Big data y open data

Wilson Peres

UAM (A) Sesión VII

Big Data

Big data

- ¿Qué es?
  - Conjuntos de datos cuyos tamaños están más allá de la capacidad de las herramientas de software de bases de datos típicas para capturar, almacenar, gestionar y analizar información.
- ¿Cómo se origina?
  - Por la explosión en la cantidad (velocidad y frecuencia) y diversidad de datos digitales generados en tiempo real como resultado del papel cada vez mayor de la tecnología en las actividades diarias (*digital exhaust*).
- ¿Para qué sirve?
  - Permite generar información y conocimiento con base en información completa en tiempo real.

Tipos de datos

- Compras y transacciones (incluyendo información de tarjetas de crédito)
- Datos de gestión empresarial
- Búsquedas (consulta, trayectoria recorrida, historia)
- Sociales (datos de identidad, información de amistades)
- Intereses personales (que me gusta, *tweets*, recomendaciones, enlaces)
- Ubicación, sensores físicos (GPS, patrones de tráfico, *Internet of Things*, etc.)
- Contenido (SMS, llamadas, e-mails)

Información generada por fuentes tradicionales, particularmente empresas e individuos en sus actividades cotidianas, y usada para un objetivo diferente de aquel para el que fue pensada

Las tres V que pueden ser cinco

- **Volumen:** Cuántos datos
- **Velocidad:** Cuán rápido se procesan
- **Variedad:** Cuántos y de qué tipos
- **Veracidad**: Cuán precisos son para predecir en el universo en consideración
- **Valor**: Cómo reducir su complejidad para hacerlos realmente valiosos

De las 3V a las 3C

- Crumbs
  - Escape digital
  - Web
  - Sensores, imágenes satelitales
- Capacidades
  - Alfabetización en datos
  - Encontrar el sentido de los datos
- Comunidad (nueva)
  - Multidisciplinaria y colaborativa

Información digital

- Google procesa 1 petabyte (mil terabytes) por hora (10^{15})
- 1200 exabytes (1 millón de terabytes) de datos digitales en 2010 (todas las palabras emitidas en la historia de la humanidad son 5 exabytes)
- Crecimiento exponencial
- Énfasis en imágenes y video (Netflix 40% de la red en USA), i. e., datos no estructurados
- Diversidad de dispositivos (>4 mil millones en 2010)
- M2M e Internet de las cosas (IoT)

What Happens in an **Internet Minute?**

And **Future Growth** is **Staggering**

intel

Implicancias

- Nueva era caracterizada por la abundancia de datos
  - Alcanza a todos los sectores económicos
  - Los datos son un nuevo factor de producción y ventaja competitiva
- Oportunidad
  - Aprender sobre el comportamiento humano para diversos fines
  - Creación de valor vía innovación, eficiencia y competitividad
  - Aumento del excedente del consumidor y del bienestar del ciudadano
- Nuevos negocios y formas de competencia
  - Almacenamiento y gestión de datos
  - Análisis de datos empresariales (más de US$100 mil millones en 2010, con crecimiento anual de 10%)

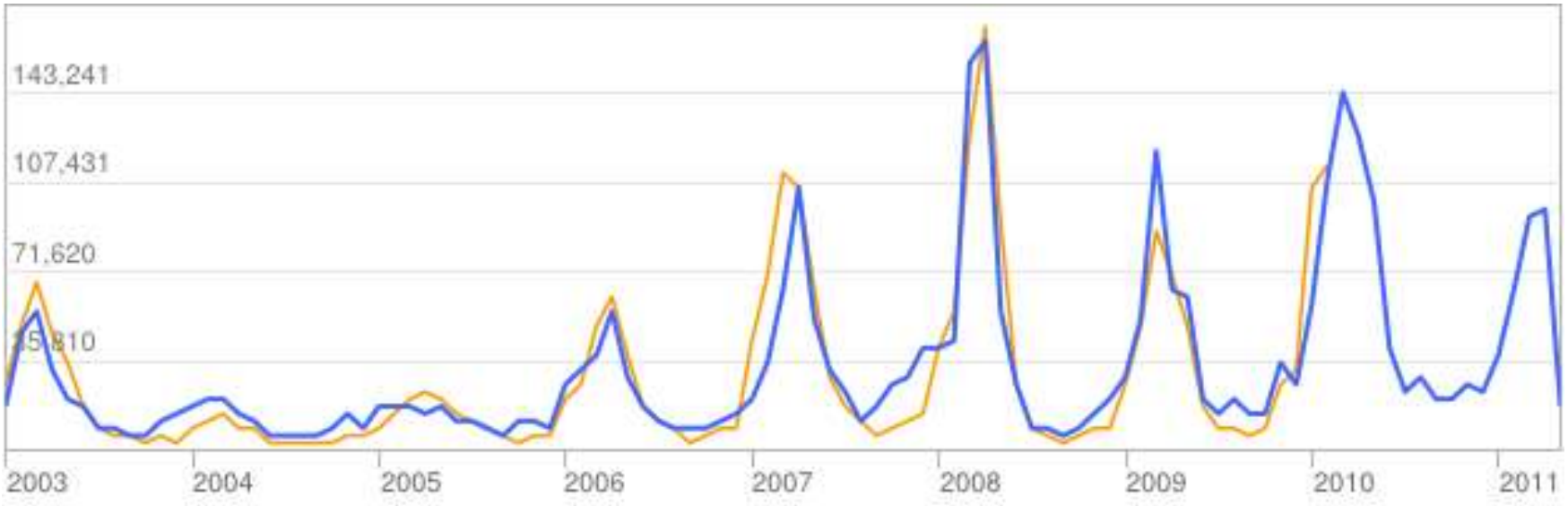
Big data para la creación de valor

1. Segmentación de mercado para personalizar acciones
2. Nuevos modelos de negocios, productos y servicios
   - Mejora de productos existentes
   - Desarrollo de nuevos productos (masa y personalización)
   - Nuevos modelos de servicio empresariales y gubernamentales, p.e. *anticipatory shipping* de Amazon
   - Gestión integral de cadenas productivas (por ejemplo, emisiones)
3. Apoyo a la toma de decisiones con *software* inteligente
4. Transparencia y eficiencia al compartir datos
5. Mejor y más oportuno análisis de desempeño de las organizaciones y ajustes en las acciones

Big data para el desarrollo

- Convertir datos imperfectos, no estructurados y complejos acerca del bienestar de las personas, en información procesable, que reduzca las brechas temporales y de conocimiento para:
  - Decisiones de política que respondan oportunamente a determinadas situaciones
  - Rápida retroalimentación sobre el impacto de esas políticas.
- Generar información inteligente que ahorre recursos en la generación de estadísticas para políticas de desarrollo.

Google: dengue en Brasil



Estimación de Google

Datos de Brasil

Ejemplos micro, macro y social

- **Micro**: Un modelo de físicos de la Northwestern University predice con más de 93% de precisión donde una persona está en un momento determinado con base en el análisis de información de teléfonos celulares generada en sus movimientos pasados
- **Macro**: El PIB de un país se puede estimar en tiempo real mediante la medida de la emisiones de luz en la noche detectadas remotamente (Helbing y Balietti, Eur. Phys. J. 2011)
- **Social**: Científicos de la John Hopkins University analizaron más de 1,6 millones de *tweets* relacionados con salud (de un total de más de 2 mil millones) en Estados Unidos entre mayo de 2009 y octubre de 2010 y hallaron una correlación de 95,8% entre la tasa de enfermos de gripe estimada con base en sus datos y la tasa oficial de engripados

Un ejemplo más detallado
(MIT, Harvard)

- Eagle, Pentland y Lazer (2009) analizaron 330 mil horas de datos sobre comportamiento en el uso de teléfonos móviles de 94 personas, y los compararon con datos de relaciones directamente reportados por los individuos
- Presentan un método para medir conductas basado en datos de proximidad y comunicación, e identifican características que les permiten predecir con 95% de precisión las relaciones de amistad recíprocas
- Usando estas señales de conducta, pueden predecir resultados individuales como satisfacción en el trabajo
- Observaciones sobre el uso de teléfonos móviles proveen indicaciones no solo sobre el comportamiento observable, sino también sobre variables como amistad y satisfacción individual

Big data para la sostenibilidad

Apoyo a la sustentabilidad

- Permite a las empresas entender, medir y actuar sobre sus mayores impactos ambientales: los que están fuera de su control
- **Nike** e **Ikea** tratan de entender el impacto de sus negocios a lo largo de toda la cadena de valor
- **BT** estudió el *carbon footprint* total de su negocio: 92% del total de emisiones fuera de su control directo; 2/3 viene de 17000 proveedores. Identificó *carbon hotspot areas*, mostrando oportunidades de negocios para reducir costos y carbono.
- **Hitachi** provee una plataforma online para que sus proveedores informen sobre su cumplimiento de los criterios de sustentabilidad (incentiva responsabilidad de los pequeños proveedores)
- Fuente: *Carbon Trust* en *The Guardian*, 31 enero 2014 y *Big Business, Big Data, Big Sustainability*, 31 octubre 2013

Tres proyectos sostenibles de grandes datos de CISCO en Dinamarca (2014)

- Colaboración con la municipalidad de Copenhagen para desarrollar soluciones sostenibles (*start-ups*).
- Alumbrado público en Albertslund.
- Una nueva ciudad en Frederikssund basada totalmente en energía renovable.

Otros ejemplos

- Monitoreo de niveles de reserva de petróleo desde el espacio con base en fotos de los tanques de depósito.
- Monitoreo de viñedos a nivel de planta.
- Monitoreo de pesquerías, por ejemplo, salmón.
- Base de datos de la totalidad del stock ganadero de un país.

BIG DATA

MORE MESSINESS

CORRELATIONS

Analytics

Data, proxies, visualization

Elementos de *Analytics*

- Herramientas y metodologías para transformar cantidades masivas de datos brutos en “datos sobre datos” con propósitos analíticos
- Se originó en biología intensiva en cómputo, ingeniería biomédica, medicina y electrónica
- Algoritmos para detectar patrones, tendencias y correlaciones, en varios horizontes temporales, en los datos
- Uso de técnicas avanzadas de visualización: datos que hacen sentido

Reality mining: sensing complex social systems (MIT)

- *Continuous data analysis over streaming data*, using tools to scrape the web (e.g., gathering product prices in real-time)
- *On-line digestion of semi-structured data and unstructured ones* (e.g., news items, product reviews)
- *Real-time correlation of streaming data (fast stream)* with slowly accessible historical data repositories

Un análisis avanzado

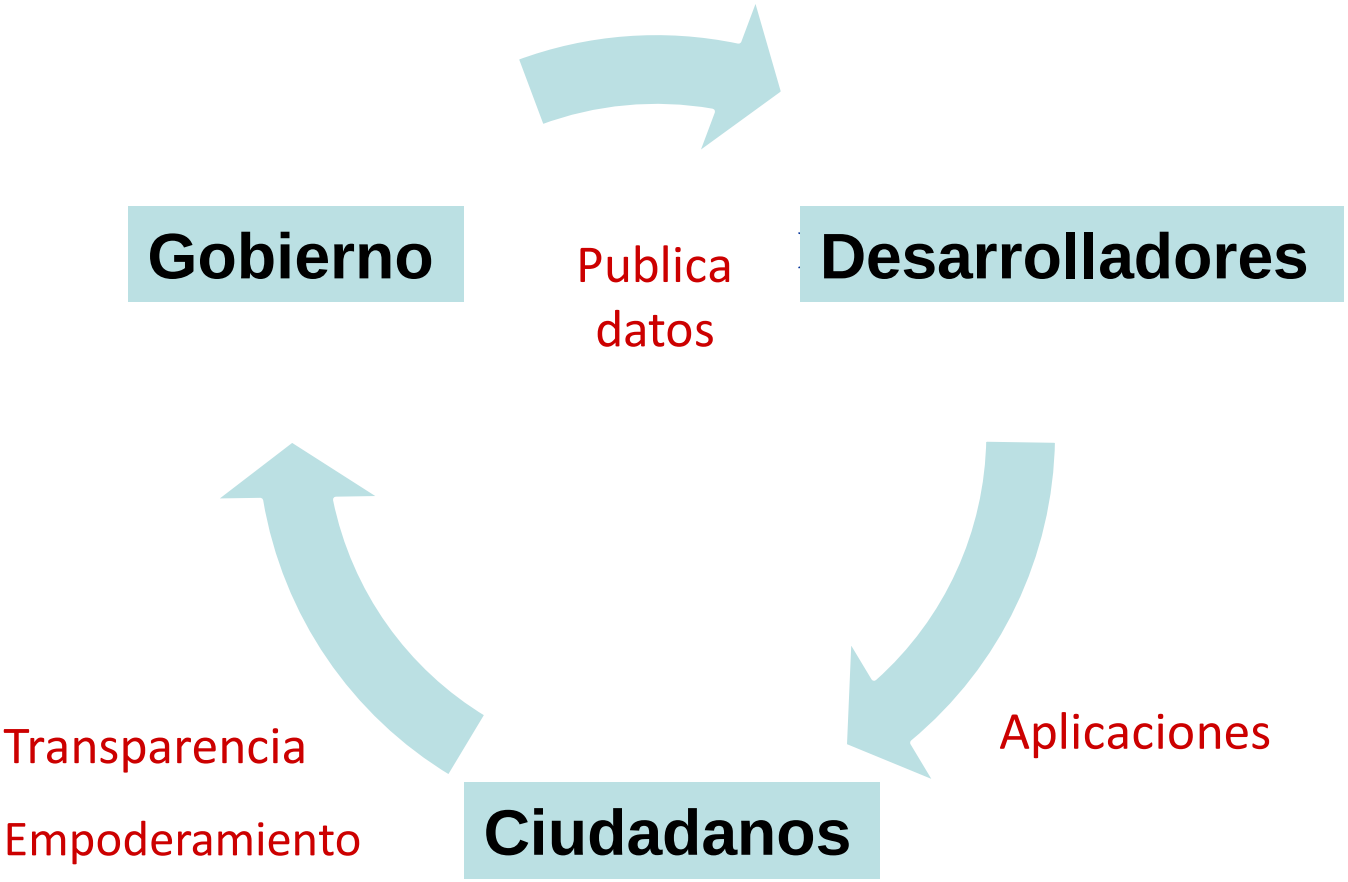
- Hal R. Varian, Chief Economist, Google
- “Big Data: New Tricks for Econometrists”, February 2014
- Hay elementos que solo se dan en *big data* que requieren instrumentos (*tools*) diferentes
  - Se necesitan instrumentos más poderosos
  - Podemos tener más predictores potenciales que los necesarios para una estimación (hay que seleccionar variables)
  - Más allá del modelo lineal”: árboles de decisiones, redes neurales, modelización de relaciones complejas, etc.
- La necesaria colaboración de la econometría (inferencia causal) con *machine learning* (predicción)

Open data

¿Qué es?

- Datos que pueden ser reutilizados y distribuidos sin restricciones de ningún tipo, en especial restricciones administrativas o tecnológicas
- 8 características
  - Completos
  - Primarios, no procesados
  - Actuales
  - Accesibles a todos los usuarios para cualquier propósito
  - Procesables por máquinas
  - No discriminatorios, abiertos a cualquiera
  - Formato no propietario
  - Libres de licencias, patentes, marcas, etc.
- Web 3, *linked* data e interoperabilidad

Actores


Gobierno
Publica datos
Desarrolladores
Transparencia
Empoderamiento
Participación
Ciudadanos
Aplicaciones

Tim Berners-Lee 2010

- W3C, *5-star rating system* for open data
  - Cero estrella. Datos no accesibles bajo una licencia abierta, aunque estén online
  - 1 estrella. Datos accesibles en la web; legibles por personas, pero no por *software* por estar en un formato cerrado. No pueden ser reutilizados fácilmente.
  - 2 estrellas. Accesibles en la web en formato estructurado legible por una máquina. Se pueden procesar, exportar y publicar fácilmente, pero dependen de *software* propietario (Word, Excel)
  - 3 estrellas. El reuso no depende de *software* propietario (CSV en lugar de Excel).
  - 4 estrellas. Los datos están in the web en lugar de on the Web con base en el uso de un *uniform resource identifier* (URI) único que permite un control fino de los datos, por ejemplo *bookmarks* o *linkages*.
  - 5 estrellas. Los datos no solo están en la web, sino *linked* a otros datos, aprovechando los efectos de red. Esto aumenta su valor pues tienen un contexto y pueden ser descubiertos por otras fuentes (p.e. a través de links a Wikipedia).

Formatos de *open data*

1 estrella	pdf
2 estrellas	Excel
3 estrellas	csv
4 estrellas	**RDFa** (Resource Description Framework in attributes) adds a set of attribute-level extensions to HTML, XHTM, and various XML-based document types for embedding rich metadata within Web documents. It enables its use for embedding RDF subject-predicate-object expressions within XHTML documents. Debe tener al menos 3 Universal Resource Indentifiers (URI)
5 estrellas	RDFa con URIs y propiedades semánticas que permiten el reuso de *linked data*

Mashups

- Un *mashup* es una aplicación que usa contenido de más de una fuente para crear un servicio nuevo que se despliega en una sola interfaz gráfica
- Gran aumento de *mashups* en la *web*:
  - Grandes empresas como Yahoo!, Google y Amazon han abierto sus datos para que sean utilizables por otras fuentes de datos sin necesidad de largas negociaciones de licencias
  - Nuevas herramientas para crear *mashups* fácilmente, con bajo requerimiento tecnológico
  - Las mashups más populares son las que tienen como base mapas (36% de las detectadas por ProgrammableWeb)

Resultados de una búsqueda en programmableweb.com

- *Environment*
  - 80 mashups, 90 APIs y 2 source codes
- *Pollution*
  - 6 mashups, 2 APIs
- C*arbon footprint*
  - 1 mashup, 2 APIs
- *Carbon emissions*
  - 3 mashups, 1 APIs

Un ejemplo de app Infoaqua, México, 2012

- Problemática
  - Falta de empatía y consciencia respecto del consumo desmedido, la contaminación y problemática general del agua, como la sequía, falta de recursos, falta de tratamiento y sobreexplotación.
- Solución planteada
  - Un interactivo que muestre a la sociedad la situación de las infraestructuras hidrológicas y cuerpos de agua en el país por medio de infográficos dinámicos en tiempo real y juegos informativos sobre la problemática del agua en el mundo
- Datos utilizados
  - **http://www.inegi.org.mx**
  - **http://www.cna.gob.mx**
  - **http://mapas.gob.mx**
- http://labs.ratio.mx/infoaqua/

Econsciencia, Bolivia, 2012

- Problemática
  - La contaminación es un problema difícil de controlar. La falta de conciencia de las personas e instituciones respecto del medio ambiente termina ocasionándole daños; algunos pueden ser reparados, pero muchas veces son irreversibles.
- Solución planteada
  - Información sobre actividades en favor y actividades que amenacen al medio ambiente estará disponible para todos, sean o no usuarios registrados
  - Mediante un mapa nacional que muestre los ríos, lagos y áreas protegidos
    - se identificarán puntos de contaminación en lugares específicos de un río o
    - se mostrará un área que indique una iniciativa de reforestación en un área protegida o
    - se podrán denunciar lugares donde se echa basura u otros residuos
  - Las instituciones serán capaces de llevar actividades en favor del medio ambiente y luego publicarlas a través del sitio; la información será accesible para todos los usuarios.
- Datos utilizados
  - **https://www.geobolivia.abc.gob.bo**
  - **https://www.wdpa.org**
- http://2012.desarrollandoamerica.org/portfolio/econsciencia/

¿Dónde reciclo? en Montevideo

- Para facilitar la clasificación y reciclaje de residuos domiciliarios, Data creó una aplicación que permite encontrar los contenedores de desechos de vidrio, plástico, pilas y metal en cualquier lugar de la ciudad.
- Obtiene la ubicación del usuario mediante el GPS de un teléfono móvil o la información de posicionamiento de una computadora y presenta un mapa con su ubicación y los puntos de reciclaje más próximos utilizando datos abiertos de la intendencia de la ciudad.
- Pese a que esos datos se podían consultar mediante una herramienta de la propia intendencia, Data decidió crear otra porque creía que podían lograr una forma más eficiente de obtener y mostrar esa información en dispositivos móviles.

- Fuente: http://datauy.org/proyecto/donde-reciclo/

D | A | T | A

APIs y su uso

- Una API (*application programming interface*) especifica la manera como algunos componentes de *software* deben interactuar entre sí.
- Un procedimiento para que un programa pueda cumplir una tarea mediante la recolección o modificación automática de datos.
- Un intermediario entre los datos y los sistemas que van a utilizarlos

Ejemplos de APIs

- *Twitter*
  - Provee un método API para casi todas las variables (*features*) que se pueden ver en su sitio web: programadores pueden usarla para hacer aplicaciones, sitios *web*, *widgets* y otros proyectos que interactúan con *Twitter*
  - Los programas se comunican con la API de *Twitter* sobre HTTP
- *Linked Brazilian Amazon Rainforest Dataset*
  - Una aplicación que se vincula con linkedscience.org que guarda y explora este *dataset* de estadísticas de deforestación
  - Los datos se subdividen en una malla de 8441 celdas de 25km x 25km cada una. Las variables que se usan son deforestación, precio de las pasturas y coordenadas para dibujar las celdas

Web crawlers

- Un w*eb crawler* es (ro)bot de Internet que revisa (*browse*) la www de manera sistemática.
- Los motores de búsqueda usan *web crawling* o *spidering software* para actualizar su contenido o indexar el contenido de otros sitios.
- *Web crawlers* pueden copiar las páginas que visitan para su procesamiento posterior por motores de búsqueda que indexan las páginas descargadas de manera que los usuarios puedan buscar más rápidamente.
- *Web scraping* es un tipo de automatización de la red que simula la navegación por humanos (comparación de precios en línea, búsqueda de contactos, monitoreo del clima, detección de cambio en páginas, investigación, *mashups*).

Metabuscadores

- Un metabuscador es un sistema que localiza información en los motores de búsqueda más usados: un buscador en buscadores.
- Retorna las búsquedas desde motores como About, Ask.com, FinWhat, Google, LookSmart, MSN Search, Teoma, Yahoo!, Bing y otros, incluyendo de audio y video.
- Carece de base de datos propia y, en su lugar, usa las de otros buscadores y muestra una combinación de las mejores páginas devueltas devueltas por cada uno.
- Ejemplos
  - **Blucora's InfoSpace** provee servicios de *metasearch* y *private-label Internet search* (soluciones para *web publishers)*su principal sitio de búsquedas es dogpile.com.
  - **Ixquick** (startpage.com) da relevancia especial a los primeros diez resultados de múltiples motores de búsqueda y usa un sistema de estrellas para calificar sus resultados. Privacidad: no registra las direcciones IP de sus usuarios.

Conclusiones

Problemas

- Disponibilidad de datos: asimetrías
  - Redes sociales generan datos abiertos
  - Gobiernos los están abriendo, pero lentamente
  - Datos de empresas siguen cerrados (¿filantropía de datos?)
- Diferentes capacidades de buscar y analizar datos
- Privacidad y los límites a la anonimización de conjuntos de datos
- Una buena parte de las nuevas fuentes de datos reflejan sólo percepciones, intenciones y deseos
- *Apophenia*: ver patrones donde no hay; cantidades masivas de datos abren conexiones en todos los sentidos (error de Tipo I)
- *More data* o *right data*