BBC

COSMIC FACTORIES: HOW STARS FORGE MATTER

#226 MARCH 2024

Sky at Night

THE UK'S BEST-SELLING ASTRONOMY MAGAZINE

THE YEAR OF THE AURORA

Discover why *NOW* is the time to see nature's greatest spectacle

2 COMETS IN 1 MONTH!

Track the progress of Pons–Brooks & PanSTARRS

AI: ESSENTIAL TO THE FUTURE OF ASTRONOMY?

THE COMMUNITIES WHO'VE PUT LIGHT POLLUTION IN THE PAST

MYSTERIES AT THE EDGE OF THE SOLAR SYSTEM

VENUS'S ACID CLOUDS COULD HARBOUR LIFE

THE RADIO TELESCOPE THAT SPANS 2 CONTINENTS

Big data at the dawn of artificial intelligence

As ever more ambitious space surveys begin to create unprecedented mountains of data, **Paul Fisher Cockburn** asks if the future of astronomy will be found in AI

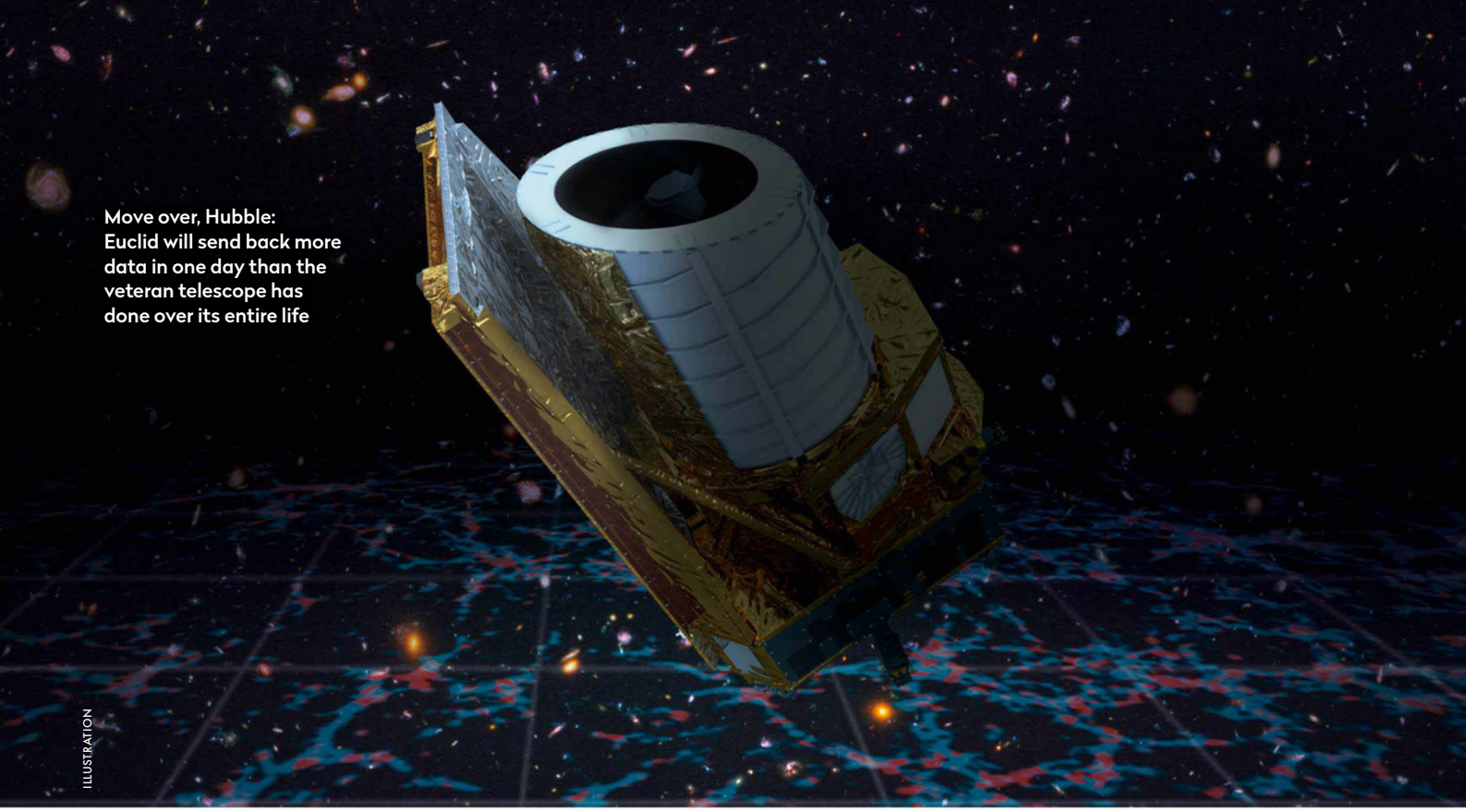
As the late Douglas Adams wrote in *The Hitchhiker's Guide to the Galaxy*, "Space is big. Really big. You just won't believe how vastly, hugely, mindbogglingly big it is."

And, to be fair, he wasn't wrong.

For astronomers – especially those focusing on cosmology – this has one obvious consequence. The more detailed and accurate their studies of such a "really big" cosmos become, the larger the amount of data they are likely to generate – and have to process. Arguably this has been a looming problem ever since astronomers first started sticking cameras onto their telescopes, but the latest digital technologies have pushed the issue to the foreground like nothing else before. ▶

ILLUSTRATION: MONSITJ/ISTOCK/GETTY IMAGES

Too much information? With a gargantuan flood of space data on its way, scientists face the Herculean task of analysing it all



Move over, Hubble: Euclid will send back more data in one day than the veteran telescope has done over its entire life

ILLUSTRATION

▶ Take the Square Kilometre Array, featured earlier in this month's issue. Once it's fully operational, it will create terabytes of data *every second*.

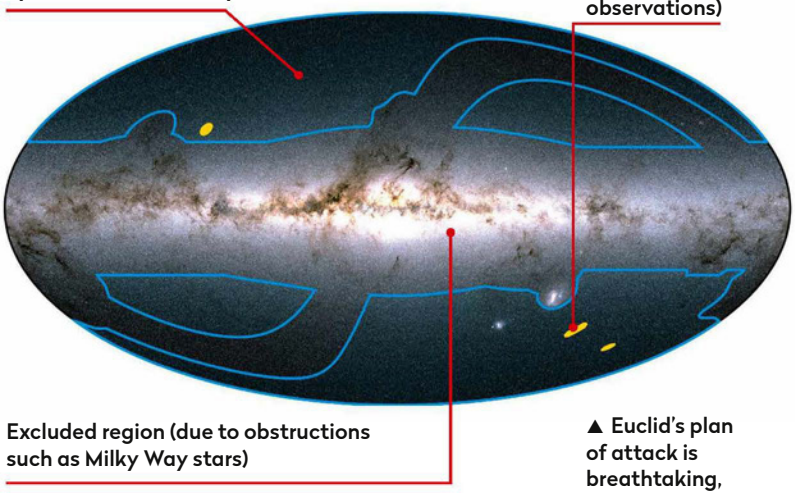
The latest European Space Agency (ESA) mission, the Euclid space telescope (launched on 1 June 2023), is another prime example. Its mission is essentially an attempt to measure the geometry of the entire Universe, improving our understanding of dark matter and dark energy. It requires the incredibly precise observation of billions of stars and galaxies. The amount of data that the mission will generate during its official six-year mission is almost inconceivable.

"What I think is interesting is that no human will look at all the Euclid data. It'll be too big, it'll never happen," explains Andrew N Taylor, professor of astrophysics at the University of Edinburgh. "There's a good chance that if anyone picks on a random piece of sky, no other human will have ever looked at that bit of sky in such detail before."

Information overload

Andrew has been part of the ESA's Euclid Consortium, which runs the space telescope, for almost 20 years, helping devise the initial concept, the design of the mission and its science goals. During the last 10 years, as the telescope and its various optics and detectors were designed, built and tested before launch, his focus – along with many others' – shifted towards data analysis.

"Euclid produces an enormous amount of data," he says. "We're downloading hundreds of gigabytes of data per day. Just to give you an idea of the volume, a good analogy is what is seen from the Hubble Space Telescope. In terms of image quality, a single image from Euclid and an image from Hubble are very similar, but if you put together all of the pictures Hubble's taken during its lifetime and laid them on the sky, it would cover an area only about 20 times the size of the full Moon. Euclid can do the equivalent of what Hubble has done, in its lifetime, in a single day. In fact *more*, because it takes not just optical images like Hubble, it also takes ones in the infrared and spectra of the galaxies and objects out there. It's just a huge step up for astronomy in the amount of information that we're going to get."

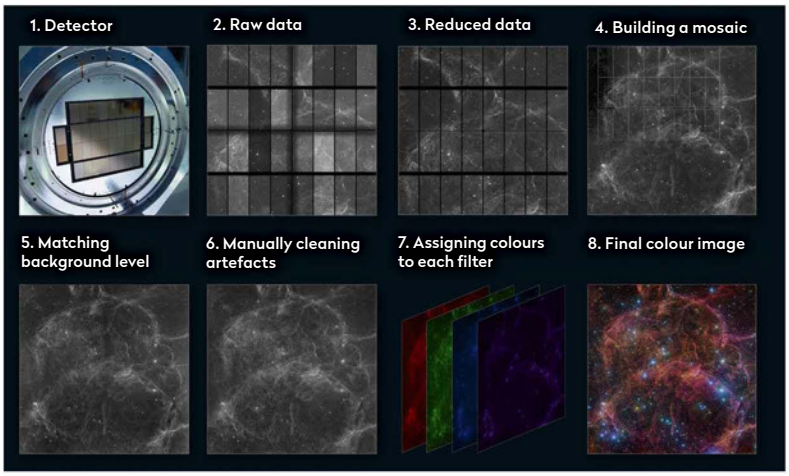


▲ **Euclid's plan of attack is breathtaking, surveying more than one-third of the sky**

This isn't just because the multinational Euclid team wanted to go out and collect lots of data. "The scientific goal is to try to understand the nature of dark matter and the dark energy Universe, and in order to test our theories we know we've got to look for very subtle little differences in things like the distribution of matter and its evolution in the Universe," he says. "It's the classic large data problem: we've got a very small signal, so we need a huge dataset to try to get the levels of precision we need to test our theories."

In order to process the data, the Euclid Consortium has developed both bespoke computer algorithms, to assist with data compression, and an IT infrastructure

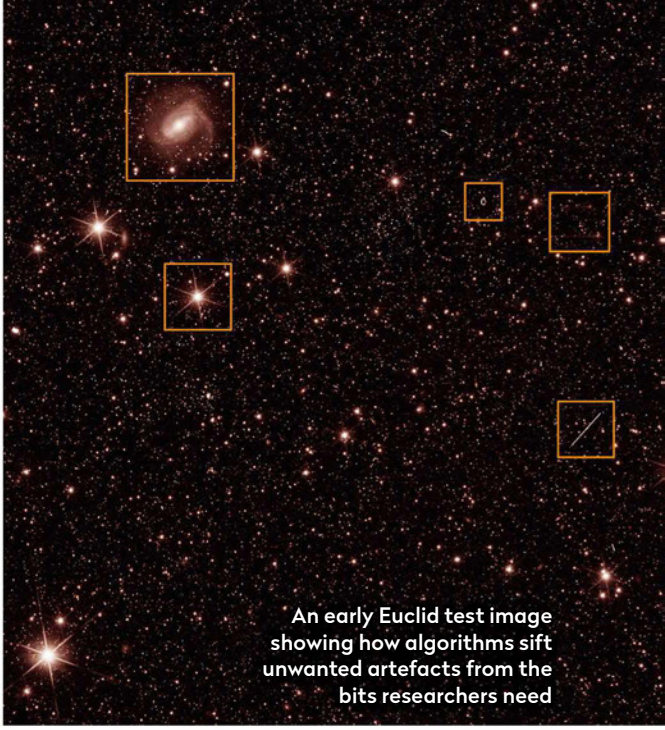
ESA/EUCLID/EUCLID CONSORTIUM/NASA. BACKGROUND GALAXIES: NASA/ESA AND S. BECKWITH (STSCI) AND THE HUDF TEAM. ESA, ESO/M KORNMESSER/VPHAS+ TEAM. ACKNOWLEDGEMENT: CAMBRIDGE ASTRONOMICAL SURVEY UNIT, EUCLID CONSORTIUM



▲ **Stages in a data 'pipeline' that wrangles unmanageably large raw data into images astronomers can interpret**

'pipeline' that takes the raw data and turns it into something astronomers can actually use and interpret. The techniques used are all based on the same traditional methods used by astronomers for hundreds of years, but the scale of the work requires a novel approach. Although all the data from Euclid will initially be received at the European Space Astronomy Centre near Madrid, that won't be where it's processed.

Work on the Euclid data will be carried out in nine science data centres across Europe, the UK centre being at the Royal Observatory Edinburgh. "The original concept," says Taylor, "was that each of the countries would take responsibility for one area of processing, and that the data would move around the countries. Pretty early on, we realised this was not feasible; the solution we came up with was that we'll split the sky up into ninths. Every data centre is doing all of the processing for their own patch of the sky."



An early Euclid test image showing how algorithms sift unwanted artefacts from the bits researchers need

Break it down

That processing will entail transforming the individual images into 'catalogues'. Algorithms will detect particular features the Euclid team are interested in, such as stars and galaxies, and then create lists of their positions, sizes and other properties. It will only be at that point, when the data is of a more manageable size, that each of the data centres will send all their data to the main centre for final compiling and processing.

This is not without its challenges. "One of the things we've had to deal with," explains Taylor, ▶

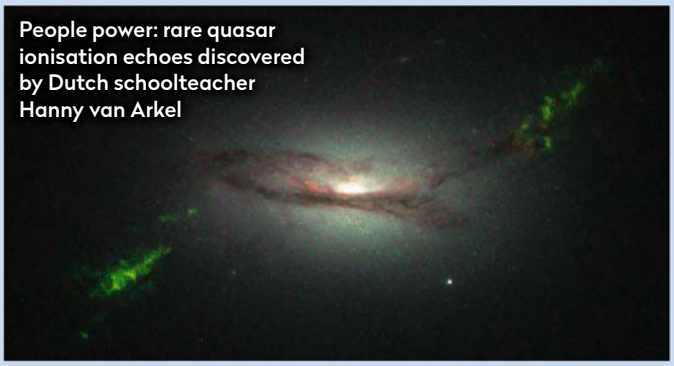
Finding the 'weird'

When it comes to spotting strange things in the data, humans are far better than machines

Can machine learning and AI help detect the unexpected in astronomical datasets? Alex Andersson, a researcher at the University of Oxford, believes it might.

"You can look at certain models of physics, like how stars work or galaxies behave, and so have things that you can predict but haven't seen yet – the sort of 'known unknowns'," he says. "And then there's the 'unknown unknowns'. I spend a lot of time working on anomaly-detection algorithms, where I try not to assume much about the physics or what I'm expecting, and just see what the data says and what comes out that looks 'weird'."

All Andersson's work involves MeerKAT in South Africa, the radio telescope that will form the core of the Square Kilometre Array (SKA). Alongside this, he runs a citizen science project called 'Burst from Space MeerKAT' on the Zooniverse platform, asking hundreds of online volunteers to search the telescope's data for unusual features. Similar citizen science projects have helped other astronomical surveys wade through vast amounts of data. Repeatedly, these volunteers have been fantastic at picking out the weird, out-of-place things in images that can lead to new discoveries. One of the earliest examples is from Zooniverse project Galaxy Zoo, where helpers found strange green blobs that turned out to be compact galaxies.



People power: rare quasar ionisation echoes discovered by Dutch schoolteacher Hanny van Arkel

But the data from Euclid and SKA will be too huge even for teams of citizen scientists to handle. Could AI be trained to search for those unknown unknowns? Andersson is testing just such anomaly-detecting algorithms on the same data his citizen scientists are trawling through, seeing if it will uncover those same 'weird' objects. His goal, though, is not to replace the citizen scientists, but rather to use AI to reduce the data to a more manageable size for them to take on.

"I personally think that discovery in the Universe is a uniquely human experience, so I don't think that machine learning will replace us in that regard," he says.

Artificial intelligence

AI trains computers to think like a human

Arguably, 2023 was the year that artificial intelligence (AI) really gained mainstream attention, inspiring – like any technological advance throughout human history – both evangelical enthusiasm for its potential benefits and apocalyptic horror at its likely downsides.

Yet AI has been around for much longer than the likes of ChatGPT. The term itself was first coined back in 1956, and scientists have been working on ways for computers to 'think' and respond to us in a more 'human' way since at least the 1960s. Without us even noticing, much of our modern life today relies on AI,

whether it's your social media activity, Netflix viewing choices or digital home assistants like Alexa and Siri.

It's all about making our technology appear to act in a more responsive way, a principle that can be employed in several different ways. One of the biggest advancements has been in the subfield of machine learning, where increasingly sophisticated software algorithms enable machines to remember their mistakes and learn from them, instead of just repeating whatever tasks they've been programmed to carry out.

Many of the concerns around the technology are based on a computer's ability to carry out many tasks – especially computations – far faster than humans. With astronomers facing ever larger datasets, however, it's likely machine learning will become a commonly-used tool.

▶ "is how do you make everything the same when the underlying computer infrastructure might not be. So there's a lot of what we call virtualisation, which is that we try to emulate the same computer everywhere."

Rise of the machines

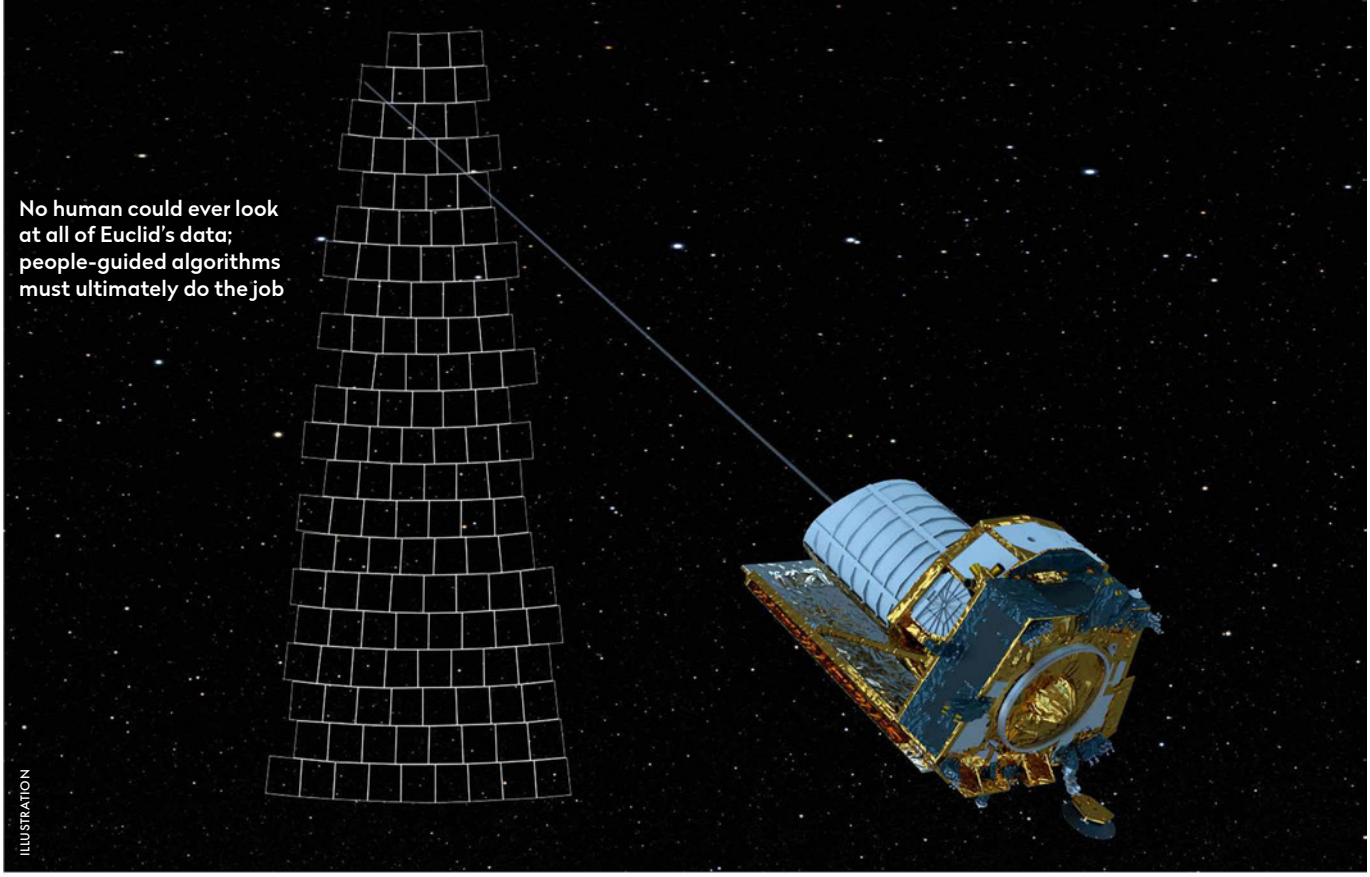
While this 'distributed' model is in part inevitable within a multinational organisation like ESA, it means the Euclid team have unexpectedly become pioneers in how to carry out massive computing projects on distributed networks, and have already been approached by large companies interested in how they did it. It's not just astronomers that have to process big datasets, after all.

With all this talk of IT systems and algorithms, how important are people when it comes to the data processing?

"We're certainly entering an age where it seems computers can do more and more, and take over more and more things," says Taylor. "I think [Euclid] was at a lucky phase, in the sense that we've been working on this for nearly 20 years. The real algorithm-writing started only about 10 years ago."

▲ Training at Goddard Space Flight Center as NASA races to embrace the challenges and opportunities of the new technology

▲ Workshops on AI and machine learning have been showing how automation will revolutionise data science across the agency



No human could ever look at all of Euclid's data; people-guided algorithms must ultimately do the job

ILLUSTRATION

"Even though we've now got a pipeline and analysis in place, the algorithm development was all done by people, using their knowledge and understanding"

There are also, however, some newer methods being used on Euclid data, such as artificial intelligence (AI), the technique of using computer programs that mimic how humans think. One particular subset of AI used in astronomy is 'machine learning', where programs learn from experience as they process the data, improving their analysis over time.

Currently, Euclid uses these newer tools in a limited way, such as helping to classify galaxies. "The tendency was to go with things that were understood, and had been shown to be reliable on previous projects," says Taylor. "A lot of things are being done more traditionally – for 'robustness', for reliability."

Keeping people in the picture

At the moment, astronomers are only just receiving the data and so are still in the early stages of processing it. As they proceed to more advanced stages of analysis, they will eventually start wanting to compare the data with what their theories predict – one of the main methods astronomers use to test their ideas about how the Universe works.

"One of the issues there, is how you generate the theoretical models," explains Taylor. "For the detailed analysis, we do want to have simulations of the Universe based on different theories of what's going on. It's a very expensive thing to run a whole simulation, so there's a lot of work going into trying to understand if we can use artificial intelligence or machine learning to provide shortcuts – not to lose precision or accuracy, but to provide rapid ways of doing it."

Ultimately, Taylor accepts that the volume of data being produced by Euclid will encourage the increased usage of machine learning, the automation of the Euclid data processing. But he sees astronomers – people – very much still at the centre of what's going on.

"At the moment, even though we've now got a pipeline and analysis in place that does this, the algorithm development was all done by people, using their knowledge and understanding of the problems to process these sorts of datasets," says Taylor.

"It's not just a case of 'Let's have some AI look at the images and it'll tell us the answers', because there are lots of details in the images that you need to understand."

One of the characteristics of AI programs is that they need data in order to learn. At the moment, AI doesn't know what Euclid data looks like, as the spacecraft has only just begun its mission to survey the sky. Even once AI does begin to process the data, it will still require human input to understand what it's looking at and to begin teaching it its first lessons.

"There's a huge amount of detail to go through and really understand," comments Taylor. "And it's people who bring that knowledge."

The issues around handling Big Data are only going to grow over time. It's only via a team effort – using both processing tools such as AI and teams of humans spread across the world – that astronomers will be able to work together towards a better understanding of the Universe around us.

Paul Fisher Cockburn is a science and astronomy writer

ANTONIO_DIAZ/ISTOCK/GETTY IMAGES, SEAN KEEFE/NASA GODDARD X 2, ESA