

Alex Selkirk, Common Data Project

Can We Have Our Cake and Eat It Too?

The Potential of a “Datatrust”
to Open Personal Data
While Protecting Privacy.



Are We Missing Out?

What %age of data is
unavailable to the public
because releasing it would
unacceptably violate
privacy?



Pushing to Release More Sensitive Data: Some examples from the U.S.

- Medicare “Personal Health Records”:
<http://www.medicare.gov/navigation/manage-your-health/personal-health-records/personal-health-records-overview.aspx>
- Individual Tax Returns
- Raw Census Data
- Surveys / Studies from the Centers for Disease Control and Dept of Health
- School Records
 - Students’ grades
 - Teachers’ results
- Criminal Records
- Social Service Records
- NYC Metrocard usage tied to credit cards
- What if we could link all of this together?



2000 U.S. Census Oopsy: Accurate versus Private

- U.S. Census uses “data-swapping” and “synthetic data” to anonymize their releases
- They made a “mistake” in the way they released 2000 data on the 65-and-older crowd
- And they can’t correct their mistake, otherwise it would compromise the anonymization work they’ve done.
- <http://freakonomics.blogs.nytimes.com/2010/02/02/can-you-trust-census-data/>

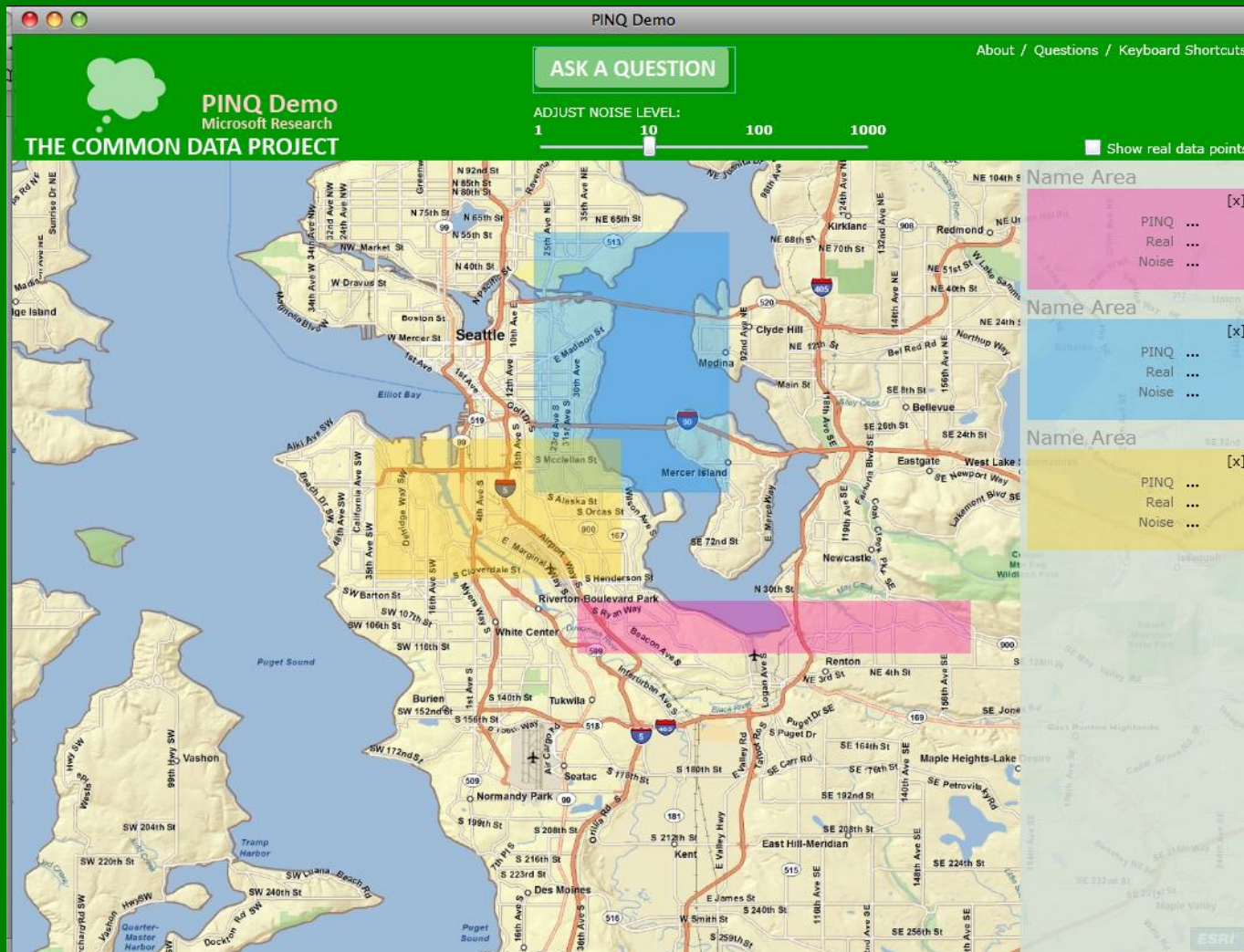


Introducing the Datatrust

- A non-profit, community “bank” for sensitive data with data deposits by donors and data withdrawals for re-use by researchers and application builders.
- Define a quantifiable privacy guarantee that wipes out individual identities, yet allows for arbitrary, high-level analysis of the raw data using technology derived from an area of crypto/statistics research called “Differential Privacy”.
- Establishes a “privacy currency” in order to measure and track privacy risk on a question-by-question basis.
- No more scrubbing, swapping, synthesizing data on a case-by-case basis.
- No pre-digested aggregates.

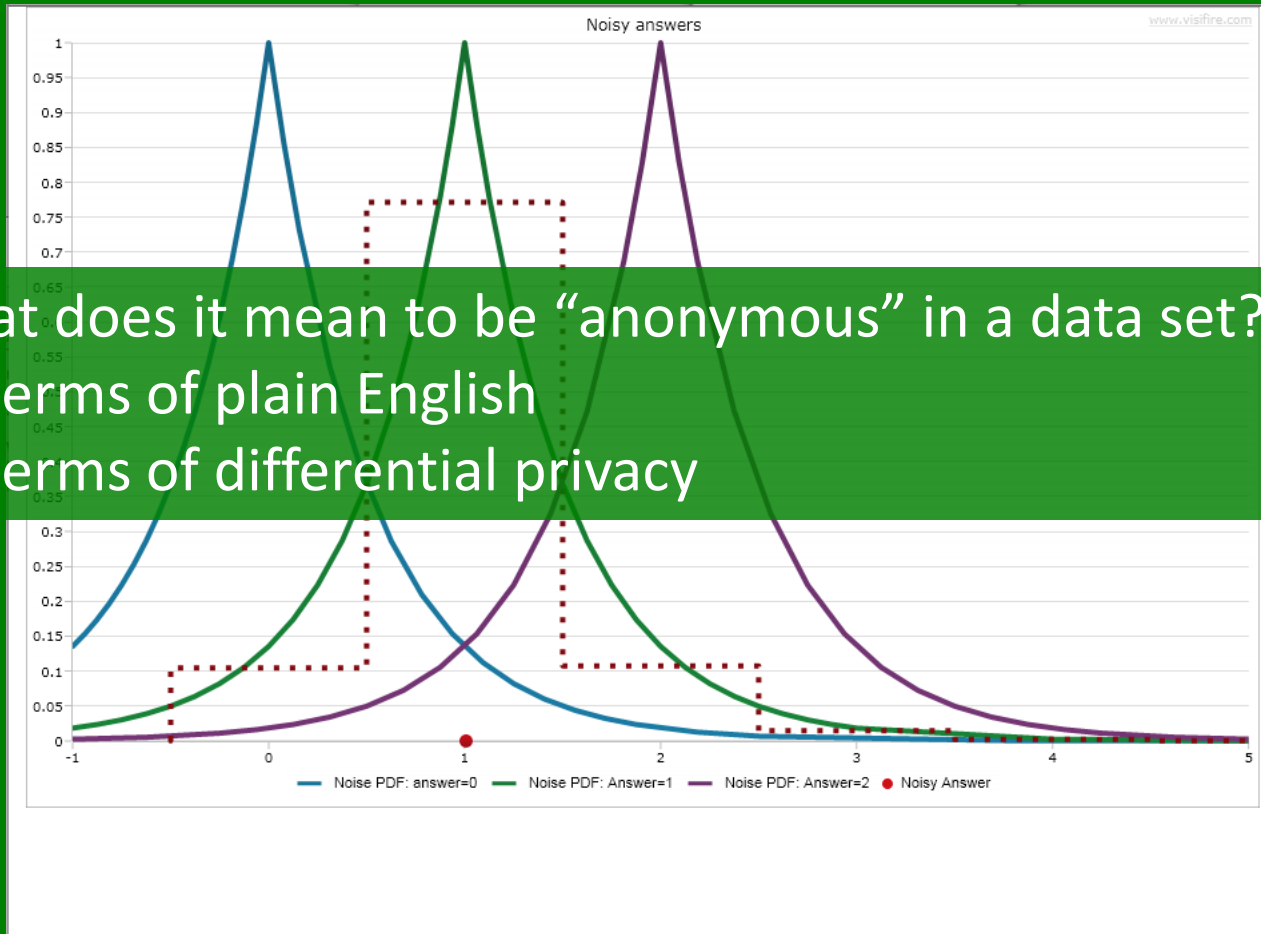


Working with Differential Privacy to Measure and Track Privacy Risk



Defining “Private” and “Anonymous”

What does it mean to be “anonymous” in a data set?
-In terms of plain English
-In terms of differential privacy



Looking for A Few (Thousand) Good (Raw and Sensitive) Data Points.

- Sensitive data
- Released in “raw” form: Not aggregated. Not sampled. Not swapped. Not “made up.” Not scrubbed.

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.

- <http://www.cdc.gov/nchs/nhanes.htm>



What We're Curious About

- What are your experiences gaining access to sensitive government data?
- U.K. National Health Services' campaign to standardize digital medical records, is that released to the public in any form?



More Resources

- Website: <http://commondataport.org>
- Blog: <http://myplaceinthecrowd.org>
- Twitter: <http://twitter.com/mpitc>
- Contact: alex.selkirk@commondataport.org

