



Dataverse and OpenDP

Tools for Privacy-Protective Analysis in the Cloud

Red Hat Research Day, September 22

Mercè Crosas, Ph.D., Harvard University
Harvard University Information Technology & Institute for Quantitative Social Science
@mercecrosas

OpenDP plans to support sensitive data from academia, government, and companies

- **Data repositories** to offer academic researchers privacy-preserving access to sensitive data.
- **Government agencies** to safely share sensitive data with researchers, data-driven policy makers, and the broader public.
- **Companies** to share data on their users and customers with academic researchers or with institutions that bring together several such datasets.

Dataverse powers Data Repositories

Dataverse Software



- Open-source software for data repositories
- 61 Dataverse repositories world-wide
- A community of users and contributors

Example: Harvard Dataverse

The screenshot shows the Harvard Dataverse website interface. At the top, there is a navigation bar with links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. The main content area includes a header for 'HARVARD Dataverse' and a description: 'Deposit and share your data. Get academic credit. Harvard Dataverse is a repository for research data. Deposit data and code here.' It displays statistics: '100,077 datasets', '19,581,653 downloads', and '4,074 dataverses'. There are two buttons: 'Add a dataset +' and 'Add a dataverse +'. Below this is a search section with the text 'Find data across research fields, preview metadata, and download files' and a search input field with a 'Find' button. A featured section highlights the 'COVID-19 Data Collection' as a 'curated collection of COVID-19 data deposited in the Harvard Dataverse repository.' At the bottom, there is a 'Browse by subject' section with a grid of subjects and their respective dataset counts: Agricultural Sciences (3,336), Computer and Information Science (1,434), Medicine, Health and Life Sciences (4,421), Arts and Humanities (1,924), Earth and Environmental Sciences (3,054), Physics (1,171), Astronomy and Astrophysics (893), Engineering (700), and Social Sciences (43,051). A 'Feedback' button is located in the bottom right corner.

- 100K searchable datasets
- 19M file downloads
- Open to all research fields

Dataverse and OpenDP together will be able to:

- Enable **search and exploration** of sensitive datasets in the repository without accessing the original data
- Produce **rich statistical summaries** of sensitive datasets to be shared widely without risk of revealing individual-level information
- Generate **differentially private “synthetic data”** that reflects many statistical properties of the original dataset
- Allow approved researchers to use the OpenDP query interface to **run differentially private analysis of interest**
- Facilitate **reproducibility** of research with sensitive datasets

Non-Sensitive vs. Sensitive data in Dataverse

Non-Sensitive DataTags (DATAVERSE TODAY)

Blue

Publicly open, no barriers

Green

Publicly open, but need to register to access

Yellow

Restricted, need to be granted permissions, but non-sensitive

Sensitive DataTags (FUTURE RELEASE)

Orange

Requires Data Use Agreement (DUA); requires data enclave
(moderate sensitive)

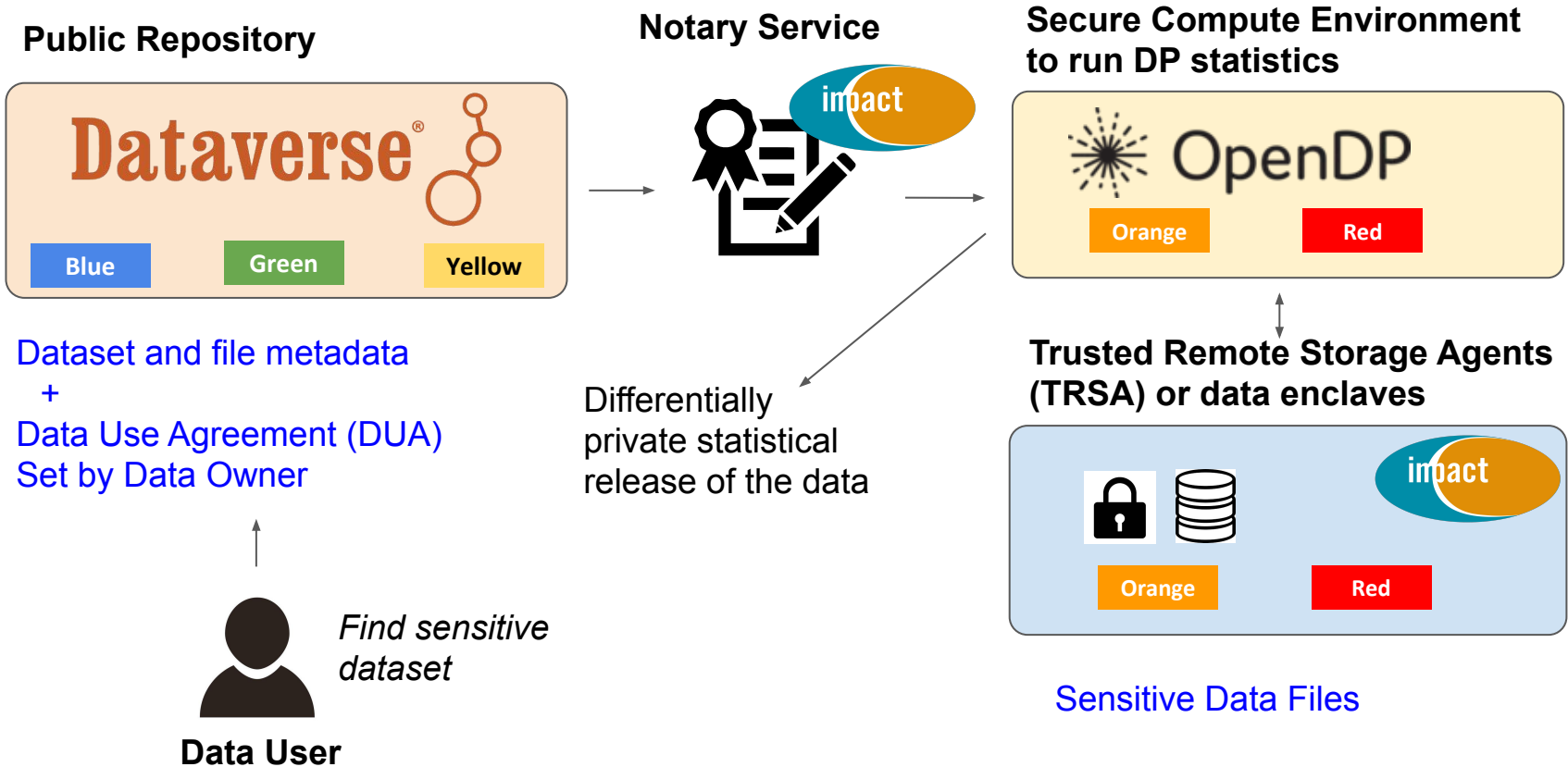
Red

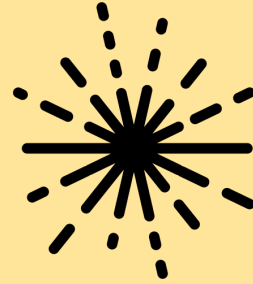
Requires DUA; stricter security requirements and audits
(high sensitive)

Crimson

Only metadata and no link to data; data stored outside network
(maximum sensitive)

Dataverse + OpenDP + Data Enclave





Thanks