Jere Pankka

A/B TESTING DESIGN AND EXAMPLES

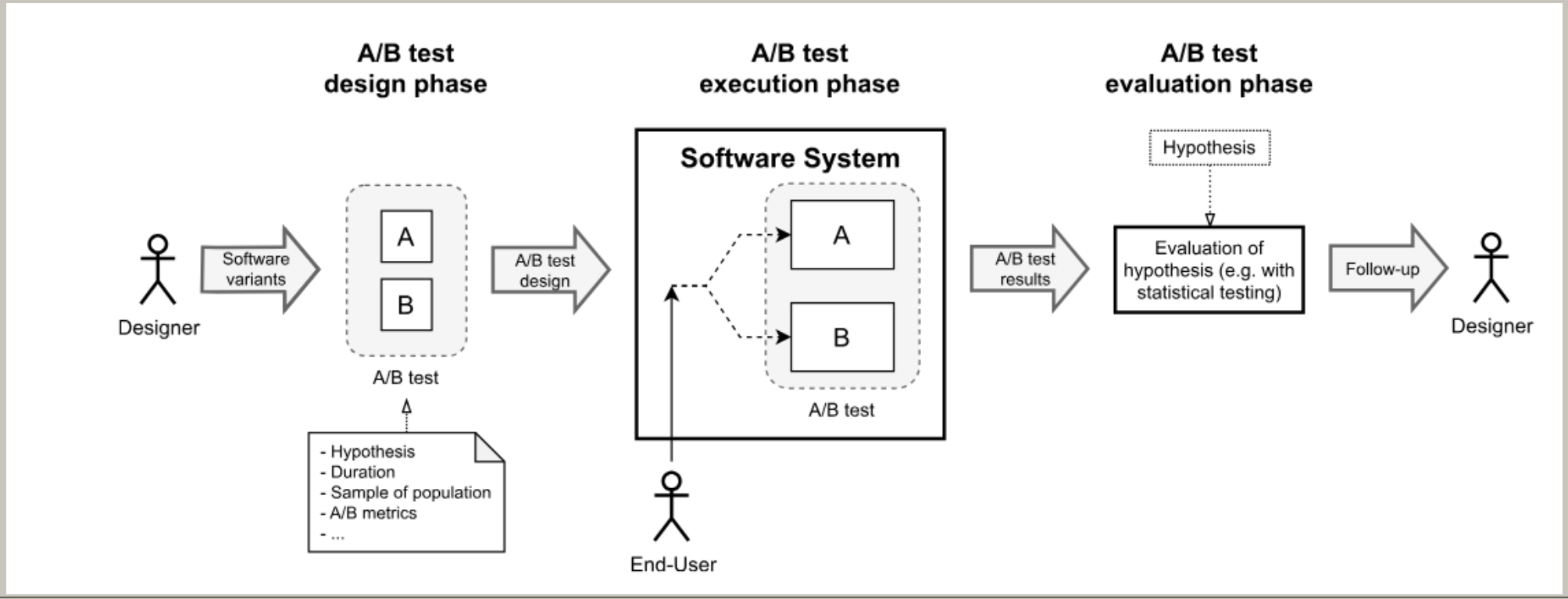
AGENDA

- What is A/B testing
- Overview of the process and a more detailed steps
- A good example from gaming industry
- Bad example from ecommerce retailer
- Common mistakes

WHAT IS A/B TESTING?

A/B TESTING AND ITS PROCESS

- A/B testing is a way to test hypothesis where two different variants of a software are evaluated (Quin et al., 2024)
- Ranges from small UI changes to fully new features (Quin et al., 2024)
- The variants are analyzed through data; this could be click rates, retention, lifetime value, for instance (Johari et al., 2017)
- Multiple use cases: ecommerce, entertainment, social media, gaming



(Quin et al., 2024)

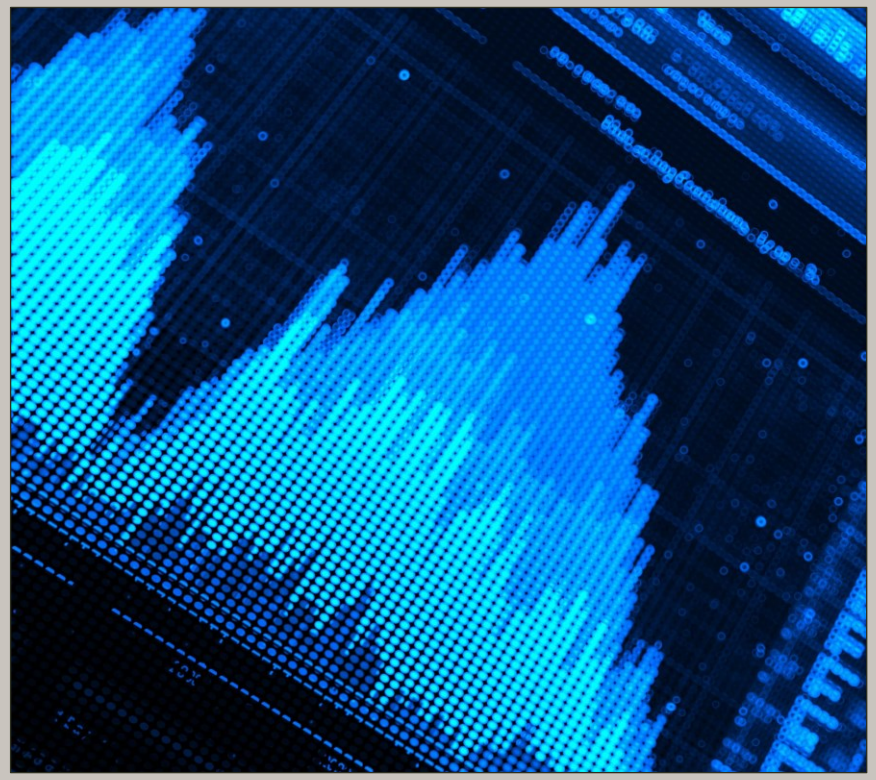
DESIGN PHASE

- The process start with defining a hypothesis
  - Should be based on some kind of user research
  - For example, in gaming the users do not finish tutorial, hence there is a low D1 retention
  - Hypothesis could be that the change will increase the tutorial conversion rate
- After forming a clear hypothesis, define what changes to make to achieve the hypothesis
  - In the example above, the changes could be clearer instructions on what to do, e.g. more audio and visual cues
  - Once again should be based on some kind of user research

DESIGN PHASE CONT.

- Defining outcome metrics: both primary metrics and secondary metrics
  - In the same example, the primary metric could be the completion rate of the tutorial
  - Secondary metrics D1 and D7 retention as well as time to complete
- Determining the duration and required sample size for the A/B testing
  - Tutorial completion rate is 55 % and minimum detectable effect 5 %, p-value 0.05
  - Based on sample size calculator, needed sample size is 4100 users
  - Daily active users is 500, so testing runs for 14 days to ensure big enough sample size
- Designing the new version to use

EXECUTION AND EVALUATION PHASE



- Execution phase
  - Running the experiment
  - Collecting data
- Evaluation phase
  - Analyze the data
  - Determine was the hypothesis met
  - Decide next steps based on the results

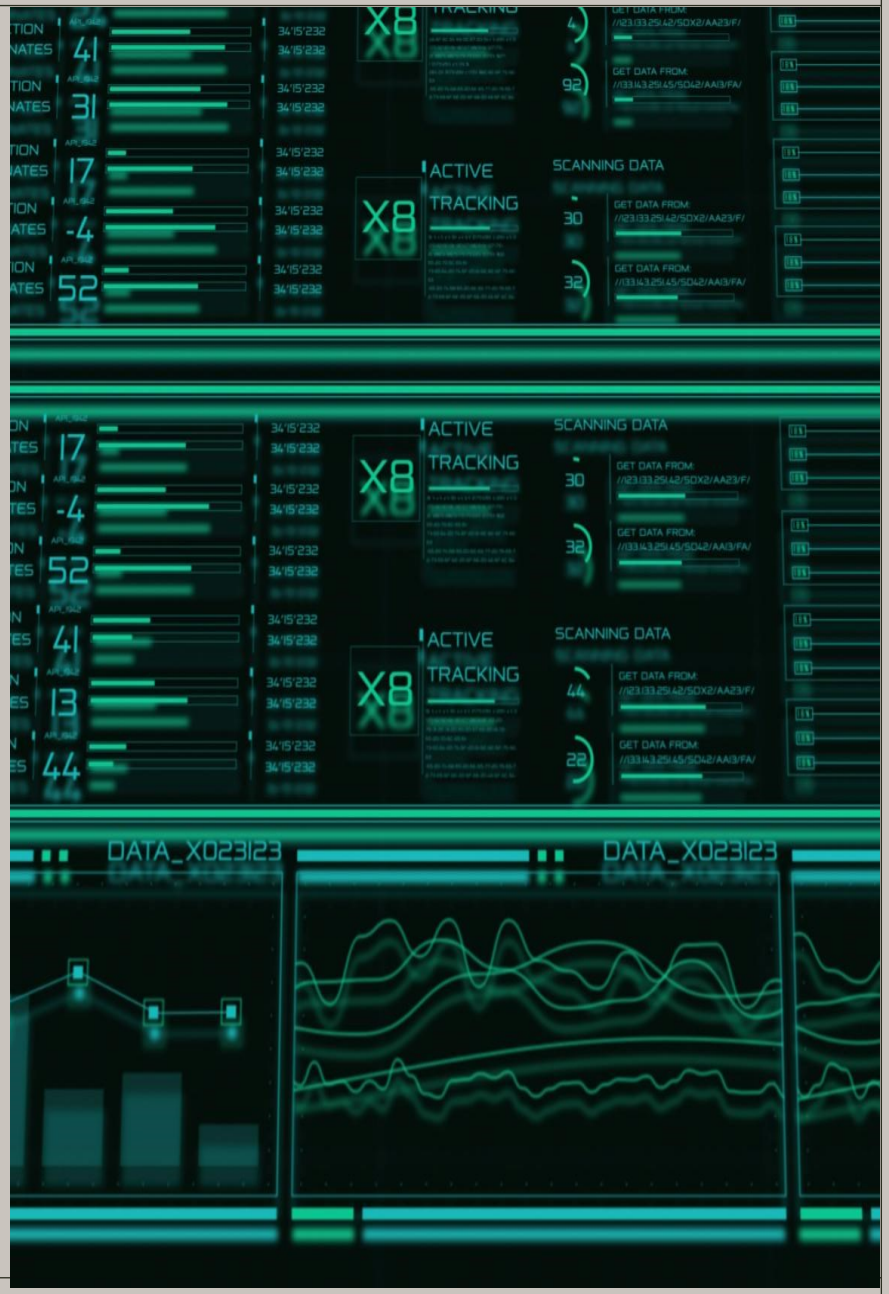
GAMING INDUSTRY EXAMPLE OF A/B TESTING

EXAMPLE REWARD TRACK

- Why A/B testing is needed?
  - Players are not progression reward track a lot
  - Completion rate is only 40 %
  - User research tells that the rewards are too scarce to be motivating
  - Special event does not increase D1 and D7 retention

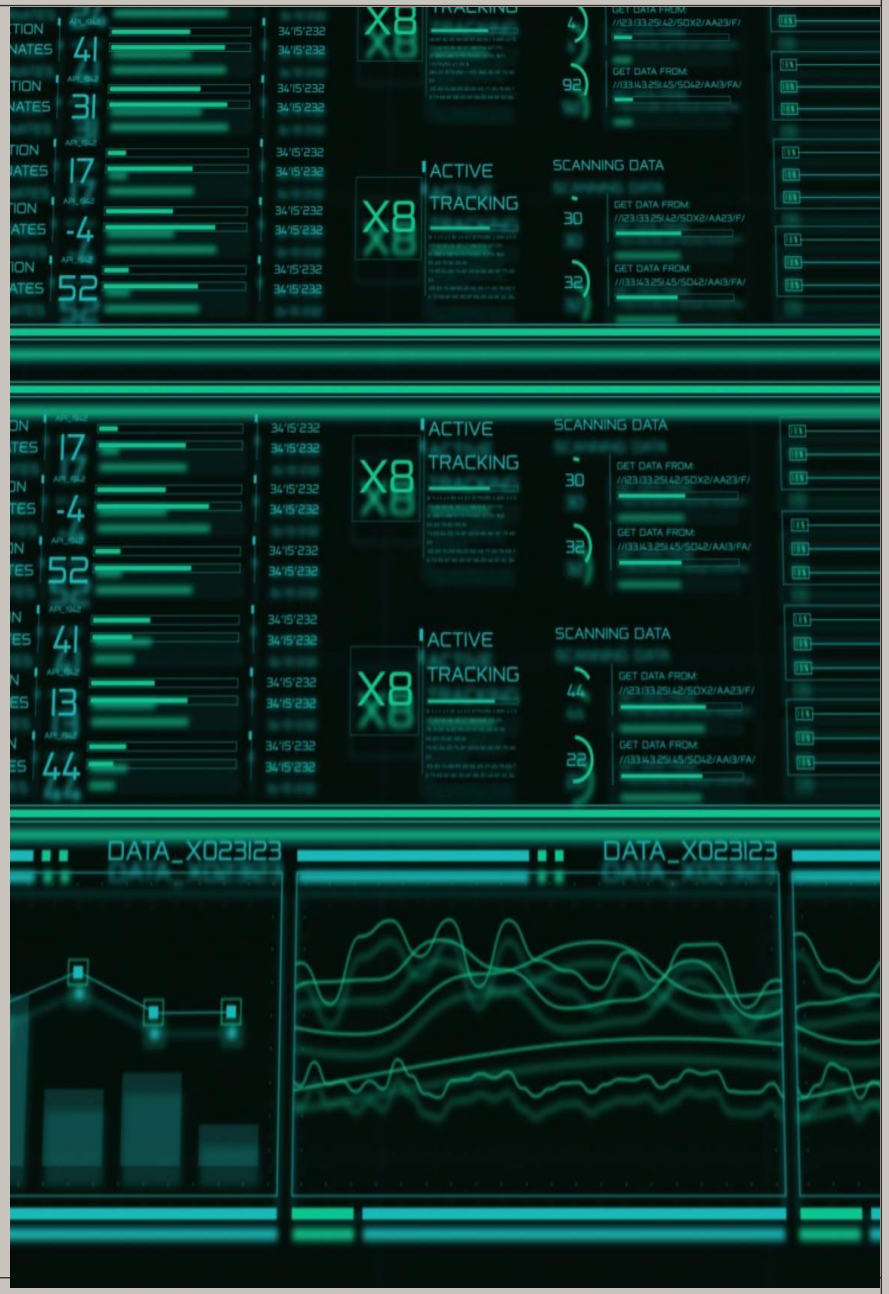
DESIGN

- Hypothesis
  - Adding more rewards will increase completion rate as players get more rewards and thus are more motivated to progress
- Control A is the original reward track, and variant B is the new reward track where players get rewards more frequently
- Primary metric is track completion rate
- Secondary metrics are track progression rate, session frequency, track revenue as well as D7 retention



DESIGN CONT.

- Duration and sample size
  - Minimum detectable effect is 5 %
  - P-value 0.05
  - Sample size 8200, DAU is 500 so experiment is run for 14 days
- Other design choices
  - Testing is focused on all players
  - Not console, OS or region dependent
  - Control A is a previous reward track, and variant B is new reward track, so it is the same for all players



MOCKUP OF EXECUTION AND EVALUATION PHASE

RECOMMENDATION

- Ship variant B
  - Clear positive effects without negative side effects
- Hypothesis is validated
  - More frequent rewards motivate the player, even though the additional total value is negligible

ECOMMERCE RETAILER EXAMPLE OF A/B TESTING

EXAMPLE PRICE ANCHORING

- Why A/B testing is needed?
  - Add-to-cart rate of products that are in sale has lowered significantly
  - Extensive research shows that users do not note the original price, leading to reduced anchoring bias
  - This is more of an example what could go wrong with A/B testing if not careful with the design

DESIGN

- Hypothesis
  - Making the original price more visually prominent will increase add-to-cart rate because the anchoring is stronger
- Control A is the original web page
- Variant B has more prominent font size and color, strikethrough of original price, and discount in percentage

DESIGN CONT.

- Primary metric is add-to-cart-rate
- Secondary metrics are conversion rate and average sale amount per purchase
- Duration and sample size
  - Minimum detectable effect is 5 %
  - P-value 0.05
  - Sample size 20000, DAU is 1000 so experiment is run for 21 days
- 50/50 split for control A and variant B

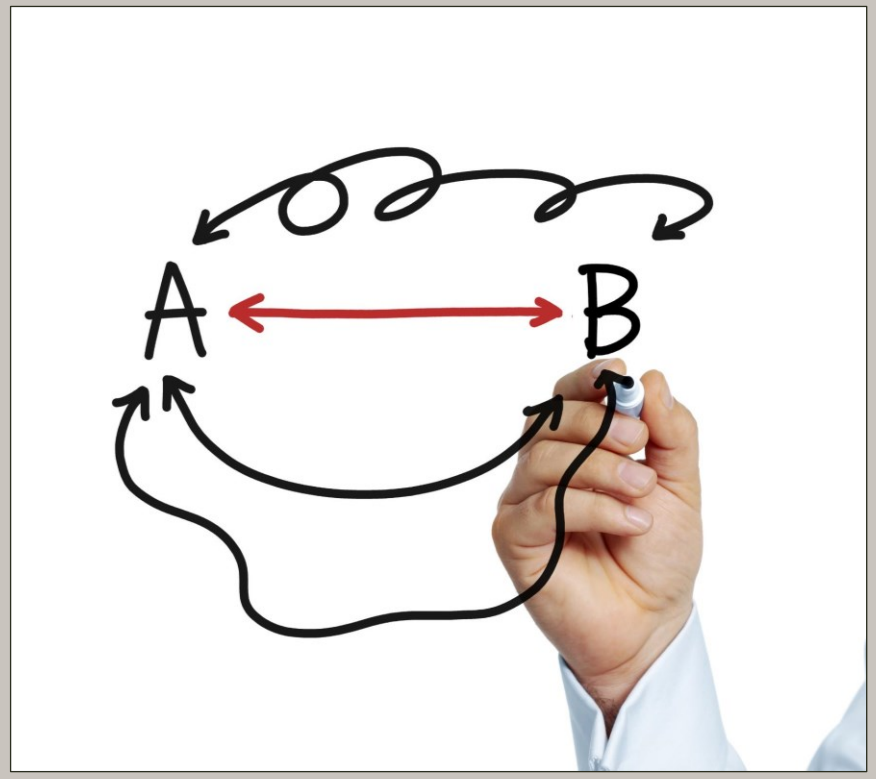
MOCKUP OF EXECUTION AND EVALUATION PHASE

RECOMMENDATION

- Do not ship variant B
- Even though the primary metric was positive and hypothesis validated, the test was designed poorly, leading to possible incorrect decisions with negative effects
- The mistakes made in this example are discussed later
- The test should be redesigned

COMMON MISTAKES IN A/B TESTING

COMMON MISTAKES IN A/B TESTING



- Sample ratio mismatch
  - Uneven exposure to control A and variant B
- Test is designed poorly or results misinterpreted
- Unexpected results caused by hidden biases, incorrect designs or unaccounted factors
- Incorrect statistical interpretation
- Running multiple experiments at the same time
- Using too many metrics or the metrics do not measure the intended outcome (Quin et al., 2024)

HOW THE MISTAKES COULD BE (OR ARE) SEEN IN THE TWO EXAMPLES

- Sample ratio mismatch
  - Due to a bug in the pricing example mobile users are more likely to see variant B: platform differences could affect the result and not the new design
- Test is designed poorly or results misinterpreted
  - Variant B in the reward track has both more frequent rewards and the rewards are more valuable: can not isolate whether improvements stem from the frequency or value
- Unexpected results caused by hidden biases, incorrect designs or unaccounted factors
  - Major competitor had a sale during the same period, where their prices were lower for the same products
- Incorrect statistical interpretation
  - Deciding to ship something based on the metric improvements, without checking statistical significance

HOW THE MISTAKES COULD BE (OR ARE) SEEN IN THE TWO EXAMPLES

- Running multiple experiments at the same time
  - For example, the new reward track and there is a new daily quest system, which will both affect user engagement
- Using too many metrics or the metrics do not measure the intended outcome
  - For instance, 10 secondary metrics is too much, and focus is lost on the relevant metrics
  - Add-to-cart rate as primary metric has a positive impact but conversion rate stays the same and average order value is down

REFERENCES

- Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017, August). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1517-1525).

- Quin, F., Weyns, D., Galster, M., & Silva, C. C. (2024). A/B testing: A systematic literature review. *Journal of Systems and Software*, *211*, 112011.