Vol. 32, No. 3, 2019

CHANCE

Using Data to Advance Science, Education, and Society

Special Issue on ASTROSTATISTICS

Including...

Statistics for Stellar Systems: From Globular Clusters to Clusters of Galaxies

ARIMA for the Stars: How Statistics Finds Exoplanets

09332480(2019)32(3)

Taylor & Francis
Taylor & Francis Group

ASA

EXCLUSIVE BENEFITS
FOR ALL ASA MEMBERS!

SAVE 30% on Book Purchases with discount code **ASA18**.
Visit the new ASA Membership page to unlock savings on the latest books, access exclusive content and review our latest journal articles!

With a growing recognition of the importance of statistical reasoning across many different aspects of everyday life and in our data-rich world, the American Statistical Society and CRC Press have partnered to develop the **ASA-CRC Series on Statistical Reasoning in Science and Society**. This exciting book series features:

- Concepts presented while assuming minimal background in Mathematics and Statistics.
- A broad audience including professionals across many fields, the general public and courses in high schools and colleges.
- Topics include Statistics in wide-ranging aspects of professional and everyday life, including the media, science, health, society, politics, law, education, sports, finance, climate, and national security.

DATA VISUALIZATION
Charts, Maps, and Interactive Graphs

Robert Grant, BayersCamp

This book provides an introduction to the general principles of data visualization, with a focus on practical considerations for people who want to understand them or start making their own. It does not cover tools, which are varied and constantly changing, but focusses on the thought process of choosing the right format and design to best serve the data and the message.

September 2018 • 210 pp • Pb: 9781138707603: ~~$29.95~~ $23.96 • www.crcpress.com/9781138707603

VISUALIZING BASEBALL

Jim Albert, Bowling Green State University, Ohio, USA

A collection of graphs will be used to explore the game of baseball. Graphical displays are used to show how measures of batting and pitching performance have changed over time, to explore the career trajectories of players, to understand the importance of the pitch count, and to see the patterns of speed, movement, and location of different types of pitches.

August 2017 • 142 pp • Pb: 9781498782753: ~~$29.95~~ $23.96 • www.crcpress.com/9781498782753

ERRORS, BLUNDERS, AND LIES
How to Tell the Difference

David S. Salsburg, Emeritus, Yale University, New Haven, CT, USA

In this follow-up to the author's bestselling classic, "The Lady Tasting Tea", David Salsburg takes a fresh and insightful look at the history of statistical development by examining errors, blunders and outright lies in many different models taken from a variety of fields.

April 2017 • 154 pp • Pb: 9781498795784: ~~$29.95~~ $23.96 • www.crcpress.com/9781498795784

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol 112, 2017

THE AMERICAN STATISTICIAN
Vol 72, 2018

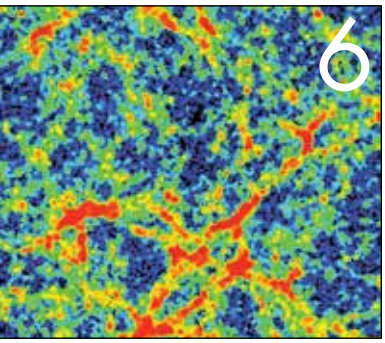
STATISTICS AND PUBLIC POLICY
Vol 5, 2018

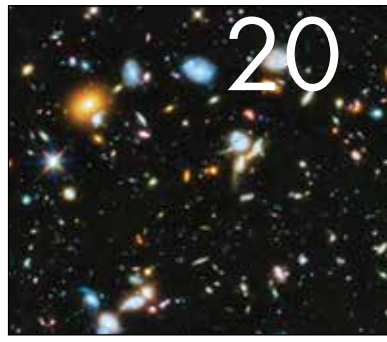
Taylor & Francis Group
an **informa** business

ASA

http://bit.ly/CRCASA2018

CHANCE
Using Data to Advance Science, Education, and Society
http://chance.amstat.org





ARTICLES

COLUMNS

DEPARTMENTS

Abstracted/indexed in Academic OneFile, Academic Search, ASFA, CSA/Proquest, Current Abstracts, Current Index to Statistics, Gale, Google Scholar, MathEDUC, Mathematical Reviews, OCLC, Summon by Serial Solutions, TOC Premier, Zentralblatt Math.

Cover design: Melissa Gotherman

AIMS AND SCOPE

CHANCE is designed for anyone who has an interest in using data to advance science, education, and society. *CHANCE* is a non-technical magazine highlighting applications that demonstrate sound statistical practice. *CHANCE* represents a cultural record of an evolving field, intended to entertain as well as inform.

SUBSCRIPTION INFORMATION

CHANCE (ISSN: 0933-2480) is co-published quarterly in February, April, September, and November for a total of four issues per year by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA, and Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

U.S. Postmaster: Please send address changes to *CHANCE*, Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

ASA MEMBER SUBSCRIPTION RATES

ASA members who wish to subscribe to *CHANCE* should go to ASA Members Only, *www.amstat.org/membersonly* and select the "My Account" tab and then "Add a Publication." ASA members' publications period will correspond with their membership cycles.

SUBSCRIPTION OFFICES

USA/North America: Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: 215-625-8900; Fax: 215-207-0050. UK/Europe: Taylor & Francis Customer Service, Sheepen Place, Colchester, Essex, CO3 3LP, United Kingdom. Telephone: +44-(0)-20-7017-5544; fax: +44-(0)-20-7017-5198.

For information and subscription rates please email s*ubscriptions@tandf.co.uk* or visit *www.tandfonline.com/pricing/journal/ucha*.

OFFICE OF PUBLICATION

American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA. Telephone: 703-684-1221. Editorial Production: Megan Murphy, Communications Manager; Valerie Nirala, Publications Coordinator; Ruth E. Thaler-Carter, Copyeditor; Melissa Gotherman, Graphic Designer.
Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: 215-625-8900; Fax: 215-207-0047.

RESPONSIBLE FOR ADVERTISEMENTS

Send inquiries and space reservations to: *advertising@taylorandfrancis.com*.
Printed in the United States on acid-free paper.

EDITOR'S LETTER

Scott Evans

Dear CHANCE *Colleagues,*

In 1961, the president of the United States John F. Kennedy wanted to land humans on the moon. For the next several years, NASA prepared to do so.

On July 16, 1969, Apollo 11 blasted off carrying astronauts Neil Armstrong, Edwin "Buzz" Aldrin and Michael Collins. Four days later, on July 20, Armstrong and Aldrin landed the lunar module called the Eagle on the moon. Astronaut Collins stayed in orbit performing experiments and taking photos. Armstrong became the first human to step on the moon. He and Aldrin walked around for three hours, conducting experiments, and picking up bits of moon dirt and rocks. They planted a United States flag and left a sign that read "Here men from the planet Earth first set foot upon the moon July 1969, A.D. We came in peace for all mankind."

The two astronauts returned to orbit, joining Collins. On July 24, 1969, all three astronauts came back to Earth safely. President Kennedy's wish came true in less than 10 years.

As humanity has continued to explore space, the collaborations between statisticians and astronomers has increased. This special issue of *CHANCE* is devoted to astrostatistics.

I wish to thank the ASA's Astrostatistics Special Interest Group, who helped us develop this special issue. Special thank you to the guest editors for this issue: **Jessi Cisewski-Kehe**, assistant professor in the Department of Statistics and Data Science at Yale University and past chair of the Astrostatistics Special Interest Group, and **Chad Schafer**, associate professor in the Department of Statistics & Data Science at the Carnegie Mellon University and chair of the Astrostatistics Special Interest Group.

Scott Evans

Special Issue on Astrostatistics

Jessi Cisewski-Kehe and Chad Schafer

Like most areas of science, astronomy has benefited from a tremendous increase in the amount of available data gathered in recent decades, from both ground- and space-based instruments. For example, the Sloan Digital Sky Survey (SDSS) commenced in 2000, at a time when studies were typically conducted with galaxy samples of sizes on the order of 1,000. By 2008, SDSS had produced a map of 930,000 galaxies, with high-resolution spectral information for each. By the time of its 15th data release in 2017, SDSS had a catalog of more than 2.5 million galaxies. The Large Synoptic Survey Telescope (LSST), currently under construction in Chile, is projected to yield a catalog of 10^{10} galaxies by the time its 10-year survey ends in 2032.

Data gathered about astronomical objects take a range of forms, but at their basic level, they consist of intensity measurements, often observed at a range of different wavelengths, at different spatial positions, and at various points in time. It is the variability across objects, wavelengths, and space and time that encodes invaluable information regarding the formation, evolution, and current state of the universe and its components. Analyses are complicated, however, by observational limitations, including measurement error, contamination, and the inherent limitations of our Earth-centered frame of reference.

The role and participation of statisticians in astronomy has increased along with the sizes of the data sets and the complexity of the inference challenges, but further collaboration is needed to address the many open problems. A core group of statisticians is committed to this effort; they are working to make astronomical data analysis known to the broader statistical community and helping to bridge any language barriers that may exist.

It has been our experience that the perceived difficulty of the subject matter is exaggerated. We typically involve undergraduate and graduate students, with no background in astronomy, in fruitful projects.

With this in mind, this special issue of *CHANCE* seeks to provide some context for common themes in astrostatistics, the types of data encountered, and the important role that statistical methods can play. Methods well-established in statistics may be little-known or used in astronomy. Astronomical problems also present unique challenges that force the development of new methods, or at least push existing methods in interesting directions.

In this issue, **Green**, **Mintz**, **Xu**, and **Cisewski-Kehe** discuss one of the recurring challenges of working with modern astronomical data—namely, that the data in their raw form are complex and not amenable to classical statistical analysis. Astronomers have often used ad hoc compression of these data, extracting low-dimensional features that are believed to encode important information, but the potential loss of information is great. Ideas from topological data analysis, a new and growing area of study in statistics, have great promise for data-driven approaches to extracting important information from data such as the large-scale structure of the universe.

Politsch and **Croft** describe the challenges and potential of working with the *Lyman-alpha forest,* another complex data set that informs our understanding of the large-scale structure of the universe. Lyman-alpha forest data are particularly interesting because of the innovative approach for collecting these data. The light from distant quasars provides access to some aspects of the intergalactic medium, which can then be used to infer properties of the distribution of gas in regions that would otherwise be inaccessible.

Although modern astronomical data sets are massive, individual observations are often of low quality. **Freeman** provides an overview of the challenges of using low-resolution photometry to estimate a

fundamental property of a celestial object: its distance from us. On cosmological scales, distance is a proxy for time, so any study of the evolution of the universe relies on accurate distance measures, typically quantified via the *redshift*.

Eadie considers a similarly fundamental problem—that of estimating the mass of gravitationally bound dynamical systems like galaxies. Such estimates are crucial to understanding the nature of *dark matter*. How the positions and velocities of stars (called *tracers*) are used to constrain a system's mass illustrates a range of inference approaches in astronomy: These observations have a complex, physically motivated relationship with the quantity of interest. Estimation in such situations presents unique challenges, and has motivated the development of a range of novel techniques.

A chief benefit of increased participation of statisticians in astronomy is the cross-pollination of ideas across fields: Statisticians work in a wide range of domains, of course, and experts in the development of statistical methods are always on the lookout for new areas of application for classic and novel approaches. **Segal** and **Segal** apply the Patient Rule Induction Method (PRIM), a less-known but potentially useful, tool for supervised learning about a classification problem in astronomy that involves identifying open clusters of stars using Gaia data.

Perhaps one of the most-exciting recent developments in astronomy is the discovery of a large number of exoplanets orbiting stars in the Milky Way, and statistical tools play a crucial role in making these discoveries. **Feigelson** provides an overview of the challenges of searching for the signatures of exoplanets in noisy time series, and discusses how classical time series models and modern machine learning can combine to develop improved approaches to this important problem.

Corliss demonstrates the potential of modern approaches to clustering in the classification of supernovae—the explosive deaths of stars—based on their observed time series. This work demonstrates that clustering, especially with data such as these, requires careful consideration of the choice of similarity measure.

With this issue as a starting point, there are many avenues for getting involved in astrostatistics.

- The American Statistical Association has an Astrostatistics Special Interest Group (*https://community.amstat.org/astrostats/home*) that welcomes and encourages the participation of statisticians and astronomers, including students, postdocs, researchers, faculty, and members of industry or government.
- The Astrostatistics and Astroinformatics Portal (ASAIP, *https://asaip.psu.edu*) is a central site that compiles information about the field relevant to astronomers, computer scientists, and statisticians. It includes information about papers, meetings, and other resources for both new and active researchers in astrostatistics.
- The Cosmostatistics Initiative (COIN, *https://cosmostatistics-initiative.org*) is an international and interdisciplinary community focused on developing stronger ties between the various fields related to astrostatistics. COIN also organizes "Residence Programs" where a small group of researchers in astronomy, computer science, cosmology, and statistics meets for a week in various destinations around the world to focus on solving several problems in the field while forming closer interdisciplinary ties. The Residence Programs have been quite productive and are a great way to make connections with those interested in astrostatistics.
- The International Astrostatistics Association (IAA, *http://iaa.mi.oa-brera.inaf.it/IAA/home.html*) seeks to bring together researchers from around the world who are interested in advancing the field of astrostatistics.

Involvement in any of these organizations can be a great way to learn more about astrostatistics. We hope many of you will join us in the exciting pursuit of exploring our universe.

Jessi Cisewski-Kehe
Guest Editor

Chad Schafer
Guest Editor

Topology of Our Cosmology with Persistent Homology

Sheridan B. Green, Abby Mintz, Xin Xu, and Jessi Cisewski-Kehe

Originally a speculative branch of natural philosophy, cosmology is an ancient field concerned with answering some of the big-picture questions about the origin and evolution of our universe. Our modern scientific understanding of physical cosmology has been under development for roughly a century, beginning around the time that Albert Einstein was writing about the astronomical implications of his theory of general relativity.

The discovery of the Cosmic Microwave Background (CMB) (the farthest structure whose light we can observe in the universe, which provides a snapshot of the universe only 400,000 years after its birth) in 1964 validated predictions of the original Big Bang model, leading to several additional decades of theoretical developments that have resulted in the current "standard model of cosmology." In this Lambda-Cold Dark Matter (ΛCDM) picture of the universe, the vast majority of matter is "dark" (i.e., it does not emit light and only interacts with regular matter via gravity) and the cosmos is expanding at an ever-increasing rate due to the abundance of dark energy (known as Λ), a mysterious substance with negative pressure that seems to permeate all of space.

In the past 20 years, the arrival of massive astronomical data sets that map out the large-scale structure (LSS) of the universe (e.g., the Sloan Digital Sky Survey [SDSS]), which is also referred to as the "cosmic web," has ushered in a data-driven, interdisciplinary era of "precision cosmology." The high nonlinearity of cosmic structure formation at and below megaparsec- (Mpc-) length scales precludes the ability to make many purely theoretical predictions of the consequences of ΛCDM (1 parsec is approximately 3.25 light years, and the largest galaxy clusters are on the order of Mpc). Fortunately, recent times have also brought about substantial increases in computational power, enabling predictions of ΛCDM to be made via cosmological *N*-body simulations (see Figure 1), one of the most-successful tools to date for making quantitative predictions of cosmic structure.

This abundance of data, both observed and simulated, has provided support for ΛCDM: Both theoretical and computational predictions of the model on large scales have been validated in statistical stress-tests against observations.

However, ΛCDM is not without its potential discrepancies and inconclusivities with the current generation of astronomical data. For instance, it remains unclear whether the structure of galaxy centers is consistent with ΛCDM, since there have been some hints that dark matter particles may actually scatter off each other (CDM is assumed to be collisionless), changing the central galactic density. We also have yet to verify whether Einstein's cosmological constant Λ, which is an unchanging property of space, does indeed describe dark energy or if the properties of dark matter instead vary across spacetime, as predicted by "quintessence" models.

Recently, a "crisis in cosmology" has been brewing, as disagreement between the value of the Hubble constant, which describes the expansion rate of the universe, calculated from temperature fluctuations in the CMB, has risen to a 4.4σ tension with the value calculated using Cepheid variable stars. These stars, which are thousands of times brighter than the sun and oscillate periodically in brightness, are known as "standard candles" because their absolute brightness can be determined directly by measuring its period. From the relationship between the absolute brightness and observed brightness of these stars, their distances can be determined, and that information can be used to trace the expansion history of the universe.

Various solutions to this apparent inconsistency with ΛCDM have been proposed, primarily including modifications to dark energy and the incorporation of the effects of the nonzero neutrino mass on large-scale structure.

To estimate free parameters and test ΛCDM, various model predictions are compared to astronomical observations and often combined using Bayesian inference

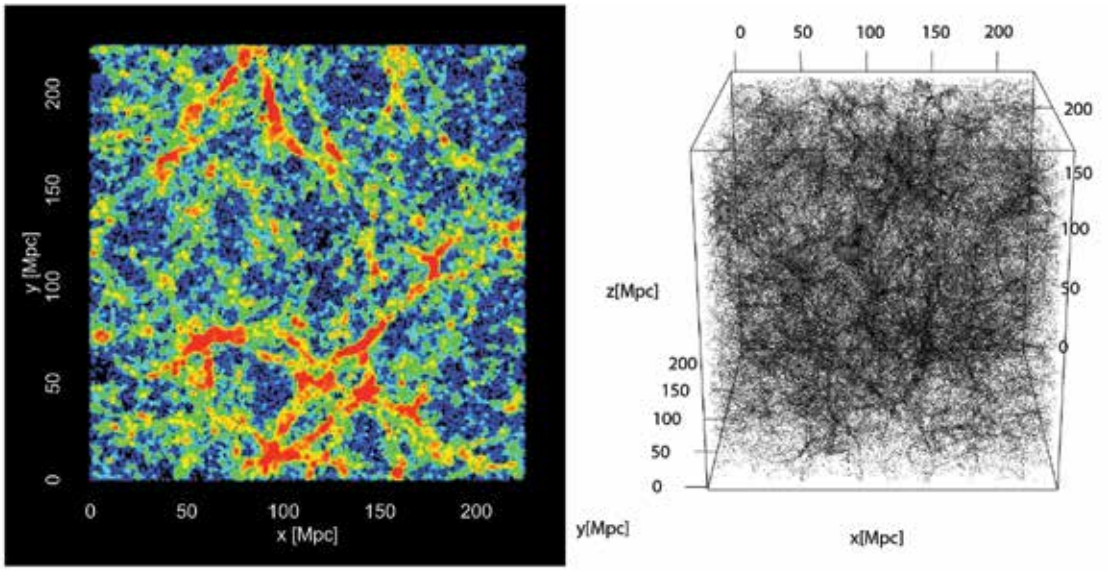


Figure 1. A simulation of the cosmic web under the ΛCDM paradigm. The large-scale structure in such simulations is both qualitatively and quantitatively consistent with the universe (right). The haloes within a thin slice of the simulation volume have points representing individual dark matter haloes, and the color represents the density of the local matter field with redder regions having higher density (left). The individual points of the 3D cosmological volume represent the locations of dark matter haloes.

techniques. Such comparisons between theory and observation use summary statistics to reduce the massive observational data sets down to several physically interpretable quantities, which are also predicted by theoretical models.

For example, one key summary statistic used for such parameter estimation is the galaxy two-point correlation function $\xi(r)$, which describes the excess probability, relative to a Poisson distribution, of finding a pair of galaxies separated by a distance r. The value of r at which this function reaches a particular local maximum plays an important role as a "cosmic ruler," with a predicted length that depends on the cosmological model and its parameters ("baryon acoustic oscillations" are responsible for this peak in $\xi(r)$).

It is expected that the massive amounts of data that will soon become available from the upcoming generation of astronomical surveys (e.g., the Large Synoptic Survey Telescope [LSST] and the Dark Energy Spectroscopic Instrument [DESI]) will provide sufficient information to discriminate between ΛCDM and various alternative models, and make it possible to gain deeper insights into the fundamental nature of our evolving universe.

To exploit the observational data optimally and unlock the maximum information content available, efforts are underway to identify additional, complementary summary statistics that are particularly sensitive to the nuanced differences between various models (e.g., the effects of dark energy or massive neutrinos on LSS).

This article explores how ideas from topological data analysis (TDA) can be used to study properties of the cosmic web. TDA provides tools for analyzing the shape of data. These tools have potential for improving the understanding of the cosmic matter field and find physically interesting summaries of the cosmic web, which can help constrain some of the unknown quantities of the universe.

The Basics of Persistent Homology

Persistent homology is a tool of TDA that can be useful in settings with data characterized by holes, such as the cosmic web, because it produces summaries of different dimensional holes in a data set that are not easily available using other methods. Persistent homology is a framework for computing the homology of actual data. In terms of homology, we speak of these

different dimensional holes by the homology groups they generate.

For example, 0-dimensional homology groups (H_0) are generated by what in statistics would be considered to be clusters, one-dimensional homology groups (H_1) are generated by loops, and two-dimensional homology groups (H_2) are generated by things like the interior of 3D balls (such as a soccer ball).

There are higher-dimensional homology groups as well, but we need only consider H_0, H_1, and H_2 for cosmology, since the universe only has three spatial dimensions. The H_0 groups can be generated by clusters of dark matter haloes (as seen in cosmological simulations) or clusters of galaxies (as observed in galaxy surveys such as the SDSS). The H_1 groups can be generated by the filaments (the weblike strings in Figure 1) that form loops, and the H_2 groups can be generated by voids (the low matter density regions of the universe).

With some notion of a connection between homology and cosmology, it is possible to get a better idea of what persistent homology provides. Consider the simulated point-cloud data set in Figure 2(a). There are points randomly sampled on three loops with a little noise and additional points scattered away from the loops. Calculating the homology of the data set would reveal that there are 200 H_0 generators (the 200 points sampled in the data set), and no H_1 or H_2 generators…but there appear to be at least three clear loops, suggesting there should be at least three H_1 generators.

This is where the "persistent" in persistent homology comes into play. For persistent homology, we create some sort of intermediate structure that changes with a specified filtration parameter.

There are different ways to do this. One approach involves growing balls around the points, where the filtration parameter is the radius of the ball; this is known as a Rips filtration. As the radius increases, eventually balls start to intersect with each other. When two balls intersect, they join to form a single connected component (H_0 generator). If a loop is present, such as those in Figure 2(a), eventually the balls connect in such a way that the loop forms. The radius of the ball (the filtration parameter value) when this loop forms is called the "birth" of the H_1 generator.

As the balls continue to grow, eventually the loop gets filled in by the larger balls; the radius at which this occurs specifies the "death" of the H_1 generator. The "persistence" of a feature is the difference between the birth and death times.

Persistent homology keeps track of the birth and death times of the different homology group generators and then plots these values on a summary diagram known as a "persistence diagram." The Rips filtration persistence diagram for this data set is displayed in Figure 2(c). The red triangles indicate the birth and death times of the H_1 generators, and the black circles indicate the birth and death times of the H_0 generators. The persistence diagrams show that the further a generator is from the diagonal (the 45-degree line), the longer that feature lasts in the filtration.

Sometimes it makes sense to consider the features that last the longest to be topological signal (e.g., if larger, fixed structures are expected), while the features near the diagonal may be considered topological noise.

From the Rips persistence diagram of Figure 2(c), there are six H_1 generators that seem well-separated from the diagonal. However, the corresponding data set shows that there seem to be only three clear loops. Additional loops can form in the Rips filtration due to the extra noise points and the layout of the loops.

Fortunately, other filtration methods tend to be more-robust to noise. One alternative filtration method uses the so-called "distance to a measure" (DTM) function, which can be thought of as estimating on a grid the average distance to some subset of the nearest neighbors within the original data set (requiring the selection of the number of nearest neighbors to use).

Another alternative is to use a kernel density estimate to smooth over the data set, which requires selection of the band width. DTM turns the point-cloud data into a function on a grid, so growing the balls for the filtration will no longer work. Instead of the radius of a ball, the filtration parameter is governed by a threshold parameter that defines lower-level sets, as explained below. As the threshold parameter increases, the lower-level sets are used to assess when homology group generators form and die.

Figure 2(b) displays an estimated DTM function. To get a sense of how, for example, an H_1 generator can form using a DTM function, imagine setting a threshold value of around 0.4 (see the color bar on the right side of Figure 2(b)) and consider the lower-level sets (the grid locations that correspond to the DTM function values that are less than 0.4). The pixels that are "activated" at that threshold would form the three loops, as shown in Figure 3.

As the threshold increases, the lower-level sets will start to fill in the loops. The DTM persistence diagram is presented in Figure 2(d). The persistence

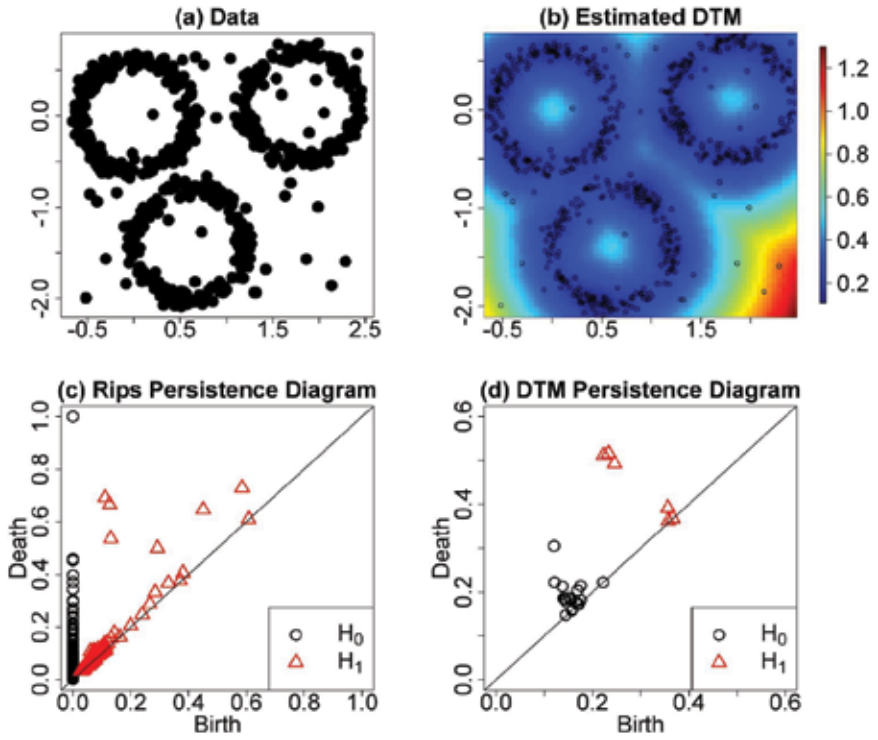


Figure 2. (a) A simulated data set. (b) An estimated DTM function for the data set displayed in (a). The corresponding persistence diagrams using a Rips filtration (c) and a DTM filtration (d). The red triangles represent the H_1 generators and the black circles represent the H_0 generators.

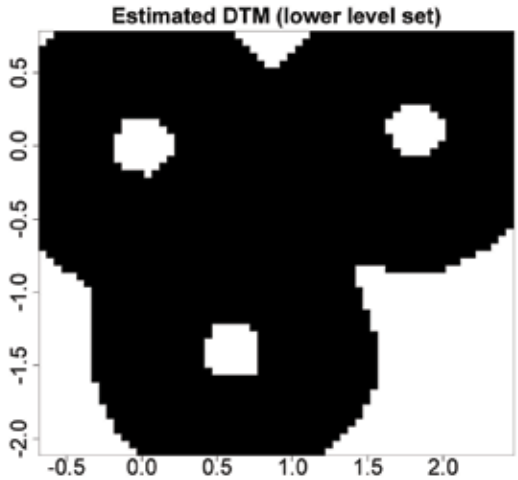


Figure 3. Lower-level set (in black) of the estimated DTM function from Figure 2(b) using a threshold of 0.4. Notice the three loops, indicating that their births occurred before 0.4 and their deaths will be after 0.4.

diagram only has three H_1 generators (red triangles) that are well-separated from the diagonal, in contrast to the six in the Rips persistence diagram of Figure 2(c).

Persistence diagrams can be useful objects to consider for data visualization, but they have also been employed for inference or prediction in settings beyond astronomy and cosmology (e.g., brain artery trees, natural language processing, biology). Many of the inference and prediction methods developed use functional summaries of persistence diagrams.

The SCHU Method

As noted previously, the homology group generators, particularly H_1 and H_2, have connections to relevant cosmological structures. In previous work, a method was developed, called "Significant Cosmic Holes in Universe" (SCHU), for finding statistically significant H_1 generators (known as "filament loops") and H_2 generators (cosmic voids) back in the original cosmological data volume using techniques available in the R package called "TDA."

In SCHU, a three-dimensional grid is built around the data volume and a DTM filtration is constructed to obtain the persistence diagram. SCHU can then assign statistical significance to generators on persistence diagrams and also locate a representation of those features back in the data volume. More specifically, SCHU provides a p-value for each cosmic feature based on a bootstrap procedure, which is computationally intensive because each bootstrap sample requires its own DTM filtration computations.

Representations of cosmological structures as defined by the homology group generators are provided by SCHU as points on the DTM grid, forming filament loops or enclosing cosmic voids. Because the homology group generators are actually mathematical groups, unique representations are not available for the generators indicated on a persistence diagram. However, obtaining a representation of these features in the cosmological volume is important for understanding the potential scientific relevance of the objects.

Other methods exist for identifying cosmological voids, so to compare the SCHU H_2 voids to such methods, estimating their physical locations is essential. For a DTM persistence diagram, the representation of a generator using SCHU tend to be the inner contour of the smallest lower-level set that forms the homology group generator. The choice of the spatial resolution (i.e., the size of the grid cells) and the number of the nearest neighbors used for computing the DTM function influences the appearance of the representations.

Using SCHU to Study Cosmological Simulations

The distribution of matter in the universe is visibly traced by galaxies, which are what are observed in galaxy surveys. In the ΛCDM model, galaxies are predicted to form within dark matter haloes. While there is not exactly a one-to-one mapping between the unseen dark matter haloes and the observed galaxies (the interested reader can look up "galaxy assembly bias," for example), the cosmic web can be studied using dark matter haloes (from large-scale cosmological simulations). Similar results can be expected when using galaxies (retrieved from real astronomical surveys).

This is helpful because it means being able to forgo running expensive, complicated cosmo-hydrodynamical simulations, which attempt to incorporate the many details of galaxy formation. Instead, the results of relatively cheap dark matter-only cosmological N-body simulations of enormous physical volumes can be used as the input to SCHU.

The large volumes studied in such simulations make it possible to use the "linear theory" of structure formation, which is only valid on large scales (i.e., above Mpc scales), while also providing information about the evolution on smaller scales that would be otherwise unobtainable through analytical techniques. These simulations also allow easily specifying, and then varying, the cosmological model under which the universe evolves. Thus, in principle, we can run simulations with widely different models of dark energy or that span a large range of the ΛCDM parameter space, profiling how the topology (encoded in the persistence diagrams) depends on the cosmology.

Currently, various simulation databases cover interesting regions of cosmological model space. For example, the Dark Energy Universe Simulation Series (DEUSS) consists of several simulations with various models of dark energy, including ΛCDM and the Ratra-Peebles and SUGRA quintessence models.

For each of these models, DEUSS contains several simulations that range over different spatial scales to capture the large spatial dynamic range over which dark energy acts.

One of the standard ΛCDM simulations in DEUSS tracks the evolution of dark matter within a cube of side length 225 megaparsecs (i.e., about

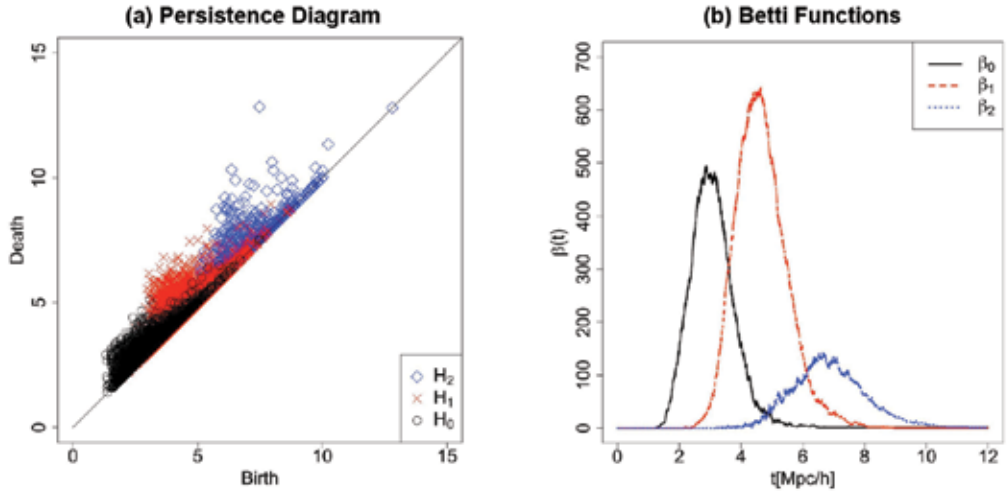


Figure 4. (a) The persistence diagram and (b) Betti functions from the $z = 0$ snapshot of the 225 Mpc ΛCDM DEUSS simulation displayed in Figure 1.

1 millionth of the volume of the observable universe) for exploring the cosmological features using SCHU. Specifically, the input data to SCHU consist of a positional catalog of dark matter haloes identified at a particular time "snapshot" during the evolution (the haloes within a thin slice of the simulation volume are displayed in Figure 1).

Motivated by the fact that the spatial extent of the largest haloes (analogous to galaxy clusters) is on the order of megaparsecs, and being primarily concerned with the much-larger filament loops and voids, a DTM with a spatial resolution for the grid of roughly one megaparsec can be employed.

Start by looking at the fully evolved simulation, which should be roughly consistent with a present-day universe (i.e., is consistent with the local environment at "redshift" $z = 0$) in a ΛCDM cosmology.

Figure 4(a) shows the plot of the persistence diagram generated by SCHU for the H_0, H_1, and H_2 of the cosmological simulation illustrated in Figure 1. The corresponding "Betti functions," which can be thought of as a functional summary of a persistence diagram, also can be displayed. For a particular dimension of homology group generators i, the ith Betti function $\beta_i(t)$ captures, as a function of the filtration threshold length scale t, the number of H_i group generators that have already been born by t and have yet to die off.

In other words, $\beta_i(t)$ represents the number of points directly above a square of side length t extending from the origin of the persistence diagram.

Figure 4(b) shows the plot of the Betti functions computed from the SCHU persistence diagrams. Functional summaries like these Betti functions can be easier to work with than persistence diagrams, especially when attempting to compare the persistence diagrams of different simulations as done below. The location of the Betti function peaks reveal the t at which the most homology group generators are active (i.e., born before and die after that t). Because we use the DTM filtration, the t can be understood as an average distance. In Figure 4(b), this suggests that the distances at which the H_1 generators form and die off are generally smaller than the distances at which the H_2 generators form and die off.

As described above, SCHU enables us to also visualize representations of these group generators in the cosmological volume. The most physically interesting generators in this setting are for H_1 and H_2 with the highest persistence (i.e., those that are furthest from the diagonal on the persistence diagram). Figure 5 plots the locations of the dark matter haloes and overlay the 10 cosmic voids and filament loops with the highest persistence in the simulation volume. The H_2 cosmic voids highlight the low-density regions of the data volume and range in both size and shape.

There are other methods for locating cosmic voids presented in the literature, but there is not yet a consensus on a definition for cosmic voids. The H_1 filament loops

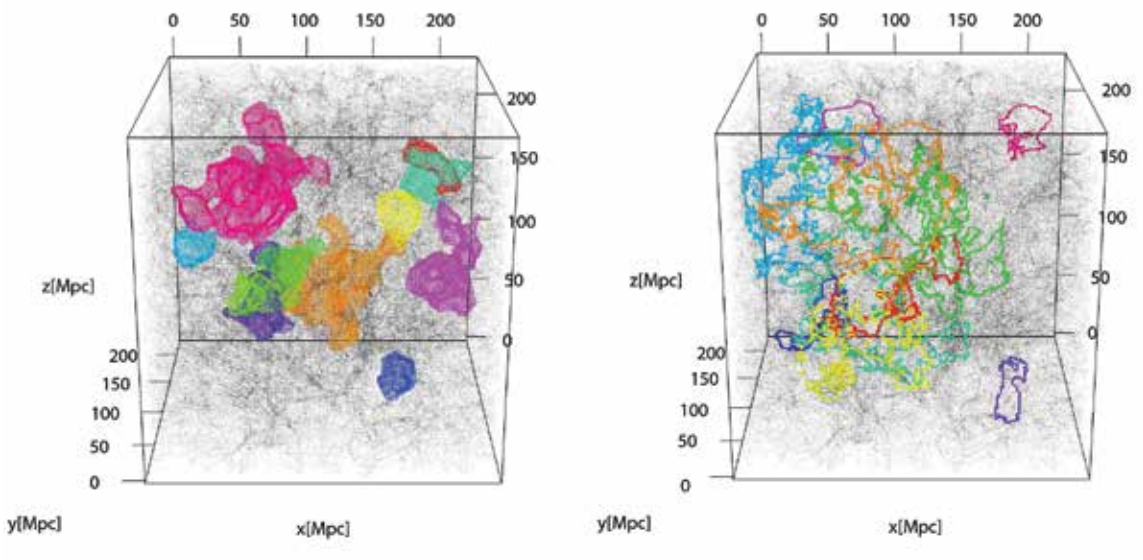


Figure 5. The most-persistent (left) voids and (right) filament loops in the $z = 0$ snapshot of the 225 Mpc ΛCDM DEUSS simulation identified using SCHU.

are a new type of cosmological structure, and we are not aware of methods, other than the combination of persistent homology and SCHU, that would be able to find such information in LSS.

The filament loops displayed in Figure 5, like the cosmic voids, vary in size and shape—some stretch and twist through a large portion of the LSS, while others are smaller and isolated to subsets of the cube. Visualizing the H_1 and H_2 structures can be useful for later analyses that may investigate, for example, how these structures form or properties of the local environment of such structures.

Another topic of interest in astronomy is how structure evolves across cosmic time. These TDA tools turn out to be useful summaries for comparing different snapshots of the LSS at different points in time. In particular, Betti functions allow for such comparisons.

Figure 6 displays the $\beta_i(t)$ functions for the ΛCDM DEUSS simulation snapshots at redshifts $z = 0, z = 1$, and $z = 4$ (corresponding to present day, and roughly 8 or 12 billion years ago, respectively). That is, the $z = 4$ snapshot in Figure 6 is from the early (simulated) universe, whereas the $z = 0$ snapshot is analogous to the present day (and the $z = 1$ snapshot is somewhere in between).

Several things can be noted when comparing these different $\beta_i(t)$ functions from Figure 6. First, the Betti function values and peak heights are lower for the early universe snapshots ($z = 4$) compared to the later ones ($z = 0$ or $z = 1$) because the early universe contains fewer dark matter haloes. The locations of the $\beta_i(t)$ function peaks also differ between the $z = 4$ and $z = 0$ or $z = 1$ snapshots, but the $z = 0$ and $z = 1$ peaks have similar locations.

Consider Figure 6(c) to illustrate what may be happening here. The $\beta_2(t)$ functions at $z = 0$ and $z = 1$ are relatively similar, so we only need to focus on the differences between, say, the $z = 0$ and $z = 4$ $\beta_2(t)$ functions. Overall, fewer H_2 generators have formed by $z = 4$, but also, most of the H_2 generators formed and died off at higher filtration thresholds. The DTM filtration threshold, t, corresponds to an average comoving distance from a point to a fixed number of nearest-neighbor dark matter haloes.

At first glance, it may seem that this implies the early universe had larger voids, but this is because of the comoving distances: The universe is constantly expanding, so we often employ comoving coordinates, where distances are measured with respect to the expansion scale at present day. There were fewer dark matter haloes overall at $z = 4$, so the average comoving distance out to a fixed number of haloes is larger than at later times, when the overall halo clustering and abundance has strengthened. If the expansion scale factor was included, the $\beta_2(t)$ function for $z = 4$ would be shifted to the left of the $z = 0$ and $z = 1$ functions.

Concluding Remarks

Topological data analysis can be a useful framework for summarizing information in complicated data

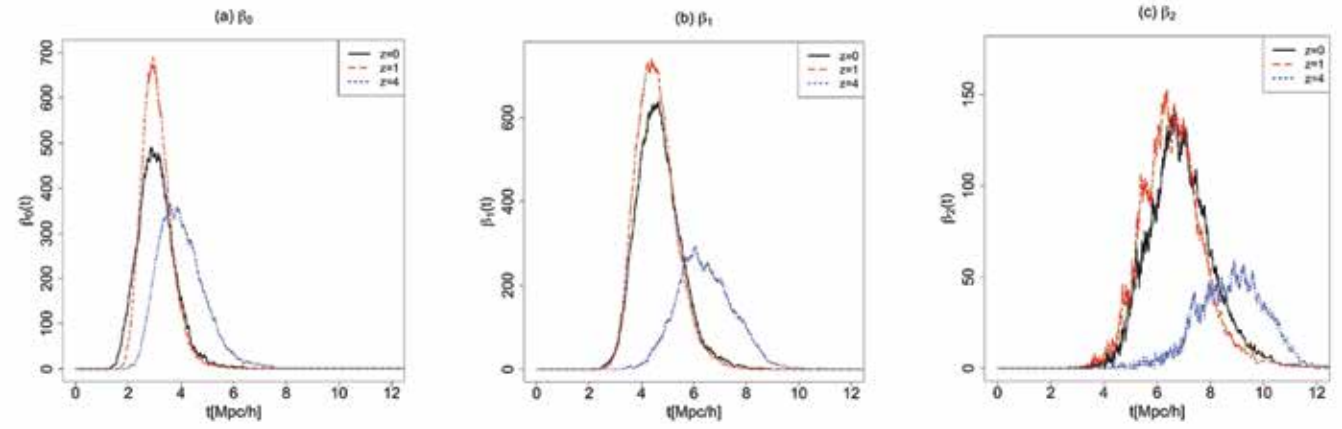


Figure 6. Plots of the (a) $\beta_0(t)$, (b) $\beta_1(t)$, and (c) $\beta_2(t)$ functions at redshifts $z = 0$, $z = 1$, and $z = 4$ (corresponding to present day, and roughly 8 or 12 billion years ago, respectively) of the 225 Mpc ΛCDM DEUSS simulation.

such as the large-scale structure of the universe. The persistent homology summaries can be used to pull potentially relevant information from the cosmic web, which can then be visualized with Significant Cosmic Holes in the Universe (SCHU) or analyzed quantitatively using functional summaries of persistence diagrams like the Betti functions.

When exploring simulation snapshots of the universe at different cosmic times, these functional summaries can help to highlight differences in structure formation. While TDA provides a set of tools natural for cosmology, other areas in science with data characterized by loops and holes may also find it useful.

Further Reading

Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. 2014. Confidence sets for persistence diagrams. *The Annals of Statistics* 42(6): 2301–2339.

Freedman, W.L. 2017. Cosmology at a crossroads. *Nature Astronomy* 1 (article id 0121).

Rasera, Y., Alimi, J-M., Courtin, J., Roy, F., Corasaniti, P-S., Fuzfa, A., and Boucher, V. 2010. Introducing the Dark Energy Universe Simulation Series. *AIP Conference Proceedings* 1241: 1134–1139. *www.deus-consortium.org*.

Springel, V., Frenk, C.S., and White, S.D. 2006. The large-scale structure of the Universe. *Nature*, 440(7088): 1137–1144.

van de Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B.J., Pranav, P., Park, C., Hellwing, W.A., Eldering, B., Kruithof, N., Bos, E.G.P., and Hidding, J. 2011. Alpha, betti, and the megaparsec universe: on the topology of the cosmic web. In *Transactions on Computational Science* XIV, 60–101. Springer-Verlag.

Xu, X., Cisewski-Kehe, J., Green, S.B., and Nagai, D. 2019. Finding cosmic voids and filament loops using topological data analysis. *Astronomy and Computing* 27: 34–52.

Wasserman, L. 2018. Topological data analysis. *Annual Review of Statistics and Its Application* 5: 501–532.

About the Authors

Jessi Cisewski-Kehe is an assistant professor in the Department of Statistics and Data Science at Yale University.

Sheridan B. Green is a PhD student in the Department of Physics at Yale University. He earned his BS in physics and mathematics in 2017 from the University of North Carolina. His research involves using numerical simulations to constrain cosmology and probe the nature of dark matter.

Abby Mintz is an undergraduate student studying astrophysics and statistics at Yale University. She has worked on projects on exoplanets, computational cosmology, and quasar absorption line spectroscopy.

Xin Xu is a PhD student in the Department of Statistics and Data Science at Yale University. She earned her BS in statistics in 2015 from Nankai University. She has worked on topological data analysis and astrostatistics.

Image courtesy of Getty Images

Mapping the Large-Scale Universe through Intergalactic Silhouettes

Collin A. Politsch and Rupert A.C. Croft

A majority of the atomic matter in the universe takes the form of a highly dilute gas that permeates the overwhelming volume of intergalactic space. Dubbed the *intergalactic medium*, this ubiquitous gas is too sparse to be observed directly, but its presence is embedded in the light of luminous background sources. Most notably, quasars—supermassive black holes that shine so brightly they can be seen billions of lightyears from Earth—allow scientists to indirectly study the structure of the intergalactic medium on vast cosmic scales.

As light travels from a distant quasar along its path to Earth, the intergalactic medium leaves an absorption signature in the light, marking the atomic elements that are present in the intervening intergalactic gas at each point along the light's path. This signature collectively reveals the presence of diffuse primordial hydrogen and helium residue in intergalactic space left over from the Big Bang, as well as a variety of metals occasionally ejected from galaxies by particularly forceful supernova explosions.

However, the bulk of the intergalactic medium is composed of electrically neutral hydrogen gas, which marks its presence by absorbing a very specific wavelength of ultraviolet light: the so-called Lyman-alpha transition at ~1,216 Angstroms.

For this reason, the *Lyman-alpha forest*—the series of absorptions

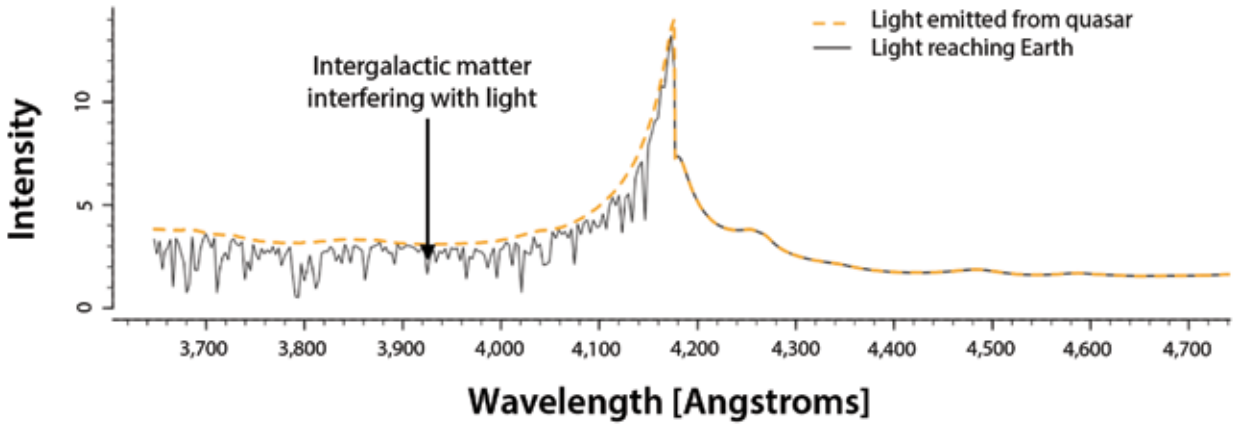


Figure 1. Simulated electromagnetic spectrum of a quasar ~10.9 billion lightyears from Earth. The orange dashed curve shows the intensity of light at each wavelength at the time that it was first emitted by the quasar and the solid black curve shows the spectrum as observed from Earth, after neutral hydrogen absorption. The absorptions appear at many different wavelengths because of the continuous stretching of light waves traversing intergalactic space caused by the expansion of the universe (Bautista, et al. 2015).

originating from the Lyman-alpha transition—is the richest source of data for scientists to study the large-scale structure of the intergalactic medium.

The original detection of the Lyman-alpha forest dates back to 1970, but scientific studies of the forest did not mature until the early 1990s, with the advent of high-resolution spectrometers—an instrument that connects to a telescope and records the intensity of the observed light as a function of its constituent wavelengths. In particular, the commissioning of the High Resolution Echelle Spectrometer (HIRES) on the Keck telescope in Mauna Kea, Hawai'i, marked the beginning of the golden age of Lyman-alpha forest cosmology.

Figure 1 shows a simulated electromagnetic spectrum of a quasar ~10.9 billion lightyears from Earth. The orange dashed curve shows the intensity of the light at each wavelength when it was emitted from the quasar and the black (wiggly) curve shows the intensity of the light as viewed from Earth.

The decrease in the intensity of the quasar's light in this region of the electromagnetic spectrum is due to intergalactic neutral hydrogen gas partially absorbing the light passing through it. This phenomenon is akin to observing a distant lighthouse through a patchy cloud of fog. When seen through a dense patch, the light is dim. When seen through a relatively thin patch, the light is bright. The intergalactic "fog" in this case is too tenuous to be observed directly, but studying its matter density distribution is fortuitously made possible by analyzing the fraction of light that is absorbed along its journey to Earth.

The fraction of absorbed light is nonlinearly related to the neutral hydrogen density, but the relation is monotonic, which allows the absorption fraction to serve as a suitable proxy for the neutral hydrogen density.

In practice, the intensity of the unabsorbed light originally emitted from the quasar (orange curve) is not known, and accurately estimating this curve represents a crucial statistical step in any Lyman-alpha forest analysis. The literature about this topic is extensive, but popular approaches include principal components analyses over a set of candidate functional shapes, interpolation of observed regions deemed to have minimal absorption, and low-order smooths of the observed spectrum subsequently scaled to match the cosmic mean absorption fraction.

As previously stated, neutral hydrogen gas absorbs at a fixed wavelength of ~1216 Angstroms. Figure 1 should therefore provoke a couple of questions, such as *Why do the absorptions appear at many different wavelengths?* and, *Why are those wavelengths significantly longer than 1,216 Angstroms?*

The answer to both questions is that the universe is expanding. When the universe began with the Big Bang ~13.7 billion years ago, it was not only all matter contained in the infinitesimal speck

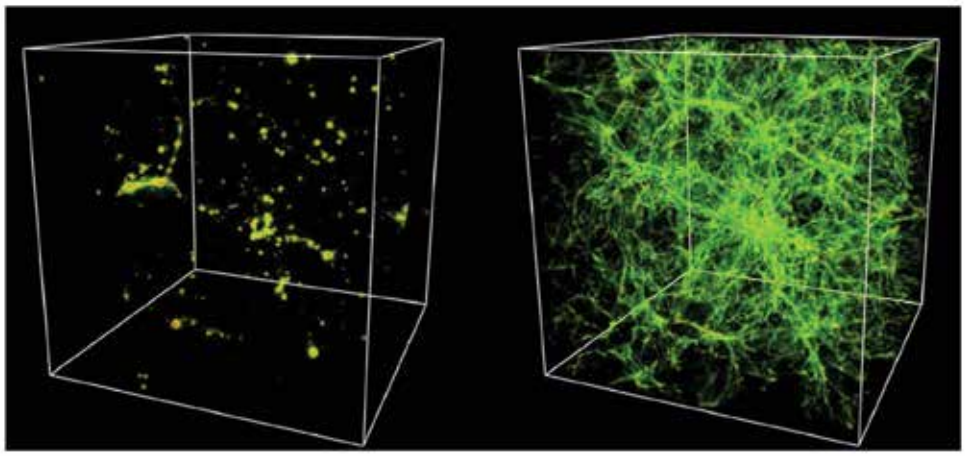


Figure 2. Simulated cubic volume of the universe's large-scale atomic structure (side length ~326 million lightyears). The distribution of galaxies in the universe (left) can be thought of as a sparse point process, while the intergalactic medium (right) has a continuous matter distribution permeating the entirety of intergalactic space. (Image: Cen and Ostriker. 2006)

that began an outward trajectory, but the fabric of space as well. Ever since that moment, space has continued to expand.

This expansion leads to a fascinating phenomenon known as *redshift*. Namely, when light travels through intergalactic space the expansion of the space, it is traveling through effectively causes the light itself to be continuously stretched to longer wavelengths.

For a (fixed) absorption line such as Lyman-alpha, the consequence is that a *forest* of absorptions becomes inscribed in the light's spectrum similarly to how a seismograph operates. Although the pen of the seismograph is stationary, the paper continuously moves underneath it, recording the amplitude of the pen's oscillations as a one-dimensional curve. Here, redshift is the mechanism that *moves the paper*, and the resulting forest effectively provides a one-dimensional map of the neutral hydrogen density along the entire path from Earth to the quasar, with longer wavelengths corresponding to the neutral hydrogen density at more distant points along the path.

Another important observational phenomenon arises due to the vast distances quasars are observed at and the finite speed at which their light travels to us. Namely, the light we observe from quasars is actually extremely old; its age is directly related to the distance it traversed on its path to Earth.

For example, an observed spectrum of a quasar 10.9 billion lightyears from Earth—such as that displayed in Figure 1—is in fact a *picture* of what that quasar looked like 10.9 billion years ago—6.4 billion years before the Earth even formed (with the various absorption lines being imprinted at different times during that period)! Mapping the matter distribution of the intergalactic medium via Lyman-alpha absorptions in spectra of distant quasars is therefore akin to charting how the adolescent universe evolved into its present-day form.

The past decade of Lyman-alpha forest observational cosmology has progressed to a point where it is now, in principle, possible to use the aggregate of currently available data to statistically reconstruct a continuous large-scale map of the intergalactic medium in all three spatial dimensions. Mapping the intergalactic medium in three dimensions can be viewed as an extension of charting the locations of the galaxies that the intergalactic medium envelopes, in the sense that the distribution of galaxies (a point process) and the distribution of intergalactic gas (a random field) together provide a complete picture of how atomic matter is distributed throughout the universe.

Scientifically, a large-scale intergalactic medium map would allow for testing cosmological models in an entirely new regime of both scale and age of the universe, potentially leading to a more-refined understanding of the parameters that govern the Big Bang cosmological model. Moreover, the scale and holistic nature of such a map could be used to locate never-before-seen regions

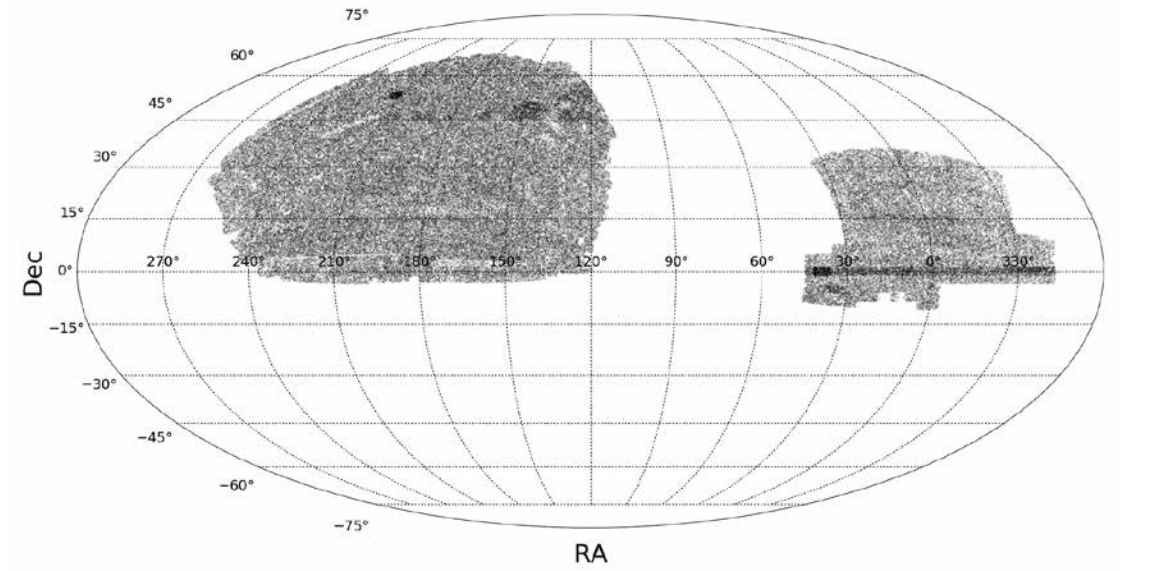


Figure 3. The locations of the ~300,000 quasars observed by the Baryon Oscillation Spectroscopic Survey (BOSS). The BOSS final data release is the most-prolific quasar catalog to date and constitutes a ~25% coverage of the sky (Alam, et al. 2015).

of the universe that would be interesting to subsequently revisit and gather more detailed observations.

A simulated illustration of the universe's large-scale structure is depicted in Figure 2, via both its galactic point process (left) and its intergalactic random field (right). Both processes arise as a result of gravity and dark energy acting on the primordial matter fluctuations left over from the nearly perfect, scale-invariant Gaussianity of the Big Bang, as revealed by the Cosmic Microwave Background radiation (CMB). The mutual gravitational forces of the matter have subsequently pulled the universe's large-scale structure into a highly non-uniform weblike distribution—appropriately dubbed the *cosmic web*. Dark matter—matter that (thus far) only reveals its presence by way of its gravitational interaction with regular matter—is also a central acting force in the Big Bang cosmological model and hydro-dynamical simulations suggest it also possesses a weblike distribution closely tracing that of the intergalactic medium and the galaxies.

The Lyman-alpha forests of quasars—and to a lesser extent, luminous star-forming galaxies—currently serve as our only window into glimpsing intergalactic neutral hydrogen. Therefore, our ability to produce high fidelity maps of the intergalactic medium is intrinsically tied to how densely these objects populate the sky at each radial distance and how effective our telescopes are at detecting them.

From 2008 to 2014, the Baryon Oscillation Spectroscopic Survey (BOSS)—part of the Sloan Digital Sky Survey at Apache Point Observatory in Sunspot, NM—collected spectra for ~300,000 quasars. This is currently the most-prolific quasar catalog to date, and has been used to show that the structure of the intergalactic medium is consistent with Einstein's general theory of relativity and the standard model of Big Bang cosmology (ΛCDM). Figure 3 shows the locations of the BOSS quasars on the celestial sphere and Figure 4 shows them as a function of their distance from Earth (in red).

The BOSS catalog constitutes a coverage of approximately 25% of the sky and, in principle, allows for a three-dimensional statistical reconstruction of the intergalactic medium over the radial range of 10.4 to 11.7 billion lightyears—*in principle* because the statistical modeling step of producing a high-fidelity three-dimensional map from a Lyman-alpha data set of this magnitude is quite nontrivial, and therefore has not yet been accomplished.

This task poses a unique statistical challenge for many reasons, but foremost among them are: the complex spatial dependence of the underlying matter distribution, the associated computational cost of

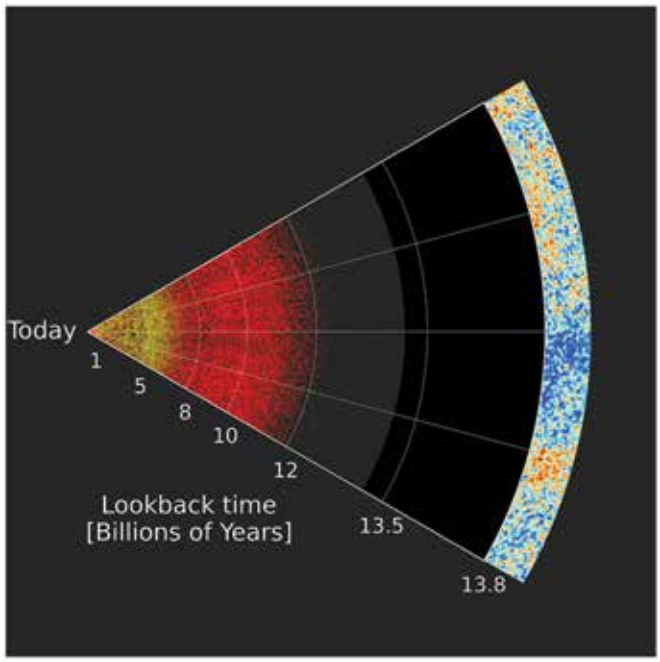


Figure 4. A slice of the radial distribution of observed BOSS sources, in terms of the lookback time to the object (analogously, the distance the light traveled to reach Earth). Earth is at the left vertex and the red pixels signify quasars observed by BOSS, while yellow pixels signify the locations of galaxies. The solid-black region represents the so-called Dark Ages between recombination and reionization when stars and galaxies had not yet formed. (Image: Anand Raichoor and the SDSS-IV/eBOSS collaboration. *www.sdss.org*.)

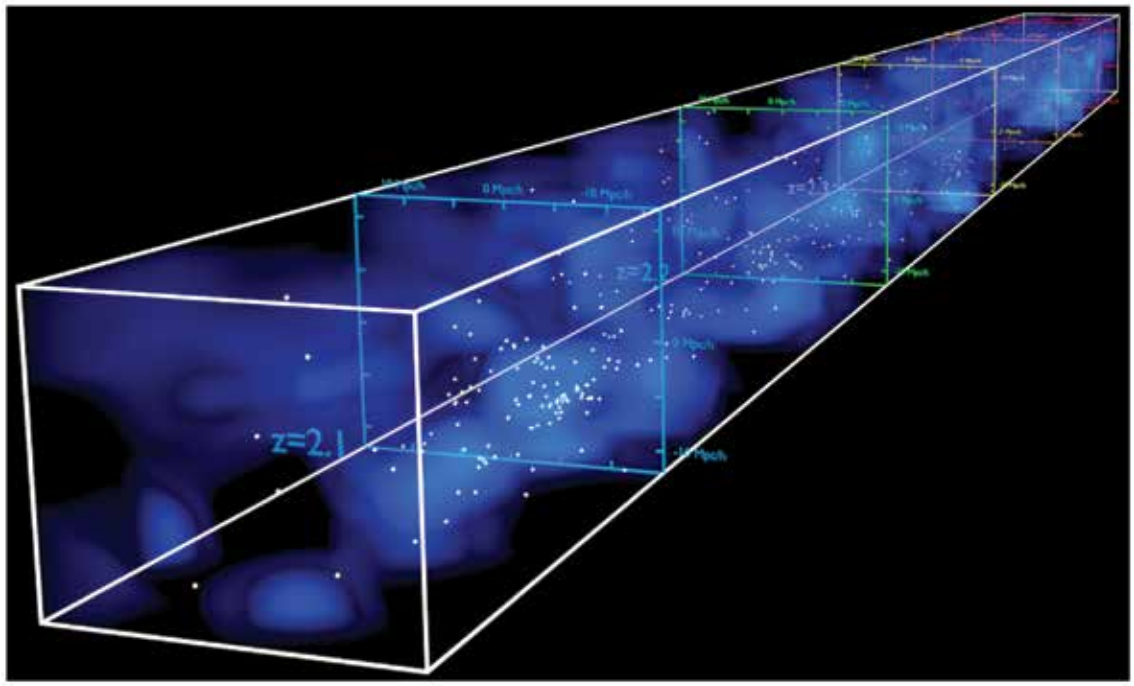


Figure 5. High-resolution 3D reconstruction of the intergalactic medium absorption field obtained via a Wiener filter on a set of 240 spectra of star-forming galaxies densely populated on a patch of sky with an area of 0.157 deg^2 and a radial distance range of ~10.6 billion lightyears to ~11.3 billion lightyears. The locations of the galaxies are indicated by the white pixels.
(Image: Khee-Gan Lee and the CLAMATO collaboration; 3D video: *www.youtube.com/watch?v=QGtXi7P4u4g*)

suitable spatial methods on such a large scale, and the highly unusual sampling design characterized by closely sampled observations along a collection of sparsely sampled line segments originating from a common vertex in three-dimensional space.

A number of methods have been proposed for producing point estimate three-dimensional reconstructions, with Wiener filtering currently the most popular among them. However, the computational cost of proposed methods has thus far limited the reconstructed intergalactic medium maps to small fractions of the total volume for which Lyman-alpha data are available.

Figure 5 displays a Wiener-filtered three-dimensional reconstruction of the intergalactic medium within a cylindrical skewer covering a 0.157 deg^2 patch of the sky and spanning a radial distance range of ~10.6 billion lightyears to ~11.3 billion lightyears (K.G. Lee, et al. 2018). This particular map was not made with BOSS data, but rather a densely sampled collection of star-forming galaxy spectra collected by the recently commissioned Cosmos Lyman-alpha Mapping and Tomography Observations (CLAMATO) survey.

Certainly, the greatest challenge in this sort of mapping—as is the case in the majority of statistical analyses—is not in producing the point estimate itself, but rather, accompanying the point estimate with reliable inference. *For any given structure appearing in the estimated map—e.g., a galaxy supercluster or a cosmic void—what is the probability that the structure indeed exists? Is there enough signal in the map to detect significant cross-correlations with other potential tracers of structure, such as the anisotropies of the CMB? Does the established inference adequately account for the uncertainty introduced by the sequential observational pipeline of analyses the data have already undergone?*

All of these are exceedingly difficult unanswered questions that rely on the development of rigorous statistical methods.

Beyond the groundbreaking work of the BOSS collaboration, the rapid influx of Lyman-alpha data is assured to continue into the foreseeable future. The next generation of the collaboration, the *extended* BOSS survey (eBOSS) began gathering data immediately after completing the previous phase, and is still carrying out its objective of adding ~500,000 new quasar spectra to its predecessor's catalog.

The Dark Energy Spectroscopic Instrument (DESI) Survey conducted on the Mayall 4 meter telescope at Kitt Peak National Observatory in Arizona is set to begin a five-year data collection later this year, with a targeted sky area of 14,000 deg^2—approximately 34% sky coverage.

The CLAMATO survey using the Low Resolution Imaging Spectrograph (LRIS) on the Keck-I telescope at Mauna Kea, Hawai'i, is actively collecting spectra of bright star-forming galaxies over comparatively smaller regions than BOSS, eBOSS, and DESI, but with a much-more densely populated target selection in the sky, allowing for higher resolution mappings. Moreover, although not a Lyman-alpha survey, the Square Kilometer Array (SKA) has begun construction on the world's largest radio telescope, with which it will pioneer the collection of observational data from the so-called Dark Ages (shown in solid black in Figure 4), using an entirely different region of the electromagnetic spectrum—the 21 cm radio emission. Mapping this completely uncharted epoch will almost surely rely heavily on the same variety of statistical methods developed for Lyman-alpha forest analyses. C

Further Reading

Alam, S., et al. 2015. The eleventh and twelfth data releases of the Sloan Digital Sky Survey: final data from SDSS-III. *The Astrophysical Journal Supplement Series* 219(1): (27 pp.).

Cen, R., and J.P. Ostriker. 2006. Where are the Baryons? II. Feedback Effects. *The Astrophysical Journal* 650(2):560-572.

Cisewski, J., et al. 2014. Non-parametric 3D map of the intergalactic medium using the Lyman-alpha forest. *Monthly Notices of the Royal Astronomical Society* 440(3):2599-2609.

Croft, R., et al. 2000. Towards a Precise Measurement of Matter Clustering: Lyα Forest Data at Redshifts 2-4. *The Astrophysical Journal* 581(1):20-52.

Lee, K.G., et al. 2018. First Data Release of the COSMOS Lyman-Alpha Mapping and Tomography Observations: 3D Lyman-α Forest Tomography at $2.05 < z < 2.55$. *The Astrophysical Journal Supplement Series* 237(2):(31 pp.).

McQuinn, M. 2016. The Evolution of the Intergalactic Medium. *Annual Review of Astronomy & Astrophysics* 1:1–55.

About the Authors

Collin A. Politsch is a PhD candidate in the Department of Statistics & Data Science and the Machine Learning Department at Carnegie Mellon University. His research interests include spatio-temporal statistics, uncertainty quantification, and applications to astronomy and cosmology.

Rupert A.C. Croft is a professor of physics in the McWilliams Center for Cosmology, which is part of Carnegie Mellon University in Pittsburgh. He is a computational cosmologist and astrophysicist. He studies structure in the Universe on the largest scales, seeking to understand the cosmic mysteries of dark matter and dark energy.

Just How Far Away is that Galaxy, Anyway?

Estimating Galaxy Distances Using Low-Resolution Photometric Data

Peter E. Freeman

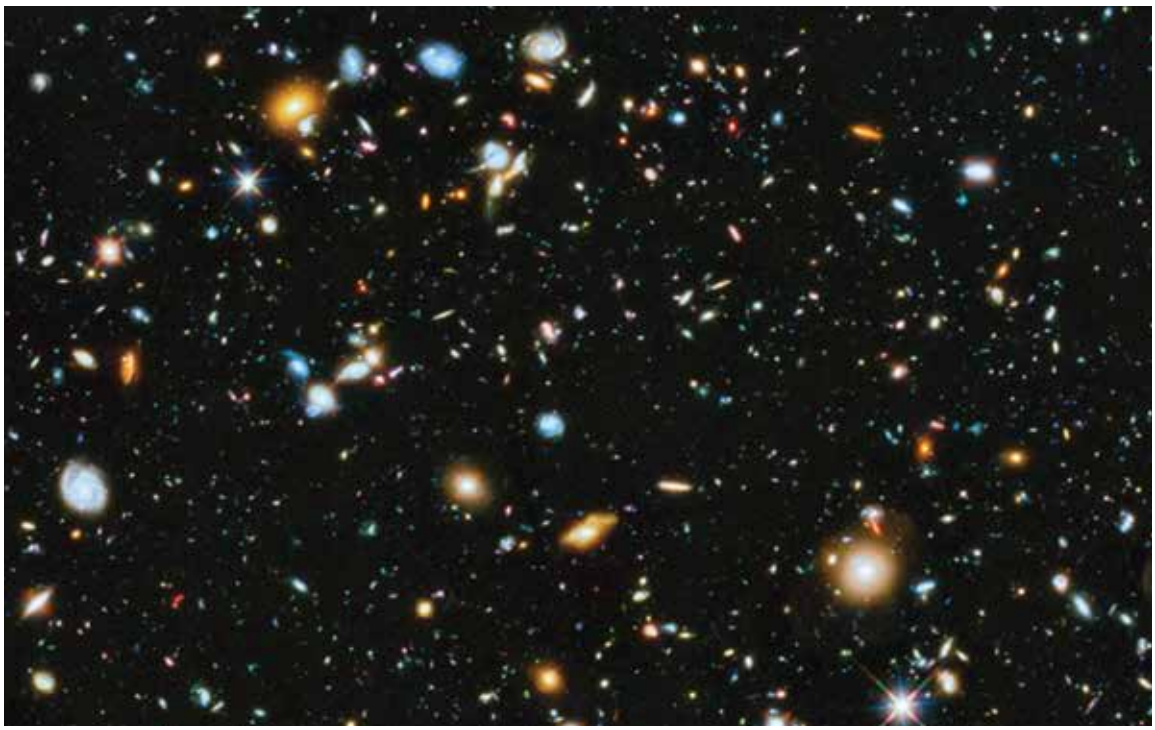


Figure 1. A portion of the night sky as observed by the Hubble Space Telescope. The majority of the objects in this image are galaxies, and for most, the image itself is the only information we have about them. How do we figure out how far away these galaxies are? (NASA Astronomy Picture of the Day. *https://apod.nasa.gov/apod/ap140605.html*)

How do we make sense of the observable universe? It is an incomprehensibly vast (and mostly empty) space containing perhaps hundreds of billions of galaxies. Luckily for us, it is also just simple enough that astrophysicists can create realistic ensembles of mock galaxies by using codes that run for months on each of thousands of nodes in a computer cluster.

Unlike real galaxies, we can trace these simulated galaxies through time, from their formation soon after the Big Bang to now. These ensembles allow us to make statistical inferences about the parameters that dictate the initial conditions and large-scale evolution of the universe, as well as about the smaller-scale physics of galaxy formation and evolution. However, to make such inferences, we need observed data.

Specifically, given images of galaxies (see Figure 1), we need to figure out how far away they are. Since the speed of light is finite, if we know the distances, we can place the galaxies in time: The farther away they are, the younger the universe was when they emitted the light that can be observed today. Once we have a set of simulated and observed galaxies from the same epoch of the universe…inference ensues.

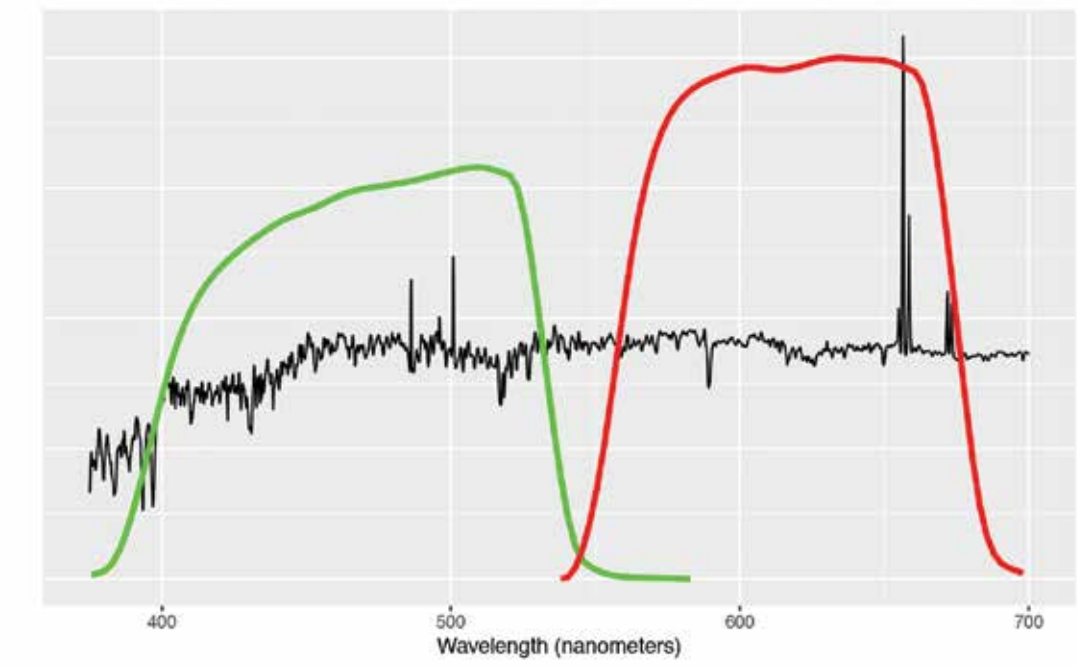


Figure 2. A theoretically generated, high-resolution galaxy spectrum (black) overlaid with two filter transmission curves (Sloan Digital Sky Survey green and red filters). For approximately 1% of the galaxies in the SDSS catalog, there are data akin to the black curve, whose spikes make redshift estimation relatively easy. For the other 99%, we have only photometric data. For instance, the photometric flux through the green filter is the summation over wavelength of the product of the spectrum and the green filter transmission curve. For SDSS, there are five filters (an ultraviolet filter to the left of green, green, red, and two infrared filters to the right of red: In photometry, all the detail in spectra is summarized in just five numbers. (*http://classic.sdss.org/dr5/algorithms/spectemplates*)

But, as is common in life, the devil is in the details. Just how do we figure out how far away galaxies are?

Before diving into these details, it is useful to talk about the expansion of the universe and an associated quantity: *redshift*. Since the Big Bang, the universe has been expanding. At first, the rate of expansion slowed with time, but now the rate of expansion is increasing with time, due to the influence of dark energy, a theoretically still-unexplained phenomenon that appears as a form of anti-gravity. Photons emitted by astronomical sources have wavelengths that are tied to the scale of the universe.

As the universe expands, the wavelengths get longer. In terms of the light that can be seen around us, it is as if it becomes less blue and more red; hence the historical term "redshift." While we cannot lay down a tape measure to determine the distance to a galaxy, we can directly infer its redshift with the right observations and a little bit of luck. Given the redshift and some assumptions about cosmology, we can finally infer the galaxy's distance.

What are the right observations? Figure 2 presents a theoretically generated galaxy spectrum that shows the intensity of light given off by the galaxy as a function of wavelength (400 to 700 nanometers—the range for which human eyes are most sensitive). While the actual details are, of course, more complicated, this spectrum can be considered as what we observe when we take the galaxy's light and pass it through a prism.

The spectrum in Figure 2 exhibits three features: a smooth portion dubbed the continuum; dips relative to the continuum that represent absorption lines; and spikes above the continuum that represent emission lines. A "line" is an indicator of a chemical transition, with particular transitions mapping to particular wavelengths.

For instance, the tallest spike in Figure 2 represents the so-called Hα transition at 656 nm, caused when electrons fall from the third energy level of hydrogen to the second energy level. (A dip at this wavelength would occur if, for instance, electrons moved *up*

from the second to third energy levels, because to have that happen the atoms would have to take in photons with wavelength 656 nm, leaving relatively fewer photons of that wavelength to travel to us.) The other observed dips and spikes are related to known transitions of electrons in hydrogen, nitrogen, oxygen, and sulfur atoms, all of which are present in the gas and stars of galaxies.

If we observe two or more lines in a spectrum, then redshift estimation is easy, because we can compare the ratios of wavelengths of observed lines to known ratios and thus identify the atomic transition responsible for each line individually. Once we know, for instance, that a particular line observed at 750 nm is Hα, we compute the ratio between that wavelength and the true wavelength of Hα (656 nm) and set that ratio equal to $1+z$, where z symbolizes redshift; here, $z = 0.143$.

To return to the question of the right observations: The right ones are those that collect spectra with sufficiently high resolution to resolve two or more individual dips and spikes. Saying that we also need a "bit of luck" refers to the fact that not all galaxies have obvious dips and spikes, although most do.

You ask, "Well, what's the issue here? Take spectra, find dips and spikes, infer redshift. Easy!"

The issue is that collecting spectral data is a time-intensive exercise. A typical SDSS spectrum shows the intensity of observed light in 3,500 separate bins, each of which is associated with a different wavelength. To have a sufficient number of photons land in each bin to clearly differentiate a galaxy from other sources in the sky around it might require observing that galaxy for hours. Even with clever spectrographs that observe hundreds, if not thousands, of galaxies at once, there simply is not enough time to record spectra for all of them. More than 200 million galaxies have been observed in images for SDSS in particular, but only for some 2 million do we have redshifts inferred from high-resolution spectra.

You may ask, "Is all hope lost?" The answer is an unequivocal no. Much statistical information is present in the 99% of galaxies without spectra, information astronomers and cosmologists would love to lay their hands on—and we can collect that information using *photometry*: the action of observing the same galaxy repeatedly with different filters in place, with each observation yielding a *magnitude*, a logarithmic measure of brightness that is believed to have first been used by the ancient Greek astronomer Hipparchus.

When the SDSS telescope observes the sky and takes images (instead of spectra, because sometimes it takes images, and sometimes it collects spectra), it does so with one of five filters placed in the light path. A given filter is made of a material that lets light of certain wavelengths pass, while absorbing light at other wavelengths. Overlaid on Figure 2 are the filter transmission curves for SDSS's green and red filters. The heights of these curves show the relative transmission efficiencies of each filter as a function of wavelength. (SDSS's three other filters include one in the ultraviolet regime, "to the left" of green in wavelength, and two in the infrared, "to the right" of red.)

The green filter is most-transparent around 500 nm, and totally opaque at 600 nm. The brightness of a galaxy observed by SDSS can thus be summarized with five numbers: one magnitude for each filter. One may think of photometry as very-low-resolution spectroscopy in which any dips and spikes that may be present are smeared out. Because of this smearing, one cannot simply look at the five numbers and infer a redshift by eye.

That's where the computers come in. Over the last two decades, astronomers and statisticians have developed a myriad of algorithms to solve the so-called "photo-z problem": predicting the redshifts of photometrically observed galaxies. We can split this myriad into two general classes, *template-based algorithms*, which compare idealized spectra to the observed data, one galaxy at a time, and so-called *empirical algorithms*, which learn statistical models relating magnitudes to redshifts given ensembles of training data.

A *template* is an artificially constructed, noiseless galaxy spectrum. Astronomers understand the physics of stars, and thus can construct accurate stellar spectra. Galaxy templates are, more or less, the cumulative spectra that result when adding stars together in different proportions. ("This template has lots of older and lighter stars and few younger and heavier stars; this other template has more younger and heavier stars; and this other template has relatively more gas, and …")

To estimate redshift, one selects a template, redshifts it by some amount (think of taking the black curve of Figure 2 and shifting it to the right by some amount, while keeping the green and red curves in place), passes the redshifted spectrum through photometric filters to get five magnitudes, and then sees how well those five numbers match those observed for the galaxy in question.

The quality of match is usually measured via the normal, or Gaussian, likelihood function, with the errors in the likelihood function provided by the uncertainties in the observed magnitudes. Different templates and different

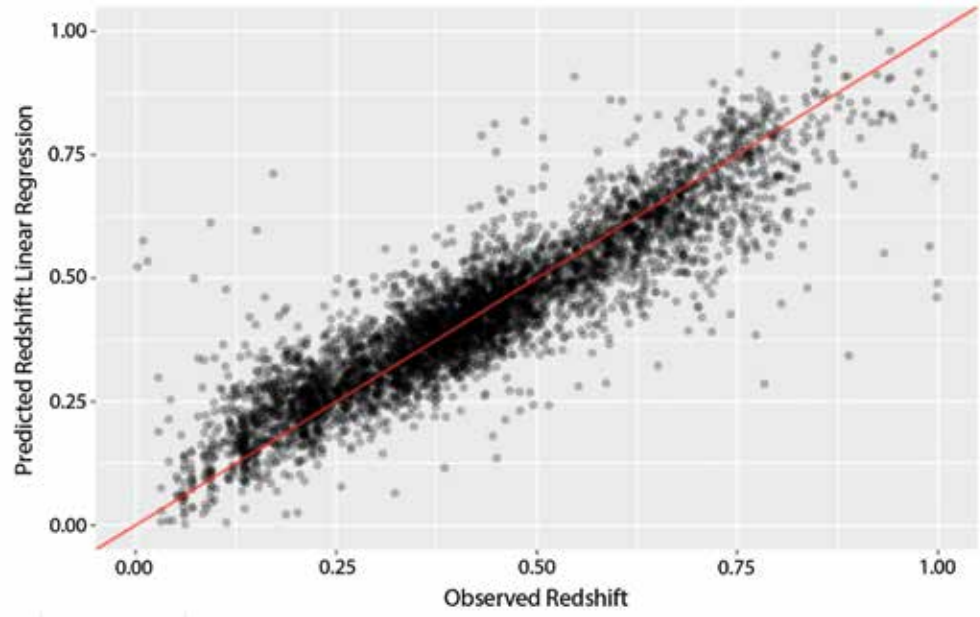


Figure 3. Estimated redshifts for 5,000 galaxies versus their observed redshifts. Multiple linear regression is used to relate five magnitude estimates for each galaxy to its observed redshift. (The adjusted R2 is 0.866.) The red line is the locus of "perfect estimation." (Data: Jeffrey Newman and Rongpu Zhou of the University of Pittsburgh.)

redshifts are used within a grid search until the template/redshift combination that maximizes the likelihood is found.

Template-based methods have the virtue of being straightforward to implement while not requiring any training data. However, they do suffer from a number of issues.

- Model misspecification: Do we have a directory of template spectra that spans all possible spectra?
- Computational intensity: If we have hundreds of template spectra in a directory, we cannot apply them all quickly when fitting any single galaxy. How do we choose the right subset of templates?
- Discreteness: Templates represent a discrete set of idealized galaxies. What if the best spectrum for a given galaxy is a linear combination of two, three, four, or more templates?

To avoid these issues, many astronomers turn to what they call empirical methods to estimate redshifts. In other words, instead of assuming complete knowledge of galaxy spectra, we let the data themselves inform the modeling process. "Empirical" is astronomer-speak for learning statistical models that relate a set of predictor variables (e.g., the magnitudes; alternatively, the *colors*, or differences in magnitudes, for each galaxy) to a response variable (the redshift).

Figure 3 displays the result of learning a multiple linear regression model that relates sets of five magnitudes to the known redshifts for 5,000 spectroscopically observed galaxies. The x- and y-axes in this figure represent the observed and predicted redshifts, respectively, and the diagonal red line is the line of "perfect estimation."

Because the points largely follow the line, with some amount of intrinsic scatter (adjusted R^2 0.866, RMSE 0.072), it can be inferred that a linear model does a good job of representing the underlying association between magnitudes and redshift. (There are issues, such as a substantial number of poorly estimated redshifts, that astronomers dub "catastrophic outliers," and the fact that predictions are on average too high at low redshift and too low at high redshift, due to measurement error in the galaxy magnitudes. This latter issue is dubbed *attenuation bias*. On the whole, though, it is fair to say that the linear model represents the underlying association well.)

Once again, we reach what we think is a stopping point: "Even if you do not have spectra, you can

apply multiple linear regression to photometric data and do a good job of predicting redshift. The end. Right?" No, not quite so fast.

Look back at Figure 1 and imagine that you are the astronomer. The image that you see is filled with galaxies that you would love to know the distances to, but you know that you have limited telescopic resources and thus cannot measure spectra for all of them. What will you do?

Chances are, you will focus your energies on the brighter galaxies, because they are the ones that have to be observed for the least amount of time to get a good spectral signal relative to the background noise. But galaxies that are brighter are generally galaxies that are either more massive or closer to us, or both, than the average galaxy in the image. Choosing the brighter galaxies for subsequent spectroscopic observations introduces *selection bias* into the redshift estimation process.

The brighter galaxies are not a random sample from the population of all the galaxies seen in Figure 1, and thus any model that is learned using these galaxies cannot necessarily be applied to the others. Stated another way, to apply your models to fainter galaxies, you would have to extrapolate them and, as we learn early in our statistics education, extrapolation is A Bad Thing to Do.

How best to deal with selection bias is an open research question.

One possibility is to mitigate its effects by applying weights to the data used to train a regression model. Let's say we have a training-set galaxy for which we have five magnitudes and a spectroscopically measured redshift. If there are many other galaxies without measured redshifts in the sample with similar magnitudes, we would give the training-set galaxy more weight when learning models, but while reweighting improves the quality of predictions for galaxies that have similar properties as the training-set galaxies, we still cannot (or, at least, should not) extrapolate our models.

Two other possibilities are data augmentation and active learning. The first involves increasing the amount of data used to train a model by resampling from the training data themselves, making use of the measurement error in their magnitudes, while the second involves increasing the amount of training data by collecting new spectra from judiciously selected galaxies—ones that give the most "bang" (increased accuracy and precision of estimates for fainter galaxies) for the "buck" (the amount of telescope time involved collecting new spectra).

Both techniques expand the applicability of a regression model by expanding the domain of the predictor variables. They will become increasingly vital in the upcoming era of the Large Synoptic Survey Telescope (LSST), when the number of photometrically observed galaxies will increase by a factor of 100 with no commensurate increase in the number of galaxies with measured spectra.

Now that we have brought up LSST, we can step back to the idea that all we have to do is apply multiple linear regression to our magnitude and redshift data to create a good statistical model. The data shown in Figure 3 lie between redshifts 0 and 1, where redshift 0 is now and redshift 1 is equivalent to a light-travel time of nearly 8 billion years. (Redshift and light-travel time are directly but nonlinearly related: Redshift 2 gets us to nearly 10.5 billion years in the past, while redshift 3 is just over 11.5 billion years. In fact, the redshift diverges to infinity as we get closer and closer to the Big Bang some 13.7 billion years ago.)

LSST will sample data from well beyond a redshift of 1. When we analyze those data, we will run into two issues: Linear models will no longer work so well, and we will observe degeneracies—sets of magnitudes that can plausibly map to more than one redshift. (We will need computationally efficient algorithms that we can apply to the tens of billions of galaxies that LSST will observe, but that is an issue for another time and another place.)

Mitigating the first issue is relatively straightforward, since we can expand our suite of possible regression models to include nonlinear ones such as random forest and extreme gradient boosting. However, these models are still, in the end, regression models: Their aim is to estimate the average value of the redshift for a given set of magnitudes. When degeneracy is present, that average may not be meaningful (see Figure 4).

To understand how to tackle degeneracies required introducing the idea of conditional density estimation (CDE). A conditional density is a probability density function of the form $f(z|\boldsymbol{x})$, where z is the redshift and $\boldsymbol{x}$ represents all the predictor variables (e.g., the galaxy magnitudes or colors). The relationship between CDE and regression is that in the former, we estimate the function $f(z|\boldsymbol{x})$ directly while making no assumptions about its shape, while in the latter, we estimate the *expected value* of z while assuming a function shape (e.g., a normal distribution with standard deviation σ, the typical assumption of linear regression).

Figure 4 shows an example of a conditional density estimate made using an extreme gradient boosting-based algorithm applied to simulated data from LSST's Data Challenge 1 (DC1). The estimate indicates that given its

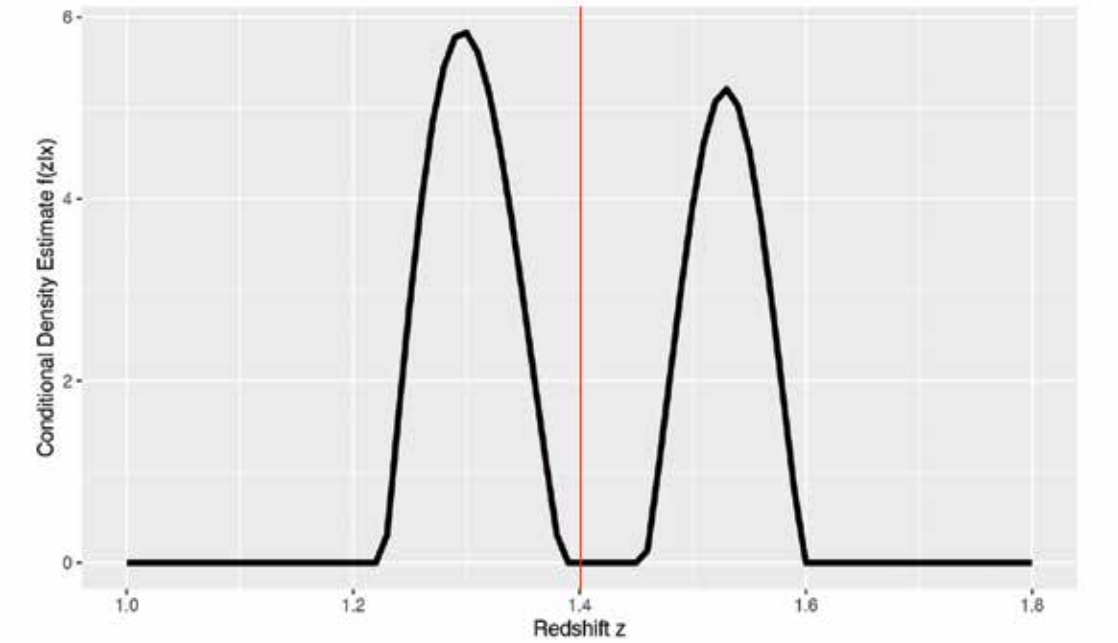


Figure 4. Example of a conditional density estimate (CDE) made by fitting an extreme gradient boosting-based estimator to data simulated for the Large Synoptic Survey Telescope Data Challenge 1 (LSST DC1). There are two viable redshifts associated with input predictor variables x (e.g., galaxy magnitudes). The regression estimate is the average of the CDE and is shown in red. It lies between the two modes and thus is not meaningful.

observed data, there are two viable redshifts for the galaxy in question: It may be a fainter, closer-to-us galaxy at a redshift of approximately 1.31, the redshift of the left mode, or a brighter, farther-from-us one at approximately 1.54, the redshift of the right mode. Note the red line. Learning a regression model and applying it to this galaxy would predict a redshift of approximately 1.4—a value not associated with either mode.

The best empirical models, thus, are computationally efficient and produce conditional density estimates while dealing properly with selection bias (if we do not take care of that via data augmentation or active learning approaches). But where do we go from here?

There are many avenues to explore. Perhaps you will walk along these avenues in the future.

The first avenue uses more information than only magnitudes or colors. For instance, the galaxy images themselves contain a wealth of information. However, working with images is difficult. An example is one of the relatively large spiral galaxies shown in Figure 1.

The first step is to extract a small, "postage-stamp image" that is centered on just that galaxy—an image that might have 64 pixels on a side. The galaxy would thus represent a single point in a 4,096-dimensional space.

Due to the so-called "curse of dimensionality," it is not feasible to analyze galaxy data in spaces with thousands of dimensions. We need to greatly reduce the dimensionality. One way to do this is to compute summary statistics for the galaxy that encode how concentrated its light is, how clumpy it appears to be, etc. (Those versed in machine learning would call this "extracting image features.") We can apply these summary statistics, which conventionally number around five to 10, as additional predictor variables in redshift estimation.

Another way to reduce dimensionality is to let an algorithm do it for us, by feeding images directly into a convolutional neural network (CNN) that extracts the image features on the fly. While the use of deep learning is intriguing and has the possibility to be a game changer in the era of LSST, it may be limited by a simple lack of training data. There are still only so many galaxies for which we know the redshift, and while that number will increase in the future, it will do so far more slowly than the overall number of observed galaxies.

Traveling along the first avenue, there is more information than just magnitudes and galaxy appearance. For instance, one could try to mine information about a galaxy's environment. If a galaxy sits in a part of space where the overall density of galaxies appears significantly higher than average, then that galaxy may be part of a galaxy cluster, and information about that cluster could be used to estimate the galaxy's redshift. If we have spectra for one or more galaxies in the cluster, then that provides, in theory, approximate redshifts for all galaxies in the cluster.

The key word above is "may," because even if a galaxy appears to lie in a high-density environment, it may be a chance superposition: The galaxy might lie far behind the cluster or far in front of it. At best, we can use the probability of cluster membership to inform CDE.

The second avenue to explore is not so much about how one estimates redshift, but whether redshift estimation can be combined with the estimation of other galaxy properties. Astronomers call this the "$p(z,\alpha)$ problem," but given the notation used above, what they really mean is that they want to estimate $f(z,\alpha|\boldsymbol{x})$, where α collectively denotes one or more of those other properties, such as mass or star-formation rate.

Currently, astronomers use physics-based codes to estimate these properties, so estimation is computationally expensive and prone to model misspecification (i.e., incomplete physics). In the upcoming era of LSST, it would be nice to bypass further implementation of such codes and use learned statistical models instead that are trained on galaxies for which we currently have property estimates.

Learning useful models will be tricky, though, because of estimator uncertainty. Spectroscopic redshifts may reasonably be viewed as "ground truth" because the "spiky" nature of the emission and absorption lines used to estimate them makes the redshift uncertainty generally negligible. However, the same cannot be said for other galaxy properties. For instance, short of putting a galaxy on a scale, any mass estimate with have significant uncertainty, both statistical and systematic. How does one properly estimate and then propagate uncertainties in derived galaxy properties to achieve optimal multivariate conditional density estimates? That is an open question awaiting a solution.

Selection bias, gathering more-informative predictor variables, the quantification and propagation of uncertainties, etc.: These are among the myriad challenges that astronomers and statisticians face when trying to answer the seemingly simple question of "Just how far away is that galaxy?"

Further Reading

Freeman, P.E., Izbicki, R., and Lee, A.B. 2017. A Unified Framework for Constructing, Tuning, and Assessing Photometric Redshift Density Estimates in a Selection Bias Setting. *Monthly Notices of the Royal Astronomical Society* 468:4556.

Large Synoptic Survey Telescope (LSST). *www.lsst.org*.

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., and Fouchez, D. 2019. Photometric Redshifts from SDSS Images Using a Convolutional Neural Network. *Astronomy & Astrophysics* 621:A26.

Pospisil, T., and Lee, A.B. 2018. RFCDE: Random Forests for Conditional Density Estimation. *https://arxiv.org/abs/1804.05753*.

Sloan Digital Sky Survey (SDSS). *www.sdss.org*.

About the Author

Peter E. Freeman is an assistant teaching professor in Carnegie Mellon University's Department of Statistics & Data Science. Since receiving his PhD in astronomy and astrophysics from the University of Chicago, he has concentrated his research in astrostatistics: the application of cutting-edge statistical methods to problems such as source detection, cosmic microwave background mapmaking, making sense of galaxy morphologies, and estimating redshifts from photometric data.

Image courtesy of Getty Images

Statistics for Stellar Systems: From Globular Clusters to Clusters of Galaxies

Gwendolyn Eadie

Astronomers often want to know the physical mass of astronomical systems that they study. Some reasons are related to understanding the system's structure and composition (*What is the fraction of binary stars in a globular star cluster? Is there a central black hole at the center of a dwarf galaxy?*), while others are related to answering fundamental questions of physics (*How much dark matter is in a galaxy? How does dark matter affect the evolution of galaxies?*).

Whatever scientific question astronomers are trying to answer, they seek a trustworthy estimate of the mass of the *dynamical system*. But how do we go about "weighing" something that floats in the vacuum of space and is millions or even billions of light years away? What kinds of data are useful for estimating the masses of dynamical systems, and what limitations, uncertainties, biases, and/or assumptions come into play?

Dynamical Systems in Astronomy

The universe is full of dynamical systems, such as globular clusters, nuclear star clusters at the centers of galaxies, spiral and elliptical galaxies, and galaxy groups and clusters. Some examples of dynamical systems from small to large scales are shown in Figure 1.

The range of scale of these systems is…astronomical! The elliptical galaxy at the top right of Figure 1 could be just one of the tiny-looking galaxies in the galaxy cluster, and the globular cluster at the top left is the smallest system shown—if it were living in the spiral galaxy (third from the left), it would barely be more than a point of light.

Globular clusters contain anywhere from tens of thousands to hundreds of thousands of stars, while spiral and elliptical galaxies can contain millions to trillions of stars or more. Galaxy clusters, such as MCS J0416.1–2403 shown

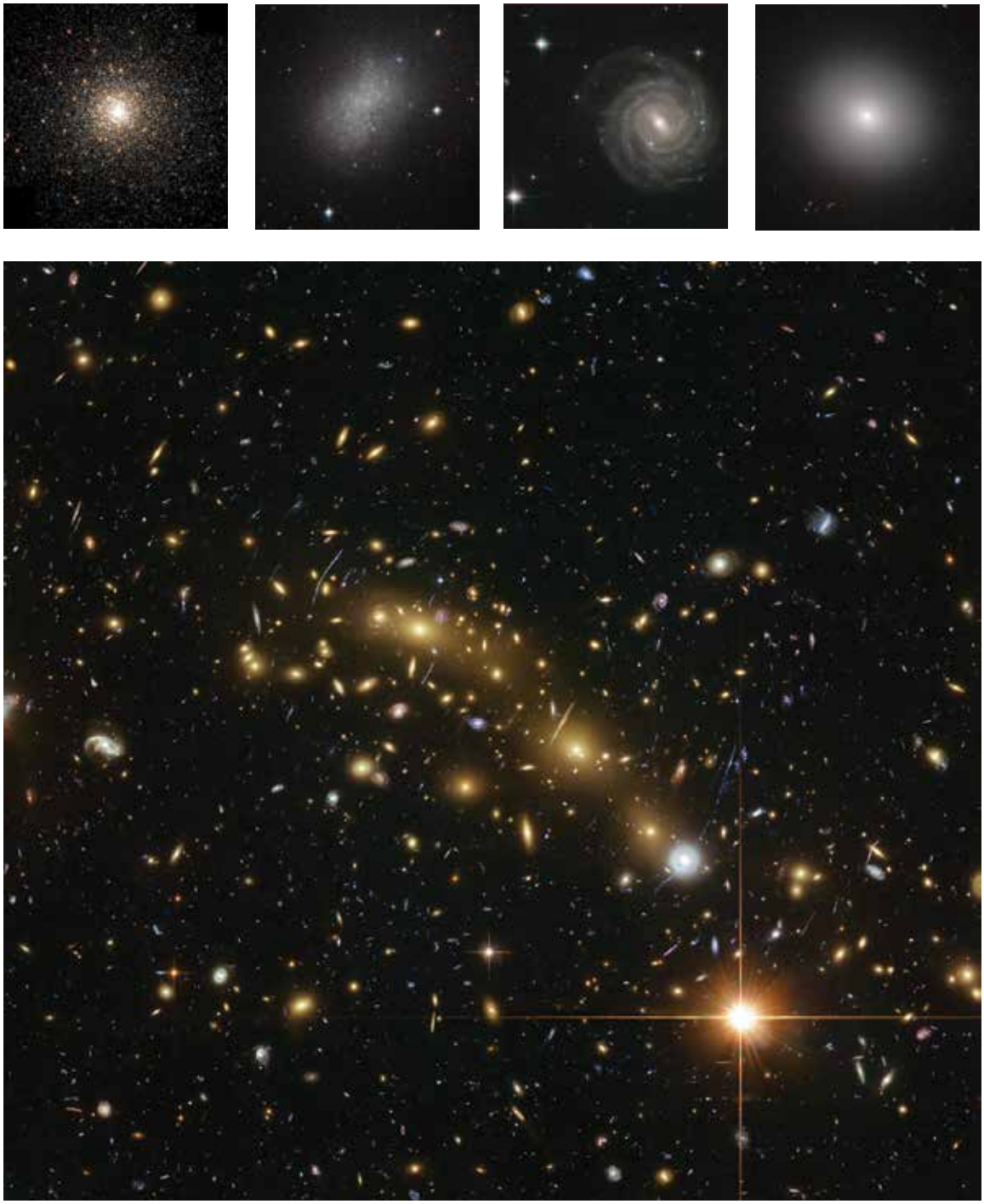


Figure 1. Examples of dynamical systems in astronomy. Top row, left to right: **globular cluster** M80 (NASA, Hubble Heritage Team, STScI, AURA), **dwarf galaxy** NGC 5264 (ESA/Hubble and NASA), **spiral galaxy** UGC 12158 (NASA/ESA Hubble Space Telescope's Advanced Camera for Surveys), and **elliptical galaxy** IC 2006 (ESA/Hubble & NASA images: Judy Schmidt and J. Blakeslee (Dominion Astrophysical Observatory)). Bottom: **galaxy cluster**, MCS J0416.1–2403 (ESA/Hubble, NASA, HST Frontier Fields Mathilde Jauzac (Durham University, UK, and Astrophysics & Cosmology Research Unit, South Africa) and Jean-Paul Kneib (École Polytechnique Fédérale de Lausanne, Switzerland)). Note: These objects vary in size by many orders of magnitude.

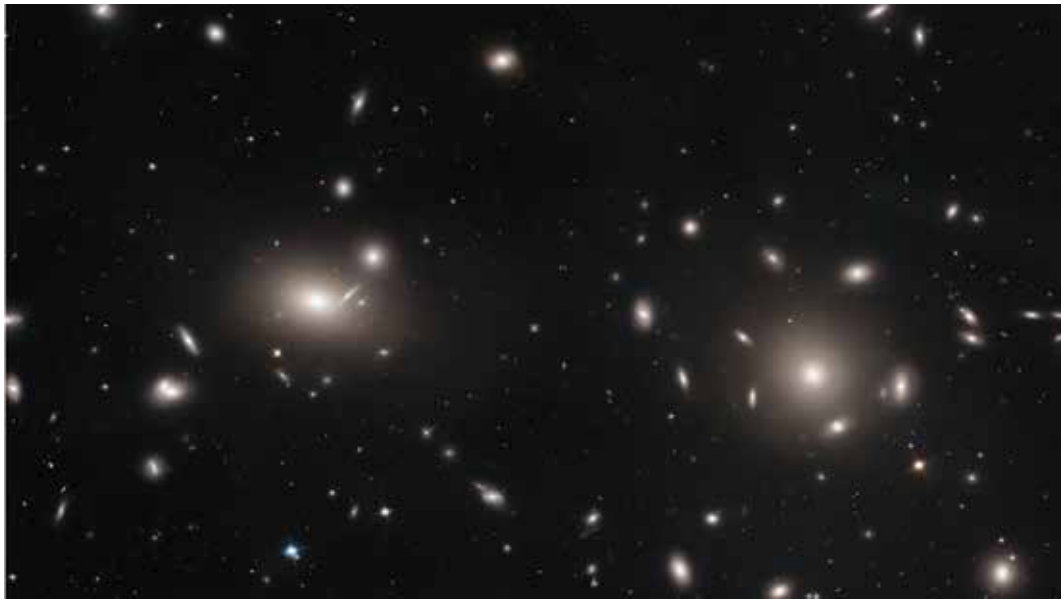


Figure 2. The Swiss astronomer Fritz Zwicky (1898–1974) is credited with coining the term *dark matter* in 1933, when he used it to describe the unseen matter responsible for holding together the Coma Cluster of galaxies, shown here. The Coma Cluster contains more than 1,000 galaxies and more than 22,000 globular clusters (not visible in the image). (Image: NASA; ESA; J. Mack, STScI; J. Madrid, Australian Telescope National Facility.)

in Figure 1, can have tens to thousands of *galaxies* within them, with each of those galaxies harboring millions to trillions of stars. Galaxies also host gas, dust, and compact objects (such as black holes, neutron stars, and white dwarfs), and are understood to sit within an enormous cloud of *dark matter*: matter whose presence is inferred by gravitational interactions with ordinary matter, but which cannot yet be measured directly (see Figure 2).

The total matter in galaxies and galaxy clusters can be composed of as much as 90% dark matter. A galaxy cluster can be so massive that it creates a *gravitational lens* to background sources of light. In the image of the galaxy cluster at the bottom of Figure 1, it is possible to see the curved arcs of light around the cluster are created by the gravitational lensing of light from background galaxies.

Determining exactly how much dark matter is in dynamical systems beyond the scale of globular clusters (beyond the scale of globular clusters, which appear to lack dark matter entirely) is motivation for understanding the total mass of these systems. Despite differences in size and number of stars, the masses of all of these different dynamical systems can be estimated using the same underlying physical principles.

The self-gravity of a dynamical system, sometimes combined with rotational support, holds the system together in a tightly or loosely bounded state. A system's compactness varies depending on the specific system, the environment in which it lives, and its past interactions with other systems.

The total gravitational potential of a dynamical system is directly related to its total mass, and also determines how objects like stars move within and around it. Thus, one approach to estimate mass is to assume a parameterized model for the gravitational potential, measure the positions and velocities of orbiting stars, and then use these data as *kinematic tracers* (or simply, tracers) to constrain the gravitational potential parameters. Once the total gravitational potential is estimated, the system's mass can be estimated. Nonparametric estimators of mass have also been explored by teams of astrophysicists and statisticians (Wang, et al. 2008).

Data for Kinematic Tracers

The type and number of kinematic tracers depends on the type of dynamical system (Table 1); some systems have multiple tracer populations (e.g., galaxies host stars, a globular cluster population, a

Table 1—Dynamical Systems and Their Tracers

System	Example Tracers	Approx. Number of Tracers
globular cluster	individual stars	10,000s–100,000s
nuclear star cluster	individual stars	≈10,000
Milky Way	outer (halo) stellar population	1,000s–10,000s
	planetary nebulae (halo)	10s–?
	globular clusters	≈150
	dwarf galaxies	10–30
	stellar streams	<10
galaxies	dwarf galaxies	10s–100s
	planetary nebulae	100s–1,000s
	globular clusters	100s–1,000s
	halo stars	1,000s–10,000s
galaxy clusters	individual galaxies	10s–1,000s
	globular clusters	>10,000s

A detailed branch of astronomy called astrometry specializes in the measurement, calibration, and uncertainty quantification of parallax, proper motions, and magnitudes (brightnesses) of stars.

dwarf galaxy population, etc.), while others have only a single population (e.g., globular clusters host stellar objects). As can be seen in the table, the size of the tracer population is sometimes on the order of only 10 tracers, but can also be as large as tens of thousands.

Every tracer in a dynamical system has three-dimensional position and velocity vectors with respect to the center of the system, making up what astrophysicists call six-dimensional *phase-space*. It takes all six phase-space components of a tracer to characterize its position and trajectory, but measuring all components is difficult and sometimes not even possible.

Take the velocity of a tracer. From the Earth-centered perspective, the velocity vector is measured in two parts: the velocity in our line of sight (v_{los}), and the projected velocity along the plane of the sky called the *proper motion* ($\vec{\mu}$). The former can readily be measured via the Doppler shift of known spectral lines; the amount of shift toward the red or blue end of the electromagnetic spectrum is a function of the tracer's movement away or toward us. The proper motion, however, is harder to measure because the tracers are so incredibly far away.

In the Milky Way, useful tracers like stars and globular clusters can be more than 45,000 light years away. To measure their proper motion requires taking images of these tracers separated over a wide span of time to observe any movement. It is best if the proper motion is measured in reference to something stationary (or approximately stationary) compared to the tracer.

For example, when measuring globular clusters, astrometrists often use background reference objects like distant galaxies and objects called *quasars*. Quasars produce incredible amounts of light, thought to be created by the accretion of gas onto supermassive black holes at the centers of extremely distant galaxies. For all intents and purposes, a background quasar is stationary with respect to a Milky Way globular cluster. These quasars are used as an anchor from which to measure the motions of stars within the cluster. The individual motions of stars can be used to study the globular cluster itself, or the mean motion of the stars can be used to estimate the proper motion of the whole cluster.

Proper motions of tracers in other galaxies are nearly impossible to obtain, and so for these systems, astronomers (and astrostatisticians) must come up with ways to deal with the incomplete velocity data.

The three-dimensional positions of tracers are a little easier to obtain, since their place in the sky is defined using an agreed-upon celestial grid of spatial coordinates called right ascension (RA) and declination.

The distance to the tracer provides the third spatial component. In the Milky Way, the RA, declination, and distance make it possible to transform the tracer's apparent

position to a coordinate system that is Galactocentric (Milky Way centered), by taking into account the sun's position in the Galaxy[1] with respect to the Galactic center, the location of the North Galactic pole, the movement of the sun through the solar neighborhood, and the rotation of the disk.

The distances to individual stars in our Galaxy can be measured using their parallax. The parallax is the apparent change of an object's position on the sky over the course of half a year, as the Earth makes its way from one side of the sun to the other. Using a small-angle approximation, the parallax is inverted to provide a distance estimate. The parallax of tracers around other dynamical systems (e.g., other galaxies, galaxy clusters) are virtually undetectable. Other distance indicators are used instead, like standard variable stars whose overall intrinsic brightness has a known relationship to the period at which they change in brightness (e.g., RR Lyrae stars).

When it comes to studying dynamical systems outside the Milky Way, the available six-dimensional phase-space information is limited to three measured components: the two-dimensional projected distance and the line-of-sight velocity. The distances to the host dynamical system are so far that obtaining high-precision distance measurements to their individual tracers is very difficult.

Statistical Inference Techniques for Dynamical Systems

Because of incomplete data, astronomers rely on point estimators to determine the mass of dynamical systems. A popular statistic for estimating the mass of a dynamical system with N tracers is the *projected mass* estimator first introduced by Bahcall & Tremaine (1981):

$$M = \frac{C_{proj}}{GN} \sum_{i=1}^{N} v_{los,i}^2 R_i, \quad (1)$$

where G is the gravitational constant, $v_{los,i}$ is the line-of-sight velocity of the i^{th} tracer, and R_i is the projected distance of the tracer from the center of the dynamical system.

This estimator is popular partly because it does not require proper motions to use it. Instead, the analyst must decide on the value for the constant C_{proj}, which is determined by assuming something about the distribution of the tracers' orbits (i.e., are they mostly circular orbits, elliptical, or some (ani)isotropic mixture of orbits?).

Figure 3. Galaxy NGC 7457. Based on observations made with the NASA/ESA Hubble Space Telescope and obtained from the Hubble Legacy Archive, a collaboration between the Space Telescope Science Institute (STScI/NASA), Space Telescope European Coordinating Facility (ST-ECF/ESA), and Canadian Astronomy Data Centre (CADC/NRC/CSA).

The estimator in Equation 1 has some nice properties; it is unbiased and has finite variance. On the flip side, it makes the unrealistic assumption that all mass in the system is centered at a point, and that the tracers are "test" particles orbiting around it. There is no way to include the shape of the dynamical system or the possibility of a non-spherical gravitational potential.

Having to choose a value for C_{proj} is also problematic since it means deciding on a distribution for the type of tracer orbits (also known as the *velocity anisotropy*). Different velocity anisotropy assumptions also can give wildly different results. An example is galaxy NGC 7457, for which the SLUGGS Survey has measured the line-of-sight velocities and projected distances for 40 globular clusters (Figure 3).

[1] Astronomers give the Milky Way a capital G to distinguish it from other galaxies.

Astronomers talk about the *velocity anisotropy* of the tracer population, which is a combination of the variances (or "dispersions") of the velocity components of the tracer population. The velocity anisotropy parameter is given by:

$$\beta = 1 - \frac{\sigma_\theta^2 + \sigma_\phi^2}{2\sigma_r^2} \qquad (2)$$

where N is the number of tracers, and the σ^2 values are the variances for each component of the velocity vectors in spherical coordinates.

Because astronomers deal with astronomically large objects, they often report mass in units of *solar mass*, $M_\odot$, which is equal to approximately 2×10^{30} kg.

Estimators such as Equation 1 also do not take into account measurement uncertainty of varying degrees, and do not allow for multiple spatial distributions for different tracer populations (Table 1). The spatial distribution of tracer populations in the Milky Way is known to vary by population, even though every population is subject to the same total gravitational potential. There is little reason to think other galaxies would be different.

Astrophysicists have derived probability density functions (pdfs) for the energy, E, and angular momentum, L, of tracers in a dynamical system, assuming a model for the gravitational potential and a model for the spatial distribution of the tracers.

Many of the details for deriving pdfs are covered in a well-known book, *Galactic Dynamics* by Binney and Tremaine (2008). These pdfs are useful components of maximum likelihood and Bayesian analyses that aim to estimate model parameters given data. However, to derive pdfs for tracers requires some assumptions about the dynamical system.

To start, tracers are assumed to be at equilibrium. This might not be true if the system recently underwent an interaction with another dynamical system (e.g., a galaxy-galaxy collision).

Another assumption that helps simplify the mathematics (and statistics) is spherical symmetry. For some dynamical systems this is a good approximation—a globular cluster, for example, is almost perfectly spherical—but for spiral galaxies, the assumption obviously breaks down. Only when considering the much-larger halo of dark matter in which the galaxy resides is spherical symmetry a more reasonable assumption. Some pdfs have been derived that allow for triaxial shapes, but their known mathematical forms are limited.

Improved Data and Methods

Like so many disciplines, astronomy is experiencing a wave of Big Data. The European Space Agency's Gaia satellite, launched in 2013, has been performing a stellar census of the Milky Way for the past seven years. On April 25, 2018, the Gaia Collaboration, et al., publicly released measurements of parallax and proper motions of more than 1 billion stars in the Galaxy. These data required advanced astrometry techniques and are revolutionizing the ability to study the stellar populations of the Milky Way (Figure 4).

The Gaia satellite has also made it possible to estimate the velocity of dynamical systems both within and just outside the Milky Way. For example, the proper motions of more than 150 Milky Way globular clusters and more than 10 dwarf galaxies have been estimated. Researchers have been using these tracer data to better understand the Milky Way's mass.

Recently, we used the globular cluster data in a hierarchical Bayesian analysis to estimate the Milky Way's total gravitational potential and mass. The hierarchical Bayesian method simultaneously accounts for measurement uncertainty, includes incomplete data (sampling the unknown components as parameters), and uses a pdf for the energies and angular momenta of the tracers based on a physical model.

The summarized results are shown in Figure 5 (adapted from Eadie and Jurić, 2019). Rather than a single point estimate for the mass, we obtain a posterior distribution and estimate the cumulative mass profile of the Milky Way with Bayesian credible regions (the gray areas in the figure).

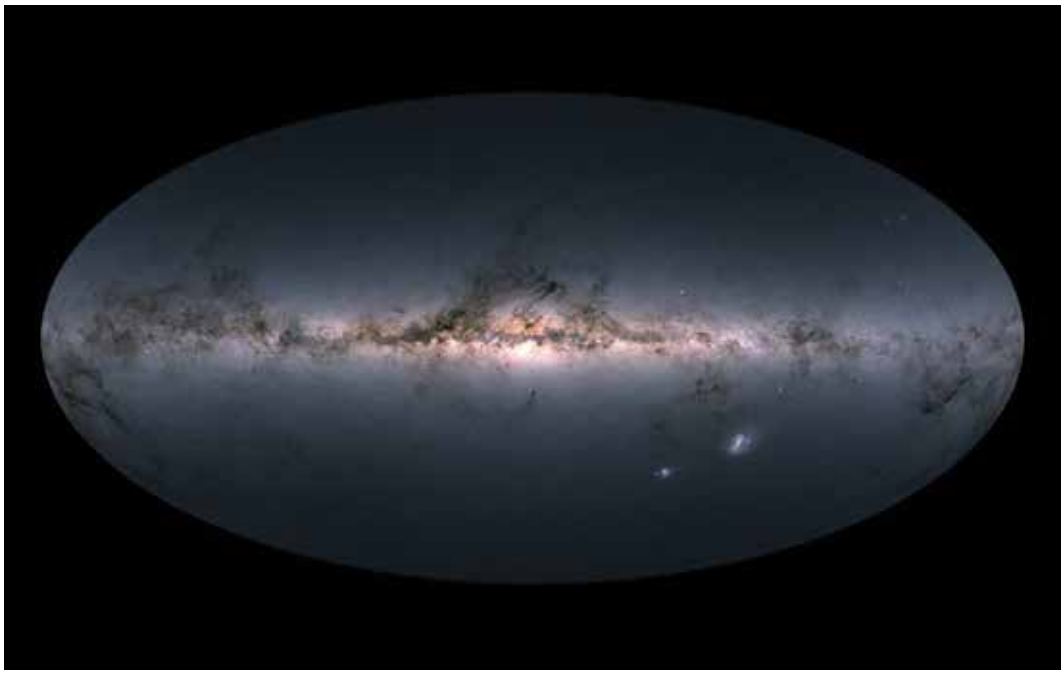


Figure 4. An all-sky view of the Milky Way Galaxy created using data from the Gaia satellite's measurements of more than 1.7 billion stars. Note that this is not a picture, but a figure. Dark portions in the plane of the galaxy correspond to dusty regions that block light, although they are only seen in this image because fewer stars were detected in these regions (ESA/Gaia/DPAC).

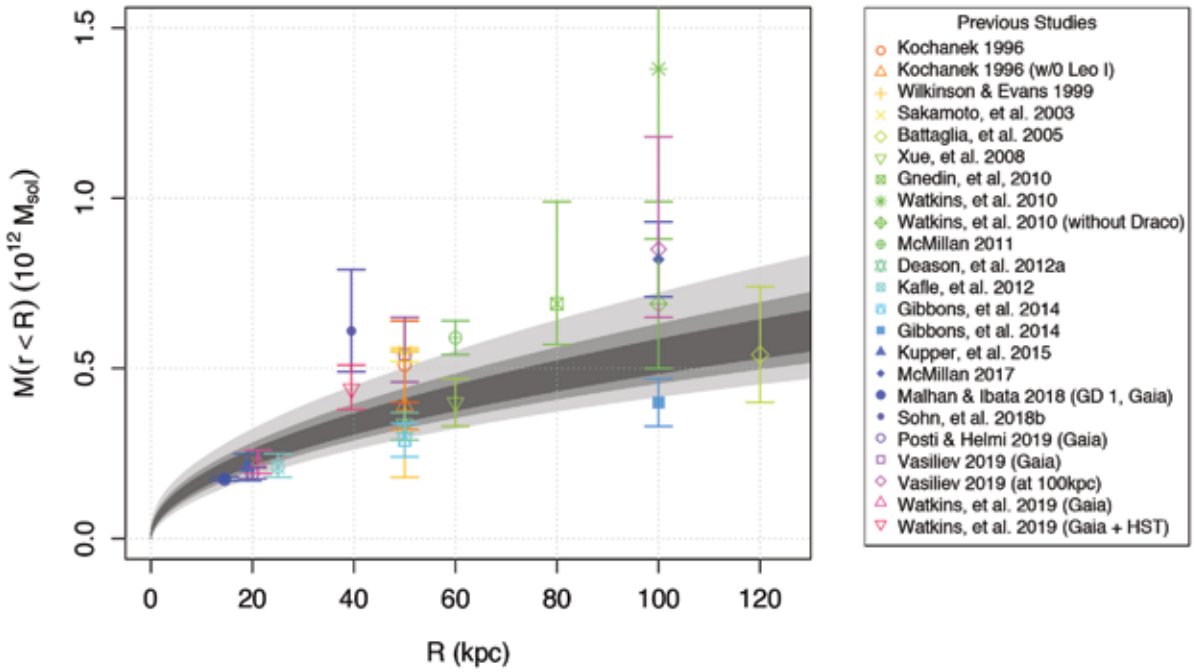


Figure 5. A cumulative mass profile estimate of the Milky Way Galaxy (grey regions are 50, 75, and 95% credible regions) using Gaia DR2 data, supplemented with additional data from the HST Proper Motion project. Other studies' mass estimates are shown as points with error bars. (Adapted from Eadie & Jurić. 2019.)

The SLUGGS Survey (*http://sluggs.swin.edu.au*) is the Study of the Astrophysics of Globular Clusters in Extragalactic Systems (SAGES) Legacy Unifying Globulars and GalaxieS Survey. This observational program used the Subaru/Suprime-Cam imager on the Subaru Telescope, and the Keck/DEIMOS spectrograph on the Keck II Telescope, both of which are near the summit of Mauna Kea on the Big Island of Hawai`i.

Use these data in Equation 1 and assume isotropic orbits ($C_{proj} = 16/\pi$), then do the same assuming extreme linear orbits ($C_{proj} = 32/\pi$). The former gives a total mass estimate of $5.68 \times 10^{10} M_{\odot}$, while the latter doubles this value. When no information is available about the tracer orbits, the mean of these two estimates is sometimes taken as the mass estimate.

The Future of Big Data in Astronomy and Dynamical Systems

As missions like Gaia improve and increase data collection, the limitations to inference about the Milky Way and its dynamical systems have more to do with statistical techniques and physical models than with missing data.

At the same time, data from large surveys have to be analyzed with scrutiny. For example, Bailer-Jones, et al. (2018) have shown that a simple inversion of the parallaxes provided by Gaia can lead to biased and incorrect estimates of the distance to stars. Some of the parallaxes measured by Gaia are negative, and a simple inversion then implies a negative distance, which is entirely unphysical. It has therefore been recommended that inference techniques such as Bayesian analysis with prior information be used to fully account for uncertainty in the nonlinear transformation from parallax to distance (see Further Reading). In short, even with Big Data, we must be careful about our statistical procedures.

In the next decade or two, enormous amounts of data on other galaxies and galaxy clusters will become publicly available through other telescope surveys and space-based missions. However, incomplete data and measurement uncertainties will remain stumbling blocks to measuring the masses of these dynamical systems beyond our Galaxy. Statistical techniques that can deal with these issues will become paramount.

With high-performance computing, methods like Bayesian analyses can incorporate measurement uncertainties, include incomplete data, and simultaneously estimate unknown parameters such as the velocity anisotropy, using large data sets. Computer simulations of galaxies combined with model emulators can also enable approaches such as approximate Bayesian computation.

As these methods increase in popularity in astrophysics and the public release of astronomical data continues, the field of astrostatistics will continue to grow.

Further Reading

Bahcall, J.N., and Tremaine, S. 1981. *The Astrophysical Journal*, 244, 805. doi: 10.1086/ 158756.

Bailer-Jones, C.A.L., Rybizki, J., Fouesneau, M., Mantelet, G., and Andrae, R. 2018. *The Astronomical Journal*, 156, 58. doi: 10.3847/ 1538–3881/aacb21.

Binney, J., and Tremaine, S. 2008. *Galactic Dynamics* 2nd ed. Princeton.

Eadie, G., and Jurić, M. 2019. *The Astrophysical Journal* 875, 159. doi: 10.3847/ 1538-4357/ab0f97.

Gaia Collaboration, Brown, A.G.A., Vallenari, A., et al. 2018. *Astronomy and Astrophysics*, 616, A1. doi: 10.1051/0004-6361/ 201833051.

Wang, X., Walker, M., Pal, J., Woodroofe, M., and Mateo, M. 2008. *Journal of the American Statistical Association*, 103, 1070. doi: 10.1198/ 016214508000000652. Statistical Association, 103, 1070. doi: 10.1198/ 016214508000000652.

About the Author

Gwendolyn Eadie is an assistant professor of astrostatistics at the University of Toronto, jointly appointed between the Department of Astronomy & Astrophysics and the Department of Statistical Sciences. In 2018, she received the J.S. Plaskett Medal from the Canadian Astronomical Society for her PhD dissertation. She is also the program chair for the ASA Astrostatistics Interest Group and co-chair of the American Astronomical Society's Working Group on Astrostatistics & Astroinformatics.

Image courtesy of Getty Images

ARIMA for the Stars: How Statistics Finds Exoplanets

Eric D. Feigelson

Planet, Planets Everywhere!

In the first century BC, the Roman poet Lucretius wrote, "you are bound to admit that in other parts of the universe there are other worlds inhabited by many different peoples and species of wild beasts." In the 17th century, Bernard de Fontenelle popularized the plurality of worlds with a best-seller among the nobility of Versailles. A generation ago, astronomer Carl Sagan entranced millions with lyrical phrases like "Think of how many stars, and planets, and kinds of life there may be in this vast and awesome universe."

From a scientific perspective, all of this was fantasy. While astronomers had established in the 19th century that the stars at night were luminous gaseous spheres like the sun, there was no evidence at all for planets orbiting the stars—until 1995. At that point, Swiss astronomers Michel Mayor and Didier Queloz had measured subtle Doppler shifts in the wavelengths of spectral lines from the starlight of 51 Pegasi, a nearby sun-like star. The observed periodic sinusoidal pattern in the star's motion indicated the star is being pulled back and forth by an invisible Jupiter-like planet orbiting every 4.2 days.

Their discovery was revolutionary, giving birth to the field of exoplanetary astronomy that is now a significant segment of all astronomical research effort.

Today, we know that most stars seen in the nighttime sky have planetary systems. Quantitative evaluations are still uncertain, but the best estimates are that stars typically have about five planets and perhaps 1% to 10% of them have Earth-like planets in Earth-like orbits where life could exist on the surface. (This fraction is known as $\eta_{\oplus}$, voiced as "eta-Earth.")

It can be inferred that there are hundreds of millions of habitable planets in the Milky Way galaxy,

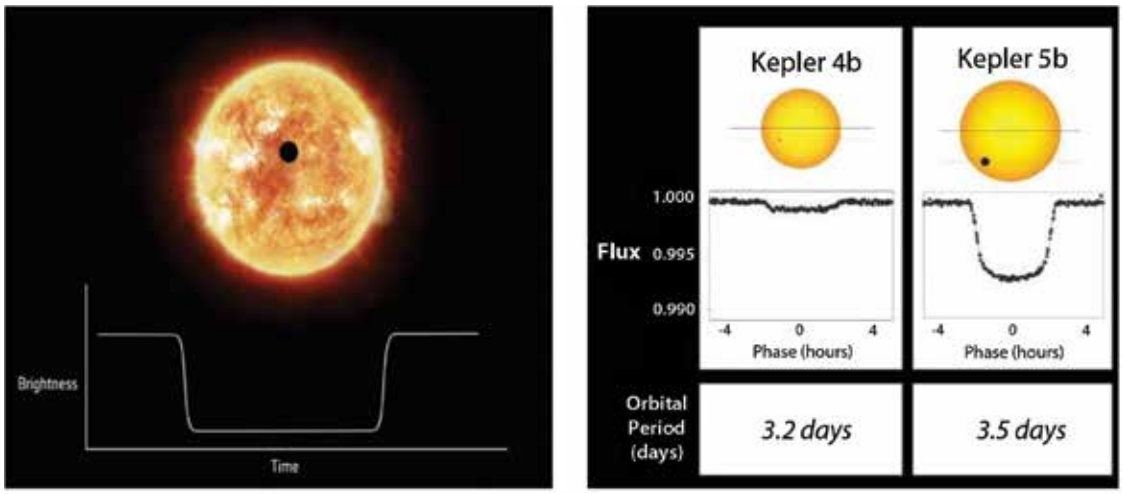


Figure 1. Diagram of an exoplanet transiting a star and the dip in brightness that occurs every orbit (left). Transit light curves from two planets found early in the Kepler mission, one a Neptune-size and the other Jupiter-size (right). Transits typically have amplitude of 0.01% to 1%, so other sources of brightness variations often overwhelm the planetary signal. (NASA/Kepler mission.)

the closest only a few lightyears away. No evidence for extraterrestrial life—Lucretius's wild beasts—has emerged, but the discovery of ubiquitous exoplanets motivates powerful research efforts directed toward this goal.

Astrostatistics and Exoplanets

Orbital models of multiplanet systems rely on sophisticated Bayesian nonlinear regression procedures where astrophysical processes constrain the prior distributions of model parameters. In part, the link between astronomy and statistics arises from the overabundance of the underlying populations. Of the billions of planets in our galaxy, we have so far sampled only a few thousand. Statistics is crucial here; a few years ago, two research groups analyzing the same data set arrived at $\eta_{\oplus}$ estimates that differed by a factor of 7 due to different statistical analysis procedures.

Statistical inference is also critical for the detection and characterization of exoplanets because the signals are so weak. One effective method for discovering exoplanets detects changes in the velocity of the star toward and away from Earth as the planet orbits the star. Earth pulls the sun only 9 centimeters per second, so detecting its Doppler signature requires a spectrograph with better than 1-in-a-million precision measurements that is stable for years.

This is still beyond current engineering capabilities, but the latest astronomical spectrographs can detect Jupiter-mass planets out to Earth-like orbits, or Earth-like planets out to Mercury-like orbits.

A second effective method for discovering planets is somewhat less-demanding on our instruments. When an orbiting planet passes (or "transits") in front of a star, a small portion of the starlight is briefly eclipsed. For a Jupiter-size planet, the light is diminished by roughly 1%; for an Earth-size planet, the diminution is around 0.01%. Measuring star brightnesses (called "stellar photometry") to this precision is quite feasible today.

Figure 1 illustrates the dip in starlight produced by a planetary transit. In addition to various mountain-top telescopes, several satellite observatories are devoted to transit searches: Kepler and TESS missions launched by the U.S. National Aeronautics and Space Agency and the Cheops mission from the European Space Agency.

Since only a small fraction of orbits produce transits when viewed from any particular direction, many stars must be photometrically monitored for months or years. A variety of large-scale transit surveys are now underway, from both space-based satellite observatories and ground-based mountaintop observatories.

The Challenge of Planetary Transit Detection

A decade ago, NASA launched the Kepler mission with high confidence it would find many Earth-like planets, thanks to its instrumental precision (around 0.003%) and planned four-year

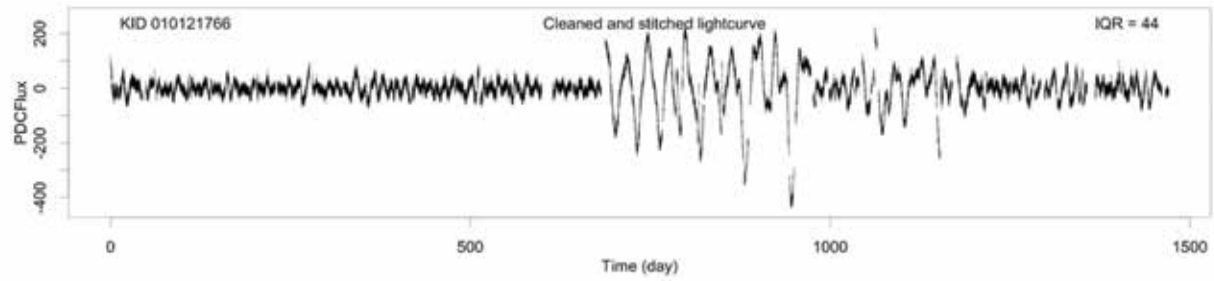


Figure 2. A four-year light curve from NASA's Kepler mission showing stellar brightness variations measured every half-hour during 2009–13. KID = Kepler identification number for this star, PDCFlux = star brightness after the "Pre-Search Data Conditioning" module is applied to reduce instrumental effects, and IQR = InterQuartile Range. In this star, the magnetic activity suddenly increased producing quasi-periodic variations from rotationally modulated starspots after about two years of observation. A low-dimensional ARFIMA model produced near-Gaussian white noise residuals but, as with most stars, no transiting planet was found here. (This plot and Figure 3 plots, obtained by Caceres, et al. 2019.)

photometric survey of ~200,000 stars. I did identify several thousand larger planets—most Kepler discoveries are super-Earth-size planets with orbital periods between two and 100 days—but hopes were foiled for the smaller planets.

It was the stellar variability that caused the problem: More stars exhibited brightness changes than expected. They mostly arise from magnetic activity and convection as seen on the sun and manifested as cool starspots, hot faculae, and flares.

While some stellar time series (called "light curves" by astronomers) can be quiet, others can show complex nonstationarity variations. The temporal behaviors are not simple: stochastic autocorrelated variations from superposed microflares, quasi-periodicities as starspots rotate in and out of view, and explosive white light flares. The phenomena cannot be modeled realistically astrophysically and are thus unpredictable. Figure 2 shows a typical nonstationary Kepler light curve.

The typical procedure has two stages: reducing the unwanted stellar variations in the time domain and searching for periodic transits in the frequency domain. The initial reduction of stellar variability in the time domain is generally performed with nonparametric procedures: wavelet analysis (as in the official Kepler pipeline) and high-dimensional local regression procedures like Gaussian Processes regression are most-often applied. A few researchers use advanced signal processing methods like Independent Component Analysis, correntropy, Empirical Mode decomposition, or Singular Spectrum Analysis.

Once the star brightness changes have been reduced, statistical procedures for finding periodic dips in brightness due to a transiting planet are well-established. Fourier transforms are not efficient because the transit shape is not sinusoidal; a "box least-squares" (BLS) matched filter for transit-shape dips is used instead. A BLS periodogram is constructed, light curves are "folded" (plotted modulo) to the most prominent spectral peak and are "vetted" by expert astronomers to make sure they exhibit properties expected of transiting planets.

This procedure is beset by four big problems:

1. It takes a time domain procedure that can treat a wide variety of unpredictable behaviors.
2. It requires a frequency domain procedure to find very faint transit-shaped periodic dips produced by small, Earth-sized planets. Periodograms have non-Gaussian power distributions that can easily give rise to spurious spectral peaks.
3. If the vetting stage is seen as a classification procedure, then the classes are badly imbalanced, with dozens of non-transiting cases for each transit. Even an occasional misidentification of a spectral peak with noise can flood the community with false alarms, wasting valuable telescope time for astronomical follow-up studies.

TERMS AND DEFINITIONS

ARIMA: In AutoRegressive Integrated Moving Average (ARIMA) time series models, the AR component treats dependence on recently past brightness values, MA component treats dependence on recently past brightness changes, and the I represents the differencing operation to reduce non stationarity. Fractional differencing in ARFIMA treats long-memory correlations corresponding to the astronomers' $1/f^{\alpha}$ "red noise." A periodic transit can be added as an exogeneous variable to the aperiodic ARIMA model using the ARIMAX formalism. We calculated model fits using the forecast CRAN package in the R statistical software environment developed by statistician Rob Hyndman and colleagues.

Box Least Squares and Transit Comb Filter: The BLS and TCF periodograms are frequency domain alternatives to the Fourier periodogram that are designed to detect periodic dip and spike shapes expected from planetary transits, rather than the periodic sinusoidal shape of Fourier analysis.

Random Forest: RF is an effective classifier using ensembles of decision trees based on "features" drawn from the data set and analysis by the scientist. Important advantages of RF for this problem include: a metric-free design permitting features with different scales and units; self-validation using out-of-bag sampling in each tree of the forest; and balanced treatment of imbalanced training sets.

Time series diagnostics: At each stage of the time series analysis, the analyst can examine how close the data are to white (uncorrelated) Gaussian noise. These include the Anderson-Darling test for normality, Ljung-Box portmanteau test for autocorrelation, and adjusted Dickey-Fuller test for stationarity.

4. Some stars have periodic variability for other reasons and disguise themselves as transiting planets. These include (quasi-)periodic variations from convection or pulsations in some stars, and eclipsing binary star systems that contaminate the target star light curve. Light curves with real, but non-planetary, periodicities are called astronomical false positives.

Penn State's Approach: AutoRegressive Planet Search (ARPS)

We have tried a different approach to stellar variability reduction based on a classic technique from time series analysis known as Box-Jenkins analysis. A low-dimensional, often-linear model is constructed where the brightness is not a function of time, but of recent past brightness levels and past changes (see sidebar). These are known as ARIMA models and are generally fit by maximum likelihood estimation, with model complexity determined by maximizing the Akaike Information Criterion.

However, the Kepler light curves are often nonstationary where the mean levels change quasi-periodically or in other peculiar ways (Figure 1). ARIMA models with a differencing operation treat many forms of non stationarity. If long-memory processes are present, ARFIMA models can be used.

The ARIMA residuals are then searched for periodic transits, but the differencing operator converts a box-like dip into ingress and egress spikes. Gabriel Caceres, while a graduate student at Penn State, developed a matched filter for a periodic sequence of double-spikes; he called this the Transit Comb Filter (TCF). The TCF periodogram is then examined for peaks representing repeated transit-like variations in the ARIMA residuals.

The ARPS procedure then faces a Big Data classification problem where the vast majority of Kepler light curves have no transit, and the small fraction with true planetary transits. Fortunately, NASA's Kepler Team recently released a "golden" list of transiting planet candidates, many confirmed with telescopic study by the broader astronomical community. These serve as a "planet training set" for a multivariate classifier, while random light curves (supplemented with sets of confirmed false positives like eclipsing binaries) serve as a "non-planet" training set.

We developed a list of several dozen "features" for a Random Forest multivariate classifier. The features are drawn from the original light curve, ARIMA residuals, TCF periodogram, folded light curve, and various time series diagnostics. Interestingly, the most-important feature emerged from an ARIMAX fit that gave a unique measure of the statistical significance of the transit depth.

ARIMAX analysis (see sidebar) is used most commonly in econometrics and had never before been applied to astronomical problems.

We optimized a Random Forest (RF) decision tree classifier using ROC curves. We carefully chose a threshold of RF probabilities on scientific grounds, balancing the recovery of true positives with the need to avoid the potentially overwhelming number of false alarms and false positives from the imbalanced classes.

Finally, we vetted the light curves satisfying the RF threshold for problems and identified promising cases. Several dozen

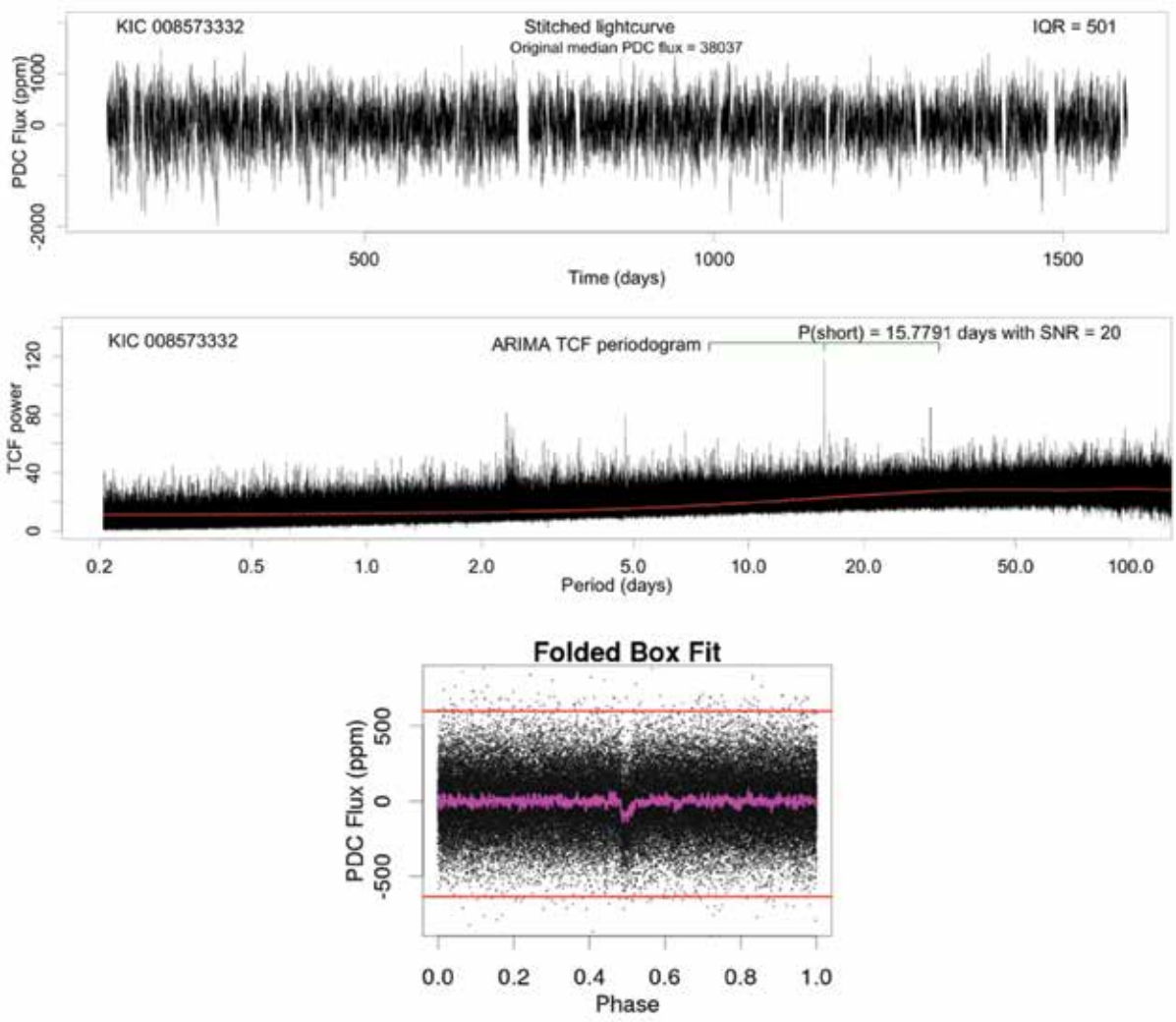


Figure 3. Light curve of a star with strong, choppy autocorrelation in units of brightness parts per million (ppm) (top). The TCF periodogram of ARIMA residuals shows a promising peak at period days (middle). The folded light curve, seen here as residuals of an ARIMAX model, shows the faint box-shaped dip expected from a planetary transit (bottom). The units of flux are ppm.

new planetary candidates emerge from this effort, most of them Earth-sized with very short orbital periods (0.2–10 days). These are planets so close to the host star that the rock surface itself is probably molten.

Figure 3 shows one of these promising discoveries. The middle panel shows a TCF spectral peak corresponding to an orbital period of 15.8 days, and the bottom panel shows a transit depth around 100 ppm corresponding to a planet with 1% the radius of the host star. If real, this Earth-sized planet has a very hot, possibly molten surface.

The Future of ARPS

From the perspective of a time series analyst, our procedure may not be particularly innovative: Box-Jenkins ARIMA-type modeling has been widely used since the 1970s, TCF can be viewed as a variant of Fourier analysis, and Random Forest classification has been successful in many fields since the 2000s. But in astronomy, this is quite an unusual approach; astronomers have weak education in applied statistical methodology and, in particular, are not familiar with ARIMA modeling.

The autoregressive planet search method outlined here can be developed further to improve the census of small planets, increasing

sensitivity and reducing false alarms and false positives. Nonlinear ARFIMA and GARCH models give better fits to stellar variations in some cases.

Preprocessing by outlier removal can clean up some noisy periodograms. The TCF algorithm can be extended to treat gradual, rather than sudden, ingress and egress spikes. Newer machine learning classifiers like XGBoost can replace the Random Forest stage, and meta-classifiers might reduce the need for expert vetting.

Research groups in Israel and the U.S. are investigating the effectiveness of Deep Learning neural networks where the full database of light curves, rather than selected features, are entered into the classifier.

We have also investigated applying the ARPS approach to ground-based transit surveys where light curves of millions of stars are constructed. The problem here is the irregular spacing of the light curve time series data. Typically, a star can be observed only about eight hours per day at night, and is visible for only about six out of 12 months per year. The daily gaps in observation can be mitigated by placing telescopes on different continents or at the South Pole. We have simulated planet detection from ground-based surveys of this type and happily find that ARIMA and TCF methods work even with 70%–90% "missing data."

The Future of Statistics and Exoplanets

Exoplanetary research is in its third decade, but the work is still in its early phases. Nearly every aspect of the observational studies requires advanced statistical methodology:

- Algorithms for reducing stellar variations, both from the observational conditions and intrinsic to the star;
- Accurate subtraction stellar emission to reveal faint planetary emission in time series, astronomical spectra, and images;
- High-dimensional parametric modeling of multiplanetary orbits fitted to sparse data sets;
- Highly imbalanced classification problems from big surveys;
- Correcting observational selection biases to quantify underlying planetary populations; and
- Searching for biosignatures.

The field of astrostatistics must lead the way in future exoplanet studies, and this requires a combined effort of astronomers and statisticians. Astronomers traditionally have little training in statistics, and few statisticians are familiar with the scientific issues or are embedded in astronomical research teams. Astronomical projects fund "software" and "data analysis" computation, but not development of the methodology that underlies the computing effort, so cross-disciplinary collaboration is still rare.

The field of astrostatistics is constantly improving, with exponential growth in the use of machine learning and Bayesian approaches in the astronomical research literature. Many challenges are being effectively addressed, and many more will be faced in the future. If we ever find the extraterrestrial "wild beasts" predicted by Lucretius, it is likely that statistics will play a crucial role in the discovery.

Further Reading

Caceres, C.A., Feigelson, E.D., Babu, G.J., Bahamonde, N., Cristen, A., Meza, C., and Cure, M. 2019. AutoRegressive Planet Search, *Astronomical Journal*, in press arxiv:1901.05116, 1905.03766 and 1905.09852.

Feigelson, E.D., Babu, G.J., and Caceres, G.A. 2018. Autoregressive time series methods for time domain astronomy, *Frontiers of Physics* 6, #80. arxiv:1901.08003.

Haswell, C.A. 2010. Transiting Exoplanets. Cambridge, UK: Cambridge University Press.

Hyndman, R.J., and Athanasopoulos, G. 2018. *Forecasting: Principles and Practices,* 2nd edition. OTexts. *otexts.com/fpp2*.

About the Author

Eric Feigelson is professor of astronomy and astrophysics, and of statistics, at Penn State University. He has been working with G. Jogesh Babu and other statisticians for 35 years developing curriculum, offering summer schools, organizing meetings, and conducting research in astrostatistics.

Identifying Milky Way Open Clusters With Extreme Kinematics using PRIM

Mark R. Segal and Jacob W. Segal

The recent (April 25, 2018) second data release (DR2) of the European Space Agency satellite *Gaia* provides a dramatic increase in the extent and precision of stellar attributes for an unprecedented number of sources in the Milky Way (MW). For instance, data on position in the sky and *proper motion* are available for more than 1.3 billion stars, with a subset of more than 7 million of the brightest stars having radial velocity measurements. (Definitions of proper motion, *radial velocity*, and other astronomical terms are provided in the glossary.)

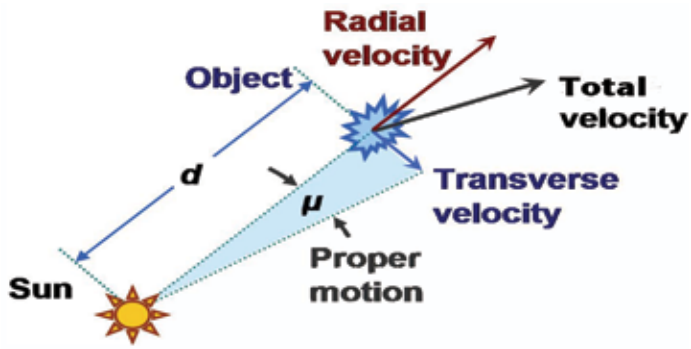


Figure 1. Cartoon depicting a star's velocity components. A star's velocity, in three-dimensional space, has three components, one in each coordinate direction. (Brews Ohare. *https://commons.wikimedia.org/wiki/File:Proper motion.JPG*.)

Investigation of *kinematics*—combined proper motion and radial velocity—is central to our exploratory analyses. The cartoon in Figure 1 illustrates how these components yield total stellar velocities. Subsequently, we describe how to transform from the depicted sun-centric viewpoint to a *Galactic Center* viewpoint. The systematic study of stellar velocities has been one of the primary applications of the *Gaia* DR2 resource, with particular interest in the identification of *hypervelocity stars* (HVS). Rewinding the paths of some of these stars has enabled speculation regarding mechanisms, often violent, about how the extreme speeds were attained.

Two of these components—depicted in Figure 1 as transverse velocity—can be obtained from the star's proper motion, μ, coupled with distance to the star, d. In turn, this distance can be obtained as inverse *parallax*, if parallax is determined precisely enough. *Proper motion* measures angular change in position over time, with position measured on the *celestial sphere* using *right ascension* (akin to longitude) and *declination* (akin to latitude).

The third component of velocity is *radial velocity*, measured along the line of sight. The magnitude of the three component velocity vector is shown as *total velocity*, designated as v_{tot} and serving as the outcome for these analyses. The sun-centric perspective is transformed to a coordinate system with origin at the Galactic Center.

Another line of enquiry using *Gaia* DR2 data has been identifying *open clusters* (OCs). Open clusters comprise from dozens to thousands of stars, formed from the same molecular cloud around the same time and loosely bound by gravitational attraction. More than 1,000 OCs have been discovered in the MW, primarily located in the *Galactic disk*. This contrasts with the rarer (<100), larger, more-luminous, accretion-formed *globular clusters* that reside throughout the *halo*, velocity measurements of which have provided evidence for the existence of dark matter.

OCs provide markers of the formation and evolution of the MW. Their ages span the entire history of the Galactic disk, ranging from the *young thin disk* (our subsequent focus) to the old

thin disk components and their attendant life cycles. These are important for explaining the creation and evolution of the MW disk and, potentially, spiral galaxies in general. Further, the spatial distribution and motion of OCs can be informative with regard to perturbations acting on the structure and dynamics of the MW.

Beyond serving as such tracers of MW structure and history, OCs are intrinsically interesting targets. As homogeneous groups of stars with the same age and same initial chemical composition, they provide a useful testbed for studying individual star formation and evolution.

Rather than search for individual HVS or OCs irrespective of velocity, these objectives can be combined by seeking *groups* of stars that exhibit (relatively) extreme velocities. The search focuses on the young thin disk because: (i) This region is where the bulk of OCs reside; (ii) the expanded *Gaia* DR2 data set has been relatively enriched in high magnitude (faint) stars, formerly obscured by gas clouds in the young thin disk; and (iii) the restriction reduces the size of the search space, which has both computational and inferential benefits.

Since the stellar population of a galactic disk exhibits (fluid-like) dynamical rotation, with little random motion due to being bound by the disk's gravitational potential, it can be anticipated that extreme velocity groups therein will not be common. Such groups, if identified, may be especially revealing in terms of originating perturbations.

None of this is to say that a broader, halo-wide search is not purposeful. Since the search process we use proceeds by successively removing the (extreme) groups detected, our approach can still recover conventional OCs. It may help to describe some current methods usd for identifying OCs and discuss associated issues.

Existing Approaches to OC Detection

Using the *Gaia* DR2 data release for OC identification—which can be framed as clustering or local mode detection based on stellar density estimates—is exceedingly challenging. This is because of (i) the large total numbers of stars, as noted above; (ii) the relatively small numbers of stars constituting an OC; (iii) differential precisions with which stellar positions and velocities are measured; and (iv) the existence of extensive local MW structure, which makes the standard use of the uniform distribution inappropriate as a null referent distribution for declaring a local mode.

Indeed, some authors have ascribed OC discovery to "serendipity" resulting from close examination of small, pre-selected regions (often corresponding to known OCs) of the MW. Accordingly, they have invoked the need for data mining methods to facilitate detection in an unbiased way on a galactic scale, which motivates our work in part.

The necessity for using data mining was made explicit in other work, with the unsupervised clustering technique DBSCAN being a frontline technique. To address the concerns about local MW structure and computational burden, researchers applied the clustering algorithm separately to predefined regions that partition the Galactic disk, a strategy we also adopt.

As with all clustering methods, DBSCAN requires a measure of similarity or distance between objects (stars). Operating in the five-dimensional space comprising three parameters defining spatial position and two proper motion (right ascension, declination) parameters, previous researchers used Euclidean distance after standardizing each parameter separately. However, as acknowledged, this choice is one of convenience and, for example, disregards the (i) strong dependencies between proper motions and parallax, as have been accommodated in identifying HVSs, and (ii) differing variation and precision of the component parameters.

Moreover, by combining spatial and velocity components, the topology induced by this metric can potentially group stars that are spatially distant, yet moving together.

Identifying such groupings, which share a common origin but have become gravitationally unbound and are termed *stellar associations*, is of interest, but our focus here is on detecting spatially proximal clusters that also move together; i.e., OCs.

We achieve this by separating spatial and velocity components. Further, we restrict to stars with radial velocity measurements, which enables derivation of total velocity per Figure 1. This permits recasting the problem in a supervised learning framework, emphasizing detection of clusters that are moving relatively rapidly or slowly.

We subsequently outline the patient rule induction method (PRIM)—the particular supervised methodology used—but first detail data sources and the preprocessing steps needed to obtain total velocity measurements in an appropriate reference frame (coordinate system).

Data Acquisition, Filtering, and Processing

The *Gaia* DR2 data set contains 7,183,262 stars with both radial velocity and astrometric parameters (position on the

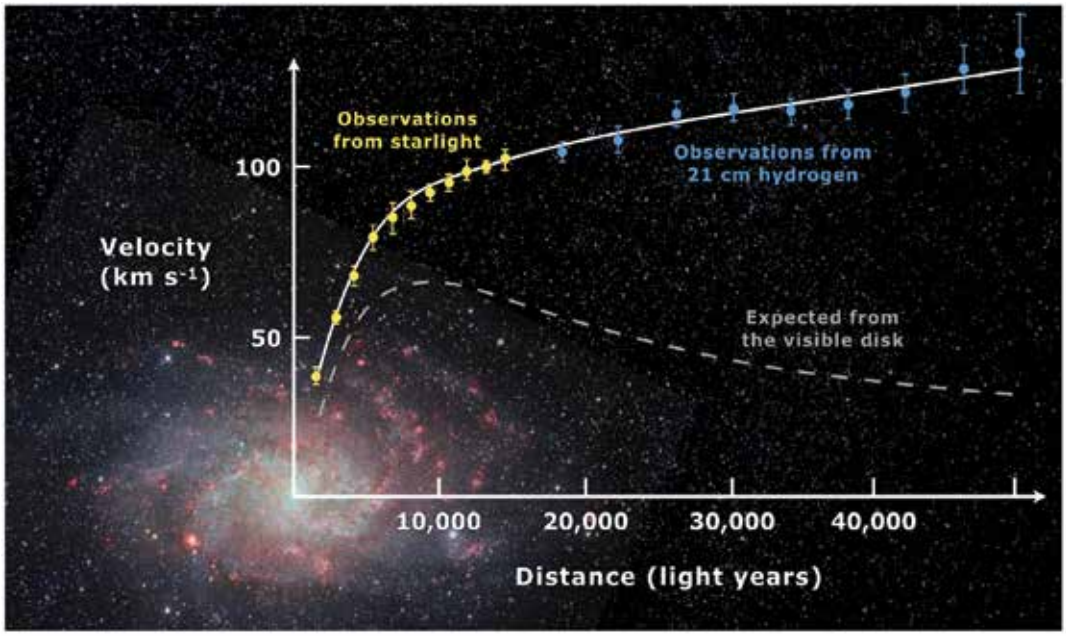


Figure 2. Theoretic and observed rotation curves for the spiral Triangulum Galaxy (Messier 33). (Mario De Leo. *https://commons.wikimedia.org/w/index.php?curid=74398525*.)

celestial sphere given by right ascension and declination, parallax, and proper motions). To obtain total velocities, it is necessary to convert parallax to distance so apparent motion of an object on the celestial sphere can be transformed to physical motion in space.

While this conversion is a matter of simple inversion, errors in parallax measurement (including negatives!) make corresponding filtering important. We follow previous work in restricting to stars with relative parallax errors between 0 and 20%, which reduces the sample to 6,376,803.

We also inherit prior determinations, from a publicly available catalog, of total velocity (v_{tot}) in a coordinate system with origin at the Galactic Center (GC). Deriving these total velocities involves correcting observed radial velocities and proper motions for the sun's position and motion relative to the GC.

We designate star position using Galactocentric (rectangular) coordinates (x_{GC}, y_{GC}, z_{GC}) with the GC as origin, since these are readily obtained from right ascension, declination, and parallax using standard transformation from spherical to Cartesian coordinates. The distance of a star from the GC, r_{GC}, is then just Euclidean distance in this coordinate system. By construction, the x_{GC} axis aligns the GC and the sun. By virtue of the sun-centric nature of *Gaia* DR2 sampling, and our focus on the young thin disk, this axis essentially coincides with the first principal component of the matrix of stellar positions in our filtered data set.

Similarly, y_{GC} and z_{GC} correspond to the second and third principal components, respectively. This agreement has bearing on our subsequent analysis.

Various rules have been advanced to define the young thin disk. Here, we use the specification that it consists of stars within 100 parsecs (*pc*; 1 *pc* ≈ 3.26 light years) of the disk as measured by z_{GC}. This restriction further reduces our data set to 1,702,932 stars, 27% of the (parallax) filtered sample.

The *rotation curve* of a spiral galaxy is a graph of stellar velocity versus radial distance from the GC, here r_{GC}. It was through study of rotation curves, and attendant discrepancies between theoretic and observed curves, that the existence of dark matter was first postulated, as illustrated in Figure 2.

To detect OCs and kinematic groups with relatively extreme velocities requires accommodation for systematic velocity variation. We effect this control by stratifying into predefined regions based on r_{GC} deciles and 5th and 95th quantiles. Such stratification limits the ability to detect clusters that straddle decile boundaries, and arguably, in view of the largely flat rotation curve for our *Gaia* DR2 subsample, is not necessary. However, by

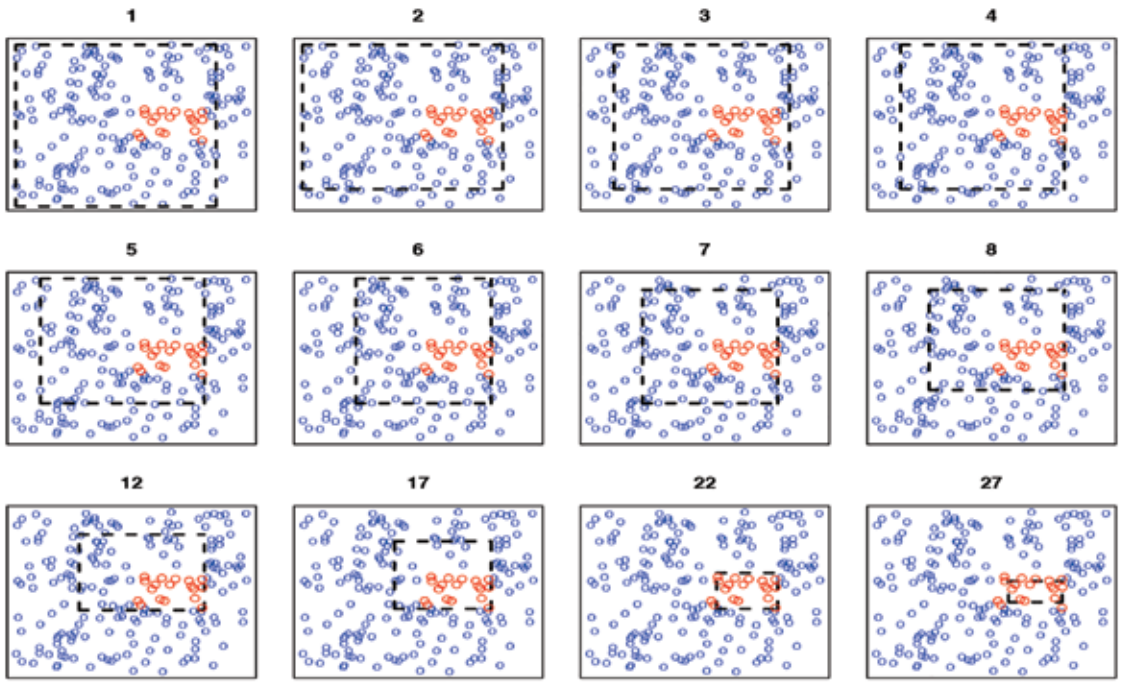


Figure 3. Illustration of PRIM algorithm. The outcome takes two values, 0 and 1, indicated by the blue and red points, respectively. The procedure starts with a rectangle (broken black lines) surrounding all of the data, then peels away points along one edge, by a prespecified proportion α (here, $\alpha = 0.1$), with the edge chosen to maximize the mean of the points remaining in the box. Starting at the top left panel, this figure shows the sequence of peelings, until a pure red region is isolated in the bottom right panel. (Hastie, T.J., Tibshirani, R.J., and Friedman, J.H. 2009. *The Elements of Statistical Learning*. New York: Springer.)

substantially reducing the computational burden, we are able to attempt permutation-based inference as a way to validate detected clusters.

The Patient Rule Induction Method

We have arrived at a setup where, we have data on the location of each star in each decile stratum (x_{GC},y_{GC}, z_{GC}) in Galactocentric coordinates and its total velocity v_{tot}. We treat velocity as an outcome and try to find local modes ("bumps")—velocity maxima or minima—in the feature space defined by position coordinates. Here, feature space is literally space!

Tree-based methods partition the feature space, using binary splits on individual features, into rectangular (box-shaped) regions to make the outcome averages in each box as different as possible. Since partitioning into boxes is recursive, the rules defining the boxes can be represented by a binary tree. Such representations enhance interpretability because they mimic decision trees.

PRIM also finds box-based clusters in feature space, but seeks boxes in which the velocity outcome average is high or low. The search for outcome maxima—fast star groups—is conducted as follows, with obvious modifications for finding slow groups.

The first step commences with a box containing all the data. This box is slightly reduced by removing a small number of observations along one face of the box, with the face chosen for this *peeling* so the mean outcome of the reduced box is maximized.

The number of observations considered for removal is governed by the peeling tuning parameter α; throughout we use the common default value $\alpha = 0.1$. The peeling procedure is then repeatedly reapplied to the reduced box until the number of points therein reaches a prescribed *minimum mass* value. This minimum mass is specified as a proportion of the initial sample size; to elicit OCs with as few as 50 stars, we use a value of 0.0005.

There is provision for expanding the resultant solution box by reverse peeling according to distinct *pasting* tuning parameter if such expansion increases the box mean; we have not deployed such expansion. Observations in the resultant box are then removed from the original data set and the entire procedure begun anew, generating a sequence of solution boxes.

Figure 3 illustrates one round of this process. While this schematic also uses peeling parameter $\alpha = 0.1$, our application differs in that (i) instead of two dimensions, we have three corresponding to star position coordinates; (ii) instead of a binary outcome, there is a continuous (total velocity) outcome; and (iii) instead of 200 data points, there are, for example, 170,293 for each decile partition.

PRIM differs from tree-based partitioning methods in that the box definitions are not described by a binary tree. This makes interpretation of the collection of rules more difficult; however, by removing the binary tree constraint, individual rules are often simpler. Since we are interested in isolating select velocity clusters, as opposed to characterizing the velocity distribution of the entire MW, this latter simplicity is preferable.

More important for our purposes is the built-in patience of PRIM, reflected by each successive peel only removing a small portion (here, 10%) of the current sample, as opposed to the much-greater data fragmentation inherent to the top-down sample splitting that arises with tree-structured methods.

By conducting peeling with reference to the given coordinates (features), it is evident that PRIM box solutions will depend on the coordinate system chosen. To both improve performance and attain coordinate system invariance, executing a principal components rotation to principal axes has been proposed before applying PRIM. However, as discussed above, no such rotation is necessary because (x_{GC}, y_{GC}, z_{GC}) already constitute principal axes for this *Gaia* DR2 subsample.

Like clustering algorithms, the PRIM procedure will generate a sequence of solution boxes regardless of the input data. Accordingly, methods are required to assess if the results are "real." If the focus is solely on the first solution box generated, then cross-validation approaches can be used to determine, on a predictive basis, an appropriate trade-off between box mean (average v_{tot} for stars in the box) and mass (number of stars in the box) based on the sequence of peeling steps.

However, since we are interested in appraising all solution boxes as potential kinematic groups, an alternate approach is needed. It is crucial to account for the adaptive search underlying PRIM solutions in any such approach. We achieve these goals by use of permutation-based inference, reapplying PRIM to scrambled data sets where, within each r_{GC} partition considered, we permute star velocity values over the (fixed) spatial positions.

Doing this a great many times generates the permutation distribution of the *sequence* of solution box mean velocity values since we are re-applying the entire PRIM procedure, generating multiple solution boxes per permutation. In addition, the individual permuted velocities of each solution box are retained, facilitating comparisons—for example, via t-test—with the velocities of the stars in the original (unpermuted) solution boxes that incorporate within-box velocity variation.

The main drawback to this strategy is computational, which can be redressed at least partly by parallelization.

Results

Figures 4–8 provide a brief but representative survey of our findings.

The histogram panels depict results from the original and permuted PRIM analyses for selected r_{GC} partitions. In each instance, we highlight findings for the top three (most extreme group velocity averages) boxes along with the median velocity box. Figure 4 pertains to searching for groups of slow moving stars in the r_{GC} band defined from the 0th to 5th r_{GC} quantiles, corresponding to 85,147 stars between 3,193 and 6,663 *pc* from the GC. The original PRIM analysis extracted a total of 1,432 boxes.

The slowest observed group (box 1, red vertical line) contains 44 stars moving with an average total velocity of $v_{tot} = 195$ km s^{-1}. This is appreciably less than the total velocities of the slowest boxes obtained by PRIM after permutation (blue histogram).

Further, using t-tests to compare the velocity distribution of the original slowest box with the velocity distribution of its permuted counterparts obtains significant results, with the p-value interquartile range (IQR) being (0.04,0.08). The differences between original and permuted findings diminish when comparing second and third boxes, and are null by the median box. When we seek fast-moving groups in this r_{GC} band, none of the results are statistically significant.

Indeed, for most bands, and for both fast- and slow-moving group detection, results are null. Figure 5 provides a representative illustration that pertains to searching for fast-moving groups in the fourth r_{GC} decile (7,921 to 8,055 *pc* from the *GC*). As discussed above, such findings are expected in view of stars in the young thin disk being bound by the disk's gravitational potential. However, that does not mean that the groups (boxes) identified are not OCs; rather, they are just not distinct with respect to total velocity.

Figure 6 showcases significant findings for fast-moving groups from the 90th to 95th r_{GC} quantile band, corresponding to stars between 8,828 and 9,349 *pc* from

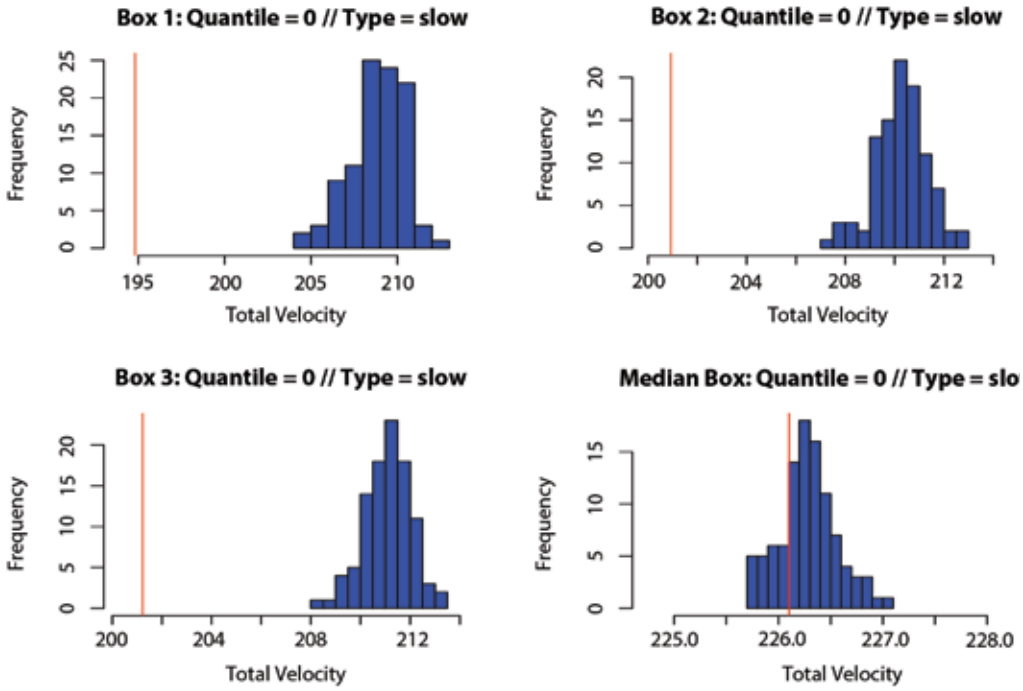


Figure 4. PRIM results for select (top three, median) slow-moving groups (boxes) for the band between 0th and 5th r_{GC} quantiles. The red vertical line indicates the observed average total velocity for the respective groups while the blue histograms give null distributions, as obtained by permuting velocities and re-applying the PRIM procedure.

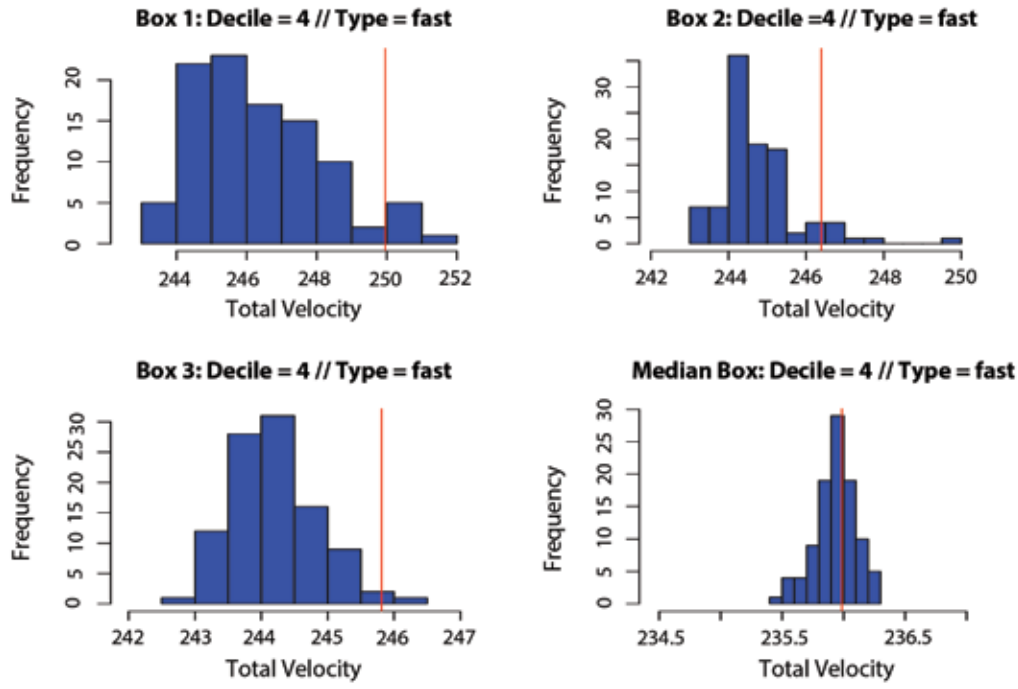


Figure 5. As for Figure 4 for select fast-moving groups (boxes) for the band between 3rd and 4th r_{GC} deciles.

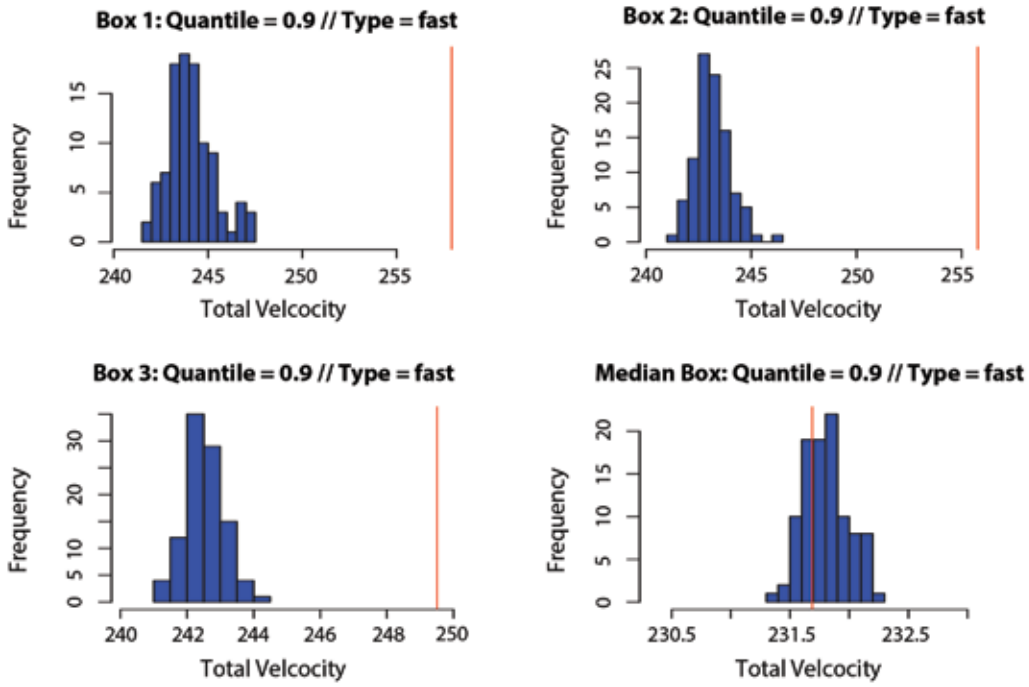


Figure 6. As for Figure 4 for select fast-moving groups (boxes) for the band between 90th and 95th r_{GC} quantiles.

the GC. The fastest observed group (from a total of 1,566 groups) contains 45 stars moving with an average total velocity of v_{tot} = 258 km s^{-1} (box 1, red vertical line), notably faster than the total velocities of the fastest boxes obtained by PRIM after permutation (blue histogram).

However, as expected, this *group average* velocity is dramatically less than the velocity (568 km s^{-1}) of a recently characterized *individual* HVS ejected from the inner disk, while possibly the three fastest stars in the MW—white dwarf companions of supernovae—have velocities exceeding 1,000 km s^{-1}. The *t*-test comparisons with total velocity distributions of permuted counterparts produce a *p*-value IQR of (0.01,0.03).

Again, as expected, differences between original and permuted findings diminish in progressing to comparing second and third boxes and are null for the median box.

Figure 7 depicts the positions of these extreme (fast and slow) groups in the portion of the Galactic disk covered by our *Gaia* DR2 subsample. Because these arise in outlying quantiles, the box extent is greater than for discoveries in more-central, confined r_{GC} bands. This is illustrated in Figure 8, which highlights the slowest box from the 8th decile (8,343 to 8,497 *pc* from the *GC*), containing 87 stars, with average total velocity v_{tot} = 215 km s^{-1} against a restricted portion of the Galactic plane.

Future Work

There are many possibilities for further study. In terms of PRIM, methods that use more-robust v_{tot} summaries and simultaneously capture variation warrant consideration. Improvements in computational efficiency are important for enabling expanded search and more-effective permutation-based inference.

In addition to isolating (rare) extreme kinematic groups, the PRIM technique also delivers numerous candidate OCs. Techniques for prioritizing and interrogating these solutions deserve attention. The nature of such interrogation depends on available data. For example, to follow up on select HVS identified from *Gaia* DR2 total velocities, use of targeted high-resolution spectroscopy enabled determination of chemical abundance patterns which, in turn, were revealing about their origins.

With putative kinematic groups and OCs, analogous follow-through is essential for verification for beyond the purely statistical and downstream interpretation. While some headway along these lines is possible using the photometric data available from *Gaia* DR2 and inspection of associated color-magnitude diagrams, detecting the characteristic signatures

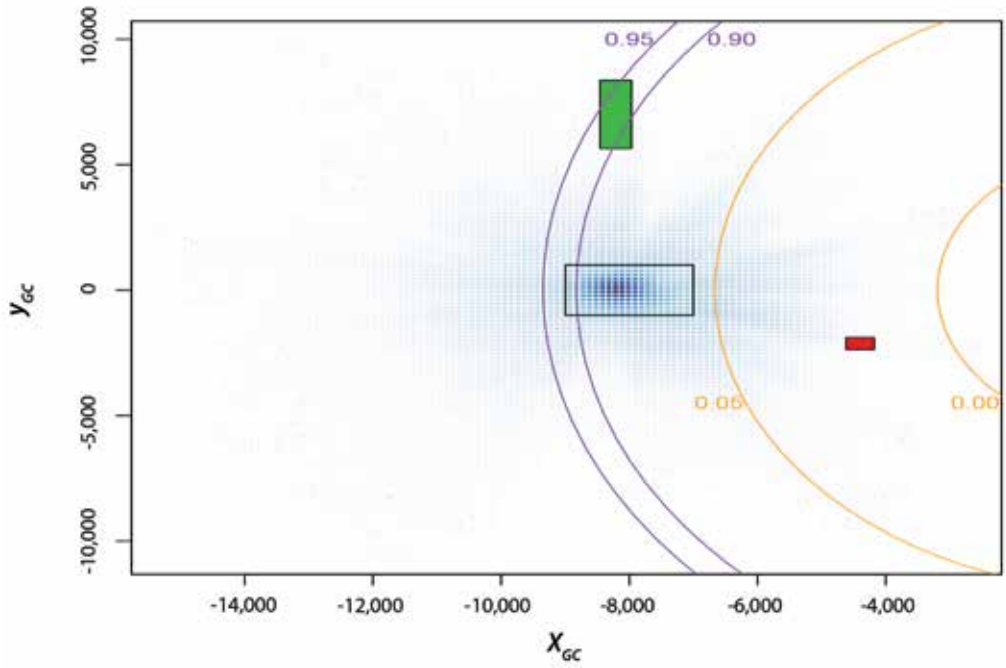


Figure 7. Smoothed scatterplot of star density in the Galactic plane (z_{GC} = 0) for Gaia DR2 subsample. Highlighted are boxes indicating locations of the fastest (green) and slowest (red) kinematic groups in their respective r_{GC} bands, depicted via semi-circles: 90th to 95th r_{GC} quantile (purple) and 0th to 5th r_{GC} quantile (orange). The fast group appears to extend beyond its band because the box is (i) projected onto the Galactic plane and (ii) a bounding box as opposed to the convex hull of contained points. Also shown (black rectangle) is the region expanded in Figure 8.

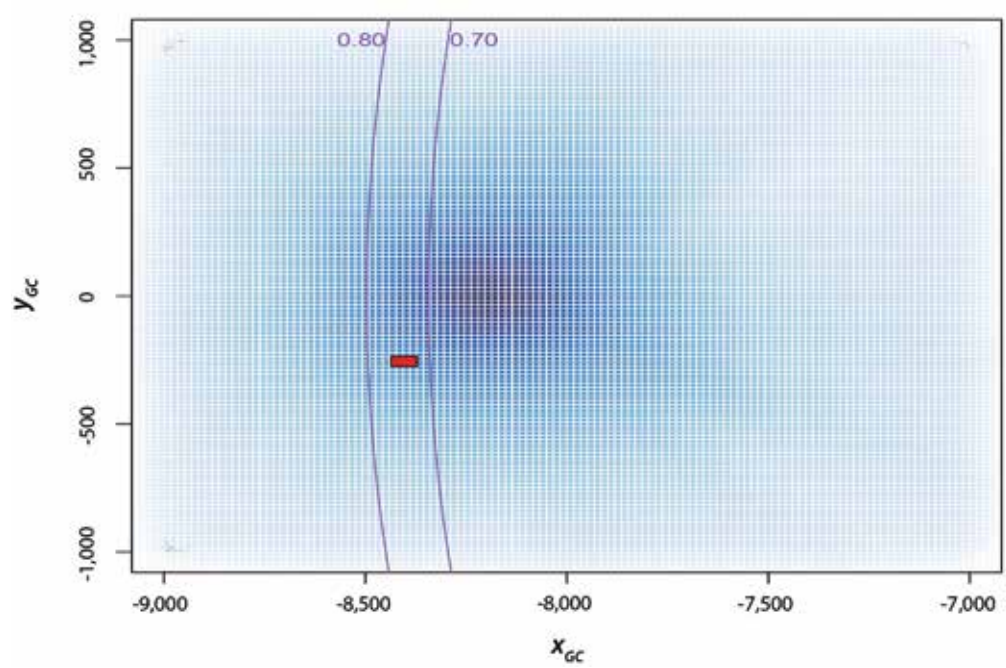


Figure 8. Smoothed scatterplot of star density in a restricted portion of our Gaia DR2 subsample. The box indicates the position of the slowest kinematic group in the band between the 70th and 80th r_{GC} quantiles, depicted by the purple semi-circles.

of OCs from these diagrams can be challenging with small cluster sizes. Some preliminary attempts to impart an algorithmic basis (using artificial neural networks on a set of manually labeled diagrams) for such detections have been made—another area that warrants additional data analytic efforts.

Glossary of Astronomical Terms

These astronomical terms and constructs are described, where appropriate from the standpoint of stars in the Milky Way, because they are the units of the analyses discussed here.

Celestial sphere: an abstract sphere that has an arbitrarily large radius and is concentric to the Earth on which all stars in the sky can be conceived as being projected. Accordingly, a star's position can be described via two coordinates analogous to latitude (declination) and longitude (right ascension).

Color–magnitude diagram: a plot of stars' brightness versus color, usually for a cluster so the stars are all at approximately the same distance, from which various attributes can be inferred.

Declination: the angular distance of a star north or south of the celestial equator, analogous to latitude. Declination and right ascension, an east-west coordinate, together define the position of a star in the sky.

Galactic Center: the rotational center of the Milky Way, approximately 8,200 parsecs away from Earth, in the direction of the constellation Sagittarius.

Galactic disk: the (relatively) thin plane containing the bulk of the Milky Way's stars.

Galaxy rotation curve: plot of orbital speeds of visible stars in the galaxy versus their radial distance from the galaxy's center.

Globular clusters: a spherical collection of tightly gravitationally bound stars, residing in the halo, and orbiting the Galactic Center.

Halo: an extended, roughly spherical component of a galaxy, comprising sparsely scattered older stars, globular clusters, and gas.

Hypervelocity stars: stars with velocities greatly exceeding the normal velocity of stars in a galaxy.

Open clusters: stellar groups comprising dozens to thousands of stars, formed from the same molecular cloud around the same time and loosely bound by gravitational attraction.

Parallax: difference in the apparent position of a star when viewed along two lines of sight, measured by the angle of inclination between these lines, from which distance to the star can be determined.

Photometry: measurement of the apparent brightness of stars and informative with respect to their temperature, distance, and age.

Proper motion: a star's angular change in position over time, measured in arc-seconds per year; it is a two-dimensional vector (since it excludes the component in the line of sight direction) defined by the angular changes per year in the star's right ascension and declination.

Radial velocity: the velocity of a star along the line of sight of an observer.

Right ascension: the angular distance of a star measured eastward along the celestial equator from the Sun at the Vernal equinox; analogous to longitude.

Stellar association: a loose star cluster containing up to 100 stars that share a common origin but, while gravitationally unbound, are still moving together through space.

Young thin disk: a structural component of the Milky Way composed of stars, gas and dust with a scale height of approximately 100 parsecs in the vertical axis perpendicular to the disk. C

Further Reading

Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., Balaguer-Núñez, L., Soubiran, C., Bossini, D., Moitinho, A., Castro-Ginard, A., Krone-Martins, A., Casamiquela, L., Sordo, R., and Carrera, R. 2018. A *Gaia* DR2 view of the open cluster population in the Milky Way. *Astronomy & Astrophysics* 618, A93.

Castro-Ginard, A., Jordi, C., Luri, X., Julbe, F., Morvan, M., Balaguer-Núñez, L., and Cantat-Gaudin, T. 2018. A new method for unveiling open clusters in *Gaia*—New nearby open clusters confirmed by DR2. *Astronomy & Astrophysics* 618, A59.

Ferreira, F.A., Santos Jr., J.F.C., Corradi, W.J.B., Maia, F.F.S., and Angelo, M.S. 2019. Three new Galactic star clusters discovered in the field of the open cluster NGC 5999 with *Gaia* DR2. Monthly Notices of the Royal Astronomical Society. *https://doi.org/10.1093/mnras/sty3511*.

Friedman, J.H., and Fisher, N.I. 1999. Bump hunting in high-dimensional data. *Statistics and Computing* 9, 123e143.

Marchetti, T., Rossi, E.M., and Brown, A.G.A. 2018. *Gaia* DR2 in 6D: Searching for the fastest stars in the Galaxy. Monthly Notices of the Royal Astronomical Society. *https://doi.org/10.1093/mnras/sty2592*.

About the Authors

Mark Segal is a professor in and head of the Division of Bioinformatics in the Department of Epidemiology and Biostatistics at the University of California San Francisco. His current research focus is primarily in computational biology.

Jacob Segal is a senior studying astrophysics at the University of Colorado, Boulder. He works in a stellar variable research group.

Time Series Clustering Methods for Analysis of Astronomical Data

David J. Corliss

Cluster analysis, often referred to as segmentation in business contexts, is used to identify and describe subgroups of individuals with common characteristics that distinguish them from the rest of the population. While segments are often identified using static characteristics, evolving systems may be better described by how things change over time. A medical patient may be classified by the amount of time since an important event such as a diagnosis, economic activity may be segmented by stages in an economic cycle, and neighborhoods grouped by stages in generational evolution. In astrostatistics, this technique is used to classify a supernova by how the amount of light it produces changes over time.

Depending on the method used, clustering time series data can identify a series of steps or phases within a time series, or identify groups of time series with a similar pattern (Similarity Analysis).

The use of cluster analysis in statistics to identify distinguishable subpopulations goes back to the 1939s, with a textbook by the behavioral psychologist Robert Tryon. Most statistical software systems support cluster analysis, with output typically including summary statistics on the final clusters, metadata on the iterations need to create them, and an output data set with a field identifying the cluster for each record.

A Common Clustering Example: The Anderson Iris Data

Cluster analysis is often taught using a data set of iris flowers collected by the American botanist Edgar Anderson. These data are often known as the Fisher Iris Data, due to their later use for discriminant analysis by the statistician Ronald Fisher. Figure 1 shows a plot of these data, based on two characteristics that tend to be distinct by species: petal length and sepal width.

Cluster Analysis Applied to Time Series Data

While most people in statistics are familiar with clustering similar to this standard example, cluster of time series data can be very different. As one example, a dynamically evolving system often goes through a series of distinct steps or stages as it changes. Figure 2 gives the example of heat applied to ice to melt it, with the temperature recorded at periodic intervals. When the data in this time series are clustered, a succession of steps can be seen.

First, the temperature of the ice increases. As the ice melts, the material remains at the freezing point. Once all the ice has melted, the resulting water is heated and the temperature begins to rise again.

As with other clustering applications, the data in each cluster share important characteristics that are distinct from the other clusters. In these time series data, however, these clusters are successive, representing stages in a process. The application of clustering methods to time series data can identify the steps in a process or stages in evolutionary development.

Time series cluster analysis can be applied to phenomena that repeat in a periodic manner, such as internet traffic over a 24-hour period or seasons in a year. However, not every event of interest has a fixed duration and some may not repeat at all. While some one-time events last for a standard length of time, the most-general case is with events that follow an exact sequence but vary in overall duration.

In astrophysics, the violent stellar eruptions known as High Velocity Absorption (HVA) events in some late B- and early A-type supergiant stars appear to follow a definite sequence but vary in duration by an order of magnitude or more. This work in time series cluster analysis originally was undertaken to identify the evolutionary stages of these one-time events of variable duration.

This analysis of events with variable duration re-normalizes the time values using the difference between two benchmark points in

Figure 1. Typical clustering example with iris flowers observed by Edgar Anderson (1936) grouped by characteristics distinctive to each species.

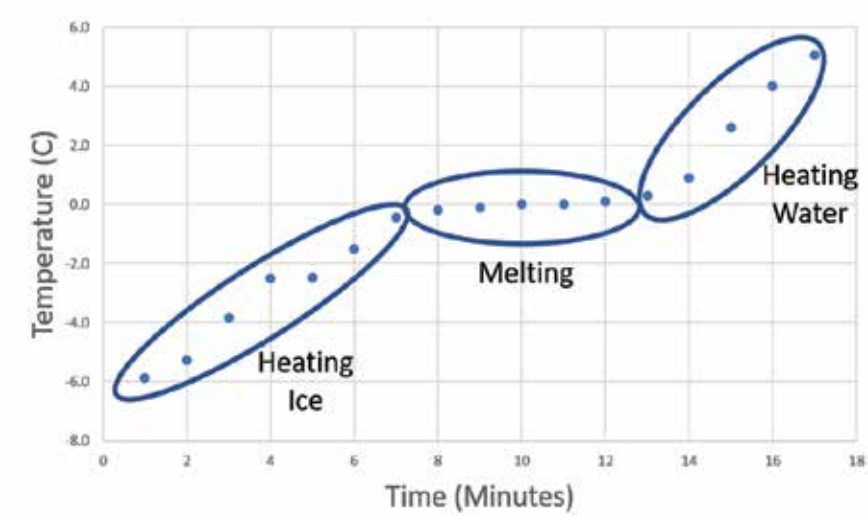


Figure 2. Time series of melting water ice.

time. Often, as in the case of HVAs, the beginning and the end of the event provide the overall duration. The time data values are then transformed to a percent of the time from beginning to end. This analysis of HVAs extracts the initial and final dates and then merges them with each record. The percent of the total time elapsed from the beginning is calculated and then used as the time variable in the cluster analysis.

The number of clusters was identified by initially generating a large number of clusters and then

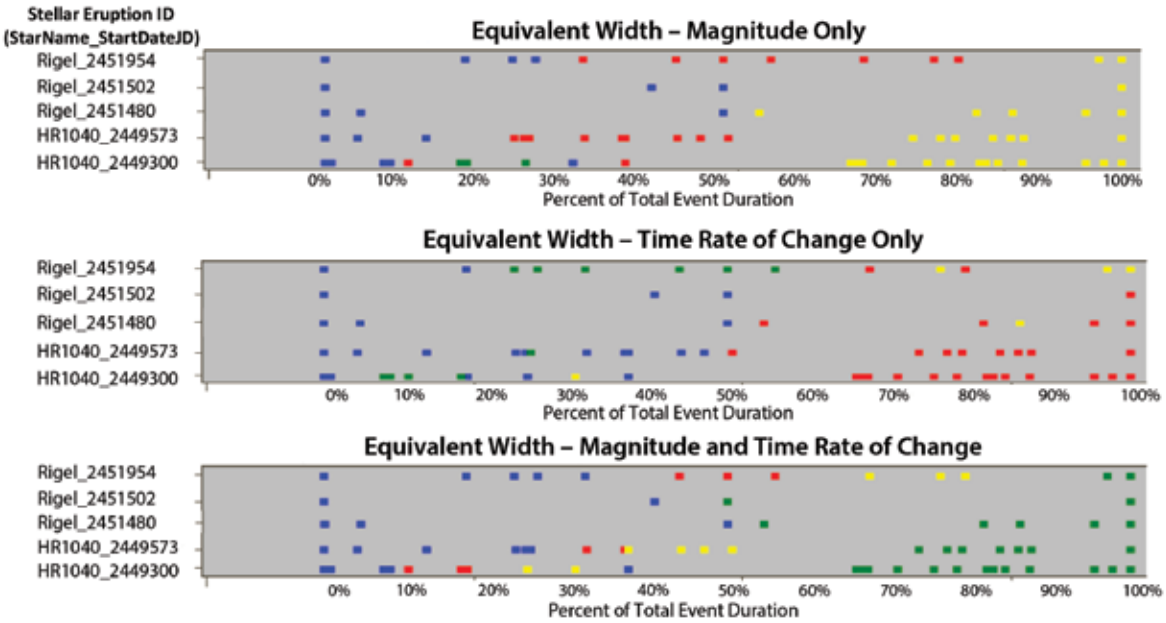


Figure 3. Testing point-in-time and rate of change variables in five HVA events. These plots test which combination of observed quantities provides the best separation of the clusters and thus defines the successive stages of these events more clearly with distinct behaviors at each stage. The system with most distinct clusters, both point-in-time and rate-of-change variables (bottom plot), has the clusters in the following order: ■ Phase 1 ■ Phase 2 ■ Phase 3 ■ Phase 4.

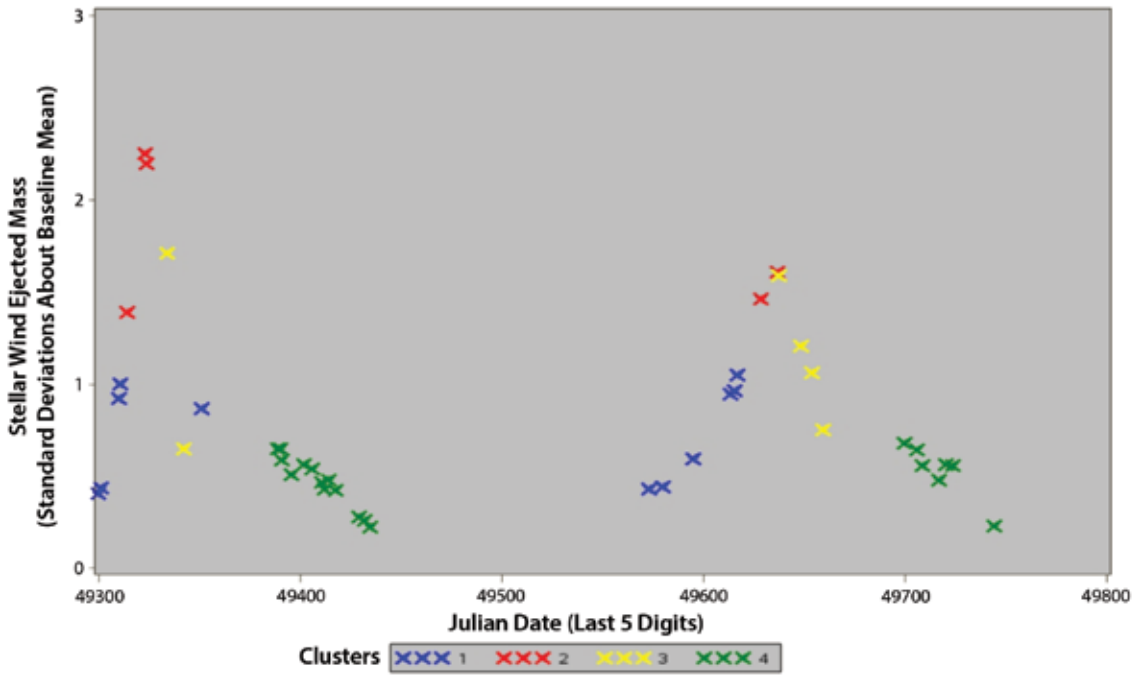


Figure 4. Time series cluster analysis of HVA events in the A0 supergiant HR 1040. The vertical scale is the number of standard deviations above the baseline mean during "quiescent" phases—with no stellar eruptions. The horizontal scale is time, given by the last five digits of the Julian date. This dating system, commonly used in astronomy time series analysis, avoids negative dates by recording the number of days since January 1, 4713 BCE, since there are few, if any, astronomical observations with individual dates before this point in time.

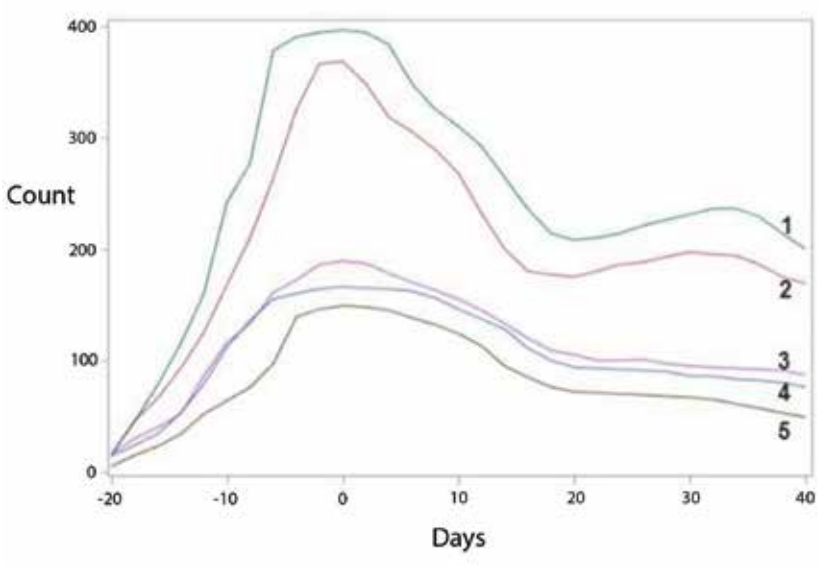


Figure 5. Infrared light curves from five Type 1a supernovae.

"pruning"—combining clusters with similar values. HVA events appear to go through up to four distinct stages, so this was set as the number of clusters in the final run of the clustering algorithm.

While this study used the SAS procedure FASTCLUS, other clustering algorithms or packages could be employed.

In this analysis, point-in-time data was used to calculate the rate of change of the variables since the previous observation. This case of HVAs in late B- and early A-type stars gave the clearest separation of clusters when both the observed values and their rates of changes were included in the analysis (Figure 3).

The first stage in these events is characterized by a rapid increase in the quantity and velocity of material ejected from the star, while the absolute amounts remain fairly small (Cluster #1), followed by a rise to a sharp peak (#2), and then a rapid decrease of 60%–70% (#3). After an interval with little change, there is a final drop back down to zero intensity (#4). This sequence of evolutionary stages is seen in two HVAs in Figure 4.

Similarity Analysis: Classifying Time Series by Pattern

Similarity analysis is a statistical method for comparing and clustering different trends or patterns over time. It uses a distance measure to quantify the difference in form or shape between different time series: Two series get a small similarity distance if their pattern is similar, but receive a large number if the patterns are different. The mutual differences between many sequences or time series can be compiled in a similarity matrix, similar in form to a correlation matrix using multiple correlations.

When cluster methods are applied to a similarity matrix, time series with similar patterns are grouped together and placed into clusters distinct from others containing times series with different patterns. This allows data captured over time to be classified into groups. This example uses similarity analysis to classify supernovae by their light curves by how the amount of light they produce changes over time.

Figure 5 shows five light curves, with each demonstrating the rise and fall in brightness from a different Type Ia supernova observed in the Sloan Digital Sky Survey (SDSS) in infrared light. (David Cinabro of Wayne State University provided a copy of the source data.)

In this example, similarity analysis reads and quantifies the differences in shape between the five lines and creates two groups. The first group, with lines 1 and 2, has a higher, faster peak and a secondary peak later on. The second group, with lines 3, 4, and 5, has a slower initial growth rate and no secondary peak.

The difference in shape can be captured and quantified by a process called Dynamic Time Warping (DTW). DTW involves drawing scale lines between two time series, resulting in a need for additional lines when difference in shape occurs. The similarity distance is the sum of the length of the lines, so shapes with only a few

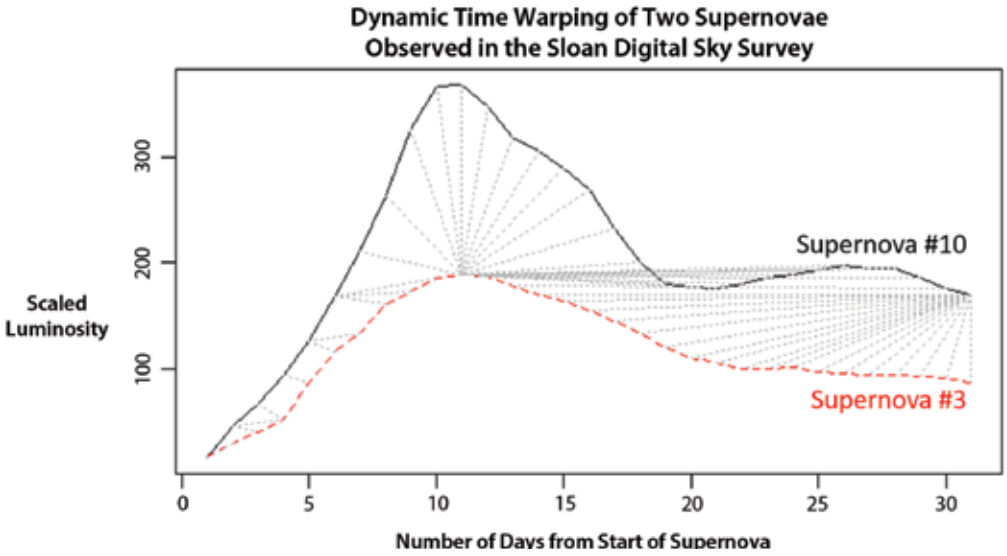


Figure 6. Dynamic time warping to calculate the similarity distance between two light curves, produced using the R package DTW.

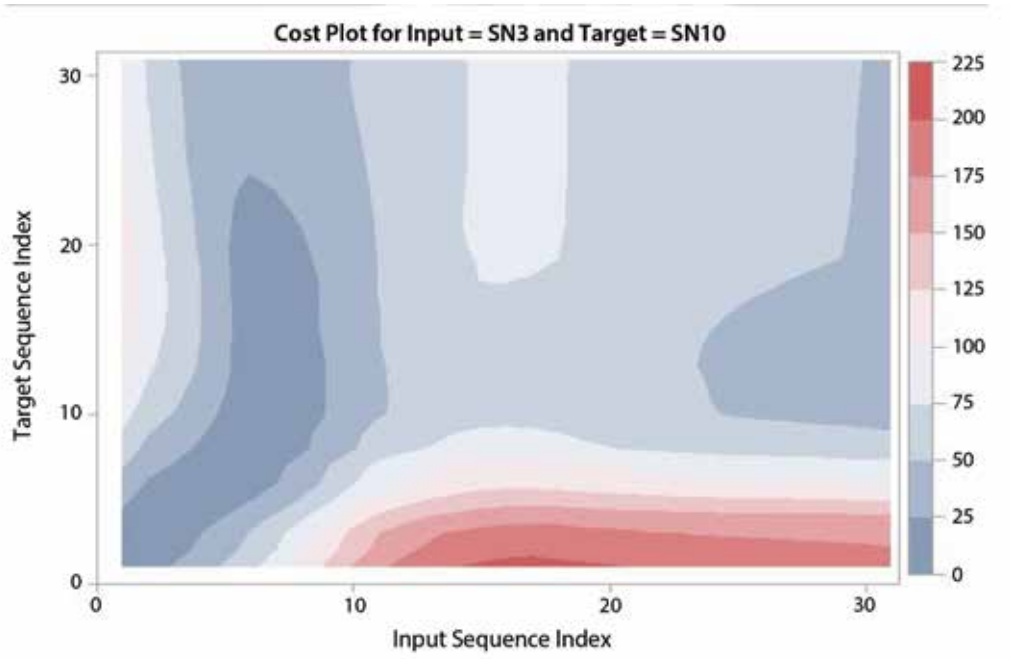


Figure 7. Dynamic time warping to calculate the similarity distance between two light curves.

Table 1—Similarity Matrix for the Infrared Light Curves of Six Supernovae in the SDSS Data

	TS1	TS2	TS3	TS4	TS5	TS6
TS1	0	817	100	469	338	307
TS2	817	0	422	644	871	895
TS3	100	422	0	92	835	818
TS4	469	644	92	0	56	206
TS5	338	871	835	56	0	378
TS6	307	895	818	206	378	0

minor differences have a smaller distance than time series with a greater difference in shape. The "warping" refers to non-linear re-scaling of the time axis to find the optimal (shortest) path independent of changes in speed.

For example, DTW works on EKG heartbeat time series data, even though heartbeats can have different durations. Two software packages capable of performing DTW are the SIMILARITY procedure in SAS, which was used in this study, and the DTW package in R. In addition to calculating the similarity distance, these also provide several important plots to visualize the data to facilitate analysis and interpretation.

A Warp Plot (Figure 6) shows the lines describing the difference in shape between two time series. The sum of the lengths of the line segments gives the similarity distance.

A Cost Plot (Figure 7) provides a different way of visualizing the difference in shape between two time series. Blue areas indicate areas of greater similarity while red ones indicate more difference; for example, at the end of these two time series, where the upper one declines monotonically while the lower one shows a second peak. This cost plot compares the same two supernova events plotted in Figure 6.

Preparing the Data and Calculating the Similarity Distance

Some quantities are highly variable, making classification using cluster analysis more difficult. Fields with larger variances receive more weight in determining clusters. Also, if the average value of a field steadily increases or decreases over time, the weight given to that field will also change. In these circumstances, a transformation of volatile field to the number of standard deviations above and below the mean before similarity analysis corrects for weighting by normalizing fields to the total amount of variation for each field in the data set.

Other data issues may arise. The time series of some quantities can be autocorrelated. For example, values in the short-term future may be correlated more strongly to values in the recent past than to those in the more-distant past. As in earlier examples, the time rate of chance of a quantity may characterize the behavior better than values observed at one moment in time. Applications of similarity analysis in econometrics may involve prices that are subject to inflationary changes over time, requiring transformation to monetary values at a specific point in time (e.g., 2019 dollars).

Gasoline prices provide an excellent example of all of the data issues: They are highly volatile and often show autocorrelation over the short term and inflation effects over the long term. With proper preparation of the data, modeling the impact of events on a volatile and sensitive commodity such as gasoline can be performed using similarity analysis to compare the event to previous price shocks.

Classification using similarity analysis involves these steps:

1. Calculate the similarity distance between the time series and placed it in a similarity matrix. Data visualizations support investigate the differences between time series.

2. Apply cluster analysis to the similarity matrix to identify clusters of time series with similar forms.

Table 1 is the similarity matrix containing the similarity distance

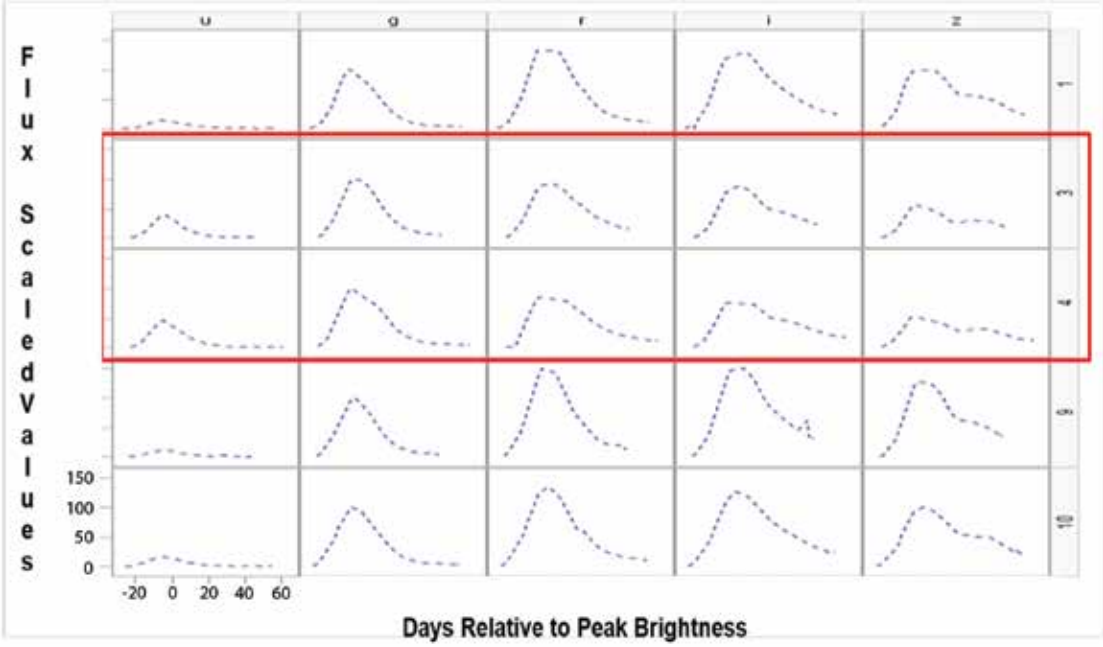


Figure 8. Additional cluster characteristics. Supernovae 3 and 4, with a second IR peak, also have a higher peak in the UV. All flux values scaled to maximum flux in g (visible) = 100.

calculated for each pair from six supernova events in the Sloan Digital Sky Survey. The matrix shows the amount of difference in shape between any two time series. For example, TS4 and TS5 are very similar in shape, so they have a small number (56). In contrast, TS4 is very different from TS1, resulting in a very large number (469). A number of visualizations can be created as this stage as well, including warp and cost plots.

Clustering the Time Series by Similarity Distance

Once the similarity distances have been calculated, the time series can be clustered using standard methods. Developing the clusters uses all time series in the SDSS data with sufficient observations to describe the shape in detail—78 time series in all.

In the case of the infrared light curves from type 1a supernovae, two distinct clusters can be seen. One cluster has a distinct *second peak in the infrared* some 35 days after the initial, maximum peak brightness. The other cluster does not have this second IT peak, fading away monotonically (Figure 5).

Once the clustering is complete, additional characteristics of the clusters may be identified. This trellis plot (Figure 8) displays data for the five supernovae plotted earlier. In addition to the infrared light (z, in the column on the right) used for the similarity analysis, it plots four other colors of light: u (ultraviolet), g (green), r (red), and I (near infrared). The similarity analysis using all the time series with sufficient data yields two clusters: one for those with a secondary peak in the far infrared one without. When all the data are plotted in the panel, additional characters of the clusters can be identified. In this case, the g, r, and i columns are very similar. However, the ultraviolet u column on the left shows a large peak in the first cluster with minimal increases in the u data for the other cluster. It may be that the mechanism or characteristics responsible for the secondary far infrared peak that first attracted attention is also responsible for the much-stronger ultraviolet radiation seen in this sub-type.

Similarity analysis is an emerging method for the classification of time series. It uses a distance measure to quantify the difference in form or shape between the two series, with a lower number indicating greater similarity. The

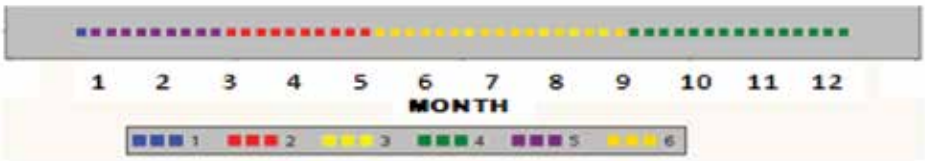


Figure 9. Seasonality in gasoline prices. This set of clusters shows a post-holiday lull in the first week of the year (cluster #1), a winter season with low prices and abundant supply (#5), a run-up of prices and supply shortages from mid-March through the end of May when refineries are changing over from winter formulations to summer (#2), a summer driving season with significant supply but higher demand and occasional spikes (#3 and #6), and a gradual decline in prices from mid-September–December (#4).

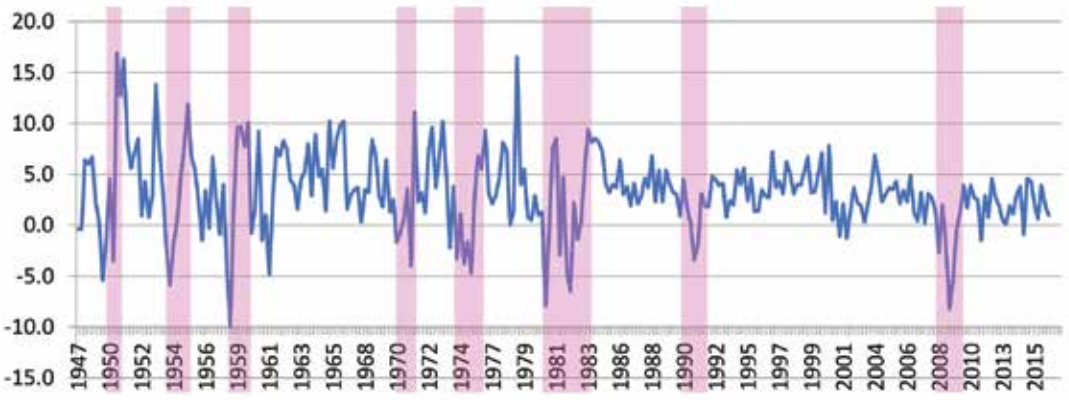


Figure 10. U.S. recession and recovery data (quarterly GDP percent change).

similarity distances for a set of time series can be arranged into a similarity matrix. Classification of time series can be achieved using standard clustering methods on the similarity distance.

In the example, SDSS light curves from a set of Type 1a supernovae indicates the presence of two subtypes: one with a fast initial growth rate and a secondary peak in the infrared z-range, while the second subtype has slower initial growth rate and no secondary infrared peak. Classification of the time series in this way led to the discovery of other differences between the two types, including a larger peak in the ultraviolet in the first sub-type.

Time Series Cluster Analyisis: Other Examples

In astrostatistics, analytic methods often borrow from and contribute to other areas of research. These time series clustering examples, developed to study hot, massive stars, can contribute to other research. In econometrics, time series cluster analysis has been applied to the evolution of market shocks, identifying seasonality in gas prices (Figure 9). Longitudinal studies in time series cluster analysis can investigate seasonal patterns in weather over time due to climate change.

Similarity analysis has been used to classify cycles in

Table 2—Similarity Matrix of U.S. Recession/Recovery Cycles

	Cycle1	Cycle2	Cycle3	Cycle4	Cycle5	Cycle6	Cycle7	Cycle8
Cycle1	0	193.67	185.09	184.12	222.47	341.67	392.48	532.5
Cycle2	193.67	0	71.34	233.86	34.16	280.74	79.89	115.73
Cycle3	185.09	71.34	0	229.11	57.79	287.56	121.01	160.87
Cycle4	184.12	233.86	229.11	0	176.47	317.19	108.08	133.58
Cycle5	222.47	34.16	57.79	176.47	0	281.18	35.96	85.93
Cycle6	341.67	280.74	287.56	317.19	281.18	0	256.74	219.47
Cycle7	392.48	79.89	121.01	108.08	35.96	256.74	0	65.64
Cycle8	532.5	115.73	160.87	133.58	85.93	219.47	65.64	0

Cycle 6, a "Double Dip" recession, is found to be an outlier due to high-similarity distances to all other events in this study.

unemployment and recession/recovery by their pattern over time, with "Double Dip" recessions as a cluster (Figure 10 and Table 2).

Applications are also found in biostatistics. Growth rates of biological systems can be classified or segmented. Similarity analysis has been used to classify Ebola outbreaks in Africa, providing comparisons for future outbreaks by identifying similar events in the past.

Best Practices for Similarity Analysis

Best practices for using similarity analysis include:

- When clustering, examine both point in time variables and rate of change data.
- When data are "unevenly spaced" (that is, the time interval between observations is not constant), treat them as an evenly spaced time series with missing data and impute the missing values.
- Where highly volatile variables are present, consider applying a data transformation such as standardization before clustering.
- Once the data have been transformed, and, in the case of similarity analysis, the similarity distances have been calculated, standard clustering best practices still apply.
- Watch out for outliers, which will appear as time series clusters with very few members.

EXPLORE MORE

Joseph M. Hilbe wrote an article about astrostatistics for the January 2014 issue of *Amstat News*. Titled *"Astrostatistics: The Re-Emergence of a Statistical Discipline,"* it is an excellent place to begin to learn more.

The Center for Astrostatistics at Penn State University has a tremendous wealth of resources for this area of research. They maintain a web portal at asaip.psu.edu. Center leaders Eric Feigelson and G. Jogesh Babu have written a textbook, *Introduction to Astrostatistics and R*, available from the Berkeley Center for Cosmological Physics.

About the Author

With a PhD in astrophysics from the University of Toledo, **David J. Corliss** performs time series analyses of evolving stars and stellar populations. He is active in the ASA, serving on the steering committee of the Conference on Statistical Practice and as past president of the Detroit chapter. Statistical methods from his astrophysics research also find application in Data for Social Good, including the monthly column he writes for *Amstat News*.

Mary W. Gray
Column Editor

Alexa Did It!

We previously examined the dangers inherent in the ubiquity of algorithms, particularly when their structure is difficult to discover and evaluate due to intellectual property protection. For example, how can a defense attorney ascertain the factors used in a secret sentencing algorithm?

What if, instead, we are dealing with the results of what a machine has learned?

What, aside from a technical definition, is machine learning? Is it only new technology or does it raise questions about its form of life? Does it just broaden the scope and depth of issues such as privacy, security, and the unethical and unlawful use of data, or are we facing a fundamental change in species?

What is the character we assign to the "machines" of machine learning? We may have scoffed at the renegade "Hal" of *2001*, but where would the responsibility lie for resulting actions were we faced with such a situation? Leaving aside the controversial question of when human personhood begins, we know that some decidedly non-human entities have been found by the U.S. Supreme Court to have certain "rights," such as to contribute to a political campaign,[1] but responsibilities? Maybe to pay taxes, but not to be guilty of murder?

Corporate structures in general may be complex and obscured, but at each step, at least in theory, we can discern a human hand. Now, though, we are talking about autonomous diagnostic and treatment tools, vehicles, identification and selection programs, weapons—who knows what next? How will the machine cope with the "unknown unknowns" of Donald Rumsfeld? What happens when adversarial attacks occur—a few pixels can change the world?

These questions go to the nature of machine learning, a nature that is also critical to the kind of intellectual property protection it receives. The U.S. Constitution provides for patent and copyright protection:

To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries.

But the history of exactly what protection should be accorded to new forms of science and useful

[1] *Citizens United v. Federal Election Commission*, 558 U.S. 310 (2010).

arts is not straightforward. Thousands of already-granted patents of genes were invalidated when, in 2013, the Supreme Court ruled that human genes cannot be patented because DNA is a "product of nature."[2] But DNA manipulated in a lab, specifically complementary DNA (cDNA), remained eligible. This synthetic DNA is produced from the messenger RNA that provides the instructions for making proteins. Is there an analogy in artificial neural networks?

To achieve patent protection, an invention or discovery must be novel and non-obvious, but also involve patentable subject matter; "a useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof may obtain a patent therefore, subject to the conditions and requirements of this title."[3]

Is machine learning patent-eligible? A patent gives its owner the legal right to exclude others from making, using, selling, and importing an invention for a limited period of years, in exchange for publishing an enabling public disclosure of the invention. The disclosure requirement can be circumvented by relying instead on trade secret law, where the essence of the protection is secrecy.

The tale of patents of software is a tangled one. In 1972, the Supreme Court decided that an algorithm for correcting binary-coded decimal numbers into pure binary numbers was an abstract idea and thus not patent-eligible.[4] Six years later,[5] the court decided that merely tying an algorithm to a process did not make the process patent-eligible; rather, the requirement was whether a transformation of an article from one state to another occurred. Aha! How about a computer program using a well-known formula (the Arrhenius Equation) to calculate when rubber was transformed from uncured to cured?[6]

A series of cases followed, based on the notion that a novel algorithm combined with a trivial physical step constitutes a novel physical device—that is, load an algorithm on a computing device to get a new patent-eligible machine. *In re Lowry*[7] held that a data structure on a computer's hard drive is similarly patent-eligible. Although the terminology used in these cases was *mathematical formula* or *algorithm*, other rulings seemed to indicate that algorithms that manipulated something other than numbers (e.g., images) might be treated differently.

Finally, however, the United States Patent and Trademark Office (USPTO) issued new guidelines that led to granting patents to a wide variety of software.

Then, in *State Street Bank v. Signature*, the Federal Circuit[8] held that a numerical calculation that produces a "useful, concrete and tangible result"—e.g., a price—is patent-eligible, opening the door to a deluge of so-called "business method patents." Perhaps the example that best shows the extent of the patent explosion is the suit, later settled, by Amazon against Barnes and Noble for infringement of Amazon's "1-click" shopping method patent. In fact, many thought that the proliferation of patents threatened the innovation the patent system was supposed to promote.

As the pendulum swung back, *State Street* and its progeny were superseded by *In re Bilski*,[9] where the U.S. Court of Appeals for the Federal Circuit Court rejected Bilski's business method patent because it was based on an abstract idea that would largely have preempted hedging as a business practice (back to "you can't patent the quadratic formula" days). In *CLS Bank International v. Alice Corp.*,[10] a business method patent involving a computer-implemented escrow service was rejected, leading to widespread invalidation of patents of software even though the Supreme Court never ruled explicitly on software patents.

But wait: In January of this year, the USPTO issued a new set of rules that would make it easier to patent software! These, too, though. will be tested in the courts.

We ask again, is machine learning software and if so, is it patent-eligible? Exactly what might the claims of the patent application be?

If neither patent nor trade secret protection, how about relying on copyright law, which protects "original works of authorship, fixed in a tangible medium

---

[2] *Association for Molecular Pathology v. Myriad Genetics, Ind.*, 569 U.S. 576 (2013).
[3] 35 U.S.C. §101.
[4] *Gottschalk v. Benson*, 409 U.S. 63 (1972).
[5] *Parker v. Flook*, 437 U. S. 583 (1978).
[6] *Diamond v. Diehr*, 450 U.S. (1981).
[7] 32 F.3d 1579 (Fed.Cir. 1994).
[8] 149 F.3d 1398 (Fed.Cir. 1998).
[9] 545 F.3d 943 (Fed.Cir. 2008)
[10] 573 U.S. 206 (2014).

including literary, dramatic, musical, artistic, and other intellectual works." However, copyrights provide much less protection than do patents; copyright covers only the expression, not the idea.

While "tangible form" may be a problematic requirement here as well as in patent law, courts have had no trouble adapting intellectual property protection to other media not contemplated by the drafters of the Constitution.

Product of nature, process, abstraction, machine in the historic sense—whatever machine learning is, whether patent-eligible or not, the question becomes who owns the patent, copyright, or trade secret it represents—the one who had the idea, the one who structured the algorithm and its training, the one who wrote the code or the "machine" itself and its learning? Millions, if not billions, can be at stake.

We can ask the same question about who is liable for the harm that might be caused when the machine acts as expected or intended, or when it does not. The wrong person is identified; the wrong treatment proscribed; the black ice is not recognized; a woman is thought to be a car; the applicant for housing, employment or educational opportunity is rejected on a discriminatory basis; privacy or security is breached; the innocent is sentenced to death; the wrong house blown up. Is the developer of the algorithm, the organizer of the data, the physician or other actor who implements the output, the coder, the machine itself liable?

A basic concept in law lets loss lie where it falls. The victim goes uncompensated unless it can be shown that the manufacturer acted negligently or unreasonably; that is, was "at fault." But with complicated electronic gear, sophisticated vehicles, complex structures of any kind, it is unreasonable to expect that the end user can ascertain the source of the failure of a product. Then the manufacturer, architect, designer, etc., has to pay even though not at fault. This strict product liability holds the producer responsible when something goes wrong, even without showing that there was any negligence, as would be the case under contract law (and the existence of a contract is also not required in the case of product liability).

Once again, we ask whether software, and in particular machine learning, is a product. Generally, "product" is defined as any movable good—but what is a good? Usually some physical presence is assumed, so as a component in a physical product, software would probably be covered by product liability—but perhaps not standing alone. Liability for risk prediction, diagnosis, and treatment by machine learning has been discussed in the medical field, but not much elsewhere. Moreover, it may be difficult for end users to invoke products liability law because there are no common standards for industry practice in machine learning development. Is it a product and if so, where does the product liability lie?

It has been difficult to avoid introducing the very discriminatory profiling into algorithms that they were purportedly developed to prevent. To guard against this is even more difficult in machine learning. How can we program our ethical and legal principles into the algorithms and be certain the machine will be guided by them? At least recognition of the problem is a first step.

Nor has there been much introspection about the enhanced nature of privacy and security protection required in the face of machine learning. For example, can certain adversarial attacks on machine learning be prosecuted under the Computer Fraud and Abuse Act? How does synthesizing data to facilitate research and enhance privacy affect intellectual property protections? Given the nature of machine learning, who can be said to own the input and the output?

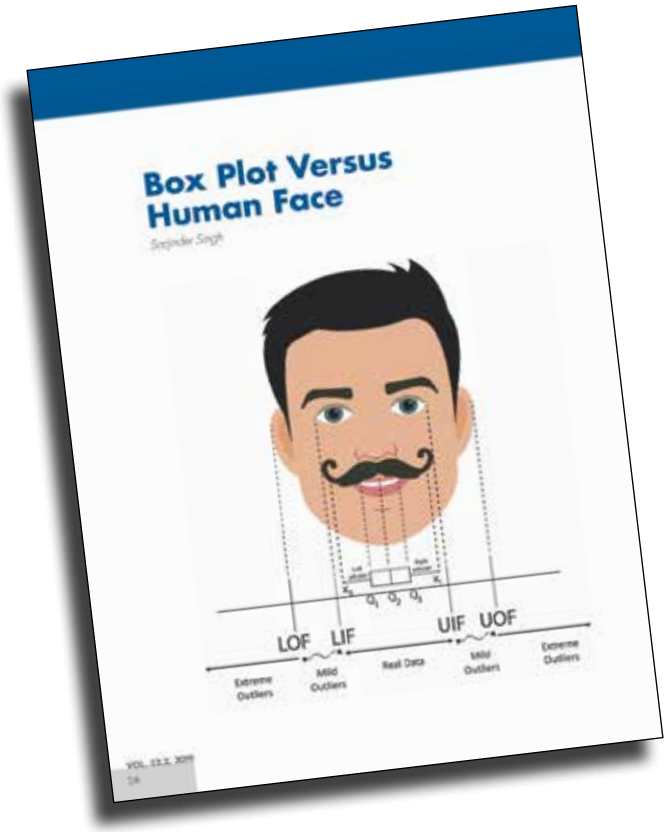
What can be agreed is that developers definitely have to consider the implications of the question of who or what is machine learning. C

About the Author

Mary Gray, who writes The Odds of Justice column, is Distinguished Professor of Mathematics and Statistics at American University in Washington, DC. Her PhD is from the University of Kansas, and her JD is from Washington College of Law at American University. She received the Elizabeth Scott Award from the Committee of Presidents of Statistical Societies and the Karl Peace Award from the American Statistical Association. Her research interests include statistics and the law, economic equity, survey sampling, human rights, education, and the history of mathematics.

Reconsidering the Human Face as Boxplot

David C. Hoaglin



I welcome Sarjinder Singh's creativity ("Box Plot Versus Human Face," *CHANCE* 32(2):28–29), but the boxplot does not align with a human face as closely as he suggests, and it does not classify observations as outliers.

In the standard boxplot (in his notation), the inner fences are:

$$\text{LIF} = Q_1 - 1.5(Q_3 - Q_1)$$
$$\text{UIF} = Q_3 + 1.5(Q_3 - Q_1)$$

and the outer fences are:

$$\text{LOF} = Q_1 - 3(Q_3 - Q_1)$$
$$\text{UOF} = Q_3 + 3(Q_3 - Q_1)$$

(For Q_1 and Q_3 the boxplot uses a particular definition, known as the "fourths.")

Importantly, all the data values are real, but some may merit investigation. As in Professor Singh's construction, the left and right whiskers end at the most-extreme data values that are inside the LIF and UIF, respectively. Data values outside the inner fences are plotted individually, and those beyond the outer fences receive larger plotting symbols.

Exploratory data analysis (EDA) does not automatically classify data values as "outliers." It leaves that judgment to the analyst, after investigation. Data values between an inner fence and the corresponding outer fence are "outside," and those beyond the outer fences are "far out." Thus, Singh's "mild outliers" are merely "outside" and his "extreme outliers" are "far out."

It is helpful to have an idea of how often samples of well-behaved (i.e., normal) data contain observations that are outside or far out. In this null situation, the percentage of random samples that contain one

or more observations beyond the inner fences is 33% for n = 5, 20% for n = 10, 23% for n = 20, 36% for n = 50, and 53% for n = 100; and it approaches 100% as n becomes large. The irregularities in this sequence of percentages arise mainly from the definition of the fourths. For the outer fences, the corresponding percentages are 13.4%, 2.9%, 1.2%, 0.5%, and 0.2%. Thus, in well-behaved data, samples containing far out observations are fairly rare, but samples containing outside observations are fairly common.

Another characteristic of the boxplot is the percentage of observations that are outside (including far out), supplemented by the percentage of observations that are far out—again, in well-behaved data. The percentage of observations beyond the inner fences is 8.6% for n = 5, 2.8% for n = 10, 1.7% for n = 20, 1.15% for n = 50, and 0.95% for n = 100, and 0.70% in a normal population. For the outer fences, the corresponding percentages are 3.3%, 0.36%, 0.074%, 0.011%, 0.002%, and 0.00023%. Thus, even normal distributions produce occasional far out observations.

In published articles, boxplots usually summarize part of a completed analysis. When they identify data values as outside, however, the analysis is generally at an early stage. Conceptually, an outlier comes from a different population or mechanism than the bulk of the data. In considering whether an outside data value may actually be an outlier, one can use a stem-and-leaf display or a dot plot to look closer at the whole batch of data. An isolated outside or, especially, far out data value may have helpful background information.

A batch may consist of distinct clusters; a boxplot may hint at such structure, but it is not designed to reveal it. If the outside data values appear at only one end of the boxplot (often the upper end), a transformation may make the batch more nearly symmetric, suggesting that the presence of outside data values is a consequence of the initial scale of the data.

Further Reading

Emerson, J.D., and Strenio, J. 1983. Boxplots and batch comparison. In *Understanding Robust and Exploratory Data Analysis.* Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds. New York: John Wiley & Sons, 58–96.

Frigge, M., Hoaglin, D.C., and Iglewicz, B. 2014. Some implementations of the boxplot. *The American Statistician* 43:50–4.

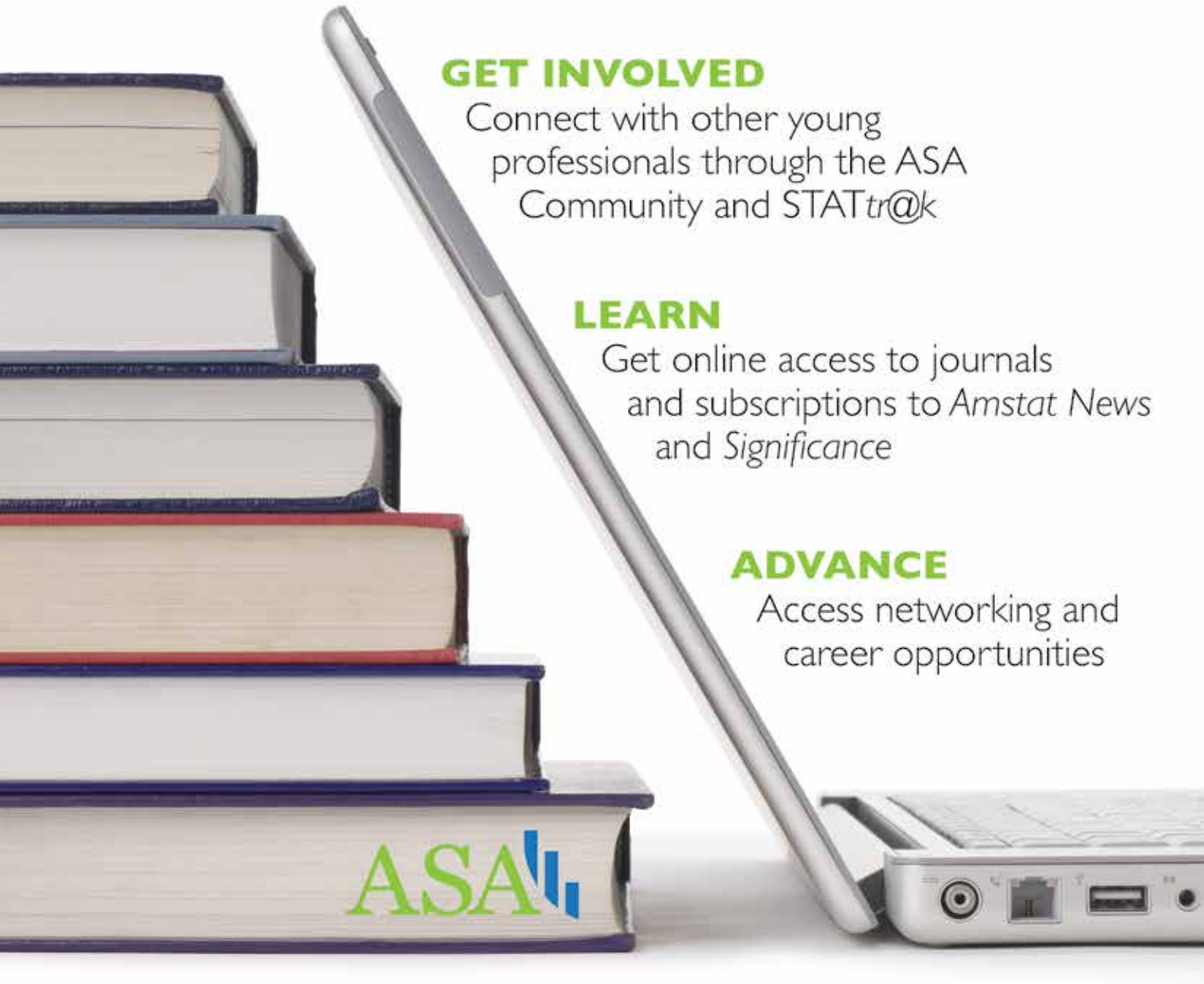
Hoaglin, D.C., Iglewicz, B., and Tukey, J.W. 1986. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association* 81:991–9.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

About the Author

David C. Hoaglin is an independent consultant based in Sudbury, Massachusetts.

STUDENTS: TAKE THE NEXT STEP AND JOIN THE AMERICAN STATISTICAL ASSOCIATION



GET INVOLVED

Connect with other young professionals through the ASA Community and STAT*tr@k*

LEARN

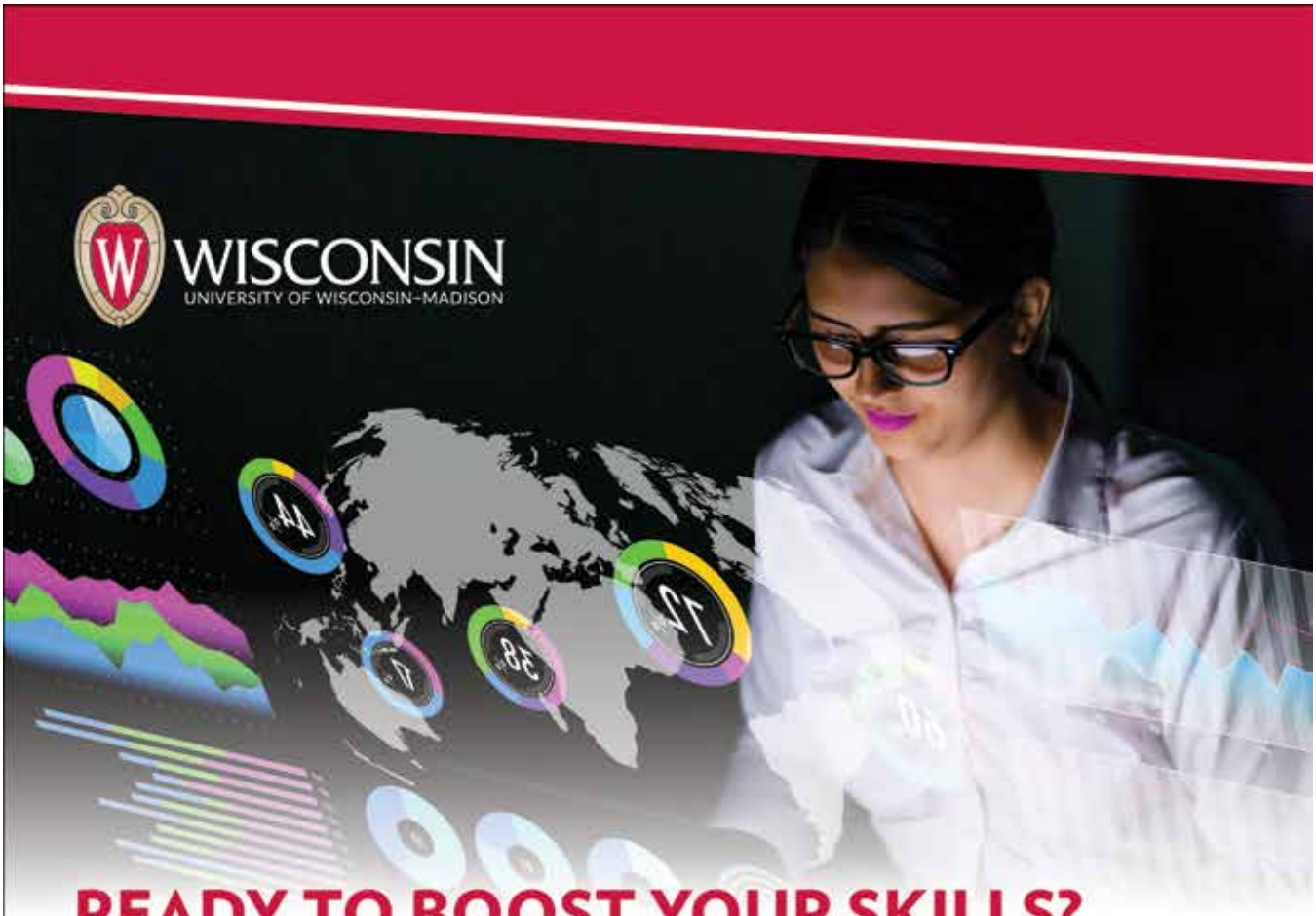
Get online access to journals and subscriptions to *Amstat News* and *Significance*

ADVANCE

Access networking and career opportunities

Student memberships are only $25!

Join today at ***www.amstat.org/join***.



READY TO BOOST YOUR SKILLS?

Transform your data career with a master's degree or certificate from UW–Madison.

Data careers are exciting, important, and lucrative. And the demand for savvy data wranglers continues to grow.

Learn GIS fundamentals
A Capstone Certificate in GIS Fundamentals will give you the core conceptual and applied underpinnings of GIS at your own pace and around your schedule.

Use new skills
You can immediately use new GIS expertise in your current job while adding to your skill tool box. Insurers are using spatial analytics to visualize, investigate, predict, and respond to catastrophe. Your GIS skills can help strengthen the insurance business.

Advance your career
We offer 13 data science and analytics programs with flexible delivery formats that fit the lives of working adults. A degree or certificate from UW–Madison can advance your career.

Visit ***go.wisc.edu/exploreuwdata*** and see how.

17018 3/19

HELP US RECRUIT THE NEXT GENERATION OF STATISTICIANS

The field of statistics is growing fast. Jobs are plentiful, opportunities are exciting, and salaries are high. So what's keeping more kids from entering the field?

Many just don't know about statistics. But the ASA is working to change that, and here's how you can help:

- Send your students to ***www.ThisIsStatistics.org*** and use its resources in your classroom. It's all about the profession of statistics.
- Download a handout for your students about careers in statistics at ***www.ThisIsStatistics.org/educators***.

If you're on social media, connect with us at ***www.Facebook.com/ThisIsStats*** and ***www.Twitter.com/ThisIsStats***. Encourage your students to connect with us, as well.

Site features:

- Videos of young statisticians passionate about their work
- A myth-busting quiz about statistics
- Photos of cool careers in statistics, like a NASA biostatistician and a wildlife statistician
- Colorful graphics displaying salary and job growth data
- A blog about jobs in statistics and data science
- An interactive map of places that employ statisticians in the U.S.