

Balancing Privacy and Utility in Synthetic Genotype Data Generation by Adaptive Optimization

Mohamed Sabri Hafidi

Anton Dignös

Johann Gamper

Ozan Kahramanoğulları

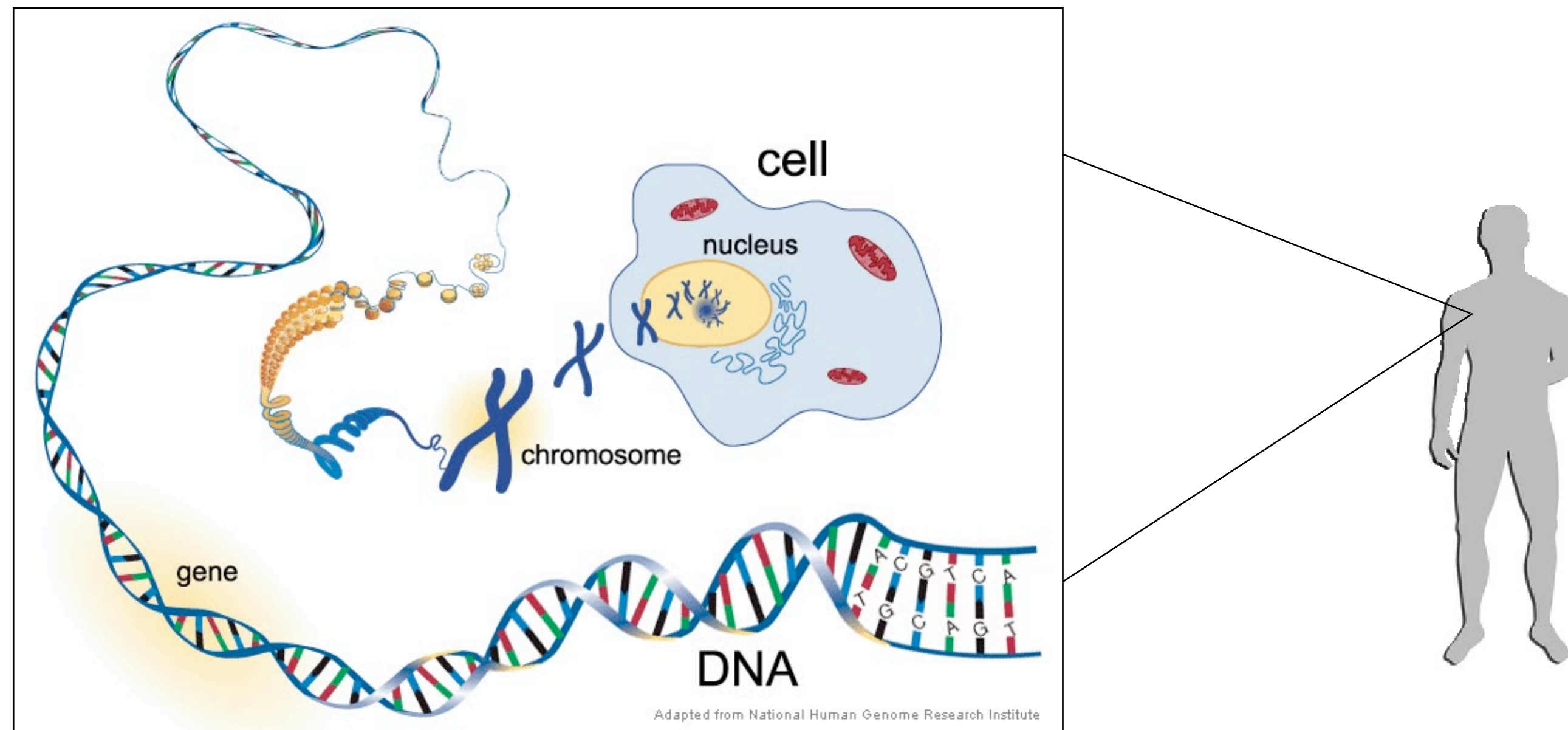
Free University of Bozen-Bolzano, Italy

This work was funded in part by the Autonomous Province of Bozen-Bolzano, call Joint Projects South Tyrol – Germany, project DyHealthNet – CUP: I53C22002980003.

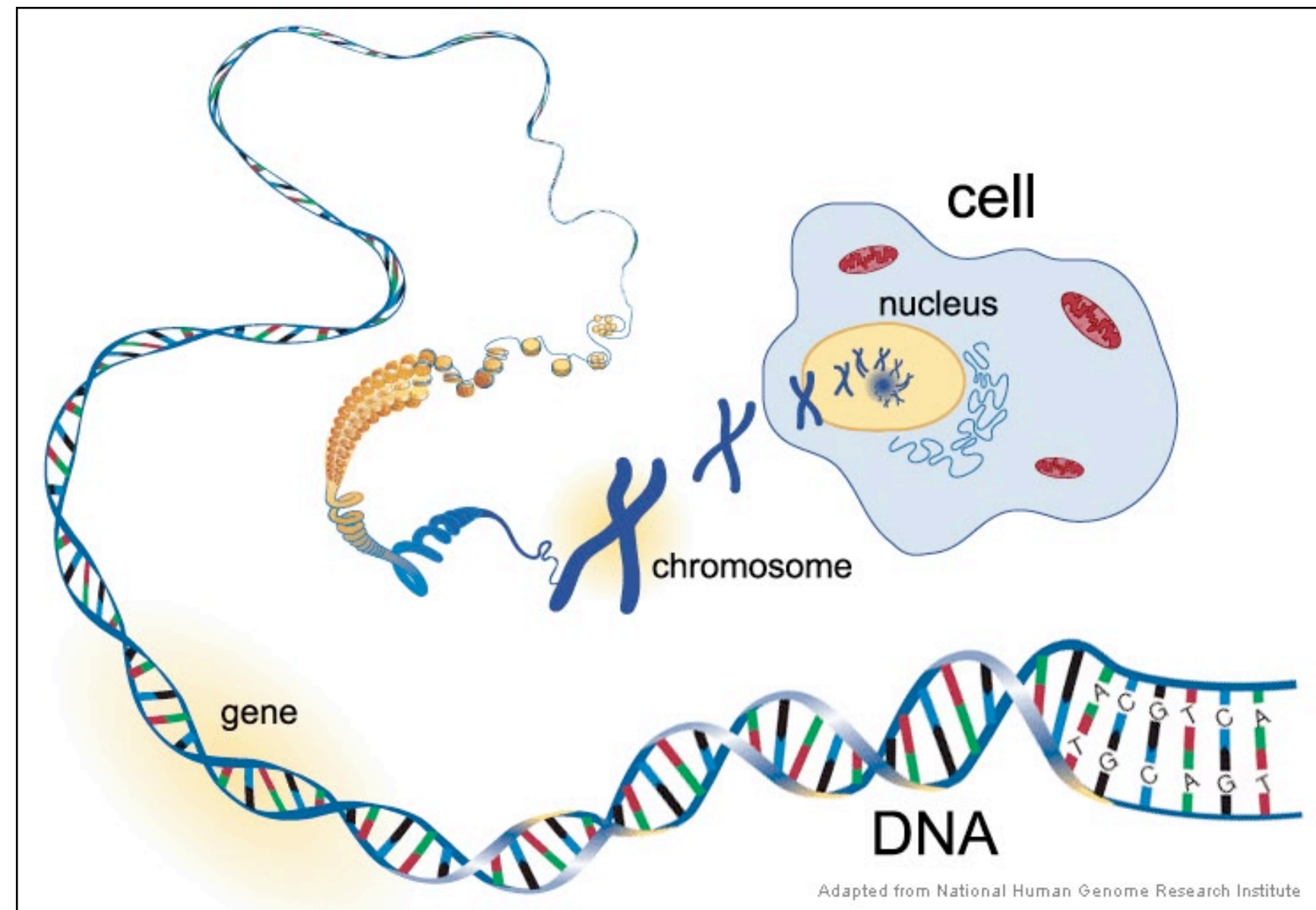


Autonome Provinz Bozen
Provincia autonoma di Bolzano
Provincia autonoma de Bulsan
SÜDTIROL • ALTO ADIGE

Understanding Genetic Data



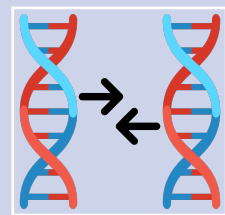
Understanding Genetic Data



The Variant Call Format - VCF



Standardized text file for genomic sequence variations



Records differences from a reference genome



Includes metadata, variant annotations, and sample genotypes

VCF File

Meta-Information

```
##fileformat=VCFv4.5
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Header

Genotypes

Samples

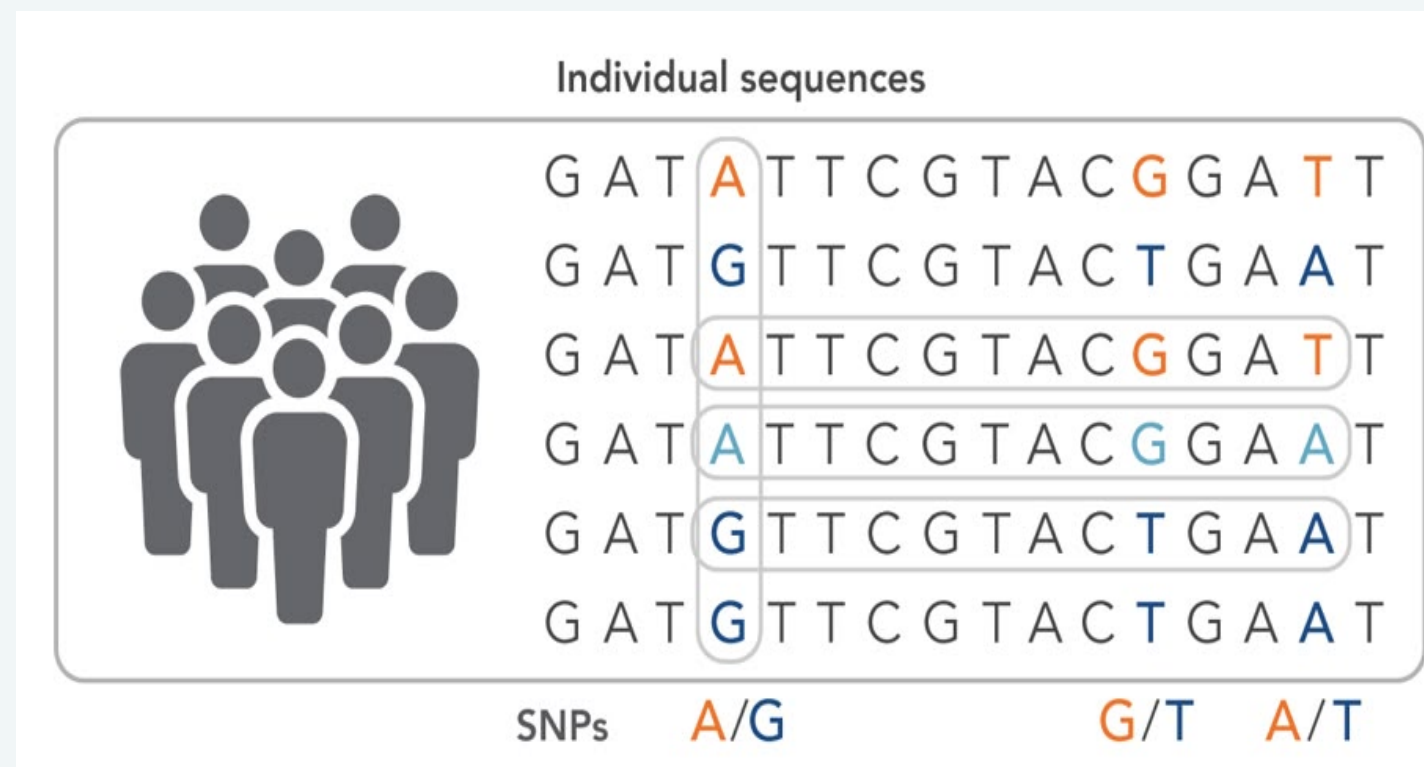
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA000001	NA000002	NA000003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variants

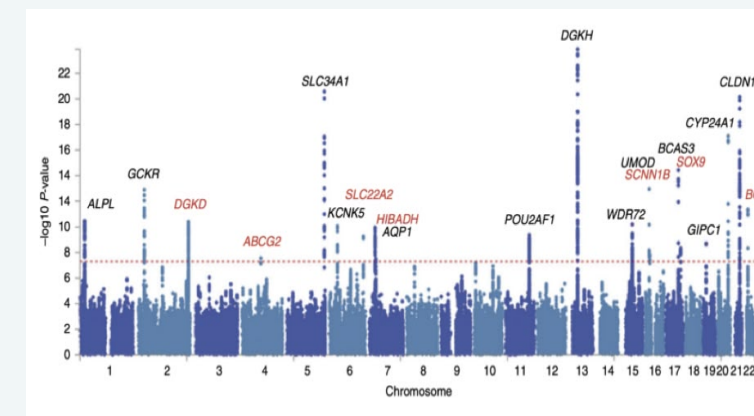
Fixed Fields

Why Genomic Data Matters

Genotyping

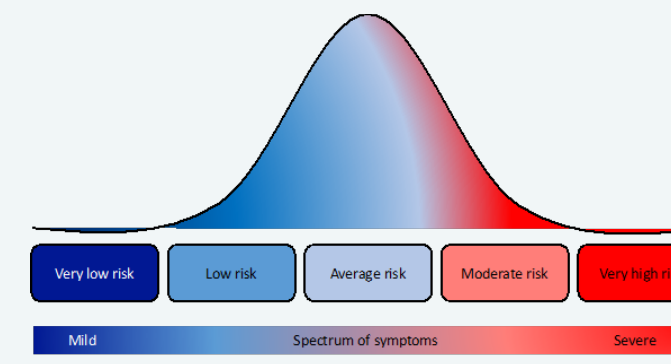


Variants across thousands of people



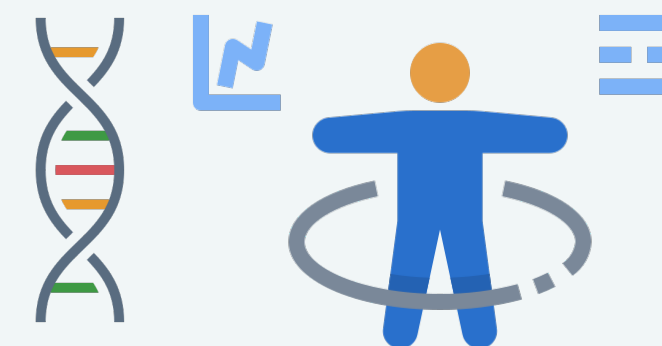
GWAS

Which variants associate with disease?



Polygenic risk score

One risk estimate per person



Personalized medicine

Prevention and treatment tailored to you

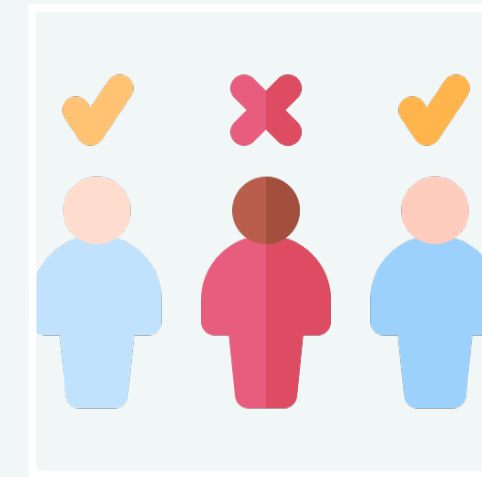
All of this depends on sharing genomes at scale.

Why Genomic Data Needs Protection



Highly personal and sensitive

Reveals health, ancestry and disease predispositions



Risk of discrimination

Misuse in insurance or employment decisions



Re-identification

Anonymized data can be linked to other datasets

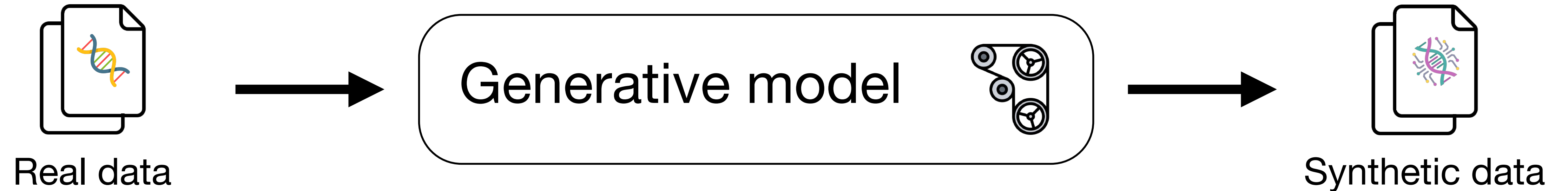


Legal and ethical duty

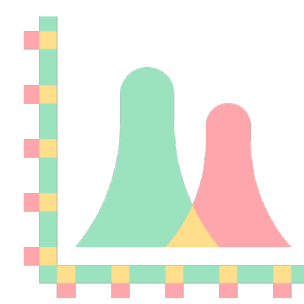


How can we share genomic data without exposing individuals?

Synthetic Genomic Data: A Path Forward

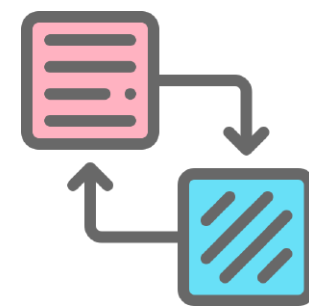


Synthetic Data Generation methods



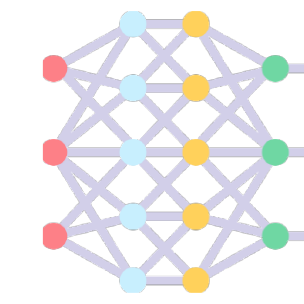
Statistical distribution

- Calculate frequencies
- Sample from the observed distribution



Algorithmic construction

- Apply invertible transformations
- e.g., shuffling



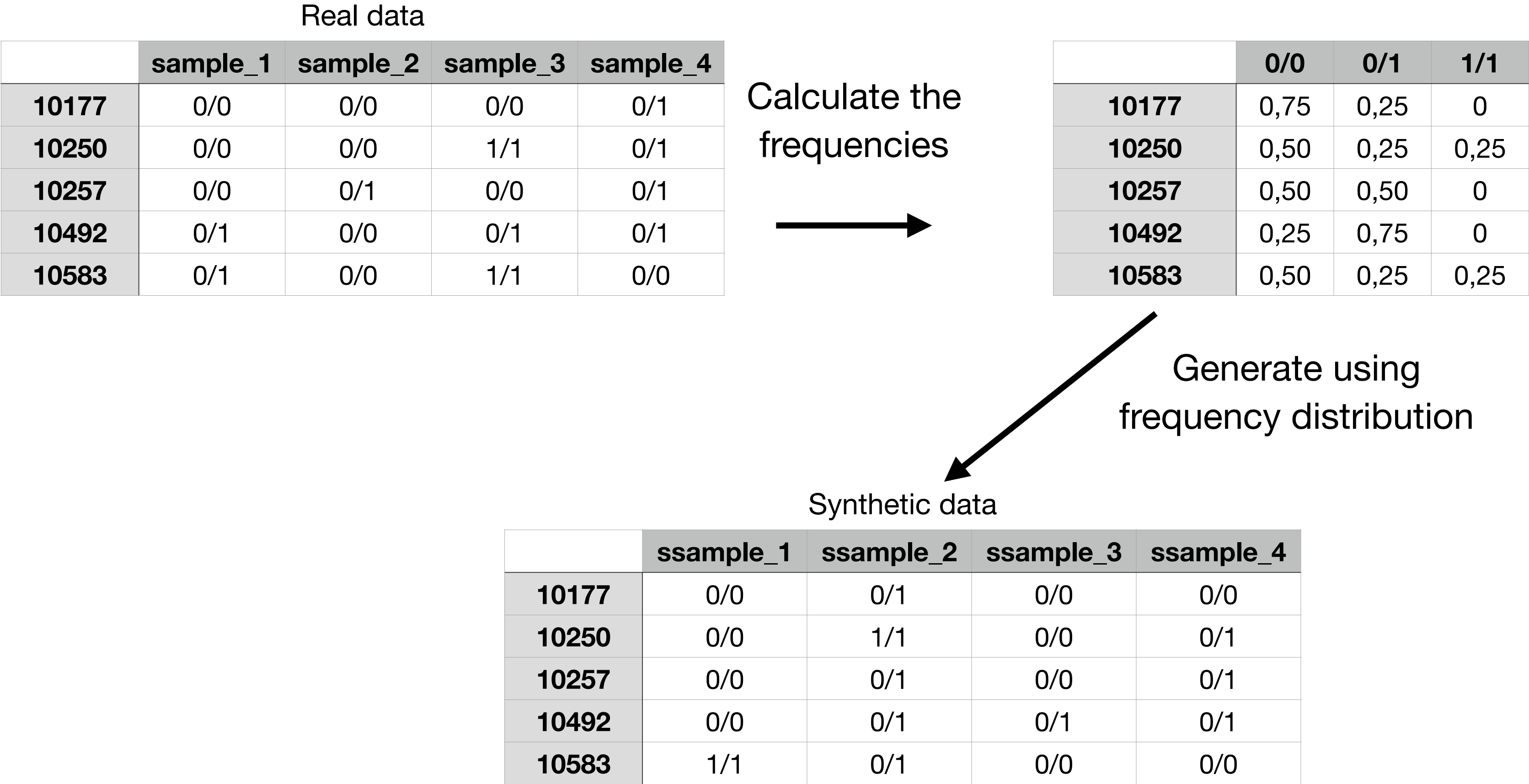
Generative Adversarial Networks

- Simulate patterns using neural network
- Generate from noise

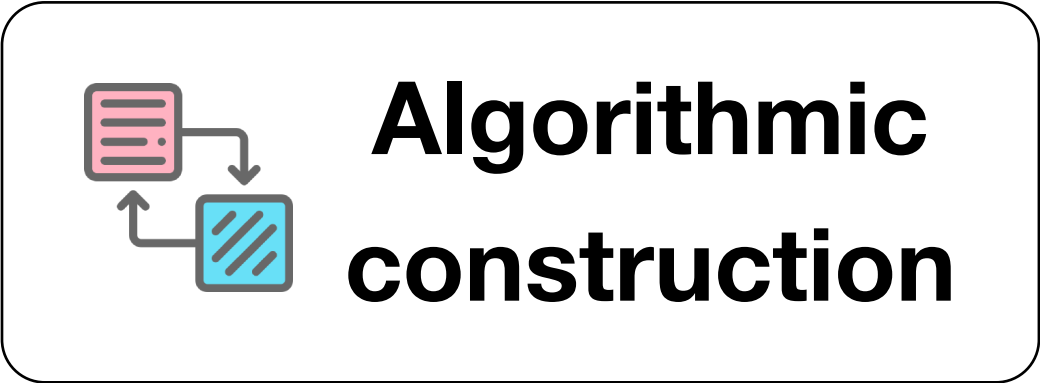
Concept



- Overview:



Concept



- **Overview:**

Real data

	RS001	RS002	RS003	RS004
10177	0/0	0/0	0/0	0/1
10250	0/0	0/0	1/1	0/1
10257	0/0	0/1	0/0	0/1
10492	0/1	0/0	0/1	0/1
10583	0/1	0/0	1/1	0/0

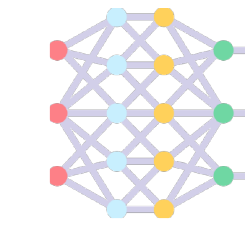
Shuffle the
genotype



Synthetic data

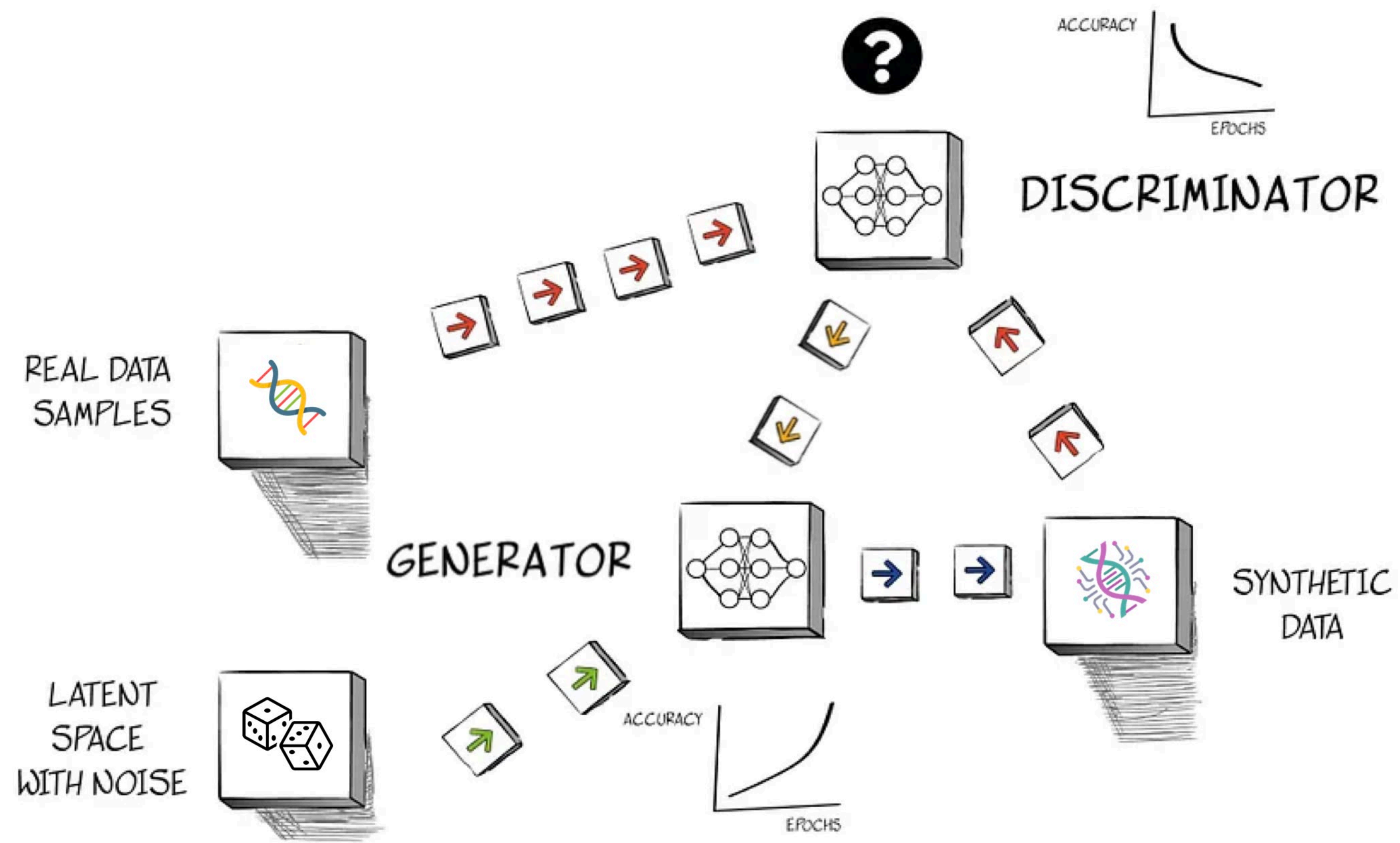
	SS001	SS002	SS003	SS004
10177	0/0	0/0	0/0	0/1
10250	1/1	0/1	0/0	0/0
10257	0/1	0/0	0/0	0/1
10492	0/1	0/1	0/0	0/1
10583	0/0	1/1	0/0	0/1

Concept



Generative Adversarial Networks

- Overview:



How to balance utility and privacy in genotype data generation?

Main Contributions



Adaptive model tuning: rollback to the best state when privacy or utility degrades



WGAN-GP for genotypes: handles discrete, high-dimensional data with stable training



Efficient worst-case privacy test: using universal shadow pool

Adaptive model tuning

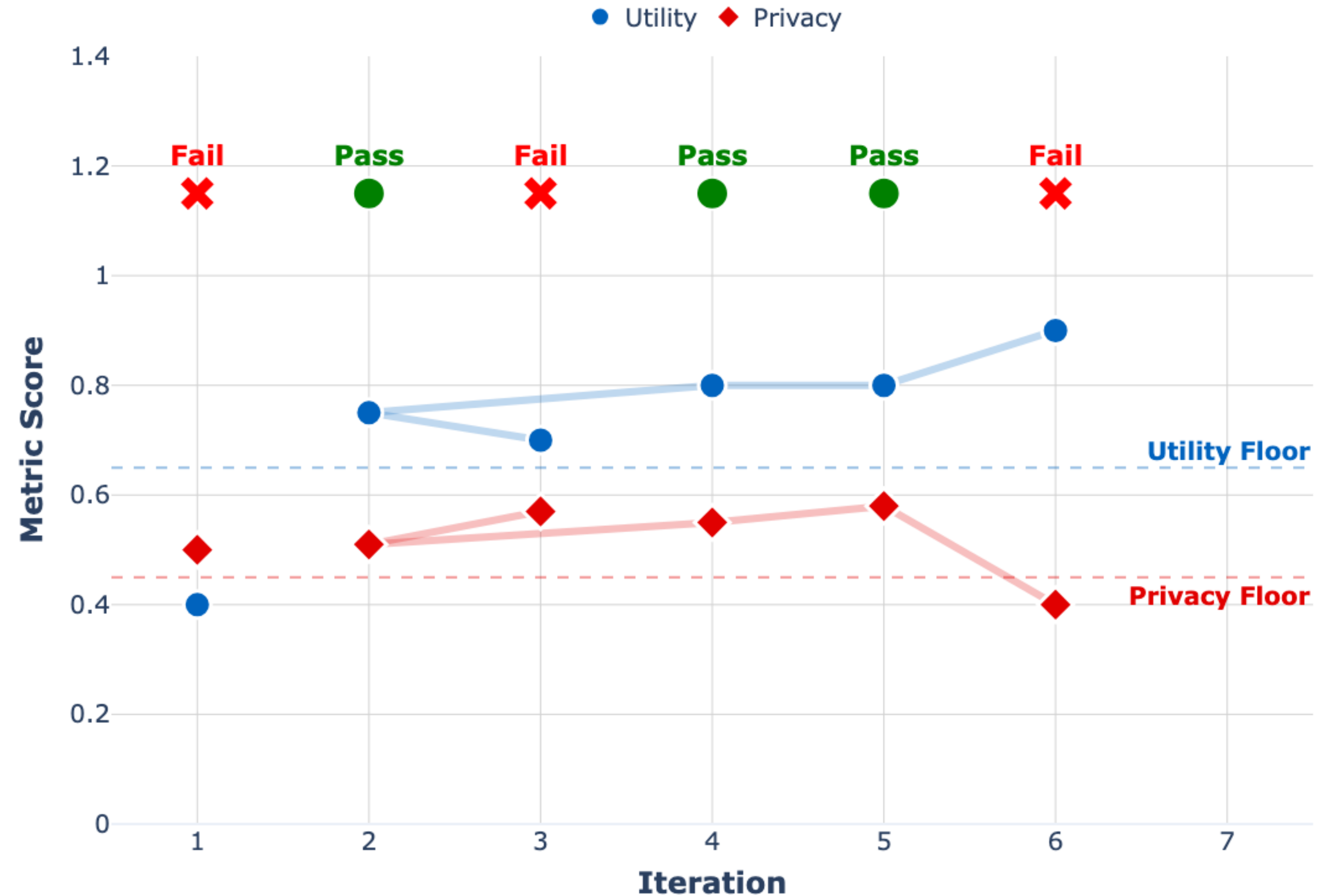
```
1:  $S \leftarrow \emptyset$ 
2:  $f_{best} \leftarrow f$                                 ▷ Store initial state as the baseline
3: for  $i = 0$  to  $N$  do
4:    $S_{new} \leftarrow f(R_{orig})$ 
5:   if  $\mathcal{P}(S_{new}) \geq \delta_p$  and  $\mathcal{U}(S_{new}) \geq \delta_u$  then
6:     if  $S = \emptyset$  or  $(\mathcal{U}(S_{new}) > \mathcal{U}(S))$  or  $(\mathcal{U}(S_{new}) = \mathcal{U}(S) \text{ and } \mathcal{P}(S_{new}) > \mathcal{P}(S))$  then
7:        $S \leftarrow S_{new}$ 
8:        $f_{best} \leftarrow f$                                 ▷ Save the model state that produced the best result
9:       Update  $f$                                 ▷ Continue optimizing from a successful state
10:    else                                ▷ Failure: Revert to the last acceptable state before updating.
11:       $f \leftarrow f_{best}$                                 ▷ Rollback to last known good configuration
12:      Update  $f$                                 ▷ Try a different tuning path
13: if  $S = \emptyset$  then
14:   return  $\emptyset, \emptyset, 0, 0$ 
15: else
16:   return  $S, f_{best}, \mathcal{P}(S), \mathcal{U}(S)$ 
```

Adaptive model tuning

```

1: Initialize:  $S_{best} \leftarrow \emptyset, f_{best} \leftarrow f_{base}$ 
2: for iteration  $i = 1 \dots N$  do
3:    $S_{new} \leftarrow \text{Generate}(f)$ 
4:   if  $S_{new}$  meets Privacy & Utility Floors then
5:     if  $S_{best} = \emptyset$  or  $\mathcal{U}(S_{new}) > \mathcal{U}(S_{best})$  or ( $\mathcal{U}(S_{new}) =$ 
 $\mathcal{U}(S_{best})$  and  $\mathcal{P}(S_{new}) > \mathcal{P}(S_{best})$ ) then
6:        $S_{best} \leftarrow S_{new}$ 
7:        $f_{best} \leftarrow f$            ▷ Update Safe Checkpoint
8:        $f \leftarrow \text{Update}(f)$      ▷ Exploit: Continue Optimization
9:   else
10:     $f \leftarrow f_{best}$            ▷ Floor Violation Detected
11:     $f \leftarrow \text{Update}(f)$      ▷ Rollback to last Safe State
12:                                     ▷ Explore: New trajectory
12: Output: Final Model  $f_{best}$ 

```



Utility Evaluation

Diversity

Are individuals as different from each other as in the real cohort?

Metric: KS distance

Variant frequencies

Is each variant as common in the synthetic data as in the real data?

Metric: MAF MAD

Relationships between variants

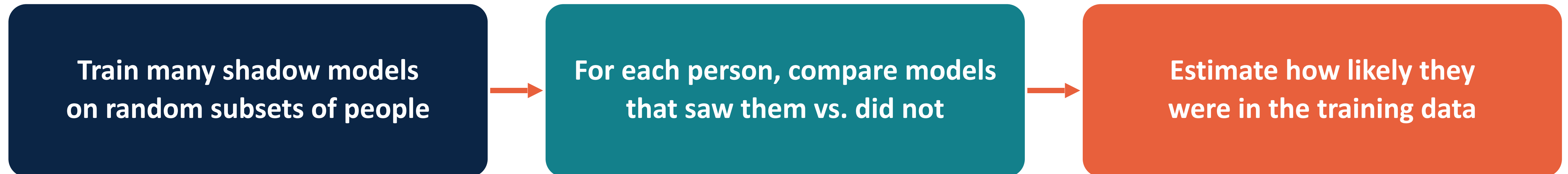
Do variants that occur together in real data also occur together in synthetic data?

Metric: LD PCC

Does the synthetic cohort look and behave like the real one?

Privacy Evaluation

Attacker's question: was this person in the training data?



Classic attack

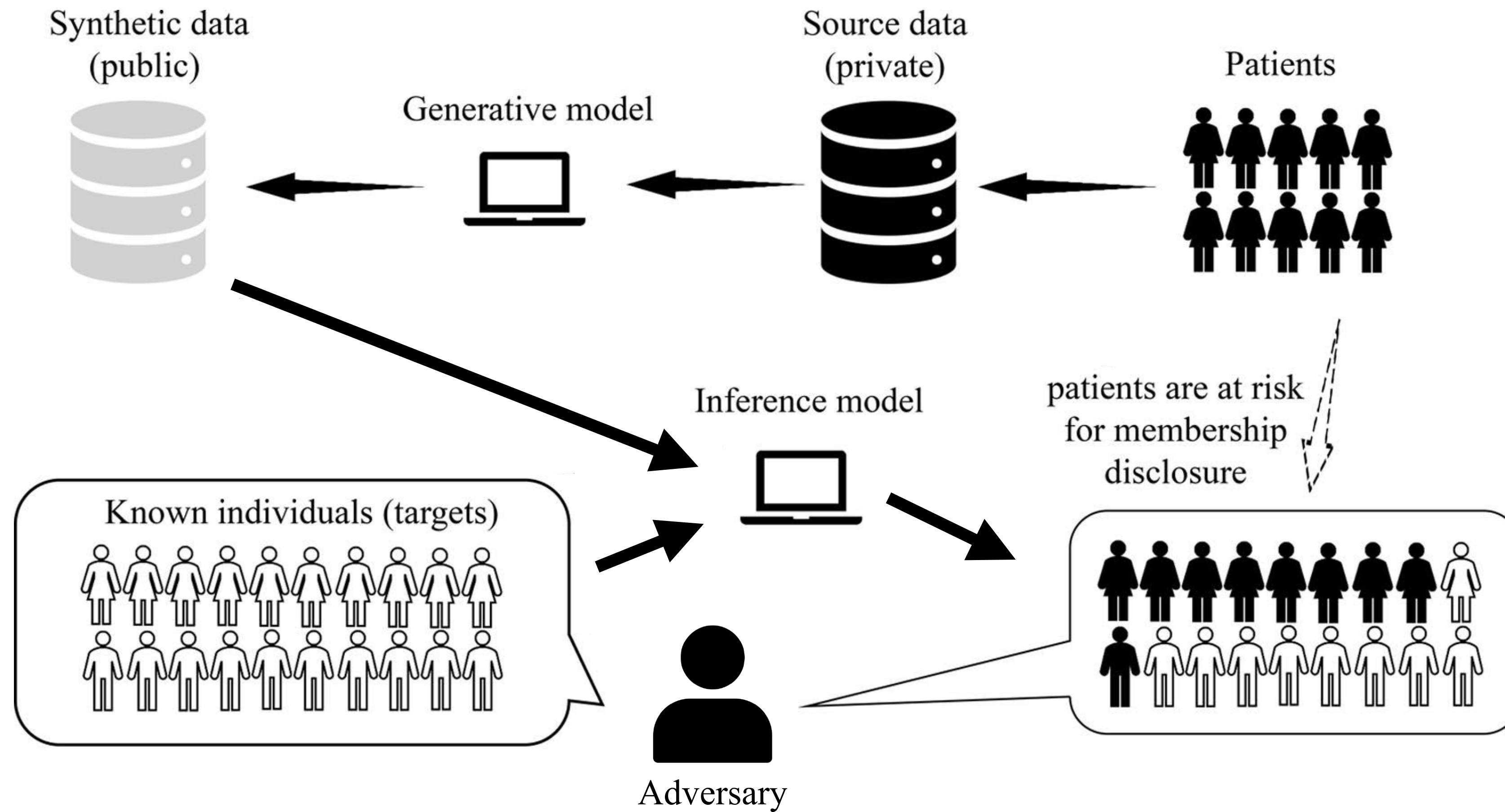
New helper models for every target person: too costly for large cohorts.

Our shared pool

Trained once and reused for everyone: the whole cohort can be checked.

One pool of models, every individual checked.

Membership Inference Attacks



Privacy Score: The Weakest Link

Dataset privacy = protection of the least-protected person



0: someone is exposed

1: strong privacy

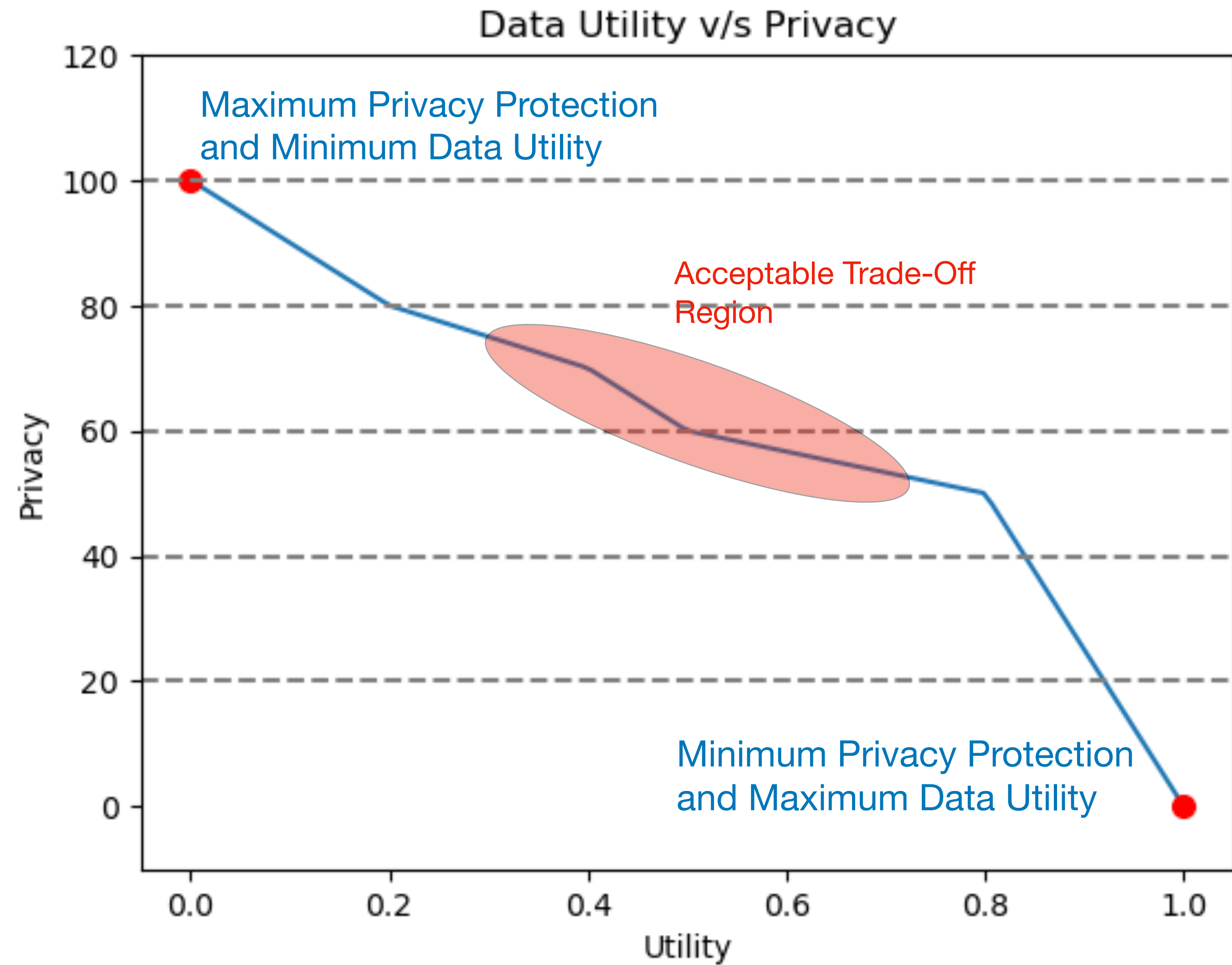
Worst case, not average

Protects the most exposed individuals in the cohort.

Acceptance rule

A candidate dataset is kept only if its score stays above a threshold (80% in our experiment).

Balancing Utility and Privacy



Experimental Setup

MGP

Mouse strains

80

individuals

42,812

variants

Missing values

None

Genotype types

3

Clean baseline

HapMap3

Human, imputed

1,184

individuals

1,387,466

variants

Missing values

Imputed

Genotype types

5

Stable, quality-filtered

1000 Genomes Chr 22

Human, real-world

2,504

individuals

1,103,547

variants

Missing values

Yes

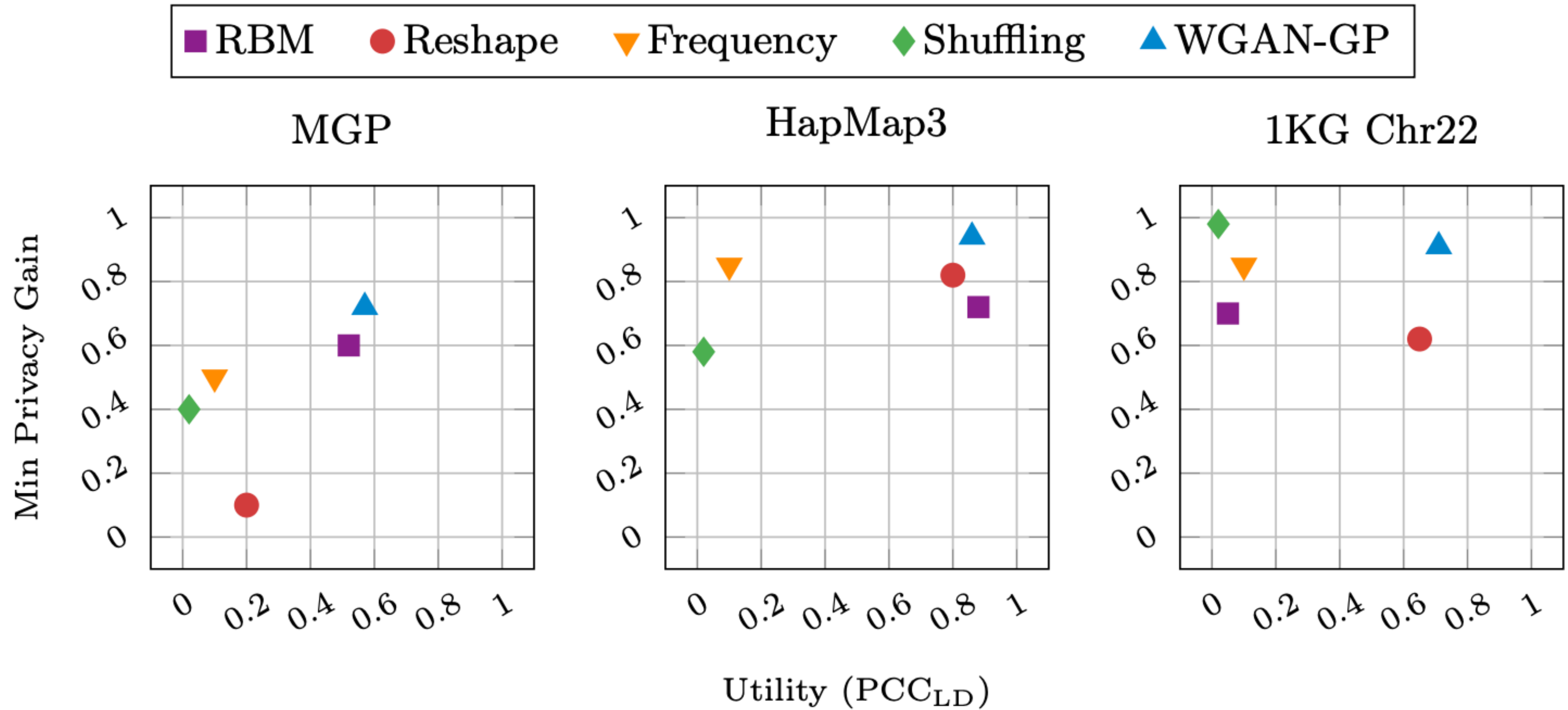
Genotype types

23

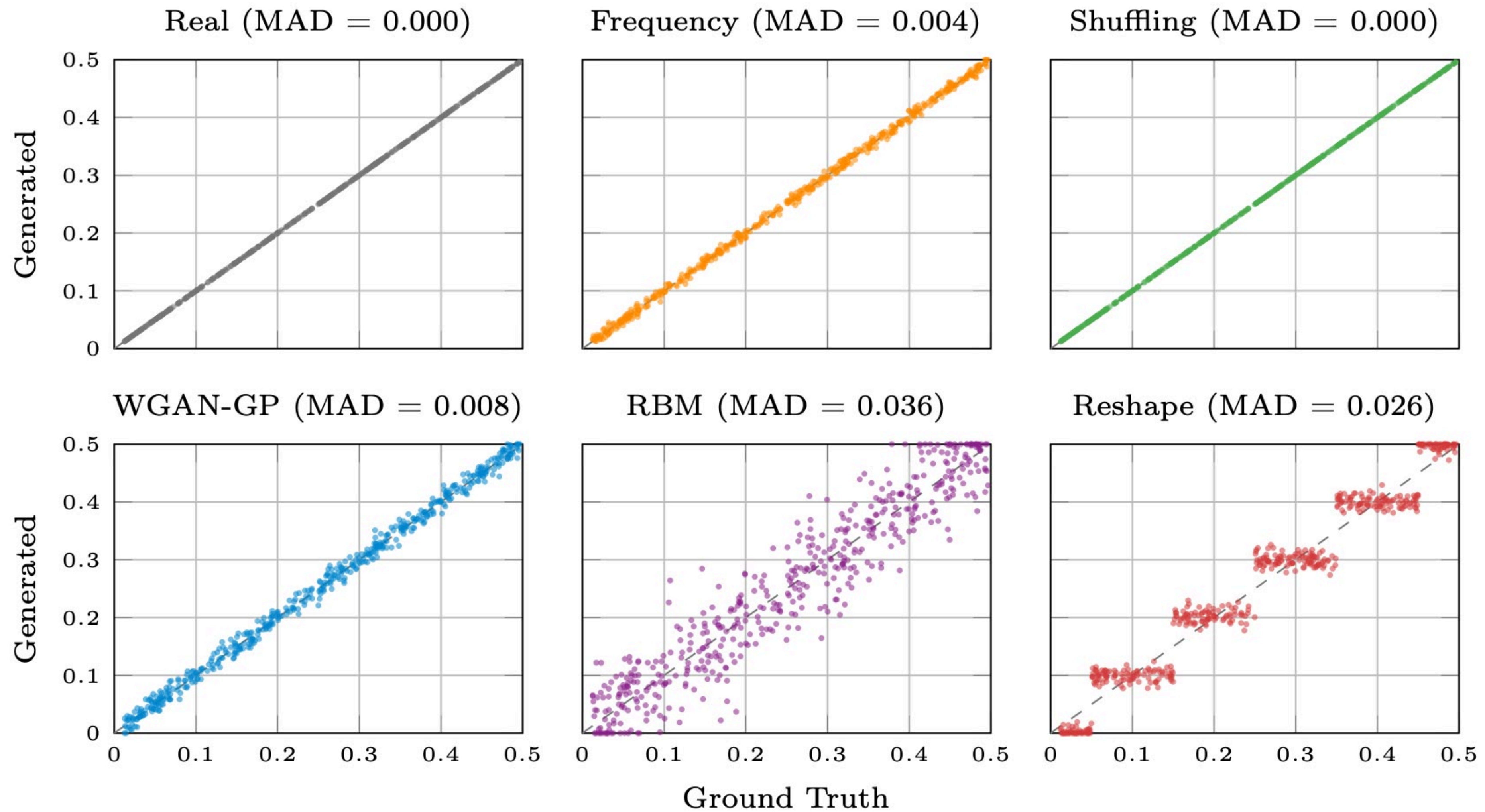
Most challenging

Increasing difficulty

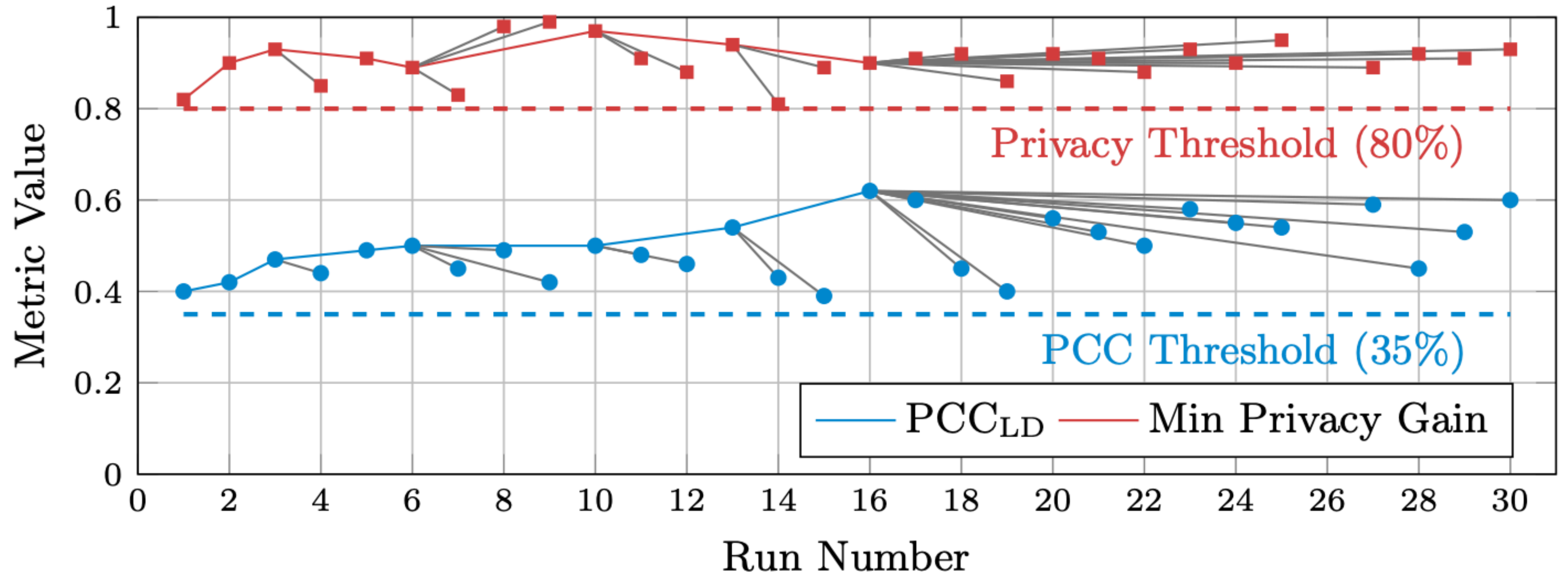
Experimental Evaluation



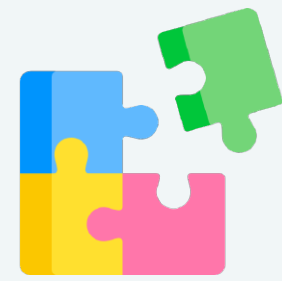
Experimental Evaluation



Experimental Evaluation



Conclusion



A plug-and-play framework

Works with any generator and any privacy or utility metric



Rollback keeps tuning stable

Recovers a good state whenever quality drops



High utility with strong privacy

Tuned WGAN-GP on three datasets, stable across repeated runs



Privacy measured at the worst case

The most exposed individual sets the score



Thank you!

Mohamed Sabri Hafidi

hmohamedsabri@unibz.it

Source code:

