CURV - connecting, uplifting, and recognizing voices

2024-01-31

Quick: think of a statistician.

Did you think of someone? Who came to mind for you? Are you surprised by who you thought of? Was it someone famous? Someone you know? Yourself? Male? Female? Person of color? Differently abled? International?

Representation Matters

When I think about the statisticians I encountered during my formal statistics education, the four individuals in Figure 1, seen on my personal coaster set (received as a gift many years ago), come to mind. Did you think of one of the four statisticians on my coaster set? Why do you think someone thought to make a coaster set including those four individuals?

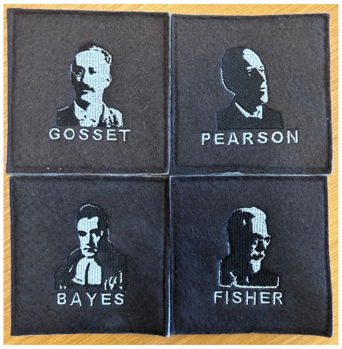
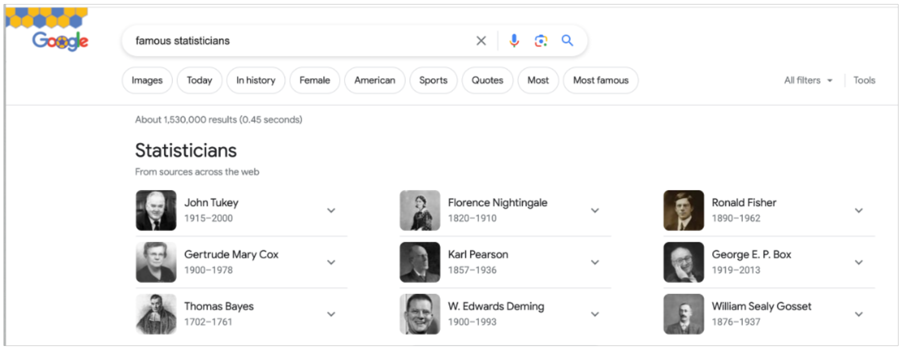


Figure 1: A set of four coasters including Gosset, Pearson, Bayes, and Fisher.

Source: Article Notebook

My coasters aside, a Google search on “famous statisticians” produces a few women on the list, but the demographics of the individuals presented are remarkably homogeneous, see Figure 2. Is your statistician listed in the Google search? Are the statisticians listed by Google somehow more representative of what it means to be a “statistician”? What does it mean to be a representative statistician? Representative of what?



famous statisticians

About 1,530,000 results (0.45 seconds)

Statisticians

From sources across the web

John Tukey 1915–2000	Florence Nightingale 1820–1910	Ronald Fisher 1890–1962
Gertrude Mary Cox 1900–1978	Karl Pearson 1857–1936	George E. P. Box 1919–2013
Thomas Bayes 1702–1761	W. Edwards Deming 1900–1993	William Sealy Gosset 1876–1937

Figure 2: In a Google search for famous statisticians, the individuals listed include Tukey, Nightingale, Fisher, Cox, Pearson, Box, Bayes, Deming, and Gosset.

Source: Article Notebook

Beyond initial thoughts, statistical tchotchkes, and Google searches, it's important to reflect on who else has made important contributions to statistics. Blatantly missing from the Google search is David Blackwell, a giant in the field.

David Blackwell

Arguably, the most famous/influential/brilliant African American statistician is David Blackwell. Statisticians may know Blackwell from the Rao-Blackwell theorem which says that after conditioning on a sufficient statistic, the new estimator will have smaller (or equal to) mean squared error than the original estimator. Indeed, the Rao-Blackwell theorem is seen in most undergraduate Statistical Theory textbooks.

Blackwell was the 1st African American elected to the National Academies of Science and the 1st African American tenured at UC Berkeley. He was the 7th African American to receive a PhD in mathematics. In 2012, President Obama posthumously awarded Blackwell the National Medal of Science.

The majority of Blackwell's career in statistics was spent at UC Berkeley (1954-1988). However, his start at UC Berkeley was postponed due to racism in the Department of Mathematics. The story is that the chair of the department (Erich Lehmann) wanted to hire Blackwell in 1942. However, the wife of another faculty member didn't want any Black members of the department because then she wouldn't feel comfortable hosting department events in her home. So the wife told her husband to tell Lehmann not to hire Blackwell. Which is what happened. You can hear Blackwell describe the situation in his own words at *www.youtube.com/watch?v=Mqpf9tw44Xw*. It is a powerful three minute video; I recommend you watch it and show it to your students.

I recount the Blackwell story to indicate that racism is critical and recent in the history of statistics. Indeed, the 2021 article by Bodmer et al. "The outstanding scientist, R.A. Fisher: his views on eugenics and race" provides a detailed description of the eugenics movement associated with R.A. Fisher and colleagues. But even more salient is notion that in the standard undergraduate curriculum, there are very few people of color who our students see as role models.

Why does representation matter?

When individuals don't feel a part of the community, their identity gets mixed up with feelings about their ability. Figure 3 shows an xkcd comic that encapsulates what can happen when individuals of the non-dominant demographic group engage with content of the course / curriculum / minor / major.

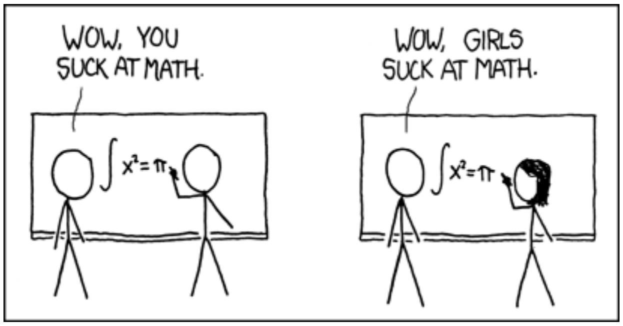


Figure 3: image credit – https://xkcd.com/385/

Source: Article Notebook

There have been numerous studies investigating the role of representation and success across STEM fields. In their 2021 article "Who Is a Scientist? The Relationship Between Counter-Stereotypical Beliefs about Scientists and the STEM Major Intentions of Black and Latinx Male and Female Students", Nguyen and Riegle-Crumb find that counter-stereotypical beliefs are associated with intentions to pursue STEM fields for Black and Latinx students. They say:

> Research indicates that many young people may be deterred from pursuing STEM fields due to prominent stereotypes regarding who best fits and belongs in such fields.

Some of the reasons that representation impacts participation in engaging in STEM include stereotypes about innate abilities and stereotypes about images in the field. The first set of stereotypes (innate abilities) describe beliefs that particular groups are not **capable** of success in STEM fields. The second set of stereotypes (images) describe beliefs that particular groups do not **belong** to STEM fields. Regardless, it seems clear that if students can see themselves in the scholars of the field, they will be more likely to continue pursuing that discipline.

Buckmire et al. (2023) write on what makes a "mathematician." They recognize that narrowly defining who should / can do mathematics is a process by which people are made to feel excluded from mathematics. In their article they present a variety of definitions for "mathematician", and each of the nine co-authors describes how they do or do not meet each of the definitions. The different representations show students a diversity of role models.

Our students' identities span the entire spectrum of backgrounds. Just like us, they are diverse with respect to race, socio-economic background, gender identity, faith, and ability. As educators, the more we can cultivate sense of belonging in our statistics or data science classroom, the more successful we will be at communicating our learning outcomes. In what follows, we present a database of role models with the hopes of introducing statisticians, data scientists, and mathematicians as individuals to whom our students can connect.

CURV - connecting, uplifting, and recognizing voices

To combat the homogeneity of representation in standard statistics content, in most of my courses, I've introduced a weekly activity of *Statistician of the Day*. The scholars who I present are those traditionally underrepresented in statistics and data science. Many of them are people of color, but they also represent individuals who are pushing boundaries to make statistics and data science more accessible and inclusive, often because they themselves have navigated a world which was not accessible or not inclusive.

The database is called **CURV - connecting, uplifting, and recognizing voices** (*hardin47.github.io/CURV)* You can see the source code and/or contribute to the database by exploring the CURV GitHub repo(*github.com/hardin47/CURV)*

Who is in CURV?

The point of the CURV database is to introduce students to scholars to whom students can relate. In that spirit, I have included:

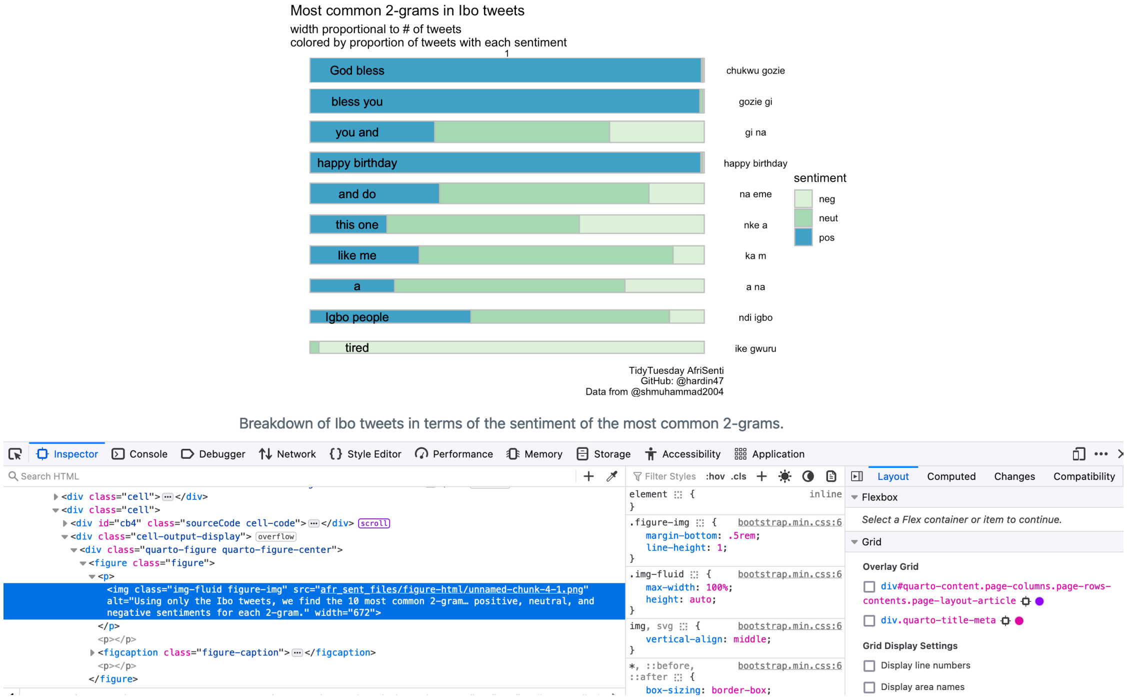
- scholars who represent the diversity of students.
- scholars working to make statistics more accessible.
- scholars using statistics to do good in the world.

In what follows, I present additional four scholars from the database. The four individuals represent a diversity of backgrounds. The hope is that not only do students seem themselves in the scholars, but that they also recognize the value of having different perspectives in problem solving.

Liz Hare

Liz Hare is not a statistician; she is a geneticist, working primarily in dog / animal genetics. However, as someone who is very active in the Minorities in R (MiR) Community, she works regularly with statisticians.

Dr. Hare is visually impaired and has focused her work on communicating the value and ease with which statisticians and data scientists can add alt text to their reports. Alt text is the words describing an image that a screen reader reads out loud as part of reading through a document. In Figure 4 you can see that the html code encapsulates the alt text describing a TidyTuesday figure on African language sentiment.



Breakdown of Ibo tweets in terms of the sentiment of the most common 2-grams.

```
<div class="cell">⋯</div>
<div class="cell">
  <div id="cb4" class="sourceCode cell-code">⋯</div> scroll
  <div class="cell-output-display"> overflow
    <div class="quarto-figure quarto-figure-center">
      <figure class="figure">
        <p>
          <img class="img-fluid figure-img" src="afr_sent_files/figure-html/unnamed-chunk-4-1.png"
          alt="Using only the Ibo tweets, we find the 10 most common 2-gram… positive, neutral, and
          negative sentiments for each 2-gram." width="672">
        </p>
        <p></p>
        <figcaption class="figure-caption">⋯</figcaption>
        <p></p>
      </figure>
```

```
element ⋮⋮ {                     inline
}
.figure-img ⋮⋮ {     bootstrap.min.css:6
  margin-bottom: .5rem;
  line-height: 1;
}
.img-fluid ⋮⋮ {      bootstrap.min.css:6
  max-width: 100%;
  height: auto;
}
img, svg ⋮⋮ {        bootstrap.min.css:6
  vertical-align: middle;
}
*, ::before,         bootstrap.min.css:6
::after ⋮⋮ {
  box-sizing: border-box;
```

Flexbox

Select a Flex container or item to continue.

Grid

Overlay Grid

☐ div#quarto-content.page-columns.page-rows-contents.page-layout-article

☐ div.quarto-title-meta

Grid Display Settings

☐ Display line numbers

☐ Display area names

Figure 4: Screen shot of a TidyTuesday analysis on African language sentiment with the alt text displayed.

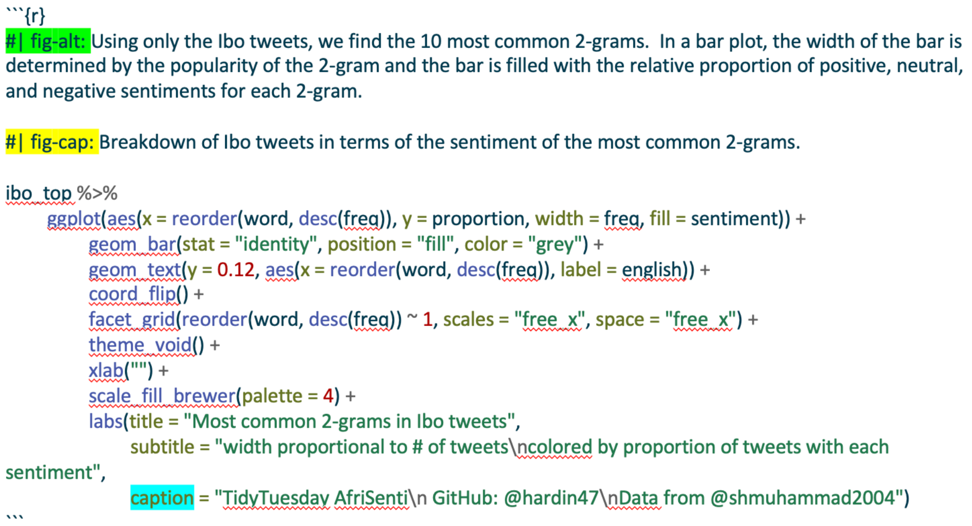
Source: Article Notebook

When using alt text, Dr. Hare asks us to consider and report:

1. What kind of graph or chart is it?
2. What variables are on the axes?
3. What are the ranges of the variables?
4. What does the appearance tell you about the relationships between the variables?

After presenting Dr. Hare and her work to my students, I am able to teach them how to include alt text in their own work, a process which is extremely straightforward if students are using R markdown or Quarto documents.

In R, including alt text is done by providing information for the relevant R chunk. After introducing students to Dr. Hare’s work, it takes very little overhead to communicate to them the ways that a graphic can be made more accessible by adding alt text to each figure. Figure 5 specifies the R code (in a Quarto document) for adding different figure descriptions. Note the difference between the (1) alt text, (2) document figure caption, and (3) ggplot figure caption. Figure 6 specifies the connection between the figure descriptions and the image itself.



~~~
```{r}
#| fig-alt: Using only the Ibo tweets, we find the 10 most common 2-grams.  In a bar plot, the width of the bar is
determined by the popularity of the 2-gram and the bar is filled with the relative proportion of positive, neutral,
and negative sentiments for each 2-gram.

#| fig-cap: Breakdown of Ibo tweets in terms of the sentiment of the most common 2-grams.

ibo_top %>%
      ggplot(aes(x = reorder(word, desc(freq)), y = proportion, width = freq, fill = sentiment)) +
            geom_bar(stat = "identity", position = "fill", color = "grey") +
            geom_text(y = 0.12, aes(x = reorder(word, desc(freq)), label = english)) +
            coord_flip() +
            facet_grid(reorder(word, desc(freq)) ~ 1, scales = "free_x", space = "free_x") +
            theme_void() +
            xlab("") +
            scale_fill_brewer(palette = 4) +
            labs(title = "Most common 2-grams in Ibo tweets",
                  subtitle = "width proportional to # of tweets\ncolored by proportion of tweets with each
sentiment",
                  caption = "TidyTuesday AfriSenti\n GitHub: @hardin47\nData from @shmuhammad2004")
```
~~~

Figure 5: R code for Ibo tweets in TidyTuesday analysis.

Source: Article Notebook

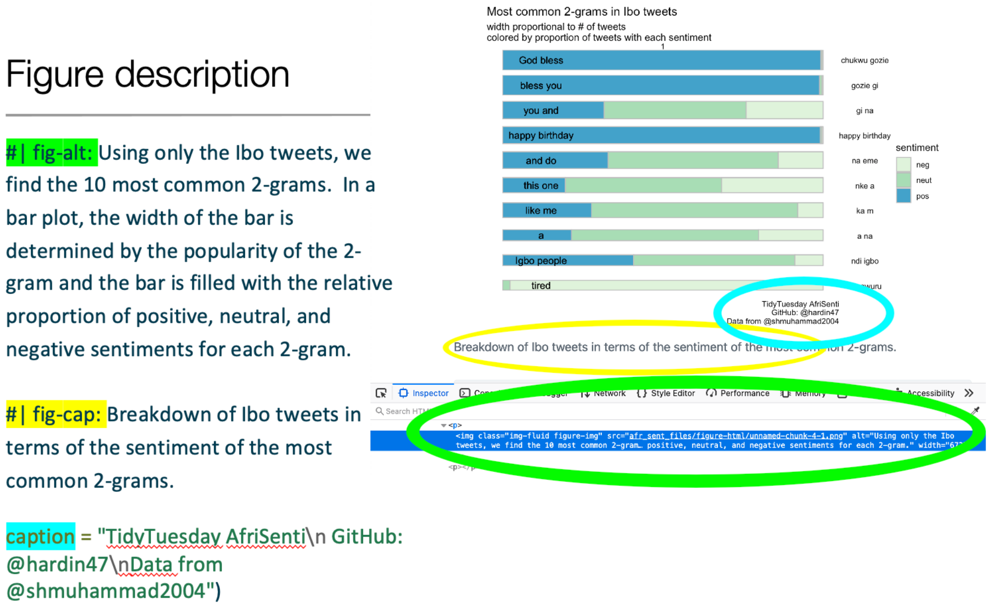


Figure description

#| fig-alt: Using only the Ibo tweets, we find the 10 most common 2-grams. In a bar plot, the width of the bar is determined by the popularity of the 2-gram and the bar is filled with the relative proportion of positive, neutral, and negative sentiments for each 2-gram.

#| fig-cap: Breakdown of Ibo tweets in terms of the sentiment of the most common 2-grams.

caption = "TidyTuesday AfriSenti\n GitHub: @hardin47\nData from @shmuhammad2004")

Breakdown of Ibo tweets in terms of the sentiment of the most common 2-grams.

```
<p>
  <img class="img-fluid figure-img" src="afr_sent_files/figure-html/unnamed-chunk-4-1.png" alt="Using only the Ibo
  tweets, we find the 10 most common 2-gram… positive, neutral, and negative sentiments for each 2-gram." width="67
<p></p
```

Figure 6: Different ways to annotate a figure include alt text, figure captions for the full file, and figure captions for the ggplot.

Source: Article Notebook

By communicating ways in which data analyses can be made more accessible to all users, we can set our students up to recognize that statistics and data science is for everyone. Students may connect to the scholars because of their own different abilities. Or they may want to work on making software more accessible. Regardless, a conversation about using alt text is always good in that the students will become better purveyors of data science.

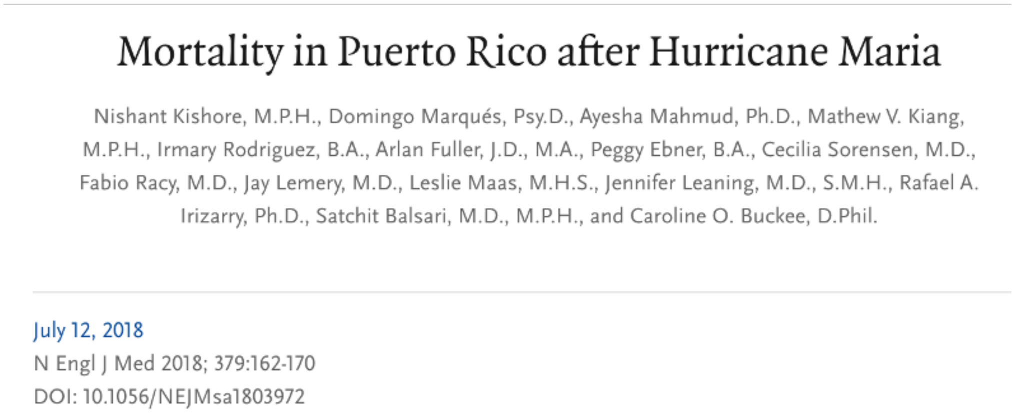
Rafael Irizarry

Rafael Irizarry is a well known biostatistician; he completed his PhD at Berkeley, worked for many years at Johns Hopkins University, and is currently running a lab at Harvard as Professor of Biostatistics and at the Dana-Farber Cancer Institute as Professor of Biostatistics and Computational Biology. He has dozens of online courses through the edX platform and over a hundred publications via Google scholar.

Relevant to the CURV database however is the work that Rafael Irizarry has done in Puerto Rico. Having graduated from the University of Puerto Rico, Professor Irizarry had a vested interest in the community that was ravaged in 2017 when Hurricane Maria, a category 5 hurricane, devastated the island. With collaborators, Professor Irizarry performed a representative stratified sample to measure neighborhoods based on how easily accessible they were in the aftermath of the hurricane.

The original news reports, months after the hurricane, were that the official death report from Hurricane Maria was 64 people. Professor Irizarry and colleagues estimated that the number of excess deaths was 4645, with a 95% confidence interval of 793 to 8498.

Their work provides a myriad of topics to unpack in a statistics classroom. Some of the discussions I have had with my students include: who is doing the work to understand climate change at a global level; how is stratified sampling different from simple random samples and why can't we always take simple random samples; and why is the CI so wide? Figure 7 provides the reference for the mortality study; Figure 8 provides the results from the abstract of the paper. I like to show students the literature so that they connect the actual publication to the research like Professor Irizarry's which can have a big impact on public policy and human lives.



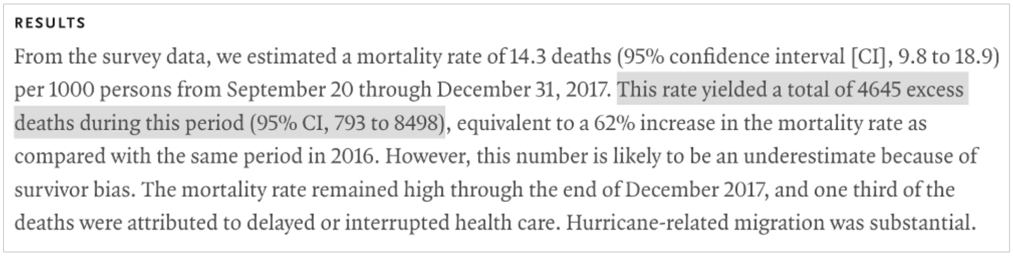
Mortality in Puerto Rico after Hurricane Maria

Nishant Kishore, M.P.H., Domingo Marqués, Psy.D., Ayesha Mahmud, Ph.D., Mathew V. Kiang, M.P.H., Irmary Rodriguez, B.A., Arlan Fuller, J.D., M.A., Peggy Ebner, B.A., Cecilia Sorensen, M.D., Fabio Racy, M.D., Jay Lemery, M.D., Leslie Maas, M.H.S., Jennifer Leaning, M.D., S.M.H., Rafael A. Irizarry, Ph.D., Satchit Balsari, M.D., M.P.H., and Caroline O. Buckee, D.Phil.

July 12, 2018
N Engl J Med 2018; 379:162-170
DOI: 10.1056/NEJMsa1803972

Figure 7

Source: Article Notebook



RESULTS

From the survey data, we estimated a mortality rate of 14.3 deaths (95% confidence interval [CI], 9.8 to 18.9) per 1000 persons from September 20 through December 31, 2017. This rate yielded a total of 4645 excess deaths during this period (95% CI, 793 to 8498), equivalent to a 62% increase in the mortality rate as compared with the same period in 2016. However, this number is likely to be an underestimate because of survivor bias. The mortality rate remained high through the end of December 2017, and one third of the deaths were attributed to delayed or interrupted health care. Hurricane-related migration was substantial.

Figure 8

Source: Article Notebook

Desi Small-Rodriguez

Dr. Small-Rodriguez is a social demographer and an Assistant Professor of Sociology and American Indian Studies at UCLA. She received a PhD in Sociology from the University of Arizona and a PhD in Demography from the University of Waikato. Dr. Small-Rodriguez is Northern Cheyenne and Chicana and grounds her work in Indigenous studies, sociology of race and ethnicity, critical demography, and health policy research. She directs the Data Warriors Lab (a mobile data sovereignty lab serving Indigenous communities) and was previously a member of the Collaboratory for Indigenous Data Governance. She is a founding member of the Global Indigenous Data Alliance.

Dr. Small-Rodriguez is passionate about Indigenous data sovereignty and Indigenous data governance. Using networks of Indigenous scholars and survey methods, she works toward the following two goals: (1) better collection and use of data on Indigenous people that has been gathered by external sources such as the census and other federal entities; (2) development of data methods and practitioners within the Indigenous community. Dr. Small-Rodriguez also works for health and economic justice on Indian Reservations.

When I introduce Dr. Small-Rodriguez to my students, I also introduce ideas of data sovereignty. I point them to the Native Nations Institute at The University of Arizona where scholars have put together policy briefs and calls to action describing how data can be used thoughtfully and respectfully.

Lester Mackey

Dr. Mackey is a machine learning researcher at Microsoft Research New England and an adjunct professor at Stanford University. He has a PhD in Computer Science and an MS in Statistics, both from UC Berkeley.

In 2006 Netflix offered $1 million to the researchers who could come up with the most accurate prediction of people's movie ratings. As undergraduates, Dr. Mackey and two friends led the competition for a few hours in its first year. Later, Dr. Mackey's group merged with a few others, forming The Ensemble. Their final analysis came in second with the exact same error rates as the winning entry. The winning entry, however, had been submitted 20 minutes prior. Sigh.

I love telling my students Dr. Mackey's Netflix story, as it gets them thinking about how *they* can do statistics and data science, even as undergraduates. We talk about Kaggle and other data competitions and also about the ways that problem solving (i.e., real data science) is so much more than models applied to data that has already been cleaned.

Dr. Mackey is involved in Stanford's initiative of Statistics for Social Good and has the following quote on his website:

> Quixotic though it may sound, I hope to use computer science and statistics to change the world for the better.

In 2023, Dr. Mackey was awarded a MacArthur Genius Grant. What an awesome person for our students to know about.

CURV in the classroom

While you surely have creative ideas on how to use CURV in the classroom, I have mostly used it as a *Statistician of the Day* project. Once a week (or sometimes at every class period), I briefly introduce a scholar and present different ways they have contributed to the field. Typically, I use only about five minutes of class time. Some scholars illicit discussion, but for most of them, I simply present their background and work, and we move on to the course material of the day.

Despite the short about of time spent on introducing the scholars, my students have been extremely positive about the engagement. In summarized student reflections, they report that the presentations of the scholars:

- is inspiring.
- is a great way to better understand the career paths of data scientists.
- is a venue for having important conversations with peers.
- highlighted diversity in the field of statistics.
- made the classroom environment feel more inclusive.

Contributing to CURV

In addition to presenting scholars to your students, you can engage the students in the conversation by having them **find** scholars to highlight. If you or your students have a name or an idea of someone aligned with the CURV goals, you can submit an issue on the CURV GitHub repo. If you would like to create an entire entry about a particular individual, please consider creating a pull request. (Creating a pull request allows students to practice data science skills while also engaging in issues of justice, equity, diversity, and inclusivity.)

CURV aligned

I developed the CURV database because I didn't know of a different resource that I could use for my *Statistician of the Day* classroom exercise. I wanted the information provided to be relevant to my students, connecting either through identity or through class content. That said, I've borrowed heavily from existing resources, and I appreciate the myriad efforts which engage students in many different ways. On the CURV website I list additional fantastic resources highlighting statisticians, data scientists, and mathematicians who are traditionally underrepresented. I encourage you and your students to check them out!

Further Reading

- Bodmer, W., Bailey, R.A., Charlesworth, B. et al. "The outstanding scientist, R.A. Fisher: his views on eugenics and race." *Heredity* 126, 565–576, 2021.

- Kennedy-Shaffer, Lee. "Teaching the Difficult Past of Statistics to Improve the Future." *Journal of Statistics and Data Science Education* 32.1, 108-119, 2024.

- Nguyen, Ursula, and Catherine Riegle-Crumb. "Who Is a Scientist? The Relationship Between Counter-Stereotypical Beliefs about Scientists and the STEM Major Intentions of Black and Latinx Male and Female Students." *International Journal of STEM Education* 8, 2021.