

Algunas consideraciones sobre el enfoque de Colecciones Masivas de Datos y el papel que puede desempeñar en la Educación

Dr. Luis Cruz Kuri
Dr. Sergio Francisco Juárez Cerrillo
Dr. Ismael Sosa Galindo
Universidad Veracruzana

Línea temática: Recursos educativos abiertos.

Palabras clave: Big Data, Ciencia de datos, reconocimiento de patrones.

Resumen

Una definición tentativa de *Colección Masiva de Datos* (Big Data) corresponde a la existencia de más datos que los que pueden almacenarse y procesarse en una sola computadora. En un congreso celebrado en Inglaterra (2014) y organizado por la UGL (Global University of London) precisamente abordando la temática sobre el papel que el enfoque de las colecciones masivas de datos pueden desempeñar en la educación, se reportan entre otros hallazgos clave que: (1) Las tecnologías del manejo de colecciones masivas de datos tienen el potencial para revolucionar el aprendizaje, pero es esencial que, junto con la innovación tecnológica, se desarrollen políticas directivas claras que protejan la privacidad de los alumnos e instituciones académicas, entre otras; (2) Los cursos masivos abiertos *online* (MOOC por sus siglas en inglés) permiten compartir en la Web grandes cantidades de información, pero éstos cursos no son un sustituto para la enseñanza. Una consideración importante para atemperar el desarrollo de tales tecnologías se refiere a que: Las colecciones masivas de datos no constituyen la interrogante, ni la respuesta, sino que simplemente proporcionan una oportunidad. En nuestro presente trabajo se exploran algunas nociones (y controversias) sobre este enfoque, dentro del cual existen traslapes con la Inteligencia Artificial, con el Reconocimiento de Patrones (PR), con la Estadística y con la Teoría de la Toma de Decisiones, entre otras disciplinas establecidas; al igual que con la supuestamente más reciente “Ciencia de los Datos”.

Introducción

Vivimos la era de los datos. En esta era, que preferimos llamar la era de la información, entendida en el sentido del sociólogo Castells en su monumental obra -Castells (2006)- los datos determinan la forma en que pensamos. Día a día dejamos nuestras huellas digitales con nuestro comportamiento online y offline en internet, redes sociales y telefonía celular. Ese rastro digital nos sigue como nuestra sombra; además aumenta con cada llamada, cada mensaje de texto, cada *tweet*, cada WhatsApp, cada compra en línea, cada click que damos. Todo queda registrado. Así, los algoritmos de aprendizaje máquina de Google y Facebook no sólo descubren nuestras preferencias e intereses sino que les dan forma y las dirigen, predictivamente, hacia el siguiente *like* que damos en internet. De acuerdo al Human Face of Big Data, cada 24 horas se genera información que es 70 veces el tamaño de la que se acumula en la biblioteca del Congreso de Estados Unidos.



La industria de los datos es la industria de este siglo, por algo Facebook vale en el mercado 200,000 millones de dólares.

En paralelo a lo anterior, el avance en la tecnología computacional ha permitido un crecimiento exponencial tanto en el volumen, como en la variedad y la complejidad de las bases de datos que se generan en la actualidad. La capacidad de generación-recolección de datos plantea problemas inéditos a la capacidad que existe para almacenarlos, validarlos, visualizarlos, analizarlos y, finalmente, transformarlos en conocimiento.

Sin embargo, el verdadero reto que plantea el análisis de datos en la actualidad no es de dimensionalidad. Es más bien un asunto relacionado con el tipo de preguntas industriales, comerciales, mercadotécnicas, tecnológicas y científicas que se tratan de responder utilizando como materia prima bases de datos masivas y de características complejas. En este contexto surge la ciencia basada en el uso intensivo de datos, o bien la llamada Ciencia de los Datos. En el sentido de Kuhn, surge como un nuevo paradigma de la metodología científica. Un paradigma de descubrimiento científico basado en el uso intensivo de datos más que en hipótesis derivadas de teorías. Esta ciencia se forma de la convergencia transdisciplinaria de la estadística, las matemáticas, la computación, y áreas de conocimiento de las ciencias sociales, las ciencias naturales y la ingeniería.

Como ciencia emergente, la Ciencia de los Datos está en sus primeros pasos, pero se dirige a convertirse en el paradigma del método científico del siglo XXI. De hecho ya hoy las ciencias (sociales y naturales) y las ingenierías, se han convertido en disciplinas basadas en uso intensivo de datos. Los descubrimientos científicos en climatología, geología, astronomía, biología, medicina, economía, finanzas, etc., están impulsados por datos, Hey et al. (2009). Esta nueva forma de hacer ciencia requiere de la colaboración de expertos en estas áreas con estadísticos y científicos de la computación. El reto que se plantea a éstos últimos está por un lado en la innovación de herramientas (computadoras y programas) y por otro en el desarrollo de nuevas técnicas de modelación matemática y algoritmos computacionales.

Nosotros, como estadísticos, debemos mencionar que la estadística, entendida como la ciencia de la extracción de conocimiento a partir de datos bajo ambientes de incertidumbre, es el instrumento de navegación de esta disciplina emergente: la Ciencia de Datos.

La revolución del uso intensivo de datos surge en la industria, los negocios, y la investigación científica. Sin embargo la educación también ha entrado a la revolución de los datos. El uso intensivo de datos es un asunto que los gestores y tomadores de decisión, llegando hasta los que diseñan las políticas públicas en educación (desde el nivel básico hasta el superior) no deben dejar pasar como una oportunidad para mejorar la calidad de la educación en sus procesos de enseñanza, aprendizaje e innovación.

En este trabajo presentamos algunas reflexiones sobre el potencial en la educación del uso intensivo de datos, en particular del *big data* el cual definiremos en la siguiente sección. Nos centramos en tres puntos: 1) Big data y Ciencias de Datos en la Academia, 2) Aprendizaje en línea y 3) Big Data para la mejora de la calidad de la educación. En la siguiente sección definimos el concepto de big data. En la Sección 3 abordamos cada uno de los puntos arriba mencionados y presentamos la experiencia de una universidad de Inglaterra con el big data.



¿Qué es Big Data?

Big Data es una evolución tecnológica producida por un lado por la capacidad actual para generar, almacenar y procesar datos de forma masiva y escalable y por otro lado por el acceso en tiempo real a estos datos gracias a dispositivos como computadoras, tabletas, y teléfonos móviles inteligentes.

Aunque aún no hay una definición formal, existe un consenso sobre lo que se entiende por Big Data. Berman (2013) caracteriza a Big Data por las tres V's: Volumen (grandes cantidades de datos), Variedad (diferentes tipos de datos) y Velocidad (los datos se acumulan a velocidad vertiginosa). El gigante de software estadístico SAS considera adicionalmente dos aspectos más, Variabilidad: los datos fluyen en períodos de tiempo muy variables; y Complejidad: los datos provienen de múltiples fuentes con o sin ninguna estructura formal además de tener correlaciones complejas junto con estructuras jerárquicas.

Big data no se debe confundir con bases de datos grandes. Big data se caracteriza además del tamaño, por la complejidad inherente de ellos. La matriz $n \times p$ que resulta de arreglar n objetos medidos en p variables, en general no representa big data. Big data en general viene sin estructura, es decir, datos no estructurados (matrices, árboles, grafos) y provenientes de diversas fuentes con diversos formatos. Por ejemplo, imágenes de Facebook, tweets, videos, y mensajes de texto. Ejemplos de big data son, véase también la Figura 1:

- Los datos que obtienen las agencias de publicidad y mercadotécnica que se dedican a monitorear redes sociales para entender cómo responde la gente a campañas publicitarias.
- Las grandes tiendas de autoservicio registran con detalle las compras de sus clientes para determinar asociaciones entre artículos comprados.
- El gobierno genera estadística oficial a diversos niveles desde el local hasta el nacional.
- Los hospitales generan y analizan datos de los registros clínicos de pacientes con diversos fines, por ejemplo para predecir que pacientes buscarán atención médica y además el tipo de atención que buscarán.
- Los equipos deportivos utilizan datos para predecir ventas de boletos.

El análisis de recursos de big data es un proceso mucho más complejo que el del análisis de matrices de datos. Sus principales etapas son extracción de datos, filtración de datos, transformación de datos y finalmente análisis el cual recurre a sofisticadas técnicas de la estadística, la inteligencia artificial, el reconocimiento de patrones, la minería de datos y el aprendizaje estadístico.





Figura 1. Blog anunciando las intenciones de Boeing para generar big data.

Fuente: <http://www.computerworlduk.com/news/data/boeing-787s-create-half-terabyte-of-data-per-flight-says-virgin-atlantic-3433595/>

Big Data dentro de la Academia

Como ya mencionamos, big data se refiere a conjuntos de datos masivos que son complejos y se recolectan a velocidades vertiginosas. Y aunque ha sido en la arena de la industria y los negocios donde se ha dado el mayor auge del big data, las universidades ya están buscando formas para adaptarse al uso del big data. Y es que caminar en cualquier campus de casi cualquier universidad permite ver a los estudiantes haciendo uso de celulares inteligentes, tabletas, computadoras portátiles, todos estos dispositivos tienen conectividad a internet, acceso a almacenamiento en nubes y uso de software gratis. Estas herramientas cambian las dinámicas dentro de las aulas y los laboratorios, la era de los datos está transformando la forma de como los estudiantes aprenden.

“El pensamiento estadístico será tan necesario para ser un ciudadano eficiente como la habilidad para leer y escribir” vaticinó acertadamente H.G. Wells hace ya poco más de un siglo. Hoy ya se habla del *Ciudadano Científico de Datos* para hacer referencia a un individuo competente en el procesamiento y análisis de datos o bien al profesional altamente especializado en la ciencia de datos. Las instituciones de educación superior deben ya adaptarse a este nuevo mundo ofreciendo a sus estudiantes las competencias para enfrentar exitosamente este nuevo paradigma del uso intensivo de datos. En este sentido y de acuerdo a Marsh, Maurovich-Horvat y Stevenson (2014), los retos de big data en educación son:

- Integración de recursos de big data así como enfocarse en la calidad y manejo de los datos.
- Big data depende de programadores y desarrolladores altamente competentes. Se necesitarán herramientas amigables para llevar a cabo proyectos de big data.
- Big data requiere de tecnología compleja lo cual restringe la disponibilidad de recursos disponibles para educación.



- Los proyectos de big data generalmente son en tiempo real, más que de datos históricos.
- Se requieren de procesos de validación estandarización de datos en los proyectos de big data.
- No hay un lenguaje común entre las diversas plataformas para realizar proyectos de big data.

Otra área relacionada con el big data en la educación es el aprendizaje en línea. Aunque esto no es nuevo, el avance tecnológico permite la transformación de la educación superior mediante esta modalidad de aprendizaje. Esto a la vez promueve una educación que cierra brechas, acorta distancias, y desaparece fronteras geopolíticas. Esta modalidad de educación ha dado lugar a lo que se conoce como *Cursos Abiertos Masivos en Línea*, aquí acuñamos el acrónimo CAMEL, (Massive Open Online Courses, MOOCs). Definitivamente, los CAMEL son la actualidad en educación superior e indudablemente están transformando la forma en que las universidades se administran y gestionan. Proporcionan otras dinámicas a los procesos de movilidad estudiantil, así como también permiten ahorrar la inversión en tiempo y costo que los estudiantes hacen en la universidad.

Dóri Sirály en <https://prezi.com/ycgyuqk9mmcm/manifiesto-mooc/> propone el siguiente manifiesto de los CAMEL:

- El aprendiz es el centro de cualquier diseño instructivo
- Considerar la comunidad de práctica y la comunidad de aprendizaje para las cuales se ha diseñado el CAMEL es beneficioso para el diseño de éste y para la propia institución que lo propone.
- Los CAMELs, como cualquier otra situación de aprendizaje, se definen en el diseño y en la acción
- Los CAMELs deben ser un elemento dentro de la estrategia digital de la institución que los propone, desde el CAMEL basado en contenidos hasta el CAMEL conectivista, pasando o no por un CAMEL basado en tareas, se relaciona con el desarrollo de la estrategia digital de la propia institución.
- En un CAMEL pueden surgir múltiples voces, tanto de los diseñadores como de los participantes en el CAMEL o de visitantes ocasionales, a partir de niveles diferentes de conocimiento experto.
- Los CAMELs pueden ser una “ruta de aprendizaje”, es decir, una secuencia de contenidos, tareas y pruebas, con o sin guía, para leer o ver simplemente o con alguna acción posterior en mente.
- Los CAMELs facilitan el debate pero este no tiene por qué ocurrir dónde establezcan los diseñadores, el lugar lo deciden los participantes. La procuración de contenidos es la forma de facilitar el acceso a la información a un número amplio de participantes.
- Los CAMELs ocurren en red pero esto no implica que no pueda haber encuentros de aprendizaje locales y cara a cara en paralelo.

Por otro lado, Dillenbourg et al. (2014) proponen *El Manifiesto Dagstuhl* sobre la Integración de los CAMELs en la educación universitaria, véase el reporte <http://drops.dagstuhl.de/opus/volltexte/2014/4786/pdf/dagman-v004-i001-p001-14112.pdf>

- Las universidades deberían de reforzar la educación residencial usando los CAMELs.
- En las universidades deberán de ser desarrollados los nuevos enfoques pedagógicos a fin de aprovechar la escala masiva.



- Las universidades deberían encaminarse hacia las prácticas de otorgar créditos oficiales para los CAMELs.
- Las universidades deberán desarrollar criterios de calidad para los CAMELs
- Las universidades deberán desarrollar normas estandarizadas para medir la participación en los CAMELs.
- El rendimiento de la inversión en los CAMELs depende en mucho de las universidades.
- Los CAMELs deberán proporcionar accesibilidad universal.
- Las nuevas políticas acerca del uso de los datos deberá ser negociada de manera colectiva.
- Big data se puede usar en admisión, presupuesto, y servicios a estudiantes. Asegurar transparencia, mejor distribución de recursos, identificación de estudiantes en riesgo.

Finalmente, construir recursos de big data con información de sus estudiantes permitiría a las universidades personalizar la experiencia educativa de cada uno de sus éstos, por ejemplo en los procesos de calificación y evaluación, en el aprendizaje personalizado y adaptativo y los procesos de tutoría. Es importante señalar aquí que con los cursos online y los CAMEL se pueden generar datos para analizarse y los cuales permitan diseñar secuencias de tópicos efectivas para un estudiante dado, acciones a tomar para mejorar desempeño de los estudiantes y detectar individuos en riesgo de deserción, determinar acciones que produzcan satisfacción y compromisos por parte de los estudiantes y determinar características del aprendizaje online que producen mejor aprendizaje de los estudiantes.

El Caso de la University College of London

La University College of London (UCL) es una universidad de mucho prestigio, la cual, según cierto ranking fidedigno ocupa el lugar número 7 a nivel mundial. Esta universidad tiene entre sus egresados o académicos a 29 Premios Nobel; también tiene a 3 recipientes de la prestigiada Medalla Fields (lo que se podría considerar el equivalente del Premio Nobel en Matemáticas). A UCL acuden estudiantes de 140 países del mundo. Está afiliada con la enorme empresa Elsevier, que maneja operaciones del orden de 500 mil millones de dólares. Elsevier es un proveedor líder mundial de soluciones de información que ayudan a tomar mejores decisiones, ofrecer una mejor atención, y frecuentemente sus entidades hacen descubrimientos innovadores en ciencia, salud y tecnología. Asimismo, se proporcionan soluciones digitales basadas en el internet dentro de las cuales se incluyen ScienceDirect, Scopus, Evolve, Knovel, Reaxys, y ClinicalKey. Por otro lado, en el terreno editorial, Elsevier publica más de 2500 revistas especializadas y tiene más de 33,000 títulos de libros publicados.

En el programa de desarrollo de UCL a 20 años (2014-2034), entre otros objetivos, se considera la creación de UGL (Global University of London) la cual, según el Presidente de Investigaciones de UCL, requerirá el apoyo del enfoque de Big Data para el desarrollo de sus programas de Docencia, de Investigación, y de servicios de diversa índole en áreas tales como las de Salud, Finanzas, Jurídica, etc. En diciembre del 2013, fue creado el Big Data Institute, en combinación con la UCL y con un apoyo substancial de Elsevier. El año pasado (mayo 2014), la UCL organizó un congreso en Londres sobre Big Data and Education. UCL tiene más de 200,000 egresados ubicados en distintas partes del mundo, con los cuales tiene seguimiento y



retroalimentación sobre sus actividades. Nuevamente, aquí el manejo con el enfoque Big Data ha jugado un importante papel.

Conclusiones

Big Data dentro de los Planes de Estudio: Hoy la tecnología nos permite capturar, organizar y analizar esta cantidad masiva de datos. Pero estos datos no son solo masivos, son también complejos y contienen muchos tipos de datos y además se generan rápidamente. Estos recursos son los que se conocen como big data. Nuevas disciplinas tales como Aprendizaje Analítico (*Analytics Learning*) y Minería de Datos Educativa están surgiendo. En estas disciplinas se aplican métodos de estadística, ciencia de la computación y aprendizaje máquina para identificar patrones y hacer predicciones. Las tecnologías computacionales y las tecnologías de la información han ocasionado que la ciencia se vuelva computacional y basada intensivamente en datos. Esto ha dado lugar al llamado *Cuarto Paradigma: Descubrimiento Científico Basado en Datos*, (véase Hey et al.(2009) y Critchlow y Van Dam (2013)). Los retos, bajo este paradigma, son capturar, analizar, modelar y visualizar información científica. Las instituciones de educación superior deben modificar los planes y programas de estudios de su oferta educativa tanto en pregrado como en posgrado para así poder preparar a profesionales competentes en el cuarto paradigma, es decir en el uso intensivo de datos y big data. Las universidades líderes en investigación deben considerar la creación de departamentos de Ciencia de Datos en donde formen profesionales especializados en los temas avanzados de la ciencia de datos. Una propuesta muy interesante en este sentido se puede ver en Cleveland (2014).

Aprendizaje en Línea: Big data está revolucionando el aprendizaje y los CAMELs son instrumentos que aprovechan las ventajas que ofrece la tecnología para implementar el aprendizaje en línea.

Proporcionar Educación de Calidad: Big data se ha usado en la industria y los negocios y ahora en la educación. Mediante la construcción de recursos de big data de sus estudiantes, las universidades pueden ayudar a confeccionar la educación a la medida de los estudiantes, una educación personalizada a las características y necesidades de los estudiantes.

Referencias

- Berman, J.J. (2013). Principles of Big Data. Preparing, Sharing, and Analyzing Complex Information. Amsterdam: Elsevier.
- Castells, M. (2006). La Era de la Información. Economía, Sociedad y Cultura. México D.F. Vols. I,II y III: Siglo Veintiuno Editores.
- Cleveland, W.S. (2014). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. Wiley Periodicals Inc. DOI: 100.1002/sam.11239.
- Critchlow, T., and Van Dam, K.K. (Eds.). (2013). Data-Intensive Science. Boca Raton: CRC Press.
- Dillenbourg, P., Fox, A., Kirchner, C., Mitchell, J., and Wirsing, M. (Eds.) (2014). Massive Open Online Courses: Current State and Perspectives. Dagstuhl Manifestos, Vol. 4, Issue 1, pp. 1-27.



Hey, T., Tansley, S., and Tolle, K. (Eds.). (2009). The Fourth Paradigm. Data-Intensive Scientific Discovery. USA: Microsoft Research.

Marsh, O., Maurovich-Horvat, L., and Stevenson, O. (2014). Big Data and Education: What 's the Big Idea? UCL Public Policy.

Semblanza

Ismael Sosa Galindo. Sus grados académicos obtenidos son: Licenciatura en Estadística, Maestría en Tecnología Educativa, Maestría en Estadística Aplicada, Doctorado en Ciencias en el Área de Alta Dirección, además tiene Especializaciones en Control de Calidad y en Métodos Estadísticos así como Diplomado en Habilidades del Pensamiento. Actualmente es catedrático de tiempo completo de la Universidad Veracruzana (UV) en la Facultad de Ciencias Administrativas y Sociales (FCAS) de las licenciaturas: Administración de Negocios Internacionales, Publicidad y Relaciones Públicas y en Relaciones Industriales. Donde imparte las experiencias educativas de: Administración de la Calidad, Estadística, Matemáticas Administrativas, Habilidades Matemáticas y Estadística Aplicada, anteriormente Bioestadística, Computación, Análisis Numérico, Administración Educativa, Matemáticas Financieras, entre otras durante más de 35 años. Sus áreas de interés son: Modelos Estadísticos, Matemáticas Aplicadas, Educación Superior y Administración de la Calidad. Anteriormente Técnico Académico e Investigador en el Instituto de Ciencias Básicas. Consejero Maestro Universitario, Consejero Técnico, Coordinador de Academias en campus de la Universidad Veracruzana en Xalapa, Coatzacoalcos y Veracruz Puerto. También ha impartido cátedra en licenciaturas y posgrados en otras Instituciones de Educación Superior. Tiene múltiples participaciones como ponencias, publicaciones, memorias, carteles en congresos y foros nacionales e internacionales, así como capítulos de libros, artículos técnicos y antologías, por citar algunas. Pertenece a la Sociedad Matemática Mexicana (SMM) y la Asociación Mexicana de Estadística (AME). Coordinador local del XVIII Congreso Nacional de Matemáticas en Xalapa (1989) con el Dr. José Carlos Gómez Larrañaga (presidente de la SMM (1989-1995). Coordinador del Primer Taller de sobre Procesos Estocásticos en el Instituto de Ciencias Básicas, U. V. con el Dr. Onésimo Hernández Lerma (CINVESTAV) y el Dr. Luis Cruz Kuri (UV). Ha desempeñado puestos en la administración pública tanto a Nivel Nacional Federal (subdirector de informática en la Dirección General de Información y Sistemas Forestales en la SARH), y Estatal Federal (delegado de Economía Agrícola en el estado de Veracruz). En el gobierno del estado de Veracruz fue Jefe de la Unidad Coordinadora de Información Agropecuaria, Forestal e Hidráulica del Programa Granero. isosa@uv.mx

Luis Cruz Kuri. Investigador titular en el Instituto de Ciencias Básicas de la Universidad Veracruzana y profesor de asignatura (actualmente en FCAS y en la Facultad de Estadística). Egresado como Matemático en la Facultad de Ciencias de la UNAM y con un doctorado (Ph.D.) con especialización en Estadística y Matemáticas, grado otorgado por la Universidad de California en Berkeley. Ha impartido cursos a niveles de licenciatura y postgrado en una variedad de instituciones locales, nacionales y del extranjero. Sus áreas de interés son: Modelos Estadísticos Multivariados y Teoría de la Probabilidad; también ha incursionado en aspectos varios de la Teoría de Juegos. Durante más de 12 años, responsable local de un proyecto conjunto NASA-UNAM-UV sobre "Detección de Patrones Espaciales y Temporales de Parámetros Climáticos y de Vegetación Alrededor de la Línea de Arboles del Pico de Orizaba". Ha participado como ponente en



congresos internacionales en varias ocasiones: en China, en Estados Unidos de Norteamérica, en Francia, Italia, Austria, España y México. Autor de varias publicaciones nacionales e internacionales en su disciplina. Se ha desempeñado como Director del ICB (1983-85); además ha participado como propietario del Honorable Consejo Técnico por el Área de Matemáticas y actualmente como Coordinador de la Academia de Matemáticas del ICB. Ha sido recipiente del *Premio al Decano*, otorgado por la Universidad Veracruzana en 2006, así como recipiente de la *Medalla al Mérito Académico*, también otorgada por la Universidad Veracruzana en el año de 1986. Cruz Kuri es ex becario del CONACYT para realización de estudios doctorales en el extranjero y regresó a México bajo el *Programa de Repatriaciones* del CONACYT. kruz2222@yahoo.com

