Mind the Gap: An incomplete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

June 25, 2023
Maleny, Australia

A complete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

June 25, 2023
Maleny, Australia

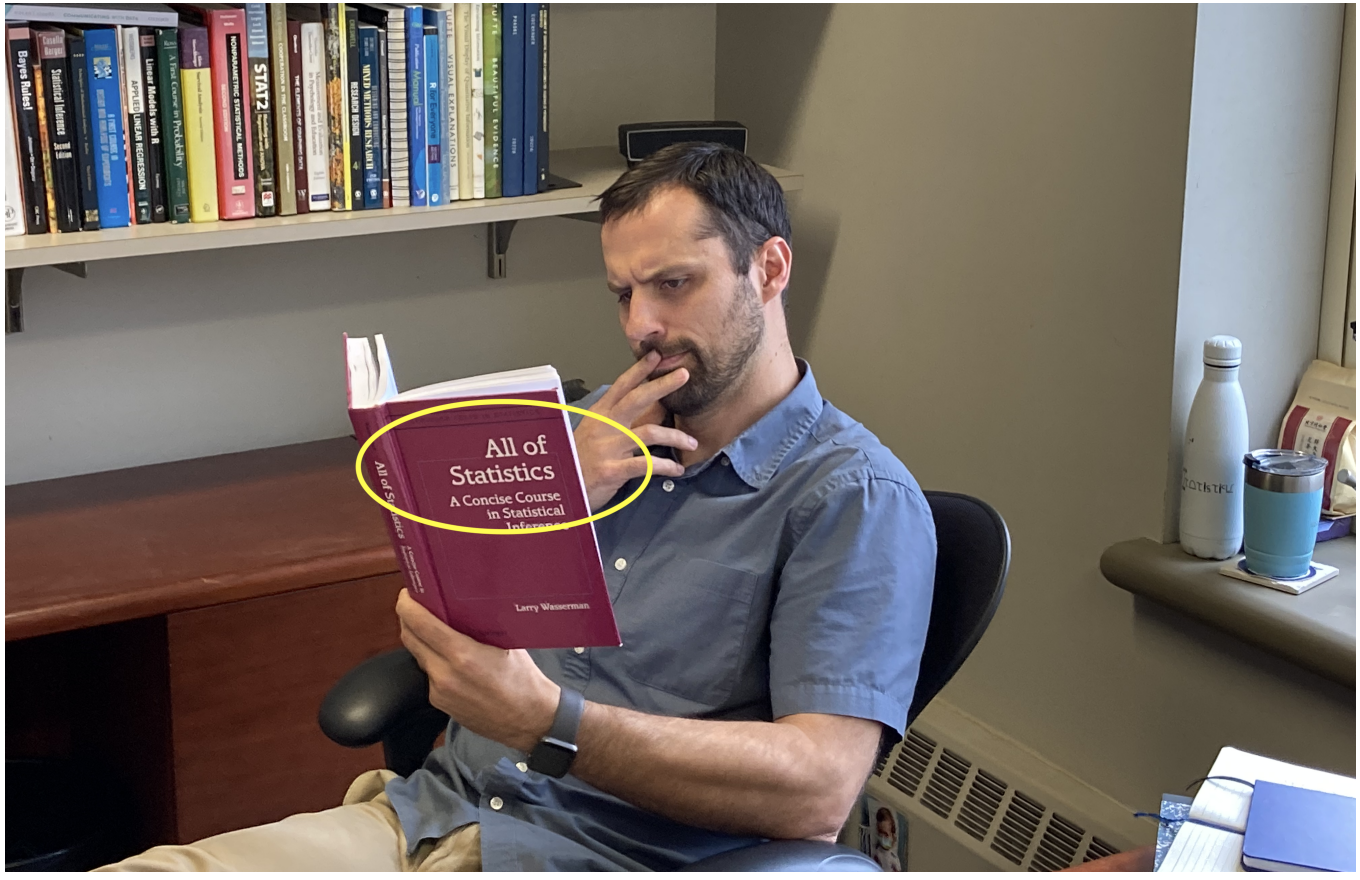


Figure 1: Getting ready to "Provide the perspective of the discipline..." as promised in the SRTL announcement.

- ...on "reconceptualising data and data-ing"
- "data-ing?"
- that wasn't in the book...

A complete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

June 25, 2023
Maleny, Australia

Mind the Gap: An incomplete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

"Reconceptualizing data & data-ing"

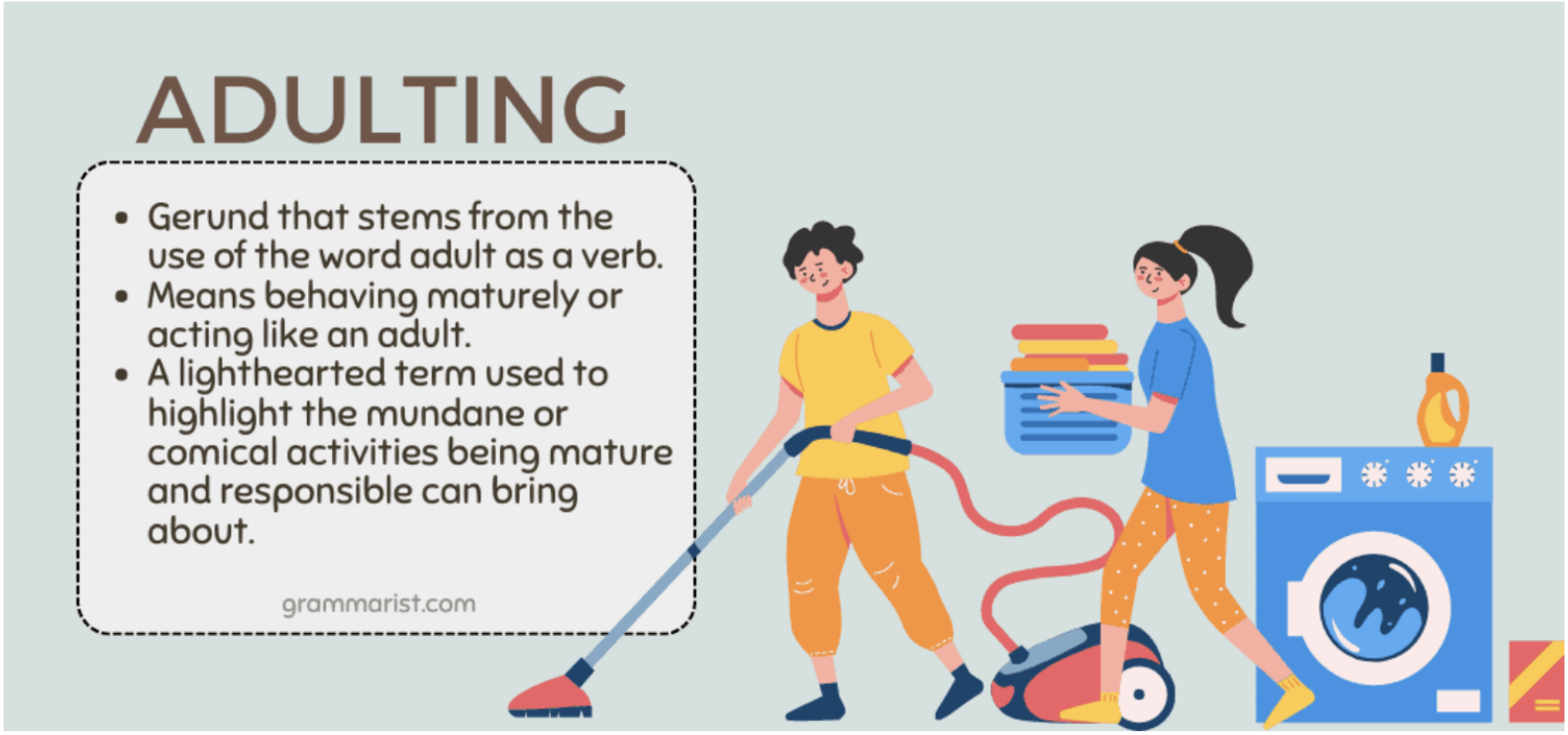
June 25, 2023
Maleny, Australia

“Data. . . ing?”

- usually not one to “verb” my nouns (then use as a gerund)
- [*checks list of participants for a “grammarian-in-residence”*]
- but I’ve heard this kind of thing happens on Twitter

ADULTING

What Is the Meaning of Adulting?



Adulting is a fairly new gerund that stems from the use of the word adult as a verb. A gerund is a verb form ending in -ing that acts as a noun. *Adulting* simply means behaving maturely or acting like an adult.

Figure 2: image credit: Grammarist. URL: https://grammarist.com/new-words/adulting/

But “Data” as a verb?

- I’ve heard of *data verbs*, just not “data” *as* a verb
- Amelia McNamara helpfully pointed out (to my surprise) this too has precedent!

Jer Thorp

Roman Makhmutov

Jer Thorp is an artist, a writer, and a teacher. He was the first data artist in residence at *The New York Times*, is a *National Geographic* Explorer, and served as the innovator in residence at the Library of Congress in 2017 and 2018. He lives under the Manhattan Bridge with his family and his awesome dog, Trapper John, MD. *Living in Data* is his first book. You can sign up for email updates here.

Figure 3: image credit: Roman Makhmutov[1]

[1]Thorp, J. (2021). Living in data: A citizen's guide to a better information future. New York: MCD, Farrar, Straus and Giroux.

Coming around. . .

"the pair data and data-ing refers to a similar conceptualization of the relation between sample and sampling, or model and modeling, where the first is the statistical concept and the second refers to the process of engaging or reasoning with this concept."

What perspective can I offer?

The birth of an academic

Figure 4: (Academic) offspring of Joan Garfield & Bob delMas enters the world (of Statistics Education)

- if any of you know me, it's probably thanks to my ~~parents~~ PhD advisors Joan Garfield & Bob delMas
- there's plenty more to my upbringing before I ever got involved with Statistics Education, including lots that has almost nothing to do with academia!

Ecolab, Inc.

Figure 10: Image credit: Patrick Kennedy, Star Tribune[2]

- Started out at Ecolab working in R&D (+ Eng)
- Interned with a small team of staff statistical consultants
- Assist teaching stat in-house (Intro, DOE, MSA, SPC, RDSA)
- Typical data for data-ing: Lots of design & analysis of experiments
- "Best way to stop yourself being *fooled* by the data is to *look* at it" –Paul Prew (Matt's ECL Mentor)

[2]Kennedy, P., (6 Feb 2023). Ecolab now selling products at Home Depot — the first time available at retail stores. Star Tribune

Medtronic, PLC

Figure 11: Medtronic Headquarters. image credit: https://asiapac.medtronic.com/xp-en/about.html

- World's largest medical technology company
- Hired due to commitment to government regulators!
- Data for Data-ing: verbatim complaints from call center, manufacturing lines, *some* clinical, sales & registration data, engineering diagnostics from in-house returned product analysis, lots more!
- Q: "What happens when a competent statistician is released into a sea of engineers?"

Medtronic, PLC

- A: "YOU GET *BUSY*!!" ~Tom Keenan (Matt's MDT mentor)
- Broader involvement including Quality, Manufacturing, R&D, Engineering, Marketing, HR/Personnel, Six Sigma

- "The company feels the mean, the customers feel the tails." –TK
- "Our job is to prosecute the war on variation!" –TK
- Regularly tasked with discussing/explaining statistical methodology and procedures to government regulators (e.g., post-market surveillance)
- Develop & teach in-house statistics courses (again!)

Nonin Medical, Inc

Figure 12: Pulse Oximeter. image credit: Michael Heisson[3]

- Senior Biostatistican ("only" statistician. . .)
- Internal & external collaborations (e.g., physician researchers, clinical trial design, etc)
- Data for Data-ing: all of it! Primarily clinical trails.

[3]Lee, E., (4 Oct 2022). The Best Pulse Oximeter for Home Use. New York Times Wirecutter. URL: https://www.nytimes.com/wirecutter/reviews/best-pulse-oximeter-for-home-use/

Back to Medtronic!?

- Back to Medtronic for a lame duck session
- Goals: clean up special projects, train new statisticians (& business analysts)...
- and *help automate my old job away*!
- Next Stop: Penn State!

Figure 13: image credit: Penn State Office of Physical Plant

Sidebar

- When I was a few years younger, people gave me puzzled looks when I described this background. . .
- Ever experience that look when someone seems to be calculating your age while you talk with them?
- When did you say you graduated?
- How old did you say your kids are?
- . . . and you worked there how many years?
- [*hmm. . . something isn't adding up*]
- Life hack: I earned my PhD *while* working full-time (and child-rearing)

Personal reflections from industry

- I often reflect on **gaps** between expectations/assumptions and the reality of my contributions on the job...
- *Expectation*: UMN Statistics Dept filled my toolbox with advanced methods & fancy models... that's what they'll expect me to do at work.
- *Mind the gap*: At work, I generally used 10-20% of the fancy things I learned in those courses
- Lots of the fancy methods that I needed at work, I *learned* at work. This is a common refrain among professional statisticians.

Personal reflections from industry

- *Mind the gap*: I noticed statisticians were regularly requested to engage with myriad issues that were **not at all statistical** in nature.

- **Bad** reason this gap might emerge: management may not understand how statisticians best contribute to the organization[4]

- **Good** reason this gap might emerge: (coming up later. . .)

[4] Deming, W. E. (2000). Out of the Crisis. MIT Press.

Mind the Gap: An incomplete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

June 25, 2023
Maleny, Australia

Mind the gap

Figure 14: image credit: https://www.stuff.co.nz/travel/78997789/mind-the-gap-voice-of-london-tube-dies

Mind the gap

- Role of statisticians at work
- Public perception of Statistics
- Student perception of EDA
- Statistics education research

Mind the gap: In the public

- *statistics vs Statistics*
- I meet students all the time that tell me they became interested in studying statistics by watching sports. . .
- a "statistic" could be regarded as a collective property of some data (e.g., a quantity calculated from a sample)
- Among the greatest contributions the field of Statistics offers the world is the means to appropriately accommodate, characterize, and even quantify **variability** and **uncertainty**

Mind the gap: statisticians at work

As a statistical consultant, the first project with a first-time client is usually a disaster. But the next one gets much better.

(Paraphrase of Sandy Weisberg)

Mind the gap: EDA[5]

- Perceived value vs potential contribution of EDA
- even (maybe especially) among a group I'll call "advanced novices"
- temptation to rush into statistical modelling and skip ahead to "get the answers"
- EDA = summary statistics & cursory plots (to appease Professor)

[5] Exploratory Data Analysis

Mind the gap: EDA

- My solution?
- maybe my students just need a framework for a careful EDA
- with an acronym to help remember it!

EDA Framework (1st attempt)

Get in "B-E-D" with your data

- **B**ecome acquainted with the data
- **E**xplore intuition for your research question(s)
- **D**iscover features in the data that impact modeling decisions

EDA Framework (2nd attempt)

Let's just reuse "E-D-A" instead!

- **Examine the data source(s)**: data provenance, variable types, coding, missingness, summary statistics/plots;
- **Discover features that influence may modeling decisions:** investigate potential outliers, consideration for recoding variables (e.g., numeric data that's functionally dichotomous), threats to modeling assumptions (*especially* independence);
- **Address research questions:** build intuition and note preliminary observations/conclusions related to each research question. Also, note observations that prompt you to refine your research questions or add new questions to investigate
- p.s. Lonneke's work included a concise summary of literature that I think improves upon the "E" step!

Mind the gap: Opportunity for research?

Mind the gap: Statistical Thinking

International Statistical Review (1999), **67**, 3, 223–265, Printed in Mexico

Statistical Thinking in Empirical Enquiry

C.J. Wild and M. Pfannkuch

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

This paper had its genesis in a clash of cultures. Chris Wild is a statistician. Like many other statisticians, he has made impassioned pleas for a wider view of statistics in which students learn "to think statistically" (Wild, 1994). Maxine Pfannkuch is a mathematics educator whose primary research interests are now in statistics education. Conception occurred when Maxine asked "What *is* statistical thinking?" It is not a question a statistician would ask. Statistical thinking is the touchstone at the core of the statistician's art. But, after a few vague generalities, Chris was reduced to stuttering.

Figure 17: Opening vignette from one of my favorite papers.[6]

[6]Wild, C. J., Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), pp 223-265.

Thinking Statistically

- Recall: *Mind the gap*: Statisticians like me were regularly requested to engage with myriad issues at Medtronic that were **not at all statistical** in nature. . .

- *Bad* reason: management unclear what contribution statisticians offer to the organization

- **Good reason**: perhaps my colleagues recognized that we don't just "do statistics" we're experts at *thinking statistically*!
- I argue that *statistical thinking transfers*
- Disciplined approach to problem solving & critical thinking
- Due consideration for uncertainty, alternative explanations, and practical implications

Mind the gap: Data & Data-ing

"the pair data and data-ing refers to a similar conceptualization of the relation between sample and sampling, or model and modeling, where the first is the statistical concept and the second refers to the process of engaging or reasoning with this concept."

- So, why verb the noun??
- Maybe because a verb for the action we're describing *doesn't exist*!

What *IS* data-ing?

- Something new to be explored?
- Something familiar by another name?
- An intersection?
- A superset?

What actions might data-ing include?

- data collection (Yannik & Susanne)
- variable creation/recognition (Amelia & Sibel)
- data cleaning (Many SRTL-ers)
- modeling and interpretation of data (Lucia)

More engaging and reasoning with data

- Andee & Michal probe evaluation of which data **needed** to achieve the scientific purposes?

- Proxy variables when we encounter a gap in the available (or accessible) data to achieve the scientific purpose of our analysis–really important part of "data-ing" as an applied statistician, but these are motivated by the scientific domain

- Carl & Kym evoke notions about data and empowering students to uncover rich (multivariate) stories
- Alyssa seeks to examine the interface between computational thinking and data-ing
- What if we favor an algorithmic rather than inferential "culture"?[7]
- Ronit challenges us to consider "big data-ing"

[7] Breiman, L. (2001). Statistical modeling: The two cultures. Statistical science, 16(3), 199-231.

How well-defined is "data"

- Is "messy" data well-defined?
  - any deviation from tidy data[8]?
  - Amelia & Kym discuss intuition of data cards
- What is "big data"??[9]
  - Volume? Velocity? Variety?[10]
  - Does "messier" make it "bigger"
  - Is the distinction absolute or relative?
  - how will this make "big data-ing" different?
- Jill & Lonneke discussed consuming and evaluating evidence—and implications of data-ing when engaged with forms of evidence more broadly concieved than has been typical for classical data analysis

[8] Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10). DOI: 10.18637/jss.v059.i10

[9] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1).

[10] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. In: Meta Group.

my two cents?

Reconceptualizing data...

- I'm grateful that the popularity of data science has lead to a wider view of data
- I suspect use of Statistics has long been tempered by a focus on just the data we know how to handle
- I'm open-minded, but I have developed strong opinions about how we qualify "big" data, and it's a relativistic rather than absolute interpretation
- I do wonder if there's similarly more to "messy" than the absolute distinction of whether or not the data are "tidy" (but perhaps not)

my two cents?

. . . and data-ing

- Was my response to the **EDA** gap an attempt to fill a **data-ing** gap?
- Intuition first, formalize with statistical models
- Is "data-ing" necessarily a *statistical* issue?
- Top of mind: some ways we engage & reason with data are not as closely linked to our statistical objectives
- Data architecture? security? privacy? reproducibility?
- Will we convene as SRTL-13 and adjourn as **D**RTL-1??

- I'm struck by the parallel to the vignette about Statistical thinking. . .
- The idea of data-ing seems to have been right in front of us all along,
- it's something we value,
- and want our students to value,
- and yet we don't yet know where it begins and ends, or even what to call it!!
- What might we learn as we refine our view of data-ing together?
- In what ways might we reconceptualize data?

"How does one speak about something that is both fish and water, means as well as end?" –Ursula Franklin[11]

- epigraph of Jer Thorp's "Living in Data"...
- and perhaps SRTL-13 / DRTL-1 as well??

[11] Franklin, U. (1990). The Real World of Technology. CBC Enterprises: Toronto

References

- Batty, M. (2015). Data about cities: Redefining big, recasting small. Paper prepared for the Data and the City workshop, Maynooth University, 31 August–1 September 2015. URL: http://www.spatialcomplexity.info/files/2015/08/Data-Cities-Maynooth-Paper-BATTY.pdf
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science, 16*(3), 199-231.
- Deming, W. E. (2000). *Out of the Crisis*. MIT Press.
- Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1).
- Laney, D. (2001) 3D data management: Controlling data volume, velocity and variety. In: Meta Group. URL: https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...
- Tukey, J. (1977). *Exploratory Data Analysis*. Vol 2.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10). DOI: 10.18637/jss.v059.i10
- Wild, C. J., Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), pp 223-265.

Mind the Gap: An incomplete picture of statistics, statisticians, & statistics education

Matthew Beckman
Penn State University

June 25, 2023
Maleny, Australia