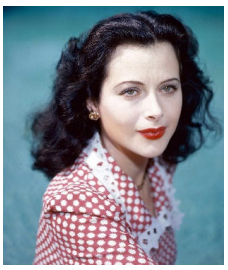
¿Habrá suficiente energía para el desarrollo global de la inteligencia artificial?

ALBERTO PRIETO ESPINOSA
Profesor Emérito del Departamento de Ingeniería de Computadores, Automática y Robótica (Univ. de Granada)



Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad de Málaga
23 de febrero 2024

Cátedra Mujer y Tecnología Hedy Lamarr de la Universidad de Málaga
Desafío Lamarr & Concurso Hilda Ericsson 2023/24

Contenido

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Concepto de Inteligencia Artificial. ¿Qué entendemos por inteligencia?
- Ejemplo: reconocimiento de dígitos decimales escritos manualmente
- Análisis de la contribución de las TIC al consumo energético
- Cómo afectan las TIC al medio ambiente
- Estimaciones sobre la evolución del consumo energético de computación y tráfico de datos digitales
- Algunos datos y situaciones prácticas sobre el consumo de programas de Inteligencia Artificial.
- Procedimientos para reducir el consumo energético en TIC
- Conclusiones

Concepto de Inteligencia artificial. ¿Qué entendemos por Inteligencia Artificial?

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

3

Qué se entiende por Inteligencia Artificial

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Definición sencilla:
  - La Inteligencia Artificial (IA) es la capacidad de un dispositivo (como un móvil, ordenador o robot) para realizar tareas que generalmente requieren inteligencia humana.
  - Definición mas rigurosa:
    - La Inteligencia Artificial es la disciplina de la Informática que desarrolla sistemas inteligentes, tomando como referencia la inteligencia humana.
      - Ejemplos de IA incluyen los sistemas de reconocimiento de voz y de caras, los coches autónomos, los buscadores en internet o los sistemas que nos recomiendan música o videos.
- Inteligencia (RAE):
  - Capacidad de entender o comprender o de resolver problemas.

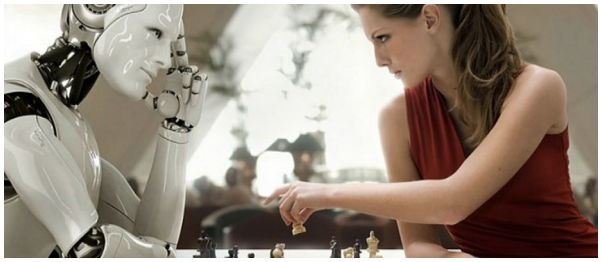
Oliver, N. (2021). Inteligencia artificial, naturalmente: un manual de convivencia entre humanos y máquinas para que la tecnología nos beneficie a todos. *Inteligencia artificial, naturalmente*, 1-161.

4

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto.
23/02/2024

¿Cuál es la diferencia entre Inteligencia Artificial y Humana?

- Actualmente la inteligencia artificial se limita a tareas específicas, muchas de las cuales se pueden hacer mejor que el mejor humano (si se entrena con millones de ejemplos).
- Sin embargo, la inteligencia humana es increíblemente flexible, capaz de aprender cientos de habilidades a partir de muy pocos ejemplos y aplicar y combinar esas habilidades en las situaciones en las que tienen sentido.



Oliver, N. (2021). Inteligencia artificial, naturalmente: un manual de convivencia entre humanos y máquinas para que la tecnología nos beneficie a todos. *Inteligencia artificial, naturalmente*, 1-161.

A. Prieto 5

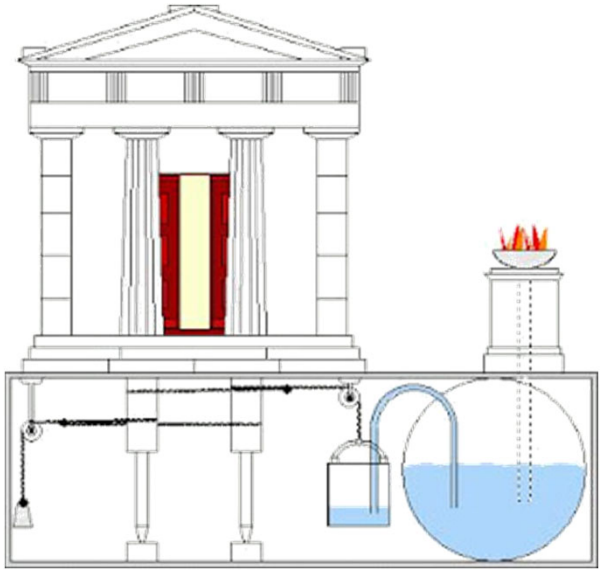
Pero, ¿qué entendemos por inteligencia?

- Ejemplos:
  - Sistemas automáticos (Templo de Heron)
  - Cálculo. Operaciones matemáticas
  - Cerebros electrónicos; cibernética.
  - Jugar al ajedrez
  - Actualidad: en cierta medida Inteligencia Artificial sinónimo de programa informático.

A. Prieto 6

¿Qué entendemos por inteligencia?

- Siglo III a de C (autómatas)
  - Puerta templo de Herón de Alejandría
- Edad Media a Siglo XVIII:
  - Autómatas, cabezas parlantes, etc.



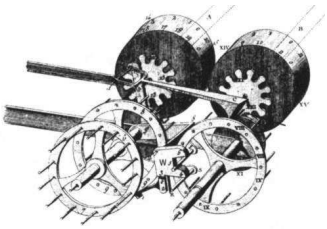
A. Prieto 7

¿Qué es inteligencia?

- Calcular
  - Hasta sólo hace poco más de un siglo sólo las personas extraordinariamente cultas sabían calcular
  - Manuscrito anónimo de época de la reina Elizabeth, 1570:
    - "La multiplicación es una vejación"
    - "La división es un mal"
    - "La regla de tres es un rompecabezas, y utilizarlos me vuelven loco"

A. Prieto

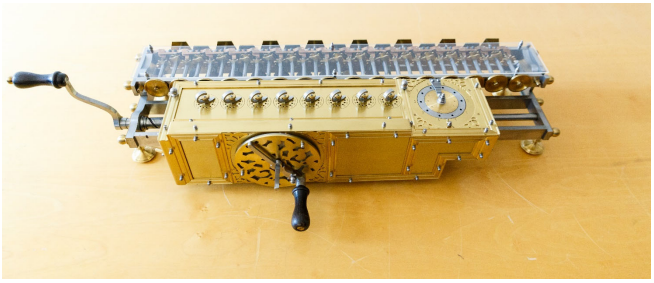
Sumas y restas: Pascalina de Blaise Pascal (1642) calculadora mecánica para + y – (ruedas dentadas y engranajes)

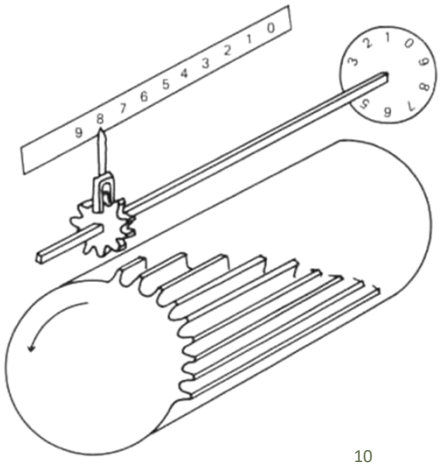


A. Prieto 9

¡Multiplicaciones y divisiones!

- Gottfried W. Leibnitz (Leipzig, 1646-1716)
  - Calculadora mecánica para +, -, x, / (rueda de Leibnitz)

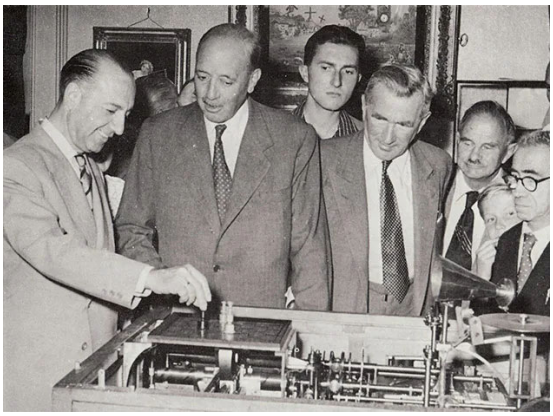


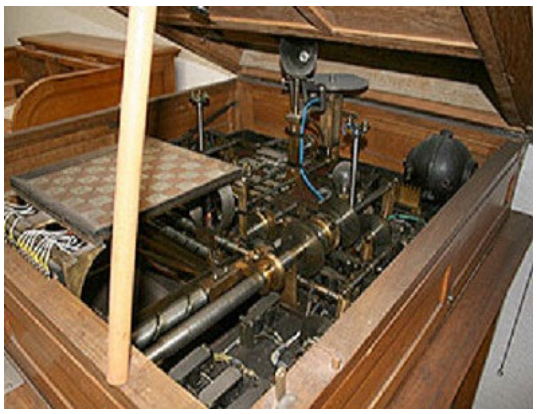


A. Prieto 10

¿Qué es inteligencia?

- Leonardo Torres Quevedo, en 1912 presentó El Ajedrecista (París, 1914)

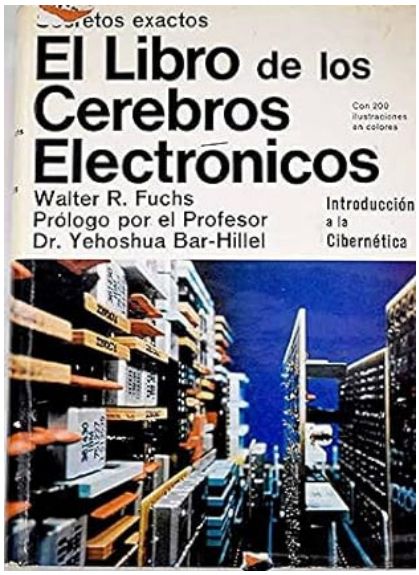




A. Prieto 11

¿Qué es inteligencia?

- ENIAC (1946). El informe sobre su diseño indicaba que se pretendía emular al cuerpo humano (cerebro, etc.).
- En las décadas de los 1950 y 1960 a los ordenadores o computadores les llamábamos "**Cerebros Electrónicos**", y a la Informática, "**Cibernética**".
  - Libro de 1969



A. Prieto 12

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto. 23/02/2024

¿Qué es inteligencia?

- Se decía hasta la década de los 1980. **"Un computador nunca superara a la inteligencia humana, y prueba de ello es que nunca será capaz de ganar al ajedrez a un humano experto"**
- Juego del ajedrez
  - 1997 Deep blue. Primer ordenador capaz de derrotar a un campeón del mundo vigente en ajedrez, Kaspárov. (El computador gano a 6 partidas por 3½ a 2½).
- En diciembre de 2017, AlphaZero, de DeepMind, venció al mejor jugador de ajedrez del mundo (programa Stockfish).
  - **Le bastó conocer las reglas del ajedrez y dedicar cuatro horas a entrenarse, jugando contra sí mismo millones de veces**.

¿Qué es inteligencia?

- Hoy día, en una multitud de aplicaciones informáticas se incluyen programas de Inteligencia Artificial, cubriendo la practica totalidad de ámbitos tan diversos como medicina, energía, transporte y educación, y, en general, investigación científica, sistemas de producción, logística, servicios digitales y prestación de servicios públicos y privados.

¿Qué es inteligencia?

- En la actualidad es tan amplio y mediático el término de **Inteligencia Artificial** que vulgarmente (y sin rigor) se esta llegando a utilizar este término como sinónimo de **programa informático**.
  - Dispositivos o sistema inteligente: cualquier dispositivo con sensores, actuadores de salida y posibilidad de comunicación a través de redes (Bluethoth, WiFi, celulares, etc.) con otros dispositivos. En ellos se programan "tomas de decisiones"
    - Internet de las Cosas, teléfonos y relojes "inteligentes", etc.
  - En los medios se dicen, sin rigor, frases como "Han hecho una Inteligencia Artificial para regular la ORA (zona azul) de aparcamientos de Granada".

Dos enfoques de la Inteligencia Artificaal

- De arriba abajo *(Top-down)*
  - Parte de la descripción precisa (algoritmizada) del proceso inteligente humano de forma que se pueda programar en un lenguaje informático para emularlo en un computador.
  - Ejemplo: sistema experto
- De abajo a arriba (conexionista) *(Down-top)*
  - El computador máquinas genera abstracciones a partir de estímulos sensoriales y otros datos.
  - Aprendizaje en máquinas, redes neuronales artificiales, etc.

Ejemplo: reconocimiento de dígitos decimales escritos manualmente

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

17

Ejemplo (conexionista): reconocimiento de dígitos decimales escritos manualmente

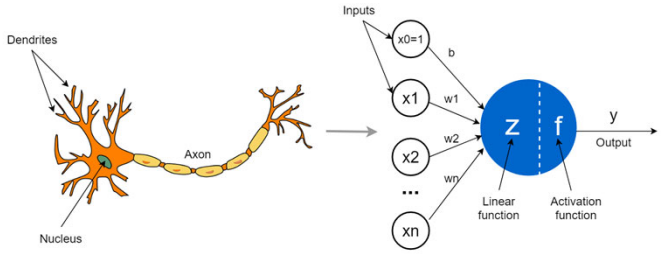
ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Reconocer si lo escrito es un 0 o un 1, etc.; independientemente de lo mal escrito que este
- Una aplicación pionera desarrollada en los EEUU para la distribución automática por los servicios de correos de cartas y paquetes según el código postal.

A. Prieto

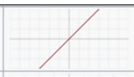
18

Elemento de elemento de computación: neurona artificial



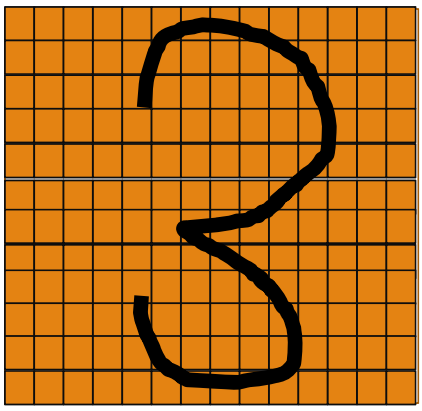
Función de activación "salto"

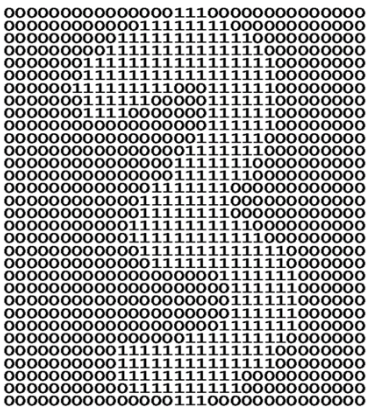
$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{threshold} \end{cases}$$

Identity		$f(x) = x$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Logistic (a.k.a. Sigmoid or Soft step)		$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$ [1]
TanH		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$

A. Prieto 19

Captación de datos





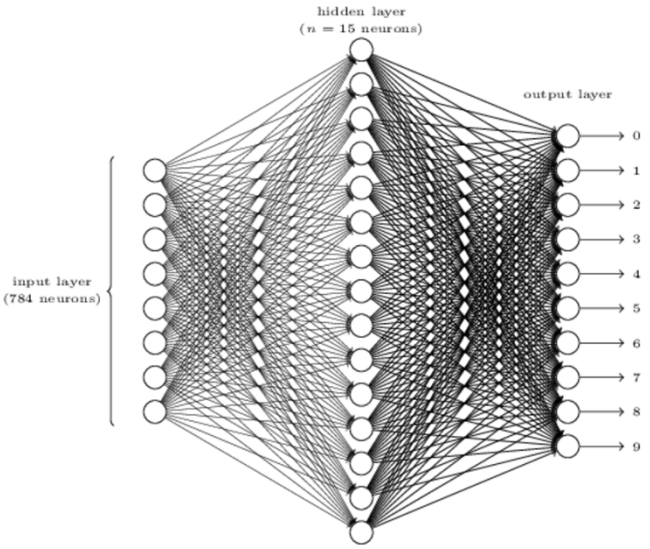
26 x 28 píxeles = 184

A. Prieto 20

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto. 23/02/2024

Perceptrón multicapa

- Algunos datos:
  - Nº de entradas: 764 (valores de píxeles)
  - Capas ocultas. Una con 5 neuronas
  - Nº de salidas: 10
  - Nº de pesos: 2.250
  - Imágenes de entrenamiento: 60.000
  - Imágenes de test; 10.000
  - (Las imágenes pueden ser de la base de datos MNIST, tomadas de 250 personas: empleados de la Oficina de Censos y de Universitarios USA)
  - Miles de millones de operaciones



Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25, pp. 15-24). San Francisco, CA, USA: Determination press. http://neuralnetworksanddeeplearning.com/chap1.html

Backpropagation, aprendizaje por descenso del gradiente (del error)

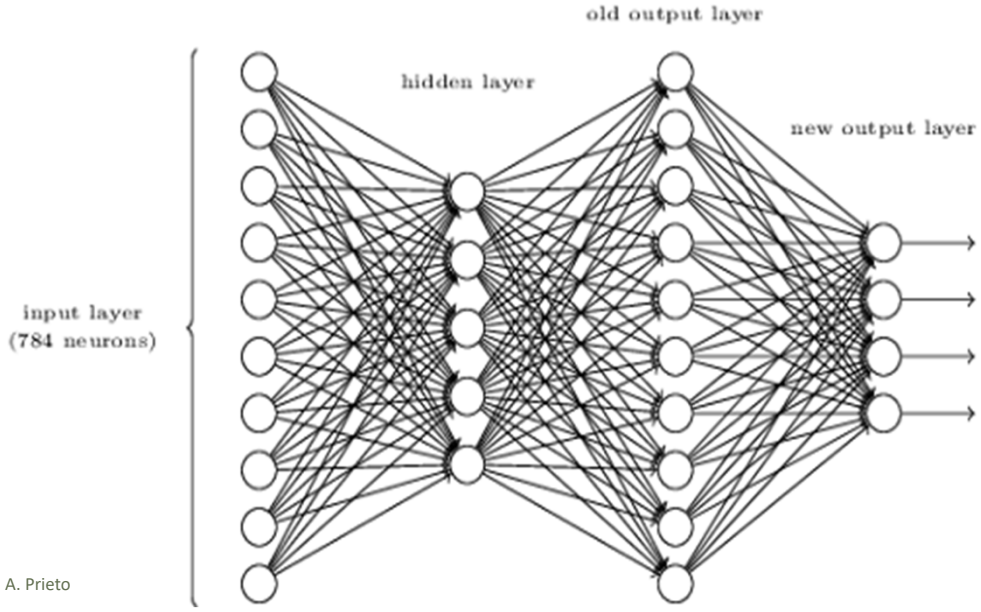
ICAR INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Las salidas son números reales de 0 a 1; si es mayor de 0,5 se considera Si: en caso contrario NO
- Se proporciona una entrada de entrenamiento (un 3, por ejemplo). Se observan si las salidas son correctas (es decir, la única salida con valor >0,5 es la 3); si no, hay error.
- Se modifican los pesos de toda la red de forma que el error disminuya
- Con el sistema backpropagation
  - Se define una función de error (C) que proporciona el error cuadrático medio entre la salida obtenida y la deseada
  - Se calcula la derivada (gradiente) de C con respecto al peso a codificar
  - Se modifica cada peso, desde la capa de salida hacia la entrada, de acuerdo con una fórmula
  - Cuando el error es menor que un valor determinado, finaliza el proceso de aprendizaje.

$$w_k \rightarrow w'_k = w_k - \frac{\eta}{m} \sum_j \frac{\partial C_{X_j}}{\partial w_k}$$

¡Miles de millones de cálculos!

Aprendizaje profundo (Deep learning): múltiples capas



A. Prieto

23

Consumo energético

24

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto. 23/02/2024

Computación verde (Green Computing) o sostenibilidad en las TIC

- Es el estudio y la práctica de diseñar, fabricar, usar y eliminar computadores, servidores y su hardware asociado para consumir energía de manera eficiente y efectiva con un impacto mínimo o nulo en la salud y el medio ambiente.

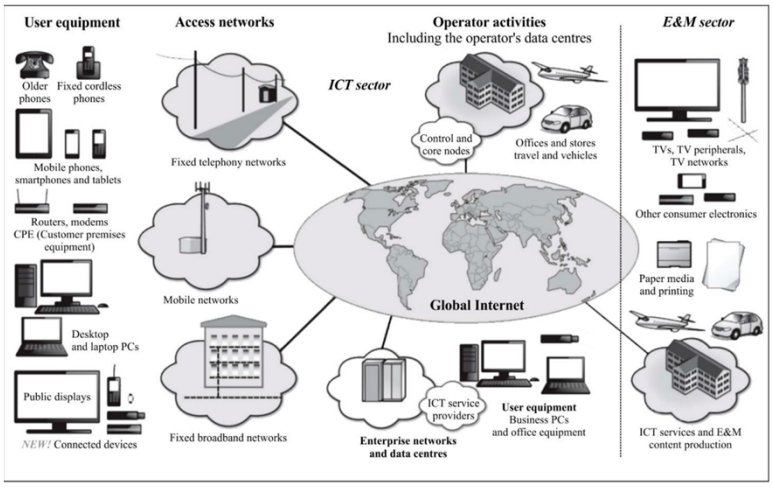
- Dhaini, M., Jaber, M., Fakhereldine, A., Hamdan, S., & Haraty, R. A. (2021). Green computing approaches-A survey. Informatica, 45(1)
- IBM (2022) Learn how green computing reduces energy consumption. https://www.ibm.com/cloud/blog/green-computing#:~:text=Green%20computing%20(also%20known%20as,consumed%20by%20manufacturers%2C%20data%20ce

- Green Computing se enmarca dentro de uno de los mayores retos de la sociedad actual, consistente en **reducir el consumo energético.**
- En general, la sociedad desconoce que las TIC, y en particular la IA, constituyen un campo relevante en el consumo de energía eléctrica, teniendo un gran impacto en las emisiones de gases de efecto invernadero.
- Todos debemos participar activamente en el reto de reducirlo.
- Además de las razones medioambientales, reducir el consumo de energía:
  - tiene fuertes implicaciones económicas y
  - mejora la autonomía de muchos dispositivos que utilizan baterías, como teléfonos inteligentes, dispositivos móviles y elementos del Internet de las Cosas

Análisis de la contribución de las TIC en el consume de energía eléctrica

- La Asociación de la Industria de Semiconductores de EE. UU. afirma:
  - Aunque la producción mundial de energía crece linealmente, la demanda de electricidad procedente de ordenadores lo hace de forma exponencial
- En el peor de los casos, las TIC podrían contribuir hasta el **23% de las emisiones globales de gases de efecto invernadero para 2030**.
- De continuar la tendencia, el consumo de energía eléctrica de la gran cantidad de equipos tecnológicos **superará la producción mundial de energía eléctrica en 2040**, lo que no sería suficiente para alimentar todos los computadores del mundo.

Sectores de las TIC involucrados en el consumo de energía



Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement. International Telecommunication Union (ITU). Recommendation ITU-T

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto.
23/02/2024

Cómo afectan las TIC al medio ambiente

- Efecto directo
- Efecto indirecto
- Efecto terciario (o de rebote)

J. Desjardins. (2018). What happens in an internet minute in 2018. Visual Capitalist. https://www. visualcapitalist. com/internetminute-2018

El efecto directo, en primer lugar, es debido a:

- La gran proliferación e incremento global del número de dispositivos electrónicos, redes de transmission y centros de datos conectados a Internet.

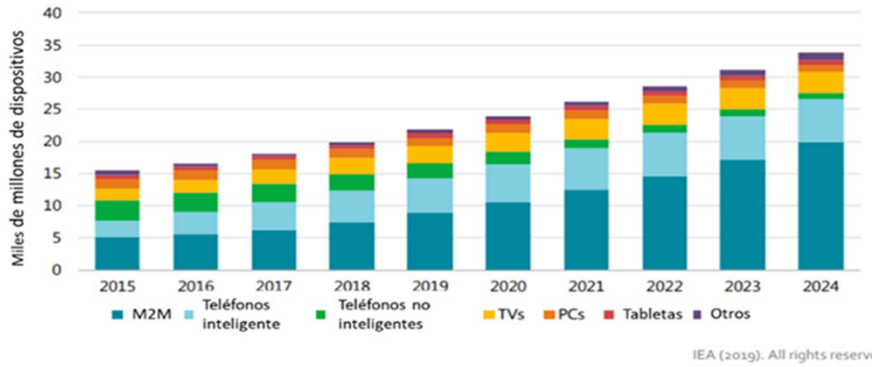


Grafico realizado por *International Energy Agency* en base al trabajo de T. Barnett y colaboradores (2019) and Cisco (2016)

El efecto directo también es debido a:

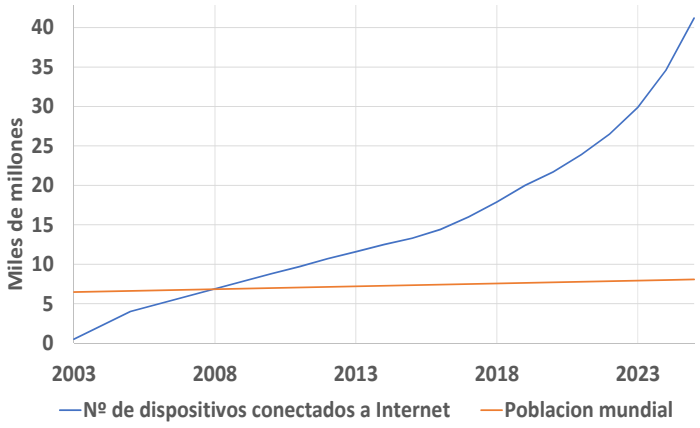
- El **incremento de aplicaciones** que constantemente usamos tanto para tareas rutinarias (telefonos inteligentes, emails, social redes sociales, ...), como para programar tareas tradicionales de computación (PCs → HPC).
- La aparición de **nuevas aplicaciones** que requieren nuevos dispositivos que, aunque individualmente consumen muy poca energía, dada su enorme cantidad su contribución global al consume es muy significativa.

J. Desjardins. (2018). What happens in an internet minute in 2018. Visual Capitalist. https://www. visualcapitalist. com/internetminute-2018.

A. Prieto 31

Ejemplo de ámbito de aplicación: Internet de las Cosas (IoT)

- Se puede considerar que IoT nació cuando el número de dispositivos conectados a Internet superó el número total de habitantes de la Tierra (finales de 2008)



A. Prieto 32

Efecto indirecto

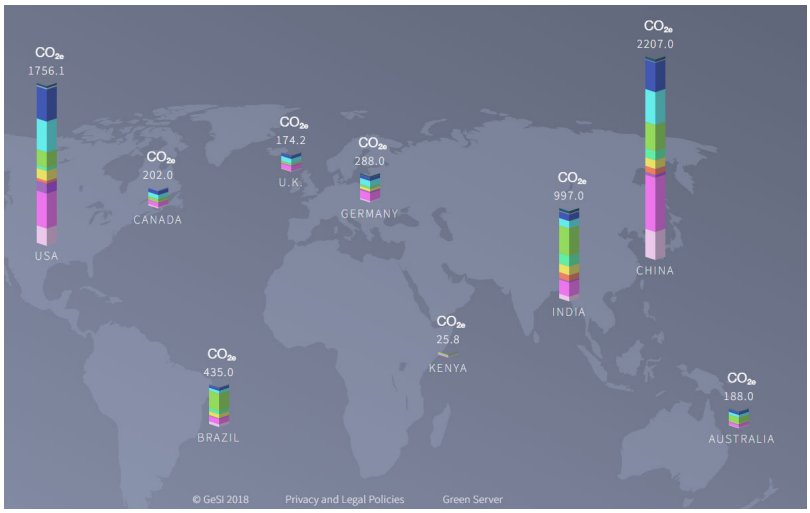
- Está provocada por aplicaciones TIC que facilitan la mejora de la eficiencia y la **reducción del consumo primario de energía** en sectores muy diversos como: construcción, industria, transporte y comercio, aportando soluciones inteligentes.
- **Es bueno para el medio ambiente**.
- En otras palabras, el aumento del consumo de TIC proviene en gran medida de su **reducción en otros sectores**, moderándose, como saldo total, el consumo global.
- **Objetivo**: identificar las diferentes aplicaciones TIC en la edificación, el transporte y la industria que redundan en una reducción del consumo energético.

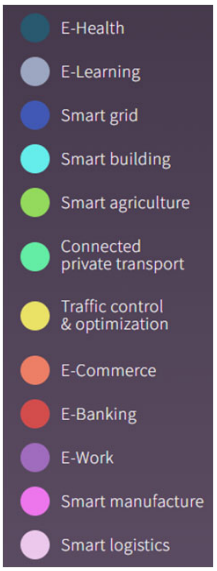
Según datos de la iniciativa GeSI los dominios TIC que están consiguiendo mayores volúmenes de reducción de CO2 son:

- E-salud
- E-enseñanza
- Redes eléctricas inteligentes
- Edificios inteligentes
- Agricultura inteligente
- Transporte privado interconectado
- Control y optimización de tráfico.
- E-comercio
- E-bancario
- E-trabajo
- Fabricación inteligente
- Logística Inteligente.

Efectos indirectos: reducción de emisiones de CO2 en millones de toneladas, gracias a las TIC

- https://smarter2030.gesi.org/explore-the-data/





Efecto terciario (o efecto rebote)

- Es un fenómeno que se produce a medida que los servicios TIC son más útiles, más baratos y eficientes energéticamente, ya que esto aumenta nuestro estilo de vida digital, lo que produce un efecto rebote: **los equipos** TIC consumen menos, pero **se utilizan mucho más**.
- A nivel mundial tiene una consecuencia **negativa.**
- Las estimaciones muestran que los posibles efectos rebote por la digitalización varían entre un **10% y un 30% de mayor consumo eléctrico**, dato que varía según el sector, la tecnología y el uso final.

GeSI. Global e-Sustainability Initiative. Accenture strategy SMARTer2030-ICT solutions. (2015) https://smarter2030.gesi.org/downloads/Full_report.pdf

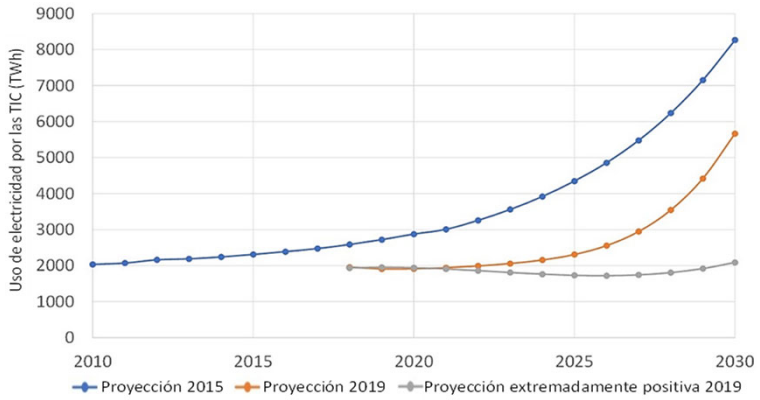
Estimaciones sobre la evolución del consumo

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

37

Proyecciones de Andrae and Edler sobre el uso de energía eléctrica por las TIC en TW·h por año (2015 y 2019)

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA



$$E = P \cdot t$$

1 W·h = 3,600 Joules

- A.S. Andrae y T. Edler. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6(1), 117-157. https://doi.org/10.3390/challe6010117
- A.S. Andrae (2019). Comparison of several simplistic high-level approaches for estimating the global energy and electricity use of ICT networks and data centers. International Journal, 5, 51. DOI: 10.30634/2414-2077.2019.05.06

A. Prieto

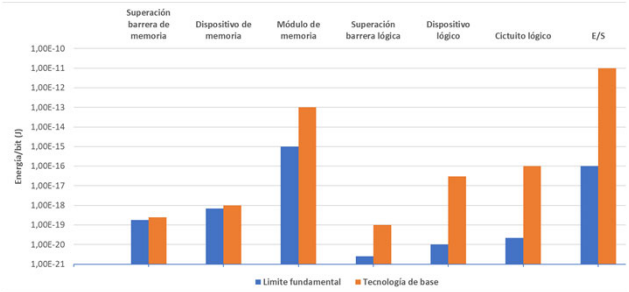
38

10^n	Prefijo	Símbolo	Equivalencia decimal
10^{18}	exa	E	1 000 000 000 000 000 000
10^{15}	peta	P	1 000 000 000 000 000
10^{12}	tera	T	1 000 000 000 000
10^{9}	giga	G	1 000 000 000
10^{6}	mega	M	1 000 000
10^{3}	kilo	k	1 000
10^{2}	hecto	h	100
10^{1}	deca	da	10
10^{0}	-	-	1

A. Prieto 39

V. Zhirnov et al. publican (2014) un artículo donde:

- Con datos reales obtienen la evolución del consumo por bit de diferentes elementos binarios (dispositivos lógicos y elementos de memoria) considerando el consumo de transistores y microprocesadores individuales y la dinámica de los procesos físicos en los diferentes componentes (capacitivos, resistivos, etc.).
- Llegan a la conclusión de que, en situaciones típicas;
  - La energía mínima requerida por transición (conmutación) de un bit es de alrededor de ≈ 10^{-14} J/bit.
  - Valor estimado como un objetivo alcanzable ≈ 10^{-17} J/bit.

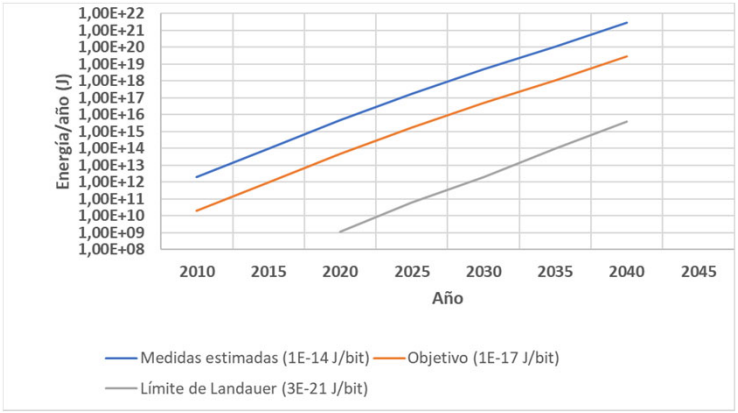


- V. Zhirnov, R. Cavin y L. Gammaitoni. (2014). Minimum energy of computing, fundamental considerations. In ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology. IntechOpen
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015). Append A4.

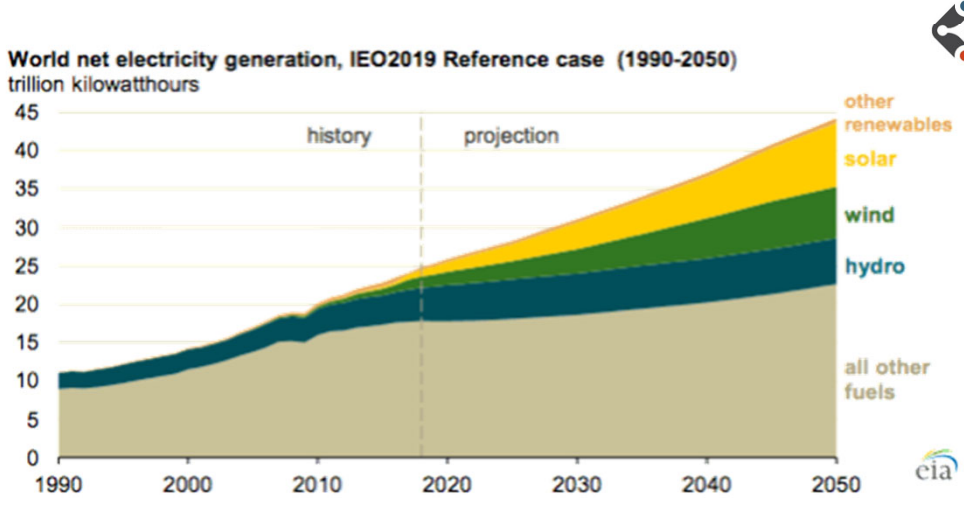
A. Prieto 40

Energía eléctrica global consumida por la informática en un año

- Valor calculado a partir de datos reales (2014): ≈10^{-14} J/bit
- Valor estimado por Zhirnov, como un objetivo a lograr ≈10^{-17} J/bit
- Límite de Landauer ≈$3 \cdot 10^{-21}$ J/bit



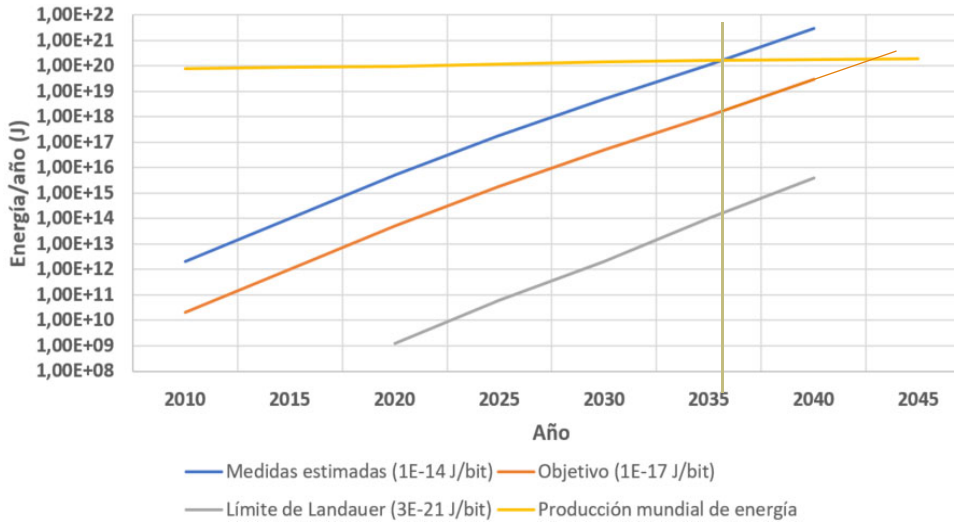
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015)



Source: U.S. Energy Information Administration, International Energy Outlook 2019

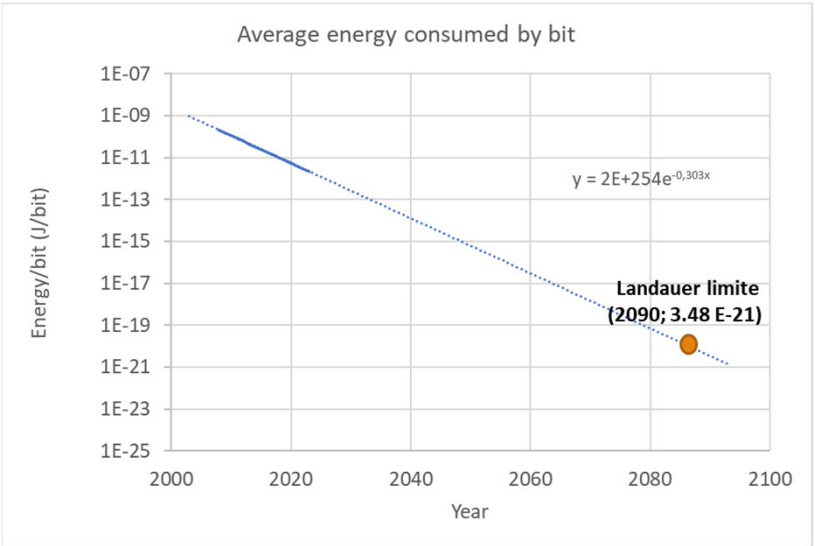
https://www.powermag.com/eia-renewables-will-account-for-half-of-global-power-generation-by-2050/

En el peor de los casos, en 2035 no habrá energía eléctrica suficiente …



A. Prieto 43

Con nuestros resultados el límite de Landauer se alcanzaría en 2090

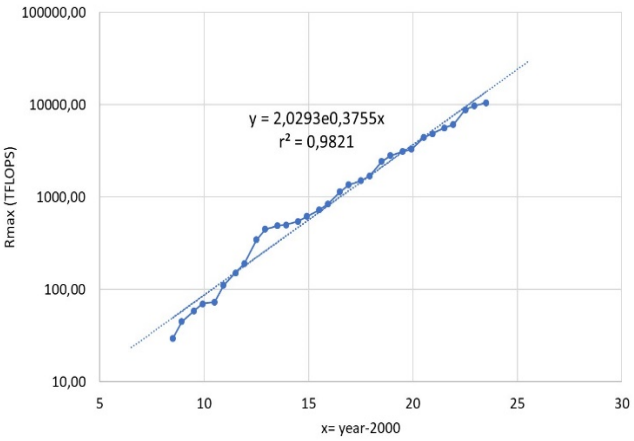


A. Prieto 44

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto. 23/02/2024

Hemos obtenido también la evolución el rendimiento de computación (últimos 15 años)

- Dobla cada 1,85 años
- PROBLEMA:
  - **La eficiencia energética crece más lentamente (2.29 años)**
  - Aunque las diferencias parecen pequeñas, doblar la eficiencia energética cada **1.85 años** significa incrementarla aproximadamente 43 veces en una década mientras que doblar cada **2.29 años** supone incrementar por 21 por década.



FLOP: operaciones con nº reales.
FLOPS: operaciones con nº reales por segundo

A. Prieto

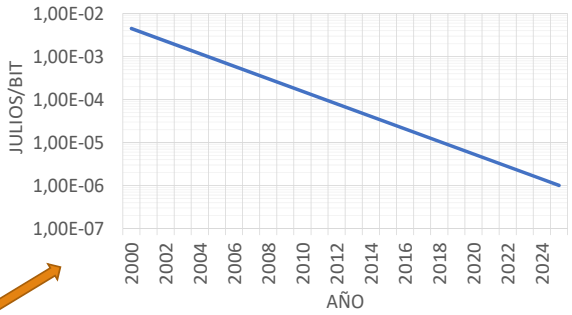
Consumo en tráfico de información digital

46

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto.
23/02/2024

Estimación del consumo de energía en el tráfico por Internet

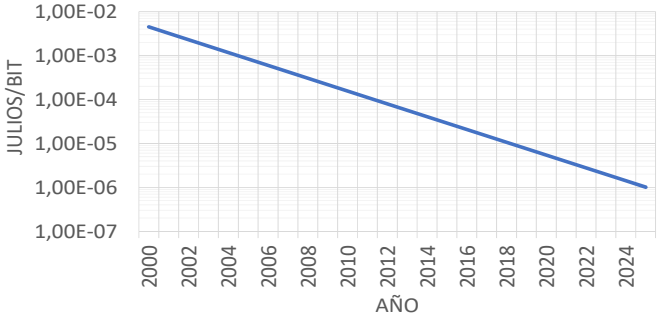
- Es difícil ya que este depende de muy diversos factores:
  - El canal de transmisión (atmosfera, cable, fibra óptica, etc.)
  - La distancia entre emisor y receptor.
  - El caudal de datos (velocidad de transmisión).
  - En el caso de un mensaje transmitido, depende además de codificación utilizada, tipo de modulación, etc.
- Estimaciones de Aslan y colbs.:
  - El consumo por bit se reduce a la mitad aproximadamente cada 2 años (en procesamiento cada 2,6 años).



- J. Aslan, K. Mayers, J. G. Koomey y C. France. (2018). Electricity intensity of internet data transmission: Untangling the estimates. Journal of industrial ecology, 22(4), 785-798.

Estimación del consumo de energía en el tráfico de 1 bit por Internet

- La energía media de transmisión de 1 bit en 2020 a través de Internet ha sido del orden de **$2{,}77 \cdot 10^{-6}$ J/bit**, lo que supone:
  - ≈ $2{,}8 \cdot 10^{11}$ veces mayor que la del procesamiento de un bit (10^{-17} J/b), si se tienen en cuenta las estimaciones de Zhirnov o de
  - 6 ordenes de magnitud mayor si se tiene en cuenta la eficiencia energética de los supercomputadores obtenida por nosotros ($5{,}5 \cdot 10^{-12}$ J/bit)



Algunos datos de situaciones prácticas

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

49

Ejemplo y ejercicio interesante: colas de emails UGR

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Texto oficial establecido por la Secretaría General de a UGR:
  - *Este mensaje ha sido generado desde una cuenta de la Universidad de Granada para los fines propios de la institución. Su contenido se considera información confidencial, por lo que queda informado de que su utilización, divulgación o copia sin autorización no está permitida. Si usted ha recibido indebidamente el correo le rogamos que advierta de ello por esta misma vía al remitente y proceda a su eliminación. Cualquier incidencia relacionada con la recepción de nuestros correos electrónicos y en particular las relativas a la seguridad y confidencialidad pueden ser comunicadas a protecciondedatos@ugr.es. Para más información al respecto, puede consultar nuestra política de privacidad .*
  - *This message has been generated from an e-mail address of the University of Granada for the institution's own purposes. Its content is considered confidential information, so it is informed that its unauthorized use, disclosure or copying is not permitted. If you have improperly received the email please warn the sender of this, same way and proceed to its removal. Any incident related to the receipt of our emails and in particular those related to security and confidentiality may be communicated to protecciondedatos@ugr.es. For more information, please refer to our privacy policy.*
- 1.285 caracteres (sin contabilizar los enlaces)

A. Prieto

50

Cálculos sobre emails de la UGR

- UGR emails (2021-2022):
- 6,819,020 mensajes enviados a Internet desde estafetas centrales o departamentales.
- 4,671,665 mensajes provenientes de Internet y entregados en buzones UGR de PAS/PDI.
- 22,278,138 mensajes provenientes de Internet y entregados en buzones UGR de Alumno.

		Mensajes enviados	Mensajes recibidos
Nº caracteres/email		1.285,00	1.285,00
Nº bits (UTF8)/carácter		16	16
Mensajes anuales		6.819.020	56.949.803
Nº de bits anuales		1,40199E+11	1,17089E+12
Energia bit transmitido (2022)	Julios/bit	3,00E-06	3,00E-06
Energia total	Julios	4,21E+05	3,51E+06
Energía total	KWh	1,17E+02	9,76E+02
MIX eléctrico	Kg CO2/KWh	0,351	0,351
Huella de carbono cola emails	KgCO2	4,10E+01	3,42E+02

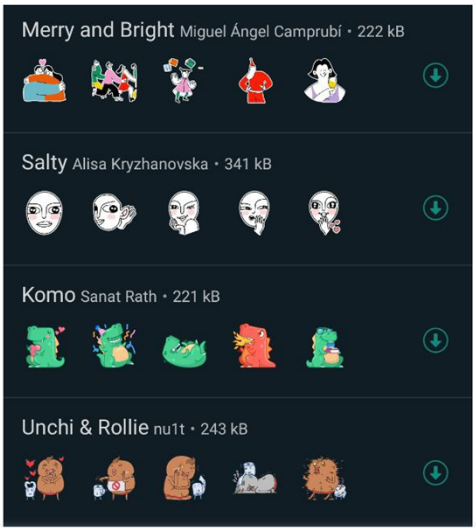
¡Casi media Tonelada de CO2 al año!

A. Prieto 51

Emoticones, emojis y pegatinas (*stickers*) en WhatsApp

- Emoji: 4 bytes
- Pegatina ≈ 50 KB, <100 KB → 0,120 g de CO2
- 20.000 millones de WhatsApp diarios.
- Si 1 de cada 10 es una pegatina → **87.500 T de CO2 al año**.

Para que os hagáis una idea, **en el caso del *pack* "Chummy Chum Chums", oficial de WhatsApp, sus 16 *stickers* solo pesan 3,7 megas**, por lo que están muy lejos de ese máximo que tenemos disponible (1MB/pegatina).



A. Prieto 52

24 OPINIÓN

Jueves 17.12.20
IDEAL

La contaminación silenciosa

ALBERTO PRIETO ESPINOSA
Academia de Ciencias de Granada

Todo proceso de transferencia o movimiento de datos, entre móviles, computadores, etc, consume energía, y muchos dispositivos, entre los que se encuentran los supercomputadores, están funcionando las 24 horas del día, siendo los consumos muy altos

Uno de los retos más importantes de la sociedad actual es reducir el consumo de energía con el objeto de mantener o hacer posible la sostenibilidad de nuestro planeta. Por hacer referencia a nuestro contexto, la Unión Europea tiene como una

y de los programas que es mucho mayor. Así, cuando enviamos un correo electrónico, además del consumo inherente a la transmisión de los bits hay que añadir el del programa que me permite editar, enviar, recibir y visualizar los emails.

Hasta ahora los parámetros que se utilizaban para medir las prestaciones de un

A. Prieto. La contaminación silenciosa. Ideal, 17/12/2020, p. 24

Aplicaciones genuinas de IA

¿Habrá suficiente energía para el desarrollo global de la IA?. A. Prieto.
23/02/2024

Algunos dominios donde se utiliza aprendizaje profundo

- (En los que se puede conocer facilmente si la respuesta es correcta o no)
- Clasificación de imágenes, ImageNet
- Detección de objetos, MS COCO
- Question answering (SQuAD 1.1) (comprensión lectora sobre articulos Wikipedia)
- Named Entity Recognition CoNLL 2003 (nombres de personas, entidades, lugares, nº de telefonos, paises, ciudades, etc.)
- Reconocimiento del habla (ASR SWB Hub500),
- Detección de caras (WIDER Face Hard),
- Generación de imágenes (CIFAR10), and
- Estimación de postura humana (MPII Human Pose).

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

A. Prieto 55

Consumo computacional y energético de aprendizaje profundo en algunas aplicaciones de referencia (proyección exponencial).

Aplicación de referencia	Tasa de error		Nº de operaciones (nº reales, flops)	Emisiones CO_2 (Tm)	Consumo anual de nº de hogares en España	Costo (€)
Clasificación de imágenes, ImageNet	2022	9.00%	1,00E+25	4,54E+03	363	9,30E+07
	Objetivo 1	5%	1,00E+31	4,54E+10	3.632.000.000	9,30E+14
	Objetivo 2	1%	1,00E+93	4,54E+71	3,63E+70	9,30E+75
Detección de objetos, MS COCO	2022	38.7%	1,00E+23	4,54E+01	4	9,30E+05
	Objetivo 1	30%	1,00E+25	4,54E+04	3.632	9,30E+08
	Objetivo 2	10%	1,00E+50	4,54E+28	3,63E+27	9,30E+32
Comprensión lectora (SQuAD 1.1)	2022	9.4%	1,00E+23	4,54E+02	36	9,30E+05
	Objetivo 1	2%	1,00E+52	4,54E+31	3,63E+30	9,30E+34
	Objetivo 2	1%	1,00E+89	4,54E+67	3,63E+66	9,30E+71
Reconocimiento de nombres de entidades (CoNLL 2003)	2022	5.4%	1,00E+25	4,54E+03	363	9,30E+07
	Objetivo 1	2%	1,00E+62	4,54E+40	3,63E+39	9,30E+44
	Objetivo 2	1%	1,00E+121	4,54E+99	3,63E+98	9,30E+103

A. Prieto 56

Aprendizaje en Chat GPT-3 175B

- Contenido:
  - **175 mil millones de parámetros** (pesos y otros valores de las redes neuronales).
  - 800 GB de memoria
  - Procesamiento de consultas de hasta 2048 tokens (palabras o subpalabras), se consideran como "ventana contextual". Eso significa que tiene 2.048 pistas a lo largo de las cuales se procesan los tokens sucesivos.
- Aprendizaje
  - Última actualización enero de 2022.
  - **Pre-entrenamiento no supervisado de red neural profunda (96 capas) con un corpus de 570 mil millones de tokens recopilados de Internet en sitios como Wikipedia (3 mil millones), Common Crawl (410 mil millones),** …
- Generación de texto
  - 96 capas de ANN transformers decoders (palabra + posición en la frase, etc.) que generan textos que simulan la redacción humana (concepto de "atención", semántica de las palabras, contexto, etc.)

A. Prieto 57

Pre-entrenamiento de GPT-3 750B

- Tiempo total de computación para el pre-entrenamiento:
  - ≈ 3.000 PFLOPS·días

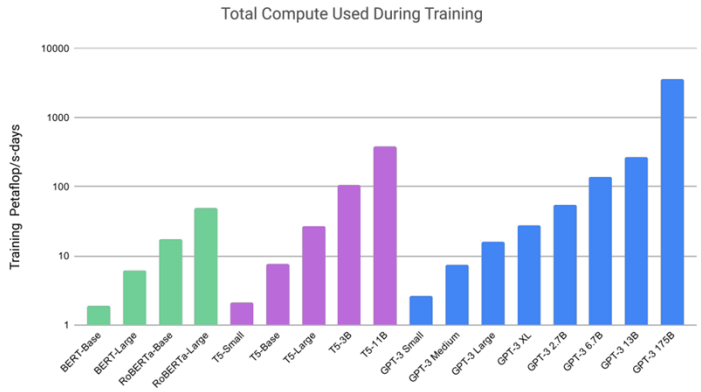


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. Cited by 12,240

A. Prieto 58

Supongamos que el entrenamiento lo efectuamos con los #1 del TOP500 y Green500 (Junio 2023) …

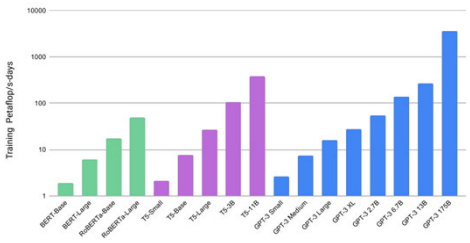
Green500 Data

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
6	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	22,703	52.592

A. Prieto 59

El preentrenamiento a gran escala requiere grandes cantidades de cálculo, lo que consume mucha energía (ChatGPT-3 175B)

	1º TOP500 (Junio 2023)	1º Green500 (Junio 2023)
	Frontier	Henri
GPT-3 entrenamiento. PetaFLOPS·dias	3.000,00	3.000,00
Gigaflops-dia	3,00E+09	3,00E+09
Eficiencia energética (GFlops/watts)	52,592	65,396
W·dia	5,70E+07	4,59E+07
KW·h	1,37E+06	1,10E+06
MIX (Kg CO2/KWh)	0,351	0,351
Huella de carbono (Kg CO2)	480.529,36	386.445,65
Prestaciones, Rmax (PFlops)	1194,00	2,88
Tiempo ejecución (s)	2,17E+05	9,00E+07
Tiempo de ejecución (dd:hh:mm:ss)	02:12:18:05	06:16:00:00

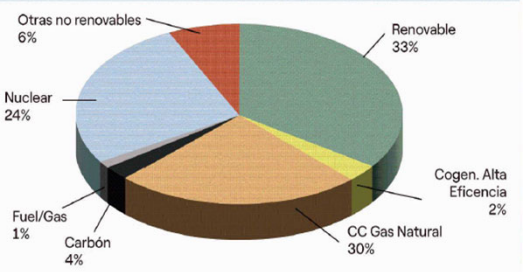


480.5 toneladas de CO2 (2 dd 12 hh) vs 386.4 ton (6 dd 16 hh)
¡38 hogares españoles durante 1 año!

A. Prieto 60

ORIGEN DE LA ELECTRICIDAD DE SU COMERCIALIZADORA. 2022.

Endesa Energía S.A.U



Origen	Endesa Energía S.A.U	Mix generación nacional
Renovable	33,0%	42,3%
Cogen. Alta Eficencia	1,5%	2,1%
CC Gas Natural	30,3%	25,7%
Carbón	3,4%	2,9%
Fuel/Gas	1,3%	1,1%
Nuclear	24,4%	20,7%
Otras no renovables	6,1%	5,2%

Emisiones de CO_2 equivalente

Endesa Energía S.A.U

Emisiones CO_2 eq. (g/kWh)	**188**
Media nacional (g/kWh)	**162**

A. Prieto 61

- El consumo de pre-entrenamiento se amortiza según se va usando.
- Una vez entrenado, en su uso normal (inferencia), generar 100 páginas de contenido a partir de un modelo entrenado puede consumir del orden de 0,4 kW-h.

0.4 KW·h x 0.351 = 0.14 Kg CO2 a la hora

- Pero llega a ser más importante que el de entrenamiento por su factor multiplicativo: ¡millones de usuarios!.
  - 2,8 millones de accesos (inferencias) consumen lo de un entrenamiento.

A. Prieto 62

Procedimientos y técnicas para incrementar la eficiencia energética en las TIC

ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

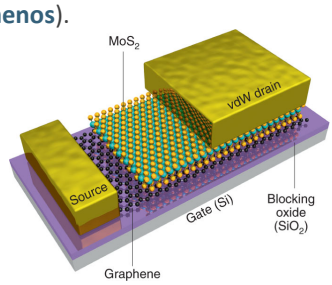
63

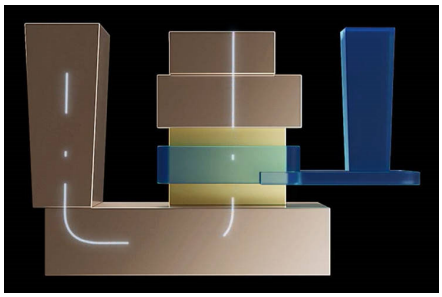
ICAR
INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Afortunadamente las previsiones más pesimistas (mediados de la década de los 2010) sobre el consumo energético asociado a las TIC no se están cumpliendo, siendo menor que el esperado.
  - La industria alrededor de las TIC es consciente del problema y se están invirtiendo recursos financieros y políticas activas para reducir el incremento del consumo, tanto en la fabricación de nuevos productos como en el consumo inherente a su uso.
- Acciones para reducir el consumo:
  - **A. Mejoras tecnológicas en los componentes electrónicos y dispositivos**
  - **B. Gestión y planificación del uso de los recursos.**
  - **C. Cambios de escala**

A. Mejoras tecnológicas en los componentes electrónicos y dispositivos …

- Cambios en los dispositivos y en la arquitectura interior de los microchips.
  - Prototipo de IBM de CI que hace posible apilar verticalmente los transistores-
  - Incremento de la densidad de integración y reducción del consumo de energía (se estima que **hasta un 85% menos**).



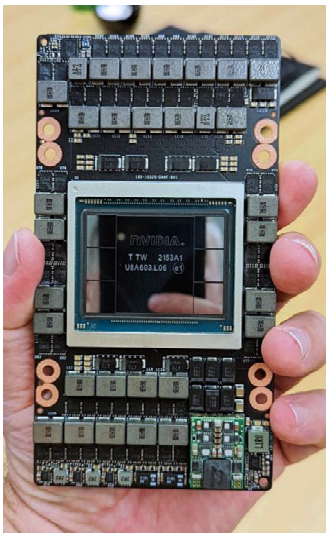


IBM's VTFET with a vertical channel (yellow) and gate-all-around (blue). Contacts are brown and the white line shows current flow.

- Zhang, J., Gao, F., & Hu, P. (2021). A vertical transistor with a sub-1-nm channel. Nature Electronics, 4(5), 325-325.
- Steve Bush, (14 diciembre 2021) IBM beats finFETs with vertical CMOS at IEDM. Electronics Weekly.com

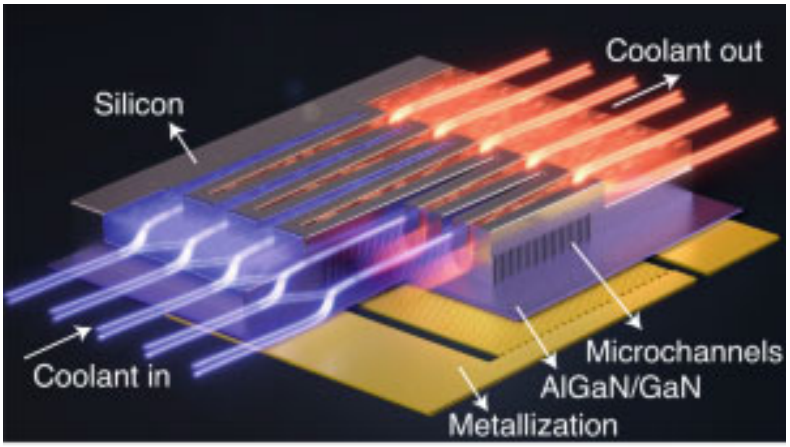
… mejoras tecnológicas en los componentes electrónicos y dispositivos …

- Inclusión de funciones de gestión de energía dentro de las CPUs con los que, dependiendo de la carga de trabajo, cambian dinámicamente entre diferentes estados de energía (modo de espera, por ejemplo).
- Desarrollo de procesadores de uso específico para ámbitos o funciones concretas, como GPU y TPU.
- Fuentes de alimentación AC/DC conmutadas; introducción de nuevos materiales, como el nitruro de galio y el carburo de silicio, que permiten diseños a más altas frecuencias.



… mejoras tecnológicas en los componentes electrónicos y dispositivos …

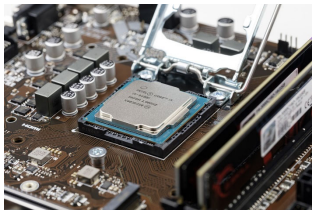
- Integración directa en los chips de sistemas de refrigeración con microfluidos, que sustituyan a los ventiladores externos.



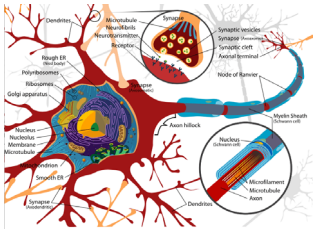
Varnava, C. Chips cool off with integrated microfluidics. Nat Electron 3, 583 (2020). https://doi.org/10.1038/s41928-020-00494-5

… mejoras tecnológicas en componentes y dispositivos electrónicos

- Cambios en la tecnología de otros dispositivos. HDD → SSD, reducción del consumo de energía superior al 50%
- Computación neuromórfica
  - El cerebro humano es uno de los sistemas más eficientes energéticamente ya que consume una potencia de unos 25 vatios (menos de la mitad de un PC portátil) y dispone de 86.000 millones de elementos informáticos (neuronas).
    - Opino que el bajísimo consumo energético se debe más que a la arquitectura a la tecnología subyacente (material) → en lugar de silicio, arseniuro de galio, etc. el del cerebro es de naturaleza biológica (bioquímica, células, tejidos,...)



pixabay



Wikimedia Commons

¿Por qué no podemos hacer en la actualidad computadores que procesen datos como lo hace el cerebro?

ICAR INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

Mammalian Brains vs Computers

Parallel distributed architecture		Serial architecture
Low power (25W), small footprint (1 liter)		High power (100MW), Large footprint (40M liters)
Asynchronous (no global clock)		Synchronous (global clock)
Analog computing, Digital communication		Digital computing and communication
Integrated memory and Computation		Memory and Computation are clearly separated
Intelligence via Learning thru BBE interactions		Intelligence via programmed algorithms/rules
Composed of noisy components and operates at low speeds (< 10 Hz)		Precision in components and operates at very high speeds (GHz)
Spontaneously active		No activity unless instructed

Cerebro completo $86 \cdot 10^9$ neuronas ; 1 chip ELSI → 10^8 transistores ;

- Avram Bar-Cohen (DARPA), "Cognitive computing, Towards the electronic brain," presentation at the Workshopon Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015
- Semiconductor Industry Association and the Semiconductor Research Corporation, Rebooting the IT Revolution: A Call 547 to Action. (2015)

A. Prieto 69

B. Gestión y planificación del uso de los recursos:

ICAR INGENIERÍA DE COMPUTADORES, AUTOMÁTICA Y ROBÓTICA

- Utilizar los distintos sistemas tratando de reducir el consumo energético global, como:
  - Hacer entrar en los **modos de suspensión o de espera** a los recursos (servidores, sistemas de memoria etc.) que en un momento dado no sean necesarios.
  - **Escalado dinámico de la tensión y de la frecuencia** (*Dynamic Voltage and Frequency Scaling,* DVFS).
    - Ejecutar lentamente los programas que no necesiten un tiempo de respuesta muy corto.
    - Se estima que, **si la frecuencia de reloj se reduce a la mitad**, el tiempo de ejecución se duplica, pero **el consumo energético se reduce a una cuarta parte**.
  - Ejecutar, en lo posible, las aplicaciones dentro de "**horas valle**" donde la producción de energía eléctrica procedente de fuentes limpias es mayor por ser más elevada la producción eólica debido al viento u horas donde la radiación solar es mayor. Doble beneficio:
    - se reduce el coste económico de la energía necesaria para la ejecución de los programas
    - se favorece el uso de las energías alternativas.

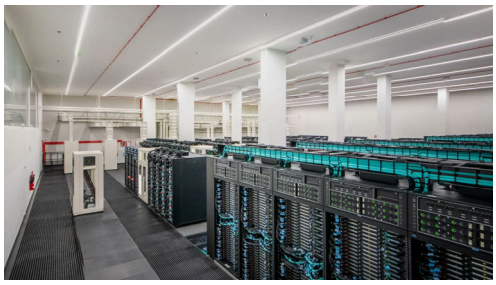
A. Prieto 70

C. Cambios de escala

- La proliferación de teléfonos inteligentes y **pequeños dispositivos móviles** da lugar a una reducción del consumo energético ya que cada uno de ellos ofrece multitud de funciones y servicios que antes realizaban dispositivos de consumo independientes.
- Planificación y asignación de tareas a los recursos hardware disponibles teniendo en cuenta su **eficiencia energética**. En particular, debe explotarse el paralelismo de las aplicaciones y de los programas buscando la mejor eficiencia posible.
  - En muchos casos, una asignación eficiente de recursos requiere el rediseño de las aplicaciones y de los algoritmos.
- **Endoso computacional** (*offloading*): los procesos que requieren tareas informáticas intensivas se transfieren (endosan) a una plataforma externa, que puede ser desde un acelerador de hardware hasta un sistema de clúster, o recursos en la nube. **Tecnologías de virtualización**.
  - Sólo es beneficioso cuando se requiere gran volumen de computación con relativamente poca cantidad de comunicación.
- Fusión o transformación de centros de datos medianos a en **centros de datos hiperescala** (mucho mayores) (Google Cloud, Amazon Web Services, Microsoft Azure, OVHCloud, o Rackspace Open Cloud), donde el consumo de energía se gestiona mucho mejor.

Centros de datos: Nacional de Supercomputación: Mare Nostrum

- "El gran reto es hacer el primer gemelo digital del cuerpo humano",
- Consumo eléctrico
  - MareNostrum4 → 2 MW
  - MareNostrum5 (202,8 millones €) en fase de inicio de producción) → 8 MW
- Varios millones de € al año.
- Top500 de junio 2023:
  - MareNostrum 5 (máquina del 2017). Eficiencia energética:
    - partición con procesadores de uso general, 3,96 Gflops/W.



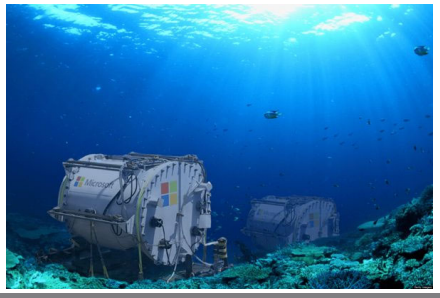
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
124	98	**MareNostrum** - Lenovo SD530, Xeon Platinum 8160 24C 2.1GHz, Intel Omni-Path, Lenovo Barcelona Supercomputing Center Spain	153,216	6.47	1,632	3.965

Planta de enfriamiento de Google en Hamina (Finlandia)

- Climatización muy critica y su funcionamiento requiere un gran consumo energético.
- Al ser un país nórdico el coste de la climatización es mucho más bajo que en países más cálidos.
- Este centro utiliza el agua del gélido mar del golfo de Finlandia para refrigerar todas sus instalaciones.

Proyecto Natick de Microsoft

- Se sumergieron durante 2 años (2018 a 2020) 864 servidores en un contenedor similar a un submarino.
- Ubicación en las Islas Orcadas, en el norte de Escocia: aguas gélidas y la red eléctrica se abastece al 100% de energía eólica, solar y marina, etc. obtenida en las cercanías. No contaba con refrigeración activa.



- https://news.microsoft.com/es-es/2020/09/15/proyecto-natick-el-futuro-de-los-centros-de-datos-bajo-el-mar-es-fiable-practico-y-sostenible/

Proyecto Natick de Microsoft

- Los servidores experimentaron una **tasa de fallos ocho veces inferior** a lo esperado en un Centro de Datos convencional, gracias, entre otras cosas, a la atmósfera de nitrógeno empleada en la cápsula sellada.
- Se rescató del fondo marino, cubierto de algas, percebes y anémonas.
- Se concluyó que **el futuro de los centros de datos bajo el mar es fiable, práctico y sostenible.**

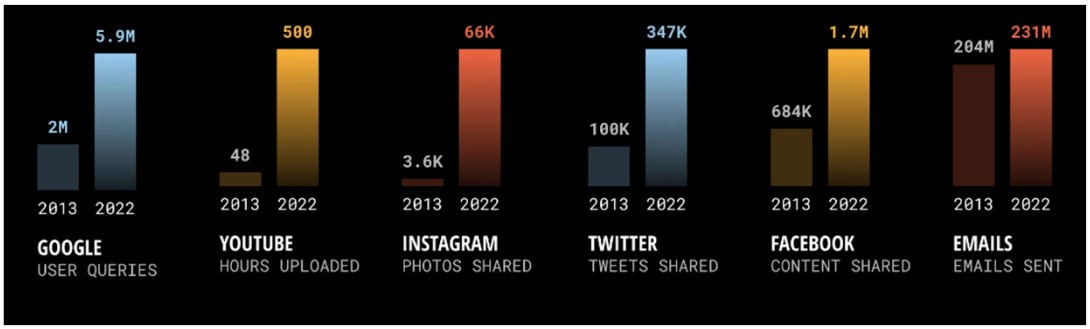
A. Prieto 75

Conclusiones

76

Conclusiones …

- Los datos nunca duermen, de forma que **cada minuto**:



Informe Data Never Sleeps 10.0, realizado por Domo. https://www.domo.com/data-never-sleeps

Conclusiones ….

- El aumento de la potencia de cálculo (FLOPS) de los computadores ha hecho posible el desarrollo de numerosas aplicaciones de IA; pero la aceleración de los requisitos es permanente mientras que el rendimiento del hardware no aumenta al mismo ritmo.
  - Se estima que los modelos de IA duplican la potencia computacional que utilizan cada 3,4 meses (0,28 años) desde 2012; mientras que
    - La potencia de cálculo, dobla cada 1,85 años
    - La eficiencia energética crece muchos más lentamente (dobla cada 2.29 años)
- Es de suma importancia forzar al aprendizaje automático hacia técnicas que requieran menos coste computacional y más eficientes desde el punto de vista energético.
  - El desarrollo de algunos modelos esta asociado al crecimiento exponencial del nº de parámetros.
- Hay establecidos hitos que hoy día son inalcanzables debido a las extraordinarias necesidades de computo y de energía requeridas.

… conclusiones …

- Se estima que en el año 2030 las tecnologías de la información consumirán aproximadamente el 13% de la electricidad mundial, y para 2050 el de los centros de datos será unas tres veces mayor que la cantidad total de energía generada en Japón.
- El **consumo de energía** debe considerarse como una medida de prestaciones tan importante como el **rendimiento computacional.**
- Por un lado, hay razones medioambientales y económicas, pero también la necesidad de mejorar la autonomía de los dispositivos que utilizan baterías.
- La reducción del consumo energético en el ámbito de las TIC es una cuestión trascendental, y debe ser afrontada desde muy distintos ámbitos (computer engineering, software engineering, dissemination-users, teaching, etc.)
- La sociedad debe estar informada de que el uso de las TIC (sea cual sea su forma), lleva implícito un consumo energético. **¡Todos debemos contribuir, desde nuestros respectivos ámbitos, al reto de lograr la sostenibilidad de nuestro planeta!**

A. Prieto 79

Agradecimientos

- Deseo agradecer a las siguientes personas su colaboración en las investigaciones que estamos realizando sobre este tema:

  - Beatriz Prieto Campos
  - Juan José Escobar
  - Miguel Damas
  - Antonio Díaz
  - Christian Morillas
  - Jesús González Peñalver
  - Andrés Ortiz (UMA)
  - Francisco Gil (UAL)
  - Francisco Illeras

Nuestras investigaciones en este ámbito actualmente se están financiando parcialmente por:

Proyecto: **PID2022-137461NB-C31** financiado por MCIN/AEI/10.13039/501100011033/ y por FEDER Una manera de hacer Europa

MINISTERIO DE CIENCIA E INNOVACIÓN
UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL
AGENCIA ESTATAL DE INVESTIGACIÓN

A. Prieto 80

¡Muchas gracias por tu atención!

- Alberto Prieto Espinosa.
- Conferencias:

  https://icar.ugr.es/informacion/directorio-personal/alberto-prieto-espinosa/web/conferencias