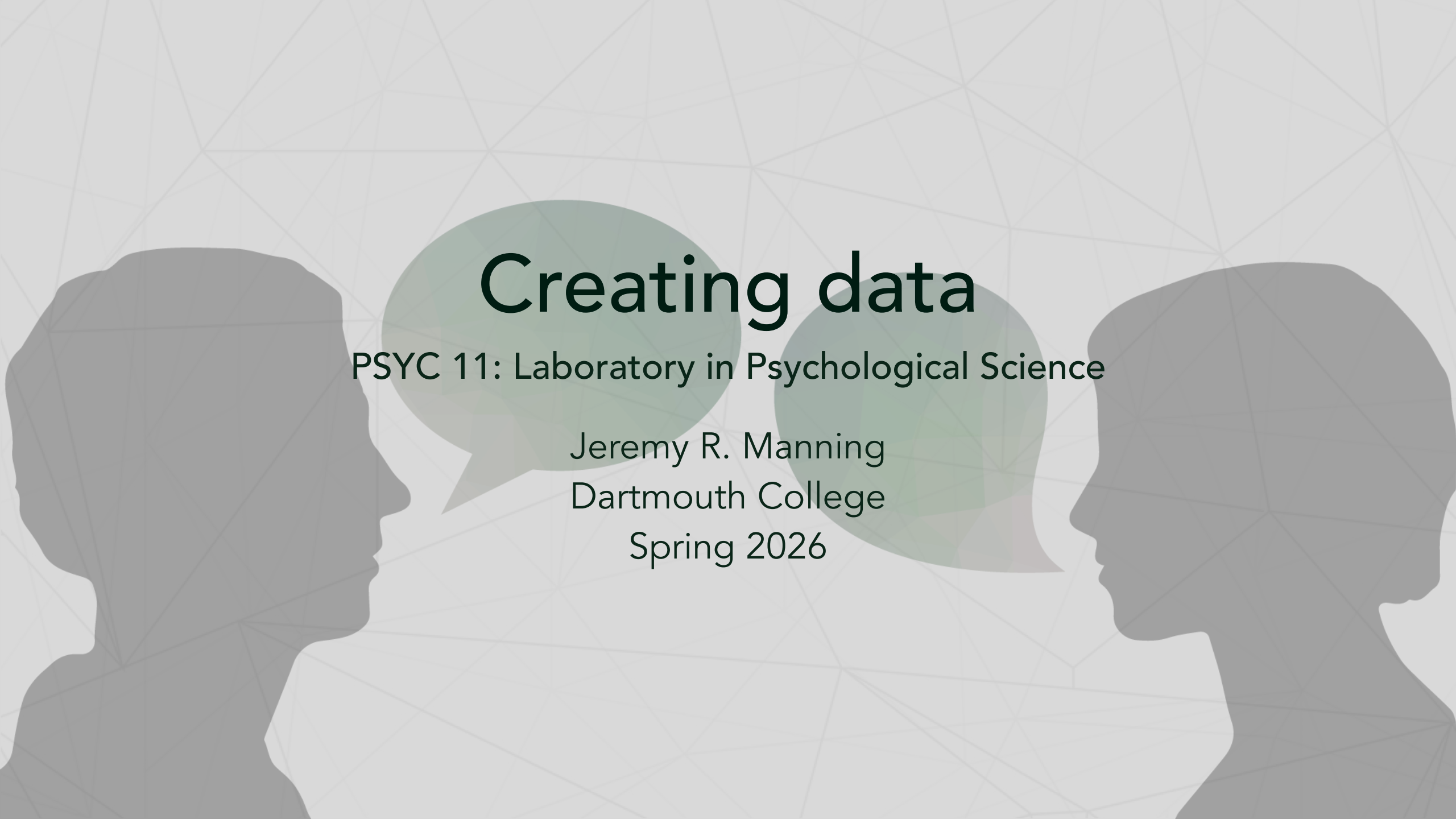


Creating data

PSYC 11: Laboratory in Psychological Science

Jeremy R. Manning
Dartmouth College
Spring 2026

What makes a good dataset?

Key properties

- Tells us about something we care about
- Has enough observations to draw conclusions
- Includes the right features to test hypotheses
- Is organized so others can work with it

What would *you* want to measure?

Think about it!

- Pick a topic you find interesting (e.g., sleep, social media, climate, sports)
- What **features** would you measure? How many observations would you need?
- What questions could you answer with that data? What questions would be **impossible** to answer?

Where do data come from?

Three approaches

- **Find existing data:** Kaggle, Google Dataset Search, FiveThirtyEight, Awesome Public Datasets
- **Generate synthetic data:** random number generators, data spoofing libraries (Faker, Mimesis)
- **Collect new data:** surveys, sensors, observations

You can use any of these in this course!

For this week's lab you'll likely want to use the first two approaches. In part II of the course you'll probably collect your own (new) data.

What can go wrong?

Common pitfalls

- Too few observations or too few features
- Missing data or inconsistent formatting
- Data that **looks** like it answers your question but actually doesn't
- Confusing correlation with causation

Answerable vs. unanswerable

Think-pair-share

- Given a dataset of 1,000 college students with: GPA, major, sleep hours, and screen time...
- Which sorts of questions **can** you answer? Which sorts of questions **can't** you answer?
- What additional features would unlock new questions?

Data sleuthing lab overview

Your two roles

- You'll play two roles: **data creator** and **data sleuth**
- **Today:** find or generate a dataset + write 5 questions about it
  - At least 1 question must be **possible** to answer with the dataset
  - At least 1 question must be **impossible** to answer
  - At least 5 features per observation, at least 500 observations
- **Thursday (X-hour):** hand off datasets (A→B, B→C, C→D, D→A) and explore
- **Friday:** wrap up analysis and discussion

What makes a good "impossible" question?

Bad: too obvious

- Asking about something completely unrelated to the dataset
- Asking about a feature that isn't in the dataset (e.g., "what's their favorite color?" when no color column exists)
- These are too easy to spot — the sleuth will know immediately

Better: subtle and interesting

- Ask about **values or patterns** that *seem* answerable but actually require data you don't have
- The sleuth should have to **analyze** the data — not just glance at the columns — to figure out whether it's possible
- The question should be interesting enough that the sleuth *wants* to answer it, even if they can't
- Examples:
  - You can only know if the question is answerable after running some analyses (e.g., groups A and B aren't distinguishable on any single feature, but they are distinguishable when you look at the interaction of two features)
  - Features co-vary in a way that makes it impossible to disentangle their effects (e.g., all the students with high screen time also have low sleep hours, so you can't tell which one is driving any observed effects on GPA)
  - The question is about a pattern that *could* be in the data but isn't (e.g., ask something about a subgroup of students with high GPA, high sleep hours, and low screen time when no such subgroup exists)

Let's get started!

Today's goal

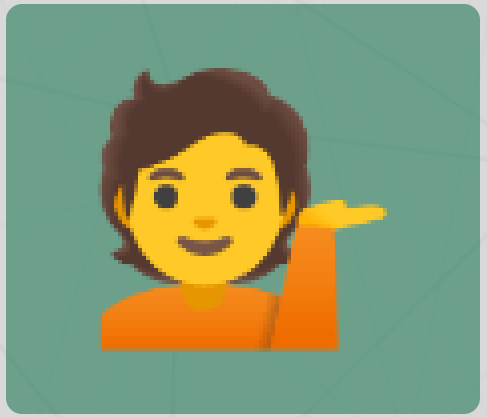
- Create your dataset and questions by the end of today's class, and **submit them using the form linked in the lab instructions**
- Think about what will make it **interesting** for the other group to explore

Lab instructions may be found here

Questions? Want to chat more?

Email me

Join our Slack



Come to office hours

Up next...

- **Wednesday**: No class (I'm away!)
- **Thursday X-hour**: Part 2 of the data sleuthing lab (exploration and analysis)
- **Friday**: Part 3 of the lab (analysis and discussion)