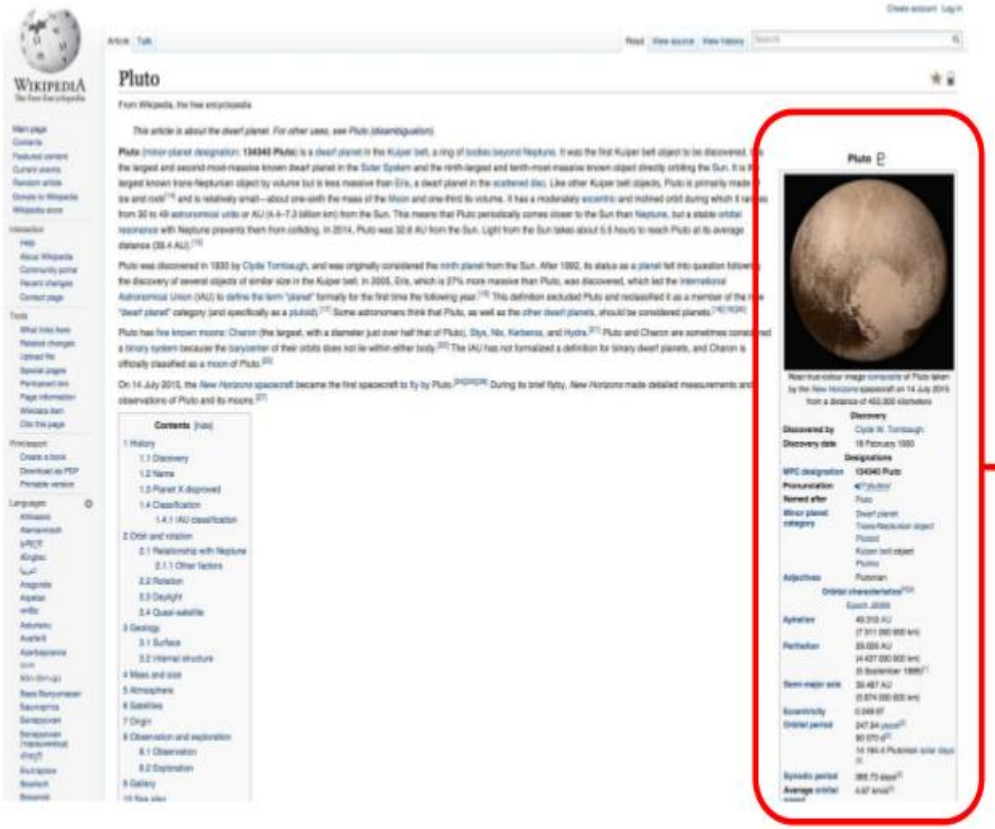


URI - Uniform Resource Identifier

Pluto ····► http://dbpedia.org/resource/Pluto

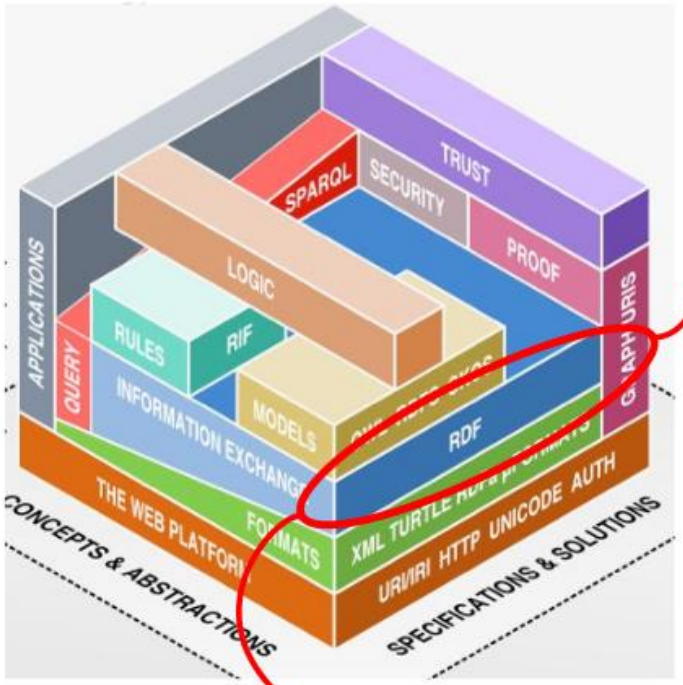
Tecnología de la Web Semántica (URI)

http://en.wikipedia.org/wiki/Pluto



http://dbpedia.org/resource/Pluto

Tecnología de la Web Semántica (RDF)



http://dbpedia.org/resource/Pluto

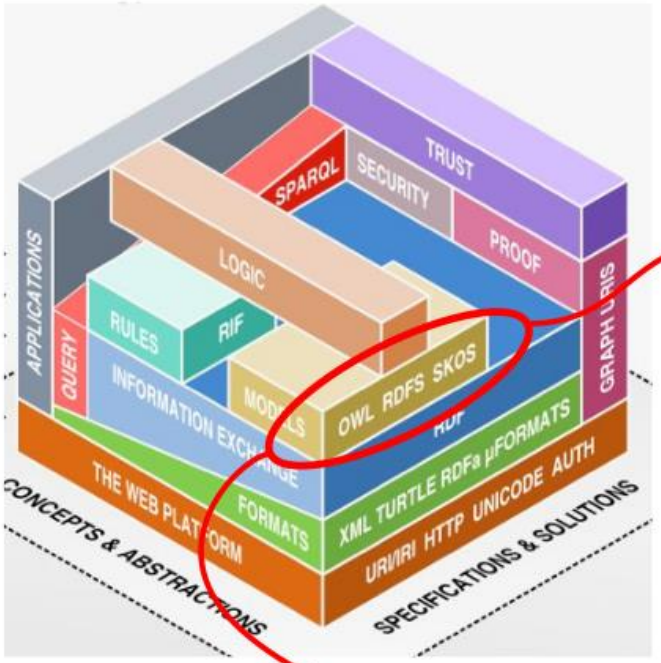
```
:Pluto rdf:type dbo:Planet .
:Pluto foaf:name "Pluto"@en .
:Pluto dbo:discoverer :Clyde_Tombaugh .
:Pluto dbo:discovered "1930-02-18"^^xsd:date .
:Clyde_Tombaugh rdf:type dbo:Person .
:Clyde_Tombaugh dbo:birthdate "1906-02-04"^^xsd:date .
...
```

RDF Resource Description Framework

RDF Triple

:Pluto	rdf:type	dbo:Planet .
RDF Subject	*RDF Property*	*RDF Object*

Tecnología de la Web Semántica (RDFs)



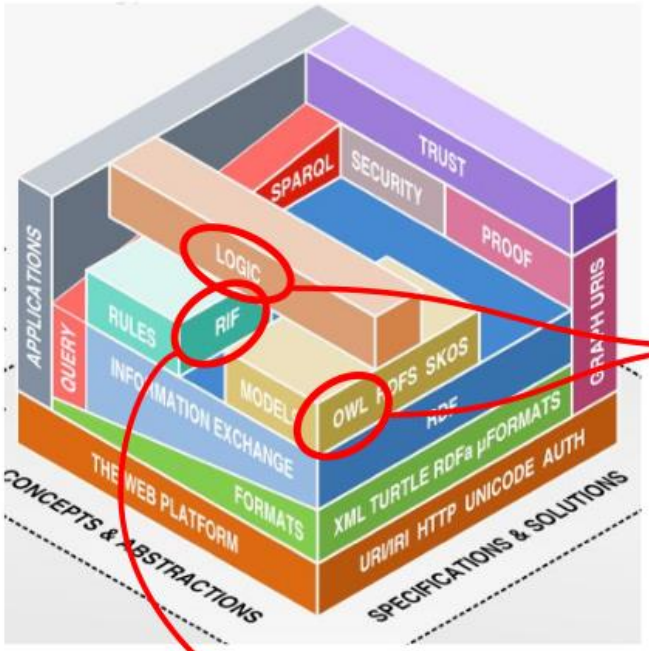
http://dbpedia.org/ontology/Planet

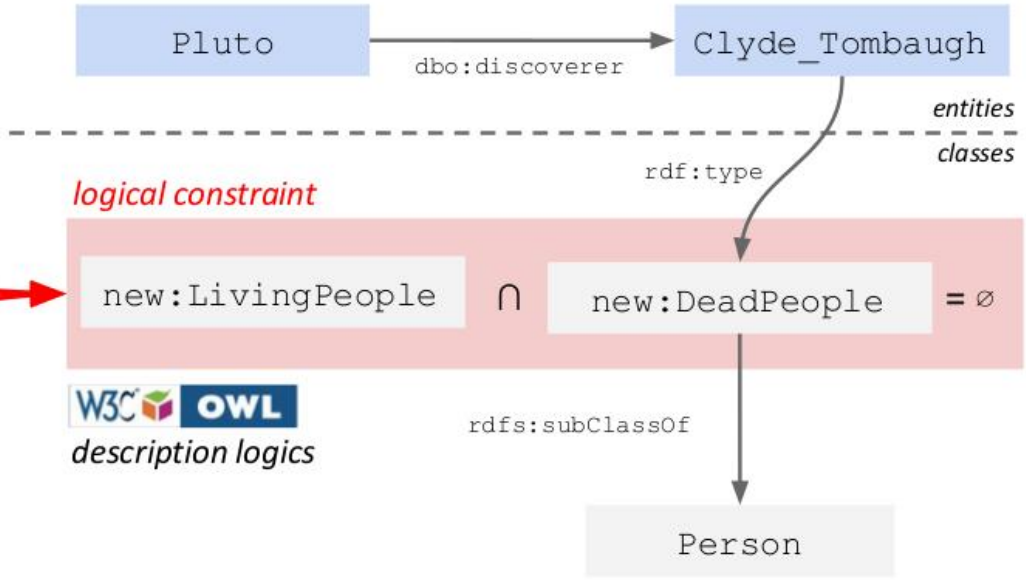
```
dbo:Planet rdf:type owl:class .
dbo:Planet rdfs:subClassOf dbo:CelestialBody .
dbo:discovered rdf:type rdf:Property .
dbo:discovered rdfs:domain owl:Thing .
dbo:discovered rdfs:range xsd:date .
dbo:discoverer rdf:type rdf: Property .
dbo:discoverer rdfs:domain owl:Thing .
dbo:discoverer rdfs:rang dbo:Person .
...
```

W3C RDFS

RDF Schema

Tecnología de la Web Semántica (OWL)

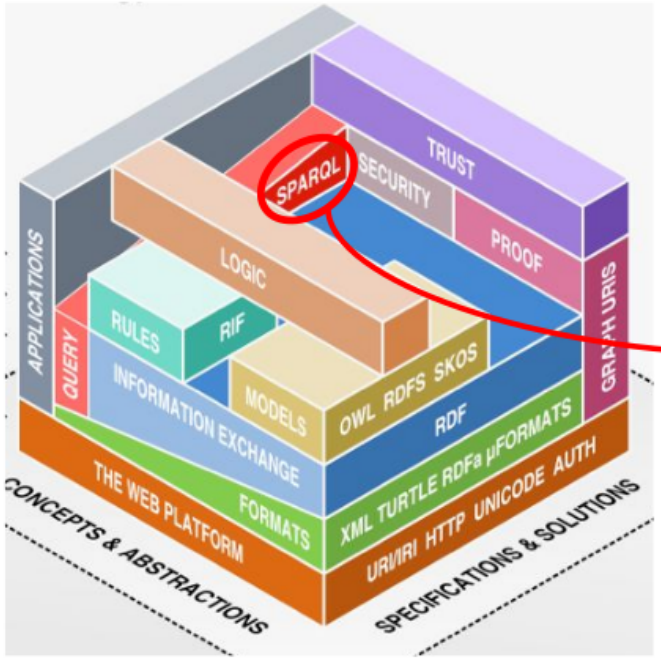




\+ *logical rules*

$$\forall x.\exists y.\mathrm{deathDate}(x,y) \wedge \mathrm{Person}(x) \wedge \mathrm{Date}(y) \rightarrow \mathbf{DeadPeople}(x)$$

Tecnología de la Web Semántica (SPARQL)



Look for all **space missions in the Solar System** which have become a **satellite** of their target

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbc: <http://dbpedia.org/resource/Category:>

SELECT distinct ?s ?o
FROM <http://dbpedia.org/>
WHERE{
?s dcterms:subject/skos:broader*
    dbc:Discovery_and_exploration_of_the_Solar_System ;
    dbp:satelliteOf ?o .
}
```

SPARQL: consulta

Buscar autores y los títulos notables de sus trabajos

```
PREFIX :     <http://dbpedia.org/resource/>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo:  <http://dbpedia.org/ontology/>

SELECT ?author_name ?title

FROM <http://dbpedia.org/>

WHERE {
     ?author rdf:type dbo:Writer .
     ?author rdfs:label ?author_name .
     ?author dbo:notableWork ?work .
     ?work rdfs:label ?title .
}
```

specifies namespaces

specifies output variables

specifies graph to be queried

specifies graph pattern to be matched

Reglas para Datos enlazados (Linked Data)

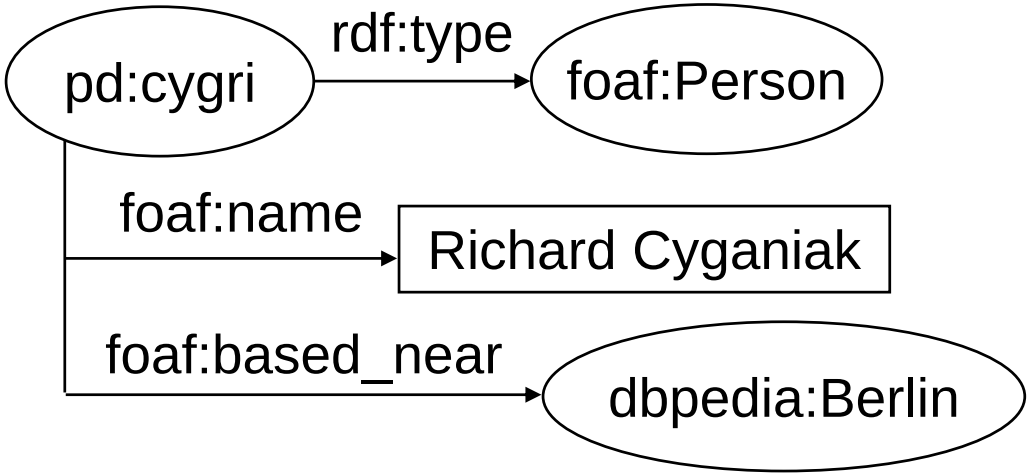
- Se debe permitir:
  - Seguir enlaces
  - Combinar la información guardada en las ontologías

- 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.

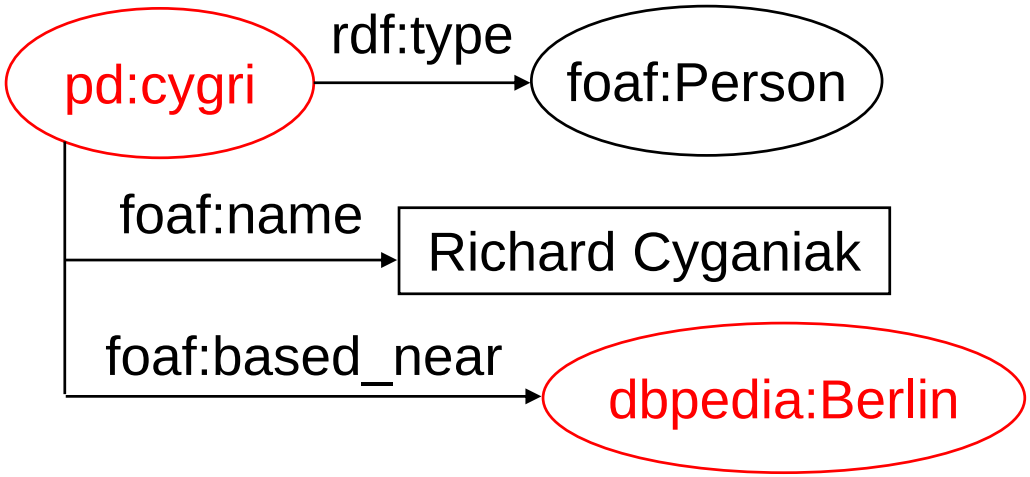
Propiedades de la Web de datos enlazados

- Cualquiera puede publicar datos
- Entidades están conectadas por enlaces
  - Un global grafo de datos que expande las fuentes de datos, descubriendo nuevas.
- Datos se auto-describen
  - Si una aplicación encuentra un dato con vocabulario no familiar, la aplicación, usando el URIs que identifica los términos del vocabulario, puede encontrar los RDFS o OWL con su definición.
- Esta Web es Open Data

Modelo RDF


pd:cygri
rdf:type
foaf:Person
foaf:name
Richard Cyganiak
foaf:based_near
dbpedia:Berlin

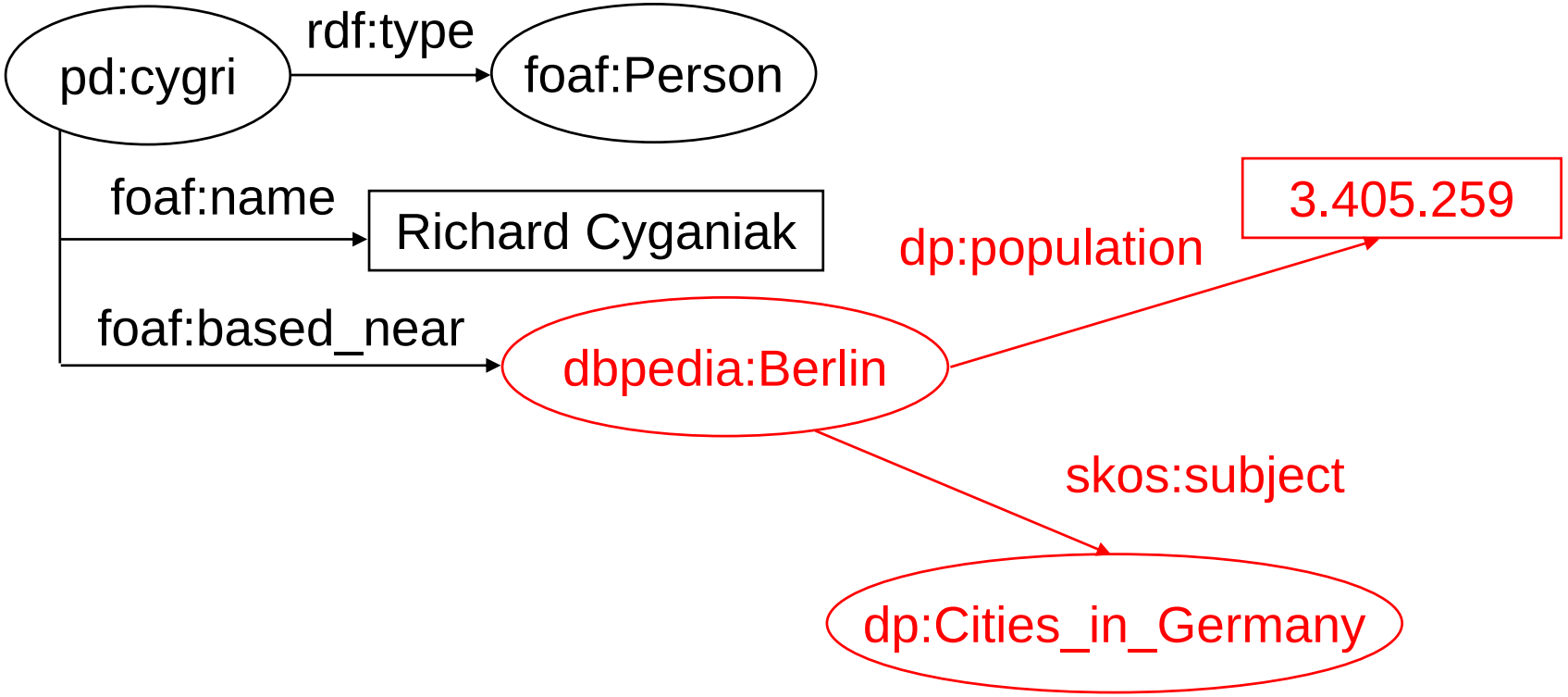
Datos son identificados con URIs http



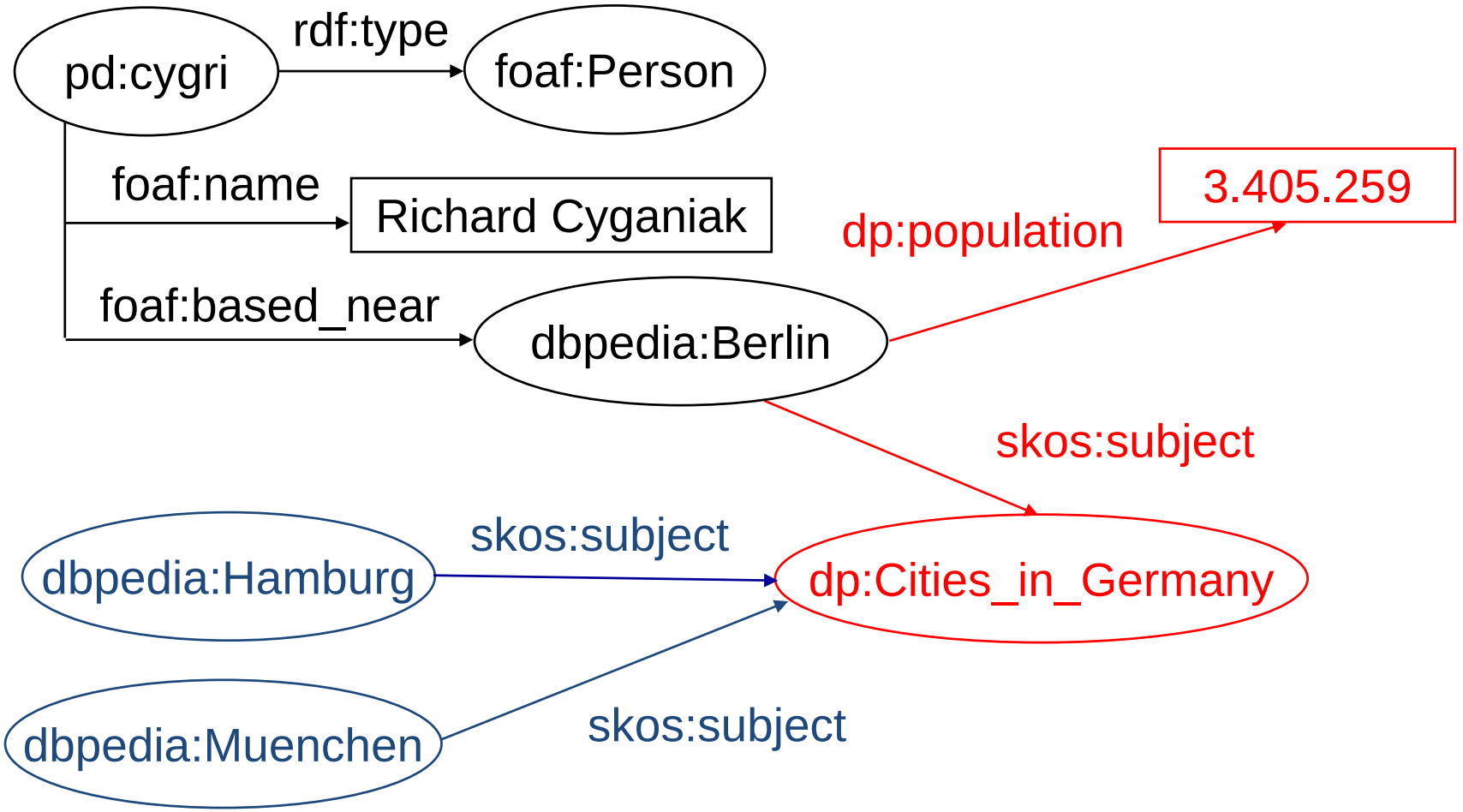
pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

dbpedia:Berlin = http://dbpedia.org/resource/Berlin

URIs en la Web

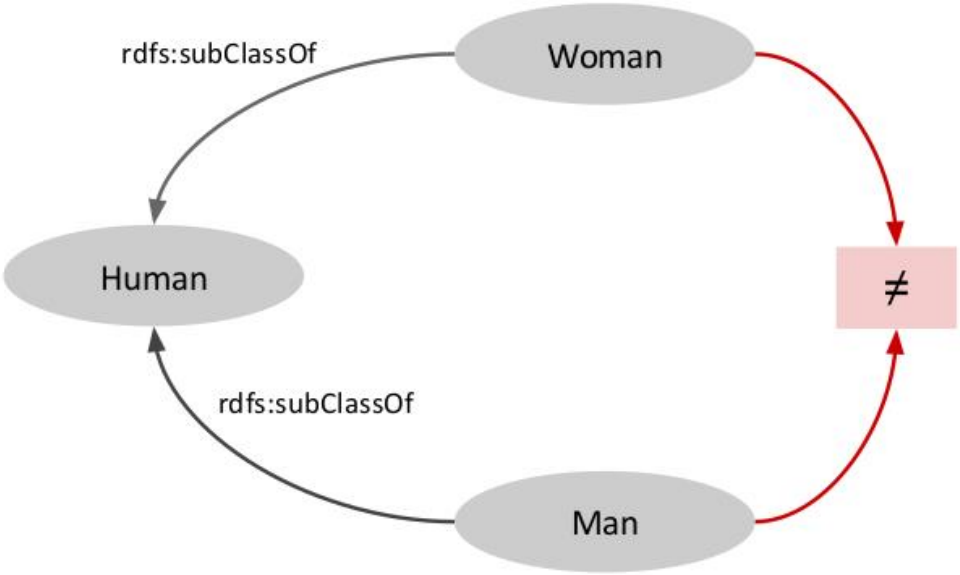


Enlazando URIs en la Web



RDFs no es suficiente...

- **Disjunctive Classes**



- RDFS Subclass relation cannot express disjunctive class (subclass) membership

Anotaciones Semánticas en la Web

En principio hay tres maneras de integrar datos estructurados con anotaciones semánticas explícitas dentro de documentos HTML

Microformatos (µFormat)

RDFa

HTML5 Microdatos (incluso schema.org)

FOAF

- Permite crear paginas Web para describir personas, vínculos entre ellos, y cosas que hacen y crean.
- Es un vocabulario RDF (http://xmlns.com/foaf/spec/) que permite tener disponible información personal de forma sencilla y simplificada para que pueda ser procesada, compartida y reutilizada.

FOAF: Conceptos fundamentales

- **Concepto básico**: <foaf:Person>
- **Propiedades simples**: name, title, familyName, nick
- **Enlaces web**: depiction (foto), homepage, workplaceHomepage
- **Relaciones entre personas**: <foaf:knows>.

Ejemplo

```
<?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>
                    Jose Aguilar
        </foaf:name>
         <foaf:mbox rdf:resourse=mailto:aguilar@ula.ve/>
     </foaf:Person>
</rdf:RDF>
```

Indicando Datos Externos

Para navegar por FOAF es por medio del esquema RDFS y su propiedad <rdfs:seeAlso>

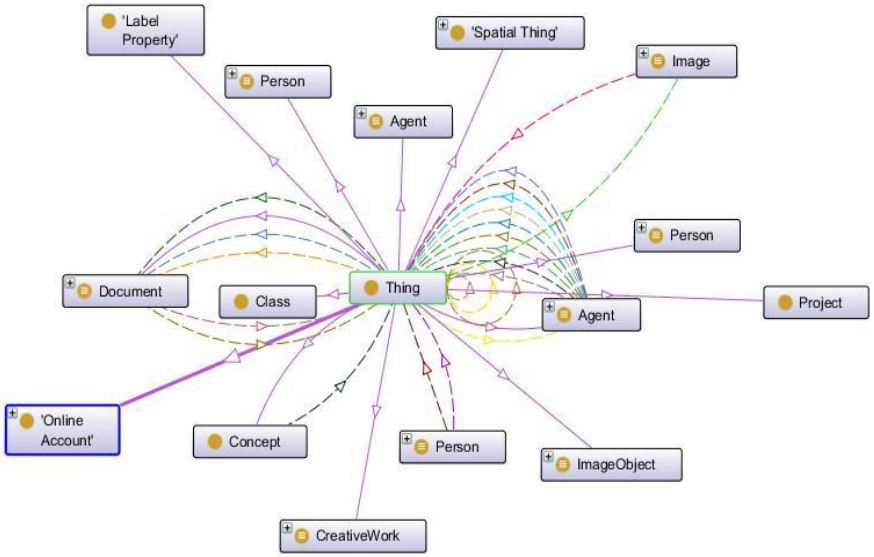
Expresando Relaciones

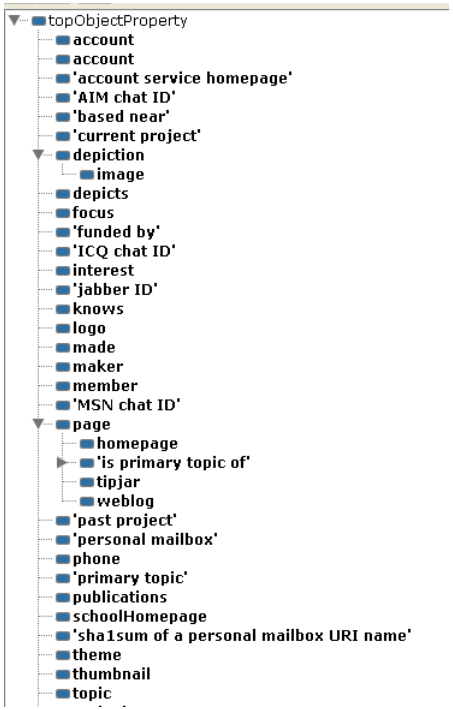
Para expresar que se conoce a alguien en FOAF es por medio de la propiedad Knows

```
<?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>
                    Jose Aguilar
        </foaf:name>
         <foaf:mbox rdf:resourse=mailto:aguilar@ula.ve/>
          <foaf:knows>
                    <foaf:Person>
                        <foaf:name> Taniana Rodriguez</foaf:name>
                        <foaf:mbox rdf:resourse=mailto:taniana@ula.ve/>
                        <rdfs:seeAlso rdf:resourse="paginawebTaniana/foaf.rdf"
                    </foaf:Person>
          </foaf:Knows>
     </foaf:Person>
</rdf:RDF>
```

Especificación de FOAF





Disco - Hyperdata Browser (About)

Richard Cyganiak

URI: http://richard.cyganiak.de/foaf.rdf#cygri Go!

Property	Value	Sources
event	...	G2
type	http://xmlns.com/foaf/0.1/Person	G1 G2 G3 G4
seeAlso	http://richard.cyganiak.de/cygri.rdf	G2
seeAlso	http://richard.cyganiak.de/foaf.rdf	G3
nearest airport	...	G1
phone	tel:+49-175-5630408	G1
sameAs	Richard Cyganiak	G1
based_near	...	G1
based_near	Berlin	G1
based_near	http://sws.geonames.org/2950159/	G1
currentProject	http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs	G3
currentProject	http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq	G3
depiction		G4
gender	male	G1
holdsAccount	cygri at del.icio.us	G1

Microformats

- Son XML tags que son incorporado dentro de las páginas Web para soportar declaraciones semánticas
- Enriquecen sitios webs con atributos, con el fin de hacer una declaración semántica, para los agentes de software.

✓ XFN (**X**HTML **F**riends **N**etwork, http://gmpg.org/xfn/), representa relaciones de personas usando hyperlinks.

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

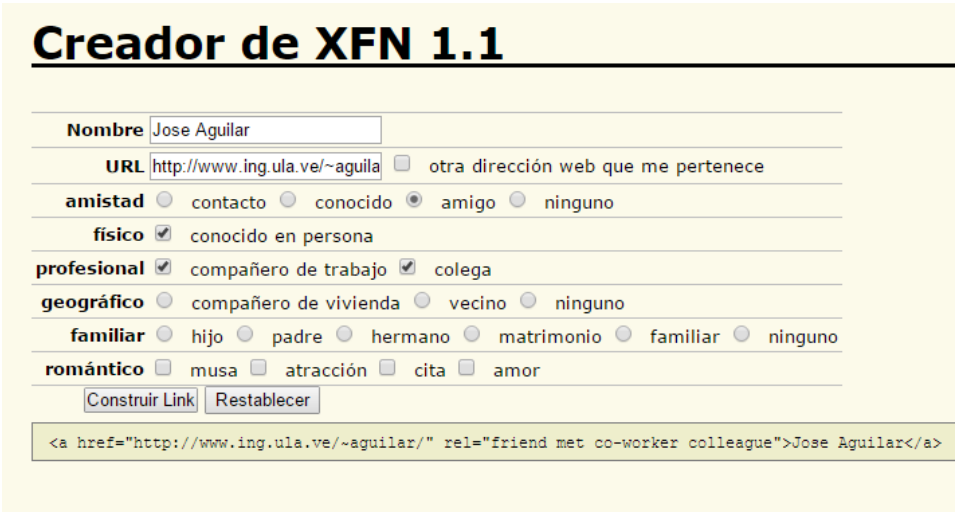
Por ejemplo:

✓ Supongamos que en la página Web de Taniana Rodríguez tiene un enlace a la página Web del Jose Aguilar

- ¿Ellos son amigo?
- ¿Ellos trabajan juntos?
- ....

✓ Si se añade explícitamente en la página Web de Taniana

- Jose Aguilar
- indica que Taniana y Jose trabajan juntos (friend como co-worker están definido en XFN microformat)



Microformats

- hCard (http://microformats.org/wiki/hcard) es un microformato que permite marcar los datos de cualquier persona o entidad

Properties

Common hCard properties (inside class `vcard`)

- `fn` - name, formatted/full. required
- `n` - structured name, container for:
  - `honorific-prefix` - e.g. Ms., Mr., Dr.
  - `given-name` - given (often first) name
  - `additional-name` - other/middle name
  - `family-name` - family (often last) name
  - `honorific-suffix` - e.g. Ph.D., Esq.
- `nickname` - nickname/alias, e.g. IRC nick
- `org` - company/organization
- `photo` - photo, icon, avatar
- `url` - home page for this contact
- `email` - email address
- `tel` - telephone number
- `adr` - structured address, container for:
  - `street-address` - street #+name, apt/ste
  - `locality` - city or village
  - `region` - state or province
  - `postal-code` - postal code, e.g. U.S. ZIP
  - `country-name` - country name
- `bday` - birthday. ISO date.
- `category` - for tagging contacts
- `note` - notes about the contact

```
<div class="vcard">
   <span class="fn">Sally Ride</span>
   (<span class="n">
      <span class="honorific-prefix">Dr.</span>
      <span class="given-name">Sally</span>
      <abbr class="additional-name">K.</abbr>
      <span class="family-name">Ride</span>
      <span class="honorific-suffix">Ph.D.</span></span>),
   <span class="nickname">sallykride</span> (IRC)
   <div class="org">Sally Ride Science</div>
   <img class="photo" src="http://example.com/sk.jpg"/>
   <a class="url" href="http://sally.example.com">w</a>,
   <a class="email" href="mailto:sally@example.com">e</a>
   <div class="tel">+1.818.555.1212</div>
   <div class="adr">
      <div class="street-address">123 Main st.</div>
      <span class="locality">Los Angeles</span>,
      <abbr class="region" title="California">CA</abbr>,
      <span class="postal-code">91316</span>
      <div class="country-name">U.S.A</div></div>
   <time class="bday">1951-05-26</time> birthday
   <div class="category">physicist</div>
   <div class="note">1st American woman in space.</div>
</div>
```

hCard Creator

hCard-o-matic

given name	Taniana Rodríguez
middle name	
family name	Taniana Josefina Rodrig
organization	Universiad de Los Andes
street	
city	Mérida
state/province	Mérida
postal code	5101
country name	Venezuela
phone	
email	taniana@ula.ve
url	http://taniana.novacorp.co/
photo url	
AIM screenname	
YIM screenname	
Jabber screenname	
Categories (comma separated)	

Restablecer Build It!

Warning - publishing your email address, phone number or instant messenger screenname on the web can open it up to abuse.

code

```
<div id="hcard-Taniana-Rodriguez-Taniana-Josefina-Rodriguez-de-Paredes" class="vcard">
<a class="url fn n" href="http://taniana.novacorp.co/"> <span class="given-name">Taniana Rodríguez</span>
 <span class="additional-name"></span>
 <span class="family-name">Taniana Josefina Rodríguez de Paredes</span>
</a>
<div class="org">Universiad de Los Andes</div>
<a class="email" href="mailto:taniana@ula.ve">taniana@ula.ve</a>
<div class="adr">
<span class="locality">Mérida</span>
```

preview

Taniana Rodríguez Taniana Josefina Rodríguez de Paredes
Universiad de Los Andes
taniana@ula.ve
Mérida , Mérida , 5101 Venezuela

This hCard created with the hCard creator.

Microformats

- hcalendar (http://microformats.org/wiki/hcalendar), es un estándar de microformat de la información de un evento, en formato iCalendar

Property List

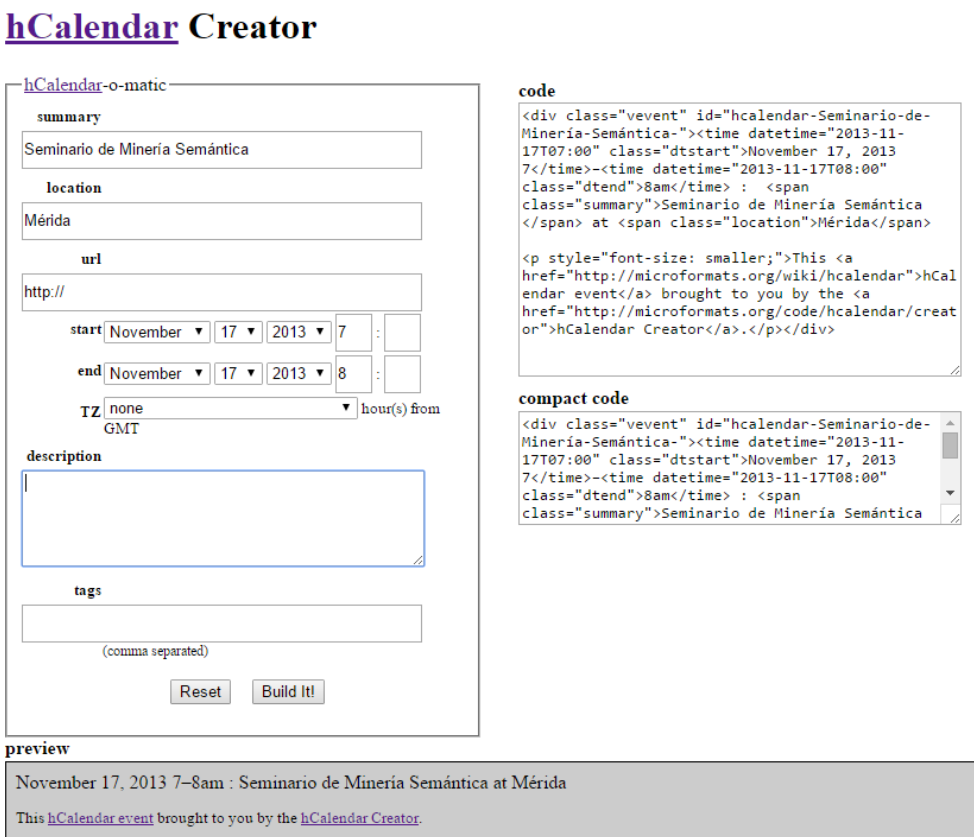
hCalendar properties (sub-properties in parentheses like this)

Required:

- **dtstart** (ISO date)
- **summary**

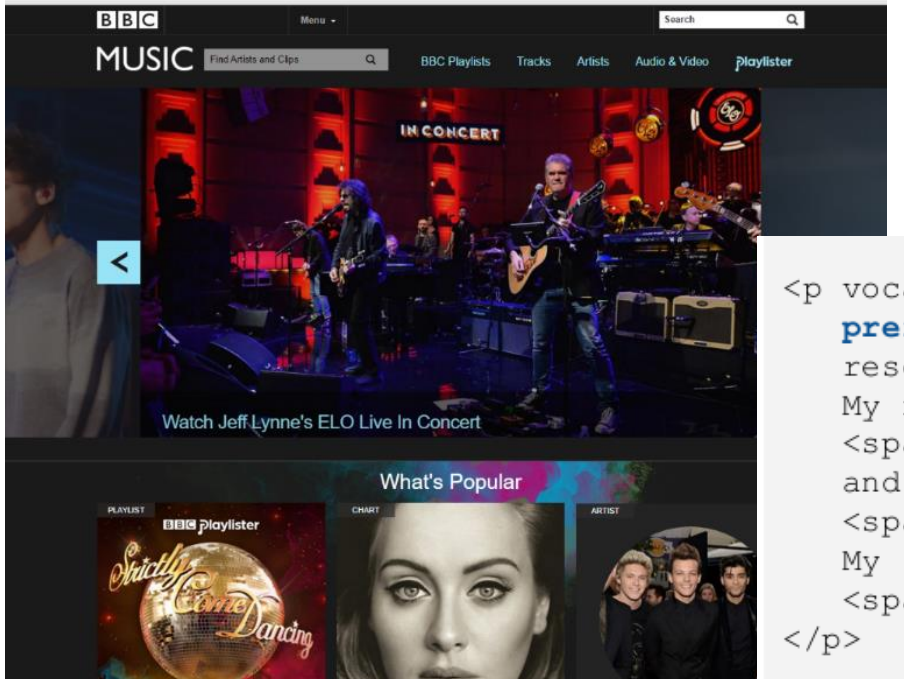
Optional:

- location
- url
- dtend (ISO date), duration (ISO date duration)
- rdate, rrule
- category, description
- uid
- geo (latitude, longitude)
- attendee (partstat, role), contact, organizer
- attach
- status
- editor's note: this list is incomplete (an incomplete list is better



RDFa

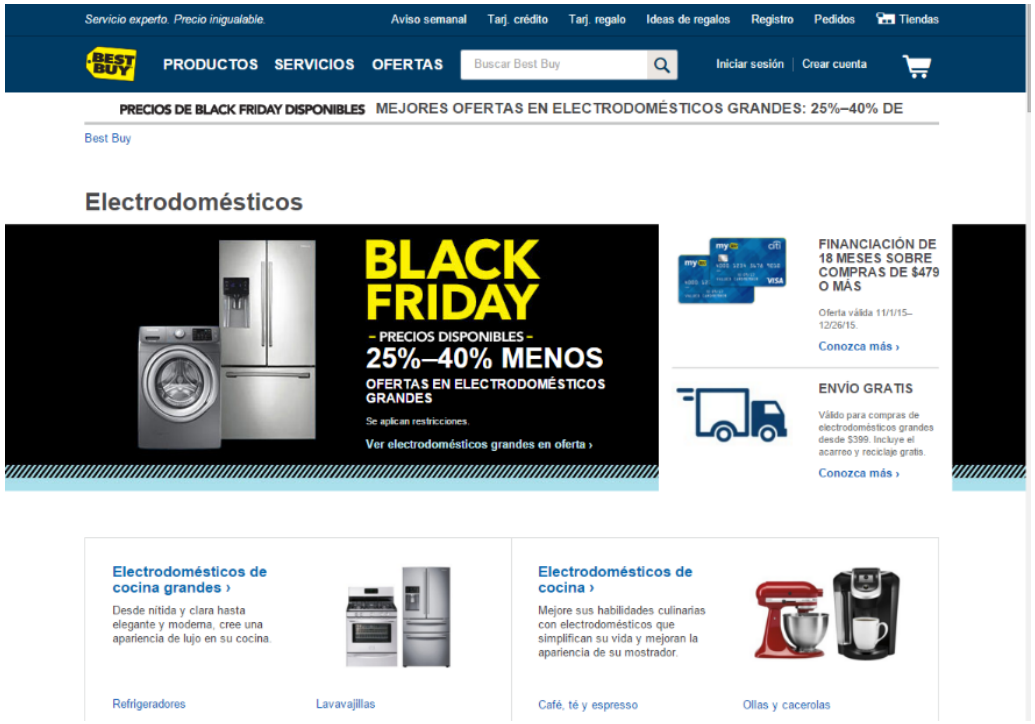
- RDFa añade semántica a las páginas Web
- RDFa define atributos para las palabras o frases que pueden ser tratadas como entidades semánticas

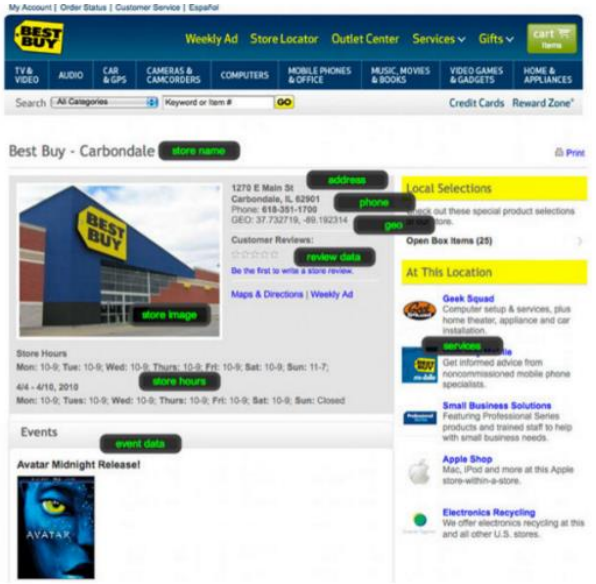


W3C RDFa

```
<p vocab="http://xmlns.com/foaf/0.1/"
   prefix="ov: http://open.vocab.org/terms/"
   resource="#harald" typeof="Person">
   My name is
   <span property="name">Harald Sack</span>
   and you can give me a ring via
   <span property="phone">1-800-555-0527</span>.
   My favorite beverage is
   <span property="ov:preferredBeverage">Green Tea</span>
</p>
```

RDFa ejemplo





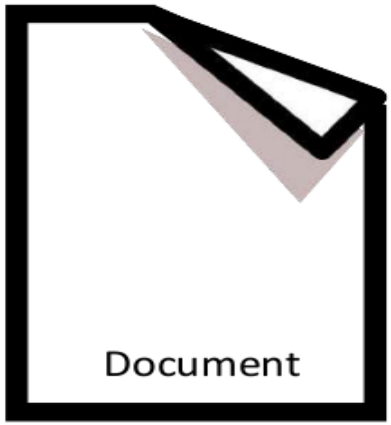
Best Buy employees entered information into the blogs every day, using online forms that output RDFa. Myers told us that the use of RDFa makes "human input from our store employees more visible on the Web."

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

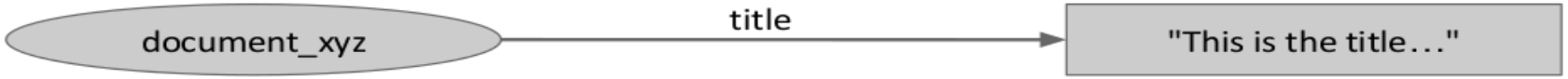
Vocabularios

Los vocabularios definen las relaciones y términos utilizados para describir y representar un área específica.

- The **Dublin Core Vocabulary** (Dublin Core Metadata Element Set)



- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights




```
PREFIX dct: <http://purl.org/dc/terms/> .

<http://example.org/document_xyz> dct:title "This is the title..."@en .
```

Aplicaciones Web con Linked Data

- SPARQL Javascript Library
  http://www.thefigtrees.net/lee/blog/2006/04/sparql_calendar_demo_a_sparql.html
- ARC for SPARQL (PHP)
  https://github.com/semsol/arc2/wiki
- dotNetRDF (C#)
  https://dotnetrdf.github.io/
- Jena/ARQ (Java)
  http://jena.apache.org/
- Sesame (Java)
  http://rdf4j.org/
- **SPARQL Wrapper (Python)**
  http://rdflib.github.io/sparqlwrapper/

Aplicaciones Web con Linked Data

Amsterdam

The Keizersgracht at dusk

Location of Amsterdam

Coordinates: 52°22'23"N 4°53'32"E

Country	Netherlands
Province	North Holland
Government	
- Type	Municipality
- Mayor	Job Cohen[1] (PvdA)
- Aldermen	Lodewijk Asscher Carolien Gehrels Tjeerd Herrema Maarten van Poelgeest Marijke Vos
- Secretary	Erik Gerritsen
Area [2][3]	
- City	219 km² (84.6 sq mi)
- Land	166 km² (64.1 sq mi)
- Water	53 km² (20.5 sq mi)
- Urban	1,003 km² (387.3 sq mi)
- Metro	1,815 km² (700.8 sq mi)
Elevation [4]	2 m (7 ft)
Population (1 October 2008)[5][6]	
- City	755,269
- Density	4,459/km² (11,548.8/sq mi)
- Urban	1,364,422
- Metro	2,158,372
- Demonym	Amsterdammer
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postcodes	1011 – 1109
Area code(s)	020

Website: www.amsterdam.nl

Extraer estructurada información desde Wikipedia

```
@prefix dbpedia
<http://dbpedia.org/resource/>.
@prefix dbterm
<http://dbpedia.org/property/>.

dbpedia:Amsterdam
  dbterm:officialName "Amsterdam" ;
  dbterm:longd "4" ;
  dbterm:longm "53" ;
  dbterm:longs "32" ;
  ...
  dbterm:leaderName dbpedia:Job_Cohen ;
  ...
  dbterm:areaTotalKm "219" ;
  ...
dbpedia:ABN_AMRO
  dbterm:location dbpedia:Amsterdam ;
  ...
```

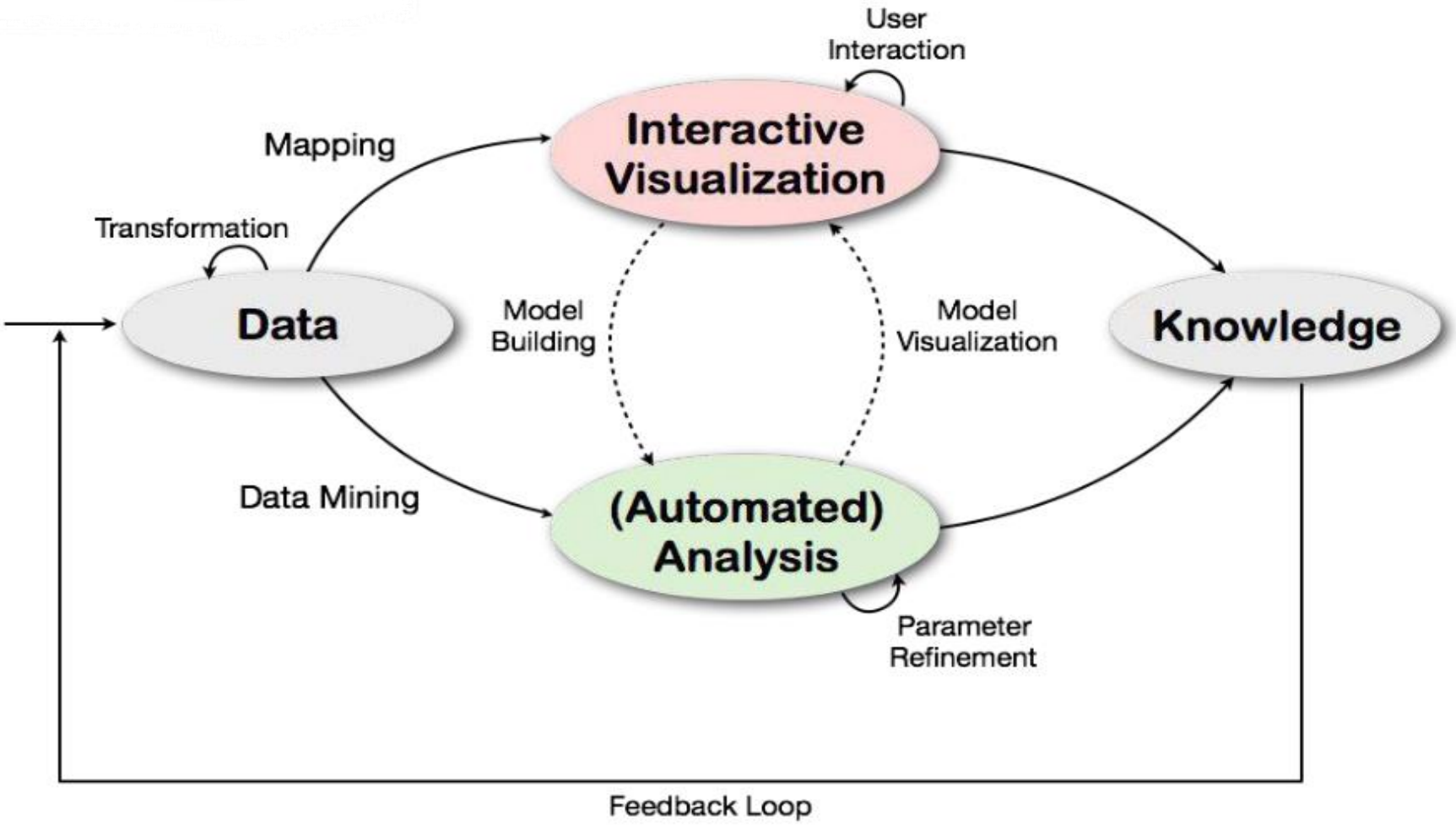
Enlazado automático

```
<http://dbpedia.org/resource/Amsterdam>
  owl:sameAs <http://rdf.freebase.com/ns/...> ;
  owl:sameAs <http://sws.geonames.org/2759793> ;
  ...
```

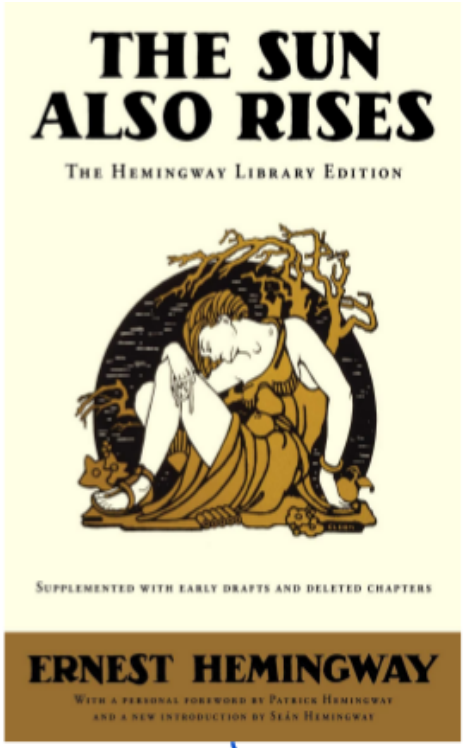
```
<http://sws.geonames.org/2759793>
  owl:sameAs <http://dbpedia.org/resource/Amsterdam>
  wgs84_pos:lat "52.3666667" ;
  wgs84_pos:long "4.8833333" ;
  geo:inCountry <http://www.geonames.org/countries/#NL> ;
 ...
```

Computador puede saltar automaticamente desde una a otra…

Análisis en Linked Data



Metadata y Anotación Semantica



La semantica es explicitamente definida con

Una ontologia que puede ser leida

automaticamente por el computador,

Basado en la tripleta

sujeto, objeto, predicado

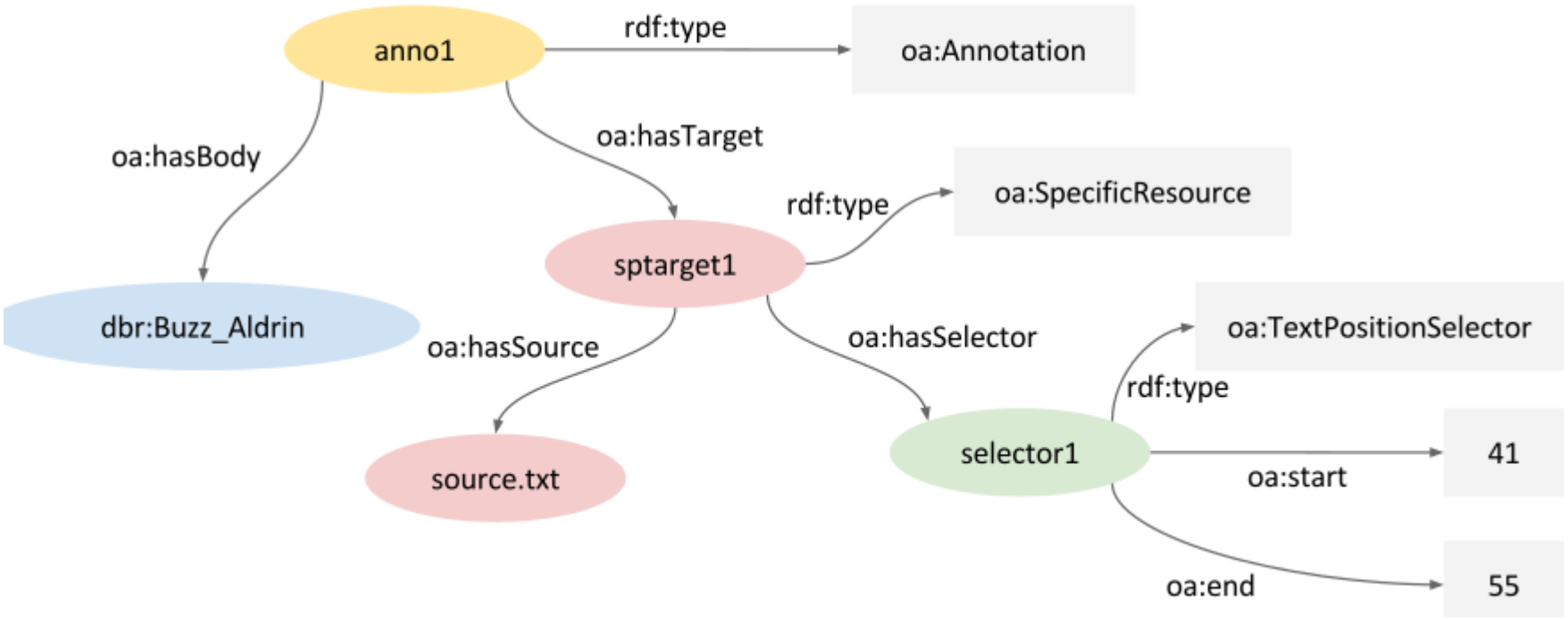
```
PREFIX dbr: <http://dbpedia.org/resource/> .
PREFIX dbo: <http://dbpedia.org/ontology/> .

:9787532717071 dbo:author dbr:Ernest_Hemingway .
```

Metadata y Anotación Semantica

On July 16, 1969, Armstrong, along with Edwin E. Aldrin, Jr., and Michael Collins, blasted off in the Apollo 11 vehicle toward the Moon.

Anotaciones Semánticas (Texto)



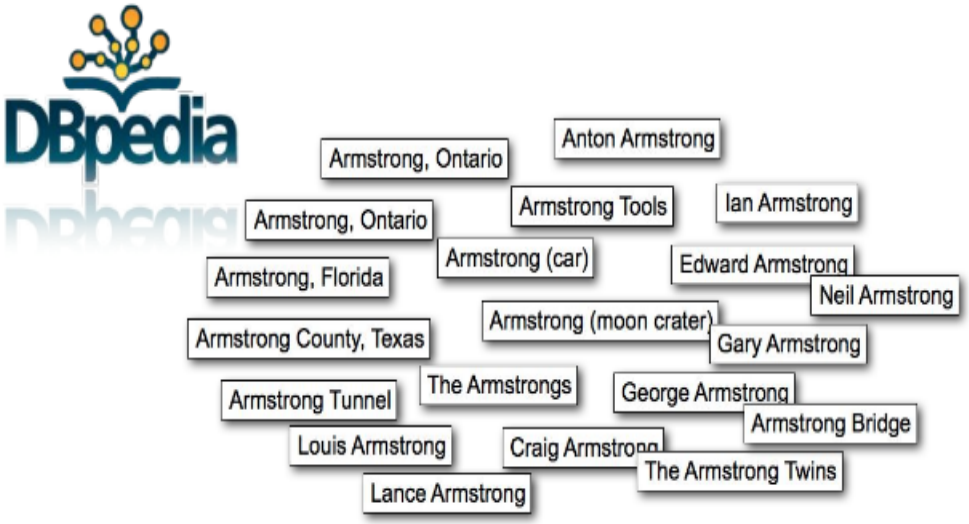
Metadata y Anotación Semantica

Named Entity Resolution

Named Entity Resolution

Armstrong landed the Eagle on the Moon

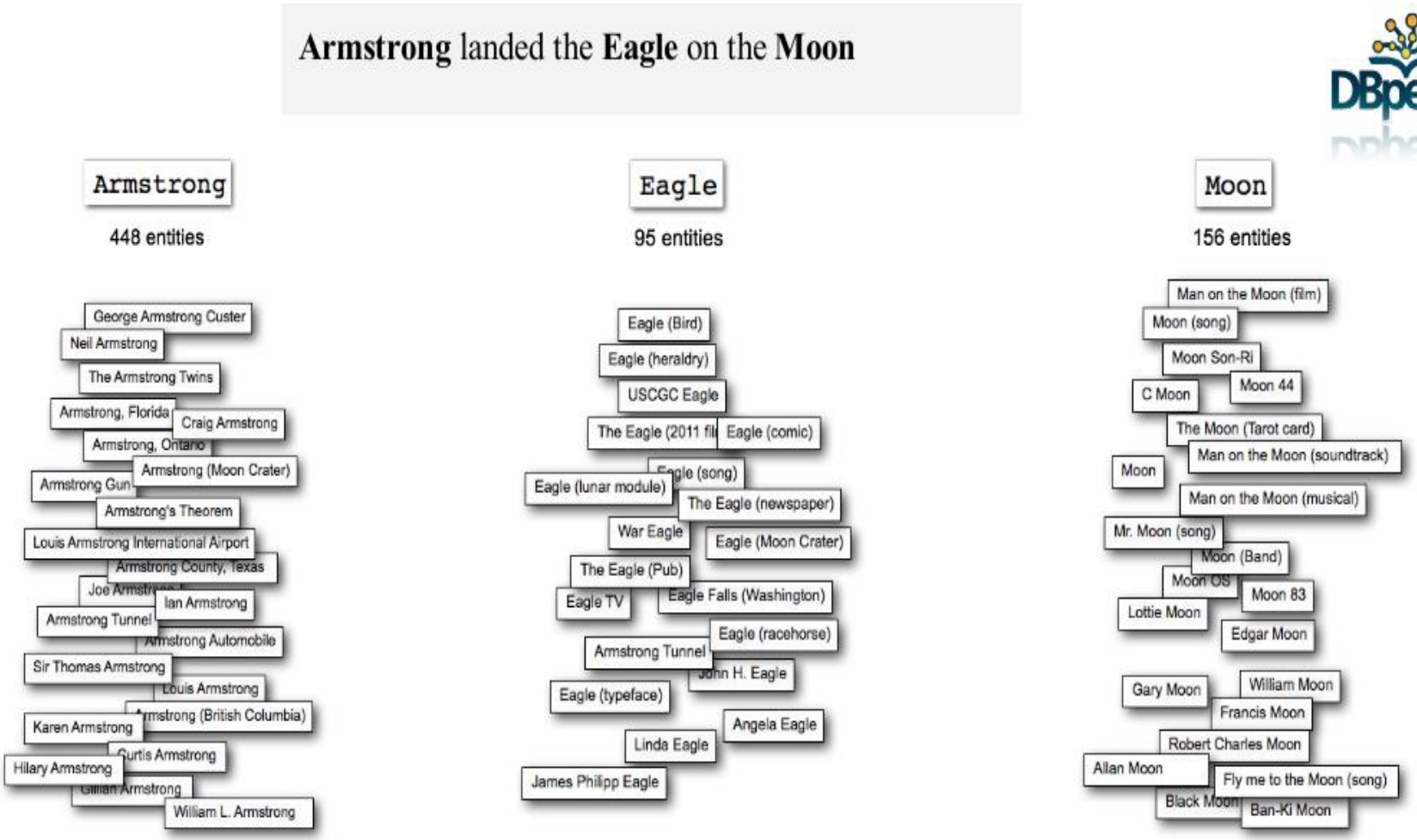
Determine all possible Entity Candidates



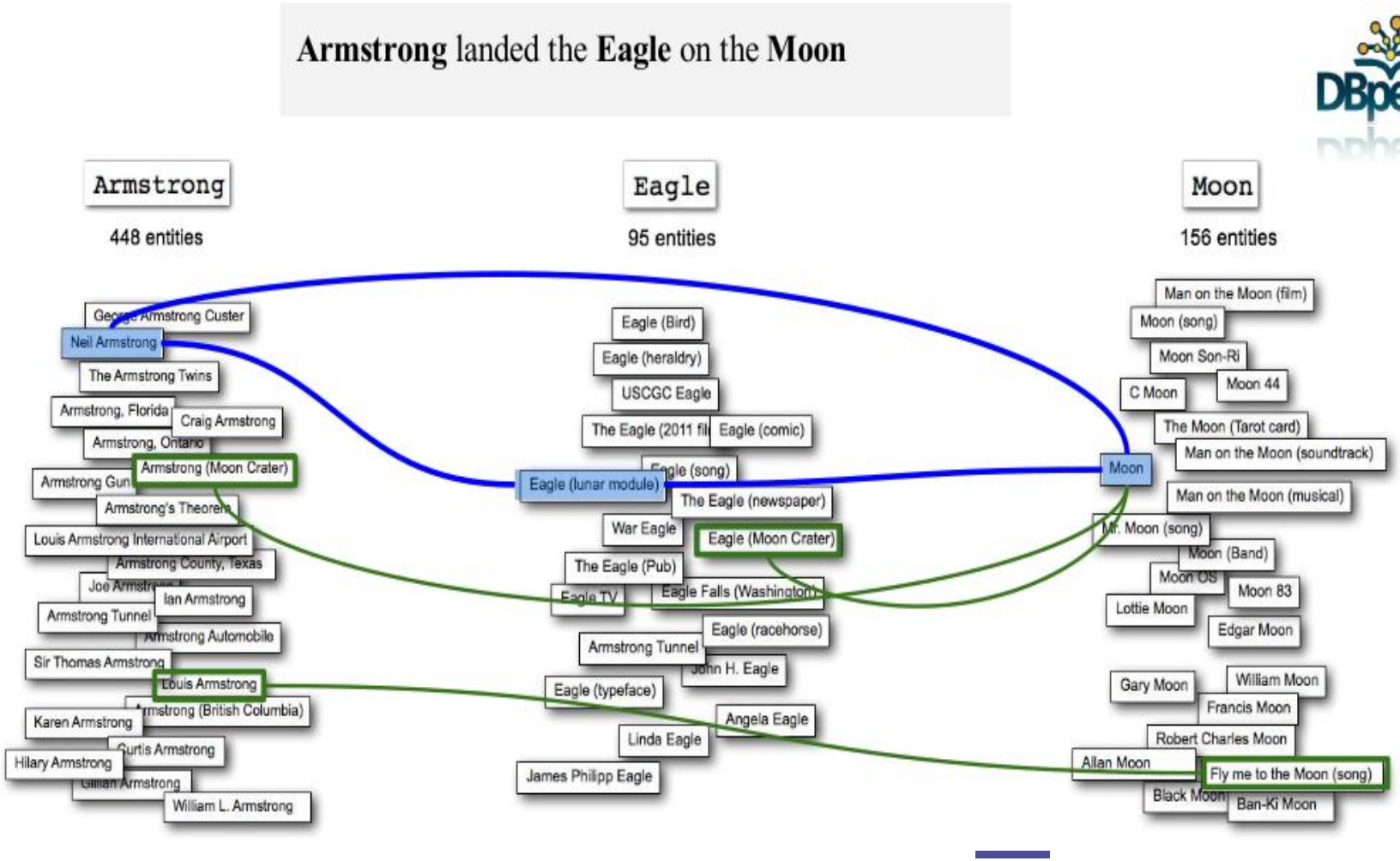
- linguistic analysis (POS tagging)
- normalization
- encoding and spelling
- special (language dependent) characters
- language dependent spellings
- abbreviations, acronyms
- type dependent spellings
- alternative names and synonyms
- fuzzy string mapping
- ...

Named Entity Resolution

Armstrong landed the **Eagle** on the **Moon**

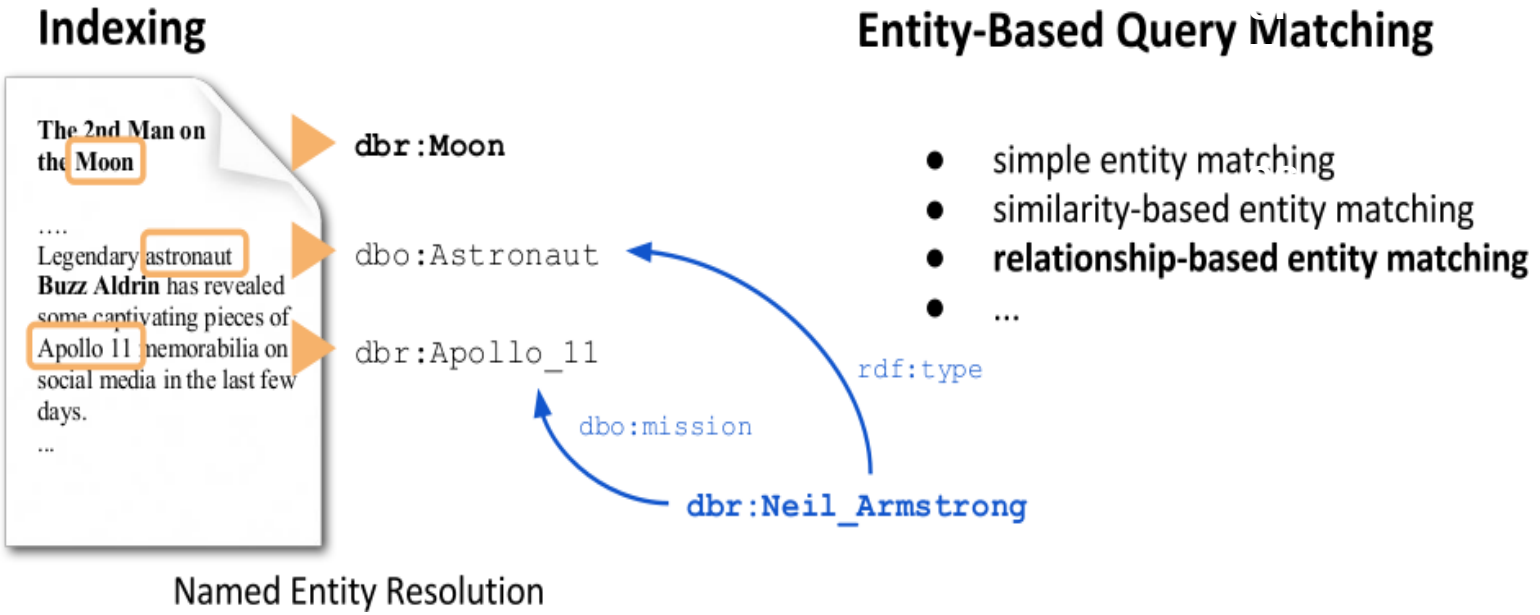


Named Entity Resolution



Búsqueda Semántica

Exploración de la información



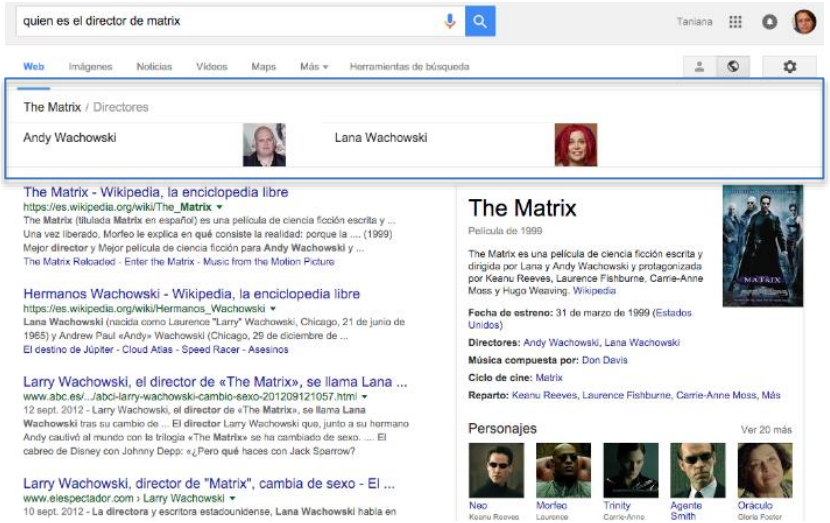
SEO (Search Engine Optimization)

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

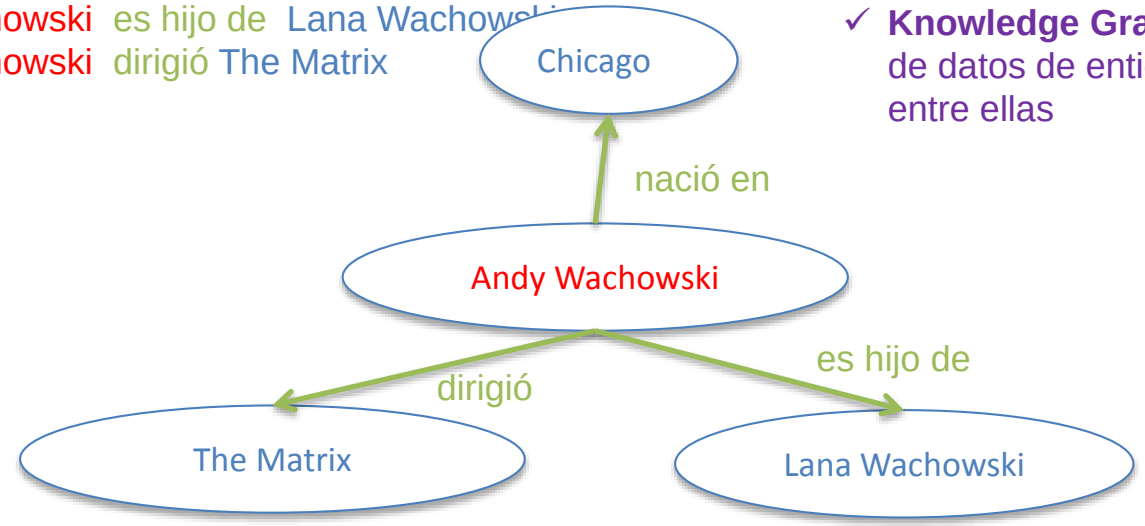
- Información puede ser extraída de diferentes fuentes: Dbpedia, IMDB, Wikipedia etc.
- Basado en una representación del conocimiento

Sujeto + Predicado +Objeto

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

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

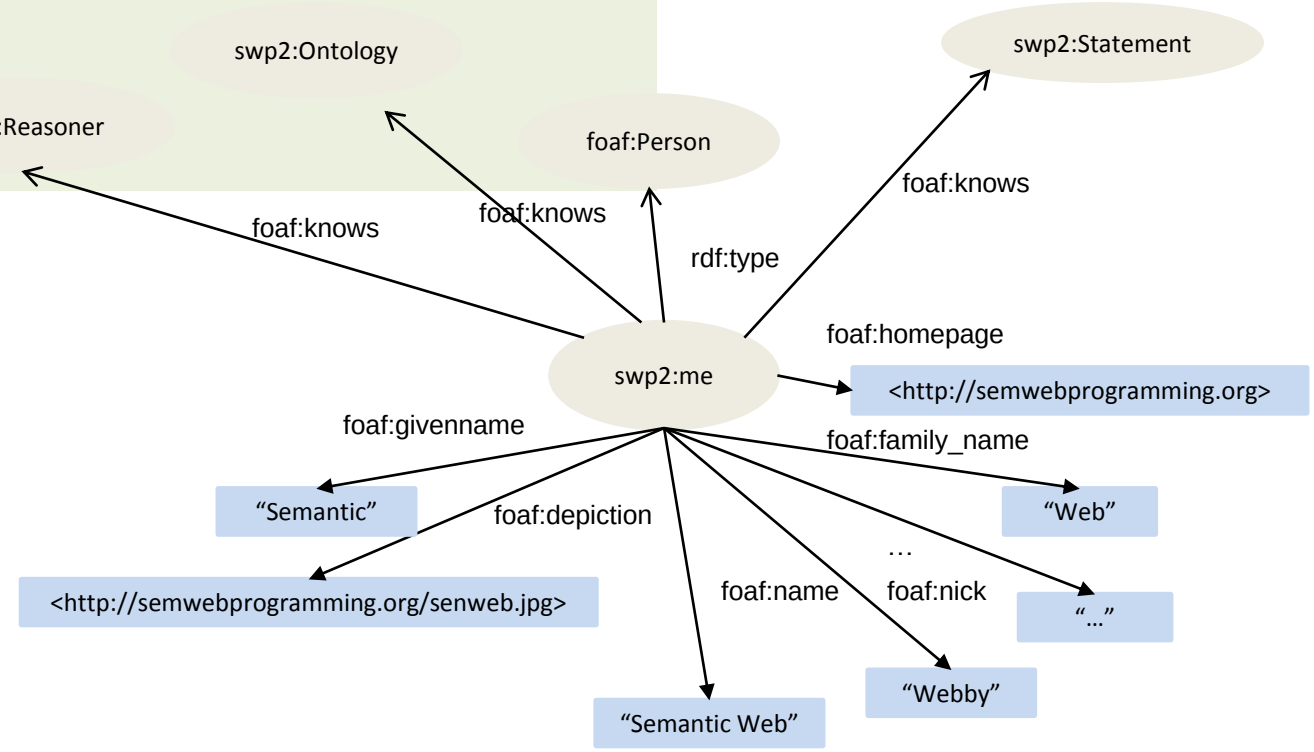
✓SEO semántico tiene como objetivo de ayudar a los buscadores a entender exactamente de qué trata tus páginas.

✓ Para ello, sigue los siguientes pasos

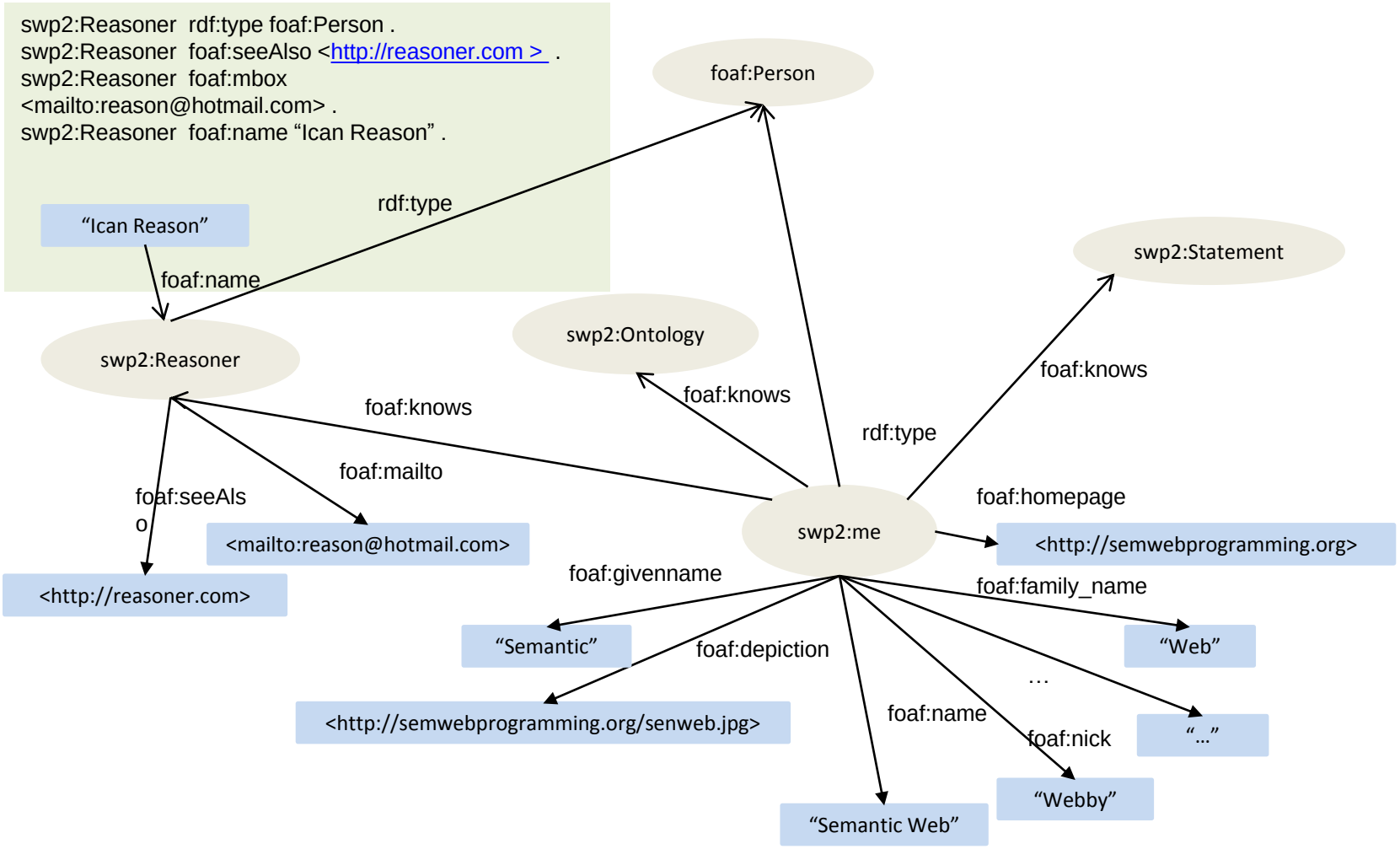
- ➢Determinar las entidades correspondientes a la página.
- ➢Desambiguarlas directamente
- ➢Desambiguarlas indirectamente.

Ejemplo

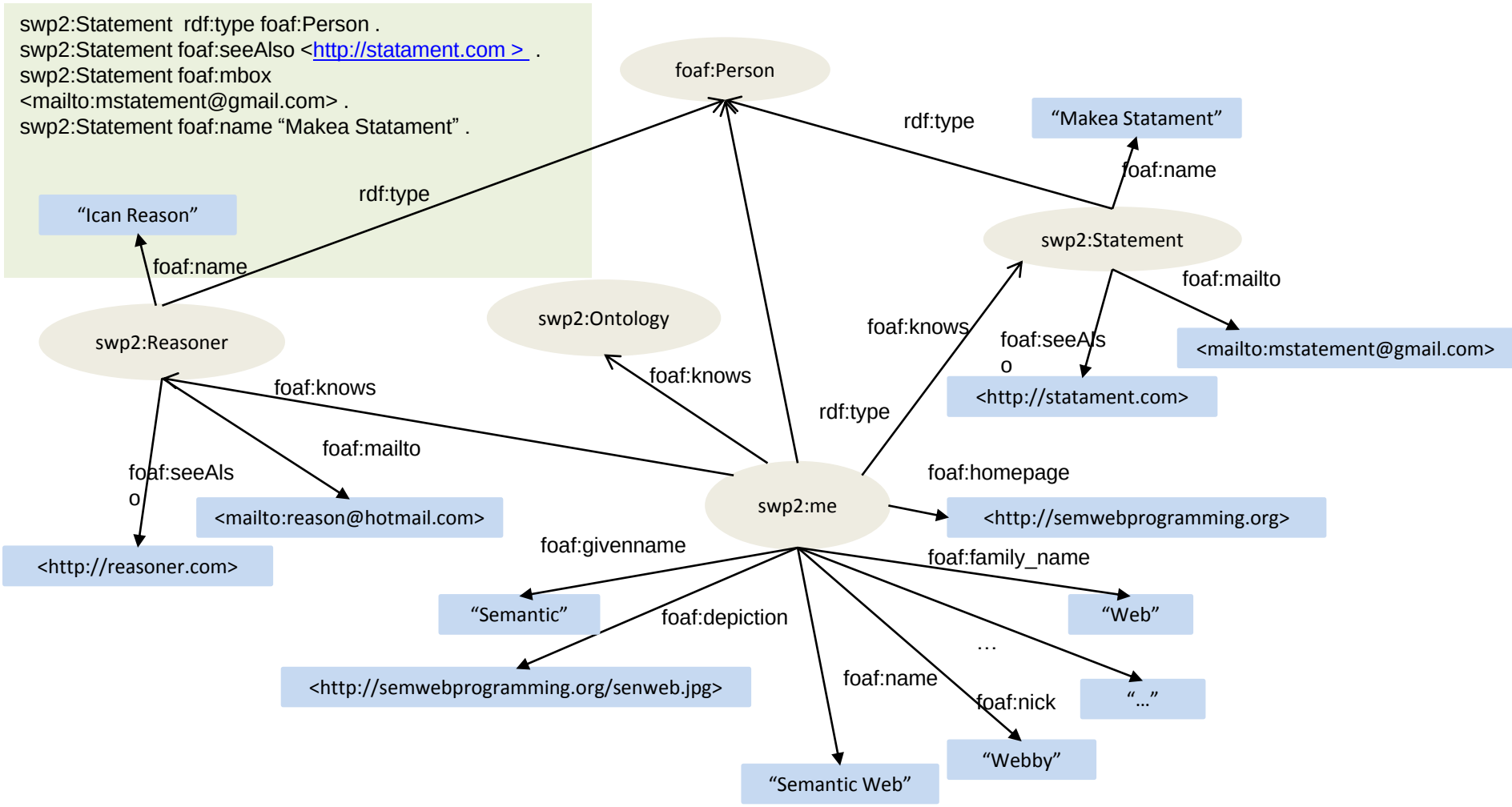
```
swp2:me  rdf:type foaf:Person .
swp2:me  foaf:depiction <http://semwebprogramming.org/senweb.jpg >.
swp2:me  foaf:family_name  “Web” .
swp2:me  foaf:givenname  “Semantic” .
swp2:me  foaf:homepage  <http://semwebprogramming.org >.
swp2:me  foaf:knows  “Reasoner” .
swp2:me  foaf:knows “Statement” .
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

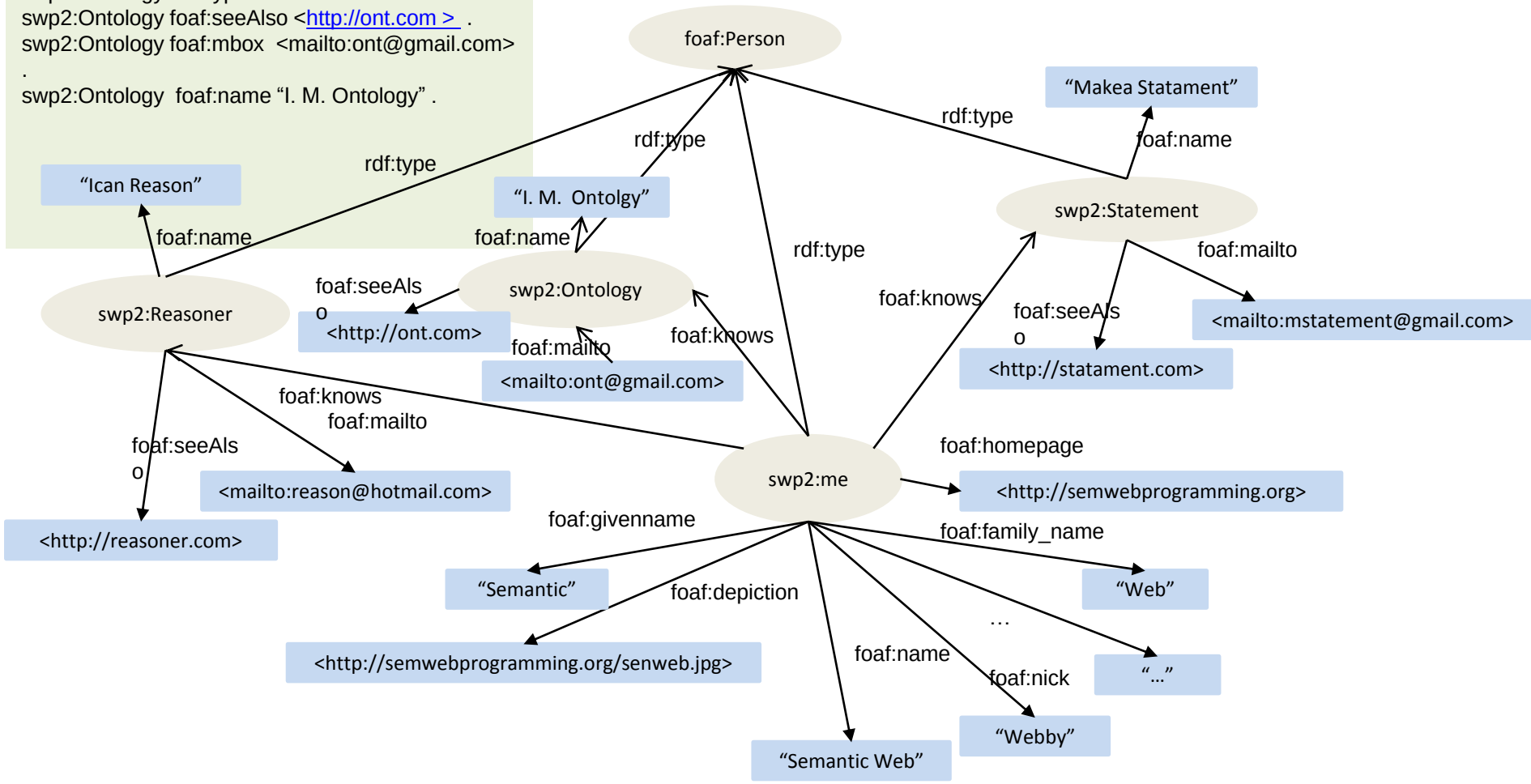

swp2:Reasoner rdf:type foaf:Person .
swp2:Reasoner foaf:seeAlso <http://reasoner.com > .
swp2:Reasoner foaf:mbox
<mailto:reason@hotmail.com> .
swp2:Reasoner foaf:name “Ican Reason” .
foaf:Person
rdf:type
“Ican Reason”
foaf:name
swp2:Statement
swp2:Ontology
swp2:Reasoner
foaf:knows
foaf:knows
foaf:knows
rdf:type
foaf:mailto
foaf:seeAlso
foaf:homepage
<mailto:reason@hotmail.com>
swp2:me
<http://semwebprogramming.org>
<http://reasoner.com>
foaf:givenname
foaf:family_name
“Semantic”
foaf:depiction
“Web”
…
<http://semwebprogramming.org/senweb.jpg>
foaf:name
foaf:nick
“…”
“Webby”
“Semantic Web”

Continuación del Ejemplo



Continuación del Ejemplo

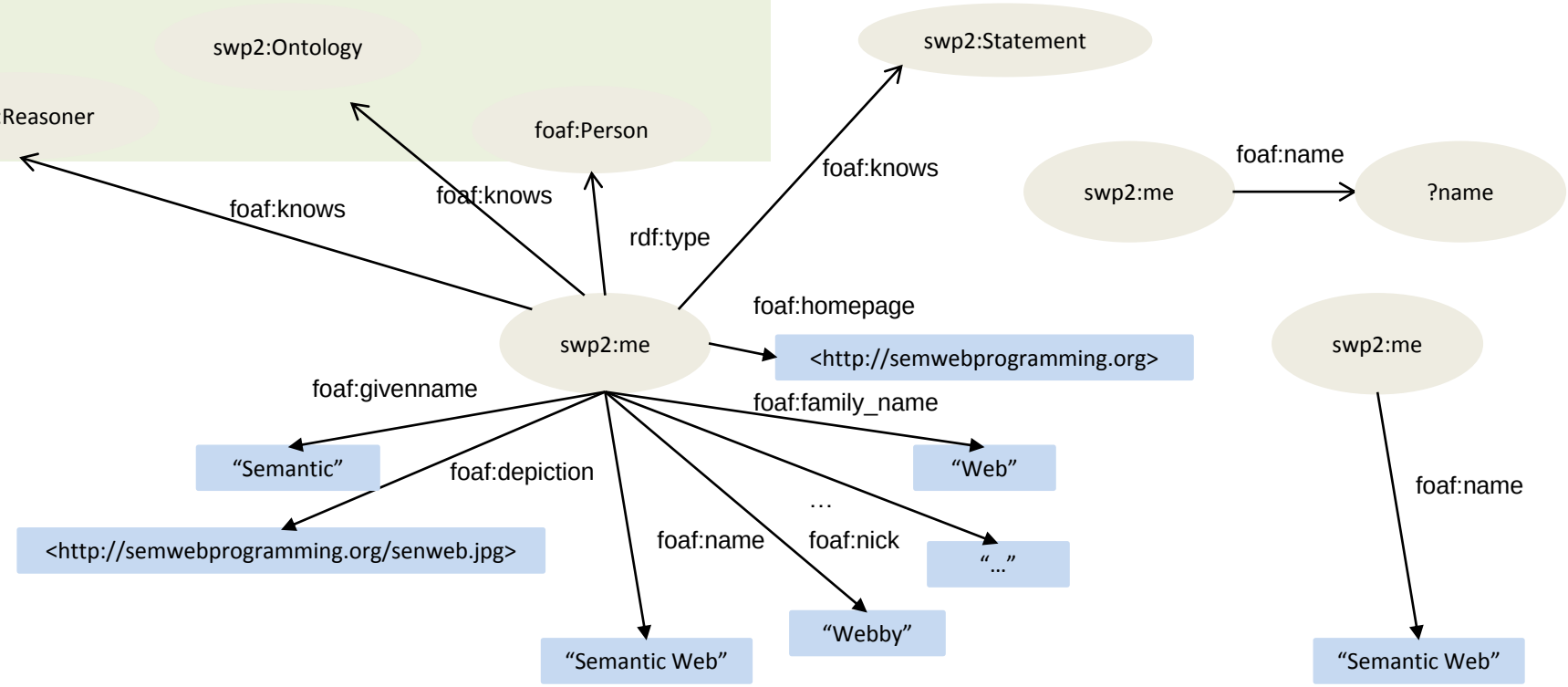
swp2:Ontology rdf:type foaf:Person .
swp2:Ontology foaf:seeAlso <http://ont.com > .
swp2:Ontology foaf:mbox <mailto:ont@gmail.com> .
swp2:Ontology foaf:name “I. M. Ontology” .



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
}
```



Sistemas Recomendadores

http://dbpedia.org/resource/From_the_Earth_to_the_Moon

DBpedia Browse using Formats Faceted Browser Sparql Endpoint

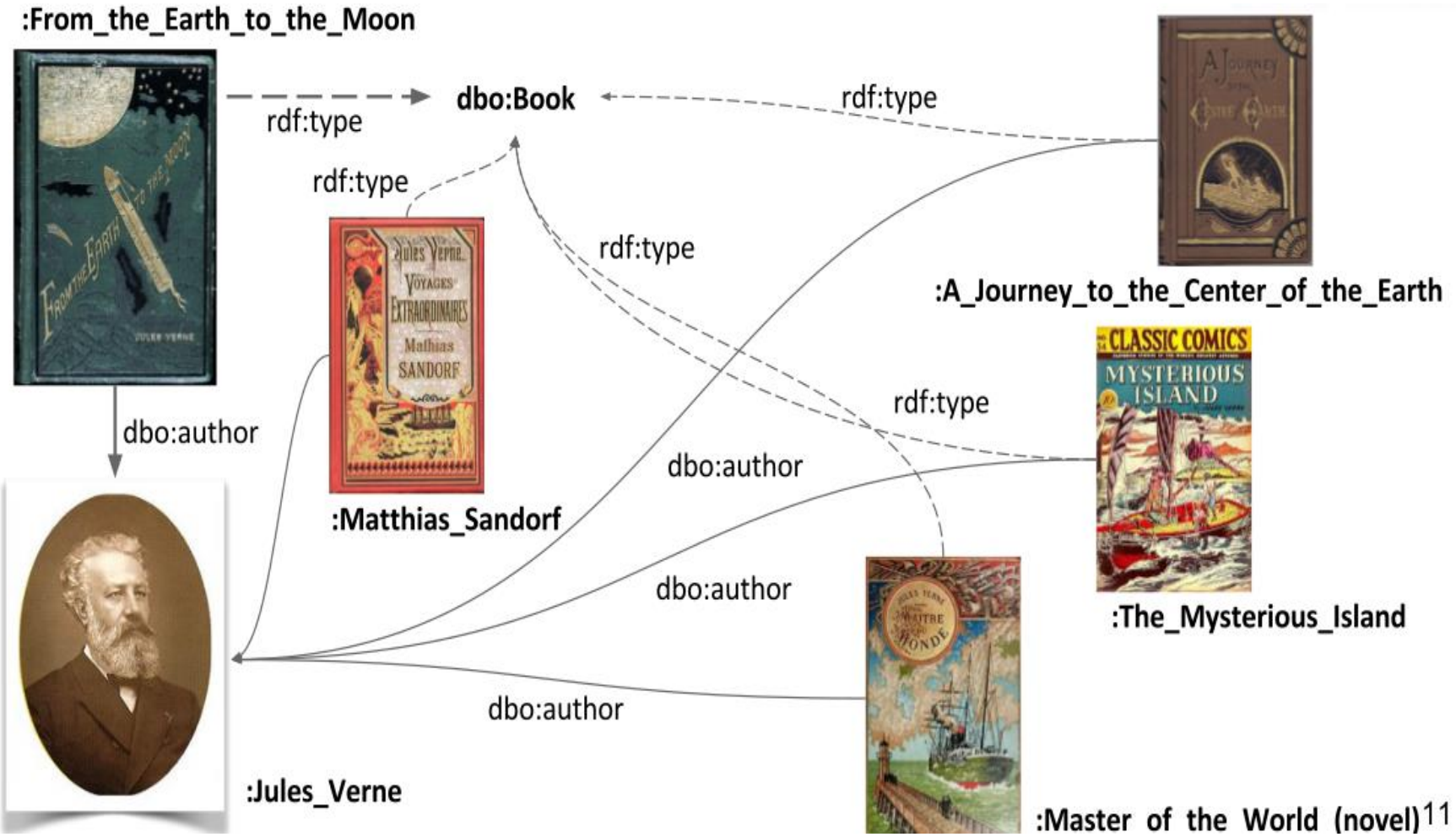
About: From the Earth to the Moon

An Entity of Type : work, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

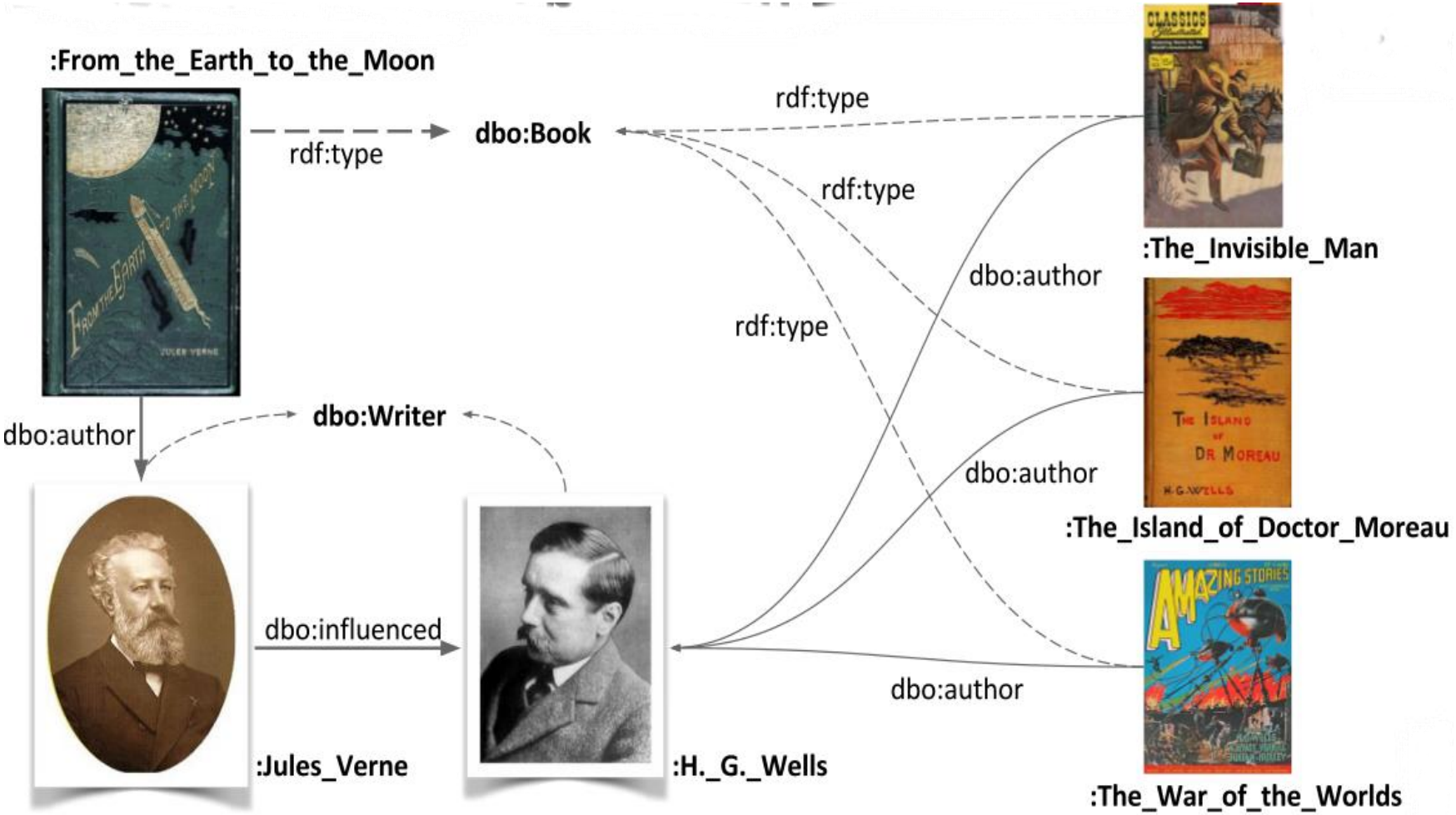
From the Earth to the Moon (French: De la terre à la lune) is an 1865 novel by Jules Verne.

Property	Value
dbo:abstract	• Von der Erde zum Mond ist ein Roman des französischen Autors Jules Verne. Der Roman wurde erstmals 1865 unter dem französischen Titel De la Terre à la Lune von dem Verleger Pierre-Jules Hetzel veröffentlicht. Die erste deutschsprachige Ausgabe erschien 1873 unter dem Titel Von der Erde zum Mond. Der englische Titel des Romans lautet From the Earth to the Moon.Es handelt sich um ein frühes Werk des Science-Fiction-Genres, das die Mondfahrt um etwa hundert Jahre vorwegnimmt. Allerdings geht es hier vor allem noch um die Vorbereitung des Abenteuers. Der Roman Reise um den Mond (Autour de la Lune) von 1870 setzte die Geschichte fort. (de) • From the Earth to the Moon (French: De la terre à la lune) is an 1865 novel by Jules Verne. It tells the story of the Baltimore Gun Club, a post-American Civil War society of weapons enthusiasts, and their attempts to build an enormous sky-facing Columbiad space gun and launch three people—the Gun Club's president, his Philadelphian armor-making rival, and a French poet—in a projectile with the goal of a moon landing.The story is also notable in that Verne attempted to do some rough calculations as to the requirements for the cannon and, considering the comparative lack of any data on the subject at the time, some of his figures are surprisingly close to reality. However, his scenario turned out to be impractical for safe manned space travel since a much longer muzzle would have been required to reach escape velocity while limiting acceleration to survivable limits for the passengers.The character of Michel Ardan, the French member of the party in the novel, was inspired by the real-life photographer Félix Nadar. (en)
dbo:author	• dbr:Jules_Verne
dbo:illustrator	• dbr:Henri_de_Montaut
dbo:literaryGenre	• dbr:Science_fiction
dbo:mediaType	• dbr:Hardcover
dbo:publisher	• dbr:Pierre-Jules_Hetzel
dbo:series	• dbr:Voyages_extraordinaires

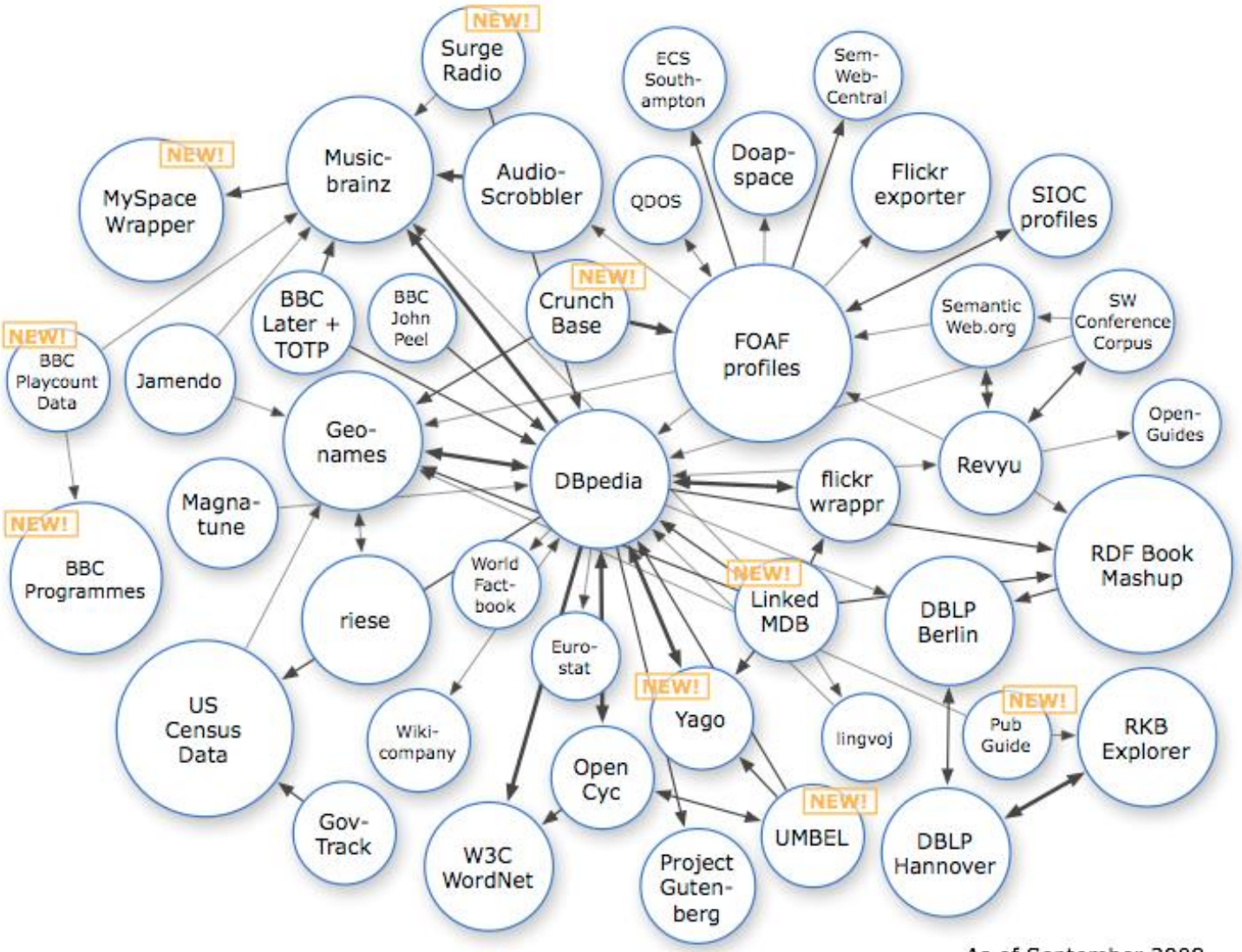
Sistemas Recomendadores



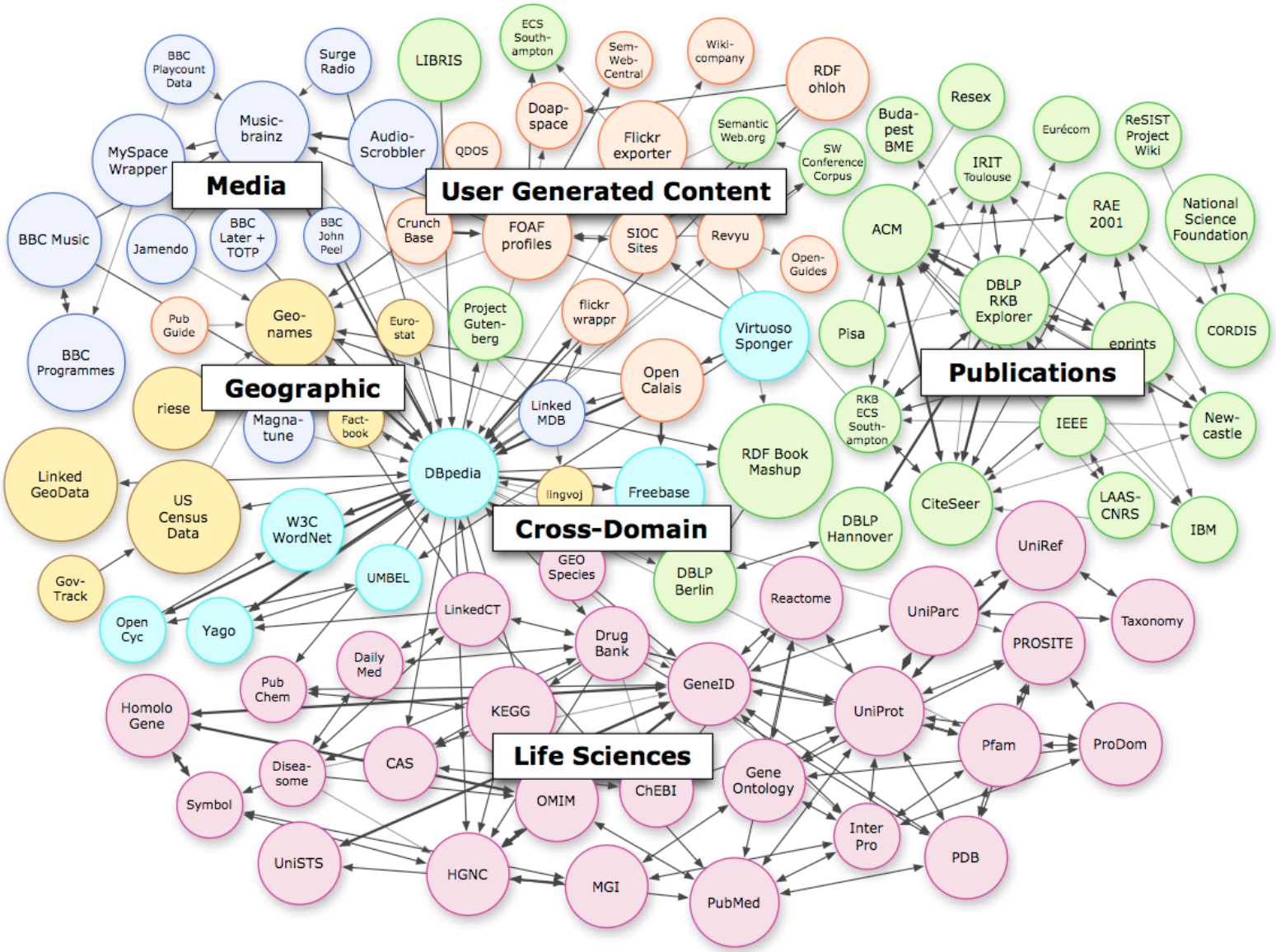
Sistemas Recomendadores



Septiembre 2008



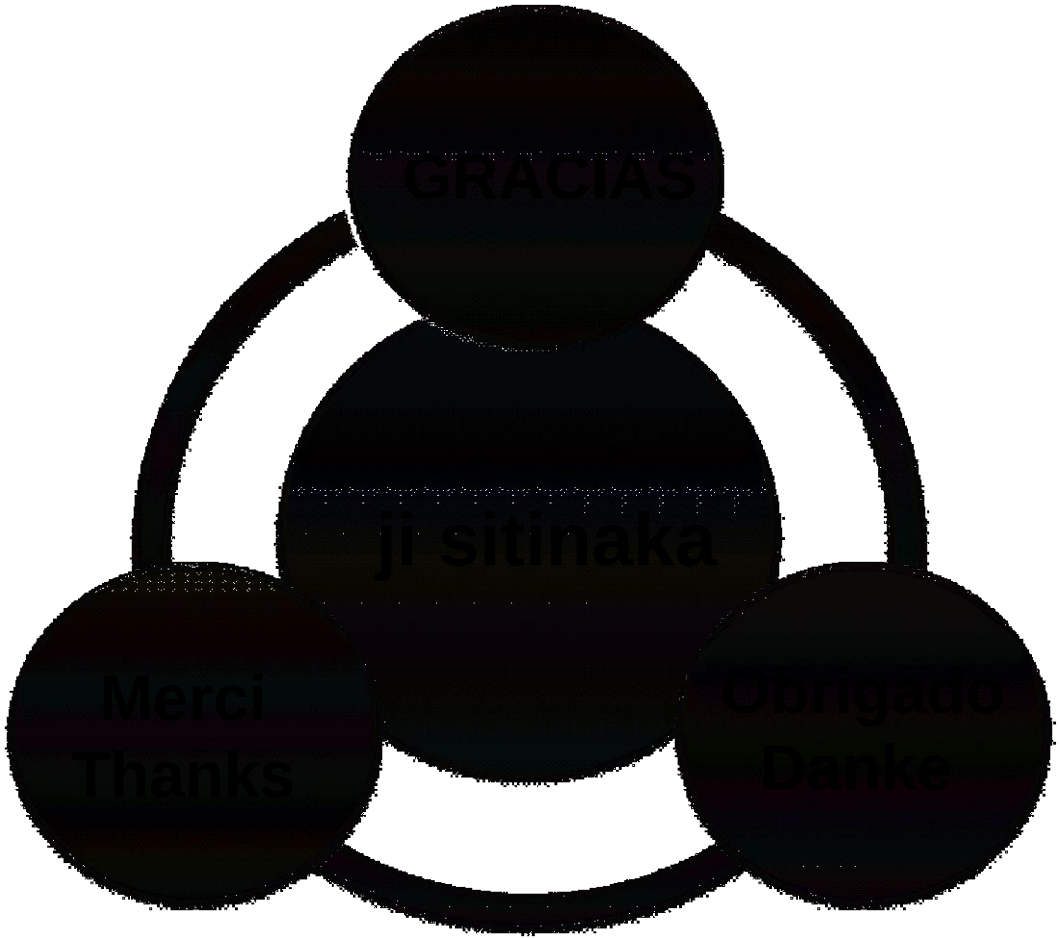
Julio 2016


Media
User Generated Content
Geographic
Publications
Cross-Domain
Life Sciences
BBC Playcount Data
Surge Radio
LIBRIS
ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Resex
Buda-pest BME
Eurécom
ReSIST Project Wiki
Music-brainz
Audio-Scrobbler
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
IRIT Toulouse
MySpace Wrapper
QDOS
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
DBLP RKB Explorer
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
eprints
CORDIS
BBC Programmes
riese
Magna-tune
Fact-book
Open Calais
Linked MDB
RKB ECS South-ampton
IEEE
New-castle
Linked GeoData
US Census Data
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
LAAS-CNRS
IBM
W3C WordNet
DBLP Hannover
UniRef
Gov-Track
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
GeneID
Homolo Gene
KEGG
UniProt
Pfam
ProDom
Disea-some
CAS
Gene Ontology
Symbol
OMIM
ChEBI
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB

Estadisticas de Julio 2016

Dominio	***No de Tripletas***	***%***	***No de enlaces***	***%***
Medios de comunicación	6.098.000.000	10,4%	10.238.000	0,8%
Publicaciones	2.012.000.000	3,2%	40.922.000	3,3%
Ciencias de la Vida	24.029.000.000	36,1%	1.033.199.000	89,4%
Datos Geográficos	30.097.000.000	46,0%	44.038.000	2,7%
Usiarios	976.000.000	1,1%	10.559.000	1,0%
Varios Dominios	2.014.000.000	3,2%	32.992.000	2,7%
Total	***>75.000.000.000***		***>1.000.000.000***	

Conclusiones



“Si buscas resultados distintos,
entonces no hagas siempre lo mismo”

A. Einstein

http://www.ing.ula.ve/~aguilar/actividad-docente/cursos.html

aguilar@ula.ve