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

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. RecSys, 2010.

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

Feifan Liu and Yang Liu. Correlation between ROUGE and human evaluation of extractive meeting summaries. ACL, 2008

why does the right metric matter?

task: recommendation

metric: user clicks

> 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.

Smitha Milli, Luca Belli, and Moritz Hardt. From optimizing engagement to measuring value. 2020.

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.

Karen Spärck Jones and Julia R. Galliers. Evaluating Natural Language Processing Systems: An Analysis and Review,, 1995.

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

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

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

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

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

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

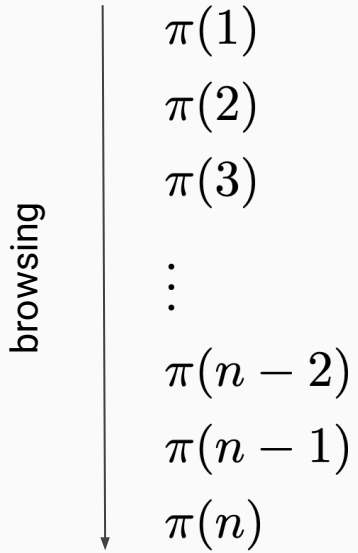
Cyril W. Cleverdon, Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, 1962

metric: recall

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

Cyril W. Cleverdon, Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, 1962

ranking



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.

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

$$\begin{aligned} \text{min-}k &\quad k\text{th smallest value} \\ \overline{\pi}(i) &\quad \text{rank position of item } i \end{aligned}$$

William S. Cooper. Expected search length,. American Documentation, 19(1):30–41, 1968.

metric: R-precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision when recall is 1.

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

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

metric: R-precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision when recall is 1.

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

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.

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

E. Voorhees and D. Tice. The trec-8 question answering track evaluation. TREC, 1999.

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.

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

E. Voorhees and D. Tice. The trec-8 question answering track evaluation. TREC, 1999.

metric: average precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision averaged over all recall levels.

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

metric: average precision

user model: in-order traversal of a ranked list, collecting all relevant items.
metric: precision averaged over all recall levels.

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

$$= \frac{1}{|\mathcal{Y}^+|} \sum_{r=1}^{|\mathcal{Y}^+|} \frac{r}{\mathrm{ESL}(\mathcal{Y}^+, \pi, r)}$$

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.

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

γ patience parameter

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. TOIS, 2002.
Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. TOIS, 2008.

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.

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

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

Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. SIGIR, 2009.

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.

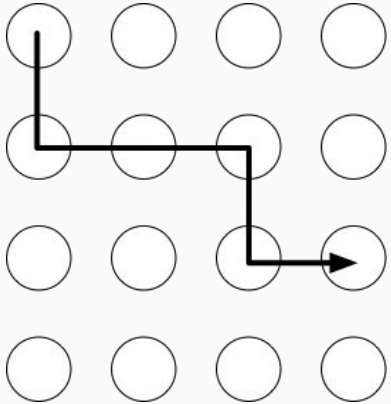
$$\mathrm{TBG}(y, \pi) = \sum_{r=1}^{n} y_{\pi(r)} \Delta(y, \pi, r)$$

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

Mark D. Smucker and Charles L.A. Clarke. Time-based calibration of effectiveness measures. SIGIR, 2012.

grid metrics

user model: model of ordered user traversal of grid items; utility independent of stopping probability.
metric: expected utility given simulated user behavior.



X Xie, J Mao, Y Liu, M de Rijke, Y Shao, Z Ye, M Zhang, S Ma. Grid-based evaluation metrics for web image search. WWW 2019.

aggregating metrics

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

Rocío Cañamares, Pablo Castells, and Alistair Moffat, Offline evaluation options for recommender systems, Information Retrieval Journal, 2020

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

Mark Claypool et al. **Implicit Interest Indicators.** IUI 2001.
Diane Kelley et al. **Implicit feedback for inferring user preference: a bibliography.** SIGIR Fourm 2003.

Explicit Feedback

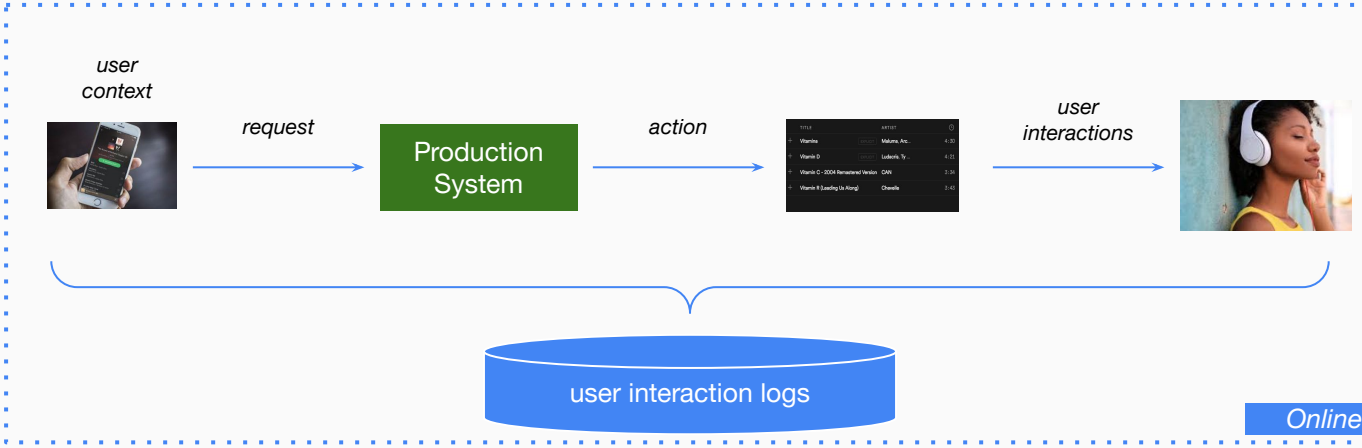
- Ask users to learn about their experience with the system (surveys, ratings, etc.) or setup annotation tasks for humans to provide labels given data points.
- Limitations of explicit feedback
  - intrusive
  - expensive
  - might not reflect ***true*** user preference
  - might be infeasible for certain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

Implicit Feedback

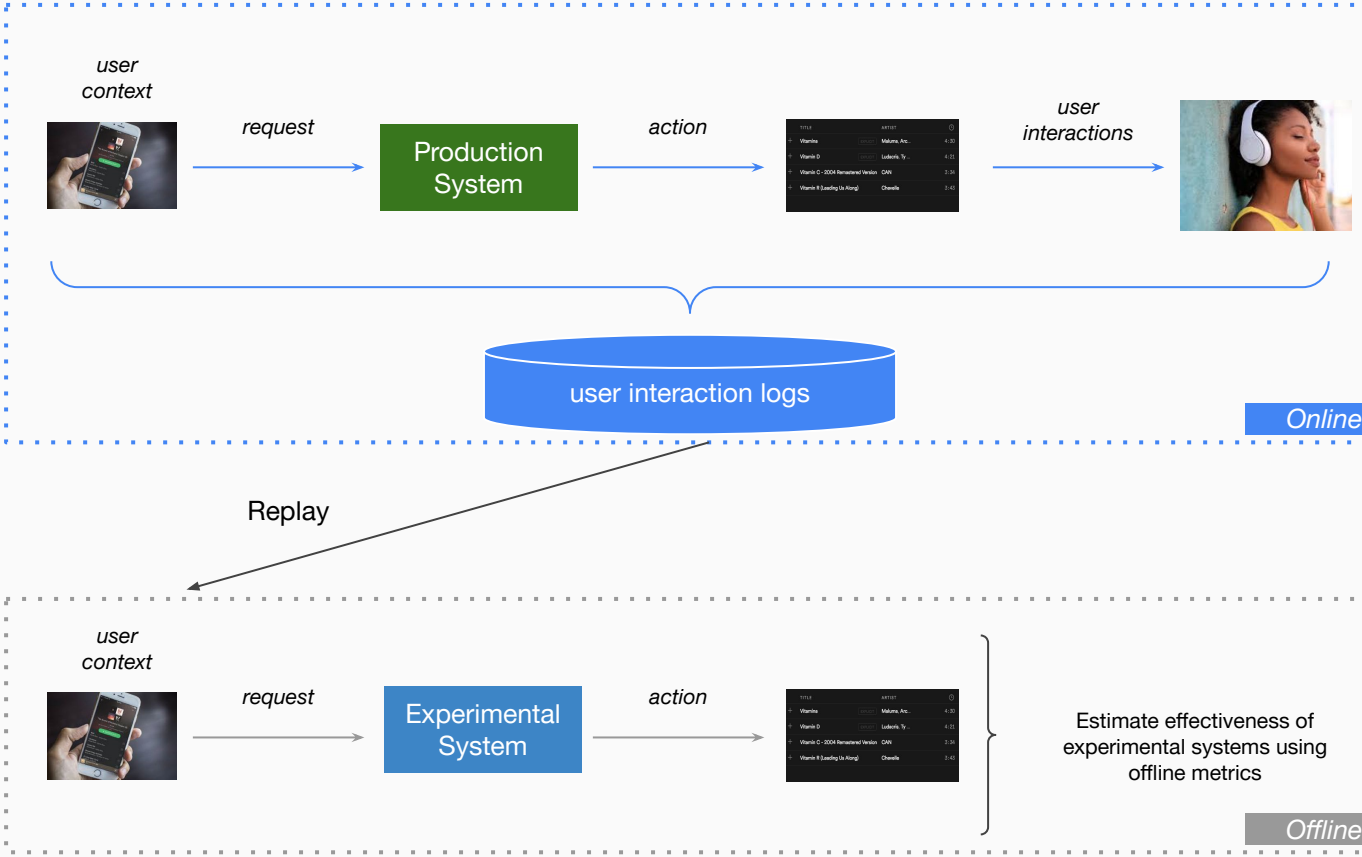
- Infer user satisfaction from behavioral systems as they interact with the system.
- Advantages over explicit
  - collected in a natural setting
  - easy & inexpensive to collect
  - provides a direct measure of user preferences
- Limitation
  - can be noisy
  - biased

Mark Claypool et al. **Implicit Interest Indicators.** IUI 2001.
Diane Kelley et al. **Implicit feedback for inferring user preference: a bibliography.** SIGIR Fourm 2003.

ML Evaluation Workflow



ML Evaluation Workflow

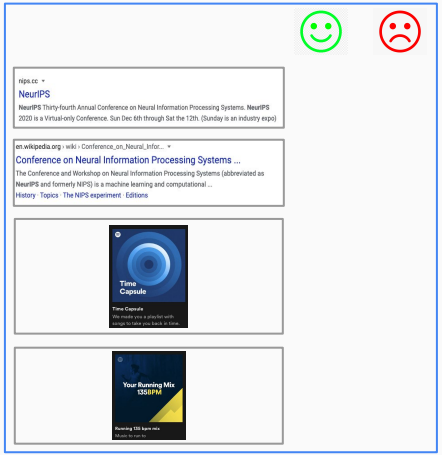


Categorizing Implicit Signals

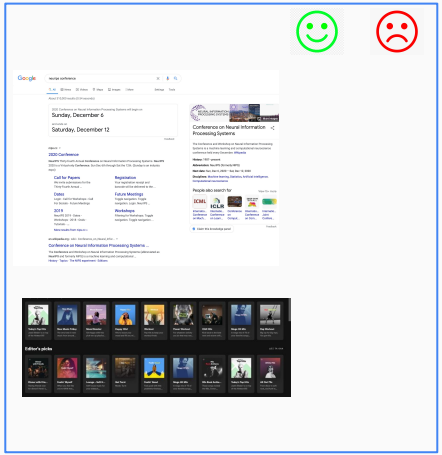
short-term

long-term

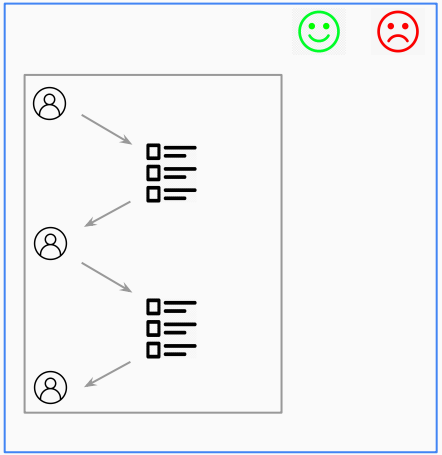
Item Level Feedback



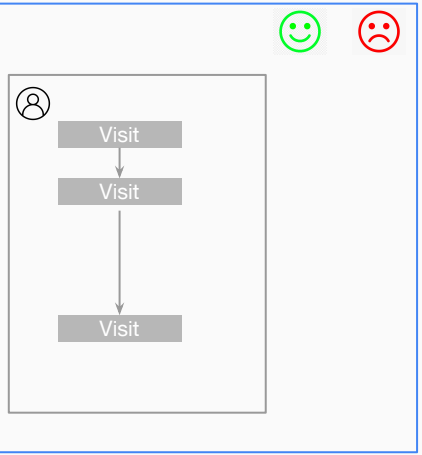
Page Level Feedback



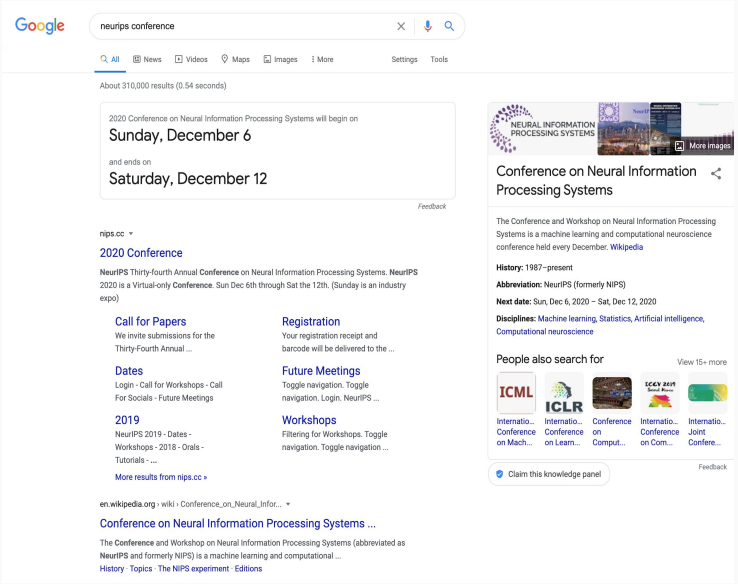
Session Level Feedback



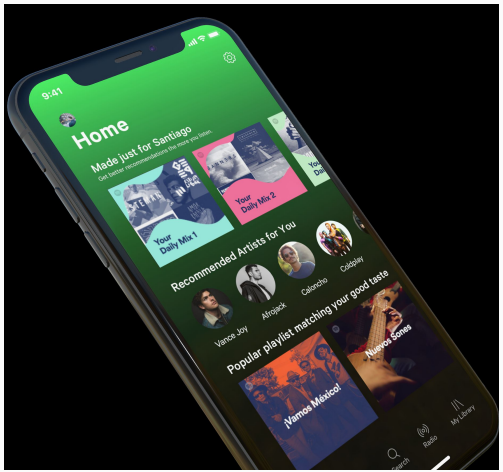
Intra-Session Feedback



Illustrative Examples



Web Search

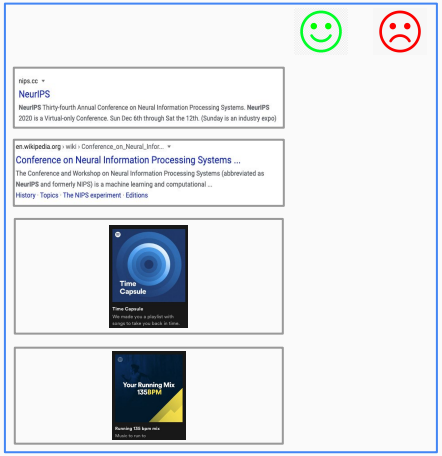


Music Recommendation

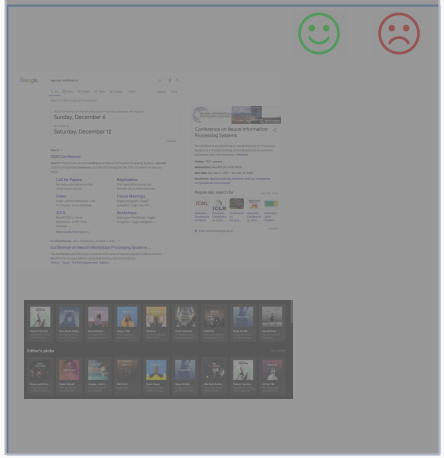
Item Level Feedback

short-term *long-term*

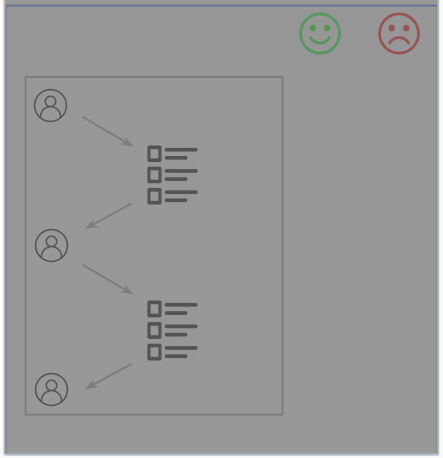
Item Level Feedback



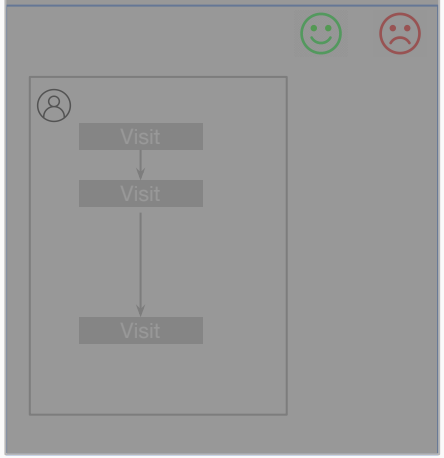
Page Level Feedback



Session Level Feedback

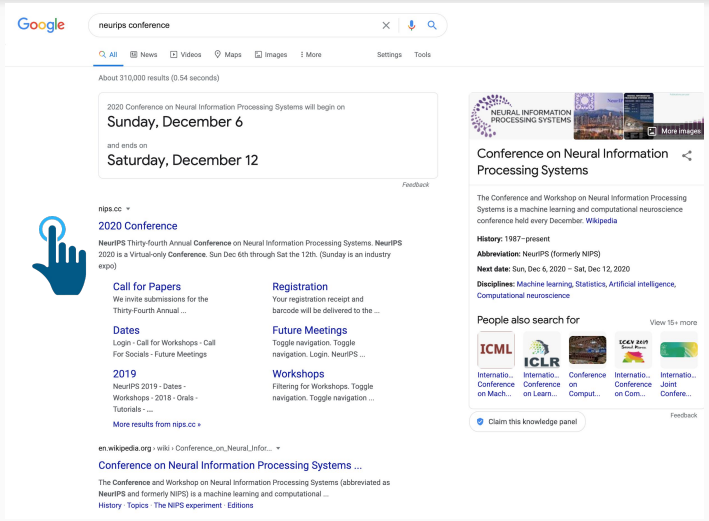


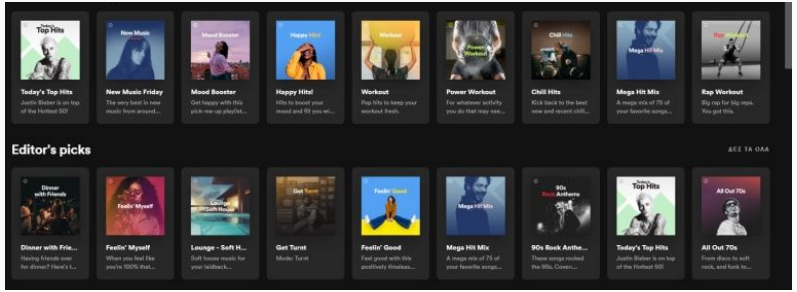
Intra-Session Feedback



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.

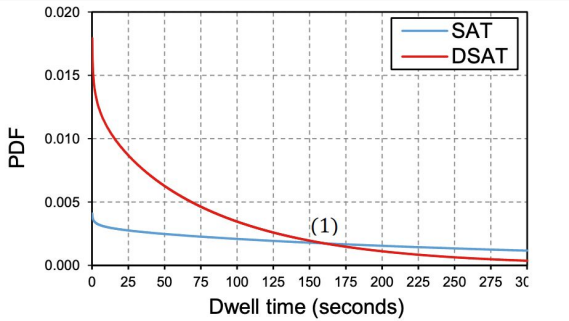




Thorsten Joachims. **Optimizing search engines using clickthrough data.** KDD 2007.

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.

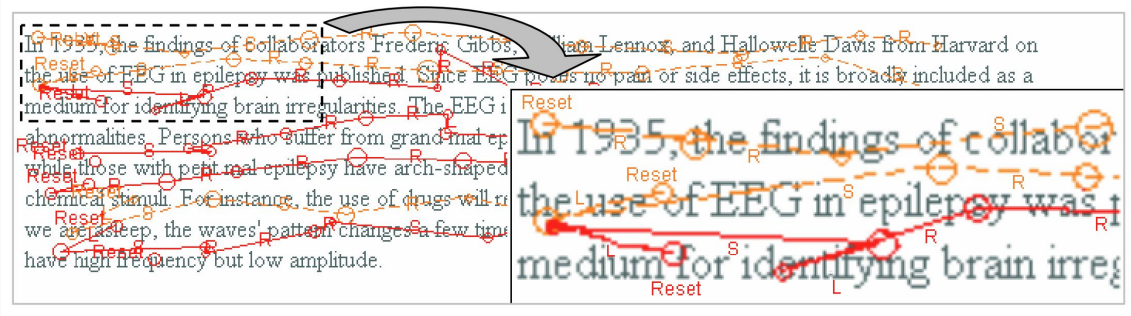


- Dwelling patterns are influenced by attributes of an item. For instance, sophisticated content (high reading level) requires more time.

Youngho Kim et al. **Modeling Dwell Time to Predict Click-level Satisfaction.** WSDM 2014.

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.



Georg Buscher et al. **Eye movements as implicit relevance feedback.** CHI 2008.

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

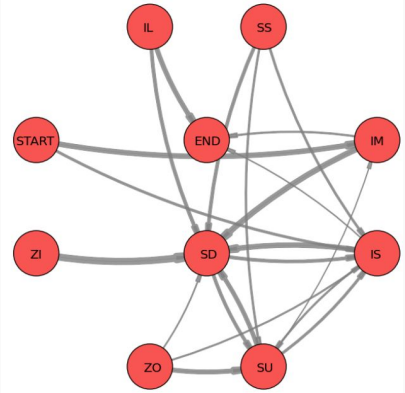
Click positions

Cursor movement positions

Jeff Huang, Ryen White and Susan Dumais. **No clicks, no problem: using cursor movements to understand and improve search.** CHI 2011.

Behavioral Signals: Touch Gestures on Mobile

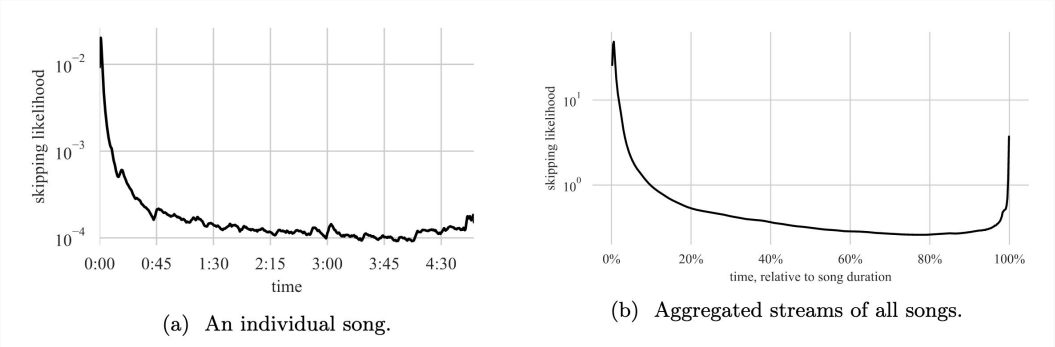
- Behavior modeling on mobile need to account for fine-grained user interactions such as
  - zooming in (ZI) and out (ZO)
  - swiping down (SD), up (SU), and horizontal (SS)
  - inactive short (IS), medium (IM), and long (IL)
- Markov state transitions can be used to predict likelihood of success for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



Qi Guo et al. **Mining touch interaction data on mobile devices to predict web search result relevance.** SIGIR 2013.

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.



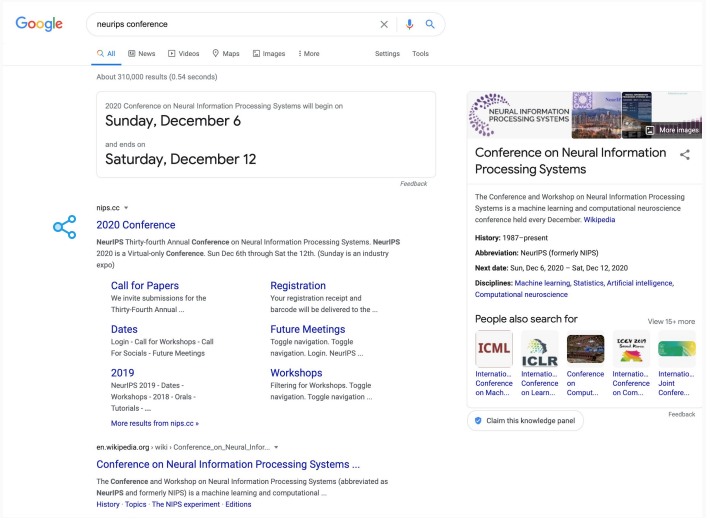
(a) An individual song.

(b) Aggregated streams of all songs.

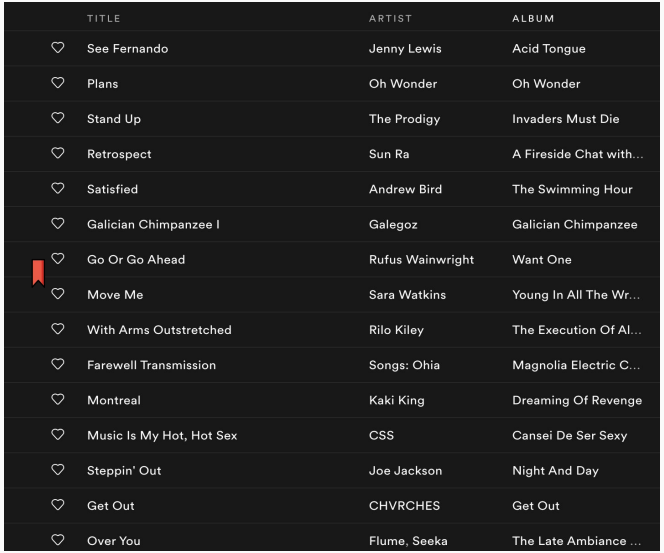
Nicola Montecchio et al. **The Skipping Behavior of Users of Music Streaming Services and its Relation to Musical Structure.** PLOS ONE 2020.

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.

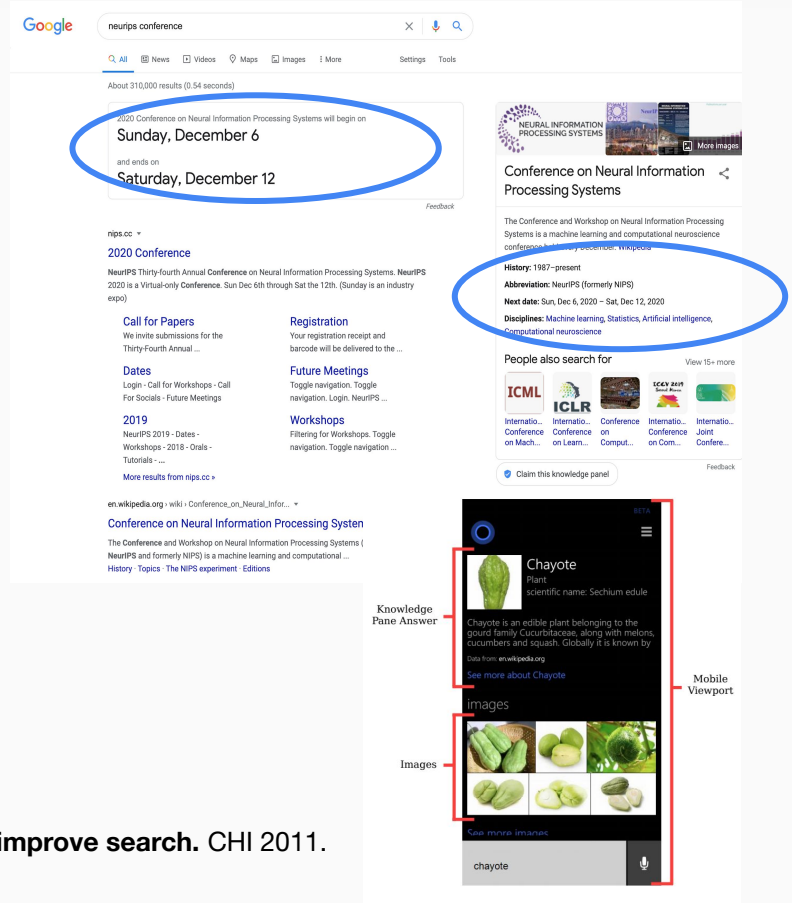


Jean Garcia-Gathright et al. **Understanding and Evaluating User Satisfaction with Music Discovery.** SIGIR 2018.

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

User Intent / Goals

Context

Heterogeneity of Items

User

Summary: Item Level Feedback

Behavioral Signals

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

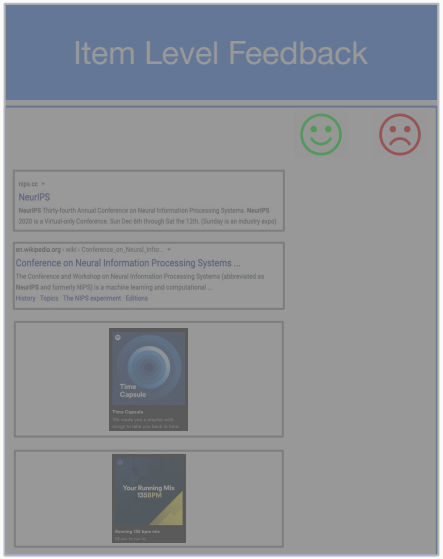
Offline metric validation

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

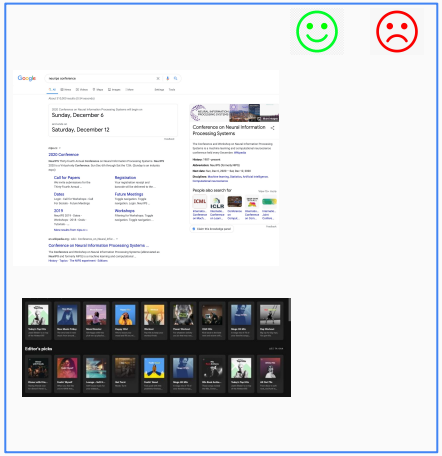
Page Level Feedback

short-term *long-term*

Item Level Feedback

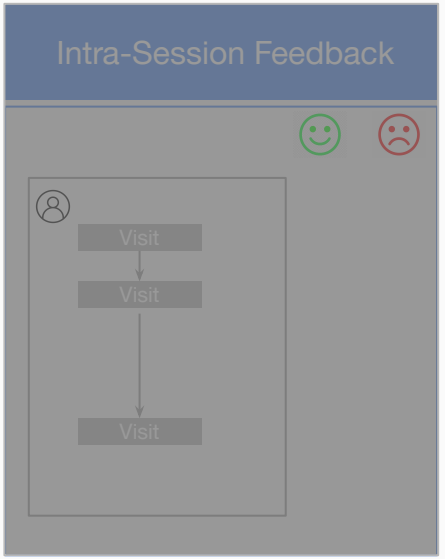


Page Level Feedback

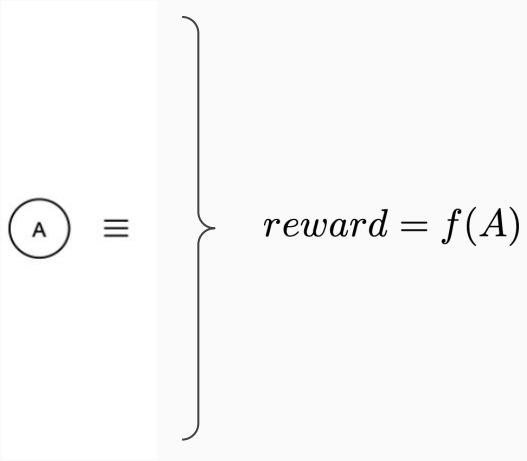


Session Level Feedback

Intra-Session Feedback



Whole Page Feedback

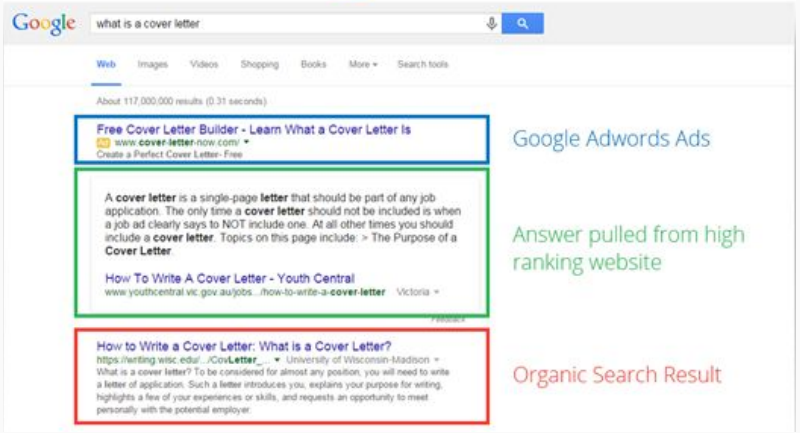


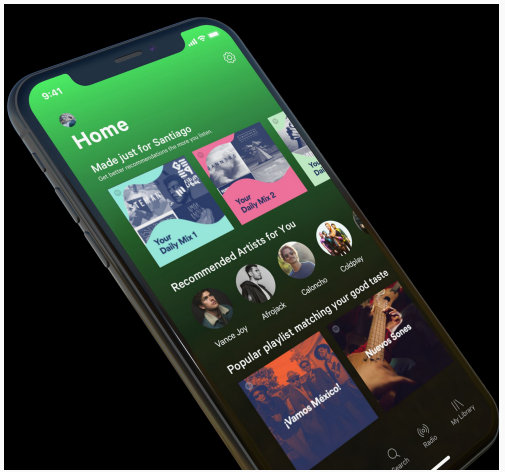
- In many application, a set of items are presented to the user for a given request.
- Whole-page feedback defines how well items presented on a page and its attributes satisfies the user request.

  Action → entire page
  Labels → success for the entire page

Peter Bailey et al. **Evaluating Whole-Page Relevance.** SIGIR 2010.

Whole Page Feedback: Examples





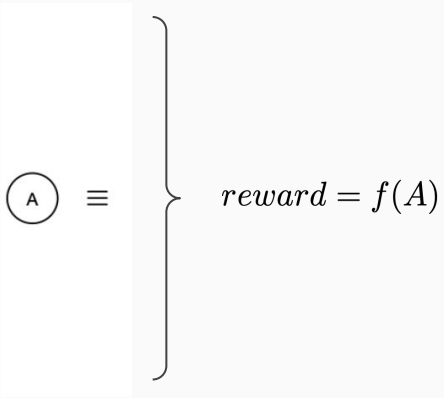
User Models for Whole Page Evaluation

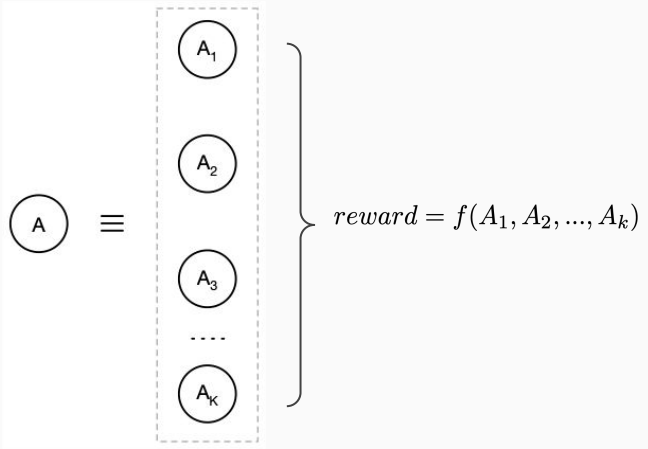
- While assigning labels/reward for the result page provides a holistic view of success, in several scenarios they are computed from rewards observed at the item-level.

  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.

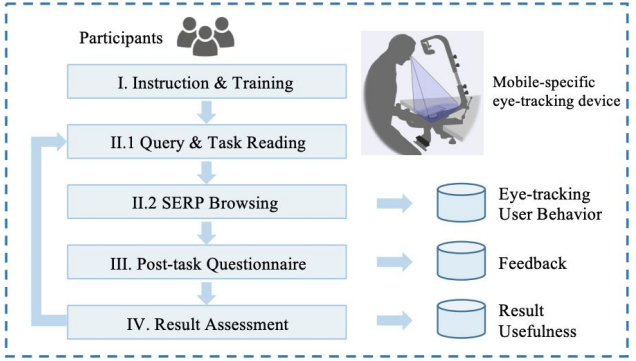
- This calls for the development of user behavioral models to accurately estimate a page-level reward



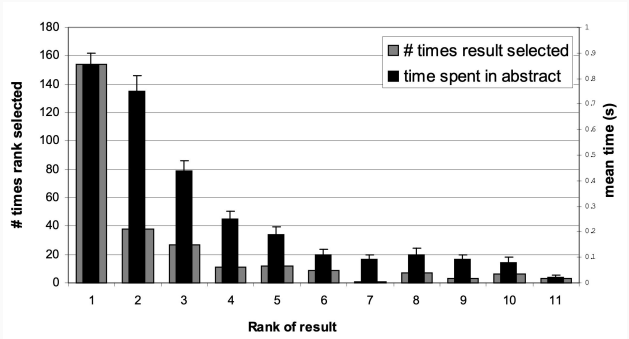


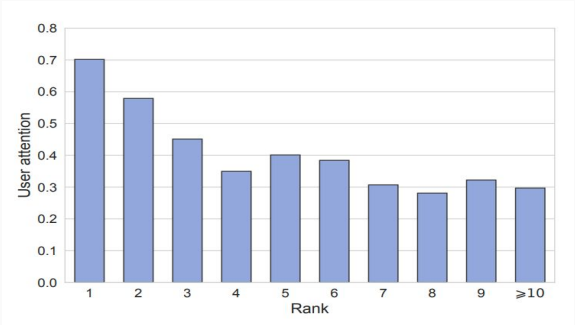
User Browsing Models: Position-Based Model

> 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***

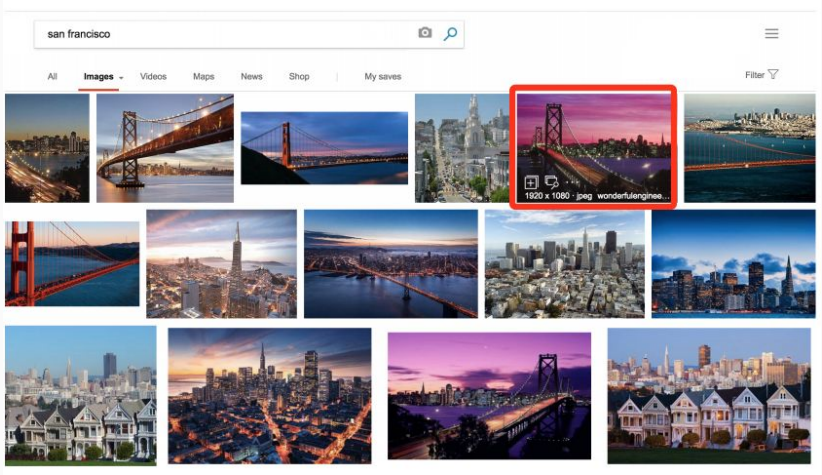




Laura A. Granka et al. **Eye-tracking analysis of user behavior in WWW search.** SIGIR 2004.
Yukun Zheng et al. **Investigating Examination Behavior in Mobile Search.** WSDM 2020.

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.

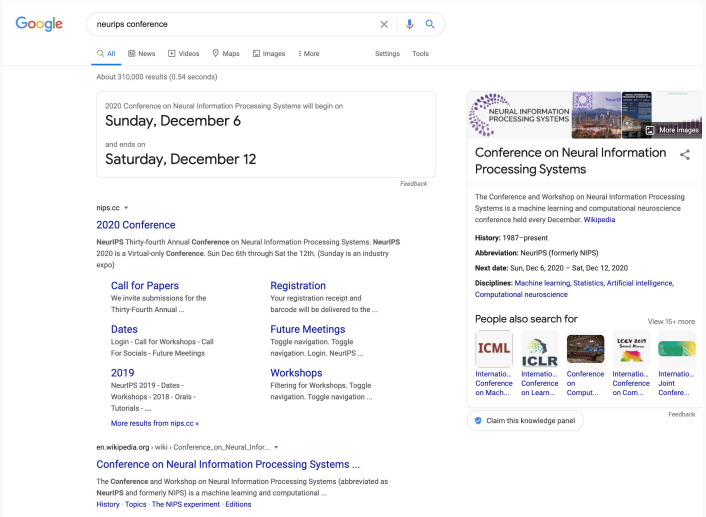


middle positions may attract more attention depending on the domain in grid-based layouts.

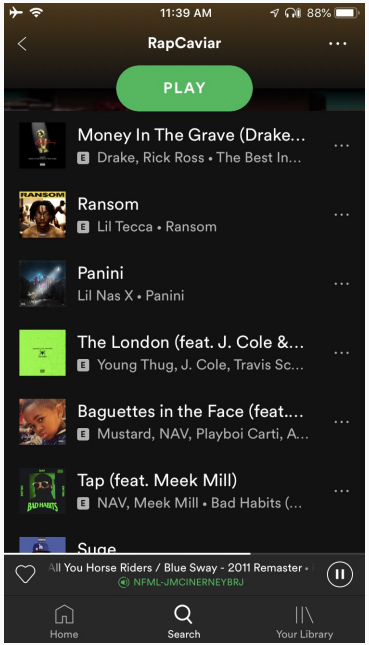
Xiaohui Xie et al. **Grid-based Evaluation Metrics for Web Image Search.** CIKM 2019.

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).

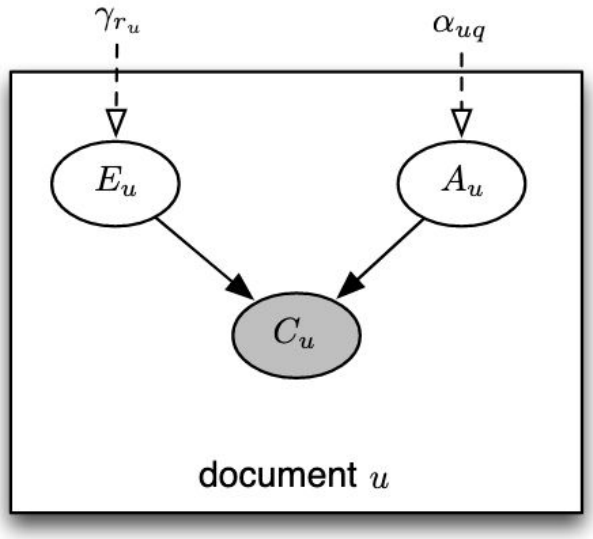


Olivier Chapelle et al. **Expected reciprocal rank for graded relevance.** CIKM 2009.

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***.

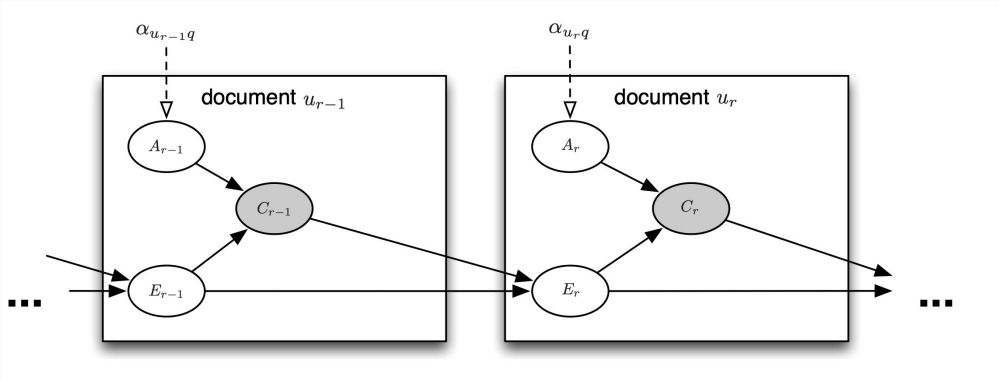
$$P(E_{r_u} = 1) = \gamma_{r_u}$$
$$P(A_u = 1) = \alpha_{uq}$$
$$P(C_u = 1) = P(E_{r_u} = 1) \cdot P(A_u = 1)$$



Aleksandr Chuklin, Ilya Markov, Maarten de Rijke. **Click Models for Web Search.** CIKM 2009.

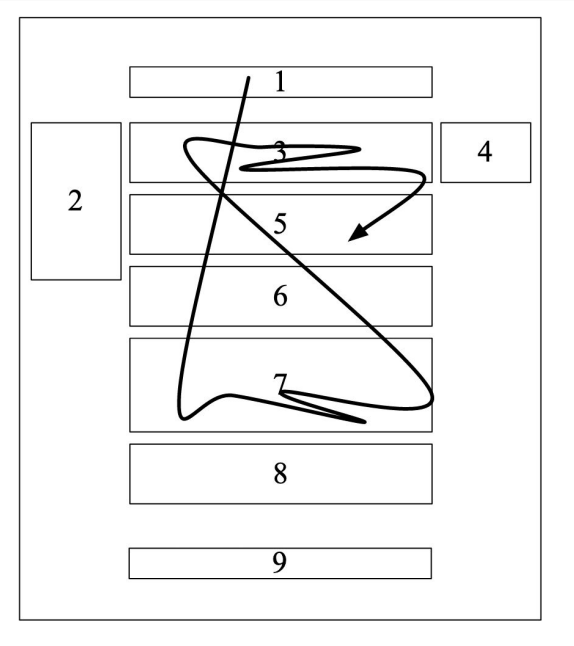
Click Models: Cascade Model

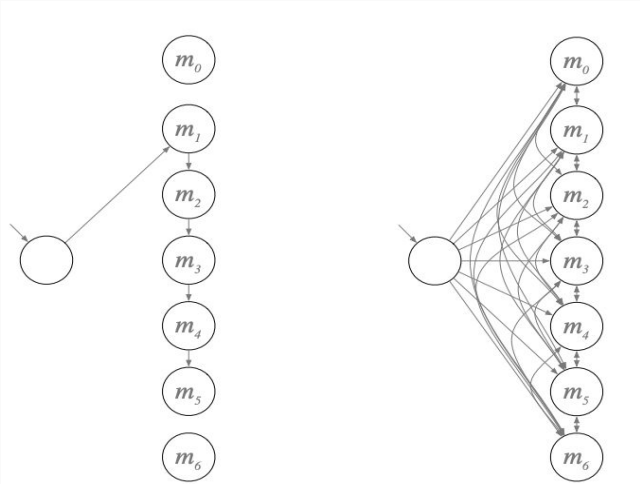
- Cascade dependency between clicks and previously examined items



Aleksandr Chuklin, Ilya Markov, Maarten de Rijke. **Click Models for Web Search.** CIKM 2009.

Browsing Models: Non-Sequential Examination Model





Fernando Diaz et al. **Robust models of mouse movement on dynamic web search results pages.** CIKM 2013.

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.

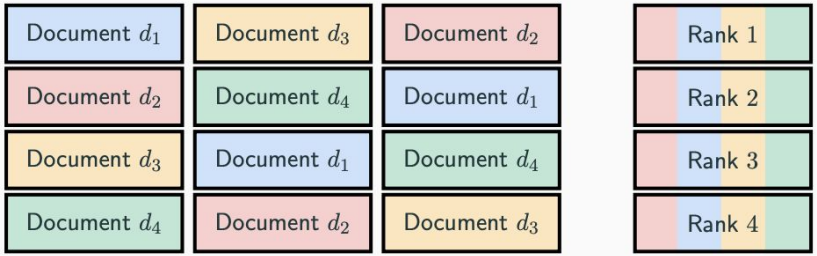


Figure Source: *Unbiased Learning to Rank: Counterfactual and Online Approaches*

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.

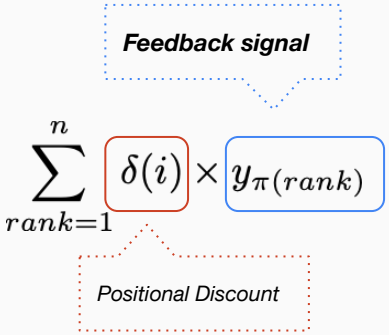
Counterfactual Evaluation from Log Data

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

Counterfactual Evaluation from Log Data

System
$\pi_{production}$

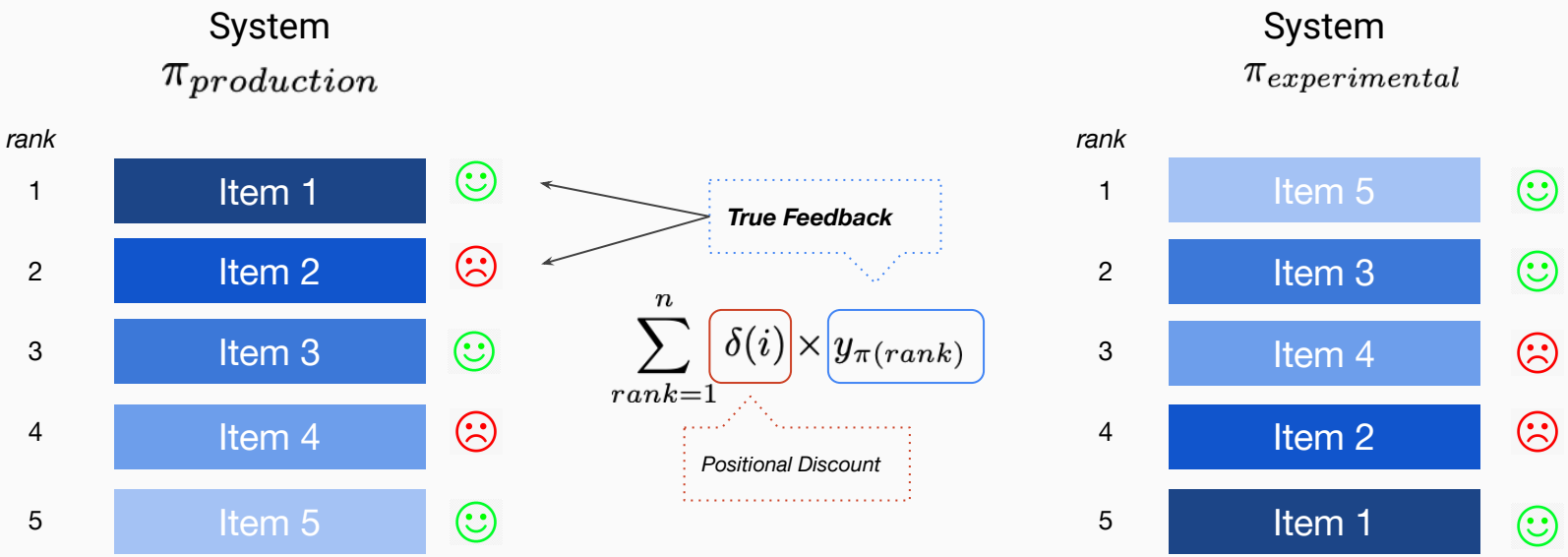
rank	
1	Item 1
2	Item 2
3	Item 3
4	Item 4
5	Item 5



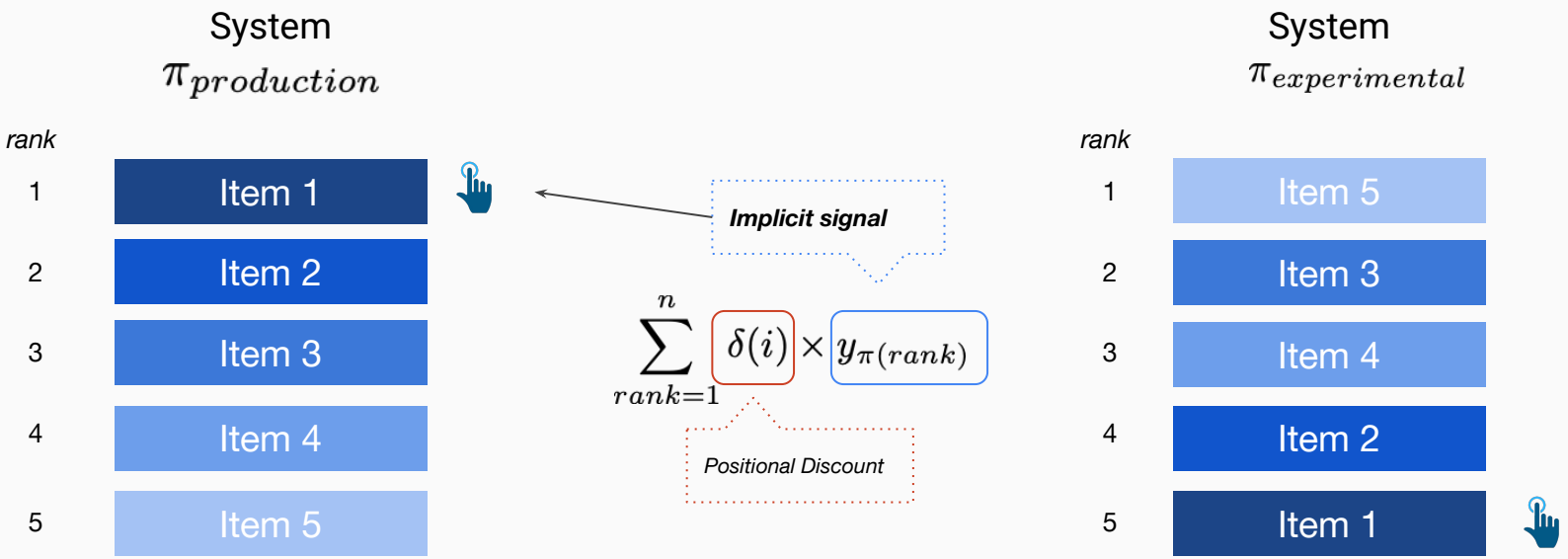
System
$\pi_{experimental}$

rank	
1	Item 5
2	Item 4
3	Item 3
4	Item 2
5	Item 1

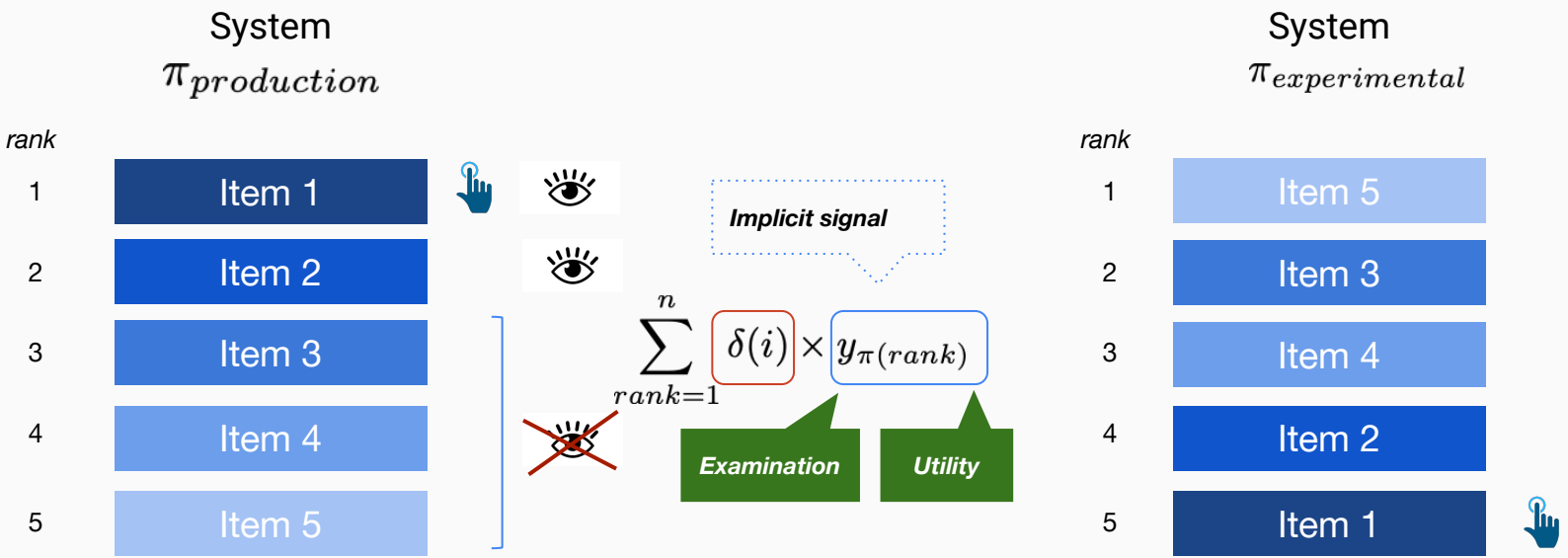
Counterfactual Evaluation from Log Data



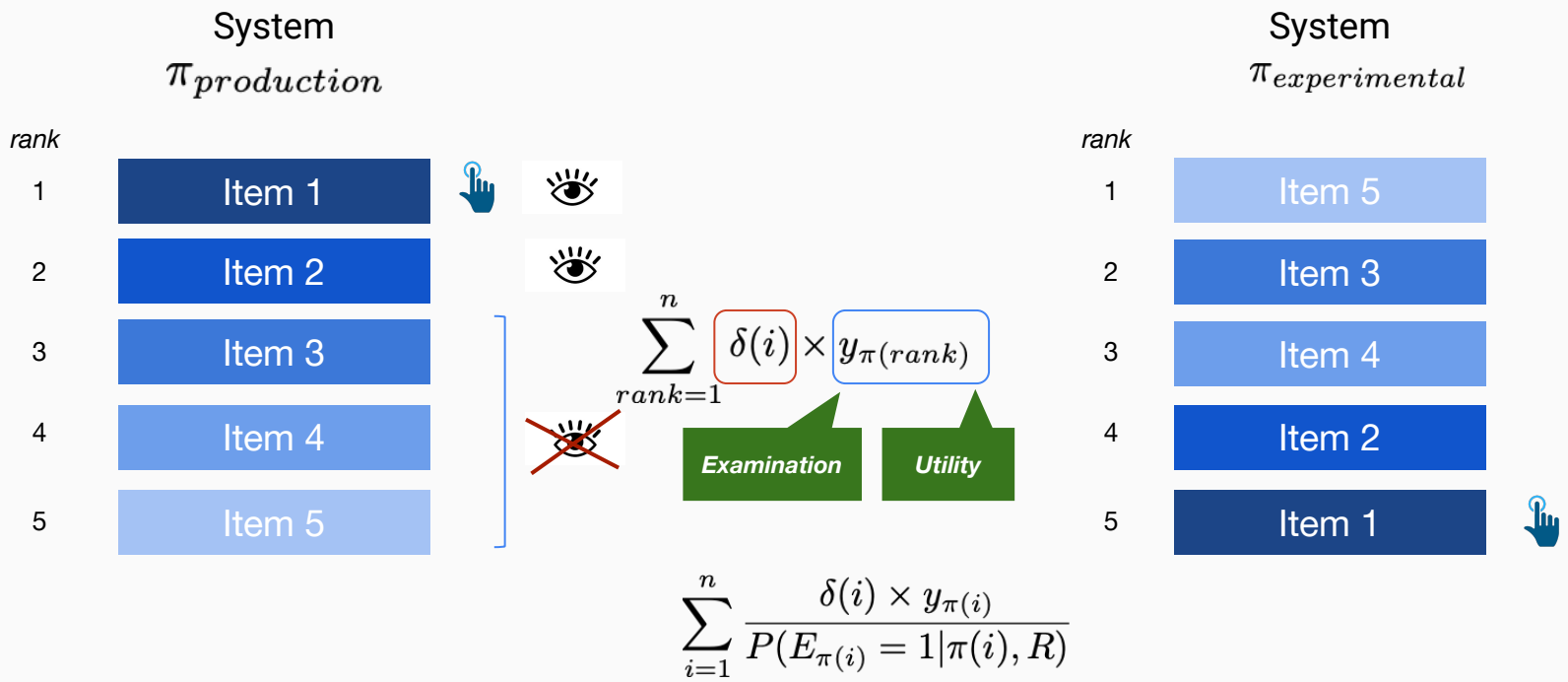
Counterfactual Evaluation from Log Data



Counterfactual Evaluation from Log Data

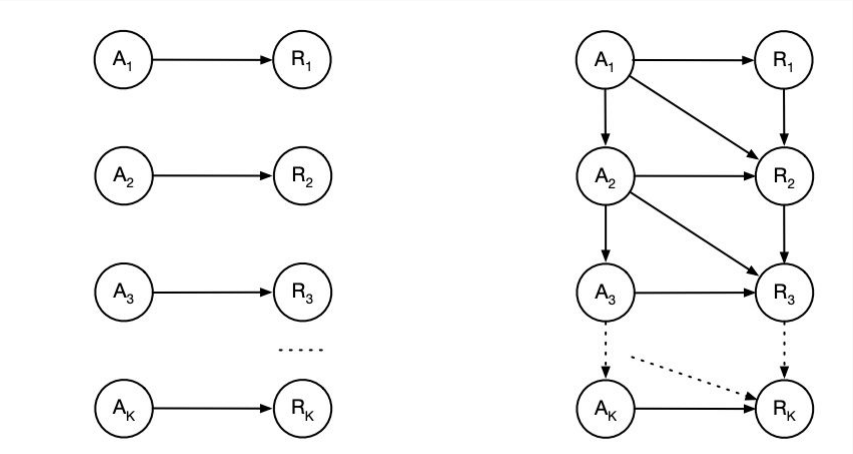


Counterfactual Evaluation from Log Data



Xuanhui Wang et al. **Learning to Rank with Selection Bias in Personal Search.** SIGIR 2016.
Thorsten Joachims et al. **Unbiased Learning-to-Rank with Biased Feedback** WSDM 2017.

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.

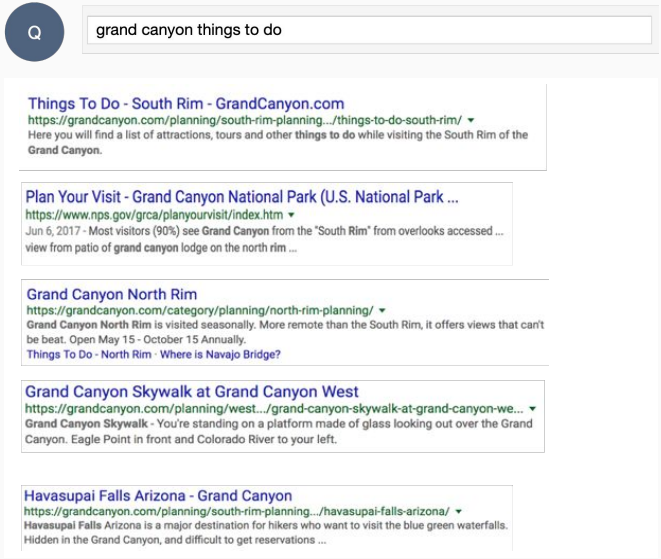
Sequential Reward Dependence Assumption

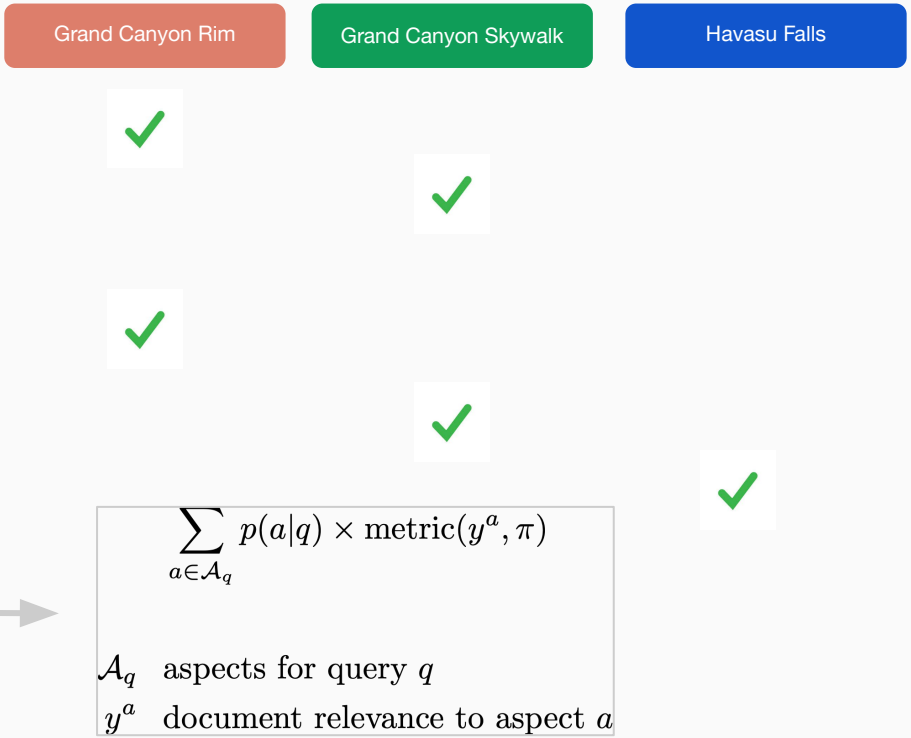
James McInerney et al. **Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions** KDD 2020.

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





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

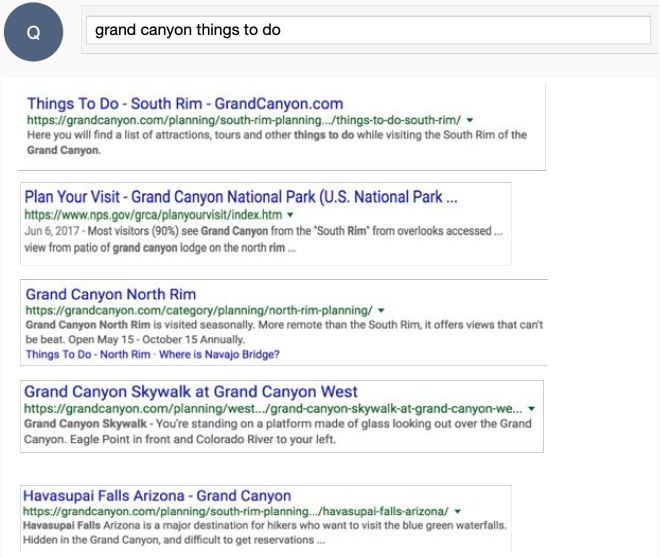
$$\sum_{a \in \mathcal{A}_q} p(a|q) \times \text{metric}(y^a, \pi)$$

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

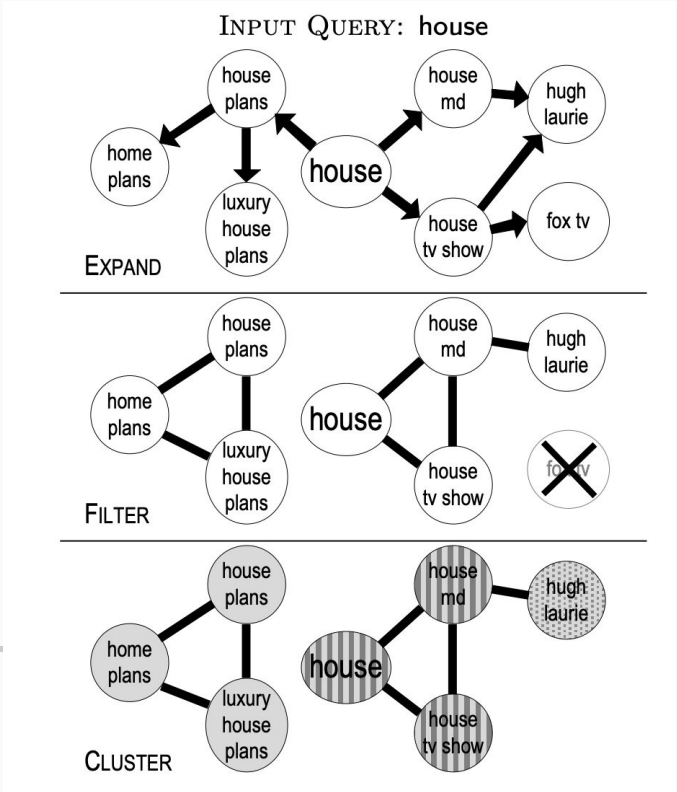
y^a document relevance to aspect a

Charles Clarke et al. **Novelty and Diversity in Information Retrieval Evaluation** SIGIR 2008.
Filip Radlinski et al. **Inferring Query Intent from Reformulations and Clicks.** WWW 2010.

Novelty & Diversity in Rankings



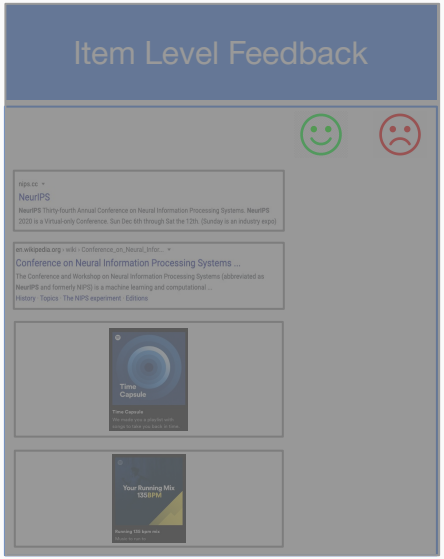
A variation of rank-based metric that rewards novelty and penalizes diversity

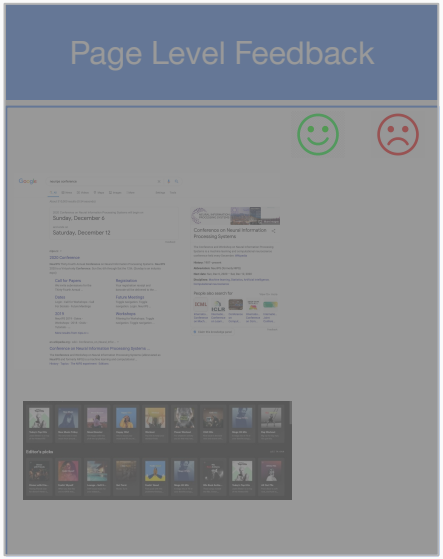


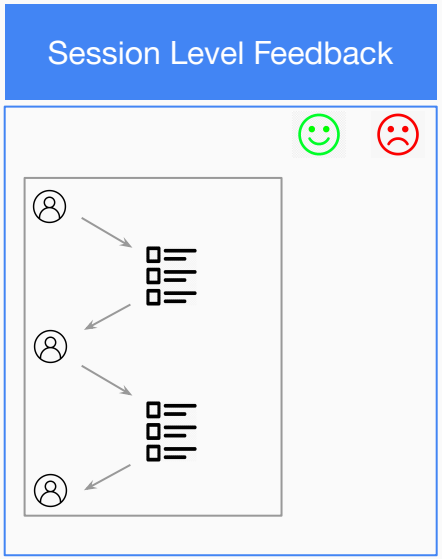
Charles Clarke et al. **Novelty and Diversity in Information Retrieval Evaluation** SIGIR 2008.
Filip Radlinski et al. **Inferring Query Intent from Reformulations and Clicks.** WWW 2010.

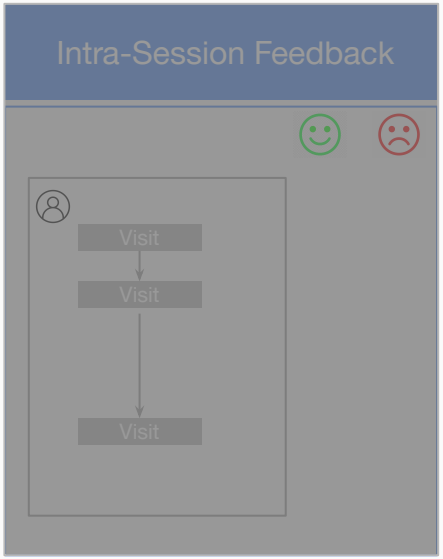
Session Level Feedback

short-term *long-term*

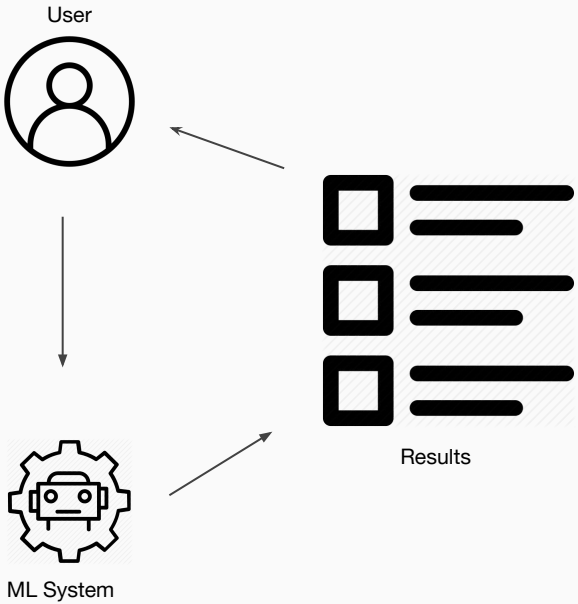




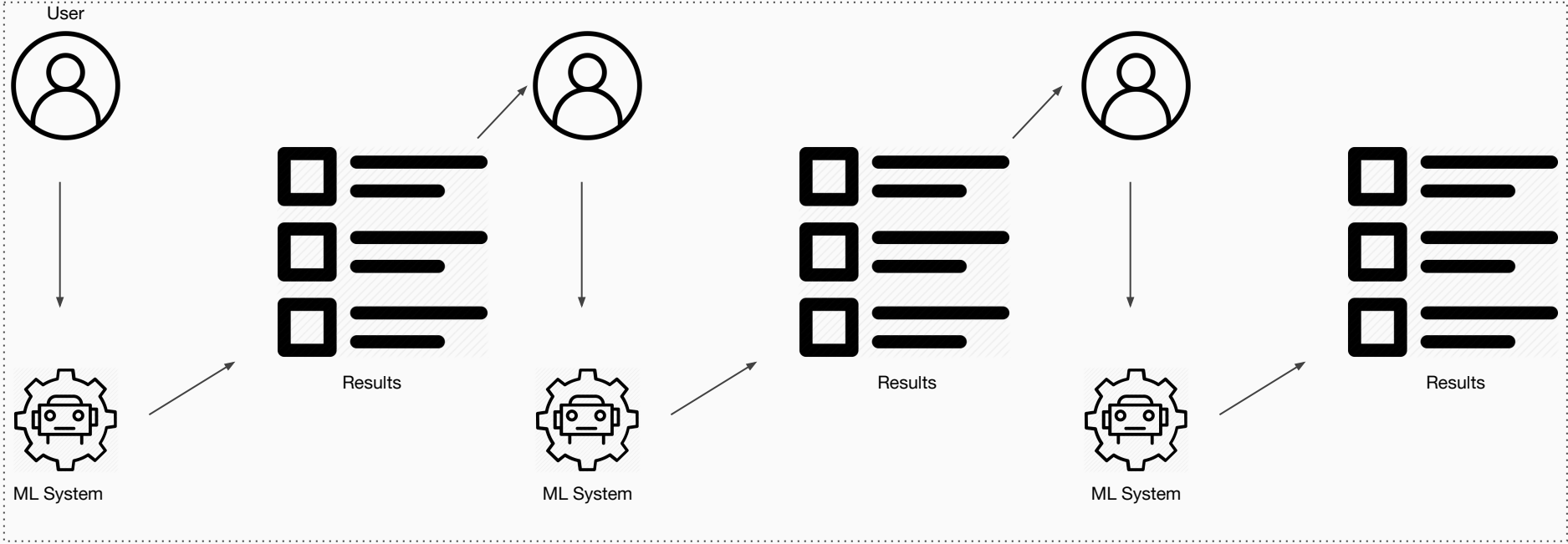




Session Level Evaluation: Introduction

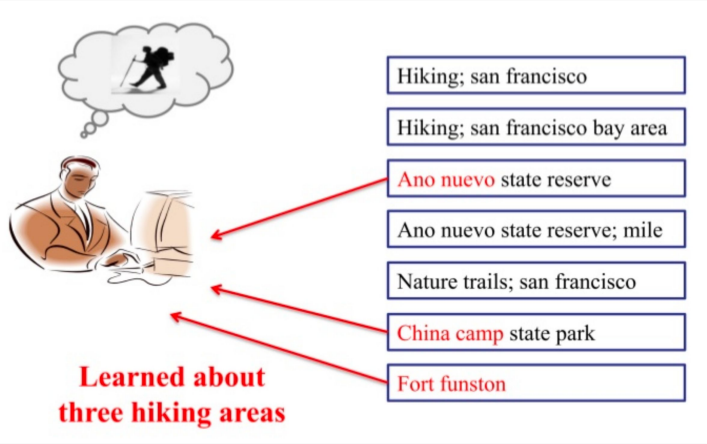


Session Level Evaluation: Introduction



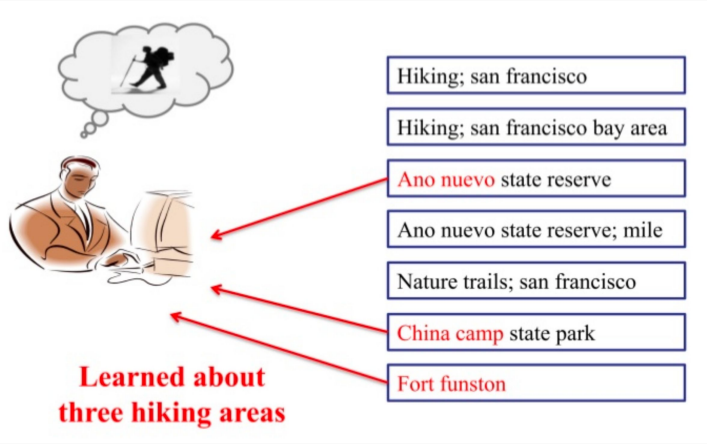
How do we evaluate the user experience over a session?

Session Level Evaluation: Examples



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

Session Level Evaluation: Examples

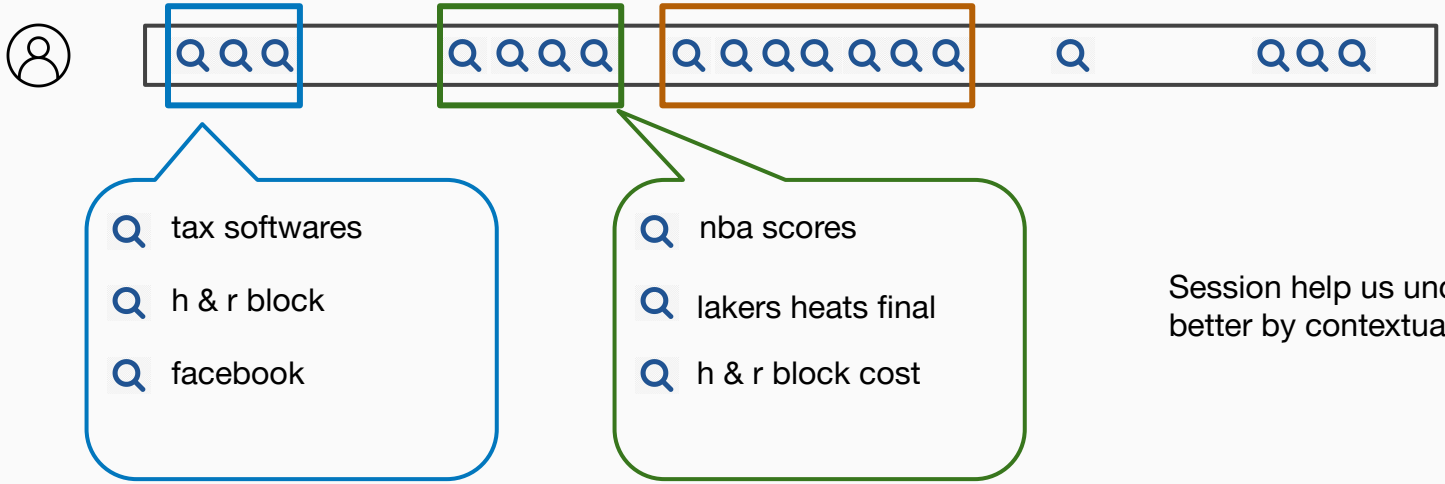


Rishabh Mehrotra et al. **Understanding & Inferring User Tasks and Need.** WWW 2018.
Shuo Zhang and Krisztian Balog. **Evaluating Conversational Recommender Systems via User Simulation.** KDD 2020.
Malte Ludewig and Dietmar Jannach. **Evaluation of session-based recommendation algorithms.** User Modeling and User Adapted Interaction 2018.

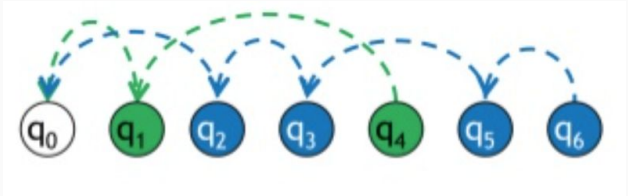
Identifying Sessions from Logs

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.

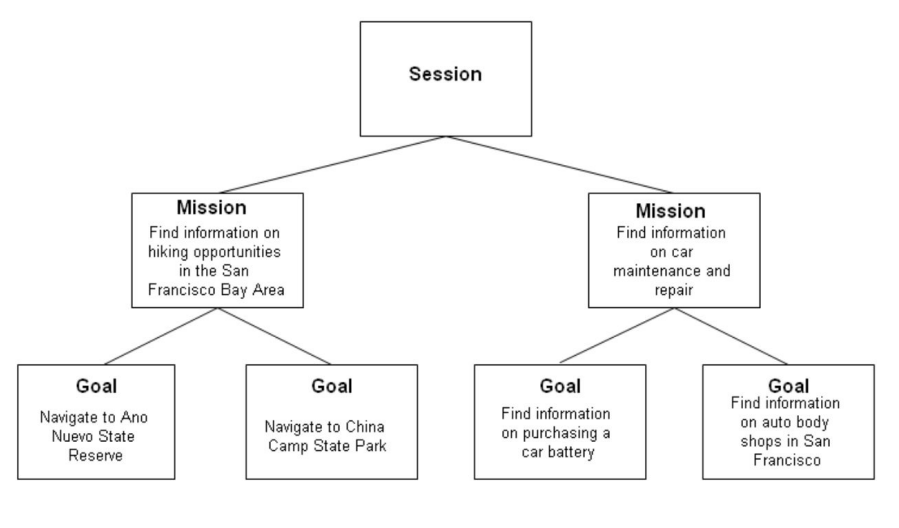


Sessions → User Goals

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



Rosie Jones & Kristina Klinkner. **Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs.** CIKM 2008.
Julia Kiseleva et al. **Understanding User Satisfaction with Intelligent Assistants.** CHIIR 2016.

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 ↓nd 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 ↓nd 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

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

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

Henry Field et al. **Predicting Searcher Frustration.** SIGIR 2010.

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.sooperarticles.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...

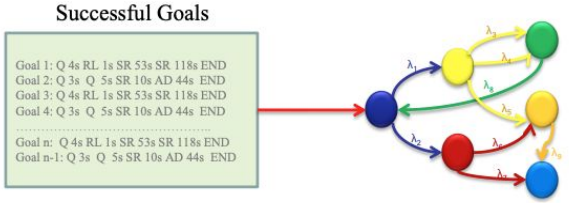
END OF SESSION

Ahmed Hassan et al. **Struggling or exploring?: disambiguating long search sessions.** WSDM 2014.

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



Time	Query	# Clicks	Avg. Dwell Time
t_1	gauage 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

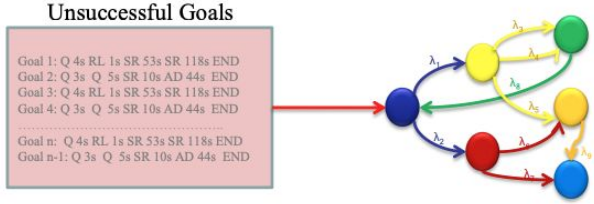


Figure Source: Understanding & Inferring User Tasks and Needs

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

Ahmed Hassan et al. **Beyond DCG: User Behavior as a Predictor of a Successful Search.** WSDM 2014.

Metric Validation: Session Level Annotations

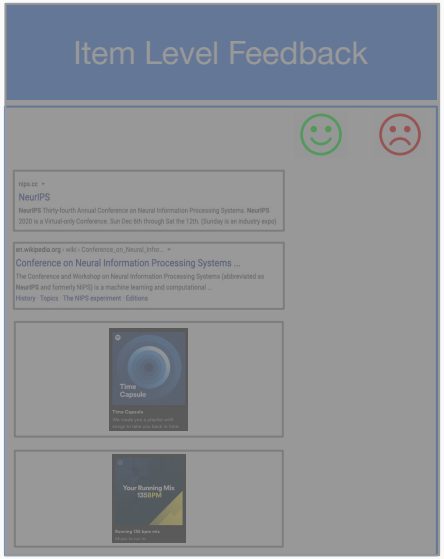
Table 3 A list of user actions that are included in utility. The row rank reflects a descending order in the absolute value of the weight associated with the events.

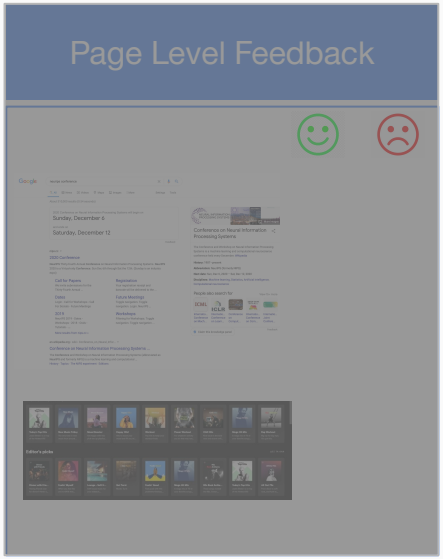
Events	Type
A click to an external page that is the last interaction in the user's session	strongly positive
A query issued by the user that is followed by a query that reformulates it	strongly negative
A click to an external page with a long dwelltime	positive
A click to an external page with a short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

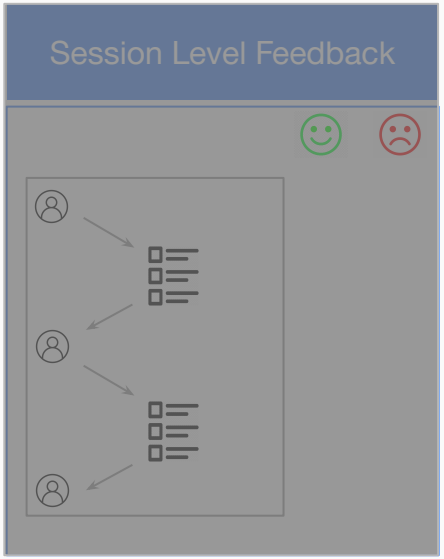
Ahmed Hassan et al. **Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments.** WSDM 2014.

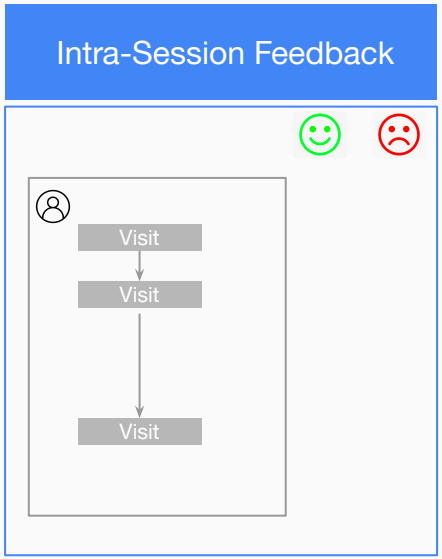
Inter-Session & Long-Term Feedback

short-term *long-term*









Inter-Session & Long-Term Feedback

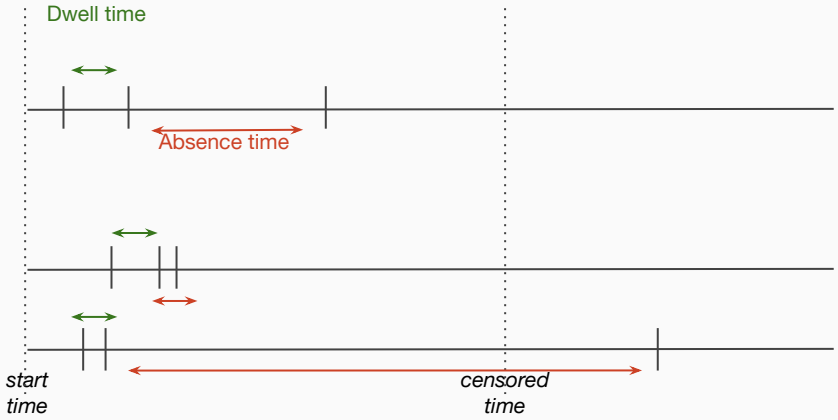
- Short-term metrics such as click can be misleading as they ignore user learning & novelty effects over time.

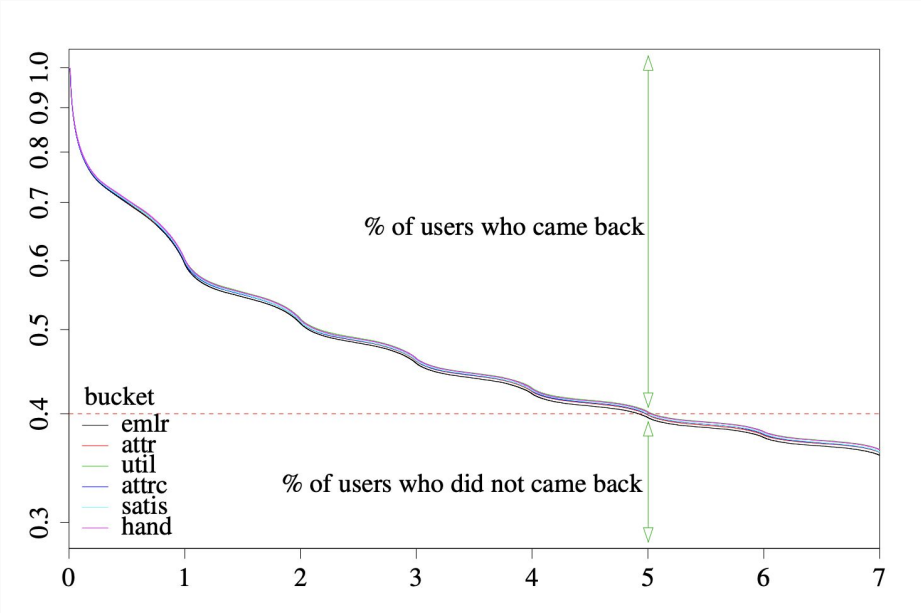
- Short-term signals ignore user's long-term goals/needs and a myopic view of success can result in "echo-chamber" effects or "filter bubbles". Kohavi et al. showed that optimizing for short-term improvements may be detrimental in the long-term.

- 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.

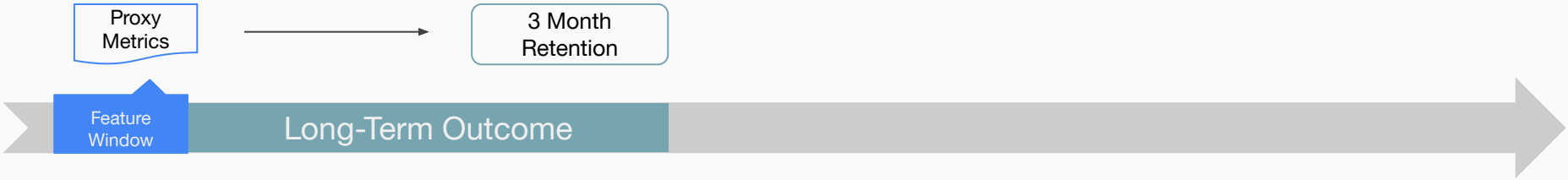




Georges Dupret and Mounia Lalmas. **Absence time and user engagement: evaluating ranking functions.** *WSDM 2013.*

Inter-Session Feedback: Surrogates & Proxies

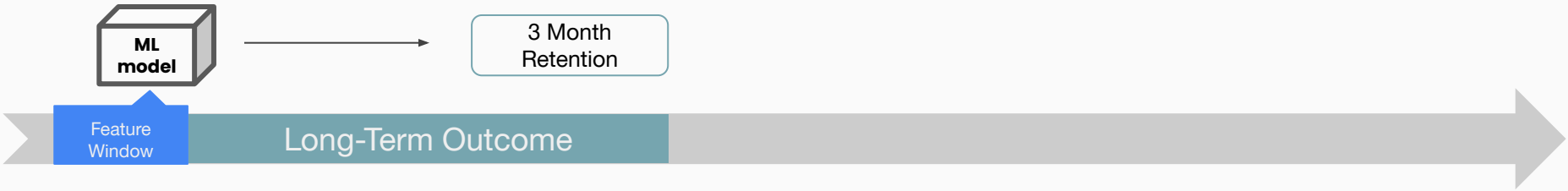
- Satisfied users coming back to the service and re-engaging with the ML system is a reasonable proxy for long term success.
- One or more short-term metrics such as Clickthrough Rate, Session success can be used as proxies to estimate long-term outcomes (e.g. user retention).



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 Mehotra 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)

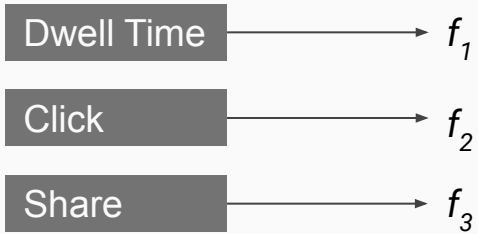
Somit Gupta et al. **Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments**. KDD 2019.

Multiple Metrics Decision

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

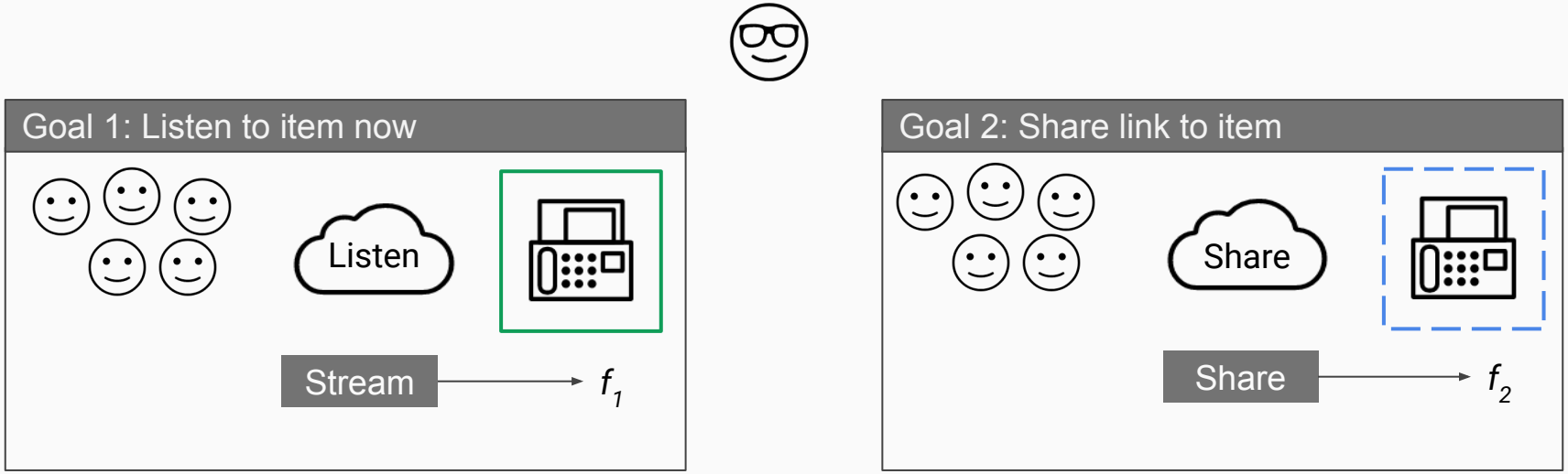



The system maker may have multiple metrics that reflect different aspects of how much the system is helping the user. For example,



Multiple Metrics - Example

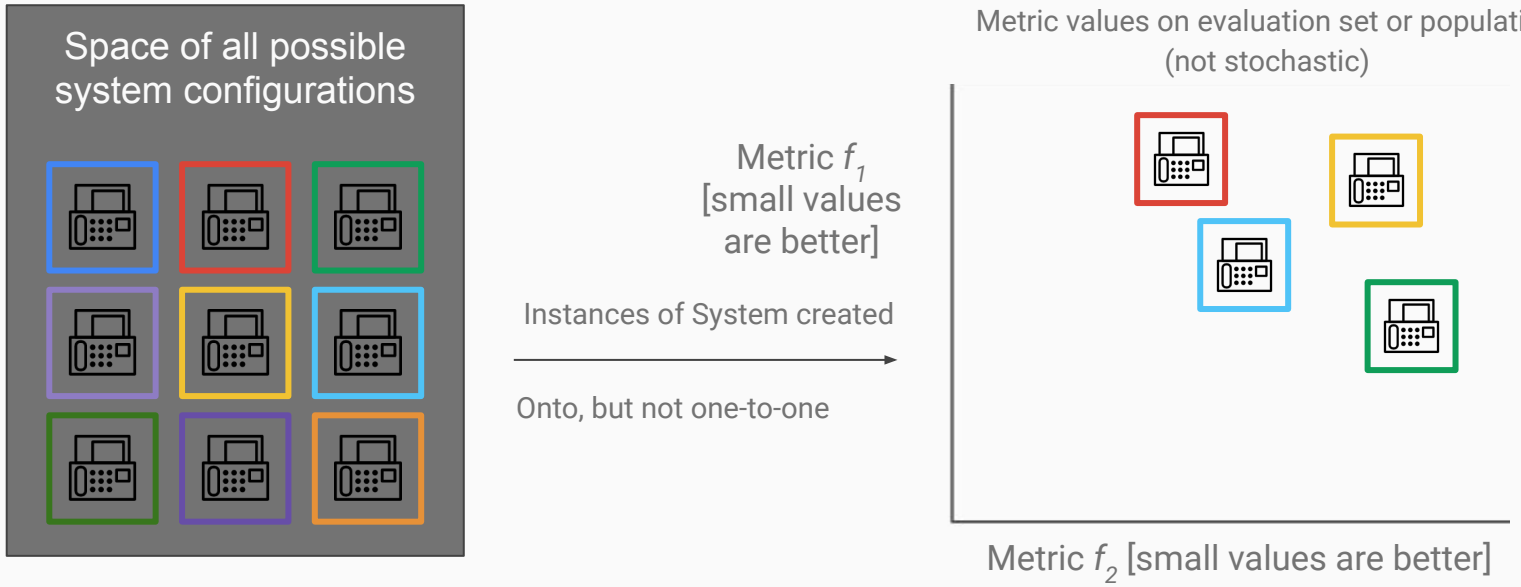
Users of an audio content search feature can have multiple goals, measurable by different metrics (Hosey et al, 2019)



Christine Hosey et al. **Just Give Me What I Want: How People Use and Evaluate Music Search**. CHI 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

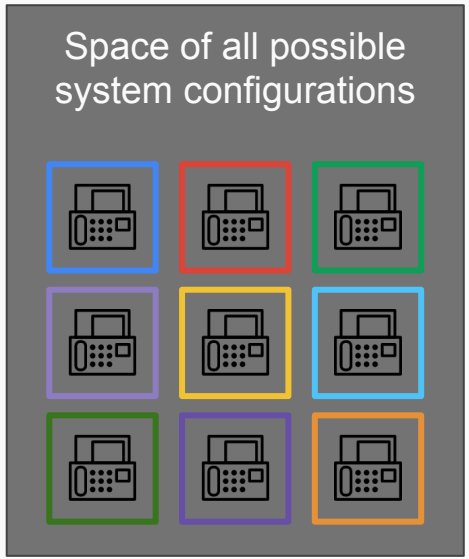


Multiple Optimization Problem (MOP)

The system maker wants to choose the best system ***x*** such that

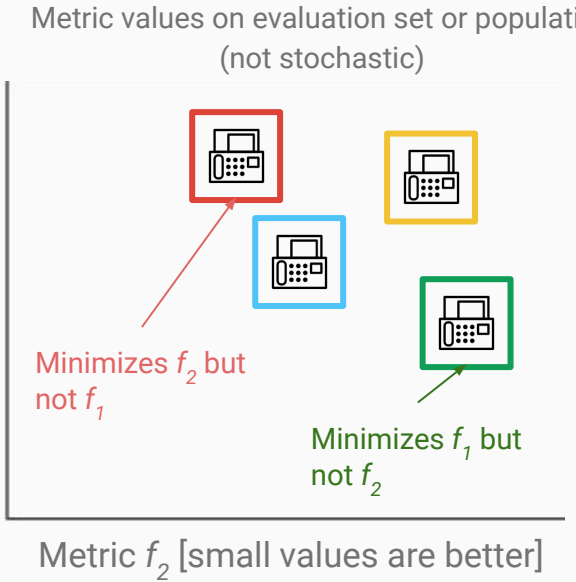
$$\text{minimize } f_1(\mathbf{x}), \ldots, \text{ minimize } f_m(\mathbf{x}), \mathbf{x} \in \mathcal{X}$$

$\mathcal{X}$



Instances of System created

Onto, but not one-to-one



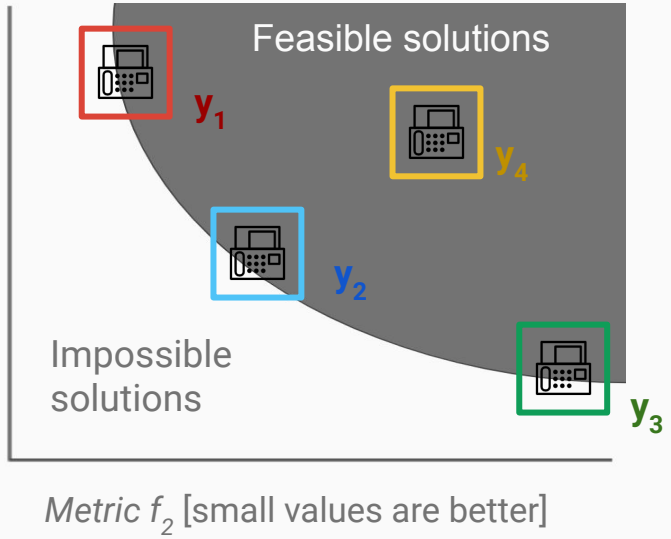
Pareto Dominance

A solution $y_1=[f_1(x), f_2(x), \ldots, f_M(x)]$ *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*.

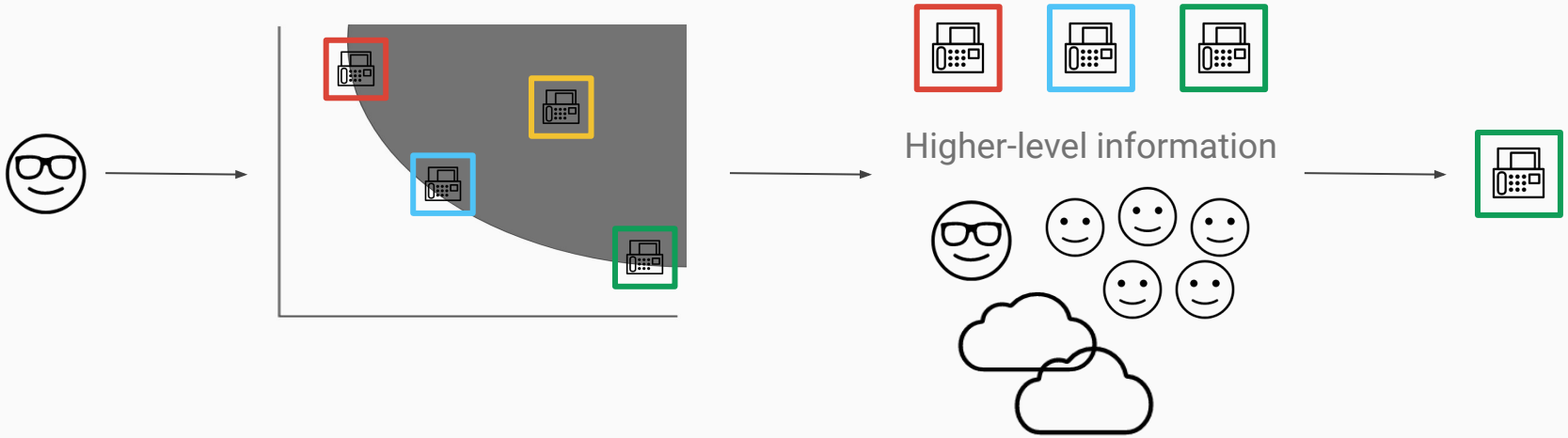
In this setting, y_1, y_2, and y_3 are on the optimal front because there are no feasible systems with better f_1 and f_2 values.



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*.



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.

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)

Tobias Schnabel et al. **Shaping Feedback Data in Recommender Systems with Interventions Based on Information Foraging Theory.** WSDM 2019.
Jean Garcia-Gathright et al. **Understanding and Evaluating User Satisfaction with Music Discovery**. SIGIR 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)

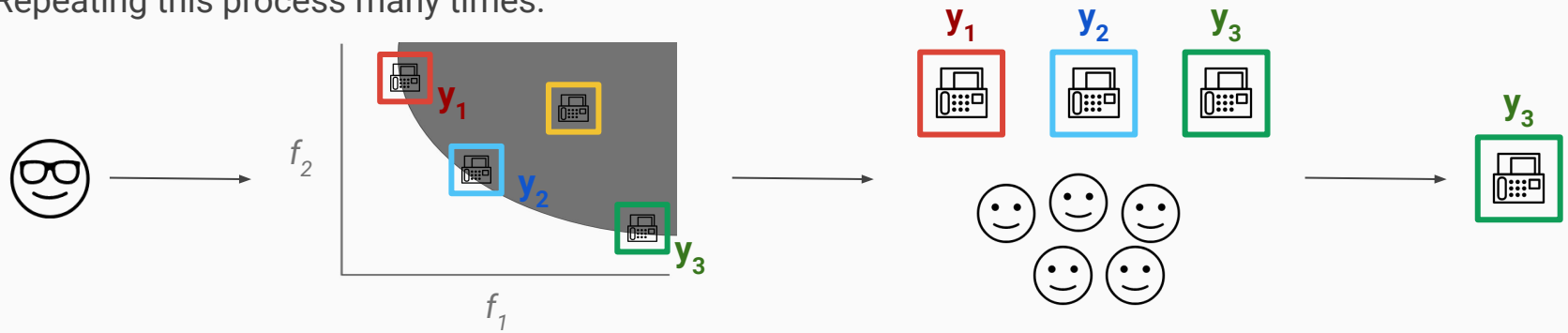
Conditioned on a user having a goal that the system can help with, does the user engage with the system again (e.g Hohnhold et al 2015, Dmitriev et al 2016, Poyarkov et al 2016, Machmouchi et al 2017)

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.

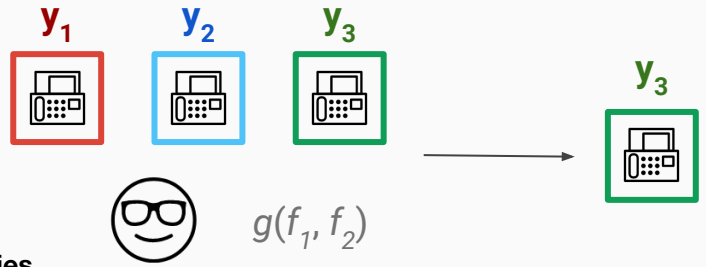
Surrogate Problems Approximate Expensive Higher-Level Information

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.



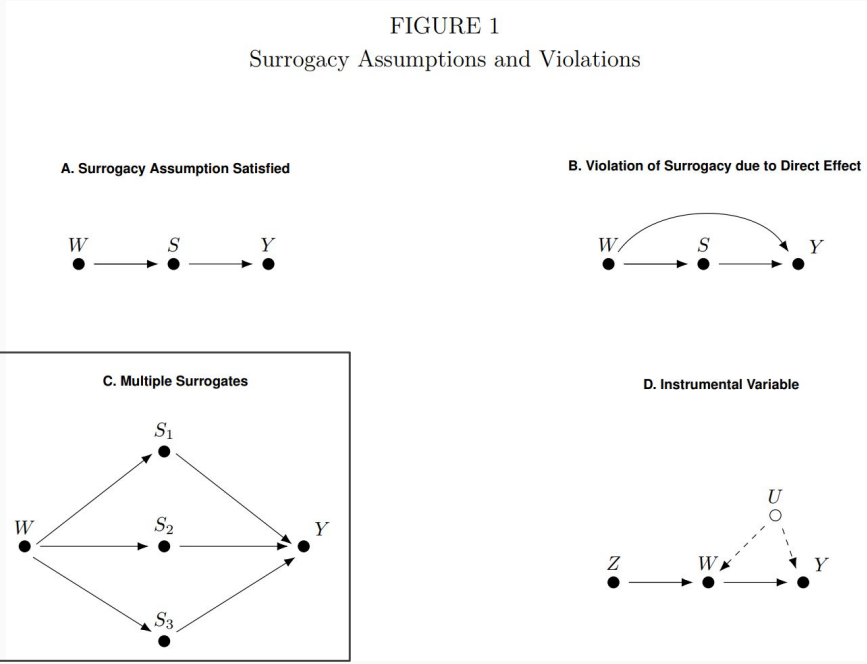
Richard Allmendinger et al. **Surrogate-Assisted Multicriteria Optimization: Complexities, Prospective Solutions, and Business Case.** J. of Multi-Criteria Decision Analyses 2017.

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:

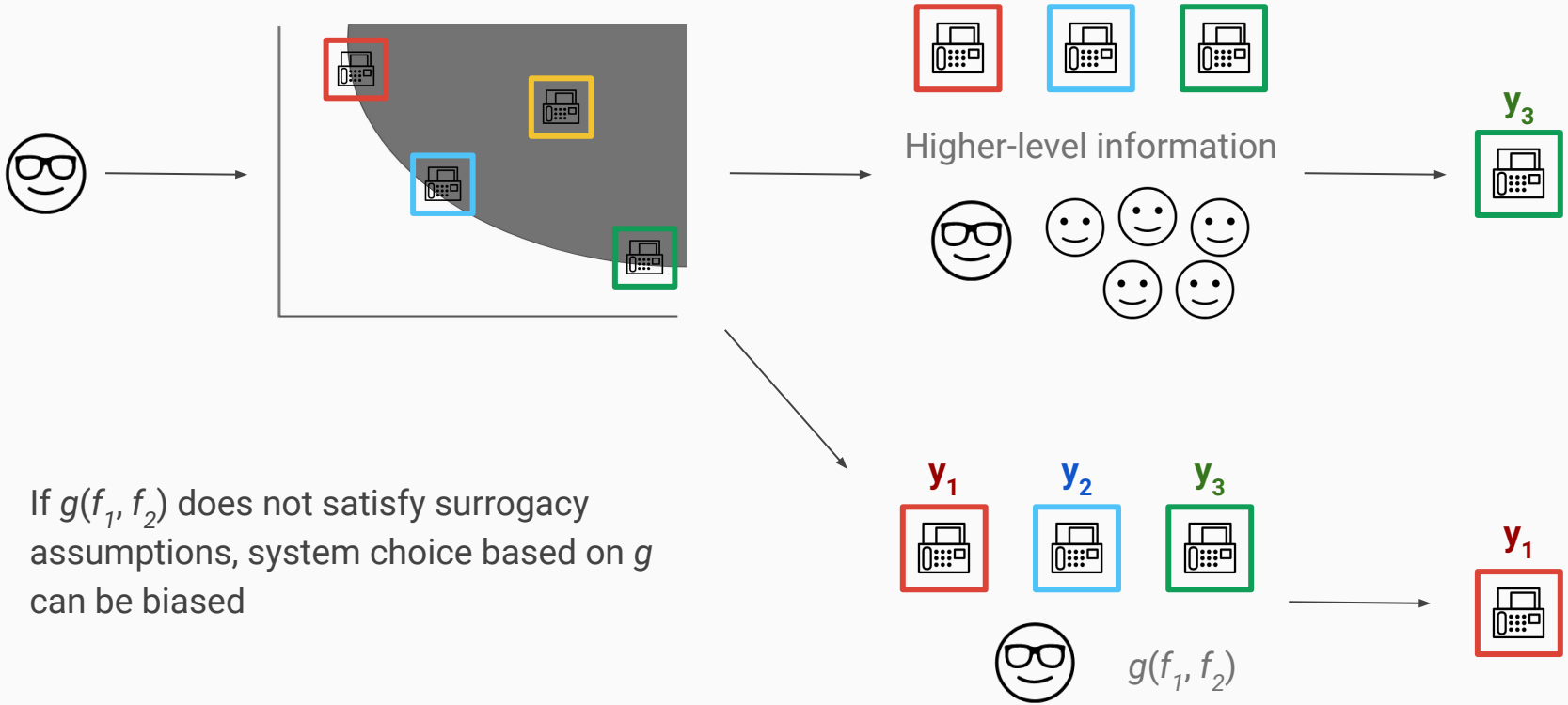
$P(Y, W \mid S) = P(Y \mid S)\, P(W \mid S)$



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

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.

Risks of Violating Surrogacy Assumptions



If $g(f_1, f_2)$ does not satisfy surrogacy assumptions, system choice based on g can be biased

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.

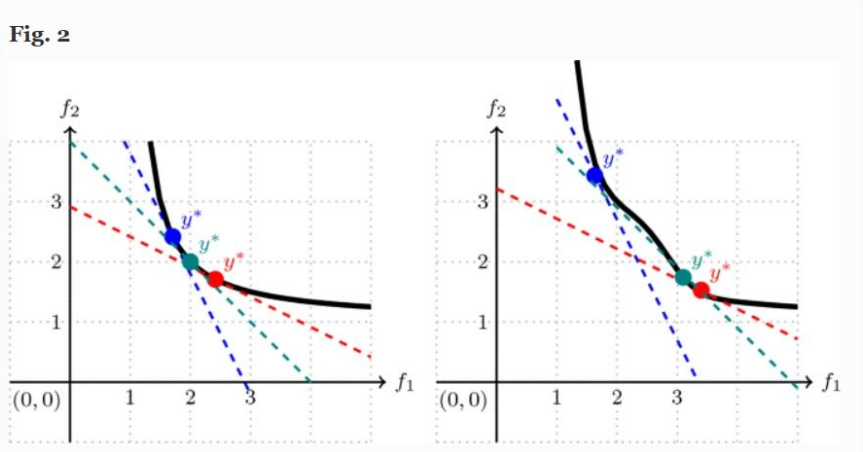
Solutions to

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

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

Concave pareto front

Convex pareto front



Michael T. M. Emmerich & André Deutz. **A tutorial on multiobjective optimization: fundamentals and evolutionary methods.** Natural Computing, 2018.

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

Learned preferences from users

$$\tilde{U}_{CTR} \approx k_1 \cdot \Delta\text{AdRelevance} + k_2 \cdot \Delta\text{LandingPageQuality} \quad (7)$$

High cost of measurement
Higher information decision criteria

Low cost of measurement metrics

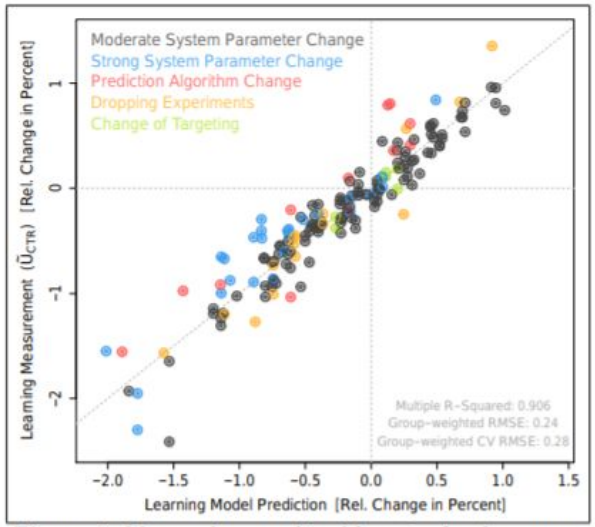


Figure 4: Measured vs. predicted learning for the current desktop macro-model.

Dmitriev et al, 2016 discuss challenges in collecting long-term decision criteria for this type of analysis

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.

Meta-modeling to construct surrogate functions

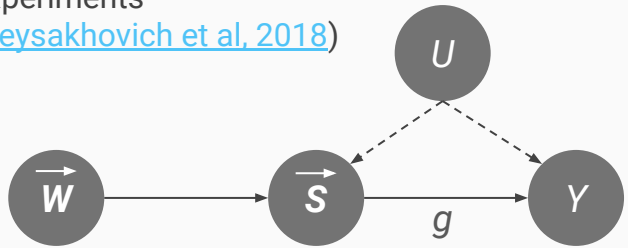
Peysakhovich et al, 2018 harvest instrumental variables from treatment assignments in many randomized experiments to learn causal relationships between low-cost of measurement metrics and decision criteria metrics

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.

Multiple surrogates (Athey et al, 2019)

Causal Effects from Randomized Experiments (Peysakhovich et al, 2018)

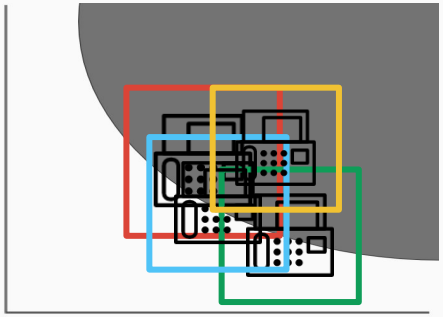


Alexander Peysakhovich & Dean Eckles. **Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables**. WWW 2018.
Zenan Wang et al. **Causal Meta-Mediation Analysis: Inferring Dose-Response Function From Summary Statistics of Many Randomized Experiments.** KDD 2020.

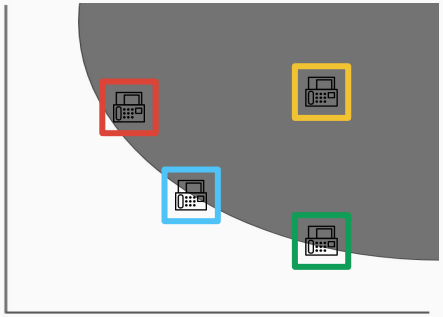
Metric Sensitivity

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

Low Sensitivity



High Sensitivity



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)

Alex Deng & Xiaolin Shi. **Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned**. KDD 2016.
Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. **Learning Sensitive Combinations of A/B Test Metrics.** WSDM 2017.

Components of Metric Sensitivity

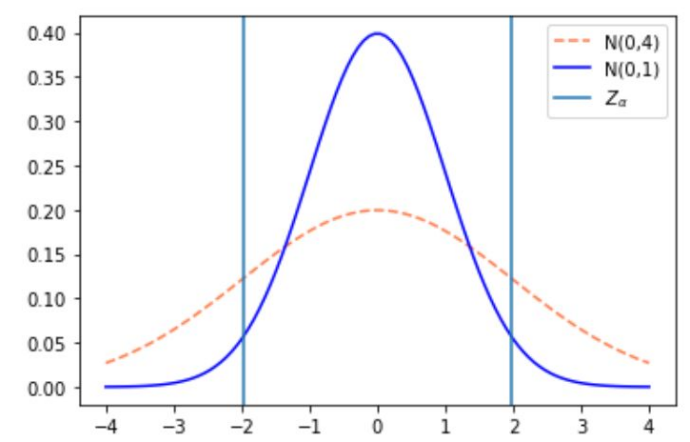
Using the decomposition of Deng 2015, the probability that there is a difference on the metric *and* that the difference is detected is

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

For

- test statistic Z
- critical value Z_α
- 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: δ~N(0,4) corresponds to a more sensitive metric than H_1: δ~N(0,1)



Alex Deng. **Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments**. WWW 2015

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

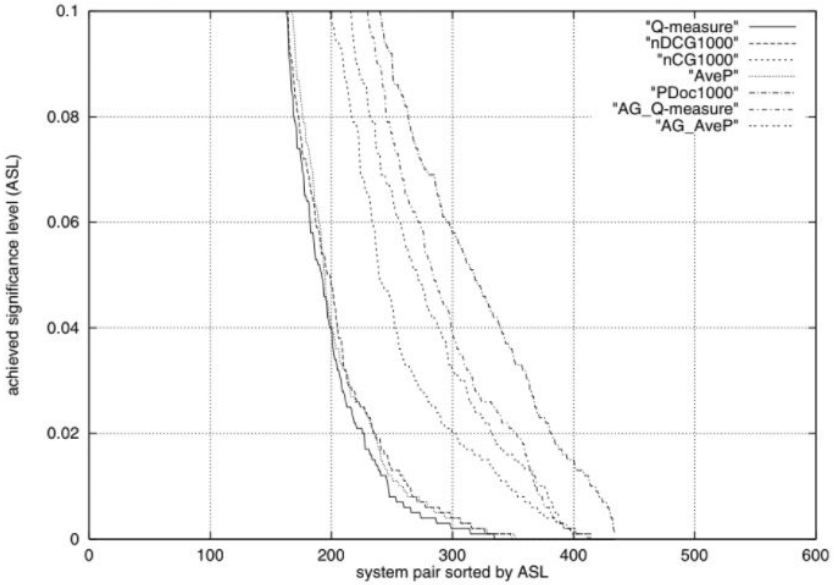


Figure 7: ASL curves based on Paired Bootstrap Hypothesis Tests (NTCIR-3 CLIR Chinese data).

Sakai 2006

Bradley Efron & Robert Tibshirani. **Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy**. Stat. Sci. 1986.
Tetsuya Sakai. **Evaluating evaluation metrics based on the bootstrap**. SIGIR 2006.
Jacques Savoy. **Statistical Inference in retrieval effectiveness evaluation.** Information Processing & Management 1997.

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

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.

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

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

Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. **Learning Sensitive Combinations of A/B Test Metrics.** WSDM 2017.

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

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

Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. **Learning Sensitive Combinations of A/B Test Metrics.** WSDM 2017.

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

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

Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. **Learning Sensitive Combinations of A/B Test Metrics.** WSDM 2017.

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

Eugene Kharitonov, Alexey Drutsa, Pavel Serdyukov. **Learning Sensitive Combinations of A/B Test Metrics.** WSDM 2017.

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:

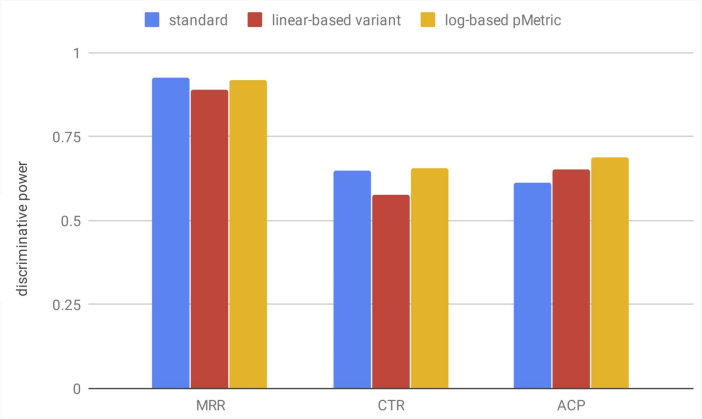
$$M = \frac{1}{n}\Sigma_{i=1}^{n} M_i$$

Personalized metric:

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

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

Figure 1: Discriminative power of metrics under the pairwise significance test with a significance level of 0.05.



Alex Deng et al. **Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data**. WSDM 2013.
Azin Ashkan & Donald Metzler. **Revisiting Online Personal Search Metrics with the User in Mind**. SIGIR 2019.

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

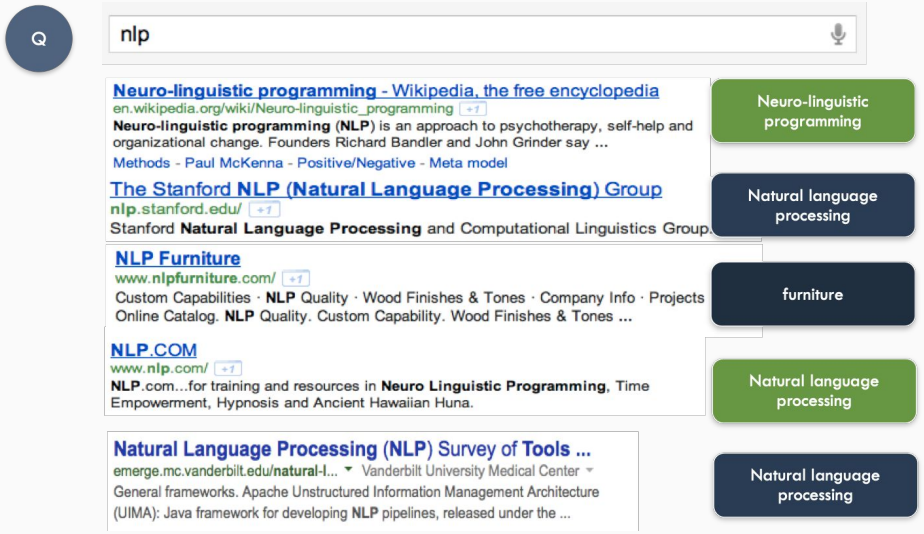
challenges in developing user-focused metrics

challenges in learning with user-focused metrics

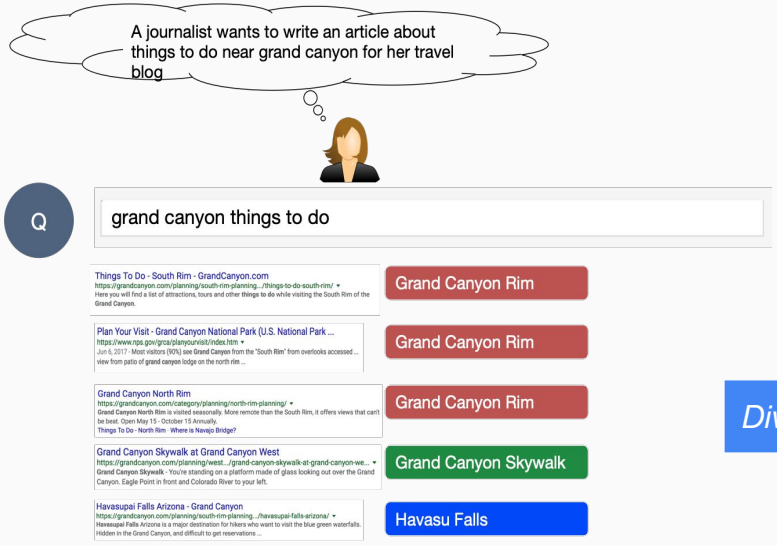
GRAVEYARD

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*



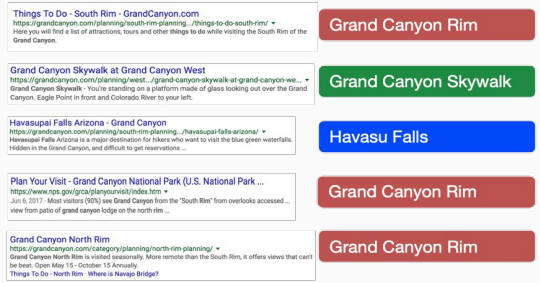
Novelty and Diversity in Rankings



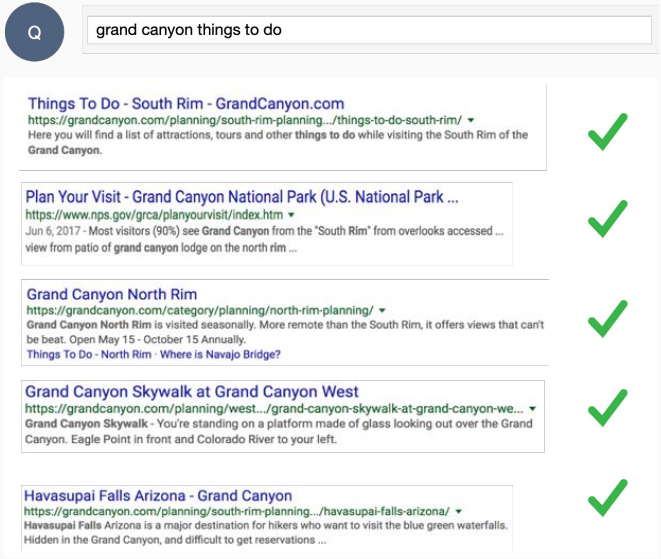
- ***Intrinsic Diversity***
  Addresses the need to reduce redundancy even when user requests are clearly specified.

  Examples: exploratory search

Diversified Ranking

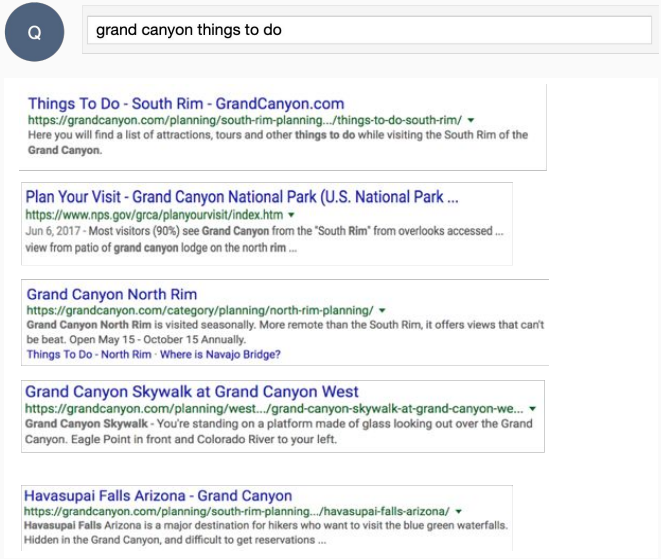


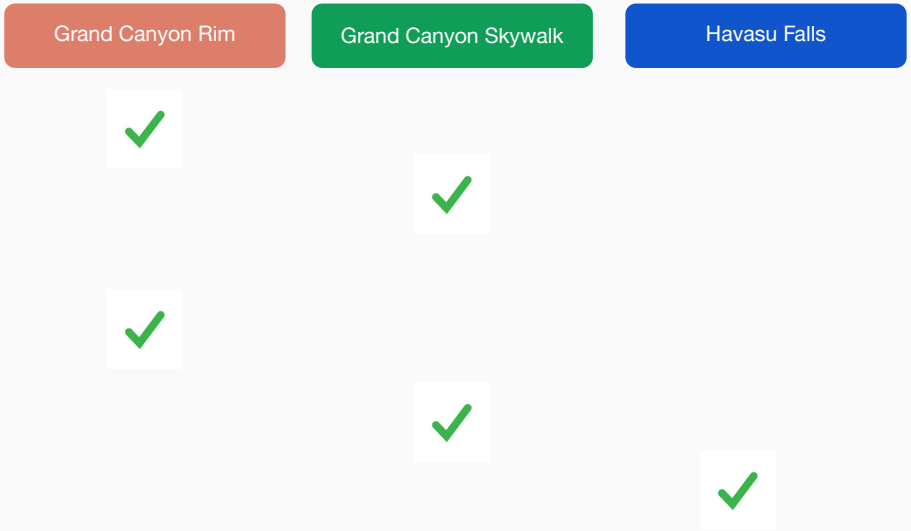
Novelty & Diversity: Subtopic-Based Evaluation



Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack et al. **Novelty and Diversity in Information Retrieval Evaluation** SIGIR 2008.

Novelty & Diversity: Subtopic-Based Evaluation





A variation of DCG metric that rewards novelty and penalizes diversity

$$\mathrm{DCG}[k] = \sum_{j=1}^{k} G[j]/\left(\log_2(1+j)\right)$$

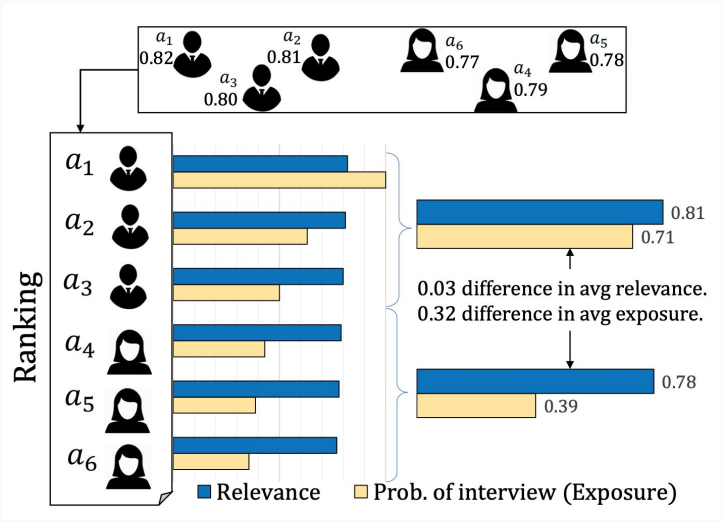
$$\mathrm{G}[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}$$

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack et al. **Novelty and Diversity in Information Retrieval Evaluation** SIGIR 2008.

Fairness in Rankings

Image Search Results for “CEO”

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



Ashudeep Singh, and Thorsten Joachims. **Fairness of Exposure in Rankings.** KDD 2018.

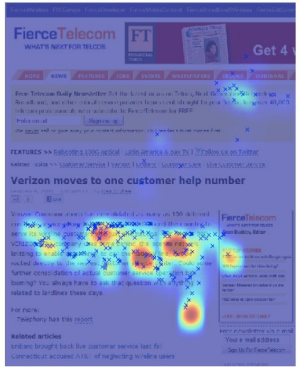
Validating Click Feedback

Also address click bait here ?

Thorsten Joachims et al. **Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search.** Info.Sys. 2007.
Yisong Yue et al. **Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data.** WWW 2010

Combining Cursor Movements & Dwell Time

dwell: time of the page view in seconds	0.167**
rank: 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
cursorcnt: num. of cursor movements	0.164**
cursorfreq: cursorcnt/dwell	-0.082*
dist: total overall distance the cursor traveled in pixels	-0.137**
xdist: total distance the cursor traveled horizontally in pixels	0.101**
ydist: total distance the cursor traveled horizontally in pixels	0.172**
speed: dist/dwell	0.101**
xspeed: xdist/dwell	-0.143**
yspeed: ydist/dwell	-0.124**
xmin: minimal x coordinate	0.112**
ymin: minimal y coordinatee	0.093*
xmax: maximal x coordinate	0.067
ymax: maximal y coordinate	0.243**
xrange: xmax-xmin	-0.006
yrange: ymax-ymin	0.172**



(a) relevant (dwell time: 30s)

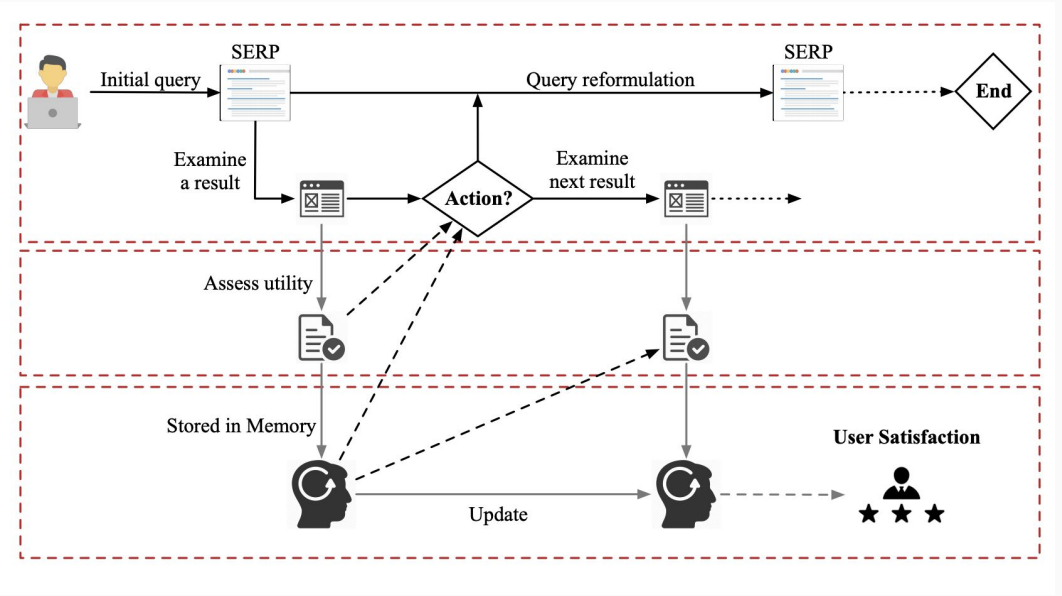
(b) non-relevant (dwell time: 30s)

Alexandra Papoutsaki et al. **WebGazer: Scalable Webcam Eye Tracking Using User Interactions.** IJCAI 2016.

Qi Guo and Eugene Agichtein. **Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior.** WWW 2012.

Directionally Explicit Signals: Lab Studies

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



Fan Zhang et al. **Cascade or Recency: Constructing Better Evaluation Metrics for Session Search.** SIGIR 2020.

Mengyang Liu et al. **Towards Designing Better Session Search Evaluation Metrics.** SIGIR 2018.

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

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search result diversification, 2015.

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
τ query sequence discount
π^q ranking for query q

K. Jarvelin, S. L. Price, L. M. L. Delcambre, M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query IR sessions. ECIR, 2008