

9<sup>th</sup> Annual Oak Ridge Postdoctoral Association Research Symposium, Poster Session

# A GAME THEORETIC MODEL FOR PRIVACY-PRESERVING GENOMIC DATA SHARING

*Zhiyu Wan, Yevgeniy Vorobeychik, Weiyi Xia, Ellen Clayton,  
Murat Kantarcioglu and Bradley Malin*

Presenter: Zhiyu Wan, Ph.D.

Health Information Privacy Lab

Vanderbilt University

[zhiyu.wan@vanderbilt.edu](mailto:zhiyu.wan@vanderbilt.edu)

07/29/2021



# ABSTRACT

Emerging scientific endeavors are creating big data repositories from millions of individuals. Sharing data in a privacy-respecting manner could lead to important discoveries, but high-profile demonstrations show that links between de-identified genomic data and named persons can sometimes be reestablished. Such re-identification attacks have focused on worst-case scenarios and spurred the adoption of data sharing practices that unnecessarily impede research. To mitigate concerns, organizations have traditionally relied upon legal deterrents, like data use agreements, and are considering suppressing or adding noise to genomic variants.

# ABSTRACT (CONT)

In this report, we use a game theoretic lens to develop more effective, quantifiable protections for genomic data sharing. This is a fundamentally different approach because it accounts for adversarial behavior and capabilities and tailors protections to anticipated recipients with reasonable resources.

# ABSTRACT (CONT)

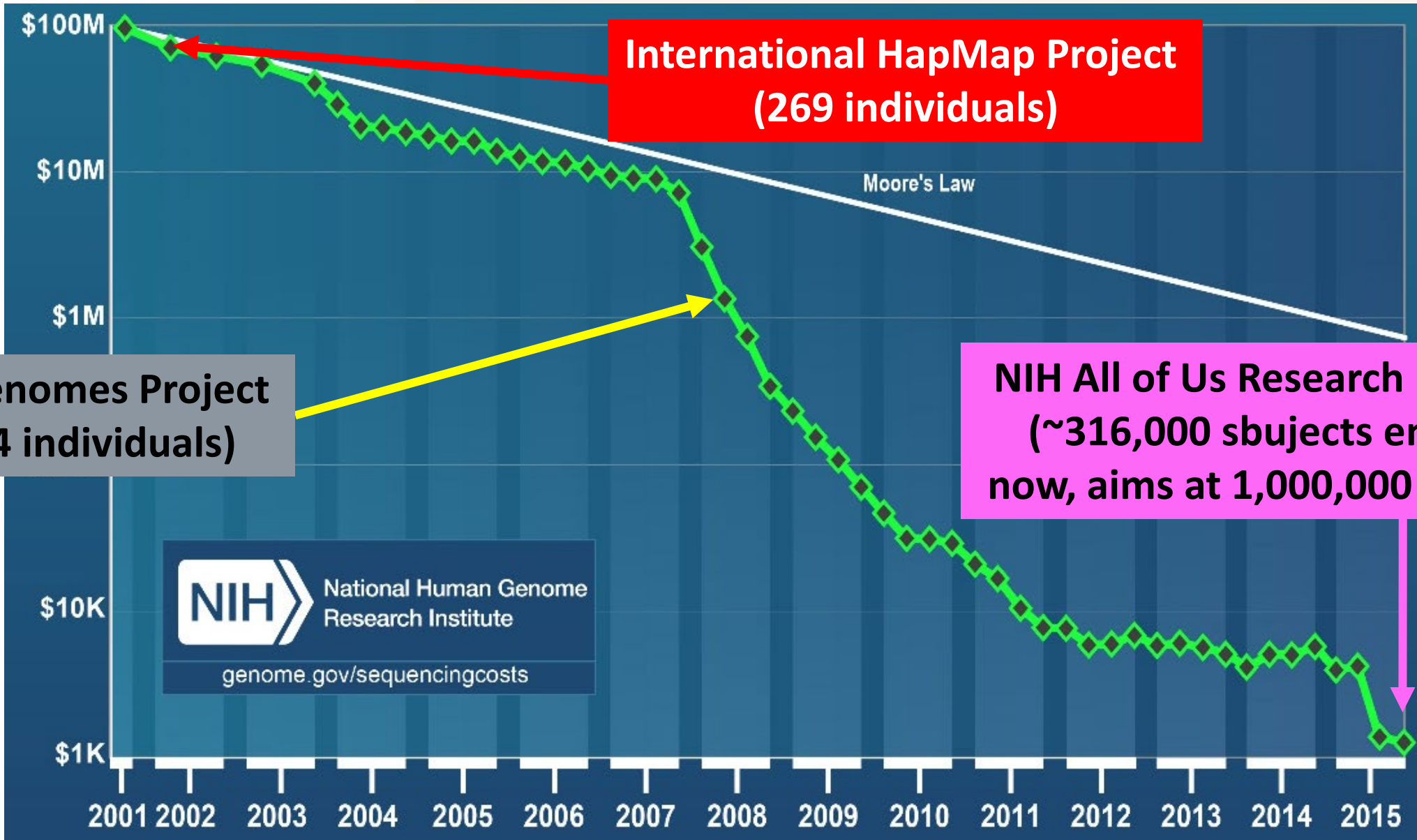
We demonstrate this approach with a public resource with genomic summary data from over 8000 individuals and show risks can be balanced against utility more effectively than traditional approaches. We further show the generalizability of this framework by applying it to other genomic data collection and sharing endeavors. Recognizing that such models are dependent on a variety of parameters, we perform extensive sensitivity analyses to show that our findings are robust to their fluctuations.

# BACKGROUND

- Why the genomic data should be shared? Sharing genomic data is beneficial to us.
  - Tests based on genomic data assists
    - Diagnosis of diseases - that are clinically actionable
    - Establishment of more effective drug regimens
  - Genomic data sharing
    - Accelerates the discovery of new associations
    - Especially for rare diseases
  - NIH-funded investigators are expected to share
    - Genomic data from studies to NIH Database of Genotypes and Phenotypes (dbGaP)
    - Data must be de-identified



# BIG GENOMIC DATA ERA



1000 Genomes Project  
(2,504 individuals)

International HapMap Project  
(269 individuals)

NIH All of Us Research Program  
(~316,000 subjects enrolled now, aims at 1,000,000 subjects)

**NIH** National Human Genome Research Institute  
[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)



VANDERBILT UNIVERSITY

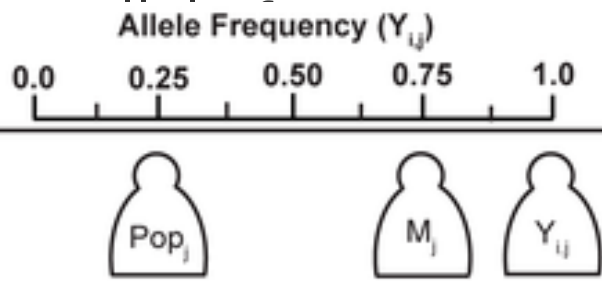
# PRIVACY RISK OF SHARING SUMMARY STATISTICS

- Sharing individual-level genomic data is useful, but risky
- Sharing allele (variant of genomic region) frequencies about a pool of genomes is still useful, but also (less) risky
- In 2008, Homer introduced an attack...

# Homer's attack in a nutshell

## The attacker knows:

- The genome of the target (her set of genomic variants)  
-  $Y_{ij}$
- The allele frequencies of the Mixture he's attacking -  $M_j$

Pop Snp	Allele Frequency ( $Y_{ij}$ )	Distance Measure	Interpretation at the given SNP
j		$D(Y_{i,j}) =  Y_{i,j} - \text{Pop}_j  -  Y_{i,j} - M_j $ $=  1.0 - 0.25  -  1.0 - 0.75 $ $= 0.75 - 0.25$ $= 0.50$	most likely to be in the Mixture



# PRIVACY RISK OF SHARING SUMMARY STATISTICS

- Sharing individual-level genomic data is useful, but risky
- Sharing allele (variant of genomic region) frequencies about a pool of genomes is useful, but also (less) risky
- In 2008, Homer introduced an attack<sup>1</sup>...
  - ... that led the NIH to removing summary statistics from dbGaP
- And more powerful attacks have emerged (e.g., Wang<sup>2</sup>, Sankararaman<sup>3</sup>)
- Technical countermeasures include SNP suppression, noise addition, etc.
- Legal deterrence includes data use agreement (DUA) and penalty

<sup>1</sup>Homer N, et al. PLoS Genetics. 2008; 4(8): e1000167.

<sup>2</sup>Wang R, et al. ACM CCS '16. 2009: 534-544.

<sup>3</sup>Sankararaman S, et al. Nature Genetics. 2009: 965-967.

# OBJECTIVES



serial model

ives

table risk level

anical and legal

- Find the best  
perfect trade  
privacy risk

sharer with a  
tility and

➤ Data sharer is also driven by (economic) incentives

# METHODS

- Model the genomic data sharing process as a one-shot Stackelberg (leader-follower) game between the data sharer and the data recipient
- The genomic data sharing process and the game model
- An illustration of the strategy profile
- Search the data sharer's strategy space using genetic algorithm

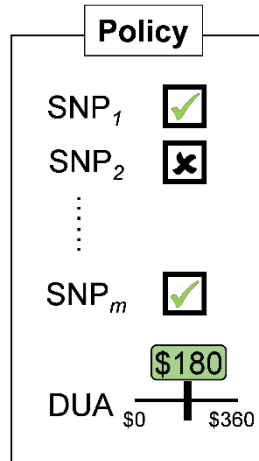
# GENOMIC DATA SHARING PROCESS

A

Sharer's Decision

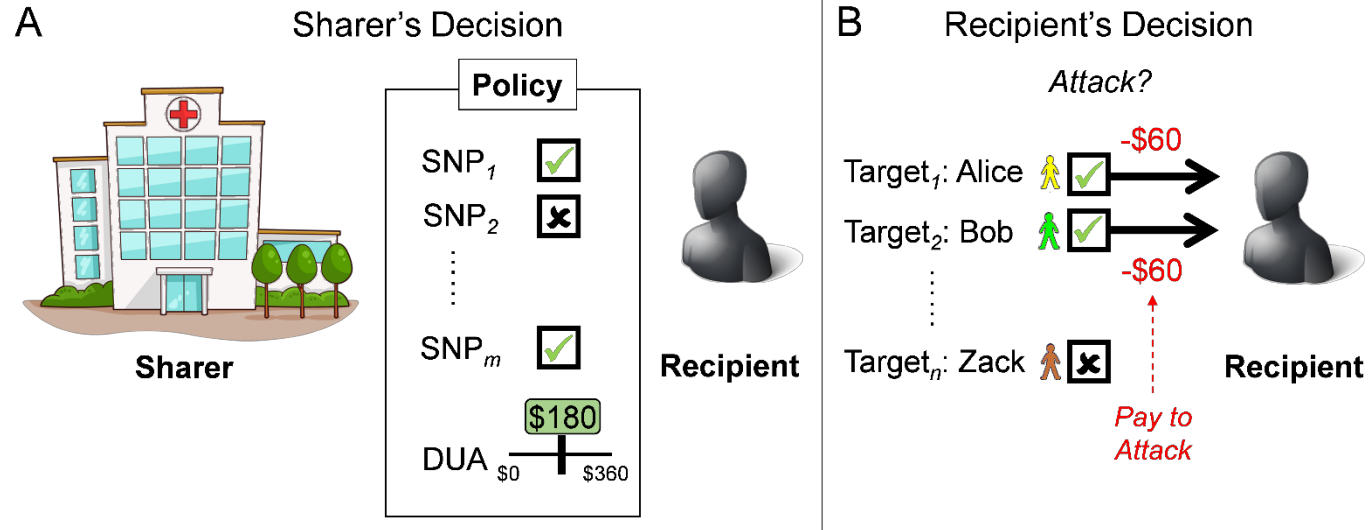


Sharer

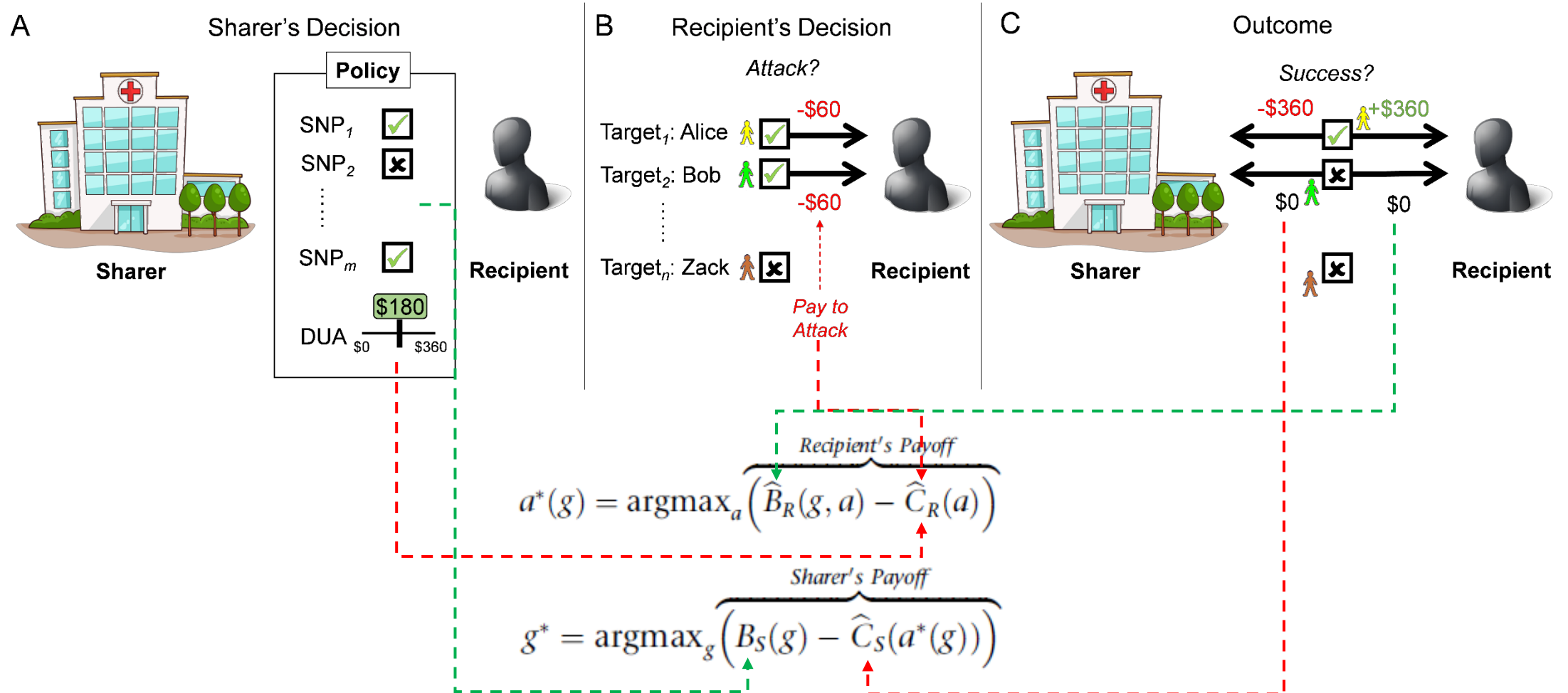


Recipient

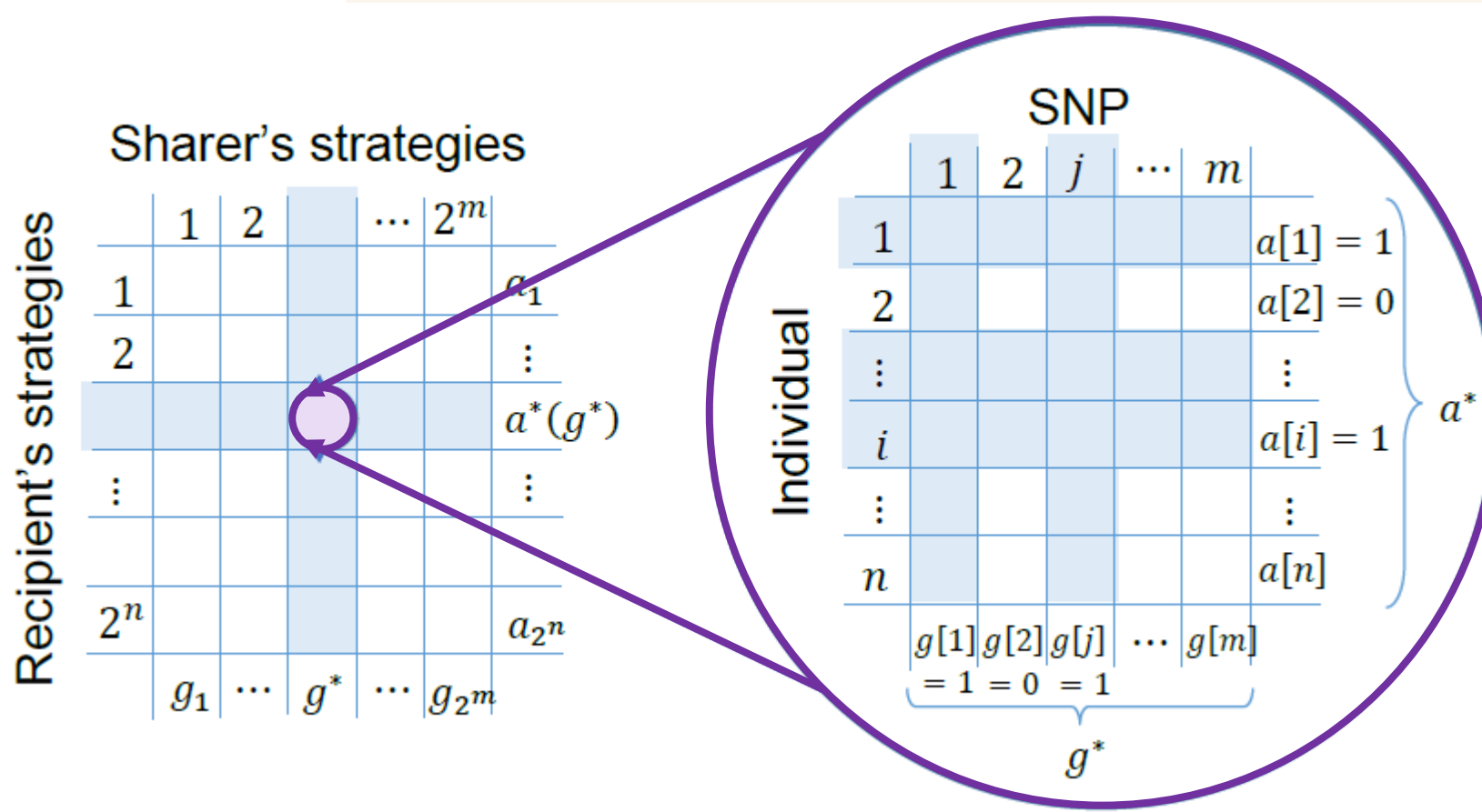
# GENOMIC DATA SHARING PROCESS



# GENOMIC DATA SHARING PROCESS & GAME MODEL



# SEARCH FOR THE DATA SHARER'S BEST STRATEGY



An illustration of the strategy profile

- Genetic Algorithm is introduced to search the strategy space

# EXPERIMENTS

- Dataset

- 8194 individuals in Sequence and Phenotype Integration Exchange (SPHINX)
- The Electronic Medical Records and Genomics – Pharmacogenomics (eMERGE-PGx) project was a multi-center pilot of implementing pharmacogenetic sequencing in clinical practice to improve health care.

<https://www.emergesphinx.org/>



Search by Gene, Drug, chr.position or rsID  Q Search

List all: [genes](#), [drugs](#)

Sequence and *Phenotype* Integration Exchange (*SPHINX*) is a web-based tool for exploring data for hypothesis generation, especially around drug response implications of genetic variation across the eMERGE Network.

	Sites	Samples	Variants	Genes	Drugs
eMERGEseq	10	24,956	62,050	794	2,055
PGRNseq	9	9,010	60,034	413	2,378
<b>Total (unique)</b>	<b>12</b>	<b>33,966</b>	<b>119,095</b>	<b>1,144</b>	<b>2,874</b>

Last update: 1/17/2021

**i** The eMERGE-PGx project was a multi-center pilot of implementing pharmacogenetic sequencing in clinical practice to improve health care. SPHINX is a searchable catalog of observed inherited variants in a 33,966 subject population, large enough to reflect even rare variation. The participants' constitutional DNA was sequenced using the PGRNseq assay, a targeted megabase of sequence in 82 PGx genes, genes identified as important for pharmacogenomics.

The eMERGEseq project was one of the major aims of the eMERGE Network during Phase III. It is aimed to identify rare variants with presumed major impact on function in a cohort of 25,000 participants across the Network. The Network created an eMERGE specific sequencing platform that is used to sequence participants at the individual sites. Baylor College of Medicine Human Genome Sequencing Center (HGSC) and Partners Healthcare with Broad Institute (the two sequencing centers) worked with the Clinical Annotation Workgroup and the Network sites to identify and validate an impactful set of genes and single nucleotide variants (SNVs) that allow for clinically actionable, pathogenic variants to be returned while providing researchers with the data needed to aid in genomic discovery. This resulted in a panel consisting of 109 genes and 1,551 SNVs.

To read more about the PGRNseq and eMERGEseq projects, visit the eMERGE network website [here](#).

#### What can I do with SPHINX?

See the lists of genes and drugs in the catalog

Search the catalog of variants by:

- Gene
- Drug interactions
- Chromosome position
- rsID (SNVid)

See for each gene:

- Observed single nucleotide variants
- Drug interactions
  - Other genes with that drug interaction
  - Other variants with that drug interaction

See for each variant:

- rsID (SNVid), where known
- Allele frequencies by European, African and Asian ancestry
- Variant category or "Type" (from SNPeff)
- Link to dbSNP and PharmGKB, where available

[Questions?](#)

**i** The sites participating in eMERGE and the eMERGE-PGx project include:

- Children's Hospital of Philadelphia
- Cincinnati Children's Hospital Medical Center with Boston Children's Hospital
- Essentia Rural Health with Marshfield Clinic and The Pennsylvania State University
- Geisinger
- Kaiser Permanente Washington with University of Washington
- Mayo Clinic
- Mount Sinai School of Medicine
- Northwestern University
- Vanderbilt University Medical Center

Sites participating in the eMERGEseq project include:

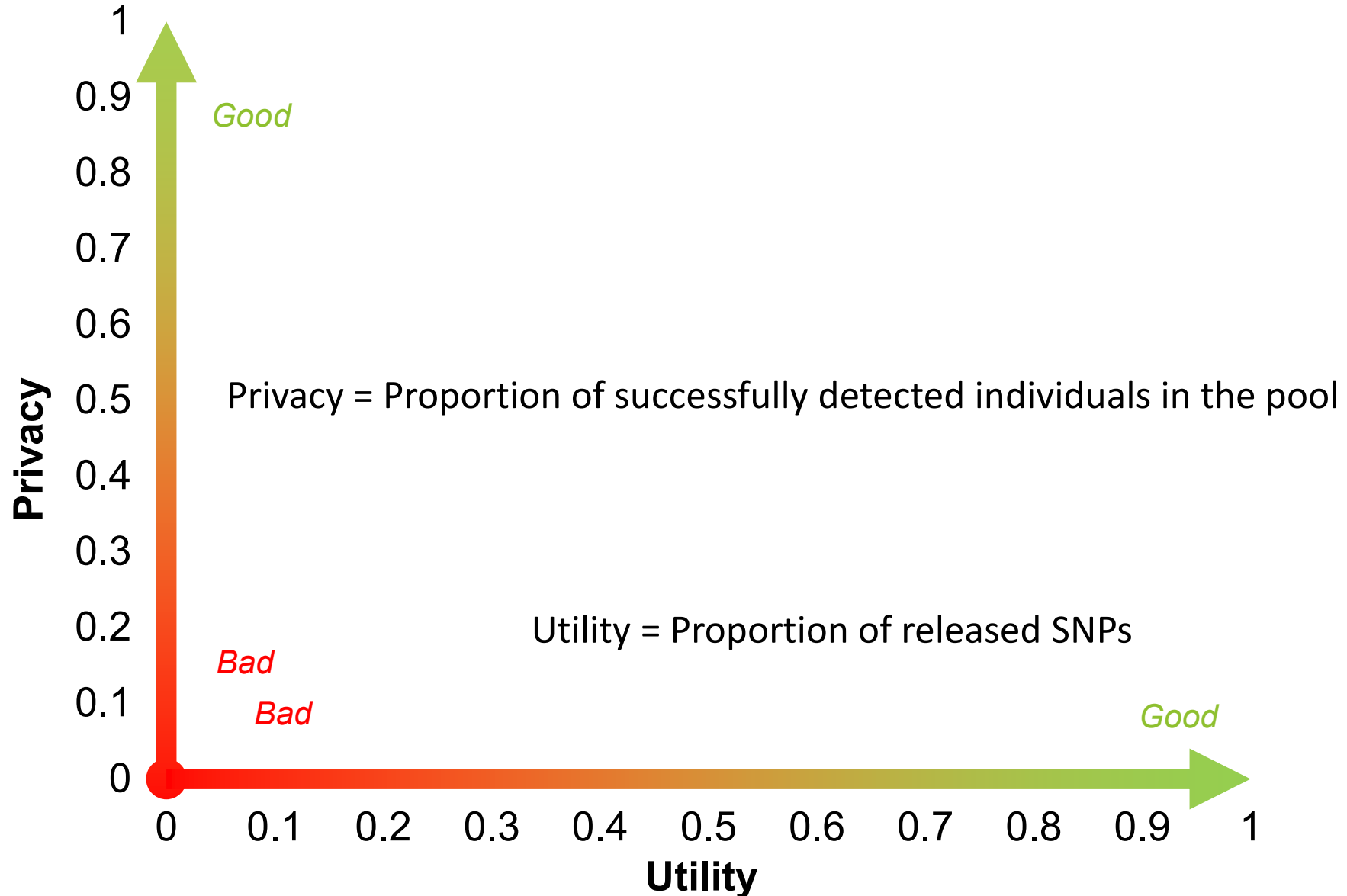
- Children's Hospital of Philadelphia
- Cincinnati Children's Hospital Medical Center with Boston Children's Hospital
- Columbia University
- Geisinger
- Harvard University
- Kaiser Permanente Washington with University of Washington
- Mayo Clinic
- Meharry Medical College
- Northwestern University
- Vanderbilt University Medical Center



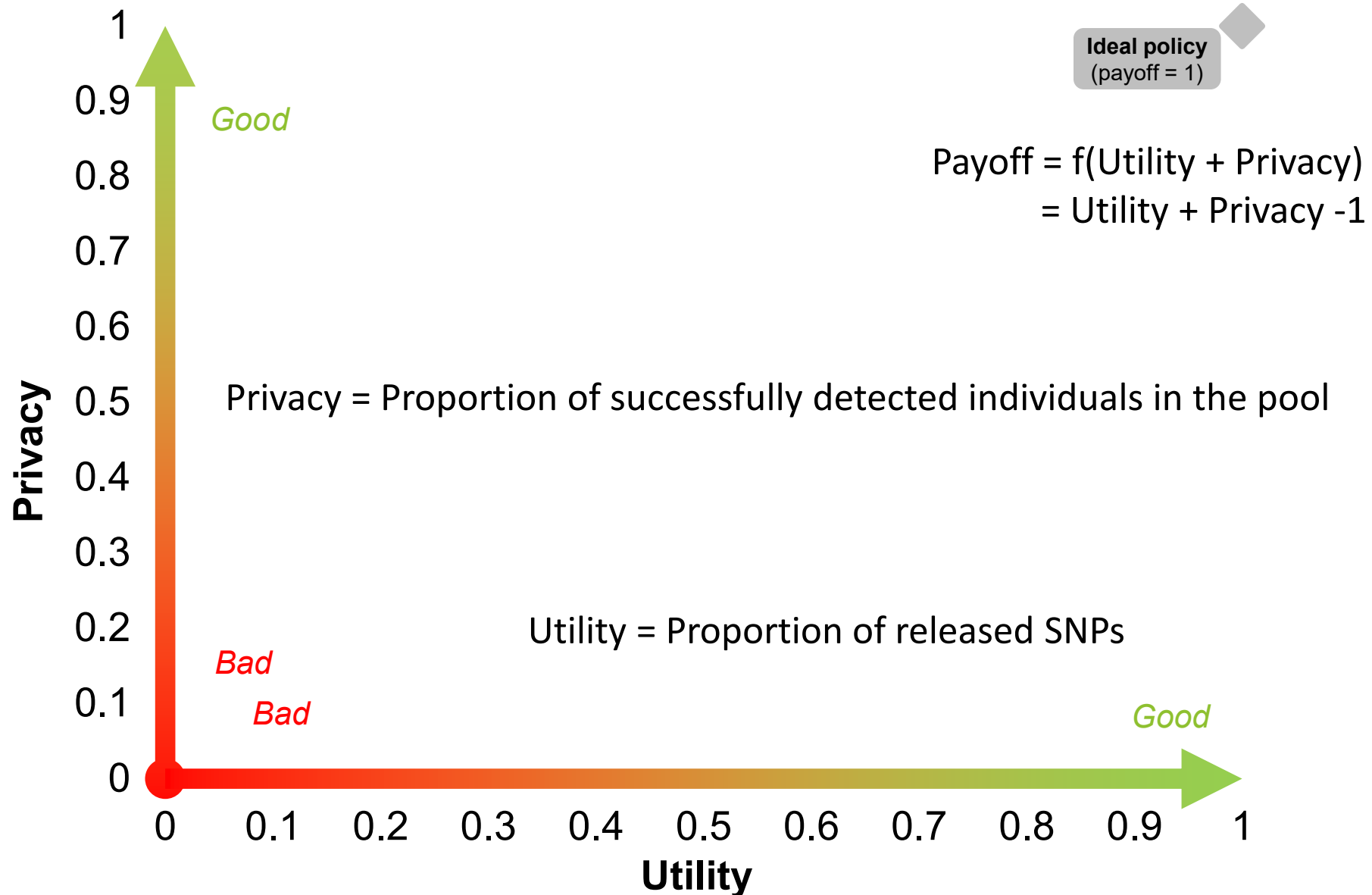
# EXPERIMENTS

- Dataset
  - 8,194 individuals in SPHINX
  - 2,504 individuals in 1000 Genome Project
  - 2,500 statistically independent SNPs to publish (total of 51,826 SNPs)
- Valuation settings:
  - \$45,000 for grant dollars (or the maximal benefit to the sharer)
  - \$360 for the benefit to the attacker for each successfully detected individual
  - \$180 for the expected penalty to the attacker per record
  - \$60 for the attacker's accessing cost per record

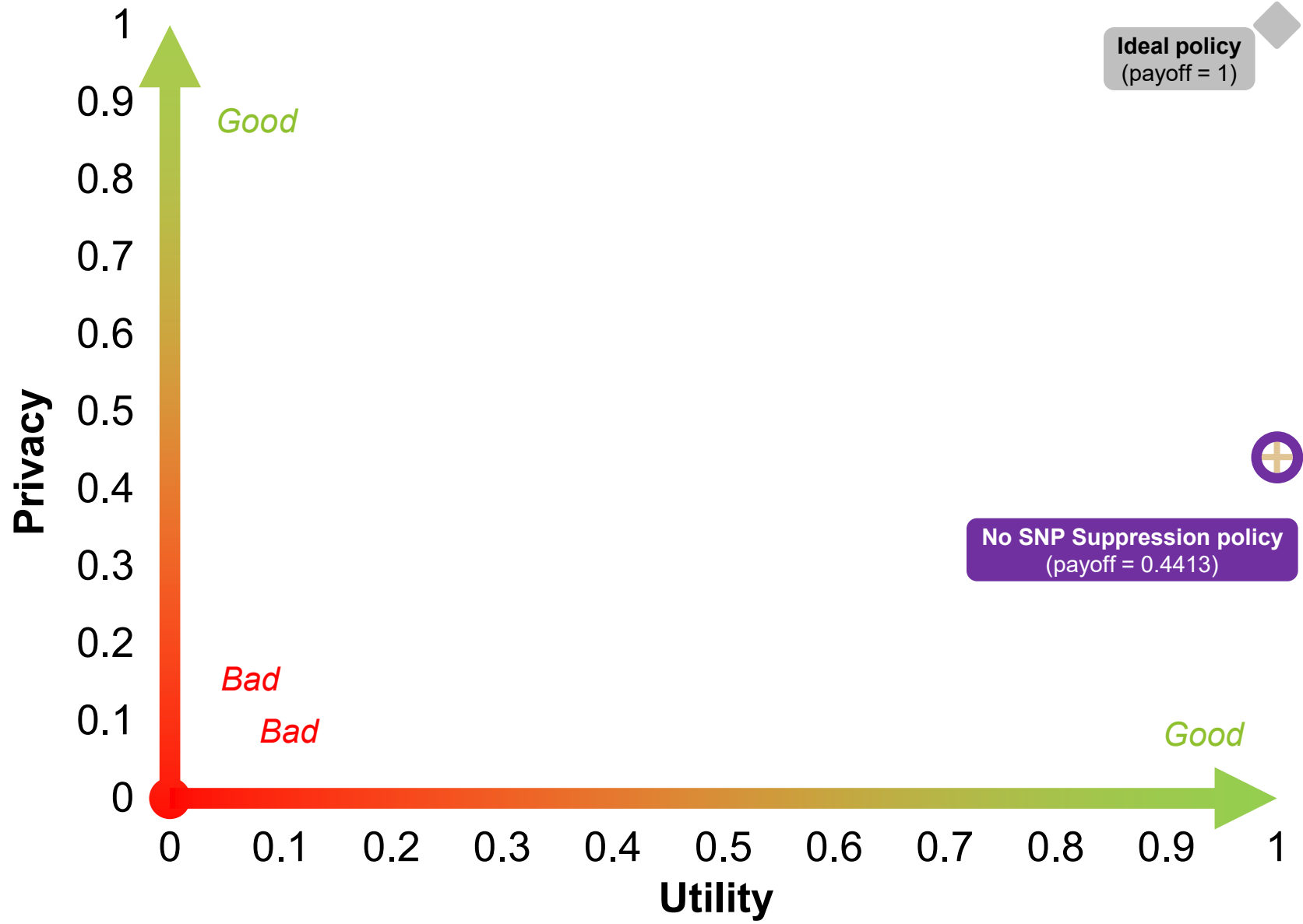
# SPHINX POLICY ANALYSIS



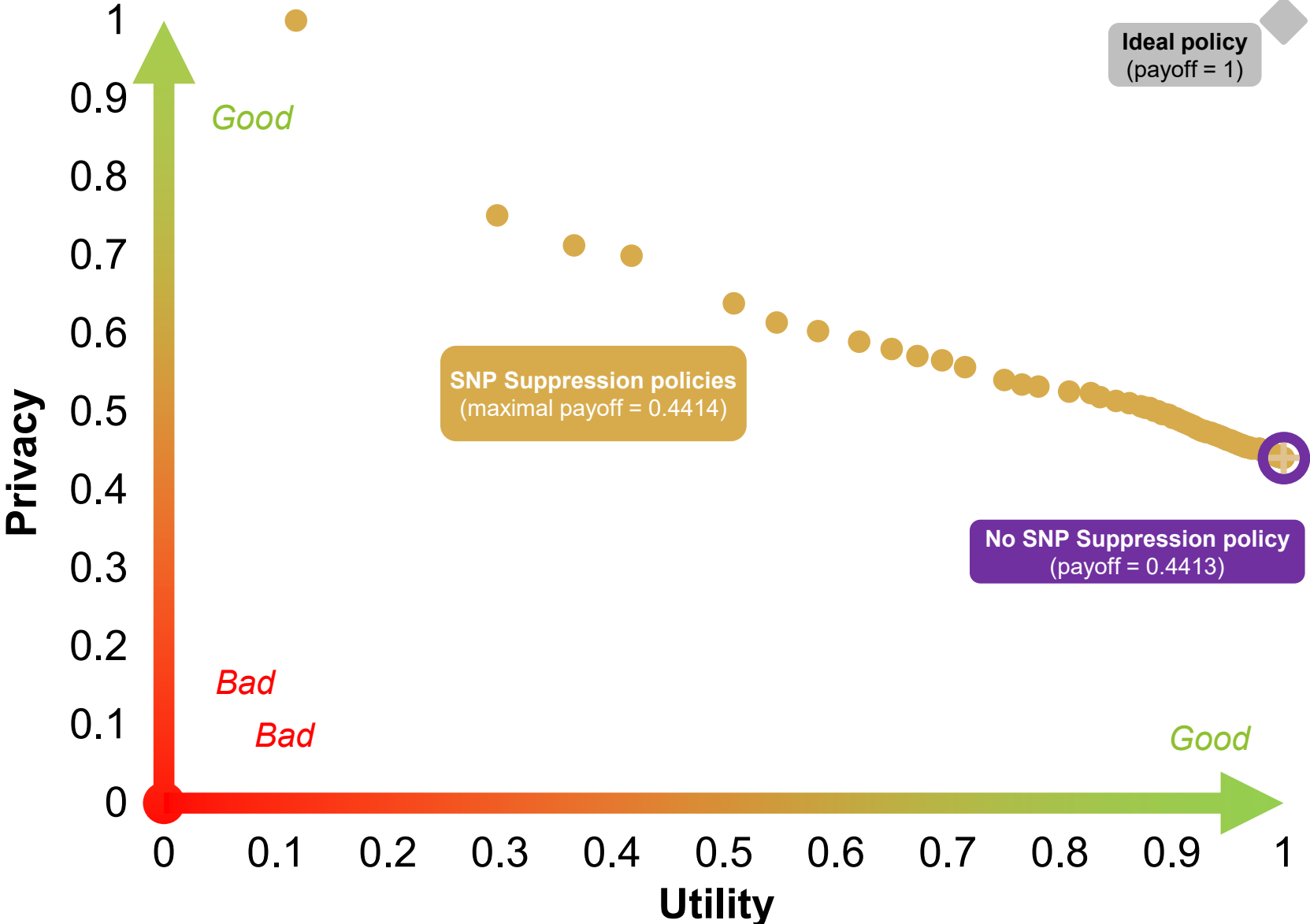
# SPHINX POLICY ANALYSIS



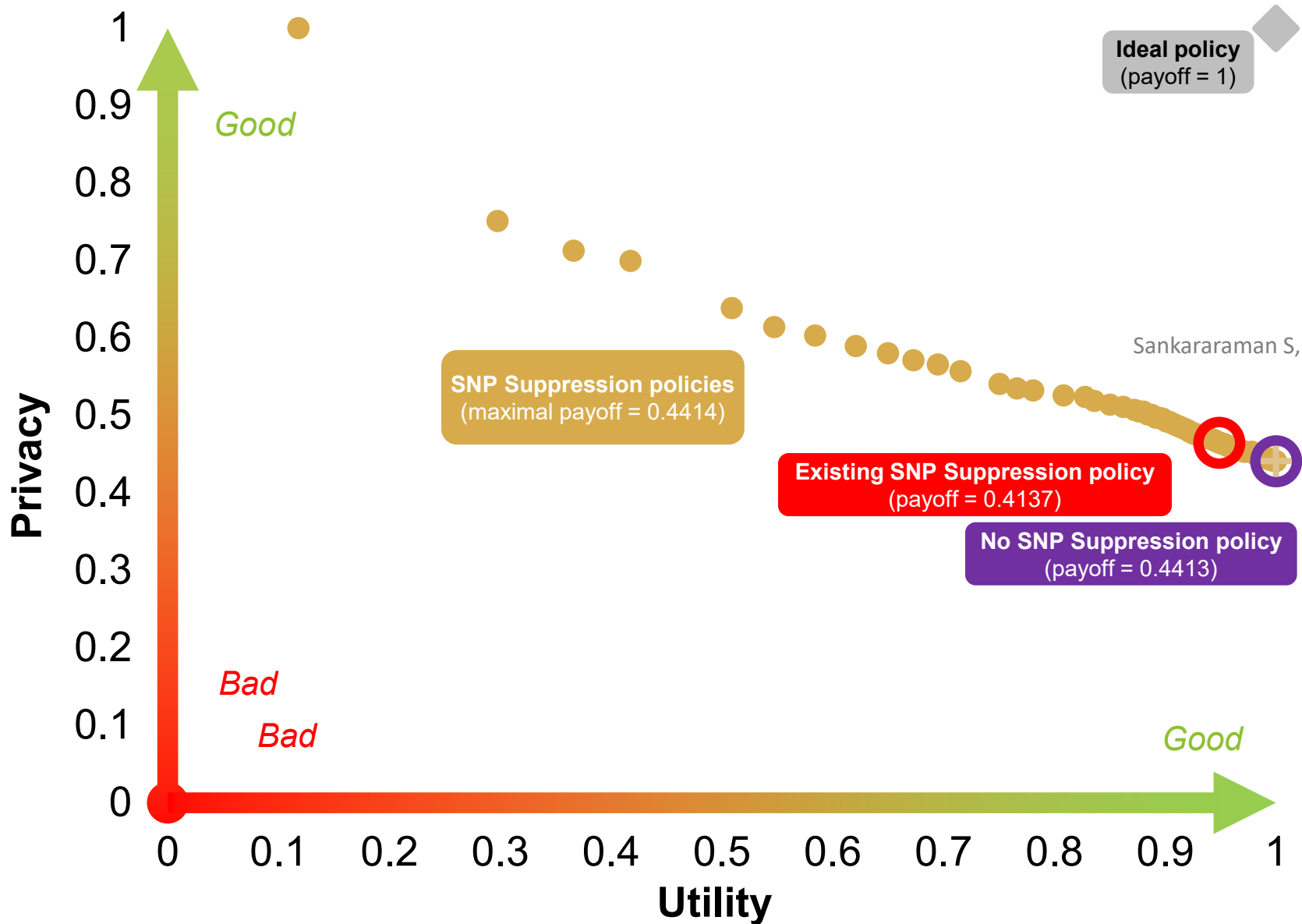
# SPHINX POLICY ANALYSIS



# SPHINX POLICY ANALYSIS

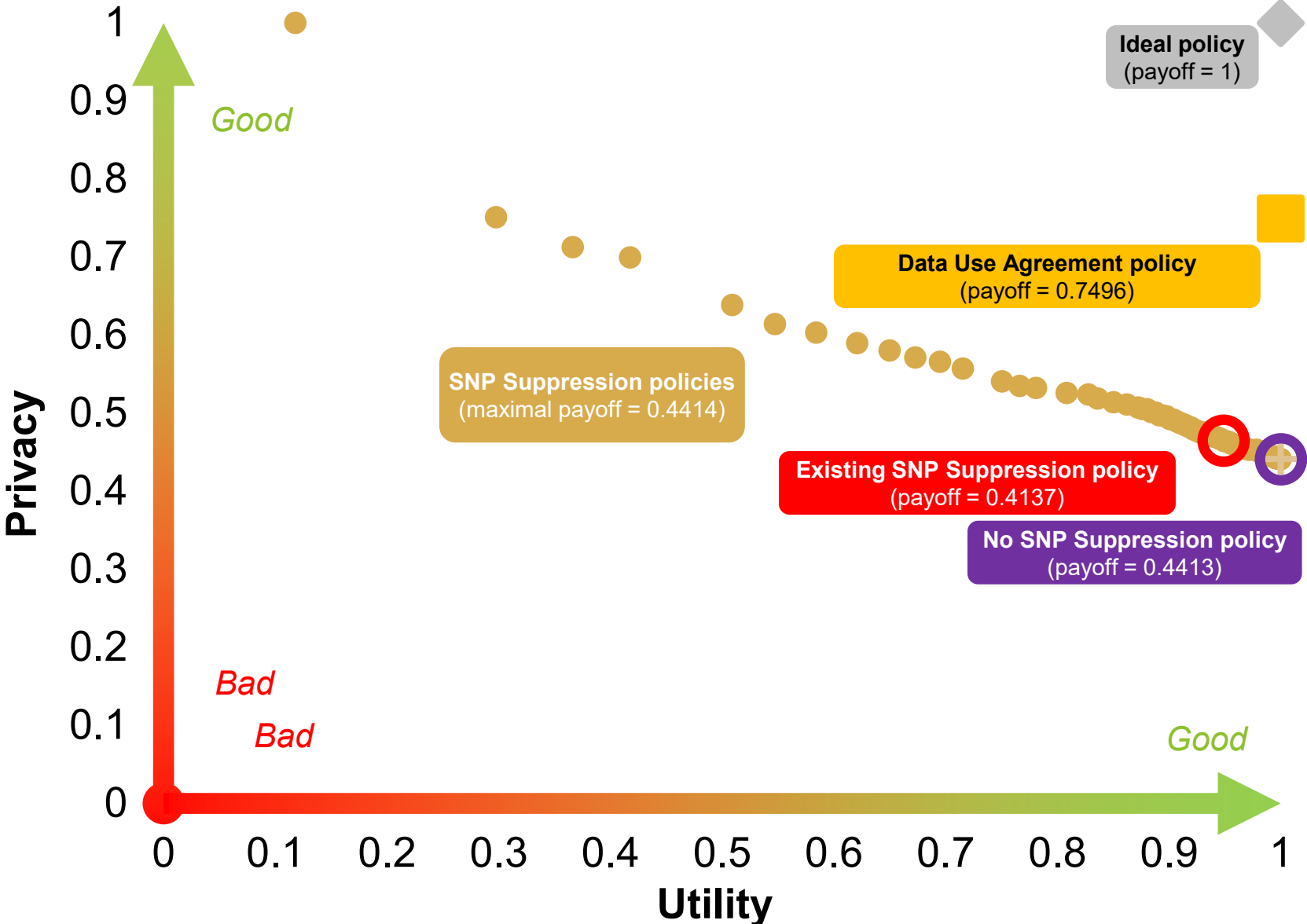


# SPHINX POLICY ANALYSIS

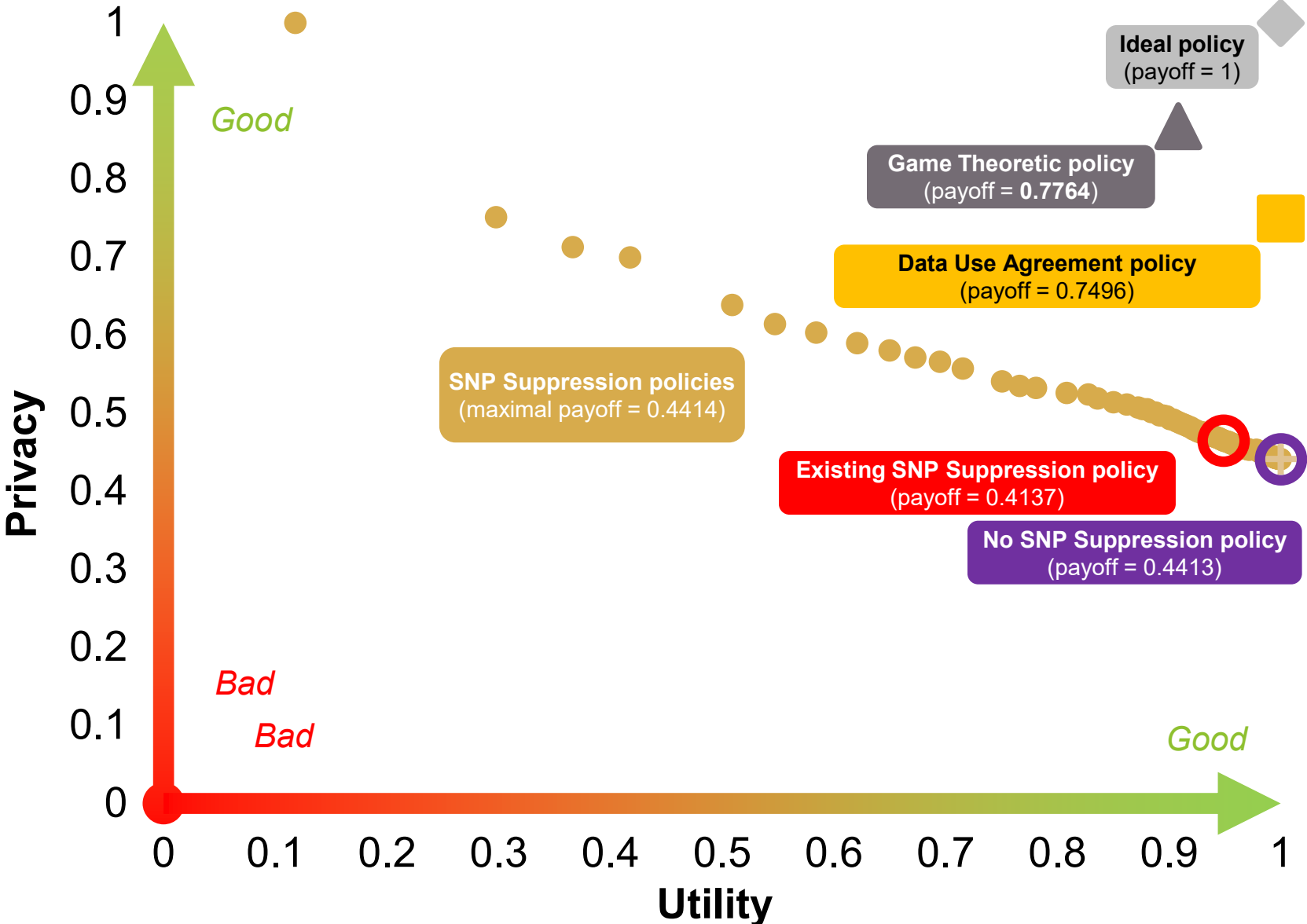


Sankararaman S, et al. Nature Genetics. 2009: 965-967.

# SPHINX POLICY ANALYSIS

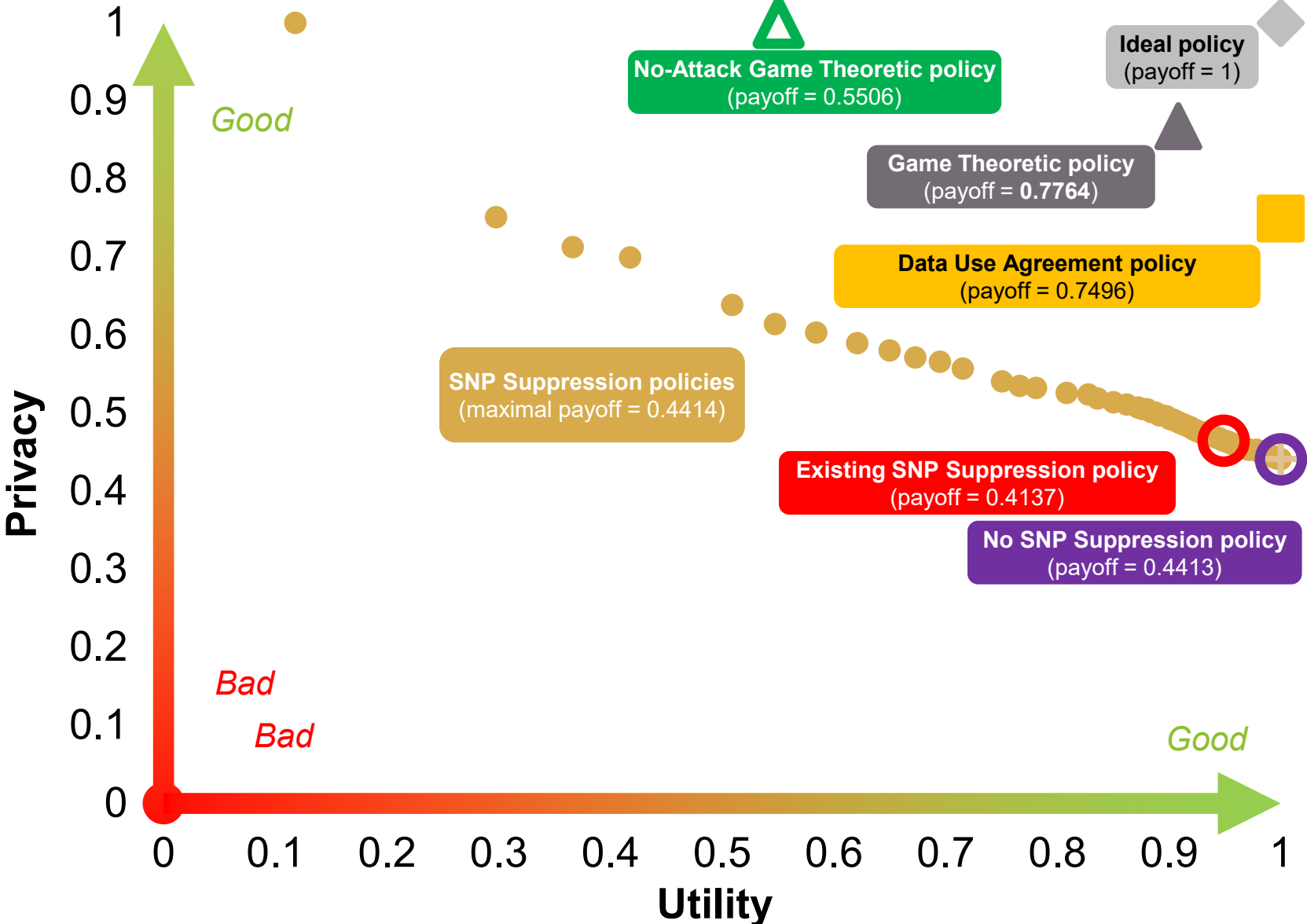


# SPHINX POLICY ANALYSIS

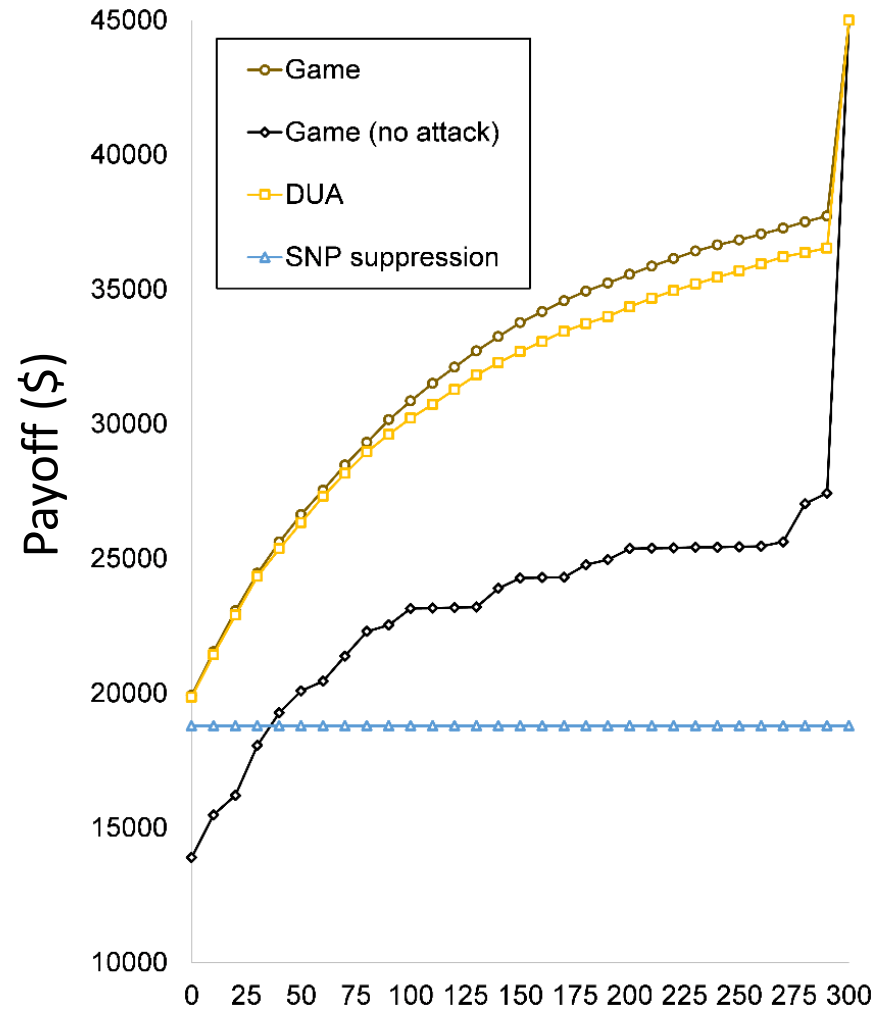




# SPHINX POLICY ANALYSIS

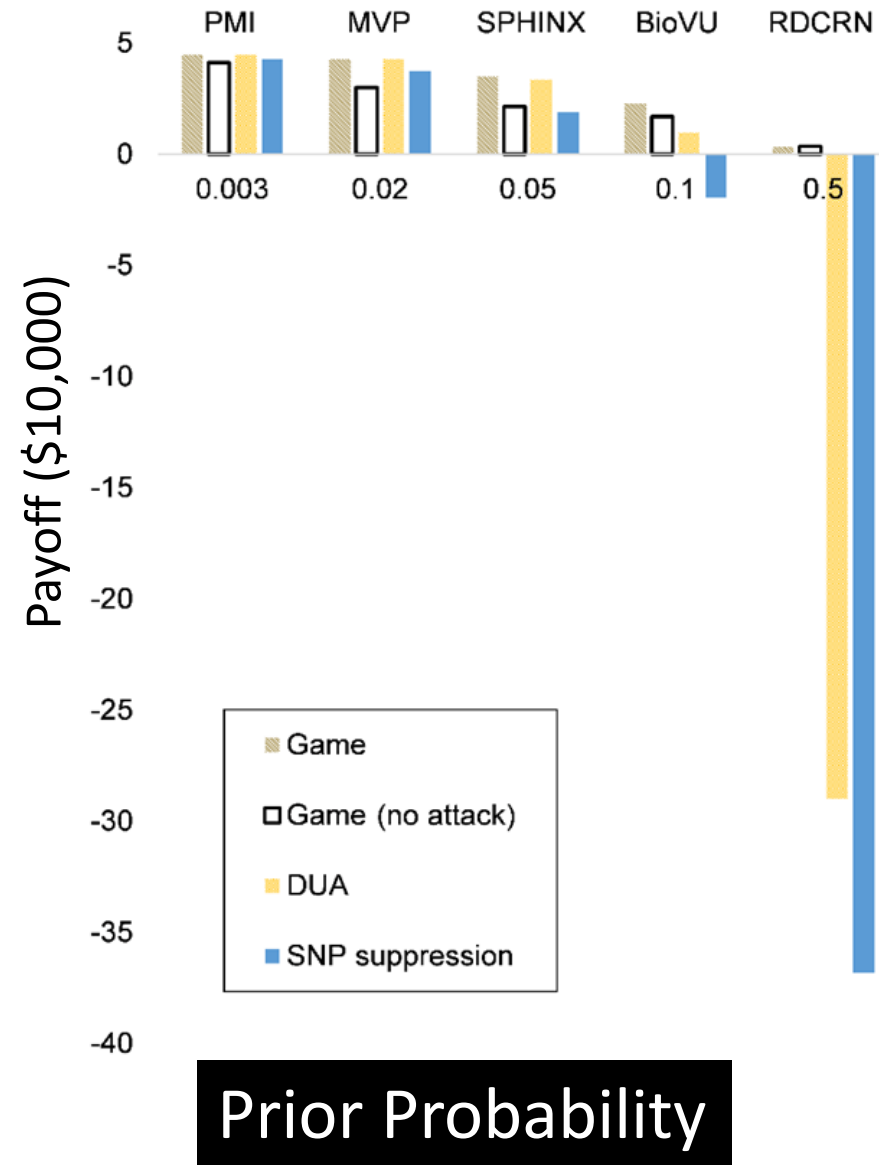


# SENSITIVITY ANALYSIS



Penalty

# SENSITIVITY ANALYSIS (CONT)



# CONCLUSIONS

- Findings

- The game-theoretic solution achieves the highest payoff for the data sharer
- The no-attack variation of the game can achieve a payoff higher than the state-of-the-art SNP-suppression strategy while eliminating privacy risk
- The game theoretic solution is not sensitive to the changes of key parameters such as the penalty and the prior probability

- Future Directions

- Valuation
- Multiple adversaries
- Irrational adversaries



Thank You!



VANDERBILT  
UNIVERSITY