Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel, *University of Copenhagen*

Modern astronomy requires big data know-how, in particular, highly efficient machine learning and image analysis algorithms. But scalability is not the only challenge: astronomy applications touch several current machine learning research questions.

One of the largest astronomical surveys to date is the Sloan Digital Sky Survey (SDSS; www.sdss.org). Each night, the SDSS telescope produces 200 Gbytes of data, and to this day close to a million field images have been acquired, in which more than 200 million galaxies, and even more stars, have been detected. Upcoming surveys will provide far greater data volumes. Another promising future survey is the Large Synoptic Survey Telescope (LSST), which will deliver wide-field images of the sky, exposing galaxies that are too faint to be seen today. A main objective of the LSST is to discover *transients*, objects that change brightness over timescales of seconds to months. These changes are due to a plethora of reasons, some of which might be regarded as uninteresting while others will be extremely rare events that can't be missed. The LSST is expected to see millions of transients per night that need to be detected in real time to allow for follow-up observations. With a staggering 30 Tbytes of images being produced per night, efficient and accurate detection will be a major challenge. Figure 1 shows how data rates have increased and will continue to increase as new surveys are initiated.

What do the data look like? Surveys usually make either *spectroscopic* or *photometric* observations (Figure 2). Spectroscopy measures the photon count at thousands of wavelengths, and the resulting spectrum allows for identifying chemical components of the observed object and thus enables determining many interesting properties. Photometry takes images using a charge-coupled device (CCD), typically acquired through only a handful of broadband filters, making photometry much less informative than spectroscopy.

Although spectroscopy provides measurements of high precision, it has two drawbacks: it

1541-1672/17/$33.00 © 2017 IEEE
Published by the IEEE Computer Society

isn't as sensitive as photometry, meaning that distant or otherwise faint objects can't be measured, and only a few objects can be captured at the same time, making it more expensive than photometry, which allows for acquiring images of thousands of objects in a single image. Photometry can capture objects that might be 10 times fainter than what can be measured with spectroscopy. A faint galaxy is often more distant than a bright one—not just in space but also in time. Discovering faint objects therefore offers the potential of looking further back into the history of the universe, over timescales of billions of years. Thus, photometric observations are invaluable to cosmologists, as they help in understanding the early universe.

Once raw observations have been acquired, a pipeline of algorithms needs to extract information from them. Much of image-based astronomy currently relies to some extent on visual inspection. A wide range of measurements are still carried out by humans but need to be addressed by automatic image analysis in light of growing data volumes—examples include 3D orientation and chirality of galaxies, and the detection of large-scale features, such as jets and streams. Challenges in these tasks include image artifacts, spurious effects, and discerning between merging galaxy pairs and galaxies that happen to overlap along the line of sight. Current survey pipelines often have trouble correctly identifying these types of problems, which then propagate into the databases.

A particular challenge is that cosmology relies on scientific analyses of long-exposure images. As such, the interest in image analysis techniques for preprocessing and de-noising is naturally great. This is particularly important for the detection of faint objects with very low signal-to-noise ratios. Automatic object detection is vital to any survey pipeline, with reliability and completeness being essential metrics. Completeness refers to the amount of detected objects, whereas reliability measures how many of the detections are actual objects. Maximizing these metrics requires advanced image analysis and machine learning techniques. Data science for astronomy is a quickly evolving field gaining more and more interest.

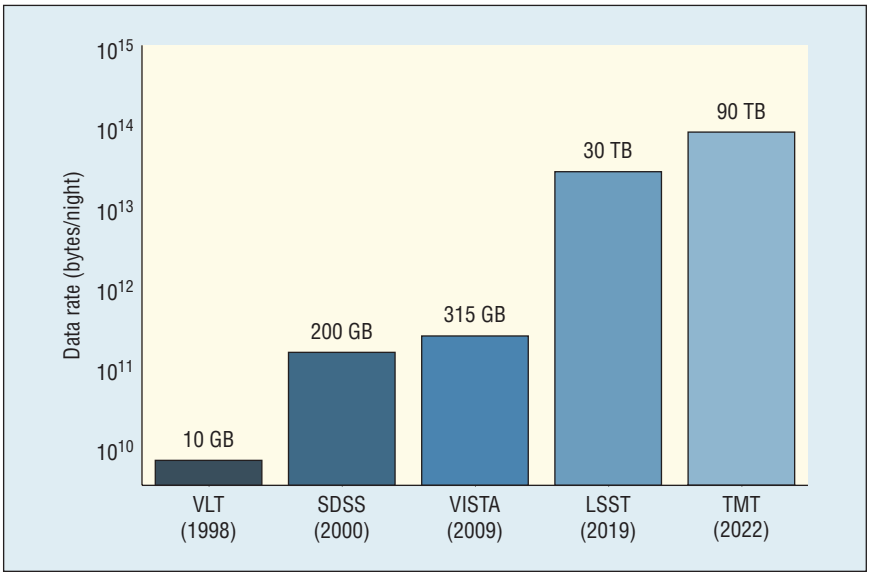

Figure 1. Increasing data volumes of existing and upcoming telescopes: Very Large Telescope (VLT), Sloan Digital Sky Survey (SDSS), Visible and Infrared Telescope for Astronomy (VISTA), Large Synoptic Survey Telescope (LSST), and Thirty Meter Telescope (TMT).

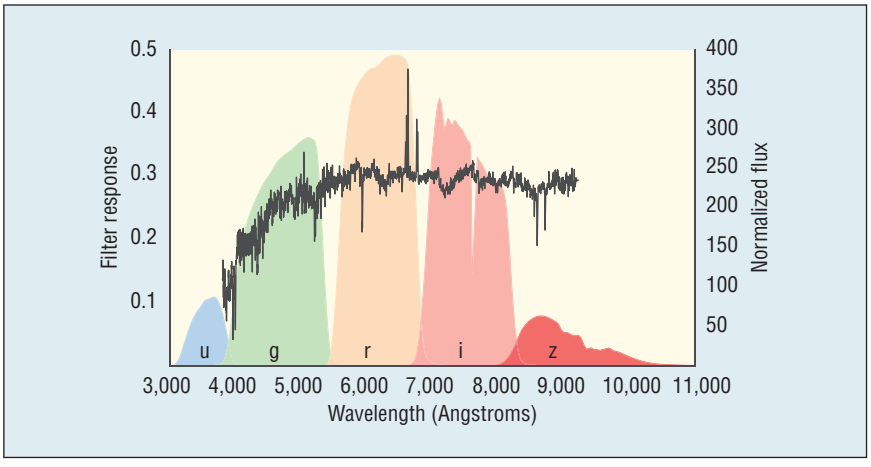

Figure 2. The spectrum of galaxy NGC 5750 (black line), as seen by the SDSS, with the survey's five photometric broadband filters u, g, r, i, and z, ranging from ultraviolet (u) to near-infrared (z). For each band, the galaxy's brightness is captured in an image.

Large-Scale Data Analysis in Astronomy

Machine learning methods can uncover the relation between input data (galaxy images) and outputs (physical properties of galaxies) based on input-output samples, and they've already proved successful in various astrophysical contexts. For example, Daniel Mortlock and colleagues[1] use Bayesian analysis to find the most distant quasar

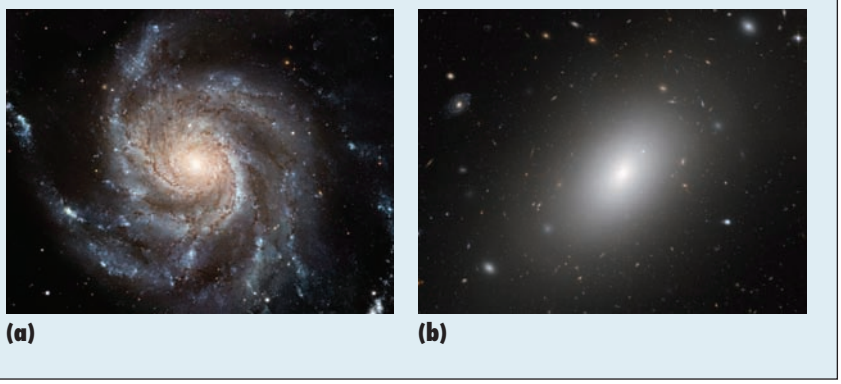

Figure 3. An example of two morphology categories: (a) the spiral galaxy M101 (credit: NASA, European Space Agency (ESA), K. Kuntz (Johns Hopkins University), F. Bresolin (University of Hawaii), J. Trauger (Jet Propulsion Lab), J. Mould (National Optical Astronomy Observatory), Y.-H. Chu (University of Illinois, Urbana), and Space Telescope Science Institute (STScI)), and (b) the elliptical galaxy NGC 1132 (credit: NASA, ESA, and the Hubble Heritage Team (STScI/Association of Universities for Research in Astronomy (AURA))-ESA/Hubble Collaboration).

to date. These extremely bright objects form at the center of large galaxies and are very rare. Bayesian comparison has helped scientists select a few most likely objects for re-observation from thousands of candidates.

In astronomy, distances from Earth to galaxies are measured by their redshifts, but accurate estimations require expensive spectroscopy. Getting accurate redshifts from photometry alone is an essential but unsolved task for which machine learning methods are widely applied.[2] However, they're far from being on a par with spectroscopy.

Another application is the measurement of galaxy morphologies. Usually, we assign a galaxy a class based on its appearance (Figure 3), traditionally via visual inspection. Lately, this has been accelerated by the citizen science project Galaxy Zoo,[3] which aims to involve the public in classifying galaxies. Volunteers have contributed more than 100 million classifications, which allow astrophysicists to look for links between galaxies' appearances (morphology) and internal and external properties. Several discoveries have been made through the use of data from Galaxy Zoo, and the classifications have provided numerous hints to the correlations between various processes governing galaxy evolution. A galaxy's morphology is difficult to quantize in a concise manner, and automated methods are high on the wish list of astrophysicists. There exists some work on reproducing the classifications using machine learning alone,[4] but better systems will be necessary when dealing with the data products of next-generation telescopes.

A growing subfield in astrophysics is the search for planets outside our solar system (exoplanets). NASA's Kepler spacecraft has been searching for exoplanets since 2009, observing light curves of stars—that is, measuring a star's brightness at regular intervals; the task is then to look for changes in the brightness, indicating that a planet might have moved in front of it. If this happens with regular duration and decrease in brightness, the source is likely to be an exoplanet. While automated software can detect such changes in brightness, the citizen science project Planet Hunters has shown that the software does miss some exoplanets. Also, detecting Earth-sized planets, arguably the most interesting, is notoriously difficult, as the decrease in brightness can be close to the noise level. For next-generation space telescopes, such as the Transiting Exoplanet Survey Satellite (TESS), scheduled to launch in 2017, algorithms for detecting exoplanets need to be significantly improved to more reliably detect Earth-sized exoplanet candidates for follow-up observations.

There are also problems that could directly affect our lives here on Earth, such as solar eruptions that, if headed toward Earth, can be dangerous to astronauts, damage satellites, affect airplanes, and, if strong enough, cause severe damage to electrical grids. Several spacecraft monitor the sun in real time to watch for these flares, and while the ultimate goal is a better understanding of the sun, the main reason is to be able to quickly detect and respond to solar eruptions. The continuous monitoring is done by automated software, but not all events are detected.[5] Solar eruptions are known to be associated with sunspots, but the connection isn't understood well enough that scientists can predict the onset or magnitude of an eruption. There could be a correlation with the complexity of the sunspots, and understanding this, as well as how the complexity develops over time, is crucial for future warning systems. While scientists are working toward a solution, for example, through the citizen science project Sunspotter (https://www.sunspotter.org), no automated method has yet been able to reliably and quantitatively measure the complexity.

This glimpse of success stories and open problems is by no means exhaustive; a much fuller overview of machine learning in astronomy appears elsewhere.[6]

Astronomy Driving Data Science

Three examples from our own work show how astronomical data analysis can trigger methodological advancements in machine learning and image analysis.

Describing the Shape of a Galaxy

Image analysis doesn't only allow for automatic classification—it can also inspire new ways to look at morphology.[7,8] For instance, we've examined how well one of the most fundamental measures of galaxy evolution, the star-formation rate, could be predicted from the *shape index*, which measures the local structure around a pixel going from dark blobs over valley-, saddle point-, and ridge-like structures to white blobs. It can thus be used as a measure of the local morphology on a per-pixel scale (Figure 4). The study showed that the shape index does indeed capture some fundamental information about galaxies, which is missed by traditional methods. Adding shape index features resulted in a 12 percent decrease in root-mean-square error (RMSE).

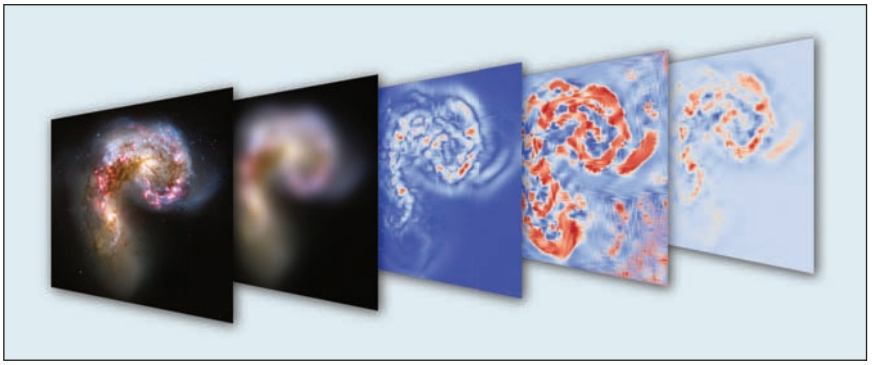

Figure 4. Shape index. From left to right, the original image of a galaxy merger, the scale-space representation of the galaxies, the curvedness (a measure of how pronounced the local structure is), the shape index, and finally the shape index weighted by the curvedness. The shape index is defined as $S(x,y;\sigma) = (2/\pi)\tan^{-1}\left(\left(-L_{xx}-L_{yy}\right)\Big/\sqrt{4L_{xy}^2+\left(L_{xx}-L_{yy}\right)^2}\right)$, where $L_{x^n y^m}(x,y;\sigma) = I * \partial^{(n+m)} G\Big/\left(\partial x^n \partial y^m\right)(x,y;\sigma)$ is the scale space representation of the image I, G is a Gaussian filter, and σ is the scale. The curvedness is defined as $C(x,y;\sigma) = \left(1/\sqrt{2}\right)\sigma^2\sqrt{L_{xx}^2+2L_{xy}^2+L_{yy}^2}$. The image shows the Antennae galaxies as seen by the Hubble Space Telescope (credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration).

Dealing with Sample Selection Bias

In supervised machine learning, models are constructed based on labeled examples—that is, observations (images, photometric features) together with their outputs (also referred to as labels, such as the corresponding redshift or galaxy type). Most machine learning algorithms are built on the assumption that training and future test data will follow the same distribution, which allows for generalization, enabling the model built from labeled examples in the training set to accurately predict target variables in an unlabeled test set. In real-life applications, this assumption is often violated in what we refer to as sample selection bias. Certain examples are more likely to be labeled than others due to factors such as availability or acquisition cost regardless of their representation in the population. Sample selection bias can be very pronounced in astronomical data,[9] and machine learning methods have to address this bias to achieve good generalization. Often, we only initially have training datasets from old surveys, while upcoming missions will probe never-before-seen regions in the astrophysical parameter space.

To correct the sample selection bias, we can resort to a technique called *importance-weighting*. The idea is to give more weight to examples in the training sample that lie in regions of the feature space underrepresented in the test sample and, likewise, give less weights to examples whose location in the feature space is overrepresented in the test set. If these weights are estimated correctly, the model we learn from the training data is an unbiased estimate of the model we would learn from a sample that follows the population's distribution. The challenge lies in estimating these weights reliably and efficiently. Given a sufficiently large sample, a simple strategy can be followed: using a nearest neighbor-based approach, we can count the number of test examples that fall within a hypersphere whose radius is defined by the distance to the Kth neighbor of a training example. The weight is then the ratio of the number of these test examples over K. This flexibly handles regions that are sparse in the training sample. In the case of redshift estimation, we could alleviate a selection bias by utilizing a large sample of photometric observations to determine the weights for the spectroscopically confirmed training set.[10]

To measure how well we approximated the true weight, we used the squared difference between true and estimated weight, that is,

$$L\left(\beta,\hat{\beta}\right) = \sum_{x\in S_{\text{train}}} \left(\beta(x)-\hat{\beta}(x)\right)^2 p_{\text{train}}(x)\,\mathrm{d}x,$$

where S_{train} is the training sample, β and $\hat{\beta}$ are true and estimated weight,

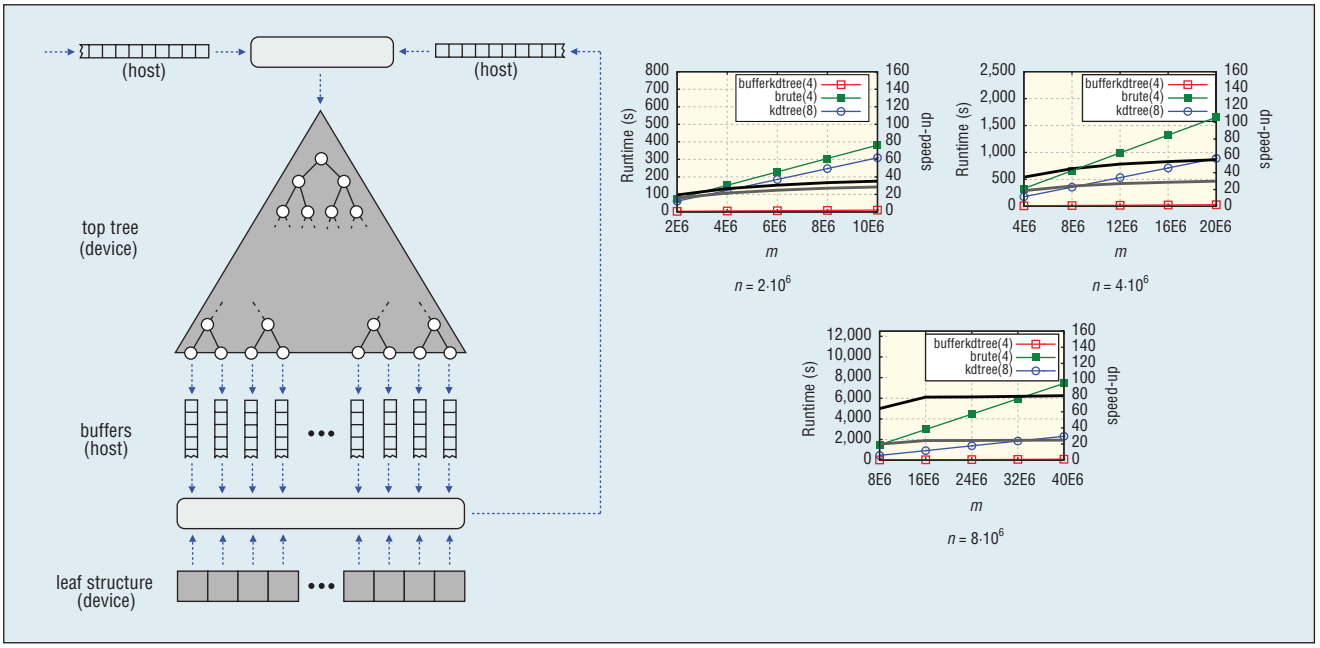

Figure 5. Nearest-neighbor searching. (a) The buffer k-d tree structure depicts an extension of classical k-d trees and can be used to efficiently process huge amounts of nearest-neighbor queries using GPUs. (b) Runtime comparison given a large-scale astronomical dataset with *n* training and *m* test examples. The speedup of the buffer k-d tree approach using four GPUs over two competitors (brute force on GPUs and a multicore k-d tree-based traversal using four cores/eight hardware threads) is shown as solid black lines.[12,13]

respectively, and p_{train} is the training density. The nearest-neighbor estimator achieved similar or lower error compared to other methods. At the same time, the estimator's running time is three orders of magnitude lower than the best competitor for lower sample sizes. Furthermore, it can scale up to millions of examples (code is available at https://github.com/kremerj/nnratio).

Scaling-Up Nearest-Neighbor Search

Nearest-neighbor methods are not only useful for addressing sample selection bias, they also provide excellent prediction results in astrophysics and cosmology—for example, they're used to generate candidates for quasars at high redshift.[11] Such methods work particularly well when the number of training examples is high and the input space is low-dimensional, making them a good choice for analyzing large sky surveys in which objects are described by photometric features (such as the five broadband filters in Figure 2). However, searching for nearest neighbors becomes a computational bottleneck in big data settings.

To compute nearest neighbors for a given query, search structures such as k-d trees are an established way to accelerate the search. If input space dimensionality is moderate (say, below 30), runtime can often be reduced by several orders of magnitude.

While approximate schemes are valuable alternatives, we're usually interested in an exact nearest-neighbor search for astronomical data. In this context, massively parallel devices, such as GPUs, show great promise. Unfortunately, nearest-neighbor search based on spatial data structures can't be parallelized in an obvious way for these devices. To this end, we developed a new tree structure that's more amenable to massively parallel traversals via GPUs (Figure 5).[12,13] The framework can achieve a significant runtime reduction at a much lower cost compared to traditional parallel architectures (code available at http://bufferkdtree.readthedocs.io). We expect such scalable approaches to be crucial for upcoming data-intensive analyses in astronomy.

Physical versus Machine Learning Models

A big concern data scientists meet when bringing forward data-driven machine learning models in astrophysics and cosmology is the lack of interpretability. There are two different approaches to predictive modeling in astronomy: physical modeling and data-driven modeling. Building physical models, which can incorporate all necessary astrophysical background knowledge, is the traditional approach. These models can be used for prediction, for example, by running Monte Carlo simulations. Ideally, this approach ensures that the predictions are physically plausible. In contrast, extrapolations by purely

data-driven machine learning models could violate physical laws. Another decisive feature of physical models is that they allow for understanding and explaining observations. This interpretability of predictions typically isn't provided when using a machine learning approach.

Physical models have the drawbacks that they're difficult to construct and that inference can take a long time (such as in the case of Monte Carlo simulations). Most importantly, the quality of the predictions depends on the quality of the physical model, which is typically limited by necessary simplifications and incomplete scientific knowledge. In our experience, data-driven models typically outperform physical models in terms of prediction accuracy. For example, a simple k nearest-neighbors model can reduce the RMSE by 22 percent when estimating star formation rates.[14,15] Thus, we strongly advocate data-driven models when accurate predictions are the main objective. And this is indeed often the case if, for example, we want to estimate properties of objects in the sky to quickly identifying observations worth a follow-up investigation or to conduct large-scale statistical analyses.

Generic machine learning methods aren't meant to replace physical modeling because they typically don't provide scientific insights beyond the predicted values. Still, we argue that if prediction accuracy is what matters, we should favor the more accurate model, whether it's interpretable or not. While the black-and-white portrayal of the two approaches might help illustrate common misunderstandings between data scientists and physicists, it is of course shortsighted. Physical and machine learning modeling aren't mutually exclusive: physical models can inform machine learning algorithms, and machine learning can support physical modeling. A simple example of the latter is using machine learning to estimate error residuals in a physical model.[7]

Dealing with uncertainties is a major issue in astronomical data analysis. Data scientists are asked to provide error bars for their predictions and to think about how to deal with input noise. In astronomy, both input and output data have (non-Gaussian) errors attached to them. Often, these measurement errors have been quantified (such as by incorporating weather conditions during observation), and it's desirable to consider these errors in the prediction. Bayesian modeling and Monte Carlo methods simulating physical models offer solutions, but they don't often scale for big data. Alternatively, we can modify machine learning methods to process error bars, as attempted for nearest-neighbor regression by modifying the distance function.[11]

Getting Started on Astronomy and Big Data

Most astronomical surveys make their entire data collection, including derived parameters, available online in the form of large databases, providing easy entry points for computer scientists wanting to get engaged in astronomical research.

The Galaxy Zoo website (https://www.galaxyzoo.org) provides data with classifications of approximately 1 million galaxies. It's an excellent resource for developing and testing image analysis and computer vision algorithms for automatic classifications of galaxies.

Much of the Kepler data for exoplanet discovery is publicly available through the Mikulski Archive for Space Telescopes (http://archive.stsci.edu/kepler). These include light curves for confirmed exoplanets and false positives, making it a valuable dataset for testing detection algorithms.

Having been monitored continuously for years, there's an incredible amount of imaging data for the sun, from archival data to near real-time images. One place to find such information is the Debrecen Sunspot Data archive (http://fenyi.solarobs.unideb.hu/ESA/HMIDD.html). These images allow for the development and testing of new complexity measures for image data or solar eruption warning systems.

Within the next few years, image analysis and machine learning systems that can process terabytes of data in near real time with high accuracy will be essential.

There are great opportunities for making novel discoveries, even in databases that have been available for decades. The volunteers of Galaxy Zoo have demonstrated this multiple times by discovering structures in SDSS images that have later been confirmed to be new types of objects. These volunteers aren't trained scientists, yet they make new scientific discoveries.

Even today, only a fraction of the images of SDSS have been inspected by humans. Without doubt, the data still hold many surprises, and upcoming surveys, such as LSST, are bound to image previously unknown objects. It won't be possible to manually inspect all images produced by these surveys, making advanced image analysis and machine learning algorithms of vital importance. Researchers could use these systems to answer questions such as how many types of galaxies there are, what distinguishes the different classes, whether the current classification scheme is good enough, and whether there are important subclasses or undiscovered classes. These questions require data science knowledge rather than astrophysical knowledge,

THE AUTHORS

Jan Kremer is a data scientist at Adform. His research interests include machine learning and computer vision. Kremer received a PhD in computer science from the University of Copenhagen. Contact him at jan.kremer@adform.com.

Kristoffer Stensbo-Smidt is a postdoctoral researcher in the Department of Computer Science, University of Copenhagen. His research interests include statistical data analysis and astrophysics. Stensbo-Smidt received a PhD in computer science from the University of Copenhagen. Contact him at k.stensbo@di.ku.dk.

Fabian Gieseke is an assistant professor in the Department of Computer Science, University of Copenhagen. His research interests lie in the field of big data analytics. Gieseke received a PhD in computer science from the University of Oldenburg. Contact him at fabian.gieseke@di.ku.dk.

Kim Steenstrup Pedersen is an associate professor in the Department of Computer Science, University of Copenhagen. His research interests include computer vision and image analysis. Pedersen received a PhD in computer science from the University of Copenhagen. Contact him at kimstp@di.ku.dk.

Christian Igel is a professor in the Department of Computer Science, University of Copenhagen. His research area is machine learning. Igel received a PhD in computer science from Bielefeld University and a Habilitation degree from Ruhr-University Bochum. Contact him at igel@diku.dk.

yet the discoveries will still help astrophysics tremendously.

In this new data-rich era, astronomy and computer science can benefit greatly from each other. There are new problems to be tackled, novel discoveries to be made, and above all, new knowledge to be gained in both fields. ⬒

References

1. D.J. Mortlock et al., "A Luminous Quasar at a Redshift of $z = 7.085$," *Nature*, vol. 474, no. 7353, 2011, pp. 616–619.
2. A.A. Collister and O. Lahav, "ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks," *Publications of the Astronomical Society of the Pacific*, vol. 116, no. 818, 2004, p. 345.
3. C.J. Lintott et al., "Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Soc.*, vol. 389, 2008, pp. 1179–1189.
4. S. Dieleman et al., "Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction," *Monthly Notices of the Royal Astronomical Soc.*, vol. 450, 2015, pp. 1441–1459.
5. E. Robbrecht and D. Berghmans, "Automated Recognition of Coronal Mass Ejections (CMEs) in Near-Real-Time Data," *Astronomy & Astrophysics*, vol. 425, 2004, pp. 1097–1106.
6. N.M. Ball and R.J. Brunner, "Data Mining and Machine Learning in Astronomy," *Int'l J. Modern Physics D*, vol. 19, no. 7, 2010, pp. 1049–1106.
7. K.S. Pedersen et al., "Shape Index Descriptors Applied to Texture-Based Galaxy Analysis," *Int'l Conf. Computer Vision (ICCV)*, 2013, pp. 2440–2447.
8. K. Polsterer et al., "Automatic Classification of Galaxies via Machine Learning Techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK)," *Astronomical Data Analysis Software and Systems XXVI*, 2015, pp. 81–86.
9. J.W. Richards et al., "Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification," *Astrophysical J.*, vol. 744, no. 2, 2012, pp. 192–210.
10. J. Kremer et al., "Nearest Neighbor Density Ratio Estimation for Large-Scale Applications in Astronomy," *Astronomy and Computing*, vol. 12, 2015, pp. 67–72.
11. K. Polsterer et al., "Finding New High-Redshift Quasars by Asking the Neighbours," *Monthly Notices of the Royal Astronomical Society*, vol. 428, no. 1, 2013, pp. 226–235.
12. F. Gieseke et al., "Buffer k-d Trees: Processing Massive Nearest Neighbor Queries on GPUs," *J. Machine Learning Research Workshop and Conf. Proc.*, vol. 32, no. 1, 2014, pp. 172–180.
13. F. Gieseke et al., "Bigger Buffer k-d Trees on Multi-Many-Core Systems," *Proc. Workshop Big Data & Deep Learning in HPC*, 2017, pp. 172–180.
14. K. Stensbo-Smidt et al., "Nearest Neighbour Regression Outperforms Model-Based Prediction of Specific Star Formation Rate," *IEEE Int'l Conf. Big Data*, 2013, pp. 141–144.
15. K. Stensbo-Smidt et al., "Sacrificing Information for the Greater Good: How to Select Photometric Bands for Optimal Accuracy," *Monthly Notices of the Royal Astronomical Soc.*, vol. 464, no. 3, 2017, pp. 2577–2596.

myCS Read your subscriptions through the myCS publications portal at **http://mycs.computer.org.**