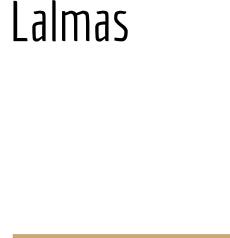


# Tutorial on Online User Engagement: Metrics and Optimization

Liangjie Hong & Mounia Lalmas



# Outline

Introduction and Scope

Metrics

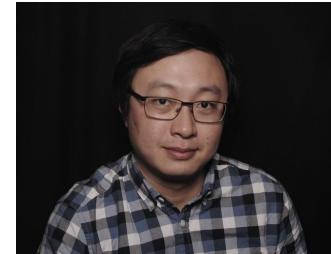
Optimization

Concluding Remarks & Future Directions

# Who we are

- Liangjie Hong, Director of Engineering - AI at LinkedIn, Sunnyvale

- Research interests: search, recommendation, advertising and now help people finding jobs
  - Website: <https://www.hongliangjie.com/>



- Mounia Lalmas, Research Director & Head of Tech Research @ Personalization at Spotify, London

- Research interests: user engagement in areas such as advertising, digital media, search, and now help people enjoying audio
  - Website: <https://mounia-lalmas.blog/>



# Acknowledgements

This tutorial is based on:

- “[Tutorial on Online User Engagement: Metrics and Optimization](#)”, The 2019 World Wide Web Conference (WWW 2019), San Francisco, May 2019.
- “[Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce](#)”, 11th ACM International Conference on Web Search and Data Mining (WSDM), Los Angeles, February 2018.
- “[Tutorial on Measuring User Engagement](#)”, 22nd International Conference on World Wide Web (WWW), Rio de Janeiro, May 2013.



# Introduction and Scope

# Introduction

Definitions

Scope

Case studies

# What is user engagement?

... Some definitions

User engagement is regarded as a **persistent** and **pervasive** cognitive affective state, not a time-specific state.

Wilmar Schaufeli, Marisa Salanova, Vicente González-romá and Arnold Bakker. **The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach**. Journal of Happiness Studies, 2002.

# What is user engagement?

... Some definitions

User engagement refers to the quality of the user experience associated with the **desire** to use a technology.

Heather O'Brien and Elaine Toms. **What is user engagement? A conceptual framework for defining user engagement with technology.** JASIST, 2008.

# What is user engagement?

... Some definitions

User engagement is **a** quality of the user experience that emphasizes the positive aspects of interaction – in particular the fact of **wanting** to use the technology **longer** and **often**.

Simon Attfield, Gabriella Kazai, Mounia Lalmas and Benjamin Piwowarski. **Towards a science of user engagement (Position Paper)**. WSDM Workshop on User Modelling for Web Applications, 2011.

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

[1] Heather O'Brien and Elaine Toms. **What is user engagement? A conceptual framework for defining user engagement with technology.** JASIST 2008.

[2] Heather O'Brien. **Defining and Measuring Engagement in User Experiences with Technology.** Doctoral thesis, Dalhousie University, 2008.

[3] Simon Attfield, Gabriella Kazai, Mounia Lalmas and Benjamin Piwowarski. **Towards a science of user engagement (Position Paper).** WSDM Workshop on User Modelling for Web Applications, 2011.

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Users must be focused to be engaged

Distortions in subjective perception of time used to measure it

Time spent can be a good proxy of focused attention

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Sensory, visual appeal of interface stimulates user and promotes focused attention

Perceived usability

Linked to design principles (e.g. symmetry, balance, saliency)

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Novelty, surprise, unfamiliarity and the unexpected; updates & innovation

Relate to serendipity, discovery and freshness

Appeal to user curiosity

Encourage inquisitive behavior and promotes repeated engagement

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Trust is a necessary condition for user engagement

Implicit contract among people and entities which is more than technological

Habit can play a role

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Emotions experienced by user are intrinsically motivating

Initial affective “hook” can induce a desire for exploration, active discovery or participation

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

People remember enjoyable, useful, engaging experiences and want to repeat them

Repetition of use, recommendation, interactivity, utility

Relate the in-the-moment experience to future experience

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Richness captures the growth potential of an activity

Control captures the extent to which a person is able to achieve this growth potential

# Characteristics of user engagement

Focused attention	Aesthetics	Novelty	Reputation, trust and expectation
Positive affect	Endurability	Richness and control	Motivation, interests, incentives and benefits

Why should users engage?

# Characteristics of user engagement

## Quality of the user experience ... durability

People remember “satisfactory” experiences  
and want to repeat them

Endurability

We need metrics to quantify the quality of the  
user experience with respect to durability

[1] Heather L. O'Brien Elaine G. Toms. **What is user engagement? A conceptual framework for defining user engagement with technology** Journal of the American Society for Information Science and Technology, Volume 59, Issue 6, February 2008.

# Why is it important to engage users?

Users have increasingly enhanced expectations about their interactions with technology

... resulting in increased competition amongst the providers of (online) services.

utilitarian factors (e.g. usability, functionality) → hedonic and experiential factors of interaction (e.g. fun, fulfillment) → user engagement

# The engagement life cycle

## Point of engagement

### **How engagement starts**

Aesthetics & novelty in sync with user interests & contexts

## Period of engagement

### **Ability to maintain user attention and interests**

Main part of engagement and usually the focus of study → focus of this tutorial

## Disengagement

### **Loss of interests leads to passive usage & even stopping usage**

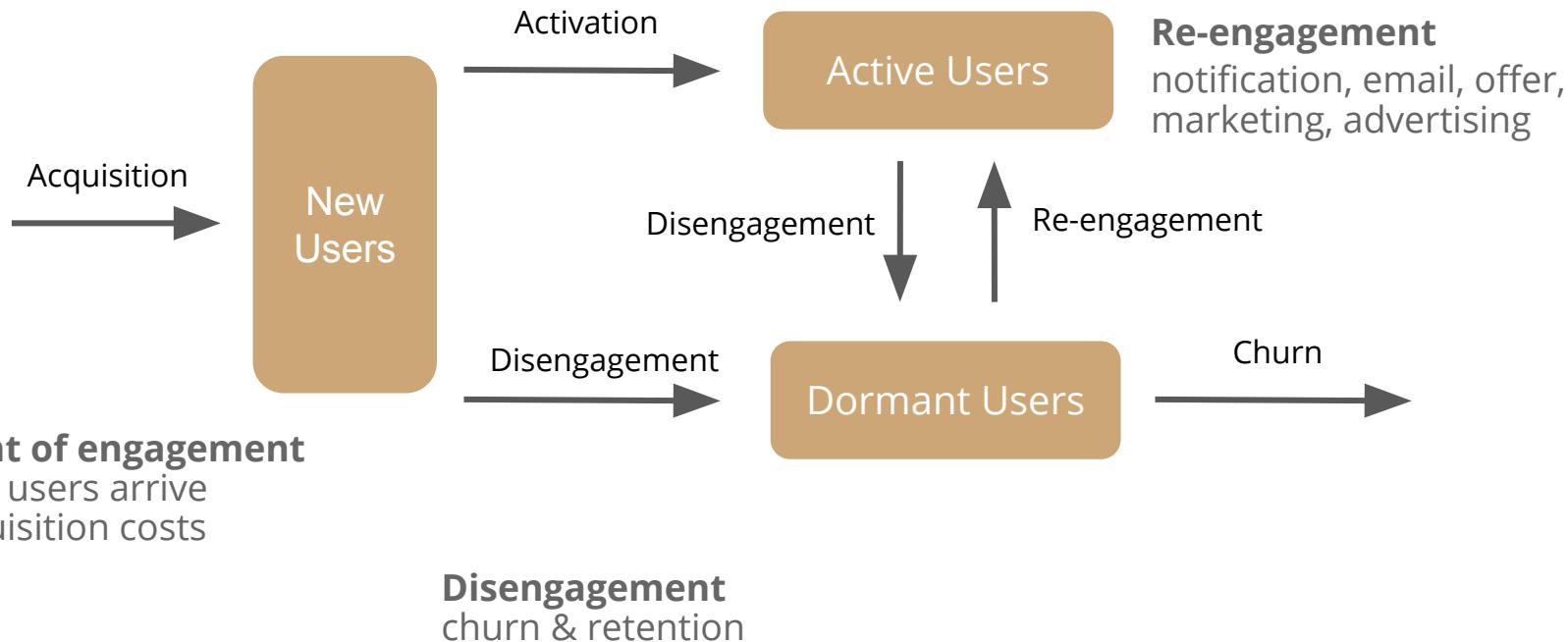
Identifying users that are likely to churn often undertaken

## Re-engagement

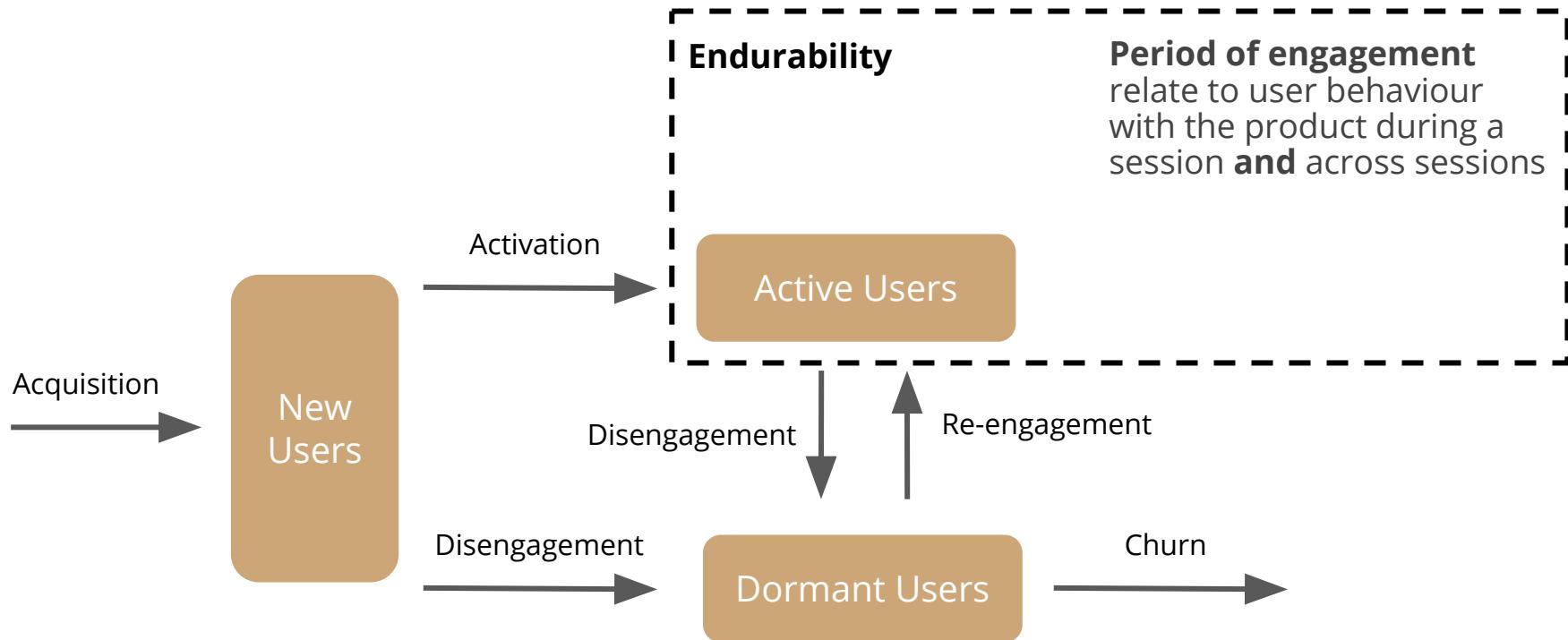
### **Engage again after becoming disengaged**

Triggered by relevance, novelty, convenience, remember past positive experience, sometimes as result of campaign strategy

# The engagement life cycle



# Endurability in the engagement life cycle



# Considerations in measuring user engagement

short term ↔ long term

laboratory ↔ “in the wild”

subjective ↔ objective

qualitative ↔ quantitative

large scale ↔ small scale

# Methods to measuring user engagement

**self-reported  
methods**  
subjective

Questionnaire,  
interview, report,  
product reaction cards

User study (lab/online)

**physiological  
measurements**  
objective

Task-based methods

Physiological

User study (lab/online)

**online  
analytics**  
objective

Data science  
Machine learning

Data study (online)

*mostly qualitative*

*mostly quantitative,  
scalability an issue*

*quantitative  
large scale*

# Scope of this tutorial

Focus on online analytics → online user engagement.

Assume that applications are “**properly designed**” in terms of usability and content.

Based on “published” work and our experience.

Focus on applications that users “chose” to engage with, widely used by “anybody” on a “large-scale” and on a mostly regularly basis.

This tutorial is not an “exhaustive” account of works in this and related areas.

# Case studies

E-commerce

Search

News

Streaming

Advertising

# E-Commerce

Amazon Prime

All liszt

Browsing History Liangjie's Amazon.com Today's Deals Gift Cards Registry Sell Help

1-16 of over 50,000 results for "liszt"

✓prime | FREE One-Day Get FREE One-Day Shipping on qualifying orders over \$35

Show results for CDs & Vinyl Classical Chamber Music Classical Sonatas Symphonies Classical Concertos See more Books Humor & Entertainment Composer & Musician Biographies Actor & Entertainer Biographies World History History See more Digital Music Classical See All 8 Departments

Refine by Delivery Date

Introducing Amazon music unlimited

Etsy wabi sabi

Jewelry & Accessories Clothing & Shoes Home & Living Wedding & Party Toys & Entertainment Art & Collectibles Craft Supplies & Tools Vintage

wabi sabi art wabi sabi ceramics wabi sabi bowl wabi sabi pottery wabi sabi necklace wabi sabi jewelry

Special offers On sale

All categories > "wabi sabi" (6,213 Results)

Kintsugi bowl, kintsugi ceramic enamel KaneloSuri \$84.41 Only 1 available ★★★★★ (77)

BIGFOOT Bowl, Charcoal Special | M... Odaka \$42.00 ★★★★★ (31)

Wabi-sabi Oversize Clutch bag for w... SCHILLERshop \$65.00 Free shipping ★★★★★ (107)

Wabi-Sabi definition, dictionary art p... footnotestudios \$4.99 ★★★★★ (87)

wabi-sabi t-shirt

Wabi Sabi: A Raw-Edge Machine Applique Quilt Pattern Karen Grana

wa-bi- sa-bi

wa-bi- sa-bi

Hi! Sign in or register | Daily Deals | Gift Cards | Help & Contact Perfect-for-Them Valentine Gifts

camera

All Categories Search Advanced

Related: digital camera security camera spy camera video camera nikon camera canon camera canon digital camera

2,241,207 results for camera Save this search

Sort: Best Match View: Group Similar Listings

Guaranteed 3 day delivery

Sell on Etsy Home Favorites You Cart

Shop by Category

Categories All Cameras & Photo Digital Cameras Film Cameras Camcorders Camera & Photo Accessories Camera Batteries More

2,241,207 results for camera Save this search

IT IN FREE 3 Delivery Fri, Feb. 9

wa-bi- sa-bi

IT IN FREE 3 Delivery Fri, Feb. 9

# E-Commerce



# Search

san Diego X | 🔍

All Maps Images News Videos More Settings Tools

About 1,030,000,000 results (1.03 seconds)

en.wikipedia.org · wiki · San\_Diego

## San Diego - Wikipedia

San Diego is a city in the U.S. state of California on the coast of the Pacific Ocean, approximately 120 miles (190 km) south of Los Angeles and immediately ...  
Named for: Saint Didacus of Alcalá  
State: California  
Area codes: 619/858  
Established: July 16, 1769  
San Diego County, California · Mayor of San Diego · History of San Diego · Vallecito

## Top sights in San Diego

Outdoors Beaches Museums Art and culture History Kid friendly Local favourites

**Balboa Park**  
4.8 ★★★★★ (54,203)  
Multi-attraction park covers 1,200 acres

**San Diego Zoo**  
4.7 ★★★★★ (37,284)  
Zoo, safari park & wildlife conservation

**SeaWorld San Diego**  
4.5 ★★★★★ (30,598)  
Aquatic theme park with shows & rides

More top sights

## People also ask

- What is San Diego famous for? ▾
- Is San Diego a dangerous city? ▾
- What salary do you need to live in San Diego? ▾
- Is San Diego expensive? ▾

Feedback

www.sandiego.gov

City of San Diego Official Website



## San Diego

City in California

San Diego is a city on the Pacific coast of California known for its beaches, parks and warm climate. Immense Balboa Park is the site of the renowned San Diego Zoo, as well as numerous art galleries, artist studios, museums and gardens. A deep harbor is home to a large active naval fleet, with the USS Midway, an aircraft-carrier-turned-museum, open to the public.

Local time: Friday 03:32

Weather: 22 °C, Wind NW at 0 mph (0 km/h), 87% Humidity

Population: 1.426 million (2018)

Mayor: Kevin Faulconer

Minimum wage: 12.00 USD per hour (1 Jan 2019)

## Plan a trip

San Diego travel guide

3-star hotel averaging £101, 5-star averaging £312

Upcoming events

Colleges and Universities: San Diego State University, MORE

Clubs and Teams: San Diego Padres, Houston Rockets, MORE

## People also search for

View 15+ more



white wine

🔍

All Shopping Images News Videos More Settings Tools

About 58,400,000 results (0.55 seconds)

## White Wines You'll Love | Drizly

https://drizly.com/white-wine/c8 ▾

Buy white wine at a great price through Drizly and have it delivered directly to your door. With the largest selection of wine online, it's easy to find the right bottle for you. Shop Chardonnays, Sauvignon Blancs, Rieslings and more.

Barefoot Pinot Grigio \$3.99 · Oyster Bay Sauvignon Blanc · Chardonnay · Riesling



## White wine

White wine is a wine whose colour can be straw-yellow, yellow-green, or yellow-gold. It is produced by the alcoholic fermentation of the non-coloured pulp of grapes, which may have a skin of any colour. [Wikipedia](#)

## Nutrition Facts

White wine ▾

Amount Per 1 serving 5 fl oz (147 g) ▾

Calories 120

% Daily Value\*

Total Fat 0 g

0%

## White wine - Wikipedia

# Search

## Search engine evaluation

- Coverage
- Speed
- Query language
- User interface
- Relevance

## User satisfaction

Users find what they want and return to the search engine for their next information need → **user engagement**

## But let us remember:

In carrying out a search task, search is a means, not an end

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. **Modern Information Retrieval: The Concepts and Technology behind Search**. ACM Press Books, 2nd Edition, 2011.

[2] Christopher Manning, Prabhakar Raghavan and Hinrich Schütze. **Introduction to Information Retrieval**. Cambridge University Press, 2008.

# News

yahoo!

Search


**Celebrity** in The Know  
**Ex-Disney Channel star joins adult content site: 'This is how bored she really is'**

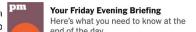
The former Disney Channel star's annot times.

Saturday, August 22, 2020



**Ad Stansberry Research**  
**Man Who Predicted Rise of**

He called bottom of stocks in '09, and now 1,800%. Now he has a surprising new p

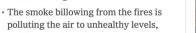


Your Friday Evening Briefing Here's what you need to know at the end of the day.



**Celebrity** SheKnows  
**Salma Hayek Brings the Heat in Teeny Bikini**

As if the summer wasn't hot enough already, thermostats a few more clicks. Currently



**U.S. Motorious**  
**Driver Successfully Outruns Live**



Why does California have so many wildfires? There are many reasons, and climate change figures prominently.



**'It's Hard to See Your Memories Burn': Loss From Fires Grows in California**

Some of the ancient, towering conifers in Big Basin Redwoods State Park are a casualty of the fires that have ravaged the state.

**DeJoy Tells Senators Election**

Sign In



Santa Clara

Today  
88° 62°  
Sat  
82° 64°  
Sun  
82° 64°

ENGLISH ESPAÑOL 中文

# The New York Times

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Subscribe

Sign In

Search

## Covid-19 Pay Cuts Coming to an End at Some Companies

Some companies are beginning to restore cuts they made to managers' salaries and bonuses, a sign that some industries—and their white-collar workers—are benefiting from glimmers of a recovery while millions of others continue to endure job and income losses.

\* New U.S. Virus Cases Stay Below 50,000

## Home Sales Reach Lofty Heights

Sales of previously owned homes in the U.S. surged in July as low interest rates and a desire for more space amid the pandemic boosted home-buyer demand. • 221

- Home-Improvement Stocks Stay the Course
- U.S. Stocks Finish Higher



KEYWORDS

SUBSCRIBE NOW

LOG IN

TODAY'S PAPER

OPINION

## Covid-19 Is Dividing the American Worker

Adoption of remote work and automation could accelerate inequalities already in place. Economists say the resulting 'K' shaped recovery will be good for professionals—and everyone else.



## Opinion

Maureen Dowd

**Joe's Fearless Weapon Against Trump: Simple Decency**  
Biden takes on Donnie Darkness and promises to bring us into the light.



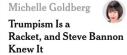
## The Editorial Board

**Another Putin Critic Apparently Poisoned**  
Will the West take action against whoever may have harmed Alexei Navalny?



**David W. Blight**  
**Obama's Call to Save America**

With the Speech of His Life, Joe Biden Becomes the Man for This Moment



**Paul Krugman**  
**Stocks Are**

Roger Cohen  
**Where Hope and**



## WEEKEND READS



TECH

**Remote Work Is Reshaping San Francisco, as Tech Workers Flee and Rents Fall**

## OPINION

**The Normalcy of Trump's Republican Party**

*By Michael Barone | Commentary*

**Joe Biden's Me-Too Covid Plan**

*By The Editorial Board | Review & Outlook*

**Democrats Make This Nonnative Restless**

*By Tunku Varadarajan | Commentary*

# News

[ENGLISH](#) [ESPAÑOL](#) [中文](#)

Tuesday, May 7, 2019

**The New York Times**

[SUBSCRIBE NOW](#) [LOG IN](#)

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Real Magazine Estate Video

**Your Tuesday Evening Briefing**  
Here's what you need to know at the end of the day.

**TRUMP'S TAXES**  
Decade in the Red:  
Trump Tax Figures Show  
Over \$1 Trillion Loss  
Donald J. Trump was able to get to the presidency, in part, by a self-spun narrative of business success and of setbacks triumphantly overcome.  
But 10 years of tax information, from 1985 to 1994, confirmed by The Times, is a far bleaker picture of his financial condition. Read our exclusive report.  
2h ago

**1 Dead and 7 Injured in Colorado School Shooting**  
Several of the students were in critical condition, the police said. Two suspects, also students, were in custody.  
But weeks after a school joined others in the Denver area in closing over security concerns as the 20th anniversary of the Columbine shooting neared.  
29m ago

**TRUMP ADMINISTRATION**  
Clash Between Trump and House Democrats Threatens Constitutional Order  
Earlier administrations had isolated themselves with congressional subpoenas. The clash between President Trump and House Democrats is turning into an all-out battle.  
Legal experts have warned that Mr. Trump's opposition to what he sees as partisan meddling could create a constitutional crisis.  
3h ago

**Sandra Bland Filmed Traffic Stop Confrontation Herself**  
A newly obtained cellphone video shows Ms. Bland's view of a traffic stop in Texas in 2015, days before she died in jail in

**Listen to 'The Daily'**  
The Chinese surveillance state, Part 2.

**In the 'Smarter Living' Newsletter**  
Why giving up is often the best way to solve a problem.

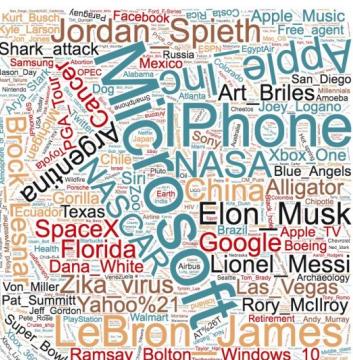
**Opinion > Google's Sundar Pichai: Privacy Should Not Be a Luxury Good**  
Yes, we use data to make products more helpful for everyone. But we also protect your information.  
1h ago 44 comments

**Lizzie Skurnick**  
**The Royal Baby and Blackness as a Badge of Honor**  
The Court Handling Trump's Lawsuit Must Move at Breakneck Speed  
Whether he "looks" black or not, I'll be glad to claim him.  
6h ago 70 comments

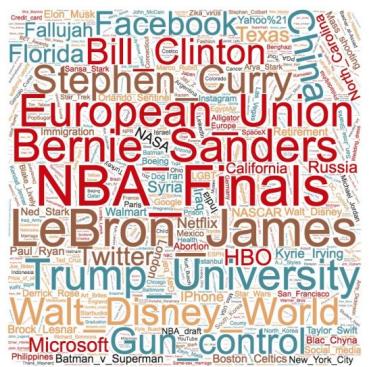
**Thomas L. Friedman**  
**How to Defeat Trump**  
Veronique de Rugy  
**Corporate Welfare Wins Again in Trump's Washington**  
The Editorial Board  
Sorry, Steve Mnuchin. Congress Has a Right to See Trump's Tax Returns

**Editors' Picks**  
A Wine-Soaked Ball Unites a Georgian Village, but Only After Dividing It

**In a cellphone video recorded by Sandra Bland, Trooper Brian Encinia is seen drawing his stun gun and heard saying, "I'm going to rip you up. Get out."**  
In every spring in Shukhuti, a black leather ball is sewn together to play Lelo Burti, a brutally physical ball game. The winners carry it



(a) Top clicked articles



(b) Top returning articles

# Streaming

amazon music

DEEZER

DLNA

iHeartRADIO

Internet Radio

KKbox

napster

pandora<sup>®</sup>

qobuz

QQ Music

RADIO.COM SiriusXM

SOUND MACHINE

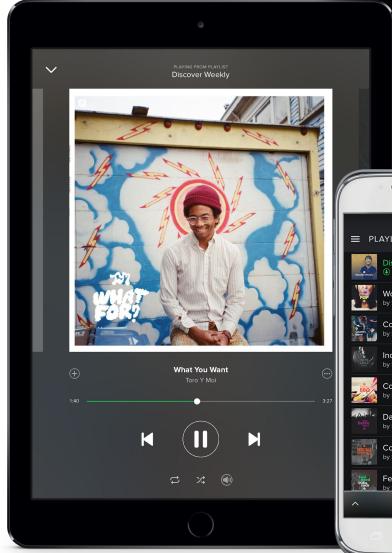
Spotify

TIDAL

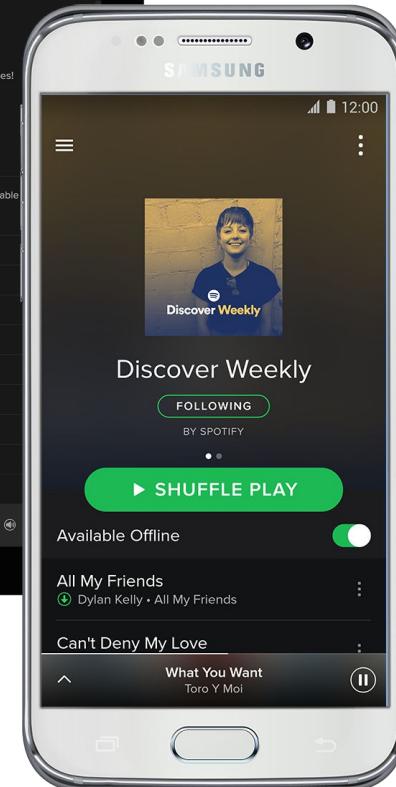
TUNE IN



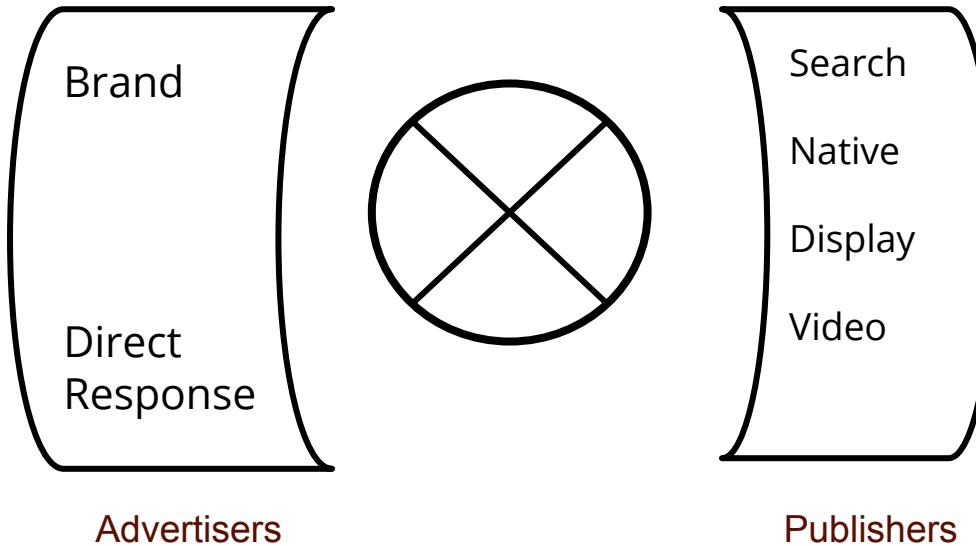
# Streaming



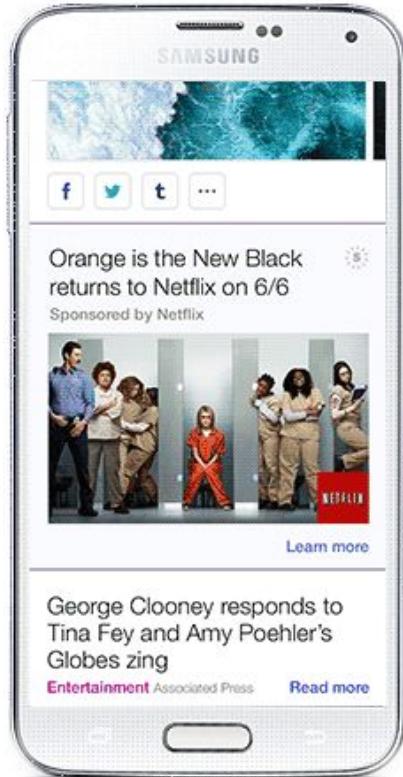
SONG	ARTIST	ALBUM	LAST LISTENED
All My Friends	Dylan Kelly	All My Friends	a day ago
Can't Deny My Love	Brandon Flowers	Can't Deny My Love	a day ago
1000	Ben Khan	1000	a day ago
Fly (FKJ Remix)	June Marieezy	Fly (Remixes) - Single	a day ago
Eventually	Tame Impala	Eventually	a day ago
Abrasive	Retstat	Abrasive	a day ago
Can't You Hear Me Knocking - Alternate Ver...	The Rolling Stones	Can't You Hear Me Knocking	a day ago
OKAGA, CA	Tyler, The Creator	Cherry Bomb	a day ago
That's Love	Oddisee	That's Love - Single	a day ago



# Advertising



# Native advertising



Visually engaging

Higher user attention

Higher brand lift

Social sharing



# Metrics

# Online metrics

Terminology, context & consideration

Intra-session metrics

Inter-session metrics

Other metrics

# Measures, metrics & key performance indicators

## Measurement:

process of obtaining one or more quantity values that can reasonably be attributed to a quantity

e.g. number of clicks

## Metric:

a measure is a number that is derived from taking a measurement ... in contrast, a metric is a calculation

e.g. click-through rate

## Key performance indicator (KPI):

quantifiable measure demonstrating how effectively key business objectives are being achieved

e.g. conversion rate

a measure can be used as metric but not all metrics are measures  
a KPI is a metric but not all metrics are KPIs

# Three levels of metrics

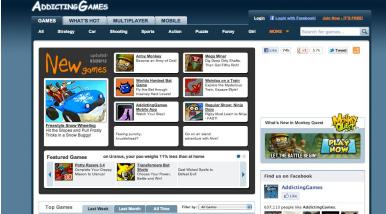
**Business metrics** -- KPIs

**Behavioral metrics** -- online metrics, analytics

**Optimisation metrics** -- metrics used to train machine learning algorithms

These three levels are connected

# Why several metrics?



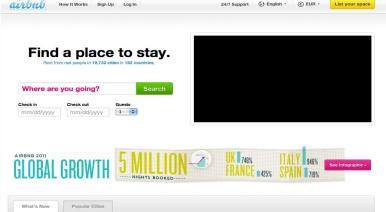
## Games

Users spend much time per visit



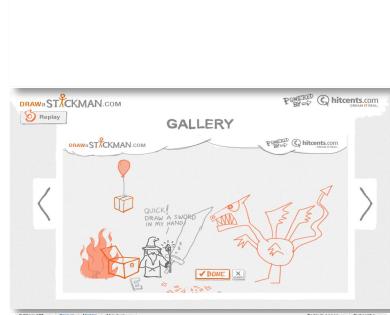
## Social media

Users come frequently & stay long



## Service

Users visit site, when needed



## Search

Users come frequently but do not stay long

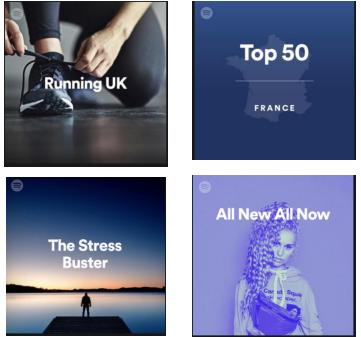
## Niche

Users come on average once a week

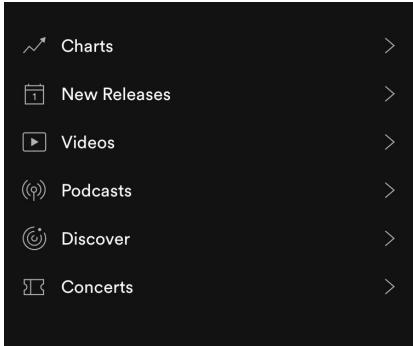
## News

Users come periodically, e.g. morning and evening

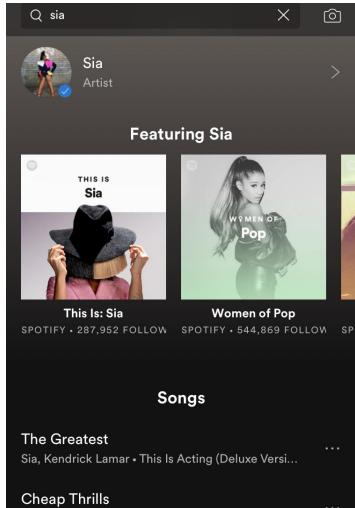
# Why several metrics?



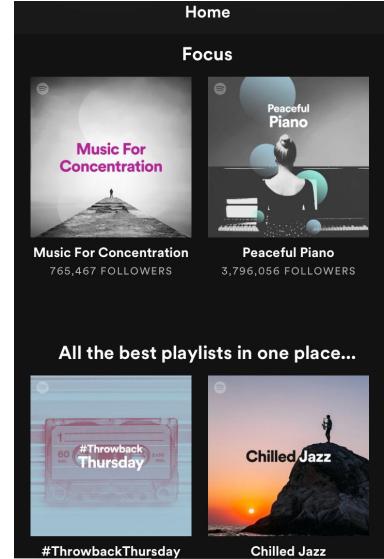
Playlists differ in their listening patterns.



Media type and freshness lead to different engagement patterns.

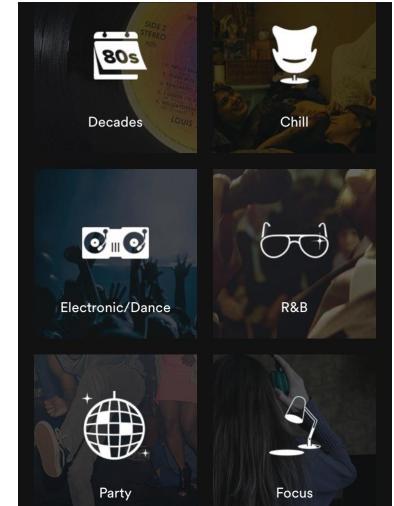


Searching has a particular engagement pattern.

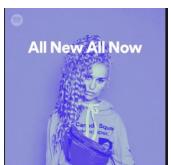
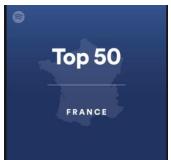


Home can be viewed as a hub with a "star" style engagement pattern.

Genres and moods can be viewed as sub-hubs, each with some common engagement patterns.



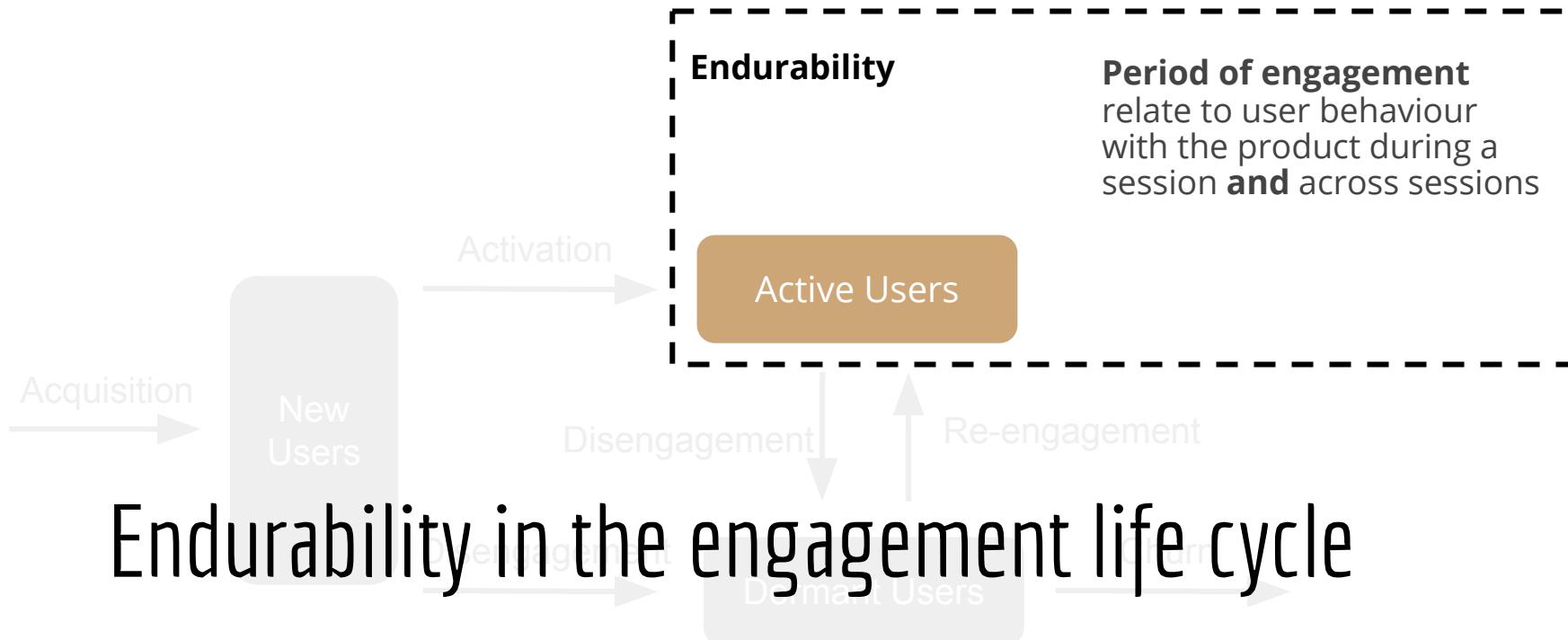
# Why several metrics?



Leaning in	Active	Occupied	Leaning back
<b>Playlists types</b> Pure discovery sets Trending tracks Fresh Finds	<b>Playlists types</b> Hits flagships Decades Moods	<b>Playlists types</b> Workout Study Gaming	<b>Playlists types</b> Sleep Chill at home Ambient sounds
<b>Playlist metrics</b> Downstreams Artist discoveries # or % of tracks sampled	<b>Playlist metrics</b> Skip rate Downstreams	<b>Playlist metrics</b> Session time Skip rate	<b>Playlist metrics</b> Session time

# Quality of the user experience

... endurability



# Three levels of engagement related to endurability

## Involvement

### **Presence of a user**

pageview, dwell time, playtime, revisit rate

## Interaction

### **Action of a user**

click-through rate, share, like, conversion rate, save, click, skip rate

## Contribution

### **Input of a user**

post, comment, create, update, reply, upload, playlist

# Three levels of engagement related to endurability

## Involvement

### **Presence of a user**

pageview, dwell time, playtime, revisit rate

## Interaction

### **Action of a user**

click-through rate, share, like, conversion rate, save, click, skip rate

## Contribution

### **Input of a user**

post, comment, create, update, reply, upload, playlist

What involvement is in application A may be interaction in application B

For interaction level, click is a “special” **action**, and is often a precursor of other levels of engagement  
Value of a click → **downstream engagement**

# Three levels of engagement related to endurability

## Involvement

### **Presence of a user**

pageview, dwell time, playtime, revisit rate

## Interaction

### **Action of a user**

click-through rate, share, like, conversion rate, save, click, skip rate

## Contribution

### **Input of a user**

post, comment, create, update, reply, upload, playlist

**Degree** of engagement in terms of “intention” increases from **involvement → interaction → contribution**

**Retention** increases from **involvement → interaction → contribution**

# From visit to session



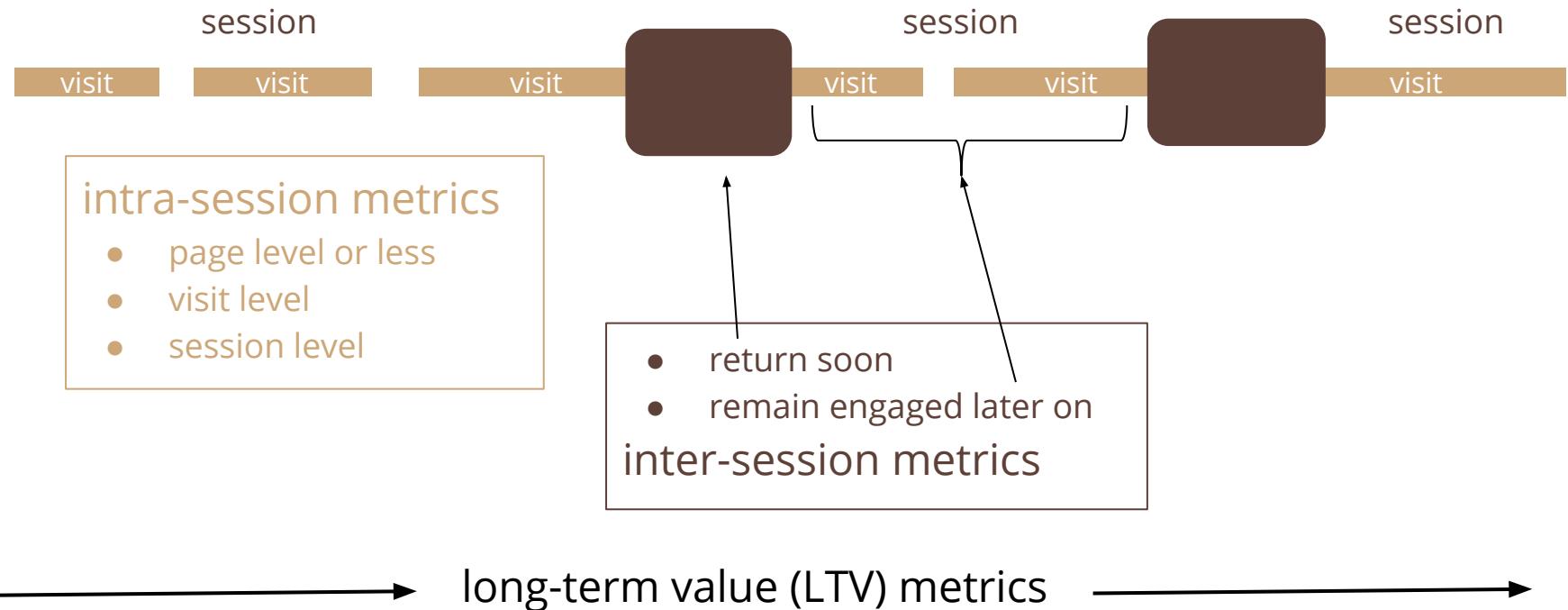
Dwell time = time spent on site (page) during a visit

Session length is amount of time user spends on site within the session

Session frequency shows how often users are coming back (loyalty)

Often 30mn is used as threshold for session boundary (desktop)

# From endurability to loyalty



# Intra- vs inter-sessions metrics

- intra-session engagement measures user activity on the site during the session → durability
- inter-session engagement measures user habit & loyalty with the site → long-term value

Intra-session (within → durability)	Granularity	inter-session (across → habit)
<b>Involvement</b> <ul style="list-style-type: none"><li>• Dwell time</li><li>• Session duration</li><li>• Page view (click depth)</li><li>• Revisit rate</li><li>• Bounce rate</li></ul>	Module ↓ Viewport ↓ Page ↓ Visit ↓ Session	<b>From one session to the next session (return soon)</b> <ul style="list-style-type: none"><li>• Time between sessions (absence time)</li></ul>
<b>Interaction</b> <ul style="list-style-type: none"><li>• Click-through rate (CTR)</li><li>• Number of shares, likes, saves</li><li>• Conversion rate</li><li>• Streamed, played</li></ul>		<b>inter-session (across → loyalty)</b> <b>From one session to a next time period such next week, or in 2 weeks time (remain engaged later on)</b> <ul style="list-style-type: none"><li>• Number of active days</li><li>• Number of sessions</li><li>• Total usage time</li><li>• Number of clicks</li><li>• Number of shares</li><li>• Number of thumb ups</li></ul>
<b>Contribution</b> <ul style="list-style-type: none"><li>• Number of replies</li><li>• Number of blog posts</li><li>• Number of uploads</li><li>• Number of playlists</li></ul>		

# Intra- vs inter-sessions metrics ... Granularity

## Intra-session metrics

Module → Viewport → Page → Visit → Session

Optimisation mostly with these metrics, with increasing complexity from “Module” to “Session”

## Inter-session metrics

Next session → Next Day → Next Week → Next Month, etc.

# Intra-session metrics

Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

# Intra-session metrics

Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

# Click-through rates (CTR)

... Interaction

Ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement

Translates to play song/video for music/video sites/formats

- Abandonment rate
- Clickbait
- Site design
- Accidental clicks (mobile)

# No click

# ... Search

www.kdd.org › kdd2020 ▾

## KDD 2020 | Virtual Conference - sigkdd

In two days, we will kick-off the KDD 2020 week! We have worked tirelessly with the organizing committee, SIGKDD executive committee, and volunteers to ...

You've visited this page many times. Last visit: 15/08/20

### Accepted Papers

KDD 2020 Accepted Papers.

Research Track Papers ...

### KDD 2020 Call for Applied ...

Key Dates · Submission: February 13, 2020 · Notification: May 15 ...

### KDD 2020 Call for Research ...

Key Dates · Submission: February 13, 2020 · Notification: May 15 ...

[More results from kdd.org »](#)

### Top stories

#### KDD 2020 Showcases Brightest Minds in Data Science and AI

insideBIGDATA · 1 day ago



## KDD 2020

Event

**Dates:** Sat, 22 Aug 2020 – Thu, 27 Aug 2020

**Location:** San Diego Convention Center, San Diego, California, United States Source

**People also search for**

ECCV 2020	2020 Internati... Confere...	2020 Internati... Joint Co...
--------------	------------------------------------	-------------------------------------

Feedback

# No click

# ... Search

**Table 3. Correlations between click and hover features and relevance judgments for queries with and without clicks.**

Result clicks or no clicks	Feature source	Correlation with human relevance judgments
Clicks (N=1194)	Clickthrough rate (c)	0.42
	Hover rate (h)	0.46
	Unclicked hovers (u)	-0.26
	Max hover time (d)	-0.15
	Combined <sup>1</sup>	<b>0.49</b>
No clicks (N=96)	Hover rate	0.23
	Unclicked hovers	0.06
	Max hover time	0.17
	Combined <sup>2</sup>	<b>0.28</b>

**Click-through rate:**

% of clicks when URL shown (per query)

**Hover rate:**

% hover over URL (per query)



**Unclicked hover:**

Median time user hovers over URL but no click (per query)

**Max hover time:**

Maximum time user hovers over a result (per SERP)

# No click

... Search

**Abandonment** is when there is no click on the search result page

User is dissatisfied (bad abandonment)

User found result(s) on the search result page (good abandonment)



858 queries (21% good vs. 79% bad abandonment manually examined)

## Cursor trail length

Total distance (pixel) traveled by cursor on SERP

Shorter for good abandonment

## Movement time

Total time (second) cursor moved on SERP

Longer when answers in snippet (good abandonment)

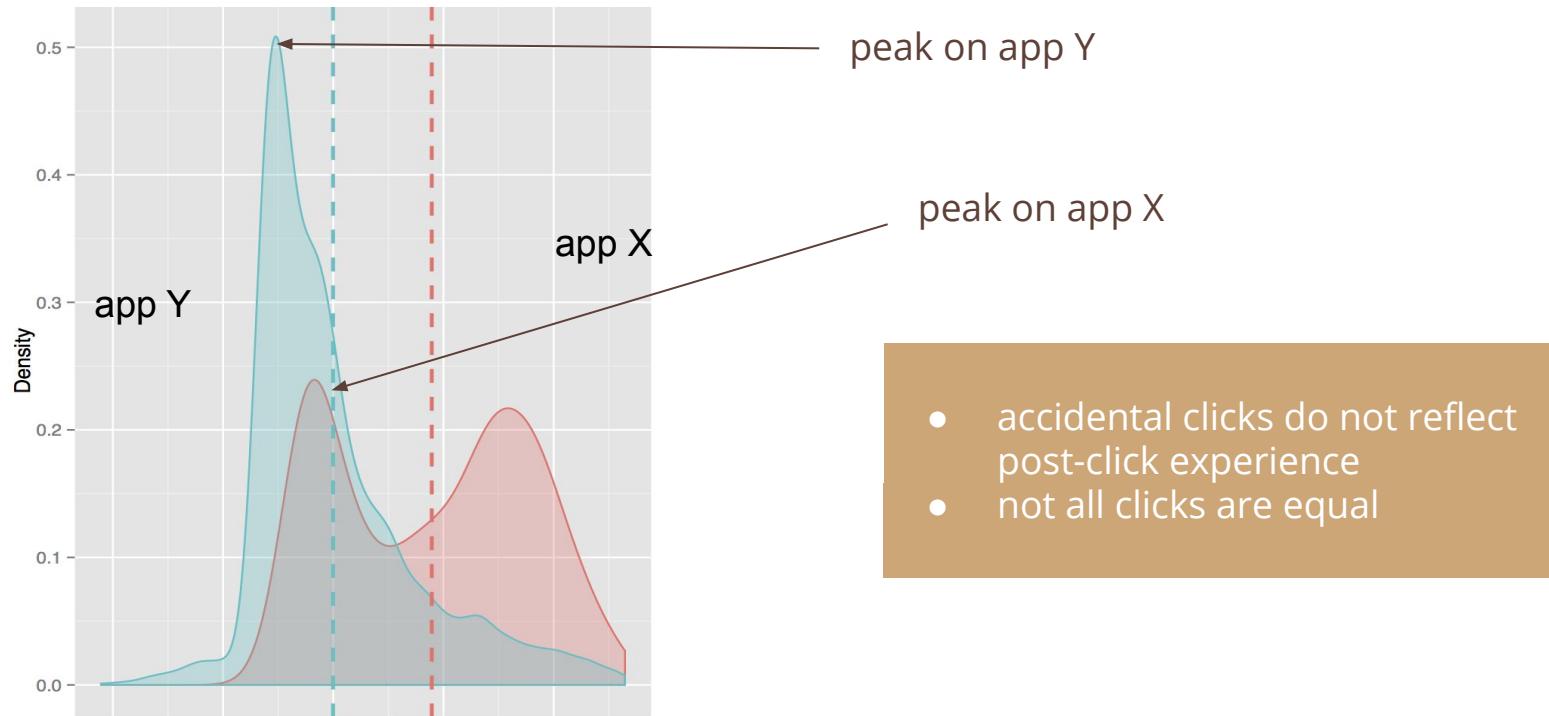
## Cursor speed

Average cursor speed (pixel/second)

Slower when answers in snippet (good abandonment)

# The quality of a click on mobile apps ... advertising

dwell time distribution of apps X and Y for given ad



# Click-through rate

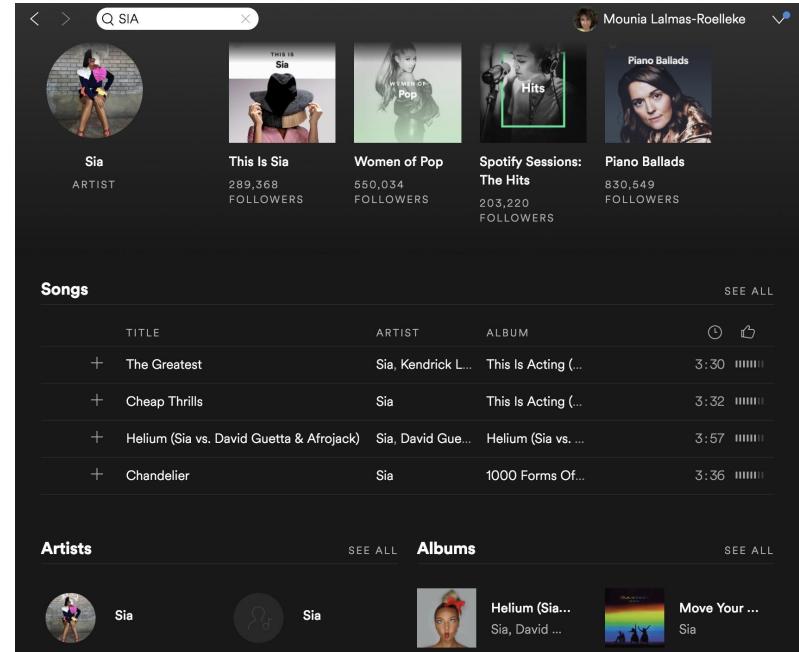
... Music

Ratio of users who click on a specific item to the number of total users who “view” that **item**

What is an item?

- Track
- Artist page
- Album
- Playlist
- ...

The value of a click  
→ downstream engagement



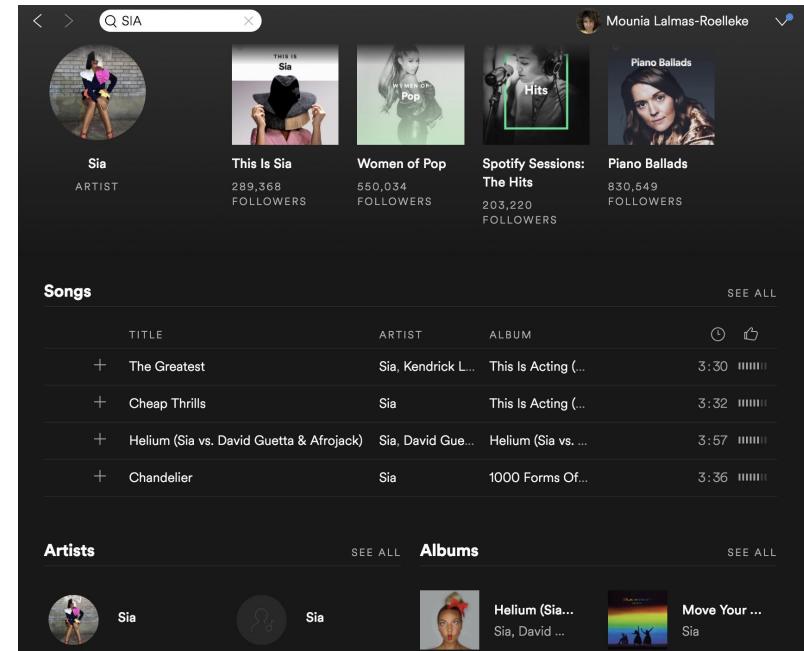
# Downstream engagement

... music

What the user does from a particular click at “place X” → downstream behaviour:

- Total number of tracks played/saved from artist contained within X
- Number of visits to album pages/artist pages contained within X
- Total time spent on album pages/artist pages contained within X
- Total number of playlists updated/created with tracks contained within X
- ...

→ building relationships



# Intra-session metrics

Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

# Dwell time

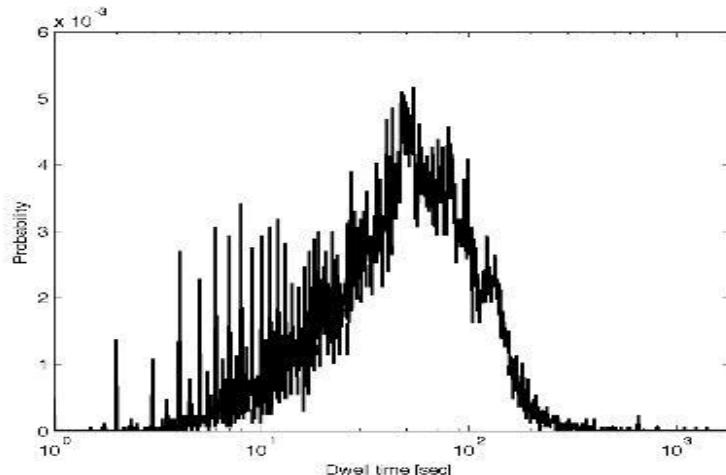
... Involvement

The contiguous time spent on a site or web (mobile) page

Good indication (proxy) of interests

Similar measure is play/streaming time for video and audio streaming services

- Not clear what user is actually looking at while on page/site
- Instrumentation issue with last page viewed and open tabs



distribution of dwell times on 50 websites

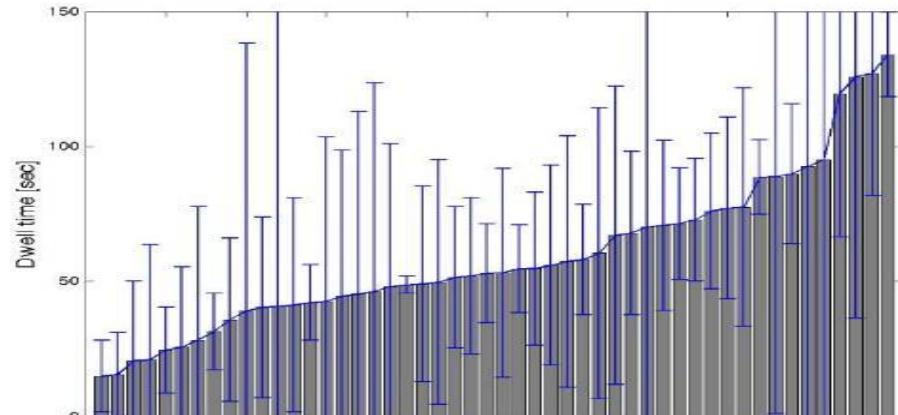
# Dwell time

... Involvement

## Dwell time varies by site type:

e.g. leisure sites tend to have longer dwell times than news

Dwell time has a relatively large variance even for the same site, maybe reflecting interests

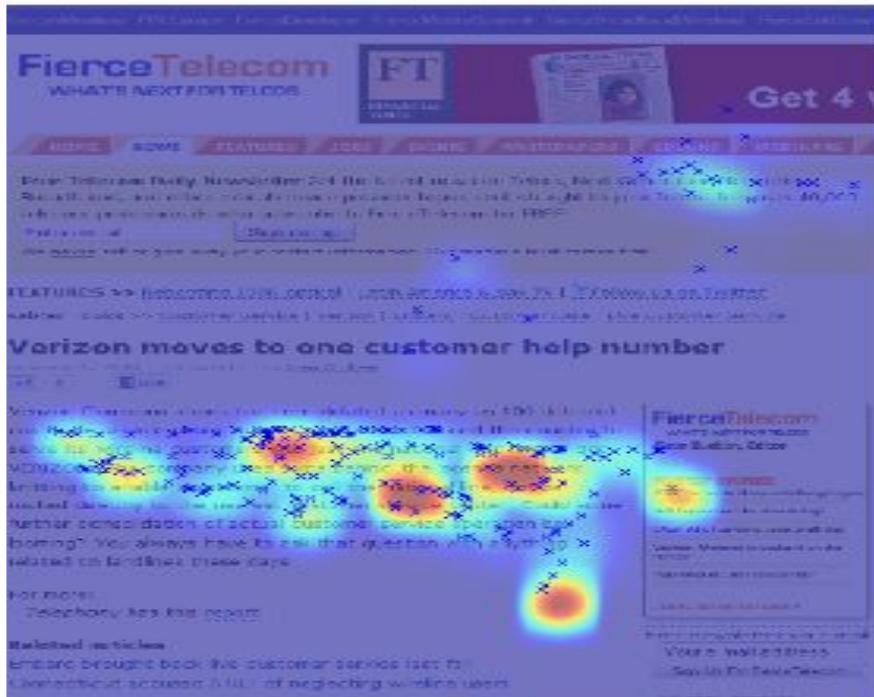


average and variance of dwell time of 50 sites

[1] Mounia Lalmas, Heather O'Brien and Elad Yom-Tov. **Measuring user engagement**. Morgan & Claypool Publishers, 2014.

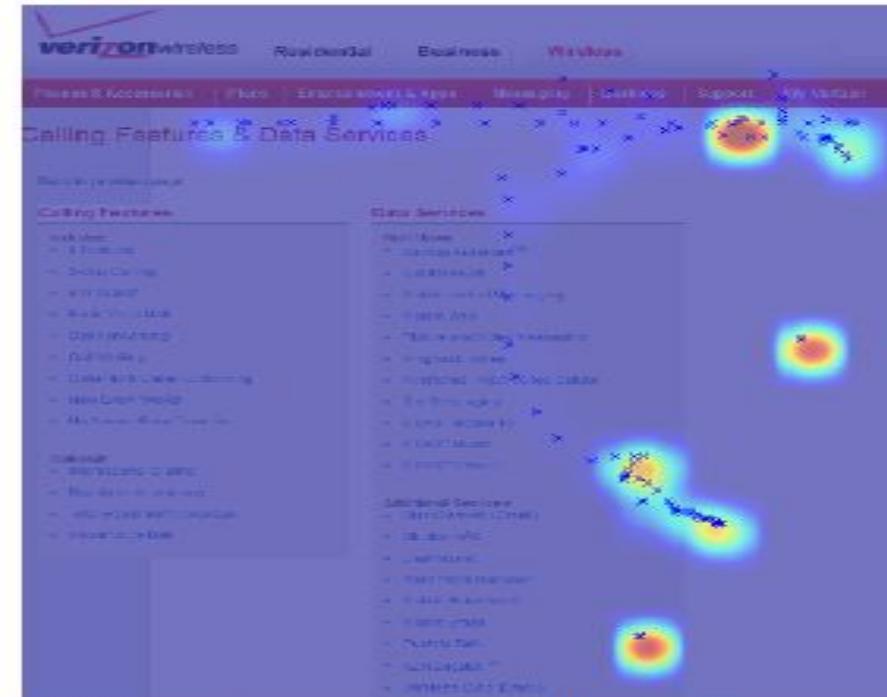
[2] Elad Yom-Tov, Mounia Lalmas, Ricardo Baeza-Yates, Georges Dupret, Janette Lehmann and Pinar Donmez. **Measuring Inter-Site Engagement**. BigData 2013.

# Dwell time ... Search



(a) relevant (dwell time: 30s)

"reading" cursor heatmap of relevant document vs "scanning" cursor heatmap of non-relevant document

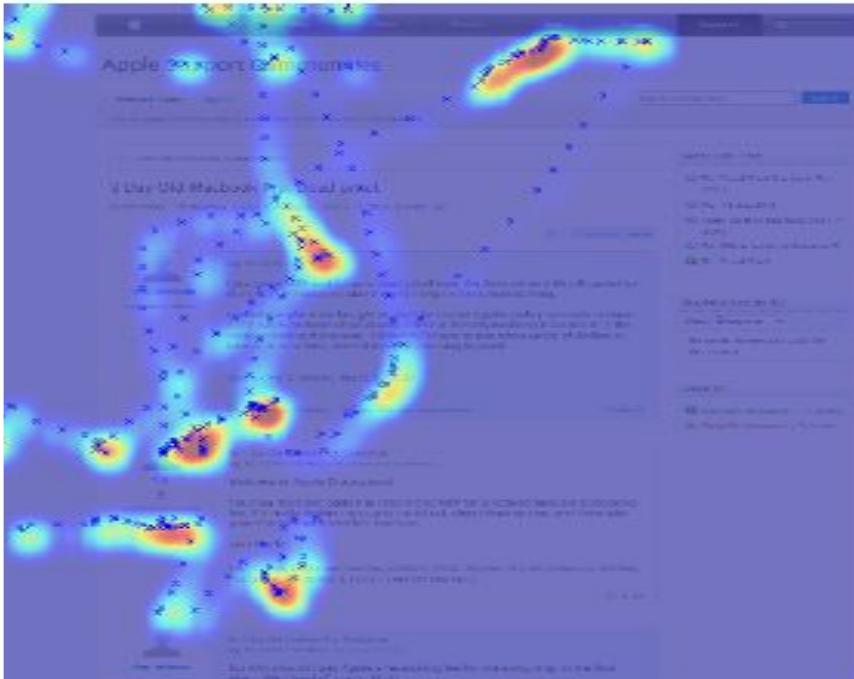


(b) non-relevant (dwell time: 30s)

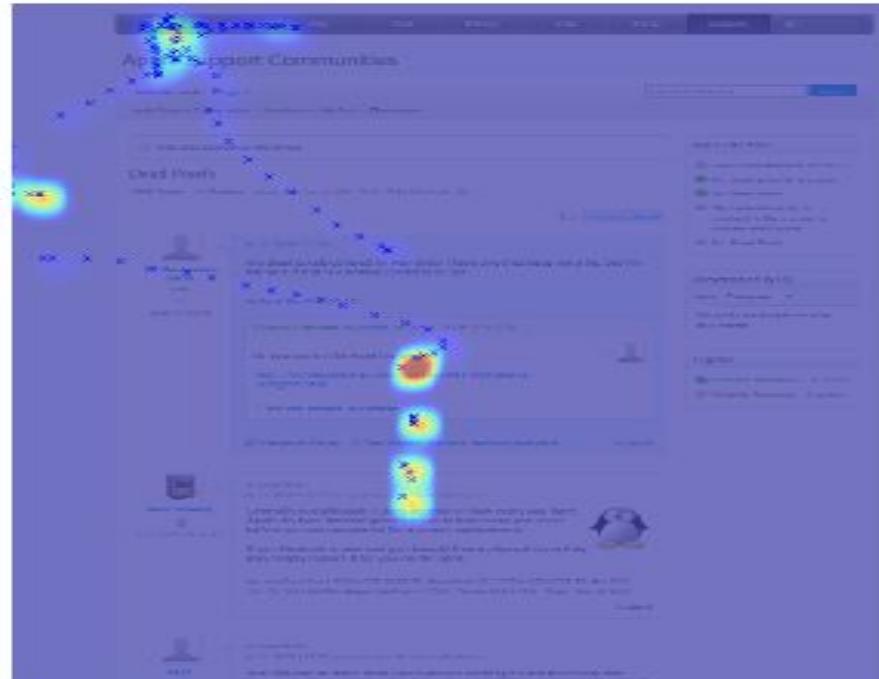
Qi Guo and Eugene Agichtein. **Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior.** WWW 2012.

# Dwell time

... Search



(a) relevant (dwell time: 70s)



(b) non-relevant (dwell time: 80s)

"reading" a relevant long document vs "scanning" a long non-relevant document

# Dwell time

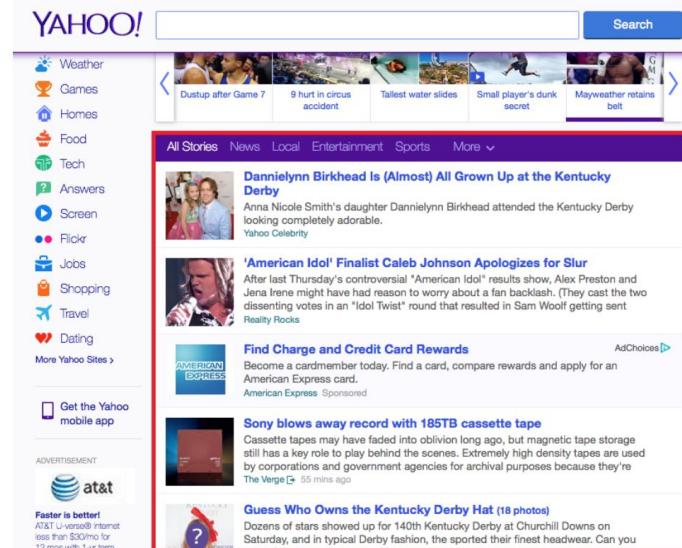
... news

Dwell time better proxy for user interest on news article in the context of personalization

Optimizing for dwell time led to increase in click-through rates

A way to reduce optimizing for click-baits

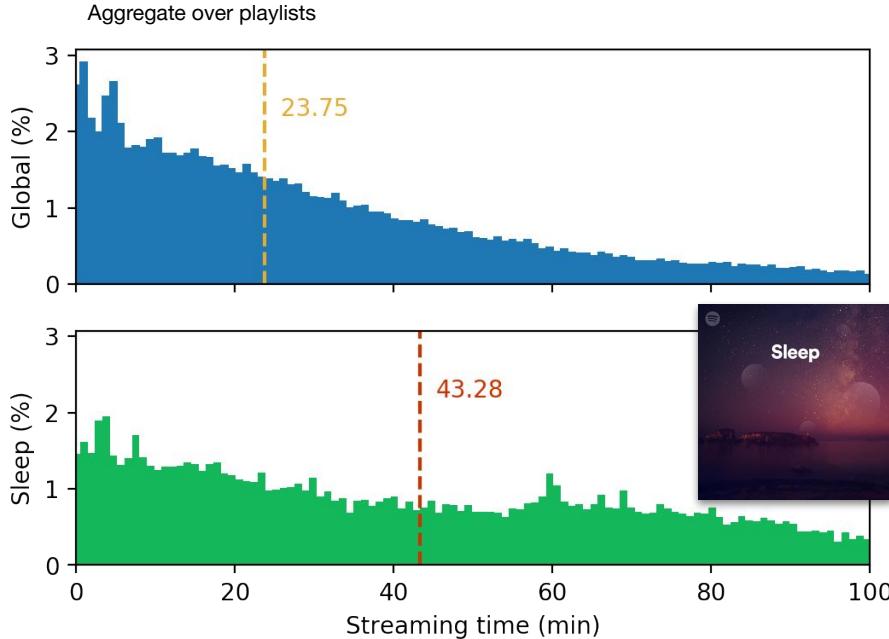
See section on Offline experiment and evaluation



**Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.**

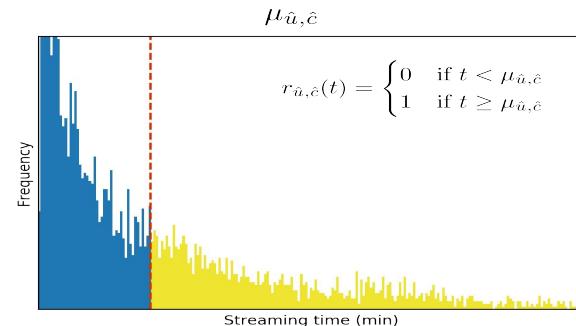
# Dwell time as streaming time

... music



Consumption time of sleep playlist longer than average playlist consumption time.

Optimizing for mean consumption time led to +22.24% in predicted stream rate compared to stream rate (equivalent to click-through rate) on Spotify Home



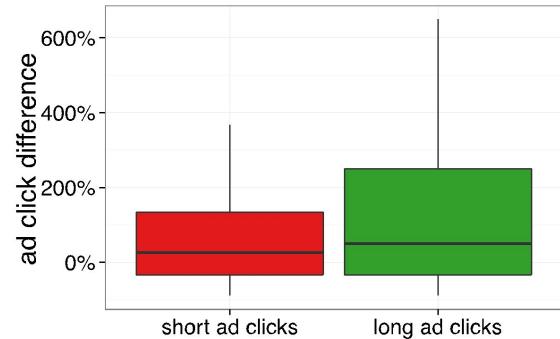
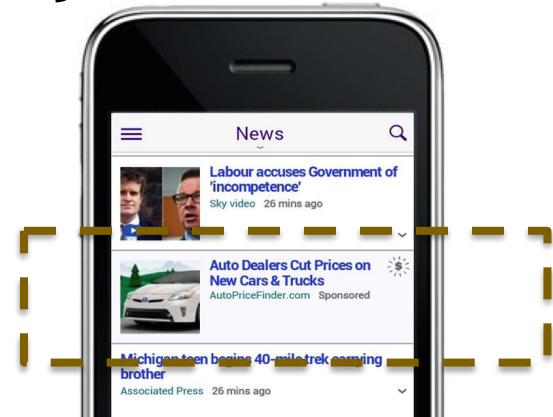
# Dwell time and ad landing page quality

## User click on an ad → ad landing page

Dwell time is time until user returns to publisher and used as proxy of quality of landing page

## Dwell time → ad click

Positive post-click experience (“long” clicks) has an effect on users clicking on ads again (mobile)



# Intra-session metrics

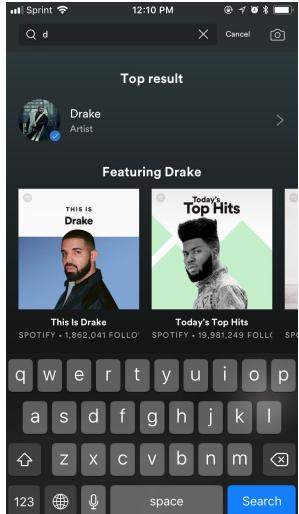
Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

# User journey in search

... Music

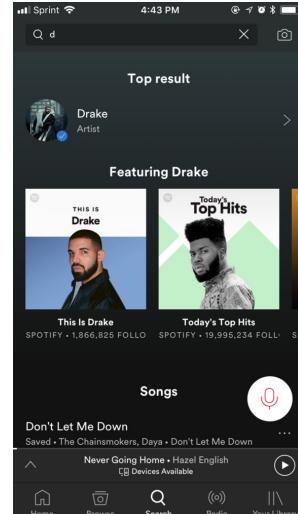
## TYPE/TALK

User communicates with us



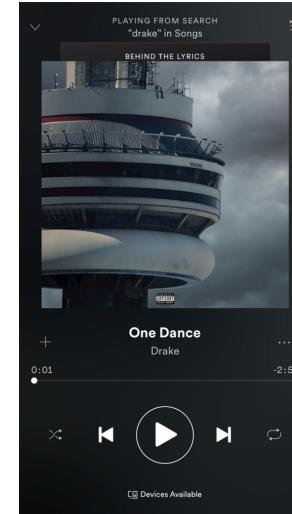
## CONSIDER

User evaluates what we show them



## DECIDE

User ends the search session



EFFORT

SUCCESS

Users evaluate their experience on search based on two main factors: **success** and **effort**

# Organize metrics

... Interaction

## "Success" metrics

### DECIDE

#### LISTEN

Have a listening session  
stream

#### ORGANIZE

Curate for future listening

add to a playlist,  
save into a collection,  
follow an artist,  
follow a playlist, ...

## "Effort" metrics

### TYPE

number of  
deletions, ...

### CONSIDER

back button  
clicks, first and  
last click  
position, ...

### Time to success

In A/B testing, success rate more sensitive than click-through rate.

[1] Praveen Ravichandran, Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas and Jenn Thom. **Developing Evaluation Metrics for Instant Search Using Mixed Methods.** SIGIR 2019.

[2] Ang Li, Jennifer Thom, Praveen Chandar, Christine Hosey, Brian St. Thomas and Jean Garcia-Gathright. **Search Mindsets: Understanding Focused and Non-Focused Information Seeking in Music Search.** WWW 2019.

# Intra-session metrics

Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

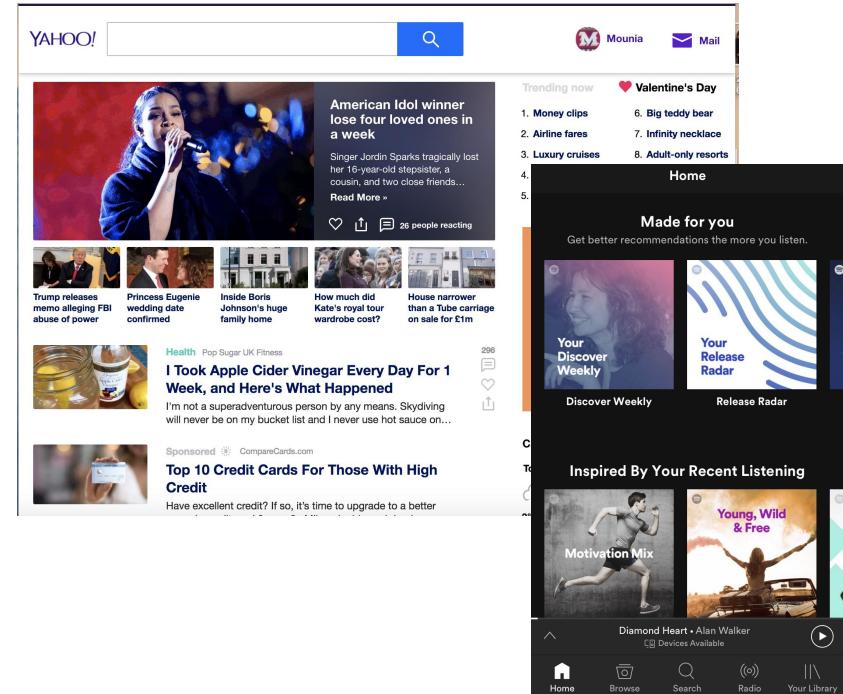
# Revisit rates ... Involvement

Number of returns to the site **within** a session or successive sessions (task) → definition of a session and a task?

Common in sites that are homepages, or contain content of regular interest to users

Useful for sites such as news aggregators, where returns indicate that user believes there may be more information to glean from the site

Where **recommender systems** must do well

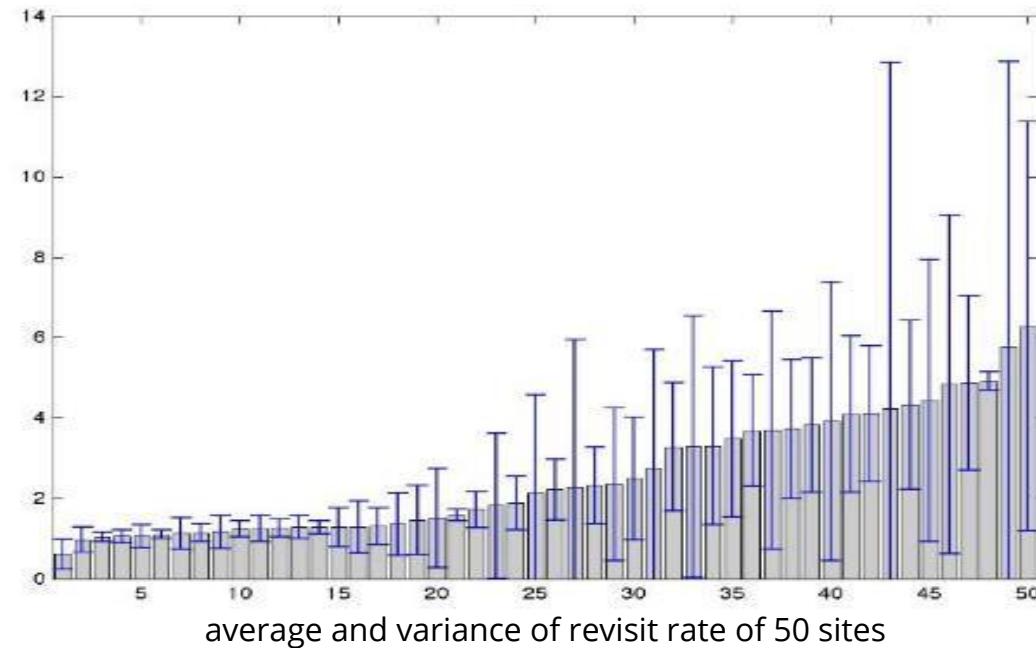


# Revisit rates

# ... Involvement

Goal-oriented sites (e.g., e-commerce) have lower revisits in a given time range observed → revisit horizon should be adjusted by site

What is a session and how does it relate to the task at hand?



# Revisit rate ... Session length

2.5M users, 785M page views, 1 month sample

Categorization of the most frequently accessed sites

11 categories (e.g. news), 33 subcategories

(e.g. news finance, news society)

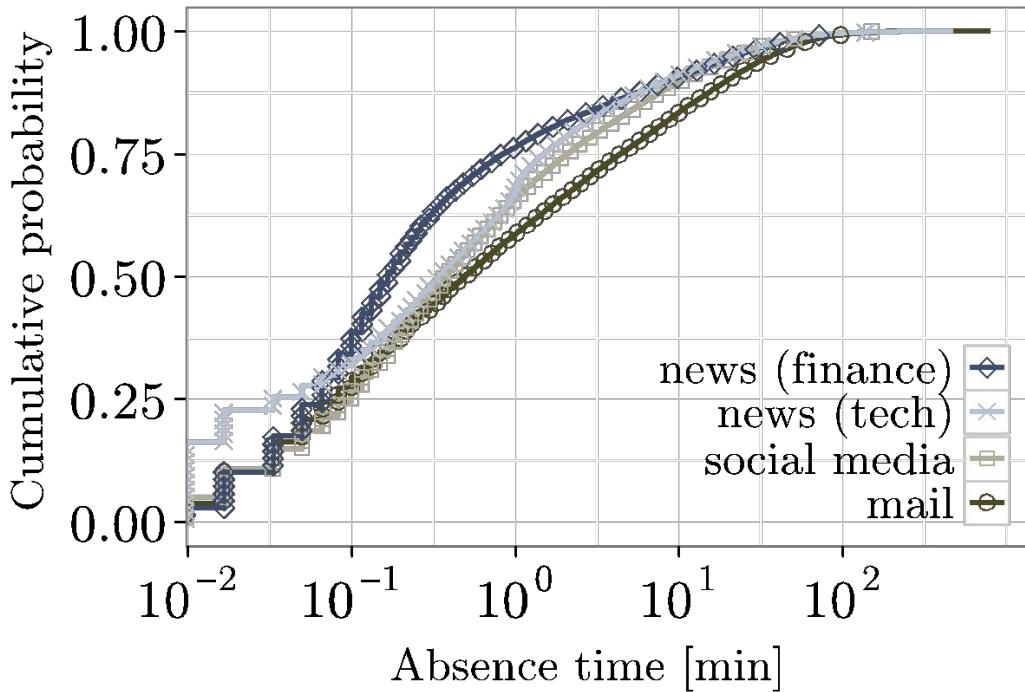
60 sites from 70 countries/regions

Cat.	Subcat.	%Sites	Description
news 22.1%	news	5.79%	
	news (soc.)	5.13%	society
	news (sport)	2.63%	
	news (enter.)	2.24%	music, movies, tv, etc.
	news (finance)	1.97%	
	news (life)	1.58%	health, housing, etc.
	news (tech)	1.58%	technology
	news (weather)	1.18%	
search 15.3%	search	12.63%	
	search (special)	1.58%	
	directory	1.05%	search for lyrics, jobs, etc.
service 11.6%	service	7.63%	translators, banks, etc.
	maps	3.03%	
	organization	0.92%	bookmarks, calendar, etc.
blogging 9.6%	blogging	3.55%	
	knowledge	3.55%	collaborative creation and collection of content
	sharing	2.50%	sharing of videos, files, etc.
navi 9.3%	navi	6.58%	
	front page	1.84%	personalized front pages
	front page (pers.)	0.92%	
support 8.7%	support	1.58%	sites that provide products and support for them
	download	7.11%	downloading software
shopping 7.9%	shopping	4.34%	
	auctions	2.11%	
	comparison	1.45%	sites to compare prices of products
leisure 5.7%	adult	2.76%	
	games	1.97%	
	entertainment	0.92%	sites with music, tv, etc.
mail 3.9%	mail	3.95%	
social media 3.0%	social media	1.97%	
	dating	1.05%	
settings 2.9%	login	1.71%	
	settings	1.18%	profile setting, site personalization

short session: average 3.01 distinct sites visited with revisit rate 10%  
long session: average 9.62 distinct sites visited with revisit rate 22%

# Time between each revisit

... online multi-tasking



50% of sites are revisited after less than 1 minute

This is likely more about online multi-tasking to perform a big task

Multi-task optimization?

# Intra-session metrics

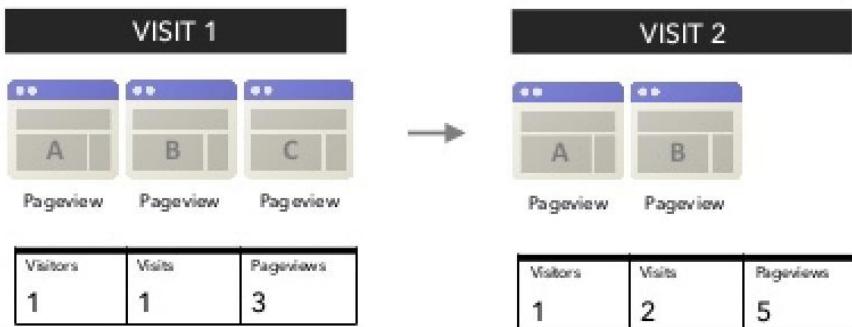
Click-through rate  
Dwell time  
“Organise” metrics  
Revisit rate  
Page view  
Conversion rate  
Social media metrics

# Pageview ... Involvement

Page view is request to load a single page

Number of pages viewed (**click depth**): average number of contiguous pages viewed during a visit → “user journey” across the application/site

Reload after reaching page → counted as additional pageview (e.g. news stream)  
If same page viewed more than once → a single unique pageview



Can be problematic with ill-designed site as high click depth may reflect users getting lost and user frustration

Site may deliberately “design” for high click depth

# Conversion rate ... Interaction

Fraction of sessions which end in a desired user action

particularly relevant to e-commerce (making a purchase) ... but also include subscribing, booking a room, free to premium conversion

Online advertising using conversion as cost model to charge advertisers

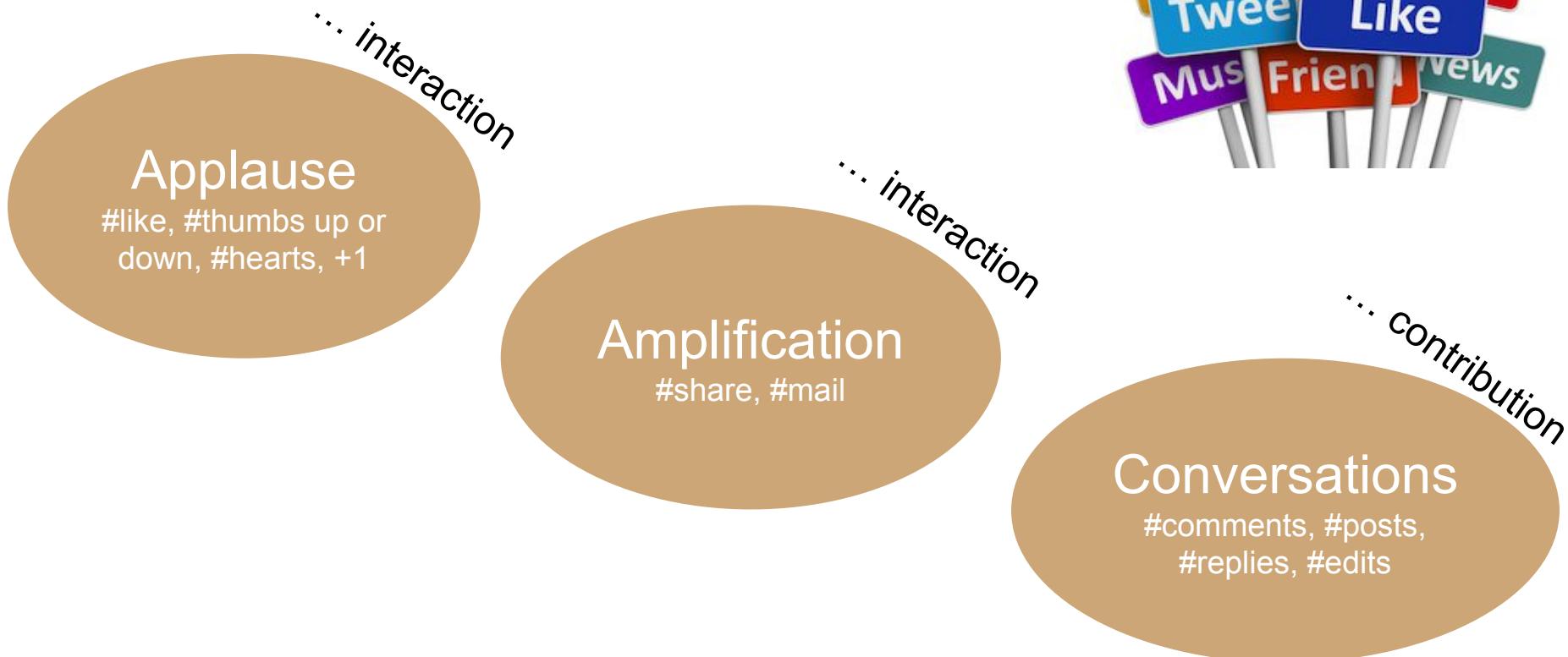
Not all sessions are expected to result in a conversion, so this measure not always informative

dwell time often used as proxy of satisfactory experience as may reflect affinity with the brand

Reference:

[1] Mihajlo Grbovic and Haibin Cheng. **Real-time Personalization using Embeddings for Search Ranking at Airbnb.** KDD 2018.

# Social media metrics



# Intra-session metrics

Some final words

What comes next

# Some final words on intra-session metrics

Metrics for smaller granularity levels such as viewport or specific section → attention

Metrics for scroll → important for stream and mobile

Whether an intra-session metric belongs to Involvement, Interaction, or Contribution depend on the expected type of engagement of the site



[1] Dmitry Lagun and Mounia Lalmas. **Understanding and Measuring User Engagement and Attention in Online News Reading.** WSDM 2016.

[2] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang and Qiaozhu Mei. **Beyond Ranking: Optimizing Whole-Page Presentation.** WSDM 2016.

[3] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster and Vidhya Navalpakkam. **Towards better measurement of attention and satisfaction in mobile search.** SIGIR 2014.

# Non intra-session metrics

**Inter-session metrics → Habit → Loyalty**

**Inter-session metrics → Loyalty**

How many users and how fast they return to the site

---

**Total use measurements → Popularity**

Total usage time

Total number of sessions

Total view time (video)

Total number of likes (social networks)

**Direct value measurement → Lifetime value**

Lifetime value, as measured by ads clicked, monetization,

Relate to return of investment (acquisition cost, value proposition)

# Inter-session metrics

Why inter-session metrics

Relationship to loyalty

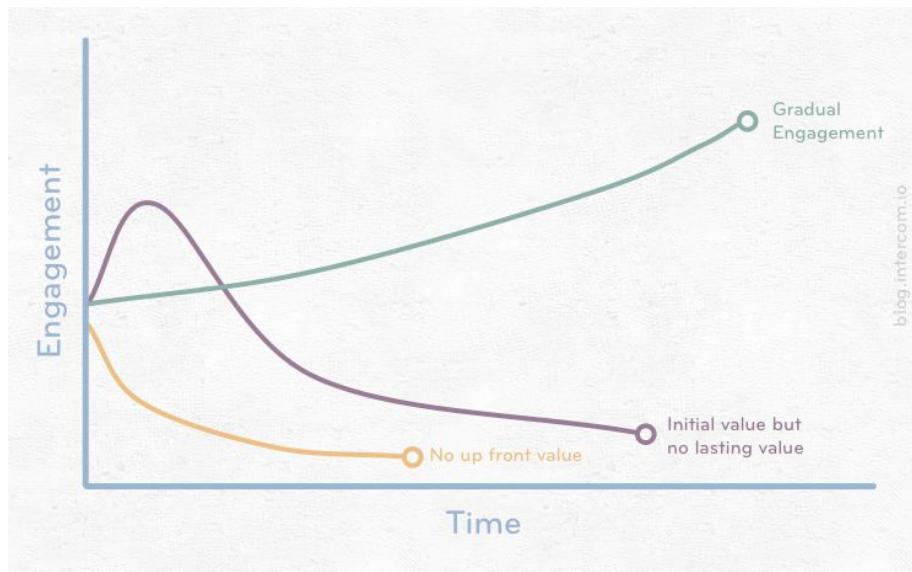
Absence time

# Why inter-session metrics?

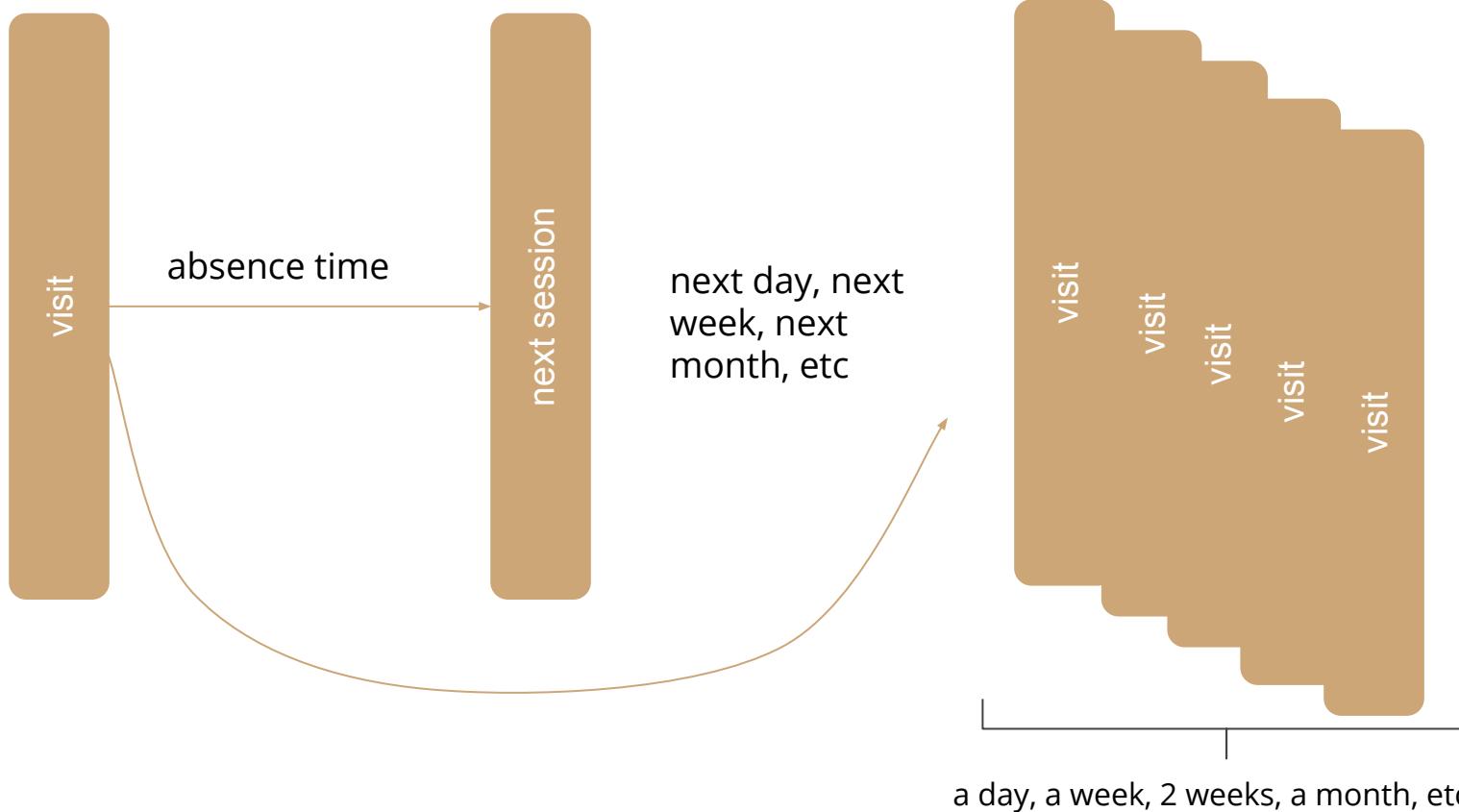
Intra-session measures can easily mislead, especially for a short time

Consider a very poor ranking function introduced into a search engine by mistake

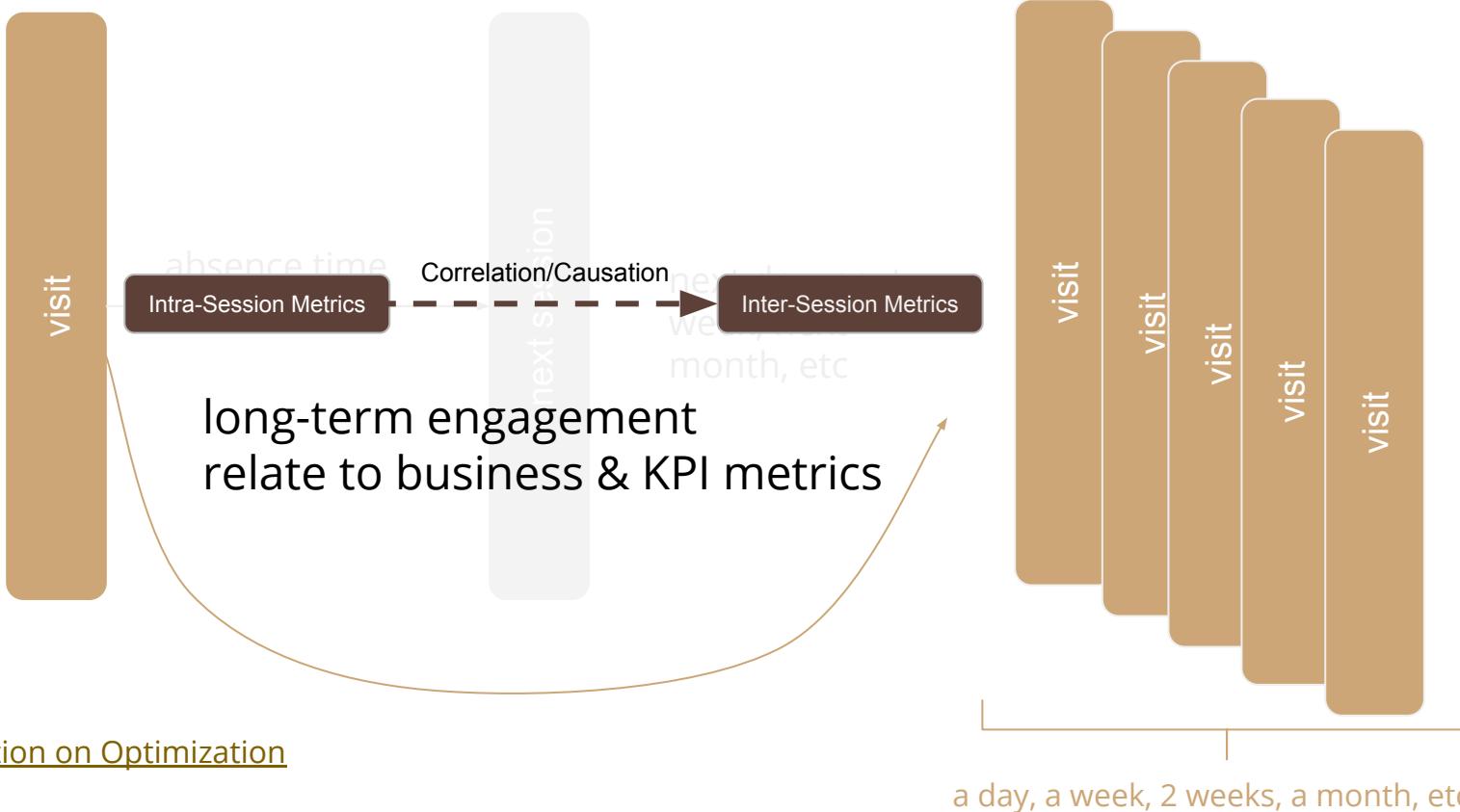
Therefore, A/B testing may provide erroneous results if only intra-session measures are used



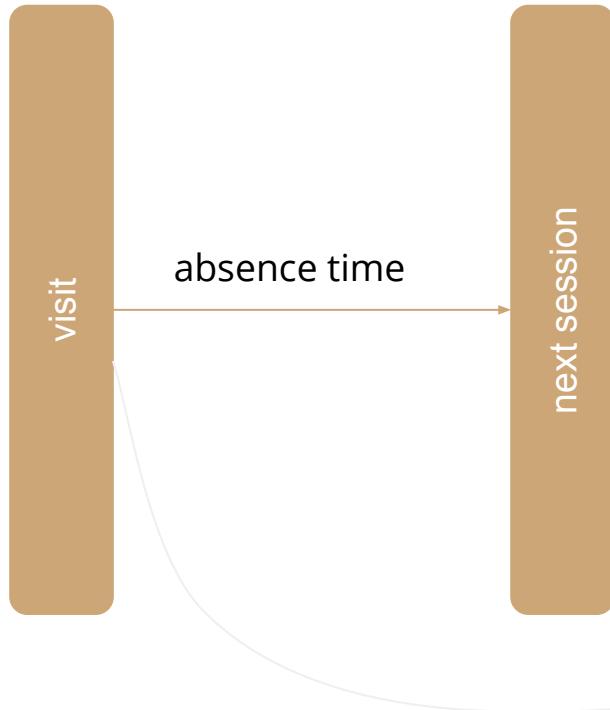
# Inter-session metrics



# Inter-session metrics



# Inter-session metrics



really mostly about endurability

next day, next  
week, next  
month, etc

habit  
periodicity  
short task/visit

absence time ≠ revisit rate  
absence time ≠ task

**Cases studies:** search and news

a day, a week, 2 weeks, a month, etc

Total number of visits or sessions  
Total number of unique visitors  
Total number of sessions per user  
Total amount of time spent ...

# Absence time applied to search

... Study I

## Ranking functions on Yahoo Answer Japan

The screenshot shows the Yahoo! Japan search interface. The search bar at the top contains the query "best sushi". Below the search bar are buttons for "Q&A" and "検索" (Search). A link "条件を指定して検索" (Specify conditions and search) is also present. The main search results are displayed under the heading "best sushi で検索した結果1~10件目 / 約59件". The results include:

- Y!社員の皆様:** 検索結果に対するフィードバックをお願いします。 | 登録方法
- スポンサードサーチ**
- すし ランキング**
- 寿司 ランキング**
- 日本橋で寿司なら《鰯門》へ**
- 【業務用】ベスト通販 納期2日**

Below these are several links related to sushi experiences and reviews from users, such as "What's your best sushi experience?", "The best sushi restaurants in Tokyo", and "My favorite sushi". Each result includes a brief description, the date it was posted, the number of answers and views, and categories like "Food & Drink" and "Japanese".

Two-weeks click data  
on Yahoo Answer  
Japan search

One millions users  
Six ranking functions

Session boundary:  
30 minutes of  
inactivity

# Examples of metrics for search

(Proxy: relevance of a search result)

Number of clicks

SAT click

Quick-back click

Click at given position

Time to first click

Skipping

Abandonment rate

Number of query reformulations

Dwell time (result vs result page)

Saffron Walden

All Maps News Images Videos More Settings Tools

About 10,400,000 results (0.67 seconds)

[www.visit'affronwalden.gov.uk](http://www.visit'affronwalden.gov.uk) Visit Saffron Walden - Saffron Walden Tourist Information Centre

Saffron Walden is a delightful market town located in north-west Essex. It has a rich heritage of old buildings and the market has been here since 1141.

Saffron Walden Market · Explore Saffron Walden · Saffron Walden Museum · History

People also ask

- What is Saffron Walden famous for?
- Is Saffron Walden a nice place to live?
- Does saffron come from Saffron Walden?
- What day is Saffron Walden Market?

Feedback

en.wikipedia.org › wiki › Saffron\_Walden

**Saffron Walden - Wikipedia**

Saffron Walden is a market town in the Uttlesford district of Essex, England, 12 miles (19 km) north of Bishop's Stortford, 15 miles (24 km) south of Cambridge and 43 miles north of London. It retains a rural appearance and some buildings of the medieval period. [Wikipedia](#)

Region: East Country: England

District: Uttlesford Fire: Essex

History · Sites and buildings of ... · Transport · Culture

Top sights in Saffron Walden

Audley End House and Gardens 4.6 ★★★★★ (2,805)  
17th-century country house & grounds

Audley End Miniature Railway and Enchant... 4.6 ★★★★★ (892)  
Train journeys through a wooded estate

Bridge End Gardens 4.7 ★★★★★ (291)  
Garden and park and garden

Audley End House and Gardens

Audley End Miniature Railway and Enchanted Wood

Bridge End Gardens

**COVID-19 alert**

Event information may be out of date due to the coronavirus (COVID-19). Confirm details with event organisers.  
[More about COVID-19](#)

Sun, 23 Aug 08:00 Hinxton, Holy Communion (BCP) St Mary and St John Church, Hinxton

Sun, 23 Aug 10:00 Duxford, All-age Service Duxford Saint Peter

Sun, 23 Aug 13:00 Picnic with ... (live) Music from Broadway Haverhill Arts Centre

The Guardian

Audley End House and Gardens Saffron Walden

Saffron Walden Town in England

Saffron Walden is a market town in the Uttlesford district of Essex, England, 12 miles north of Bishop's Stortford, 15 miles south of Cambridge and 43 miles north of London. It retains a rural appearance and some buildings of the medieval period. [Wikipedia](#)

Weather: 19 °C, Wind W at 12 mph (19 km/h), 63% Humidity

Hotels: 3-star averaging £82. [View hotels](#)

Shire county: [Essex](#)

Population: 15,504 (2011)

Postcode district: CB10, CB11

Sovereign state: United Kingdom

Local time: Sunday 12:18

Events

COVID-19 alert

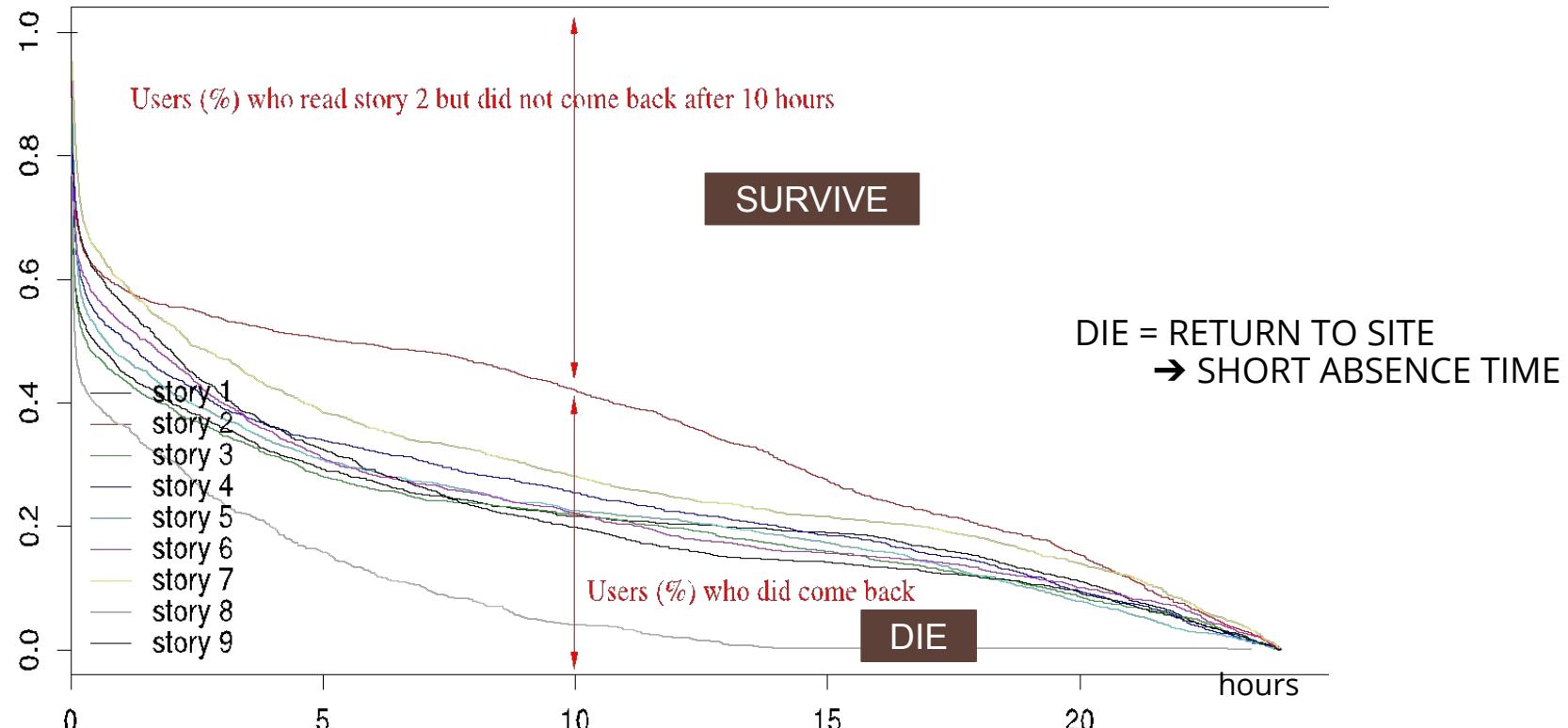
Event information may be out of date due to the coronavirus (COVID-19). Confirm details with event organisers.  
[More about COVID-19](#)

Sun, 23 Aug 08:00 Hinxton, Holy Communion (BCP) St Mary and St John Church, Hinxton

Sun, 23 Aug 10:00 Duxford, All-age Service Duxford Saint Peter

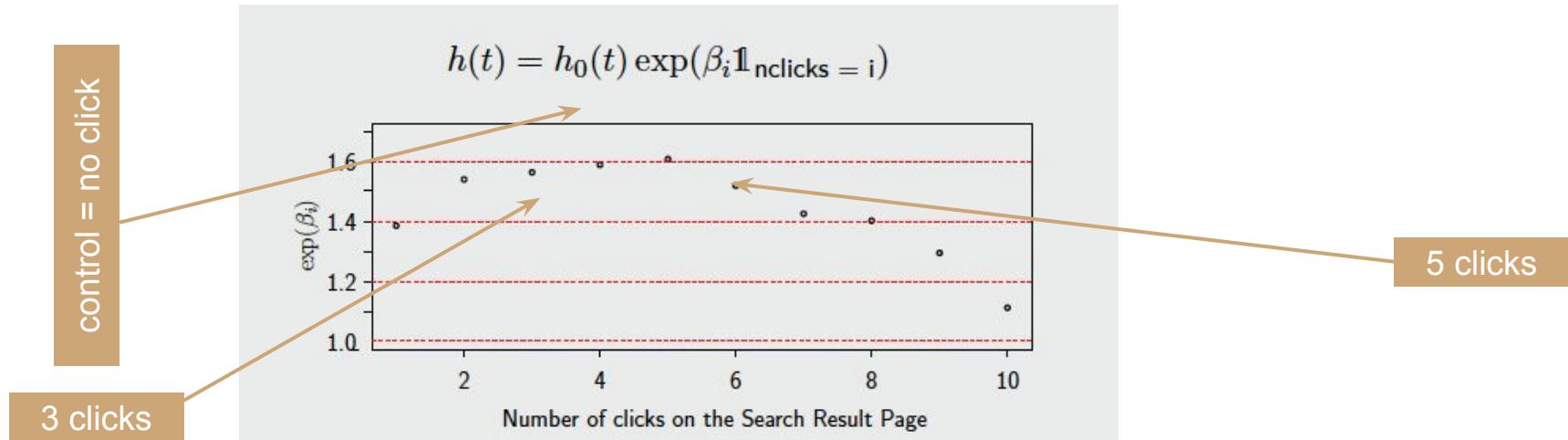
Sun, 23 Aug 13:00 Picnic with ... (live) Music from Broadway Haverhill Arts Centre

# Absence time and survival analysis



# Absence time and number of clicks

survival analysis: high hazard rate (die quickly) = short absence



No click means a bad user search session ... in Yahoo Japan search

Clicking between 3-5 results leads to same user search experience

Clicking on more than 5 results reflects poor user search session; users cannot find what they are looking for

# Absence time and search session

... What else?

intra-session search metrics → absence time



- Clicking lower in the ranking (2<sup>nd</sup>, 3<sup>rd</sup>) suggests more careful choice from the user (compared to 1<sup>st</sup>) → shorter absence time
- Clicking at bottom is a sign of low quality overall ranking → longer absence time
- Users finding their answers quickly (time to 1<sup>st</sup> click) return sooner to the search application → shorter absence time
- Returning to the same search result page is a worse user experience than reformulating the query → longer absence time

# DCG versus absence time to evaluate five ranking functions



## DCG@1

Ranking Alg 1  
Ranking Alg 2  
Ranking Alg 3  
Ranking Alg 4  
Ranking Alg 5

## DCG@5

Ranking Alg 1  
Ranking Alg 3  
Ranking Alg 2  
Ranking Alg 4  
Ranking Alg 5

## Absence time

Ranking Alg 1  
Ranking Alg 2  
Ranking Alg 5  
Ranking Alg 3  
Ranking Alg 4

# Absence time and search experience

... Study II

intra-session search metrics → absence time



From 21 experiments carried out through A/B testing, using absence time agrees with 14 of them (66% which one is better)

## Positive

- One more query in session
- One more click in session
- SAT clicks
- Query reformulation

## Negative

- Abandoned session
- Quick-back clicks

# Absence time and search experience ... Studies I & II

intra-session search metrics → absence time ← proxy of durability

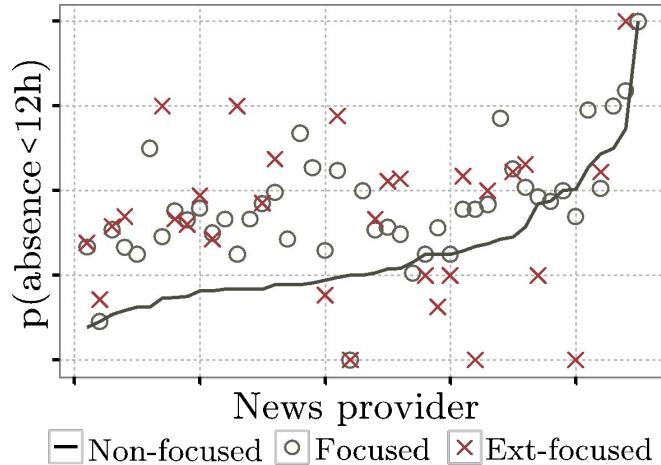
Demonstrated that absence time is an appropriate inter-session metric for search because of the correlation & predictive power of known indicators of a positive search experience

→ absence time as a metric to compare A/B test in search

These known indicators could act as intra-session metrics, which could be optimised by the ranking algorithms

They can also be used as features in the ranking algorithms themselves

# Absence time & focused news reading



For 70% of news sites that provide links to off-site content, probability that users return within 12 hours increases by 76%

## Ukraine crisis: 'Dozens killed' in east as Minsk talks held



Ukrainian troops are trying to defend the key transport hub of Debaltseve

At least 40 people have been reported killed as fighting between Ukrainian troops and pro-Russian rebels rages on in the east of the country.

Ukrainian officials say 15 soldiers and 12 civilians died in the past 24 hours. The rebels report 13 casualties.

The separatists also claim to have seized the town of Vuhlehirsk and surrounded the key hub of Debaltseve, but the Ukrainian military denies this.

Meanwhile, urgent truce talks ended in Belarus, but no deal was signed.

Representatives of Ukraine and Russia, as well as rebel envoys and members of the Organization for Security and Co-operation (OSCE), took

### Around the Web

[H Peace in Ukraine depends on America](#)

[Ukraine Crisis Map](#)

[Explosion in Ukraine](#)

[Casualties of the Ukrainian crisis](#)

[You Exclusive interview with President Putin](#)

[100](#)

[Casualties of the Ukrainian crisis](#)

[Exclusive interview with President Putin](#)

[100](#)

[Casualties of the Ukrainian crisis](#)

[Exclusive interview with President Putin](#)

[100](#)

[Casualties of the Ukrainian crisis](#)

[Exclusive interview with President Putin](#)

[100](#)

Related off-site content

# Other metrics

- Popularity
- Long-term value (LTV)

# Popularity metrics

## With respect to users

- MAU (monthly active users), WAU (weekly active users), DAU (daily active users)
- Stickiness (DAU/MAU) measures how much users are engaging with the product
- Segmentation used to dive into demographics, platform, recency, ...

## With respect to usage

- Absolute value metrics (measures) → aggregates over visits/sessions  
total number of clicks; total number of sessions; total number of time spent per day, month, year
- Usually correlate with number of active users

# Long-term value (LTV) metrics

How valuable different users are based on lifetime performance → value that a user is expected to generate over a given period time, e.g. such as 12 months

- Services relying on advertising for revenue:
  - based on a combination of forecasted average pageviews per user, actual retention & revenue per pageview
- Services relying on actual purchases (e-commerce):
  - based on total amount of purchases
- Services relying in content being consumed (entertainment)
  - account for cost on producing exclusive content and/or licensing content

e.g. CAC = customer acquisition cost

**LTV > CAC** =   
**CAC > LTV** = 

Help analyzing strategy (acquisition, content, etc) and estimate further strategy costs

# Recap

Online engagement & metrics

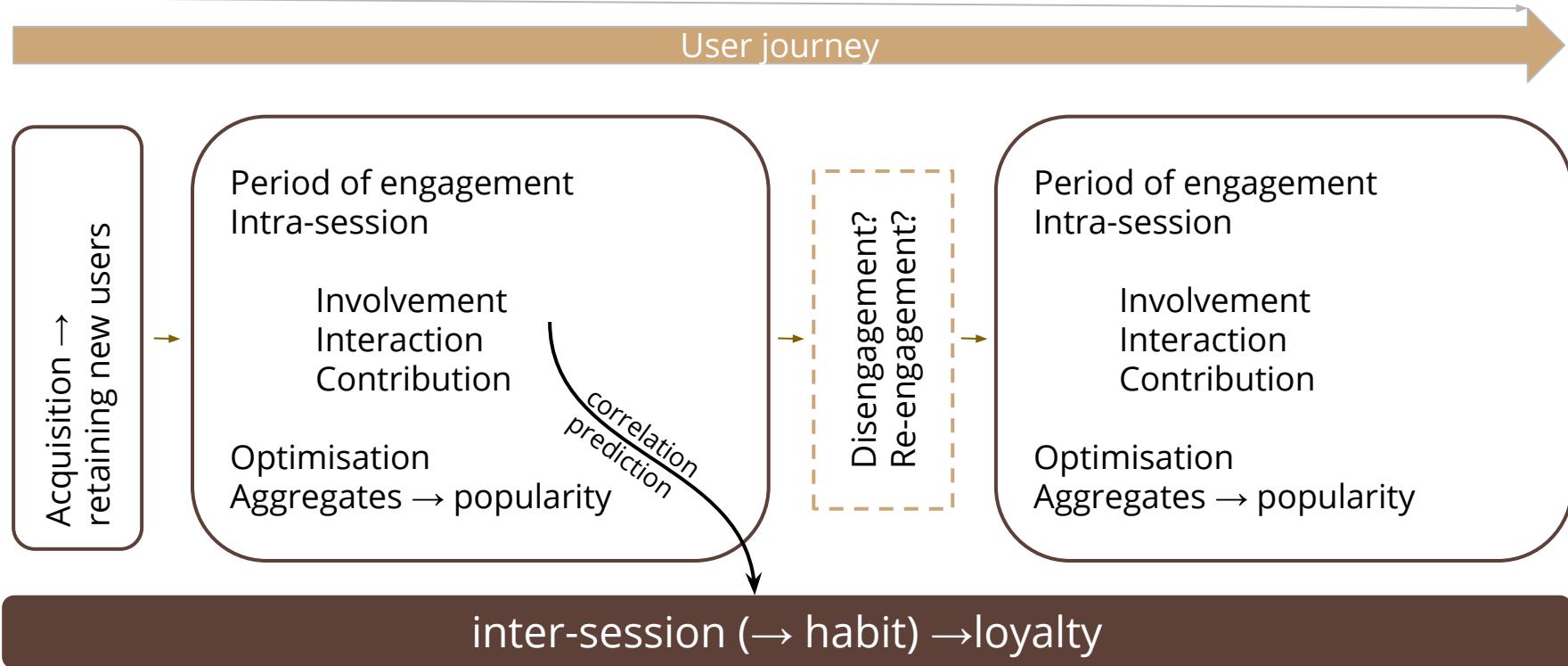
How it all fits together

# Online engagement & metrics

... recap

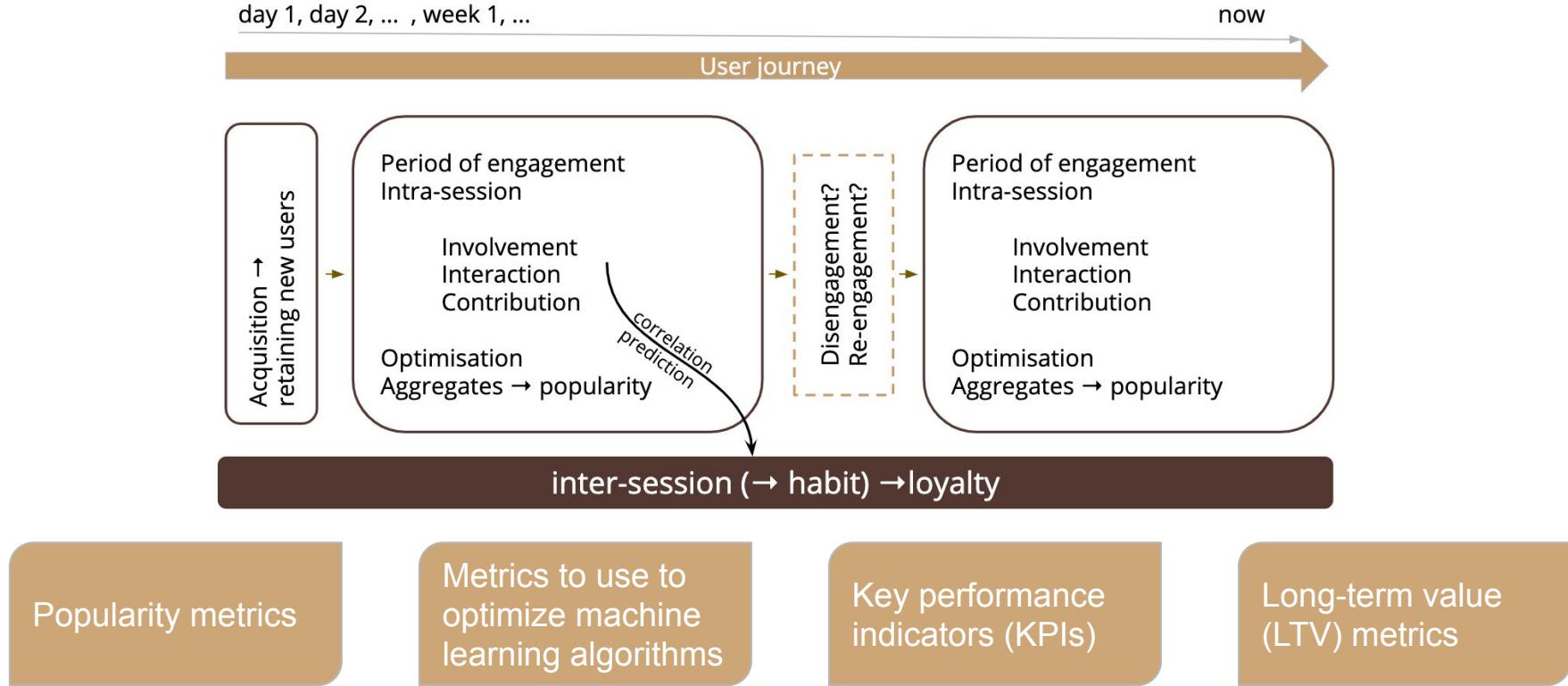
day 1, day 2, ... , week 1, ...

now



# Online engagement & metrics

## ... all together





# Optimization

# Optimization

Manual & Semi-Manual Optimization

Automatic Optimization

Combining Two Camps

# Two Camps of Optimizations

- **Manual and Semi-Manual Optimization**
  - e.g. The classic Hypothesis-Experiment-Evaluation Cycle
- **Automatic Optimization**
  - e.g., Online Learning, Multi-armed Bandits, Reinforcement Learning...

# Two (Three?) Camps of Optimizations

- **Manual and Semi-Manual Optimization**
  - e.g. The classic Hypothesis-Experiment-Evaluation Cycle
- **Automatic Optimization**
  - e.g., Online Learning, Multi-armed Bandits, Reinforcement Learning...
- **Combining Two Camps**

# Manual and Semi-Manual Optimization

Online Experiments and Evaluation  
Offline Experiments and Evaluation  
Observational Study

# Manual Optimization

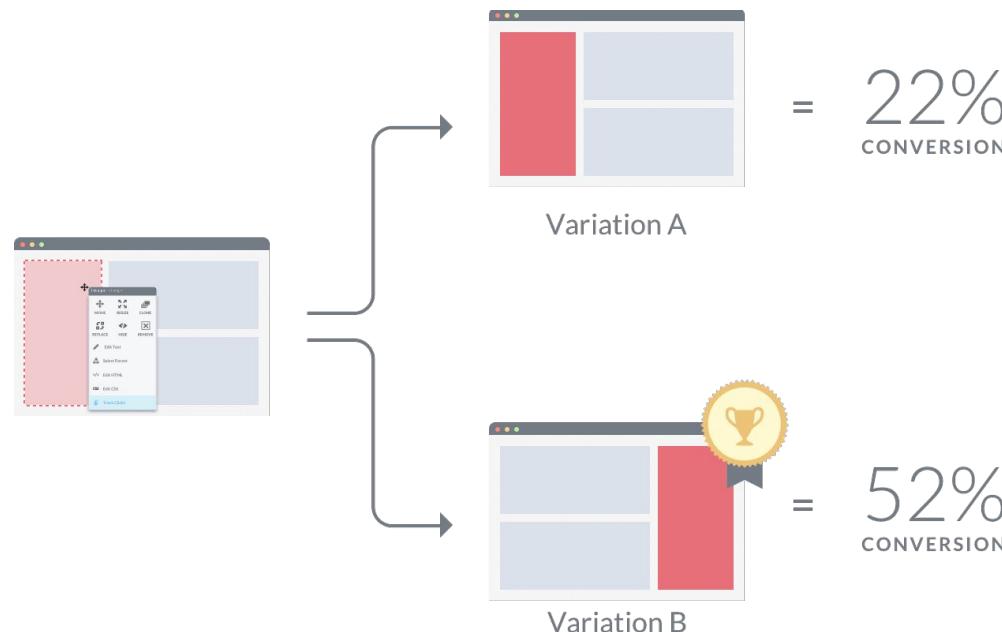
1. Choose a hypothesis to improve a metric.
2. Choose a realization of the hypothesis.
3. Launch an A/B online experiment to test the realization.
4. Monitor, measure and conclude the A/B online experiment.
5. If the realization improves the metric, go to Step 1, otherwise with certain probability go to Step 2, or go to Step 1.

# Manual Optimization

1. **Choose a hypothesis** to improve a metric.
  2. **Choose a realization** of the hypothesis.
  3. **Launch an A/B online experiment** to test the realization.
  4. **Monitor, measure** and **conclude** the A/B online experiment.
  5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.
- 
- Step 2-4 Inner Loop
  - Step 1-5 Outer Loop (*Data Scientist Ascent*)

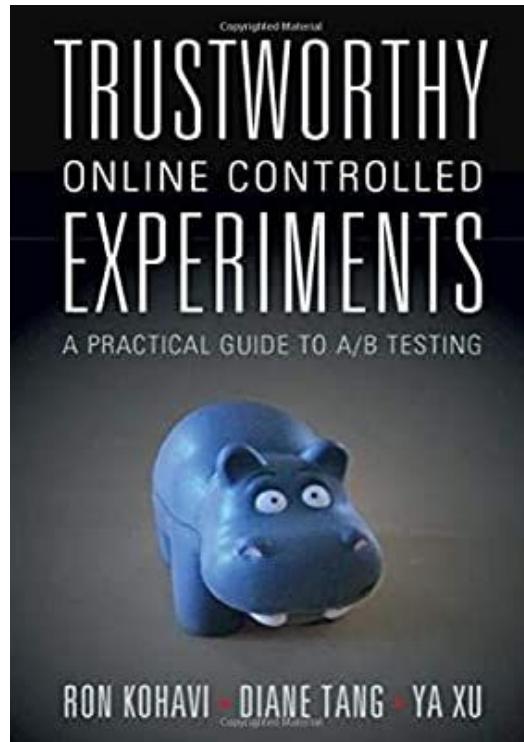
# Online Experiments and Evaluation

A/B Tests or Bucket Tests or Online Controlled Experiments



# Online Experiments and Evaluation

A/B Tests or Bucket Tests or Online Controlled Experiments



# Online Experiments and Evaluation

## Selected References

1. Somit Gupta, Xiaolin Shi, Pavel Dmitriev, Xin Fu, and Avijit Mukherjee. **Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.** WSDM 2020 Tutorial.
2. Somit Gupta, Xiaolin Shi, Pavel Dmitriev, and Xin Fu. **Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.** WWW 2020 Tutorial.
3. Xiaolin Shi, Pavel Dmitriev, Somit Gupta, and Xin Fu. **Challenges, Best Practices and Pitfalls in Evaluating Results of Online Controlled Experiments.** KDD 2019 Tutorial.
4. Somit Gupta, Ronny Kohavi, Alex Deng, Jeff Omhover, and Paweł Janowski. **A/B Testing at Scale: Accelerating Software Innovation.** WWW 2019 Tutorial.
5. Alexey Drutsa, Gleb Gusev, Eugene Kharitonov, Denis Kulemyakin, Pavel Serdyukov, and Igor Yashkov. **Effective Online Evaluation for Web Search.** SIGIR 2019 Tutorial
6. Tetsuya Sakai. **Conducting Laboratory Experiments Properly with Statistical Tools: An Easy Hands-on Tutorial.** SIGIR 2018 Tutorial.
7. Alex Deng, Pavel Dmitriev, Somit Gupta, Ron Kohavi, Paul Raff, and Lukas Vermeer. **A/B Testing at Scale: Accelerating Software Innovation.** SIGIR 2017 Tutorial.
8. Ben Carterette. **Statistical Significance Testing in Information Retrieval: Theory and Practice.** SIGIR 2017 Tutorial.

# Online Experiments and Evaluation

## Benefits from Running Controlled Online Experiments

- Statistical tools and software packages are available to monitor, measure and conclude the classic hypothesis testing setup.
- The difference of the main metric between the control and the treatment group could link to *Average Treatment Effect (ATE)* in Causal Inference and hence might explain the causal effects of a hypothesis on an outcome.
- It is *easy* to implement and easy to explain to practitioners, executives and large audience.

# Online Experiments and Evaluation

## Challenges from Running Controlled Online Experiments

- There are non-trivial pitfalls and challenges to conduct valid and meaningful online experiments.
- It is very easy to violate basic assumptions of running and monitoring an online experiment, hence obtaining invalid results (e.g., p-value hacking, peeking and etc.)
- It is sometimes puzzling to interpret results from an online experiment and therefore, hard to make a conclusion.
- It is even more challenging to run many series of experiments due to *false discovery rate* and other issues.
- ...

# Online Experiments and Evaluation

## Challenges from Running Controlled Online Experiments

- There are non-trivial pitfalls and challenges to conduct valid and meaningful online experiments.
- It is very easy to violate basic assumptions of running and monitoring an online experiment, hence obtaining invalid results (e.g., p-value hacking, peeking and etc.)
- It is sometimes puzzling to interpret results from an online experiment and therefore, hard to make a conclusion.
- It is even more challenging to run many series of experiments due to *false discovery rate* and other issues.
- ...

## References:

[1] Aaditya Ramdas. **Foundations of Large-Scale Sequential Experimentation**. In KDD 2019.

# Online Experiments and Evaluation

## Challenges from Running Controlled Online Experiments

- There are non-trivial pitfalls and challenges to conduct valid and meaningful online experiments.
- It is very easy to violate basic assumptions of running and monitoring an online experiment, hence obtaining invalid results (e.g., p-value hacking, peeking and etc.)
- It is sometimes puzzling to interpret results from an online experiment and therefore, hard to make a conclusion.
- It is even more challenging to run many series of experiments due to *false discovery rate* and other issues.
- ...

**It is not easy at all.**

# Online Experiments and Evaluation

## Challenges from Running Controlled Online Experiments

- There are non-trivial pitfalls and challenges to conduct valid and meaningful online experiments.
- It is very easy to violate basic assumptions of running and monitoring an online experiment, hence obtaining invalid results (e.g., p-value hacking, peeking and etc.)
- **It is sometimes puzzling to interpret results from an online experiment and therefore, hard to make a conclusion.**
- It is even more challenging to run many series of experiments due to *false discovery rate* and other issues.
- ...

**It is not easy at all.**

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

The image displays three screenshots of an online search ecosystem, likely from an e-commerce platform like Etsy, illustrating the flow of user interaction between search results, product details, and user-generated content.

**Top Left:** A screenshot of the search results page for "wedding dress". The results are filtered by category (Jewelry & Accessories) and include various items like boho wedding dresses and beach wedding dresses. The interface includes a sidebar for filters (e.g., shipping, special offers) and a main search area with a "Search" button.

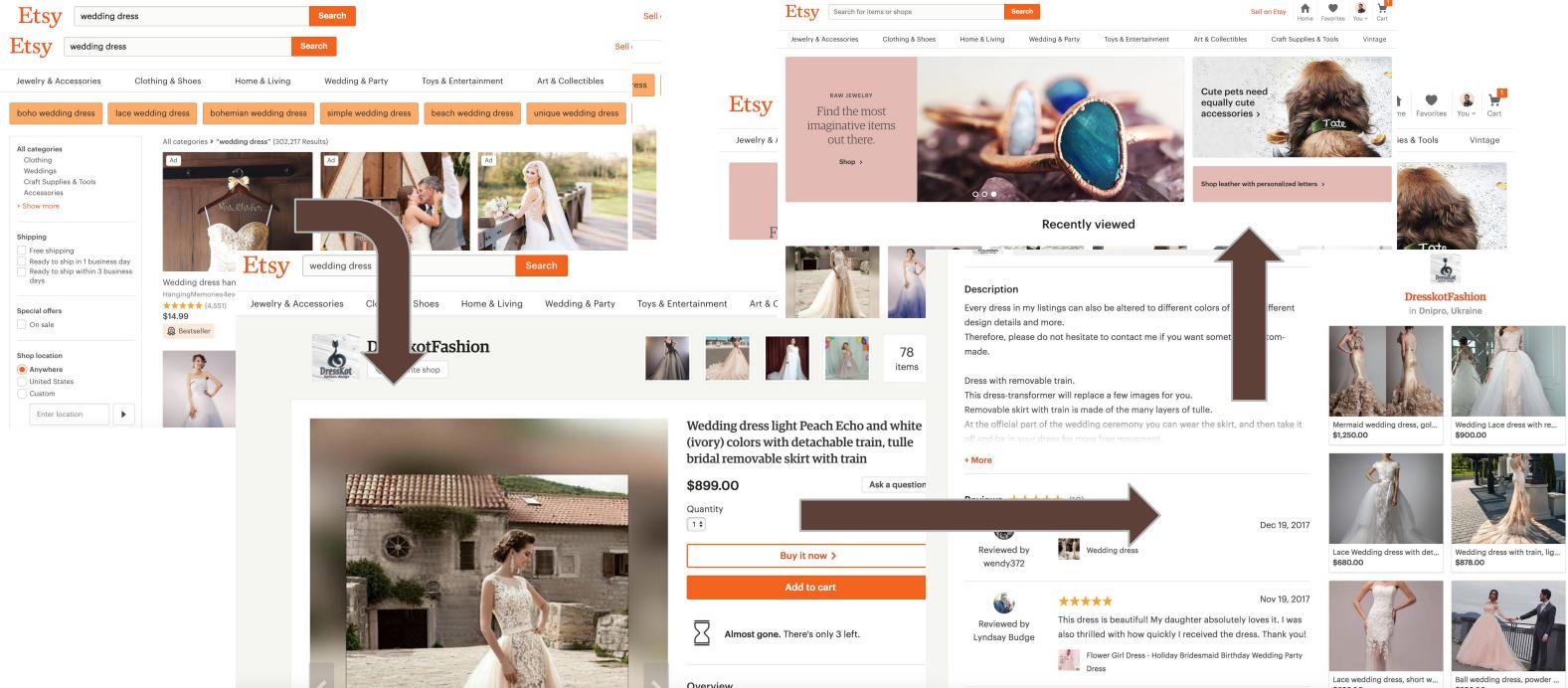
**Bottom Left:** A detailed view of a specific product listing for a "Wedding dress light Peach Echo and white (ivory) colors with detachable train, tulle bridal removable skirt with train". The product is priced at \$899.00, and the listing includes a "Buy it now" button, an "Add to cart" button, and a note about the item being almost gone. The product image shows a woman in a white wedding dress standing in front of a building.

**Top Right:** A screenshot of the homepage or a sidebar section. It features a "RAW JEWELRY" advertisement with a call to action "Find the most imaginative items out there. Shop >". Below this are sections for "Recently viewed" items and a "DresskotFashion" shop. The DresskotFashion shop page shows several wedding dresses and a review from a user named Wendy372.

**Bottom Right:** A close-up view of the "Recently viewed" section of the sidebar. It shows a review for a "Wedding dress" from a user named Lyndsay Budge, dated Nov 19, 2017. The review text reads: "This dress is beautiful! My daughter absolutely loves it. I was also thrilled with how quickly I received the dress. Thank you!" Below the review are other recommended products like "Lace Wedding dress with det... \$680.00" and "Ball wedding dress, powder ... \$999.00".

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems



Xuan Yin and Liangjie Hong. **The Identification and Estimation of Direct and Indirect Effects in Online A/B Tests through Causal Mediation Analysis.** In KDD 2019.

# Online Experiments and Evaluation

**Online Controlled Experiments on Search and Recommendation Ecosystems**

**Example I: An A/B Test Result for A New Recommendation Algorithm**

	% Change
<b>Recommendation Clicks</b>	+5%
<b>Search Clicks</b>	-3%
<b>Revenue</b>	~

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Example I: An A/B Test Result for A New Recommendation Algorithm

	% Change
<b>Recommendation Clicks</b>	+5%
<b>Search Clicks</b>	-3%
<b>Revenue</b>	~

- Improvements might come as a result of a series of A/B testing results.
- Not shipping early cornerstone results might lead to a sub-optimal user experience in a long run.
- Shipping placebo results might lead to a sub-optimal user experience in a long run.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Example II: An A/B Test Result for A New Recommendation Algorithm

	% Change
<b>Recommendation Clicks</b>	-10%
<b>Search Clicks</b>	+5%
<b>Revenue</b>	+1%

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Example II: An A/B Test Result for A New Recommendation Algorithm

	% Change
<b>Recommendation Clicks</b>	-10%
<b>Search Clicks</b>	+5%
<b>Revenue</b>	+1%

- Deteriorations might come as a result of a series of A/B testing results.
- Once damage is done, it might impact machine learning algorithms in many ways.
- Not shipping early cornerstone results might lead to a sub-optimal user experience in a long run.
- Shipping placebo results might lead to a sub-optimal user experience in a long run.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

We need to understand **the interplay** between recommendation and search modules as well as their **whole ecosystem** to create a coherent user experience and optimize user engagement.

- **Opportunity 1:** Understand experimental results while multiple teams work on different recommendation and search modules.
- **Opportunity 2:** Develop and implement strategies to improve multiple modules and possibly optimize overall user engagement.
- **Opportunity 3:** Develop machine learning models to directly optimize user engagement from a whole user journey perspective.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

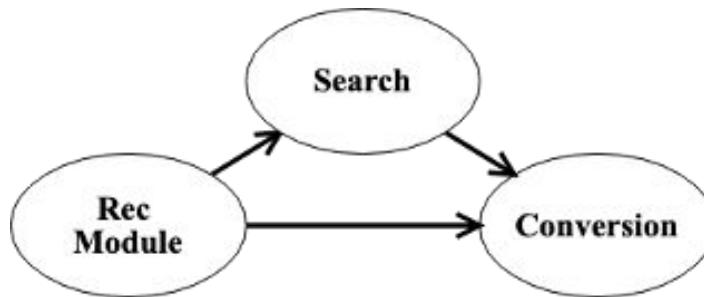
We need to understand **the interplay** between recommendation and search modules as well as their **whole ecosystem** to create a coherent user experience and optimize user engagement.

- **Opportunity 1: Understand experimental results while multiple teams work on different recommendation and search modules.**
- **Opportunity 2:** Develop and implement strategies to improve multiple modules and possibly optimize overall user engagement.
- **Opportunity 3:** Develop machine learning models to directly optimize user engagement from a whole user journey perspective.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

A product change could *induce* changes in user interaction with other products.



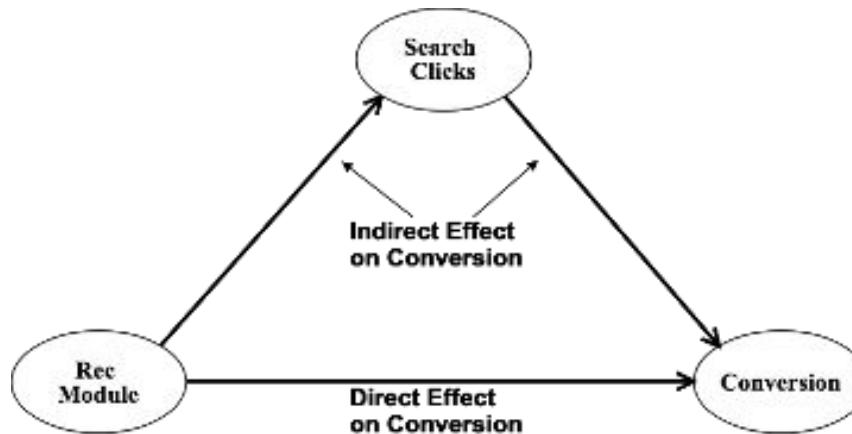
- An improved recommendation module could effectively suggest items that satisfy users' needs so that users don't need to search as much as usual.
- The overall performance of an improved recommendation module could be cannibalized by the induced reduction of user engagement in search.
- The performance of search could be cannibalized by an improved recommendation module.

# Online Experiments and Evaluation

Online Controlled Experiments on Search and Recommendation Ecosystems

Common Solution:

Splitting Average Treatment Effect (ATE) into Two Parts: Direct Effect and Indirect Effect



# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

**Common Solution:**

**Splitting Average Treatment Effect (ATE) into Two Parts: Direct Effect and Indirect Effect**

- Causal Mediation Analysis (**CMA**) is a formal statistical framework to conduct such analysis.
- Average Direct Effect (**ADE**) is the direct impact of new recommendations while keeping search behavior fixed.
- Average Causal Mediation Effect (**ACME**) is the impact of induced changes in search behavior due to changes in recommendation algorithm.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Common Solution:

#### Splitting Average Treatment Effect (ATE) into Two Parts: Direct Effect and Indirect Effect

- **ATE, ADE and ACME** has been studied extensively in the literature.
- Existing methodologies cannot be easily utilized due to violations of the key assumptions in the literature: no unmeasured causally-dependent mediator.
- A typical E-commerce site could have hundreds of web-pages and modules, and all of them could be mediators. It is difficult to measure all of them.
- We extended **ADE** and **ACME** to Generalized **ADE (GADE)** and Generalized **ACME (GACME)** respectively.
- It is easy to implement and only requires solving two linear regression equations simultaneously.
- Git Repo: <https://github.com/xuanyin/causal-mediation-analysis-for-ab-tests>

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Case I: RecSys Listing Page Same-Shop Experiment

Effect	% Change	
	Conversion Rate	GMV
<b>GADE</b> Direct Effect of the Change of Rec Module	<b>0.4959%*</b>	0.1681%
<b>GACME</b> The Effect of the Induced Change of Search	-0.2757%***	-0.4200%***
<b>ATE</b>	0.2202%	-0.2518%

#### Notes:

1. % Change = Effect/Mean of Control
2. \*\*\* p<0.001, \*\* p<0.01, \* p<0.05, .' p<0.1. Two-tailed p-value is derived from z-test for  $H_0$ : the effect is zero, which is based on asymptotic normality.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

### Case II: RecSys Listing Page Internal-Bottom Desktop Experiment

Effect	% Change	
	Conversion Rate	GMV
<b>GADE</b> Direct Effect of the Change of Rec Module	<b>0.3448%*</b>	0.0659%
<b>GACME</b> The Effect of the Induced Change of Search	-0.0570%.	-0.0926%.
<b>ATE</b>	<b>0.2878%.</b>	-0.0267%

#### Notes:

1. % Change = Effect/Mean of Control
2. \*\*\* p<0.001, \*\* p<0.01, \* p<0.05, . p<0.1. Two-tailed p-value is derived from z-test for  $H_0$ : the effect is zero, which is based on asymptotic normality.

# Online Experiments and Evaluation

## Online Controlled Experiments on Search and Recommendation Ecosystems

- Understanding direct vs. indirect effects enables us to understand the competition between recommendation modules and search results; and give more informed decisions during roll-outs.
- Develop better recommendation strategies such as suggesting items and categories not searched organically or diverse information shown in different surfaces.
- Develop better offline evaluation framework to incorporate both search and recommendation results.

# Recap: Online Experiments and Evaluation

## Benefits from Running Controlled Online Experiments

- Statistical tools and software packages are available to monitor, measure and conclude the classic hypothesis testing setup.
- The difference of the main metric between the control and the treatment group could link to **ATE** in Causal Inference and hence might explain the causal effects of a hypothesis on an outcome.
- It is *easy* to implement and easy to explain to practitioners, executives and large audience.

## Challenges from Running Controlled Online Experiments

- There are non-trivial pitfalls and challenges to conduct valid and meaningful online experiments.
- It is very easy to violate basic assumptions of running and monitoring an online experiment, hence obtaining invalid results (e.g., p-value hacking, peeking and etc.)
- It is sometimes puzzling to interpret results from an online experiment.
- It is even more challenging to run many series of experiments due to *false discovery rate* and other issues.

# Manual Optimization

1. **Choose a hypothesis** to improve a metric.
2. **Choose a realization** of the hypothesis.
3. **Launch an A/B online experiment** to test the realization.
4. **Monitor, measure** and **conclude** the A/B online experiment.
5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

# Manual Optimization

1. **Choose a hypothesis** to improve a metric.
2. **Choose a realization** of the hypothesis.
3. **Launch an A/B online experiment** to test the realization.
4. **Monitor, measure** and **conclude** the A/B online experiment.
5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

## Major Challenges:

- Limited throughout and bounded by traffic and time.
- Cannot re-use data.

# Semi-Manual Optimization

1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline experiments**.
3. Launch an **A/B online experiment** to test the realization.
4. **Monitor, measure** and **conclude** the A/B online experiment.
5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

# Semi-Manual Optimization

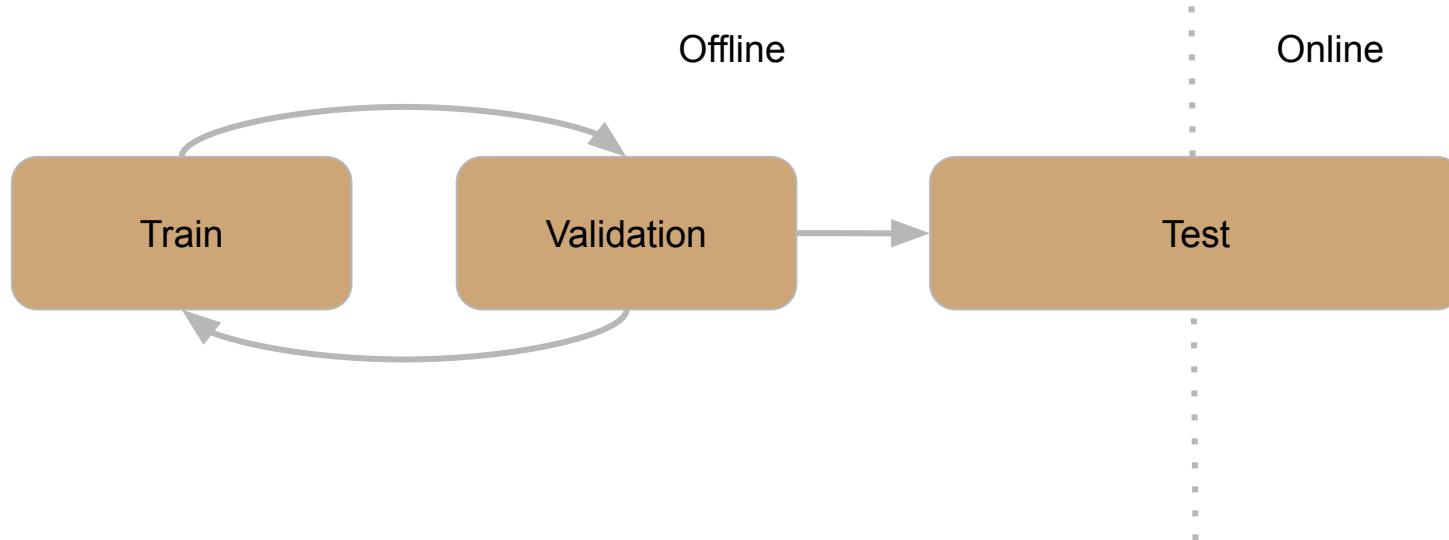
1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline experiments**.
3. Launch an **A/B online experiment** to test the realization.
4. **Monitor, measure** and **conclude** the A/B online experiment.
5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

## Main Ideas:

- Improve data reuse (e.g., offline experiments need datasets.)
- Improve throughout.
- Machine learning textbook scenario.

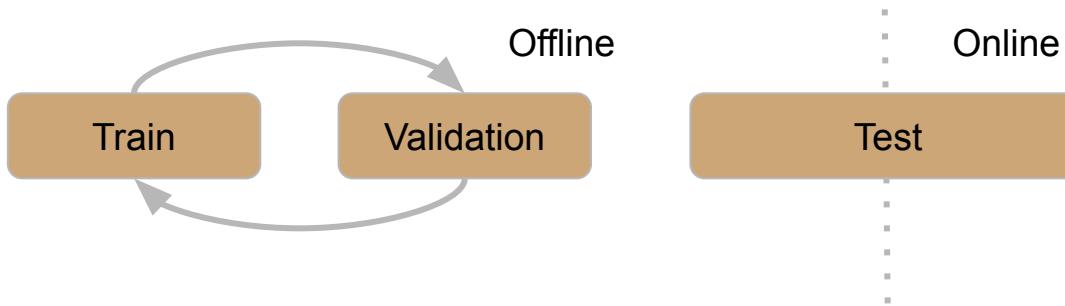
# Offline Experiment and Evaluation

## Traditional Offline Dataset/Collection Experiment



# Offline Experiment and Evaluation

- Supervised Learning
- Cross-validation
- View online experiments as extension to offline optimization (testset)



# Offline Experiment and Evaluation

## Optimizing Intra-Session Metrics

If inter-session metrics can be explicitly modeled or write them down in their clear form, you can use online optimization tools to directly optimize them.

# Offline Experiment and Evaluation

## Optimizing Intra-Session Metrics

If inter-session metrics can be explicitly modeled or write them down in their clear form, you can use online optimization tools to directly optimize them.

- This is usually **difficult** or **impossible** because of
  - Complexity of inter-session metrics (you can't really write them down or hard).
  - You don't have data.
  - You have extremely sparse data.
  - Hard to deploy such systems.

...

# Offline Experiment and Evaluation

The image displays two side-by-side screenshots of the Etsy search interface.

**Left Screenshot (Search for "wabi sabi"):**

- Header:** Shows the Etsy logo, search bar with "wabi sabi", a "Search" button, and navigation links for "Sell on Etsy", "Home", "Favorites", "You", and "Cart".
- Breadcrumbs:** "Jewelry & Accessories" → "Clothing & Shoes" → "Home & Living" → "Wedding & Party" → "Toys & Entertainment" → "Art & Collectibles" → "Craft Supplies & Tools" → "Vintage".
- Search Results:** "All categories > \"wabi sabi\" (6,213 Results)".
  - Product 1:** "Kintsugi bowl, kintsugi ceramic er Kanelasuri" by "Ad ST3jewellery". Price: \$84.41. Rating: ★★★★☆ (77). Description: "Free shipping".
  - Product 2:** "wabi-sabi" t-shirt by "Ad ST3jewellery".
- Filters and Sidebar:** Includes sections for "Special offers" (checkbox for "On sale"), "All categories" (with "Home & Living", "Art & Collectibles", "Jewelry", "Craft Supplies & Tools" options), "Shipping" (checkboxes for "Free shipping", "Ready to ship in 1 business day", "Ready to ship within 3 business days"), and "Shop location" (radio buttons for "Anywhere", "United States", "Custom", with "Enter location" input).

**Right Screenshot (Search for "jewelry box"):**

- Header:** Shows the Etsy logo, search bar with "jewelry box", a "Search" button, and navigation links for "Sell on Etsy", "Home", "Favorites", "You", and "Cart".
- Breadcrumbs:** "Jewelry & Accessories" → "Clothing & Shoes" → "Home & Living" → "Wedding & Party" → "Toys & Entertainment" → "Art & Collectibles" → "Craft Supplies & Tools" → "Vintage".
- Search Results:** "All categories > \"jewelry box\" (241,017 Results)".
  - Product 1:** "Raven box, handmade boxes, steampunk" by "Ad ST3jewellery". Price: \$30.95. Rating: ★★★★☆ (35).
  - Product 2:** "Bridesmaid Gift / Popular Bridesmaid..." by "SugarAndChicShop". Price: \$45.00. Rating: ★★★★☆ (1,208).
  - Product 3:** "Matte Black Custom Branded Laser..." by "Uzbeams". Price: \$85.00. Rating: ★★★★☆ (162).
  - Product 4:** "Personalized Memory Box, Keepsake ..." by "EngraveMyMemories". Price: \$29.95. Rating: ★★★★☆ (6,548). Description: "Eligible orders get 10% off".
- Filters and Sidebar:** Includes sections for "Sort by: Relevancy", "Special offers" (checkbox for "On sale"), "All categories" (with "Jewelry", "Home & Living", "Craft Supplies & Tools", "Weddings" options), "Shipping" (checkboxes for "Free shipping", "Ready to ship in 1 business day", "Ready to ship within 3 business days"), and "Shop location" (radio buttons for "Anywhere", "United States", "Custom", with "Enter location" input).

Liang Wu, Diane Hu, Liangjie Hong and Huan Liu. **Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce**. SIGIR 2018.

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

- Expected GMV

$$GMV = \underbrace{\sum_{\forall s \in S}}_{\text{A search session}} \underbrace{\sum_{\forall i^s}}_{\text{An item in } s} \underbrace{Price(i^s)}_{\text{Price of } i^s} \underbrace{Pr(\Phi = 1 | i^s, q^s)}_{\text{Prob of purchase}},$$

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

- Purchase Decision Process

rosy wedding dress

All categories > "rosy wedding dress" (72 Results)



60 Colors Chiffon Rosy Long Party Dress Evening...  
CHARMINGJOY  
★★★★★ (650)  
\$51.50



Rosy brown dress chiffon party dress ros...  
LovelyMellodyClothing  
★★★★★ (1,597)  
\$39.00

Rosy brown dress chiffon party dress ros...  
brown prom dress chiffon cocktail dress  
bow back dress rosy brown bridesmaid  
dresses chiffon dress

\$39.00

Style



Ivory Mauve Flower Girl Dress - Flower girl...  
BloomsNBugs  
★★★★★ (857)  
\$69.00



Rosy Mauve Satin Bridal Sash - Rosy Mauve We...  
encoreastoreNatalie  
★★★★★ (29)  
\$14.00

Ivory Mauve Flower Girl Dress - Flower girl...  
Dress Rosy Mauve - Flower Girl Dress - Dress  
for Flower Girls - flower girls Pink Mauve

\$69.00+

Only 1 available

Size

Search Page

Product Page

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

- Click Decision(s) from Search-Result-Page (SERP)
- Purchase Decision(s) from Listing Page

$$Pr(\Phi = 1|i, q) = \underbrace{Pr(\Psi = 1|i, q)}_{\text{click model}} \underbrace{Pr(\Phi = 1|\Psi = 1, i, q)}_{\text{purchase model}},$$

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

- Click Decision(s) from Search-Result-Page (SERP)

$$NDCG_K(\varrho) = N_{max}^{-1} \sum_{r=0}^{K-1} \frac{2^{l(r^{-1})}}{\log(1+r)},$$



$$\mathcal{L}_c = N_{max}^{-1} \sum_{i=1}^m \frac{2^{l(i)}}{\log(1 + \sum_{i_b=1, i_b \neq i_a}^m \sigma(f_c(x_a) - f_c(x_b)))},$$

$f_c$  is learned by a neural-network model through back-prop.

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

- Purchase Decision from Listing Page

$$\mathcal{L}_p = \sum_{i=1}^N Price(i) \log\{1 + \exp[-l'_i(w_p x_i)]\} + ||w_p||^2,$$

Price-Weighted Logistic Regression

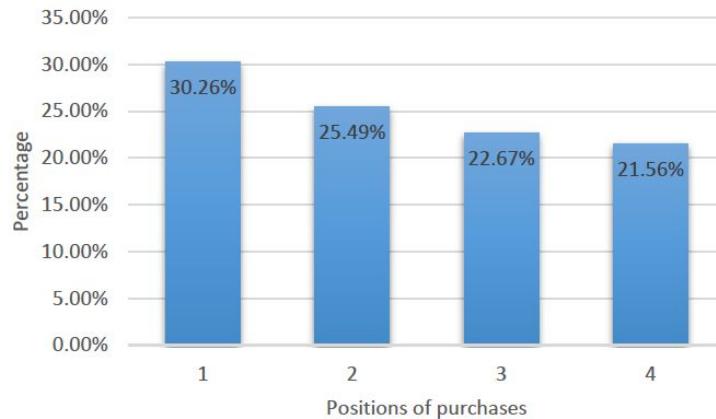
# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

Sessions	Queries	Items	Avg. Items per Session
334,931	239,928	6,347,251	19.0
Keywords	Buyers	Sellers	Avg. Items per Query
631,778	270,239	550,025	26.5

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search



**Figure 2: Position distribution of items being purchased in the top 4 spots of a search result page. The first position achieves the most purchases, while nearly 70% of purchases are in the lower positions.**

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

Relevance	Low Level	Sum of TF
		Sum of Log TF
		Sum of Normalized TF
		Sum of Log Normalized TF
		Sum of IDF
		Sum of Log IDF
		Sum of ICF
		Sum of TF-IDF
		Sum of Log TF-IDF
		TF-Log IDF
		Length
		Log Length
	High Level	BM25
		Log BM25
		LM <sub>DIR</sub>
		LM <sub>JM</sub>
		LM <sub>ABS</sub>
Revenue		Price
		Price - Cat.Mean
		(Price - Cat.Mean)/Cat.Mean

Click	RankNet [1]	RNet
	RankBoost [10]	RBoost
	AdaRank [39]	ARank
	LambdaRank [2]	LRank
	ListNet [3]	LNet
	MART [12]	MART
	LambdaMART [38]	LMART
Purchase	SVM [4]	SVM
	Logistic Regression [28]	LR
	Random Forest [22]	RM
Both	Weighted Purchase [44]	WT
	LMART+RM	LMRM
	LETORIF	LETORIF

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

Category	Method	Click NDCG@5			Purchase NDCG@5			Revenue NDCG@5		
		Train	Vali	Test	Train	Vali	Test	Train	Vali	Test
Click	RNet	0.1743	0.1731	0.1378**	0.1672	0.1721	0.1676**	0.1692	0.1700	0.1356**
	RBoost	0.2150	0.1768	0.1323**	0.2150	0.1768	0.1715**	0.2150	0.1768	0.1311**
	ARank	0.1718	0.1711	0.1351**	0.1718	0.1711	0.1706**	0.1718	0.1711	0.1358**
	LRank	0.1694	0.1688	0.1360**	0.1678	0.1711	0.1672**	0.1713	0.1719	0.1366**
	LNet	0.1665	0.1703	0.1355**	0.1601	0.1682	0.1620**	0.1646	0.1696	0.1348**
	MART	0.2700	0.1758	0.1380**	0.2155	0.1803	0.1796*	0.2696	0.1688	0.1408**
	LMART	0.3056	0.1777	<b>0.1412</b>	0.3056	0.1777	0.1717**	0.3056	0.1777	0.1370**
Purchase	SVM	0.1785	0.1772	0.1336**	0.1831	0.1754	0.1755**	0.1816	0.1752	0.1320**
	LR	0.1978	0.1739	0.1310**	0.1978	0.1739	0.1782**	0.1978	0.1739	0.1332**
	RM	0.3359	0.1698	0.1363**	0.3329	0.2305	0.1798**	0.3327	0.1685	0.1376**
Both	WT	0.1970	0.1682	0.1334**	0.1815	0.1763	0.1761**	0.1781	0.1648	0.1375**
	LMRM	0.2943	0.2597	0.1354**	0.3087	0.2530	0.1688**	0.2943	0.2594	0.1332**
	LETORIF	0.1765	0.1550	0.1351**	0.2731	0.1841	<b>0.1801</b>	0.2039	0.1698	<b>0.1494</b>

Symbol \* indicates that the method is outperformed by the best one by 0.05 statistical significance level, \*\* indicates 0.01.

# Offline Experiment and Evaluation

## Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search

Category	Method	Rev@1	Rev@2	Rev@3	Rev@4	Rev@5	Rev@6	Rev@7	Rev@8	Rev@9	Rev@10
Click	RNet	4.47**	4.69**	4.89**	4.91*	5.06**	5.23**	5.21**	5.33**	5.46**	5.55**
	RBoost	4.57**	4.69**	4.69**	4.76**	4.97**	5.17**	5.23**	5.36**	5.49**	5.57**
	ARank	4.37**	4.66**	4.76**	4.90**	5.06**	5.20*	5.33**	5.47**	5.59**	5.67**
	LRank	4.38**	4.61**	4.74**	4.86**	5.07**	5.25**	5.42**	5.42**	5.67**	5.78**
	LNet	4.30**	4.59**	4.78**	4.99**	5.16**	5.35**	5.49**	5.61**	5.63**	5.63**
	MART	<b>4.62</b>	4.72**	4.86**	5.04**	5.26**	5.47**	5.47**	5.64**	5.74**	5.86**
	LMART	4.46*	4.54**	4.73**	5.10**	5.31**	5.56**	5.75**	5.90*	6.01**	6.14**
Purchase	SVM	4.41**	4.54**	4.76**	4.77**	4.95**	5.16**	5.34**	5.50**	5.64**	5.77**
	LR	4.29**	4.65**	4.65**	4.69**	4.74**	4.81*	4.94**	4.97**	5.11**	5.11**
	RM	4.52**	4.82**	4.86**	5.02**	5.18**	5.33*	5.50**	5.66**	5.79**	5.92**
Both	WT	4.52**	4.69**	4.80**	4.85**	5.01**	5.07**	5.23**	5.32**	5.35**	5.41**
	LMRM	4.42**	4.50**	4.72**	5.08**	5.23**	5.41**	5.57**	5.60**	5.73**	5.85**
	LETORIF	4.58**	<b>4.90</b>	<b>5.08</b>	<b>5.47</b>	<b>5.64</b>	<b>5.85</b>	<b>6.02</b>	<b>6.19</b>	<b>6.40</b>	<b>6.54</b>

Symbol \* indicates that the method is outperformed by the best one by 0.05 statistical significance level, \*\* indicates 0.01.

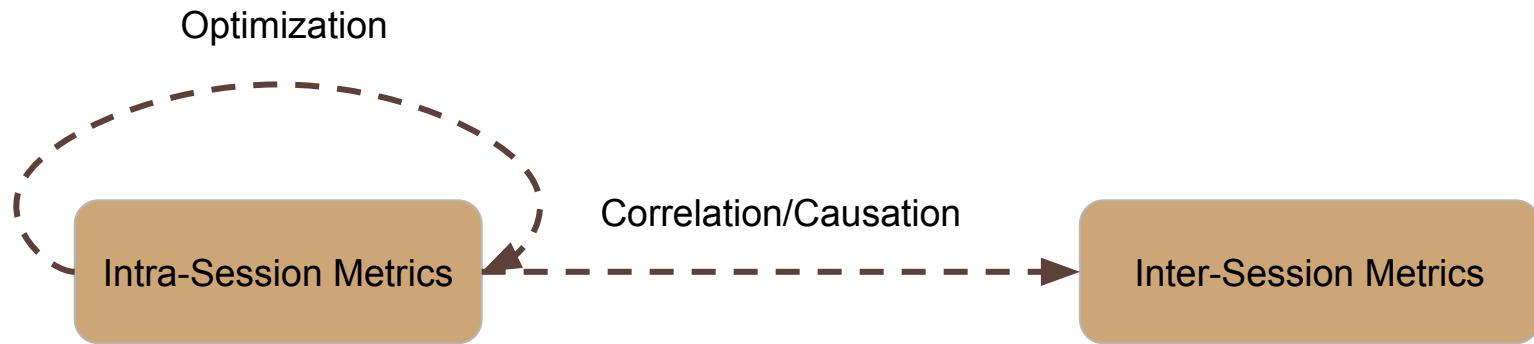
# Offline Experiment and Evaluation

## **Optimizing Gross-Merchandise-Value (GMV) in E-commerce Search**

- This work is about optimizing GMV in Session
  - How about long-term GMV?
  - How about other discovery?
- ...
- First step in optimizing user engagements in E-commerce search.

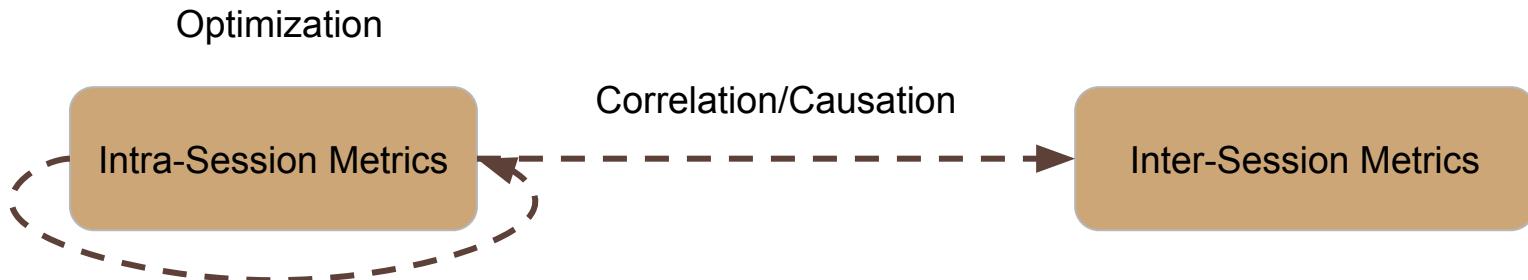
# Offline Experiment and Evaluation

## Optimizing Inter-Session Metrics

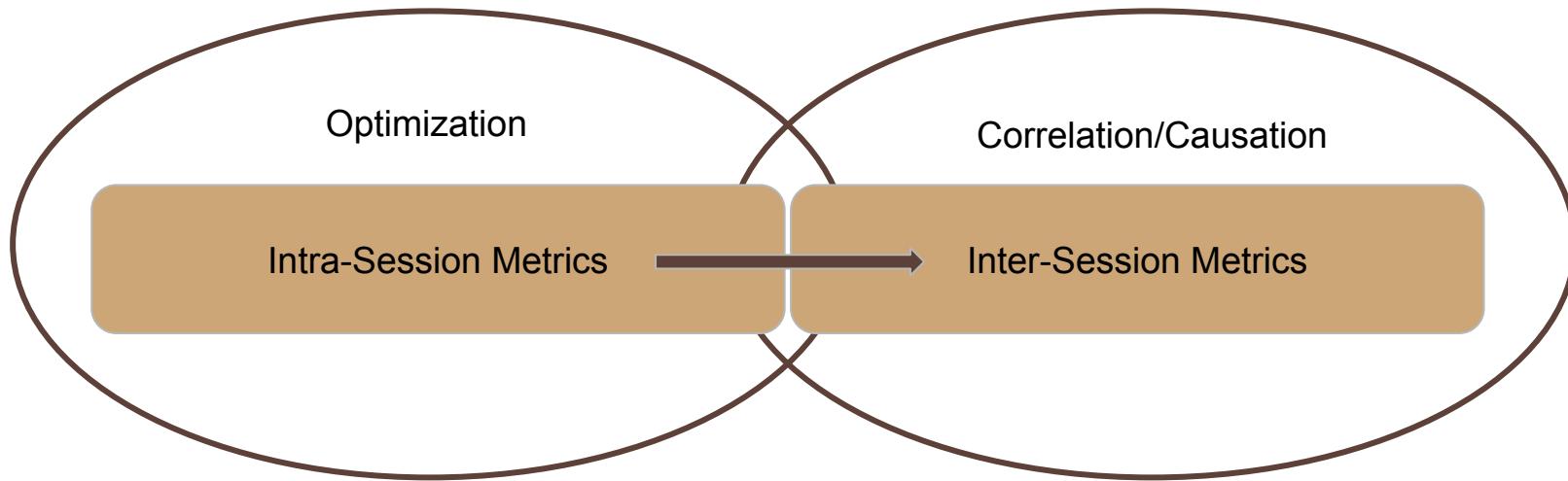


# Offline Experiment and Evaluation

1. Intra-Session and Inter-Session Correlation
2. Optimization Intra-Session as Surrogate
3. Finding (*Better*) Proxy Metrics



# Offline Experiment and Evaluation



# Offline Experiment and Evaluation

## Beyond Clicks: Dwell Time in Personalization

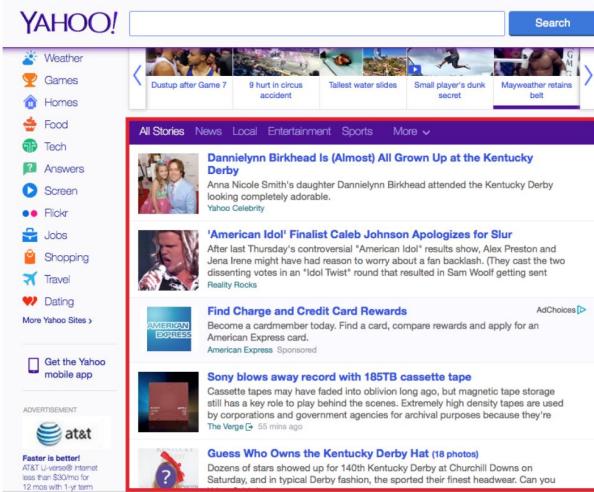
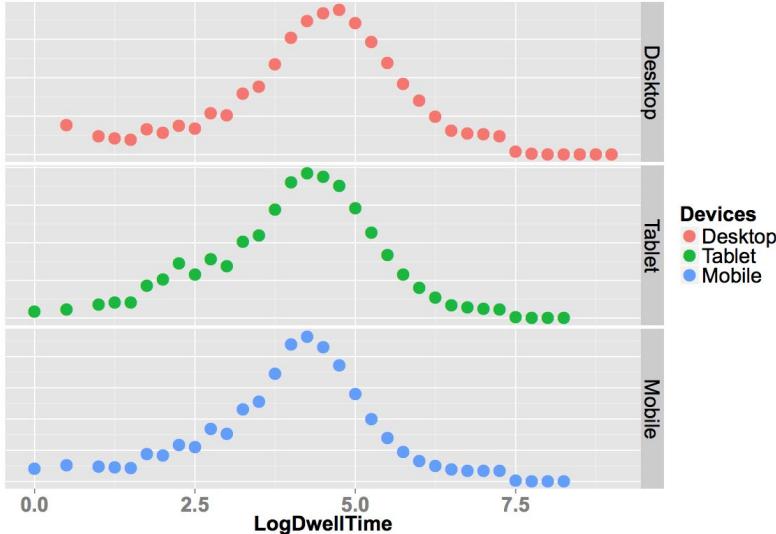


Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.

# Offline Experiment and Evaluation

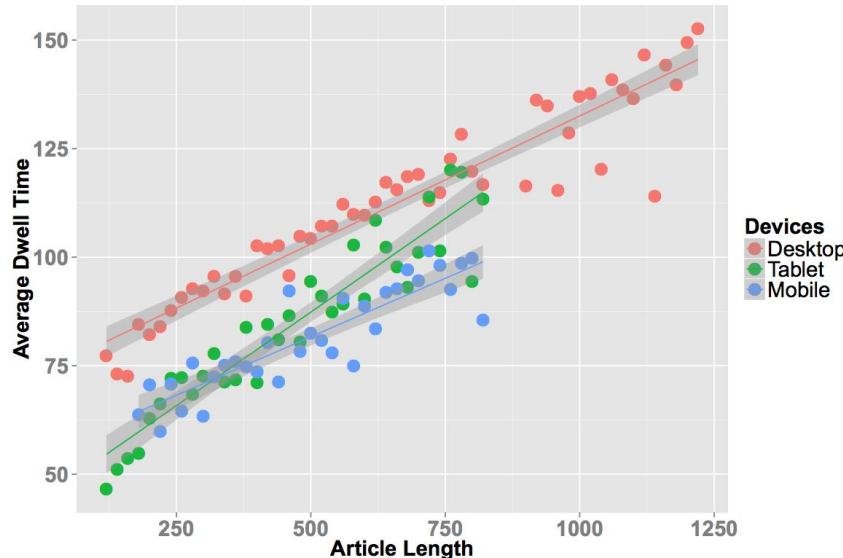
## Beyond Clicks: Dwell Time in Personalization



**Figure 2:** The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

# Offline Experiment and Evaluation

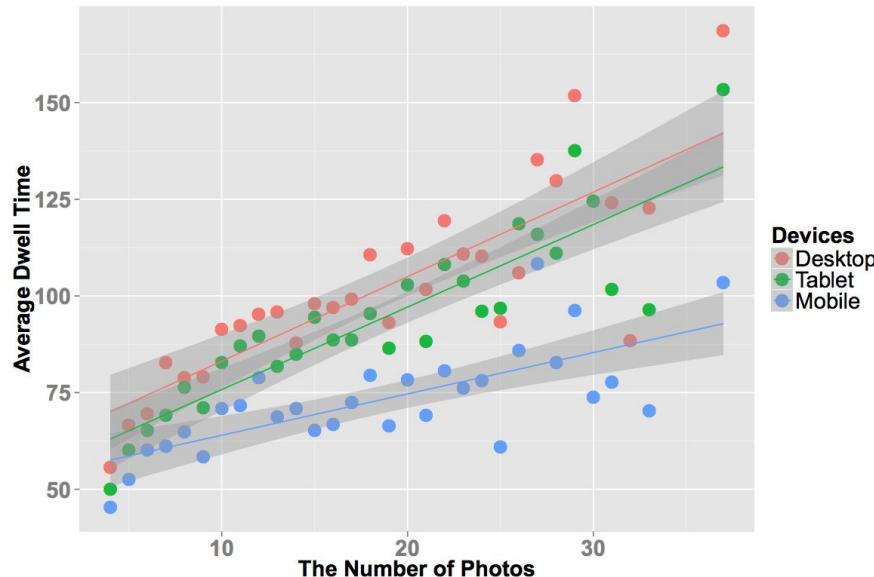
## Beyond Clicks: Dwell Time in Personalization



**Figure 3:** The relationship between the average dwell time and the article length where X-axis is the binned article length and the Y-axis is binned average dwell time.

# Offline Experiment and Evaluation

## Beyond Clicks: Dwell Time in Personalization



**Figure 4:** The relationship between the average dwell time and the number of photos on a slideshow where X-axis is the binned number of photos and the Y-axis is binned average dwell time.

# Offline Experiment and Evaluation

## Beyond Clicks: Dwell Time in Personalization

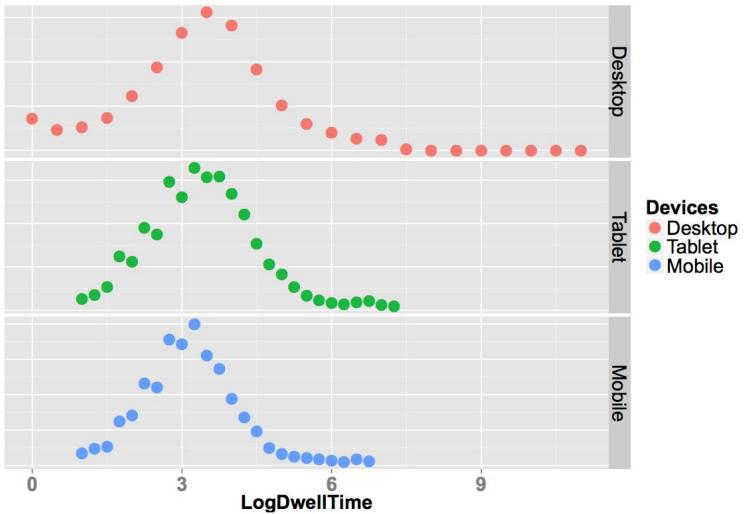


Figure 5: The (un)normalized distribution of log of dwell time for slideshows across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

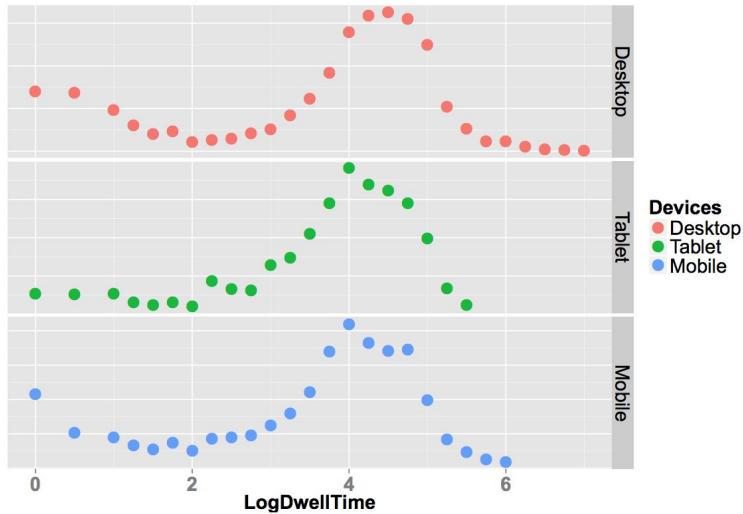


Figure 6: The (un)normalized distribution of log of dwell time for videos across different devices. The X-axis is the log of dwell time and the Y-axis is the counts.

# Offline Experiment and Evaluation

## Beyond Clicks: Dwell Time in Personalization

Table 4: Offline Performance for Learning to Rank

Signal	MAP	NDCG	NDCG@10
Click as Target	0.4111	0.6125	0.5680
Dwell Time as Target	0.4210	0.6201	0.5793
Dwell Time as Weight	0.4232	0.6226	0.5820

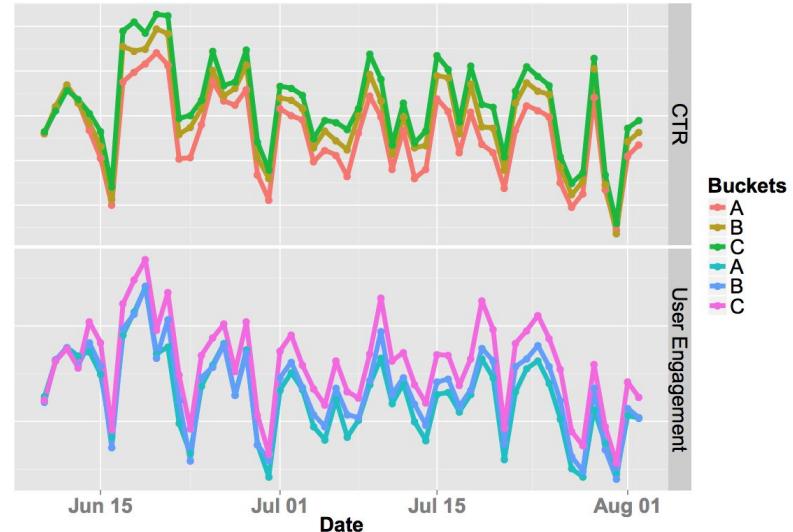


Figure 7: The relative performance comparison between three buckets. The top figure shows the relative CTR difference and the bottom figure shows the relative user engagement difference.

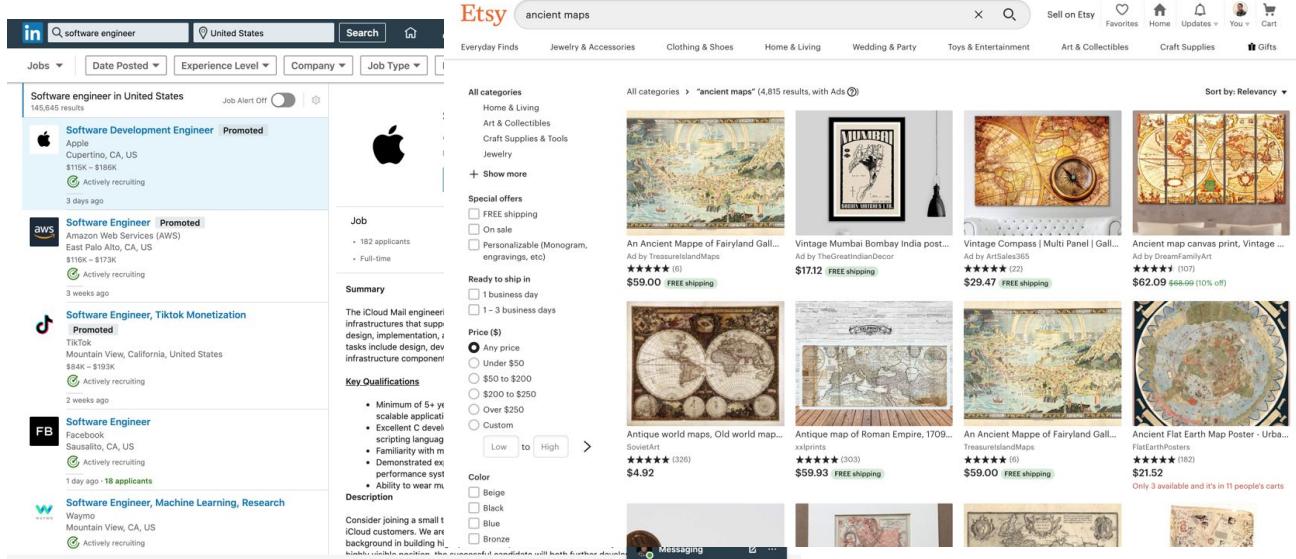
# Offline Experiment and Evaluation

## Beyond Clicks: Dwell Time in Personalization

- Optimizing Dwell-Time becomes the *de-facto* method to drive user engagement in Yahoo News Stream.
- The inter-session user engagement metric is a variant of dwell-time on sessions, considering the depth of the session.
- They correlate very well in quarterly basis.

# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis The Causal Path from Algorithms to Business

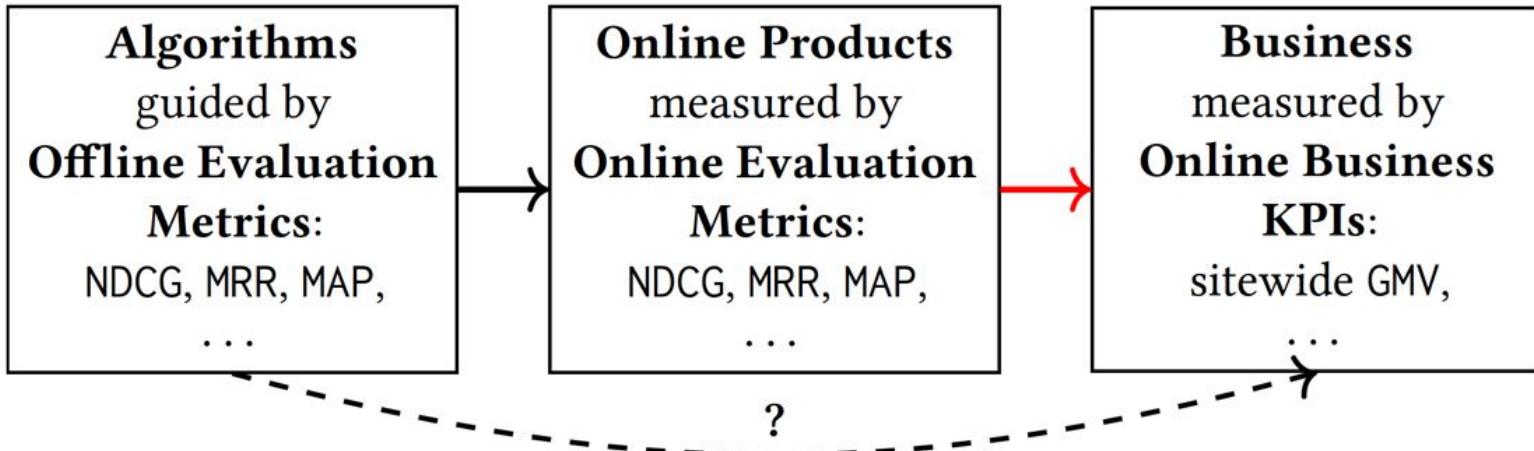


Z. Wang, X. Yin, T. L and L. Hong. **Causal Meta-Mediation Analysis: Inferring Dose-Response Function From Summary Statistics of Many Randomized Experiments.** KDD 2020.

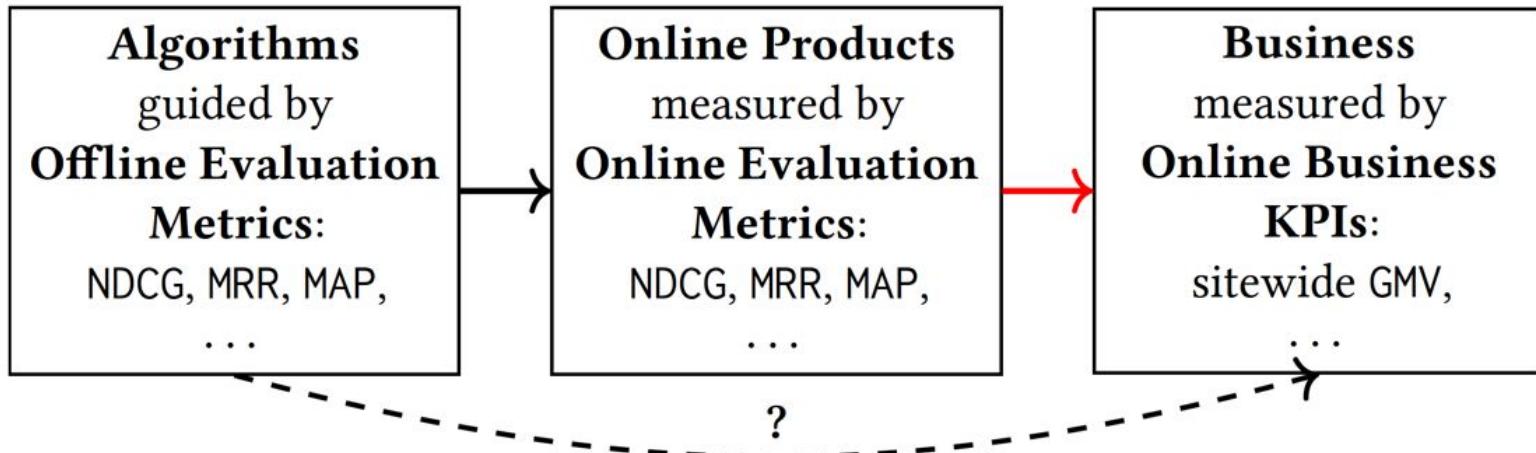
# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis

### The Causal Path from Algorithms to Business

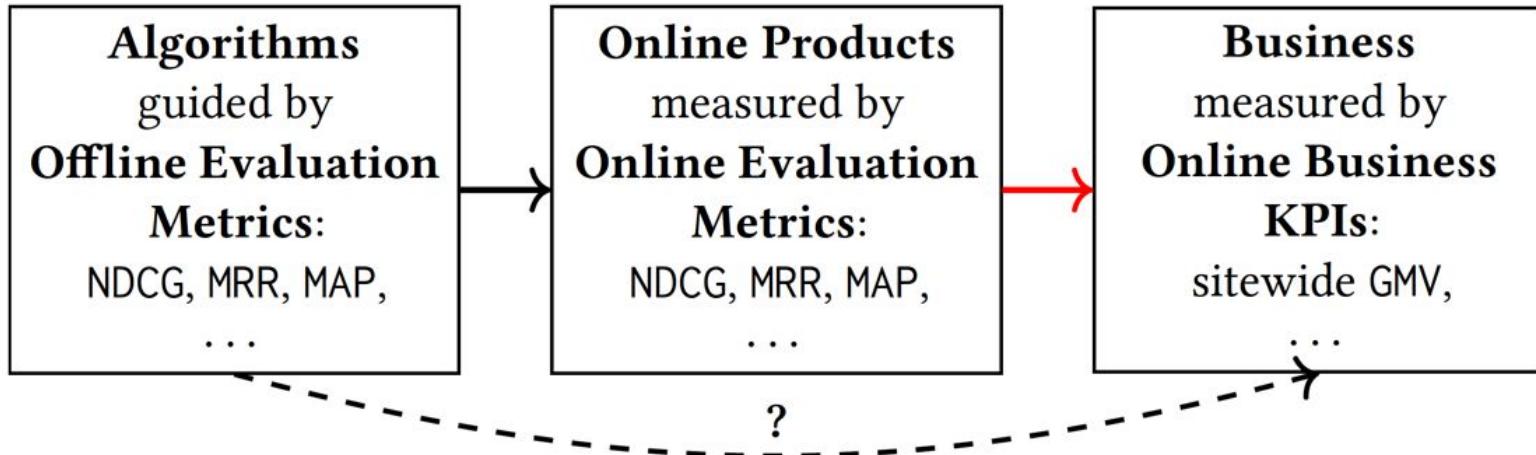


# Offline Experiment and Evaluation



- First part (the black arrow): counterfactual estimators of offline evaluation metrics to bridge the inconsistency between changes of offline and online evaluation metrics.
- Second part (the red arrow): the causality between online products (assessed by online evaluation metrics) and the business (assessed by online business KPIs).  
e.g. how business KPIs would change for a 10% increase in an online evaluation metric.

# Offline Experiment and Evaluation



- We model the causality between online evaluation metrics and business KPIs by *dose-response function (DRF)* in potential outcome framework.
- Instead of conducting online tests, we use results from historical A/B experiments to conduct Meta-Analysis.
- Online evaluation metrics could be mediators that (partially) transmit causal effects of treatments on business KPIs in experiments where treatments are not necessarily algorithm-related.

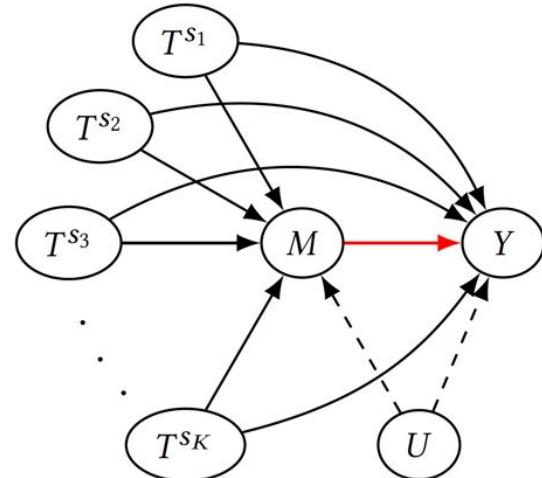
# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis

### The Causal Path from Algorithms to Business

#### Key Ideas:

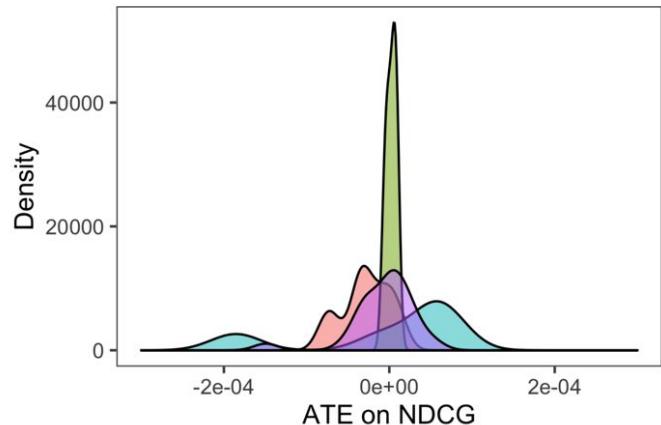
- $T$ s are treatments;  $M$  is a mediator;  $Y$  is a outcome;  $U$  is unobserved and unmeasured.  $M$  is online evaluation metric.  $Y$  is online business KPI.



# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis

### The Causal Path from Algorithms to Business



#### Data:

- 190 experiments from different teams.
- The figure shows that basic assumptions used by the method holds: enough variations.

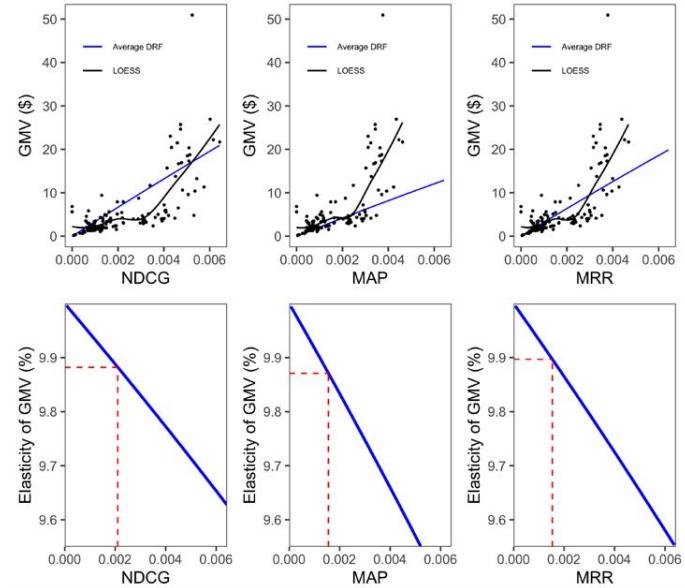
# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis

### The Causal Path from Algorithms to Business

#### Results:

- NDCG, MAP, MRR all have positive causal relationships with GMV.
- We could utilize elasticity to choose a better metric.



# Offline Experiment and Evaluation

## Causal Meta-Mediation Analysis

### The Causal Path from Algorithms to Business

- We model the causality between online evaluation metrics and business KPIs by dose-response function (DRF) in potential outcome framework.
- Instead of conducting online tests, we use results from historical A/B experiments to conduct Meta-Analysis.
- From 190 experiments' data, we have established positive causal relationships between offline metrics and business KPIs and also could choose which metric is better.

# Recap: Offline Experiment and Evaluation

## Benefits from Running Offline Experiments

- Offline experiments can be treated as classic machine learning scenarios.
- A family of hypotheses can be easily evaluated due to data reuse.
- Offline experiments can test potentially highly risk ideas without real harm.

## Challenges and Opportunities from Running Offline Experiments

- Offline experiments typically do not generalize to online settings due to biases, concept drifts and etc.
- Optimizing user engagement metrics usually require working with proxy metrics.
- Obtaining causal relationships between offline settings and online settings is hard.

[1] Mark Sanderson. **Test Collection Based Evaluation of Information Retrieval Systems**. Foundations and Trends® in Information Retrieval: Vol. 4: No. 4, 2010

[2] Donna Harman. **Information Retrieval Evaluation**. Synthesis Lectures on Information Concepts, Retrieval, and Services 3:2, 2011.

# Semi-Manual Optimization

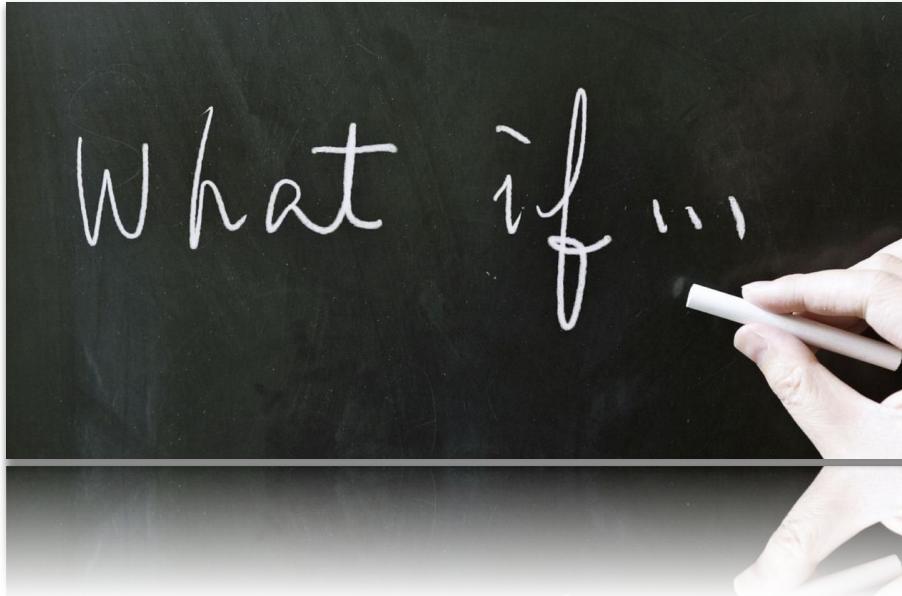
1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline experiments**.
3. Launch an **A/B online experiment** to test the realization.
4. **Monitor, measure** and **conclude** the A/B online experiment.
5. **If the realization improves the metric**, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

# Semi-Manual Optimization

1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline experiments with (some) guarantees**.
3. Launch an A/B online experiment to test the realization.
4. Monitor, measure and conclude the A/B online experiment.
5. If the realization improves the metric, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

# Offline A/B Experiment and Evaluation

Counterfactual Offline Reasoning/Experiment



# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

### Logging Policy

- Uniform-randomly show items.
- Gather user feedbacks (rewards).

### New Policy

- Show items according to a model/algorithm.
- Accumulate rewards if item matches history pattern.

[1] Lihong Li, Wei Chu, John Langford and Xuanhui Wang. **Unbiased Online Evaluation of Contextual-bandit-based News Article Recommendation Algorithms**. WSDM 2011.

[2] Alexander Strehl, John Langford, Lihong Li and Sham Kakade. **Learning from Logged Implicit Exploration data**. NIPS 2010.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment



Figure 1: A snapshot of the “Featured” tab in the Today Module on the Yahoo! Front Page [14]. By default, the article at F1 position is highlighted at the story position.

Lihong Li, Wei Chu, John Langford and Xuanhui Wang. **Unbiased Online Evaluation of Contextual-bandit-based News Article Recommendation Algorithms**. WSDM 2011.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

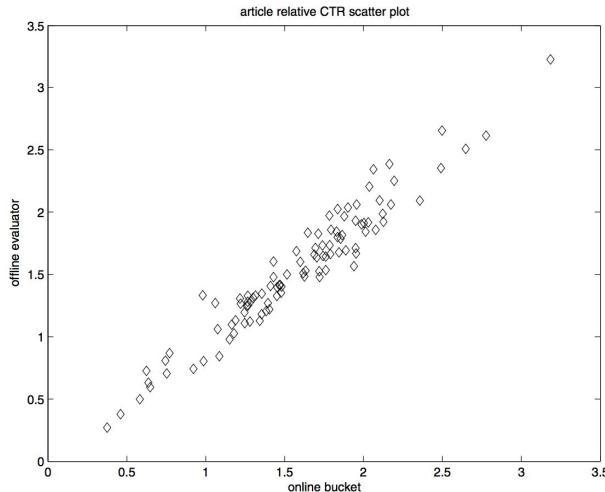


Figure 2: Articles' CTRs in the online bucket versus offline estimates.

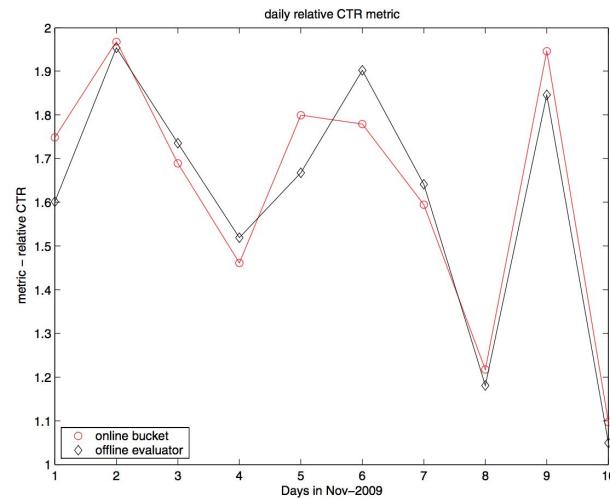


Figure 3: Daily overall CTRs in the online bucket versus offline estimates.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

- Address data bias
- Causality
- Reusable
- Some good theories

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

Etsy

Sell on Etsy [Home](#) [Favorites](#) [You](#) [Cart](#)

Jewelry & Accessories Clothing & Shoes Home & Living Wedding & Party Toys & Entertainment Art & Collectibles Craft Supplies & Tools Vintage

jewelry box wood wooden jewelry box large jewelry box small jewelry box jewelry box vintage personalized jewelry box

All categories Jewelry Home & Living Craft Supplies & Tools Weddings + Show more

Shipping  Free shipping  Ready to ship in 1 business day  Ready to ship within 3 business days

Special offers  On sale

Shop location  Anywhere  United States  Custom Enter location

Did you mean the shop [JewelryBox?](#)

All categories > "Jewelry box" (241,017 Results)

 Ad Raven box, handmade boxes, steampunk style ST3jewellery **★★★★★** (35) \$30.95

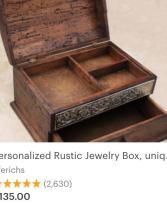
 Ad Bridesmaid Gift / Popular Bridesmaid... SugarAndChicShop **★★★★★** (1,208) \$45.00

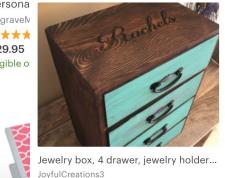
 Ad Matte Black Custom Branded Laser... Izbbeams **★★★★★** (162) \$85.00

 Ad Flower Girl or Bridesmaids Gift Box J... ZCreateDesign **★★★★★** (17,835) \$16.20 ~~\$18.00~~ (10% off)

 Wall Jewelry Box with metal mesh do... HookedOnWoodwork **★★★★★** (252) \$98.00

 Built on Order - Reclaimed Wood Box... IndependentBoxWorks **★★★★★** (554) \$40.00

 Personalized Rustic Jewelry Box, uniqu... dierichs **★★★★★** (2,630) \$135.00

 Jewelry box, 4 drawer, jewelry holder... JoyfulCreations3 **★★★★★** (216) \$35.00

 Rustic Jewelry Box - Custom engrave... Dustyroadgurl **★★★★★** (1,954) \$28.00

 Wood Jewelry Box candntreasures **★★★★★** (40) \$10.00

 Personalized Rustic Wedding Wood B... danielcuostromade **★★★★★** (724) \$15.00

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

$$\hat{v}_1(\pi) := \frac{1}{n} \sum_{i=1}^n \frac{\pi(a_i|q_i)}{p_i} r_i$$

The screenshot shows the Etsy search interface with the query "jewelry box" entered. The results page displays various items related to jewelry boxes, such as a hand-painted raven box, a gold frame with a monogram, and a black matte box for whiskey. The interface includes filters for categories like Jewelry & Accessories, shipping options, and special offers, along with a sidebar for shop location.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

- Need logging and an exploration strategy
- In development, emerging topic

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

**How to effectively gather data that minimize hurting user engagement metrics?**

[1] Liangjie Hong and Adnan Boz. **An Unbiased Data Collection and Content Exploitation/Exploration Strategy for Personalization.** CoRR abs/1604.03506, 2016.

[2] Tobias Schnabel, Paul N. Bennett, Susan Dumais and Thorsten Joachims. **Short-Term Satisfaction and Long-Term Coverage: Understanding How Users Tolerate Algorithmic Exploration.** WSDM 2018.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

**How to effectively gather data that minimize hurting user engagement metrics?**

- Uniform-random greatly *hurts* user engagement and *nobody* is doing this.
- Classic Thompson Sampling and Upper-Confidence-Bound would eventually *converge*.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

**How to effectively gather data that minimize hurting user engagement metrics?**

- Uniform-random greatly *hurts* user engagement and *nobody* is doing this.
- Classic Thompson Sampling and Upper-Confidence-Bound would eventually *converge*.

## Requirements:

- Provide **randomness** and **do not** converge.
- User-friendly.

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

**How to effectively gather data that minimize hurting user engagement metrics?**

---

**Algorithm 3** Thompson Sampling for Bernoulli Ranked-list Bandit

---

**Require:**  $\alpha, \beta$  prior parameters of a Beta distribution

$S_i = 0$  and  $F_i = 0, \forall i$  {Success and failure counters}

**for**  $t = 1, \dots, T$  **do**

**for**  $i = 1, \dots, K$  **do**

        Draw  $\theta_i$  according to  $\text{Beta}(S_i + \alpha, F_i + \beta)$ .

**end for**

    Compute  $\mathbf{p}$  such that  $p_k = \frac{\theta_k}{\sum \theta_k}$ .

    Sample  $N$  items from  $\text{Mult}(\mathbf{p})$ .

    Observe  $N$  rewards  $\mathbf{r}_t$ .

    Update  $S$  and  $F$  for those  $N$  items according to  $\mathbf{r}_t$ .

    Logging  $N$  items,  $\mathbf{p}$  and  $\mathbf{r}_t$ .

**end for**

---

# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

How to effectively gather data that minimize hurting user engagement metrics?

---

**Algorithm 3** Thompson Sampling for Bernoulli Ranked-list Bandit

---

**Require:**  $\alpha, \beta$  prior parameters of a Beta distribution

$S_i = 0$  and  $F_i = 0, \forall i$  {Success and failure counters}

**for**  $t = 1, \dots, T$  **do**

**for**  $i = 1, \dots, K$  **do**

        Draw  $\theta_i$  according to  $\text{Beta}(S_i + \alpha, F_i + \beta)$ .

**end for**

    Compute  $\mathbf{p}$  such that  $p_k = \frac{\theta_k}{\sum \theta_k}$ .

**Sample  $N$  items from  $\text{Mult}(\mathbf{p})$ .**

    Observe  $N$  rewards  $\mathbf{r}_t$ .

    Update  $S$  and  $F$  for those  $N$  items according to  $\mathbf{r}_t$ .

    Logging  $N$  items,  $\mathbf{p}$  and  $\mathbf{r}_t$ .

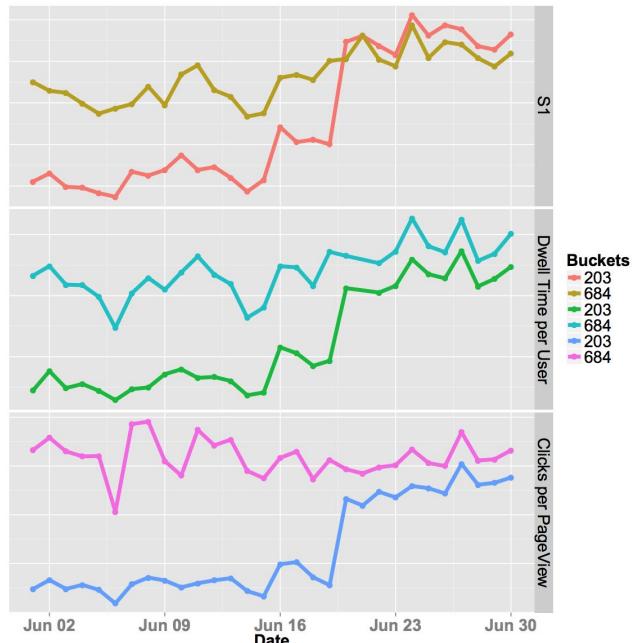
**end for**

---

# Offline A/B Experiment and Evaluation

Counterfactual Offline Reasoning/Experiment

How to effectively gather data that minimize hurting user engagement metrics?



# Offline A/B Experiment and Evaluation

## Counterfactual Offline Reasoning/Experiment

How to effectively gather data that minimize hurting user engagement metrics?

Algorithm	Metrics	Skewness	Mean	Median
New Algorithm	View Distribution	6.76	10,868.46	2,500.00
Old Algorithm		9.65	2,328.70	441.50
New Algorithm	Click Distribution	14.46	1,059.25	64.00
Old Algorithm		14.64	241.17	7.00
New Algorithm	CTR Distribution	2.28	0.04	0.03
Old Algorithm		3.87	0.03	0.02
New Algorithm	Item Cold-Start Distribution	1.15	37.26	13.86
Old Algorithm		3.47	100.02	13.05

# Offline A/B Experiment and Evaluation

## Generic Idea:

1. Rewrite the objective function with inverse propensity scoring.
2. Try to optimize or approximate the new objective.
3. Optimization under counterfactual setting, simulating A/B testing

## References:

- [1] Xuanhui Wang, Michael Bendersky, Donald Metzler and Marc Najork. **Learning to Rank with Selection Bias in Personal Search**. SIGIR 2016.
- [2] Thorsten Joachims, Adith Swaminathan and Tobias Schnabel. **Unbiased Learning-to-Rank with Biased Feedback**. WSDM 2017.
- [3] Thorsten Joachims and Adith Swaminathan. **Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement**. SIGIR 2016 Tutorial.
- [4] Adith Swaminathan and Thorsten Joachims. **Counterfactual risk minimization: learning from logged bandit feedback**. ICML 2015.
- [5] Lihong Li, Jinyoung Kim and Imed Zitouni. **Toward Predicting the Outcome of an A/B Experiment for Search Relevance**. WSDM 2015.
- [6] Adith Swaminathan et al. **Off-policy evaluation for slate recommendation**. NIPS 2017.
- [7] Tobias Schnabel, Adith Swaminathan, Peter Frazier and Thorsten Joachims. **Unbiased Comparative Evaluation of Ranking Functions**. ICTIR 2016.
- [8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham and Simon Dollé. **Offline A/B testing for Recommender Systems**. WSDM 2018.

# Recap: Offline A/B Experiment and Evaluation

## Summary

- Causality
- Reusable
- Need logging and an exploration strategy
- In development, emerging topic

## References:

- [1] Xuanhui Wang, Michael Bendersky, Donald Metzler and Marc Najork. **Learning to Rank with Selection Bias in Personal Search**. SIGIR 2016.
- [2] Thorsten Joachims, Adith Swaminathan and Tobias Schnabel. **Unbiased Learning-to-Rank with Biased Feedback**. WSDM 2017.
- [3] Thorsten Joachims and Adith Swaminathan. **Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement**. SIGIR 2016 Tutorial.
- [4] Adith Swaminathan and Thorsten Joachims. **Counterfactual risk minimization: learning from logged bandit feedback**. ICML 2015.
- [5] Lihong Li, Jinyoung Kim and Imed Zitouni. **Toward Predicting the Outcome of an A/B Experiment for Search Relevance**. WSDM 2015.
- [6] Adith Swaminathan et al. **Off-policy evaluation for slate recommendation**. NIPS 2017.
- [7] Tobias Schnabel, Adith Swaminathan, Peter Frazier and Thorsten Joachims. **Unbiased Comparative Evaluation of Ranking Functions**. ICTIR 2016.
- [8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham and Simon Dollé. **Offline A/B testing for Recommender Systems**. WSDM 2018.

# Semi-Manual Optimization

1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline experiments with (some) guarantees**.
3. Launch an A/B online experiment to test the realization.
4. Monitor, measure and conclude the A/B online experiment.
5. If the realization improves the metric, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1-p)$  go to Step 1.

# Semi-Manual Optimization

1. Choose a **family of hypotheses** to improve a metric.
2. Choose a realization from the family **via offline/online experiments with (some) guarantees**.
3. ~~Launch an A/B online experiment to test the realization.~~
4. ~~Monitor, measure and conclude the A/B online experiment.~~
5. ~~If the realization improves the metric, go to Step 1, otherwise with probability  $p$  go to Step 2, and probability  $(1 - p)$  go to Step 1.~~
6. Launch the realization.

# Observational Study

**Sometimes, even offline experiments may not be feasible or practical.**

# Observational Study

Sometimes, experiments may not be feasible or practical.

- **Example 1:**

We want to test which “Add to Cart” button may lead to more Monthly-Active-Users (MAUs).

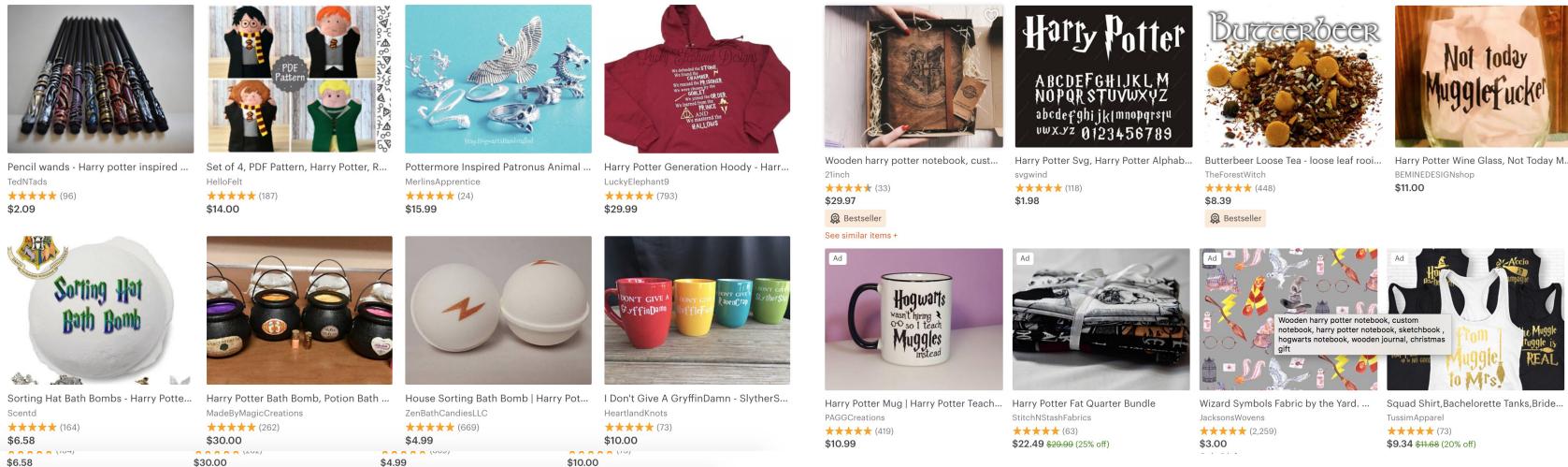
 California CA State Cutout - Large & Small - Pick Size - Laser Cut Unfinished Wood Cutout ... By CraftCutConcept...	\$0.25	<a href="#">Add to Cart</a>
 STATE BOOK CUTOUTS • Choose A Book To Make Into A Custom State Cut-Out • states • California... By AguiarDesign	\$18.99	<a href="#">Add to Cart</a>
 Customizable California State Pillow with Personalized Embroidered Patch By lovecalifornia	\$99.95	<a href="#">Add to Cart</a>
 South Carolina Cutout By SoutherlandDesi...	\$30.00	<a href="#">Add to Cart</a>
 Tactical California State Flag Patch By Patches4You	\$9.00	<a href="#">Add to Cart</a>
 Virginia State Cutout Wall Art - Repurposed Rustic Pallets & LED Lights By JoePallet	\$140.00	<a href="#">Add to Cart</a>

# Observational Study

Sometimes, experiments may not be feasible or practical.

- Example 2:

We want to test which search ranking algorithm may lead to higher Year-Over-Year Changes of user search sessions.



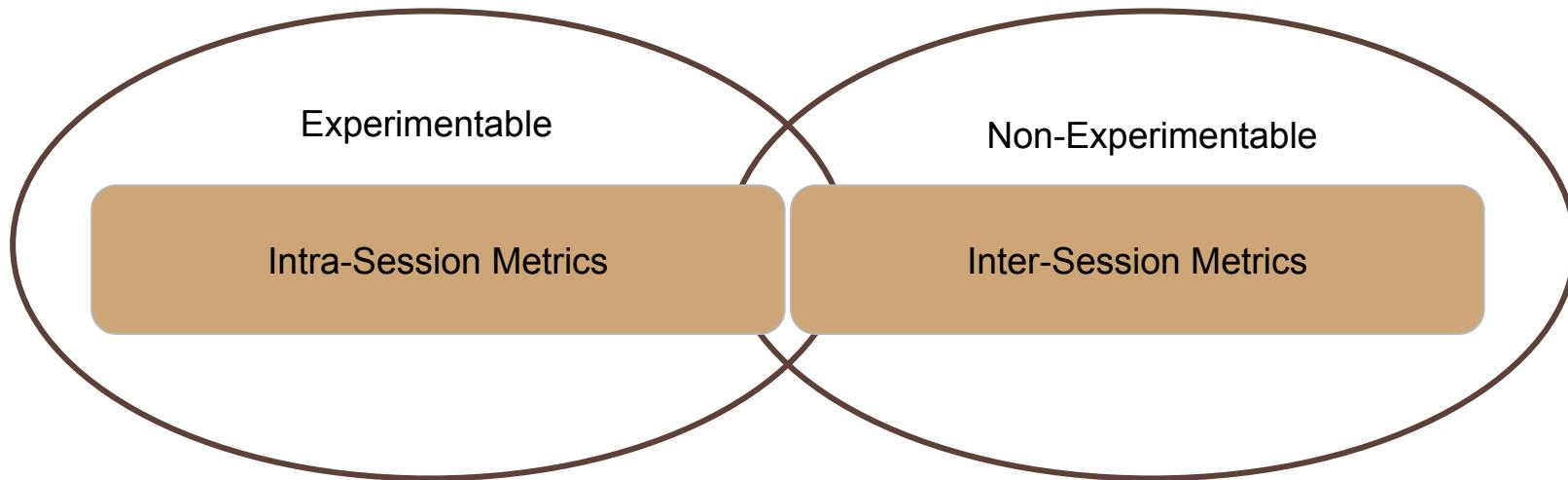
# Observational Study

**Sometimes, experiments may not be feasible or practical.**

- **Example 3:** Holiday marketing campaigns, policy changes, etc.



# Observational Study



# Recap: Observational Studies

## Statistical Relationship

- Emerging topics between statistics and machine learning
- Well grounded theory for classic cases
- Easy for simple cases
- Not well studied in a lot of online settings
- Difficult for complex scenarios
- Almost always strong assumptions

[1] David Sontag and Uri Shalit. **Causal Inference for Observational Studies**. ICML 2016 Tutorial.

[2] Lihong Li, Wei Chu, John Langford and Xuanhui Wang. **Unbiased Online Evaluation of Contextual-bandit-based News Article Recommendation Algorithms**. WSDM 2011.

[3] Lihong Li, Jin Young Kim and Imed Zitouni. **Toward Predicting the Outcome of an A/B Experiment for Search Relevance**. WSDM 2015.

# Summary: Manual and Semi-Manual Optimization

## Key Ideas

- Both manual and semi-manual optimization rely on the concept of experiments.
- There is a clear distinction between online settings and offline settings.
- Online experiments are good starting points to help decision making to improve metrics.
- Offline experiments are effective to improve the overall throughout and might avoid risks.
- The barrier between online and offline settings is non-trivial.
- There are promising results to link offline experiments and online experiments.
- Observational studies have strong assumptions.

# Metrics, Evaluation and Experiments

The relationships between metrics, evaluation and experiments

- **Requiring certain user behaviors**
  - e.g., NDCG, AUC, Precision, Recall,...

# Metrics, Evaluation and Experiments

## The relationships between metrics, evaluation and experiments

- **Requiring certain user behaviors**
  - e.g., NDCG, AUC, Precision, Recall,...
- **Decomposition assumption**
  - e.g., Conversion Rate, Click-Through-Rate,...

# Metrics, Evaluation and Experiments

## The relationships between metrics, evaluation and experiments

- **Requiring certain user behaviors**
  - e.g., NDCG, AUC, Precision, Recall,...
- **Decomposition assumption**
  - e.g., Conversion Rate, Click-Through-Rate,...
- **Naturally missing/partial data**
  - e.g., Dwell-time, View, Scroll,...

# Automatic Optimization

Online Learning

Multi-armed Bandits

Reinforcement Learning

# Automatic Optimization

1. **Choose a hypotheses** to improve a metric.
2. **Launch** a realization of the hypothesis **via Automatic Optimization techniques**.

# Automatic Optimization

1. **Choose a hypotheses** to improve a metric.
  2. **Launch** a realization of the hypothesis **via Automatic Optimization techniques**.
- 
- An offline experiment stage is optional to select better realizations.
  - An online experiment stage is optional to monitor and measure results.
  - But the key idea is to *optimize* versus *experiment*.

# Automatic Optimization

- Have a clear objective/reward/utility/loss
- Emphasize on *Maximization/Minimization*
- Three classes of Automatic Optimization techniques
  - Online Learning/Optimization
  - Multi-armed Bandit
  - Reinforcement Learning

# Online Learning

## Online Learning

```
for  $t = 1, 2, \dots$ 
    receive question  $\mathbf{x}_t \in \mathcal{X}$ 
    predict  $p_t \in D$ 
    receive true answer  $y_t \in \mathcal{Y}$ 
    suffer loss  $l(p_t, y_t)$ 
```

- The learner's ultimate goal is to minimize the cumulative loss suffered along its run.
- Theoretical analysis is around *Regret Minimization*.

# Online Learning

$$\mathbf{w}_{t+1} = \arg \min_{\mathbf{w}} \left( \mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2} \sum_{s=1}^t \sigma_s \|\mathbf{w} - \mathbf{w}_s\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 \right)$$

---

**Algorithm 1** Per-Coordinate FTRL-Proximal with  $L_1$  and  $L_2$  Regularization for Logistic Regression

---

# With per-coordinate learning rates of Eq. (2).

**Input:** parameters  $\alpha, \beta, \lambda_1, \lambda_2$

( $\forall i \in \{1, \dots, d\}$ ), initialize  $z_i = 0$  and  $n_i = 0$

**for**  $t = 1$  **to**  $T$  **do**

    Receive feature vector  $\mathbf{x}_t$  and let  $I = \{i \mid x_i \neq 0\}$   
     For  $i \in I$  compute

$$w_{t,i} = \begin{cases} 0 & \text{if } |z_i| \leq \lambda_1 \\ -\left(\frac{\beta + \sqrt{n_i}}{\alpha} + \lambda_2\right)^{-1}(z_i - \text{sgn}(z_i)\lambda_1) & \text{otherwise.} \end{cases}$$

Predict  $p_t = \sigma(\mathbf{x}_t \cdot \mathbf{w})$  using the  $w_{t,i}$  computed above

Observe label  $y_t \in \{0, 1\}$

**for** all  $i \in I$  **do**

$g_i = (p_t - y_t)x_i$  #gradient of loss w.r.t.  $w_i$

$\sigma_i = \frac{1}{\alpha} \left( \sqrt{n_i + g_i^2} - \sqrt{n_i} \right)$  #equals  $\frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}$

$z_i \leftarrow z_i + g_i - \sigma_i w_{t,i}$

$n_i \leftarrow n_i + g_i^2$

**end for**

**end for**

---

# Recap: Online Learning

## Online Learning

- Easy to understand and implement.
- Do not have a notion of multiple competing hypotheses
- In general, do not know how good/bad

[1] Elad Hazan. **Introduction to Online Convex Optimization**. Foundations and Trends® in Optimization: Vol. 2: No. 3-4, 2016.

[2] Shai Shalev-Shwartz. **Online Learning and Online Convex Optimization**. Foundations and Trends® in Machine Learning: Vol. 4: No. 2, 2012.

# Multi-armed Bandits

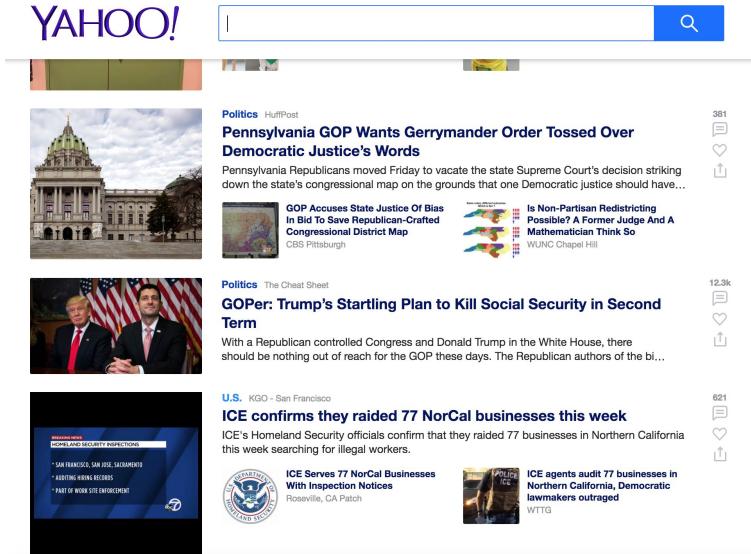
Formally, we define by  $\mathcal{A} = \{1, 2, \dots, K\}$  a set of  $K$  arms, and a contextual-bandit algorithm  $A$  interacts with the *world* in discrete trials  $t = 1, 2, 3, \dots$ . In trial  $t$ :

1. The world chooses a feature vector  $\mathbf{x}_t$  known as the *context*. Associated with each arm  $a$  is a real-valued reward  $r_{t,a} \in [0, 1]$  that can be related to the context  $\mathbf{x}_t$  in an arbitrary way. We denote by  $\mathcal{X}$  the (possibly infinite) set of contexts, and  $(r_{t,1}, \dots, r_{t,K})$  the reward vector. Furthermore, we assume  $(\mathbf{x}_t, r_{t,1}, \dots, r_{t,K})$  is drawn i.i.d. from some unknown distribution  $D$ .
2. Based on observed rewards in previous trials and the current context  $\mathbf{x}_t$ ,  $A$  chooses an arm  $a_t \in \mathcal{A}$ , and receives reward  $r_{t,a_t}$ . It is important to emphasize here that *no* feedback information (namely, the reward  $r_{t,a}$ ) is observed for *unchosen* arms  $a \neq a_t$ .
3. The algorithm then improves its arm-selection strategy with all information it observes,  $(\mathbf{x}_{t,a_t}, a_t, r_{t,a_t})$ .

- The learner's ultimate goal is to maximize the cumulative reward along its run.
- Theoretical analysis is around *Regret Minimization*.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics



Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. **Returning is Believing: Optimizing Long-term User Engagement in Recommender Systems.** In CIKM 2017. ACM, New York, NY, USA, 1927-1936.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

- Most algorithms focus on intra-session effects (e.g., clicks, dwell, etc.).

[1] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. **Google news personalization: scalable online collaborative filtering.** In WWW 2007. ACM, New York, NY, USA, 271-280.

[2] Yehuda Koren, Robert Bell and Chris Volinsky. **Matrix Factorization Techniques for Recommender Systems.** Computer 42(8):2009.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

- Most algorithms focus on intra-session effects (e.g., clicks, dwell, etc.).

[1] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. **Google news personalization: scalable online collaborative filtering**. In WWW 2007. ACM, New York, NY, USA, 271-280.

[2] Yehuda Koren, Robert Bell, and Chris Volinsky. **Matrix Factorization Techniques for Recommender Systems**. Computer 42(8):2009.

- Users may leave because of boredom from popular items.

Komal Kapoor, Karthik Subbian, Jaideep Srivastava, and Paul Schrater. **Just in Time Recommendations: Modeling the Dynamics of Boredom in Activity Streams**. In WSDM 2015. ACM, New York, NY, USA, 233-242.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

- Users may have high immediate rewards but *accumulate linear regret* after they leave.
- Predict a user's immediate reward, but also project it onto *future clicks*, making recommendation decisions dependent over time.
- Rapid change of environment requires this kind of decisions *online*.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

Some more related work about *modeling users' post-click behaviors*:

- [1] Nicola Barbieri, Fabrizio Silvestri and Mounia Lalmas. **Improving Post-Click User Engagement on Native Ads via Survival Analysis.** WWW 2016.
- [2] Mounia Lalmas, Jane.e Lehmann, Guy Shaked, Fabrizio Silvestri and Gabriele Tolomei. **Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users.** KDD Industry Track 2015.
- [3] Nan Du, Yichen Wang, Niao He, Jimeng Sun and Le Song. **Time-Sensitive Recommendation From Recurrent User Activities.** NIPS 2015.
- [4] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava and Tao Ye. **A Hazard Based Approach to User Return Time Prediction.** KDD 2014.

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**Balance between**

- 1. Maximize immediate reward of the recommendation**

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**Balance between**

1. **Maximize immediate reward of the recommendation**
2. **Explore other possibilities to improve model estimation.**

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Balance between

1. Maximize immediate reward of the recommendation
2. Explore other possibilities to improve model estimation.
3. Maximize expected future reward by keeping users in the system.

To maximize *the cumulative reward* over time, the system has to **make users click more** and **return more often**.

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**Main Idea**

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Main Idea

- **Model how likely an item would yield an immediate click:**  
[1] At iteration  $i$ , if we recommend item  $a$ , how likely it is going to be clicked by user  $u$ .

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Main Idea

- **Model how likely an item would yield an immediate click:**  
[1] At iteration  $i$ , if we recommend item  $a$ , how likely it is going to be clicked by user  $u$ .
- **Model future visits after seeing this item and their expected clicks:**  
[2] At iteration  $i+1$ , what do we recommend.  
[3] How that decision would impact the click behavior at  $i+1$   
[4] Future return probability at  $i+2$ , and  
So on...

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Main Idea

- **Model how likely an item would yield an immediate click:**  
[1] At iteration  $i$ , if we recommend item  $a$ , how likely it is going to be clicked by user  $u$ .
- **Model future visits after seeing this item and their expected clicks:**  
[2] At iteration  $i+1$ , what do we recommend.  
[3] How that decision would impact the click behavior at  $i+1$   
[4] Future return probability at  $i+2$ , and  
So on...

**Can be formulated in a reinforcement learning setting.**

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**A Major Challenge:**

future candidate pool undefined, thus **standard reinforcement learning** can't apply.

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**A Major Challenge:**

future candidate pool undefined, thus **standard reinforcement learning** can't apply.

**Need approximations.**

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**Approximations**

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

## **Approximations**

1. Future clicks depend on users. (Strong? or not)

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

## Approximations

1. Future clicks depend on users. (Strong? or not)
2. Only model finite steps in future, or even just one step ahead.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Approximations

1. Future clicks depend on users. (Strong? or not)
2. Only model finite steps in future, or even just one step ahead.
3. Only model whether the user return in a finite horizon.

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

## Approximations

1. Future clicks depend on users. (Strong? or not)
2. Only model finite steps in future, or even just one step ahead.
3. Only model whether the user return in a finite horizon.

**New Objective:**  $P(C_{u,i} = 1 | a_i) + \epsilon_u P(\Delta_{u,i} \leq \tau | a_i)$

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Model Summary

1. Use **Generalized Linear Model (Bernoulli)** to model how likely a user  $u$  would click on an item  $a_i$  at iteration  $i$ .

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Model Summary

1. Use **Generalized Linear Model (Bernoulli)** to model how likely a user  $u$  would click on an item  $a_i$  at iteration  $i$ .
2. Use **Moving Average** to model a user  $u$ 's marginal click probability.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Model Summary

1. Use **Generalized Linear Model (Bernoulli)** to model how likely a user  $u$  would click on an item  $a_i$  at iteration  $i$ .
2. Use **Moving Average** to model a user  $u$ 's marginal click probability.
3. Use **Generalized Linear Model (Exponential)** to model a user  $u$ 's return time intervals.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Model Summary

1. Use **Generalized Linear Model (Bernoulli)** to model how likely a user  $u$  would click on an item  $a_i$  at iteration  $i$ .
2. Use **Moving Average** to model a user  $u$ 's marginal click probability.
3. Use **Generalized Linear Model (Exponential)** to model a user  $u$ 's return time intervals.
4. Use **Upper Confidence Bound (UCB)** on top of [1-3].

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Model Summary

1. Use **Generalized Linear Model (Bernoulli)** to model how likely a user  $u$  would click on an item  $a_i$  at iteration  $i$ .
2. Use **Moving Average** to model a user  $u$ 's marginal click probability.
3. Use **Generalized Linear Model (Exponential)** to model a user  $u$ 's return time intervals.
4. Use **Upper Confidence Bound (UCB)** on top of [1-3].

Note that both [1] and [3]'s coefficients are personalized.

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

---

**Algorithm 1**  $r^2$ Bandit

---

- 1: **Inputs:**  $\eta > 0, \tau > 0, \delta_1 \in (0, 1)$
- 2: **for**  $i = 1$  to  $N$  **do**
- 3:     Receive user  $u$
- 4:     Record current timestamp  $t_{u,i}$
- 5:     **if** user  $u$  is new: **then**
- 6:         Set  $\mathbf{A}_{u,1} \leftarrow \eta \mathbf{I}, \hat{\boldsymbol{\theta}}_{u,1} \leftarrow \mathbf{0}^d, \hat{\boldsymbol{\beta}}_{u,1} \leftarrow \mathbf{0}^d, \hat{\epsilon}_{u,1} \sim U(0, 1);$
- 7:     **else**:
- 8:         Compute return interval  $\Delta_{u,i-1} = t_{u,i} - t_{u,i-1}$
- 9:         Update  $\hat{\boldsymbol{\beta}}_{u,i}$  in user return model using MLE.
- 10:      **end if**
- 11:     Observe context vectors,  $\mathbf{x}_a \in \mathbb{R}^d$  for  $\forall a \in I(t_{u,i})$
- 12:     Make recommendation  $a_{u,i} = \arg \max_{a \in I(t_{u,i})} P(C_{u,i} = 1 | \mathbf{x}_a, \hat{\boldsymbol{\theta}}_{u,i}) + \hat{\epsilon}_{u,i} P(\Delta_{u,i} \leq \tau | \mathbf{x}_a, \hat{\boldsymbol{\beta}}_{u,i}) + \alpha_{u,i} \|\mathbf{x}_a\|_{\mathbf{A}_{u,i}^{-1}}$
- 13:     Observe click  $C_{u,i}$
- 14:      $\mathbf{A}_{u,i+1} \leftarrow \mathbf{A}_{u,i} + \mathbf{x}_{a_{u,i}} \mathbf{x}_{a_{u,i}}^\top$
- 15:     Update  $\hat{\boldsymbol{\theta}}_{u,i+1}$  in user click model using MLE.
- 16:     Update  $\hat{\epsilon}_{u,i+1} = \sum_{j \leq i} C_{u,j} / i$
- 17: **end for**

---

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

---

**Algorithm 1**  $r^2$ Bandit

---

- 1: **Inputs:**  $\eta > 0, \tau > 0, \delta_1 \in (0, 1)$
- 2: **for**  $i = 1$  to  $N$  **do**
- 3:     Receive user  $u$
- 4:     Record current timestamp  $t_{u,i}$
- 5:     **if** user  $u$  is new: **then**
- 6:         Set  $\mathbf{A}_{u,1} \leftarrow \eta \mathbf{I}, \hat{\boldsymbol{\theta}}_{u,1} \leftarrow \mathbf{0}^d, \hat{\boldsymbol{\beta}}_{u,1} \leftarrow \mathbf{0}^d, \hat{\epsilon}_{u,1} \sim U(0, 1);$
- 7:     **else**:
- 8:         Compute return interval  $\Delta_{u,i-1} = t_{u,i} - t_{u,i-1}$
- 9:         Update  $\hat{\boldsymbol{\beta}}_{u,i}$  in user return model using MLE.
- 10:      **end if**
- 11:     Observe context vectors,  $\mathbf{x}_a \in \mathbb{R}^d$  for  $\forall a \in I(t_{u,i})$
- 12:     Make recommendation  $a_{u,i} = \arg \max_{a \in I(t_{u,i})} P(C_{u,i} = 1 | \mathbf{x}_a, \hat{\boldsymbol{\theta}}_{u,i}) + \hat{\epsilon}_{u,i} P(\Delta_{u,i} \leq \tau | \mathbf{x}_a, \hat{\boldsymbol{\beta}}_{u,i}) + \alpha_{u,i} \|\mathbf{x}_a\|_{\mathbf{A}_{u,i}^{-1}}$
- 13:     Observe click  $C_{u,i}$
- 14:      $\mathbf{A}_{u,i+1} \leftarrow \mathbf{A}_{u,i} + \mathbf{x}_{a_{u,i}} \mathbf{x}_{a_{u,i}}^\top$
- 15:     Update  $\hat{\boldsymbol{\theta}}_{u,i+1}$  in user click model using MLE.
- 16:     Update  $\hat{\epsilon}_{u,i+1} = \sum_{j \leq i} C_{u,j} / i$
- 17: **end for**

---

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

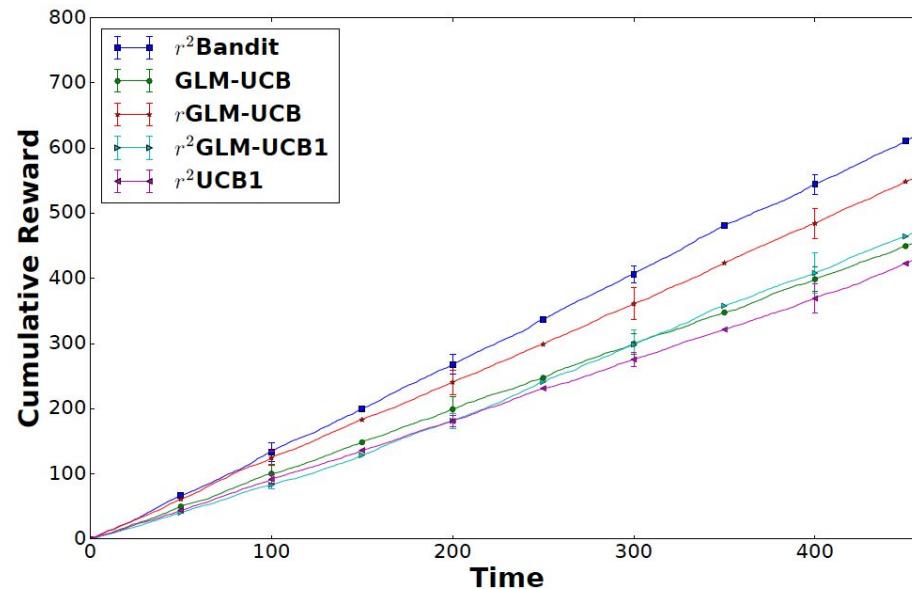
### Simulations

1. **Type 1:** items with **high** click probability but **short** expected return time;
2. **Type 2:** items with **high** click probability but **long** expected return time;
3. **Type 3:** items with **low** click probability but **short** expected return time;
4. **Type 4:** items with **low** click probability and **long** expected return time.

# Multi-armed Bandits

How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

## Simulations

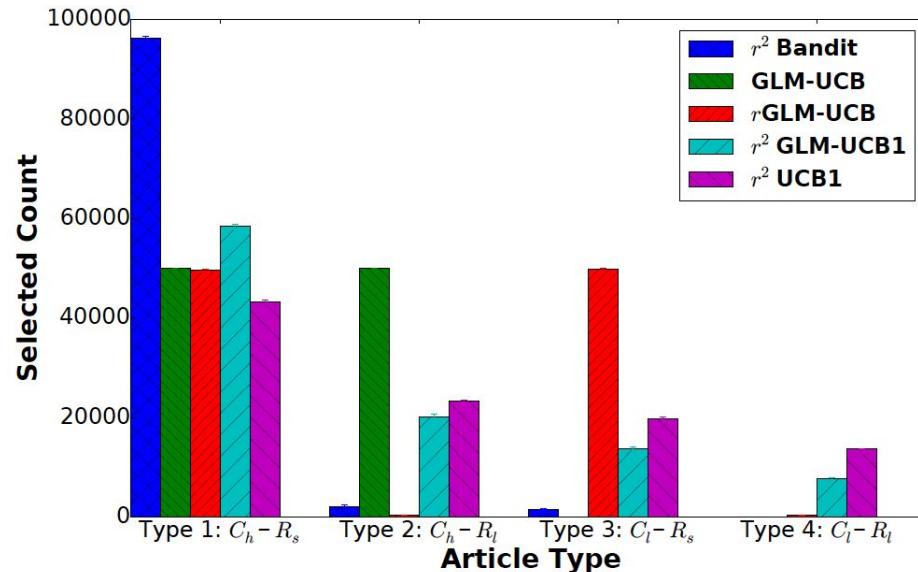


(a) Cumulative clicks over time

# Multi-armed Bandits

How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

## Simulations

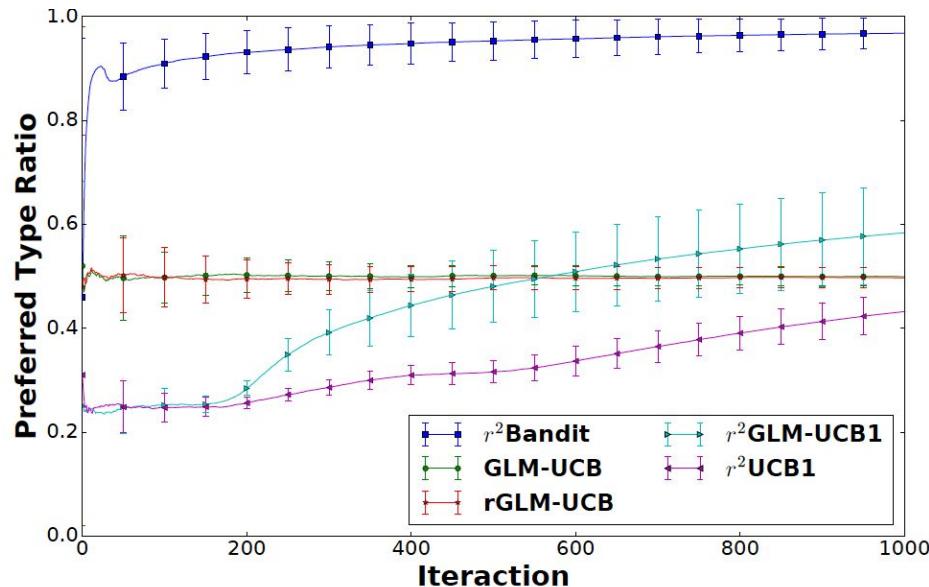


(b) Distribution of selected item types

# Multi-armed Bandits

How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

## Simulations



(c) Evolution of preferred item type ratio

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

### Real-World Dataset

- Collect 4 weeks of data from Yahoo news portal.
- Reduce features into 23 by PCA.
- Sessionized the data by 30 mins.
- Return time is computed by time interval between two sessions.
- Total:
  - 18,882 users,
  - 188,384 articles
  - 9,984,879 logged events, and
  - 1,123,583 sessions.

# Multi-armed Bandits

How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

Real-World Dataset

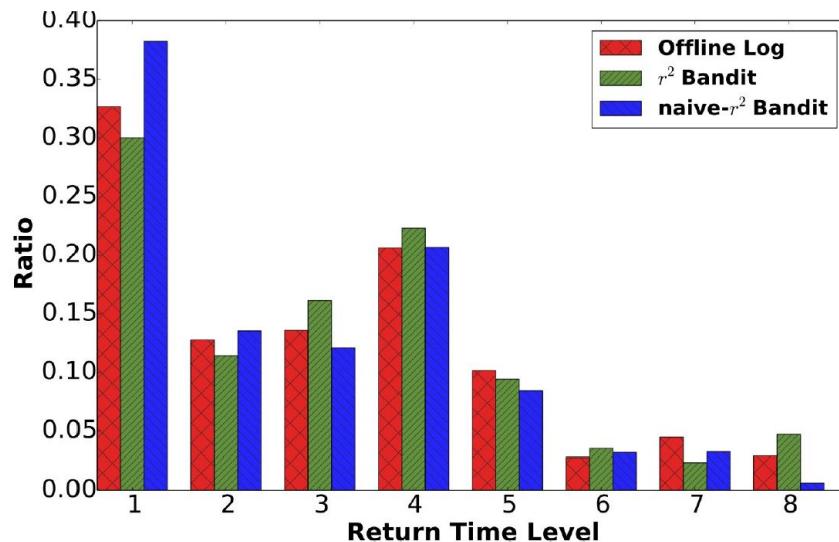


Figure 2: Discretized user return time distribution.

# Multi-armed Bandits

**How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics**

**Real-World Dataset: Evaluation**

- Cumulative clicks over Time
- Click-through Rate (CTR)
- Average Return Time
- Return Rate
- Improved User Ratio
- No return Count

# Multi-armed Bandits

## How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

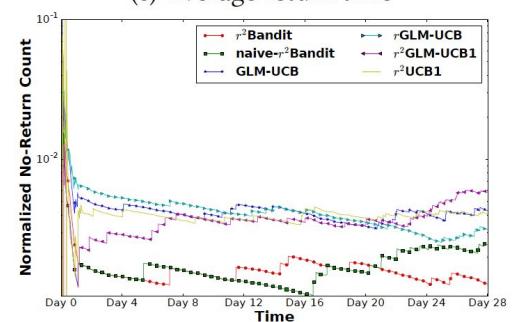
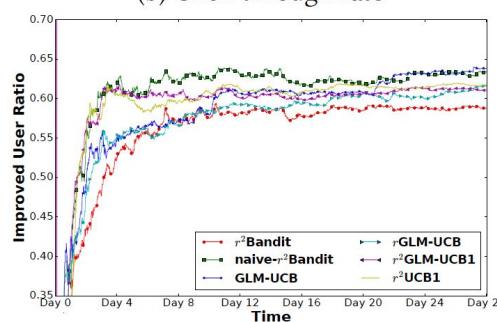
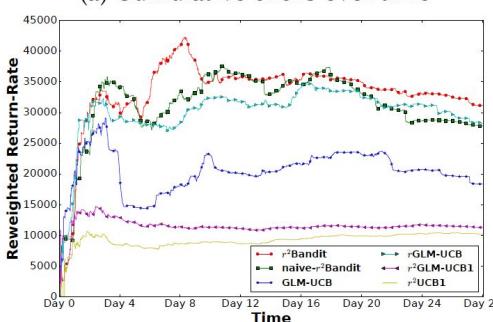
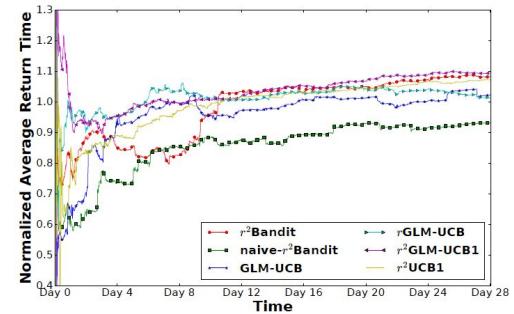
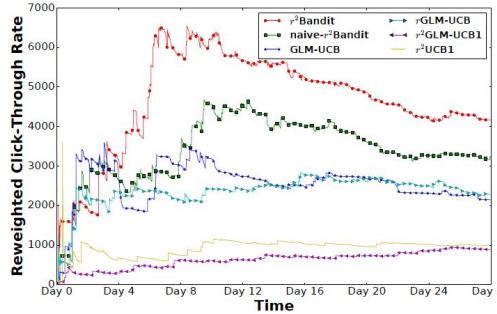
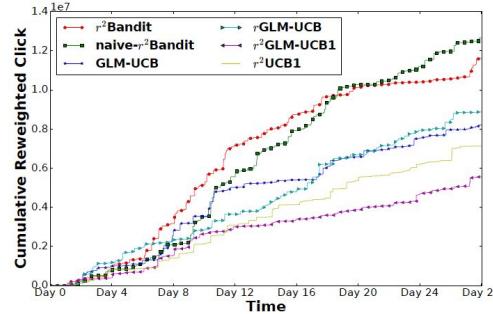
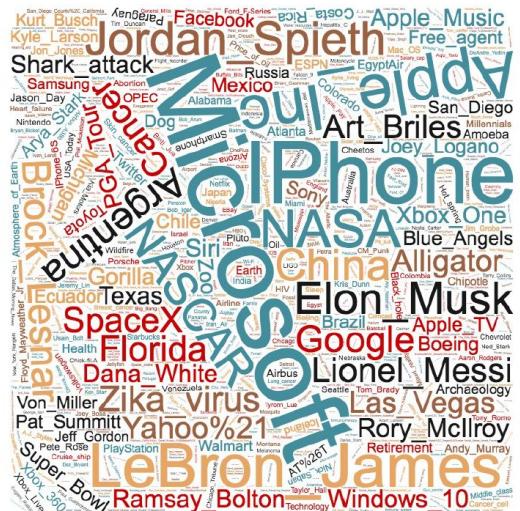


Figure 3: Experiment results on real-world news recommendation log data.

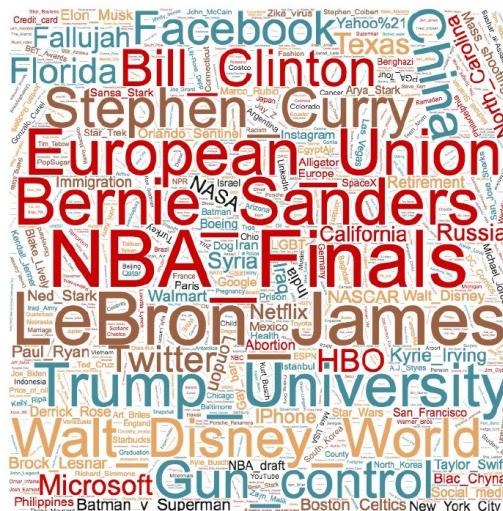
# Multi-armed Bandits

How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

Real-World Dataset: Word Cloud



(a) Top clicked articles



(b) Top returning articles

Figure 4: Word cloud of algorithm selected article content.

# Recap: Multi-armed Bandits

## Multi-armed Bandits

- Easy to understand and implement.
- Challenge to scale to millions/billions.
- In general, do not know how good/bad

[1] Lihong Li, Wei Chu, John Langford and Robert Schapire. **A contextual Bandit Approach to Personalized News Article Recommendation.** WWW 2010.

[2] Lihong Li, Wei Chu, John Langford and Xuanhui Wang. **Unbiased Online Evaluation of Contextual-bandit-based News Article Recommendation Algorithms.** WSDM 2011.

# Reinforcement Learning

A Markov decision process is a **4-tuple**  $(S, A, P_a, R_a)$ , where

- $S$  is a finite set of states,
- $A$  is a finite set of actions (alternatively,  $A_s$  is the finite set of actions available from state  $s$ ),
- $P_a(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$  is the probability that action  $a$  in state  $s$  at time  $t$  will lead to state  $s'$  at time  $t + 1$ ,
- $R_a(s, s')$  is the immediate reward (or expected immediate reward) received after transitioning from state  $s$  to state  $s'$ , due to action  $a$

The goal is to choose a policy  $\pi$  that will maximize some cumulative function of the random rewards, typically the expected discounted sum over a potentially infinite horizon:

$$\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1}) \quad (\text{where we choose } a_t = \pi(s_t), \text{i.e. actions given by the policy})$$

where  $\gamma$  is the discount factor and satisfies  $0 \leq \gamma \leq 1$ . (For example,  $\gamma = 1/(1+r)$  when the discount rate is  $r$ .)  $\gamma$  is typically close to 1.

# Reinforcement Learning

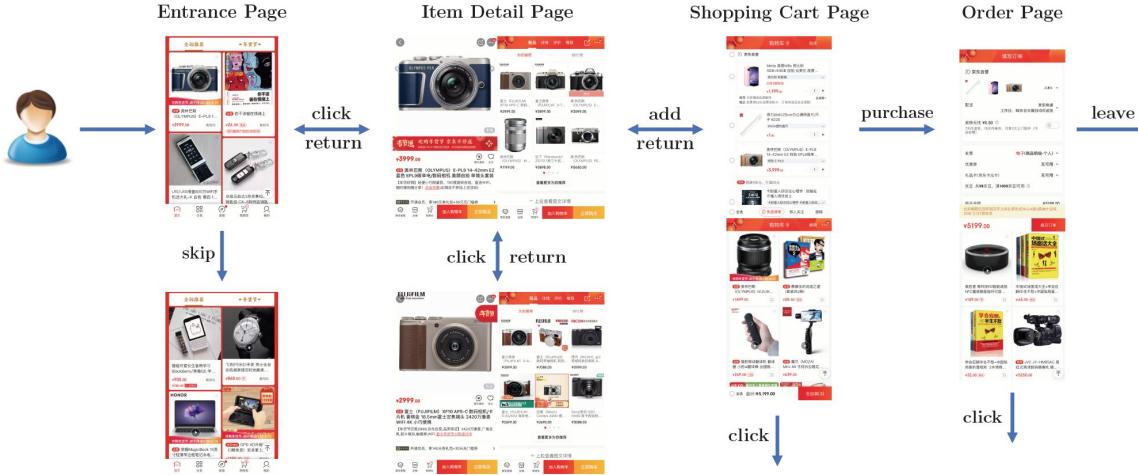


Figure 1: An example of whole-chain recommendations.

## Early Attempts:

[1] Xiangyu Zhao, Long Xia, Yihong Zhao, Dawei Yin and Jiliang Tang. **Model-Based Reinforcement Learning for Whole-Chain Recommendations**. CoRRabs/1902.03987, 2019.

[2] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. **Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems**. KDD 2019.

# Recap: Reinforcement Learning

## Reinforcement Learning

- Intuitive to understand and difficult to implement.
- Challenge to scale to millions/billions.
- In general, do not know how good/bad an algorithm is.

[1] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin and Jiliang Tang. **Deep Reinforcement Learning for Page-wise Recommendations**. RecSys 2018.

[2] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang and Dawei Yin. **Recommendations with Negative Feedback via Pairwise Deep Reinforcement Learning**. KDD 2018.

[3] Di Wu, Xiujun Chen, Xun Yang, Hao Wang, Qing Tan, Xiaoxun Zhang, Jian Xu and Kun Gai. **Budget Constrained Bidding by Model-free Reinforcement Learning in Display Advertising**. CIKM 2018.

[4] Xiangyu Zhao, Long Xia, Yihong Zhao, Dawei Yin and Jiliang Tang. **Model-Based Reinforcement Learning for Whole-Chain Recommendations**. CoRRabs/1902.03987, 2019.

[5] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. **Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems**. KDD 2019.

# Summary: Automatic Optimization

## Key Ideas

- Automatic optimization is to optimize an objective function or a metric *versus* to experiment and compare two ideas.
- Automatic optimization is typically an iterative learning process.
- There is no clear way to understand how good or how bad an idea is in general.
- There is a clear distinction between online settings and offline settings.
- The barrier between online and offline settings is non-trivial.

# Combining Two Camps

# Two Main Camps of Optimization

- **Manual and Semi-Manual Optimization**
  - e.g. The classic Hypothesis-Experiment-Evaluation Cycle
- **Automatic Optimization**
  - e.g., Online Learning, Multi-armed Bandits, Reinforcement Learning...

# Two Main Camps of Optimization

- **Manual and Semi-Manual Optimization**

Pros: Have deep roots in Statistics, Economics and etc

Cons: Concerning with ATE (or similar) and slow & costly to operate

- **Automatic Optimization**

Pros: Have deep roots in ML, Control and etc.

Cons: Concerning with maximizing/minimizing rewards/loss

## Combining Two Camps

Can we maximize/minimize rewards while concerning ATE?

# Combining Two Camps

## Two Challenges for Standard A/B Testing:

- **Time Cost**  
Product evolution pushes its shareholders to consistently monitor results from online A/B experiments, which usually invites peeking and altering experimental designs as data collected.
- **Opportunity Cost**  
A static test usually entails a static allocation of users into different variants, which prevents an immediate roll-out of the better version to larger audience or risks of alienating users who may suffer from a bad experience.

# Combining Two Camps

## Contributions:

1. Propose an imputed sequential Girshick test for Bernoulli model with a fixed allocation.
2. Use simulations to demonstrate that the test procedure also applies to an adaptive allocation such as Thompson sampling with a small error inflation.
3. Conduct a regret analysis of A/B tests from the Multi-armed Bandit (MAB) perspective.
4. Conduct extensive studies including simulations as well as experiments on an industry-scale experiment, demonstrating the effectiveness of the proposed method and offering practical considerations.

Nianqiao Ju, Diane Hu, Adam Henderson and Liangjie Hong. **A Sequential Test for Selecting the Better Variant: Online A/B testing, Adaptive Allocation, and Continuous Monitoring**. WSDM 2019.

# Combining Two Camps

Sequential analysis [2] studies experiments where the number of observations required is not determined in advance and at each stage of the experiment a decision is made to accept some hypothesis, reject it, or take more observations.

Setup:  $X \sim f_\theta(\cdot)$  where  $\theta \in \Theta \subset \mathbb{R}$  and with two simple hypotheses  $H_0 : \theta = \theta_0$  and  $H_1 : \theta = \theta_1$  (assuming  $\theta_0 < \theta_1$  without loss of generality).

Based on our risk tolerance  $\delta$ , we choose some number  $AB$  according to desired Type-I error and Power of the test. Then at each stage of the experiment, the **Sequential Probability Ratio Test** compute the probability ratio

$$\frac{p_{1m}}{p_{0m}} = \frac{f_{\theta_1}(x_{1:m})}{f_{\theta_0}(x_{1:m})}.$$

We continue the experiment and take more observations if  $B < \frac{p_{1m}}{p_{0m}} < A$ ; if  $\frac{p_{1m}}{p_{0m}} > A$ , then the process terminates with a decision to reject  $H_0$ ; and if  $\frac{p_{1m}}{p_{0m}} < B$  then we terminate with acceptance of  $H_0$ .

# Combining Two Camps

**Girshick's Double Dichotomy Test** goes as follows: fix some  $\delta > 0$  and at time  $t$ , we would have  $t$  pairs of data and the log likelihood ratio is

$$Z_t = \log \left( \frac{p_{1t}}{p_{0t}} \right) = \underbrace{-\delta}_{\text{risk tolerance}} \times \overbrace{t}^{\text{sample size}} \times \underbrace{(\bar{Y}_t - \bar{X}_t)}_{\text{difference in empirical averages}}.$$

In real experiments, we cannot observe both  $x_t$  and  $y_t$  because a customer is either in control group or in treatment group with fixed probability  $\rho$  and  $1 - \rho$ . To this end we design an **imputed Girshik Test** with the imputed log likelihood ratio test statistic

$$\widehat{Z}_t = \widehat{\log} \left( \frac{p_{1t}}{p_{0t}} \right) = \underbrace{-\delta}_{\text{risk tolerance}} \times \overbrace{\frac{2mn}{t}}^{\text{effective sample size}} \times \underbrace{(\bar{Y}_n - \bar{X}_m)}_{\text{difference in empirical averages}}$$

Note that in this case is still unbiasedly estimating the average treatment effect.

# Combining Two Camps

## Imputed Girshin Test for Adaptive Allocation

To address opportunity cost of experiments even further, we use Thompson sampling [1] for an adaptive allocation of customers, which results in a time-varying  $\rho_t$ . As data is collected, the posterior distribution  $p_1, p_2$  is sequentially updated. After  $t$  data points  $D_{1:t}$  are collected, the next customer is assigned to group 1 based on the probability of the 1st group being the optimal one, given the current data, calculated from the posterior distribution of rewards through

$$\mathbb{P}(p_1 > p_2 | X_{1:t}) = \int \mathbb{I}(p_1 \geq p_2) \pi(p_1, p_2 | D_{1:t}) dp_1 dp_2.$$

Because of stopping time concerns, we use the geometric mean  $\sqrt{mn}$  as the effective pair size for Thompson Sampling. To approximate the treatment effect, we would still use the empirical average, although this estimator is consistent but no longer unbiased.

$$\widetilde{Z}_t = \log \widetilde{\left( \frac{p_{1,t}}{p_{0,t}} \right)} = (-\delta) \times \underbrace{\sqrt{mn}}_{\text{effective sample size}} \times (\bar{Y}_n - \bar{X}_m).$$

# Combining Two Camps

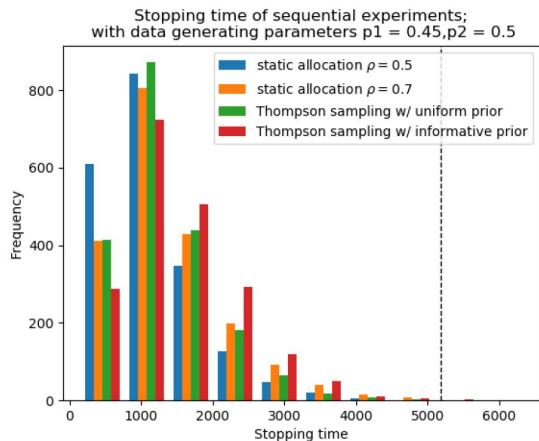


Figure 4: A histogram of stopping times for the imputed sequential Girshick test using different allocation schemes, corresponding to Table 1. The dashed black line is the sample size required by a fixed-time proportion test. There is a vanishingly small number of simulations where the sequential test requires more samples than the fixed-time proportion test.

	static allocation		Thompson sampling	
	$\rho = 0.5$	$\rho = 0.7$	Unif. priors	inform. priors
$\mathbb{P}(\text{accept} \omega_a)$	99.8 %	99.75%	97.7%	99.55%
average $\tau$	1165.26	1383.86	1300.47	1537.59
min	186	148	263	235
median $\tau$	1024	1194	1140	1376
max	5622	6214	4952	6329

**Table 1: Comparison of number of observations required by the imputed Girshick test using different allocation schemes. For the same set up  $p_1 = 0.45, p_2 = 0.5, \alpha = 0.05, \beta = 0.05$ , a fixed-time two-sample proportion test needs 2589.479 observations in each group.**

# Recap: Combining Two Camps

- Sequential Test from Statistics + Multi-armed Bandit from ML
- Challenges:
  - Biased v.s. Unbiased
  - Deriving valid p-values
  - Provide practical benefits
- Emerging Topics

- [1] Alex Deng. **Objective bayesian two sample hypothesis testing for online controlled experiments.** WWW 2015.
- [2] Alex Deng, Jiannan Lu and Shouyuan Chen. **Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing.** DSAA 2016.
- [3] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. **Peeking at A/B Tests: Why It Matters, and What to Do About It.** KDD 2017.
- [4] Steven L Scott. **Multi-armed bandit experiments in the online service economy.** Applied Stochastic Models in Business and Industry 31, 1:2015.
- [5] Minyong R Lee and Milan Shen. **Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments.** KDD 2018.
- [6] Aaditya Ramdas. **Foundations of Large-Scale Sequential Experimentation.** In KDD 2019.



# Concluding remarks and future direction

# Metrics: Concluding Remarks

## Main takeaways:

Three levels of engagement, involvement, interaction and contribution, where contribution better predictor of high engagement: challenge is to make a user **becoming** a contributor.

Not all clicks are equal: look at the **value** of a click and relate it to **downstream engagement**, e.g. spend time, purchase, organize, as this often leads to better measurement of engagement.

Understand the relationship between intra-session and inter-session metrics: ensure that optimizing for the former **drives** business metrics in the right direction.

How users engage is **one** (big) part of the engagement lifecycle: don't forget to measure point of engagement (beginning) and disengagement (end).

# Metrics: Concluding Remarks

## Opportunities:

- How to systematically discover new metrics, through for example the quantification of users' holistic feelings or by learning them.
- How to use mixed methods to elicit hypotheses of what engagement means and inspire metric development.
- How to consider non engagement metrics (e.g diversity, discovery, revenue) when measuring online engagement.
- How to ensure we have the right set of guardrail metrics?

# Metrics: Concluding Remarks

## **Challenges:**

How to account for bias when measuring and optimizing for given metrics.

How to account for intent, segmentation and diversity.

How to incorporate negative signals.

# Optimizations: Concluding Remarks

## **Opportunities:**

Emerging topics of utilizing and combining techniques, methodologies and ideas from Machine Learning, Statistics, Economics, Control Theory and more fields.

# Optimizations: Concluding Remarks

## Opportunities:

Emerging topics of utilizing and combining techniques, methodologies and ideas from Machine Learning, Statistics, Economics, Control Theory and more fields.

## Challenges:

- Still early stage, a lot of heuristics, require more active research
- Costly to practice and involve institution commitments
- Offline and online barriers still exist
- Optimizing for multiple (*possibly competing*) metrics
- Optimize under *FATE* (Fairness, Accountability, Transparency, and Ethics)

# Optimizations: Concluding Remarks

## Opportunities:

Reinforcement learning has shown promising results in many AI domains. However, it is still in early stage to apply to long-term metric optimization.

## Challenges:

- Hard to scale to millions and billions of users and items
- Early results focusing on better predictions comparing to classic methods
- Difficult to simulate real-world applications
- What metrics to optimize



# Thank you

Website:<https://onlineuserengagement.github.io/>