Cogs 9
Discussion Section

FA22 Week 2
Will McCarthy

If you do not have a group: please come and tell us now!

Upcoming due dates

Thursday Oct 6th (tomorrow!):

Quiz 1 (on reading 1)

Friday Oct 7th:

Group Assignment 1;

group signup google sheets (extra credit)

Groups

Does anyone *not* have a group?

Does any group have 3 or fewer members?

Has anyone not signed up on the sign up sheet?

In the meantime, talk to someone from outside your group about your data science question. Can you think of any ethical considerations that might arise in relation to their question?

Any questions about the course?

This week’s content

Forming a good data science question

Reading 1: 50 Years of Data Science, David Donoho (for **Quiz 1**)

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data’s Day of Reckoning

All (I hope) will help with **Assignment 1**.

Tips for good data science questions

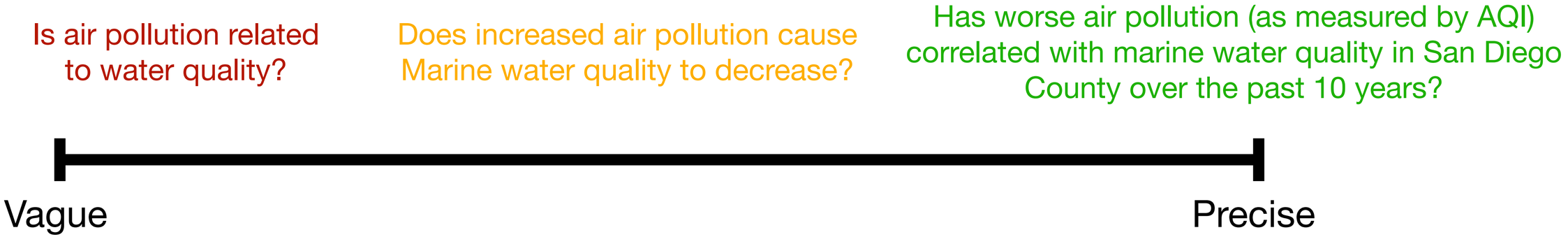
Questions should be *precise*

Why? Because vague questions lead to vague answers

If you are wrong in a quantifiable way then you learn

If you are right in a vague way you often don't

You can increase precision by *adding detail* to your question (measures, time, place)



Good questions are precise but have *generalizable* answers

Constraining a vague question or topic

Availability of data: a preliminary search can help you (part of assignment 1)

Existing measures: it's hard to come up with new measures

Stakeholder interest: pick things that you find interesting, but also pick things that other people would find interesting

Tip: look at every concept in your question and ask "can we make this more precise?"

Avoid: "How is x related to y?" "what can x tell me about y?" "What happens when x?" "What can we do with x?"

Reading 1: 50 Years of Data Science, David Donoho

What was the purpose of this paper?

Provide a historical account of data science (in relation to statistics)?

Define data science?

Presenting a case for a new science of data?

Describe what data science could be?

Provides a good answer to the question: *isn't data science just statistics?*

Four major influences act on data analysis today* (Tukey)

1. The formal theories of statistics
2. Accelerating developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever wider variety of disciplines

***1962!**

Tukey “called for a reformation” of academic statistics: a science of data analysis

Cleveland's 6 foci of activity (and allocations of effort)

Multidisciplinary investigations (25%)

Models and Methods for Data (20%)

Computing with Data (15%)

Pedagogy (15%)

Tool Evaluation (5%)

Theory (20%)

"An action plan to expand the technical areas of statistics focuses on the data analyst… . The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly."

Generative vs. Predictive cultures

Generative: understand the process giving rise to the data

Predictive: predict new data

Breiman says: stats should include both (not just generative) **Why?**

Common Task Framework

(a) A publicly available training dataset

(b) A set of enrolled competitors

(c) A scoring referee

“All the competitors share the *common task* of training a prediction rule which will receive a good score; hence the phase *common task framework*.”

What makes this the “secret sauce” of predictive culture?

Greater Data Science: 6 Divisions

Data Exploration and Preparation

Data Representation and Transformation

Computing with Data

Data Modeling

Data Visualization and Presentation

Science about Data Science

Similar to Cleveland's… **how do they differ?**

Assignment 1

Prep for final project! (unlike assignments 2, 3, 4 which are standalone)

First attempt at: formulating a data science question, doing some background research, finding a dataset

More about documenting your efforts than having something polished

Ethical considerations > 50% of points

Think *very carefully* about this part

Even If you think everything “should be fine”, you need to justify why.

Tip: don’t pick a project because you think ethical issues won’t apply. (Obviously don’t pick something unethical.. but if you have something to discuss about ethics it’s actually easier for you)

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data's Day of Reckoning

Paper structure

- Ethics and security training
- Developing guiding principles
- Building ethics into a data-driven culture
- Regulation
- Building our future

—

Ethics is compared with "Security" throughout— meaning cybersecurity

A lot of this is targeted towards companies that make "products"

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data's Day of Reckoning

Ethics and security training

"In many fields, ethics is an essential part of professional education. This isn't true in computer science, data science, artificial intelligence, or any related field."

(If you think CogSci or Psychology are related, this isn't true)

"Software and security ethics frequently go hand in hand…

and our current practices for teaching security provide an example of what not to do." (Elective, isolated from dev classes)

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data's Day of Reckoning

Developing guiding principles

Very useful for Assignment

(Remember, 20 points!)

Lots of questions won't immediately apply, but think of similar questions

- ❑ Have we listed how this technology can be attacked or abused?
- ❑ Have we tested our training data to ensure it is fair and representative?
- ❑ Have we studied and understood possible sources of bias in our data?
- ❑ Does our team reflect diversity of opinions, backgrounds, and kinds of thought?
- ❑ What kind of user consent do we need to collect and use the data?
- ❑ Do we have a mechanism for gathering consent from users?
- ❑ Have we explained clearly what users are consenting to?
- ❑ Do we have a mechanism for redress if people are harmed by the results?
- ❑ Can we shut down this software in production if it is behaving badly?
- ❑ Have we tested for fairness with respect to different user groups?
- ❑ Have we tested for disparate error rates among different user groups?
- ❑ Do we test and monitor for model drift to ensure our software remains fair over time?
- ❑ Do we have a plan to protect and secure user data?

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data's Day of Reckoning

Building ethics into a data-driven culture

Integrating ethics into corporate culture is harder than security. **Why?**

Reading 2(a): Loukides M, Mason, H & Patil DJ, Data's Day of Reckoning

Regulation

In some industries ethical standards that exist (e.g. in psych/cogsci)

Institutional Review Boards (IRBs)

In some, regulatory bodies enforce standards (NRC, FDA etc.)

"One challenge of developing a policy framework is that the policy development process nearly always lags behind the pace of innovation"

Group work / questions