Modern Astrophysics At The Forefront of Data Science

dAtAres UCLA DataRes Jun 1, 2019 · 11 min read ★

By Mason MacDougall

Introduction

When people think of astronomy, they usually think of stars and planets — galaxies and black holes — the vast expanse of space. They imagine ancient philosophers staring at the heavens through hand-held telescopes, mapping out the motions of the celestial objects visible in the night sky. Phases of the moon, constellations, and something about Kepler's Laws are all people

tend to remember from brief astronomy lessons in their middle school science classrooms.

Although what comes to mind when most people hear the word “astronomy” is not necessarily incorrect, it is a little outdated. In the 1920’s, Edwin Hubble calculated the distance to Andromeda with pen and paper while staring at photographic plates from the Mt. Wilson 100- inch telescope (the largest in the world until 1949). This was the first definitive evidence that the Milky Way was not the only galaxy in the universe. Today, Hubble is well-known as the namesake for the space-based telescope that took the clearest optical images of deep space ever taken before the 21 st century (**Figure 1**).

Figure 1. Hubble Telescope Deep Field optical image. Credit: NASA; ESA; G. Illingworth, D. Magee, and P. Oesch, University of California, Santa Cruz; R. Bouwens, Leiden University; and the HUDF09 Team.

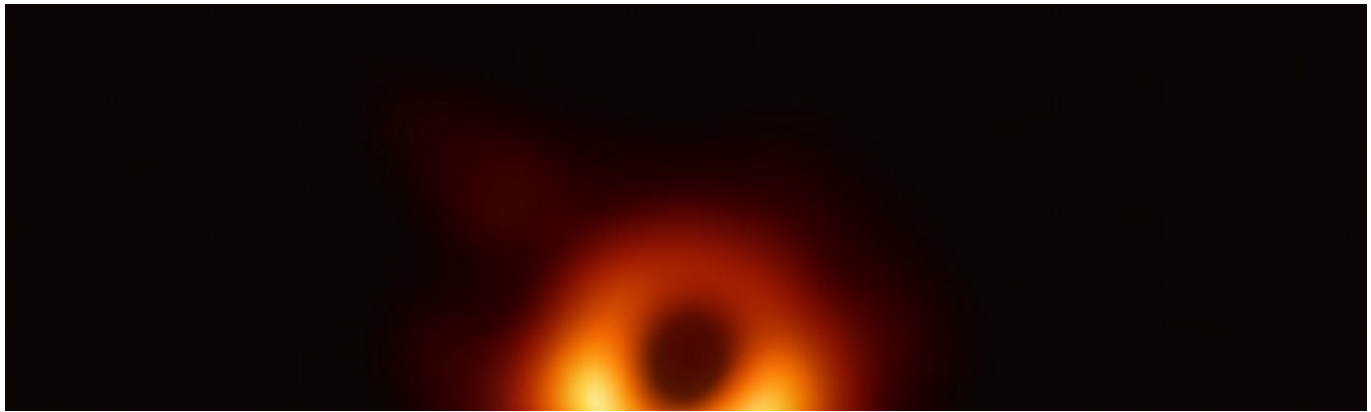
In less than 100 years, the field of astronomy has made leaps and bounds in progress — both philosophically and technologically. The 20th century took us from being unaware of the existence of other galaxies to hunting for Earth 2.0 among an ever-growing list of extra-solar planets. Correspondingly, we went from pen-and-paper and photographic plate images to supercomputers and petabytes of digital data. In order to keep up with the vast amount of data being collected by today's large assembly of telescopes, complex models, database frameworks, and processing pipelines continue to be developed and expanded.

Recently, major feats of astrophysics research have made their way into mainstream media such as the discovery of Trappist-1 in the search for extraterrestrial life, the detection of gravitational waves, and the imaging of a supermassive black hole in galaxy M87. The public sees easily digestible graphs, heavily processed images, and fanciful artistic conceptions without gaining a full understanding of the amount of time and effort that goes into producing such findings. Although the massive, highly advanced telescopes are usually what takes center-stage when explaining how astrophysical

discoveries are made, this is only part of the story. The rest of the story lies in the data analysis and processing that takes a telescope measurement and turns it into a confirmed discovery to be written about in a paper.

Black Hole In M87

About a month ago, in April 2019, a paper was published in The Astrophysical Journal by The Event Horizon Telescope (EHT) Collaboration regarding the imaging of the supermassive black hole at the center of galaxy M87. The groundbreaking paper, titled *First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole*, has deservedly emerged into the public eye with one particular figure making its way into news articles and media pieces around the world. This image, shown in **Figure 2**, is the culmination of several years of work by hundreds of scientists around the world — yet, despite the incredible feat of physics, engineering, and computing, many critics deemed this image "blurry," "unsatisfying," and "underwhelming."



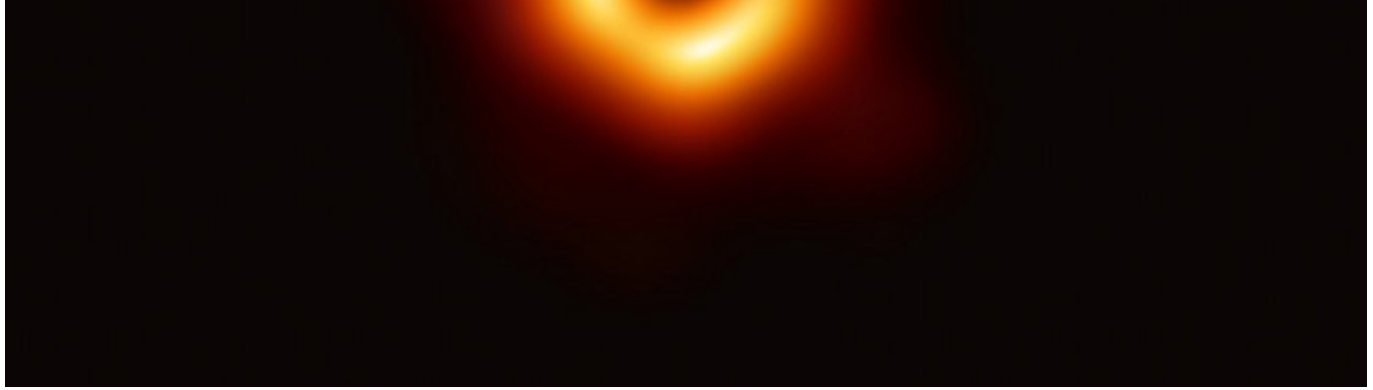


Figure 2. The Event Horizon Telescope network has obtained the first image of a black hole at the center of the galaxy M87. Credit: EHT Collaboration.

To put this image in perspective, the galaxy M87 is roughly 53.5 *million* light years away from Earth. That means that it takes light (the fastest thing in the universe) about 53.5 *million years* to travel from M87 to Earth. 53.5 million light years is equivalent to 314 quintillion miles or 506 quintillion kilometers. If you're not getting the picture yet, it is *really* far away.

Even more difficult is imaging the black hole at the center of this galaxy. Not only do we have to figure out where the blackhole is and how to see it through all of the gas, dust, and light obscuring our view, but we also have to resolve its relatively small size in the vast expanse of the night sky. This is roughly equivalent to trying to read a single 11-pt font letter on a piece of paper placed on the surface of the moon from here on Earth.

To accomplish such an impossible task, the Event Horizon Telescope was designed to essentially turn the entire planet Earth into a massive telescope.

This was accomplished through a technique known as very-long-baseline interferometry which essentially stitches together observations from radio telescopes all around the world to produce combined images with extremely high resolution, creating a "telescope" whose effective aperture is the diameter of the Earth. As one might imagine, this requires an incomprehensible amount of engineering and computational effort to perfect, especially on a global scale.

What most people do not see or understand is that this image comes from a scientific paper that is the 4th paper in a series of 6 where every aspect of the instrumentation, observations, and computation is described in detail. In paper *III. Data Processing and Calibration*, the EHT team illustrates the process of taking the *petabytes* of raw data produced by initial observations and reducing it into comprehensible scientific data.

The storage of this raw data is explained in a line from Section 3 of Paper III, reading "Through the receiver and backend electronics at each telescope, the sky signal is mixed to baseband, digitized, and recorded directly to hard disk, resulting in petabytes of raw VLBI voltage signal data." From here, a large team of astronomers and computer scientists were tasked with designing pipelines and tools to remove bad data points, reduce errors, and filter out noise through statistical modeling before even beginning to identify real astronomical signals among heaps of unusable

data. The tedium involved in this process from raw data to result is laid out in an example below from my own astrophysics research.

Modeling Orbital Parameters for Exoplanet Systems Using MCMC Technique

Currently, I am working in the field of observational exoplanetary astrophysics –essentially, I am looking for Earth 2.0 while also trying to understand the wide variety of planets that exist. My project is being led by Dr. Erik Petigura who is wrapping up his post-doctoral work at Caltech before transitioning to an assistant professor position here at UCLA this summer.

To carry out our work, I examine data from exoplanet observations by the Kepler and TESS spacecrafts to try to figure out the properties of planet candidates. The challenge with this field is that exoplanets cannot usually be directly observed due to their lack of strong emission (relative to their host stars) and their relatively minuscule size. For example, we can barely see Pluto in our own solar system from here on Earth, so it is understandably difficult to see planets in other solar systems lightyears away.

Since stars are much more easily detectable by modern telescopes, we look at how a planet around a given star produces observable changes in the

perceived properties of the star. One such property that we focus on is a star's brightness and how this brightness changes over time when a planet directly passes between a star and the Earth (**Figure 3**). This type of occurrence is known as a "transit" and observations of transits have yielded several thousand candidate exoplanets to date — which results in a lot of data! So, what can we do with this data?

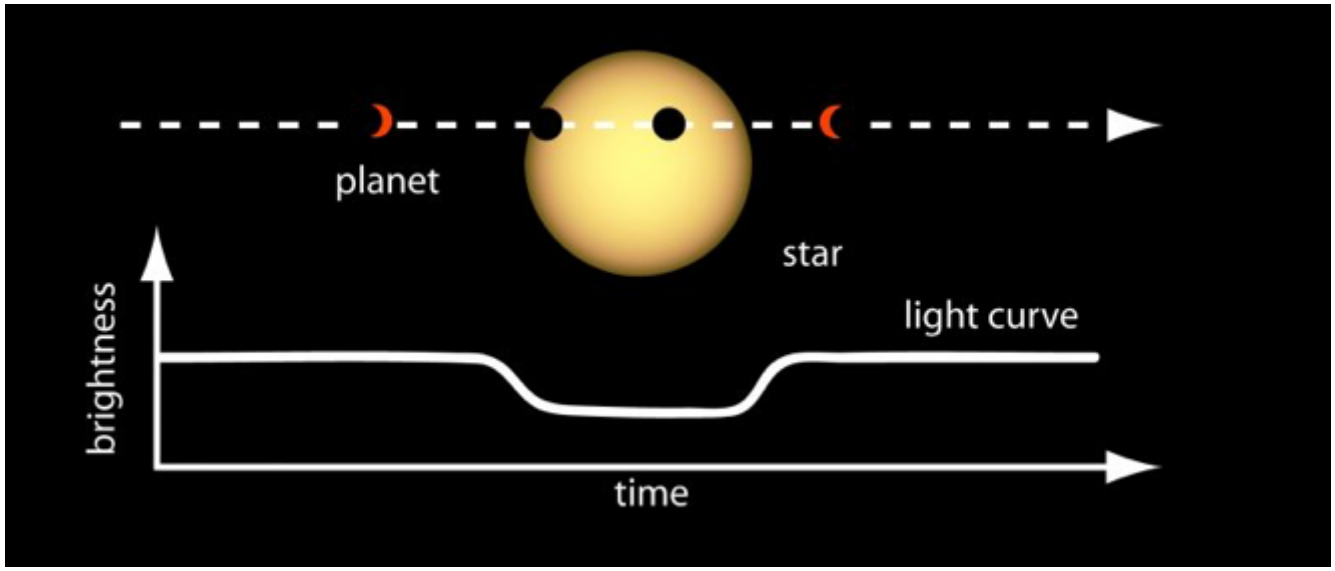


Figure 3. Simplified diagram showing exoplanet transiting in front of the host star and its effect on the star's apparent brightness over time. Credit: NASA.

Transit data primarily consists of a transit "light curve" displaying stellar brightness over time. However, this data must be heavily reduced and cleaned up before any science can be done. This is where data science comes in to save the day by taking the raw observational data and finding the signals hidden within the noise. In my project, the light curves used

were measured with the Kepler Space Telescope and extracted from the raw data archives on MAST (Mikulski Archive for Space Telescopes). These look something like **Figure 4** before any analysis or reduction has been done, but this is not even close to the ideal flat line with a dip seen in **Figure 3**.

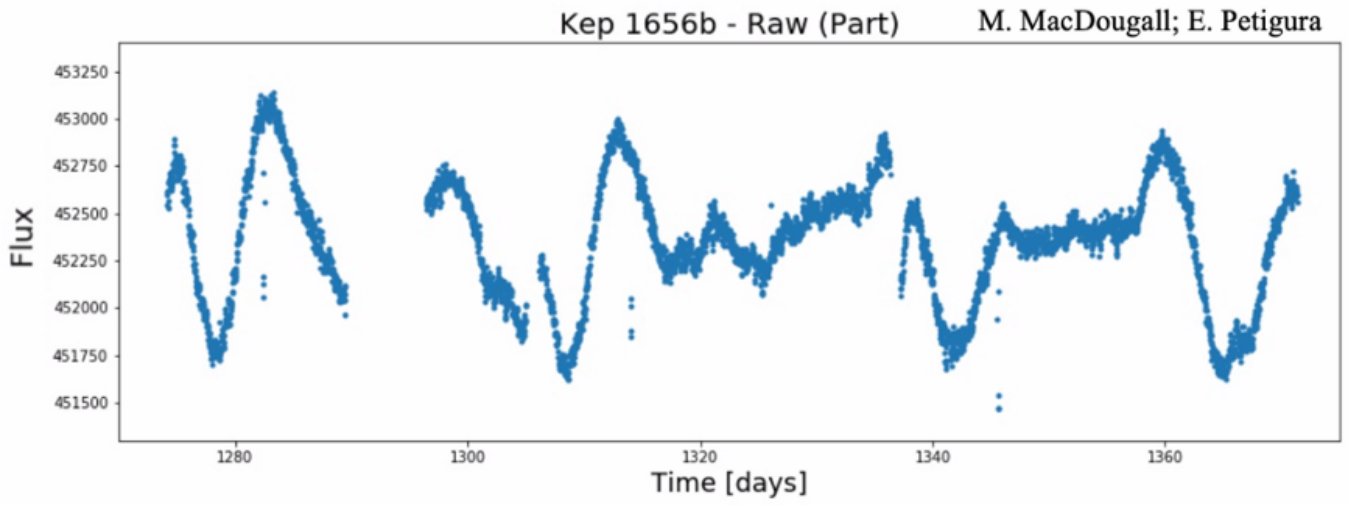


Figure 4. Example raw light curve for exoplanet Kepler 1656b showing high noise and variability before reduction. Credit: Kepler Space Telescope; MAST; M. MacDougall; E. Petigura; Brady et al. 2018.

Before attempting to extract any scientific measurements from this light curve, we normalize the data and mask out bad data points that recorded non-real, null, or infinite brightness values. We then detrend our data by removing the low-frequency background noise trend causing the strong flux variations using a **Savitzky-Golay filter**. With this filter, we successfully flatten our data via convolution over successive sub-sets of adjacent points in our data set to fit these points with a low-degree polynomial via linear least squares fitting.

The final step in the reduction process is removing significant outliers that might substantially skew statistical modeling results. This is done through both a sigma cutoff of 6-s and manual clipping of high variance regions of post-reduction data. The pipeline that I created to successfully run the raw data through took about a month to finalize, and even then, I have only checked its success with one exoplanet light curve. In the end, we have a flattened, normalized, higher signal-to-noise light curve with clearly defined transits occurring periodically **Figure 5**.

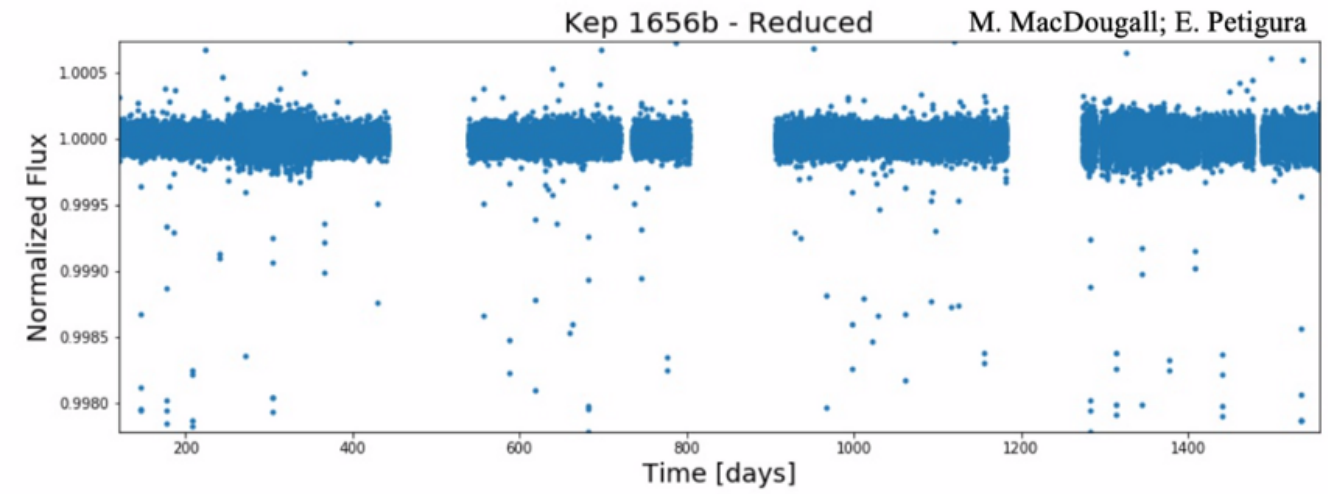


Figure 5. Example of reduced light curve for Kepler 1656b showing normalized flux with periodic dimming. Credit: M. MacDougall; E. Petigura; Brady et al. 2018.

From the data contained in this light curve, the best-constrained property of the planet that can be derived is an estimate of the planet's radius based on how much stellar light is blocked during transit. We can also determine the orbital period of a transiting planet by noting how far apart in time

transits occur. To ensure that all potential transits are actually related events occurring at a fixed orbital period, we must phase-fold our data around the transit midpoint and see if all transit candidates stacked on top of one another have the same shape and depth (**Figure 6**).

Figure 6. Example of the phase-folded light curve for Kepler 1656b — Period of 31.578659 days. Credit: M. MacDougall; E. Petigura; Brady et al. 2018.

Once a period and radius have been estimated, we can approximate a variety of other orbital parameters including inclination, semi-major axis (distance between planet and star), and eccentricity (how circular or elongated the orbit is). However, there is no simple equation that we can plug values into and suddenly know everything about the planetary system. Although the physics of orbital dynamics is well understood, many of the parameters that are considered are degenerate with one another so a wide range of combinations could give similar results.

In order to home in on the parameter values that produce a model which best fits our data, we must set up a statistical modeling program to optimize the fit. The modeling software used in this process is a Python package known as **BATMAN** (Bad-Ass Transit Model cAlculatioN — this is the actual name) which takes in various orbital parameters and produces an idealized light curve model based on the inputs.

Thus, given a guess of the optimal parameters, we can produce an accompanying model and compare it to the actual data to assess the fit. This assessment is made on the basis of chi-squared testing to determine the level of correlation between each observed data point and each modeled data point. The better the fit, the lower the total chi-squared value for a model derived from a particular set of guessed parameters.

Figure 7. Example of attempting to fit a BATMAN model to the phase-folded light curve of Kepler 1656b. Credit: M. MacDougall; E. Petigura; Brady et al. 2018.

The catch here is that the fastest way to find the best fit model is to already have a pretty good idea of what the best parameter estimates are. This problem is made even worse by the fact that the Kepler Space Telescope long cadence data is measured only every 30 minutes, so for a transit that occurs over the course of about 3 hours we have at most 6 data points per transit. With such low resolution, it's difficult to precisely determine the shape of the transit — including depth, flattening, steepness, and symmetry which are all used to infer the orbital parameters.

It is much easier to get a rough estimate by modeling the phase-folded light curve as seen in **Figure 7**, but ultimately we must rely on statistical modeling to be able to properly model the entire unfolded light curve at each time stamp in the original data. Luckily, this particular system was already studied in depth by Brady et al. 2018, where the planet was found to have an eccentricity of roughly 0.84 among other precise estimates. We take this information (planet radius, distance from the host star, period, inclination, eccentricity, position along an orbit, and time of first transit) as the initial condition to an optimization tool known as a Markov Chain Monte Carlo (MCMC).

An MCMC is designed to take as its input a set of guessed parameter values, create a model from these values, and use chi-squared testing to compare the model to the data. Once this process has completed, the program creates a new set of guessed parameters that are slightly perturbed from the

initial set. Then, given the chi-squared value of the last fit, the program weighs whether or not it is valuable to take the proposed step to this new set of parameters or stay where it is. We run this for 106 steps from 20 different initial guesses (walkers) simultaneously, with each step slowly getting closer to the optimal fit. We would ideally end up with a Gaussian distribution of our walkers' final parameters, but so far we have not been able to achieve this convergence, unfortunately. This is likely a matter of poorly constrained prior assumptions, bad initial guesses, or too few steps — all of which we are still looking into. Nevertheless, we still have fairly well-constrained estimates of all of the orbital parameters in this system which we can use to model the light curve data to a substantial level of accuracy (**Figure 8**).

Figure 8. Example plotting a best-fit BATMAN model with the reduced light curve of Kepler 1656b demonstrating strong agreement. Credit: M. MacDougall; E. Petigura; Brady et al. 2018.

Although our optimization method has not yet been finalized, we have made significant progress towards accomplishing our goal of automatedly identifying the best fit light curve model for a given planetary system using BATMAN. We will continue to work on getting our MCMC up and running well while also looking into other statistical techniques that might better assess how good a fit is. We aim to test our program on a larger sample of Kepler planet candidates with previously estimated orbital parameters to improve our ability to properly model these already well-studied light curves. Once our technique has demonstrated consistently high accuracy, we will begin to use it to model new TESS candidates in an attempt to better understand the orbital characteristics of the population of planets being observed. With such knowledge, we may be able to gain new knowledge regarding the habitability of distant worlds and the likelihood of finding life beyond Earth.

Science Astrophysics Datares Data Science Black Holes