The data just got real

Preparing students to use statistics beyond the classroom

eCOTS
June 12, 2024

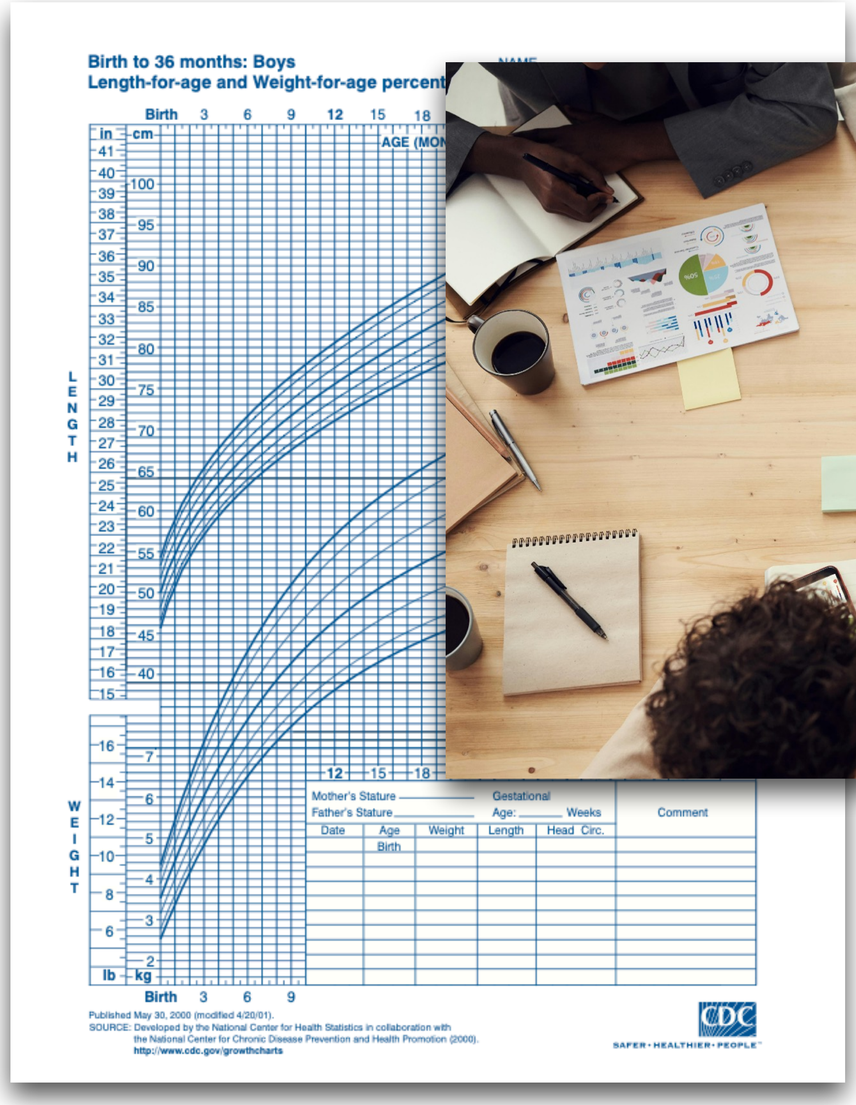
Maria Tackett
Duke University

🔗 bit.ly/ecots24-beyond-the-classroom

Photo by javier trueba on Unsplash

Source: MTV

“...when data stop being polite and start getting real.”



Center for Disease Control

Photo by fauxels

How to opt out of having your data 'train' ChatGPT and other AI chatbots

The reality: You are helping AI learn, whether you want to or not.

By Shira Ovide

May 31, 2024 at 12:30 p.m. EDT

Washington Post

sident_data	spend_per_resident_points	basketball_data
$62.00	30	2.4
$65.00	27.5	2.2
$62.00	10	2.2
$58.00	6	2.3
$59.00	6	2.2
32	$68.00	7	2.2
32	$63.00	6	NA
32	$64.33	6	NA
32	$60.11	5	NA
50	$115.00	63	1.4

Trust for Public Land

Student experience in statistics/ data science course(s)

Student experience in the “real world”

Student experience in statistics/ data science course(s)

Student experience in the “real world”

Methods to process, visualize, & analyze data

....and more!

the “other” skills

Curriculum guidlines: More than methods

- "...facile with **professional statistical software and other appropriate tools** for data exploration, cleaning, validation, analysis, and communication."*
- "... write clearly, speak fluently, and construct effective visual displays and compelling written summaries....They should be able to **communicate complex statistical methods in basic terms** to managers and other audiences and visualize results in an accessible manner."*
- "...demonstrate ability to **collaborate in teams** and to organize and manage projects."*
- "...exposure to and **ethical training** in areas such as citation and data ownership, security and sensitivity of data, consequences and privacy concerns of data analysis, and the professionalism of transparency and reproducibility."**

* ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*

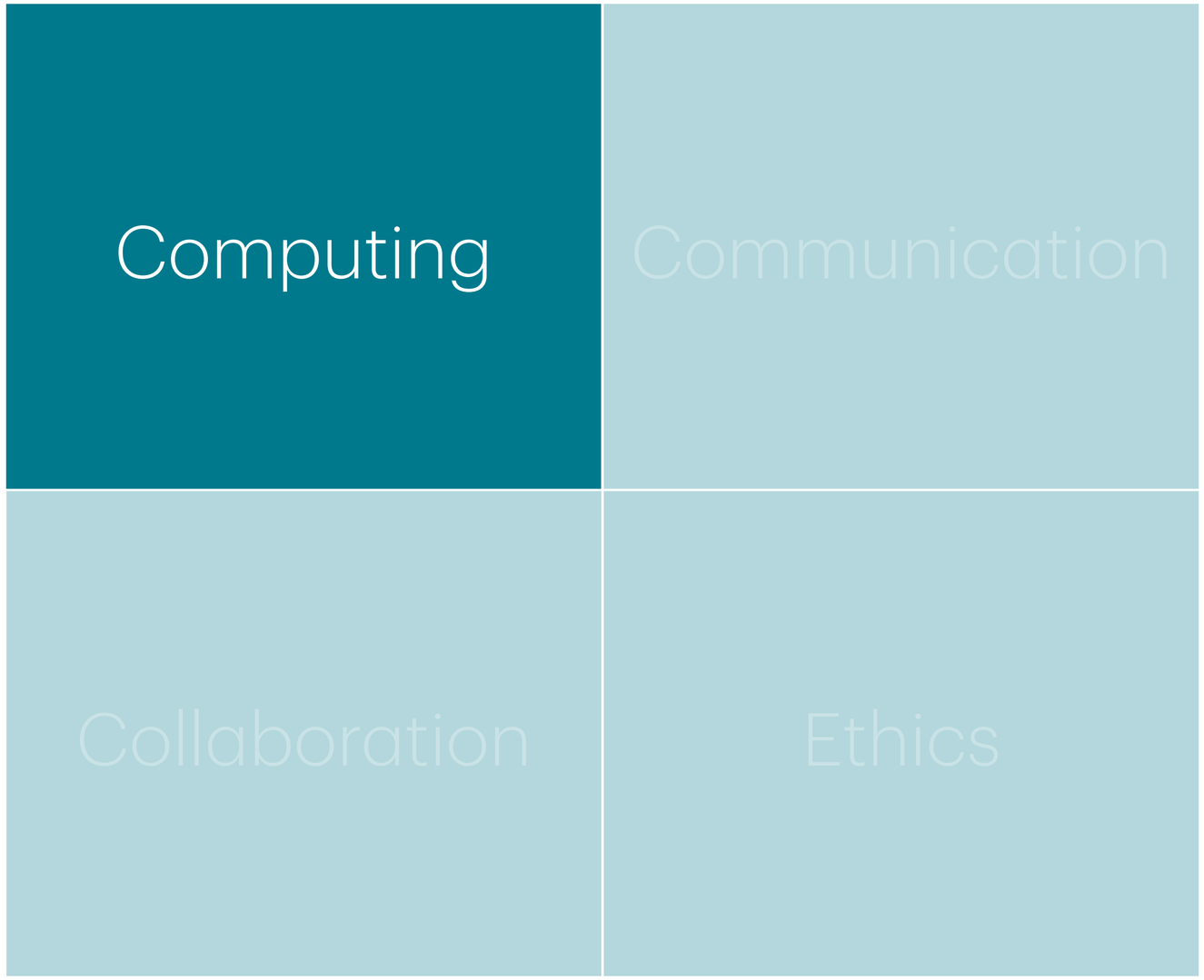
** De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

Increased enrollment in AP statistics courses and introductory statistics courses at two-year colleges.

"This shift reflects the belief that **statistics is a universal discipline**, not just needed for a handful of students, but required for a number of disciplines and recommended for many others."

ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*


Computing
Communication
Collaboration
Ethics


Computing
Communication
Collaboration
Ethics

"The goal of teaching computing and information technologies is to remove obstacles to engagement with a problem."

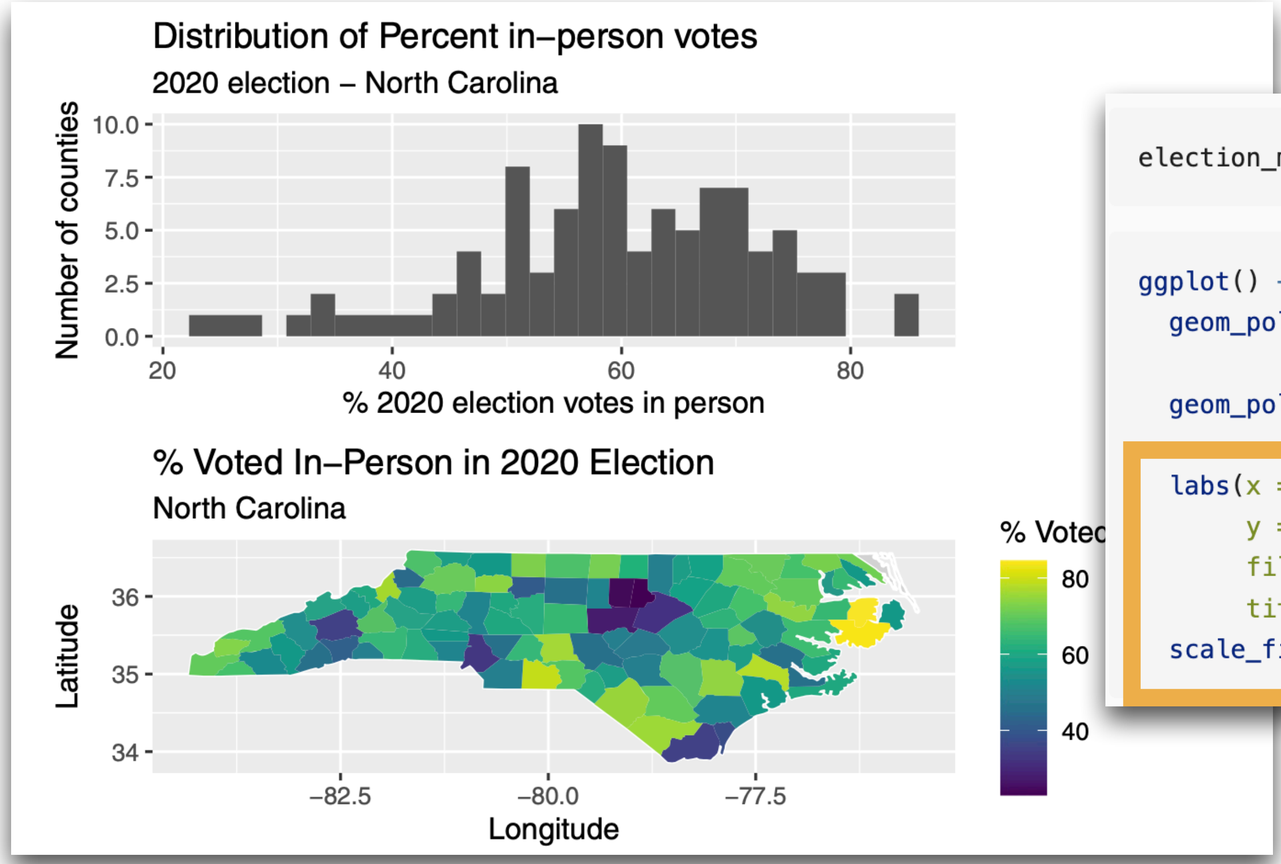
Nolan, D., & Temple Lang, D. (2010). Computing in the Statistics Curricula. The American Statistician, 64(2), 97–107. https://doi.org/10.1198/tast.2010.09132

Teaching computing

- Opportunity for students to gain experience using **professional computing tools**
  - Ability to work with more **realistic and complex data**
  - Develop a **reproducible workflow** while learning statistical methods
- Give students experience with (exposure to) computing in **early courses**
  - Incorporate **more advanced skills** in subsequent courses
  - Reach a **broad and diverse** student population

Example: Assessing spatial dependence

Students visualize the distribution of the percentage of in-person votes in North Carolina counties in the 2020 election.



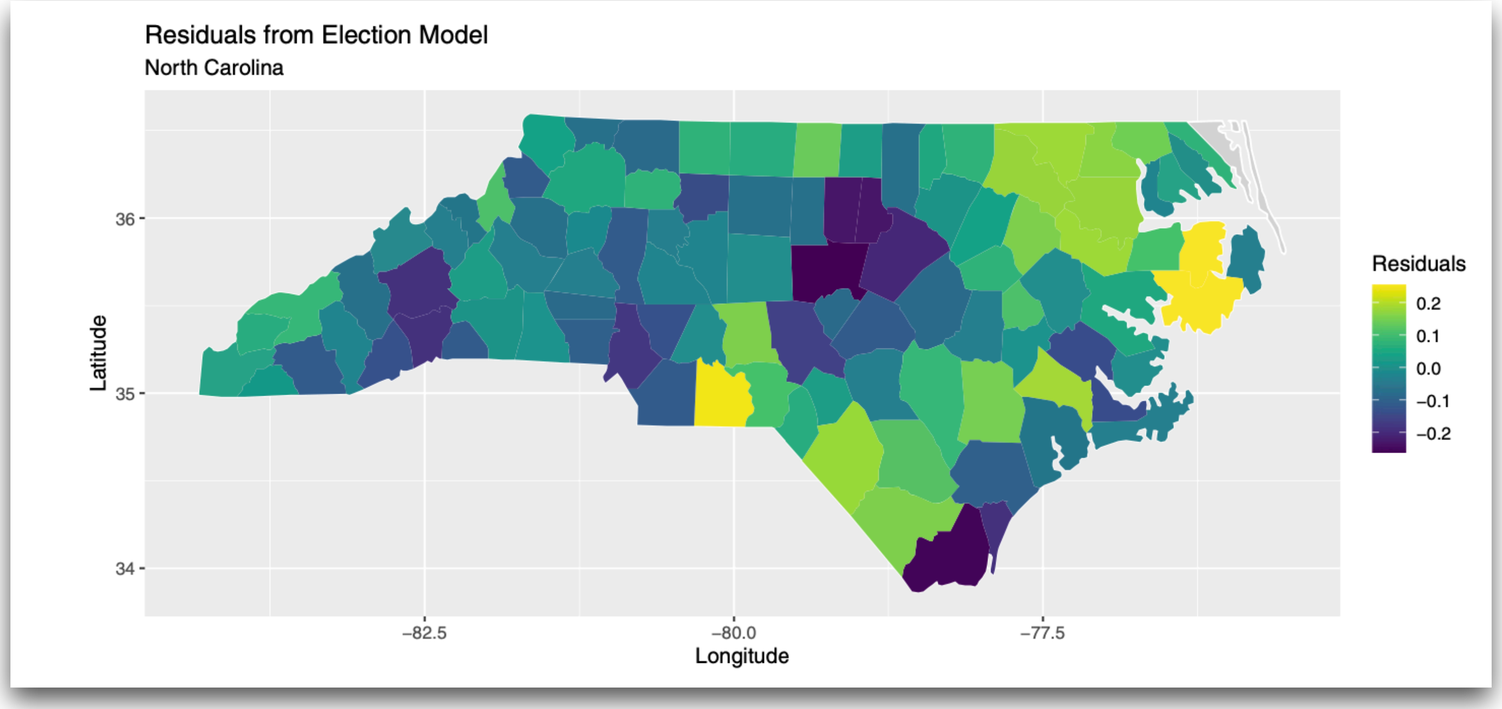

```
election_map_data <- left_join(election_nc, county_map_data)


ggplot() +
  geom_polygon(county_map_data, mapping = aes(x = long, y = lat, group = group),
               fill = "lightgray", color = "white") +
  geom_polygon(election_map_data, mapping = aes(x = long, y = lat, group = group,
                                                fill = inperson_pct)) +
  labs(x = "_____",
       y = "_____",
       fill = "_____",
       title = "_____") +
  scale_fill_viridis()
```

Full assignment at sta210-fa21.netlify.app/hw/hw-01

Example: Assessing spatial dependence

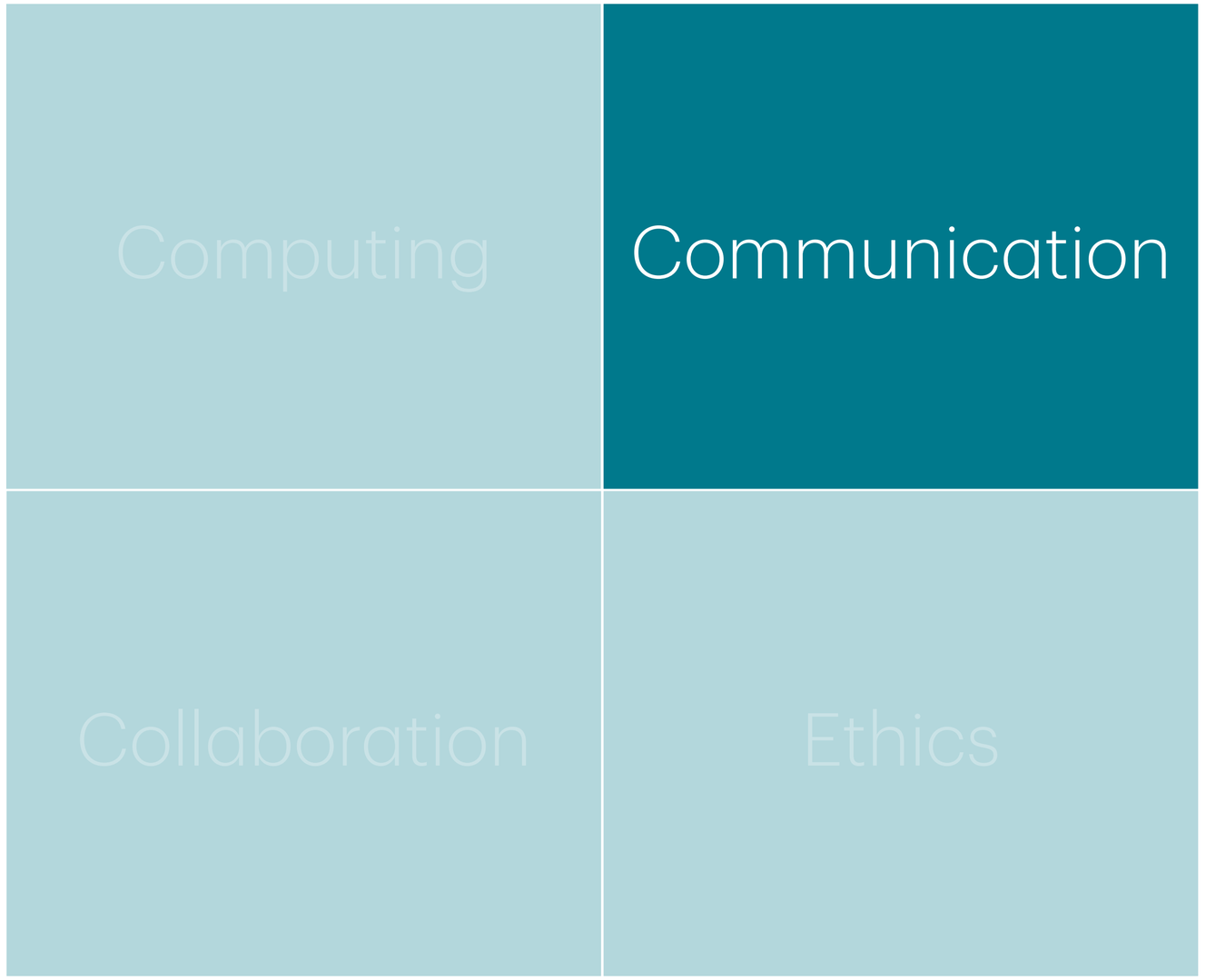
Students fit a linear regression model of the relationship between political leaning and percentage of in-person votes.

They plot the residuals on a map to assess the independence condition.



"Hint: Start with the code from Exercise 2."

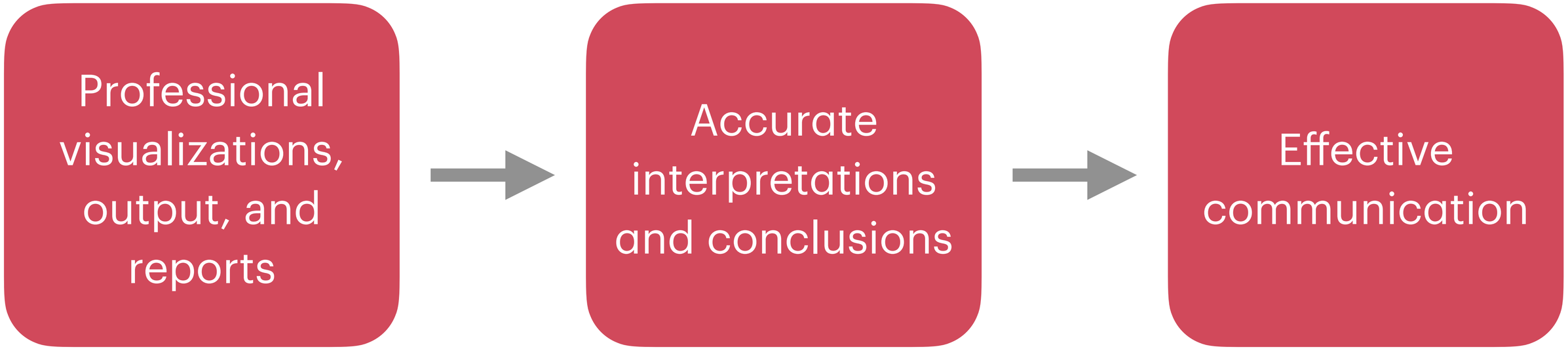
Full assignment at sta210-fa21.netlify.app/hw/hw-01


Computing
Communication
Collaboration
Ethics

"...communication is the competency that is often the difference between a truly high performing data scientist and a data scientist who is simply good."

"The most underrated skill in data science" https://medium.com/intercom-rad/the-most-underrated-skill-in-data-science-communication-7ed2fab82801

Teaching written communication



Example: Analyze data and communicate results

Analysis objective

Analyze data from the FiveThirtyEight article “Why Many Americans Don’t Vote” to examine the relationship between political party identification and voting behavior of adults in the United States.

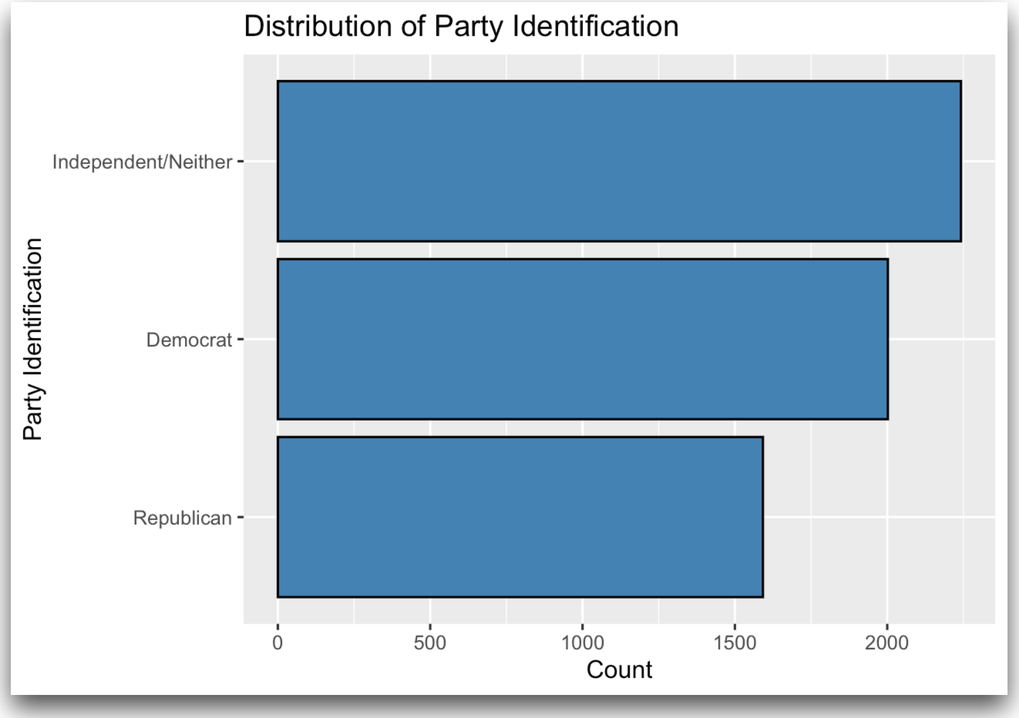
Full assignment at sta210-fa23.netlify.app/hw/hw-04

Example: Analyze data and communicate results

Students collapse the variable **`party_id`** into three levels and visualize the results.

```
voter_data <- voter_data |>
  mutate(party_id = case_when(
    Q30 == 1 ~ "Republican",
    Q30 == 2 ~ "Democrat",
    TRUE ~ "Independent/Neither"
  ))
```

Data processing



Professional visualizations

Full assignment at sta210-fa23.netlify.app/hw/hw-04

Example: Analyze data and communicate results

After fitting an initial model, students conduct a hypothesis test to determine whether to include **`party_id`** in the model.

term	df.residual	residual.deviance	df	deviance	p.value
frequent_voter ~ ppage + educ + race + gender + income_cat	4366	5072.595	NA	NA	NA
frequent_voter ~ ppage + educ + race + gender + income_cat + party_id	4364	5052.151	2	20.444	0

Full assignment at sta210-fa23.netlify.app/hw/hw-04

Example: Analyze data and communicate results

After fitting an initial model, students conduct a hypothesis test to determine whether to include **`party_id`** in the model.

"Our p-value is approximately 0. ***Since it is very small, we reject*** H_0***.*** *The data provides sufficient evidence that at least one of the coefficients from `party_id` is not equal to 0.* ***Therefore, we should add `party_id` to the model.***"

- Student response (emphasis mine)

Accurate interpretations

Full assignment at sta210-fa23.netlify.app/hw/hw-04

Example: Analyze data and communicate results

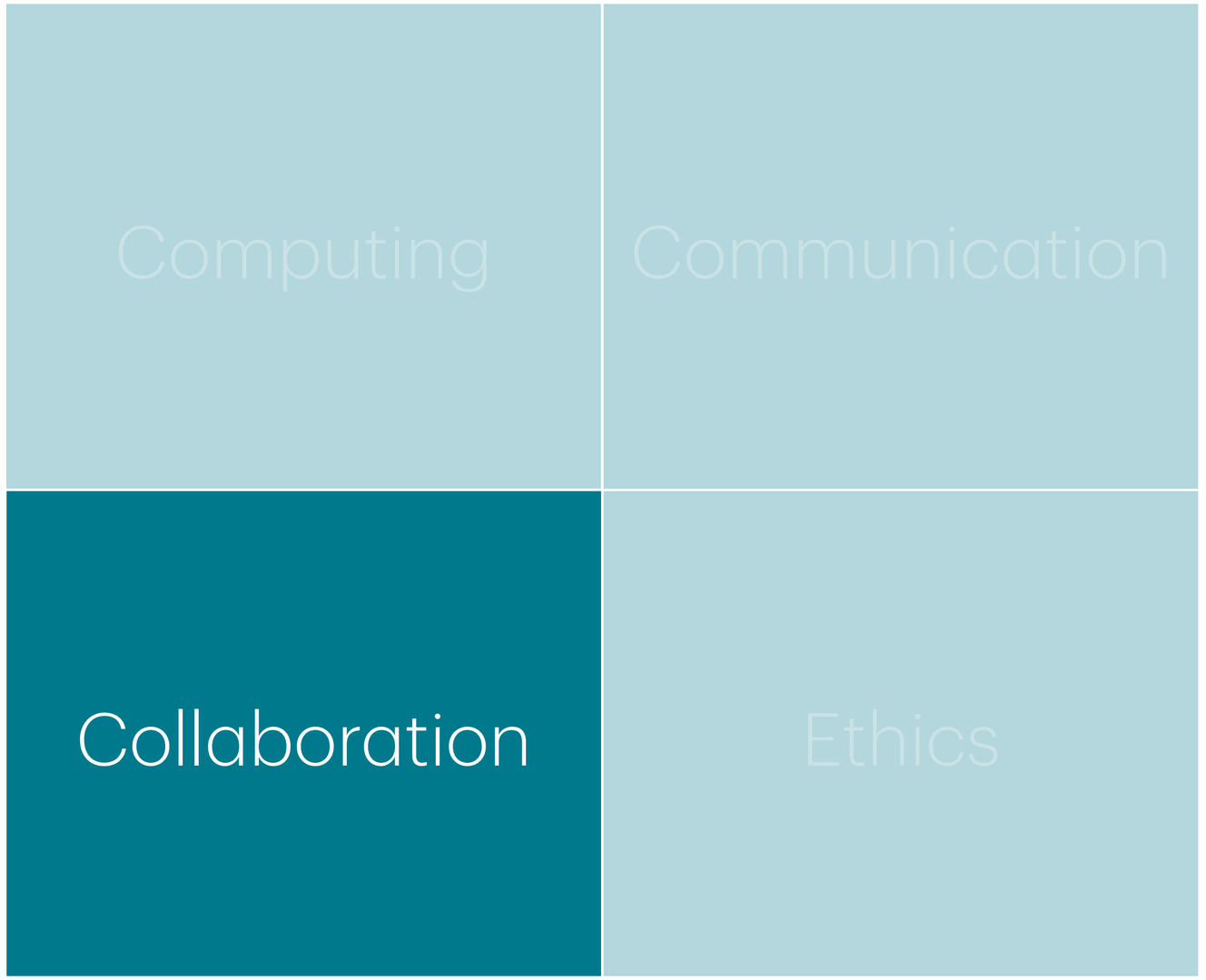
Students write a short paragraph summarizing the association between party identification and voting behavior.

"...The biggest difference between the odds is in the Independent/Neither party_id with the Republican and Democratic party_id respectively. ***This tells us that voters who identify as Independent or Neither are much less likely to be frequent voters than their Republican and Democratic counterparts.*** *Whereas the odds between frequent voters who identify as being* ***Republican and Democrat are similar.****"*

- Student response (emphasis mine)

Effective communication

Full assignment at sta210-fa23.netlify.app/hw/hw-04


Computing
Communication
Collaboration
Ethics

IBM: Data science is a team sport.

Why incorporate teamwork?

Model collaborative and interdisciplinary nature of data-driven work in practice

Skills developed

- Communication
- Interpersonal
- Computing for collaborative work

Skills enhanced

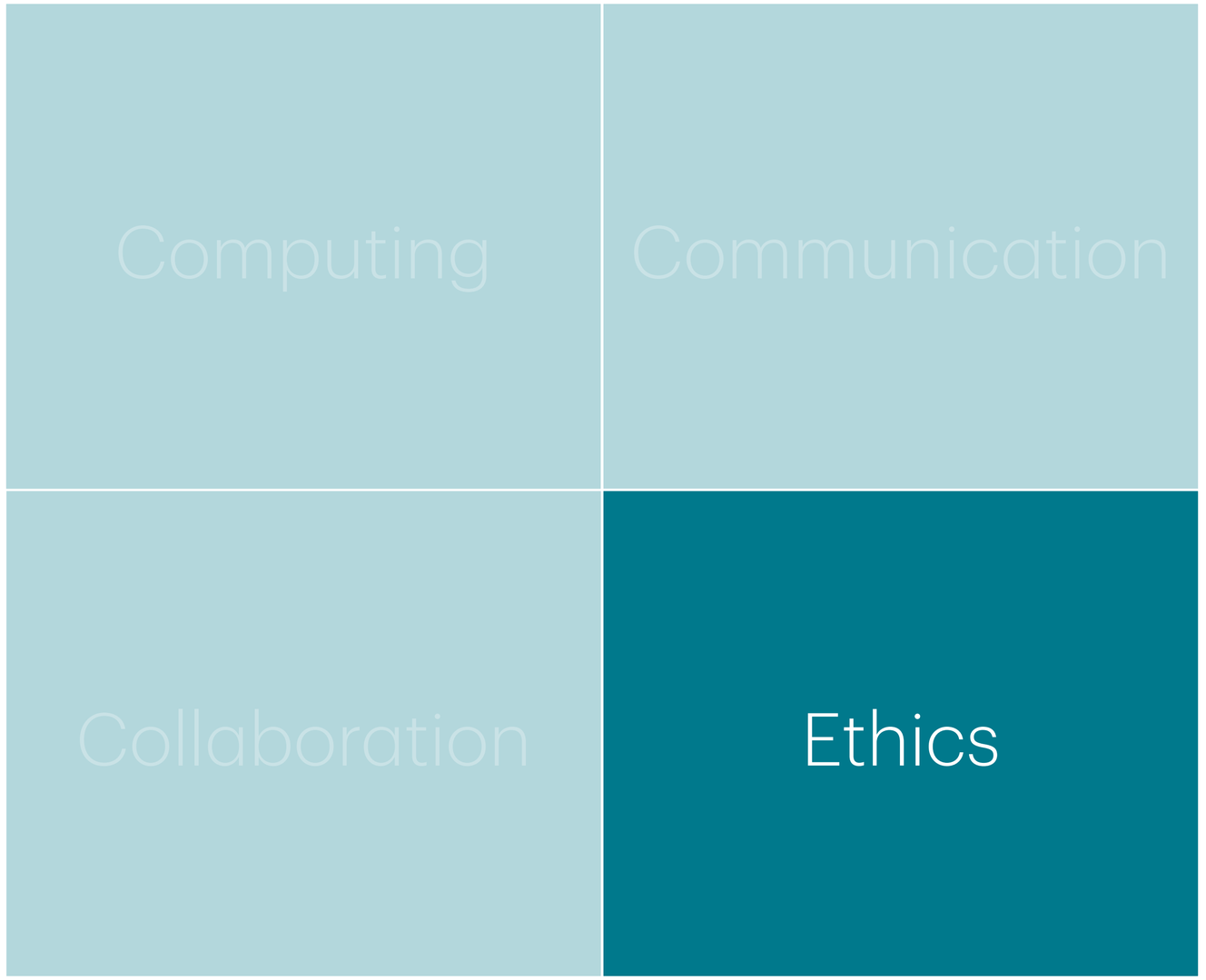
- Statistical reasoning
- Problem solving
- Critical thinking
- Conceptual understanding

Adapted from Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in Learning and Teaching Statistics. Journal of Statistics Education, 16(1). https://doi.org/10.1080/10691898.2008.11889557

Strategies for effective teamwork

- Form **diverse teams** and avoid isolating students
- Keep the teams **consistent**
- Provide **guidance** on how to work on a team
- Incorporate regular **peer feedback and self-reflection**
- Dedicate time for teams to **work together during class**
- Start with a **team agreement**

Adapted from Carnegie Mellon University Eberly Center. *Using Group Projects Effectively*: cmu.edu/teaching/designteach/design/instructionalstrategies/groupprojects


Computing
Communication
Collaboration
Ethics

TECHNOLOGY

Cambridge Analytica made “ethical mistakes” because it was too focused on regulation, former COO says

“It felt like, well, once that was done, then we’ve done what we needed to do, and we forgot to pause and think about, ethically, what was going on.”

by **Eric Johnson**
Jul 31, 2019, 6:20 AM EDT

Vox

TECHNOLOGY

How big data is helping states kick poor people off welfare

“These systems make our values visible to us in a way that calls us to a moral reckoning.”

by **Sean Illing**
Feb 6, 2018, 9:00 AM EST

Vox

AI researchers uncover ethical, legal risks to using popular data sets

The Data Provenance Initiative analyzed data sets used to build generative AI and found confusion surrounding licensing and fair use

By Nitasha Tiku

October 25, 2023 at 12:01 p.m. EDT

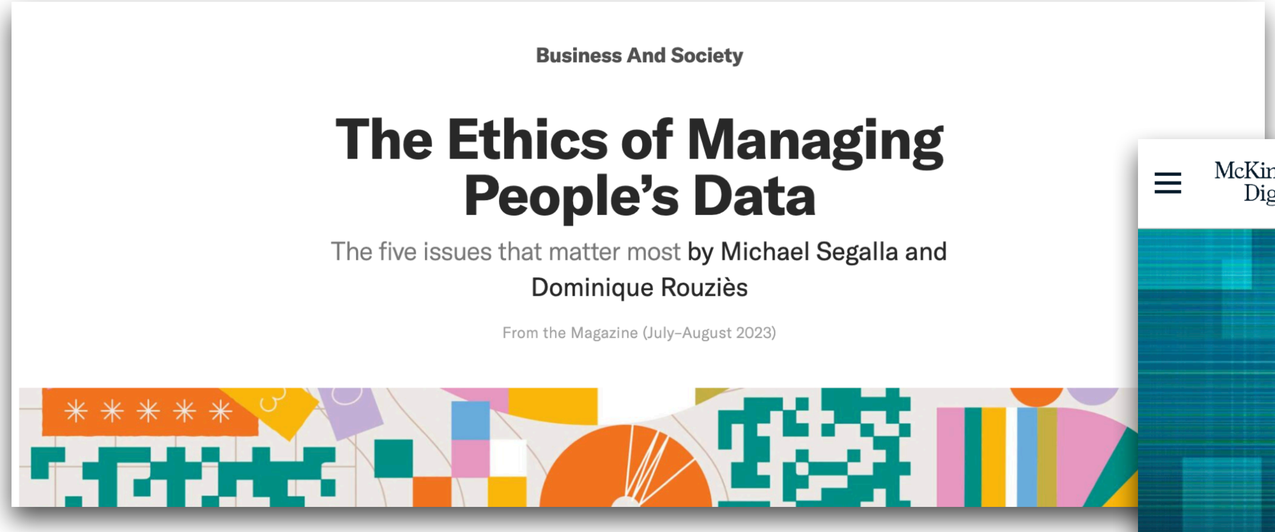
Washington Post

Opinion

Boeing’s manufacturing, ethical lapses go back decades

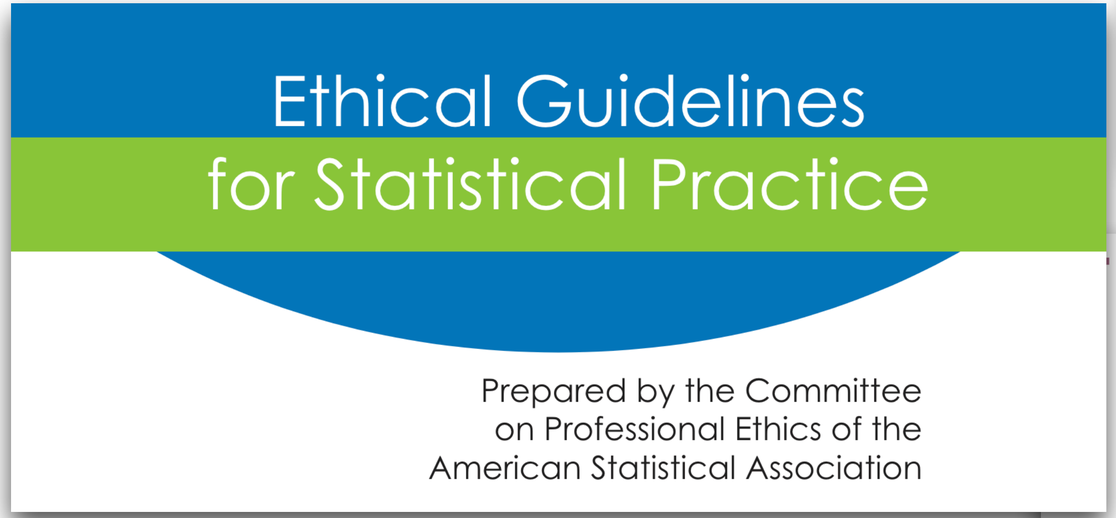
Jan. 22, 2024 at 2:19 pm | *Updated Jan. 22, 2024 at 3:19 pm*

The Seattle



Harvard Business Review

McKinsey



American Statistical Association

Slate

Ethical considerations

What you're doing

- Data security and privacy
- Honest reporting
- Algorithmic bias
- ...

How you're doing it

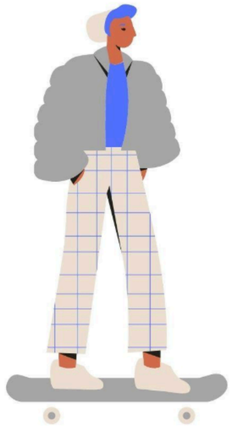
- Reproducibility
- Transparency
- Treatment of others
- ...

What you're doing: Proxy variables

Redlining Analogy

Individual 1

- Credit score: 650
- Income: $70,000/yr
- Hair color: green
- Shirt color: green



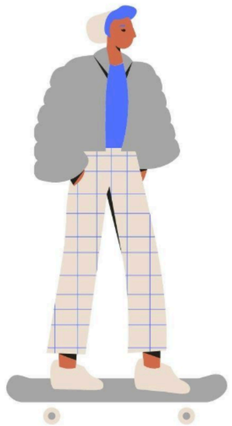
Data Science Ethics Deep Dive

What you're doing: Proxy variables

Redlining Analogy

Individual 2

- Credit score: 650
- Income: $70,000/yr
- Hair color: blue
- Shirt color: blue



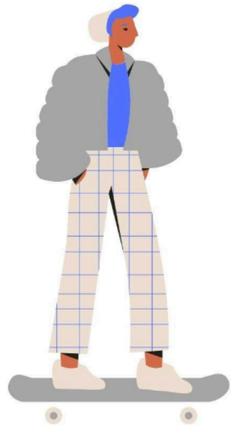
Data Science Ethics Deep Dive

What you're doing: Proxy variables

Redlining Analogy

Individual 2

- Credit score: 650
- Income: $70,000/yr
- **Race**: blue
- **Zip code**: blue



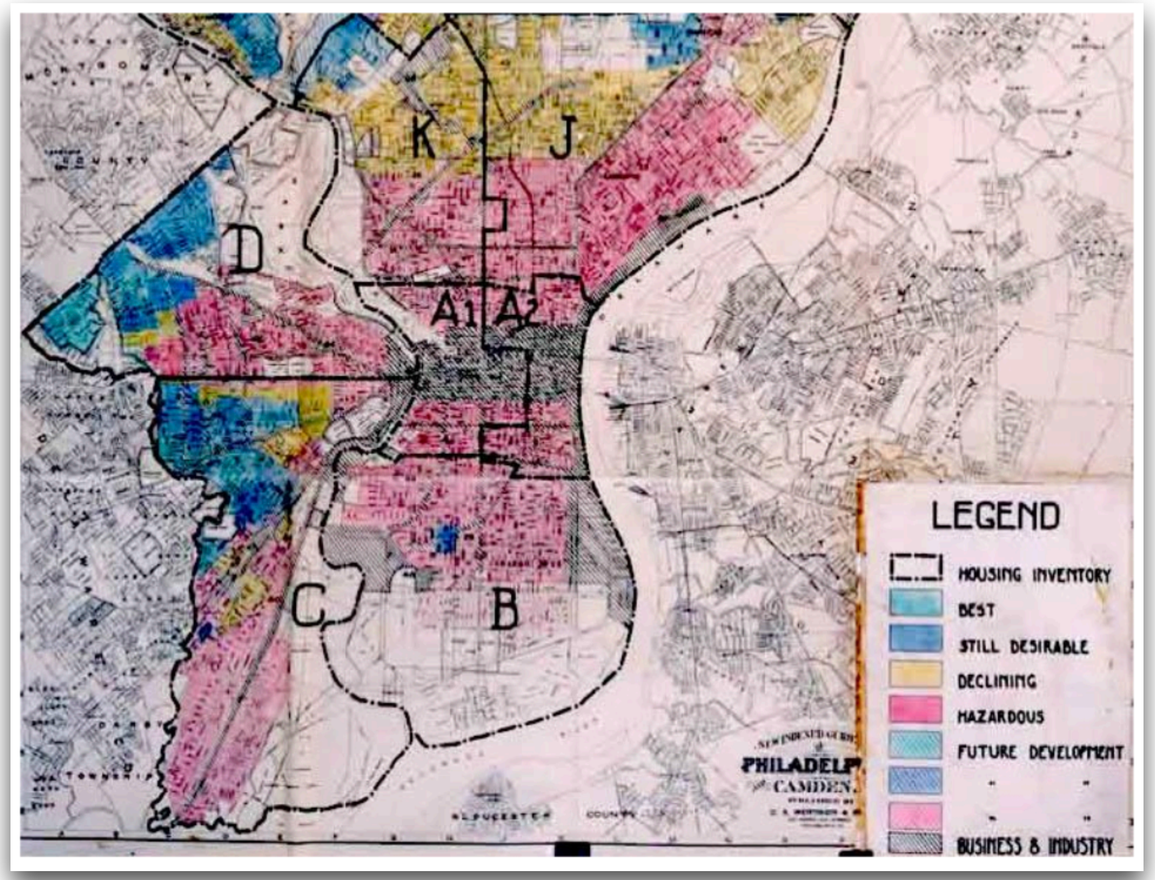
Data Science Ethics Deep Dive

What you're doing: Proxy variables

"What are the ethical considerations of including hair color and shirt color in a model?"

What you're doing: Proxy variables

Redlining: practice of denying people access to credit because of where they live, even if they are personally qualified for loans.*



Data Science Ethics Deep Dive

* Definition from federalreservehistory.org

How you're doing it: Course policy on using AI

- **Use of artificial intelligence (AI)**: You should treat AI tools, such as ChatGPT, the same as other online resources. There are two guiding principles that govern how you can use AI in this course:[2] (1) *Cognitive dimension:* Working with AI should not reduce your ability to think clearly. We will practice using AI to facilitate—rather than hinder—learning. (2) *Ethical dimension*: Students using AI should be transparent about their use and make sure it aligns with academic integrity.
    - **AI tools for code:** You may make use of the technology for coding examples on assignments; if you do so, you must explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. You may use these guidelines for citing AI-generated content.
    - **No AI tools for narrative:** Unless instructed otherwise, AI is **not** permitted for writing narrative on assignments. In general, you may use AI as a resource as you complete assignments but not to answer the exercises for you. You are ultimately responsible for the work you turn in; it should reflect your understanding of the course content.

sta210-fa23.netlify.app/syllabus

Guiding principles*

Cognitive dimension: AI should facilitate, not hinder, learning

Ethical dimension: Be transparent about use and make sure it aligns with academic integrity

*Adapted from Policies related to ChatGPT and other AI tools by Joel Gladd

Final thoughts

- **Start small**: Slowly incorporate the “other” skills in your course
- **Be intentional**: Include these skills in the course learning objectives
- **Explain why**: Get student buy-in
- **Be open:** Try new things and get student feedback early

Incrementally

Bridge the gap between the classroom and the "real world"

Thank you!

maria.tackett@duke.edu

🔗 bit.ly/ecots24-beyond-the-classroom

References

- **Curriculum guidelines**
  - American Statistical Association Undergraduate Guidelines Workgroup. 2014. *2014 curriculum guidelines for undergraduate programs in statistical science*. Alexandria, VA: American Statistical Association.
  - De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Ye, P. (2017). *Curriculum guidelines for undergraduate programs in data science*. Annual Review of Statistics and Its Application, 4, 15-30.
  - National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

- **Computing**
  - Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with Git and GitHub as a learning objective in statistics and data science courses. *Journal of Statistics and Data Science Education*, *29*(sup1), S132-S144.
  - Nolan, D., & Temple Lang, D. (2010). Computing in the Statistics Curricula. *The American Statistician*, 64(2), 97–107. https://doi.org/10.1198/tast.2010.09132
  - 2021 Journal of Statistics and Data Science Education special issue *Computing in the Statistics and Data Science Curriculum*. https://www.tandfonline.com/toc/ujse21/29/sup1
  - 2022 Journal of Statistics and Data Science Education special issue *Teaching Reproducibility*. https://www.tandfonline.com/toc/ujse21/30/3

References

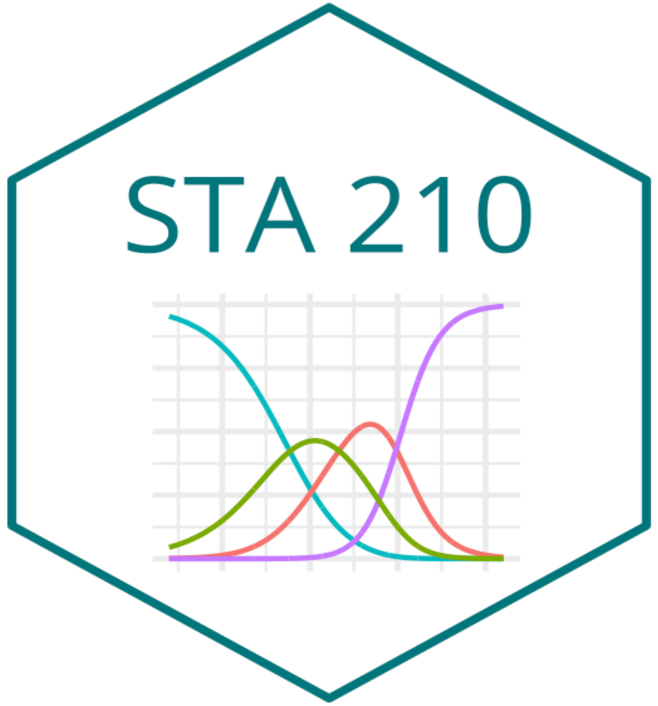
- **Communication**
  - Cline, K. S. (2008). A writing-intensive statistics course. *Primus*, *18*(5), 399-410.
  - Nolan, D., & Stoudt, S. (2021). *Communicating with data: The art of writing for data science*. Oxford University Press.

- **Collaboration**
  - Carnegie Mellon University Eberly Center. Using Group Projects Effectively: cmu.edu/teaching/designteach/design/instructionalstrategies/groupprojects
  - Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in Learning and Teaching Statistics. Journal of Statistics Education, 16(1). https://doi.org/10.1080/10691898.2008.11889557
  - Vance, E. A. (2021). Using team-based learning to teach data science. *Journal of Statistics and Data Science Education*, *29*(3), 277-296.

- **Ethics**
  - Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M., & Ott, M. Q. (2022). Integrating Data Science Ethics Into an Undergraduate Major: A Case Study. *Journal of Statistics and Data Science Education*, *30*(1), 15–28. https://doi.org/10.1080/26939169.2022.2038041
  - Glanz, H., Hardin, J., Horton, N. (2020) *Teach Data Science*: https://teachdatascience.com/tags/ethics/

Courses I teach

[Introduction to Data Science]



[Regression Analysis*]

[Generalized Linear Models]

* Described in Tackett, M. (2023). Three principles for modernizing an undergraduate regression analysis course. *Journal of Statistics and Data Science Education*, *31*(2), 116-127.