9.S918: Statistical Inference in Brain and Cognitive Sciences

Week 1 Day 1: Introduction to the class

Roger Levy
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology

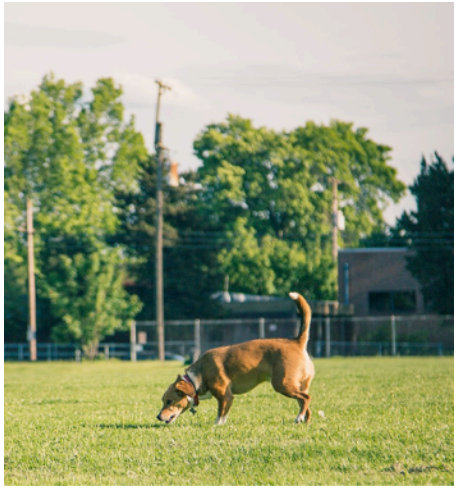
April 2, 2024

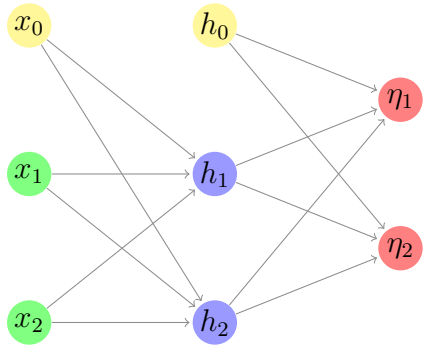
Goal of this class

- Today's datasets – experimental and observational alike – are increasingly diverse and complex
- Even simple questions often demand analyses requiring substantial background & training to really understand
- This class will help give you that background & training
- The material covered here will be of interest to many fields

Prediction versus scientifically interpretable data analysis

- Modern machine learning has brought us powerful **predictive** models – but their parameters are hard to **interpret**





"A dog in a park"

- In this class, in contrast, we'll focus primarily on approaches that give us **interpretable** fitted models

Q: *"What is the contribution of a word's probability in context to how long someone tends to spend reading it, beyond context-invariant word properties like length and frequency?"*

→ ***"About 3 milliseconds per bit of log-probability"***

Q: *"Based on existing observational data, what effect does early intervention have on childhood cognitive development?"*

→ ***"Decreases risk ratio by a factor of 0.8"***

Q: *"How should I understand the relationship between personality assessments A and B?"*

→ ***"A's Factor 1 and B's Factor 1 are similar; A's Factor 2 maps to a mix of B's factors 2 and 3"***

Main topics covered in this class

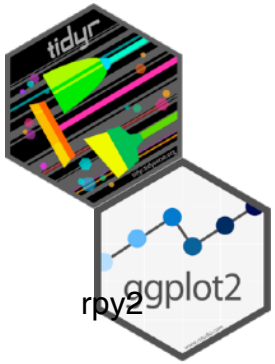
- **Causal inference:** how to distinguish between correlation and causation; counterfactual inference; estimating effects of hypothetical interventions
- **Hierarchical/mixed-effects/multi-level regression:** appropriately taking repeated measures (within subjects, items, settings...) into account by simultaneously modeling effects at multiple levels of structure in data
- **High-dimensional modeling**: dimensionality reduction; regularization; modeling correspondences across datasets
- **Throughout:** best open science practices: identifying and minimizing researcher degrees of freedom; power analysis, experiment design, and preregistration; dealing with missing or unmeasurable variables; good scientific communication

Class logistics

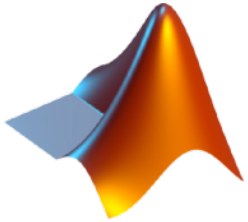
- Syllabus:
- https://rlevy.github.io/statistical-inference-spring-2024/
- We meet here in 46–3310 TuTh 9–10:30am
- Class will be primarily lecture plus discussion + Q&A, with occasional practicum sessions; bring your laptops!
- Evaluation:
  - Weekly problem sets
  - A class project (your own data analysis, experiment design, simulation, and/or something similar; start thinking now!)
- Generative AI policy: use GenAI tools however you can to assist your learning and growth, but remember they can't be trusted!

Software in class

- The official class language is `R` (including some parts of the tidyverse), plus the probabilistic programming language `Stan` for some applications



- You may be more comfortable in Python or Matlab: these languages are fine, but some of what we will do will require calling `R` from these languages



Rpy2: https://pypi.org/project/rpy2/

Rcall: https://www.mathworks.com/matlabcentral/fileexchange/104945-rcall

Questions about overall class content or logistics?