Lessons from the Toronto Data Workshop

Rohan Alexander, University of Toronto

Joint Statistical Meetings, Toronto, 8 August 2023

Outline

1. The what and how of reproducibility (in data science)?
2. Rigour (in data science)?
3. Data science and its constituent parts?
4. What are, and how do we teach, the foundational skills of data science?
5. What does the relationship between industry and academia look like?

department

To my ~~family~~. Please don't read this.

1. What is reproducibility (in data science) and how do we teach it?

"Increasingly, **we consider a paper to be an advertisement,** and for the associated **code, data, and environment to be the actual work**"

Buckheit and Donoho (1995)

What does reproducibility look like (in data science)?

Literate programming and version control

Literate programming:

- **Mine Çetinkaya-Rundel (2021)**, “In the beginning was R Markdown“
- **Mine Çetinkaya-Rundel (2022)**, “Reproducible authoring with Quarto”
- **Debbie Yuster**, “Reproducible Student Reports with Python + Quarto”

Version control:

- **Maria Tackett**, “Knit, Commit, and Push: Teaching version control in undergraduate statistics courses”
- **Colin Rundel**, “Teaching Statistical computing with Git and GitHub”

Let’s grade ourselves

Code: 🥇

- Quarto, Python/R, Git/GitHub, R Studio/VS Code

Data: 🥈 or 🥉

- Data sharing is (more) common in research practice, but not taught (much).
- Parquet is standard in industry, but not taught (much).

Environment: 🗑️🔥

- “The package universe feels a lot more brittle than it did a year ago because I’m trying to use renv and lock files”
- “Docker is a nightmare”

2. What is rigour (in data science)?

We know what rigor looks like in statistical theory: theorems are accompanied by proofs... innovation is needed to ensure that the same level of rigor characterizes claims based on data and code.

Horton et al. 2022

End-to-end workflow

- **Emily Giambalvo and Ence Morse**, The Washington Post, “How the NFL blocks Black coaches”
- **Meggie Debnath** and **Maitreyee Sidhaye**, St. Michael’s Hospital, Unity Health Toronto, “The Things We Learned from Deploying AI in Healthcare”
- **Jacob Matson**, Simetric, Inc., “From data to dashboard”
- **Ijeamaka Anyene**, Kaiser Permanente, “Taking the next step past standard charts”

3. Data science ↔ constituent parts?

Data science ↔ constituent parts?

- **Karen Chapple**, “Data science + geography”
- **Fedor Dokshin**, “Data science + sociology”
- **Tegan Maharaj**, “Data science + information”
- **Josh Speagle**, “Data science + astronomy”
- **Yun William Yu**, “Data science + math”
- **Radu Craiu**, “Data science + statistics"
- **Kieran Campbell**, “Data science + biomedicine”
- **Leanne Trimble**, “Data science + libraries”
- **Nathan Taback**, “Teaching data science”

Data science ↔ constituent parts?

Two examples

Core Research

What is our “Hilbert list”?

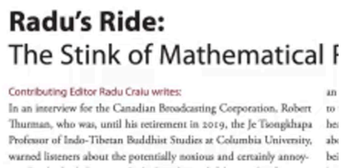
Who are we writing for?

How to deal with “more of the same” issue?

How to/Should we measure impact?

Methodology: new trade-offs

The Core Statistician’s Dilemma: Adaptation vs Amalgamation?


Radu’s Ride:
The Stink of Mathematical F

Contributing Editor Radu Craiu writes:

In an interview for the Canadian Broadcasting Corporation, Robert Thurman, who was, until his retirement in 2019, the Je Tsongkhapa Professor of Indo-Tibetan Buddhist Studies at Columbia University, warned listeners about the potentially noxious and certainly annoy-

Radu Craiu

Questions (*I think about when I think about teaching data science*)

- Should learning objectives/goals include both backend/frontend data science? (e.g., is data literacy enough for students in some programs?)
- What is a good balance of “data science stack” versus “science” to teach?
- How closely should academic data science mirror industrial data science?
- What are the ethical and social implications for training students to work within surveillance capitalism?
- What should students in social sciences, sciences, humanities learn about data science?
- How do we help faculty develop data science skills?
- What roles should statisticians, and computer scientists play in this education?

Nathan Taback


Data science
Statistics
Engineering
Math
Computer science
Economics

4. What are the foundational skills (and how do we teach them)?

What are the foundational skills of data science?

(And how do we teach them?)

- There is emerging agreement on the foundational data science skills:
  - computational thinking
  - sampling design/framework
  - statistical modeling
  - graphs (some interactive)
  - some Git and GitHub
  - APIs and SQL
  - command line basics
  - cleaning data
  - Python, (some) R, (& Julia?)
  - ethics
  - writing
  - workflows
- But we have very little agreement on how best to teach it.

How do we teach data science?

Much work on code, version control, less work on writing, ethics, workflows

2021:

- **Tiffany Timbers**, "Teaching reproducibility: Motivation, direct instruction and practice"
- **Tyler Girard** "Replication as a Pedagogical Tool"

2022:

- **Aneta Piekut,** "Integrating reproducibility into the curriculum of an undergraduate social sciences degree"
- **Aya Mitani,** "Reproducible, reliable, replicable? In-class exercise using peer-reviewed studies"
- **Shannon Ellis,** "Structuring & Managing Group Projects in Large-Enrollment Undergraduate Data Science Courses"
- **Lars Vilhuber**, "Teaching for large-scale Reproducibility Verification"
- **Debbie Yuster**, "Infusing Reproducibility into Introductory Data Science"

2023

- **Mine Dogucu**, "Reproducible Teaching in Statistics and Data Science Curricula"

5. What does the relationship between industry and academia look like?

Lessons from industry

- **Brittany Witham**, Geopolitica, “Data science at a startup”
- **Meg Risdal**, Kaggle, “Lessons from running data science competitions”
- **David Shor**, Blue Rose Research, “Data Science and US Politics”
- **Laura Bronner**, 538, “Quantitative editing”
- **Jacob Matson**, Simetric, Inc., “From data to dashboard”
- **Kathy Ge**, Uber, “Experimentation and product design”
- **Emily Riederer**, Capital One, “Observational causal inference”
- **Lucas Cherkewski**, Canadian Digital Service, “Using publicly-available data”

Different strengths and needs

- What does industry do well?
  - Performant code that scales
    - SQL, DuckDB, et al
  - Data engineering - reliable and correct pipelines
  - Understanding needs of users
  - Evolving and changing
    - Parquet
- What does academia do well?
  - Acknowledgement of model limitations (ideally).
  - Opportunity for perfectionism
    - Stan
  - Uncertainty quantification

Lessons for (data science) academia?

- Critical skills: writing, Python, R, Git and GitHub, SQL, DuckDB et al, Parquet.
- Critical frameworks:
  - Thinking in an algorithmic way.
  - Sampling design.
- Two-way communication with stakeholders - what do they know that we do not?
- Data engineering - the need for reliable and correct pipelines.
- R Studio needs to be complemented with VS Code. Notebooks need to be complemented with scripts.

Thank you!

rohanalexander.com

rohan.alexander@utoronto.ca

Upcoming Fall 2023:

- **Saloni Dattani**, Our World in Data
- **Nima Sarajpoor**, Manulife
- **Lindsay Katz and Zane Schwartz**, The Investigative Journalism Foundation
- **Tom Cardoso**, The Globe and Mail
- **Marzieh Fadaee**, Cohere for AI
- **Wendy Foster**, Shopify
- **Apoorva Lal**, Netflix
- **Annie Collins**, Giving Tuesday
- **Adam McAskill**, Evidence BI

Free, all welcome, sign up: https://forms.gle/sXbEixoa1iJR4Q7A8

References

- Barba, Lorena, 2018, “Terminologies for Reproducible Research”, https://arxiv.org/abs/1802.03311.
- Buckheit, Jonathan, and David Donoho. 1995. “Wavelab and Reproducible Research.” In Wavelets and Statistics, 55–81. Springer. https://doi.org/10.1007/978-1-4612-2544-7_5.
- Gelman, Andrew and Eric Loken, 2012, “Statisticians: When We Teach, We Don’t Practice What We Preach”, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics2.pdf.
- Horton, Nicholas, Rohan Alexander, Micaela Parker, Aneta Piekut, and Colin Rundel. 2022. “The Growing Importance of Reproducibility and Responsible Workflow in the Data Science and Statistics Curriculum.” Journal of Statistics and Data Science Education 30 (3): 207–8. https://doi.org/10.1080/26939169.2022.2141001.
- Wallach, Hanna. 2018. “Computational Social Science Computer Science + Social Data.” Communications of the ACM 61 (3): 42–44. https://doi.org/10.1145/3132698.