# Code of Ethics - Provenance Working Group

DRAFT IN-PROGRESS

Last updated: January 23nd, 2018

## *Introduction*

These community principles for provenance are being developed by data science professionals and researchers. This initiative is part of Data for Democracy project (in collaboration with Bloomberg and BrightHive) to develop a data science community code of ethics.

## *Principles for Provenance and Ownership*

*Clear data provenance promotes trust, aids transparency, and reduces barriers to sharing data. Collecting, propagating, storing, and curating provenance incurs costs, but these costs improve the value of the data and the decisions that the data inform. Maintaining data provenance is a key responsibility, and it should be done in a way that is mindful of the privacy trade-offs.*

### *Principle #1: Provenance as a responsibility*

Provenance-equipped datasets should be the norm rather than the exception. As a data collector, I will be responsible to record provenance; as a data publisher, I will be responsible to propagate provenance; as a data scientist, I will be responsible to review, consider, and declare what I know about data provenance.

### *Principle #2: Provenance should promote trust*

Provenance should be used to enact data access and usage control mechanisms, ensuring the rights of data subjects. Provenance should promote trust of data consumers, data scientists and the public in data analysis results.

### *Principle #3: Ignorance is not an excuse*

As a data scientist, prior to performing analysis, building models, generating conclusions, I have an obligation to understand where the data I use originated.

Usage of inappropriate, biased, or dirty datasets can lead to incorrect conclusions. It is my responsibility when I am consuming downstream data and when I provide data to others to demand provenance, and use it during the interpretation of the results.

### *Principle #4: Provenance should be fit to purpose*

While a perfect, complete, fine-grained record of every recordable measurement at all stages of the data science lifecycle may be a theoretically complete provenance, it is likely to include far too much information for most users in most situations. Ideally, provenance should inform stakeholders exactly what they need to know for the purpose at hand. As a data scientist, I will promote and maintain accurate reduced representations of provenance to make my actions and uses interpretable while protecting data subjects' privacy.

### *Principle #5: Provenance can evolve*

It may seem that that there is only one chance to record provenance, but refinements, either manual or automated, can improve the value of provenance after data is collected. Uncertain, incomplete, unstructured provenance is of enormous value, even though it is suboptimal. Perfect is the enemy of good. In addition to recording provenance as I have received it, I will refine, additionally curate, and better represent data provenance as new reliable information and trusted techniques for so doing become available.

# Misc

## Key topics:

a. *Ownership*
b. *Quality*
c. *Transparency*
d. *Value*
e. *Trust*
f. *Transformation*

## Personas:

*(Use Cases from Principles) [Needs to be one unified list across all the topics/groups]*

- Data subject (the person, object, or context the data is about)
- Data collector (
    - Responsibility: record sources of bias, etc.
- Data publisher
    - Responsibility: support for provenance to remain associated with data during upload, release, search. If the provenance exists, make sure there's a place to put it. Allow flexibility: free-text is better than nothing, key-values may be better than free-text.
- Data gatherer (they who find a dataset, transform it as needed, and assert its relevance for a particular task)
    - Responsibility: Review available provenance, think critically about the implied biases, insist on provenance for crucial applications that affect people, assess the quality of the provenance, provide feedback
- Data scientist (they who use a found dataset to answer a particular question)
    - Responsibility: Work to control for biases in source data implied by provenance information.
- Auditor
- Regulator
    - Responsibility: develop policy for provenance for crucial applications that affect people
- The general public
    - Responsibility: think critically about the sources of data; don't just "blame the algorithm."