

# Big Data, Minería y Aprendizaje: Conceptos y Herramientas para Economistas

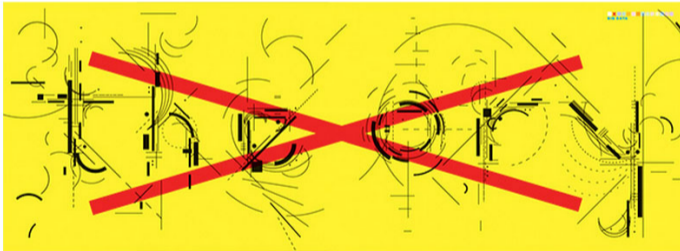
Walter Sosa-Escudero

Universidad de San Andrés y CONICET

# Big data: revolucion o cerveza artesanal?

SCIENCE : DISCOVERIES 

# The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08*Illustration: Marian Bantjes*

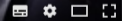
# 2014 Significance Lecture

## The Big Data Trap

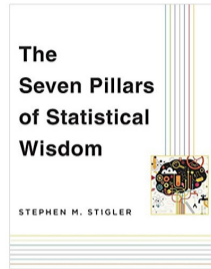
Tim Hartford, Economist, journalist and broadcaster  
Chair: Brian Tarran, Editor, *Significance*



▶ ⏪ 🔊 1:36 / 58:08



RSS 2014 Significance Lecture - The Big Data trap



found after an accident that he could remember absolutely everything. He could reconstruct every day in the smallest detail, and he could even later reconstruct the reconstruction, but he was incapable of understanding. Borges wrote, “To think is to forget details, generalize, make abstractions. In the teeming world of Funes there were only details.”<sup>2</sup> Aggregation can yield great gains above the individual components. Funes was big data without Statistics.

When was the arithmetic mean first used to summarize a data set, and when was this practice widely adopted? These are two very different

# Big data: ¿Otra vez arroz?

## DEBATE

2 opiná

142 shares



11



131



**Walter Sosa  
Escudero**  
Profesor  
Asociado, Udesa

Creo conservar, en algún recóndito lugar de mi casa, mi paleta de paddle de cuando en los noventa pensaba que el juego del presente se transformaría en el deporte del futuro. También disfruto de los vinilos de mi adolescencia que escucho casi a diario. Y por alguna razón exótica guardo celosamente una caja de diskettes de mis comienzos con la computación personal, allá en los ochenta.

# Tres casos

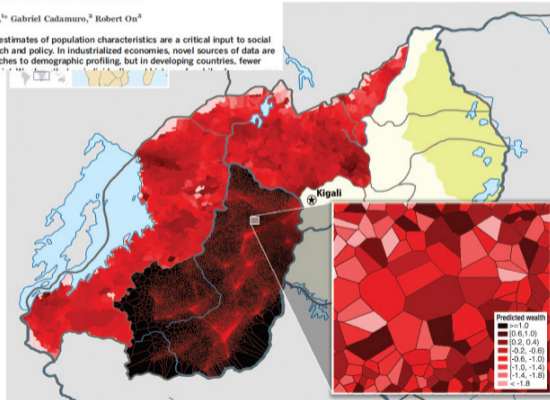
# Pobreza en Rwanda (predecir)

ECONOMICS

## Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,<sup>1a</sup> Gabriel Cadamuro,<sup>2</sup> Robert On<sup>3</sup>

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer





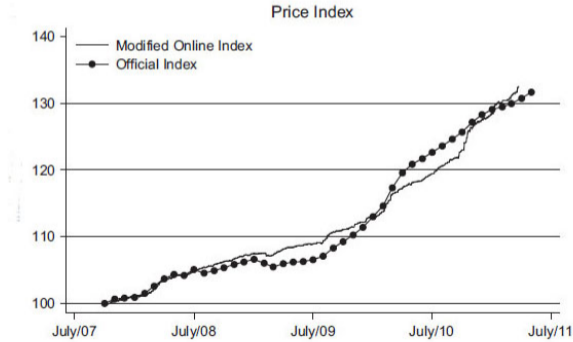
# Precios en Argentina (medir)



## Online and official price indexes: Measuring Argentina's inflation

Alberto Cavallo\*

Massachusetts Institute of Technology, Sloan School of Management, 75 Massachusetts Ave 55D-512, Cambridge, MA 02139, USA



## **Sales Taxes and Internet Commerce**

Liran Einav

Dan Knoepfle

Jonathan Levin

Neel Sundaresan

AMERICAN ECONOMIC REVIEW

VOL. 104, NO. 1, JANUARY 2014

(pp. 1-26)

- Datos: no encuesta ni experimento.
- De interactuar espontaneamente con 'entidades' interconectadas.
- Masivos
- Desestructurados

# Big data

## Small data (estadística clásica)


- Extraer lo máximo de **pocos** datos
- Solución: **estructura** (muestreo, modelo)
- Enfoque: muestreo complejo aproxima muestreo al azar (**lento** y **caro**, pero bueno). Teoría, experimentos.

## Big data

- **Muchos** datos (Volumen)
- Muchos datos **no estructurados** (Variedad)
- Muchos datos no estructurados e **inmediatos** (Velocidad)
- 'Condicionales baratos'.

## Big Data

Fenomeno de datos masivos, observacionales, no estructurados, producto de interactuar con objetos (fisicos o virtuales) interconectados



Estadística, econometría,  
machine/statistical learning,  
inteligencia artificial

$$Y = f(X) + u$$

- Interes en  $f(\cdot)$ . Efecto causal
- Modelo: Teoria, experimento.
- Probabilidades (error estandar, tests)
- Bueno?: insesgado, varianza minima, inferencia valida.



$$Y = f(X) + u$$

- Interes en  $Y$ : predecir, clasificar, medir.
- Modelo: modelo?. Lo **aprendemos**.
- Prediccion puntual (no inferencia).
- Bueno?: Performance predictiva fuera de la muestra.

- Etiqueta estadística: ex-ante. Teoría, identificación 'limpia' (consistencia). Inferencia robusta.
- Machine learning: ex-post, iterativa. Cross validation.
- Machine learning **construye** el modelo más que lo estima, en base a la performance predictiva **fuera de la muestra**. Adios al  $R^2$  (y a MCO? Mmm...).

# Jerga, desafíos y oportunidades

- Prediccion fuera de la muestra
- Muestra de entrenamiento y de evaluacion
- Aprendizaje
- Aprendizaje supervisado y no supervisado
- Regresion y clasificacion

- Dependencias (realmente tenemos big data?. Trump effect)
- Choice based sampling.
- Contracticos (podemos tener *todos* los datos?).
- Falacia de la correlacion.
- Transparencia / privacidad.
- Comunicabilidad. Caja negra (deep learning, forests, etc.)
- Consenso social/politico.

- New mas que big.
- Complejidad, heterogeneidad. No linealidades. Maldicion de la dimensionalidad.
- Rapido (crucial para la politica). Google Flu Trends. Price scrapping.
- Oportunidad para diseño de experimentos.
- Complementa a las estadisticas oficiales (no reemplaza).
- Cobertura. Rural vs. urbano, etc..

## Lecturas

- Hastie, Tibshirani, Friedman (2009)
- James, Witten, Hastie and Tibshirani (2014).
- Murphy (2012, Machine Learning)
- Varian (2014)
- Edición especial de JEP sobre Big Data (JEP, 2014)
- Papers: Keely and Tan (2008, Journal of Public Economics), Bajari et al. (2015, American Economic Review), Cavallo and Rigobon (2013, Journal of Monetary Economics).
- Mayer-Schonberger y Cukier (Big Data, 2013).

- Charla de Tim Harford sobre 'The Big Data Trap'.
- Nota en Clarin (6/4/2014)
- Computer intensive
- Olvidense de Stata



# big data

walter sosa escudero



breve manual para conocer la ciencia de datos  
que ya invadió nuestras vidas

XXI siglo veintiuno  
ediciones

ciencia que ladra...  
será muy útil

'... su antepasado no creía en un tiempo uniforme, absoluto. Creía en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. ... No existimos en la mayoría de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En este, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravesar el jardín, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma.'

*El jardín de senderos que se bifurcan*

'Ireneo tenía diecinueve años; había nacido en 1868; me pareció monumental como el bronce, más antiguo que Egipto, anterior a las profecías y a las pirámides. Pense que cada una de mis palabras (que cada uno de mis gestos) perduraria en su implacable memoria; me entorpeció el temor de multiplicar ademanes inútiles.'

*Funes, el memorioso*