Running IR Experiments with Real Users - Common Practices and Challenges

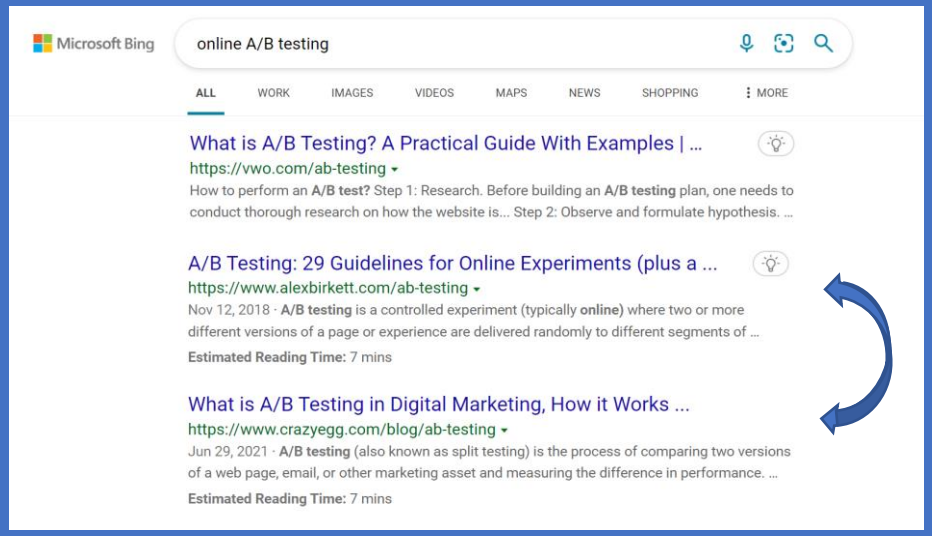
Georg Buscher

A Short Intro

- Thesis in IR about measuring attention with eye tracking
- 2010-2016: building “online” metrics for Bing
- 2016-2020: building “online”, “offline”, survey metrics for FB search
- 2020-2022: experimentation platform in FB/Meta
- Since 2022: measurement for Bing

What Do You Want to Measure?

Search results ranking improvements!



How Do You Measure?

DCG!

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)}$$

Result 1
rel = 5 (highly relevant)

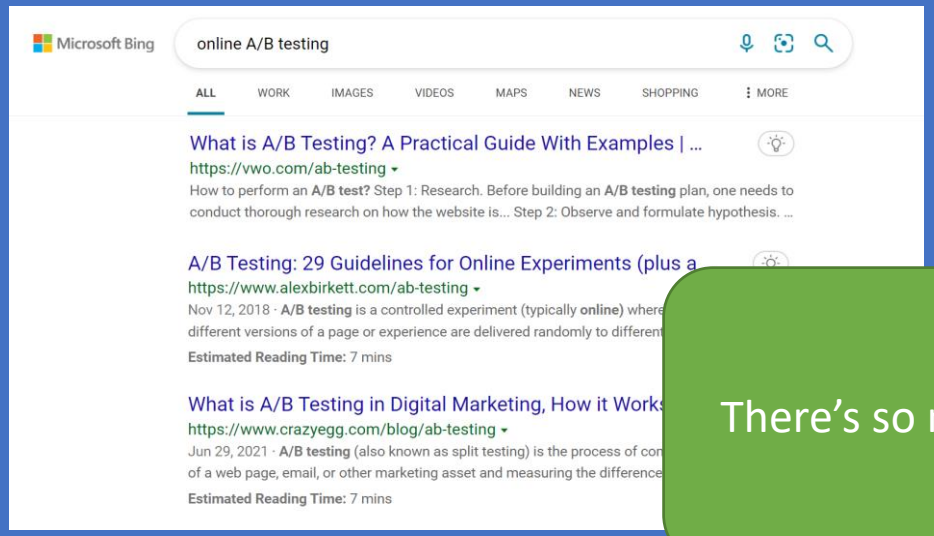
Result 2
rel = 1 (junk)

Result 3
rel = 4 (relevant)

Result 4
rel = 4 (relevant)

What Do You Want to Measure?

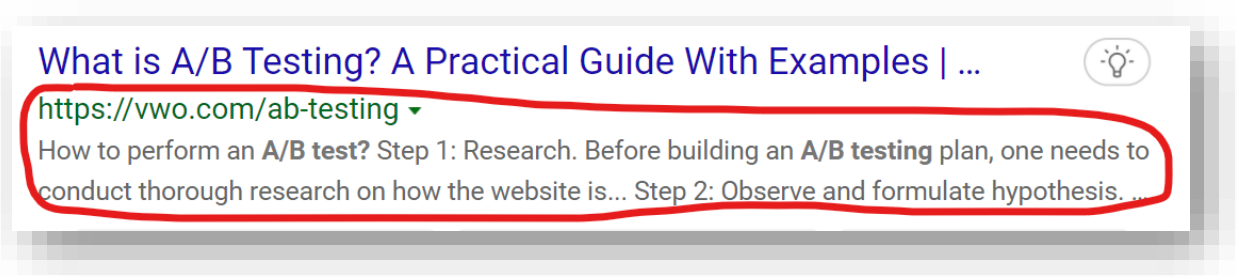
Search results ranking improvements!



There’s so much more!

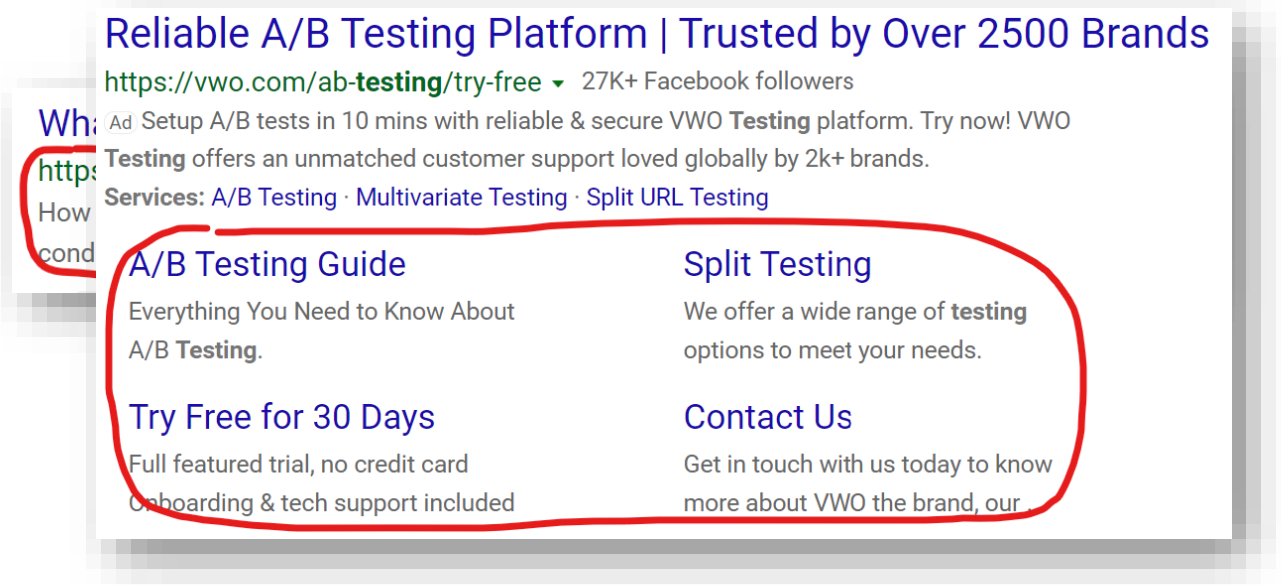
What Do You Want to Measure?

- Snippets



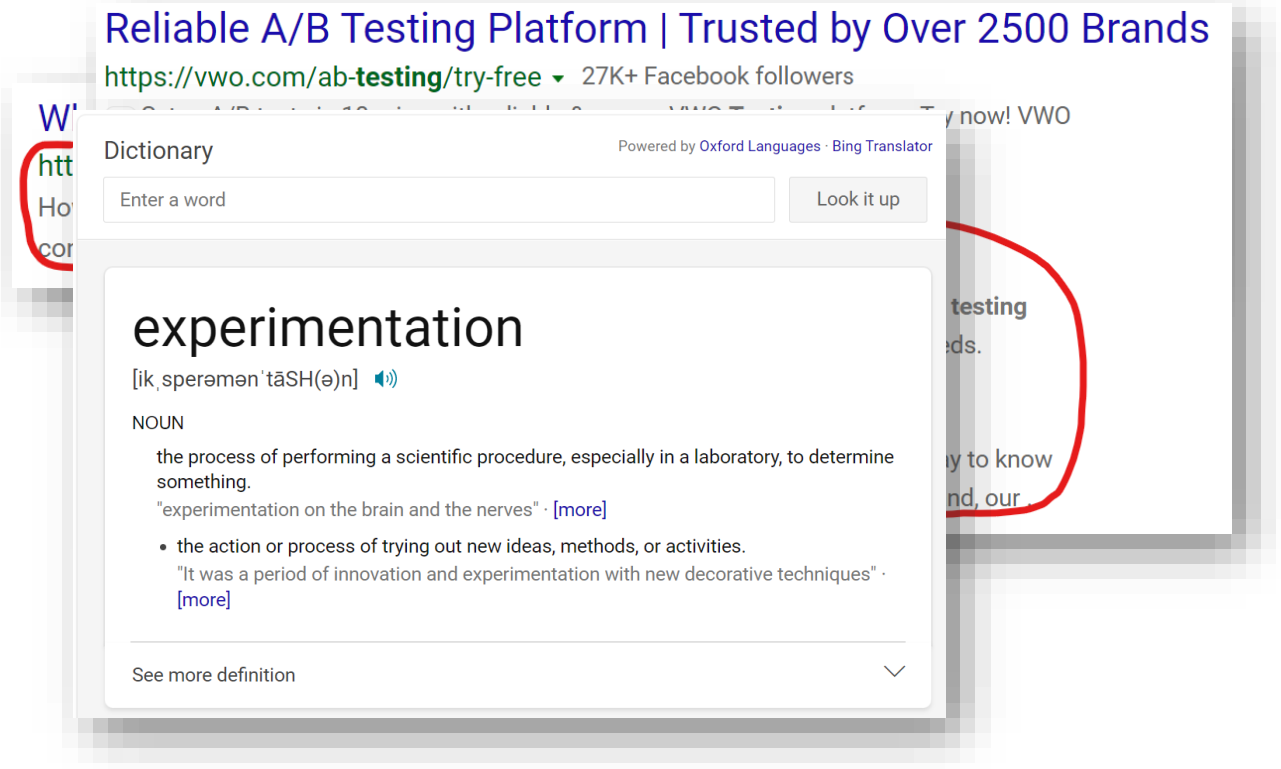
What Do You Want to Measure?

- Snippets
- Deep links



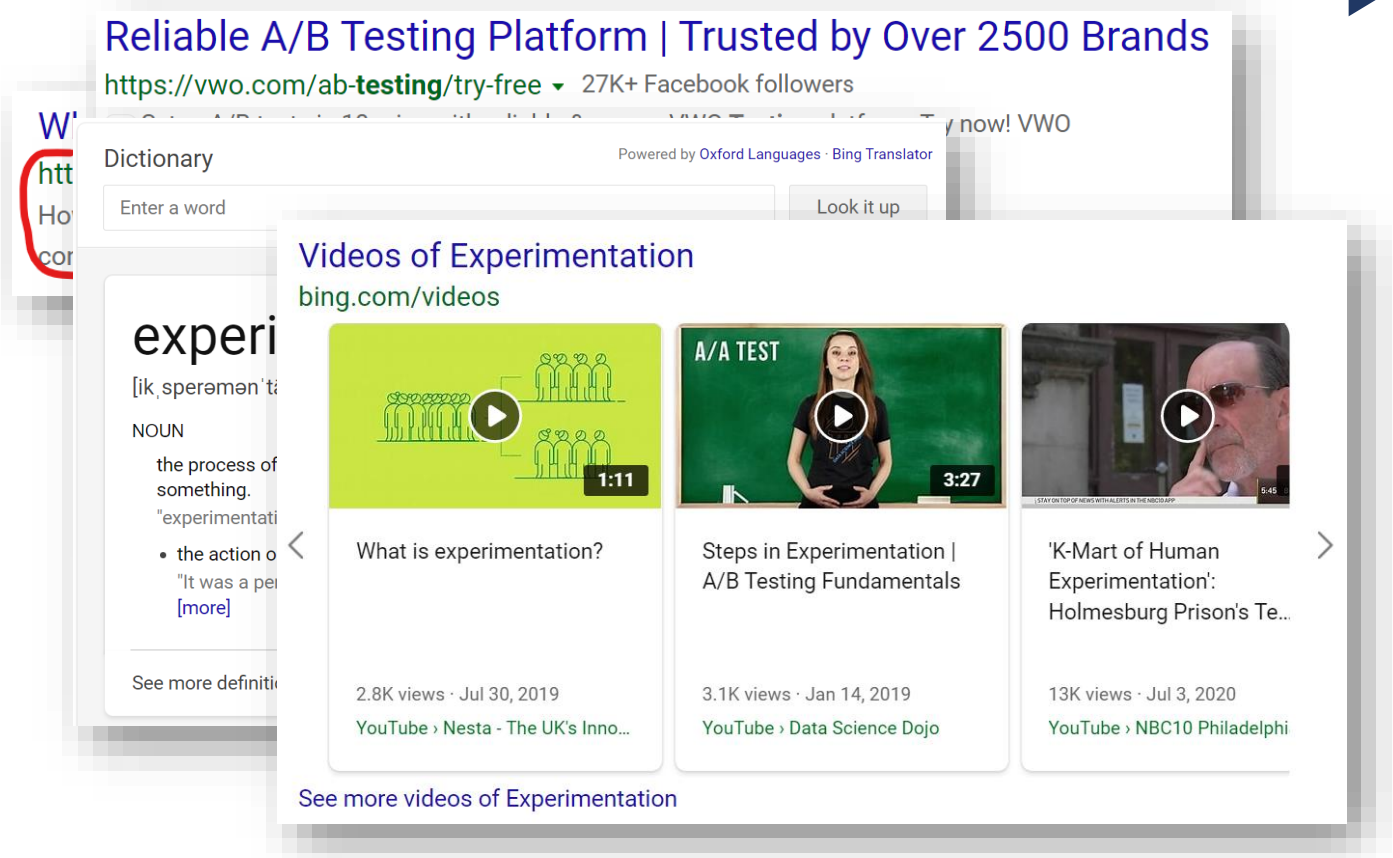
What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers



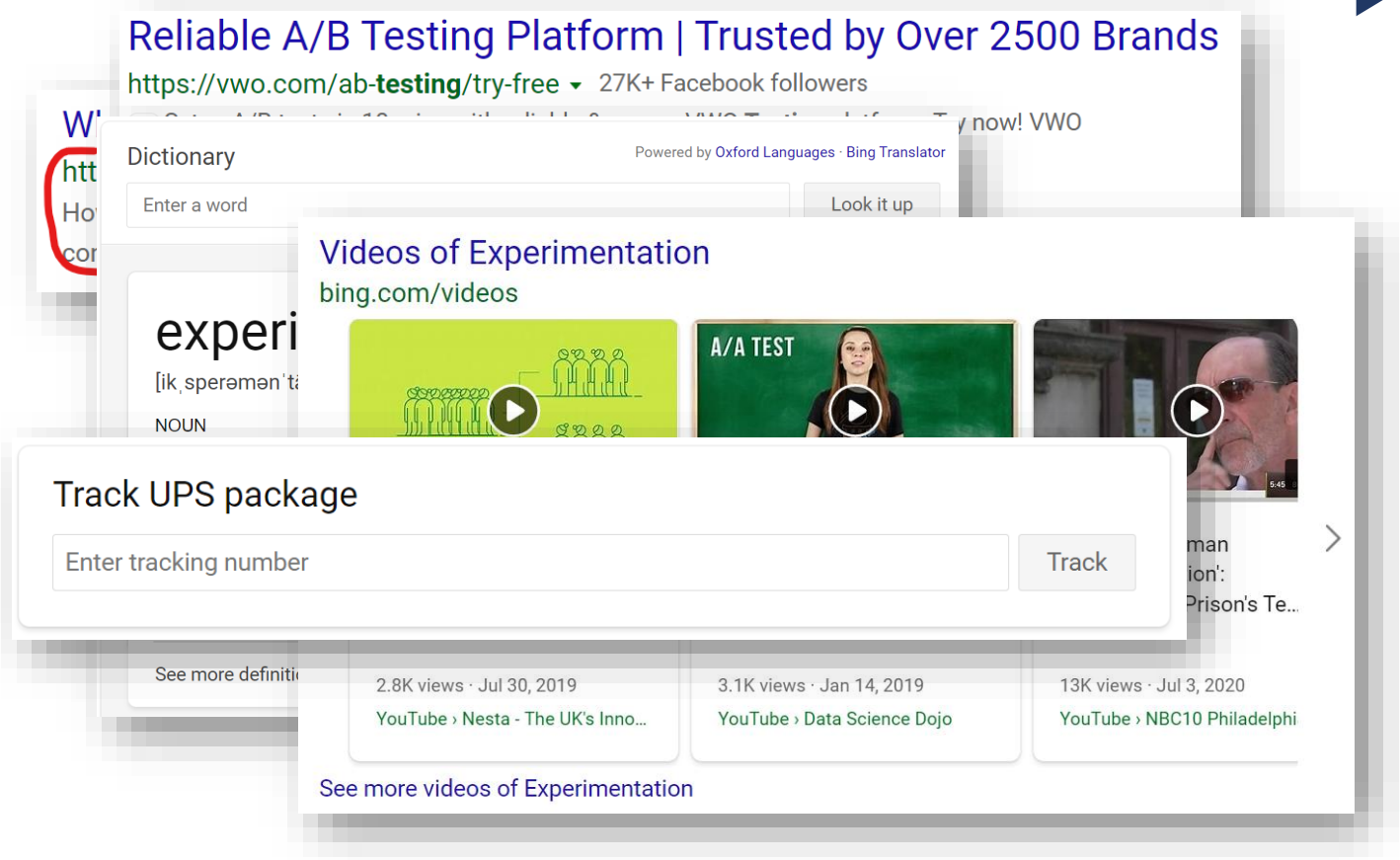
What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers



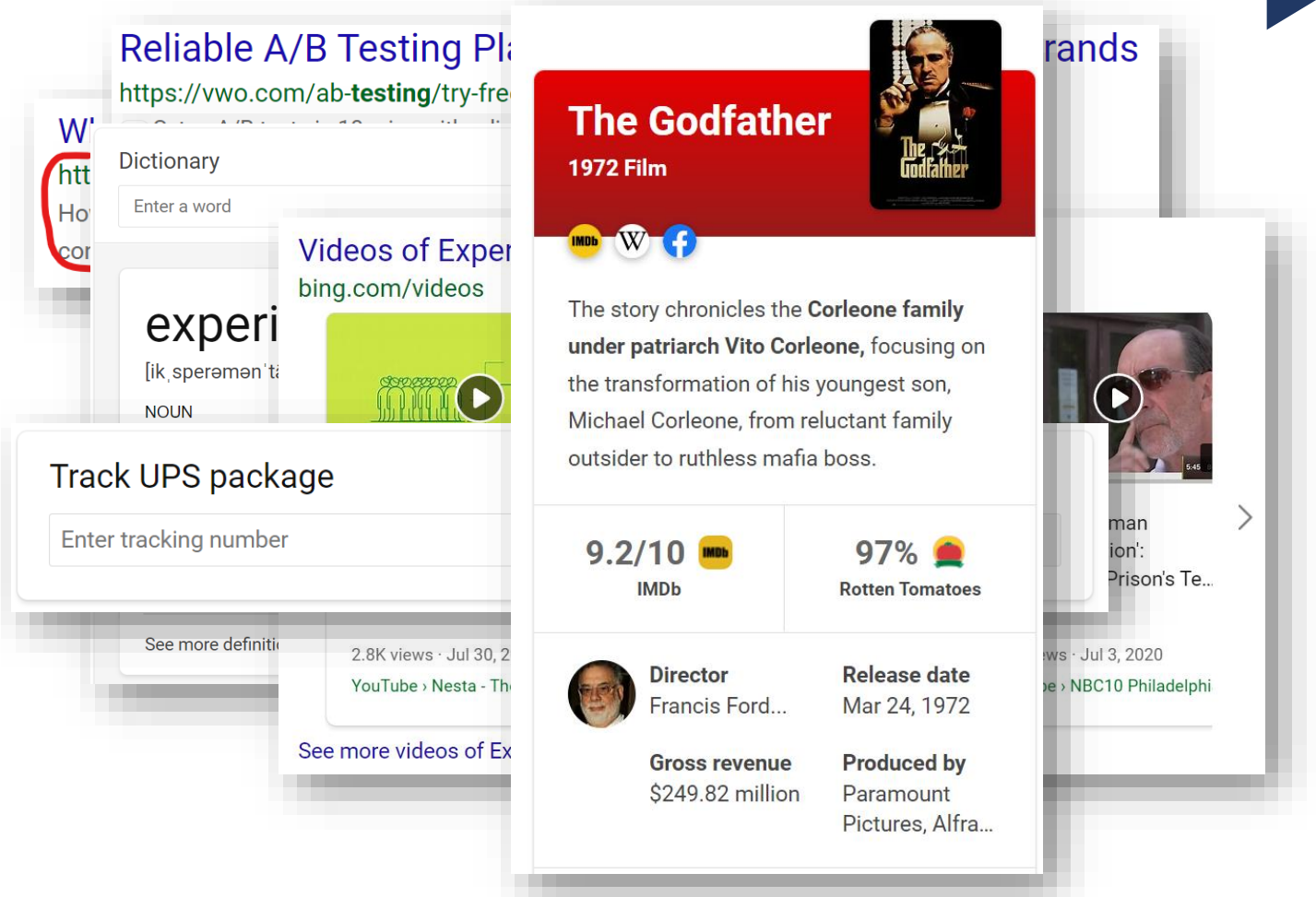
What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers



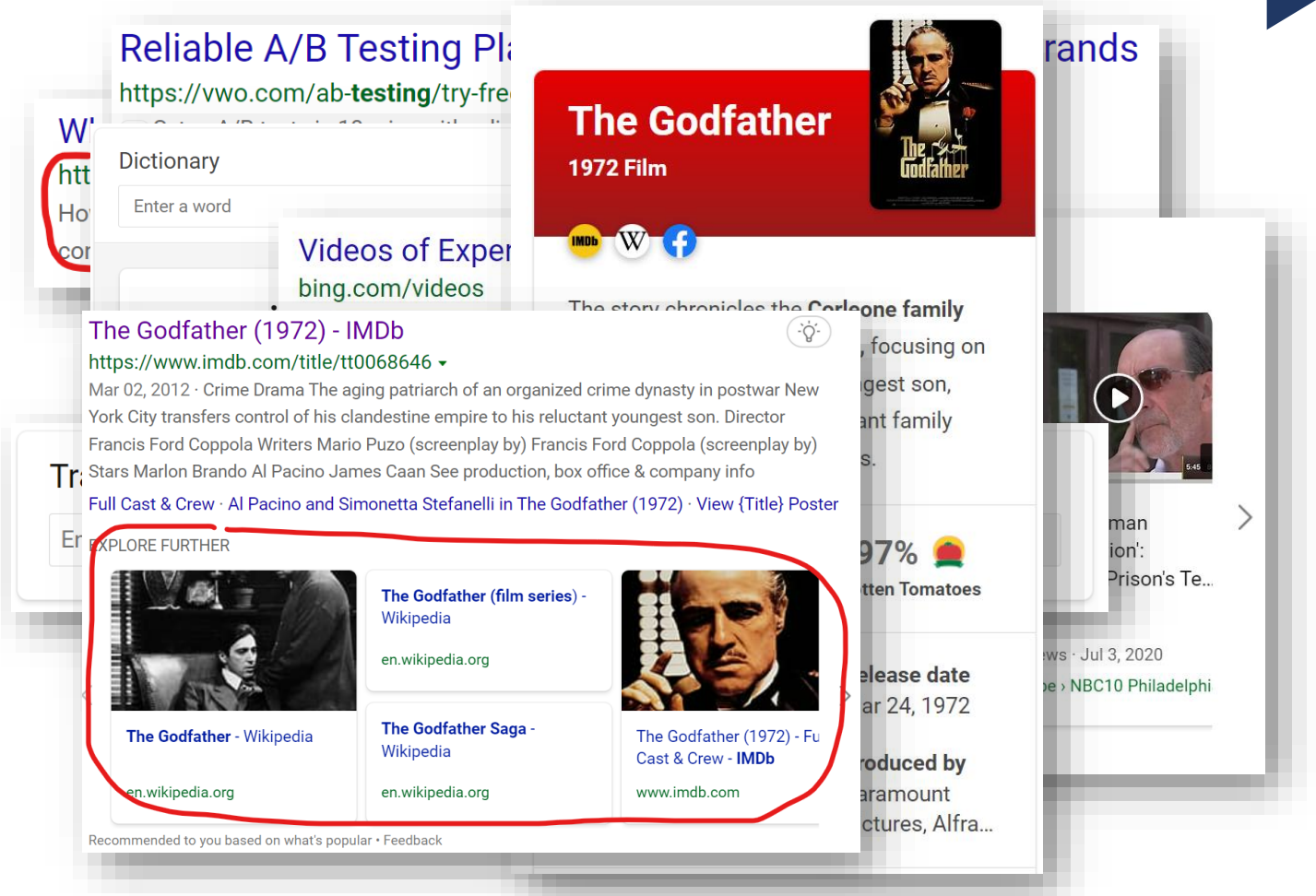
What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content



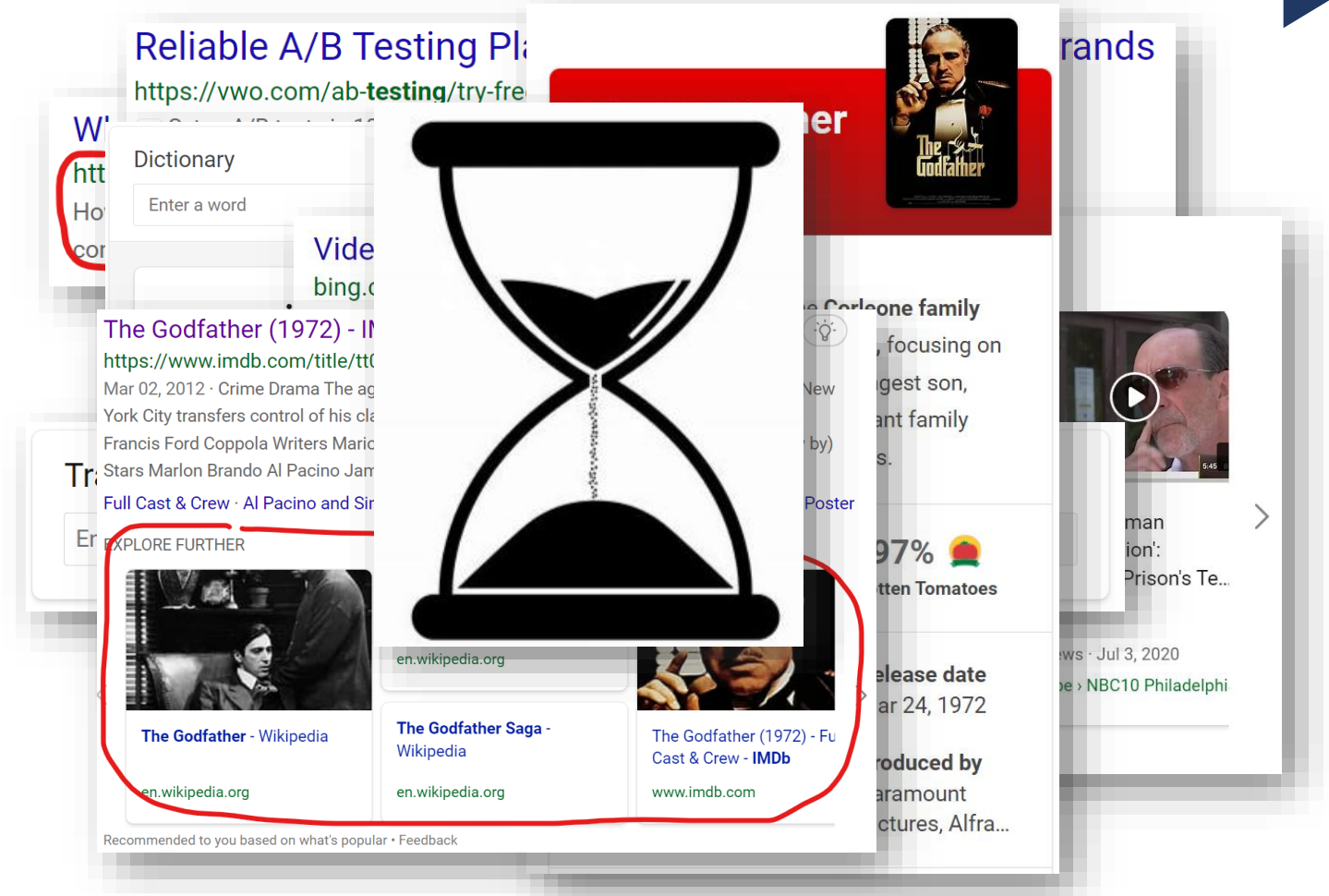
What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content
- Dynamic content



What Do You Want to Measure?

- Snippets
- Deep links
- Direct answers
- Media answers
- Utility answers
- Right/left rail content
- Dynamic content
- Performance
- …



How Do You Measure All This?

Online Controlled Experimentation

How Does It Work?

In a nutshell:

- Split real users into 2 (or more) groups: treatment, and control that get different experiences
- Run the experiment for a while and log whatever you can, importantly: any interactions the user is doing
- Compute metrics based on logged data

What can be re-used?

DCG-Style Evaluation

- Ground truth!
- Labeled results
- Metric definition

What can be re-used?

DCG-Style Evaluation

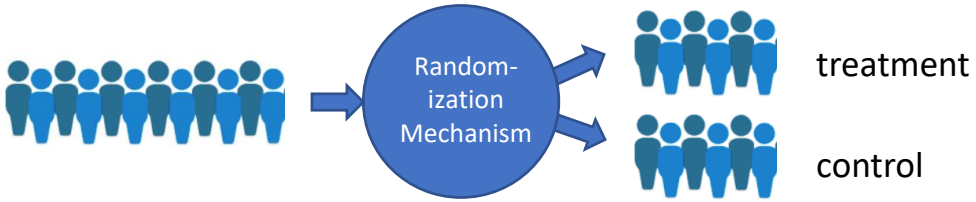
- Ground truth!
- Labeled results
- Metric definition

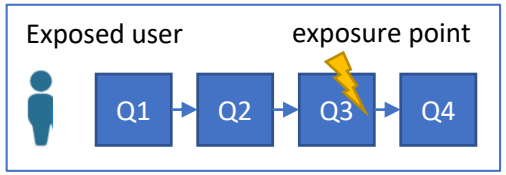
Controlled A/B Experimentation

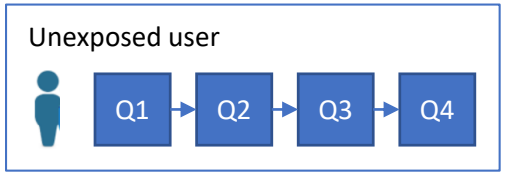
- No ground truth!
- Measurement framework
- Experimentation protocols
- Metrics
- Metric evaluation guidance

Measurement Framework

- Randomizing users into treatment/control
  - Multiple ways of doing this that can influence chances for false positive metric movements (seed finding!).
- Logging exposure to the intervention
  - Including exposed users only can drastically increase your ability to detect changes.
- Logging system responses
  - Everything from page load time to what results and UI elements were shown to the user.
- Logging user actions
  - Any explicit action taken by the user with timestamp.








```
<Resultspage loadtime="1">
   <Algo pos="1" title="abc" url="http://abc.de">
      <Deeplink title="def" url="..."/>
      <Deeplink title="def" url="..."/>
   </Algo>
   ...
</Result>
```

- Clicks on URLs, dynamic content
- Typing
- Query submission
- Viewport changes
- Mouse cursor movement

Experimentation Protocols

Setup

- Declare a hypothesis before starting the experiment
  - What will show the experiment is working as expected?
  - What will show the treatment is better than control?
  - What are important guardrails?
- Size the experiment to properly power the hypothesized metrics

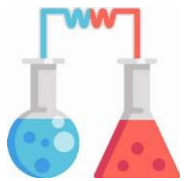
Start / Monitor

- Figure out appropriate time period to run on
  - Multiples of 1 week to avoid day-of-week effects
  - Long enough so users get over an initial period with novelty effects
- Monitor the health of your experiment
- No accidental changes to the experiment setup while running

Stop

- Double-check metric power to make sure you gathered enough data

Metrics



1. Experiment validity

Does the experiment produce valid data?

- User balance
- Log size
- Invalid log events (e.g., wrong order, etc.)

2. System Behavior

Is your IR system behaving as expected?

- Performance
- No results
- Ranking of certain result types
- Client errors

3. Behavioral / implicit feedback

Do users react to it favorably?

- Good vs bad clicks and interactions
- Search effort (reformulations, time, …)
- Abstractions on session and user-day level

4. Survey / explicit feedback

Do users like it?

- User satisfaction
- Net promoter score

Decision Making Guidance

Check Experiment Validity

- Balanced assignment of users to treatment and control?
- Any metric movements that are surprising could indicate other unfairness

Abide by hypothesis

- Did the system respond as expected?
- Did users change their behavior as expected?
- Were any guardrails violated?

Ideal: Back-test experiment

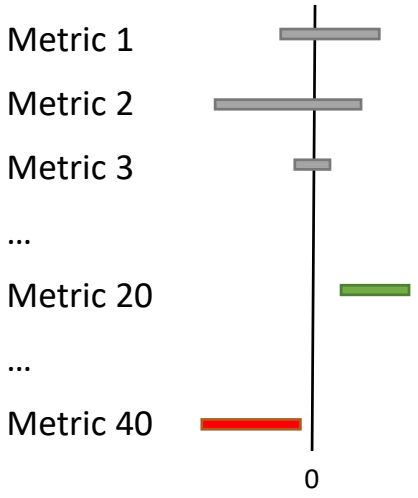
- to confirm gains, particularly when experiment has been of exploratory nature
  - e.g., multiple treatments tested in parallel

Common Pitfalls and Challenges

Reporting Results from Online Controlled Experiments can Easily Go Wrong!

Cherry Picking Metrics

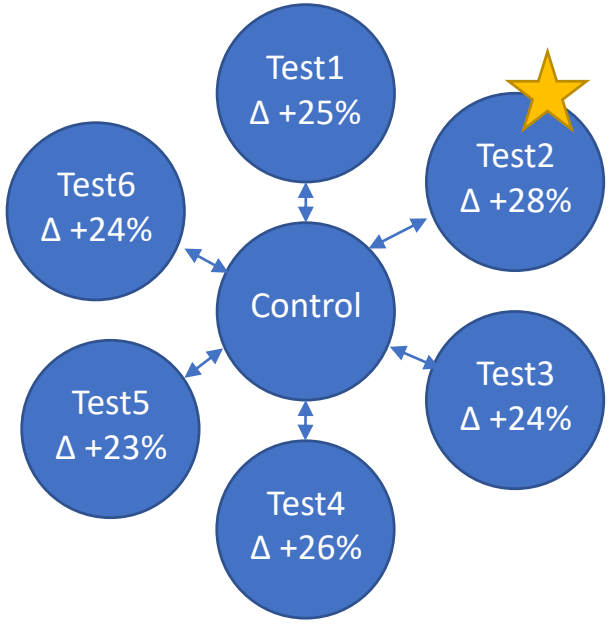
- The more metrics you will look at, the more of them will move.
  - Using Pval of < 0.05 (or 95% conf interval): 1 in 20 metrics expected to move by chance.
  - What is real, what is noise?
- Failure mode: cherry-pick what's green, explain away the red as noise.
- Stick to your hypothesis!



Selecting the Best of Multiple Treatments

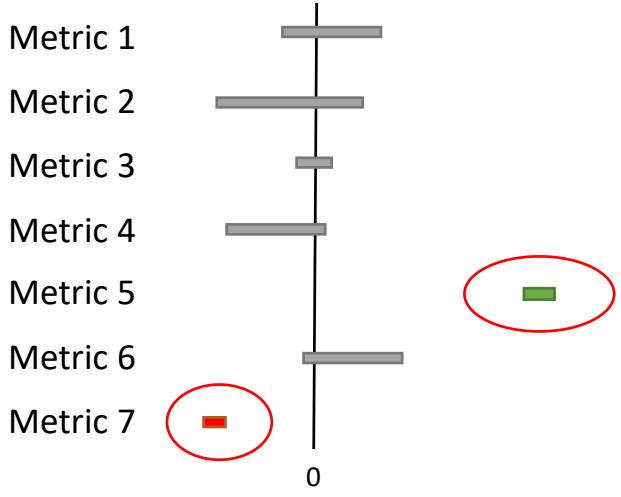
- Challenge when running many treatment variants: selection bias
  - E.g., parameter optimization, to find the best parameter setting.
- Selecting the best run from many does not necessarily yield the best setting.
  - Thought experiment: even if you ran multiple parallel A/A tests, one would be the "best"!

Best practice: Re-run experiment for winning treatment to confirm effect.

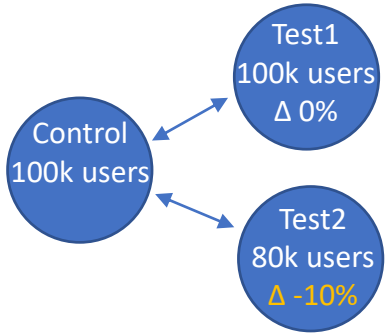


Surprising Metric Movements

- Surprisingly strong metric movements usually indicate that something is wrong.
  - Never take at face value!
  - Debug and confirm you understand the root cause in all its facets.

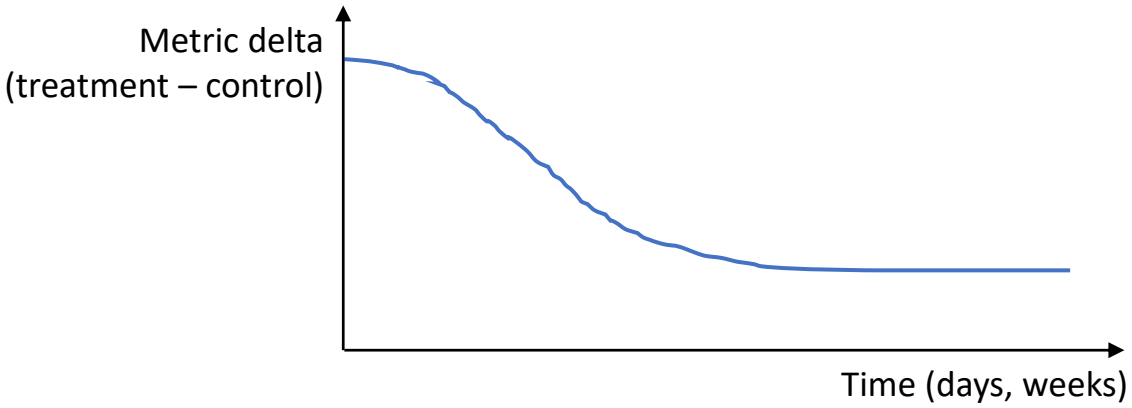


- Example:
  - Two treatments, one control
  - Test 2 has fewer users than the others by design
    - Nothing wrong with it per se…
    - But surprising metric movement!
  - → Turns out Test1, Test2, Control each used their own search result cache.
  - → Makes the comparison unfair as their caches warm up at different rates.



Novelty Effects

- Depending on the nature of the treatment, it may cause a novelty effect.
  - Users react differently for a while ("kick the tires"), then adopt a routine

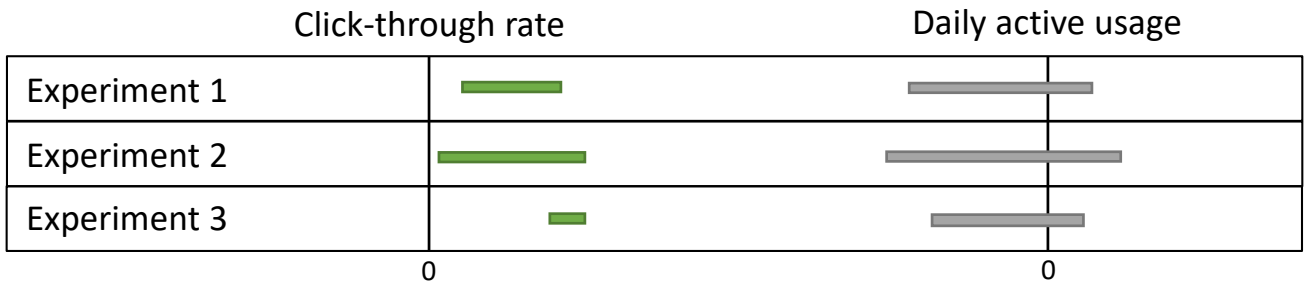


- Awareness for potential novelty effects and explicit investigation whether they exist for a particular experiment are necessary.

Sensitive Target Metrics vs Insensitive Guardrails

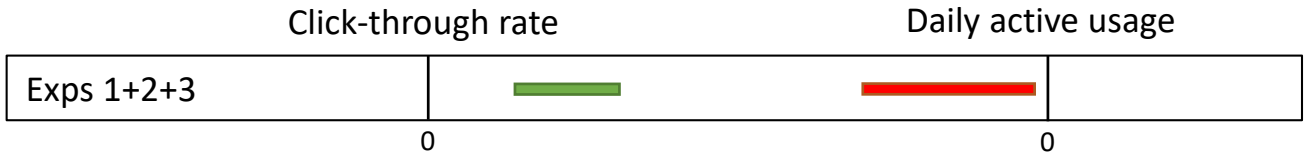
Example

- Target: increase click-through-rate.
- Guardrail: daily active usage should not regress.



Each experiment increased the target while "not moving" the guardrail. That's usually enough to declare a win.

- Power analysis really matters!
- Cumulative holdouts can help, if appropriate.



Summary

Re-Use in Controlled A/B Experimentation

- Measurement framework
- Experimentation protocols
- Metrics
- Metric evaluation guidance

Common Pitfalls

- Cherry picking metrics
- Selecting from multiple treatments
- Handling surprising metric movements
- Novelty effects
- Different levels of statistical sensitivity