MÈTODE Science Studies Journal (2014). Universitat de València.
DOI: 10.7203/metode.83.3590
Artículo recibido: 29/04/2014, aceptado: 02/06/2014.

MACRODATOS Y ESTADÍSTICA

LA PERSPECTIVA DE UN ESTADÍSTICO

DAVID ROSSELL

Los macrodatos (*big data*) representan un recurso sin precedentes para afrontar retos científicos, económicos y sociales, pero también incrementan la posibilidad de caer en conclusiones engañosas. Por ejemplo, el uso de enfoques basados exclusivamente en datos y que se despreocupan de comprender el fenómeno en estudio, que se orientan a un objetivo escurridizo y cambiante, que no tienen en cuenta problemas cruciales en la recopilación de datos, que resumen o «cocinan» inadecuadamente los datos y que confunden el ruido con la señal. Repasaremos algunos casos exitosos e ilustraremos cómo pueden ayudar los principios de la estadística a obtener una información más fiable de los datos. También abordaremos los retos actuales que requieren estudios metodológicos dinámicos como las estrategias de eficiencia computacional, la integración de datos heterogéneos, extender los fundamentos teóricos a cuestiones cada vez más complejas y, quizás lo más importante, formar una nueva generación de científicos capaces de desarrollar e implantar estas estrategias.

Palabras clave: macrodatos, estadística, estudios de caso, trampas, retos.

¿QUÉ SON LOS MACRODATOS?

En los últimos años se ha producido un incremento significativo en nuestra capacidad de recopilar, almacenar y compartir datos. Según IBM, el 90 % de los datos del mundo se ha generado en los últimos dos años (International Business Machines Corporation, 2011). Estos datos proceden de internet (búsquedas, redes sociales, blogs, imágenes), teléfonos de última generación, estudios científicos (genómica, imágenes cerebrales, epidemiología, medio ambiente), negocios (datos de clientes, transacciones, indicadores financieros), administración (población, salud, clima, sensores automáticos) y otras fuentes.

La importancia estratégica de los macrodatos no radica en la cantidad sino en las aplicaciones potenciales que ofrecen. Por ejemplo, la caracterización de enfermedades complejas a escala molecular combinadas con el historial médico y de tratamiento y con pruebas diagnósticas o de imagen ofrece oportunidades sin precedentes para personalizar la medicina. El Gran Colisionador de Hadrones registra datos 40 millones de veces por segundo para comprobar las teorías de la física. Los sitios web generan cada día millones de recomendaciones, y comparan nuevos productos y sus precios. Los datos pueden ayudar a gestionar ciudades o recursos naturales, a estudiar el cambio climático o a promover el desarrollo de regiones. Las notas en blogs y redes sociales se aprovechan para diseñar estrategias políticas y para estudiar cómo se difunden las ideas.

> «LA EXPERIENCIA HA ENSEÑADO A LOS ESTADÍSTICOS QUE LOS DATOS PUEDEN SER ENGAÑOSOS Y, LO QUE ES PEOR, DAR UNA SENSACIÓN ERRÓNEA DE OBJETIVIDAD»

Gracias al amplio alcance de todo este potencial, los medios de comunicación, el mundo académico y el de los negocios han acogido los macrodatos con un entusiasmo rayando a veces el sensacionalismo. Términos como *avalancha de datos* o *tsunami* se han hecho comunes. El Foro Económico Mundial de 2012 declaró los datos como un nuevo tipo de activo económico comparable a la moneda o al oro (Foro Económico Mundial, 2012). Las profesiones relacionadas con el manejo de datos encabezan constantemente muchas clasificaciones. Dejando aparte el bombo publicitario, revisaremos tanto los logros como

las limitaciones y destacaremos las lecciones aprendidas y los retos pendientes. Aunque los macrodatos requieren un enfoque pluridisciplinar, adoptaremos un punto de vista estadístico. La estadística es una disciplina dedicada específicamente a recopilar, analizar e interpretar datos. Es decir, nos lleva de las preguntas a los datos, de los datos a la información y de la información al conocimiento y a la toma de decisiones. Puede parecer sorprendente, pues, que los estadísticos hayan sido relativamente cautelosos a la hora de acoger los macrodatos como una fuerza todopoderosa. Yo creo que la explicación es sencilla. La experiencia ha enseñado a los estadísticos que los datos pueden ser engañosos y, lo que es peor, dar una sensación errónea de objetividad. Aunque sean poderosos, los macrodatos también abren la puerta a muchas confusiones. Debido a la variedad de aplicaciones (los macrodatos a menudo se definen como las tres V: volumen, velocidad y variedad), no podremos abarcar todo lo referente a los macrodatos, por eso me limitaré a abordar algunos de los principales problemas y a poner algunos ejemplos.

Maximilien Brice (2009 CERN)

La importancia estratégica de los datos no radica en la cantidad sino en los usos potenciales. Por ejemplo, el Gran Colisionador de Hadrones registra datos 40 millones de veces por segundo para poner a prueba las teorías de la física.

LOS DATOS Y EL PROCESO SUBYACENTE

El relato de cómo el gerente de béisbol Billy Beane aplicó indicadores de rendimiento y análisis de datos para formar un equipo competitivo (Lewis, 2003) se ha convertido ya en todo un clásico contemporáneo de los casos de éxito en el aprovechamiento de los datos, tanto que incluso dio lugar a una película de Hollywood bastante popular. El mérito más notable de Bean es que su equipo jugaba mejor que rivales con mayor presupuesto y dirigidos por expertos en béisbol. Los sondeos electorales británicos (Curtice y Firth, 2008) y estadounidenses (Silver, 2012), cuya extraordinaria precisión trituró las previsiones de los analistas políticos, son otros éxitos recientes. Otros casos son los de las predicciones meteorológicas que anunciaban catástrofes naturales (Silver, 2012), o la explosión de las tecnologías *-ómicas* en las que se basan muchos, si no la mayoría, de los avances recientes en biomedicina.

«LAS NUEVAS TECNOLOGÍAS SON INÚTILES A MENOS QUE CIENTÍFICOS BRILLANTES PLANTEEN PREGUNTAS RELEVANTES E INTERPRETEN LOS RESULTADOS EN EL CONTEXTO QUE CORRESPONDA»

Estas historias pueden haber dado la falsa impresión de que con los datos basta. Por ejemplo, en una entrevista publicada por *The New York Times* se afirmaba que los datos pueden reemplazar la experiencia y la intuición, lo que facilita un enfoque más científico (Lohr, 2012). No podría estar menos de acuerdo con este punto de vista, que ilustra un posible problema de los macrodatos. Si bien es cierto que las opiniones no contrastadas con datos pueden conducir a conclusiones erróneas, también los análisis ciegos llevan a error con frecuencia. Disponer de datos fiables y de sólidos conocimientos, lejos de oponerse, se complementan. En los anteriores ejemplos, las predicciones tuvieron éxito porque estudiaban sistemas fundamentalmente reproducibles, e implicaban la comprensión del fenómeno que estudiaban. Las variables elegidas para predecir el rendimiento en el béisbol se prestaban a una interpretación natural de la materia de estudio. Y los pronósticos de Silver aprovechaban sus conocimientos sobre la política norteamericana. Las predicciones meteorológicas se basan en simulaciones informáticas y leyes físicas, que los meteorólogos corrigen posteriormente para eliminar las imprecisiones sistemáticas. Las nuevas tecnologías son inútiles a menos que científicos

Leaders (Executive Sport Ltd)

El relato de cómo el gerente de béisbol Billy Beane aplicó indicadores de rendimiento y análisis de datos para formar un equipo competitivo se ha convertido ya en todo un clásico contemporáneo de las historias de éxito del aprovechamiento de los datos, tanto que incluso dio lugar a una película de Hollywood de bastante éxito.

brillantes planteen preguntas relevantes e interpreten los resultados en el contexto que corresponda.

Un mantra de la estadística indica que la correlación no implica causalidad. Nathan Eagle se adelantó en la predicción del cólera en Ruanda a partir de los datos de movilidad que extrajo de las llamadas de teléfonos móviles (Shaw, 2014). Eagle observó que la movilidad estaba correlacionada con los brotes de cólera y que, por tanto, podía ayudar a predecirlos. Después descubrió que la movilidad realmente predecía las inundaciones, que reducen la movilidad e incrementan a corto plazo el riesgo de brotes de cólera. Actualmente incorpora información sobre la actividad de las poblaciones en sus predicciones. No hay nada que pueda reemplazar a la comprensión del fenómeno que se estudia, es decir, el proceso de generación de datos, para poder analizarlo.

DINÁMICA DE DATOS

Los Centros de Control y Prevención de Enfermedades (CDC) de los EE UU remiten semanalmente el número de visitas médicas por enfermedades de tipo gripal, pero los resultados van con tres semanas de retraso, que es lo que cuesta procesarlos. Google Flu Trends (GFT) utiliza el número de búsquedas en Internet relacionadas con la gripe para predecir el eventual informe de los CDC para la semana en curso, proporcionando un seguimiento en tiempo real que en una ocasión se consideró más preciso que los informes de los propios CDC. Aunque GFT no lo pretendía, se ha convertido en el buque insignia de la sustitución de los métodos tradicionales por macrodatos. Sin embargo, Lazer *et al*. (2014), entre otros, han averiguado que las predicciones de GFT no son tan certeras. Aunque al principio eran precisas, desde entonces las visitas reales siempre se han sobreestimado. Predecir simplemente una semana a partir de los informes de los CDC de tres semanas atrás da mejores resultados. Lazer *et al*. argumentan que la caída en la precisión de GFT se debe sobre todo a los cambios en el motor de búsqueda de Google. Este ejemplo ilustra otra trampa importante. En el caso del béisbol y en el resto de ejemplos anteriores, el proceso subyacente que genera los datos suele permanecer constante a lo largo del tiempo. El béisbol tiene unas reglas fijas, la intención de voto no varía mucho a corto plazo, y las leyes de la naturaleza son constantes. Por el contrario, los cambios en los buscadores alteran el proceso de generación de los datos que se introducen en GFT y por consiguiente modifican la relación con el resultado que intentamos predecir.

«LA TEORÍA NOS ENSEÑA QUE, EN PRINCIPIO, TENER MUCHOS DATOS SIEMPRE ES BUENO. UNA TRAMPA TENTADORA CONSISTE EN FORZAR LOS DATOS HASTA QUE PAREZCAN APOYAR UNA IDEA PRECONCEBIDA»

Esta incertidumbre, en la literatura estadística, se conoce como *sistema dinámico* y requiere técnicas especiales para incorporar su peculiar estructura y poder reflejar la incertidumbre de manera fidedigna. Las predicciones se basan en los datos observados y, por tanto, un supuesto implícito es que los datos futuros serán similares o al menos evolucionarán de una forma previsible. Cuando pueden darse cambios repentinos, la confianza en nuestras predicciones disminuye. Consideremos el fracaso a la hora de prever los impagos de hipotecas en la Gran Recesión. El riesgo de impago se calculaba a partir de los datos recopilados durante un período de crecimiento económico generalizado. En estos períodos el riesgo de que los individuos A y B dejen de pagar sus hipotecas no presenta ninguna correlación en particular. Por tanto el riesgo de impagos generalizados se considera bajo y aunque algunos individuos dejen de pagar, seguramente otros continuarán siendo solventes. Sin embargo, en períodos de crisis los impagos están estrechamente correlacionados. Si la economía va mal y el precio de la vivienda cae, mucha gente se volverá insolvente a la vez y las posibilidades de una crisis generalizada serán mucho mayores (Gorton, 2009). Este ejemplo ilustra un sesgo conocido como extrapolación. Incluso cuando sabemos algo sobre el proceso de generación de datos, es arriesgado

hacer predicciones en situaciones en las que hay pocos o ningún dato disponibles. La mayoría de métodos están diseñados para producir predicciones que sean válidas en general, pero aunque la mayoría de las predicciones sean precisas, las que se desarrollan en escenarios poco frecuentes (por ejemplo, pacientes con una variante rara de una enfermedad) pueden fallar completamente. Es necesario, por tanto, examinar cuidadosamente el problema que nos ocupa.

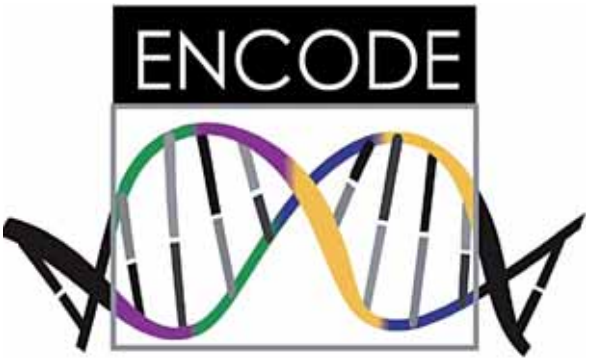
SEÑAL, RUIDO Y SESGO

La teoría nos enseña que, en principio, tener muchos datos siempre es bueno. Con nuevos datos se incrementa el potencial para obtener más información y, si este no fuese el caso, siempre se podría descartar el dato. No parece que tener más datos sea perjudicial. La pega de este razonamiento es que en la práctica no descartamos datos sino que tratamos de buscarles algún patrón. Una trampa tentadora consiste en forzar los datos hasta que parezcan apoyar una idea preconcebida. Eso no quiere decir que el análisis de datos no pueda ser motivado por una hipótesis previa, sino que se necesita una estrategia adecuada para reducir la probabilidad de obtener resultados no reproducibles. Las últimas décadas han mostrado avances apasionantes en métodos estadísticos orientados a distinguir la señal del ruido en los datos masivos. Pero estos avances todavía no han calado en los análisis rutinarios de datos. Nuzzo (2014) considera que, al observar un valor *p* de 0,01 para una hipótesis con diecinueve probabilidades contra una de no ser cierta, la probabilidad de que se trate de un falso positivo es del 0,89. Con los macrodatos a menudo se registran datos simplemente porque los podemos obtener, no porque se espere incrementar sustancialmente la señal. La probabilidades, por tanto, son muy superiores a diecinueve contra una y las posibilidades de falsos positivos se disparan.

Otra cuestión fundamental es que los macrodatos a menudo proceden de diferentes fuentes, se han obtenido mediante diferentes técnicas o presentan diferentes formatos. No necesariamente tienen que ser comparables o presentar la misma calidad y a menudo están sometidos a varios sesgos sistemáticos. Por ejemplo, el proyecto Encode es una de las mayores iniciativas posteriores al Proyecto Genoma Humano. Los datos se recolectaron en laboratorios repartidos por todo el mundo, usando múltiples tecnologías y procedimientos experimentales. Cuando desarrollamos un sistema para visualizar estos macrodatos, encontramos sesgos sistemáticos entre los microbiochips y las tecnologías de secuenciación que se debían corregir para evitar interpretaciones erróneas (Font-Burgada *et al*., 2013).

MÈTODE

Los Centros de Control y Prevención de Enfermedades (CDC) de los EE UU remiten semanalmente el número de visitas médicas por enfermedades de tipo gripal, pero los resultados van con tres semanas de retraso, que es lo que cuesta procesarlos. Google Flu Trends (GFT) utiliza el número de búsquedas en Internet relacionadas con la gripe para predecir el eventual informe de los CDC para la semana en curso, proporcionando un seguimiento en tiempo real que ha sido calificado de más preciso que los informes de los propios CDC.



ENCODE PROJECT

Los macrodatos a menudo proceden de diferentes lugares, se han obtenido mediante diferentes técnicas o presentan diferentes formatos. No siempre son comparables o presentan la misma calidad y a veces están sometidos a varios sesgos sistemáticos. Este es el tipo de problemas que experimenta el proyecto Encode, una de las iniciativas más importantes posteriores al Proyecto Genoma Humano. Los datos se recolectaron en laboratorios repartidos por todo el mundo, usando múltiples tecnologías y procedimientos.

Más en general, visualizar datos heterogéneos que sean fáciles de entender es un reto, pero se están haciendo progresos. Por ejemplo, con las técnicas de visualización del flujo sanguíneo ideadas por Michelle Borkin y sus tutores se incrementaba de un 39 a un 91 % la capacidad de los médicos para diagnosticar obstrucciones arteriales (Shaw, 2014). En el pasado, los métodos de metaanálisis se concibieron para combinar indicios de diferentes estudios siguiendo un procedimiento riguroso. Los macrodatos requieren nuevos métodos para poder integrar y visualizar los datos de manera fiable.

MÈTODE

Las sugerencias de películas que hace Netflix utilizan un modelo que promedia 107 predicciones. La teoría de la decisión puede ayudar a evaluar las ventajas de algoritmos complejos en un contexto dominado por la incertidumbre y los objetivos contrapuestos; por ejemplo, el grado de satisfacción de los clientes también puede depender de la diversidad de las sugerencias.

PLANIFICAR

Los macrodatos están cambiando la forma de recopilar las pruebas. En lugar de diseñar cuidadosamente un estudio, la tendencia suele ser registrar todos los datos que sea posible, aceptando de forma implícita que cualquier patrón que se observe en ellos seguramente será relevante. Esta idea falsa es una trampa muy problemática. La representatividad de los datos no depende del tamaño de la muestra sino de la forma de recopilarlos. Importa más la calidad que la cantidad. Un ejemplo clásico es un estudio británico en el que se evaluaron en 20.000 niños los beneficios de la leche pasteurizada. William Gosset, más conocido como Student, señaló que, por culpa de la distribución aleatoria inadecuada, un estudio con solo seis gemelos habría sido más fiable (Student, 1931). Un factor que contribuye a la falta de atención que se presta al diseño del estudio puede ser el exceso de fe en las nuevas tecnologías. Por ejemplo, la comunidad científica ha recibido con entusiasmo la irrupción de la secuenciación de alto rendimiento (HTS). He conocido reputados investigadores que argumentan que con una sola muestra estos estudios son tan buenos como las tecnologías anteriores con docenas de muestras. Aunque la HTS sea precisa, una sola muestra no puede medir la variabilidad para comparar poblaciones. Otra anécdota es que algunos centros de HTS procesan dos muestras en diferentes fechas cuando deberían procesar en paralelo para evitar sesgos. Como resultado, experimentos muy caros han dado resultados prácticamente inútiles.

> «LA REPRESENTATIVIDAD DE LOS DATOS NO DEPENDE DEL TAMAÑO DE LA MUESTRA SINO DE LA FORMA DE RECOPILARLOS. IMPORTA MÁS LA CALIDAD QUE LA CANTIDAD»

La extensión de la teoría sobre el diseño de experimentos formulada por Ronald Fisher a los macrodatos ha sido en su mayoría desatendida, pero hay notables excepciones. Dado que vamos hacia la medicina personalizada, Berry (2012) ha defendido los ensayos clínicos adaptados a grupos cada vez más reducidos y la toma de decisiones adaptadas a cada paciente. Müller *et al*. (2004) han propuesto diseños rigurosos para estudios de comprobación de hipótesis masivas. También han tenido éxito propuestas de diseño de estudios observacionales. Para mostrar las ventajas del seguro público de salud en México, King *et al*. (2009) elaboraron un estudio que comparaba las comunidades con este seguro y las que no lo tenían. Como estas mostraban características similares, las diferencias entre los resultados en salud se pueden atribuir más al seguro que a factores externos.

UN CASO PARA LA ESTADÍSTICA

De forma similar a las bases que sentaron pioneros como Ronald Fisher, William Gosset o Harold Jeffreys en la aplicación de los datos a la ciencia, los negocios y la política, el paradigma de los macrodatos se alimenta de contribuciones metodológicas. El algoritmo PageRank utilizado por Google se basa en las cadenas de Markov. Las sugerencias de películas que hace Netflix utilizan un modelo que promedia 107 predicciones. La teoría de la decisión puede ayudar a evaluar las ventajas de complejos algoritmos en un contexto dominado

por la incertidumbre y los objetivos contradictorios, por ejemplo, el grado de satisfacción de los clientes también puede depender de la diversidad de las sugerencias.

Ya hemos expuesto la necesidad de explorar nuevos métodos para separar la señal del ruido, capturar procesos dinámicos, diseñar experimentos e integrar datos heterogéneos. Los métodos computacionales que combinan potencia de procesamiento con estrategias inteligentes para resolver problemas complejos son otro de los temas centrales, ya que es poco probable que tengan éxito los enfoques exhaustivos o de fuerza bruta. Otros retos son la recuperación y el resumen de datos. Los métodos automáticos para detectar y dar formato a los datos no estructurados (como imágenes o blogs) pueden descartar información o inducir sesgos. Otro problema es que actualmente generamos más datos de los que podemos almacenar (Hilbert, 2012), lo que obliga a resumir los datos. Y los resúmenes implican el riesgo de perder información. Como ejemplo, hace poco informamos de que la estrategia que actualmente se aplica para recapitular los datos de la secuenciación de ARN descarta tanta información que ciertos detalles se escapan aunque la cantidad de datos vaya creciendo hasta el infinito (Rossell *et al*., 2014). Un tema relacionado es el de la toma de muestras. Almacenar una muestra apropiada obtenida de todos los datos puede incrementar la velocidad y reducir costes, con una pérdida insignificante en la precisión. Fan *et al*. (2014) y Jordan (2013) han abordado cuestiones relativas a la estadística y el procesamiento de macrodatos.

La estadística, como disciplina que combina razonamiento científico, teoría de la probabilidad y matemáticas, es un componente necesario para que la revolución de los macrodatos alcance todo su potencial. Sin embargo, la estadística no puede funcionar de forma aislada sino que necesita la colaboración de conocimientos técnicos, de la informática y de otras disciplinas relacionadas. Como reflexión final, el principal obstáculo para superarse bien puede ser la falta de profesionales con la combinación adecuada de capacidades. La selección y la formación de jóvenes talentos dispuestos a participar en esta excitante aventura debería ser una prioridad.

REFERENCIAS

Berry, D., 2012. «Adaptive Clinical Trials in Oncology». *Nature Reviews Clinical Oncology*, 9: 199-207. DOI: <10.1038/nrclinonc.2011.165>.

Curtice, J. y D. Firth, 2008. «Exit Polling in a Cold Climate: the BBC-ITV Experience Explained». *Journal of the Royal Statistical Society A*, 171(3): 509-539. DOI: <10.1111/j.1467-985X.2007.00536.x>.

Fan, J.; Han, F. y H. Liu, 2014. «Challenges of Big Data Analysis». *National Science Review*, 1(2): 293-314. DOI: <10.1093/nsr/nwt032>.

Font-Burgada, J.; Reina, O.; Rossell, D. y F. Azorín, 2013. «ChroGPS, a Global Chromatin Positioning System for the Functional Analysis and Visualization of the Epigenome». *Nucleic Acids Research*, 42(4): 1-12. DOI: <10.1093/nar/gkt1186>.

Foro Económico Mundial, 2012. *Big Data, Big Impact: New Possibilities for International Development*. Foro Económico Mundial. Cologny, Suiza. Disponible en: <www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf>.

Gorton, G., 2009. «Information, Liquidity, and the (Ongoing) Panic of 2007». *American Economic Review*, 99(2): 567-572. DOI: <10.1257/aer.99.2.567>.

Hilbert, M., 2012. «How Much Information Is There in the "Information Society"?». *Significance*, 9(4): 8-12. DOI: <10.1111/j.1740-9713.2012.00584.x>.

International Business Machines Corporation, 2011. *IBM Big Data Success Stories*. International Business Machines Corporation. Armonk, NY. Disponible en: <http://public.dhe.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf>.

Jordan, M., 2013. «On Statistics, Computation and Scalability». *Bernoulli*, 19(4): 1378-1390. DOI: <10.3150/12-BEJSP17>.

King, G. *et al*., 2009. «Public Policy for the Poor? A Randomized Assessment of the Mexican Universal Health Insurance Programme». *The Lancet*, 373: 1447-1454. DOI: <10.1016/S0140-6736(09)60239-7>.

Lazer, D.; Kennedy, R.; King, G. y A. Vespignani, 2014 «The Parable of Google Flu: Traps in Big Data Analysis». *Science*, 343(6176): 1203-1205. DOI: <10.1126/science.1248506>.

Lewis, M., 2003. *Moneyball. The Art of Winning an Unfair Game*. W. W. Norton & Company. Nueva York.

Lohr, S., 2012. «The Age of Big Data». *The New York Times*, 11 de febrero de 2012. Disponible en: <www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>.

Müller, P.; Parmigiani, G.; Robert, C. y J., Rousseau, 2004. «Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays». *Journal of the American Statistical Association*, 99(468): 990-1001. DOI: <10.1198/016214504000001646>.

Nuzzo, R., 2014. «Scientific Method: Statistical Errors». *Nature*, 506: 150-152. DOI: <10.1038/506150a>.

Rossell, D.; Stephan-Otto Attolini, C.; Kroiss, M. y A. Stöcker, 2014. «Quantifying Alternative Splicing from RNA-Sequencing Data». *The Annals of Applied Statistics*, 8(1): 309-330. DOI: <10.1214/13-AOAS687>.

Shaw, J., 2014. «Why "Big Data" Is a Big Deal». *Harvard Magazine*, 3: 30-35, 74-75. Disponible en: <http://harvardmag.com/pdf/2014/03-pdfs/0314-30.pdf>.

Silver, N., 2012. *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*. Penguin Press. Nueva York.

Student, 1931. «The Lanarkshire Milk Experiment». *Biometrika*, 23(3-4): 398-406. DOI: <10.2307/2332424>.

ABSTRACT

Big Data and Statistics: A Statistician's Perspective.

Big Data brings unprecedented power to address scientific, economic and societal issues, but also amplifies the possibility of certain pitfalls. These include using purely data-driven approaches that disregard understanding the phenomenon under study, aiming at a dynamically moving target, ignoring critical data-collection issues, summarizing or preprocessing the data inadequately and mistaking noise for signal. We review some success stories and illustrate how statistical principles can help obtain more reliable information from data. We also touch upon current challenges that require active methodological research such as strategies for efficient computation, integration of heterogeneous data, extending the underlying theory to increasingly complex questions and, perhaps most importantly, training a new generation of scientists who can develop and deploy these strategies.

Keywords: Big Data, statistics, case studies, pitfalls, challenges.

AGRADECIMIENTOS:
Trabajo parcialmente financiado por NIH grant R01 CA158113-01.

David Rossell. Profesor del departamento de Estadística. Universidad de Warwick (Reino Unido).