Fight the HIPPO

Get A/B experiments right

Ning Zhou

Oslo, 2020-04-16

WOMEN IN DATA SCIENCE

A bit about myself

- PM at Microsoft, working with enterprise search
- Previously researcher and PM in Schibsted, Microsoft (1st time!) and SONY
- My relationship with A/B experiments
  - My first A/B experiment in 2011
  - 9 A/B experiments year to date
  - Formalized the experimentation process for Microsoft Search in SharePoint and Office.com in 2019 with my colleague Natalia An
  - Reviewing ~10 A/B experiments every week

What is HIPPO?

HIghest Paid Person's Opinion

First used by Avinash Kaushik in his book Web Analytics: An Hour a Day

Why HIPPO is **not** the way to go

- Suboptimal decision making
- Discourage different opinions
- Hinder creativity and innovation

What is AB experimentation

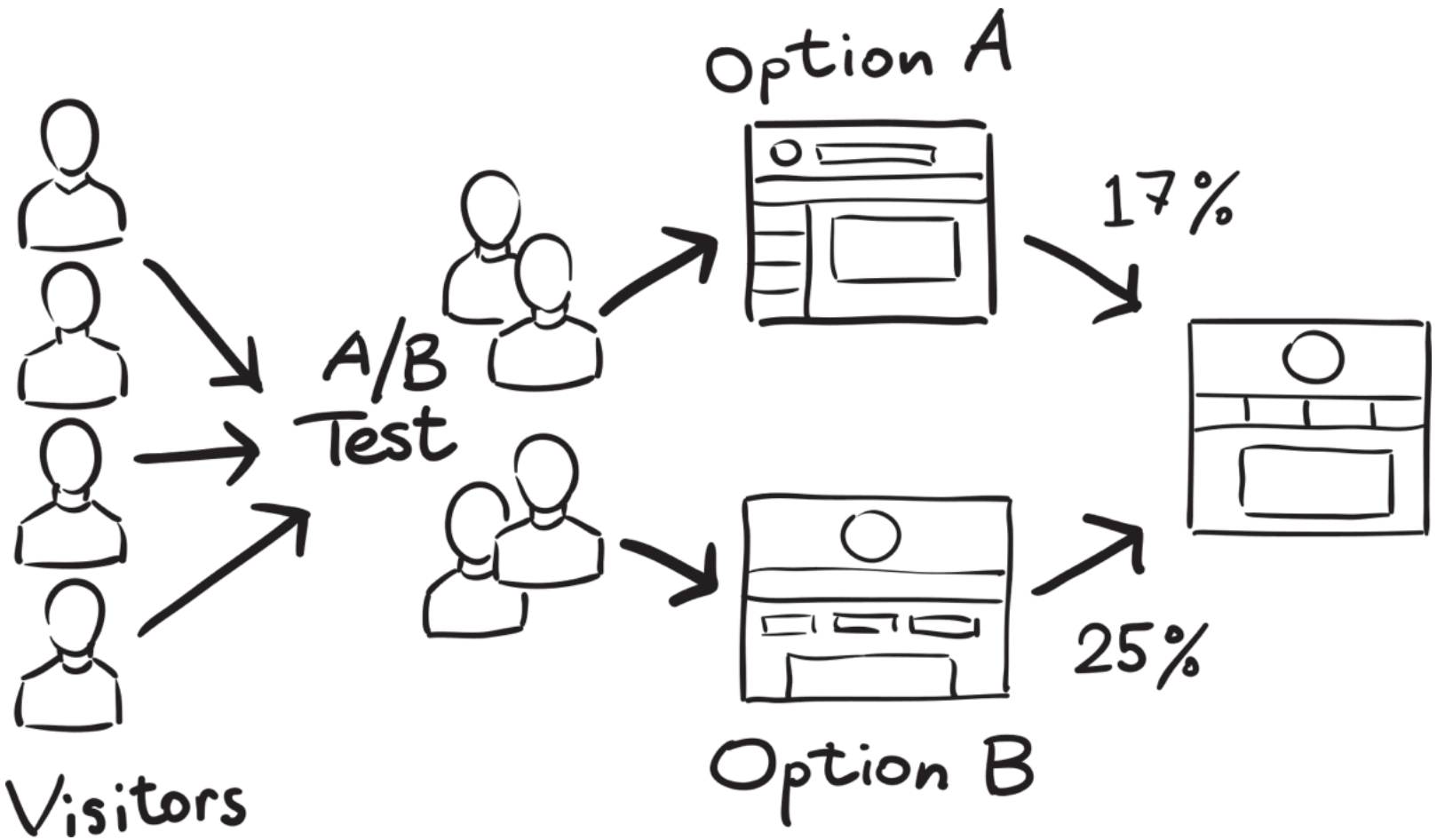
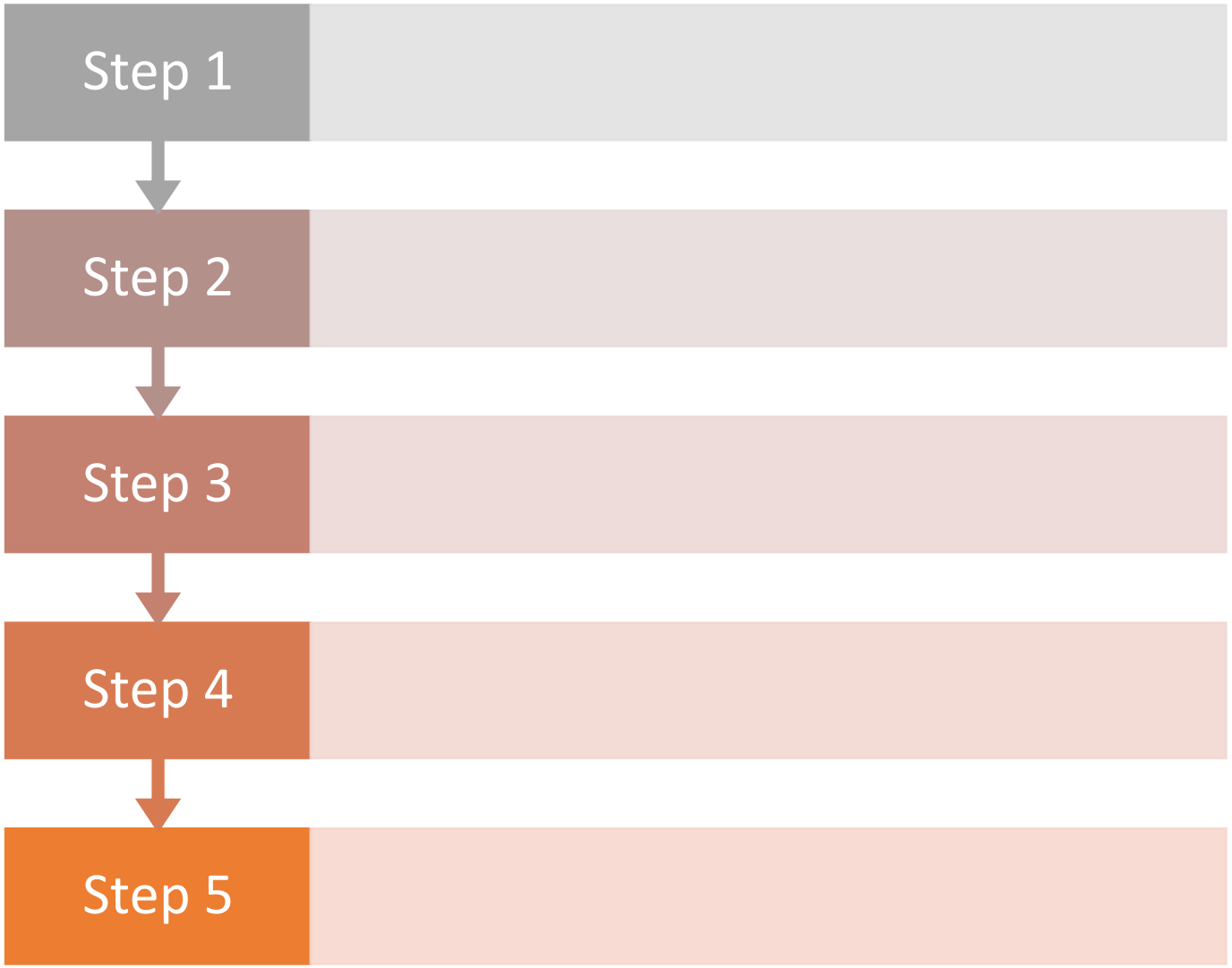


Image from https://www.ab-tests.net/

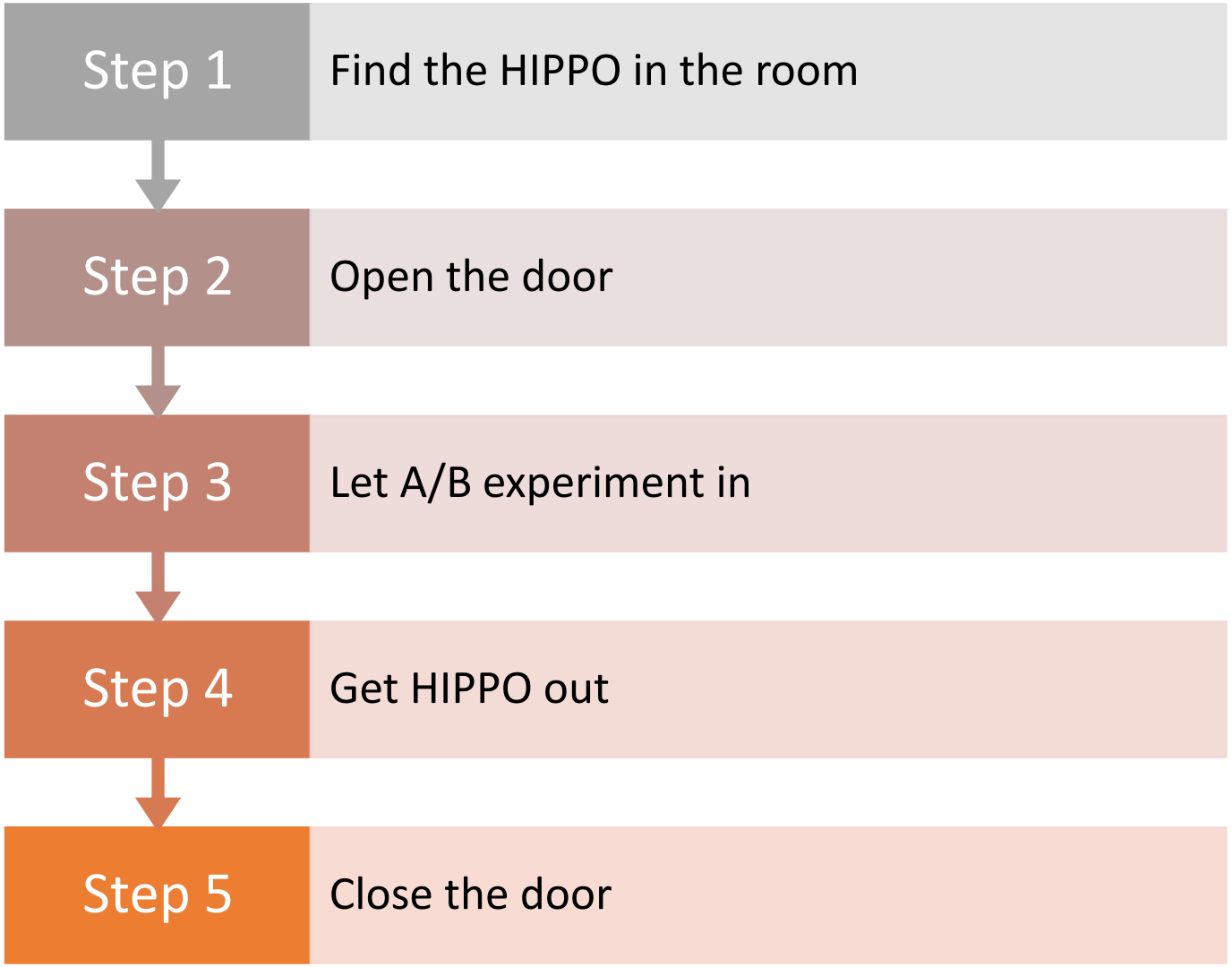
How can A/B experiments help you?

- Give customers their voice
- Use data instead of your own opinion to fight the HIPPO
- Prevent yourself becoming a HIPPO

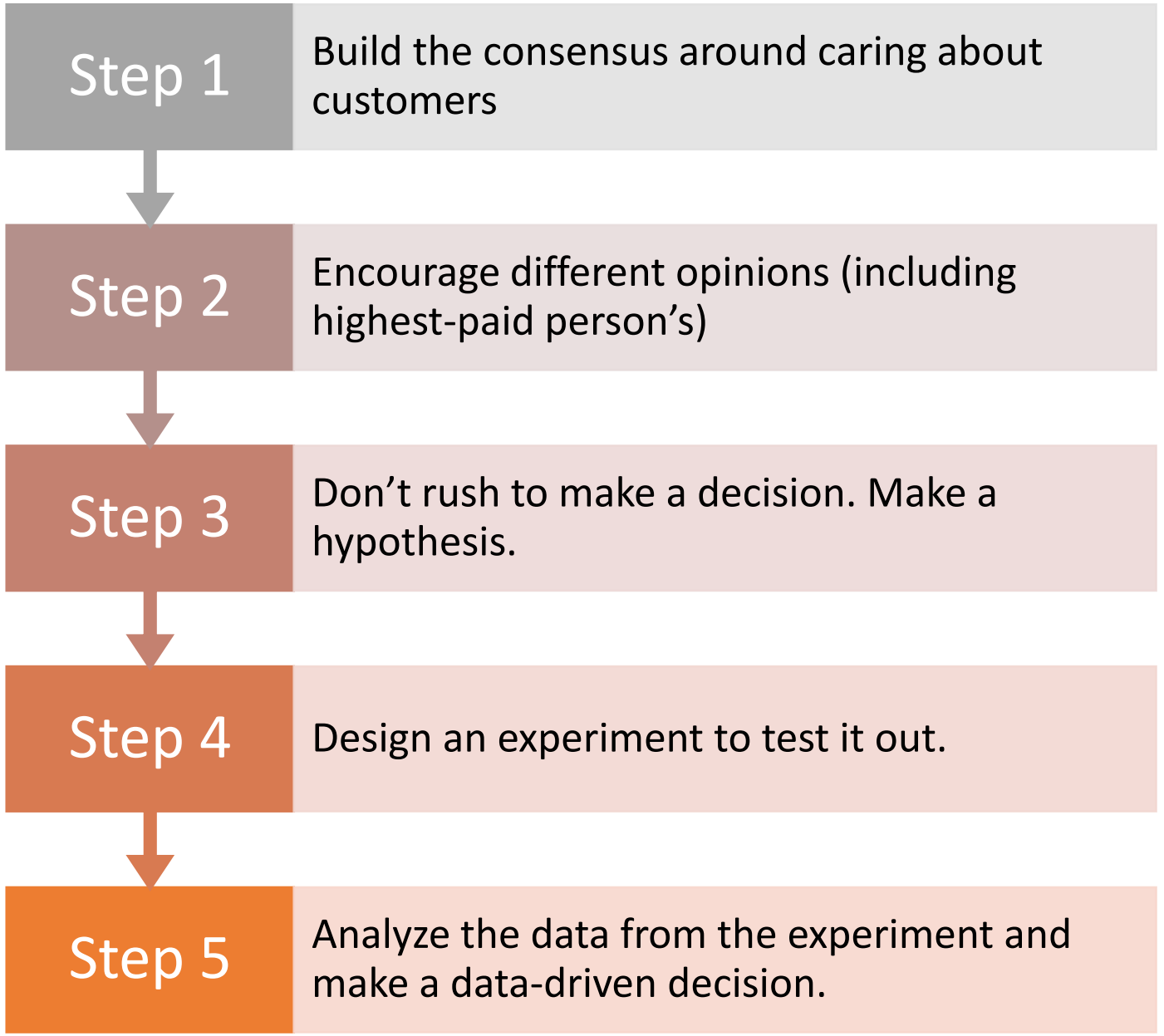
5 steps to get HIPPO out of the door



5 steps to get HIPPO out of the door



5 steps to get HIPPO out of the door



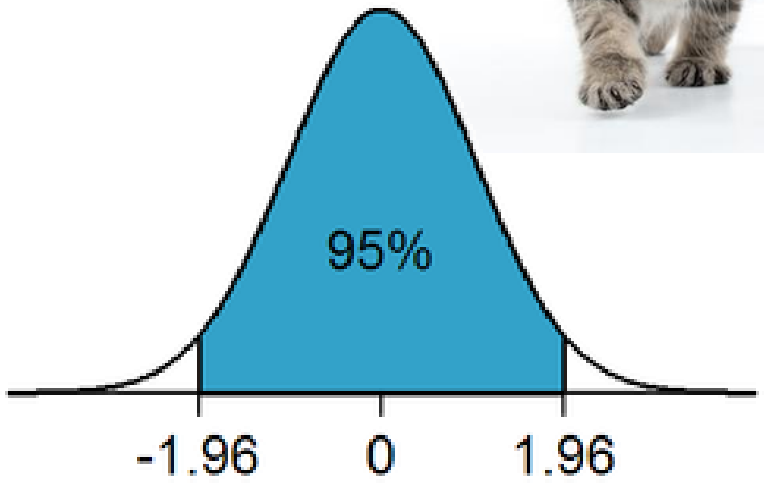
Basic concepts of A/B experiments

According to Wikipedia:

- **Null hypothesis:** “there is no relationship between two measured phenomena or no association among groups”
- **Statistical significance:** “a result has statistical significance when it is very unlikely to have occurred given the null hypothesis”
- ***p*-value:** “the probability of obtaining a result at least as extreme, given that the null hypothesis were true”

Basic concepts of A/B experiments – Example

- **Null hypothesis**: Adding a kitten picture will not change the quality of my presentation
- **A/B experiment**: Half of the audience seeing the kitten and half not, and measure the rating of my presentation from the two groups
- **Data**: With a p-value < 0.05 that the average rating from the group having seen the kitten is higher.
- **Decision**: I'll by default include a kitten picture in all my presentations.



Kitten from https://www.petmd.com/cat/behavior/how-socialize-kitten-new-people

What can be improved with my experiments?

- Very unlikely to obtain statistically significant results of p-value < 0.05 without large audience
- Different metrics to define presentation quality
- Different ways of separating the audience
- Try a different kitten picture
- Try more kitten pictures
- Try HIPPO pictures
- …

Challenge 1: Define success

Is rating a good metric for presentation quality?

- Only a small number of participants fill in the survey
- People with strong opinions are more likely to fill in the survey
- People don’t always say their true opinion in the survey

Choosing the right metrics is critical for a trustworthy experiment.

Challenge 2: Too small audience

- A true example: An experiment was running for two weeks, collected 120 sessions with 24 clicks. Similar features normally have a 10% click through rate. Is it successful?
- Answer: Inconclusive.

Products in early stage should either test strong impact or route to other ways for collecting data for decision making.

Challenge 3: Make trade-off

- Conversion rate vs Cost
- Relevance vs Latency
- Rich content vs Page loading time
- …

There is rarely a *perfect* decision. It is very common that you improve on some metrics while hurting some others.

Challenge 4: Causation vs Correlation

- Metrics improved after your change **!=** Metrics improved because of your change
- The division of audience can be biased, e.g. divide by different channels or regions. You might see differences without doing anything different.

Doing the experiment is only half of the work. Analyzing the results is the other half.

Challenge 5: The temptation of testing every option…

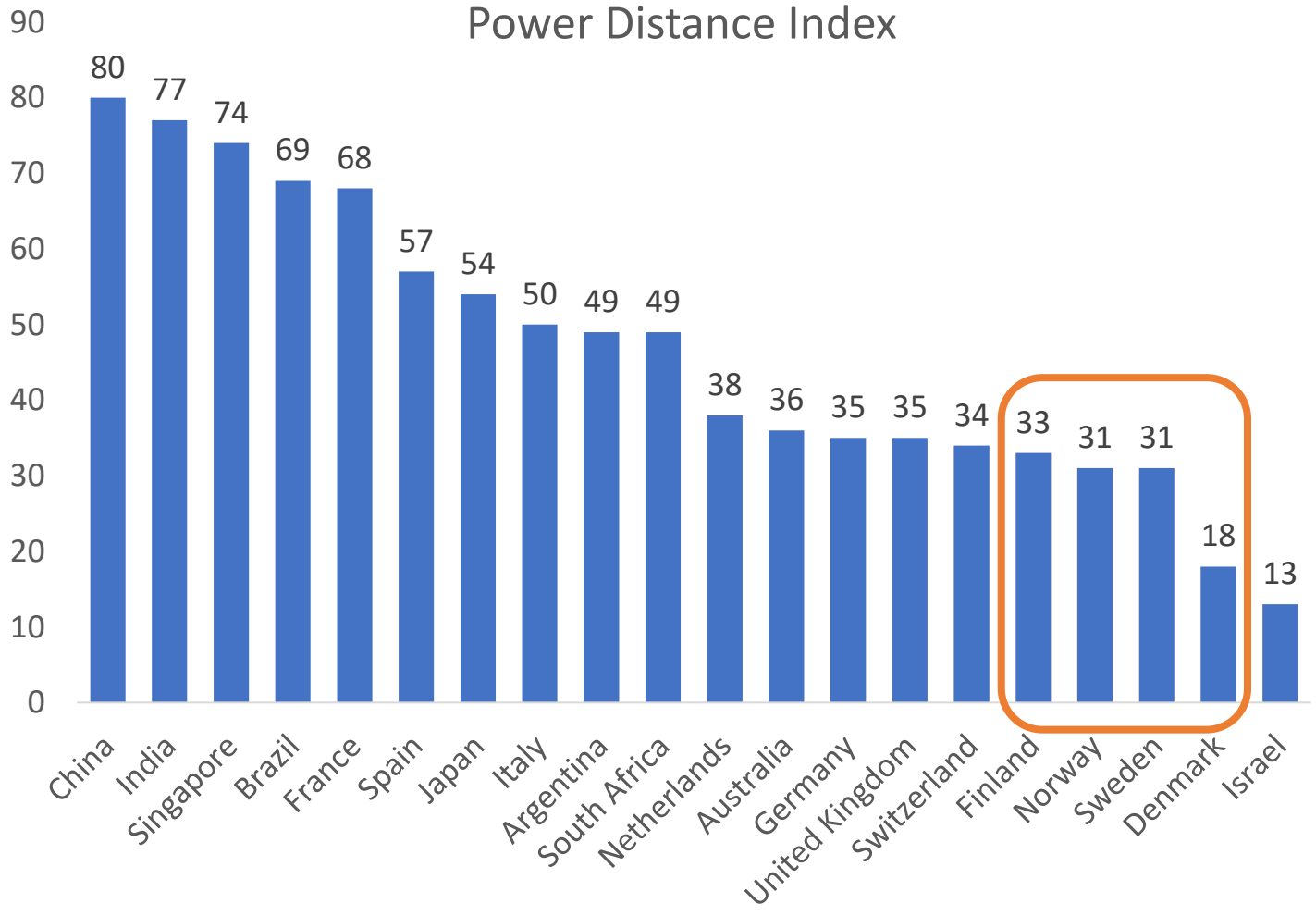
- A/B experiment is not free – bandwidth of audience available for testing, time and engineering cost for conducting and analyzing an experiment, user tolerance on testing, etc.
- A/B experiment cannot give you a perfect answer of all questions.

Don’t hesitate to use A/B experiment, but use it wisely.

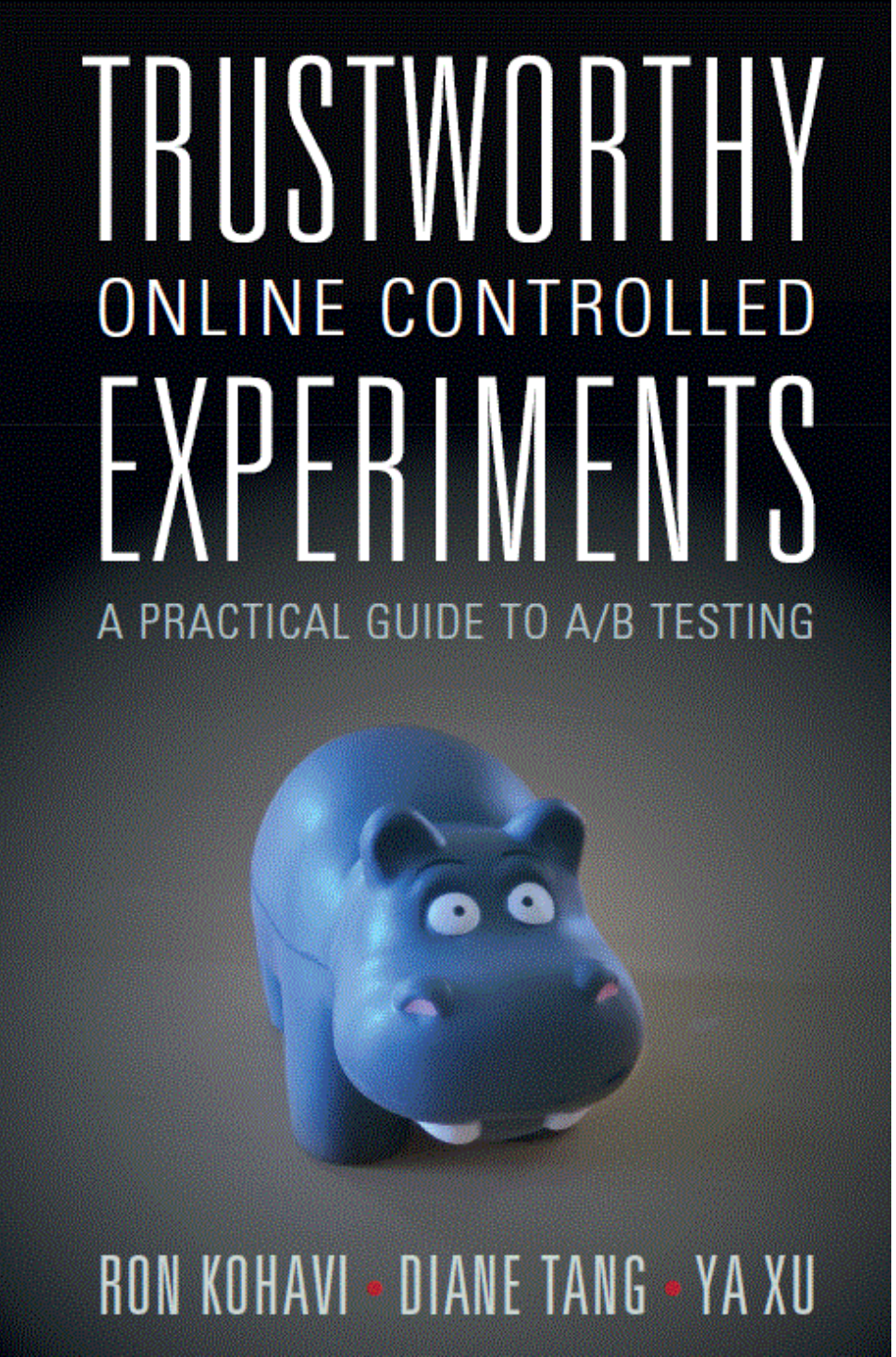
The culture aspect

It may be not that bad in Nordic, but…

- Beware of different cultures of your clients, suppliers and colleagues
- HIPPO can also manifest in other forms than salary or title



Source: https://clearlycultural.com/geert-hofstede-cultural-dimensions/power-distance-index/



Extended readings

- Ron's Harvard Business Review article in 2017: A/B Testing: How to Get it Right
- https://exp-platform.com/
- https://experimentguide.com/

Thank you!
Questions?

Email: ning@duoja.com