Using “git” to leverage development and external collaboration

William Sexton

GASP 2018, Washington D.C.
October 24, 2018

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

2020 Disclosure Avoidance System (DAS)

1. Science project.
    - Implement a DAS for the 2020 Decennial Census that protects the confidentiality of each individual respondent's microdata record by adhering to the leading disclosure limitation standard; differential privacy. Data publication systems that satisfy differential privacy must balance the cost of increased accuracy against foregone privacy.
    - Optimize the differentially private publication technology to push the privacy-loss – accuracy trade-off frontier outward.
2. Engineering project.
    - Develop a turnkey system; one that does not require any human interaction.
    - Develop a carefully documented, well-curated code base that may be publicly released.

2020 DAS Team

- The application to the Census Bureau's 2020 publication system incorporates work by John Abowd (Associate Director and Chief Scientist of Research and Methodology), Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Gerome Miklau, Brett Moran, Edward Porter, Anne Ross, William Sexton, Lars Vilhuber, and Pavel Zhuravlev.

Division of Developers

- Small internal team of developers that actively contribute to the 2020 DAS code base.
- Internal developers contribute to both the science project and the engineering project.
- Small external team of academics that engage in mutually beneficial research efforts to achieve the scientific goals.
- Git is a development tool we leverage for internal delevopment and external collaboration. Git helps us achieve both the scientific and engineering goals.

What is Git?

- A version-control system.

- A distributed version-control system meaning each working copy of the code is itself a repository that stores a complete history of all changes.

- One of the most widely-used version-control systems.
  - The large community means it's easy to get support when needed.

Our Git Setup

- We use the Github Enterprise development platform.
- A team of Technical Integrators control the network where our main remote repository is housed.
  - Manage permissions and access control
  - Handle security configurations
- We have the ability to set the visiblity of projects (public/private) and control who has permissions to make commits.

Advantages of Git

- Ease of use
- Development Flexibility
- Git hooks
- Submodules
- Branches
- External Collaboration

My version-control software ate my homework

- It’s hard to make irreversible mistakes; almost every operation can be undone.
  - Recover deleted files.
  - Rollback ”broken” commits.
  - Undo merges.

- Easy to pick up the basics.

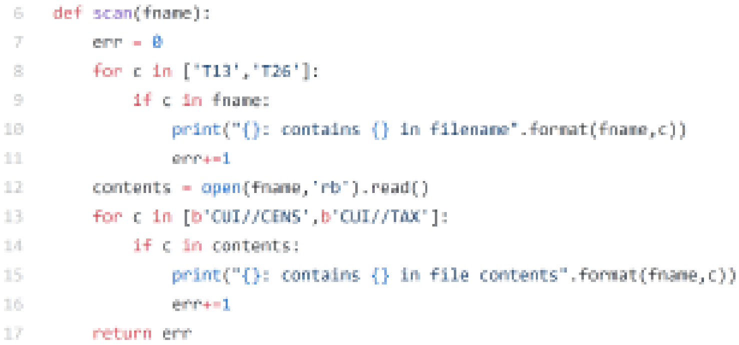
- Tons of online resources and it is easy to seek answers on community forums.

Flexibility

- We use Amazon Web Services (AWS) linux-based development environment.
  - Technical Integrators controls access, security, cluster start-up.
- We have ability to request the spin up/shut down of Elastic Map Reduce (EMR) clusters.
- Git enables fast setup on new clusters.
  - $ git clone git@github.com:example.git
- Git protects against unexpected crashes; EMR clusters are relatively easy to accidentally crash on large jobs. EC2 nodes are more resistant.

Git Hooks

- A hook is a script that Git runs before or after an event like a commit or a push.



```python
def scan(fname):
    err = 0
    for c in ['T13','T26']:
        if c in fname:
            print("{}: contains {} in filename".format(fname,c))
            err+=1
    contents = open(fname,'rb').read()
    for c in [b'CUI//CENS',b'CUI//TAX']:
        if c in contents:
            print("{}: contains {} in file contents".format(fname,c))
            err+=1
    return err
```

Submodules

- Submodules allow us to actively develop infrastructure while simultaneously using the infrastructure in other projects.

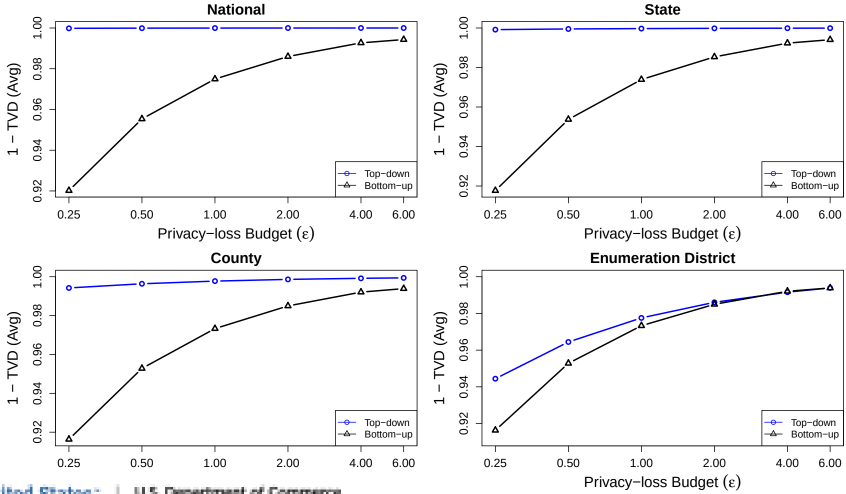
- The DAS framework is a driver program we've developed that abstracts the data flow into fundamental processing steps.
  - input reader
  - privacy algorithm engine
  - output writer

- The 2020 Decennial DAS, 2018 Census End to End test DAS, and the Business Dynamics Statistics (BDS) DAS use the DAS framework as a submodule.

Branches

- Branch off master → develop new feature → Confirm it is stable → Merge back into master.
- As the code base grows in complexity, branches allow developers to test refactoring code to improve modularity or to conform to coding standards.
- Allows us to “freeze” stable code for 2018 end to end Census test.
- Allows researchers to experiment simultaneously with the same code base while preserving the integrity of the master branch.
    - Develop a new feature.
    - Experiment with a new algorithm.
    - Experiment with subtle tweaks to the current algorithm.

Example: Branching to experiment with 1940 Census data



U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

External Collaboration: An Economist, Statistician, and Operations Researcher walk into a bar...

- to ask computer scientists for directions?
- Academic collaborators maintain a private external github organization.
- Internal repositories are periodically synced with external copies.
- Collaborators are able to test the code with public use data such as the 1940 Census. https://doi.org/10.18128/D010.V8.0.EXT1940USCB.
- Allows internal developers to get feedback from privacy experts.
- Independent external research projects can be pulled back in if desired.

References

1. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D010.V8.0.EXT1940USCB.

Thanks!

william.n.sexton@census.gov