

# **Modulo “Análisis de Datos Científicos y Geográficos”**

## **INTRODUCCIÓN**

**ITBA**



## OBJETIVO BÁSICO DE ESTE MODULO...

➤ **Soy parte de una compañía?**  
Puedo enriquecer las decisiones



➤ **Soy un consultor independiente?**  
Puedo detectar sectores donde ofrecer mis servicios





## BIG DATA: UNA REVOLUCIÓN QUE TRANSFORMARÁ CÓMO VIVIMOS, TRABAJAMOS Y PENSAMOS

*“El Big Data es **ilimitado** y no estructurado; **impreciso** pero **predecible**, y no puede demostrar relaciones de causalidad, pero pueden mostrar **correlaciones**”*

*(Viktor Mayer-Schönberger y Kenneth Cukier)*

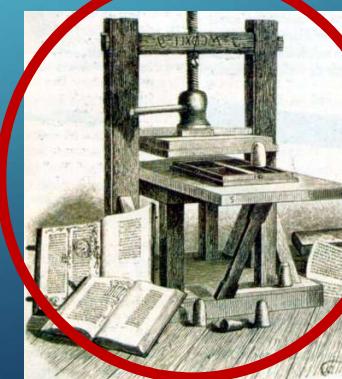
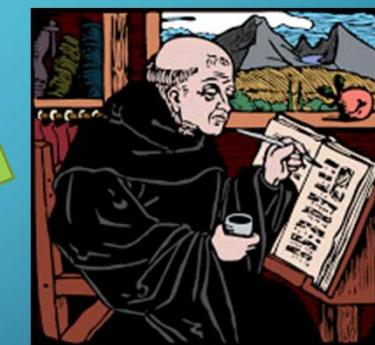
# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...

DATOS  
ORALES

Cantidad limitada  
de información  
con baja precisión



# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...

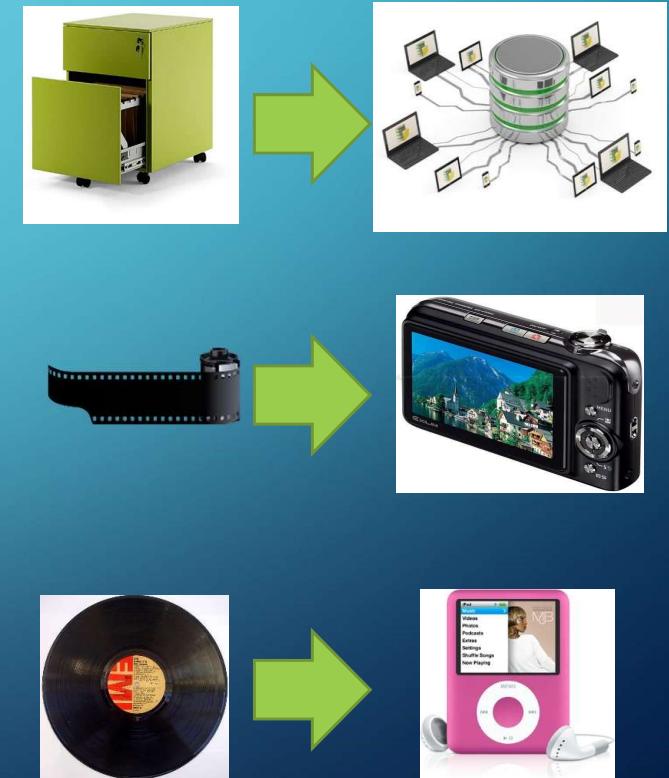


Aumenta la cantidad  
de registros y su  
accesibilidad

# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...



Aumenta  
indescriptiblemente la  
cantidad de registros



# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...



Datos Propios  
+ Metadatos !

# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...

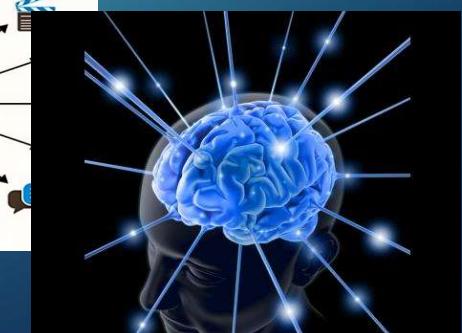


## Consecuencia

Instante a instante se tiene una colección heterogénea de enorme cantidad de datos y metadatos...  
Correlacionarlos y analizarlos se vuelve altamente complejo !

**VOLUMEN + VARIEDAD + VELOCIDAD**

# LA EVOLUCION DE LOS REGISTROS DE EXPERIENCIAS...

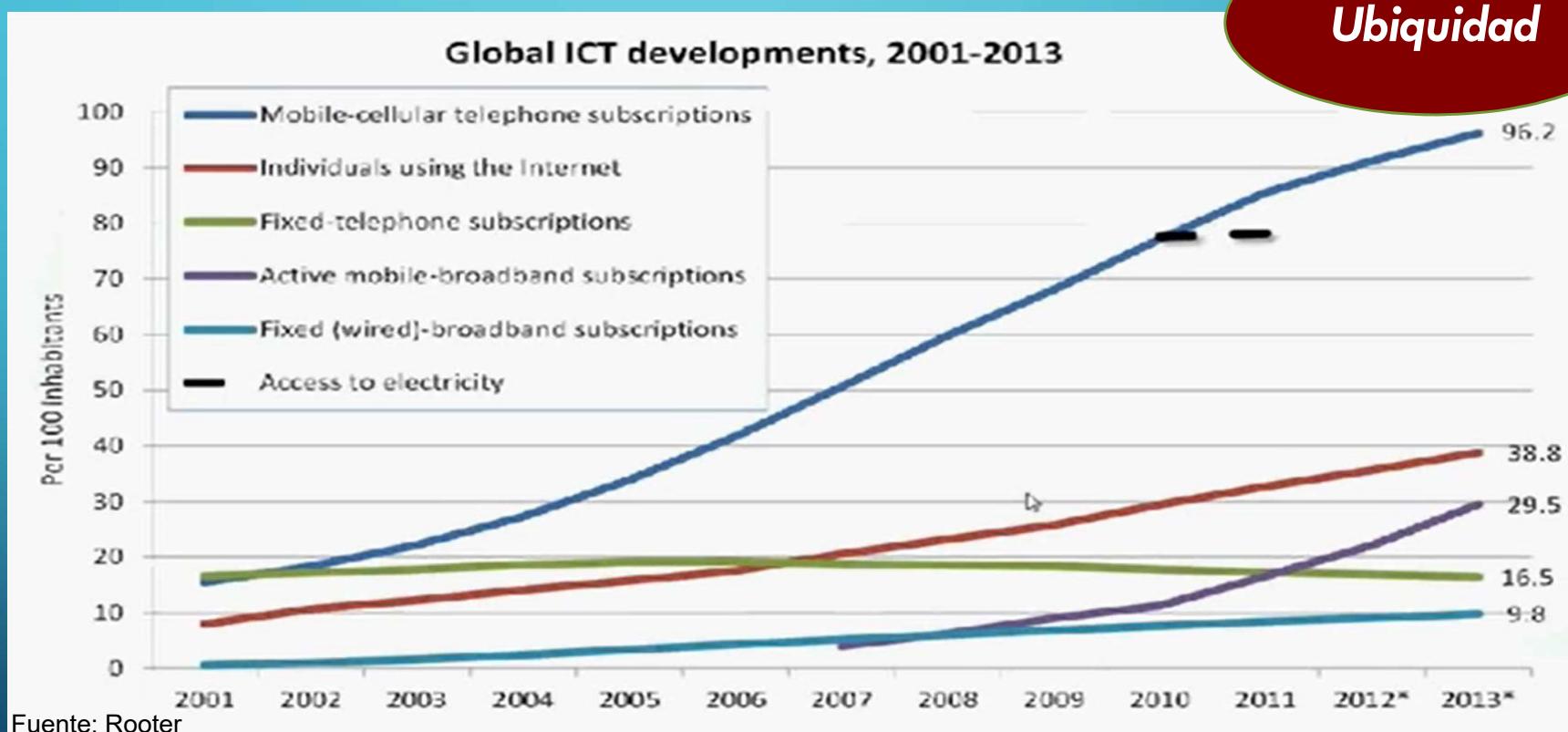


## LA ERA BIG DATA

**TECNOLOGIA CAMBIA →  
CAMBIAN NUESTROS HABITOS,  
NUESTRO COMPORTAMIENTO...**



## LA ERA BIG DATA





## LA ERA BIG DATA

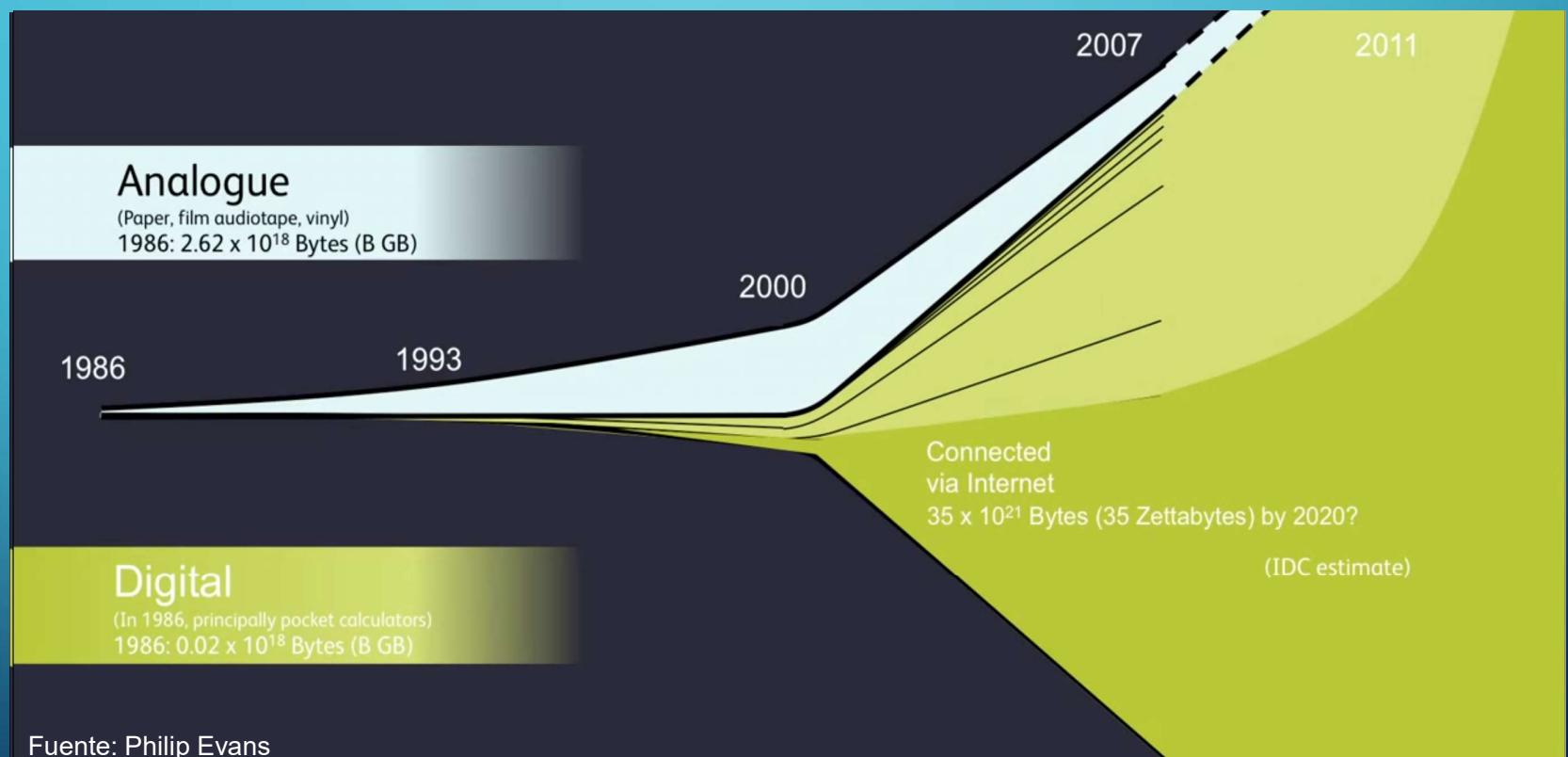
La revolucion digital ha cambiado nuestra forma de crear conocimiento.

La produccion analogical crece 5% cada año...  
pero la produccion digital crece un 30% por año

El acceso a celulares incrementa aun mas esta diferencia

Fuente: Philip Evans

## LA ERA BIG DATA



# LA ERA BIG DATA



**Información Estructurada**

**Consultas en Internet**

**Blogs y RRSS**

**Comercio electrónico**

**Imágenes y videos**

**Sonidos**

**Acciones frente a Buscadores**

**Ubicaciones (GPS+acelerómetro)**

...

## LA ERA BIG DATA

Podemos estimar en tiempo real  
para hacer predicciones ?



Los datos son incompletos...  
... pero hay tantos que no importa?



Muestreo  $\cong$  Universo?





## LA ERA BIG DATA

- **Analizo el PASADO → busco CAUSALIDAD**
- **Futuro se parece al Pasado → PREDIGO**
- **Creo un MODELO → cambio variables y sigo prediciendo**
- **Simulacion Computacional vs. Modelos Matematicos ?**



EJEMPLO 1:



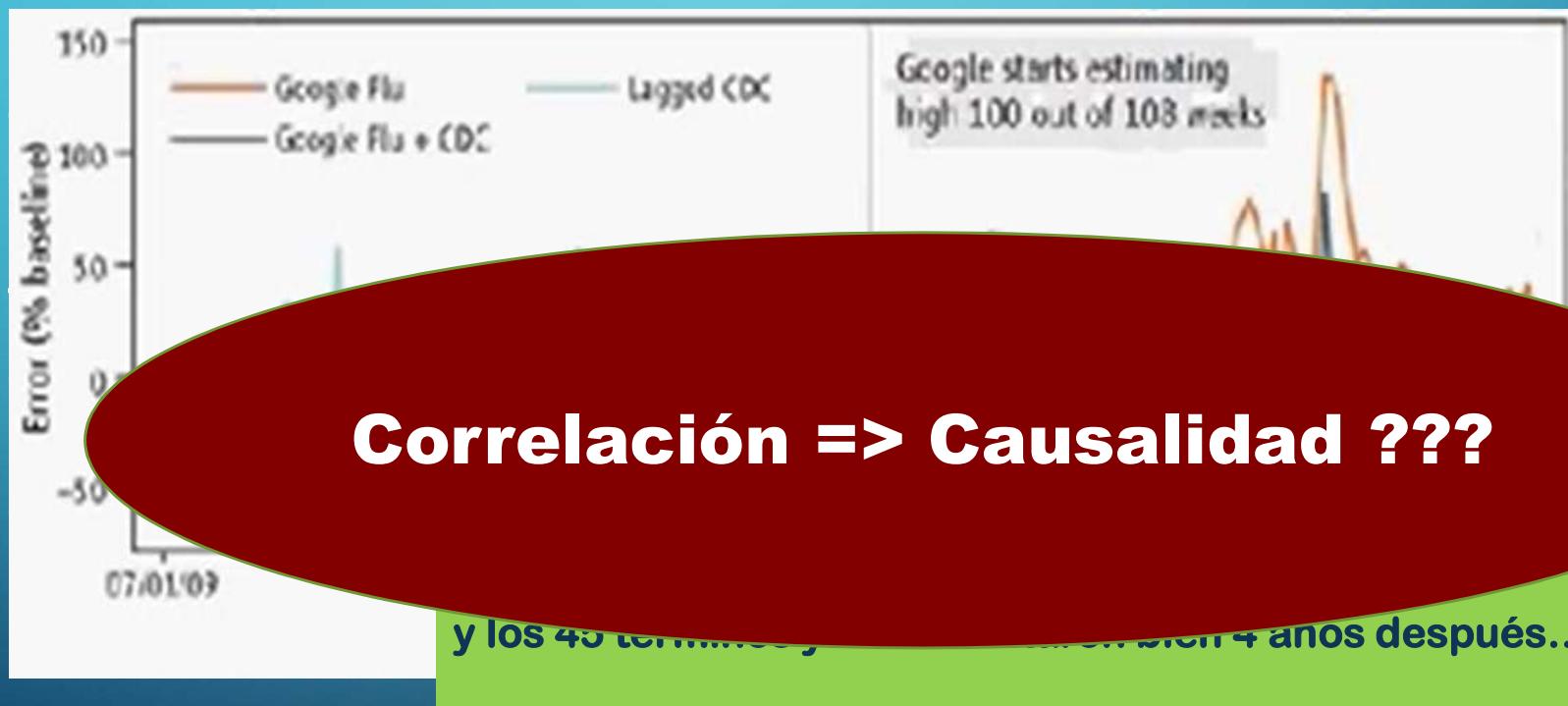
Google estimó perímetro de gripe haciendo una correlación entre las consultas su buscador sobre palabras relacionadas con gripe (tos, dolor de garganta, etc).

Tomaron 50 millones de términos de búsqueda y los correlacionaron con datos oficiales sobre la gripe.  
Corrieron 450 mil de modelos matemáticos y se quedaron con los 40 términos que mejor predijeron la gripe

## EJEMPLO 1:



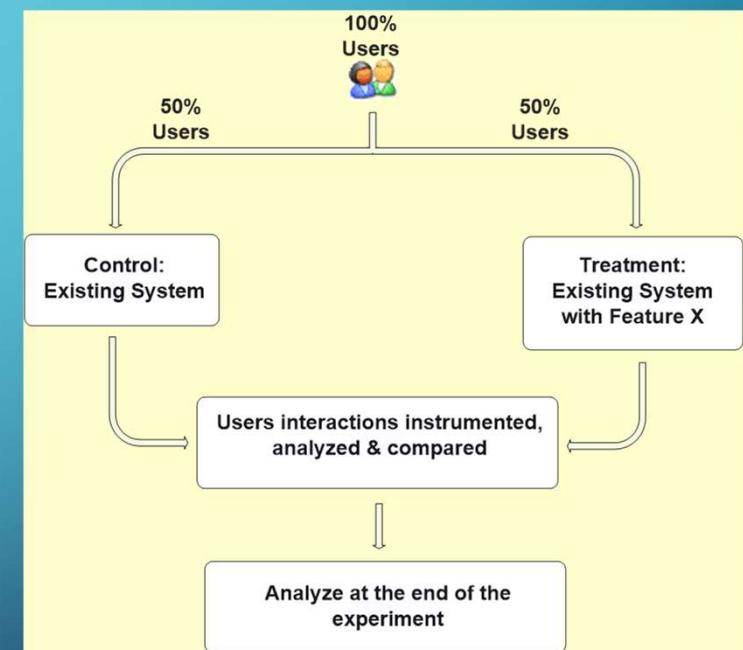
EJEMPLO 1:



## EJEMPLO 2: TEST A/B



- Los usuarios se dividen **aleatoriamente** entre dos o más grupos:
  - A (**Control → sistema sin cambios**)
  - B (**Tratamiento → el cambio que se quiere probar**)
- Se miden todas las actividades de los usuarios y del sistema
- Se calcular métricas de interés
- Se analizan los resultados y se comparar el delta de las métricas entre grupo A y B

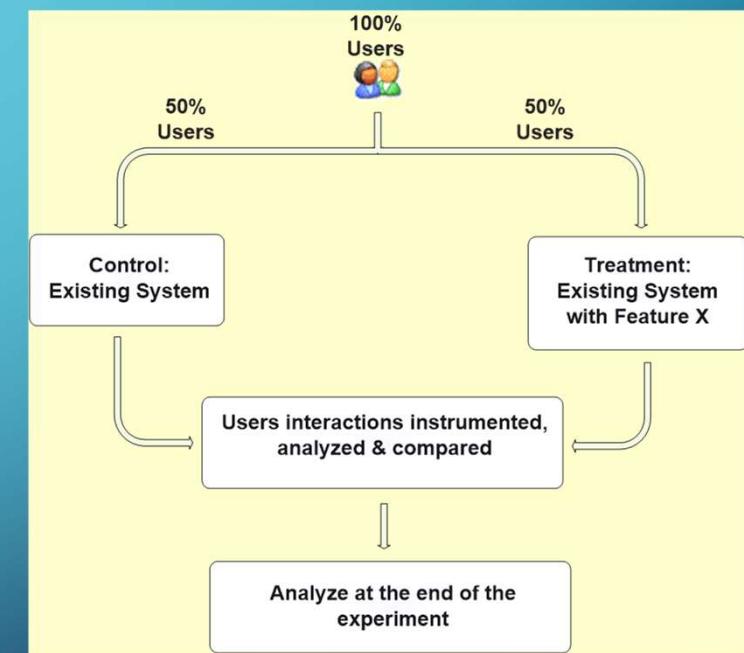


Fuente: Isidro Hegouaburu – Microsoft Azure

## EJEMPLO 2: TEST A/B

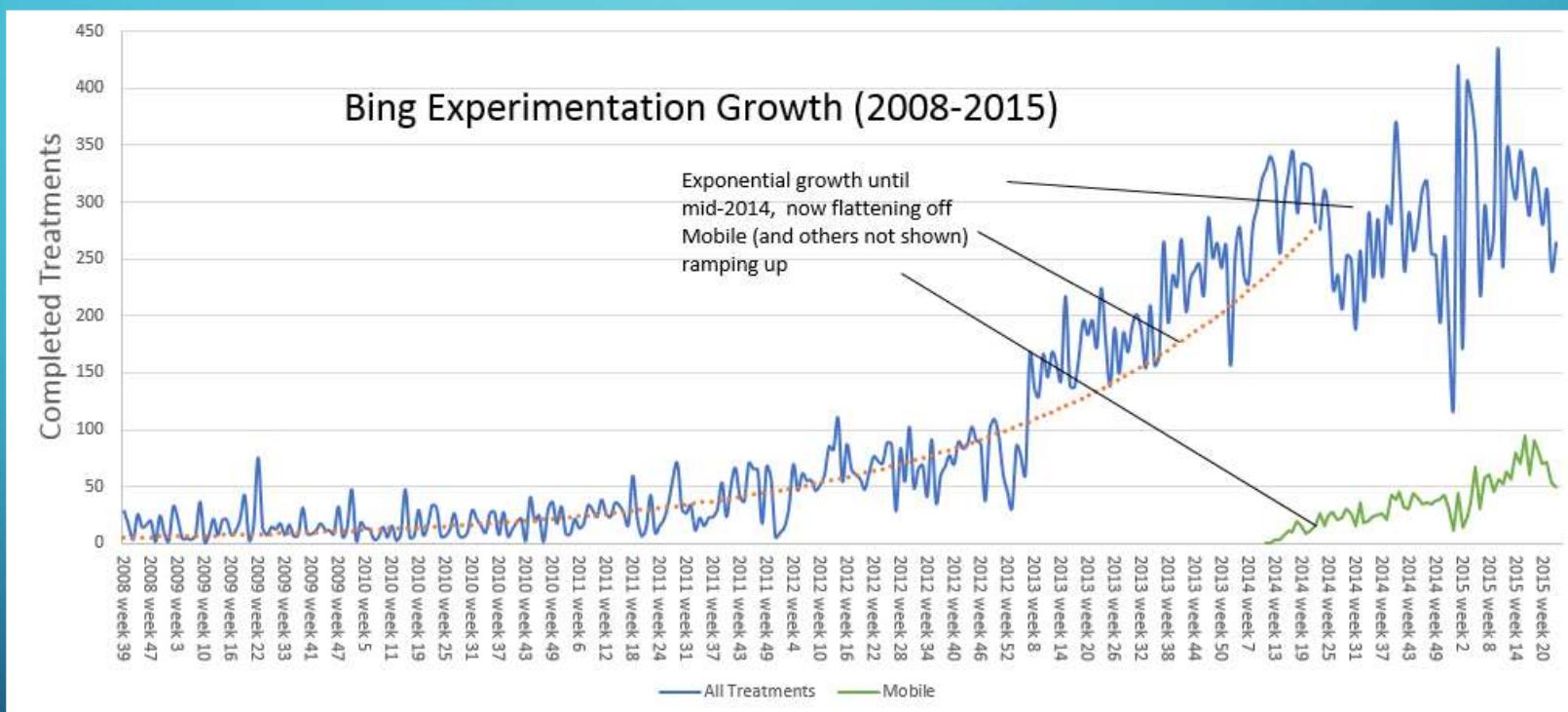


- Se usan **usuarios reales** del sistema
- Se deben correr tests estadísticos para confirmar que las diferencias observadas no son debido al azar
- El test A/B es la forma más simple de experimento controlado, la mejor forma científica de probar causalidad



Fuente: Isidro Hegouaburu – Microsoft Azure

## EJEMPLO 2: TEST A/B



- En 2017 completaban ~300 experimentos por semana en Bing
- 90% de los 150M de usuarios están en múltiples experimentos (15 promedio, con 5 variantes c/u)
- Hay  $5^{15} = 30.000$  millones de versiones de Bing en todo momento!

## EJEMPLO 2: TEST A/B



- En 2017 completaban ~300 experimentos por semana en Bing
- 90% de los 150M de usuarios están en múltiples experimentos (15 promedio, con 5 variantes c/u)
- Hay  $5^{15} = 30.000$  millones de versiones de Bing en todo momento!

## EJEMPLO 2 A: SEARCH BOX



A:

Web | MSN | Images | Video | News | Maps | Shopping

Live Sea

Popular Searches: Fireworks safety | Rihanna | Campaign patriotism

B:

Web | MSN | Images | Video | News | Maps | Shopping

Live Sea

Fireworks safety | Rihanna | Campaign patriotism

## EJEMPLO 2B: DEEP LINKS



A

[Esurance® Auto Insurance - You Could Save 28% with Esurance.](#) Ads  
www.esurance.com/California  
Get Your Free Online Quote Today!

B

[Esurance® Auto Insurance - You Could Save 28% with Esurance.](#)  
www.esurance.com/California  
Get Your Free Online Quote Today!  
Get a Quote · Find Discounts · An Allstate Company · Compare Rates



- La opción B es 5 mseg más lenta (más cómputo y página un poco más pesada)
- El límite de píxeles significa un promedio de 4 anuncios tipo “A” vs 3 anuncios tipo “B”
- **El experimento muestra que brinda mas ganancias!**

## EJEMPLO 3: LINKS SUBRAYADOS



A

Bing search results for "amazon". The results page shows several underlined links, such as "Amazon.com® Official Site - Amazon", "Amazon.com", "Amazon.com - Company", "Amazon.com - Kindle eBooks", "Amazon.com - Books", "Amazon.com - Stock price", "Amazon.com - Recent post", and "News about Amazon". A large green octagonal button with a white thumbs-up icon is overlaid on the left side of the results.

B

Bing search results for "amazon". The results page shows several non-underlined links, such as "Ad www.amazon.com", "Amazon.com - Company", "Amazon.com - Kindle eBooks", "Amazon.com - Books", "Amazon.com - Stock price", "Amazon.com - Recent post", and "News about Amazon". The overall design is cleaner and more modern than version A.

Pese a que queda demostrado que es mejor, los diseños modernos eliminan los links subrayados para ser mas "cool": **Bing pierde \$25m/año por no subrayar los links!**



## HOY EN DIA... CADA MINUTO....

- DATOS ESTRUCTURADIS (LONGITUD Y FORMATO DEFINIDOS)  
Y NO ESTRUCTURADOS (VIDEOS, AUDIO, IMAGENES, TEXTO LIBRE)
- DATOS INTERNOS Y EXTERNOS (APROX 75% NO ESTRUCTURADO)
- FUENTES: PERSONAS A MAQUINAS  
PERSONAS A PERSONAS  
MAQUINAS A MAQUINAS



## HOY EN DIA... CADA MINUTO....

Algunas cifras representativas:

90%

de la información en el mundo fue creada en los dos años anteriores  
(Bringing Big data to the Enterprise, IBM, 2012)

Twitter

Número promedio de tweets por día: 58 millones  
Usuarios activos por mes: 115 millones

Facebook

Usuarios activos en el mes: 1.440 millones  
Links compartidos: 1 millón cada 20 minutos

Whatsapp

Usuarios activos (enero 2015): 700 millones  
Mensajes acumulados (abril 2014): 64.000 millones

Google

Búsquedas (2014): 5.740 millones por día

Fuente: [www.statisticsbrain.com](http://www.statisticsbrain.com)

Fuente: Rooter



HOY EN DIA... CADA MINUTO....

**Tons of Spatial data out there...**



**twitter**  Geotagged Microblogs  Geotagged Pictures

 Medical Data  Smart Phones  Sensor Networks

 VGI  Satellite Images  Traffic Data

UNIVERSITY OF MINNESOTA **27**



## HOY EN DIA... CADA MINUTO....

**Aumenta la cantidad de sensores  
y la capacidad de los dispositivos  
que generan los datos**

**Se comparten datos a través de  
la comunidad y se incrementan  
los grupos multidisciplinarios**

**Aumenta la velocidad de  
transmisión (5G, 6G?)**





## HOY EN DIA... CADA MINUTO....

Aumenta la cantidad de sensores  
y la capacidad de los dispositivos  
que generan los datos



Aumenta la velocidad de  
transmisión (5G, 6G?)

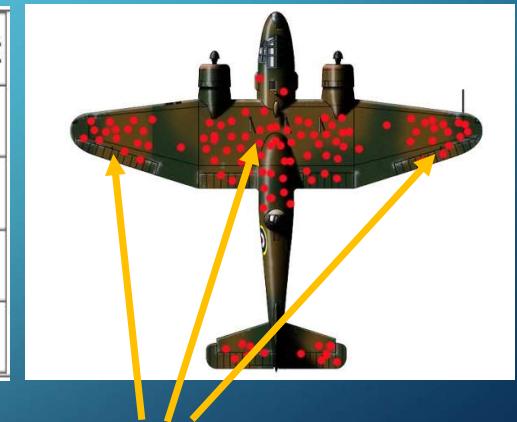
# MUCHOS DATOS + ANALISIS CORRECTO

Segunda Guerra Mundial - EEUU : Dónde mejorar el blindaje en los bombarderos?

DATOS: Los agujeros de bala de los aviones que regresaban no se distribuían uniformemente



Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8



Blindemos estas zonas !!!

Abraham Wald  
(matemático, teoría de la decisión):

**NOOOO, blindemos el resto !!!**

*Fuente: How Not to be Wrong - Jordan Ellenberg*



## ANALISIS EN BIG DATA

**Las Telefonicas saben tus posiciones, tus ingresos por como recargas tu celular... etc...**

**Los Servicios saben tus gustos, tus elecciones....**



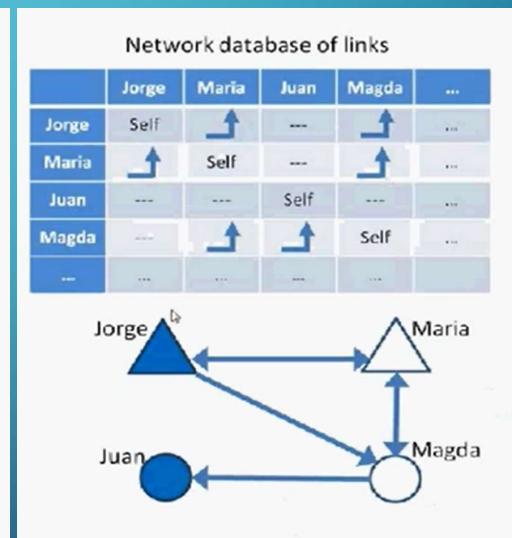
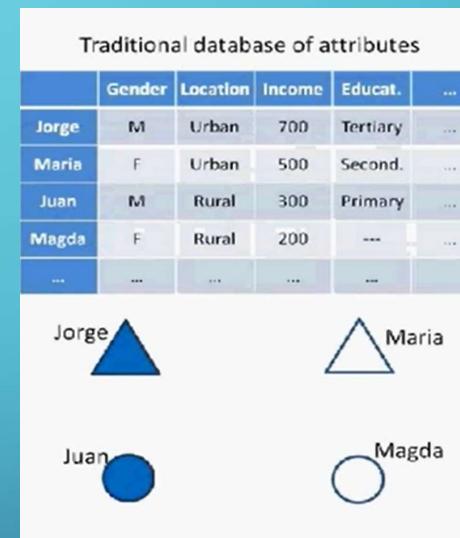
**Con 8 clasificaciones de películas ( 2 de las cuales pueden ser completamente equivocadas) y las fechas en que se hicieron (con un error de 14 días).... el 99% de los registros pueden ser identificados**

# ANALISIS EN BIG DATA

Miles de datos automaticos

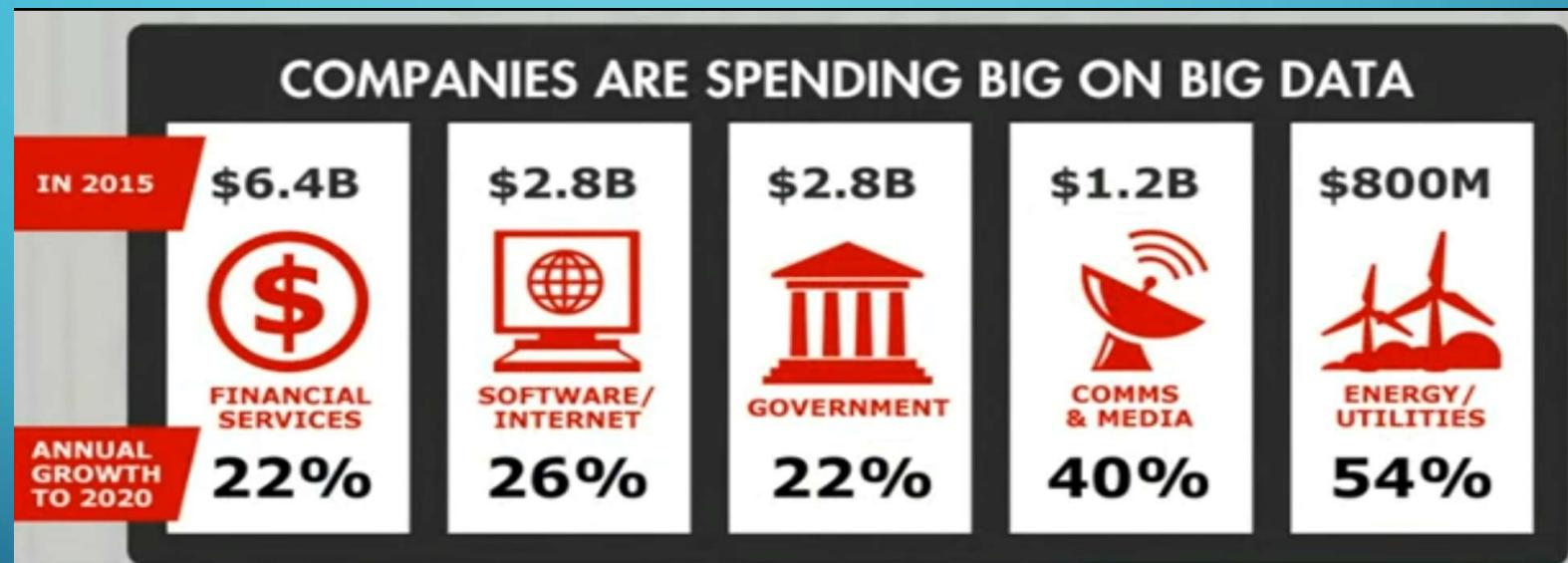
+

Datos importantes de redes  
interpersonales

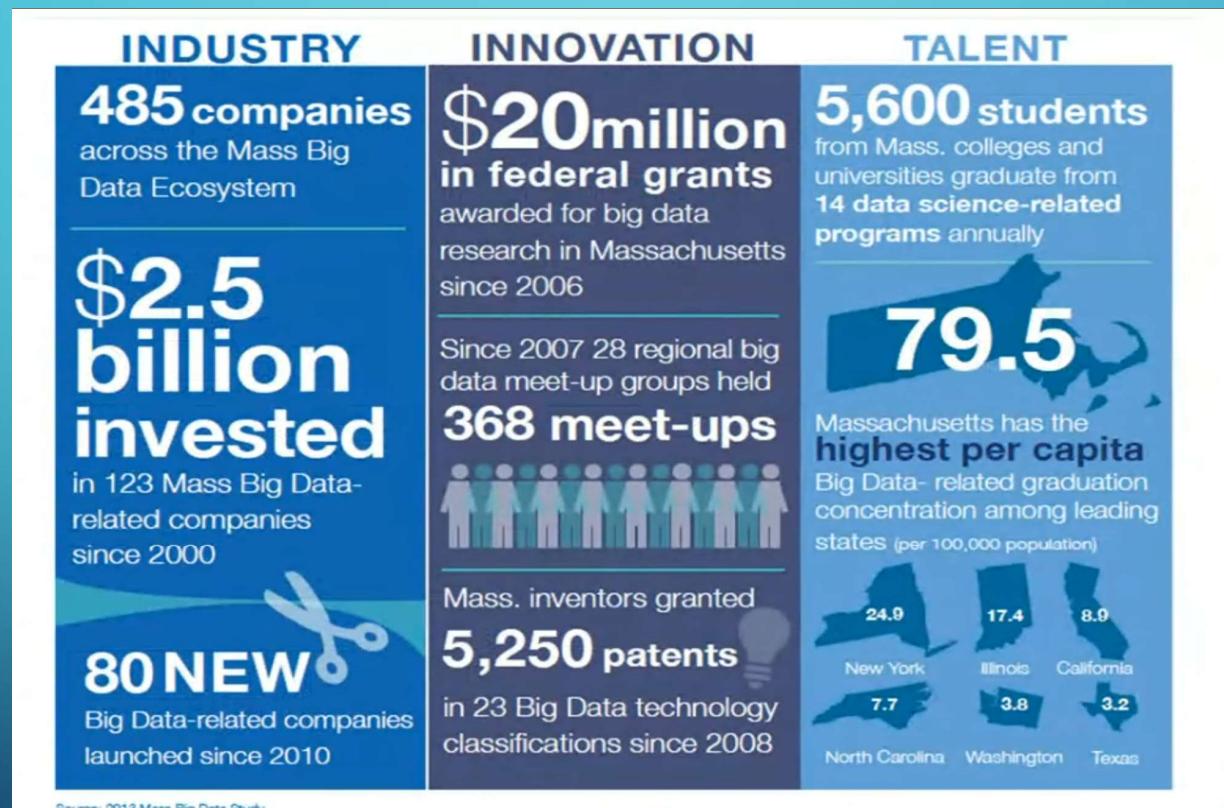


Predicciones mucho mas reales....

## CIENCIA DE DATOS.... DONDE ?



# CIENCIA DE DATOS.... DONDE ?



# CIENCIA DE DATOS.... DONDE ?



## APLICACIONES DE CIENCIA DE DATOS

DATOS..... o ..... PERSONAS ?

AMBOS!

Mejorar Bienestar

Web

Logs

IoT

RRSS



## HAGAMOS ALGUNOS EJERCICIOS...

Fuente de Datos	Volumen	Variedad	Velocidad	Viabilidad	Veracidad
Tweets					
Facebook					
Imágenes Satelitales					
Tomografías Computadas					
Registros Uber					
Rendimiento académico					
Listas de reproducción Spotify					
Compras en MercadoLibre					
Registros médicos de un hospital					

## HAGAMOS ALGUNOS EJERCICIOS...

Un candidato a Elecciones Presidenciales te contrata para hacer un estudio de tendencia, basado en Tweets. Para realizar un diseño inicial de esquema de trabajo:

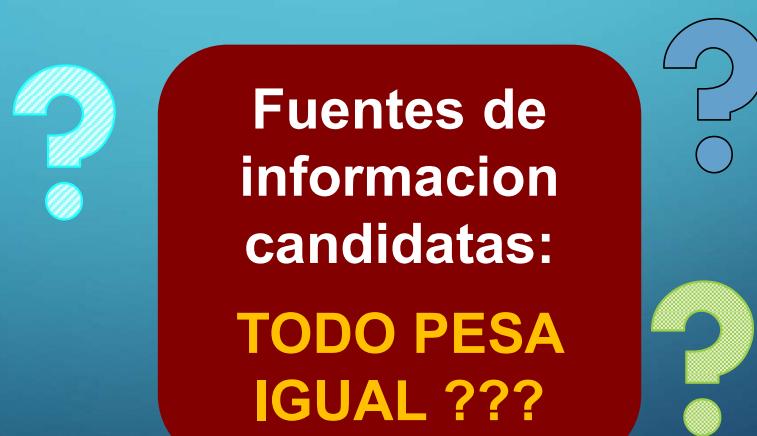
- Qué información tomarías de cada tweet?
- Qué búsqueda aplicarías sobre la información coleccionada?
- Con los resultados, qué tipo de informe estas pensando ofrecerle al candidato?





## HAGAMOS ALGUNOS EJERCICIOS...

Esquematiza el tipo de información necesaria, y cómo obtenerla, para diseñar un sistema de gestión de tráfico que minimice embotellamientos, disminuya el consumo de combustible y los costos de transporte.



## HAGAMOS ALGUNOS EJERCICIOS...

Esquematiza el tipo de información necesaria, y cómo obtenerla, para diseñar un sistema de gestión de tráfico que minimice embotellamientos, disminuya el consumo de combustible y los costos de transporte.

60% ?



30% ?



10% ?





HAGAMOS ALGUNOS EJERCICIOS...

Analista BI    ?    Científico de Datos

# CIENCIA DE DATOS.... DÓNDE ?

- ◆ Marketing
- ◆ Finanzas
- ◆ Salud
- ◆ Genetica
- ◆ Psicologia
- ◆ Astronomia
- ◆ Agricultura



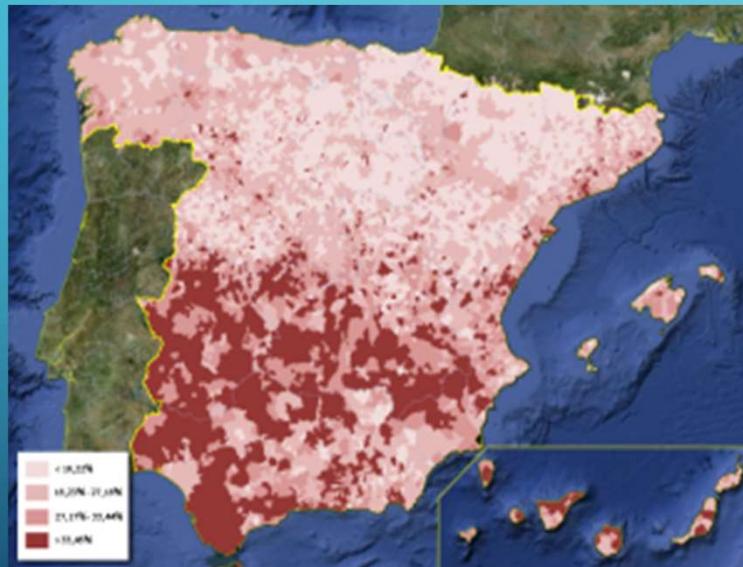
- ◆ Medio ambiente
- ◆ Ciberseguridad
- ◆ Educacion
- ◆ Deportes
- ◆ Politica
- ◆ Turismo
- ◆ ...

CIENCIA DE DATOS.... DÓNDE ?



# IMPORTANTE FUENTE: DATOS GEOGRAFICOS

Ejemplo: Geomarketing aplicado al lanzamiento de productos



¿QUIÉN?

Al aportar una clasificación de los tipos de hogares, facilita la definición del target.

¿DÓNDE?

Habits Big Data sitúa estas familias geográficamente sobre el mapa. Ofrece su precisa localización en el territorio deseado)...

¿CUÁNTOS?

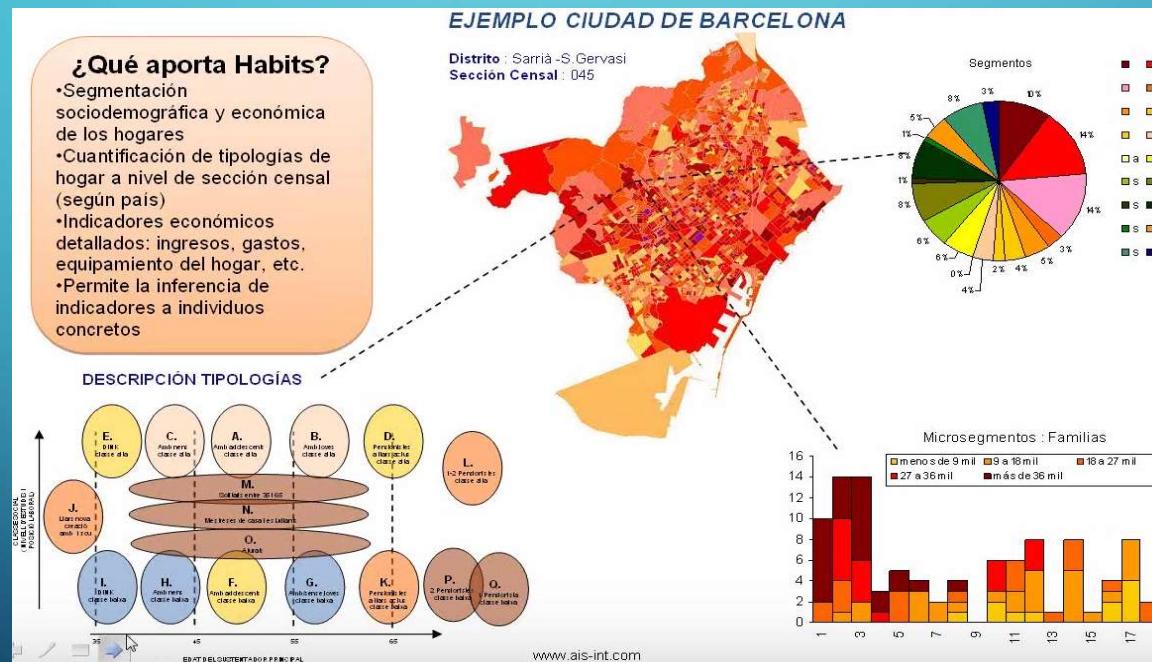
Habits Big Data no se limita a describir cuál es la tipología de familia dominante en un área geográfica determinada, sino que ofrece la densidad de cada una de las tipologías.

¿CÓMO?

Habits Big Data dispone de los datos sobre hábitos de consumo, niveles de renta y otros indicadores sociodemográficos, por lo que brinda información sobre cómo se comporta cada tipología de familia, cada target.

# IMPORTANTE FUENTE: DATOS GEOGRAFICOS

Ejemplo: Geomarketing aplicado al lanzamiento de productos





## DATOS GEOGRAFICOS

# Cómo almacenar y consultar Datos Geográficos ?

*Coming soon...*