ETHICAL STATISTICS AND DATA SCIENCE

Rochelle Tractenberg ~ Georgetown University
former vice chair & chair of Committee on Professional Ethics
Chair of Working Groups on ASA Ethical Guidelines revisions
2016; 2018; 2021 (co-chair)

UC Irvine Stats 5 Seminar in Data Science

26 January 2022

It is a pleasure to be here at UC Irvine (where I did my first PhD and Master's degrees!) to chat with you all about ethical practice of statistics and data science.

TAKE HOME MESSAGES

- "ASA ETHICAL GUIDELINES FOR STATISTICAL PRACTICE" WILL HELP ANY STATISTICIAN, DATA SCIENTIST, OR INDIVIDUAL USING STATISTICS OR DATA SCIENCE TO BE A *MORE ETHICAL* USER OF THE QUANTITATIVE SCIENCES;
- THE PURPOSE IS NOT TO CORRECT UNETHICAL HABITS, BUT TO INCULCATE A RESPECT FOR, AND AWARENESS OF, THE COMPLEXITIES OF ETHICAL STATISTICAL & DATA SCIENCE PRACTICE IN THE MODERN DATA ANALYSIS LANDSCAPE.

OUTLINE

- What are the ASA Ethical Guidelines for Statistical Practice?
- Why should I bother?
- What is "the Statistics & data science pipeline"?
  - How do ethics pertain along the statistics & data science pipeline?
- DISCLOSURE: these materials & activities are featured in two books I've written & am in the process of getting published.

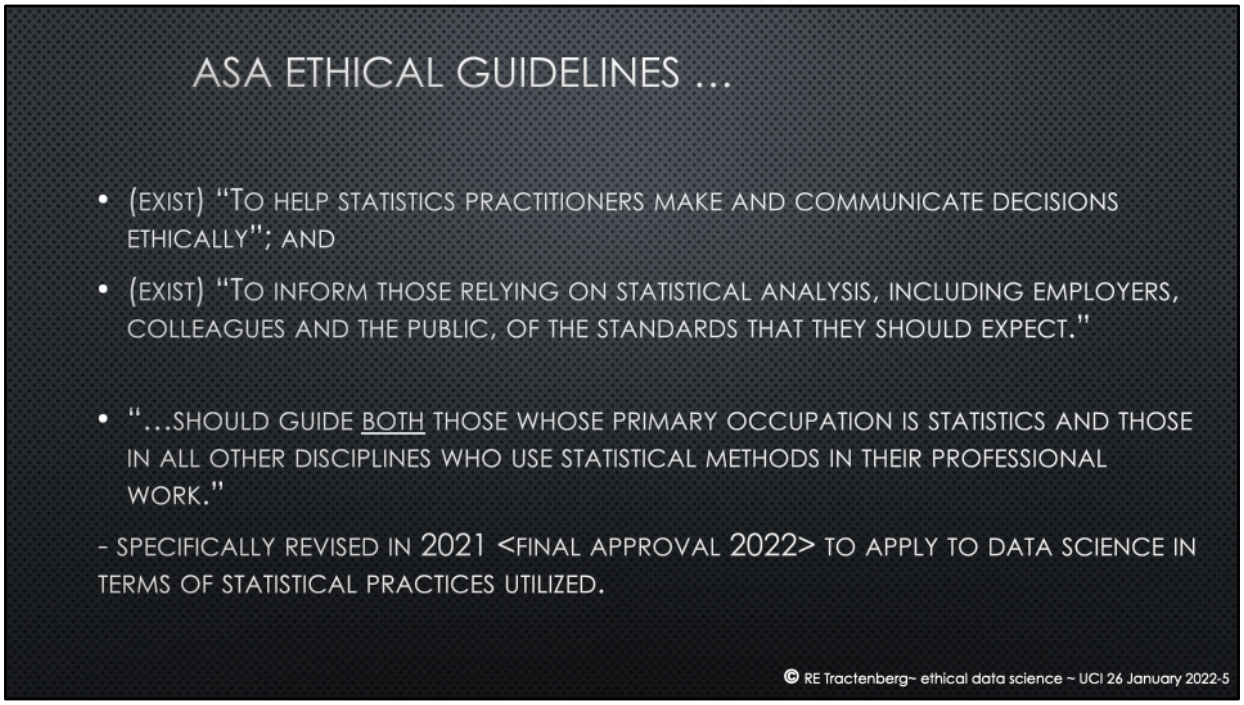
The ASA Ethical Guidelines have been revised for 2022. Ethical statistics and data science are *dynamic*, not fixed.

https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

The eight principles of the ASA Ethical Guidelines for Statistical Practice (A-H) have 60 elements, plus 12 additional ones in the Appendix**– everything in this talk applies to the ASA Ethical Guidelines from 2016, 2018, and 2022, no matter what year you're looking!! The specific examples are from 2022.**

Principle B outlines the practitioner's responsibilities relating to data – Big or Small. Since data science involves data, as well as analysis (methods), these Guidelines are applicable to both statistics and data science.

All of the NIH topics for responsible conduct of research are also addressed in the ASA Ethical Guidelines – but, the Guidelines are much more relevant to all practice with data. This talk is intended to show how that is true.



ASA Ethical Guidelines for Statistical Practice are NOT JUST FOR STATISTICIANS!! Instead, they "…should guide both those whose primary occupation is statistics and those in all other disciplines who use statistical methods in their professional work."

WHY SHOULD I BOTHER?

OK, ETHICAL GUIDELINES EXIST AND ARE UPDATED… SO WHAT?

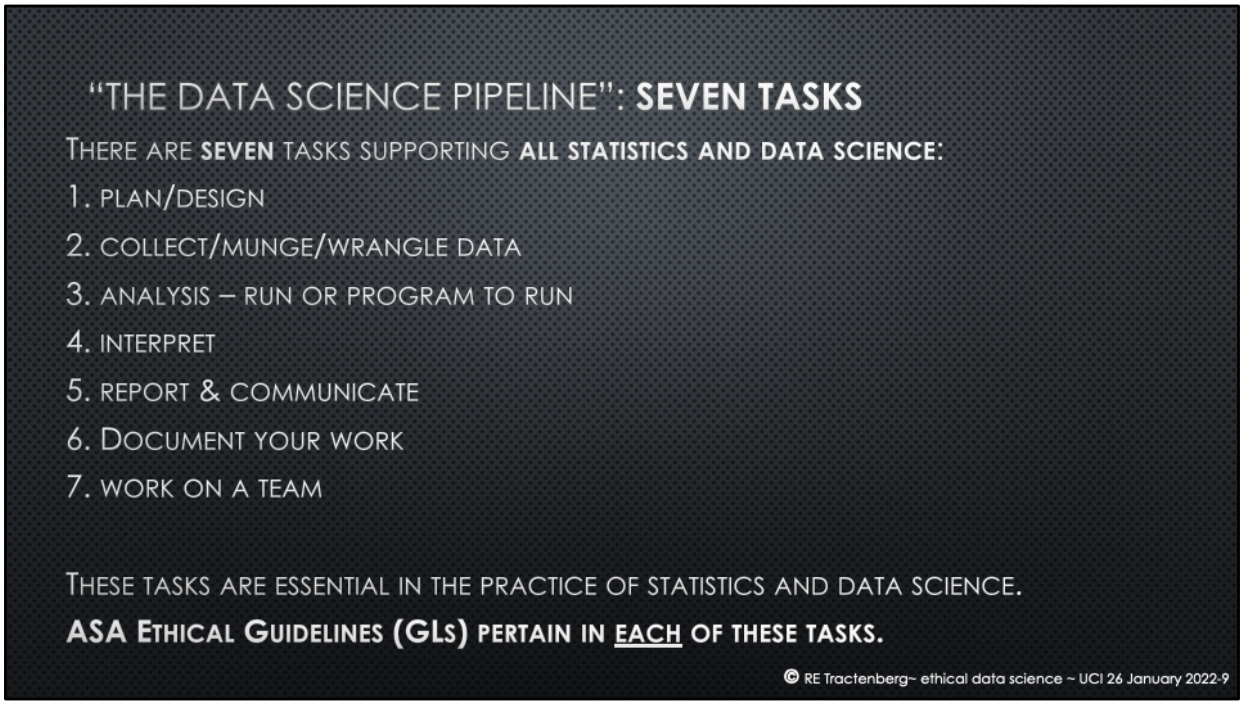
IN THE NEWS (2018) "CAMBRIDGE ANALYTICA" (CA)

- CAMBRIDGE ANALYTICA (CA) SNUCK AN ALGORITHM INTO FACEBOOK TO STEAL DATA – EVEN FROM THOSE WHO SPECIFICALLY DID NOT WANT THEIR DATA SHARED.
- STEALING DATA IS **ILLEGAL, NOT (ONLY) UNETHICAL.** IMPORTANTLY, A LOT OF OTHER UNETHICAL DECISIONS HAPPENED, NOT A SINGLE ONE "TO STEAL DATA".
- APPLYING THE ASA GUIDELINES **MIGHT HAVE STOPPED THIS THEFT AND SUBSEQUENT SCANDAL AT *7 DIFFERENT POINTS*.**
- THE MORE PEOPLE KNOW AND UNDERSTAND THEIR RESPONSIBILITIES, THE MORE LIKELY IT IS THAT ILLEGAL – AND UNETHICAL – ACTIVITIES CAN BE STOPPED!

SEVEN POINTS TO HAVE STOPPED THE CA SCANDAL

Plan/Design	Collect	Analyze	Interpret	Document	Report	Team
You are asked to design a system that scrapes data automatically from Facebook.	You build a system that scrapes Facebook only AFTER the user opts-in –prior to every scrape. However, your opt-in mechanism is removed.	Piles of data begin to arrive from "the company data scraper" for you to analyze. No consent info is available. You are told that "consent" is personal data – so, not available.	Results suggest a strong likelihood for bias in the results, but leaders interpret the results as "proof they can change how people think".	You are told not to document your system, the changes, or your suggestions to ensure valid results/decisions based on the data/analyses. "Too time consuming".	You submit your fully documented report. You later discover that none of your content made it into the "final" report, which is claimed to be "complete".	Leadership informs your team that they bought an algorithm you will be "using" –first they want you to remove all the consent pop ups ("they will ruin the user experience")

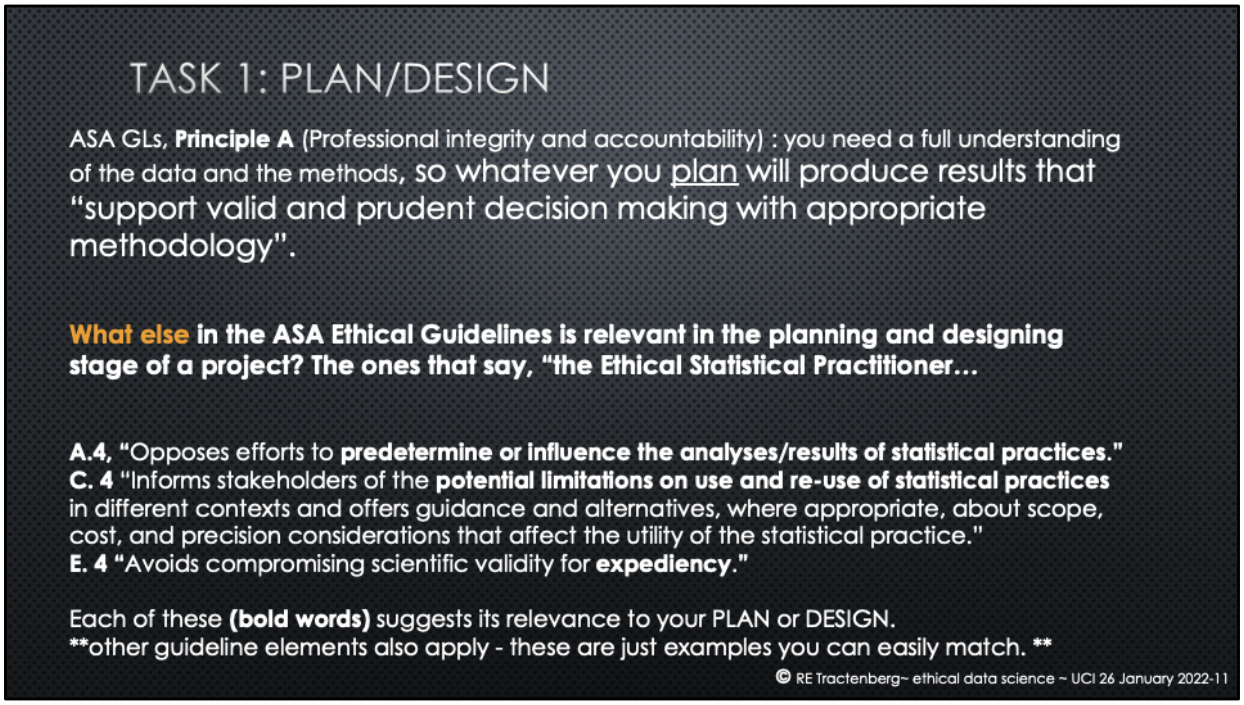
- WHAT DECISIONS COULD HAVE BEEN MADE (INSTEAD OF "DOING NOTHING") AT EACH OF THESE STEPS?
- BREAKING DOWN "WORK" INTO THESE SEVEN TASKS MAKES IT MORE STRAIGHTFORWARD TO IDENTIFY "WRONG" OR POTENTIALLY UNETHICAL BEHAVIORS BY OTHERS – AND AVOID THEM YOURSELF!
- **ASA GLS HELP YOU IDENTIFY WHEN TO MAKE DECISIONS IF YOU THINK "SOMETHING IS WRONG"**– AS WELL AS WHAT TO DO, AND WHY.



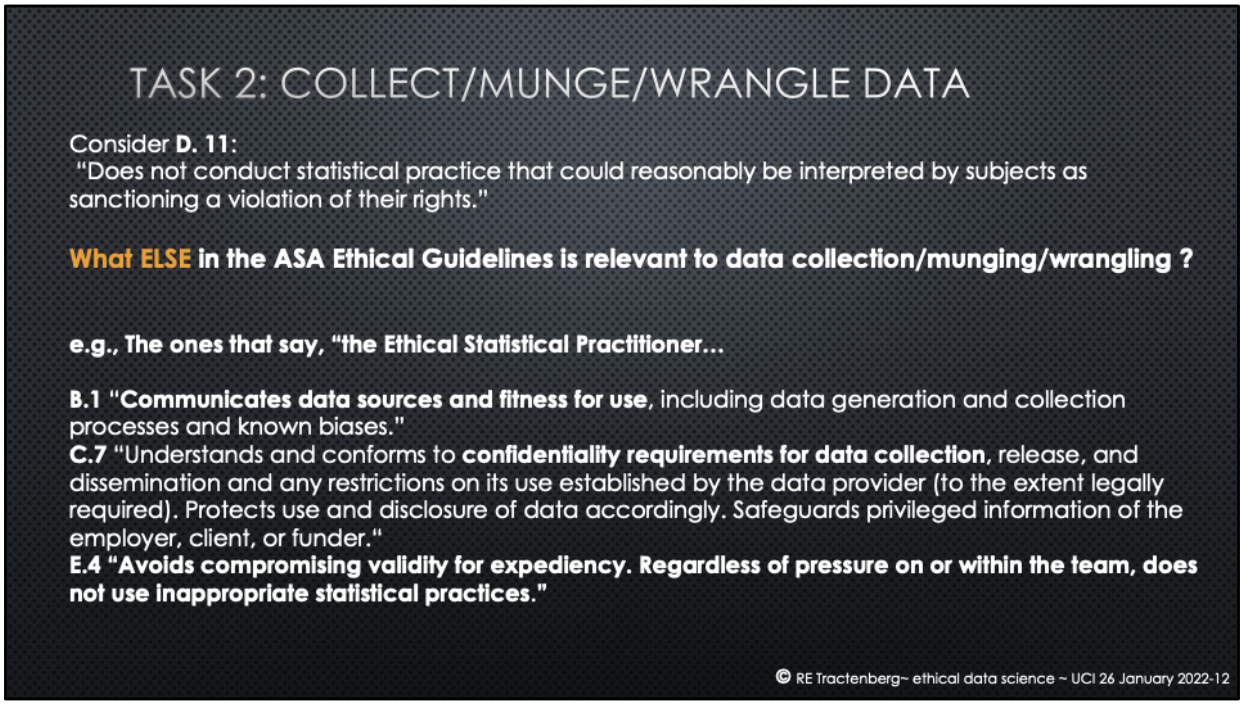
Note that the NIH responsible conduct of research (RCR) topics may be relevant in each (or any) of these tasks. Each task can (should) be done ethically (responsibly).

ETHICAL *PRACTICE*
MATCHING ASA GLS TO TASK

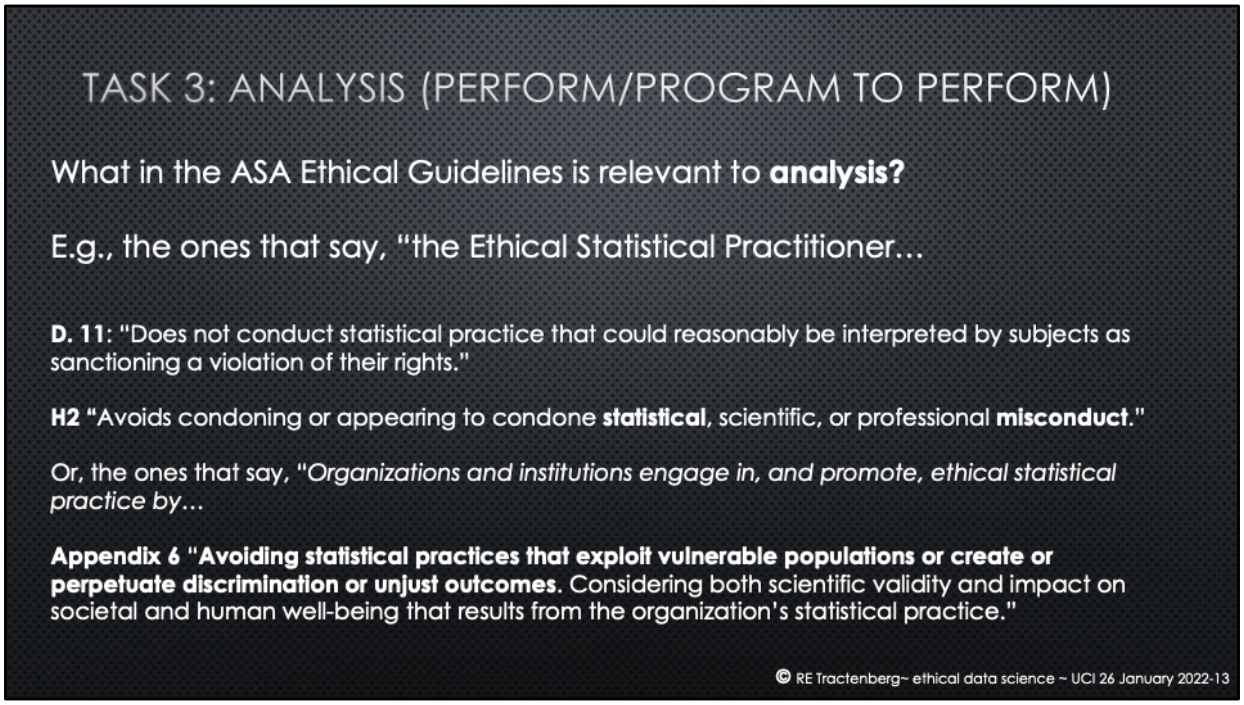
STEP 1: SCAN THE GUIDELINES TO IDENTIFY WHAT IS RELEVANT TO YOUR TASK!



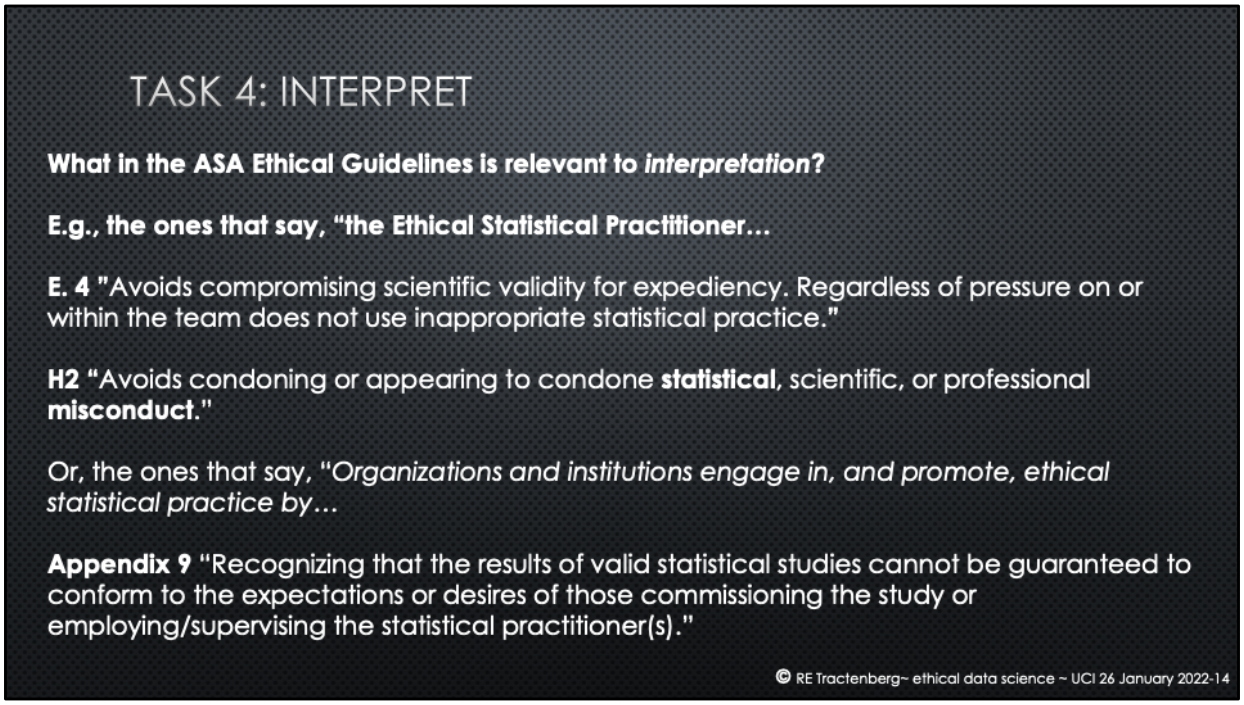
The 2022 ASA Ethical Guidelines (GLs) have many elements that are relevant to support ethical statistical practice at the plan/design phase. This is true whether you're planning an experiment or clinical trial, or planning to scrape data.



Although Principle D is all about Responsibilities to data contributors and those affected by statistical practices, many other principles pertain to the collection, munging, and wrangling of data.



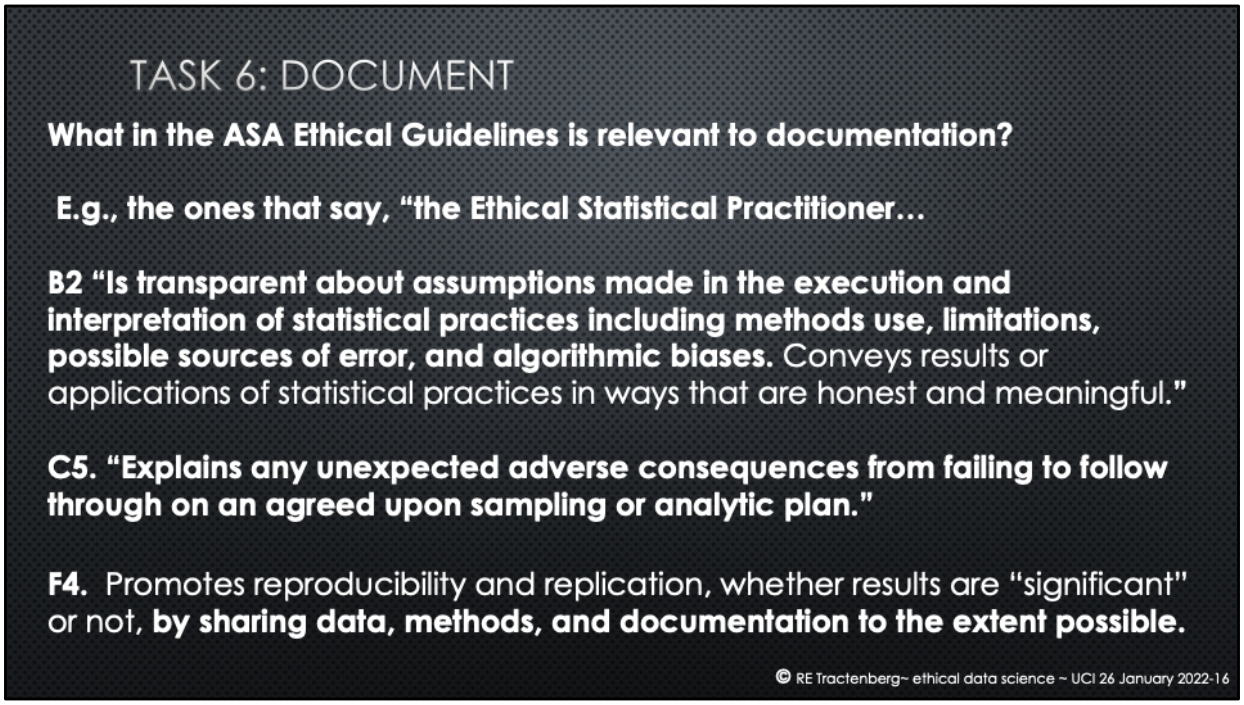
The most obvious matches to Analysis are in Principles B (methods & data) and A (integrity and accountability as a practitioner using statistics or data science). But those are not the only relevant guidance, as you can see here. D11 (Principle D: Responsibilities to Research Subjects, Data Subjects, or those directly affected by statistical practices) and H2 (Principle H: Responsibilities regarding potential misconduct) are also relevant. Analysis is one of the key "statistical practices" of the statistician and the data scientist, so a lot of the Guideline Principles can pertain to the task of analysis.



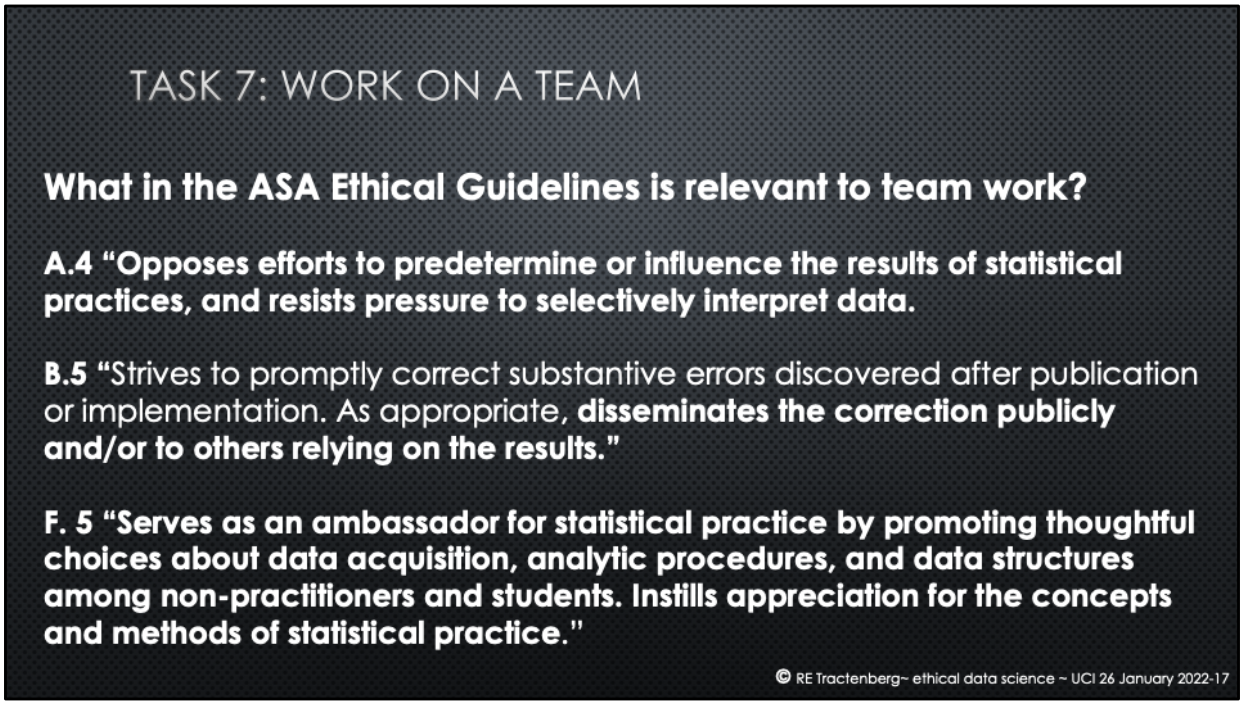
Interpretation is clearly part of statistical practice, and is sometimes part of data science (depends on the definition of data science you use, and where you are in a project). The simple visual match/search for the word "interpret" in the Guidelines won't identify all – or even many - of the relevant Guideline elements, though. The elements from E (responsibilities to research team colleagues), H (responsibilities regarding allegations of misconduct), and the Appendix (responsibilities of organizations/institutions) are important for this task – even though they are not explicitly about interpretation. When results are cherry picked, or p-hacked, interpretations will not be reproducible or valid. Ethical practice requires that we avoid compromising scientific validity for expediency (E4). This is true even if over-interpreting, or failing to appropriately describe the uncertainty of an interpretation of analyses, will strengthen a grant application or paper, the ethical practitioner. In every case, the ethical practitioner "Ensures all discussion and reporting of statistical design and analysis is consistent with these Guidelines." (E3)

It is also important to notice that "statistical practice" is not limited to designing/planning the data collection, collecting/munging the data, and its analysis.

Communication is a key component of research – but whenever the statistical practitioner or data scientist completes an analysis project, the results need to be communicated. All practitioners must be transparent, to promote understanding of what was done and the strengths and limitations of the results. Hint: In addition to obvious cues like "report" and "communicate", wherever the word "transparency" appears, it is a signal that communication is required, and that communication should be transparent (honest).



In addition to communicating and reporting, practitioners must document what they did – including assumptions that were made – so that others can appreciate the evidence that results from statistical and data science practice. Sometimes documentation remains internal (and a "report" or "communication" (paper, poster, etc.) to the outside/external stakeholders is a distinct activity). In many cases, the documentation is the most detailed description of your work, while a report may be more selective in terms of methods and results, providing just a summary of that effort. Documentation should be sufficient for replication of what you did.



Most of the Guidelines are relevant for team work – since many statisticians and data scientists work with other people from different –or the same- disciplines on any given project. Many of the 6 tasks we've discussed involve team work/team members. There is an entire Principle (E) dedicated to Responsibilities to Members of Interdisciplinary Teams. Transparency, and respect for human and animal data contributors as well as others on the team, are fundamental to ethical practice in statistics and data science.

However, sometimes you are on a team of people with your same training. In these cases, Principle F (Responsibilities to Fellow Statistics Practitioners and the Profession) are also relevant.

Note that the ethical practitioner has responsibilities beyond simply being competent at their job (Principle A1 – one of 67 items).

SEVEN WAYS YOU COULD NOW STOP THE NEXT CA SCANDAL

Plan/Design	Collect	Analyze	Interpret	Document	Report	Team
You are asked to design a system that scrapes data automatically from Facebook.	You build a system that scrapes Facebook only AFTER the user opts-in –prior to every scrape. However, your opt-in mechanism is removed.	Piles of data begin to arrive from "the company data scraper" for you to analyze. No consent info is available. You are told that "consent" is personal data – so, not available.	Results suggest a strong likelihood for bias in the results, but leaders interpret the results as "proof they can change how people think".	You are told not to document your system, the changes, or your suggestions to ensure valid results/decisions based on the data/analyses. "Too time consuming".	You submit your fully documented report. You later discover that none of your content made it into the "final" report, which is claimed to be "complete".	Leadership informs your team that they bought an algorithm you will be "using" –first they want you to remove all the consent pop ups ("they will ruin the user experience")

- EACH TASK IS AN OPPORTUNITY FOR YOU TO PRACTICE ETHICALLY.
- IGNORING AN UNETHICAL (OR POTENTIALLY) UNETHICAL REQUEST OR ACT IS NOT ETHICAL. BUT IT IS A DECISION. THE GLS EXIST TO HELP YOU MAKE THESE, AND OTHER, DECISIONS ETHICALLY.
- YOU SHOULD BE ABLE TO UTILIZE THE GLS NOW TO EXAMINE YOUR- AND OTHERS' – DECISIONS ON EACH TASK!

SUMMARY

ASA Ethical Guidelines for Statistical Practice (**GL**) are relevant for all of the tasks in the DS Pipeline.

The GLs support ethical practice of both statistics AND data science.

Many GLs are relevant for >1 task
All tasks have support from elements from more than one GL Principle.

Noticing an unethical assignment or act could possibly prevent illegal behaviors (as well as additional unethical ones).