Tutorial on Online User Engagement:
Metrics and Optimization

Liangjie Hong & Mounia Lalmas

THE WEB CONFERENCE

Outline

Introduction and Scope

Metrics

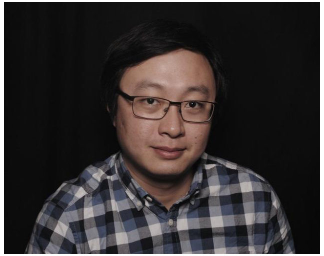
Optimisation

Concluding Remarks & Future Directions

Who we are

- 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 audio
  - Website: https://mounia-lalmas.blog/

- Liangjie Hong, Director of Engineering - Data Science and Machine Learning at Etsy, New York City
  - Research interests: search, recommendation, advertising and now hand-craft goods
  - Website: https://www.hongliangjie.com/



Acknowledgements

This tutorial is based on "**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.

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

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

Appeal to user curiosity; encourages 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

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

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?

Quality of the user experience … endurability

People remember "satisfactory" experiences and want to repeat them

Endurability

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

[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.

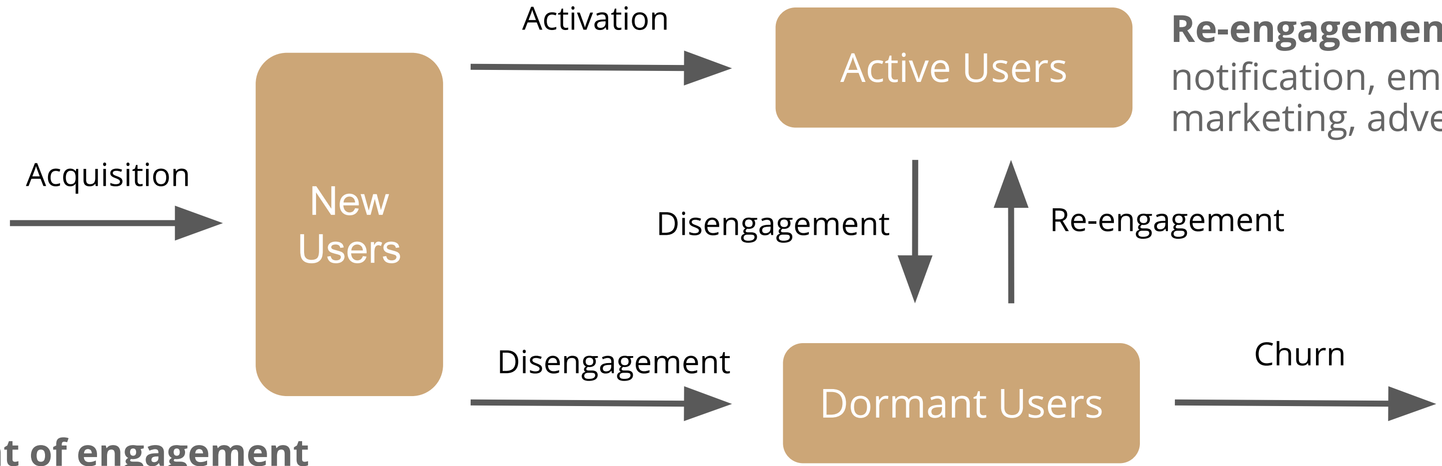
utilitarian factors (e.g. usability) → hedonic and experiential factors of interaction (e.g. fun, fulfillment) → user engagement

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

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
Disengagement	Loss of interests lead 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

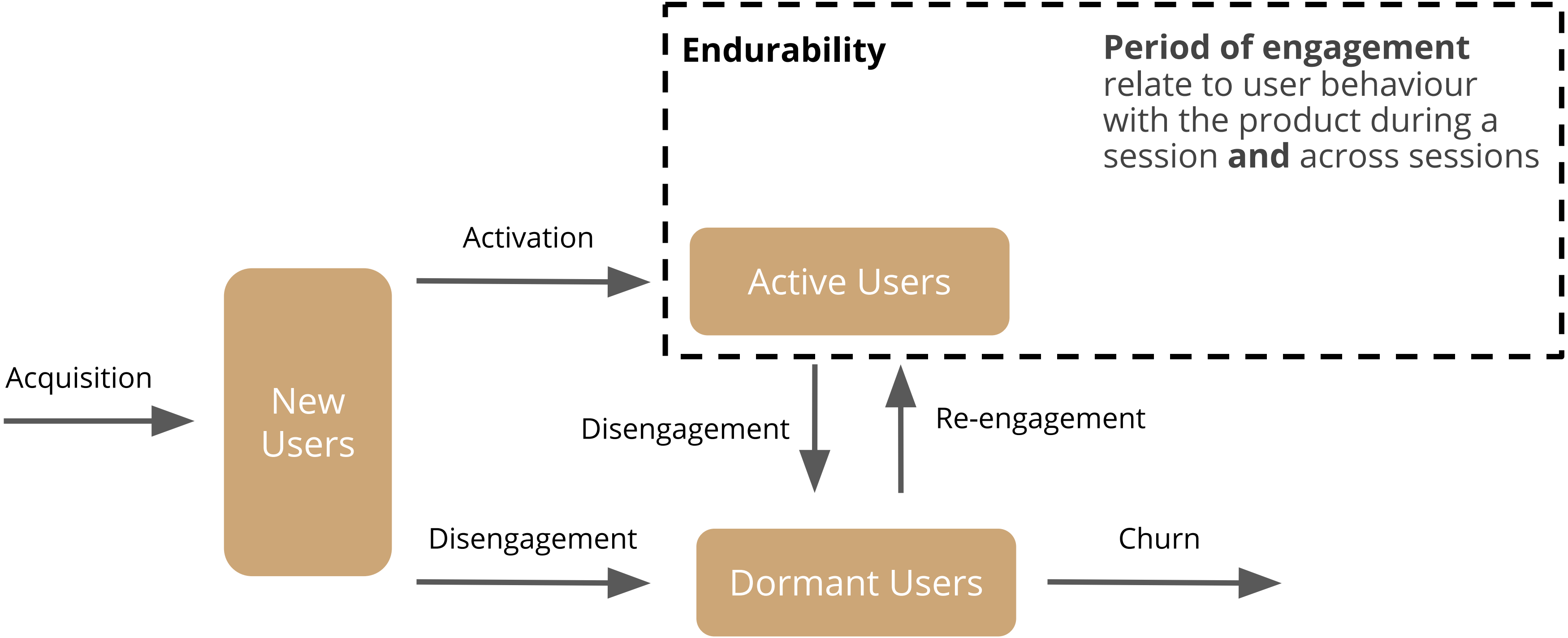


Re-engagement
notification, email, offer,
marketing, advertising

Point of engagement
how users arrive
acquisition costs

Disengagement
churn & retention

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

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

Methods to measuring user engagement

self-reported methods
subjective

Questionnaire, interview, report, product reaction cards

User study (lab/online)

mostly qualitative

physiological measurements
objective

- Task-based methods
- Neurological
- Physiological

User study (lab/online)

mostly quantitative, scalability an issue

online analytics
objective

Data science
Machine learning

Data study (online)

quantitative large scale

Scope of this tutorial

Focus on online analytics → online user engagement.

Assume that applications are “properly designed”.

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

Search

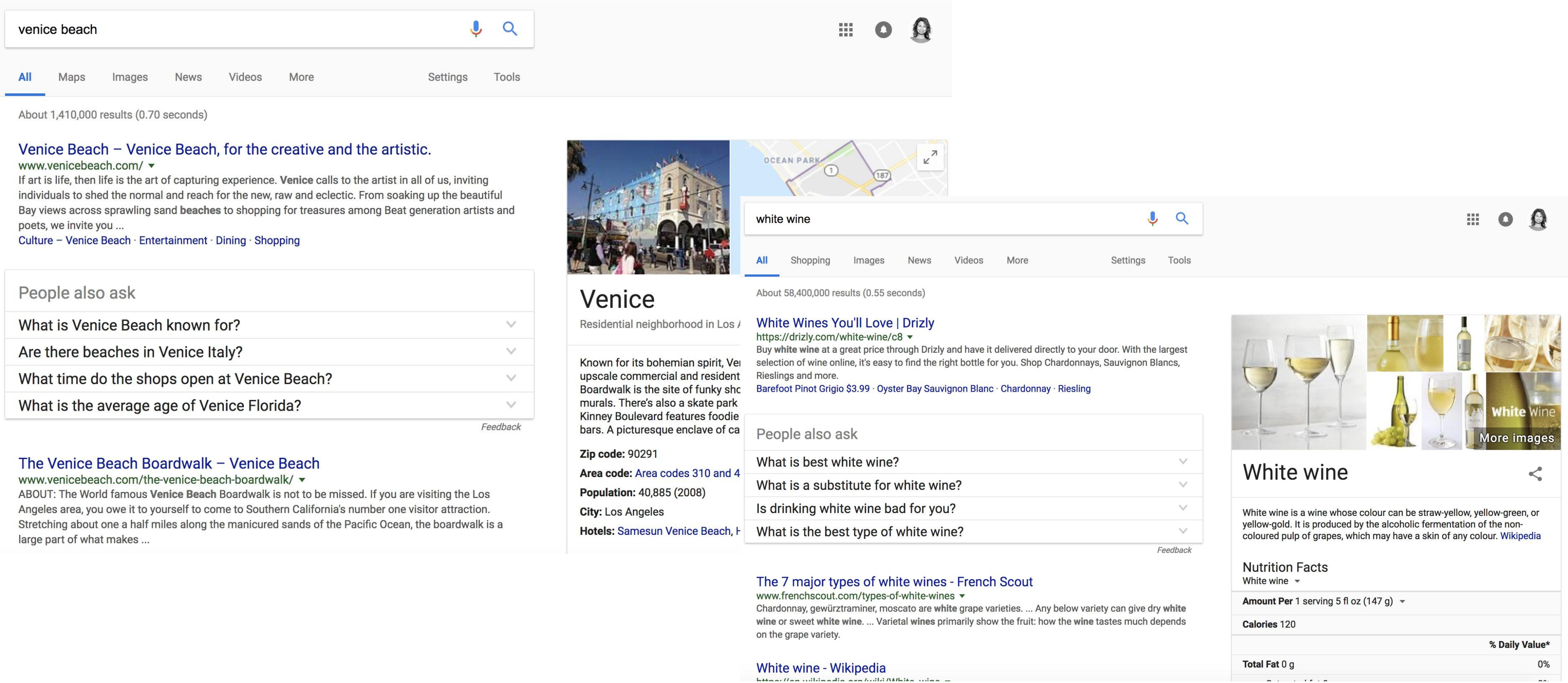
News

E-commerce

Entertainment

Advertising

Search


venice beach
All
Maps
Images
News
Videos
More
Settings
Tools
About 1,410,000 results (0.70 seconds)
Venice Beach – Venice Beach, for the creative and the artistic.
www.venicebeach.com/
If art is life, then life is the art of capturing experience. Venice calls to the artist in all of us, inviting individuals to shed the normal and reach for the new, raw and eclectic. From soaking up the beautiful Bay views across sprawling sand beaches to shopping for treasures among Beat generation artists and poets, we invite you ...
Culture – Venice Beach · Entertainment · Dining · Shopping
People also ask
What is Venice Beach known for?
Are there beaches in Venice Italy?
What time do the shops open at Venice Beach?
What is the average age of Venice Florida?
Feedback
The Venice Beach Boardwalk – Venice Beach
www.venicebeach.com/the-venice-beach-boardwalk/
ABOUT: The World famous Venice Beach Boardwalk is not to be missed. If you are visiting the Los Angeles area, you owe it to yourself to come to Southern California's number one visitor attraction. Stretching about one a half miles along the manicured sands of the Pacific Ocean, the boardwalk is a large part of what makes ...
OCEAN PARK
Venice
Residential neighborhood in Los
Known for its bohemian spirit, Ve
upscale commercial and resident
Boardwalk is the site of funky sho
murals. There's also a skate park
Kinney Boulevard features foodie
bars. A picturesque enclave of ca
Zip code: 90291
Area code: Area codes 310 and 4
Population: 40,885 (2008)
City: Los Angeles
Hotels: Samesun Venice Beach, H
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
People also ask
What is best white wine?
What is a substitute for white wine?
Is drinking white wine bad for you?
What is the best type of white wine?
Feedback
The 7 major types of white wines - French Scout
www.frenchscout.com/types-of-white-wines
Chardonnay, gewürztraminer, moscato are white grape varieties. ... Any below variety can give dry white wine or sweet white wine. ... Varietal wines primarily show the fruit: how the wine tastes much depends on the grape variety.
White wine - Wikipedia
White Wine
More images
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%

Search

Search engine evaluation

- Coverage
- Speed
- Query language
- User interface

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

ENGLISH ESPAÑOL 中文

SUBSCRIBE NOW LOG IN

The New York Times

Tuesday, May 7, 2019

Today's Paper

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

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

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

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

S&P 500	-1.65%
Dow	-1.79%
Nasdaq	-1.96%

60°F 70° 53° New York, NY

TRUMP'S TAXES

Decade in the Red: Trump Tax Figures Show Over $1 Billion in Losses

- Donald J. Trump was propelled 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, obtained by The Times paints a far bleaker picture of his financial condition. Read our exclusive report.

2h ago

Donald J. Trump in 1986, his career marked by acqu
Ted Thai/The LIFE Picture C

Here are five takeaways of what the numbers show.

3h ago

Mr. Trump's state tax

Opinion ›

Google's Sundar Pichai: Privacy Should Not Be a Luxury Good

Yes, we use data to make products more helpful for

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.
- Just weeks ago, the school joined others in the Denver area in closing over security concerns as the 20th anniversary of the Columbine shooting neared.

15m ago

YAHOO!

1 student dead in Colorado school shooting

Man's racist Facebook comment lands him in trouble

Brendan Fraser on his tragic experience filming 'The...

Disney's big announcement after moving studios

Congres hamme

Celebrity Women's Health

Um, People Aren't Sure Where Kim Kardashian's Internal Are In Her Met Gala Look

Is a corset that tight even safe?

To make things clear
1.This dress is corseted BUT
2.Kim trains her ass off 6 days a fucking week, she wakes up early af and is dedicated.
3.I paved the road for her but SHE did the work!

Met Gala 2019: Jared Leto carries a replica of his own head as an accessory
Yahoo Style UK

Kendall Jenner and He Harry Styles Had a Mo Gala
Elle

THE WALL STREET JOURNAL.

Subscribe | Sign In

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

What's News

Stocks Sink as Trade Concerns Intensify

The stock market's declines deepened, with the Dow sliding more than 450 points, as investors braced for the increased likelihood the U.S. will raise tariffs on Chinese goods later this week.

436

- Some See Buying Opportunity in Rare Dip

China Agrees to Resume U.S. Trade Negotiations

China is sending its top trade envoy to Washington to resume negotiations and confront U.S. demands that Beijing detail the laws it would change as a part of a trade deal.

- U.S. Consumers Face Hit in Trade Fight

U.S. Lifts Sanctions on Venezuelan General Who Broke With Maduro

The Trump administration lifted sanctions on a Venezuelan general who

Occidental CEO Battles Oil-Field Giant to Rule the Permian Basin

Vicki Hollub goes all-in to best mighty Chevron for the prize of Anadarko, seeking to bulk up in a region that is the epicenter of U.S. shale production.

- Anadarko Says Occidental's Offer 'Superior' to Chevron
- Rivals Vie for Mastery Over America's Hottest Oil Field

Watchdog Probes FBI Reliance on Dossier in Surveillance of Trump Aide

The Justice Department's watchdog, close to concluding its inquiry into steps the FBI took in its

Disney Reveals Movie Lineup Through 2027

DAVID PIERCE

Pixel 3a: Google's $400 Challenge to the $1,000 Phone

Markets

U.S. EUROPE ASIA FX RATES FUTURES

1D 5D 3M 6M 1Y

DJIA	25965.09	-473.39	-1.79%
S&P 500	2884.05	-48.42	-1.65%
Nasdaq	7963.76	-159.53	-1.96%
Russell 2000	1582.31	-32.66	-2.02%
DJ Total Mkt	29734.82	-511.59	-1.69%

May 7 '19, 5:10 PM EDT MARKETS →

Opinion

Motive Matters in Trump Spygate
By Holman W. Jenkins, Jr. | Business World

The Pseudo-Impeachment
By The Editorial Board | Review & Outlook

In Praise of Great Professors
Future View

Liberty Mutual.

News

ENGLISH ESPAÑOL 中文

SUBSCRIBE NOW LOG IN

The New York Times

Tuesday, May 7, 2019

Today's Paper

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

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

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

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

S&P 500 -1.65% ↓
Dow -1.79% ↓
Nasdaq -1.96% ↓

60°F 70° 53°
New York, NY

TRUMP'S TAXES

Decade in the Red: Trump Tax Figures Show Over $1 Billion in Losses

- Donald J. Trump was propelled 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, obtained by The Times paints a far bleaker picture of his financial condition. Read our exclusive report.

2h ago

Donald J. Trump in 1986, during a tumultuous period of his career marked by acquisition and collapse. Ted Thai/The LIFE Picture Collection, via Getty Images

Here are five takeaways of what the numbers show.
3h ago

Mr. Trump's state tax returns could be released under a bill being advanced by New York lawmakers.
18m 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.
- Just weeks ago, the school joined others in the Denver area in closing over security concerns as the 20th anniversary of the Columbine shooting neared.

29m ago

Mahesh Anandan picked up his three children after a shooting at STEM School Highlands Ranch near Denver, where eight students were shot, one fatally, on Tuesday. Nick Cote for The New York Times

TRUMP ADMINISTRATION

Clash Between Trump and House Democrats Threatens Constitutional Order

- Earlier administrations had isolated skirmishes 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

Trump Administration Can Keep Sending Asylum Seekers to Mexico, Court Rules
It was an unusual victory for the Trump administration in a liberal-leaning court.
43m ago

The secretary of state made an unexpected trip to Iraq a day after the U.S. warned of new threats from Iran.
2h 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

In a cellphone video recorded by Sandra Bland, Trooper Brian Encinia could be seen drawing his stun gun and heard saying, "I will light you up. Get out. Now." Investigative

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
Whether he "looks" black or not, I'll be glad to claim him.
6h ago 70 comments

Neil Eggleston and Joshua A. Geltzer
The Court Handling Trump's Lawsuit Must Move at Breakneck Speed

Carl Levin
Trump Is Defying Congress. Will They Hold Him in Contempt?

Thomas L. Friedman
How to Defeat Trump

Jesse Moss
The Abandoned Vans of Atlanta

Veronique de Rugy
Corporate Welfare Wins Again in Trump's Washington

Ross Douthat
The One-Income Trap

The Editorial Board
Sorry, Steve Mnuchin. Congress Has a Right to See Trump's Tax Returns.

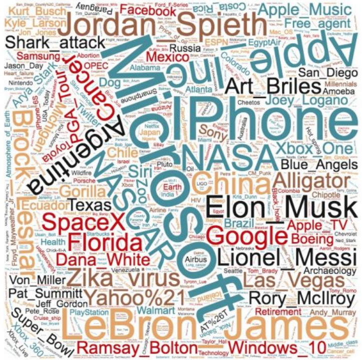
Charlie Warzel
We Are Drowning in Data

Mark Gevisser
Why I Am Voting for South Africa's Ramaphosa

Editors' Picks

A Wine-Soaked Ball Unites a Georgian Village, but Only After Dividing It

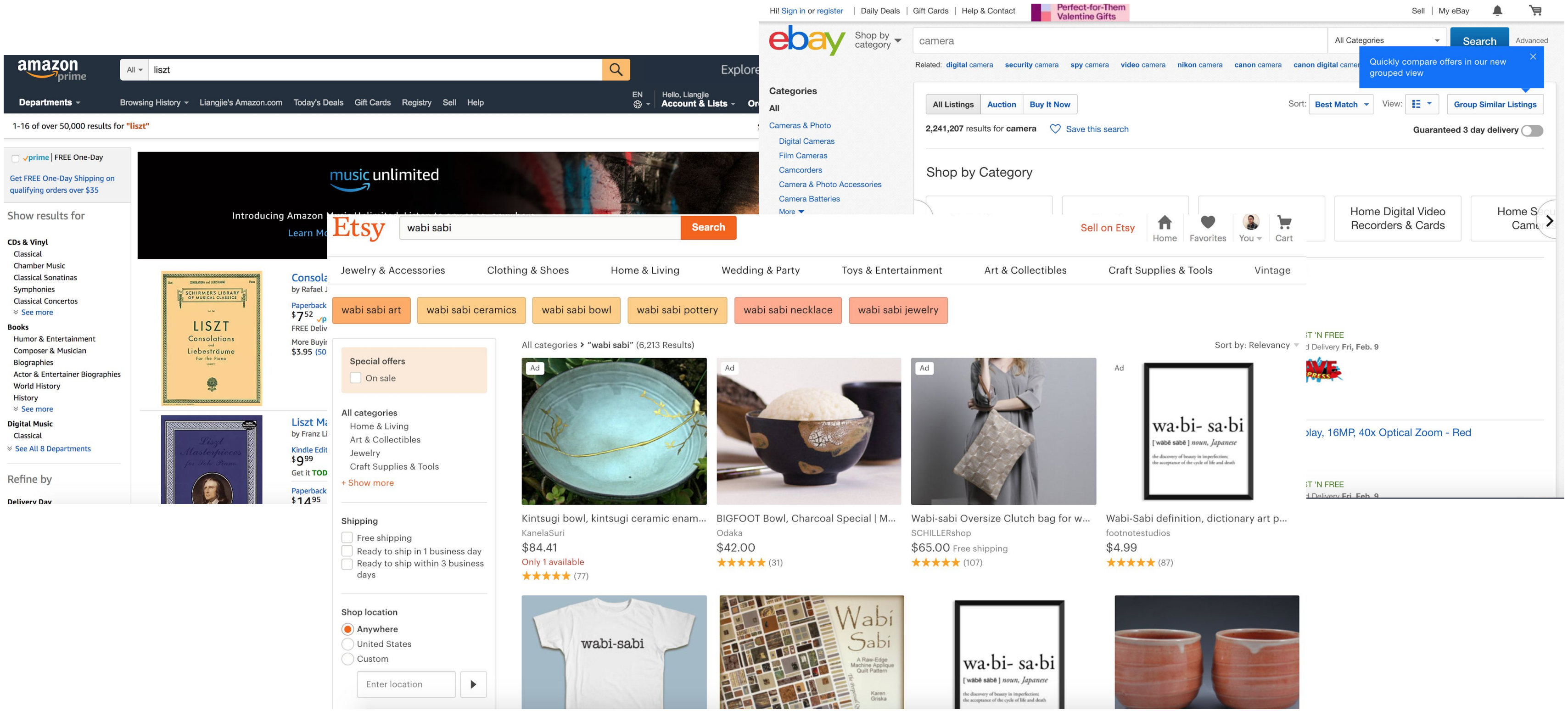
Every spring in Shukhuti, a black leather ball is sewn together to play Lelo Burti, a brutally physical folk game. The winners carry it



(a) Top clicked articles

(b) Top returning articles

E-Commerce



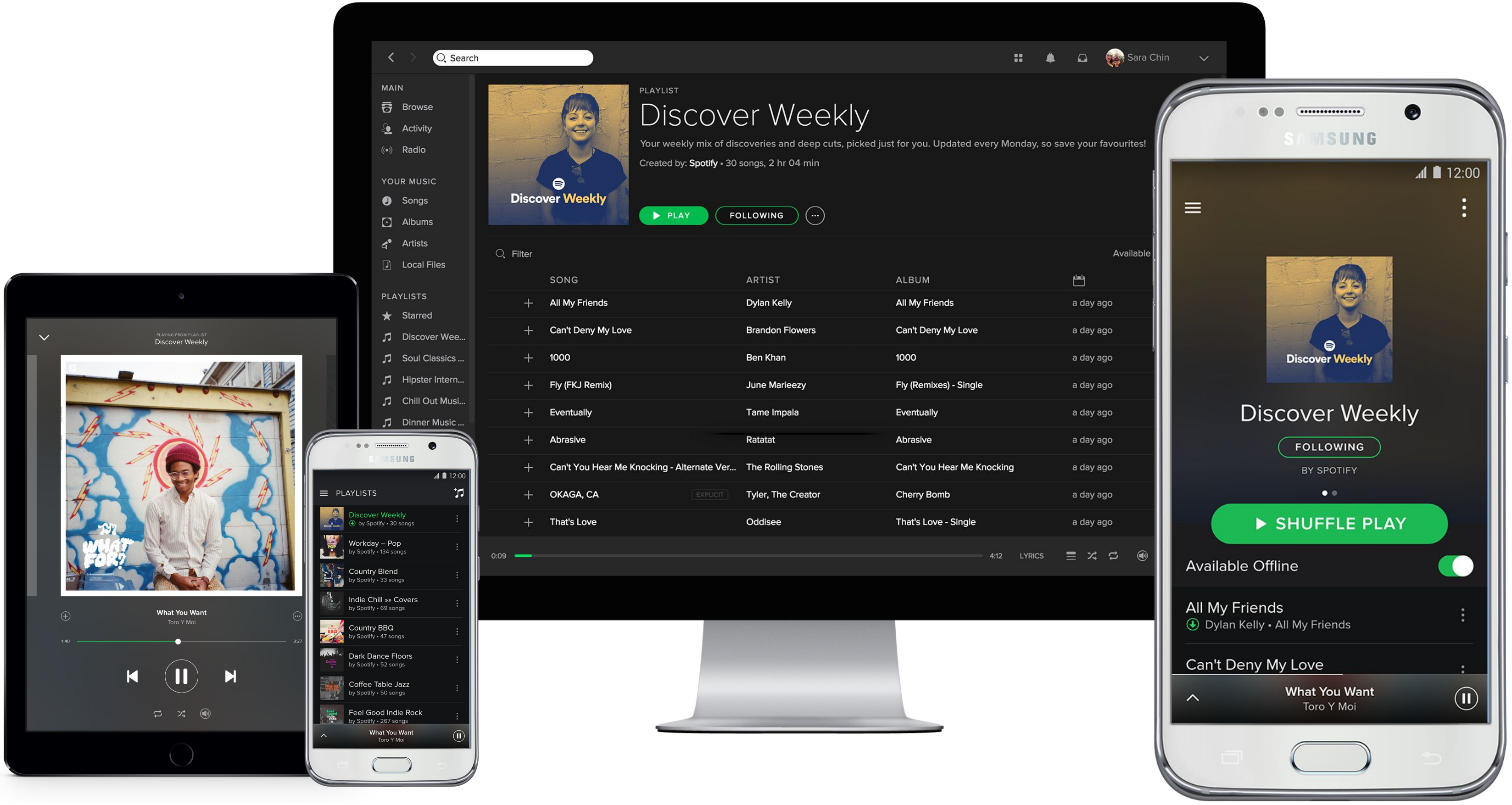
amazon prime
All
liszt
Explore
Departments
Browsing History
Liangjie's Amazon.com
Today's Deals
Gift Cards
Registry
Sell
Help
EN
Hello, Liangjie
Account & Lists
1-16 of over 50,000 results for "liszt"
prime | FREE One-Day
Get FREE One-Day Shipping on qualifying orders over $35
music unlimited
Show results for
CDs & Vinyl
Classical
Chamber Music
Classical Sonatinas
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
LISZT
Consolations
Liebesträume
Hi! Sign in or register
Daily Deals
Gift Cards
Help & Contact
Perfect-for-Them Valentine Gifts
Sell
My eBay
ebay
Shop by category
camera
All Categories
Search
Advanced
Related: digital camera security camera spy camera video camera nikon camera canon camera
Quickly compare offers in our new grouped view
Categories
All
Cameras & Photo
Digital Cameras
Film Cameras
Camcorders
Camera & Photo Accessories
Camera Batteries
More
All Listings
Auction
Buy It Now
Sort:
Best Match
View:
Group Similar Listings
2,241,207 results for camera
Save this search
Guaranteed 3 day delivery
Shop by Category
Home Digital Video Recorders & Cards
Etsy
wabi sabi
Search
Sell on Etsy
Home
Favorites
You
Cart
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
All categories › "wabi sabi" (6,213 Results)
Sort by: Relevancy
Special offers
On sale
All categories
Home & Living
Art & Collectibles
Jewelry
Craft Supplies & Tools
+ Show more
Shipping
Free shipping
Ready to ship in 1 business day
Ready to ship within 3 business days
Shop location
Anywhere
United States
Custom
Enter location
Kintsugi bowl, kintsugi ceramic enam...
KanelaSuri
$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)
wa·bi- sa·bi
wabi-sabi
Wabi Sabi

E-Commerce

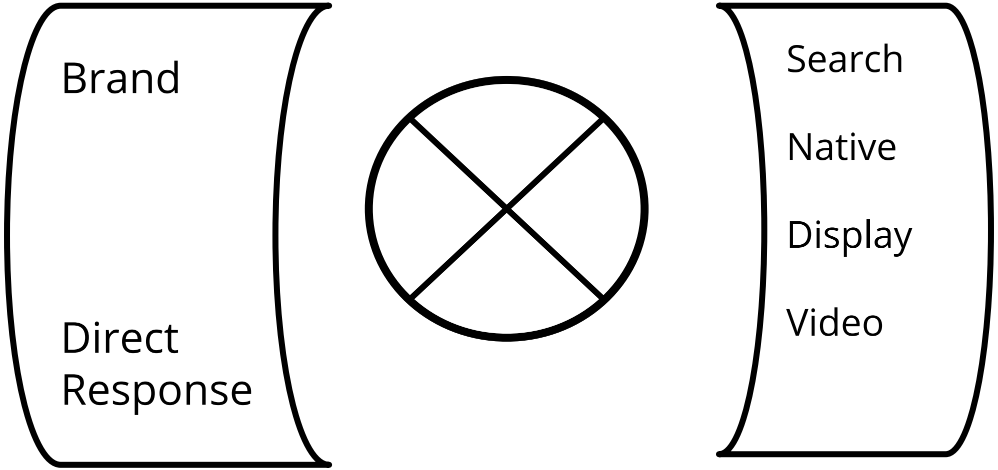
Entertainment

Spotify
iTunes
MUSIC
Google play
DEEZER
TIDAL
SHAZAM
amazon music
anghami
beatport PRO
7digital
Rhapsody
emusic
junodownload
247

Entertainment



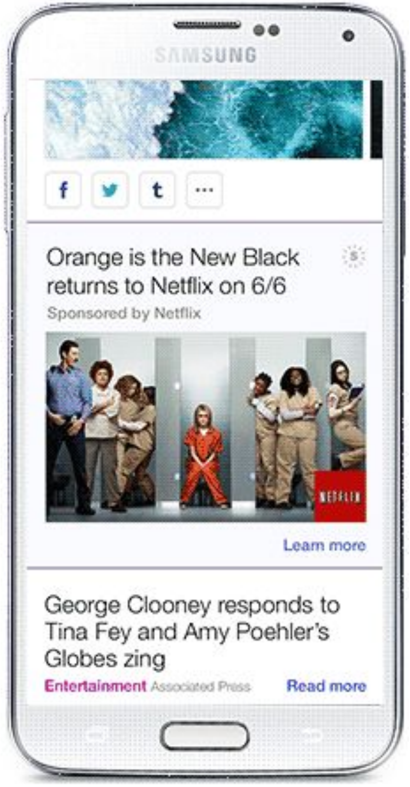
Advertising



DEMAND
(advertisers)

SUPPLY
(publishers)

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

https://www.klipfolio.com/blog/kpi-metric-measure

Three levels of metrics

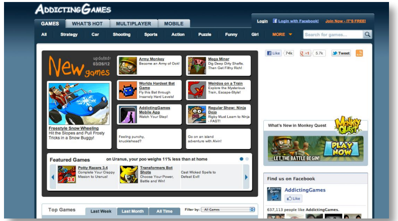
Business metrics -- KPIs

Behavioral metrics -- online metrics, analytics

Optimisation metrics -- metrics used to train machine learning algorithms

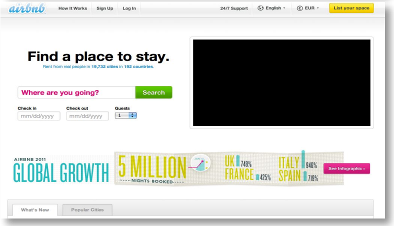
These three levels are connected

Why several metrics?

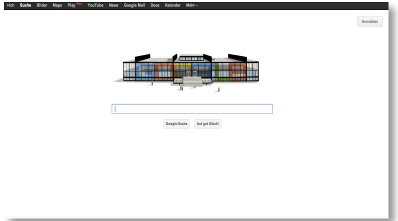


Games
Users spend much time per visit

Social media
Users come frequently & stay long

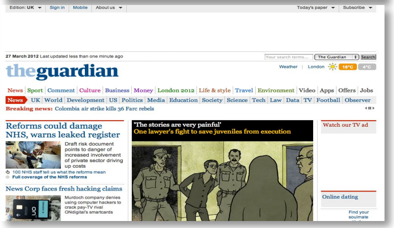


Service
Users visit site, when needed



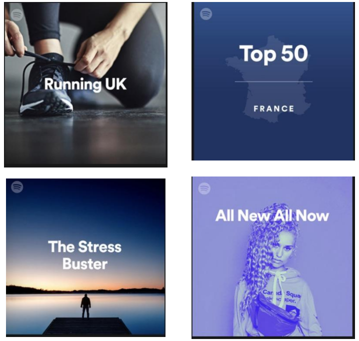
Search
Users come frequently but do not stay long

Niche
Users come on average once a week



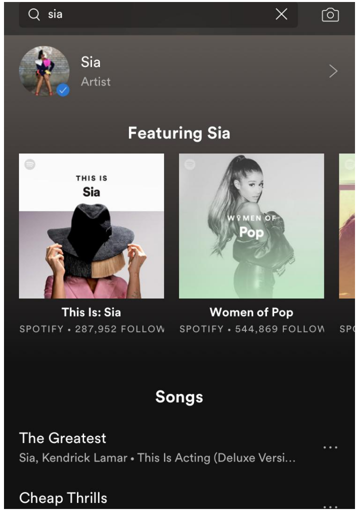
News
Users come periodically, e.g. morning and evening

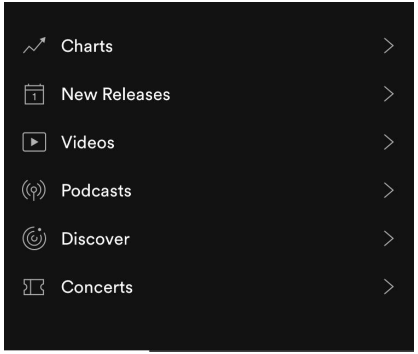
Why several metrics?



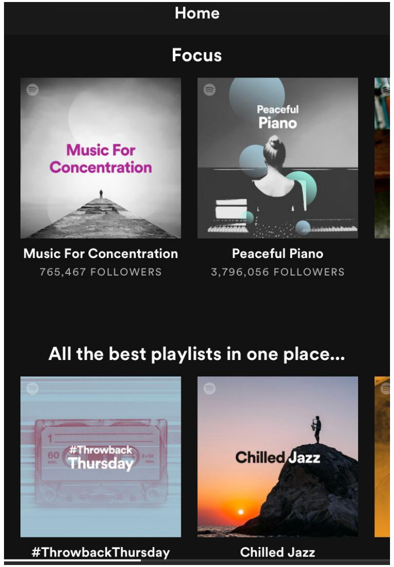
Playlists differ in their listening patterns.

Searching has a particular engagement pattern.



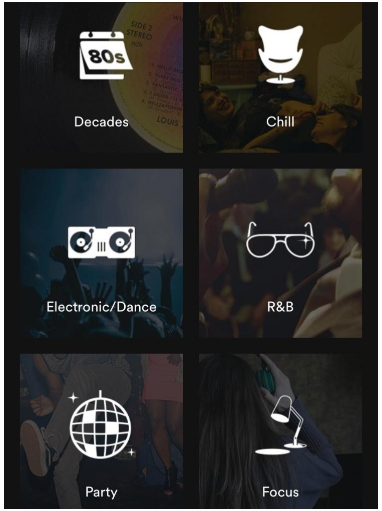


Media type and freshness lead to different engagement patterns.

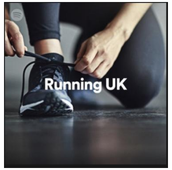


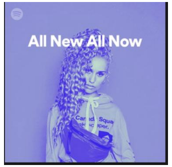
Home can be viewed as a hub with a “star” style engagement pattern.

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



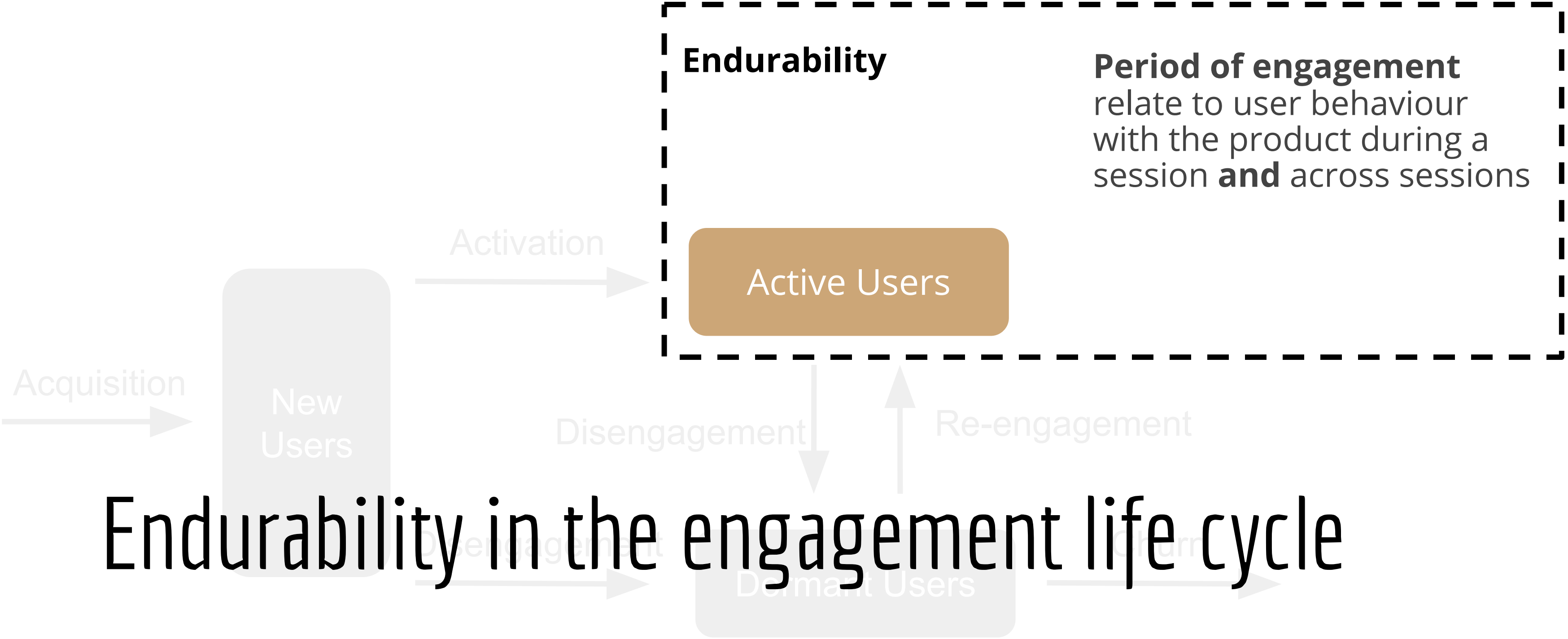
Why several metrics?





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

Quality of the user experience … endurability



Endurability in the engagement life cycle

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, likes, conversion rate, save, click, skip rate
Contribution	**Input of a user** post, comment, create, update, reply, upload

What involvement is in application A may be interaction in application B
Degree of engagement in terms of “intention” increases from **involvement** → **interaction** → **contribution**

From visit to session

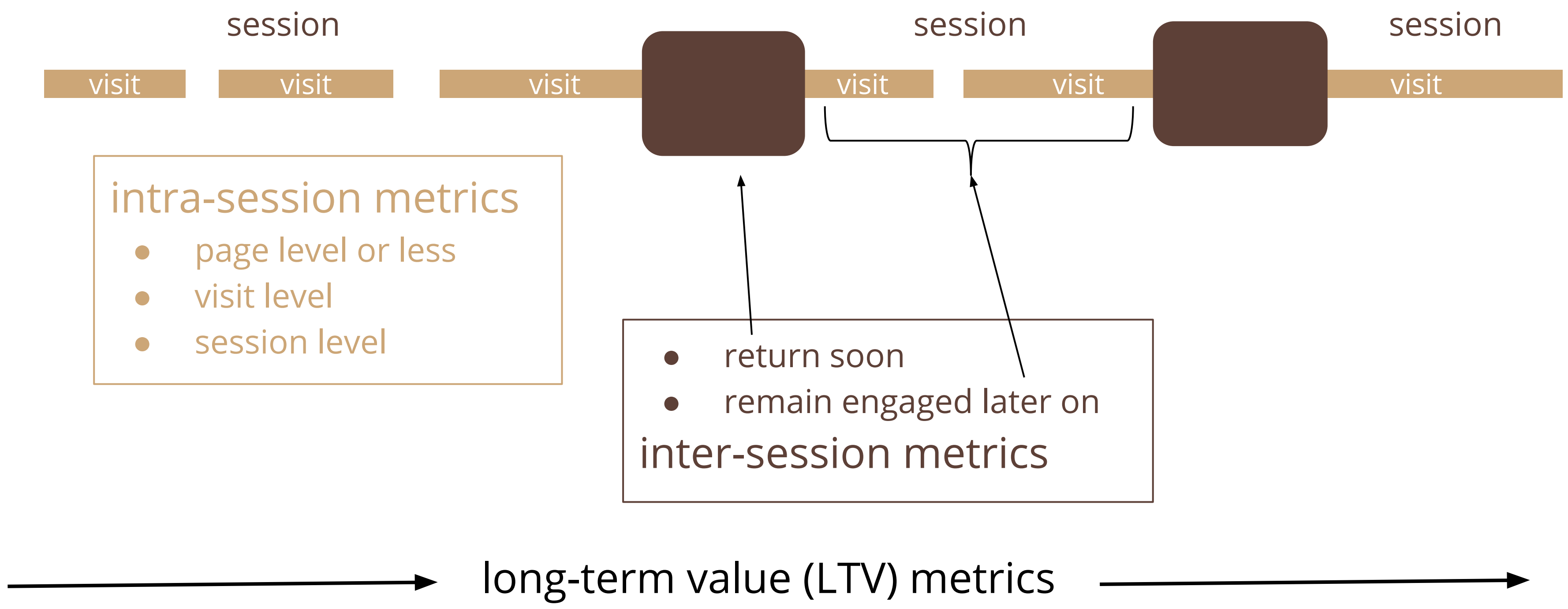
Dwell time = time spent on site (page) during a visit

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 → endurability
- inter-session engagement measures user habit & loyalty with the site → long-term value

Intra-session (within → endurability)		inter-session (across → habit)
Involvement • Dwell time • Session duration • Page view (click depth) • Revisit rate • Bounce rate **Interaction** • Click-through rate (CTR) • Number of shares, likes, saves • Conversion rate • Streamed, played **Contribution** • Number of replies • Number of blog posts • Number of uploads	**Granularity** Module ↓ Viewport ↓ Page ↓ Visit ↓ Session	**From one session to the next session (return soon)** • Time between sessions (absence time) **inter-session (across → loyalty)** **From one session to a next time period such next week, or in 2 weeks time (remain engaged later on)** • Number of active days • Number of sessions • Total usage time • Number of clicks • Number of shares • Number of thumb ups

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

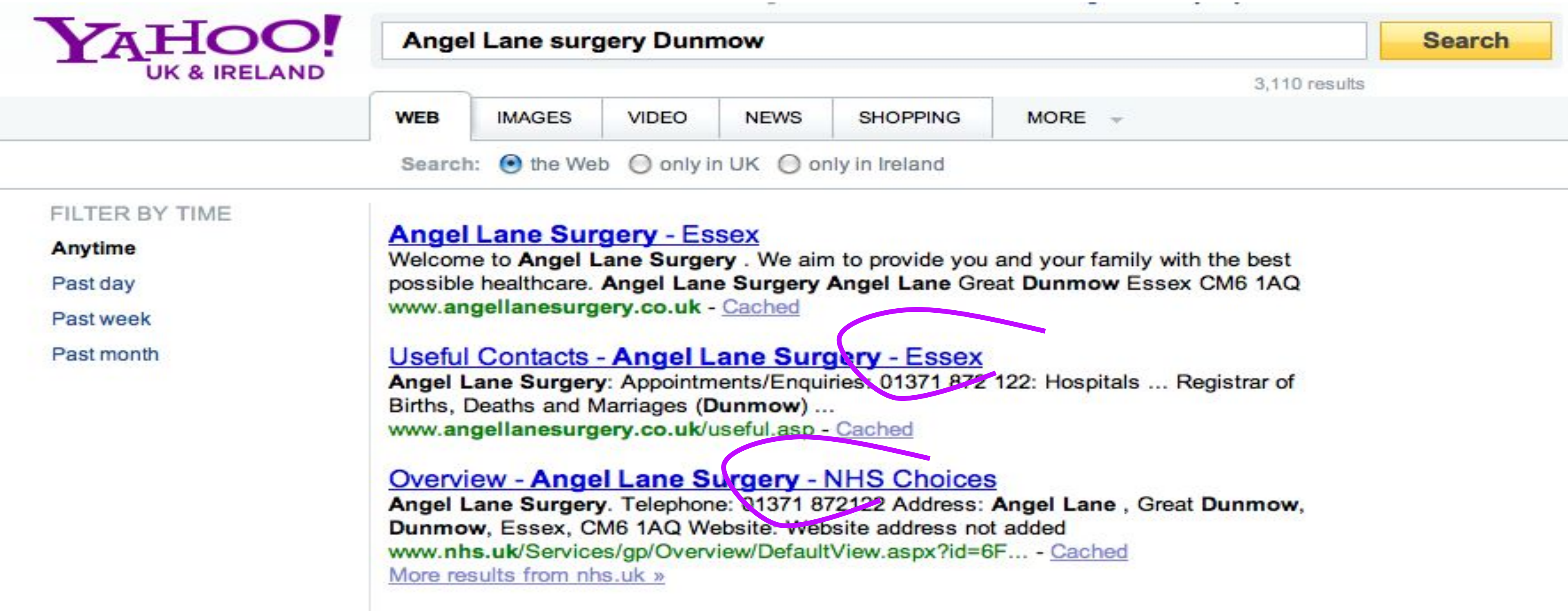
Click-through rates (CTR) ... Interaction

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



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[1]	**0.49**
No clicks (N=96)	Hover rate	0.23
	Unclicked hovers	0.06
	Max hover time	0.17
	Combined[2]	**0.28**

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)

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

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% 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

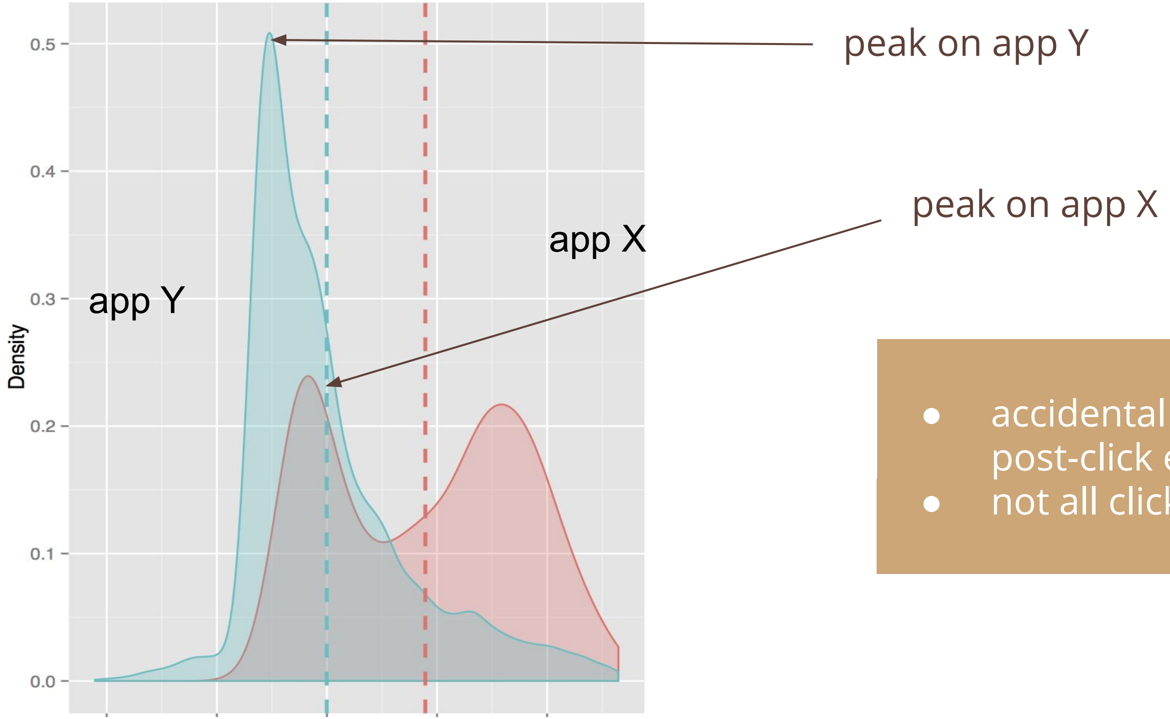
Average cursor speed (pixel/second)

Slower when answers in snippet (good abandonment)

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

The quality of a click on mobile apps … advertising

dwell time distribution of apps X and Y for given ad



- accidental clicks do not reflect post-click experience
- not all clicks are equal

Gabriele Tolomei, Mounia Lalmas, Ayman Farahat and Andy Haines. **Data-driven identification of accidental clicks on mobile ads with applications to advertiser cost discounting and click-through rate prediction.** Journal of Data Science and Analytics, 2018.

Click-through rate … Music

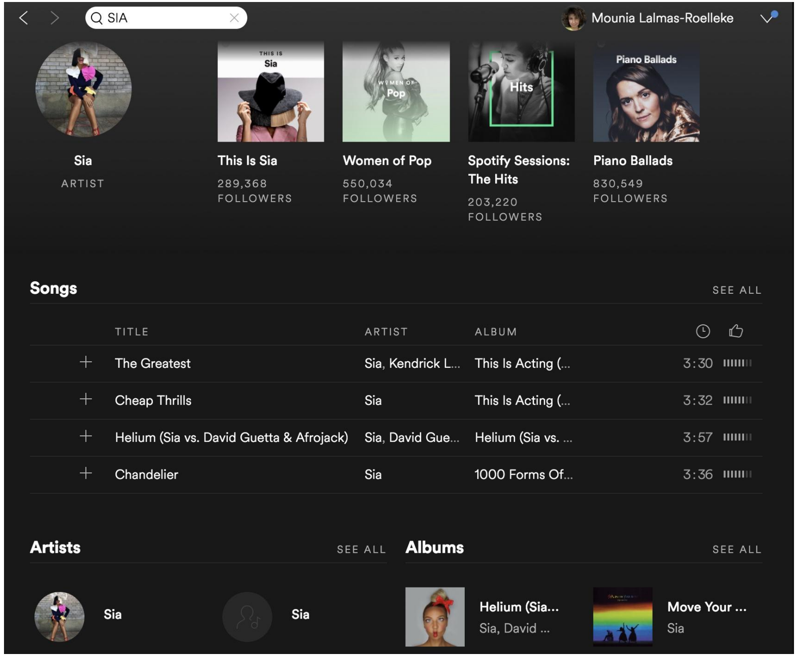
Ratio of users who click on a specific item to the number of total users who “view” that **item**

What is an item?

- Track
- Artist page
- Album
- Playlist
- …

The value of a click

→ downstream engagement

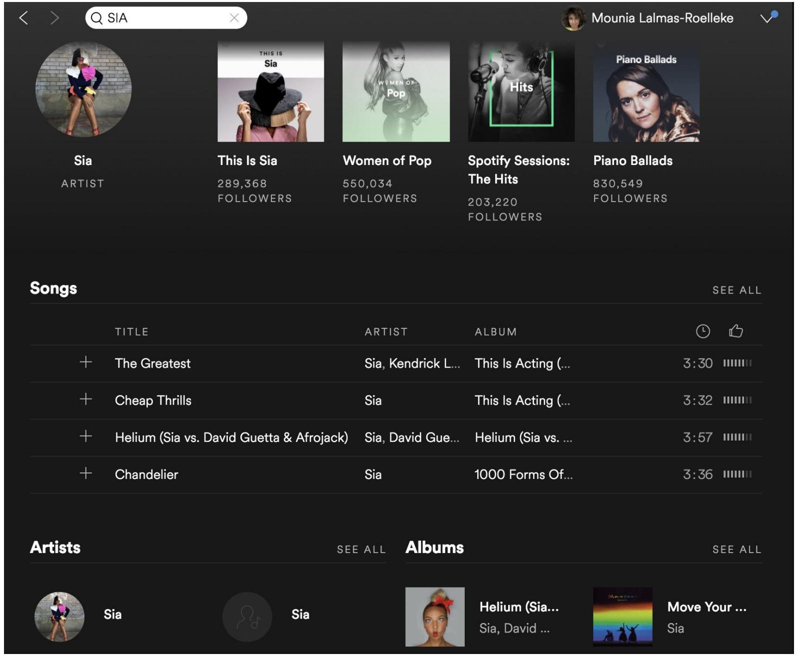


Downstream engagement … music

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**



Rishabh Mehrotra, Mounia Lalmas, Doug Kenney, Tim Lim-Meng and Golli Hashemian. **Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations.** WWW 2019.

Intra-session metrics

Click-through rate
Dwell time
"Organise" metrics
Revisit rate

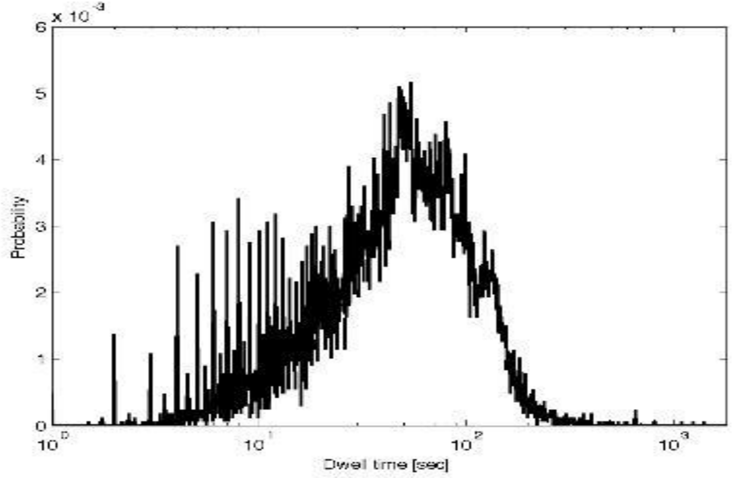
Page view
Conversion rate
Social media metrics

Dwell time ... Involvement

The contiguous time spent on a site or web page

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

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



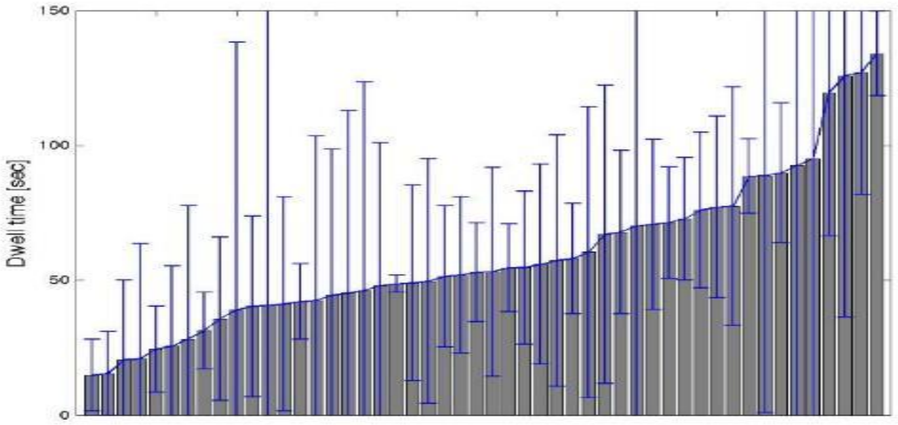
distribution of dwell times on 50 websites

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

Dwell time ... Involvement

Dwell time varies by site type: leisure sites tend to have longer dwell times than news, e-commerce, etc.

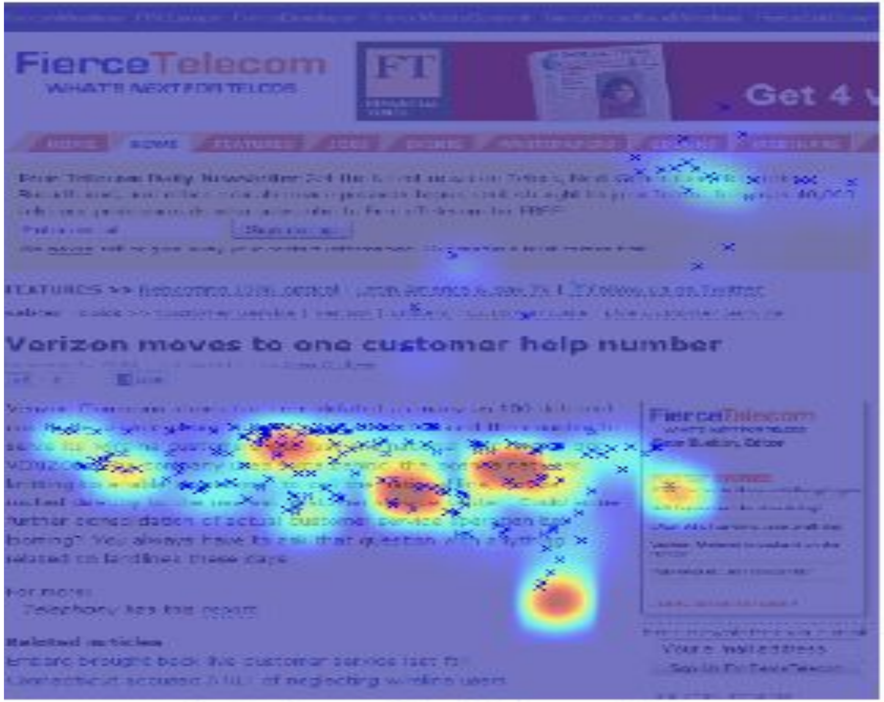
Dwell time has a relatively large **variance** even for the same site



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)

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

“reading” cursor heatmap of relevant document vs “scanning” cursor heatmap of non-relevant document

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

Dwell time

(a) relevant (dwell time: 70s)

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

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

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

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 and optimize for click-baits

See section on Offline experiment and evaluation

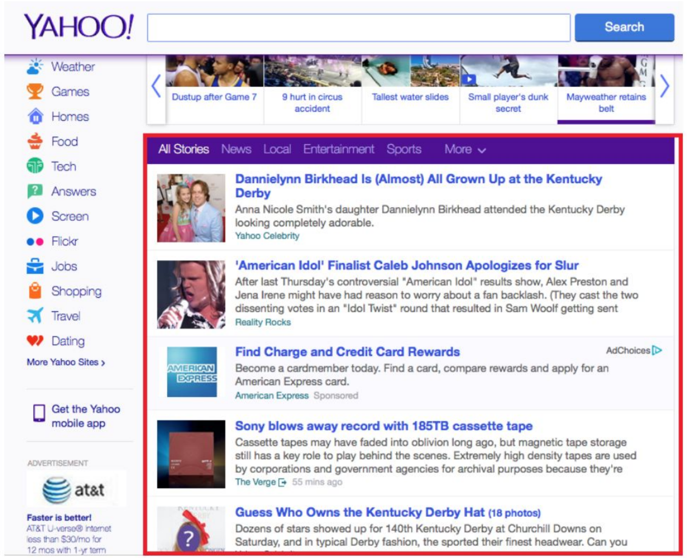
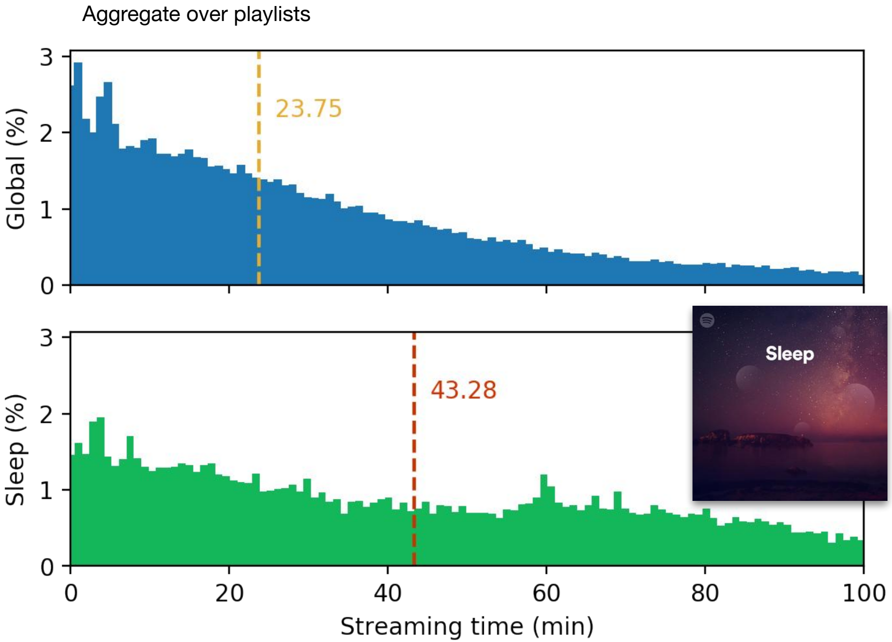


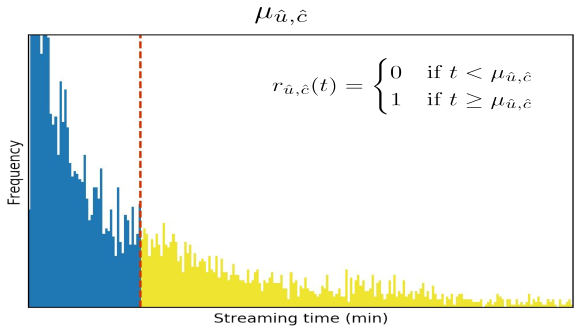
Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu and Suju Rajan. **Beyond Clicks: Dwell Time for Personalization**. RecSys 2014.

Dwell time as streaming time … music



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



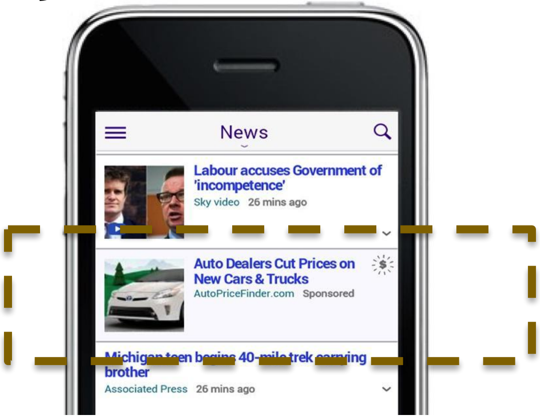
Consumption time of leep playlist longer than average playlist consumption time.

Paolo Dragone, Rishabh Mehrotra and Mounia Lalmas. **Deriving User- and Content-specific Rewards for Contextual Bandits.** WWW 2019.

Dwell time and ad landing page quality

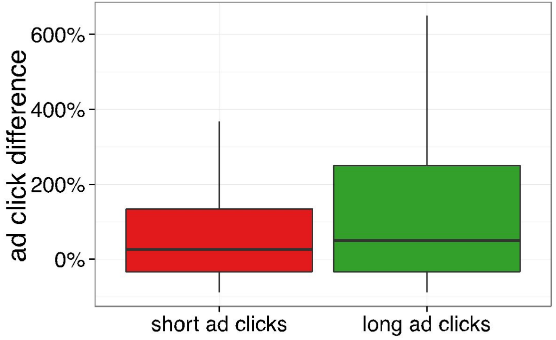
User click on an ad → ad landing page

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)



Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri and Gabriele Tolomei. **Promoting Positive Post-click Experience for In-Stream Yahoo Gemini Users.** KDD Industry track 2015.

Intra-session metrics

Click-through rate
Dwell time
“Organise” metrics
Revisit rate

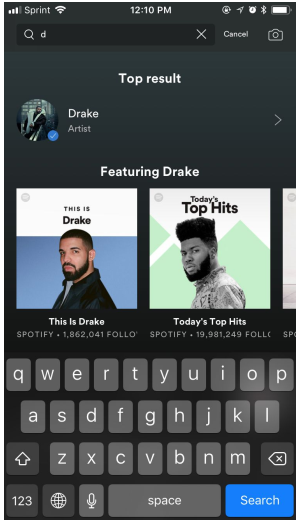
Page view
Conversion rate
Social media metrics

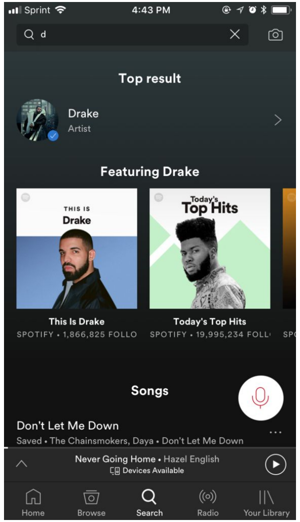
User journey in search … Music

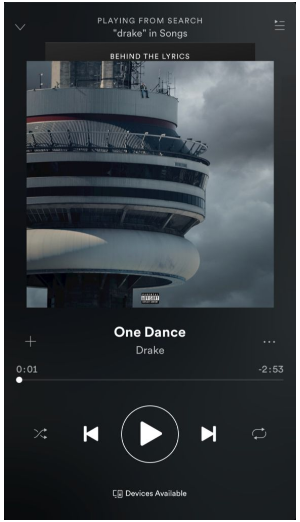
TYPE/TALK
User communicates with us

CONSIDER
User evaluates what we show them

DECIDE
User ends the search session







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

EFFORT

SUCCESS

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.

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

Intra-session metrics

Click-through rate
Dwell time
“Organise” metrics
Revisit rate

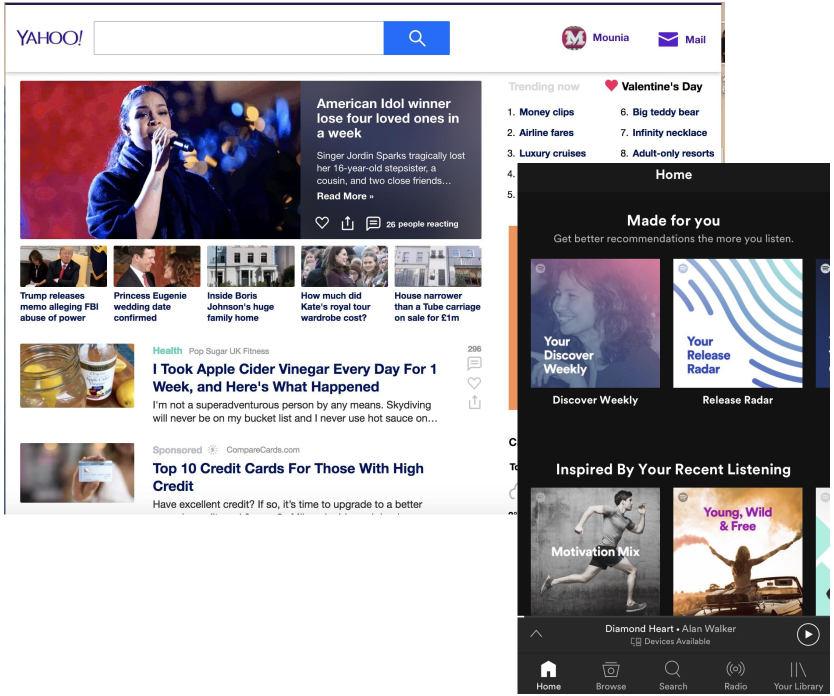
Page view
Conversion rate
Social media metrics

Revisit rates … Involvement

Number of returns to the website **within** a session → definition of a session?

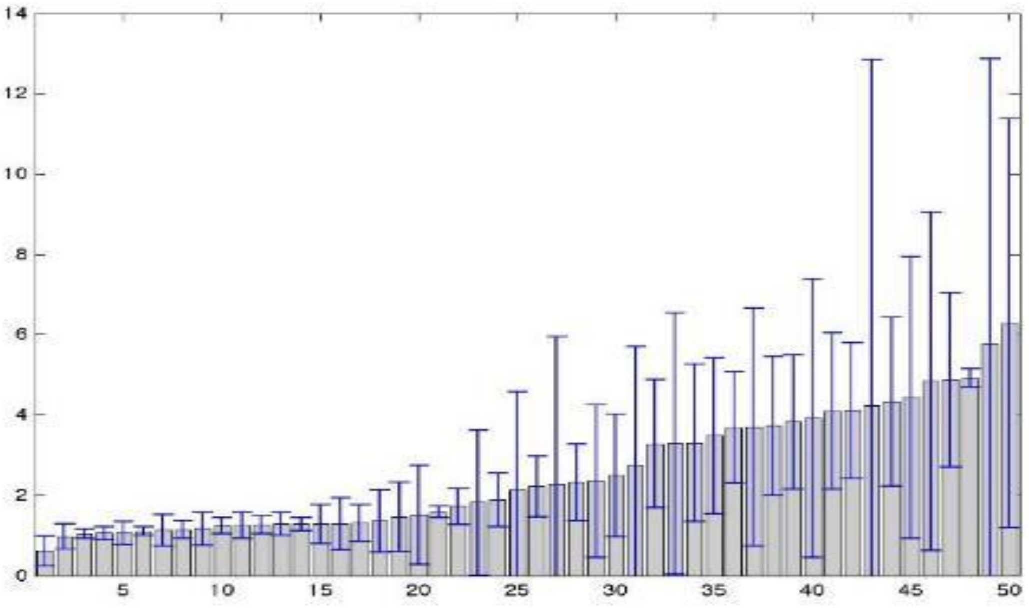
Common in sites which may be browser homepages, or contain content of regular interest to users.

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



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



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

Revisit rate … Session length

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

Categorization of the most frequently accessed sites

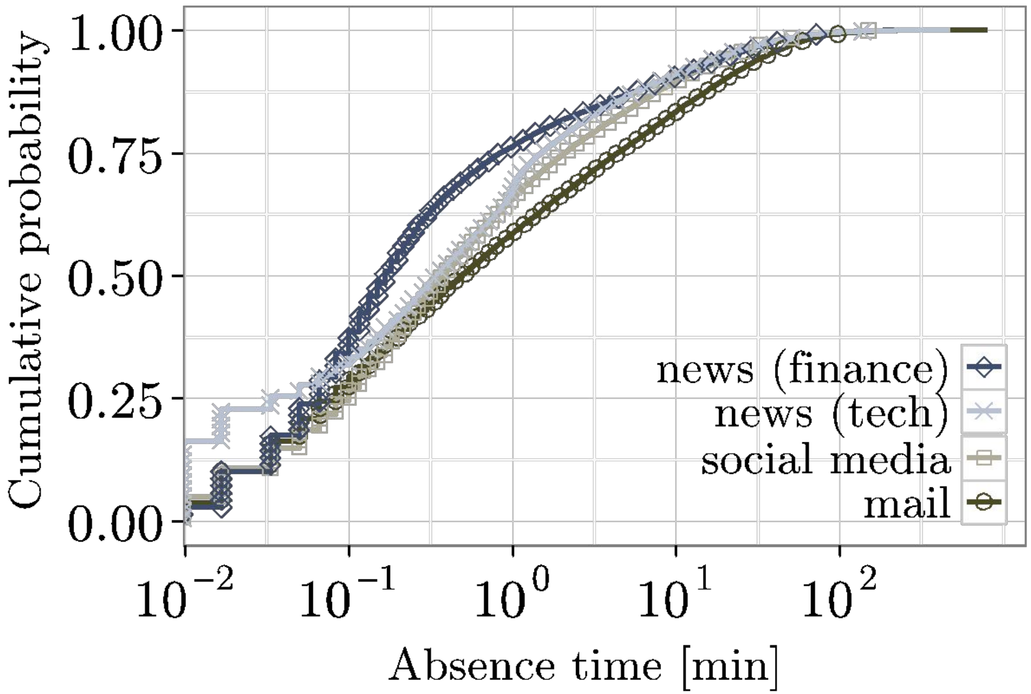
- 11 categories (e.g. news), 33 subcategories
- (e.g. news finance, news society)
- 60 sites from 70 countries/regions

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%	*search for lyrics, jobs, etc.*
	directory	1.05%	
service 11.6%	service	7.63%	*translators, banks, etc.*
	maps	3.03%	
	organization	0.92%	*bookmarks, calendar, etc.*
sharing 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%	front page	6.58%	
	front page (pers.)	1.84%	*personalized front pages*
	sitemap	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 3.0%	social media	1.97%	
	dating	1.05%	
settings 2.9%	login	1.71%	
	settings	1.18%	*profile setting, site personalization*

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

Janette Lehmann, Mounia Lalmas, Georges Dupret and Ricardo Baeza-Yates. **Online Multitasking and User Engagement.** CIKM 2013.

Time between each revisit … online multi-tasking



50% of sites are revisited after less than 1 minute

Janette Lehmann, Mounia Lalmas, Georges Dupret and Ricardo Baeza-Yates. **Online Multitasking and User Engagement.** CIKM 2013.

Intra-session metrics

Click-through rate
Dwell time
“Organise” metrics
Revisit rate

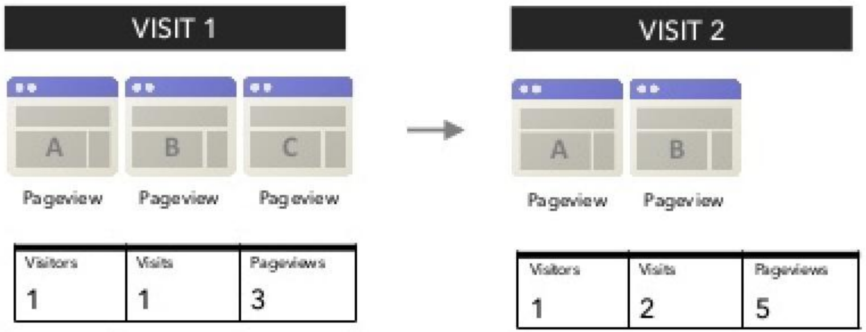
Page view
Conversion rate
Social media metrics

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

Reload after reaching the page → counted as additional pageview
If same page viewed more than once → a single unique pageview



https://www.slideshare.net/timothylelek/google-analytics-for-dummies

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

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, free to premium user 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

Social media metrics

… interaction

Applause
#like, #thumbs up or down, #hearts, +1

… interaction

Amplification
#share, #mail

… contribution

Conversations
#comments, #posts, #replies, #edits

https://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/

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 may 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] 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 → 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, etc.

Inter-session metrics

Why inter-session metrics

Relationship to loyalty

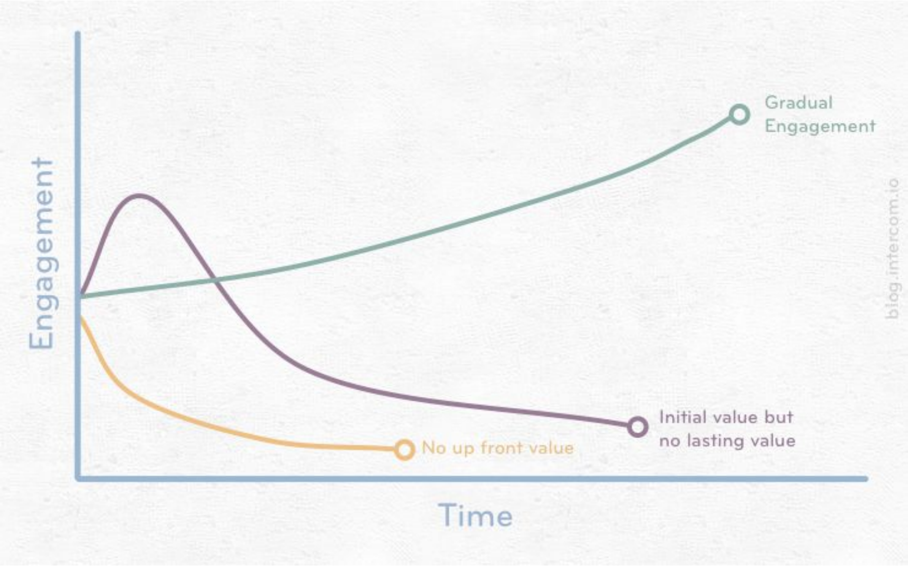
Absence time

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

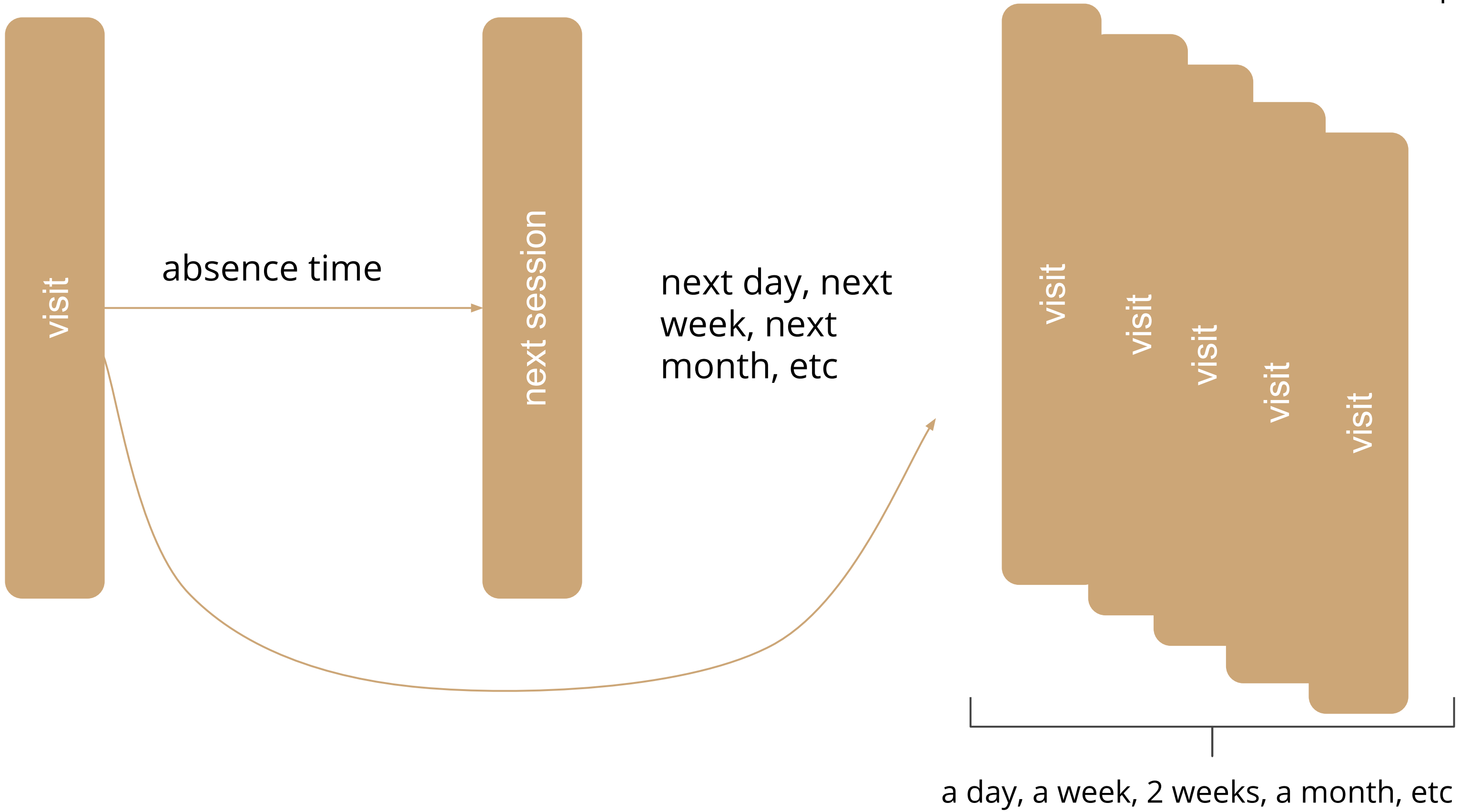
Therefore, bucket testing may provide erroneous results if only intra-session measures are used



Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker and Ya Xu. **Trustworthy online controlled experiments: Five puzzling outcomes explained.** KDD 2012.

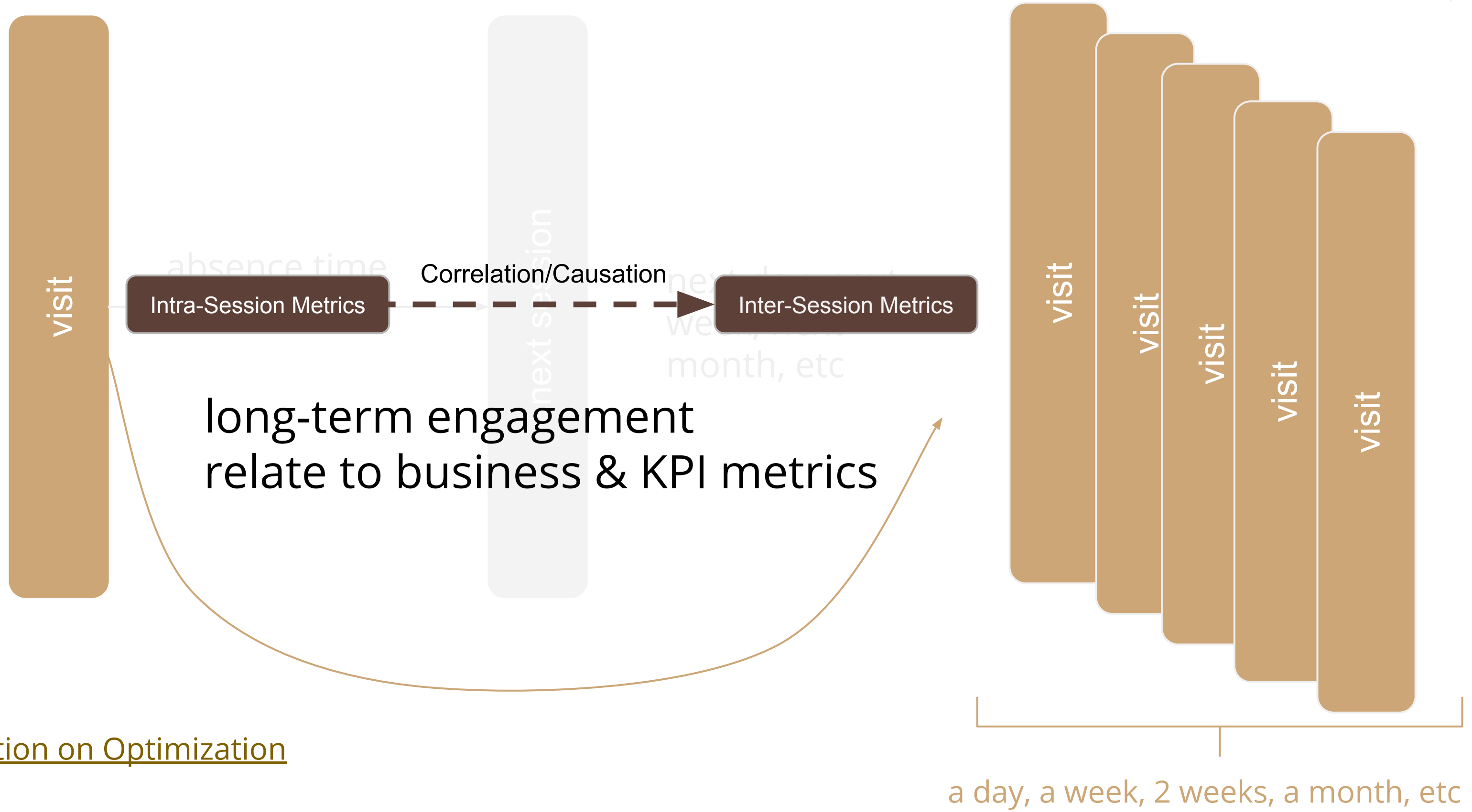
Inter-session metrics

Total number of sessions
Total number of days active
Total number of clicks
Total amount of time spent ...



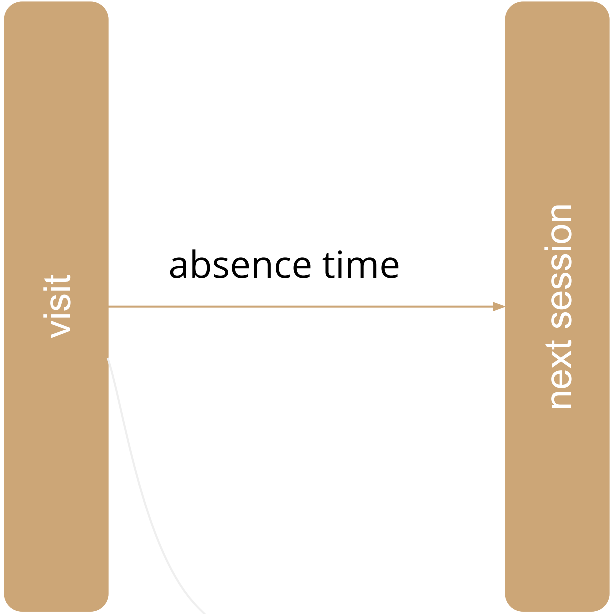
Inter-session metrics

… loyalty



See section on Optimization

Inter-session metrics ... loyalty



really mostly about endurability

habit
periodicity
short task/visit

absence time ≠ revisit rate

Cases studies: search and news

Absence time applied to search … Study I

Ranking functions on Yahoo Answer Japan

Two-weeks click data on Yahoo Answer Japan search

One millions users
Six ranking functions

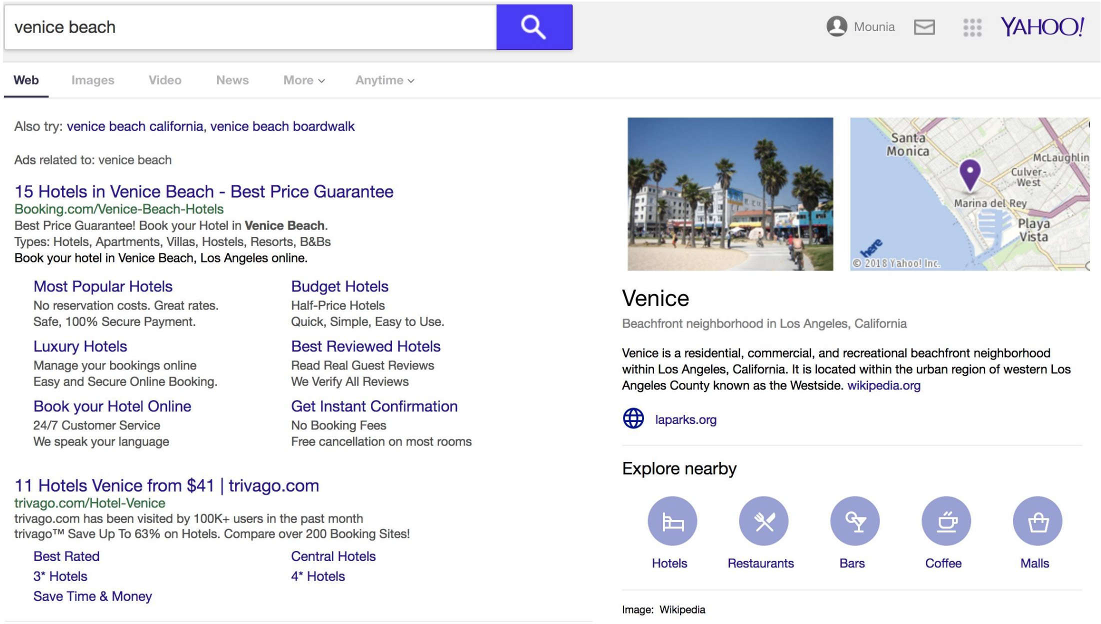
Session boundary:
30 minutes of inactivity

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

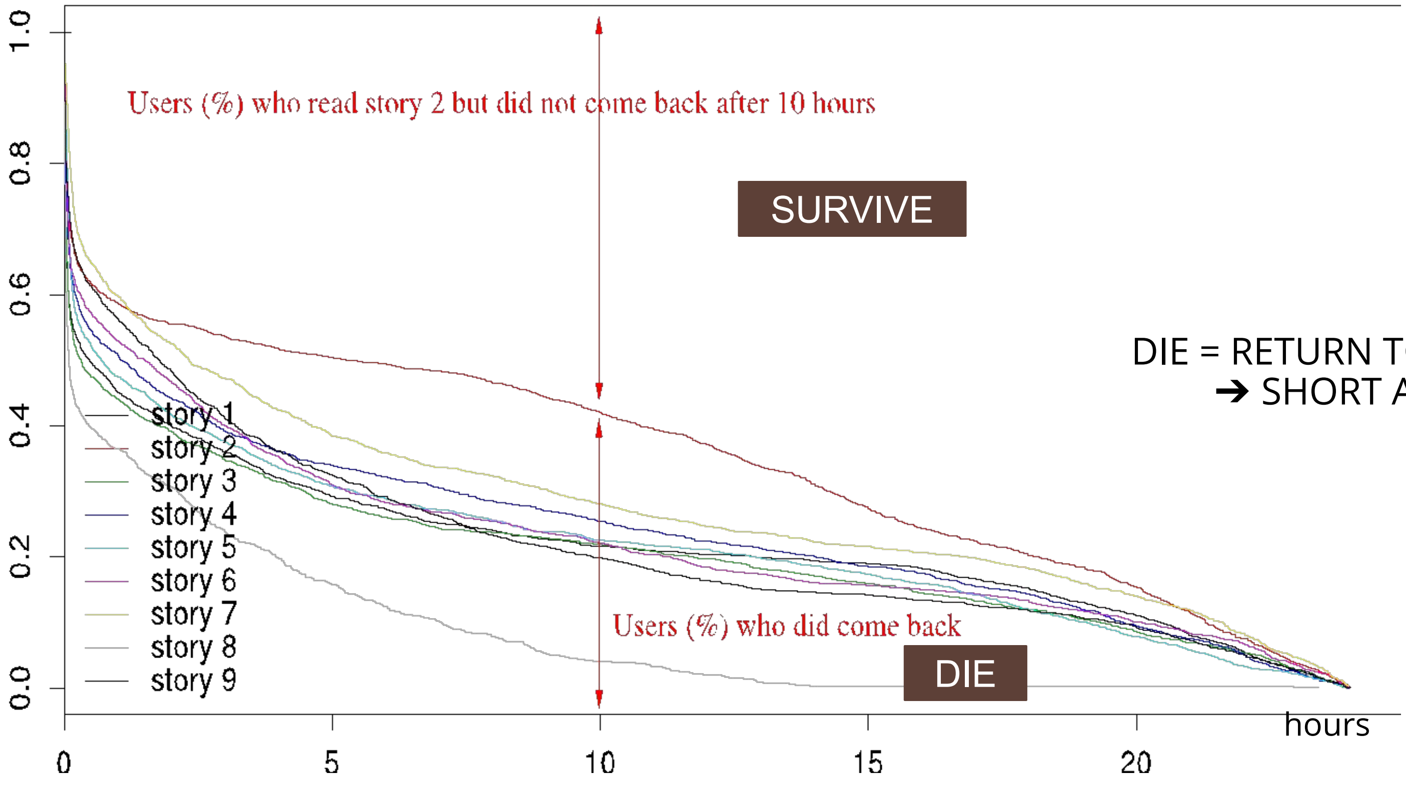
Examples of metrics for search

(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)



Absence time and survival analysis



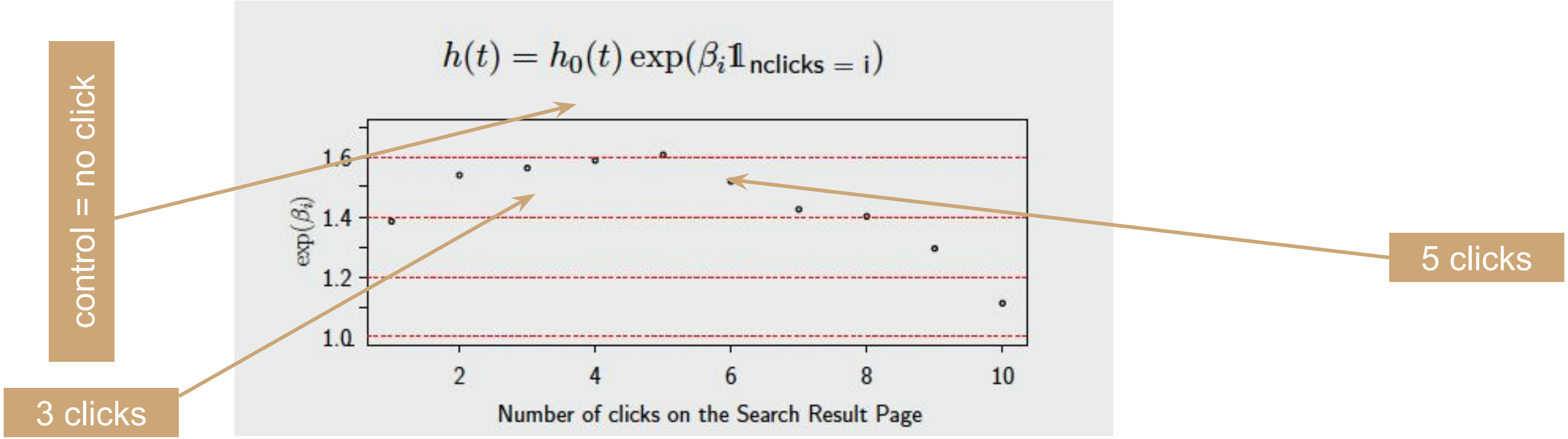
DIE = RETURN TO SITE

➔ SHORT ABSENCE TIME

Odd Aalen, Ornulf Borgan and Hakon Gjessing. **Survival and Event History Analysis: A Process Point of View.** Statistics for Biology and Health, 2008.

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

DCG versus absence time to evaluate five ranking functions

YAHOO! JAPAN

DCG@1	DCG@5	Absence time
Ranking Alg 1	Ranking Alg 1	Ranking Alg 1
Ranking Alg 2	Ranking Alg 3	Ranking Alg 2 Ranking Alg 5
Ranking Alg 3	Ranking Alg 2	Ranking Alg 3
Ranking Alg 4	Ranking Alg 4	Ranking Alg 4
Ranking Alg 5	Ranking Alg 5	

Absence time and search session … What else?

intra-session search metrics → absence time

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

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

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 (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

Sunandan Chakraborty, Filip Radlinski, Milad Shokouhi and Paul Baecke. **On Correlation of Absence Time and Search Effectiveness.** SIGIR 2014.

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

intra-session search metrics → absence time

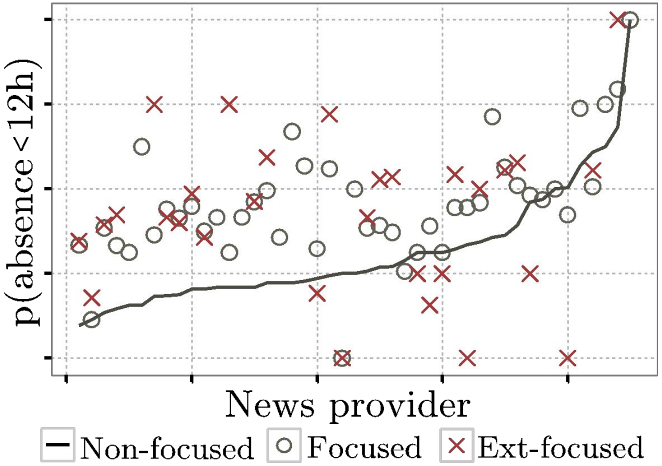
Demonstrated that absence time is an appropriate inter-session metric for search because of the correlation & predictive power of known indicators of a positive search experience

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

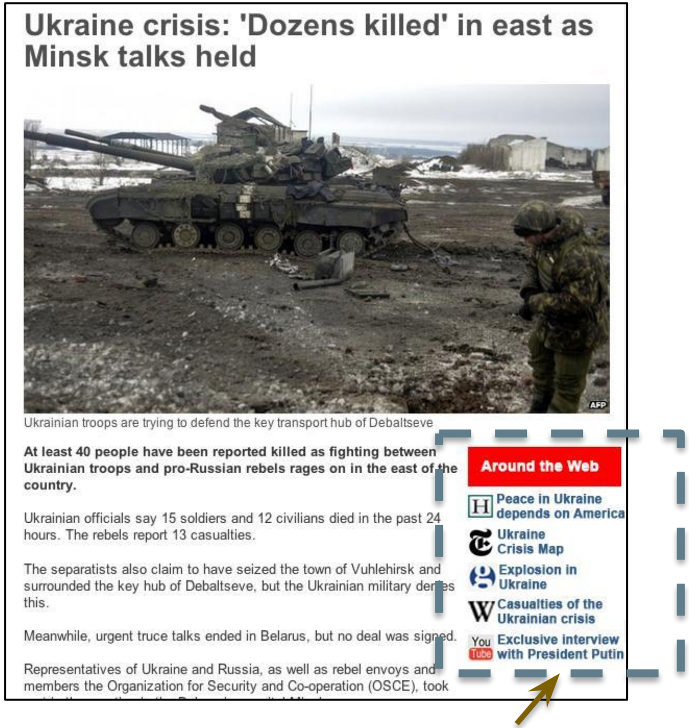
These known indicators could act as intra-session metrics, which could be optimised by the ranking algorithms

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%



Related off-site content

Janette Lehmann, Carlos Castillo, Mounia Lalmas and Ricardo Baeza-Yates. **Story-focused Reading in Online News and its Potential for User Engagement.** JASIST 2016.

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
- E-commerce relying on actual purchases:
  - based on total amount of purchases

Help analyzing acquisition strategy (customer acquisition cost) and estimate further marketing costs

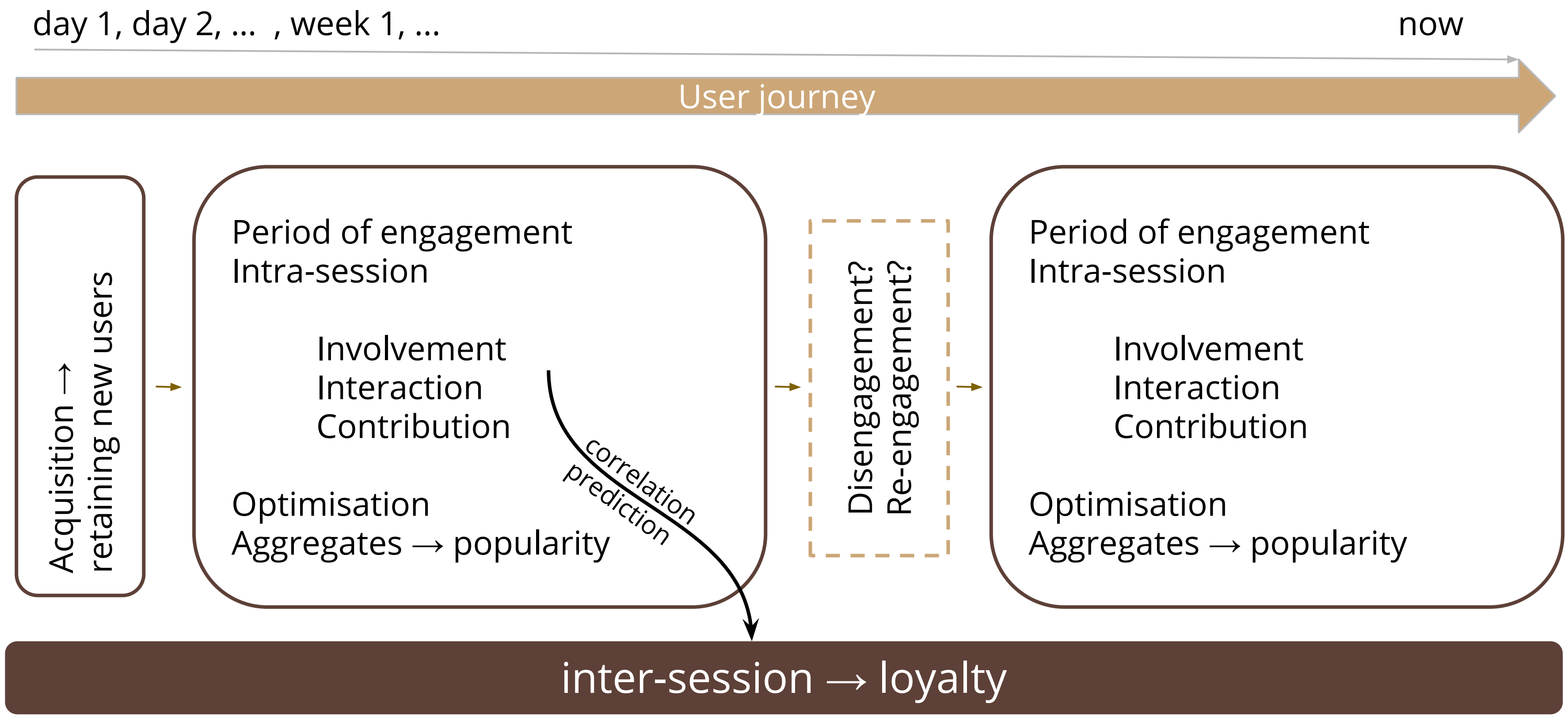
https://www.apptamin.com/blog/lifetime-value-mobile-customer/

Recap

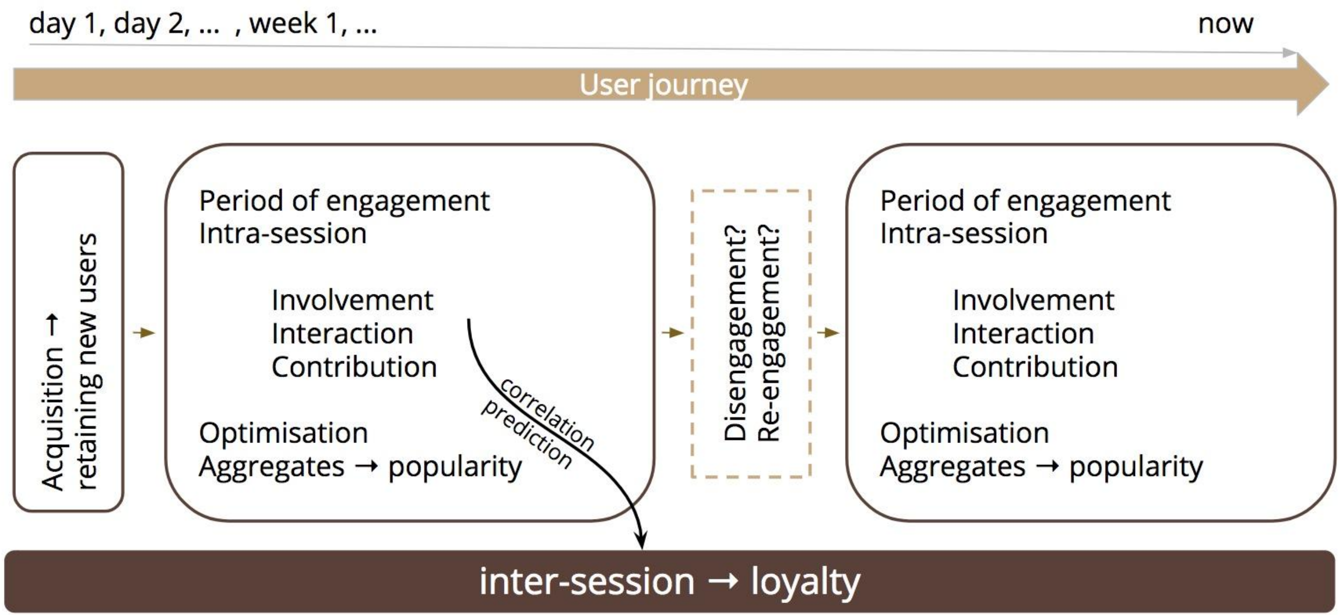
Online engagement & metrics

How it all fits together

Online engagement & metrics … recap



Online engagement & metrics … all together



- Popularity metrics
- Metrics to use to optimize machine learning algorithms
- Key performance indicators (KPIs)
- Long-term value (LTV) metrics

Optimization

Optimization

Manual/Semi-Manual Optimization

Automatic Optimization

Combining Two Camps

Two Camps of Optimizations

- **Manual/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/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/Semi-Manual Optimization

Online Experiments and Evaluation

Offline Experiments and Evaluation

Observational Study

Manual/Semi-Manual Optimization

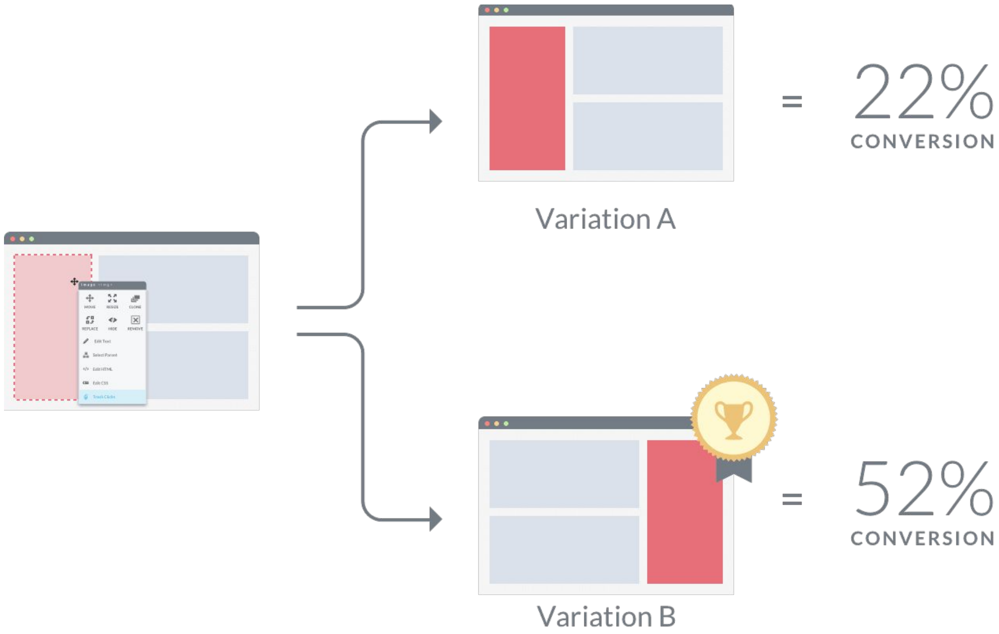
Algorithm 1 Better Data Scientist Descent

```
procedure BETTER DATA SCIENTIST DESCENT
loop:
    Design metrics around company goals
    Create event predictors
    Search through value functions with automatic A/B tests
    goto loop.
```

Introduced by Jason Gauci from Facebook

Online Experiments and Evaluation

A/B Tests or Bucket Tests or Online Controlled Experiments



Online Experiments and Evaluation

- A lot of statistical tools offer measuring the difference between control and treatment
- Link to *Average Treatment Effect* (ATE) in Causal Inference
- Sometimes the only way to understand causal effects
- *Easy* to implement and easy to explain

[1] Ben Carterette. **Statistical Significance Testing in Information Retrieval: Theory and Practice**. SIGIR 2017 Tutorial.
[2] Tetsuya Sakai. **Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015**. SIGIR 2016.
[3] Tetsuya Sakai. **The Probability that Your Hypothesis Is Correct, Credible Intervals, and Effect Sizes for IR Evaluation**. SIGIR 2017.
[4] Benjamin A. Carterette. **Multiple Testing in Statistical Analysis of Systems-based Information Retrieval Experiments**. ACM Trans. Inf. Syst. 30, 1, Article 4, 2012.

Online Experiments and Evaluation

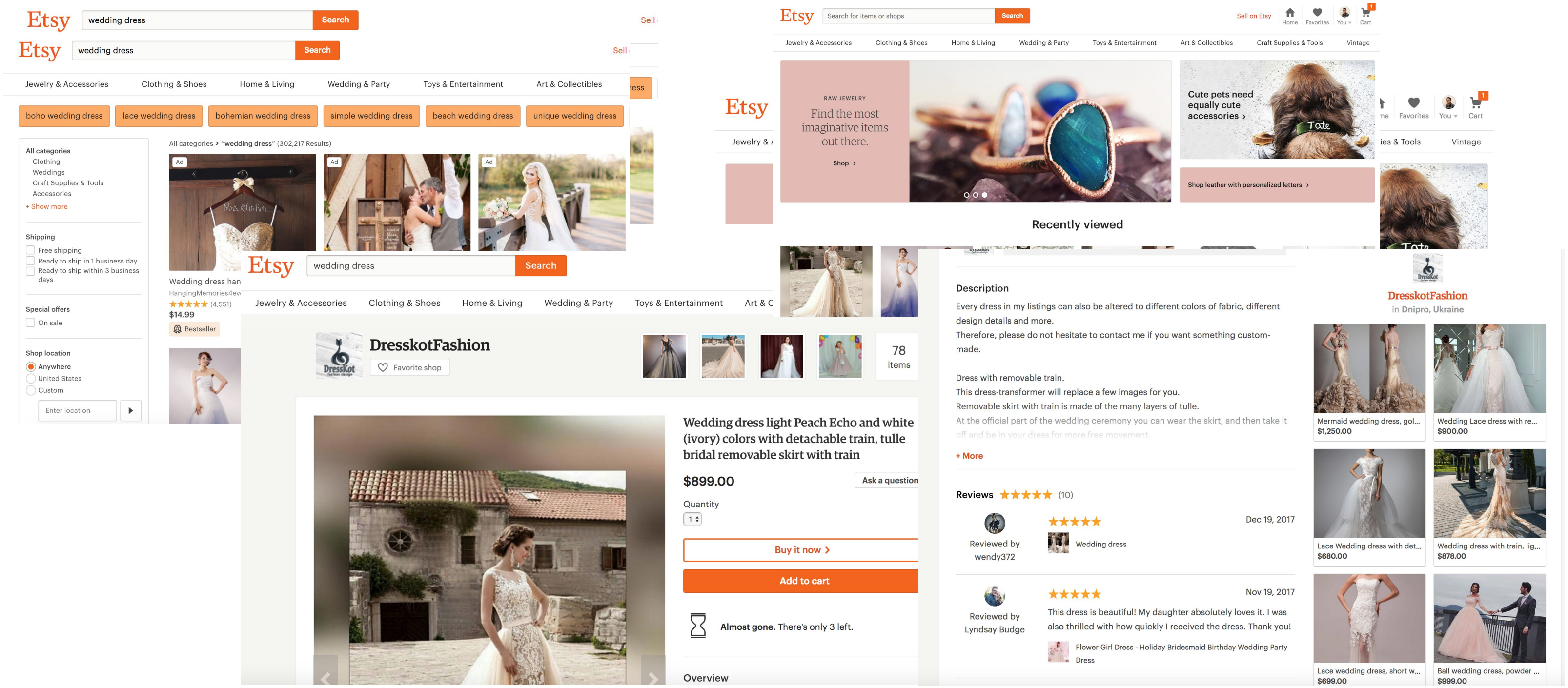
- A lot of statistical tools offer measuring the difference between control and treatment
- Link to *Average Treatment Effect* (ATE) in Causal Inference
- Sometimes the only way to understand causal effects
- *Easy* to implement and easy to explain

- Not well studied in a lot of online settings
- Gold standard for statistical difference
- Weak for practical difference

[1] Ben Carterette. **Statistical Significance Testing in Information Retrieval: Theory and Practice**. SIGIR 2017 Tutorial.
[2] Tetsuya Sakai. **Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015**. SIGIR 2016.
[3] Tetsuya Sakai. **The Probability that Your Hypothesis Is Correct, Credible Intervals, and Effect Sizes for IR Evaluation**. SIGIR 2017.
[4] Benjamin A. Carterette. **Multiple Testing in Statistical Analysis of Systems-based Information Retrieval Experiments**. ACM Trans. Inf. Syst. 30, 1, Article 4, 2012.

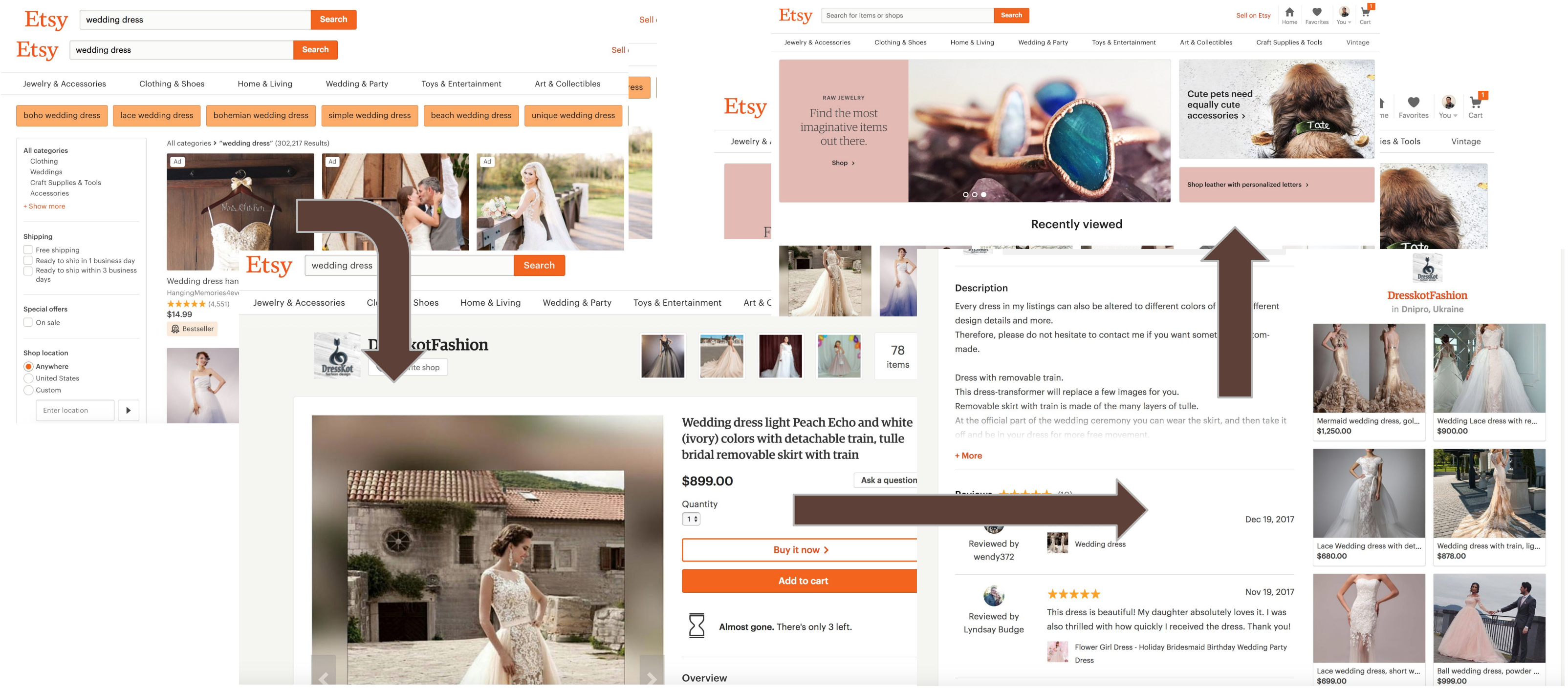
Online Experiments and Evaluation

A/B Tests or Bucket Tests or Online Controlled Experiments



Online Experiments and Evaluation

A/B Tests or Bucket Tests or Online Controlled Experiments



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 <u>Controlled</u> Experiments and Evaluation**
  - Pros:
    - A lot of statistical tools offer measuring the difference between control and treatment
    - Link to *Average Treatment Effect* (ATE) in Causal Inference
    - Sometimes the only way to understand causal effects
    - *Easy* to implement and easy to explain
  - Cons:
    - Live traffic is limited (100%)
    - Power differences need time (days to weeks)
    - Cycles and number of innovations are bounded
    - Might hurt user engagement
    - Engineering cost
    - Cannot re-use
    - Nuances to get *more accurate* insights

Online Experiments and Evaluation

Metrics for Online Experiments

- **Directional**
  Have correlations with inter-session metrics and KPIs.

Online Experiments and Evaluation

Metrics for Online Experiments

- **Directional**
  Have correlations with inter-session metrics and KPIs.
- **Sensitivity**
  Easily detect changes.

Online Experiments and Evaluation

Summary

- Direct and dynamic
- Causality
- Metrics for online experiments
- Impacts (e.g, user engagement, traffic, set-up and etc.)
- Cannot re-use

[1] Ron Kohavi, Roger Longbotham, Dan Sommerfield and Randal M. Henne. 2009. **Controlled Experiments on the Web: Survey and Practical Guide**. DMKD 18, 1 (February 2009).
[2] Alex Deng and Xiaolin Shi. 2016. **Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned**. KDD 2016.
[3] Pavel Dmitriev, Somit Gupta, Dong Woo Kim and Garnet Vaz. 2017. **A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments**. KDD 2017.

Offline Experiment and Evaluation

Traditional Offline Dataset/Collection Experiment

- **High risk experiments**.
  It may drive users away.

Offline Experiment and Evaluation

Traditional Offline Dataset/Collection Experiment

- **High risk experiments**.
  It may drive users away.
- **Learn more insights & highly reusable**.
  Easy to gather data and easy to compute metrics and compare.

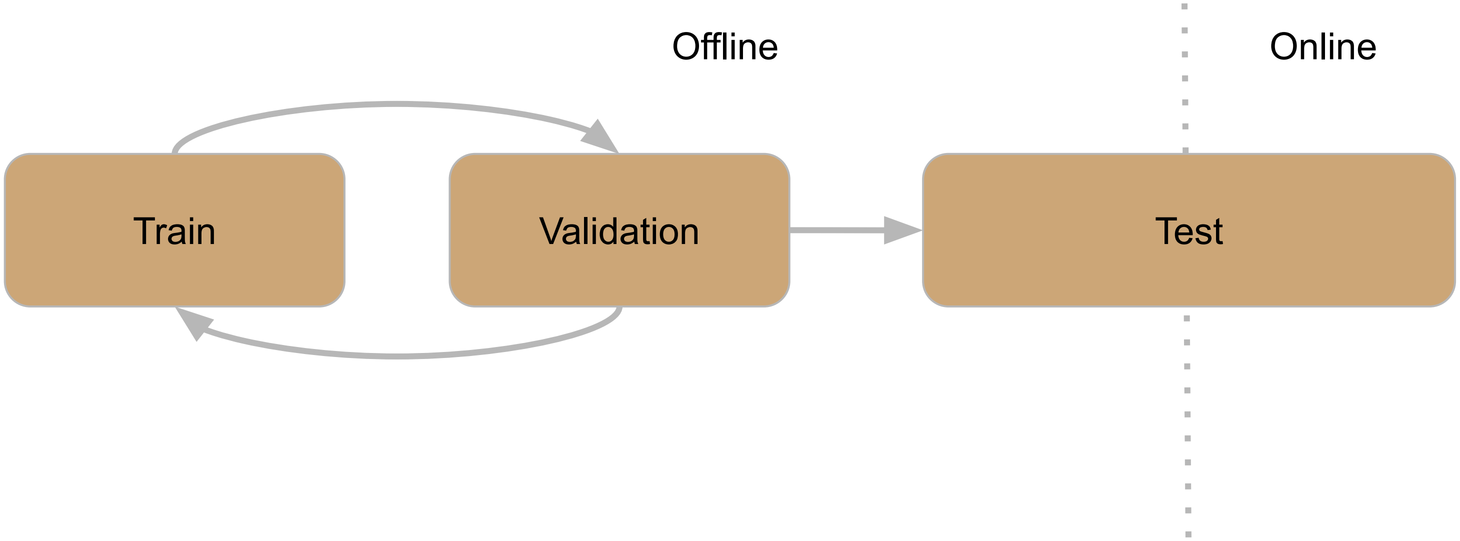
Offline Experiment and Evaluation

Traditional Offline Dataset/Collection Experiment

- **High risk experiments**.
  It may drive users away.
- **Learn more insights & highly reusable**.
  Easy to gather data and easy to compute metrics and compare.
- **Machine learning theory of generalization**.
  Textbook scenario.

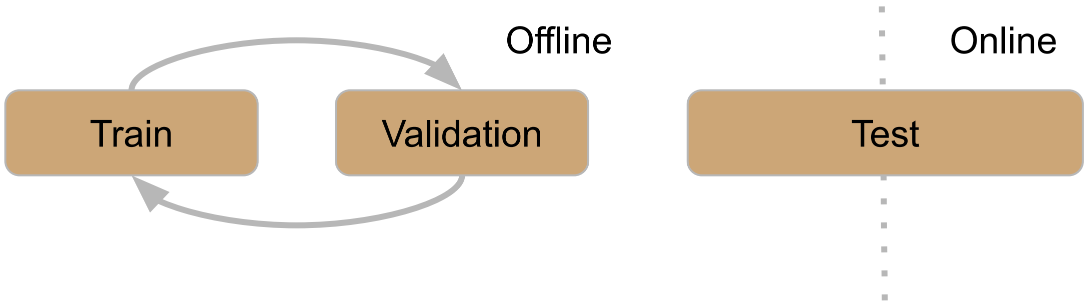
Offline Experiment and Evaluation

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 Inter-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 Inter-Session Metrics

Approach I

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 Inter-Session Metrics

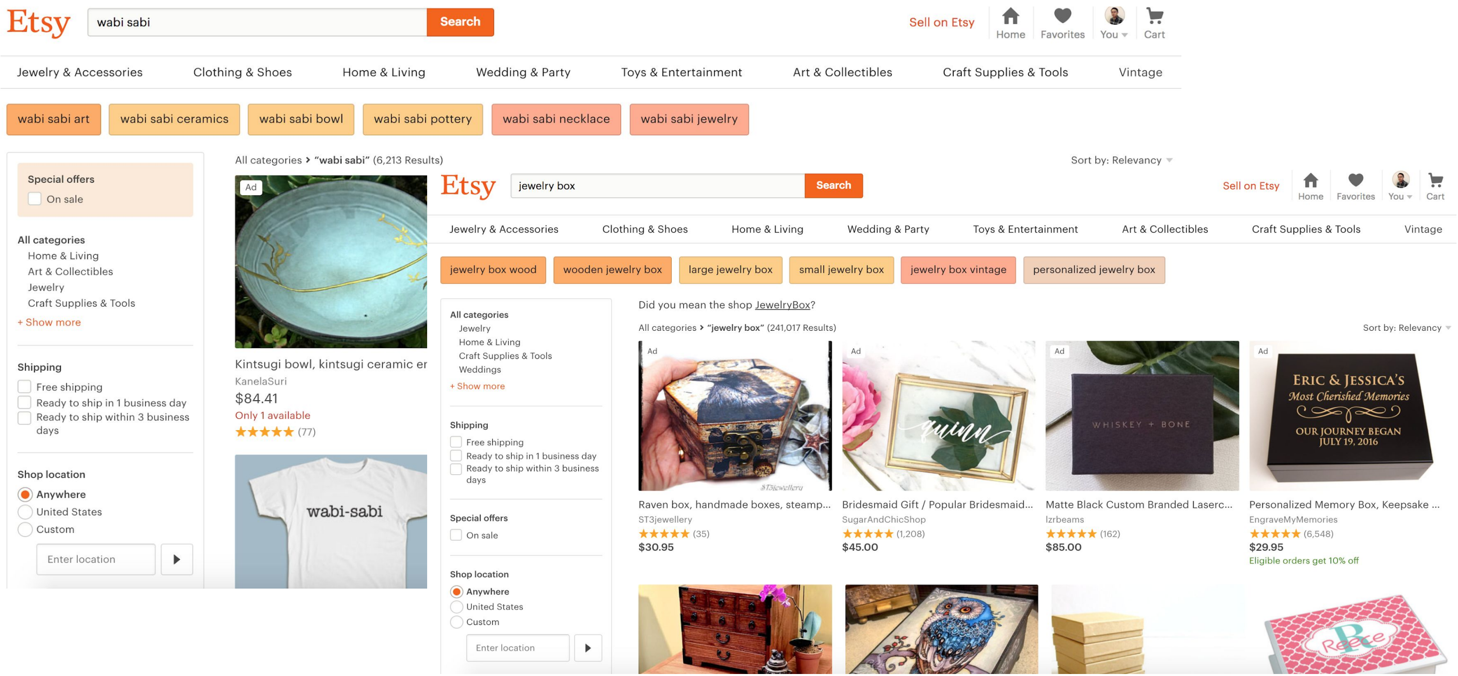
Approach I

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.
  - Your have extremely sparse data.
  - Hard to deploy such systems.

...

Offline Experiment and Evaluation



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

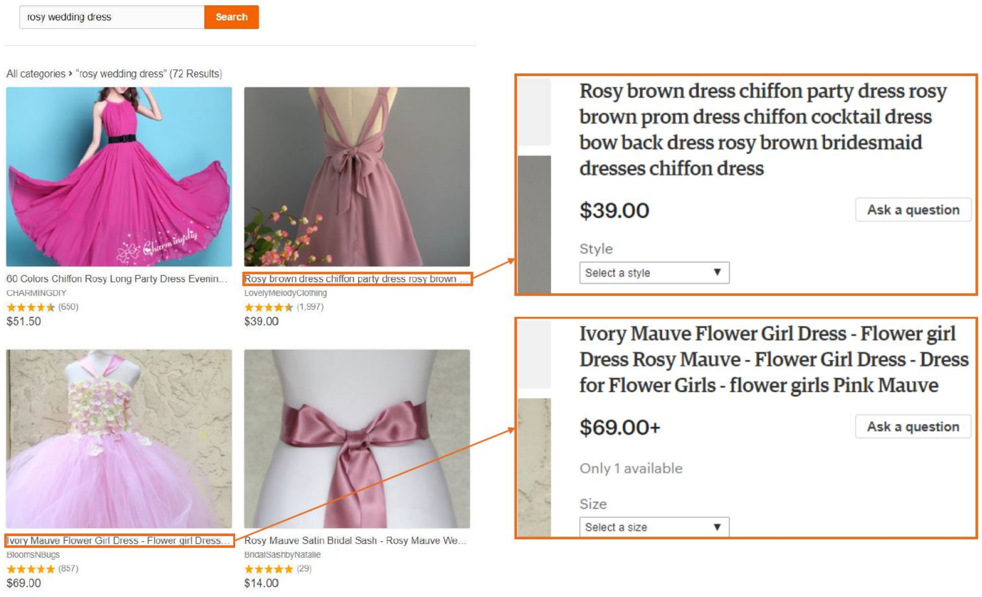
- **Expected GMV**

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

Offline Experiment and Evaluation

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

- **Purchase Decision Process**



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

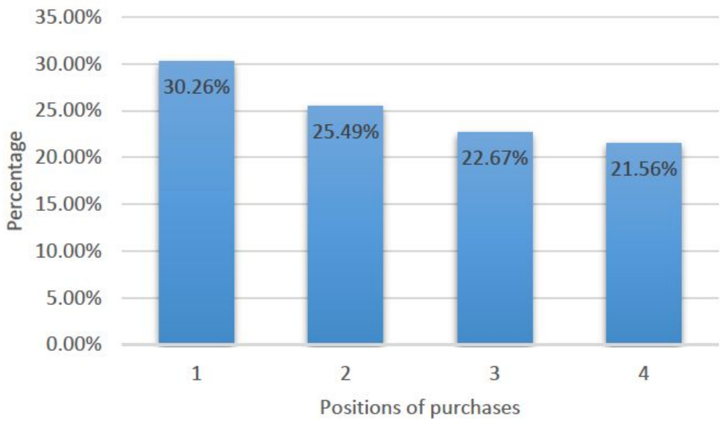


Figure 2: Position distribution of items being purchased in the top 4 spots of a search result page. The first position achieves the most purchases, while nearly 70% of purchases are in the lower positions.

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_{DIR}
		LM_{JM}
		LM_{ABS}
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	**0.1412**	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	**0.1801**	0.2039	0.1698	**0.1494**

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	**4.62**	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**	**4.90**	**5.08**	**5.47**	**5.64**	**5.85**	**6.02**	**6.19**	**6.40**	**6.54**

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?

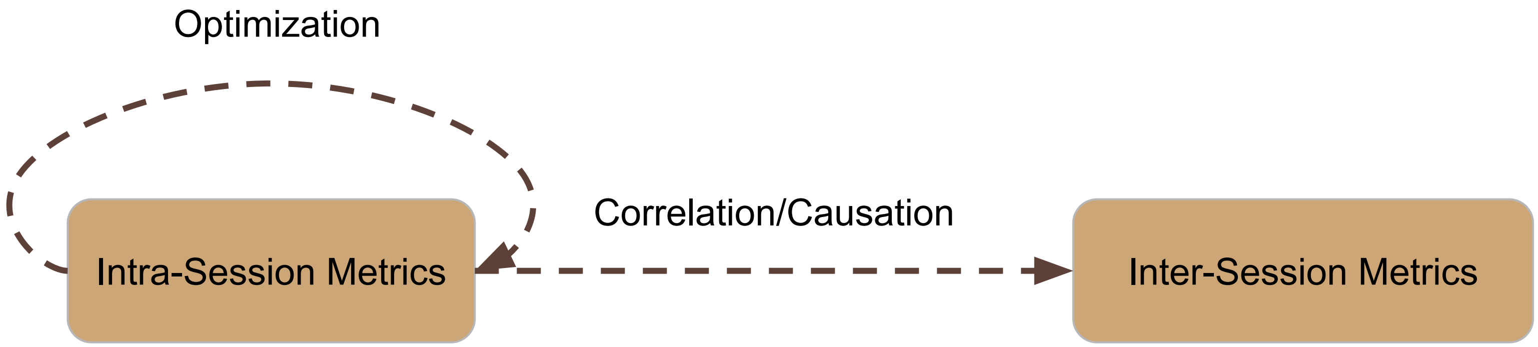
...

- First step in optimizing user engagements in E-commerce search.

Offline Experiment and Evaluation

Optimizing Inter-Session Metrics

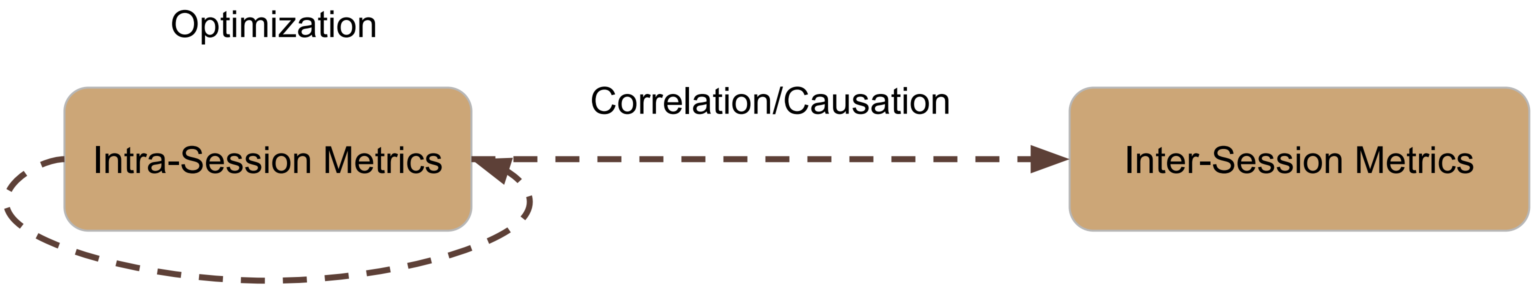
Approach II



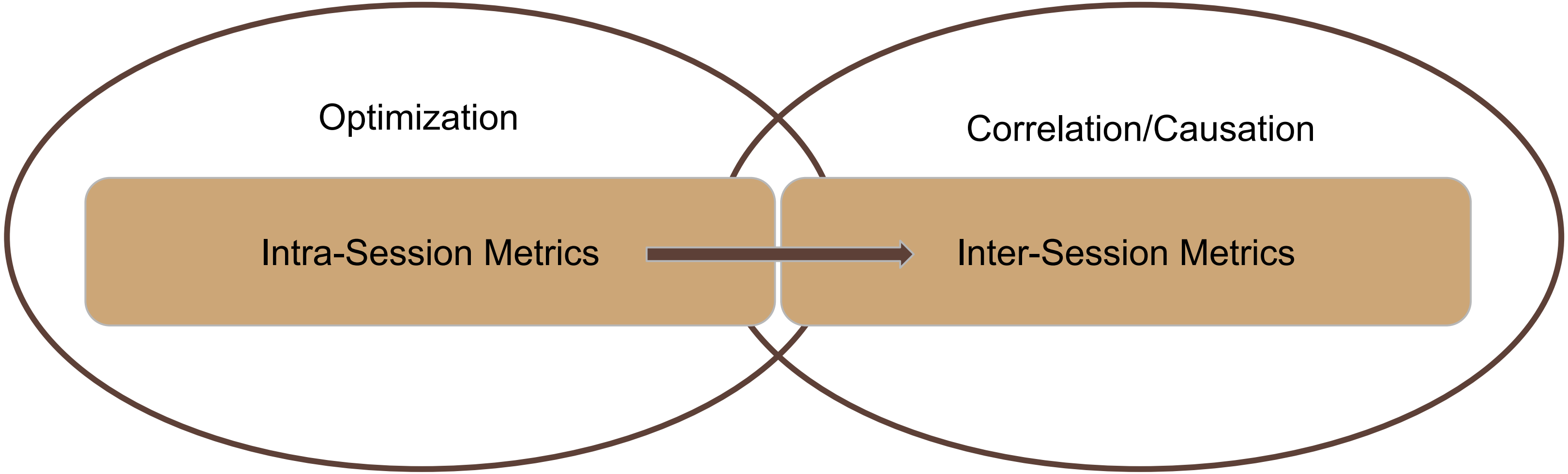
Offline Experiment and Evaluation

Approach II

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

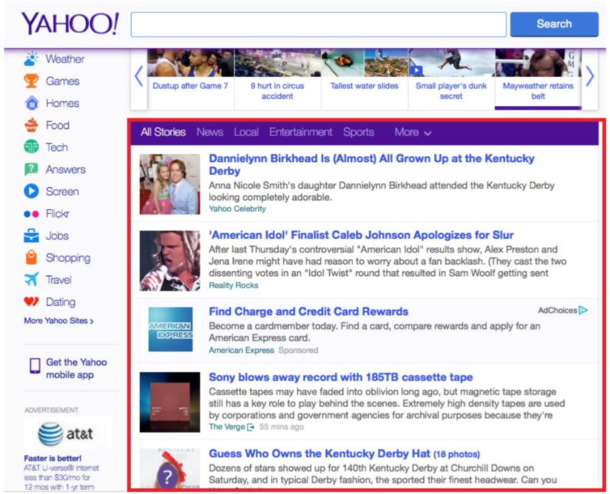


Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu and Suju Rajan. **Beyond Clicks: Dwell Time for Personalization**. RecSys 2014.

Offline Experiment and Evaluation

Beyond Clicks: Dwell Time in Personalization

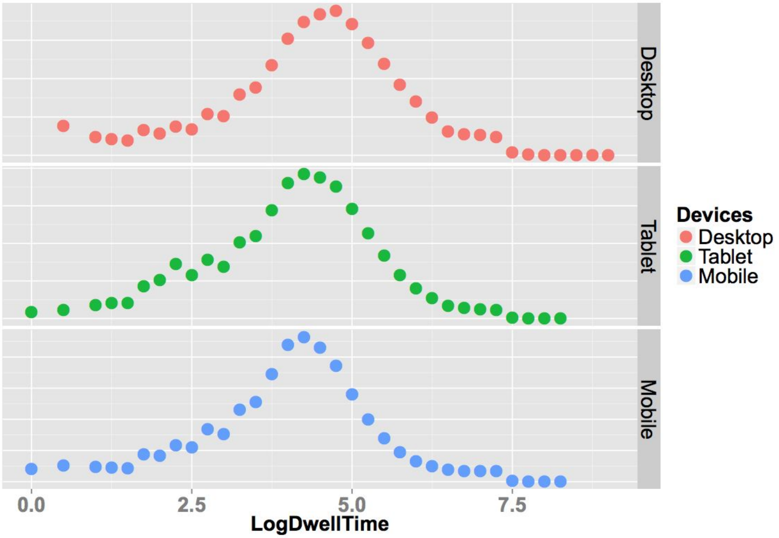


Figure 2: The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

Offline Experiment and Evaluation

Beyond Clicks: Dwell Time in Personalization

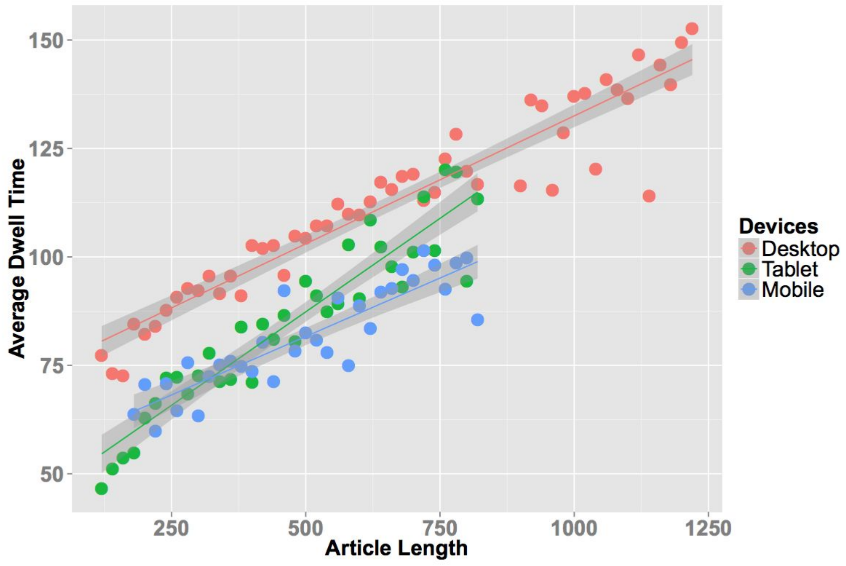


Figure 3: The relationship between the average dwell time and the article length where X-axis is the binned article length and the Y-axis is binned average dwell time.

Offline Experiment and Evaluation

Beyond Clicks: Dwell Time in Personalization

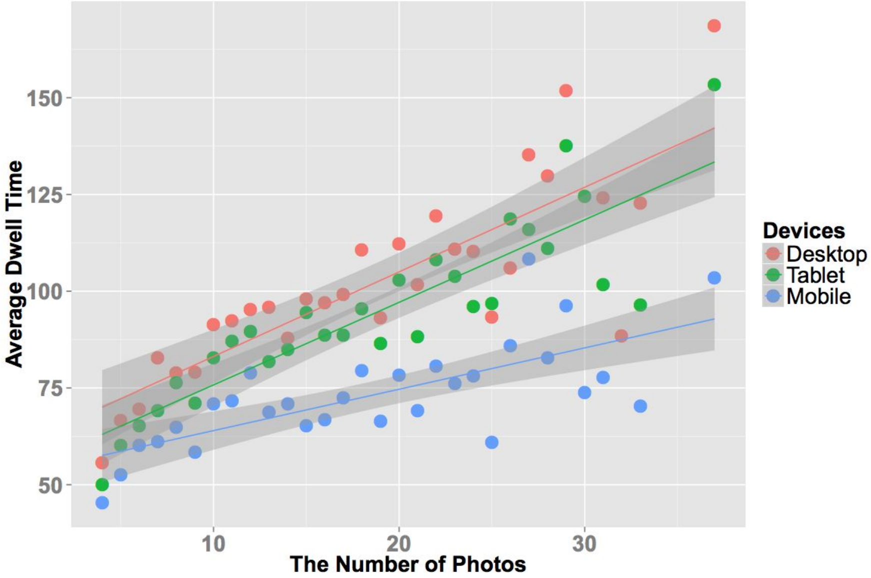


Figure 4: The relationship between the average dwell time and the number of photos on a slideshow where X-axis is the binned number of photos and the Y-axis is binned average dwell time.

Offline Experiment and Evaluation

Beyond Clicks: Dwell Time in Personalization

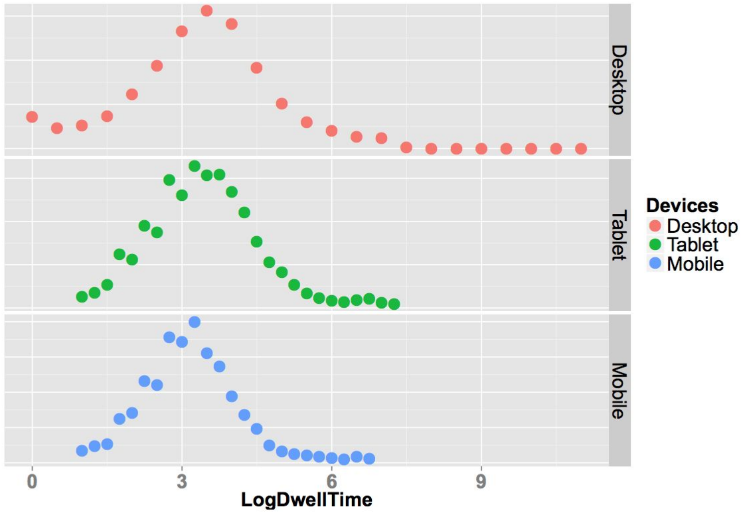


Figure 5: The (un)normalized distribution of log of dwell time for slideshows across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

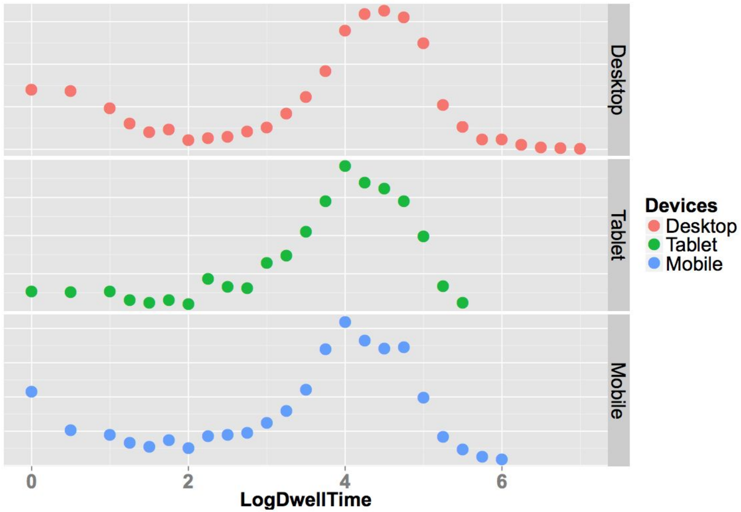


Figure 6: The (un)normalized distribution of log of dwell time for videos across different devices. The X-axis is the log of dwell time and the Y-axis is the counts.

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

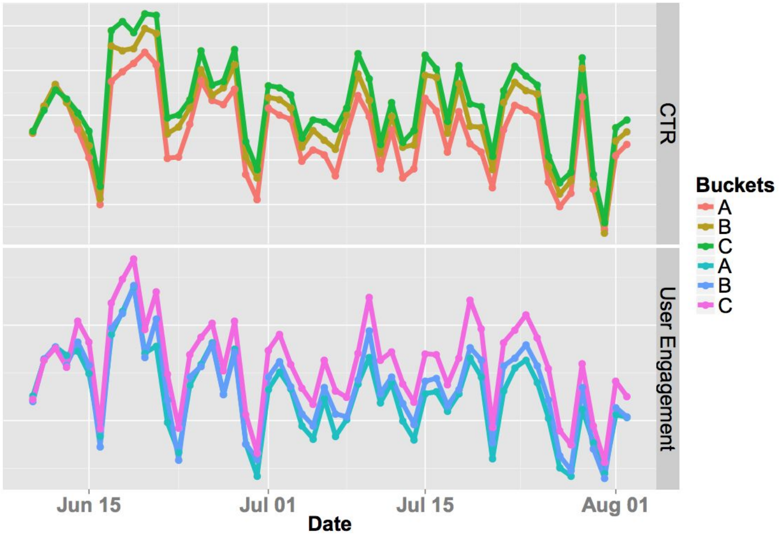


Figure 7: The relative performance comparison between three buckets. The top figure shows the relative CTR difference and the bottom figure shows the relative user engagement difference.

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

Summary

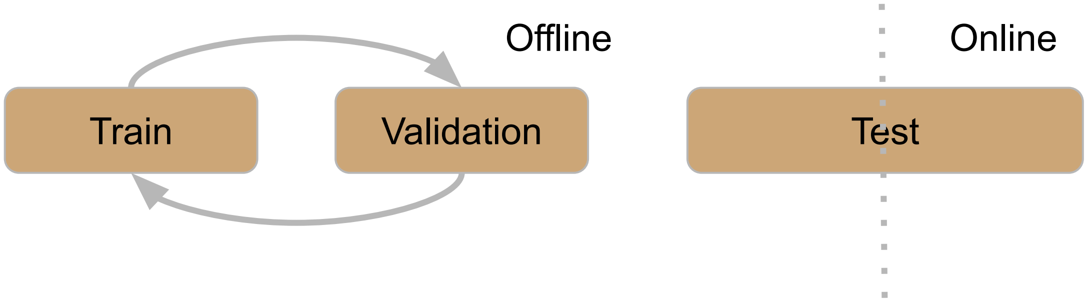
- **Approach I, Direct Optimization**
- **Approach II, Correlation and Optimization**

Offline Experiment and Evaluation

It doesn’t work or it doesn’t work smoothly.

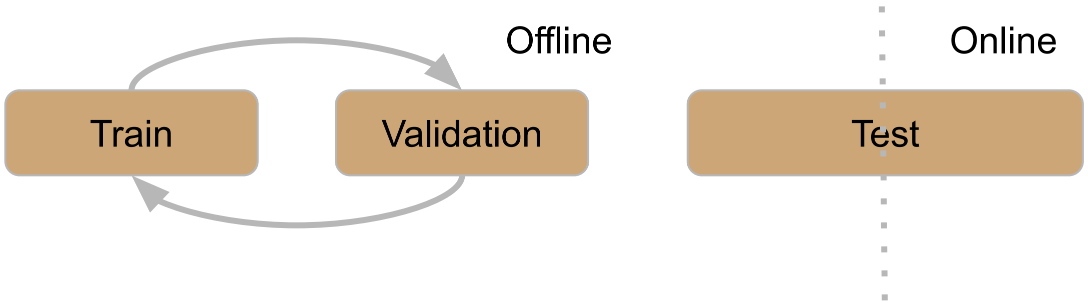
Offline Experiment and Evaluation

- **Bias**
  Examples: presentation bias, system bias...



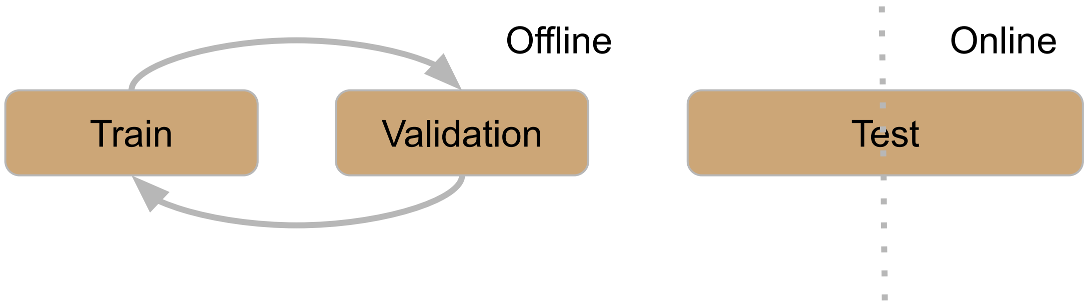
Offline Experiment and Evaluation

- **Concept Drifts**
  Examples: seasonal, interest shift...



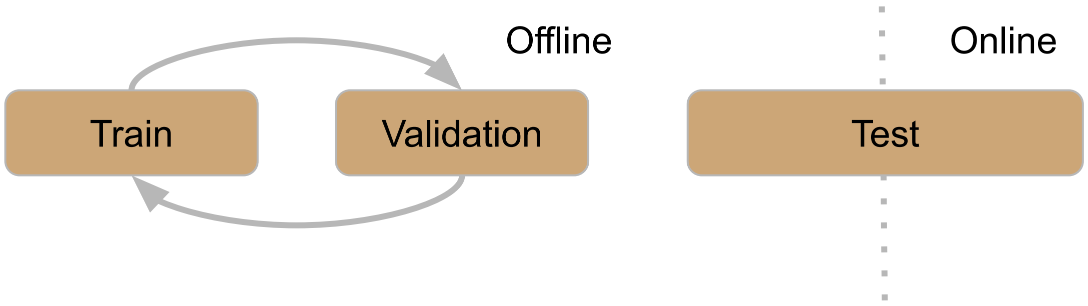
Offline Experiment and Evaluation

- **Different of offline metrics and online metrics**
  Examples: AUC/nDCG versus DAU...



Offline Experiment and Evaluation

- **Bias**
- **Concept Drift**
- **Different of offline metrics and online metrics**



Offline Experiment and Evaluation

- **Selection/sampling bias**
  e.g. presentation bias, system bias
- **Correlation**
  e.g. hard to control everything
- **Static**
  e.g., temporal dynamics, lacking "new" user behaviors

Offline Experiment and Evaluation

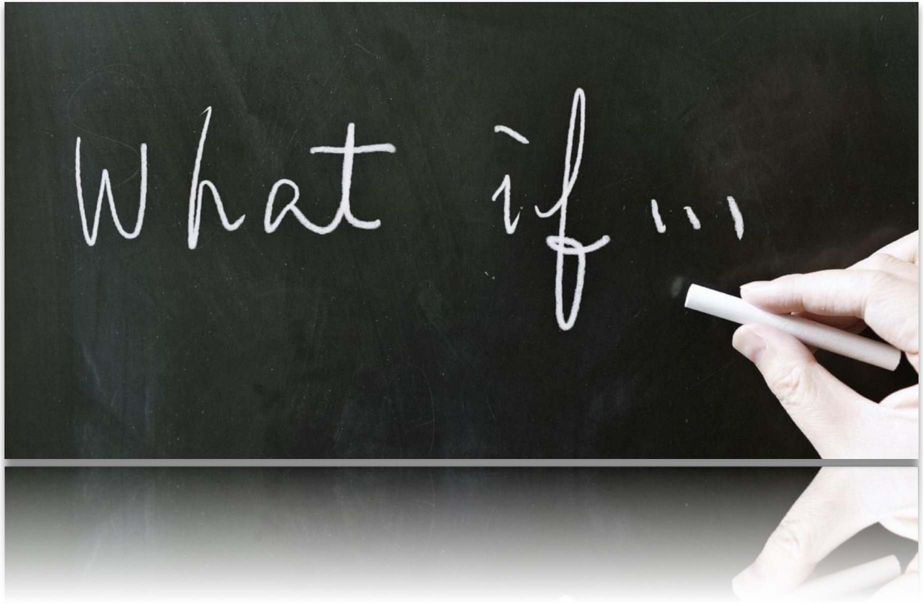
Summary

- Indirect and can be reused
- Good machine learning theories
- Correlation
- Static

[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.

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

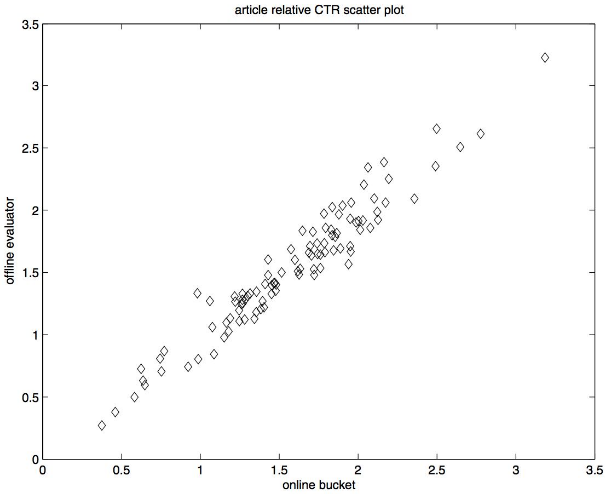


Figure 2: Articles' CTRs in the online bucket versus offline estimates.

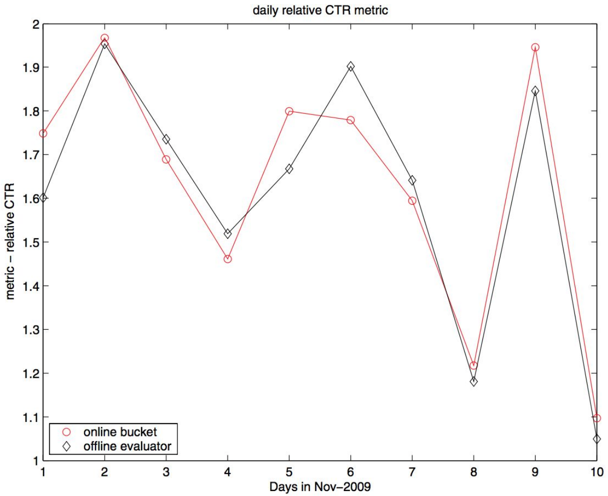


Figure 3: Daily overall CTRs in the online bucket versus offline estimates.

Offline A/B Experiment and Evaluation

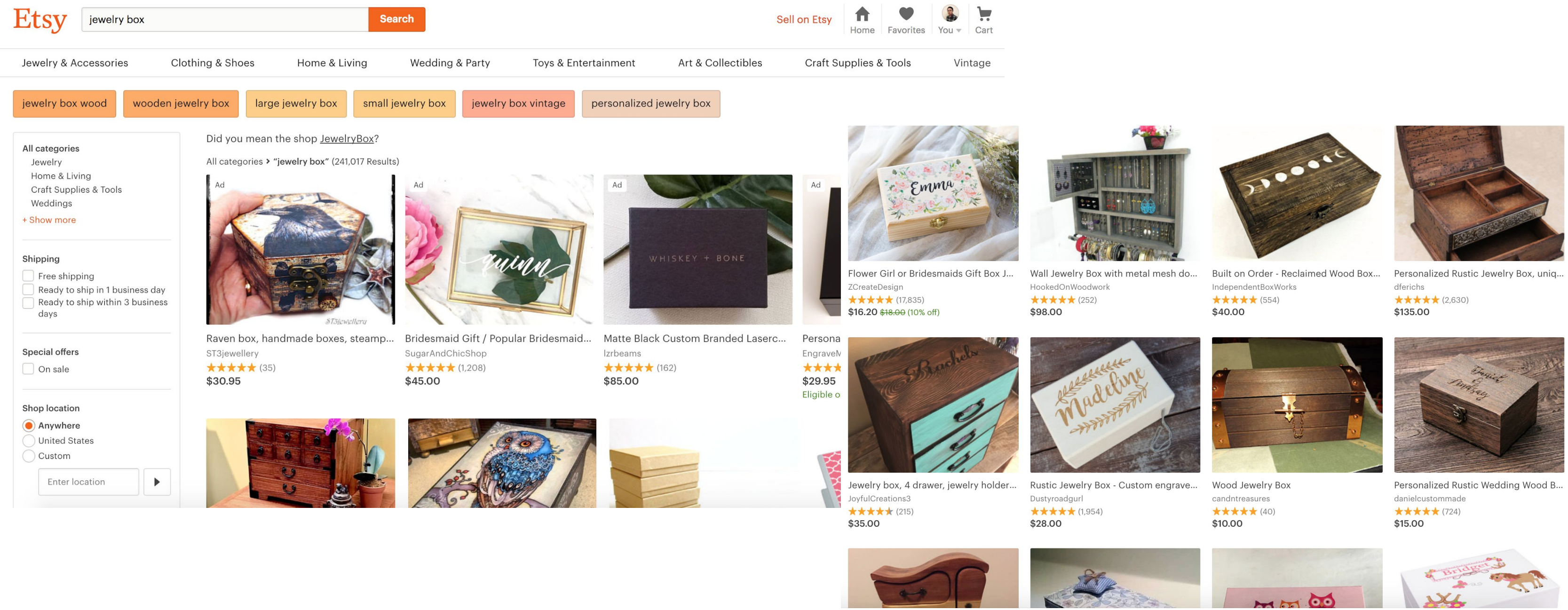
Counterfactual Offline Reasoning/Experiment

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

Offline A/B Experiment and Evaluation

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

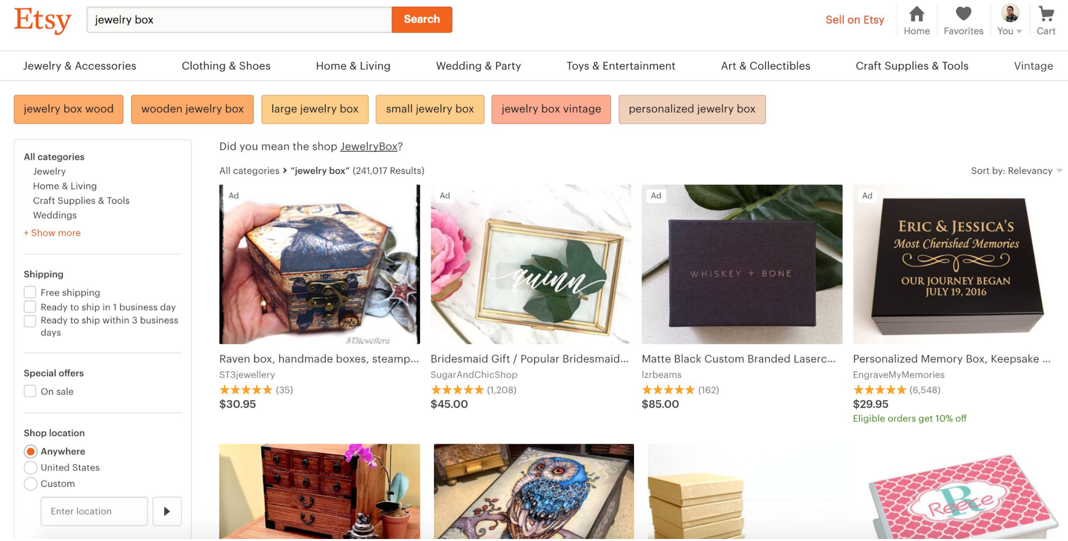


Offline A/B Experiment and Evaluation

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

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



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: α, β prior parameters of a Beta distribution
$S_i = 0$ and $F_i = 0, \forall i$ {Success and failure counters}
for $t = 1, \cdots, T$ **do**
  for $i = 1, \cdots, K$ **do**
    Draw θ_i according to Beta$(S_i + \alpha, F_i + \beta)$.
  end for
  Compute p such that $p_k = \frac{\theta_k}{\sum \theta_k}$**.**
  Sample N **items from 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?

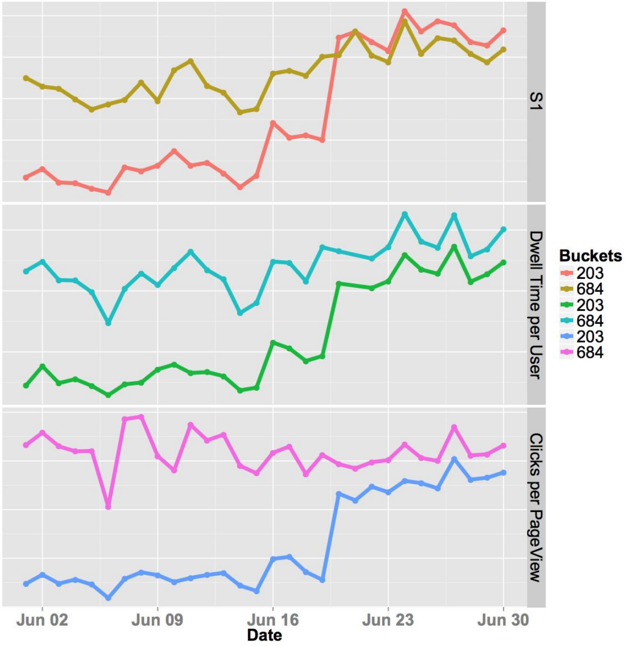
Algorithm 3 Thompson Sampling for Bernoulli Ranked-list Bandit

Require: α, β prior parameters of a Beta distribution
$S_i = 0$ and $F_i = 0, \forall i$ {Success and failure counters}
for $t = 1, \cdots, T$ **do**
  for $i = 1, \cdots, K$ **do**
    Draw θ_i according to Beta$(S_i + \alpha, F_i + \beta)$.
  end for
  Compute p such that $p_k = \frac{\theta_k}{\sum \theta_k}$**.**
  Sample N **items from Mult.(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	View Distribution	9.65	2,328.70	441.50
New Algorithm	Click Distribution	14.46	1,059.25	64.00
Old Algorithm	Click Distribution	14.64	241.17	7.00
New Algorithm	CTR Distribution	2.28	0.04	0.03
Old Algorithm	CTR Distribution	3.87	0.03	0.02
New Algorithm	Item Cold-Start Distribution	1.15	37.26	13.86
Old Algorithm	Item Cold-Start Distribution	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.

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.

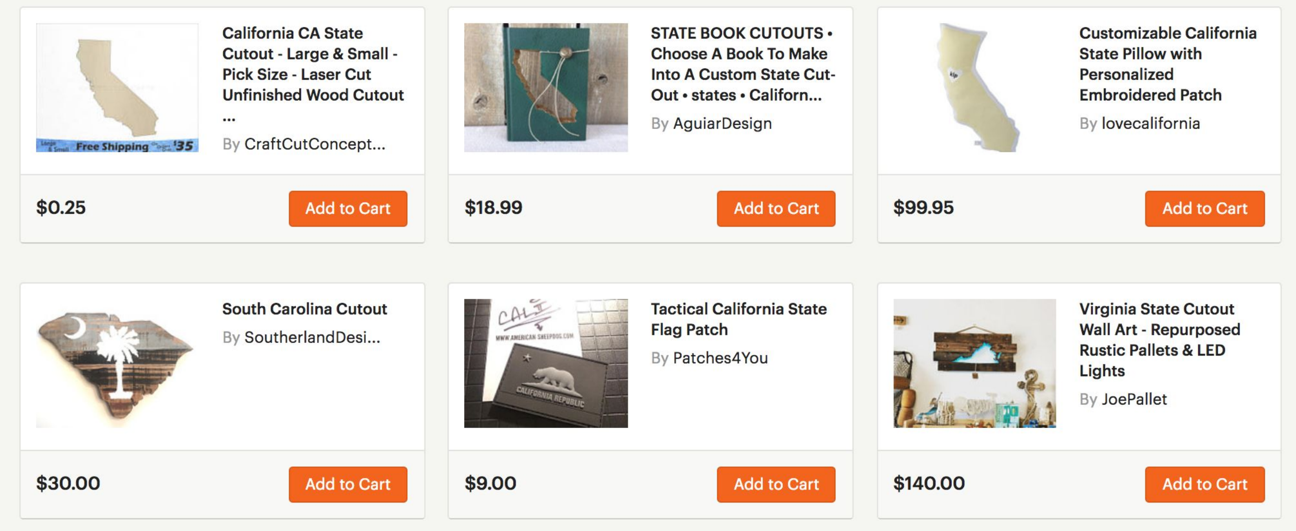
Observational Study

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

Observational Study

Sometimes, experiments may not be feasible or practical.

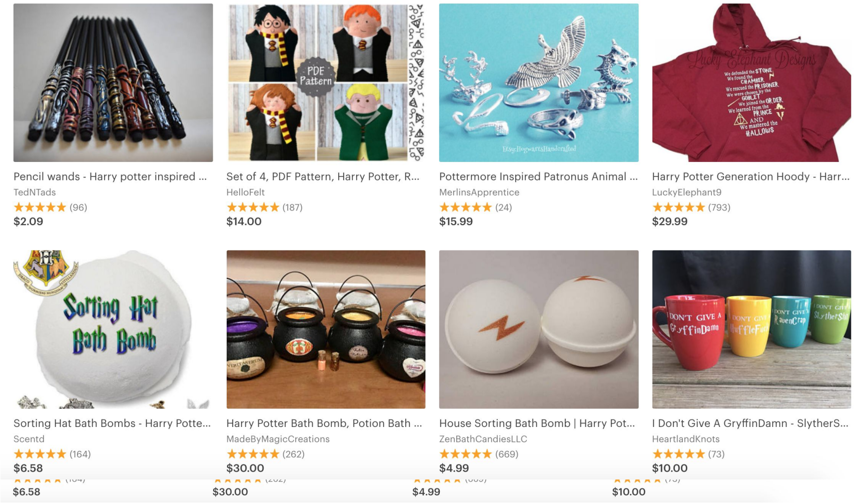
- **Example 1**:
  We want to test which “Add to Cart” button may lead to more <u>Monthly-Active-Users</u> (MAUs).

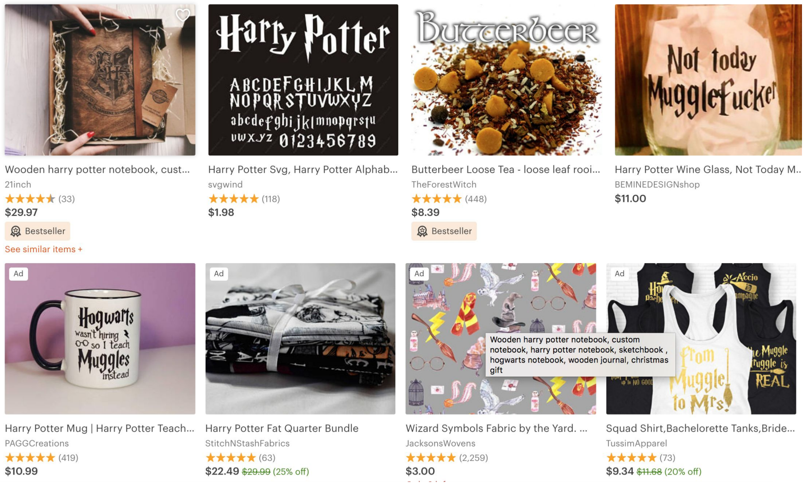


Observational Study

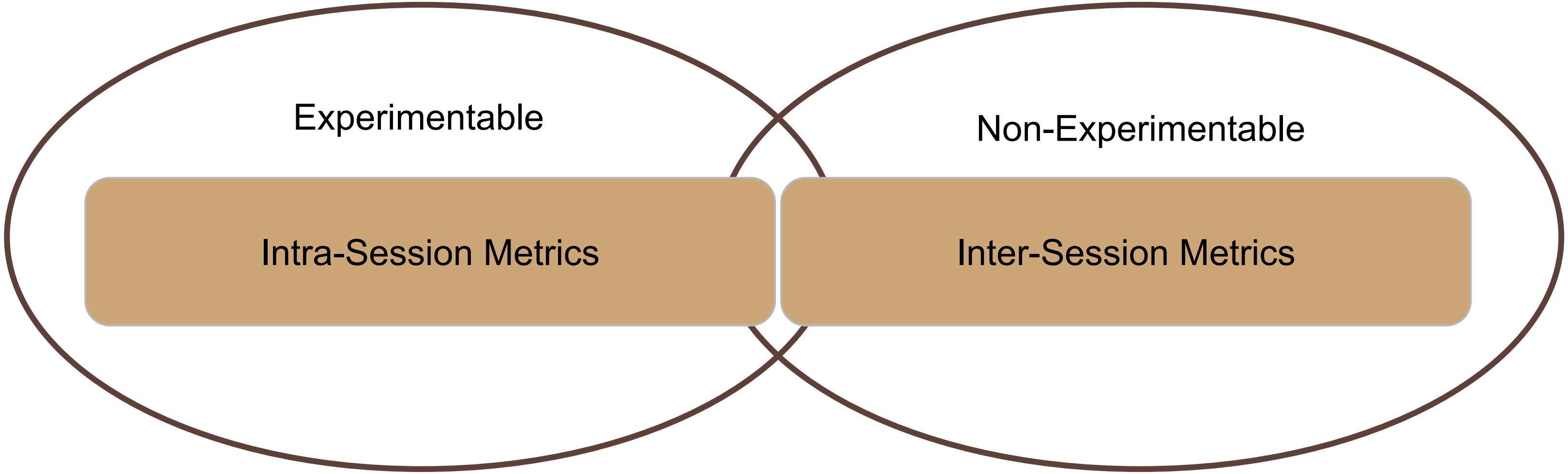
Sometimes, experiments may not be feasible or practical.

- **Example 2**:
  We want to test which search ranking algorithm may lead to higher Year-Over-Year Changes of user search sessions.





Observational Study



Causal Inference

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

[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.

Experiments v.s. Observational Study

Summary

- Run experiments as much as possible.
- Understand experimentable and non-experimentable.

Experiments v.s. Observational Study

Summary

- Run experiments as much as possible.
- Understand experimentable and non-experimentable.

- **Bias**: almost always indicates temporal, spatial and population sampling.
- **Conclusions**: almost always needs inference.

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

- 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 $t = 1, 2, \ldots$

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 α, β, λ_1, λ_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

H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson Tom Boulos, and Jeremy Kubica. **Ad click prediction: a view from the trenches.** KDD 2013.

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

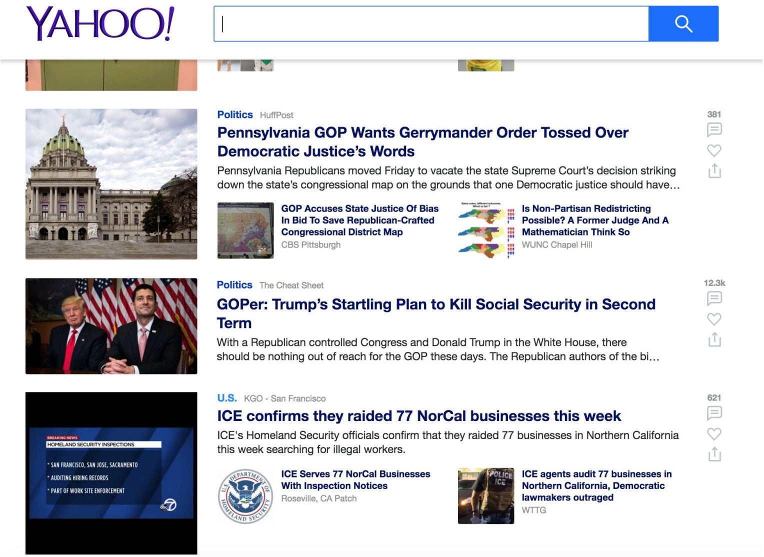
Formally, we define by $\mathcal{A} = \{1, 2, \ldots, K\}$ a set of K arms, and a contextual-bandit algorithm A interacts with the *world* in discrete trials $t = 1, 2, 3, \ldots$. In trial t:

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 $(r_{t,1}, \ldots, r_{t,K})$ the reward vector. Furthermore, we assume $(\mathbf{x}_t, r_{t,1}, \ldots, r_{t,K})$ 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 *accumate 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_i, 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_i, 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_i, 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^2Bandit

```
Inputs: η > 0, τ > 0, δ_1 ∈ (0, 1)
for i = 1 to N do
    Receive user u
    Record current timestamp t_{u,i}
    if user u is new: then
        Set A_{u,1} ← ηI, θ̂_{u,1} ← 0^d, β̂_{u,1} ← 0^d, ε̂_{u,1} ~ U(0, 1);
    else:
        Compute return interval Δ_{u,i−1} = t_{u,i} − t_{u,i−1}
        Update β̂_{u,i} in user return model using MLE.
    end if
    Observe context vectors, x_a ∈ ℝ^d for ∀a ∈ I(t_{u,i})
    Make recommendation a_{u,i} = arg max_{a∈I(t_{u,i})} P(C_{u,i} = 1|x_a, θ̂_{u,i}) + ε̂_{u,i} P(Δ_{u,i} ≤ τ|x_a, β̂_{u,i}) + α_{u,i} ||x_a||_{A^{-1}_{u,i}}
    Observe click C_{u,i}
    A_{u,i+1} ← A_{u,i} + x_{a_{u,i}} x^T_{a_{u,i}}
    Update θ̂_{u,i+1} in user click model using MLE.
    Update ε̂_{u,i+1} = Σ_{j≤i} C_{u,j}/i
end for
```

Multi-armed Bandits

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

Algorithm 1 r^2Bandit

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

Multi-armed Bandits

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

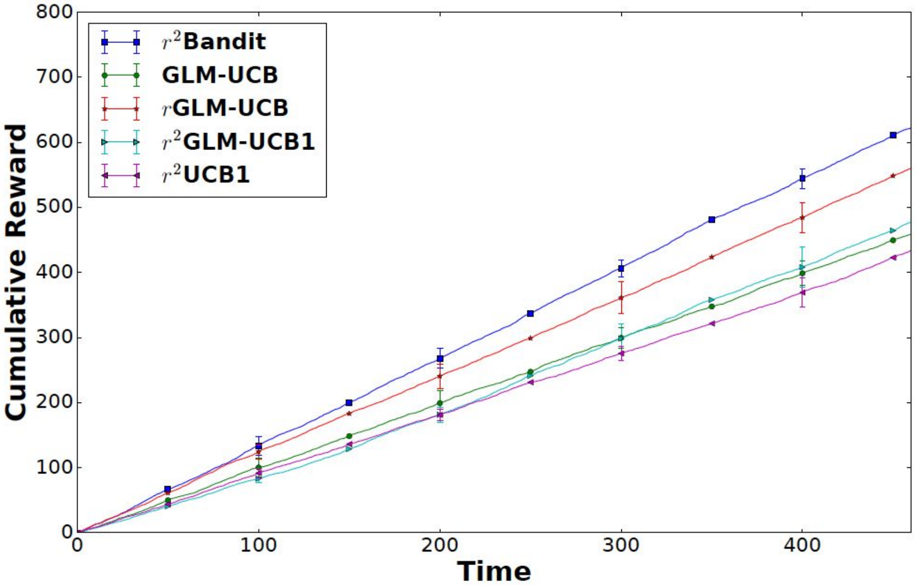
Simulations

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

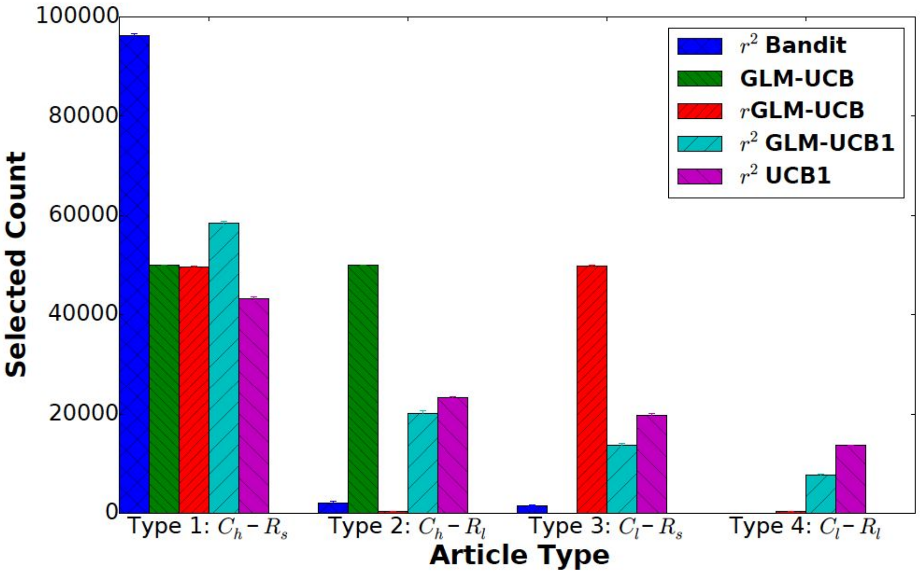


(a) Cumulative clicks over time

Multi-armed Bandits

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

Simulations

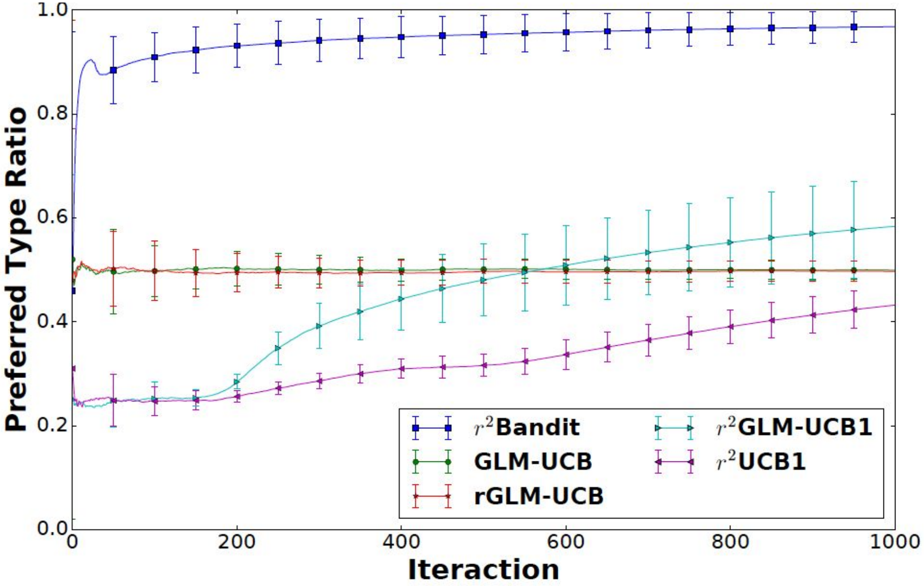


(b) Distribution of selected item types

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

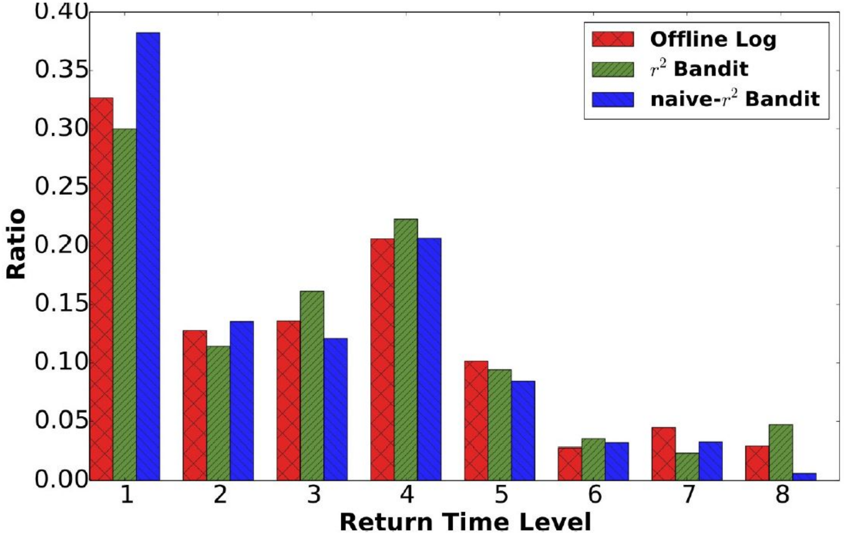


Figure 2: Discretized user return time distribution.

Multi-armed Bandits

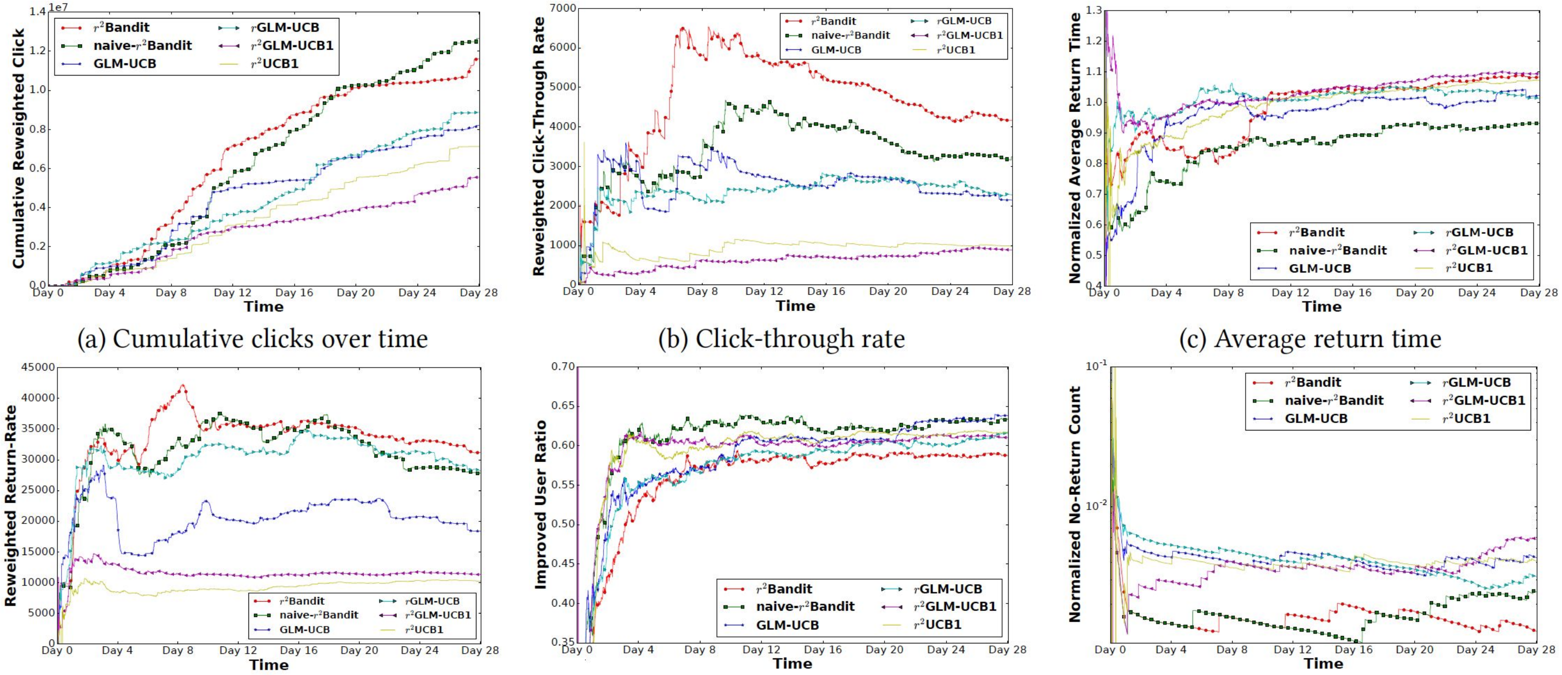
How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

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



(a) Cumulative clicks over time (b) Click-through rate (c) Average return time

(d) Return rate (e) Improved user ratio (f) No return count

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.

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),
- $P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ 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 π 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 γ is the discount factor and satisfies $0 \leq \gamma \leq 1$. (For example, $\gamma = 1/(1+r)$ when the discount rate is r.) γ is typically close to 1.

Reinforcement Learning

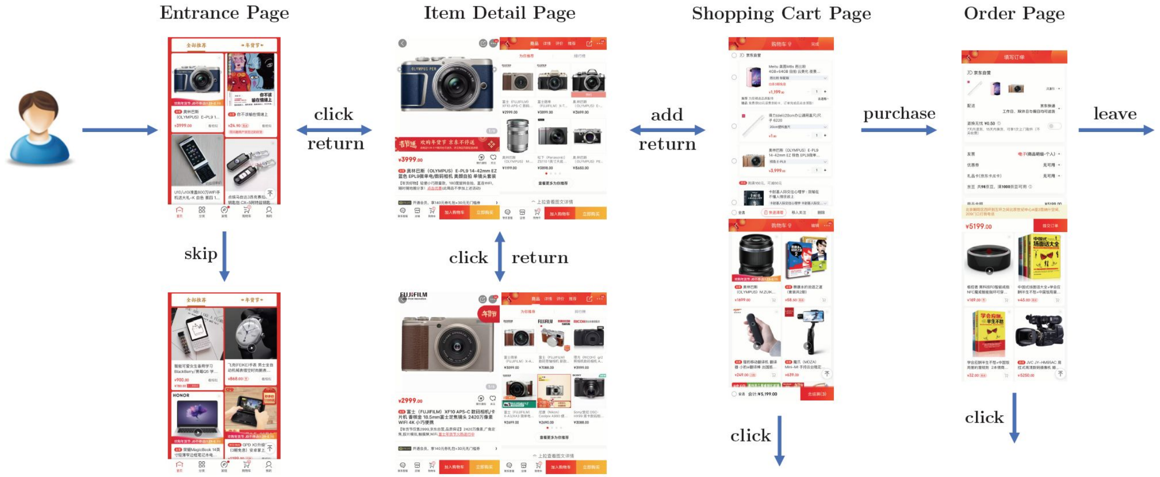


Figure 1: An example of whole-chain recommendations.

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.** CoRR abs/1902.05570, 2019.

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

[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.

Combining Two Camps

Two Main Camps of Optimization

- **Manual/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/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.

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

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 δ, 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 termiante with acceptance of H_0.

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

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

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

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 ρ and $1 - \rho$. To this end we design an **imputed Girshik Test** with the imputed log likelihood ratio test statistic

$$\widehat{Z_t} = \widehat{\log\left(\frac{p_{1t}}{p_{0t}}\right)} = \underbrace{-\delta}_{\text{risk tolerance}} \times \frac{\overbrace{2mn}^{\text{effective sample size}}}{t} \times \underbrace{(\overline{Y_n} - \overline{X_m})}_{\text{difference in empirical averages}}$$

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

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

Imputed Girshit 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 ρ_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

$$\mathbb{P}(p1 > p2|X_{1:t}) = \int \mathbb{I}(p1 \geq p_2)\pi(p_1, p_2|D_{1:t})dp_1 dp_2.$$

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 consisten but no longer unbiased.

$$\widetilde{Z}_t = \widetilde{\log\left(\frac{p_{1,t}}{p_{0,t}}\right)} = (-\delta) \times \underbrace{\sqrt{mn}}_{\text{effective sample size}} \times \left(\overline{Y_n} - \overline{X_m}\right).$$

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**. In WSDM 2019.

Combining Two Camps

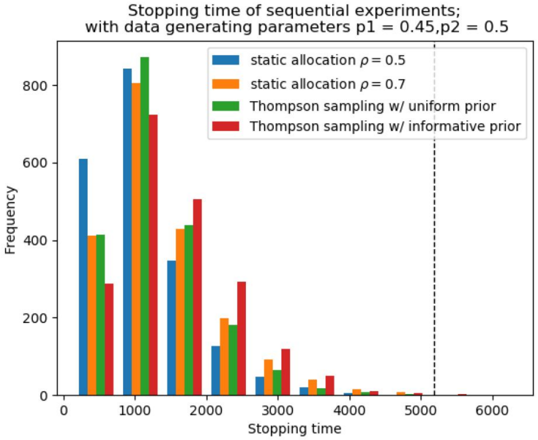


Figure 4: A histogram of stopping times for the imputed sequential Girshick test using different allocation schemes, corresponding to Table 1. The dashed black line is the sample size required by a fixed-time proportion test. There is a vanishingly small number of simulations where the sequential test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

	static allocation		Thompson sampling	
	$\rho = 0.5$	$\rho = 0.7$	Unif. priors	inform. priors
$\mathbb{P}(\text{accept}\mid\omega_a)$	99.8 %	99.75%	97.7%	99.55%
average τ	1165.26	1383.86	1300.47	1537.59
min	186	148	263	235
median τ	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.

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 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.

Concluding remarks and future direction

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, revenue) when measuring online engagement.

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
- Optimizing for multiple (*possibly competing*) metrics
- Optimize under *FATE* (Fairness, Accountability, Transparency, and Ethics)

Thank you

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