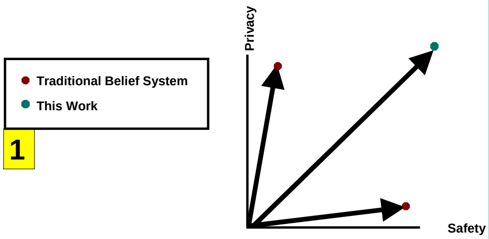
Carnegie Mellon
DATA PRIVACY LAB

Privacy-Preserving Bio-terrorism Surveillance

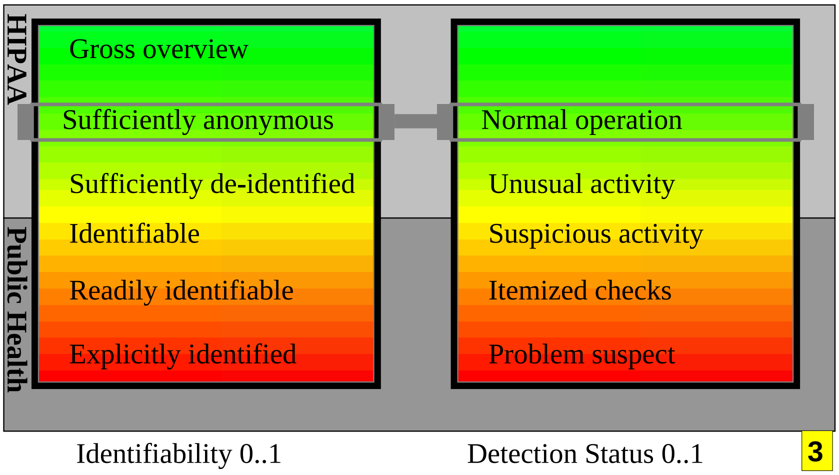


Following the events of September 11, 2001, many in the American public falsely believe they must choose between safety and privacy. This work presents technology that allows medical data to be shared for bio-terrorism surveillance such that the shared data have provable assurances of privacy protection while remaining practically useful. The result allows the American public to enjoy both safety and privacy (1). The conditions for providing such data are listed in (2).

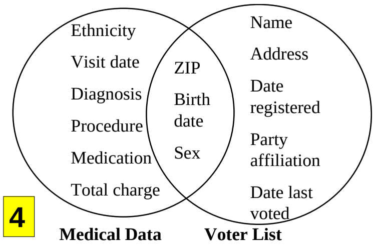
Privacy Conditions for Databases

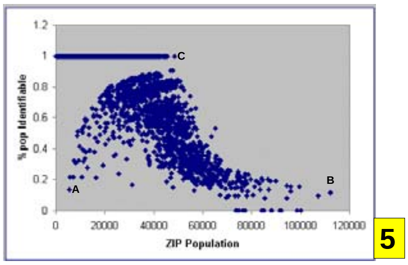
- No person whose information is contained in a Database can be re-identified.
- Investigators can access necessary information contained in a Database freely and easily.
- Results from qualified investigations are equivalent to results in the absence of privacy protection.

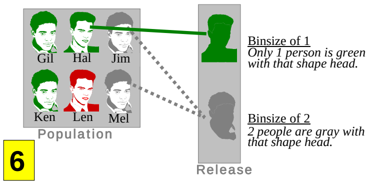
2



The solution is to provide data with a sliding scale of identifiability (3), where the level of anonymity matches the scientifically derived need based on suspicious occurrences appearing within the data. Bio-terrorism surveillance begins with data sufficiently de-identified in accordance to HIPAA. If evidence presents itself, a "drill-down" providing increasing more identifiable data commences in accordance to public health law. So, the goal is to prove that the data are anonymous yet remains useful.

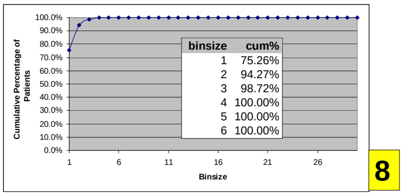




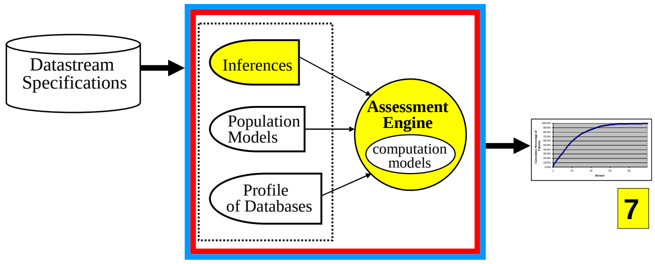


Binsize of 1
Only 1 person is green with that shape head.

Binsize of 2
2 people are gray with that shape head.

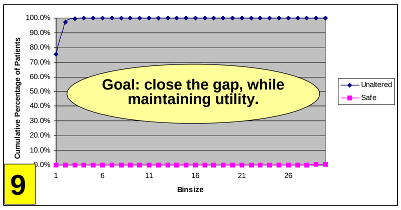


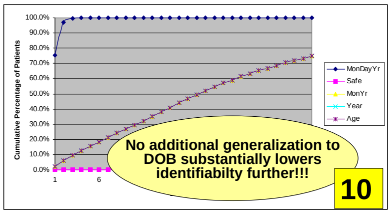
binsize	cum%
1	75.26%
2	94.27%
3	98.72%
4	100.00%
5	100.00%
6	100.00%

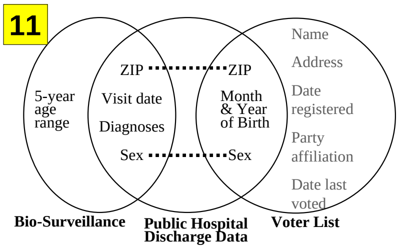


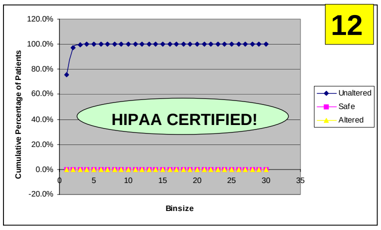
Block diagram of the Risk Assessment Server (7), [invented by Sweeney, 1998, now licensed to Privcaert, Inc. www.privacert.com]. Given a description of a dataset, it uses a population model, its knowledge of available databases and data inferences, and an inference engine to relate the number of people who could be identified in each record in the data (8 output, 6 decription of output).

Data needed by bio-terrorism surveillance can be linked to a population register to re-identify people by name (4). Overall, 87% of the population of the USA is uniquely identified by {date of birth, sex, 5-digit ZIP} (5). Therefore, the data are not sufficiently de-identified! (9) shows the gap. Changing date of birth to month and year of birth improved results (10), but still not sufficient. *Despite intuition, no further generalizations of date of birth make any further improvements!* Using the Risk Assessment Server, we see a 2-stage inference attack was found (11) in which the data is linked to publicly available hospital data to re-learn month and year of birth, irregardless of other aggregations of date of birth in the medical data. To solve this problem, the diagnoses codes are aggregated (12) to syndromes making the data sufficiently de-identified, while remaining useful for surveillance !









Latanya Sweeney
http://privacy.cs.cmu.edu/dataprivacy/projects/riskassess/index.html