Ethical Data Science Practice

Marianne Huebner

Center for Statistical Training and Consulting (CSTAT)

Department of Statistics and Probability

Image credit: Emily Kuo, Techburst
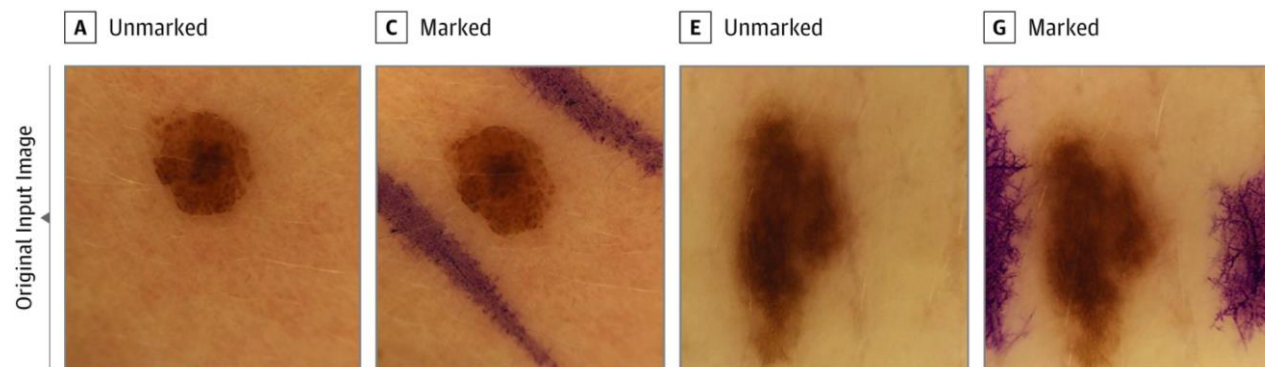
Where things go wrong

Data scientists make data-driven decisions that require the collection of data and model building approaches that can have serious implications for health, security, politics, society.

Diagnose melanoma: How are classifiers trained?

Winkler et al. JAMA 2019

Asthma drug development: AI classified black patients as lower risk vs equally sick white patients. Depends on whether patients were treated aggressively.

Holzhauer et al. Pharmaceutical statistics 2022

Ethical data science practice

Ethical data science takes a holistic approach that includes people from a range of perspectives (computational, statistical, scientific, ethical and legal, human) who collectively and collaboratively solve problems.

- Combine statistical thinking with the computational perspective:
  - Think about variation, uncertainty, sources of bias (measurement, representation, aggregation, evaluation, algorithmic, etc)
  - Think about data sources, data quality, initial data analysis framework
  - Think about reproducibility of methodological approaches
  - Think about unconscious bias of the data scientist (reflexivity statement)
- Reproducibility and reporting practices
- Downstream use of datasets

American Statistical Association 2022 Guidelines

Examples related to data ethics and AI

- Netflix documentary: Coded Bias
- Evil AI cartoons: https://www.evilaicartoons.com
- Data equity framework: https://weallcount.com
- DataEthics, not-for-profit, politically independent based in Denmark: https://dataethics.eu
- The Moral Machine Experiment, Nature 2018, https://www.nature.com/articles/s41586-018-0637-6
- Safiya Noble: Algorithms of Oppression, NYU Press 2018
- Data & Society: https://datasociety.net
- Professional organizations: Royal Statistical Society, American Statistical Associations

Abstract

Data scientists make data-driven decisions that require the collection of data and model building approaches that can have serious implications for health, security, politics, society.

The computational perspective is not enough. AI and Machine learning raise questions of accountability. Data quality standards may not have been considered. The outcome may be determined by the way the classifier is trained. Who takes responsibility when things go wrong?

Ethical data science takes a holistic approach that includes people from a range of perspectives (computational, statistical, scientific, ethical and legal, human) who collectively and collaboratively solve problems.

Elements of data ethics include respect privacy and confidentiality, attend to downstream use of data sets, promote transparency and accountability, follow applicable regulations and professional practices.