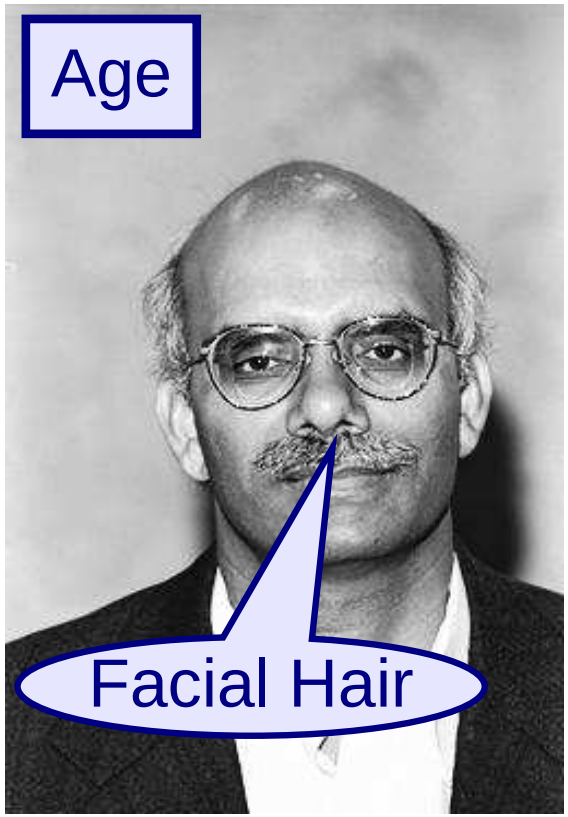
# GLMMs for Analyzing Verification Performance

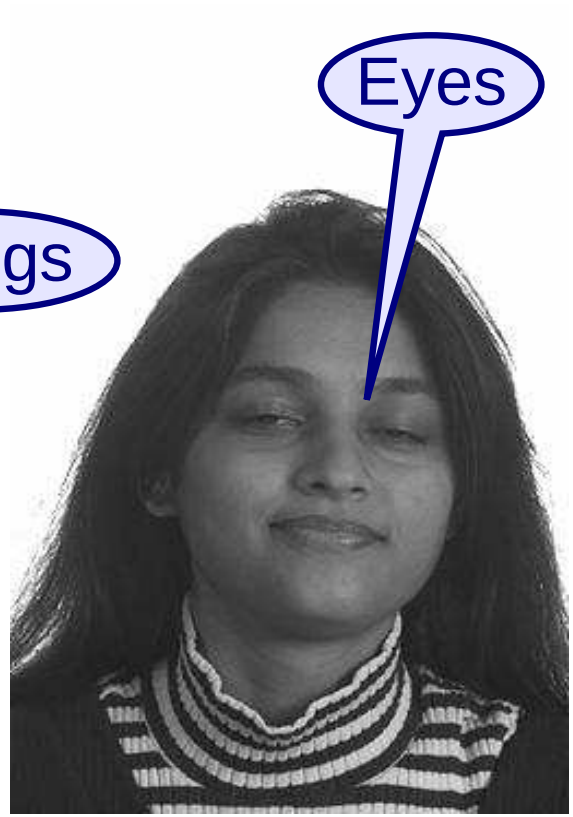
**Jonathon Phillips*, Geof Givens**, Ross Beveridge & Bruce Draper**

***National Institute of Standards and Technology**
****Statistics, Colorado State University**
**Computer Science, Colorado State University**

# Context - What Makes Subjects Harder or Easier to Recognize

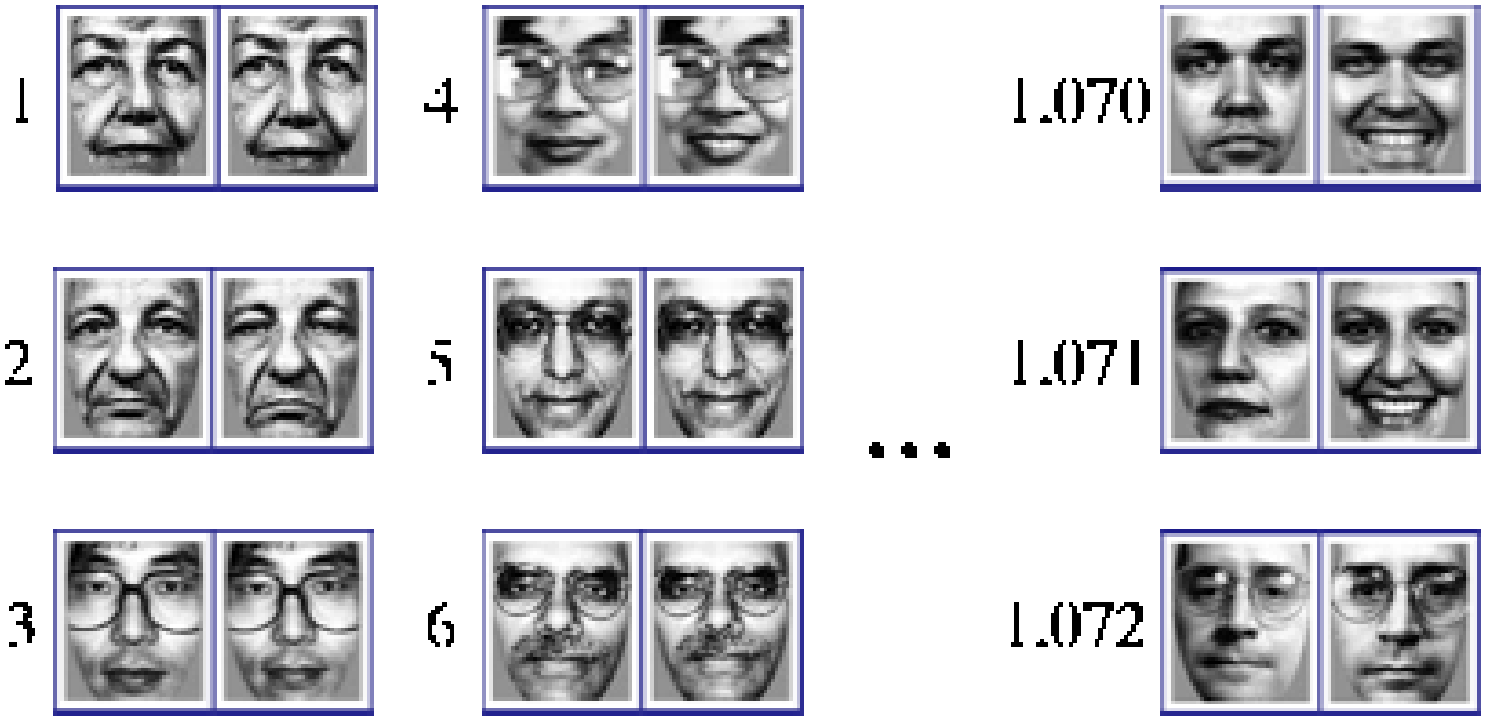




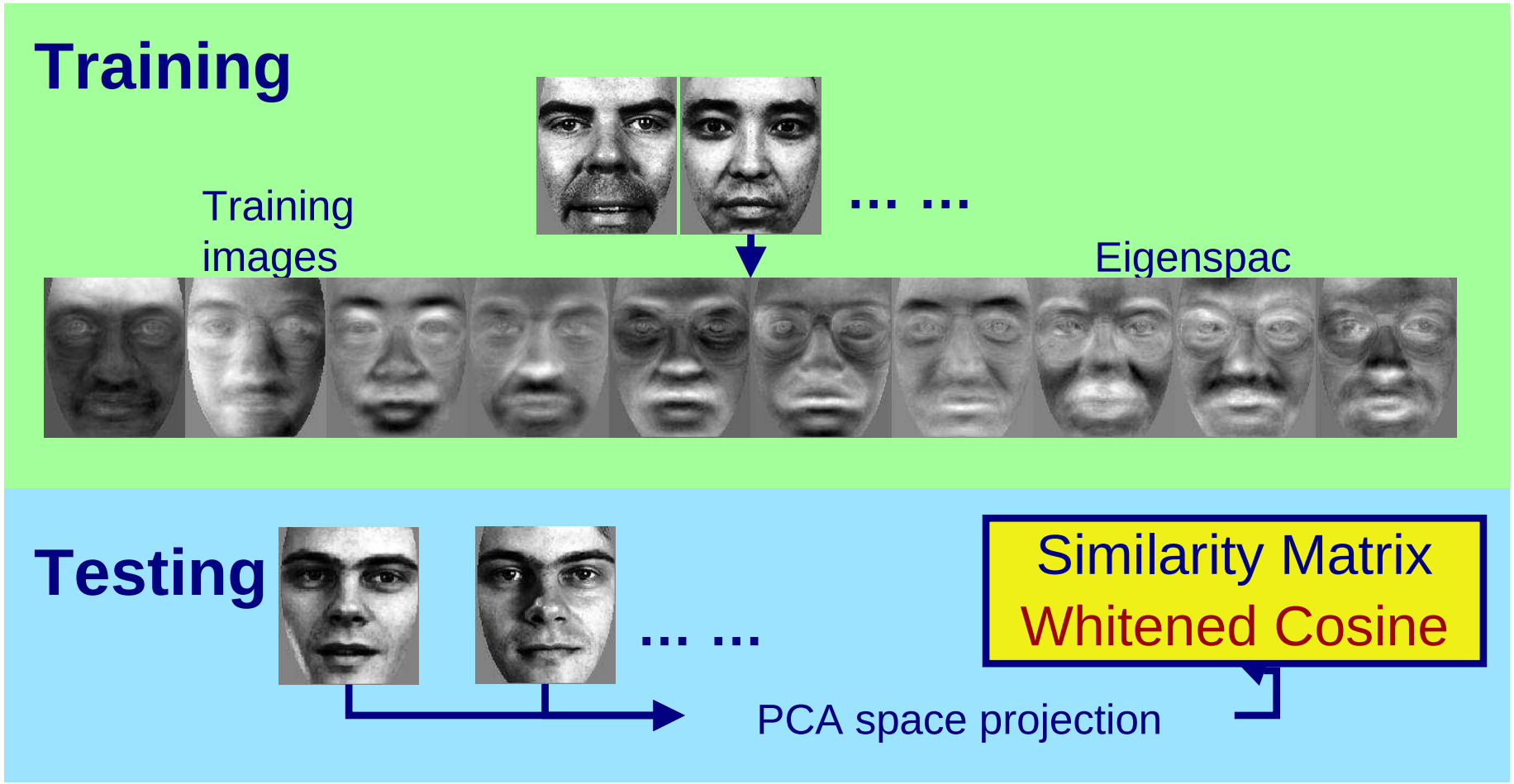


# Image Data for Study One

- 1,072 Human Subjects from the FERET Data
- 2,144 FERET Images
- Exactly 2 images per subject, taken on same day



# Algorithm for this Study:
# PCA with Whitened Cosine

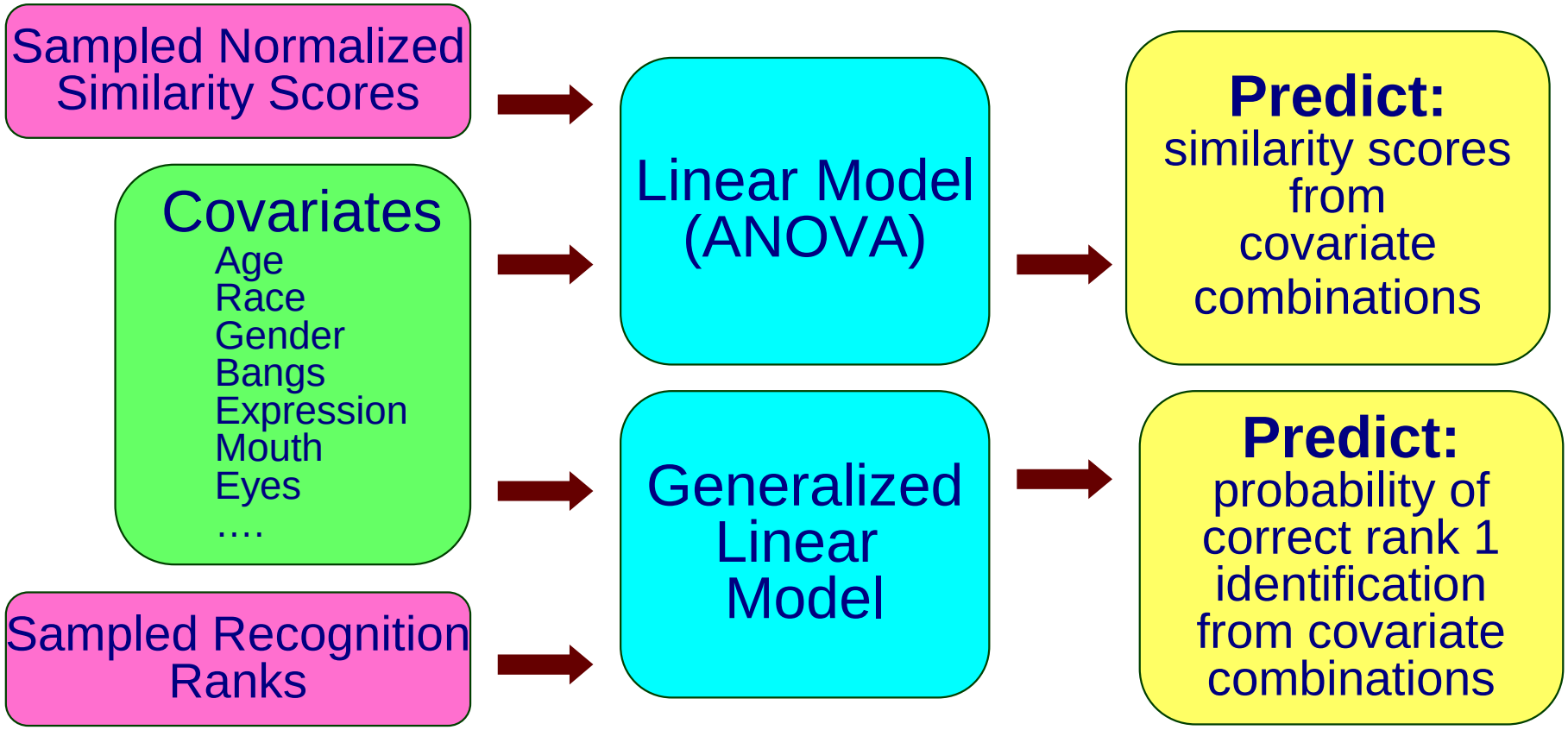



*M. A. Turk & A. P. Pentland. Face Recognition Using Eigenfaces, CVPR 1991*

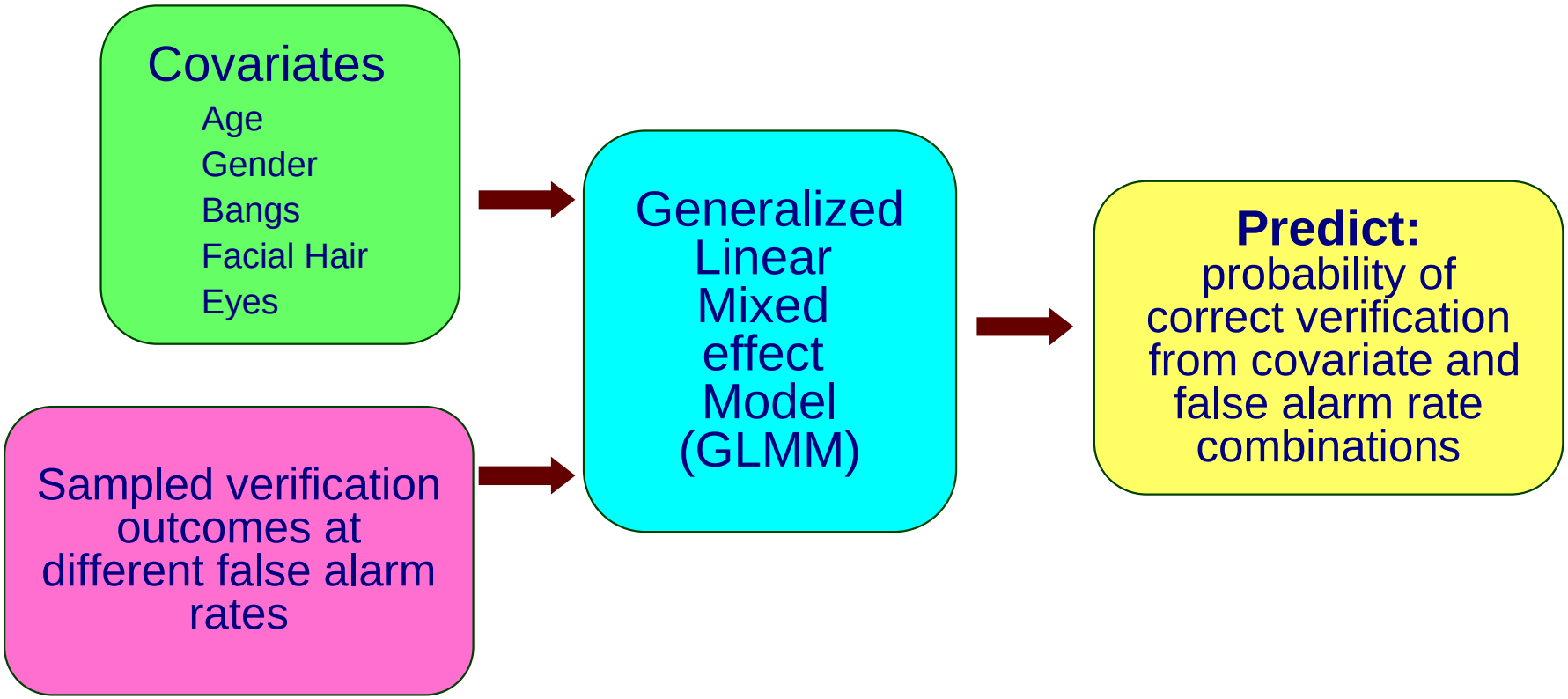
# Algorithm for this Study: PCA with Whitened Cosine

- View the algorithm as a stand-in.
- Most important feature of this talk …
  - *A new way to study verification performance.*

# Past Work - Modeling Face Identification Performance



# This Work - Modeling Face Verification Performance



# Verification Indicator Variable and FAR settings

- Our study - 1,072 x 1,072 similarity matrix
  - 1,072 match scores, 1,148,112 non-match scores
- Sort non-match scores, greatest to least
  - Find similarity score yielding desired false alarm rate (FAR)
- For the 1,072 FERET subjects
  - Correctly verified when similarity score above threshold

Indicator Variable *Y* for each subject for each FAR setting:

1 verified
0 otherwise

7 settings total

| Setting | FAR ($\alpha$) | Rate per 10,000 |
|---|---|---|
| 1 | 1/10,000 | 1 |
| 2 | 1/5,000 | 2 |
| 3 | 1,2,500 | 4 |
| 4 | 1/1,000 | 10 |
| 5 | 1/500 | 20 |
| 6 | 1/250 | 40 |
| 7 | 1/100 | 100 |

# Covariates for This Study

| Factor | Levels | Count |
|---|---|---|
| Age | Young | 774 |
| | Old | 298 |
| Gender | Male | 624 |
| | Female | 448 |
| Bangs | Absent in both images | 812 |
| | Present in both images | 228 |
| | Absent in one image and present in the other | 32 |
| Facial Hair | Absent in both images | 912 |
| | Present in both images | 149 |
| | Absent in one image and present in the other | 11 |
| Eyes | Open in both images | 964 |
| | Closed in both images | 11 |
| | Open in one image and closed in the other | 97 |

# Random Effects are Important
# GLMM vs. GLM

- Some people are harder to recognize then others
- But, we don't care who specifically is hard or easy

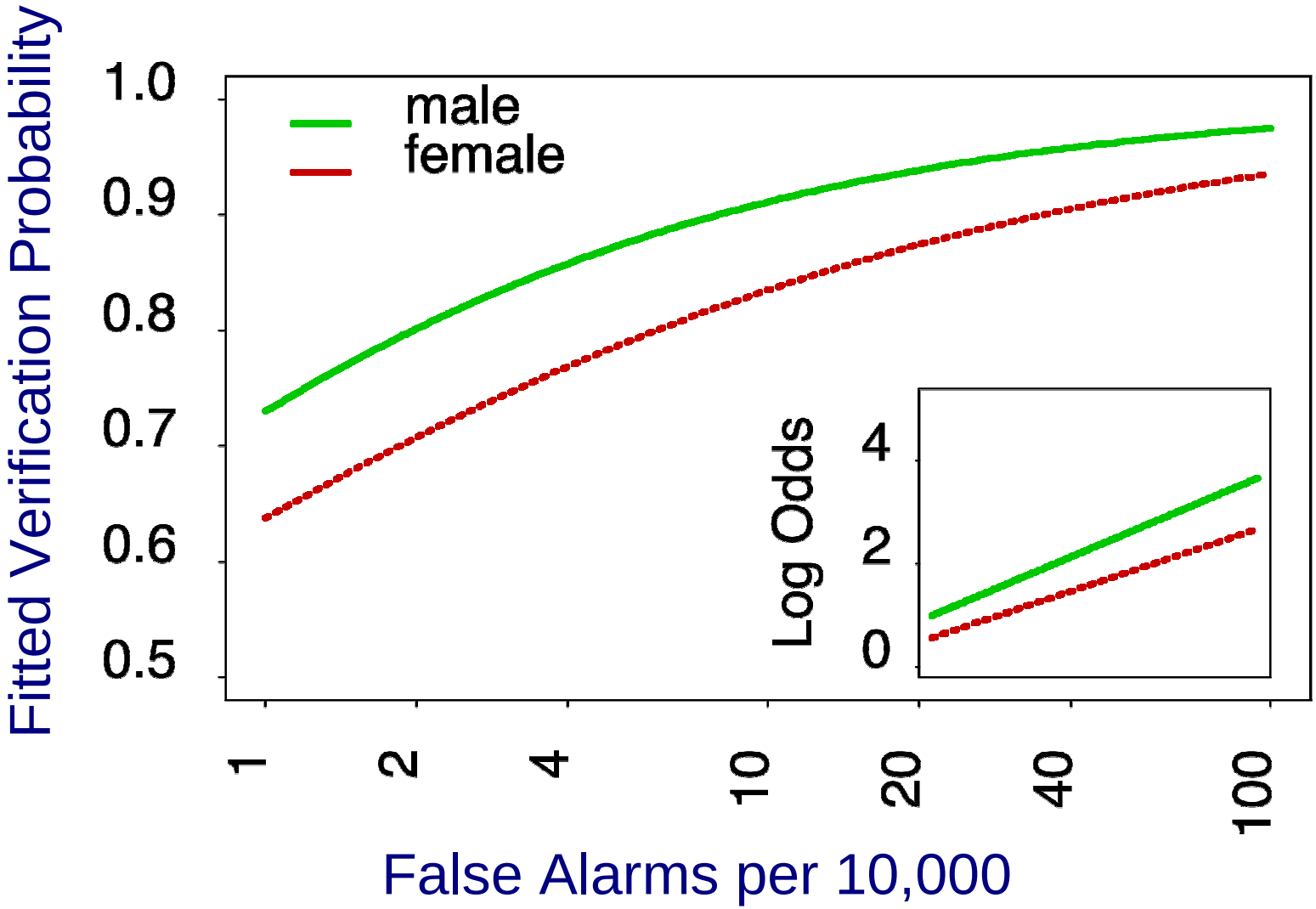
*Removing the "noise" of random effects reveals other significant effects of interest*

In an earlier study:

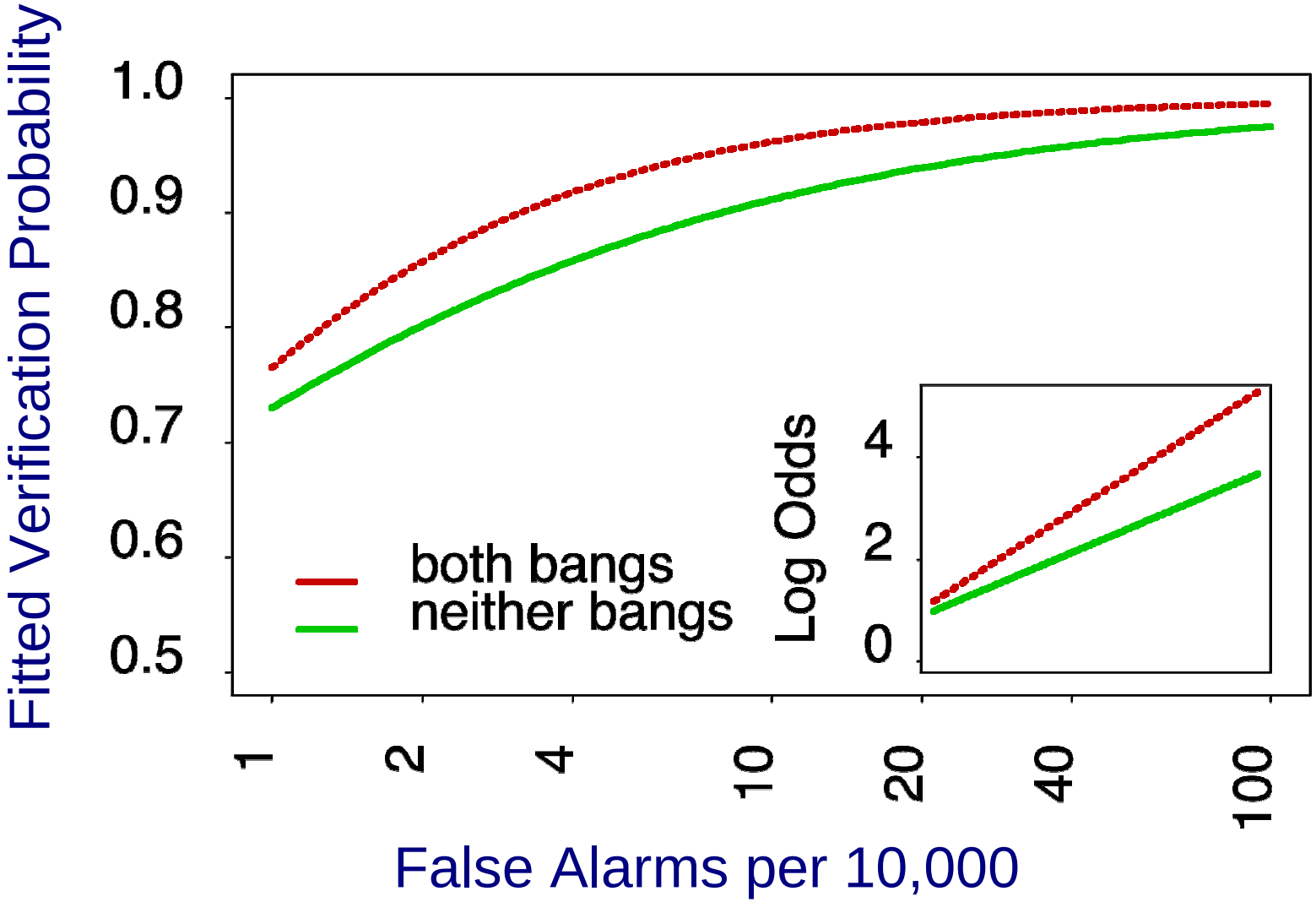
*About 70% of the random variation in rank 1 recognition probability is due to subject variation, not training variation or pure randomness*

*(Sorry - I don't have corresponding number for this study)*

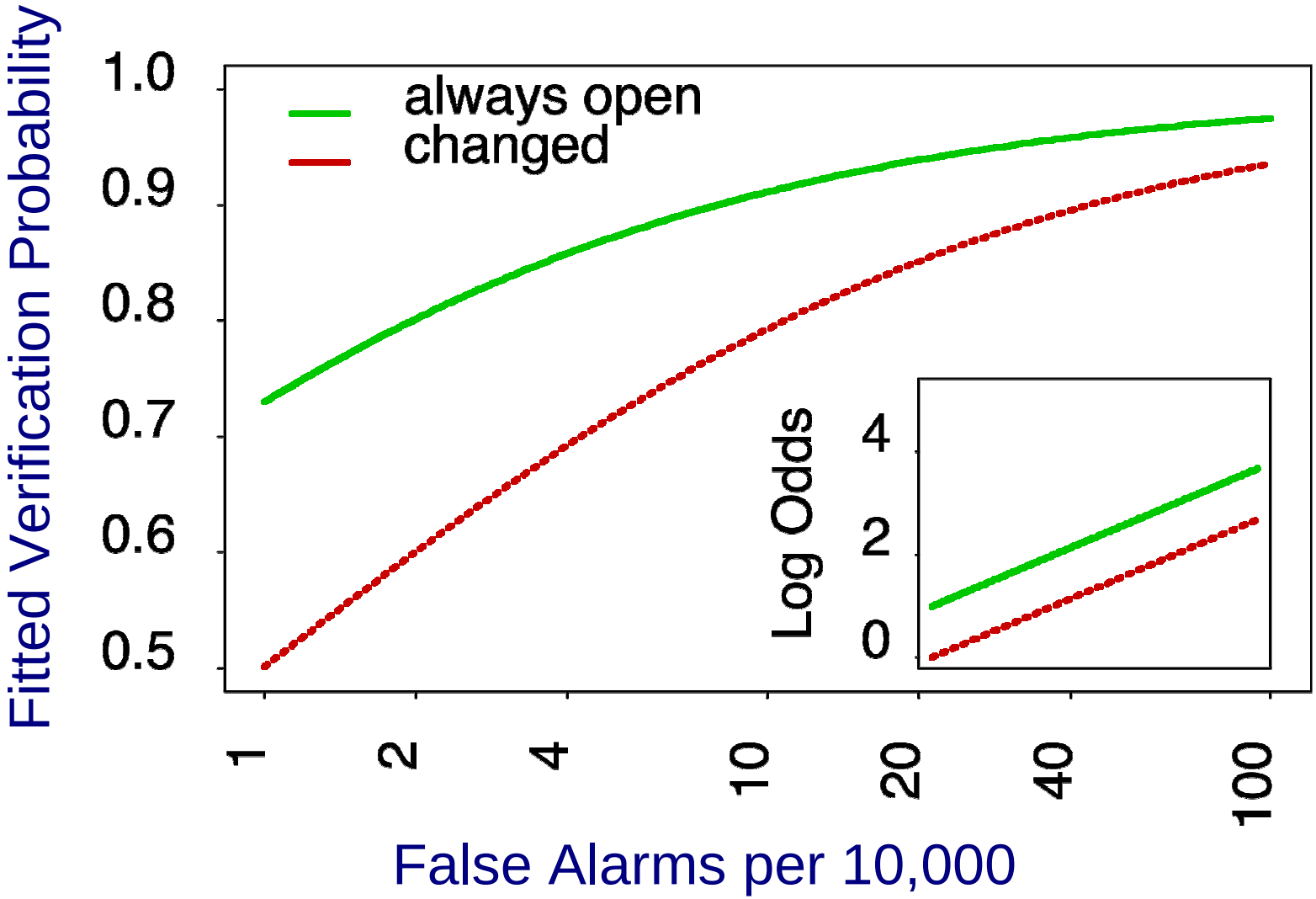
# Results of the Model - Gender



# Results of the Model - Bangs



# Results of the Model - Eyes



# Study Two

- Face Recognition Grand Challenge (FRGC)
  - State-of-the-art algorithms
  - Subject covariates
  - Image covariates

# Overview

**Goal:** *Quantitative analysis of when, and by how much, properties of subjects and images improve/degrade verification performance.*

**Setup:**

- FRGC Version 2.0 Experiment 4.
- People - 466
- Query Images - 8,014 uncontrolled lighting stills
- Target Images -16,028 controlled lighting stills
- Three algorithms - A, B and C
- Generalized Linear Mixed Effect (GLMM) Model

# The Statistical Model (GLMM Overview)

- Performance Variable
  - Was a subject correctly verified at a given False Accept Rate?
- Fixed Effect Covariates
  - Subject Covariates
    - Age, Gender, Race, Expression (Smile, Neutral), …
  - Image Covariates
    - Image-size, Rotation-of-face, Focus, …
  - False Accept Rate
    - 1/100, 1/200,1/350, 1/700, 1/1000, 1/2500, 1/5000 and 1/10,000
- Random Effect
  - Subject identity

# Data

- An observation/outcome for the model is:
  - A query-target image pair for a given subject
- Associated with each outcome is:
  - Performance variable, 1 if subject correctly verified
  - Which algorithm: A, B or C
  - An FAR setting
  - A listing of subject covariates: age, gender, …
  - A listing of image covariates: image-size, face-rotation, …
- How much data?
  - 134,760 observations*
  - 352 people
  - Outcomes balanced by person: 128 each

* Originally 135,168. 384 were removed due to problems with associated covariates.

# Findings

- General
  - Covariates matter & Covariates Interact
  - Covariate effects can be asymmetric
  - Few universal effects, i.e. same for all algorithms
- Surprises
  - Expression effect is asymmetric
  - Significant image-size algorithm interaction
  - Rotation matters - for all algorithms
- Mounting Evidence
  - Older males without glasses are easier
  - Race: Non-majority races (Asian, Hispanic) easier
  - Glasses: Hinders verification
- Trends
  - Top performer is more sensitive to covariates

# Findings: Expression

Neutral Neutral (NN)

Neutral Smiling (NS)

Smiling Neutral (SN)

Smiling Smiling (SS)

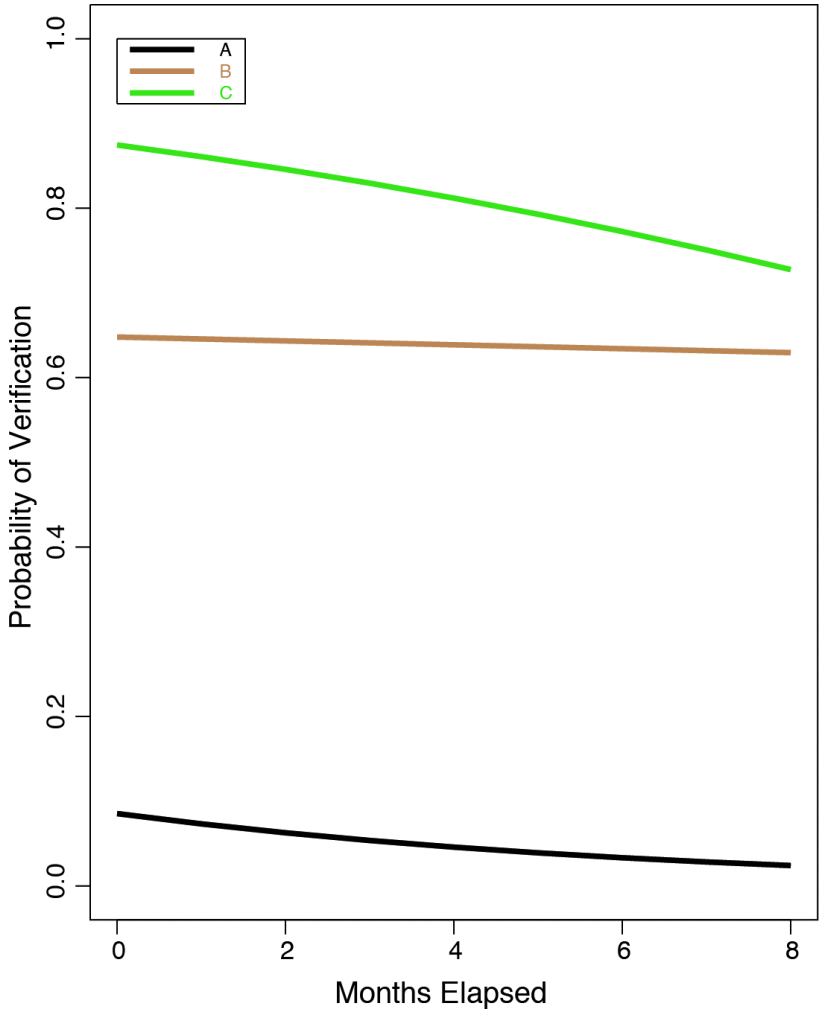
# Findings: Expression

Predicted Probability of Correct Verification by expression and algorithm.

| | A | B | C |
|---|---|---|---|
| NN | .065 | .644 | .850 |
| NS | .021 | .602 | .783 |
| SN | .033 | .571 | .718 |
| SS | .075 | .723 | .870 |

**Punchine: Have subjects smile during enrollment.**

# Trends: Robustness Vs. Performance




- Consider elapsed time between when the query and target images are acquired.
- For algorithms A and C, this time matters.
- For algorithm B it does not.
- Algorithm C, although better, is also more sensitive to changes.
- This is a general trend.

# Conclusions

- Successful extension of our previous work
  - From identification to verification.
  - Modeling probability of correct verification.
- Pleasant Surprise
  - Verification framework is in some respects superior.
  - Repeated measures allows shift from GLM to GLMM.
  - Better handles intrinsic differences between people.
- Interactions between FAR and other covariates.
  - Existence of interactions interesting - both directions.
  - Suggestive for algorithm designers.
- Future work
  - Study of similar nature but larger scope using FRGC results.

# Conclusions - Future Work Example

A previous talk showed the following results for 3D expression experiment:

- Histogram shows a subset of marginal effects over 3 covariates:
  - Algorithm 12 (alg1, alg2, …)
  - Modality 3 (Both, Shape, Texture)
  - Expression 2 (Neutral, Change)
- Using the approach just illustrated, we could capture these, plus
  - FAR 7 (1/10000, 1/5000, 1/1000, …)
- Individual contribution of factors made clear
- Interaction as well as subject variability taken into account
  - All
  - Neutral vs Neutral
  - Neutral vs NonNeutral
  - Expression Covariate

# Thank You

# Generalized Linear Mixed Model (GLMM)

Analysis is: ***Mixed Effects Logistic Regression with Repeated Measures on People.***

- Let **A** and **B** be 2 factors that might influence algorithm performance. For example, age and gender.
  - Example factor settings A=**a** and B=**b**.
- Let $\alpha$ be factor inducing a log linear effect.
- $Y_{pab\alpha}$ is
  - 1 if Person p is verified correctly,
  - 0 otherwise.
- $Y_{pab\alpha}$ depends on:
  - person **p**,
  - factors **A** and **B**, and
  - false alarm rate $\alpha$.

# GLMM Model Continued …

$Y_{pab\alpha}$ is Bernoulli R.V. with success probability $p_{pab\alpha}$

$$\log\left(\frac{p_{pab\alpha}}{1-p_{pab\alpha}}\right) = \mu + A_a + B_b + \gamma_\alpha \log(\alpha) + A_a \gamma_{a\alpha} \log(\alpha) + \pi_p$$

| | | |
|---|---|---|
| $\mu$ | = | grand mean |
| $A_a$ | = | effect of setting $a$ of factor $A$ |
| $B_b$ | = | effect of setting $b$ of factor $B$ |
| $\gamma_\alpha \log(\alpha)$ | = | linear dependence upon log of FAR $\alpha$ |
| $A_a \gamma_{a\alpha} \log(\alpha)$ | = | potential $A_a$ FAR interaction effect |
| $\pi_p$ | = | subject id. random effect (next page) |

# Subject Variation - The Mixed in Generalized Linear Mixed effect Model

$$\left[\pi_1, \ldots, \pi_{1,072}\right]^T \sim \text{Multivariate Normal where}$$

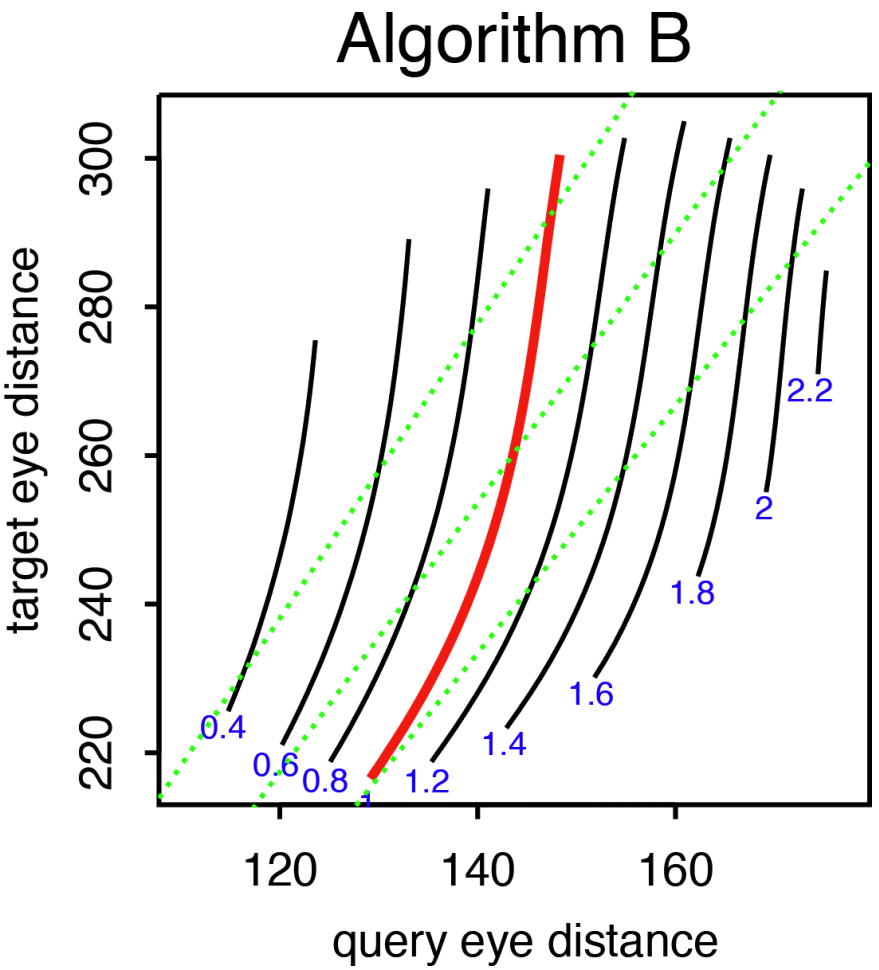
$$\mathrm{E}\left(\pi_p\right) = 0, \ \mathrm{Var}\ \pi_p = \sigma_\pi^2,$$

$$\mathrm{Cor}\left(y_{pab\alpha}, y_{p'a'b'\alpha'}\right) = \begin{cases} \gamma & \text{if} \quad p = p' \\ 0 & \text{if} \quad p \neq p' \end{cases}$$

**This means:**

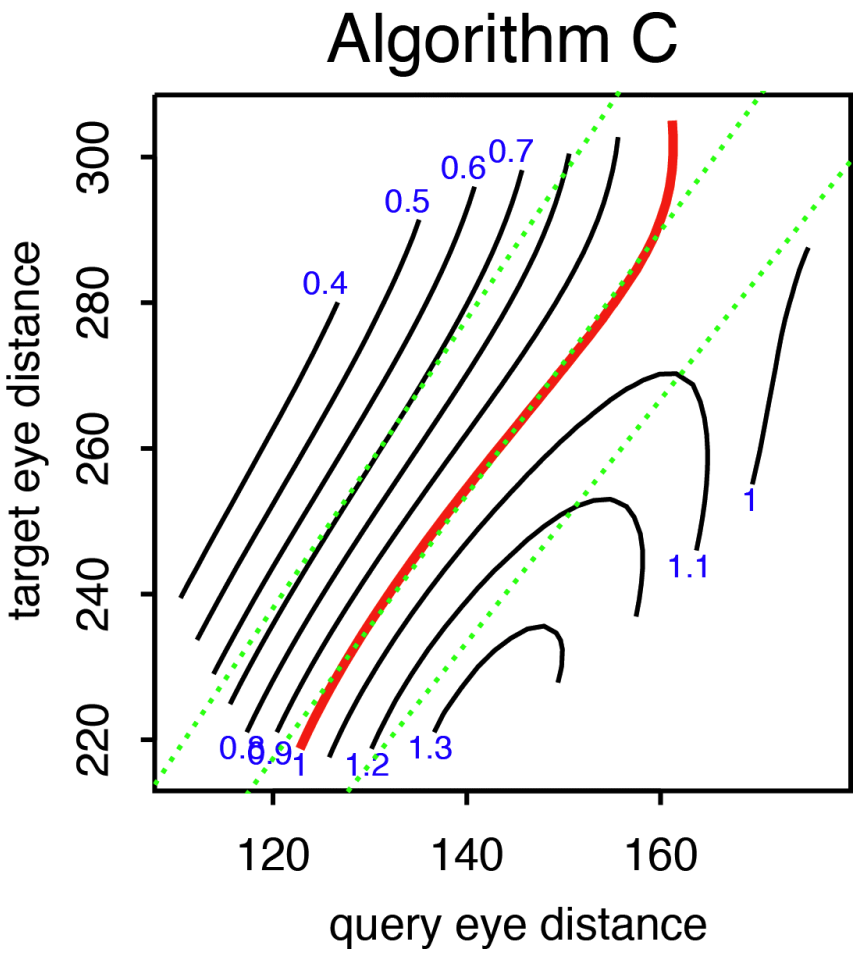
*The outcomes, i. e. verification success/failure, are uncorrelated when testing different people but correlated when testing the same person under different configurations.*

# Findings: Image Size




- Red line is baseline performance
- Contours indicate odds ratio.
- Larger means better
- Query image eye distance dominates algorithm performance
- Target image eye distance relatively unimportant
- ***Algorithm B wants as many pixels on face as possible***

# Findings: Image Size




- Unlike algorithm B, dependence of algorithm C on relative sizes of query and target images is complex
- ***Algorithm C wants relatively smaller target images***
- ***Algorithm C has a sweet spot in relative size of query and target images***