Introducción a Ciencia de Datos

Organización del curso

- Introducción a Ciencia de Datos y Big Data
- Técnica de Map-Reduce
- Hadoop/Sparc (a definir)
- Aprendizaje Automático (Machine Learning)
- Visualización

Agradecimientos

- Este curso es una versión libre del curso on-line Introducción a la Ciencia de Datos (Bill Howe, Univ. de Wasshington) https://www.coursera.org/course/datasci
- El Dr. Esteban Feuerstein nos facilitó parte del material introductorio

Definición informal

"La Ciencia de Datos es un área de trabajo nueva vinculada a la recolección, adaptación, análisis, visualización y preservación de grandes volúmenes de datos" (An Introduction to Data Science, Jeffrey Stanton)

La Ciencia de Datos aparece en el cruce de 3 áreas:

- Procesamiento de datos (depuración y formateo de datos)
- Técnicas de aprendizaje automático
- Visualización (comunicar los resultados)

La Ciencia de Datos y los *Productos de Datos*

La Ciencia de Datos no sólo se ocupa de responder preguntas también se ocupa de hacer aplicaciones (*Productos de Datos*)

Algunos ejemplos:

- Aplicaciones de Datos (correctores ortográficos, traductores)
- Visualizaciones interactivas (Google Maps, Mapa de la Gripe)
- Bases de Datos on-line (Sloan Digital Sky Survey)

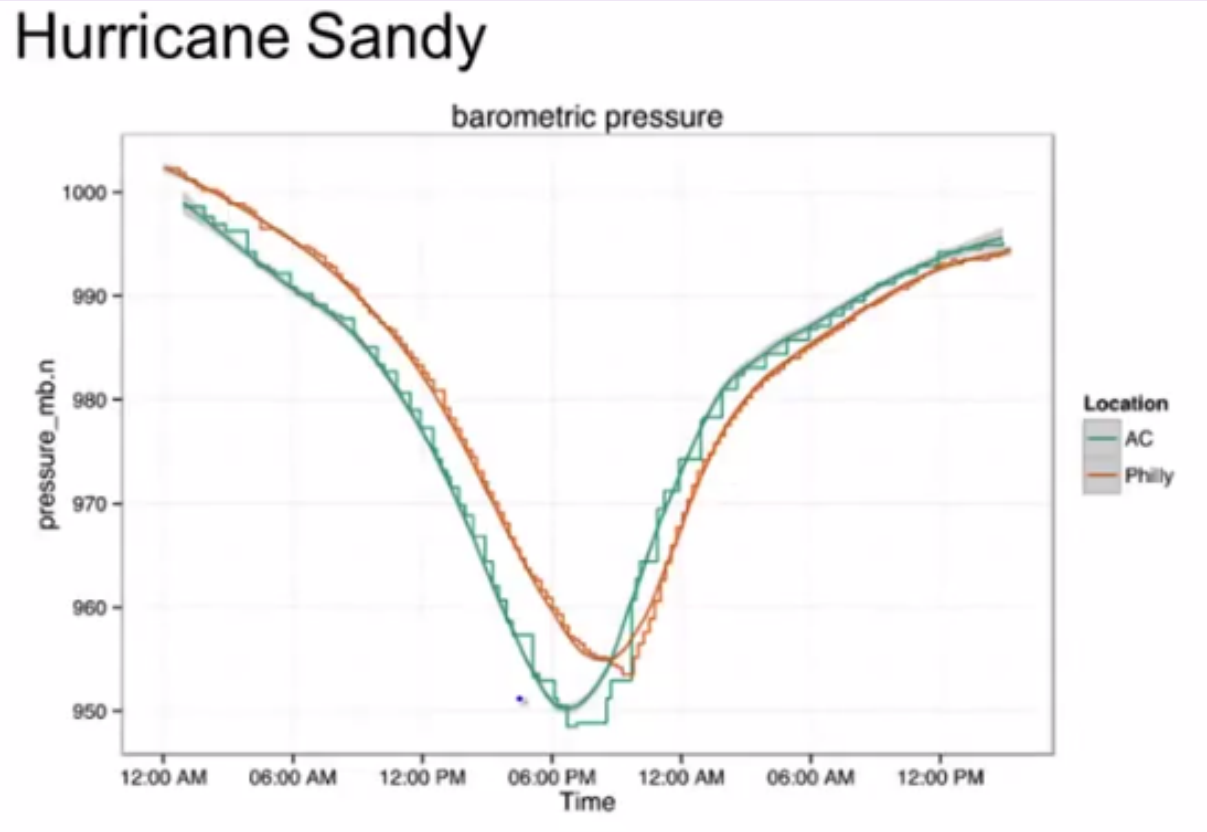
Campañas electorales

En la actualidad se emplean metodologías de análisis de datos para definir los distintos grupos socioculturales a los cuales se trata de influir durante las campañas electorales

Resignificar la información

Emplear la información disponible en internet de formas novedosas

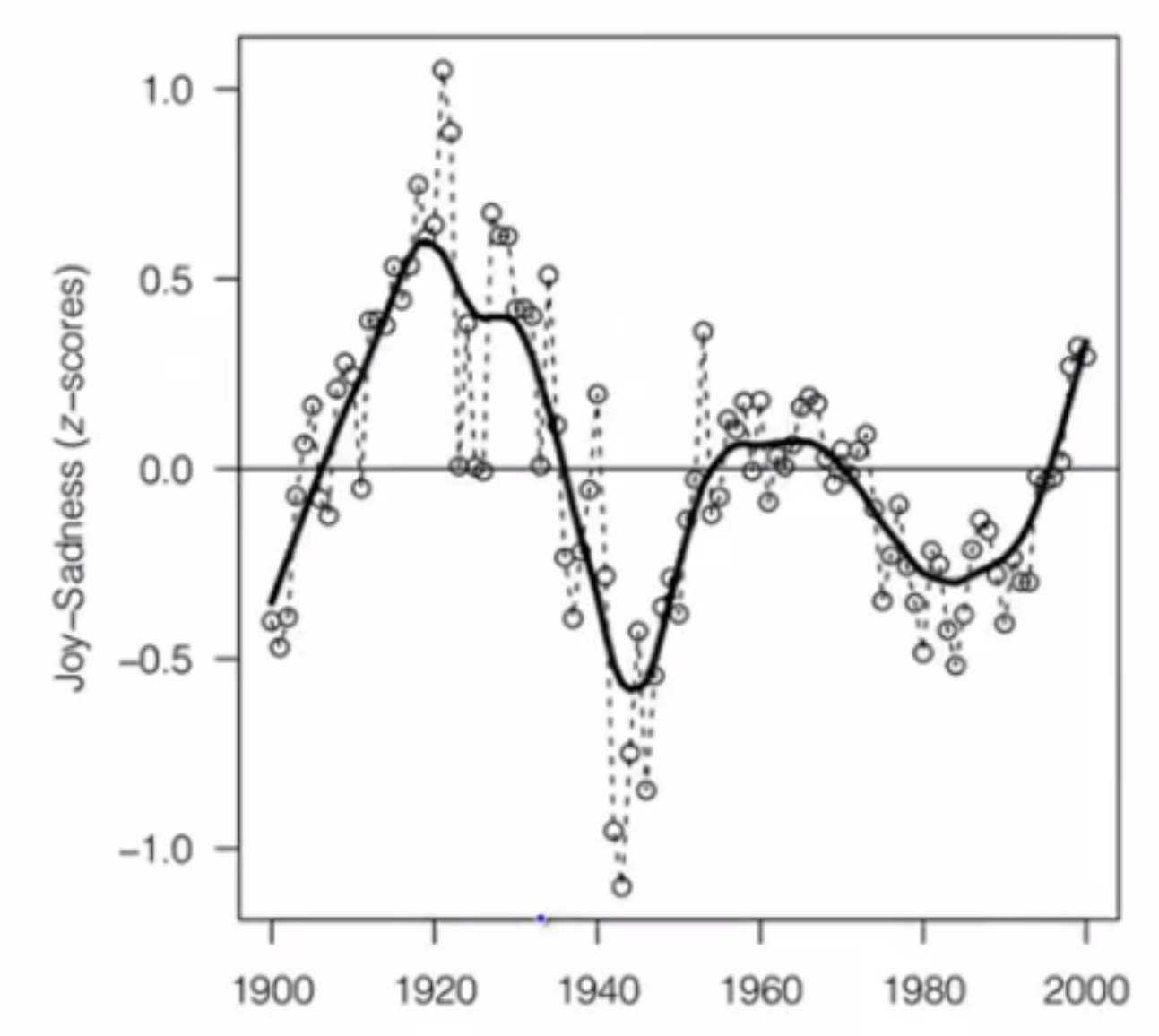
- Ejemplo: Huracán Sandy tomar información de estaciones meteorológicas y producir una visualización en tiempo real del paso del huracán



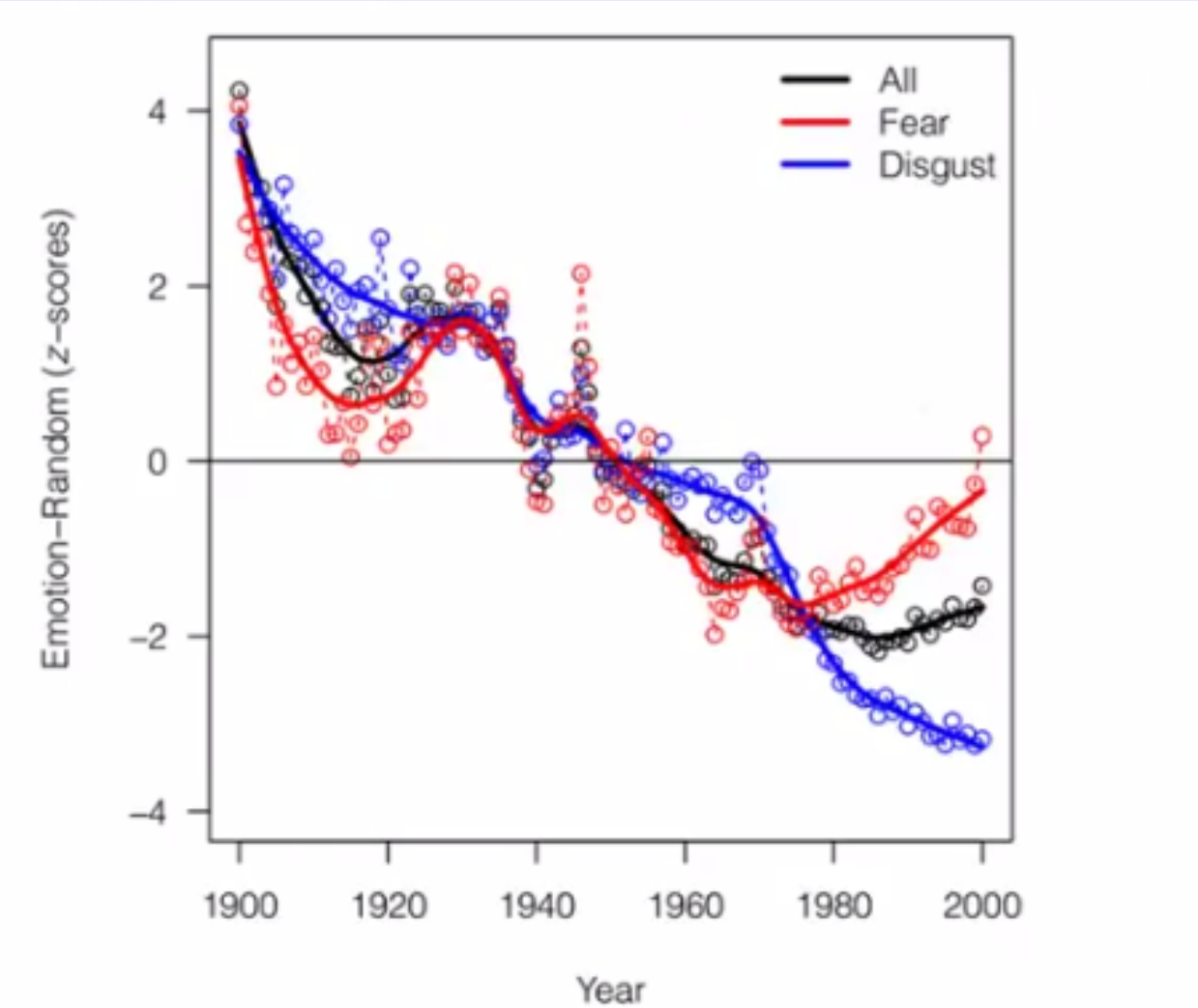
Análisis de emociones en textos

1. Convertir todos los libros digitalizados a n-gramas (Google https://books.google.com/ngrams)
   - 1-grama: "Inflación"
   - 6-grama: "El costo de vida aumenta diariamente"
2. Asignarle a cada 1-grama un puntaje "emocional" (WordNet https://wordnet.princeton.edu/)
3. Contar y normalizar las ocurrencias de cada palabra

Análisis de emociones en textos



Análisis de emociones en textos



La Ciencia de Datos está revolucionando otras áreas

- Métodos novedosos de análisis de datos se emplean en ciencias sociales: historia, antropología, lingüística
- El periodismo de investigación (Wikileaks)

Casos de error

Se produjo un error en la estimación del pico de gripe a partir de las consultas en Google porque la prensa le dio difusión y se magnificó la intensidad

“El mayor obstáculo en el procesamiento de datos proviene de la gran variedad de formatos existentes, de la información no estructurada y de las fuentes contradictorias.” (Doug Laney)

eScience

eScience es una nueva perspectiva en las ciencias empíricas según la cual se capturan datos masivamente para emplearlos en la verificación de las hipótesis científicas.

Gracias a los avances tecnológicos recientes se abarató mucho el costo de la obtención de datos.

El ritmo al que se producen los datos supera ampliamente nuestra capacidad actual para analizarlos.

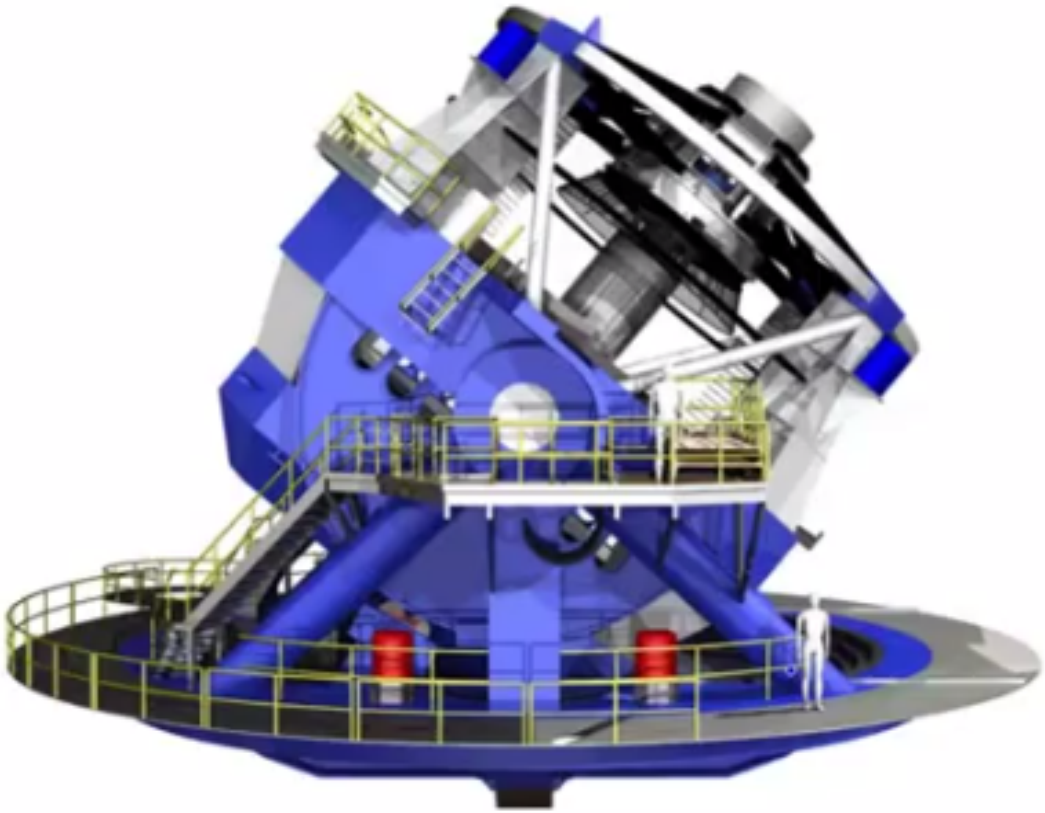
eScience: Astronomía

Es posible hacer mapeos periódicos de la galaxia en alta resolución

Large Synoptic Survey Telescope (LSST)

40TB/day (an SDSS every two days), 100+PB in its 10-year lifetime

400mbps sustained data rate between Chile and NCSA

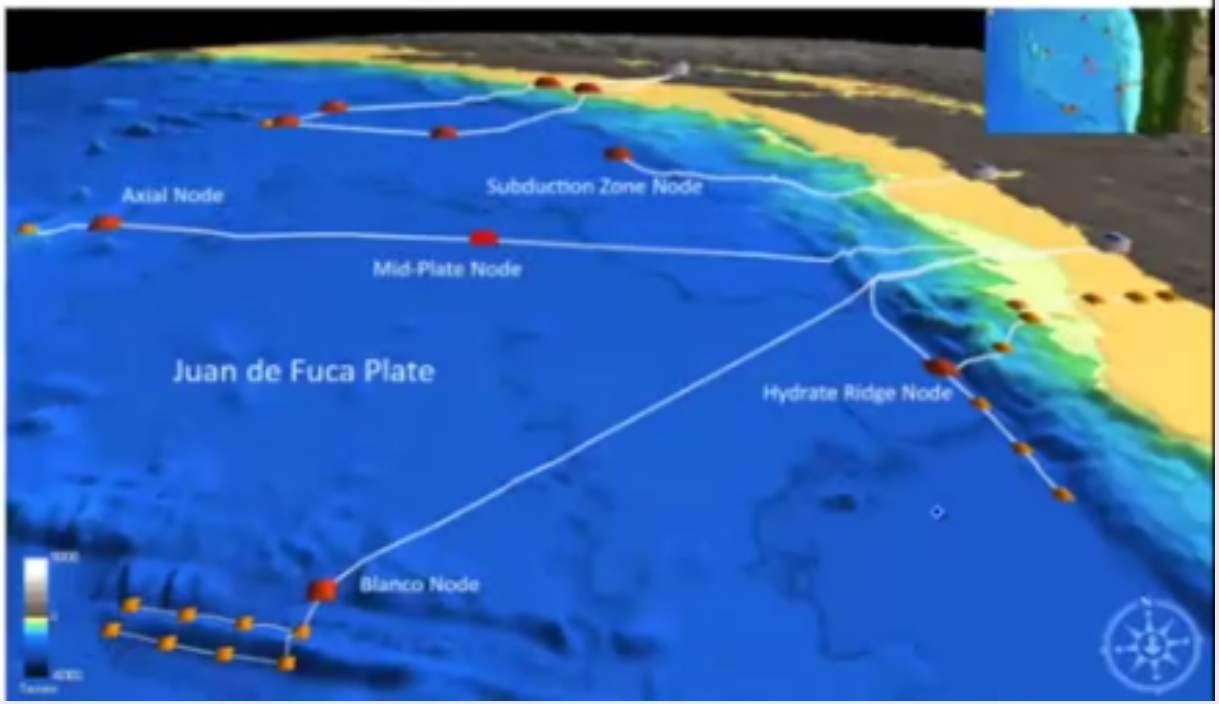


eScience: Oceanografía

Modelos de alta resolución, sensores baratos, satélites

Regional Scale Nodes of the NSF Ocean Observatories Initiative

1000 km of fiber optic cable on the seafloor, connecting thousands of chemical, physical, and biological sensors

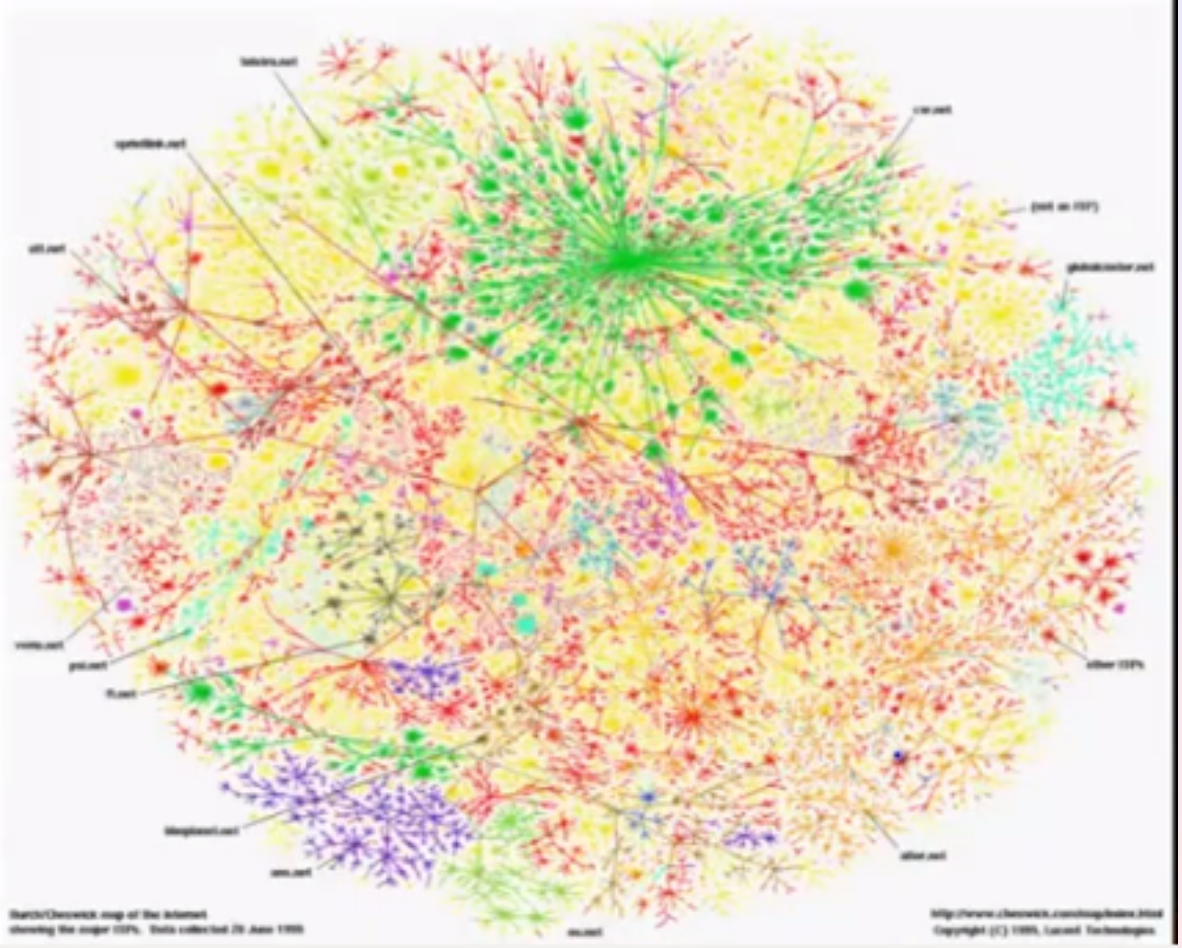


Internet

The Web

20+ billion web pages x 20KB = 400+TB

One computer can read 30-35 MB/sec from one disk => 4 months just to read the web



Big Data: una definición informal

"*Big Data* es cualquier conjunto de datos costoso de mantener y difícil de procesar." (Michael Franklin, Univ. Berkeley)

Big Data: las 3 V's

- Volumen: el tamaño de los datos
- Velocidad: la relación entre el procesamiento de datos y la creciente necesidad de interactividad (tiempo real)
- Variedad: la diversidad de fuentes de datos, formatos, estructuras, etc.