# The Art and Science of A/B Testing: A Deep Dive for Aspiring Data Scientists

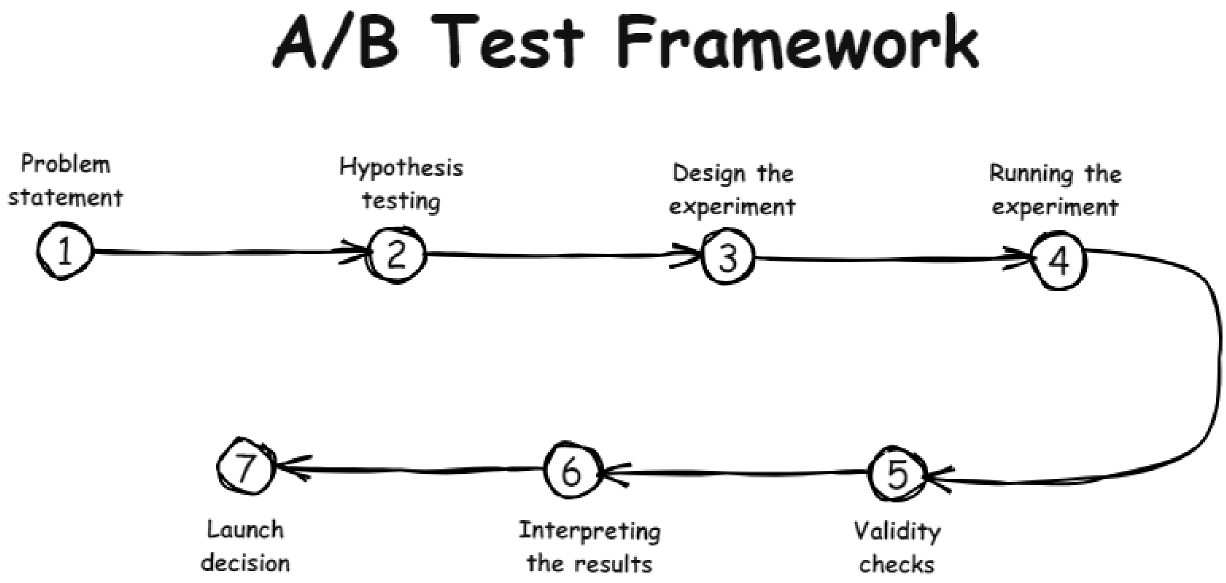
A/B testing is one of the critical components in the daily responsibilities of a data scientist. Understanding this concept can dramatically influence the decision-making process in business through data-driven insights. This article will guide you through the essentials of A/B testing, leveraging real-life examples and insights drawn from credible sources on the internet.

## Understanding the Purpose of A/B Testing

A/B testing, at its core, is a method for comparing two versions of a webpage or app against each other to determine which one performs better. It involves two groups: a control group and a treatment group. The goal is to identify whether the changes made in the treatment group produce a significant improvement over the control group.

Why is this important? In tech companies, data scientists use A/B tests to validate the impact of new features or changes on user behavior and business metrics. This method helps ensure that decisions made are not based on assumptions or random chance, but on solid, statistical evidence.

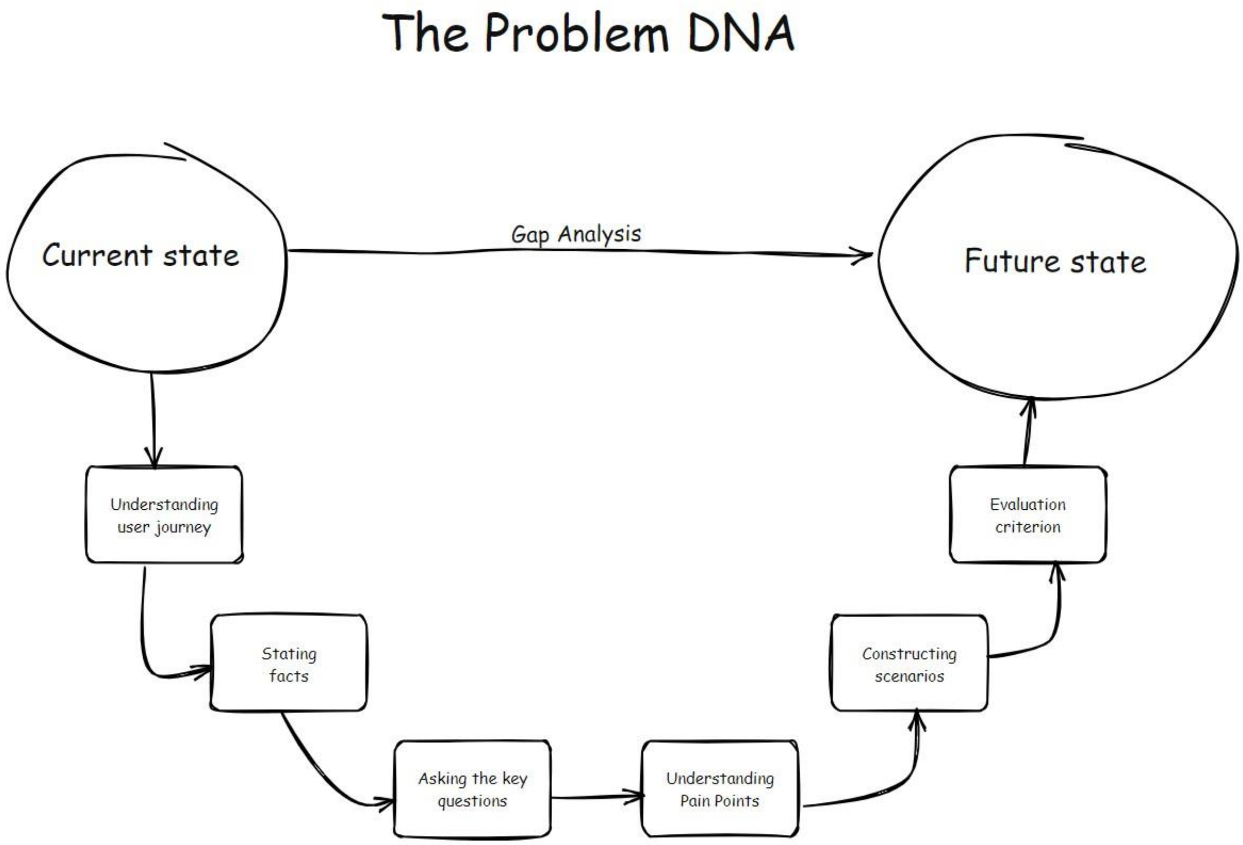
## Setting Up for Success: The A/B Testing Framework



Outlined below is a systematic approach to conducting an A/B test, which can be broken down into several key steps:

1. Problem Understanding: The first step involves clearly defining the problem. What is the business trying to improve? Which part of the user experience is under scrutiny? This

phase includes determining the success metrics and understanding the user journey. One way to do this is through building the problem DNA.


The Problem DNA

The DNA is a way to visualize the problem with a 30K ft view. It covers the essential segments that should be considered to fill in the gap from the current state and reach the target state, i.e., the state where we want to be at. This also helps nail down the answers to 4 qualifying questions such as below to identify and assess the success metric/s early in the process:

- Is the metric measurable with the existing instrumentation platform?
- Is our metric attributable, i.e., is there a clear linkage that the cause (the treatment) that we applied to the platform has led to the effect that we saw?
- Is our metric sensitive, i.e., statistically speaking, does the metric have low variability?
- Is the metric timely, i.e., can we measure the success behavior in a short term?

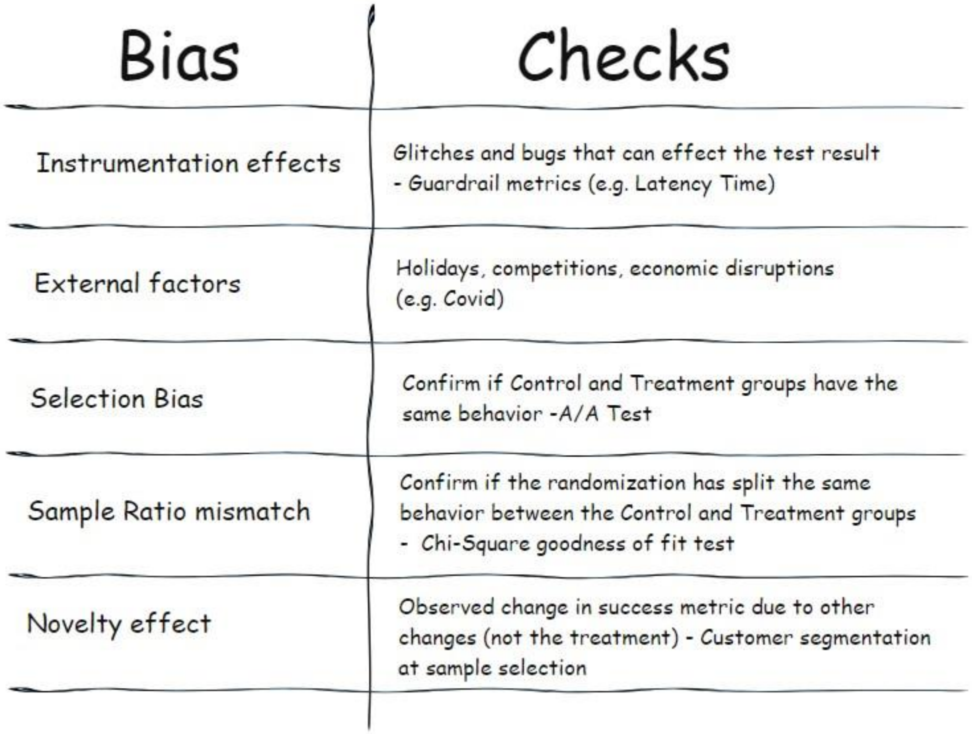
2. Hypothesis Formulation: Before any actual testing begins, it's crucial to formulate a hypothesis. This involves a few things:
    - Setting up a null hypothesis (no change or effect) and an alternative hypothesis (presence of a change or effect).

- Setting the significance level for accepting the alternate hypothesis or refuting the null hypothesis. It is basically the decision threshold. If the probability of observing a particular event is very low, then it is deemed statistically significant.
- Setting up the statistical power. It is basically the causality probability. If the statistical power is 0.8, let's say, then there's 80% probability of detecting an effect given that the alternative hypothesis is true.
- Finally, setting the practical significance, i.e., the minimum detectable effect (usually kept at 1% lift)

3. Designing the Experiment: This step involves planning how the test will be conducted. Key considerations include deciding on the randomization unit (e.g., users or user sessions), which user population of the user segment to target, determining the sample size, and the duration of the experiment.

4. Execution: Running the experiment requires collecting data in a way that is free from biases and other external influences that could skew the results. Note: We should not make any decision in terms of whether we're going to launch or not while the experiment hasn't been completed yet as there's a chance of Early adopter bias (This is the tendency to overestimate the feedback from the first and most enthusiastic users).

5. Validity Checks: Before analyzing the results, conducting sanity checks ensures that the data collected are reliable and the experiment setup was correct.



| Bias | Checks |
|---|---|
| Instrumentation effects | Glitches and bugs that can effect the test result<br>- Guardrail metrics (e.g. Latency Time) |
| External factors | Holidays, competitions, economic disruptions<br>(e.g. Covid) |
| Selection Bias | Confirm if Control and Treatment groups have the same behavior -A/A Test |
| Sample Ratio mismatch | Confirm if the randomization has split the same behavior between the Control and Treatment groups<br>- Chi-Square goodness of fit test |
| Novelty effect | Observed change in success metric due to other changes (not the treatment) - Customer segmentation at sample selection |

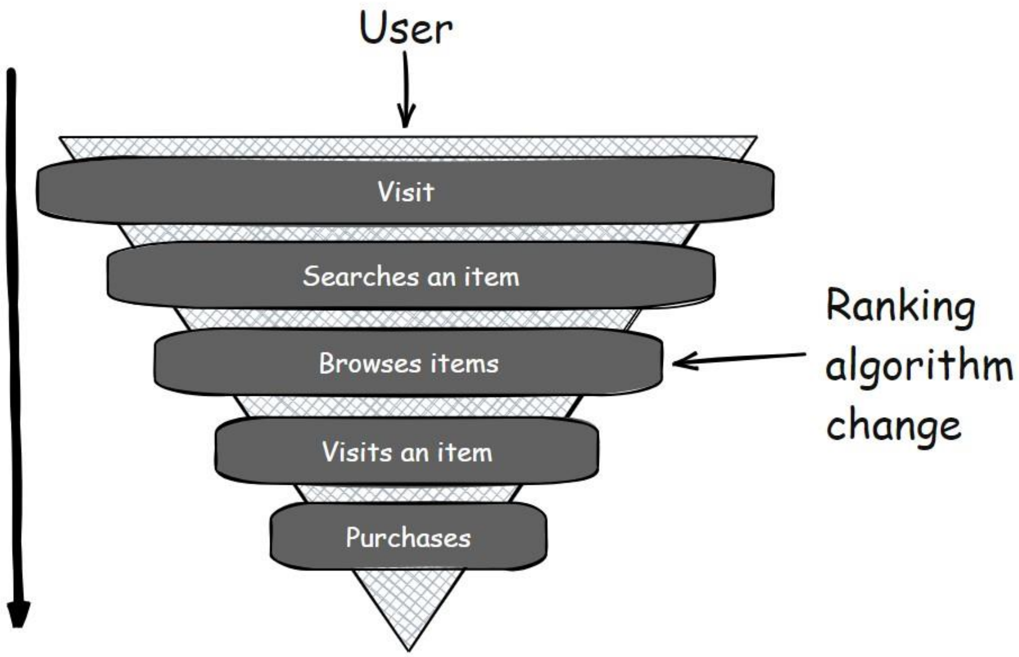
6. Interpreting Results: Analyzing the data to see if there is a statistically significant difference between the control and treatment groups. This includes looking at metrics like the p-value and confidence intervals.

7. Business Decision: The final step involves using both the statistical results and business context to decide whether to implement the change on a larger scale.

## Applying the Framework: A Real-World Example

Consider a hypothetical e-commerce store that wants to test a new algorithm for recommending products. The store hypothesizes that the new algorithm will increase user engagement and, consequently, sales.

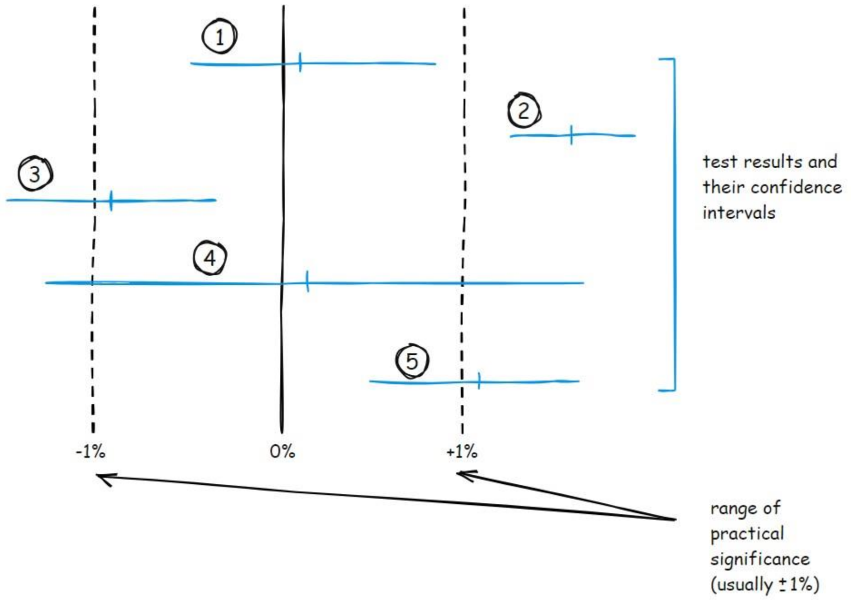
- User Journey: Customers visit the site, search for products, browse through the results, and make purchases. The experiment would ideally target users right from the search phase to ensure they are exposed to the new algorithm.



- Success Metrics: The primary metric based on the 4 qualifying questions could be revenue per day per user, as it directly reflects the effectiveness of the new algorithm in promoting sales.

- Experiment Design: Users are randomly assigned to either the control group (old algorithm) or the treatment group (new algorithm). The test would run for a sufficient period to gather actionable data, avoiding any special events or sales that might influence user behavior unnaturally.

- Launching decision: As you can see, among the 5 scenarios, only the second case has the lift (MDE) in bounds, the confidence interval is practically significant. So, this provides a strong support that we should make a launch. The rest of the cases, we either change the algorithm further, iterate the idea again, or scrap the change altogether.


1
2
3
4
5
test results and their confidence intervals
-1%
0%
+1%
range of practical significance (usually ±1%)

## Key Considerations and Pro Tips

- Sample Size and Duration: It's essential to calculate the right sample size and test duration to achieve statistically reliable results. This ensures that the findings are robust and can be generalized to the entire user base.

- Avoiding Early Conclusions: Data scientists must resist the temptation to make premature decisions based on initial trends in the data. Decisions should only be made after the full data set is analyzed post-experiment.

- Addressing Novelty Effects: Sometimes, users might react positively to a new feature simply because it's new, not necessarily better. Segmenting users into new versus returning groups can help identify and control for this novelty effect.

## Conclusion

A/B testing is more than just a technical skill—it's a critical thinking framework that requires understanding both the statistical underpinnings and the business implications of the data. For data scientists, mastering A/B testing means being able to guide business decisions with precision, ensuring that innovations truly enhance the user experience and contribute positively to the company's goals. Aspiring data scientists should approach A/B testing not just as a tool for their toolkit but as a fundamental skill that underpins much of what makes digital businesses successful today.