Artículo

Revista Ciencia y Naturaleza

La ciencia de los datos y su reproducibilidad

David A. Paz García

1019

Artículo

La ciencia de los datos y su reproducibilidad

Cómo citar este artículo: Paz-García DA. 2023. La ciencia de los datos y su reproducibilidad. Revista Ciencia y Naturaleza 01 (1019): 19-30. https://doi.org/10.5281/zenodo.14478040


Folder Trinity corrida (Fc)
Folder reads muestars para hacer mapping (Freads, ver arriba)
SAM1=${Freads}Unmatched_reads_NWAfree.fq.1.gz
SAM2=${Freads}Unmatched_reads_NWAfree.fq.2.gz
To=${Fc}Trinity.fasta
Tbow=${Fbow}Trinity.fasta
cp ${Trin} ${Tbow}
bowtie2-build ${Tbow} ${Tbow}

La ciencia actual de los grandes datos (“Big Data”)

Muchas de las investigaciones que realizan actualmente los científicos incluyen una enorme cantidad de datos. Estos pueden tratarse de millones de secuencias de ADN para reconstruir el genoma de una especie extinta como el mamut, el análisis de la regulación y la expresión de miles de genes de especies ante la respuesta a fluctuaciones climáticas, o el análisis de las variables ambientales para predecir las condiciones climáticas futuras, entre otros.

La capacidad de la humanidad de reunir y analizar grandes cantidades de información es una de las mayores revoluciones de nuestro tiempo.

La obtención de información también conlleva el desarrollo y la utilización de programas y equipo especializado de gran capacidad de memoria para probar hipótesis, crear simulaciones y modelos predictivos. En ningún momento de la historia del ser humano, se había tenido la capacidad de adquirir tal cantidad de información. Lo anterior representa una oportunidad única en nuestro tiempo para poder realizar investigación integral en diversos niveles y desde diferentes áreas de conocimiento. Así se tiene la necesidad de desarrollar diversas técnicas y automatizaciones para realizar un estudio científico más sólido.


Vertigo3d/Getty Images Signature

La ciencia de los datos es un campo relativamente nuevo, también conocido como "*Big Data*", y se refiere al conjunto de información tan grande y compleja que es difícil de analizar utilizando técnicas tradicionales de procesamiento. Este campo interdisciplinario combina la estadística, la informática y el aprendizaje automático.

Un tipo de inteligencia artificial conocido como **aprendizaje automático**, permite que las computadoras aprendan de la información sin ser programadas explícitamente para ello.


deemakdaksina

La **estadística** es la ciencia de recopilar, analizar y obtener conclusiones de los datos, a través de técnicas matemáticas para comprender patrones.

Por su parte, la **informática** es el área de la ciencia que utiliza diferentes métodos para almacenar y procesar datos digitales que deben recopilarse, administrarse y analizarse de una manera que los convierta en información útil, lo que contempla grandes retos.


Giuseppe Ramos J

El avance de la tecnología ha cambiado las investigaciones, ahora se puede analizar cantidades inmensas de información que antes era impensable.

El primer paso en el *BigData* es **recopilar información**. Esta puede provenir de sensores ambientales como velocidad del viento, temperatura, humedad, corrientes oceánicas, entre otros. Una vez que se recopila la información, es necesario estandarizar y **organizar** para poder analizar los datos. Posteriormente, los científicos utilizan una variedad de técnicas para encontrar **tendencias y patrones**. Los resultados se pueden utilizar para tomar decisiones y hacer **predicciones** (Figura 1).

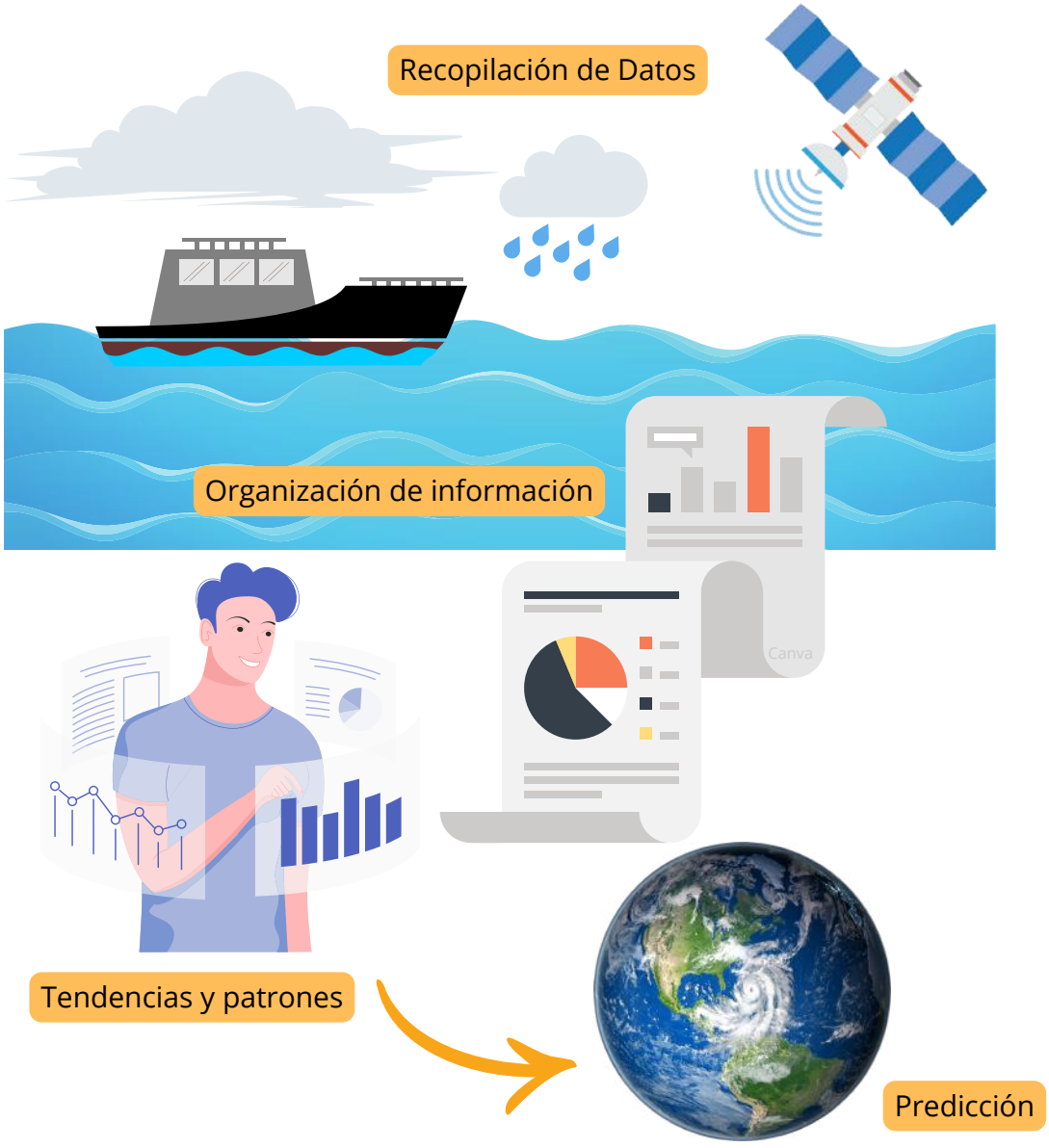


Figura 1. Ejemplo de la ciencia de datos desde la recopilación de la información, organización, análisis y predicción.

La ciencia de datos es un esfuerzo de colaboración que requiere diferentes habilidades. Los investigadores deben comunicar sus hallazgos a quienes no son expertos en el campo y necesitan trabajar con otros para diseñar e implementar soluciones. La demanda de habilidades en la ciencia de datos está creciendo rápidamente y existe escasez de científicos calificados para ello.

El desarrollo de nuevas tecnologías para obtener y analizar la inmensa cantidad de datos ha sido abrumador. Actualmente, esta demanda se traduce en una necesidad sin precedentes de contar con recursos humanos con una formación de análisis de datos especializado. Por lo tanto, es evidente la necesidad de proporcionar herramientas, materiales de capacitación y la generación de nuevas técnicas. Una de las principales preocupaciones de la comunidad científica internacional sobre el crecimiento desproporcionado de datos es que la investigación pueda ser replicable y reproducible.

¿Qué significa que una investigación sea replicable y reproducible?

Una parte clave de los estudios es que puedan corroborar los hallazgos científicos previos, lo que se conoce como **replicabilidad**. Un ejemplo de replicabilidad sería si se arrojaran distintos objetos desde un edificio y se observaran los efectos de la gravedad cuando caen hacia el suelo. Este experimento podría replicarse y diferentes grupos de investigación de todo el mundo podrían obtener resultados similares y confirmar los descubrimientos científicos (Figura 2).

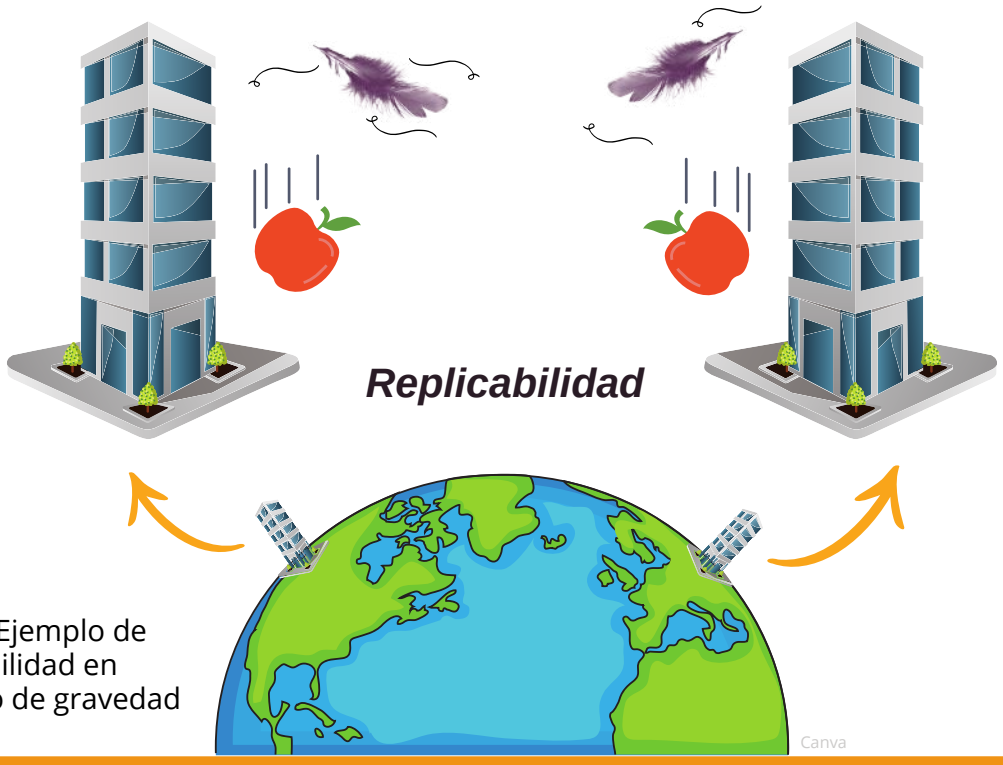


Figura 2. Ejemplo de replicabilidad en experimento de gravedad

Por otro lado, debido a la gran cantidad de datos, es necesario que el procesamiento y las condiciones del análisis sean consistentes para conseguir resultados similares por parte de otros. Esto es conocido como **reproducibilidad**. Una analogía para esto sería seguir una receta y usar la misma cantidad de ingredientes para obtener una hamburguesa exactamente idéntica (con tiempo de cocción para la carne y disposición de ingredientes idéntico, figura 3).

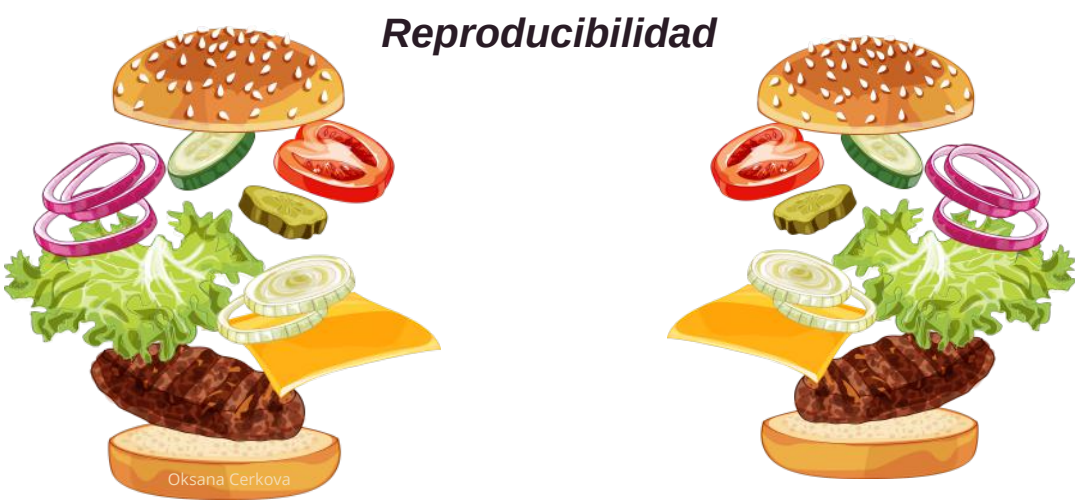


Figura 3. Ejemplo de Reproducibilidad.

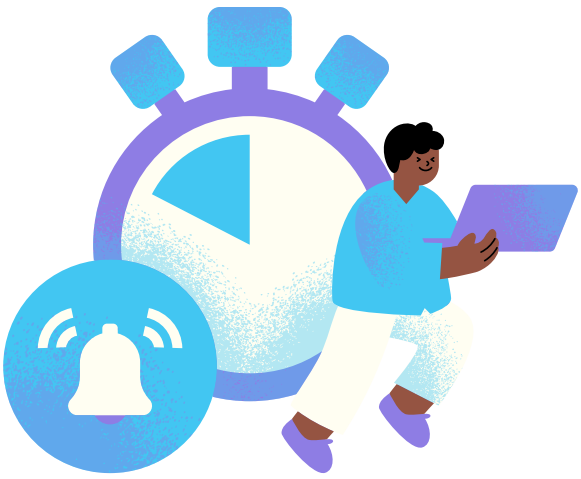
Una característica importante de los estudios científicos es la reproducibilidad, esto es que cualquier persona tenga la posibilidad de recrear los métodos o experimentos desarrollados en la investigación. Así la replicabilidad y reproducibilidad son una parte central de la ciencia y beneficia a los avances y desarrollo de tecnología en todos los ámbitos.

Importancia de la reproducibilidad

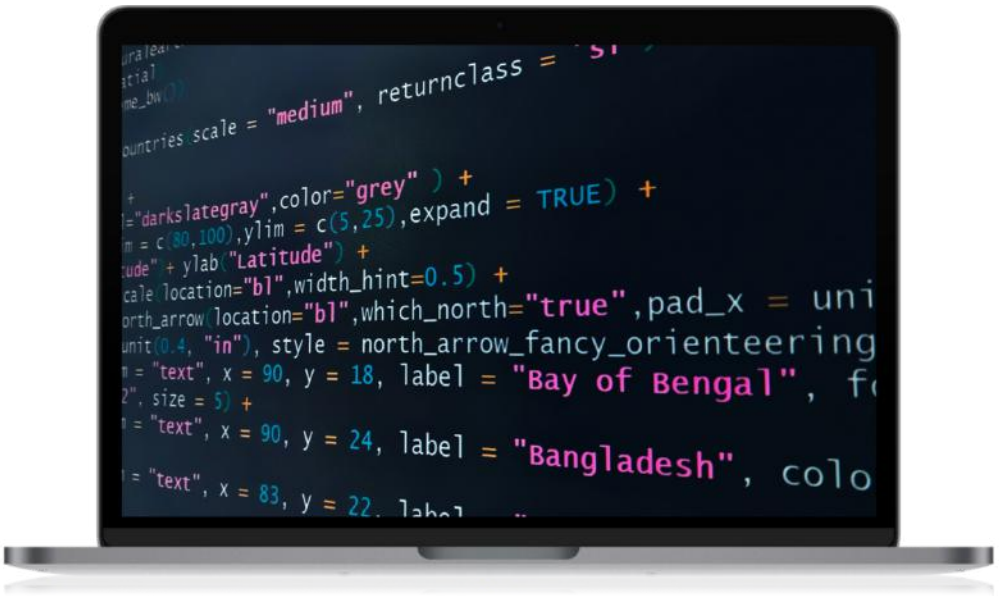
En la actualidad, el uso de los recursos computacionales es sumamente crítico para el análisis de los datos. Es importante reconocer que la formación de personal capacitado hacia las ciencias computacionales, creación de **códigos** y **bioinformática** se encuentra limitado en comparación con la demanda que existe. Una carrera profesional con formación académica en análisis masivo de datos se convertirá en una necesidad indispensable en el futuro.

Es primordial conocer las diferentes pautas básicas para desarrollar mejores prácticas computacionales para la creación de códigos. La ausencia de estas puede resultar en la pérdida de datos e ineficiencia en el uso de recursos computacionales y por tanto la obtención de los resultados y el desarrollo de la investigación lleven mucho más tiempo de lo necesario.

Los análisis informáticos deben contener un flujo de trabajo como el que se realiza en los laboratorios, con las bitácoras, códigos empleados y los pasos que se llevan a cabo para desarrollar la investigación. Esto conlleva también a las anotaciones de los errores en los códigos y las versiones utilizadas en los mismos para asegurarse que cuando se comparten funcionan correctamente.



El flujo de trabajo y documentos reproducibles, como los documentos **markdown en lenguaje R,** incluyen los datos crudos sin procesar pasando por la conversión y exploración de la información mediante gráficos, creación de modelos y análisis estadísticos hasta la obtención de los archivos finales del artículo científico. Estos documentos deben de contener la información detallada de los materiales empleados. Un código bien documentado permite su reutilización por otras investigaciones y hace que su verificación sea sencilla. Algunos de los beneficios al compartir estos materiales reproducibles incluyen una mayor comprensión y reutilización de los mismos, además de que pueden examinarse y mejorarse.



Para Consulta

Camargo-Vega et al. 2015. Conociendo Big Data. Revista Facultad de Ingeniería. 24 (38). http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006

Allen C, Mehler DMA (2019) Open science challenges, benefits and tips in early career and beyond. PLoS Biol 17(5): e3000246. https:// doi.org/10.1371/journal.pbio.3000246.

QuantumFracture. 2019. ¿Por qué una Pluma y un Martillo Caen a la Vez? https://youtu.be/EzcyW0naDLw

David A. Paz García

Doctor en Ciencias en Uso, manejo y preservación de los recursos naturales. Investigador CONACyT-Centro de Investigaciones Biológicas del Noroeste. Contacto: dpaz@cibnor.mx