# What Does Reproducibility Mean for Federal Statistics? An Academic Perspective

H.V. Jagadish

Univ. of Michigan

http://www.eecs.umich.edu/~jag

# Reproducibility

- The ability of an entire analysis of an experiment or study to be duplicated, either by the same researcher or by someone else working independently.
- Cornerstone of the scientific method.
- Human judgment and introspective intuition is not reproducible.
- Processing by defunct software is not (easily) reproducible.

# Talk Outline

- Achieving Reproducibility
- Managing Cost

# Achieving Reproducibility

- Ideally, a different method also leads to the same results.
- But at least someone else should be able to use (substantially) the same method to get the same results.
- Requires that every step be documented precisely.

# In a Recipe

Bake at 400 degrees for 7 minutes ✔

Cook uncovered until lightly browned ✔?

"Cooking" by Baron Valium CC-BY-SA

Add salt to taste ✖

# In Computing Research

- Can experimental results (e.g. performance evaluation of a new algorithm) reported in a research paper be reproduced?
- Volunteer “reproducibility committee” attempts reproduction and awards a badge to work that could be reproduced.
- Challenges include unstated code dependencies, unstated assumptions,…

# Survey Data

- Paradata are recorded, and reported.
- While not part of the “headline” conclusion, these are often critical for scholarly study, for reconciliation of differences, and so on.
- Additional processing of survey data is also recorded, but spottily.
  - E.g. manual error correction

# Big Data

- Administrative, business, and other data increasingly being re-purposed to compute statistics.
  - https://www.nap.edu/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy
- Recording “paradata” becomes even more important.
  - Meanings of variables may subtly differ.
  - But paradata and metadata may be limited.
- Likely to involve much more processing.

# Provenance

- Metadata associated with data that informs from where and how the data came about.
- At data set level:
  - Record workflow that created the data set.
- At data item level:
  - Explain how that individual item was derived,
  - List what source data items it depends on.

# We Know How to Do This

- Record provenance for workflow.
- Place software in versioned repo.
  - Managing dependencies can be tricky
  - But we (mostly) know how to do this, too.
- Note version used for all software, along with all invocation parameters.
- For non-software operations, have to record all changes.
- Need to retain access to source data.

# But There Are Challenges

- A full provenance dump is usually too much for a user to deal with.
- Provenance exploration methods are subject of current research.
- Records the What, Where, How; but does not record the Why.
  - Need to record assumptions.

# Talk Outline

- Achieving Reproducibility

- Managing Cost
  - Cost to producer
    - Will discuss tomorrow
  - Cost to consumer

# Expensive => Delegated

- Reproduction is typically expensive and requires skill.
- Most consumers of scientific conclusions (or statistics) cannot afford reproduction and do not have the skill.
- But still (should) care about reproducibility.

=> Rely on experts to reproduce and verify.

# National Statistics

- Additionally, consumers cannot (often) be given access to source data, due to privacy concerns.
- But still need explicit documentation of metadata to enable reproducibility
- Particularly as new sources of data are used, and new methodologies are introduced.

# Trust

- Verifiability is central to trust.
- The most straightforward way to verify a national statistic is to reproduce it.
- Reproducibility requires adequate metadata, and logging of all actions.
- Metadata/logging is critical for trustworthy national statistics.

# Acknowledgment


- NSF Grant 1250880