# Detecting the Unexpected: Discovery in the Era of Astronomically Big Data

J. E. G. Peek, jegpeek[at]stsci.edu

## Abstract

The era of big data is dawning in astronomy. Many have focused on the engineering challenges that come with petabyte-scale data sets and trillion-row databases. In this Institute workshop we explored a different side of big data: the scientific challenges to enabling the process of discovery in data sets too large to easily explore.

## Detecting the Unexpected

It is a truism the greatest discoveries from ground-breaking astronomical instruments aren't what was proposed in the whitepapers, but from surprising new results. In the era of *Hubble*'s launch, these discoveries were made by inspecting the data by eye. New correlations, new objects, and new physics were all discovered by manipulating small data sets, finding inconsistencies, and tracking them down with follow-up observation. How is this process going to work when datasets are far to large to be inspected "by eye" or even loaded into memory? How can we enhance the process of *discovery* in the era of astronomically big data?

*Detecting the Unexpected: Discovery in the Era of Astronomically Big Data* was a workshop held at the Institute in February 2017 aimed at precisely this question. We focused on what discoveries have been made with large data sets and what methods have been deployed to deal with these kinds of problems. We defined "Astronomically Big Data" broadly; not by bytes, but by a scope of data large enough that the classical methods of scientific exploration break down. *Mocking the Universe,* led by Dr. Molly Peeples, was an Institute workshop focused on how we can properly perform *hypothesis testing* against large surveys using powerful simulations of mock universes. In *Detecting the Unexpected* we focused instead on data sets large enough that our usual processes of *hypothesis generation* needed a fresh look.

## Scientific Themes & Data Methods

Hypothesis generation in big-data astronomy is a broad topic; we focused down onto a few main scientific themes, and a few data methods. Scientifically, we centered our discussions around the time-domain and large spectroscopic surveys. The time domain is a quickly-growing area of interest in astronomy with major projects across the electromagnetic spectrum. Presentations were made about how to find truly variable objects in real time in enormous data sets full of systematic errors, as well as how to characterize these objects, and find outliers. Spectroscopic observations are also in a period of dramatic growth, particularly with the explosion of new ground- and space-based facilities for Integral Field Unit and Multi-Object Spectroscopy, leading to very interesting discoveries. Talks focused on new surprising discoveries in vast spectroscopic databases, and the methods by which these discoveries were made.

Methodologically, we focused on three tools. The first was machine learning, and data-driven approaches to classification and outlier detection. Machine learning was shown to be especially important in real-time discovery systems like the Palomar Transient Factory, and for finding

extremely rare objects in large spectroscopic databases. Deep learning, or artificial neureal networks, were also showcased, and shown to be powerful tools in dealing with large quantities of hard-to-characterize data. The second tool was citizen science. Many presentations were given on Zooniverse projects, including *Gravity Spy, Radio Galaxy Zoo,* and *Disk Detective.* The fusion of machine learning and citizen science is proving to be a key avenue for discovery. Finally we explored data visualization, and how new visualization techniques and technologies can help us make sense of big data.

## New Workshop Formats for the Institute

*Detecting the Unexpected* broke new ground at the Institute by incorporating a few new ways of holding a workshop. Each day we held "Unconferences," which are participant-led breakout sessions on topics that the participants themselves decide on. Unconferences covered topics as diverse as how to change our field-support junior scientists engaged in data exploration and the details of how particular machine learning algorithms are implemented. We also had a sequence of sessions called the "Data & Methods Bazaar" in which presenters led short explorations of their data sets and software, and participants went from station to station, learning about particular tools. Finally, we held the Institute's first "Hack Day," in which participants pitched ideas they wanted to build in a single day. Groups were formed and over the course of the day we saw many different "Hacks" move forward. There were citizen science hacks on the interstellar medium, deep neural net hacks on stellar spectroscopy, and hacks for visualizing massive data sets over a network.

## The Future

Many came away from *Detecting the Unexpected* with new tools, new ideas, and a new appreciation for how tricky the process of astronomy will become as we enter a new realm of data. Chris Lintott, PI of of *Zooniverse* and professor of astrophysics at Oxford, will be carrying the *Detecting the Unexpected* torch to Oxford for *Detecting the Unexpected 2* in March, 2018.

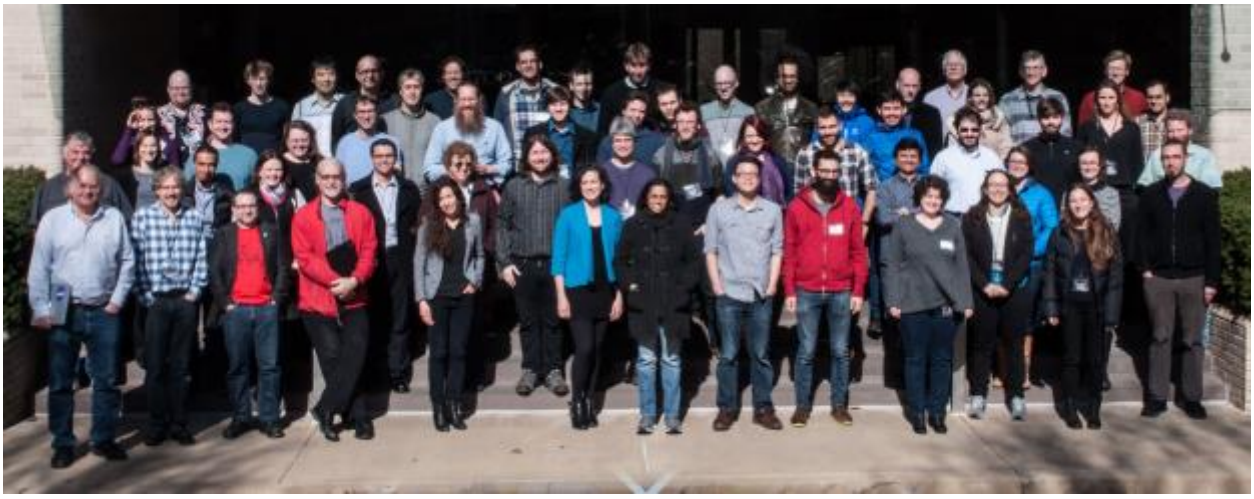

**Figure 1:** The attendees of *Detecting the Unexpected*. Image credit: Chad Smith.