Astronomy generates mountains of data—that's perfect for AI

May 30 2024, by Evan Gough

A drone's view of the Rubin Observatory under construction in 2023. The 8.4-meter telescope is getting closer to completion and first light in 2025. The telescope will create a vast amount of data that will require special resources to manage, including AI. Credit: Rubin Observatory/NSF/AURA/A. Pizarro D

Consumer-grade AI is finding its way into people's daily lives with its ability to generate text and images and automate tasks. But astronomers need much more powerful, specialized AI. The vast amounts of

observational data generated by modern telescopes and observatories defies astronomers' efforts to extract all of its meaning.

A team of scientists is developing a new AI for astronomical data called AstroPT. They've presented it in a new paper titled "AstroPT: Scaling Large Observation Models for Astronomy." The paper is available on the *arXiv* preprint server, and the lead author is Michael J. Smith, a data scientist and astronomer from Aspia Space.

Astronomers are facing a growing deluge of data, which will expand enormously when the Vera Rubin Observatory (VRO) comes online in 2025. The VRO has the world's largest camera, and each of its images could fill 1,500 large-screen TVs. During its 10-year mission, the VRO will generate about 0.5 exabytes of data, which is about 50,000 times more data than is contained in the U.S.'s Library of Congress.

Other telescopes with enormous mirrors are also approaching first light. The Giant Magellan Telescope, the Thirty Meter Telescope, and the European Extremely Large Telescope combined will generate an overwhelming amount of data.

The VRO's need for multiple sites to handle all of its data is a testament to the enormous volume of data it will generate. Without effective AI, that data will be stuck in a bottleneck. Credit: NOIRLab

Having data that can't be processed is the same as not having the data at all. It's basically inert and has no meaning until it's processed somehow. "When you have too much data, and you don't have the technology to process it, it's like having no data," said Cecilia Garraffo, a computational astrophysicist at the Harvard-Smithsonian Center for Astrophysics.

This is where AstroPT comes in.

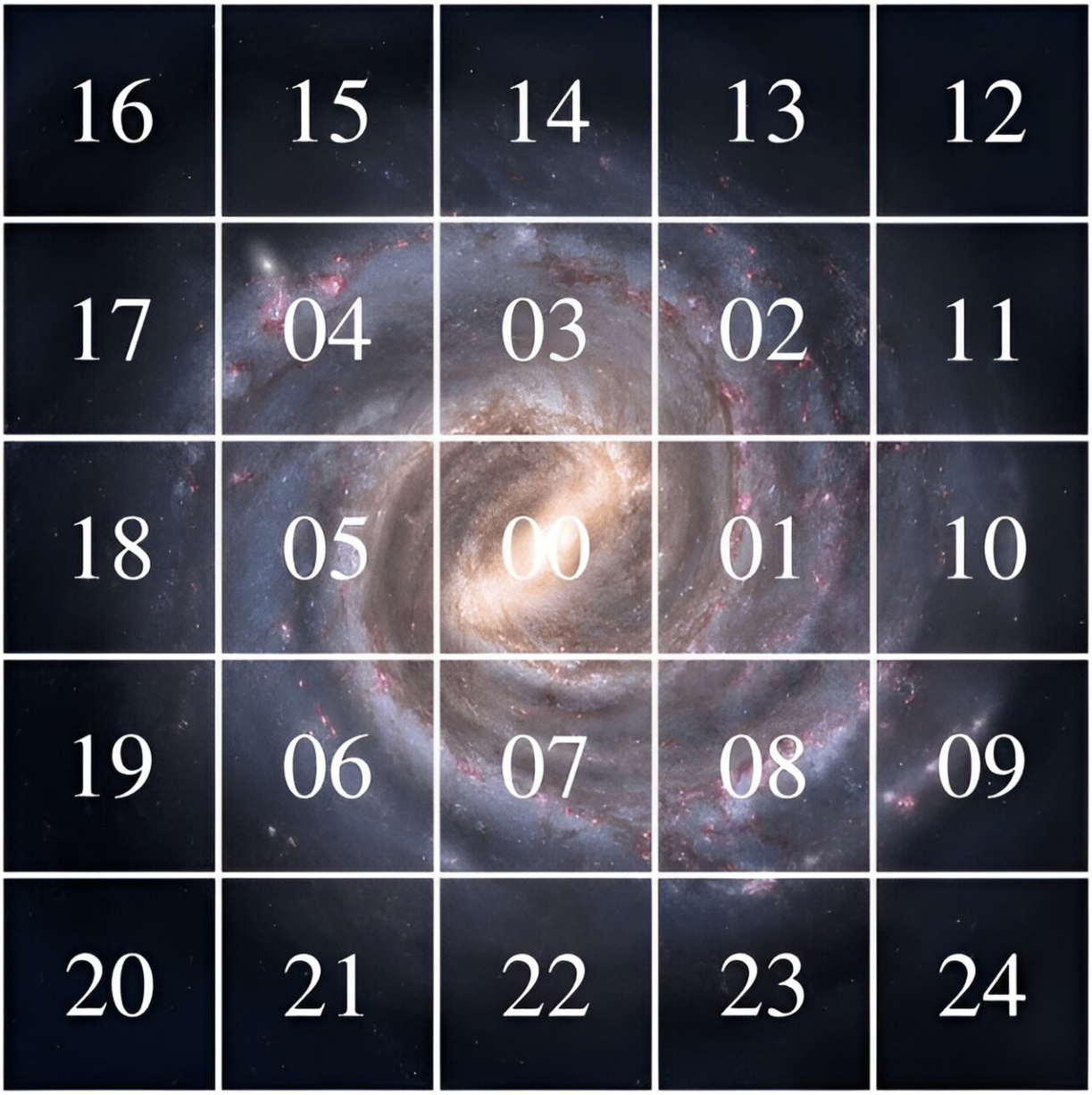
AstroPT stands for Astro Pretrained Transformer, where a transformer is a particular type of AI. Transformers can change or transform an input sequence into an output sequence. AI needs to be trained, and AstroPT

has been trained on 8.6 million 512 x 512-pixel images from the DESI Legacy Survey Data Release 8. DESI is the Dark Energy Spectroscopic Instrument. DESI studies the effect of Dark Energy by capturing the optical spectra from tens of millions of galaxies and quasars.

AstroPT and similar AI deal with "tokens." Tokens are visual elements in a larger image that contain meaning. By breaking images down into tokens, an AI can understand the larger meaning of an image. AstroPT can transform individual tokens into coherent output.

AstroPT has been trained on visual tokens. The idea is to teach the AI to predict the next token. The more thoroughly it's been trained to do that, the better it will perform.

"We demonstrated that simple generative autoregressive models can learn scientifically useful information when pre-trained on the surrogate task of predicting the next 16 × 16 pixel patch in a sequence of galaxy image patches," the authors write. In this scheme, each image patch is a token.



This image illustrates how the authors trained AstroPT to predict the next token in a 'spiralized' sequence of galaxy image patches. It shows the token feed order. "As the galaxies are in the center of each postage stamp, this set up allows us to seamlessly pretrain and run inference on differently sized galaxy postage stamps," the authors explain. Credit: Smith et al, 2024

One of the obstacles to training AI like AstroPT concerns what AI scientists call the "token crisis." To be effective, AI needs to be trained on a large number of quality tokens. In a 2023 paper, a separate team of researchers explained that a lack of tokens can limit the effectiveness of some AI, such as LLMs or Large Language Models. "State-of-the-art LLMs require vast amounts of internet-scale text data for pre-training," they wrote. "Unfortunately, … the growth rate of high-quality text data on the internet is much slower than the growth rate of data required by LLMs."

AstroPT faces the same problem: a dearth of quality tokens to train on. Like other AI, it uses LOMs or Large Observation Models. The team says their results so far suggest that AstroPT can solve the token crisis by using data from observations. "This is a promising result that suggests that data taken from the observational sciences would complement data from other domains when used to pre-train a single multimodal LOM, and so points towards the use of observational data as one solution to the 'token crisis.'"

AI developers are eager to find solutions to the token crisis and other AI challenges.

Without better AI, a data processing bottleneck will prevent astronomers and astrophysicists from making discoveries from the vast quantities of data that will soon arrive. Can AstroPT help?

The authors are hoping that it can, but it needs much more development. They say they're open to collaborating with others to strengthen AstroPT. To aid that, they followed "current leading community models" as closely as possible. They call it an "open to all project."

"We took these decisions in the belief that collaborative community development paves the fastest route towards realizing an open source

web-scale large observation model," they write.

"We warmly invite potential collaborators to join us," they conclude.

It'll be interesting to see how AI developers will keep up with the vast amount of astronomical data coming our way.

More information: Michael J. Smith et al, AstroPT: Scaling Large Observation Models for Astronomy, *arXiv* (2024). DOI: 10.48550/arxiv.2405.14930

Provided by Universe Today