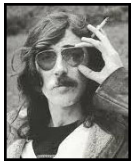
Big Data: Desafíos y Perspectivas para las Ciencias Sociales

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

Preludio



'Todo es GMM. Si sabes GMM, sabes todo'

(alguien, circa 1993)

El fin del mundo

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

SCIENCE : DISCOVERIES

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

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

'Basta de una vez con la teoría del comportamiento humano, desde la linguística a la sociología. Olvídense de la taxonomía, la ontología y la psicología. ¿Quién sabe por qué la gente hace lo que hace? El punto es que lo hacen, y que podemos rastrearlo y medirlo con una precision antes impensable. Con suficientes datos los numeros hablan por si mismos.'

Chris Anderson, Wired, 23/6/2008

'Quizás llegue el dia en que se pueda reunir una cantidad suficiente de datos para responder todas las preguntas sociales que nos preocupan. Lo dudo. Siempre habrán contiendas digitales, lugares oscuros y sesgos en cómo se usa y produce la información y la tecnología.'

Mark Graham, The Guardian, 9/3/2012

Clarín | Noticias | Deportes | ¡ExtraShow! | Estilo | Ser

iEco | Mercados • Indicadores • Empresas y Negocios | Economía

Clarin.com › iEco › Economía › 06/04/14

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.

Jerga

NoSQL, Hadoop, mineria, aprendizaje, visualización, fat models, funciones de pérdida, riesgo de Bayes, LASSO, CART, GARROTE, sobreajuste, muestra de entrenamiento, aprendizaje supervisado y no supervisado, validación cruzada, árboles, bosques, podas, nodos, clusters, boosting, bagging, matriz de confusion, curva ROC, regularización, shrinkage, Bayes, incertidumbre de modelo, promedio de modelos, error reproducible, prediccion fuera de muestra, bases, splines, GAM, support vector machines, subset selection, scrapping, networks, phyton, predictive analytics.

Demasiado Spanglish...

Problemas

- Decidir si un email es spam.
- Identificar codigos postales escritos a mano.
- Relevar precios.
- Asignar tratamiento para pacientes con cancer.
- Identificar factores que puedan identificar a los pobres.
- Encontrar el nombre de una canción a partir de un silbido o tarareo.
- Traducir un texto.
- Predecir preferencias por la redistribucion ante una eleccion.
- Encontrar la clase media.
- Recomendar libros o peliculas.

Babel

- Aprender, hurgar, encontrar patrones, clasificar, reducir dimensionalidad, visualizar, resumir, decidir.
- Manipular, mover, organizar, guardar, recuperar, explorar datos masivos.
- Data mining, statistical learning, machine learning, predictive analytics.

Más Spanglish...

Experiencias

- Epidemia de gripe A
- Iphones lentos
- Google translate
- El desafio de Netflix
- Mil millones de precios

El paradigma predictivo

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

- Objetivo: predicir Y en base a X, sin observar u y sin saber $f(.)$ ('aprender' f).
- Datos de entrenamiento: usados para construir / estimar un modelo predictivo.
- Datos de test /evaluación: usados para validar y evaluar.
- Objetivo: predecir bien en los datos de *test*.

Dinamica

- Decidir qué puede ser f.
- Aprender (estimar) f. Volver al paso inicial. Iterar.
- Decidir la dimensión de X.

Cuidado (para econometristas): el objetivo es predicion fuera de la muestra!.

Exito

Error cuadratico medio:

$$E\big[(Y - \hat{f}(X)^2)\big]$$

Validacion cruzada

1. Dejar 1/5 afuera. Estimar. Computar el ECM para los datos que quedaron afuera.
2. Repetir con otro 1/5
3. Computar el ECM total.

Elegir modelo con minimo ECM

El 'trade off' sesgo varianza

$$\text{ECM} = \text{Sesgo}^2 + \text{Varianza}$$

- Cuanto mas complejo es el modelo, menor el sesgo y mayor la varianza.
- ¿Es realmente relevante que la estimacion sea insesgada? Enorme diferencia con el enfoque clasico.
- El rifle de los economistas. Espacio para la estimación sesgada.

El paradigma clásico / frecuentista

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

- Objetivo: conocer $f(\ )$ o sus caracteristicas (derivadas, por ejemplo).
- Exito: insesgadez, varianza minima, consistencia, etc.
- ¿f? Viene de una estructura (teoria) o representa un experimento (¿teoria?).
- Paradigma de estimacion insesgada / consistente y varianza minima.
- Validez interna (no hay test data).

Lo bueno

- Rol de la induccion y la descripción en las ciencias sociales (y en la ciencia).
- ¿Big data o new data?
- Experimentos
- Construccion de contrafactuales.
- Complejidad, alta frecuencia.
- Re-evaluar la 'tirania de la insesgadez'.
- ¡Bayes!

Lo malo

- ¿Small data? ¿Realmente hay un problema de pocos datos en las ciencias sociales?
- Mas datos no es necesariamente mejor.
- Choice based sampling. Papelones historicos. El origen de Gallup.
- Small data es un problema poblacional: no observabilidad de contrafactuales.
- Polítcas: ¿prohibir paraguas? ¿regalar televisores?
- Daniel Heymann y el PBI diario. Datos de la frecuencia correcta. Raices unitarias.

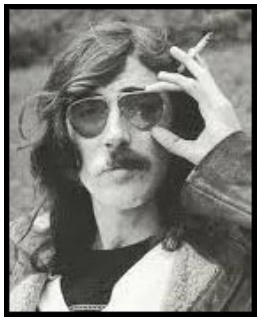
Lo feo

¿El fin de la teoria? Intentos 'ateoricos' en economia.

- VAR y econometria dinamica.
- La 'revolucion de credibilidad': experimentos, instrumentos. Mostly Harmless.
- ¿Big data? Mejoras sustanciales con la teoria (linguistica basada en viejas traducciones, meteorologia). Acciones de politica. Analisis causal.

Teoria y conocimiento causal como necesidad atavica.

Perspectivas



'... yo ya soy parte del mar'

- Posturas extremas sobre el tema.
- Discusion filosofica produnda sobre el rol de la teoria y la induccion.
- ¿Cambios en la enseñanza?

Caja de herramientas

Lecturas

- Hastie, Tibshirani,Friedman (2009)
- James, Witten, Hastie and Tibshirani (2014).
- Murphy (2012, Machine Learning)
- Varian (2014)
- Edicion especial de JEP sobre Big Data (JEP, 2014)
- Papers: Keely and Tan (2008, Journal of Public Econommics), 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
- Curso (learning): Hastie and Tibshirani (Stanford)
- Libros gratis!

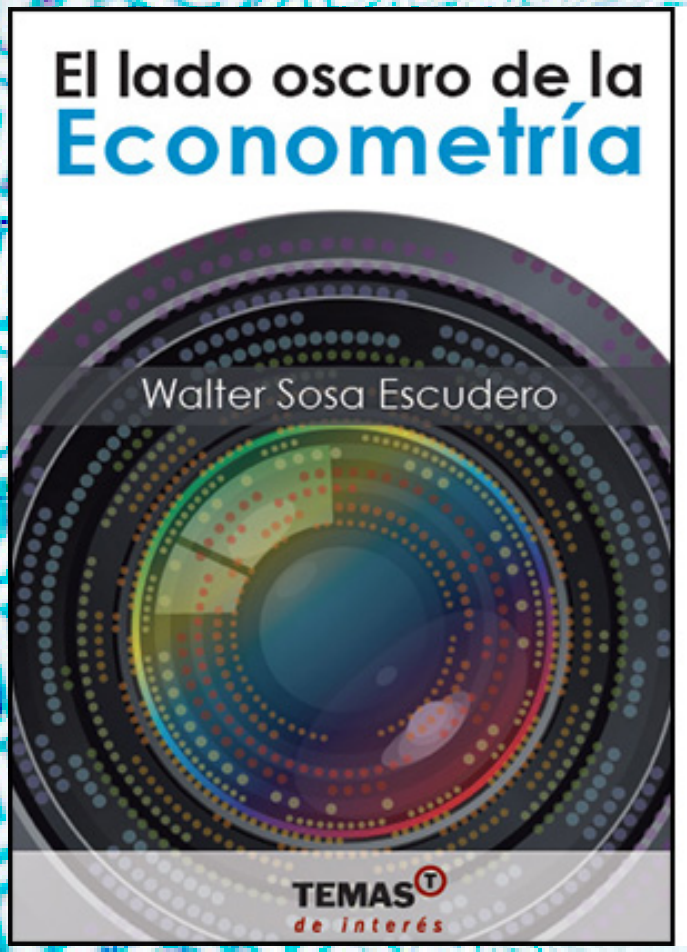
JLB

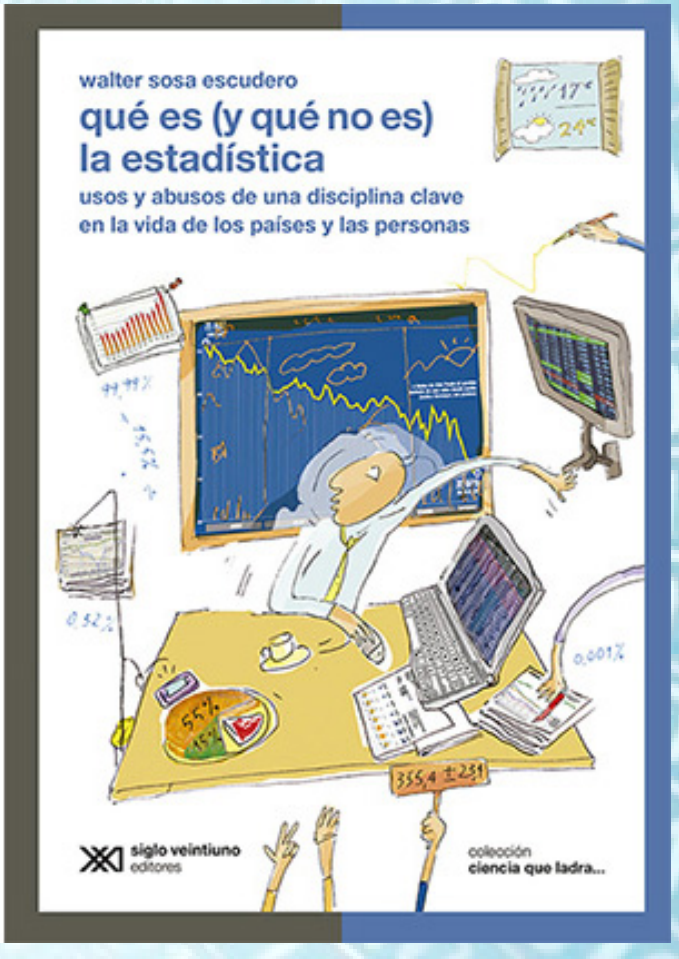
'... su antepasado no creia en un tiempo uniforme, absoluto. Creia en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. ... No existimos en la mayoría de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los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

El jardin de senderos que se bifurcan

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

Funes, el memorioso


El lado oscuro de la
Econometría
Walter Sosa Escudero
TEMAS
de interés


walter sosa escudero
qué es (y qué no es)
la estadística
usos y abusos de una disciplina clave
en la vida de los países y las personas
siglo veintiuno
editores
colección
ciencia que ladra...