

Behavior-Based Metrics

Praveen Chandar, Spotify

Explicit Feedback

- Ask users to learn about their experience with the system (surveys, ratings, etc.) or setup annotation tasks for humans to provide labels given data points.
- Limitations of explicit feedback
 - intrusive
 - expensive
 - might not reflect **true** user preference
 - might be infeasible for certain applications (e.g., personalization)

Implicit Feedback

- Infer user satisfaction from behavioral systems as they interact with the system.
- Advantages over explicit
 - collected in a natural setting
 - easy & inexpensive to collect
 - provides a direct measure of user preferences
- Limitation
 - can be noisy
 - biased

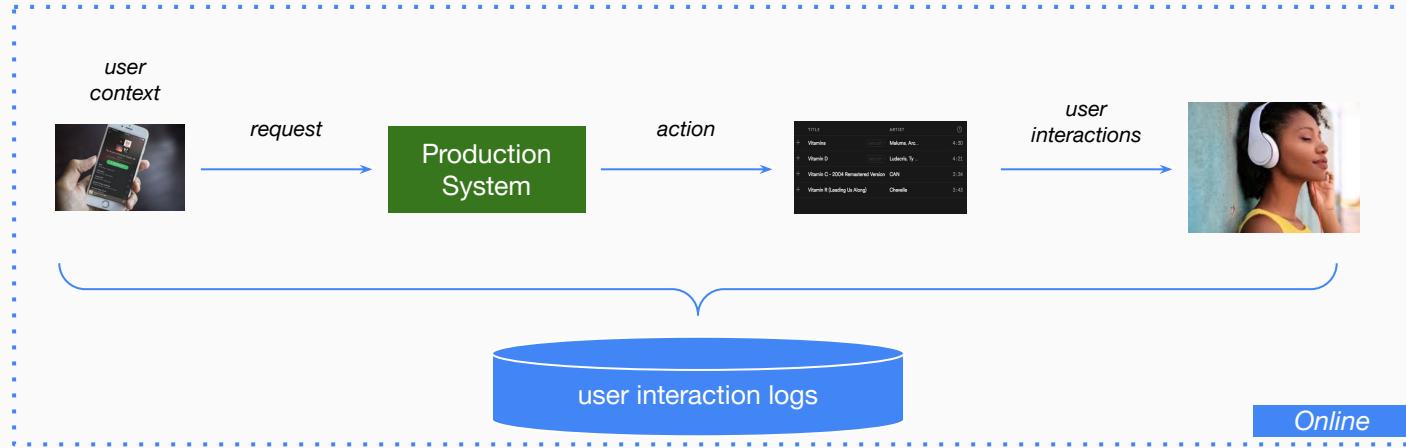
Explicit Feedback

- Ask users to learn about their experience with the system (surveys, ratings, etc.) or setup annotation tasks for humans to provide labels given data points.
- Limitations of explicit feedback
 - intrusive
 - expensive
 - might not reflect **true** user preference
 - might be infeasible for certain applications (e.g., personalization)

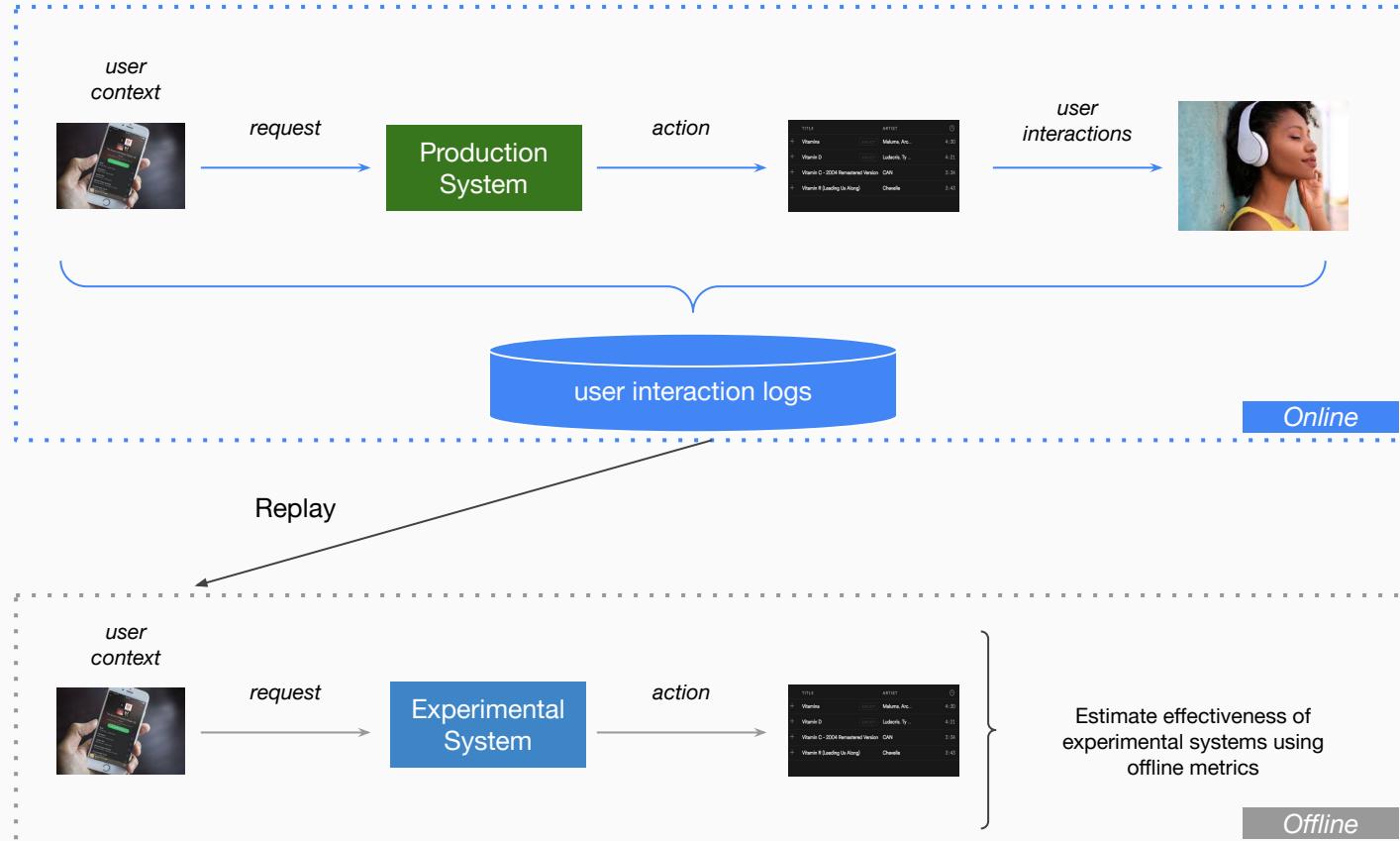
Implicit Feedback

- Infer user satisfaction from behavioral systems as they interact with the system.
- Advantages over explicit
 - collected in a natural setting
 - easy & inexpensive to collect
 - provides a direct measure of user preferences
- Limitation
 - can be noisy
 - biased

ML Evaluation Workflow



ML Evaluation Workflow



Categorizing Implicit Signals

short-term

Item Level Feedback



NeurIPS
NeurIPS Thirty-fourth Annual Conference on Neural Information Processing Systems. NeurIPS 2020 is a Virtually Only Conference. Sun Dec 6th through Sat the 12th. (Sunday is an industry expo)

en.wikipedia.org/w/index.php?title=Conference_on_Neural_Information_Processing_Systems&oldid=961111122

Conference on Neural Information Processing Systems ...

The Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS) is a machine learning and computational ...

History Topics · The NIPS experiment · Editors

Time Capsule

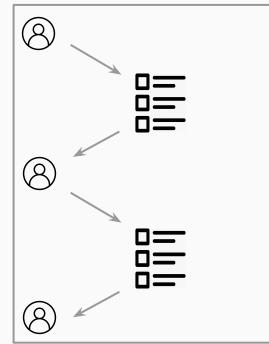
Your Running Mix 10:58PM

Page Level Feedback

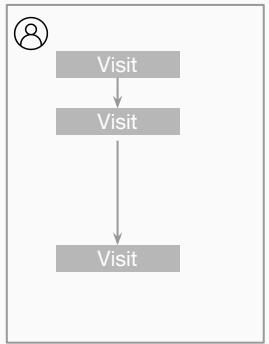


Google search results for "NeurIPS Conference". The top result is a link to the NeurIPS website, which shows the conference schedule from December 6 to 12, 2020, and various sessions like "Keynote", "Workshop", and "Poster". Below the search results are several thumbnail images of video clips or presentations.

Session Level Feedback



Intra-Session Feedback



long-term

Illustrative Examples

Google search results for "neurips conference".

Search bar: neurips conference

Navigation: All, News, Videos, Maps, Images, More, Settings, Tools

Results count: About 310,000 results (0.54 seconds)

Top result:

2020 Conference on Neural Information Processing Systems will begin on Sunday, December 6

and ends on Saturday, December 12

[Feedback](#)

nips.cc

2020 Conference

NeurIPS Thirty-fourth Annual Conference on Neural Information Processing Systems. NeurIPS 2020 is a Virtual-only Conference. Sun Dec 6th through Sat the 12th. (Sunday is an industry expo)

Call for Papers

We invite submissions for the Thirty-Fourth Annual ...

Dates

Login - Call for Workshops - Call For Socials - Future Meetings

2019

NeurIPS 2019 - Dates - Workshops - 2018 - Oral ...

Tutorials - ...

More results from nips.cc +

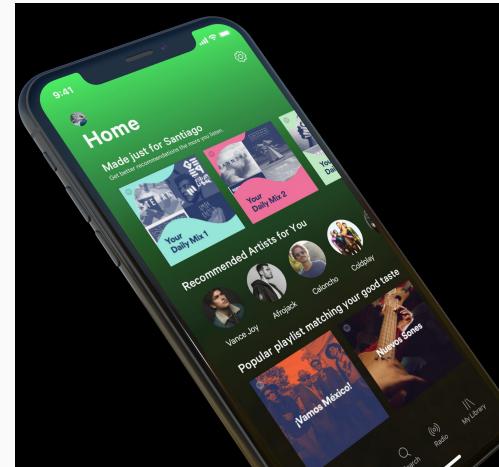
[en.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems](#)

Conference on Neural Information Processing Systems

The Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS) is a machine learning and computational ...

History · Topics · The NIPS experiment · Editions

Web Search



Music Recommendation

Item Level Feedback

short-term

Item Level Feedback



neurips.cc
NeurIPS
NeurIPS Thirty-fourth Annual Conference on Neural Information Processing Systems. NeurIPS 2020 is a Virtually Only Conference. Sun Dec 6th through Sat the 12th. (Sunday is an industry expo)

en.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems...
Conference on Neural Information Processing Systems ...

The Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS) is a machine learning and computational ...
History Topics · The NIPS experiment · Editors

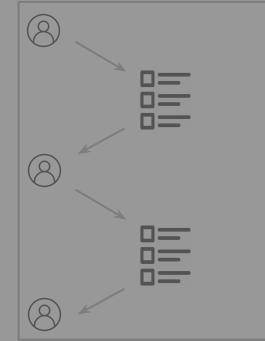


Page Level Feedback

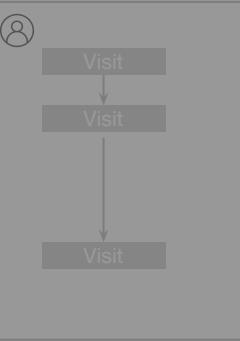


Google search results for 'NeurIPS Conference'. The top result is a link to the NeurIPS website, which shows the conference schedule from December 6 to 12, 2020, and various session details. Below the search results are several thumbnail images of conference papers or presentations.

Session Level Feedback



Intra-Session Feedback



long-term

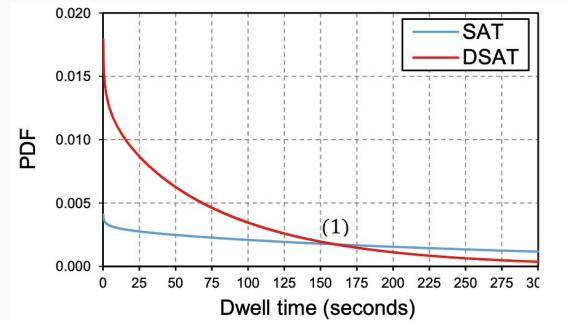
Behavioral Signals: Clicks

- User **clicks** are widely used behavioral signal for predicting satisfaction but are often noisy.
- Clicks are useful for learning personalized ML models in several applications.
- We **assume** that users are not randomly clicking on items, but make a informed choice.
- Clicks are considered independent of other user actions and items.

The image consists of two main parts. On the left is a screenshot of a Google search results page for the query "neurips conference". The top result is a snippet from nips.cc about the 2020 NeurIPS conference, which starts on Sunday, December 6, and ends on Saturday, December 12. To the right of this snippet is a detailed description of the conference, including its history (1987-present), abbreviation (NeurIPS), next date (Sun, Dec 6, 2020 - Sat, Dec 12, 2020), disciplines (Machine learning, Statistics, Artificial intelligence, Computational neuroscience), and related conferences like ICML, ICLR, and NeurIPS. On the right is a screenshot of a Spotify 'Editor's picks' page. It features a grid of 20 music tracks, each with a thumbnail image and a title. The tracks include "Today's Top Hits", "New Music Friday", "Mood Booster", "Happy Hits", "Workout", "Power Workout", "Chill Hits", "Rap Hits", "Mega Hit Mix", "Rap Workout", "Dinner with Friends", "Feelin' Myself", "Lounge - Soft & Mellow", "Get Turnt", "Feelin' Good", "Mega Hit Mix", "90s Rock Anthems", "Today's Top Hits", and "All Out Tops". A large blue hand icon is overlaid on the Spotify screenshot, pointing towards the top-left corner where the tracks are displayed.

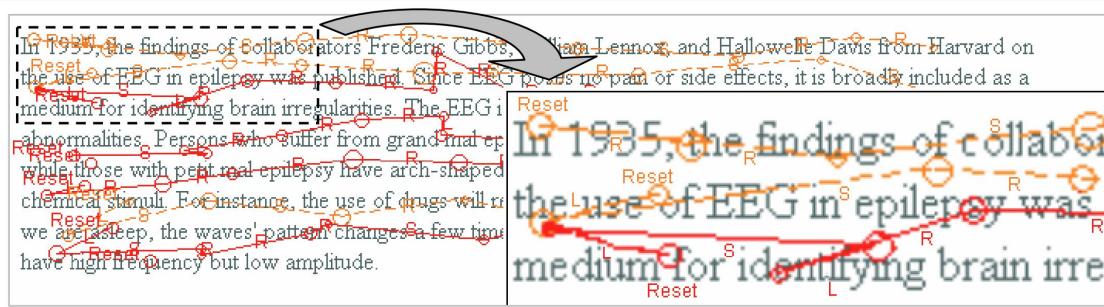
Behavioral Signals: Dwell Time

- **Dwell Time** is the time spent by the user on a page or item after clicking on it.
- Longer dwell time (30 secs or more in search applications) is one common approach to reduce noise in click signals.
- Empirical analyses have shown differences in dwell time distributions for satisfied & dissatisfied clicks.
- Dwelling patterns are influenced by attributes of an item. For instance, sophisticated content (high reading level) requires more time.



Behavioral Signals: User Eye Movements

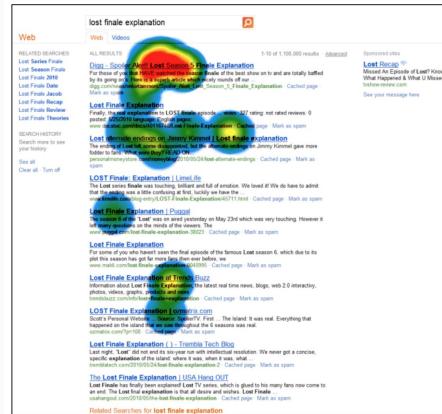
- > User eye movement exhibits unique behavior when reading. The eye fixation and saccade could be used to infer satisfaction.
- > **Fixation:** 200-250 ms steady gaze at one point.
- > **Saccade:** Rapid eye movement from one fixation to the next.



Behavioral Signals: Cursor Movements

- Collecting eye-tracking data can be intrusive and expensive.
- Studies have shown a correlation between **cursor** & **gaze** position.
- Signals from cursor movements can be combined to predict user satisfaction
 - **Click-through rate:** % of clicks when item is shown
 - **Hover rate:** % hover over items
 - **Unclicked hover:** Time of hover over item w/o click

Click positions

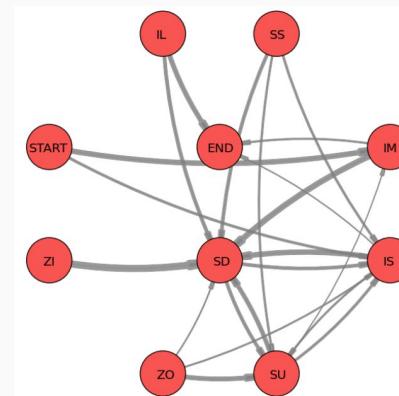


Cursor movement positions



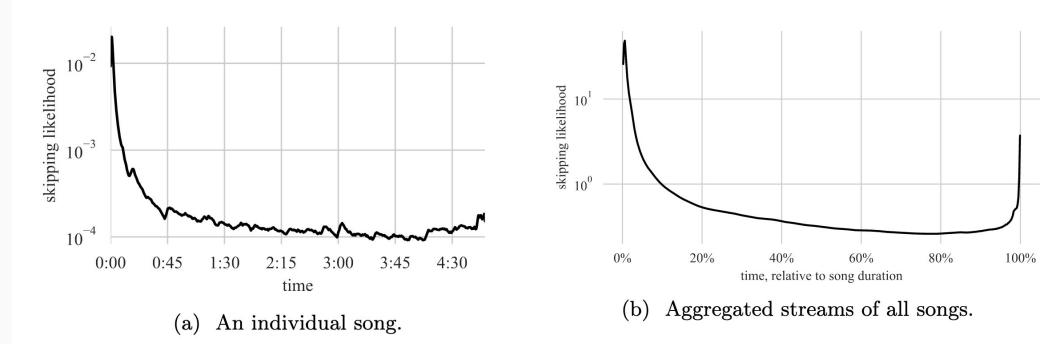
Behavioral Signals: Touch Gestures on Mobile

- Behavior modeling on mobile need to account for fine-grained user interactions such as
 - zooming in (ZI) and out (ZO)
 - swiping down (SD), up (SU), and horizontal (SS)
 - inactive short (IS), medium (IM), and long (IL)
- Markov state transitions can be used to predict likelihood of success for items (i.e., relevance of a page).



Behavioral Signals: Streams & Skip Behavior

- About 25% of streamed songs are skipped within first 5 secs and only about half of all songs are listened to in their entirety [[Lamere 2018](#)]
- There exists a correlations between skipping & musical structure and can be used to understand user satisfaction.



Behavioral Signals: Bookmarks, Saving & Shares

- Social interactions such as sharing with friends and followers could be used to measure satisfaction.
- Behavioral signals such as bookmarking, saving could measure satisfaction.

Google search results for "nips conference". The snippet shows the conference will begin on Sunday, December 6 and end on Saturday, December 12. Below the snippet is the official NIPS website with sections for Call for Papers, Registration, Future Meetings, Workshops, and more. A knowledge panel at the bottom right lists related conferences like ICML, ICLR, ECML, and NeurIPS.

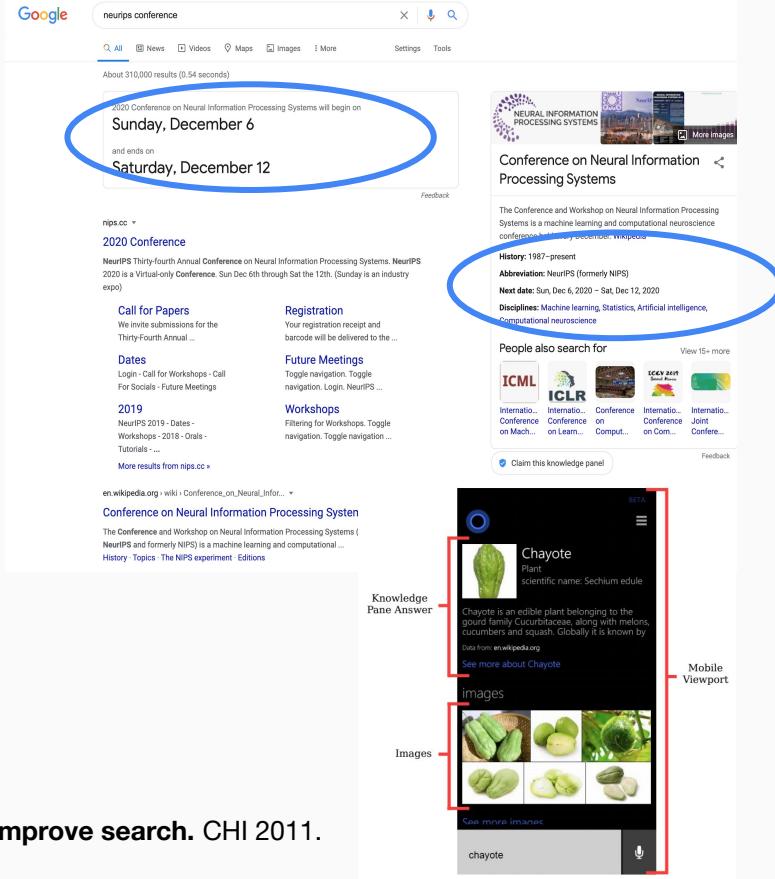
TITLE	ARTIST	ALBUM
See Fernando	Jenny Lewis	Acid Tongue
Plans	Oh Wonder	Oh Wonder
Stand Up	The Prodigy	Invaders Must Die
Retrospect	Sun Ra	A Fireside Chat with...
Satisfied	Andrew Bird	The Swimming Hour
Galician Chimpanzee I	Galegoz	Galician Chimpanzee
Go Or Go Ahead	Rufus Wainwright	Want One
Move Me	Sara Watkins	Young In All The Wr...
With Arms Outstretched	Rilo Kiley	The Execution Of Al...
Farewell Transmission	Songs: Ohia	Magnolia Electric C...
Montreal	Kaki King	Dreaming Of Revenge
Music Is My Hot, Hot Sex	CSS	Cansei De Ser Sexy
Steppin' Out	Joe Jackson	Night And Day
Get Out	CHVRCHES	Get Out
Over You	Flume, Seeka	The Late Ambiance ...

Jean Garcia-Gathright et al. **Understanding and Evaluating User Satisfaction with Music Discovery**. SIGIR 2018.

Anlei Dong et al. **Time is of the essence: improving recency ranking using Twitter data**. WWW 2010.

Good Abandonment

- Scenarios in which user does not interact with the results for a given request but are satisfied is referred to as ***good abandonment***.
- In contrast ***bad abandonment*** happens when user needs are not satisfied and results in no interaction with the results.
- Clicks and dwell time signals alone are often insufficient to distinguish between the two. The following signal have been used to predict ***good abandonment*** in mobile and desktop:
 - properties of the request
 - user session
 - gaze and viewport tracking



Jeff Huang et al. No clicks, no problem: using cursor movements to understand and improve search. CHI 2011.

Kyle Williams et al. Detecting Good Abandonment in Mobile Search. WWW 2016.

Summary: Item Level Feedback

Behavioral Signals

- Clicks
- Gestures on Mobile
- Dwell Time
- Streams & Skip Behavior
- Eye-Tracking
- Bookmarks, Saving & Shares
- Cursor Movements

User Intent / Goals

Context

Heterogeneity
of Items

User

Summary: Item Level Feedback

Behavioral Signals

- Clicks
- Gestures on Mobile
- Dwell Time
- Streams & Skip Behavior
- Eye-Tracking
- Bookmarks, Saving & Shares
- Cursor Movements

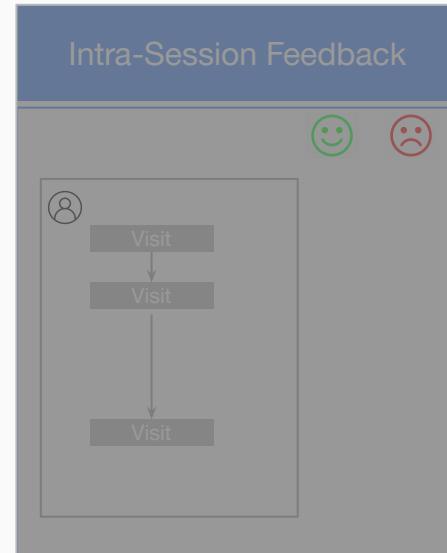
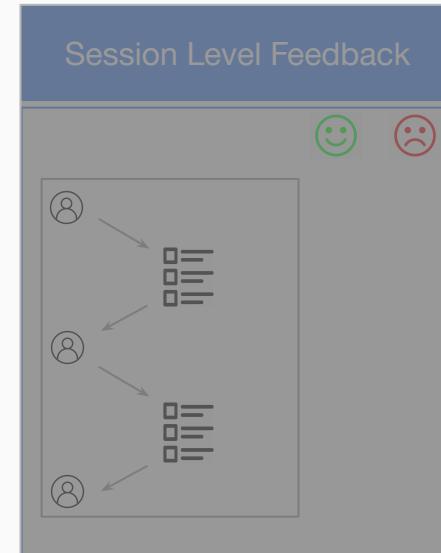
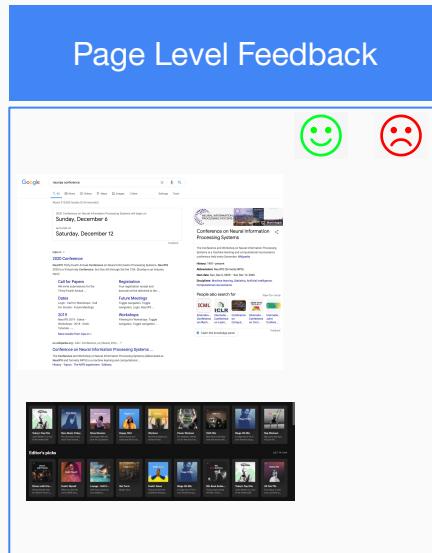
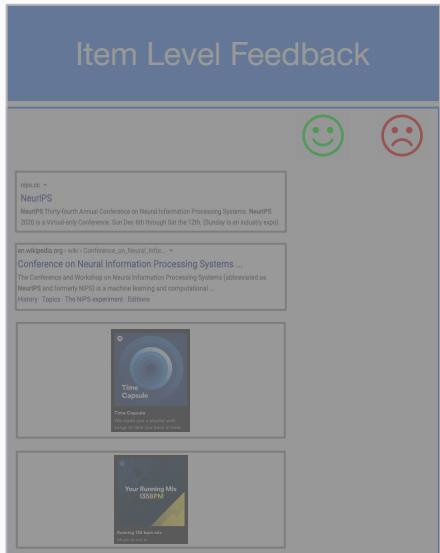


*Offline
metric
validation*

- AUC
- Precision
- Akaike Information criterion (AIC)
- Bayesian information criterion (BIC)

Page Level Feedback

short-term



long-term

Whole Page Feedback

$$A \equiv \left\{ \begin{array}{l} \\ \\ \\ \end{array} \right.$$

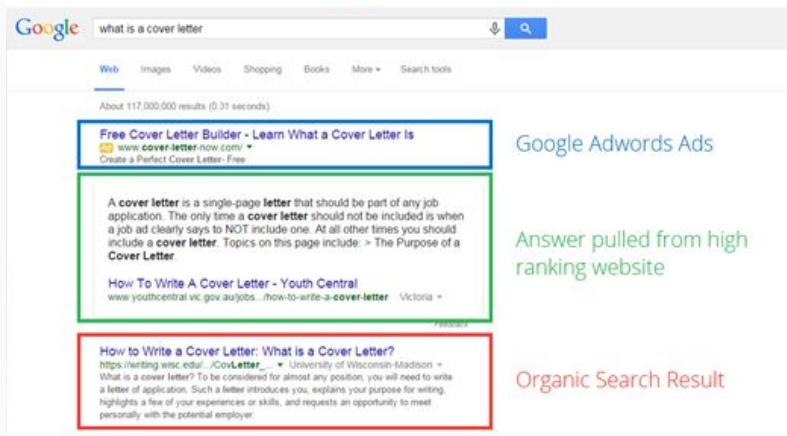
$$reward = f(A)$$

- In many application, a set of items are presented to the user for a given request.
- Whole-page feedback defines how well items presented on a page and its attributes satisfies the user request.

Action → entire page

Labels → success for the entire page

Whole Page Feedback: Examples



Google search results for "what is a cover letter". The results include a Adwords Ad, an organic search result from Youth Central, and another organic search result from University of Wisconsin-Madison.

Google Adwords Ads

Answer pulled from high ranking website

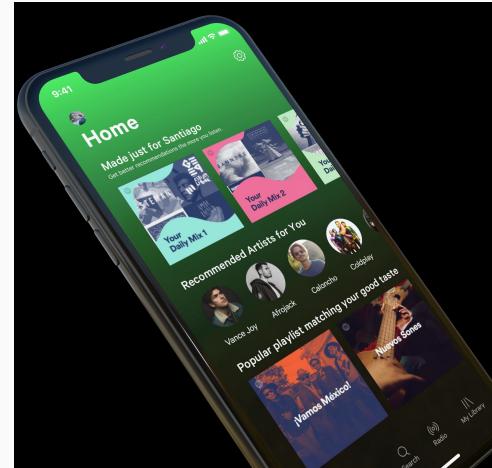
Organic Search Result

Free Cover Letter Builder - Learn What a Cover Letter Is
www.cover-letter-now.com/ • Create a Perfect Cover Letter - Free

A **cover letter** is a single-page **letter** that should be part of any job application. The only time a **cover letter** should not be included is when a job ad clearly says to NOT include one. At all other times you should include a **cover letter**. Topics on this page include: > The Purpose of a **Cover Letter**

How To Write A Cover Letter - Youth Central
www.youthcentral.vic.gov.au/jobs.../how-to-write-a-cover-letter Victoria •

How to Write a Cover Letter: What is a Cover Letter?
<https://writing.wisc.edu/.../CoverLetter...> ▾ University of Wisconsin-Madison ▾ What is a cover letter? To be considered for almost any position, you will need to write a letter of application. Such a letter introduces you, explains your purpose for writing, highlights a few of your experiences or skills, and requests an opportunity to meet personally with the potential employer.



User Models for Whole Page Evaluation

- While assigning labels/reward for the result page provides a holistic view of success, in several scenarios they are computed from rewards observed at the item-level.
 - Action* → decompose action into sub-actions
 - Labels* → aggregation of sub-action rewards
- However, the observed rewards at the item level could be biased due to various reasons relating to the characteristics of the page and ignoring them would lead to unreliable estimates.
- This calls for the development of user behavioral models to accurately estimate a page-level reward

$$\textcircled{A} \equiv \left\{ \begin{array}{l} \\ \end{array} \right. \quad \text{reward} = f(A)$$

$$\textcircled{A} \equiv \left\{ \begin{array}{l} \textcircled{A}_1 \\ \textcircled{A}_2 \\ \textcircled{A}_3 \\ \dots \\ \textcircled{A}_k \end{array} \right\} \quad \text{reward} = f(A_1, A_2, \dots, A_k)$$

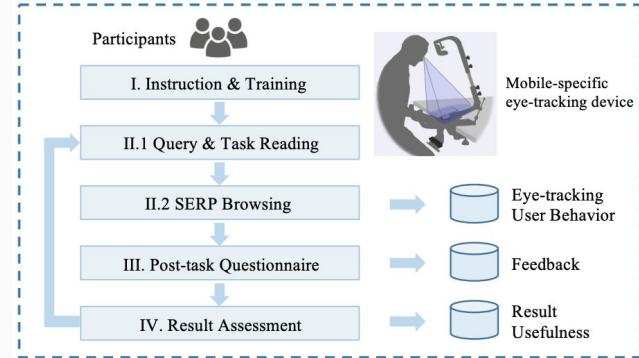
User Browsing Models: Position-Based Model

How do users interact with the list of ranked results of search engines?

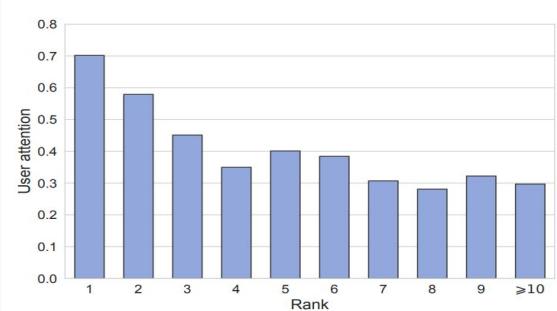
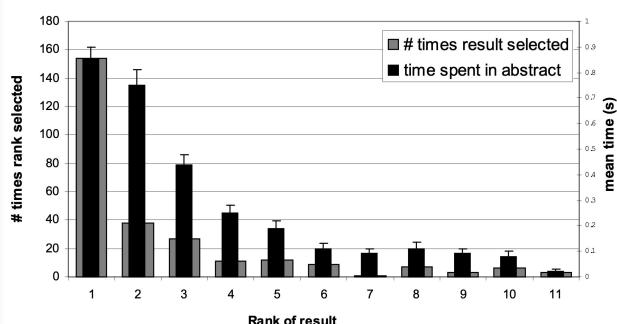
Do they read the abstracts sequentially from top to bottom, or do they skip links?

How many of the results do users evaluate before clicking on a link or reformulating the search?

- Granka et al. 2004



User Attention per rank on *Desktop* vs. *Mobile*

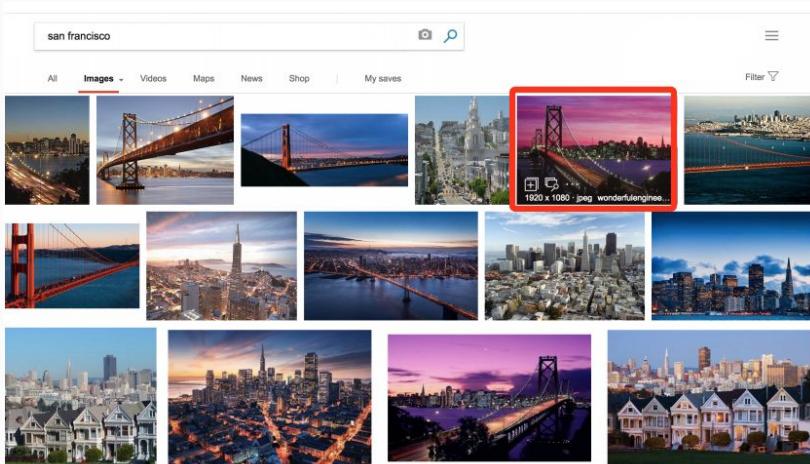


Laura A. Granka et al. Eye-tracking analysis of user behavior in WWW search. SIGIR 2004.

Yukun Zheng et al. Investigating Examination Behavior in Mobile Search. WSDM 2020.

Grid-based Model

- In movie recommendations, image search, etc., the results are presented in a grid view, and users examine them both vertically and horizontally.

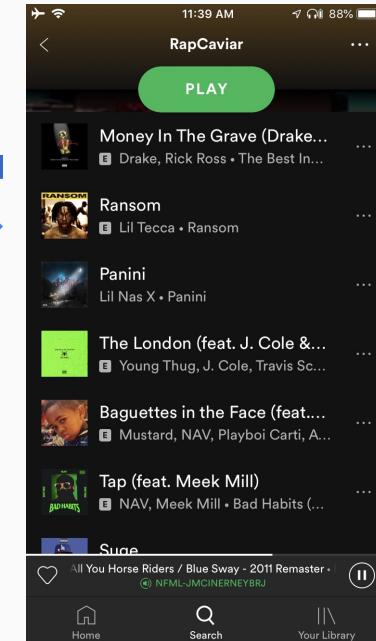
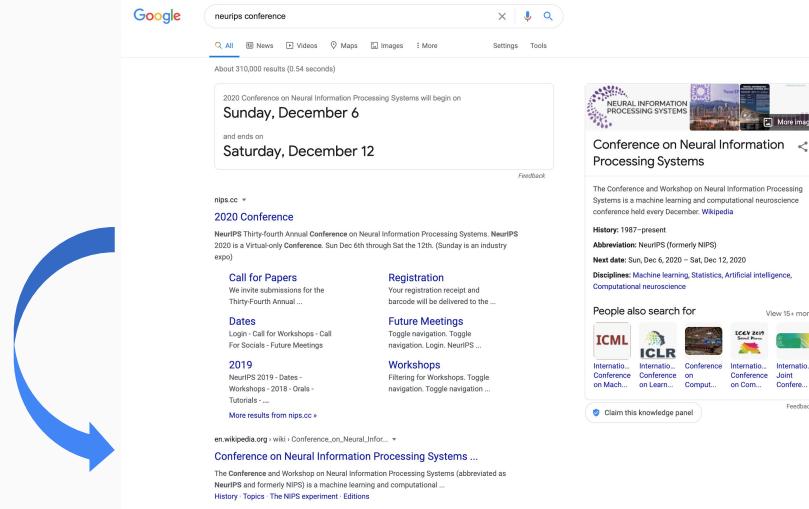


middle positions may attract more attention depending on the domain in grid-based layouts.



User Browsing Models: Cascade Model

- Likelihood of a user examining a document at rank i depends on how satisfied they were with previously observed documents.
- Users' interaction with a track in a playlist could be influenced by previous track(s).



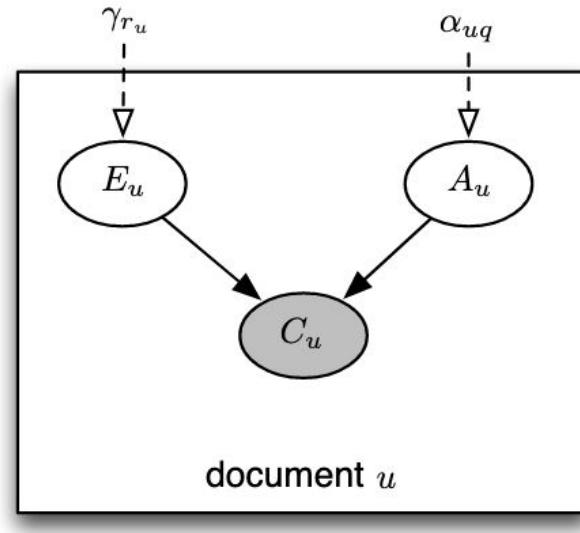
Click Models: Position-Based Model

- Click models is a way to represent user clicks that makes it easy to quantify user behavior. Many existing click models use probabilistic graphical models.
- A user examining an item at position r does not depend on examinations and clicks above r .
- Clicks depend on probability of examination & user-item attractiveness or ***utility***.

$$P(E_{r_u} = 1) = \gamma_{r_u}$$

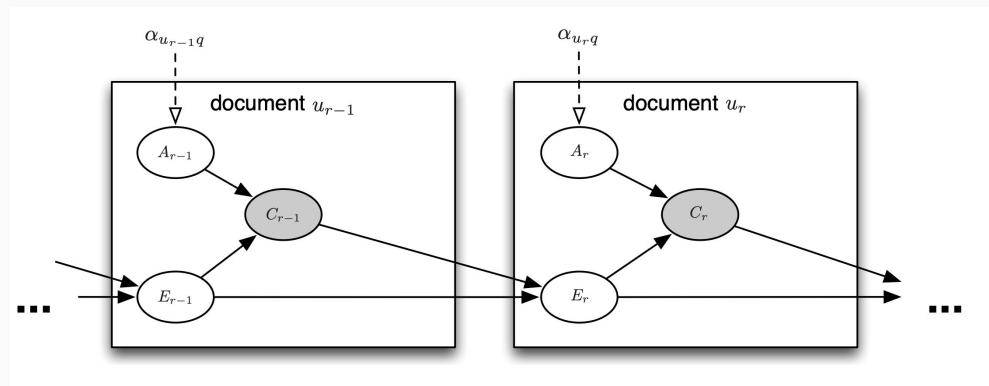
$$P(A_u = 1) = \alpha_{uq}$$

$$P(C_u = 1) = P(E_{r_u} = 1) \cdot P(A_u = 1)$$

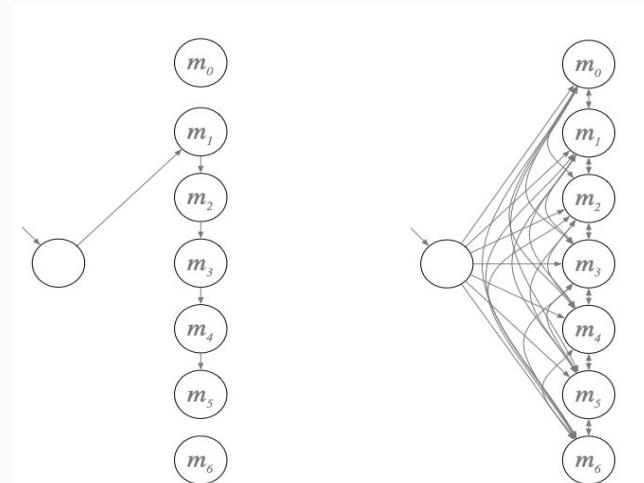
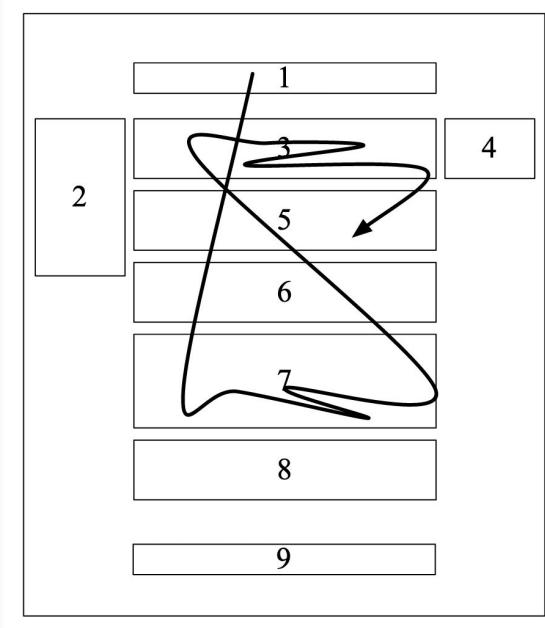


Click Models: Cascade Model

- Cascade dependency between clicks and previously examined items



Browsing Models: Non-Sequential Examination Model



Estimating Position Bias from Log Data

- **Position Bias** Higher ranked items are more likely to be examined resulting in higher interaction.

- Interventions to address position bias
 - Randomize items (if policy is not stochastic)
 - Aggregate interaction per position
 - Infer propensity of observation conditional on position
- Randomization Techniques
 - *Randomize Top K*
 - *Swap Method*
Swap items from random position with pivot rank.
 - *Intervention Data*
Exploit the randomness in historical intervention data (i.e., A/B tests).

Assumption

Probability of examination is dependent only on the rank of the item.

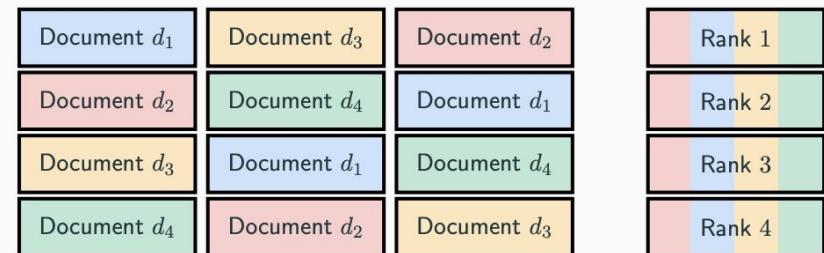


Figure Source: [Unbiased Learning to Rank: Counterfactual and Online Approaches](#)

Lihong Li et al. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. WSDM 2011

Xuanhui Wang et al. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. WSDM 2018.

Aman Agarwal et al. Estimating Position Bias without Intrusive Interventions. WSDM 2019.

Counterfactual Evaluation from Log Data

System $\pi_{production}$	
rank	
1	Item 1
2	Item 2
3	Item 3
4	Item 4
5	Item 5

System $\pi_{experimental}$	
rank	
1	Item 5
2	Item 4
3	Item 3
4	Item 2
5	Item 1

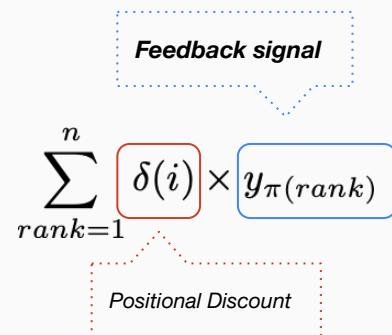
Counterfactual Evaluation from Log Data

System $\pi_{production}$	
rank	
1	Item 1
2	Item 2
3	Item 3
4	Item 4
5	Item 5

$$\sum_{rank=1}^n \delta(i) \times y_{\pi(rank)}$$

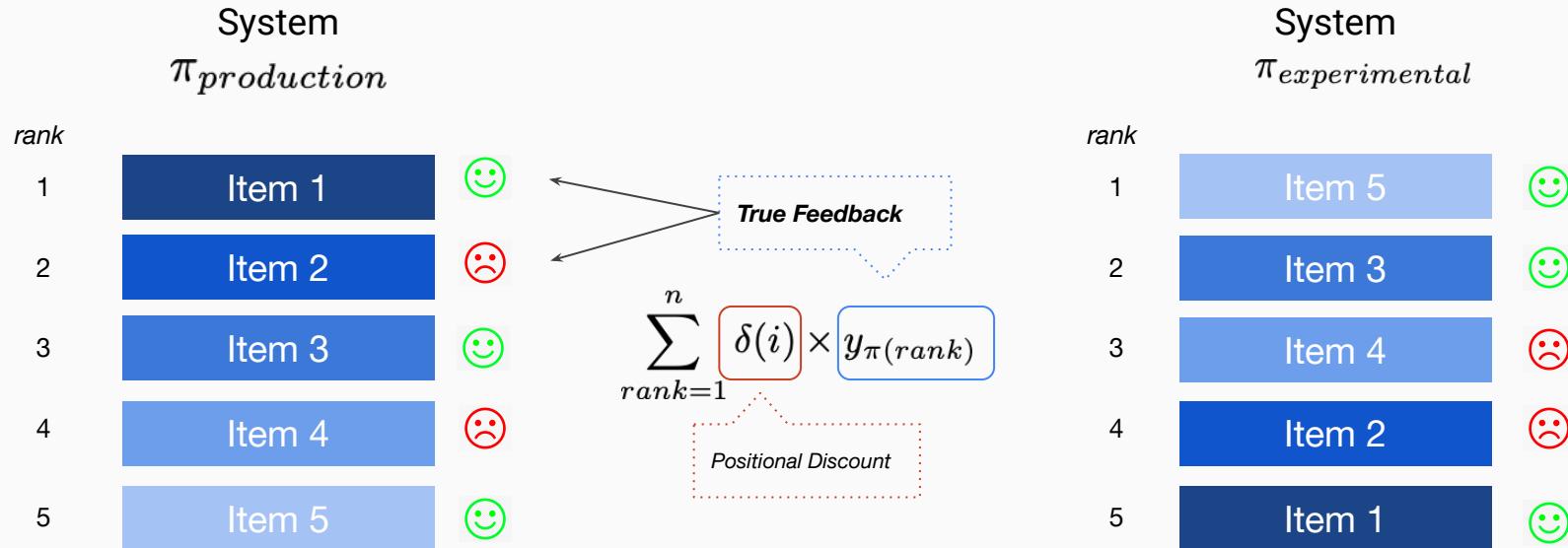
Feedback signal

Positional Discount

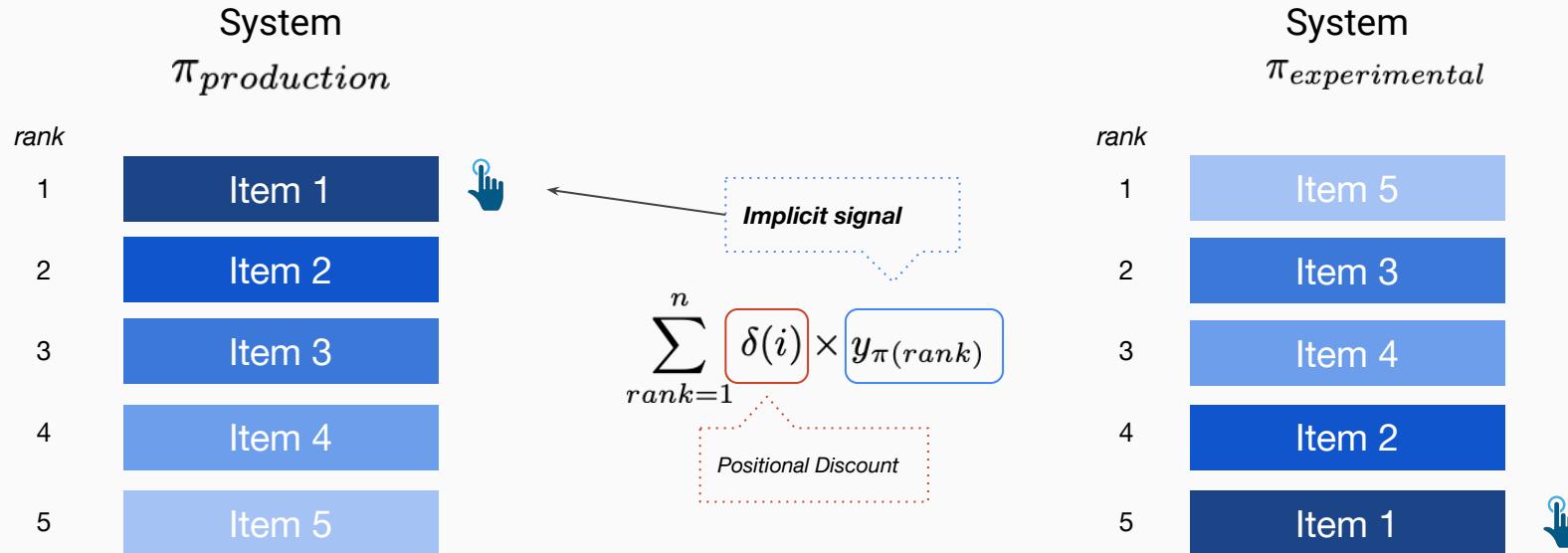


System $\pi_{experimental}$	
rank	
1	Item 5
2	Item 4
3	Item 3
4	Item 2
5	Item 1

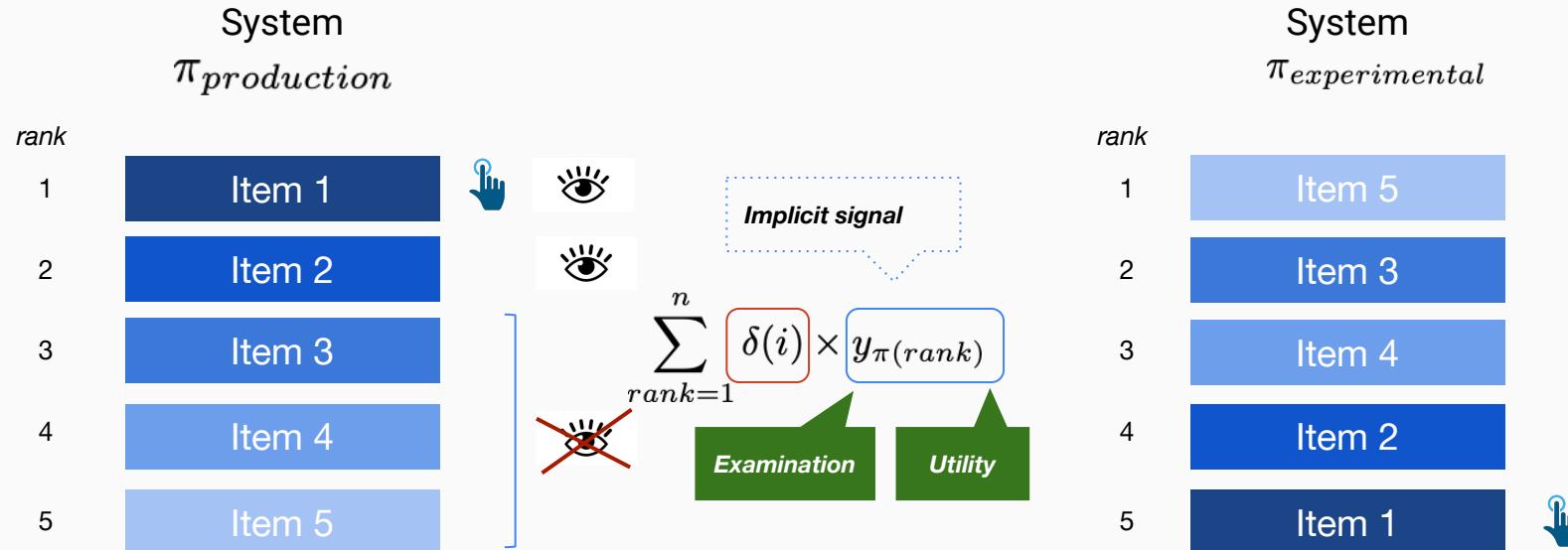
Counterfactual Evaluation from Log Data



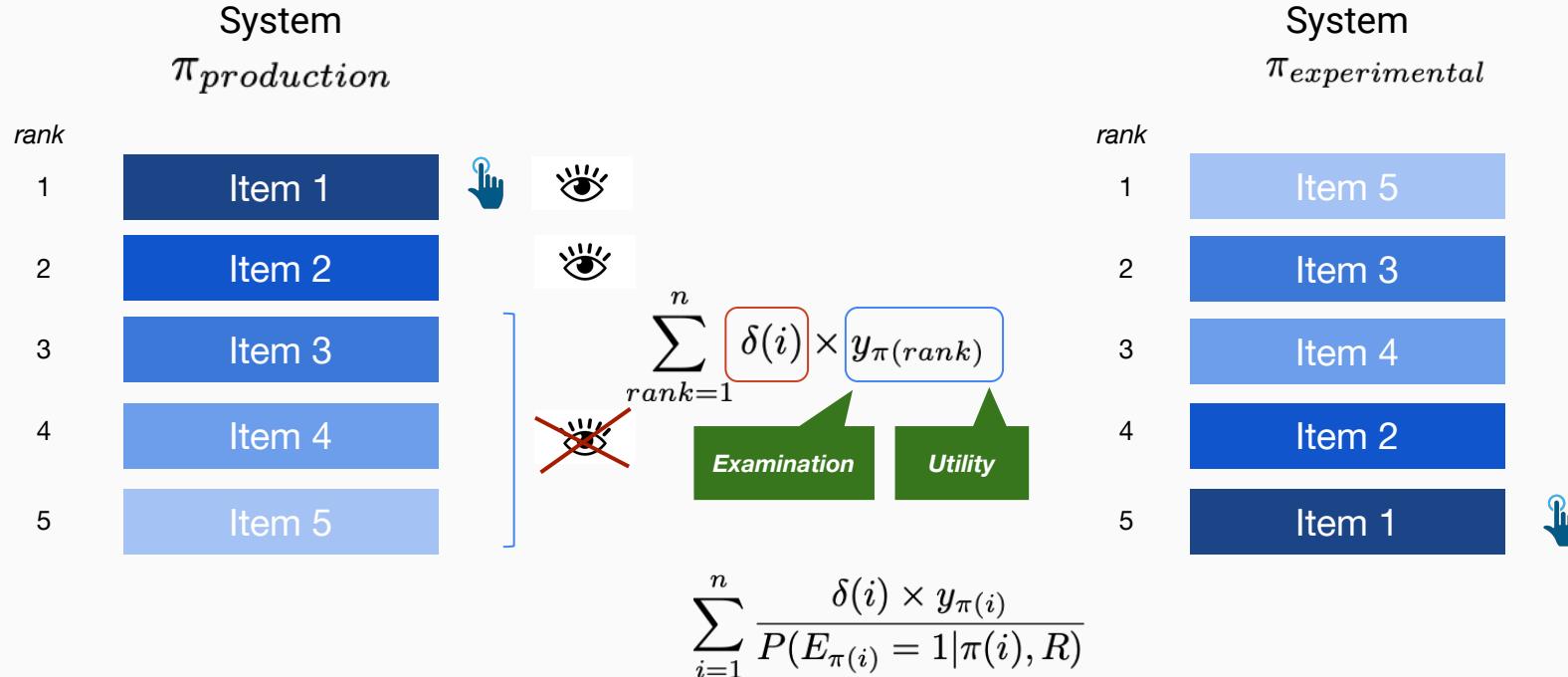
Counterfactual Evaluation from Log Data



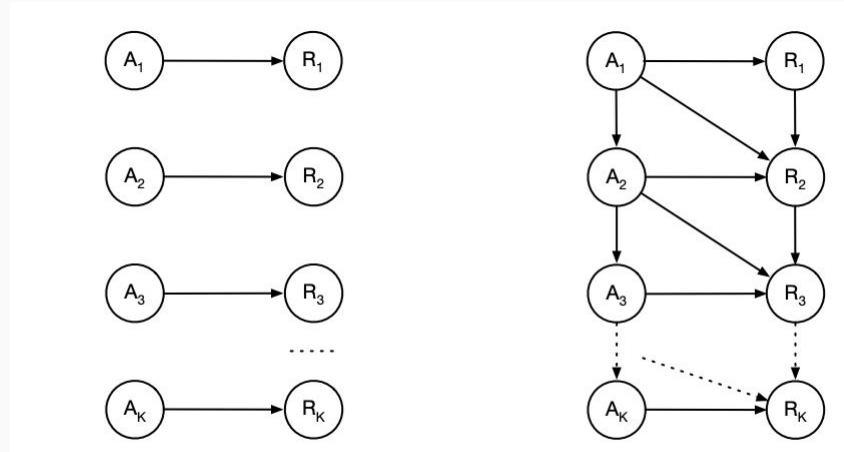
Counterfactual Evaluation from Log Data



Counterfactual Evaluation from Log Data



Counterfactual Slate Evaluation



Linear Additive Rewards
when sub-action rewards
are not available.

Adith Swaminathan et al.
Off-policy evaluation for slate recommendation.
NeurIPS 2017.

Independence
Assumption

Shaui Li et al. **Offline evaluation of ranking policies with click models.** KDD 2018.

Sequential Reward
Dependence Assumption

James McInerney et al.
Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions
KDD 2020.

Counterfactual Evaluation: Grid-based & Beyond

- *Ruocheng Guo et al.* **Debiasing Grid-based Product Search in E-commerce.** KDD 2020.
- *Jiawei Chen et al.* **Bias and Debias in Recommender System: A Survey and Future Directions.** TKDE 2020

Novelty & Diversity in Rankings

Q grand canyon things to do

Things To Do - South Rim - GrandCanyon.com
<https://grandcanyon.com/planning/south-rim-planning.../things-to-do-south-rim/> ▾
Here you will find a list of attractions, tours and other things to do while visiting the South Rim of the Grand Canyon.

Plan Your Visit - Grand Canyon National Park (U.S. National Park ...
<https://www.nps.gov/grca/planyourvisit/index.htm> ▾
Jun 6, 2017 - Most visitors (90%) see Grand Canyon from the "South Rim" from overlooks accessed ... view from patio of grand canyon lodge on the north rim ...

Grand Canyon North Rim
<https://grandcanyon.com/category/planning/north-rim-planning/> ▾
Grand Canyon North Rim is visited seasonally. More remote than the South Rim, it offers views that can't be beat. Open May 15 - October 15 Annually.
Things To Do - North Rim - Where is Navajo Bridge?

Grand Canyon Skywalk at Grand Canyon West
<https://grandcanyon.com/planning/west.../grand-canyon-skywalk-at-grand-canyon-we...> ▾
Grand Canyon Skywalk - You're standing on a platform made of glass looking out over the Grand Canyon. Eagle Point in front and Colorado River to your left.

Havasupai Falls Arizona - Grand Canyon
<https://grandcanyon.com/planning/south-rim-planning.../havasupai-falls-arizona/> ▾
Havasupai Falls Arizona is a major destination for hikers who want to visit the blue green waterfalls. Hidden in the Grand Canyon, and difficult to get reservations ...

A variation of rank-based metric that rewards novelty and penalizes diversity



Grand Canyon Rim Grand Canyon Skywalk Havasu Falls



$$\sum_{a \in \mathcal{A}_q} p(a|q) \times \text{metric}(y^a, \pi)$$

\mathcal{A}_q aspects for query q

y^a document relevance to aspect a

Novelty & Diversity in Rankings

Q grand canyon things to do

Things To Do - South Rim - GrandCanyon.com
<https://grandcanyon.com/planning/south-rim-planning.../things-to-do-south-rim/> ▾
Here you will find a list of attractions, tours and other things to do while visiting the South Rim of the Grand Canyon.

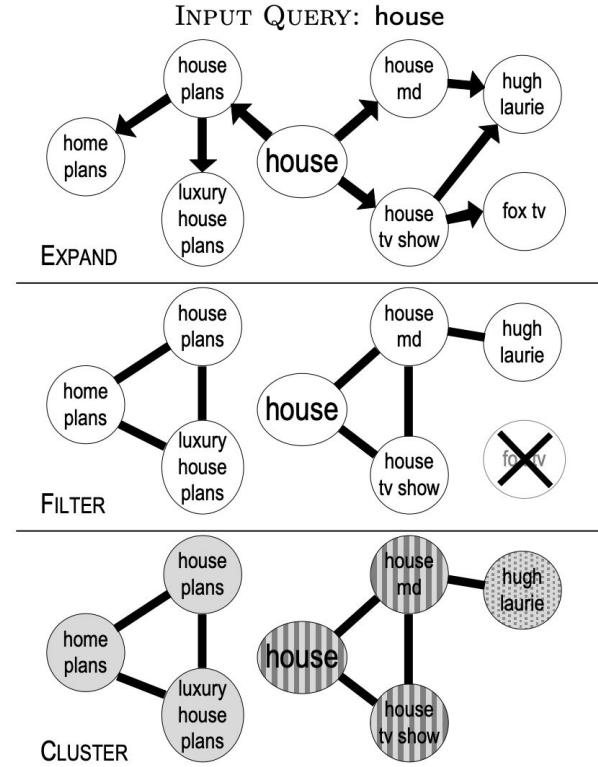
Plan Your Visit - Grand Canyon National Park (U.S. National Park ...
<https://www.nps.gov/grca/planyourvisit/index.htm> ▾
Jun 6, 2017 - Most visitors (90%) see Grand Canyon from the "South Rim" from overlooks accessed ... view from patio of grand canyon lodge on the north rim ...

Grand Canyon North Rim
<https://grandcanyon.com/category/planning/north-rim-planning/> ▾
Grand Canyon North Rim is visited seasonally. More remote than the South Rim, it offers views that can't be beat. Open May 15 - October 15 Annually.
Things To Do - North Rim - Where is Navajo Bridge?

Grand Canyon Skywalk at Grand Canyon West
<https://grandcanyon.com/planning/west.../grand-canyon-skywalk-at-grand-canyon-we...> ▾
Grand Canyon Skywalk - You're standing on a platform made of glass looking out over the Grand Canyon. Eagle Point in front and Colorado River to your left.

Havasupai Falls Arizona - Grand Canyon
<https://grandcanyon.com/planning/south-rim-planning.../havasupai-falls-arizona/> ▾
Havasupai Falls Arizona is a major destination for hikers who want to visit the blue green waterfalls. Hidden in the Grand Canyon, and difficult to get reservations ...

A variation of rank-based metric that rewards novelty and penalizes diversity



Session Level Feedback

short-term



ripi.cc
NeurIPS

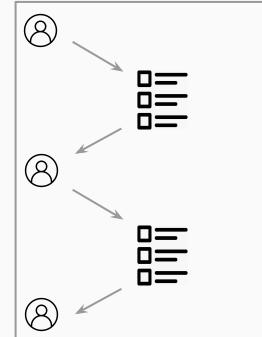
en.wikipedia.org › wiki › Conference_on_Neural_Information_Processing_Systems ...
Conference on Neural Information Processing Systems ...
The Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS) is a machine learning and computational ...
History Topics The NIPS experiment Editions

A screenshot of a mobile application titled "Time Capsule". The screen features a large blue circular graphic in the center. Below the graphic, the word "Time Capsule" is written in white, sans-serif font. At the bottom of the screen, there is a thin horizontal bar containing the text "We made you a play list with..." followed by three small, illegible icons.



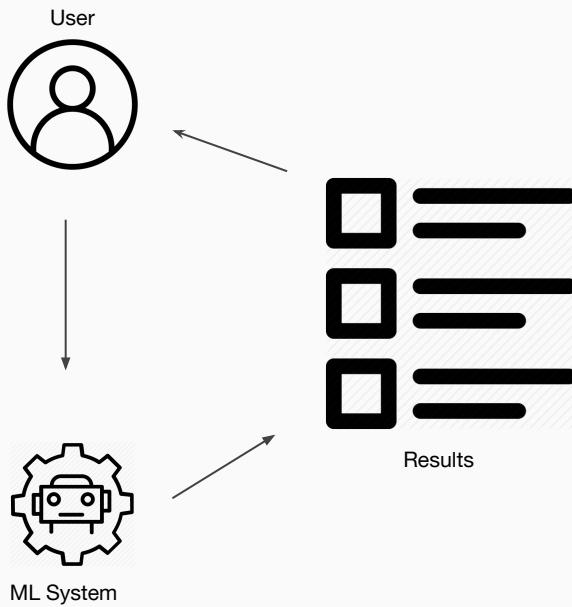
A screenshot of the National Geographic website's 'Top 100' page. The page features a large, bold title 'TOP 100' at the top left. Below it is a sub-section titled 'The Best Destinations in the World'. A prominent heading reads 'Discover the most inspiring travel destinations from around the globe'. The main content is a grid of 100 small thumbnail images, each representing a different travel destination. The thumbnails are arranged in a grid format, with some rows having more columns than others. Each thumbnail includes a small caption below it, such as 'Santorini, Greece' or 'Tulum, Mexico'. The overall layout is clean and modern, with a white background and a mix of black and blue text.

Session Level Feedback

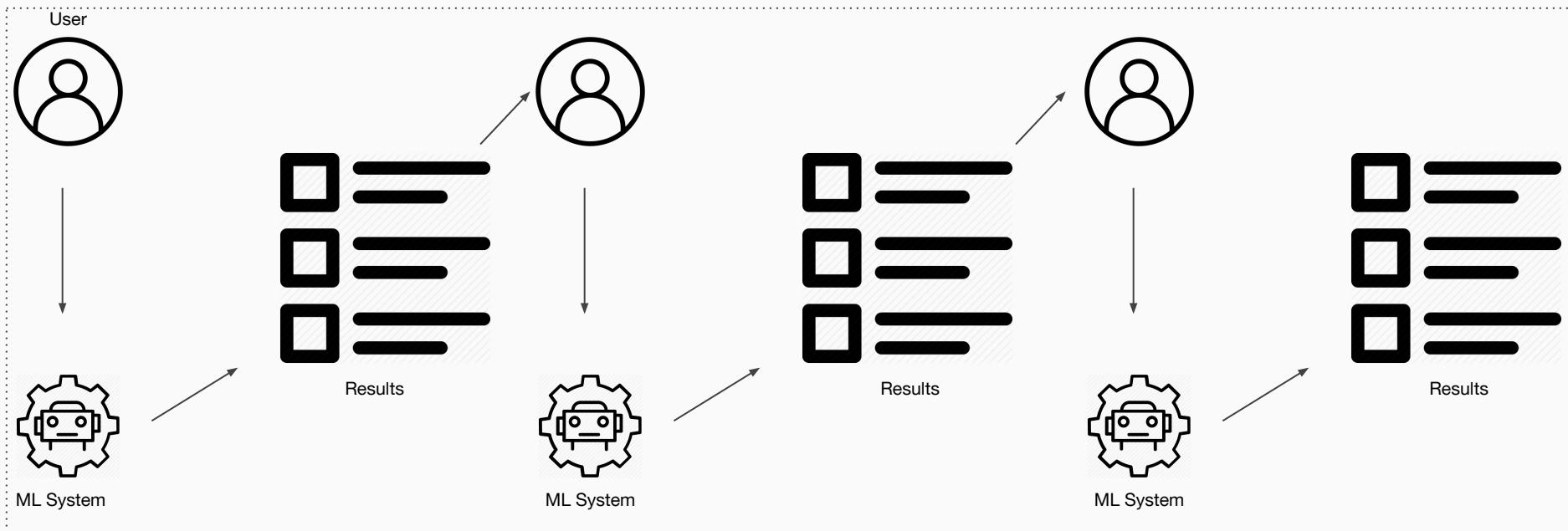


long-term

Session Level Evaluation: Introduction

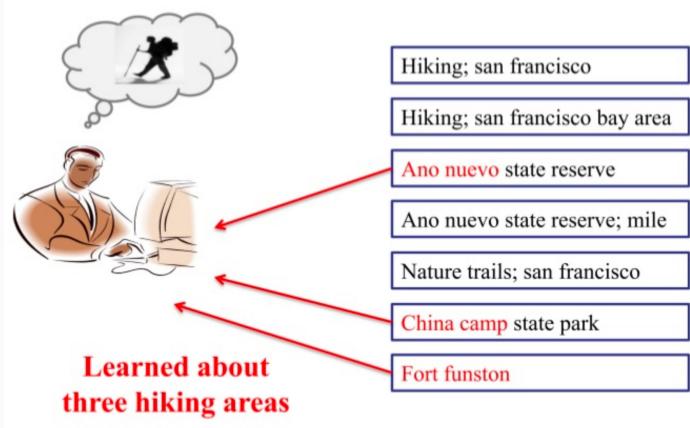


Session Level Evaluation: Introduction

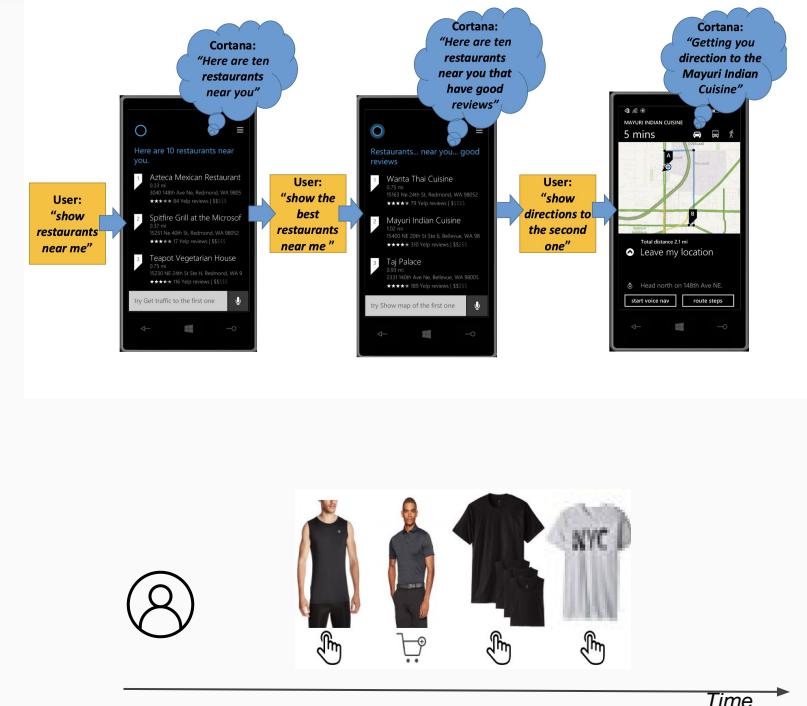
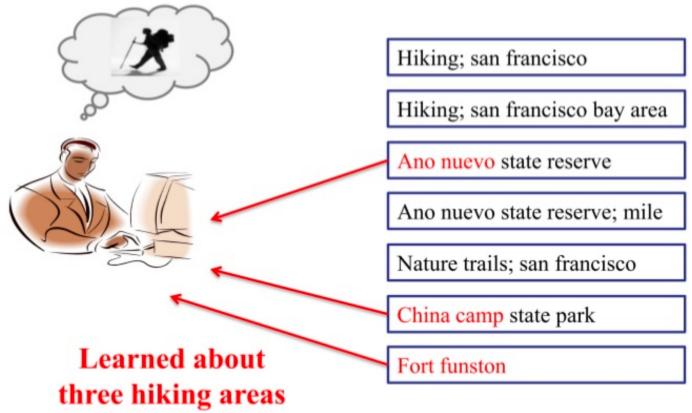


How do we evaluate the user experience over a session?

Session Level Evaluation: Examples



Session Level Evaluation: Examples



Rishabh Mehrotra et al. **Understanding & Inferring User Tasks and Need**. WWW 2018.

Shuo Zhang and Krisztian Balog. **Evaluating Conversational Recommender Systems via User Simulation**. KDD 2020.

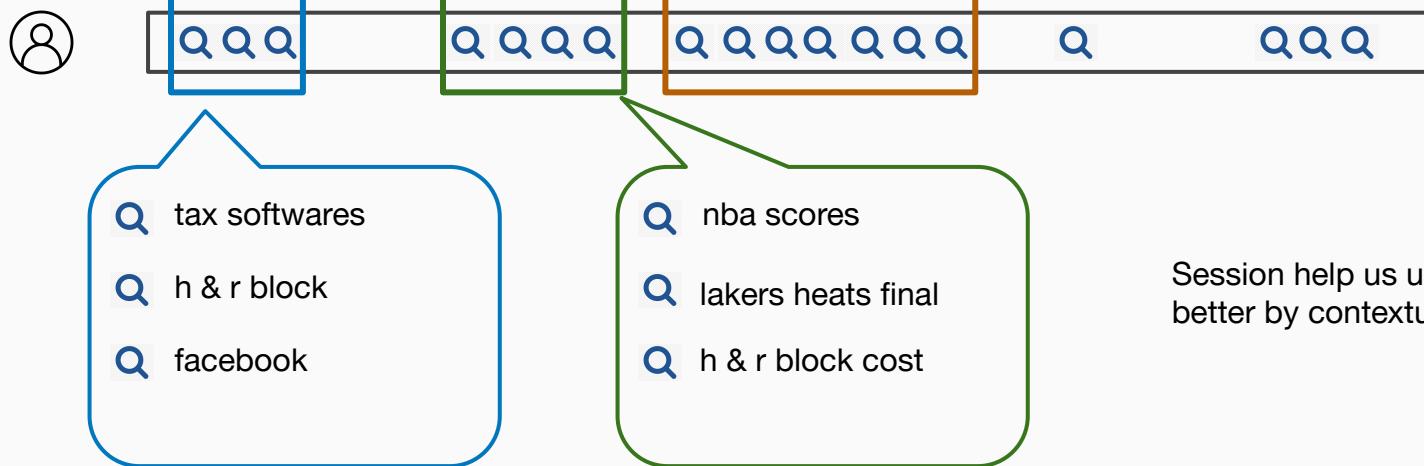
Malte Ludewig and Dietmar Jannach. **Evaluation of session-based recommendation algorithms**. User Modeling and User Adapted Interaction 2018.

Identifying Sessions from Logs

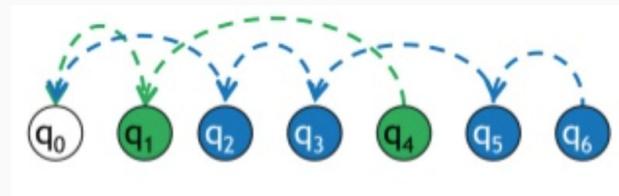


time between requests
could be reasonable
proxies to identify sessions

Identifying Sessions from Logs



Session help us understand user goals better by contextualizing their requests.



Sessions → User Goals

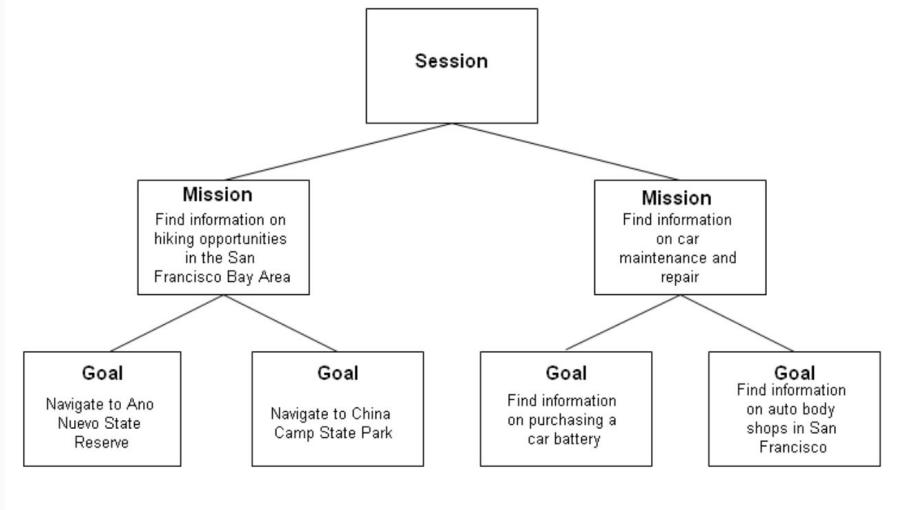
User Sessions

All user activity within a fixed time window.

User Goals

Atomic piece of user activity accomplished by issuing one or more requests to the system.

Web Search Example



Identifying User Goals

QUERY and TIMESTAMP	GOAL #	MISSION #	DESCRIPTION
hiking; san francisco Tue Apr 17 23:43:17 2007 (4m 17s)	1	1	MISSION 1: Find info on hiking opportunities in and around San Francisco
hiking; san francisco bay area Tue Apr 17 23:47:34 2007 (4m 59s)	1	1	GOAL 1: Find info on hiking trails in San Francisco and the Bay Area
ano nuevo state reserve Tue Apr 17 23:52:33 2007 (7m 54s)	2	1	GOAL 2: Navigate to Ano Nuevo State Reserve and find out about distances
ano nuevo state reserve; miles Wed Apr 18 00:00:27 2007 (3m 34s)	2	1	
nature trails; san francisco Wed Apr 18 00:04:01 2007 (16m 15s)	1	1	
lobos creek trail; Wed Apr 18 00:20:16 2007 (0m 3s)	3	1	GOAL 3: Navigate to Lobos Creek Trail
china camp state park; san rafael Wed Apr 18 00:20:19 2007 (2m 35s)	4	1	GOAL 4: Navigate to China Camp, San Rafael and find out about distances
china camp; miles Wed Apr 18 00:22:54 2007 (20m 2s)	4	1	
hike; san francisco Wed Apr 18 00:42:56 2007 (3m 19s)	1	1	
fort funston Wed Apr 18 00:46:15 2007 (1h 51m 26s)	5	1	GOAL 5: Navigate to Fort Funston
			MISSION 2: Find info on car maintenance and repair
brake pads Wed Apr 18 03:36:47 2007 (16m 36s)	6	2	GOAL 6: Find info on brake pads
auto repair Wed Apr 18 03:53:23 2007 (8m 0s)	7	2	GOAL 7: Find info on an auto body shop in San Francisco
auto body shop Wed Apr 18 04:01:23 2007 (3m 31s)	7	2	
batteries Wed Apr 18 04:04:54 2007 (0m 29s)	8	2	
car batteries Wed Apr 18 04:05:23 2007 (2m 8s)	8	2	GOAL 8: Find info on purchasing a car battery
auto body shop; san francisco Wed Apr 18 04:07:31 2007 (3m 33s)	7	2	
buy car battery online free shipping Wed Apr 18 04:11:04 2007	8	2	

Extracting Goals from User Logs in Search

About 75% of the user goals are accomplished by issuing queries across sessions.

Approaches to extracting user goals from log data:

- Clustering
- Structure Learning
- Hawkes Processes
- Entity-based Extraction

Why Sessions?

Understand User Behavior

Predict Success

Validate Metrics

User Behavior: Modeling User Frustration

- Frustration is not the same as success.
- A user can be successful in accomplishing their task but still be frustrated with their experience.



What was the best selling TV in 2008?

television set sales 2008
"television set" sales 2008
"television" sales 2008
google trends
"television" sales statistics 2008



user got frustrated
starting here



*Extracted from an actual user
study conducted at UMass*

User Behavior: Exploring vs. Struggling

Struggling Session

- Query** can you use h & r block software for more than one year
- Query** how do I file 2012 taxes on hr block
- Click** <http://www.hrblock.com>
- Query** can you only use h & r block one year
- Click** http://www.consumeraffairs.com/finance/hr_block_free.html
 - Click** <http://financialsoft.about.com/od/taxcut/gr/HR-Block-At-Home-...>
- Query** do I have to buy new tax software every year
- Click** http://financialsoft.about.com/od/simpletips/f/upgrade_yearly.htm...
 - Click** <http://askville.amazon.com/buy-version-Tax-Software-year/Answer...>

END OF SESSION

Exploring Session

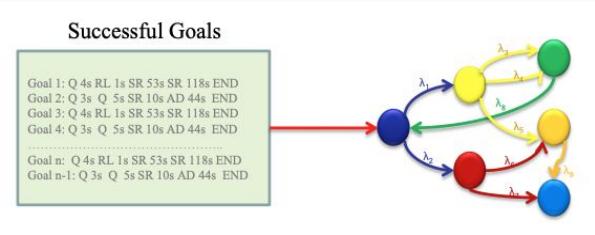
- Query** career development advice
- Click** <http://www.soperarticles.com/business-articles/career-devel...>
- Query** employment issues articles
- Click** <http://jobseekeradvice.com/category/employment-issues/...>
- Query** professional career advice
- Click** <http://ezinearticles.com/?Career-Advice-and-Professional-Ment...>
 - Click** <http://askville.amazon.com/buy-version-Tax-Software-year/Answer...>
- Query** what is a resume
- Click** <http://en.wikipedia.org/wiki/R%C3%A9sum%C3%A9>

END OF SESSION

Success Prediction: Search Trails

Time	Query	# Clicks	Avg. Dwell Time
t_1	sea bass in oven	1	Short
t_2	baked sea bass	1	Short
t_3	baked sea bass recipe	6	Long

Table 1: Example of a Successful Goal



Predict success with Markov chains by using time distributions to model each transition.

Time	Query	# Clicks	Avg. Dwell Time
t_1	gauge mod for rfactor	0	NA
t_2	gauges for rfactor	1	Short
t_3	new gauges for rfactor	0	NA
t_4	gauges mod for rf	0	NA
t_5	new tacks for rfactor	1	Short
t_6	rfactor gauge plugin	0	NA

Table 2: Example of an Unsuccessful Goal

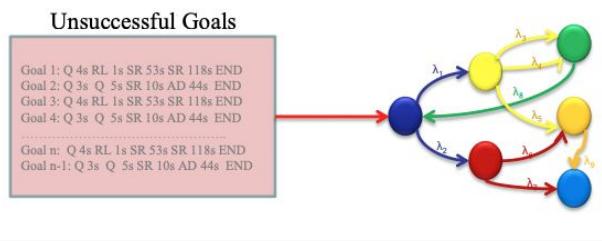


Figure Source: [Understanding & Inferring User Tasks and Needs](#)

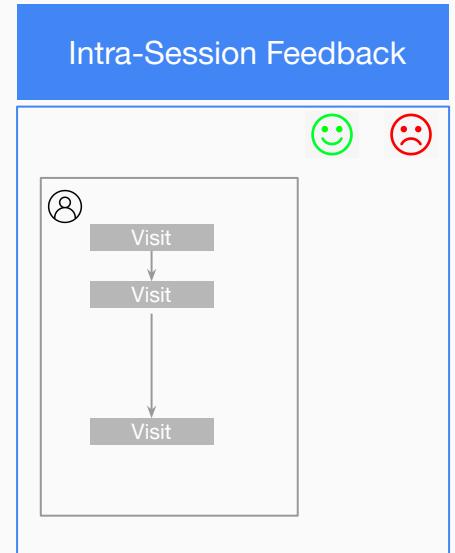
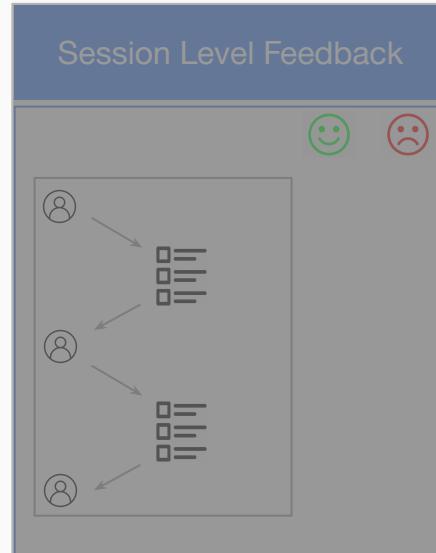
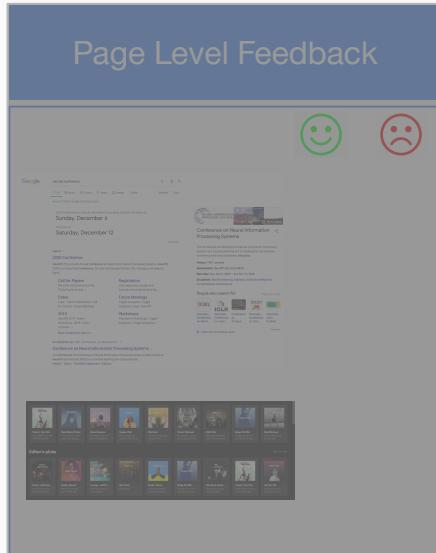
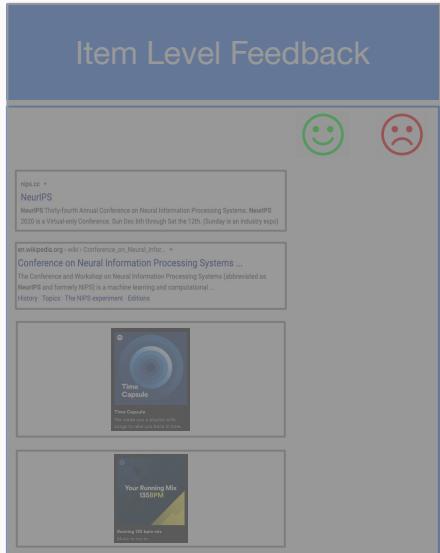
Metric Validation: Session Level Annotations

Table 3 A list of user actions that are included in utility. The row rank reflects a descending order in the absolute value of the weight associated with the events.

Events	Type
A click to an external page that is the last interaction in the user's session	strongly positive
A query issued by the user that is followed by a query that reformulates it	strongly negative
A click to an external page with a long dwelltime	positive
A click to an external page with a short dwelltime	negative
A query issued by the searcher that is not followed by a reformulation	weakly negative

Inter-Session & Long-Term Feedback

short-term



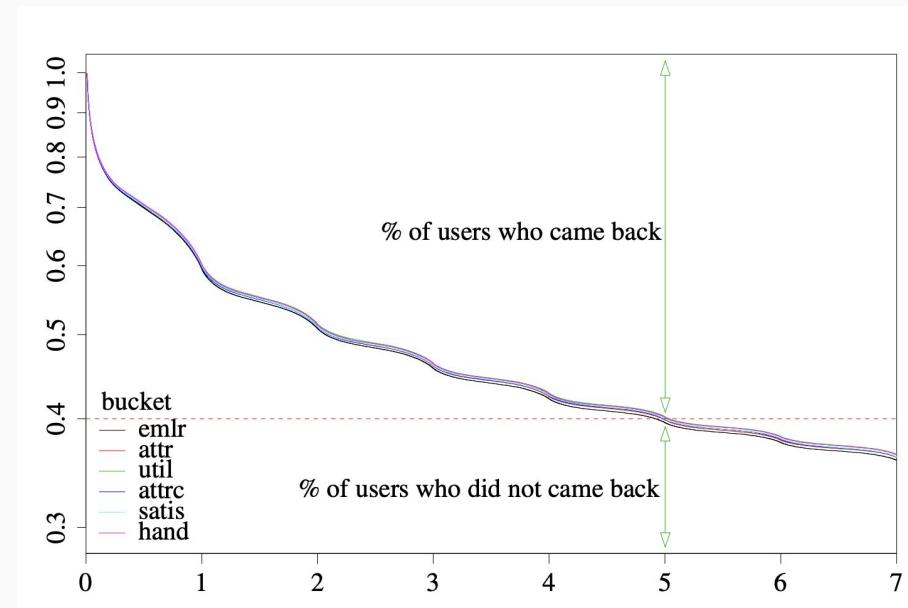
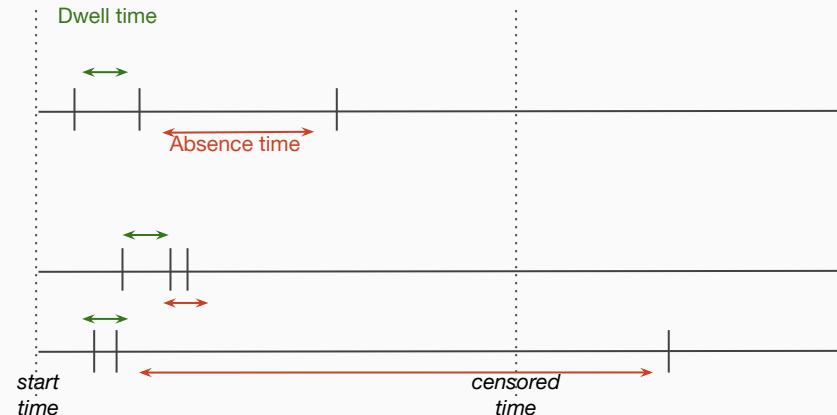
long-term

Inter-Session & Long-Term Feedback

- Short-term metrics such as click can be misleading as they ignore user learning & novelty effects over time.
- Short-term signals ignore user's long-term goals/needs and a myopic view of success can result in "echo-chamber" effects or "filter bubbles". Kohavi et al. showed that optimizing for short-term improvements may be detrimental in the long-term.
- Measuring long-term success is challenging and complicated. We discuss two approaches
 - Predicted Metrics
 - Surrogates or Proxies

Inter-Session Feedback: Absence Time

- The time between successive sessions can be used to measure satisfactions (i.e., quicker the users return to the service → more satisfied).
- Modeling absence time using survival analysis (Cox model) to predict future outcome can be used to compare ML systems.



Inter-Session Feedback: Surrogates & Proxies

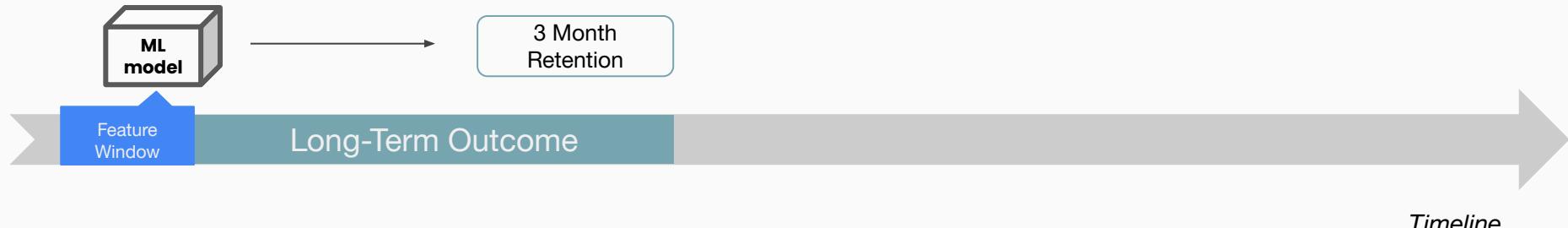
- Satisfied users coming back to the service and re-engaging with the ML system is a reasonable proxy for long term success.
- One or more short-term metrics such as Clickthrough Rate, Session success can be used as proxies to estimate long-term outcomes (e.g. user retention).



Susan Athey, Raj Chetty, Guido W. Imbens, Hyunseung Kang. **The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely.** NBER 2019.

Inter-Session Feedback: Surrogates & Proxies

- Satisfied users coming back to the service and re-engaging with the ML system is a reasonable proxy for long term success.
- One or more short-term metrics such as Clickthrough Rate, Session success can be used as proxies to estimate long-term outcomes (e.g. user retention).
- Alternatively, modeled metrics that predict long-term success could be used to compare ML systems.



Susan Athey, Raj Chetty, Guido W. Imbens, Hyunseung Kang. **The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely.** NBER 2019.

Additional Reading Materials and References

Relevant Tutorials

- Harrie Oosterhuis et al. **Unbiased Learning to Rank: Counterfactual and Online Approaches.** WWW 2020
- *Liangjie Hong & Mounia Lalmas Tutorial on Online User Engagement: Metrics and Optimization.* KDD 2020
- *Jean Garcia-Gathright et al. Mixed methods for evaluating user satisfaction.* RecSys 2019
- *Rishabh Mehrotra et al. Understanding & Inferring User Tasks and Need.* WWW 2018.
- *Thorsten Joachims & Adith Swaminathan Tutorial on Counterfactual Evaluation and Learning.* SIGIR 2016
- *Aleksandr Chuklin et al. Click models for web search and their applications to IR.* WSDM 2016.
- *Et al. Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.* KDD 2019

Survey Papers

- *Jiawei Chen et al. Bias and Debias in Recommender System: A Survey and Future Directions.* TKDE 2020