CAPÍTULO 1

¿Qué es la ciencia de datos?

La ciencia de datos abarca un conjunto de principios, definiciones de problemas, algoritmos y procesos para extraer patrones no obvios y útiles de grandes conjuntos de datos. Muchos de los elementos de la ciencia de datos se han desarrollado en campos relacionados, como el aprendizaje automático y la minería de datos. De hecho, los términos *ciencia de datos*, *aprendizaje automático* y *minería de datos* a menudo se usan indistintamente. Lo que comparten estas disciplinas es el enfoque de mejorar la toma de decisiones a través del análisis de datos. Sin embargo, aunque la ciencia de datos toma prestado de estos otros campos, tiene un alcance más amplio. El aprendizaje automático se centra en el diseño y la evaluación de algoritmos para extraer patrones de los datos. La minería de datos generalmente se ocupa del análisis de datos estructurados y a menudo implica un énfasis en las aplicaciones comerciales. La ciencia de datos tiene en cuenta todas estas consideraciones, pero también aborda otros desafíos, como la captura, limpieza y transformación de redes sociales y datos web no estructurados; el uso de tecnologías del *big data* para almacenar y procesar grandes conjuntos de datos no estructurados; y preguntas relacionadas con la ética y la regulación de datos.

Mediante la ciencia de datos podemos extraer diferentes tipos de patrones. Por ejemplo, podríamos querer extraer patrones que nos ayuden a identificar grupos de clientes que exhiben comportamientos y gustos similares. En la jerga empresarial, esta tarea se conoce como *segmentación de clientes*, y en la terminología de la ciencia de datos se llama *agrupamiento*. Alternativamente, podríamos querer extraer un patrón que identifique los productos que se compran frecuentemente juntos,

un proceso llamado *minería de reglas de asociación*. O podríamos querer extraer patrones que identifiquen eventos extraños o anormales, como reclamos de seguro fraudulentos, un proceso conocido como *anomalía o detección de valores atípicos*. Finalmente, podríamos querer identificar patrones que nos ayuden a clasificar las cosas. Por ejemplo, la siguiente regla ilustra cómo se vería un patrón de clasificación extraído de un conjunto de datos de correo electrónico: *Si un correo electrónico contiene la frase "Hacer dinero fácilmente", es probable que sea correo no deseado.* Identificar estos tipos de reglas de clasificación se conoce como *predicción*. La palabra *predicción* puede parecer una elección extraña porque la regla no predice lo que sucederá en el futuro: el correo electrónico ya es o no es un correo no deseado. Por lo tanto, es mejor pensar que los patrones de predicción predicen el valor faltante de un atributo en lugar de predecir el futuro. En este ejemplo, estamos prediciendo si el atributo de clasificación de correo electrónico debe tener el valor "correo no deseado" o no.

Si un experto humano puede crear fácilmente un patrón en su propia mente, entonces no vale la pena el tiempo y el esfuerzo que requiere la ciencia de datos para "descubrirlo".

Aunque podemos usar la ciencia de datos para extraer diferentes tipos de patrones, siempre queremos que los patrones sean no obvios y útiles. El ejemplo de la regla de clasificación de correo electrónico del párrafo anterior es tan simple y obvia que si fuera la única regla extraída por un proceso de ciencia de datos, quedaríamos decepcionados. Por ejemplo, esta regla de clasificación de correo electrónico verifica solo un atributo: ¿contiene la frase "ganar dinero fácilmente"? Si un experto humano puede crear fácilmente un patrón en su propia mente, entonces no vale la pena el tiempo y el esfuerzo que requiere la ciencia de datos para "descubrirlo". En general, la ciencia de datos se vuelve útil cuando tenemos una gran cantidad de ejemplos de datos y cuando los patrones son demasiado complejos para que los humanos los descubran y extraigan manualmente. Como límite inferior, podemos tomar una gran cantidad de ejemplos de datos para definir que supere lo que un experto humano puede verificar fácilmente. Con respecto a la complejidad de los patrones, podemos definirla en relación con las habilidades humanas. Los humanos somos razonablemente buenos para definir reglas que marcan un, dos, cientos, miles y, en casos extremos, millones de atributos.

Los patrones que extraemos mediante la ciencia de datos son útiles solo si nos dan una idea del problema que nos permite hacer algo para ayudar a resolverlo. La frase *conocimiento procesable* a veces se usa en este contexto para describir lo que queremos que nos den los patrones extraídos. El término *conocimiento* destaca que el patrón debería proporcionarnos información relevante sobre el problema que no sea obvia. El término *procesable* destaca que la información que obtenemos también debe ser algo que tengamos la capacidad de usar de alguna manera. Por ejemplo, imagina que estamos trabajando para una compañía de teléfonos celulares que está tratando de resolver un problema de *abandono* de clientes, es decir, demasiados clientes se están cambiando a otras compañías. Una forma en que se podría utilizar la ciencia de datos para abordar este problema es extraer patrones de los datos sobre clientes anteriores que nos permitan identificar a los clientes actuales que tienen riesgos de abandono y luego contactar a estos clientes e intentar convencerlos de que se queden con nosotros. Un patrón que nos permite identificar a los posibles clientes que abandonarían es útil para nosotros solo si (a) los patrones identifican a los clientes con suficiente anticipación para que podamos contactarlos antes de que abandonen y (b) nuestra empresa pueda formar un equipo para contactarlos. Ambas cosas son necesarias para que la empresa pueda actuar según el conocimiento que nos brindan los patrones.

Una breve historia de la ciencia de datos

El término *ciencia de datos* tiene una historia específica que se remonta a la década de 1990. Sin embargo, los campos en los que se basa tienen una historia mucho más larga. Un aspecto en esta historia más larga es la historia de la recopilación de datos; otro es la historia del análisis de datos. En esta sección, revisaremos los principales desarrollos en estos aspectos y describiremos cómo y por qué convergieron en el campo de la ciencia de datos. Por necesidad, esta revisión introduce una nueva terminología a medida que describimos y nombramos las innovaciones técnicas importantes a medida que vayan surgiendo. Para cada nuevo término proporciorenamos una breve explicación de su significado. Más adelante en el libro volveremos a muchos de estos términos y proporcionaremos una explicación más detallada de ellos. Comenzaremos con la historia de la recopilación de datos, luego presentaremos la historia del análisis de datos y, finalmente, cubriremos el desarrollo de la ciencia de datos.

La historia de la recopilación de datos

Los primeros métodos para registrar datos pueden haber sido marcas en palos para registrar el paso de los días o postes clavados en el suelo para marcar el amanecer en los solsticios. Con el desarrollo de la escritura, sin embargo, nuestra capacidad de registrar nuestras experiencias y los eventos en nuestro mundo aumentó enormemente la cantidad de datos que recopilamos. La primera forma de escritura se desarrolló en Mesopotamia alrededor del 3.200 a. C. y se utilizó para mantener registros comerciales. Este tipo de mantenimiento de registros captura lo que se conoce como *datos transaccionales*. Los datos transaccionales incluyen información de eventos como la venta de un artículo, la emisión de una factura, la entrega de bienes, el pago con tarjeta de crédito, las reclamaciones de seguros, etc. Los *datos no transaccionales*, como los datos demográficos, también tienen una larga historia. Los primeros censos conocidos tuvieron lugar en el Egipto faraónico alrededor del año 3.000 a. C. La razón por la cual los primeros estados pusieron tanto esfuerzo y recursos en grandes operaciones de recolección de datos fue que estos estados necesitaban aumentar los impuestos y los ejércitos, lo que demuestra la afirmación de Benjamin Franklin de que solo hay dos cosas ciertas en la vida: la muerte y los impuestos.

En los últimos 150 años, el desarrollo del sensor electrónico, la digitalización de datos y la invención de la computadora han contribuido a un aumento masivo en la cantidad de datos que se recopilan y almacenan. Un hito en la recopilación y el almacenamiento de datos ocurrió en 1970 cuando Edgar F. Codd publicó un artículo que explicaba el *modelo de datos relacionales*, que fue revolucionario en términos de establecer cómo se almacenaban, indexaban y recuperaban (en ese momento) los datos de las bases de datos. El modelo de datos relacionales permitió a los usuarios extraer datos de una base de datos mediante consultas simples que definían qué datos deseaba el usuario sin requerir que se preocupara por el estándar internacional de estructura subyacente para definir consultas de base de datos. Las bases de datos relacionales almacenan datos en tablas con una estructura de una fila por instancia y una columna por atributo. Esta estructura es ideal para almacenar datos porque puede descomponerse en atributos naturales.

Las bases de datos son la tecnología natural que se utiliza para almacenar y recuperar datos *transaccionales* u operativos estructurados (es decir, el tipo de datos generados por las operaciones diarias de una empresa). Sin embargo, a medida que las compañías se han vuelto más grandes y más automatizadas, la cantidad y variedad de datos generados por diferentes partes de estas compañías han aumentado dramáticamente. En la década de 1990, las empresas se dieron cuenta de que a pesar de que estaban acumulando enormes cantidades de datos, se encontraban repetidamente con dificultades para analizar esos datos. Parte del problema era que los datos a menudo se almacenaban en numerosas bases de datos separadas dentro de una organización. Otra dificultad era que las bases de datos estaban optimizadas para el almacenamiento y la recuperación de datos, actividades caracterizadas por altos volúmenes de operaciones simples, como SELECCIONAR, INSERTAR, ACTUALIZAR y ELIMINAR. Para analizar sus datos, estas compañías necesitaban tecnología que pudiera reunir y conciliar los datos de bases de datos dispares y que facilitara las operaciones de datos analíticos más complejos. Este desafío empresarial condujo al desarrollo de *almacenes de datos*. En un almacén de datos, los datos se toman de toda la organización y se integran, lo que proporciona un conjunto de datos más completo para el análisis.

En las últimas décadas, nuestros dispositivos se han vuelto móviles y conectados en red, y muchos de nosotros pasamos muchas horas en línea todos los días usando tecnologías sociales, juegos de computadora, plataformas de medios y motores de búsqueda web. Estos cambios en la tecnología y en cómo vivimos han tenido un impacto dramático en la cantidad de datos recopilados. Se estima que la cantidad de datos recopilados durante los cinco milenios desde la invención de la escritura hasta 2003 es de aproximadamente 5 exabytes. Desde 2013, los humanos generan y almacenan esta misma cantidad de datos *todos los días*. Sin embargo, no solo es la cantidad de datos recopilados lo que ha crecido dramáticamente sino también la variedad de datos. Solo considera la siguiente lista de fuentes de datos en línea: correos electrónicos, blogs, fotos, *tweets*, me gusta, recursos compartidos, búsquedas en la web, carga de videos, compras en línea, podcasts. Y si consideramos los metadatos (datos que describen la estructura y las propiedades de los datos brutos) de estos eventos, podemos comenzar a comprender el significado del término *big data*. El *big data* a menudo se define en términos de las tres V: el *volumen* extremo de datos, la *variedad* de los tipos de datos y la *velocidad* a la que deben procesarse los datos.

La llegada del *big data* ha impulsado el desarrollo de una gama de nuevas tecnologías de bases de datos. Esta nueva generación de bases de datos a menudo se conoce como *"bases de datos NoSQL"*. Por lo general, tienen un modelo de datos más simple que las bases de datos relacionales tradicionales. Una base de datos NoSQL almacena datos como objetos con atributos, utilizando un lenguaje de notación de objetos como el *JavaScript Object Notation* (JSON). La ventaja de usar una representación de datos de objetos (en contraste con un modelo basado en tablas relacionales) es que el conjunto de atributos para cada objeto está encapsulado dentro del objeto, lo que resulta en una representación flexible. Por ejemplo, puede ser que uno de los objetos en la base de datos, en comparación con otros objetos, solo tenga un subconjunto de atributos. Por el contrario, en la estructura de datos tabular estándar utilizada por una base de datos relacional, todos los puntos de datos deben tener el mismo conjunto de atributos (es decir, columnas). Esta flexibilidad en la representación de objetos es importante en contextos donde los datos no pueden (por variedad o tipo) descomponerse naturalmente en un conjunto de atributos estructurados. Por ejemplo, puede ser difícil definir el conjunto de atributos que deberían usarse para representar texto libre (como *tweets*) o imágenes. Sin embargo, aunque esta flexibilidad de representación nos permite capturar y almacenar datos en una variedad de formatos, estos datos aún deben extraerse en un formato estructurado antes de que se pueda realizar un análisis en ellos.

La existencia del *big data* también ha llevado al desarrollo de nuevos marcos de procesamiento de datos. Cuando se trata de grandes volúmenes de datos a altas velocidades, puede ser útil desde una perspectiva computacional y de velocidad distribuir los datos en varios servidores, procesar consultas calculando resultados parciales de una consulta en cada servidor y luego combinar estos resultados para generar la respuesta a la consulta. Este es el enfoque adoptado por el marco de *MapReduce* en Hadoop. En el marco de MapReduce, los datos y las consultas se asignan a (o se distribuyen en) varios servidores, y los resultados parciales calculados en cada servidor se reducen (fusionan).

La historia del análisis de datos

La estadística es la rama de la ciencia que se ocupa de la recopilación y el análisis de datos. El término *estadística* originalmente se refería a la

recopilación y análisis de datos sobre el Estado, como datos demográficos o datos económicos. Sin embargo, con el tiempo se amplió el tipo de datos a los que se aplicaba el análisis estadístico, de modo que hoy las estadísticas se utilizan para analizar todo tipo de datos. La forma más simple de análisis estadístico de datos es el resumen de un conjunto de datos en términos de *estadísticas de resumen* (*descriptivas*) (incluidas medidas de una tendencia central, como la *media aritmética*, o medidas de variación, como el *rango*). Sin embargo, en los siglos XVII y XVIII, el trabajo de personas como Gerolamo Cardano, Blaise Pascal, Jakob Bernoulli, Abraham de Moivre, Thomas Bayes y Richard Price sentó las bases de la teoría de la probabilidad, y a lo largo del siglo XIX muchos estadísticos comenzaron a utilizar distribuciones de probabilidad como parte de su kit de herramientas analíticas. Estos nuevos desarrollos en matemáticas permitieron a los estadísticos ir más allá de las estadísticas descriptivas y comenzar a hacer *aprendizaje estadístico*. Pierre Simon de Laplace y Carl Friedrich Gauss son dos de los matemáticos más importantes y famosos del siglo XIX, y ambos hicieron importantes contribuciones al aprendizaje estadístico y la ciencia de datos moderna. Laplace tomó las intuiciones de Thomas Bayes y Richard Price y las desarrolló en la primera versión de lo que ahora conocemos como la *regla de Bayes*. Gauss, en su búsqueda del planeta enano desaparecido Ceres, desarrolló el *método de mínimos cuadrados*, que nos permite encontrar el mejor modelo que se ajusta a un conjunto de datos de modo que el error en el ajuste minimice la suma de las diferencias al cuadrado entre los puntos de datos en el conjunto de datos y el modelo. El método de mínimos cuadrados proporcionó la base para los métodos de aprendizaje estadístico como la *regresión lineal* y la *regresión logística*, así como el desarrollo de modelos de *redes neuronales artificiales* en inteligencia artificial (volveremos a los mínimos cuadrados, análisis de regresión y redes neuronales en el capítulo 4).

Entre 1780 y 1820, casi al mismo tiempo que Laplace y Gauss estaban haciendo sus contribuciones al aprendizaje estadístico, un ingeniero escocés llamado William Playfair estaba inventando gráficos estadísticos y sentando las bases para la *visualización de datos* y el *análisis exploratorio de datos* modernos. Playfair inventó el *gráfico de líneas* y el *gráfico de área* para datos de series temporales, el *gráfico de barras* para ilustrar comparaciones entre cantidades de diferentes categorías y el *gráfico circular* para ilustrar proporciones dentro de un conjunto. La ventaja de visualizar datos cuantitativos es que nos permite usar nuestras poderosas habilidades visuales para resumir, comparar e interpretar datos. Es cierto que es

difícil visualizar conjuntos de datos grandes (muchos puntos de datos) o complejos (muchos atributos), pero la visualización de datos sigue siendo una parte importante de la ciencia de datos. En particular, es útil para ayudar a los científicos de datos a explorar y comprender los datos con los que están trabajando. Las visualizaciones también pueden ser útiles para comunicar los resultados de un proyecto de ciencia de datos. Desde la época de Playfair, la variedad de gráficos de visualización de datos ha crecido constantemente, y hoy en día hay investigaciones en curso sobre el desarrollo de enfoques novedosos para visualizar grandes conjuntos de datos multidimensionales. Un desarrollo reciente es el algoritmo de *incrustación de vecino estocástico distribuido en t* (t-SNE), que es una técnica útil para reducir datos de alta dimensión a dos o tres dimensiones, lo que facilita la visualización de esos datos.

Los desarrollos en la teoría de la probabilidad y las estadísticas continuaron hasta el siglo XX. Karl Pearson desarrolló pruebas de hipótesis modernas, y R. A. Fisher desarrolló métodos estadísticos para el *análisis multivariado* e introdujo la idea de la *estimación de máxima verosimilitud* en la inferencia estadística como un método para sacar conclusiones basadas en la probabilidad relativa de eventos. El trabajo de Alan Turing en la Segunda Guerra Mundial condujo a la invención de la computadora electrónica, que tuvo un impacto dramático en las estadísticas porque permitió cálculos estadísticos mucho más complejos. A lo largo de la década de 1940 y las décadas posteriores, se desarrollaron varios modelos computacionales importantes que todavía se usan ampliamente en la ciencia de datos. En 1943, Warren McCulloch y Walter Pitts propusieron el primer modelo matemático de una *red neuronal*. En 1948, Claude Shannon publicó "Una teoría matemática de la comunicación" y al hacerlo fundó la *teoría de la información*. En 1951, Evelyn Fix y Joseph Hodges propusieron un modelo para el *análisis discriminatorio* (lo que ahora se llamaría un problema de *clasificación* o *reconocimiento de patrones*) que se convirtió en la base de los modelos de vecinos más cercanos modernos. Estos desarrollos posguerra culminaron en 1956 con el establecimiento del campo de la *inteligencia artificial* en un taller en Dartmouth College. Incluso en esta etapa temprana del desarrollo de la inteligencia artificial, el término *aprendizaje automático* estaba comenzando a usarse para describir programas que le daban a una computadora la capacidad de aprender de los datos. A mediados de la década de 1960, se hicieron tres contribuciones importantes al aprendizaje automático. En 1965, el libro de Nils Nilsson titulado *Learning Machines* mostró cómo las redes neuronales podían usarse para

aprender modelos lineales para clasificar. Al año siguiente, Earl B. Hunt, Janet Marin y Philip J. Stone desarrollaron el marco del sistema de aprendizaje de conceptos, que fue el progenitor de una importante familia de algoritmos del aprendizaje automático que indujeron modelos de árbol de decisión a partir de datos según un modelo descedente. Casi al mismo tiempo, varios investigadores independientes desarrollaron y publicaron versiones tempranas del algoritmo de agrupamiento *k-means*, ahora el algoritmo estándar utilizado para la segmentación de (clientes) datos.

El campo del aprendizaje automático está en el núcleo de la ciencia de datos moderna porque proporciona algoritmos que pueden analizar automáticamente grandes conjuntos de datos para extraer patrones potencialmente interesantes y útiles. El aprendizaje automático ha seguido desarrollándose e innovando hasta el día de hoy. Algunos de los desarrollos más importantes incluyen *modelos de conjunto*, donde las predicciones se realizan utilizando un conjunto (o comité) de modelos, con cada modelo votando en cada consulta, y *redes neuronales de aprendizaje profundo*, que tienen múltiples (es decir, más de tres) capas de neuronas. Estas capas más profundas de la red pueden descubrir y aprender representaciones de atributos complejos (compuestos de múltiples atributos de entrada interactivos que han sido procesados por capas anteriores), que a su vez permiten a la red aprender patrones que se generalizan a través de los datos de entrada. Debido a su capacidad para aprender atributos complejos, las redes de aprendizaje profundo son particularmente adecuadas para datos de alta dimensión y, por lo tanto, han revolucionado una serie de campos, incluida la *visión artificial* y el *procesamiento del lenguaje natural*.

Como discutimos en nuestra revisión de la historia de la base de datos, los primeros años de la década de 1970 marcaron el comienzo de la tecnología de base de datos moderna con el modelo de datos relacionales de Edgar F. Codd y la posterior explosión de la generación y el almacenamiento de datos que condujeron al desarrollo del almacenamiento de datos en la década de 1990 y más recientemente al fenómeno del *big data*. Sin embargo, mucho antes de la aparición del *big data*, a fines de los años ochenta y principios de los noventa, era evidente la necesidad de un campo de investigación dirigido específicamente al análisis de estos grandes conjuntos de datos. Fue alrededor de esta época que el término *minería de datos* comenzó a usarse en las comunidades de bases de datos. Como ya hemos discutido, una respuesta a esta necesidad fue el desarrollo de almacenes de datos. Sin embargo, otros investigadores de bases de

datos respondieron contactándose con otros campos de investigación, y en 1989 Gregory Piatetsky-Shapiro organizó el primer taller sobre *descubrimiento de conocimiento en bases de datos* (KDD). El anuncio del primer taller de KDD resume claramente cómo el taller se centró en un enfoque multidisciplinario para el problema del análisis de grandes bases de datos.

El descubrimiento de conocimiento en bases de datos plantea muchos problemas interesantes, especialmente cuando las bases de datos son grandes. Dichas bases de datos suelen ir acompañadas de un conocimiento sustancial del dominio que puede facilitar significativamente el descubrimiento. El acceso a grandes bases de datos es costoso, de ahí la necesidad de muestreo y otros métodos estadísticos. Finalmente, el descubrimiento de conocimiento en bases de datos puede beneficiarse de muchas herramientas y técnicas disponibles de varios campos diferentes, incluidos sistemas expertos, aprendizaje automático, bases de datos inteligentes, adquisición de conocimiento y estadísticas.[1]

De hecho, los términos *descubrimiento de conocimiento en bases de datos* y *minería de datos* describen el mismo concepto, la distinción es que la minería de datos es más frecuente en las comunidades empresariales y el KDD es más frecuente en las comunidades académicas. Hoy en día, estos términos se usan indistintamente,[2] y muchos de los principales lugares académicos usan ambos. De hecho, la principal conferencia académica en este campo es la Conferencia Internacional sobre Descubrimiento de Conocimiento y Minería de Datos.

El surgimiento y la evolución de la ciencia de datos

El término *ciencia de datos* adquirió importancia a fines de la década de 1990 en discusiones relacionadas con la necesidad de que los estadísticos se unieran a los científicos informáticos para aportar rigor matemático al análisis computacional de grandes conjuntos de datos. En 1997, la conferencia pública de C. F. Jeff Wu "¿Estadísticas = Ciencia de datos?" destacó una serie de tendencias prometedoras para la estadística, incluida la disponibilidad de conjuntos de datos grandes/complejos en bases de datos masivas y el creciente uso de algoritmos y modelos computacionales. Concluyó la conferencia pidiendo que se cambiara el nombre de las estadísticas a "ciencia de datos".

En 2001, William S. Cleveland publicó un plan de acción para crear un departamento universitario en el campo de la ciencia de datos (Cleveland 2001). El plan enfatizaba la necesidad de que la ciencia de datos fuera una asociación entre las matemáticas y la informática. También enfatizaba la necesidad de que la ciencia de datos se entiendiera como un esfuerzo multidisciplinario y que los científicos de datos aprendieran cómo trabajar y relacionarse con expertos en la materia. En el mismo año, Leo Breiman publicó "Statistical Modeling: The Two Cultures" (2001). En este documento, Breiman caracteriza el enfoque tradicional de las estadísticas como una cultura de modelado de datos que considera que el objetivo principal del análisis de datos es identificar el modelo de datos estocástico (oculto) (por ejemplo, *regresión lineal*) que explica cómo se generaron los datos. Contrasta esta cultura con la cultura de modelado algorítmico que se enfoca en usar algoritmos de computadora para crear modelos de predicción que sean precisos (en lugar de explicativos, en términos de cómo se generaron los datos). La distinción de Breiman entre un enfoque estadístico en modelos que explican los datos versus un enfoque algorítmico en modelos que pueden predecir con precisión los datos destaca una diferencia central entre los estadísticos y los investigadores de aprendizaje automático. El debate entre estos enfoques todavía está en curso dentro de las estadísticas (véase, por ejemplo, Shmueli 2010). En general, hoy en día la mayoría de los proyectos de ciencia de datos están más alineados con el enfoque de aprendizaje automático de construir modelos de predicción precisos y menos preocupados por el enfoque estadístico en la explicación de los datos. Entonces, aunque la ciencia de datos se hizo prominente en las discusiones relacionadas con las estadísticas y todavía toma prestados métodos y modelos de las estadísticas, con el tiempo ha desarrollado su propio enfoque distinto para el análisis de datos.

Desde 2001, el concepto de ciencia de datos se ha ampliado mucho más allá de la redefinición de las estadísticas. Por ejemplo, en los últimos 10 años ha habido un enorme crecimiento en la cantidad de datos generados por la actividad en línea (venta minorista en línea, redes sociales y entretenimiento en línea). La recopilación y preparación de estos datos para su uso en proyectos de ciencia de datos ha resultado en la necesidad de que los científicos de datos desarrollen las habilidades de programación y piratería para extraer, fusionar y limpiar datos (a veces no estructurados) de fuentes web externas. Además, la aparición del *big data* ha significado que los científicos de datos deben poder trabajar con tecnologías del *big data*, como Hadoop. De hecho, hoy en día el papel de un científico de datos

se ha vuelto tan amplio que existe un debate continuo sobre cómo definir la experiencia y las habilidades necesarias para llevar a cabo esta función.[3] Sin embargo, es posible enumerar la experiencia y las habilidades que son relevantes para el rol en las que la mayoría de las personas están de acuerdo, que son las que se muestran en la Figura 1. Es difícil para un individuo dominar todas estas áreas y, de hecho, la mayoría de los científicos de datos generalmente tienen un conocimiento profundo y experiencia real en solo un subconjunto de ellos. Sin embargo, es importante comprender y estar al tanto de la contribución de cada área a un proyecto de ciencia de datos.

Los científicos de datos deberían tener cierta experiencia en el dominio. La mayoría de los proyectos de ciencia de datos comienzan con un problema específico del dominio del mundo real y la necesidad de diseñar una solución basada en datos para este problema. Como resultado, es importante que un científico de datos tenga suficiente experiencia en el dominio para comprender el problema, por qué es importante y cómo una solución de ciencia de datos al problema podría encajar en los procesos de una organización. Esta experiencia en el dominio guía al científico de datos mientras trabaja para identificar una solución optimizada.

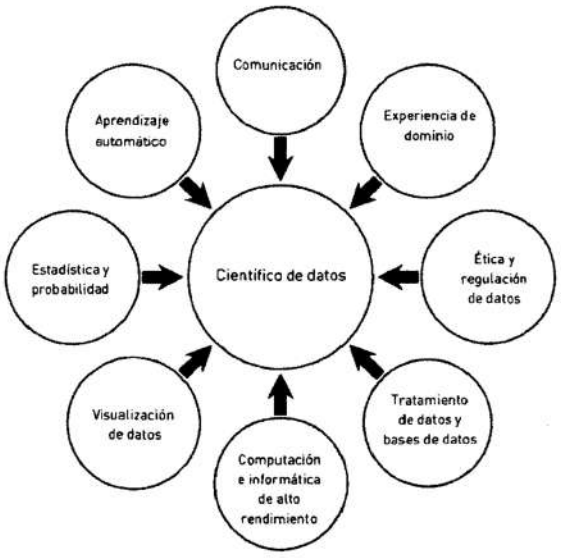

Figura 1. Un desiderátum de conjunto de habilidades para un científico de datos.

También le permite interactuar con expertos en dominios reales de una manera significativa para que pueda ilicitar y comprender el conocimiento relevante sobre el problema subyacente. Además, tener cierta experiencia en el dominio del proyecto le permite al científico de datos aportar sus experiencias al trabajar en proyectos similares en los mismos dominios y otros relacionados para definir el enfoque y el alcance del proyecto.

Los datos están en el centro de todos los proyectos de ciencia de datos. Sin embargo, el hecho de que una organización tenga acceso a los datos no significa que pueda usarlos legal o éticamente. En la mayoría de las jurisdicciones, existe una legislación antidiscriminatoria y de protección de datos personales que regula y controla el uso de la utilización de datos. Como resultado, un científico de datos necesita comprender estas regulaciones y también, en términos más generales, tener una comprensión ética de las implicaciones de su trabajo si quiere usar los datos de manera legal y adecuada. Volveremos a este tema en el capítulo 6, en el que discutimos las regulaciones legales sobre el uso de datos y las cuestiones éticas relacionadas con la ciencia de datos.

En la mayoría de las organizaciones, una parte importante de los datos provendrá de las bases de datos de la organización. Además, a medida que crece la arquitectura de datos de una organización, los proyectos de ciencia de datos comenzarán a incorporar datos de una variedad de otras fuentes de datos, que comúnmente se conocen como "fuentes de *big data*". Los datos en estas fuentes de datos pueden existir en una variedad de formatos diferentes, generalmente una base de datos de alguna forma: relacional, NoSQL o Hadoop. Todos los datos en estas diversas bases de datos y fuentes de datos deberán integrarse, limpiarse, transformarse, normalizarse, etc. Estas tareas tienen muchos nombres, como *extracción, transformación y carga, organización de datos, tratamiento de datos, fusión de datos, procesamiento de datos*, etc. Al igual que los datos de origen, los datos generados a partir de las actividades de ciencia de datos también deben almacenarse y administrarse. Una vez más, una base de datos es la ubicación de almacenamiento típica para los datos generados por estas actividades porque luego se pueden distribuir y compartir fácilmente con diferentes partes de la organización. Como consecuencia, los científicos de datos necesitan tener las habilidades para interactuar y manipular datos en bases de datos.

Una gama de habilidades y herramientas informáticas permite a los científicos de datos trabajar con grandes datos y procesarlos en información nueva y significativa. La *informática de alto rendimiento* (HPC en inglés) implica agregar potencia informática para ofrecer un rendimiento superior al que se puede obtener de una computadora independiente. Muchos proyectos de ciencia de datos funcionan con un conjunto de datos muy grande y algoritmos de aprendizaje automático que son costosos informáticamente. En estas situaciones, es importante tener las habilidades necesarias para acceder y utilizar los recursos de HPC. Más allá de HPC, ya hemos mencionado la necesidad de que los científicos de datos puedan eliminar, limpiar e integrar datos web, así como manejar y procesar texto e imágenes no estructurados. Además, un científico de datos también puede terminar escribiendo aplicaciones internas para realizar una tarea específica o alterar una aplicación existente para sintonizarla con los datos y el dominio que se está procesando. Finalmente, también se requieren habilidades informáticas para poder comprender y desarrollar los modelos de aprendizaje automático e integrarlos en la producción o aplicaciones analíticas o de fondo en una organización.

La presentación de datos en un formato gráfico hace que sea mucho más fácil ver y comprender lo que sucede con los datos. La visualización de datos se aplica a todas las fases del proceso de ciencia de datos. Cuando los datos se inspeccionan en forma de tabla, es fácil pasar por alto cosas como valores atípicos o tendencias en las distribuciones o cambios sutiles en los datos a través del tiempo. Sin embargo, cuando los datos se presentan en la forma gráfica correcta, estos aspectos de los datos pueden resaltar. La visualización de datos es un campo importante y en crecimiento, y recomendamos dos libros, *The Visual Display of Quantitative Information* de Edward Tufte (2001) y *Show Me the Numbers: Designing Tables and Graphs to Enlighten* de Stephen Few (2012) como una excelente introducción a los principios y técnicas de visualización efectiva de datos.

Métodos de estadística y probabilidad se utilizan en todo el proceso de ciencia de datos, desde la recopilación inicial y la investigación de los datos hasta la comparación de los resultados de diferentes modelos y análisis producidos durante el proyecto. El aprendizaje automático implica el uso de una variedad de técnicas avanzadas de estadística e informática para procesar datos para encontrar patrones. El científico de datos que participa en los aspectos aplicados del aprendizaje automático no tiene que escribir sus propias versiones de algoritmos de aprendizaje

automático. Al comprender estos algoritmos, para qué se pueden usar, qué significan los resultados que generan y qué tipo de algoritmos de datos particulares se pueden ejecutar, el científico de datos puede considerar los algoritmos de aprendizaje automático como un cuadro gris. Esto le permite concentrarse en los aspectos aplicados de la ciencia de datos y probar los diversos algoritmos para ver cuáles funcionan mejor para el escenario y los datos que le interesan.

Finalmente, un aspecto clave de ser un científico de datos exitoso es poder comunicar la historia en los datos. Esta historia podría descubrir el conocimiento que ha revelado el análisis de los datos o cómo los modelos creados durante un proyecto se ajustan a los procesos de una organización y el probable impacto que tendrán en el funcionamiento de la misma. No tiene sentido ejecutar un proyecto brillante de ciencia de datos a menos de que se utilicen y comuniquen los resultados de este de tal manera que los colegas con antecedentes no técnicos puedan comprenderlos y confiar en ellos.

¿Dónde se usa la ciencia de datos?

La ciencia de datos impulsa la toma de decisiones en casi todos los aspectos de las sociedades modernas. En esta sección, describimos tres estudios de caso que ilustran el impacto de la ciencia de datos: las compañías de consumo que usan la ciencia de datos para ventas y marketing; los gobiernos que utilizan la ciencia de datos para mejorar la salud, la justicia penal y la planificación urbana; y las franquicias deportivas profesionales que utilizan ciencia de datos en el reclutamiento de jugadores.

Ciencia de datos en ventas y marketing

Walmart tiene acceso a grandes conjuntos de datos sobre las preferencias de sus clientes mediante el uso de sistemas de punto de venta, rastreando el comportamiento del cliente en el sitio web de Walmart y los comentarios de las redes sociales sobre Walmart y sus productos. Durante más de una década, Walmart ha estado utilizando la ciencia de datos para optimizar los niveles de stock en las tiendas, un ejemplo bien conocido es cuando en 2004 reabasteció con Pop-Tarts de fresas sus tiendas en la ruta del huracán Francis en base a un análisis de datos de ventas previos al huracán

Charley, que había golpeado unas semanas antes. Más recientemente, Walmart ha utilizado la ciencia de datos para impulsar sus ingresos minoristas en términos de introducir nuevos productos basados en el análisis de las tendencias de las redes sociales, el análisis de la actividad de las tarjetas de crédito para hacer recomendaciones de productos a los clientes y la optimización y personalización de la experiencia en línea de los clientes en el sitio web de Walmart. Walmart atribuye un aumento del 10% al 15% en las ventas en línea a las optimizaciones de ciencia de datos (DeZyre 2015).

El equivalente de ventas superiores y ventas cruzadas en el mundo en línea es el "sistema de recomendación". Si has visto una película en Netflix o has comprado un artículo en Amazon, sabrás que estos sitios web utilizan los datos que recopilan para proporcionar sugerencias sobre lo que debes ver o comprar a continuación. Estos sistemas de recomendación se pueden diseñar para guiarte de diferentes maneras: algunos te guían hacia éxitos de taquilla y bestsellers, mientras que otros te guían hacia artículos de nicho que son específicos para tus gustos. El libro de Chris Anderson, *La Economía Long Tail* (2008), argumenta que a medida que la producción y la distribución se vuelven menos costosas, los mercados pasan de vender grandes cantidades de un pequeño número de artículos exitosos a vender cantidades más pequeñas de un mayor número de artículos de nicho. Esta compensación entre impulsar las ventas de productos exitosos o de nicho es una decisión de diseño fundamental para un sistema de recomendación y afecta los algoritmos de ciencia de datos utilizados para implementar estos sistemas.

Gobiernos que usan ciencia de datos

En los últimos años, los gobiernos han reconocido las ventajas de adoptar la ciencia de datos. En 2015, por ejemplo, el gobierno de Estados Unidos nombró al Dr. D. J. Patil como el primer científico de datos en jefe. Algunas de las mayores iniciativas de ciencia de datos encabezadas por el gobierno de Estados Unidos han estado en salud. La ciencia de datos está en el centro de las iniciativas "Cancer Moonshot"[4] y "Precision Medicine". La iniciativa "Precision Medicine" [Medicina de precisión] combina la secuenciación del genoma humano y la ciencia de datos para diseñar medicamentos para pacientes individuales. Una parte de la iniciativa es el programa "All of Us" [Todos nosotros],[5] que recopila datos ambientales,

de estilo de vida y biológicos de más de un millón de voluntarios para crear los conjuntos de datos más grandes del mundo para la medicina de precisión. La ciencia de datos también está revolucionando la forma en que organizamos nuestras ciudades: se utiliza para rastrear, analizar y controlar los sistemas ambientales, de energía y de transporte e informar la planificación urbana a largo plazo (Kitchin 2014a). Volveremos al tema de la salud y las ciudades inteligentes en el capítulo 7, en el que discutiremos cómo la ciencia de datos será aún más importante en nuestras vidas en las próximas décadas.

La iniciativa de datos policiales del gobierno de EE.UU.[6] se centra en el uso de la ciencia de datos para ayudar a los departamentos de policía a comprender las necesidades de sus comunidades. La ciencia de datos también se está utilizando para predecir los puntos críticos del crimen y la reincidencia. Sin embargo, los grupos de libertad civil han criticado algunos de los usos de la ciencia de datos en la justicia penal. En el capítulo 6, discutiremos las preguntas de privacidad y ética planteadas por la ciencia de datos, y uno de los factores interesantes en esta discusión es que las opiniones que las personas tienen en relación con la privacidad personal y la ciencia de datos varían de un dominio a otro. Muchas personas que están contentas de que sus datos personales sean utilizados para investigaciones médicas financiadas con fondos públicos tienen opiniones muy diferentes cuando se trata del uso de datos personales para la vigilancia y la justicia penal. En el capítulo 6, también discutiremos el uso de datos personales y ciencia de datos para determinar las primas de seguros de vida, salud, automóvil, hogar y viajes.

Ciencia de datos en deportes profesionales

La película *Moneyball* (Bennett Miller 2011), protagonizada por Brad Pitt, muestra el creciente uso de la ciencia de datos en los deportes modernos. La película se basa en el libro del mismo título (Lewis 2004), que cuenta la verdadera historia de cómo el equipo de béisbol Oakland Athletics utilizó la ciencia de datos para mejorar su reclutamiento de jugadores. La gerencia del equipo identificó que las estadísticas de porcentaje en base y el poder de un bateador eran indicadores más informativos del éxito ofensivo que las estadísticas tradicionalmente enfatizadas en el béisbol, como el promedio de bateo de un jugador. Esta idea permitió a Oakland Athletics reclutar una lista de jugadores infravalorados y tener un desempeño por

encima de su presupuesto. El éxito de Oakland Athletics con la ciencia de datos ha revolucionado el béisbol, y la mayoría de los otros equipos de béisbol ahora integran estrategias similares basadas en datos en sus procesos de reclutamiento.

La historia de *Moneyball* es un ejemplo muy claro de cómo la ciencia de datos puede dar a una organización una ventaja en un espacio de mercado competitivo. Sin embargo, desde una perspectiva de ciencia de datos pura, quizás el aspecto más importante de la historia de *Moneyball* es que destaca que a veces el valor principal de la ciencia de datos es la identificación de atributos informativos. Una creencia común es que el valor de la ciencia de datos está en los modelos creados a través del proceso. Sin embargo, una vez que conocemos los atributos importantes en un dominio, es muy fácil crear modelos basados en datos. La clave del éxito es obtener los datos correctos y encontrar los atributos correctos.

La clave del éxito es obtener los datos correctos y encontrar los atributos correctos.

En *Freakonomics: Un economista politicamente incorrecto explora el lado oculto de lo que nos afecta*, Steven D. Levitt y Stephen Dubner ilustran la importancia de esta observación en una amplia gama de problemas. Como lo expresaron, la clave para entender la vida moderna es "saber qué medir y cómo medirlo" (2009, 14). Mediante la ciencia de datos podemos descubrir los patrones importantes en un conjunto de datos, y estos patrones pueden revelar los atributos importantes en el dominio. La razón por la cual la ciencia de datos se usa en tantos dominios es que no importa cuál sea el dominio del problema: si los datos correctos están disponibles y el problema se puede definir claramente, entonces la ciencia de datos puede ayudar.

¿Por qué ahora?

Varios factores han contribuido al reciente crecimiento de la ciencia de datos. Como ya hemos mencionado, la aparición del *big data* ha sido impulsada por la relativa facilidad con la que las organizaciones pueden recopilar datos. Ya sea a través de registros de transacciones de punto de venta, clics en plataformas en línea, publicaciones en redes sociales, aplicaciones en teléfonos inteligentes u otros miles de canales, las compañías

ahora pueden crear perfiles mucho más ricos de clientes individuales. Otro factor es la mercantilización del almacenamiento de datos con economías de escala, lo que hace que almacenar datos sea más barato que nunca. También ha habido un tremendo crecimiento en la potencia informática. Las tarjetas gráficas y las unidades de procesamiento gráfico (GPU en inglés) se desarrollaron originalmente para hacer una representación gráfica rápida para juegos de computadora. La característica distintiva de las GPU es que pueden llevar a cabo multiplicaciones rápidas de matrices. Sin embargo, las multiplicaciones de matrices son útiles no solo para la representación gráfica, sino también para el aprendizaje automático. En los últimos años, las GPU se han adaptado y optimizado para el uso del aprendizaje automático, lo que ha contribuido a grandes aceleraciones en el procesamiento de datos y el entrenamiento de modelado. También se han vuelto disponibles herramientas de ciencia de datos fáciles de usar y se han reducido las barreras para ingresar a la ciencia de datos. En su conjunto, estos desarrollos significan que nunca ha sido tan fácil recopilar, almacenar y procesar datos.

En los últimos 10 años también ha habido avances importantes en el aprendizaje automático. En particular, ha surgido el aprendizaje profundo y ha revolucionado la forma en que las computadoras pueden procesar el lenguaje y los datos de imágenes. El término *aprendizaje profundo* describe una familia de modelos de redes neuronales con múltiples capas de unidades en la red. Las redes neuronales han existido desde la década de 1940, pero funcionan mejor con conjuntos de datos grandes y complejos y requieren una gran cantidad de recursos informáticos para entrenar. Por lo tanto, la aparición del aprendizaje profundo está relacionada con el crecimiento en el *big data* y la potencia informática. No es una exageración describir el impacto del aprendizaje profundo en una variedad de dominios como nada menos que extraordinario.

El programa informático AlphaGo[7] de DeepMind es un excelente ejemplo de cómo el aprendizaje profundo ha transformado un campo de investigación. Go es un juego de mesa que se originó en China hace 3.000 años. Las reglas de Go son mucho más simples que el ajedrez; los jugadores se turnan para colocar piezas en un tablero con el objetivo de capturar las piezas de su oponente o el territorio vacío circundante. Sin embargo, la simplicidad de las reglas y el hecho de que Go usa un tablero más grande significa que hay muchas más configuraciones de tablero posibles que en ajedrez. De hecho, hay más configuraciones de tablero posibles en Go

que átomos en el universo. Esto hace que Go sea mucho más difícil que el ajedrez para computadoras debido a su espacio de búsqueda mucho más grande y a la dificultad de evaluar cada una de estas posibles configuraciones de tablero. El equipo de DeepMind utilizó modelos de aprendizaje profundo para permitir a AlphaGo evaluar las configuraciones de tablero y seleccionar el siguiente movimiento a realizar. El resultado fue que AlphaGo se convirtió en el primer programa informático en vencer a un jugador profesional de Go, y en marzo de 2016 AlphaGo venció a Led Sedol, el 18 veces campeón mundial de Go, en un partido visto por más de 200 millones de personas en todo el mundo. Para poner en contexto el impacto del aprendizaje profundo en Go: en 2009, el mejor programa informático Go en el mundo fue calificado en el extremo inferior de los aficionados avanzados; siete años después AlphaGo venció al campeón mundial. En 2016, se publicó un artículo que describía los algoritmos de aprendizaje profundo detrás de AlphaGo en la revista científica académica más prestigiosa del mundo, *Nature* (Silver, Huang, Maddison *et al.* 2016).

El aprendizaje profundo también ha tenido un impacto masivo en una gama de tecnologías de consumo de alto perfil. Facebook ahora utiliza el aprendizaje profundo para el reconocimiento de rostros y para analizar textos con el fin de publicitar directamente a las personas en función de sus conversaciones en línea. Tanto Google como Baidu utilizan el aprendizaje profundo para el reconocimiento de imágenes, subtítulos y búsqueda, y traducción automática. Las asistentes virtuales Siri de Apple, Alexa de Amazon, Cortana de Microsoft y Bixby de Samsung utilizan el reconocimiento de voz basado en el aprendizaje profundo. Huawei está desarrollando actualmente un asistente virtual para el mercado chino, y también utilizará el reconocimiento de voz de aprendizaje profundo. En el capítulo 4, "Introducción al aprendizaje automático", describiremos las redes neuronales y el aprendizaje profundo con más detalle. Sin embargo, aunque el aprendizaje profundo es un desarrollo técnico importante, quizás lo más significativo en términos del crecimiento de la ciencia de datos es la mayor conciencia de las capacidades y beneficios de la ciencia de datos y la aceptación de las organizaciones, que ha sido producto de estas historias de éxito de alto perfil.

Mitos sobre la ciencia de datos

La ciencia de datos tiene muchas ventajas para las organizaciones modernas, pero también hay una gran expectación en torno a ella, por lo que debemos entender cuáles son sus limitaciones. Uno de los mitos más importantes es la creencia de que la ciencia de datos es un proceso autónomo que podemos hacer correr en nuestros datos para encontrar las respuestas a nuestros problemas. En realidad, la ciencia de datos requiere una supervisión humana especializada en las diferentes etapas del proceso. Se necesitan analistas humanos para enmarcar el problema, diseñar y preparar los datos, seleccionar qué algoritmos de aprendizaje automático son los más apropiados, interpretar críticamente los resultados del análisis y planificar la acción adecuada a tomar en función de la información que el análisis ha revelado. Sin supervisión humana calificada, un proyecto de ciencia de datos no podrá cumplir sus objetivos. Los mejores resultados de la ciencia de datos ocurren cuando la experiencia humana y la potencia informática trabajan juntas, como lo expresaron Gordon Linoff y Michael Berry: "La minería de datos permite que las computadoras hagan lo que mejor saben hacer: excavar entre una gran cantidad de datos. Esto, a su vez, permite que las personas hagan lo que mejor hacen, que es configurar el problema y comprender los resultados" (2011, 3).

El uso generalizado y creciente de la ciencia de datos significa que hoy el mayor desafío de la ciencia de datos para muchas organizaciones es localizar analistas humanos calificados y contratarlos. El talento humano en la ciencia de datos es muy importante, y el abastecimiento de este talento es actualmente el principal cuello de botella en la adopción de la ciencia de datos. Para poner en contexto este déficit de talento, en 2011 un informe del Instituto Global McKinsey proyectó un déficit en Estados Unidos de entre 140.000 y 190.000 personas con habilidades de análisis y ciencia de datos y un déficit aún mayor de 1,5 millones de gerentes con la capacidad de comprender la ciencia de datos y procesos analíticos a un nivel que les permita interrogar e interpretar los resultados de la ciencia de datos de manera adecuada (Manyika, Chui, Brown *et al.* 2011). Cinco años después, en su informe de 2016, el instituto seguía convencido de que la ciencia de datos tiene un enorme potencial de valor sin explotar en una gama cada vez mayor de aplicaciones, pero que el déficit de talento se mantendrá, con un déficit previsto de 250.000 científicos de datos a corto plazo (Henke, Bughin, Chui y col. 2016)

El segundo gran mito de la ciencia de datos es que cada proyecto de ciencia de datos necesita *big data* y utilizar el aprendizaje profundo. En general, tener más datos ayuda, pero tener los datos *correctos* es el requisito más importante. Los proyectos de ciencia de datos se llevan a cabo con frecuencia en organizaciones que tienen significativamente menos recursos en términos de datos y potencia informática que Google, Baidu o Microsoft. Los ejemplos indicativos de la escala de los proyectos de ciencia de datos más pequeños incluyen la predicción de reclamos en una compañía de seguros que procesa alrededor de 100 reclamos por mes; predicción de abandono estudiantil en una universidad con menos de 10.000 estudiantes; predicción de abandono de membresía en un sindicato con varios miles de miembros. Por lo tanto, una organización no necesita manejar terabytes de datos o tener recursos informáticos masivos a su disposición para beneficiarse de la ciencia de datos.

Un tercer mito de la ciencia de datos es que el software moderno de ciencia de datos es fácil de usar, por lo que la ciencia de datos es fácil de hacer. Es cierto que el software de ciencia de datos se ha vuelto más fácil de usar. Sin embargo, esta facilidad de uso puede ocultar el hecho de que hacer ciencia de datos correctamente requiere tanto el conocimiento de dominio apropiado como la experiencia con respecto a las propiedades de los datos y los supuestos que sustentan los diferentes algoritmos de aprendizaje automático. De hecho, nunca ha sido tan fácil hacer mal la ciencia de datos. Como todo lo demás en la vida, si no comprendes lo que haces cuando haces ciencia de datos, cometerás errores. El peligro con la ciencia de datos es que la tecnología puede intimidar a las personas y hacerlos confiar en cualquier resultado que el software les presente. Sin embargo, pueden haber enmarcado el problema mal, haber ingresado los datos incorrectos o haber utilizado técnicas de análisis con suposiciones inapropiadas. Por lo tanto, es probable que los resultados que presenta el software sean la respuesta a la pregunta incorrecta o que se basen en los datos incorrectos o en un cálculo incorrecto.

El último mito sobre la ciencia de datos que queremos mencionar aquí es la creencia de que la ciencia de datos se amortiza rápidamente. La verdad de esta creencia depende del contexto de la organización. La adopción de la ciencia de datos puede requerir una inversión significativa en términos de desarrollo de infraestructura de datos y contratación de personal con experiencia en ciencia de datos. Además, la ciencia de datos no dará resultados positivos en cada proyecto. A veces no hay una gema oculta de

conocimiento en los datos, y a veces la organización no está en condiciones de actuar sobre el conocimiento revelado por el análisis. Sin embargo, en contextos donde hay un problema comercial bien entendido y los datos apropiados y la experiencia humana están disponibles, entonces la ciencia de datos puede (a menudo) proporcionar una visión procesable que le da a una organización la ventaja competitiva que necesita para tener éxito.