UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Social and Decision Analytics Division
Data Science Project Ethics Tool

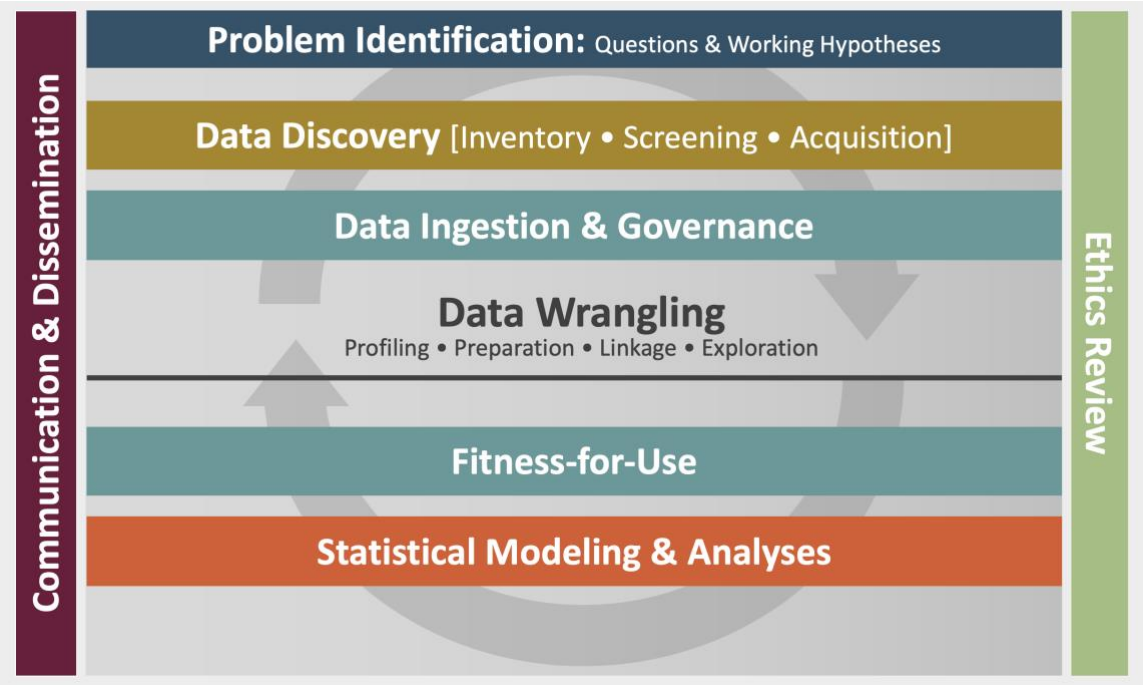
Sallie Ann Keller, Division Director, Social and Decision Analytics, Distinguished Professor in Biocomplexity, Professor of Public Health Sciences, University of Virginia, sak9tr@virginia.edu

Stephanie S. Shipp, Division Deputy Director and Professor, sss5ssc@virginia.edu

Aaron D. Schroeder, Research Associate Professor, ads7fg@virginia.edu

Data science teams bring together researchers and application stakeholders across many areas of expertise. Each with their own set of research integrity norms and habits. This requires that ethics be woven into every aspect of doing data science. The Data Science Framework reinforces this as data science ethics touches every component and step in the practice of data science.

Using an ethics tool is the first step for researchers to agree on a set of principles. The ***data science ethics tool template*** provided here can be adapted to specific data science projects. It provides a set of guiding principles and questions to address the ethical decisions across the data lifecycle.



References

Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data-driven governance: Creating a new foundation for democracy. Statistics and Public Policy, 4(1), 1-11. https://doi.org/10.1080/2330443X.2017.1374897

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.2d83f7f5

Source: Biocomplexity Institute, https://biocomplexity.virginia.edu/data-science-ethics

Social and Decision Analytics Division
Data Science Project Ethics Tool

Problem Initiation
Recognize and affirm that all project plans incorporate regular checks, discussion, and documentation to ensure adherence to the ethical principles of research.

Problem Identification
Establish the ethical basis for undertaking the project as well as the project requirements for both the protection of research participants and the equitable allocation of all potential project benefits and risks.

- ☐ What are the expected benefits of the project to the 'public good,' and do they outweigh potential risks to certain populations?
- ☐ Are there implicit assumptions and biases in the framing of the project regarding the studied communities and how will they be addressed?
- ☐ What type of institutional review board approval process is needed? Has the team reviewed the protocol?

Data Discovery, Inventory, Screening, & Acquisition
Consider potential biases that may be introduced through the choice of datasets and variables.

- ☐ Do the data include disproportionate coverage for different communities under study? Do
- ☐ data have adequate geographic coverage?
- ☐ Have checks and balances been established to identify and address implicit biases in the data?

Data Ingestion and Governance
Put in place data platforms and processes to ensure data transfer, storage, and database development adheres to data governance agreements and best practices for data quality assurance.

- ☐ Have team members reviewed standard operating procedures (SOPs) and data management plans?
- ☐ Do additional procedures need to be defined for this project?

Data Wrangling
Cleaning, transforming, linking, and exploratory analysis are critical steps in understanding data quality, how representative the data are, and potential biases in the data.

- ☐ What is the quality of the data?
- ☐ How representative are the data? What populations are covered, not covered?
- ☐ Are your assumptions correct?

Fitness-for-Use Assessment
Critically assess the overall utility of the results in achieving the predicted benefits of the study, to be transparent about potential limitations of the study, and to ensure that unintended biases haven't been introduced as a result of data choice and model refinement.

- ☐ What are the limitations of the results? Are the results useful given the purpose of the study?
- ☐ Do the statistical results support the potential benefits of the study previously stated?

- ☐ Do the statistical results support the mitigation of the potential risks of the study previously stated?
- ☐ Do any of the data require revisiting the question of potential biases being introduced through the choice of datasets and variables?

Statistical Modeling & Analysis

Establish transparency in methods, results and limitations.

- ☐ Have project methods and outputs been made as transparent as possible?
- ☐ Are the potential limitations of the research clearly presented?
- ☐ Should the research be used as the basis for policy action? Have the predicted benefits and social costs to all potentially affected communities been considered?

Communication and Dissemination

Summarize ethics-related questions and actions taken, to reinforce the process of ethical consideration in continuing and future projects. Refine protocols for replication and expansion of the research findings, and information dissemination.

- ☐ Did key ethical questions arise during the research and, if so, how were they addressed? How could they be addressed differently in future projects?
- ☐ Are research protocols, methods and data available to other researchers? If so, in what way, and, if not, what factors are limiting the ability to do so?

After Project Debriefing

Summarize questions and actions taken to reinforce the process of ethical consideration on all continuing and future projects. Establish protocols for replication and expansion of the research findings, and information dissemination.

- ☐ Did key ethical questions arise during the research and, if so, how were they addressed? How could they be addressed differently on future projects?
- ☐ Are research protocols, methods and data available to other researchers? If so, in what way, and, if not, what factors are limiting the ability to do so?