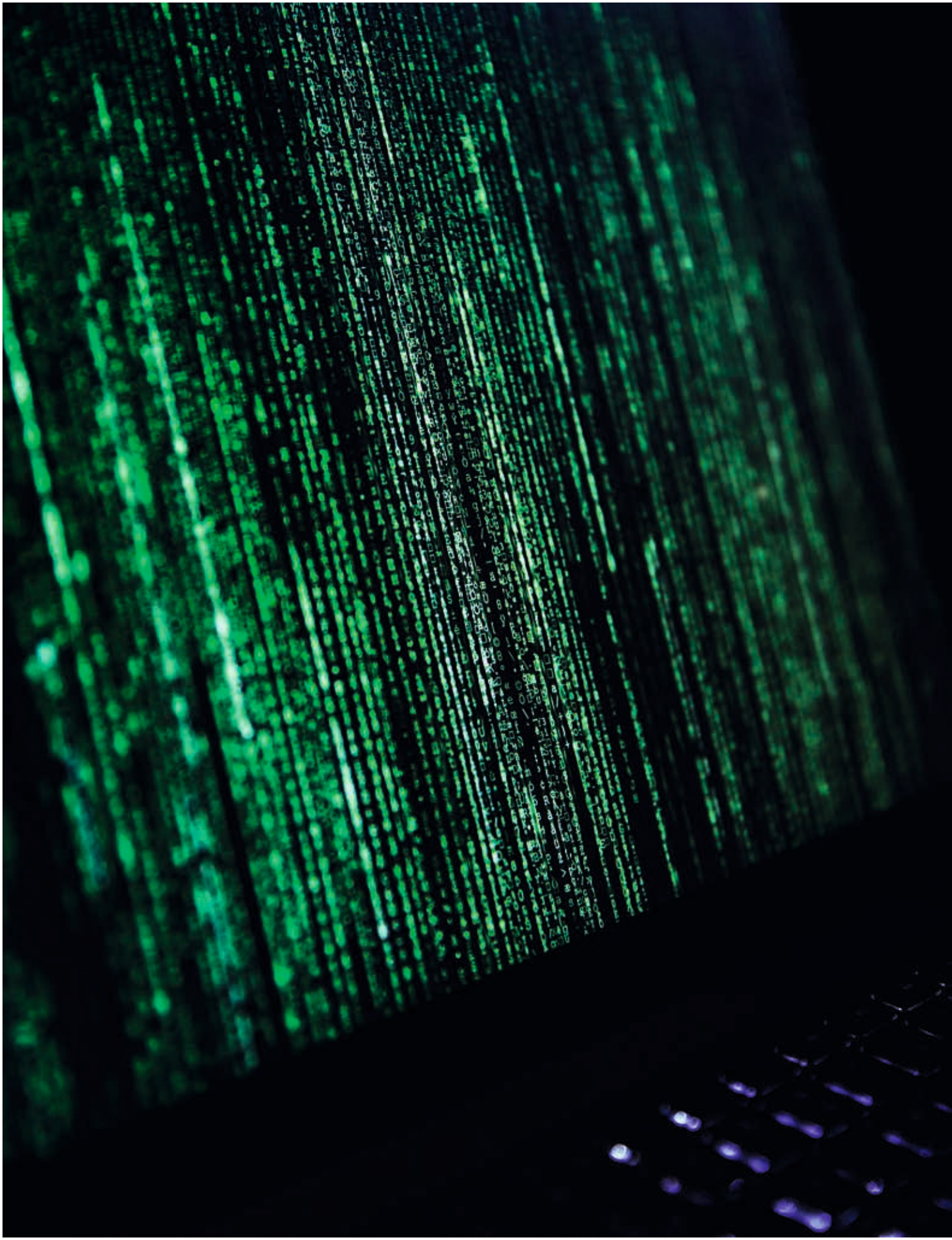


Pxhere

TODO LO QUE SABEN DE NOSOTROS

¿SE PUEDE NAVEGAR CON PRIVACIDAD POR EL OCÉANO DEL 'BIG DATA'?

Xavier Duran

«El uso de ordenadores en la investigación sobre el estatus financiero, médico, mental y moral de los ciudadanos es una gran –y creciente» industria. La frase parece provenir de alguno de los textos que desde hace cierto tiempo llaman la atención sobre el negocio que significa la búsqueda y procesamiento de datos personales. Pero se escribió hace muchos años. Se encuentra en el libro *Electronic nightmare. The new communications and freedom* ("La pesadilla electrónica. Las nuevas comunicaciones y la libertad"), publicado en 1981 por John Wicklein (Wicklein, 1981, p. 191), destacado periodista y experto en comunicación que había sido redactor y editor en *The New York Times*, y que había tenido responsabilidades en programas de noticias en varias cadenas de televisión. La cita sirve para demostrar que el problema de la privacidad amenazada por la informática y las telecomunicaciones no es un tema nuevo. Pero si en 1981 preocupaba solo a personas bien informadas como Wicklein –y muchos gobiernos–, en los últimos años la potencia de las herramientas informáticas y la facilidad de la transmisión de datos le han dado una nueva dimensión y la preocupación se ha extendido.

«HOY, EN TODO EL MUNDO, EN UN SOLO DÍA CIRCULA MUCHÍSIMA INFORMACIÓN, MUCHA MÁS QUE EN CUALQUIER OTRO MOMENTO DE LA HISTORIA»

En la obra, Wicklein no solo hace una exposición muy rigurosa e informada sobre los inicios de la televisión interactiva, el videotexto o los diarios electrónicos, debatiendo aspectos legales, económicos, políticos y sociales, sino que se muestra extraordinariamente lúcido para detallar una situación hipotética (Wicklein, 1981, p. 28–30). La sitúa en 1994, «cuando el cable de dos sentidos, interactivo, se ha extendido a centenares de ciudades por todo el país» –en eso se quedó corto–. Habla de una ficticia Martha Johnson, candidata a desbancar al alcalde de la no menos ficticia ciudad de San Serra, al sur de California. Aprovechando la televisión interactiva, Johnson encarga un libro que aboga por abolir las leyes que dificultan la actividad sexual consentida entre adultos. También compra un desodorante en espray. Cuando se va, su marido utiliza el mismo aparato para responder a una encuesta, donde se muestra contrario a permitir que las lesbianas enseñen en escuelas públicas. Después, contrata una película pornográfica. También hace una transferencia, porque ha recibido el aviso de que no pueden enviar el libro encargado por su mujer porque tienen pagos atrasados.

La compañía de cable ha ido reuniendo información. Y tiene personas a las que les interesa. El alcalde y contrincante de Martha Johnson se entera de que el marido de esta ha respondido una encuesta en la que muestra una opinión contraria a la que ella defiende, y que ha contratado una película pornográfica. Sabe que ha sido él porque el monitor instalado frente a la casa de los Johnson le había permitido saber que la mujer había salido y no había vuelto. Otro cliente es la editorial de una revista porno, que envía al señor Johnson un ejemplar en un sobre discreto. Un grupo ecologista sabe ahora que la candidata compra desodorantes en espray que destruyen la capa de ozono. Y una empresa de calificación crediticia se entera de que se ha rechazado una compra a los Johnson, lo que incluirá en su expediente sobre el matrimonio.

En 1981, la hipótesis podría parecer exagerada. En 1994 quizá también, pero no tanto. Y el 2018 vemos que, aún siendo un buen visionario, Wicklein incluso se quedó corto.

¿CUÁNTA INFORMACIÓN CIRCULA POR LA RED Y QUÉ SE PUEDE HACER?

No sabemos cuánta información circulaba por el mundo en 1981, pero solo una pequeña parte debía viajar en formato electrónico. Sí que tenemos una idea aproximada de cuánta información había en todo el mundo en 1997 gracias al informático Michael Lesk. Calculó que en total había 12.000 *petabytes* (PB) de información (Lesk, 1997) –12 millones de gigas, para hacerlo un poco más comprensible o menos incomprensible–. Actualmente

El concepto de *big data* comprende no solo la recolección y acumulación de los datos que los usuarios comparten en Internet, sino también el procesamiento de estos. La creciente potencia de las herramientas informáticas que lo hacen posible y la facilidad con las que se transmite esta información genera posibilidades inmensas, pero también muchos riesgos para la privacidad de la ciudadanía.

solo en una hora se transmiten 500 PB, equivalentes a 6.600 años de vídeo de alta definición. Eso significa que en un solo día ya se transmiten los 12.000 PB calculados por Lesk. Los formatos son muy diversos. Cada minuto se envían unos 200 millones de correos electrónicos y se cuelgan medio millón de comentarios en Facebook. En un día se publican más de setenta millones de fotos en Instagram. Añadamos consultas en Internet, mensajes y llamadas de teléfonos móviles, tuits, otras redes sociales, imágenes grabadas por cámaras de vigilancia, por cámaras particulares, imágenes de satélite, transacciones bancarias, datos clínicos... Quizá no haga falta averiguar si el cálculo del número de *petabytes* es bastante aproximado o no. La conclusión simplemente es que hoy, en todo el mundo, en un solo día circula muchísima información, mucha más que en cualquier otro momento de la historia.

Erik Lucatero

La cantidad ingente de información que circula por la red adopta muchos formatos: cada minuto se envían unos 200 millones de correos electrónicos y se escriben medio millón de comentarios en Facebook. En un día se cuelgan más de setenta millones de fotos en Instagram. A eso habría que añadir consultas en buscadores, comentarios en otras redes sociales, transacciones bancarias, mensajería instantánea...

¿Qué se puede hacer con tanta información? Cosas muy positivas. Y también cosas muy negativas o que requieren fuertes controles. Este gran número de datos y su procesamiento se llama *big data*. A menudo, la expresión no se traduce, y cuando se hace, el equivalente es *macrodatos* o *datos masivos*. El *big data* no comprende solo los datos, sino también las herramientas informáticas para procesarlos. Una cosa es inútil sin la otra.

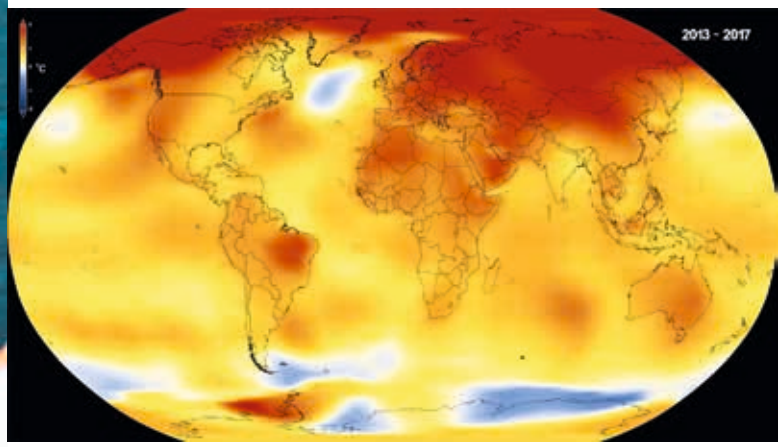
No hay una definición concreta de *big data*. No está claro a partir de cuándo hay macrodatos y cuándo hay, simplemente, muchos datos. En todo caso, al *big data* sí que se le otorgan unas características. Las tres V del *big data* señalan que tiene que tener volumen, velocidad y variedad. La primera condición se cumple claramente, como hemos visto. La segunda, con las tecnologías actuales de producción y transmisión de datos, también. Y diríamos que tampoco hay que extenderse en justificar la tercera. A menudo se añaden dos V más: veracidad y valor. Es necesario que los datos sean fiables, lo que no siempre se produce. Y es necesario que tengan valor, y eso último es más fácil de demostrar.

Así, el *big data* tiene una gran importancia en los ámbitos más diversos. Cualquier campo donde sea necesario tener muchos datos y procesarlos se beneficia de las técnicas de *big data*. Ordenadores potentes aumentan la velocidad de cálculo y algoritmos complejos permiten transformar los datos en información útil. Los datos pueden ser muy científicos –interacciones entre partículas en un acelerador, posiciones y datos astrofísicos de estrellas y galaxias...– o muy cotidianos –la gestión del tráfico o del servicio de limpieza en una ciudad–. Todos los ámbitos de la investigación se benefician de los macrodatos. Un estudio sobre las mutaciones genéticas que se dan en un cierto tipo de cáncer y que no aparecen en pacientes sanos, relacionadas también con estilos de vida y otras características de cada persona, no se podría hacer sin el *big data* y sin un supercomputador. Los estudios climáticos serían mucho menos útiles si no pudiésemos procesar con cierta velocidad inmensas cantidades de datos.

DATOS QUE PROPORCIONAMOS SIN SABERLO

Pero los datos tienen valor en muchos ámbitos más, y no siempre se conoce quién los utiliza, cómo y para qué. Si con una tecnología aún limitada y poco extendida Wicklein imaginaba aquel perjuicio para los Johnson, imaginemos qué se puede saber hoy de nosotros.

Se pueden averiguar cosas sobre nosotros a partir de datos que proporcionamos alegremente y de forma voluntaria. La mayoría hace búsquedas en Internet con buscadores que preservan los datos de lo que hemos buscado y qué webs hemos visitado –son las llamadas *cookies*–, reenviamos tuits que pueden ser vistos por millones de personas y que pueden quedar registrados, colgamos comentarios en Facebook o fotos en Instragram. No todo el mundo hace todo eso ni muy a menudo, pero en todo caso son datos que entregamos para poder tener



Estudio de Visualizaciones Científicas de la NASA

El análisis de datos masivos es una herramienta necesaria en muchos ámbitos de la investigación científica; por ejemplo, en los estudios climáticos. En la imagen, gráfico elaborado por el Instituto Goddard de Estudios Espaciales de la NASA, que muestra el aumento de temperaturas a escala global en el período entre 2013 y 2017. El gráfico forma parte de un vídeo donde pueden visualizarse los cambios de temperatura desde 1880, año en que se iniciaron los registros. En la elaboración de esta infografía se incorporaron los datos provenientes de 6.300 estaciones meteorológicas de todo el mundo.

o proporcionar información, para comentar cosas o por pura moda o narcisismo. Un consejo es ser discreto y cauto. Sin embargo, como veremos más adelante, eso no es suficiente. El uso que hizo la consultora Cambridge Analytics de los perfiles de 87 millones de usuarios de Facebook es un buen ejemplo de que no sabemos dónde irá a parar la información sobre nosotros.

En 2015, unos investigadores quisieron crear un algoritmo que pudiese juzgar nuestra personalidad a través del rastro que dejamos en Facebook (Youyou, Kosinski y Stilwell, 2015). Lo compararon con unos cuestionarios rellenados por compañeros, amigos, parejas y familiares. Y vieron que basándose solo en diez «Me gusta», el algoritmo conocía al usuario tan bien como un compañero de trabajo. Con 70 «Me gusta», ya era tan preciso como lo podía ser un compañero de habitación. Con 150, ya le conocía tan bien como un progenitor o un hermano. Y con 300 era tan exacto como lo podía ser su pareja. Solo con eso, con una pequeña parte del rastro que dejamos en Internet, ya se puede elaborar un perfil con bastante aproximación.

Quizá eso parecerá poco importante comparado con un estudio anterior (Kosinski, Stilwell y Graepei, 2013) donde se había demostrado la capacidad de estos algoritmos para adivinar otras cosas. Elaborado con más de 58.000 voluntarios, permitió constatar que el algoritmo era capaz de predecir, en un 88 % de los casos, si el usuario era heterosexual o homosexual, en un 95 % si era afroamericano o caucásico, en un 85 % si era republicano o demócrata, en un 82 % si era cristiano o musulmán. Incluso se creyeron capaces de acertar en un 78 % de los casos si era inteligente y en un 68 % si era emocionalmente inestable.

Cuando se alerta sobre las amenazas a la privacidad, mucha gente aduce que recibir publicidad o propuestas comerciales no es un problema grave. Dicen que puedes hacer caso o no y ya está. Es posible, aunque puede llegar a ser un poco engorroso. Pero con los datos masivos ya no se trata de eso, sino de información mucho más sensible y que todo el mundo tiene derecho a mantener en secreto si lo desea, como la ideología, la religión o las preferencias sexuales. Hay países donde la homosexualidad es delito y puede significar incluso una condena a muerte. Que un algoritmo pueda predecir si una persona es homosexual no resulta inocente ni inocuo, y lo mismo pasa con la religión o con las opiniones políticas o el activismo social.

«CUALQUIER CAMPO DONDE SEA NECESARIO TENER MUCHOS DATOS Y PROCESARLOS SE BENEFICIA DE LAS TÉCNICAS DE ‘BIG DATA’»

INFORMACIÓN PARA MUTUAS, ASEGURADORAS, EMPRESAS...

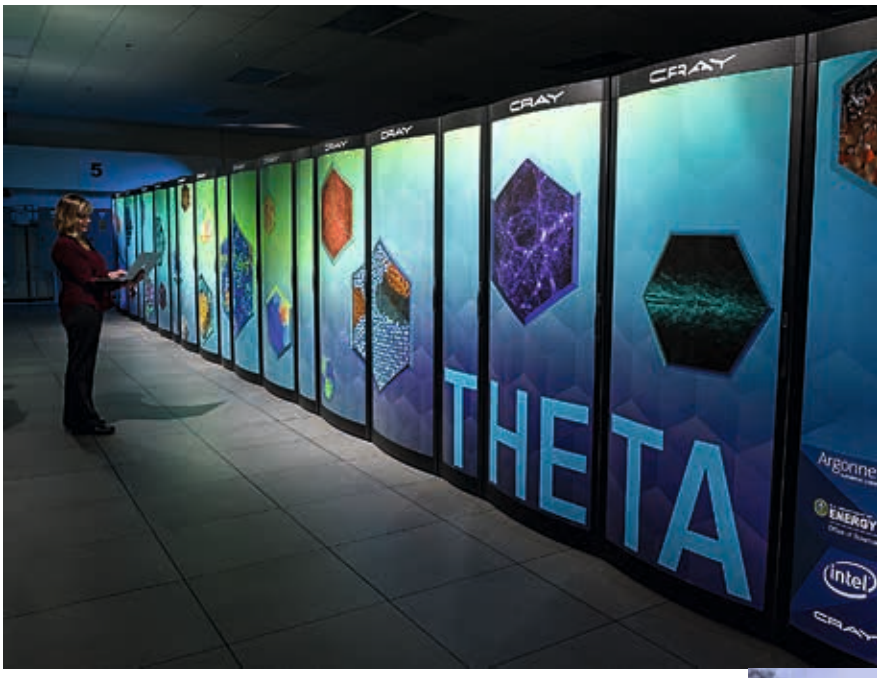
Hay muchas otras cosas que se pueden deducir de nuestros datos que no afectan solo a un colectivo, como el homosexual. Todos generamos de una manera u otra información médica. Ya no se trata de todo lo que tienen de nosotros en sus archivos médicos y hospitales y que en cierto momento puede ser robado por ciberdelincuentes. Hay otras posibilidades. Si hacemos una búsqueda sobre una enfermedad determinada, Google sabe que la hemos hecho. Si compramos un libro sobre cierta enfermedad, Amazon sabe que la hemos hecho. Si opinamos sobre determinado trastorno o buscamos gente que lo sufre o familias de gente que lo sufre, Internet lo sabe.

Esta información que circula y que alguien conserva, relacionada con muchas otras informaciones sobre nuestras actividades, puede hacer creer a un algoritmo que tenemos la enfermedad o que nos preocupa poderla tener en el futuro. Quizá simplemente lo buscamos por curiosidad o por otras razones. Pero si la información llega –y puede llegar perfectamente– a aseguradoras o mutuas, podemos tener problemas para suscribir pólizas o seguros de vida. Y si llega a empresas en general, pueden tener mucha información sobre los posibles problemas médicos de un candidato a un puesto de trabajo. Hay empresas que acumulan datos médicos y no médicos de millones de per-

sonas y que las utilizan para aceptar o denegar una solicitud de asistencia médica (Allen, 2018). Utilizan algoritmos que no son infalibles.

Un algoritmo es una serie de instrucciones ordenadas que permiten afrontar un determinado problema u operación. Los hay muy sencillos, como el que permite comprobar si hemos hecho bien una división: cociente por divisor más residuo tiene que dar el dividendo. Los hay muy complejos, que tienen cientos de líneas de códigos informáticos. El buscador de Google debe tener unos miles de millones de líneas de código.

Los algoritmos no piensan. Los algoritmos siguen instrucciones y asimilan datos. Pero alguien tiene que haberles dicho qué datos deben asimilar. Los datos los encontrarán en la red o en archivos o publicaciones que reflejan aproximadamente lo que es la sociedad (Duran, 2018, p. 147–165). Los diseñan y elaboran personas y tienen los tics de los que los elaboran. Aprenden a partir de la información que se proporciona o que un sistema inteligente puede recoger. No entienden de matices si no se los entrena para tenerlos en cuenta, y por eso pueden llegar a conclusiones curiosas (véanse algunos ejemplos en O'Neil, 2016/2018).



Laboratorio Nacional de Argonne (EE UU)

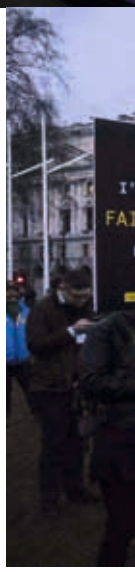
Los datos necesitan ser procesados para resultar útiles y es aquí donde entran en juego los supercomputadores como el de la imagen. Se trata del supercomputador Theta, propiedad del Laboratorio Nacional de Argonne en los Estados Unidos. Equipos científicos de todo el país pueden acudir a estas instalaciones a hacer uso de sus recursos en estudios donde el análisis de datos es clave.

Además, muchos algoritmos no están validados. Y las empresas los suelen esconder porque consideran que son una propiedad intelectual que hay que preservar de la competencia. El resultado es que son cajas negras (Pasquale, 2015) donde se nos juzga por lo que dicen nuestros datos y no con una certeza absoluta, sino simplemente con probabilidades. Juzgan nuestro posible gasto médico según el barrio donde vivimos, los estudios de nuestros padres o nuestras amistades y los comentarios que estas hacen. No somos individuos, sino estadísticas procesadas. Y sobre eso se toman decisiones vitales.

DEL ROSTRO A LOS GENES

Las posibilidades del *big data* para aportar cosas positivas crecen cada día. Hay investigaciones en los ámbitos más diversos que hacen progresar la ciencia y la gestión del ámbito público o que nos pueden hacer la vida más fácil. Pero también aumentan las posibilidades tanto de captar datos sin que lo sepamos como de aplicarlos con intenciones poco claras que, además, desconocemos. Y el desconocimiento causa indefensión.

Así, los sistemas de reconocimiento de rostros son cada vez más precisos: pueden identificar gente en fotos o vídeos –gente que quizá no quiere ser identificada– y relacionar el rostro y el nombre con otras informaciones que se encuentran en la red. A veces, ni tan solo hacen falta estos sistemas. Facebook pide a los usuarios que etiqueten a la gente que conozcan que aparece en las imágenes. ¡Ya no hace falta ni utilizar Facebook para estar etiquetado en Facebook! Hay sistemas que pueden identificar, a partir de la voz y el discurso, si una persona presenta señales de Parkinson o Alzheimer, incluso antes de que sea posible un diagnóstico médico. Eso es muy positivo para aplicar la prevención o el diagnóstico precoz, pero lo es menos si hay empresas que lo utilizan para averiguar si algún trabajador –o alguien que quiere contratar una póliza– presentará estos síntomas en un futuro más o menos próximo.



La genética añade una dimensión más a las posibilidades y a los riesgos. Esta ciencia ha avanzado mucho, pero eso, además de hacer conocer muchos genes ligados a enfermedades o a la predisposición a sufrirlas, también ha permitido constatar su gran complejidad. Ya no hace falta tener en cuenta solo los genes, sino también la epigenética, es decir, las alteraciones que pueden afectar la expresión de los genes a causa de la alimentación, la contaminación o traumas vitales y que incluso se pueden transmitir a la descendencia.

Por ello, algunos genes determinan, pero muchos otros, no. Que se pueda ceder información genética presenta un riesgo según en qué manos caiga. Pero, además, afecta a nuestro entorno familiar. Podemos evitar que haya datos nuestros en archivos, pero si un pariente directo los ha proporcionado, está ofreciendo también una parte de nuestra identidad genética. Y ahora no es tan

difícil que alguien lo haga. Empresas como 23andMe ofrecen pruebas genéticas para conocer la predisposición a ciertas enfermedades o para buscar, en su inmensa base de datos, posibles parientes más o menos directos.

CONCLUSIONES

Desde que en 1895, Wilhelm Roentgen descubrió los rayos X y abrió el camino a la radiología, nuestro cuerpo no ha dejado de hacerse cada vez más transparente. (Duran, 2016). Y el *big data* ha conseguido que no solo sea transparente el cuerpo, sino también nuestros actos y pensamientos. Eso genera posibilidades inmensas, pero también muchos riesgos (Duran, 2018).

No basta con la prudencia de los usuarios –que también es imprescindible– para evitar estos peligros. Por ello, tiene que haber una legislación que proteja sus derechos y unos tecnólogos y unas empresas que comprendan la necesidad de poner límites a las herramientas informáticas. El Reglamento Europeo de Protección de Datos, obligatorio en toda la Unión desde mayo de 2018, pretende garantizar que todo el mundo pueda elegir qué datos dona, a quién y con qué objetivo. Y que pueda incluso reclamar que se le explique qué criterios ha considerado un algoritmo que haya ayudado a tomar una decisión –sobre un crédito, una trabajo, un contencioso...

Estos derechos son un elemento esencial. Pero también lo es estudiar bien lo que se desprende de los modelos matemáticos para no obtener conclusiones sesgadas que condicionan el presente y el futuro. Pueden ser herramientas potentes para hacer una sociedad más justa, siempre que estemos dispuestos no solo a cantar sus excelencias sino también a admitir y contrarrestar sus limitaciones. Como decía en un artículo Joshua Blumenstock (2018), director del Data-Intensive Development Lab y profesor de la Universidad de California en Berkeley, el uso del *big data* para el desarrollo ofrece muchas posibilidades, pero requiere una versión de la ciencia de datos «considerablemente más humilde que la que ha captado la imaginación popular». La humildad es a menudo necesaria por alcanzar el éxito, porque ayuda a corregir o evitar errores. Por ello, habrá que incluirla y ejercitarla en el algoritmo que nos tiene que llevar a la sociedad de datos del futuro.

John Lubbock

El acceso a información sensible de usuarios por parte de terceras partes interesadas es un tema candente. A principios de 2018, estallaba el escándalo de Cambridge Analytica, cuando se descubrió que esta consultora británica había utilizado información de 87 millones de usuarios de Facebook para elaborar mensajes publicitarios personalizados –incluso *fake news*– favorables a la victoria de Donald Trump en la campaña electoral estadounidense de 2016. Inmediatamente, el papel de Cambridge Analytica en otras campañas electorales fue sometido a escrutinio público. Los lazos entre la consultora y el sector del *Vote Leave* ("vota salir") en el referéndum celebrado en 2016 en el Reino Unido sobre la permanencia del país en la Unión Europa hicieron que muchos ciudadanos saliesen a la calle para protestar por lo que consideraban un proceso manipulado.

«SOLO CON UNA PEQUEÑA PARTE DEL RASTRO QUE DEJAMOS EN INTERNET, YA SE PUEDE ELABORAR UN PERFIL CON BASTANTE APROXIMACIÓN»

REFERENCIAS

Allen, M. (2018, de 17 julio). Health insurers are vacuuming up details about you — And it could raise your rates. *ProPublica*. Consultado en https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates

Blumenstock, J. (2018). Don't forget people in the use of big data for development. *Nature*, *561*, 170–172. doi: 10.1038/d41586-018-06215-5

Duran, X. (2016). *L'individu transparent. Dels raigs X al 'big data'*. Lleida: Pagès editors.

Duran, X. (2018). *L'imperi de les dades. El 'big data': Oportunitats i amenaces*. Alzira: Bromera.

Kosinski, M., Stilwell, D., & Graepei, T. (2013). Private traits and attributes from digital records of human behaviour. *PNAS*, *110*(5), 5802–5805. doi: 10.1073/pnas.1218772110

Lesk, M. (1997). *How much information is there in the world?* Consultado en http://www.lesk.com/mlesk/ksg97/ksg.html

O'Neil, C. (2018). *Armas de destrucción matemática*. (V. Arranz, Trad.). Madrid: Capitán Swing. (Trabajo original publicado en 2016).

Pasquale, F. (2015). *The black box society. The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Wicklein, J. (1981). *Electronic nightmare. The new communications and freedom*. Nueva York: The Viking Press.

Youyou, W., Kosinski, M., & Stilwell, D. (2015). Computer-based personality judgements are more accurate tan those made by humans. *PNAS*, *112*(4), 1036–1040. doi: 10.1073/pnas.1418680112

Xavier Duran. Licenciado en Química y doctor en Ciencias de la Comunicación. Periodista científico de TV3, ensayista y divulgador (Barcelona). Ha publicado varios libros, entre los que se encuentra *La ciència en la literatura* (Universidad de Barcelona, 2015). Su obra *L'imperi de les dades. El 'big data', oportunitats i amenaces* (Bromera, 2018) ganó el XXIII Premio Europeo de Divulgación Científica Estudi General.