



Bringing the Night Sky Closer: Discoveries in the Data Deluge

ALYSSA A. GOODMAN
Harvard University
CURTIS G. WONG
Microsoft Research

THROUGHOUT HISTORY, ASTRONOMERS have been accustomed to data falling from the sky. But our relatively newfound ability to store the sky’s data in “clouds” offers us fascinating new ways to access, distribute, use, and analyze data, both in research and in education. Here we consider three inter-related questions: (1) What trends have we seen, and will soon see, in the growth of image and data collection from telescopes? (2) How might we address the growing challenge of finding the proverbial needle in the haystack of this data to facilitate scientific discovery? (3) What visualization and analytic opportunities does the future hold?

TRENDS IN DATA GROWTH

Astronomy has a history of data collection stretching back at least to Stonehenge more than three millennia ago. Over time, the format of the information recorded by astronomers has changed, from carvings in stone to written records and hand-drawn illustrations to photographs to digital media.

While the telescope (c. 1600) and the opening up of the electromagnetic spectrum beyond wavelengths visible to the human eye (c. 1940) led to qualitative changes in the nature of astronomical investigations, they did not increase the volume of collected data nearly as much as did the advent of the Digital Age.

Charge-coupled devices (CCDs), which came into widespread use by the 1980s, and equivalent detectors at non-optical wavelengths became much more efficient than traditional analog media (such as photographic plates). The resulting rise in the rate of photon collection caused the ongoing (and potentially perpetually accelerating) increase in data available to astronomers. The increasing capabilities and plummeting price of the digital devices used in signal processing, data analysis, and data storage, combined with the expansion of the World Wide Web, transformed astronomy from an observational science into a digital and computational science.

For example, the Large Synoptic Survey Telescope (LSST), coming within the decade, will produce more data in its first year of operation—1.28 petabytes—than any other telescope in history by a significant margin. The LSST will accomplish this feat by using very sensitive CCDs with huge numbers of pixels on a relatively large telescope with very fast optics ($f/1.234$) and a wide field of view (9.6 square degrees), and by taking a series of many shorter exposures (rather than the traditional longer exposures) that can be used to study the temporal behavior of astronomical sources. And while the LSST, Pan-STARRS, and other coming astronomical mega-projects—many at non-optical wavelengths—will produce huge datasets covering the whole sky, other groups and individuals will continue to add their own smaller, potentially more targeted, datasets.

For the remainder of this article, we will assume that the challenge of managing this explosive growth in data will be solved (likely through the clever use of “cloud” storage and novel data structures), and we will focus instead on how to offer better tools and novel technical and social analytics that will let us learn more about our universe.

A number of emerging trends can help us find the “needles in haystacks” of data available over the Internet, including crowdsourcing, democratization of access via new browsing technologies, and growing computational power.

CROWDSOURCING

The Sloan Digital Sky Survey was undertaken to image, and measure spectra for, millions of galaxies. Most of the galaxy images had never been viewed by a human because they were automatically extracted from wide-field images reduced in an automated pipeline. To test a claim that more galaxies rotate in an anticlockwise direction than clockwise, the Sloan team used custom code to create a Web page that served up pictures of galaxies to members of the public willing to play the online Galaxy Zoo game, which consists primarily of classifying the handedness of the

galaxies. Clever algorithms within the “Zoo” serve the same galaxy to multiple users as a reference benchmark and to check up on players to see how accurate they are.

The results from the first year’s aggregated classification of galaxies by the public proved to be just as accurate as that done by astronomers. More than 50 million classifications of a million galaxies were done by the public in the first year, and the claim about right/left handed preference was ultimately refuted. Meanwhile, Hanny Van Arkel, a schoolteacher in Holland, found a galaxy that is now the bluest known galaxy in the universe. It has come under intense scrutiny by major telescopes, including the Very Large Array (VLA) radio telescope, and will soon be scrutinized by the Hubble Space Telescope.

DEMOCRATIZING ACCESS VIA NEW BROWSING TECHNOLOGIES

The time needed to acquire data from any astronomical object increases at least as quickly as the square of the distance to that object, so any service that can accumulate custom ensembles of already captured images and data effectively brings the night sky closer. The use of archived online data stored in a “data cloud” is facilitated by new software tools, such as Microsoft’s WorldWide Telescope (WWT), which provide intuitive access to images of the night sky that have taken astronomers thousands and thousands of hours of telescope time to acquire.

Using WWT (shown in Figure 1 on the next page), anyone can pan and zoom around the sky, at wavelengths from X-ray through radio, and anyone can navigate through a three-dimensional model of the Universe constructed from real observations, just to see what’s there. Anyone can notice an unusual correspondence between features at multiple wavelengths at some position in the sky and click right through to all the published journal articles that discuss that position. Anyone can hook up a telescope to the computer running WWT and overlay live, new images on top of online images of the same piece of sky at virtually any wavelength. Anyone can be guided in their explorations via narrated “tours” produced by WWT users. As more and more tours are produced, WWT will become a true “sky browser,” with the sky as the substrate for conversations about the universe. Explorers will navigate along paths that intersect at objects of common interest, linking ideas and individuals. Hopping from tour to tour will be like surfing from Web page to Web page now.

But the power of WWT goes far beyond its standalone ability. It is, and will continue to be, part of an ecosystem of online astronomy that will speed the progress of both “citizen” and “professional” science in the coming years.



FIGURE 1.

WorldWide Telescope view of the 30 Doradus region near the Large Magellanic Cloud.

Image courtesy of the National Optical Astronomy Observatory/National Science Foundation.

Microsoft, through WWT, and Google, through Google Sky, have both created API (application programming interface) environments that allow the sky-browsing software to function inside a Web page. These APIs facilitate the creation of everything from educational environments for children to “citizen science” sites and data distribution sites for professional astronomical surveys.

Tools such as Galaxy Zoo are now easy to implement, thanks to APIs. So it now falls to the astronomical and educational communities to capitalize on the public’s willingness to help navigate the increasing influx of data. High-school students can now use satellite data that no one has yet analyzed to make real discoveries about the Universe, rather than just sliding blocks down inclined planes in their physics class. Amateur astronomers can gather data on demand to fill in missing information that students, professionals, and other astronomers ask for online. The collaborative and educational possibilities are truly limitless.

The role of WWT and tools like it in the professional astronomy community will

also continue to expand. WWT in particular has already become a better way to access all-sky surveys than any extant professional tool. WWT, as part of international “virtual observatory” efforts, is being seamlessly linked to quantitative and research tools that astronomers are accustomed to, in order to provide a beautiful contextual viewer for information that is usually served only piecemeal. And it has already begun to restore the kinds of holistic views of data that astronomers were used to before the Digital Age chopped up the sky into so many small pieces and incompatible formats.

GROWING COMPUTATIONAL POWER

In 10 years, multi-core processors will enhance commodity computing power two to three orders of magnitude beyond today’s computers. How will all this computing power help to address the data deluge? Faster computers and increased storage and bandwidth will of course enable our contemporary approaches to scale to larger datasets. In addition, fully new ways of handling and analyzing data will be enabled. For example, computer vision techniques are already surfacing in consumer digital cameras with face detection and recognition as common features.

More computational power will allow us to triage and potentially identify unique objects, events, and data outliers as soon as they are detected and route them to citizen-scientist networks for confirmation. Engagement of citizen scientists in the alerting network for this “last leg” of detection can be optimized through better-designed interfaces that can transform work into play. Interfaces could potentially connect human confirmation of objects with global networks of games and simulations where real-time data is broadly distributed and integrated into real-time massive multiplayer games that seamlessly integrate the correct identification of the objects into the games’ success metrics. Such games could give kids the opportunity to raise their social stature among game-playing peers while making a meaningful contribution to science.

VISUALIZATION AND ANALYSIS FOR THE FUTURE

WWT offers a glimpse of the future. As the diversity and scale of collected data expand, software will have to become more sophisticated in terms of how it accesses data, while simultaneously growing more intuitive, customizable, and compatible.

The way to improve tools like WWT will likely be linked to the larger challenge of how to improve the way visualization and data analysis tools can be used together in all fields—not just in astronomy.

Visualization and analysis challenges are more common across scientific fields than they are different. Imagine, for example, an astronomer and a climate scientist working in parallel. Both want to study the properties of physical systems as observed within a spherical coordinate system. Both want to move seamlessly back and forth between, for example, spectral line observations of some sources at some specific positions on a sphere (e.g., to study the composition of a stellar atmosphere or the CO₂ in the Earth's atmosphere), the context for those positions on the sphere, and journal articles and online discussions about these phenomena.

Today, even within a discipline, scientists are often faced with many choices of how to accomplish the same subtask in analysis, but no package does all the subtasks the way they would prefer. What the future holds is the potential for scientists, or data specialists working with scientists, to design their own software by linking componentized, modular applications on demand. So, for example, the astronomer and the climate scientist could both use some generalized version of WWT as part of a separate, customized system that would link to their favorite discipline- or scientist-specific packages for tasks such as spectral-line analysis.

CONCLUSION

The question linking the three topics we have discussed here is, “How can we design new tools to enhance discovery in the data deluge to come in astronomy?” The answer seems to revolve around improved *linkage* between and among existing *resources*—including citizen scientists willing to help analyze data; accessible image browsers such as WWT; and more customized visualization tools that are mashed up from common components. This approach, which seeks to more seamlessly connect (and reuse) diverse components, will likely be common to many fields of science—not just astronomy—in the coming decade.