Big Data y Machine Learning: ¿futuro o presente en las UCIs?

Pedro Rascado Sedes
Facultativo Especialista de Área de Medicina Intensiva. Servicio de Medicina Intensiva
Xerencia de Xestión Integrada de Santiago de Compostela. A Coruña. España
e-mail: pedrorascado@hotmail.com

El objetivo de este artículo es hacer una revisión sobre el concepto de *Big Data* e inteligencia artificial aplicada a la sanidad, haciendo hincapié en lo relacionado con el aprendizaje automático o *Machine Learning*.

La utilización del término *Big Data* es relativamente reciente. La mayoría de los autores aceptan que la primera vez que se usa Big Data es en un artículo publicado en 1997 por investigadores de la NASA (Michael Cox y David Ellsworth) titulado *Application-Controlled Demand Paging for Out-of-Core Visualization*. En él hacen referencia a la dificultad para manejar y almacenar grandes volúmenes de datos. Los mismos autores publican en 1999 el que se considera por otros la primera publicación académica sobre *Big Data: Visually Exploring Gigabyte data sets in real time*.

Aunque éste se puede considerar el inicio del concepto, es antiguo el interés, sobre todo a nivel empresarial, por un análisis de tipo predictivo que busca la extracción de conocimiento de los datos, en forma de patrones, tendencias o modelos que permitan una cierta certeza sobre el resultado de potenciales acciones futuras. Para denominar este tipo de análisis, a finales de los 80 surge la expresión Minería de Datos (*Data Mining*), que podemos considerar el antecedente histórico del actual *Big Data*.

El desarrollo del *Big Data* discurre paralelo al crecimiento y la popularización de internet (*World Wide Web*) que ha generado un volu-

men de datos imposibles de procesar con las herramientas y técnicas tradicionales y un interés de las empresas por maximizar beneficios analizando esa información.

Desde el punto de vista sanitario, la generalización de la historia clínica electrónica ha supuesto el almacenamiento de una cantidad ingente de variables que podrían ser explotadas con el objetivo del generar nuevo conocimiento. Dentro del sistema sanitario, la Unidad de Cuidados Intensivos (UCI) se puede considerar un caso especialmente paradigmático de uso de *Big Data* para mejorar el cuidado de los pacientes. Por una parte, la generación de manera continua de gran cantidad de información. Además de la incluida tradicionalmente en las historias clínicas electrónicas, alguna más específica: monitores de cabecera, respiradores, bombas de medicación, equipos de hemofiltración.... Sin olvidarnos de los datos no siempre disponibles, pero cada vez más frecuentes, resultado de las -ómicas (genómica, transcriptómica, proteómica). Por otro lado, la complejidad de la atención en el paciente crítico hace demasiado simplista la expectativa de que una única intervención pueda modificar el pronóstico lo que provoca más atractiva, si cabe, la aproximación mediante *Big Data*.

Podemos medir la utilización de *Big Data* en ciencias de la salud por el número de publicaciones indexadas en Pubmed que hacen referencia al término. En el año 2008 la revista *Nature* dedica un número monográfico al tema, pero no es hasta los últimos años en los que se ve un crecimiento exponencial (Figura 1),

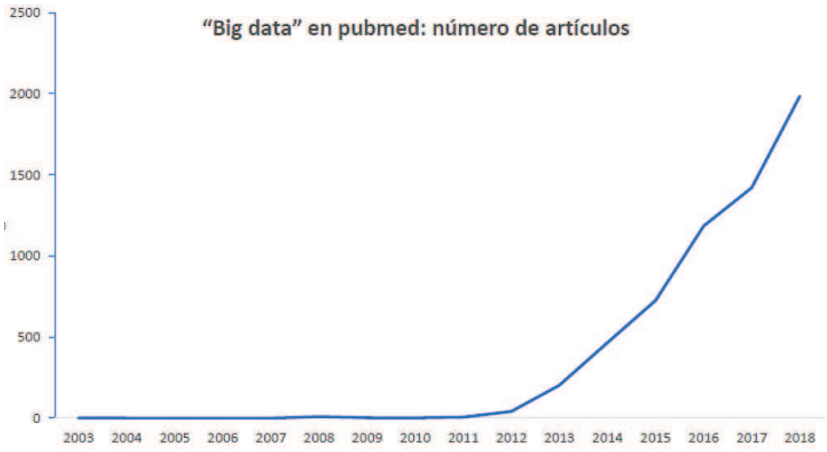


Figura 1. *Big Data* en Pub Med

con casi 2000 referencias los 11 primeros meses de 2018 frente a las poco más de 1400 en 2017, o las 42 de 2013.

Aunque puede resultar intuitivo el concepto de *Big Data*, no hay una definición aceptada de manera general. Algunas definiciones se centran en responder a la pregunta ¿qué es? y otras, sin embargo, dan respuesta a ¿qué hace? Mayoritariamente, las definiciones incluyen de una u otra manera la 3V´s (Volumen, Variedad y Velocidad) propuestas por Doug Laney en 2001 como las tres dimensiones a considerar en el manejo de los datos, de las que hablaremos más adelante.

Así nos encontramos con las siguientes definiciones:

"*Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*"

Garner Inc.

"*Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.*"

TechAmerica Foundation's Federal Big Data Commission

"*Big data es un término común bajo el que se agrupan toda clase de técnicas de tratamiento de grandes volúmenes de datos, fuera de los análisis y herramientas clásicas. Este concepto engloba muchas ideas y aproximaciones, pero todas con un objetivo común: extraer información de valor de los datos, de forma que pueda ser de ayuda para las decisiones y procesos de negocio*"

Instituto de Ingeniería del conocimiento

Describiremos ahora más profundamente las 3V´s que definen *Big Data*:

-**Volumen**: Quizá la primera magnitud que condicionó un cambio de concepto el manejo de datos. Big Data se define por el enorme volumen de datos, que aumenta además de manera exponencial. Aunque no hay un tamaño de referencia se puede hablar de *Big Data* a partir de un petabyte (1015 bytes, aproximadamente mil millones de fotos). La tecnología para guardar y procesar ha avanzado paralelamente por lo que el mayor problema no es ahora el tamaño de los datos.

-**Variedad**: La variedad se refiere a los tipos diferentes de datos. Estos pueden ser estructurados, no estructurados o semiestructurados.
Los datos estructurados son aquellos en los que se conoce el número y el tamaño de los campos. Forman las bases de datos clásicas (SPSS, Excel,...), sin embargo, representan no más del 10% de los datos incluidos en los análisis de *Big Data*.
Los no estructurados se presentan en el formato tal y como fueron recolectados, carecen de una estructura conocida de antemano y no pueden ser almacenados en tablas. Pertenecen a estos los documentos de texto, pdf, fotografías, e-mail, etc... Uno de los grandes avances en las técnicas de *Big Data* es el desarrollo de la tecnología para analizar estos datos no estructurados. Refiriéndonos a las aplicaciones sanitarias, la tecnología actual permite el reconocimiento de voz y la incorporación de textos escritos no estructurados al análisis. Conocidas como técnicas de procesamiento del lenguaje natural, su desarrollo ha permitido que se incorporen en los algoritmos de análisis de *Big Data* todos los textos de escritura libre que componen la historia clínica.
Los datos semiestructurados son aquellos que no se limitan a campos determinados, pero que contienen marcadores para separar los distintos elementos. Por ejemplo, HTML o XML (lenguajes utilizados en la web).
-**Velocidad**: Se refiere a la velocidad a la que se generan los datos y consecuentemente a la velocidad que deberían ser analizados. Se estima que en 2011 el tamaño global de datos generados relacionados con los cuidados sanitarios fue de 150 exabytes (1018).
Aunque no incluidas en las definiciones clásicas de *Big Data*, la mayoría de los autores incluyen al menos una cuarta, y hasta 7 V´s en el concepto.
-**Veracidad**: Se refiere al sesgo, al ruido (datos no relevantes o erróneos) y a la alteración de los datos. De especial interés en el caso de las aplicaciones sanitarias que utilizan datos de monitorización no validados. Existen técnicas que eliminan el ruido de los datos, pero conviene utilizar fuentes fidedignas para evitar conclusiones inadecuadas.
-**Valor**: Hablamos del valor que generan los datos una vez analizados y convertidos en información. La mejor manera de generar valor es realizar la pregunta correcta y evitar análisis ciegos.

Aportaciones de *Big Data*

Las aportaciones de *Big Data* se relacionan con la elaboración de conclusiones en base al análisis de datos. Entroncan directamente con los conceptos de Inteligencia Artificial y *Machine Learning* (Aprendizaje automático).

Entendemos por inteligencia artificial la capacidad de las máquinas para realizar comportamientos que puedan parecer humanos. Estamos muy lejos de conseguir máquinas inteligentes, considerando la inteligencia global, pero sí se ha desarrollado software y hardware para realizar actividades muy concretas. Por ejemplo: reconocimiento de voz, o análisis de imágenes que permite reconocer y clasificar objetos. Más cercano a nuestro entorno por ejemplo analizar pruebas radiológicas, sin intervención humana.

El *machine learning* es una forma de inteligencia artificial en la que las máquinas aprenden de manera automática.

Analítica descriptiva

Consiste en la descripción objetiva de la población de estudio. Es frecuente buscar correlaciones para crear grupos homogéneos con características similares. Se puede utilizar para seleccionar pacientes que responden de mejor manera a un tratamiento determinado o aquellos que tienen mayor a menor mortalidad.

Figura 2

En este punto es importante considerar la diferencia con la estadística clásica. En ésta, se selecciona una muestra de la población, se es-

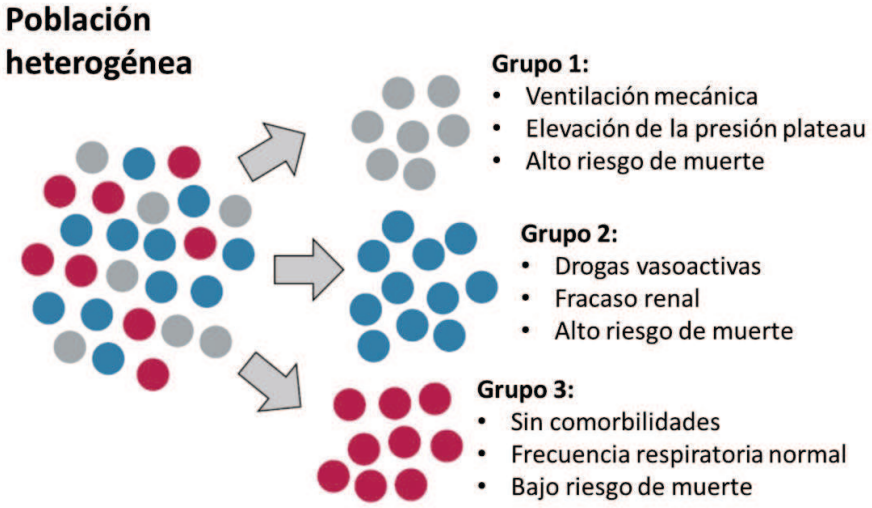


Figura 2. Analítica descriptiva. Tomado de: Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. Chest. 2018;154:1239-48.

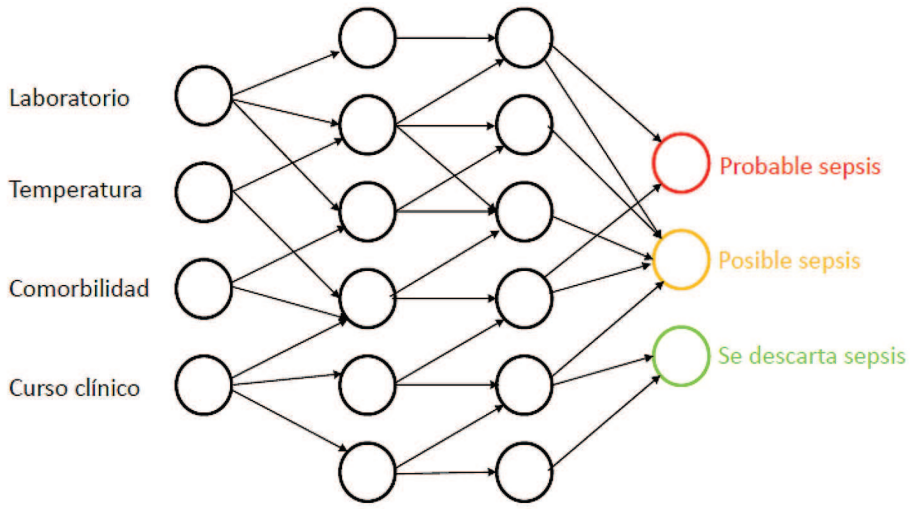
tudia, se infieren conclusiones que posteriormente se extrapolan a la población global. En *Big Data* se analiza toda la población en búsqueda de correlaciones. Es necesario analizar posteriormente la viabilidad de las correlaciones y valorar la causalidad.

Analítica predictiva

Se construye sobre la analítica descriptiva y usa modelos estadísticos avanzados para añadir a nuestra base de información datos que no conocemos.

Los ejemplos más habituales son los que desarrollan algoritmos que predicen la probabilidad de muerte o que analizan variables que detectan el riesgo de desarrollar una sepsis o de precisar ingreso en UCI.

El *machine learning*, con la utilización de redes neuronales artificiales es una de las formas más desarrolladas de realizar predicciones de manera automática (Figura 3). Las redes neuronales pretenden simular el cerebro humano. Consiste en varias capas de neuronas artificiales interconectadas. Las conexiones son ciegas para el usurario, que sólo conoce las entradas y salidas. Cuando el algoritmo presenta múltiples capas de neuronas se conoce como *Deep Learning*. Las neuronas de cada capa están unidas virtualmente con las neuronas de la capa siguiente mediante conexiones que tienen un peso determinado. Este peso no es más que un número que se utiliza para realizar operaciones matemáticas (funda-



3. Red neuronal artificial.

mentalmente modelos de regresión lineal modificados) que en función del resultado "activan" en mayor o menor medida las neuronas de la capa siguiente. El proceso sigue por las diferentes capas hasta que da como resultado la "activación" de alguna de las neuronas de la capa de salida. El proceso de aprendizaje se realiza sobre un conjunto de datos con resultados conocidos (lo que se conoce como aprendizaje supervisado). De manera automática la máquina modifica los pesos de las interconexiones para disminuir el error. Una vez aprendido se pueden realizar las predicciones sobre nuevos datos.

Big Data en la práctica clínica

En el entorno de cuidados intensivos lo más desarrollado son los modelos de aprendizaje supervisado para predecir el pronóstico de pacientes ingresados en la UCI y para el diagnóstico precoz de la sepsis.

Se ha demostrado analizando bases de datos que los modelos basados en *machine learning* predicen mejor el pronóstico que los clásicos modelos de regresión.

Horng y col. demostraron que es posible predecir la sepsis en triaje aplicando técnicas de aprendizaje automático. El modelo se ensayó en una base de datos encontrando que el modelo que mejor detecta la sepsis en triaje es aquel que incluye entre las variables el análisis de texto libre (AUC 0,85).

Más recientemente, Mao y col. han desarrollado un algoritmo basado exclusivamente en signos vitales que predice el desarrollo de shock séptico 4 horas antes de que se produzca con un AUC de 0,95.

Por el momento, no se han publicado soluciones basadas en inteligencia artificial que hayan demostrado prospectivamente modificar el pronóstico de pacientes.

En el futuro, grandes cantidades de datos clínicos, de laboratorio, variables fisiológicas, -ómicas, etc... serán estudiadas por sistemas inteligentes que las analizarán y presentarán las conclusiones al clínico para facilitar la toma de decisiones. Modelos pronósticos y de detección precoz añadirán valor a la atención. La colaboración con otras disciplinas ajenas a la medicina: ingeniería, informática, análisis de datos..., nos permitirá el aprovechamiento eficiente de todos los datos que se generan diariamente.

Big data y *Machine Learning* están aquí y han venido para quedarse.

Más información en:

Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015;35:137–44.

Sanchez-Pinto LN, Luo Y, Churpek MM. Big Data and Data Science in Critical Care. Chest. 2018;154:1239-48.

Horng S, Sontag DA, Halpern Y, et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One. 2017;12:e0174708.

Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8:e017833.

Álvaro Barbero. La inteligencia artificial son los padres. Instituto de Ingeniería del Conocimiento. [Video] Junio 2018(Acceso en Diciembre 2018). Disponible en: https://www.youtube.com/watch?v=XCVEK2HXkFU

Los autores de este artículo declaran no tener conflicto de intereses