Intro to Epi/Biostats

September 11, 2023

PBHLTH 198, Fall 2023 @ UC Berkeley

Andrew O'Connor

Class Outline

- Introductions

- Syllabus

- What is Public Health?

- Epidemiology vs. Biostatistics

- Activity: Exploring Careers in Public Health

- Data Analysis Frameworks

Andrew O'Connor

- He/Him/His
- 4th Year
- Majors: Public Health & Data Science
  - concentrations: biostatistics, human & population health
- Hometown: San Diego, CA
- Hobbies: studying foreign languages
- Research Interests: Indigenous & Rural Health, Transportation Planning, Spatial Analytics
- Fun Fact: I like kpop!

INTRODUCTIONS + ROSE/THORN

- Name & Pronouns
- Year
- Major
- Hometown
- Favorite place/spot in Berkeley
- Why are you taking this course?

ANNOUNCEMENTS

SYLLABUS

Syllabus

| Category	Percent of Grade	
Attendance	25%	
Worksheets/Labs	30%	Due by 5PM the next class
Pair Projects	25%	
Final Presentation	20%	Monday 11/27/23
Extra Credit	5%	
TOTAL	105% (possible)	

Public Health

Public health is "the science and art of **preventing** disease, **prolonging** life, and **promoting** health through the organized efforts and informed choices of society, organizations, public and private communities, and individuals." - CDC

Key Words:
- **Upstream**
- **Disease Prevention**
- **Primary Prevention**
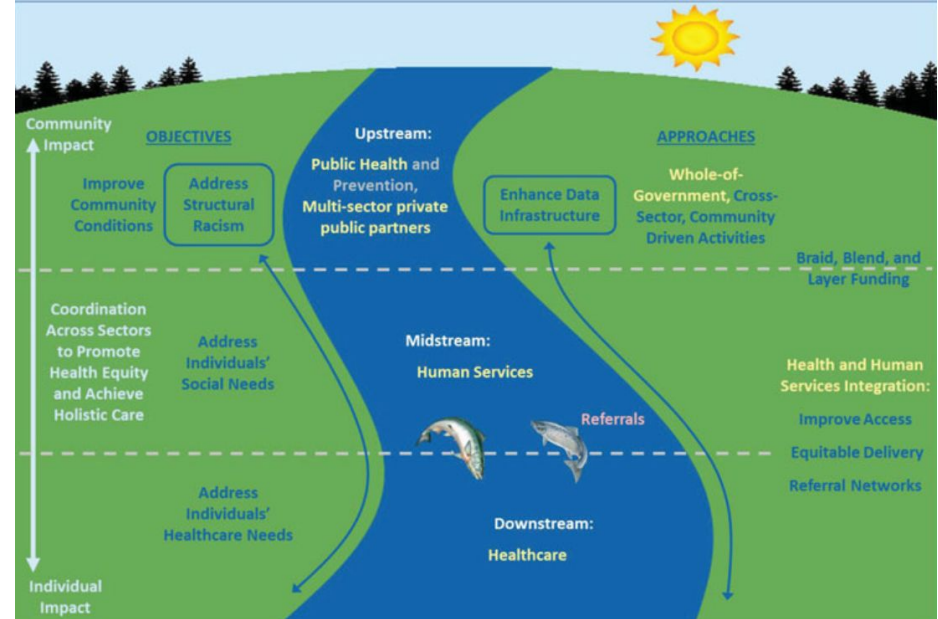
Public Health: How can we prevent people from going to the doctor's office? How can we encourage people to live healthy and fulfilling lives? How can we make this equitable for all?

Public Health Concentrations

Biostatistics	⊕
Community Health and Human Development	⊕
Environmental Health Sciences	⊕
Epidemiology	⊕
Health Policy and Management	⊕
Infectious Diseases	⊕

PBHLTH 142
− related to −
DATA C8/C100/C140
STAT 20/C88S
STAT 134/135

PBHLTH 150A

Epidemiology vs. Biostatistics

"Epidemiology is the study of the distribution and determinants of **health-related states or events in specified populations,** and the application of this study to the **control** of health problems" (CDC)

"Study the factors that **determine the distribution of health** and disease in human populations, and improve the **understanding of data** that are relevant to issues in public health." (Berkeley SPH)

Epidemiology: How can I study disease in relation to populations and design studies in ways that understand how, why and where the disease occurs?
Biostatistics: How can I analyze health data using advanced statistical methods to come up with valid conclusions?

Epidemiology vs. Biostatistics

The confluence of the modern epidemiologist
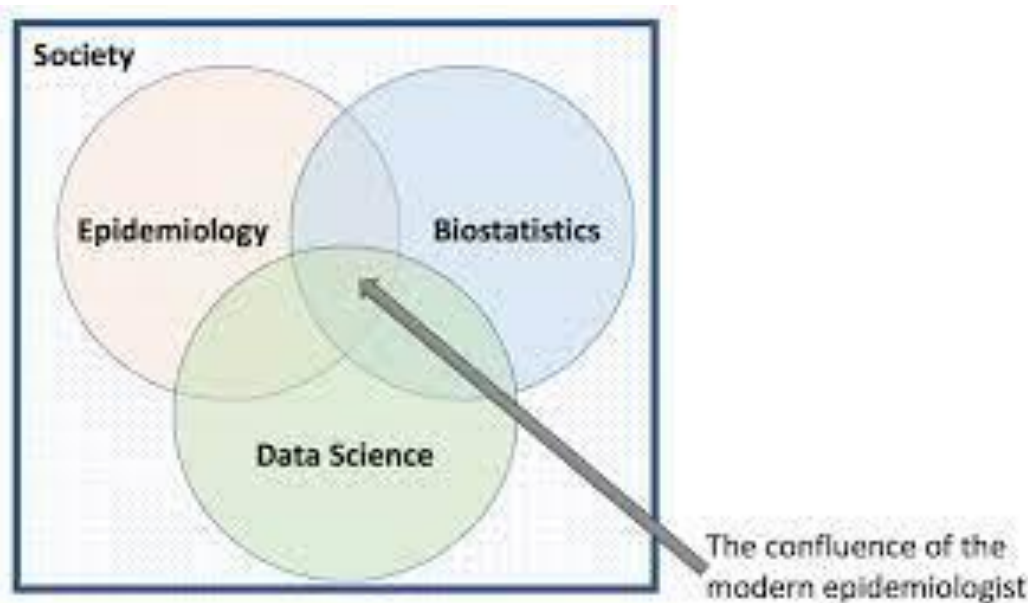
Epidemiology and Biostatistics are intimately related
Data Science and Computing are making their way into both fields

Career Paths

Bachelors → Masters → PhD → Work

Bachelors → Masters → Work

Bachelors → Work → Masters → Work → PhD

Bachelors → Work → Masters → Work

Masters, PhD degrees, and years of work experience are basically required to be a biostatisician or an epidemiologist. It's important to consider (regardless of any field) if pursuing a higher degree is worth it to you.

ACTIVITY 1
JOB SEARCHING

Data-Driven Decisions
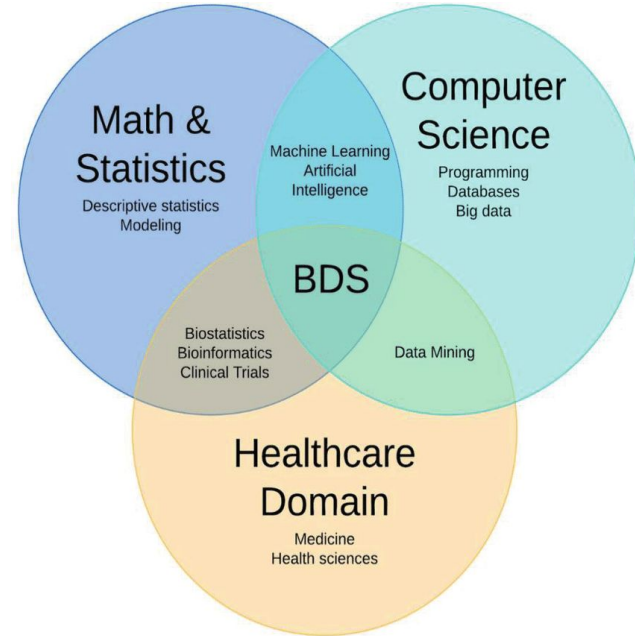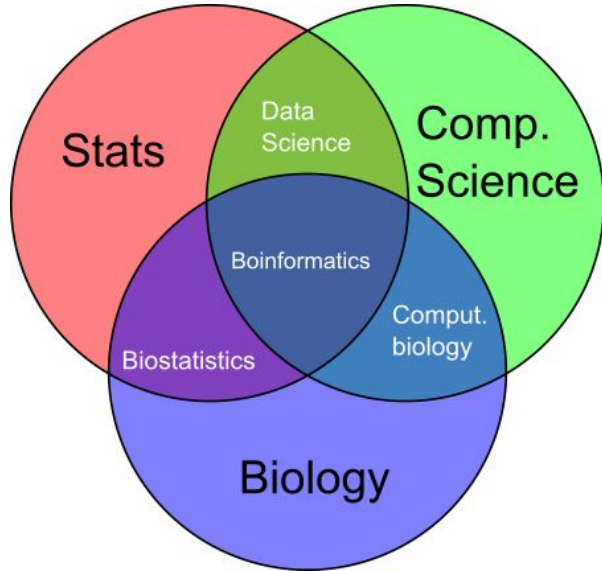
the reason why you see so much overlap with statistics and data science is because they're all quantitative fields. they all take different perspectives on certain problems but ultimately use the same statistical techniques to answer questions

Tools

the most valuable tools you can learn as an undergraduate
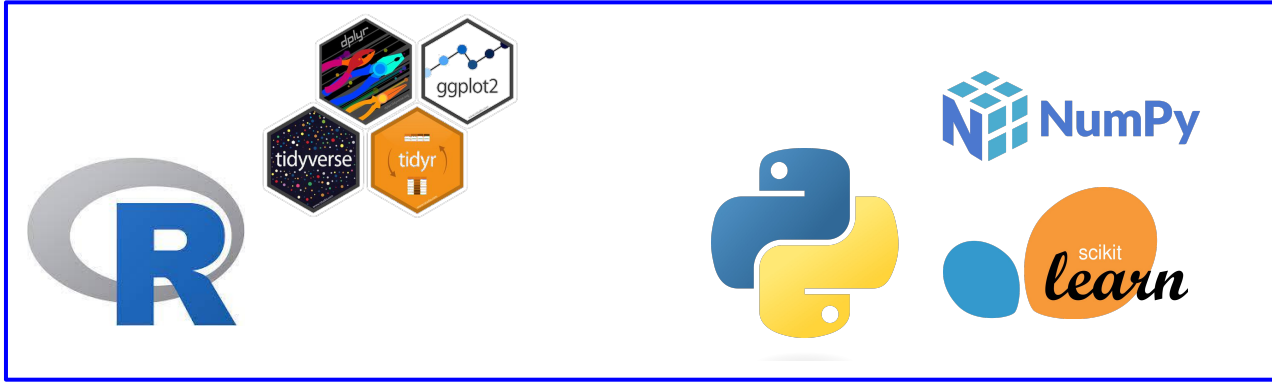
Different software exist for different types of analyses
Listed here: R, Python, SAS, SPSS, STATA, ArcGIS, QGIS, Excel, (and much more!)

Techniques
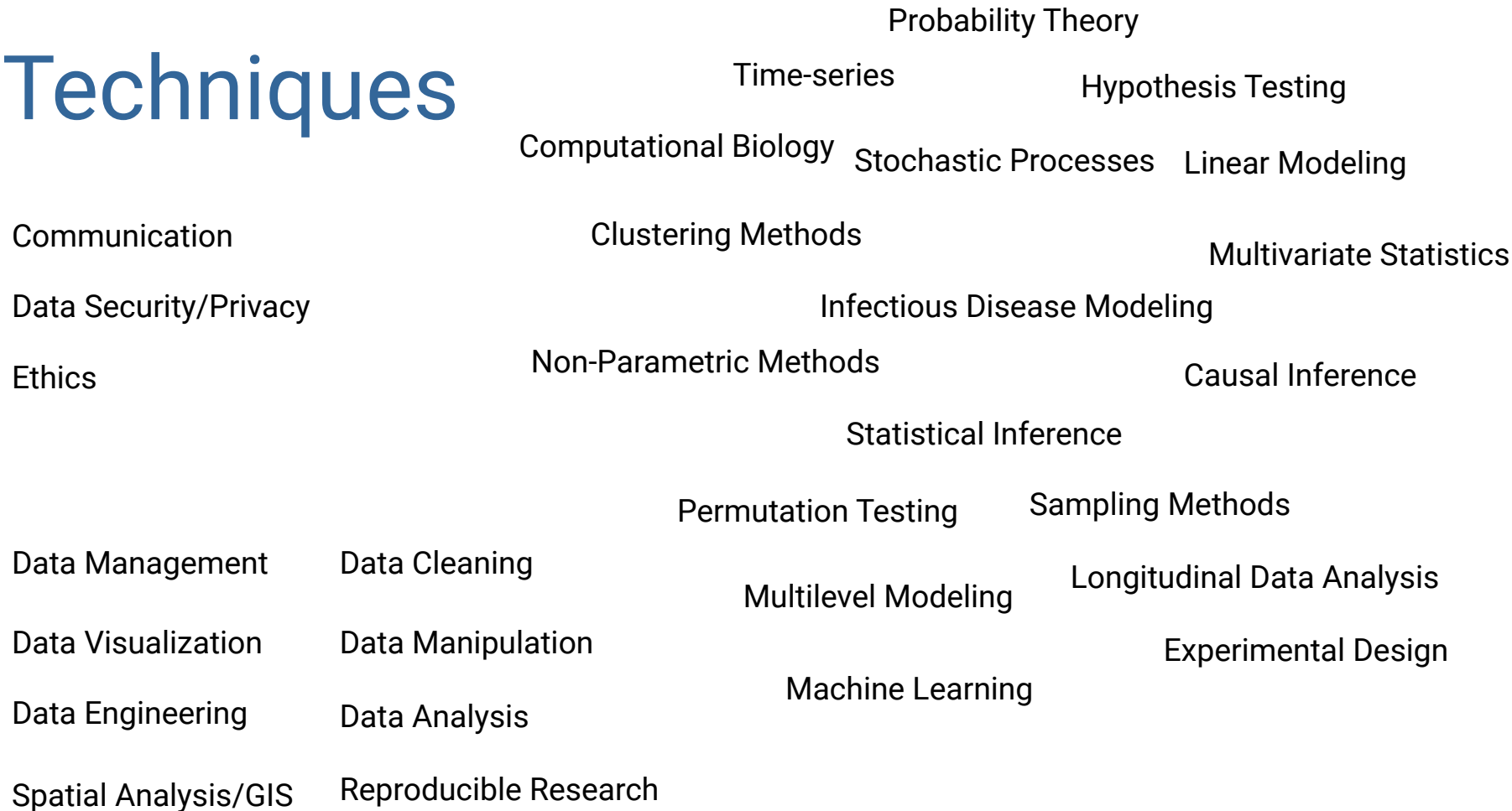
Probability Theory

Time-series

Hypothesis Testing

Computational Biology

Stochastic Processes

Linear Modeling

Communication

Clustering Methods

Multivariate Statistics

Data Security/Privacy

Infectious Disease Modeling

Ethics

Non-Parametric Methods

Causal Inference

Statistical Inference

Permutation Testing

Sampling Methods

Data Management

Data Cleaning

Longitudinal Data Analysis

Multilevel Modeling

Data Visualization

Data Manipulation

Experimental Design

Data Engineering

Data Analysis

Machine Learning

Spatial Analysis/GIS

Reproducible Research

Techniques

Probability Theory

Time-series

Hypothesis Testing

Computational Biology   Stochastic Processes   Linear Modeling

Communication   Clustering Methods   Multivariate Statistics

Data Security/Privacy   Infectious Disease Modeling

Ethics   Non-Parametric Methods   Causal Inference

Statistical Inference

Permutation Testing   Sampling Methods

Data Management   Data Cleaning

Multilevel Modeling   Longitudinal Data Analysis

Data Visualization   Data Manipulation

Experimental Design

Data Engineering   Data Analysis

Machine Learning

Spatial Analysis/GIS   Reproducible Research

the most valuable concepts and skills you can learn as an undergraduate

Topics Covered

Week 1: Intro

Week 2: Stats Fundamentals I

Week 3: Stats Fundamentals II

Week 4: Probability

Week 5: Epi I

Week 6: Epi II

Week 7: Stats Fundamentals IV

Week 8: ML I (Basic concepts)
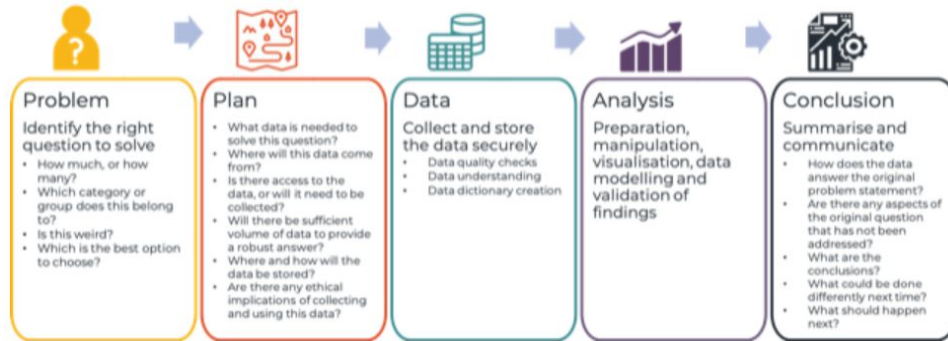
Week 9: ML II (Regression)

Week 10: ML III (Random Forests)
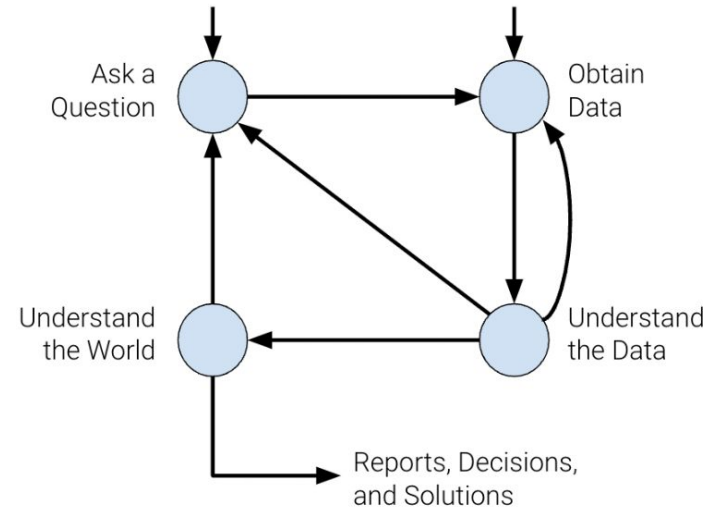
Week 11: Project Walkthrough

Week 12: Final Presentations
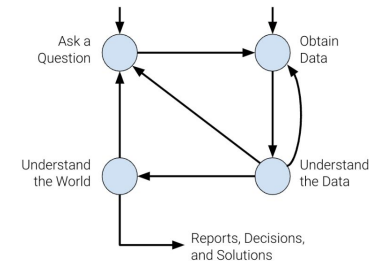
Data Analysis Framework

Public Health

Problem
Identify the right question to solve
- How much, or how many?
- Which category or group does this belong to?
- Is this weird?
- Which is the best option to choose?

Plan
- What data is needed to solve this question?
- Where will this data come from?
- Is there access to the data, or will it need to be collected?
- Will there be sufficient volume of data to provide a robust answer?
- Where and how will the data be stored?
- Are there any ethical implications of collecting and using this data?

Data
Collect and store the data securely
- Data quality checks
- Data understanding
- Data dictionary creation

Analysis
Preparation, manipulation, visualisation, data modelling and validation of findings

Conclusion
Summarise and communicate
- How does the data answer the original problem statement?
- Are there any aspects of the original question that has not been addressed?
- What are the conclusions?
- What could be done differently next time?
- What should happen next?

Data Science

Ask a Question

Obtain Data

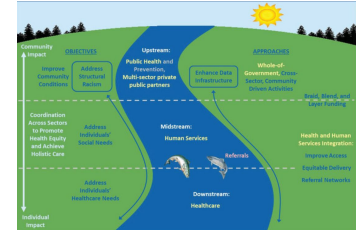Understand the World

Understand the Data

Reports, Decisions, and Solutions

The data science lifecycle is a high-level overview of the data science workflow. It's a cycle of stages that a data scientist should explore as they conduct a thorough analysis of a data-driven problem.

Recap

- **Public Health** is about disease prevention in order to prevent disease, prolong health, promote health
  - Many ways to do this; Can do it through health policy, community engagement, getting involved with politics, or **using statistics and epidemiological theory to help public officials make informed decisions to ultimately eliminate disease (this class!)**
- **Epidemiology** seeks to understand disease within populations and design studies to explore the how, why, and where of disease occurrence
- **Biostatistics** aims to analyze health data using advanced statistical methods to derive valid conclusions
- **Data analysis frameworks** provide a structured and efficient approach to data analysis, ensuring consistency and quality in results while facilitating collaboration and compliance with regulations regarding privacy
  - Examples: **PPDAC**, **Data Science Lifecycle** (we will focus on this more)

See you next week!