Understanding the Universe with AI

Professor Roberto Trotta

23 November 2020

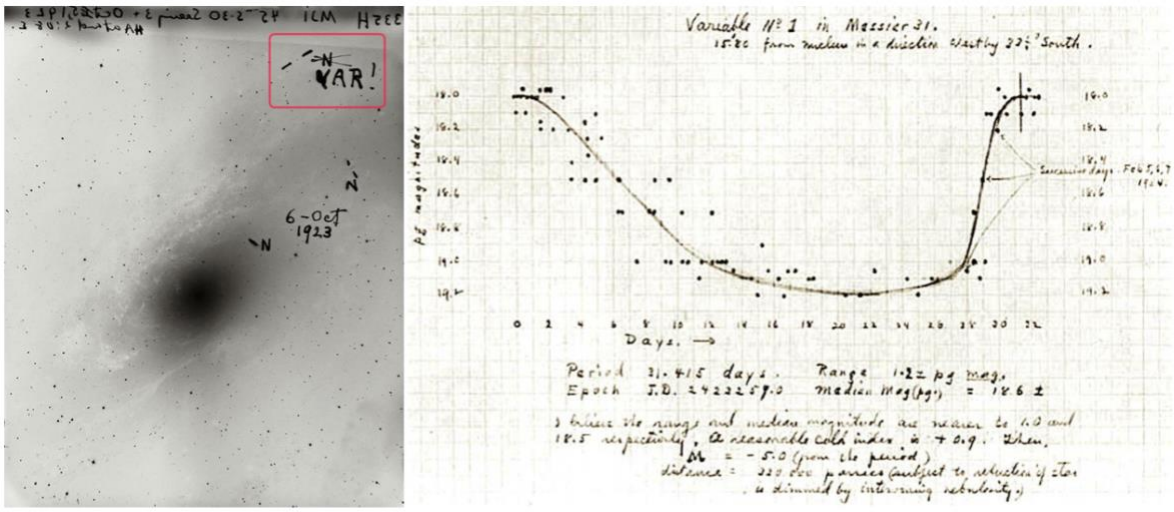
Digital technology from the early 1990s onwards produced an exponential increase in astronomical data. Within our lifetime, the entirety of the visible universe will have been mapped out: we will have seen everything there is to see. The question will then be: what does it all mean?

Solving the mysteries of dark matter and dark energy (which together account for 95% of the universe) and finding life elsewhere in the universe won't be possible without statistical and data analysis methods that have yet to be invented. No human eye will ever inspect all the 50 billion galaxies in the visible universe, nor the 7,500 billion potentially habitable planetary systems: we need machines to do it for us.

Beginnings

On a bitterly cold day, in February 1947, Alan Turing walked under the grand arch leading to the magnificent courtyard of the Royal Academy in Piccadilly, London. The British mathematician was due to give a lecture at the London Mathematical Society, in which he intended to outline his vision of the future of computing machines. At a time when the first electronic calculators filled entire rooms with their oversized valves, mercury lines memories and punchcard readers, Turing imagined the possibility that computers could be programmed "to learn from experience". Rather than merely executing a series of pre-determined instructions, he argued in his lecture, computers should be imparted with the ability to deviate from their programming "if good reason arose", and hence exhibit new behaviours that could surpass the imagination and capability of their masters. To achieve such "intelligent machines", he proposed to mimic the child's brain, by providing computers with an education of punishments and rewards that would, over time, teach them how to achieve the desired outcomes. In so doing, Alan Turing foreshadowed the central idea of many algorithms now at the core of the artificial intelligence revolution, a decade before anybody even used that term.

Around the same time, physical cosmology was making its first, giant strides into the vastness of the universe. In the 1920s, the American astronomer Edwin Hubble had dramatically enlarged our vistas on the universe with two momentous discoveries. With the help of the 100-inch Hooker telescope at Mount Wilson Observatory, at the time the biggest in the world, he had been scrutinising the Andromeda galaxy, our nearest cosmic neighbour. In the early 1920s, nobody knew whether the "Andromeda nebula", as it was then called, was a blob of gas inside our own galaxy, the Milky Way, or whether it was a galaxy in its own right, and thus much further away. This is because measuring distances in the cosmos is far from simple: a star of a given apparent brightness in the sky, for example, may be dim and nearby or bright and far away – we simply cannot tell unless we have an independent mean of knowing its true brightness (which varies greatly among various types of star).



Credit: Carnegie Observatories

Figure 1: Left: A negative image of the Andromeda galaxy taken by Edwin Hubble on Oct 6th 1923, showing in the upper right corner the variable star which he would use to measure the distance to the galaxy. Right: Hubble's observations of the period of variability of the star.

Hubble was building on the trailblazing work of Henrietta Swan Leavitt, whom we would today describe as an astronomer but who at the time could only work as poorly paid woman “computer” at Harvard Observatory. Charged to examine thousands of photographic plates looking for variable stars, Leavitt identified 25 stars of a particular kind whose periodic changes in brightness, she discovered, were related to their luminosity: the longer the period, the brighter the star, a relationship today known as “Leavitt’s law”. With this powerful tool in hand, Hubble looked for the right kind of variable star in Andromeda, with a view of measuring its period and via Leavitt’s law determine the distance of the nebula. On Oct 6th, 1923 he finally succeeded in find one such variable star (Figure 1), and he converted the 31 days period into a measurement of the distance to Andromeda. He concluded that Andromeda was about 1 million km from us, a value that, while 2.5 times smaller than the correct distance, clearly put the nebula well outside our own galaxy: the field of cosmology was thus born.

But Hubble had an even more ambitious aim: by continuing his observational campaign, and building on data collected over the previous decade by Vesto Slipher, he measured the distance to several further galaxies, and discovered a simple linear relation between their speed and their distance from us. Firstly, he established that the vast majority of the galaxies was moving away from us (differently from Andromeda, which is on the contrary moving towards us, leading to a probable merge with the Milky Way in about 5 billion years’ time). Secondly, he found that the more distant galaxies had a greater speed of recession. Taken together, these observations could only be explained by the expansion of the universe itself, which is stretching the fabric of space between us and other galaxies – more so for more distant galaxies, which therefore recede faster. This was exactly what the Belgian priest, mathematician and astronomer Georges Lemaître had found a few years earlier, by analysing Einstein’s equations of General Relativity and applying them to the universe as a whole. In honour of the contributions of both men, the proportionality constant between distance and speed is today called “the Hubble-Lemaître constant”. If the universe is expanding, then it follows that it was smaller in the past, and hence there is a point in time when the whole observable universe was concentrated in a point: the Big Bang. The inverse of the Hubble-Lemaître constant gives an estimate of the time elapsed since the Big Bang, and Hubble’s data implied an age of the universe of 2 billion years – far too young, since the Solar System is about 5 billion years old. We know today that the right value is about 14 billion years, ample time for the Sun to form and life to evolve on Earth.

Three Questions at The Frontier

Almost a century later, the two apparently distinct endeavours of physical cosmology and artificial intelligence have become unexpected allies in our quest to understand the cosmos. Modern research at the frontiers of astrophysics and cosmology produces vast quantities of complex data, from ground-based telescopes, space observatories and even, for example, dark matter detectors in underground laboratories or neutrino telescopes buried in the Antarctic icecap. The sheer volume of data, their intricate interdependencies, and their complex relationship with the physical systems we wish to probe and understand mean that classical statistical and data analysis methods are rapidly becoming obsolete. The bottleneck in our capability to decrypt the mysteries of the universe is already today our ability to process, analyse and interpret the exponentially increasing amount of data from our telescopes and observatories – and artificial intelligence is our only hope.

There are many important open questions at the research frontiers of cosmology and astrophysics. To settle them, we need to gather more data from the cosmos – a programme that will only be successful to the extent that we manage to understand what the data actually mean. To select just a few exemplar topics, let us focus our attention on three of the big open questions of today: the nature of dark matter, the cause of cosmic acceleration, and the existence of life elsewhere in the universe.

Gravitational effects of an otherwise invisible distribution of mass have been observed in the cosmos for decades, in a variety of systems ranging from dwarf galaxies to the largest observable scales in the universe. The evidence for the existence of this “dark matter” has mounted to the point that it is today almost irrefutable: the way galaxies spin, how they form, the interaction of clusters of galaxies, the distribution of imperfections in the leftover light from the Big Bang and many other observations agree in indicating that dark matter makes up about 25% of the universe – 5 times as abundant as normal matter. We remain however in the dark as to its fundamental nature: perhaps a new type of massive particle beyond the Standard Model of particle physics, perhaps an ultra-light field pervading all space, perhaps a consequence of the existence of a 5^{th} dimension, perhaps a fundamental error in the laws of gravity; there are many avenues to explore and they all require more and better data to be followed to their eventual conclusion: either a discovery or a definitive rejection of the idea.

The second big question is to determine what is responsible for the accelerated expansion of the universe. In 1998, two teams of astronomers stunned the world with their announcement that the study of distant stellar explosions had shown that the universe has been picking up speed it its expansion for the past 6 billion years. This cannot be explained if the universe is filled only with matter (both visible and dark) and radiation (light and neutrinos): all of these generate gravity and thus slow the expansion of the universe down. To explain acceleration, we need to invoke dark energy, perhaps a property of empty space itself, which Einstein first envisaged 1917 as a way of exactly balancing the gravitational attraction of matter with an exotic repulsive force and thus leading to a static, non-expanding universe. After Hubble discovered the expansion of the universe, Einstein abandoned the notion. However, present-day observations are consistent with Einstein’s idea of dark energy, but the uncertainty is large and more data are needed to rule out even more outlandish alternatives.

The third looming question is the existence of life elsewhere in the cosmos, something that humans have wondered about from time immemorial. With the discovery of thousands of exoplanets (i.e., planets orbiting other stars than the Sun) in the last decade, we are now poised to make a momentous breakthrough that will change forever our outlook on ourselves and the cosmos. We

have reasons to believe that planets are very common in the galaxy – there are likely more planets than stars in the Milky Way! – and the chance that life has evolved elsewhere is non-negligible. Whether we are able to find it is a question that depends on our technological capability of scanning more and more exoplanetary systems, all the while refining our analysis capabilities to detect the faint whisper of alien life (in the form of chemical signatures in the atmosphere of an exoplanet or even an exomoon) subtly hidden in the astrophysical noise of the cosmos.

The Era of Machine Learning

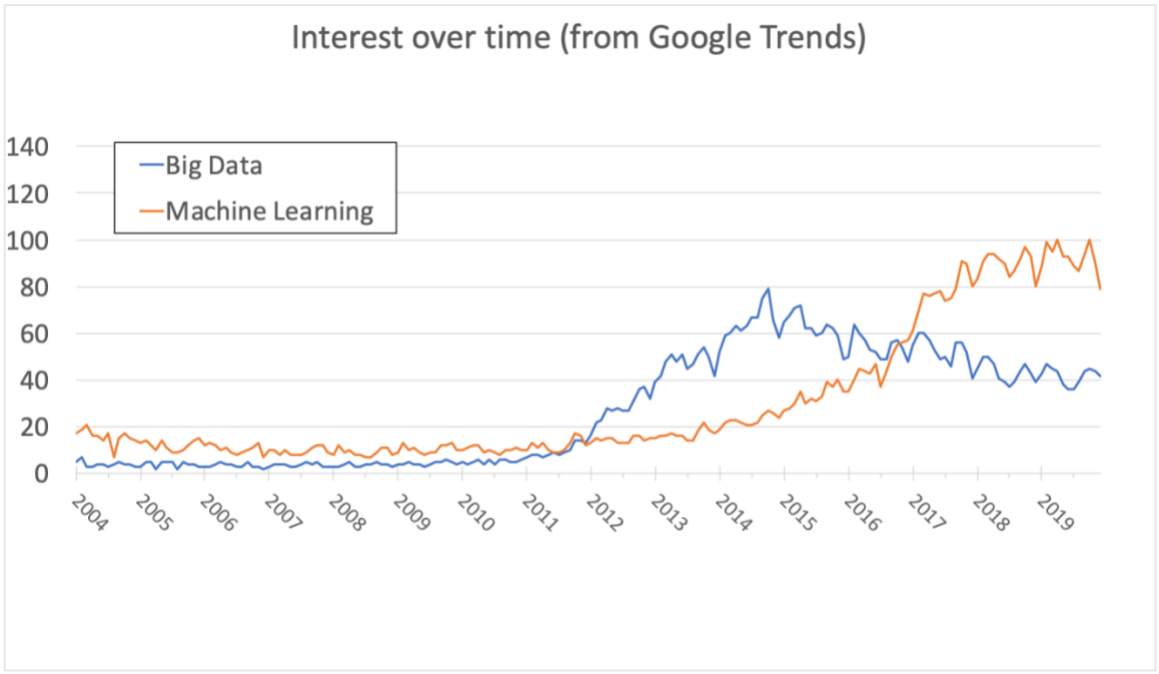


Figure 2: Interest over time in Google searches for the term "Big Data" (blue) vs "Machine Learning" (orange). Machine learning has overtaken big data as the buzzword of the moment.

Thanks to the revolution of digital imaging and robotic telescopes in the 1990s, the amount of data we collect about the universe is growing exponentially. For example, the Vera Rubin Observatory, a new ground-based telescope that will open its 8.4 main mirror to the sky sometimes in 2021-22, will produce 200,000 images per year – no human eye will ever inspect them. We will need intelligent machines to do it for us. The number of stellar explosions used to investigate dark energy will also increase by orders of magnitude, making it impossible to characterize in painstaking detail each of them as was the case until the present day – yet necessary if we want to use them to better understand the puzzling accelerated expansion. Only a machine learning-assisted statistical approach will be feasible. Surveying thousands of stars looking for exoplanets and potential life there will only succeed if we learn to reliably disentangle the faint signals of interest from the omnipresent noise.

This is where machine learning – a branch of artificial intelligence that encapsulates Turing's vision of "machines that can learn from experience" – will be indispensable. Indeed, the era of Big Data that raged at the beginning of the 2010s is rapidly giving way to the era of machine learning (Figure 2), as we realize that vast amount of data are almost useless unless we also command the algorithms that can extract meaning from them.

While it is difficult to define machine learning in a nutshell, we can learn to recognize it at work, as it has rapidly colonized many aspects of our lives, often without us noticing. Machine learning is behind the recommender systems that suggest our next purchase in online store or guess what we

might want to watch on a streaming service; it is the core of autonomous driving systems that one day might drive us around (though many obstacles remain before this can be reliably achieved); it helps doctors diagnose illnesses faster and with more accuracy; it's the heart of the virtual assistants that understand our vocal commands and reply in an almost human fashion. But machine learning has also the potential of being misused, especially when it is deployed surreptitiously and without democratic oversight, individual consent nor accountability. The spread of facial recognition software raises ethical questions about privacy and control, while the correlation of large databases of our abundant digital traces (credit card purchases, social media, smartphone location, fitness trackers, online activities and so on) might lead to the kind of surveillance society that authoritarian regimes could in the past only dream of. Autonomous lethal weapons are being developed that might one day have the capability to attack targets without human control nor oversight, which creates the danger of military escalation and has profound moral implications. It is therefore important to realize that machine learning is a powerful technology that, not unlike atomic energy, can be used both for the benefit of humanity as much as to create a dystopian future. Differently from atomic energy, which can only be controlled and deployed by the organization of powerful nation states, machine learning is extremely nimble and spreads rapidly, often leading to unforeseen applications that go well beyond the original intentions of its creators.

From The Big Bang To AI, and Back

Machine learning and cosmology can help each other in a virtuous circle: on the one hand, machine learning methods are indispensable tools to interpret future large data sets about the cosmos; on the other, cosmological problems of the kind I have described here act as challenging testing grounds to catalyse novel machine learning solutions, which, if successful, can then be deployed with confidence to solve other, more applied problems. The result is a mutually beneficial circle of development leading to better education, advances in research as well as societal applications, for example in the world of business. Let us look at a couple of examples of this virtuous circle in action.

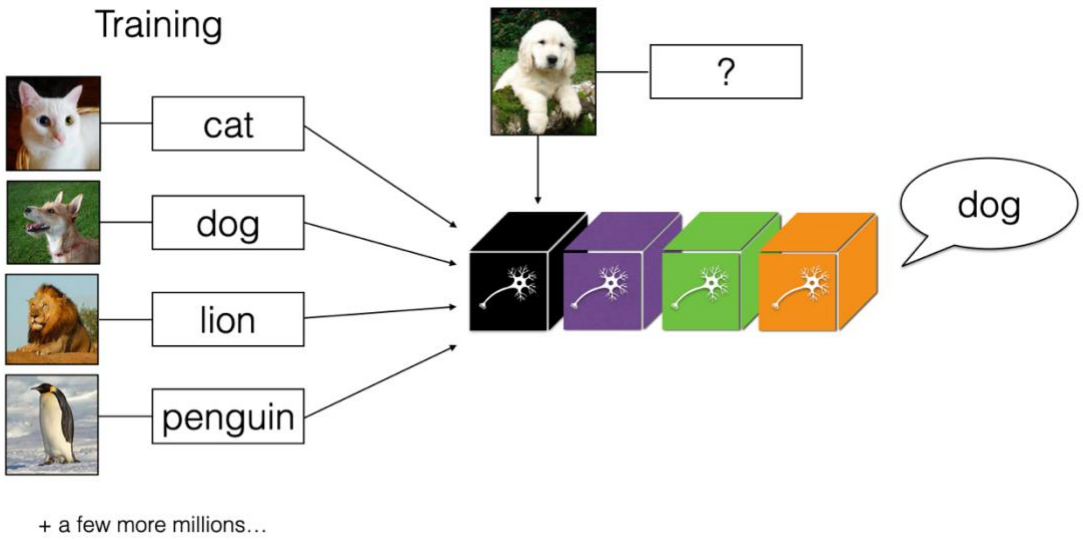


Figure 3: schematic illustration of the training of a machine learning system on example images.

Within machine learning, computer vision has made huge strides forward in the last few years, powered by a class of algorithms known as "deep learning"[1]. To make a concrete example of a popular and common approach, let us consider the task of learning to recognize an animal given a

[1] There are of course many different types of machine learning, some of which can "learn from experience" without the need for a training data set nor human input. This example serves the purpose of making the present discussion more concrete.

picture, as in Figure 3. (While this may appear as a frivolous example, this kind of task is a central pillar for many applications, ranging from self-driving cars that need to recognize their surroundings to breast cancer screening). Animals come in all shape and forms, of course, and pictures can be taken from a variety of angles, light conditions, backgrounds and so on. Following Turing's ideas, we would like an algorithm that is capable of learning from experience, rather than being pre-programmed with characteristics of each animal. For example, specifying that a certain animal is characterized by "four legs + brown fur + hooves" would fail if the picture does not show all of the elements, and of course we would need a very long list of variations to capture the essence of what a human child easily would recognize as a "horse". The machine learning solution is to assemble a large set of sample pictures (often in the millions, or more), each one showing the animal and a label with its name and feed it to the system. As the machine learning system ingests more and more labelled pictures, it adjusts its internal "knobs" to learn from the data themselves what makes a picture of a cat, a cat – without the need for humans to give an explicit description. When presented with a picture of an animal that it has never seen before, the trained machine learning system will output its best guess (in a probabilistic sense) of its name.

Once trained, the same machinery can be instructed to generate new, entirely original pictures of animals, or faces, or landscapes, or whatever objects it has learned to recognize. Since the machine has learnt the characteristics of the objects from real images, its newly generated ones will appear entirely plausible to us. This technology can be deployed very effectively in cosmology, where the machine learning system is trained to learn, for example, the distribution of dark matter and galaxies in the universe. Such distribution is obtained by simulating in a supercomputer the evolution of the dark matter, gas, stars and galaxies in the universe as a function of time, from a few hundred million years after the Big Bang to today. However, this simulation process requires very large computational resources, and can thus only be carried out for a few choices of physical conditions in the universe. In order to explore a large number of possible physical conditions in a much shorter amount of time, cosmologists use the trained machine learning system to create new dark matter simulations at a fraction of the computational cost and time that it would take to carry out the real simulation, yet with almost identical results. Another use of machine learning is in future large surveys of galaxies: with hundreds of thousands of galaxies imaged every year, humans are just not up to the task of inspecting them all and isolating the objects of scientific interest. Machine learning systems can be trained to do just that, and for example are able to group images of galaxies according to their shape, without the need for human input to define the different categories.

Overcoming Machine Learning Bias

But machine learning has got its limitations, too, and they can be crippling, especially when deployed in real-life situations when the health and well-being of people might be at stake. A major problem occurs when the trained system we introduced earlier, which has learnt to recognize images of animals, is presented with a new image of an animal that it has never seen before, or only rarely. Not being sufficiently familiar with this new input, the algorithm will then give a wrong result (for example, by judging a fish to be a lion), without realizing its mistake. If we have come to rely on the accuracy of machine learning conclusions for any application, this kind of undetected errors can have catastrophic consequences. This "bias" of machine learning is being increasingly observed and studied in many real-life applications: from image recognition software classifying images of African-American as "gorillas", to face detection systems exhibiting poor accuracy with dark skinned female faces or skin cancer screening tools that fare much worse for dark skinned patients, racial and gender biases worm their way into supposedly "objective" algorithms by way of their biased or incomplete training data. In other words, far from being fairer and unbiased, often machine learning systems learn and intensify our own biases, thus leading to poorer or outright dangerous outcomes for certain segments of the population, particularly minorities.

Cosmology can, perhaps surprisingly, help with this and other limitations of machine learning, by providing well-defined problems that share this same methodological bias. The observation of stellar explosions used to map out the expansion history of the universe suffer from a source of bias that is similar, from a statistical point of view, to the bias that plagues the above examples. By developing novel solutions in a cosmological setting, which provides a cleaner testbed for ideas and is free from the ethical and moral considerations associated with more applied problems, astrophysicists can help advance machine learning research and improve its capabilities for the benefits of the whole of society.

The importance and impact of machine learning and AI will only grow in the future. Cosmology and astrophysics stand to hugely benefit from these powerful methods, which will enable new discoveries that would otherwise be out of our reach. At the same time, the shortcomings of machine learning can be overcome thanks to novel developments arising from the complex and challenging problems faced by cosmologists, thus ensuring that AI leads to a better society for everyone rather than to a dystopic future where its power is harnessed only to the advantage a few.

Further reading

- Angwin, Julia & Larson, Jeff & Mattu, Surya & Kirchner, Lauren. 2016. *Machine Bias*. ProPublica, May 23, 2016 online.
- Bahcall, N. A. (2015). Hubble's Law and the expanding universe. *Proceedings of the National Academy of Sciences, 112*(11), 3173-3175. doi:10.1073/pnas.1424299112
- Broussard, Meredith, Artificial Unintelligence: How Computers Misunderstand the World, MIT Press (2018)
- Fluke, CJ, Jacobs, C. Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. *WIREs Data Mining Know Discov*. 2020; 10:e1349. https://doi.org/10.1002/widm.1349
- Tasker, E., The Planet Factory: Exoplanets and the Search for a Second Earth, Bloomsbury Sigma (2017).
- Turin, Alan M., 1947, 'Lecture to the London Mathematical Society on 20 February 1947', in: The Collected Works of A.M. Turing. Volume 1: Mechanical Intelligence, D.C. Ince (ed.), Amsterdam: North-Holland, 1992.
- Turing, Alan M., 1950. Computing Machinery and Intelligence. Mind, 1950. LIX(236): p. 433-460.