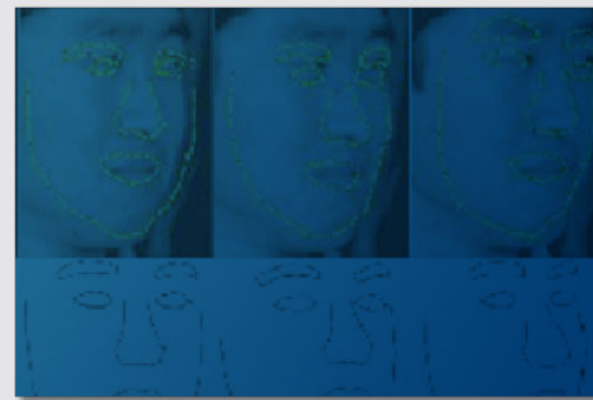
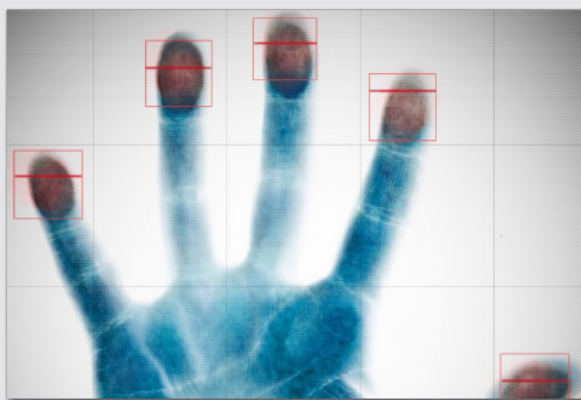
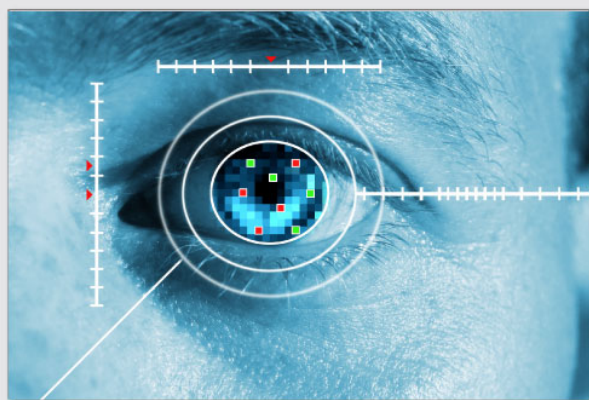


Multi-Stage Stratified Sampling for the Design of Large Scale Biometric Systems

Jad Ramadan, Mark Culp, Ken Ryan, Bojan Cukic

West Virginia University



Problem

- How to create a set of biometric samples for research?
 - How many subjects to include in a sample?
 - How are subjects chosen?
- Performance prediction requires adequate population samples too.
 - Convenience sampling introduces strong bias.
 - Alternative sampling methods have cost and practicality implications for data collections.



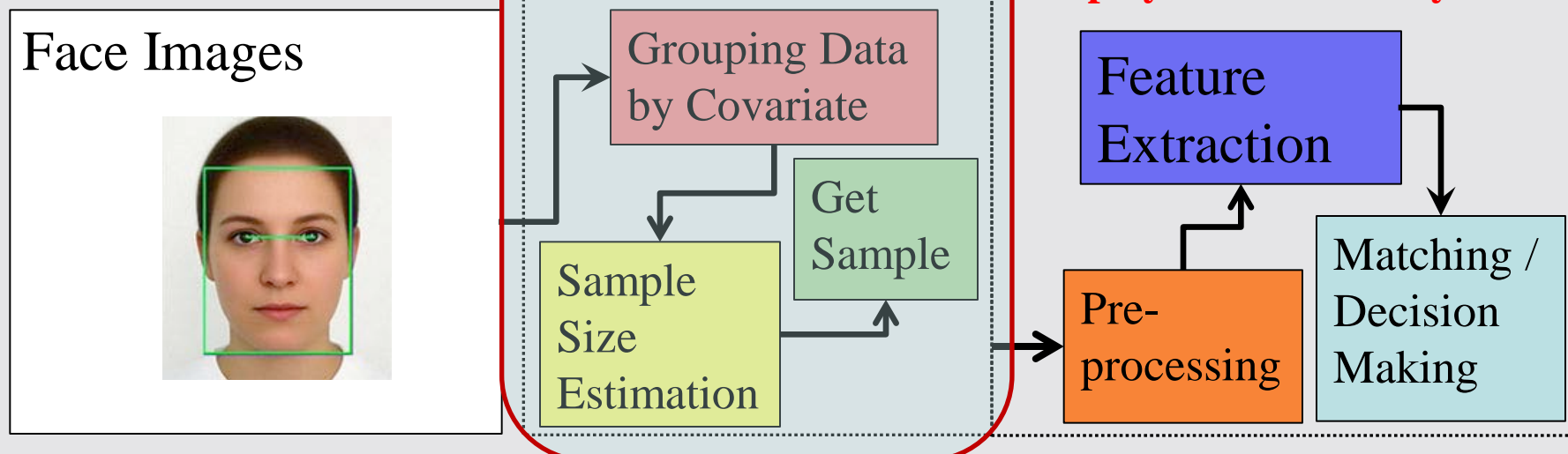
Stratification Benefits

- Stratification - the process of dividing population into homogeneous, mutually exclusive subgroups.
- Multi-stage stratified sampling design increases “trustworthiness” of match rate estimates
 - Lower costs and smaller performance prediction errors.
- We address the following specific questions:
 1. How can a researcher use existing large datasets to generate stratified samples for the purpose of biometric performance prediction?
 2. What are practical benefits of stratification?



Our Approach

- The process:

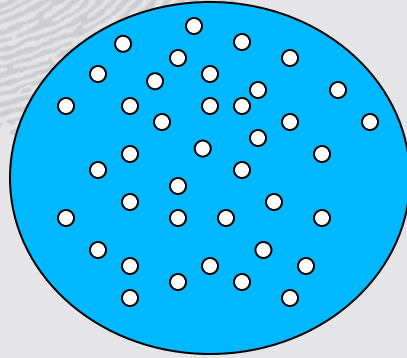


- We investigate the **Performance Prediction** phase.
 - Sample size estimation approach for Rank 1 identification rate estimation.

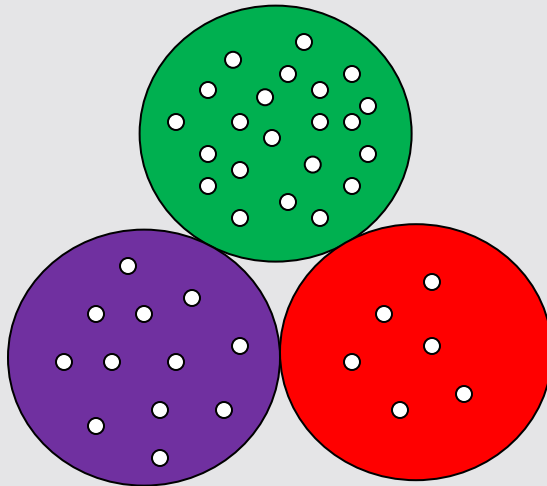
Stratified Sampling in Biometrics

- Stratified Sampling first partitions the population into L available groups (e.g. males, females).
 - Within each group, a sample is created by taking an *independent simple random sample*.
- Goal: Participants within each group are as similar as possible.
 - *Individual stratum variances are minimized.*
- What is the criteria for effective grouping?
 - There should be *clear differences in match rates between strata*.
 - May be algorithm dependent!
 - Strata based on eye color, facial hair or hair color do not exhibit this.
 - In face recognition, age group, ethnicity and gender *could be used as strata*.

Stratified and Simple Random Sampling: Difference



- *Simple Random Sampling* takes a sample from a population in a way so that each sample has the same chance of being selected.



- In *stratified random sampling*, the population is first separated into non-overlapping strata . A sample is created by simple random sampling from each stratum.
- *Sample size from each strata may differ.*

Intuition:

How tall are NBA players?

- **# Players: 434; Mean height: 79.04in; Variance: 12.9 in²**
- How many players must be sampled to *estimate the average height to within one inch*?
- Grouping the players by position reduces variance
 - 5.94 in.² (guards), 2.32 in.² (forwards), 1.85 in.² (centers)
- **Simple random sampling: 47 observations.**
- **Stratified sampling: 13** (optimally allocated) observations.
- A stratified sample of 7 guards, 4 forwards, and 2 centers selected from *any* NBA season will yield an estimate of the mean height from that season, within an inch, 95% of the time.



Large Face Data Sets

- In large data sets, the number of false matches tends to increase.
 - Imposter score correlations *close to 0* within each cluster helps reduce the FMR.
- We investigated imposter score correlations within the strata (e.g. African American females, Caucasian males).
 - Pinellas County Sherriff's Office data set.
 - Most of the subjects are white males.
 - 2.5K each for male/female and black/white demographics.
 - Experiment: FaceVACS 8.6.0, 10,000x10,000 match scores.



CITeR

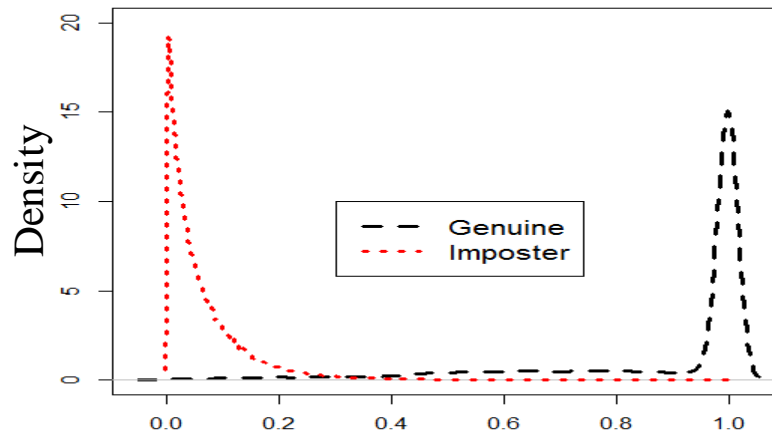
CENTER FOR
IDENTIFICATION
TECHNOLOGY
RESEARCH

8

© CITeR

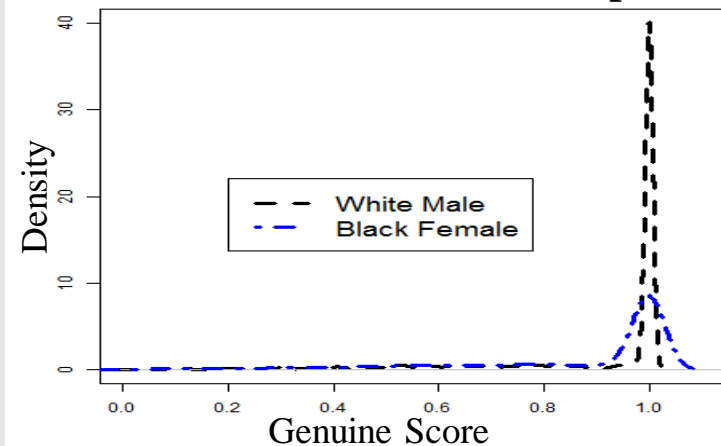
Genuine/Imposter Score Distributions

FaceVACS Similarity Score Distribution

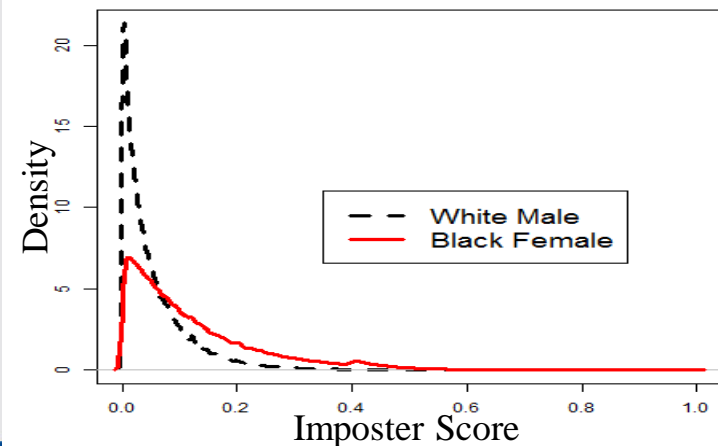


- Score distributions change with demographic information.
- Black female similarity scores exhibit a larger variance.
 - Added uncertainty will have a significant impact in matching.

FaceVACS Genuine Score Comparison

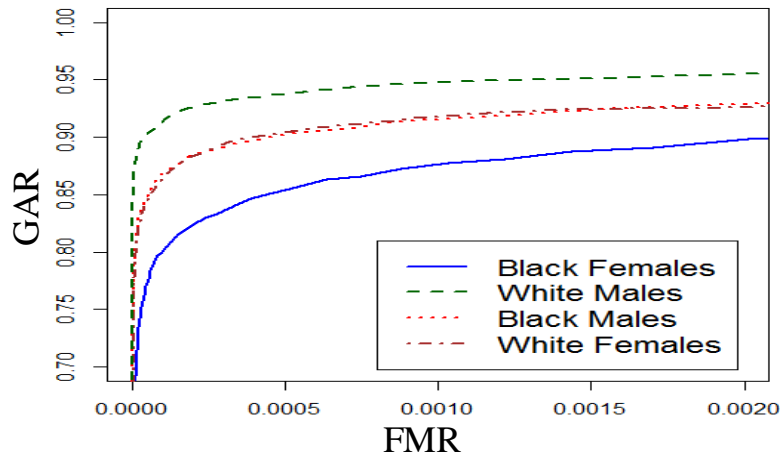


FaceVACS Imposter Score Comparison

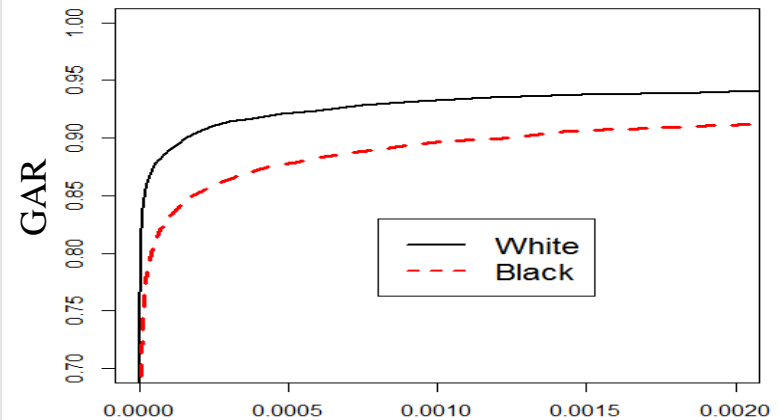


Cohort Interactions

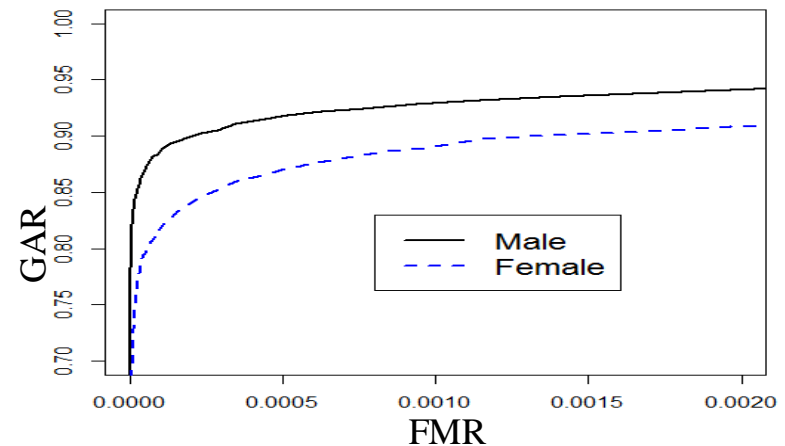
ROC curve for 4 cohort combinations



Black/White ROC Curve Comparison



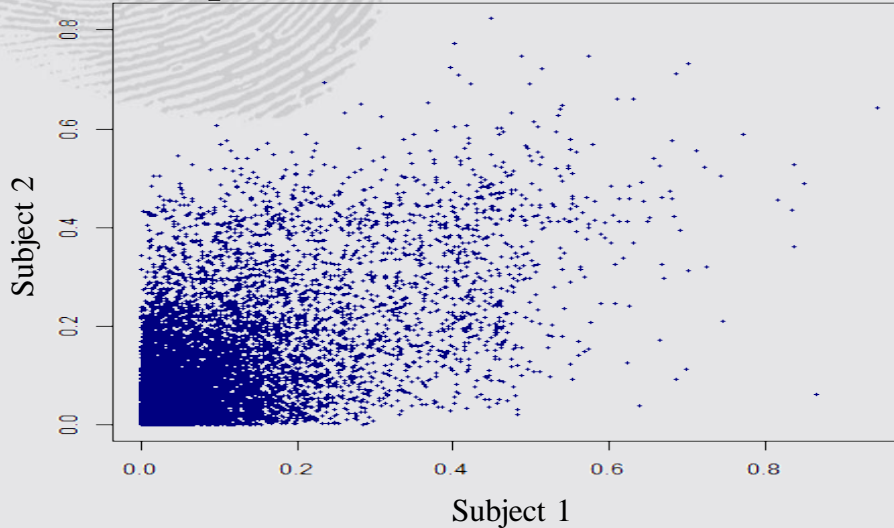
Male/Female ROC Curve Comparison



- If the variation in similarity scores among black females is reduced,
And
- If no imposter score correlation existed,
 - Black females would become more identifiable.

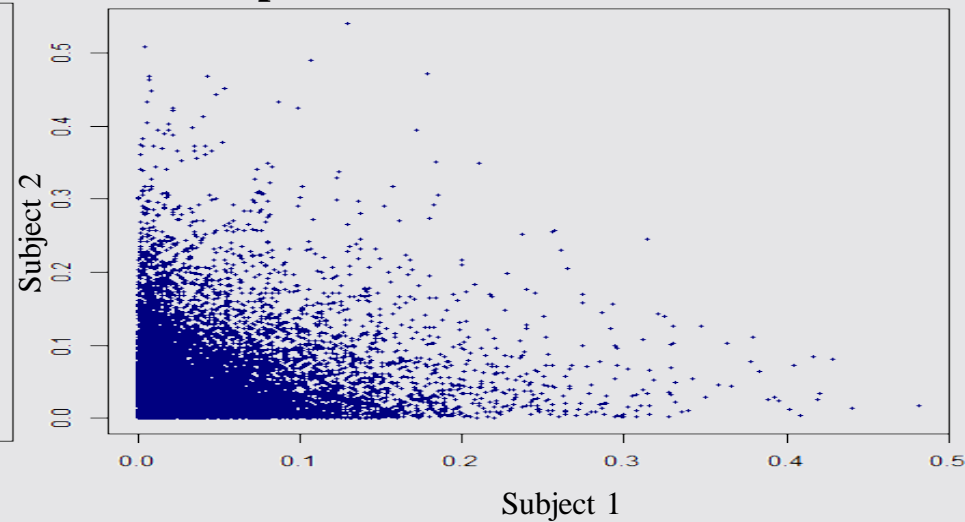
Impostor Score Correlation

Impostor Scores of 2 Black Females



- The correlation coefficient above is **0.595**.
- Similarity scores between different, unrelated subjects exhibit almost a linear relationship.
- The matcher has difficulties differentiating between the black females.

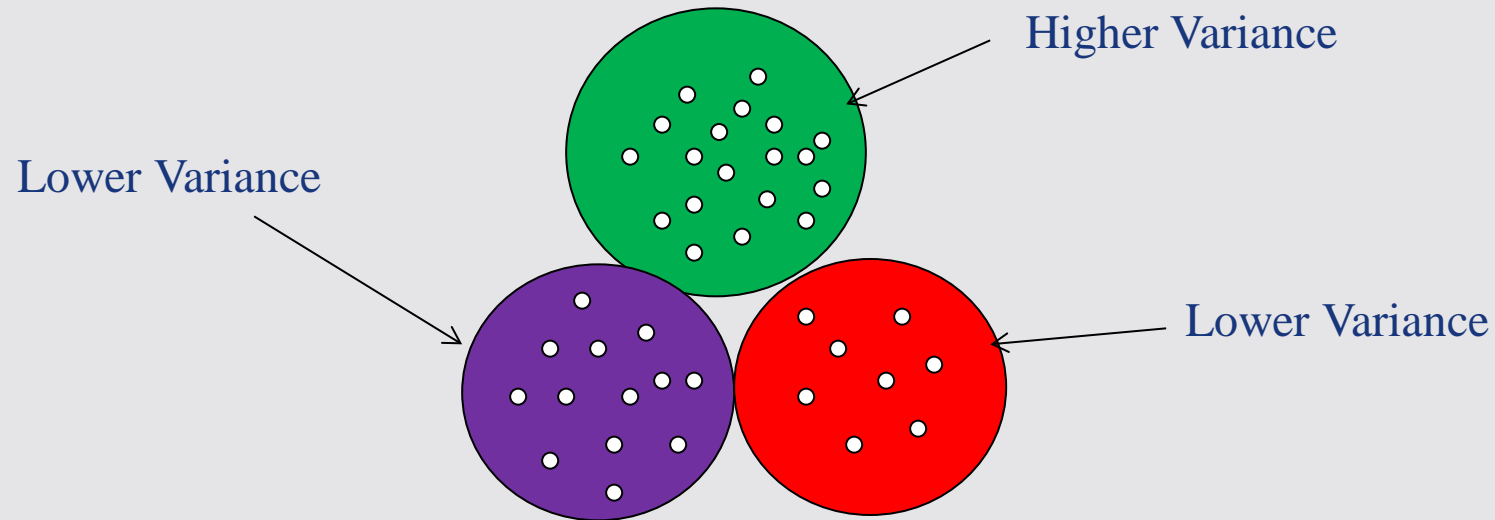
Impostor Scores of 2 White Females



- The correlation coefficient here is **-0.006** (no relationship).
- This is desirable because the matcher is having a much easier time differentiating the individuals in these images.

Stratified Sampling, Correlation, Large Samples

- Stratified sampling assigns a higher “weight” to cohorts that are seen to cause difficulties in facial recognition



- Correlations among imposter scores of black females likely due to insufficient training with black female samples [Klare et al.]

Match Results

Table: Genuine Accept Rates (GAR) at a fixed False Accept Rate of 0.01%

<i>Demographic</i>	Black Females	Black Males	White Females	White Males	Overall
<i>GAR</i>	0.8048	0.87	0.8684	0.916	0.853

Grouped by Gender

Male	Female
94.4%	89.5%

Grouped by Ethnicity

Black	White	Hispanic
88.7%	94.4%	95.7%

Grouped by Age

18-30	30-50	50-70
91.7%	94.6%	94.4%

- Large difference in GAR between black females and white males.
- *Face Recognition Performance: Role of Demographic Information* [Klare et al.]
 - There seem to be extra interactions with gender and ethnicity that increase differences in match rates.
 - Dynamic face matcher selection.



Sample Size Equations

- B represents a chosen bound.
- N (N_k) is overall sample (strata) size.
- p (p_k) is the GAR at FAR 0.01%.

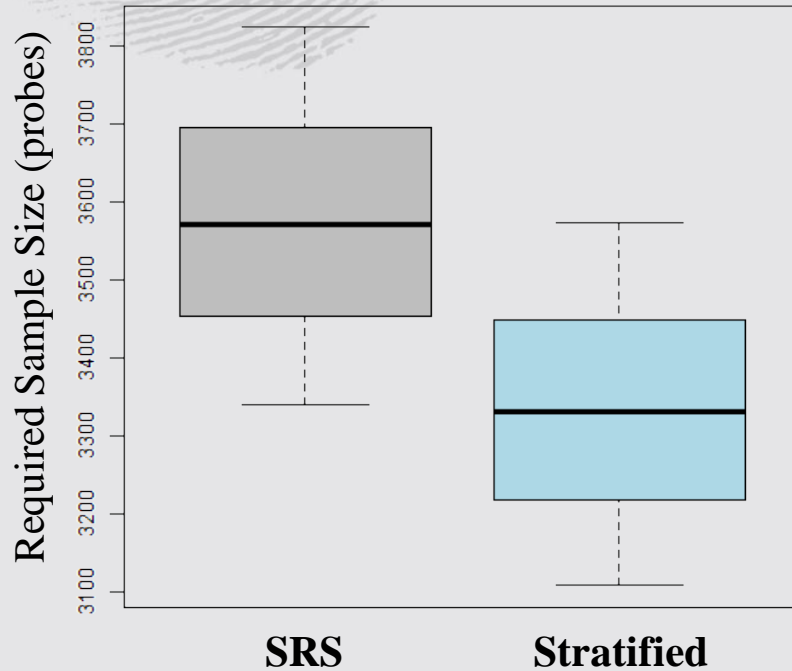
Stratified Random Sampling:
$$n = \frac{4 \left(\sum_{k=1}^L N_k \sqrt{p_k (1 - p_k)} \right)^2}{N^2 B^2 + 4 \sum_{i=1}^L N_i p_i (1 - p_i)}$$

Simple Random Sampling:
$$n = \frac{4 N p (1 - p)}{(N - 1) B^2 + 4 p (1 - p)}$$



Data Stratification

Results using our 4 Cohorts as Strata



Note: The error bound for the plot above ranges from 0.9% to 1%.

Table: Allocation of the sample based on Stratification

Black Female	Black Male	White Female	White male
~30%	~25%	~25%	~20%

- Stratified sampling, using the 4 cohorts of interest, now allows for 230 fewer tests to estimate performance within 1%.
- An added bonus:
 - The next collection may emphasize the sampling of black females, the most troublesome cohort.

Results

- The total sample sizes below were obtained using an error bound of 1%

Simple Random Sampling

- Total Size: 3341
 - Allocation:
 - 843 black females
 - 834 black males
 - 842 white females
 - 822 white males
- Estimated GAR of 85.3% at an FMR of 0.01% .

Stratified Random Sampling

- Total Size: 3109
 - Allocation:
 - 933 black females
 - 777 black males
 - 777 white females
 - 622 white males
- Estimated GAR of 85.3% at an FMR of 0.01% .

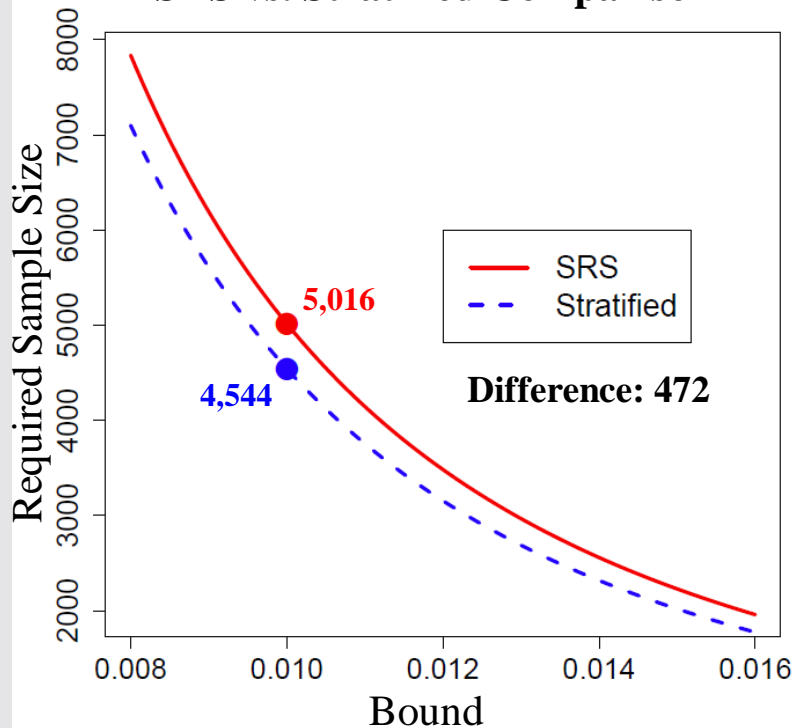
- Stratified random sampling achieved the same performance using 232 fewer subjects.



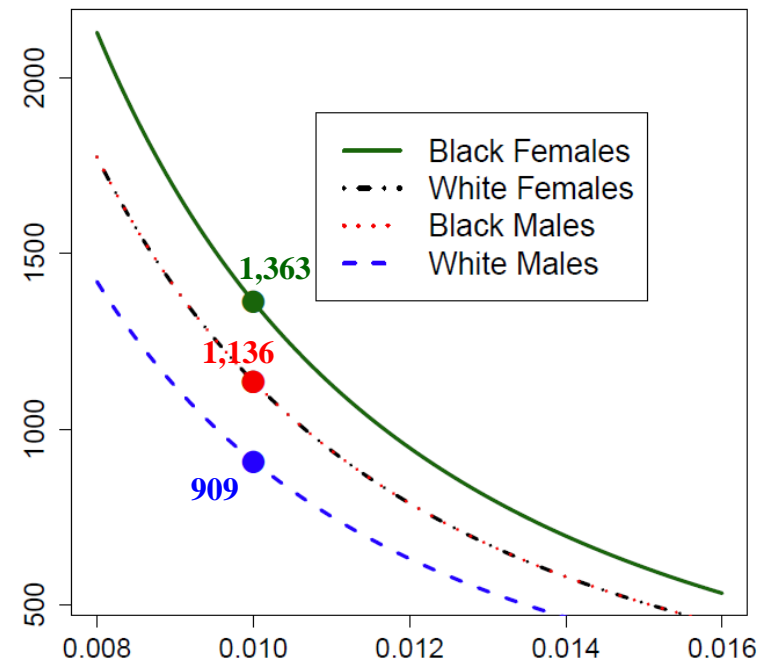
Data Extrapolation

- The total sample sizes below were obtained using an error bound of 1%.
- Differences when predicting a population of one billion. From previous studies, we are assuming a GAR of 85.3% at .01% FMR

SRS vs. Stratified Comparison



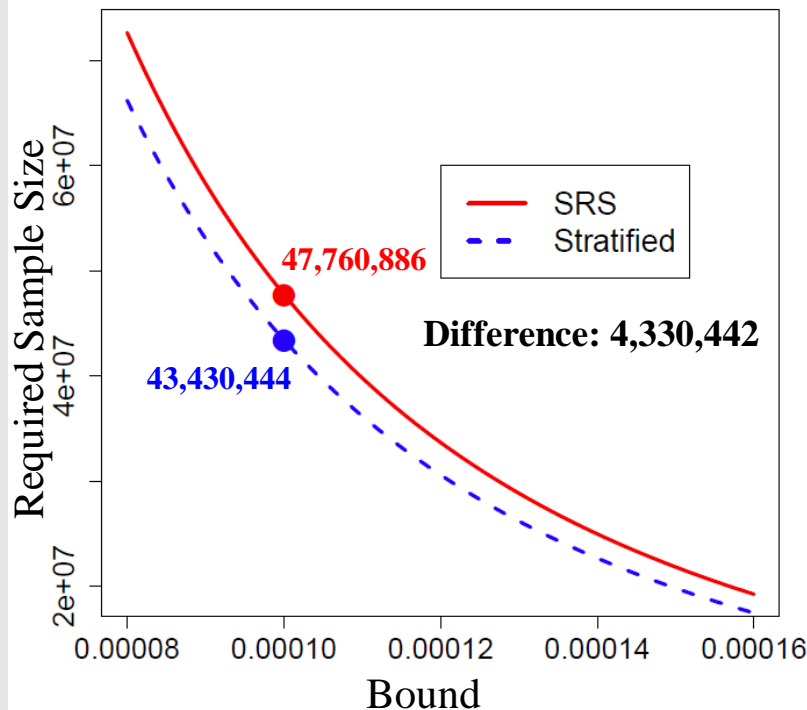
Allocation for Stratified Sampling



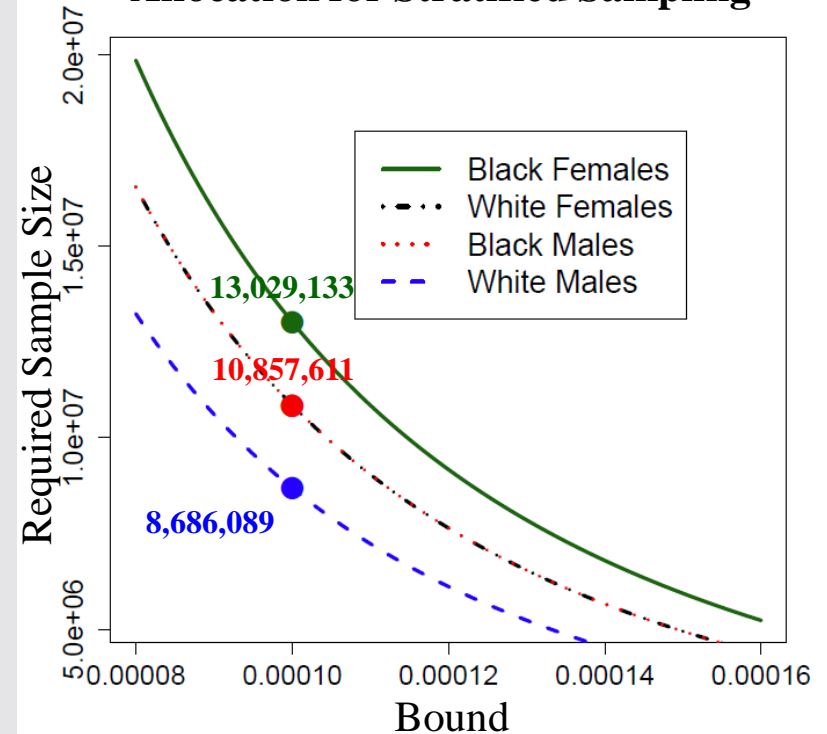
Data Extrapolation

- Now, we calculate the necessary sample size using an error bound of 0.01%.
- How many samples from a population of 1 billion would we need to estimate the GAR at 0.01% FMR to within 0.0001?

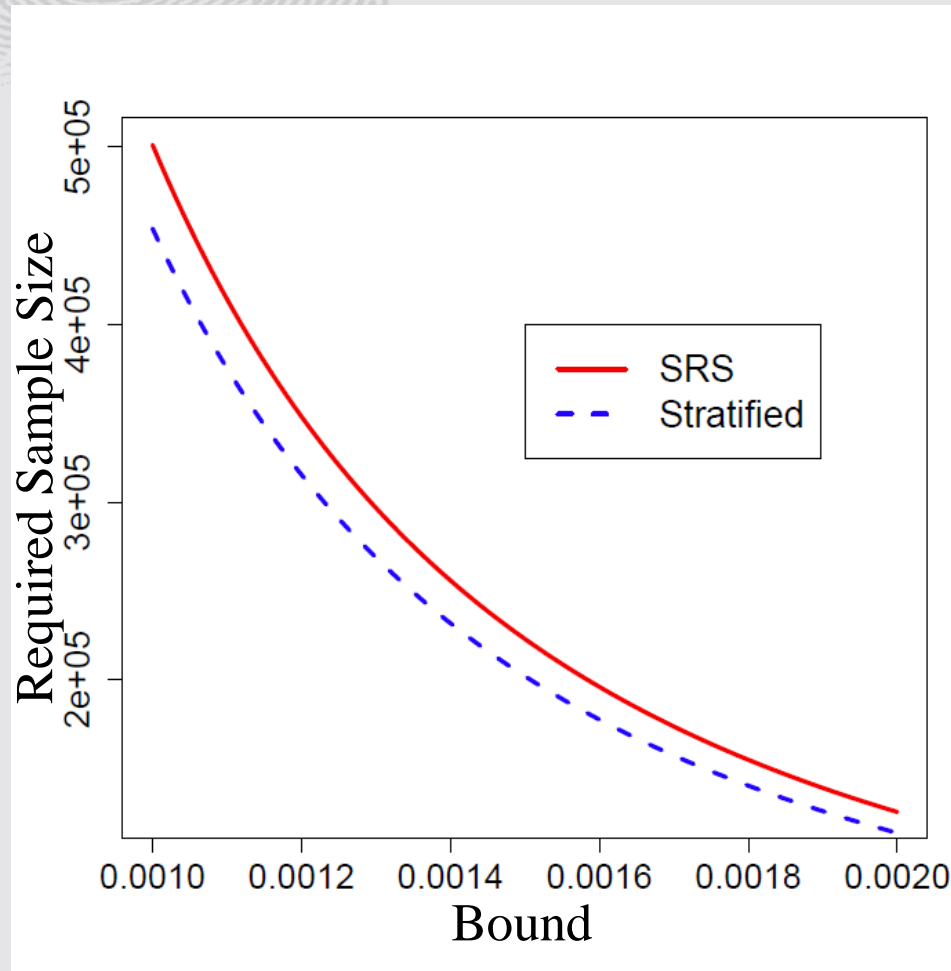
SRS vs. Stratified Comparison



Allocation for Stratified Sampling



Sample Size Reduction



- Stratified sampling requires around 10% fewer subjects to achieve the same performance estimate, regardless of the chosen error bound.
- In general, *the choice of error bound will not have an impact on the sample size reduction due to stratification.*

The Effect of Errors in Stratification

- In a simulation, 10% and 33% of African American population was reclassified as white and vice versa.

Simulate the effects of an incorrect classification by an algorithm or experimenter.

- Results (baseline is the leftmost table):

GAR at .01% FMR (no errors)

Black	White
83.3 %	89.02 %

GAR at .01% FMR (10% errors)

Black	White
83.4 %	88.26 %

33% errors

Black	White
84.3 %	86.06 %

Differences:

Black	White
+0.1 %	-0.76%

Black	White
+1.0 %	-2.96 %



Summary

- **Applied a stratified sampling design to face recognition.**
Approach offers savings in performance prediction for large systems.
Offers guidance for performance prediction from existing collections.
- **Unbiased performance predictions from a stratified sample.**
Given valid assumptions, performance predictions are accurate
The reward comes from the ability to allocate the sample.
- **Investigated the effect of errors in demographic information.**
The strata seem robust to small strata misclassification.
- **Should be extended to other biometric modalities.**
The role of matching algorithms.

