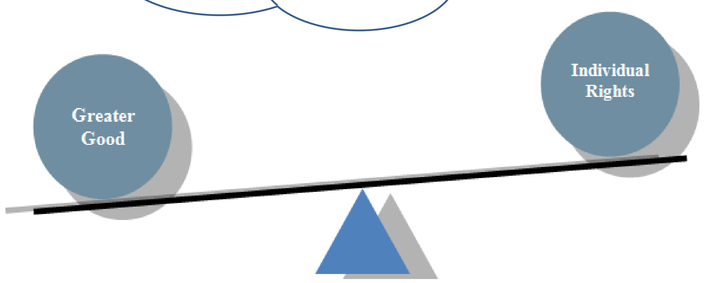
Anonymity & Security in Genomic Datasets

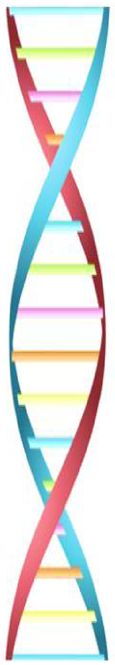
Christopher E. Martinez, *University of Washington;* Derrick Roach, *DePauw University;* and Kristine Arthur, *Purdue University*
Advisor: Dr. John A. Springer, *Purdue University*

Define best practices for preserving anonymity and security in genomic datasets.



Research

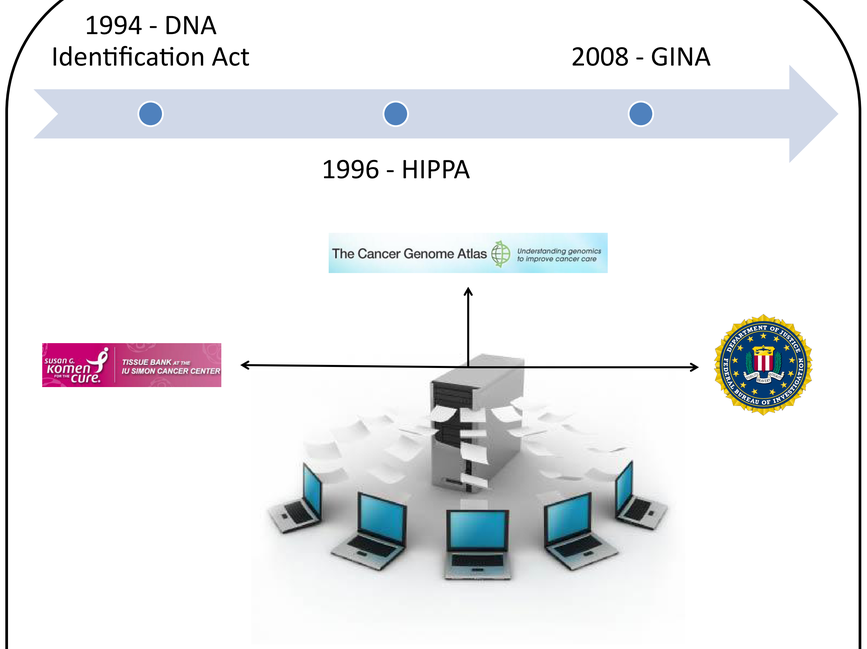
- Decreased Healthcare Costs
- Increased Medicine Effectiveness
- More Accurate Drug Dosages
- Advanced Screening for Disease
- Better Vaccines



Privacy

- Decreased Risk of Disclosure of Personal Information
- Improve Existing Laws for Better Protection
- Prevention of Genoism
- Deterrence of Identity Theft

This material is based upon work supported by the National Science Foundation under grant #1062970. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

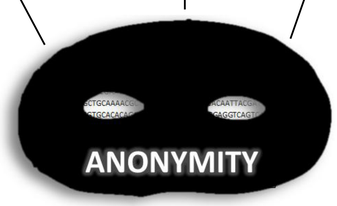


De-Identification Methods

k-Anonymity
Records are arranged into equivalence classes with at least k-1 other records

l-Diversity
Each equivalence class represents at least l distinct sensitive attributes

t-Closeness
The distribution of sensitive attributes in each equivalence class is within a threshold t of the overall distribution



Given a table of patient records, various techniques can be applied to reduce the likelihood of patient re-identification.

Gender	Age	DNA
Female	<25	CTGA
Female	<25	CTGA
Male	<25	ACTG
Male	<25	TCGA

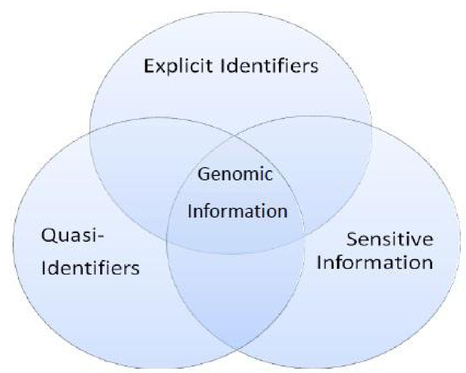
k-Anonymity alone does not sufficiently protect against attribute disclosure

Gender	Age	DNA
Male	<25	CTGA
Male	<25	CTGA
Male	<25	ACTG
Male	<25	CTGA

l-Diversity can misrepresent the occurrence of an attribute in a population

Gender	Age	DNA
Female	>25	CTGA
Female	>25	GTAC
Female	>25	ACTG
Female	>25	CGTA

t-Closeness releases data which is close in distribution to the overall population



Genomic information can be placed into all three categories of patient identifiers