Dataverse:
Recent and Upcoming Features

Gustavo Durand
Dataverse Technical Lead / Architect

Dataverse

Overview

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard’s Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- 15 on the core team - developers, designers, UI/UX, metadata specialists, curation manager

Dataverse Features - Data

- Persistent IDs / URLs
  - DataCite
  - Handle
- Automatically Generated Citations with attribution
- Compliant with FAIR and data citation principles
- Domain-specific Metadata
- Versioning
- File Storage
  - Local
  - Swift (OpenStack)
  - S3 (Amazon)

Dataverse Features - Users

- Multiple Sign In options
  - Native
  - Shibboleth
  - OAuth (ORCID)
- Dataverses within Dataverses
- Branding
- Widgets

Dataverse Features - Workflows

- Permissions
- Access Controls and Terms of Use
- Publishing Workflows
- Private URLs
- Upload / Download Workflows
  - Browser
  - Dropbox
  - Rsync (for big data “packages”)

Dataverse Features - Interoperability

- APIs
  - SWORD
  - Native
- Harvesting (OAI-PMH)
  - Client
  - Server

Glassfish Server 4.1

Java SE8

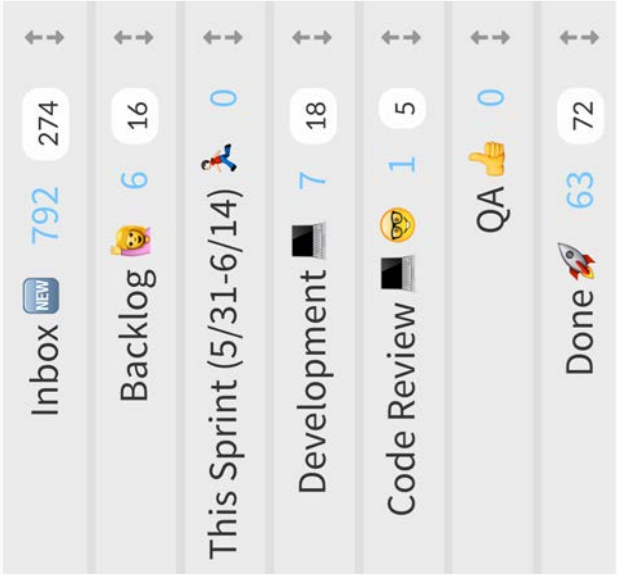
Java EE7

- Presentation: JSF (PrimeFaces), RESTful API
- Business: EJB, Transactions, Asynchronous, Timers
- Storage: JPA (Entities), Bean Validation

Storage: Postgres, Solr, File System / Swift / S3

Dataverse Development Process

- Inbox
- Backlog
- This Sprint
- Development
- Code Review
- QA
- Done



https://waffle.io/IQSS/dataverse

(some) Collaborations

- SBGrid Data
  - Large Data and Support
- Massachusetts Open Cloud
  - Big Data Storage and Compute Access (OpenStack)
- DANS/CIMMYT
  - Handles Support
- ResearchSpace
  - API Java Client Library
- Provenance
  - W3C PROV

Dataverse Community

- 33 installations around the world

Dataverse Community

- 50+ code contributors outside of the Core Team
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
  - Dataverse Google Group
  - Dataverse Community Calls
  - Dataverse Community Meeting
- Global Dataverse Community Consortium

Recent / Upcoming Features

File Persistent Identifiers

Authors, Published Year, “File Name”, *Dataset Title*, **File Persistent Identifier**, Repository Name, Version, Universal Numerical Fingerprint (UNF)

- JOINT DECLARATION OF DATA CITATION PRINCIPLES #7: Specificity and Verifiability: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim.
- Important for Provenance
- APIs now use PIDs
- Format can be related to Dataset PID or independent (configurable by installation)

Provenance

Provenance means “a record of ownership of a work of art or an antique, used as a guide to authenticity or quality.”

Provenance for Data means “the information required to accurately document the history of data, including how it was created and how it was transformed.”

Big Data Support and Multiple Storage Locations

Files | Metadata | Terms | Versions

1 File

3
Dataverse Package - 1.9 GB - Apr 13, 2015
SHA-1: 3697862ac80b0986b02ef83f0a0f81ebadf1aba1

This data file can be accessed through a terminal window, using the commands below. For more information about downloading and verifying data, see our User Guide.

Local Access

```
/programs/datagrid/10.15785/SBGRID/3
```

Download Access

```
rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/3
```
(Harvard Medical School, USA)

```
rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/3
```
(Uppsala University, Sweden)

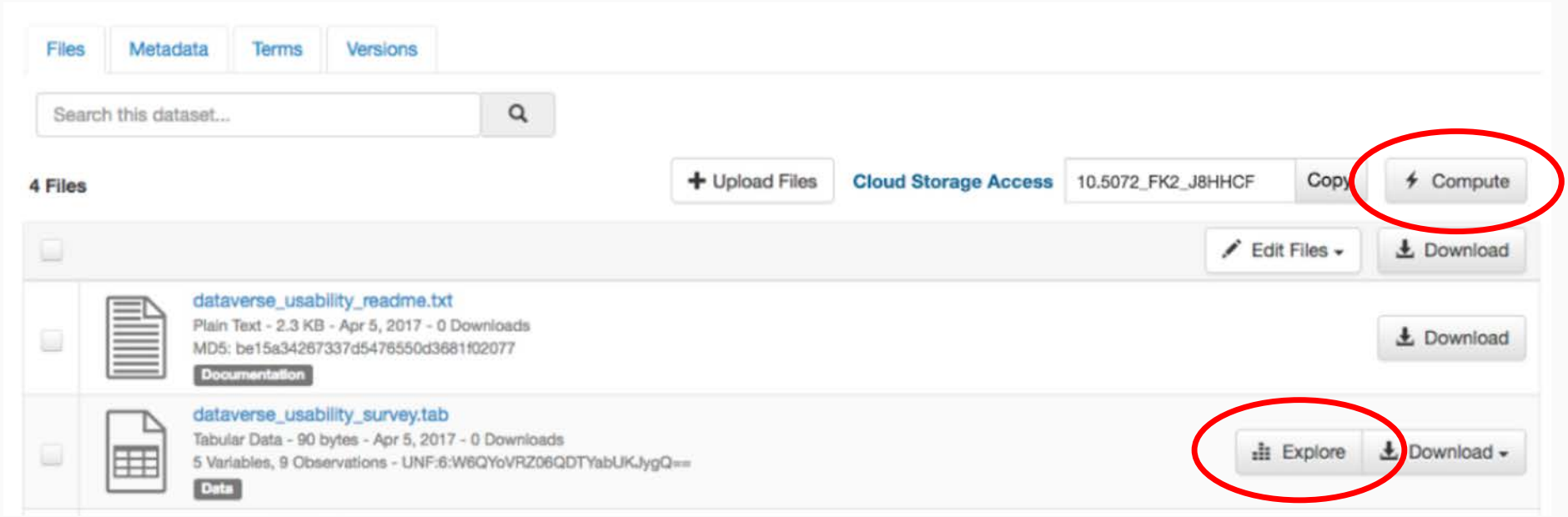
```
rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/3
```
(Institut Pasteur deMontevideo, Uruguay)

```
rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/3
```
(Shanghai Institutes for Biological Sciences, China)

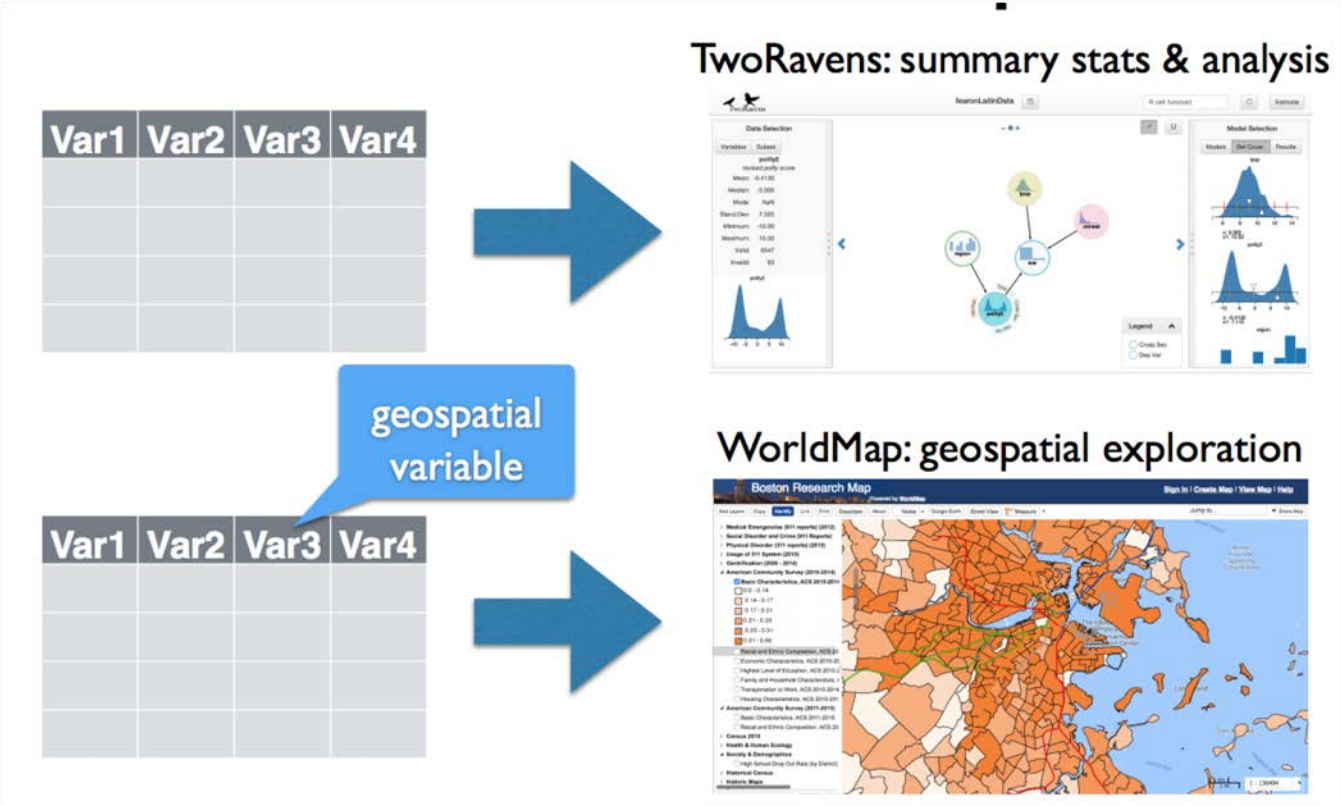
Verify Data

```
cd 3 ; shasum -c files.sha
```

Compute/Explore Access



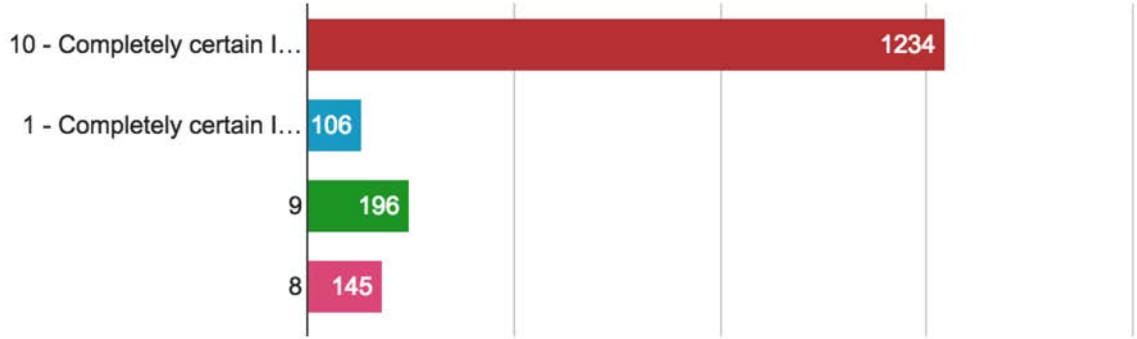
External Tools: Two Ravens and World Map



External Tools: Data Explorer

Chart View | Table View

Variable GRIDTURNOUT3_1_TURNOUT3: TURNOUT3. In October, 2015, the next Federal Election will be held. Using a 1-to-10 scale, where 10 means you are completely certain you will vote and 1 means you are completely certain you will NOT vote, how likely are you to vote in the upcoming Federal



A **datatag** is a set of security features and access requirements for file handling.

A **datatags repository** is one that stores and shares data files in accordance with a standardized and ordered level of security and access requirements.

https://datatags.org

DataTags Levels

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Differential Privacy

$$Pr[T(M(X)) = 1] \leq e^{\epsilon} Pr[T(M(X')) = 1] + \delta, \qquad \forall\, T.$$

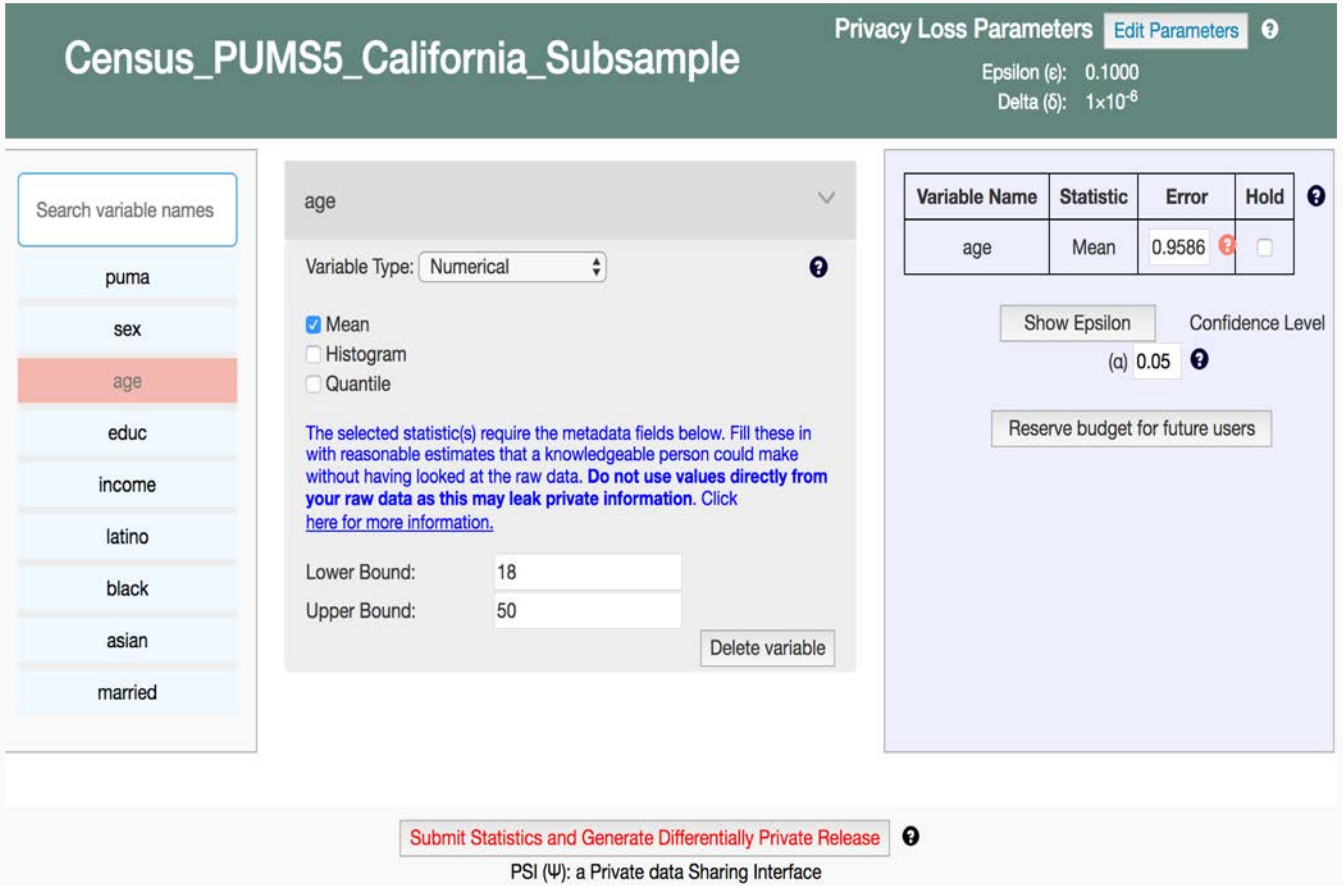
Differential Privacy is a formal, mathematical conception of privacy preservation. It **guarantees** that any reported result does not reveal information about any one single individual, regardless of auxiliary information.

https://privacytools.seas.harvard.edu/differential-privacy

https://privacytools.seas.harvard.edu/psi

External Tools: PSI Budgeteer

The budgeteer allows users to select which statistics they would like to calculate and are given estimates of how accurately each statistic can be computed. They can also redistribute their privacy budget according to which statistics they think are most valuable in their dataset.



Thank you!

Please get in touch with us!

Google Group, Github, IRC, Twitter - dataverse.org/contact

support@dataverse.org