CLOUD ENGINEERING

A/B Testing

Ashish Pujari

Lecture Outline

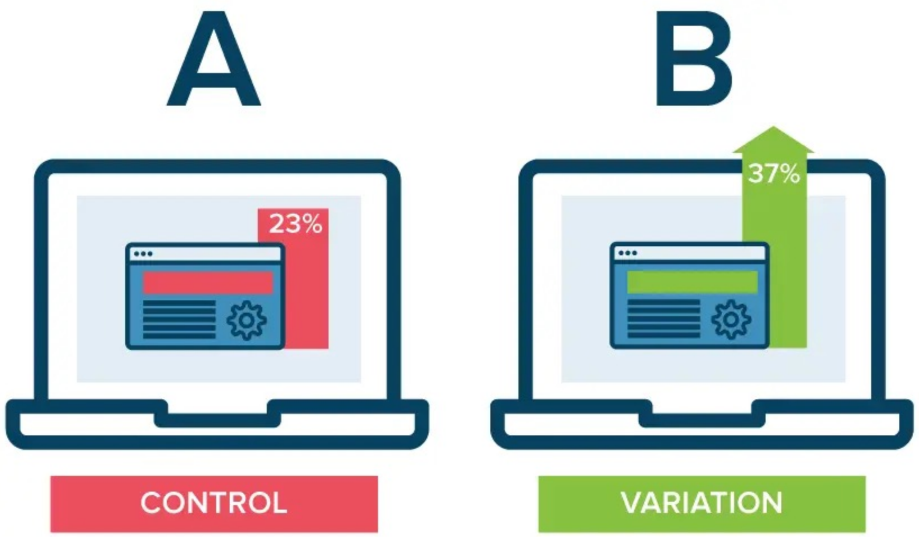
- A/B Testing
- Testing Process
- Multi-armed Bandits

A/B TESTING

Overview

A/B Testing

- Practice of making randomized experiments for optimizing business decisions
- Helps us learn which variation is more effective and make improvements accordingly.
- E.g., between two versions of a web page or a ranking algorithm and which one attracts more visitors or generates more sales.



Source: Optimizely

A/B Testing: Applications

- E-commerce
- Software Development
- Digital Advertising
- Content Publishing
- Mobile App Development
- Email Marketing
- Financial Services
- Recommender Systems

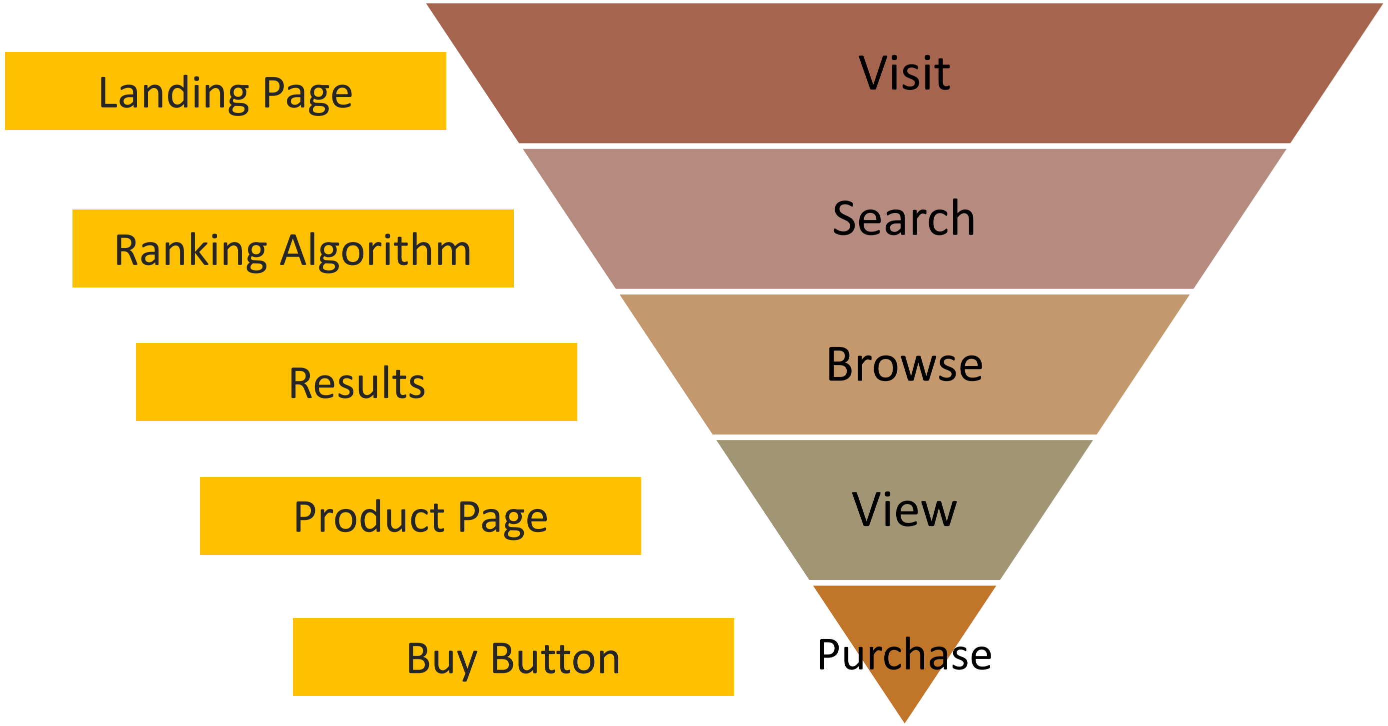
Machine Learning A/B Testing

- Recommendation Systems
- Search Engine Ranking
- Fraud Detection
- Ad Targeting
- Email Campaign Optimization

Example: Recommender Systems

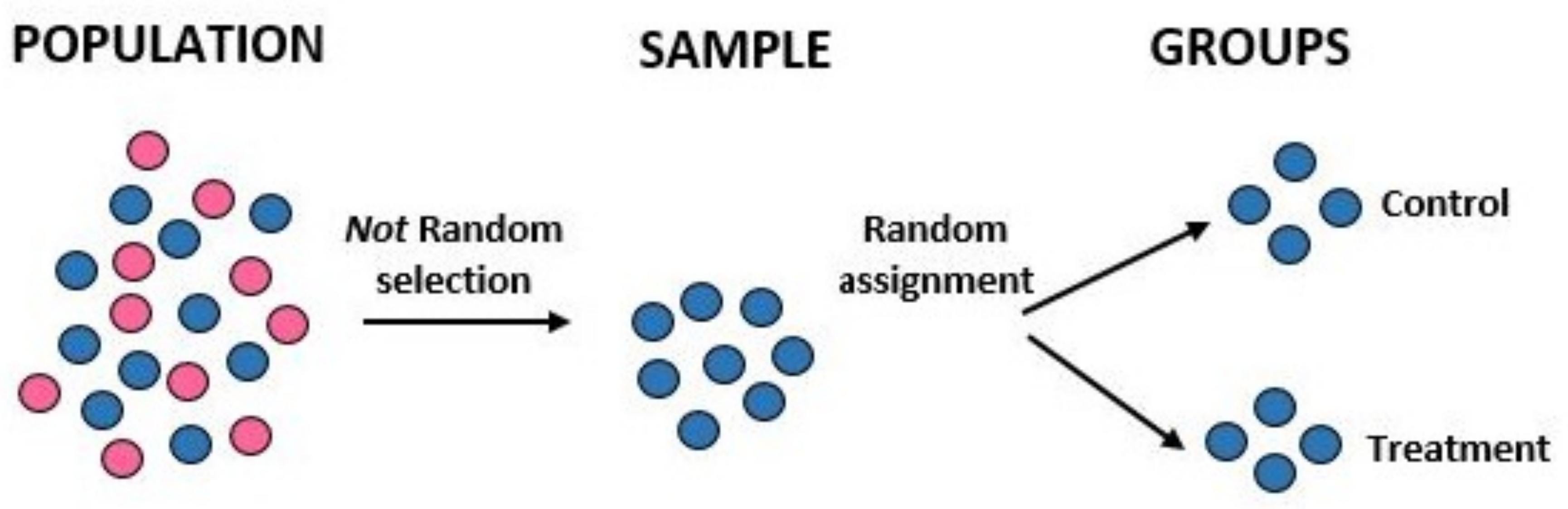
- Experiments
  - Recommendation Display
  - User Segmentation
  - Ranking Strategies
  - Parameter Optimization
- Benefits
  - Continuously improve the recommendation effectiveness
  - Deliver personalized, relevant, and engaging recommendations
  - Enhance user experience and drive desired user actions

User Journey Metrics



Control and Treatment Groups

- The control group (A) is the group that does not receive the treatment or change.
- The treatment group (B) is the group that receives the treatment or change.



Considerations

- Sample Size
  - Sufficient sample sizes are needed to obtain reliable results.
- Randomization
  - Ensures that each participant has an equal chance of being assigned to either group.
- Test Duration
  - Duration of the test should be long enough to capture variations in user behavior.
- Statistical Significance
  - Helps determine if the observed differences are likely due to chance or if they are meaningful.

A/B TESTING PROCESS

Formulate Hypothesis

- Null Hypothesis
  - Assumes no effect or difference
  - E.g., Average revenue per day between the baseline and variant ranking algorithms are the same; any observed difference is due to randomness
- Alternative Hypothesis
  - Assumes an effect or difference
  - E.g., Average revenue per day between the baseline and variant ranking algorithms are different.

2. Define Metrics

- Metric
  - Quantity used to measure the impact of your change
  - Should either be a KPI or directly related to a KPI
  - E.g., Conversion Rates, Mobile signups, Sales, Revenue, etc.

- Guiding Principles
  - Measurable
    - Can the behavior be tracked from the data collected
  - Attributable
    - Can the behavior be assigned to the treatment
  - Sensitive
    - Does the metric have low variability that can be measured reliably

A/B Testing Process

1. Formulate Hypothesis
2. Define Metrics
3. Experiment Design
4. Collect Data
5. Analyze Results
6. Launch Decision

Statistical Analysis

- Estimation and Inference
- Confidence Intervals
- p-values
- Multiple Comparisons

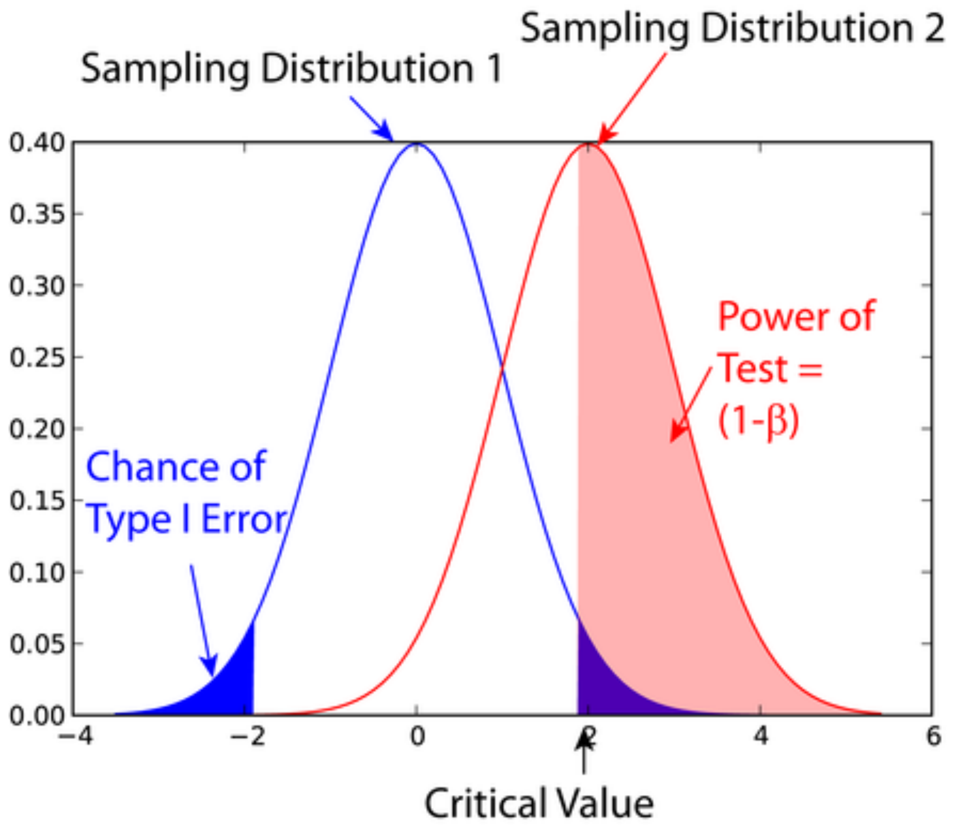
Significance Level

- How likely it is that the difference between control and test version isn't due to error or random chance
- Typically set to 95%

	Reject H0	Fail to Reject H0
Reality: H0 is True	Type I error (probability = α)	Probability = 1-α
Reality: H0 is False	Power (1-β)	Type II error (probability = β)

Power Analysis

- Determines the sample size required to detect an effect of a given size with a given degree of confidence.
- Statistical power $(1 - \beta)$ is the inverse of the probability of making a Type II error (β)
- Function of four factors:
  - Sample size
  - Minimum Effect of Interest (MEI, or Minimum Detectable Effect)
  - Significance level (α)
  - Desired power level (implied Type II error rate)



Lift

- *Lift* is the percent improvement of a target metric
- Easy to understand and explain but does not take randomness into account

$$lift = \frac{m_2 - m_1}{m_1}$$

  - m_1 Average of the first (or control) group
  - m_2 Average of the second (or test) group

Effect Size

- Effect size is the statistical strength of our result by controlling for randomness
- Cohen's d is one way to increase explanatory power through the use of standard deviation

$$Effect\ Size = \frac{m_2 - m_1}{s_{pooled}}$$

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2}{n_1 + n_2 - 2}}$$

- n_1 Size of the first (or control) group
- n_2 Size of the second (or test) group
- s_1 Sample standard deviation of the first (or control) group
- s_2 Sample standard deviation of the second (or test) group

Cohen's d	Effect size
0.01	Very small
0.2	Small
0.5	Medium
0.8	Large
1.2	Very large
2.0	Huge

Minimum Detectable Effect (MDE)

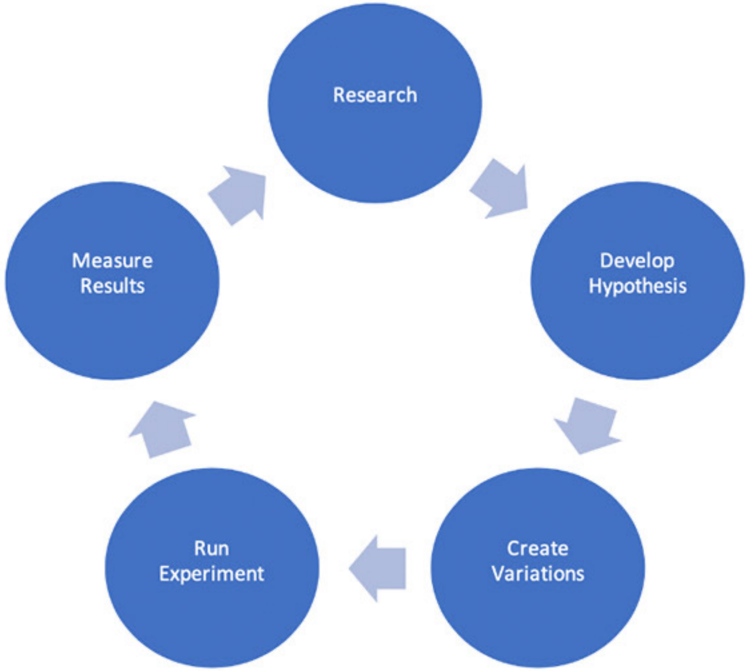
- Minimum effect size that should be detected with a certain probability
- MDE is inversely related to sample size is necessary to calculate the minimum required sample size

$$MDE = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)\sqrt{\frac{2p(1-p)}{n}}$$

- $\mathrm{Z}(k)$ is a critical value to reject hypothesis with probability k
- α is the significance level
- $(1-\beta)$ is the power of the test
- n is the sample size per group
- p is the baseline proportion (or probability of success) in the control group

3. Experiment Design

- Experimental Unit
  - Smallest unit you are measuring the change over
  - E.g., Individual users make a convenient experimental unit
- Target Population
  - E.g., Visitors who have searched for products
- Sample Size
  - Use sample size calculator
- Experiment Duration
  - Long enough to derive meaningful results
  - E.g., 1-2 weeks



Online vs. Offline Testing

Online Testing	Offline Testing
Real-time data collection	Uses historical data
Dynamic environment	Simulated environment
Captures actual user behavior	Controlled variables
Immediate feedback	No impact on real users
Realistic conditions	Cost-effective
Accurate, relevant results	Preliminary insights
Risk of negative impact on users	Less realistic
Resource-intensive	Historical bias
Ethical considerations	Limited scope
Used for website design, app features, pricing strategies	Used for model validation, algorithm comparison, initial hypothesis testing

4. Collect Data

- Set up data pipelines
- Set up instrumentation
- Run Experiment
- Avoid peeking p-values
- Test Validation

A/B Statistical Tests

- Test if there is a statistically significant difference between two groups in terms of a specific metric.
- Depends on the nature of the data, assumptions, and requirements of the A/B test.
- Tests
  - Chi-squared test
  - Student's t-test
  - Welch's t-test
  - Mann-Whitney U (Wilcoxon rank-sum) test
  - Bootstrap test
  - Bayesian methods

A/A Test

- Helps validate the experimental setup
- By comparing two identical groups, it helps identify and address any biases, errors, or inconsistencies in the testing framework
- Expected Outcome:
  - p-value should be greater than the significance level, indicating no significant difference between the two groups.
- Unexpected Outcome:
  - p-value less than the significance level would indicate a significant difference between the identical groups, suggesting potential issues with the randomization process, data collection, or other aspects of the experimental setup.

Frequentist vs Bayesian Approach

Frequentist

Control
Treatment
Hypothesis

Experiment

Calculate Test Statistic and P-value

Accept/Reject Null Hypothesis

Bayesian

Control
Treatment
Define Priors

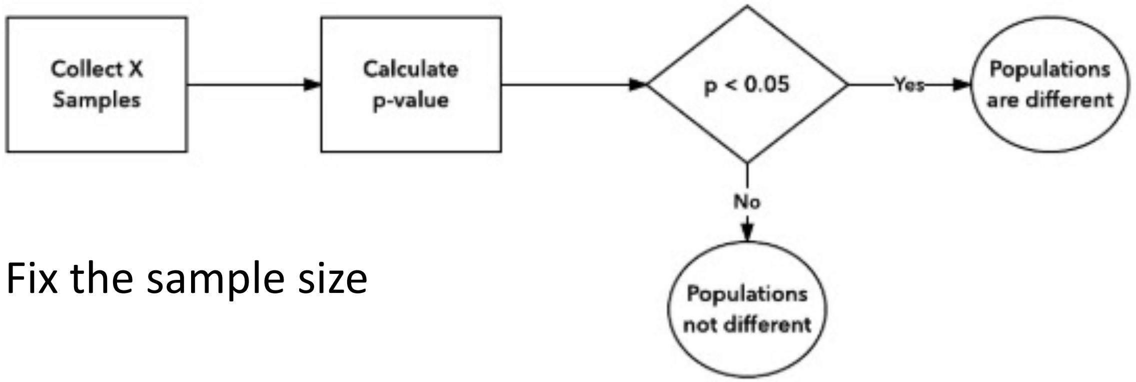
Experiment

Calculate posterior distributions for control and treatment

X% Confident that the lift is Y%

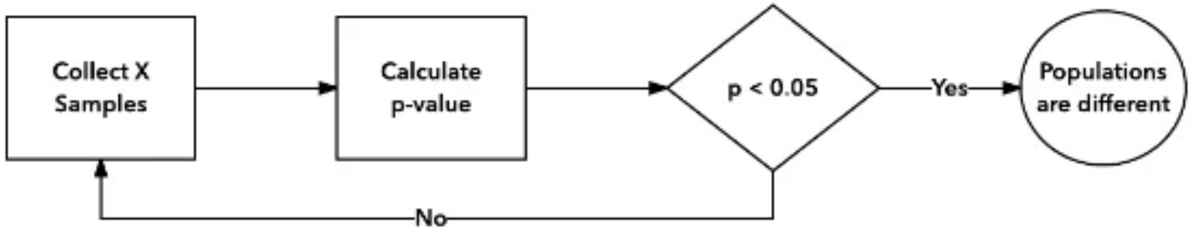
A/B Testing Process

Correct A/B testing



Fix the sample size

Incorrect A/B testing



https://gopractice.io/data/peeking-problem/

Test Validation

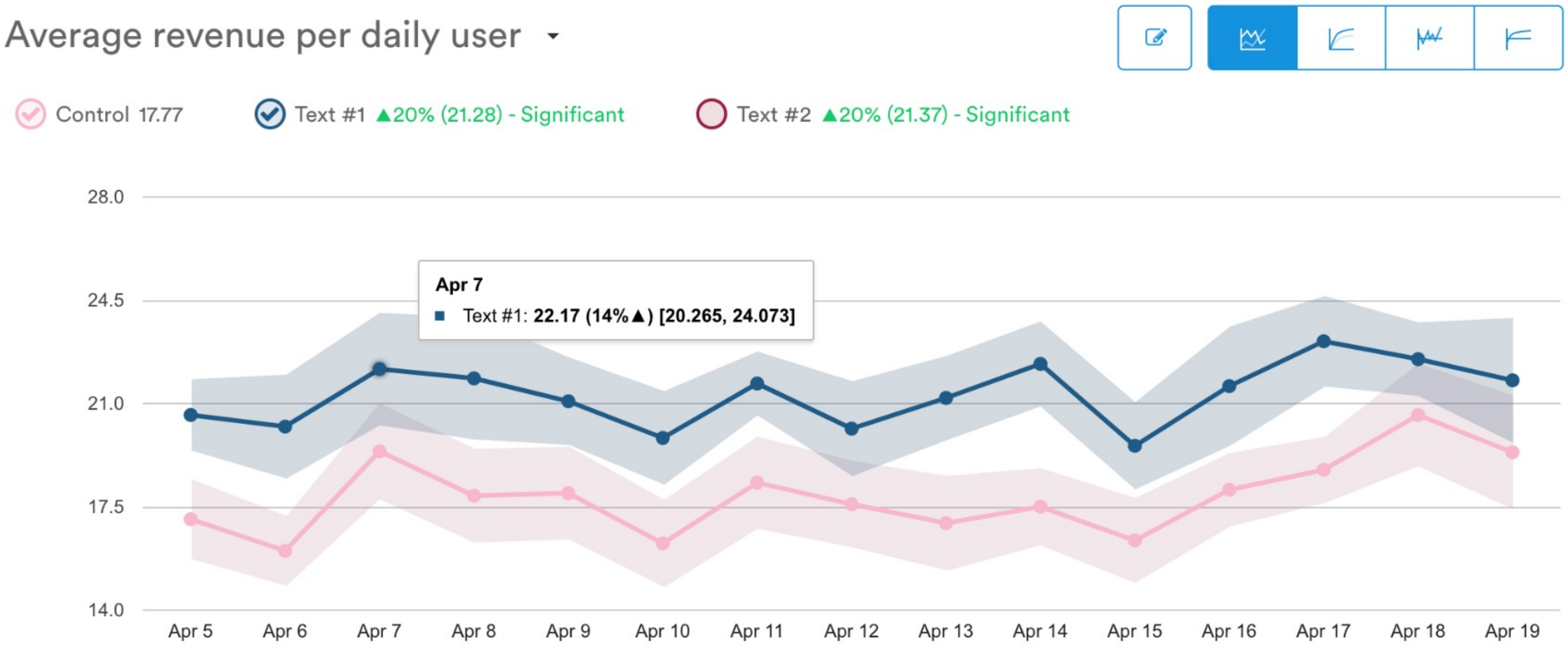
- Instrumentation Effects
  - Testing tool
  - Bugs
- External Factors
  - Seasonality
  - Holidays
  - Competition
  - Adverse Events
- Selection Bias
  - A/A Testing
- Sample Ratio Mismatch
  - Chi Squared Goodness of Fit
- Novelty Effect
  - User segmentation old vs new

Ethical Considerations

- Informed Consent
- Data Privacy
- Fairness and Bias
- Monitoring and Stopping Rules

5. Analyze Results

- Sample Dashboard



https://docs.leanplum.com/docs/ab-test-analytics

6. Launch Decision

- Metric Tradeoffs
  - Primary vs Secondary metrics
- Cost of launching
  - Implement Winning Variation

Summary

- Requires a very good understanding of the business problem
- A/B testing is a way to test your own assumptions
- A/B tests heavily depend on sample size which should be decided in advance
- A/B tests are difficult to design and execute
- Could take weeks to show results
- Statistical significance does not indicate that variation is better than control or the magnitude of the result

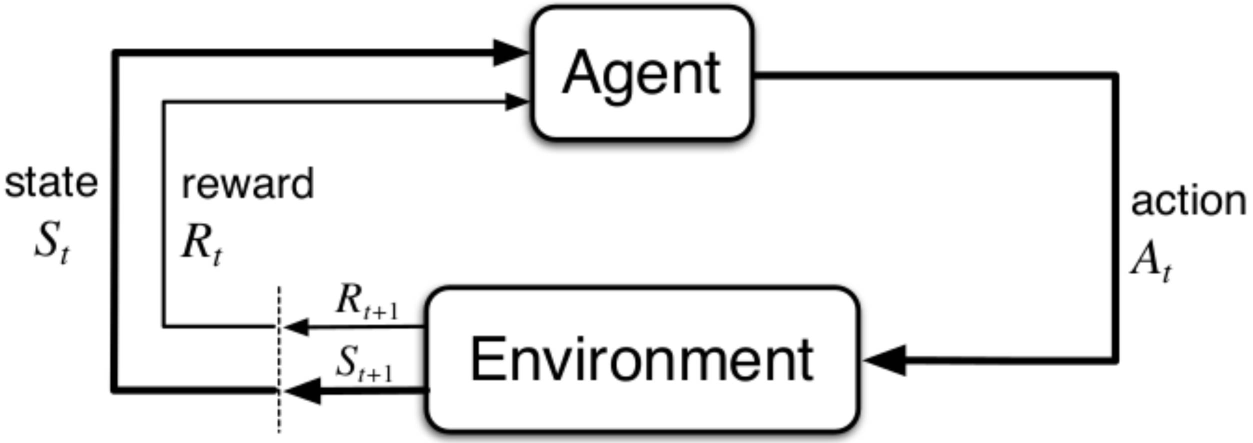
MULTI ARMED BANDITS

Other Situations

- Dynamic Environments
  - E.g, Personalized recommendations where user preferences evolve over time
- Sequential Decision Making
  - E.g, A multi-step user journey, such as onboarding processes,
- Complex Reward Structures
  - E.g, Retention strategies where the goal is to maximize long-term user engagement
- Exploration vs. Exploitation
  - E.g, Strategies where exploring new ad placements might uncover higher-performing options
- Contextual and Personalized Policies
  - E.g, Personalized marketing campaigns where the best action varies between user segments.

Reinforcement Learning

- An agent in a current state (S_t) takes an action (A_t) to which the environment reacts and responds, returning a new state (S_{t+1}) and reward (R_{t+1}) to the agent.
- Given the updated state and reward, the agent chooses the next action, and the loop repeats until an environment is solved or terminated.



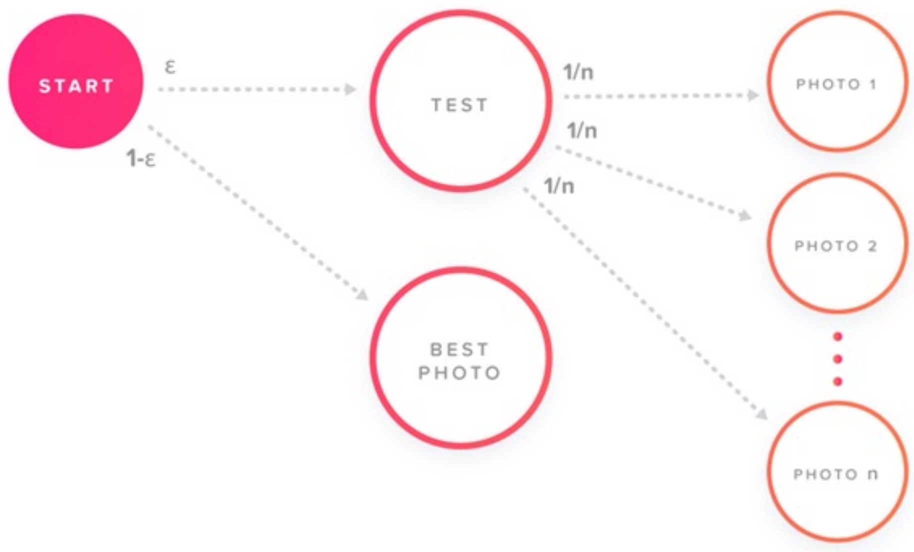
Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto

Multi-Armed Bandits

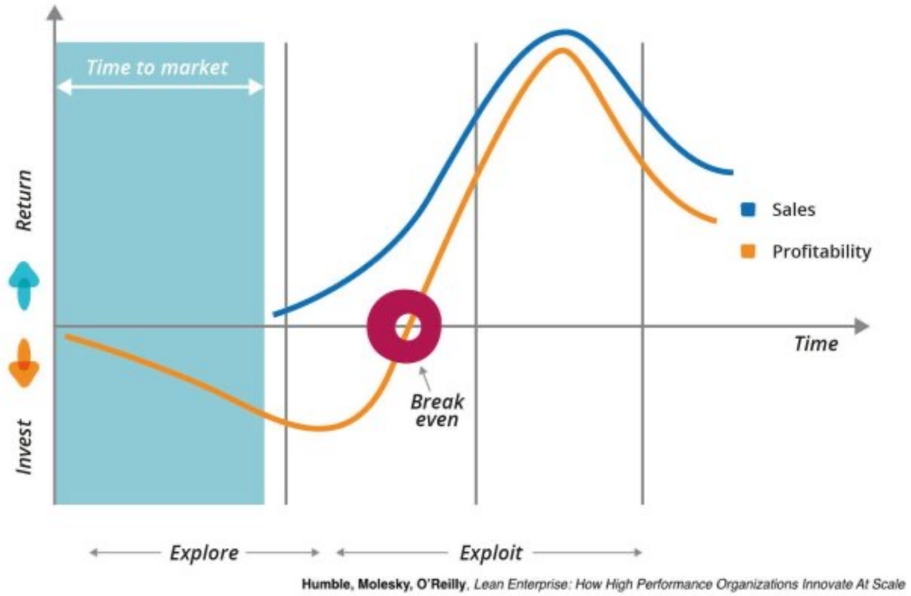
- Allows for adaptive (dynamic) allocation of traffic based on the performance of each arm in real-time.
- Benefits
  - Allows for more effective decision-making and optimization of experiments.
  - Reduces potential loss of performance by quickly identifying and exploiting better-performing variations while continuing to explore other possibilities.

Slot machines are called bandits

Exploration vs Exploitation



Profile Picture - Epsilon-Greedy Algorithm



Business Strategy - Exploration vs Exploitation

Epsilon (ε) : The agent takes random actions for probability ε and greedy action for probability (1-ε)

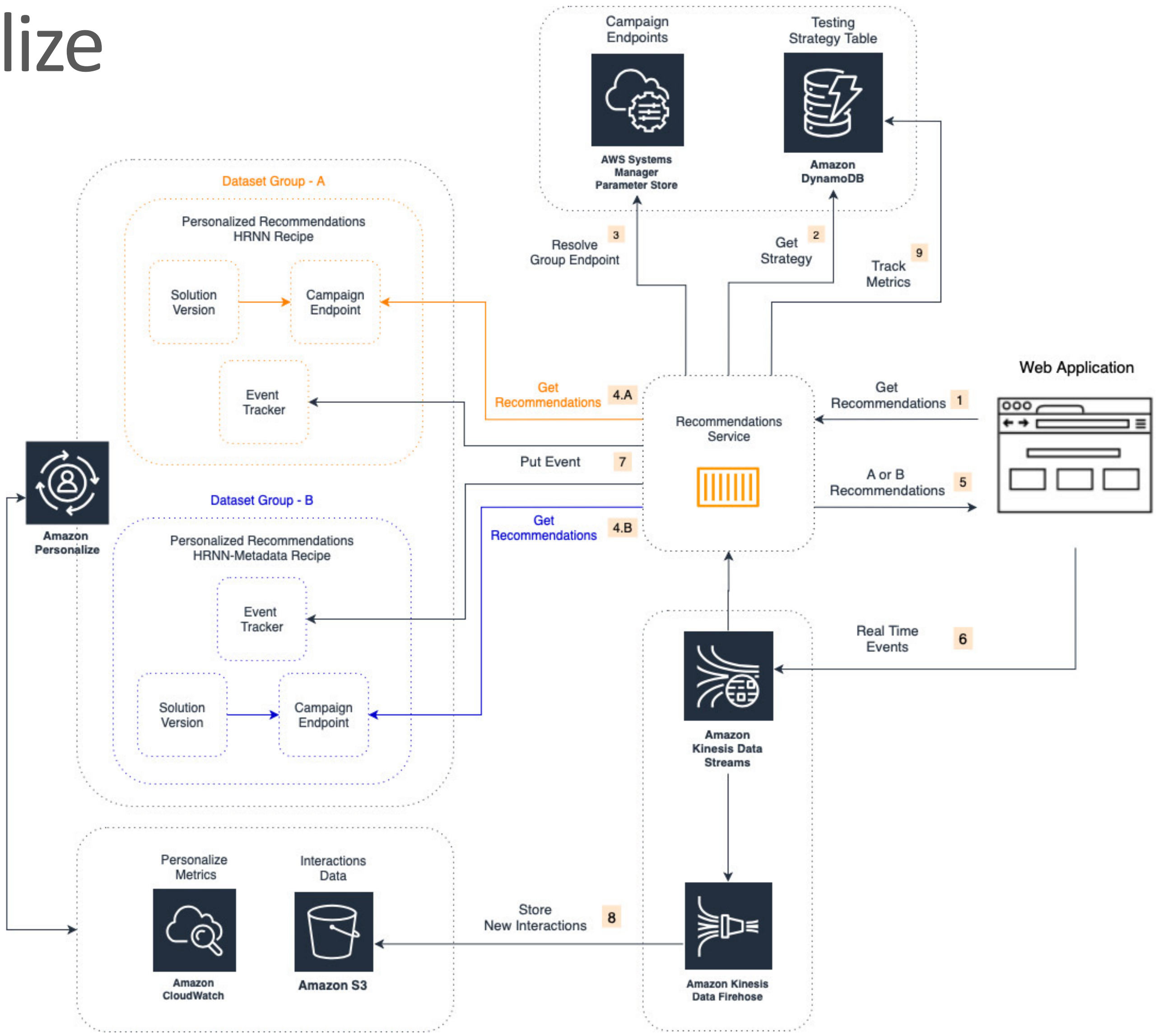
Multi-Armed Bandits: Process

1. Set up variations (arms)
2. Start with exploration
3. Track performance
4. Exploit high-performing arms
5. Continue exploration and exploitation
6. Adapt allocation over time
7. Convergence to optimal arm

Multi-Armed Bandits Strategies

- Epsilon Greedy
- Upper Confidence Bound (UCB1)
- Bayesian Bandits
  - UCB Tuned Bandit
  - Thompson Sampling
- Softmax Exploration
- Gradient Bandits

AWS Personalize



Microservices-based implementation of an A/B test between two Amazon Personalize campaigns