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

HARVARD Dataverse
IQSS The Institute for Quantitative Social Science
OpenDP
HARVARD UNIVERSITY Information Technology
The Dataverse® Project

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	Computer and Information Science 3,798	Medicine, Health and Life Sciences 10,548
Arts and Humanities 36,716	Earth and Environmental Sciences 9,554	Physics 1,760
Astronomy and Astrophysics 1,350	Engineering 2,292	Social Sciences 64,287
Business and Management 2,341	Law 5,849	
Chemistry 955	Mathematical Sciences 722	

IQSS

What is Dataverse?

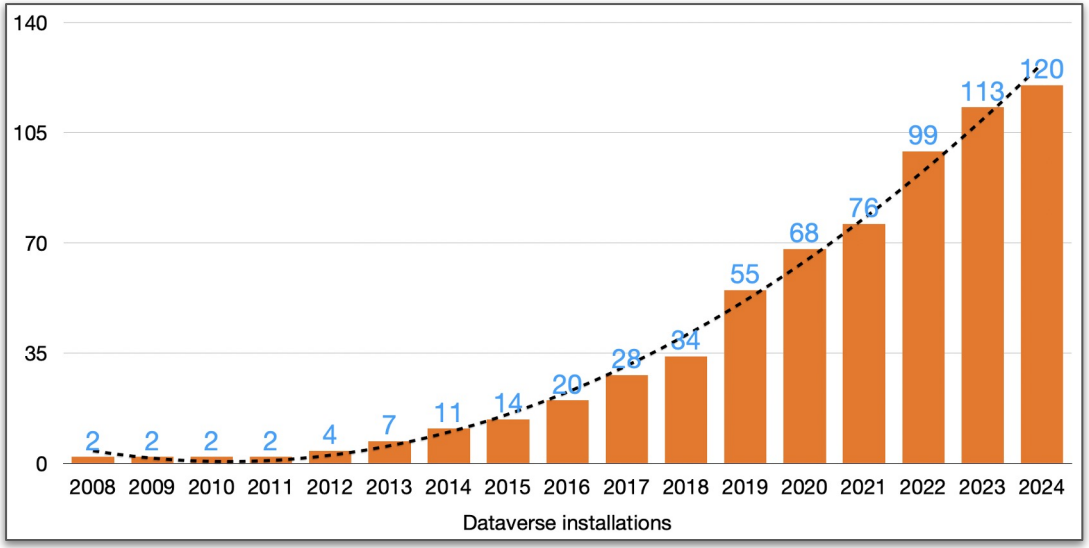
An **open-source** platform that provides a **generalist** repository to **publish**, **cite**, and **archive research data**

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	Computer and Information Science 3,798	Medicine, Health and Life Sciences 10,548
Arts and Humanities 36,716	Earth and Environmental Sciences 9,554	Physics 1,760
Astronomy and Astrophysics 1,350	Engineering 2,292	Social Sciences 64,287
Business and Management 2,341	Law 5,849	
Chemistry 955	Mathematical Sciences 722	

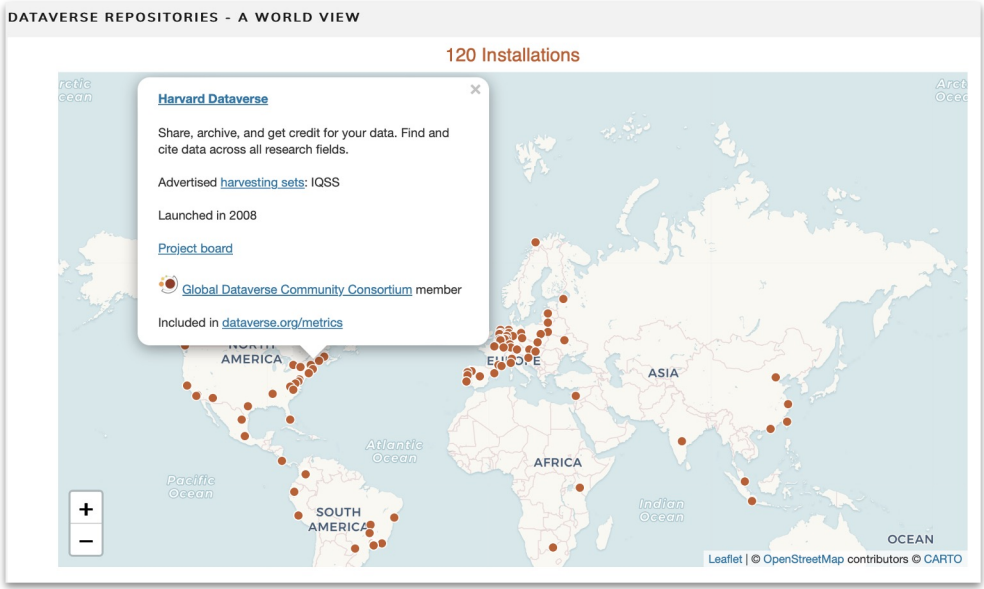
IQSS

What is Dataverse?

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

Soon at BSC ?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Follow this link if you want to know more about Dataverse and its history

IQSS

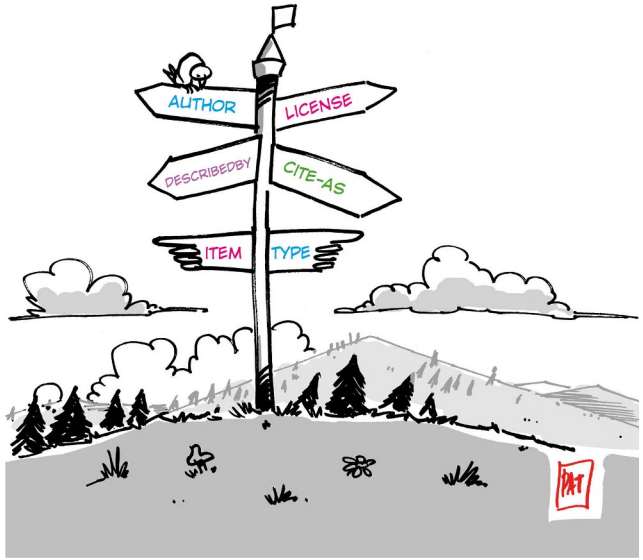
Signposting and Discoverability @ Dataverse

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

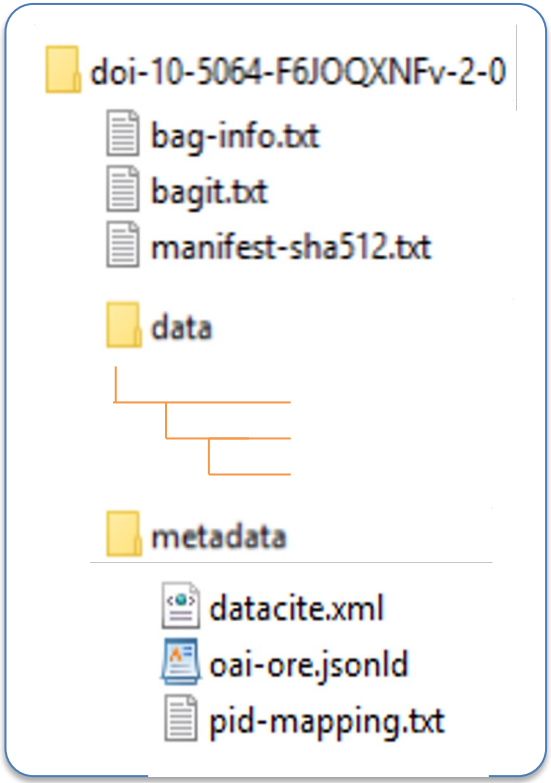
Higher granularity of information

IQSS

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).

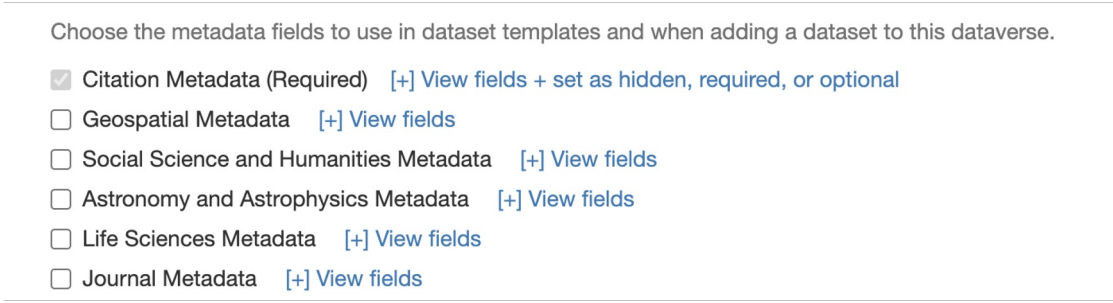
It is meant to be able to export/encode **ALL** of the metadata available for a dataset.

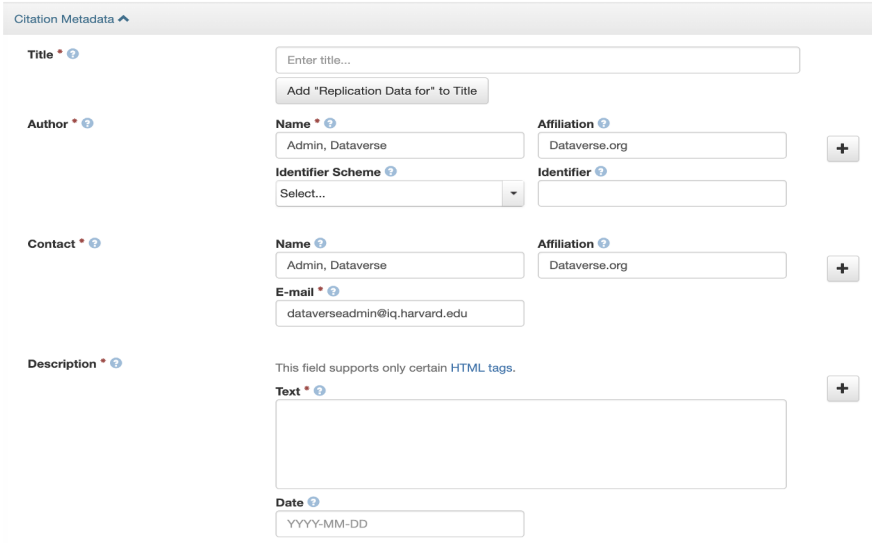
it is used in Dataverse APIs for developing export plugins, as one of the carriers of metadata that the plugin code can further manipulate.

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





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.

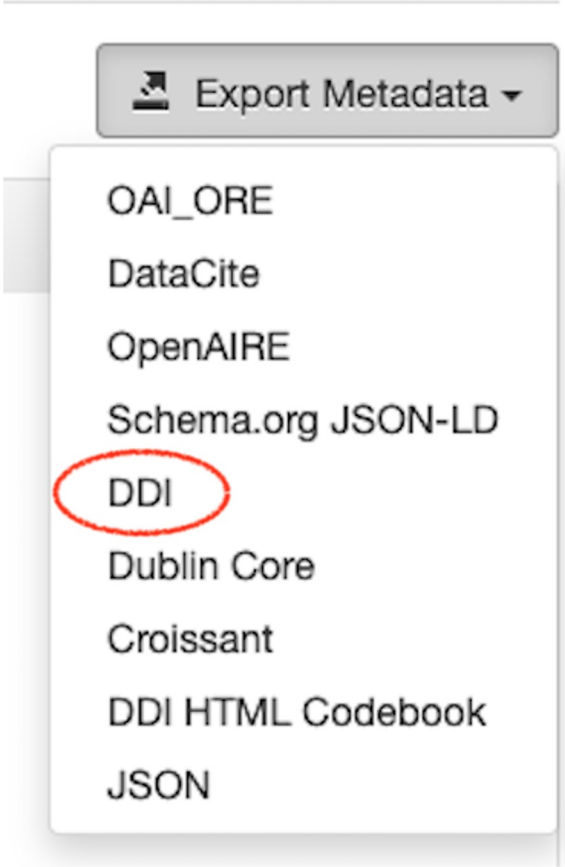
DDI Codebook metadata

```
…
<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.

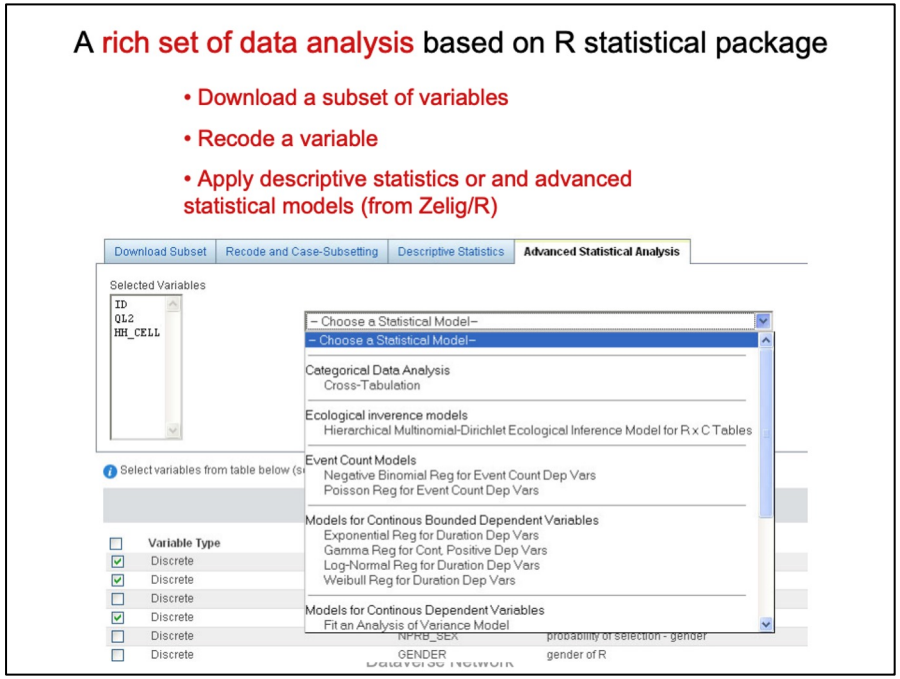
Exporters can be added as plug-ins



Use of the DDI Codebook Metadata

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

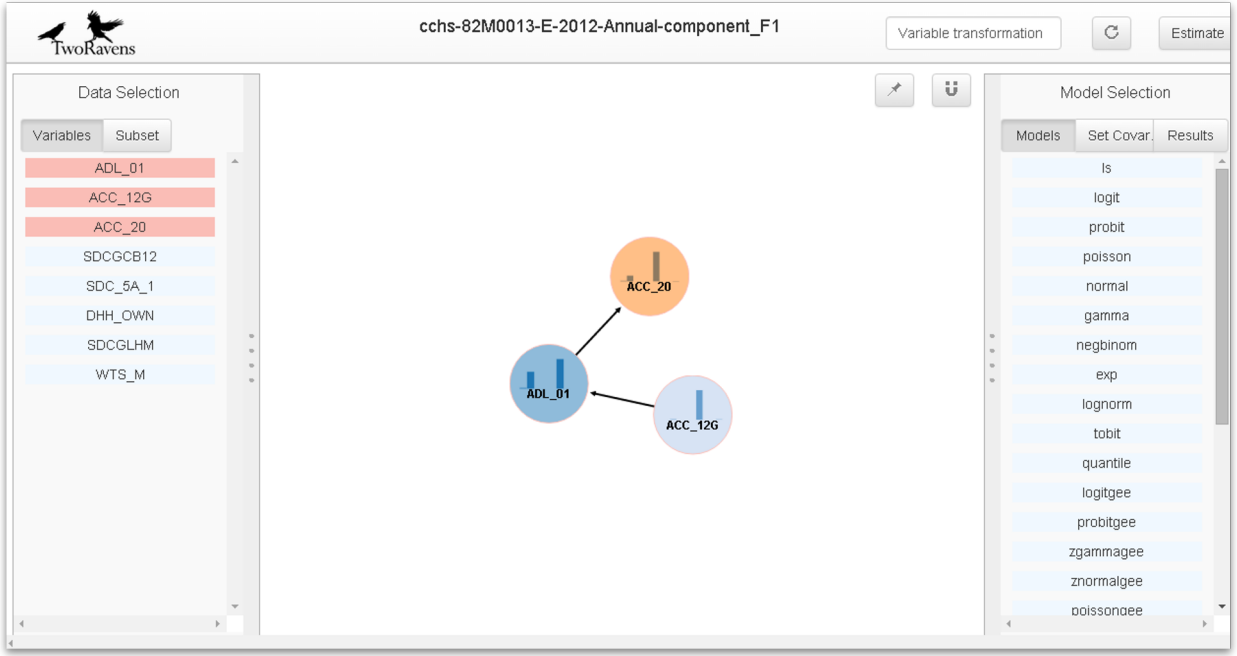
from an early version (v3.0):



Use of the DDI Codebook Metadata

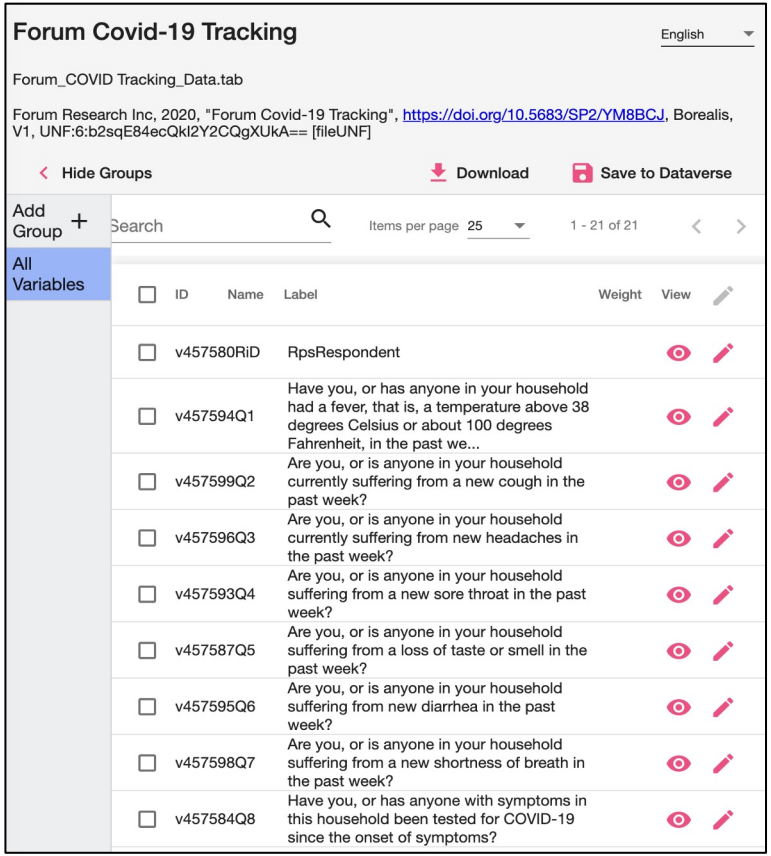
Integration with **Two Ravens**

v.4.0 (ca 2018):



Use of the DDI Codebook Metadata

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



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 | Ac

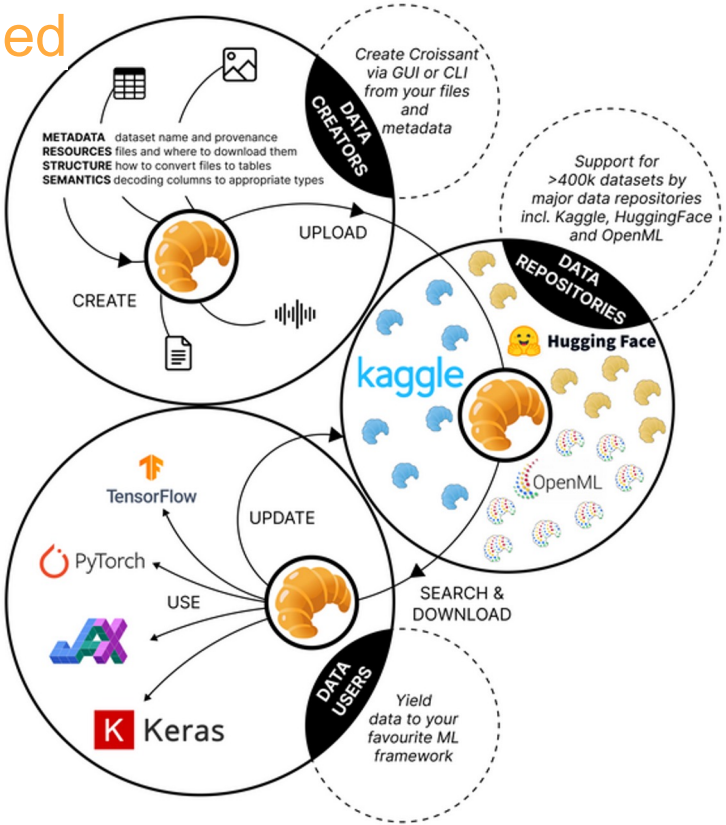
Computer Science > Machine Learning

[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-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry 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

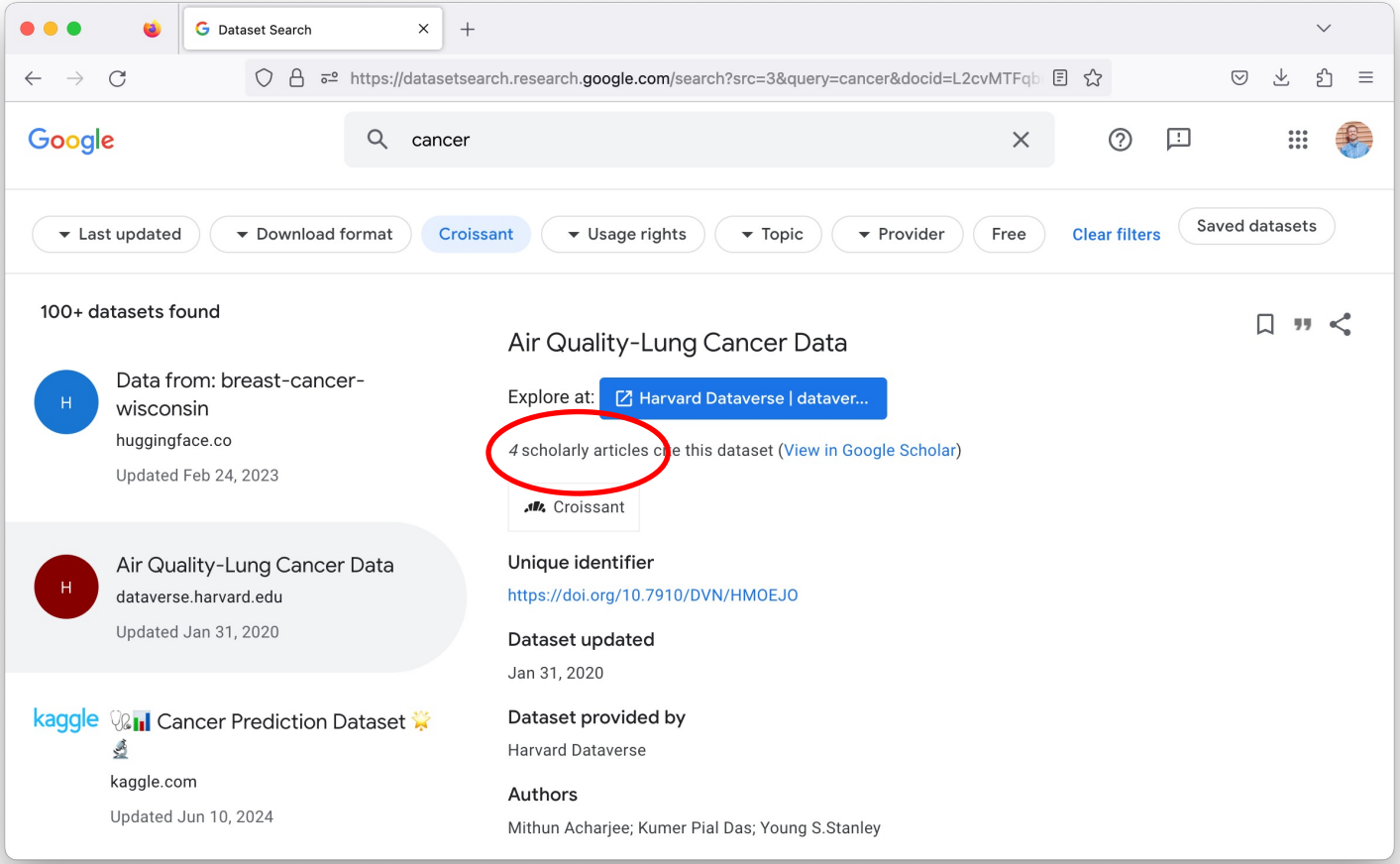
IQSS/DV contributed to definition of this standard for ML/AI workflows. HDV exposes CroissantML metadata to crawlers to increase discoverability of this type of data and has a built-in exporter.

IQSS

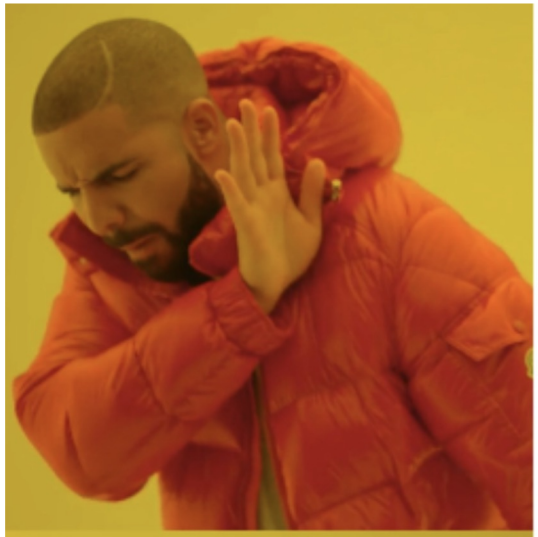
Dataverse dataset now being exposed to the ML/CS community

Thanks to Croissant-ML support.

Now in production.



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…

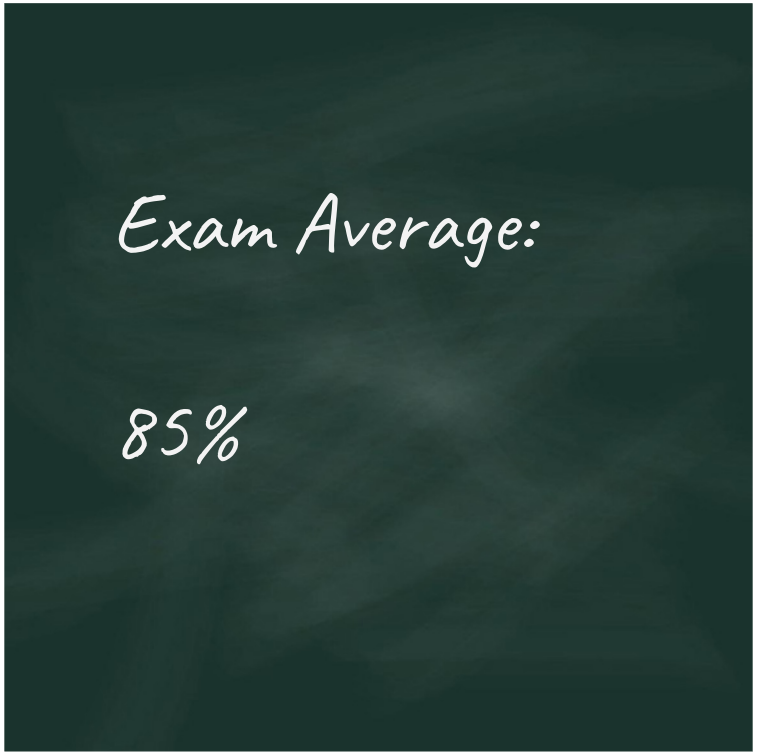
“If you publish too many statistics and too accurately, ***no one,*** *in the input database,* ***can be given any reasonable assurance of the confidentiality*** *of their input data.”*

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%

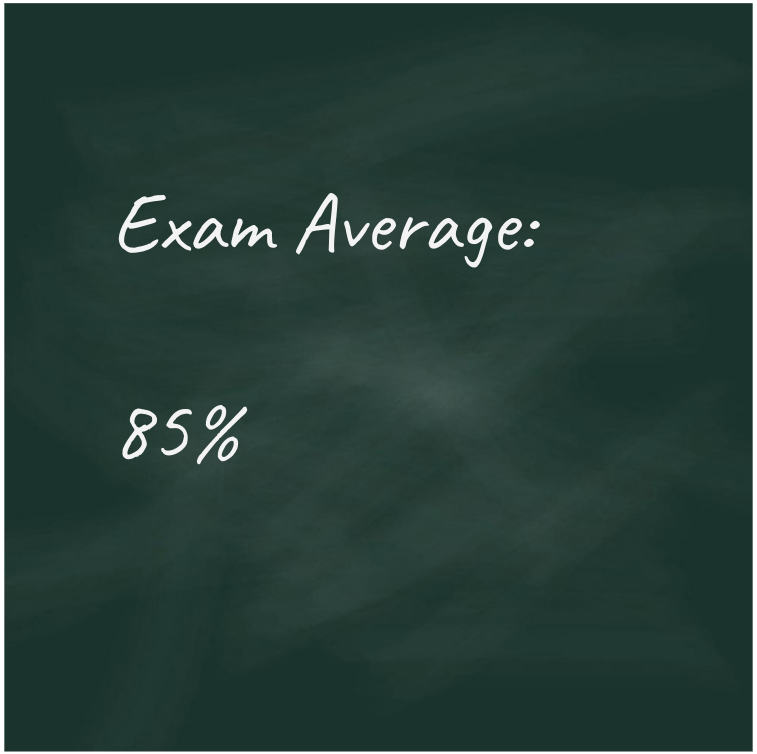


(Example courtesy of Ethan Cowan, former OpenDP team member and author of upcoming O'Reilly book, Hands-On Differential Privacy: Introduction to the Theory and Practice Using OpenDP)

Sharing Data

Class has 5 students

Your friend joins the circus and drops the class

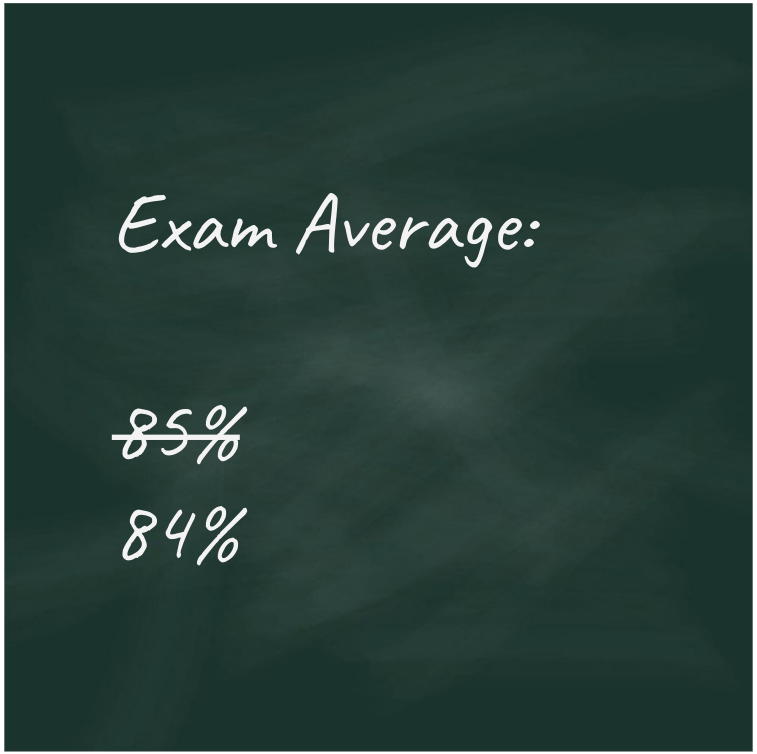


Sharing Data

Class has 5 students

Your friend joins the circus and drops the class

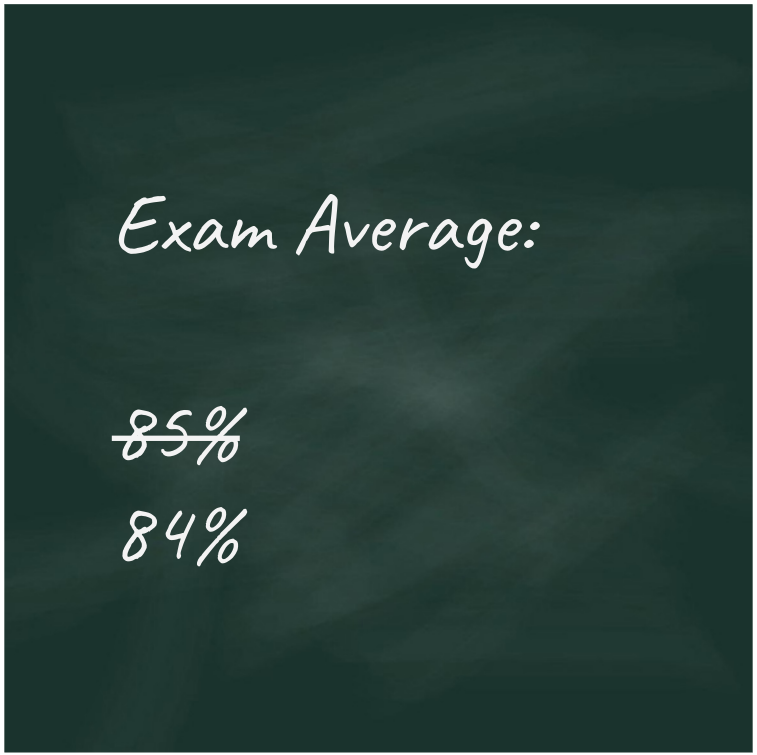
Has your friend's privacy been violated?



Has your friend's privacy been violated?

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

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



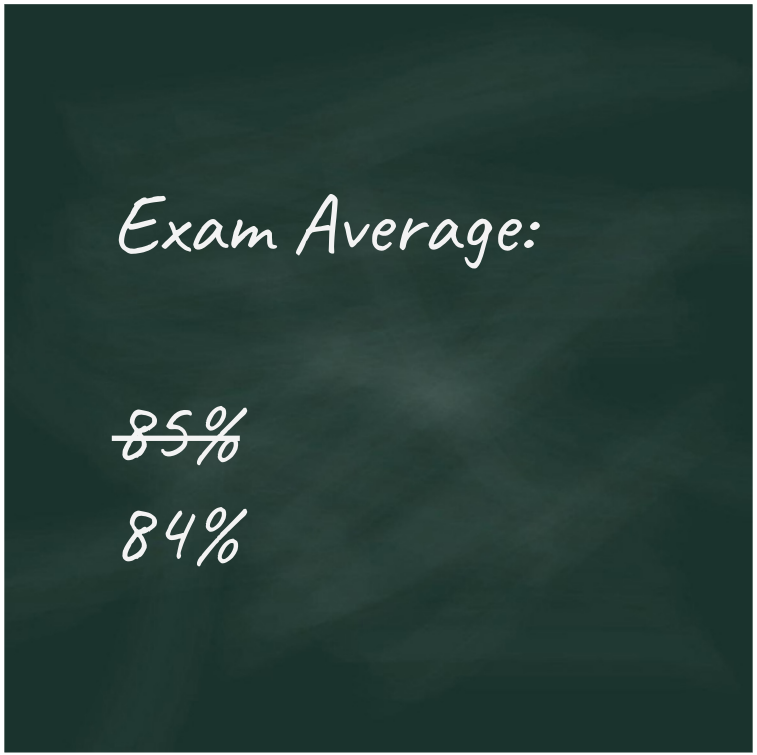
Has your friend's privacy been violated?

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

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

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

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



Has your friend's privacy been violated?

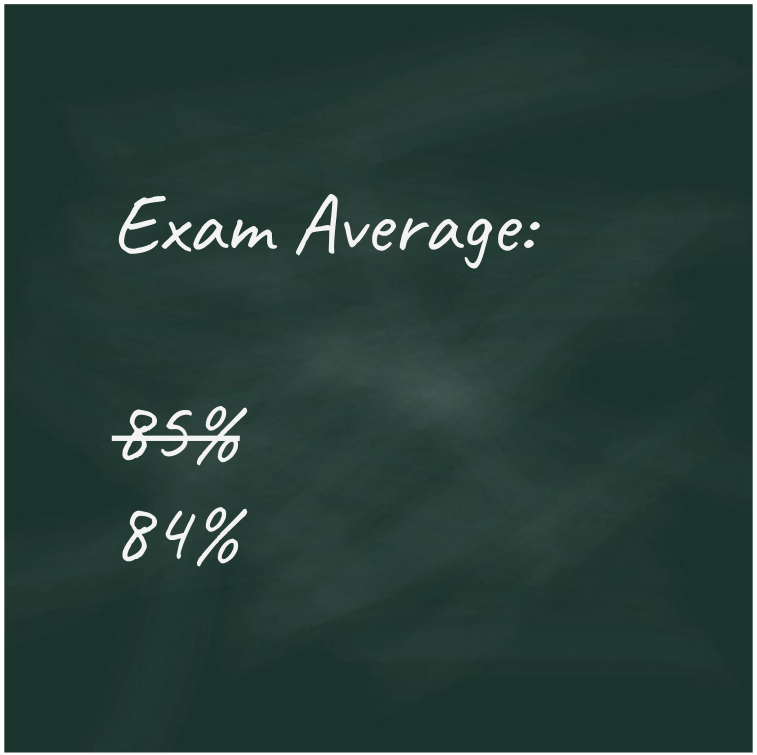
$$\frac{Q + R + S + T + X}{5} = 85$$

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

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

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

X = 425 - 336 = 89 ⇐ your friend's score !

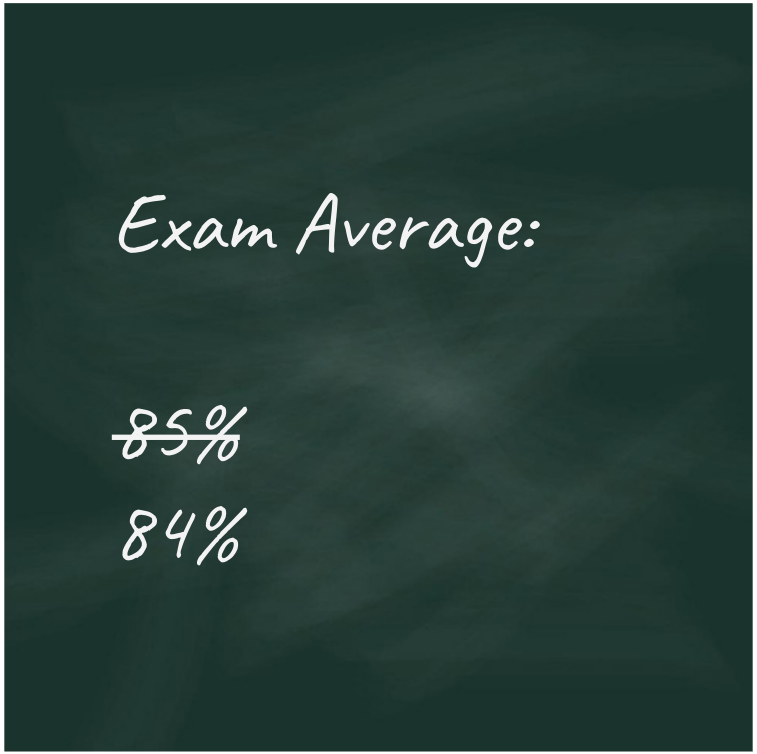


Has your friend's privacy been violated?

Yes!

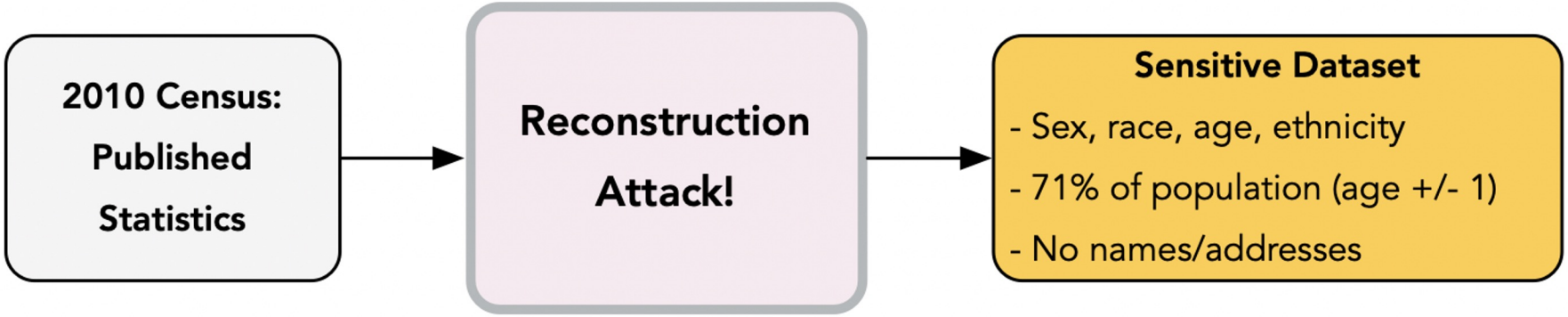
The teacher has disclosed your friend's exam score to the entire class

Releasing statistics can tell us information about individuals



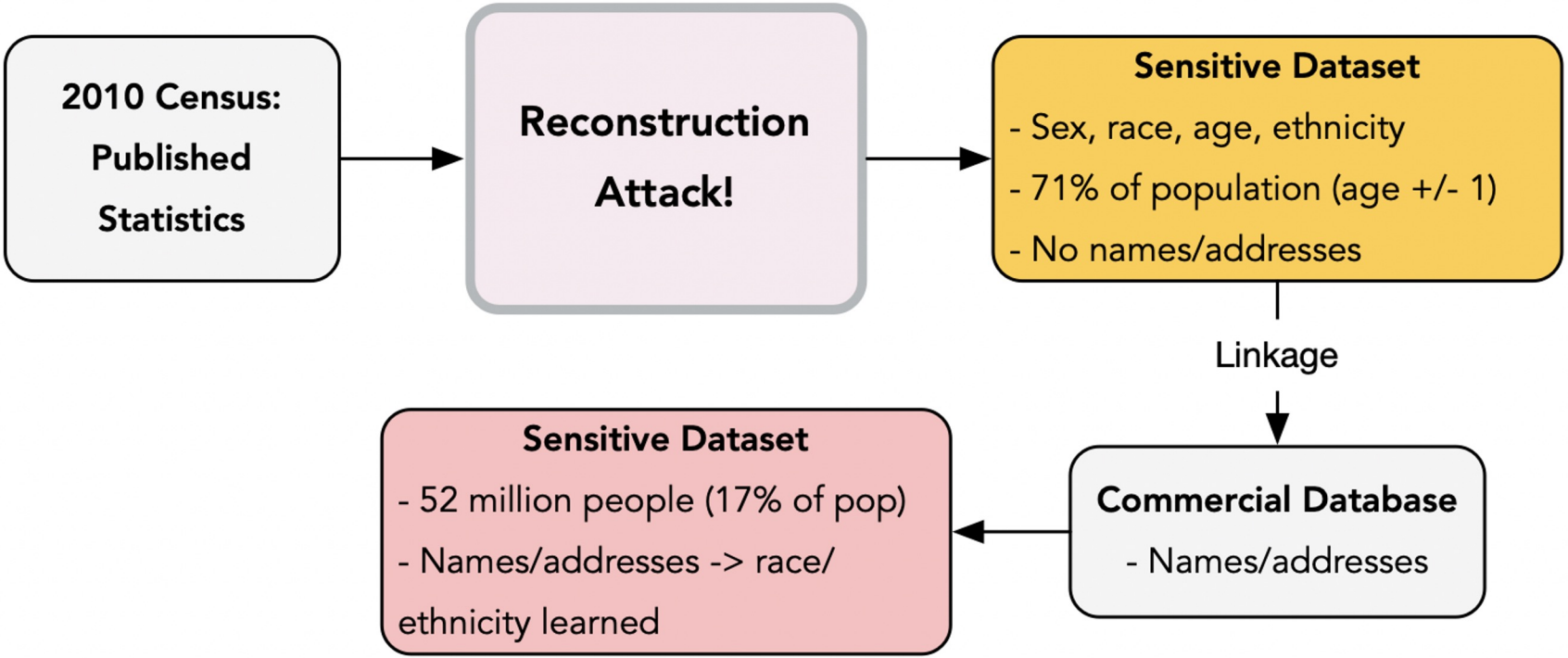
Implications: 2010 Census

- 308,745,538 individuals



Source: John M. Abowd, Chief Data Scientist, U.S. Census, June 12, 2019, ICML Conference

Implications: 2010 Census

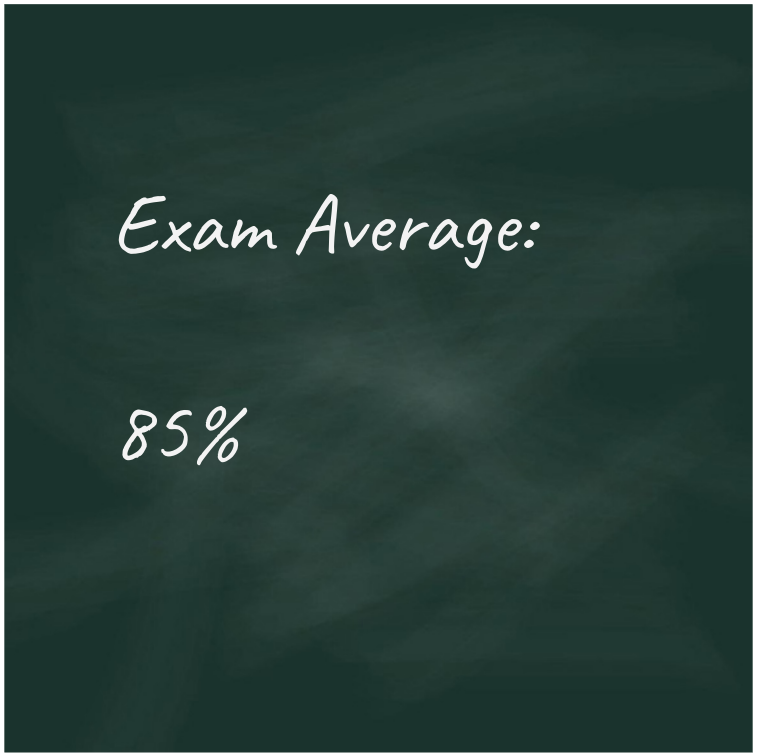


Source: John M. Abowd, Chief Data Scientist, U.S. Census, June 12, 2019, ICML Conference

How could this be avoided?

Using letter grades (less precise)

Not updating value (less dynamic)

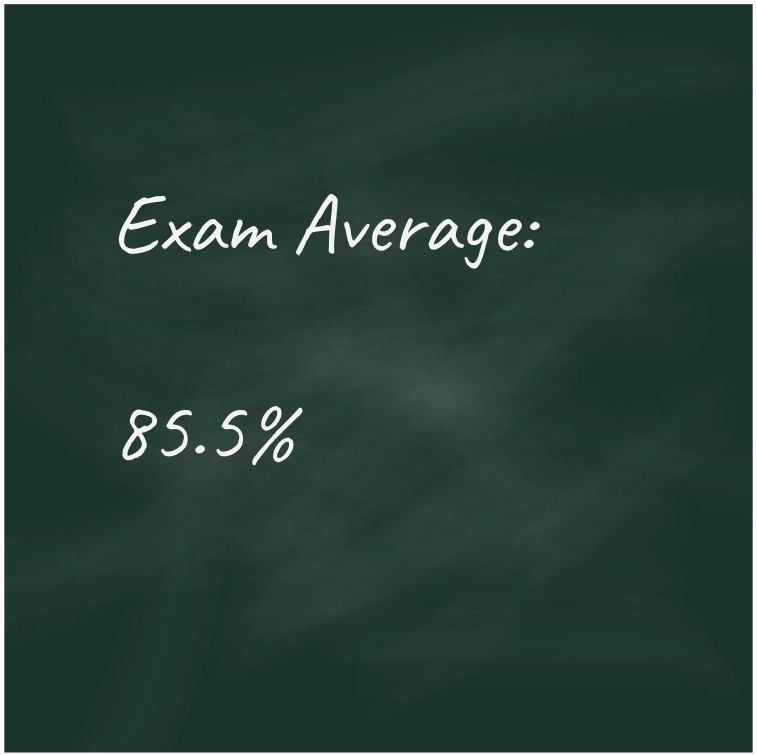


Or… What if we add noise?

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

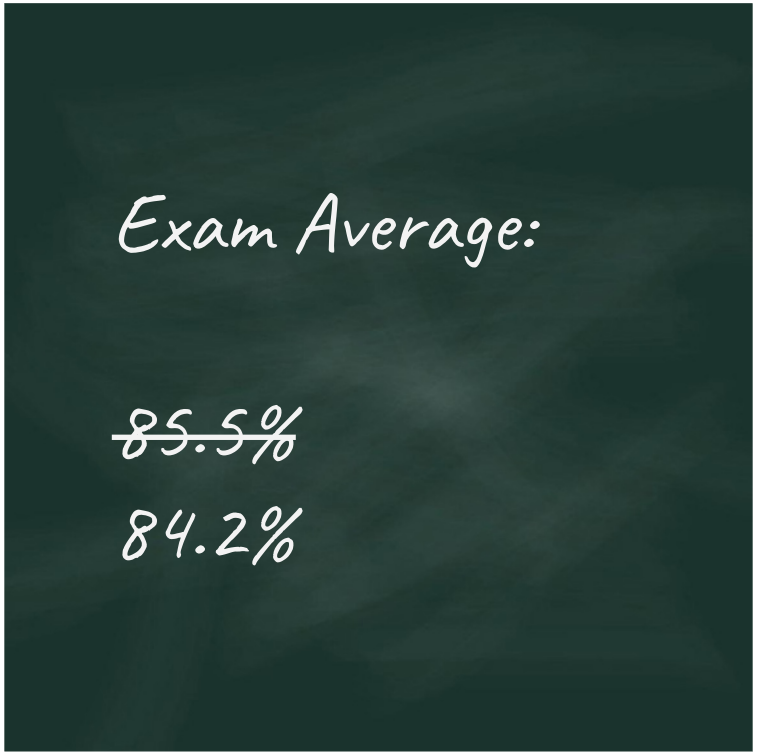


What if we add noise?

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



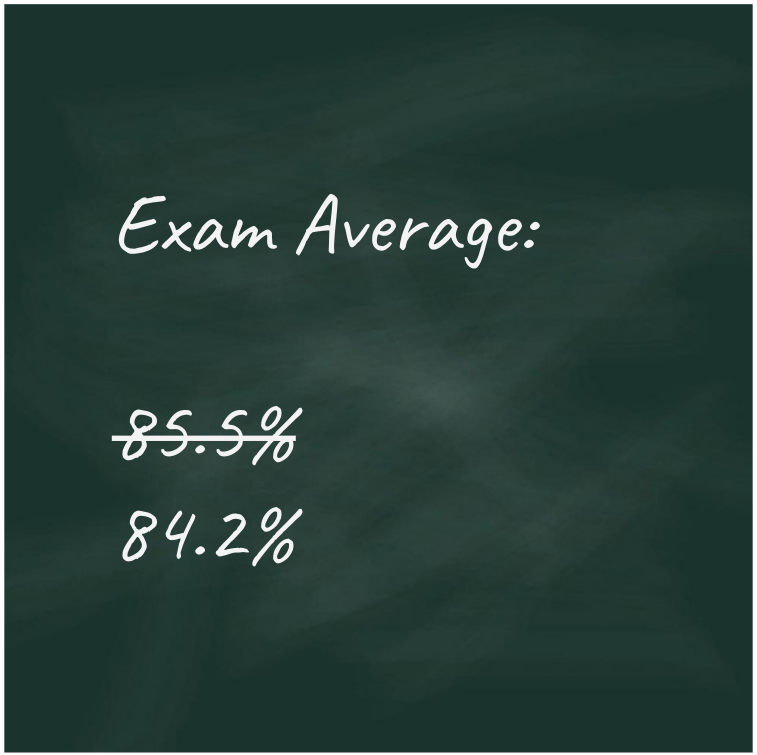
Has your friend's privacy been violated?

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

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

$$Q + R + S + T + X = 85.5 * 5 = (Y + noise1) * 5$$

$$Q + R + S + T = 84.2 * 4 = (Z + noise2) * 4$$



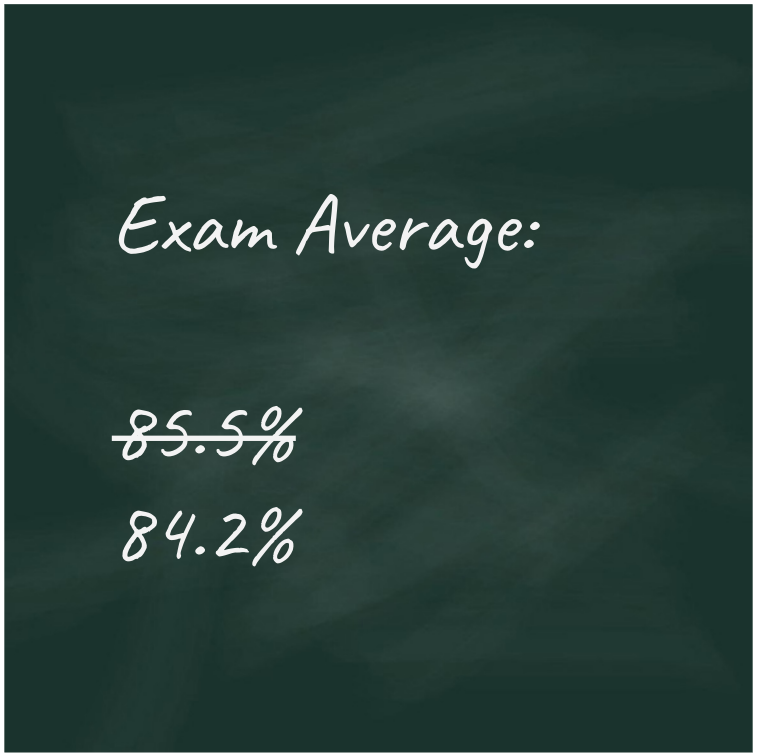
Has your friend's privacy been violated?

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

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

$$Q + R + S + T + X = 85.5 * 5 = (Y + noise1) * 5$$

$$Q + R + S + T = 84.2 * 4 = (Z + noise2) * 4$$



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

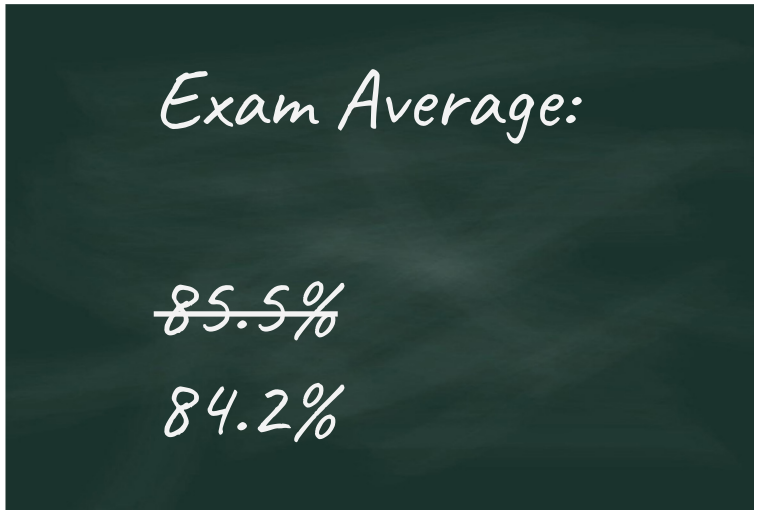
Privacy has been preserved

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



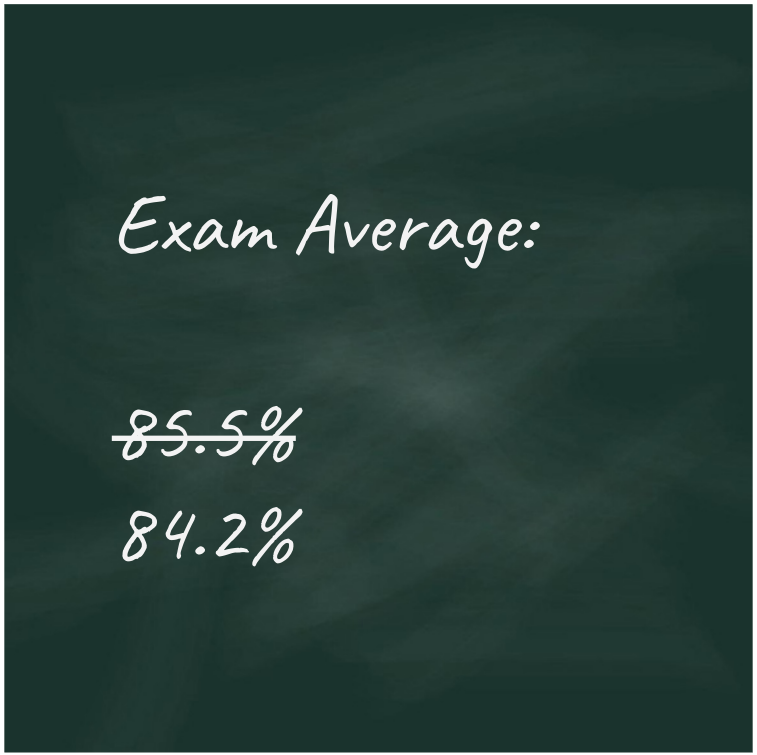
Differential Privacy

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

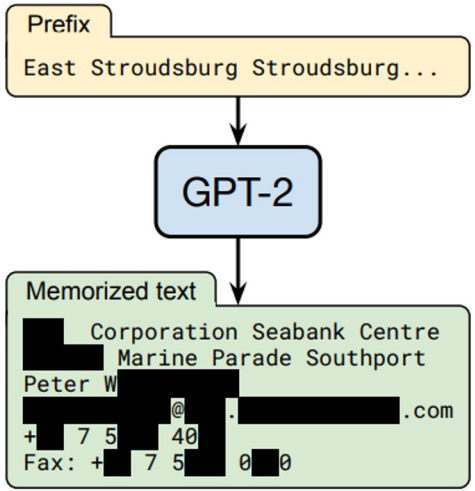
Seeing **DP-mean** of 85.5 vs 84.2 gives us information about your friend's *possible grades*

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

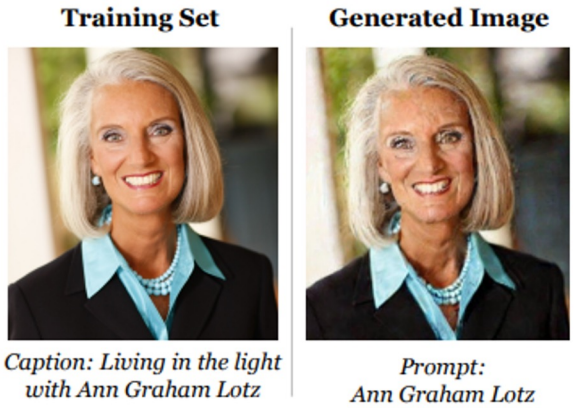
Epsilon - defines how much noise to add



Extracting Training Data from ML/LLM Models



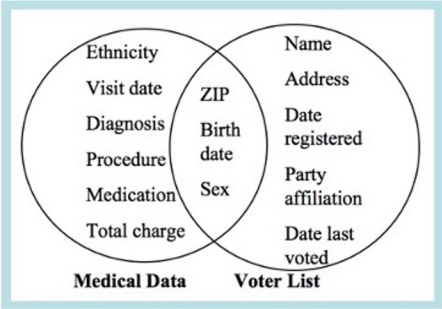
[Carlini, Tramèr, Wallace et al. 2021]



[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]

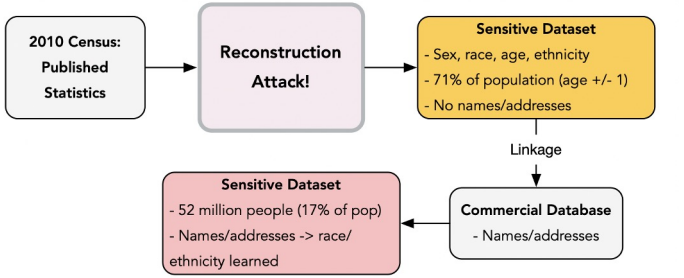
Attacks on "Aggregate" Statistics

Why Differential Privacy?

❌ Releasing statistics leaks privacy

❌ Releasing lots of statistics leaks lots of privacy.

- 2010 census was reverse engineered



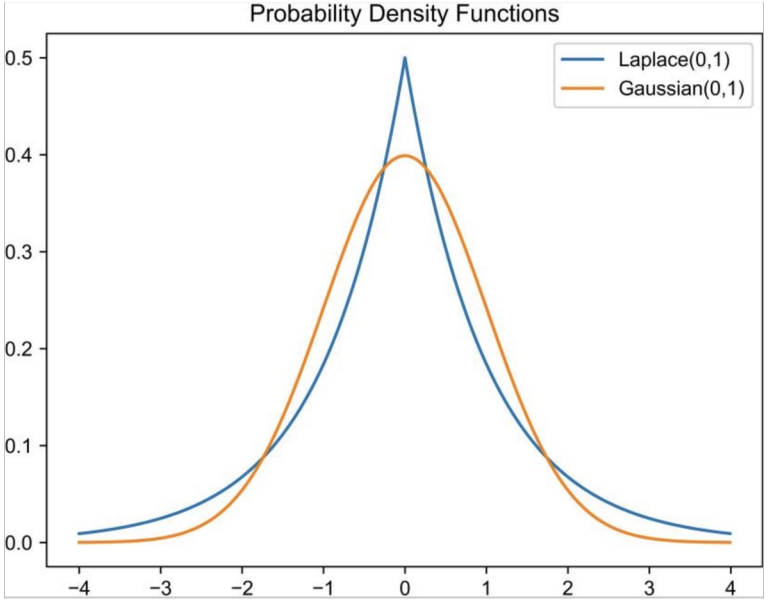
❌ Traditional protection techniques don’t always work

- 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

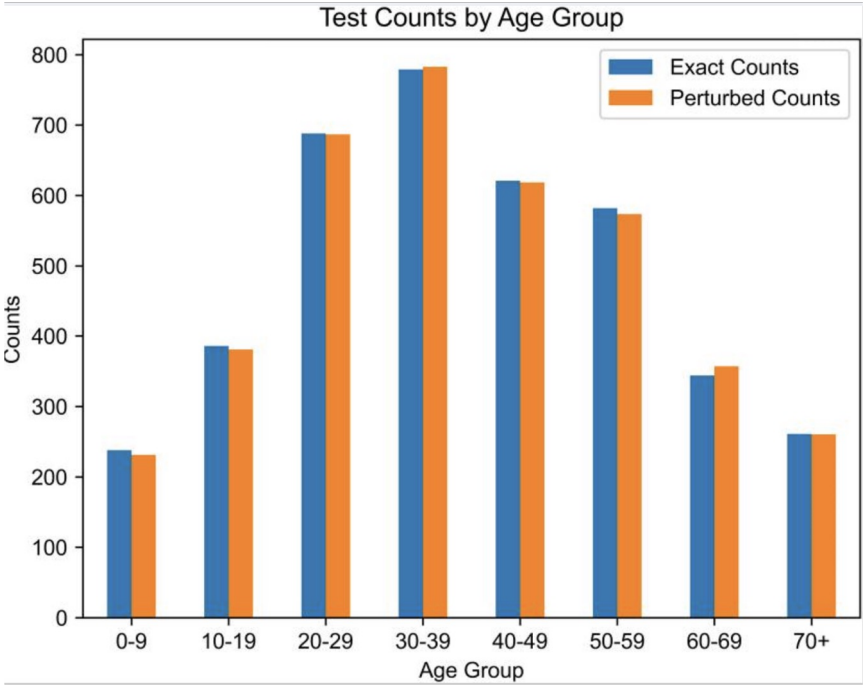
How Differential Privacy works in practice ?



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)

Patters *et al* 2021 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8662814/

How Differential Privacy works in practice ?

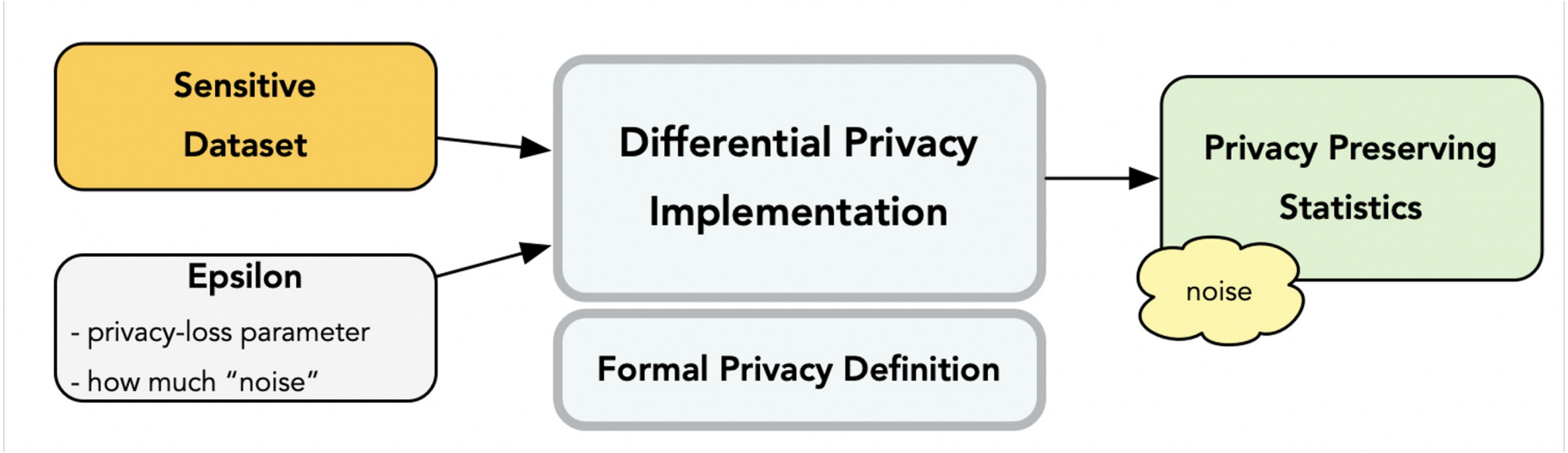


On histograms: adding independent noises to each bin protects privacy.

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

Patters *et al* 2021 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8662814/

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 D1 and D2, where D1 and D2 differ only by 1 row, the probability of having some statistical result $\mathcal{A}$(D1) is the same as $\mathcal{A}$(D2) but with

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

The (Inherent) Privacy-Utility Trade-off

Every statistical release incurs some privacy loss ε_i.

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

With multiple queries, the privacy loss accumulates.

- Overall privacy loss $\leq \varepsilon_1+\varepsilon_2+\cdots+\varepsilon_k$

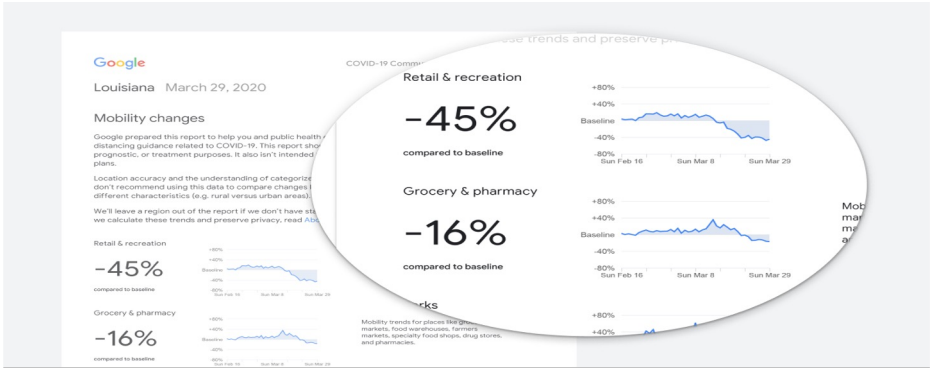
  There are better composition theorems for differential privacy.

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

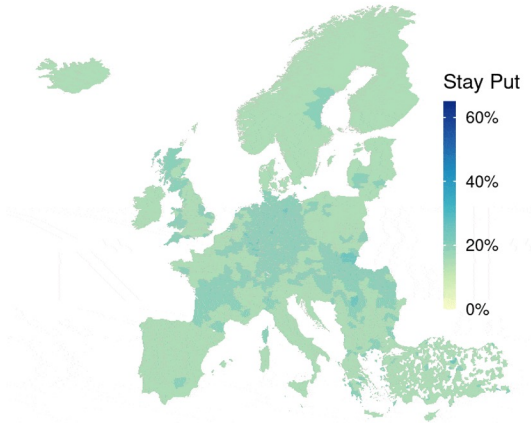
Recommended use: set an overall budget ε (e.g. $\varepsilon=.1$)

Stop answering queries when budget reached.

DP and the Pandemic



Date: 2020-03-04



GOOGLE COVID-19 Community Mobility Reports [Aktay et al., April 2020]
https://www.google.com/covid19/mobility/

Facebook Movement Range Data [Herdağdelen et al., April 2020]
https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/

OpenDP

About ▾ Opportunities ▾ Community ▾ Software ▾ People ▾ Events Blog

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/

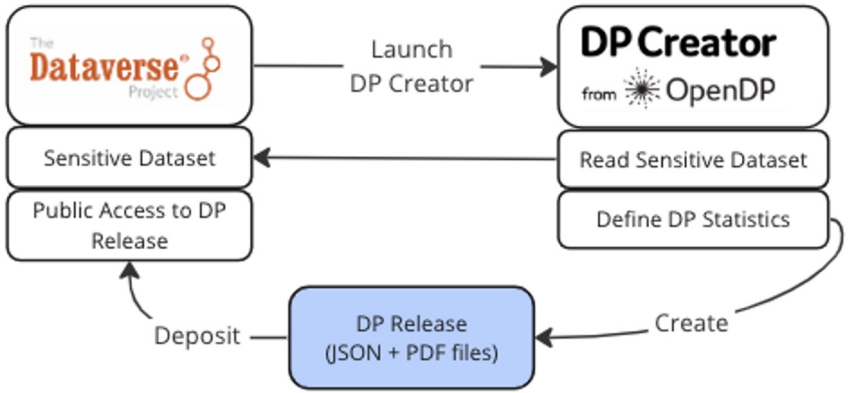
DP and Dataverse: The DPCreator Application

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

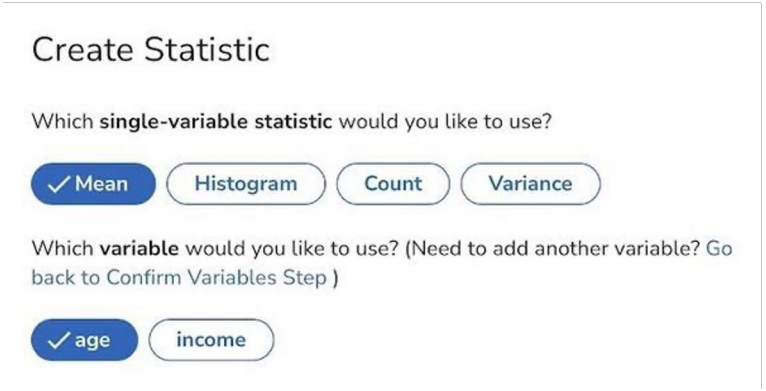
- Integrating user studies and feedback
- “Desktop”/client side version
  - Sensitive data stays with the user
- “Learning Mode” to further help users

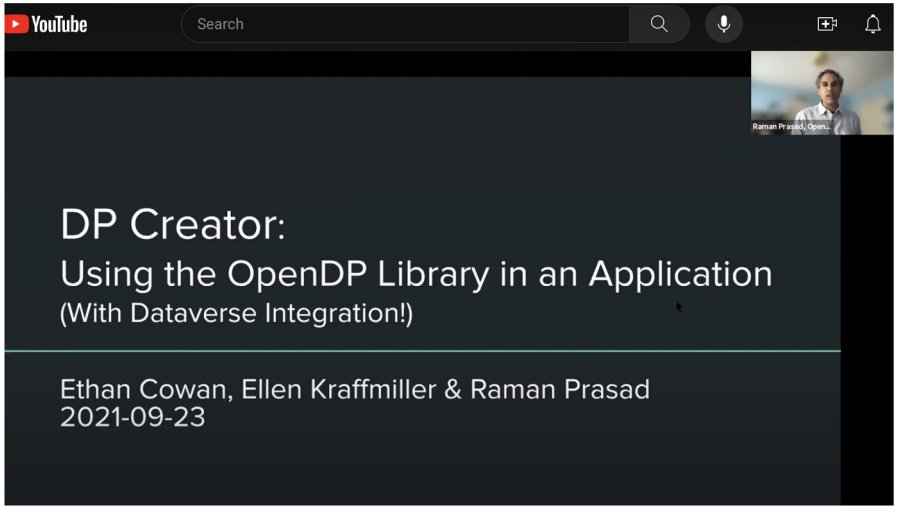


OpenDP

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.

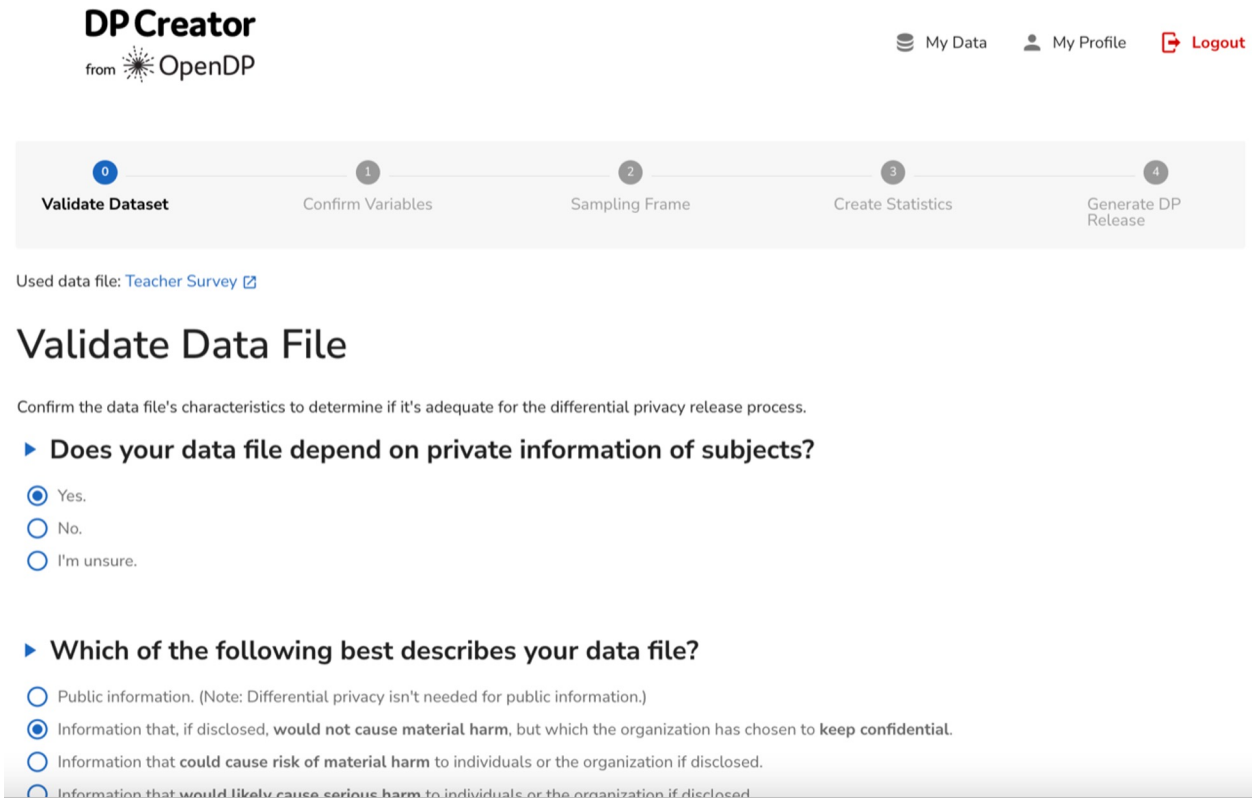
Through a series of questions, the user is helped in the process of creating privacy guaranteed summaries of the data.



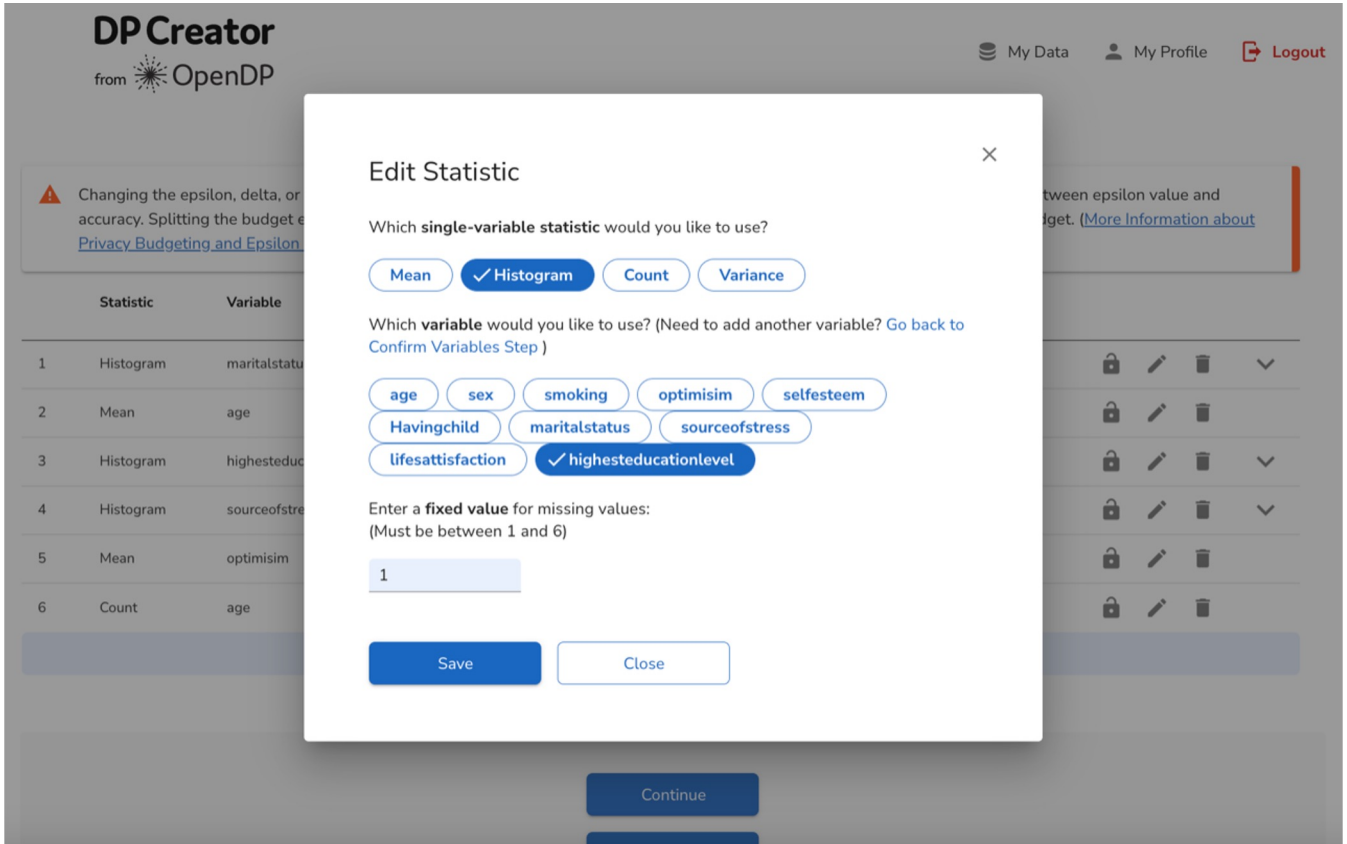


Youtube video of DPCreator

DP Creator: a Tool for Non-Experts



Creating Statistics



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

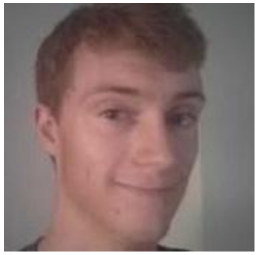
Andrew Vyrros
Senior Library Architect

Annie Wu
Program Director

Sharon Ayalde
Director of Community Engagement

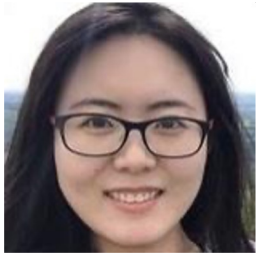
Lindsay Froess
Project Coordinator

Raman Prasad
Technical Lead for Research Software



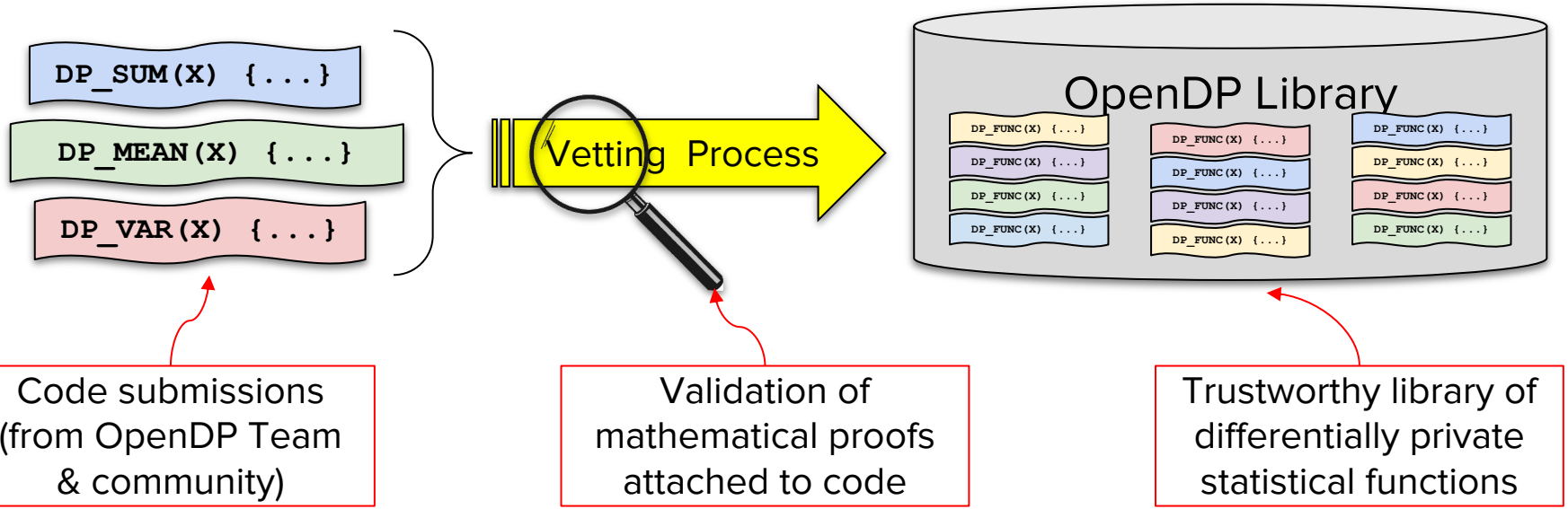
Michael Shoemate
Senior Software Developer

Vikrant Singhal
Research Associate



Wanrong Zhang
Research Associate

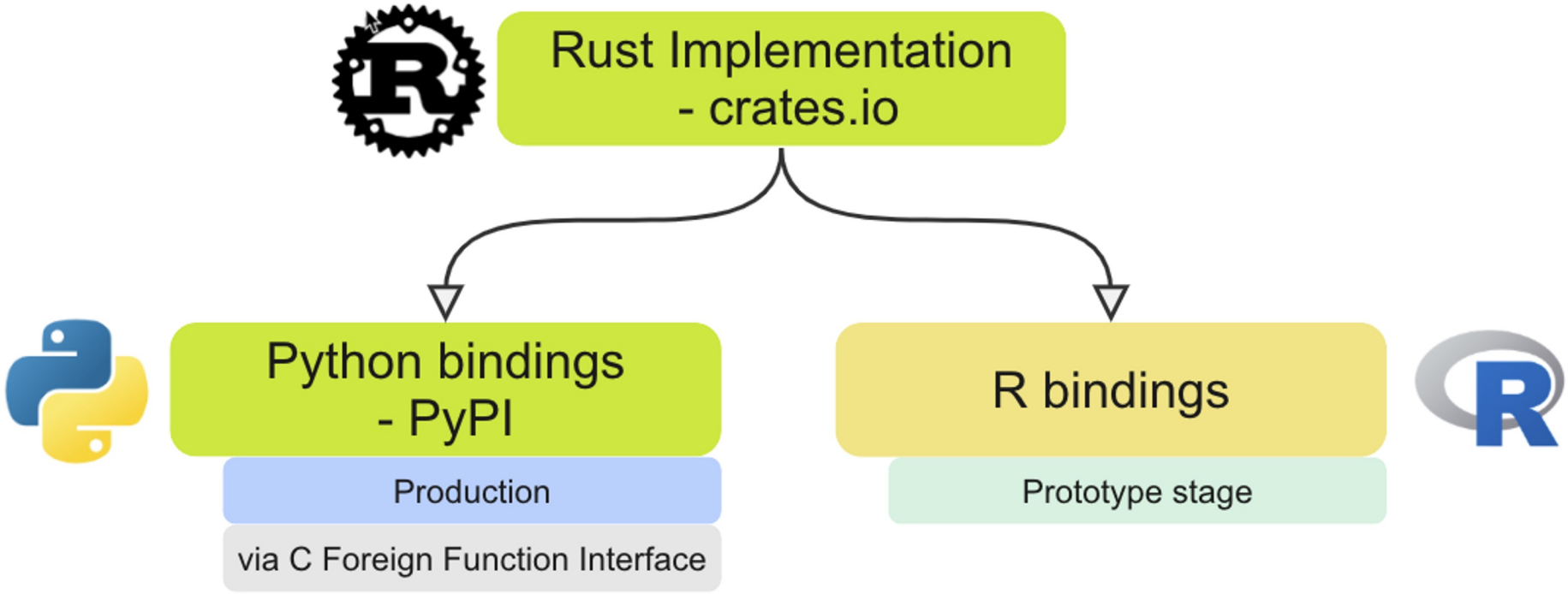
Building a Trustworthy Library



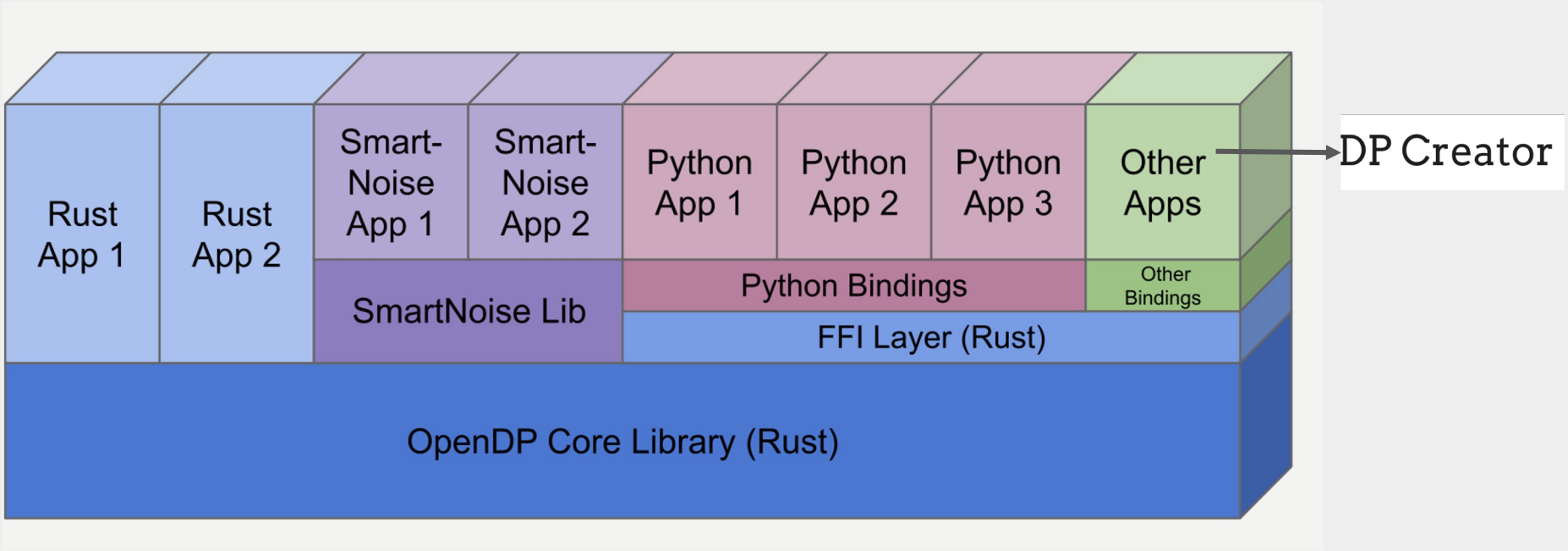
Use Cases

- **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



HARVARD UNIVERSITY

IQSS

Thanks!

With contribution by & credits to the teams of

The Dataverse® Project

OpenDP

siacus@iq.harvard.edu