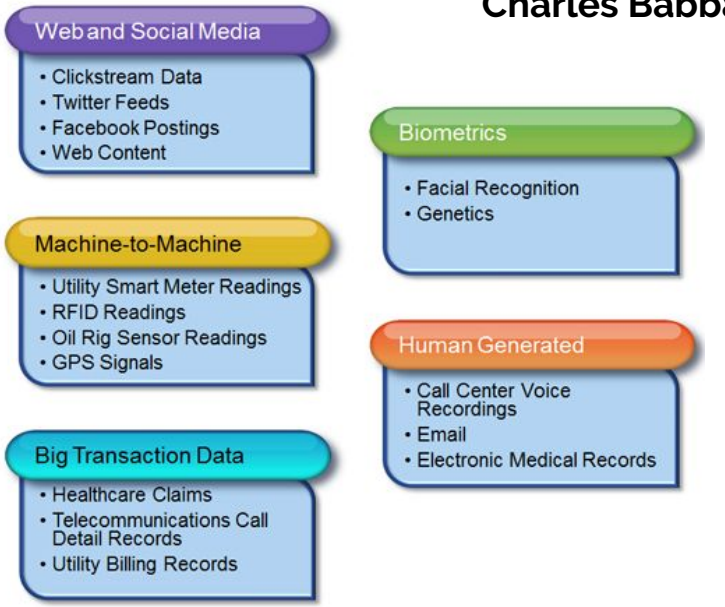
"The world is one big data problem."

Rodrigo García

# “Errors using inadequate data are much less than those using no data at all”

**Charles Babbage.**




Big Data no se refiere a alguna cantidad en específico.

Además del gran ***volumen*** de información, esta existe en una gran ***variedad*** de datos que pueden ser representados de diversas maneras en todo el mundo. Junto a **Velocidad** hacen las 3 *V* del Big Data.

# “Big data is not about the data”

**Gary King.**

- Data:
  - becoming commoditized
  - easy to come by; often a free byproduct of IT improvements
  - Ignore it & your company will still have more every year
  - With a bit of effort: huge data production increases
- Where the Value is: the Analytics
  - Output can be highly customized
  - Moore’s law (doubling speed/power every 18 months) v. 1000x increase with one algorithm
  - $2M computer v. 2 hours of algorithm design
  - Low cost; little infrastructure; mostly human capital needed
  - Innovative analytics: enormously better than off-the-shelf approaches

# “ With big data, researchers have brought cherry-picking to an industrial level.”

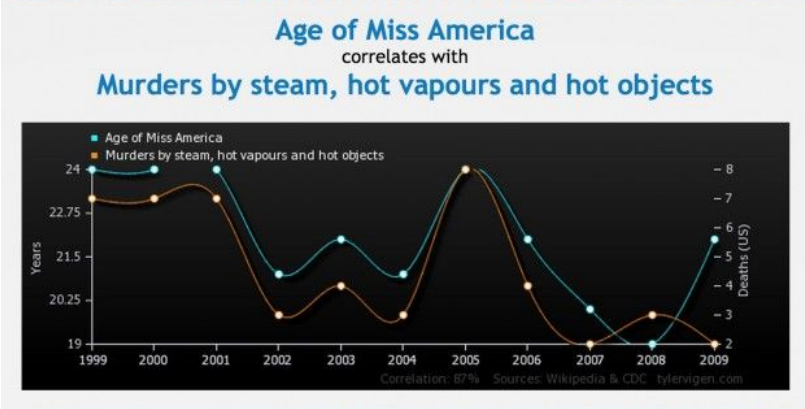
**Nassim N. Taleb**

La **falacia de prueba incompleta**, **supresión de pruebas**, o por su designación en inglés ***cherry picking*** (seleccionar lo mejor de algo, o bien, seleccionar lo peor de algo, o bien, seleccionar algo "a la medida"), es la acción de citar casos individuales o datos que parecen confirmar la verdad de una cierta posición o proposición, a la vez que se ignora una importante cantidad de evidencias de casos relacionados o información que puede contradecir la proposición.
Actualmente se posee demasiadas variables, pero poca información por variable, por lo que las relaciones crecen demasiado, más rápido que la información real. En otras palabras: Big Data puede significar más información, pero también significa falsa información.

# "Big Data is not a substitute for Big Ideas"

**Joseph Coughlin**

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?



# "Big Data is not a substitute for Big Ideas"

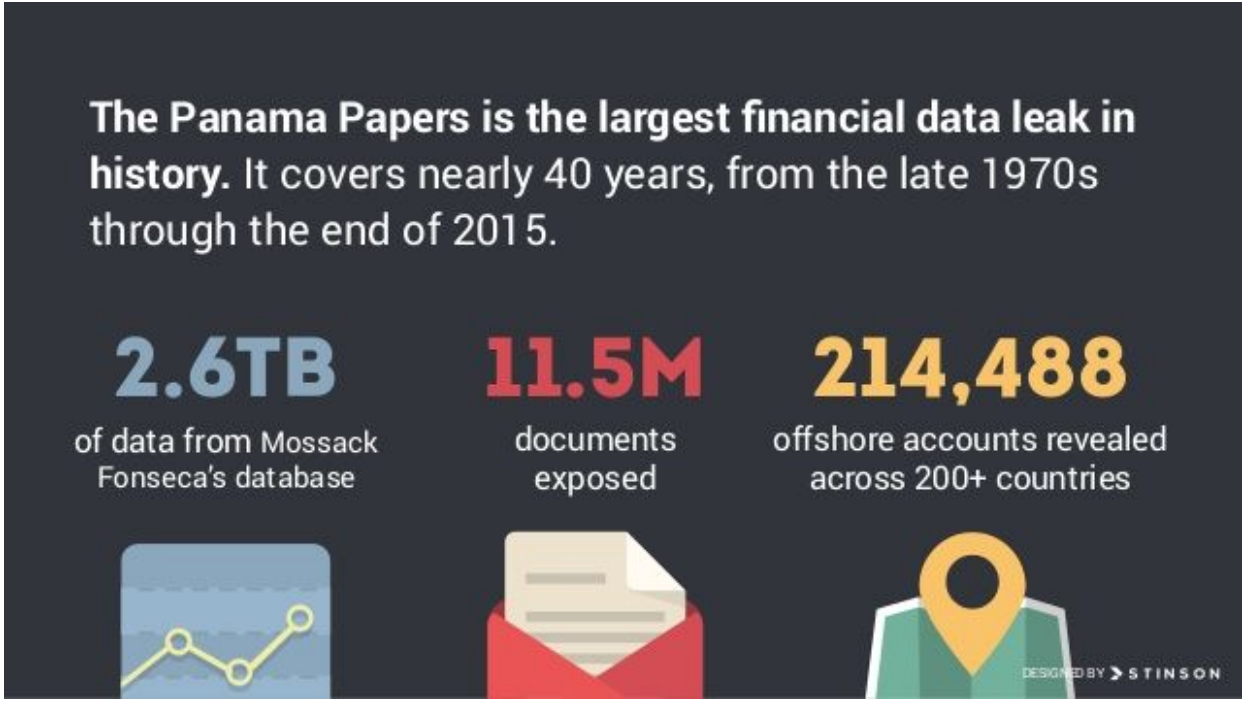
## Joseph Coughlin

- **Lineberger Comprehensive Cancer Center - Bioinformatics Group** utiliza Hadoop y HBase para analizar datos producidos por los investigadores de The Cancer Genome Atlas(TCGA) para soportar las investigaciones relacionadas con el cáncer.
- El **PSG College of Technology**, India, analiza múltiples secuencias de proteínas para determinar los enlaces evolutivos y predecir estructuras moleculares. La naturaleza del algoritmo y el paralelismo computacional de Hadoop mejora la velocidad y exactitud de estas secuencias.
- La **Universidad de Maryland** es una de las seis universidades que colaboran en la iniciativa académica de cómputo en la nube de IBM/Google. Sus investigaciones incluyen proyectos en la lingüística computacional (machine translation), modelado del lenguaje, bioinformática, análisis de correo electrónico y procesamiento de imágenes.

# "The world is one big data problem."

**Andrew McAfee**

Los 11.5 millones de documentos enviados al Süddeutsche Zeitung y al International Consortium of Investigative Journalists (ICIJ) incluían 5 millones de emails, 3 millones de archivos de database, 2 millones de PDFs, 1 millón de imágenes, 230 mil documentos de texto y más de 2 mil archivos no calificados.

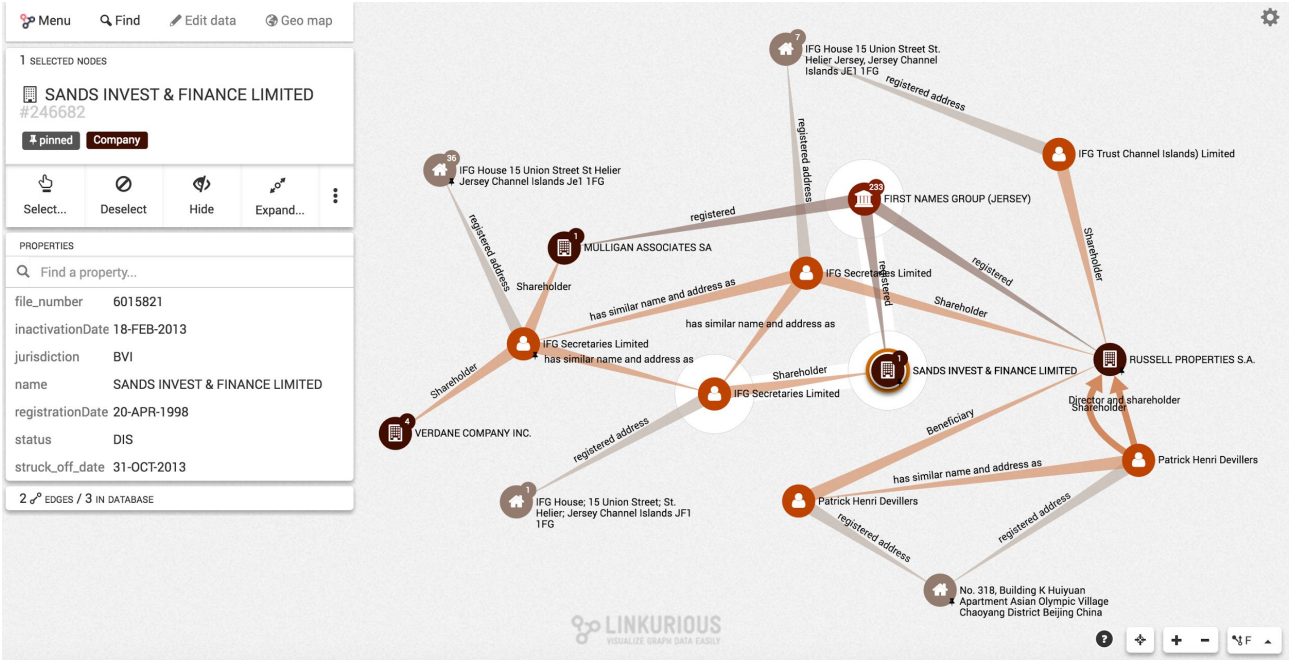


# "The world is one big data problem."

## Andrew McAfee

El ICIJ contrató a la firma de analytics Nuix, quienes utilizan una herramienta de indexación que extrae textos y metadatos de los documentos, para luego poder hacer consultas y encontrar la relación entre datos.

El proceso que tomó 2 semanas en hacer de la data algo consultable para encontrar conexiones, patrones y relaciones.



# "The world is one big data problem."

**Andrew McAfee**

Los Panama Papers son una muestra del camino a seguir. Ahora es posible recolectar y analizar información más rápido que nunca mediante el uso de métodos del machine learning, como el deep learning.
Información no estructurada, como textos en posteos social media, emails, notas, etc., representan una gran oportunidad para los negocios que puedan aprovecharlos.

**Aún hay margen para futuras revelaciones cuando los periodistas e investigadores consigan añadir más criterios de búsqueda y encontrar nuevas relaciones entre los nombres y los datos.**

**“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**

GRACIAS!

Rodrigo García
rodrigo.garcia@globallogic.com