

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 ...

- (EXIST) "TO HELP STATISTICS PRACTITIONERS MAKE AND COMMUNICATE DECISIONS ETHICALLY"; AND
 - (EXIST) "TO INFORM THOSE RELYING ON STATISTICAL ANALYSIS, INCLUDING EMPLOYERS, COLLEAGUES AND THE PUBLIC, OF THE STANDARDS THAT THEY SHOULD EXPECT."
 - "...SHOULD GUIDE BOTH THOSE WHOSE PRIMARY OCCUPATION IS STATISTICS AND THOSE IN ALL OTHER DISCIPLINES WHO USE STATISTICAL METHODS IN THEIR PROFESSIONAL WORK."
- SPECIFICALLY REVISED IN 2021 <FINAL APPROVAL 2022> TO APPLY TO DATA SCIENCE IN TERMS OF STATISTICAL PRACTICES UTILIZED.

© RE Tractenberg- ethical data science - UCI 26 January 2022-5

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.

"THE DATA SCIENCE PIPELINE": SEVEN TASKS

THERE ARE **SEVEN** TASKS SUPPORTING **ALL STATISTICS AND DATA SCIENCE**:

1. PLAN/DESIGN
2. COLLECT/MUNGE/WRANGLE DATA
3. ANALYSIS – RUN OR PROGRAM TO RUN
4. INTERPRET
5. REPORT & COMMUNICATE
6. DOCUMENT YOUR WORK
7. WORK ON A TEAM

THESE TASKS ARE ESSENTIAL IN THE PRACTICE OF STATISTICS AND DATA SCIENCE.

ASA ETHICAL GUIDELINES (GLs) PERTAIN IN EACH OF THESE TASKS.

© RE Tractenberg- ethical data science - UCI 26 January 2022-9

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!

TASK 1: PLAN/DESIGN

ASA GLs, **Principle A** (Professional integrity and accountability) : you need a full understanding of the data and the methods, so whatever you plan will produce results that "support valid and prudent decision making with appropriate methodology".

What else in the ASA Ethical Guidelines is relevant in the planning and designing stage of a project? The ones that say, "the Ethical Statistical Practitioner..."

A.4, "Opposes efforts to **predetermine or influence the analyses/results of statistical practices.**"

C. 4 "Informs stakeholders of the **potential limitations on use and re-use of statistical practices** in different contexts and offers guidance and alternatives, where appropriate, about scope, cost, and precision considerations that affect the utility of the statistical practice."

E. 4 "Avoids compromising scientific validity for **expediency.**"

Each of these (**bold words**) suggests its relevance to your PLAN or DESIGN.

**other guideline elements also apply - these are just examples you can easily match. **

© RE Tractenberg - ethical data science - UCI 26 January 2022-11

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.

TASK 2: COLLECT/MUNGE/WRANGLE DATA

Consider D. 11:

"Does not conduct statistical practice that could reasonably be interpreted by subjects as sanctioning a violation of their rights."

What ELSE in the ASA Ethical Guidelines is relevant to data collection/munging/wrangling ?

e.g., The ones that say, "the Ethical Statistical Practitioner..."

B.1 "Communicates data sources and fitness for use, including data generation and collection processes and known biases."

C.7 "Understands and conforms to confidentiality requirements for data collection, release, and dissemination and any restrictions on its use established by the data provider (to the extent legally required). Protects use and disclosure of data accordingly. Safeguards privileged information of the employer, client, or funder."

E.4 "Avoids compromising validity for expediency. Regardless of pressure on or within the team, does not use inappropriate statistical practices."

© RE Trachtenberg - ethical data science - UCI 26 January 2022-12

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.

TASK 3: ANALYSIS (PERFORM/PROGRAM TO PERFORM)

What in the ASA Ethical Guidelines is relevant to **analysis**?

E.g., the ones that say, "the Ethical Statistical Practitioner..."

D. 11: "Does not conduct statistical practice that could reasonably be interpreted by subjects as sanctioning a violation of their rights."

H2 "Avoids condoning or appearing to condone **statistical**, scientific, or professional **misconduct**."

Or, the ones that say, "*Organizations and institutions engage in, and promote, ethical statistical practice by...*"

Appendix 6 "Avoiding statistical practices that exploit vulnerable populations or create or perpetuate discrimination or unjust outcomes. Considering both scientific validity and impact on societal and human well-being that results from the organization's statistical practice."

© RE Tractenberg—ethical data science ~ UCI 26 January 2022-13

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.

TASK 4: INTERPRET

What in the ASA Ethical Guidelines is relevant to *interpretation*?

E.g., the ones that say, "the Ethical Statistical Practitioner..."

E. 4 "Avoids compromising scientific validity for expediency. Regardless of pressure on or within the team does not use inappropriate statistical practice."

H2 "Avoids condoning or appearing to condone **statistical**, scientific, or professional **misconduct**."

Or, the ones that say, "*Organizations and institutions engage in, and promote, ethical statistical practice by...*"

Appendix 9 "Recognizing that the results of valid statistical studies cannot be guaranteed to conform to the expectations or desires of those commissioning the study or employing/supervising the statistical practitioner(s)."

© RE Tractenberg—ethical data science ~ UCI 26 January 2022-14

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.

TASK 5: REPORT AND COMMUNICATE

What in the ASA Ethical Guidelines is relevant to reporting and communication?

E.g., the ones that say, "the Ethical Statistical Practitioner..."

B.1 "Communicates data sources and fitness for use, including data generation and collection processes and known biases."

C2. "Regardless of personal or institutional interests or external pressures, **does not use statistical practices to mislead any stakeholder."**

E 3 "Ensures all communications about statistical practices are consistent with these guidelines. Promotes transparency in all statistical practices."

© RE Tractenberg- ethical data science - UCI 26 January 2022-15

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).

TASK 6: DOCUMENT

What in the ASA Ethical Guidelines is relevant to documentation?

E.g., the ones that say, "the Ethical Statistical Practitioner...

B2 "Is transparent about assumptions made in the execution and interpretation of statistical practices including methods use, limitations, possible sources of error, and algorithmic biases. Conveys results or applications of statistical practices in ways that are honest and meaningful."

C5. "Explains any unexpected adverse consequences from failing to follow through on an agreed upon sampling or analytic plan."

F4. Promotes reproducibility and replication, whether results are "significant" or not, by sharing data, methods, and documentation to the extent possible.

© RE Trachtenberg—ethical data science ~ UCI 26 January 2022-16

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.

TASK 7: WORK ON A TEAM

What in the ASA Ethical Guidelines is relevant to team work?

A.4 “Opposes efforts to predetermine or influence the results of statistical practices, and resists pressure to selectively interpret data.

B.5 “Strives to promptly correct substantive errors discovered after publication or implementation. As appropriate, disseminates the correction publicly and/or to others relying on the results.”

F. 5 “Serves as an ambassador for statistical practice by promoting thoughtful choices about data acquisition, analytic procedures, and data structures among non-practitioners and students. Instills appreciation for the concepts and methods of statistical practice.”

© RE Trachtenberg—ethical data science ~ UCI 26 January 2022-17

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).

© RE Tractenberg - ethical data science - UCI 26 January 2022-19