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

- 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

<table>
<thead>
<tr>
<th>Non-Sensitive DataTags</th>
<th>Sensitive DataTags</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(DATAVERSE TODAY)</strong></td>
<td><strong>(FUTURE RELEASE)</strong></td>
</tr>
<tr>
<td>Blue</td>
<td>Orange</td>
</tr>
<tr>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>Yellow</td>
<td>Crimson</td>
</tr>
</tbody>
</table>

- **Blue**: Publicly open, no barriers
- **Green**: Publicly open, but need to register to access
- **Yellow**: Restricted, need to be granted permissions, but non-sensitive
Dataverse + OpenDP + Data Enclave

Public Repository

**Dataverse**

Dataset and file metadata + Data Use Agreement (DUA) Set by Data Owner

Data User

Find sensitive dataset

Notary Service

Differentially private statistical release of the data

Secure Compute Environment to run DP statistics

OpenDP

Trusted Remote Storage Agents (TRSA) or data enclaves

Sensitive Data Files

Data User
Thanks

@mercecrosas