Statistics By Default

Lecture 1: A/B Testing

Andrew Lizarraga

What this course IS NOT:

- This is not tutoring!

What this course IS NOT:

- This is not tutoring!
- I won’t have time to go into every detail.

What this course IS NOT:

- This is not tutoring!
- I won’t have time to go into every detail.
- I’m happy to answer questions.

What this course IS NOT:

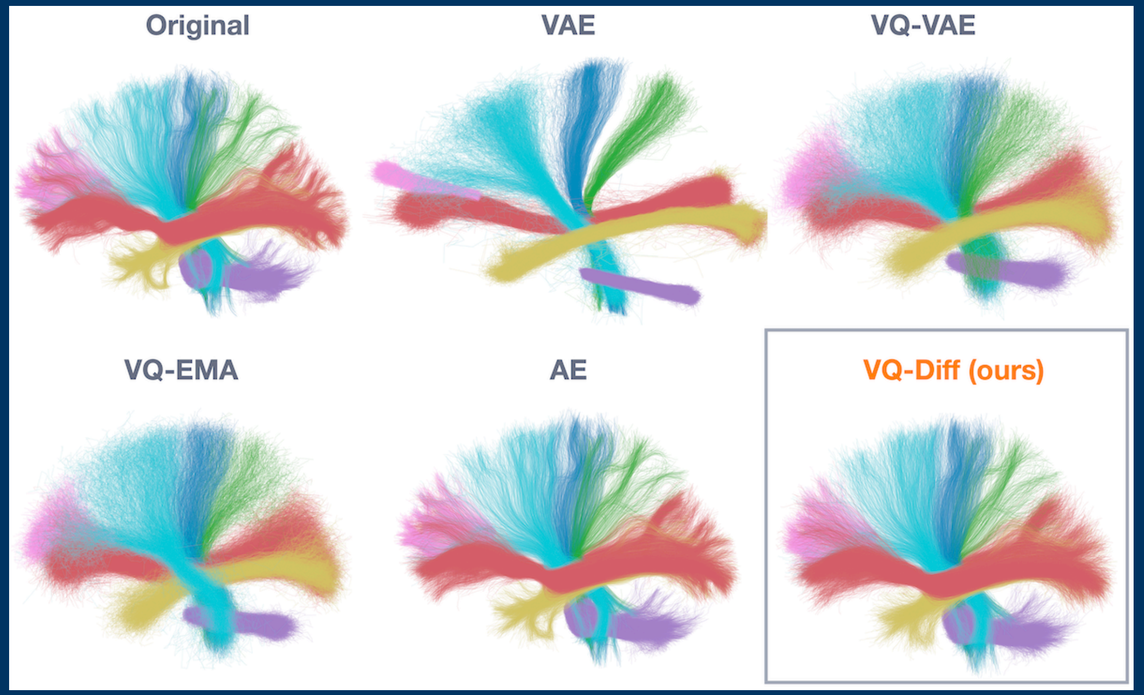
- This is not tutoring!
- I won’t have time to go into every detail.
- I’m happy to answer questions.
- However you get as much as you put in.

What this course IS NOT:

- This is not tutoring!
- I won’t have time to go into every detail.
- I’m happy to answer questions.
- However you get as much as you put in.
- Be sure to help your peers!

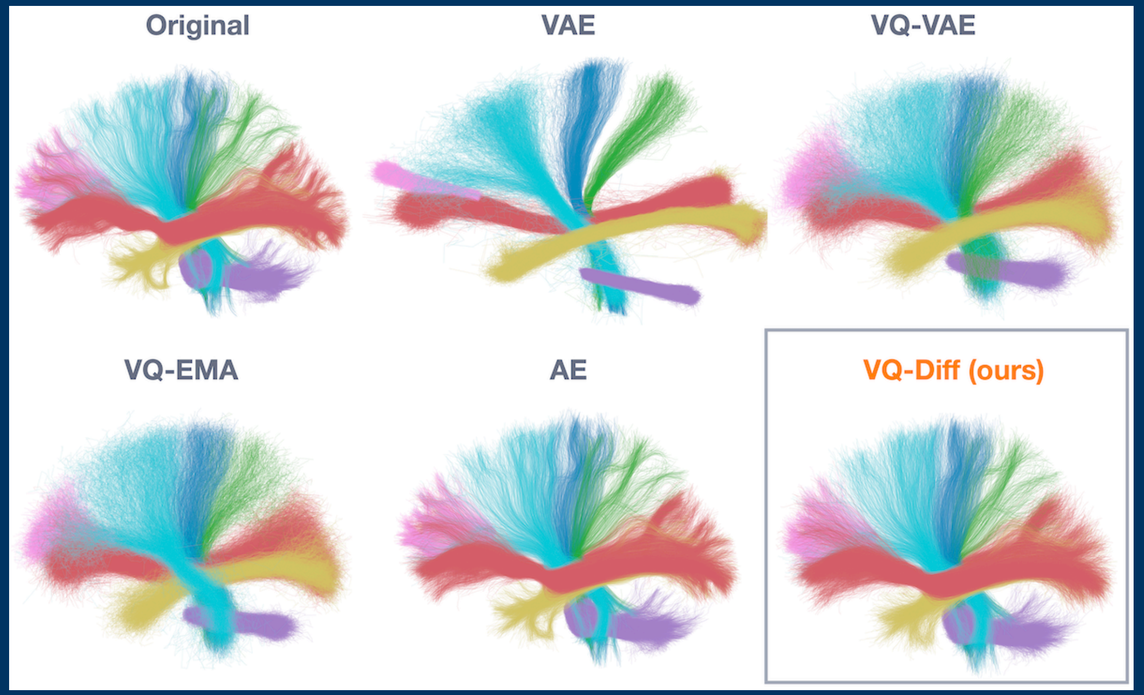
Quick About Me

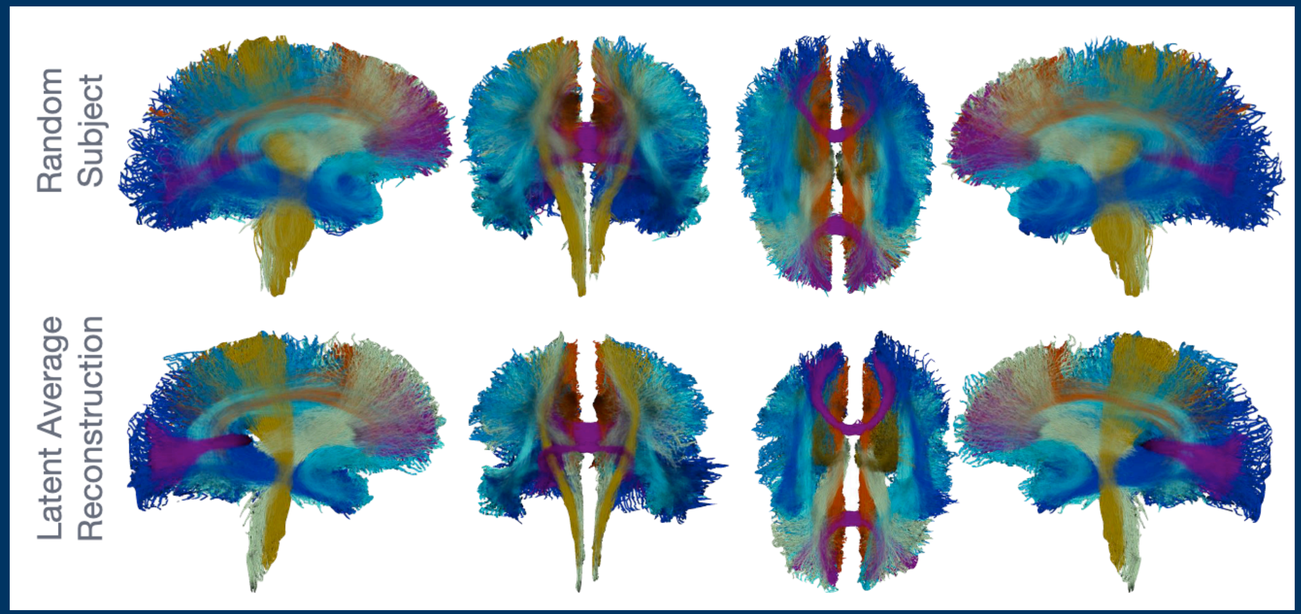
- **Research - AI4Science / Generative Models / Latent Modeling**



Quick About Me

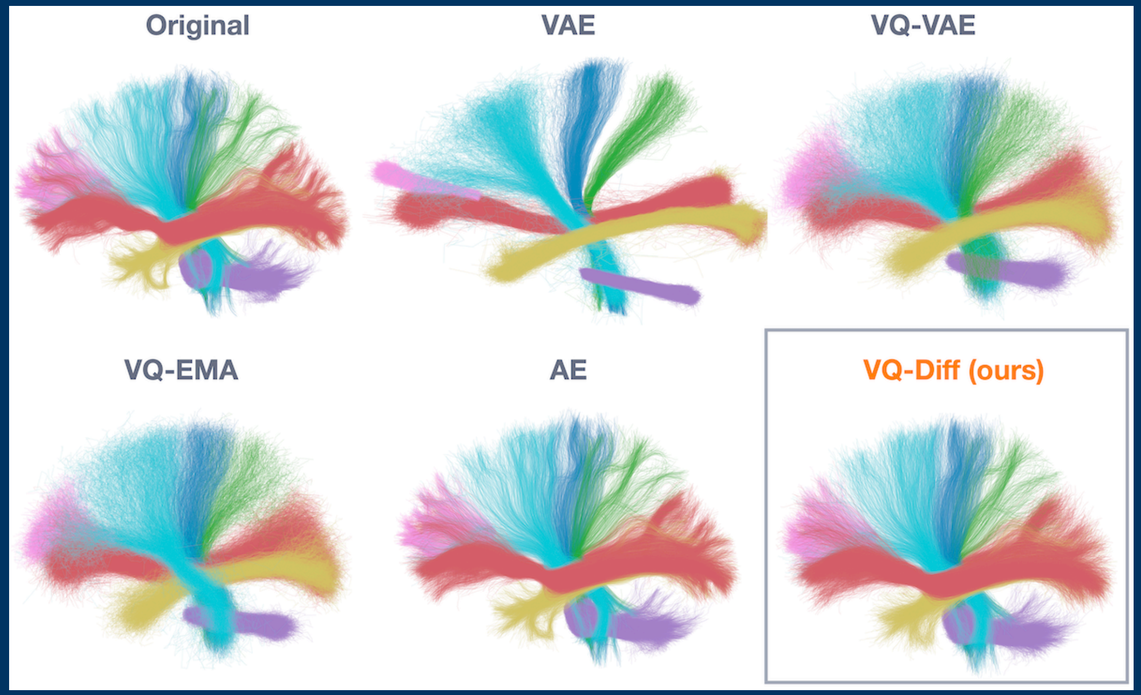
- **Research - AI4Science / Generative Models / Latent Modeling**

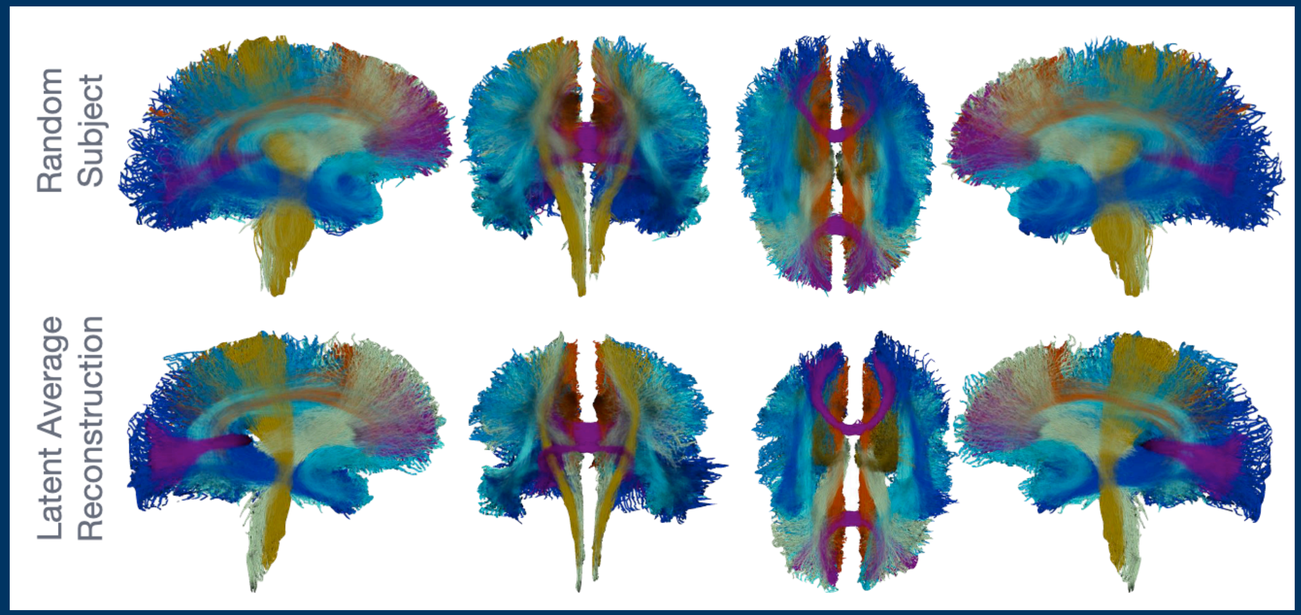


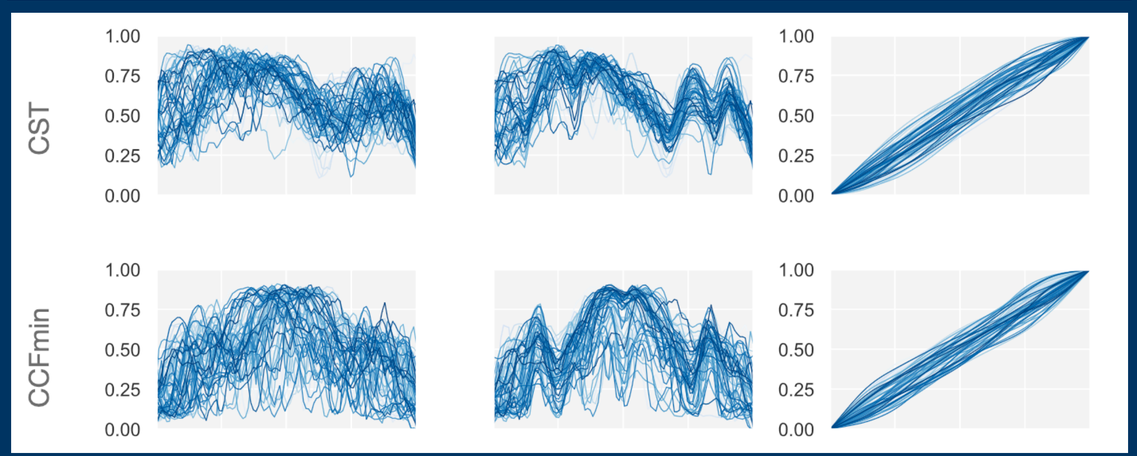


Quick About Me

- **Research - AI4Science / Generative Models / Latent Modeling**

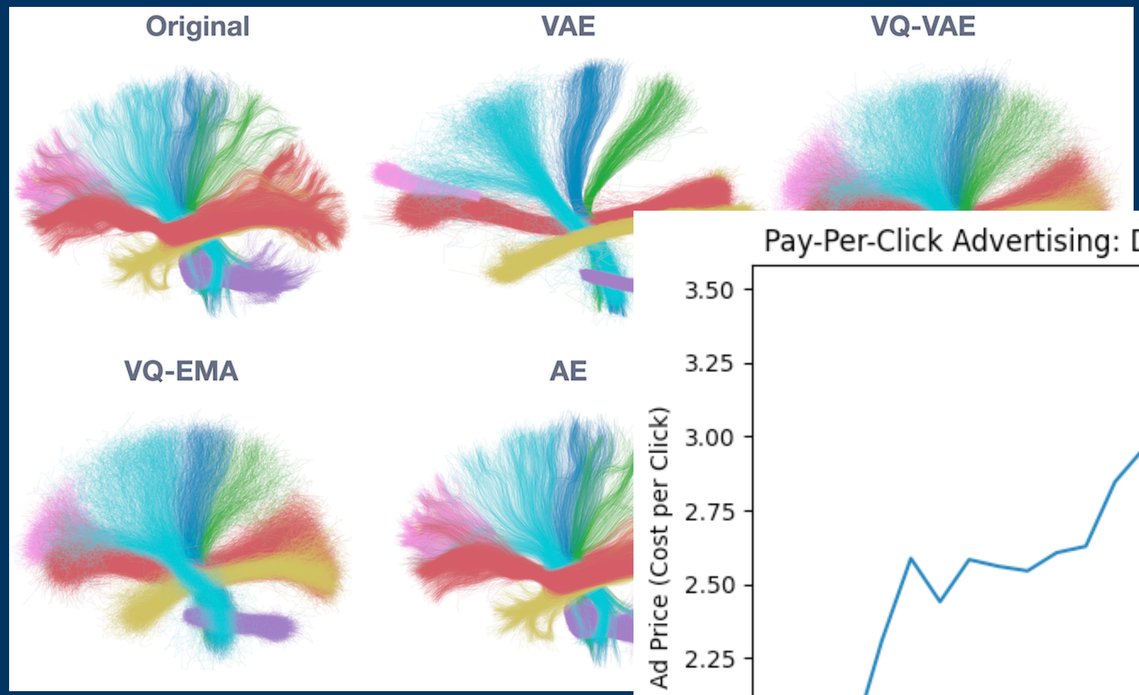


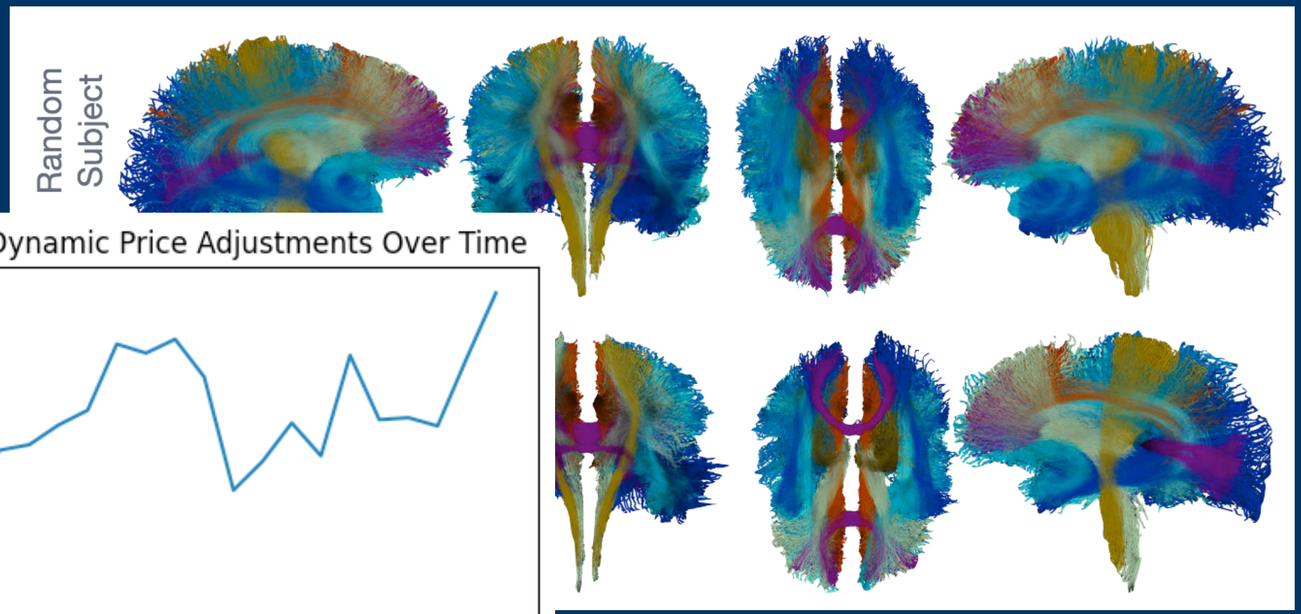


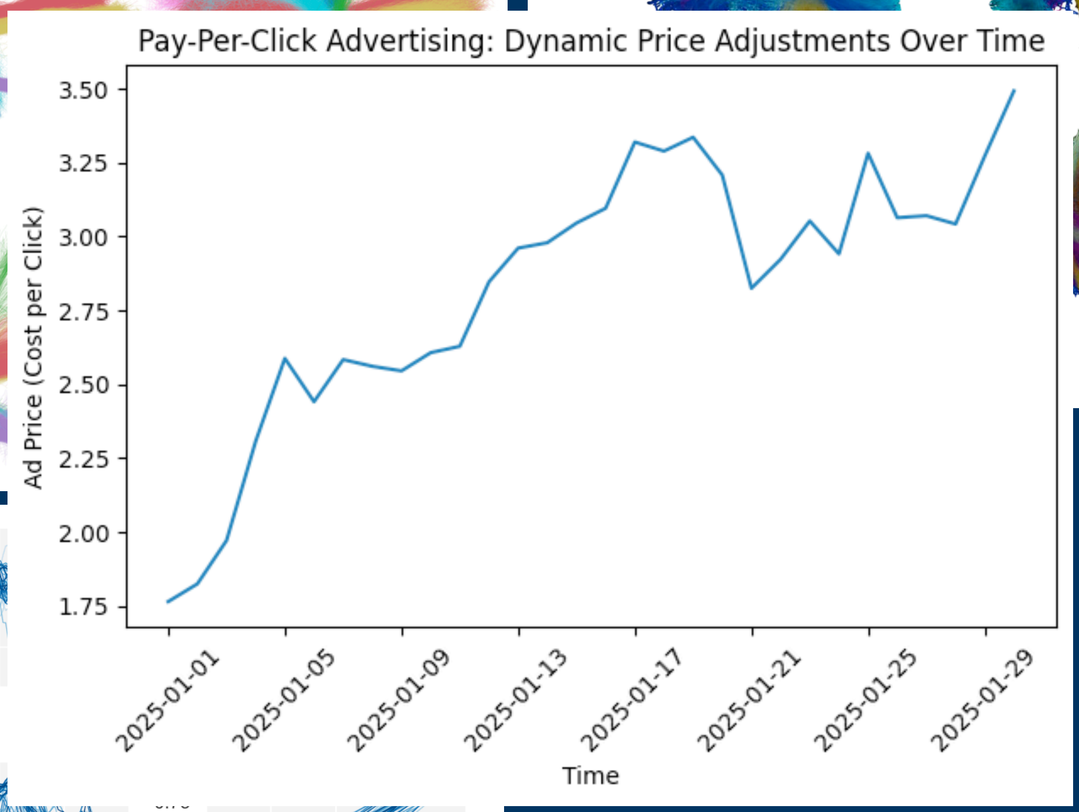


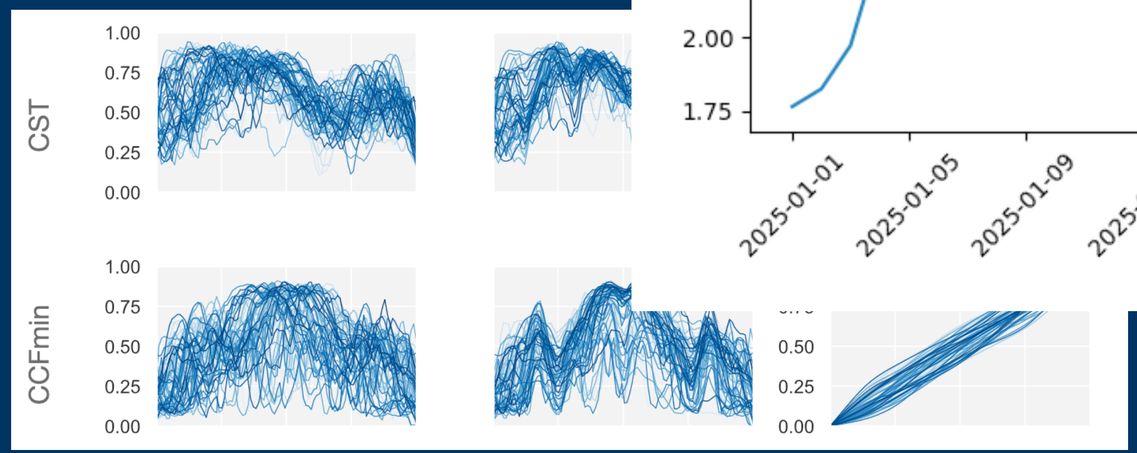
Quick About Me

- **Research - AI4Science / Generative Models / Latent Modeling**



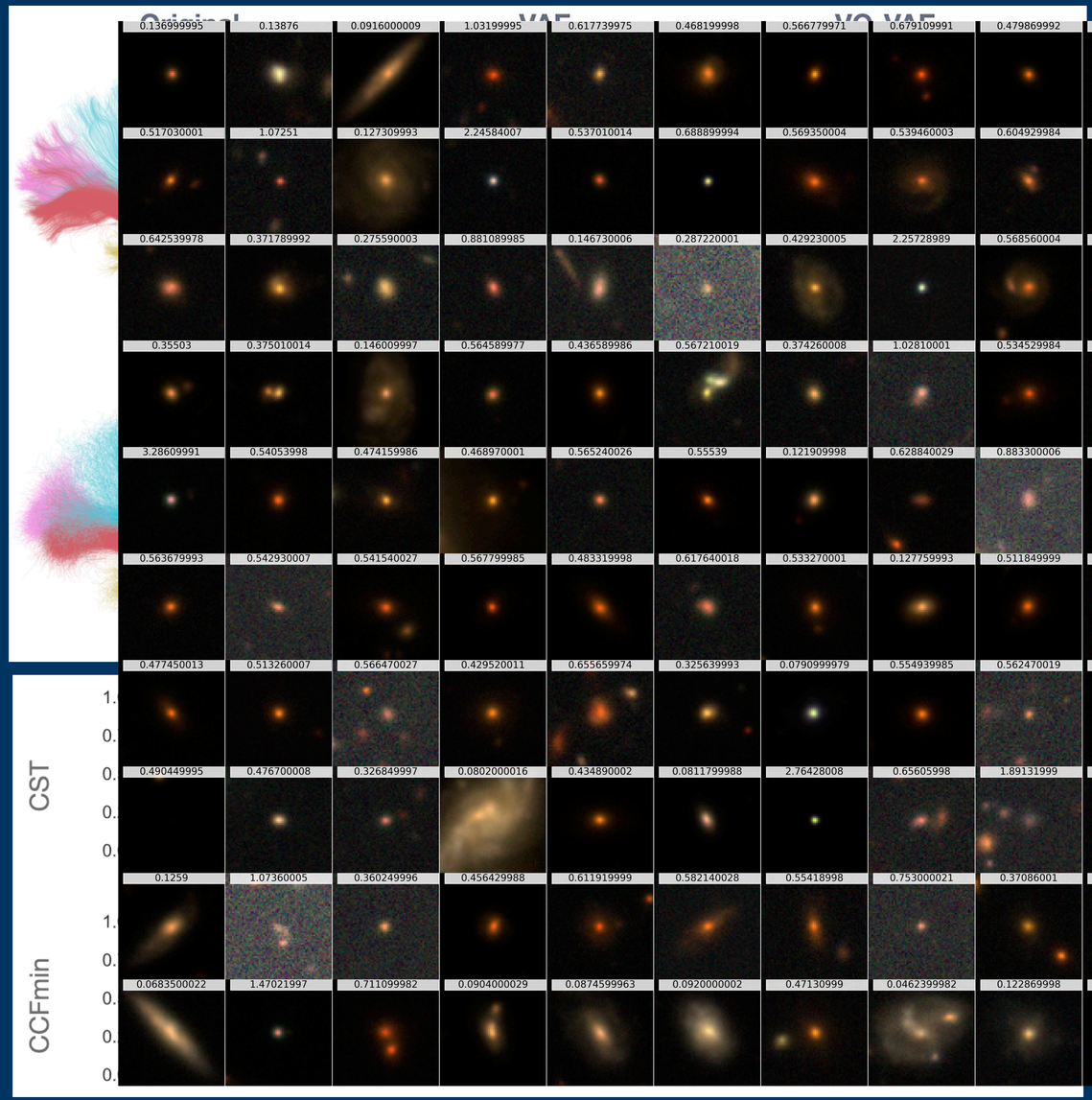


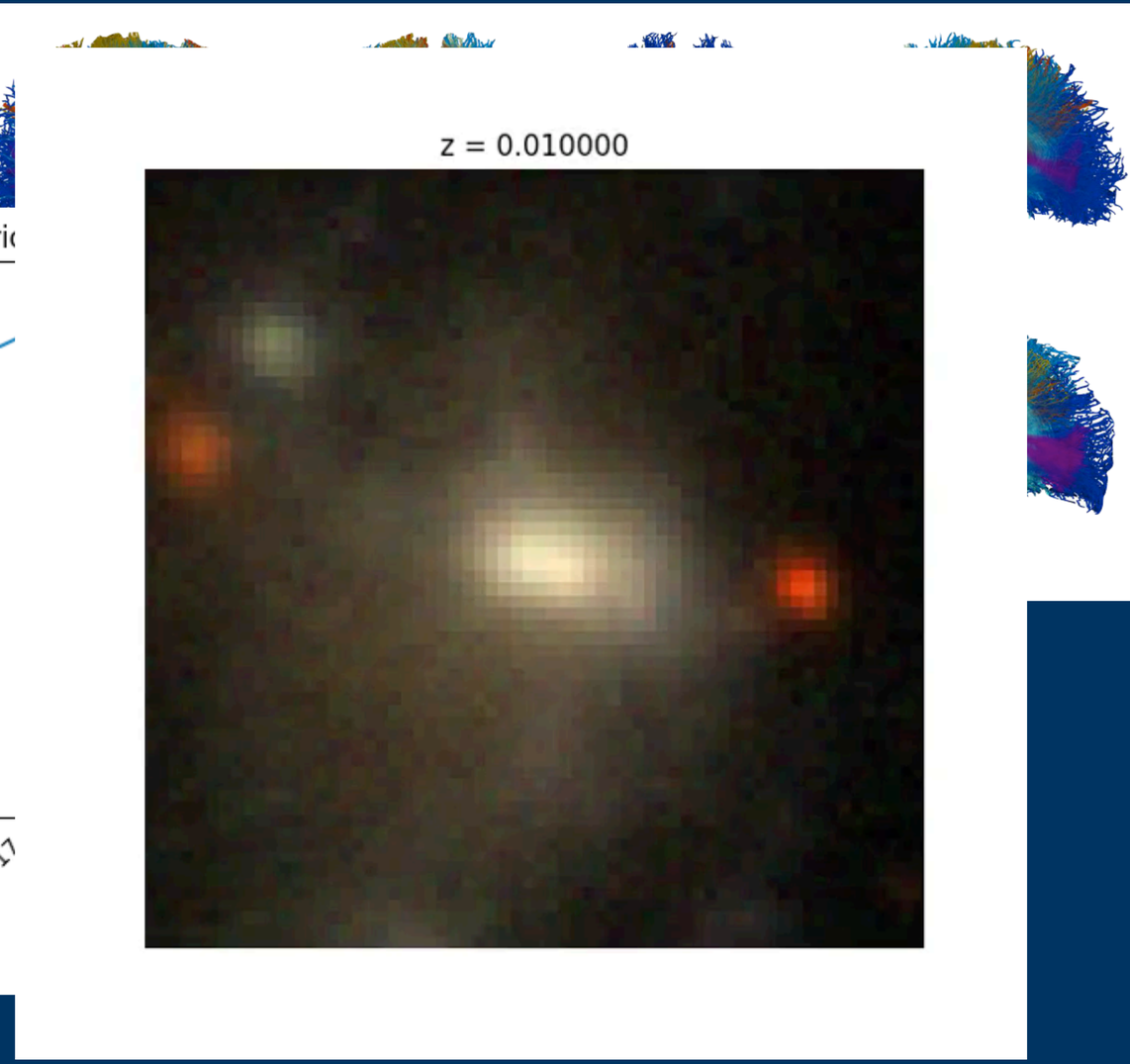




Quick About Me

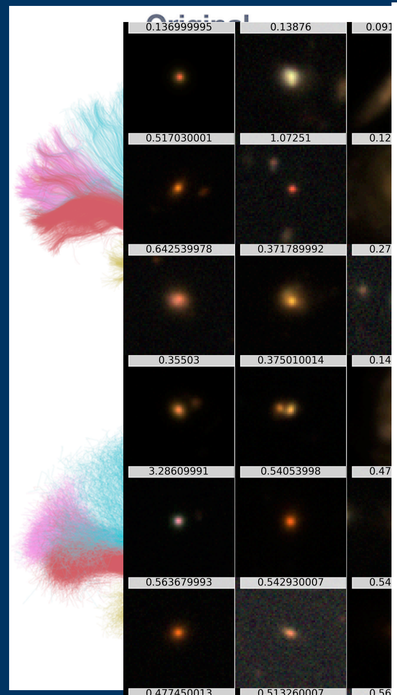
- **Research - AI4Science / Generative Models / Latent Modeling**

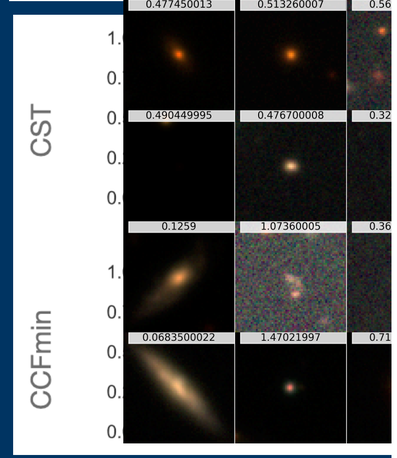




Quick About Me

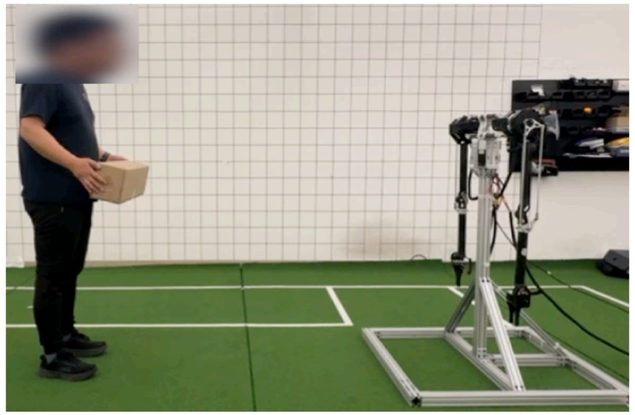
- **Research - AI4Science / Generative Models / Latent Modeling**



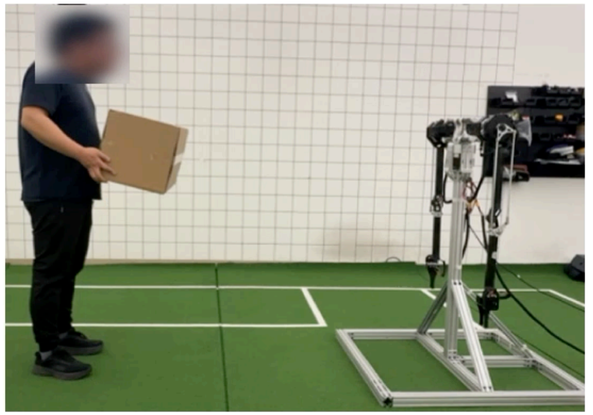


Latent Adaptive Planner: Hardware Testing

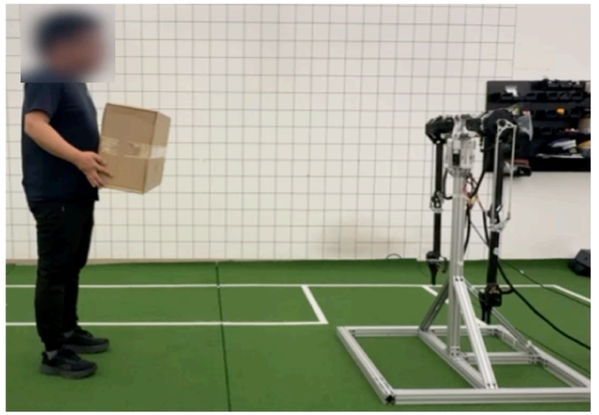
Baseline: Diffusion Policy **(Best Selected Cases)**



Robot A – Box 1



Robot A – Box 2



Robot A – Box 3

Only learned final pose with fine success rates due to task simplicity, with highly noisy execution

What’s this about?

What’s this about?

- **Students are struggling with applications:**

What’s this about?

- Students are struggling with applications:
  - Lack of practical programming.

What’s this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.

What’s this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.
  - Lack of engineering maturity:

What’s this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.
  - Lack of engineering maturity:
    - What’s practical?

What's this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.
  - Lack of engineering maturity:
    - What's practical?
    - When should you apply a technique or not?

What's this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.
  - Lack of engineering maturity:
    - What's practical?
    - When should you apply a technique or not?
    - Having too complicated of a view on a problem.

What's this about?

- Students are struggling with applications:
  - Lack of practical programming.
  - Lack of putting theory to practice.
  - Lack of engineering maturity:
    - What's practical?
    - When should you apply a technique or not?
    - Having too complicated of a view on a problem.
    - Having too simple of view on a complicated problem.

Assumptions:

Assumptions:

- I’m assuming you have background in stats, linear algebra, and programming (you may not have this background).

Assumptions:

- I’m assuming you have background in stats, linear algebra, and programming (you may not have this background).

- I want everyone to be comfortable with answering question and actually stating your thought process when answering questions.

A/B - Testing

A/B - Testing

- What is an A/B test in intuitive terms?

A/B - Testing

- What is an A/B test in intuitive terms?
- When should you conduct an A/B test?

A/B - Testing

- What is an A/B test in intuitive terms?
- When should you conduct an A/B test?
- When can you perform an A/B test?

A/B - Testing

- **What is an A/B test in intuitive terms?**
  - **A statistically sound approach to compare two things:**

A/B - Testing

- What is an A/B test in intuitive terms?
  - A statistically sound approach to compare two things:
    - Is drug B better than drug A?

A/B - Testing

- What is an A/B test in intuitive terms?
  - A statistically sound approach to compare two things:
    - Is drug B better than drug A?
    - Do customers prefer product B over product A?

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):
    - When it’s not easy to tell if there is a difference between two objects of interest.

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):
    - When it’s not easy to tell if there is a difference between two objects of interest.
      - Is NYSE a better exchange than NASDAQ?

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):
    - When it’s not easy to tell if there is a difference between two objects of interest.
      - Is NYSE a better exchange than NASDAQ?
      - Is drug B safer than drug A (even marginally so)?

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):
    - When it’s not easy to tell if there is a difference between two objects of interest.
      - Is NYSE a better exchange than NASDAQ?
      - Is drug B safer than drug A (even marginally so)?
      - Is computer B faster than computer A?

A/B - Testing

- When should you conduct and A/B test?
  - (Many Answers):
    - When it’s not easy to tell if there is a difference between two objects of interest.
      - Is NYSE a better exchange than NASDAQ?
      - Is drug B safer than drug A (even marginally so)?
      - Is computer B faster than computer A?
      - Is LLM B better than LLM A at mathematics?

A/B - Testing

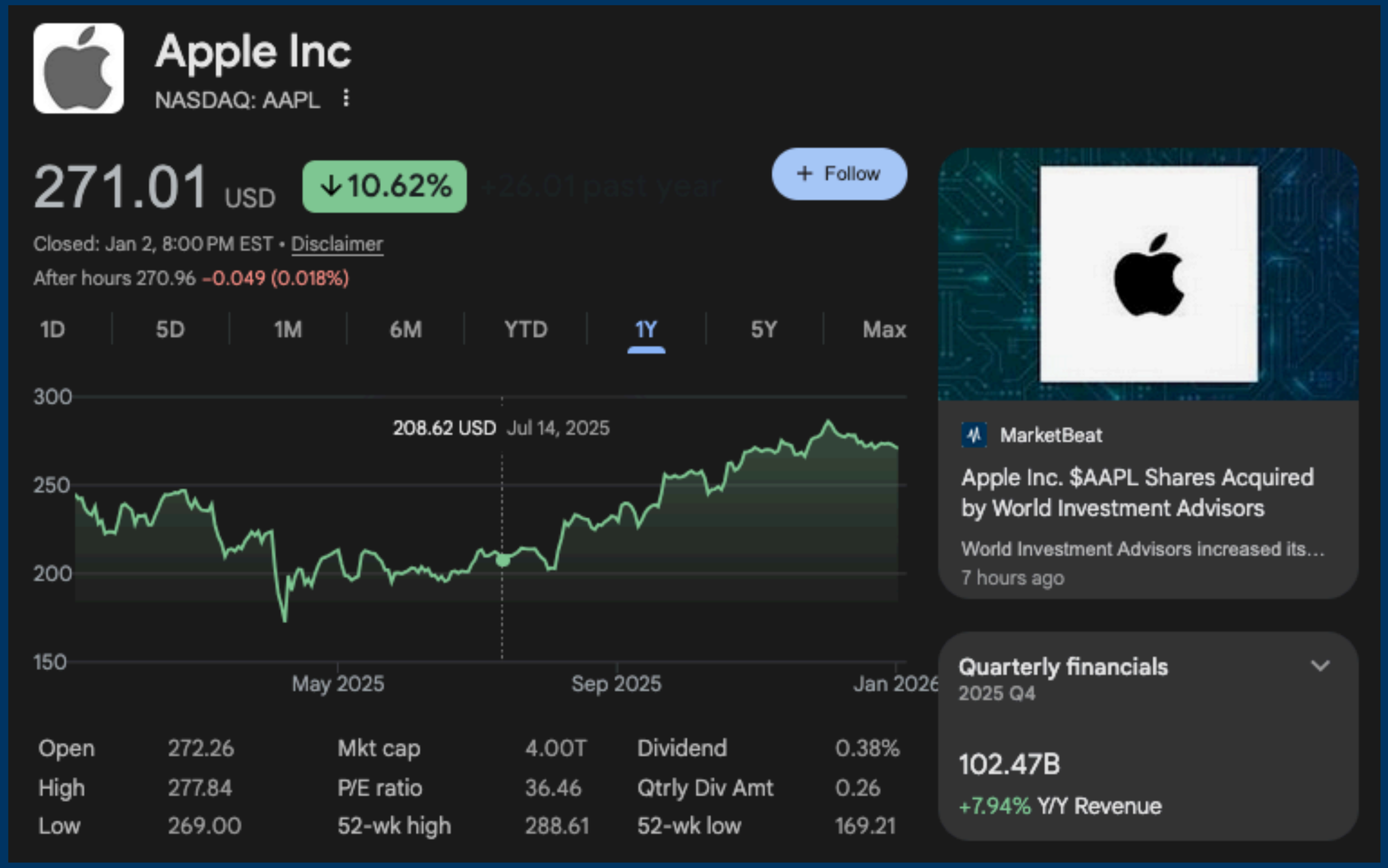
- When can you perform an A/B test?

A/B - Testing

- When can you perform an A/B test?
    - We assume your data is i.i.d. or collected in i.i.d fashion.

A/B - Testing

- **Do you think this data is i.i.d?**

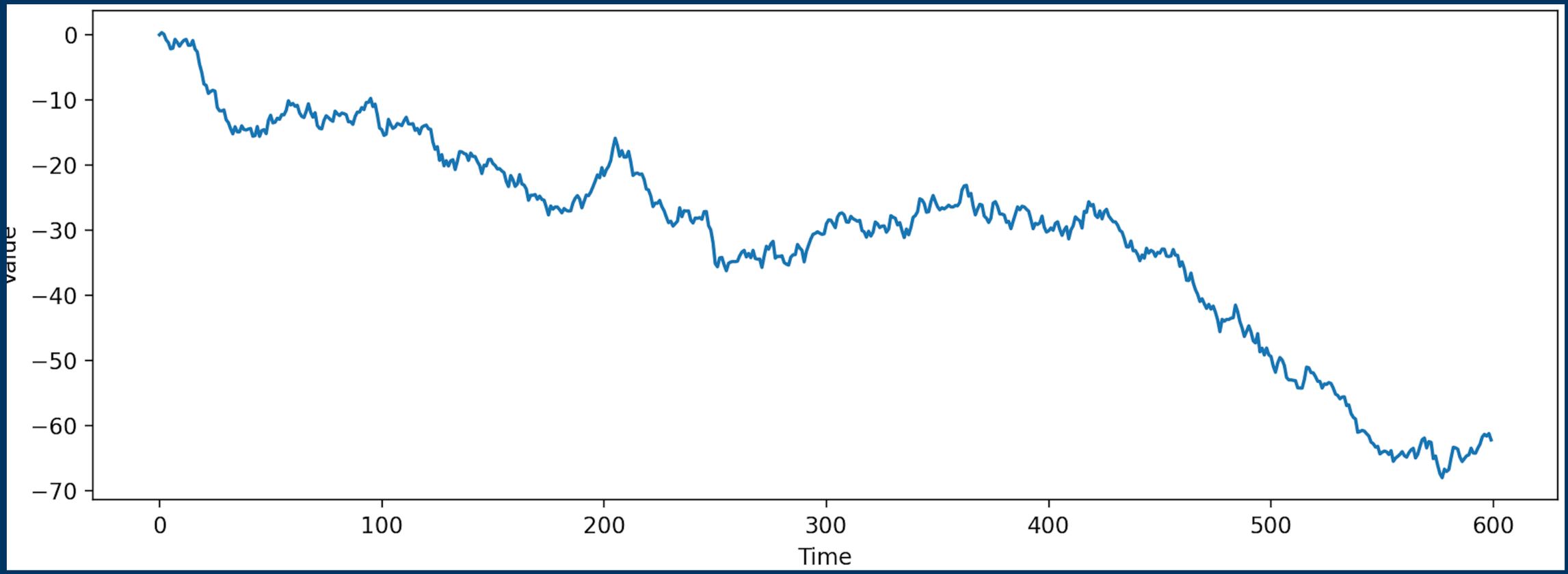


A/B - Testing

- When can you perform an A/B test?
    - We assume your data is i.i.d. or collected in i.i.d fashion.
    - We assume the data is “stationary”, and most business metrics are not stationary, rendering your A/B test dubious at best.

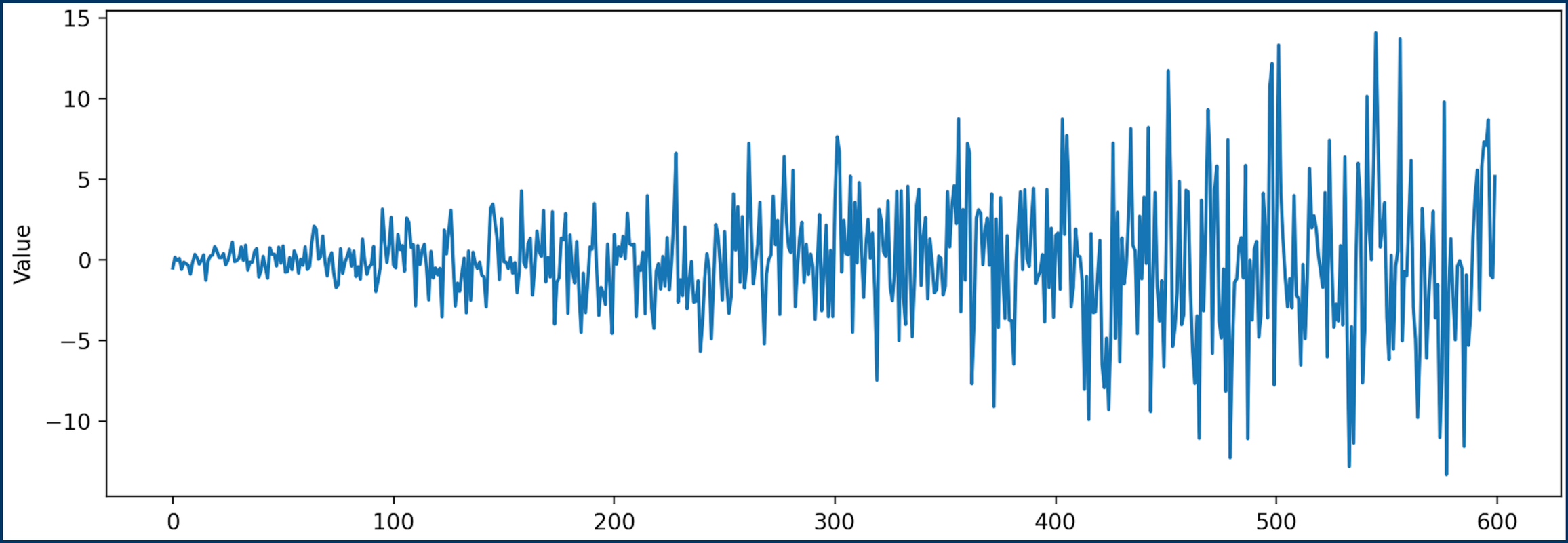
A/B - Testing

- **Even if you don’t know what it means for data to be “Stationary”, I want you to tell me which of these series seems stationary?**



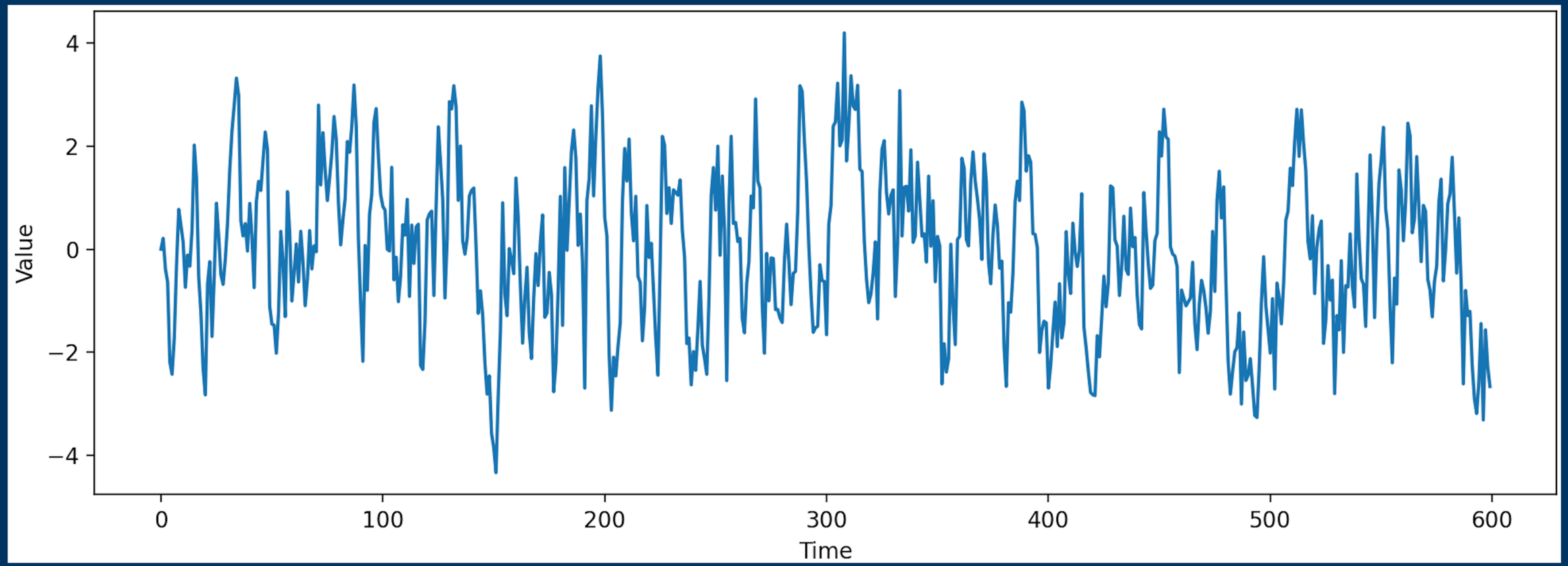
A/B - Testing

- **Even if you don’t know what it means for data to be “Stationary”, I want you to tell me which of these series seems stationary?**



A/B - Testing

- **Even if you don’t know what it means for data to be “Stationary”, I want you to tell me which of these series seems stationary?**

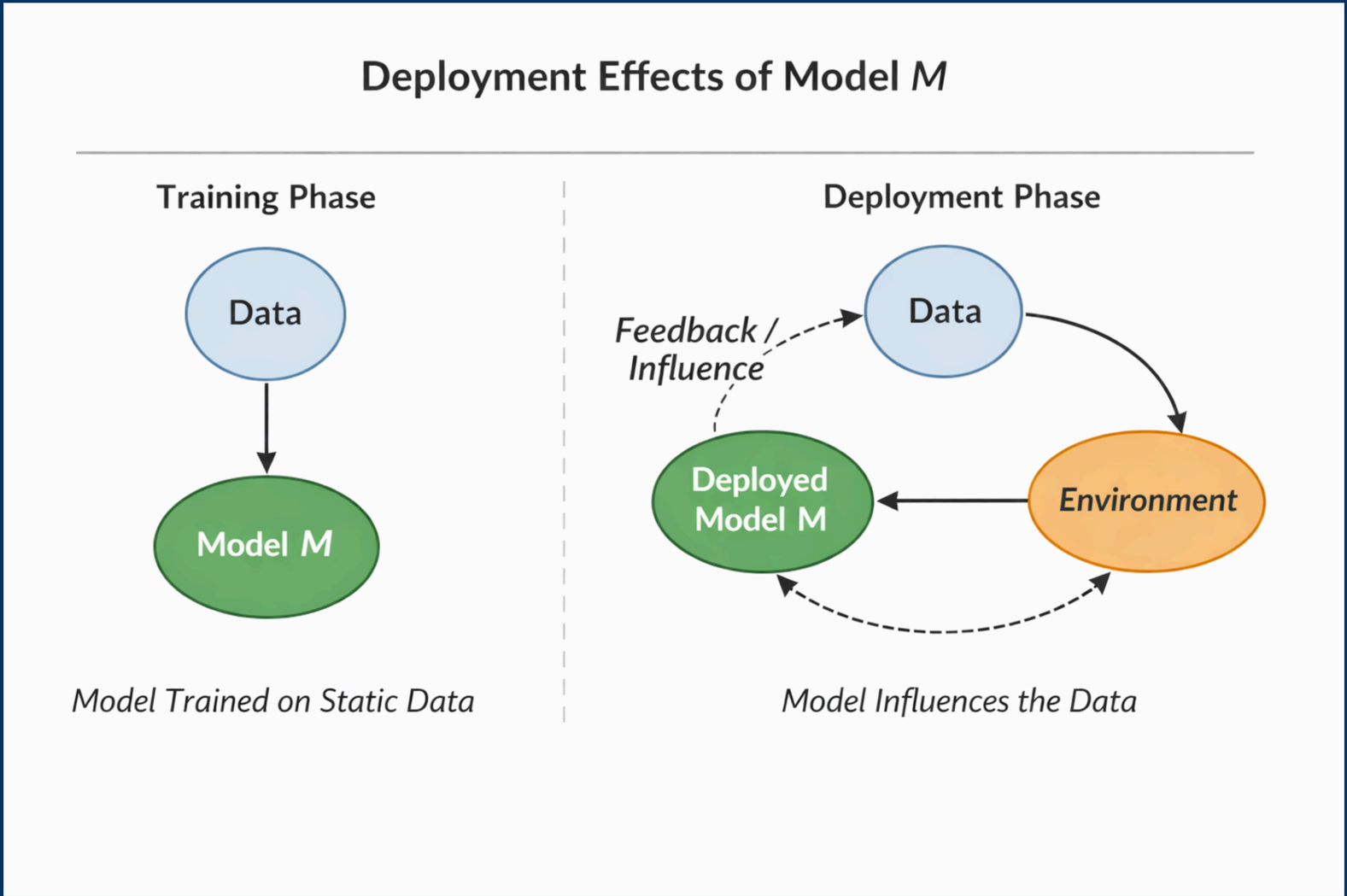


A/B - Testing

- When can you perform an A/B test?
    - We assume your data is i.i.d. or collected in i.i.d fashion.
    - We assume the data is “stationary”, and most business metrics are not stationary, rendering your A/B test dubious at best.
    - We assume no transient effects.

A/B - Testing

- **You train your model on data when your model didn’t exist and couldn’t influence the data distribution.**



A/B - Testing

- When can you perform an A/B test?
  - We assume your data is i.i.d. or collected in i.i.d fashion.
  - We assume the data is “stationary”, and most business metrics are not stationary, rendering your A/B test dubious at best.
  - We assume no transient effects.
  - We assume there is enough signal in the data to draw a meaningful conclusion.

A/B - Testing

- **Not enough signal?**

- https://www.tum.de/en/news-and-events/all-news/press-releases/details/40-percent-of-mri-signals-do-not-correspond-to-actual-brain-activity
- https://neurosciencenews.com/fmri-neural-activity-30057/

A/B - Testing

- **Not enough signal?**

- https://www.tum.de/en/news-and-events/all-news/press-releases/details/40-percent-of-mri-signals-do-not-correspond-to-actual-brain-activity
- https://neurosciencenews.com/fmri-neural-activity-30057/

A/B - Testing

- **Not enough signal?**

- https://www.tum.de/en/news-and-events/all-news/press-releases/details/40-percent-of-mri-signals-do-not-correspond-to-actual-brain-activity
- https://neurosciencenews.com/fmri-neural-activity-30057/

A/B - Testing

- **Not enough signal?**

Key Facts:

- **Mismatch Revealed:** In roughly 40% of cases, higher fMRI signals were linked to lower neural activity.
- **Oxygen Efficiency Shift:** Brain regions often meet extra energy demand by extracting more oxygen instead of increasing blood flow.
- **Clinical Impact:** fMRI findings in depression, Alzheimer's, and aging may reflect vascular differences rather than true neural activation changes.

- https://www.tum.de/en/news-and-events/all-news/press-releases/details/40-percent-of-mri-signals-do-not-correspond-to-actual-brain-activity
- https://neurosciencenews.com/fmri-neural-activity-30057/

Let’s Begin

A/B - Testing (Stock Exchange Rates)

- **We consider a hypothetical example to see if trading AAPL has a lower execution cost on ASDAQ v.s. BYSE.**

- We follow examples (with some modifications) from: `Experimentation for Engineers (From A/B testing to Bayesian optimization`, by Tim Sweet

A/B - Testing (Stock Exchange Rates)

- We consider a hypothetical example to see if trading AAPL has a lower execution cost on ASDAQ v.s. BYSE.

- We perform a very simple simulation to demonstrate what an A/B test looks like under ideal conditions.

- We follow examples (with some modifications) from: `Experimentation for Engineers (From A/B testing to Bayesian optimization`, by Tim Sweet

A/B - Testing (Stock Exchange Rates)

- We consider a hypothetical example to see if trading AAPL has a lower execution cost on ASDAQ v.s. BYSE.

- We perform a very simple simulation to demonstrate what an A/B test looks like under ideal conditions.

- For the sake of this experiment, we suspect that it might be cheaper to trade on BYSE

- We follow examples (with some modifications) from: `Experimentation for Engineers (From A/B testing to Bayesian optimization`, by Tim Sweet

A/B - Testing (Stock Exchange Rates)

- **Before experimenting on real data and potentially losing money. How can we simulate this experiment?**

```python
def trading_system(exchange: str) -> float:
    if exchange == "ASDAQ":
        execution_cost = 12.0
    elif exchange == "BYSE":
        execution_cost = 10.0
    else:
        raise ValueError("Exchange Not supported")
    execution_cost += np.random.normal()
    return execution_cost
```

A/B - Testing (Stock Exchange Rates)

- **Running the exchange once. Can we conclude that B is better than A?**

```
np.random.seed(17)
print(trading_system("ASDAQ"))
print(trading_system("BYSE"))
```

```
12.276265890021 32
8.145371921193496
```

A/B - Testing (Stock Exchange Rates)

- **Running the exchange once. Can we conclude that B is better than A?**

```
np.random.seed(17)
print(trading_system("ASDAQ"))
print(trading_system("BYSE"))
```

```
12.276265890021328
8.145371921193496
```

```
np.random.seed(18)
print(trading_system("ASDAQ"))
print(trading_system("BYSE"))
```

```
12.079428443806204
12.190202357414222
```

A/B - Testing (Stock Exchange Rates)

- If a single comparison isn’t enough, what should we do?

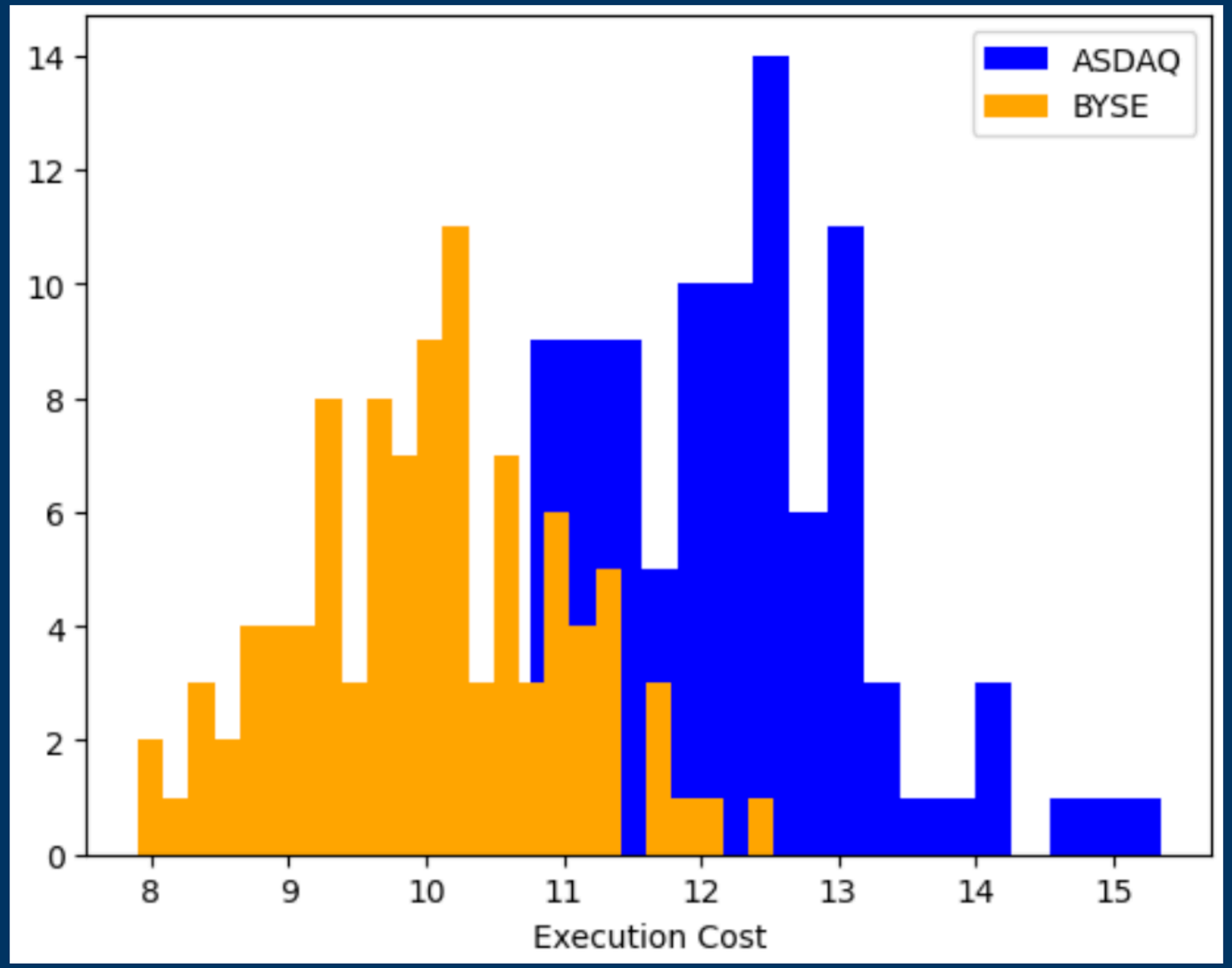
A/B - Testing (Stock Exchange Rates)

- **If a single comparison isn’t enough, what should we do?**

```
1  np.random.seed(17)
2  a = np.array([trading_system("ASDAQ") for _ in range(100)])
3  b = np.array([trading_system("BYSE") for _ in range(100)])
4
5  plt.hist(a, 25, color ='blue')
6  plt.hist(b, 25, color ='orange')
7  plt.legend(["ASDAQ", "BYSE"])
8  plt.xlabel("Execution Cost")
9  plt.show()
```

A/B - Testing (Stock Exchange Rates)

- If a single comparison isn't enough, what should we do?



A/B - Testing (Stock Exchange Rates)

- What can we conclude?

A/B - Testing (Stock Exchange Rates)

- What can we conclude?

  - Perhaps the averages of the distributions are different, suggesting that BYSE is cheaper.

A/B - Testing (Stock Exchange Rates)

- What can we conclude?
  - Perhaps the averages of the distributions are different, suggesting that BYSE is cheaper.
- What assumptions did we make? Is this an appropriate way to run an A/B test?

A/B - Testing (Stock Exchange Rates)

- What can we conclude?
  - Perhaps the averages of the distributions are different, suggesting that BYSE is cheaper.
- What assumptions did we make? Is this an appropriate way to run an A/B test?
  - We are assuming our samples are i.i.d.

A/B - Testing (Stock Exchange Rates)

- What can we conclude?
  - Perhaps the averages of the distributions are different, suggesting that BYSE is cheaper.
- What assumptions did we make? Is this an appropriate way to run an A/B test?
  - We are assuming our samples are i.i.d.
  - We are assuming that we are sampling a stationary distribution.

A/B - Testing (Stock Exchange Rates)

- What can we conclude?
  - Perhaps the averages of the distributions are different, suggesting that BYSE is cheaper.
- What assumptions did we make? Is this an appropriate way to run an A/B test?
  - We are assuming our samples are i.i.d.
  - We are assuming that we are sampling a stationary distribution.
    - This is true because we construct the exchanges to have fixed rates, and the variance introduce by the random noise is fixed.

A/B - Testing (Stock Exchange Rates)

- Are there any problems with our analysis so far?

A/B - Testing (Stock Exchange Rates)

- Are there any problems with our analysis so far?

- Consider a simple example how the time of day can affect our collection of samples:

A/B - Testing (Stock Exchange Rates)

- Are there any problems with our analysis so far?

- Consider a simple example how the time of day can affect our collection of samples:

  - Suppose we need a massive amount of measurements to detect the signal:

A/B - Testing (Stock Exchange Rates)

- Are there any problems with our analysis so far?

- Consider a simple example how the time of day can affect our collection of samples:

  - Suppose we need a massive amount of measurements to detect the signal:

    - So we take many measurements of BYSE in the morning.

A/B - Testing (Stock Exchange Rates)

- Are there any problems with our analysis so far?

- Consider a simple example how the time of day can affect our collection of samples:

  - Suppose we need a massive amount of measurements to detect the signal:

    - So we take many measurements of BYSE in the morning.

    - Once that process finishes, we then take many measurements of ASDAQ (now in the afternoon)

- **What if trading regardless of exchange is cheaper in the afternoon?**

A/B - Testing (Stock Exchange Rates)

- The main point is when performing an A/B test, it’s critical to take a random sample of the data.

A/B - Testing (Stock Exchange Rates)

- The main point is when performing an A/B test, it’s critical to take a random sample of the data.

- How do we simulate the stock exchanges with this time-of-day (TOD) bias?

A/B - Testing (Stock Exchange Rates)

- **The main point is when performing an A/B test, it's critical to take a random sample of the data.**

- **How do we simulate the stock exchanges with this time-of-day (TOD) bias?**

```python
def trading_system_tod(exchange: str, time_of_day: str) -> float:
    '''
    Trading system with time of day (tod) effect
    '''
    if time_of_day == "morning":
        bias = 2.5
    elif time_of_day == "afternoon":
        bias = 0.0
    else:
        raise ValueError('Not a valid time of day!')
    return trading_system(exchange) + bias
```

A/B - Testing (Stock Exchange Rates)

- **Compare the exchanges with TOD effect.**

A/B - Testing (Stock Exchange Rates)

- **Compare the exchanges with TOD effect.**

```
np.random.seed(17)
print(np.array([trading_system_tod("BYSE", "morning") for _ in range(100)]).mean())
print(np.array([trading_system_tod("ASDAQ", "afternoon") for _ in range(100)]).mean())
```

A/B - Testing (Stock Exchange Rates)

- **Compare the exchanges with TOD effect.**

```
np.random.seed(17)
print(np.array([trading_system_tod("BYSE", "morning") for _ in range(100)]).mean())
print(np.array([trading_system_tod("ASDAQ", "afternoon") for _ in range(100)]).mean())
```

```
12.611509794247766
12.008382946497411
```

A/B - Testing (Stock Exchange Rates)

- If we didn’t think about the TOD effect we’d conclude that ASDAQ is cheaper than BYSE (which is wrong).

A/B - Testing (Stock Exchange Rates)

- If we didn’t think about the TOD effect we’d conclude that ASDAQ is cheaper than BYSE (which is wrong).

- This is known as a confounder bias.

A/B - Testing (Stock Exchange Rates)

- If we didn’t think about the TOD effect we’d conclude that ASDAQ is cheaper than BYSE (which is wrong).
- This is known as a confounder bias.
- The remedy for this is to introduce randomization.

A/B - Testing (Stock Exchange Rates)

- If we didn't think about the TOD effect we'd conclude that ASDAQ is cheaper than BYSE (which is wrong).
- This is known as a confounder bias.
- The remedy for this is to introduce randomization.
- To mitigate bias effects that we may or may not be aware of, it's always a good idea to compare random samples of the data. (This is a very common mistake I see in a lot of businesses).

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

- Every time the system makes a trade, it flips a coin:

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

- Every time the system makes a trade, it flips a coin:
  - H —> Sample from ASDAQ.

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

- Every time the system makes a trade, it flips a coin:
  - H —> Sample from ASDAQ.
  - T —> Sample from BYSE.

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

- Every time the system makes a trade, it flips a coin:
  - H —> Sample from ASDAQ.
  - T —> Sample from BYSE.
  - So about half of the morning trades are on A, and the other half are on B.

A/B - Testing (Stock Exchange Rates)

- But it takes a long time to run our system, If we take the A samples first, it will take all morning, before we can start sampling B. How do we fix this?

- Every time the system makes a trade, it flips a coin:
  - H —> Sample from ASDAQ.
  - T —> Sample from BYSE.
  - So about half of the morning trades are on A, and the other half are on B.
  - Likewise about half of the trades on A are in the afternoon, and the other half are on B.

A/B - Testing (Stock Exchange Rates)

- **Randomizing Measurements:**

```
def randomized_measurement():
    asdaq_measurement = []
    byse_measurement  = []
    for tod in ["morning", "afternoon"]:
        for _ in range(100):
            if np.random.randint(2) == 0:
                asdaq_measurement.append(trading_system_tod("ASDAQ", tod))
            else:
                byse_measurement.append(trading_system_tod("BYSE", tod))
    return (np.array(asdaq_measurement).mean(),
                    np.array(byse_measurement).mean()
                    )
```

A/B - Testing (Stock Exchange Rates)

- **Randomizing Measurements:**

```
np.random.seed(17)
print(randomized_measurement())
```

A/B - Testing (Stock Exchange Rates)

- **Randomizing Measurements:**

```
np.random.seed(17)
print(randomized_measurement())
```

```
(np.float64(13.395888706238 52), np.float64(11.259639285763223))
```

A/B - Testing (Stock Exchange Rates)

- **Randomizing Measurements:**

```
np.random.seed(17)
print(randomized_measurement())
```

```
(np.float64(13.395888706238 52), np.float64(11.259639285763223))
```

- **We see that ASDAQ is more expensive than BYSE (which is correct).**

A/B - Testing (Stock Exchange Rates)

- **Randomizing Measurements:**

```
np.random.seed(17)
print(randomized_measurement())
```

```
(np.float64(13.395888706238 52), np.float64(11.259639285763223))
```

- **We see that ASDAQ is more expensive than BYSE (which is correct).**

- **We don’t need to be aware of the TOD effect if we simply introduce randomization.**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**
  - **We set the avg. exchange rate of ASDAQ to be 12.0.**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**
  - **We set the avg. exchange rate of ASDAQ to be 12.0.**

  - **Let’s consider what happens if we sample 1 measurement from ASDAQ and compare it to the true rate of 12.0.**

A/B - Testing (Stock Exchange Rates)

- Means and Variance:
  - We set the avg. exchange rate of ASDAQ to be 12.0.
  - Let’s consider what happens if we sample 1 measurement from ASDAQ and compare it to the true rate of 12.0.
  - Then let’s take three samples and then average the samples and then compare to 12.0

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

```
1  np.random.seed(17)
2  measurements = np.array([trading_system("ASDAQ") for _ in range(3)])
3  print(measurements - 12.0)
4  print(measurements.mean() - 12.0)
```

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

```
1  np.random.seed(17)
2  measurements = np.array([trading_system("ASDAQ") for _ in range(3)])
3  print(measurements - 12.0)
4  print(measurements.mean() - 12.0)
```

```
[ 0.27626589 -1.85462808  0.62390111]
-0.3181536924862769
```

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

- **We see the mean is very close to the true exchange rate of 12.0**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

- **We see the mean is very close to the true exchange rate of 12.0**
- **The individual samples can be as far away as 1.85**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

- **We see the mean is very close to the true exchange rate of 12.0**
- **The individual samples can be as far away as 1.85**
- **What happened?**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

- **We see the mean is very close to the true exchange rate of 12.0**
- **The individual samples can be as far away as 1.85**
- **What happened?**
  - **The variance of the mean of three samples is smaller than the variance of individual samples.**

A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**

- **The function below let's us take a specified number of measurements and then sample an exchange of choice:**

```python
def aggregate_measurement(exchange: str, num_individual_measurements: int) -> float:
    individual_measurements = np.array([
        trading_system(exchange)                    # take an individual measurement
        for _ in range(num_individual_measurements) # multiple times
    ])
    return individual_measurements.mean()          # then average
```

A/B - Testing (Stock Exchange Rates)

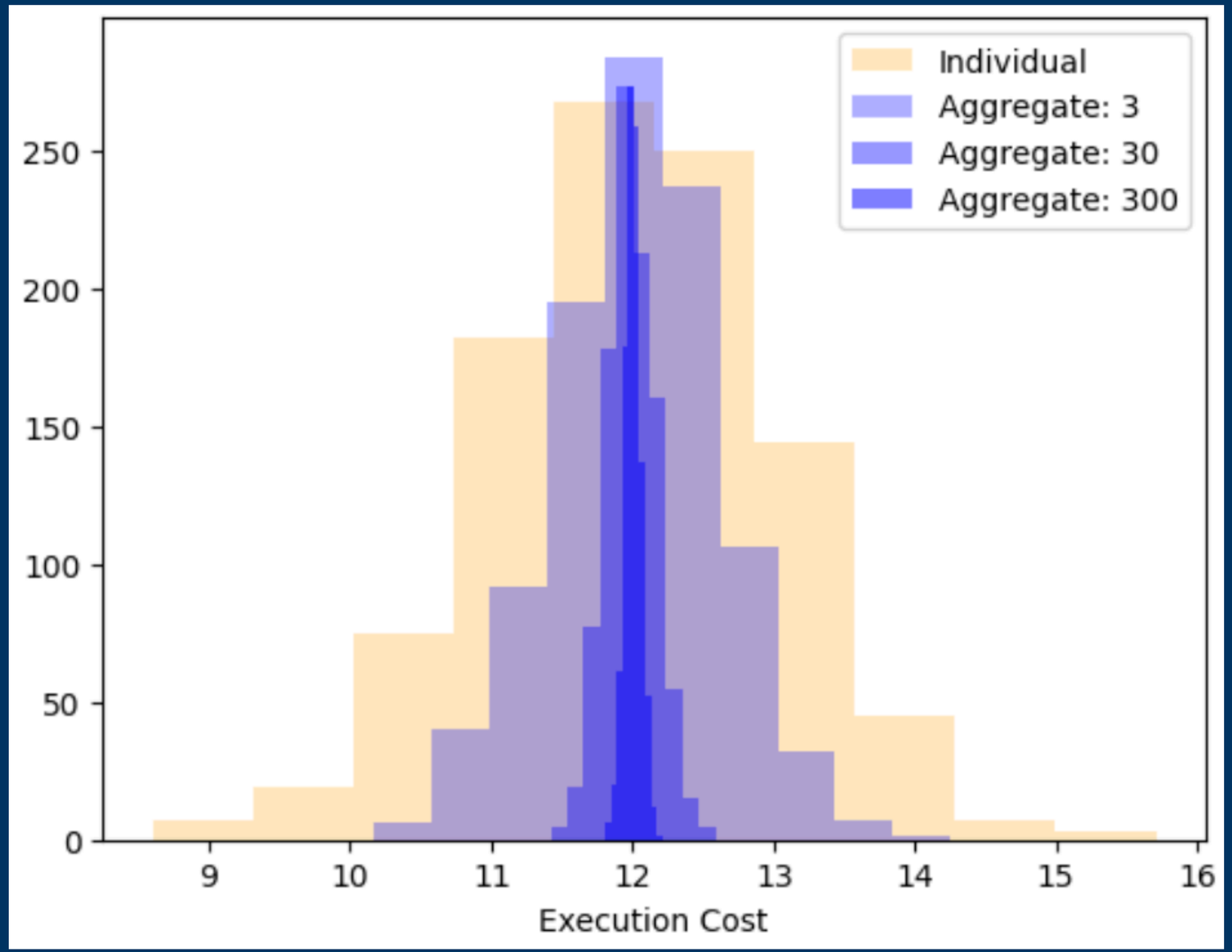
- **Means and Variance:**

- **The function below let's us take a specified number of measurements and then sample an exchange of choice:**

```
np.random.seed(17)
plt.hist(np.array([trading_system("ASDAQ")              for _ in range(1000)]), color = 'orange', alpha = 0.25)
plt.hist(np.array([aggregate_measurement("ASDAQ", 3)   for _ in range(1000)]), color = "blue", alpha = 0.30)
plt.hist(np.array([aggregate_measurement("ASDAQ", 30)  for _ in range(1000)]), color = "blue", alpha = 0.40)
plt.hist(np.array([aggregate_measurement("ASDAQ", 300) for _ in range(1000)]), color = "blue", alpha = 0.50)
plt.xlabel("Execution Cost")
plt.legend(["Individual", "Aggregate: 3", "Aggregate: 30", "Aggregate: 300"])
plt.show()
```

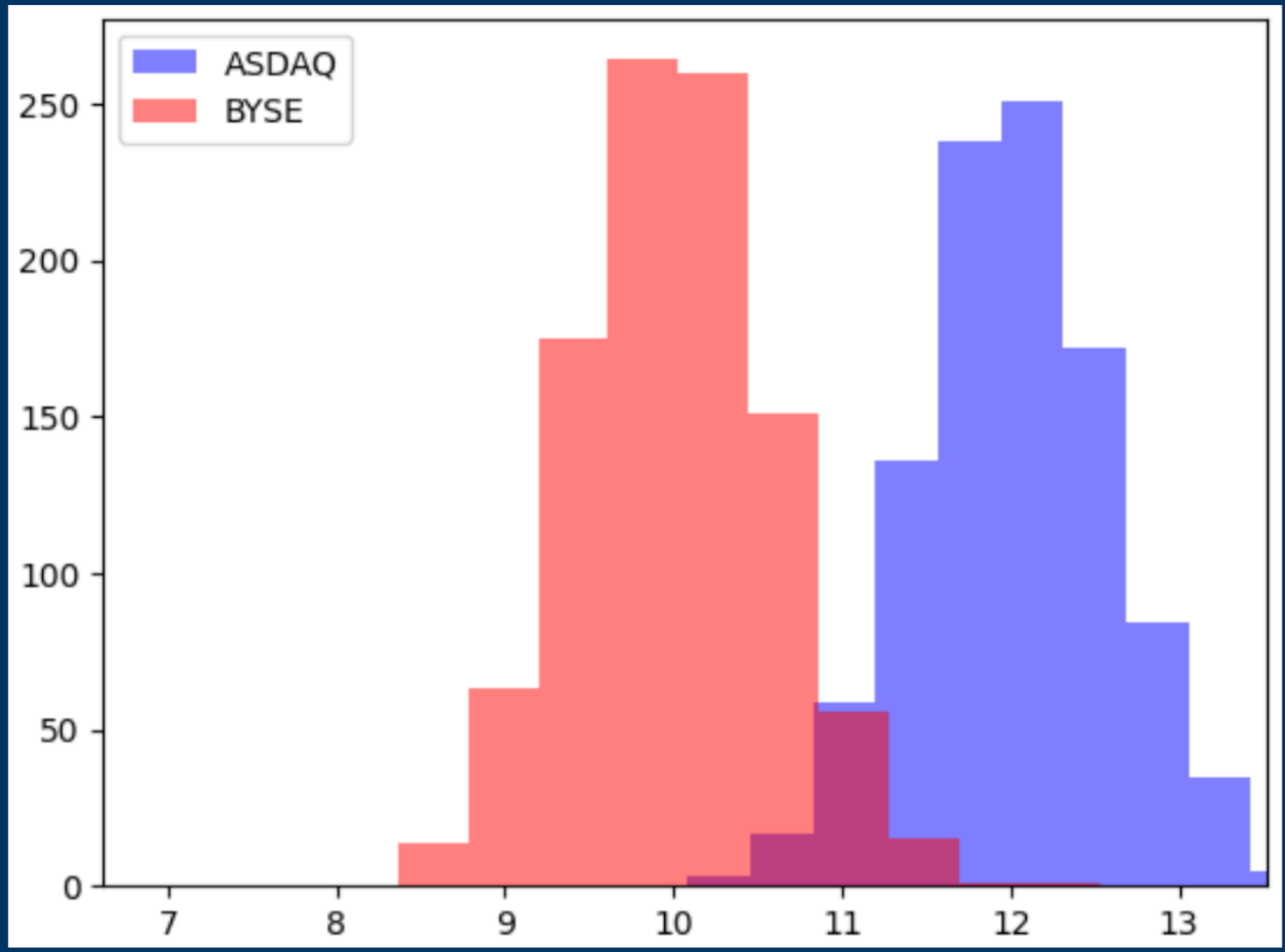
A/B - Testing (Stock Exchange Rates)

- **Means and Variance:**



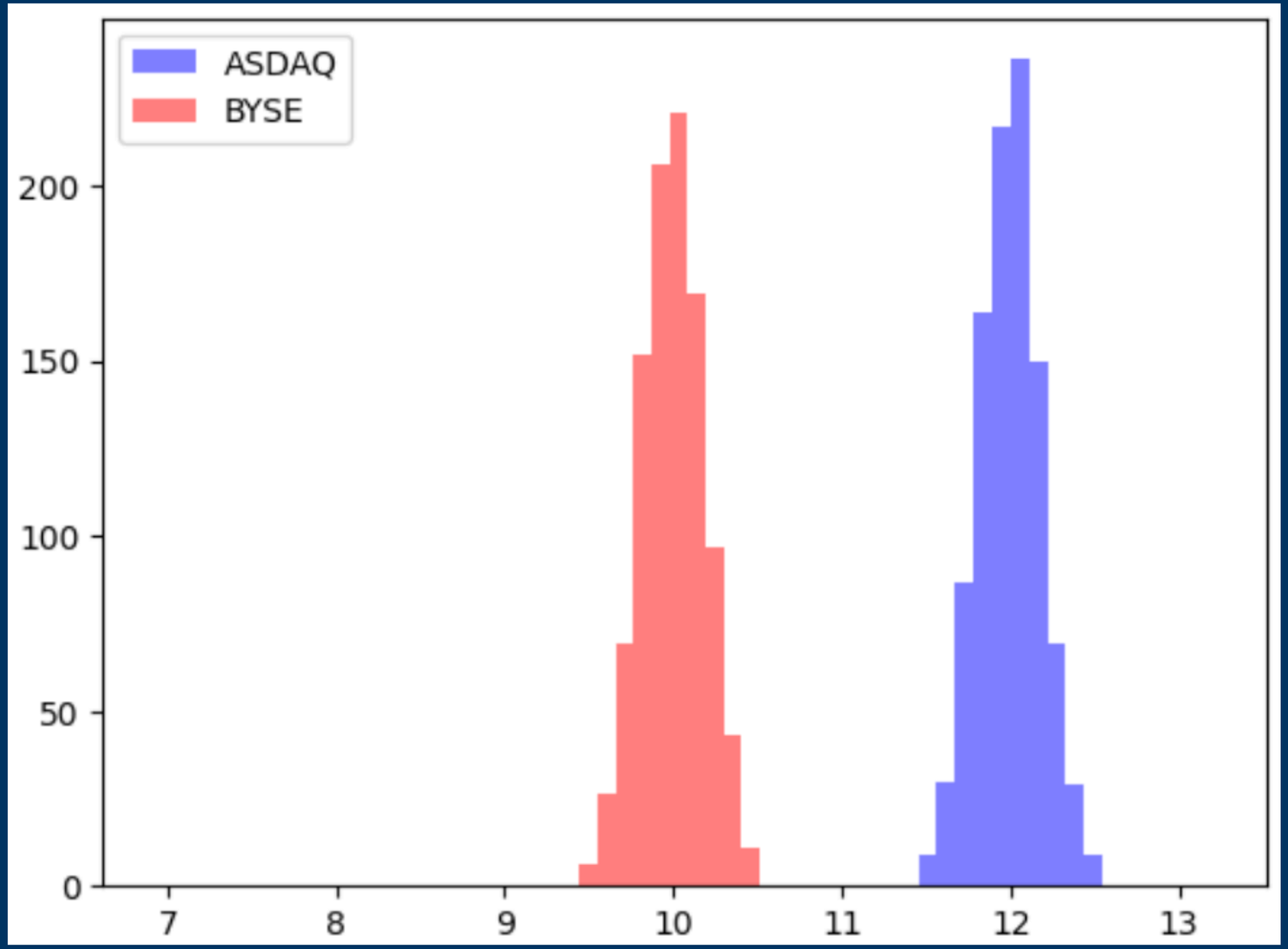
A/B - Testing (Stock Exchange Rates)

- **Comparing Two exchanges 3 samples:**



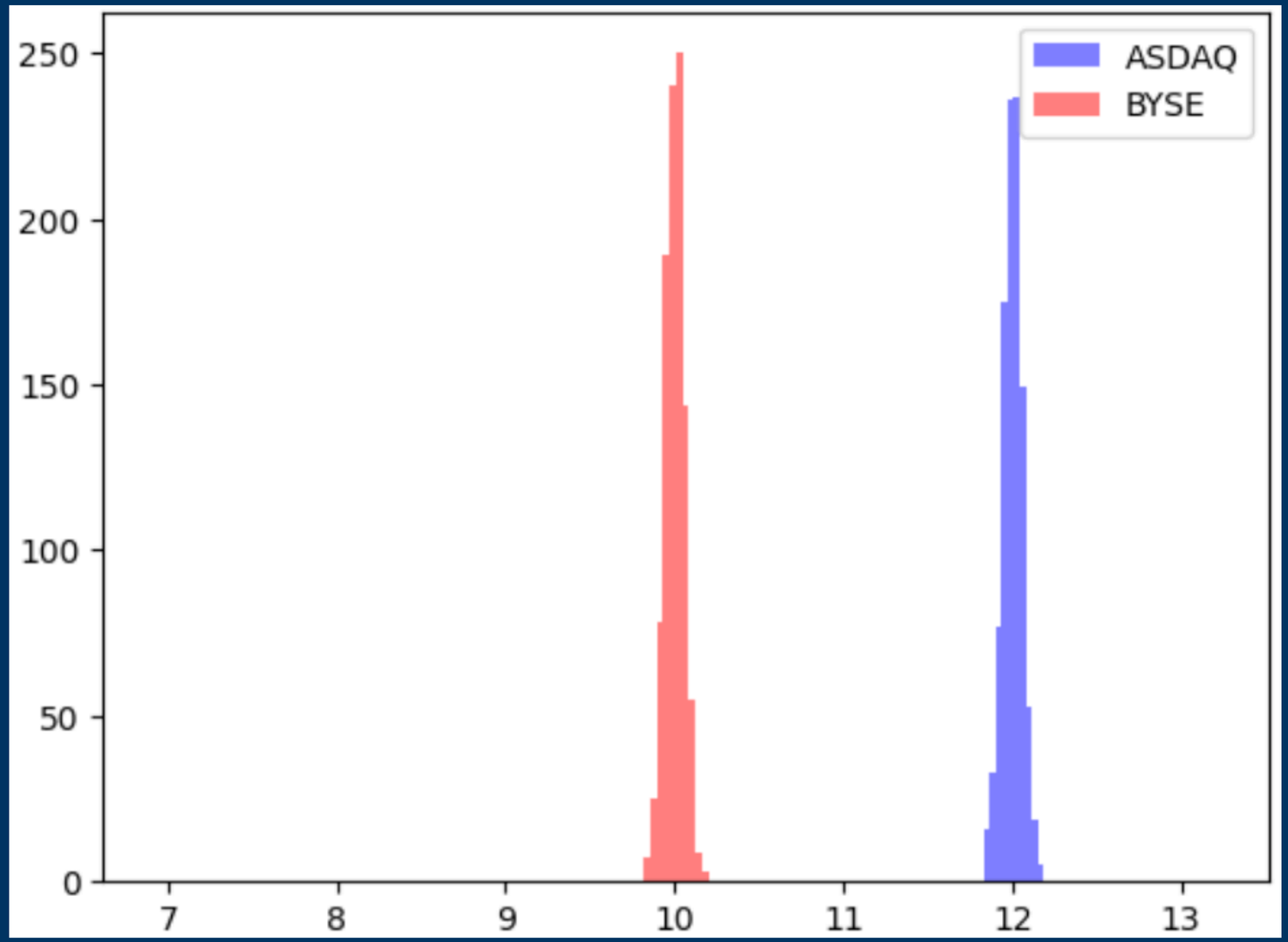
A/B - Testing (Stock Exchange Rates)

- **Comparing Two exchanges 30 samples:**



A/B - Testing (Stock Exchange Rates)

- **Comparing Two exchanges 300 samples:**



Thank You

(Up next: SE, z-scores, False Positives, FDR-corrections)

Contact Me: andrewlizarraga@g.ucla.edu