RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory

Ensuring security and privacy of large-scale datasets: a synthetic data perspective

Jaideep Vaidya

Management Science and Information Systems

jsvaidya@rutgers.edu

Ensuring security and privacy of large-scale datasets

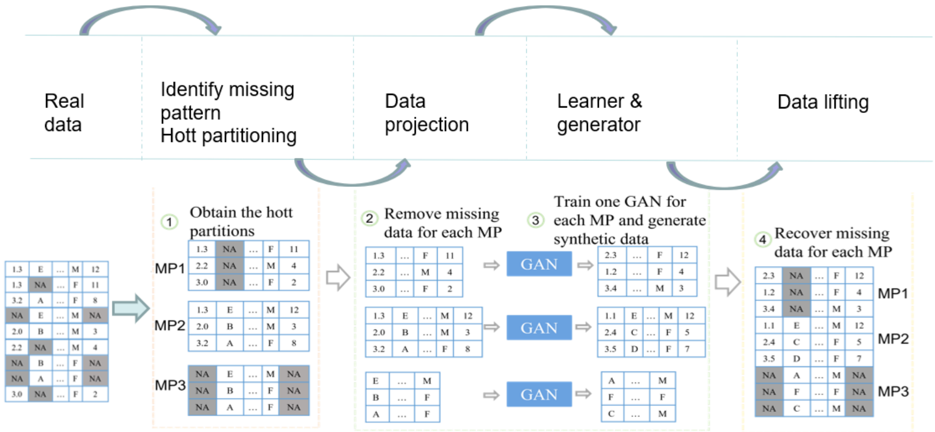
RUTGERS–NEW BRUNSWICK
Artificial Intelligence
and Data Science Collaboratory

- What can you learn from data?
- Lots and lots!
    - Identifying personal genomes by surname inference [GMGHE'13]
    - Identifying participation in complex DNA mixture [H et al.'08]
    - EEG and MEG data can leak financial and identity related information [IHE'18]
    - Modeling 3D facial shape from DNA [C et al.'14]
    - Identifying participation from a summary statistic [DSSUV'15]
    - Identifying participants in the Personal Genome Project by Name [SAW'13]
    - Identifying users in Netflix prize data [NS'08]
    - Identifying "complete patient's info." from aggregate and anonymized queries [VSJO'13]
- The list goes on…

- What can you do about this?
    - Access control
    - Differential privacy
    - Secure Multiparty Computation
    - Synthetic Data
    - …

Recent work: Generate synthetic datasets ***conditioned on the missing patterns*** that preserves both data and missing data distribution

Approach 1: HottGAN → P $(Xo|M)$P (M)


Real data
Identify missing pattern Hott partitioning
Data projection
Learner & generator
Data lifting
1 Obtain the hott partitions
MP1
MP2
MP3
2 Remove missing data for each MP
GAN
GAN
GAN
3 Train one GAN for each MP and generate synthetic data
4 Recover missing data for each MP
MP1
MP2
MP3

1.3	E	…	M	12
1.3	NA	…	F	11
3.2	A	…	F	8
NA	E	…	M	NA
2.0	B	…	M	3
2.2	NA	…	M	4
NA	B	…	F	NA
NA	A	…	F	NA
3.0	NA	…	F	2

MP1	1.3	NA	…	F	11
	2.2	NA	…	M	4
	3.0	NA	…	F	2
MP2	1.3	E	…	M	12
	2.0	B	…	M	3
	3.2	A	…	F	8
MP3	NA	E	…	M	NA
	NA	B	…	F	NA
	NA	A	…	F	NA

1.3	…	F	11
2.2	…	M	4
3.0	…	F	2

1.3	E	…	M	12
2.0	B	…	M	3
3.2	A	…	F	8

E	…	M
B	…	F
A	…	F

2.3	…	F	12
1.2	…	F	4
3.4	…	M	3

1.1	E	…	M	12
2.4	C	…	F	5
3.5	D	…	F	7

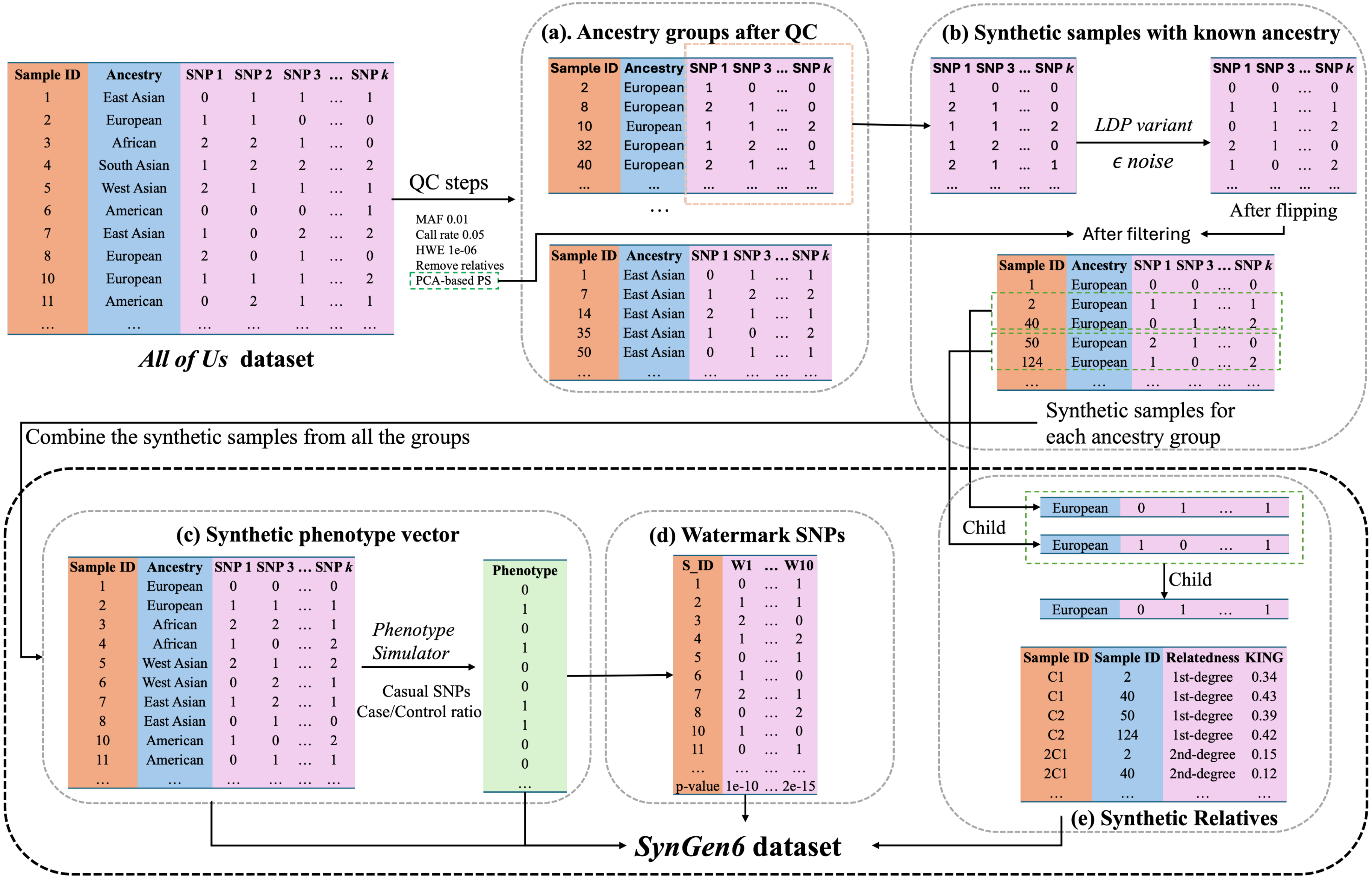
A	…	M
F	…	F
C	…	M

2.3	NA	…	F	12	
1.2	NA	…	F	4	MP1
3.4	NA	…	M	3	
1.1	E	…	M	12	
2.4	C	…	F	5	MP2
3.5	D	…	F	7	
NA	A	…	M	NA	
NA	F	…	F	NA	MP3
NA	C	…	M	NA	

More details: X. Wang, H. Asif, and J. Vaidya. 2023. Preserving Missing Data Distribution in Synthetic Data. WWW '23 https://doi.org/10.1145/3543507.3583297

Ensuring security and privacy of large-scale datasets

Genomic Data: Deep generative models (DNADiffusion, CTGAN, …)



Sample ID	Ancestry	SNP 1	SNP 2	SNP 3	…	SNP *k*
1	East Asian	0	1	1	…	1
2	European	1	1	0	…	0
3	African	2	2	1	…	0
4	South Asian	1	2	2	…	2
5	West Asian	2	1	1	…	1
6	American	0	0	0	…	1
7	East Asian	1	0	2	…	2
8	European	2	0	1	…	0
10	European	1	1	1	…	2
11	American	0	2	1	…	1
…	…	…	…	…	…	…

***All of Us* dataset**

QC steps

MAF 0.01
Call rate 0.05
HWE 1e-06
Remove relatives
PCA-based PS

(a). Ancestry groups after QC

Sample ID	Ancestry	SNP 1	SNP 3	...	SNP *k*
2	European	1	0	...	0
8	European	2	1	...	0
10	European	1	1	...	2
32	European	1	2	...	0
40	European	2	1	...	1
...	...	...	...	...	...

…

Sample ID	Ancestry	SNP 1	SNP 3	…	SNP *k*
1	East Asian	0	1	…	1
7	East Asian	1	2	…	2
14	East Asian	2	1	…	1
35	East Asian	1	0	…	2
50	East Asian	0	1	…	1
…	…	…	…	…	…

(b) Synthetic samples with known ancestry

SNP 1	SNP 3	...	SNP *k*
1	0	...	0
2	1	...	0
1	1	...	2
1	2	...	0
2	1	...	1
...	...	...	...

LDP variant

ϵ noise

SNP 1	SNP 3	...	SNP *k*
0	0	…	0
1	1	…	1
0	1	…	2
2	1	…	0
1	0	…	2
...	...	...	...

After flipping

After filtering

Sample ID	Ancestry	SNP 1	SNP 3	…	SNP *k*
1	European	0	0	…	0
2	European	1	1	…	1
40	European	0	1	…	2
50	European	2	1	…	0
124	European	1	0	…	2
…	…	…	…	…	…

Synthetic samples for each ancestry group

Combine the synthetic samples from all the groups

(c) Synthetic phenotype vector

Sample ID	Ancestry	SNP 1	SNP 3	…	SNP *k*
1	European	0	0	…	0
2	European	1	1	…	1
3	African	2	2	…	1
4	African	1	0	…	2
5	West Asian	2	1	…	2
6	West Asian	0	2	…	1
7	East Asian	1	2	…	1
8	East Asian	0	1	…	0
10	American	1	0	…	2
11	American	0	1	…	1
…	…	…	…	…	…

Phenotype Simulator

Casual SNPs
Case/Control ratio

Phenotype
0
1
0
1
0
0
1
1
0
0
…

(d) Watermark SNPs

S_ID	W1	…	W10
1	0	…	1
2	1	…	1
3	2	…	0
4	1	…	2
5	0	…	1
6	1	…	0
7	2	…	1
8	0	…	2
10	1	…	0
11	0	…	1
…	…	…	…
p-value	1e-10	…	2e-15

Child

European	0	1	…	1
European	1	0	…	1

Child

European	0	1	…	1

Sample ID	Sample ID	Relatedness	KING
C1	2	1st-degree	0.34
C1	40	1st-degree	0.43
C2	50	1st-degree	0.39
C2	124	1st-degree	0.42
2C1	2	2nd-degree	0.15
2C1	40	2nd-degree	0.12
…	…	…	…

(e) Synthetic Relatives

***SynGen6* dataset**

More details: Wang, X., Min, S. and Vaidya, J., 2024. Descriptor: Synthetic Genomic Dataset with Diverse Ancestry (SynGen6). *IEEE Data Descriptions*.