mammograms to draw the interpreter's attention to areas that might be abnormal. However, analysis of a large sample of clinical mammography interpretations from the US Breast Cancer Surveillance Consortium registry demonstrated that there was no improvement in diagnostic accuracy with CAD[3]. Moreover, that study revealed that the addition of CAD worsened sensitivity (the performance of radiologists in determining that cancer was present), thus increasing the likelihood of a false negative test. CAD did not result in a significant change in specificity (the performance of radiologists in determining that cancer was not present) and the likelihood of a false positive test[3].

It has been speculated that CAD was not as useful in the clinic as experimental data suggested it might be because radiologists ignored or misused its input owing to the high frequency of marks on the images that were not findings suggestive of cancer. This outcome was attributed by some to the limited processing power available for CAD, which meant that comparisons with previous imaging studies of the same person were not possible[4]. Thus, CAD might mark regions that were not changing over time and that could be easily dismissed by expert readers. Another factor that limited CAD is that it was developed using the performance of human-based diagnosis. It was trained using mammograms in which humans had found signs of cancer and others that were false negatives – cases in which humans could not see signs of cancer although the disease was indeed present[4]. Similar pitfalls could be encountered with AI-based decision aids, too.

A system by which AI finds abnormalities that humans miss will require radiologists to adapt to the use of these types of tool. Imagine a system in which an algorithm marks a dense breast area on a screening mammogram and the human radiologist cannot see anything that looks potentially malignant. With CAD, radiologists scrutinize the areas marked, and if they decide the mark is probably not cancer, they assign the mammogram as being negative for malignancy. However, if AI algorithms are to make a bigger difference than CAD in detecting cancers that are currently missed, an abnormality detected by the AI system, but not perceived as such by the radiologist, would probably require extra investigation. This might result in a rise in the number of people who receive callbacks for further evaluation. A clinical trial would show the effect of the AI system on the detection of cancer and the rate of false positive diagnoses, while also allowing the development of effective clinical practice in response to mammograms flagged as abnormal by AI but not by the radiologist.

In addition, it would be essential to develop a mechanism for monitoring the performance of the AI system as it learns from cases it encounters, as occurs in machine-learning algorithms. Such performance metrics would need to be available to those using these tools, in case performance deteriorates over time.

It is sobering to consider the sheer volume of data needed to develop and test AI algorithms for clinical tasks. Breast cancer screening is perhaps an ideal application for AI in medical imaging because large curated data sets suitable for algorithm training and testing are already available, and information for validating straightforward clinical end points is readily obtainable. Breast cancer screening programmes routinely measure their diagnostic performance – whether cancer is correctly detected (a true positive) or missed (a false negative). Some areas found on mammograms might be identified as abnormal but turn out on further testing not to be cancerous (false positives). For most women, screening identifies no abnormalities, and when there is still no evidence of cancer one year later, this is classified as a true negative.

Most other medical tasks have more-complicated clinical outcomes, however, in which the clinician's decision is not a binary one (between the presence or absence of cancer), and thus further signs and symptoms must also be considered. In addition, most diseases lack readily accessible, validated data sets in which the 'truth' is defined relatively easily. Obtaining validated data sets for more-complex clinical problems will require greater effort by readers and the development of tools that can interrogate electronic health records to identify and annotate cases representing specific diagnoses.

To achieve the promise of AI in health care that is implied by McKinney and colleagues' study, anonymized data in health records might thus have to be treated as precious resources of potential benefit to human health, in much the same way as public utilities such as drinking water are currently treated. Clearly, however, if such AI systems are to be developed and used widely, attention must be paid to patient privacy, and to how data are stored and used, by whom, and with what type of oversight.

Etta D. Pisano is at the American College of Radiology, Philadelphia, Pennsylvania 19103, USA, and at Beth Israel Lahey Medical Center, Harvard Medical School, Boston, Massachusetts.
e-mail: episano@bidmc.harvard.edu

1. McKinney, S. M. *et al. Nature* **577**, 89–94 (2020).
2. Neri, E. *et al. Insights Imaging* **10**, 44 (2019).
3. Lehman, C. D. *et al. JAMA Intern. Med.* **175**, 1828–1837 (2015).
4. Kohli, A. & Jha, S. *J. Am. Coll. Radiol.* **15**, 535–537 (2018).

Astronomy

Galaxy cluster illuminates the cosmic dark ages

Nina A. Hatch

Observations of a distant cluster of galaxies suggest that star formation began there only 370 million years after the Big Bang. The results provide key details about where and when the first stars and galaxies emerged in the Universe. **See p.39**

Shortly after the Big Bang, the Universe was completely dark. Stars and galaxies, which provide the Universe with light, had not yet formed, and the Universe consisted of a primordial soup of neutral hydrogen and helium atoms and invisible 'dark matter'. During these cosmic dark ages, which lasted for several hundred million years, the first stars and galaxies emerged. Unfortunately, observations of this era are challenging because dark-age galaxies are exceptionally faint[1]. On page 39, Willis *et al.*[2] provide a glimpse of what happened during the dark ages by doing some galactic archaeology. By measuring the ages of stars in one of the most distant clusters of galaxies known, the authors located galaxies that formed stars in the dark ages, close to the earliest possible time that stars could emerge.

A galaxy cluster is a group of thousands of galaxies that orbit each other at speeds[3] of about 1,000 kilometres per second. They are prevented from flying apart by the gravitational pull of the accompanying dark matter, which has the equivalent total mass of about one hundred trillion Suns[4]. Astronomers use these clusters as laboratories for many experiments in astrophysics, such as measuring the composition of the Universe, testing theories of gravity and determining how galaxies form. Willis *et al.* used one of the

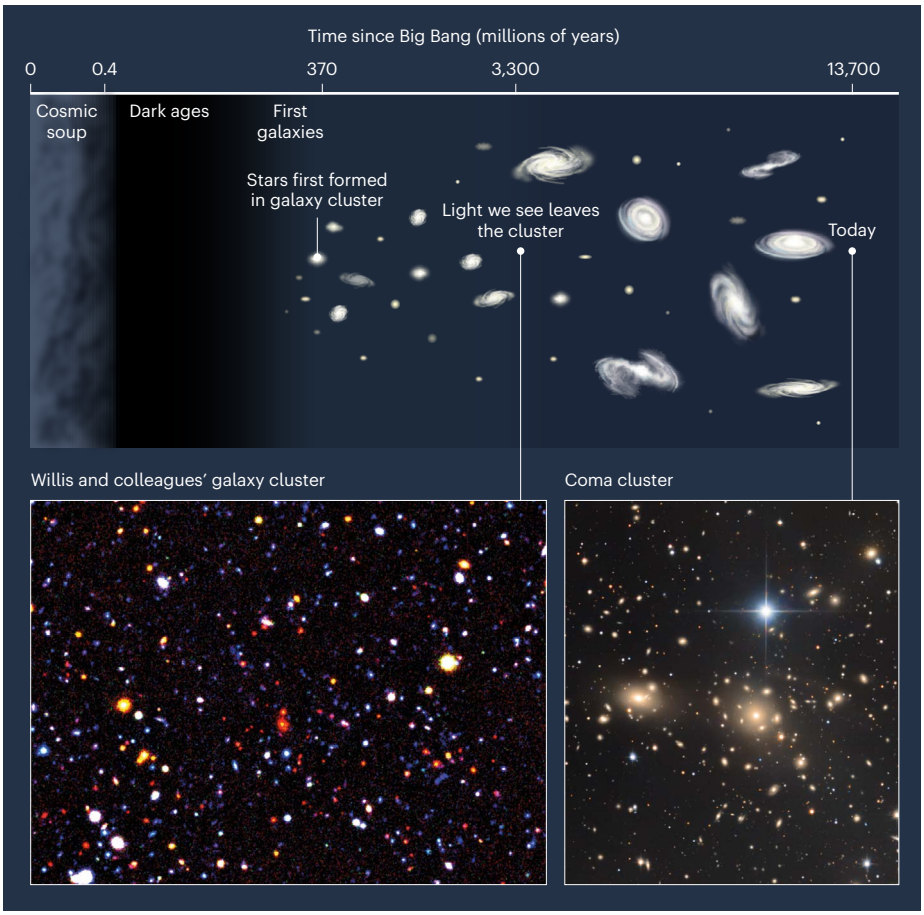

Figure 1 | Chronology of the Universe. After the Big Bang, the Universe consisted of a cosmic soup of radiation and matter. About 400,000 years later, it entered an era known as the cosmic dark ages in which it was devoid of light. The first stars and galaxies began to emerge a few hundred million years later, and gradually provided the Universe with light. Willis *et al.*[2] report that star formation in a distant cluster of galaxies began roughly 370 million years after the Big Bang. The light that we see from this galaxy cluster was emitted when the Universe was about 3.3 billion years old. The cluster is likely to have become one of the largest structures in the present-day Universe, comparable in mass to the Coma cluster. (Image credits: Willis and colleagues' galaxy cluster: N. A. Hatch; Coma cluster: Russ Carroll, Rob Gendler, Bob Franke/Dan Zowada Memorial Observatory, Wayne State Univ.)

most distant clusters known to study when the most massive galaxies in the Universe began to produce stars.

Although nearby clusters, such as the Coma cluster, are easier to observe than those farther away, we cannot measure their ages precisely because the galaxies are extremely old. It is difficult to differentiate between, for example, a galaxy that is 7 billion years old and one that is 13 billion years old[5]. Therefore, to obtain a precise date for when clusters first formed their stars, Willis and colleagues used NASA's Hubble Space Telescope to look at one of the most distant clusters they could find.

Because light travels at a finite speed, the most distant clusters we can see are also those in the earliest stages of the Universe that we can see. The light from the cluster examined by Willis *et al.* has been travelling for 10.4 billion years before it reaches Earth, which means that we are looking at a cluster as it was just 3.3 billion years after the Big Bang. Consequently, this cluster acts as a keyhole through which we can peer into the early Universe (Fig. 1).

Willis and colleagues found that the cluster contains several galaxies that have similar red colours. The colour of a galaxy can be used to estimate its age because younger stars are bluer than their older, redder counterparts. As a result, galaxies that have red colours formed their stars a long time ago[5]. By comparing the colours of the cluster galaxies with those of models, the authors estimated that the stars of these galaxies started to emerge when the Universe was only 370 million years old. This epoch is when we expect the first stars to have formed in the cosmic dark ages[6].

"The galaxy cluster acts as a keyhole through which we can peer into the early Universe."

One particularly intriguing point is that Willis *et al.* identified at least 19 galaxies in the cluster that have similar colours, which means that the galaxies have similar ages. At the time when these galaxies formed their stars, they would have been well spread out, so it is a conundrum as to why they all began producing stars at approximately the same time. Were they influenced by their environment? Alternatively, did the star formation in one galaxy somehow trigger a chain reaction, leading to star formation in nearby gas clouds? We do not currently have the answer, but what is clear from the authors' work is that these distant clusters are full of the oldest galaxies in the Universe.

In my opinion, Willis and colleagues' age estimates are the best ones possible, given the limited data that the authors have from the Hubble telescope. However, determining ages from the colours of galaxies is a relatively crude method that is subject to large uncertainties. For example, a young galaxy that contains a lot of astronomical dust can have the same colour as an old galaxy containing little dust. Therefore, although the authors' results are tantalizing, they should be treated with caution until NASA's James Webb Space Telescope (JWST) is launched in the next few years.

The JWST will measure spectra of the light emitted by these galaxies. A comparison of the spectra with models will be a much more accurate way to determine the ages of the stars than using the colours of galaxies. Furthermore, because it is easier to measure the ages of earlier galaxies than those of more recent ones[5], it makes sense to target galaxies in the progenitors of these galaxy clusters in the early Universe. Willis and colleagues' results make a strong case for these distant clusters being some of the first targets that the JWST should observe.

Nina A. Hatch is in the School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK.
e-mail: nina.hatch@nottingham.ac.uk

1. Stark, D. P. *Annu. Rev. Astron. Astrophys.* **54**, 761–803 (2016).
2. Willis, J. P. *et al. Nature* **577**, 39–41 (2020).
3. Struble, M. F. & Rood, H. J. *Astrophys. J. Suppl. Ser.* **125**, 35–71 (1999).
4. Bahcall, N. A. *Annu. Rev. Astron. Astrophys.* **15**, 505–540 (1977).
5. Bruzual, G. & Charlot, S. *Mon. Not. R. Astron. Soc.* **344**, 1000–1028 (2003).
6. Planck Collaboration. *Astron. Astrophys.* **596**, A108 (2016).