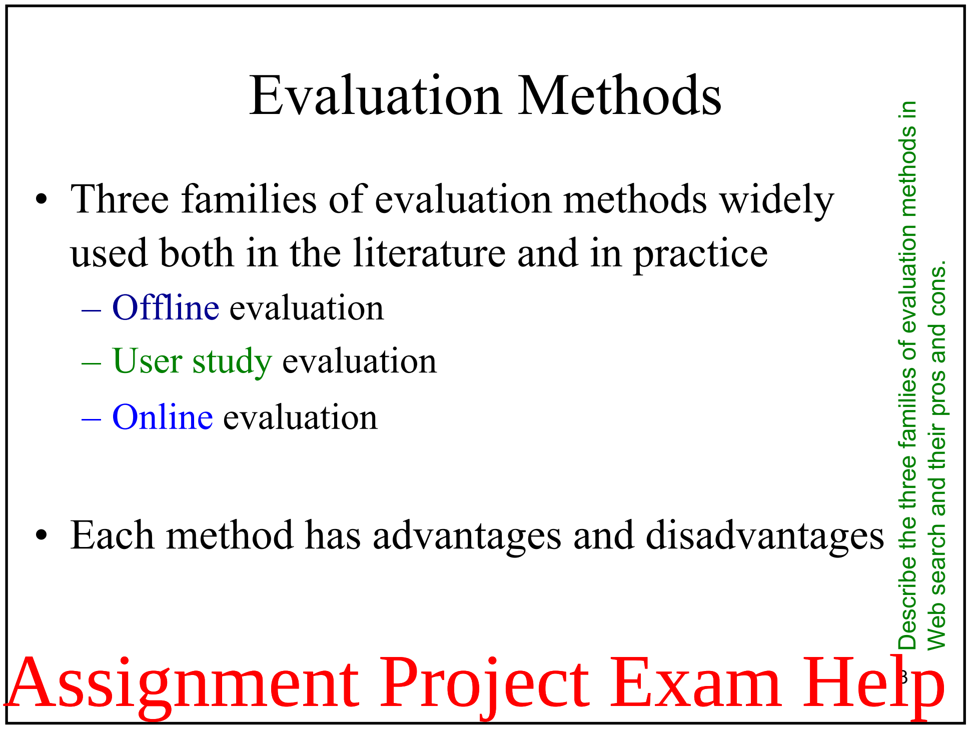


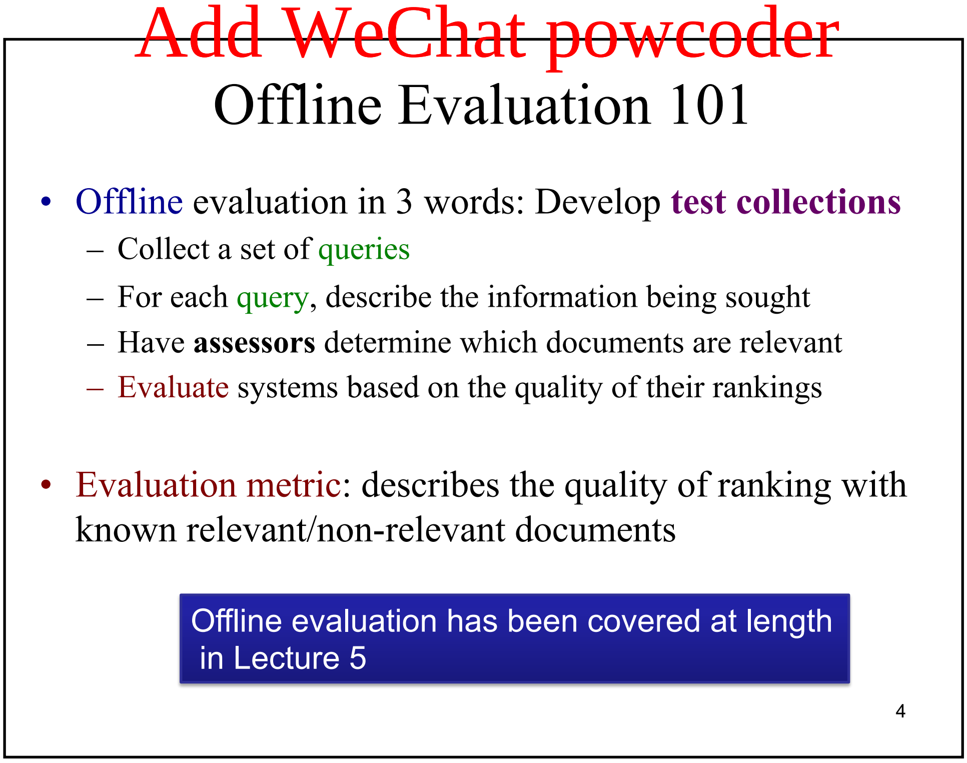
Evaluation Methods

- Three families of evaluation methods widely used both in the literature and in practice
  - Offline evaluation
  - User study evaluation
  - Online evaluation

- Each method has advantages and disadvantages

Describe the three families of evaluation methods in Web search and their pros and cons.

3



Offline Evaluation 101

- Offline evaluation in 3 words: Develop **test collections**
  - Collect a set of queries
  - For each query, describe the information being sought
  - Have **assessors** determine which documents are relevant
  - Evaluate systems based on the quality of their rankings

- Evaluation metric: describes the quality of ranking with known relevant/non-relevant documents

Offline evaluation has been covered at length in Lecture 5

4

Offline Evaluation Cont.

Advantages

- The experimental condition is fixed; same queries, and same relevance judgements
- Evaluations are **reproducible**; keeps us "honest"
- By experimenting on the same set of queries and judgements, we can better understand how one system is better than another

Disadvantages

- Human assessors that judge documents relevant/non-relevant are **expensive**
- Human assessors are not the user; judgements are made out of context
- Assumes that relevance is the same for every user

User Study Evaluation 101

User studies in 3 bullets:

- Provide a small set of users with several retrieval systems
- Ask them to complete several (potentially different) search tasks
- Learn about a system performance by
  - Observing what they do
  - Asking why they did it

The usual evaluation methods and techniques from HCI apply

D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3(1-2): 1-224, 2009

User Study Evaluation Cont.

Advantages

- **Detailed** data about users, and their reaction to systems
- In reality, a search is done to accomplish a higher-level **task**
- In user studies, this task can be manipulated and studied
- In other words, the experimental "starting-point" does not need to be the query

Disadvantages

- User studies are **expensive** (pay users/subjects, scientist's time, data coding, etc)
- Difficult to **generalise** from small studies to broad populations
- Environments where they are conducted are not necessarily the user's normal environment
- Need to re-run experiment every time a new system is considered

7

Online Evaluation 101

- See how users **interact** with your **live** retrieval system when just using it
- Treat some of the users by a changed version of the system
- Based on their behavior (e.g. implicit feedback), infer if they are more likely to prefer the changed system
- Examples of observed behavior: clicks, skips, saves, forwards, bookmarks, "likes", etc.
- Two main approaches
  - **A/B testing**: Have x/2% of query traffic use system A and x/2% of query traffic use system B where x is about 5% of traffic
  - **Interleaving**: Expose a **combination** of system versions to users

8

Online Evaluation

- **Assumption**: Observable user behavior reflects relevance
- This assumption gives us "high fidelity"
  - Real users replace the assessors: No ambiguity in information need; Users actually want results; Measures performance on real queries
- But introduces a major challenge …
  - We cannot train the users: How do we know when they are happy? Real user behavior requires careful design, metrics and evaluation
- … and noticeable drawbacks:
  - We need a lot of user data to compensate for **noisy** user interaction (e.g. clicks are noisy)
  - Data isn't trivially *reusable* later

Describe online evaluation, its main assumption, and its pros and cons.

Online Data & Bias

- A variety of data can describe online behavior
  - Queries, Results and Clicks
  - Mouse movement: Clicks, selections, hover
  - Eye tracking
- Can we simply interpret clicked results as relevant? A variety of biases make this difficult:
  - **Position Bias**: Users are more inclined to examine and click on higher-ranked results
  - **Contextual Bias**: Whether users click on a result depends on other nearby results
  - **Attention Bias**: Users click more on results which draw attention to themselves
  - **Accidental Clicking**; **Malicious Clicking**, etc.

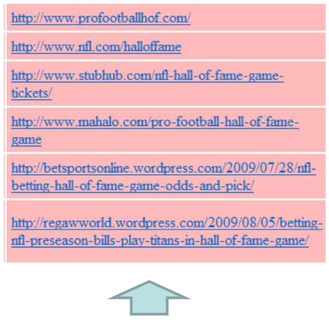
Describe five possible biases affecting users' interaction in Web search

Online Evaluation

A/B TESTING

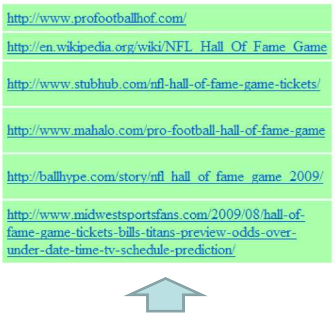
A/B Testing (1)

- Each user is assigned to one of two conditions
- They might see the left *or* the right ranking



1. http://www.profootballhof.com/
2. http://www.nfl.com/halloffame
3. http://www.stubhub.com/nfl-hall-of-fame-game-tickets/
4. http://www.mahalo.com/pro-football-hall-of-fame-game
5. http://betsportsonline.wordpress.com/2009/07/28/nfl-betting-hall-of-fame-game-odds-and-pick/
6. http://regawworld.wordpress.com/2009/08/05/betting-nfl-preseason-bills-play-titans-in-hall-of-fame-game/

Ranking A



1. http://www.profootballhof.com/
2. http://en.wikipedia.org/wiki/NFL_Hall_Of_Fame_Game
3. http://www.stubhub.com/nfl-hall-of-fame-game-tickets/
4. http://www.mahalo.com/pro-football-hall-of-fame-game
5. http://ballhype.com/story/nfl_hall_of_fame_game_2009/
6. http://www.midwestsportsfans.com/2009/08/hall-of-fame-game-tickets-bills-titans-preview-odds-over-under-date-time-tv-schedule-prediction/

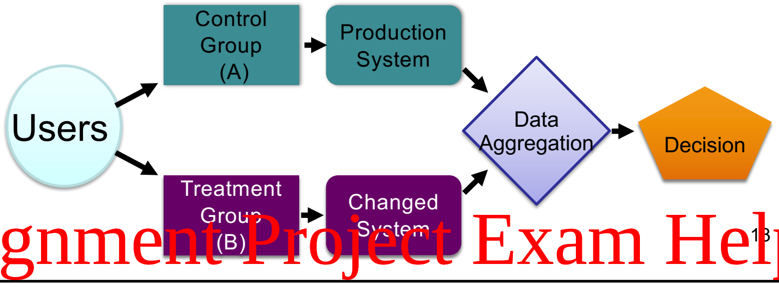
Ranking B

Describe the A/B testing procedure, its pros and cons.

- **Measure** user interaction with their system (e.g. clicks)
- Look for differences between the populations

A/B Testing (2)

- Concept is fairly trivial: Randomly split traffic between two (or more) versions
  - A (Control) & B (Treatment, i.e. Alternative System)
- Collect **metrics** of interest & Analyse
- Run **statistical tests** to confirm differences are not due to chance
- Best scientific way to demonstrate *causality* – changes in metrics are caused by changes introduced in the treatment(s)



13

A/B Testing Metrics

- Examples of online **metrics** used:
  - Abandonment Rate (% of queries with no click)
  - Mean Reciprocal Rank (mean of 1/rank for all clicks)
  - User Engagement (e.g. clicks per Query; Time to First Click, Time to Last Click, etc)
  - Sessions per User
  - Probability of Switching to another search engine
- **A/B tests** are used by many web companies such as Google, Bing, Facebook, etc.
  - Use of special experimental platforms allowing to run A/B tests at large-scale (e.g. 100s per-day)
- A/B experiments are not the panacea for everything (c.f. survey by Kohavi et al. 2009)
  - They can take a long time to complete

14

Advantages of A/B Testing

- When the variants run **concurrently**, only two things could explain a **change** in metrics:
  1. The "feature(s)" (A vs. B)
  2. Random chance

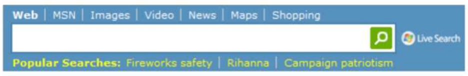
- Everything else happening affects both variants

- For #2, conduct **statistical tests** for significance (e.g. Student's t-test)

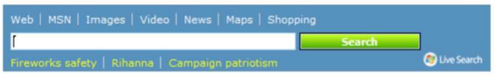
A/B Testing Guidelines

- Perform many sanity checks
  - E.g. too many unsuccessful tests means that lots of users are experiencing degraded performance
- Run an **A/A** test!
- If something is "*amazing*", find the **flaw**
  - Look for confounding variables – e.g. ensure that not too many variables have changed

□ **A**


Web | MSN | Images | Video | News | Maps | Shopping
Live Search
Popular Searches: Fireworks safety | Rihanna | Campaign patriotism

□ **B**


Web | MSN | Images | Video | News | Maps | Shopping
Search
Fireworks safety | Rihanna | Campaign patriotism
Live Search

Kohavi et al., 2013

  - etc.

Describe three A/B testing guidelines that help increase confidence in the results.

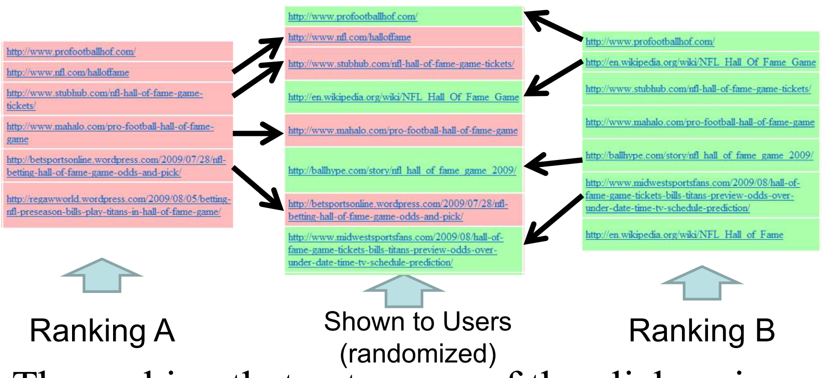
Online Evaluation

INTERLEAVING

17

Online Evaluation: Interleaving

- A within-user online ranker comparison
  - Presents results from both rankings to every user



Ranking A

- http://www.profootballhof.com/
- http://www.nfl.com/halloffame
- http://www.stubhub.com/nfl-hall-of-fame-game-tickets/
- http://www.mahalo.com/pro-football-hall-of-fame-game
- http://betsportsonline.wordpress.com/2009/07/28/nfl-betting-hall-of-fame-game-odds-and-pick/
- http://regawworld.wordpress.com/2009/08/05/betting-nfl-preseason-bills-play-titans-in-hall-of-fame-game/

Shown to Users (randomized)

- http://www.profootballhof.com/
- http://www.nfl.com/halloffame
- http://www.stubhub.com/nfl-hall-of-fame-game-tickets/
- http://en.wikipedia.org/wiki/NFL_Hall_Of_Fame_Game
- http://www.mahalo.com/pro-football-hall-of-fame-game
- http://ballhype.com/story/nfl_hall_of_fame_game_2009/
- http://betsportsonline.wordpress.com/2009/07/28/nfl-betting-hall-of-fame-game-odds-and-pick/
- http://www.midwestsportsfans.com/2009/08/hall-of-fame-game-tickets-bills-titans-preview-odds-over-under-date-time-tv-schedule-prediction/

Ranking B

- http://www.profootballhof.com/
- http://en.wikipedia.org/wiki/NFL_Hall_Of_Fame_Game
- http://www.stubhub.com/nfl-hall-of-fame-game-tickets/
- http://www.mahalo.com/pro-football-hall-of-fame-game
- http://ballhype.com/story/nfl_hall_of_fame_game_2009/
- http://www.midwestsportsfans.com/2009/08/hall-of-fame-game-tickets-bills-titans-preview-odds-over-under-date-time-tv-schedule-prediction/
- http://en.wikipedia.org/wiki/NFL_Hall_of_Fame

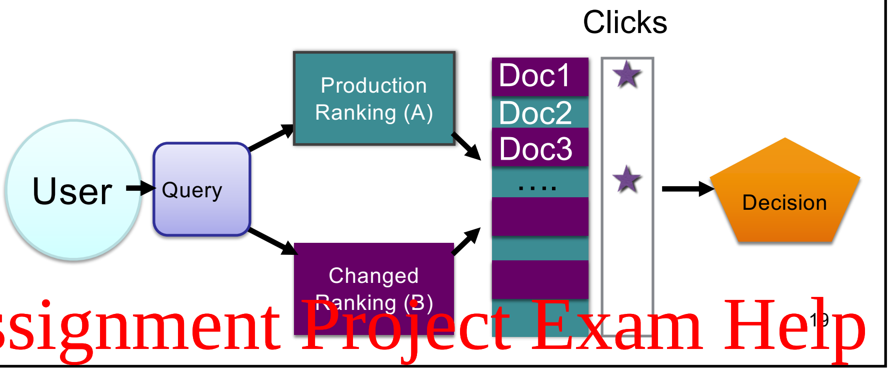
Describe the interleaving process.

- The ranking that gets more of the clicks wins
  - Designed to be unbiased, and much more sensitive than A/B

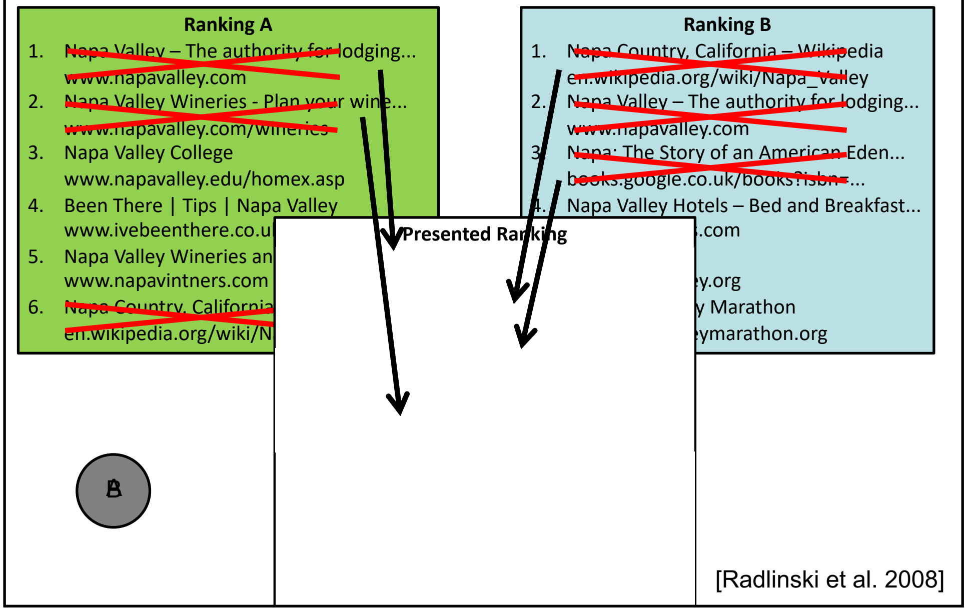
18

Interleaving

- Procedure:
  - Generate interleaved result list
  - Keep track of **assignments** (which ranker contributed which document)
  - Observe user behavior (e,g., clicks)
  - **Credit** clicks to original rankers to infer outcome



Team Draft Interleaving



Ranking A

1. ~~Napa Valley – The authority for lodging...~~ ~~www.napavalley.com~~
2. ~~Napa Valley Wineries - Plan your wine...~~ ~~www.napavalley.com/wineries~~
3. Napa Valley College www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley www.ivebeenthere.co.u
5. Napa Valley Wineries an www.napavintners.com
6. ~~Napa Country, California~~ ~~en.wikipedia.org/wiki/N~~

Ranking B

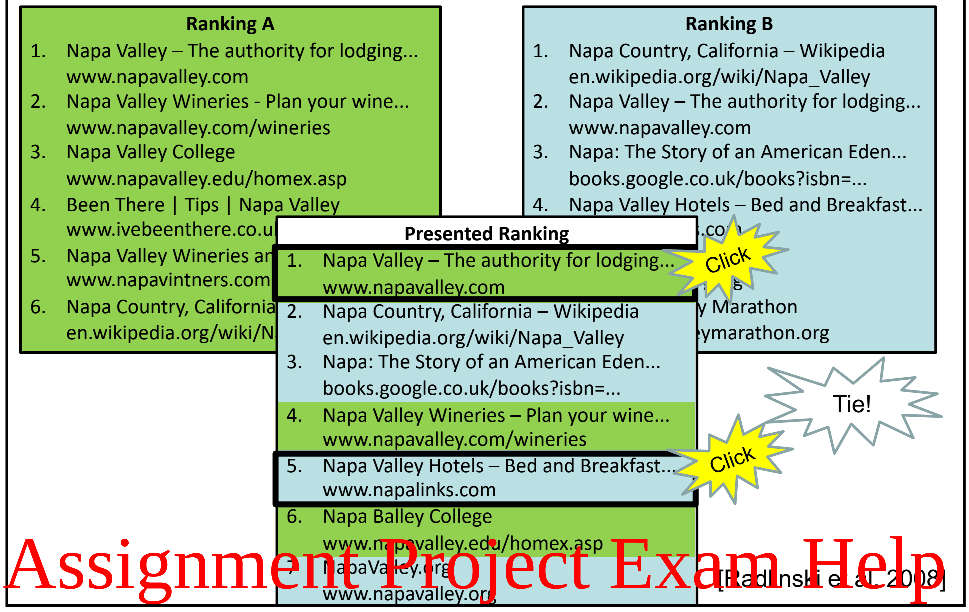
1. ~~Napa Country, California – Wikipedia~~ ~~en.wikipedia.org/wiki/Napa_Valley~~
2. ~~Napa Valley – The authority for lodging...~~ ~~www.napavalley.com~~
3. ~~Napa: The Story of an American Eden...~~ ~~books.google.co.uk/books?isbn=...~~
4. Napa Valley Hotels – Bed and Breakfast... .com

y.org

y Marathon ymarathon.org

[Radlinski et al. 2008]

Team Draft Interleaving



Ranking A

1. Napa Valley – The authority for lodging... www.napavalley.com
2. Napa Valley Wineries - Plan your wine... www.napavalley.com/wineries
3. Napa Valley College www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley www.ivebeenthere.co.u
5. Napa Valley Wineries ar www.napavintners.com
6. Napa Country, California en.wikipedia.org/wiki/N

Ranking B

1. Napa Country, California – Wikipedia en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging... www.napavalley.com
3. Napa: The Story of an American Eden... books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast... .com

y Marathon eymarathon.org

Presented Ranking

1. Napa Valley – The authority for lodging... www.napavalley.com
2. Napa Country, California – Wikipedia en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden... books.google.co.uk/books?isbn=...
4. Napa Valley Wineries – Plan your wine... www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast... www.napalinks.com
6. Napa Balley College www.napavalley.edu/homex.asp
7. NapaValley.org www.napavalley.org

[Radlinski et al. 2008]

Scoring Interleaved Evaluations

- Clicks credited to "owner" of result
  - Ranking r_1
  - Ranking r_2
  - Shared: A & B share top K results when they have identical results at each rank 1…K
- Ranking with more **credits** wins
- Needs a **statistical test**: e.g. Binomial test

$$\left(\mathbf{E}\left[\frac{C_A - C_B}{C}\right] > 0\right) \rightarrow (\mathrm{A} \succ \mathrm{B})$$

$$\left(\mathbf{E}\left[\frac{C_A - C_B}{C}\right] < 0\right) \rightarrow (\mathrm{B} \succ \mathrm{A})$$

C_i total clicks on results from i

C total clicks

Interleaving

- Examples of **metrics** used:
  - Relative difference in number of clicks received by the results from A and B
  - Ratio of the sessions with the results from B getting more clicks
- Allows to directly compare two rankings A & B. Deals with issues of position bias and user calibration.
- However, there are a number of issues:
  - **Reusability**: Can only elicit pairwise preferences for specific pairs of ranking functions
  - **Interpretation**: Doesn't tell us much about document-level assessments and user behavior

Describe the pros and cons of interleaving.

A/B vs. Interleaving

	A/B tests	**Interleaving**
Idea	Treat different users with different modifications of the search engine	Treat the same user with a combination of the results from both alternatives
Applicability	Very general (UI, ranking, new products, verticals, …)	Ranking only
Metrics used	Click-based, session-based, user-based, etc	Click-based only (somewhat restrictive)

So why do we need interleaving?

Online Evaluation Efficiency

- It turns out that:
  - *Interleaving is more **sensitive*** = evaluating the same change using interleaving requires 10x-100x times **less** data than the corresponding A/B test
  - It requires less data = allows us to use the resource of user sessions more **efficiently**

- Intuitive explanation:
  - In A/B tests, **different** users are treated with different systems
  - In interleaving, **the same user** compares the systems
    - ✓The **noise** due to user variance is removed

A/B vs Interleaving

	A/B tests	**Interleaving**
Idea	Treat different users with different modifications of the search engine	Treat the same user with a combination of the results from both alternatives
Applicability	Very general (UI, ranking, new products, verticals, …)	Ranking only
Metrics used	Click-based, session-based, user-based, etc	Click-based only (somewhat restrictive)
Efficiency	Not too efficient	**Very efficient**

What is the main advantage of using interleaving?

Why Efficiency is Important? (1)

- «*At Microsoft's Bing, the use of controlled experiments has **grown exponentially** over time, with over 200 concurrent experiments now running on any given day*» Kohavi et al., Online Controlled Experiments at Large Scale, KDD 2013

- Running 200 experiments:
  - 10% of the query traffic per experiment for two weeks = 5 experiments per week = 40 weeks*
  - 5% of the query traffic per experiment for two weeks = 10 experiments per week = 20 weeks*

* Only a motivational example: sometimes the same user might participate in several experiments at the same time + the number of the experiments reported by Bing might span several months.

Why Efficiency is Important? (2)

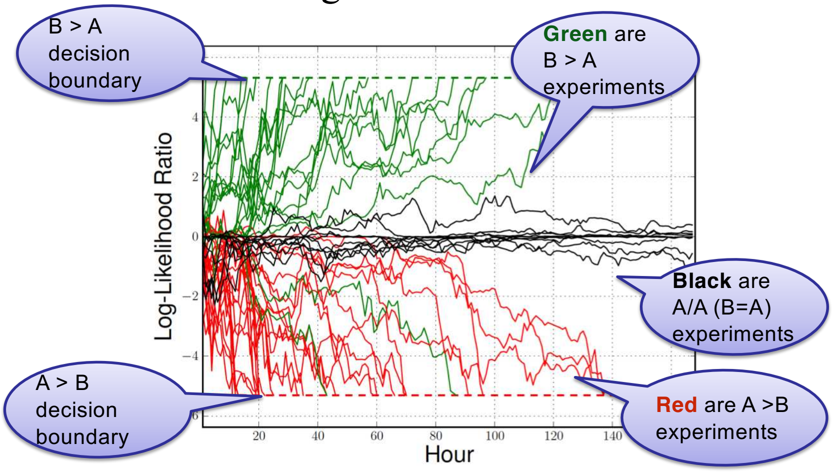
- Number of experiments grows
- Each experiment consumes some resources (user sessions)
- The duration of the experiments limits the **evolution** of the search engine
  - The faster a change is evaluated, the faster it can be deployed

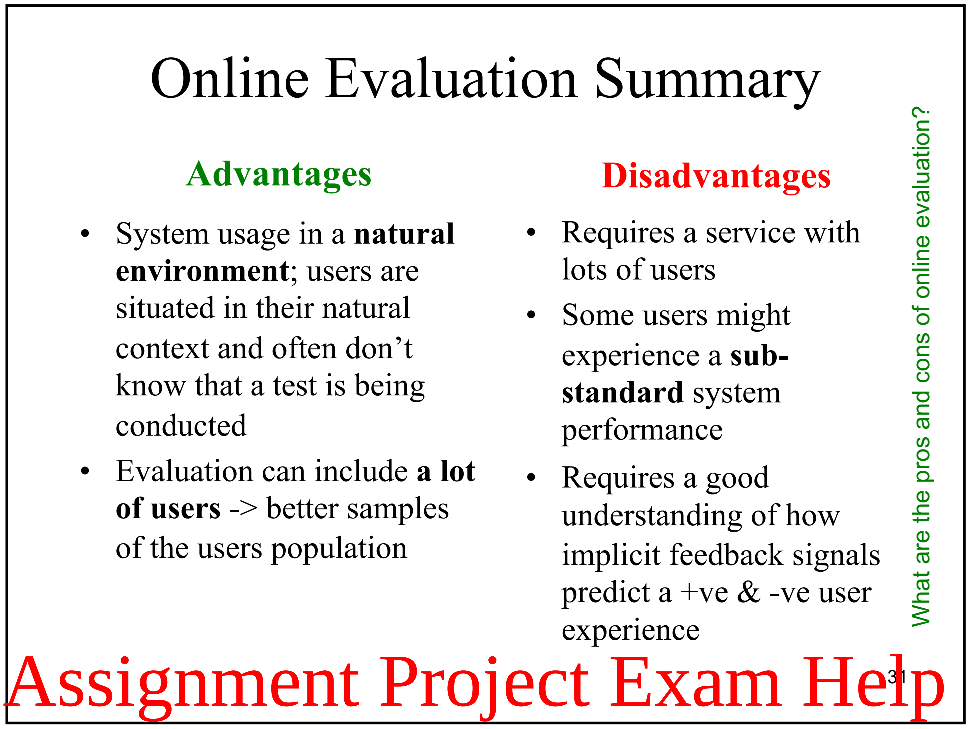
Why Efficiency is Important? (3)

- More than a half of the tested changes are either useless or harmful
- On average, the users who participate in an A/B or an interleaving experiment where the tested change B is worse than the production system A, suffer a somehow degraded experience
- Reducing the duration of the online experiments, i.e. increasing the online evaluation **efficiency** is important since:
  - SEs do not want to harm their users' experience
  - SEs want to evolve their system as fast as possible

Recent Research in Interleaving Efficiency (In Collaboration with Yandex, SIGIR 2015)

- Sequential testing: can we terminate a sequential test as soon as significance is reached?





Online Evaluation Summary

Advantages

- System usage in a **natural environment**; users are situated in their natural context and often don't know that a test is being conducted
- Evaluation can include **a lot of users** -> better samples of the users population

Disadvantages

- Requires a service with lots of users
- Some users might experience a **sub-standard** system performance
- Requires a good understanding of how implicit feedback signals predict a +ve & -ve user experience

What are the pros and cons of online evaluation?