

Enhancing FAIR-ness in Harvard Dataverse with Variable-Level Metadata and Differential Privacy

Stefano M. Iacus, Senior Research Scientist

Director of Data Science and Product Research @ IQSS, Harvard University

Affiliate Faculty of the Kempner Institute for the Study of Natural and Artificial Intelligence

*Computational Social Science Conference: innovative methods, research workflows and data stewardship,
28-29 October 2024, Barcelona @ Universitat Politècnica de Catalunya*



What is Dataverse?

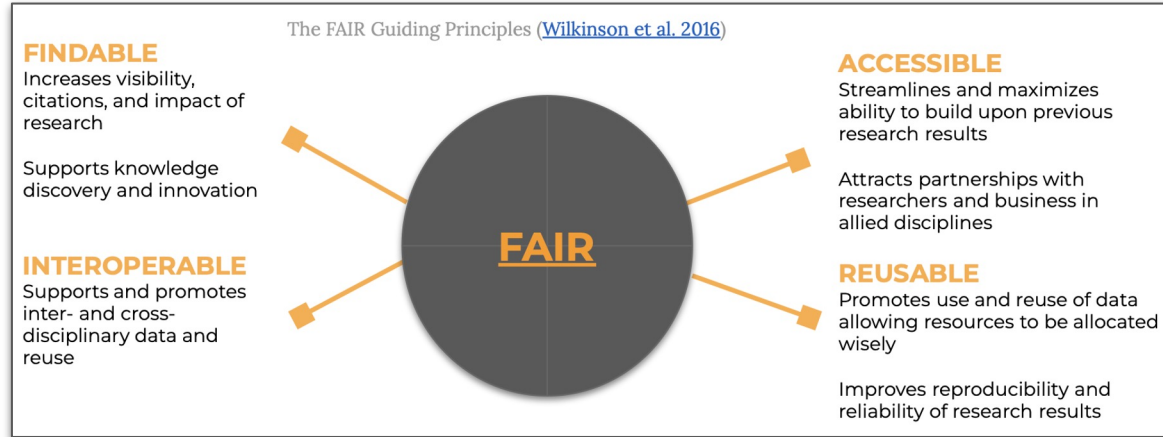
An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.



Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

Physics 1,760

Social Sciences 64,287

What is Dataverse?

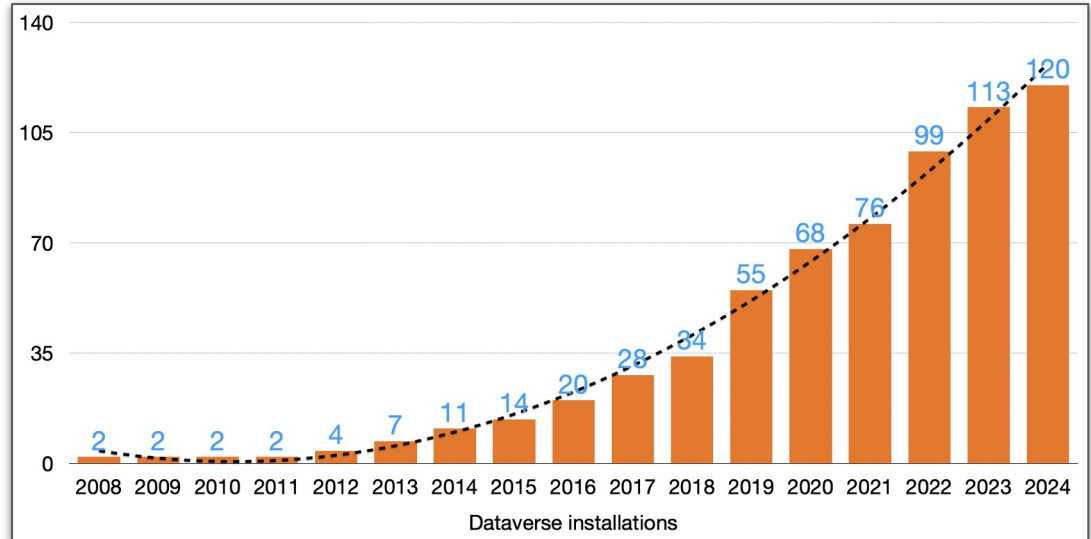
An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Built to support **multiple types of data, users, and workflows**

Supports **FAIR** principles and **Signposting**.

Developed mainly at Harvard's Institute for Quantitative Social Science (IQSS) since 2006 + key contributors from our large community

Started as a data sharing platform for the social science now **covers a wide range of disciplines**.



Agricultural Sciences 4,904

Arts and Humanities 36,716

Astronomy and Astrophysics 1,350

Business and Management 2,341

Chemistry 955

Computer and Information Science 3,798

Earth and Environmental Sciences 9,554

Engineering 2,292

Law 5,849

Mathematical Sciences 722

Medicine, Health and Life Sciences 10,548

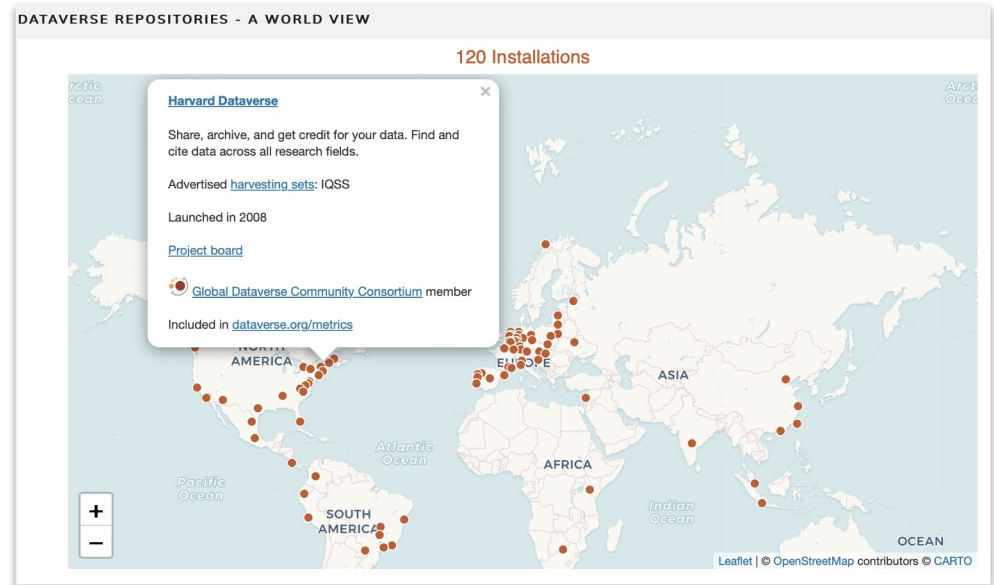
Physics 1,760

Social Sciences 64,287

What is Dataverse?

An **open-source** platform that provides a **generalist** repository to **publish, cite, and archive research data**

Soon at BSC ?



Follow this [link](#) if you want to know more about Dataverse and its history



Signposting and Discoverability @ Dataverse

Repositories **web pages** are **not optimized** for use by **machine agents** that navigate the scholarly web.

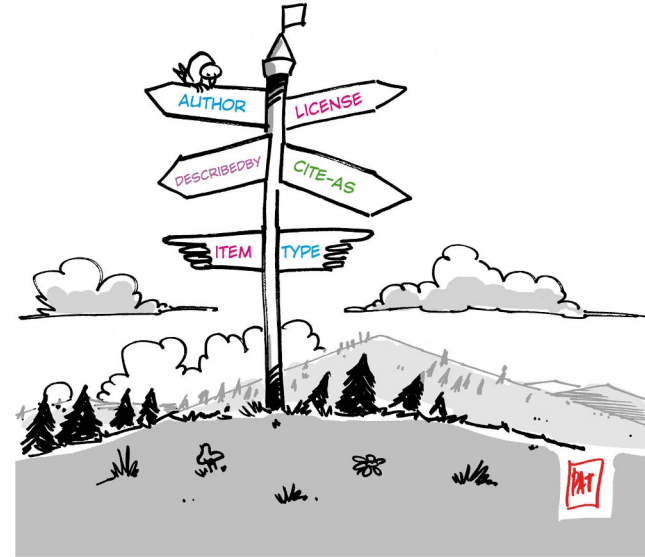
How can a robot determine which links on a landing page lead to content and which to metadata?

How can a bot distinguish those links from the myriad of other links on the page?

Signposting exposes these info to bots in in a standards-based way.

Release 5.14 added [Signposting](#) support to Dataverse to improve machine discoverability of datasets and files. To date, Dateverse is the only generalist repository supporting it.

More discoverability features here: <https://guides.dataverse.org/en/5.14/admin/discoverability.html>



Metadata Types

1. **Citation Metadata:** any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset;
1. **Domain Specific Metadata:** with specific support currently for Social Science, Life Science, Geospatial, and Astronomy datasets;
1. **File-level Metadata:** varies depending on the type of data file and include options like file tags, descriptions, variable names, and hierarchy preservation.

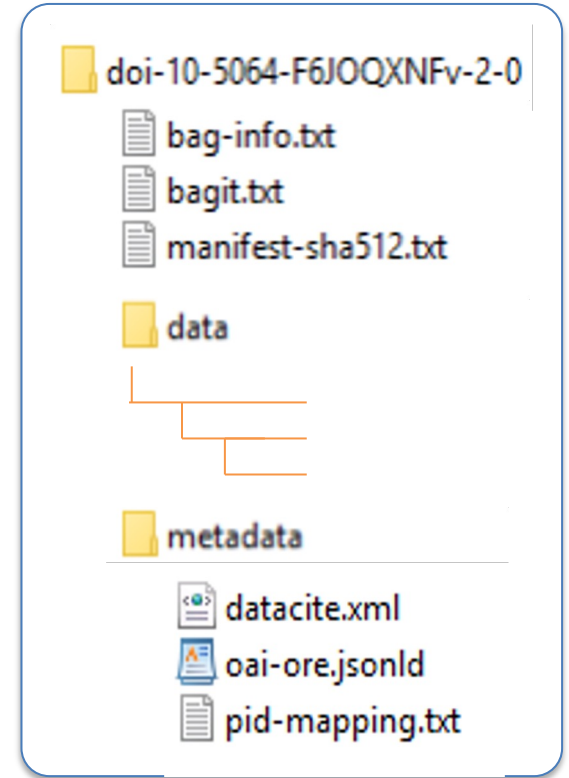
Provenance,
authorship, etc

Higher
granularity of
information

BagIt (Dataset Level Metadata)

About BagIt

- Standards-based approach to dataset exchange and archiving
- Hierarchical structure for storing data & metadata for preservation
- Includes required descriptive metadata & fixity information



[Detailed Video](#) (8 mins)

Dataverse & BagIt

- Generates BagIt zip file with complete metadata & all datafiles for a Dataverse dataset
- Conforms to RDA recommendations & includes complete JSON-LD/RDF metadata using [OAI-ORE](#) structure
- Imports BagIt packages as datasets, providing round-trip, export/import capability.

Dataverse Archives Using BagIT

- [Texas Data Repository](#)
(Duracloud/Chronopolis)
- [Qualitative Data Repository](#)
(Google Cloud)
- [Harvard Data Commons](#)
Workflow developed for depositing archival bags into the Harvard DRS repo (not in production yet)

Dataverse & OAI_ORI

OAI_ORI is one of the built-in metadata export formats that Dataverse supports (used also for BagIT).

It is meant to be able to export/encode **ALL** of the metadata available for a dataset.

it is used in Dataverse APIs for developing export plugins, as one of the carriers of metadata that the plugin code can further manipulate.

<https://guides.dataverse.org/en/latest/developers/metadataexport.html>

Dynamic (Domain Specific) Metadata

- **Metadata is defined dynamically** at the database level, allowing for modularly adding new Metadata blocks
- Supports:
 - single or **multiple** values
 - simple or **compound** values
 - controlled vocabularies
 - **external** vocabularies

Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- Citation Metadata (Required) [\[+\] View fields + set as hidden, required, or optional](#)
- Geospatial Metadata [\[+\] View fields](#)
- Social Science and Humanities Metadata [\[+\] View fields](#)
- Astronomy and Astrophysics Metadata [\[+\] View fields](#)
- Life Sciences Metadata [\[+\] View fields](#)
- Journal Metadata [\[+\] View fields](#)

Citation Metadata [^](#)

Title * ⓘ

Author * ⓘ

Name * ⓘ <input type="text" value="Admin, Dataverse"/>	Affiliation ⓘ <input type="text" value="Dataverse.org"/>	<input type="button" value="+"/>
Identifier Scheme ⓘ Select... <input type="text"/>	Identifier ⓘ <input type="text"/>	

Contact * ⓘ

Name ⓘ <input type="text" value="Admin, Dataverse"/>	Affiliation ⓘ <input type="text" value="Dataverse.org"/>	<input type="button" value="+"/>
E-mail * ⓘ <input type="text" value="dataverseadmin@iq.harvard.edu"/>		

Description * ⓘ

This field supports only certain [HTML tags](#).

Text * ⓘ

Date ⓘ

DDI Codebook support (Variable Level Metadata)

- An important Dataverse feature (unique at the time it was) was that, at **submission time**, the platform attempted an **automatic extraction of data- and variable-level tabular metadata** from supported statistical file formats (Stata, SPSS, R, etc.).
- A user could upload a Stata .dta file and the data is automatically converted into archival/non-proprietary tabular, column-subsettable format.
- Descriptive statistics are calculated and variable metadata are extracted from the file and stored in the database.
- The database schema is largely based on the **DDI Codebook** specifications.

DDI Codebook metadata



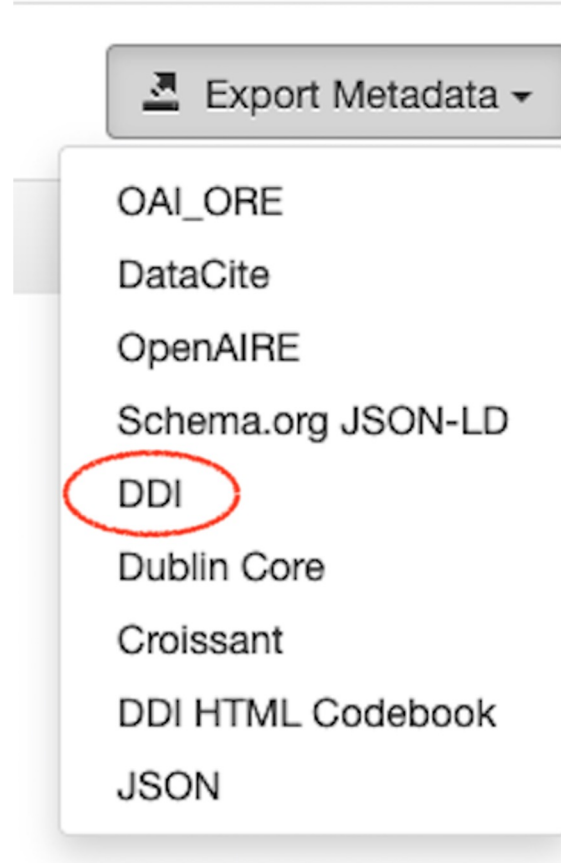
```
...
<var ID="v402" name="region" intrvl="contin">
<location fileid="f9243"/>
<labl level="variable">country's region, based on MAR project</labl>
<sumStat type="min">0.0</sumStat>
<sumStat type="mean">4.108623</sumStat>
<sumStat type="stdev">2.467951</sumStat>
<sumStat type="vald">6610</sumStat>
<sumStat type="medn">5.0</sumStat>
<sumStat type="mode">.</sumStat>
<sumStat type="invd">0</sumStat>
<sumStat type="max">7.0</sumStat>
<catgry>
<catValu>5</catValu>
<labl level="category">n. africa and the middle east</labl>
<catStat type="freq">910</catStat>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">e. europe and the former soviet union</labl>
<catStat type="freq">646</catStat>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">asia</labl>
<catStat type="freq">1096</catStat>
</catgry>
...
```

Variable-level metadata and DDI Codebook 2.5



From the internal stored version of the metadata Dataverse provides different exporters.

Exporters can be added as plug-ins



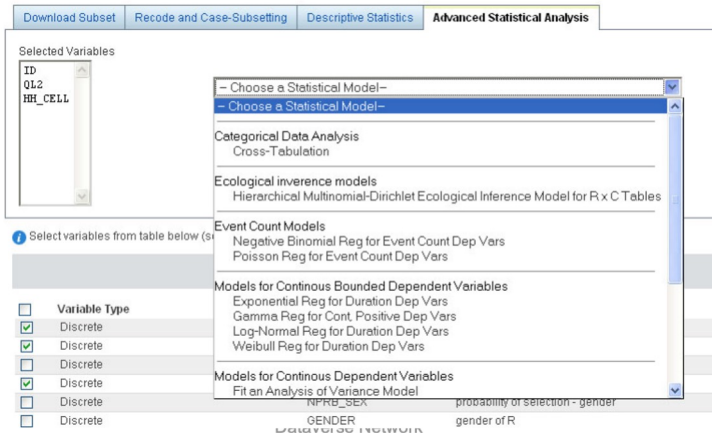
Use of the DDI Codebook Metadata

Utilized by various data exploration and visualization external tools available to the Dataverse users

from an early version (v3.0):

A rich set of data analysis based on R statistical package

- Download a subset of variables
- Recode a variable
- Apply descriptive statistics or and advanced statistical models (from Zelig/R)

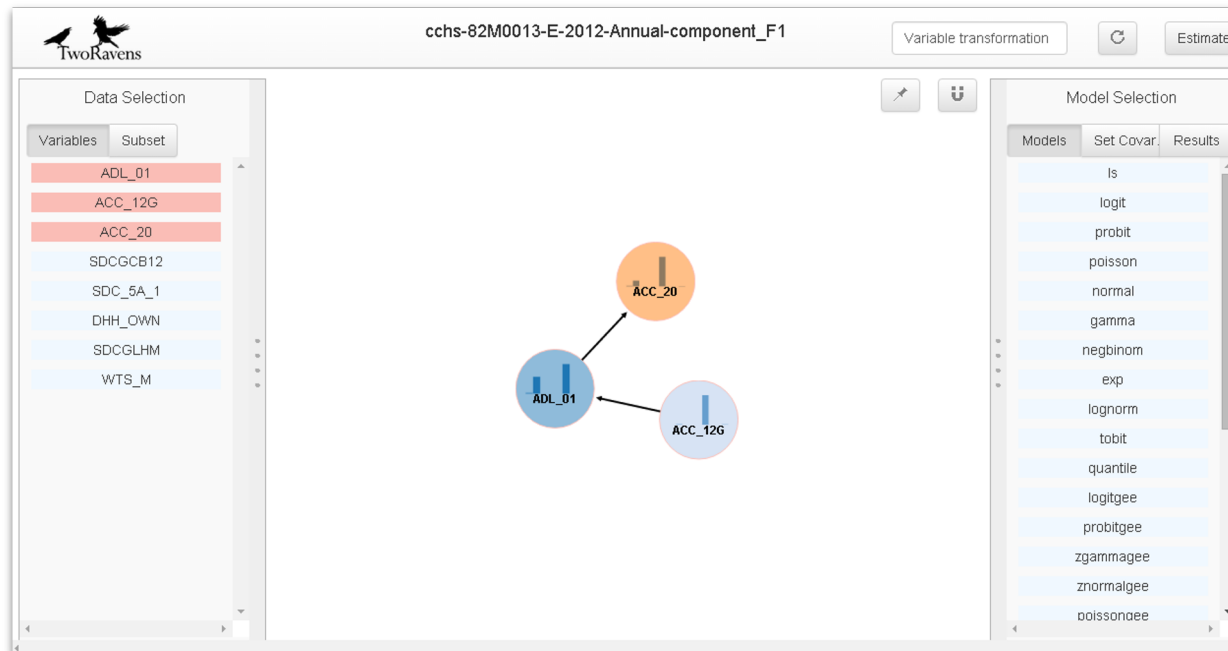


The screenshot displays the 'Advanced Statistical Analysis' tab in the DDI software. It features a 'Selected Variables' list containing 'ID', 'Q12', and 'HH_CELL'. Below this list is a 'Variable Type' section with checkboxes for 'Discrete'. A large dropdown menu is open, showing a list of statistical models including 'Categorical Data Analysis', 'Ecological Inference Models', 'Event Count Models', 'Models for Continuous Bounded Dependent Variables', and 'Models for Continuous Dependent Variables'. At the bottom of the interface, a table shows variables such as 'GENDER' and 'probability of selection - gender'.

Use of the DDI Codebook Metadata

Integration with Two Ravens

v.4.0 (ca 2018):



The screenshot displays the Two Ravens software interface for the dataset 'cchs-82M0013-E-2012-Annual-component_F1'. The interface is divided into several sections:

- Data Selection:** A list of variables is shown, with 'ADL_01', 'ACC_12G', and 'ACC_20' highlighted in red. Other variables include SDCGCB12, SDC_5A_1, DHH_OWN, SDCGLHM, and WTS_M.
- Model Selection:** A list of statistical models is displayed, including 'Is', 'logit', 'probit', 'poisson', 'normal', 'gamma', 'negbinom', 'exp', 'lognorm', 'tobit', 'quantile', 'logitgee', 'probitgee', 'zgammagee', 'znormalgee', and 'poissondee'.
- Diagram:** A central diagram shows three nodes: 'ADL_01' (blue circle), 'ACC_12G' (light blue circle), and 'ACC_20' (orange circle). Arrows point from 'ADL_01' and 'ACC_12G' to 'ACC_20', indicating a causal or predictive relationship.

Use of the DDI Codebook Metadata



Data Explorer (by Borealis/Open Scholar) as integrated with the current Dataverse v6

Forum Covid-19 Tracking English

Forum_COVID Tracking_Data.tab

Forum Research Inc, 2020, "Forum Covid-19 Tracking", <https://doi.org/10.5683/SP2/YM8BCJ>, Borealis, V1, UNF:6:b2sqE84ecQk12Y2CQgXUkA== [fileUNF]

< Hide Groups Download Save to Dataverse

Add Group + Search Items per page 25 1 - 21 of 21

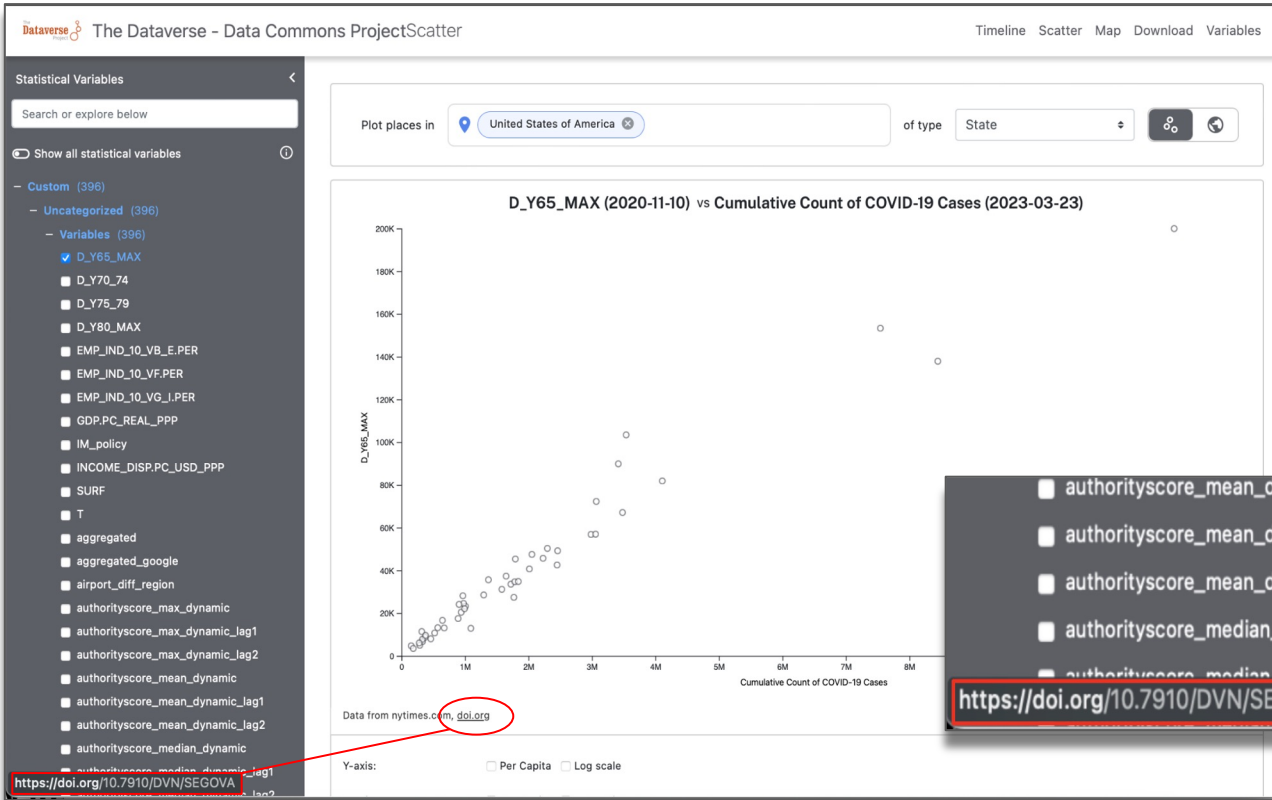
All Variables

<input type="checkbox"/>	ID	Name	Label	Weight	View	
<input type="checkbox"/>	v457580RiD	RpsRespondent				
<input type="checkbox"/>	v457594Q1		Have you, or has anyone in your household had a fever, that is, a temperature above 38 degrees Celsius or about 100 degrees Fahrenheit, in the past we...			
<input type="checkbox"/>	v457599Q2		Are you, or is anyone in your household currently suffering from a new cough in the past week?			
<input type="checkbox"/>	v457596Q3		Are you, or is anyone in your household currently suffering from new headaches in the past week?			
<input type="checkbox"/>	v457593Q4		Are you, or is anyone in your household suffering from a new sore throat in the past week?			
<input type="checkbox"/>	v457587Q5		Are you, or is anyone in your household suffering from a loss of taste or smell in the past week?			
<input type="checkbox"/>	v457595Q6		Are you, or is anyone in your household suffering from new diarrhea in the past week?			
<input type="checkbox"/>	v457598Q7		Are you, or is anyone in your household suffering from a new shortness of breath in the past week?			
<input type="checkbox"/>	v457584Q8		Have you, or has anyone with symptoms in this household been tested for COVID-19 since the onset of symptoms?			

Google DataCommons & Harvard Dataverse Project (WiP)



Join data across **Dataverse** and with other public data sources through **space** and **time** using schema.org metadata and (temporarily) Google Data Commons software to visualize the results.



Example: crossing NYT & Dataverse data

Try it at gdc.dataverse.org

DDI-CDI and Dataverse

Dataverse has long since expanded into multidisciplinary research data outside of Quantitative Social Science.

This means having to handle

- Domain-specific metadata that's outside of what can be described using the original DDI Codebook-based vocabulary
- Data files (including Big Data) that are not necessarily rectangular, “wide” variable-observations tables (for example, streams of data as produced by experimental instruments or other types of “long” data files)

... and this is where we see **DDI-CDI** in our future

Generalist Repository Ecosystem Initiative (NIH)

Development of support for DDI-CDI in Dataverse is funded in part by the GREI NIH grant.

From Aim 2:

Increase support for biomedical and cross-domain metadata standards and controlled vocabularies

- 1#. expand DDI support to include the recently released DDI-Cross-Domain Integration (DDI-CDI) schema



Adopting and implementing DDI-CDI will allow Dataverse to align even better with the FAIR data principles.

For more information, see the Project issue <https://github.com/IQSS/dataverse-pm/issues/286>

Interest in DDI-CDI within Dataverse Community

Members of the community from across the world have expressed potential interest in the technology

Dataverse Community Group thread on DDI-CDI:

<https://groups.google.com/g/dataverse-community/c/sjiY3-OQPhc/m/QwYrHEEyAwAJ>

A Zulip channel dedicated to DDI-CDI:

<https://dataverse.zulipchat.com/#narrow/stream/450733-ddi-cdi>

Nov 5, 2024 10am Eastern, **Arofan Gregory** presenting @ **Dataverse Community Call:**

To join: <https://harvard.zoom.us/j/98965964030> (Passcode: community)

CroissantML: AI/ML-Ready Datasets Simplified

arXiv > cs > arXiv:2403.19546v1 Search... Help | Ad

Computer Science > Machine Learning

[Submitted on 28 Mar 2024 (this version), latest version 30 May 2024 (v2)]

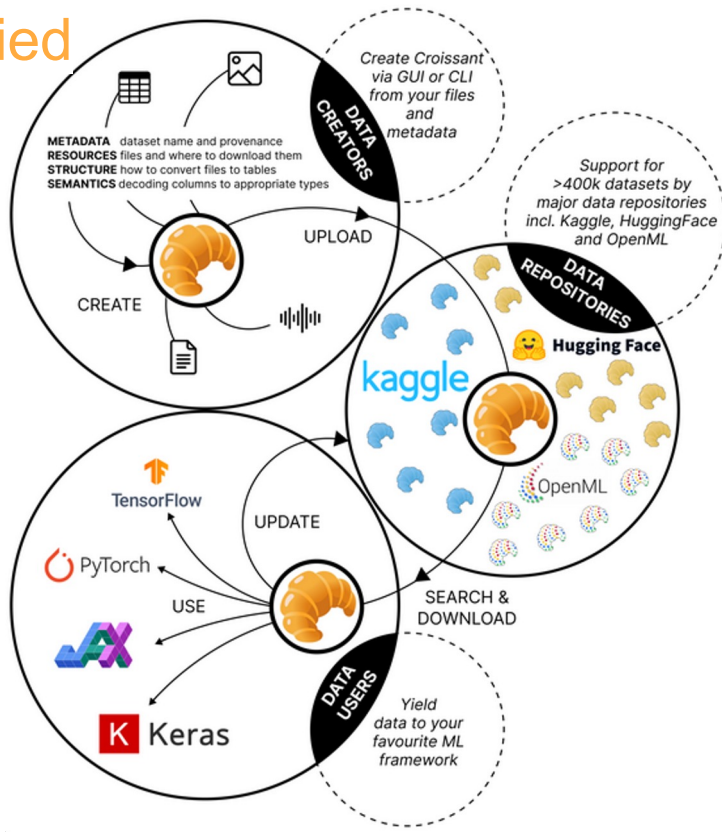
Croissant: A Metadata Format for ML-Ready Datasets

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Joan Giner-Miguel, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruysen, Rajat Shinde, Elena Simperl, Geoffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Steffen Vogler, Carole-Jean Wu

Data is a critical resource for Machine Learning (ML), yet working with data remains a key friction point. This paper introduces Croissant, a metadata format for datasets that simplifies how data is used by ML tools and frameworks. Croissant makes datasets more discoverable, portable and interoperable, thereby addressing significant challenges in ML data management and responsible AI. Croissant is already supported by several popular dataset repositories, spanning hundreds of thousands of datasets, ready to be loaded into the most popular ML frameworks.

<https://arxiv.org/abs/2403.19546v1>

IQSS/DV contributed to definition of this standard for ML/AI workflows. HDV exposes CroissantML metadata to crawlers to increase discoverability of this type of data and has a built-in exporter.



Dataverse dataset now being exposed to the ML/CS community

Thanks to Croissant-ML support.

Now in production.

The screenshot shows a browser window with the URL <https://datasetsearch.research.google.com/search?src=3&query=cancer&docid=L2cvMTFq>. The search results are filtered by 'Croissant'. The top result is 'Air Quality-Lung Cancer Data' from 'dataverse.harvard.edu', updated on Jan 31, 2020. A red circle highlights the text '4 scholarly articles cite this dataset (View in Google Scholar)'. Below this, there are sections for 'Unique identifier' (https://doi.org/10.7910/DVN/HMOEJO), 'Dataset updated' (Jan 31, 2020), 'Dataset provided by' (Harvard Dataverse), and 'Authors' (Mithun Acharjee; Kumer Pial Das; Young S.Stanley). Other results include 'Data from: breast-cancer-wisconsin' and 'Cancer Prediction Dataset' from Kaggle.

Dataverse and Sensitive Data



**I have
sensitive
data. Can
I share it?**



**Sure! I
just need to
remove personal
identifiers.**

Let's go together through a
simple example...

*“If you publish too many statistics and too accurately, **no one**, in the input database, **can be given any reasonable assurance of the confidentiality** of their input data.”*

2019, John Abowd, Chief Scientist, U.S. Census, 2016 to 2022

Example: Sharing Data

Class has 5 students

Teacher writes the average exam score on the board: 85%

Exam Average:

85%

(Example courtesy of Ethan Cowan, former OpenDP team member and author of upcoming O'Reilly book, [Hands-On Differential Privacy: Introduction to the Theory and Practice Using OpenDP](#))

Sharing Data

Class has 5 students

Your friend joins the circus and drops the class

Exam Average:

85%



Sharing Data

Class has 5 students

Your friend joins the circus and drops the class

Has your friend's privacy been violated?



Exam Average:

~~85%~~

84%

Has your friend's privacy been violated?

$$\frac{Q + R + S + T + X}{5} = 85$$

$$\frac{Q + R + S + T}{4} = 84$$

Exam Average:

~~85%~~

84%

Has your friend's privacy been violated?

$$\frac{Q + R + S + T + X}{5} = 85$$

$$\frac{Q + R + S + T}{4} = 84$$

$$Q + R + S + T + X = 85 * 5 = 425$$

$$Q + R + S + T = 84 * 4 = 336$$

Exam Average:

~~85%~~

84%

Has your friend's privacy been violated?

$$\frac{Q + R + S + T + X}{5} = 85$$

$$\frac{Q + R + S + T}{4} = 84$$

$$Q + R + S + T + X = 85 * 5 = 425$$

$$Q + R + S + T = 84 * 4 = 336$$

$$\mathbf{X = 425 - 336 = 89}$$

← your friend's score !

Exam Average:

~~85%~~

84%

Has your friend's privacy been violated?

Yes!

The teacher has disclosed your friend's exam score to the entire class

Releasing statistics can tell us information about individuals

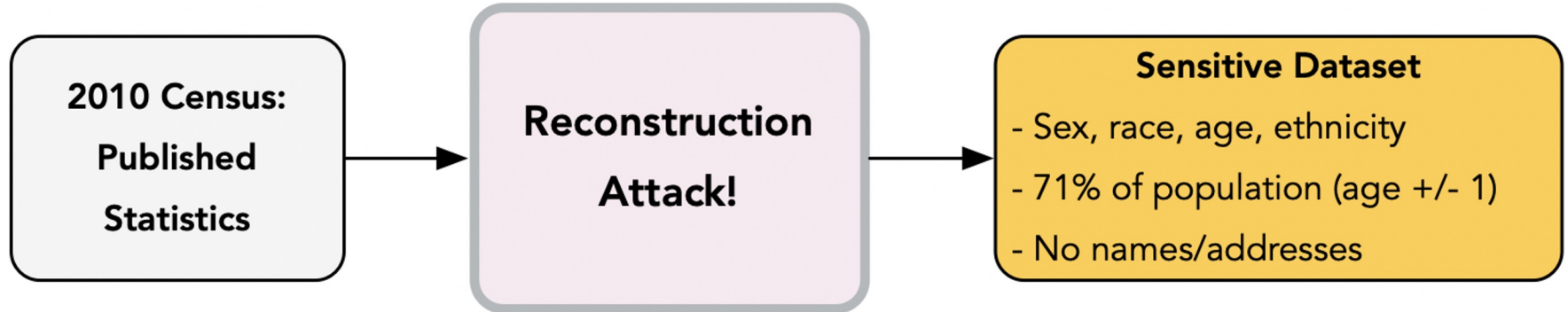
Exam Average:

~~85%~~

84%

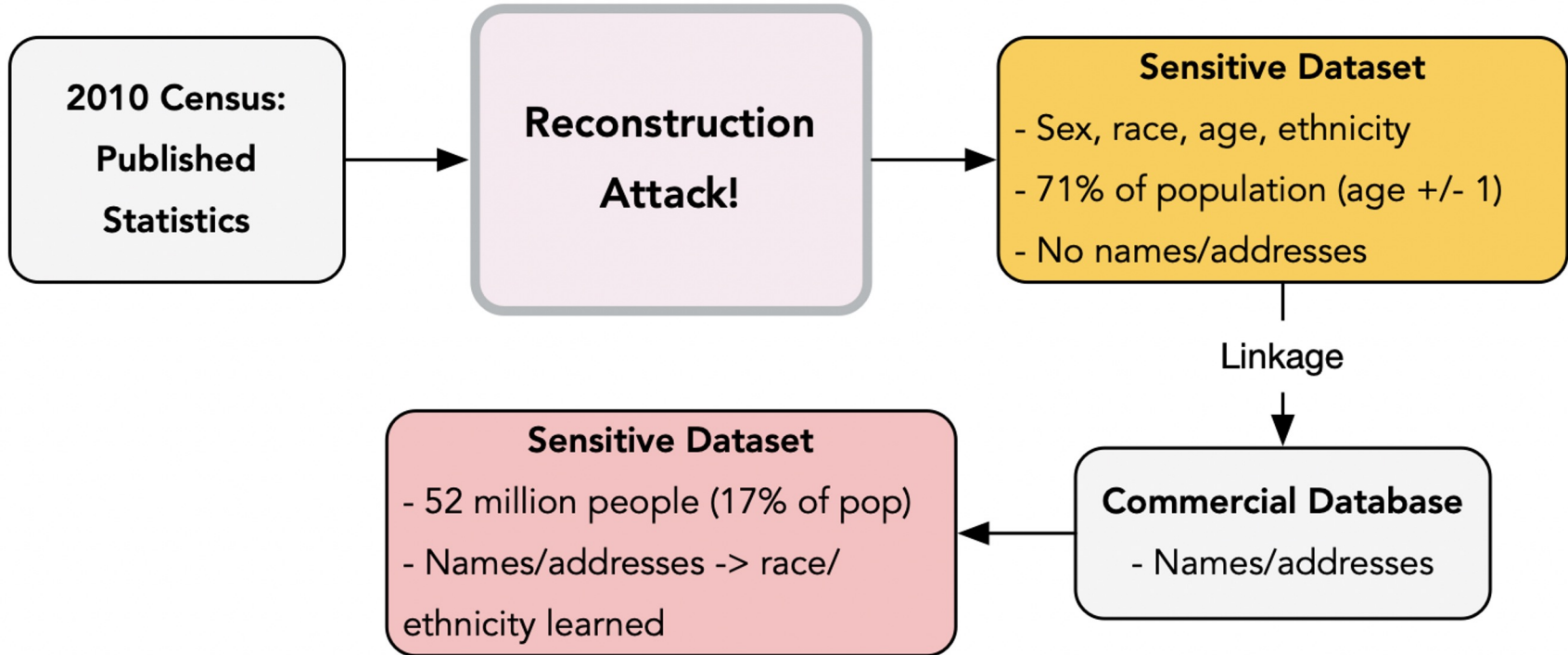
Implications: 2010 Census

- 308,745,538 individuals



Source: [John M. Abowd, Chief Data Scientist, U.S. Census, June 12, 2019, ICML Conference](#)

Implications: 2010 Census



Source: [John M. Abowd, Chief Data Scientist, U.S. Census, June 12, 2019, ICML Conference](#)

How could this be avoided?

Using letter grades (less precise)

Not updating value (less dynamic)

Exam Average:

85%

Or... What if we add noise?

Teacher calculates average,
adds a random number

Random number might be
negative, might be positive

Tells the class that a random
number was added, but doesn't
tell students what the number is

Exam Average:

85.5%

What if we add noise?

Teacher calculates average,
adds a random number

Random number might be
negative, might be positive

Tells the class that a random
number was added, but doesn't
tell students what the number is

Exam Average:

~~85.5%~~

84.2%

Has your friend's privacy been violated?

$$\frac{Q + R + S + T + X}{5} = 85.5 = Y + \text{noise1}$$

$$\frac{Q + R + S + T}{4} = 84.2 = Z + \text{noise2}$$

$$Q + R + S + T + X = 85.5 * 5 = (Y + \text{noise1}) * 5$$

$$Q + R + S + T = 84.2 * 4 = (Z + \text{noise2}) * 4$$

Exam Average:

~~85.5%~~

84.2%

Has your friend's privacy been violated?

$$\frac{Q + R + S + T + X}{5} = 85.5 = Y + \text{noise1}$$

$$\frac{Q + R + S + T}{4} = 84.2 = Z + \text{noise2}$$

$$Q + R + S + T + X = 85.5 * 5 = (Y + \text{noise1}) * 5$$

$$Q + R + S + T = 84.2 * 4 = (Z + \text{noise2}) * 4$$

You can't reverse engineer your friend's grade!

Privacy has been preserved

Exam Average:

~~85.5%~~

84.2%

Differential Privacy: Accuracy / Privacy Balance

Adding noise adds uncertainty to the final value

Final value is still useful

Cannot be used to learn anything about individuals

Exam Average:

~~85.5%~~

84.2%

Differential Privacy

Seeing mean of 85 vs 84 **gives** us **information** about your friend's grade

Seeing **DP-mean** of 85.5 vs 84.2 gives us information about your friend's *possible grades*

The distance between these possible distributions should be small so that we can't effectively guess the true value

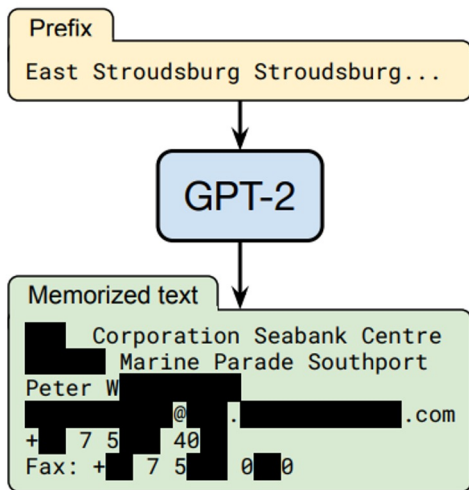
Epsilon - defines how much noise to add

Exam Average:

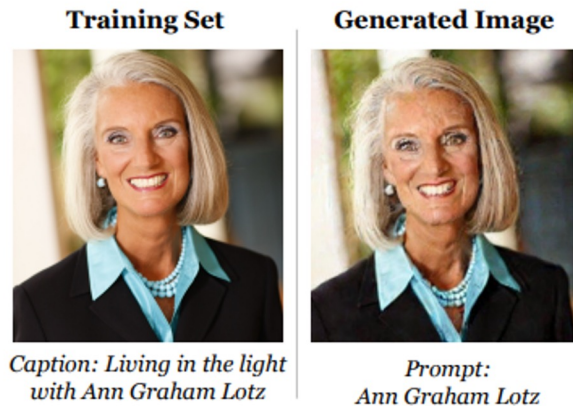
~~85.5%~~

84.2%

Extracting Training Data from ML/LLM Models



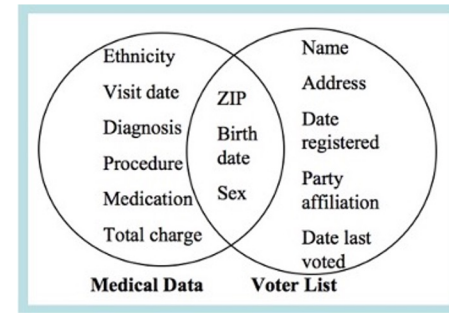
[Carlini, Tramèr, Wallace et al. 2021]



[Carlini, Hayes, Nasr et al. 2023]

Typical attacks on privacy

- **Re-identification:** determining *who-is-who* even after “PII” removed
 - Applied to medical data [Sweeney `97, Teague et al.`16], Netflix challenge [Narayanan-Shmatikov `08], ...
- **Database Reconstruction:** reconstructing almost the entire underlying dataset [Dinur-Nissim `03,...]
 - Applied to Census releases [Garfinkel et al. `18] and Diffix [Cohen-Nissim `19].
- **Membership Inference:** determining whether a target individual is in the dataset [Dwork-Smith-Steinke-Ullman-V. `15]
 - Applied to genomic data [Homer et al. `08,...] and ML as a service [Shokri et al. `17,...].



[Sweeney `97]

Attacks on
“Aggregate”
Statistics

Why Differential Privacy?

✘ Releasing statistics leaks privacy

✘ Releasing lots of statistics leaks lots of privacy.

- [2010 census was reverse engineered](#)

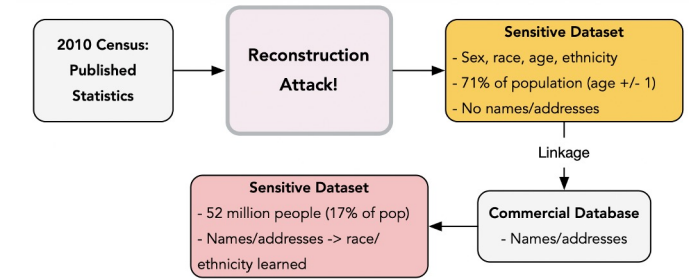
✘ Traditional protection techniques don't always work

- Harvard-MIT EdX public dataset of student stats. Used k-anonymity; vetted by experts.
- BUT [researchers connected stats of people who failed an EdX course w/ LinkedIn profiles](#)

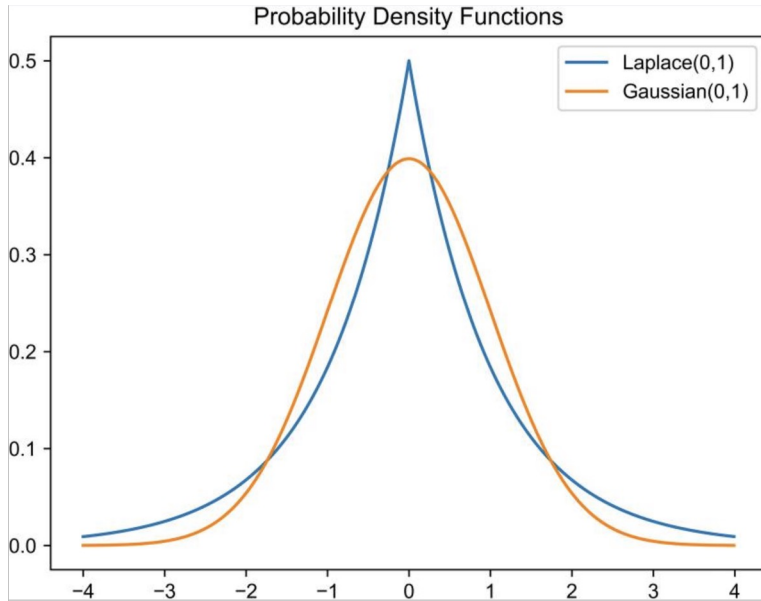


Differential Privacy adds calibrated noise to statistics to protect individuals

- You cannot tell if any one individual was in the dataset but the statistics are still useful

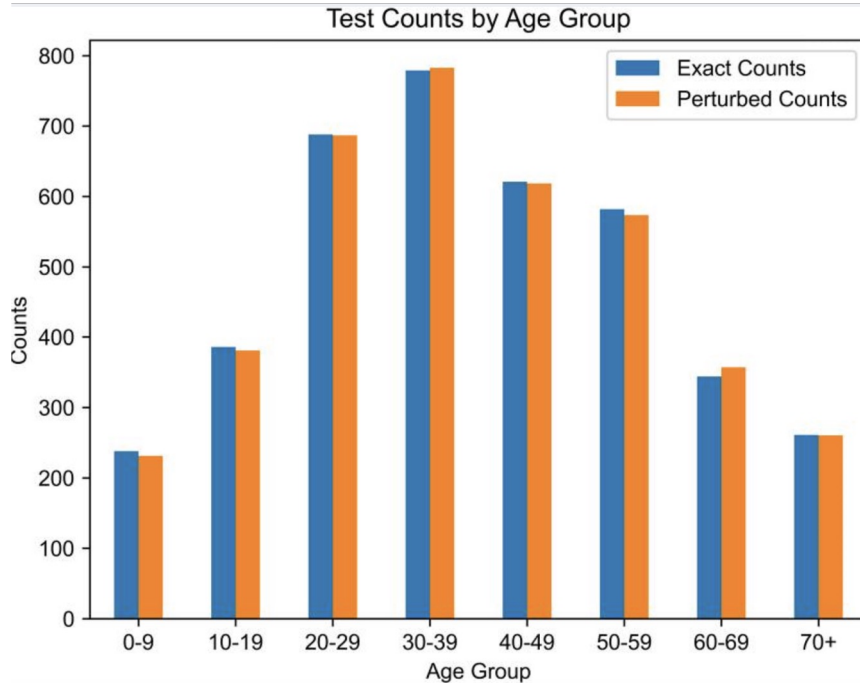


How Differential Privacy works in practice ?



Differential privacy adds a **Laplace** (instead of Gaussian) **noise** to statistics: heavy tails and high peak at zero, i.e. perturbed data far from true value (compared to Gaussian noise)

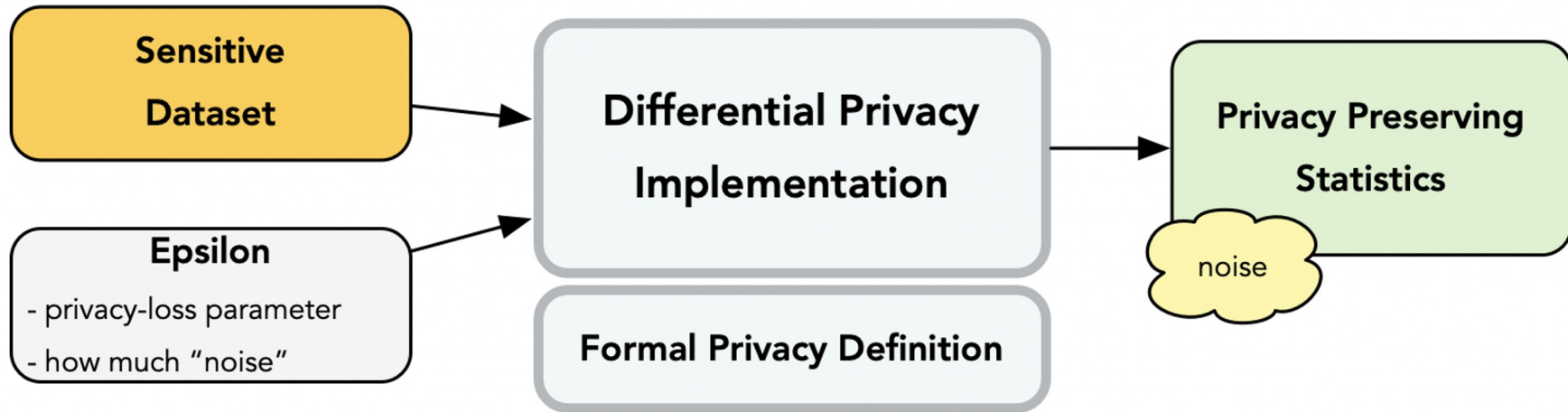
How Differential Privacy works in practice ?



On histograms: adding independent noises to each bin protects privacy.

good news: the scale of the noise does not depend on the number of bins.

Differential Privacy: “Safe-to-Share” Stats



Mathematical bases of ϵ -Differential Privacy

The 2006 Dwork, McSherry, Nissim and Smith article introduced the concept of **ϵ -differential privacy**.

Differential privacy is designed to give each individual roughly the same privacy that would result from having their data removed.

Individual privacy depends on the sample size. The less the observations, the more the privacy is at risk.

An algorithm \mathcal{A} is ϵ -DP if, given two data sets D_1 and D_2 , where D_1 and D_2 differ only by 1 row, the probability of having some statistical result $\mathcal{A}(D_1)$ is the same as $\mathcal{A}(D_2)$ but with

$$\Pr[\mathcal{A}(D_1) \in \mathcal{S}] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in \mathcal{S}]$$

The (Inherent) Privacy-Utility Trade-off

Every statistical release incurs some privacy loss ϵ_i .

- More noise \Rightarrow more privacy (smaller ϵ_i), less accuracy
- Less noise \Rightarrow less privacy (larger ϵ_i), more accuracy
- Tradeoff is less stark on larger populations ($n \rightarrow \infty$)

With multiple queries, the privacy loss accumulates.

- Overall privacy loss $\leq \epsilon_1 + \epsilon_2 + \dots + \epsilon_k$

There are better composition theorems for differential privacy.

[Dwork-Rothblum-V. 09, Kairouz-Oh-Viswanath '15, Murtagh-V. '16, ...]

Recommended use: set an overall budget ϵ (e.g. $\epsilon = .1$)

Stop answering queries when budget reached.



[OpenDP Library v0.6 Released!](#) | [Application Open for the 2023 Fellows Program](#) | [We Are Hiring](#)

Developing Open Source Tools for Differential Privacy

OpenDP is a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data. These tools, which we call OpenDP, will offer the rigorous protections of [differential privacy](#) for the individuals who may be represented in confidential data and statistically valid methods of analysis for researchers who study the data.

Join Us on [Slack](#), [Github](#), [Mailing List](#)!

[Learn more about us](#)



<http://opendp.org/>

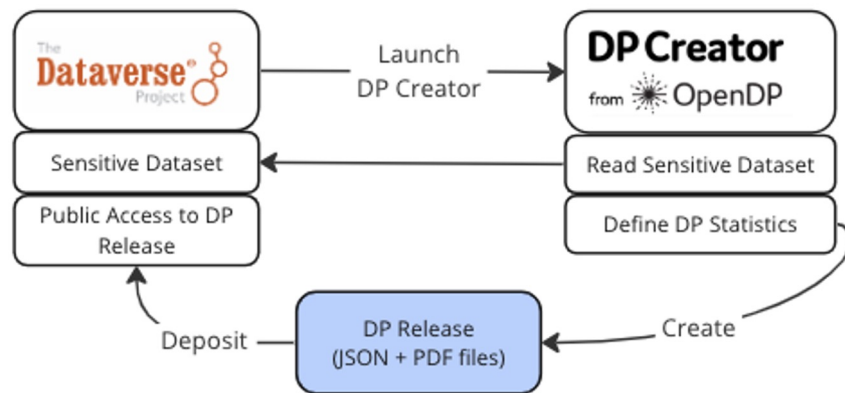
DP and Dataverse: The DPCreator Application

Goal: Allow non-experts to use Differential Privacy (via the OpenDP Library)

- Connect to social science repositories (**Dataverse**)
- Guided user experience
- Initial versions created/tested

Next steps

- Integrating user studies and feedback
- “Desktop”/client side version
 - Sensitive data stays with the user
- “Learning Mode” to further help users



DP Creator is web application that uses the OpenDP library to produce DP statistics of the data without writing any code and depositing it to **Dataverse**. The application incorporates step-by-step instructions to guide new users in learning the basics of tuning parameters to generate useful DP statistics.

Through a series of questions, the user is helped in the process of creating privacy guaranteed summaries of the data.

Create Statistic

Which **single-variable statistic** would you like to use?

✓ Mean

Histogram

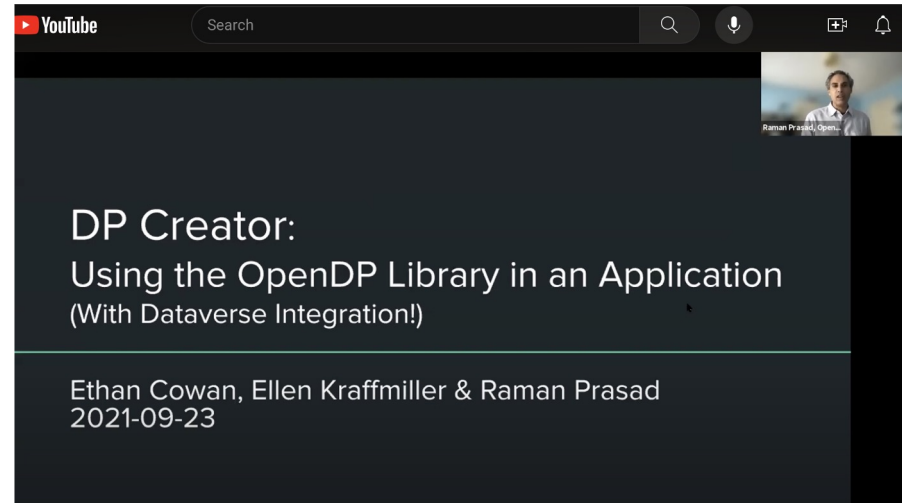
Count

Variance

Which **variable** would you like to use? (Need to add another variable? Go back to Confirm Variables Step)

✓ age

income

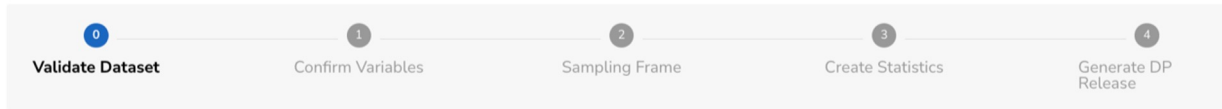


[Youtube video of DPCreator](#)

DP Creator: a Tool for Non-Experts

DP Creator
from  OpenDP

 My Data  My Profile  Logout



Used data file: [Teacher Survey](#) 

Validate Data File

Confirm the data file's characteristics to determine if it's adequate for the differential privacy release process.


▶ Does your data file depend on private information of subjects?

- Yes.
- No.
- I'm unsure.

▶ Which of the following best describes your data file?

- Public information. (Note: Differential privacy isn't needed for public information.)
- Information that, if disclosed, **would not cause material harm**, but which the organization has chosen to **keep confidential**.
- Information that **could cause risk of material harm** to individuals or the organization if disclosed.
- Information that **would likely cause serious harm** to individuals or the organization if disclosed.

Creating Statistics

DP Creator
from  OpenDP

My Data My Profile Logout

Edit Statistic [X]

Which **single-variable statistic** would you like to use?

Mean Histogram Count Variance

Which **variable** would you like to use? (Need to add another variable? [Go back to Confirm Variables Step](#))

age sex smoking optimism selfesteem
 Havingchild maritalstatus sourceofstress
 lifesatisfaction highesteducationlevel

Enter a **fixed value** for missing values:
(Must be between 1 and 6)

Statistic	Variable
1	Histogram maritalstatus
2	Mean age
3	Histogram highesteduc
4	Histogram sourceofstre
5	Mean optimism
6	Count age

Changing the epsilon, delta, or accuracy. Splitting the budget e
[Privacy Budgeting and Epsilon](#)

between epsilon value and
tget. [\(More Information about](#)

OpenDP

A **community effort** to build a **trustworthy** and **open-source** suite of differential privacy tools that can be **easily adopted** by custodians of sensitive data to make it available for **research and exploration** in the public interest.

Why?

- Channel our collective advances on science & practice of DP
- Enable wider adoption of DP
- Address high-demand, compelling use cases
- Provide a starting point for custom DP solutions
- Identify important research directions for the field

OpenDP Leadership



Gary King
Faculty Director



Salil Vadhan
Faculty Director



James Honaker
Chief Privacy
Engineer



Stefano Iacus
Director of Data
Science, IQSS



Andrew Vyrros
Senior Library
Architect



Annie Wu
Program Director



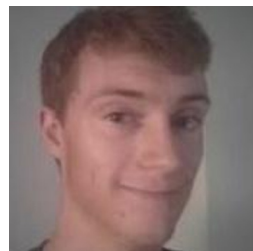
Sharon Ayalde
Director of
Community
Engagement



Lindsay Froess
Project
Coordinator



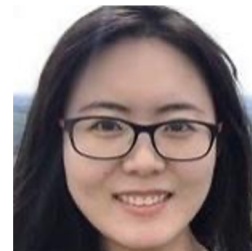
Raman Prasad
Technical Lead for
Research Software



Michael Shoemate
Senior Software
Developer

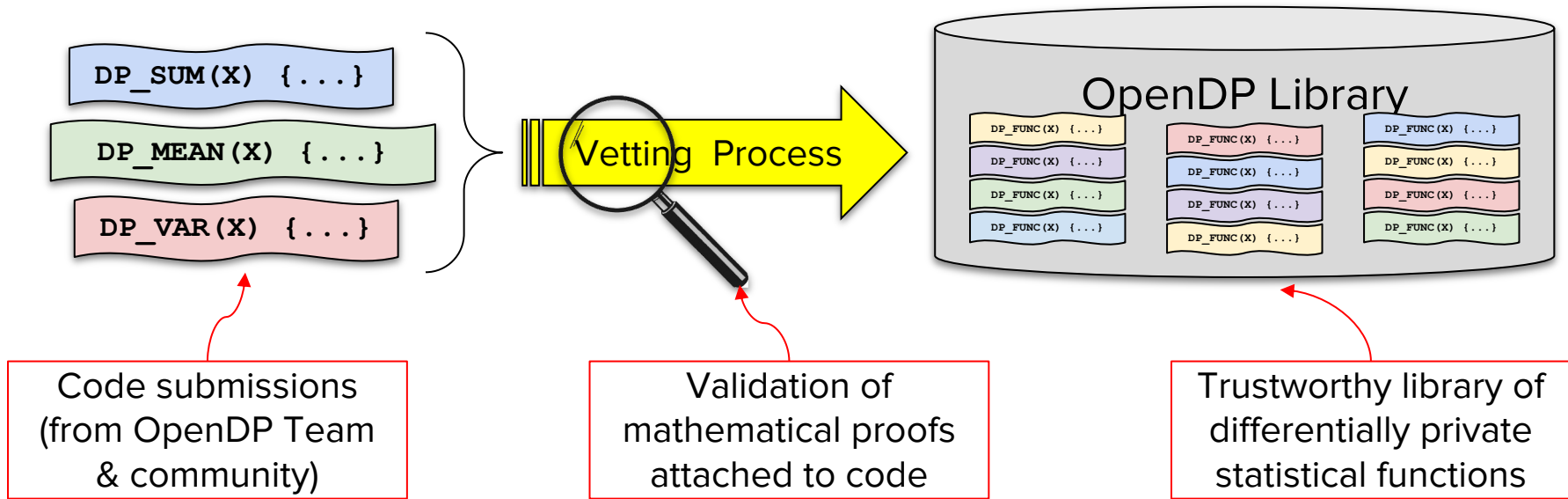


Vikrant Singhal
Research
Associate



**Wanrong
Zhang**
Research
Associate

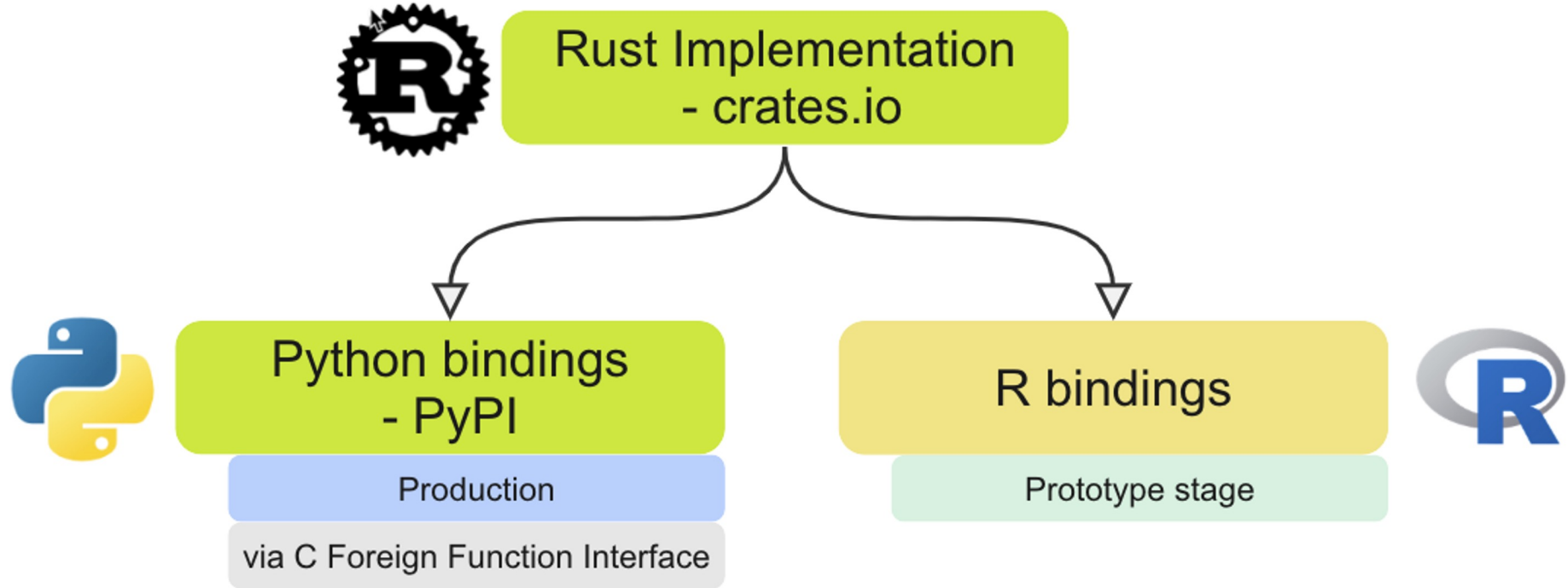
Building a Trustworthy Library



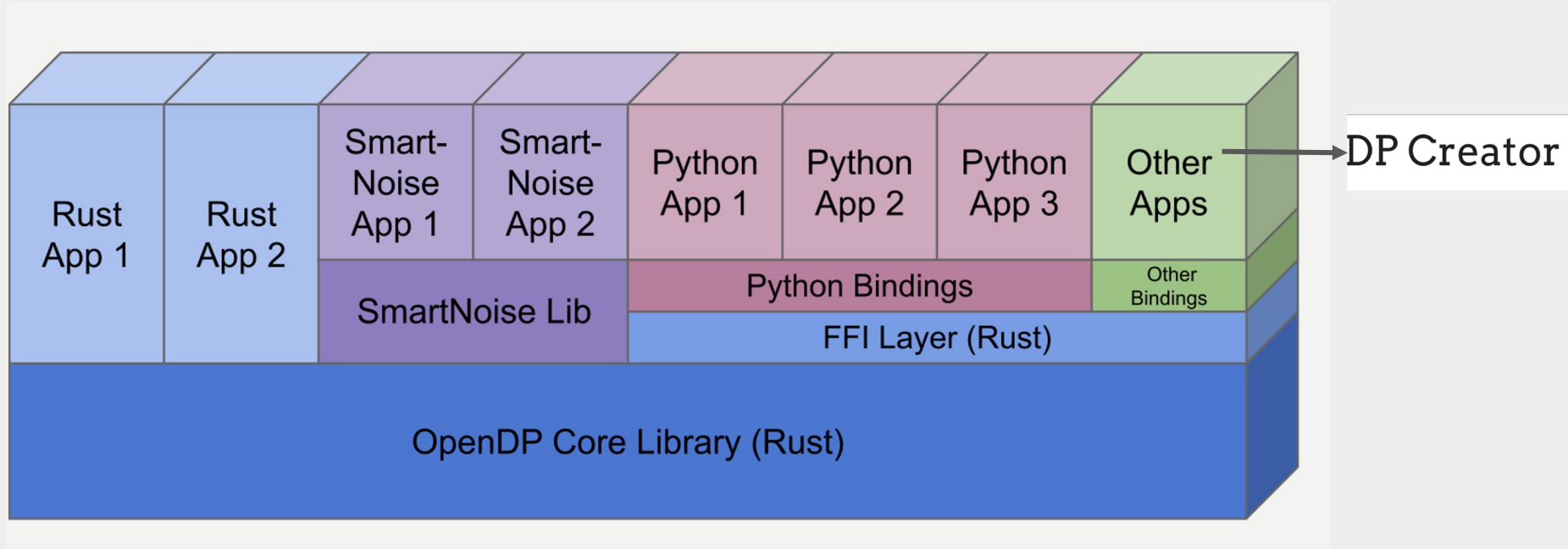
Use Cases

- Archival data repositories (e.g. [Dataverse](#), ICPSR, Zenodo) enabling secondary reuse and replication.
- Government agencies making data available to the public, both for official statistics and open data mandates.
- Data for good programs at companies, sharing data on customers with public and researchers
- Analytics on customer data, internally & with partners
- Machine learning on customer data

The OpenDP Library v0.8



The OpenDP Library ecosystem





Thanks!



*With contribution by & credits to the
teams of*

The
Dataverse[®]
Project 

 OpenDP

siacus@iq.harvard.edu