Large-scale data fusion in astronomy

openaccessgovernment.org/article/large-scale-data-fusion-in-astronomy/182435

19 September 2024

Peter Melchior, Assistant Professor of Statistical Astronomy from Princeton University, provides a compelling analysis of large-scale data fusion in astronomy

Observational astronomy constantly pushes the limits of what can be observed. We seek to peer deeper into the Universe, extend the wavelength range, or increase the spatial, spectral, and temporal resolution. Observing with the premier telescopes and the latest instruments leads to the groundbreaking advances that move astronomy forward. The discipline is defined by its instruments.

Enterprise-level astronomy

Because these instruments are expensive to build and operate, significant specialization and increasingly large scientific collaborations are needed to extract the best science from the data.

For example, the Vera C. Rubin Observatory is currently being built and commissioned by about 130 full-time employees, about half of whom form its Data Management team, which creates and maintains a codebase of about 500,000 lines of code in Python and C++. The resulting data products enable eight science collaborations with approximately 2,500 scientists. The data from the Euclid satellite, which launched in July 2023, is processed and analyzed by a consortium that comprises a similar number of scientists.

Efforts for these flagship facilities form the basis of a finely calibrated hardware and software ecosystem upon which the demanding studies for which these instruments were designed can successfully be carried out. Individual researchers and their students benefit from the creation and continuous improvement of data products they simply could not produce on their own. However, each of the large projects supports only its own instrument(s). Working with another data source becomes difficult. One has to leave the ecosystem with its dedicated software and data archives as well as scores of conventions and calibrations, only to reenter another project's ecosystem, where almost all elements are different, and details thereof may not be easy to find in public documentation.
In business terms: the large observational projects unintentionally promote lock-in.

It leads to a form of tunnel vision where solutions are pursued within the respective ecosystem that could better be addressed by leaving it. A prototypical example is a phenomenon called "blending". The latest crop of ground-based telescopes have become so sensitive that they pick up the emissions of many sources in any area of the sky. That is the goal because more sources increase the statistical power of scientific studies.

Blending challenge

But, with limited image sharpness caused by the blurring effects from Earth's atmosphere, it becomes increasingly common that multiple adjacent sources overlap. At the depth of Rubin Observatory's upcoming Legacy Survey of Space and Time (LSST), LARGE-SCALE DATA FUSION about 60% of all detected source have noticeable levels of blending, and about 15-20% are blended so severely that they appear as one source. [1] The sky gets crowded, and the idea that one can see a star or a galaxy as distinct and separate from its neighbors becomes largely obsolete. But separability is baked into almost all data processing algorithms and data products: we want the flux, size, ellipticity…of this galaxy, without the contributions of that galaxy.

To make matters worse, source separation is an under-constrained inverse problem: in general, we do not have enough information to unambiguously determine how a blended galaxy would look unblended. This challenge has led to a considerable amount of research on how machine-learning methods can be utilized to improve source separation [2,3,4]. Such methods learn from examples of isolated galaxies how to reconstruct unblended galaxies from blended images, that is, these methods learn a (often implicit) prior of how galaxies look. The origin of the training data is important. While we can create large samples of simulated galaxy images with simplified analytic profiles, none of them describe galaxies realistically. Still, high-quality samples of actual unblended galaxies are limited, on the order of 10,000 – 100,000, so training of deep-learning architectures becomes difficult.

We may not depend on priors for source separation as much as one could fear. Given the upcoming era of massive wide-field surveys, it is becoming increasingly likely that the same galaxy is observed by multiple instruments. Of particular importance in this context are space telescopes like Euclid because their data remain unperturbed by Earth's atmosphere. They can attain image resolutions that are out of reach from the ground, unless adaptive optics are employed, which can only be done over small sky areas. But, as rocket-delivered cargo, space telescopes and their components (mirrors, lenses, electronics) need to be kept small and light to keep the launch costs acceptable. What space telescopes gain in image resolution; they usually lack in sensitivity.

Advances from data fusion

My group designs techniques to exploit the evident synergy between space- and ground-based telescopes. We seek to capitalize on the overlap between the observed sky areas and the investments in high-quality data products to reconstruct galaxies from all available instruments, a process often called "Data Fusion". To be more concrete, we want to build a "virtual" telescope that combines the advantages of all instruments that can be integrated, so that the strengths of one data source can compensate for the relative weaknesses of others. From a Bayesian perspective, our proposal means a relaxation of the current strong reliance on priors (how does any galaxy look) by strengthening the likelihood (how does this galaxy look).

And it also allows us to see each galaxy in context. The model we want to build combines spatial, spectral, and temporal information for a comprehensive picture of a galaxy's properties and behavior. And because of the data volumes and substantial overlap of LSST and Euclid (their intersections cover about one-fourth of the entire sky), we will be able to do so for billions of galaxies. This approach will form the new basis of scientific studies that would otherwise be impossible, for example, on the complex interplay between astrophysical processes in the optical and infrared part of the spectrum or the importance of multi-modal deep learning [(5)] to establish common representations for galaxy images and spectra. To fully exploit the combined power of the large astronomical projects of the 2020s and beyond, investment should be made in education and long-term career paths that advance and leverage the confluence between observational astronomy and machine learning.

References

1. http://dx.doi.org/10.1038/s42254-021-00353-y
2. http://dx.doi.org/10.1093/mnras/staa3062
3. http://dx.doi.org/10.3847/1538-4357/aca1b8
4. http://dx.doi.org/10.1016/j.ascom.2024.100875
5. https://arxiv.org/abs/2310.03024

Please Note: This is a Commercial Profile