# Random Forests for Hypothesis Testing

## Development and Application to Cancer Detection

**Sambit Panda**

*December 2, 2024*

*Department of Biomedical Engineering, Johns Hopkins University*

# I would like to thank my family, …

# my future wife, …

# my friends and lab mates, …

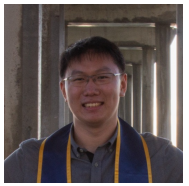
# and my coauthors and thesis committee!

**NeuroData Lab**


Cencheng Shen


Adam Li


Yuxin Bai


Hao Xu


Suki Ogihara


Ronan Perry

**Ludwig Cancer Institute**


Sam Curtis


Chris Douville


Nick Papadopolous


Bert Vogelstein

**Thesis Committee**


Joshua Vogelstein


Carey Priebe

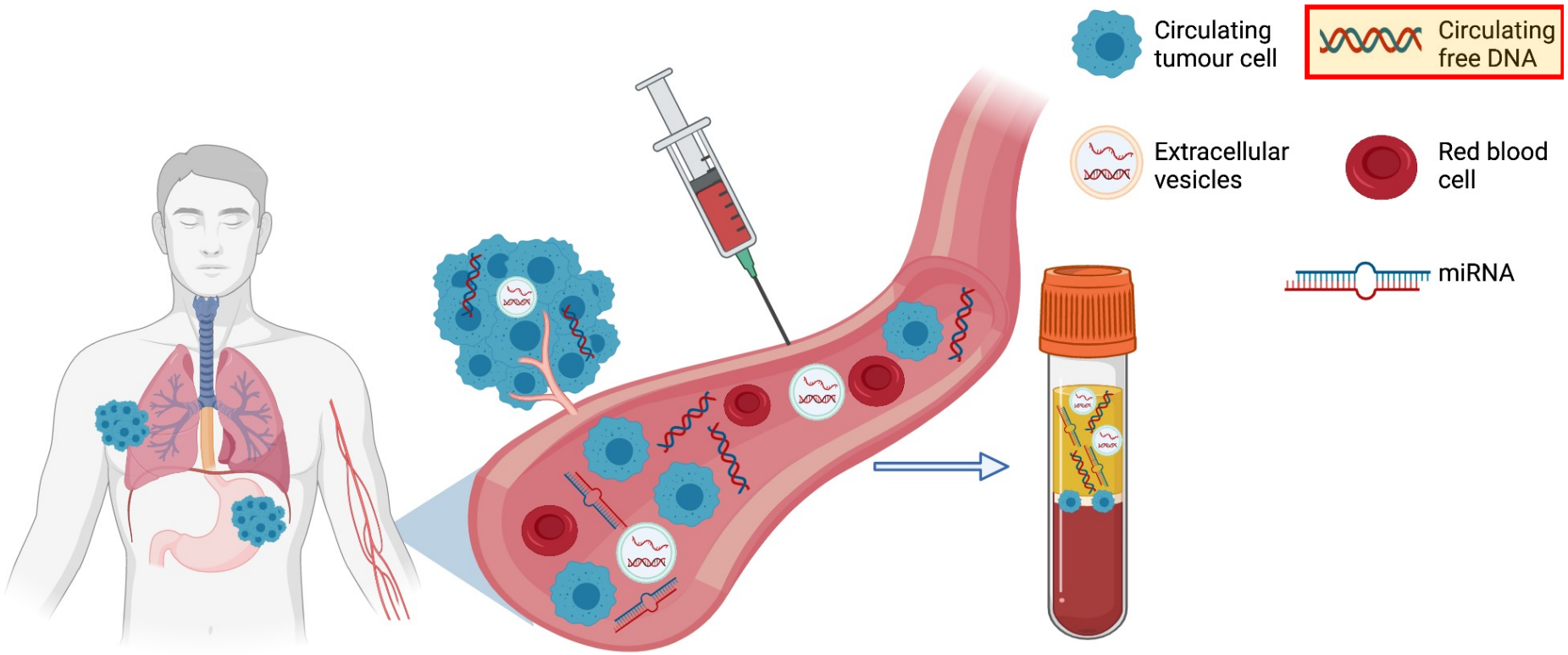

Eliza O'Reilly

# Outline

1. Motivation
2. Aim 1: K-Sample Testing via Dependence Measures
3. Aim 2: KMERF Algorithm
4. Aim 3: MIGHT and CoMIGHT Algorithms
5. Conclusion

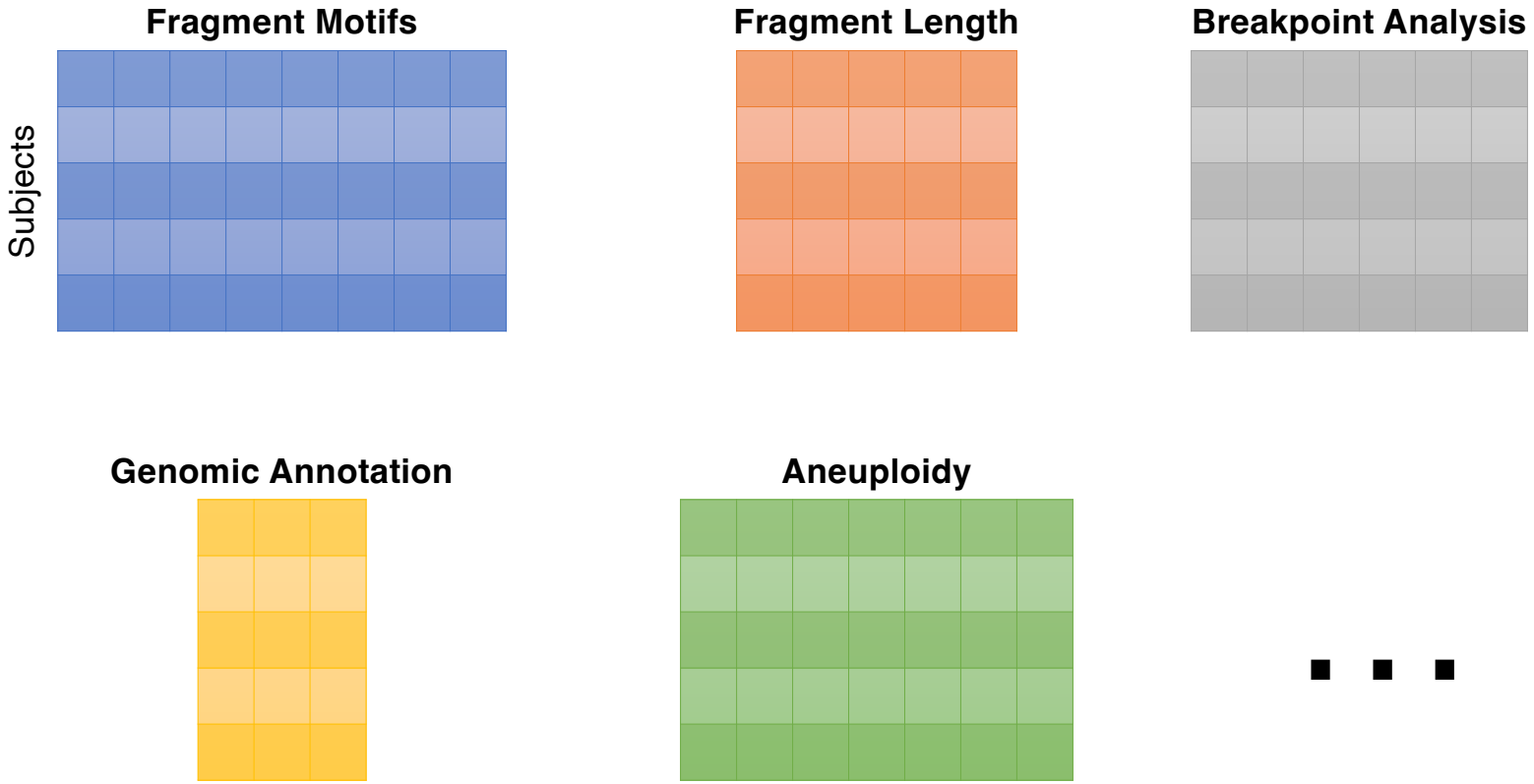
# Motivation

# How can we detect cancer early?

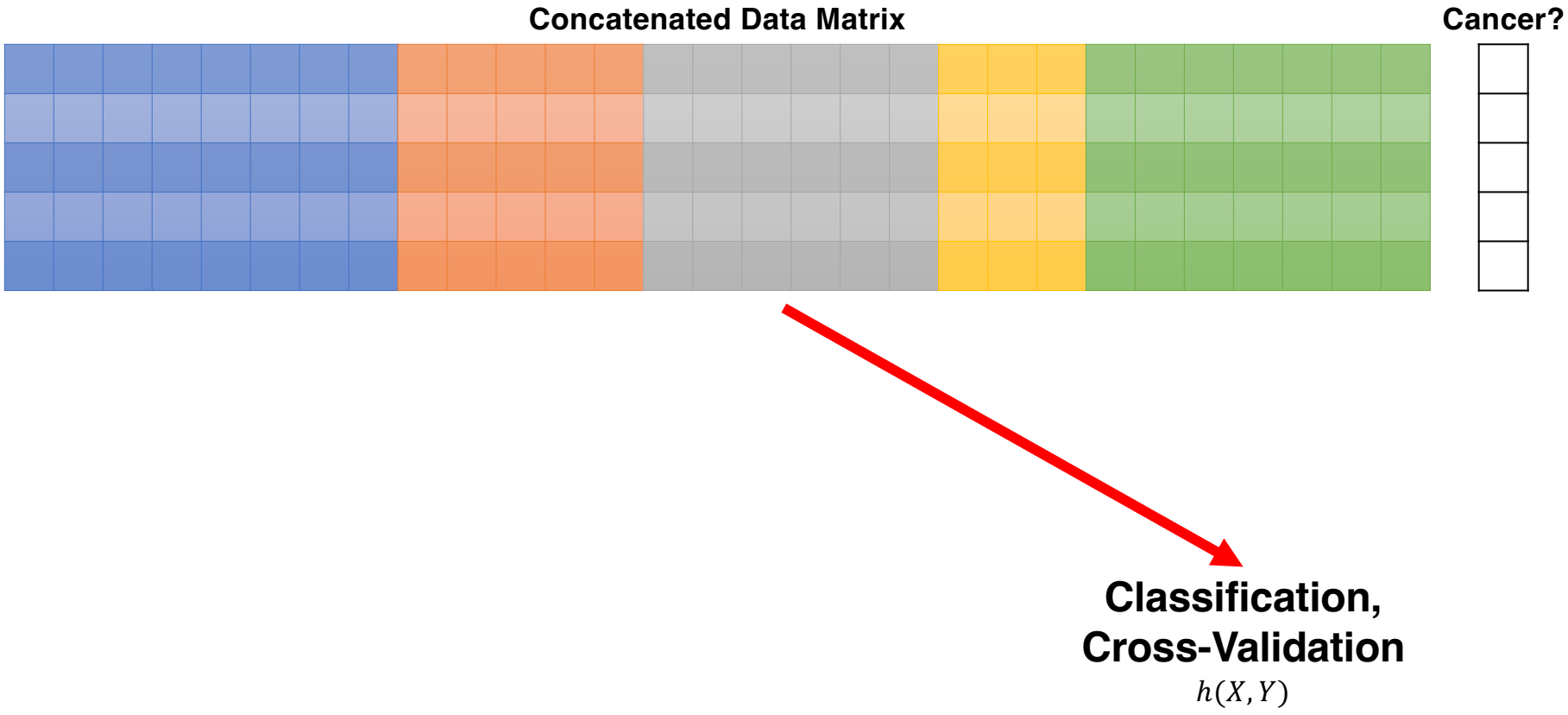



BioRender; R. L. Siegel, A. N. Giaquinto, A. Jemal. CA (2024); S. Connal et al. J. Transl. Med. (2023)

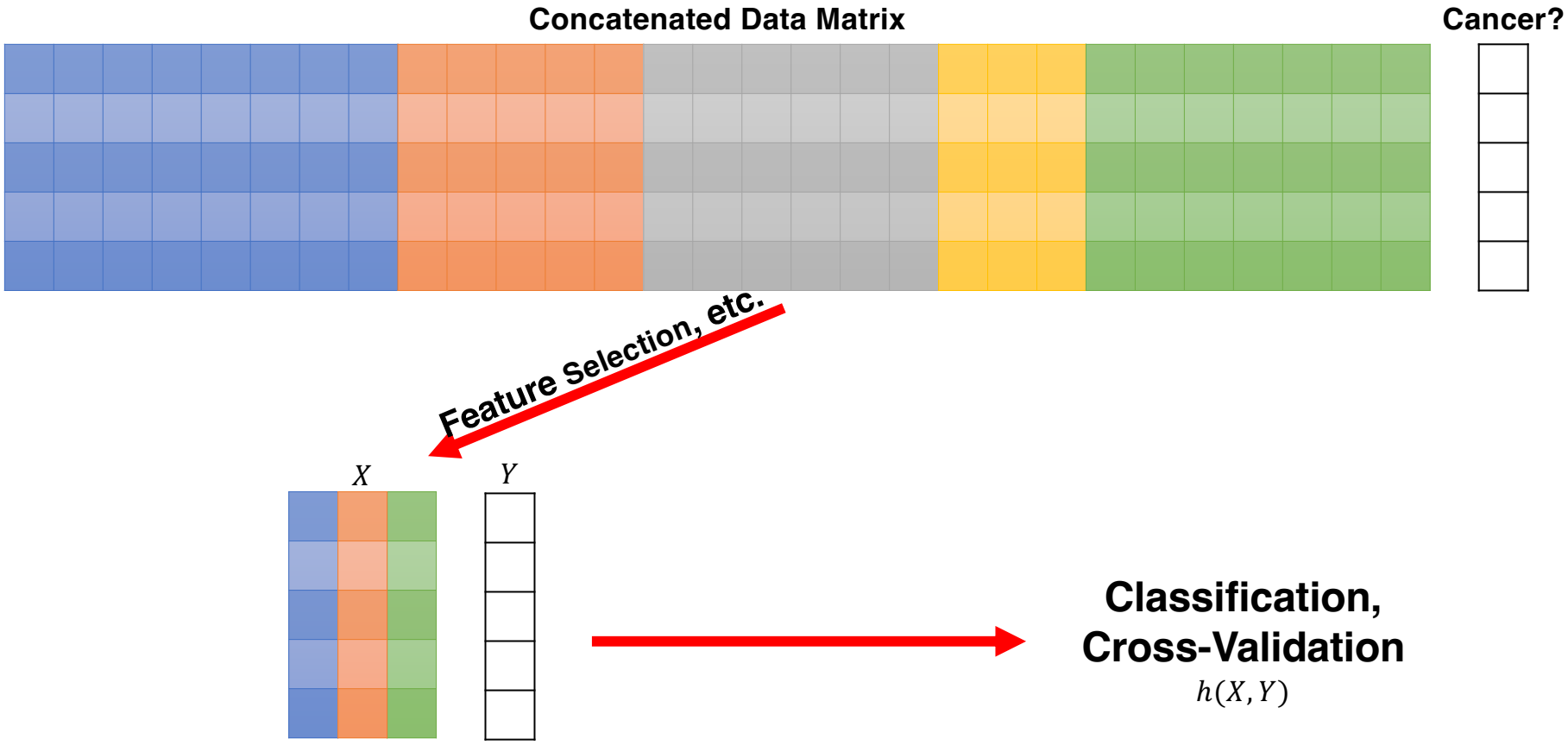
# Liquid biopsy generates lots of views



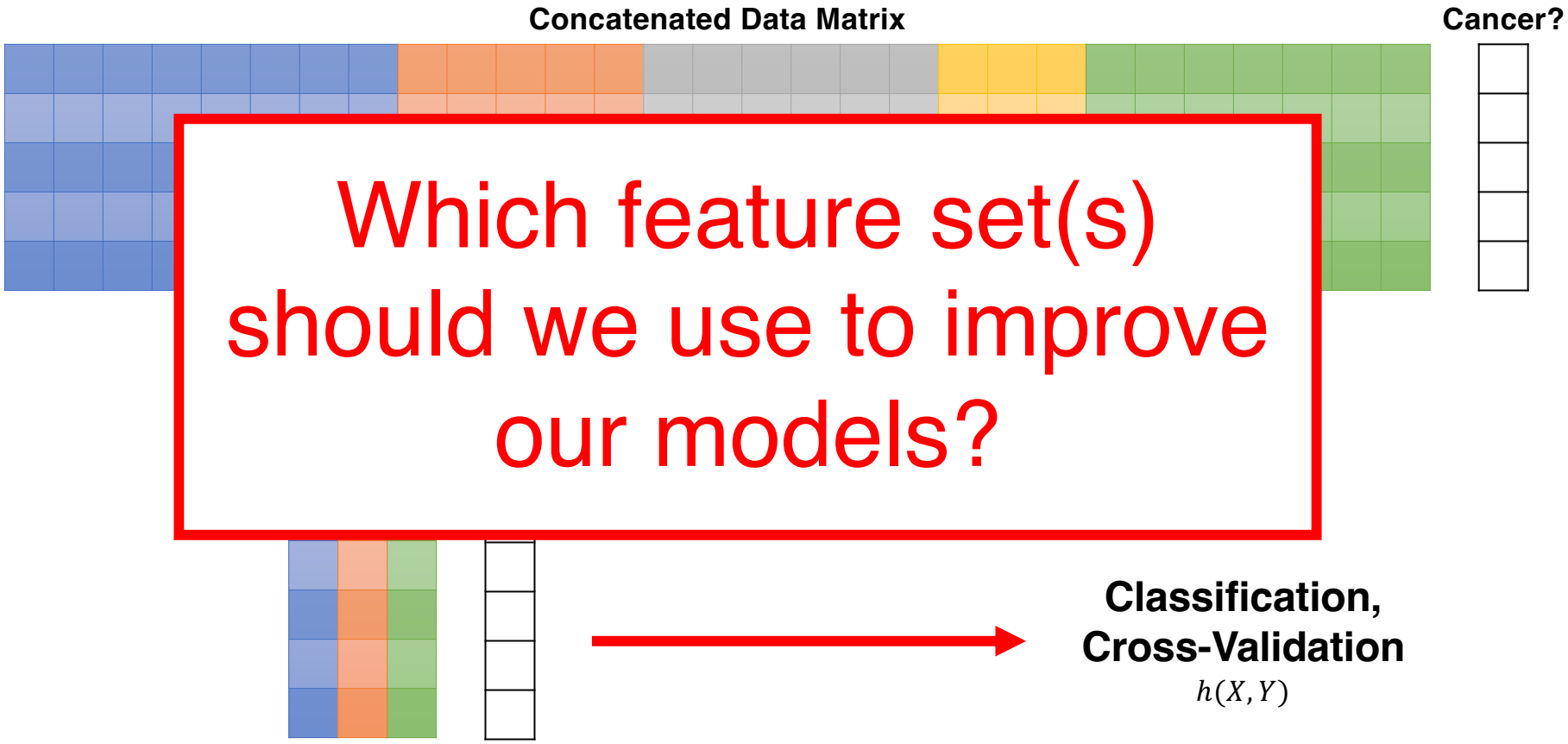
# How predictive models are currently made



# How predictive models are currently made



# How predictive models are currently made



# Our approach

Devise algorithms that find…

- which sets of features are informative
- which of the informative sets are most predictive
- whether combinations of sets of features are more predictive than single sets

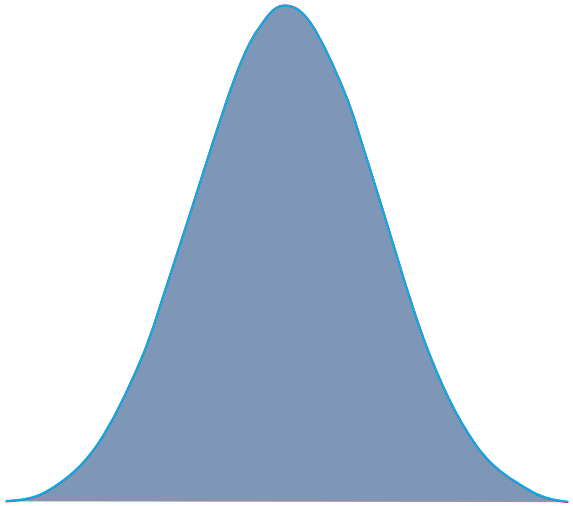
# Aim 1: K-Sample Testing via Dependence Measures

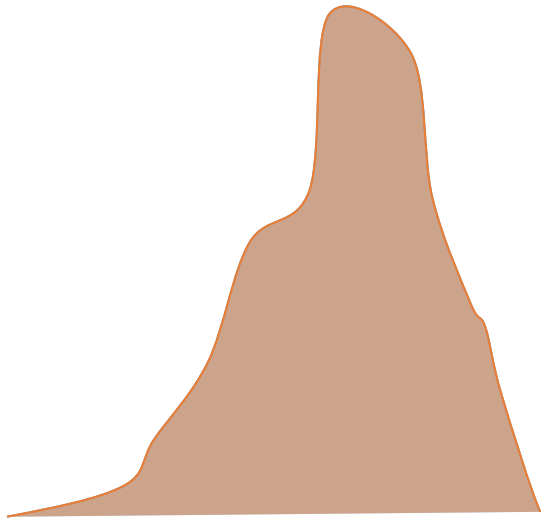
S. Panda et al. Stat & Prob Letters (2025)

# Question: Are all the groups the same or not?

$$H_0 : \mu_1 = \cdots = \mu_k$$
$$H_A : \exists\, j \neq j' \;\; \mu_j \neq \mu_{j'}$$

$$H_0 : F_1 = \cdots = F_k$$
$$H_A : \exists\, j \neq j' \;\; F_j \neq F_{j'}$$





# There are a few two/k-sample tests

**Parametric Tests**

- T-tests
- ANOVA
- MANOVA
- etc.

**Non-parametric Tests**

- Mann Whitney U
- Wilcoxon Rank Sum
- HHG
- Energy
- MMD
- DISCO
- etc.

# There are even fewer k-sample tests

**Parametric Tests**

- T-tests
- **ANOVA**
- **MANOVA**
- etc.

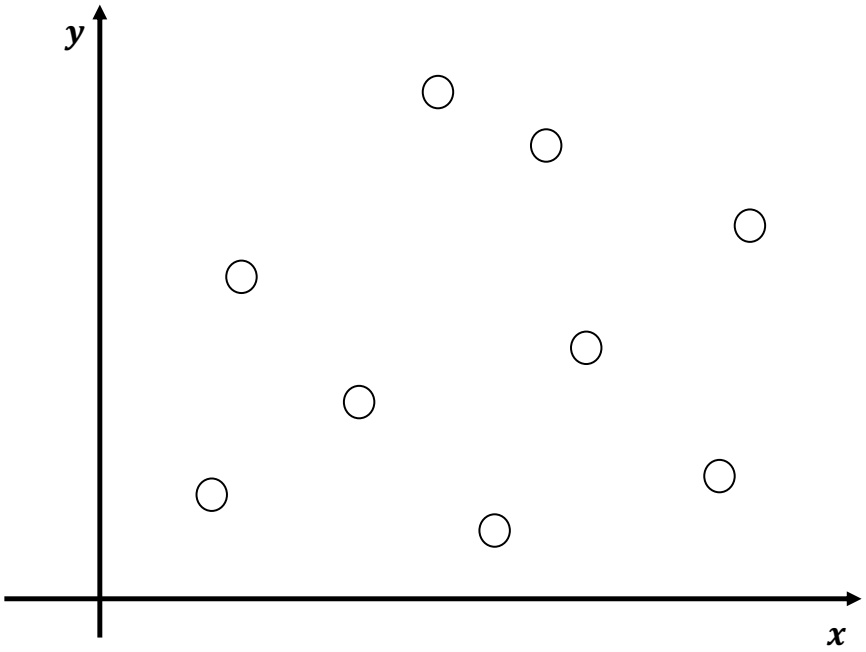
**Non-parametric Tests**

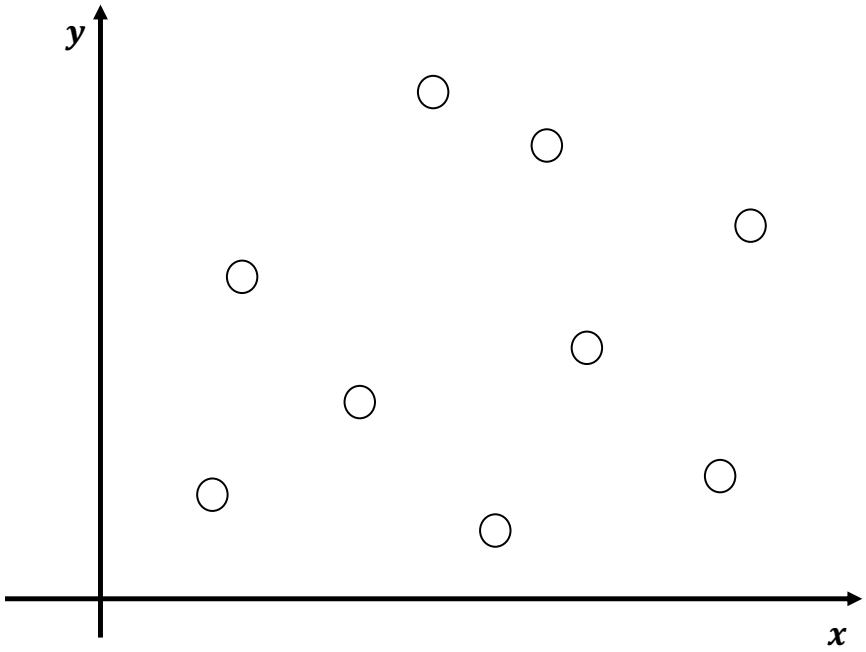
- Mann Whitney U
- Wilcoxon Rank Sum
- HHG
- Energy
- MMD
- **DISCO**
- etc.

# Question: Is there a relationship?

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0$$

$$H_0 : F_{XY} = F_X F_Y$$
$$H_A : F_{XY} \neq F_X F_Y$$



# There are a lot more independence tests

**Parametric Tests**

- Pearson's Correlation
- Spearman Rank Correlation
- Kendall Tau
- RV
- CCA
- etc.

**Non-parametric Tests**

- HHG
- Dcorr
- Hsic
- MGC
- etc.

**Question:** Is there a relationship?

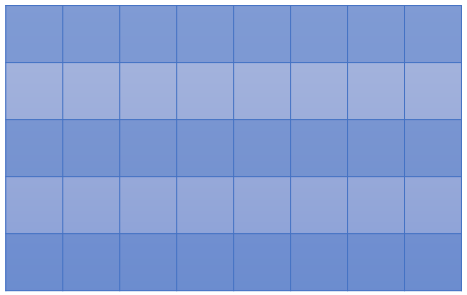
=

**Question:** Are they different?

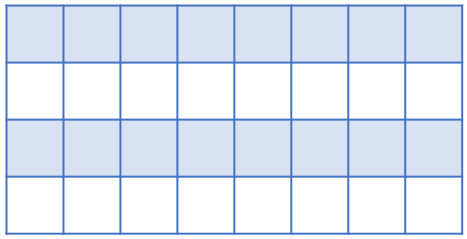
*(via a transformation of the data)*

# How the transformation works

$U_1$

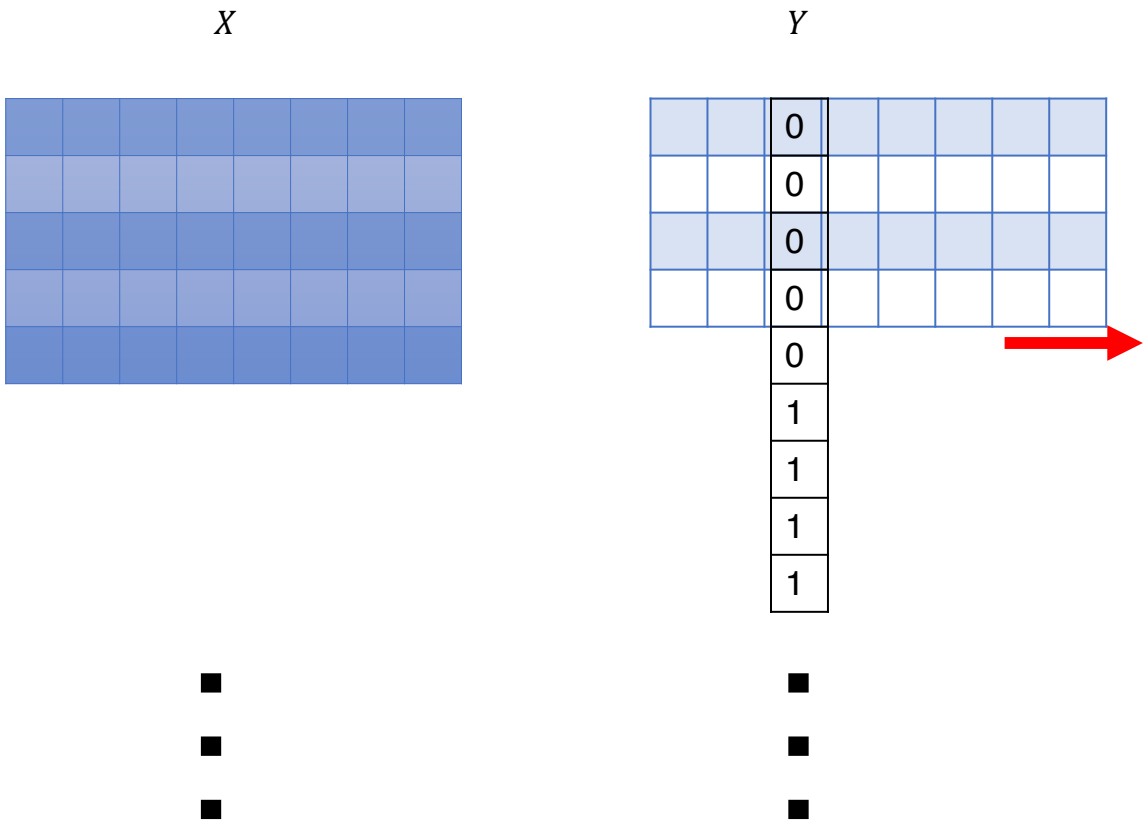


$U_2$

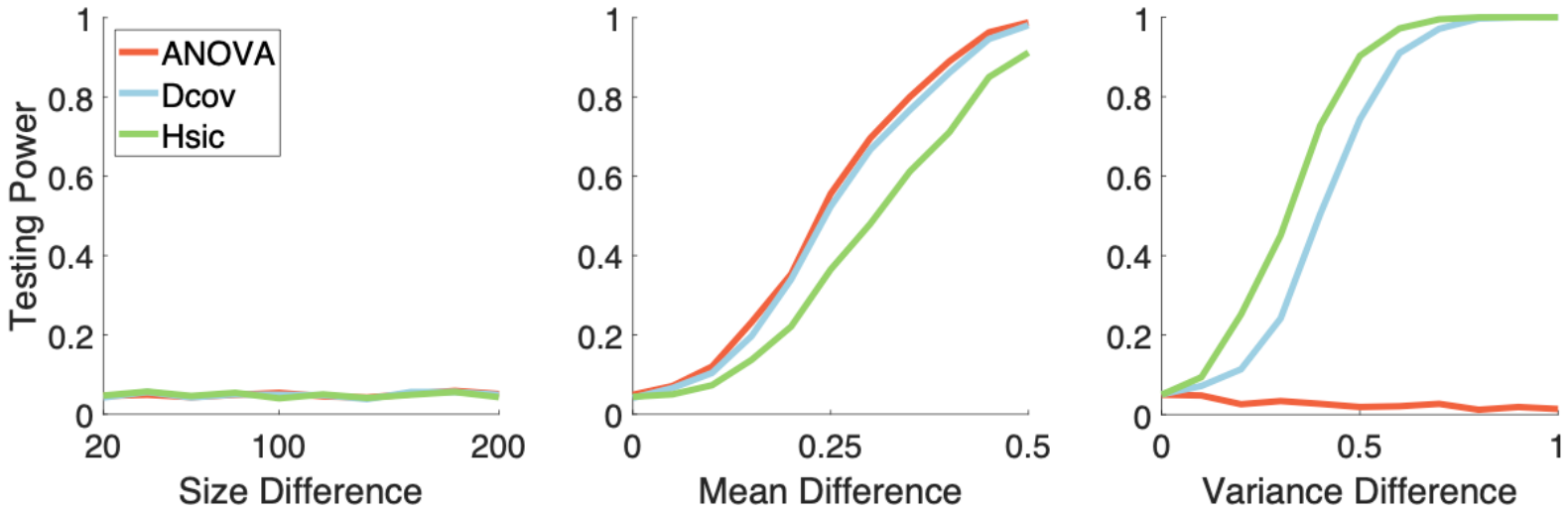


. . .

# How the transformation works



# Our framework performs better than ANOVA



# Can we use this framework to make better tests?

**Independence Tests**

- RV/CCA
- Dcorr/Hsic
- HHG
- MGC
- etc.

**Two/K-Sample Tests**

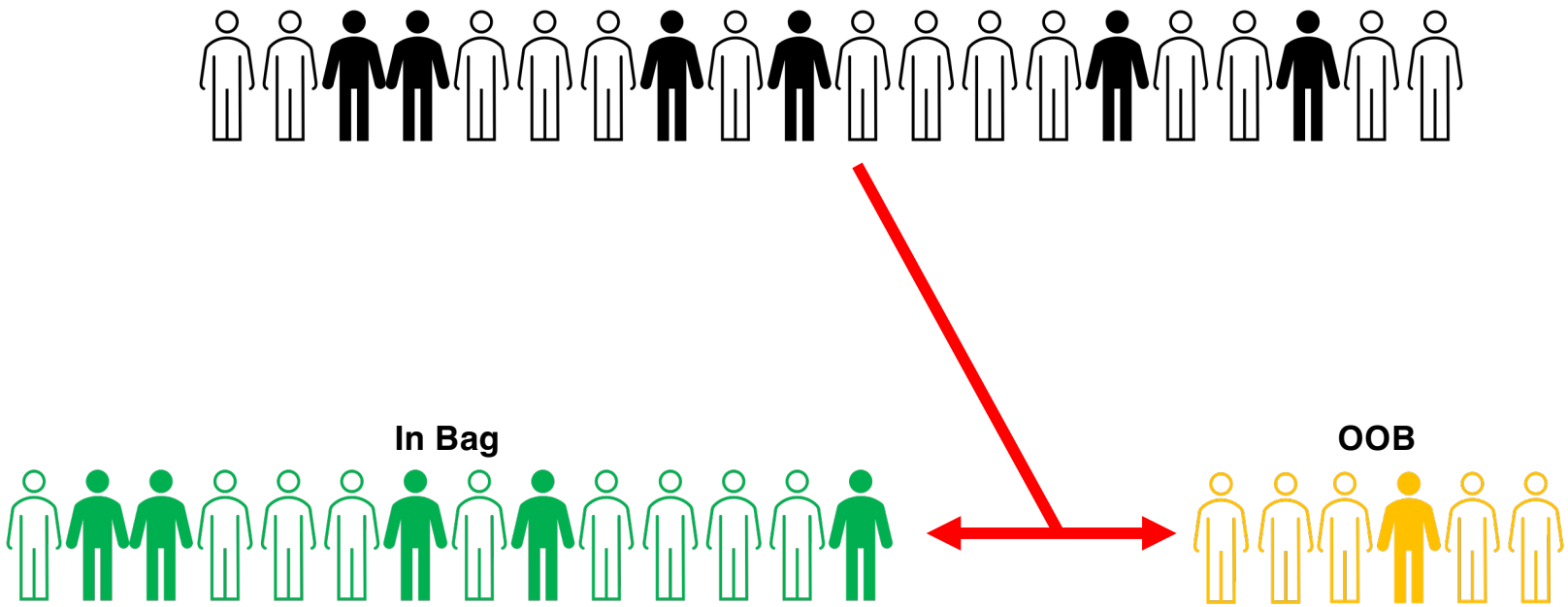
- Hotelling's $T^2$/MANOVA
- Energy/MMD
- HHG
- Our K-Sample papers
- etc.

- **Using a data adaptive kernel like random forest may improve finite sample testing power in high dimensions (or # of variables).**

# Aim 2: Kernel Mean Embedding Random Forest (KMERF)

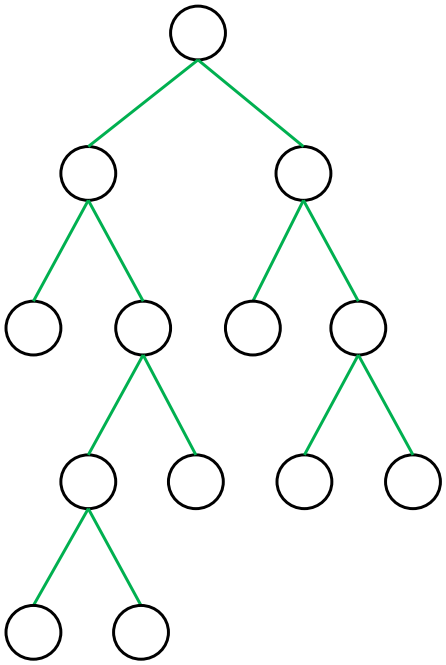
S. Panda et al. In Preparation ICML 2025.

# **Step 1**: Train a Random Forest (RF)




L. Breiman. Machine Learning (2001)

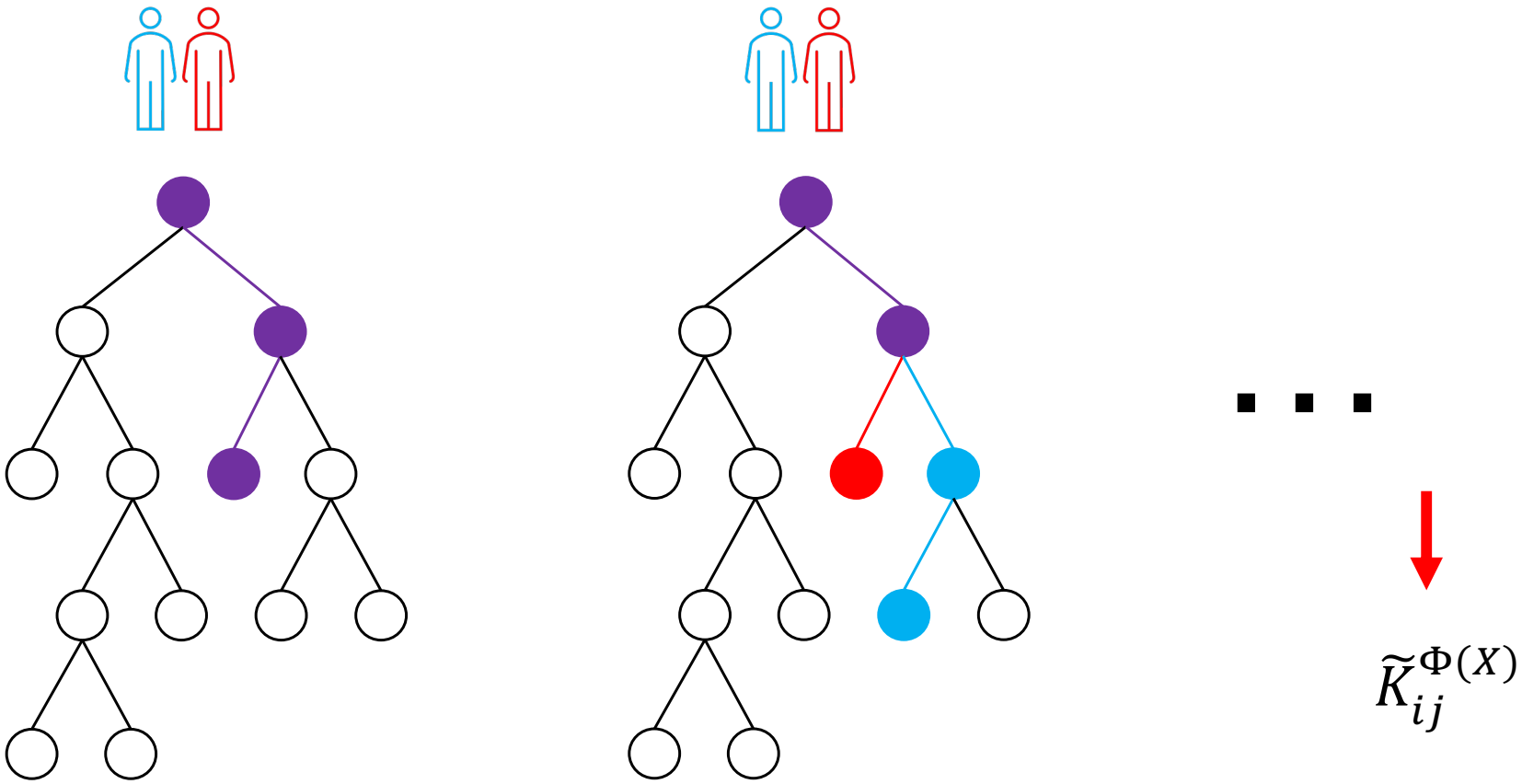
# **Step 1**: Train a Random Forest (RF)



. . .

L. Breiman. Machine Learning (2001)

# **Step 2**: Estimate RF Kernel



$$\widetilde{K}_{ij}^{\Phi(X)}$$

L. Breiman. Stat. Sci. (2000)

# **Step 3**: Compute Unbiased Dcov

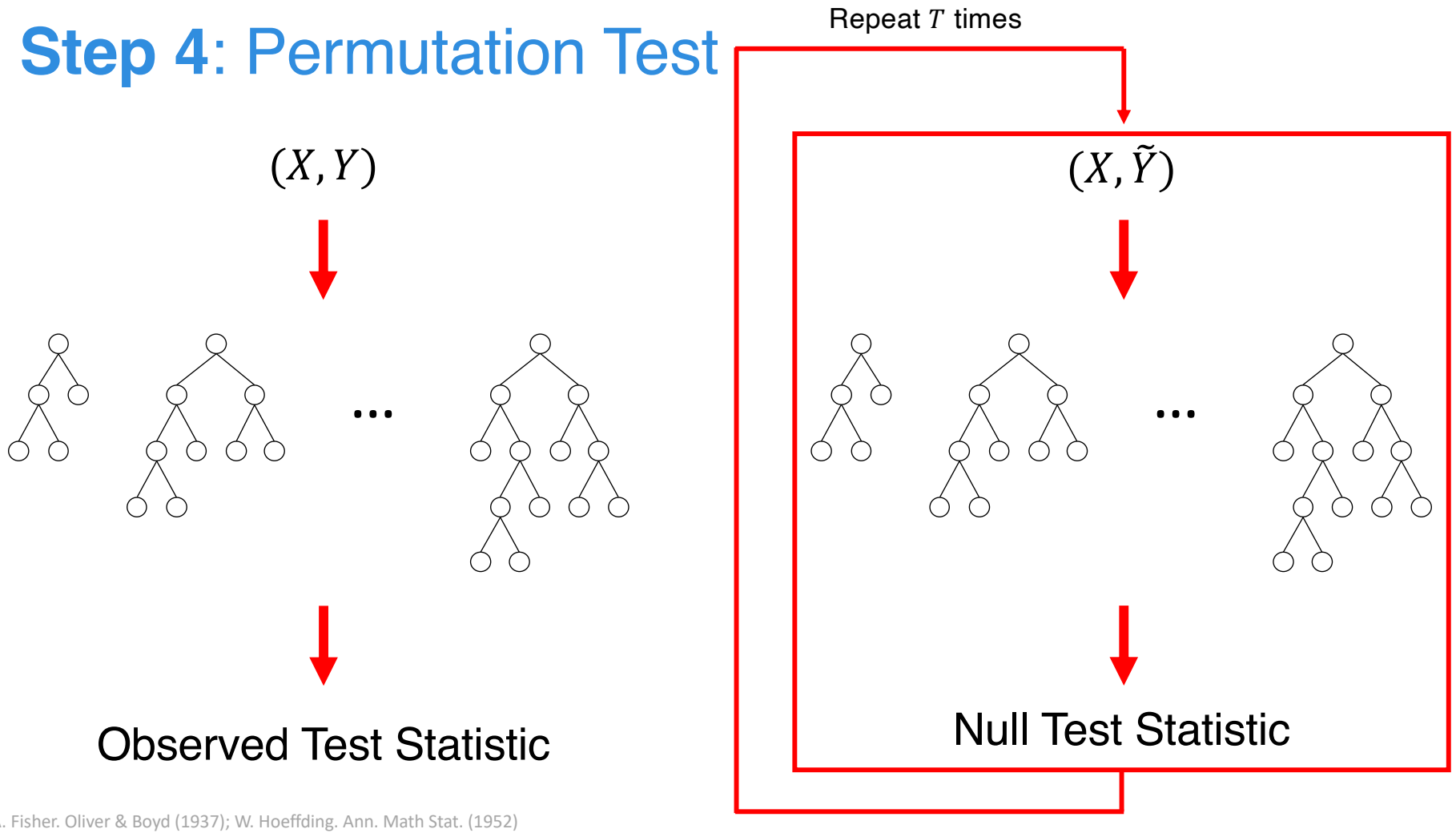
$$L_{ij}^{X} = \begin{cases} \widetilde{K}_{ij}^{\Phi(X)} - \dfrac{1}{n-2}\sum_{t=1}^{n}\widetilde{K}_{it}^{\Phi(X)} - \dfrac{1}{n-2}\sum_{s=1}^{n}\widetilde{K}_{sj}^{\Phi(X)} + \dfrac{1}{(n-1)(n-2)}\sum_{s,t=1}^{n}\widetilde{K}_{st}^{\Phi(X)} & i \neq j \\ 1 & i = j \end{cases}$$

$$D^{X} = 1 - L^{X}$$

$$c_n(X,Y) = \frac{1}{n(n-3)}\operatorname{tr}(D^{X}D^{Y})$$

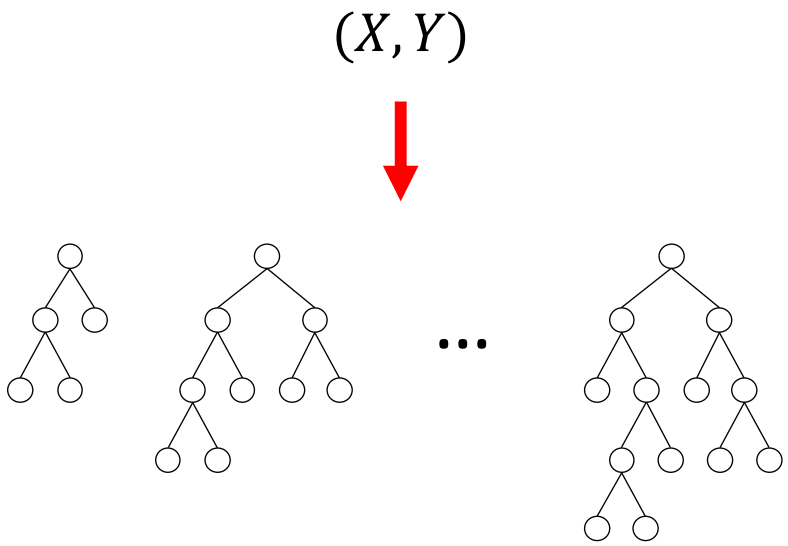
G. J. Székely & M. L. Rizzo. Ann. Stat. (2014); C. Shen & J. T. Vogelstein AStA Adv. Stat. Anal. (2020)

# **Step 4**: Permutation Test




R. A. Fisher. Oliver & Boyd (1937); W. Hoeffding. Ann. Math Stat. (1952)

# **Step 4**: Chi-Square Approximation to Dcorr



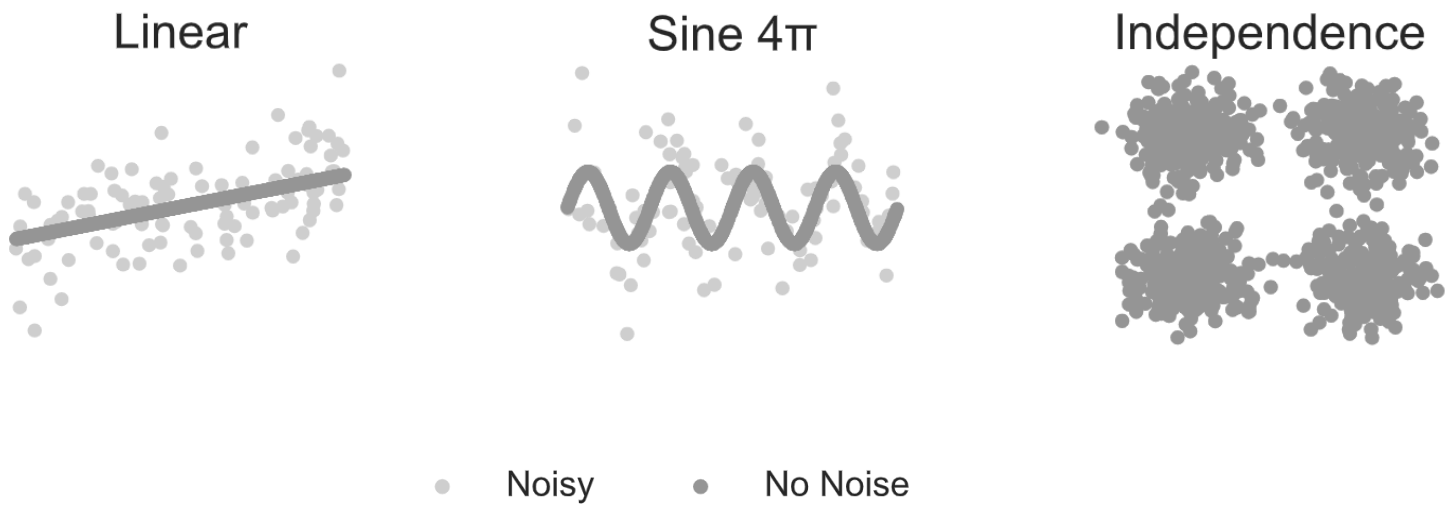
$(X, Y)$

Observed Test Statistic

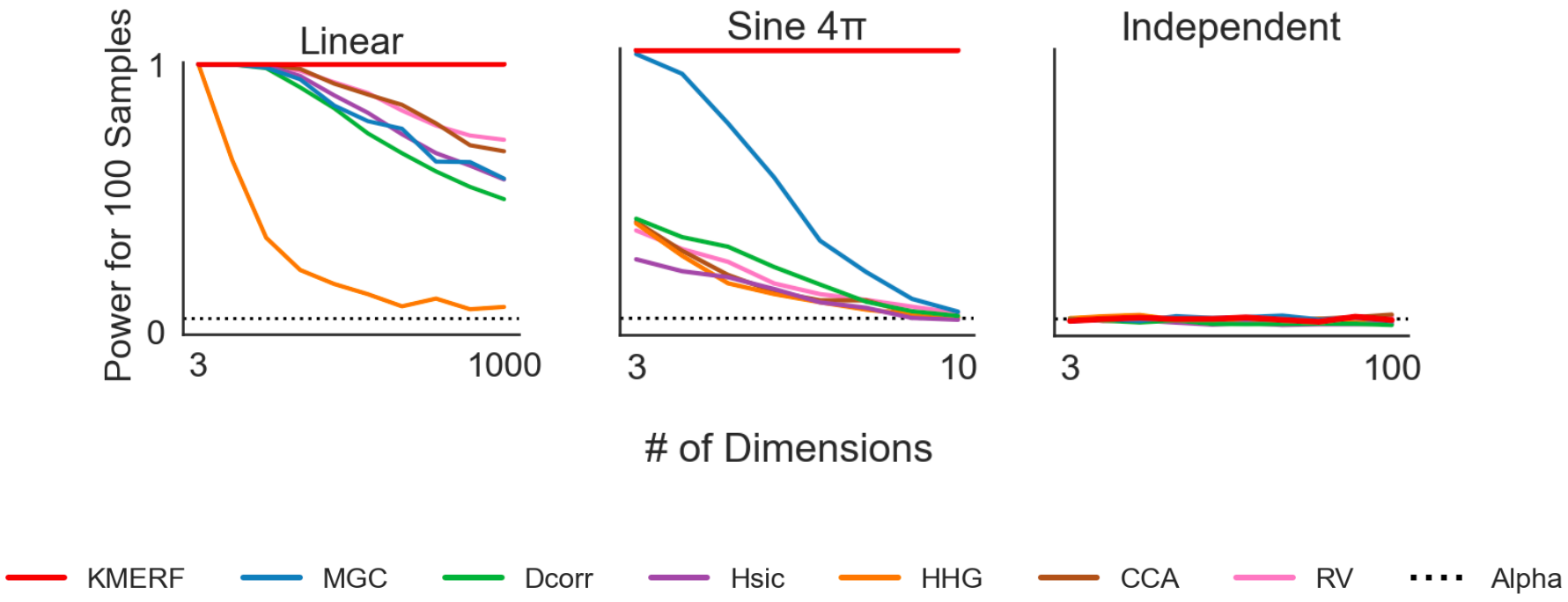
$$p = 1 - F_{\chi^2_{1^2-1}}\left(n \cdot \frac{c_n(X,Y)}{\sqrt{c_n(X,X) \cdot c_n(Y,Y)}}\right)$$

C. Shen, S. Panda & J. T. Vogelstein. JCGS (2022)

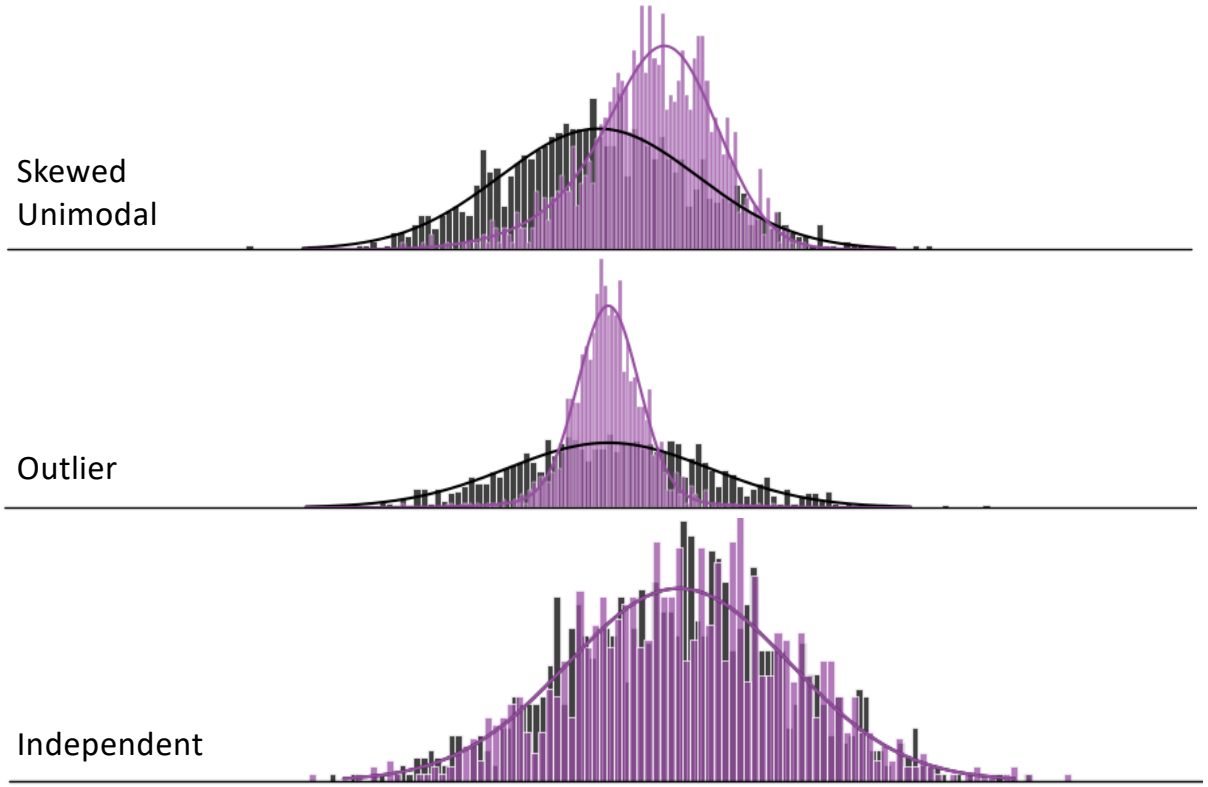
# Independence testing simulation settings



# KMERF dominates other tests in these settings

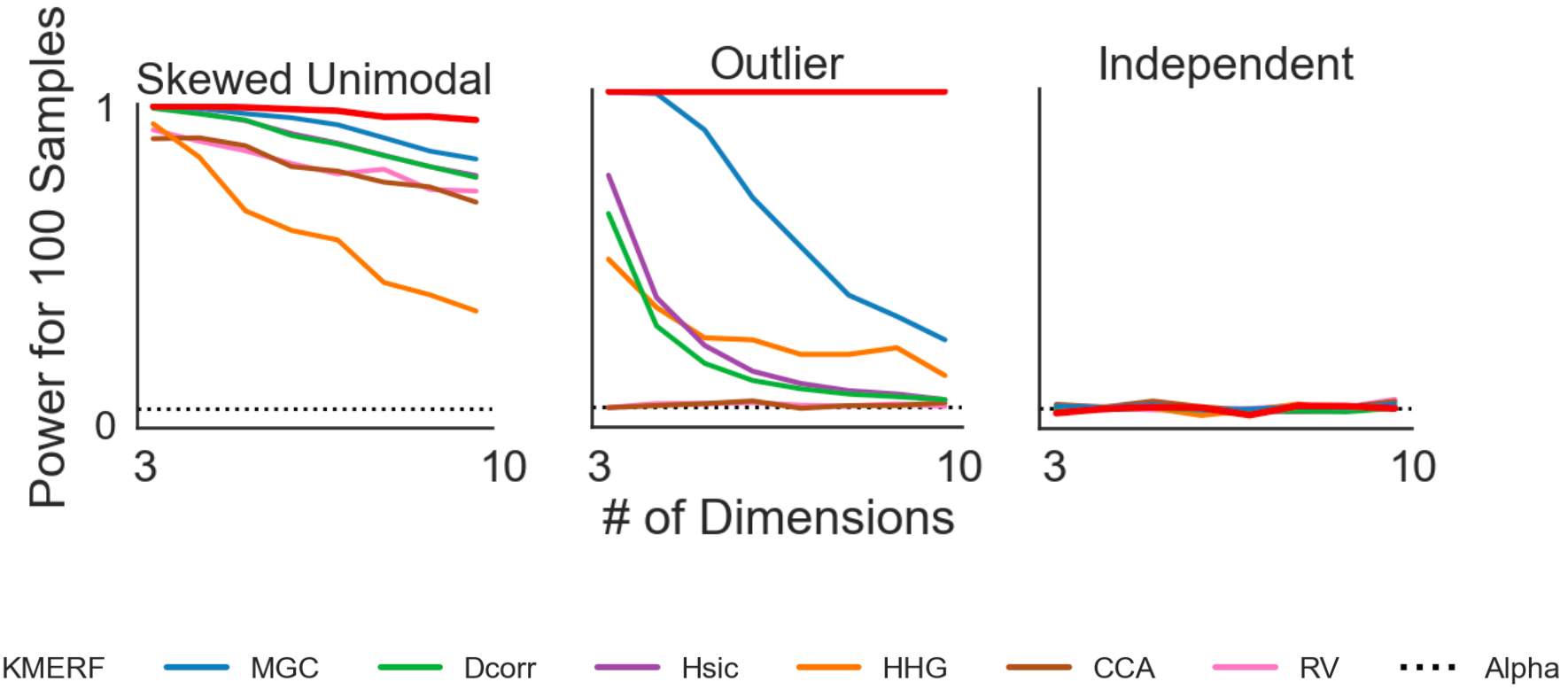


# Two-sample testing simulation settings




S. Marron & M. Wand. Ann. Stat (1992)

# KMERF dominates other tests in these setting



# KMERF can also be improved

**Independence Tests**

- RV/CCA
- Dcorr/Hsic
- HHG
- MGC
- KMERF
- etc.

**Two/K-Sample Tests**

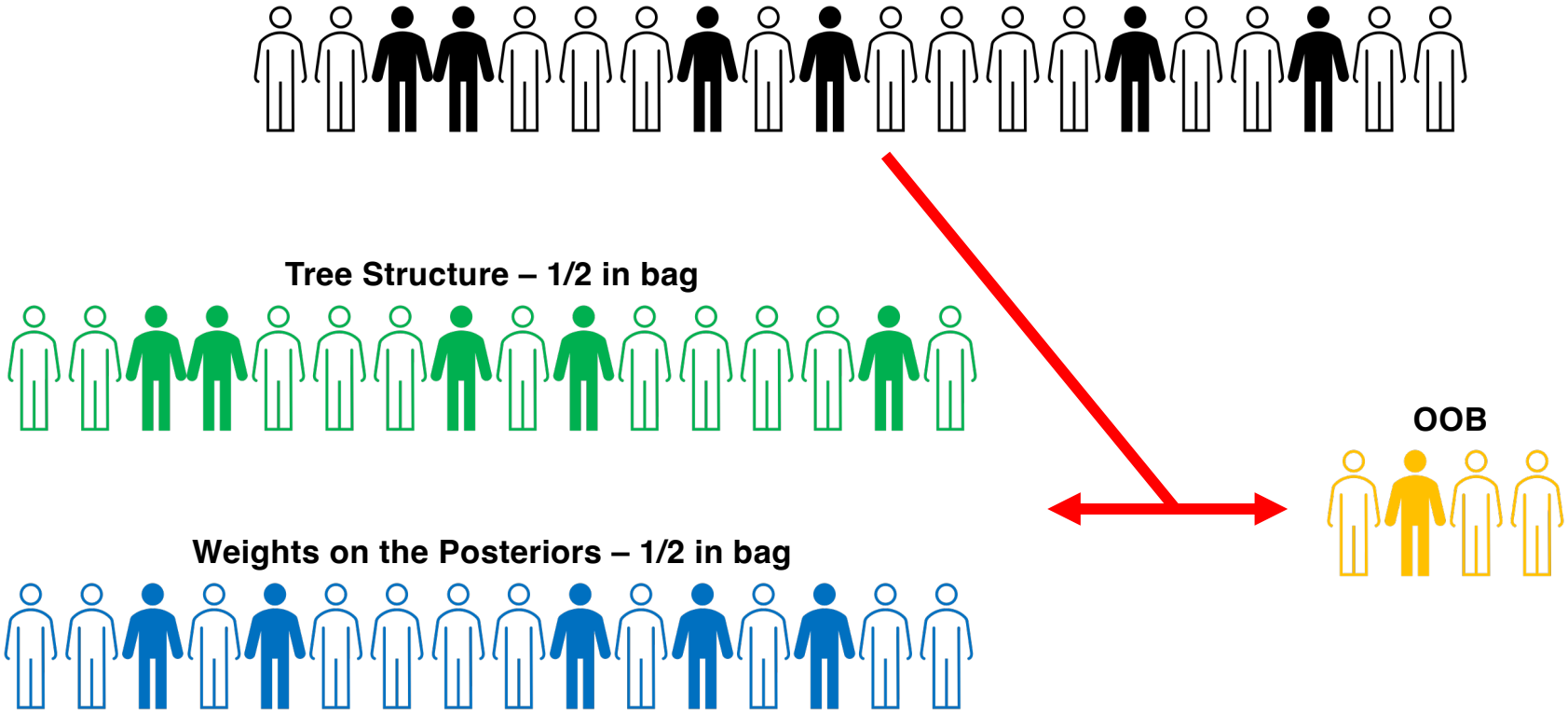
- Hotelling's $T^2$
- Energy/MMD
- HHG
- Our K-Sample papers
- etc.

- **Estimating quantities directly from random forest may especially improve high-dimensional performance**
- **Additionally, none of these tests can tell us whether an informative set or combination of sets is more predictive**

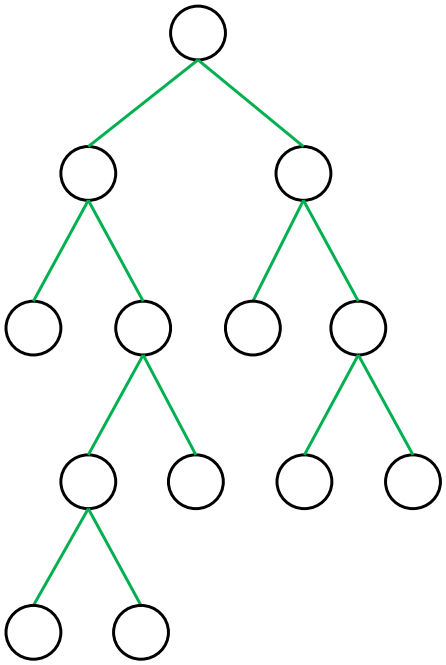
# Aim 3: Multidimensional Informed Generalized Hypothesis Technology (MIGHT)

S. Panda et al. In Preparation PNAS.

# **Step 1**: Bootstrap and get 3 sets for each tree

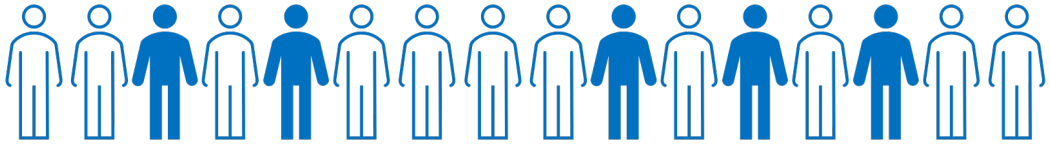


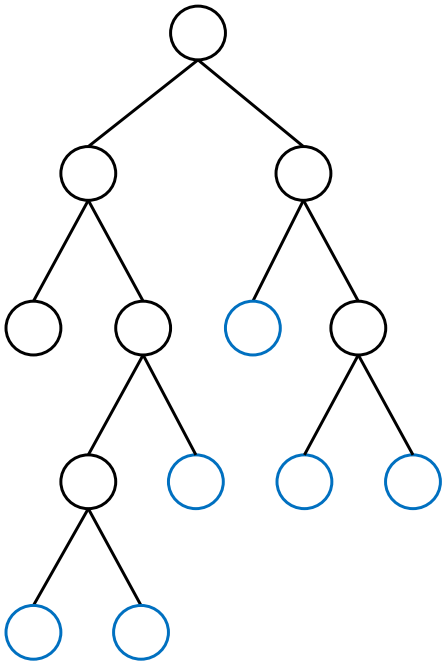
## **Step 2**: Grow the trees



. . .

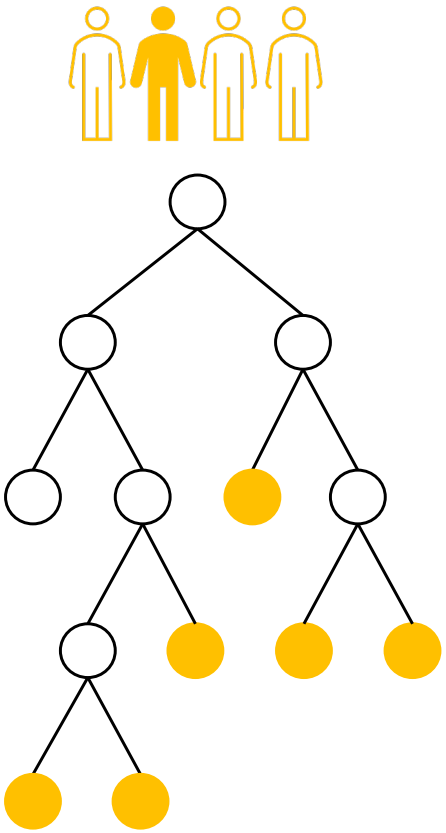
## **Step 3**: Estimate posterior weighting function





. . .

# **Step 4**: Compute posteriors and test statistics

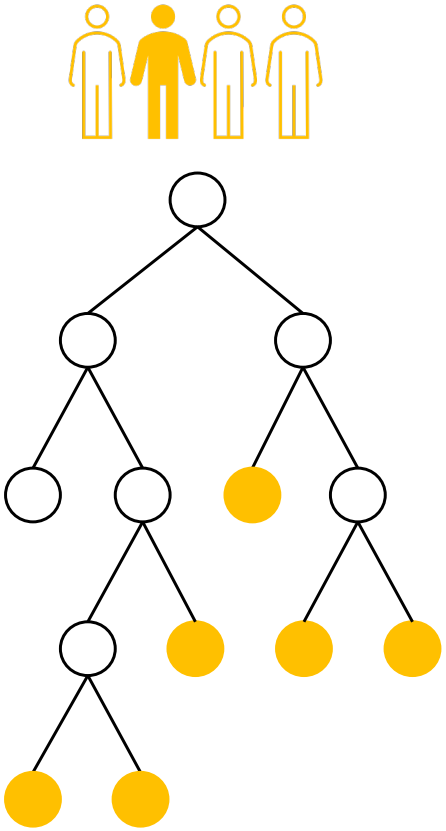


■ ■ ■

**Compute Test Statistics**

1. Classification Accuracy (Acc)
2. Mutual information (MI)
3. Area under the curve (AUC)
4. Sensitivity at k% Specificity (S@k)

# **Step 4**: Compute posteriors and test statistics

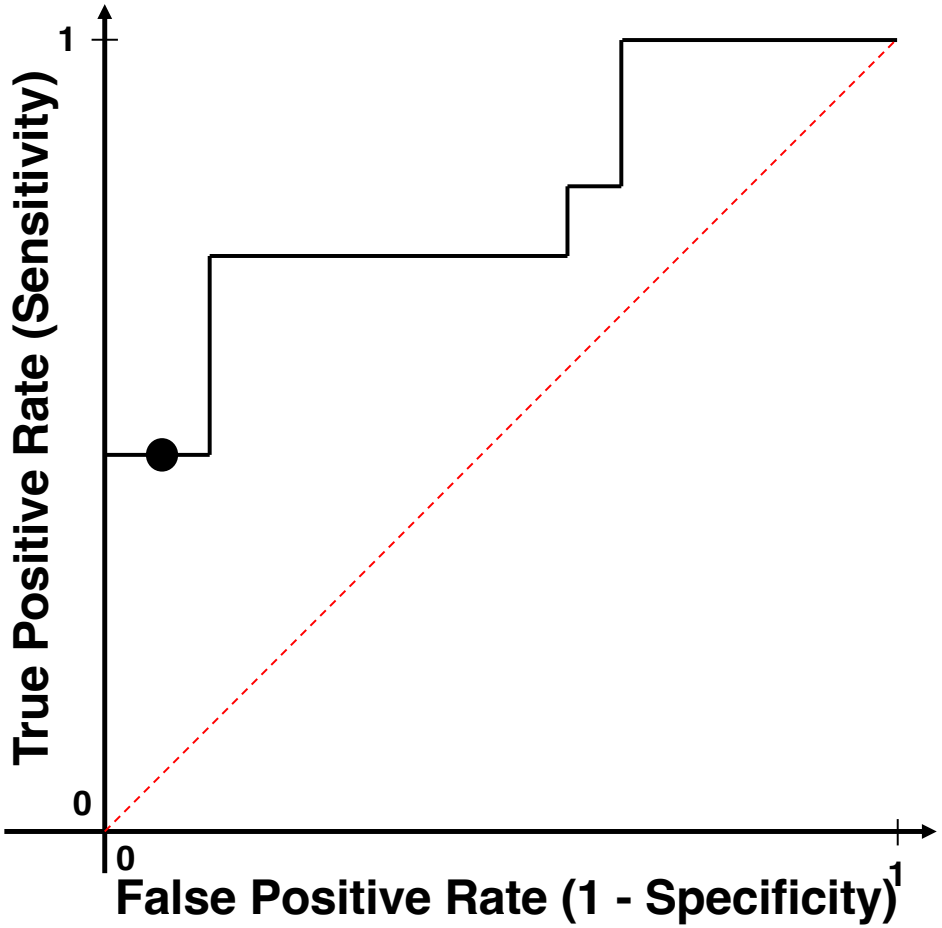


■ ■ ■

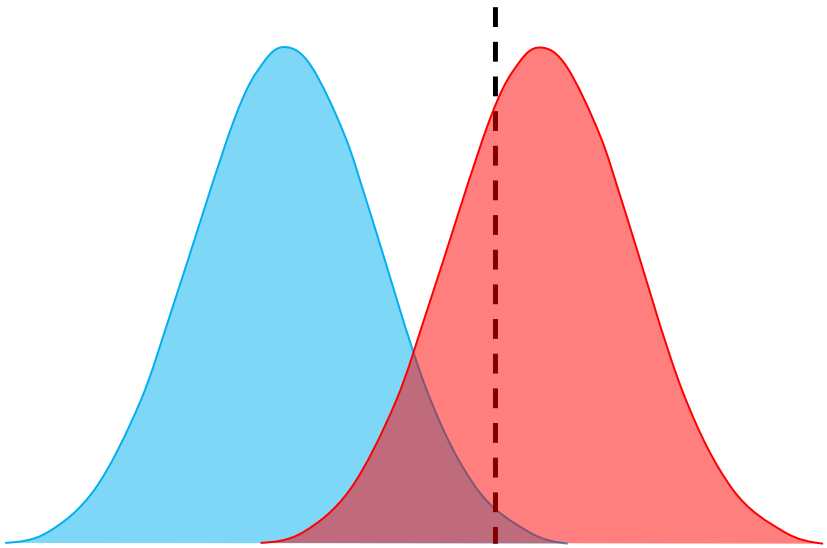
**Compute Test Statistics**

1. Classification Accuracy (Acc)
2. Mutual information (MI)
3. Area under the curve (AUC)
4. **Sensitivity at k% Specificity (S@k)**

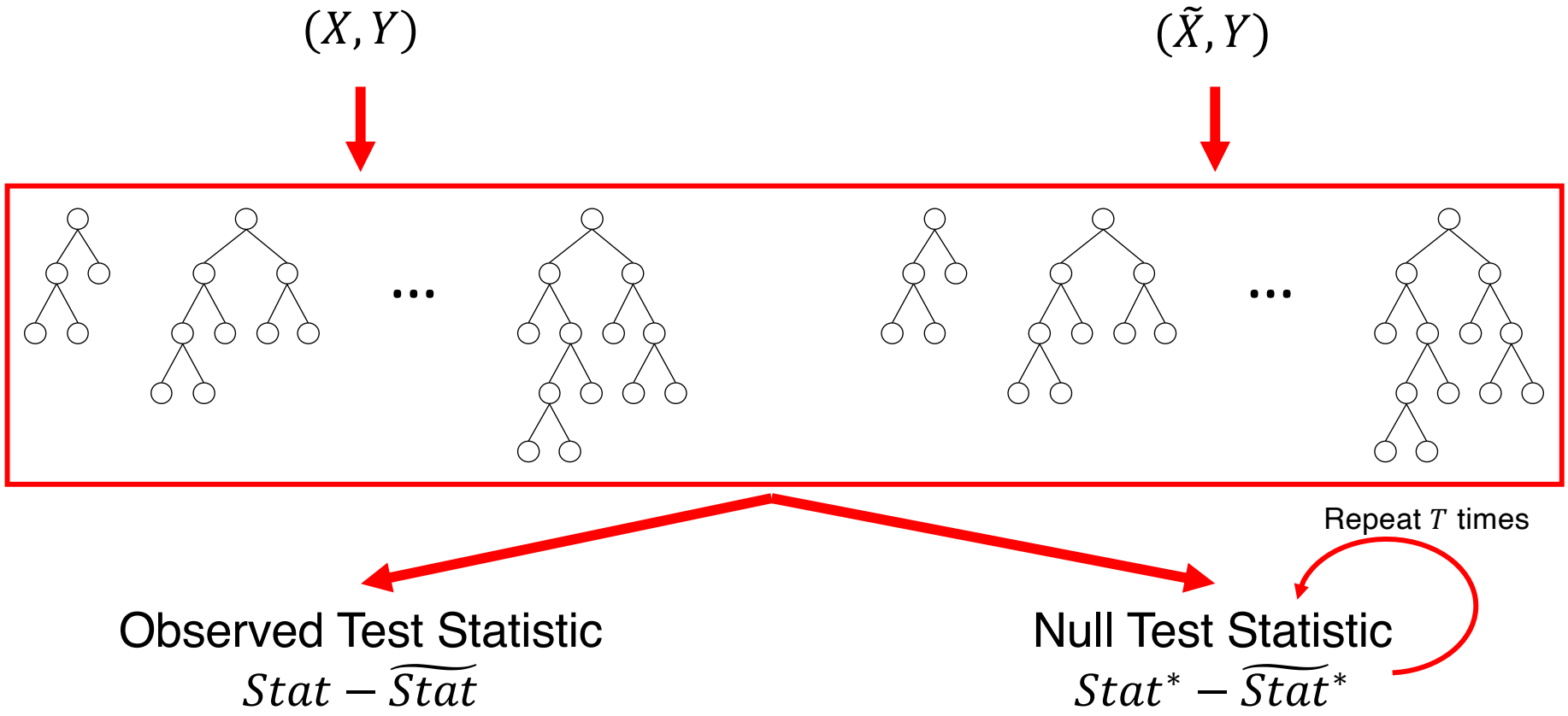
# Sensitivity at k% Specificity (S@k)




$$S@k = \mathbb{P}\{\eta(X) > T_k | Y = 1\}$$

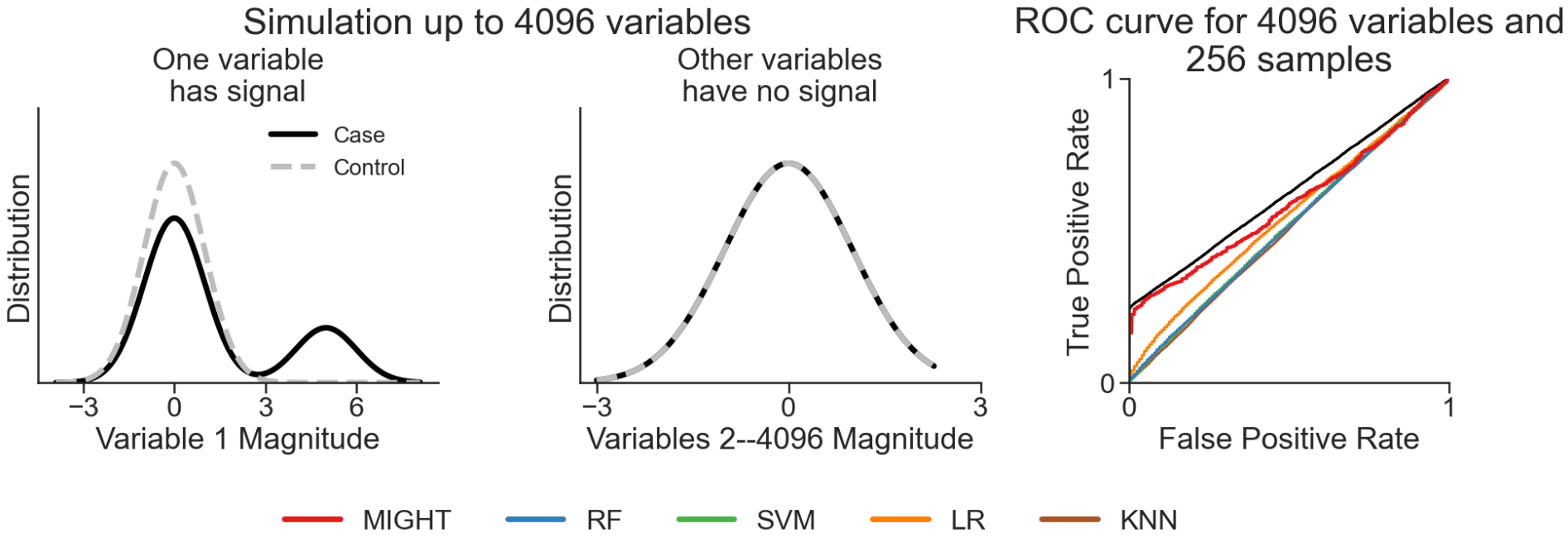


# **Step 5**: Fast Permutation Test

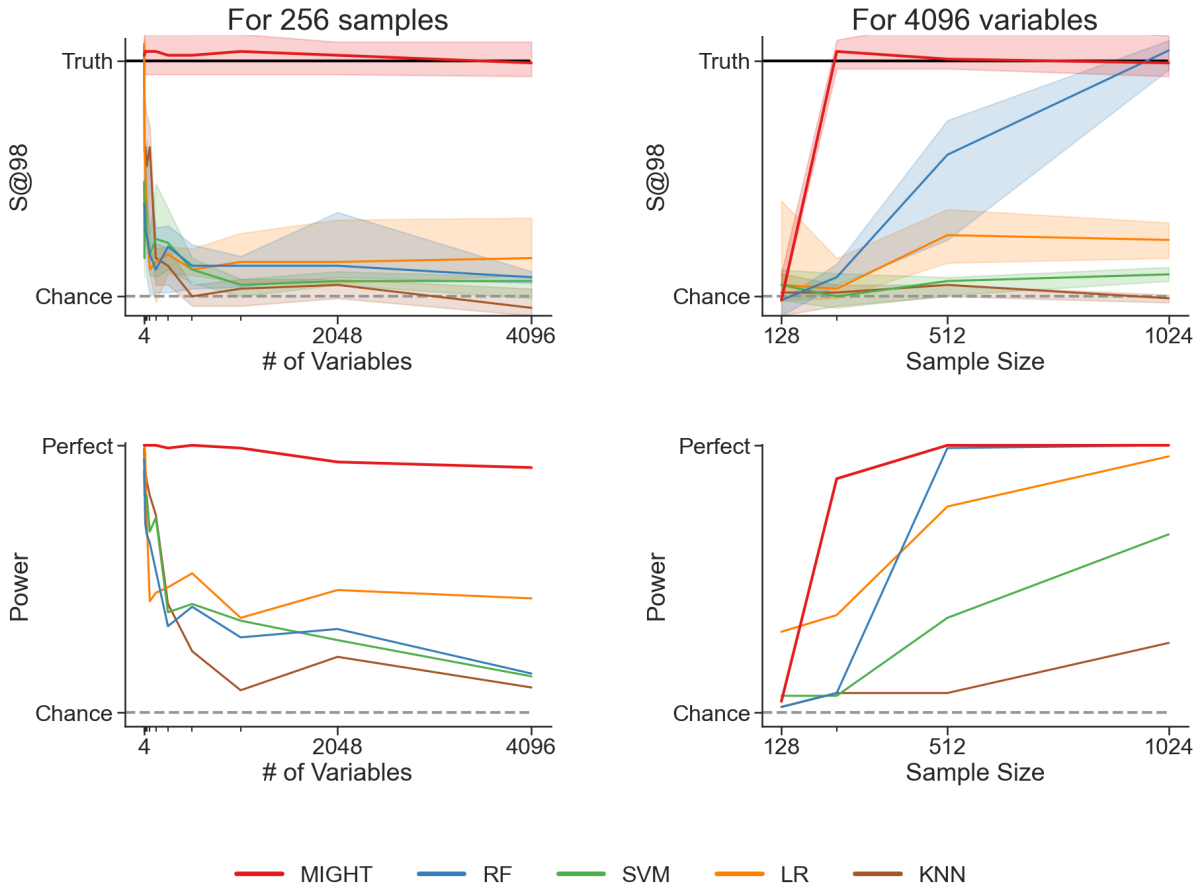



T. Coleman, W. Peng, L. Mentch. JMLR (2022)

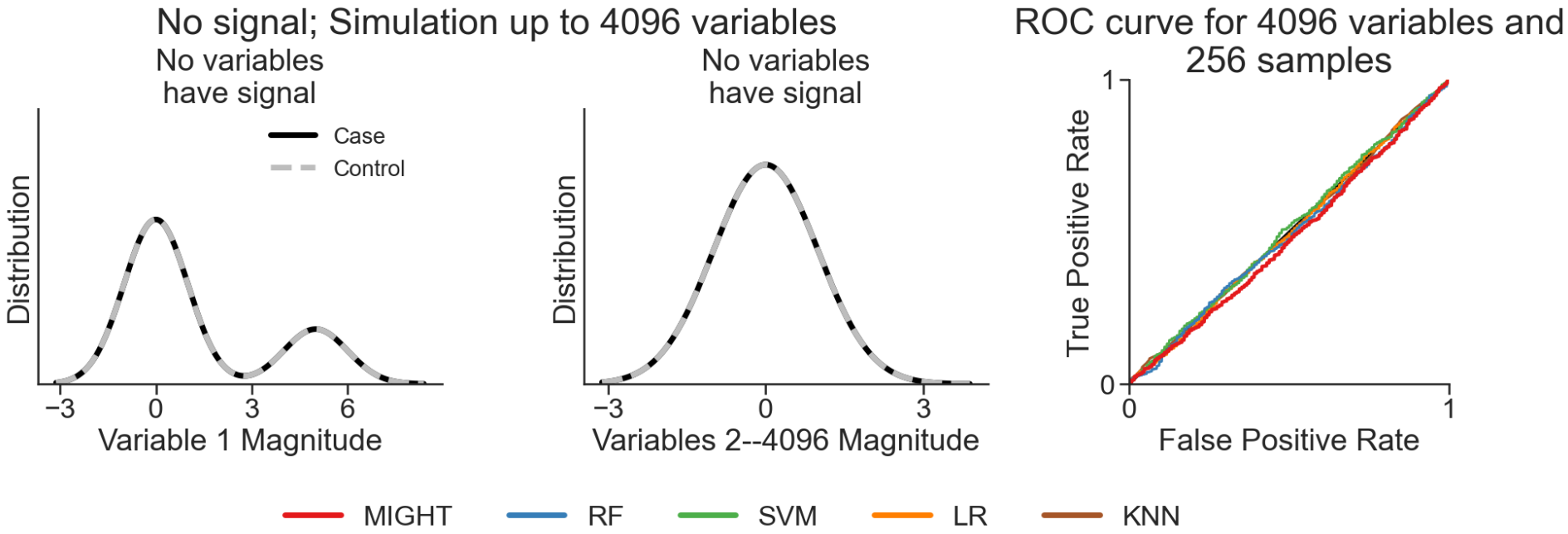
# How well does MIGHT do when signal exists?



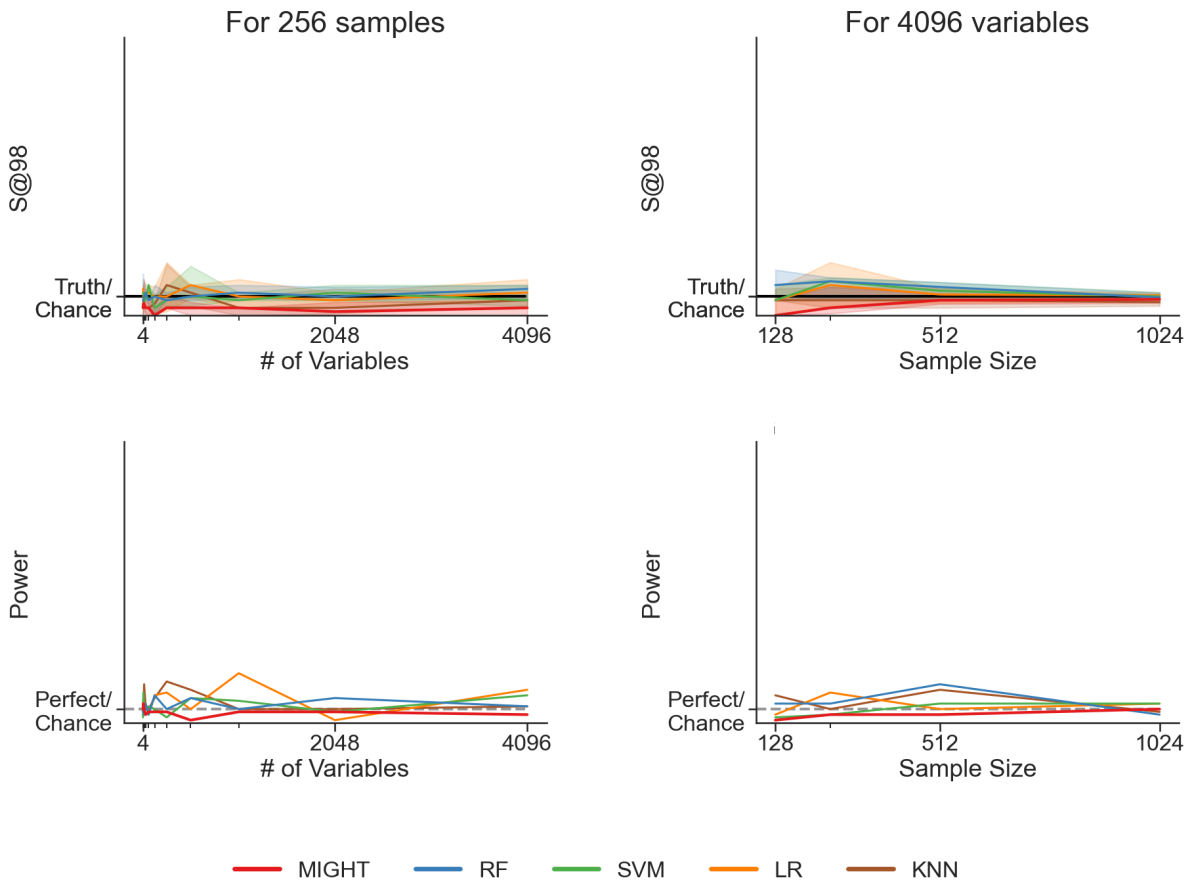
# MIGHT detects signal better than others



# What about when there is no signal?

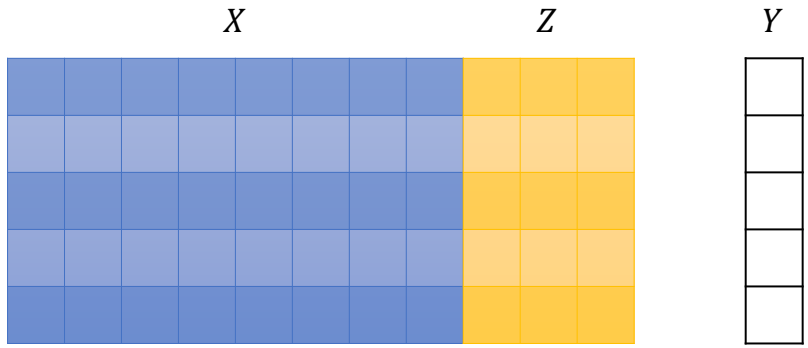


# All approaches work well when there is no signal

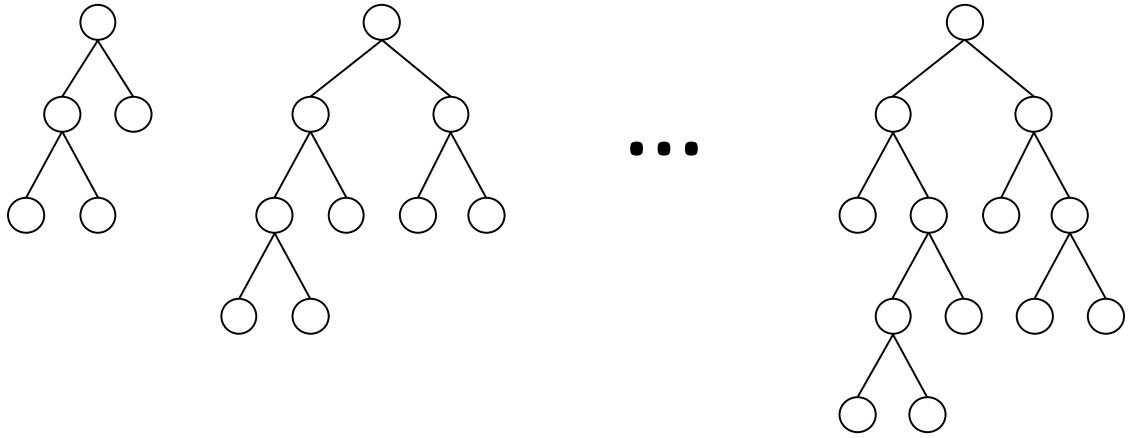


# CoMIGHT tests if sets add additional information

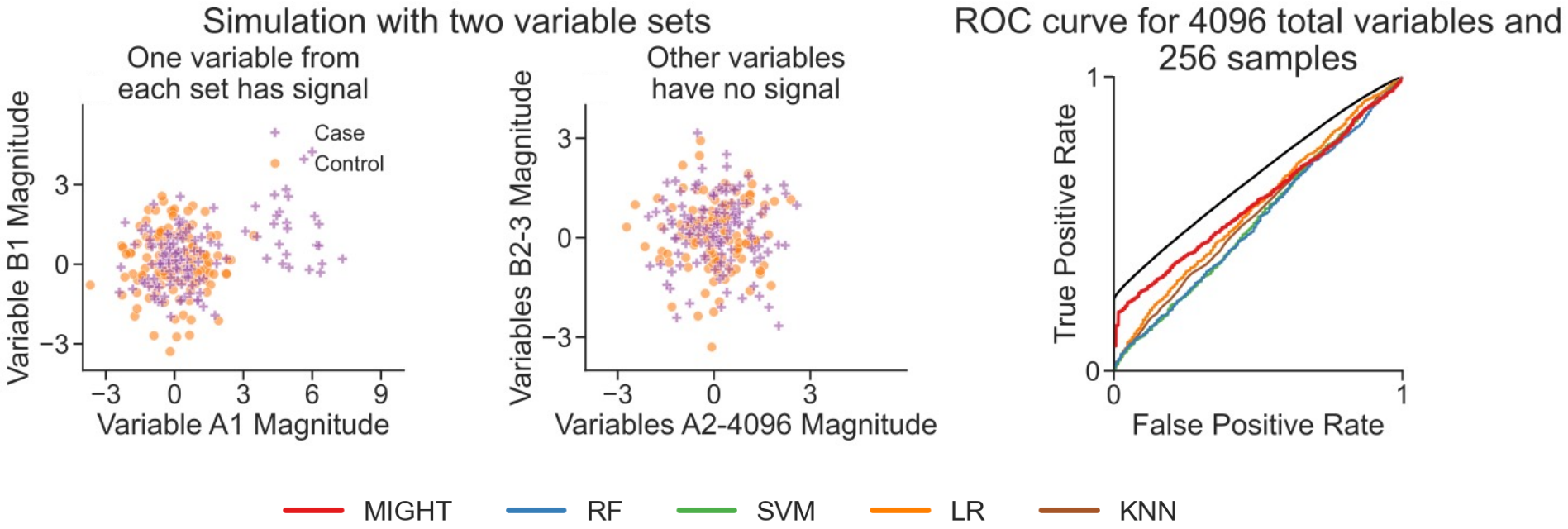
$$H_0 : F_{X,Y|Z} = F_{X|Z} F_{Y|Z}$$
$$H_A : F_{X,Y|Z} \neq F_{X|Z} F_{Y|Z}$$



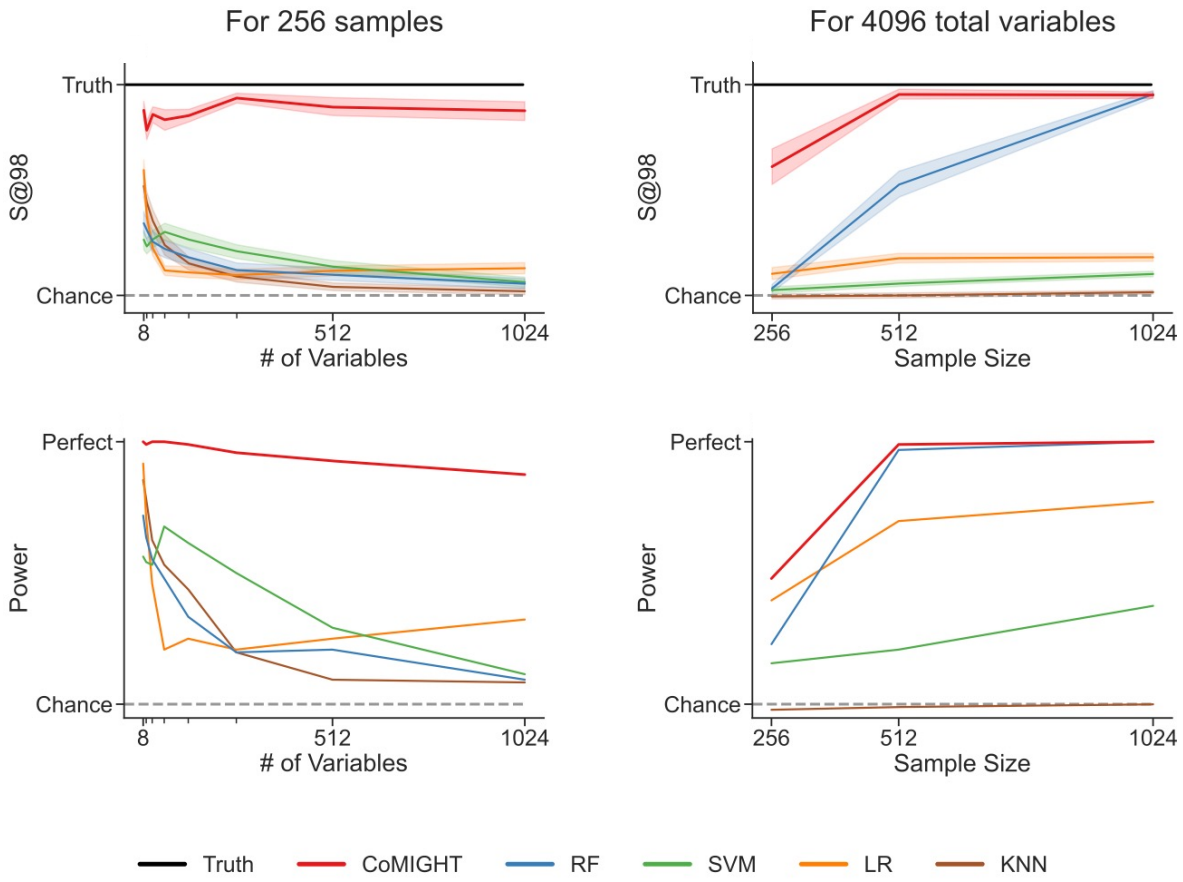

P-value computation permutes columns of $Z$



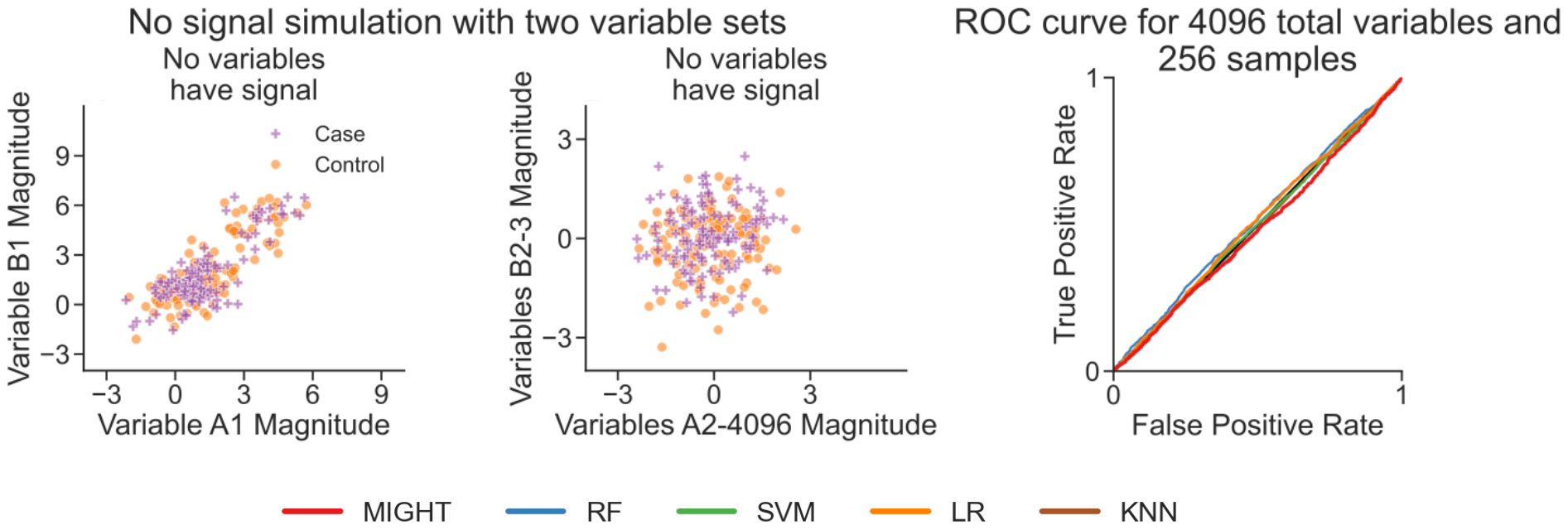
# How well does CoMIGHT do when signal exists?



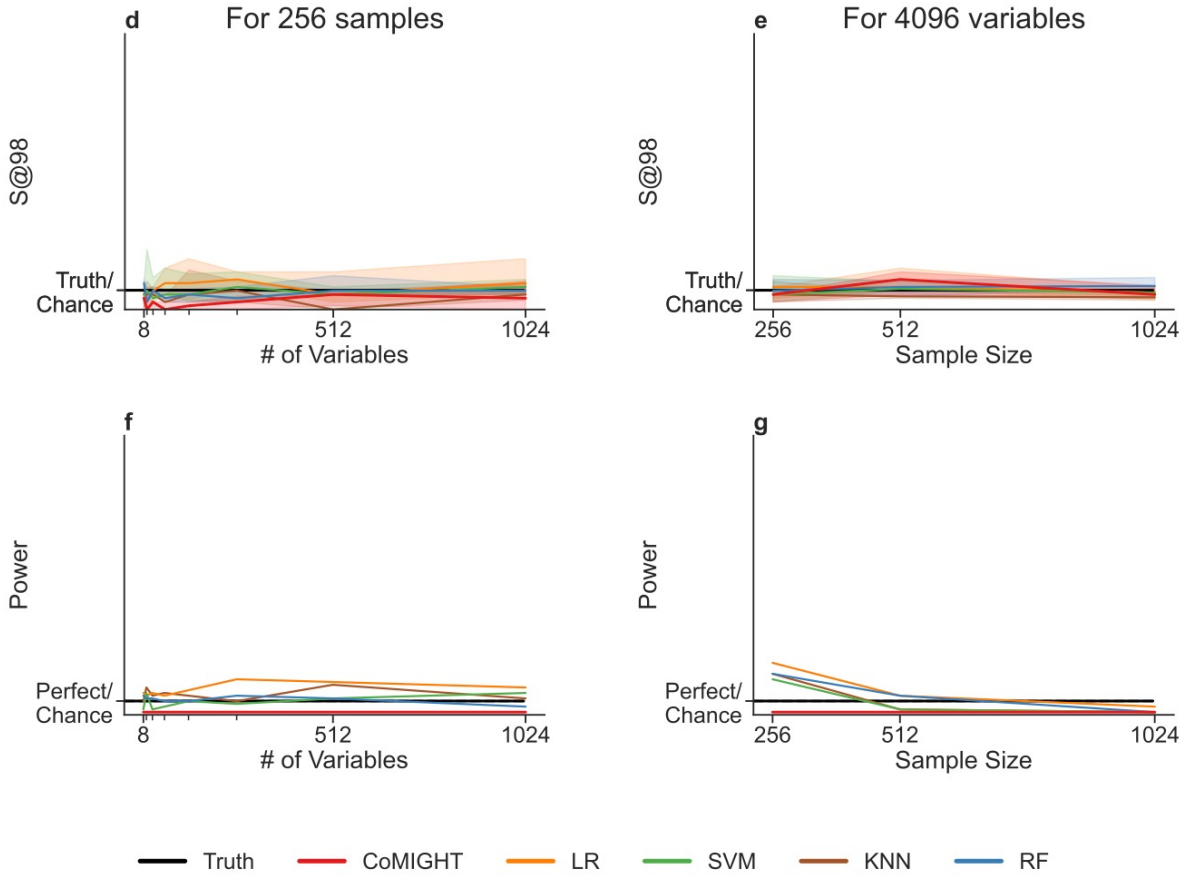
# CoMIGHT detects signal better than others



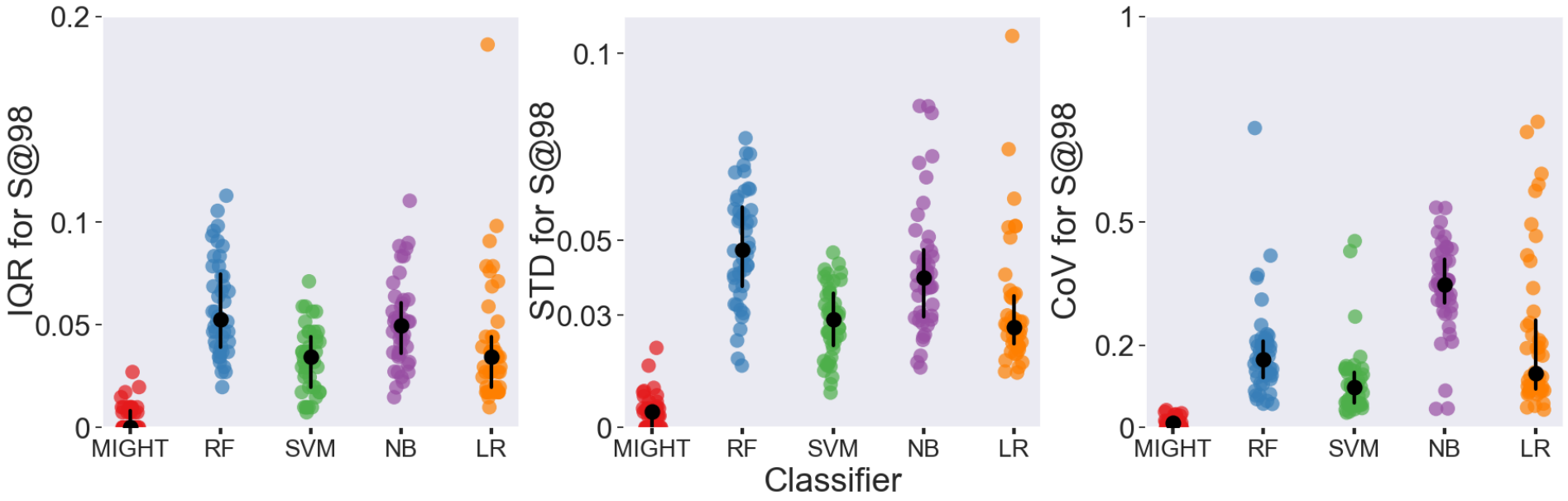
# What about when there is no signal?



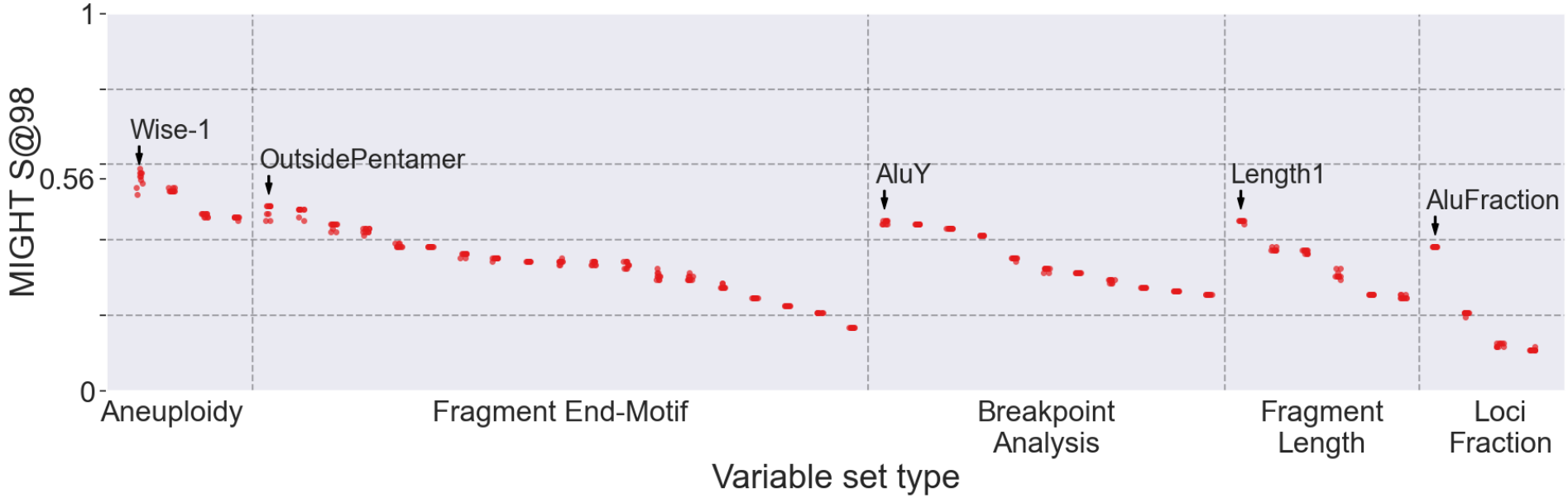
# Only CoMIGHT is valid at low sample sizes



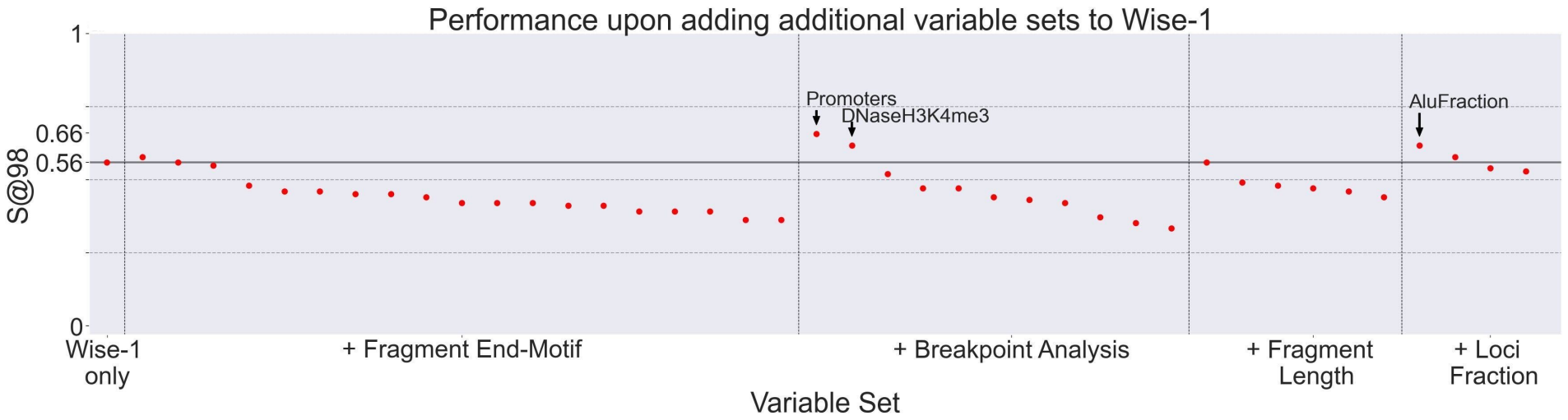
# MIGHT estimates are more precise than others



# MIGHT estimates are more precise than others



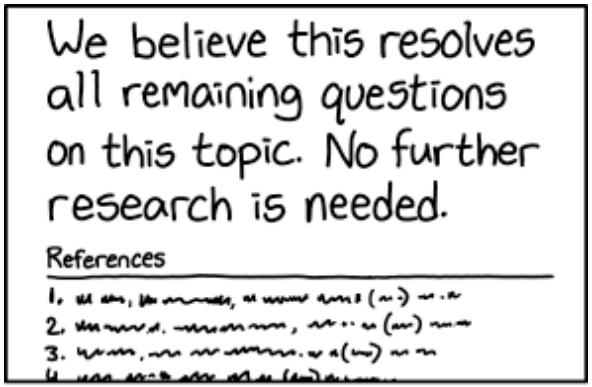
# Variable sets don’t always add information



# Conclusion

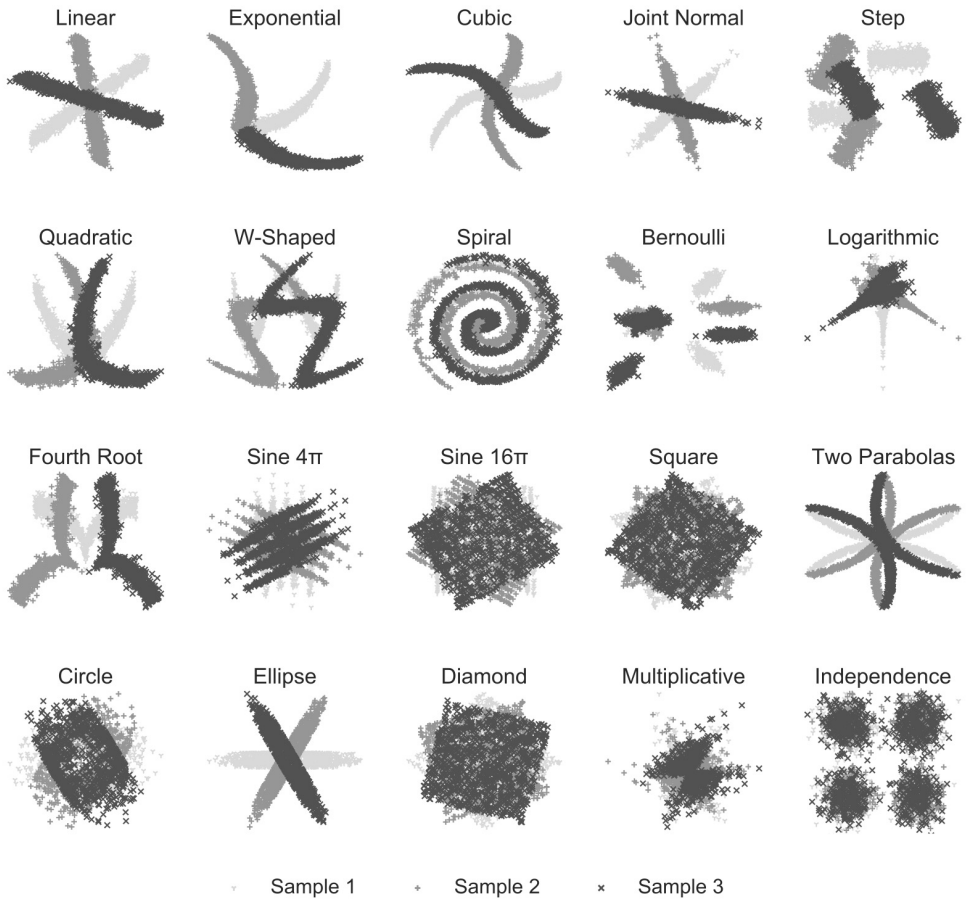
- **Aim 1**: We developed a new framework for k-sample testing
- **Aim 2**: KMERF has shown strong finite sample testing power
- **Aim 3**: MIGHT and CoMIGHT can determine information within feature sets using statistics practitioners care about

# Questions?

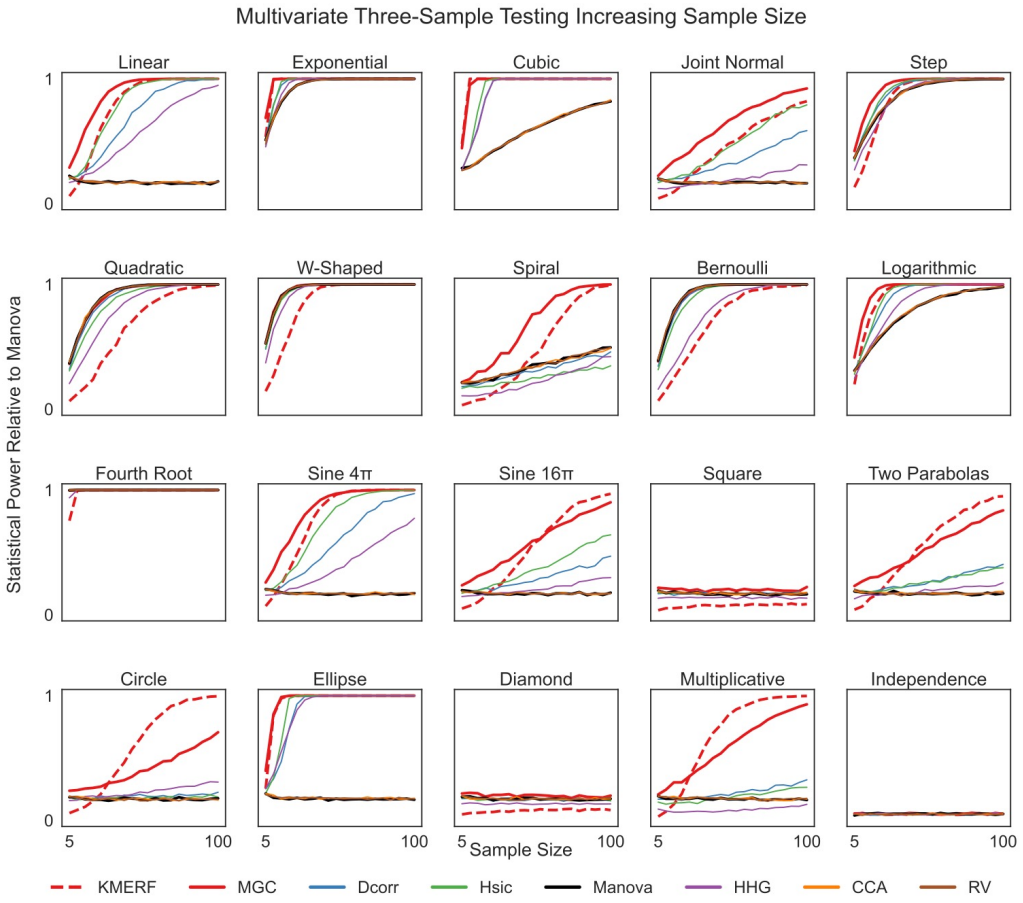



JUST ONCE, I WANT TO SEE A RESEARCH PAPER WITH THE GUTS TO END THIS WAY.

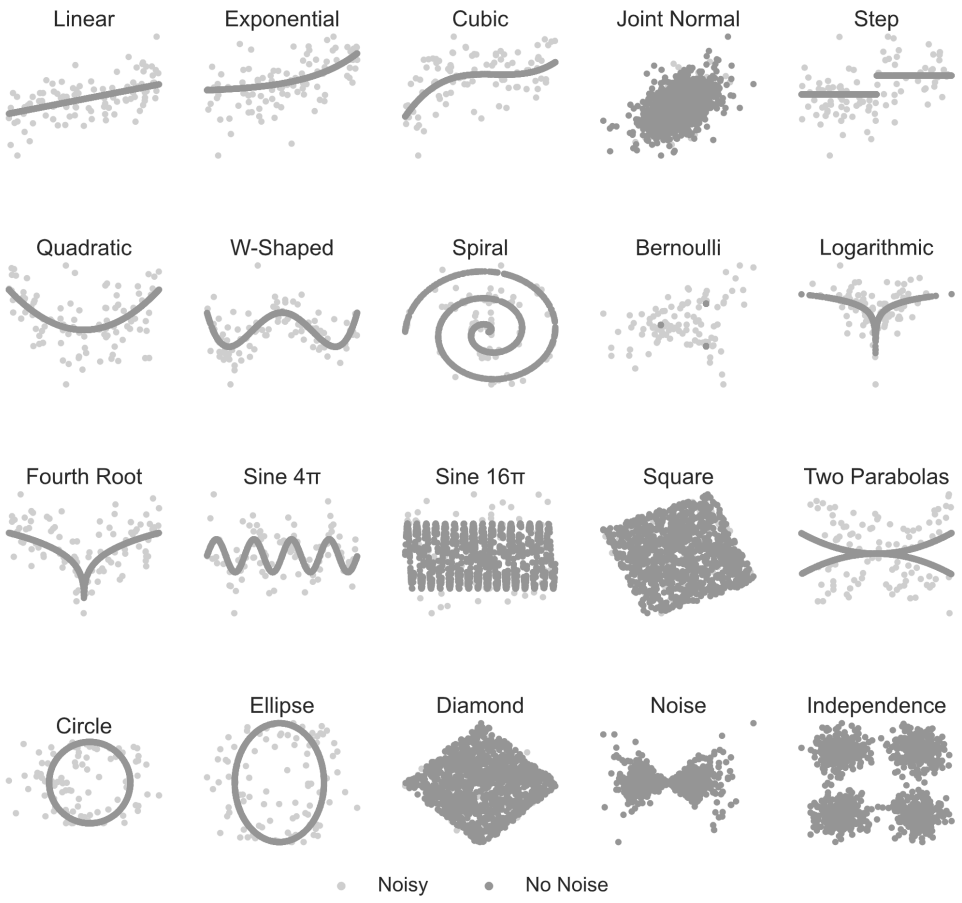
# 3-sample testing simulation settings



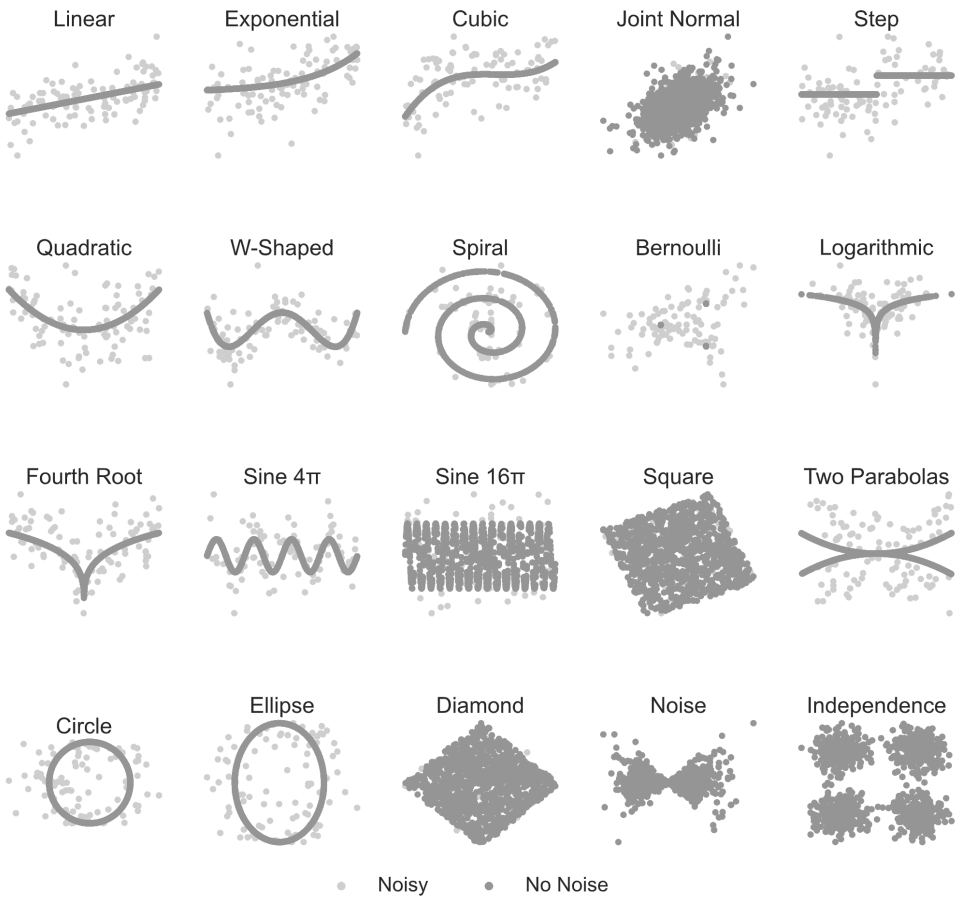
# 3-sample Testing vs. Sample Size



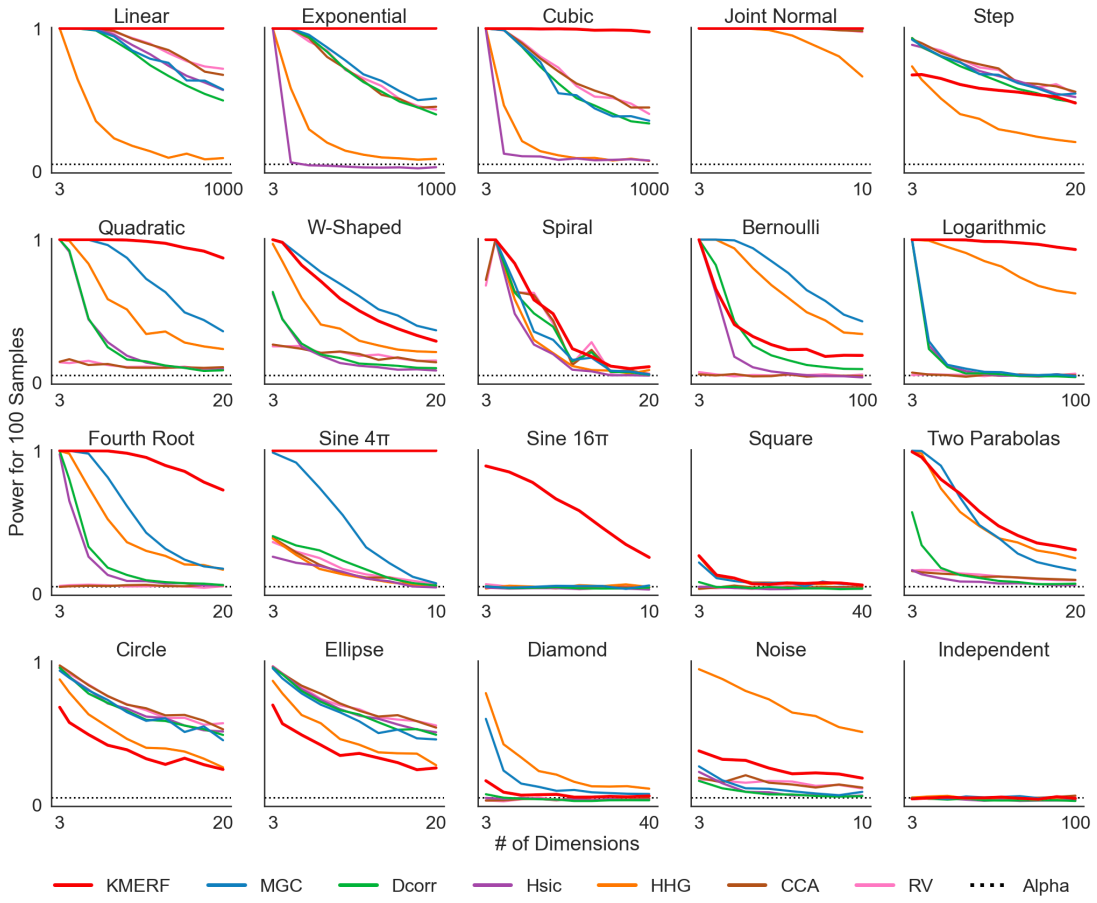
# Independence testing simulation settings



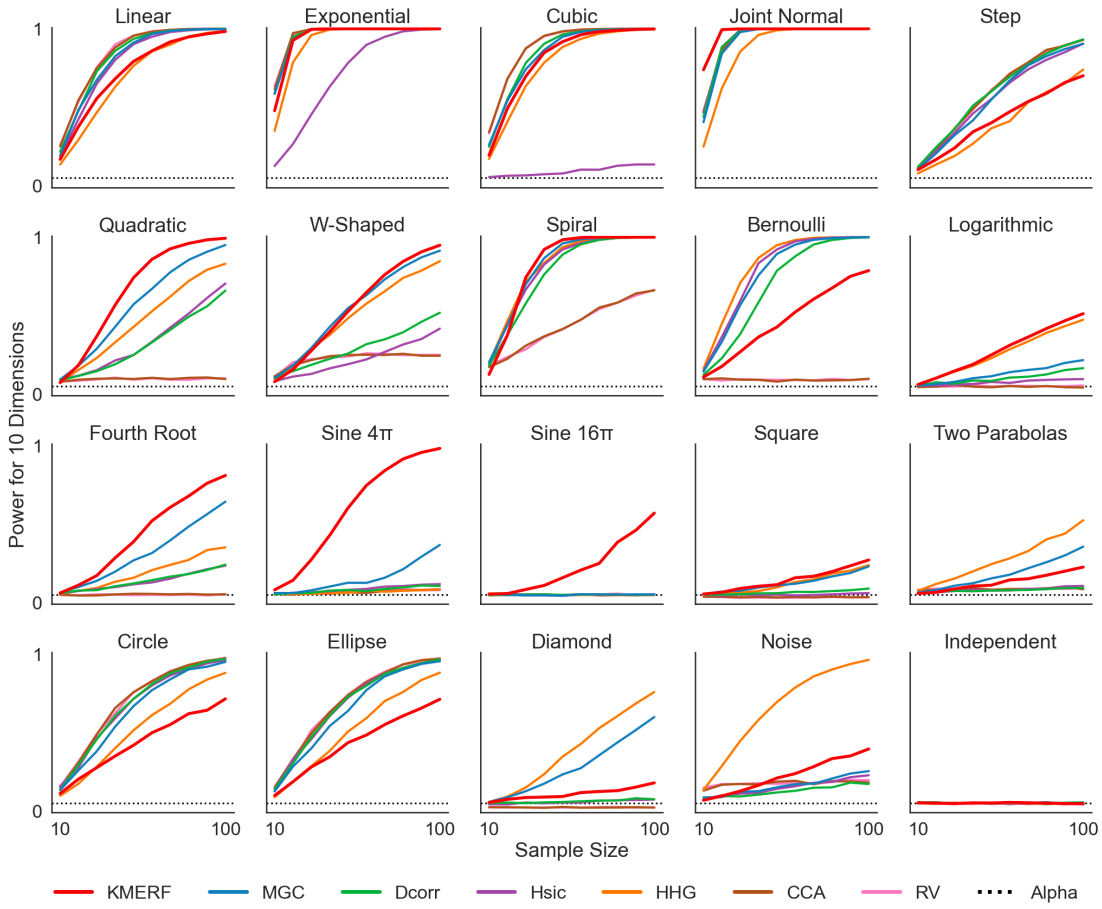
# Independence testing simulation settings



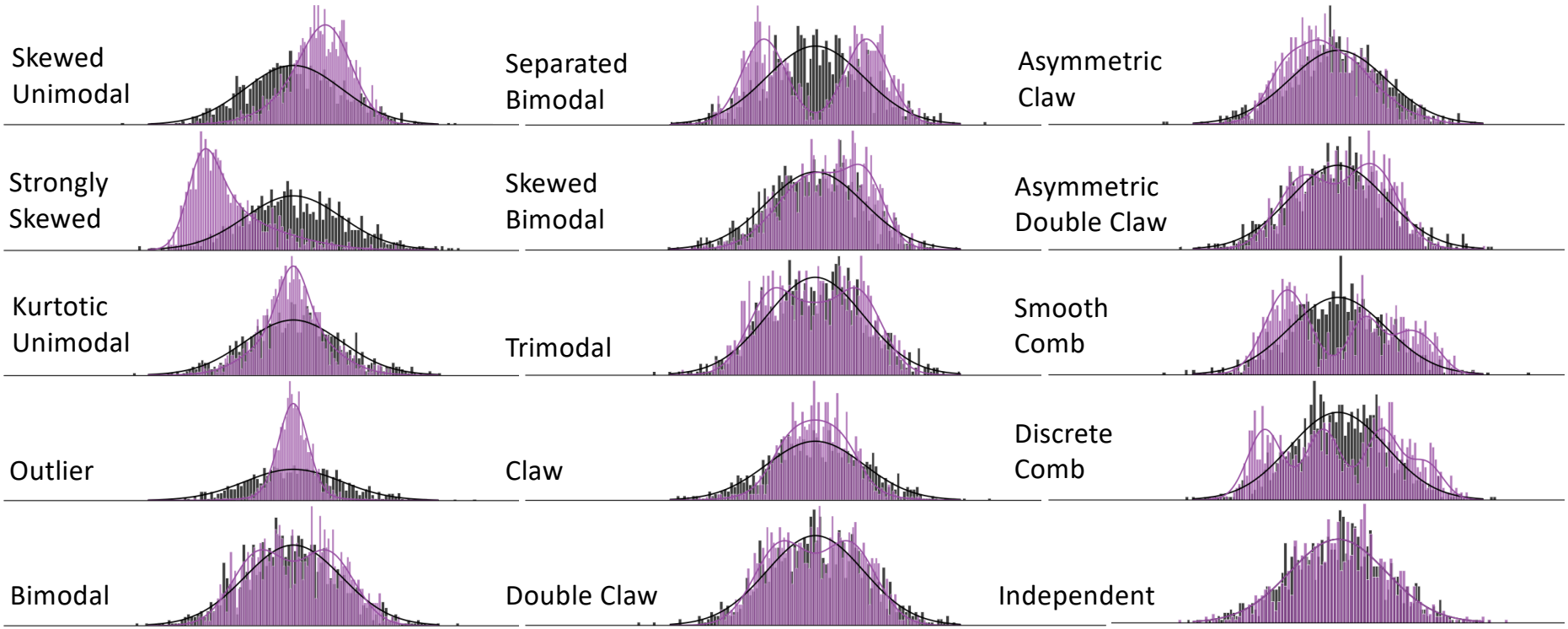
# Independence KMERF vs. Dimension



# Independence KMERF vs. Sample Size

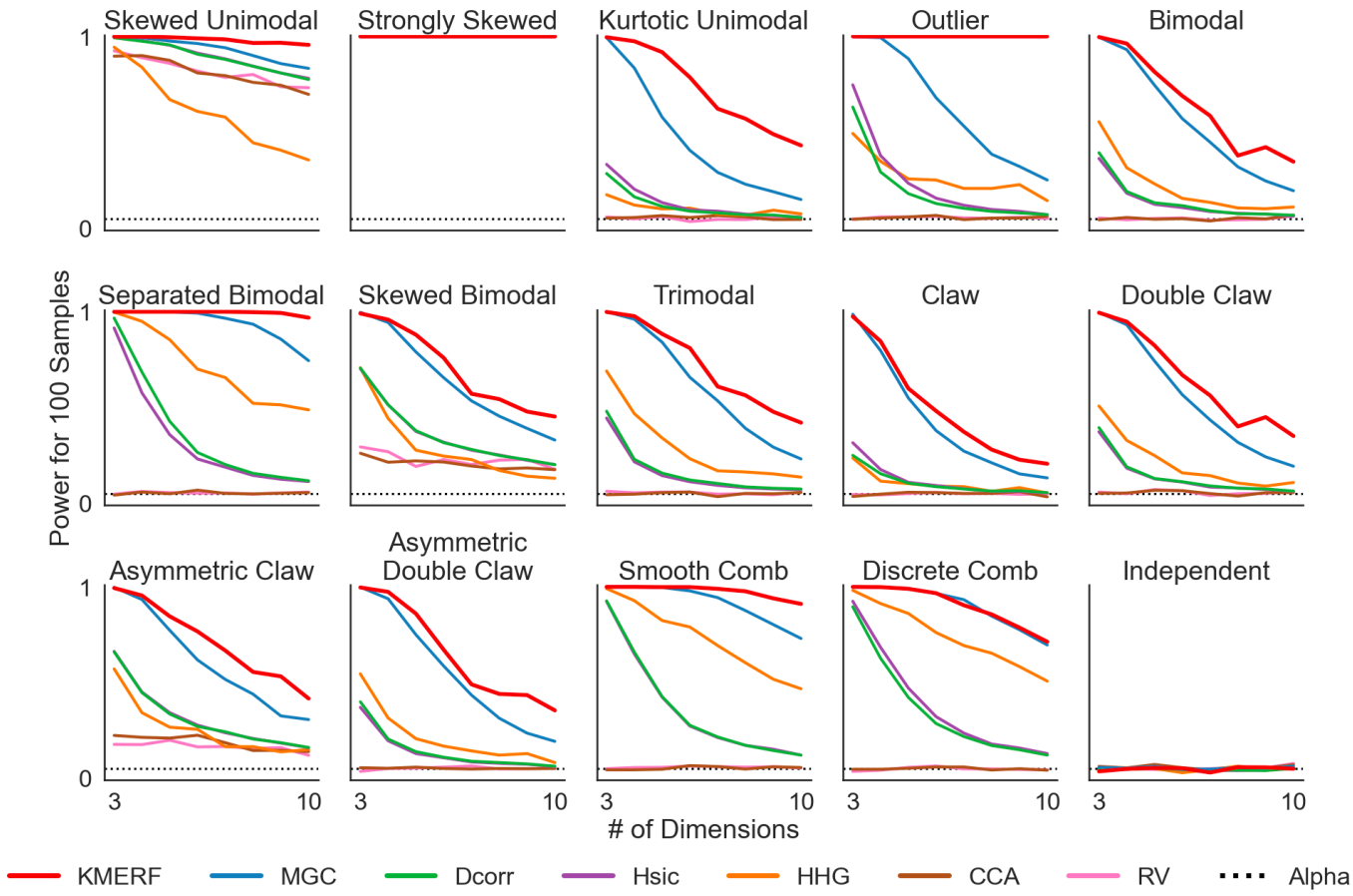


# Two-sample testing simulation settings

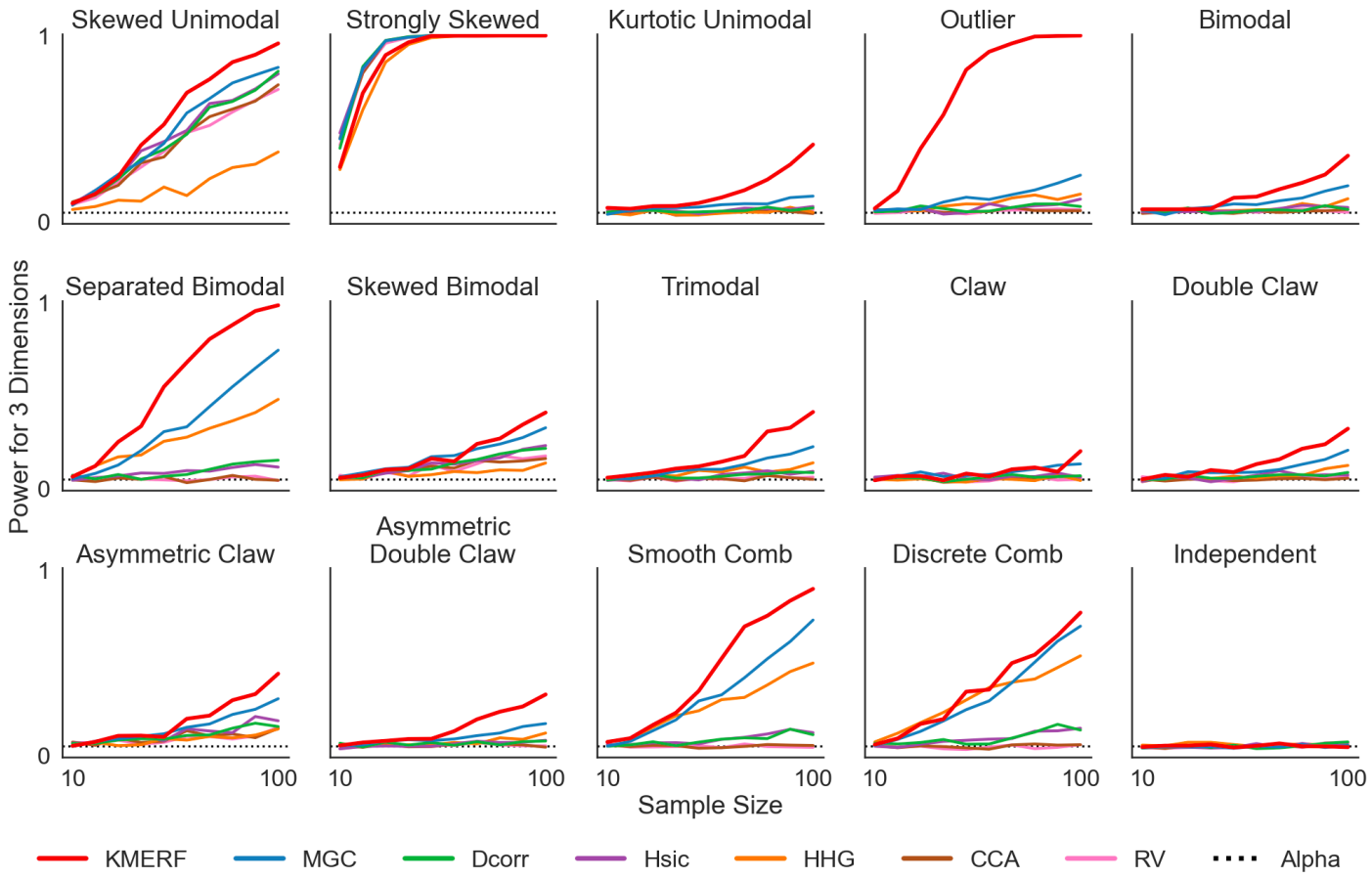



S. Marron & M. Wand. Ann. Stat (1992)

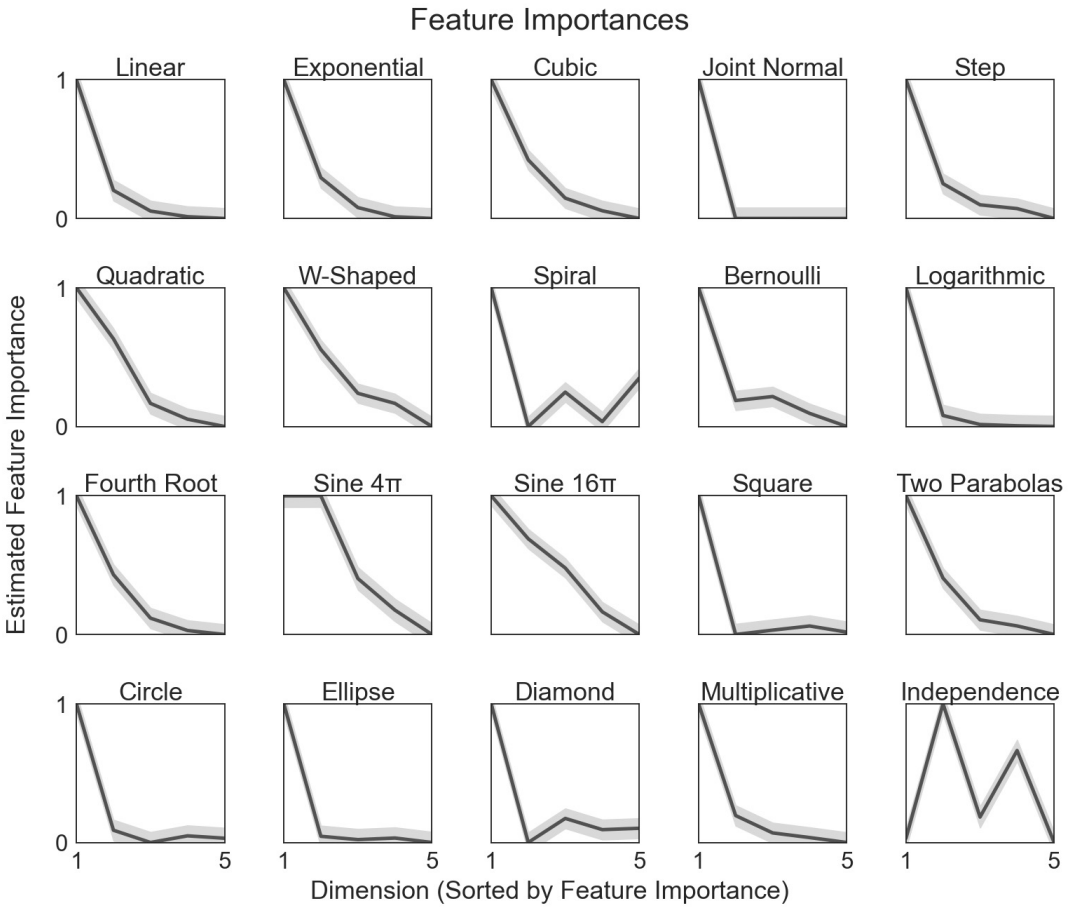
# Two-Sample KMERF vs. Dimension



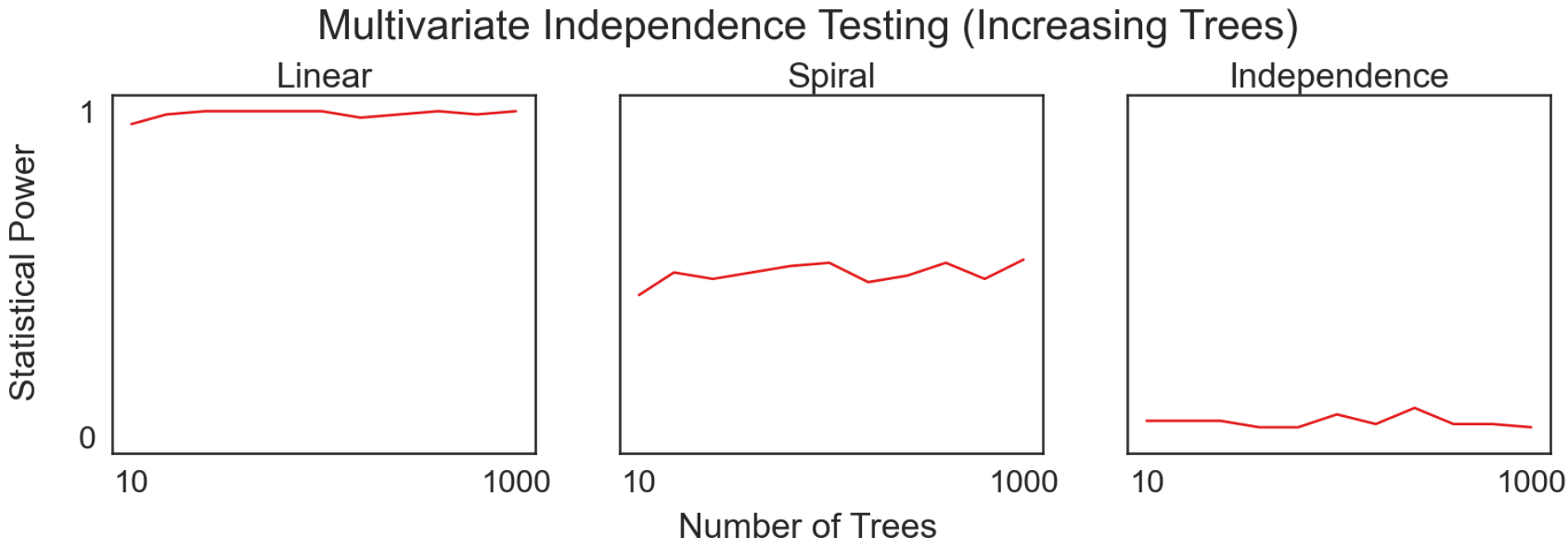
# Two-Sample KMERF vs. Sample Size



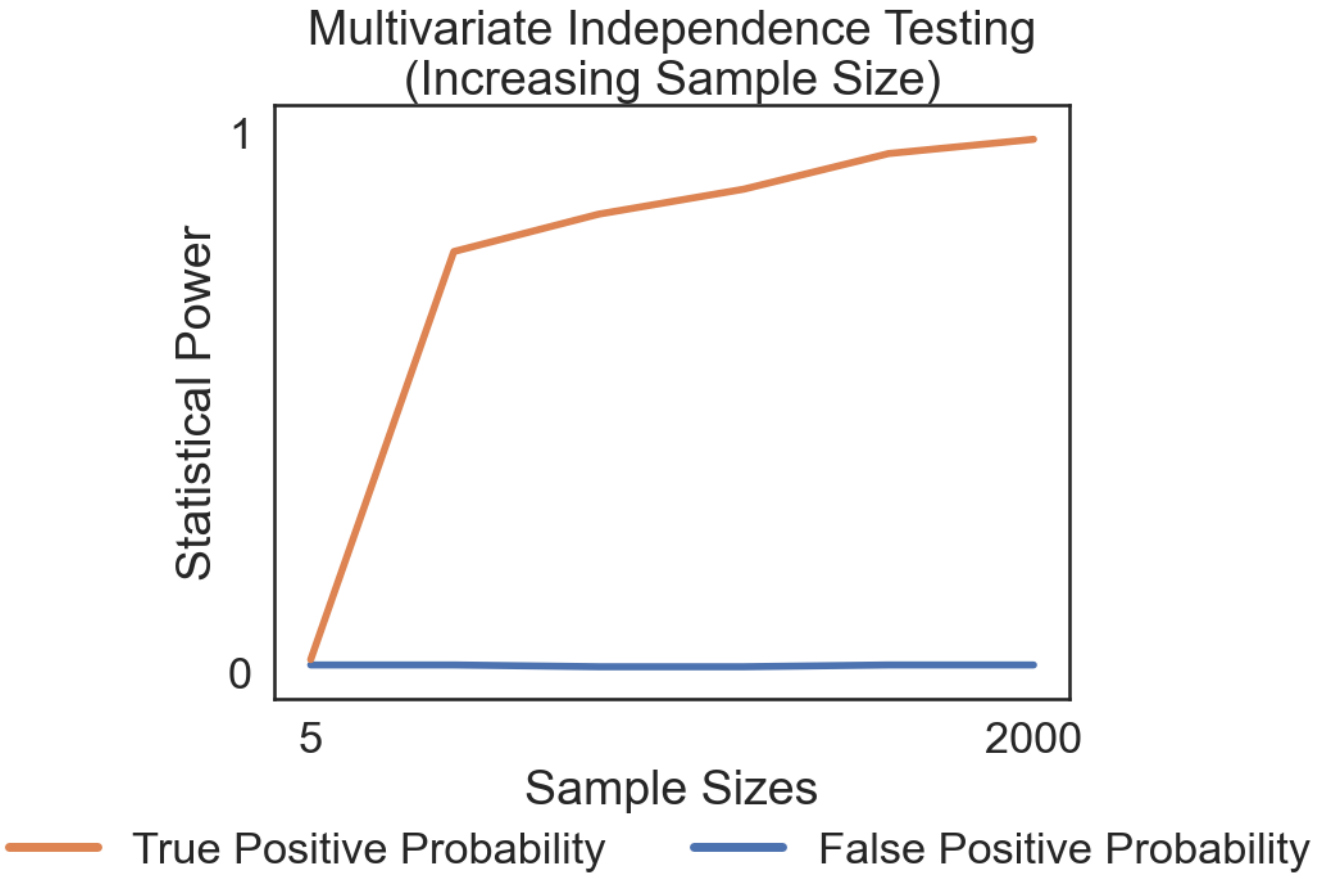
# KMERF correctly determines feature importance



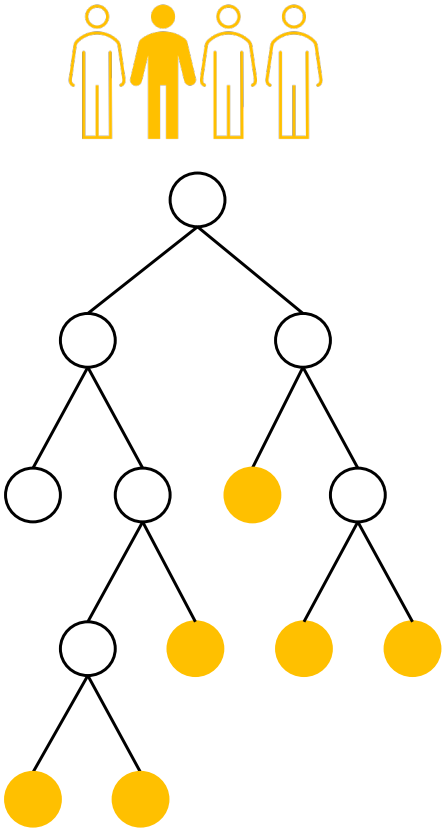
# KMERF is relatively invariant to the # of trees



# KMERF correctly controls False Discovery Rate



# **Step 4**: Compute posteriors and test statistics



■ ■ ■

**Compute Test Statistics**

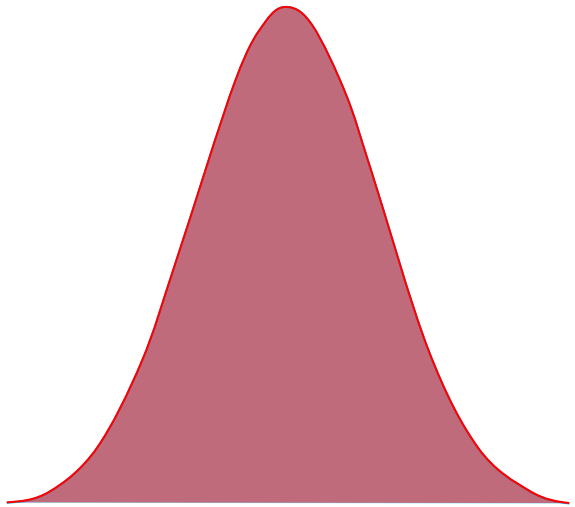
1. Classification Accuracy (Acc)
2. **Mutual information (MI)**
3. **Area under the curve (AUC)**
4. **Sensitivity at k% Specificity (S@k)**

# Mutual Information (MI)

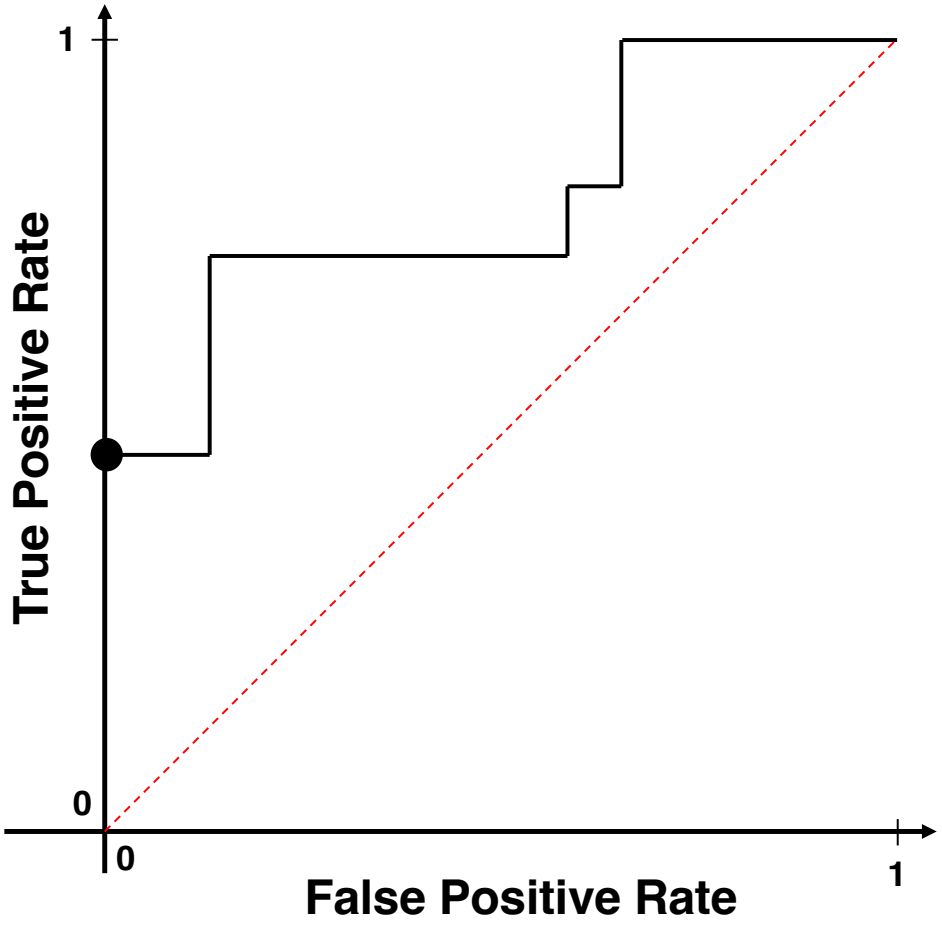
$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

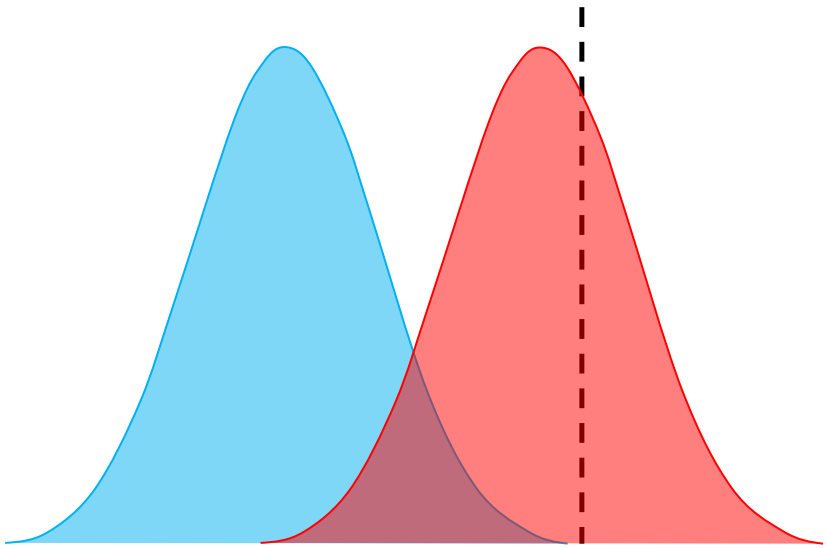
$$I(X;Y) = H(Y) - H(Y|X)$$



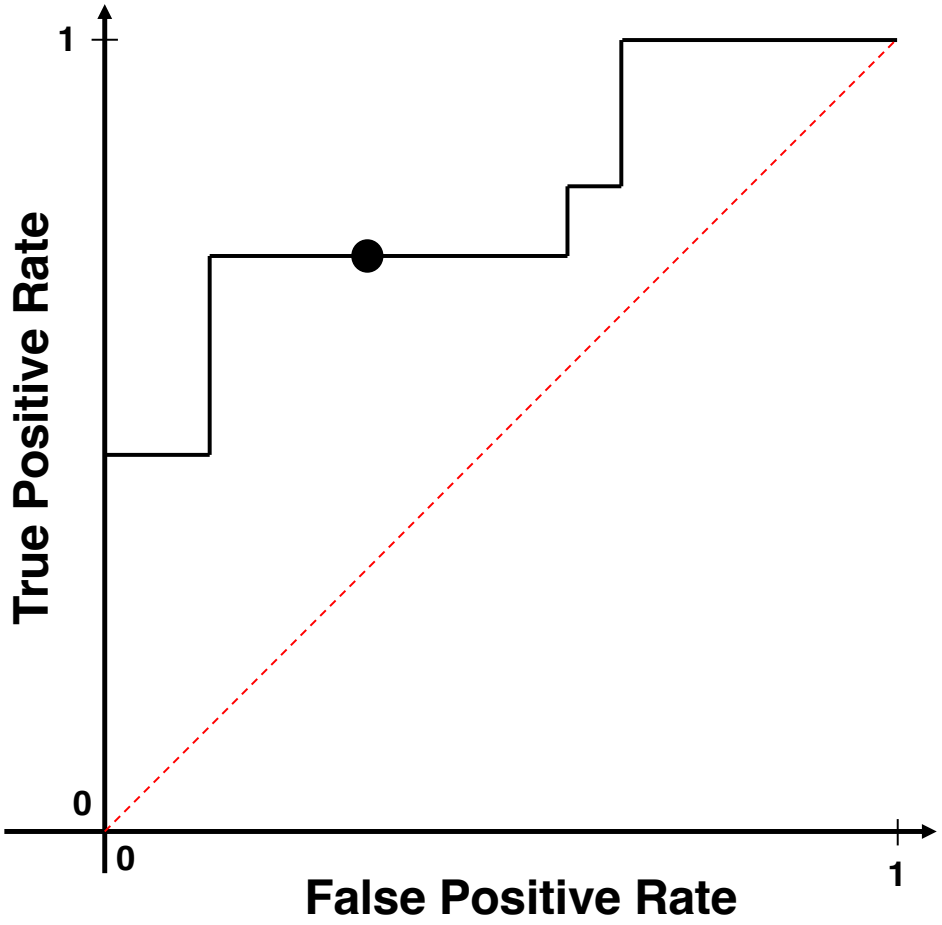
# Sensitivity at k% Sensitivity (S@k)



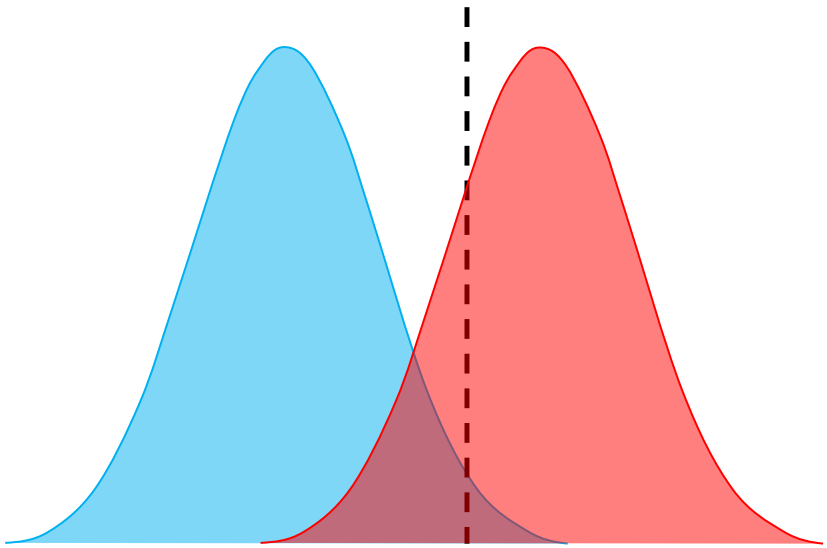

$$S@k = \mathbb{P}\{\eta(X) > T_k | Y = 1\}$$



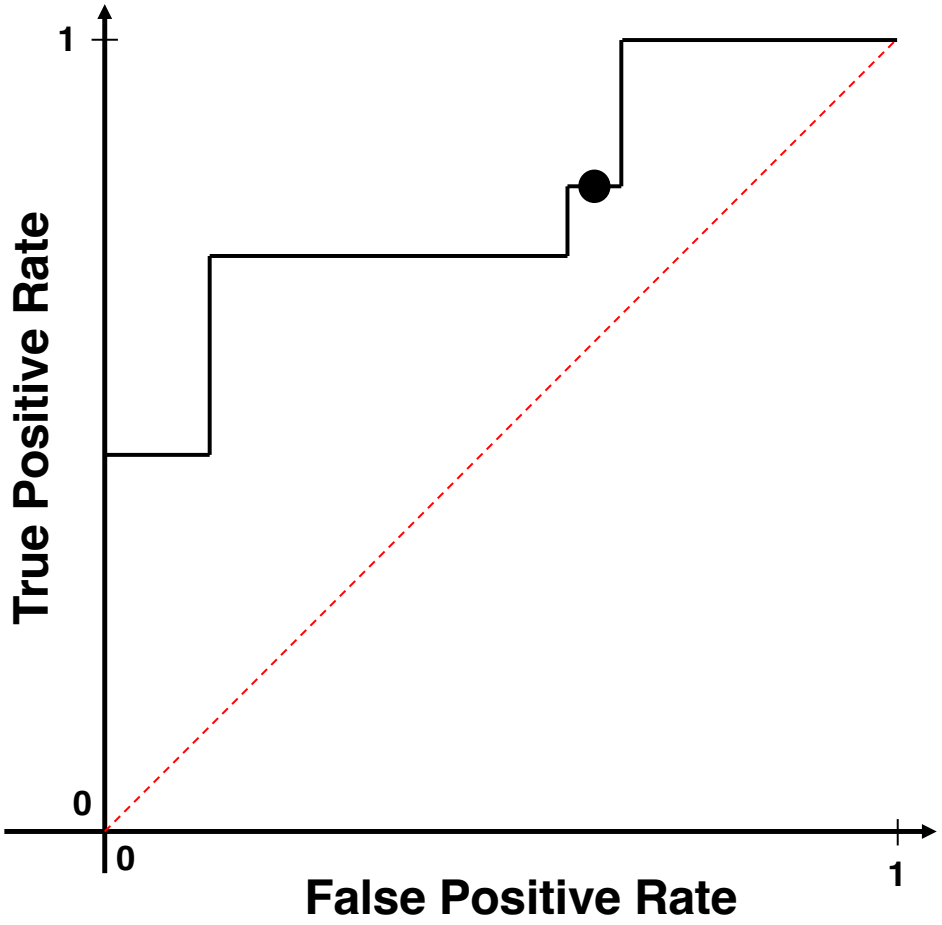
# Sensitivity at k% Sensitivity (S@k)



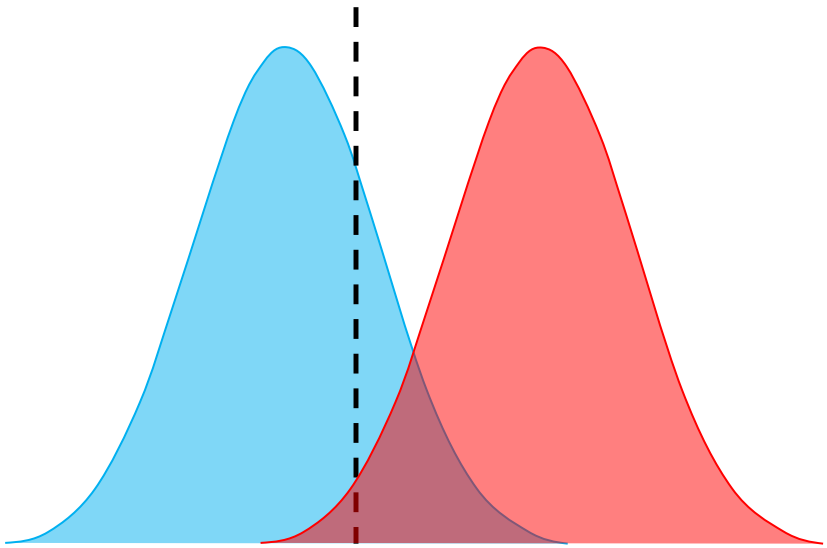

$$S@k = \mathbb{P}\{\eta(X) > T_k | Y = 1\}$$



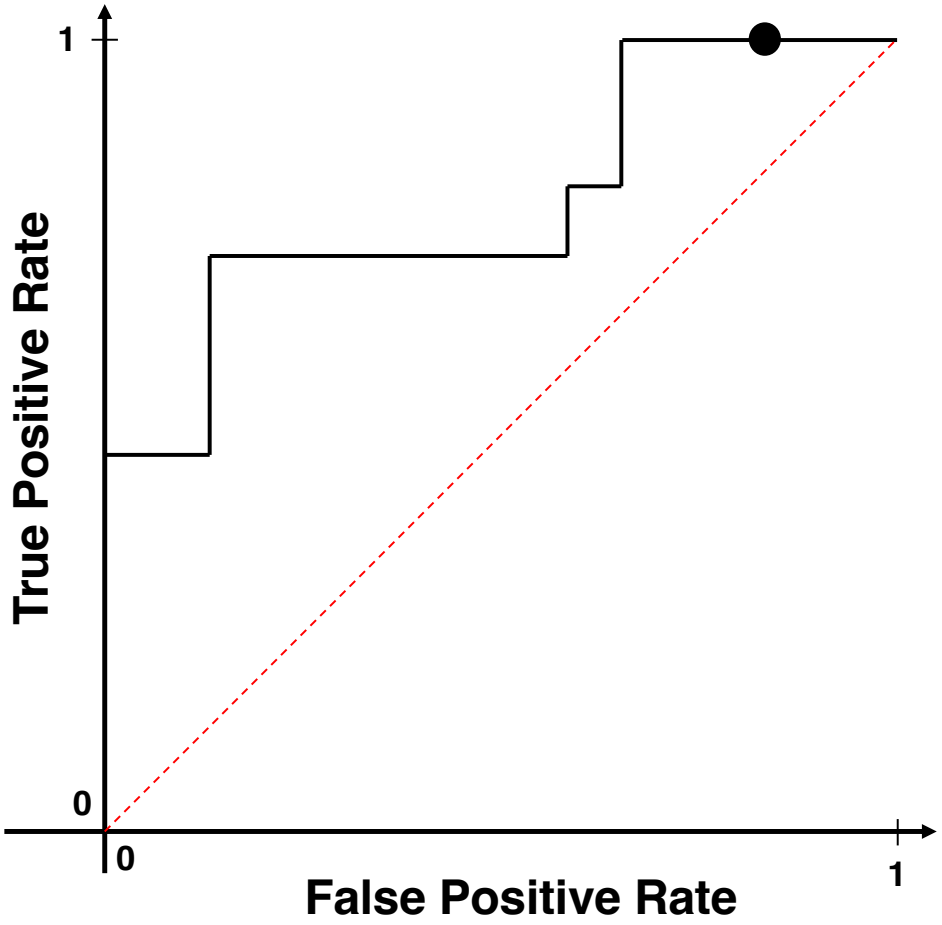
# Sensitivity at k% Sensitivity (S@k)



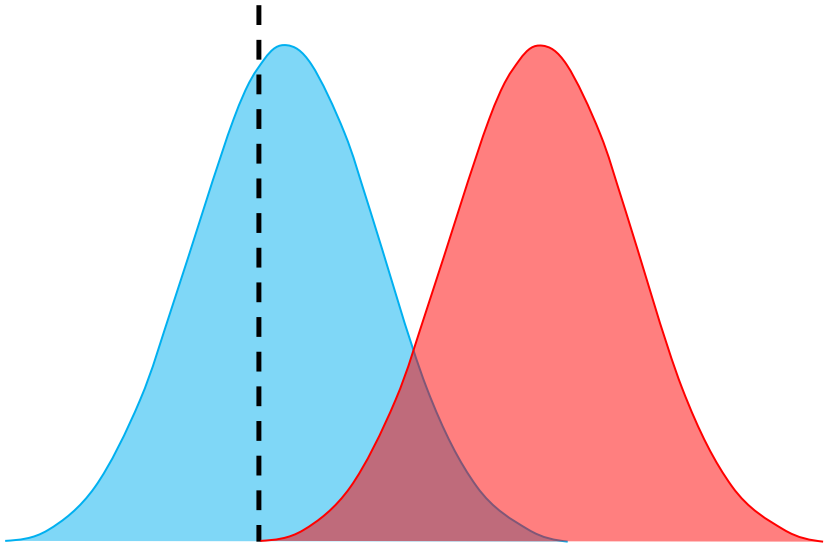

$$S@k = \mathbb{P}\{\eta(X) > T_k | Y = 1\}$$



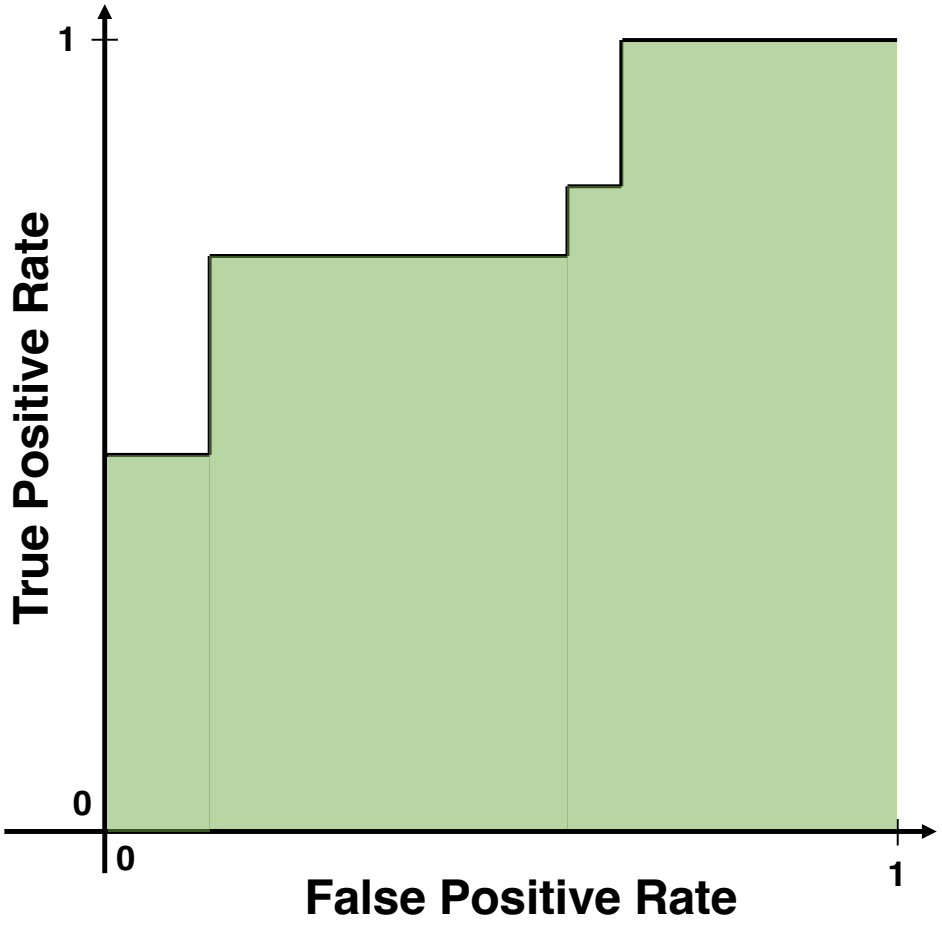
# Sensitivity at k% Sensitivity (S@k)



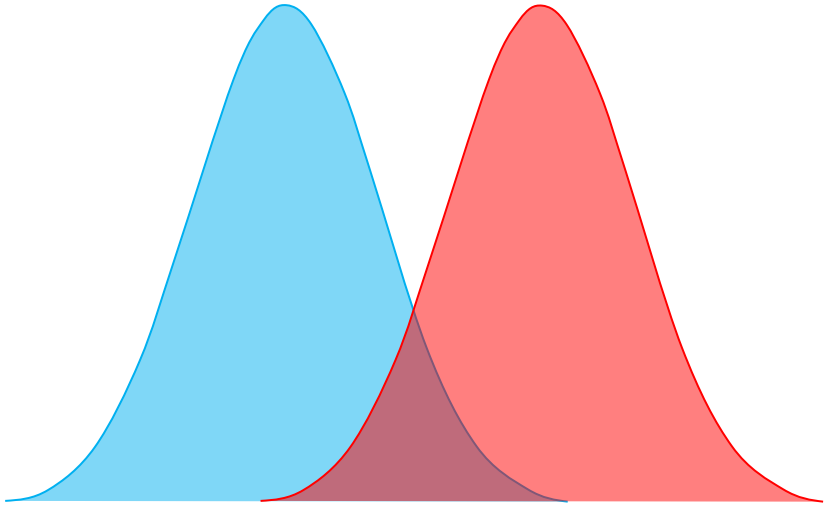

$$S@k = \mathbb{P}\{\eta(X) > T_k | Y = 1\}$$



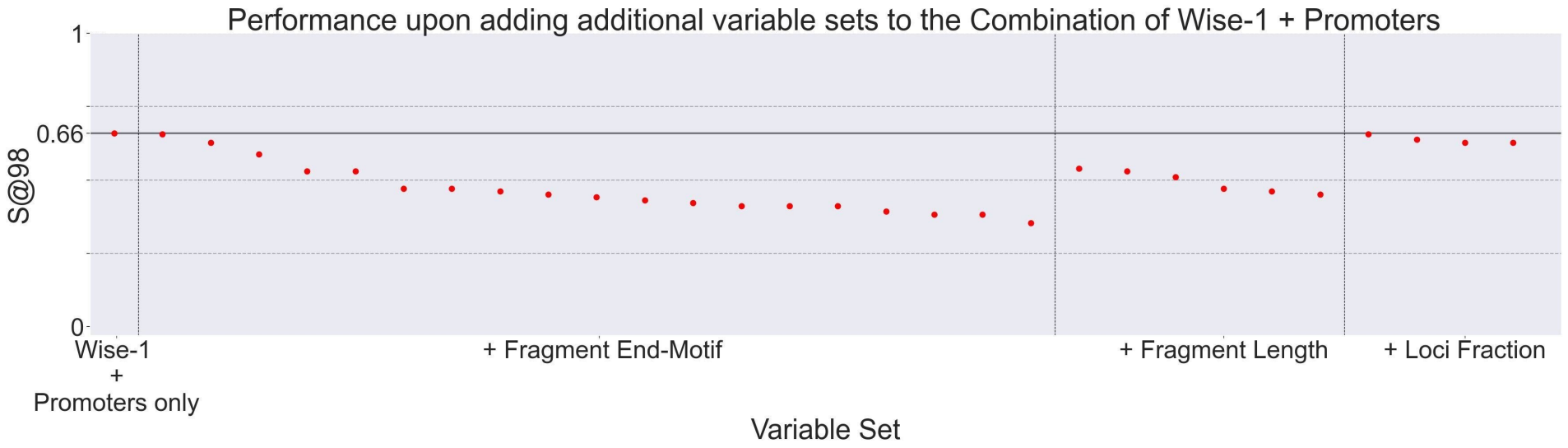
# Area Under the Curve (AUC)



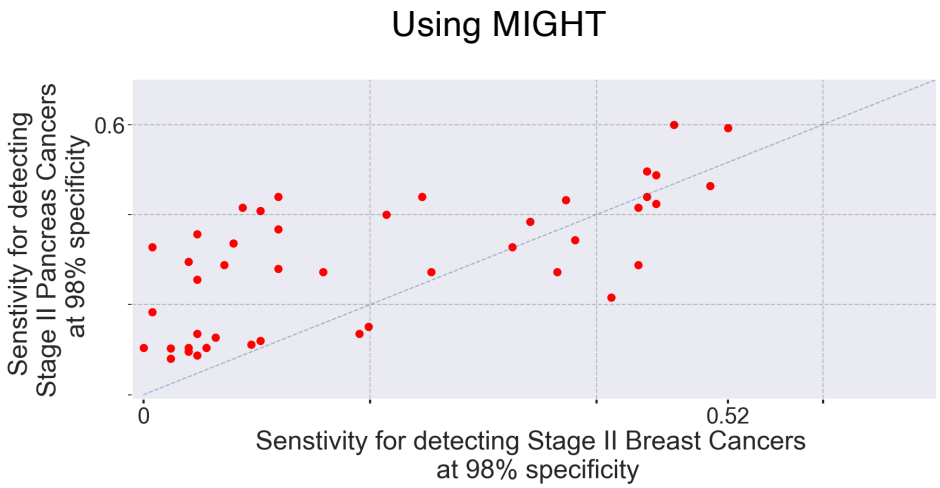

$$AUC = \mathbb{P}\{\eta(X_1) > \eta(X_0)\}$$

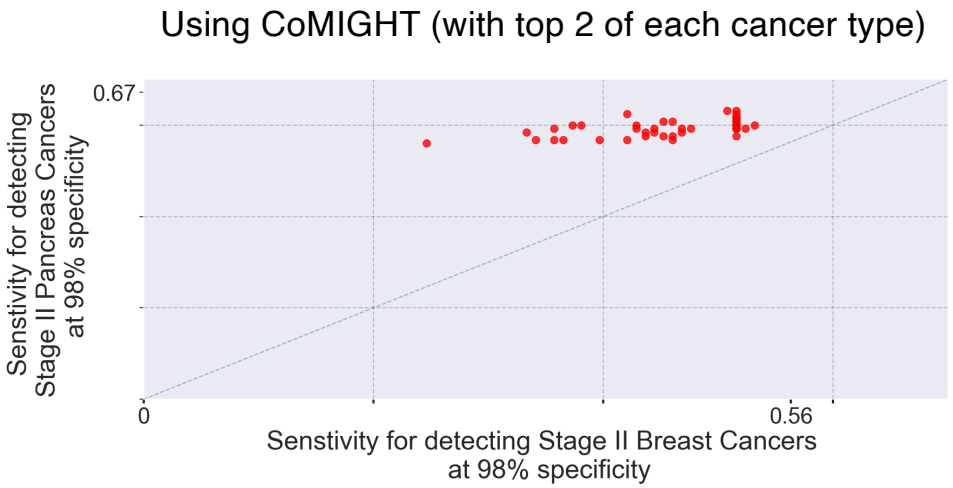


# Variable sets don’t always add information

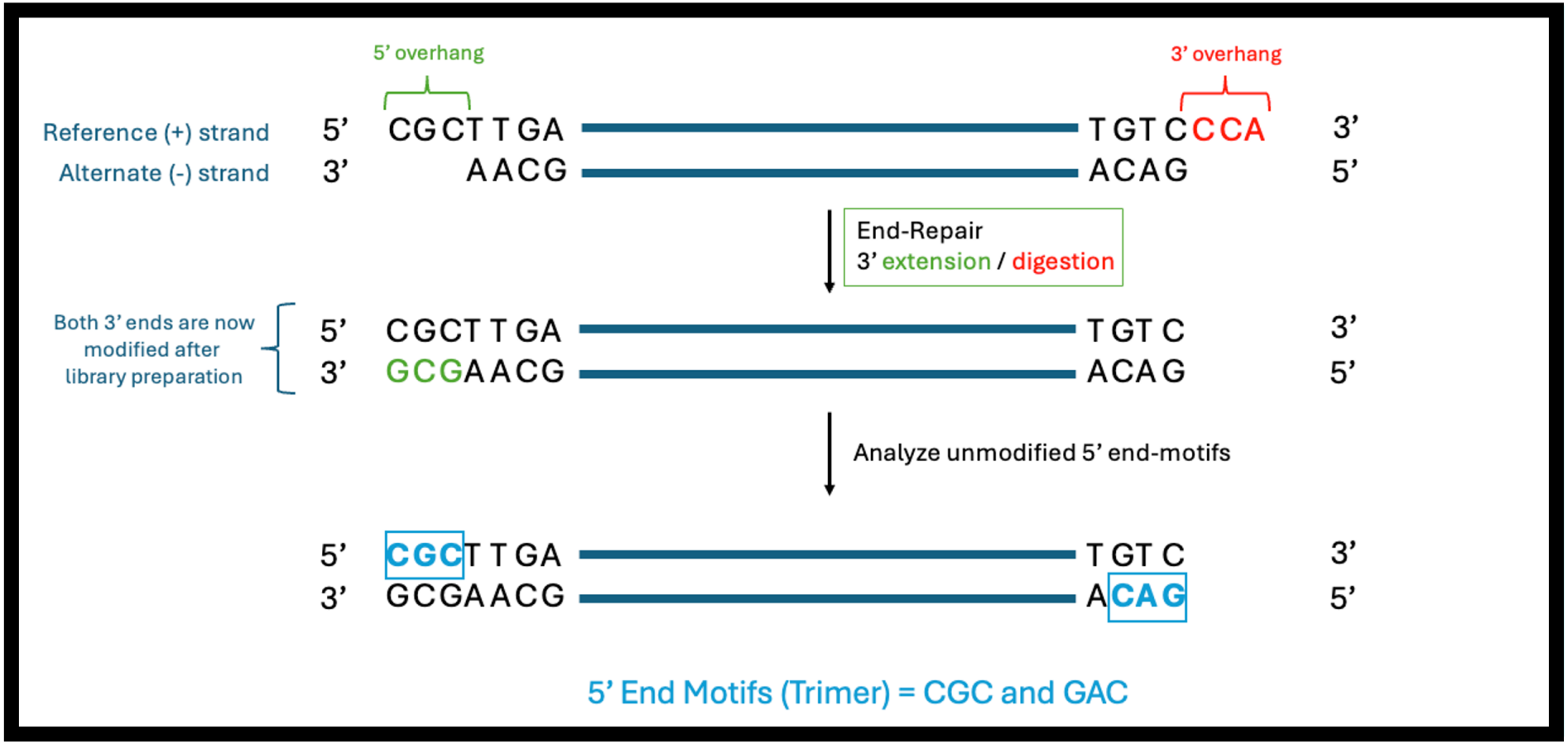


# Less information in breast vs. pancreatic cancer

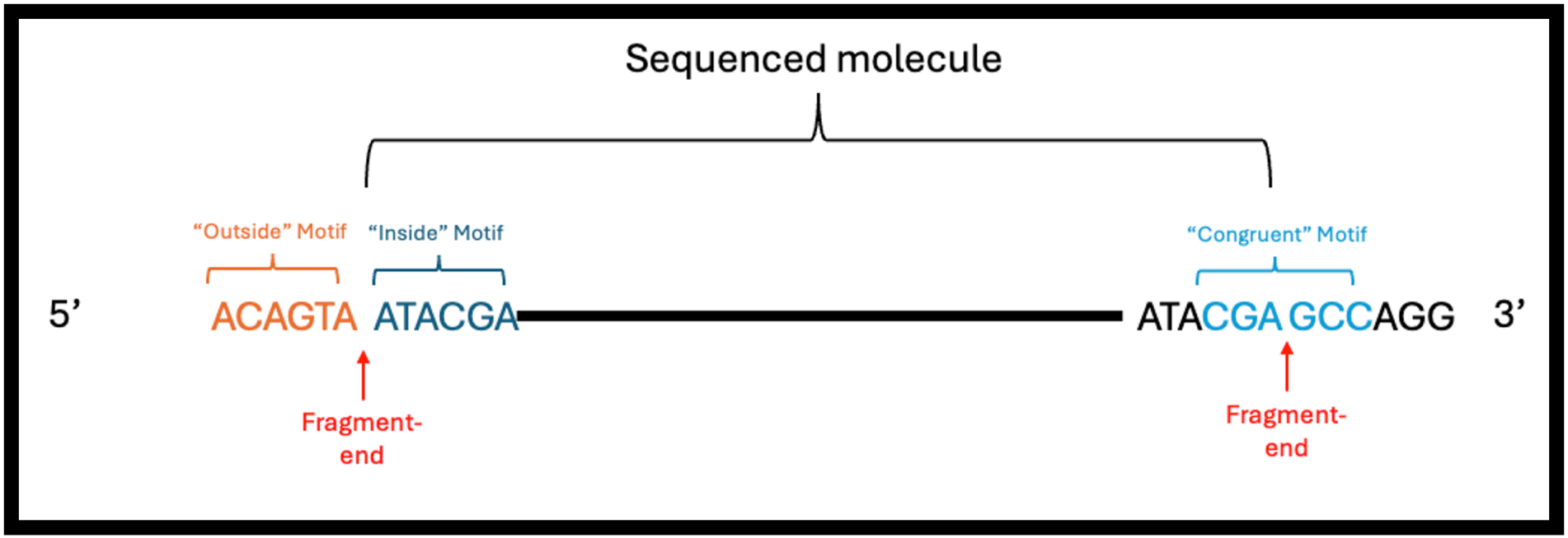


# Definition of Inside End Motifs



# Definition of Outside Motifs



# How breakpoint ratios are calculated