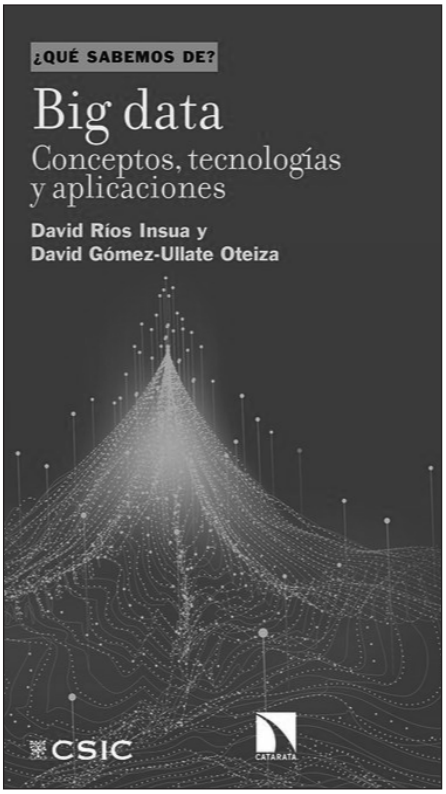


# *BIG DATA* CONCEPTOS, TECNOLOGÍAS Y APLICACIONES

David Ríos Insua y David Gómez-Ullate Oteiza
CSIC-CATARATA (2019)

Esta obra de David Ríos (AXA-ICMAT Chair en IC-MAT-CSIC y miembro de Real Academia de Ciencias Exactas, Físicas y Naturales), junto con David Gómez-Ullate Oteiza (investigador de la Universidad de Cádiz), aporta una visión sintética de lo qué es el Big Data, qué tecnologías se emplean para almacenarlo, procesarlo y qué resultados se están obteniendo actualmente en diferentes ámbitos de alcance nacional. El 29 de Septiembre de 2021 tuve la oportunidad de escuchar a David Ríos en su Discurso Inaugural del año académico 2021-2022 de la mencionada Real Academia, "Luces y sombras del Big Data y la Inteligencia Artificial" que extracta y comprime las ideas centrales de su libro y que desarrolla en el primer artículo del presente monográfico.

La capacidad técnica actual de capturar, recopilar y procesar todo tipo de datos (sociales, mercado, sensores y operaciones) nos ayuda a entender mejor cómo actuamos, cómo nos sentimos, cómo nos movemos e interactuamos y cómo respondemos frente a las políticas de los gobiernos y las decisiones de las empresas. Este aprovechamiento de la información contribuye a una mejora de las relaciones con los ciudadanos, los clientes, personalizando servicios o productos o automatizando todo tipo de procesos. También es cierto que se ponen de manifiesto algunas de las exageraciones establecidas sobre el Big Data como "el Diluvio de los Datos vuelve obsoleto el método científico".

El libro nos introduce en el fenómeno del Big Data identificando los pilares estadístico-matemáticos e informático-tecnológicos que sostienen este tipo de proyectos, además de revisar las tecnologías principales y su evolución. Se identifica claramente la necesidad de acompañar al almacenamiento del Big Data con su procesamiento, que aplicado al mundo de los negocios sería el *Business Analytics*. Se identifica también la diferencia entre el *Business Analytics* (aprovechamiento analítico de todo el Big Data para los negocios) con el *Business Intelligence* (empleo de la información corporativa altamente agregada de una compañía para consumo de su alta dirección para medir su rendimiento y asistir en la elaboración de planes estratégicos).

Las famosas V's del Big Data (Volumen, Velocidad y Variedad) son definidas correctamente, pero también se hace una mención especial a la cuarta V (Valor), sin duda la más relevante a la hora de acometer este tipo de inversiones. Se ofrece un correlato con las tecnologías que soportan el almacenamiento, el procesamiento y la visualización del Big Data.

En la exploración ofrecida de los métodos estadísticos y de aprendizaje automático se muestra el nuevo horizonte multidisciplinar de la *ciencia de datos*, campo en el que los autores son unos de los principales expertos del país. Se muestran las diferencias entre el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo, señalando algoritmos de ejemplo por cada categoría de aprendizaje. Los autores reflexionan sobre la relevancia del dilema sesgo-varianza expresado como una búsqueda de la complejidad óptima que debe tener el algoritmo para poder soslayar tanto su infraajuste como su sobreajuste. Más adelante, el libro se adentra en el Aprendizaje Profundo, que los autores consideran con razón, un auténtico *rebranding* científico (en realidad parte del Aprendizaje Automático). De nuevo se muestran los entornos de trabajo (*frameworks)* más empleados por la Industria y la Investigación con los que construir estos modelos, y algunos campos de aplicación como la visión artificial, el procesamiento del lenguaje natural o la inteligencia de negocio.

Una vez fijados los conceptos propedéuticos de esta disciplina, los autores nos muestra el empleo actual de la combinación del Big Data y la Inteligencia Artificial en tres grandes dominios: la promoción de las políticas públicas y la comunicación política, su empleo en la salud pública y, por último, la ciberseguridad.

En relación con la aplicación en Política y Administración se citan casos como la salud pública, las infraestructuras, el ahorro energético, la administración electrónica, o la gestión de ciudades inteligentes. También se cubren los aspectos relacionados con la mercadotecnia política relatando el caso de *Cambridge Analytica* analizando los datos de Facebook para poder mejorar el impacto de los mensajes a perfiles o colectivos seleccionados. Sorprende cómo la correcta selección de una muestra puede hacer que un mensaje tenga un impacto suficiente como para aumentar el éxito de una campaña electoral, o incluso de ganar unas elecciones, e.g. caso de *The Cave* en las elecciones de Obama de 2002. Más adelante se realiza un análisis de un problema relacionado con el anterior: la propagación de bulos en redes sociales, y se relata detalladamente el caso Trump en las elecciones de 2008. Estas técnicas no son recientes, sino que parten de la Economía del Comportamiento del premio Nobel Thaler, en la que se trata de dirigir las decisiones de compra de los consumidores, incluyendo casos en los que se pueda poner en riesgo la salud pública. Ante esta amenaza se plantean nuevos métodos para detectar estos bulos: *aproximaciones lingüísticas* que intentan localizar señales de engaño, y *aproximaciones en red* que clasifican las fuentes en creíbles o engañosas.

La salud es uno de los campos donde la aplicación del Big Data va a tener más impacto en un futuro cercano, tanto apoyando la diagnosis (determinar adecuadamente la enfermedad de un paciente dados unos síntomas y un historial clínico) como la prognosis (estimar el comportamiento futuro empleando observaciones actuales de la evolución del paciente). En ambos casos, se trata de problemas con razonamiento bajo incertidumbre para lo cual los autores plantean el empleo de la inferencia bayesiana que es más rigurosa en el tratamiento de la incertidumbre que los métodos tradicionales de corte frecuentista o de máxima verosimilitud. También se ahonda en la aplicación del aprendizaje profundo si se dispone de suficientes casos como para entrenar el modelo de manera satisfactorio. En este caso se podría llegar a mejorar incluso la capacidad diagnóstica humana. Es muy relevante la distinción que se establece entre predicción y descripción. Los modelos de aprendizaje profundo presentan una altísima precisión en sus capacidades de predicción, pero como *caja negra* que son, no ofrecen descripción alguna de las causas (síntomas, dieta, meteorología, etc.) que concurren en un efecto determinado (enfermedad). También se detalla el empleo de tecnologías como el Procesamiento del Lenguaje Natural en la creación de asistentes virtuales que ayuden a los profesionales de la salud en la búsqueda de información o en la formación de un diagnóstico; también se hace referencia al empleo del aprendizaje profundo para reconocimiento de imágenes. Toda esta algoritmia necesita vastas cantidades de información para poder entrenarse de manera estable, y es por ello que se añade un análisis pormenorizado de fuentes resultado de la transformación digital del sector como son los sensores, las redes sociales o la información de la sanidad pública. Son reseñables algunos dispositivos como los *ponibles* (*wearables*), prendas u otros adminículos que puede llevar encima un usuario y transmitir todo tipo de información biométrica o de comportamiento. Llamativo es el caso del análisis de información producida en tiempo real en redes sociales (*nowcasting)* que permite crear predicciones sobre epidemias o pandemias (recordemos el caso de Google Flu Trends).

El tercer caso de aplicación es la ciberseguridad. Los ciberataques son cada día más frecuentes y dañinos. Al transformarse digitalmente la sociedad, los cibersistemas comienzan a proliferar suponiendo que un agente externo pueda tomar el control no ya de un sistema de información, sino de un sistema de operación, e.g. un vehículo autónomo, el sistema de gestión de una central nuclear o un sistema de control de tráfico aéreo. Se estudia la *darknet* como un espacio no vigilable empleado para actividades de tráfico de armas, drogas o pedopornografía

(aunque también para emisión de noticias de manera libre dentro de un país no democrático) al igual que el impacto del *hacktivismo* (amenazas graves a la seguridad pública que bordean el terrorismo) o el *malware* que tiene como objetivo atacar un sistema de información. Analizados los problemas se muestran las actuales metodologías existentes que buscan una solución pero que están basadas en matrices de riesgos que puedan generar asignaciones de riesgos inadecuadas. Como aproximación de solución se plantea la monitorización del Big Data que circula por las redes de información y los sistemas de gestión de incidencias y eventos (SIEM) que monitorizan quién entra y qué hace dentro de una red. Dada esta infraestructura preexistente, los autores describen una mejora de estos sistemas con la implantación de un sistema basado en *Análisis de Riesgos Adversarios (ARA)*. Estos modelos de aprendizaje automático adversario se basan en que un adversario pueda engañar al modelo clasificador para obtener un beneficio. Un ejemplo interesante es que un adversario puede engañar al sistema de guiado de un vehículo autónomo modificando algunos píxeles de la imagen capturada del entorno para conseguir que el sistema guía "no vea" y el vehículo colisione. Estos sistemas basados en ARA superan la restricción de disponer de "conocimiento común" entre todos los jugadores (habitual en Teoría de Juegos No Cooperativa), lo cual se considera realista en este escenario en el que existe un incentivo para esconder información a los otros participantes.

La referencia a la cuestión ética del uso de la IA es insoslayable discurriendo sobre la responsabilidad del tratamiento del *Big Data.* Comenzando con la descripción de las *cookies* y su empleo por la mercadotecnia de precisión para aumentar la respuesta de campañas, se muestra cómo las empresas recurren a artificios que recopilan información personal del usuario (es apabullante lo que puede saberse de nosotros sólo con nuestra dirección IP) ofreciendo servicios gratuitos para poder obtener nuestros datos ("cuando Vd. no paga por los productos/servicios, Vd. es el producto"). Inquieta el caso del autómata aspirador *Roomba* y su capacidad de dibujar planos detallados de viviendas. En cuanto a la publicidad se desarrolla la idea de que construyendo "modelos de atribución" se obtienen rasgos característicos que definen un público objetivo óptimo con el que optimizar la respuesta de campañas promocionales. Del marketing se continua al ámbito público donde se relata el caso Snowden y el programa PRISM que recopilaba datos personales de usuarios de las grandes empresas tecnológicas. Hace reflexionar sobre el empleo de datos de tránsito y otros de comportamiento social como los que elabora el programa del *crédito social* del gobierno chino, país donde se encuentran el 42% de las cámaras de circuito cerrado de TV. La unión de la información captada por las cámaras transportada en tiempo de latencia muy bajo por sus redes 5G y analizada con aprendizaje profundo de reconocimiento visual parece una auténtica realización de la pesadilla de Orwell. Finalmente, tras plantear un estudio de los sesgos cognitivos que pueden impedir que el juicio de un autómata sea objetivo, suponiendo claras discriminaciones, se recomienda la aplicación de algunas soluciones algebraicas sencillas que pueden evitarlo. La reflexión final es la necesidad de articular reglamentos de protección de datos personales como el europeo (Reglamento General de Protección de Datos) que establecen un marco garantista al respecto.

Como conclusión la obra plantea las siguientes reflexiones:

- El análisis del Big Data ha sido sobrevendido por algunas empresas tecnológicas
- Se debe efectuar un claro análisis coste-beneficio de la complejidad de una solución analítica
- Es necesario disponer de Big Data, pero también de una correcta estrategia e infraestructura de cómo explotarlo, es decir, de una capacidad de Ciencia de Datos
- De las famosas V's del Big Data, los autores se quedan con la más relevante: el valor que se pueda derivar de su análisis y que justifique las inversiones de su infraestructura
- Los métodos bayesianos ofrecen un mayor rigor en el tratamiento de problemas bajo incertidumbre que los métodos frecuentistas, sin embargo, aún no escalan adecuadamente con Big Data
- Desde el ángulo ético se debe promocionar una mayor concienciación a nivel social del valor de los datos y una mejor regulación de los aspectos de privacidad de datos

En su conferencia en la Academia el autor planteó dos iniciativas a considerar, ambas relativas a la mejora del posicionamiento de nuestro país en Ciencia de Datos. La primera, es la reforma educativa de los estudios de Matemáticas y, en especial, los de Estadística y Probabilidad; campos que suelen quedar al final de los programas educativos y que no acaban de cubrirse en su totalidad. La segunda, la creación de un Instituto Nacional de Ciencia de Datos que tuviese como objetivo la mejora de la toma de decisiones públicas, y el desarrollo de fundamentos matemáticos de las metodologías desarrolladas, y así mejorar la disciplina.

**Wolfram Rozas**