Security Risks to Third-Party Genetic Genealogy Services

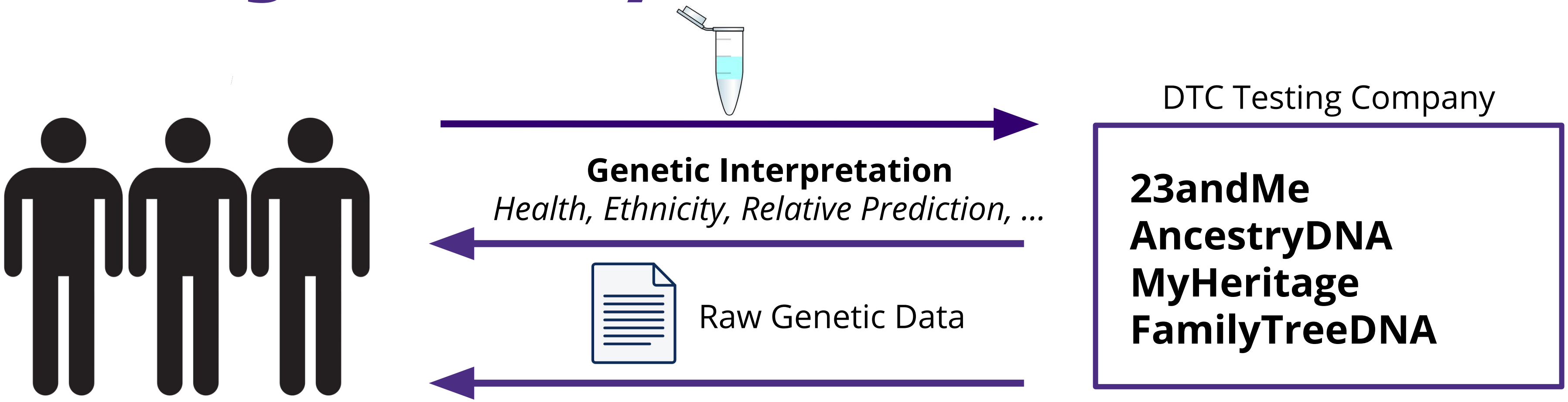
Peter Ney, Luis Ceze, Tadayoshi Kohno

BE BOUNDLESS

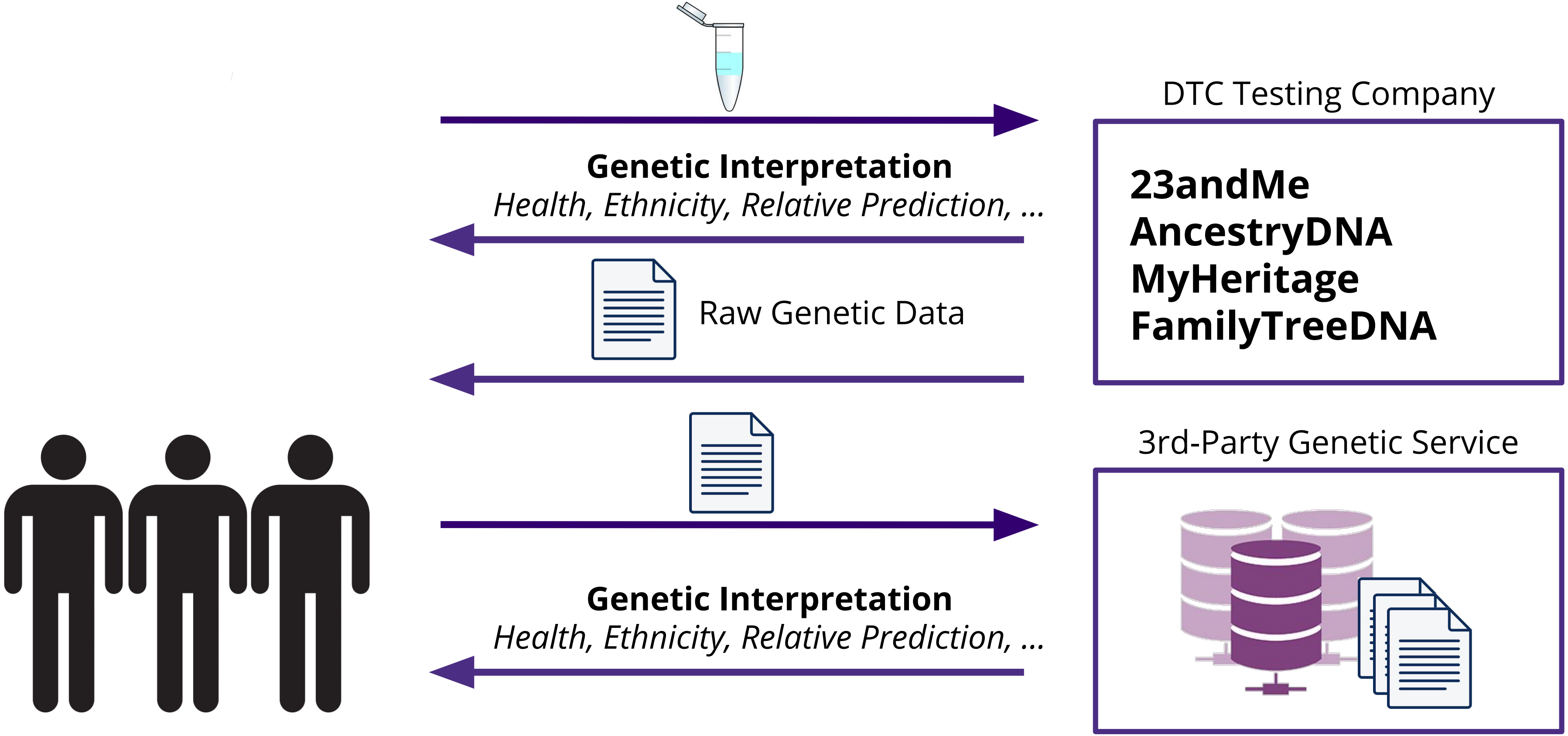
SECURITY & PRIVACY
RESEARCH LAB
UNIVERSITY OF WASHINGTON

MISL

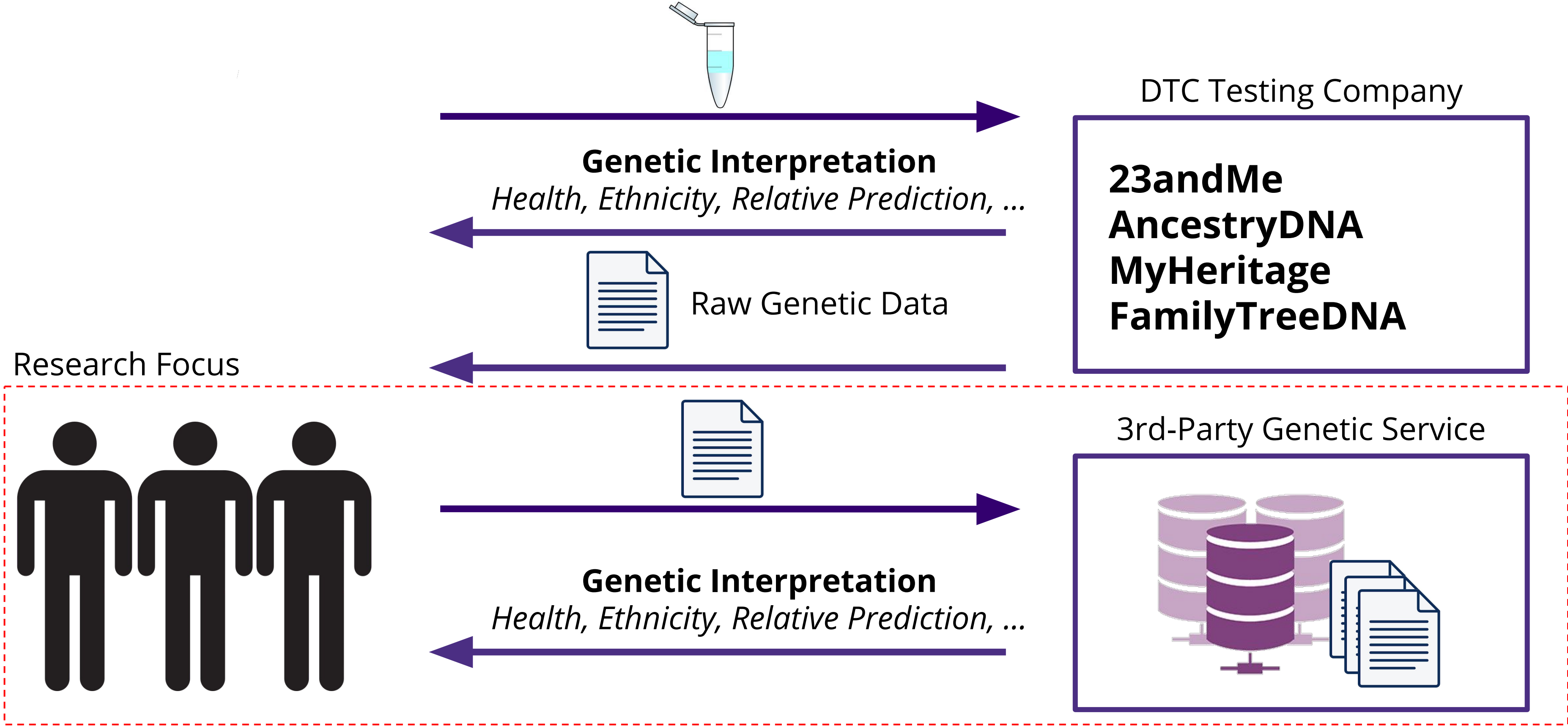
Direct-to-Consumer (DTC) Genetic Testing and Analysis



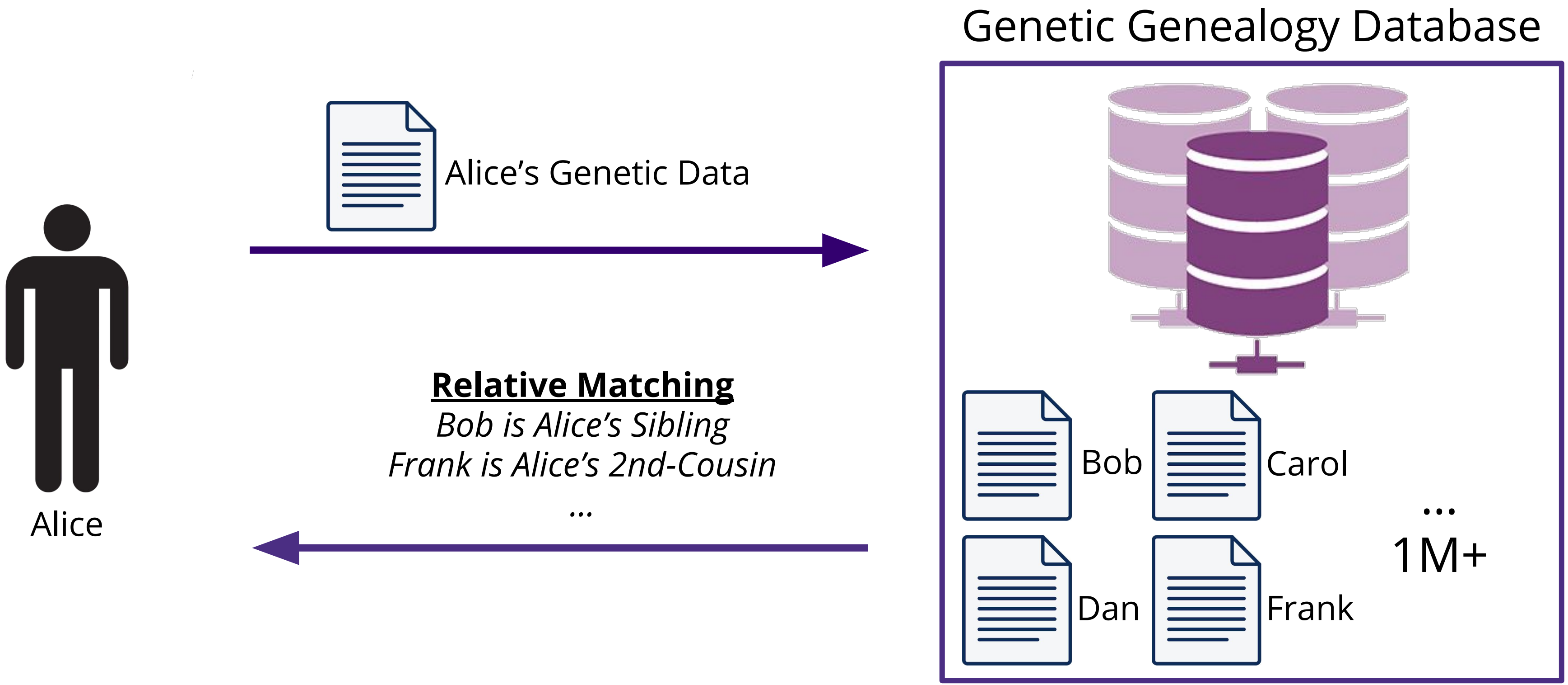
Direct-to-Consumer (DTC) Genetic Testing and Analysis



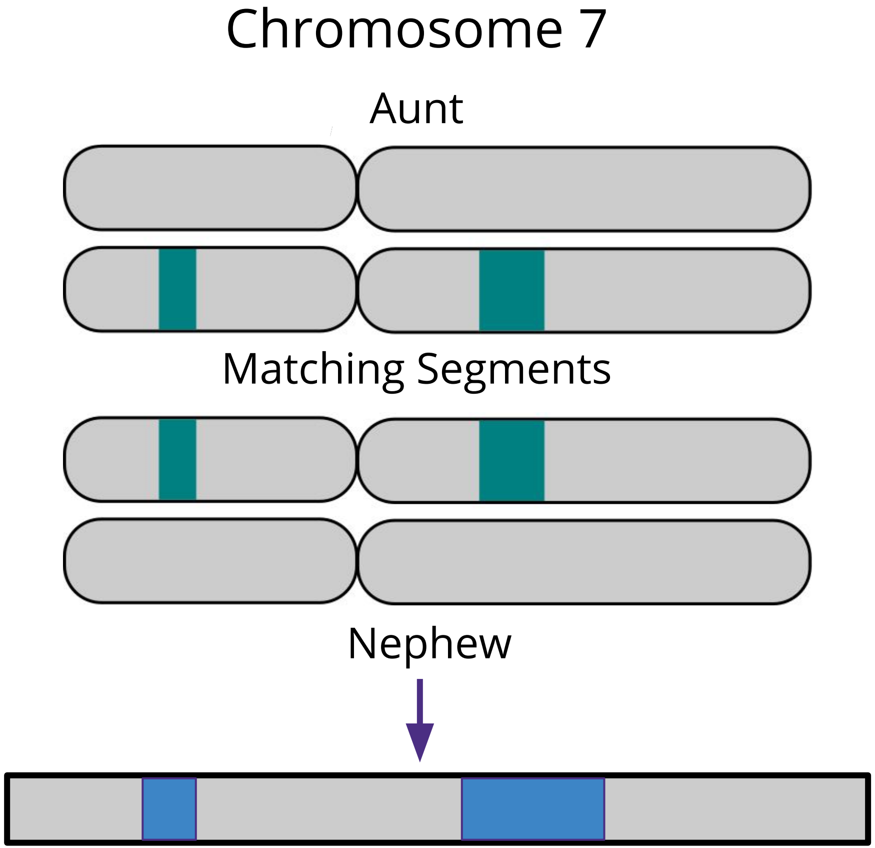
Direct-to-Consumer (DTC) Genetic Testing and Analysis



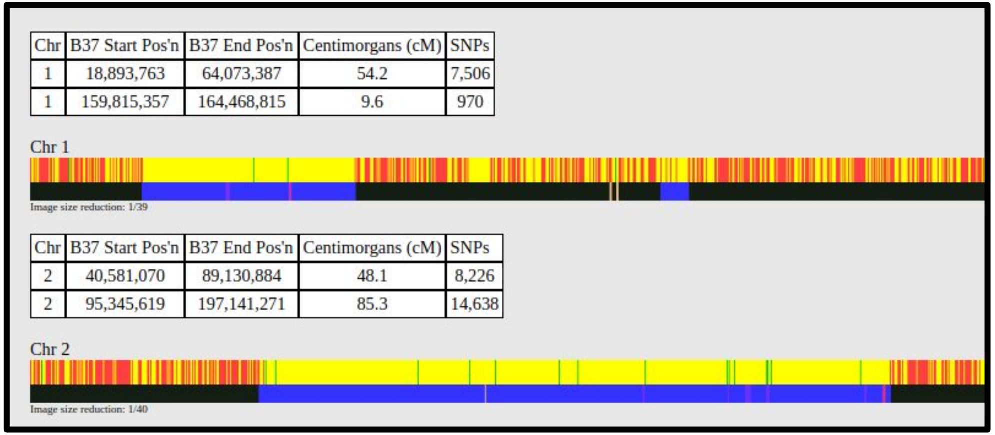
Third-Party Genetic Genealogy Services



Relative Matching Algorithms



- Long shared segments of DNA are indicative of recent shared ancestry
- More and longer shared segments means a closer relationship
- Relative matching algorithms try to identify these shared segments between users



Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Chr 1

Image size reduction: 1/39

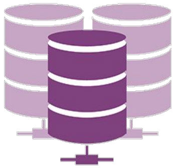
Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
2	40,581,070	89,130,884	48.1	8,226
2	95,345,619	197,141,271	85.3	14,638

Chr 2

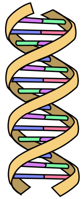
Image size reduction: 1/40

Prior Attacks Against Genetic Genealogy Services: Identity Inference

Anonymous DNA sample or genetic data



Research Dataset

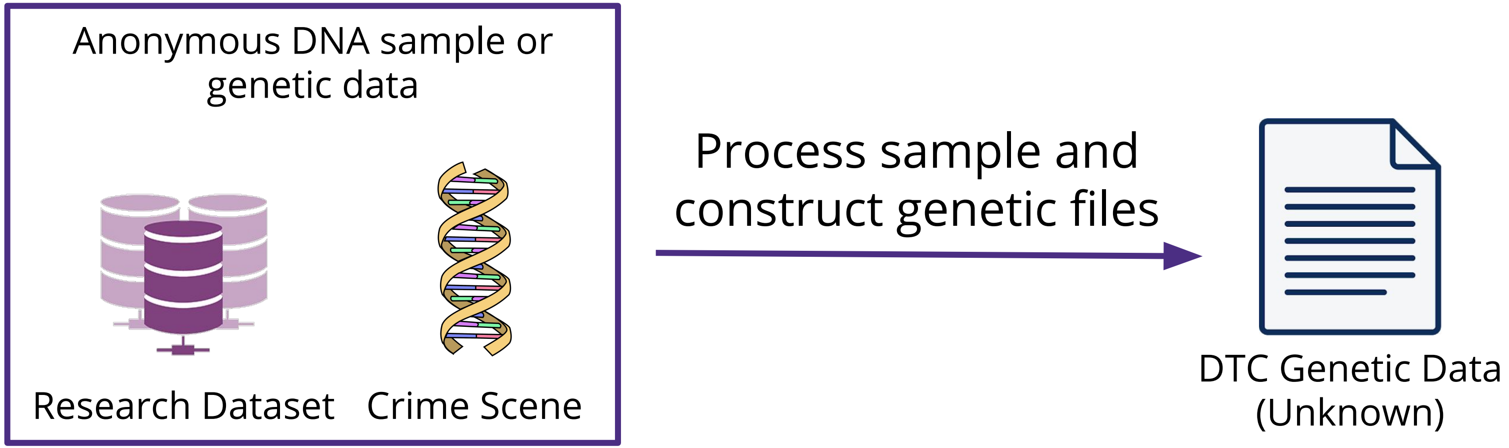


Crime Scene

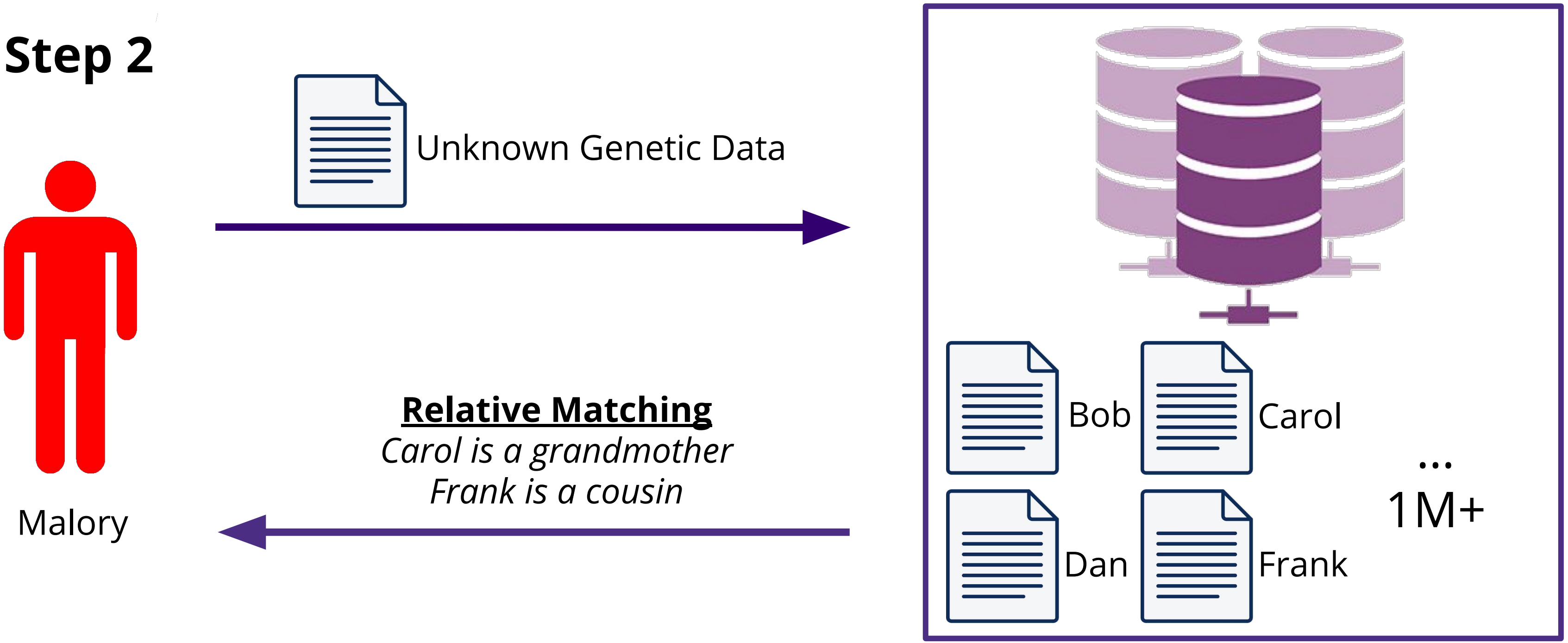
Goal: identify the source (person) of an anonymous DNA sample or genetic data

Prior Attacks Against Genetic Genealogy Services: Identity Inference

Step 1

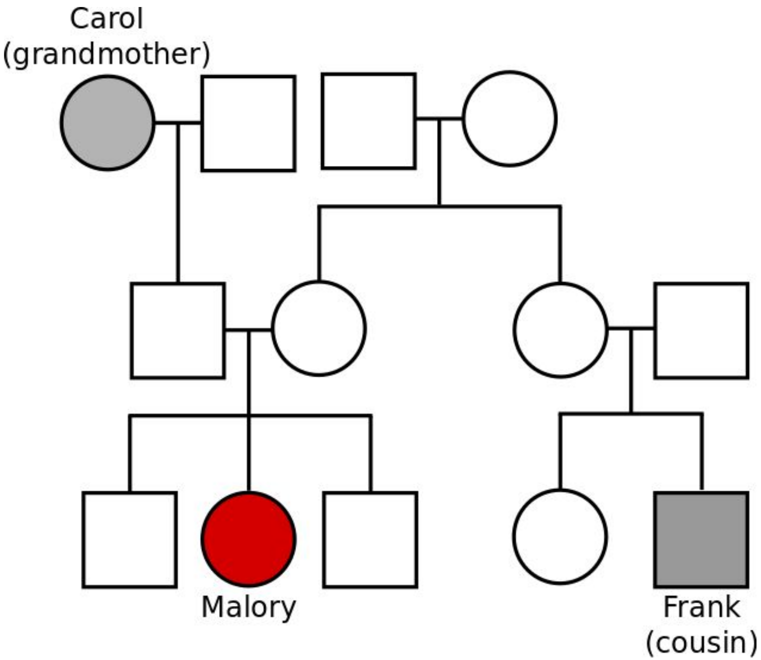


Prior Attacks Against Genetic Genealogy Services: Identity Inference



Prior Attacks Against Genetic Genealogy Services: Identity Inference

Step 3: Combine the relatives with other sources of information like genealogies to identify the source of the sample or data



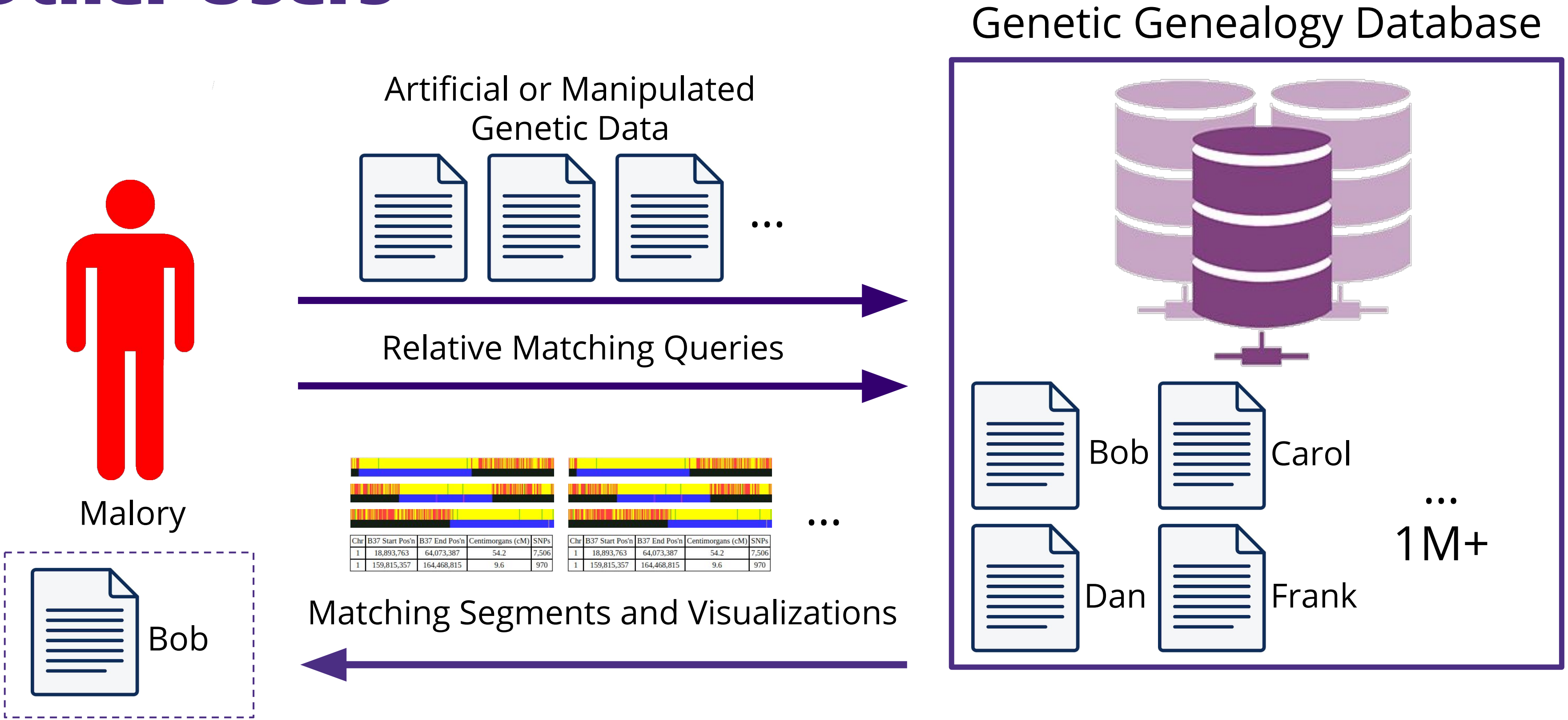
Law enforcement

- 100+ samples identified from crimes and unknown remains
- Suspected Golden State Killer

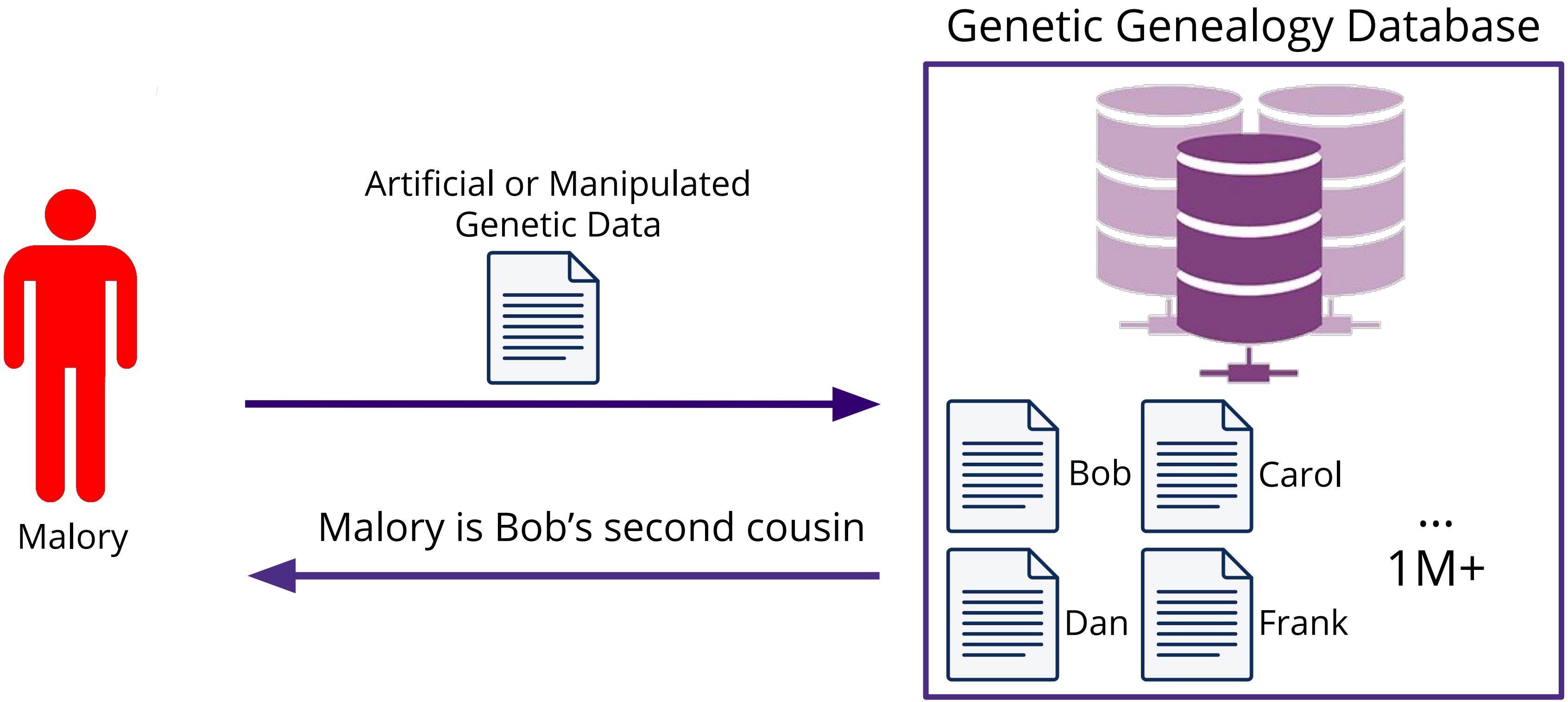
Anonymous research data

- Ex: 1000 Genomes Data (*Erlich et al. Science. 2018)*

Attack 1: Extract Genetic Markers from Other Users



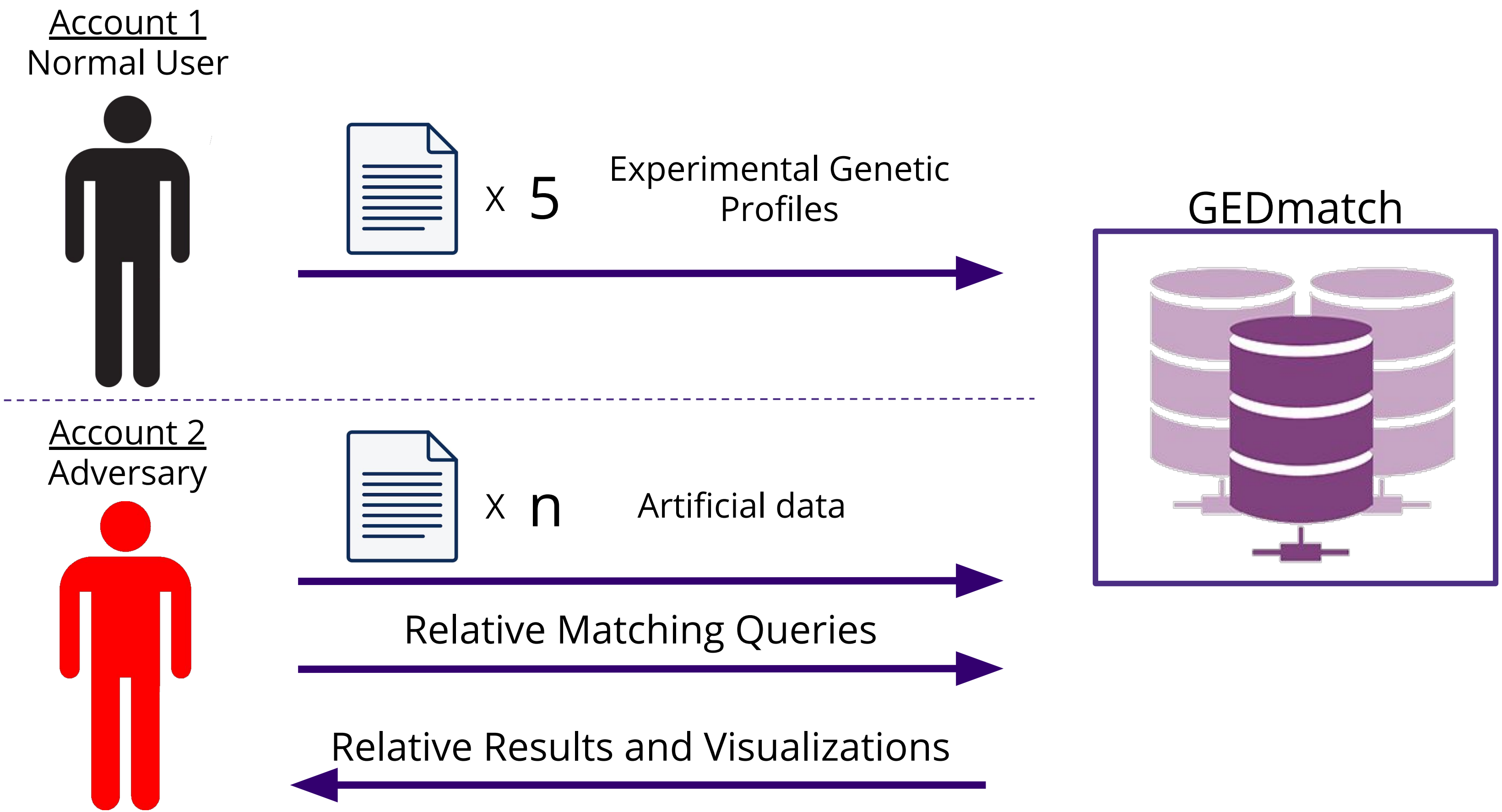
Attack 2: Forge Genetic Relationships



Case Study on GEDmatch

- GEDmatch runs the largest third-party DTC genetic genealogy service
  - Over 1.2 millions files have been uploaded
- Used extensively by law enforcement
  - Used to solve Golden State Killer case
  - Government contracting (Parabon Nanolabs)
  - Unidentified remains (DNA Doe Project)
- Identity inference attacks demonstrated on GEDmatch (*Erlich et al. Science. 2018)*
- Goal is to evaluate the feasibility of these new attacks on GEDmatch

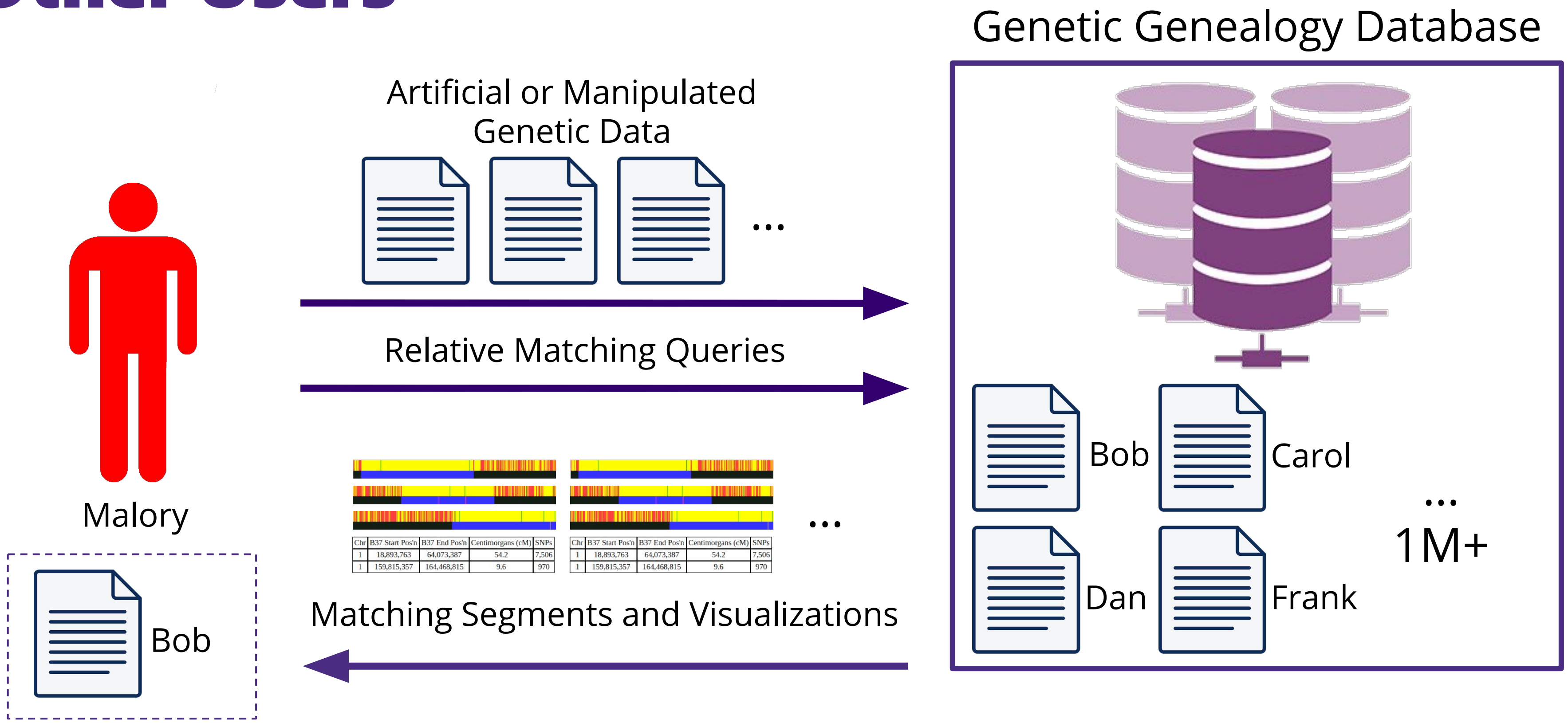
Experimental Setup



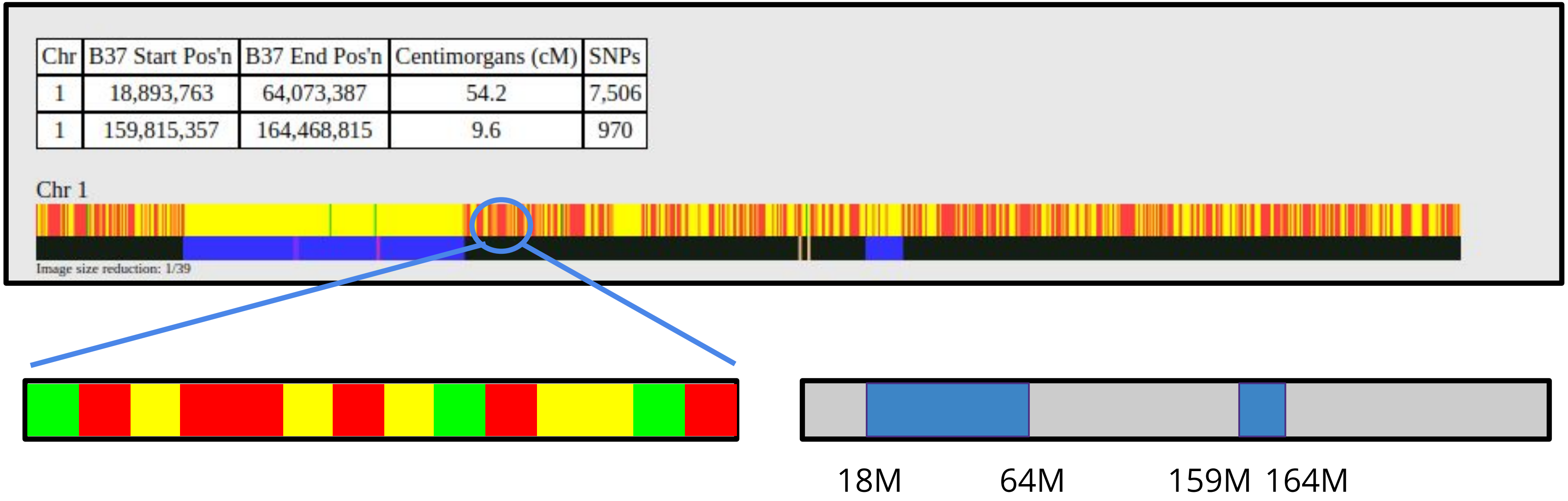
Ethics of Data Uploads and Queries

- Uploaded all data to a sandboxed “Research” setting so that the uploaded files would not interact with real GEDmatch users
- Only ran queries with and analyzed results from data that we uploaded
  - GEDmatch let’s you target relative matching queries against specific data files
- ToS allowed artificial data uploads if:
  - (1) Intended for research
  - (2) Not used to identify anyone in the database
- IRB determined that research was exempt from review because the experimental data was derived from public sources with no identifiers

Attack 1: Extract Genetic Markers from Other Users



GEDmatch Visualizations and Segments

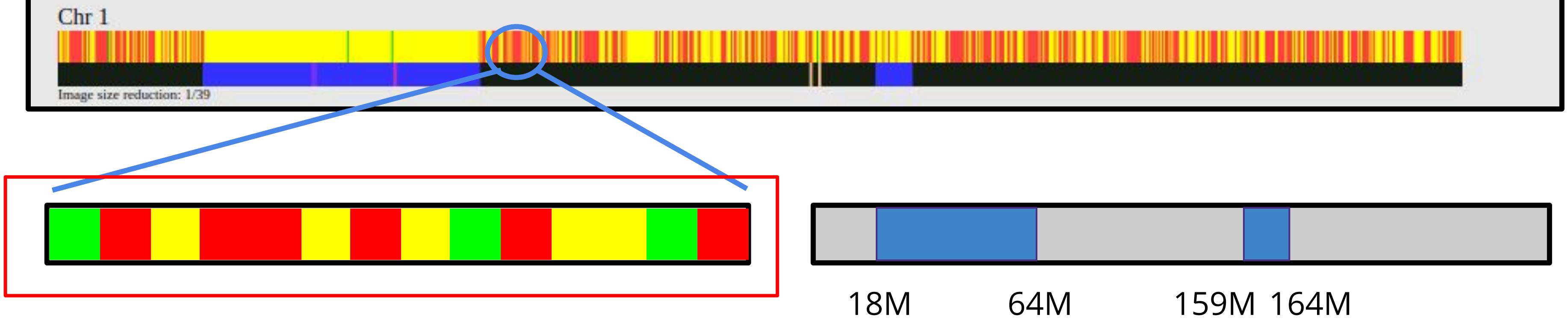


Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Both visualizations leak information about the underlying DNA markers in other genetic files.

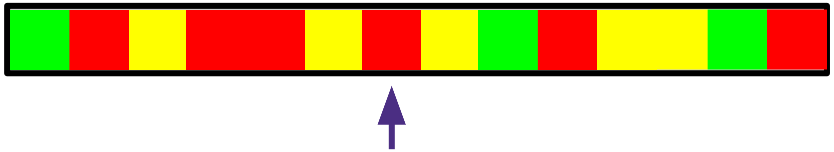
GEDmatch Visualizations and Segments

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970



Both visualizations leak information about the underlying DNA markers in other genetic files.

Genetic Extraction via Marker Visualizations



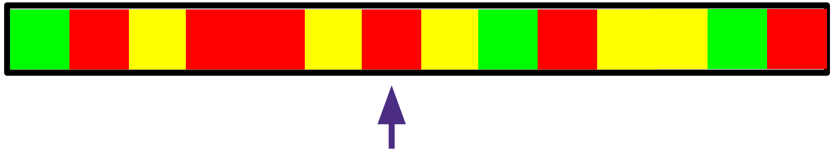
Each pixel corresponds to a single genetic marker (many are missing)

Markers same

Markers half-match

Markers different

Genetic Extraction via Marker Visualizations

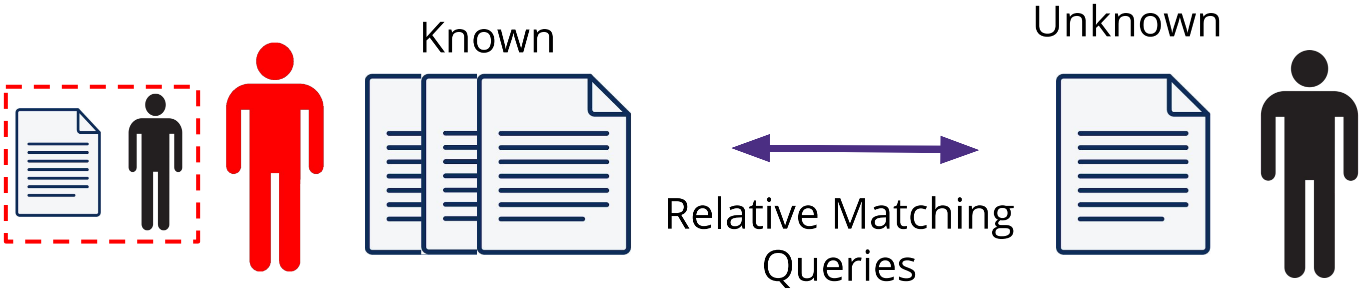


Each pixel corresponds to a single genetic marker (many are missing)

Markers same

Markers half-match

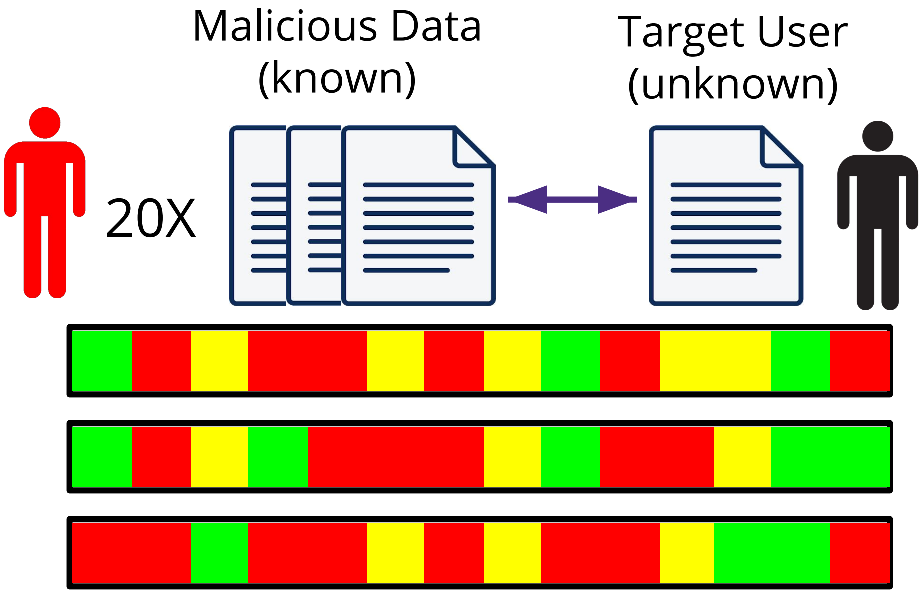
Markers different



Genetic Extraction via Marker Visualizations

Step 1

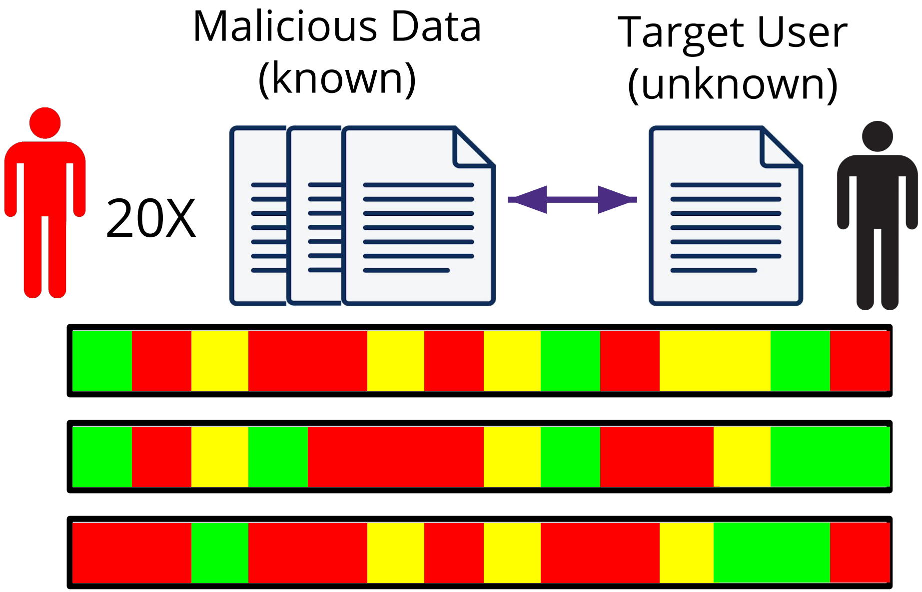
Run 20 relative matching queries against a target and gather visualizations



Genetic Extraction via Marker Visualizations

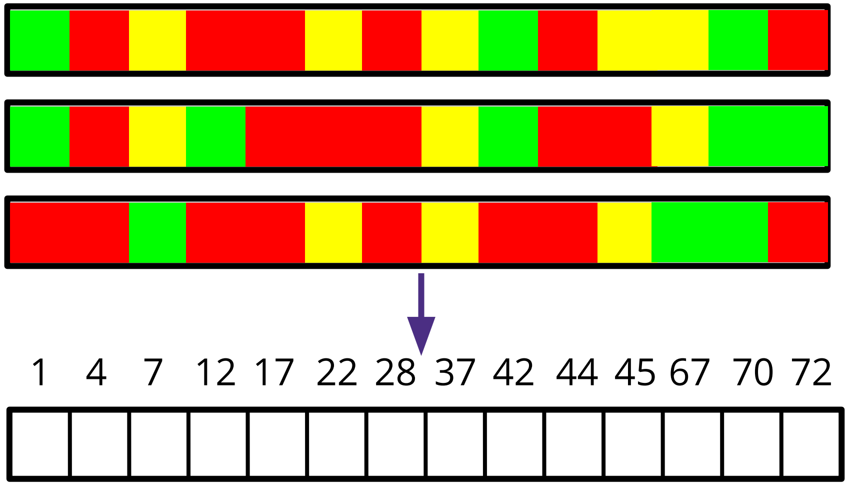
Step 1

Run 20 relative matching queries against a target and gather visualizations



Step 2

Use mastermind-like algorithm to determine which pixels correspond to specific markers. (Similar to *Goodrich. S&P. 2009.* DNA sequence extraction via DNA sequence alignment scores.)



Genetic Extraction via Marker Visualizations

Step 3

Combine known artificial genetic markers with visualizations to infer target's genetic markers

1	4	7	12	17	22	28	37	42	44	45	67	70	72
A	A	G	T	T	G	G	G	C	A	T	T	A	C
A	C	C	T	C	C	G	G	G	A	G	T	C	C

Malicious File +

1	4	7	12	17	22	28	37	42	44	45	67	70	72
A	A	G	C	T	C	G	G	C	C	T	T	A	T
A	A	G	C	C	C	T	A	G	C	G	G	C	T

Target File

Genetic Extraction via Marker Visualizations

Step 3

Combine known artificial genetic markers with visualizations to infer target's genetic markers

1	4	7	12	17	22	28	37	42	44	45	67	70	72
A	A	G	T	T	G	G	G	C	A	T	T	A	C
A	C	C	T	C	C	G	G	G	A	G	T	C	C

Malicious File +

1	4	7	12	17	22	28	37	42	44	45	67	70	72
A	A	G	C	T	C	G	G	C	C	T	T	A	T
A	A	G	C	C	C	T	A	G	C	G	G	C	T

Target File

Step 4

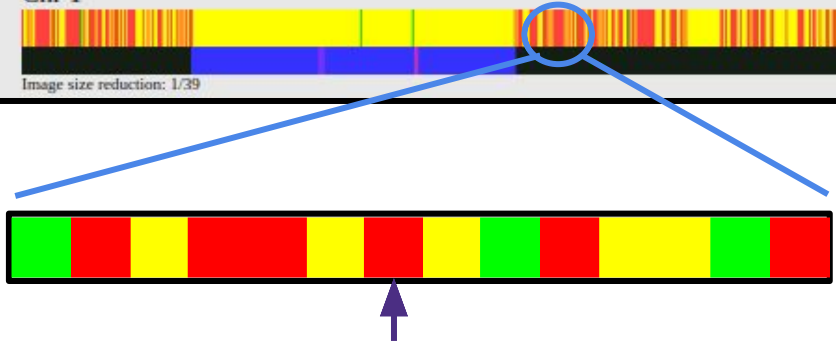
Fill in the gaps with genetic imputation (statistical technique)

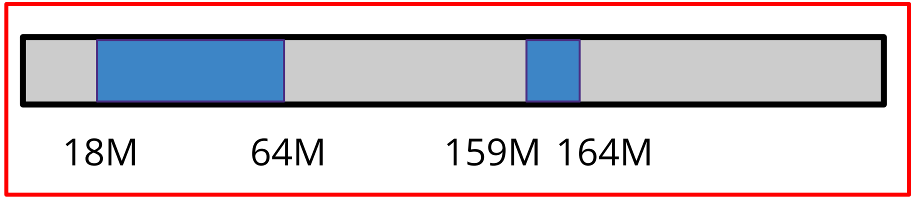
1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	A	G	C	T	C	G	G	C	C	T	T	A	T
A	A	G	C	C	C	T	A	G	C	G	G	C	T

In total we were able to extract an **average of 92% of the genetic markers with 98% accuracy** from the 5 test files.

GEDmatch Visualizations and Segments

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

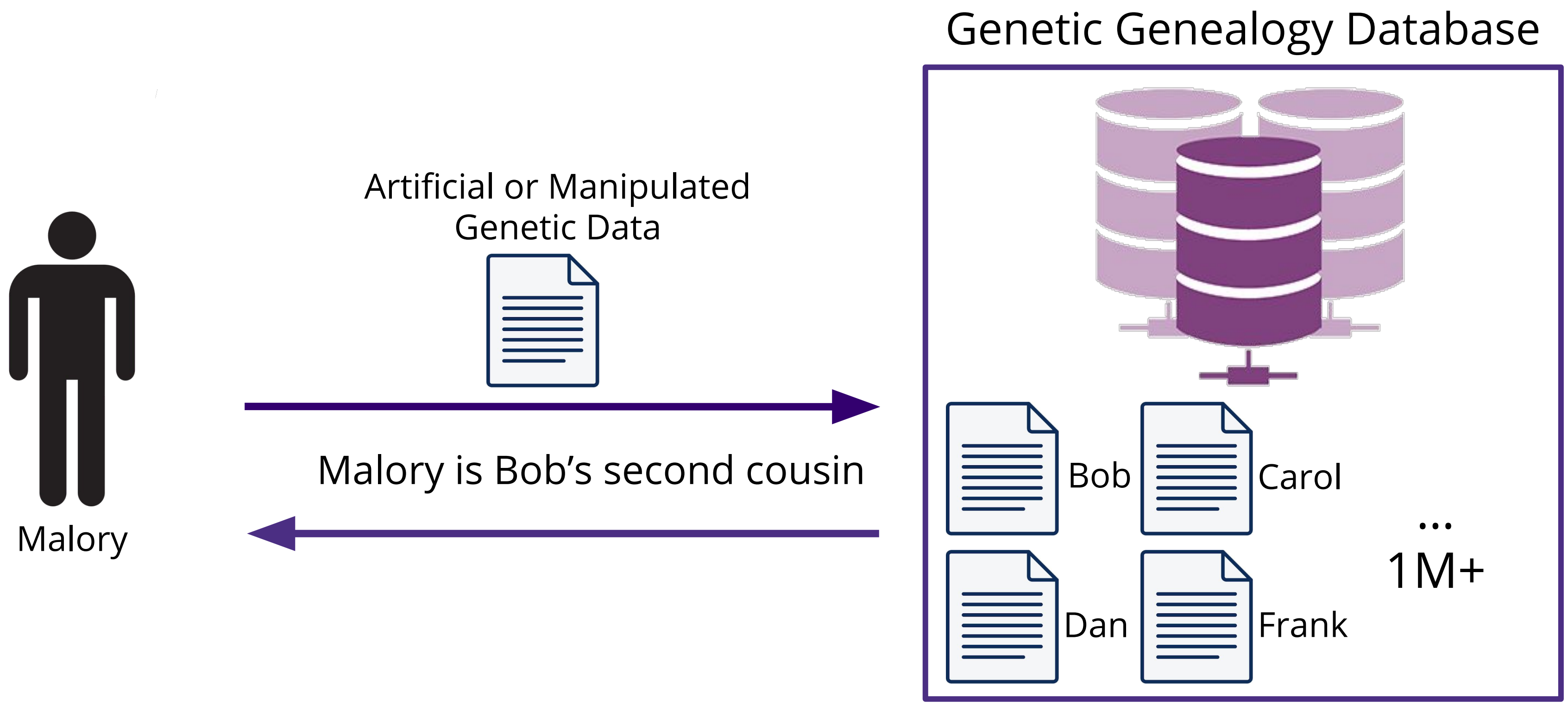




Edge and Coop. eLife. 2020. (independently discovered).

Both visualizations leak information about the underlying DNA markers in other genetic files.

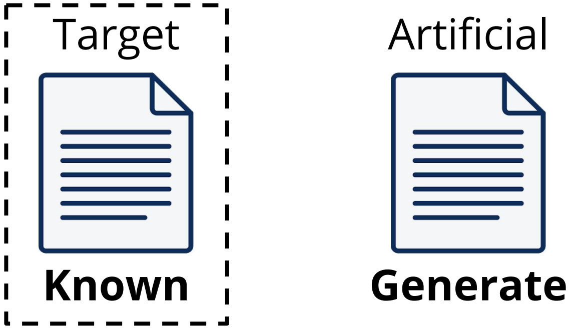
Attack 2: Forge Genetic Relationships



Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction

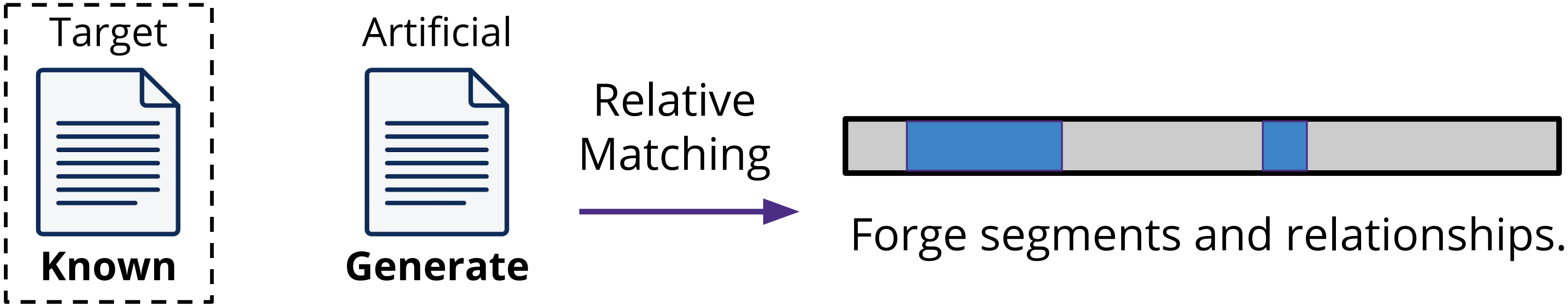
- Parent/Child: 50%
- 1st cousin: 12.5%



Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction

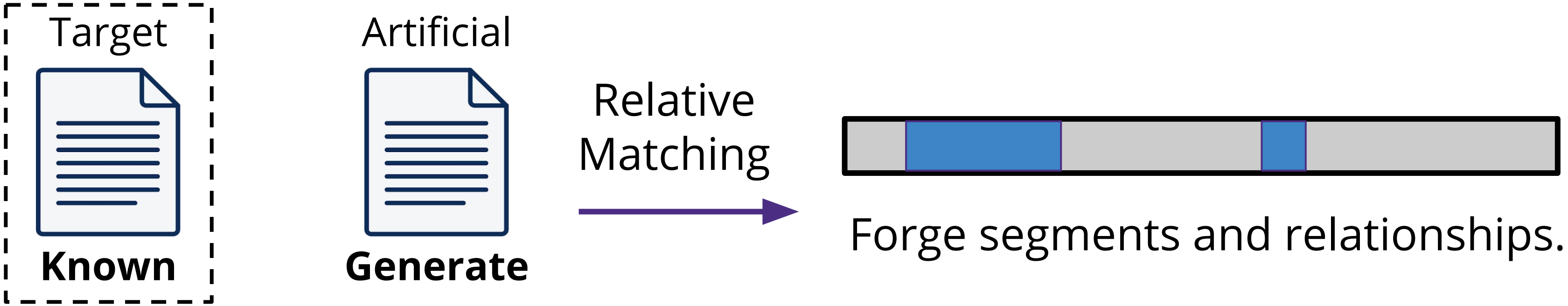
- Parent/Child: 50%
- 1st cousin: 12.5%



Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction

- Parent/Child: 50%
- 1st cousin: 12.5%

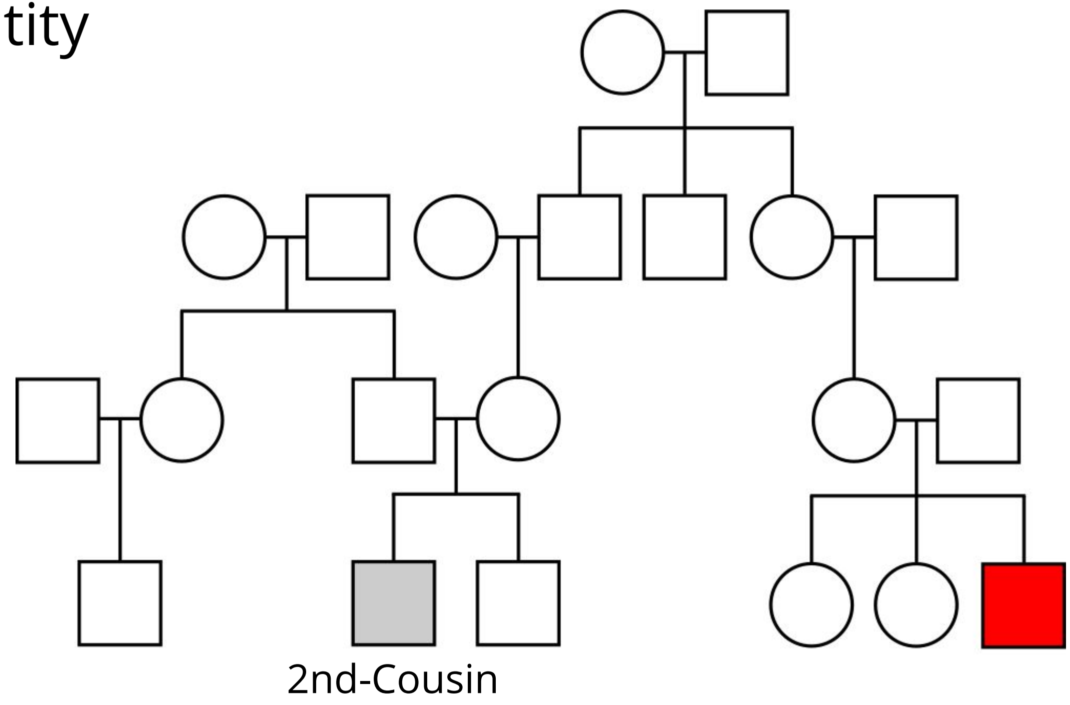


Discover target's genetic profile using:

1) Genetic extraction attacks (shown earlier). *Tested on GEDmatch.*
2) Gather DNA sample surreptitiously and sequence it.
3) Adversary wants to forge relative for themselves.

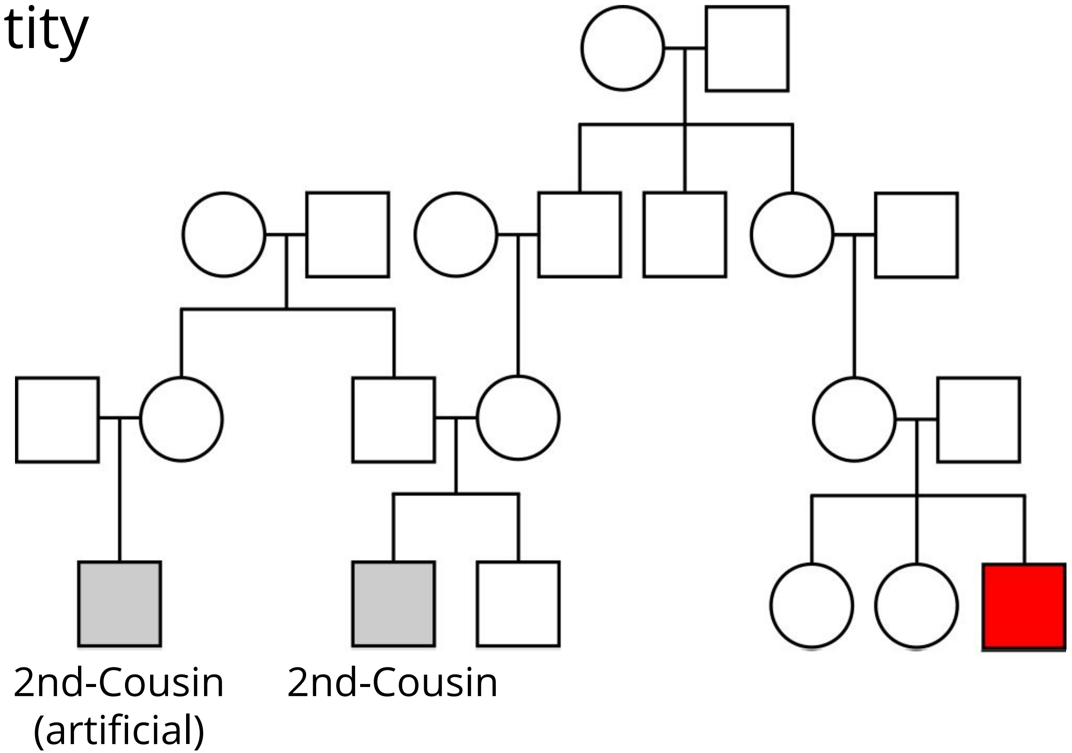
Why Make Artificial Relatives?

1) "Long lost relative." Not uncommon in genetic genealogy because of misidentified paternity.
2) Change inferred identity



Why Make Artificial Relatives?

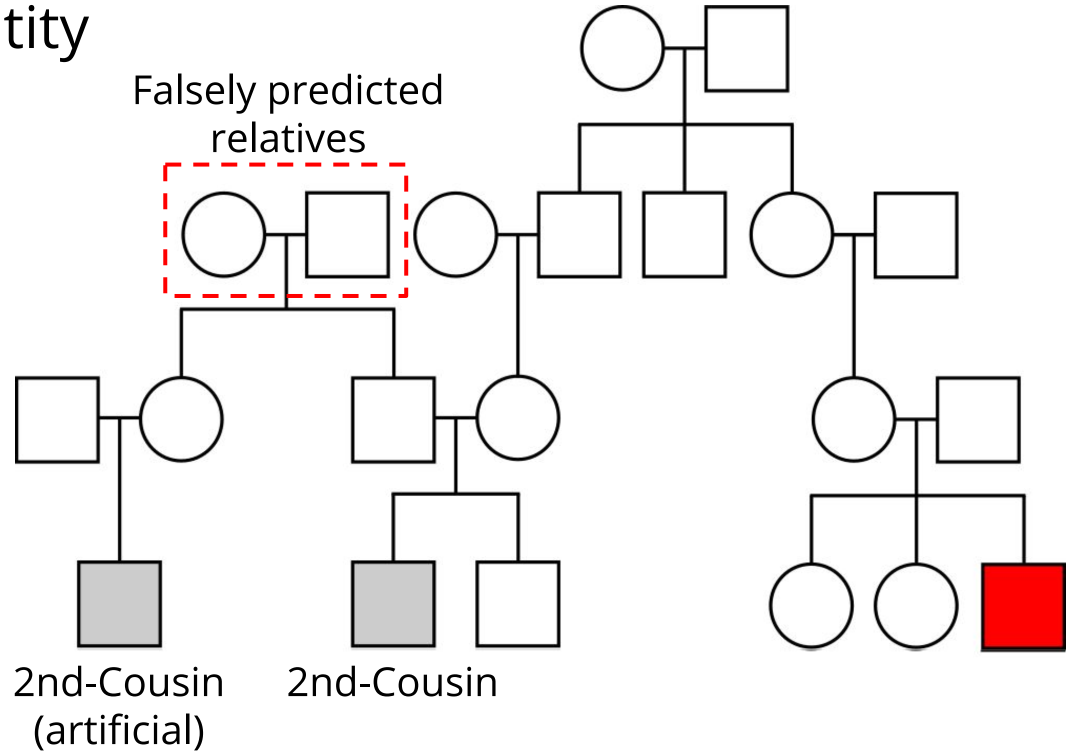
1) "Long lost relative." Not uncommon in genetic genealogy because of misidentified paternity.
2) Change inferred identity



Why Make Artificial Relatives?

1) "Long lost relative." Not uncommon in genetic genealogy because of misidentified paternity.
2) Change inferred identity

Search occurs on wrong branch of tree



Open question is how this could affect import inferences, like law enforcement, which is currently an expert driven and manual process

Responsible Disclosure

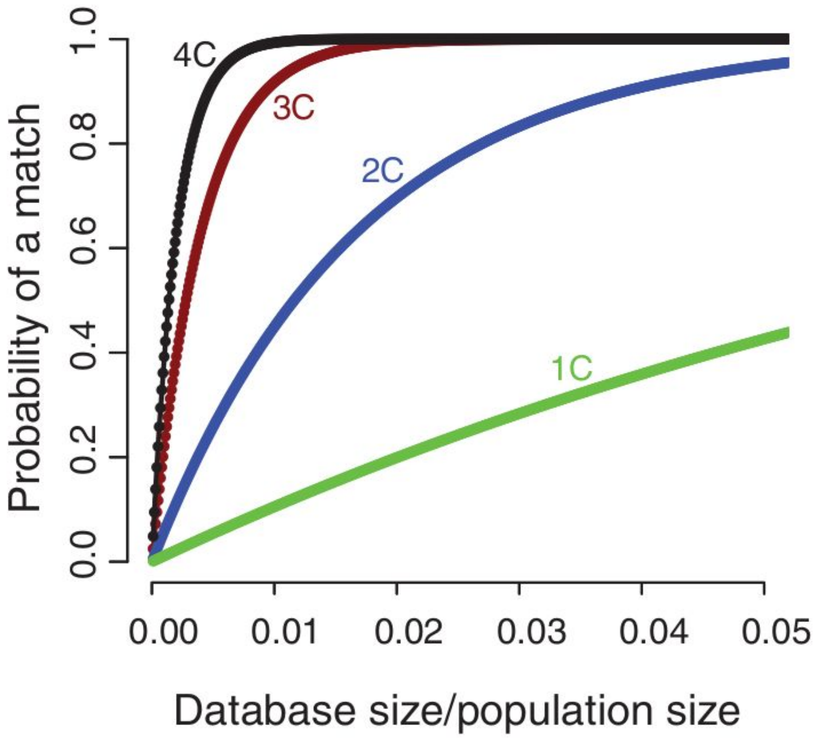
Poor API and design choices on GEDmatch contributed significantly to the vulnerabilities we uncovered:

- Lack of data authentication / integrity checks
- High resolution visualizations
- Ability to target specific users and direct queries
- Algorithms somewhat vulnerable by design

Responsibly disclosed results to GEDmatch, who modified their visualization algorithms to mitigate data extraction attacks.

Long term changes in the DTC industry, especially data authentication, are needed to prevent attacks via malicious data uploads and ensure long term security.

Security and the Future of Genetic Genealogy



Source: Erlich et al. Identity Inference of Genomic Data Using Long-Range Familial Searches. Science. 2018.

Consumer genetic genealogy databases have major implications for genetic privacy:

- Used to solve crimes and results are used in court
- Relevant to genetic surveillance and anonymous genetic data
  - 1M+ database: identification is possible by not easily scalable
  - 10M+: identification is simple

Encourage the community to develop methods to make genetic genealogy more secure by design