UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Information Retrieval

Online Evaluation

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Ranking evaluation

Lots of alternative solutions

- Which one to choose?
- How to improve upon them?

Evaluation enables an informed choice

- Rigor of science
- Efficiency of practice

Evaluation methodology

Feedback

- Implicit
- Explicit

Mode

- Retrospective
- Prospective

<table>
<tr><th></th><th>retrospective</th><th>prospective</th></tr>
<tr><th>implicit</th><td>counterfactual evaluation</td><td rowspan="2">online evaluation</td></tr>
<tr><th>explicit</th><td>offline evaluation</td></tr>
</table>

Offline evaluation

Retrospective experiments

- How well can we predict (hidden) ***past preferences***?

Benchmarked using static test collections

- High throughput
- High reproducibility


Credit: NIST

Offline evaluation limitations

Scalability

- Relevance judgments are costly
- More so if expert judgments are needed

Realism

- Hired judges aren’t real users
- Laboratory studies aren’t naturalistic

Offline results often don’t hold live

Features are built because we believe they are useful

- Most experiments show that features fail to move the metrics they were designed to improve

Observations based on experiments at Microsoft [Kohavi et al., 2009]

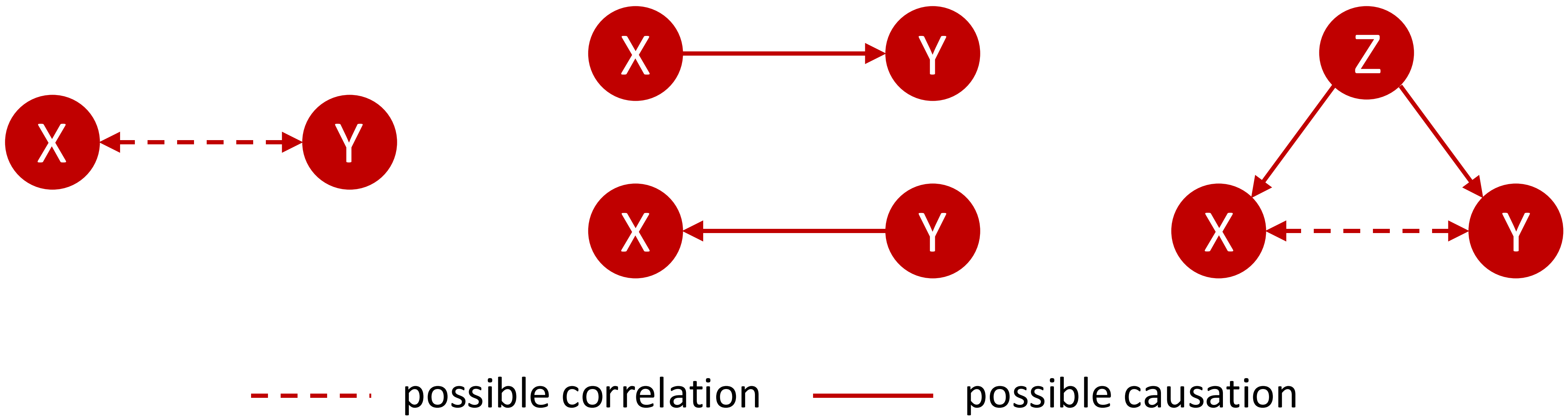
- 1/3 good, 1/3 bad, 1/3 neutral ideas

Why do offline and online eval disagree?

Causality

Offline data allows for mining correlations

- But correlation does not imply causation!



Example flawed analysis

Observation (highly stat-sig)

- Palm size negatively correlates with life expectancy
- The larger your palm, the less you will live

Gender is the common cause

- Women have smaller palms and live 6 years longer than men on average

Online evaluation

Focus on implicit user feedback

- Derived from observable user activity
- Captured during natural interaction

Implicit signals with various levels of noise

- Clicks, dwell-times, purchase decisions

Allows for detecting causation

Controlled experiments

> “ *An experiment is a procedure carried out to support, refute, or validate a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated.*
>
> ◦ http://en.wikipedia.org/wiki/Experiment

Controlled experiments

When different variants run concurrently, only two things could explain a change in metrics

- #1: their “feature(s)” (A vs. B)
- #2: random chance

Everything else happening affects both the variants

- For #2, we conduct statistical tests for significance

Hypotheses and variables

Example hypothesis

- H: increasing the weight given to document recency in the ranking will increase user click-through rate

Variables of interest

- X: independent variable (recency weight)
- Y: dependent variable (user click-through rate)

Hypotheses and variables

Alternative hypothesis

- H_1: increasing X will increase Y

Corresponding null hypothesis

- H_0: increasing X will **not** increase Y

How to support H_1?

- Show that H_0 is improbable!

Unit of experimentation

Defines the granularity of the experiment

- User (most typical), query, user+day

Smaller units (e.g., queries)

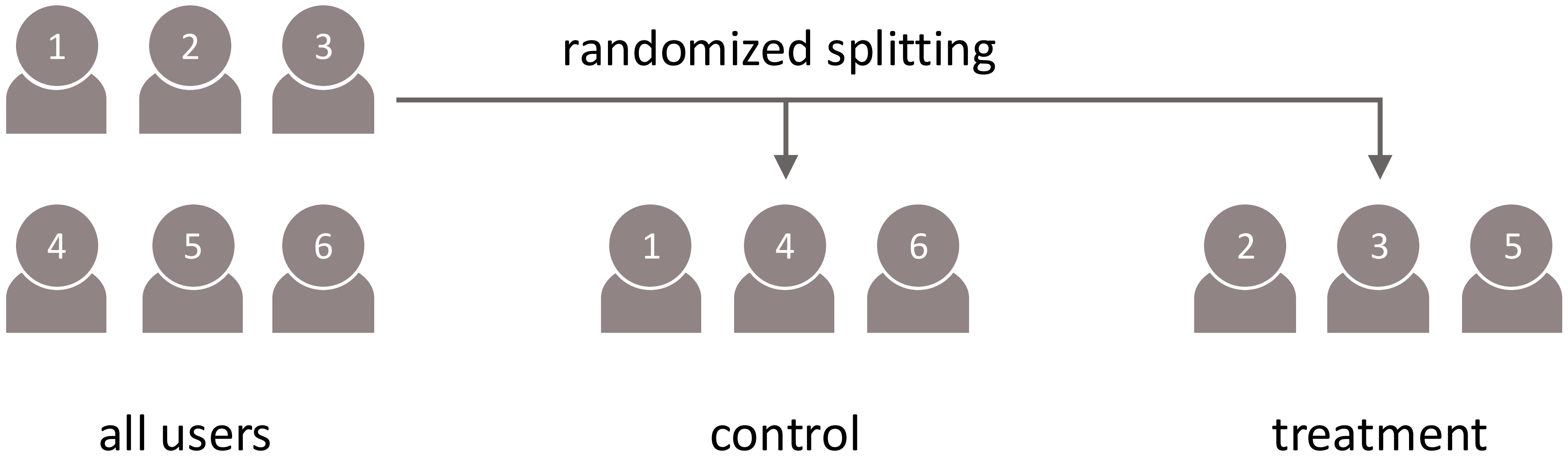
- Reduced data requirements

Larger units (e.g., users)

- Reduced risk of network effects

Between-subject experiments

Each user is exposed to a single variant



A/B test

Randomly split traffic between two (or more) versions

- A (control, typically the existing system)
- B (treatment)

Collect metrics of interest

Analyze

A/B test

A/B/n is common in practice

- Compare A/B/C/D/…, not just two variants
- Sensitive to small changes (given large samples)

Equivalent names

- Flights (Microsoft), 1% tests (Google), bucket tests (Yahoo!), randomized clinical trials (medicine)

Pre-test validation

A/A tests used to validate splitting

- Same approach (A) applied to different user groups

Ideally, no significant difference should be observed

- Outliers in either partition may introduce bias

In practice, A/A test over multiple splittings

- Significant differences should rarely occur (under 5%)

Absolute metrics

Document-level

- Click rate, click models

Ranking-level

- Reciprocal rank, CTR@k, time-to-click, abandonment

Session-level

- Queries per session, session length, time to first click

Relative metrics

Absolute document-level metrics are biased

- Position bias: top ranked document favored
- Presentation bias: highlighted documents favored

Relative document-level metrics are less affected

- Click-skip, fair pairs

Can we compare rankings to each other?

Within-subject experiments

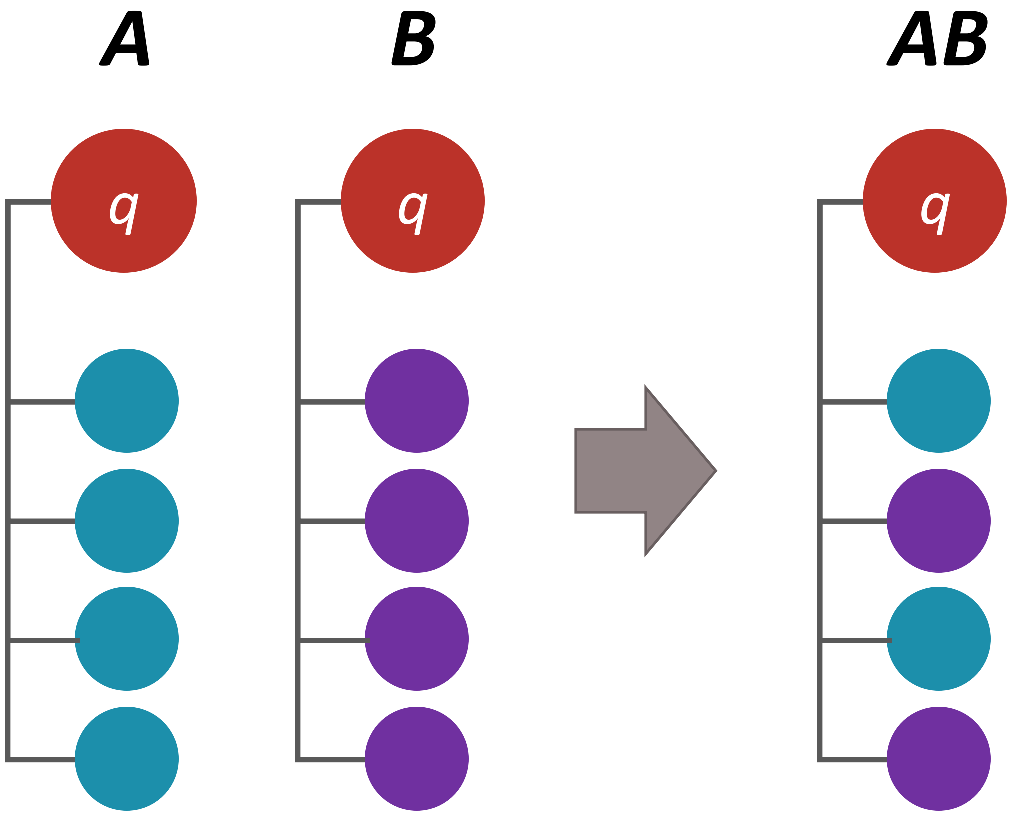
Side-by-side experiments are common in lab studies

- Not naturalistic to run in production systems though

Solution: interleaving

- Mix results from different rankings
- Observe user feedback (e.g., clicks)
- Credit feedback to original rankers

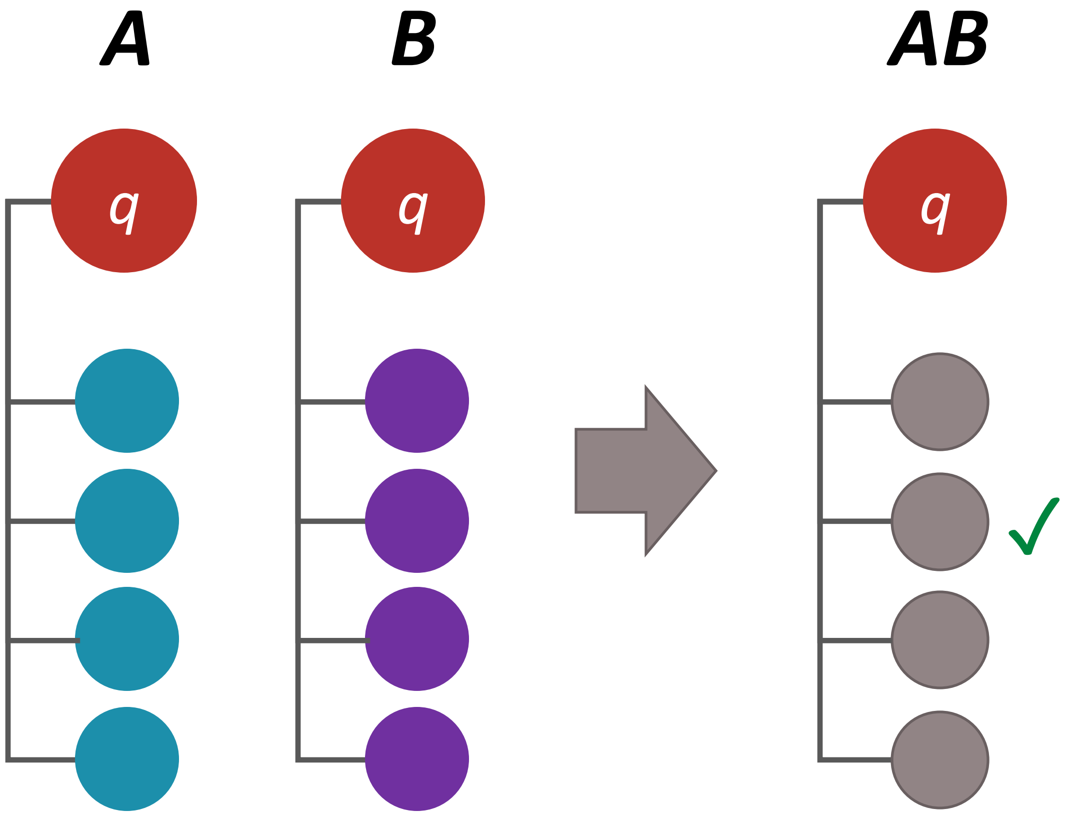
Interleaved comparisons
[Joachims, KDD 2002]



Blend results from both conditions into a single ranking

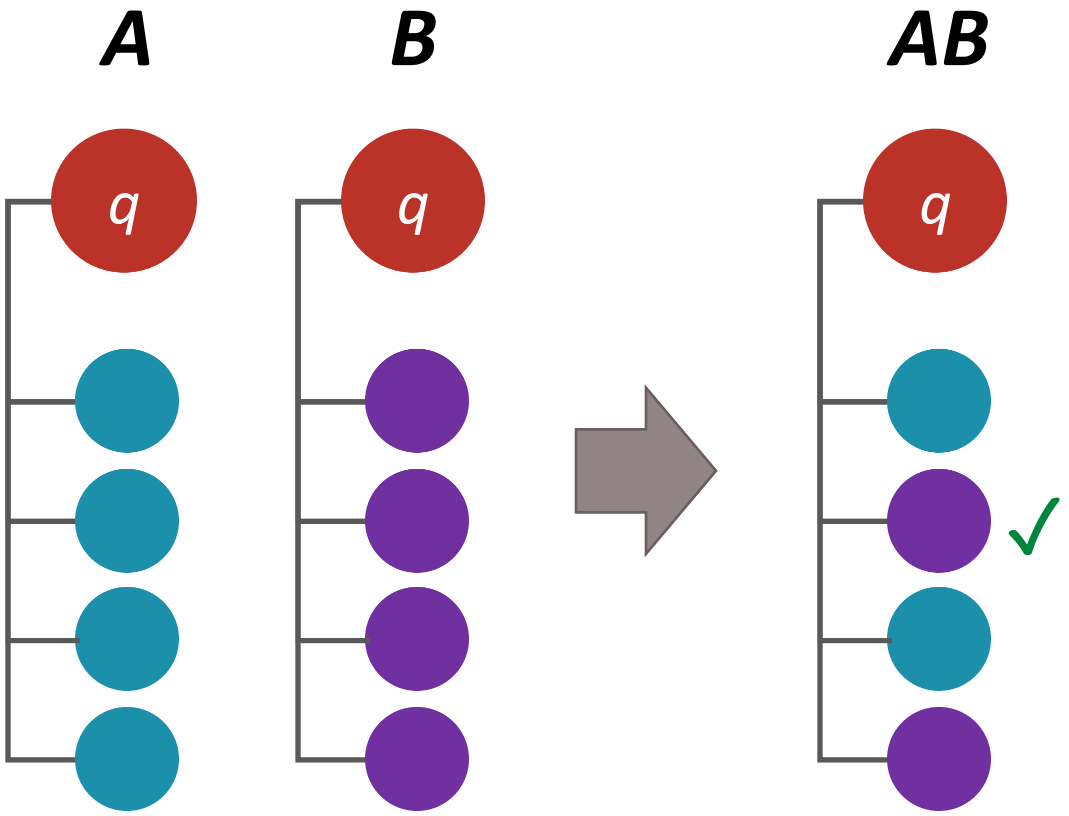
Interleaved comparisons

[Joachims, KDD 2002]



Hide provenance from the user and collect feedback

Interleaved comparisons
[Joachims, KDD 2002]



Assign credit:
e.g., B wins over A

Balanced Interleaving
[Joachims, KDD 2002]

ALGORITHM 1: Balanced Interleaving, following [Chapelle et al. 2012].

Input: $\mathbf{l}_1, \mathbf{l}_2$
$\mathbf{l} = [\,]; i_1 = 0; i_2 = 0$
$\mathit{first\_1} = \mathit{random\_bit}()$ — decide who gets priority
while $(i_1 < \mathit{len}(\mathbf{l}_1)) \wedge (i_2 < \mathit{len}(\mathbf{l}_2))$ **do** — if not end of *A* or *B*
  if $(i_1 < i_2) \vee ((i_1 == i_2) \wedge (\mathit{first\_1} == 1))$ **then** — if A least explored or A has priority
    if $\mathbf{l}_1[i_1] \notin \mathbf{l}$ **then**
      $\mathit{append}(\mathbf{l}, \mathbf{l}_1[i_1])$ — append next *A* result
    $i_1 = i_1 + 1$
  else
    if $\mathbf{l}_2[i_2] \notin \mathbf{l}$ **then**
      $\mathit{append}(\mathbf{l}, \mathbf{l}_2[i_2])$ — append next *B* result
    $i_2 = i_2 + 1$
  // present $\mathbf{r}$ *to user and observe clicks* $\mathbf{c}$*, then infer outcome (if at least one click was observed)*
d_{max} = lowest-ranked clicked document in $\mathbf{l}$
$k = \min\left\{j : (d_{max} = \mathbf{l}_1[j]) \vee (d_{max} = \mathbf{l}_2[j])\right\}$ — earliest rank of d_{max} in *A* or *B*
$c_1 = \mathit{len}\left\{i : c[i] = \mathit{true} \wedge \mathbf{l}[i] \in \mathbf{l}_1[1..k]\right\}$
$c_2 = \mathit{len}\left\{i : c[i] = \mathit{true} \wedge \mathbf{l}[i] \in \mathbf{l}_2[1..k]\right\}$ — count clicks in *A* and *B* up to position *k*
return -1 **if** $c_1 > c_2$ **else** 1 **if** $c_1 < c_2$ **else** 0

Balanced Interleaving
[Joachims, KDD 2002]

Each query produces a single comparison result

- Either A or B wins, or there is a tie

Degree of preferences computed across queries

- $\Delta_{AB} = \frac{wins(A) + 0.5\ ties(A,B)}{wins(A) + wins(B) + ties(A,B)} - 0.5$

Interleaving extensions

Team draft interleaving [Radlinski et al., CIKM 2008]

◦ Randomizes provenance of duplicate documents

Probabilistic interleaving [Hofmann et al., CIKM 2011]

◦ Sample over probabilistic input rankings

(Probabilistic) multileaving [Schuth et al., CIKM 2014, SIGIR 2015]

◦ Mix multiple (possibly infinitely many) rankers

Long-term metrics
[Hohnhold et al., KDD 2015]

Measuring short-term effects is straightforward

- i.e., just run an A/B or interleaved test

Search engines are evaluated on market share (distinct queries per month) and revenue as long-term goals

- How can we measure (and influence) these?

Long-term metrics
[Hohnhold et al., KDD 2015]

Revenue can be broken down according to

- $$\frac{Revenue}{Period} = \underbrace{\frac{Users}{Period}}_{1} \underbrace{\frac{Sessions}{User}}_{2} \underbrace{\frac{Queries}{Session}}_{3} \underbrace{\frac{Ads}{Query}}_{4} \underbrace{\frac{Clicks}{Ad}}_{5} \underbrace{\frac{Cost}{Click}}_{6}$$

- 1 and 2: harder to influence
- 4 and 6: easier to influence, but negative impact
- 3 and 5: easier to influence, with positive impact

Long-term metrics
[Hohnhold et al., KDD 2015]

Revenue can be broken down according to

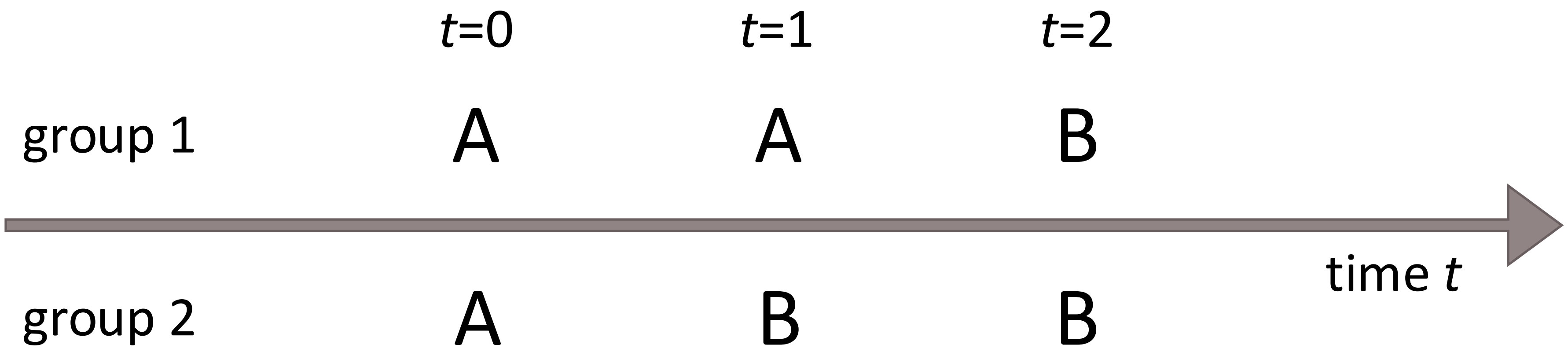
- $$\frac{Revenue}{Period} = \frac{Users}{Period}\frac{Sessions}{User}\frac{Queries}{Session}\frac{Ads}{Query}\frac{Clicks}{Ad}\frac{Cost}{Click}$$

1 2 3 4 5 6

Are any of these impacts persistent in the long run?

Long-term metrics
[Hohnhold et al., KDD 2015]

Long-term impact of B



A1=A2

B1=B2 ***no impact***

B1≠B2 ***long-term impact***

The cultural challenge
[Deng et al., SIGIR 2017, KDD 2017]

“ *It is difficult to get a man to understand something when his salary depends upon his not understanding it.*

- Upton Sinclair

The cultural challenge
[Deng et al., SIGIR 2017, KDD 2017]

Why people/orgs avoid controlled experiments

- Some believe it threatens their job as decision makers
- Proposing several alternatives and admitting you don’t know which is best is hard
- Failures of ideas may hurt professional standing
- “We know what to do. It’s in our DNA!”

The cultural challenge
[Deng et al., SIGIR 2017, KDD 2017]

Dismissing controlled experiments as a guiding mechanism means following the HiPPO

- HiPPO = Highest Paid Person's Opinion

Summary

Online evaluation via controlled experiments

◦ Crucial to measure causal effects on user behavior

Several methods proposed for ranking evaluation

◦ Between-subject, within subject experiments

Can be leveraged to guide learning to rank

◦ Incremental learning from user interactions

References

[Online evaluation for information retrieval]
Hofmann et al., FnTIR 2016

[A/B testing at scale: accelerating software innovation]
Deng et al., SIGIR 2017 / KDD 2017

References

Trustworthy online controlled experiments: five puzzling outcomes explained
Kohavi et al., KDD 2012

Focusing on the long-term: it's good for users and business
Hohnhold et al., KDD 2015

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next...

Online Learning to Rank

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br