Learning Where to Look:
Reliable Certificates Under Scarce Ground Truth

Gautam Dasarathy

Arizona State University | Amazon

http://gautamdasarathy.com

CNI Seminar, IISc
November 2025

Learning Where to Look:
Reliable Certificates Under Scarce Ground Truth

Or

Label-efficient Two-Sample Testing

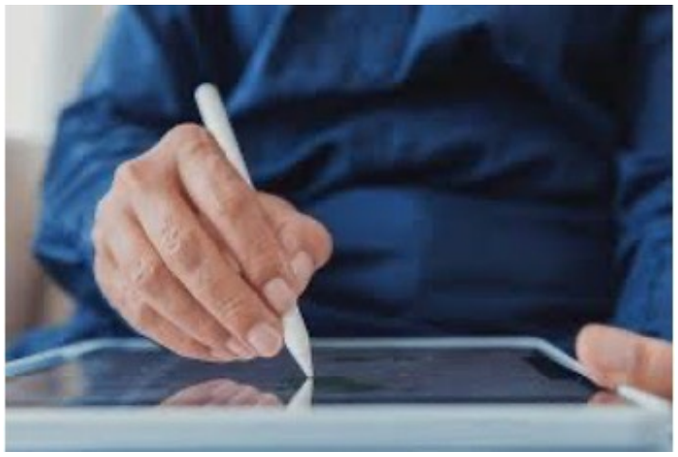
Gautam Dasarathy

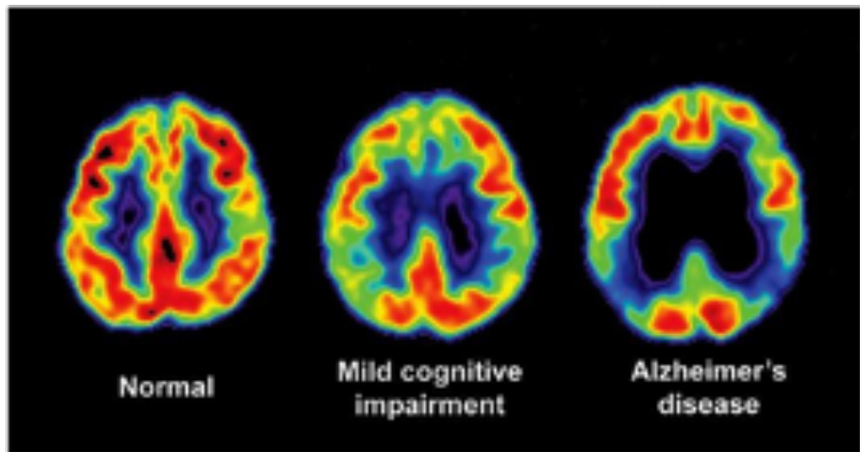
Arizona State University | Amazon

http://gautamdasarathy.com

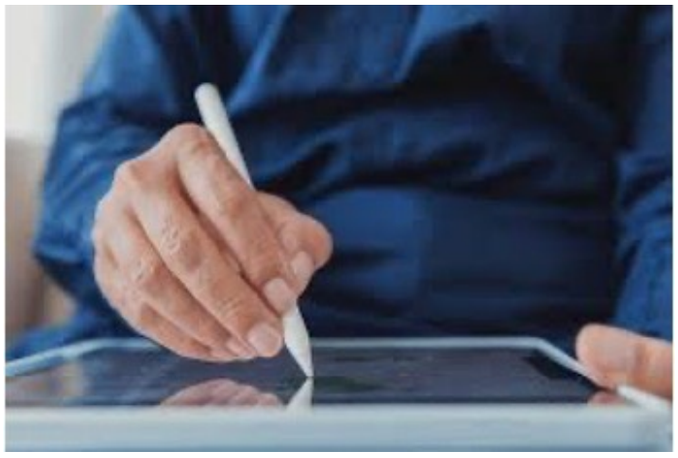
CNI Seminar, IISc
November 2025

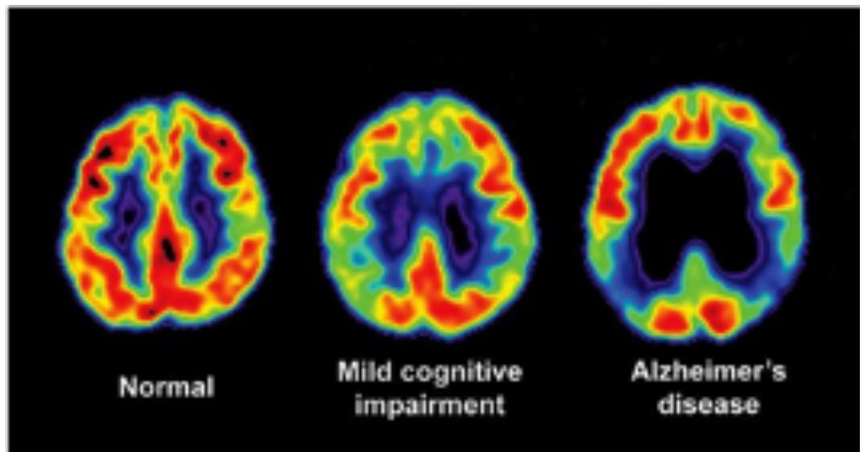
Can Digital Tests Stand in for PET Scans?





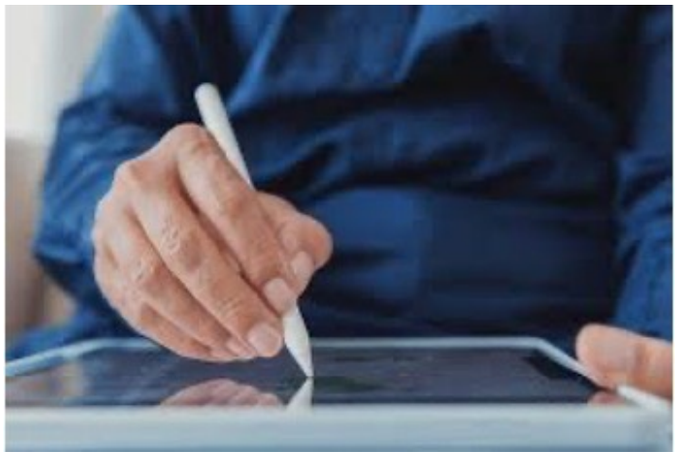
Can Digital Tests Stand in for PET Scans?



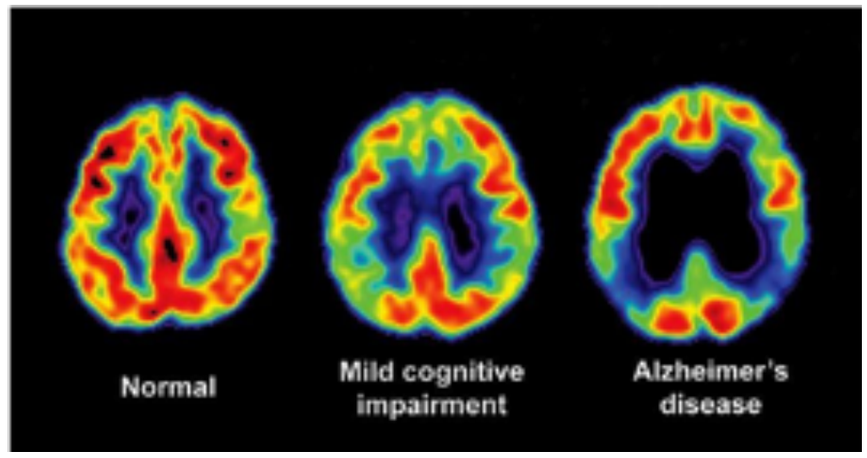


You invent a **new digital test**
for Alzheimer's.

Can Digital Tests Stand in for PET Scans?

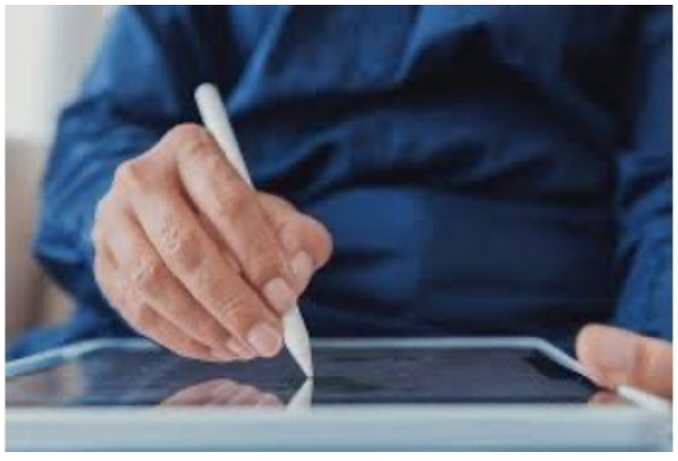


You invent a **new digital test** for Alzheimer's.



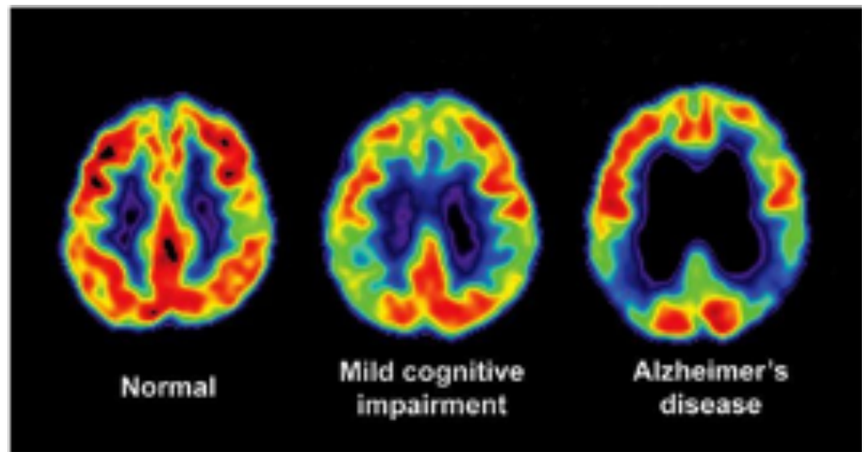
PET scan measuring amyloid build up

Can Digital Tests Stand in for PET Scans?



You invent a **new digital test** for Alzheimer's.

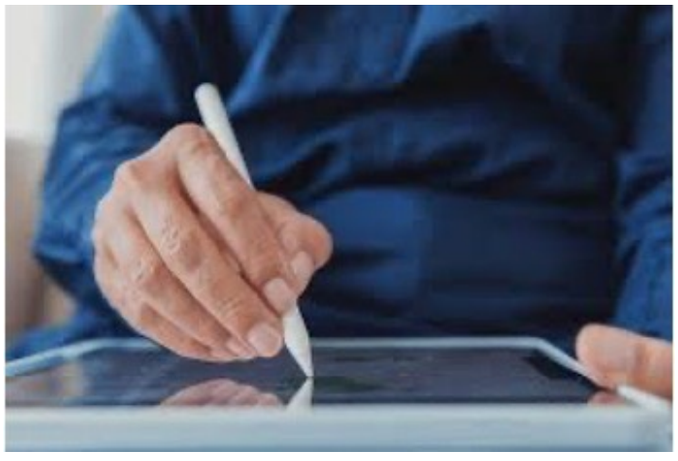
cheap, scalable



PET scan measuring amyloid build up

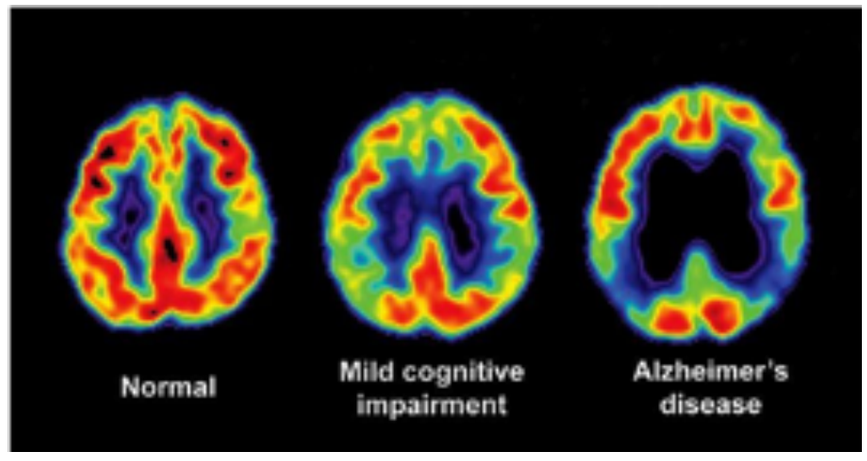
costly, invasive

Can Digital Tests Stand in for PET Scans?



You invent a **new digital test** for Alzheimer's.

cheap, scalable



PET scan measuring amyloid build up

costly, invasive

Is this any good?
Do the **digital test distributions differ** between high- and low-amyloid groups?

Can Digital Tests Stand in for PET Scans?

Can Digital Tests Stand in for PET Scans?

Can Digital Tests Stand in for PET Scans?

high amyloid group

low amyloid group

Can Digital Tests Stand in for PET Scans?

high amyloid group

test results

low amyloid group

Can Digital Tests Stand in for PET Scans?

high amyloid group

test results

Same or different distributions?

low amyloid group

Can Digital Tests Stand in for PET Scans?

high amyloid group

test results

Same or different distributions?

low amyloid group

bad stand in

good stand in

Can Digital Tests Stand in for PET Scans?

high amyloid group

test results

Two-Sample Testing

low amyloid group

bad stand in

good stand in

Can Digital Tests Stand in for PET Scans?

high amyloid group

test results

Two-Sample Testing

low amyloid group

bad stand in

good stand in

Two-sample testing: Given samples $X_1, \ldots, X_m \sim P$ (high-amyloid) and $Y_1, \ldots, Y_n \sim Q$ (low-amyloid), Test: $H_0 : P = Q$ vs $H_1 : P \neq Q$

Two Sample Testing is Everywhere

Two Sample Testing is Everywhere

Digital health sensor validation

Two Sample Testing is Everywhere

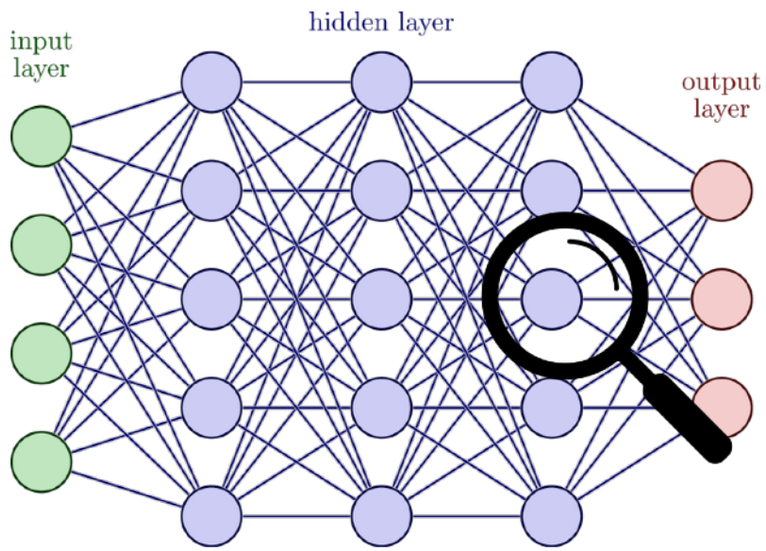
Digital health sensor validation

Fraud intervention efficacy

Two Sample Testing is Everywhere

Digital health sensor validation

Fraud intervention efficacy



Model Monitoring / ML OPs:
data drift relative to training?

2ST: More than a Century of Data-Driven Science

2ST: More than a Century of Data-Driven Science

Gosset (1908)

2ST: More than a Century of Data-Driven Science

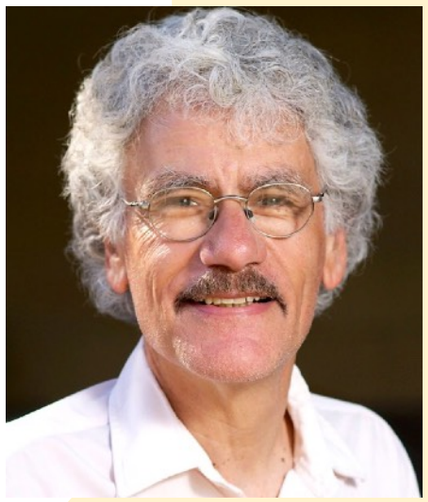
Gosset (1908)

Wald-Wolfowitz (1940s)

2ST: More than a Century of Data-Driven Science

Gosset (1908)

Wald-Wolfowitz (1940s)



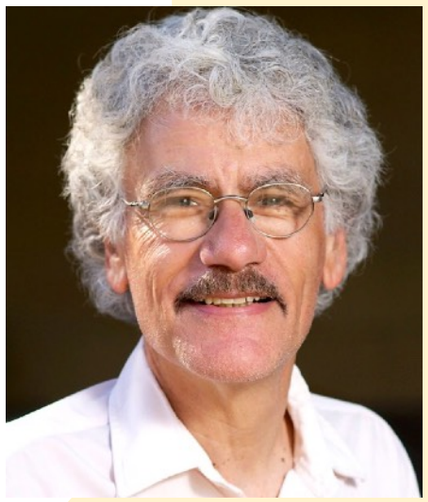
Friedman-Rafsky (1979)

2ST: More than a Century of Data-Driven Science

Gosset (1908)

Modern era

Wald-Wolfowitz (1940s)



Friedman-Rafsky (1979)

Setting Up The Two Sample Testing Problem

Setting Up The Two Sample Testing Problem

Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

Setting Up The Two Sample Testing Problem

Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

$$H_0 : P = Q$$
$$H_1 : P \neq Q$$

Setting Up The Two Sample Testing Problem

Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

$$H_0 : P = Q$$
$$H_1 : P \neq Q$$

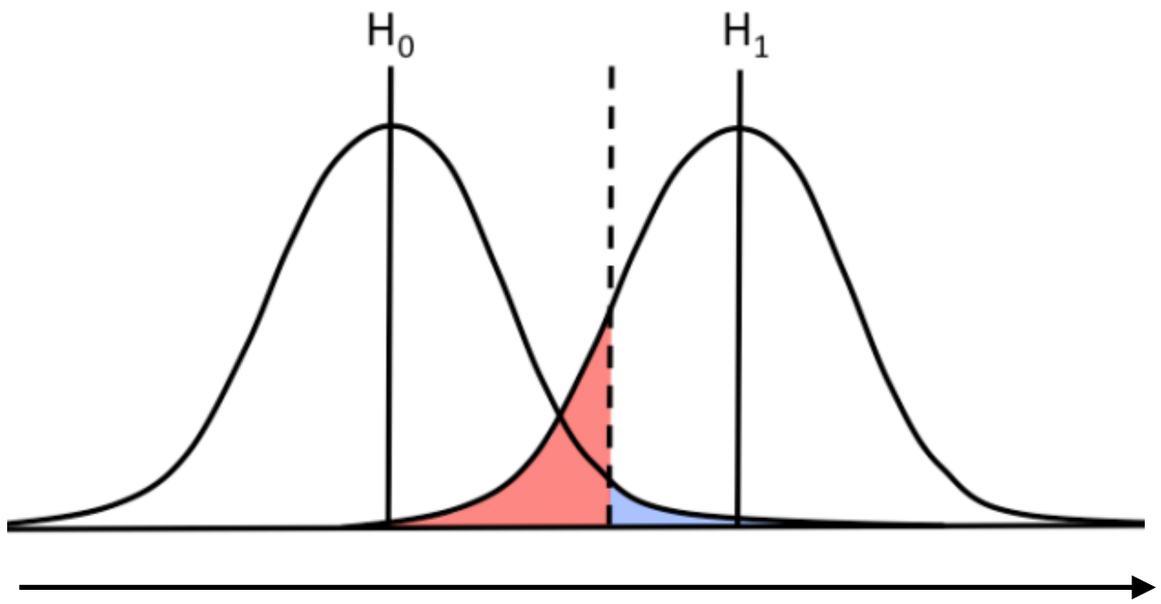
We usually compute a **statistic from the data**, and reject H_0 if the value is *too extreme*

Setting Up The Two Sample Testing Problem

Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

$$H_0 : P = Q$$
$$H_1 : P \neq Q$$

We usually compute a **statistic from the data**, and reject H_0 if the value is *too extreme*



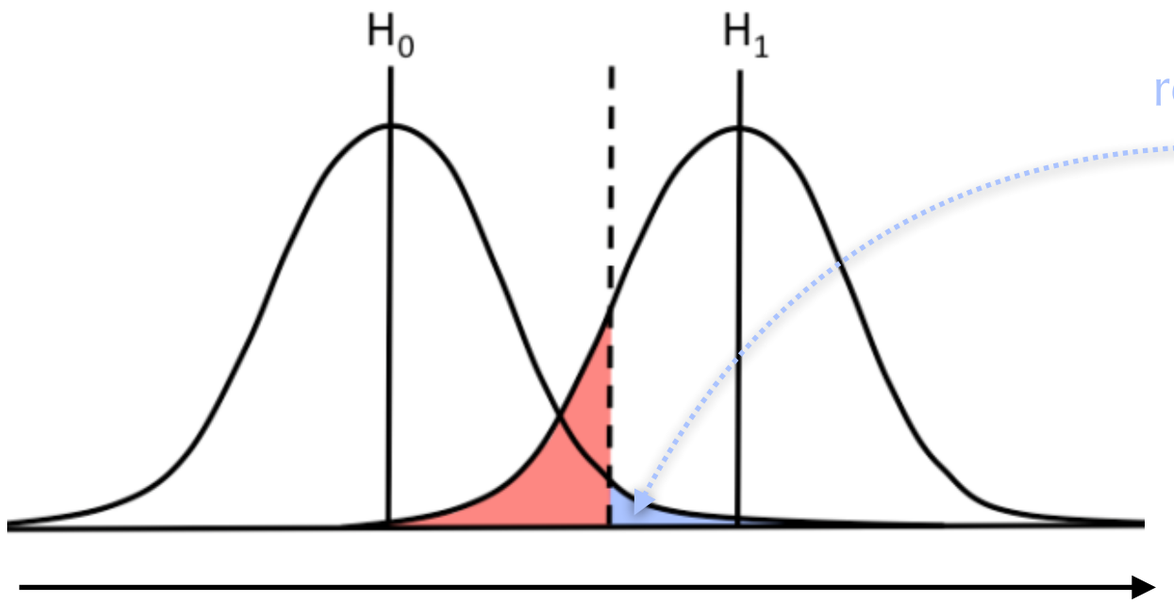
test statistic distribution under
H_0 and H_1

Setting Up The Two Sample Testing Problem

Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

$$H_0 : P = Q$$
$$H_1 : P \neq Q$$

We usually compute a **statistic from the data**, and reject H_0 if the value is *too extreme*



Type I error (α)
reject H_0 when it is true

Setting Up The Two Sample Testing Problem

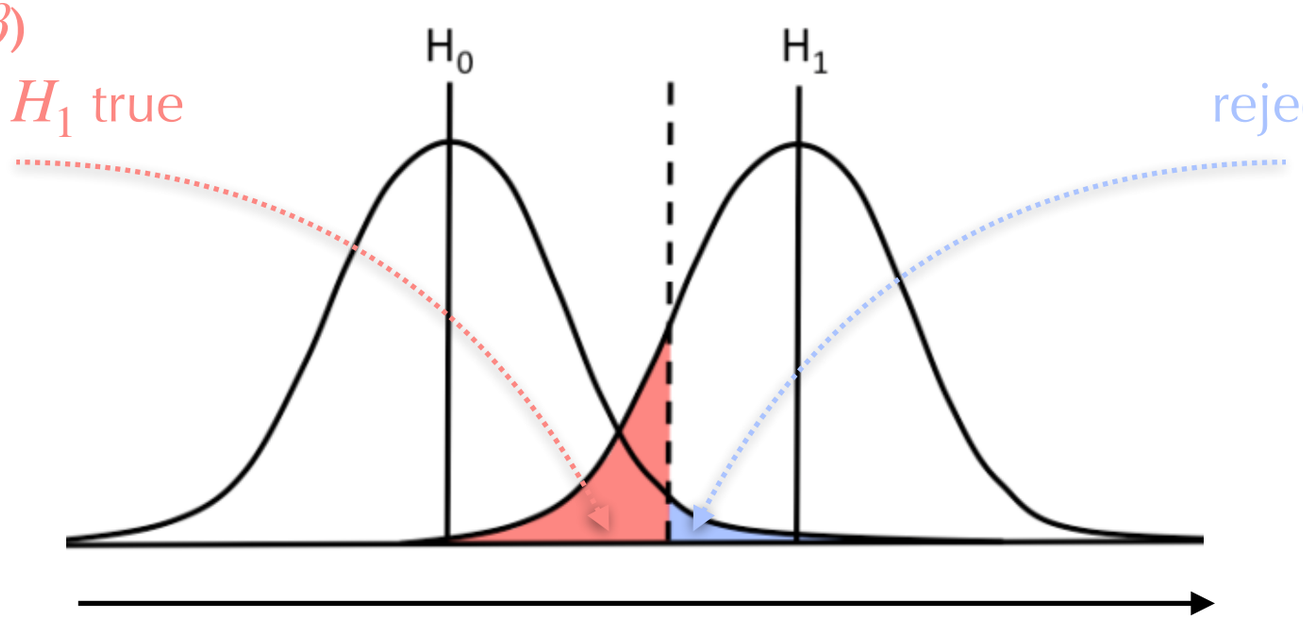
Given: $X_1, \ldots, X_m \sim P$ and $Y_1, \ldots, Y_n \sim Q$ (iid). Perform the following **hypothesis test**

$$H_0 : P = Q$$
$$H_1 : P \neq Q$$

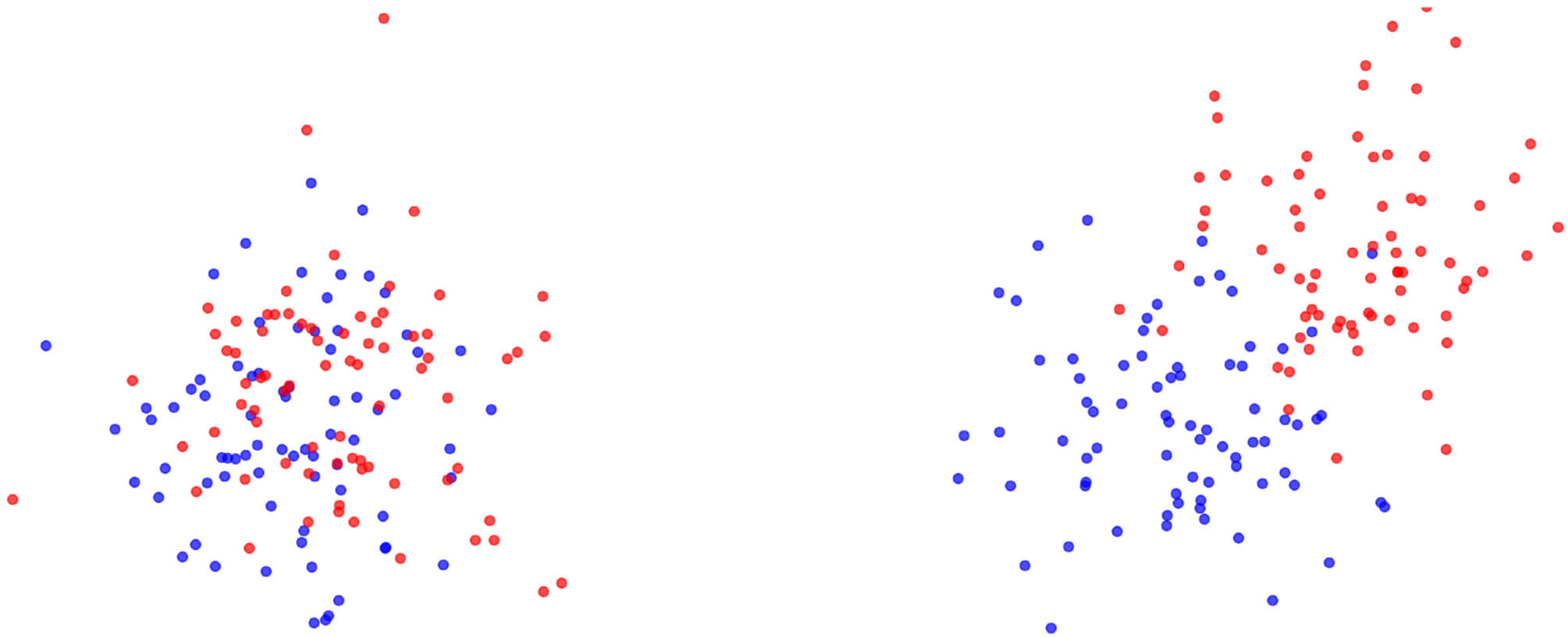
We usually compute a **statistic from the data**, and reject H_0 if the value is *too extreme*

Type II error (β)
fail to reject H_0 when H_1 true

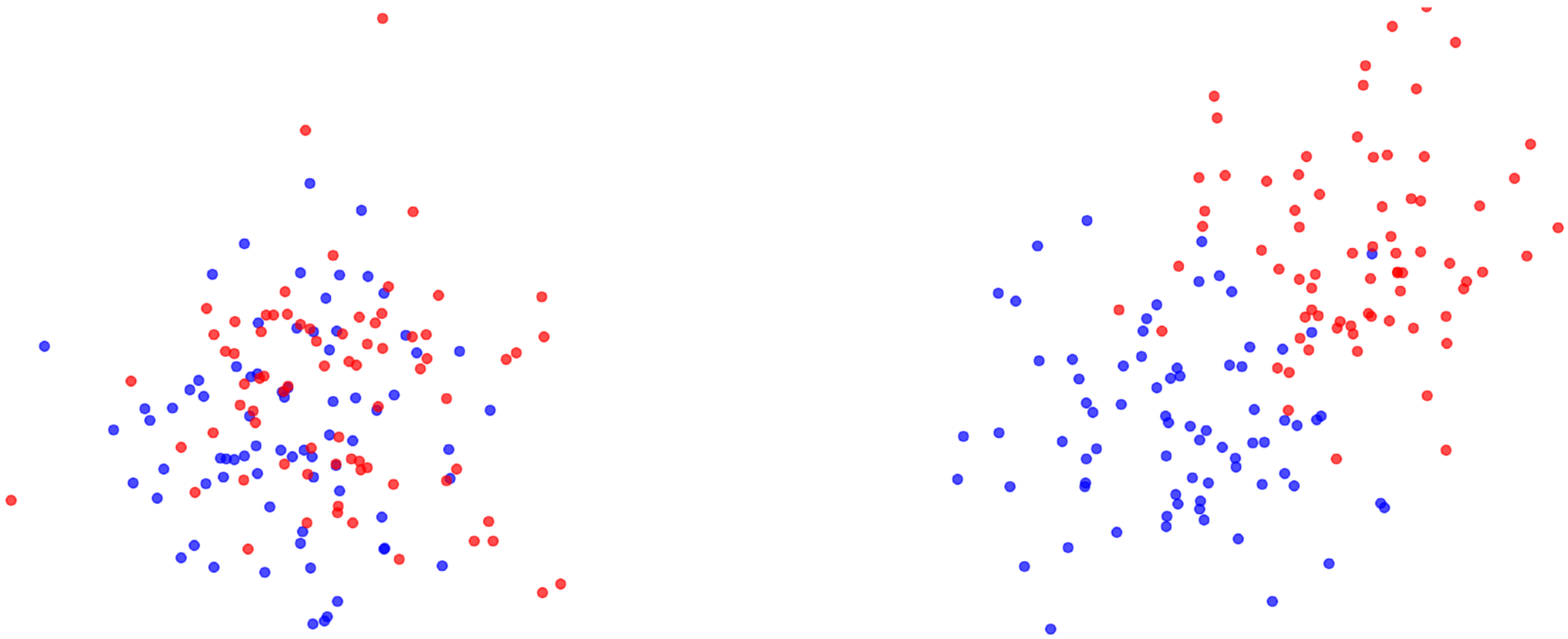
Type I error (α)
reject H_0 when it is true



Classical Two Sample Test in Action: Friedman-Rafsky

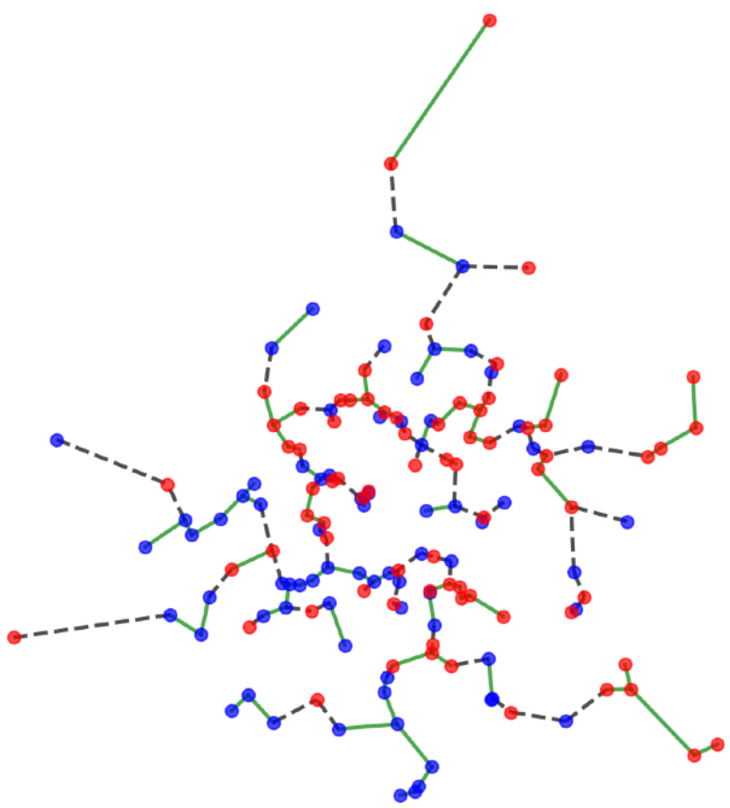


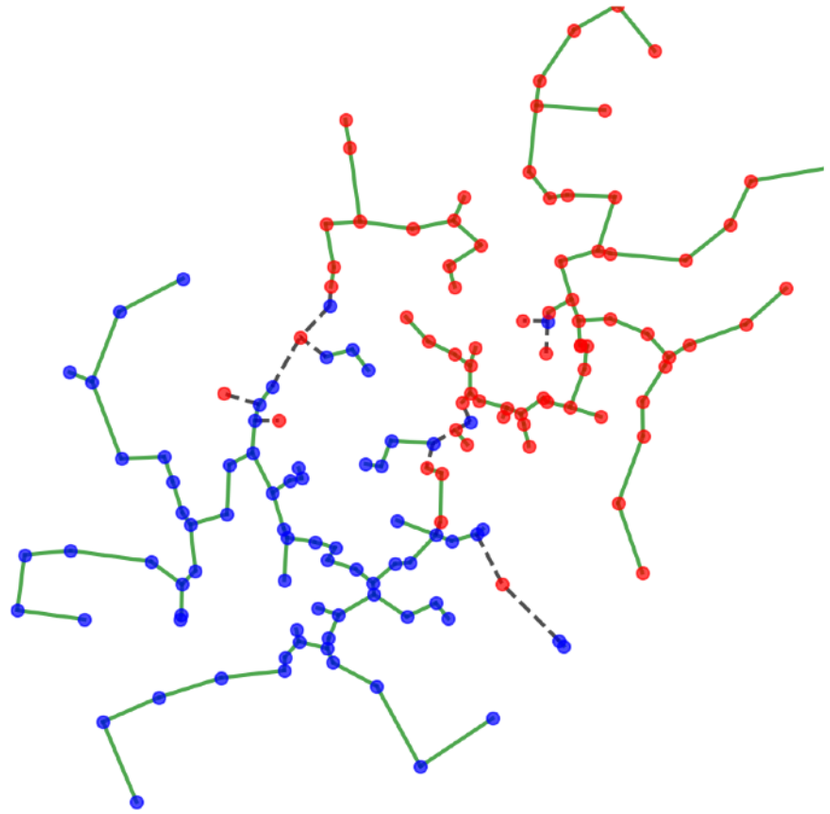
Classical Two Sample Test in Action: Friedman-Rafsky



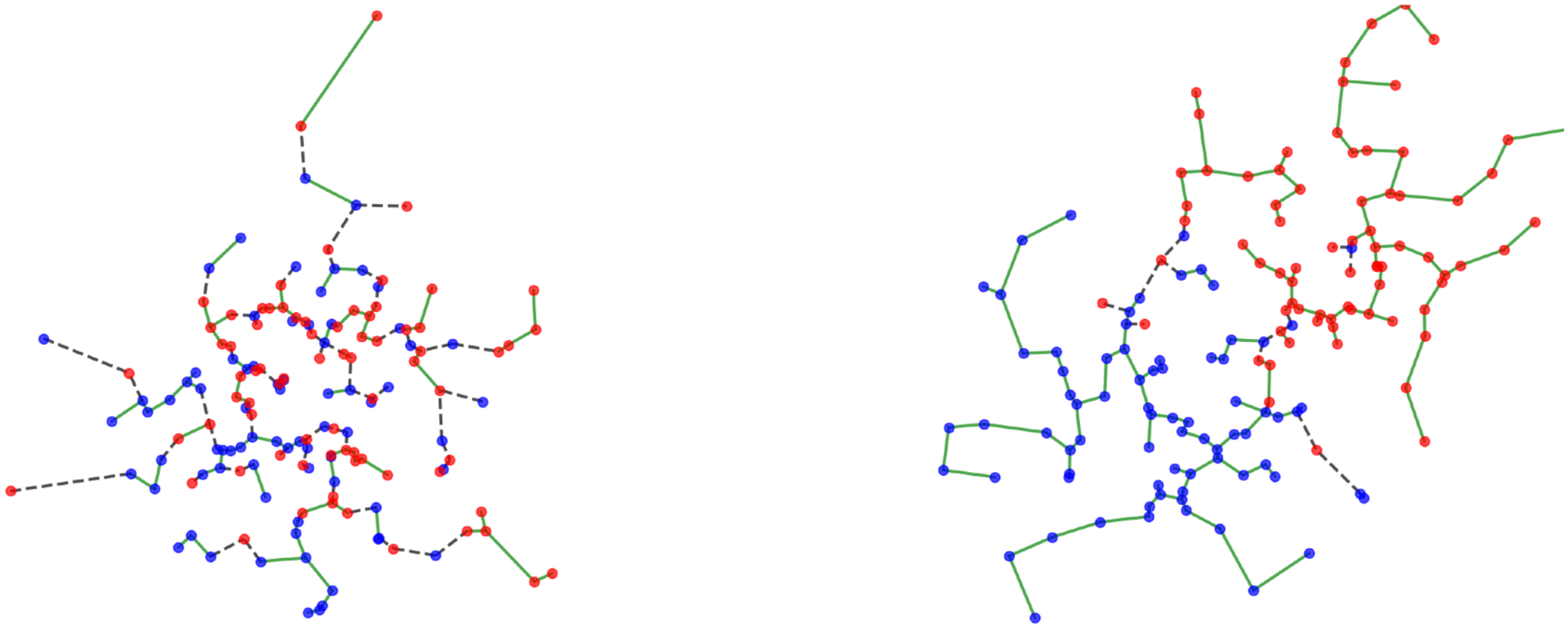
Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.

Classical Two Sample Test in Action: Friedman-Rafsky



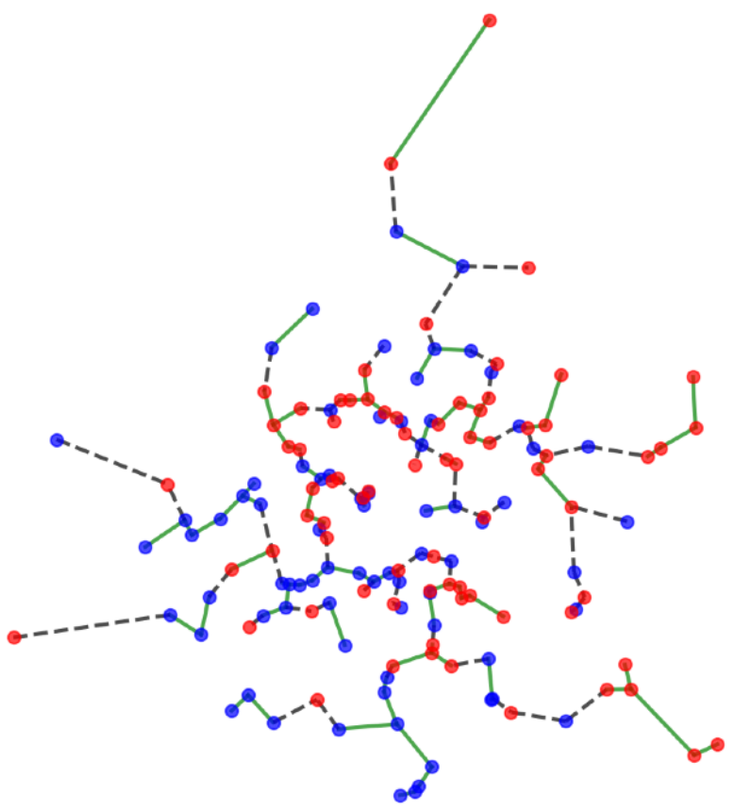


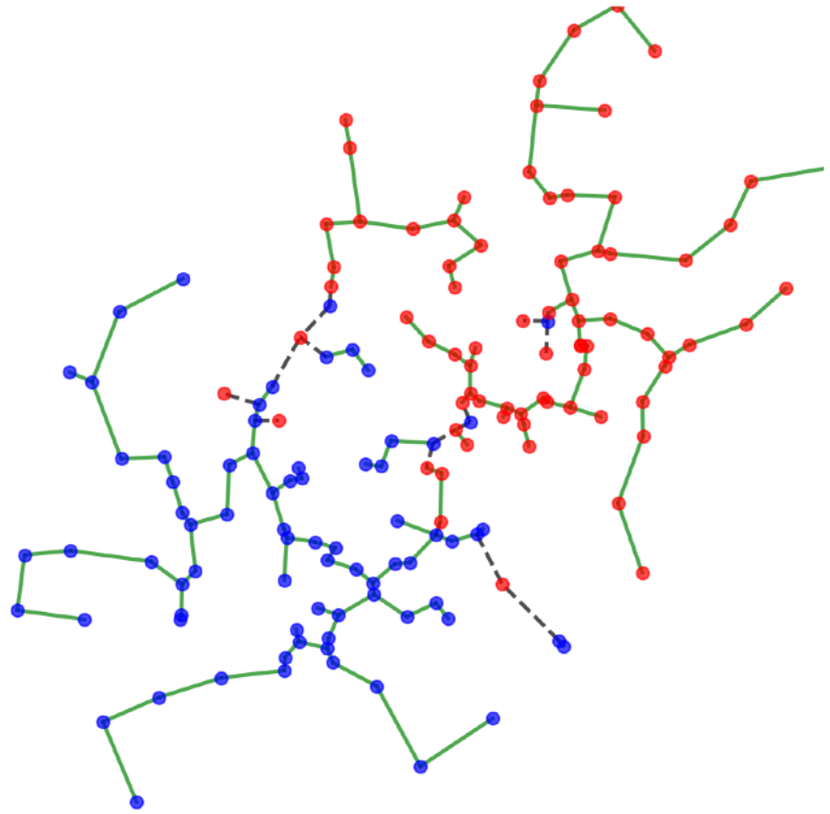
Classical Two Sample Test in Action: Friedman-Rafsky



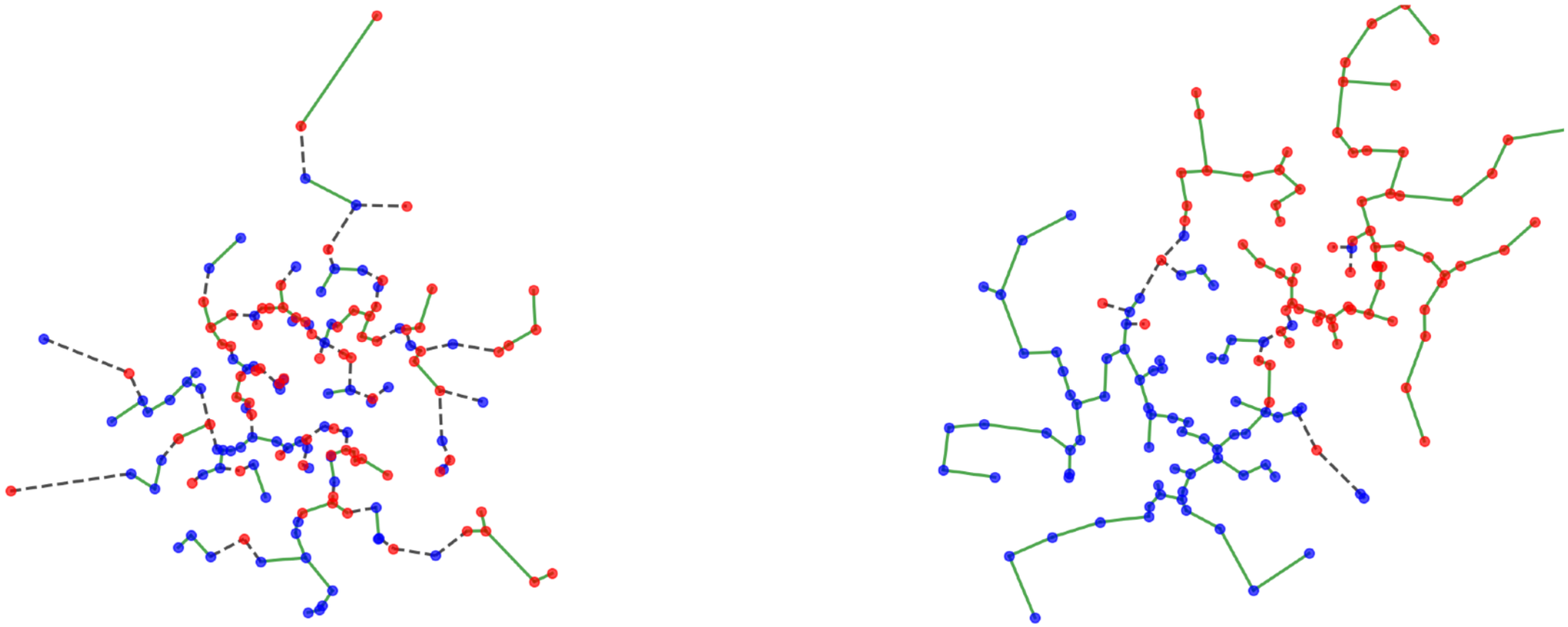
Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.

Classical Two Sample Test in Action: Friedman-Rafsky



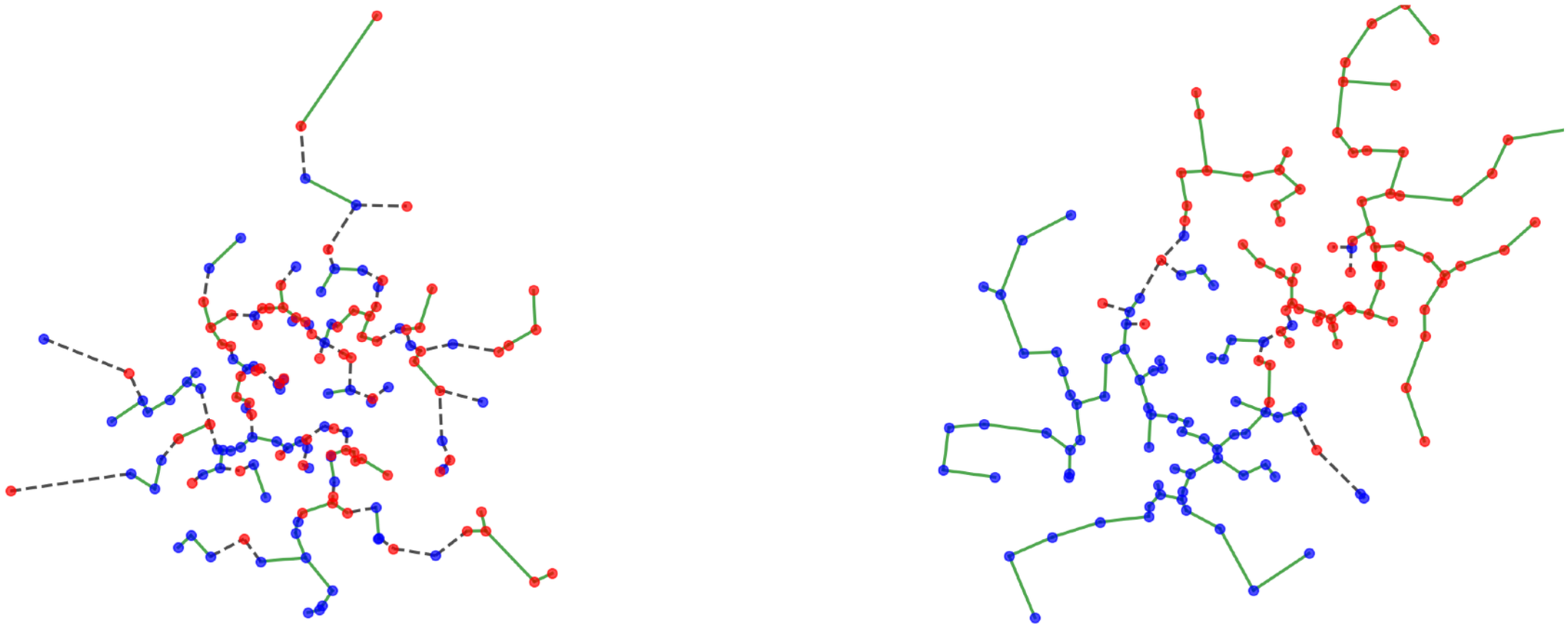


Classical Two Sample Test in Action: Friedman-Rafsky



Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.
Henze, Norbert, and Mathew D. Penrose. "On the multivariate runs test." Annals of statistics (1999): 290-298.

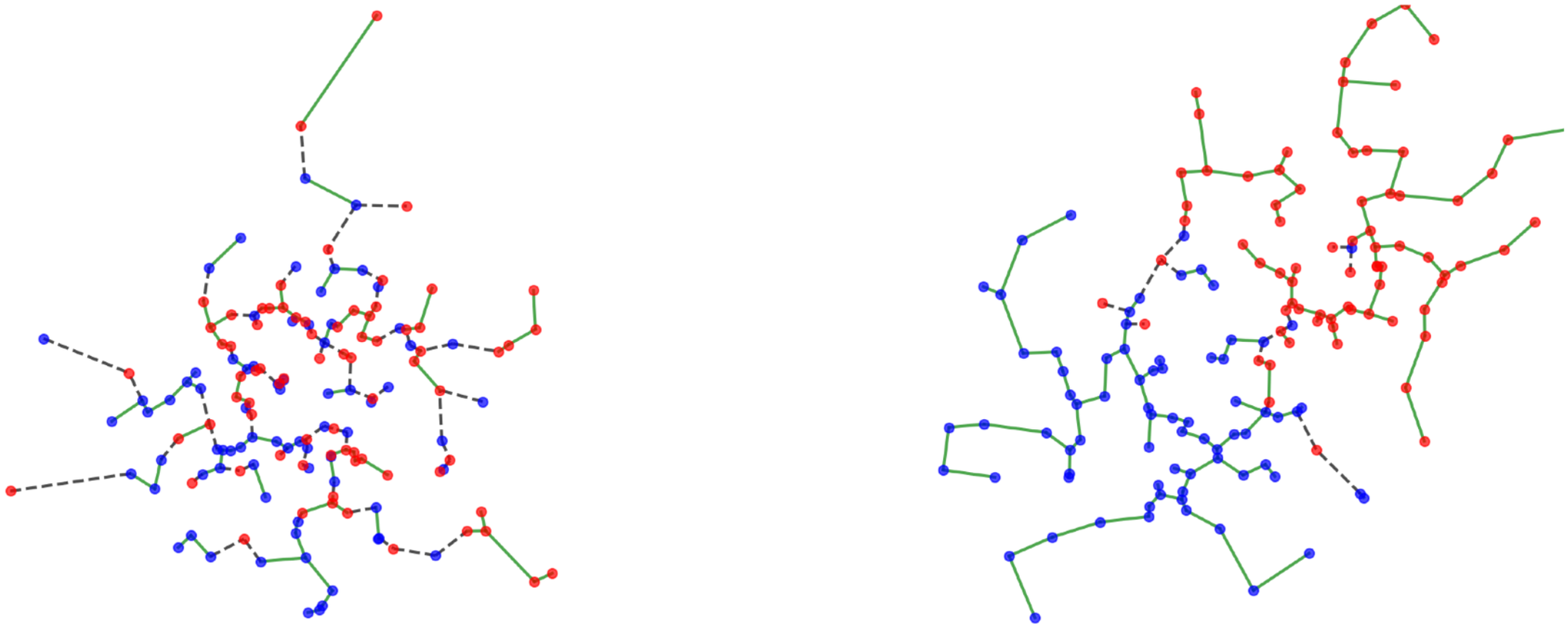
Classical Two Sample Test in Action: Friedman-Rafsky



Theorem 1 (FR `79) The normalized cut-edge count R is **asymptotically normal** under H_0 ; it's mean and variance have analytical expressions —> can construct a ***permutation test***

Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.
Henze, Norbert, and Mathew D. Penrose. "On the multivariate runs test." Annals of statistics (1999): 290-298.

Classical Two Sample Test in Action: Friedman-Rafsky

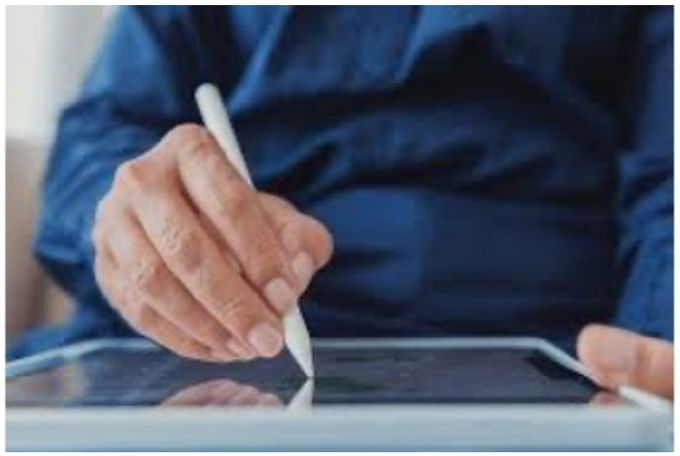


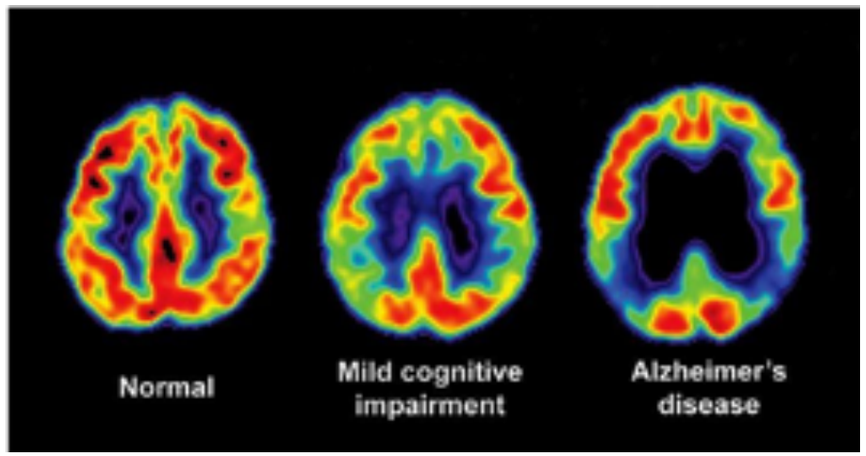
Theorem 1 (FR `79) The normalized cut-edge count R is **asymptotically normal** under H_0 ; it's mean and variance have analytical expressions —> can construct a ***permutation test***

Theorem 2 (HP `99) The FR test is **consistent.** In particular $R/m+n \to c\left[1-D_f\left(P\|Q\right)\right]$

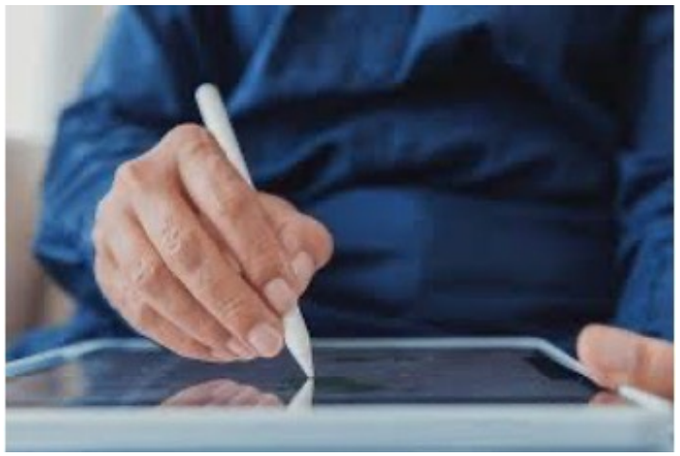
Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.
Henze, Norbert, and Mathew D. Penrose. "On the multivariate runs test." Annals of statistics (1999): 290-298.

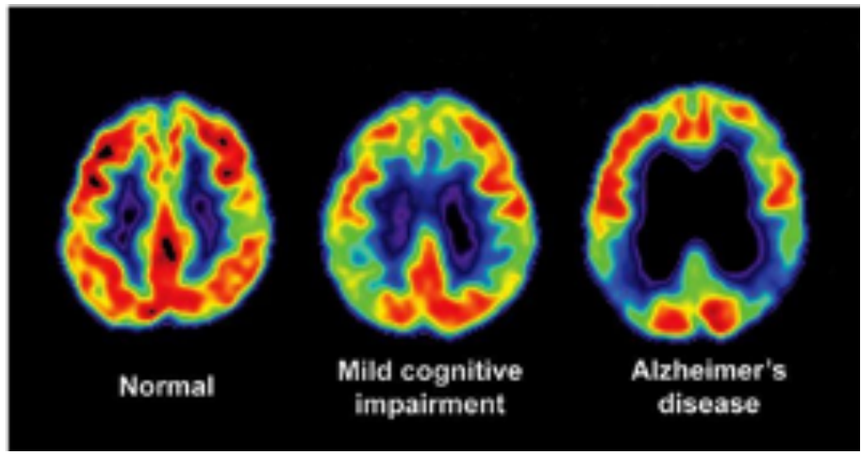
The Catch:
Group Memberships are Often Expensive to Determine





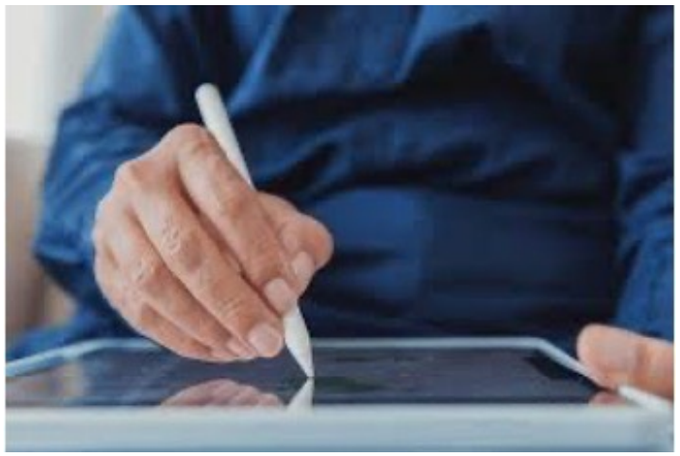
The Catch:
Group Memberships are Often Expensive to Determine



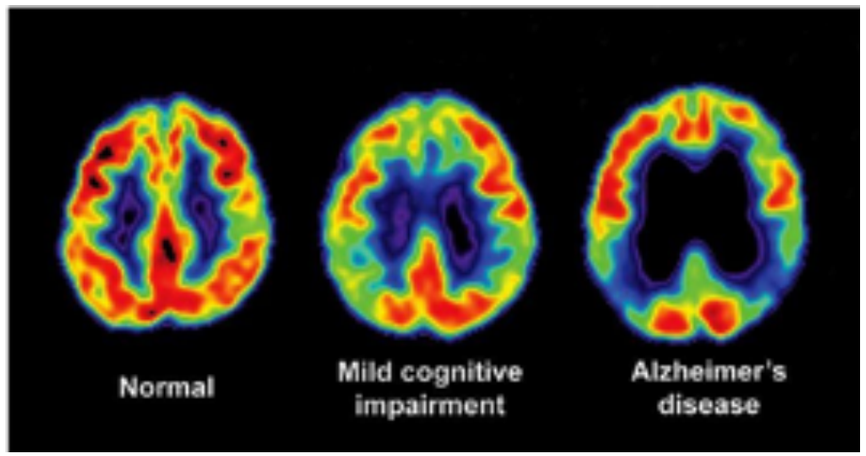


You invent a **new digital test** for Alzheimer's.

The Catch:
Group Memberships are Often Expensive to Determine



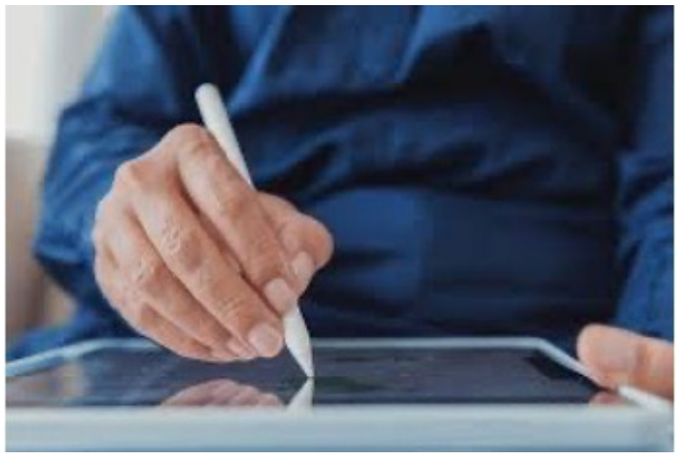
You invent a **new digital test** for Alzheimer's.



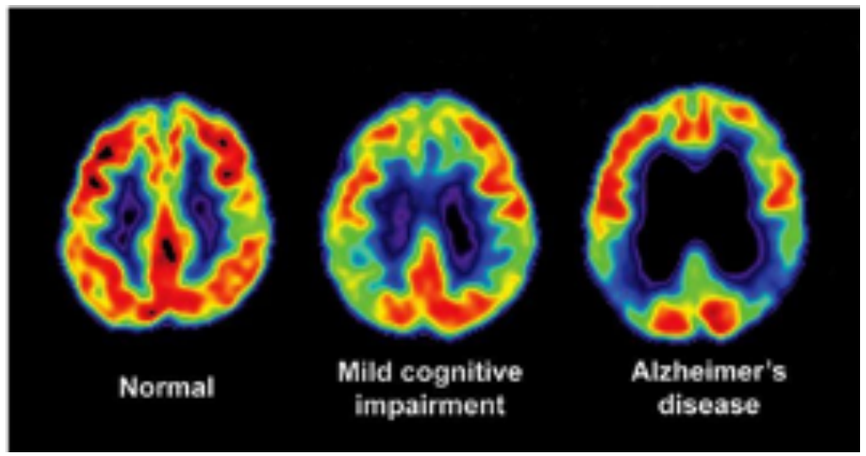
PET scan measuring amyloid build up

The Catch:
Group Memberships are Often Expensive to Determine

Back to our example: validating digital biomarkers for AD



You invent a **new digital test** for Alzheimer's.



PET scan measuring amyloid build up

The Catch:
Group Memberships are Often Expensive to Determine

a population of digital test results
cheap

exhaustive/ random sampling →

reveal amyloid amount by PET scans

very expensive!

→

high amyloid group

low amyloid group

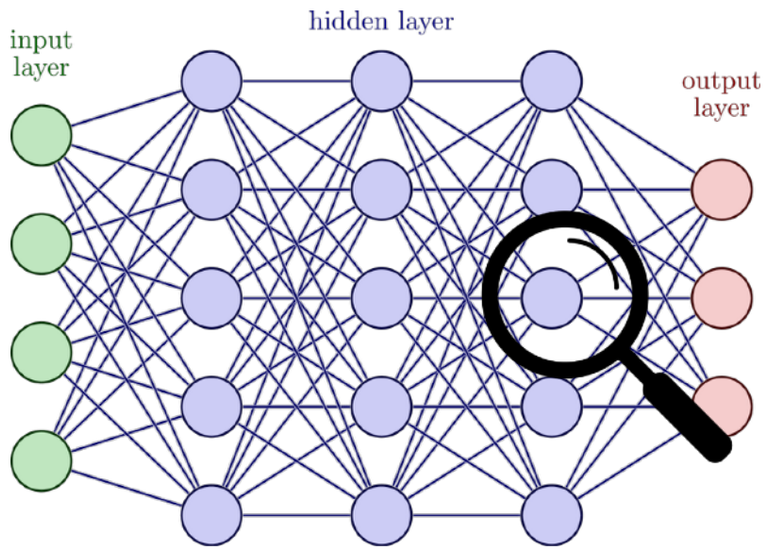
The Catch:
Group Memberships are Often Expensive to Determine

Digital health sensor validation

digital health data easy;
lab tests expensive

Financial Fraud Detection

transactions features easy to obtain;
classifying is expensive



Model Monitoring / ML OPs:
data drift relative to training?

model outputs/perf easy;
post-deploy groundtruth hard

Active Querying for Two-Sample Testing

a population
of digital test
results
cheap

Active selection

exhaustive/
~~random~~
sampling

reveal amyloid amount by PET scans

very expensive!
only need a limited number

high amyloid group

low amyloid group

Active Querying for Two-Sample Testing

a population of digital test results

cheap

Active selection

exhaustive/ random sampling

very expensive!

reveal amyloid amount by PET scans

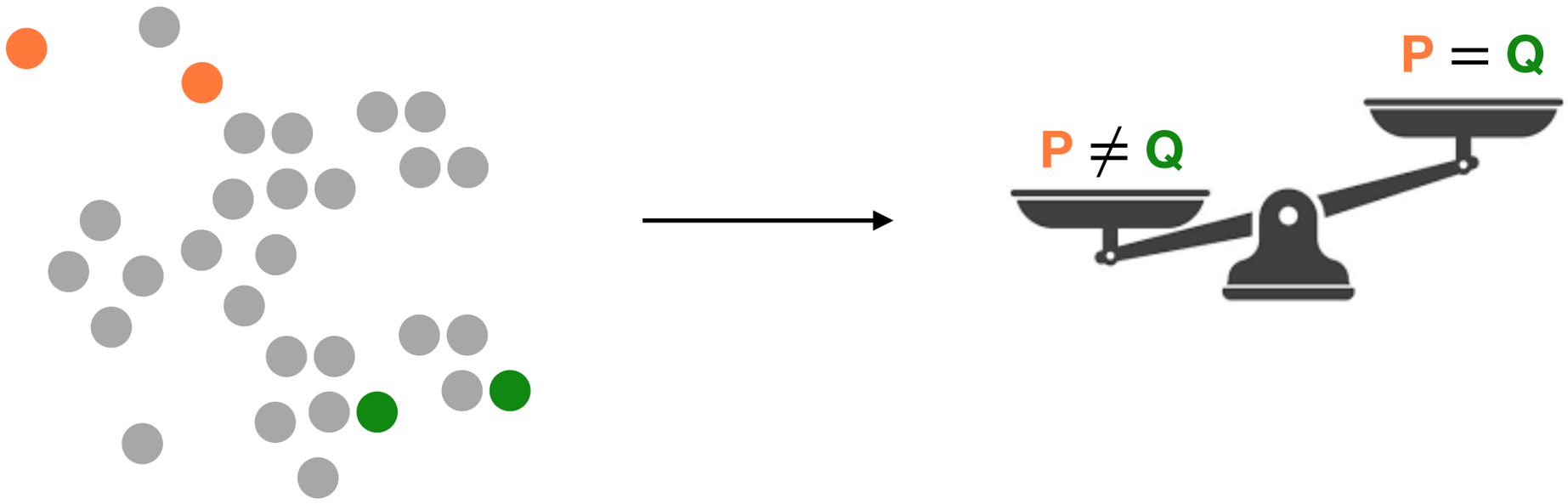
only need a limited number

high amyloid group

low amyloid group

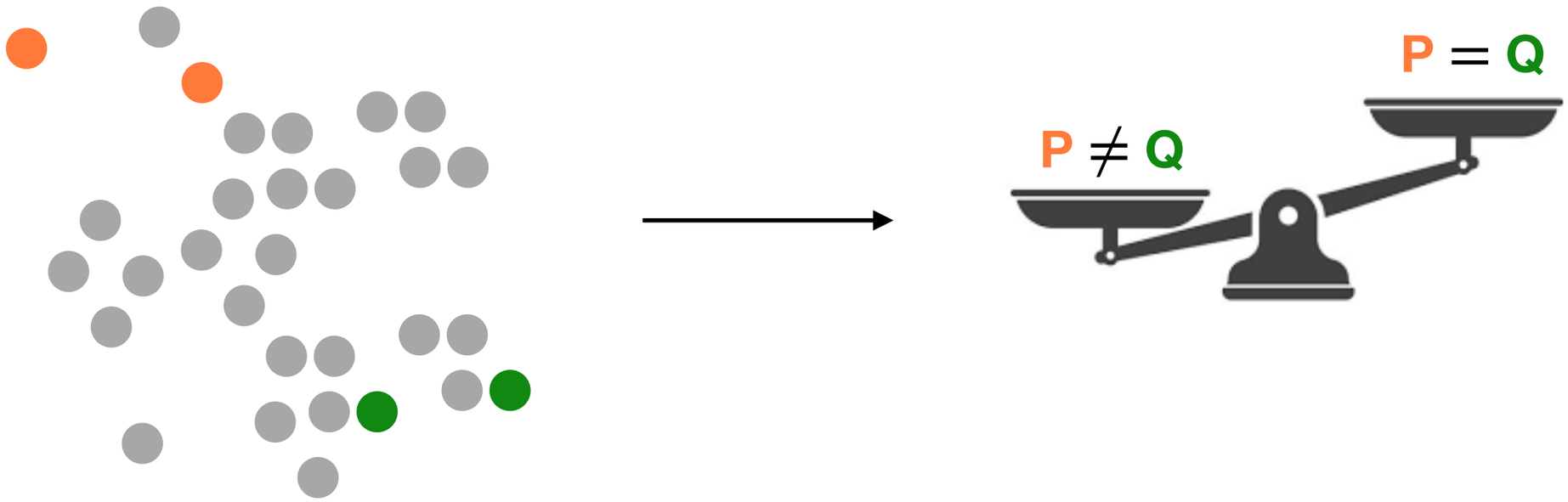
Idea: **Carefully (and adaptively)** select digital test results (features) and query their group memberships (i.e., PET scans)

A New Two-Sample Testing Problem



A New Two-Sample Testing Problem

Given a **large population of sample features** and a **limited labeling (group-membership ascertaining) budget**, our goal is to develop a **label-efficient two-sample test** to determine **whether the two samples are drawn from the same or different distributions.**



A Different Perspective on Two-Sample Testing

Data Model

A Different Perspective on Two-Sample Testing

Data Model

$$Z \sim \text{Ber}(\theta)$$

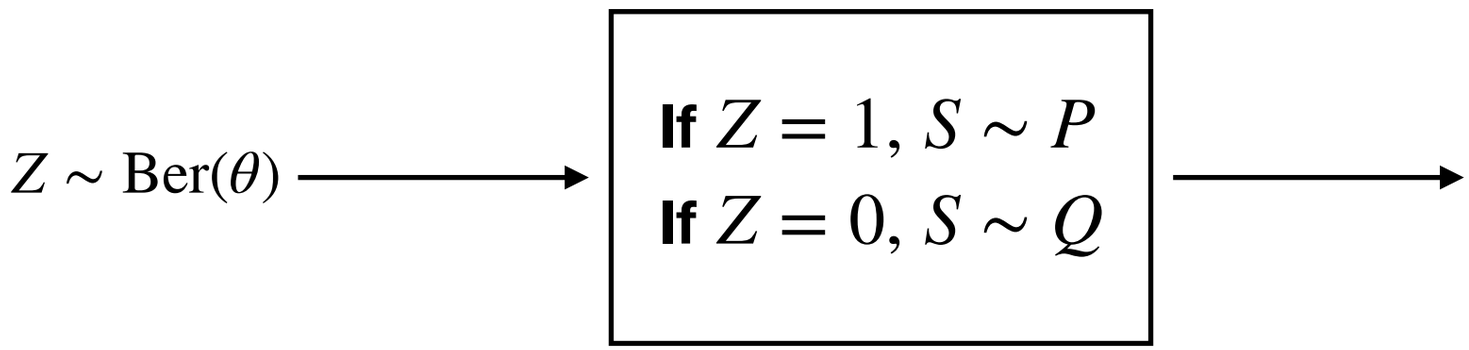
A Different Perspective on Two-Sample Testing

Data Model

$Z \sim \text{Ber}(\theta) \longrightarrow$

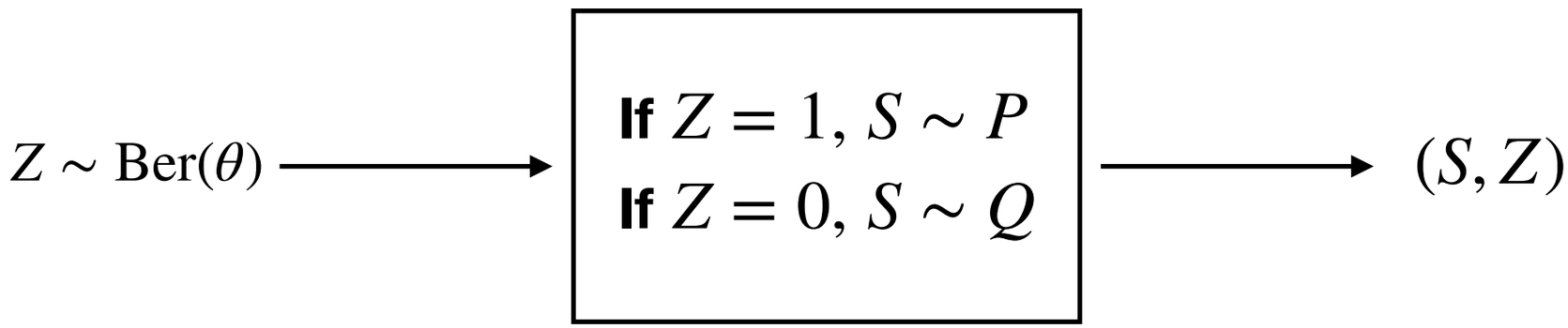
A Different Perspective on Two-Sample Testing

Data Model



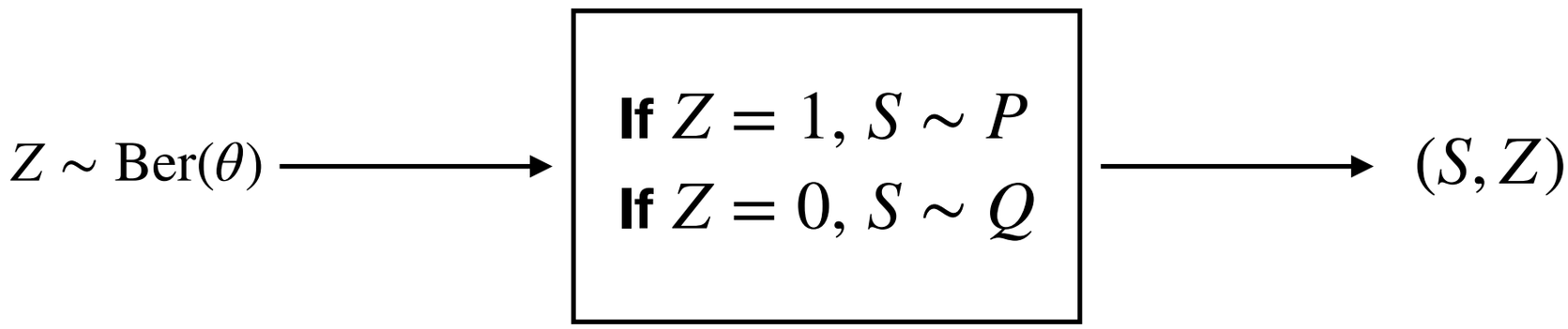
A Different Perspective on Two-Sample Testing

Data Model



A Different Perspective on Two-Sample Testing

Data Model



The two-sample testing problem can be recast as an **independence test** here.

$$H_0 : p(S \mid Z = 0) = p(S \mid Z = 1) \text{ or } S \perp\!\!\!\perp Z$$
$$H_1 : p(S \mid Z = 0) \neq p(S \mid Z = 1) \text{ or } S \not\perp\!\!\!\perp Z$$

Without Further Ado: The Bimodal Query Algorithm

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

1. **Construct a training set:** Randomly select a set of features and reveal their labels

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

1. **Construct a training set:** Randomly select a set of features and reveal their labels

2. **Classifier training:** Train a binary classifier using the training set to obtain $\hat{P}(Z \mid S)$, an estimate of the conditional label probability.

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

1. **Construct a training set:** Randomly select a set of features and reveal their labels

2. **Classifier training:** Train a binary classifier using the training set to obtain $\hat{P}(Z \mid S)$, an estimate of the conditional label probability.

3. **Bimodal Query:** using the rest of the label budget, **query** the labels corresponding **to high** $\hat{P}(Z = 0 \mid S)$ **and** $\hat{P}(Z = 1 \mid S)$ **— the modes!**

Without Further Ado: The Bimodal Query Algorithm

Bimodal Query Algorithm

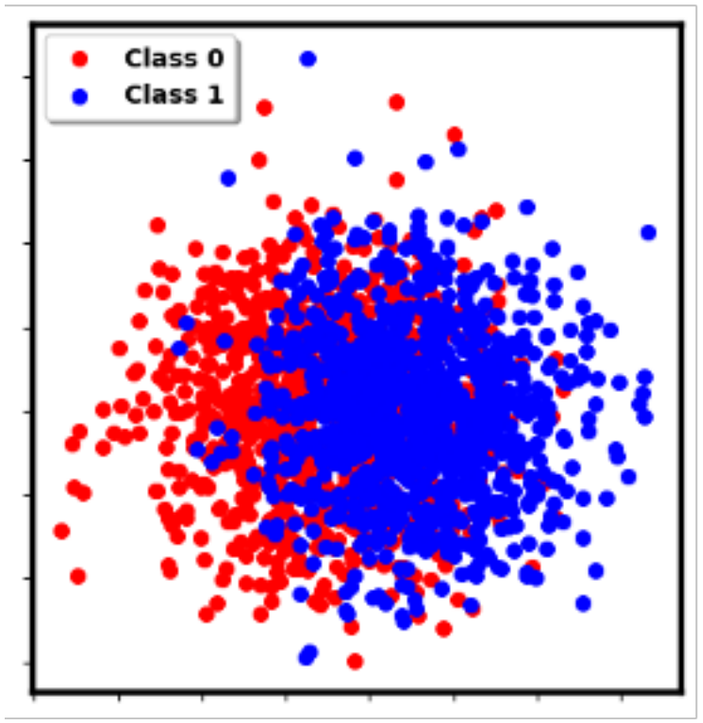
1. **Construct a training set:** Randomly select a set of features and reveal their labels

2. **Classifier training:** Train a binary classifier using the training set to obtain $\hat{P}(Z \mid S)$, an estimate of the conditional label probability.

3. **Bimodal Query:** using the rest of the label budget, **query** the labels corresponding **to high $\hat{P}(Z = 0 \mid S)$ and $\hat{P}(Z = 1 \mid S)$ — the modes!**

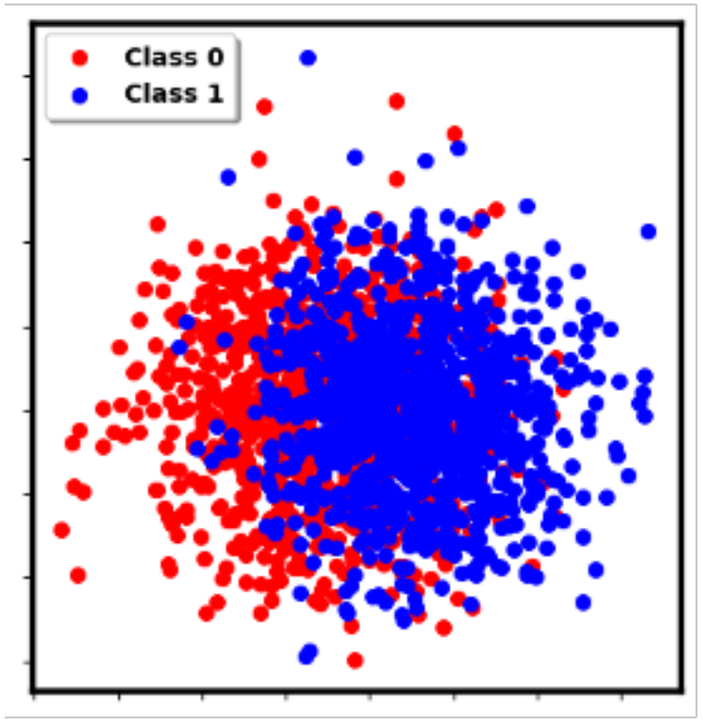
4. **Two-sample testing**: Construct a two-sample test (e.g., FR test) on the resulting two samples

What does this do?

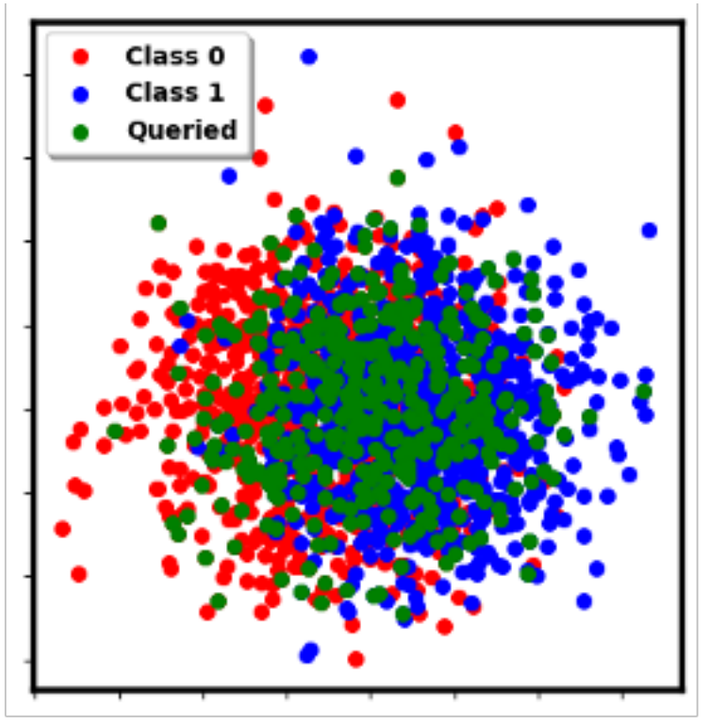


Sample features (the labels are unknown)

What does this do?

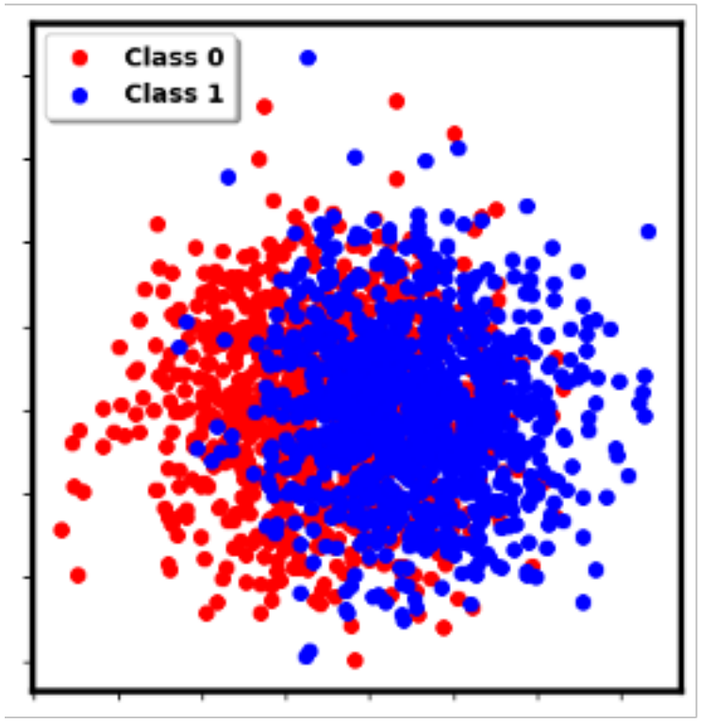


Sample features (the labels are unknown)

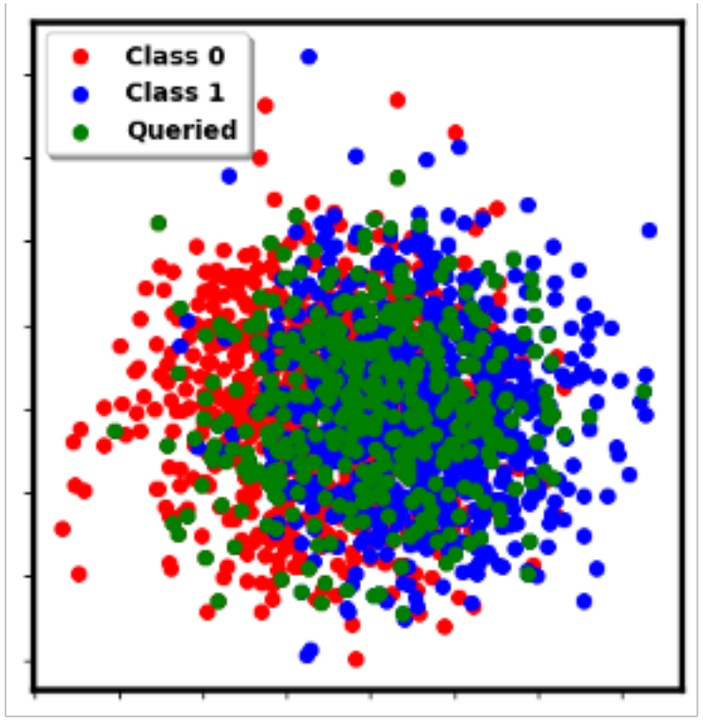


Random sampling

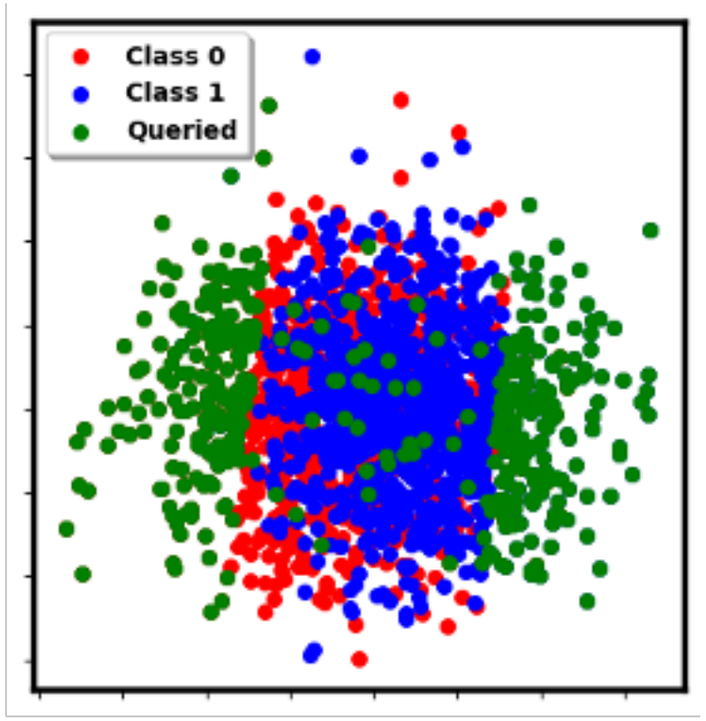
What does this do?



Sample features (the labels are unknown)



Random sampling



Bimodal query

What does this *really do?*

Theorem (LKSRDB 24/LDRB 22). Assuming an **appropriate classifier (e.g., KNN) is used**, then under H_1, the distribution of features s selected by the bimodal query **converges to** $p^*(s)$-- the distribution that makes the FR statistic **maximally powered**.

Li, W, Kadambi, P, Saidi, P, Ramamurthy, K, **GD**, and Berisha, V. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, **GD**, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)

What does this *really do?*

Theorem (LKSRDB 24/LDRB 22). Assuming an **appropriate classifier (e.g., KNN) is used**, then under H_1, the distribution of features s selected by the bimodal query **converges to** $p^*(s)$-- the distribution that makes the FR statistic **maximally powered**.

Proof idea:

Li, W, Kadambi, P, Saidi, P, Ramamurthy, K, **GD**, and Berisha, V. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, **GD**, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)

What does this *really do?*

Theorem (LKSRDB 24/LDRB 22). Assuming an **appropriate classifier (e.g., KNN) is used**, then under H_1, the distribution of features s selected by the bimodal query **converges to** $p^*(s)$-- the distribution that makes the FR statistic **maximally powered**.

Proof idea:

- We first show a structural result: the FR statistic (R/n) **converges to a function of** $\int p(Z=0\,|\,S)p(Z=1\,|\,S)dp(S)$.

Li, W, Kadambi, P, Saidi, P, Ramamurthy, K, **GD**, and Berisha, V. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, **GD**, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)

What does this *really do?*

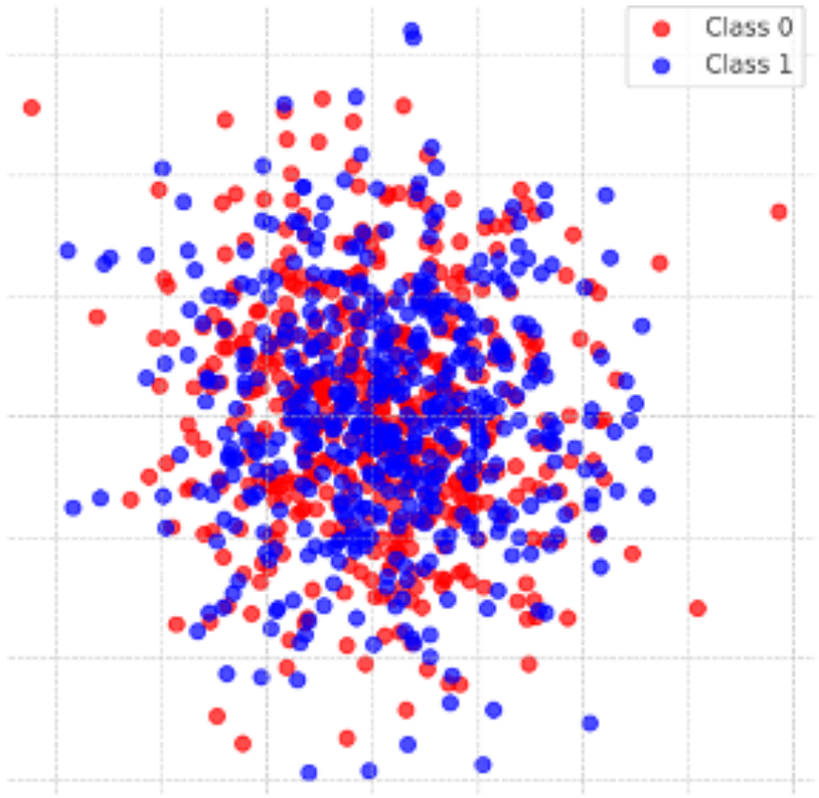
Theorem (LKSRDB 24/LDRB 22). Assuming an **appropriate classifier (e.g., KNN) is used**, then under H_1, the distribution of features s selected by the bimodal query **converges to** $p^*(s)$-- the distribution that makes the FR statistic **maximally powered**.

Proof idea:

- We first show a structural result: the FR statistic (R/n) **converges to a function of** $\int p(Z=0\,|\,S)p(Z=1\,|\,S)dp(S)$.
- We then show that this function is **minimized (asymptotically)** by our Bimodal Query. LP in p(s) —> optima at extremes, roughly.

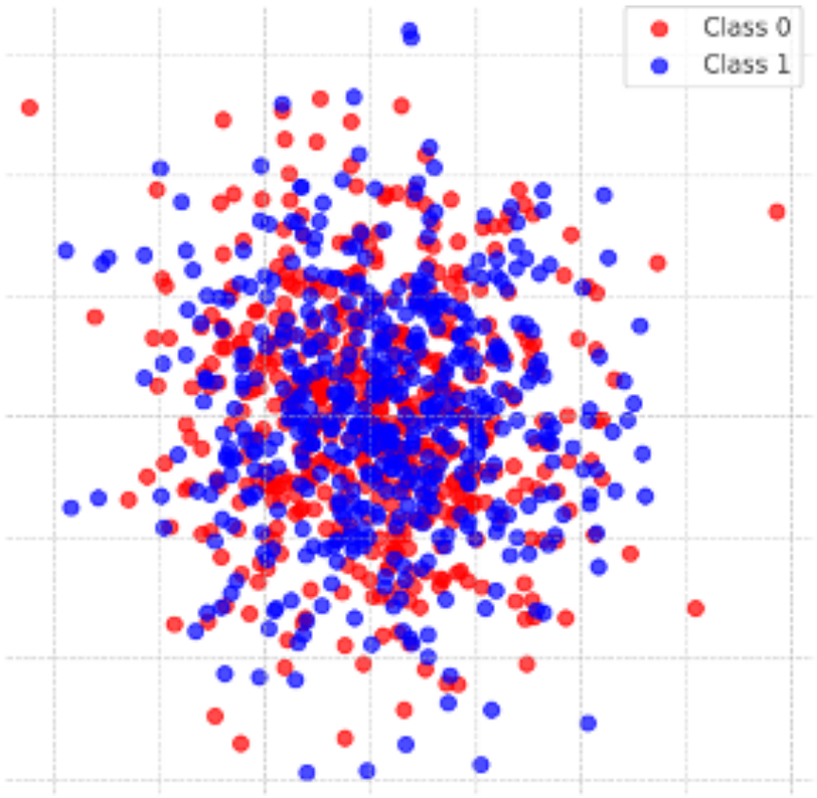
Li, W, Kadambi, P, Saidi, P, Ramamurthy, K, **GD**, and Berisha, V. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, **GD**, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)

Type I Error Control



Theorem 2 (LKSRDB 24). Under H_0, $p(S \mid Z = 0)$ and $p(S \mid Z = 1)$ are identical. Consequently our procedure (built on, say FR) **controls the Type I error** at the specified level.

Type I Error Control

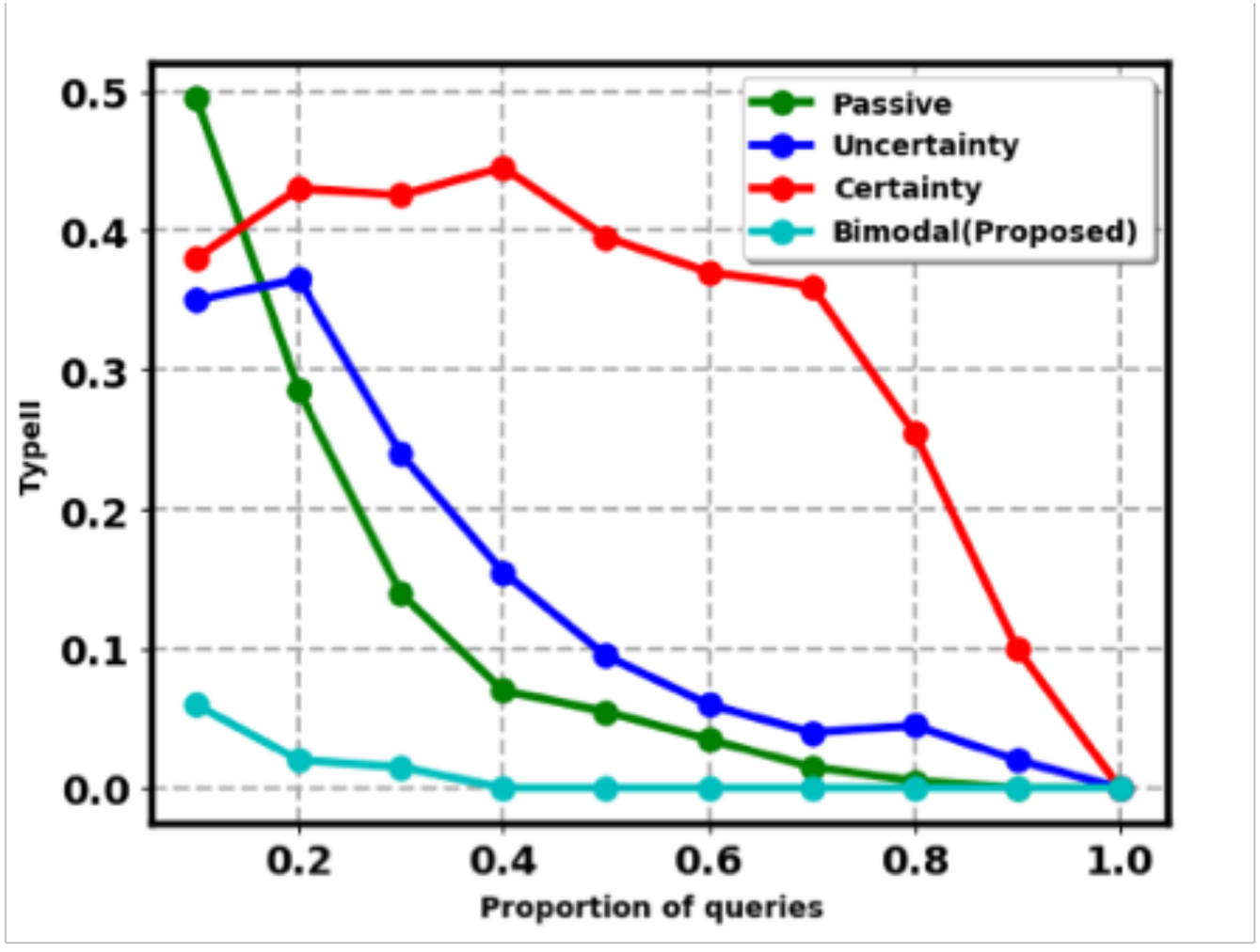


Theorem 2 (LKSRDB 24). Under H_0, $p(S \mid Z = 0)$ and $p(S \mid Z = 1)$ are identical. Consequently our procedure (built on, say FR) **controls the Type I error** at the specified level.

Li, W, Kadambi, P, Saidi, P, Ramamurthy, K, **GD**, and Berisha, V. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, **GD**, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)

Alzheimer's Disease Neuroimaging Initiative (ADNI)

Alzheimer's Disease Neuroimaging Initiative (ADNI)



5 cognition measure scores + PET based Amyloid Scores

Beyond Fixed Budgets: Sequential and Design Extensions

Beyond Fixed Budgets: Sequential and Design Extensions

Li, W, Kadambi, P., Saidi, P., Ramamurthy, K, N., **GD**, and Berisha, Visar. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li W, **GD**, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Beyond Fixed Budgets: Sequential and Design Extensions

Sequential Two Sample Testing

- Label acq can be **done sequentially**
- Maintain a **running test-statistic**
- Produces **anytime-valid p-values** — Type-I control holds *even under adaptive stopping*
- Enables **early stopping** once evidence is sufficient
- Test statistic converges to **mutual information** between features and group labels
- In practice: similar power with **fewer** queried verifications

Li, W, Kadambi, P., Saidi, P., Ramamurthy, K, N., **GD**, and Berisha, Visar. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li W, **GD**, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Beyond Fixed Budgets: Sequential and Design Extensions

Sequential Two Sample Testing

- Label acq can be **done sequentially**
- Maintain a **running test-statistic**
- Produces **anytime-valid p-values** — Type-I control holds *even under adaptive stopping*
- Enables **early stopping** once evidence is sufficient
- Test statistic converges to **mutual information** between features and group labels
- In practice: similar power with **fewer** queried verifications

Extends classical sequential testing (à la Wald) to **nonparametric, label-limited** settings.

Li, W, Kadambi, P., Saidi, P., Ramamurthy, K, N., **GD**, and Berisha, Visar. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li W, **GD**, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Beyond Fixed Budgets: Sequential and Design Extensions

Sequential Two Sample Testing

- Label acq can be **done sequentially**
- Maintain a **running test-statistic**
- Produces **anytime-valid p-values** — Type-I control holds *even under adaptive stopping*
- Enables **early stopping** once evidence is sufficient
- Test statistic converges to **mutual information** between features and group labels
- In practice: similar power with **fewer** queried verifications

Extends classical sequential testing (à la Wald) to **nonparametric, label-limited** settings.

Active Matched Pair Experiment Design

- Same principle: query / enroll where **information gain is highest**
- Actively select **pairs** (treatment–control) from covariate space
- Target regions of **large predicted treatment effect**
- Guarantee: enrolled region encloses true **responder set**
- Achieves **provably early detection** of heterogeneous effects
- Retains **valid Type-I inference** while improving sample efficiency

Li, W, Kadambi, P., Saidi, P., Ramamurthy, K, N., **GD**, and Berisha, Visar. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li W, **GD**, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Beyond Fixed Budgets: Sequential and Design Extensions

Sequential Two Sample Testing

- Label acq can be **done sequentially**
- Maintain a **running test-statistic**
- Produces **anytime-valid p-values** — Type-I control holds *even under adaptive stopping*
- Enables **early stopping** once evidence is sufficient
- Test statistic converges to **mutual information** between features and group labels
- In practice: similar power with **fewer** queried verifications

Extends classical sequential testing (à la Wald) to **nonparametric, label-limited** settings.

Active Matched Pair Experiment Design

- Same principle: query / enroll where **information gain is highest**
- Actively select **pairs** (treatment–control) from covariate space
- Target regions of **large predicted treatment effect**
- Guarantee: enrolled region encloses true **responder set**
- Achieves **provably early detection** of heterogeneous effects
- Retains **valid Type-I inference** while improving sample efficiency

Bridges **testing and design:** both are adaptive inference under verification constraints.

Li, W, Kadambi, P., Saidi, P., Ramamurthy, K, N., **GD**, and Berisha, Visar. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li W, **GD**, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Weizhi Li
LLNL

Prad Kadambi
ASU/Mayo

Karthi Ramamurthy
IBM

Pouria Saidi
Mayo

Visar Berisha
ASU

CNS-2003111, and CCF-2048223

N00014-21-1-2615

20240065DR

Takeaways

Takeaways

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

- Not discussed (in detail) today

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

- Not discussed (in detail) today
  - **Sequential (active) version**, where one does testing sequentially

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

- Not discussed (in detail) today
  - **Sequential (active) version**, where one does testing sequentially
    - We adapt Wald-esque SPRT to create an **anytime valid** sequential test
    - **Provably consistent and better than passive** sequential counterparts

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel ″active″ algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

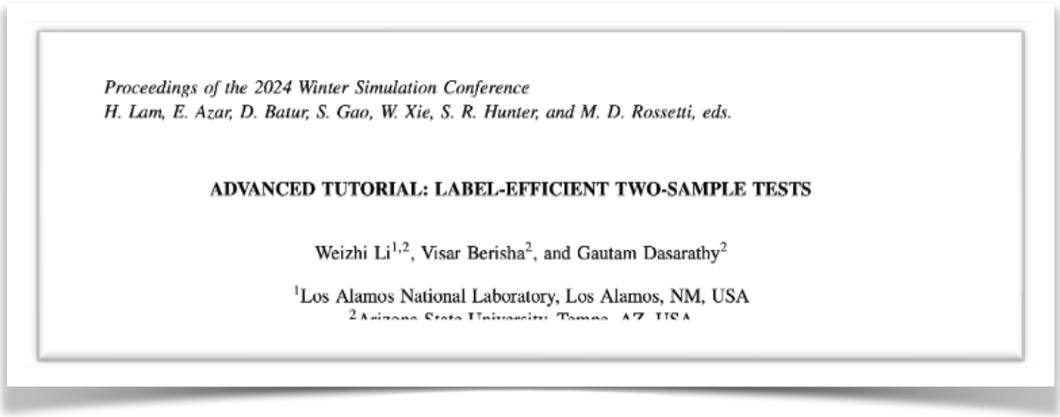
- Not discussed (in detail) today
  - **Sequential (active) version**, where one does testing sequentially
    - We adapt Wald-esque SPRT to create an **anytime valid** sequential test
    - **Provably consistent and better than passive** sequential counterparts
  - **Cohort enrichment + Matched-Pair experiment design:**
    - Find a sub-group with significant effect

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, “A Label-Efficient Two-Sample Test”. Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)

Takeaways

- A **new take** on a **classical hypothesis testing problem**: Label Efficient Two-Sample Tests.
- Proposed a novel "active" algorithm that is:
  - **asymptotically valid** (Type I Error is correct)
  - **consistent**
  - **provably better** than passive sampling

https://gautamdasarathy.com



Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

ADVANCED TUTORIAL: LABEL-EFFICIENT TWO-SAMPLE TESTS

Weizhi Li[1,2], Visar Berisha[2], and Gautam Dasarathy[2]

[1]Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Arizona State University, Tempe, AZ, USA

- Not discussed (in detail) today
  - **Sequential (active) version**, where one does testing sequentially
    - We adapt Wald-esque SPRT to create an **anytime valid** sequential test
    - **Provably consistent and better than passive** sequential counterparts
  - **Cohort enrichment + Matched-Pair experiment design:**
    - Find a sub-group with significant effect

Li, Weizhi, Prad Kadambi, Pouria Saidi, Karthikeyan Natesan Ramamurthy, Gautam Dasarathy, and Visar Berisha. "Active Sequential Two-Sample Testing." Transactions on Machine Learning Research (2024).
Li, Dasarathy, Ramamurthy, Berisha, "A Label-Efficient Two-Sample Test". Uncertainty in AI (2022)
Li W, GD, Berisha, V., Matched-Pair Experiment Design with Active Learning. arXiv:2509.10742v2 (2025)