CB 10

BIG DATA Y LAS MATEMÁTICAS

Pedro A. Willging

Facultad de Ciencias Exactas y Naturales
Universidad Nacional de La Pampa
Av. Uruguay 151 – Santa Rosa – La Pampa
pedro@exactas.unlpam.edu.ar

Palabras Clave: big data, correlación, minería de datos, estadística.

RESUMEN
Se presenta el concepto de "*big data*" y sus posibles implicancias en el ámbito de la enseñanza de la matemática, particularmente en el área del análisis estadístico y la modelización de los datos. Además de presentar los ideas centrales relacionados con este concepto novedoso, se incluye la discusión de su importancia en el futuro próximo, particularmente para la educación matemática.

INTRODUCCIÓN
Con un título provocador: "El fin de la teoría: el diluvio de datos hace obsoleto el método científico", Chris Anderson – escritor, físico y periodista- sacudía al mundo científico con su artículo en la revista Wired[1] en 2008. Anderson, reconocido en el ámbito de las innovaciones tecnológicas, entre otras cosas por acuñar el término y la teoría de "la larga estela" ("*long tail*" en inglés[2]) hizo esta afirmación basándose en los logros obtenidos en la predicción de fenómenos diversos de manera muy acertada por la aplicación de herramientas analíticas en el tratamiento de grandes volúmenes de datos. En el 2009, cuando se produce la epidemia del virus H1N1[3], un grupo de investigadores de la compañía Google fue capaz de predecir con rapidez, certeza y en tiempo real donde y como se estaba dispersando la gripe. La herramienta que utilizaron se denomina *Google Flu Trends*, que busca correlaciones entre las frecuencias de búsquedas en Internet y la difusión espacial y temporal de la gripe. El descubrimiento de Google demostró ser mas rápido que el sistema de los Centros de Control y Prevención de Enfermedades de los EEUU y fue anunciado como un logro destacado en la revista Nature (Ginsberg *et al*., 2009). Entra de este modo en escena, con grandes expectativas una nueva tendencia en negocios, ciencia y tecnología: "*Big Data*". El entusiasmo ha sido tan generalizado que algunos periodistas se han preguntado: "¿Qué puede aprender la ciencia de

1 La revista Wired es una publicación mensual norteamericana que existe desde 1993 y a la vez un sitio web de noticias (wired.com) donde se tratan temas relacionados con las tecnologías y el modo en que ellas afectan a la cultura, la educación, la economía y la política. Chris Anderson fue jefe de editores de Wired desde 2001 hasta 2012.

2 Ver el artículo original de Anderson, o el más elaborado libro subsiguiente: Anderson, Chris. "The Long Tail" Wired, October 2004 (http://archive.wired.com/wired/archive/12.10/tail.html). Anderson, Chris. (2006). The Long Tail: Why the future of business is selling less of more (New York: Hyperion Books).

3 El H1N1 es un nuevo virus detectado por primera vez en los humanos en el 2009 que se propagó rápidamente alrededor del mundo. Durante la temporada de gripe del 2010-2011, el virus H1N1 no causó infecciones generalizadas. No se ha necesitado una vacuna aparte, aunque este virus aún es uno de los tres virus incluidos en la vacuna (estacional) regular.

Google?" (Anderson, 2008)

¿QUÉ ES *BIG DATA*[4]?

Como suele ocurrir con los neologismos relacionados con las innovaciones tecnológicas y que se difunden en los medios por cuestiones de moda, el término "*big data*" es aún un término vago, sin una definición rigurosa (y de hecho no tenemos una traducción en nuestro idioma que se utilice más que el término en su idioma original). Se asocia con este concepto a los datos recolectados en experimentos como el Colisionador de Hadrones del CERN[5] o el viaje del telescopio Hubble[6] que generan cantidades enormes de información (i.e. datos en diferentes formatos). También la actividad en sitios web como Facebook, Twitter, o las búsquedas en Google. Estas actividades generan cantidades gigantescas de transacciones, que solo es posible analizar con herramientas de minería de datos, las cuales se han mejorado en los últimos tiempos aprovechando los avances en la velocidad de procesamiento de las computadoras y en la disponibilidad de mayor almacenamiento. La disponibilidad de muchos de estos datos es cada vez más sencillo de obtener, entre otros motivos por la creciente digitalización de las transacciones, y por la ubicuidad de la información. Este período de disponibilidad exponencial de información ha dado en llamarse *la era Petabyte*. El byte es la unidad de información en dispositivos de almacenamiento de datos. Un petabyte es 10 elevado a la 15 bytes. Ver la Tabla 1 para comparar con otras unidades de medición de almacenamiento de datos.

Múltiplo (símbolo)	Bytes
Kilobyte (KB)	10^3
Megabyte (MB)	10^6
Gigabyte (GB)	10^9
Terabyte (TB)	10^{12}
Petabyte (PB)	10^{15}
Exabyte (EB)	10^{18}
Zetabyte(ZB)	10^{21}
Yotabyte(YB)	10^{24}

Tabla 1: Unidades de información/almacenamiento de datos

Para dar algunos ejemplos concretos de lo que son los volúmenes de información actual, se puede mencionar: las observaciones del Hubble, incluyendo unas 500.000 fotografías, ocupan 1.420 discos ópticos de 6,66 GB, Google procesa más de 24 petabytes de datos por día y Facebook tiene 60 mil millones de imágenes, es decir 1,5 petabytes. ¡Esto es big data! La cantidad de datos acumulados en los últimos 2 años, proveniente de actividades de la vida diaria (como los teléfonos y las tarjetas de crédito, y los sensores inteligentes de los edificios o los medios de transporte entre otros), que totaliza un zetabyte, empequeñece todos los registros previos de la humanidad. Estamos frente a una nueva revolución de la información, pero que no tiene que ver solo con que tenemos una cantidad enorme de datos, sino que ahora

4 Big data="grandes datos".

5 El Gran Colisionador de Hadrones -Large Hadron Collider (LHC)- es el acelerador de partículas mas grande y poderoso del mundo. Es parte de los proyectos que lleva a cabo el CERN (Comision Europea de Investigaciones Nucleares) . Ver: http://home.web.cern.ch/topics/large-hadron-collider

6 El telescopio espacial Hubble (HST por sus siglas en inglés), es un telescopio que orbita en el exterior de la atmósfera a 593 km sobre el nivel del mar. Ver: http://hubblesite.org/the_telescope/

se sabe cómo hacer algo con esos datos. Y esa revolución, afirma Shaw (2014) tiene más que ver con la mejora en los métodos estadísticos y computacionales que con el crecimiento exponencial de la capacidad de almacenamiento o inclusive del poder computacional. Lo que antes requería el trabajo de días en una costosa super-computadora, ahora lo puede realizar una notebook en minutos con el algoritmo adecuado. Un ejemplo más de esta aceleración en los análisis de grandes volúmenes de datos lo proporciona el caso del genoma humano. La primera secuencia completa del genoma humano se completó en 2001 luego de 15 años de trabajo a un costo de entre 1.000 a 3.000 millones de dólares. Hoy una secuencia del genoma se logra en 24 horas y cuesta 1.000 dólares.

MODELOS, CORRELACIÓN Y DESPROLIJIDADES

El método científico se basa en la posibilidad de comprobar hipótesis. Los modelos mentales creados por los científicos se testean por medio de experimentos que o bien confirman o rechazan las teorías de cómo funciona el mundo. Con estos procedimientos de validación la ciencia ha trabajado por siglos. Pero la disponibilidad masiva de datos podría estar desafiando la aproximación a la ciencia consistente en hipótesis, modelo, test. La era petabyte nos permitiría prescindir de los modelos y solamente buscar correlaciones en los datos con poderosos algoritmos que descubran los patrones que subyacen. "La nueva disponibilidad de enormes cantidades de datos, junto con las herramientas estadísticas para estrujar esos números, ofrece un modo completamente nuevo de entender el mundo. La correlación sustituye a la causalidad, y la ciencia puede avanzar aun sin modelos coherentes, teorías unificadas, o realmente ninguna explicación mecánica por completo." (Anderson, 2008).

Dentro de las aplicaciones actuales de *big data*, se pueden citar las estrategias de marketing de sitios como Amazon, que sugieren compras a los clientes basándose en los gustos e intereses de millones de clientes previos. También el sitio Farecast, que aconsejaba sobre el precio más conveniente de un pasaje de avión, indicándoles a sus clientes el momento exacto en el cual debían comprarlo (hasta que la compañía fue comprada por Microsoft para incorporarla a su navegador Bing y dejó de ofrecer sus servicios). Los cálculos de Farecast se hacían con millones de registros de precios de vuelos, obtenidos en la web y de bases de datos de la industria de la aviación (Etzioni, Tuchinda, Knoblock, & Yates, 2003). Con el desarrollo de herramientas para el estudio de billones de pares de bases del genoma humano fue posible identificar los genes que se convierten en prominentes más rápidamente en la evolución humana, lo cual sirve para determinar, por ejemplo, la resistencia a enfermedades como la malaria (Grossman *et al.*, 2013). En África, los investigadores han realizado varios experimentos con datos de telefonía celular para monitorear patrones de desplazamiento de personas y predecir estallidos de enfermedades infecciosas como el cólera, o para establecer el nivel de suministro del banco de sangre de las zonas rurales (Wesolowski *et al*., 2012; Hill, Banser, Berhan, & Eagle, 2010).

Uno de los cambios de paradigma que introduce *big data* es que es posible vivir con la inexactitud y la desprolijidad de los datos. Los científicos han mejorado y optimizado los instrumentos y procedimientos a efectos de lograr exactitud en sus mediciones. Cuando se toma un número limitado de muestras de una población, el error en las mediciones puede llevar a la falta de precisión en los resultados generales. Es por ello que se toma tanto cuidado en obtener las mejores y más exactas medidas de unos pocos, pero costosos datos. A diferencia de lo que ocurre en el mundo de pocos datos, donde reducir los errores y asegurarse de la más alta calidad en los datos es esencial, cuando se dispone de una fuente enorme de datos relacionados con un fenómeno, es posible descuidar la precisión y la exactitud. Esto es notablemente demostrado con los avances recientes en la traducción de lenguajes. La traducción de lenguajes con máquinas se inicia durante la Segunda Guerra Mundial, y no

tuvo grandes progresos hasta 1990, cuando el proyecto *Candide* de IBM[7] utilizó los registros parlamentarios canadienses de 10 años, unas 3 millones de pares de sentencias en francés e inglés para alimentar a una computadora. Esto se convirtió en traducción de máquina estadística, convirtiendo a la traducción en un problema matemático. De pronto las traducciones mejoraron, pero aún no lo suficiente. Hasta que en 2006, Google tomó en sus manos el problema, y decidió alimentar a las computadoras con cantidades de documentos de texto muchísimo mayores que *Candide*. El diccionario de Google incluyó todo el material disponible en la web, con sus desprolijidades y de pobre calidad literaria. *Google Translate* tiene a su disposición un cuerpo de 95.000 millones de sentencias en inglés, puede traducir 60 lenguajes, incluyendo entradas en formato de audio (Franz & Brants, 2006). A diferencia del texto de las transcripciones parlamentarias canadienses, totalmente libres de errores gramaticales y ortográficos, los textos de la web de *Google Translate* contienen oraciones incompletas, errores de tipeo y ortografía, deficiencias gramaticales y otros defectos. Pero el hecho de que es millones de veces más extenso, hace que las deficiencias no resulten decisivas a la hora de obtener traducciones mucho mejores que las que se habían logrado hasta ahora.

IMPLICANCIAS PARA EL ÁMBITO EDUCATIVO

Las herramientas que se desarrollan y utilizan en el análisis de *big data* pueden, en su mayoría, emplearse en disciplinas tan dispares como la astronomía o la medicina. Por ello es que algunas instituciones educativas ya están tomando acciones para desarrollar nuevos programas, por ejemplo una maestría en biología computacional y genética cuantitativa, y otra en ciencia de los datos (Harvard). Más aún hay quien afirma que "*Big data* está teniendo un impacto transformativo a través de virtualmente todas las disciplinas académicas- es tiempo de que la ciencia de los datos sea integrada en los cursos básicos de todos grados universitarios" (N. Eagle citado en J. Shaw, 2014).

Una de las herramientas más poderosas para facilitar la comprensión de inmensas cantidades de datos es la representación visual o generación de visualizaciones. El cerebro humano es capaz de entender relaciones y fenómenos complejos y descubrir patrones, cuando están representados visualmente, en un modo que aun las computadoras no logran hacerlo. Pero las computadoras son capaces de manipular las cuantiosas cantidades de datos para convertirlas en representaciones compactas que el científico puede interpretar y significar. Para ilustrar con un caso: un grupo de médicos, físicos y científicos computacionales crearon una representación visual para modelar el flujo de sangre en las arterias coronarias (Borkin *et al.*, 2011). El objetivo de la visualización era proveer a los médicos con una herramienta para localizar los taponamientos de las arterias. Se crearon simulaciones 3D utilizando datos de 300 millones de células rojas de sangre en 12 arterias coronarias principales. Esto ha permitido a lo médicos mejorar el diagnóstico de los bloqueos arteriales en un 62% (ver imágenes en http://hemo.seas.harvard.edu/images-movies).

Toman valor entonces los cursos donde se estudie como convertir grandes masas de datos en visualizaciones. Conceptos como agregación de datos, filtrado y agrupamiento (*clustering*) son centrales para reducir o simplificar los datos y lograr que tengan sentido para el investigador que los analiza. Las herramientas de visualización que se utilizan en medicina de imágenes pueden adaptarse para su uso en astronomía u otros campos.

Ahora bien: ¿Cómo nos increpa esta revolución de los datos y qué nos obliga a responder en el ámbito particular de la educación matemática?

Una consecuencia directa que aparece cuando tenemos la posibilidad de administrar grandes

7 El proyecto "Candide" fue un sistema experimental de traducción mecánica desarrollado en el Centro de investigación IBM TJ Watson a principio de los 90's.

cantidades de datos es que el concepto de muestreo pierde sentido. No necesitamos conformarnos con una pequeña muestra cuando es posible tener a mano la totalidad de los datos, es el caso: N=total.

La posibilidad de tener *todos* los datos de la muestra sacude intensamente al área de las ciencias sociales, donde los especialistas recurren frecuentemente a estudios de muestreo y cuestionarios. Con la disponibilidad de herramientas que colectan información pasivamente, mientras la gente realiza sus actividades cotidianas, se eliminan los prejuicios asociados con la selección de la muestra. Las herramientas técnicas para manipular estos datos ya han cambiado, y drásticamente, pero los métodos de análisis estadísticos y nuestro pensamiento aún está acomodándose. Los autores Mayer-Schönberger & Cukier (2013) afirman que:

"Adentrarse en un mundo de *big data* requerirá que cambiemos nuestro pensamiento acerca de los méritos de la exactitud. Aplicar la mentalidad convencional de medición al mundo digital conectado del siglo 21 es perder un punto crucial. Como se ha mencionado antes, la obsesión con la exactitud es un artefacto de la era analógica, escasa de información. Cuando los datos eran escasos, cada punto era crítico, y por lo tanto se tomaba gran cuidado para evitar que algún punto parcialice el análisis.

Hoy no vivimos en esa situación de escasez de información. Tratando con conjuntos de datos cada vez más exhaustivos, los cuales capturan no solo una pequeña tajada del fenómeno en estudio sino mucho más del todo, no tenemos que preocuparnos demasiado de que los puntos individuales parcialicen el análisis general. En lugar de apuntar a acabar con cada pedacito de inexactitud a un costo cada vez más elevado, calculamos con las desprolijidades en mente" (p. 40).[8]

Si bien *big data* ya ha producido resultados sorprendentes con descubrimientos en situaciones donde se aplicaron correlaciones sin arriesgar ningún modelo previo, la idea de ver *qué ocurre* sin importar *por qué ocurre* puede llevar a grandes fallas. Como lo demostraron las sobre-estimaciones de *Google Flu* para el invierno de 2012 en casi un factor de dos, habiendo sido tan acertadas en predecir la expansión de la epidemia de resfrío por varias temporadas previas. El modelo libre de teoría, rico en datos se equivocó. El análisis de correlaciones sin teoría es frágil: si no se sabe que hay detrás de la correlación, es difícil saber que puede hacer que se desplome. En el caso de *Google Flu* puede ser que algún trascendido en las noticias disparara búsquedas en Internet de gente que estaba sana, o que el algoritmo diagnosticará erróneamente cuando las personas ingresaban síntomas médicos.

CONCLUSIÓN Y ADVERTENCIAS

Este artículo estaría incompleto si no se advirtiera que la novedad y el entusiasmo que genera la aparición de descubrimientos espectaculares y mediáticos como resultado de las técnicas de *big data* no deben impedirnos ser cautos y analizar con cuidado los lados débiles de esta nueva aproximación a la realidad. Sin dudas, las futuras investigaciones tratarán estas deficiencias, que incluyen los errores en la determinación de que es estadísticamente significante, el problema de las comparaciones múltiples, y los problemas éticos relacionados con la manipulación de los datos (seguridad, privacidad y transparencia).

Cuando se presentan los resultados de los trabajos de investigación, suele informarse que "los resultados son significantes con un 95% de confianza", lo cual implica que 5 de cada 100 o 1 de cada 20 veces es posible encontrar resultados que siendo rotulados como estadísticamente significantes surgen de pura casualidad. Si bien este peligro está latente en todos los resultados estadísticos, en el caso de *big data* se magnifica sustancialmente el riesgo de encontrar correlaciones espurias.

8 Traducción propia.

Esta situación ha sido mencionada recientemente en la publicación de Ioannidis (2005), donde se apunta al problema de las comparaciones múltiples, que aparece cuando un investigador busca demasiados patrones posibles en los datos. Cuando se examinan patrones en los datos, los que surgen como estadísticamente significantes son aquellos que es poco probable que aparezcan de modo aleatorio, pero si se intentan todas las posibles comparaciones en cantidades enormes de datos, es muy probable que asomen patrones falsos con mayor frecuencia.

Un modo de resolver el problema de las comparaciones múltiples es transparentar los procedimientos y las hipótesis de modo que los estudios puedan ser replicados, y allí sí poder establecer qué resultados tienen sentido. Lamentablemente aparece aquí el problema de la privacidad y seguridad de los datos. Muchas de las investigaciones más recientes relacionadas con *big data* están siendo conducidas por compañías que esperan obtener ventajas competitivas, y no están dispuestos a compartir sus bases de datos con el público, ni transparentar los procedimientos.

No existe suficiente cantidad de personas capacitadas para trabajar y analizar el volumen de datos implicados en *big data*, menos aún en ámbitos que no se relacionen con las ciencias de la computación o la matemática. Es por ello que se necesita popularizar y difundir este tipo de técnicas y conocimientos a las distintas áreas de la ciencia. La oportunidad que brinda *big data* de actualizar y desarrollar nuevos métodos estadísticos y de análisis de datos está aquí, disponible para que los científicos la aprovechen. Estos avances son necesarios y se harán posible construyendo (y no ignorando) sobre los conocimientos afianzados de la estadística y la computación.

REFERENCIAS

- Anderson, C. (2008). The end of theory: el diluvio de datos hace obsoleto el método científico, Wired Magazine, 16(7), http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory
- Borkin, M., Gajos, K., Peters, A., Mitsouras, D., Melchionna, S., Rybicki, F., Feldman, C., & Pfister, H. (2011). Evaluation of artery visualizations for heart disease diagnosis, Proceedings of Information Visualization 2011, IEEE Transactions on Visualization and Computer Graphics, 17(12).
- Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A. (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. En Proceedings de 9na Conferencia Internacional sobre Descubrimiento del Conocimiento y Minería de Datos. ACM, New York, NY, USA, 119-128. http://doi.acm.org/10.1145/956750.956767
- Franz, A., & Brants, T. (2006). All our N-gram are belong to you, Google blog post, http://googleresearch.blogspot.com.ar/2006/08/all-our-n-gram-are-belong-to-you.html
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, Nature 457, 1012-1014. doi:10.1038/nature07634
- Harford, T. (28 de Marzo, 2014). Big data: are we making a big mistake?, Financial Times Magazine, http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html
- Hill, S., Banser, A., Berhan, G., & Eagle, N. (2010). Reality Mining Africa, Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium, 45-50.
- Ioannidis, J. P. A. (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

- Mayer-Schönberger, V., & cukier, K. (2013). Big data. A revolution that will trasnform how we live, work and think. London: John Murray Publishers.
- Shaw, J. (marzo-abril 2014). Why "Big Data" is a big deal, Harvard Magazine. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal
- Wesolowski, A., Eagle, N., Noor, A., Snow, R., & Buckee, C. (2012). Heterogenous mobile phone ownership and usage patterns in Kenya, PLoS ONE, 7(4): e35319.