Integración de Bases de Conocimiento

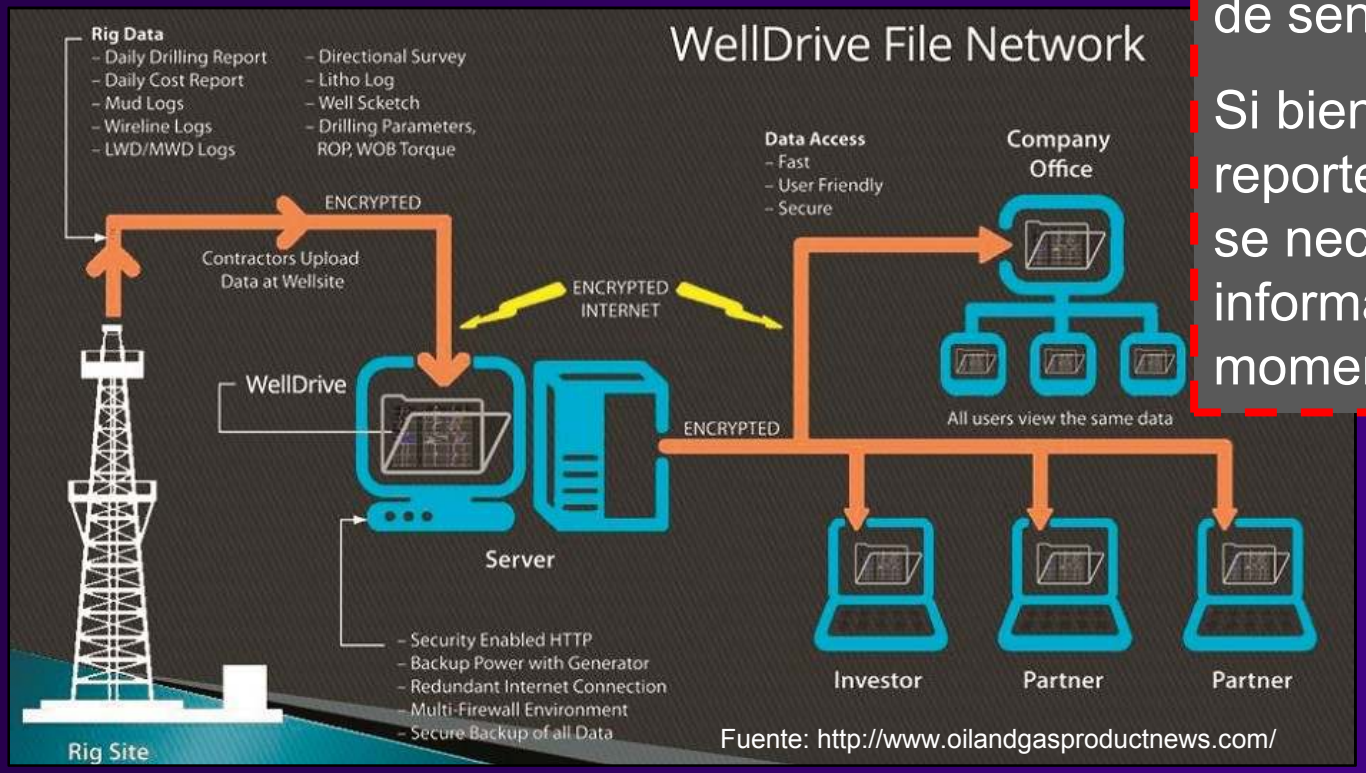
Clase I – Introducción

Profesores: Maria Vanina Martinez y Ricardo Rodriguez

La era de Big Data…

Estamos viviendo en una época en la que la información nos *rodea por completo*

Producción de petróleo y gas natural:



Fuente: http://www.oilandgasproductnews.com/

Un pozo petrolero puede generar hasta 1TB *por día* al tomar información de sensores cada 4ms.

Si bien se producen reportes diarios, también se necesita analizar la información en el momento.

La era de Big Data…

Estamos viviendo en una época en la que la información nos *rodea por completo*

Mercados financieros:

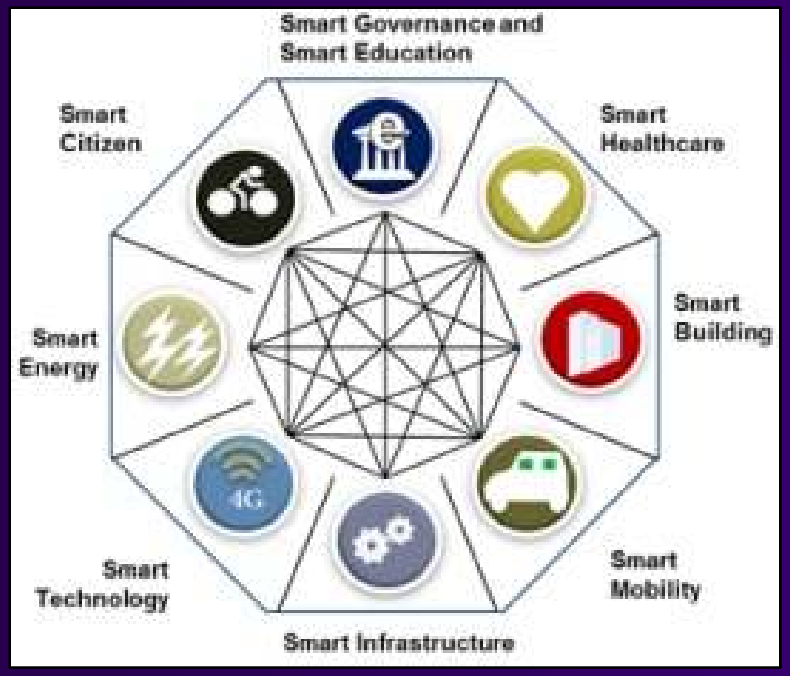
En los últimos 20 años se produjeron unos 20PB de información sobre millones de instrumentos financieros en más de 400 mercados, con datos tomados cada µs.

Fuente: Thomson Reuters

La era de Big Data…

Estamos viviendo en una época en la que la información nos *rodea por completo*

Ciudades inteligentes y gobierno digital



Estas aplicaciones requieren soluciones al problema de la fusión de datos para ser efectivos: tránsito (aéreo, trenes, autos, camiones), horarios de transporte público, cámaras, sensores ambientales, estacionamiento, manifestaciones, ...

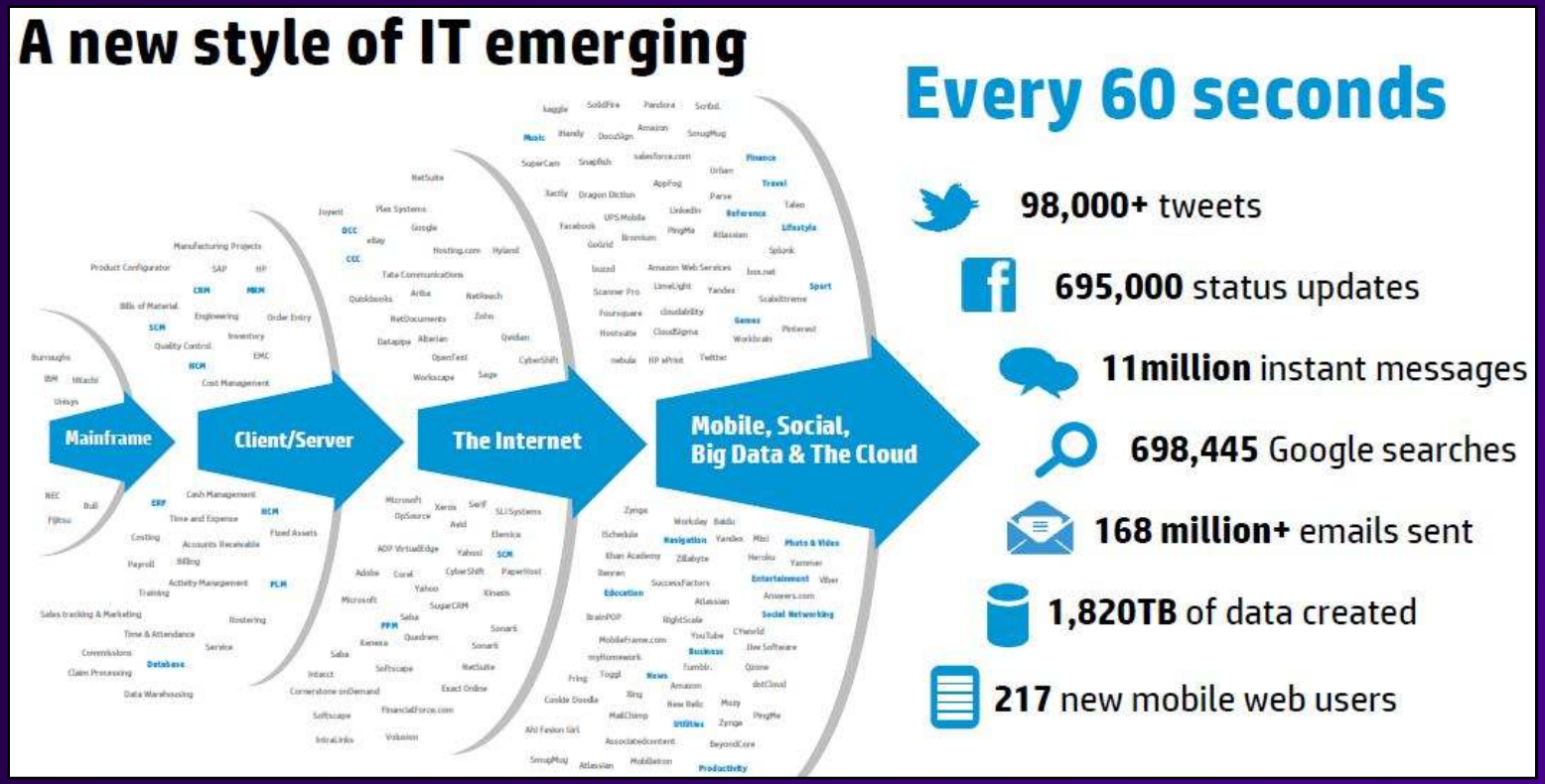
Fuente: www.fmlink.info

La era de Big Data…

Estamos viviendo en una época en la que la información nos *rodea por completo*

Redes sociales

Fuente: www.hp.com (info de 2012)

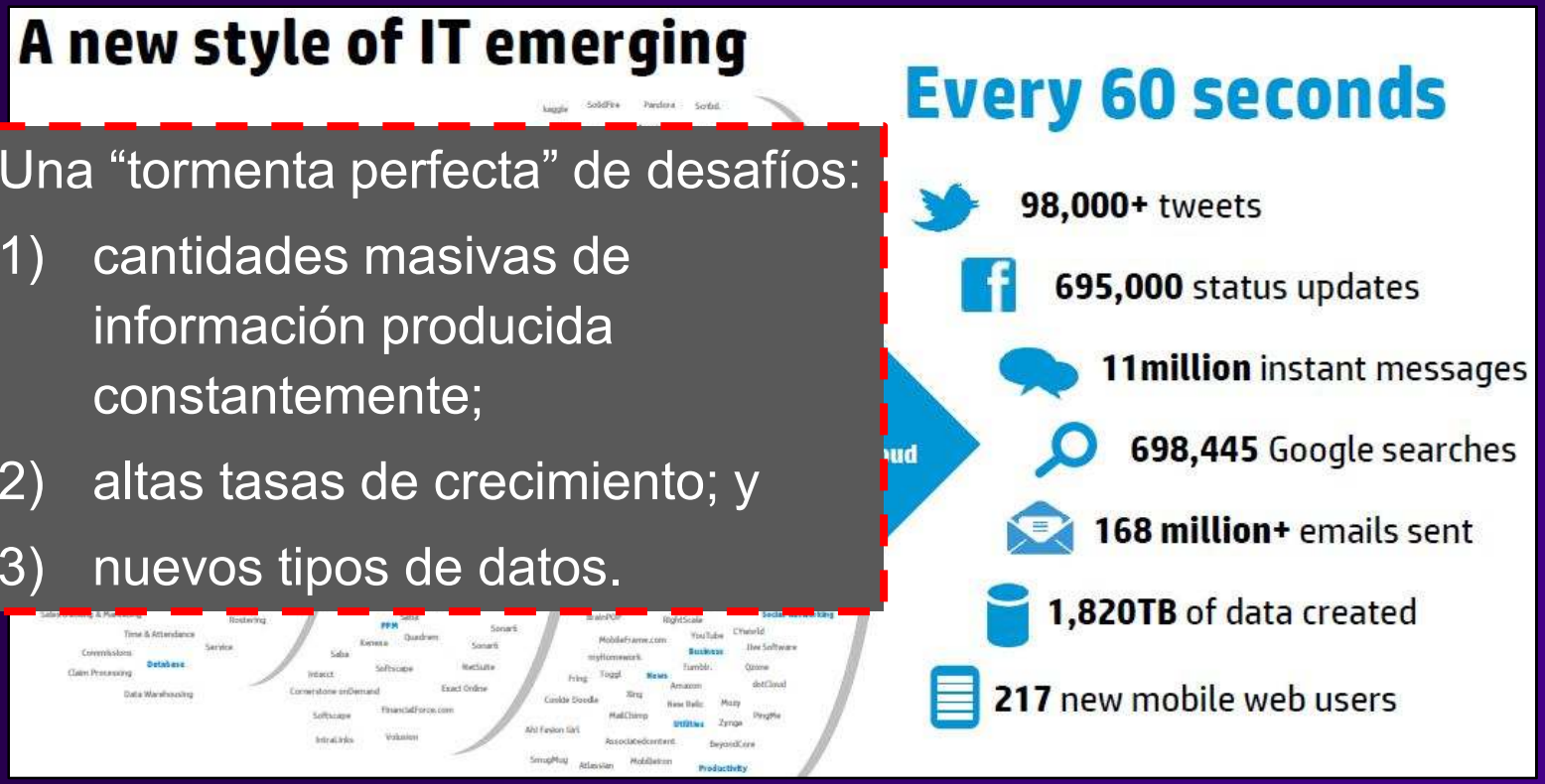


La era de Big Data…

Estamos viviendo en una época en la que la información nos *rodea por completo*

Redes sociales

Fuente: www.hp.com (info de 2012)



Una “tormenta perfecta” de desafíos:

1) cantidades masivas de información producida constantemente;
2) altas tasas de crecimiento; y
3) nuevos tipos de datos.

La era de Big Data…

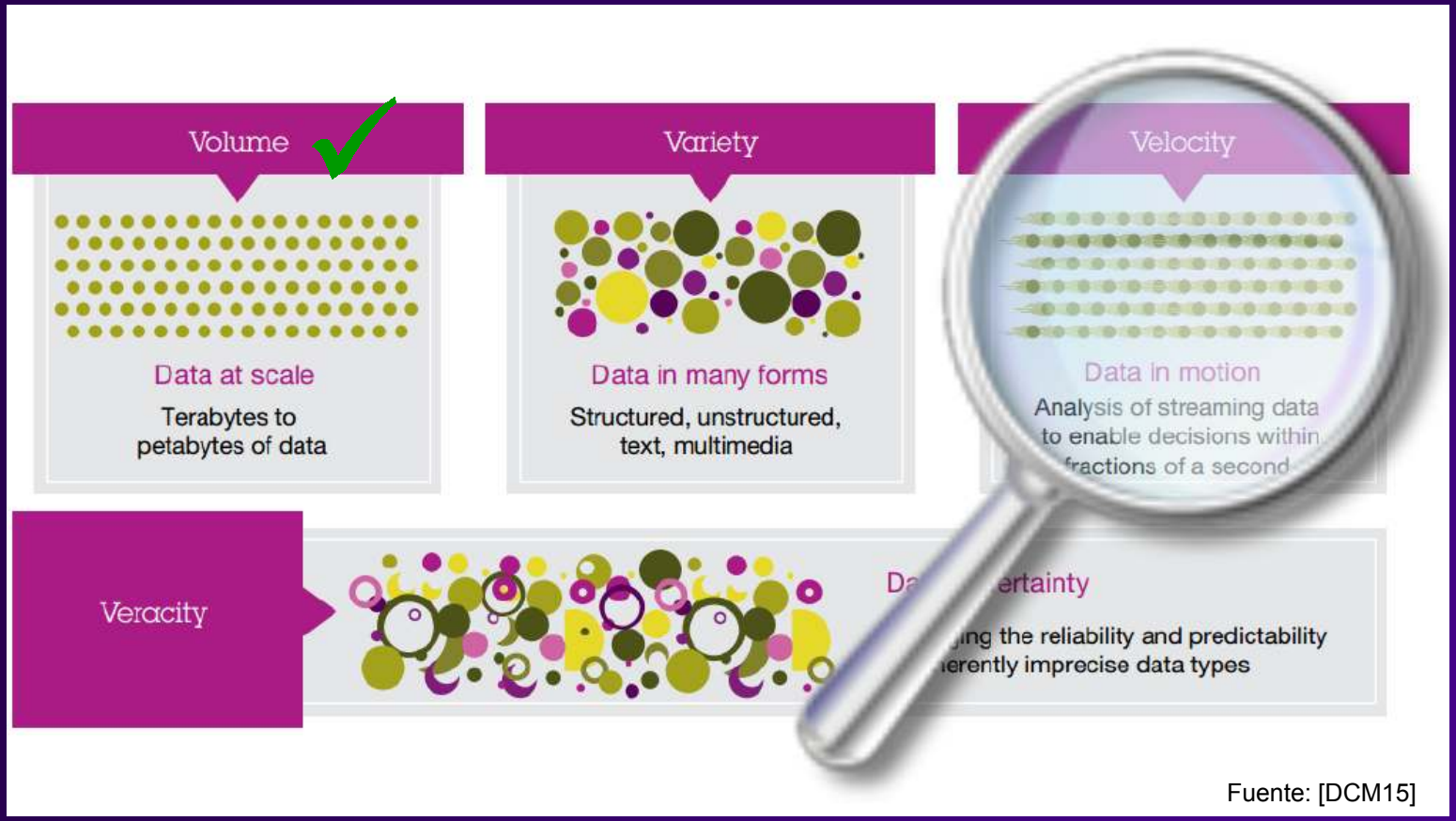
Fuente: Practical Analytics (info de 2012)

Tamaño de los datos

Nota: suele existir confusión entre unidades obtenidas a partir de multiplicaciones por 1000 vs. por 1024.

Velocidad

Este aluvión de información debe procesarse de manera eficiente y efectiva para aprovechar su contenido.



Fuente: [DCM15]

Cantidad y Velocidad: *¿Adiós a los DBMS*?

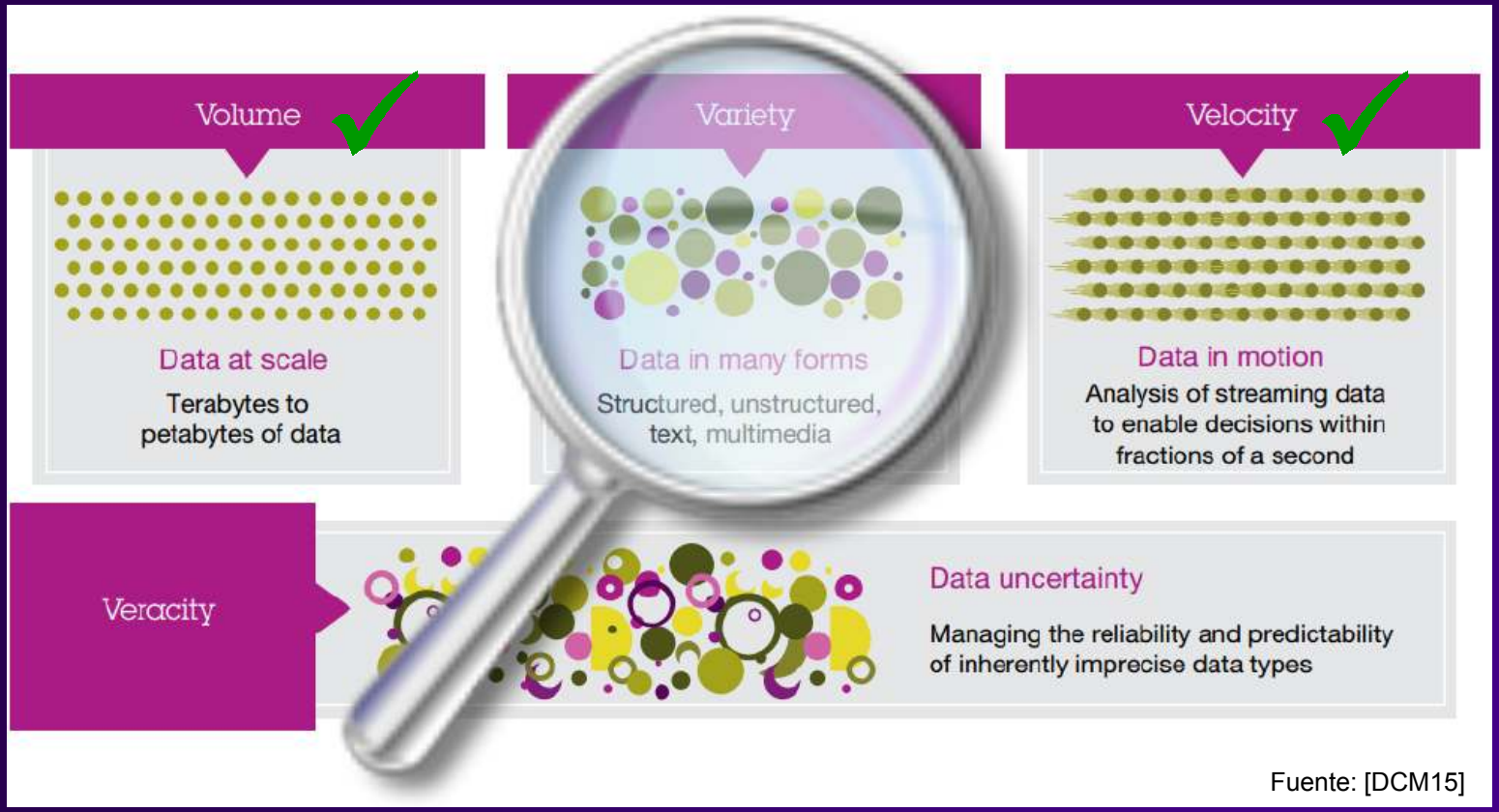
- Ya hace más de 10 años que se planteó la pregunta, y las respuestas han sido contundentes:
  - Se pasó de procesar datos de un sólo negocio a un conjunto, con requerimientos heterogéneos.
  - Nuevas características: "*shared nothing*" y alta disponibilidad.
  - "*No knobs*": los DBMS son altamente configurables porque antes el personal era barato y el hardware costoso; hoy es a la inversa.
  - *Multi-threading* y control de recursos: artificios pensados para esconder la latencia; resulta artificial en un entorno inherentemente *single-thread*.
  - Nuevas tecnologías: grandes capacidades de memoria, "*hot standbys*", la Web, etc.

Cantidad y Velocidad: *¿Adiós a los DBMS*?

- Esto redunda en:
  - La caída del modelo "*one size fits all*": los diferentes problemas de manejo de datos se pueden resolver con arquitecturas de software especializadas.
  - La falla de las implementaciones del modelo relacional para los mercados modernos.
  - La necesidad de repensar tanto los modelos de datos como los lenguajes de consulta; cada aplicación especializada tiene la posibilidad de tomar las decisiones más convenientes.
  - Cada vez mas las aplicaciones requieren consultar y procesar información de distintos repositorios de datos de naturaleza (e información) heterogenea.

Variedad

- Los datos son en general manifestaciones de *eventos*.
- El objetivo es procesar dichos eventos...



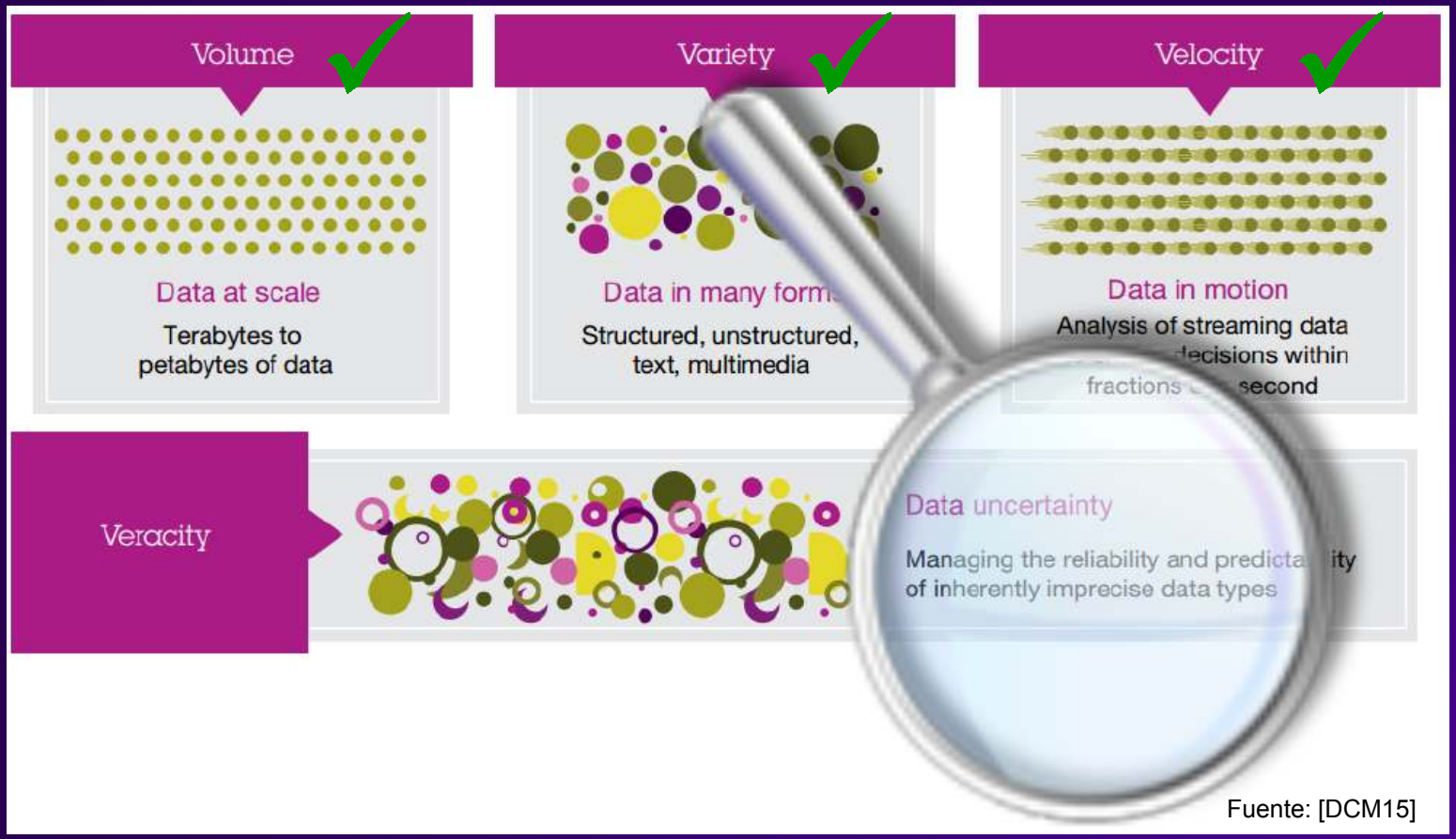
Fuente: [DCM15]

Variedad: Eventos

- Los datos surgen de múltiples fuentes, y por lo tanto se encuentran representados en diferentes formatos.
- La Web Semántica se propuso como solución a este problema representar el conocimiento uniformemente:
  - RDF
  - OWL
  - SPARQL
  - Ontologías: lógicas de descripción, Datalog+/–, etc.
- Lamentablemente, el mundo no es estático, lo cual complica la *aplicación directa* de estas tecnologías.

Veracidad

La información proveniente del mundo real generalmente es *incierta* por una o más razones.

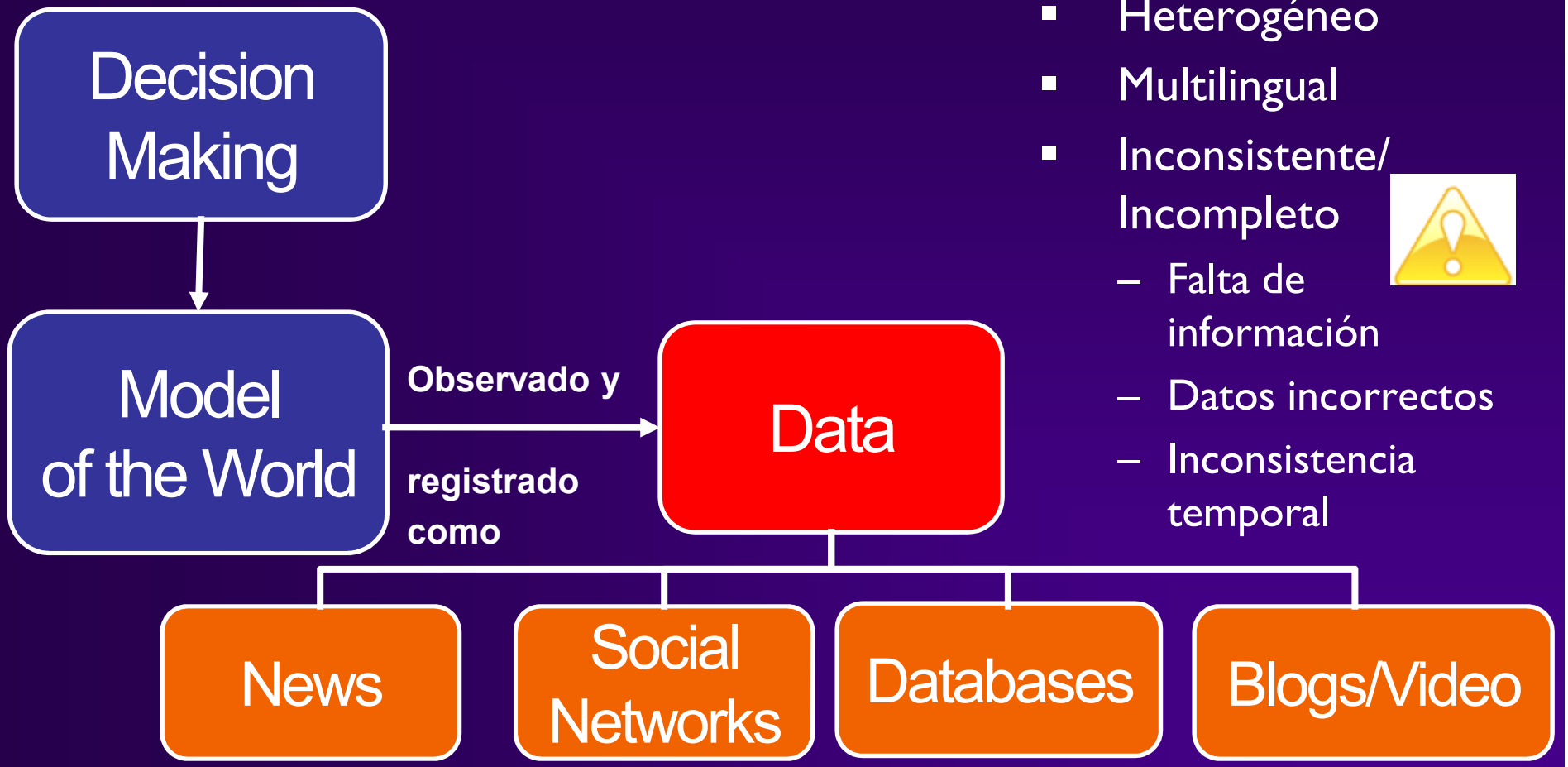


Fuente: [DCM15]

Veracidad: Soporte para incertidumbre

- Muchas aplicaciones deben lidiar con fuentes de datos imprecisas o poco confiables:
  - Sensores ruidosos
  - Incompletitud
  - Inconsistencia
  - Información inherentemente incierta (clima, mercados, etc.)
- La habilidad de cuantificar la incertidumbre asociada con la información es una parte esencial de la expresividad.
- Dos aspectos ortogonales:
  - Soporte para entradas con incertidumbre
  - Soporte para salidas con incertidumbre

El problema de la *integración de conocimiento* tiene como desafío lidiar con la heterogeneidad de las bases de conocimiento con el objetivo de proveer una visión unificada.



- Heterogéneo
- Multilingual
- Inconsistente/ Incompleto
  - Falta de información
  - Datos incorrectos
  - Inconsistencia temporal

Ejemplo: Fragmento de una tabla relacional de un sistema de información de bancos

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S

Ejemplo: Fragmento de una tabla relacional de un sistema de información de bancos

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-200					195000,00	N
140904	15-mag-2001	15-giu-2005	33060	N	N	230600,00	N
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S

Valor negativo indica un retiro de dinero

Ejemplo: Fragmento de una tabla relacional de un sistema de información de bancos

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
140904	15					230600,00	N
124589	5-					195000,00	S
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S

S significa que el cliente es líder del grupo al que pertenece

Ejemplo: Fragmento de una tabla relacional de un sistema de información de bancos

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
			0	N	N	230600,00	N
			6	N	S		
-452901	13-mag-2001	27-lug-2004	92770	S	N		
129008	10-mag-2001	1-gen-9999	62010	N	S		

S significa que el cliente es líder del grupo al que pertenece

S significa que el cliente es "la cabeza" del grupo al que pertenece

Ejemplo: Fragmento de una tabla relacional de un sistema de información de bancos

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
140904						230600,00	N
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S

N significa que el campo FATTURATO no es válido

El problema de integración

- Este ejemplo muestra que en los sistemas del mundo real, el significado de los datos en las tablas puede ser ambiguo.
- Es crucial entender el significado de los datos si queremos manejar de manera "correcta" la información en las tablas y extraerla.
  - Fuertemente ligado a como los datos su usan regularmente y poder entenderlo requiere de la experticia de dominio (background knowledge) los usuarios que lo consumen.
- Además…en general, los sistemas de información usan diferentes fuentes de datos heterogéneas, internas y externas a la organización.

Más problemas…

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N
124589	5-mag-2001	30-lug-2004	92736				
-452901	13-mag-2001	27-lug-2004	92770				
129008	10-mag-2001	1-gen-9999	62010				

TRAN_ID	TRAN_DATE	ID_GRUP	AMOUNT	TRAN_ST
12975402	1-gen-9999	92736	-124589	N
19351128	15-giu-2005	35060	-140900	N
15622008	30-lug-2004	92736	-124589	S
16321377	27-lug-2004	92770	452901	N
17475506	1-gen-9999	62010	-129008	S

Más problemas de integración

- La integración de conocimiento debe lidiar inevitablemente con problemas de *incertidumbre* y/o *inconsistencia*.
- En un sistema de información, se espera que la resolución (o no) de esos problemas sea *(semi-)automática*:
  - Esto es no sólo deseable sino imprescindible en sistemas que manejan *grandes cantidades* de datos (por ej., provenientes de la Web).
  - Estos mecanismos o métodos deben además tener una correspondencia con los métodos (o resultados) que un ser *humano* utilizaría al enfrentar la tarea.

En esta materia…

- Estudiar distintas perspectivas al problema de integración de bases de conocimiento: OBDA, Data Integration vrs Data Exchange, Revisión de Creencias, “Merging” de Creencias.
- Estudiar los principales problemas que acarrea la integración de datos (Inconsistencia e Incertidumbre):
  - Semánticas de respuesta a consultas tolerantes a la inconsistencia, Argumentación.
  - Modelos probabilísticos: input y output incierta.
  - Agregación de información utilizando técnicas de Social Choice.

Referencias

*El ejemplo principal de esta clase fue tomado del tutorial: "*Methods and Tools for Developing Ontology-Based Data Access Solutions Concepts for Ontology-Based Data Access" dictado por Giuseppe De Giacomo, Domenico Lembo, Antonella Poggi, Valerio Santarelli and Domenico Fabio Savo, en ISWC 2017: https://sites.google.com/a/dis.uniroma1.it/mt4obda/

Parte del contenido de este curso está basado en trabajo de investigación realizado en colaboración con Thomas Lukasiewicz, Georg Gottlob, V.S. Subrahmanian, Avigdor Gal, Andreas Pieris, Giorgio Orsi, Livia Predoiu y Oana Tifrea-Marciuska.