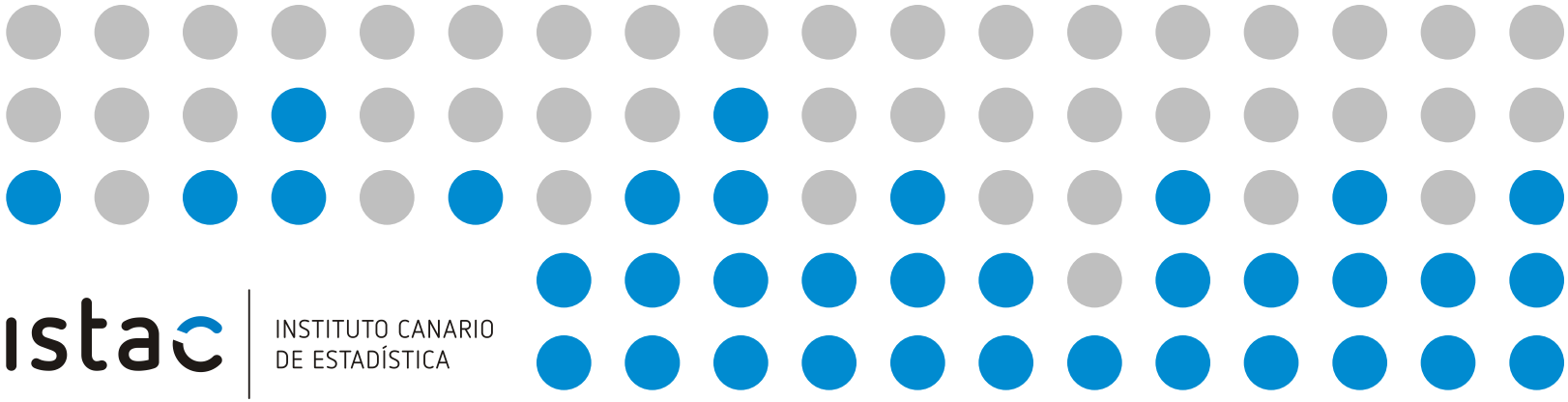
DIVULGACIÓN ESTADÍSTICA

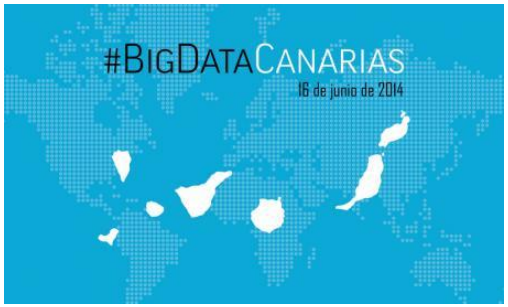
BIG DATA
Nuevos retos para la estadística pública



istac | INSTITUTO CANARIO DE ESTADÍSTICA

DIVULGACIÓN ESTADÍSTICA

BIG DATA
Nuevos retos para la estadística pública



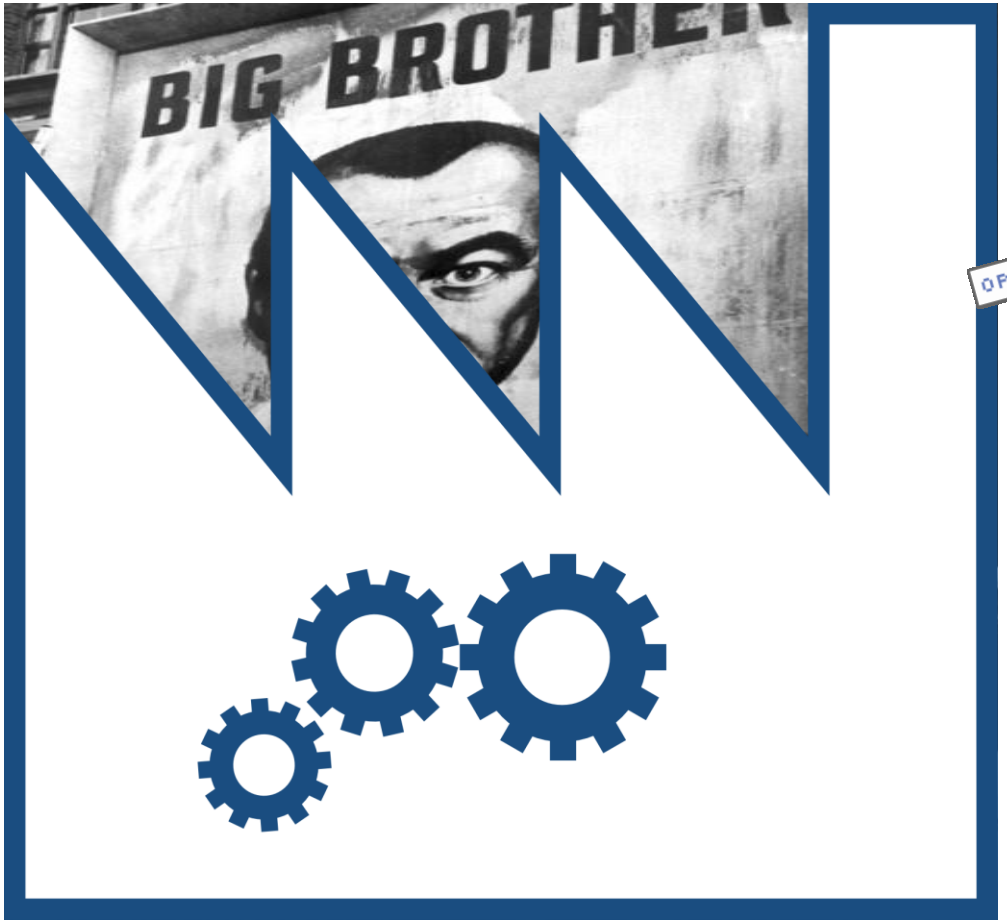
#BigDataCanarias
La Laguna (Tenerife) – 16 de junio de 2014

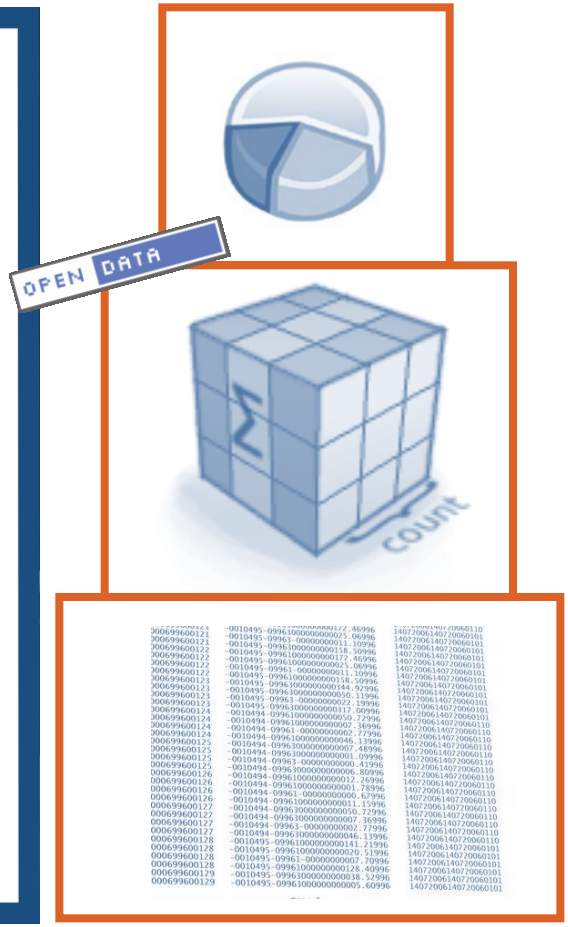
Universidad de La Laguna
Escuela Técnica Superior de Ingeniería Informática
Grupo Taro

Alberto González Yanes
Jefe de Servicio de Estadísticas Económicas
jgonyanp@gobiernodecanarias.org
@agonzalezyanes

istac | INSTITUTO CANARIO DE ESTADÍSTICA

¿Qué es una Oficina Central de Estadística?

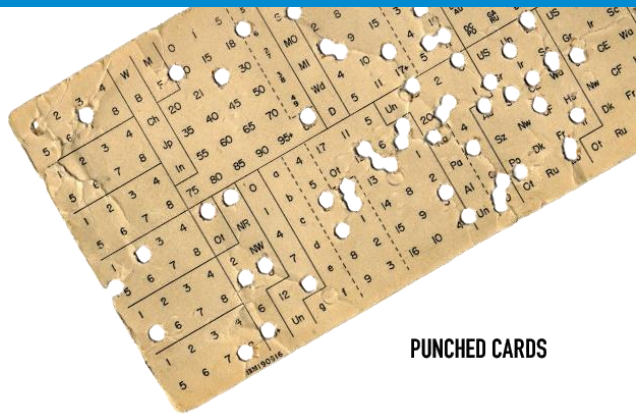




ENCUESTA	INDUSTRIALIZACIÓN	INDEPENDENCIA	MACRODATO
REGISTROS	INVESTIGACIÓN	INNOVACIÓN	MICRODATO

¿Qué hay de nuevo, viejo?

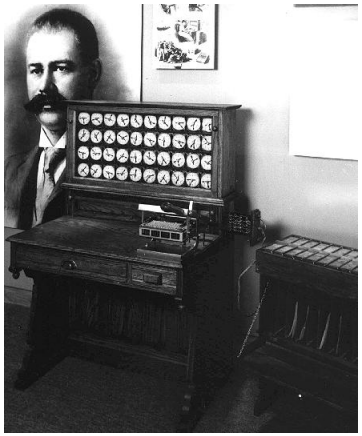
En 1880 comenzó a realizarse el censo en EEUU y debido a la cantidad de personas que lo formaba, tardó 8 años en terminarse. Incluso habían variables que no se llegaron a tabular. Por este motivo, el gobierno de los EEUU convocó un concurso para encontrar la mejor forma de realizar censos posteriores. En 1885 **Herman Hollerith** construye la máquina censadora o tabuladora, que por medio de **tarjetas perforadas** reducía el tiempo de realización del censo.



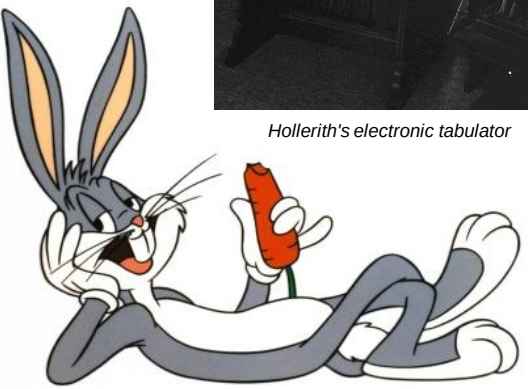
PUNCHED CARDS

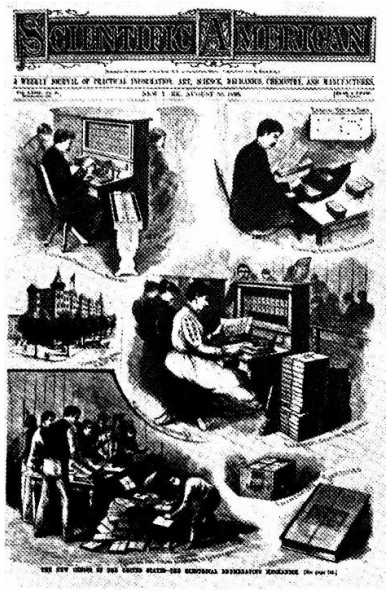
PRUEBA: Procesar los datos del censo 1880 de cuatro áreas en St Louis, MO. Tres candidatos:

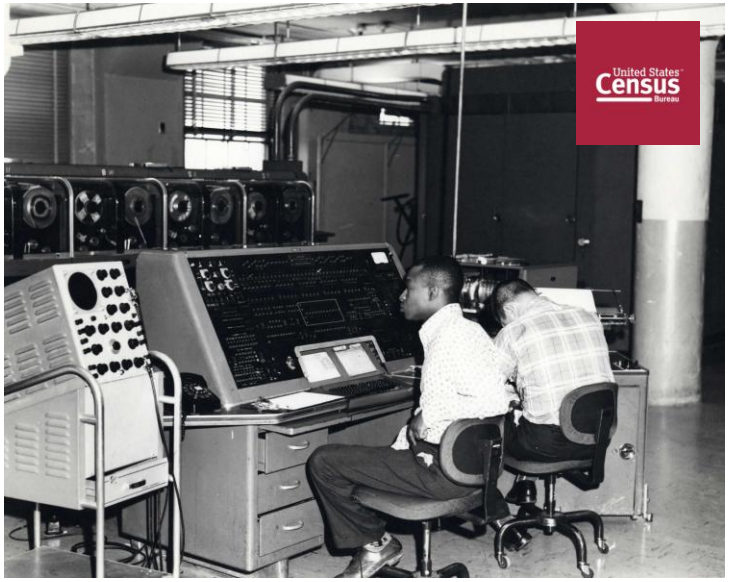
CAPTURA DE DATOS: 144,5 horas - 100,5 horas - **72,5 horas**.
PREPARAR DATOS PARA TABULACIÓN: 44,5 horas - 55,5 horas - **5,5 horas**



Hollerith's electronic tabulator



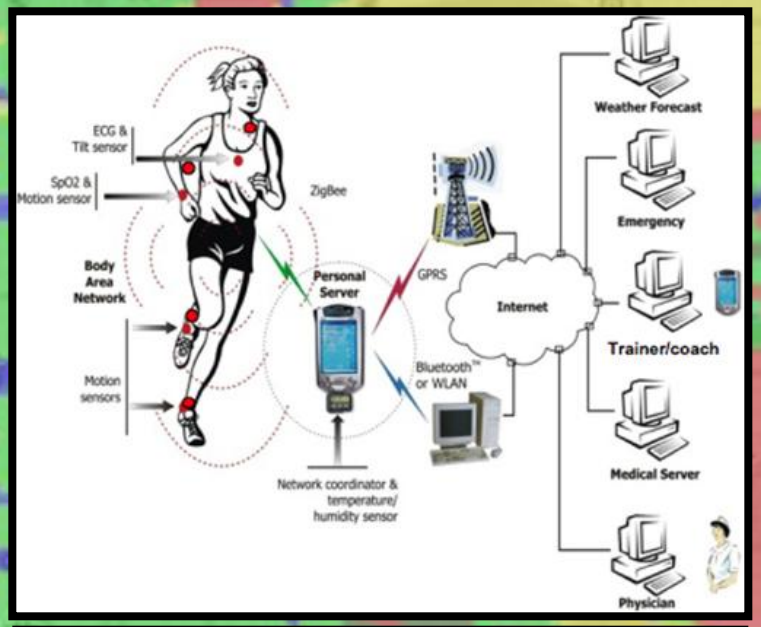




A UNIVAC computer at the Census Bureau, ca. 1960.


#SMART_CITIES
#INTERNET_OF_THINGS
#SOCIAL_DATA
DATIFICACIÓN
#DATA_SCIENTIST
#LINKED_DATA
#OPEN_DATA
#BIG_DATA
#DATA_DRIVEN_JOURNALISM
#DATA_VISUALIZATION

¿Qué hay de nuevo, viejo?



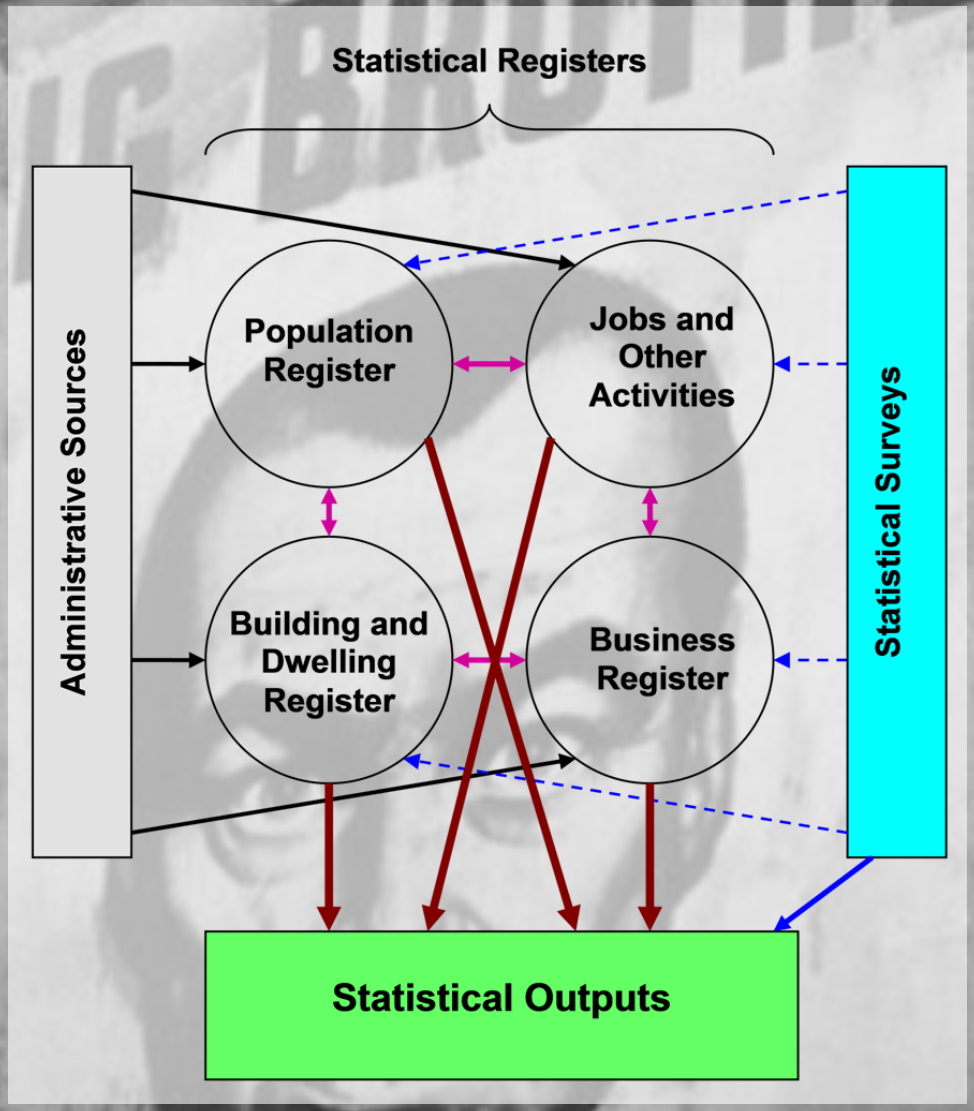
SENSORIZACIÓN

INTERNET

Automatización masiva de recogida de datos a ¿bajo coste?

Datificación ¿completa? de la actividad humana

¿Qué hay de nuevo, viejo?



¿Qué hay de nuevo, viejo?

¿Qué hay de nuevo, viejo?

Sample survey	Census	Register-based survey
Not included in register system	Included in register system – can be used for other register-based surveys	
Uses the register system to define populations and as a source for variables		
Sample design, estimation, measures of uncertainty	System-based thinking and coordination with other register-based surveys are important	
Own data collection – produce own questionnaries		Uses others- administrative registers
Editing – can contact respondents		Editing – can contact register-providing authority
Nonresponse – reminders, when to stop data collection?		Mismatch related to missing values or undercoverage
Quality flaws – sampling errors, measurement errors	Quality flaws - measure – ment errors	Quality flaws – relevance errors, lack of comparability
Small tables – cannot give estimates for small groups	Presentation – large tables with many cells	

Nuevas fuentes

- Uses others- administrative registers
- Editing – can contact register-providing authority
- Mismatch related to missing values or undercoverage
- Quality flaws – relevance errors, lack of comparability

¿Qué hay de nuevo, viejo?

$$ECM = b^2 + v^2$$

BIG ≠ ALL

BIG ≠ OWN

BIG ≠ EVERYWHERE

BIG ≠ FREE

BIG ≠ ALLWAYS

¿Qué hay de nuevo, viejo?

El pago a través de tarjetas de crédito o débito supone una parte de los pagos totales realizados en un comercio, dado que aproximadamente el 50% del gasto en comercios se realiza mediante dinero en efectivo.

Este porcentaje fluctúa, entre otros, en función de la categoría del comercio y su entorno, pero también por sesgos culturales inherentes a la nacionalidad del usuario.

En este informe ninguno de los resultados presentados es una extrapolación para deducir el gasto total llevado a cabo por los turistas extranjeros, las cifras reflejadas son en todo caso las recabadas por los medios de pago electrónico BBVA, y no deben tomarse como cifras absolutas de gasto realizado por cualquier medio de pago.

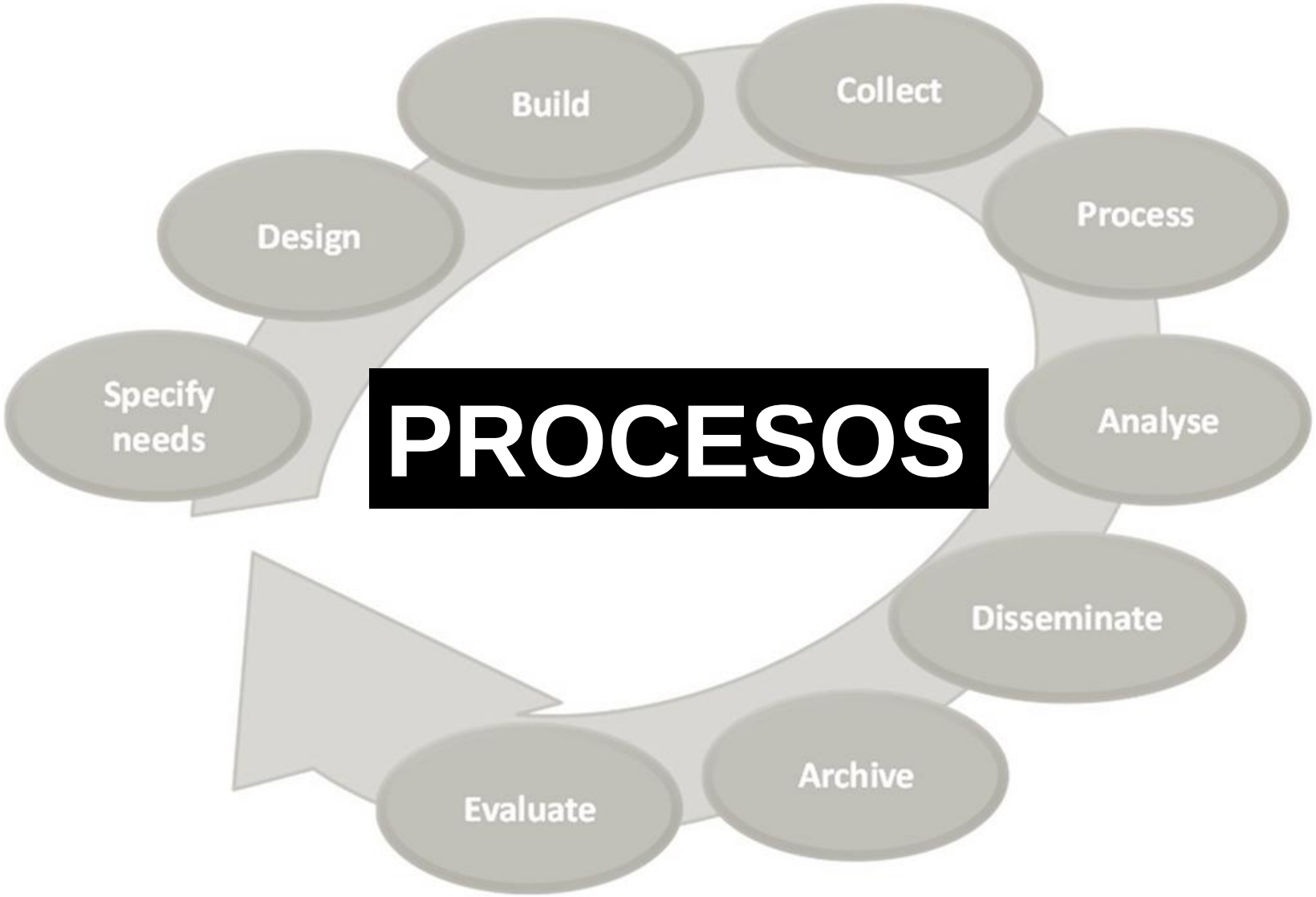
¿Qué hay de nuevo, viejo?

Como todos los datasets, éste también presenta ciertas limitaciones que conviene conocer. **La situación de los teléfonos no es totalmente precisa, ya que la que en realidad se tiene es la de la antena**. En entornos urbanos eso no es demasiado problemático puesto que la densidad de antenas es lo bastante alta como para ofrecer una precisión razonable; pero puede serlo en zonas rurales.

Otra limitación se puede producir a la hora de extrapolar datos totales a partir de la información que se obtiene. **Por poner un ejemplo concreto, no todos los teléfonos de los turistas rusos que visitan España se conectarán a la red de Telefónica**, lo que implica que si se quiere conocer el total de teléfonos rusos hay que realizar ciertas extrapolaciones que pueden introducir ciertos errores.

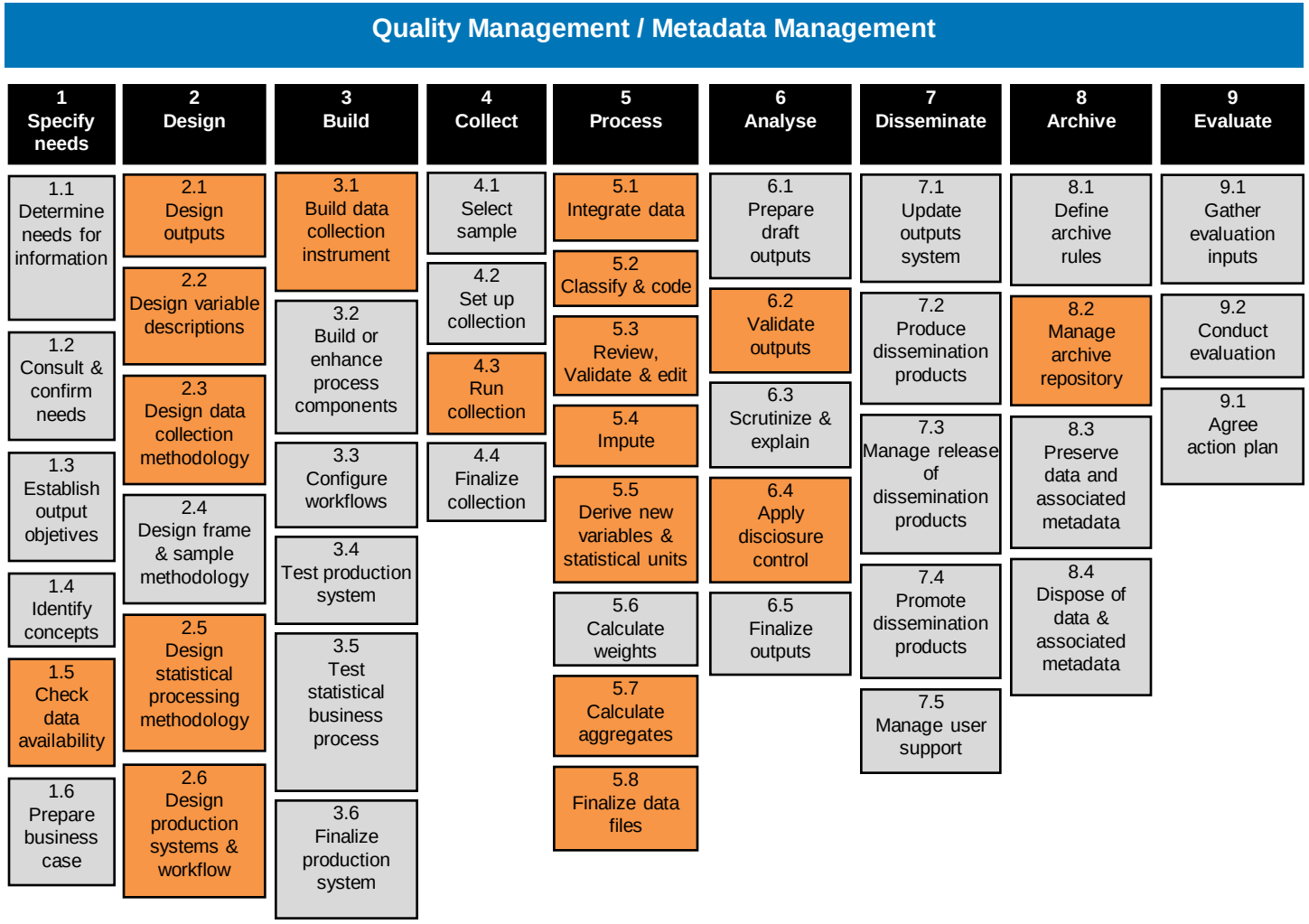
En este informe todos los datos que se presentan no están extrapolados, así que no deben tomarse como absolutos. Pero creemos que aun así pueden dar una idea bastante clara de situación.

¿Qué hay de nuevo, viejo?

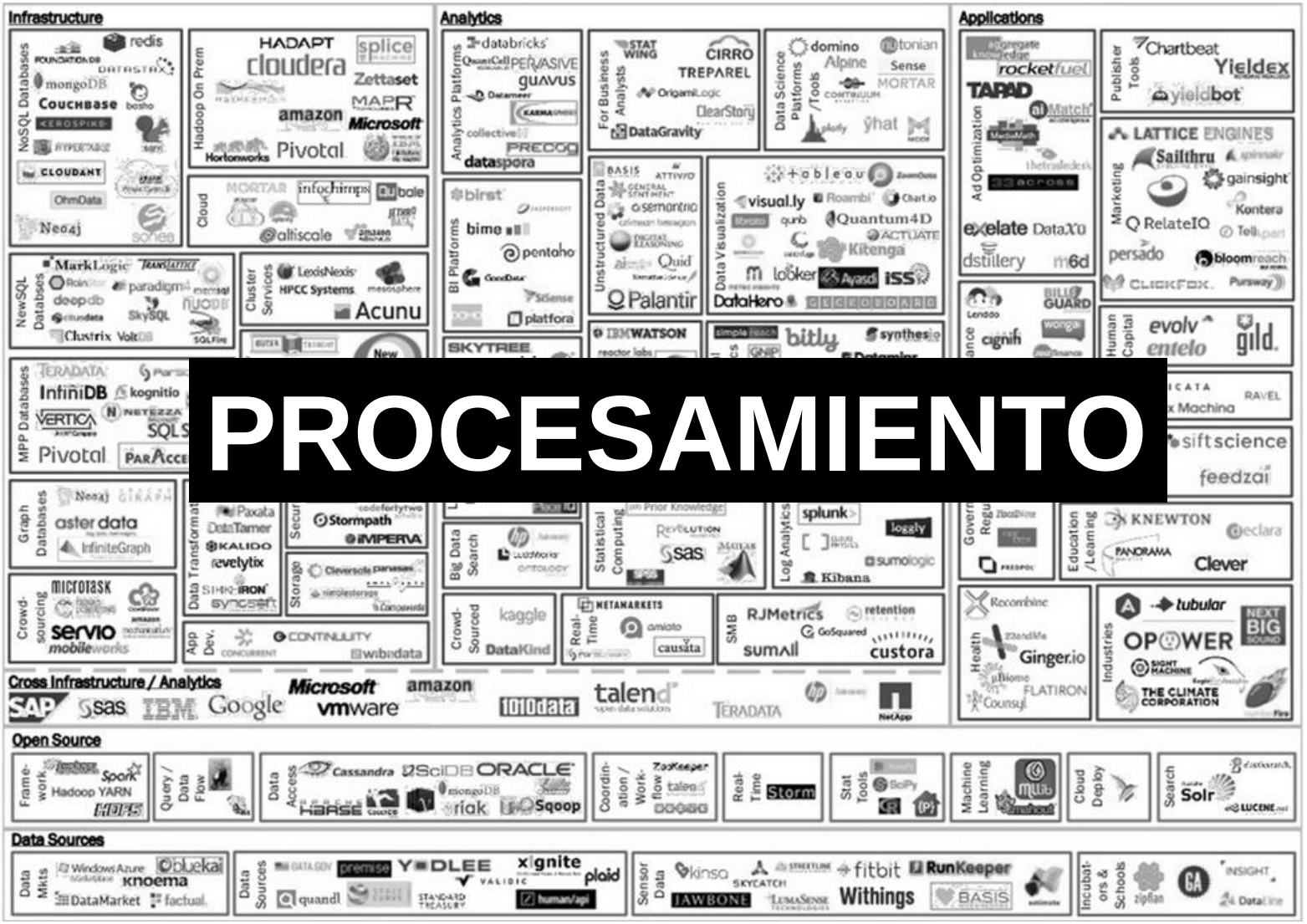


¿Qué hay de nuevo, viejo?

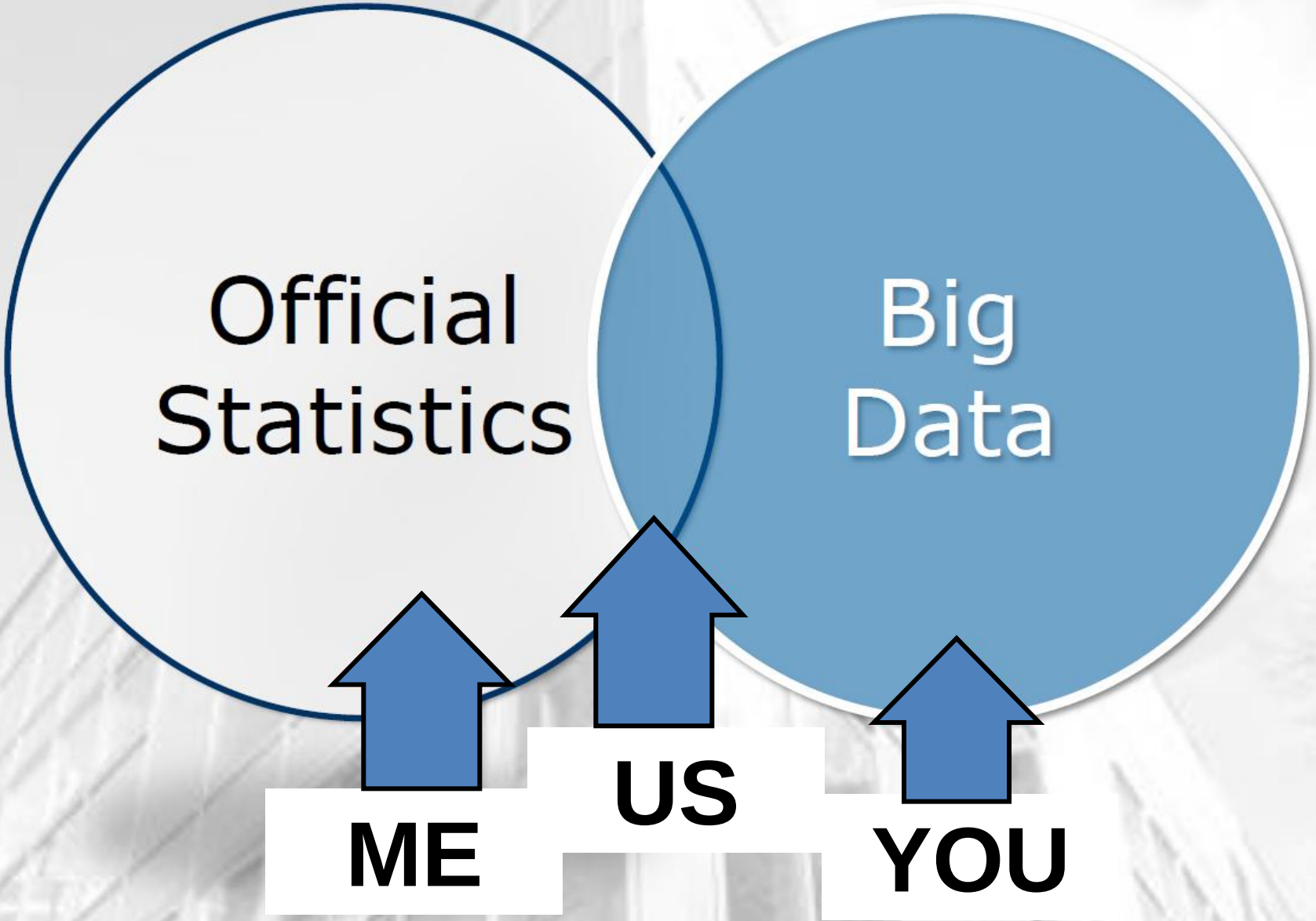
The Generic Statistical Business Process Model (GSBPM)



¿Qué hay de nuevo, viejo?



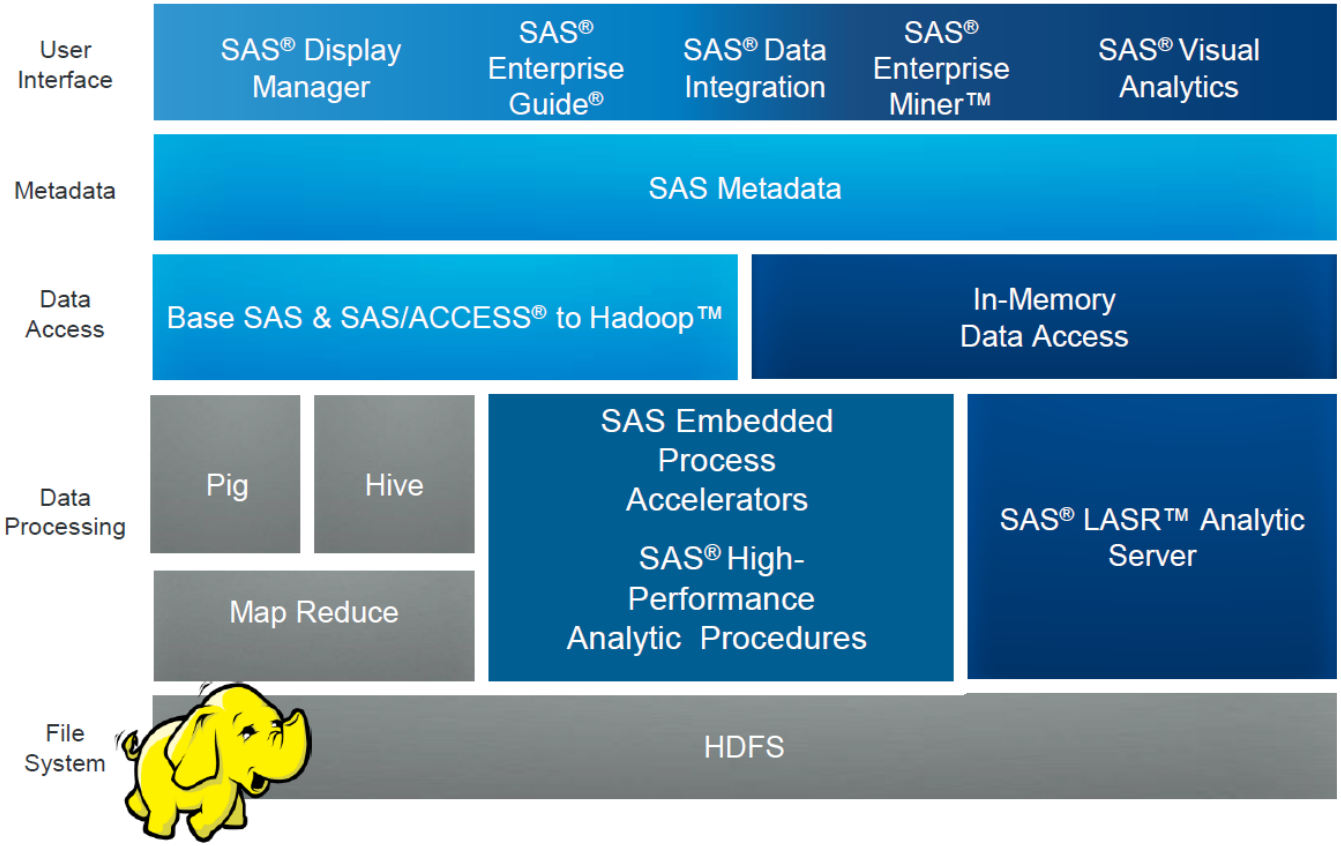
¿Qué hay de nuevo, viejo?



¿Qué hay de nuevo, viejo?

SAS & HADOOP | SAS® DENTRO DEL ECOSISTEMA HADOOP

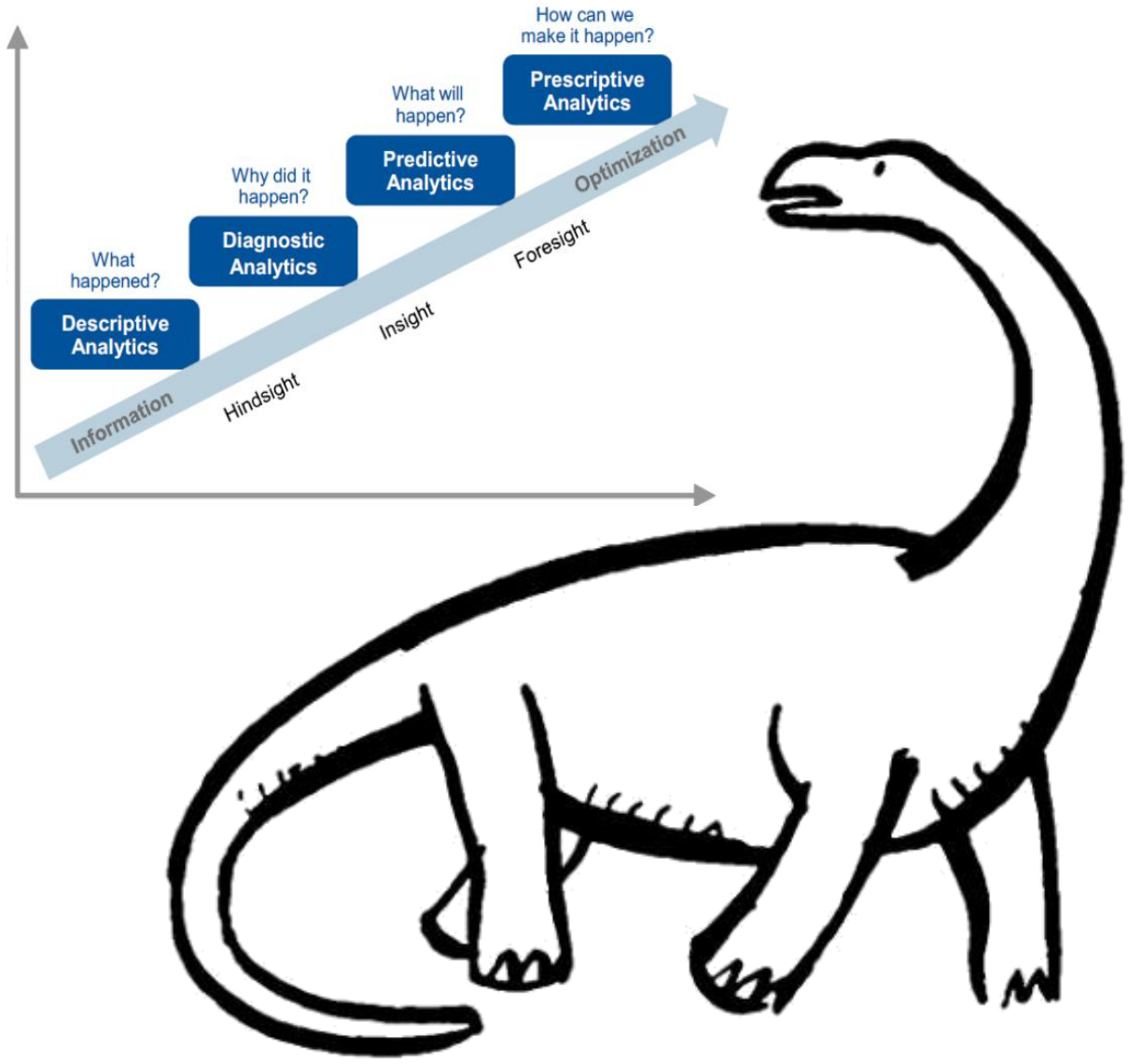
SAS® User



Next-Generation SAS® User

What happens when official statistics meets… BIG DATA

What happens when official statistics meets… BIG DATA



BIG

HEAVY

SLOW

Scheveningen Memorandum
Big Data and Official Statistics

EUROPEAN STATISTICAL SYSTEM

1. Reconocer que el Big Data representa **nuevas oportunidades y desafíos** para las estadísticas oficiales, y por lo tanto fomentar al Sistema Estadístico Europeo y sus socios a examinar efectivamente el potencial del BIG DATA en ese sentido. **> RECONOCIMIENTO**

2. Reconocer que Big Data es un fenómeno que está afectando a muchos ámbitos. Por tanto, es esencial desarrollar una **"Estrategia de estadísticas oficiales basadas en Big Data"** y examinar el lugar y las interdependencias de esta estrategia en el contexto más amplio de una estrategia global del gobierno a nivel nacional, así como a nivel de la UE. **> ESTRATEGIA**

3. Reconocer las implicaciones del Big Data en la **legislación de protección de datos y derechos de la persona** (por ejemplo, acceso a fuentes de datos en poder de terceros), implicaciones que deben ser abordadas apropiadamente como un asunto prioritario. **> LEGISLACIÓN**

4. Tener en cuenta que varios institutos nacionales de estadística están iniciando actualmente o considerando los diferentes usos del Big Data en un contexto nacional. Es necesario compartir las experiencias obtenidas en los proyectos Big Data concretos y colaborar dentro de la ESS y más allá, en un nivel global. **> COMPARTIR EXPERIENCIAS**

Scheveningen Memorandum
Big Data and Official Statistics

EUROPEAN STATISTICAL SYSTEM

5. Reconocer que el **desarrollo de las capacidades y habilidades** necesarias para explorar con eficacia los Big Data es esencial para su incorporación en el Sistema Estadístico Europeo. Esto requiere esfuerzos sistemáticos, como los cursos de formación adecuados y el establecimiento de comunidades dedicadas, incluyendo académicos, para el intercambio de experiencias y mejores prácticas. **> FORMACIÓN**

6. Reconocer que el **carácter multidisciplinar** del Big Data, lo que requiere sinergias y asociaciones entre los expertos y las partes interesadas de diversos dominios, incluyendo gobierno, universidades y titulares de las fuentes de datos privadas. **> COOPERACIÓN**

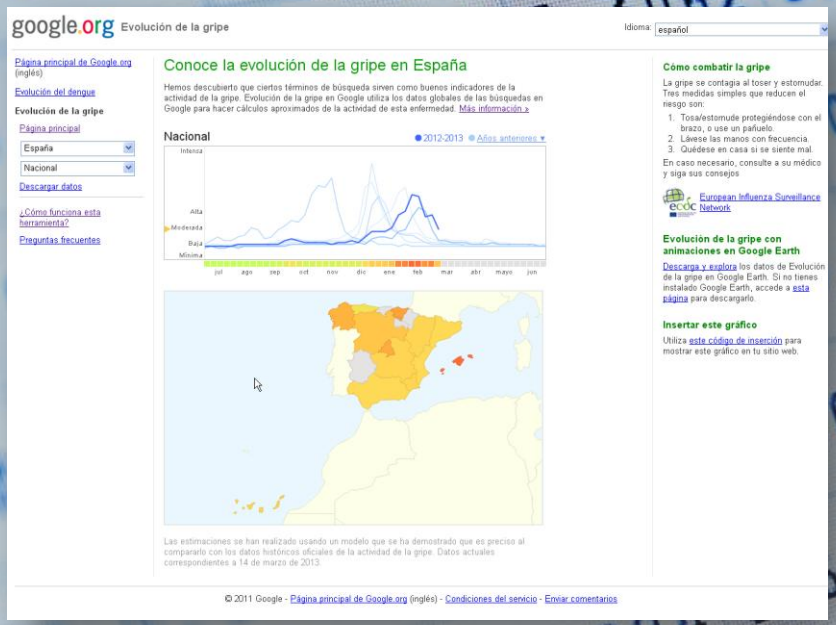
7. Reconocer que el uso de grandes volúmenes de datos en el contexto de las estadísticas oficiales requiere **nuevos desarrollos metodológicos**, de evaluación de la calidad y de abordaje de los problemas de TI relacionados. La Sistema Estadístico Europeo debería hacer un esfuerzo especial para apoyar esos desarrollos. **> INNOVACIÓN METODOLÓGICA**

8. Coinciden en la importancia de dar seguimiento a la implementación de este memorando, y por lo tanto se adopta un plan de acción y plan de trabajo del SEE. **> PLAN DE ACCIÓN**

What happens when official statistics meets… BIG DATA

PRIMARIA

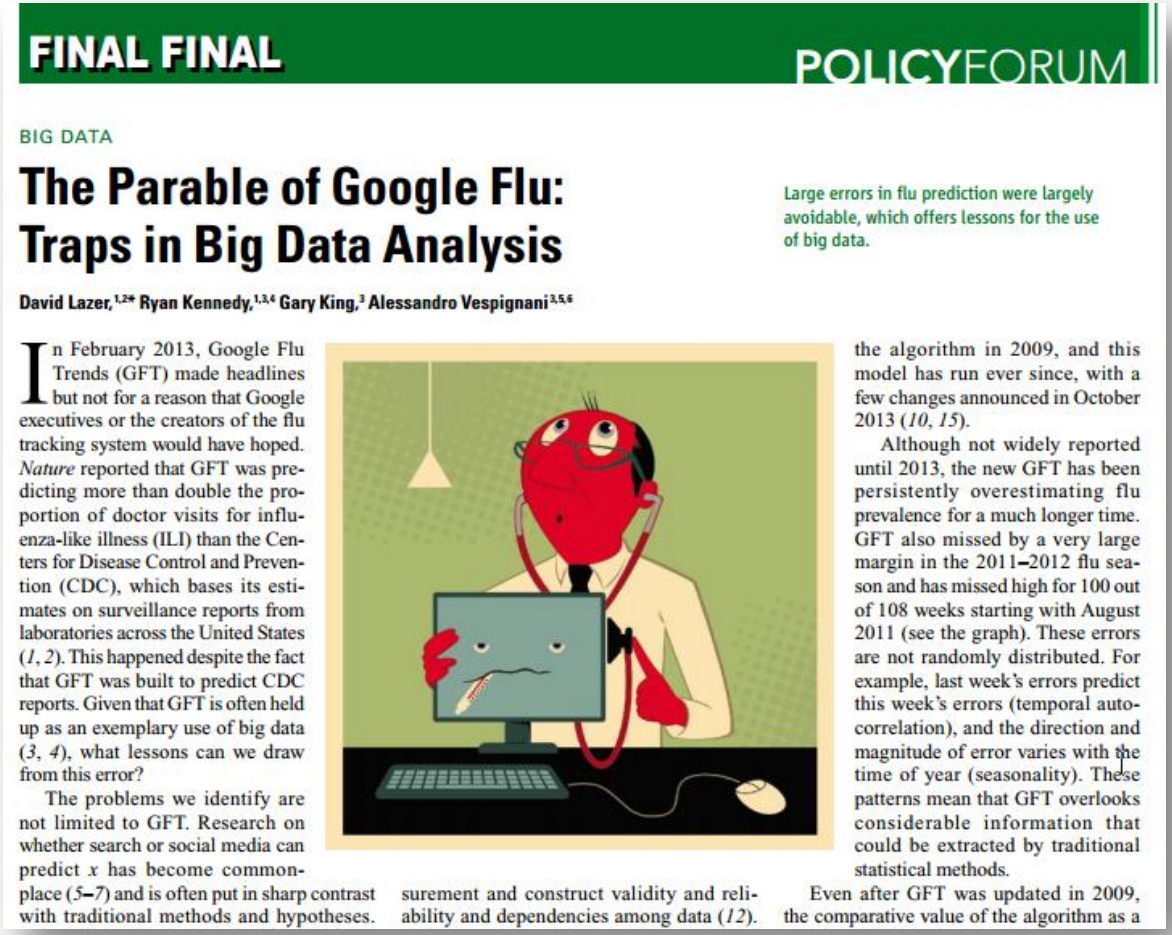
SECUNDARIA



Cifras contrastadas con la estadística pública

What happens when official statistics meets… BIG DATA

PRIMARIA



FINAL FINAL

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses.

surement and construct validity and reliability and dependencies among data (*12*).

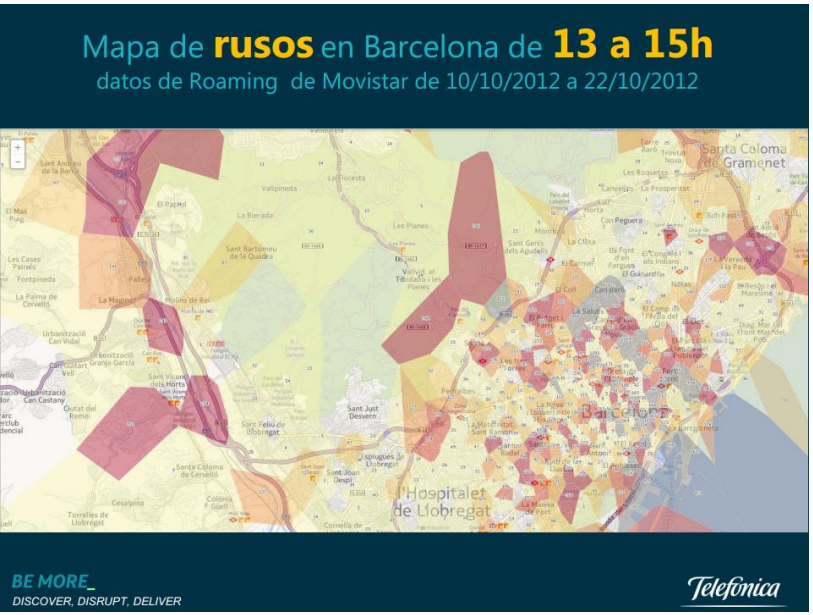
the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

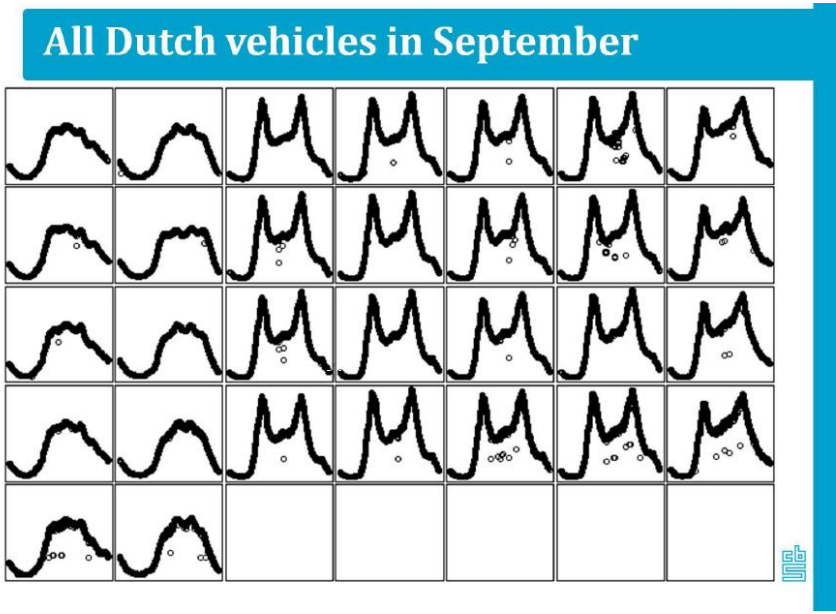
Even after GFT was updated in 2009, the comparative value of the algorithm as a

What happens when official statistics meets… BIG DATA

PRIMARIA

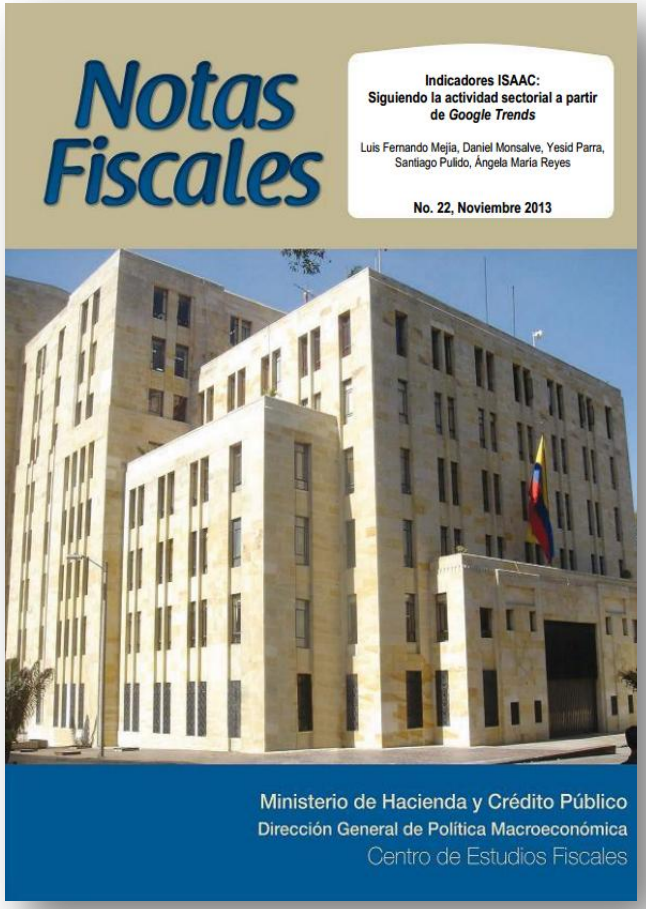


ajena



propia

SECUNDARIA



Statistics Netherlands
Division of Process development, IT and Methodology
Methodology sector

Heerlen
The Netherlands

Social Media Sentiment and Consumer Confidence

Piet J.H. Daas and Marco J.H. Puts

Table 1. Social media messages properties for various platforms and their correlation with consumer confidence

Social media platform	Number of social media messages[1]	Number of messages as percentage of total (%)	Correlation coefficient of monthly sentiment index and consumer confidence (r)[2]
All platforms combined	3,153,002,327	100	0.75
Facebook	334,854,088	10.6	0.81*
Twitter	2,526,481,479	80.1	0.68
Hyves	45,182,025	1.4	0.50
News sites	56,027,686	1.8	0.37
Blogs	48,600,987	1.5	0.25
Google+	644,039	0.02	-0.04
Linkedin	565,811	0.02	-0.23
Youtube	5,661,274	0.2	-0.37
Forums	134,98,938	4.3	-0.45

[1] period covered June 2010 untill November 2013
[2] confirmed by visual inspecting scatterplots and additional checks (see text)
*cointegrated

GRACIAS POR SU ATENCIÓN

Más información:
www.gobiernodecanarias.org/istac
www.slideshare/istac
@istac_es

istac | INSTITUTO CANARIO DE ESTADÍSTICA