Confidence Intervals for 1:1 Matching Tasks

Riccardo Fogliato

arXiv: https://arxiv.org/abs/2306.01198
GitHub: https://github.com/awslabs/cis-matching-tasks

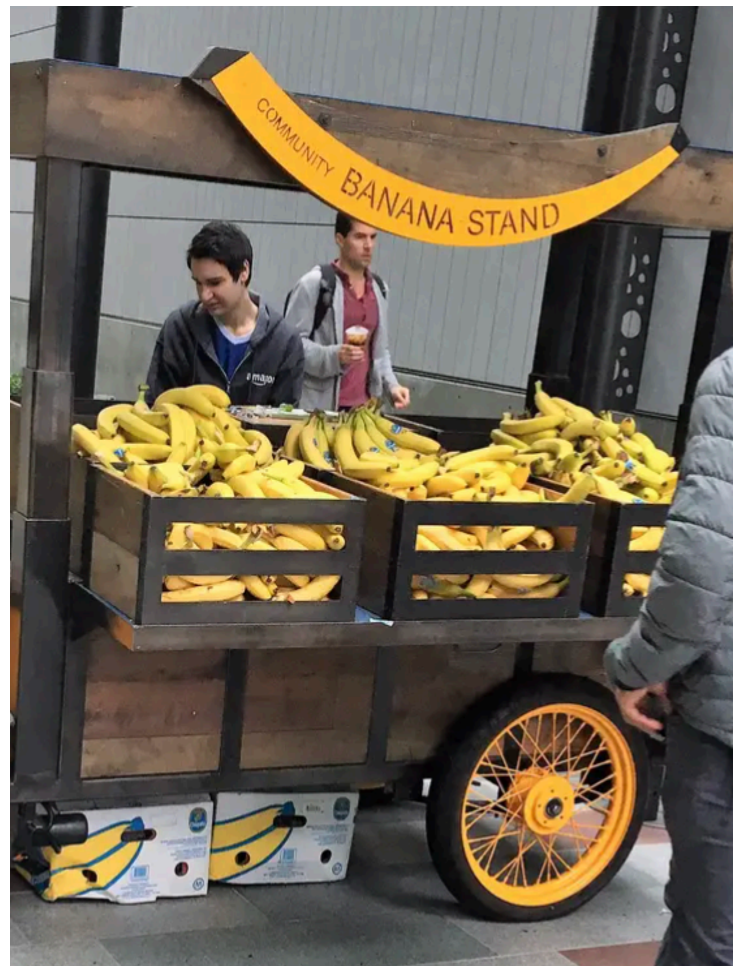
Background

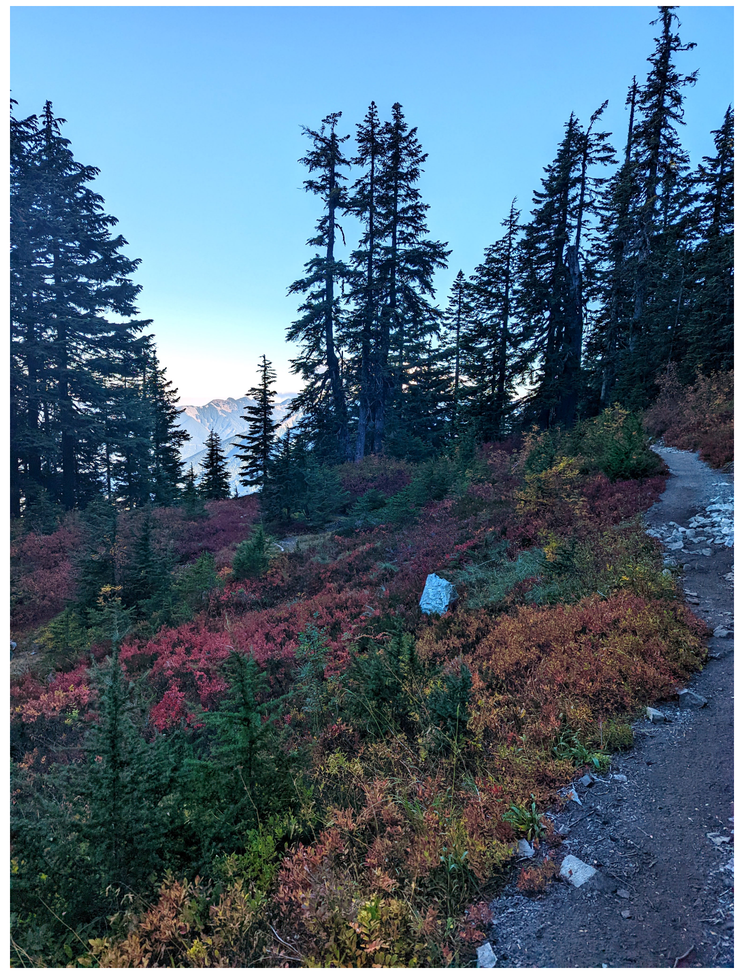
Undergrad @ Uni Padova

Master @ Uni Torino

PhD @ CMU

Applied Scientist @ AWS





Joint work with

Pratik Patil

UC Berkeley

Pietro Perona

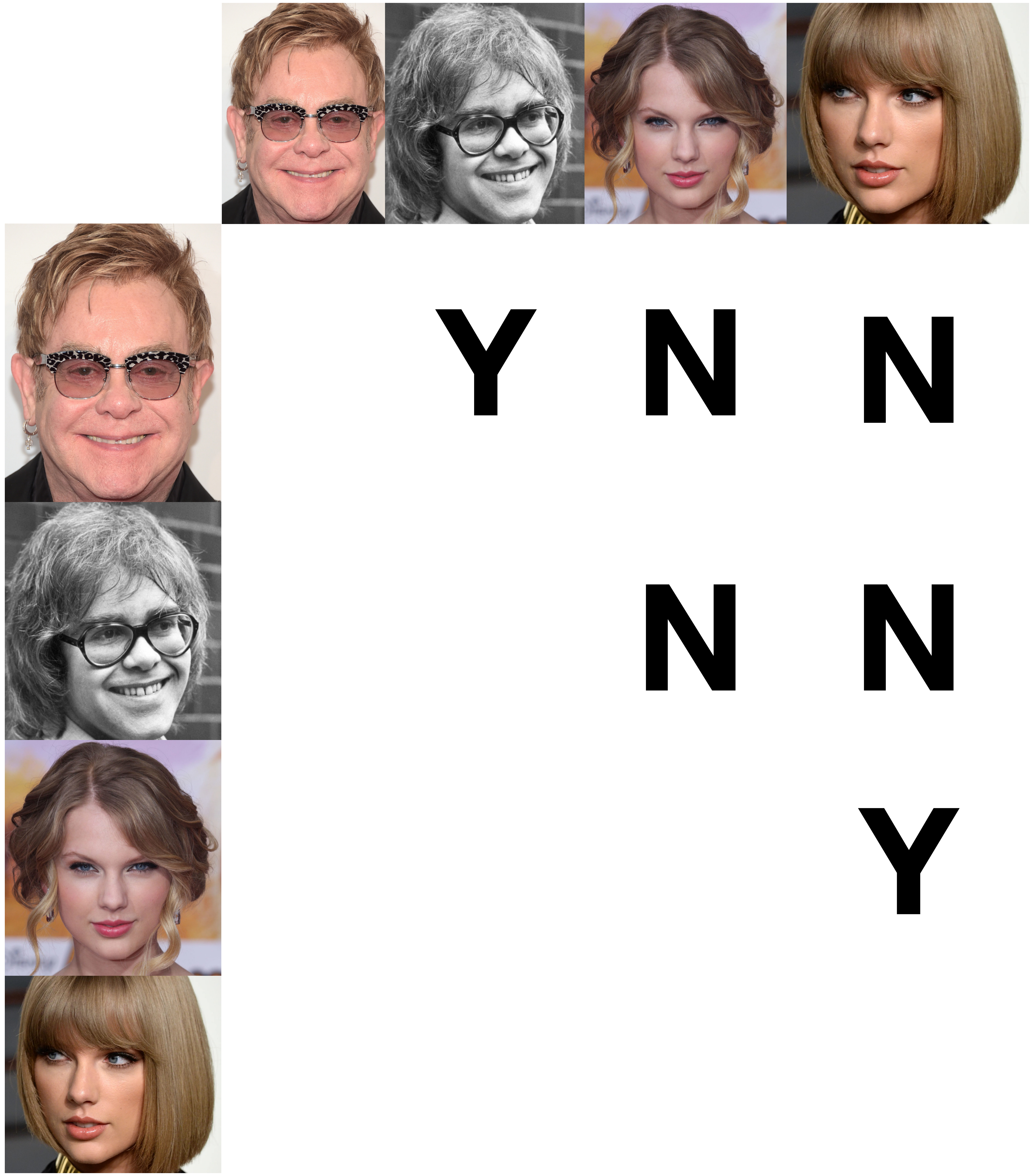
Caltech & AWS

1:1 matching tasks involve comparing two items to verify a match

Are the people in the two images the same person or not?

Y: Same person

N: Different person



Confidence Intervals for 1:1 Matching Tasks

Why do we need uncertainty?

TPM

Can you check the performance of our facial recognition service on this customer's data?

Scientist

Sure! Let me generate the predictions

Why do we need uncertainty?

TPM

Can you check the performance of our facial recognition service on this customer's data?

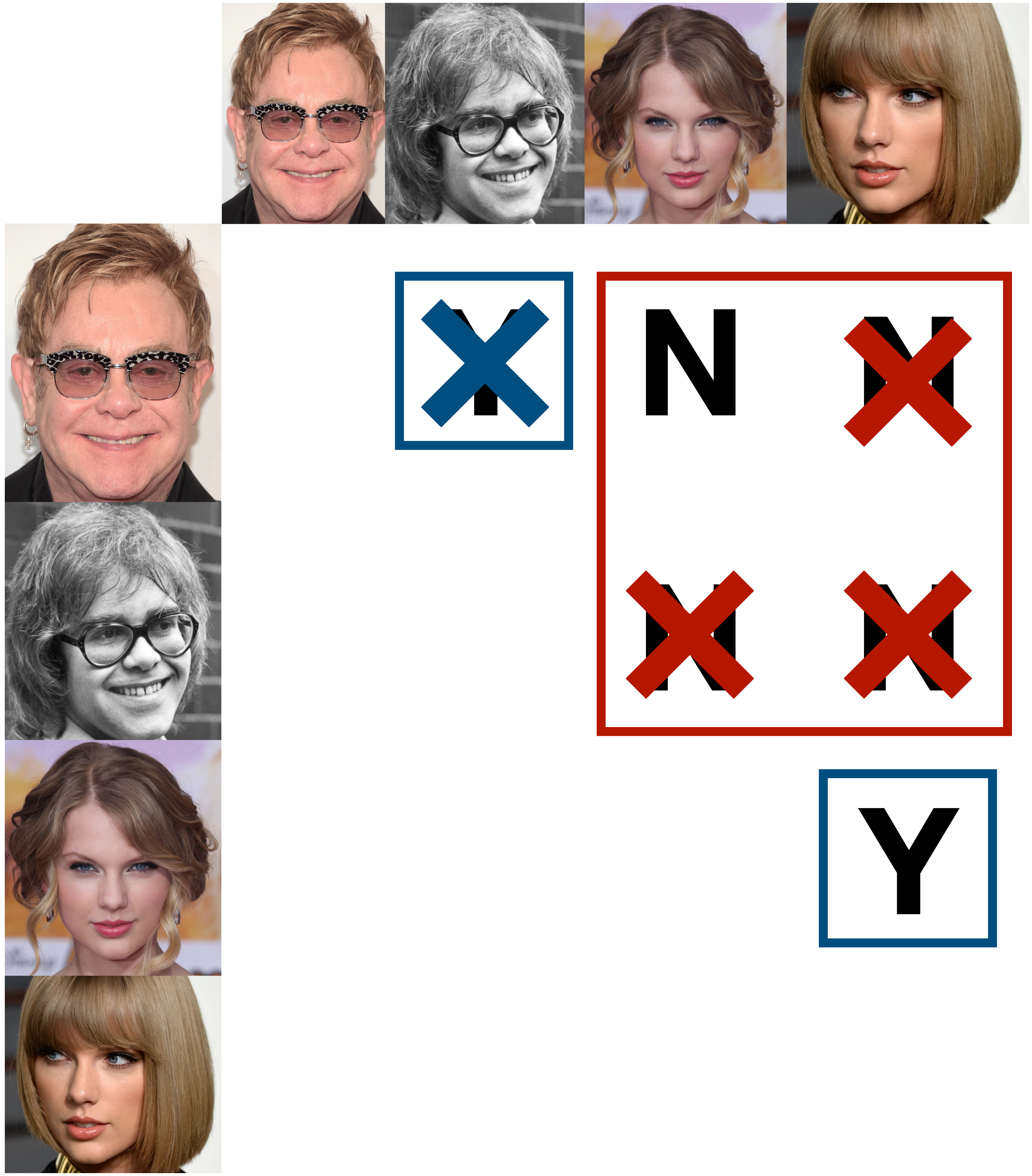
Scientist

Sure! Let me generate the predictions

Here are the results!
False Accept Rate (FAR) = 75%
False Reject Rate (FRR) = 50%

TPM

Oh no this is horrible!!!!!



Why do we need uncertainty?

Scientist

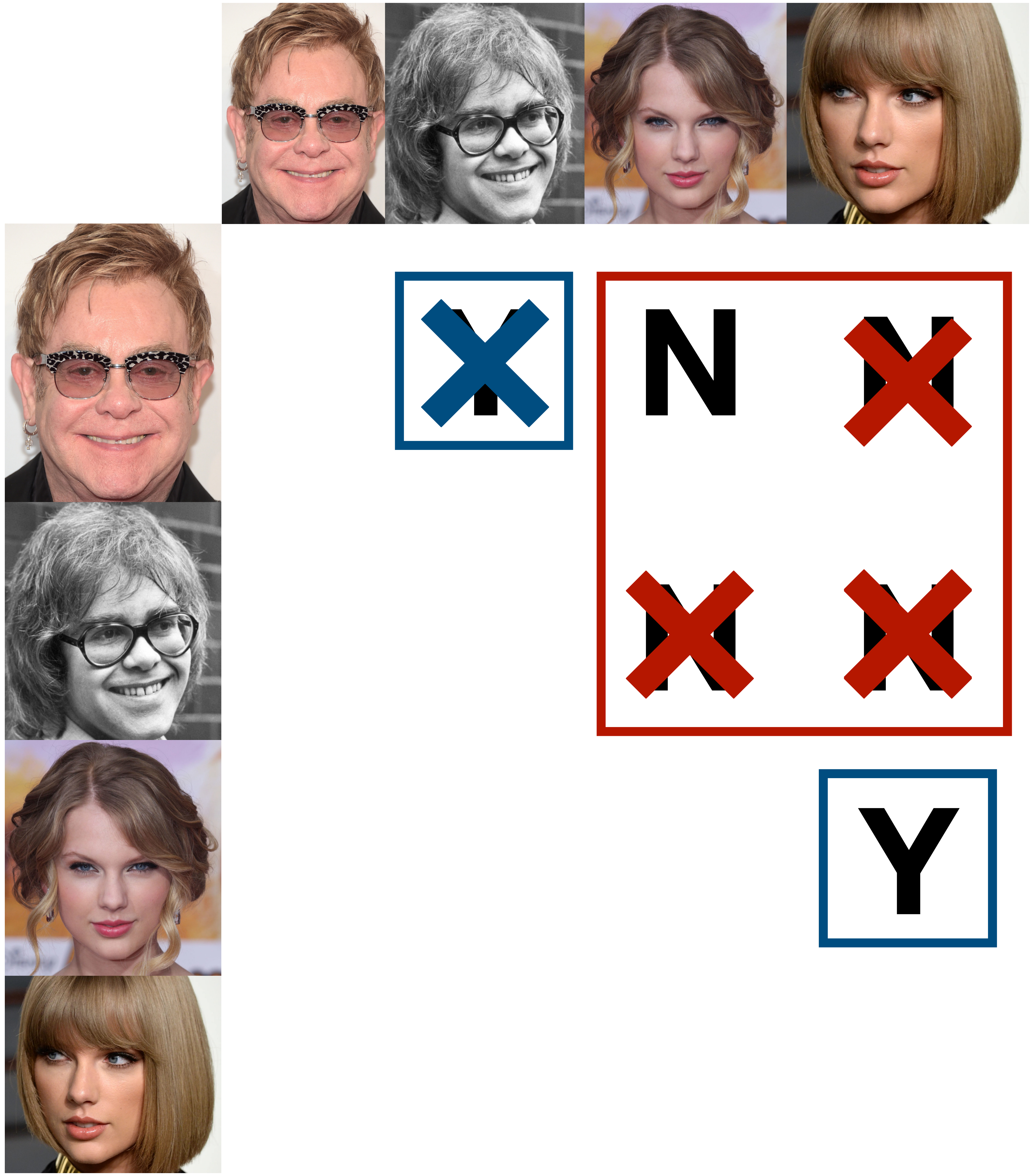
Oops I forgot the 95% confidence intervals!
False Accept Rate (FAR) =75% (10%, 80%)
False Reject Rate (FRR) = 50% (10%, 60%)

So much uncertainty…

TPM

What a relief…!

We need higher standards! We'll make reporting uncertainty estimates mandatory from now on.



How do we construct confidence intervals in 1:1 matching tasks?

Commonly used approaches

Wald and (naive) Wilson intervals based on the Normal approximation of the maximum likelihood estimator

Assumptions:

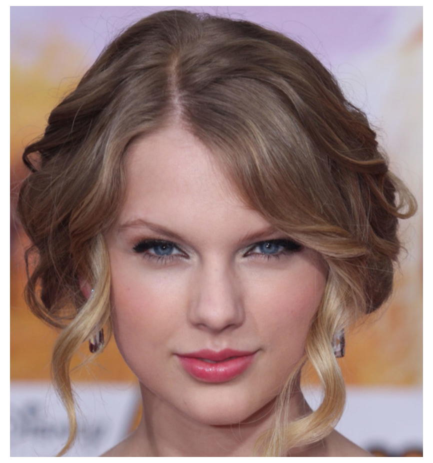
- Independent data
- Identically distributed data
- Finite mean and variance
- Large sample size

Standard central limit theorem assumptions do not hold in the context of 1:1 matching tasks

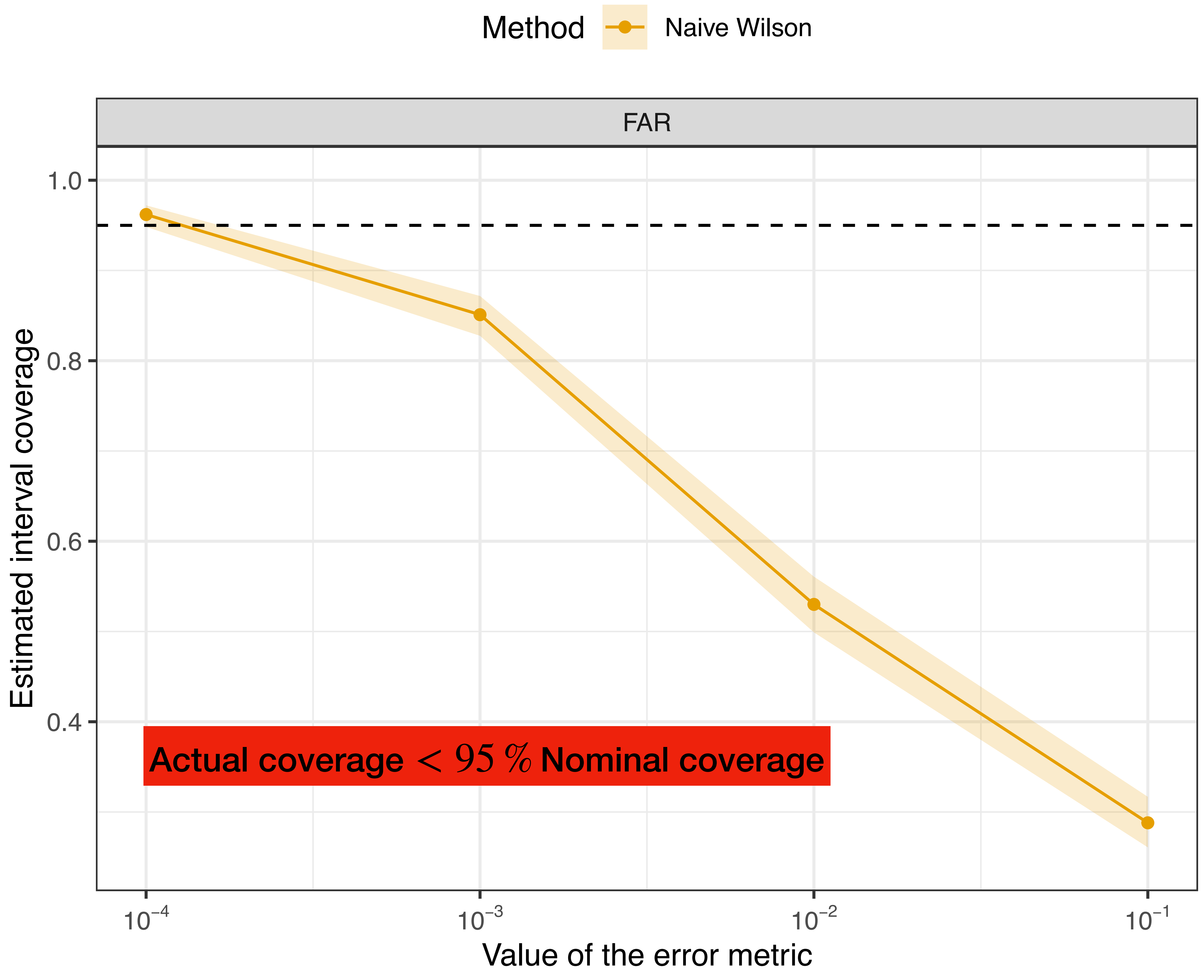
Assumptions:

- ~~Independent data~~
- Identically distributed data
- Finite mean and variance
- Large sample size

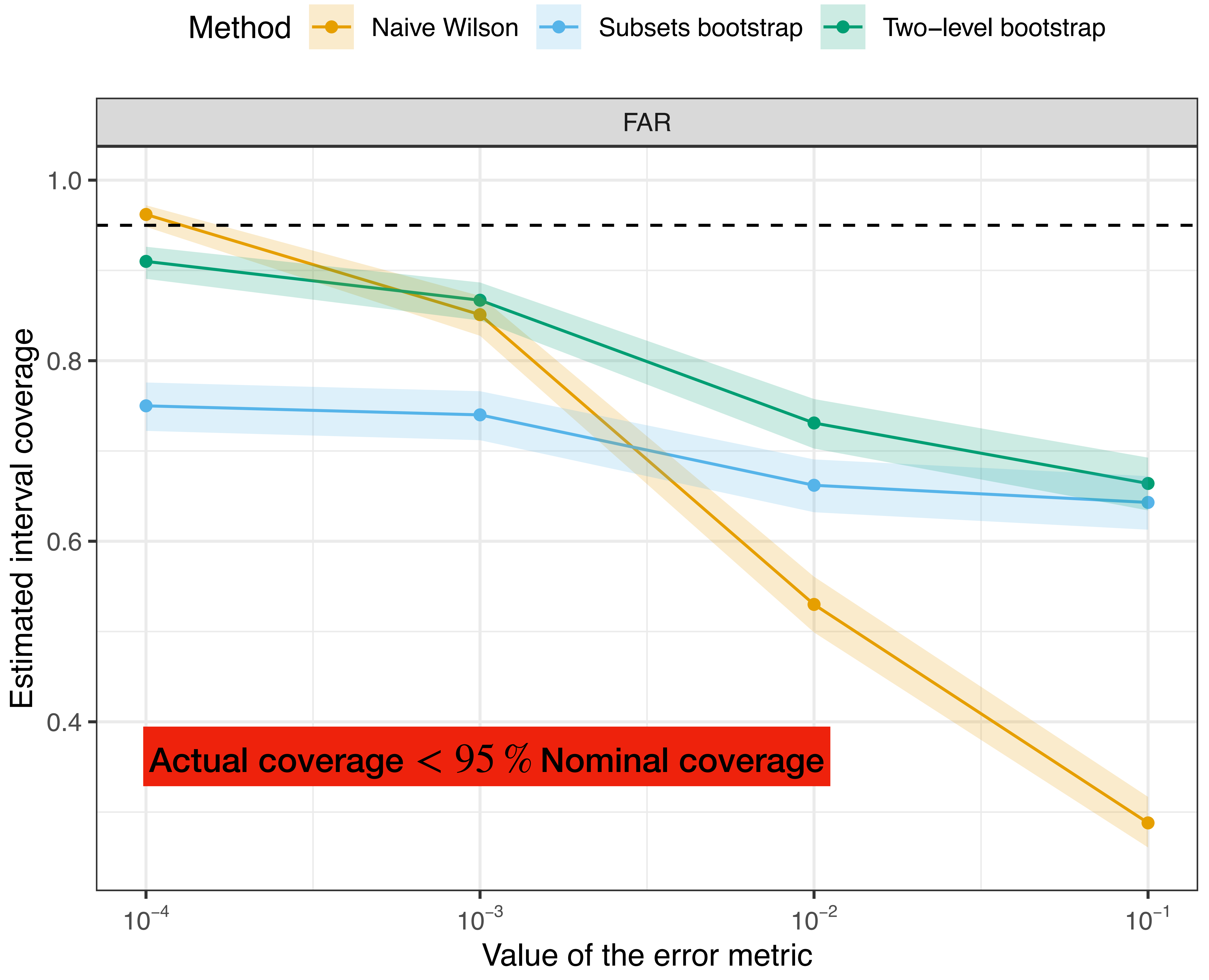
vs.



Naive Wilson intervals for the FAR are too narrow

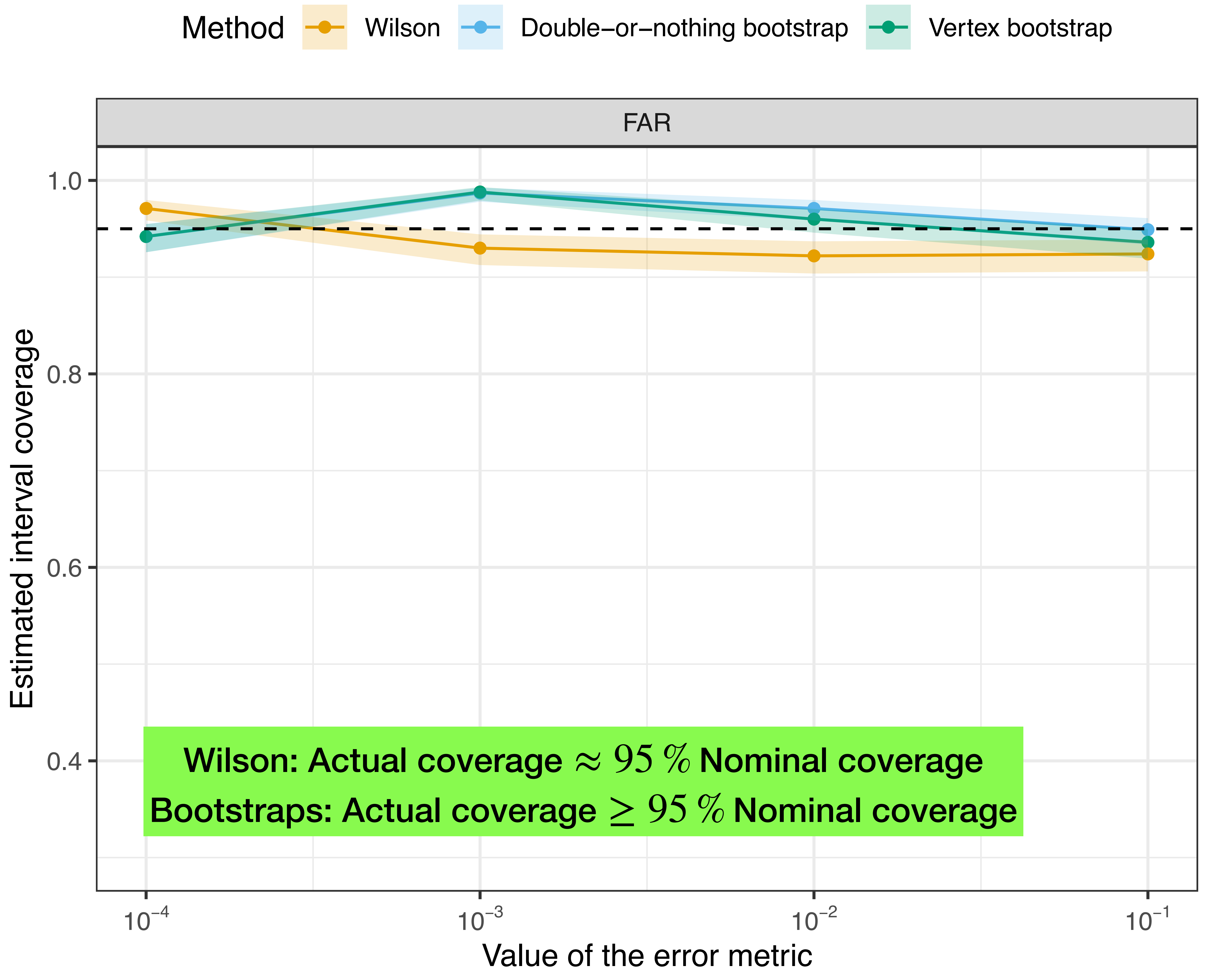


Bootstrap methods used in Facial Recognition produce FAR intervals that are too narrow



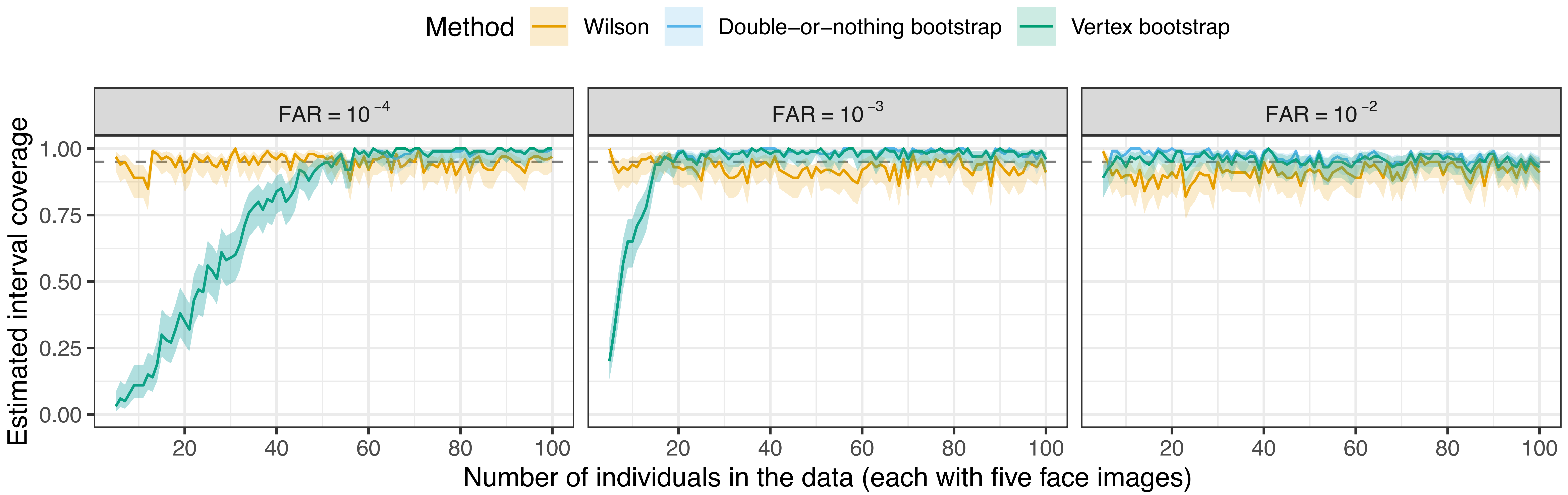
Bolle, Ruud M., Nalini K. Ratha, and Sharath Pankanti. "Error analysis of pattern recognition systems—the subsets bootstrap." *Computer Vision and Image Understanding* 93.1 (2004): 1-33.
Poh, Norman, Alvin Martin, and Samy Bengio. "Performance generalization in biometric authentication using joint user-specific and sample bootstraps." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.3 (2007): 492-498.

(Improved) Wilson, double-or-nothing and vertex bootstrap produce FAR intervals that mostly achieve nominal coverage



Snijders, Tom AB, and Stephen P. Borgatti. "Non-parametric standard errors and tests for network statistics." *Connections* 22.2 (1999): 161-170.
Owen, Art B., and Dean Eckles. "Bootstrapping data arrays of arbitrary order." *Annals of Applied Statistics* (2012): 895-927.

Bootstrap intervals are inadequate when error rates are too small



Wilson: Actual coverage $\approx 95\,\%$ Nominal coverage
Bootstraps: Actual coverage $\geq 95\,\%$ Nominal coverage* when FAR $\gg 0$

How do we construct intervals that achieve nominal coverage for FAR in 1:1 matching tasks?

Constructing Wilson intervals

Naive and (correct) Wilson intervals for FAR are given by

$$\frac{\widehat{\text{FAR}}\,\hat{N}^*_{FAR} + \frac{1}{2}z^2_{1-\alpha/2}}{\hat{N}^*_{FAR} + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}\sqrt{\hat{N}^*_{FAR}}}{\hat{N}^*_{FAR} + z^2_{1-\alpha/2}} \sqrt{\widehat{\text{FAR}}\,(1-\widehat{\text{FAR}}) + z^2_{1-\alpha/2}/(4\hat{N}^*_{FAR})}$$

where $\hat{N}^*_{FAR} = (\widehat{\text{FAR}}\,(1-\widehat{\text{FAR}}))/\text{Var}(\widehat{\text{FAR}})$

(Correct) Wilson intervals

$$\text{Var}(\widehat{\text{FAR}}) = \frac{1}{3}\left[\text{Var}(\widehat{\text{FAR}}_{12}) + \text{Var}(\widehat{\text{FAR}}_{13}) + \text{Var}(\widehat{\text{FAR}}_{23})\right]$$ } Naive Wilson

$$+\frac{2}{3}\left[\text{Cov}(\widehat{\text{FAR}}_{12}, \widehat{\text{FAR}}_{13}) + \text{Cov}(\widehat{\text{FAR}}_{12}, \widehat{\text{FAR}}_{23}) + \text{Cov}(\widehat{\text{FAR}}_{12}, \widehat{\text{FAR}}_{23})\right]$$

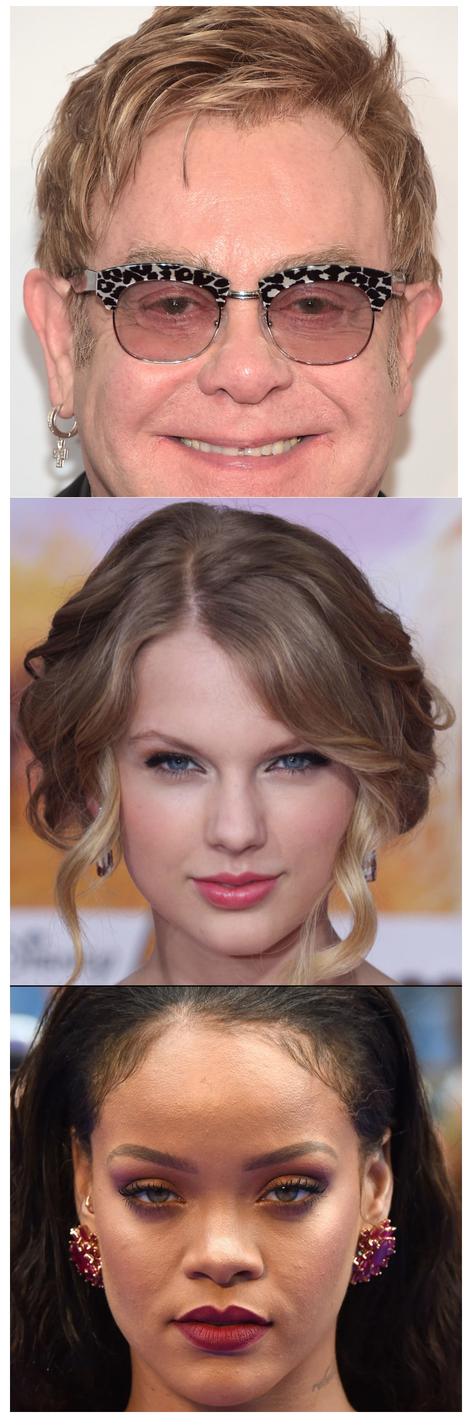
Constructing double-or-nothing bootstrap intervals

Percentile bootstrap recipe for $1-\alpha$ FAR intervals

For repetition $b = 1,\ldots,B$:

- sample $w_i \sim$ Uniform$\{0,2\}$ for $i = 1,\ldots,G$ and compute $\widehat{\text{FAR}}^{b} = \sum_{i \neq j} w_i w_j \widehat{\text{FAR}}_{ij} / \sum_{i \neq j} w_i w_j$

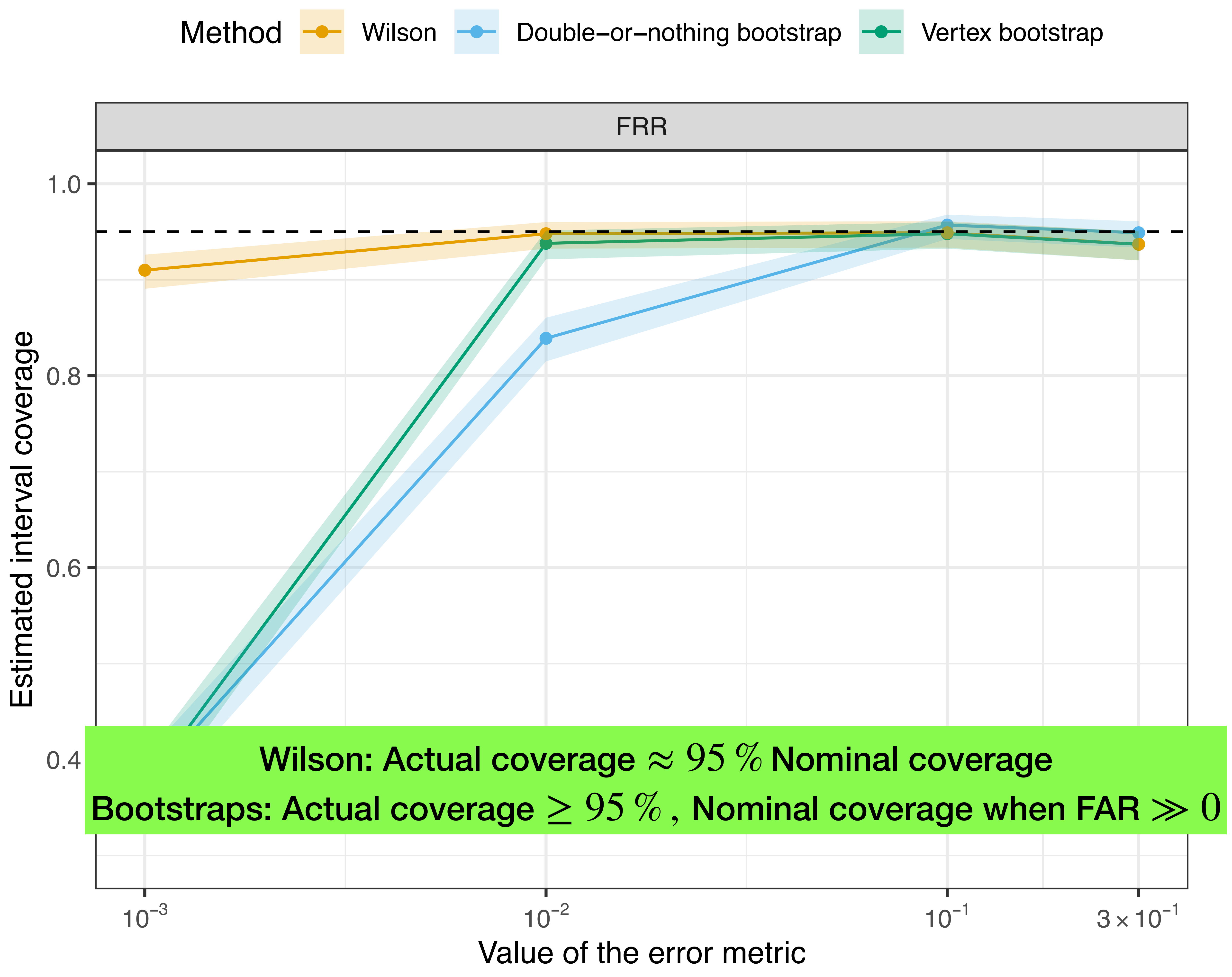
Then take the $(\alpha/2, 1-\alpha/2)$ quantiles of the $\widehat{\text{FAR}}_b$ estimates



	b=1	b=2	b=B
	$w_1 = 2$	$w_1 = 0$	$w_1 = 2$
	$w_2 = 2$	$w_2 = 2$	$w_2 = 2$
	$w_3 = 0$	$w_3 = 2$	$w_3 = 2$
	$\widehat{\text{FAR}}^{1} = \widehat{\text{FAR}}_{12}$	$\widehat{\text{FAR}}^{2} = \widehat{\text{FAR}}_{23}$	$\widehat{\text{FAR}}^{B} = \widehat{\text{FAR}}$

How do we construct intervals that achieve nominal coverage for FRR and ROC in 1:1 matching tasks?

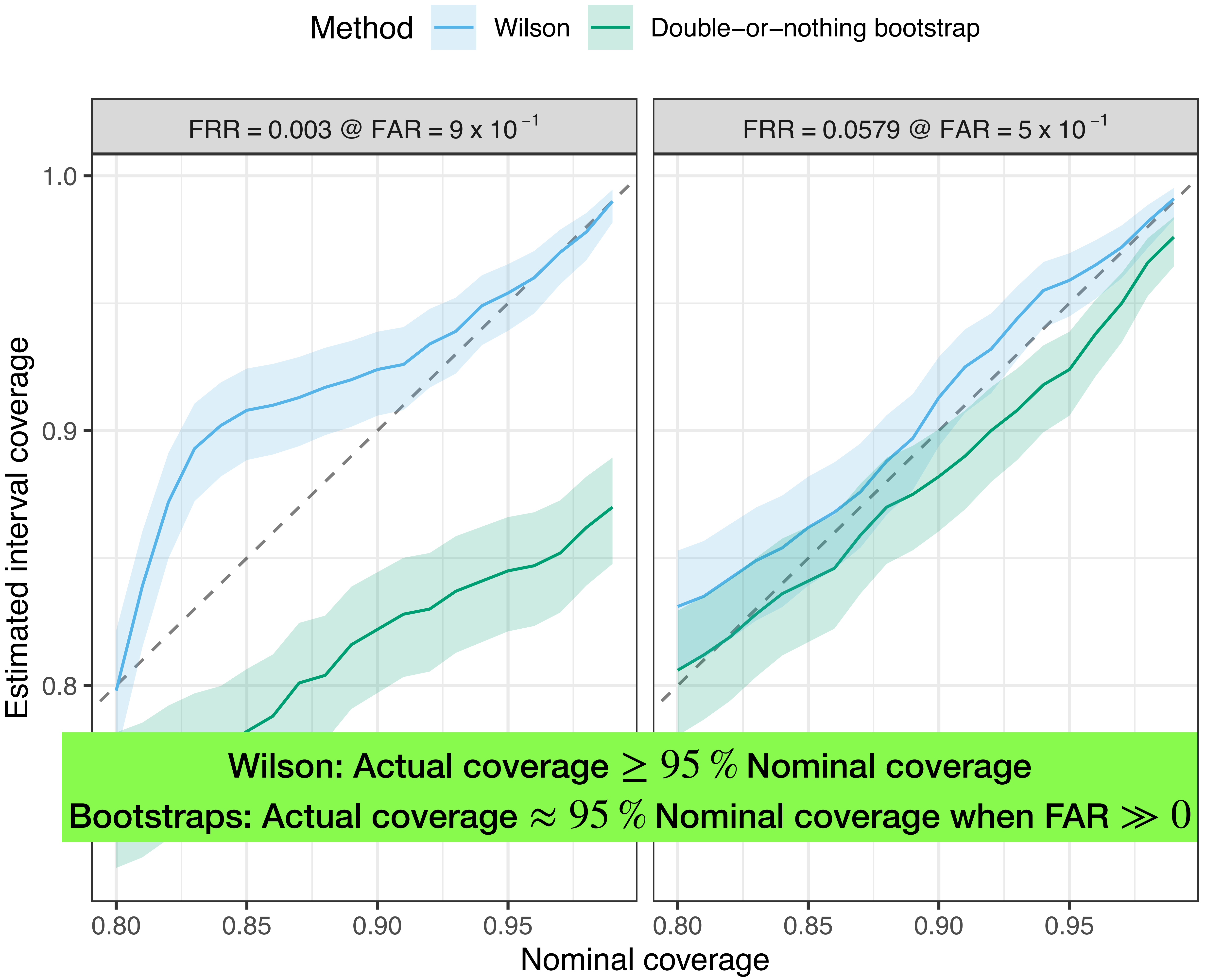
Wilson, double-or-nothing, and vertex bootstrap produce FRR intervals that mostly achieve nominal coverage



Wilson: Actual coverage $\approx 95\,\%$ Nominal coverage

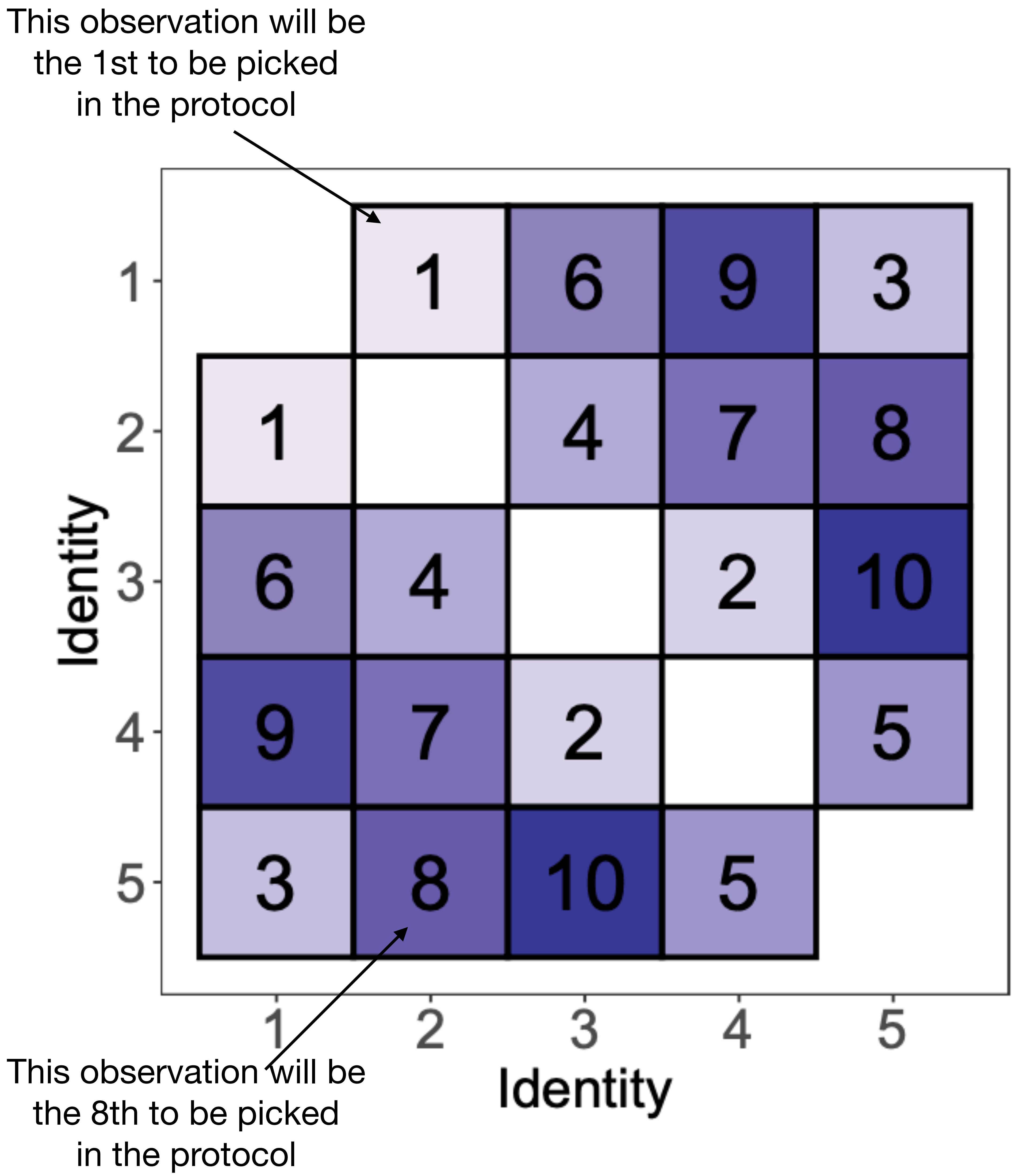
Bootstraps: Actual coverage $\geq 95\,\%$, Nominal coverage when FAR $\gg 0$

Wilson-based intervals for the ROC are conservative, while double-or-nothing bootstrap intervals achieve nominal coverage for larger error metrics



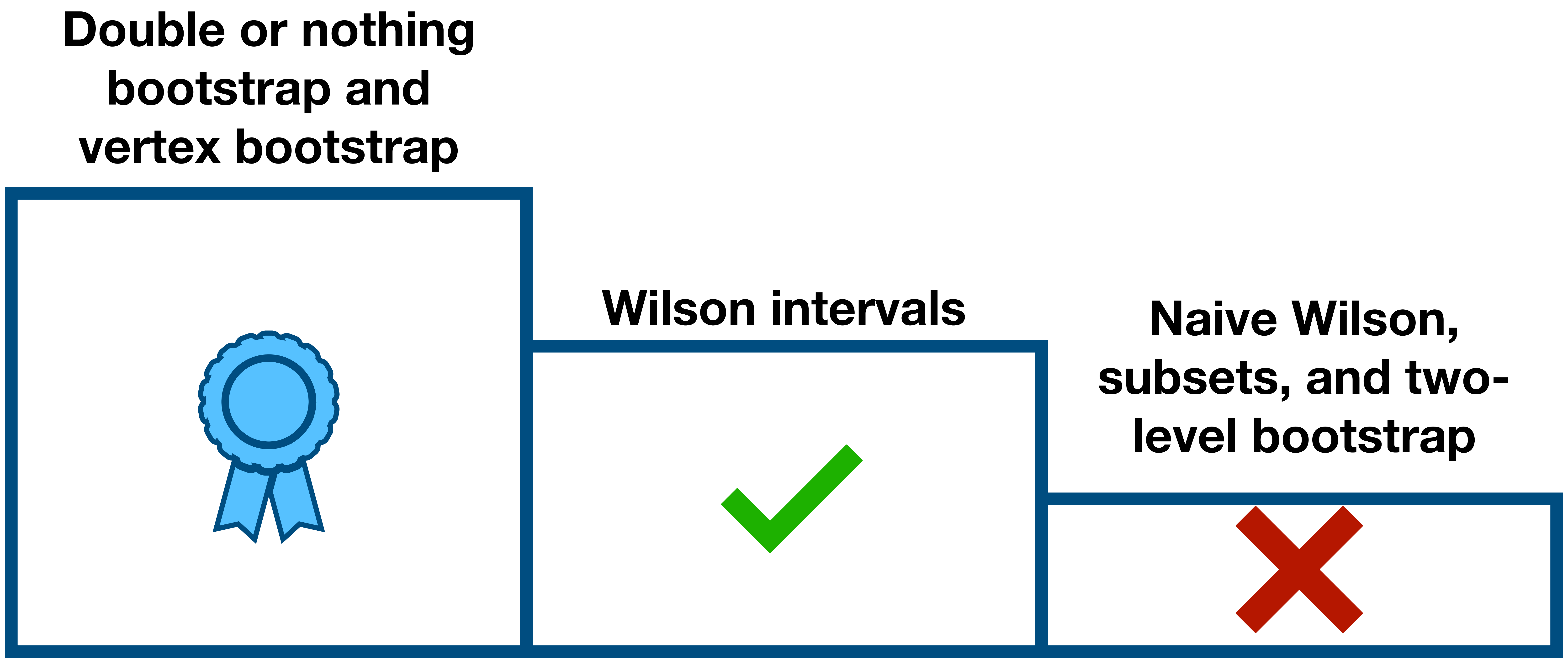
Conti, Jean-Rémy, and Stéphan Clémençon. "Assessing Performance and Fairness Metrics in Face Recognition-Bootstrap Methods." *arXiv preprint arXiv:2211.07245* (2022).

Massive dataset and constrained resources? To minimize the variance of FAR and FRR estimates, protocols should consider independent observations



Takeaway for FAR/FRR intervals

Takeaway for ROC/AUC intervals



Code for reviewed methods: github.com/awslabs/cis-matching-tasks

README.md

Confidence Intervals for Error Rates in Matching Tasks

paper arXiv license Apache-2.0

This repository houses the code to construct confidence intervals for performance metrics in tasks such as 1:1 face and speaker verification. See our paper for a description of the methods.

Tutorial on MORPH

In this tutorial, we will assess the performance of a facial recognition system in a 1:1 face verification task on the MORPH dataset. We have obtained the embeddings generated by the system for the images in the data and stored them in a dictionary `df[identity name][image name] = embedding`. Below we load the dictionary.

```
import json
from utils import *

df_main = json.load(open('../data/morph/embeddings.json', 'r'))
len(df_main)  # number of identities in the data
```

63548

We analyze the system performance in two settings:

- *small datasets*: We assess the system performance on all pairwise comparisons between the images in the data.
- *large datasets*: We compute the system performance on a subset of all pairwise comparisons between the images in the data.

General tutorials: github.com/awslabs/cis-matching-tasks

README.md

Confidence Intervals for ML Performance Metrics

Assessing the uncertainty of performance metrics of machine learning (ML) models is critical. By quantifying the uncertainty in our performance estimates, we can better gauge the effectiveness of ML models on real-world customer data. This repository provides brief tutorials on how to construct confidence intervals for the performance metrics of machine learning models such as binary accurracy, log loss, F1 score, AUC, MSE, and others

Classification Tasks

This notebook covers the construction of confidence intervals and hypothesis testing for metrics typically employed to evaluate the performance of ML models in binary classification tasks. These methods are model agnostic, in that they apply to any model that outputs a confidence score for each prediction.

Problem Setup

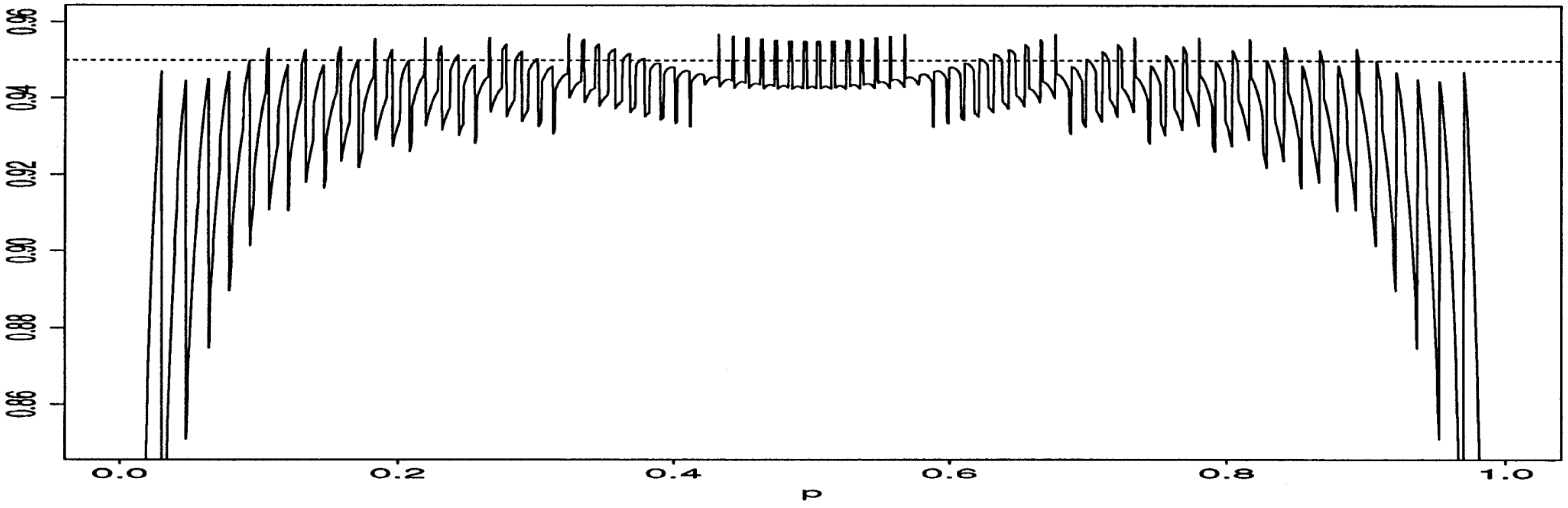
Overview

We have a dataset with n observations (X_i, Y_i), where each pair is independently and identically distributed (IID) from a probability distribution P. Here, X_i is a vector of features, and Y_i is a binary outcome. The outcome Y is defined as:

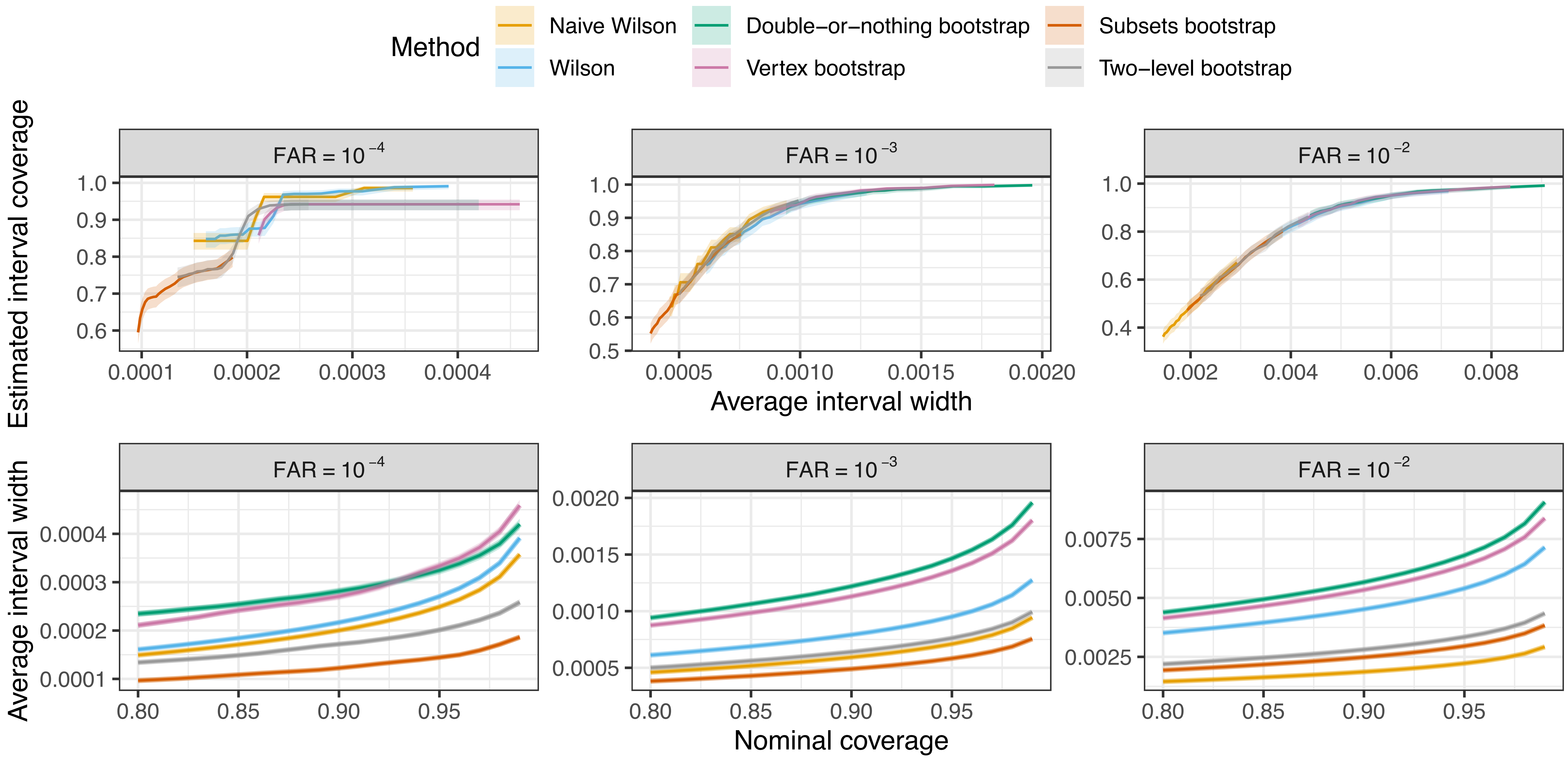
$$Y = \begin{cases} 1 & \text{with probability } \mathbb{E}_P[Y|X], \\ 0 & \text{with probability } 1 - \mathbb{E}_P[Y|X] \end{cases}$$

Thank you!

Wald intervals fail in 1:1 matching tasks when error rates are low



Width vs. coverage



Width of various algorithms