OpenDP: an Overview

Salil Vadhan
Harvard University
salil_vadhan@harvard.edu

OpenDP Community Meeting
13-15 May 2020

Supported by: Microsoft

Alfred P. Sloan Foundation

+

NSF Google

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

Goals of Differential Privacy

[…, Dwork-McSherry-Nissim-Smith `06]

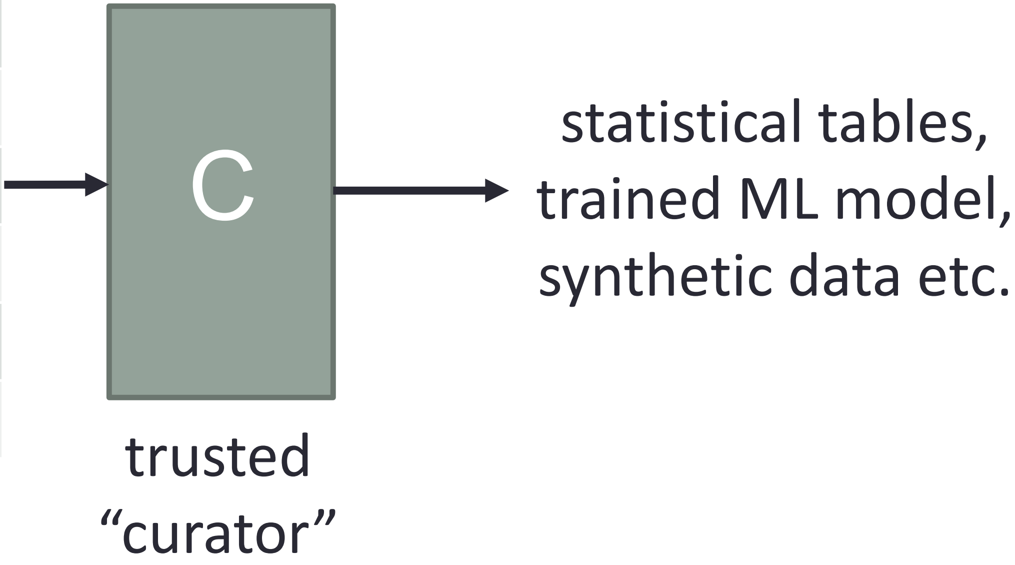
- Utility: enable “statistical analysis” of datasets

- Privacy: protect “individual-level” data

[See appendix of OpenDP whitepaper for a brief primer on DP]

Statistical Releases

Name	Sex	Blood	…	HIV?
Chen	F	B	…	Y
Jones	M	A	…	N
Smith	M	O	…	N
Ross	M	O	…	Y
Lu	F	A	…	N
Shah	M	B	…	Y



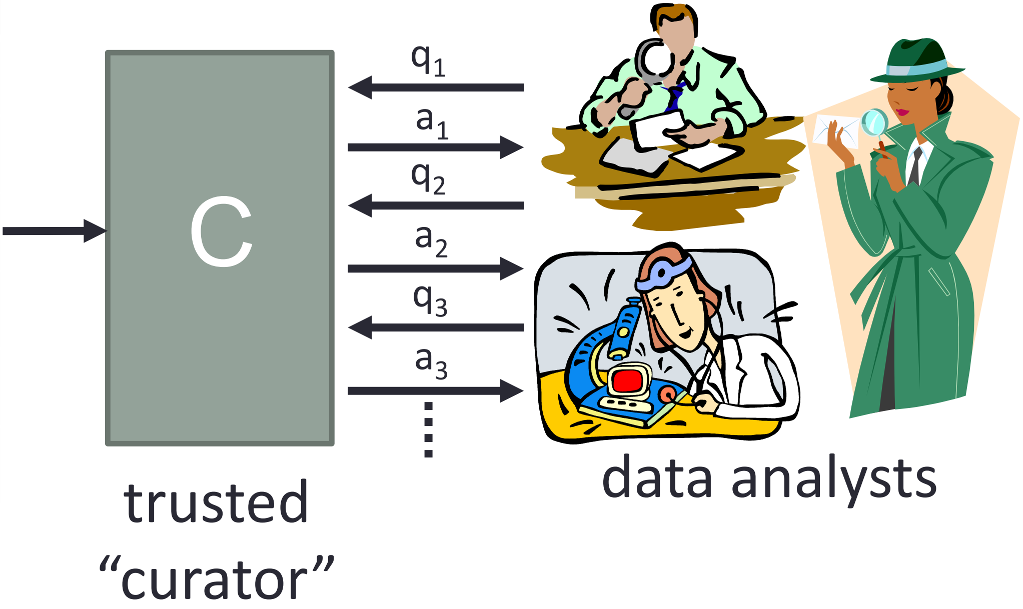
C

trusted
“curator”

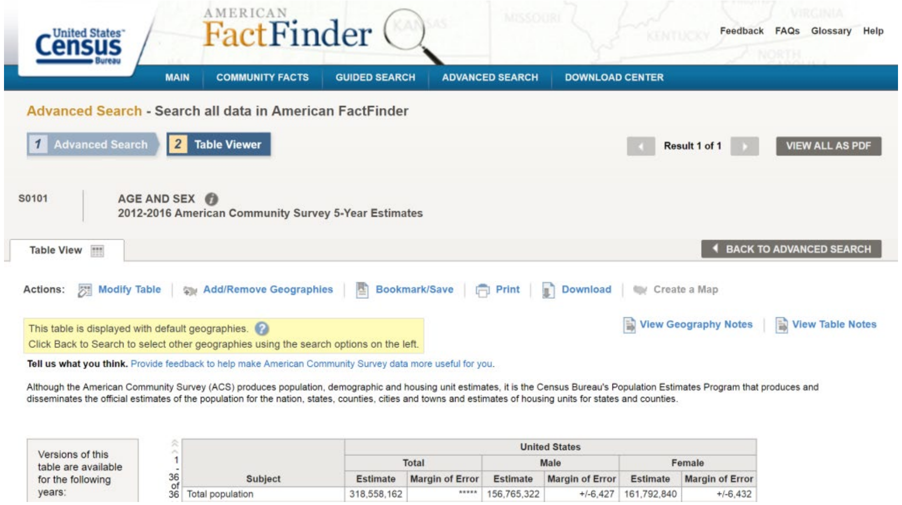
statistical tables,
trained ML model,
synthetic data etc.

Statistical Query Systems

Name	Sex	Blood	⋯	HIV?
Chen	F	B	⋯	Y
Jones	M	A	⋯	N
Smith	M	O	⋯	N
Ross	M	O	⋯	Y
Lu	F	A	⋯	N
Shah	M	B	⋯	Y



Existing Query Interfaces


United States Census Bureau
AMERICAN FactFinder
Feedback FAQs Glossary Help
MAIN COMMUNITY FACTS GUIDED SEARCH ADVANCED SEARCH DOWNLOAD CENTER

Advanced Search - Search all data in American FactFinder

1 Advanced Search
2 Table Viewer
Result 1 of 1
VIEW ALL AS PDF

S0101 AGE AND SEX
2012-2016 American Community Survey 5-Year Estimates

Table View
BACK TO ADVANCED SEARCH
Actions: Modify Table | Add/Remove Geographies | Bookmark/Save | Print | Download | Create a Map
View Geography Notes | View Table Notes

This table is displayed with default geographies.
Click Back to Search to select other geographies using the search options on the left.

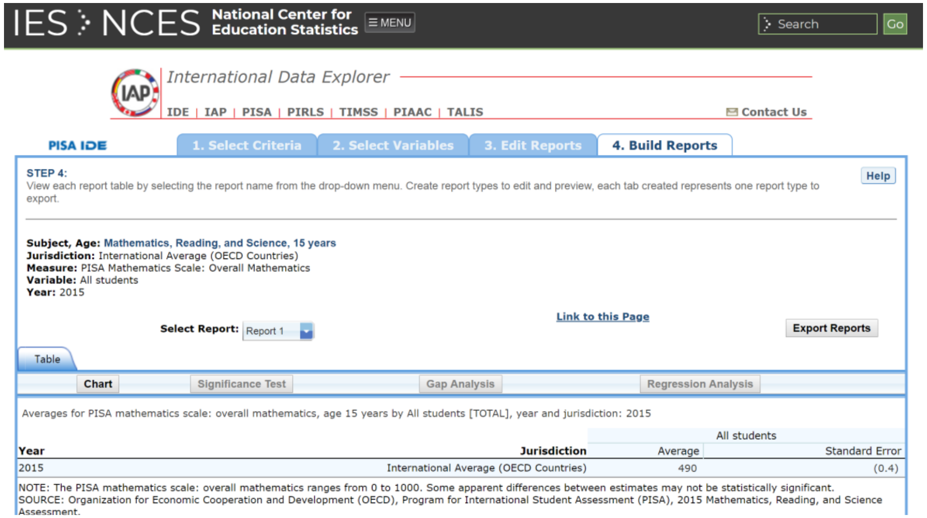
Tell us what you think. Provide feedback to help make American Community Survey data more useful for you.

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Versions of this table are available for the following years:

1 - 36 of 36

	United States					
	Total		Male		Female	
Subject	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
Total population	318,558,162	*****	156,765,322	+/-6,427	161,792,840	+/-6,432


IES NCES National Center for Education Statistics
MENU
Search Go
IAP International Data Explorer
IDE | IAP | PISA | PIRLS | TIMSS | PIAAC | TALIS
Contact Us
PISA IDE
1. Select Criteria
2. Select Variables
3. Edit Reports
4. Build Reports
Help

STEP 4:
View each report table by selecting the report name from the drop-down menu. Create report types to edit and preview, each tab created represents one report type to export.

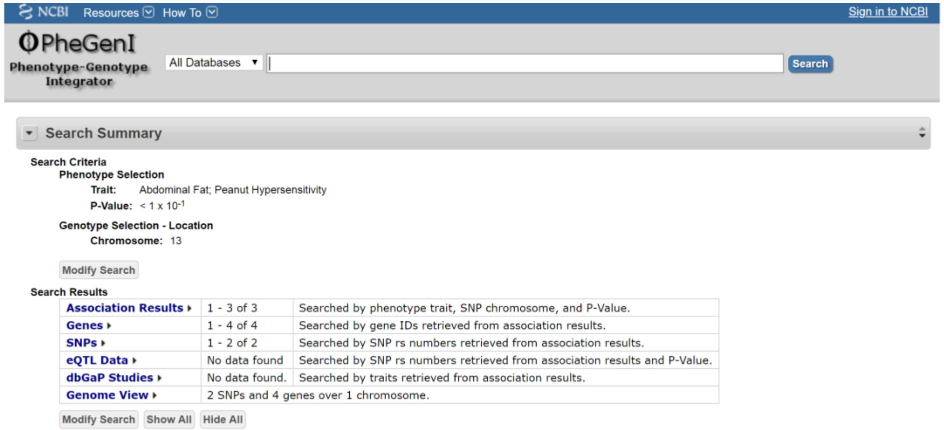
Subject, Age: Mathematics, Reading, and Science, 15 years
Jurisdiction: International Average (OECD Countries)
Measure: PISA Mathematics Scale: Overall Mathematics
Variable: All students
Year: 2015

Select Report: Report 1

Averages for PISA mathematics scale: overall mathematics, age 15 years by All students [TOTAL], year and jurisdiction: 2015

		All students	
Year	Jurisdiction	Average	Standard Error
2015	International Average (OECD Countries)	490	(0.4)

NOTE: The PISA mathematics scale: overall mathematics ranges from 0 to 1000. Some apparent differences between estimates may not be statistically significant.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2015 Mathematics, Reading, and Science Assessment.

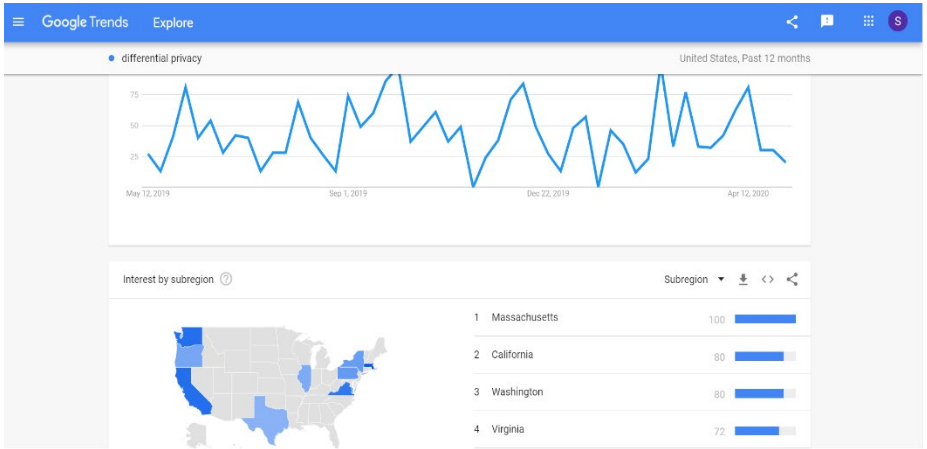


Search Summary

Search Criteria
Phenotype Selection
Trait: Abdominal Fat; Peanut Hypersensitivity
P-Value: $< 1 \times 10^{-1}$
Genotype Selection - Location
Chromosome: 13

Search Results

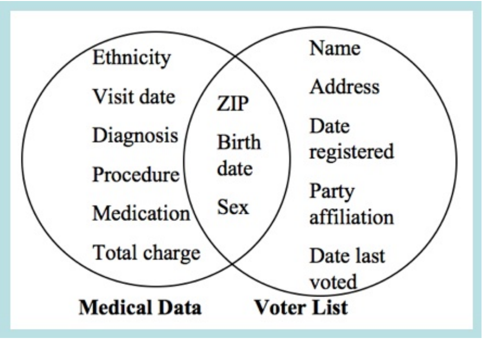
Association Results	1 - 3 of 3	Searched by phenotype trait, SNP chromosome, and P-Value.
Genes	1 - 4 of 4	Searched by gene IDs retrieved from association results.
SNPs	1 - 2 of 2	Searched by SNP rs numbers retrieved from association results.
eQTL Data	No data found	Searched by SNP rs numbers retrieved from association results and P-Value.
dbGaP Studies	No data found.	Searched by traits retrieved from association results.
Genome View	2 SNPs and 4 genes over 1 chromosome.	



1	Massachusetts	100
2	California	80
3	Washington	80
4	Virginia	72

Why DP? Attacks on Privacy

- Re-identification: determining who is who even after "PII" removed
  - Applied to medical data [Sweeney `97], Netflix challenge [Narayanan-Shmatikov `08], …



[Sweeney `97]

- 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,…].

Attacks on "Aggregate" Statistics

Goals of Differential Privacy

[…, Dwork-McSherry-Nissim-Smith `06]

- Utility: enable “statistical analysis” of datasets
  - e.g. inference about population, ML training, descriptive statistics, synthetic data
- Privacy: protect “individual-level” data
  - against “all” attack strategies, background info.
  - now accepted as a “gold standard” for protection

How to achieve?

- Inject “small” amount of random noise into statistical computations

[See appendix of OpenDP whitepaper for a brief primer on DP]

Differentially Private Algorithms circa 2014

- histograms [DMNS06]
- contingency tables [BCDKMT07, GHRU11, TUV12, DNT14],
- machine learning [BDMN05,KLNRS08],
- regression & statistical estimation [CMS11,S11,KST11,ST12,JT13]
- clustering [BDMN05,NRS07]
- social network analysis [HLMJ09,GRU11,KRSY11,KNRS13,BBDS13]
- approximation algorithms [GLMRT10]
- singular value decomposition [HR12, HR13, KT13, DTTZ14]
- streaming algorithms [DNRY10,DNPR10,MMNW11]
- mechanism design [MT07,NST10,X11,NOS12,CCKMV12,HK12,KPRU12]
- synthetic data [BLR08,HR10,GGHRW14]
- …

Differential Privacy Deployed

U.S. Census Bureau

- “OnTheMap” commuter data [Machanavajjhala et al. `06]
- Planned: all public-use products from 2020 Decennial Census [Abowd `18]

Tech Industry

- RAPPOR for Chrome Statistics [Erlingsson et al. `14]
- Tensorflow Privacy [Abadi et al. `16,…]
- iOS10 and Safari [2016]
- Windows 10 [Ding et al. `17]
- …

Research Community

- Numerous prototypes from individual groups

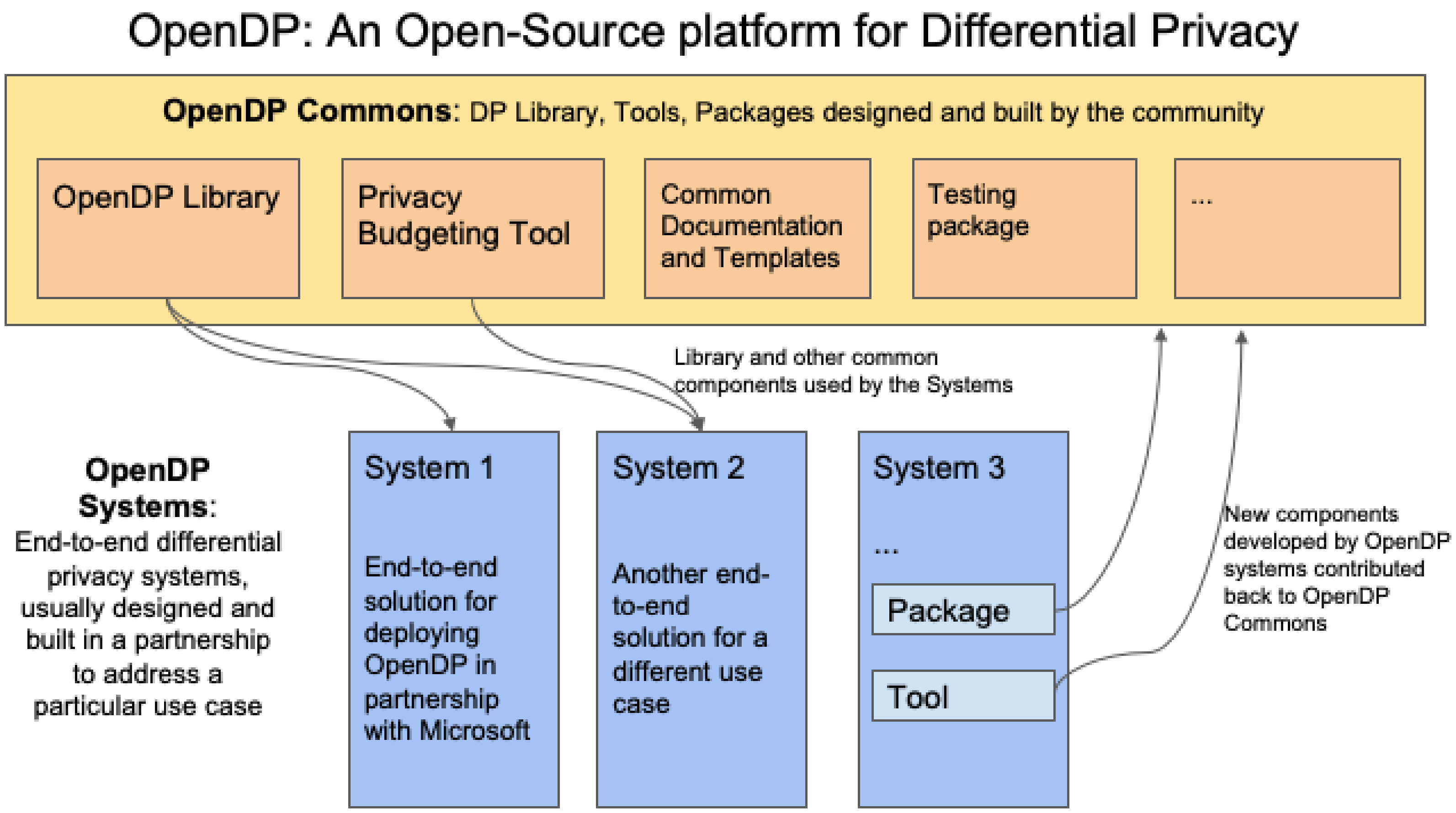
OpenDP

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

Planned Structure



OpenDP: An Open-Source platform for Differential Privacy

OpenDP Commons: DP Library, Tools, Packages designed and built by the community

OpenDP Library

Privacy Budgeting Tool

Common Documentation and Templates

Testing package

...

Library and other common components used by the Systems

OpenDP Systems: End-to-end differential privacy systems, usually designed and built in a partnership to address a particular use case

System 1

End-to-end solution for deploying OpenDP in partnership with Microsoft

System 2

Another end-to-end solution for a different use case

System 3

...

Package

Tool

New components developed by OpenDP systems contributed back to OpenDP Commons

Key Elements

- Use Cases
- Governance
- Programming Framework
- Statistical Functionality
- System Integrations
- Collaborations
- Community!

More details in plenaries, breakouts, and the whitepaper.

How we got here

Spring/Summer 2019

- Pitch to DP community @ Simons Institute
- Proposal to Sloan Foundation
- Funding received
- Microsoft collaboration starts

Fall/Winter 2019-2020

- Ad Hoc Design Committee meetings & workshop
- OpenDP staff hired
- Software development advances with Microsoft

Spring 2020

- Programming Framework & other elements fleshed out
- First version of system with Microsoft near completion
- Advisory Board formed
- OpenDP Community Meeting!

Where we’re going

Summer 2020:

- Absorb community feedback
- Implement DP library in OpenDP Commons
- Form Ad Hoc Security Review Committee
- Find DP Applications Leader(s), COVID-19 use case
- Establish partnership model, more collaborations
- Fundraising

Fall 2020:

- Launch the OpenDP Commons with working library
- Establish Editorial Board & Committers to review contributions
- Release MVP of 1st OpenDP System, with Dataverse integration
- Second OpenDP Community Meeting

Beyond:

- Expand functionality and deployments
- Form Steering Committee
- Sustainability through community commitment

What can you do?

Follow our plans

- Many more details in the whitepapers at http://opendp.io/
- Watch for emails and posts from us

Contribute

- Participate in breakout discussions
- Send feedback & suggestions to info@opendp.io anytime
- Stay tuned for more opportunities

Collaborate

- See Collaborations plenary & breakout