Welcome to Data Science **in Practice**:

Tools and Applications

(ACMS 40876/60876)

Today’s Roadmap

1. **Introductions**
2. **Course overview**
3. **What is the data science life cycle?**

Introductions

- Name
- Where are you from?
- Major/Program
- Year of Study
- Research area/Advisor (if applicable)
- Why did you choose to learn about statistics/data science?

  or

  What do you hope to gain from this course?

Why statistics (beyond the math, methods, and $$)?

Contribute to scientific knowledge

Ex. Discovering gene-gene interactions that drive cardiovascular disease

Opportunity for high impact

Ex. Predicting COVID-related cases/deaths to decide where to send PPE to do the most good

Collaboration

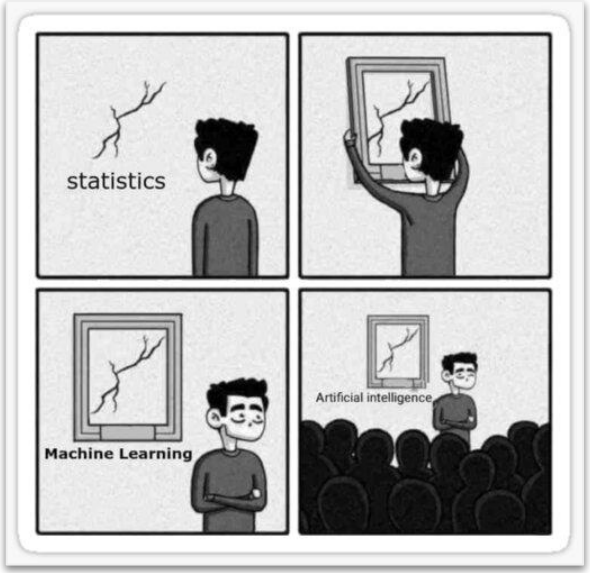
Ex. With other statisticians, domain experts, non-profits, government agencies, etc.

What do *I* hope you gain from this course?

My (babbling) view of statistics/data science

1. **Statistics** and **data science** go hand-in-hand.

- + A good statistician practices good data science
- + A good data scientist practices good statistics



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

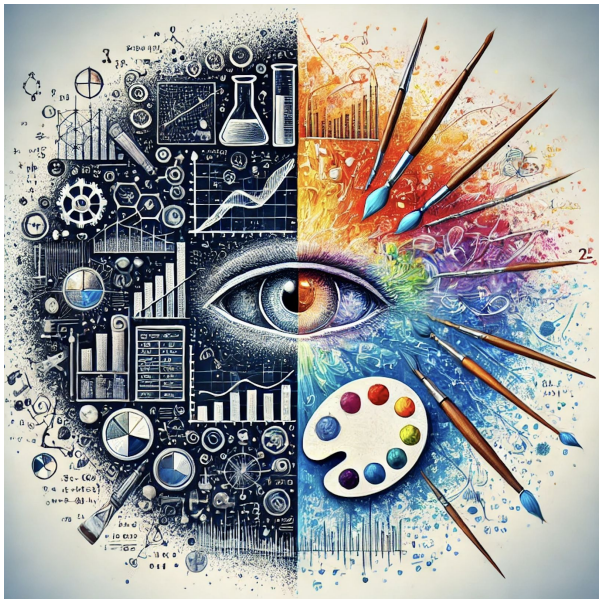
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current prob-

My (babbling) view of statistics/data science

2. Statistics is both a **science** and an **art!**

- The scientific perspective:
  - Good statistics adheres to the scientific method.
  - A good scientist **accumulates** multiple pieces of evidence **iteratively**.
  - We are scientists!
- The artistic perspective:
  - **There is no formula to a "good" statistical analysis**
  - Given the same data set, two statisticians can perform two completely different analyses, each with their own merits but of similarly high-quality.
  - There is more than one "right" answer.

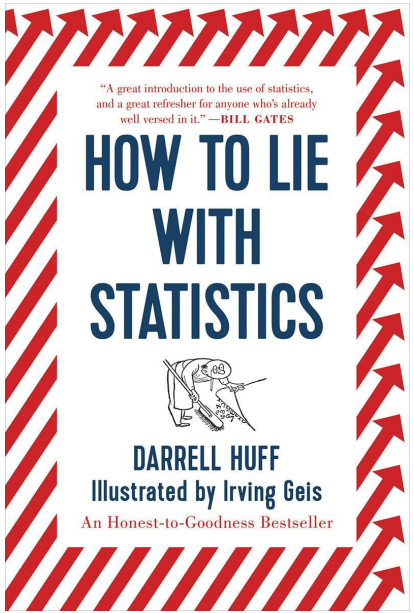


* Generated via Dall-E

My (babbling) view of statistics/data science

3. Everyone can tell a story using data,
but we want to tell a **truthful** story using data

- We, as statisticians, wield incredible power - we can not only spread false information but can do so with loads of “evidence”.
- Be extremely conscious of this power and wield it responsibly!

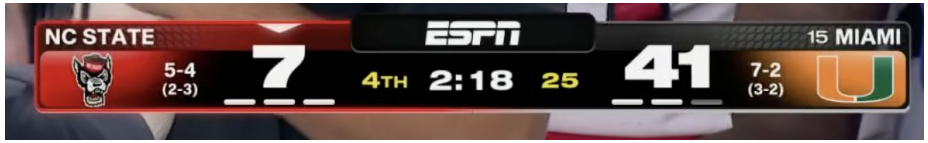


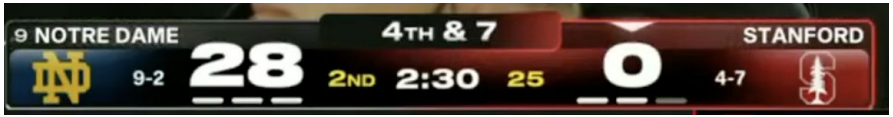
How to lie with statistics

10 Miami Hurricanes (7 - 2)	38	-	10	Syracuse Orange (3 - 7)
10 Miami Hurricanes (6 - 1)	42	-	7	Stanford Cardinal (3 - 5)
10 Miami Hurricanes (10 - 2)	38	-	7	Pittsburgh Panthers (8 - 4)
10 Miami Hurricanes (8 - 2)	41	-	7	NC State Wolfpack (5 - 5)

11 Notre Dame Fighting Irish (9 - 2)	70	-	7	Syracuse Orange (3 - 8)
11 Notre Dame Fighting Irish (10 - 2)	49	-	20	Stanford Cardinal (4 - 8)
11 Notre Dame Fighting Irish (8 - 2)	37	-	15	Pittsburgh Panthers (7 - 3)
11 Notre Dame Fighting Irish (4 - 2)	36	-	7	NC State Wolfpack (4 - 3)

The same, but **truthful** side of the story





My (babbling) view of statistics/data science

4. We can use statistics to make an **impact** on people, so why not?

- Out of all the questions that we can help answer with statistics, let's ask the *responsible* ones that *matter*.
- Why stop at the theory/modeling, rather than deploying into the real world?
- Let's operate under the assumption that we can translate data-driven insights into action.

Course Syllabus

Data science life cycle

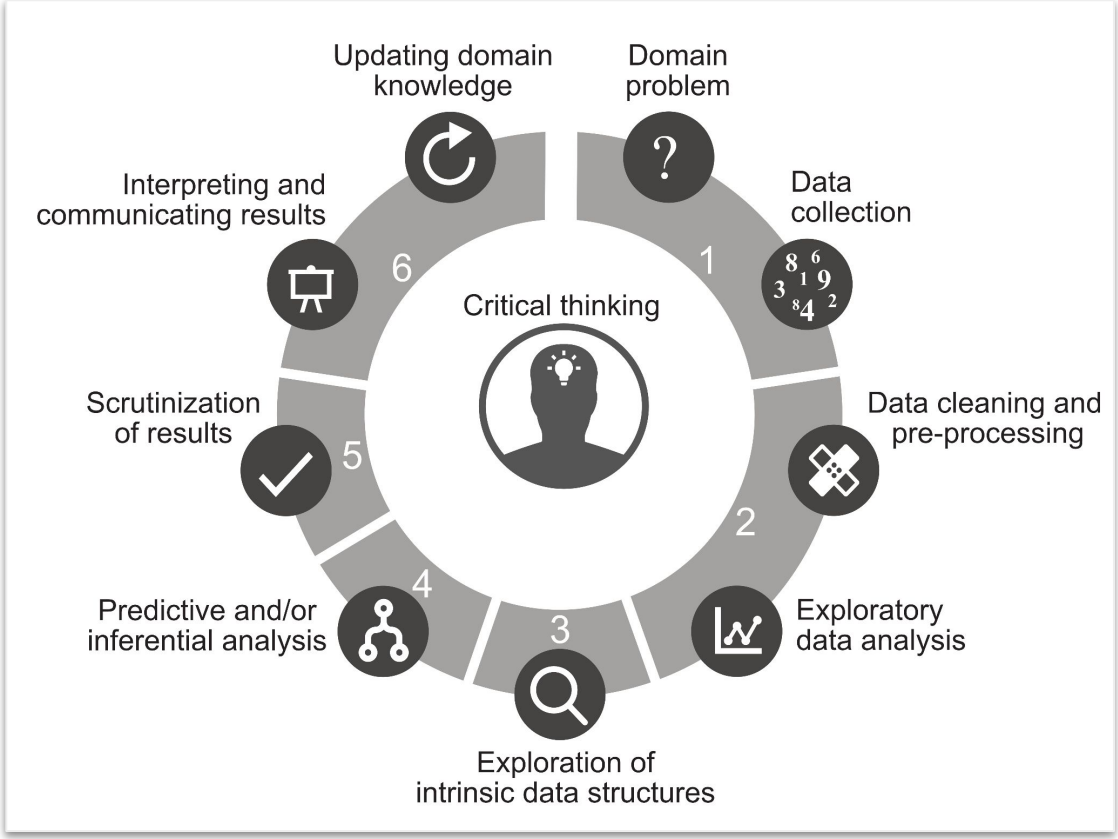
Activity: What are the stages of the data science life cycle?

Imagine you are the sole statistician/data scientist, starting a new scientific collaboration.

Brainstorm a list of tasks that you might find yourself doing in this role.

You might find it helpful to draw on some prior experiences and think sequentially.

Data Science Life Cycle



Credits: Veridical Data Science (VDS) Textbook

Resources and Next Time

- **Supplemental Resource:** Veridical Data Science (VDS) Textbook (Ch 1-2)
- **Lab 0 (optional):** Tutorial and review of R/tidyverse and python/pandas
  - Found on Canvas in lab0/ folder
  - Lab 0 is NOT graded
  - Solutions will be released on Friday
- Be on the lookout for a "tools installation" Canvas announcement
- Next time:
  - Reproducible workflow tools: git, GitHub, renv, conda, quarto

Resources and Next Time

- **Supplemental Resource:** Veridical Data Science (VDS) Textbook (Ch 1-2)
- **Lab 0 (optional):** Tutorial and review of R/tidyverse and python/pandas
  - Found on Canvas in lab0/ folder
  - Lab 0 is NOT graded
  - Solutions will be released on Friday
- Be on the lookout for a "GitHub setup" Canvas announcement
- Next time:
  - What do **problem formulation** and **data collection** look like in reality?
  - We will discuss **case studies**, focusing on these beginning stages of the data science life cycle