El camino desde la Inteligencia Artificial al *Big Data*

Antonio Berlanga
Doctor en Ingeniería Informática.
Profesor Titular de Ciencias de la Computación e Inteligencia Artificial.
Dpto. de Informática de la Universidad Carlos III de Madrid.

Un recorrido por el desarrollo de la Inteligencia Artificial y sus paradigmas, en especial el Aprendizaje Automático, que va a proveer de muchos de los algoritmos que se aplican en el tratamiento automático de grandes volúmenes de datos no estructurados, conocido con el término de "*Big Data*".

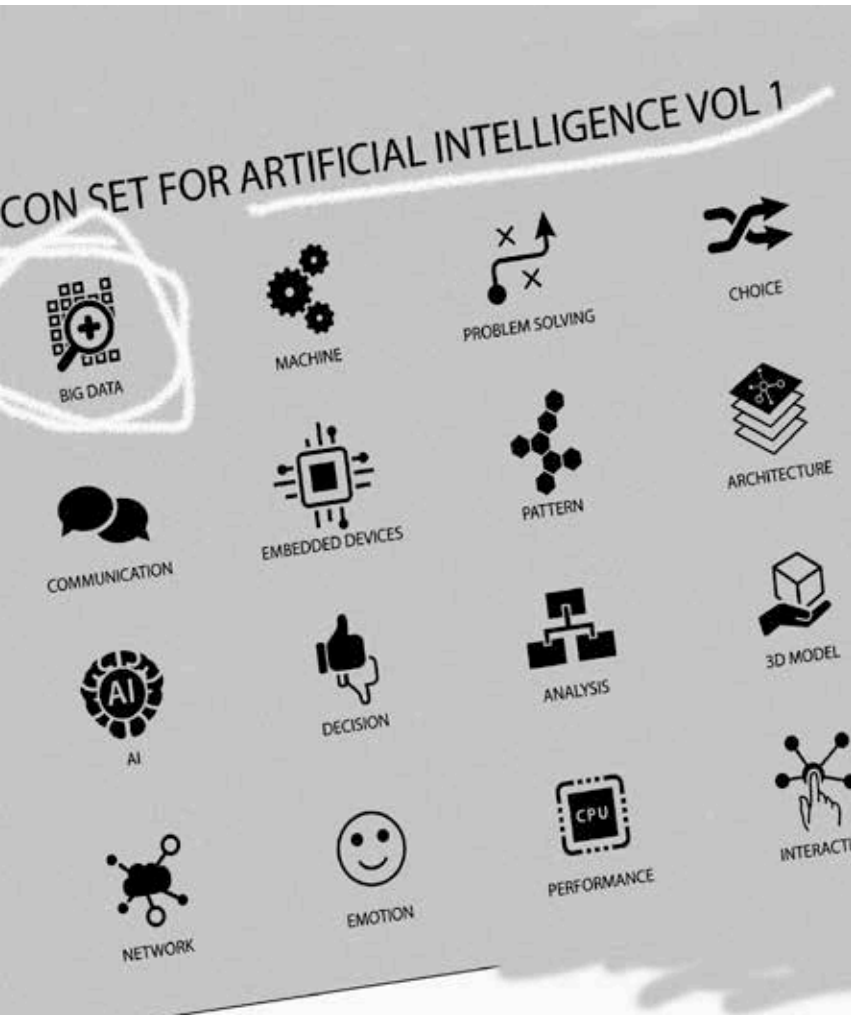
Se cumplen 60 años desde que John McCarthy acuñó el término "Inteligencia Artificial"[1]. Fue durante la reunión de 10 jóvenes investigadores, organizada por el propio McCarthy en Dartmouth, interesados en estudiar los aspectos de la inteligencia que podían simularse algorítmicamente[2]. Se anunció a su término, nada más y nada menos, el nacimiento de una nueva ciencia. El perfil académico y la carrera profesional de sus creadores marcaron su desarrollo posterior que alcanza el presente. Física, biología, economía, psicología cognitiva y en especial matemática e ingeniería, fueron las ramas de conocimiento embrionarias que han desarrollado una, todavía aún más interdisciplinaria ciencia. En las décadas siguientes, esta nueva disciplina experimentó un crecimiento extraordinario, fruto del interés que suscitó tanto académico como empresarial. Se intuían grandes posibilidades de aplicación y muy pronto también, se apreciaron sus limitaciones.

Uno de los criterios por los que se clasifican las técnicas de Inteligencia Artificial (IA) obedece a la forma cómo se construyen. Así, se distinguen entre las técnicas que utilizan una aproximación "*top-down*" de las que utilizan un esquema "*bottom-up*". Las técnicas "*top-down*" se basan en la llamada "*The physical symbol system hypothesis*"[3] que postula que un sistema físico puede realizar acciones inteligentes si está dotado de la apropiada representación simbólica de conocimiento. Mediante procesos, esta representación puede combinarse para obtenerse estructuras complejas de símbolos, estando estos procesos representados asimismo mediante símbolos. Este enfoque dio lugar a las aproximaciones basadas en la lógica y sistemas de reglas que tuvo en los sistemas expertos su máximo exponente[4]. Un sistema experto requiere de la incorporación del conocimiento que tienen especialistas en el dominio del problema que se va a resolver y de un sistema que automáticamente pueda modificar y ampliar este conocimiento. Los primeros inconvenientes para realizar una aplicación práctica de los sistemas expertos surgen de esos requerimientos. Por un lado, la forma de representar el conocimiento introducirá un sesgo sobre el alcance del razonamiento; por otro, los especialistas deben aportar conocimiento completo y libre de errores, ya que el sistema será incapaz de corregirlos salvo en casos de ambigüedades triviales. Otro problema surge de la dificultad para evaluar el desempeño de los sistemas expertos[5]. Pero el problema crítico se encuentra en la construcción del algoritmo que permite extraer nuevo conocimiento. El sistema experto razona con conocimiento de alto nivel, pero generalmente opera con datos de muy bajo nivel, a menudo con poca o ninguna estructura. Hacer la transición de los datos al cono-

> *Cuando las técnicas que se aplican sobre los datos proceden de la estadística cuantitativa y cualitativa clásica, se usa el término* "Data Analytics", *reservando* "Big Data" *para cuando se utilizan técnicas inductivas, las propias del Aprendizaje Automático.*

cimiento es una tarea sin cerrar y objeto de interés por parte de la comunidad de la inteligencia artificial simbólica y otras áreas de conocimiento afines. Este enfoque dominó la inteligencia artificial desde los años 50 hasta finales de los 80 del siglo pasado. Se desarrollaron muchos sistemas expertos, aplicándose a entornos donde la información está bien estructurada y el conocimiento puede organizarse, de forma natural, con sistemas de reglas, como por ejemplo en el área del diagnóstico médico, en ingeniería para el control y test de sistemas automáticos, en economía para la detección de fraude y cálculo de riesgo crediticio, etc.[6-10].

En el declive de la Inteligencia Artificial simbólica participó la emergencia del enfoque "bottom-up". La idea de modelar el razonamiento por deducción, similar al humano, es reemplazado por la imitación de pequeños comportamientos inteligentes, obtenidos por inducción, de los que pueden emerger comportamientos cada vez más complejos. Se rescata un campo de investigación formulado a finales de los años 60, el Aprendizaje Automático[11], que estudia la construcción de algoritmos que pueden extraer conocimiento a partir de conjuntos de datos. El resultado de un algoritmo de Aprendizaje Automático va a ser un modelo que explica las regularidades existentes en los datos. Los árboles y reglas de decisión, redes bayesianas, sistemas ocultos de Markov, clasificadores lineales, algoritmos evolutivos, máquinas de soporte vectorial, redes de neuronas artificiales son algunos de los algoritmos investigados y aplicados desde el Aprendizaje Automático. Todos comparten la característica de contar con métricas que cuantifican la bondad del modelo que construyen, ya sea para clasificar, agrupar o predecir. En función de esa medida reajustan los parámetros del modelo con el fin de optimizar su desempeño. Un elemento clave es el requisito de obtener modelos generales. Un modelo general inducido a partir de un conjunto de datos permite ser aplicado sobre otros datos distintos. Si tomamos como ejemplo la construcción de un sistema de reconocimiento de caracteres manuscritos, el modelo deberá reconocer otros caracteres en textos que no han sido utilizados para su construcción. Por tanto, se presenta la cuestión acerca de cómo debe ser el conjunto de datos, tanto en tamaño y características. Siguiendo con el ejemplo del reconocedor de caracteres, utilizar textos en castellano sesgará el modelo, reconociendo con diferente acierto los distintos caracteres, ya que hay letras con una frecuencia de aparición muy alta frente a otras de muy baja. Respecto del tamaño de la muestra, se observó que una técnica pobre con un conjunto de datos grande tendría mejor comportamiento que una buena técnica con pocos datos[12].



A finales de los años 80 no se disponía de grandes recursos para computación y el almacenamiento de volúmenes masivos de datos a bajo coste, lo que en gran parte (también aparecieron problemas formales en algunas técnicas que posteriormente se han superado) hizo que las técnicas de Aprendizaje Automático no se popularizasen más allá del entorno académico. Simultáneamente, se desarrollaron metodologías para poder extraer conocimiento de las bases de datos. El estudio de esas metodologías adoptó el nombre de Descubrimiento de Conocimiento en Bases de Datos, o más conocido por sus siglas en inglés KDD *(Knowledge Discovery in Databases)*. Un paso en las metodologías implicaba la aplicación de técnicas que revelasen estructuras y relaciones entre los datos, fue llamado Minería de Datos[13]. La Minería de Datos comparte una gran cantidad de técnicas con el Aprendizaje Automático, a tal forma que resulta, hoy día, difícil realizar una distinción clara entre ambas. El consenso actual es considerar al Aprendizaje Automático como el es-

tudio de técnicas, con las características que han sido mencionadas anteriormente, que pueden incorporarse a un proceso de Minería de Datos[14-15]. Durante las décadas siguientes, la Minería de Datos fue incorporándose, primero desplazando, en las áreas en las que se habían aplicado, a los sistemas expertos y, posteriormente con la explosión en los 90 de internet y los sistemas de información, a una gran variedad de actividades[16].

En 1997, se publica el primer artículo en una conferencia internacional en el que se realiza una definición del término "*Big Data*"[17]. Hace referencia al problema de tener que procesar un conjunto de datos con un tamaño superior a la memoria del ordenador y al del almacenamiento en disco, incluso si este es remoto. Empiezan a emerger situaciones en las que la acumulación de datos es tan grande que es necesario definir nuevos procedimientos para poder aplicar las técnicas de extracción de conocimiento. Pocos años después, en 2001 , se realiza una definición para "*Big Data*" que ha sido ampliamente aceptada[18], aunque ampliada y revisada con posterioridad. Conocida como las "3Vs", hace referencia a las características de volumen, velocidad y variedad de los datos. Cuando las técnicas que se aplican sobre los datos proceden de la estadística cuantitativa y cualitativa clásica, se usa el término "*Data Analytics*", reservando "*Big Data*" para cuando se utilizan técnicas inductivas, las propias del Aprendizaje Automático.

Es a partir de 2010 cuando el interés acerca del "*Big Data*" crece de forma exponencial[19]. El coste de almacenamiento ha caído en 10 años casi a su milésima parte y el de la computación a la centésima. Han surgido nuevos conceptos de aplicación que hacen un uso masivo de datos no estructurados; "*Smart Cities*", "Internet de la cosas", "*Smart Health*", "Industria 4.0" son solo algunos ejemplos[20-22]. Las grandes corporaciones internacionales anuncian aplicaciones y servicios basados en "*Big Data*"; cualquier usuario de redes sociales, sin saberlo, está haciendo uso de las facilidades y recursos que proporcionan estas técnicas.

El "*Big Data*" está llamado a revolucionar el mundo como lo hizo Internet, tendrá que evolucionar, como lo hizo la red de redes. Hay muchos desafíos a futuro[23], uno muy importante será el de crear especialistas en este campo con una formación académica híbrida entre la estadística, las ciencias de la computación, los sistemas de información, la computación de altas prestaciones; en definitiva multidisciplinar, tal como lo fue en su origen la Inteligencia Artificial.

Referencias

1. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach. Third edition*. 2014.
2. R. R. Kline, "Cybernetics, automata studies, and the dartmouth conference on artificial intelligence," *IEEE Ann. Hist. Comput.*, vol. 33, no. 4, pp. 5–16, 2011.
3. A. Newell, "Physical Symbol Systems*," *Cogn. Sci.*, vol. 4, no. 2, pp. 135–183, Apr. 1980.
4. P. V. S. Ponnapalli, "Introduction to expert systems. 2nd edn, P. Jackson, Addison-Wesley, Wokingham, 1990, ISBN 0-201-17578-9, xi + 526pp. £21.95," *Int. J. Adapt. Control Signal Process.*, vol. 6, no. 1, pp. 65–67, Jan. 1992.
5. J. Biegel and U. G. Gupta, "Expert systems in manufacturing: Promises and perils," *Comput. Ind. Eng.*, vol. 19, no. 1–4, pp. 127–130, Jan. 1990.
6. P. SZOLOVITS, R. S. PATIL, and W. B. SCHWARTZ, "Artificial Intelligence in Medical Diagnosis," *Ann. Intern. Med.*, vol. 108, no. 1, p. 80, Jan. 1988.
7. Z. Z. Zhang, G. S. Hope, and O. P. Malik, "Expert systems in electric power systems-a bibliographical survey," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1355–1362, 1989.
8. B. K. Wong and J. A. Monaco, "Expert system applications in business: A review and analysis of the literature (1977–1993)," *Inf. Manag.*, vol. 29, no. 3, pp. 141–152, Sep. 1995.
9. Shu-Hsien Liao, "Expert system methodologies and applications—a decade review from 1995 to 2004," *Expert Syst. Appl.*, vol. 28, no. 1, pp. 93–103, 2005.
10. T. Gupta and B. K. Ghosh, "A survey of expert systems in manufacturing and process planning," *Comput. Ind.*, vol. 11, no. 2, pp. 195–204, Jan. 1989.
11. T. M. Mitchell, *Machine Learning*. McGraw-Hill, Inc., 1997.
12. P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012.
13. J. Han, M. Kamber, and J. (Computer scientist) Pei, *Data mining: concepts and techniques*. Elsevier/Morgan Kaufmann, 2012.
14. I. Bose and R. K. Mahapatra, "Business data mining — a machine learning perspective," *Inf. Manag.*, vol. 39, no. 3, pp. 211–225, 2001.
15. I. H. (Ian H. . Witten and E. Frank, *Data mining : practical machine learning tools and techniques*. Morgan Kaufman, 2005.
16. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.
17. M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, pp. 235–244, 1997.
18. D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety," 2001.
19. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, pp. 137–144, 2015.
20. M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *Eur. Phys. J. Spec. Top.*, vol. 214, no. 1, pp. 481–518, Nov. 2012.
21. M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.
22. S. Yin and O. Kaynak, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, Feb. 2015.
23. W. Fan and A. Bifet, "Mining big data," *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, p. 1, Apr. 2013.