OBS Business School

Inteligencia artificial, inteligencia computacional y análisis inteligente de datos

José Ángel Olivas Varela
Colaborador de OBS Business School

Marzo, 2021

Partners Académicos:

Universitat de Barcelona

ulc barcelona

Autor

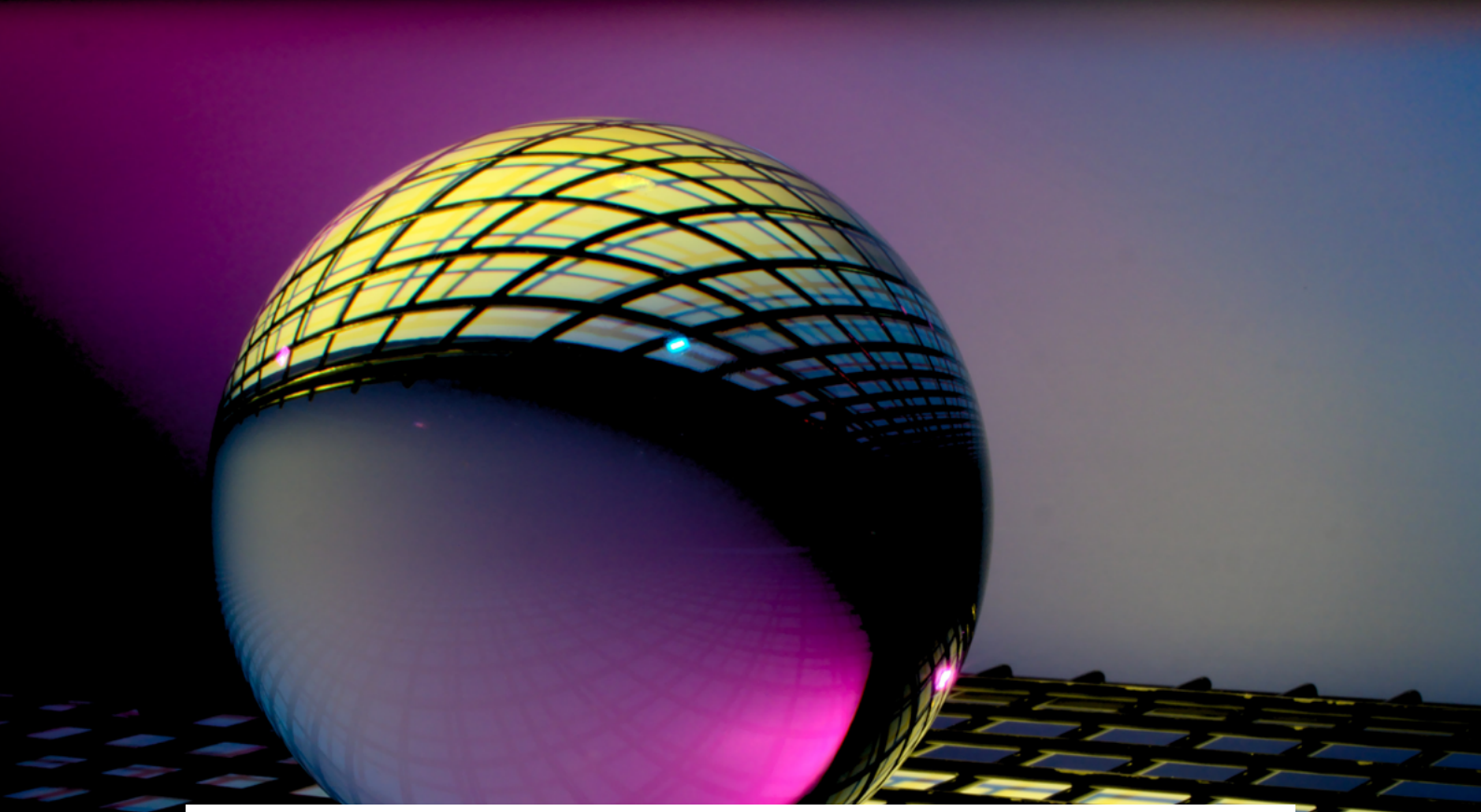
José Ángel Olivas Varela

Colaborador de
OBS Business School

Nacido en 1964 en Lugo, España, se licenció en Filosofía (especialidad Lógica) en 1990 (Universidad de Santiago de Compostela), obtuvo un Máster en Ingeniería del Conocimiento del Dpto. de Inteligencia Artificial de la Universidad Politécnica de Madrid en 1992, y un Doctorado en Ingeniería Informática en 2000 (Universidad de Castilla–La Mancha). En 2001 fue Postdoc Visiting Scholar en el BISC (Berkeley Initiative in Soft Computing) con Lotfi A. Zadeh (creador de la Lógica Borrosa), University of California-Berkeley, EE. UU. Desde entonces sigue colaborando activamente con el BISC y con el Centro de Inteligencia Artificial del SRI Internacional de la Universidad de Stanford (Prof. Richard Waldinger). Sus principales líneas de investigación actuales son el uso de técnicas de Soft Computing para la Recuperación de Información, el análisis de datos y las aplicaciones en Inteligencia Artificial e Ingeniería del Conocimiento. Recibió, entre otros, el premio en la modalidad de Investigación y/o Desarrollo de Productos Científicos en el XI concurso sobre Medio Ambiente del Ayuntamiento de Madrid (2002) por su tesis doctoral, y el premio al mejor artículo científico de la Asociación Española Para la Inteligencia Artificial (AEPIA) en 2015. Autor del libro *Búsqueda eficaz de información en la Web* y de más de 300 publicaciones científicas, es Doctor *honoris causa* por la Universidad Nacional de La Plata, Buenos Aires, Argentina, desde 2020.

Colaborador de OBS Business School desde 2015, Director del grupo de investigación SMILe-UCLM (*Soft Management of Internet and Learning*), Coordinador del Programa de Doctorado y del Programa Oficial de Posgrado en Tecnologías Informáticas Avanzadas de la UCLM, el Dr. Olivas ha colaborado con empresas como INSA (Ingeniería y Servicios Aeroespaciales, NASA), SOUTHCO, DANONE, VODAFONE, DOCPATH, INDRA, MOVISTAR, MOLLITIAM INDUSTRIES, ATT, y muchas otras.

Índice



Capítulo 1

Introducción

Hoy en día es frecuente oir en todos los medios de comunicación y en otros ámbitos hablar a todas horas de inteligencia artificial, en la mayoría de los casos vinculándola únicamente al procesamiento y análisis de datos. Pero si se analizan con rigor los orígenes y el objeto de esta disciplina, se verá que esta concepción es una simplificación peligrosa que puede llevar a equívocos en la sociedad en general. En este informe de investigación se intentarán aclarar diversos conceptos relacionados con la inteligencia artificial y se presentará un mapa de las técnicas y aplicaciones que suelen ser consideradas dentro de esta disciplina, tratando de diferenciar en lo posible cuáles deben ser consideradas realmente como inteligencia artificial (IA).

En primer lugar, en IA nada es tan nuevo como a veces se transmite o pueda parecer. Los modelos y técnicas más de moda en este momento surgieron hace ya varias décadas y apenas han sufrido variaciones desde su concepción.

Por ejemplo, los modelos basados en redes neuronales artificiales, en particular los denominados de *Deep Learning* (algoritmos que buscan relaciones profundas en datos numéricos, usando estructuras de redes neuronales artificiales y una computación muy intensa), tienen su origen en propuestas presentadas en los años siguientes a la Segunda Guerra Mundial, y los algoritmos de aprendizaje que utilizan fueron desarrollados en las décadas de los 80 y 90 del siglo pasado. Los *Random Forests* (algoritmos que pueden ser considerados 'bosques' de árboles de decisión, factibles actualmente gracias a la potencia computacional disponible) son desarrollos de esa misma época del profesor de estadística de la UC Berkeley, Leo Breiman (fallecido en 2005), que fueron presentados en el año 2001. La lógica borrosa (*fuzzy logic*) se sigue utilizando prácticamente de la misma forma en la que fue concebida por el profesor Zadeh (UC Berkeley, fallecido en 2017) en 1965. Realmente, en IA no ha habido nuevos paradigmas en los últimos 30 años. Lo que sí supone un cambio sustancial es el aumento de la potencia computacional y la capacidad de manipular grandes volúmenes de datos. Esto nos lleva a que estas técnicas, que en el momento de su presentación no eran tan eficientes, ahora sí lo sean, lo que permite que puedan ser aplicadas a grandes volúmenes de datos o a sistemas muy complejos en tiempos de computación mucho más razonables.

Pero no solo se hace IA a partir de datos. El "aprendizaje automático" (*machine learning*) es solo una parte de la IA. Además, este "aprendizaje" poco tiene que ver con cómo aprendemos los humanos: simplemente suele consistir en escalar operaciones estadísticas básicas sencillas a un conjunto grande de datos (estructurados, numéricos y normalizados…), lo que permite generalizar regularidades numéricas entre estos y encontrar patrones de comportamiento. Esto nos lleva al gran cuello de botella de los algoritmos de aprendizaje automático: no es fácil (ni posible) representar una situación real solamente con datos numéricos estructurados y normalizados, que es la única entrada que permiten estos métodos computacionales. Por lo tanto, IA no es equivalente a aprendizaje automático (ML), y mucho menos a *deep learning* (DL):

IA ≠ ML ≠ DL

En IA deben tenerse en cuenta otros modelos de simulación de la inteligencia humana, provenientes de áreas como la lógica, la lingüística, la sociología, la antropología social, la filosofía o la psicología cognitiva, como por ejemplo modelos de:

- Razonamiento, inferencia.
- Incertidumbre, imprecisión.
- Causalidad, heurística, categorización conceptual, almacenamiento de experiencias.
- Abstracción, lenguaje natural.
- Sentimientos, emociones.

Por ello, más allá de la matemática y la estadística, hay otras áreas de la IA en las que se debe profundizar y a las que se debe prestar más atención, como se mostrará en diversos ejemplos y aplicaciones a lo largo de este informe:

- Ingeniería del conocimiento.
- Procesamiento de Lenguaje Natural (PLN).
- Reconocimiento de patrones.
- Razonamiento aproximado, inferencia.
- Modelos causales, heurísticas, representación del conocimiento, almacenamiento de experiencias.

Áreas que, indudablemente, están profundamente vinculadas al *Big Data* (aprovechamiento inteligente de datos masivos). En otras palabras, se deben desarrollar sistemas de IA que tengan en cuenta no solo los datos existentes sobre el fenómeno a tratar, sino también el conocimiento humano sobre ese dominio (las hipótesis, interpretaciones y heurísticas de los expertos en ese campo y otros muchos elementos "cognitivos" más), para crear sistemas "sofisticados" que simulen el razonamiento humano, y no solo aplicar técnicas estadísticas básicas que nos permitan distinguir automáticamente fotografías de "perritos" de las de "gatitos" con un alto porcentaje de aciertos, pero con un coste computacional muy alto, tarea que un niño de pocos años hace con una tasa nula de error y sin ser considerado especialmente "inteligente".

Para profundizar en estos temas, son recomendables el libro de 2019 de Melanie Mitchell: *Artificial Intelligence: A Guide for Thinking Humans* [1], que habla por ejemplo de cómo la IA tuvo y sigue teniendo el problema del "burro y el palo con la zanahoria", pues en los años 50 la gente decía que en cuestión de 10, 15, 25 años íbamos a tener una IA completa y capaz de sustituir al ser humano en todos los empleos, o el artículo de agosto de 2020 de investigadores de universidades de Finlandia y Austria titulado "*Artificial Intelligence*: *A Clarification of Misconceptions, Myths and Desired Status*" [2], en el que los autores presentan diversos mitos y errores en la concepción actual de la IA en esta línea, acerca de la imposibilidad real de crear máquinas que se comporten como humanos "óptimos".

En este informe de investigación se abordarán los conceptos de IA y aprendizaje automático, y se ofrecerá un mapa de las técnicas más habituales en este campo, y otras tales como la computación suave y el razonamiento aproximado. Posteriormente se hará una aproximación crítica al análisis inteligente de datos, haciendo especial hincapié en el dominio de la medicina y el análisis de sentimientos y opiniones. El informe se cerrará con un apartado de conclusiones, orientadas principalmente al futuro profesional vinculado a la IA y a las diversas oportunidades que ofrece, y a cómo la IA está provocando un cambio significativo en la sociedad (digital) actual.

Capítulo 2

Qué son la Inteligencia Artificial y el análisis inteligente de datos, y cómo afectan a nuestra vida cotidiana

La **Inteligencia Artificial** (IA/AI –siglas en inglés–) se puede ver como la disciplina del ámbito de la computación y los sistemas de información que pretende simular computacionalmente comportamientos humanos que pueden ser considerados como inteligentes. Hay diversas ramas dentro de la IA, como la Visión Artificial o la Robótica, pero en este informe nos centraremos en el aprendizaje automático (AA/ML *Machine Learning*) y en lo que se suele denominar "Ingeniería del Conocimiento" (IC/KE), que se ocupa del desarrollo de Sistemas Basados en el Conocimiento (SBC/KBS), como pueden ser los Sistemas de Ayuda a la Decisión (DSS *Decision Support Systems*). La tradición de los SBC comenzó en los años 80 del siglo pasado con los denominados "Sistemas Expertos", sistemas computacionales que tratan de emular las capacidades de un experto en un tema determinado, un proceso que se basa en la extracción del conocimiento del propio experto (o grupo de expertos) y su posterior transmisión al sistema. Con la proliferación del almacenamiento y uso de datos de forma masiva, los SBC actuales suelen apoyarse en ambos pilares: expertos y datos.

Para la gestión del conocimiento experto hay diversas metodologías que consisten básicamente en la adquisición, representación e implantación de dicho conocimiento (IC). Estos sistemas suelen usar bases de reglas del tipo "si el paciente tiene los síntomas A, B y C, entonces con probabilidad o creencia X tiene la enfermedad E", para almacenar y utilizar este conocimiento con el objeto de inferir nuevos consejos de **ayuda en la decisión**.

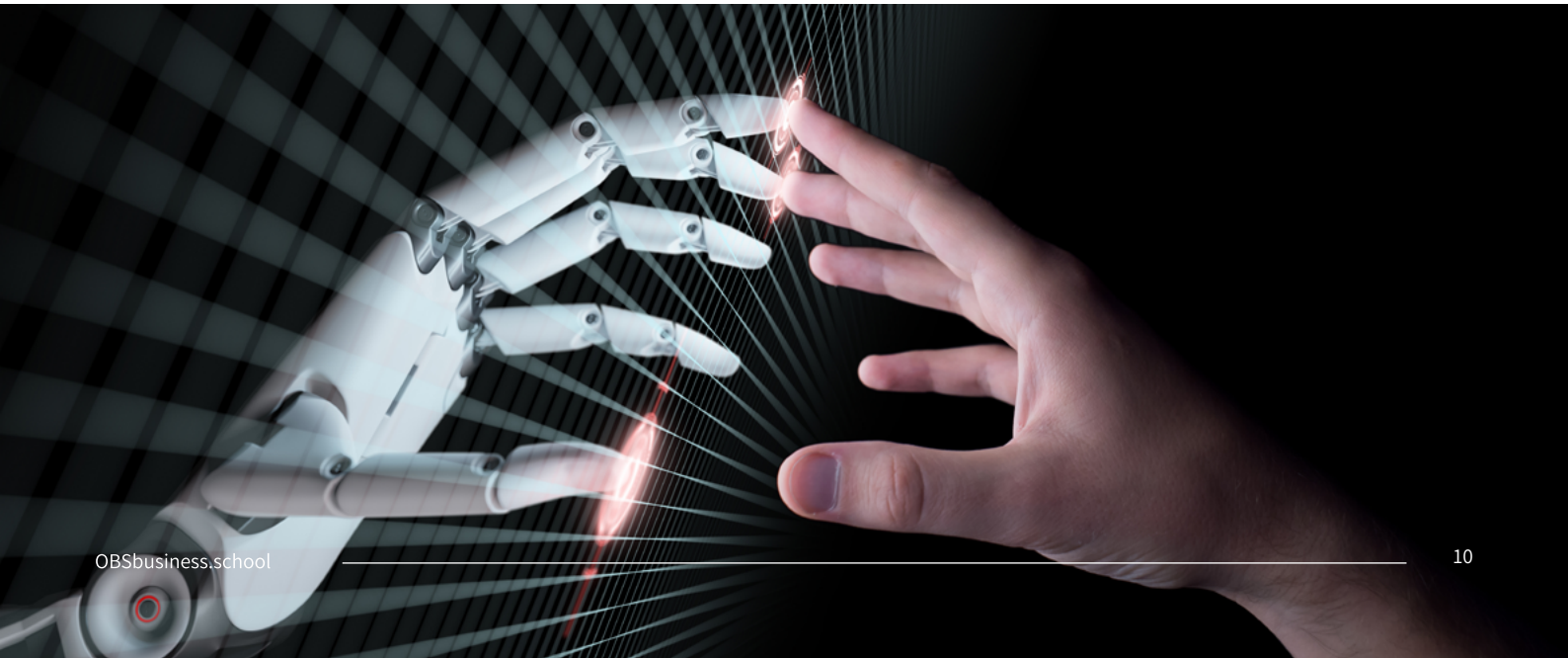
Cuando se tienen en cuenta datos (como historias clínicas de pacientes, casos anteriormente tratados, registros de incidencias de enfermedades, datos de factores que pueden provocar determinadas dolencias, etc.), entonces es necesario recurrir a lo que en IA se llaman técnicas de aprendizaje automático o ***Machine Learning*** y, por supuesto, a técnicas provenientes de la matemática y la estadística, como por ejemplo las de regresión.

La mayor parte de las técnicas de *Machine Learning* (AA), rama de la IA en la que se diseñan mecanismos para dotar a los sistemas computacionales de capacidad de aprendizaje, en el sentido de la capacidad de descubrir **regularidades (patrones)** en datos o situaciones anteriores y aplicarlos a nuevos problemas o situaciones análogas, provienen de la estadística, y permiten encontrar y caracterizar estas regularidades numéricas en los datos considerados

Aunque muchas de estas propuestas son propias tanto de la estadística como del AA, se puede considerar este **mapa completo** de los paradigmas y grupos de técnicas

- **Paradigma Analógico (Aprendizaje por analogía)**: Se pretende encontrar una solución a un problema que se presenta ahora usando el mismo procedimiento utilizado en la resolución de uno similar que se presentó en una ocasión anterior. Si dos problemas son similares en algún aspecto de su formulación, entonces pueden serlo también en sus soluciones. Nuevos problemas pueden ser abordados reduciéndolos a problemas análogos resueltos. Ejemplos: analogía por transformación, analogía por derivación, "razonamiento basado en casos", etc.
- **Paradigma Inductivo**: Árboles de decisión, algoritmos de inducción pura, etc.
- **Paradigma Conexionista**: Redes Neuronales Artificiales, etc.
- **Paradigma Evolutivo**: Algoritmos Genéticos, otros métodos de optimización, colonias de insectos, *Particle Swarm Optimization*, descenso estocástico del gradiente, etc.
- **Modelos gráficos probabilistas**: Bayesianos, cadenas de Markov, Filtros de Kalman, redes de creencia, Máquinas de Soporte Vectorial (SVM), Metaheurísticas, etc.

Además, según el tipo y contenido de los datos disponibles, y el objetivo que se persigue en su análisis, las técnicas de AA se suelen clasificar en técnicas de **agrupamiento (clustering)** y técnicas de **clasificación**.



Las **técnicas de *clustering*** (aprendizaje no supervisado) consisten en agrupar los elementos de una colección en subconjuntos (clases, categorías, *clusters*), nítidos o borrosos, en base a su similitud. Es aprendizaje no supervisado porque las clases o categorías no se conocen a priori, sino que serán determinadas por las propias similitudes entre los elementos (es decir, los datos no están "etiquetados" ni contienen información sobre la clase a la que pertenece cada registro). Por lo tanto, estas técnicas se centran en una "medida de similitud" entre elementos, de la que puede haber infinidad de variantes: estadísticas, distancias euclídeas, distancias vectoriales (coseno), distancias borrosas, etc. Los principales exponentes de este tipo de algoritmos en los paradigmas enumerados anteriormente son:

- **Paradigma Conexionista**: Redes Neuronales Artificiales: SOM (*Self Organized Maps*, Mapas de Kohonen, toolbox de Matlab SOM).
- **Modelos estadísticos y probabilistas**: *K-means, c-means, K-nearest neighbours* (KNN), *Mean shift* (ventanas circulares con un centroide), *Dirichlet process* (estocásticos basados en distribuciones de probabilidad), LDA (*Latent Dirichlet Allocation*), Modelos Gaussianos, etc.
- **Extensiones basadas en Lógica Borrosa (fuzzy logic)**: *Fuzzy K-means, Fuzzy c-means*, Isodata, etc.

Las **técnicas de clasificación** (aprendizaje supervisado), consisten en asignar una clase a un nuevo elemento en base a un conjunto de categorías previamente establecidas (por ejemplo, evaluando los síntomas de un nuevo paciente para decidir si tiene o no gripe –clase previamente establecida). Se basan en un entrenamiento en base a ejemplos con la solución conocida (supervisado) para crear modelos que permitan clasificar nuevos casos análogos:

- **Paradigma Inductivo**: Árboles de decisión [3]: ID3 [4], CART [5], C4.5 [6], See5, *Random Forest* (de moda en Big Data, introducidos por Leo Breiman en 2001), etc.
- **Paradigma Conexionista**: Redes Neuronales Artificiales: Perceptrón Multicapa (con *backpropagation*), Convolutivas, Neocognitrones, Redes de Hopfield, Redes recurrentes, Adaline, *Deep Learning (*de moda en Big Data), etc.
- **Modelos estadísticos y probabilistas**: Redes Bayesianas, Naive-Bayes, Máquinas de Soporte Vectorial (SVM), Metaheurísticas, etc.

1 Computación suave y razonamiento aproximado

El **razonamiento aproximado** se enmarca dentro de la **inteligencia computacional** (*computational intelligence*), esencialmente dentro de lo que se suele denominar **computación suave** (*soft computing*), por el hecho de que abarca técnicas tolerantes a la imprecisión y la incertidumbre inherentes a la representación del conocimiento e inferencia de cualquier sistema computacional inteligente. Principalmente, la lógica borrosa o difusa (*fuzzy logic*) [7] provee de mecanismos adecuados para formalizar este razonamiento aproximado de una forma completa y precisa. El razonamiento aproximado puede entenderse como un conjunto de técnicas que permiten describir, representar y hacer inferencias considerando y aprovechando la imprecisión e incertidumbre inherentes a todos los desarrollos actuales en IA.

Desde que en el año 1965 el profesor Lotfi A. Zadeh introdujo la lógica borrosa [8], muchas han sido sus aplicaciones, las más importantes en el campo del control industrial (lavadoras Bosch con sistema Eco-Fuzzy, ABS de Nissan, aire acondicionado Mitsubishi...). Pero, ya desde sus inicios, la intención del profesor Zadeh era introducir un formalismo capaz de representar y manipular la incertidumbre e imprecisión inherentes al lenguaje natural y, sobre todo, al razonamiento aproximado.

Por otro lado, la mayor parte de la inmensa cantidad de información contenida en Internet está almacenada en documentos textuales (lenguaje natural escrito) y en multitud de idiomas. Además, muchos de los datos de lo que hoy en día se denomina *big data* tienen una gran carga de imprecisión e incertidumbre.

El profesor Zadeh también acuñó el término *soft-computing*, que se puede traducir al español como *computación suave o blanda*. La computación suave se diferencia de la computación convencional (dura) en que, a diferencia de ella, es tolerante a la imprecisión, la incertidumbre y la verdad parcial. El modelo a seguir para la computación suave es la mente humana y su principio rector es aprovechar la tolerancia a la imprecisión, la incertidumbre y la verdad parcial para lograr trazabilidad, robustez y bajo coste en las soluciones.

Las ideas básicas que subyacen a la computación suave en su estado actual tienen vínculos con muchas influencias anteriores, entre ellas, los conjuntos difusos introducidos en los años 60 del siglo pasado, los trabajos de los años 70 sobre el análisis de sistemas complejos y procesos de decisión, y los de los años 80 sobre teoría de posibilidades y análisis de datos blandos. La inclusión de la teoría de redes neuronales en la computación suave llegó en un momento posterior. En esta coyuntura, los principales componentes de la computación suave (CS) son la lógica borrosa (LB), la teoría de redes neuronales (RN) y el razonamiento probabilístico (RP), con este último subsumiendo las redes de creencias, los algoritmos genéticos, la teoría del caos y partes de la teoría del aprendizaje.

Lo que es importante tener en cuenta es que el CS no es una mezcla de LB, RN y RP. Se trata más bien de una asociación en la que cada uno de los socios aporta una metodología distinta para abordar los problemas en su ámbito. Desde esta perspectiva, las principales contribuciones de LB, RN y RP son complementarias y no competitivas.

La complementariedad de LB, RN y RP tiene una consecuencia importante: en muchos casos un problema puede resolverse de forma más eficaz utilizando LB, RN y RP en combinación y no exclusivamente. Un ejemplo llamativo de una combinación particularmente efectiva es lo que se ha llegado a conocer como sistemas neuroborrosos (*neurofuzzy*), que permiten que mediante el uso combinado de una red neuronal artificial con algunos de sus parámetros y mecanismos algorítmicos "borrosificados", el ABS y el "control de velocidad de crucero" de Nissan o Mazda sean más suaves, que las lavadoras Bosch con sistema *eco-fuzzy* sequen mejor la ropa y salga menos arrugada, que las imágenes de video sean más estables o que los aires acondicionados de Misubishi mantengan la temperatura más uniformemente. Lo que es particularmente significativo es que, tanto en los productos de consumo como en los sistemas industriales, el empleo de técnicas de *soft computing* conduce a sistemas que tienen un alto cociente de inteligencia de máquina (IM).

En gran medida, es la alta IM de los sistemas basados en CS lo que explica el rápido crecimiento en el número y la variedad de aplicaciones de la computación suave y especialmente de la lógica difusa. La estructura conceptual de la computación suave sugiere que los estudiantes deben ser entrenados no solo en teoría de redes neuronales, lógica borrosa y razonamiento probabilístico, sino en todas las metodologías asociadas, aunque no necesariamente en el mismo grado. Lo mismo se aplica a las revistas, los libros y las conferencias. Hay muchas revistas científicas y libros con este término en su título. Una tendencia similar es perceptible en los títulos de las conferencias y congresos.

2 Qué se entiende habitualmente por ‘analítica de datos’

La “analítica de datos” [9] se suele enmarcar como una parte esencial de lo que se denomina “Inteligencia de Negocio” (*Business Intelligence*, BI) y se suele definir como la capacidad de transformar datos [10] en información para ayudar a gestionar una empresa, que consiste en los procesos, aplicaciones y prácticas que apoyen la toma de decisiones ejecutivas. La BI se suele dividir en dos grandes grupos:

- **BI Operacional**: trata de los informes estándar, descripciones de los datos (información), funciones al nivel operacional con los trabajadores, clientes, usuarios, socios…
- **BI Analítica y Táctica/Estratégica**: pretende dar soporte a los ejecutivos y a los gestores en niveles tácticos que contribuyen a la estrategia global de la empresa o institución. Suele contemplar análisis estadístico, modelos predictivos y de extrapolación, pronósticos, optimización…

La **analítica descriptiva** está orientada a la generación de un resumen claro y fácil de entender de una colección de datos. Este es el fundamento y concepto más básico de todas las estadísticas. Se centra en la visualización de datos para entender el pasado y el presente. Se describen los datos con tablas o gráficos, mostrando descripciones numéricas de la variabilidad y la posición. También se suele llamar modelización descriptiva.

Para el **análisis predictivo** se suelen usar técnicas **estadísticas**. Las más utilizadas son las de **regresión**, que expresan la formalización de una relación significativa entre dos o más variables para calcular pronósticos a partir del conocimiento de los valores en un individuo concreto. Entre otros tipos de regresión, para el análisis de datos se suelen usar principalmente la lineal, la múltiple o la logística. Merece una mención especial el citado método CART (*Classification and Regression Trees*, fundamento de los *Random Forests*, introducido en 2001 por Leo Breiman), porque es uno de los ejemplos más claros de las diversas técnicas que se pueden englobar tanto en técnicas estadísticas como en técnicas de aprendizaje automático. En otras palabras, la mayoría de las técnicas de aprendizaje automático se basan en mecanismos estadísticos.

También suelen usarse técnicas de extrapolación de funciones (estimaciones o líneas de tendencia, pero sin capacidad de pronóstico, hechos/cambios puntuales) o métodos para establecer correlaciones entre variables (demasiado evidente, no suele funcionar de forma muy fina). Además, se buscan patrones en los datos que puedan ser aplicados a situaciones futuras (KDD *Knowledge Discovery in Databases* y Minería de Datos) [11, 12, 13], donde se suelen emplear los métodos de *clustering* y clasificación enumerados anteriormente.

Dentro del análisis predictivo juega un papel fundamental el **Análisis de Series Temporales**, las cuales se suelen clasificar en estacionarias (medias y/o variabilidad se mantienen constantes) y no estacionarias (medias y/o variabilidad no se mantienen constantes, cambios de varianza/tendencias). También es importante señalar que se usan otros métodos para diferentes necesidades en el análisis de series temporales, como por ejemplo para anticipar las tendencias (método de mínimos cuadrados, tendencias evolutivas, diferenciación estacional), para la predicción de comportamientos futuros (regresión o alisadores exponenciales), o para la interpolación, que consiste en predecir datos faltantes.

El análisis predictivo se centra en un escenario futuro. El **prescriptivo** se centra en múltiples alternativas. Por lo tanto, un modelo prescriptivo puede ser considerado como una combinación de modelos predictivos (uno por cada posible escenario) que se ejecutan en paralelo. El objetivo es encontrar la mejor opción posible: OPTIMIZACIÓN. Por ejemplo, la elección del tratamiento más adecuado en oncología. Las principales técnicas usadas provienen de la Investigación Operativa o de la IA y en particular del AA, como los algoritmos genéticos (computación evolutiva), técnicas estocásticas o algunas metaheurísticas.

Conviene distinguir entre **predicción** y **pronóstico**. Aunque es frecuente su uso de forma inversa según los campos de aplicación, en estadística la segunda tiene que ver con la estimación, con la extrapolación, la continuidad; por ejemplo, podemos estimar (pronosticar) cuántos habitantes tendrá Madrid en 2025 en base a la tendencia demográfica. En cambio, en otros dominios el pronóstico consiste en anticiparse a un hecho puntual en base a un conjunto pequeño de alternativas; por ejemplo, en el reverso del impreso de una quiniela de fútbol pone “marque con una X su pronóstico” y las alternativas son 1-X-2, o un terremoto se debe pronosticar en base a las alternativas SI-NO. Los sistemas para abordar una u otra opción son claramente distintos. Los primeros se basan en la extrapolación y los segundos en una serie de características para anticiparse al hecho puntual (por ejemplo, en las quinielas saber si alguno de los equipos se juega algo importante como el descenso, si hay un jugador importante lesionado o quién va a ser el árbitro del partido).

Esta distinción tiene una fuerte repercusión en el tipo de sistemas inteligentes a desarrollar: los de estimación, cuando los fenómenos y los datos que los describen siguen una tendencia, es suficiente con el uso de técnicas estadísticas o de aprendizaje automático de las citadas anteriormente. En cambio, cuando debemos anticiparnos a hechos puntuales que suponen una ruptura de la tendencia es necesario el uso de técnicas más cognitivas, como por ejemplo acudir al conocimiento de expertos, o métodos que consideren relaciones causales en fenómenos anteriormente descritos (como por ejemplo electorales) y sean capaces de extrapolar estas relaciones a situaciones futuras en las condiciones específicas actuales.

3 Críticas sobre la percepción habitual del análisis de datos: se debe ir hacia un análisis 'inteligente' de datos

Desde el punto de vista del autor de este informe, la visión habitual de las posibilidades y expectativas de la BI es demasiado restringida, y mucho más las prácticas habituales en las empresas e instituciones. Se habla por ejemplo de "transformar datos en información", o de "apoyar la toma de decisiones", pero estas son expresiones muy imprecisas y, además, ¡hay muchas otras cosas que se pueden hacer!

Si nos fijamos en las posibles salidas propuestas, vemos que nuevamente el ámbito es demasiado restringido. Normalmente la "información" es solo "visualización", es decir, una forma de representar datos originales que están en bruto de una forma más ordenada y resumida, como por ejemplo el típico 'gráfico de quesera'. Pero esto es muy poco "inteligente", pues al fin y al cabo será el examinador el responsable de tomar las decisiones, basándose en la información de la que dispone. La calidad de estas decisiones dependerá exclusivamente de la capacidad y preparación del decisor humano.

Por ello, la tendencia debería ser hacia la generación y el manejo de "conocimiento" en el sentido anteriormente descrito. Se pueden diseñar sistemas sofisticados de muy diversos tipos, como los Sistemas de Ayuda a la Decisión (DSS), los Sistemas Recomendadores (*Recommender Systems*, como los que usa *Amazon* o *TripAdvisor*) o el Análisis de Series Temporales (Predicción vs. Pronóstico), por ejemplo, para tareas automáticas de Segmentación (segmentación de clientes, de pacientes, de productos…), recomendación de otros productos, pronóstico de terremotos, incendios u otros fenómenos naturales, puntos de inflexión en la bolsa o el cambio de moneda... y solo las grandes empresas punteras se esfuerzan en diseñar sistemas de este tipo. En cambio, en países como España o los de Latinoamérica es muy raro encontrar empresas o entidades que afrontan con interés y rigor estos objetivos.

Como se ha dicho, el interés de todo ello pasa por encontrar regularidades en los datos o su evolución (¡patrones!), y ser capaz de formalizarlos o expresarlos en alguna estructura formal computable para que puedan ser usados por estos sistemas. Las salidas esperadas deben condicionar todo el proceso de análisis de datos. Por ejemplo, no se diseña de la misma forma un sistema de **predicción** (en que los datos no mantienen una tendencia y el comportamiento futuro del fenómeno estudiado depende de factores puntuales, y por tanto es más cognitivo) que uno de **pronóstico** (en los que el fenómeno sí sigue una tendencia clara y se pueden usar las técnicas estadísticas comentadas). Así, la práctica habitual es ir "a ciegas" hacia delante en la aplicación de los sistemas computacionales, es decir, partir de los datos más o menos en bruto y aplicarles algún algoritmo o herramienta de análisis (por ejemplo: estadística o de aprendizaje automático) y tratar de interpretar la salida generada por si puede ser útil. Desgraciadamente, este es un enfoque claramente erróneo del análisis y minería de datos, y suele ser el más habitual.

4 Algunas aplicaciones a la medicina y otros ámbitos

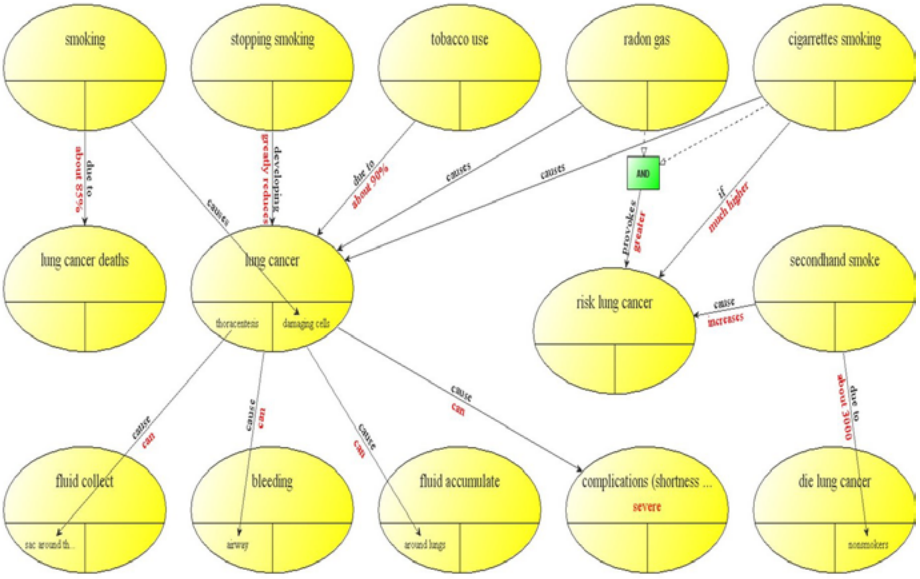
Como se ha visto, los Sistemas de Ayuda a la Decisión (*Decision Support Systems* DSS) son sistemas computacionales que proporcionan consejos para la mejora en la toma de decisiones. Hay una gran cantidad de aplicaciones de este tipo de sistemas. Si nos centramos en sistemas de medicina, podemos citar tres pequeños ejemplos (que se pueden estudiar en detalle siguiendo las referencias bibliográficas), desarrollados en el marco del grupo de investigación liderado por el autor de este informe.

El primero tiene que ver con el concepto de "enfermedades borrosas" (*fuzzy diseases*), implantado en un sistema para diagnosticar y tratar fibromialgia, que obtuvo el premio al mejor trabajo en el congreso de la Asociación Española para la Inteligencia Artificial en 2015 [14]. En este trabajo se propone el uso del concepto de "prototipo deformable borroso" para caracterizar enfermedades que pueden ser confundidas o que incluso no están perfectamente caracterizadas o asumidas por toda la comunidad médica.

En el segundo ejemplo se muestra el diseño de un sistema de ayuda a la decisión en oncología a partir de la detección, clasificación y uso de las expresiones causales y condicionales en textos médicos (en este caso de Mayo Clinic y Mount Sinai Hospital). En la figura 1 se puede ver el grafo causal de la pregunta "¿qué causa cáncer de pulmón?" [15]. Se parte de un programa que permite identificar frases causales y condicionales en documentos de texto y que las clasifica entre 20 tipos posibles de las tres estructuras condicionales básicas del idioma inglés (*first, second,* y *third*), así como entre otro tipo de estructuras del estilo de "*due to*" o "*causes*". Una vez identificadas y clasificadas estas oraciones caudales, se genera de forma automática el grafo causal, que puede ser de ayuda en la toma de decisiones, tanto en el diagnóstico como en el tratamiento, al igual que en la explicación o en cualquier otro tipo de conocimientos relacionados. Este mecanismo también permite responder a preguntas del tipo "¿Cuál es la principal causa del cáncer de pulmón en determinado contexto?", o diseñar sistemas para la administración inteligente de fármacos a lo largo del tiempo en procesos donde el tratamiento no está pautado de forma precisa ni es predecible de antemano, como puede ser la insulina en determinados tipos de diabetes.

Figura 01 ⟶ GRAFO CAUSAL DE LA PREGUNTA "¿QUÉ CAUSA CÁNCER DE PULMÓN?"

Fuente: Sobrino et al. (2014)



En el tercer caso se presenta un estudio inicial sobre el desarrollo de un sistema para la estimación de probabilidades de sufrir determinados tipos de cáncer en base al estudio del genograma del paciente, llevado a cabo con el servicio de oncología de un conocido hospital español [16]. El manejo de genogramas es un ejemplo claro de tratamiento de imprecisión e incertidumbre ("el abuelo creo que falleció de cáncer de intestino"), pero es mucho más ventajoso tratar de aprovechar esta información que desecharla por no saber cómo explotarla. Esta capacidad de establecer inferencias en condiciones de imprecisión e incertidumbre permite, por ejemplo, recomendar al paciente algún test genético o aconsejar una mastectomía preventiva.

Se han propuesto y diseñado aplicaciones a otros diversos campos basadas en modelos sofisticados de IA (en este caso los mencionados "**prototipos deformables borrosos**", propuestos por el autor de este informe en 2000 en su tesis doctoral), como las relacionadas con:

- Prevención de **incendios forestales** (en Galicia, Siberia, Canadá, EE.UU.) [17]: a partir de series temporales de "ciclos de incendios" entre periodos de lluvias el sistema permite aconsejar sobre la optimización de recursos (en cantidad, prioridades u horarios), en base a la capacidad de anticiparse al número de incendios esperable en cada zona concreta.

- **Ingeniería del software** y sistemas de información [18, 19]: para el tratamiento de datos sobre diversos elementos y factores considerados en metodologías y ciclos de vida de desarrollo de software, como por ejemplo la "comunicación entre equipos" o la "dificultad de la tarea", difíciles de estimar de forma precisa.
- **Control de tráfico** [20]: descubriendo y estableciendo patrones de los flujos de tráfico a partir de los datos normalmente muy imprecisos de los que se dispone, con el fin, por ejemplo, de regular la sincronización de los semáforos de una determinada zona.
- **Historias clínicas electrónicas y gestión documental** [21]: la historia clínica de un paciente puede contener muchos documentos (se estima que en una persona "normal" de mediana edad pueden ser más de 300) de naturaleza heterogénea (imágenes con diferentes niveles de definición, informes de consulta, informes de urgencias, tratamientos recomendados, analíticas…) y que además suelen estar en diferentes formatos según las diferentes zonas geopolíticas. Entre otras muchas cosas, se han desarrollado sistemas para la generación automática de resúmenes de estas historias clínicas que permitan, por ejemplo, disponer de información precisa en situaciones de urgencia sin tener que consultar todos los documentos: "el paciente es diabético, hipertenso, alérgico a algunos tipos de anestesia y ha tenido dos cirugías en la zona abdominal…".
- **Recuperación de información** y búsqueda Web [22, 23]: se han implementado diversos **metabuscadores** que permiten añadir una capa "inteligente" antes de enviar las consultas a los buscadores Web habituales como *Google* o *Bing*. Esto permite considerar por ejemplo sinónimos o variedades diatópicas de un idioma cuando se realiza una consulta: encontrar páginas de "pantallas no muy caras" cuando se busca "monitores baratos" o recetas de "frutillas con crema chantilly" cuando se buscan "fresas con nata".
- **Ciencias sociales** [24]: etc.

5 El análisis de sentimientos y opiniones en redes sociales

El **análisis de sentimientos** o **minería de opiniones** (*opinion mining*) es uno de los temas de investigación más recientes en el ámbito de la IA. Hoy en día es uno de los campos más importantes, difíciles y demandados por la repercusión que tiene tanto para las empresas como para la sociedad en general. En las investigaciones desarrolladas por el grupo de investigación del autor de este informe, se han propuesto diversas aplicaciones y métodos. El artículo "*Sentiment analysis: A review and comparative analysis of web services*" [25] ofrece una descripción del estado actual de la investigación y las aplicaciones en análisis de sentimientos y opiniones.

Las técnicas de **recuperación de información textual**, **minería de textos** o **procesamiento de lenguaje natural (PLN)** se suelen centrar en el procesamiento, la búsqueda o la extracción de información objetiva. Los hechos tienen un componente objetivo; sin embargo, también hay otros elementos textuales que expresan características subjetivas. Estos elementos son principalmente opiniones, sentimientos, valoraciones, actitudes y emociones, que son el foco del análisis de sentimientos. Todos ellos están estrechamente relacionados, aunque presentan ligeras diferencias. Esto implica el nacimiento de muchas tareas relacionadas con este nuevo campo de investigación, como la minería de opiniones, el análisis de subjetividad, la detección de emociones o la detección del *spam* de opinión, entre otros.

El análisis de sentimientos ofrece muchas oportunidades para desarrollar nuevas aplicaciones, especialmente debido al gran crecimiento de las herramientas disponibles, como por ejemplo las recomendaciones sobre los temas propuestos en fuentes como blogs y redes sociales. El sistema de recomendaciones puede funcionar teniendo en cuenta aspectos tales como las opiniones positivas o negativas sobre esos temas o productos. Los sitios web de opiniones pueden recopilar información de diferentes fuentes con el fin de resumir o componer una opinión global sobre un candidato, producto, etc., lo que puede sustituir a los sistemas que requieren explícitamente opiniones o resúmenes.

Los **sistemas de pregunta y respuesta** representan otro campo en el que las opiniones desempeñan un papel importante. La detección de preguntas sobre opiniones y sus posibles respuestas, así como su tratamiento, son esenciales para generar buenas respuestas. La detección de información subjetiva es realmente importante en campos relacionados con la argumentación, en los que las frases objetivas suelen ser más valiosas. Aun así, uno de los campos en los que el análisis de sentimientos tiene un mayor impacto es la industria. Pequeñas y grandes empresas, al igual que otras organizaciones, como los gobiernos, tienen mucho interés en saber lo que la gente siente sobre sus marcas, productos o miembros.

Como se ha dicho, el análisis de sentimientos es un concepto que abarca muchas tareas, como la extracción de sentimientos, la clasificación de sentimientos, la clasificación de subjetividad, el resumen de opiniones o la detección de *spam* de opinión, entre otros. Para llevar a cabo cualquiera de estas actividades, el análisis de sentimientos tiene que lidiar con muchos desafíos.

El primero es la definición de los elementos que intervienen. Por lo tanto, es necesario definir claramente conceptos como **opinión**, **subjetividad** o **emoción** (una tarea nada fácil). Por ejemplo, de una manera sencilla, la opinión de un usuario puede ser considerada como un sentimiento positivo o negativo acerca de una entidad o de un elemento de esa entidad. Por otro lado, la subjetividad no implica necesariamente un sentimiento, sino que permite expresar sentimientos o creencias y, específicamente, nuestros propios sentimientos, creencias y emociones.

Estas definiciones tienen que ser **formalizadas mediante expresiones matemáticas** que puedan ser calculadas y utilizadas como entradas para los algoritmos de IA. Por lo tanto, el éxito del análisis de sentimientos depende principalmente de la capacidad de extraer la información necesaria de esas definiciones a partir de los textos para llevar a cabo esas tareas.

A pesar de la complejidad y la dificultad de este problema, muchas empresas y universidades están desarrollando **nuevas herramientas e instrumentos y servicios web** que tratan varios de los temas mencionados. Estos servicios podrían incluirse, especialmente para la investigación, en otras aplicaciones o plataformas sin necesidad de ser experto en análisis de sentimientos.

Los términos **minería de opiniones**, **análisis de sentimientos** y **análisis de subjetividad** (*opinion mining, sentiment analysis* y *subjectivity analysis* respectivamente) se usan frecuentemente como sinónimos. Sin embargo, sus orígenes son diversos y algunos autores consideran que cada concepto presenta connotaciones diferentes, al igual que otros estrechamente relacionados, como el **análisis afectivo**.

Surgen muchas **tareas vinculadas al análisis de sentimientos.** Algunas de ellas están estrechamente relacionadas y es difícil separarlas claramente porque comparten muchos aspectos. Las más importantes son:

- **Clasificación de los sentimientos** (también llamada orientación de sentimientos, orientación de opinión, orientación semántica o polaridad de sentimientos). Se basa en la idea de que un documento o texto puede expresar una opinión de una persona sobre una entidad y trata de medir su sentimiento hacia dicha entidad. Por lo tanto, esta tarea consiste básicamente en clasificar las opiniones en tres categorías principales: positivo, negativo o neutro.
- **Clasificación de subjetividad.** Consiste principalmente en detectar si una frase dada es subjetiva o no. Una frase objetiva expresa información factual, mientras que una frase subjetiva puede expresar otro tipo de información personal, como opiniones, evaluaciones, emociones y creencias. Además, las frases subjetivas pueden expresar lo positivo o lo negativo, pero no todas lo hacen. Esta tarea puede ser vista como un paso previo a la clasificación de los sentimientos. Una buena clasificación de subjetividad puede asegurar una mejor clasificación de sentimientos. Incluso se considera como un proceso más difícil que el de distinguir entre sentimientos positivos, neutros o negativos.

- **Resumen de opiniones.** Se centra especialmente en la extracción de las características principales de una entidad compartida en uno o varios documentos y los sentimientos sobre ella. Por lo tanto, se pueden distinguir dos perspectivas en esta tarea: resumen de un solo documento o de varios documentos. La integración de documentos individuales consiste en analizar los hechos presentes en el documento analizado, por ejemplo, cambios en la orientación de los sentimientos a lo largo del documento o vínculos entre las diferentes entidades o características encontradas, y mostrar principalmente aquellos textos que mejor las describen. Por otro lado, en los resúmenes multidocumento, una vez detectadas las características y entidades, el sistema tiene que agrupar y/o ordenar las diferentes frases que expresan sentimientos relacionados con esas entidades o rasgos. Al final puede ser presentado como un gráfico o un texto que muestre las principales características o entidades y cuantifique de alguna manera el sentimiento en cada uno de ellos, por ejemplo, agregando intensidades de sentimientos o contando el número de sentimientos positivos o negativos.

- **Recuperación de opiniones.** Intenta recuperar documentos que expresan una opinión sobre una consulta determinada. En este tipo de sistemas se necesitan dos puntuaciones para cada documento: la puntuación de relevancia frente a la consulta y la puntuación de opinión sobre la consulta. Normalmente se utilizan ambos para clasificar los documentos.

- **Sarcasmo e ironía.** Se centra en detectar afirmaciones con contenido irónico y sarcástico. Esta es una de las tareas más complicadas en este campo, especialmente, debido a la falta de acuerdo entre los investigadores sobre cómo la ironía o el sarcasmo pueden ser formalmente representadas y definidas.

- **Otros.** Además de las actividades mencionadas anteriormente, se pueden destacar otras tareas relacionadas con el análisis de sentimientos, como la detección de género o autoría, que intenta determinar el género o la persona que ha escrito un texto u opinión, o la detección de *spam* de opinión, que trata de detectar opiniones o reseñas que incluyen contenidos no confiables publicados para distorsionar la opinión pública hacia las personas, empresas o productos.

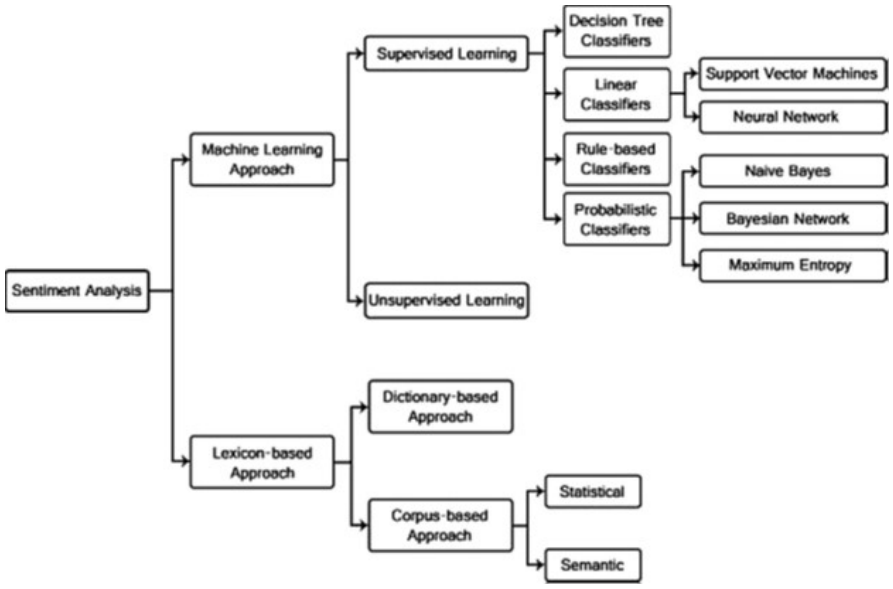
Las **técnicas de IA** que se usan habitualmente están basadas en aprendizaje automático, uso de diccionarios, estadística y semántica.

Como se ha visto, las técnicas de aprendizaje automático se pueden agrupar en tres categorías principales: cuando es posible se usan técnicas **supervisadas**, mientras que las técnicas **semisupervisadas** y **no supervisadas** se proponen cuando no es posible tener un conjunto inicial de documentos u opiniones etiquetados para clasificar el resto de los ítems. El éxito de estas técnicas se basa principalmente en la selección y extracción del conjunto apropiado de características utilizadas para detectar sentimientos. En esta tarea, las técnicas de procesamiento del lenguaje natural desempeñan un papel muy importante. En este campo, algunos de los elementos a tener en cuenta son:

- Los **términos** (palabras o n-gramas) y su frecuencia.
- La **información** de la **categoría gramatical**. Los adjetivos desempeñan un papel importante pero los sustantivos pueden ser significativos.
- Las **negaciones**, que pueden cambiar el significado de cualquier oración.
- Las **dependencias sintácticas** (análisis de árbol), que pueden determinar el significado de una oración.

Figura 2 ⟶ TÉCNICAS USADAS HABITUALMENTE PARA EL ANÁLISIS DE SENTIMIENTOS

Fuente: Serrano et al. 1015



Además, los **enfoques híbridos**, que combinan técnicas supervisadas y no supervisadas, o incluso técnicas semisupervisadas, pueden ser útiles para clasificar los sentimientos.

Otras técnicas se basan principalmente en un **léxico de sentimientos** (es decir, en una colección de sentimientos conocidos y precompilados), términos, frases e incluso modismos, desarrollados para los géneros tradicionales de comunicación, como el *Opinion Finder*. Estructuras aún más complejas, como ontologías o diccionarios, que miden la orientación semántica de las palabras o frases, pueden ser usadas para este propósito. Aquí se pueden encontrar dos subclasificaciones: enfoques basados en diccionarios y en corpus.

Los **enfoques basados en diccionarios** se apoyan generalmente en el uso de un conjunto inicial de términos (semillas), que normalmente se recogen y anotan de forma manual. Este conjunto crece buscando los sinónimos y antónimos en un diccionario. Un ejemplo de ese diccionario podría ser *WordNet*, que se utilizó para desarrollar un tesauro llamado *SentiWordNet*. El principal inconveniente de este tipo de enfoques es la incapacidad de hacer frente a orientaciones específicas de dominio y contexto. Aun así, podría ser una solución interesante dependiendo del problema.

Las técnicas basadas en **corpus** surgen con el objetivo de proporcionar diccionarios relacionados con un dominio específico. Estos diccionarios se generan a partir de un conjunto de términos de opinión de semillas que crecen a través de la búsqueda de palabras relacionadas mediante el uso de técnicas estadísticas o semánticas. En los métodos basados en estadística como el análisis semántico latente, simplemente se puede utilizar la frecuencia de aparición de las palabras dentro de una colección de documentos. Por otro lado, los métodos semánticos, como el uso de sinónimos y antónimos, o las relaciones de tesauro, como *WordNet*, también pueden representar una solución interesante.

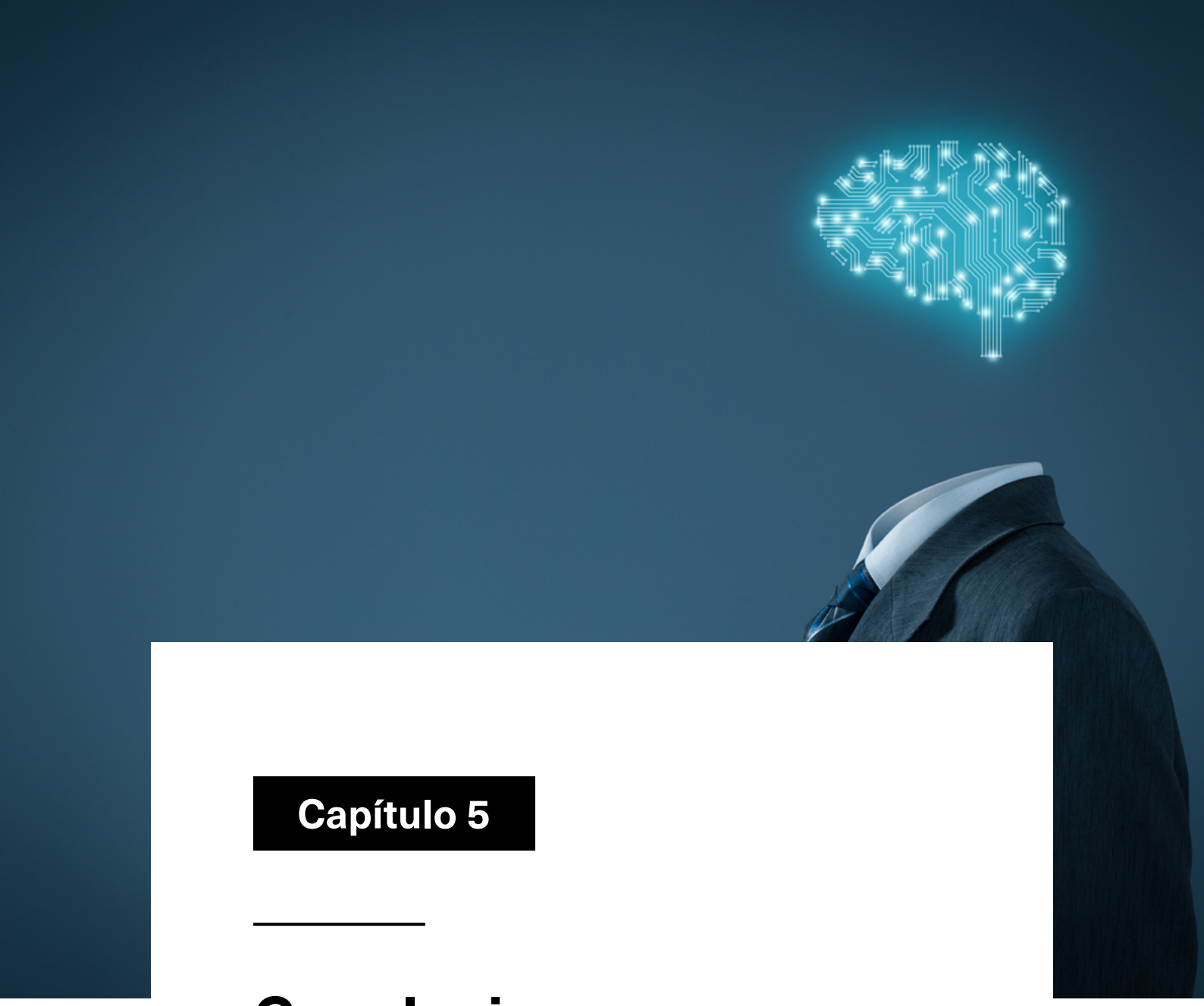
6 El científico de datos, la profesión más demandada en IA

Hoy en día el "científico de datos" (*data scientist*) es una figura muy demandada pero escasa en ambientes profesionales, científicos o académicos, puesto que, como se esboza en este documento, un "científico de datos" debe poseer conocimientos de computación, bases de datos, IA, Aprendizaje Automático, estadística, visualización, reconocimiento de patrones, sociología, psicología, KDD y Minería de Datos… y debe ser capaz de seleccionar y guiar las herramientas y técnicas más adecuadas para cada problema y objetivos concretos. Evidentemente, hoy en día no es fácil encontrar profesionales con un perfil tan completo.

La mejor forma de paliar esta dificultad debería ser la formación de equipos multidisciplinares que cubran todas estas necesidades, pero esto tampoco es frecuente y por ello gran parte de los proyectos de análisis de datos que llevan a cabo diferentes instituciones y entidades no consiguen los resultados esperados.

Otro problema frecuente e importante a la hora de desarrollar este tipo de proyectos es la tendencia a aplicar las herramientas, métodos o algoritmos que el equipo mejor conoce ("al que solo dispone de un martillo, todos los problemas le parecen clavos") de una forma "ciega" sobre las bases de datos de las que se dispone (normalmente solo una), sin importar en principio cuál es el propósito final de ese análisis de datos. Esta aproximación (errónea) podría mejorarse en gran medida disponiendo de un buen "científico de datos", en el sentido anteriormente descrito, capaz de guiar todo el proceso con criterios bien sustentados.

Las habilidades más demandadas para los trabajos vinculados con la IA y el AA son el conocimiento de determinados lenguajes y entornos de programación, en particular **R** (para estadística) y **Python** (Anaconda, Scikit, para AA e IA en general) o **Scala** (multiparadigma); diferentes herramientas específicas para entornos *big data* como **Mahout** (*scalable machine learning and data mining*), **Spark ML/Lib** (*machine learning library*) o **H2O** (*data science in H2O*), así como el propio ecosistema **Apache hadoop** (basado en el paradigma **mapreduce** implantado en Google desde 2001); o herramientas específicas para la implementación de redes neuronales artificiales y modelos de *deep learning*, como **Keras** o **Tensorflow**.



Capítulo 5

Conclusiones

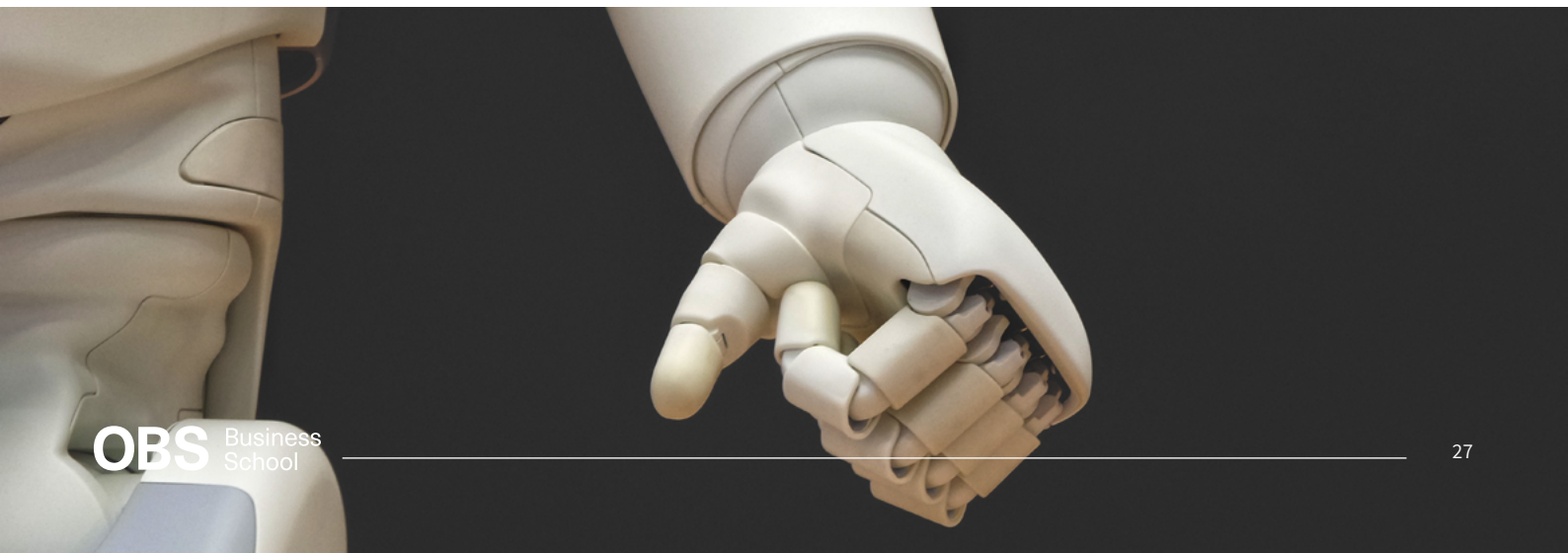
En este informe de investigación se ha intentado plasmar una visión crítica sobre lo que suele entenderse hoy en día por inteligencia artificial. También se ha propuesto una visión general de lo que suele ser entendido por IA (que va más allá del aprendizaje automático y la analítica de datos). Se han presentado diversas aplicaciones de la IA a diferentes campos (en particular a la medicina) y se ha abordado uno de los principales temas de investigación y de preocupación en todas las empresas de hoy: el análisis de sentimientos y opiniones.

Como hemos visto, la necesidad de un análisis inteligente de datos en todas las empresas y entidades provoca la demanda de diversos profesionales de múltiples ámbitos. En particular, son especialmente requeridos titulados en matemáticas y estadística, por su capacidad para entender los rudimentos formales de la mayoría de los algoritmos de aprendizaje automático y poder diseñar y modificar algoritmos de este tipo. Sin embargo, resulta claro que los directores de proyectos de esta naturaleza deben tener una formación más amplia, incluso más allá de la ciencia de datos. Deberían disponer de un "maletín de herramientas" que les permita seleccionar aquellas más adecuadas para cada problema específico, y no solo en lo que concierne al análisis de datos. Por ejemplo, para un sistema de anticipación a un resultado electoral, no solo se deben tener en cuenta los datos, sino que hay que considerar teorías e hipótesis de la sociología o la antropología social, como por ejemplo el efecto o teorema de Thomas. De la misma forma, para la implementación de un sistema de análisis de sentimientos u opiniones, sería conveniente contar con expertos en lingüística computacional y en psicología clínica.

Esto significa que cada vez se requerirán más profesionales de diversos ámbitos con formación orientada hacia los desarrollos en inteligencia artificial. Por otra parte, desde el punto de vista puramente de ingeniería, son necesarios profesionales capaces de mantener y diseñar sistemas robotizados o automáticos, que en muchos casos van a sustituir a otro tipo de mano de obra humana, lo que está provocando cambios sustanciales en los modelos sociales de trabajo.

La inteligencia artificial está cambiando y cambiará todavía más nuestras vidas y nuestro modo de desenvolvernos en el mundo. Pero hay que tener mucho cuidado con no trivializar y a cualquier automatismo sencillo llamarle "de inteligencia artificial". Por ejemplo, que un reloj sea capaz de medir nuestras pulsaciones y generar un gráfico para estimar cómo van a comportarse en un futuro… es discutible que esto deba ser considerado inteligencia artificial; que un asistente de voz responda correctamente a la orden "mañana despiértame con música romántica", o que un sistema domótico abra las persianas a determinada hora o bajo determinadas condiciones, tampoco deben ser vistos como tal.

En conclusión, debemos pensar en una inteligencia artificial que emule en determinados aspectos el comportamiento racional humano, y no solo en sistemas que analicen un conjunto de datos numéricos desde un punto de vista estadístico.



Referencias bibliográficas

1 Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking*. Penguin Books, London, UK.

2 Emmert-Streib, F., Yli-Harja, O., Dehmer, M. (2020). Artificial Intelligence: A Clarification of Misconceptions, Myths and Desired Status, Frontiers in Artificial Intelligence 3: 91. https://www.frontiersin.org/article/10.3389/frai.2020.524339

3 Quinlan, J. R. (1979). Discovering Rules by Induction from a Large Collection of Examples. En D. Michie (ed.) Expert Systems in the Microelectronic Age, Edinburgh University Press.

4 Quinlan, J. R. (1983). Induction of Decision Trees. Machine Learning 1, 81-106.

5 Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

6 Quinlan, J. R. (1988). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo CA.

7 Olivas, J. A. (2002). La Lógica Borrosa y sus aplicaciones. BOLE.TIC, nº 24, pp. 21 - 28.

8 Zadeh, L. A. (1987). *Fuzzy Sets and Applications* (Selected Papers, edited by R. R. Yager, S. Ovchinnikov, R. M. Tong, H. T. Nguyen), John Wiley, Nueva York, 1987.

9 Siegel, E. (2013). *Analítica predictiva. Predecir el futuro utilizando Big Data*. Anaya Multimedia-Anaya Interactiva.

10 Mayer-Schönberger, V., Cukier, K. (2013). *Big data. La revolución de los datos masivos*. Turner.

11 Piatetsky-Shapiro, G., Frawley, W. (1991). *Knowledge Discovery in Databases*. AAAI/MIT Press, Cambridge MA.

12 Agrawal, D., Das, S., Abbadi, A. E. (2010). Big Data and Cloud Computing: New Wine or Just New Bottles?. Proc. of the VLDB 2010, Vol. 3, No. 2.

13 Agrawal, D., Das, S., Abbadi, A. E. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. Proc. of the ETDB 2011, Uppsala, Sweden.

14 Romero-Cordoba, R., Olivas, J. A., Romero, F. P., Alonso-Gonzalez, F., Serrano-Guerrero, J. (2017). An Application of Fuzzy Prototypes to the

Diagnosis and Treatment of Fuzzy Diseases. Int. Journal of Intelligent Systems 32(2), 194-210.

15 Sobrino, A., Puente, C., Olivas, J. A, (2014). Extracting Answers from causal mechanisms in a medical document. Neurocomputing 135, 53–60.

16 Calatrava, C., Oruezabal, M. J., Olivas, J. A., Romero, F. P., Serrano-Guerrero, J. (2015). A Decision Support System for Risk Analysis and Diagnosis of Hereditary Cancer, Proc. of the 2015 International Conference on Artificial Intelligence IC-AI'2015, Las Vegas, NV, CSREA Press, USA.

17 Olivas, J. A. (2003). Forest Fire Prediction and Management using Soft Computing. Proc. of the 1st IEEE-INDIN'03, IEEE International Conference on Industrial Informatics: 338 – 344.

18 Genero, M., Olivas, J. A., Piattini, M., Romero, F. P. (2001). Using metrics to predict OO information systems maintainability. In Dittrich, Geppert and Norrie (eds): Advanced Information Systems Engineering CAISE2001, Springer, LNCS 2068: 388–401.

19 Peralta, A., Romero, F. P., Olivas, J. A., Polo, M. (2010). Knowledge extraction of the behaviour of software developers by the analysis of time recording logs. Proc. of the 2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'10).

20 Angulo, E.; Romero, F. P.; García, R.; Serrano-Guerrero, J.; Olivas, J. A. (2011). An adaptive approach to enhanced traffic signal optimization by using soft-computing techniques. Expert Systems with Applications 38 (3): 2235-2247.

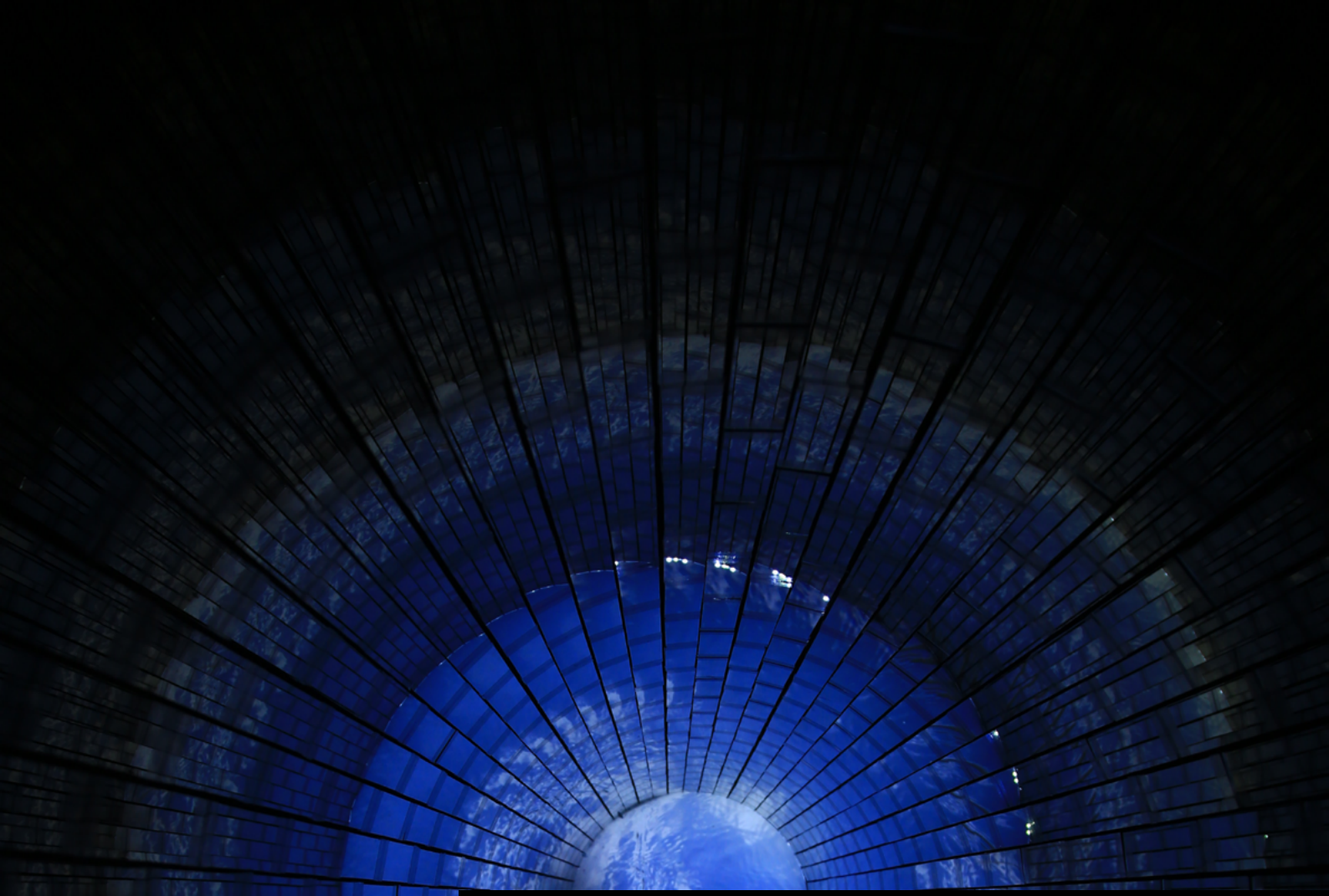
21 Romero F. P.; Caballero, I.; Serrano-Guerrero, J.; Olivas, J. A. (2012). An approach to web-based Personal Health Records filtering using fuzzy prototypes and data quality criteria, Information Processing & Management, 48 (3): 451-466.

22 Garcés, P., Olivas, J. A., Romero, F. P. (2006): Concept-matching IR systems versus Word-matching IR systems: Considering fuzzy interrelations for indexing web pages. Journal of the American Society for Information Science and Technology JASIST, 57 (4): 564-576.

23 Olivas, J. A. (2011). *Búsqueda eficaz de información en la Web*, Edulp, La Plata, Argentina. http://hdl.handle.net/10915/18401.

24 Sobrino, A.; Olivas, J. A.; Puente, C. (2010). Causality and Imperfect Causality from texts: a frame for causality in social sciences. Proc. of the 2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'10).

25 Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences 311, 18–38.



OBS Business School

School of **Business Administration & Leadership**

School of **Innovation, & Technology Management**

School of **Health Management**

De:

Planeta Formación y Universidades