IA eñ™ • Publicaciones de autores

Los datos como eje principal en el "Estado del arte de la ciencia de datos en el idioma español y su aplicación en el campo de la Inteligencia Artificial"

**Celia Medina Lloret, Adrián Hernández González,
Jose Alejandro Orozco Leal, Diana Díaz Raboso, Marta Riquelme,
Ángel Manuel Pérez López**

Published on: May 14, 2021

DOI: 10.21428/39829d0b.2512bcb7

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Abstracto

Este trabajo está orientado a la ciencia de datos y la inteligencia artificial, especialmente al manejo de los datos en el idioma español. El interés por desarrollarlo tiene como base nuestra anterior publicación: *El estado de la ciencia de datos en el idioma español y su aplicación en el campo de la Inteligencia Artificial*, <https://doi.org/10.21428/39829d0b.f5257ea7>. El principal objetivo de este documento se divide en dos partes: por un lado, verificar si realmente existe un sesgo cultural entre el castellano y el inglés en la ciencia de datos. Por otro lado, analizar cómo se gestiona la privacidad de los datos en aplicaciones que utilizan, entre otras, técnicas como el contact tracing, además de dispositivos electrónicos como relojes “inteligentes”. Para llevar a cabo la primera parte del estudio se ha indagado sobre la disponibilidad de conjuntos de datos en inglés y español, la estructura técnica que poseen, y la creación de una base de datos que confirmará la discriminación cultural que existe entre ambas lenguas. Para la segunda, se ha profundizado en temas de *smartwatches* en menores, añadiendo una tabla comparativa de privacidad de las grandes marcas. Tras la evaluación realizada, se observa que existe un sesgo cultural entre el castellano y el inglés. De hecho, los resultados que ofrecen el 70% de *datasets* analizados están en inglés. Existen motivos que justifican la carencia de una educación tecnológica digna para países de habla hispana; por ejemplo, la inversión en innovación tecnológica. Precisamente, un artículo publicado por el diario El Economista con fecha 14 de mayo de 2020, hace una lista de los países más implicados con una innovación tecnológica. Entre estos países no destaca España, pues la inversión que realiza es el 1,24% sobre el total del PIB (Producto Interior Bruto). Esto quiere decir que no llega ni a la media europea, impulsada por cumplir el objetivo de alcanzar, al menos el 3% en 2020. En el lado opuesto, se encuentran países como Corea del Sur, Dinamarca y Suecia, alcanzando la barrera del 3 e incluso rebasando el 4% ([Economista, 2020](#)). Por consecuencia, esto tiene repercusiones negativas para la educación tecnológica. Una prueba de ello son los resultados del índice PISA (Programa para la Evaluación Internacional del Estudiantes), donde los españoles sacan menos nota que la media de países de la OCDE (Organización para la Cooperación y el Desarrollo Económicos) en dos ramos de conocimiento tan importantes para la tecnología como son: ciencia y matemáticas. Además de España; Chile, México y Colombia se encuentran a la cola de la lista ([Epdata, 2020](#)). Este documento, además, muestra unos datos que nos advierte de que el avance tecnológico afecta cada vez más a la sociedad, y consecuentemente a su privacidad. La

centralización de las decisiones en organizaciones internacionales colaborativas podría ser una solución, aplicando: estrategias eficientes, éticas y deontológicas.

Palabras clave: ciencia de datos, inteligencia artificial, idioma español y ciberseguridad.

Abstract

The main objective of this paper is divided into two. First, we need verify if there exist a cultural bias between Spanish and English language in data science. Second, looking how data privacy is controlled in applications which use some techniques such as contact tracing, smartwatches, and other electronics devices. The indicators that were used for this assessment cover security normative of European and Spanish regulation (GRDP), as well as other international rules. To carry out the first part, we have investigated about the accessibility of datasets in Spanish and English language; the technique structure that present, and we have also created a database which offers a valuation of cultural discrimination between these two languages. By the second issue, cybersecurity, we have looked for some topics, for example, smartwatches for minors and there is a comparative table of security on major brands. Following the evaluation carried out, it is observed that there is a cultural bias between Spanish and English language. The analysis of results throws that around 70% of datasets "speaks" English. There is reason to justify the lack of technological education to Spanish speakers countries, such as, technological innovation inversion. Precisely, an article published by the newspaper The Economist, it made a list of the countries most involved with technology on date May 14, 2020. Among these countries, Spain does not stand out because it makes an investment of 1,25% of the total GDP (Gross Domestic Product). Europe aims with reaching at least 3% in 2020, so it means that Spain does not reach the European average. On contrary, there are other countries such as: South Korea, Denmark and Sweden that even exceed the barrier of 3%. Consequently, these data have negative impact for technology education. As well as that, the results of PISA (Program for International Student Assessment) inform reveal that Spanish students' knowledge of science and mathematics are below average for OECD (Organization for Economic Cooperation and Development) countries. Furthermore, Chile, México and Colombia are at the bottom of the list. This document also shows some data that technological advances increasingly affect society and, consequently, privacy. The focus of decisions must be regulated by a collaborative international organization to develop a solution in basis of: efficiency, ethic, and deontological strategies.

Key words: data science, artificial intelligence, spanish language and cibersecurity.

1. Introducción

Las capacidades de la inteligencia artificial hoy son una realidad gracias a algoritmos y datos que entrenan a esos algoritmos. Y es que si no existieran datos observables, bien de disponibilidad pública, o bien, privada, la inteligencia artificial no existiría como tal, pues hablar de inteligencia artificial, definitivamente es hablar de datos ([BBVA, 2018](#)).

Siempre que trabajamos con datos en inteligencia artificial es importante evaluar la calidad de estos datos. Jordi Calvera Sagué, *mánager* regional de *InterSystems* para España, Portugal, Israel, Grecia, Turquía y Latinoamérica, defiende la idea de que para desarrollar un buen proyecto basado en inteligencia artificial o aprendizaje automático es necesario que los datos sean de calidad. A estas palabras se suman las de los profesionales de esta ciencia; los científicos de datos, quiénes confirman que la manipulación de datos sin depurar se ha convertido en un auténtico desafío para su trabajo ([Calvera, 2020](#)). Pero, ¿a qué se refieren con la depuración de los datos? Comúnmente conocida como *Scrubbing*, ésta es una técnica de limpieza de datos cuya finalidad es modificar o eliminar datos incorrectos, incompletos, y duplicados de una base de datos ([TechTarget, 2019](#)).

Además de comprobar el estado de los datos, es muy importante mantener la integridad de los datos que se utilizan, sobre todo si son de carácter personal. A partir de la década de los 90, la preocupación por la desprotección de la información en los sistemas informáticos en red se vio justificada por el incremento de secuencias de comandos utilizados para violar a gran escala información personal de los clientes de una corporativa, ataques a infraestructuras civiles y ciber espionaje dirigido a documentos secretos ([Mulligan & Schneider, 2011](#)). Conociendo estos datos, lo que nos queda claro es que la privacidad y la protección de datos es fundamental en esta era tecnológica, donde los ciberataques son cada vez más frecuentes, sobre todo en grandes entidades y organismos políticos y económicos ([Banafa, 2018](#)). Esto no significa que cualquier individuo que tenga acceso a un dispositivo electrónico no esté en riesgo, pues la privacidad de sus actos puede quedar visiblemente expuesta, perdiendo así todo tipo de intimidad.

2. Disponibilidad de conjuntos de datos en inglés y castellano

En este apartado se discute sobre los *datasets* y las lenguas castellana e inglesa. La metodología para la realización del estudio consiste en: seleccionar algunos de los

motores de búsqueda que aparecen en Internet sobre los conjuntos de datos ([Stacy Stanford, 2018](#)): Dataset Search de Google ([Google, 2018](#)), Kaggle ([Goldbloom, 2010](#)) y el Banco Mundial de Datos ([BancoMundialdeDatos, 1944](#)). Después, se filtran palabras de diversos campos de estudio, dando prioridad a la inteligencia artificial, señalando la disponibilidad de idiomas en los datos y acentuando el detalle del castellano, así como de las culturas predominantes. Para comenzar, analizamos los conjuntos de datos obtenidos sobre lingüística, redes sociales, turismo, tecnología y globalización.

Image 1
Conjuntos de datos *Datasets Search Google*

En primer lugar, el motor de búsqueda empleado como medidor ha sido Dataset Search de Google. Este gráfico nos muestra un predominio del inglés en todas las áreas, aunque no representa una diferencia desproporcionada de los valores, a excepción del término turismo en inglés, en el que se observa una diferencia de casi 200 *datasets* entre ambas lenguas. Si centramos nuestra atención en la palabra globalización en castellano, encontramos varios conjuntos de datos que hablan sobre la globalización y las relaciones internacionales, las infraestructuras del proceso de globalización, así como de las fronteras y los mercados ilegales en la era de la globalización. Cuando comparamos los datos en porcentaje sobre los idiomas, el 62.93% de los *datasets* son en inglés y el restante 37.07% en español. Aunque mayoritariamente los conjuntos de

datos hacen referencia a España, también es cierto que aproximadamente un 20% de la información es de índole internacional con especial foco en países latinoamericanos como México y Guatemala.

Image 2
Conjuntos de datos *Datasets Kaggle*

Este esquema es bastante similar al anterior, pues las variables representadas son las mismas. En este caso, el motor de búsqueda utilizado ha sido Kaggle. Sin embargo, los resultados que arrojan son dispares. Una vez más, el idioma por excelencia es el inglés. El castellano ni siquiera tiene representación gráfica, obteniendo unos valores entre 0 y 4 en cada una de las selecciones. En valores de tanto por ciento, los *datasets* encontrados en español no superarían el 2.2% sobre el total, mientras que el inglés ocupa un 97.8% . En este punto llama la atención que no se recoja ningún conjunto de datos sobre globalización ni redes sociales en castellano, especialmente porque son temas actuales y de elevado interés social. Al contrario de lo que sí ocurre con la lengua inglesa, pues encontramos conjuntos de datos sobre la actividad de móviles en una ciudad o la falsedad en la detección de rostros. Además, la fecha de actualización de los *datasets* encontrados en inglés es más reciente que los de español, lo que nos indica una mayor implicación del inglés con la tecnología.

Los tres siguientes esquemas son una muestra de los artículos que hablan de ciencia y tecnología en el Banco Mundial de Datos, Dataset Search de Google y Kaggle. Comparamos y analizamos las diferencias y similitudes que existen.

Image 3
Banco Mundial de Datos Tecnología

Esta gráfica de barras muestra los conteos de las secciones de tecnología en el Banco Mundial de Datos para verificar las proporciones del castellano y el inglés. En la escala de 0 a 100, la lengua inglesa roza los 90, marcando una diferencia de casi 80 puntos con el español. Teniendo como principal base la tecnología, estos datos representados en la gráfica hacen referencia a los artículos en publicaciones científicas y técnicas en el año 2018.

Image 4
Datasets Search Google Tecnología

Siguiendo la línea de la imagen anterior, valoramos los resultados ofrecidos por Dataset Search de Google. En este caso, no se aprecia una desproporción de los números, sino que las barras representadas en ambos idiomas están igualadas, existiendo una mínima disimilitud. En la selección del español, nos parece interesante destacar algunos conjuntos de datos como la brecha de género de personas tituladas en carreras profesionales del área de tecnología por región o el número de equipos médicos de alta tecnología en Asturias por tipo. También hay que destacar que en la búsqueda del castellano no solamente aparecen *datasets* de España, sino que encontramos un alto porcentaje de datos gestionados por países latinoamericanos como Chile y Panamá, entre otros.

Image 5
Kaggle Tecnología

Esta última gráfica es muy similar a la del Banco Mundial de Datos, puesto que presenta una diferencia abismal de los datos. Podríamos decir que el castellano representa el 0% de la información, mientras que el inglés el 100%. Esta es otra certeza de que el mundo digital se expande a pasos agigantados y el inglés es la llave maestra en esta andadura.

Los dos últimos esquemas son específicos del área que estamos trabajando: la inteligencia artificial.

Image 6
Dataset Search Google: Inteligencia Artificial

El primero hace referencia a Dataset Search de Google. Los cuatro campos de estudio seleccionados presentan una enorme distancia entre los resultados en castellano e inglés. A pesar de los datos tan mínimos que recoge la lengua española, la inteligencia artificial es el más destacado con más de 24 conjuntos de datos. Probablemente, esto se debe a que es un ámbito más general de la tecnología y en constante evolución. En esta búsqueda, encontramos conjuntos de datos muy interesantes que tratan sobre la clasificación de patrones en imágenes sísmicas aplicando inteligencia artificial.

También abarca datos sobre las oportunidades y los desafíos de la inteligencia artificial y la ciberseguridad, la facturación mundial del mercado de procesamiento del lenguaje natural 2017-2025 y los coches de segunda mano a la venta en España gestionados por la ciencia de datos. Como se puede observar, existen *datasets* de todas las nacionalidades latinoamericanas, donde la multiculturalidad y la variedad lingüística es notable. A pesar de ello, el inglés es con una marcada diferencia el principal motor de la inteligencia artificial y todo lo relacionado con ella. De hecho, el 96.9% de los *datasets* son en inglés, dejando un 3.1% para el español.

Image 7
Kaggle: Inteligencia Artificial (Escala Logarítmica)

Los datos que muestra Kaggle en ambos idiomas son por lo general muy escasos. En el caso de la lengua castellana, podríamos destacar la presencia de conjuntos de datos sobre inteligencia artificial y ciencia de datos (aproximadamente el 1% de los encontrados), mientras que en la lengua inglesa todas las selecciones tienen una representación, llegando casi a alcanzar el 100% de los resultados. Cabe destacar el número de *datasets* sobre ciencia de datos en inglés, que sin duda es el único dato relevante del gráfico. En los dos conjuntos de datos en español, uno de ellos procede de la variante latinoamericana del castellano, concretamente de Paraguay. El mismo caso se repite con el único *dataset* encontrado en español sobre ciencia de datos: la donación de órganos en México, donde el castellano de España se ausenta una vez más.

Como valoración final, llegamos a la conclusión de que el inglés es clave para el desarrollo de los conjuntos de datos generales, referentes a la tecnología y, sobre todo, a la inteligencia artificial. Mientras, al castellano le queda un largo camino por recorrer, pues las variantes latinoamericanas están solventando, en cierta parte, el vacío del español en un área tan importante en nuestro día a día como es la tecnología.

3. Estructura técnica de los conjuntos de datos

En base al análisis de la tabla del final de sección acerca de los conjuntos de datos de la comunidad Kaggle, nos percatamos aún más del notable sesgo cultural que existe entre el castellano y el inglés. Para este estudio, la metodología llevada a cabo se centra en la búsqueda de conjuntos de datos en inglés y español por medio de la filtración de los siguiente términos:

- Inteligencia artificial - *Artificial Intelligence*
- Tecnología - *Technology*
- Aprendizaje automático - *Learning Machine*
- Procesamiento del lenguaje natural - *Natural Language Programming*
- Ciencia de datos - *Data Science*
- Científico de datos - *Data Scientist*
- Globalización - *Globalisation*
- Lingüística - *Linguistics*
- Redes sociales - *Social Networks*
- Turismo - *Tourism*

De los **23.771** conjuntos de datos que se encontraron con fecha de consulta 12/04/2021, tan solo 10 se encontraban en castellano, representando un **0.042% del total**. Mientras que el resto, **23.761** “hablaban” inglés.

En cuanto al formato de exportación, más de la mitad de los conjuntos de datos se exportan en formato **CSV**, concretamente **14.931** conjuntos de datos o lo que es igual; **60.85% sobre el total**.

Image 8
Comparativa formatos de exportación

Pero realmente, ¿a qué llamamos formato CSV ? Se trata de un formato de texto en el que los valores están separados por comas. Estos archivos son muy útiles cuando se elaboran *datasets* porque podemos convertirlos en tablas de manera automática y además ocupan muy poco espacio al ser texto plano. Lo habitual es que dicho archivo tenga los campos separados por comas en cada columna y utilice un salto de línea para declarar una nueva fila, pero puede ocurrir que encontremos archivos separados por “;” o por “-”. En cualquier caso, se convertiría de manera automática en una tabla al abrir el archivo en cualquier *software* que lo soporte. De la misma forma, destacamos los formatos de imagen más utilizados, **JPG y PNG**, con **3.070 datasets (12.51% sobre el total)**.

En cuanto a los tipos de bases de datos, **un 86%** de las mismas son **bases de datos estáticas**, es decir, bases de datos que no se actualizan desde hace 90 días o más. Al contrario de lo que ocurre con las **bases de datos dinámicas**, en las que el **14%** restantes son constantemente actualizadas.

En la tabla que se encuentra adjunta como repositorio están presentes los valores al detalle de cada uno de los conjuntos de datos, desglosados por idioma y formato de exportación.

Visit the web version of this article to view interactive content.

4. Software y herramientas empleadas en la ciencia de datos

En este punto del estudio exploramos qué lenguajes de programación se utilizan más en la ciencia de datos. Durante la revisión, se emplean recursos de búsqueda como Scielo, Academia.edu, Google Academic, Kaggle y WorldWideScience de los que obtenemos resolución mayoritariamente en inglés. Esto reafirmaría el gran poder de la lengua inglesa en la ciencia de datos.

Tras el análisis de los resultados determinamos que el lenguaje de programación R es por excelencia uno de los más aplicados en ciencia de datos. Según la Revista Internacional de Investigación de Ingeniería y Tecnología (IRJET) ([Tecnología, 2020](#)), R es un lenguaje de código abierto, lo que implica disponer de una mayor flexibilidad, seguridad y calidad a la hora de ofrecer un análisis estadístico de los datos. De la misma forma, R proporciona varias estadísticas y capacidades gráficas útiles para la clasificación y pruebas estadísticas. A esto hay que añadir algunas de las organizaciones y entidades públicas y privadas que trabajan con R Data Science: International Journal of Information and Education Technology, Kearney, Elaine de College of Arts and Sciences y Department of Computer Sciences. ([Yamamoto et al., 2021](#)).

Otro de los lenguajes de programación más destacados en ciencia de datos es Python. Éste es un lenguaje versátil que destaca por su código legible y limpio ([Robledano, 2019](#)). Al igual que ocurre con R, Python cuenta con una licencia de código abierto que permite su aplicación en cualquier escenario. Además, posee diferentes librerías que se especializan en la manipulación de datos, manejo de vectores, matrices y operaciones matemáticas a gran velocidad, así como la generación de visualizaciones interactivas; todo ello, atendiendo a un conjunto de reglas. Entre las más destacadas están: Pandas, Numpy y Plotly ([Bagnato, 2020](#)).

De acuerdo con el doctor en economía de la Universidad de Sao Paulo, A. Días Porto Chiavegatto Filho:

“Los lenguajes de programación R y Python van de la mano.” ([Chiavegatto Filho, 2015](#)).

Esta expresión significa que ambos tienen puntos en común respecto al manejo de información; recopilación, exploración, modelado y visualización de datos ([Parada, 2021](#)). La principal diferencia es que R está orientado al análisis estadístico en ciencia de datos y Python es multifuncional en otros campos de desarrollo *web* y además, permite integrar datos que provienen de diferentes plataformas ([Tecnología, 2020](#)).

A parte de estos, existe otro *software* muy potente en ciencia de datos: Anaconda. Ésta es una de las mayores plataformas de código abierto que facilita el uso de otros lenguajes de programación abiertos, como R y Python, para el procesamiento de datos a gran escala, los análisis predictivos y la computación científica. Por esta serie de características se convierte en un gran implemento para el desarrollo de la ciencia de datos ([Anaconda, 2021](#)).

A colación de lo que veníamos explicando sobre el *software* o lenguajes de programación empleados en ciencia de datos, es muy interesante recopilar también información sobre las herramientas auxiliares de las que esta ciencia se vale.

Escribir un informe en ciencia de datos prescinde de mecanismos de trabajo con denominadores en común como son: escribir código y procesar datos. R Markdown, LaTeX o Jupyter notebook ([Randles et al., 2017](#)) son algunos de los *software* más destacados que se pueden utilizar para la ciencia de datos ([Ding, 2016](#)).

Markdown fue creada por John Gruber en 2004 y distribuida bajo la licencia de BSD. Una de las ventajas que ofrece es la simplicidad de su sintaxis y el fácil acceso a los dispositivos móviles ([Cristóbal, 2016](#)). Pero, ¿por qué Markdown es una herramienta útil en la ciencia de datos? En primer lugar, ofrece colaboración con otros científicos de datos además de un entorno para hacer ciencia de datos, pues no se emplea únicamente como una libreta para escribir lo que estás haciendo, sino también lo que estás pensando. En el aspecto de la colaboración se pueden utilizar herramientas de control de versiones como Git y Github ([Wickham, 2016](#)).

De acuerdo con el proyecto Jupyter, ésta es una aplicación de código abierto destacada por la transformación y visualización de datos y por el aprendizaje automático. Asimismo, tiene otras extensiones como JupyterLab en la que una de sus vinculaciones con la ciencia de datos es configurar y planificar la interfaz de usuario para apoyar una amplia gama de trabajos en la ciencia de datos. ([Jupyter, 2021](#)) Agregar que Jupyter Notebook debe su nombre a los tres principales lenguajes que soporta (Julia, Python y R). Este cuaderno utiliza una lista ordenada de celdas de entrada y salida que pueden contener código, fórmulas matemáticas y texto, entre otras.

Ya, para terminar, hablamos de Sweave, un componente del lenguaje de programación R que permite trabajar en conjunto con LaTeX. LaTeX es uno de los programas de tipografía capaz de producir complejas ecuaciones matemáticas ([Zhang, 2017](#)). De hecho, durante muchos años ha sido utilizado para gestar distintas revistas de ciencia, ingeniería y matemáticas ([Britannica, 2021](#)). Podríamos decir que Latex es utilizado para escribir fórmulas científicas en publicaciones, es más en el entorno puro de la ciencia LaTeX es forzoso conocerlo o al menos saber de él.

En el próximo apartado centraremos la atención en las medidas de protección de la privacidad digital que se aplican para gestionar los datos.

5. Ciberseguridad en datos

En la actualidad existe una amplia disponibilidad de dispositivos móviles o 'inteligentes' que todos tenemos a nuestro alcance desde cámaras fotográficas digitales pasando por tabletas o agendas digitales hasta relojes 'inteligentes'. Recientemente, leíamos una publicación en el periódico ABC del 11 de mayo de 2017 sobre la falta de información que tienen los menores acerca de los riesgos que supone Internet. De hecho, varios estudios realizados en Proyecto Hombre demuestran que el 90% de los jóvenes con edades comprendidas entre los 10 y los 14 años tienen un dispositivo portátil y la mayoría lo utilizan para chatear y difundir imágenes y vídeos en las redes sociales sin saber cómo se gestiona ([Setién, 2017](#)). El problema es aún mayor cuando el desconocimiento sobre el funcionamiento de los dispositivos electrónicos se extiende tanto a adolescentes como a adultos.

Cuando nos referimos a que un aparato es 'inteligente' no debemos olvidar que para que algo aporte beneficios y soluciones en determinadas situaciones, ese algo debe conocer previamente lo que está ocurriendo. Todo esto se traduce en una única palabra; datos. Pongamos como ejemplo los *smartwatches* o relojes 'inteligentes' y veamos las ventajas que ofrecen. Entre otras funcionalidades, un *smartwatch* permite medir la presión arterial y la temperatura, interactuar con actividades de reconocimiento, controlar todos los dispositivos de nuestra casa que estén conectados a Internet, notificar la llegada de un correo y leerlo desde tu muñeca sin acceder al *smartphone*, enviar mensajes de voz, gestionar tus redes sociales y contestar llamadas ([Yañez, 2015](#)).

Para saber más acerca del manejo de los relojes 'inteligentes' y su correlación con el seguimiento de la actividad humana, tomamos como referencia el artículo de investigación: An Analysis of Human Activities Recognition using Smartwatches

Dataset publicado por la International Journal of Advanced Computer Science and Applications en diciembre de 2020 ([Karim et al., 2020](#)). Los *smartwatches* tienen sensores que identifican modelos o patrones del comportamiento humano junto con técnicas de aprendizaje automático, el teorema de Bayes, el procesamiento de datos o el también conocido como método k de los vecinos más cercanos. Estos procedimientos generan un gran volumen de información con el que se pretende cumplir el objetivo de adquirir precisión en los resultados emitidos. Estos sensores son muy útiles para, por ejemplo, monitorizar la salud humana y ofrecer los servicios necesarios al paciente, pues son capaces de medir distintas actividades físicas como: caminar, ciclismo, *running* y subir o bajar escaleras. Los detectores a los que nos estamos refiriendo no son otros que: *Global positioning system* (GPS), *Wireless Fidelity* (Wi-Fi) y *Near Field Communication* (NFC). Estas tecnologías que a veces conectamos o desconectamos sin saber realmente lo que estamos haciendo, proporcionan referencias muy valiosas para el dispositivo sobre nosotros o en otras palabras, alimentamos sin querer la 'inteligencia' del aparato.

El servicio GPS no solamente está incluido en *smartwatches*, sino que su uso es muy frecuente entre los dispositivos de rastreo a menores. Sin embargo, existe un amplio desconocimiento en este tema. Los hábitos de caminos que siguen nuestros hijos no son conocidos únicamente por sus padres, amigos o familiares cercanos, sino que las empresas de tecnología también pueden recibir este tipo de información y, es más, una mala gestión o brecha de seguridad podría llevar los datos a manos de la persona equivocada y acabar en un secuestro, por lo que estaríamos hablando de un delito de gravedad ([Judd, 2020](#)). A colación de estos datos, nos parece interesante hablar sobre el caso Anna, un hecho real que narró el ingeniero Maik Morgenstern en el blog Internet of Things en noviembre de 2019. Anna es una niña de corta edad que vive con sus padres en el distrito de Lücklemburg pero durante las vacaciones la pequeña se va con sus abuelos a Norderney porque sus padres tienen que trabajar. Durante el periodo vacacional, Anna vive en una acogedora casa en el distrito de Fischerhafen. La niña sale a dar cortos paseos sola por el vecindario sin la compañía de sus abuelos. A simple lectura, parece una historia contada por cualquier persona, ¿verdad? Pues no. Realmente toda esta información la sabemos gracias a un reloj "inteligente" ([Morgenstern, 2019](#)).

5.1 Smartwatches en menores

Hemos elegido el caso de Anna, pero como Anna podríamos mencionar cientos de casos más. Los fallos de seguridad más importantes de estos relojes de bajo coste se

producen en las aplicaciones y en la conexión con los servidores que almacenan los datos más que en el propio dispositivo, lo cuál es más grave aún, ya que no se necesita acceder físicamente al aparato para poder aprovechar las vulnerabilidades. Los puntos débiles más comunes suelen estar relacionados con el encriptado de credenciales o el cifrado de las comunicaciones entre la aplicación y el servidor que aloja los datos. Otro aspecto a tener en cuenta es el bajo coste del dispositivo. Esto se traduce en descuidar aspectos muy importantes en la protección de datos, máxime si hablamos de dispositivos que van a ser utilizados por menores de edad. Si a esto le añadimos que ninguno de los casos de estudio cuenta con el RGPD (Reglamento General de Protección de Datos) accesible, estaríamos ante un grave delito contra la seguridad privada. Sin embargo, en dispositivos de marcas más reconocidas como Nokia ([Clausing, 2018](#)), Samsung o Huawei ([Clausing, 2018](#)), estos problemas no ocurren con frecuencia, ya que se realizan conexiones cifradas. De esta forma, los datos están completamente encriptados en la mayoría de los casos y los ataques por fuerza bruta quedarían limitados ([Stykas, 2019](#)).

Entre las marcas más utilizadas por menores destacamos: Carl Kids Watch, hellOO! Children's Smart watch, SMA-WATCH-M2 y GATOR Watch. El problema de Carl Kids Watch reside más en su aplicación que en el propio reloj. La aplicación falla en la implementación de certificados para conexiones seguras HTTPS, por lo que prácticamente cualquier certificado es aceptado. Igualmente, se utilizan conexiones no cifradas y se almacena un fichero en la tarjeta SD no cifrado en el que aparecen, entre otras cosas, la contraseña de la cuenta en texto plano. A esto se le suma que la comunicación entre usuario y servidor se realiza mediante conexión HTTP sin cifrar, incluyendo el registro o inicio de sesión a una cuenta ([Clausing, 2018](#)).

La información relativa al registro electrónico en hellOO! Children's Smart watch está sin cifrar en la carpeta del *software*, por lo que si se logra tener acceso a ella, los atacantes tendrían el control total sobre los datos del usuario; llamadas, mensajes y posición GPS ([Henke, 2018](#)).

Otra de las marcas diseñadas para "proteger" a los menores mediante el rastreo GPS es SMA-WATCH-M2 y es, precisamente, el dispositivo que Anna, la niña de la que hablábamos anteriormente, portaba en su muñeca. Este aparato obtiene una puntuación negativa en las pruebas de ciberseguridad porque los delincuentes podrían obtener la ubicación y escuchar o manipular conversaciones confidenciales. El fallo de seguridad de este reloj "inteligente" se produce en el portal *web* que el fabricante ofrece, ya que no está cifrado al completo y el servidor no verifica el *token* de

autenticación. Se puede acceder sin problema al identificador de usuario con toda la información relativa al uso del dispositivo y las coordenadas actuales. Igualmente, todo puede reproducirse en la propia aplicación del fabricante y tener control total de las funcionalidades del reloj como si fuéramos el usuario legítimo ([Morgenstern, 2019](#)).

Con los dispositivos de la marca Gator Watch se logró acceder, utilizando ataques IDOR a datos sensibles de los usuarios y a funciones específicas como: coordenadas reales del GPS, llamadas bidireccionales y unidireccionales así como notas de voz. También se recuperó información personal como: el nombre, la edad, el peso y la altura. De esto se sigue que la página *web* que monitoriza los dispositivos tiene una vulnerabilidad relacionada con el uso de servidores proxy, es decir, permite revisar la solicitud que se envía al servidor *web* y obtener el acceso total a la plataforma. ¿Cómo lo consigue? Cambiando un par de valores en la petición al servidor como si fuéramos administradores. De esta forma, una vez que tengamos todos los privilegios de administrador de la plataforma, tendremos acceso a los datos de todos los usuarios ([OZA, 2020](#)).

Y es que, finalmente, llegamos a la conclusión de que estos aparatos “inteligentes” son capaces, como ya vemos, de escribir tu vida al pie de la letra. Por eso, no debemos olvidar que los únicos responsables de proporcionar nuestros datos personales a los demás, somos nosotros mismos al tener un dispositivo “inteligente” adherido a nuestro cuerpo. Otro aspecto que debemos de tener en cuenta es que esto ocurre en mayor o menor medida dependiendo del sistema de seguridad que cada marca desarrolle. Por este motivo, al final de este punto incluiremos una tabla comparativa de las diferentes marcas de relojes y la confianza que deberíamos depositar en cada una de ellas.

5.2 Comparativa de seguridad

Se han analizado cuatro de las grandes marcas que se pueden encontrar en el mercado para el acceso a dispositivos *wearables*: Samsung ([Foundation_Mozilla, 2020](#)), Apple ([Foundation_Mozilla, 2020](#)), Fitbit ([Foundation_Mozilla, 2020](#)) y Garmin ([Foundation_Mozilla, 2020](#)). Las marcas se han elegido en función de ventas para Samsung, Apple, Garmin, ya que la idea es tener controladas las que más venden y añadir una con muchos modelos en el mercado de cara a la base de datos como es Fitbit.

Las conclusiones que sacamos de este estudio y que se recogen, además, en la tabla del final de sección, es que todas las marcas realizan una recopilación de datos no anónima; almacenando datos del GPS y la salud del usuario.

La parte positiva de esta forma de recopilar datos la encontramos en Samsung, pues no recopila información *online* de menores de 13 años. No ocurre así con las otras tres marcas, Apple, Fitbit y Garmin, que no hacen distinción de edad.

Otra de las cuestiones muy importantes a considerar en la seguridad de estos dispositivos es analizar si se comparten los datos con terceros. Desafortunadamente la respuesta es: Sí. Actualmente estas cuatro marcas comparten datos con otras empresas, aunque no todas de la misma forma. En todas ellas los datos se comparten con terceros para analizar métricas y comparar resultados, pero solo en Samsung los datos del usuario son vendidos o alquilados. Al contrario de lo que ocurre con Apple; pues es la única de las empresas analizadas que no comparte los datos con fines de publicidad y/o márketing.

SAMSUNG		APPLE	
Recopilación de Datos		Recopilación de Datos	
¿Se realiza de forma anónima?	No	¿Se realiza de forma anónima?	No
¿Se almacenan datos del GPS?	Si	¿Se almacenan datos del GPS?	Si
¿Datos de salud?	Si	¿Datos de salud?	Si
Información en línea de menores de 13 años	No	Información en línea de menores de 13 años	Si
Datos Compartidos		Datos Compartidos	
La información recopilada se comparte con terceros	Si	La información recopilada se comparte con terceros	Si
Los datos se comparten para análisis	Si	Los datos se comparten para análisis	No disponible
Los datos se comparten para publicidad y / o marketing de terceros	Si	Los datos se comparten para publicidad y / o marketing de terceros	No
Los datos se venden o alquilan a terceros.	Si	Los datos se venden o alquilan a terceros.	No
¿Los terceros están autorizados a conocer la identidad del usuario?	Si	¿Los terceros están autorizados a conocer la identidad del usuario?	Si
Seguridad		Seguridad	
Protección de cuenta mediante 2FA	No.	Protección de cuenta mediante 2FA	Si
¿Todos los datos son encriptados en su envío? (Smartwatch - Smartphone)	No, solo algunos	¿Todos los datos son encriptados en su envío? (Smartwatch - Smartphone)	Si
¿Todos los datos son encriptados en su envío? (Smartphone - Servidor)	Si	¿Todos los datos son encriptados en su envío? (Smartphone - Servidor)	Si
¿Se almacenan todos los datos encriptados?	No, solo algunos	¿Se almacenan todos los datos encriptados?	Si
¿Se avisa al usuario si se produce una brecha de seguridad?	No.	¿Se avisa al usuario si se produce una brecha de seguridad?	Si
¿Se realizan auditorías de seguridad en el producto?	No disponible	¿Se realizan auditorías de seguridad en el producto?	No disponible

Image 9
Comparativa de seguridad Samsung y Apple

FITBIT		GARMIN	
Recopilación de Datos		Recopilación de Datos	
¿Se realiza de forma anónima?	No	¿Se realiza de forma anónima?	No
¿Se almacenan datos del GPS?	Si	¿Se almacenan datos del GPS?	Si
¿Datos de salud?	Si	¿Datos de salud?	Si
Información en línea de menores de 13 años	Si	Información en línea de menores de 13 años	Si
Datos Compartidos		Datos Compartidos	
La información recopilada se comparte con terceros	Si	La información recopilada se comparte con terceros	Si
Los datos se comparten para análisis	Si	Los datos se comparten para análisis	Si
Los datos se comparten para publicidad y / o marketing de terceros	Si	Los datos se comparten para publicidad y / o marketing de terceros	No disponible
Los datos se venden o alquilan a terceros.	No	Los datos se venden o alquilan a terceros.	No
¿Los terceros están autorizados a conocer la identidad del usuario?	No disponible	¿Los terceros están autorizados a conocer la identidad del usuario?	No disponible
Seguridad		Seguridad	
Protección de cuenta mediante 2FA	No disponible	Protección de cuenta mediante 2FA	No disponible
¿Todos los datos son encriptados en su envío? (Smartwatch - Smartphone)	Si	¿Todos los datos son encriptados en su envío? (Smartwatch - Smartphone)	No, solo algunos
¿Todos los datos son encriptados en su envío? (Smartphone - Servidor)	No disponible	¿Todos los datos son encriptados en su envío? (Smartphone - Servidor)	No disponible
¿Se almacenan todos los datos encriptados?	No disponible	¿Se almacenan todos los datos encriptados?	No disponible
¿Se avisa al usuario si se produce una brecha de seguridad?	No disponible	¿Se avisa al usuario si se produce una brecha de seguridad?	No disponible
¿Se realizan auditorías de seguridad en el producto?	No disponible	¿Se realizan auditorías de seguridad en el producto?	No disponible

Image 10
Comparativa de seguridad Fitbit y Garmin

5.3 ¿Cómo se podría atacar a las distintas marcas de *smartwatches*?

¿Y si hablamos de seguridad? Tanto Fitbit como Garmin son bastante opacas en este aspecto y no nos ofrecen ningún tipo de información, salvo en el caso de la encriptación de los datos en su envío, que es total en el caso de Fitbit ([Program, 2020](#)) y Apple ([C. S. P. Program, 2020](#)).

Tanto Garmin ([Program, 2020](#)) como Samsung ([Program, 2020](#)) encriptan sus datos parcialmente en el envío entre *smartwatch* y *smartphone*. Apple, además, es la única de las cuatro empresas analizadas que almacena los datos de manera encriptada y alerta si se produce alguna brecha de seguridad. Eso sí, en todas las marcas los datos de los usuarios están sujetos al RGPD. Por añadidura, en ninguna de las empresas sabemos si se realizan auditorías de seguridad externas en los productos.

En cuanto a los posibles ciberataques que podrían realizarse, se han analizado cuatro tipologías en este primer acercamiento: Man in the middle, Phishing, Inyección SQL y

Ataque Drown.

- **Man in the middle:** Esta forma de ciberataque supone que el atacante se encuentra entre el cliente / usuario, en este caso el *smartwatch*, y el servidor, teléfono móvil que recibe los datos. El atacante es capaz de leer, modificar o insertar datos a su voluntad. Tanto Samsung como Garmin están expuestos a estos ataques ya que envían los datos desde el *smartwatch* al teléfono móvil sin cifrar. Tanto Apple como Fitbit los envían totalmente cifrados.
- **Phishing:** Esta técnica consiste en engañar al usuario para que normalmente ingrese sus datos de autenticación en una *web / app* muy similar a la original pero cuya propiedad la tiene el atacante. El atacante captura esos datos y los puede utilizar para acceder a la *web* o aplicación real. Samsung es vulnerable ya que no se puede realizar una autenticación *2FA* o de dos factores. Para Fitbit y Garmin no se posee información al respecto y Apple estaría más protegido ya que si dispone de *2FA*.
- **Inyección SQL:** Este ataque consiste en aprovechar una vulnerabilidad del sistema para ingresar sentencias SQL maliciosas en la base de datos del sistema atacado. Hay diferentes tipos de ataque por SQL pero se basan en conseguir el acceso a las tablas de la base de datos objetivo. Samsung y Apple han tenido problemas en el pasado con la inyección SQL a los servidores que almacenan sus datos, por lo que es posible que se encuentren nuevos problemas de seguridad en el futuro. Para Garmin y Fitbit no se dispone de información.
- **Ataque Drown:** Esta es una vulnerabilidad que afecta al protocolo SSLv2 y permite descifrar comunicaciones seguras realizadas, utilizando el protocolo TLS entre un cliente y un servidor. Todas las marcas elegidas están protegidas contra este tipo de ataques ya que sus servidores no soportan SSLv2, el cual puede permitir descifrar las conexiones TLS entre servidor y cliente

SAMSUNG		APPLE	
Formas Ataque		Formas Ataque	
Man in the middle	Si (Sólo de Smartwach a Smartphone).	Man in the middle	No
Phishing (Relacionado con 2FA)	Si	Phishing (Relacionado con 2FA)	No
Inyeccion SQL	Posible	Inyeccion SQL	Posible
Ataque DROWN	No	Ataque DROWN	No
¿Posibilidad de ataque a 3º con nuestros datos?	Si	¿Posibilidad de ataque a 3º con nuestros datos?	Si

FITBIT		GARMIN	
Formas Ataque		Formas Ataque	
Man in the middle	No	Man in the middle	Si (Sólo de Smartwach a Smartphone).
Phishing (Relacionado con 2FA)	No disponible	Phishing (Relacionado con 2FA)	No disponible
Inyeccion SQL	No disponible	Man in the middle	No disponible
Ataque DROWN	No	Ataque DROWN	No
¿Posibilidad de ataque a 3º con nuestros datos?	No disponible	¿Posibilidad de ataque a 3º con nuestros datos?	No disponible

Image 11
Posibilidad de ataques a las distintas marcas (Samsung, Apple, Fitbit, Garmin)

5.4 Estándares de privacidad

“El recurso más valioso del mundo ya no es el petróleo, sino los datos”. Esta frase fue recogida en la conocida publicación semanal londinense, The Economist ([Economist, 2017](#)), con el objeto de alertar a la población del inminente poder de los datos. Y es que, cuando hablamos de inteligencia artificial no hablamos de otra cosa que de datos y algoritmos. Esto quiere decir que la IA aprende datos y por consecuencia, ofrece soluciones y/o recomendaciones. Es por este motivo que la seguridad en los datos debe estar regulada para no causar efectos devastadores, ya no solamente a nivel general, empresas y compañías, sino también a nivel individual; personas. Pero realmente cuando hablamos de ciberseguridad : ¿a qué nos estamos refiriendo? De acuerdo con el ingeniero electrónico, Juan A. Lloret Egea:

“La ciberseguridad es un área enfocada a proteger lo que extorsiona a distintas entidades mediante el uso de medios tecnológicos y agrede su intimidad,

privacidad, dinero y libertad.”

Los ataques cibernéticos cada vez son más frecuentes, de hecho, Juniper Research ([Research, 2021](#)), una entidad que proporciona servicios de investigación y análisis del sector tecnológico, emitía un informe reciente que informaba del coste producido por brechas de seguridad en Estados Unidos. La cantidad de la que estaríamos hablando sería de casi 4 millones de dólares y 400 billones; el coste para la economía global. ([Villas, 2020](#)).

No podríamos o al menos, no deberíamos hablar de manejo de datos, inteligencia artificial y seguridad sin mencionar la mayor autoridad mundial que existe: Openmined. Pero, ¿qué es Openmined y cuáles son sus funcionalidades? De acuerdo con su página oficial:

“Openmined es una comunidad de código abierto cuyo objetivo es hacer que el mundo preserve la privacidad al disminuir la entrada a las tecnologías privadas de IA.” ([Openmined, 2021](#))

Andrew Trask, el líder de Openmined, planteaba preguntas muy interesantes sobre la ciberseguridad y la ciencia de datos. Una de ellas estaba relacionada con el procesamiento y la gestión de datos sensibles como historiales médicos, estados financieros y hábitos privados en la que Trask insistía en que gran cantidad de datos que las personas producimos a través de *smartphones* o cualquier dispositivo electrónico necesitan ser reforzados, limitados y de difícil acceso para asegurar su privacidad. Para garantizar la seguridad y actuar de forma certera con la ciencia de datos en dominios privados, Openmined se apoya en las leyes de privacidad de datos, desarrollando el campo de la seguridad en la ciencia de datos ([Emma Bluemke & Kang, 2019](#)).

5.5 Normativa española, europea, latinoamericana e internacional del uso de datos

La gestión de datos, sobre todo los de carácter personal, deben estar regulados por una normativa o ley de seguridad. Para ello, todas las áreas o entidades geopolíticas tienen un reglamento que trata este tema. En este punto expondremos cómo trabajan la protección de datos la Unión Europea, Estados Unidos y Latinoamérica.

De acuerdo con la Comisión Europea, en el código se expone: qué datos son vulnerables, cómo se autoriza el tratamiento de datos, qué regulación existe para los menores y, cuáles son las multas correspondientes a las infracciones cometidas. En Europa, la seguridad de los datos está recogida en el Reglamento General de

Protección de Datos (RGPD), donde la Unión Europea establece unos requisitos a las organizaciones para gestionar los datos personales de los ciudadanos del continente. ([Reglamento, 2021](#)).

En este párrafo en concreto vamos a detallar los aspectos que nos parecen más importantes dentro de la regulación europea. En el rango de datos personales se incluyen: nombres y apellidos, dirección del protocolo Internet (IP) y datos en poder de hospitales o médicos. Según las normas de protección de datos de la UE, una información pasa a ser manipulable cuando: el interesado ha dado su consentimiento, existe una obligación legal, se protegen los intereses vitales del interesado y se tratan temas para una misión de interés público. Otro de los asuntos fundamentales concierne a las normas para menores; entre las que destacan el consentimiento parental en el uso de redes sociales y descargas de ciertos contenidos de Internet.

Ya, para terminar con el decreto europeo, referenciamos la gestión de la infracción de normas y sanciones de la que se obtiene la siguiente información:

“El incumplimiento del Reglamento General de Protección de Datos puede dar lugar a multas de hasta 20 millones de euros o del 4% del volumen de negocios mundial de la empresa, en determinadas infracciones.” ([Comisión Europea, 2021](#))

El consentimiento europeo se aplica en todos los países miembros de la UE. Sin embargo, cada país presenta una legislación propia consolidada. Por ejemplo, España tiene un diario oficial nacional dedicado a gestionar este tipo de leyes; el Boletín Oficial del Estado. El BOE publica la Ley Orgánica 3/2018, del 5 de diciembre sobre Protección de Datos Personales y garantía de los derechos digitales. Aunque la mayoría de los artículos recogidos en el BOE sobre la protección de datos quedan escudados por el Reglamento de la UE, encontramos algunos matices incluidos a modo de iniciativa, concretamente, el artículo 7, que trata sobre el consentimiento de los menores de edad. Aquí se puede leer cómo el tratamiento de los datos personales de un menor de edad únicamente podrá fundarse en su consentimiento cuando sea mayor de 14 años, marcando el factor edad como límite ([BOE, 2018](#)). Otros países de la UE como Alemania ha desarrollado su ley de protección de datos que, aunque es una adaptación del RGPD, el nuevo BDSG, que así se llama, podría cambiar la privacidad de datos alemanes ([Protección de datos en Alemania, 2018](#)).

Entre los cambios se desarrollan y especifican ciertas cuestiones tales como: un régimen sancionador para aquellas conductas en el ámbito de la protección de datos de carácter personal que excedan del ámbito del RGPD, lo que implicaría una sanción

de un máximo de 50.000 euros y obligaciones específicas de videovigilancia, relaciones laborales o evaluación de perfiles. ([Alemania_RGPD, 2017](#)).

“Gap es una de las miles de compañías preparadas para la nueva ley de privacidad del consumidor que requiere más transparencia sobre cómo se gestionan los datos del cliente.”

Éste es el pie de foto que acompaña una imagen de la tienda Gap en The Wall Street Journal. Y es que puede sorprender el título pero, “La primera ley de privacidad en línea de Estados Unidos entra en vigor en California a principios de 2020” tal y como recoge el periódico La Vanguardia. Por sus siglas en inglés se conoce como CCPA (ley de Privacidad del Consumidor) y es un símil al Reglamento General de Protección de Datos de la UE. ([Renter, 2020](#)).

Esta ley afectará principalmente a los gigantes tecnológicos como: Amazon, Google, Facebook y Microsoft además de las empresas que no tengan presencia física en California pero sí ofrezcan servicios al estado ([Haggin, 2019](#)). De acuerdo con las declaraciones de Molins Renter, esta norma dicta que si una empresa compra o vende datos de un mínimo de 50.000 residentes del estado de California en un año o bien, el 50% de sus ingresos económicos dimanen de la venta de datos vulnerables de sus clientes, la entidad debe emitir un informe sobre qué datos se manejan y lo más crucial, qué se está haciendo con esa información. Si todas estas pautas no se cumplen, la CCPA emitirá multas de 7.500 dólares por violación de la ley. Fue entonces cuando el conglomerado Facebook empezó a tener problemas, pues según recoge The Wall Street Journal:

“Facebook debe pagar una multa de 5 billones de dólares después de que la Comisión Federal de Comercio de Estados Unidos descubriera que la compañía engañó a los usuarios y administró incorrectamente sus datos personales.”

Otras multinacionales como Microsoft informaban de la aplicación de esta normativa a todo el país.

El Reglamento General de Protección de Datos de la Unión Europea fue pionero en favorecer la seguridad en la gestión de información vulnerable de los ciudadanos, velando tanto por sus derechos como deberes. A partir de ahí, diversas instituciones políticas latinoamericanas se vieron comprometidas a desarrollar y/o actualizar una ordenanza que amparara esta cuestión. El primer país en adaptar una regulación semejante al RGPD fue Brasil, con la Ley General de Protección de Datos (LGPD) en

2018, entrando en vigor en febrero de 2020 ([“La Visión de América Latina Sobre El Reglamento General de Protección de Datos,” 2020](#)). Pero, ¿en qué consiste la ley aprobada por Brasil? A pesar de su inspiración en el RGPD, la LGPD presenta un conjunto de perceptos diferentes en lo que a la base legal para el procesamiento de datos se refiere, por ejemplo: realiza estudios por entidades de investigación que aseguren, siempre que sea posible, el anonimato de los datos personales; protege la vida o la seguridad física del interesado o de un tercero y cumple con una obligación legal o reglamentaria del controlador ([BOE, 2020](#)). De la misma forma, Argentina también elaboró su propia Ley de Protección de Datos Personales (PDPA) ([LeyDatosArgentina, 2021](#)). El principal objetivo de estos países latinoamericanos es lograr una idoneidad en las transferencias de datos con la UE . Pocos son los países latinos que quedaron relegados a esta regulación, aunque Ecuador y Paraguay se encuentran en vías de desarrollo de una ley.

5.6 Privacidad y derecho de los ciudadanos en el conjunto de datos

Existen multitud de motores de búsqueda en la *web* que recopilan conjuntos de datos variados sobre casi cualquier tema de interés: política, economía, moda, medicina, música, cine, astrología, gastronomía, tecnología y educación, entre otros. En particular, Google Dataset Search tiene alrededor de 25 millones de conjuntos de datos. ([Heras, 2020](#)) En algunos casos, como en el ámbito sanitario, se proporcionan datos para tratar de identificar a personas, es decir, datos sometidos a vulnerabilidad.

Desde que empezó la pandemia del COVID-19, se ha visto incrementada la necesidad del seguimiento a los enfermos así como a sus contactos. A esto hay que añadir que un individuo afectado por COVID-19 puede haber contagiado a muchas otras personas sin que lo sepa y por tanto, no pueden ser fácilmente identificadas. Este tipo de situaciones podemos observarlas en la cola de un supermercado o cuando vamos a comprar ropa a una tienda. Fue a partir de ahí cuando los investigadores centraron su atención en la tecnología y singularmente en los *smart devices*, por su habilidad para la localización y proximidad a través de sensores como: GPS, WiFi y Bluetooth ([Ahmed et al., 2020](#)). Es por esto por lo que se desarrolló el contact tracing o rastreo de contactos.

“El contact tracing se fundamenta en encontrar a aquellas personas no reportadas como infectadas, rastreando a quien haya contraído la enfermedad de un caso positivo.” ([Leonie Reichert*, 2020](#))

Sin duda, tener una aplicación basada en el *contact tracing* puede ayudar en el control de la expansión de la pandemia pero, ¿qué hay de la privacidad del usuario? De conformidad con el IEEE, la arquitectura del sistema adoptada está estrechamente ligada a la forma de recabar y gestionar los datos recibidos. En las *apps* destinadas al uso del rastreo de infectados por COVID-19 se distinguen principalmente dos arquitecturas: centralizada y descentralizada, aunque también existe la arquitectura híbrida, una mezcla entre las dos anteriores. Los gráficos añadidos se toman como referencia de: A Survey of COVID-19 Contact Tracing Apps, publicado por el IEEE Access en julio 2020 ([Ahmed et al., 2020](#)).

Según la publicación tecnológica Xataka, un sistema centralizado identifica a los usuarios a nivel individual por medio de un servidor central que controlan las autoridades sanitarias y desde el que se maneja toda la información recibida ([Fernández, 2020](#)).

La siguiente figura muestra el funcionamiento del contact tracing en una aplicación. El diseño de arquitectura centralizada está basado en BlueTrace. BlueTrace es un protocolo que preserva la privacidad en el proceso de contact tracing, utilizando la conexión por Bluetooth, ya que permite una interoperabilidad global. BlueTrace fue creado para el registro de proximidad descentralizado y además complementa el rastreo de contactos centralizado y por parte de las autoridades sanitarias ([Jason Bay, 2020](#)).

El boceto se divide en ocho pasos. Para empezar, un usuario debe descargar la aplicación y registrar sus datos: nombre, número de teléfono, rango de edad en la que se encuentre y código postal en el servidor. Luego, el servidor verifica el número de teléfono, enviando una contraseña de un único uso (*Once Time Password*) por SMS (paso 1). Sobre esta verificación, el servidor registra un TempID, el cual está encriptado y solamente es válido durante 15 minutos. El TempID y el tiempo de

expiración son transferidos a la aplicación del usuario (paso 2). Una vez que un cliente entra en contacto con otra aplicación de usuario, intercambian el denominado mensaje de encuentro, por medio de Bluetooth. En este mensaje no se revela ningún tipo de información privada del usuario debido a la codificación del código TempID (paso 3). Llegados a este punto, ¿qué ocurre cuando hay un infectado por COVID-19? Las autoridades sanitarias confirman si el/la infectado/a tiene la aplicación instalada y si es así y esta persona da autorización, los datos quedan registrados en el servidor (paso 4 y 5). Iteración del servidor a través de la lista de los mensaje de encuentro, empleando valores de aproximación (paso 6). De esta forma, el servidor adquiere la información para ser procesada por las autoridades sanitarias (paso 7). En la última etapa, los centros médicos y hospitales alertan al servidor y posteriormente a los usuarios afectados de su posible exposición a la enfermedad (paso 8).

Conforme a las palabras del periodista tecnológico, Enrique Pérez, la relación entre usuarios y afectados por COVID-19 en los sistemas descentralizados es de uso privativo, esto es, las autoridades sanitarias no identifican los usuarios de forma personal y el usuario dispone de las funcionalidades básicas de la aplicación ([Pérez, 2020](#)).

Image 13

La descripción de la arquitectura descentralizada está basada en el protocolo de seguimiento de contactos automatizado privado (PACT). Como podemos observar, esta gráfica está dividida en 11 pasos. Previamente, el usuario debe registrarse en la aplicación (paso 1). Los dispositivos electrónicos generan semillas que utilizan en combinación con la hora marcada para generar pseudónimos y chirridos de aproximadamente un minuto. De esta forma, se preservará la privacidad (paso 2). Estos chirridos se intercambian con otros dispositivos de estrecho contacto periódicamente (pasos 3 y 4). Cuando se diagnostica un usuario positivo en COVID-19, (paso 5) las semillas generadas pueden cargarse al servidor si el usuario lo permite a través de un código de verificación que se envía al dispositivo del usuario (pasos 6, 7 y 8). De tal forma que el servidor sería un intermediario donde recoge qué usuarios están o no infectados. En el paso 9, el servidor comienza a almacenar los datos recibidos por parte de los usuarios. Después, otro usuario diferente puede registrarse en la aplicación, descargar los primeros datos de la base de datos del servidor (paso 10) y hacer su propio chequeo (paso 11) ([Ahmed et al., 2020](#)).

Para terminar con la explicación de los tipos de arquitecturas que utilizan las aplicaciones de rastreo, nos centramos en el sistema híbrido. Este es una combinación de la arquitectura centralizada y descentralizada. Por una parte, la generación y gestión del TempID es controlado por los dispositivos para asegurar la privacidad del usuario. Por otra, el análisis de riesgos y notificaciones queda a cargo del servidor. El objetivo de este sistema es dividir las funcionalidades entre el servidor y los dispositivos para lograr una mayor eficacia y sobre todo, seguridad en el rastreo de contactos ([Ahmed et al., 2020](#)).

En el continente europeo, la idea era crear una aplicación unificada para la prevención del COVID-19, en la que el rastreo por Bluetooth privado y el uso voluntario fueran las bases. Por supuesto, destacaron los valores de privacidad de los usuarios y el cumplimiento del Reglamento General de Protección de Datos. ([Fernández, 2020](#)) Pero, ¿qué dice España sobre este tema? La agencia española de protección de datos publicaba en mayo de 2020 un programa que analiza los costes y beneficios sobre el uso de las tecnologías en la lucha contra el COVID-19: geolocalización en redes sociales, geolocalización mediante la información recogida por los operadores de telecomunicaciones, cámaras infrarrojas, *chatbots*, aplicaciones de seguimiento de contactos y pasaportes digitales de inmunidad. Esto no solamente supone una amenaza para la economía, sino también para la seguridad, pues el anonimato no está completamente asegurado ([Datos, 2020](#)). Siguiendo esta línea, Manuel Carro, director

del IMDEA Software Institute de Madrid se pronunciaba con respecto a la seguridad de las *apps* de rastreo:

“La bomba es que ningún modelo garantiza toda la privacidad. ¿Privacidad hasta qué punto? Hay cosas que son inherentes a los modelos de contactos, sean centralizados o descentralizados.” ([Carro, 2020](#)).

Esta recopilación de información nos lleva a la conclusión de que todo el rendimiento que aporta la tecnología en el control y freno de la pandemia se convierte en temerario si detrás se esconde una mala gestión de los datos.

5.7 Uso privado de los datos: aprendizaje federado

Parafraseando al ingeniero, Juan A. Lloret Egea, *hay autoridades como gobiernos y hospitales que controlan los datos para su posterior inclusión en un dataset. El problema siempre de usar datos privados de personas, aunque sea con fines de salud pública e interés social, es que en la mayoría de los casos se pueden encontrar puntos de unión para terminar identificando a las personas. Por ejemplo, si unimos los términos descriptores: hombre, alto, tez morena, complexión delgada y menor de 30 años, al final acabaríamos diciendo un nombre. Pero, ¿cuáles son las iniciativas que se están planteando para el futuro en privacidad? el MIT Technology Review plantea una perspectiva que propone un uso de los datos anónimo y garante* ([Conner-Simons, 2020](#)). Todo empezó en el año 2017, cuando Google ofreció un nuevo enfoque del aprendizaje automático, de tal forma que su algoritmo aprendía de distintas fuentes distribuidas de datos en varios dispositivos. El aprendizaje federado o FL, como así denomina McMahan & Ramage (2017) a este proceso, es un enfoque general de “llevar el código a los datos, en vez de los datos al código” y aborda problemas fundamentales de privacidad, propiedad y localidad de los datos ([Bonawitz et al., 2019](#)).

Realmente, la estrategia del aprendizaje federado aplicada al acceso de la inteligencia artificial en el ámbito médico podría ser de gran ayuda para mantener un equilibrio entre la administración de la información y la privacidad de los pacientes, pues como afirma el profesor del Instituto Tecnológico de Massachussets y director del grupo de investigación Camera Culture del MIT Media Lab, Ramesh Raskar:

“Hay una falsa dicotomía entre la privacidad de los datos del paciente y la utilidad de estos para la sociedad. La gente no se da cuenta de que la arena se está moviendo bajo sus pies y que, de hecho, ahora se puede obtener privacidad y utilidad al mismo tiempo.” ([Hao, 2019](#)).

De hecho, existen varios artículos que corroboran la argumentación de R. Raskar, por ejemplo:

“El aprendizaje federado diferencialmente privado para la predicción del cáncer.”
([Beguiet et al., 2021](#)).

Esta publicación compendia los resultados tras lograr ecuanimidad entre el rendimiento de la predicción y el presupuesto de privacidad. Para ello, se entrena un modelo supervisado en la predicción de la aparición del cáncer de mama a partir de datos genómicos, divididos en dos centros virtuales. “Aprendizaje federado para la predicción del teclado de los móviles” ([Hard et al., 2019](#)) es otra publicación que narra las múltiples ventajas del aprendizaje federado. En este modelo, el algoritmo federado permite una mayor calidad en el conjunto de datos, logrando así una mejor recuperación de la predicción. En términos generales, el aprendizaje federado brinda a los usuarios un mayor control sobre el uso de sus datos porque lo único que se transfieren son los algoritmos y no los datos sin procesar. ([Hao, 2019](#)).

6. Ejemplos de APPs que utilizan datos, idiomas de disponibilidad y repercusión de la tecnología en el comportamiento humano

La inteligencia artificial y el Big Data se han convertido en un arma muy poderosa contra la pandemia mundial del coronavirus, permitiendo realizar grandes avances en poco tiempo. Sin embargo, un uso inadecuado de ellas podría suponer un gran riesgo para la población.

Ya, en el siglo XX, algunos de los pensadores y escritores más ilustres de la época, como George Orwell y Aldous Huxley, pronosticaban tópicos que actualmente ya están ocurriendo. Orwell en su obra de ciencia ficción, *1984*, hablaba de una sociedad completamente digitalizada y controlada por la tecnología, capaz de apoderarse de vidas humanas. De igual forma, Aldous Huxley en su utópica obra, *Un mundo feliz*, describía una sociedad inmersa en la felicidad y tecnológicamente avanzada. ([Ruiza, 2004](#)).

Según un estudio de la fundación BBVA, actualmente un 60% de los españoles consideran Internet una pieza fundamental en sus vidas, frente al 18% en el año 2008. Además el 90% hace uso de Internet a diario, mientras que 1 de cada 3 personas reconoce estar conectada a la red casi todo el día, ([FundaciónBBVA, 2021](#)). ¿Acaso no estamos delante de una sociedad completamente digitalizada? Aunque quizás, la cuestión más sensata que deberíamos plantearnos es: la tecnología ya domina nuestra vida, ¿vamos a permitir que algunos lo hagan a través de ella?

China por su parte cree que sí. De hecho, el Gobierno está promoviendo el denominado crédito social que surgió en 2018 y que aún se está probando a niveles locales. Pero, ¿qué es esto del crédito social chino? y ¿cómo afecta a su población? Este tipo de crédito incluye a los ciudadanos en una lista negra en función de cómo actúen, es decir, primero, recopilan datos de sus actos, después, en base a un algoritmo, se genera una puntuación y se decide si entras a formar parte de esa lista negra (la cual te privará de ciertos derechos) y durante cuánto tiempo. Por si esto no fuera suficiente, gigantes tecnológicos como Alibaba se unen a este tipo de iniciativa con Zhima Credit. Este es otro tipo de crédito que nace para premiar a los usuarios por su buenas acciones. Las recompensas se reciben en forma de entradas a salas VIPs en aeropuertos o incluso pedir préstamos a un interés mucho más favorable. La cuestión real es que Zhima Credit no sólo recopila información de cómo te mueves, por dónde y en qué establecimientos pagas, sino que también tiene acceso a tu saldo en la cuenta bancaria y a las posibles deudas, en caso de que existieran. Este tipo de actos permiten sumar puntos y formar parte de la red de contactos “ciudadanos ejemplares”. ([Finacial, n.d.](#))

Otro de los casos que llama especialmente nuestra atención es el sistema de códigos QR que China propone usar a nivel global para contener la COVID-19. Con este método, los usuarios introducen la información solicitada y reciben en su *smartphone* un código de salud en tres colores: el verde permite la libre circulación y el naranja y rojo obligan a que la persona tenga que estar en cuarentena hasta dos semanas. Teóricamente, ésta era una forma de combatir la pandemia, en cambio, la ciudad de Hangzhou ya planea una versión permanente que permita asignar una puntuación basada en el historial médico, chequeos y hábitos de estilo de vida. ([Mundo, 2020](#))

En cualquier caso, es normal que todo este control poblacional lo estemos viendo de manera más incipiente en China y es que según informa Growth from Knowledge, tan solo el 8% de los internautas chinos no cederían sus datos a cambio de una recompensa. ([Statista, 2017](#))

% de Internautas que no cederían sus datos personales por una recompensa

Image 14
% de Internautas que no cederían sus datos personales por una recompensa

7. El español en datos

Durante el desarrollo de este trabajo, hemos incluido múltiples datos así como la vinculación que tienen con las lenguas castellana e inglés. Sin embargo, no podíamos concluir sin incluir cómo está la lengua española en el mundo; en datos y gráficos.

Con el paso de los años, observamos como el inglés es uno de los idiomas más importantes del mundo, convirtiéndose en una lengua franca y universal. Europa Press ([Epdata, 2021](#)) hace público un dato que confirma la revista Ethnologue sobre que el español se encuentra entre las lenguas más habladas del mundo. Y es que, a pesar de no ser declarado idioma oficial en más de 20 países, el español es la segunda lengua más hablada en el mundo, así lo dice Ethnologue:

Image 15
Lenguas maternas más habladas en el mundo

A diferencia del castellano, el inglés tiene establecido la propiedad de idioma oficial en más de 50 países. Esto no solamente influye en la expansión de la lengua, sino que quizás, éste puede ser uno de los puntos clave que le permita a la lengua inglesa abarcar en demasía, con respeto a otras lenguas; importantes y novedosos asuntos como es el estudio y la investigación en el campo de la inteligencia artificial. Siguiendo esta línea, es interesante conocer la evolución de los hablantes nativos de español en el mundo e indagar en el número de hablantes del castellano, tanto en países hispanohablantes como los países no hispanohablantes. De esta forma, se podrá verificar si el castellano está quedando relegado ante la invasión del inglés. Esto es lo que nos dice el Instituto Cervantes ([InstitutoCervantes, 2019](#)):

Image 16
Evolución de los hablantes nativos de español en el mundo

Alrededor de un 7% de la población mundial habla español como lengua nativa y esta cifra se ha visto incrementada a lo largo de los últimos ocho años. Esto quiere decir que el castellano no es una lengua que se encuentre en decadencia ([Epdata, 2021](#)). De hecho, en la siguiente figura vemos como el número de hablantes de español crecerá en el futuro, superando incluso al número de hablantes nativos de inglés y francés, de acuerdo con el Instituto Cervantes:

Con los datos en la mano y sabiendo que el español es una de las lengua más poderosas y ricas que existen, ¿por qué se halla un enorme sesgo cultural en áreas como la educación tecnológica y la investigación en el campo de la IA? Si alrededor del mundo hay más hispanohablantes, ¿por qué es predominante la lengua inglesa en ciencia de datos, en machine learning, en programación del lenguaje natural, en deep learning, en redes neuronales y en ciberseguridad? ¿Por qué los estudiantes hispanoamericanos no tenemos acceso a una educación digna en estos temas, si el español es la segunda lengua más hablada del planeta después del chino? Ante la falta de soluciones, es importante conocer qué o quiénes nos está/n llevando a esta demoledora situación. Las respuestas a estas preguntas están regidas por la regla de las ocho faltas: falta de liderazgo político, falta de liderazgo tecnológico, falta de líderes concienciados, falta de recursos económicos, falta de herramientas en educación, falta de líderes en posiciones estratégicas donde se toman las grandes

decisiones, falta de concienciación social y falta de interés.

8. Aspectos generales

Teniendo en cuenta que nuestra sociedad se dirige cada vez más a una sociedad tecnológica, y como acabamos de ver, esa tecnología, o en otras palabras, los datos que recoge y maneja, pueden ser utilizados en ocasiones para fines poco éticos, con un uso lucrativo o incluso criminal. Uniendo este aspecto a que estamos relegando que toda esta información se utilice y maneje mayoritariamente en Inglés, deberíamos plantearnos seriamente, ¿qué control nos deja esto de nuestros datos y de esta tecnología? Si no somos capaces ni de formarnos, ni de formar a nuestra sociedad en su propia lengua materna, ¿quién o qué asumirá el control de las decisiones que tiene que tomar la tecnología para seguir avanzando?, ¿hacia dónde nos dirigimos?, ¿coincide con hacia adónde nos queremos dirigir?

9. Conclusiones

La ciberseguridad es un área vital en la inteligencia artificial porque mejora los servicios y sistemas que ofrece, proporcionándole estabilidad y privacidad. Por su parte, la ciencia de datos analiza conjuntos de datos con el objetivo de precisar un resultado, por lo que la convierte en un área muy exacta y necesaria en la IA. Con los resultados en la mano de nuestro anterior proyecto y a sabiendas de que el inglés es la lengua por excelencia en temas de tecnología, en general, e inteligencia artificial, en particular, ponemos el punto de mira en la herramienta esencial que necesita un científico de datos para trabajar; los *datasets*. ¿qué disponibilidad cultural hay entre los principales motores de búsqueda que albergan millones de conjuntos de datos? Para dar respuesta a esta pregunta, hemos diseñado una base de datos a partir de la que se han generado dos conjuntos de datos en español e inglés, respectivamente. Estos *datasets* se agrupan por varios campos: título, tema, idioma, fecha de modificación y url.

Visit the web version of this article to view interactive content.

Conjunto de datasets en inglés

Visit the web version of this article to view interactive content.

Conjunto de datos en español

La evidencia de los resultados demuestra que existe una gran discriminación cultural entre el inglés y el español. Para conocer más sobre la metodología y los objetivos de estos conjuntos de datos, visiten este enlace: <https://github.com/Itvia/Los-datos-como-eje-principal-de-la-ciencia-de-datos-en-el-idioma-espa-ol-y-la-IA>

Ya, para cerrar esta conclusión y dando paso a nuevas oportunidades, nos parece muy oportuno citar el nuevo reglamento sobre inteligencia artificial que presentó la Comisión Europea el día 22 de abril de 2021. Un día inolvidable para la Unión Europea ya que esta preceptiva brinda apoyo al despliegue de la inteligencia artificial en Europa. ([CENCENELEC, 2021](#))

Citations

1. Economista, E. (2020). *La innovación tecnológica se atasca en España y despunta en Europa*. <https://www.eleconomista.es/ranking-innovacion-universidades/noticias/10537379/05/20/La-innovacion-tecnologica-se-atasca-en-Espana-y-despunta-en-Europa.html> ↵
2. Epdata. (2020). *España en la prueba PISA, en datos y gráficos*. <https://www.epdata.es/datos/espana-pisa-datos-graficos/484> ↵
3. BBVA. (2018). *¿Qué es la inteligencia artificial y cómo funciona?* <https://www.bbva.com/es/que-es-la-inteligencia-artificial/> ↵
4. Calvera, J. (2020). *Los datos son el combustible de la inteligencia artificial*. <https://www.computerworld.es/tendencias/los-datos-son-el-combustible-de-la-inteligencia-artificial> ↵
5. TechTarget, Colab. (2019). *Los datos son el combustible de la inteligencia artificial*. <https://searchdatacenter.techtarget.com/es/definicion/Scrubbing-o-depuracion-de-datos> ↵
6. Mulligan, D. K., & Schneider, F. B. (2011). Doctrine for Cybersecurity. *Daedalus*, 140(4), 70–92. https://doi.org/10.1162/daed_a_00116
<https://direct.mit.edu/daed/article/140/4/70/26918/Doctrine-for-Cybersecurity> ↵

7. Banafa, A. (2018). *Inteligencia Artificial en ciberseguridad: retos*.
<https://www.bbvaopenmind.com/tecnologia/inteligencia-artificial/inteligencia-artificial-en-ciberseguridad-retos/> =
8. Stacy Stanford, R. I. (2018). <https://sitiobigdata.com/2019/12/24/datasets-publicos-para-aprendizaje-automatico/#> =
9. Google. (2018). <https://datasetsearch.research.google.com/> =
10. Goldbloom, A. (2010). <https://www.kaggle.com/> =
11. BancoMundialdeDatos. (1944). <https://datos.bancomundial.org/> =
12. Tecnología, I. (2020). <https://www.unir.net/ingenieria/revista/r-vs-python/> =
13. Yamamoto, N., Ishida, A., Ogitsuka, K., Oishi, N., & Murakami, J. (2021). Development of Online Learning Material for Data Science Programming Using 3D Puzzle. *International Journal of Information and Education Technology*, 11(4), 154-163. <https://doi.org/10.18178/ijiet.2021.11.4.1505> =
14. Robledano, Á. (2019). <https://openwebinars.net/blog/que-es-python/> =
15. Bagnato, J. I. (2020). <https://blogs.solidq.com/es/poder-del-dato/10-librerias-python-para-data-science-y-machine-learning/> =
16. Chiavegatto Filho, A. D. P. (2015). Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, 24(2), 325-332. <https://doi.org/10.5123/s1679-49742015000200015> =
17. Parada, M. (2021). <https://openwebinars.net/blog/r-vs-python-en-que-se-diferencian-y-en-que-se-parecen/> =
18. Tecnología, I. (2020). <https://www.unir.net/ingenieria/revista/r-vs-python/> =
19. Anaconda. (2021). <https://www.anaconda.com/products/individual> =
20. Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017, June). Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). <https://doi.org/10.1109/jcdl.2017.7991618> =
21. Ding, I.-K. (2016). <https://www.quora.com/profile/I-Kang-Ding> =

22. Cristóbal, J. (2016). <https://markdown.es/> ↵
23. Wickham, H. (2016). *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. O'Reilly Media. ↵
24. Jupyter. (2021). <https://jupyter.org/> ↵
25. Zhang, Y. T. (2017). Is LaTeX a useful skill for Data Science? In *Quora*.
<https://www.quora.com/Is-LaTeX-a-useful-skill-for-Data-Science> ↵
26. Britannica. (2021). <https://www.britannica.com/technology/Java-computer-programming-language> ↵
27. Setién, M. (2017). https://www.abc.es/familia/educacion/abci-mayoria-menores-no-conscientes-riesgos-internet-201705111609_noticia.html ↵
28. Yañez, C. (2015). <https://www.ceac.es/blog/que-ventajas-o-desventajas-tienen-los-smartwatches> ↵
29. Karim, S., Aqil, S., & Mahmood, N. (2020). An Analysis of Human Activities Recognition using Smartwatches Dataset. *International Journal of Advanced Computer Science and Applications*, 11(12).
<https://doi.org/10.14569/ijacsa.2020.0111241>
<https://thesai.org/Publications/ViewPaper?Volume=11&Issue=12&Code=IJACSA&SerialNo=41> ↵
30. Judd, B. (2020). https://www.abc.net.au/news/2020-01-08/gps-trackers-for-children-what-do-you-need-to-know/11627208?utm_source=abc_news_web&utm_medium=content_shared&utm_content=link&utm_campaign=abc_news_web ↵
31. Morgenstern, M. (2019). <https://www.iot-tests.org/2019/11/product-warning-chinese-childrens-watch-reveals-thousands-of-childrens-data/> ↵
32. Clausing, E. (2018). <https://www.iot-tests.org/2018/08/nokia-steel-hr-no-weaknesses-in-the-security-check/> ↵
33. Clausing, E. (2018). <https://www.iot-tests.org/2018/08/huawei-band-2-pro-fit-and-secure/> ↵
34. Stykas, V. (2019). <https://www.pentestpartners.com/security-blog/gps-watch-issues-again/> ↵

35. Clausing, E. (2018). <https://www.iot-tests.org/2018/02/carl-kids-deceptive-security-promise/> =
36. Henke, D. (2018). <https://www.iot-tests.org/2018/02/helloo-childrens-smart-watch/> =
37. Morgenstern, M. (2019). <https://www.iot-tests.org/2019/11/product-warning-chinese-childrens-watch-reveals-thousands-of-childrens-data/> =
38. OZA, S. (2020). <https://spanning.com/blog/insecure-direct-object-reference-web-based-application-security-part-6/> =
39. Foundation_Mozilla. (2020). <https://foundation.mozilla.org/es/privacynotincluded/samsung-galaxy-watch-3/> =
40. Foundation_Mozilla. (2020). <https://foundation.mozilla.org/es/privacynotincluded/apple-watch-6/> =
41. Foundation_Mozilla. (2020). <https://foundation.mozilla.org/es/privacynotincluded/fitbit-versa-3/> =
42. Foundation_Mozilla. (2020). <https://foundation.mozilla.org/es/privacynotincluded/garmin-forerunner-series/> =
43. Program, C. S. (2020). <https://privacy.commonsense.org/privacy-report/Fitbit> =
44. Program, C. S. (2020). <https://privacy.commonsense.org/privacy-report/Apple-Watch> =
45. Program, C. S. P. (2020). <https://privacy.commonsense.org/privacy-report/Garmin-Vivofit-Jr>. =
46. Program, C. S. P. (2020). <https://privacy.commonsense.org/privacy-report/Samsung-Galaxy-Watch> =
47. Economist, T. (2017). <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> =
48. Research, J. (2021). <https://www.juniperresearch.com/home> =
49. Villas, M. (2020). <https://sherpa.ai/blog/es/la-importancia-de-cyber-security-data-science/> =

50. Openmined. (2021). <https://www.openmined.org/> ↵
51. Emma Bluemke, A. T., Antonio Lopardo, & Kang, N. (2019). ↵
52. Reglamento. (2021). <https://truben.no/latex/bibtex/> ↵
53. Comisión Europea. (2021). https://ec.europa.eu/info/index_es ↵
54. BOE. (2018). <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673> ↵
55. Protección de datos en Alemania. (2018). <https://leyderecho.org/proteccion-de-datos-en-alemania/> ↵
56. Alemania_RGPD. (2017). <https://www.lexology.com/library/detail.aspx?g=e9d3d53c-dd6b-4a81-a543-64bf221ee745> ↵
57. Renter, A. M. (2020).
<https://www.lavanguardia.com/vida/20200105/472713380363/california-estados-unidos-privacidad-consumidor-e-commerce-comercio-electronico.html#:~:text=La%20CCPA%20establece%20multas%20de,hizo%20en%20la%20Uni%C3%B3n%20Europea.> ↵
58. Haggin, P. (2019). <https://www.wsj.com/articles/businesses-across-the-board-scramble-to-comply-with-california-data-privacy-law-11567947602> ↵
59. La Visión de América Latina sobre el Reglamento General de Protección de Datos. (2020). *Comentario Internacional*, 99-112.
<https://doi.org/10.32719/26312549.2019.19.4>
<https://revistas.uasb.edu.ec/index.php/comentario/article/view/1546> ↵
60. BOE. (2020). <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673> ↵
61. Ley Datos Argentina. (2021). <https://docs.microsoft.com/es-es/compliance/regulatory/offering-pdpa-argentina#:~:text=De%20acuerdo%20con%20la%20Constituci%C3%B3n,de%20datos%20p%C3%BAblicas%20y%20privadas.> ↵
62. Heras, J. M. (2020). <https://www.iartificial.net/google-dataset-search-descubre-conjuntos-de-datos/> ↵
63. Ahmed, W., Bertsch, P. M., Bivins, A., Bibby, K., Farkas, K., Gathercole, A., Haramoto, E., Gyawali, P., Korajkic, A., McMinn, B. R., Mueller, J. F., Simpson, S. L.,

- Smith, W. J. M., Symonds, E. M., Thomas, K. V., Verhagen, R., & Kitajima, M. (2020). Comparison of virus concentration methods for the RT-qPCR-based recovery of murine hepatitis virus, a surrogate for SARS-CoV-2 from untreated wastewater. *Science of The Total Environment*, 739, 139960.
<https://doi.org/10.1016/j.scitotenv.2020.139960>
<https://www.sciencedirect.com/science/article/pii/S004896972033480X> ↵
64. Leonie Reichert*, B. S. **, Samuel Brack*. (2020).
<https://eprint.iacr.org/2020/375.pdf> ↵
65. Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., Seneviratne, A., Hu, W., Janicke, H., & Jha, S. K. (2020). A Survey of COVID-19 Contact Tracing Apps. *IEEE Access*, 8, 134577–134601.
<https://doi.org/10.1109/access.2020.3010226>
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9144194> ↵
66. Fernández, S. (2020). <https://www.xatakamovil.com/aplicaciones/asi-app-que-quiere-gobierno-espana-para-frenar-coronavirus-uso-voluntario-unificada-toda-europa> ↵
67. Jason Bay, T. A. Q., Joel Kek, Alvin Tan, Chai Sheng Hau, Lai Yongquan, Janice Tan. (2020). https://bluetrace.io/static/bluetrace_whitepaper-938063656596c104632def383eb33b3c.pdf ↵
68. Pérez, E. (2020). <https://www.xataka.com/privacidad/descentralizado-vs-centralizado-que-dos-sistemas-rastreo-proximidad-diferentes-uso-nuestros-datos> ↵
69. Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., Seneviratne, A., Hu, W., Janicke, H., & Jha, S. K. (2020). A Survey of COVID-19 Contact Tracing Apps. *IEEE Access*, 8, 134577–134601.
<https://doi.org/10.1109/access.2020.3010226> ↵
70. Fernández, S. (2020). <https://www.xatakamovil.com/aplicaciones/asi-app-que-quiere-gobierno-espana-para-frenar-coronavirus-uso-voluntario-unificada-toda-europa> ↵
71. Datos, Agencia. Española. P. (2020). *EL USO DE LAS TECNOLOGÍAS EN LA LUCHA CONTRA EL COVID19. UN ANÁLISIS DE COSTES Y BENEFICIOS*.
<https://www.aepd.es/sites/default/files/2020-05/analisis-tecnologias-COVID19.pdf> ↵

72. Carro, M. (2020). <https://maldita.es/malditatecnologia/20200430/contact-tracing-rastreo-contactos-apps/> ↵
73. Conner-Simons, A. (2020). <https://news.mit.edu/2021/mit-launches-data-privacy-focused-initiative-fod-0420> ↵
74. Bonawitz, et.al. K. (2019). TOWARDS FEDERATED LEARNING AT SCALE: SYSTEM DESIGN. *ArXiv*, Vol2(1902.01046v2 [cs.LG]), 1. <https://arxiv.org/pdf/1902.01046.pdf> ↵
75. Hao, K. (2019). <https://www.technologyreview.es/s/11017/aprendizaje-federado-la-nueva-arma-de-ia-para-asegurar-la-privacidad> ↵
76. Beguier, et.al., C. (2021). Differentially Private Federated Learning for Cancer Prediction. *ArXiv*, 2101.02997v2, 1. <https://arxiv.org/pdf/2101.02997.pdf> ↵
77. Hard, et.al., A. (2019). FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION. *ArXiv*, 1811.03604v2, 1. <https://arxiv.org/pdf/1811.03604.pdf> ↵
78. Ruiza, E., M. J. Fernández, T. y Tamaro. (2004). *George Orwell*. <https://www.biografiasyvidas.com/biografia/o/orwell.htm> ↵
79. FundaciónBBVA. (2021). *El proceso de digitalización de la sociedad española se consolida y se amplía en el marco de la pandemia con nuevos usos como el teletrabajo*. <https://www.fbbva.es/noticias/el-proceso-de-digitalizacion-de-la-sociedad-espanola-se-consolida-y-se-amplia-en-el-marco-de-la-pandemia-con-nuevos-usos-como-el-teletrabajo/> ↵
80. Financial, A. (n.d.). *Zhima Credit*. <https://gw.alipayobjects.com/os/rmsportal/nGVZldPeZtcyBkbhojbi.pdf> ↵
81. Mundo, B. N. (2020). *Covid-19: el sistema de códigos QR que China propone usar a nivel global para contener el coronavirus (y por qué causa polémica)*. <https://www.bbc.com/mundo/noticias-55049848> ↵
82. Statista. (2017). <https://es.statista.com/grafico/7860/venderias-tus-propios-datos-personales-si-pudieras/> ↵
83. Epdata. (2021). <https://www.epdata.es/datos/lengua-espanola-mundo-datos-graficos/513> ↵

84. InstitutoCervantes. (2019).

https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2019.pdf ⇐

85. CENCENELEC. (2021). *European Standards support the EU ambitions on*

Artificial Intelligence. https://www.cencenelec.eu/News/Brief_News/Pages/TN-2021-020.aspx ⇐