Big Data

…GRANDES RETOS Y OPORTUNIDADES

Big Data

"Big Data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it"

Dan Ariely, Duke University

Director or the Center for Advanced Hindsight

Big Data

El término fue introducido por John Mashey ”, director científico en *Silicon Graphics* en los 1990’s. En una presentación con una transparencia titulada “Big Data and the Next Wave of InfraStress”.

“… I was using one label for a range of issues, and I wanted the simples, shortest phrase to convey that the boundaries of computing keep advancing…”

Un concepto que evoluciona

El termino “Big Data” ya ha estado circulando por algún tiempo, pero todavía prevalece confusión sobre lo que realmente significa. En realidad, el concepto está evolucionando continuamente ya que es el concepto detrás de muchas olas de transformación digital y científica y metodológica.

Una definición, inicial es

“Big Data” se refiere al conjunto de datos que es tan grande o complejo que no puede ser percibido, adquirido, gestionado y procesado con tecnologías tradicionales (Cheng et al. 2014).

Las cuatro V's de Big Data

Volumen (Volume)

Esta características es la más asociada a "Big Data" y se refiere a la cantidad de información que se tiene, la cual puede alcanzar proporciones casi incomprensibles.

Facebook: Con más usuarios que habitants de China, la plataforma alberga aproximadamente 250 mil millones (250×10^9) fotografías.

Sensor de temperature (weather underground): Con un solo sensor reportando c/minuto se tienen 525,950 observaciones en un año. Las **250,000** estaciones personales (en EU), entonces generarían 132×10^9 lecturas solo de temperatura!

Las cuatro V’s de Big Data

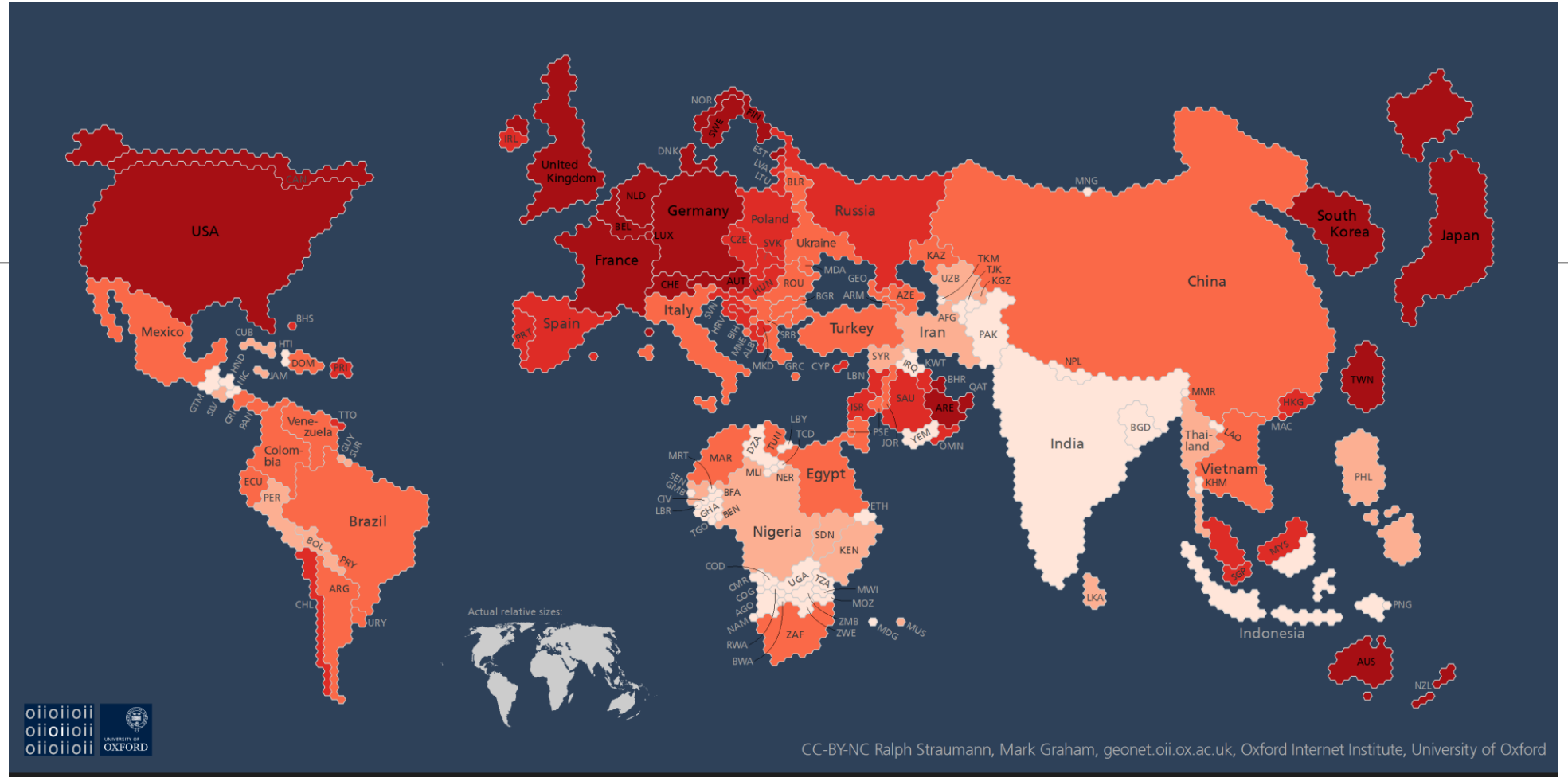
Velocidad (Velocity)

Se asocia a la cantidad de información que se genera por unidad de tiempo. **Cada minuto**:

YouTube: Los usuarious suben 72 horas de videos

Transacciones: En EU hay 100,000 transacciones de tarjetas de crédito

Búsquedas de Google: Recibe más de 2 millones de consultas.



The World Online

Percentage of people online

80–100% 60–80% 40–60% 20–40% 0–20%

Number of people online

One ● represents roughly 470,000 people online.

The countries are scaled proportionally to the number of Internet users in that country. Countries with fewer than 470,000 people online have been removed from the map. The shading indicates the percentage of the population that is online.

The visualization uses 2013 data from the World Bank's Worldwide Development Indicators project and from Natural Earth.

Las cuatro V’s de Big Data

Variedad (Variety)

Se refiere a las diferentes presentaciones y complejidades de la información. Las fuentes de información son las estructuradas (como las bases de datos) pero la información es tan rica como imágenes, voz y video.

Veracidad (Veracity)

Se refiere a la incertidumbre de los datos. En muchos casos la calidad y precisión de la información es menos controlable. Por ejemplo los mensajes de Twitter (abreviaciones, errores tipográficos, lenguaje coloquial)

Las ~~cuatro~~ cinco V’s de Big Data

La quinta:

Valor (Value)

Se relaciona con el potencial de transformar la información en ganancia.

Se enfoca en la ciencia de datos, tal como herramientas y métodos estadísticos y analíticos para la extracción de conocimientos y toma de decisiones.

Tan sólo una gran base?

Big data tiene características que no suelen compartirse in bases de datos pequeñas:

1. Alta dimensionalidad
2. Heterogeneidad
3. Escala
4. Oportuno (Tiempo Real)
5. Requerimiento de seguridad y privacidad

Algunos retos de Big Data

En relación con Data pipeline

- Adquisición/ Almacenamiento: Se deben encontrar formas de almacenar la información más eficientemente.
- Transmisión: Mejorar la rapidez de la comunicación.

- Limpieza/adecuación de la información puede ser ruidosa o transformada a una base más estructurada.
- El análisis de la información puede ser muy complejo. Aún si es posible, puede tomar mucho tiempo.
- Procesamiento en tiempo real. La información es dinámica y los resultados de los análisis se deben actualizar.

Tecnologías de Información

Big Data ha dado origen a nuevas tecnologías. Ellas incluyen

- Data Centers
- Cloud computing
- Hadoop
- Internet of Things (IoT)

Aplicaciones de Big Data

Algunos de los problemas de Big Data que involucran fuertemente matemáticas son:

1. Análisis de redes

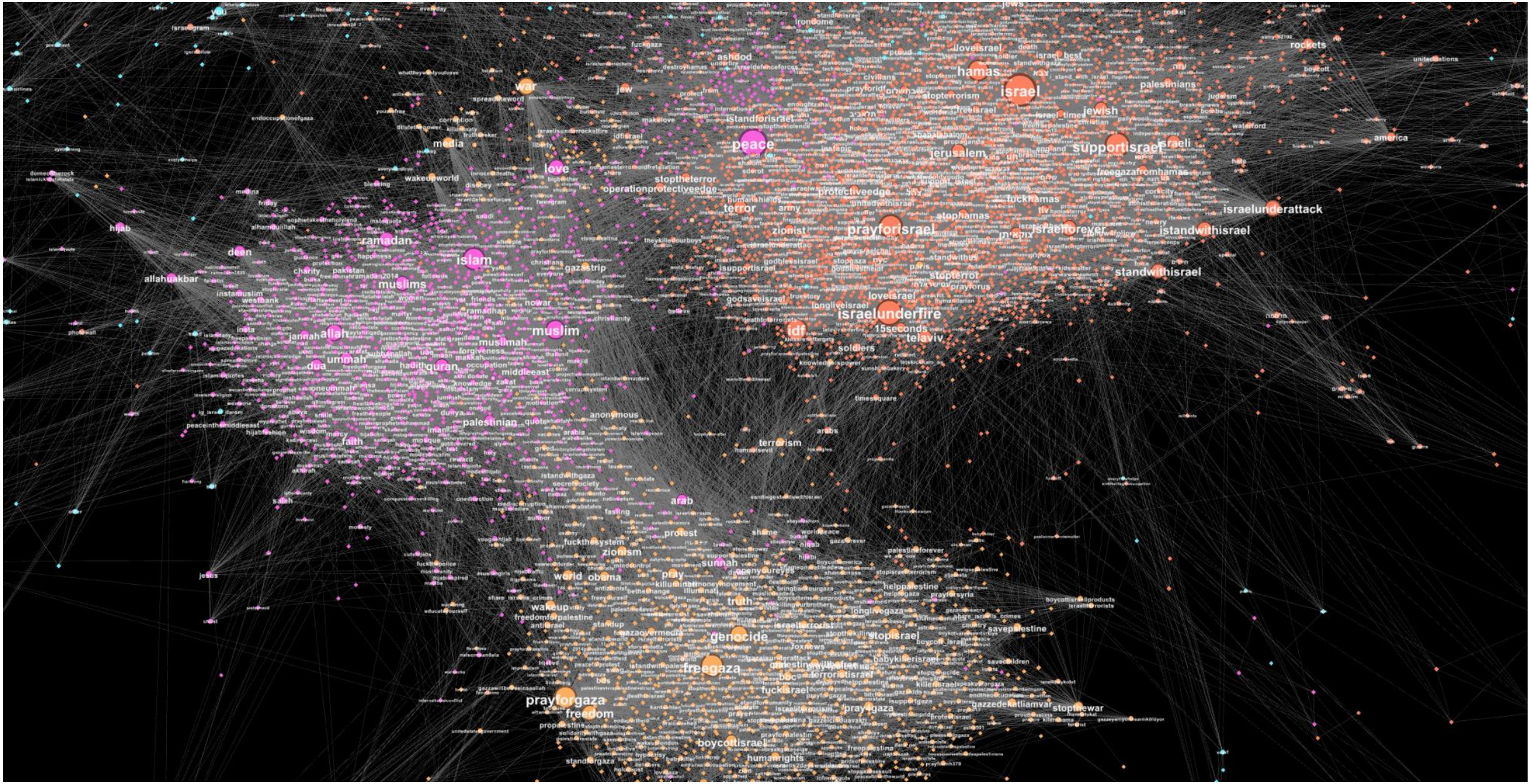
- Detección de grupos (K means, por ejemplo)
- Detección de comunidades (Machine learning)
- Detección de topologías. Como el de los nodos más influyentes. Por ejemplo

Topología de estrella es útil para encontrar los influyentes que pueden motivar a un gran número de personas

Topología de anillo puede ayudar a determinar los nodos críticos para transferencia de información

Topología de red puede ayudar en categorizar comunidades activas que son candidatas en promover ofertas, ideas, información (Krishna, 2012).

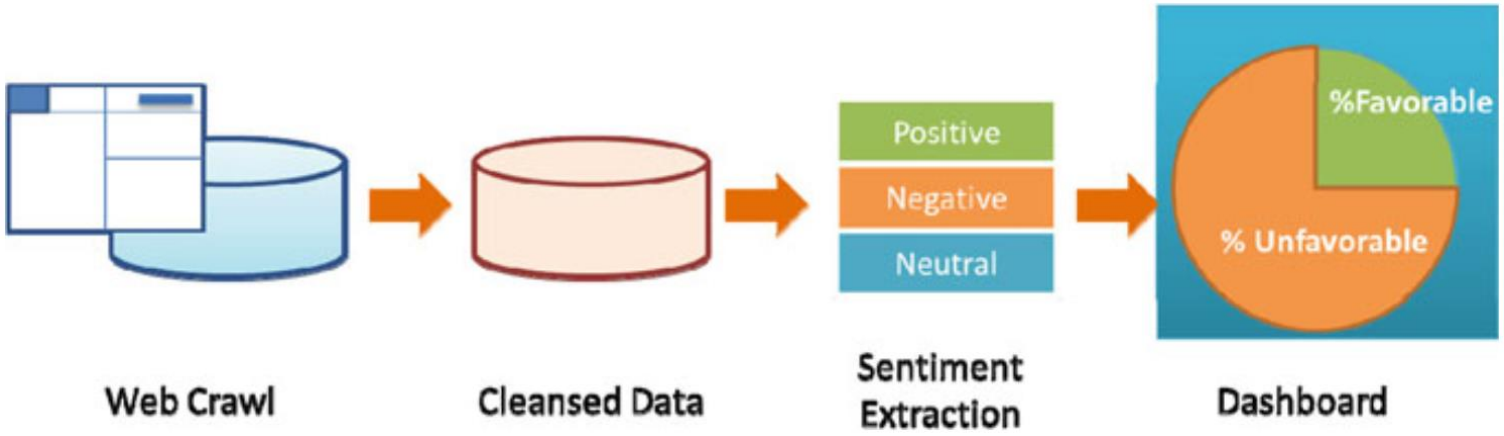
Detectar los rasgos topológicos más importantes ayuda a seleccionar una visualización más informativa.



fuente:
http://globalvoicesonline.org/2014/08/04/israel-gaza-war-data-the-art-of-personalizing-propaganda/

Aplicaciones de Big Data

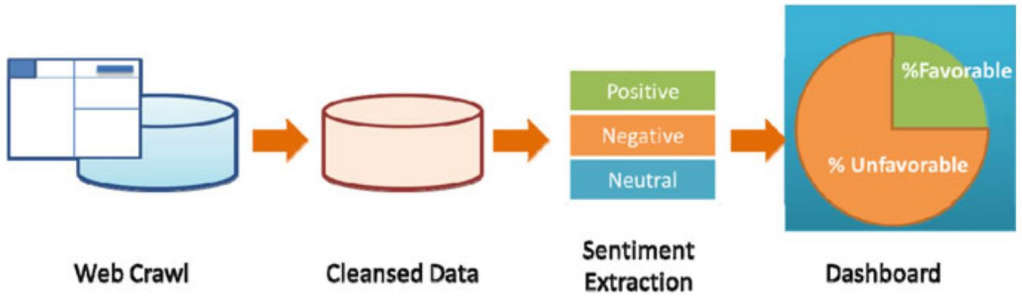
2. Sentiment mining (Análsis de sentimiento).



Mohanty, et al. (2015).

Aplicaciones de Big Data

2. Sentiment mining (Análsis de sentimiento).



Natural language processing. Es un área de ciencias de la computación e IA para la interacción entre computadoras y humanos. Retos: reconocimiento de voz (speech recognition), comprensión de lectura (natural language understanding) y generación de lenguaje natural.

Text analysis (text mining). Es el proceso de extraer información de alta calidad a partir de texto.

Computational linguistics. Se relaciona con la estadística para modelar el lenguaje natural. Se desarrolla en equipos Interdisciplinarios que suelen incluir lingüistas y científicos de la computación.

Biometrics. Se relaciona con la obtención y análisis de medidas corporales.

Aplicaciones de Big Data

3. En Salud

Genomics (y otros omics)

Red médica

Herramientas de vigilancia epidémica basadas en internet

Monitorieo de condiciones de salud y calidad de comestibles. La incorporación de aplicaciones y sensores que monitorean la salud y los comestibles.

EVA

Julian Ríos y Antonio Torres

Aplicaciones de Big Data

4. En ciencias ambientales
   - Predicción basada en información satelital
   - Predicción, tendencias de variables usando diversas fuentes de información
5. En la banca y finanzas
   - Detección de Fraudes
   - Lavado de Dinero
   - Análisis de Riesgo
6. En el comercio
   - Recomendaciones de productos
   - Predicción de tendencias
7. En la manufactura
   - Mantenimiento preventivo
   - Predicción de la demanda
8. En la prevención del delito

Actualmente, qué es Big Data?

Es una área del conocimiento multidisciplinaria. Dentro de las matemáticas involucra fuertemente áreas como la ciencia de la computación, estadística, probabilidad, la geometría, topología y el análisis numérico, entre otras.

Big Data es más grande que los datos grandes.

Con que temas entran cada una de las áreas?

Computación

Inteligencia Artificial:

Redes Neuronales, Machine learning, Deep learning

Estadística

Análisis de datos de gran dimensión (HDDA) -> (problema n<<p):

Selección del modelo

Métodos para hacer más eficiente la inferencia o predicción:

Métodos basados submuestreo, Leveraging (randomised numerical linear algebra), Divide and conquer

Con que temas entran cada una de las áreas?

Probabilidad

Matrices y grafos aleatorios

Topología

TDA. Está en la intersección de la **topología algebraica** y la **estadística** con el principal objetivo de describir la “forma” de objetos multidimensionales y descubrir patrones o cualidades.

Redes complejas

Una nota sobre la seguridad y privacidad

Preocupación por las posibles violaciones a la privacidad y confidencialidad.

No solo se refiere al uso no autorizado y la filtración de información (Ej. Google, Facebook).

Sino que puede incluir la predicción que se ha intentado hacer por instancias para predecir información personal, como su estatus.

Reiter (2012) detalla los recursos estadísticos para proteger la confidencialidad.

Google enfrenta demanda por invasión de privacidad

Revistas científicas de Big Data

* **Journal of Big Data**. Springer Open. https://journalofbigdata.springeropen.com/

* **Big Data & Society.** Sage. http://journals.sagepub.com/home/bds

Big Data Research. ScienceDirect. Elsevier. https://www.sciencedirect.com/journal/big-data-research

* **Big Data Analytics.** Springer Nature. https://bdataanalytics.biomedcentral.com/

IEEE Transactions on Big Data.
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6687317

* **Big Data Mining and Analytics**. IEEE Xplore
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=8254253

Big Data…una nueva disciplina

Diebold (2012) opina que

Big Data no es sólo un término de moda ni un fenómeno, sino una disciplina ya que el fenómeno que le motiva es “nuevo” y muy real. Aunque se puede considerar que disciplinas tradicionales son capaces de enfrentar este fenómeno, los métodos alrededor de Big Data son más que la suma de las partes que de ellas vienen. Big Data está motivado nuevas investigaciones y llevándonos a nuevos lugares que eran inimaginables hace algunas décadas. Desde computo en la nube o paralelo masivo a métodos para lidiar con millones de pruebas de hipótesis.

“…Big Data is emerging as a major interdisciplinary triump”.

Bibliografía

- S. Pyne; B.L.S. P. Rao y S.B. Rao (2016) *Big Data Analytics: Methods and Applications*. Springer.
- M. Chen, S. Mao, Y. Zhang, and V. C. Leung (2014) *Big data: related technologies, challenges and future prospects.* Springer.
- R. Krishna, P., Indukuri, K.V., Syed, S. (2012) A generic topology discovery approach for huge social networks. In: ACM COMPUTE 2012, 23–24
- Mohanty, H., Bhuyan, P., & Chenthati, D. (Eds.). (2015). *Big Data: A Primer* (Vol. 11). Springer.
- Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and challenges of big data computing in health sciences. *Big Data Research*, *2*(1), 2-11.
- Reiter, JP (2012) Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. Public Opin Q 76(1):163–181
- Wang, Chun, et al. (2016) Statistical methods and computing for big data. *Statistics and its interface* 9.4: 399.