Data Science Concepts and Analysis

Week 1: Welcome to PSTAT 100!

- Course introduction
- Course structure
- Getting started

This week

- Course introduction
  - Perspectives on data science
  - Scope and topics
- Course structure
  - Format and schedule
  - Materials and resources
  - Assignments and assessment
  - Course policies

Course introduction

- Perspective on data science: “lifecycle”
- Course scope

What’s data science?

Currently understood, “data science” encompasses a wide range of activities that involve *uncovering insights from quantitative information*.

Data scientists typically combine specific interests (“domain knowledge”, *e.g.*, biology) with computation, mathematics, and statistics and probability to contribute to knowledge in their communities.

Data science lifecycle

There is an emerging consensus that doing data science involves proceeding through a **lifecycle**: *a repeated sequence of steps.*

- Less consensus at the moment about how many steps and what they are (google ‘data science lifecycle’ and check out all the flowcharts).

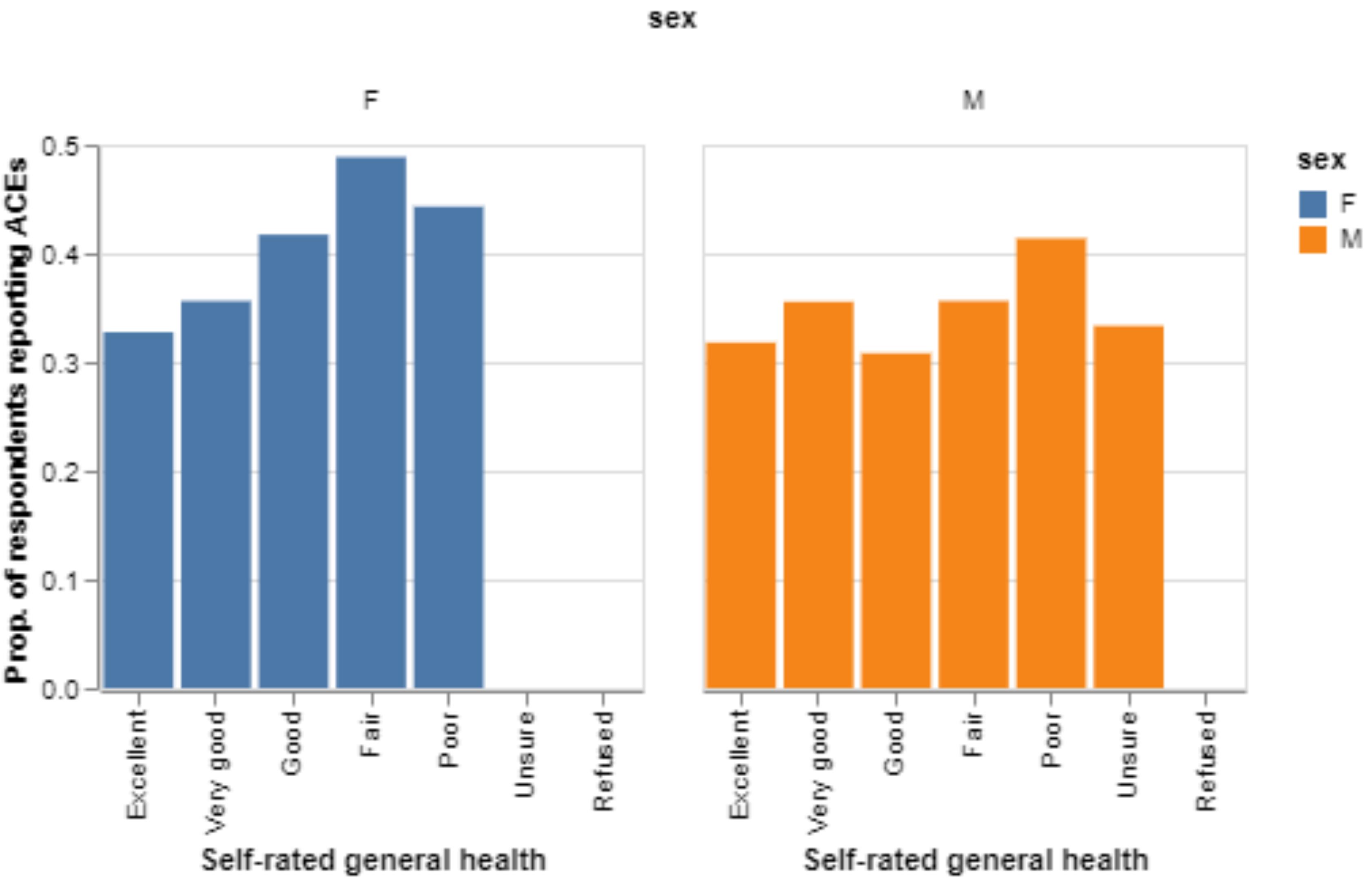
Most versions of the ‘data science lifecycle’ involve a few categories of steps:

- Project planning
- Data collection and organization
- Exploration
- Analysis
- Communication and interpretation

(Perhaps it is this idea of a lifecycle that characterizes data science as distinct from other quantitative fields.)

Case studies: a preview

Case study 1: ACE and health



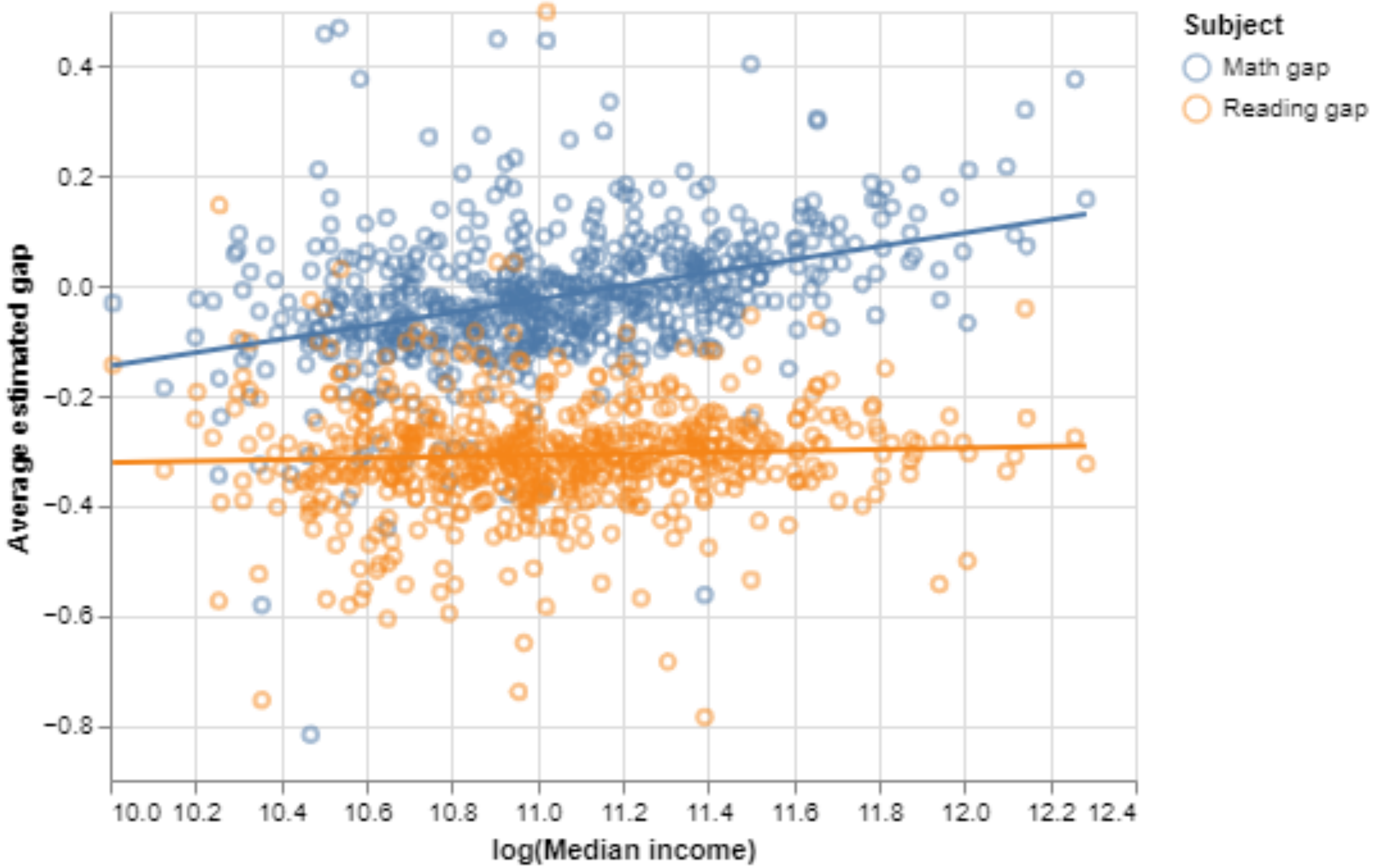
Association between adverse childhood experiences and general health, by sex.

Case study 1: ACE and health

You will:

- process and recode 10K survey responses from CDC's 2019 behavior risk factor surveillance survey (BRFSS)
- cross-tabulate health-related measurements with frequency of adverse childhood experiences

Case study 2: SEDA



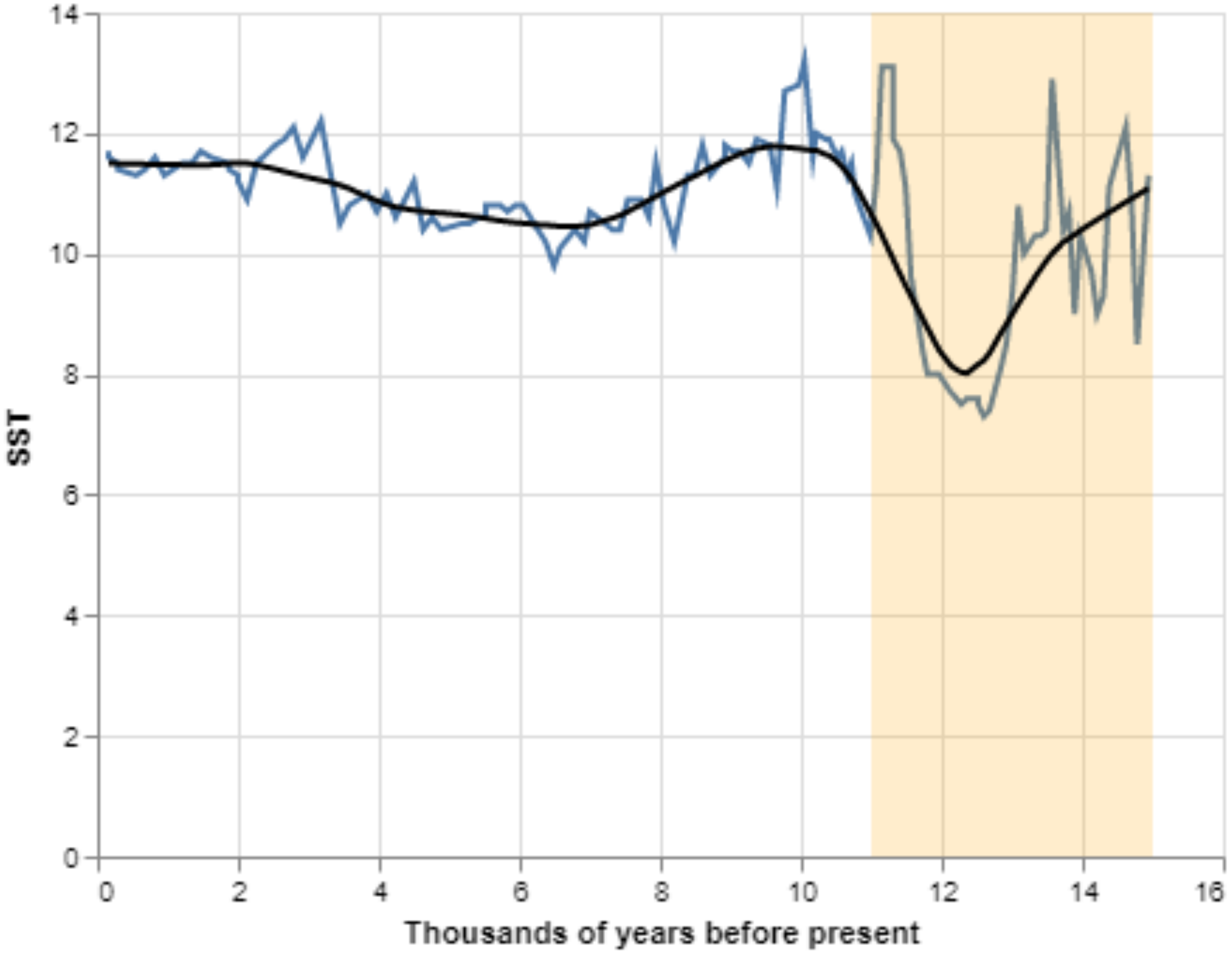
Education achievement gaps as functions of socioeconomic indicators, by gender.

Case study 2: SEDA

You will:

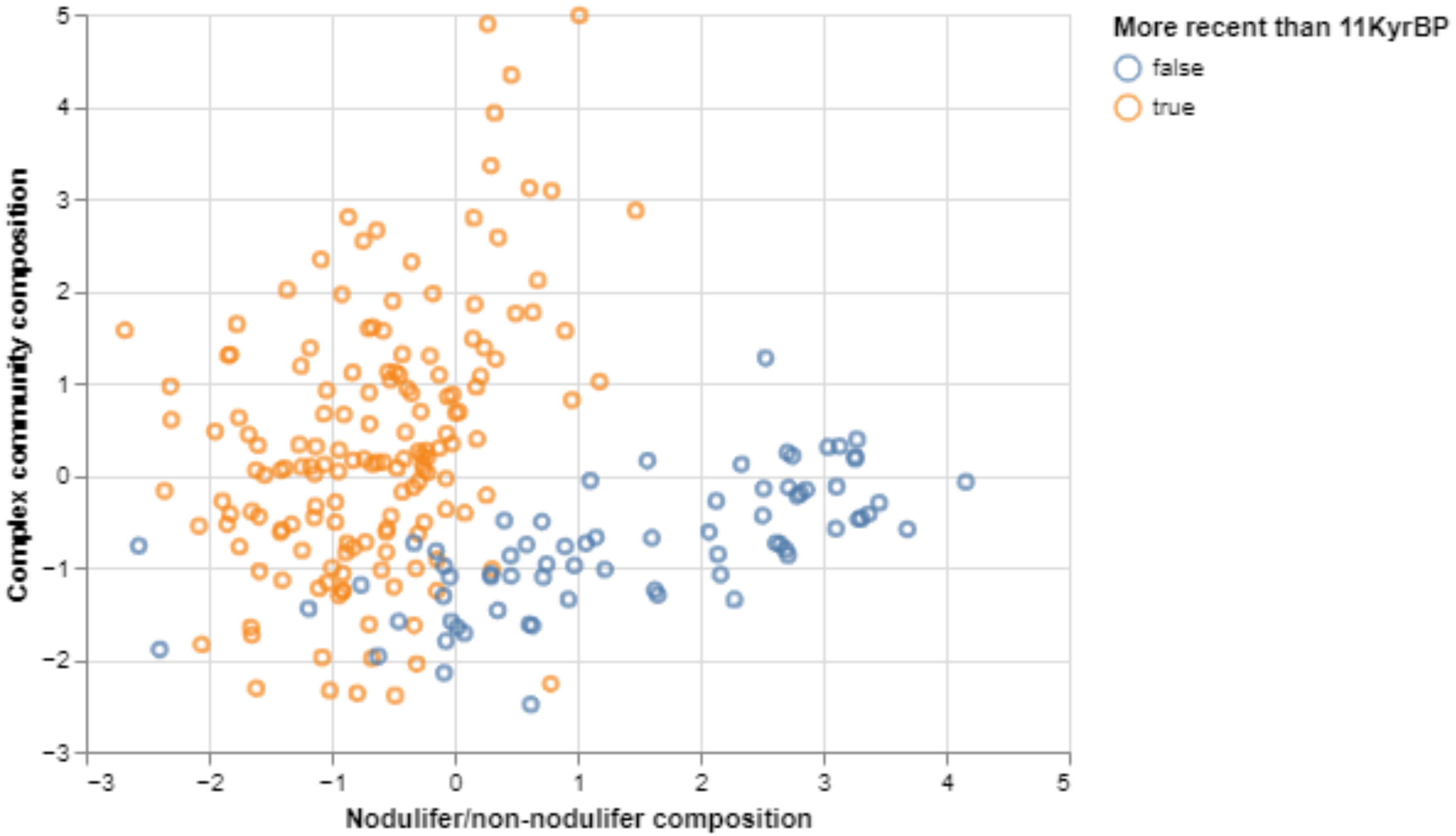
- merge test scores and socioeconomic indicators from the 2018 Standford Education Data Archive by school district
- visually assess correlations between gender achievement gaps among grade schoolers and socioeconomic indicators across school districts in CA

Case study 3: Paleoclimatology



Sea surface temperature reconstruction over the past 16,000 years.

Case study 3: Paleoclimatology



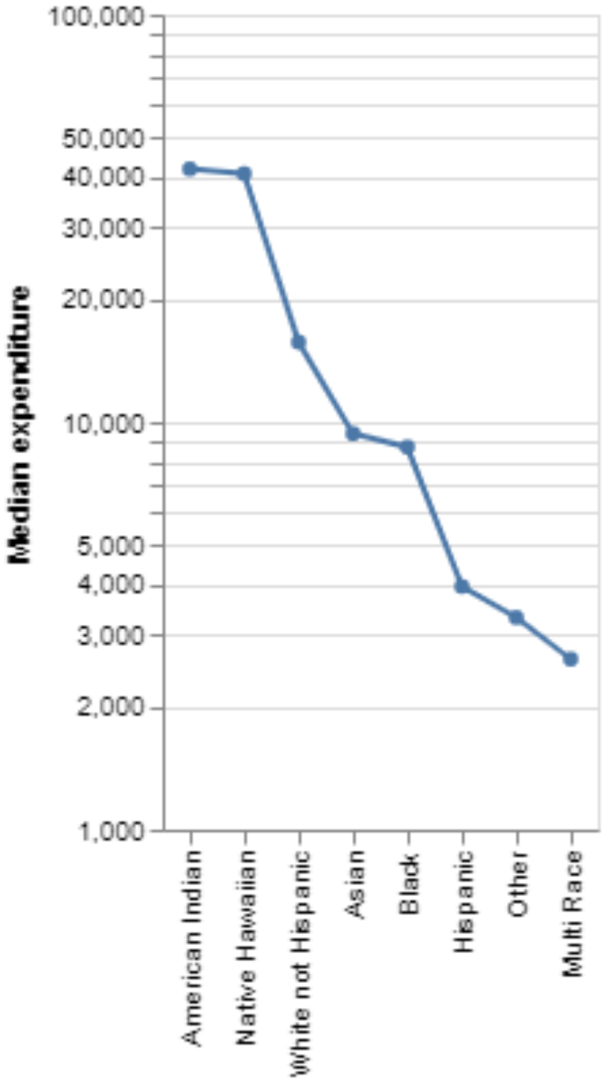
Clustering of diatom relative abundances in pleistocene (pre-11KyBP) vs. holocene (post-11KyBP) epochs.

Case study 3: Paleoclimatology

You will:

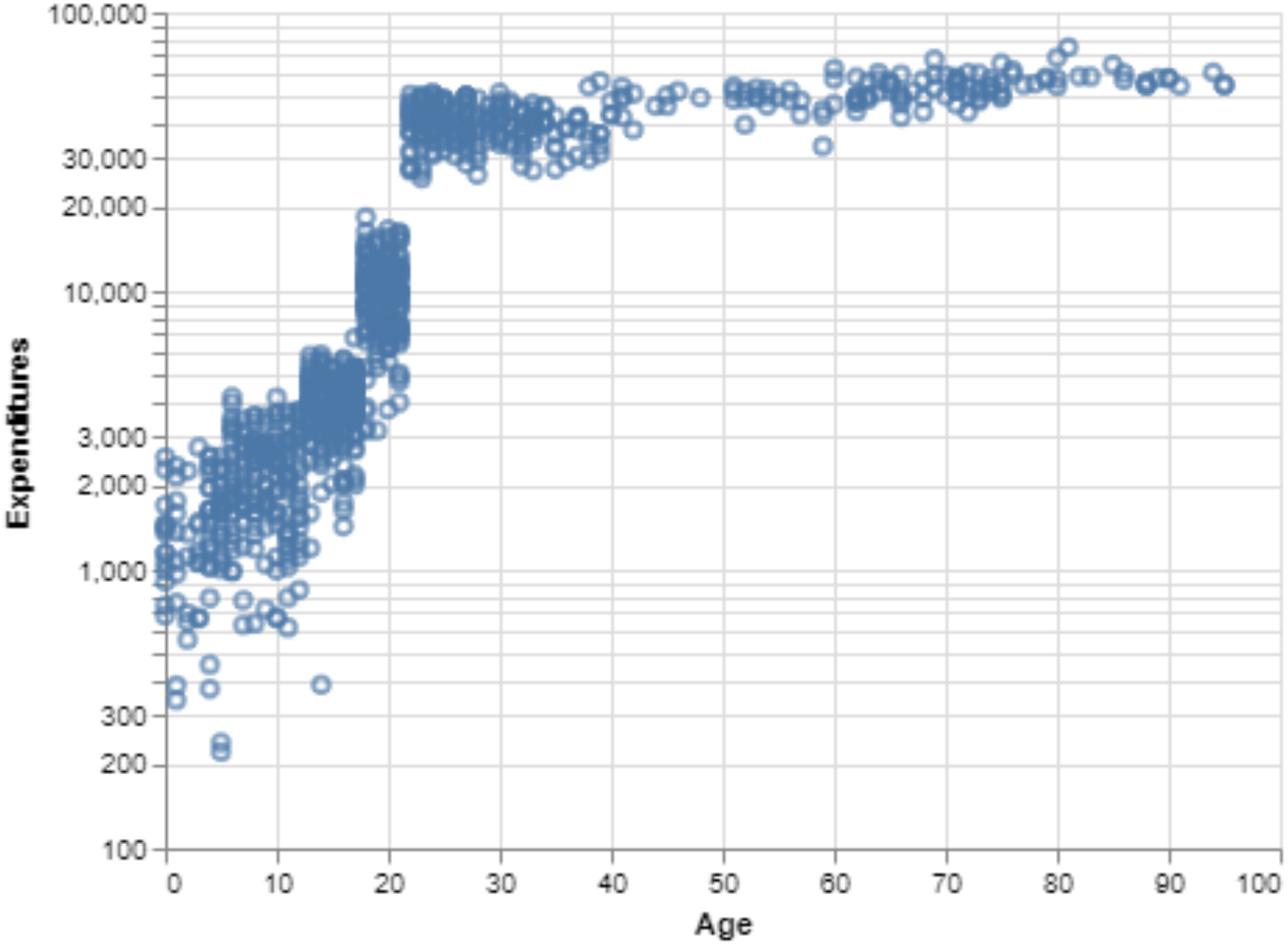
- explore ecological community structure from relative abundances of diatoms measured in ocean sediment core samples spanning ~15,000 years
- use dimension reduction techniques to obtain measures of community structure
- identify shifts associated with the transition from pleistocene to holocene epochs

Case study 4: Discrimination at DDS?



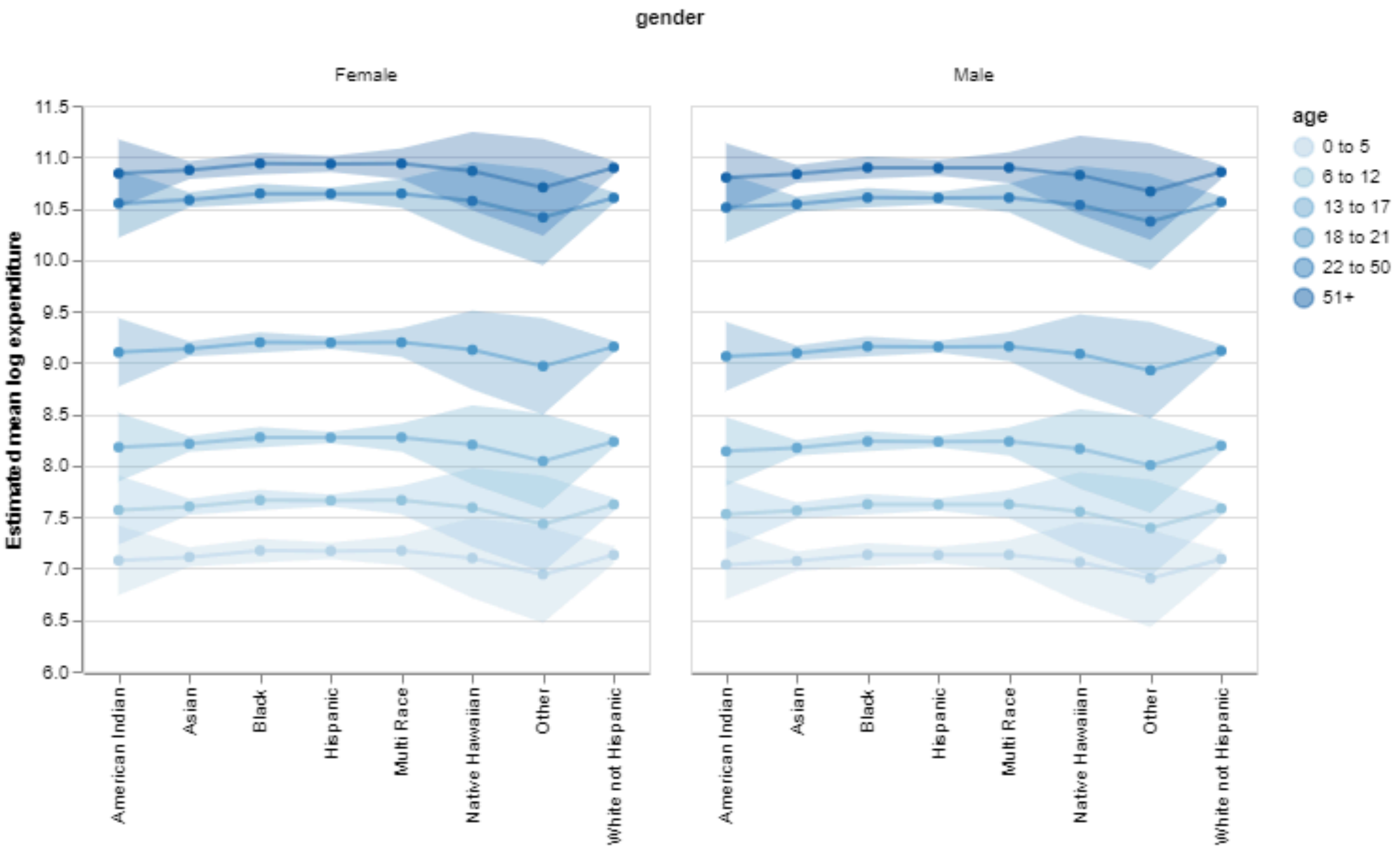
Apparent disparity in allocation of DDS benefits across racial groups.

Case study 4: Discrimination at DDS?



Expenditure is strongly associated with age.

Case study 4: Discrimination at DDS?



Correcting for age shows comparable expenditure across racial groups.

Case study 4: Discrimination at DDS?

You will:

- assess the case for discrimination in allocation of DDS benefits
- identify confounding factors present in the sample
- model median expenditure by racial group after correcting for age

About the course

Scope

This course is about developing your data science toolkit with foundational skills:

1. Core competency with R data science libraries
2. Critical thinking about data
3. Visualization and exploratory analysis
4. Application of statistical concepts and methods in practice
5. Communication and interpretation of results
6. Ethical data science

What’s unique about PSTAT100?

There are a few distinctive aspects:

- multiple end-to-end case studies
- question-driven rather than method-driven
- emphasis on project workflow
- data storytelling and communication

Limitations

There are also some things we probably *won't* cover:

- Predictive modeling or machine learning (PSTAT 131)
- Algorithm design and implementation (CS)
- Techniques and methods for big data (PSTAT 135)
- Theoretical basis for methods

Weekly Pattern

We’ll follow a simple weekly pattern:

- **Mondays**
  - Lecture
  - Assignments due 11:59pm PST
- **Tuesdays**
  - Section
- **Wednesdays**
  - Lecture
  - Late work due 11:59pm PST

Pages

Course page	Primary use
Canvas	Announcements and links to content
tinyurl.com/pstat100	Computing and distribution
pstat100.lsit.ucsb.edu	Computing

Tentative schedule

Week	Topic	Subjects	Lifecycle
1	Introduction	What’s data science?	
2	Tidy data	Import and organization	Collect/Acquiant/Tidy
3	Sampling	Informative vs. uninformative data	Collect/Acquaint/Tidy
4	Visualization	Plot types, aesthetics, principles	Explore
5	Exploratory analysis	Density estimation and descriptive statistics	Explore/Analyze

Tentative schedule

Week	Topic	Subjects	Lifecycle
6	Exploratory analysis	Dimension Reduction	Explore/Analyze
7	Regression and causality	Linear regression	Analyze/Interpret
8	Regression and causality	Non-linear models	Analyze/Interpret
9	Classification	Logistic regression etc	Analyze/Interpret
10	TBD	TBD	

The last week is flex time to explore other topics or extend coverage of previous topics.

Assessments

- **Labs** (20% final grade weight, 10 pts each)
    - Short/moderate-length guided programming assignments
    - Given weekly through week 8
    - Collaboration encouraged; individual submissions required

***Lab objective**: introduce and develop core skills with data science libraries in R.*

Assessments

- **Homeworks** (50% final grade weight, 50 pts each)
  - Applications of course ideas and lab skills to analyses of real datasets
  - 4 assignments, released/due biweekly
  - Collaboration encouraged and group submissions allowed

***Homework objective**: practice workflow and explore case studies.*

Assessments

- **Project** (30% final grade weight)
  - Open-ended data analysis based on your interests
  - final report due end-of-quarter
  - Collaboration expected

***Project objective:** apply learned skills to a problem of your choosing.*

Policies

- **Deadlines and late work**
  - One-hour grace period on all deadlines
  - One free late on any assignment (except final project report)
  - 75% partial credit thereafter
  - No late work beyond 72 hours after deadline without instructor permission

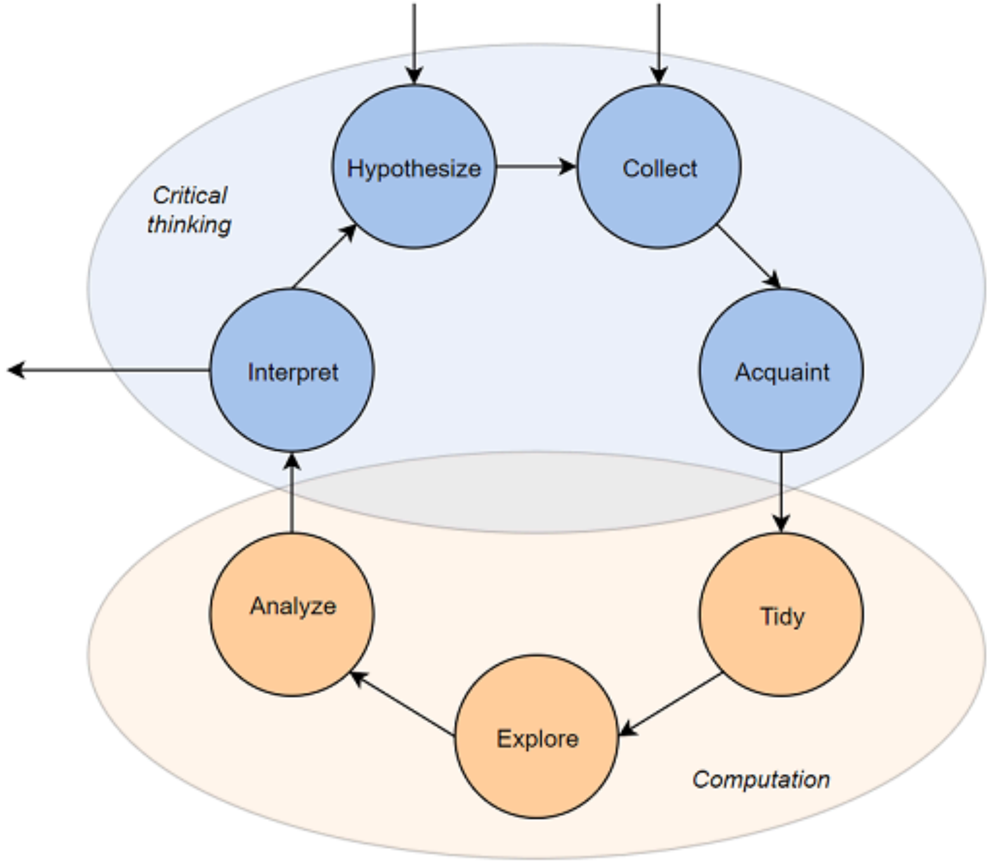
PSTAT 100 lifecycle

In this course, we’ll articulate the lifecycle in terms of the following steps.

0. **H**ypothesize: question formulation/refinement.
1. **C**ollect: go out and sample or acquire data ‘second-hand’.
2. **Ac**quaint: get to know your dataset; make friends!
3. **T**idy: clean up and organize your data.
4. **E**xplore: search for patterns and structure.
5. **An**alyze: seek to understand.
6. **I**nterpret: explain the meaning of your analysis.

PSTAT 100 lifecycle

No data science lifecycle would be complete without a flowchart!



Notice the multiple entry points – some projects start with a focused question; others, with a dataset.