

# Beyond Accuracy

Grounding Evaluation Metrics for Human-Machine Learning Systems

# Introduction

Fernando Diaz, Google

# Artificial Intelligence in Decision Support

**Consumer domains:** web search, music recommendation, etc

**Specialized domains:** legal discovery, clinical decision support, etc.

# Accuracy is Not Enough

reducing  $\frac{1}{n} \sum_{i=1}^n (y_i - f_\theta(x_i))^2$

does not necessarily improve **user satisfaction**

# user-focused metrics

**objective:** emphasize development of tools to help individuals with their goals.

**metric definition problem:** how to develop a metric that accurately reflects how much an individual is being helped

# why does the right metric matter?

Evaluation of recommender has long been divided between accuracy metrics (e.g., precision/recall) and error metrics (notably, RMSE and MAE). The mathematical convenience and fitness with formal optimization methods, have made error metrics like RMSE more popular, and they are indeed dominating the literature. However, it is well recognized that accuracy measures may be a more natural yardstick, as they directly assess the quality of top-N recommendations.

This work shows, through an extensive empirical study, that the convenient assumption that an error metric such as RMSE can serve as good proxy for top-N accuracy is questionable at best. There is no monotonic relation between error metrics and accuracy metrics. This may call for a re-evaluation of optimization goals for top-N systems. On the bright side we have presented simple and efficient variants of known algorithms, which are useless in RMSE terms, and yet deliver superior results when pursuing top-N accuracy.

**task:** recommendation

**metric:** minimize error of predicted ratings

# why does the right metric matter?

In this paper, we have made a first attempt to systematically investigate the correlation of automatic ROUGE scores with human evaluation for meeting summarization. Adaptations on ROUGE setting based on meeting characteristics are proposed and evaluated using Spearman's rank coefficient. Our experimental results show that in general the correlation between ROUGE scores and human evaluation is low, with ROUGE SU4 score showing better correlation than ROUGE-1 score. There is significant improvement in correlation when disfluencies are removed and speaker information is leveraged, especially for evaluating system-generated summaries. In addition, we observe that the correlation is affected differently by those factors for human summaries and system-generated summaries.

**task:** summarization

**metric:** target summary word recall

# why does the right metric matter?

Most recommendation engines today are based on predicting user engagement, e.g. predicting whether a user will click an item or not. However, there is potentially a large gap between engagement signals and a desired notion of *value* that is worth optimizing for [Ekstrand and Willemsen, 2016]. Just because a user engages with an item doesn't mean they value it. A user might reply to an item because they are angry about it, or click an item in order to gain more information about it [Wen et al., 2019], or watch addictive videos out of temptation.

**task:** recommendation

**metric:** user clicks

# intrinsic vs. extrinsic evaluation

**intrinsic evaluation:** measuring the performance of a *component* of a system, independent of how it contributes to the end task (and often independent of the user)

**extrinsic evaluation:** measuring the performance of a *component* of a system, in the context of how it contributes to the end task and user.

# situated versus simulated environments

**simulated evaluation:** some aspects of evaluation are constructed in controlled settings (e.g. user models, labels, log data); most “offline evaluation”.

**situated evaluation:** almost all aspects of the evaluation are observed only as the system is used (e.g. production tests); most “online evaluation”.

# task-oriented domains

summarization

extraction

retrieval

recommendation

# what we will cover

offline metrics

online/behavioral metrics

multiple metrics

open problems

# what we will not cover

optimization of behavior toward a metric

aggregation population level metrics

experimentation

# Offline Metrics

Fernando Diaz, Google

# what are offline metrics?

system evaluation that replaces real user behavior with some proxy for that behavior.

## proxies

- user models
- annotations
- log data

# why use offline metrics

- faster than running a real experiment
- repeatable
- avoids exposing users to ineffective systems

# **example:** search and recommendation

**task:** given an a query/context and a repository of items, find relevant information

# **example:** search and recommendation

**task:** given an a **query/context** and a repository of items, find relevant information

# **example:** search and recommendation

**task:** given an a query/context and a **repository of items**, find relevant information

# **example:** search and recommendation

**task:** given an a query/context and a repository of items, find **relevant information**

# instances

$$\mathcal{D}_q = \{\langle x, y \rangle_1, \dots, \langle x, y \rangle_n\}$$

- $x$  query-document features
  - (e.g. term matches,
  - item popularity, context)
- $y$  relevance

*all metric definition for a single query or context*

# **metric:** accuracy

$$\sum_{i=1}^n (y_i - f_\theta(x_i))^2$$

$f$  document scoring function

$y_i$  relevance label for document  $i$

$x_i$  query-document features for document  $i$

# example: search and recommendation

**task:** given an a query/context and a repository of items, find relevant information

## questions

- *how is information presented?*
- *how do users consume the information?*

# **metric:** precision

$$\text{Prec}(\mathcal{Y}^+, \mathcal{F}_\theta^+(X)) = \frac{|\mathcal{Y}^+ \cap \mathcal{F}_\theta^+(X)|}{|\mathcal{F}_\theta^+(X)|}$$

$\mathcal{Y}^+$  relevant document set

$\mathcal{F}_\theta^+(X)$  predicted relevant document set

# **metric: recall**

$$\text{Rec}(\mathcal{Y}^+, \mathcal{F}_\theta^+(X)) = \frac{|\mathcal{Y}^+ \cap \mathcal{F}_\theta^+(X)|}{|\mathcal{Y}^+|}$$

# ranking

browsing


$$\begin{matrix} \pi(1) \\ \pi(2) \\ \pi(3) \\ \vdots \\ \pi(n-2) \\ \pi(n-1) \\ \pi(n) \end{matrix}$$

# **metric:** expected search length

**user model:** in-order traversal of a ranked list, collecting up to  $k$  items.

**metric:** number of nonrelevant documents skipped before reaching  $k$  relevant items.

$$\text{ESL}(\mathcal{Y}^+, \pi, k) = \min-k_{i \in \mathcal{Y}^+} \bar{\pi}(i)$$

$\min-k$   $k$ th smallest value

$\bar{\pi}(i)$  rank position of item  $i$

# **metric:** R-precision

**user model:** in-order traversal of a ranked list, collecting all relevant items.

**metric:** precision when recall is 1.

$$\text{RPrec}(\mathcal{Y}^+, \pi) = \text{Prec}(\mathcal{Y}^+, \pi_{1:k^*})$$

$$k^* = \max_{i \in \mathcal{Y}^+} \bar{\pi}(i)$$

# **metric:** R-precision

**user model:** in-order traversal of a ranked list, collecting all relevant items.

**metric:** precision when recall is 1.

$$\begin{aligned}\text{RPrec}(\mathcal{Y}^+, \pi) &= \text{Prec}(\mathcal{Y}^+, \pi_{1:k^*}) \\ &= \frac{|\mathcal{Y}^+|}{\text{ESL}(\mathcal{Y}^+, \pi, |\mathcal{Y}^+|)}\end{aligned}$$

# **metric:** reciprocal rank

**user model:** in-order traversal of a ranked list, satisfied by one item.

**metric:** inverse of the number of documents skipped before reaching the relevant item.

$$\text{RR}(\mathcal{Y}^+, \pi) = \max_{i \in \mathcal{Y}^+} \frac{1}{\bar{\pi}(i)}$$

# **metric:** reciprocal rank

**user model:** in-order traversal of a ranked list, satisfied by one item.

**metric:** inverse of the number of documents skipped before reaching the relevant item.

$$\begin{aligned} \text{RR}(\mathcal{Y}^+, \pi) &= \max_{i \in \mathcal{Y}^+} \frac{1}{\bar{\pi}(i)} \\ &= \frac{1}{\text{ESL}(\mathcal{Y}^+, \pi, 1)} \end{aligned}$$

# **metric:** average precision

**user model:** in-order traversal of a ranked list, collecting all relevant items.

**metric:** precision averaged over all recall levels.

$$\text{AP}(\mathcal{Y}^+, \pi) = \frac{1}{|\mathcal{Y}^+|} \sum_{i \in \mathcal{Y}^+} \text{Prec}(\mathcal{Y}^+, \pi_{1:\bar{\pi}(i)})$$

# **metric:** average precision

**user model:** in-order traversal of a ranked list, collecting all relevant items.

**metric:** precision averaged over all recall levels.

$$\begin{aligned} \text{AP}(\mathcal{Y}^+, \pi) &= \frac{1}{|\mathcal{Y}^+|} \sum_{i \in \mathcal{Y}^+} \text{Prec}(\mathcal{Y}^+, \pi_{1:\bar{\pi}(i)}) \\ &= \frac{1}{|\mathcal{Y}^+|} \sum_{r=1}^{|\mathcal{Y}^+|} \frac{r}{\text{ESL}(\mathcal{Y}^+, \pi, r)} \end{aligned}$$

# **metric:** rank-biased precision

**user model:** in-order traversal of a ranked list; utility independent of stopping probability.

**metric:** expected utility given simulated user behavior.

$$\text{RBP}(y, \pi) = (1 - \gamma) \sum_{r=1}^n y_{\pi(r)} \gamma^{r-1}$$

$\gamma$  patience parameter

# **metric:** expected utility

**user model:** in-order traversal of a ranked list, gains utility of 1 for each relevant item.

**metric:** expected utility given simulated user behavior.

$$\text{EU}(y, \pi) = \sum_{r=1}^n y_{\pi(r)} \phi(y_{\pi(r)}) \underbrace{\gamma^{r-1} \prod_{r'=1}^{r-1} (1 - \phi(y_{\pi(r')}))}_{\text{examination probability}}$$

$\phi(i)$  probability that user stops given relevance of item  $i$

# **metric:** time-based gain

**user model:** in-order traversal of a ranked list; utility independent of stopping probability; utility based on time to reach the position.

**metric:** expected utility given simulated user behavior.

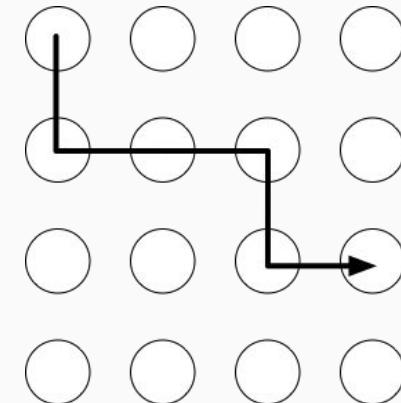
$$\text{TBG}(y, \pi) = \sum_{r=1}^n y_{\pi(r)} \Delta(y, \pi, r)$$

$\Delta(y, \pi, r)$  discount based on time to reach rank  $r$

# grid metrics

**user model:** model of ordered user traversal of grid items; utility independent of stopping probability.

**metric:** expected utility given simulated user behavior.



# aggregating metrics

- traffic-weighted
- unique query/context
- usage-segmented
- group-segmented

# further reading

IR evaluation: Modeling user behavior for measuring effectiveness

Charles L.A. Clarke, Mark D. Smucker, and Emine Yilmaz

SIGIR, 2015

Ian Soboroff. Building Test Collections, SIGIR 2017.

Offline evaluation options for recommender systems

Rocío Cañamares, Pablo Castells, and Alistair Moffat

Information Retrieval Journal, 2020

# 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

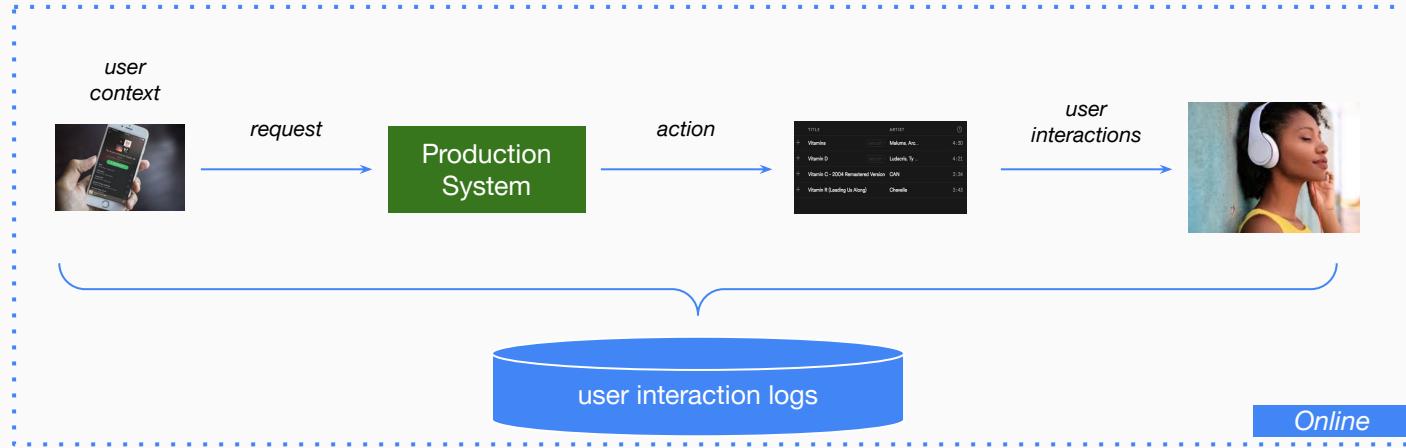
## Explicit Feedback

- 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)

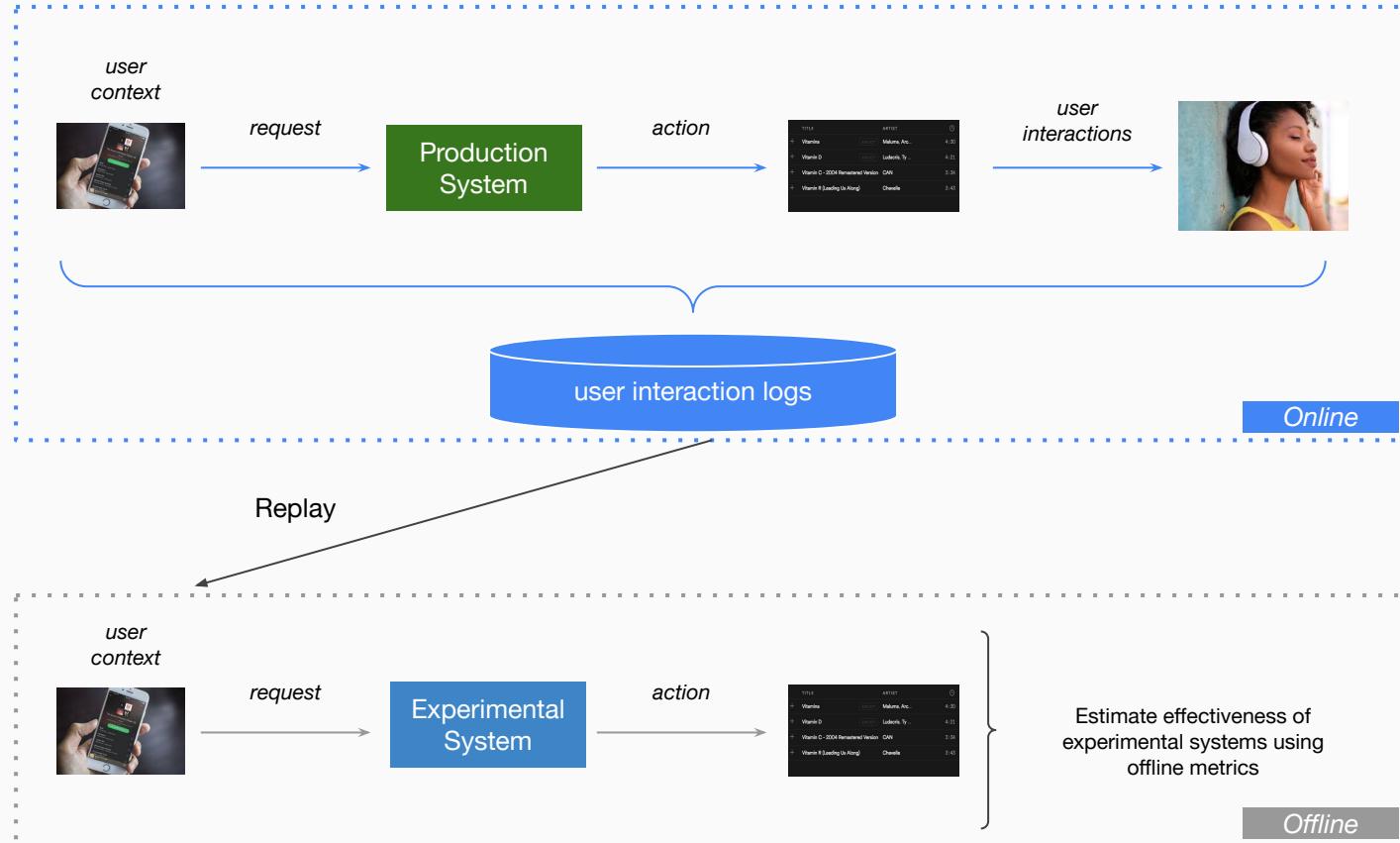
## Implicit Feedback

- 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

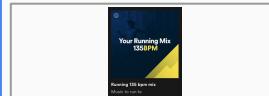
*short-term*

## Item Level Feedback



ripc.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



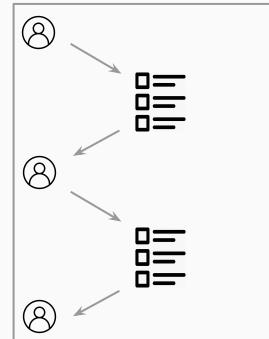
## Page Level Feedback



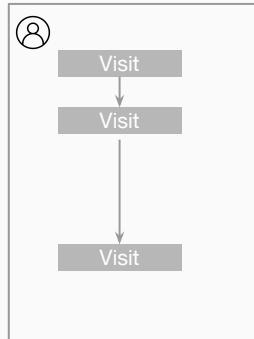
Google search results for "NeurIPS Conference". The top result is a link to the NeurIPS website, which shows the conference schedule from December 6 to 12, 2020, and various sessions like "Keynote", "Workshop", and "Poster". Below the search results are several thumbnail images of conference papers or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

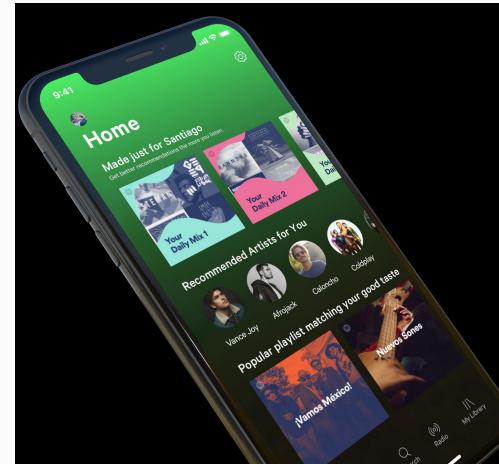
[en.wikipedia.org/wiki/Conference\\_on\\_Neural\\_Information\\_Processing\\_Systems](#)

**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 Virtual-only Conference. Sun Dec 6th through Sat the 12th. (Sunday is an industry expo)

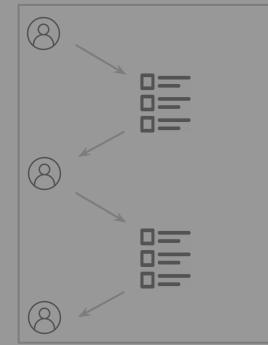
[en.wikipedia.org / wiki / Conference\\_on\\_Neural\\_Infro... \\*](https://en.wikipedia.org/wiki/Conference_on_Neural_Infro...)  
**Conference on Neural Information Processing Systems ...**  
The Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS) is a machine learning and computational ...  
History · Topics · The NIPS experiment · Editions

A screenshot of a mobile application titled "Time Capsule". The main screen features a large blue circular icon with concentric rings, resembling a target or a clock face. Below the icon, the word "Time Capsule" is displayed in white capital letters. At the bottom of the screen, there is a dark bar containing the text "Time Capsule" and "We made you a playlist with...".



The screenshot shows the homepage of the NeurIPS conference. At the top, there's a navigation bar with links for Home, Schedule, Papers, Tutorials, and About. Below the navigation is a search bar. The main content area features a large image of a brain, followed by the text "Conference on Neural Information Processing Systems". There are sections for "Keynote Speakers", "Program Committees", "Organizers", and "Sponsors". A sidebar on the left lists "About", "Call for Papers", "Registration", "Program Committees", "Keynote Speakers", "Organizers", and "Sponsors". At the bottom, there's a footer with links for "About", "Call for Papers", "Registration", "Program Committees", "Keynote Speakers", "Organizers", and "Sponsors".

A screenshot of the Kindle app's 'New Books' section. The top row shows ten book covers from left to right: 'Natural Psychology' by Daniel Siegel, 'How Birds Fly' by Steve Parker, 'Breakthrough' by Michael S. Lewis, 'Pretty Bird' by Michael S. Lewis, 'The Art of War' by Sun Tzu, 'Power Without Influence' by Robert H. Schuller, 'How to Win Friends and Influence People' by Dale Carnegie, 'The Big Picture' by Edward O. Wilson, and 'The Five Dysfunctions of a Team' by Patrick Lencioni. Below this is a header 'Editor's choice'. The bottom row shows five more book covers: 'The Five Dysfunctions of a Team' by Patrick Lencioni, 'The Art of War' by Sun Tzu, 'How Birds Fly' by Steve Parker, 'Natural Psychology' by Daniel Siegel, and 'The Big Picture' by Edward O. Wilson.



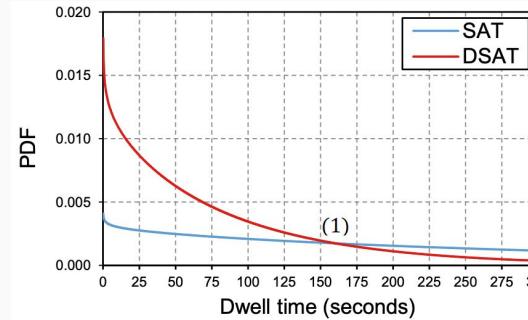
# Behavioral Signals: Clicks

- 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 advertising the "2020 Conference on Neural Information Processing Systems" held on December 6-12, 2020. To the right of the 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, NeurIPS, and NeurIPS 2019. On the right is a screenshot of a Spotify interface showing a grid of "Editor's picks" tracks. A large blue hand icon is overlaid on the left side of the Spotify screen, pointing towards the track thumbnails.

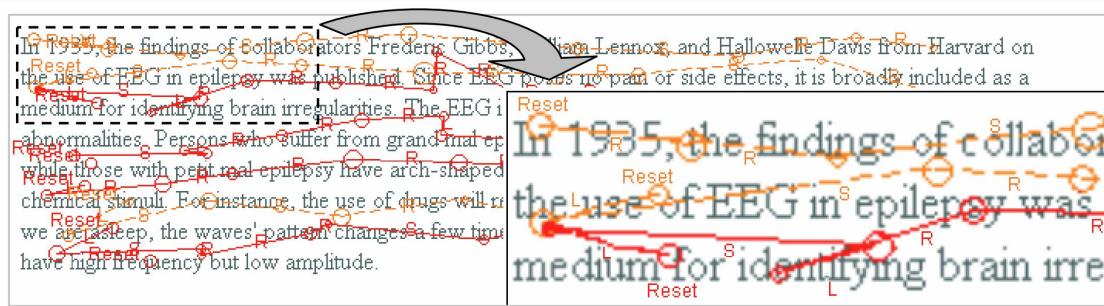
# Behavioral Signals: Dwell Time

- **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

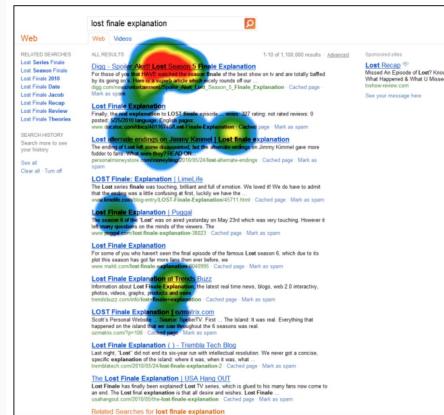
- User eye movement exhibits unique behavior when reading. The eye fixation and saccade could be used to infer satisfaction.
  - **Fixation:** 200-250 ms steady gaze at one point.      ➤ **Saccade:** Rapid eye movement from one fixation to the next.



# 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

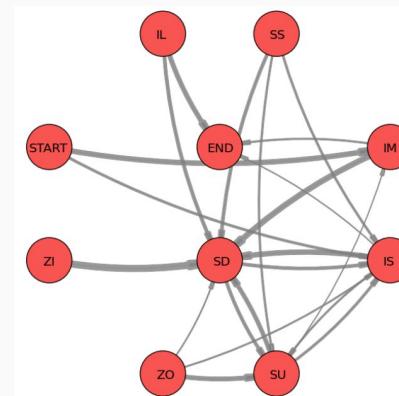


Cursor movement positions



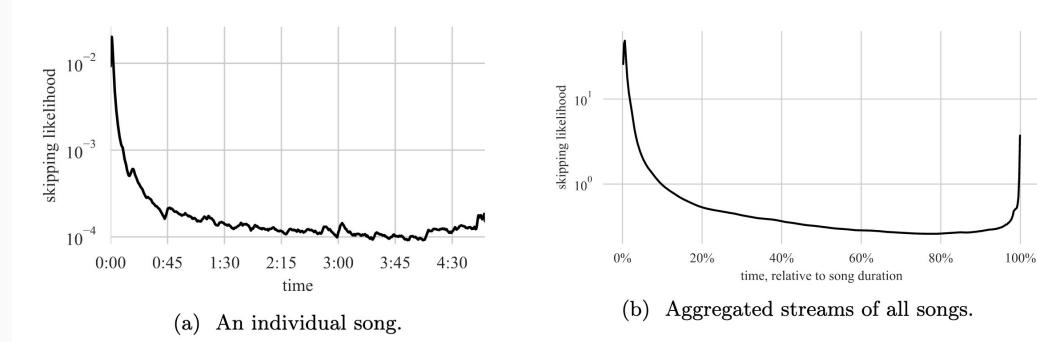
# Behavioral Signals: Touch Gestures on Mobile

- 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.

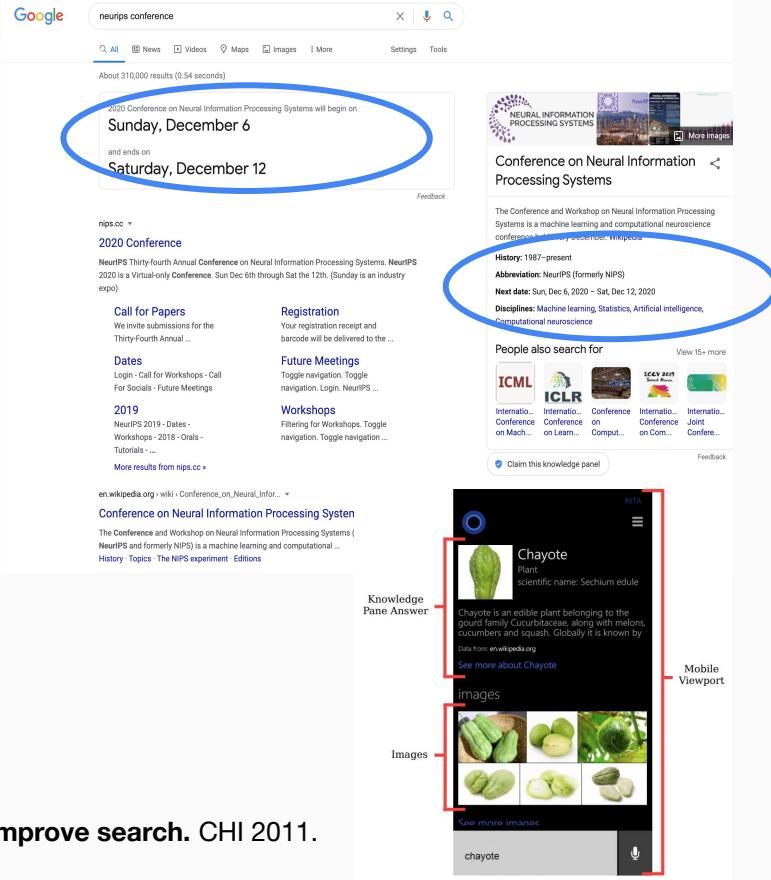
TITLE	ARTIST	ALBUM
See Fernando	Jenny Lewis	Acid Tongue
Plans	Oh Wonder	Oh Wonder
Stand Up	The Prodigy	Invaders Must Die
Retrospect	Sun Ra	A Fireside Chat with...
Satisfied	Andrew Bird	The Swimming Hour
Galician Chimpanzee I	Galegoz	Galician Chimpanzee
Go Or Go Ahead	Rufus Wainwright	Want One
Move Me	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

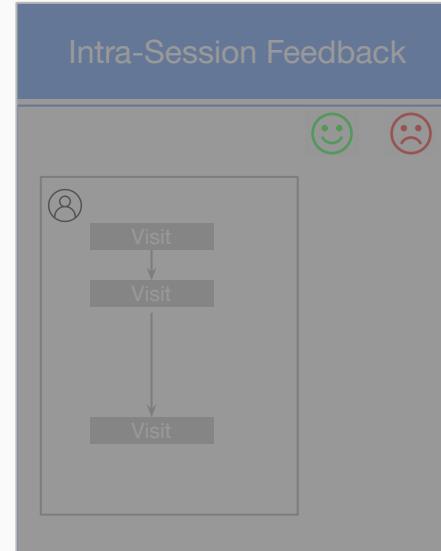
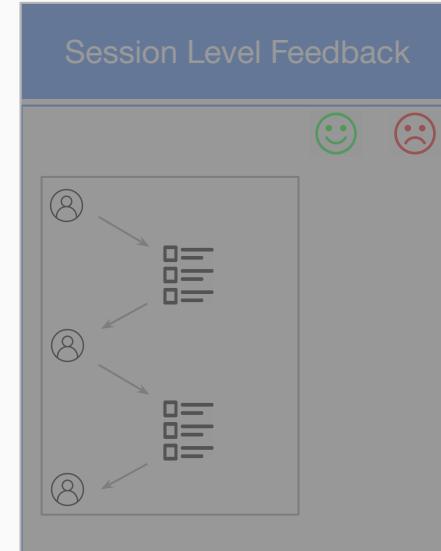
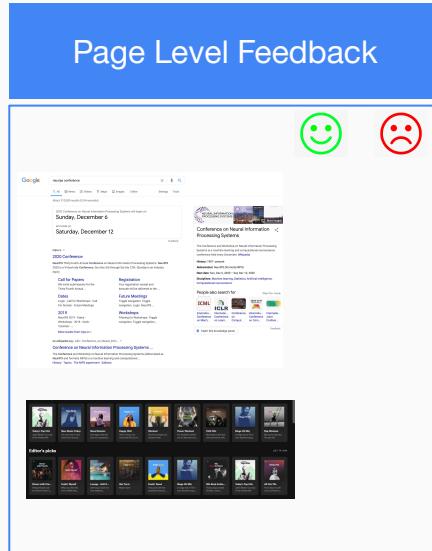
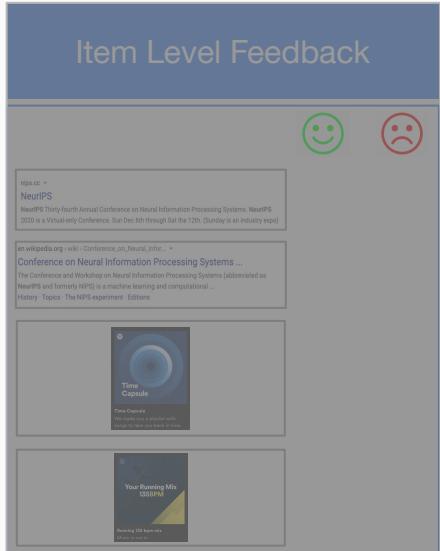


*Offline  
metric  
validation*

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

# Page Level Feedback

*short-term*



*long-term*

# Whole Page Feedback

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

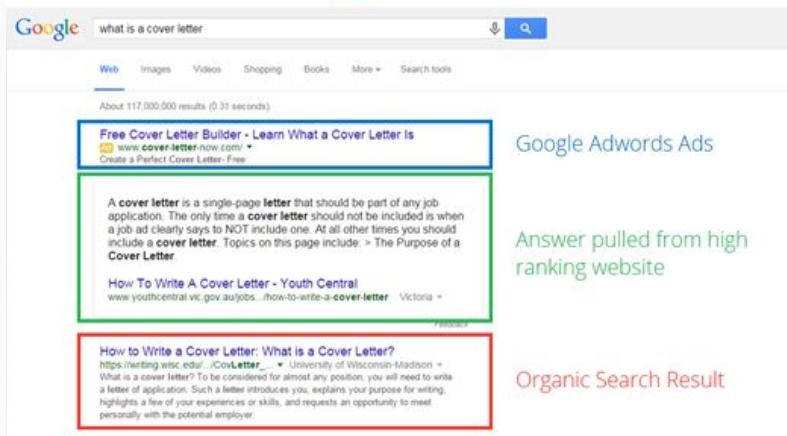
$$reward = f(A)$$

- 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**

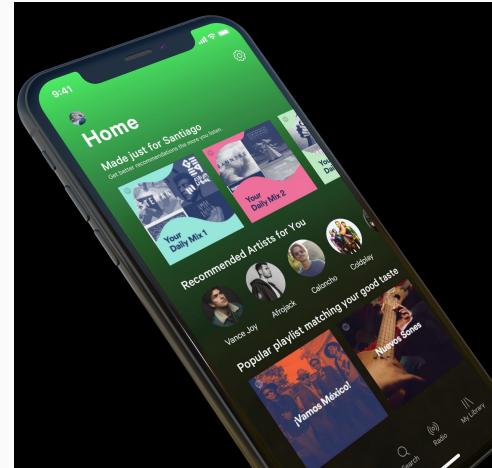
**Organic Search Result**

**Free Cover Letter Builder - Learn What a Cover Letter Is**  
[www.cover-letter-now.com/](http://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](http://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

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

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

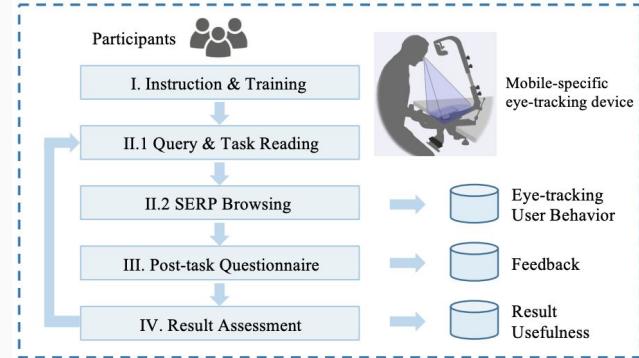
# User Browsing Models: Position-Based Model

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

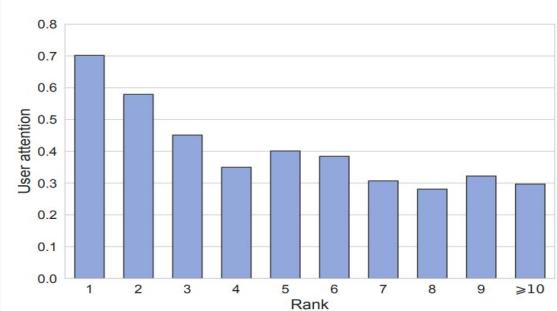
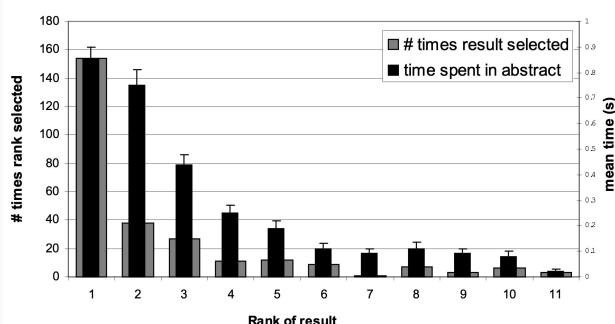
Do they read the abstracts sequentially from top to bottom, or do they skip links?

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*

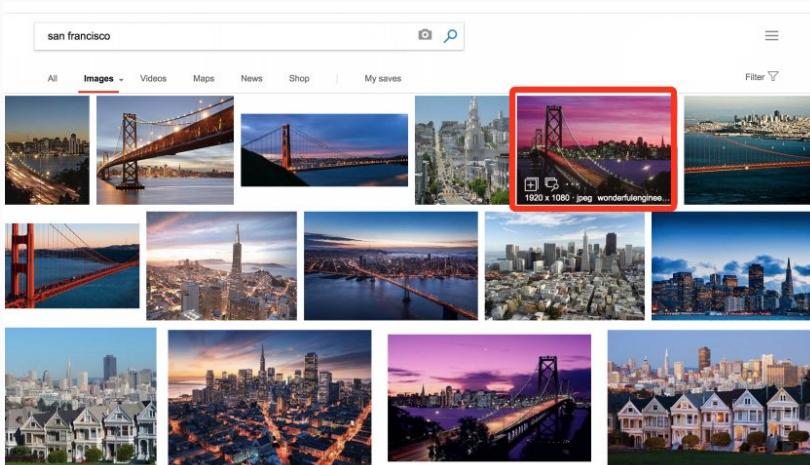


Laura A. Granka et al. Eye-tracking analysis of user behavior in WWW search. SIGIR 2004.

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.

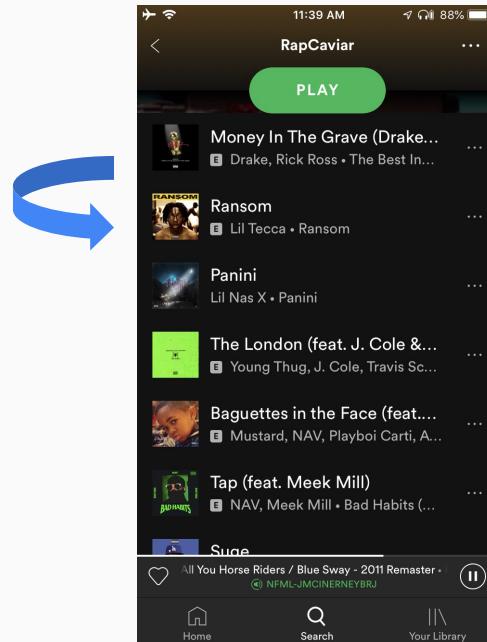
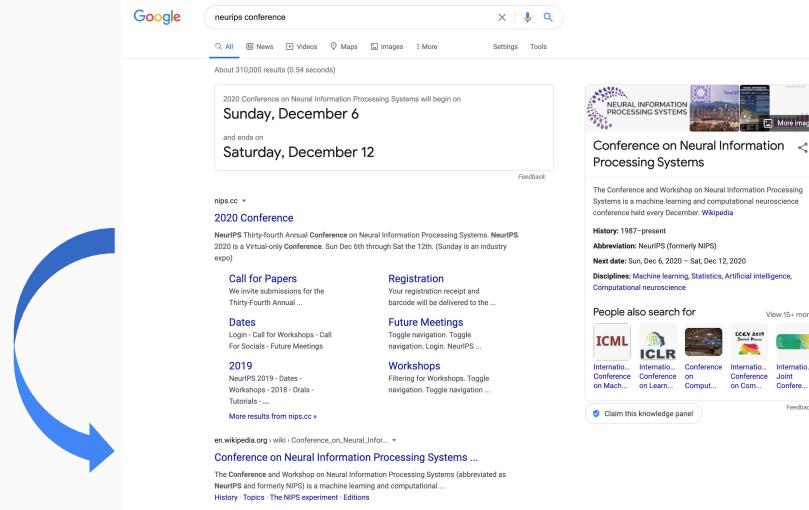


***middle positions*** may attract more attention depending on the domain in grid-based layouts.



# 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).



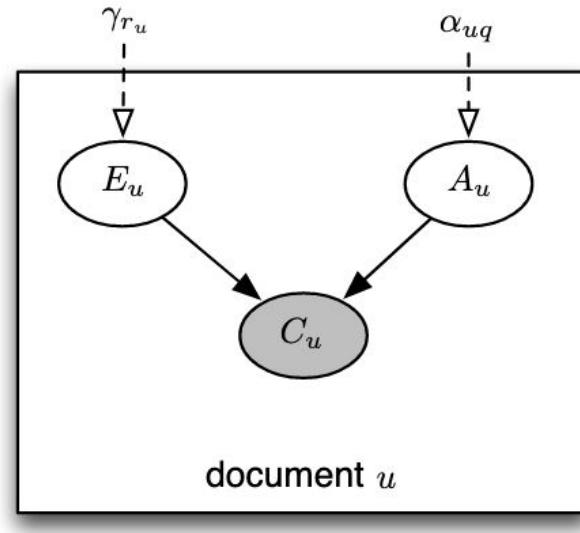
# Click Models: Position-Based Model

- 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***.

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

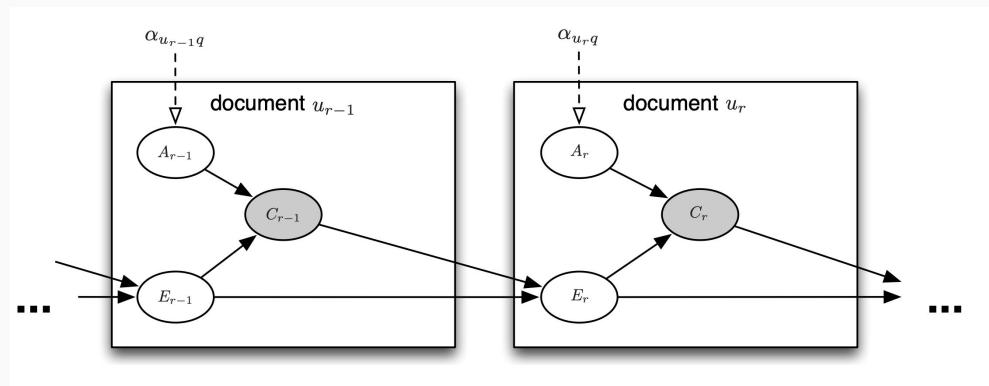
$$P(A_u = 1) = \alpha_{uq}$$

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

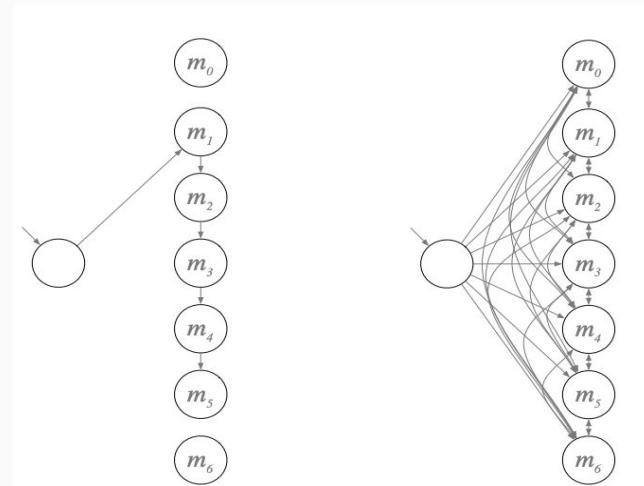
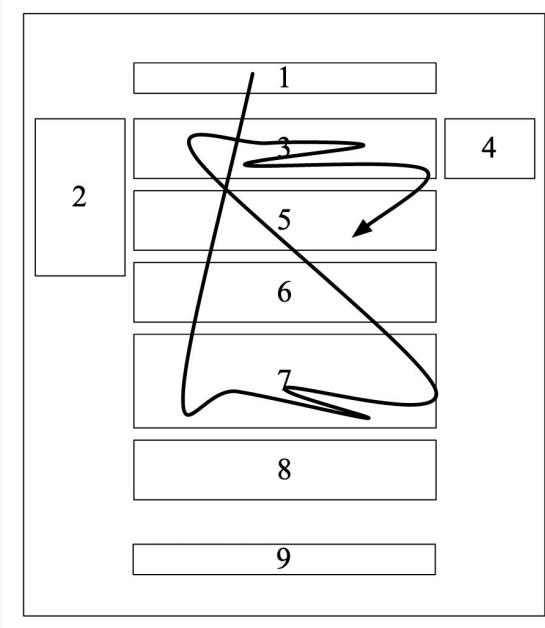


# Click Models: Cascade Model

- 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.

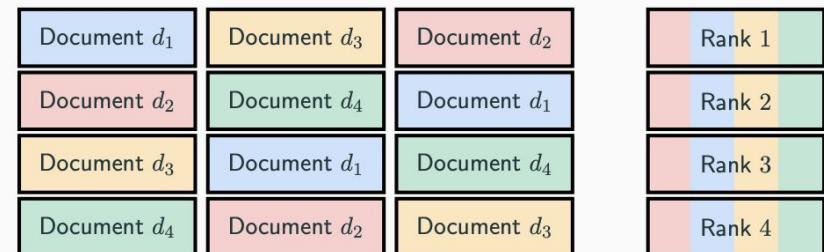


Figure Source: [Unbiased Learning to Rank: Counterfactual and Online Approaches](#)

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

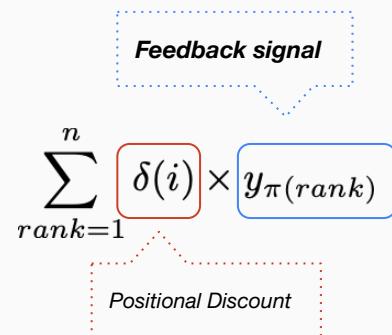
# Counterfactual Evaluation from Log Data

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

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

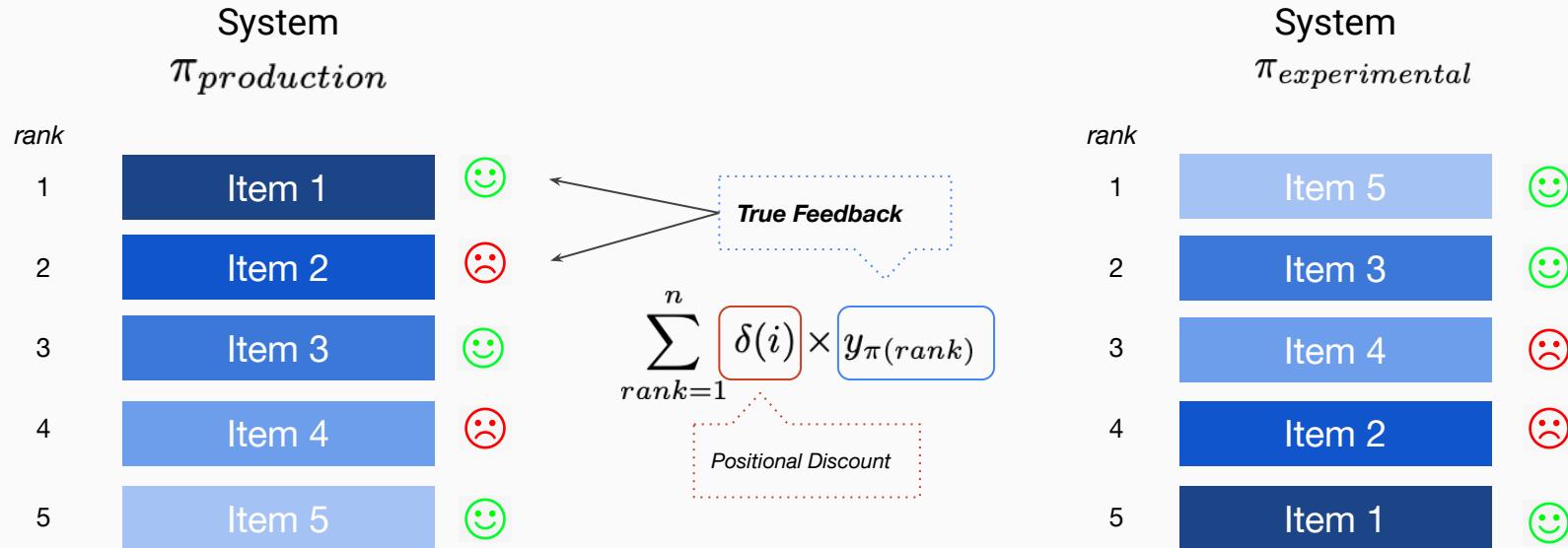
*Feedback signal*

*Positional Discount*

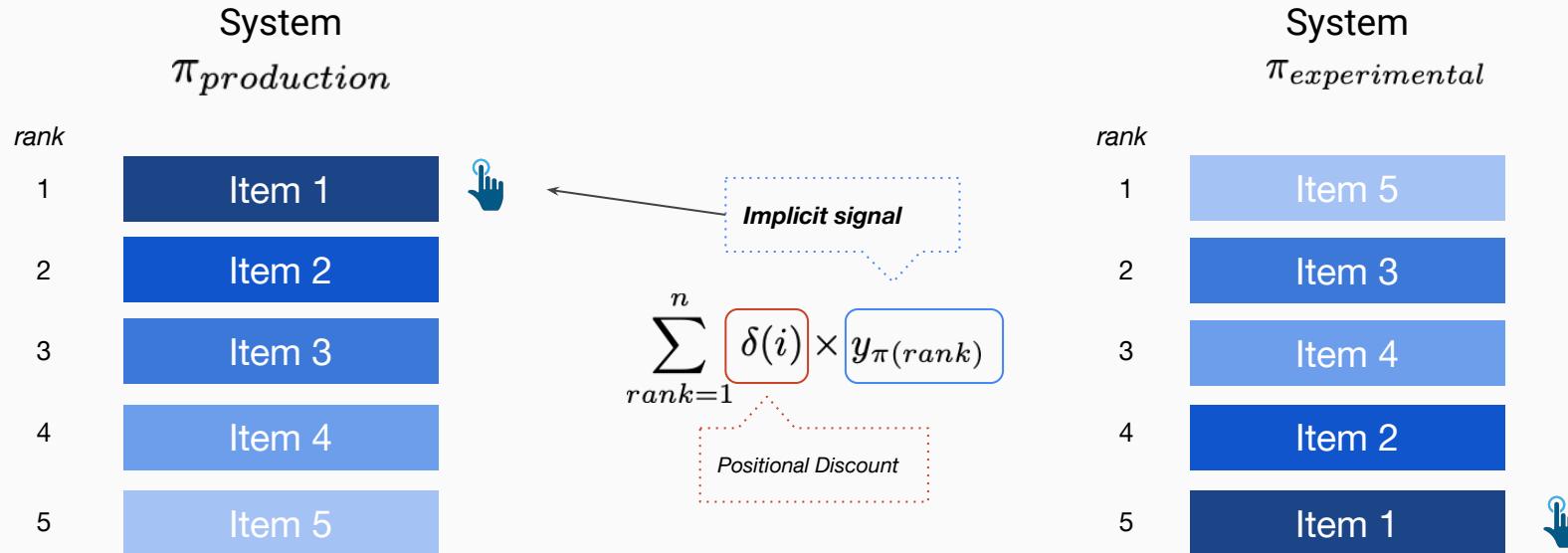


System $\pi_{experimental}$	
rank	
1	Item 5
2	Item 4
3	Item 3
4	Item 2
5	Item 1

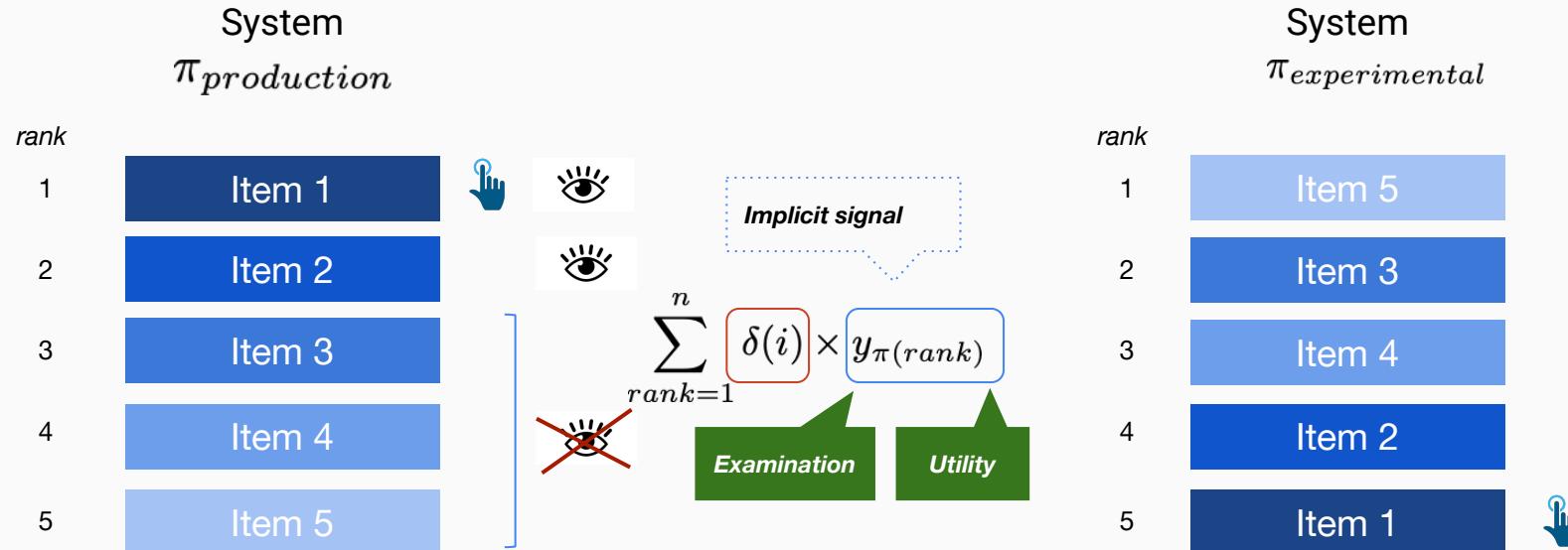
# Counterfactual Evaluation from Log Data



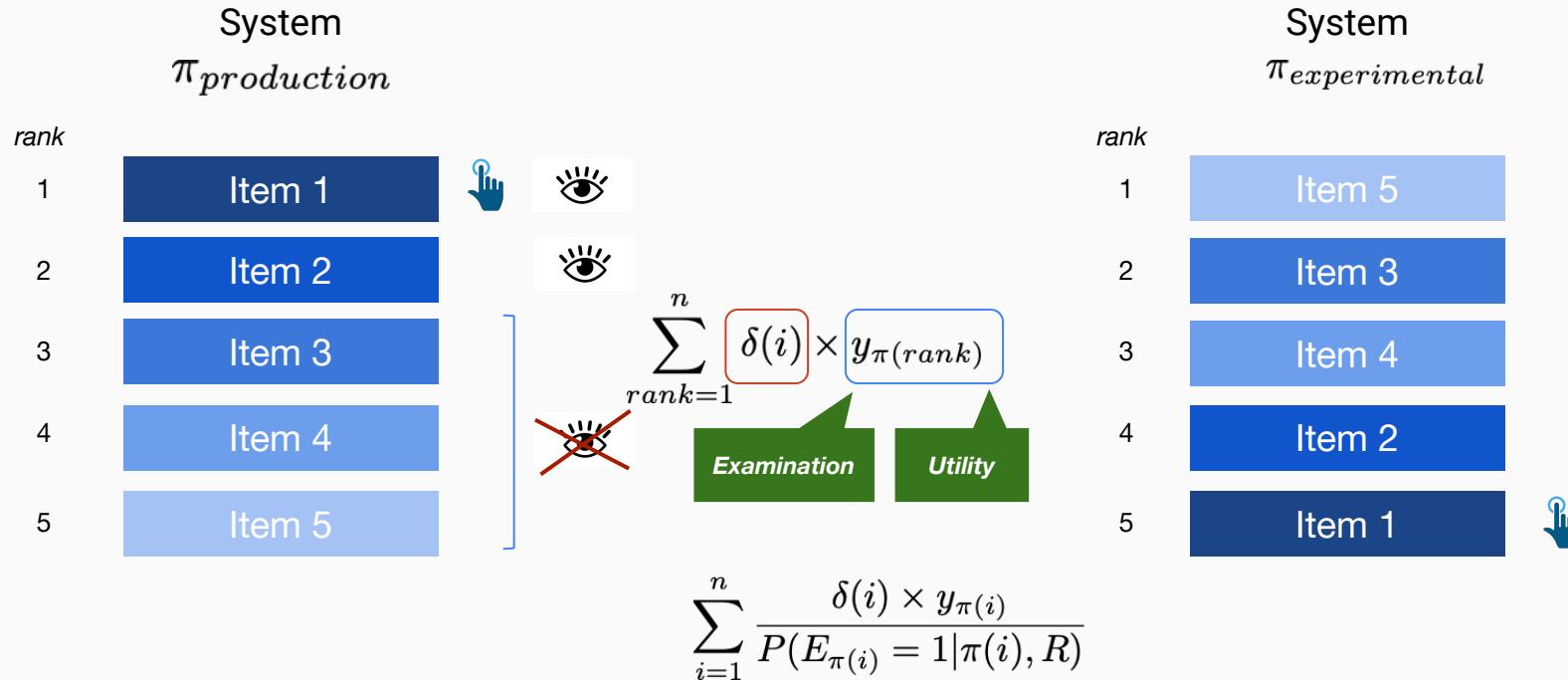
# Counterfactual Evaluation from Log Data



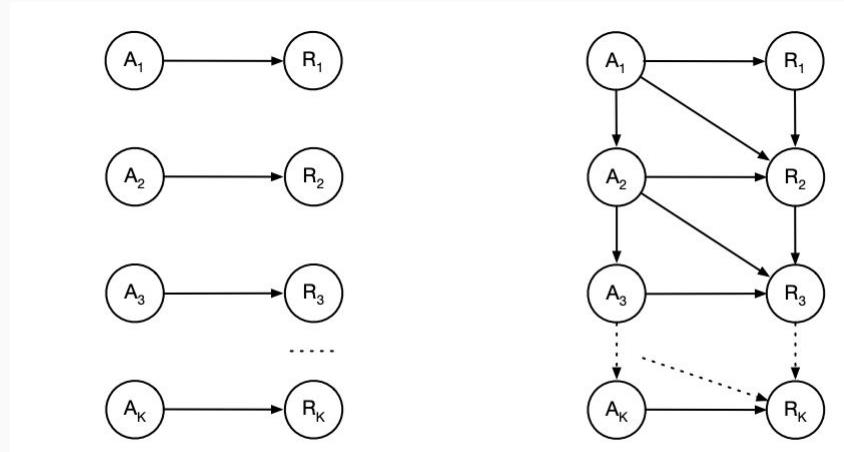
# Counterfactual Evaluation from Log Data



# 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?

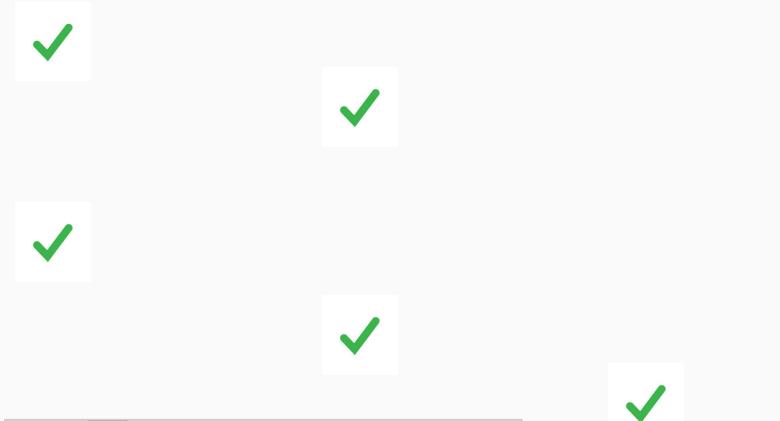
Grand Canyon Skywalk at Grand Canyon West  
<https://grandcanyon.com/planning/west.../grand-canyon-skywalk-at-grand-canyon-we...> ▾  
Grand Canyon Skywalk - You're standing on a platform made of glass looking out over the Grand Canyon. Eagle Point in front and Colorado River to your left.

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.

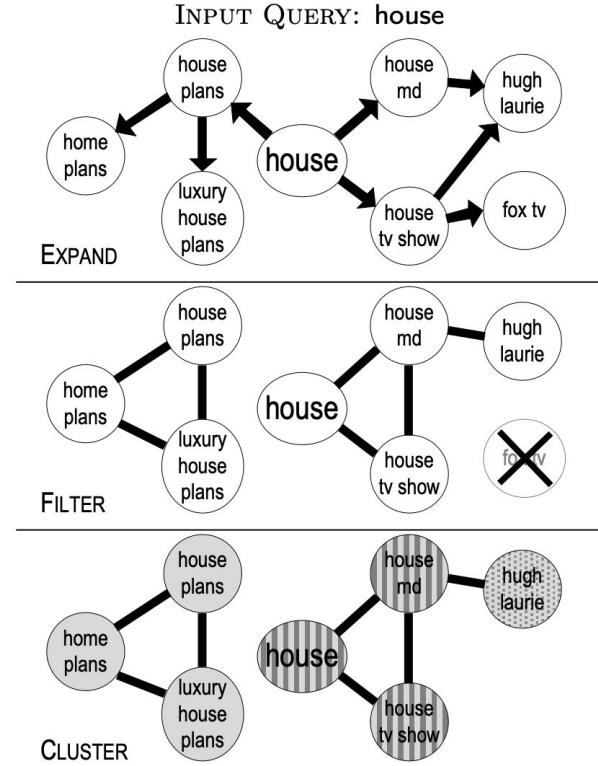
Plan Your Visit - Grand Canyon National Park (U.S. National Park ...  
<https://www.nps.gov/grca/planyourvisit/index.htm> ▾  
Jun 6, 2017 - Most visitors (90%) see Grand Canyon from the "South Rim" from overlooks accessed ... view from patio of grand canyon lodge on the north rim ...

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*



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

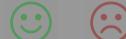
en.wikipedia.org : wiki : Conference\_on\_Neural\_Information\_Processing\_Systems ...

# 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 ...

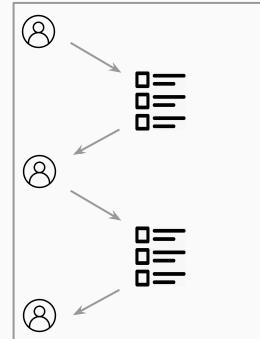
History Topics The NIPS experiment . Editions

A small image of a CD cover titled "Your Running Mix 135BPM" featuring a yellow triangle graphic.



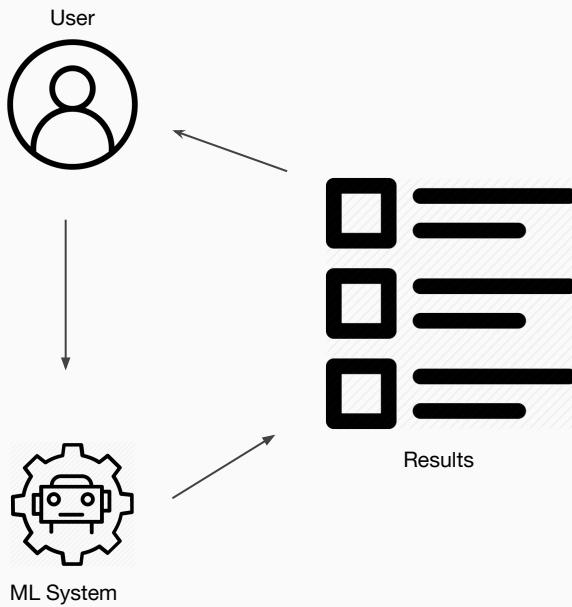
A screenshot of the 'Editor's picks' section on the National Geographic website. The page features a grid of 15 small article cards, each with a thumbnail image, the article title, and a brief description. The titles include 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', 'Travel: 10 Best Places to See the Northern Lights', and 'Travel: 10 Best Places to See the Northern Lights'. Each card has a small 'Read more' link at the bottom.

## Session Level Feedback

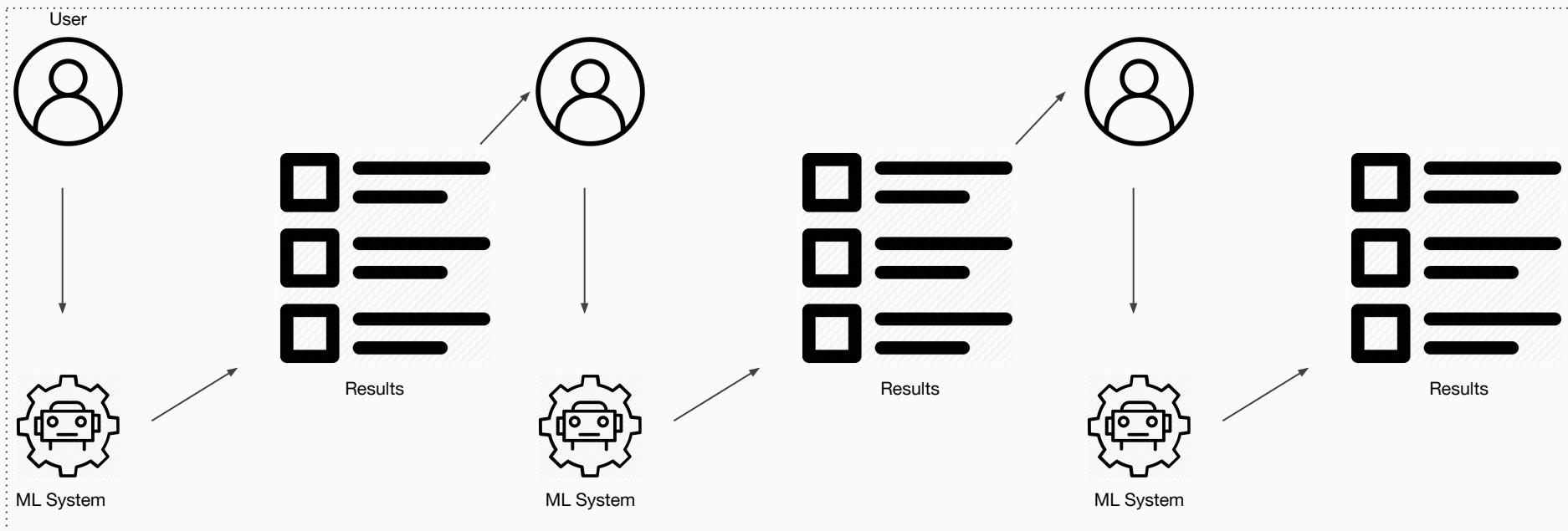


*long-term*

# Session Level Evaluation: Introduction

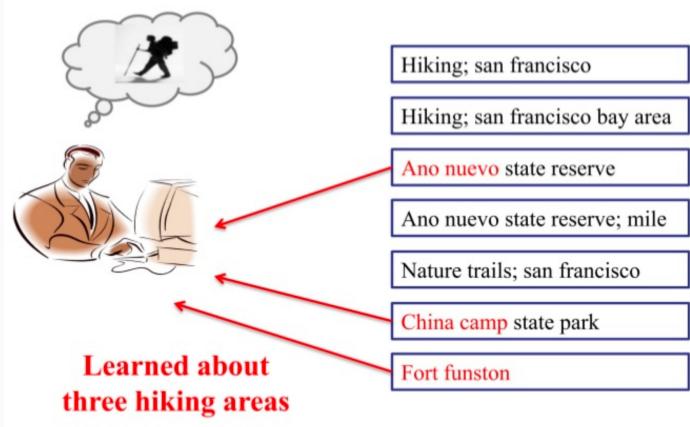


# Session Level Evaluation: Introduction

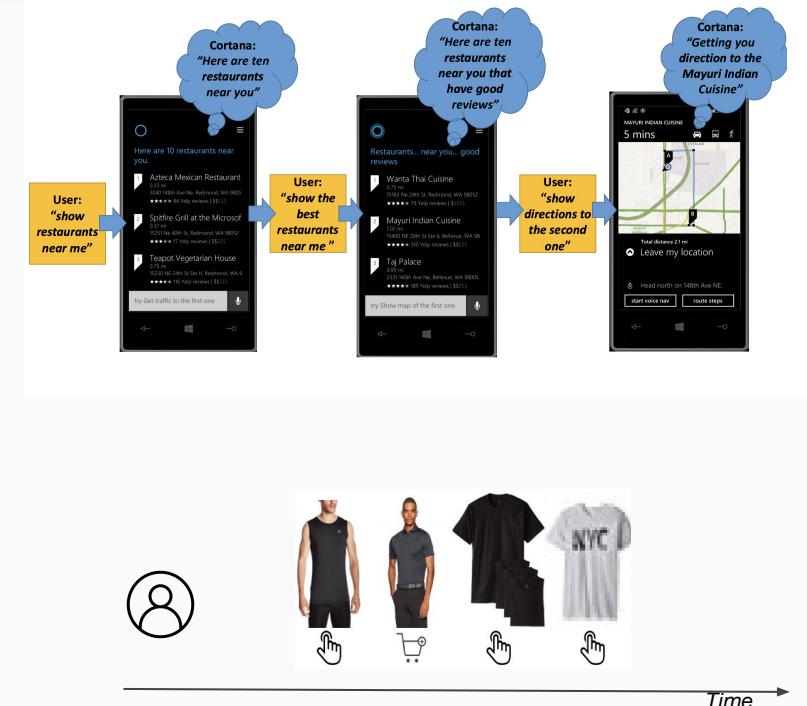
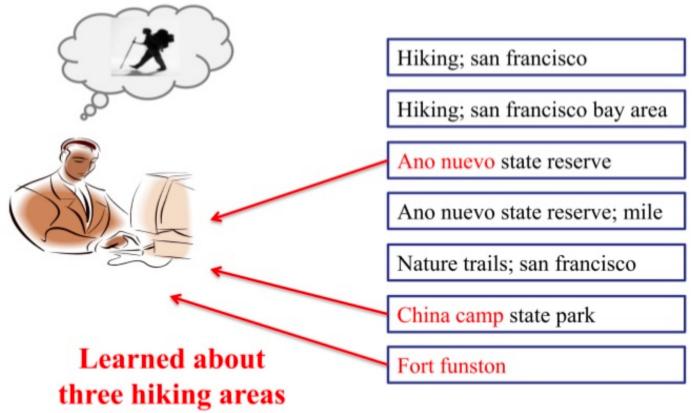


***How do we evaluate the user experience over a session?***

# 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.

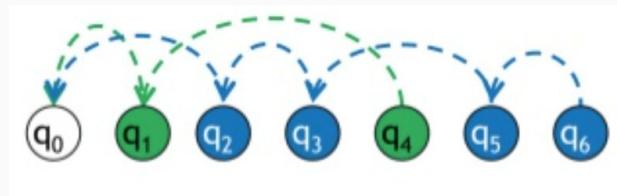
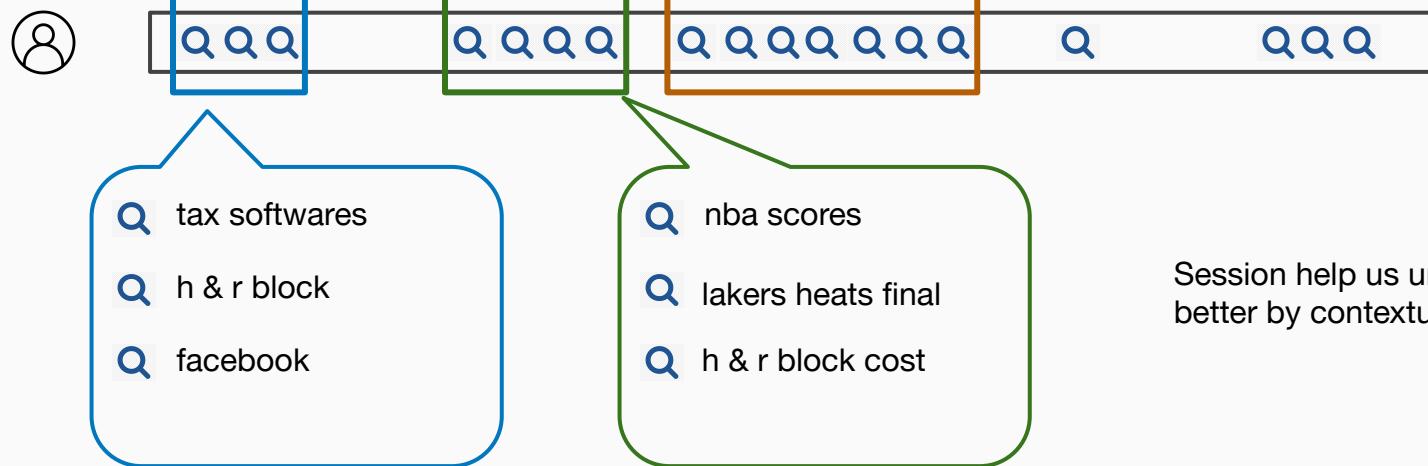
Malte Ludewig and Dietmar Jannach. **Evaluation of session-based recommendation algorithms**. User Modeling and User Adapted Interaction 2018.

# Identifying Sessions from Logs



time between requests  
could be reasonable  
proxies to identify sessions

# Identifying Sessions from Logs



# Sessions → User Goals

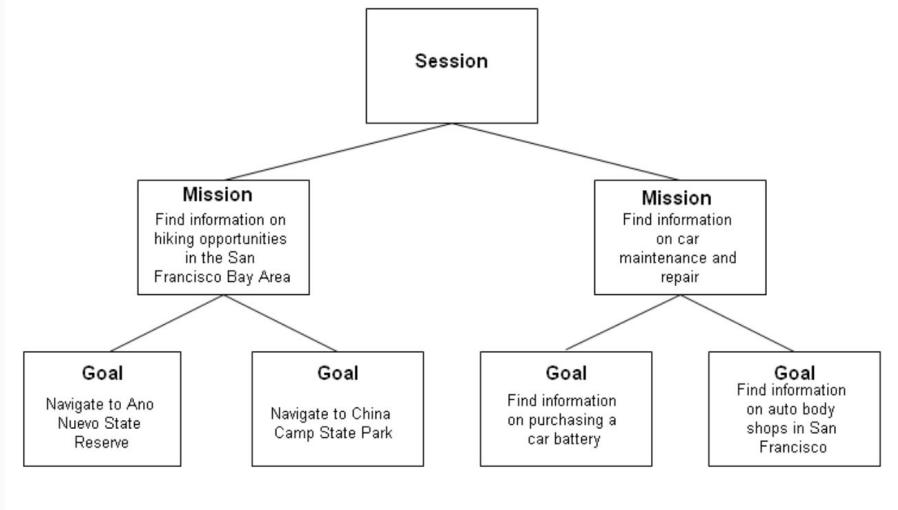
## User Sessions

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](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...](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

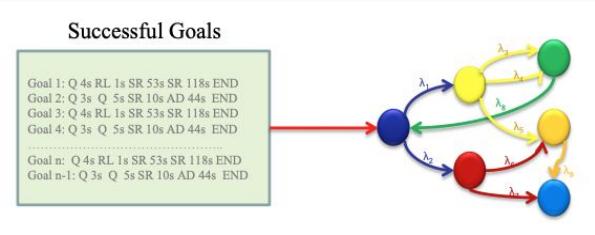
- Query** career development advice
- Click** <http://www.soperarticles.com/business-articles/career-devel...>
- Query** employment issues articles
- Click** <http://jobseekeradvice.com/category/employment-issues/...>
- Query** professional career advice
- Click** <http://ezinearticles.com/?Career-Advice-and-Professional-Ment...>
  - 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

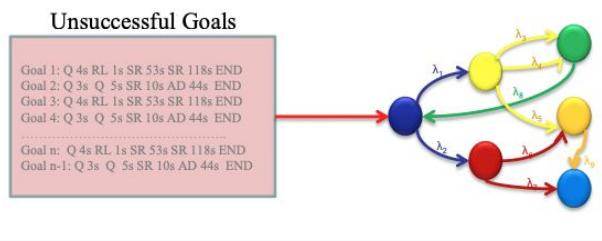


Figure Source: [Understanding & Inferring User Tasks and Needs](#)

# 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

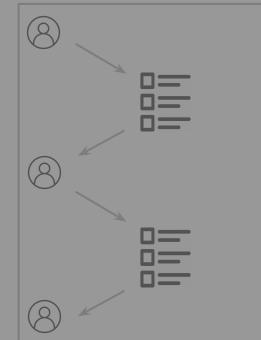
*short-term*



A screenshot of the Shazam mobile application interface. At the top, there's a dark header with the Shazam logo and a play button icon. Below the header, a large blue rectangular box contains the text "Your Burning Mix" in white, with "500 BPM" underneath it. In the bottom right corner of this box, there's a small yellow triangle icon. The background of the app shows a blurred image of a person's face.



A screenshot of the 'Editor's picks' section on the National Geographic website. The page features a grid of twelve travel-related articles, each with a small thumbnail image and a title. The titles include: 'Travel: The Best Cities in Asia', 'How to Get the Most Out of Your Trip', 'National Geographic Traveler: Your Guide to the World', 'Travel: The Best Beaches in the Americas', 'National Geographic Traveler: Your Guide to the World', 'Travel: The Best Beaches in the Americas', and 'Travel: The Best Beaches in the Americas'. The layout is clean and organized, with a white background and a clear grid structure.



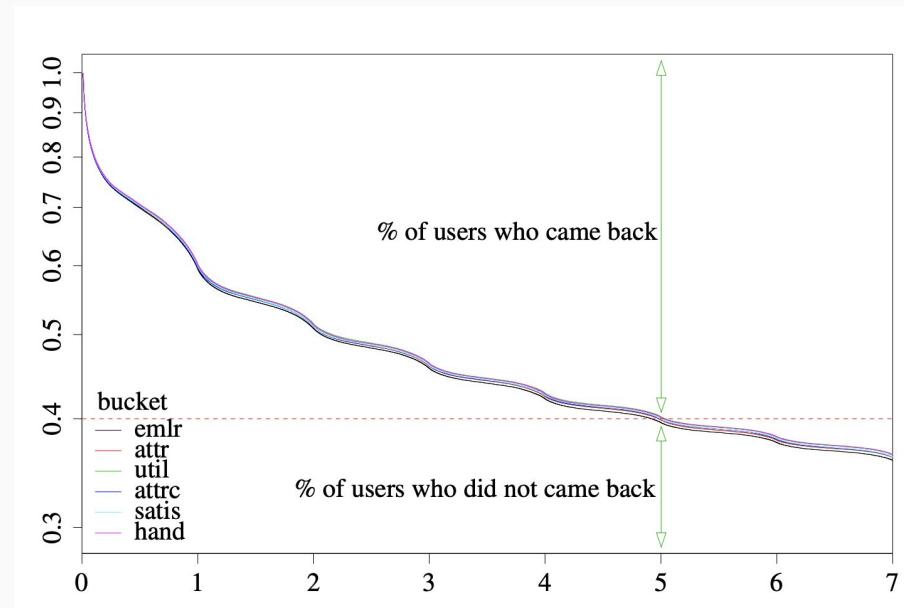
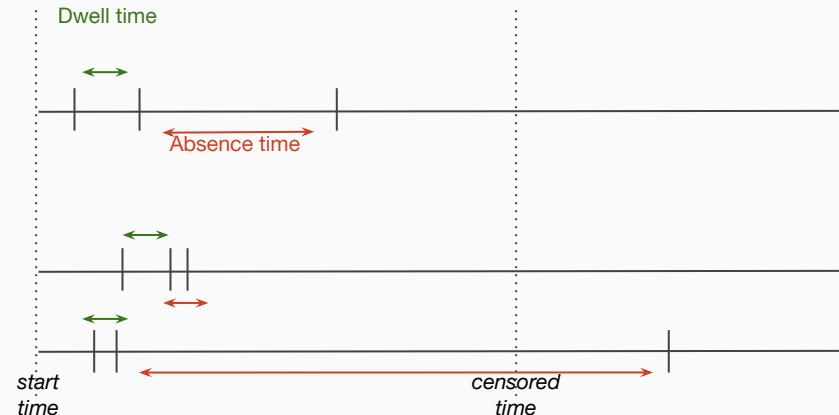
*long-term*

# 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

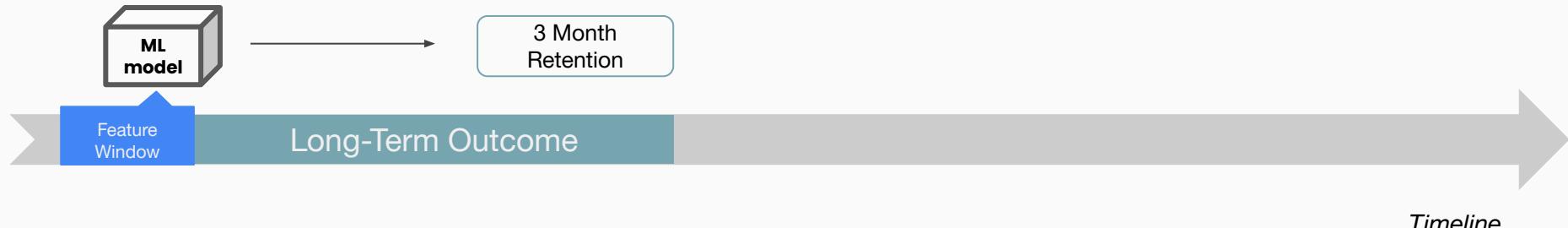
- Satisfied users coming back to the service and re-engaging with the ML system is a reasonable proxy for long term success.
- One or more short-term metrics such as Clickthrough Rate, Session success can be used as proxies to estimate long-term outcomes (e.g. user retention).



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

# Multiple Metrics

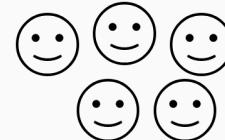
Brian St. Thomas, Spotify

# Section Outline

- Decisions with multiple metrics
- Role of surrogate functions
- Statistical surrogacy
- Examples of statistical surrogates
- Comparing sensitivity

# Single Metric Decisions

**Objective:** System maker wants to make a tool to help individuals with their goals



The system maker can choose the system that performs the best according to a single metric



Ideal metrics are **STEDI** (Sensitive, Trustworthy, Efficient, Debuggable, Interpretable)

All emojis designed by [OpenMoji](#) – the open-source emoji and icon project. License: [CC BY-SA 4.0](#)

E. Jäger. <https://openmoji.org/library/#group=smileys-emotion&emoji=1F60E>  
M. Wahl. <https://openmoji.org/library/#search=machine&emoji=1F4E0>  
E. Jäger. <https://openmoji.org/library/#group=smileys-emotion&emoji=1F642>  
V. Boutzikoudi. <https://openmoji.org/library/#search=cloud&emoji=2601>

**Somit Gupta et al. Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.** KDD 2019.

# Multiple Metrics Decision

**Objective:** System maker wants to make a tool to help individuals with their goals

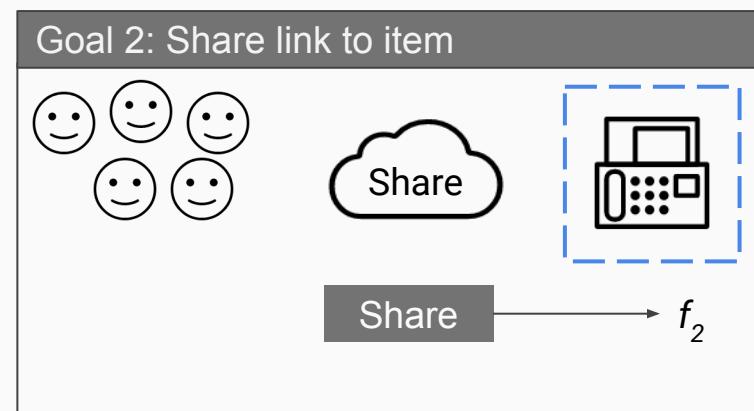
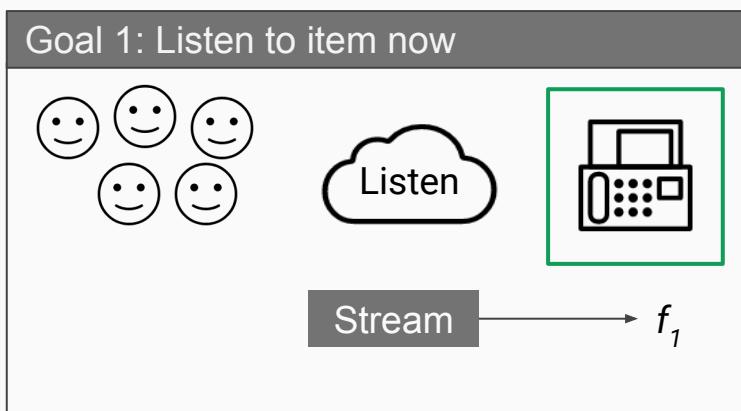


The system maker may have multiple metrics that reflect different aspects of how much the system is helping the user. For example,



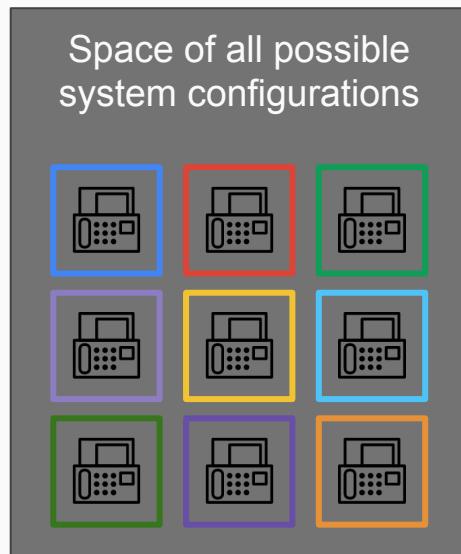
# Multiple Metrics - Example

Users of an audio content search feature can have multiple goals, measurable by different metrics ([Hosey et al, 2019](#))



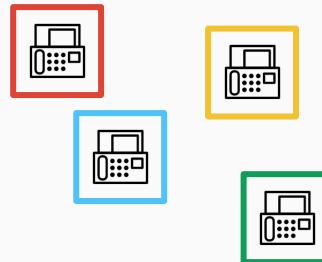
# Multiple Metrics in Evaluation

The system maker wants to ensure that the tool helps individuals with their goals in the different ways reflected by each metric



Metric  $f_1$   
[small values  
are better]

Metric values on evaluation set or population  
(not stochastic)

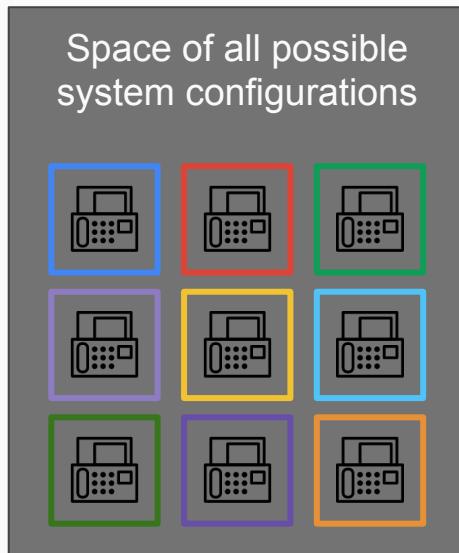


# Multiple Optimization Problem (MOP)

The system maker wants to choose the best system  $\mathbf{x}$  such that

$$\text{minimize } f_1(\mathbf{x}), \dots, \text{minimize } f_m(\mathbf{x}), \mathbf{x} \in \mathcal{X}$$

$\mathcal{X}$

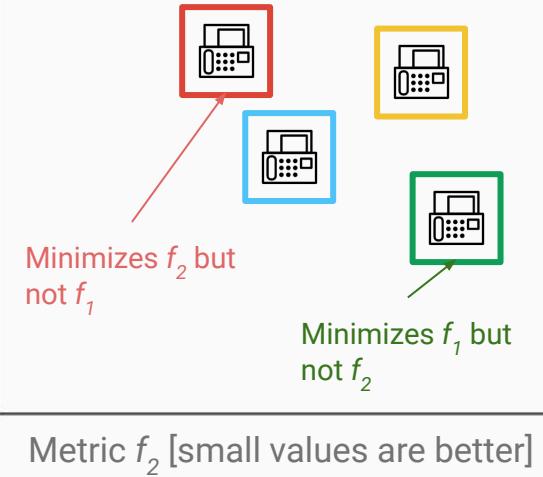


Metric  $f_1$   
[small values are better]

Instances of System created

Onto, but not one-to-one

Metric values on evaluation set or population  
(not stochastic)



# Pareto Dominance

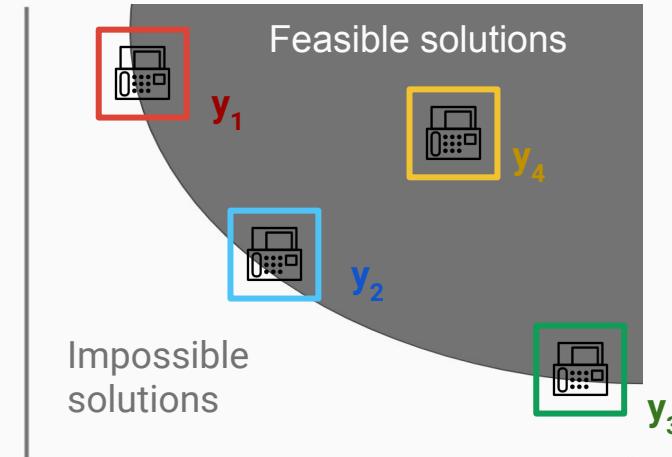
A solution  $y_1 = [f_1(x), f_2(x), \dots, f_M(x)]$  pareto dominates another solution  $y_2$  if

- $y_1[i] \leq y_2[i]$  for all  $i$  in  $1:M$  and ...
- ... there exists  $j$  such that  $y_1[j] < y_2[j]$ .

The set of points that are not pareto dominated by any other point is the pareto (optimal) front.

In this setting,  $y_1$ ,  $y_2$ , and  $y_3$  are on the optimal front because there are no feasible systems with better  $f_1$  and  $f_2$  values.

Metric  $f_1$   
[small values  
are better]

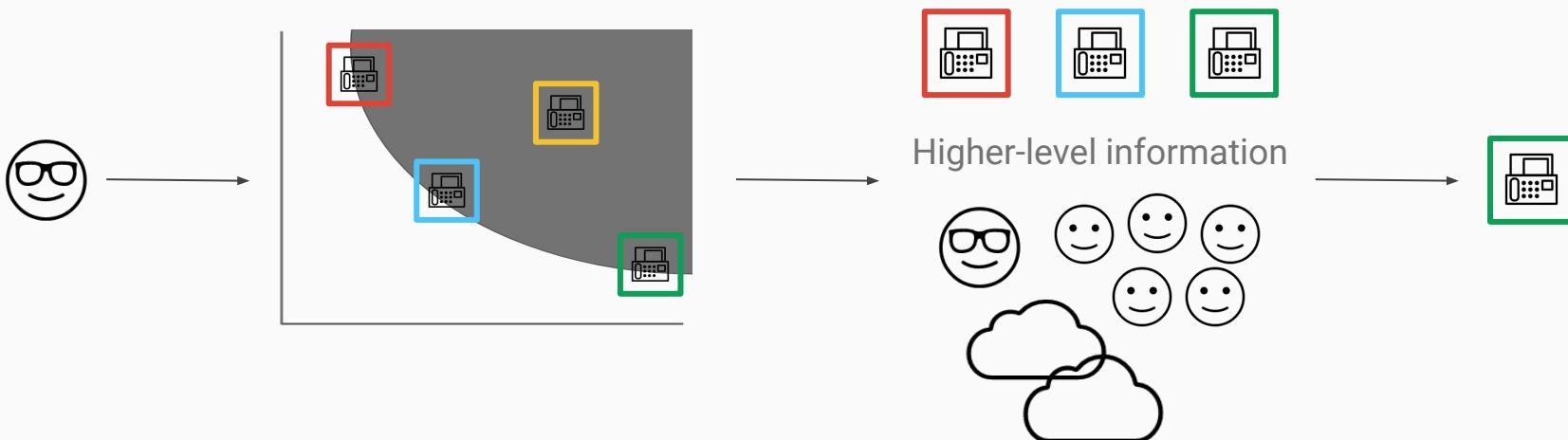


Metric  $f_2$  [small values are better]

# Principles of Optimizing with Multiple Metrics

[Deb 2011](#) highlights two steps in finding solutions to (evolutionary) MOPs

1. Find multiple non-dominated points as close to the pareto front as possible that reflect a wide trade-off among objectives
2. Choose one of the obtained points using *higher-level information*.



## *Higher-Level Information* for Human-ML Systems

Higher level information to choose the preferred system can come from a variety of sources



System maker chooses based on *a priori* estimates of user preferences

## Higher-Level Information for Human-ML Systems

Higher level information to choose the preferred system can come from a variety of sources



System maker chooses based on *a priori* estimates of user preferences



User Satisfaction, measured through survey or evaluative feedback  
([Schnabel et al 2019](#), [Garcia-Gathright et al 2018](#))

# Higher-Level Information for Human-ML Systems

Higher level information to choose the preferred system can come from a variety of sources



System maker chooses based on *a priori* estimates of user preferences



User Satisfaction, measured through survey or evaluative feedback  
([Schnabel et al 2019](#), [Garcia-Gathright et al 2018](#))



Conditioned on a user having a goal that the system can help with, does the user engage with the system again (e.g. [Hohnhold et al 2015](#), [Dmitriev et al 2016](#), [Poyarkov et al 2016](#), [Machmouchi et al 2017](#))

*Henning Hohnhold et al. Focusing on the Long-term: It's Good for Users and Business.* KDD 2015.

*Pavel Dmitriev et al. Pitfalls of Long-Term Online Controlled Experiments.* IEEE Big Data 2016.

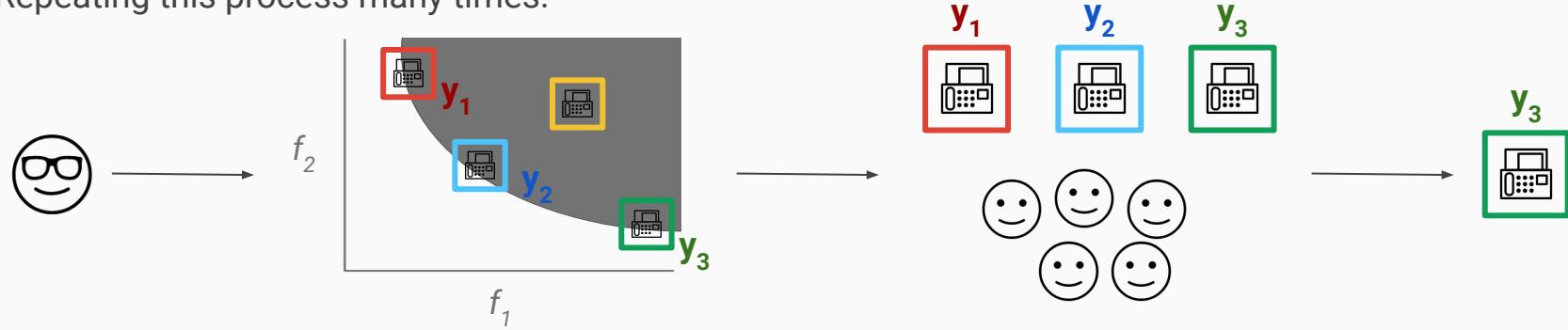
*Alexey Poyarkov et al. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments.* KDD 2016.

*Widad Machmouchi et al. Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments.* CIKM 2017.

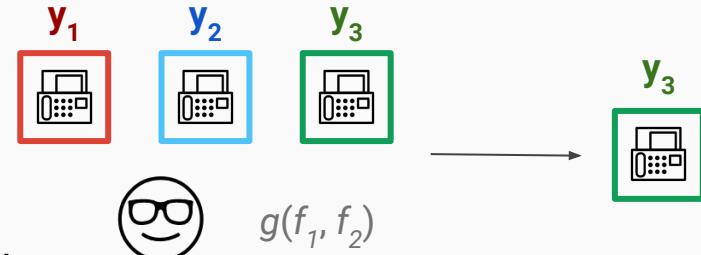
# Surrogate Problems Approximate Expensive Higher-Level Information

Collecting the ideal decision criteria can be either expensive or slow, so instead it's useful to construct a *surrogate problem* (e.g. [Allmendinger et al 2017](#))

Repeating this process many times:



We can learn a surrogate function  $g$ , such that  $g(f_1, f_2)$  is largest when the decision maker has highest preference for the corresponding system.



# Statistical Surrogacy

[Athey et al 2019](#) shows the conditions under which an effect on the surrogate function is an unbiased estimate of an effect on the decision criteria.

*Statistical Surrogacy:*

$$P(Y, W | S) = P(Y | S) P(W | S)$$

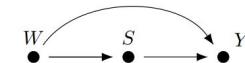
With many metrics, they are likely to lie near the causal chain between the difference in systems and the decision criteria

FIGURE 1  
Surrogacy Assumptions and Violations

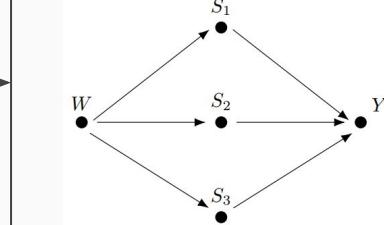
A. Surrogacy Assumption Satisfied



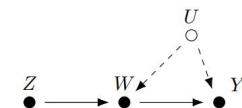
B. Violation of Surrogacy due to Direct Effect



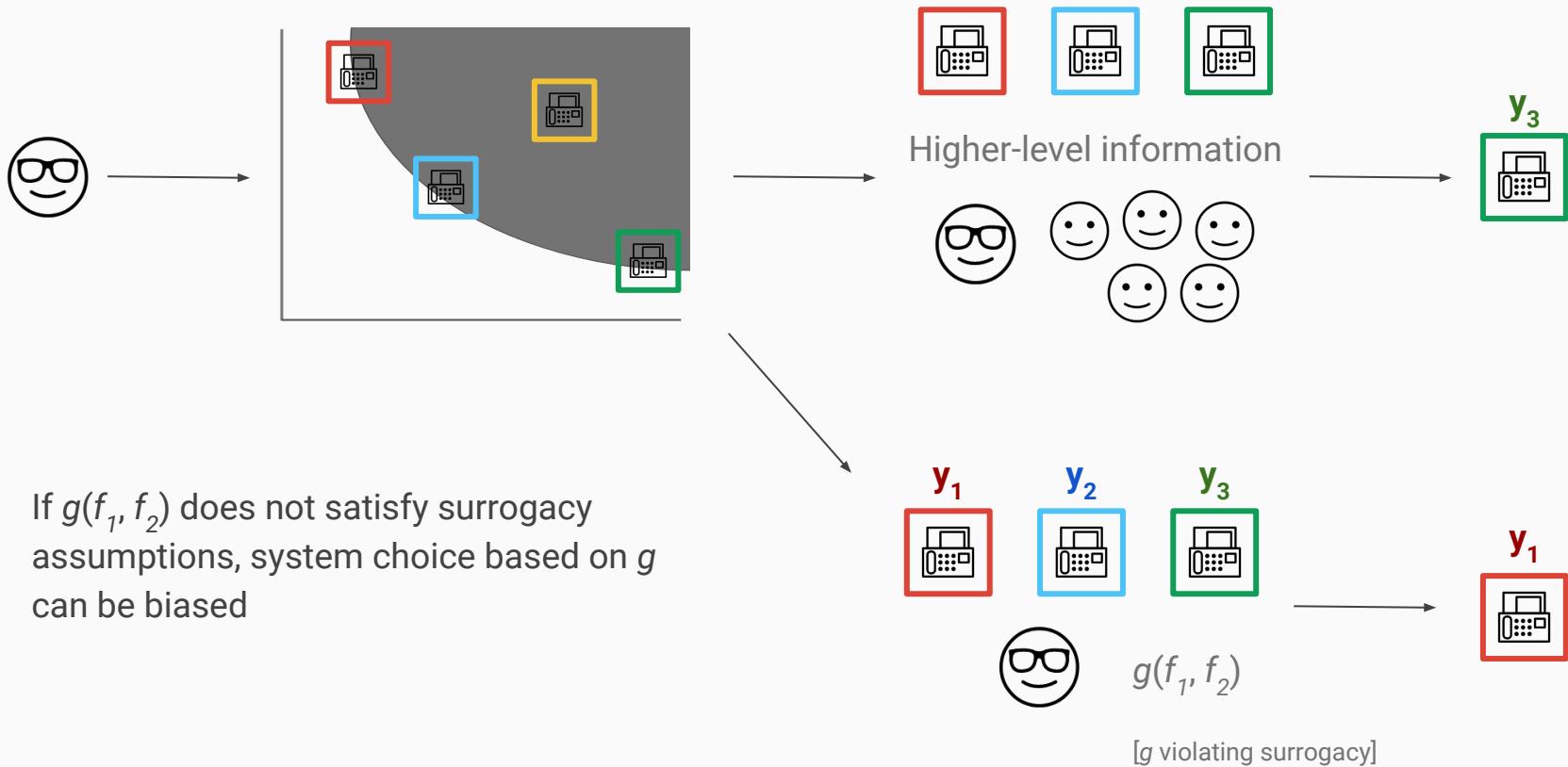
C. Multiple Surrogates



D. Instrumental Variable



# Risks of Violating Surrogacy Assumptions



# Linear Combination Surrogate Functions

Solving MOPs is worth its own tutorial (e.g. [Emmerich & Deutz, 2018](#)), but a consequence makes linear combinations a useful functional form for surrogate functions.

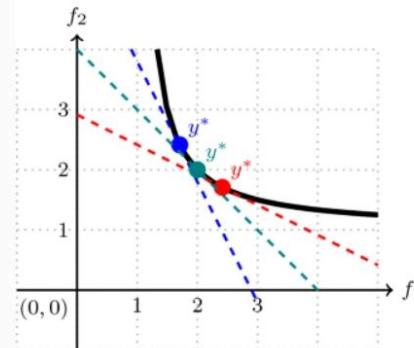
Solutions to

$$\text{minimize } \sum_{i=1}^m w_i f_i(x), x \in \mathcal{X}.$$

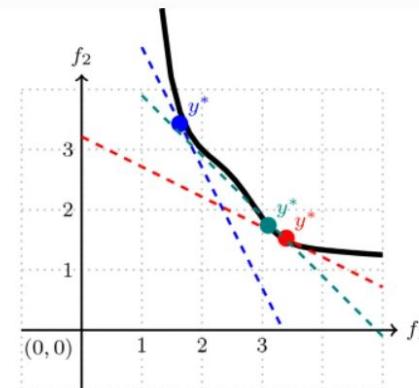
will always return points on the concave portions of the pareto front (however not convex portions)

Concave pareto front

Fig. 2

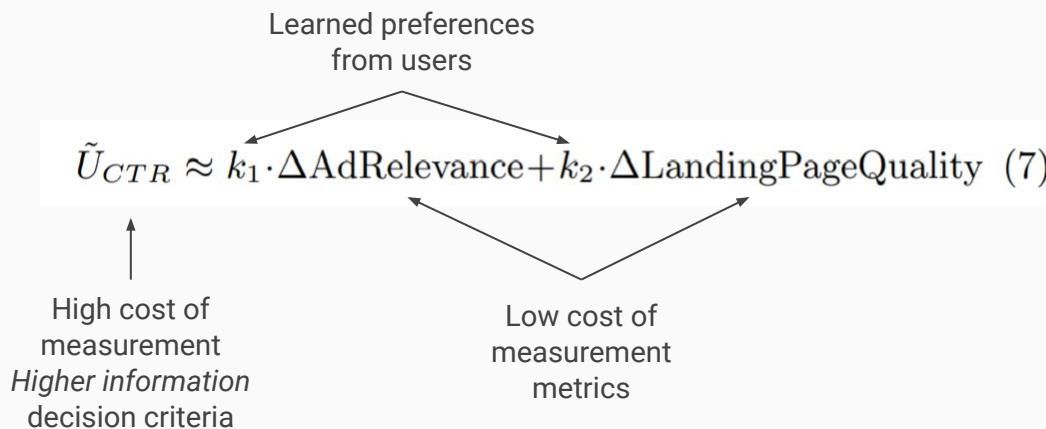


Convex pareto front



# Example: Meta-Analysis of Google Web Search Experiments

In [Hohnhold et al 2015](#), long-term decision criteria for web search is approximated using a surrogate of the form



[Dmitriev et al. 2016](#) discuss challenges in collecting long-term decision criteria for this type of analysis

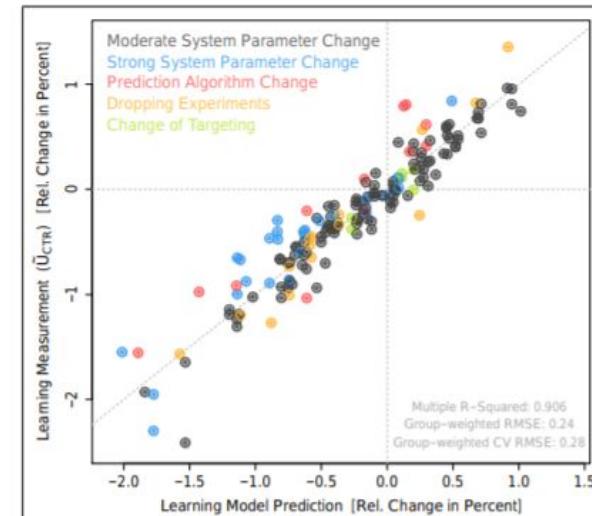


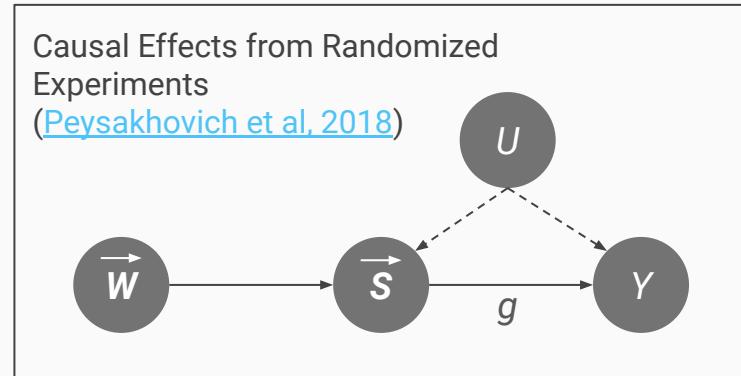
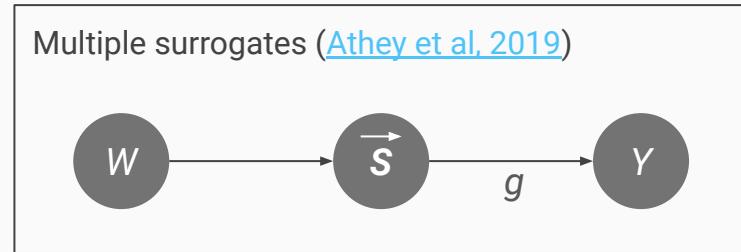
Figure 4: Measured vs. predicted learning for the current desktop macro-model.

# Meta-modeling to construct surrogate functions

[Peysakhovich et al, 2018](#) harvest instrumental variables from treatment assignments in many randomized experiments to learn causal relationships between low-cost of measurement metrics and decision criteria metrics

These causal relationships can be used to construct the surrogate function  $g$

This work is extended by [Wang et al, 2020](#) to create an estimate from the summary statistics of the experiments.



# Metric Sensitivity

Sensitivity of a metric is the ability to detect differences between the underlying systems, given that a true difference exists.



Sensitivity of input metrics impact the sensitivity of the surrogate function, but is also an important property of the surrogate function itself ([Deng et al 2016](#), [Kharitonov et al 2017](#))

# Components of Metric Sensitivity

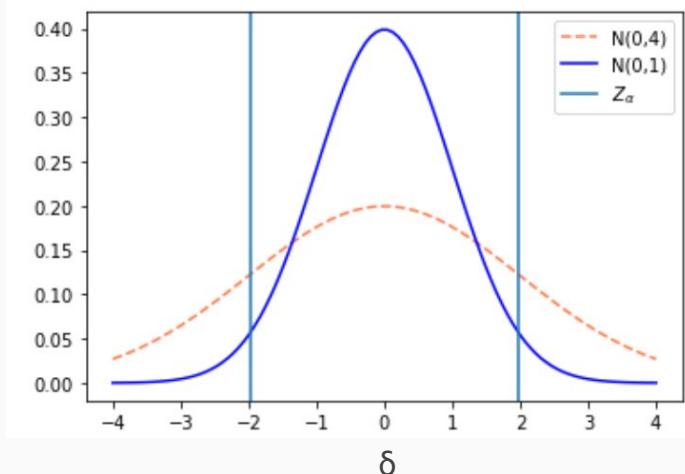
Using the decomposition of [Deng 2015](#), the probability that there is a difference on the metric *and* that the difference is detected is

$$P(H_1) \times P(Z > Z_\alpha | H_1)$$

For

- test statistic  $Z$
- critical value  $Z_\alpha$
- Alternative hypothesis  $H_1$  for generating treatment effects
  - [Deng 2015](#) learn empirical prior on  $P(H_1)$  and parameters of  $H_1$  using historical online experiments

$H_1: \delta \sim N(0,4)$  corresponds to a more sensitive metric than  $H_1: \delta \sim N(0,1)$



True difference in metric between  
systems

# Offline Metric Sensitivity

Metric sensitivity for offline metrics can be measured through using [Efron & Tibshirani 1986](#) Bootstrap Hypothesis Testing (e.g. [Sakai 2006](#), [Savoy 1997](#))

Bootstrap Hypothesis Tests can be constructed around the desired metric and the Null hypothesis there is no difference between pairs of systems.

Conduct Bootstrap Hypothesis Tests between pairs of different systems using each metric, and compare the rate at which differences occur

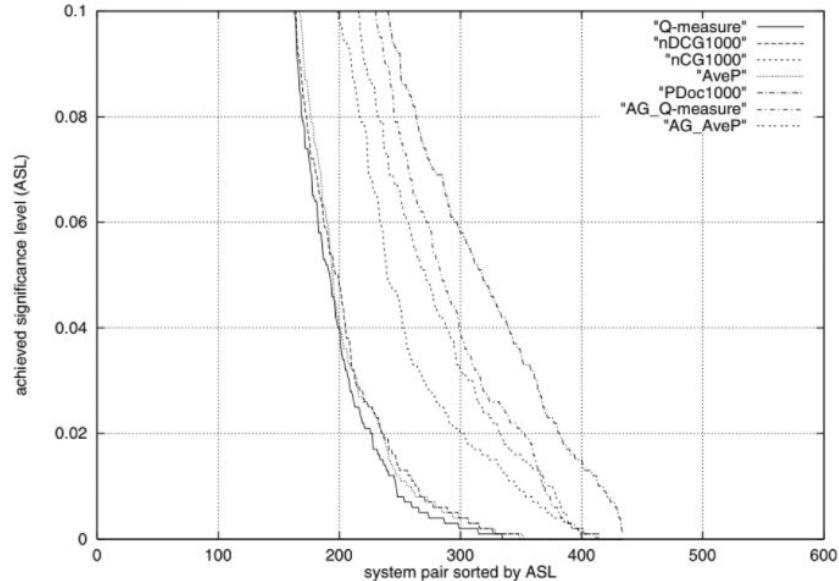


Figure 7: ASL curves based on Paired Bootstrap Hypothesis Tests (NTCIR-3 CLIR Chinese data).  
[Sakai 2006](#)

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric	P(True Movement)	$V^2/(1/N_E)$	P(Detect True Movement)
Count Metric A(whole page)	3%	25.1	2.1%
Count Metric A(sub region)	17%	37.2	12.8%
Metric B	31%	2.8	9.8%
Metric B with Utility Weights	45%	3.3	15.6%
Metric C	12%	9.2	5.9%
Metric C with VR	11%	19.7	8.0%
Metric D	13%	29.3	9.4%
Metric D Capped	16%	38.6	12.1%

**Table 1: Examples of detailed sensitivity decomposition. Results align with theory and experience.**

[Deng et al 2016](#)

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric functions can be optimized directly for sensitivity ([Kharitonov et al 2017](#)) given historical A/B and A/A tests

$$J(w) = \frac{1}{|\mathbb{E}|} \sum_{e \in \mathbb{E}} D(w; A_e, B_e) - \alpha \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} |D(w; A_c, B_c)| + \beta \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} |D(w; A_u, B_u)|$$

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric functions can be optimized directly for sensitivity ([Kharitonov et al 2017](#)) given historical A/B and A/A tests

Maximize difference  
without changing sign over  
experiments with known  
effect



$$J(w) = \frac{1}{|\mathbb{E}|} \sum_{e \in \mathbb{E}} D(w; A_e, B_e) - \alpha \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} |D(w; A_c, B_c)| + \beta \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} |D(w; A_u, B_u)|$$

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric functions can be optimized directly for sensitivity ([Kharitonov et al 2017](#)) given historical A/B and A/A tests

Maximize difference without changing sign over experiments with known effect

Minimize difference over A/A tests



$$J(w) = \frac{1}{|\mathbb{E}|} \sum_{e \in \mathbb{E}} D(w; A_e, B_e) - \alpha \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} |D(w; A_c, B_c)| + \beta \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} |D(w; A_u, B_u)|$$

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric functions can be optimized directly for sensitivity ([Kharitonov et al 2017](#)) given historical A/B and A/A tests

Maximize difference without changing sign over experiments with known effect

Minimize difference over A/A tests

$$J(w) = \frac{1}{|\mathbb{E}|} \sum_{e \in \mathbb{E}} D(w; A_e, B_e) - \alpha \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} |D(w; A_c, B_c)| + \beta \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} |D(w; A_u, B_u)|$$

Maximize difference over experiments with unknown effect

# Online Metric Sensitivity

Online metrics can be compared through sensitivity decomposition ([Deng 2015](#)) using historical experiments.

Metric functions can be optimized directly for sensitivity ([Kharitonov et al 2017](#)) given historical A/B and A/A tests

Pre-experiment data can be used to improve sensitivity ([Deng et al 2013](#), [Ashkan & Metzler 2019](#))

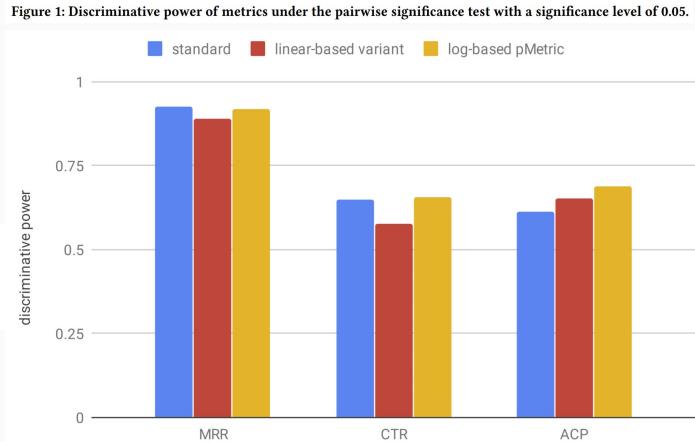
Standard metric:

$$M = \frac{1}{n} \sum_{i=1}^n M_i$$

Personalized metric:

$$w_i = f(\bar{s}_i, s_i, \rho)$$

$$pM = \frac{\sum_{i=1}^n w_i M_i}{\sum_{i=1}^n w_i}$$



# Additional Reading Materials and References

- Richard Allmendinger et al. Surrogate-Assisted Multicriteria Optimization: Complexities, Prospective Solutions, and Business Case.* J. of Multi-Criteria Decision Analyses 2017.
- Azin Ashkan & Donald Metzler. Revisiting Online Personal Search Metrics with the User in Mind.* SIGIR 2019.
- 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.
- Kalyanmoy Deb. Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction.* In L. Wang (Ed.), A. Ng (Ed.), K. Deb (Ed.) *Multi-Objective Evolutionary Optimization for Product Design and Manufacturing*. 2011.
- Alex Deng et al. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data.* WSDM 2013.
- Alex Deng. Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments.* WWW 2015
- Alex Deng & Xiaolin Shi. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned.* KDD 2016.
- Pavel Dmitriev et al. Pitfalls of Long-Term Online Controlled Experiments.* IEEE Big Data 2016.
- Bradley Efron & Robert Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.* Stat. Sci. 1986.
- Michael T. M. Emmerich & André Deutz. A tutorial on multiobjective optimization: fundamentals and evolutionary methods.* Natural Computing, 2018.
- Jean Garcia-Gathright et al. Understanding and Evaluating User Satisfaction with Music Discovery.* SIGIR 2018.
- Somit Gupta et al. Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.* KDD 2019.
- Henning Hohnhold et al. Focusing on the Long-term: It's Good for Users and Business.* KDD 2015.
- Christine Hosey et al. Just Give Me What I Want: How People Use and Evaluate Music Search.* CHI 2019.
- Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. Learning Sensitive Combinations of A/B Test Metrics.* WSDM 2017.
- Widad Machmouchi et al. Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments.* CIKM 2017.
- Alexander Peysakhovich & Dean Eckles. Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables.* WWW 2018.
- Alexey Poyarkov et al. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments.* KDD 2016.
- Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap.* SIGIR 2006.
- Jacques Savoy. Statistical Inference in retrieval effectiveness evaluation.* Information Processing & Management 1997.
- Tobias Schnabel et al. Shaping Feedback Data in Recommender Systems with Interventions Based on Information Foraging Theory.* WSDM 2019.
- Zenan Wang et al. Causal Meta-Mediation Analysis: Inferring Dose-Response Function From Summary Statistics of Many Randomized Experiments.* KDD 2020.

# Open problems

# challenges in developing user-focused metrics

# challenges in learning with user-focused metrics



GRAVEYARD



# Novelty and Diversity in Rankings

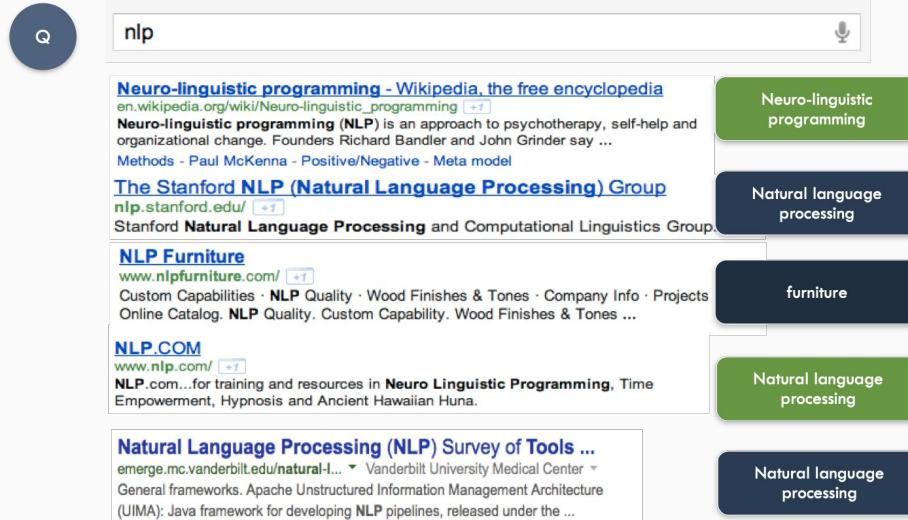
- **Diversification of Results** increases the chance of users satisfying their specific need and lead to higher utility by reducing redundancy.
- Different Types of Diversity

- **Extrinsic Diversity**

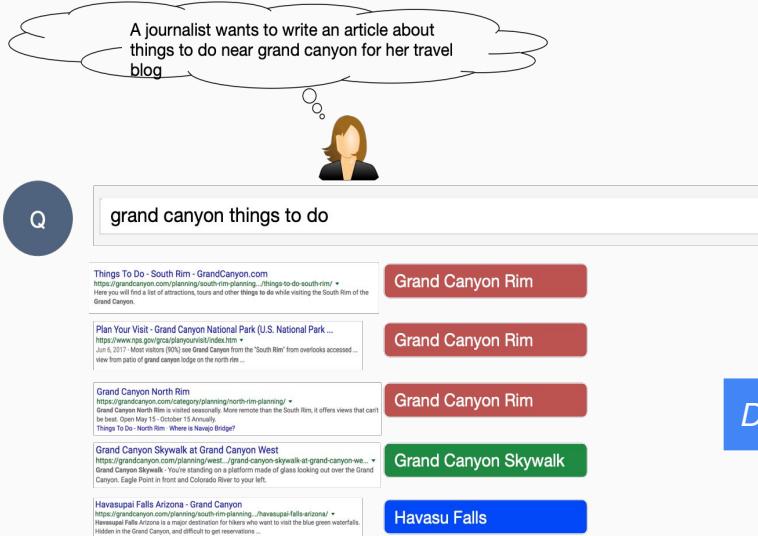
User goals are uncertain from the request and the system diversifies results to maximize success.

Example: ambiguous queries in search

- **Intrinsic Diversity**



# Novelty and Diversity in Rankings



## *Intrinsic Diversity*

Addresses the need to reduce redundancy even when user requests are clearly specified.

## Examples: exploratory search

## *Diversified Ranking*



# Novelty & Diversity: Subtopic-Based Evaluation

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.

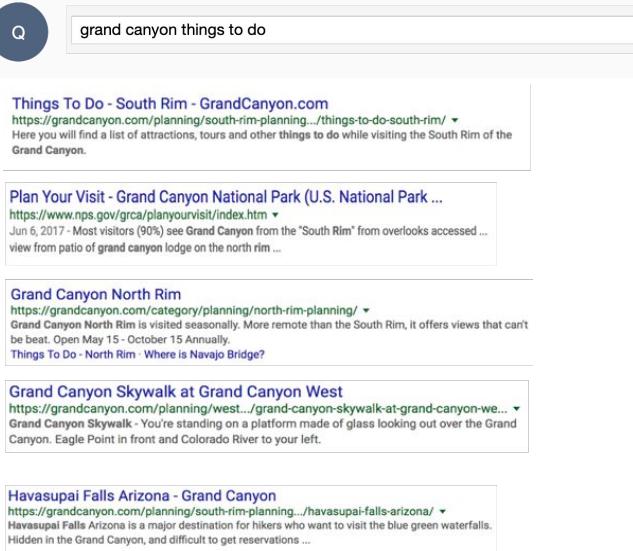
Plan Your Visit - Grand Canyon National Park (U.S. National Park ...  
<https://www.nps.gov/grca/planyourvisit/index.htm> ✓  
Jun 6, 2017 - Most visitors (90%) see Grand Canyon from the "South Rim" from overlooks accessed ... view from patio of grand canyon lodge on the north rim ...

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.

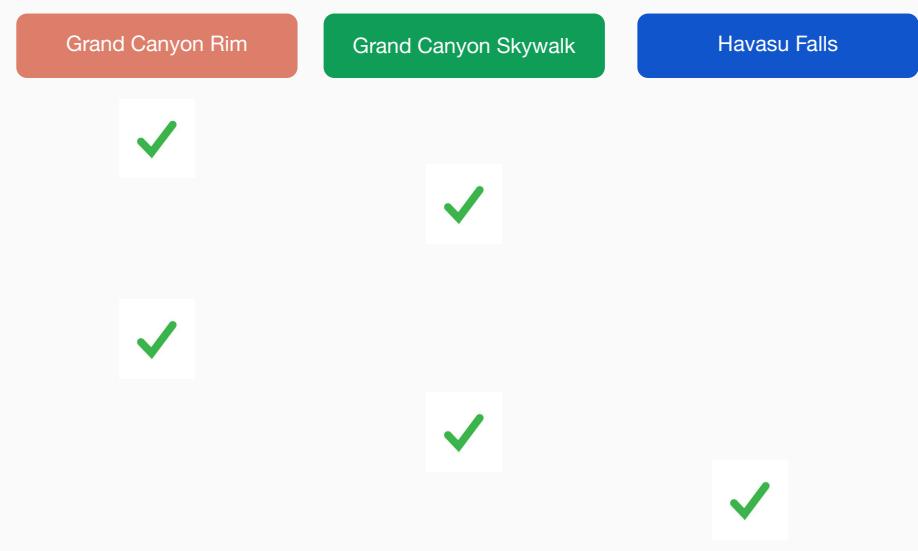
Havasupai Falls Arizona - Grand Canyon  
<https://grandcanyon.com/planning/south-rim-planning.../havasupai-falls-arizona/> ✓  
Havasupai Falls Arizona is a major destination for hikers who want to visit the blue green waterfalls. Hidden in the Grand Canyon, and difficult to get reservations ...

# Novelty & Diversity: Subtopic-Based Evaluation



A variation of DCG metric that rewards novelty and penalizes diversity

→

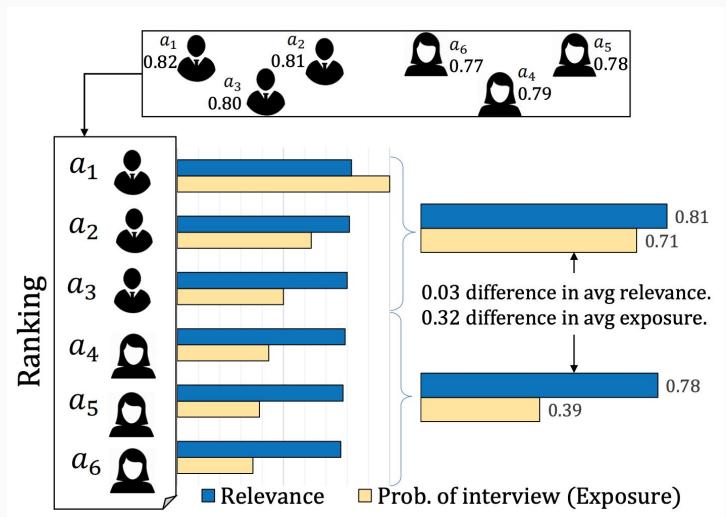

$$\text{DCG}[k] = \sum_{j=1}^k G[j] / (\log_2(1 + j))$$
$$G[k] = \sum_{i=1}^m J(d_k, i) (1 - \alpha)^{r_{i,k-1}}$$

# Fairness in Rankings

Image Search Results for “CEO”



small differences in predicted success scores  
can lead to a large difference in exposure

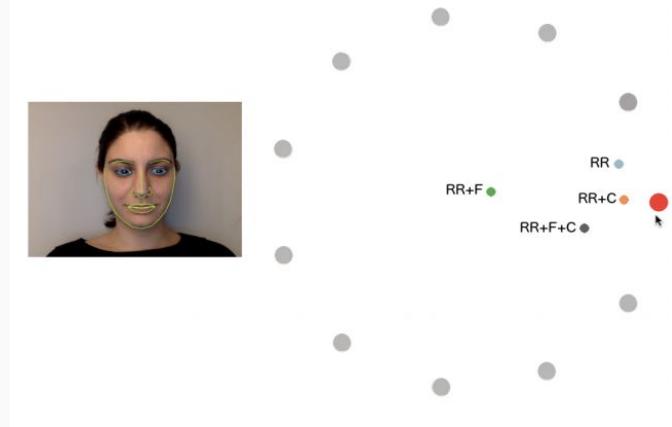


# Validating Click Feedback

Also address click bait here ?

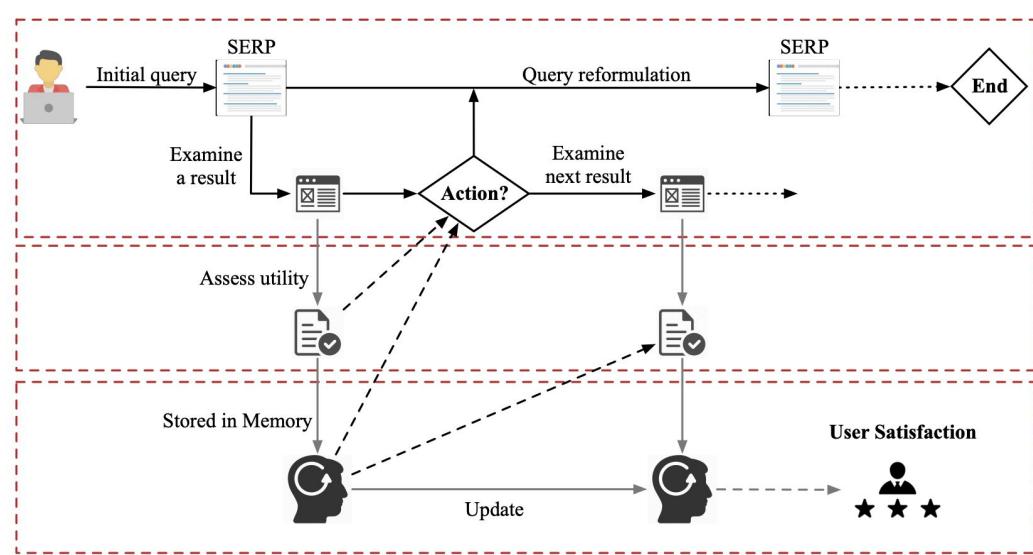
# Combining Cursor Movements & Dwell Time

<i>dwell</i> : time of the page view in seconds	0.167**
<i>rank</i> : the rank of the document or the rank of the origin (i.e., the landing page) of the search trail that the document is on if its rank is not available	-0.073
<i>cursorcnt</i> : num. of cursor movements	0.164**
<i>cursorfreq</i> : cursorcnt/dwell	-0.082*
<i>dist</i> : total overall distance the cursor traveled in pixels	-0.137**
<i>xdist</i> : total distance the cursor traveled horizontally in pixels	0.101**
<i>ydist</i> : total distance the cursor traveled vertically in pixels	0.172**
<i>speed</i> : dist/dwell	0.101**
<i>xspeed</i> : xdist/dwell	-0.143**
<i>yspeed</i> : ydist/dwell	-0.124**
<i>xmin</i> : minimal x coordinate	0.112**
<i>ymin</i> : minimal y coordinate	0.093*
<i>xmax</i> : maximal x coordinate	0.067
<i>ymax</i> : maximal y coordinate	0.243**
<i>xrange</i> : xmax-xmin	-0.006
<i>yrange</i> : ymax-ymin	0.172**



# Directionally Explicit Signals: Lab Studies

- Simulate user experience in a lab setting
- Collect data at various granularities



# intent-aware metrics

**user model:** in-order traversal of a ranked list; utility dependent on subtopic of higher ranked documents; independent of stopping probability.

**metric:** expected utility given simulated user behavior.

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

$\mathcal{A}_q$  aspects for query  $q$

$y^a$  document relevance to aspect  $a$

# session metrics

**user model:** sequence of queries .

**metric:** expected utility given  
simulated user behavior.

$$\sum_{t=1}^T \tau_t \times \text{metric}(y, \pi^{q_t})$$

$T$  number of queries in session

$\tau$  query sequence discount

$\pi^q$  ranking for query  $q$