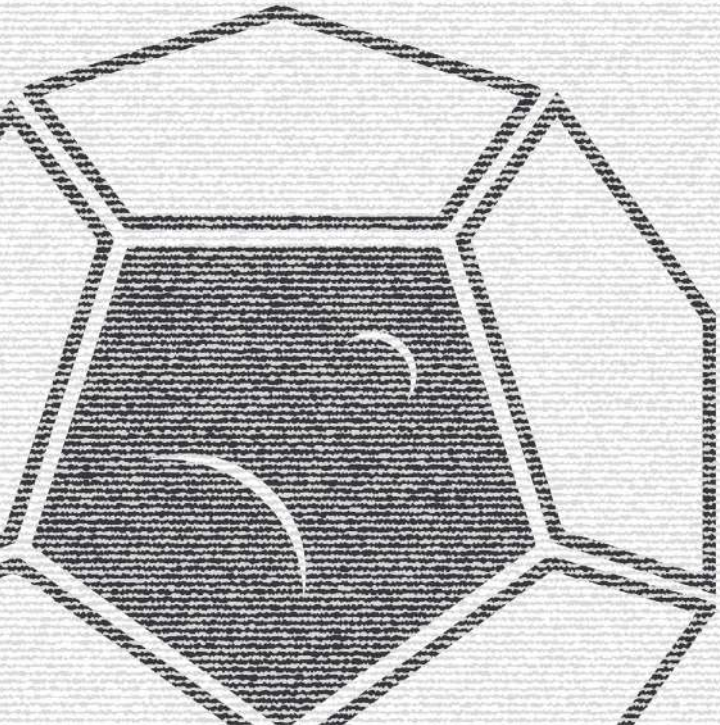
Maximiliano Bozzoli
Luis Salvatico
David Merlo
(Eds.)

Epistemología e Historia de la Astronomía

Volumen I



Epistemología e Historia de la Astronomía

Volumen I

Maximiliano Bozzoli
Luis Salvatico
David Merlo
(Eds.)

Colecciones del CIFFyH

Epistemología e historia de la Astronomía / Maximiliano Bozzoli ... [et al.]; compilación de Luis Salvatico; David C. Merlo. - 1a ed. - Córdoba : Universidad Nacional de Córdoba. Facultad de Filosofía y Humanidades, 2023.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-33-1721-1

1. Astronomía. 2. Historia. 3. Epistemología. I. Bozzoli, Maximiliano. II. Salvatico, Luis, comp. III. Merlo, David C., comp.
CDD 520.3

Publicado por
Área de Publicaciones de la Facultad de Filosofía y Humanidades - UNC
Córdoba - Argentina
1º Edición

Área de
Publicaciones

Diseño de portadas: Manuel Coll y María Bella
Diagramación: María Bella
Imagen portada: "*JEHA (Jornadas de Epistemología e Historia de la Astronomía)*" (2021), de Maximiliano Bozzoli

2023

Expertos en la oscuridad. Datos, archivos y cómo usarlos

Julián Reynoso*

Resumen

La posibilidad de recolectar y acumular grandes volúmenes de datos en el auge de lo que se supo denominar *data deluge* vino acompañada también un despliegue de infraestructura y estrategias para poder manejar semejante diluivio sin precedentes. Esto también trajo aparejado numerosos problemas, tanto epistemológicos como metodológicos. En este sentido, son numerosas las propuestas que señalan la necesidad de crear y formar profesionales cuya principal responsabilidad sea la de curar los datos recopilados y almacenados, no como una sub-tarea de la investigación sino como una especialidad en sí misma. Schembera & Duran (2020) proponen la creación de un Chief Data Officer en los centros de supercomputacion, mientras que Leonelli (2013, 2018) por su parte propone la necesidad de incorporar las habilidades necesarias para la curación de datos no como una habilidad de un técnico sino como una rama de la investigación científica.

Palabras clave: *archivo, big data, experticia humana, curación de datos, datos oscuros.*

Abstract

Having the possibility of collecting and accumulating large volumes of data, christened "big data", was accompanied by a deployment of infrastructure and strategies to handle such an unprecedented volume of data. This also brought with it several problems, both epistemological and methodological. In this sense, there are numerous proposals that point to the need to create and train professionals whose main responsibility is to

* Centro de Investigaciones de la Facultad de Filosofía y Humanidades (UNC) – Instituto de Humanidades (CONICET).

curate the data collected and stored, not as a sub-task of research but as a specialty. Schembera & Duran (2020) propose the creation of a Chief Data Officer in supercomputing centers, while Leonelli (2013, 2018) proposes the need to incorporate the necessary skills for data curation not as a technician's skill but as a branch of scientific research.

Keywords: *archive, big data, human expertise, data curation, dark data.*

Introducción

La astronomía no es ninguna extraña a las complejidades que implica el manejo de grandes volúmenes de datos, con múltiples procedencias y amplios períodos temporales. Desde la temprana modernidad ha sido posible observar un proceso paulatino de estandarización, que se consolidó a mediados del siglo XX. Un informe de 1998 del Sistema de Datos Astrofísicos (ADS) de la NASA catalogaba como "antiguos" aquellos datos que hubieran sido obtenidos antes de 1940 (cf. Kurtz y Eichhorn, 1998, p. 293)

El proceso de acumulación de datos, registros y observaciones se incrementó exponencialmente gracias a las posibilidades que abrieron los numerosos avances en la tecnología de sensores, capacidad de cómputo y almacenamiento que comenzaron en la segunda mitad del Siglo XX. Un verdadero "diluvio de datos" que empapó a prácticamente todas las disciplinas científicas. La astronomía no solo no fue la excepción, sino que en muchos casos estuvo a la vanguardia del proceso.

En la última década, sin embargo, han surgido numerosas promesas de sistemas automáticos que permitirían obtener resultados inesperados y novedosos a partir del mero *input* de datos y simulaciones. El advenimiento del "diluvio de datos", sumado a avances en *machine learning* y otras áreas de inteligencia artificial generaron grandes expectativas en torno a una potencial revolución en la manera de hacer ciencia.

En el presente trabajo propongo analizar algunas de estas promesas a la luz de la noción de experticia humana, con la intención de rescatar algunas de las dificultades que la automatización por vía de *machine learning* no pueden aún sortear.

Experticia

La noción de experticia ha recibido muchos aportes desde distintos pensadores que han trabajado en áreas como la filosofía de la mente, teoría de la acción y teoría del conocimiento. Gilbert Ryle, por ejemplo, contribuyó con la distinción entre "saber qué" y "saber cómo", en sus trabajos a finales de la década de 1940.

Polanyi (1958) contribuyó a su estudio, partiendo de la idea de que pareciera haber un "conocimiento tácito", donde vertió estudios desde la sociología, la psicología y la filosofía. Harry Collins fue, sin embargo, quién más elaboró a partir de los trabajos de Polanyi y propuso dos definiciones sobre el conocimiento tácito. Una versión débil, como "aquello que no es explícito ni expresado de manera directa". La versión más fuerte ya impone una cláusula de imposibilidad: aquello que no *puede* ser explícito o explicitado.

En trabajos posteriores en colaboración con Evans (H. M. Collins y Evans, 2007) vinculan fuertemente la noción de experticia con la socialización lingüística, con la idea wittgensteniana de "formas de vida", pero que podríamos resumir en el conjunto de prácticas, hábitos y costumbres que se dan en el centro de investigación. Así, plantean un giro sociológico dentro del campo de estudio de la experticia al defender que el aprendizaje y la práctica constante que llevan a un novato a convertirse en experto solo son posibles si se dan en una comunidad social de expertos. En consecuencia, el experto es quien que logra convertirse en un miembro de un determinado grupo social, compartiendo un lenguaje común y aprendiendo las maneras de actuar apropiadas y aceptadas por tales. (cf Collins y Evans, 2018; Mondaca y Reynoso, 2022).

Otro gran afluente de estos estudios proviene de intentar comprender la experticia humana a la luz de la experticia en las máquinas, un campo que se conoce como *expert systems* en el que se intenta emular las capacidades de toma de decisión de expertos humanos en dominios específicos. Esto significa que requería un gran esfuerzo programar estos sistemas y su aplicabilidad está muy constreñida al dominio, lo que los pone en relativa desventaja en relación a otros medios automatizados de trabajo con pretensiones de generalidad. Los sistemas expertos deben ser cuidadosamente ajustados para cada dominio, lo que hace difícil reutilizarlos luego en otros dominios.

La práctica de los sujetos expertos se encuentra siempre inserta en una práctica social más amplia. De tal manera, los sistemas expertos que han mostrado un nivel de rendimiento igual o superior al de sujetos expertos no pueden considerarse como ejemplos de experticia hasta tanto no adquieran el nivel de socialización propio de expertos humanos -lo cual, para Collins, no puede ser alcanzado por máquinas (Mondaca y Reynoso, 2022).

> La noción de experticia interactiva se termina de comprender en 2013, donde Collins presenta un artículo muy sugerente que reúne sus tesis de los últimos veinte años bajo lo que el autor denomina el modelo tridimensional de la experticia. (Mondaca en prensa, 2022)

En filosofía de la ciencia, sin embargo, estas discusiones sobre experticia no han sido incorporadas al canon de estudio sino hasta algunos años, por lo que resulta una noción fértil para incorporar en estudios centrados en prácticas científicas.

Trabajo de archivo en astronomía

La historia del material de archivo en astronomía es larga y no sería una exageración ligarla al desarrollo de la escritura cuneiforme en Babilonia. Los estudiosos en la antigua Grecia primero, y en el imperio romano después es maravillarían por el enorme caudal de observaciones que habían acumulado. Es difícil no preguntarse, entonces, qué pensarían de las enormes cantidades de datos que la disciplina recolecta y acumula en nuestros días. Según el trabajo realizado por Kremer et al. (2017), se estima que por cada noche de observación, el *Very Large Telescope* del Observatorio Europeo Austral era capaz de recolectar hasta 10 GB. El *Sloan Digital Sky Survey*, operativo apenas dos años después que el VLT en el año 2000, ya alcanzaba los 200 GB por noche. Si bien aún no operativos, el Observatorio Vera C. Rubin podría ser capaz de generar hasta 30 TB[1], mientras que el *Thirty Meter Telescope* podría generar hasta 90 TB. Es notorio, sin embargo, poder trazar ciertos paralelos entre el estado de la astronomía a mediados del siglo XVII y la actualidad. El desarrollo de nuevas tecnolo-

1 [^1]: A modo de referencia, 1 TB (terabyte) de almacenamiento equivale a 1024 GB (gigabytes).

gías permitió fabricar nuevas lentes e instrumentos que abrieron la puerta a un caudal de nuevas observaciones, que ponían en cuestión el "estado del arte" de los conocimientos en aquella época. Hsia (2017) Argumenta que esto también obligó a un desarrollo en las formas en las que dicha información era almacenada, curada y puesta en a disposición (en la mayoría de los casos) a otros investigadores ávidos de corroborar las mediciones que habían realizado sus colegas. Nuevos estándares fueron fijados, a través de una red intrincada de correspondencia personal en la que se intercambiaban protocolos de observación, datos en varias maneras y detalles del instrumental utilizado. Así, lentamente, se fue constituyendo un *canon* archivístico de observaciones, incrementado generación tras generación de astrónomos. La estandarización de los registros y las ansias por socializar Esta tendencia continúa hoy en día, con un caudal de información vastamente superior. En palabras de Hsia:

> Sin embargo, un reto clave para los astrónomos contemporáneos que trabajan en la vanguardia de la "ciencia intensiva en datos" -el llamado "cuarto paradigma" de la investigación científica- es uno con el que sus predecesores han luchado durante mucho tiempo: cómo capturar, conservar y analizar datos empíricos. (Hsia, 2017, p. 37)

El gran diluvio de datos que lo empapa todo

El concepto de "*big data*" ha sido empleado de muchas maneras muy distintas como una forma de agrupar una serie de avances en materia de captura, almacenamiento y procesamiento de datos de procedencias muy diversas.

En torno a dicho concepto se elaboraron múltiples promesas sobre cómo revolucionaría la comprensión del mundo, optimizaría toma de decisiones, mitigaría los riesgos e incluso mejoraría la calidad de los datos obtenidos. *Big data* fue presentado como un hito histórico a la altura de la revolución industrial del siglo XIX o la revolución científica del siglo XVIII.

El impacto más grande fue en los sectores tecnológicos e industriales, donde la promesa de mitigación de riesgos y optimización caló hondo, sin embargo, la fanfarria fue tal que el despliegue también llegó a disciplinas científicas.

Expresiones maximalistas, por no decir exageradas, fueron planteadas (y rápidamente ignoradas por la comunidad científica) como las de Chris Anderson quien llegó a afirmar que gracias al big data sería posible prescindir del método científico. Esta versión radical, se monta en la idea de que -Ley de Los Grandes Números mediante- con suficientes muestras, los errores y la incerteza deberían tender a cero. Hubo, sin embargo, expresiones mas moderadas como "El Cuarto Paradigma" de Jim Gray, una metodología basada en la noción de ciencia *data-intensive*, que consiste en la captura, curación y análisis de grandes cantidades de datos (Hey et al., 2009). Aquí "paradigma" no está usado en un sentido kuhniano, menos aún si tenemos en cuenta que muchos de quienes proponen esta metodología la consideran complementaria a los "antiguos paradigmas" que Gray considera: el empírico-observacional, teórico-analítico y un tercero de simulaciones a gran escala.

El atractivo es evidente a primera vista. Succi y Coveney (2019) lo resumen en cuatro puntos:

1) El crecimiento exponencial en la capacidad de producción, adquisición y navegación de los datos.

2) Rastrear patrones de bases de datos complejas a través de algoritmos de búsqueda "inteligente" promete ser más rápido y reveladora que la modelización del comportamiento subyacente (es decir, usar teorías)

3) Puede aplicarse a cualquier disciplina (y la astronomía es un hermoso ejemplo). Pero incluso en aquellas que tradicionalmente no han sido muy susceptibles al tratamiento matemático. (Succi y Coveney dicen que es una forma de sugerir que estos dominios son muy complejos para ser modelados).

4) Pueden tener aplicación "inmediata" (lo que trae todo un conjunto de complicaciones éticas, en las que no me explayaré a los efectos de este trabajo)

Sin embargo, las promesas de semejante revolución probaron ser, en el mejor de los casos, una expresión de deseo que hasta ahora no ha estado

a la altura del revuelo que causó. Las disciplinas científicas han mostrado ser bastante menos permeables a "comprar" las promesas de las propuestas más radicales en torno a big data. Es importante aclarar también que esto no significa que no se hayan adoptado un gran conjunto de métodos y técnicas de filtrado que hacen una primera gran "zaranda" de datos, que luego son puestos a disposición de los investigadores. Y es aquí donde se introduce una gran fuente de nuevos problemas que no son susceptibles de ser tratados por vías automatizadas.

Dark data y archivos fragmentarios

La noción de "experticia" que expuse en el primer apartado resulta de gran utilidad, puesto que permite comprender algunas implicancias intrínsecas al trabajo que llevan adelante los investigadores en sus centros. Menciono esto porque la literatura que ha trabajado big data en contextos científicos se ha centrado en ensalzar los beneficios, sin tener en cuenta algunas consecuencias no deseables. De manera contraintuitiva es posible observar cómo la digitalización de los registros y los datos ha causado que, en muchos casos, se pierdan o queden inutilizables por distintos aspectos de la infraestructura. En un artículo reciente Schembera y Durán (2020) exponen algunas de estas dificultades. Big data, dicen los autores, se enfoca en aquellos datos que están disponibles a los distintos usuarios. Sin embargo, el caudal de datos es tal que en muchas ocasiones queda inaccesibles a los usuarios y se los denomina "*dark data*", término que los autores toman de (Heidorn, Stahlman, y Steffen, 2018), pero señalan que no es apta para entornos de computación de alto rendimiento (como sería el caso de los grandes observatorios ya mencionados). Esto obedece a múltiples razones, algunas tan mundanas como fallas en los equipos que los almacenan o que el investigador proceda de manera descuidada, sin usar correctamente las etiquetas (metadatos) que permiten clasificarlos y ordenarlos. El fenómeno del *bit rot*, la paulatina desintegración de los datos que quedan guardados en distintos medios de almacenamiento es también otro factor a tener en cuenta. El CERN, por ejemplo, mantiene grandes colecciones de datos en cintas magnéticas por las ventajas que ofrece en almacenamiento a largo plazo[2]. La diferencia fundamental entre la noción

2 Cf.: https://home.cern/science/computing/storage

de Heidorn, Stahlman, y Steffen (2018) y la de Schembera y Durán (2020) radica en el interés del primero en la "larga estela de la ciencia", en un conjunto grande de proyectos más bien pequeños cuya producción de datos puede ser más manejable siguiendo ciertos protocolos bien establecidos. Para los segundos, en cambio, el foco está puesto en investigación de Centros de Computación de Alto Rendimiento en los que el equipamiento y la infraestructura utilizada es específica y determinada.

Un aspecto no menor a tener en cuenta en torno al trabajo cotidiano de los investigadores es la potencial pérdida y fragmentación de los archivos que cada uno almacena y custodia, a veces sin tener plena conciencia de ello, en las nubes y computadoras personales de cada uno. La digitalización de los archivos, a pesar de sus innumerables ventajas, ha llevado a que muchos de estos archivos queden "congelados" o atrapados en las cuentas personales de almacenamiento en la nube. Esto es un fenómeno relativamente novedoso, dado que no siempre es fácilmente accesible para terceras personas, por razones de privacidad y custodia de datos personales.

No da lo mismo un sistema experto que una persona experta

En base a lo visto hasta ahora, es posible afirmar que el excesivo optimismo inicial que hubo en estrategias automatizadas para procesar los grandes volúmenes de datos que continúan produciéndose en las diversas disciplinas científicas no fue acompañado de un gran éxito en la tarea. Esto no quiere decir que los métodos de *machine learning* no sean de utilidad, como hemos visto en las secciones anteriores no sería posible manejar tal diluvio de datos si no hubiera filtros algorítmicos y otros métodos automatizados para procesarlos.

Sin embargo, quiero dedicar las últimas secciones de este trabajo para rescatar la noción de experticia que mencioné anteriormente para poner en otro contexto la mesura respecto de las posibilidades que desarrollos en IA ofrecen en contextos científicos. Puntualmente, han surgido diversas propuestas en relación a la necesidad de respaldar, tanto con financiamiento como también con líneas de formación, trayectos de carrera especializados en manejo de datos.

Una característica central de la noción de dark data que mencionamos anteriormente la forma en la que plantea la necesidad de convertir el manejo de archivos en una prioridad para las instituciones, en lugar de

delegar esa tarea a los investigadores. El *Scientific Data Officer* ("Responsable de Datos Científicos") [^2] que proponen (Schembera y Durán, 2020) será un nuevo sendero profesional dentro de las instituciones científicas, que actuará como intermediario entre investigadores, administrativos y directivos con la principal responsabilidad de almacenar, curar y presevar los datos dentro de un centro de computación de alto rendimiento. Esto permitiría relevar a los investigadores de esta tarea, que tiende a percibirse como de segundo orden, dada la presión cada vez mayor en concentrarse en publicaciones. Si bien cada vez más organismos de financiación requieren poner a disponibilidad de la comunidad los datos primarios empleados en las investigaciones, tal requerimiento rara vez se acompaña con infraestructura adecuada en términos institucionales para realizar esa tarea. [^2]: La traducción es mía.

Por su parte, aunque si bien basada en su experiencia como investigadora en ciencias de la vida, Leonelli (2014) rescata la figura del *curador* como responsable del correcto etiquetado de los metadatos que sirven para catalogar los datos que surgen de las investigaciones. La autora pone gran énfasis en la interoperabilidad y puesta a disposición dela comunidad científica a través de la iniciativa FAIR que apunta a optimizar el manejo y la administración de datos, de forma tal de hacerlos. La sigla significa *Findability, Accessibility, Interoperability, and Reuse* (encontrables, accesibles, interoperables y reutilizables).

Estas iniciativas apuntan a brindar a la comunidad científica la mayor cantidad de datos, debidamente curados y etiquetados, aprovechando al máximo el diluvio sin precedentes que se está generando diariamente. La posibilidad de desarrollar nuevos trayectos profesionales dentro de los centros descongestiona la agenda de los investigadores que ya no deberán encargarse del archivo, lo que disminuirá las posibilidades de sub-utilización de los recursos, al tiempo que permitirá una mayor interoperabilidad entre grupos de investigación. Aquí vuelve a reflotar la dimensión social de la noción de experticia de Collins y la necesidad de incorporar aspectos idiosincráticos a la hora de diseñar e implementar sistemas de archivo y socialización de datos.

Bibliografía

Collins, H. M., y Robert Evans. 2007. *Rethinking Expertise.* Chicago: University of Chicago Press.

Collins, Harry, y Robert Evans. 2018. "A Sociological/Philosophical Perspective on Expertise: The Acquisition of Expertise Through Socialization." In *The Cambridge Handbook of Expertise and Expert Performance*, edited by A. Mark Williams, Aaron Kozbelt, K. Anders Ericsson, y Robert R. Hoffman, 2nd ed., 21–32. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316480748.002.

Heidorn, P. Bryan, Gretchen R. Stahlman, y Julie Steffen. 2018. "Astrolabe: Curating, Linking, and Computing Astronomy's Dark Data." *The Astrophysical Journal Supplement Series* 236 (1): 3. https://doi.org/10.3847/1538-4365/aab77e.

Hsia, Florence. 2017. "Astronomy After the Deluge." In *Science in the Archives: Pasts, Presents, Futures*, edited by Lorraine Daston, 17–52. University of Chicago Press. https://doi.org/10.7208/9780226432533-003.

Kremer, Jan, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, y Christian Igel. 2017. "Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy." *IEEE Intelligent Systems* 32 (02): 16–22. https://doi.org/10.1109/MIS.2017.40.

Kurtz, Michael J., y Guenther Eichhorn. 1998. "The Historical Literature of Astronomy, via ADS" 153 (January): 293. https://ui.adsabs.harvard.edu/abs/1998ASPC..153..293K.

Leonelli, Sabina. (2014). "What difference does quantity make? On the epistemology of Big Data in biology". *Big Data & Society*, 1(1), 205395171453439. https://doi.org/10.1177/2053951714534395

Mondaca, Sofía, y Julián Reynoso. 2022. "Experticia Humana y Opacidad Epistémica En Contextos de Prácticas Científicas." In *Filosofía de La*

Ciencia Por Jóvenes Investigadores Vol. 2, editado por María Paula Buteler, Ignacio Heredia, Santiago Marengo, y Sofía Mondaca, 2:117–26.

Polanyi, Michael. 1958. *Personal knowledge; towards a post-critical philosophy*. Chicago: University of Chicago Press.

Schembera, Björn, y Juan M. Durán. 2020. "Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer." *Philosophy & Technology* 33 (1): 93–115. https://doi.org/10.1007/s13347-019-00346-x.

Succi, Sauro, y Peter V. Coveney. 2019. "Big Data: The End of the Scientific Method?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 377 (2142): 20180145. https://doi.org/10.1098/rsta.2018.0145.