Phenotypic Data and Individual Privacy

Arif Harmanci

What is a linking attack? Case of Netflix Prize

Movie ratings database

Anonymized Netflix Prize Training Dataset made available to contestants

100 million ratings
500,000 users

200 movie ratings/user
5,000 users/movie rating

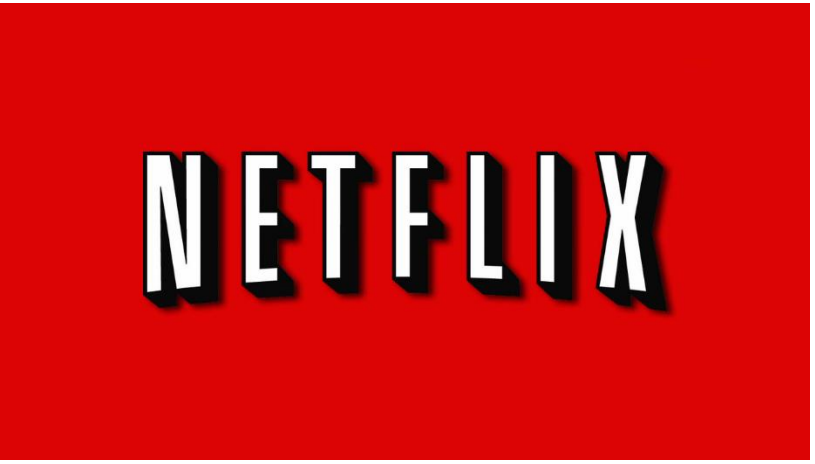
User (ID)	Movie (ID)	Date of Rating	Rating [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

What is a linking attack? Case of Netflix Prize

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

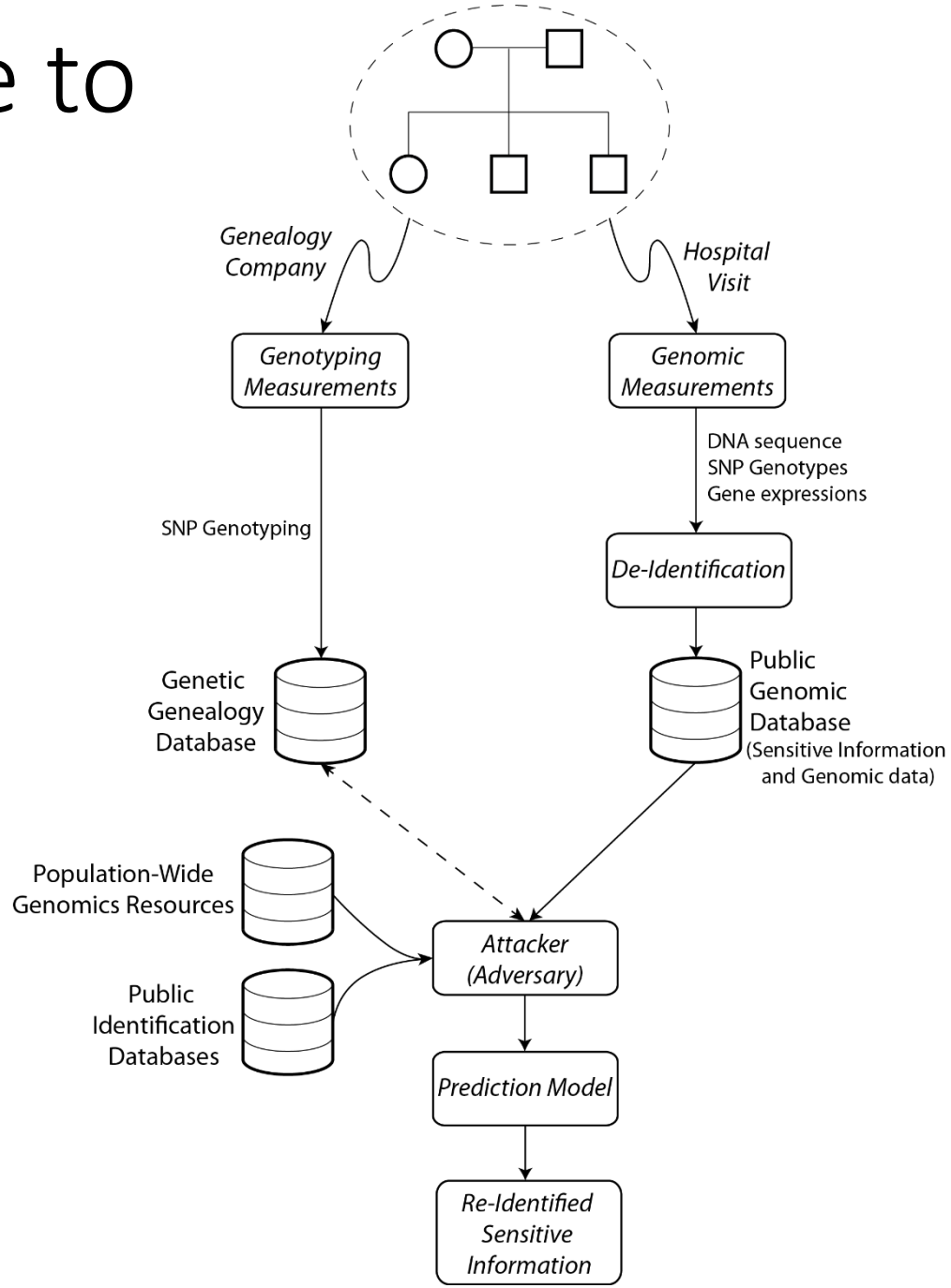
Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

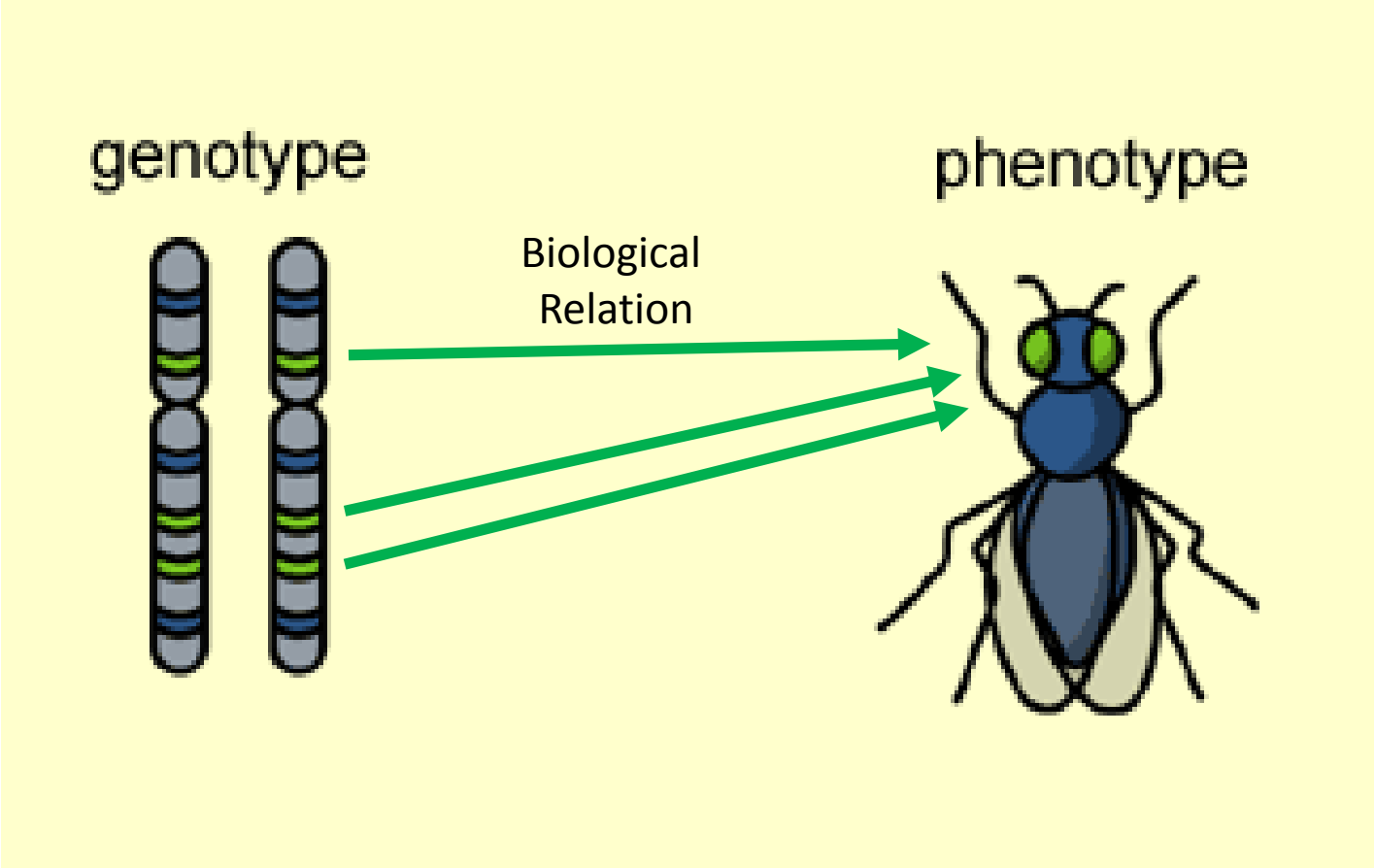


1. Very large datasets
2. A lot of users have a Netflix and an IMDb account
3. A user rates similar scores to a movie in Netflix and IMDb
4. A user rates a particular movie around the same date in Netflix and IMDb

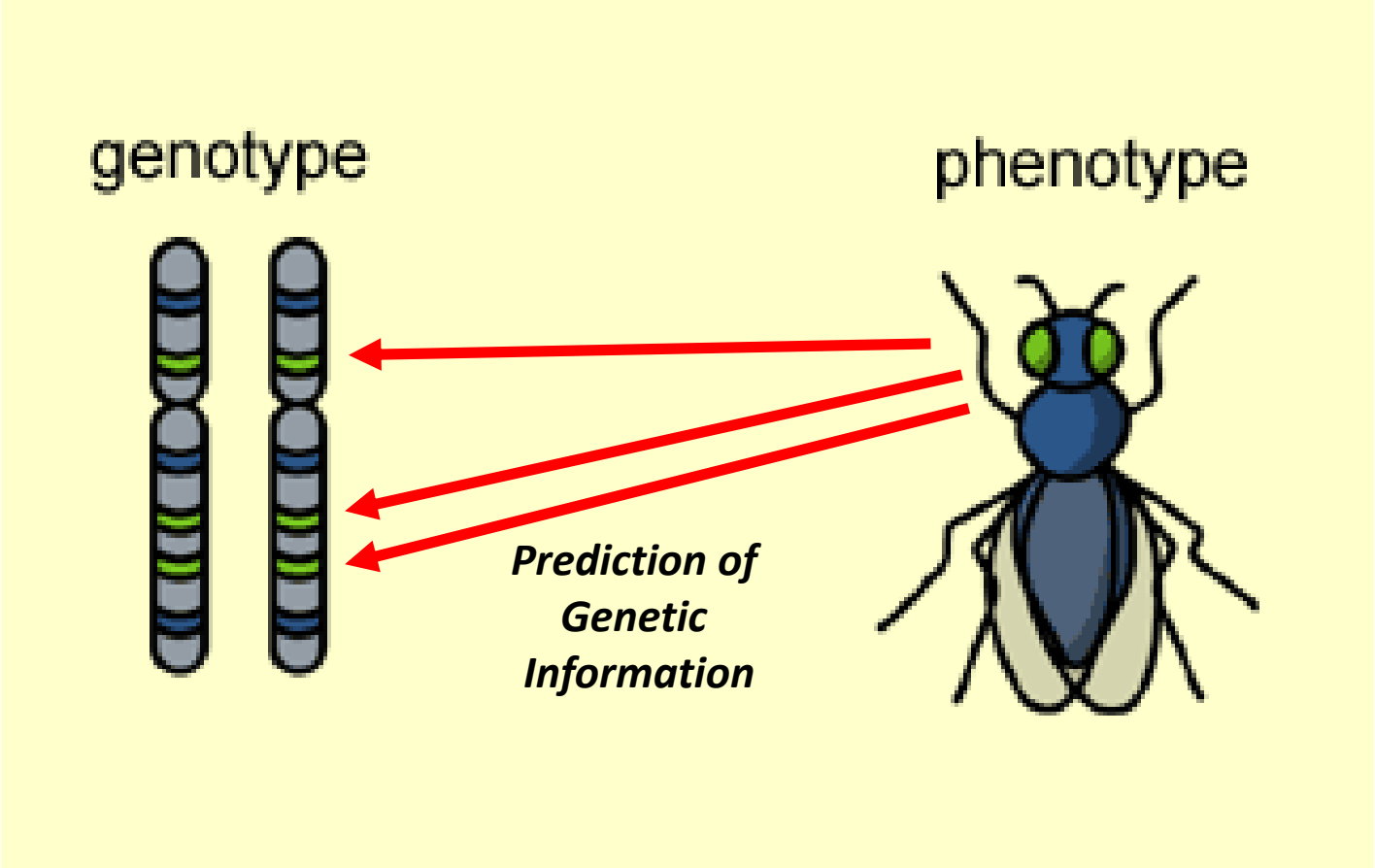
How does this relate to genomic privacy?



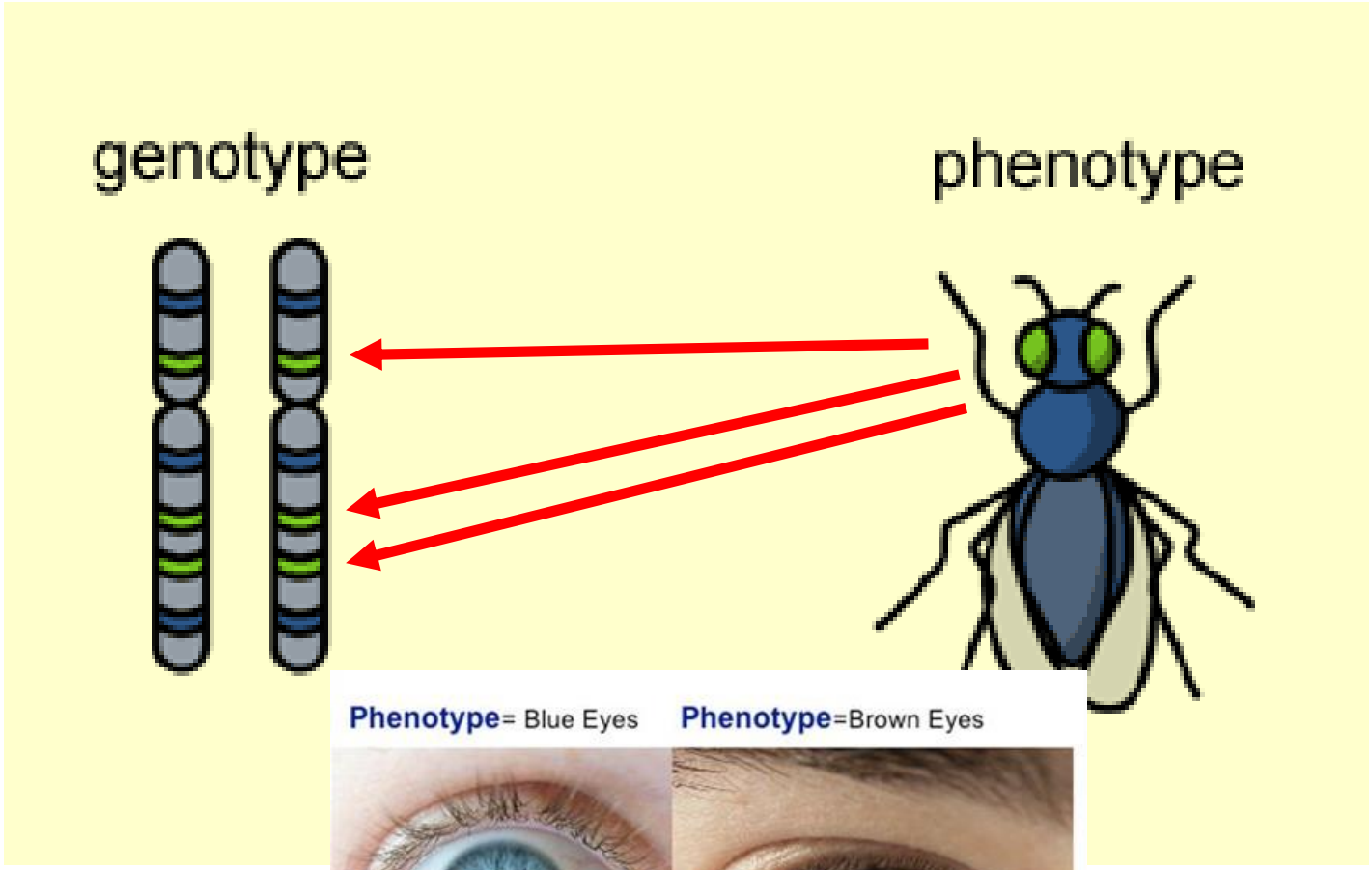
How does this relate to genomic privacy?



How does this relate to genomic privacy?



How does this relate to genomic privacy?



Linking Attacks: Phenotypic Data

Phenotype Dataset (Public)

Phenotype ID	HIV Status	Phenotype 1	Phenotype 2	⋯	Phenotype q
PID-*1*	HIV+	0.1	-2.7	⋯	90.3
PID-*2*	HIV-	0.5	8.6	⋯	63.5
⋮	⋮	⋮	⋮	⋮	⋮
PID-n	HIV-	-0.2	5.4	⋯	50.3

Phenotype-Genotype Correlation Dataset

Phenotype 1 ⟷ Genotype 1
Phenotype 2 ⟷ Genotype 2
⋯
Phenotype q ⟷ Genotype q

Genotype Dataset (Stolen/Hacked/Queried)

Genotype ID	Genotype 1	Genotype 2	⋯	Genotype q
GID-*1*	0	1	⋯	1
GID-*2*	2	1	⋯	0
⋮	⋮	⋮	⋮	⋮
GID-m	1	2	⋯	1

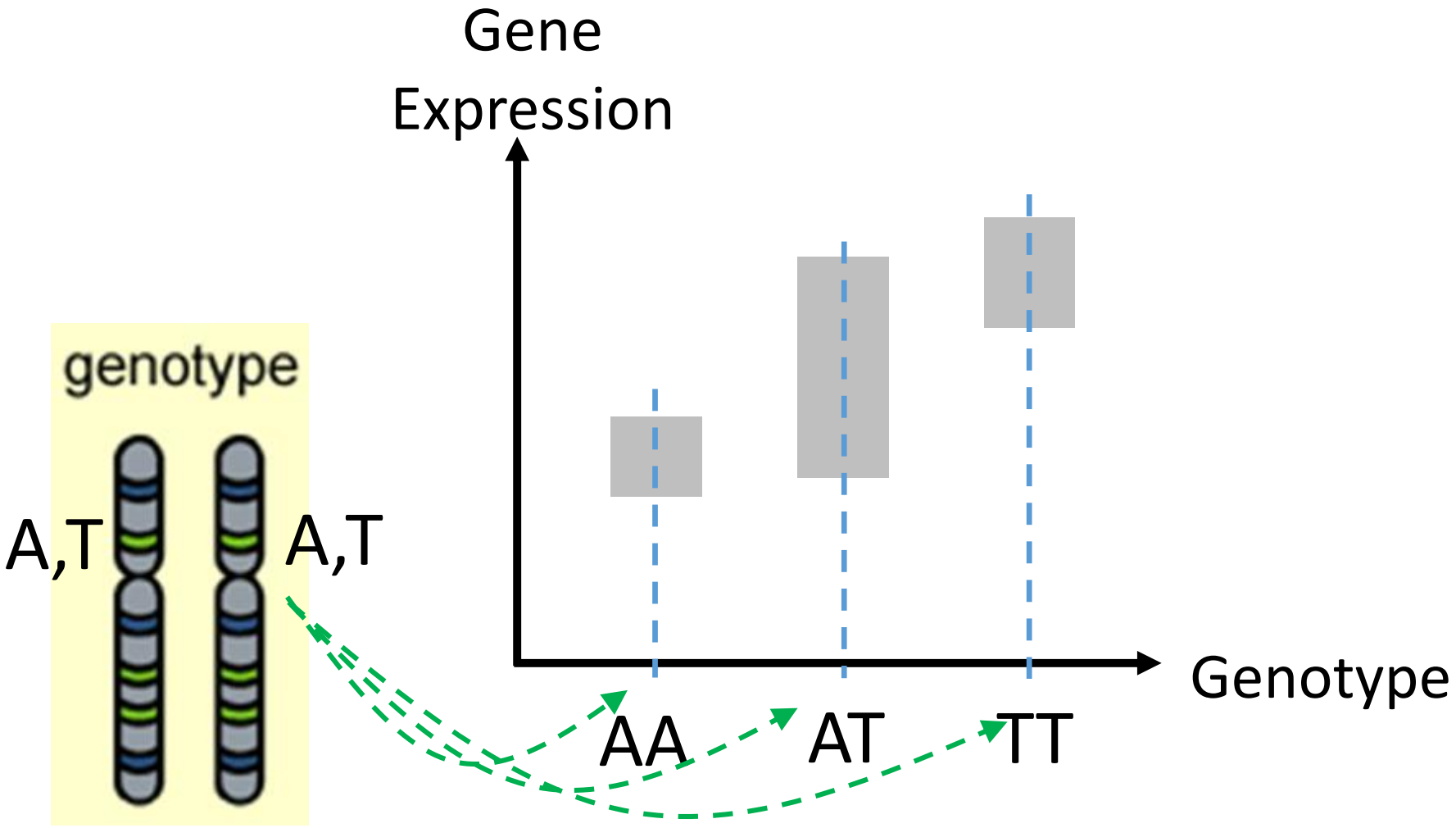
Genotype Prediction

Phenotype ID	HIV Status	Predicted Genotypes: Genotype 1	Genotype 2	⋯	Genotype q
PID-*1*	HIV+	1	0	⋯	2
PID-*2*	HIV-	2	2	⋯	1
⋮	⋮	⋮	⋮	⋮	⋮
PID-n	HIV-	0	1	⋯	1

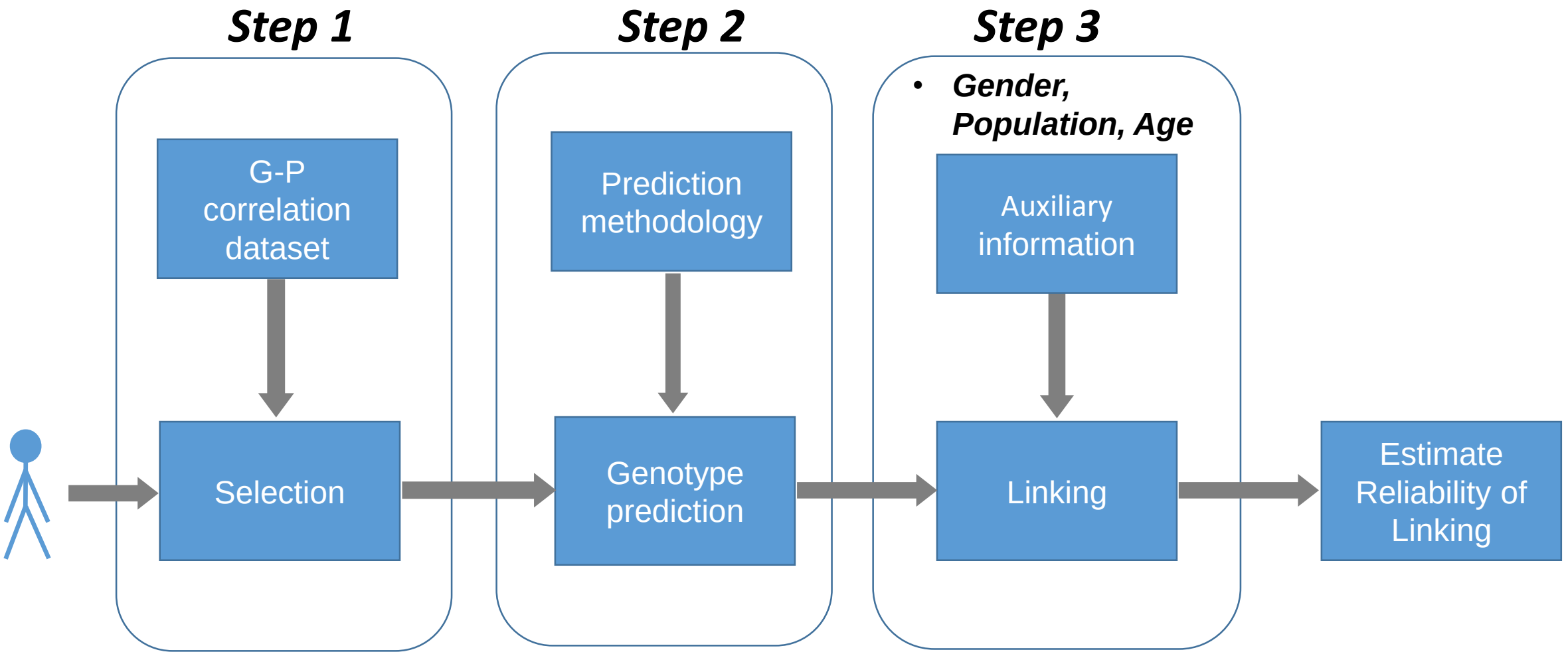
Genotype Comparison and Matching

Genotype ID	Phenotype ID	HIV Status	Predicted/Matched Genotypes: Genotype 1	Genotype 2	⋯	Genotype q
GID-*1*	PID-*8*	HIV+	0/0	1/1	⋯	1/1
GID-*2*	PID-*3*	HIV-	2/2	1/1	⋯	0/0
GID-*3*			1/0	1/0	⋯	0/2
GID-*4*	PID-*1*	HIV+	1/1	0/0	⋯	2/2
GID-*5*			0/1	1/1	⋯	2/1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Gene expression-genotype linking: expression Quantitative Trait Loci



3-Steps of a Linking Attack



Take Home Messages

- Your genetic information (4 letters) can be guessed somewhat accurately from your physical characteristics!
  - Height, eye color, hair color, weight…
- There are 3 steps in a linking attack!
- Phenotype and genotype datasets can be reliably linked to reveal sensitive information!
- Auxiliary information is very important!
  - Your gender and where you are from can be used against you in a linking attack
- Relatives are also at risk of privacy breach!