# El fenómeno Big Data y los títulos en Estadística en España.

Daniel Peña

Rector

Universidad Carlos III de Madrid

V Conferencia Interuniversitaria sobre Titulaciones en Estadística
UCM, enero 2014

# Indice

1. Introducción
2. El fenómeno Big Data
3. Su efecto en las titulaciones de Estadística
4. Conclusiones

# Introducción

La Estadística actual fue creado por Pearson y Fisher para tratar con pequeñas muestras. Muchos departamentos de todo el mundo siguen enseñando  bajo esta influencia.

- Muestras pequeñas (n<200)
- Descripción univariante
- Inferencia: homogeneidad de los datos y utilización óptima de la información
- Modelos paramétricos simples y escuetos
- Estimación óptima (suficiencia, eficiencia) y contraste de hipótesis

# Introducción

- Contrastes de ajuste
- Datos categóricos, tablas de contingencia
- Modelos de regresión lineal
- Modelos multivariantes  lineales bajo hipótesis de normalidad
- Series temporales lineales univariantes

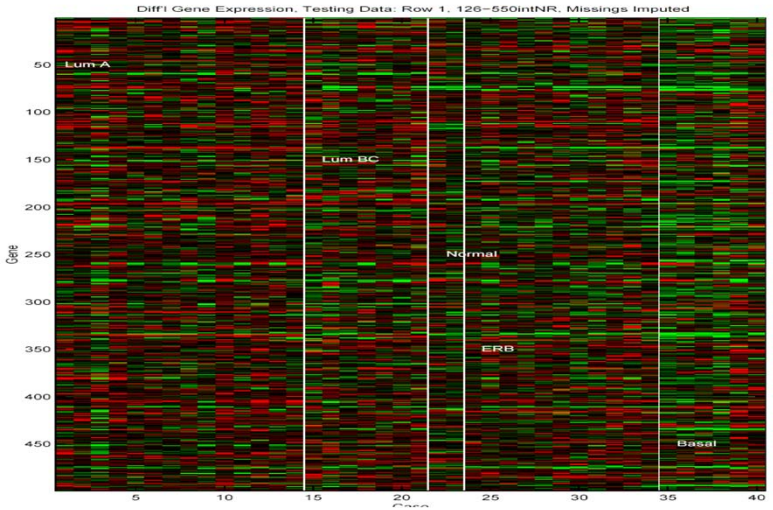
# Introducción

Las limitaciones de estos métodos para una muestra de 100,000 datos y 500 variables:

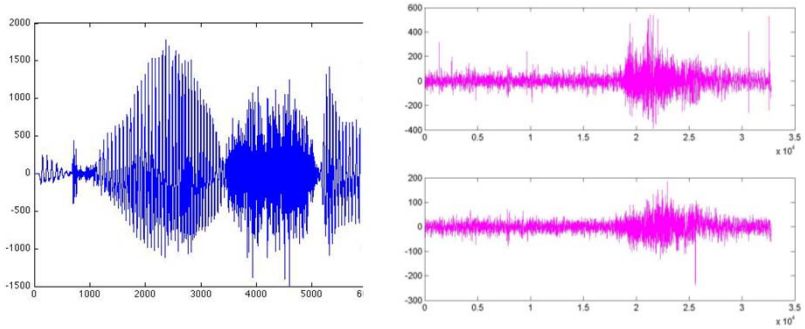
- Heterogeneidad: diferentes modelos en diferentes zonas del espacio
- Ajuste del modelo: Se rechaza cualquier contraste de ajuste o modelo paramétrico simple
- Eficiencia irrelevante y más importante robustez, grupos heterogeneidad; relaciones entre las variables
- Necesitamos nuevos métodos automáticos de selección y comparación de modelos

# Introducción

- Los datos pueden ser objetos y no solo mediciones (Digital information, matrices/images, surfaces in many dimensions, texts, social network messages,…)




- Se requieren modelos complejos y heterogéneos (Dynamic, Multivariate, Non parametric/semiparametric) para Big Data.



# El fenómeno Big Data

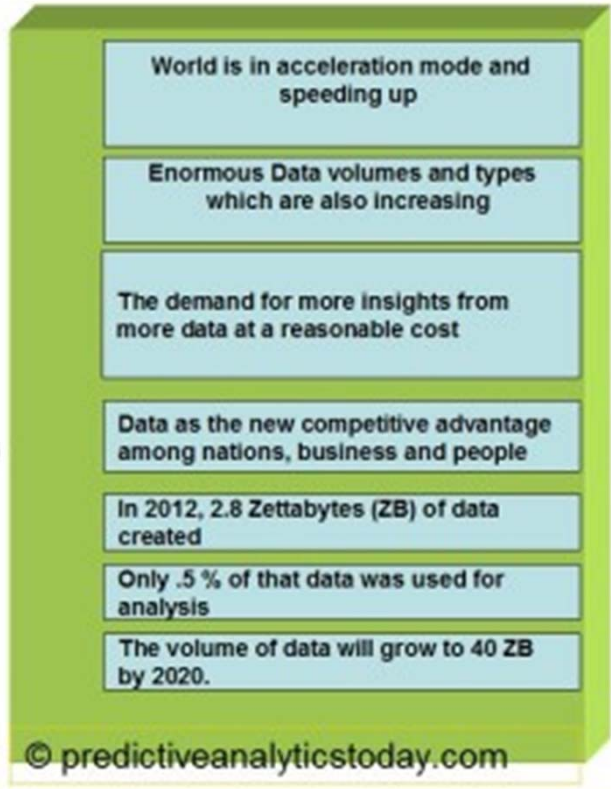
En pocos años el crecimiento de los bancos de datos es exponencial.

- Cualquier aparato digital genera **gratis** mucha información sobre su uso.(Medidores, aparatos TIC, redes, etc)
- Muestras de miles de datos y de variables comienzan a ser habituales

# El fenómeno Big Data

En Google:

Statistics: 3.600 M (Estadística 8,4M)
Big Data 2.530 M
Mathematics 278 M (Matemática,18,4 M)
Computer Science 1.400 M (informática 87 M)




No es un fenómeno pasajero

Su interés crece con gran velocidad

# El fenómeno Big Data

- 1. World Data Centre for Climate El WDCC (Centro Mundial de datos para el clima), base de datos más grande del mundo. Almacena unos 400 terabytes de información sobre el clima en todo el mundo.

- 2. National Energy Research Scientific Computing Center El NERSC investiga distintos tipos de energía. Su base de datos tiene 2.8 Petabytes.

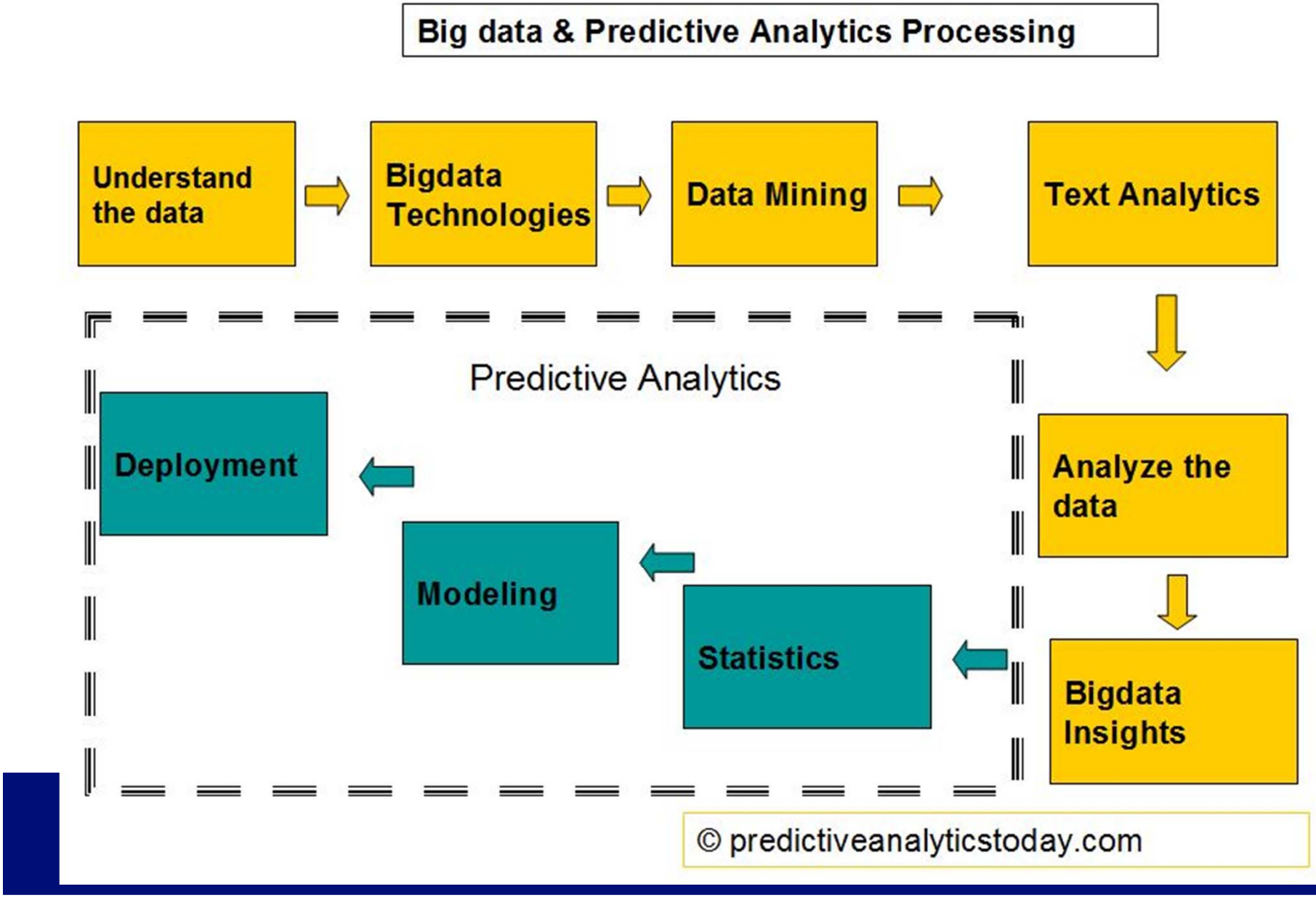
- 3. AT&T. compañía de telecomunicaciones. almacena 350 terabytes de información.

  4. Google Recibe más de 100 millones de consultas al día. Se supone que almacena cientos de terabytes de información.

**Terabyte** (TB)=$10^{12}$ Petabyte (PB)=$10^{15}$ **Exabyte** (EB)=$10^{18}$ **Zettabyte** (ZB)= $10^{21}$
•La colección impresa de la biblioteca del congreso de los EE.UU=10Terabytes

# El fenómeno Big Data

- No hay que transmitir para su uso estas grandes masas de datos, podremos operar con Cloud computing
- Necesitamos procedimientos automáticos de construir modelos y compararlos
- Se podrá analizar datos desde el teléfono en cualquier momento


Big data & Predictive Analytics Processing
Understand the data
Bigdata Technologies
Data Mining
Text Analytics
Predictive Analytics
Analyze the data
Bigdata Insights
Statistics
Modeling
Deployment
© predictiveanalyticstoday.com

# Las titulaciones de Estadística

- Reorientar la forma de enseñar la estadística: menos teoría univariante de modelos simples y más práctica con datos multivariantes y dinámicos.
- Tratar no solo con mediciones sino con imágenes, textos, categorías etc.
- La computación parte central de la Estadística
- Debemos colaborar con otros departamentos para el análisis de datos masivos: informática, ingeniería de telecomunicación.

# Las titulaciones de Estadística

- Nuevos Grados en Data Science, Data Analytics, Data Mining, Data Engineering…
- Enseñamos estadística one-shot cuando vamos a tener que combinar muchas herramientas estadísticas en un mismo problema
- Cómo encontrar relaciones y patrones,  clasificar, hacer grupos, etc será clave
- Cómo reducir la dimensión eficazmente
- Statistical Learning (Hastie Tibshirani) clave para el futuro.

# From Big Data to Big Statistics
# John Sall, SAS

Now that we have lots of data and can process it amazingly fast, we still need ways to look at it without being overwhelmed. We don't want to look at 10,000 graphs--we want one graph that shows the bright spots among 10,000 graphs. We need volcano plots and false-discovery-rate plots. We want the computer and software to do the work of finding what is most interesting and bringing it to our attention. We want our results sorted and summarized, but with access to the detail we need to understand it. Also, when we look at the most significant of thousands of statistical tests, we want to know if we are seeing random coincidence selected out of thousands, or if we are seeing real effects

# Conclusión

- Es una oportunidad para resituar la Estadística en el centro de la adquisición de conocimiento
- Debemos cooperar con los científicos que entienden los datos y con los que saben como transmitirlos y manipularlos eficazmente
- Se va a necesitar una gran cantidid de científicos para este tema.

# Conclusión

- La información de los métodos estadísticos un estudiante la encuentra gratis en internet y en los MOOCs.
- Tenemos que aportar experiencia, confianza, capacidad de intregración de conceptos…y para eso necesitamos profesores que hayan trabajo con datos masivos y hagan investigación sobre ese campo.
- Si no damos esta formación nosotros lo harán otros