

La mayoría de los campos de la astronomía están cambiando rápidamente. Existen, o pronto estarán disponibles, grandes cantidades de datos observacionales sin precedentes. Ante este incremento de datos, la Inteligencia Artificial o Aprendizaje Automático se ha convertido a lo largo de la última década en un ingrediente cada vez más habitual en el análisis científico en astrofísica. El objetivo de esta reseña es proporcionar una visión global y relativamente superficial de la evolución del campo en los últimos años así como de los desarrollos presentes y futuros. Es imposible abarcar todas las aplicaciones en todos los campos de la astrofísica, por ello me centro casi exclusivamente en el uso de la IA para astronomía extragaláctica, evolución de galaxias y cosmología, campos que conozco de primera mano.



Marc Huertas-Company
Instituto de Astrofísica de Canarias
marc.huertas.company@gmail.com

Aunque las aplicaciones del aprendizaje automático a la astronomía existen desde hace al menos treinta años, los últimos años han sido testigos de un aumento sin precedentes de los métodos de aprendizaje profundo traducido en un incremento exponencial de las publicaciones (Figura 1). Este resurgimiento se ve impulsado por importantes avances en el campo del Aprendizaje Automático desde la popularización de las Redes Neuronales Convolucionales (CNNs) hace más de una década (Krizhevsky et al., 2012).

VOLUMEN Y COMPLEJIDAD DE DATOS SIN PRECEDENTES

Estudios espectrofotométricos modernos como el Legacy Survey of Space and Time (LSST; Ivezić et al., 2019) o Euclid (Laureijs et al., 2011) proporcionarán o están ya proporcionando espectros e imágenes de alta calidad de cientos de millones de galaxias. Los cartografiados espectroscópicos de campo integral a bajo y alto desplazamiento al rojo están alcanzando tamaños estadísticamente relevantes (por ejemplo, MaNGA - Bundy et al., 2015) que permiten resolver la estructura interna de las galaxias más allá de las propiedades integradas.

Además, nuevas instalaciones como el telescopio espacial James Webb (JWST) están abriendo la ventana a un desplazamiento al rojo y un régimen de masas estelares completamente nuevos, tanto en imágenes como en espectroscopia, y podremos ser testigos de la aparición de las primeras galaxias del universo. Las instalaciones de rayos X y radio (por ejemplo, SKA, Athena) sondearán el gas frío y caliente de las galaxias con una resolución mejorada. En el ámbito teórico, la potencia de cálculo ha evolucionado hasta el punto de que ahora podemos generar simulaciones realistas de galaxias en un contexto cosmológico que abarca la mayor parte de la historia del Universo (por ejemplo, TNG - Pillepich et al., 2018) que reproducen adecuadamente un gran número de propiedades observables. En este contexto de creciente complejidad y rápido aumento de los volúmenes de datos, se ha convertido en un nuevo reto para la comunidad combinar y extraer con precisión información científicamente relevante de estos conjuntos de datos.

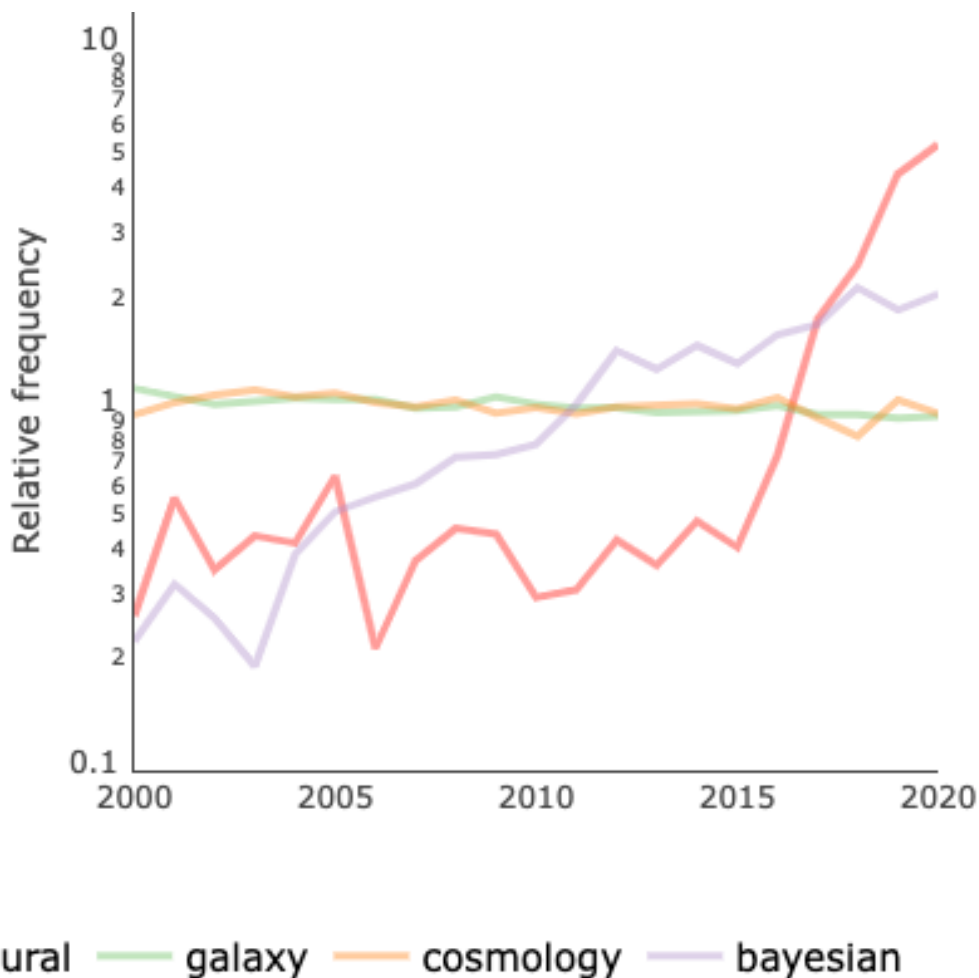


Figura 1. Cambio relativo del número de publicaciones en arXiv:astro-ph con diferentes palabras clave en el abstract a lo largo de los últimos años. El número de trabajos que mencionan redes neuronales ha experimentado un crecimiento sin precedentes en los últimos años, significativamente más pronunciado que otros ámbitos de la astrofísica. Fuente: ArXivSorter. Figura tomada de Huertas-Company & Lanusse (2023)

APRENDIZAJE AUTOMÁTICO Y ASTROFÍSICA: UNA HISTORIA DE MÁS DE 30 AÑOS

Hay una larga historia de aplicaciones de aprendizaje automático en astronomía que comenzó mucho antes del más reciente boom del aprendizaje profundo. Diferentes tipos de algoritmos de aprendizaje automático, incluidas las primeras redes neuronales artificiales (ANN), árboles de decisión (DT), bosques aleatorios (RF) o algoritmos kernel como las máquinas de vectores de soporte (SVM), se han aplicado

a diferentes áreas de la astrofísica desde la segunda mitad del siglo pasado. Por ejemplo, las ANN, los árboles de decisión y los mapas autoorganizativos (SOM) se han aplicado ampliamente a la clasificación de estrellas y galaxias (Odewahn et al., 1992; Weir et al., 1995; Miller et al., 1996; Bazell et al., 1998; Andreon et al., 2000; Qin et al., 2003; Ball et al., 2006). El problema de la clasificación de la morfología de las galaxias también ha sido objeto de una cantidad significativa de trabajos relacionados con

el aprendizaje automático (Storrie-Lombardi et al., 1992; Lahav et al., 1995, 1996; Odewahn et al., 1996; Naim et al., 1997; Madgwick et al., 2003; Cohen et al., 2003). Ball et al. (2004) es probablemente el primer trabajo en utilizar ANNs para clasificar galaxias en el SDSS. En la primera década del presente siglo las SVM se hicieron más populares y también se utilizaron para proporcionar catálogos de morfología de galaxias (Huertas-Company et al., 2008, 2011). Los Árboles de Decisión también se han aplicado a otras tareas de clasificación como la separación AGN/galaxia (White et al., 2000; Gao et al., 2008). Más allá de la clasificación, el aprendizaje automático, y especialmente las ANNs, se han aplicado ampliamente al problema de la estimación de corrimientos al rojo fotométricos (D'Abrusco et al., 2007; Li et al., 2007; Banerji et al., 2008). Remitimos al lector a Ball et al. (2010) y Baron et al. (2019) para una revisión completa y extensa de las técnicas de aprendizaje automático pre-deep learning aplicadas a la astronomía.

LA REVOLUCIÓN DEL APRENDIZAJE PROFUNDO O DEEP LEARNING

El aprendizaje profundo designa generalmente las técnicas de optimización basadas en gradientes de arquitecturas modulares de complejidad variable; es, por tanto, un subcampo de la disciplina más general del aprendizaje automático. En muchos aspectos, el aprendizaje profundo representa un cambio en la forma de abordar el análisis de datos.

El acceso a grandes conjuntos de datos junto a la disponibilidad de recursos informáticos lo suficientemente potentes, han permitido durante la última década pasar de un enfoque centrado en algoritmos y basado en características diseñadas manualmente a un enfoque de aprendizaje de características no supervisado y totalmente basado en datos. Esto implica que, en lugar de desarrollar algoritmos avanzados específicos de cada dominio para cada tarea, nos basamos en un algoritmo de optimización genérico para extraer las características más significativas en un bucle de entrenamiento end-to-end. Se trata de un nuevo enfoque de los datos en astrofísica y en la ciencia en general.

Desde mi punto de vista, y esto es evidentemente una visión subjetiva, el aporte principal del aprendizaje profundo a la ciencia ha sido la capacidad para

muestrear distribuciones de probabilidad de muy alta dimensionalidad, tarea que hasta ahora resultaba prohibitiva o directamente imposible. Esto tiene implicaciones profundas en la manera en la que se analizan los datos y se extrae información. Es el objetivo de esta reseña dar una breve descripción de esta transición en curso.

LA CLASIFICACIÓN: UNA HISTORIA DE ÉXITO

Un ejemplo paradigmático de cómo el aprendizaje profundo está cambiando el análisis científico es la clasificación de fuentes. Es un paso básico de procesamiento de primer orden en la mayoría de los grandes cartografiados pero resulta muy ilustrativo. La rápida penetración del aprendizaje profundo puede explicarse de forma natural porque se trata, posiblemente, de la aplicación «lista para usar» más sencilla. De hecho, el aprendizaje profundo comenzó a atraer la atención de la comunidad de visión por ordenador cuando las redes neuronales convolucionales ganaron por primera vez el concurso ImageNet de clasificación de imágenes (Krizhevsky et al., 2012).

Una de las primeras tareas que realizan los científicos cuando se enfrentan a un problema complejo es identificar objetos morfológicamente similares. En astronomía extragaláctica, la clasificación de objetos puede ser de distintos tipos. En el caso de las imágenes, las aplicaciones más comunes son la clasificación morfológica de galaxias, la separación estrella-galaxia y la detección de lentes gravitacionales fuertes.

En todas estas aplicaciones, el aprendizaje profundo se ha convertido en menos de una década en el método privilegiado para la clasificación de objetos en grandes conjuntos de datos. Catálogos en grandes cartografiados basados en aprendizaje profundo son cada vez más habituales y se usan de modo rutinario para el análisis científico. Esto se debe fundamentalmente a que las redes neuronales proporcionan a la vez un incremento significativo de la calidad de la clasificación a la vez que reducen el tiempo de cálculo.

Aunque es un caso muy específico, es un claro ejemplo de cómo el deep learning ha transformado un campo de investigación y actuado como acelerador de la investigación científica. Hace poco más de una

década no había una respuesta clara a cómo clasificar fuentes en la nueva generación de cartografiados como Euclid o LSST. Se sabía que se había llegado al límite de la ciencia ciudadana (Lintott et al., 2008). Por ejemplo, clasificar todas las galaxias de Euclid mediante ciencia ciudadana tomaría cerca de 100 años. Hoy en día, a pocos meses de la publicación de los primeros datos del telescopio espacial Euclid, la clasificación de cientos de millones de galaxias prácticamente de manera instantánea es posible gracias al aprendizaje profundo (Huertas-Company et al., 2015; Walmsley et al., 2022).

INFERENCIA BAYESIANA Y APRENDIZAJE PROFUNDO: ¿UN MATRIMONIO PERFECTO?

Dada su flexibilidad y la capacidad del aprendizaje profundo para muestrear distribuciones de probabilidad de alta dimensionalidad, su utilización en astrofísica para tareas más complejas y específicas es cada vez más habitual.

En los últimos años hemos asistido a un rápido desarrollo del denominado *simulation-based inference* (Cranmer et al., 2019), una aproximación que fusiona la inferencia bayesiana, tan importante en astrofísica durante los últimos veinte años, con el aprendizaje profundo. La inferencia bayesiana, aunque extremadamente útil para constreñir propiedades físicas, es de manera general muy costosa desde un punto de vista computacional. El muestreo tradicional con cadenas de Markov se torna impracticable en casos donde la dimensionalidad del espacio a explorar es grande. Por otro lado, se requiere de manera general una aproximación gaussiana a la distribución de verosimilitud (*likelihood*), lo que no siempre es una buena aproximación en casos altamente no lineales.

El aprendizaje profundo permite sortear de manera muy eficaz estos escollos, abriendo la posibilidad de realizar inferencias en espacios de muy alta dimensionalidad (¡incluido el espacio de las imágenes!) y relajando la aproximación de gaussianidad. Sin embargo, se requiere un ingrediente indispensable: una simulación fidedigna. Es decir, si uno es capaz de simular el proceso físico en cuestión hasta el nivel más bajo de la observación incluyendo ruido y características de los detectores, el aprendizaje profundo abre las puertas a la inferencia bayesiana en situaciones sin precedentes.

«Se sabía que se había
llegado al límite de
la ciencia ciudadana.
Por ejemplo, clasificar
todas las galaxias de
Euclid mediante ciencia
ciudadana tomaría cerca
de 100 años.

Hoy en día, a pocos
meses de la publicación
de los primeros datos del
telescopio espacial Euclid,
la clasificación de cientos
de millones de galaxias
prácticamente de manera
instantánea es posible
gracias al aprendizaje
profundo.»

Este matrimonio entre la inferencia bayesiana y el aprendizaje profundo ha permitido aplicaciones que hasta ahora eran imposibles, como por ejemplo la inferencia de parámetros cosmológicos a partir de las imágenes directamente (Jeffrey et al., 2021) o de una única galaxia (Villaescusa-Navarro et al., 2022) o la estimación de la historia de fusiones de una galaxia a partir de su cinemática (Pillepich et al., 2024).

Quedan bastantes escollos fundamentales que resolver antes de poder aplicar este tipo de análisis a cualquier conjunto de datos. El más importante es probablemente saber cómo identificar situaciones en las que las simulaciones son casi perfectas pero lo suficientemente diferentes como para introducir sesgos (Aurrekoetxea et al., 2023; Dayal et al., 2023). Por otro lado, se requiere de manera general un conjunto de simulaciones diferentes para cada situación o tipo de dato, lo que, dependiendo de la situación, puede ser un problema en cuanto a tiempo de cómputo.

Sin embargo, mi opinión es que esta manera de analizar datos se va a convertir en cada vez más habitual en los próximos años.

¿HACIA UNA IA FUNDACIONAL?

Como se ha mencionado en los apartados anteriores, uno de los mayores problemas de las aplicaciones de IA es que, de manera general, se requiere un entrenamiento específico para cada situación. Esto implica ya sea un etiquetado o una simulación. Las características diferentes de cada conjunto de datos/instrumentos combinado con el carácter altamente no lineal de las redes neuronales acentúan este problema. Por lo tanto, en la última década hemos visto aparecer un gran número de trabajos que desarrollan o entrenan modelos de IA específicos para sus necesidades y que, por lo tanto, son de difícil exportación. Sin contar que, de manera general, la proporción de datos con etiquetas es pequeña y/o sesgada.

Para hacer frente a estas limitaciones ha surgido en los últimos años, especialmente en el ámbito privado, el concepto de modelo fundacional. Se trataría de un modelo de IA capaz de generalizar lo suficiente como para no solo poderse aplicar a diferentes conjuntos de datos, sino también para diferentes tareas. El ejemplo más exitoso de este tipo de modelo es sin duda ChatGPT y los modelos de lenguaje en general.

¿Cómo se entrena esta IA? La idea es que estos modelos no se entrenan para una tarea específica, sino que se entrenan para entender la estructura de los datos. Por ejemplo, los modelos de lenguaje se entrenan para predecir la palabra siguiente de un texto dado un contexto. Esto tiene al menos una ventaja fundamental y es que se elimina la necesidad de etiquetas y, por lo tanto, se puede entrenar con todos los datos disponibles. De esta manera se consiguen hacer modelos más robustos, que generalicen mejor y que puedan efectuar múltiples tareas.

Los ejemplos en astrofísica son todavía bastante limitados (Huertas-Company et al., 2023), pero es probablemente una cuestión de tiempo. La astrofísica es un campo ideal para este tipo de aprendizaje dada la gran cantidad de datos diversos y generalmente de uso público sin restricciones éticas. Iniciativas como el MultiModalUniverse (MMU) están construyendo enormes bases de datos con datos astronómicos especialmente construidas para el entrenamiento de modelos fundacionales. Esto es un claro ejemplo de un cambio de tendencia. Está por ver dónde nos lleva y qué otras sorpresas nos esperan en este cambio en plena ebullición y que progresa a una velocidad vertiginosa.

REFERENCIAS

- Andreon S., Gargiulo G., Longo G., Tagliaferri R., Capuano N., 2000, *Monthly Notices of the Royal Astronomical Society*, 319, 700
- Angeloudi E., Falcón-Barroso J., Huertas-Company M., Sarmiento R., Pillepich A., Walo-Martín D., Eisert L., 2023, , 523, 5408
- Angeloudi E., Falcón-Barroso J., Huertas-Company M., Boecker A., Sarmiento R., Eisert L., Pillepich A., 2024, *Nature Astronomy*, 8, 1310
- Ball N. M., Brunner R. J., 2010, *International Journal of Modern Physics D*, 19, 1049
- Ball N. M., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R. J., 2004, *Monthly Notices of the Royal Astronomical Society*, 348, 1038
- Ball N. M., Brunner R. J., Myers A. D., Tchong D., 2006, *The Astrophysical Journal*, 650, 497
- Banerji M., Abdalla F. B., Lahav O., Lin H., 2008, *Monthly Notices of the Royal Astronomical Society*, 386, 1219
- Baron D., 2019, Technical report, Machine Learning in Astronomy: a practical overview, <https://ui.adsabs.harvard.edu/abs/2019arXiv190407248B>. <https://ui.adsabs.harvard.edu/abs/2019arXiv190407248B>
- Bazell D., Peng Y., 1998, *The Astrophysical Journal Supplement Series*, 116, 47
- Bundy K., et al., 2015, *The Astrophysical Journal*, 798, 7
- Cohen S. H., Windhorst R. A., Odewahn S. C., Chiarzenza C. A., Driver S. P., 2003, *The Astronomical Journal*, 125, 1762
- Cranmer K., Brehmer J., Louppe G., 2019, Technical report, The frontier of simulation-based inference, <https://ui.adsabs.harvard.edu/abs/2019arXiv191101429C>. <https://ui.adsabs.harvard.edu/abs/2019arXiv191101429C>
- D'Abrusco R., Staiano A., Longo G., Brescia M., Paolillo M., De Filippis E., Tagliaferri R., 2007, *The Astrophysical Journal*, 663, 752
- de Santi N. S. M., et al., 2023, arXiv e-prints, p. arXiv:2310.15234
- Gao D., Zhang Y.-X., Zhao Y.-H., 2008, *Monthly Notices of the Royal Astronomical Society*, 386, 1417
- Huertas-Company M., Lanusse F., 2023, , 40, e001
- Huertas-Company M., Rouan D., Tasca L., Soucail G., Le Fèvre O., 2008, , 478, 971
- Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, , 525, A157
- Huertas-Company M., et al., 2015, , 221, 8
- Huertas-Company M., Sarmiento R., Knapen J. H., 2023, *RAS Techniques and Instruments*, 2, 441
- Ivezić Z., et al., 2019, *The Astrophysical Journal*, 873, 111
- Jeffrey N., Alsing J., Lanusse F., 2021, *Monthly Notices of the Royal Astronomical Society*, 501, 954
- Krizhevsky A., Sutskever I., Hinton G. E., 2012, in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., <https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b-76c8436e924a68c45b-Abstract.html>
- Lahav O., et al., 1995, *Science*, 267, 859
- Lahav O., Naim A., Sodré Jr. L., Storrie-Lombardi M. C., 1996, *Monthly Notices of the Royal Astronomical Society*, 283, 207
- Laureijs R., et al., 2011, Technical report, Euclid Definition Study Report, <https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L>. <https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L>
- Li L.-L., Zhang Y.-X., Zhao Y.-H., Yang D.-W., 2007, *Chinese Journal of Astronomy and Astrophysics*, 7, 448
- Lintott C. J., et al., 2008, *Monthly Notices of the Royal Astronomical Society*, 389, 1179
- Madgwick D. S., 2003, *Monthly Notices of the Royal Astronomical Society*, 338, 197
- Miller A. S., Coe M. J., 1996, *Monthly Notices of the Royal Astronomical Society*, 279, 293
- Naim A., Ratnatunga K. U., Griffiths R. E., 1997, arXiv e-prints, pp astro-ph/9704012
- Odewahn S. C., Stockwell E. B., Pennington R. L., Humphreys R. M., Zumach W. A., 1992, *The Astronomical Journal*, 103, 318
- Odewahn S. C., Windhorst R. A., Driver S. P., Keel W. C., 1996, *The Astrophysical Journal*, 472, L13
- Pillepich A., et al., 2018, *Monthly Notices of the Royal Astronomical Society*, 473, 4077
- Qin D.-M., Guo P., Hu Z.-Y., Zhao Y.-H., 2003, *Chinese Journal of Astronomy and Astrophysics*, 3, 277
- Storrie-Lombardi M. C., Lahav O., Sodre Jr. L., Storrie-Lombardi L. J., 1992, *Monthly Notices of the Royal Astronomical Society*, 259, 8P
- Villaescusa-Navarro F., et al., 2021, *The Astrophysical Journal*, 915, 71
- Villaescusa-Navarro F., et al., 2022
- Walmsley M., et al., 2022, , 509, 3966
- Weir N., Fayyad U. M., Djorgovski S., 1995, *The Astronomical Journal*, 109, 2401
- White R. L., et al., 2000, *The Astrophysical Journal Supplement Series*, 126, 133