




Tutorial on Metrics of User Engagement Applications to Search & E-Commerce

Mounia Lalmas & Liangjie Hong



Outline

1. Introduction and scope
2. Towards a taxonomy of metrics
3. Experimentation and evaluation of metrics
4. Optimisation for metrics
5. Applications
 - a. Search
 - b. E-commerce
6. Recap and open challenges
7. References ... to come

Acknowledgements

- This tutorial uses some material from a tutorial “**Measuring User Engagement**” given at **WWW 2013**, Rio de Janeiro (with Heather O’Brien and Elad Yom-Tov).
- M. Lalmas, H. O’Brien and E. Yom-Tov. “**Measuring User Engagement**”, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 2014.



Introduction and scope

Introduction and scope

Who we are

What is user engagement

Approaches to measure user engagement

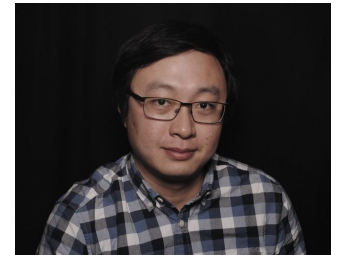
The focus of this tutorial

... Outline

Who we are

- Mounia Lalmas, Research Director at Spotify, London
 - Research interests: user engagement in areas such as advertising, digital media, search, and now music
 - Website: <https://mounia-lalmas.blog/>

- Liangjie Hong, Head of Data Science at Etsy, New York
 - Research interests: search, recommendation, advertising and now hand-craft goods
 - Website: <https://www.hongliangjie.com/>



What is user engagement?

... Some definitions

User engagement is regarded as a **persistent** and **pervasive** cognitive affective state, not a time-specific state (Schaufeli et al., 2002)

User engagement refers to the quality of the user experience associated with the **desire** to use a technology (O'Brien and Toms, 2008)

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** (Attfield et al., 2011).

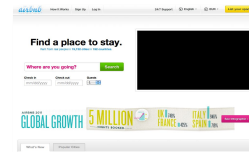
All the above can translate into the “emotional, cognitive and behavioural **connection** that exists, at any point in time **and** over time, between a user and a technological resource” (O'Brien, Lalmas & Yom-Tov, 2013)

Why is it important to engage users?

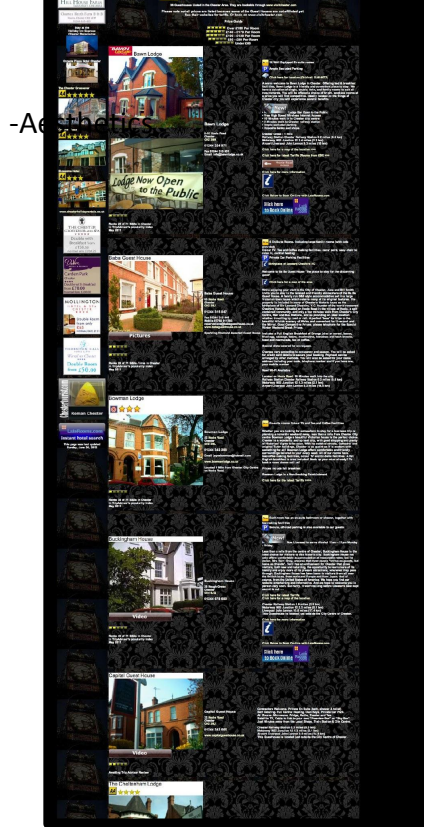
Users have increasingly enhanced expectations about their interactions with technology

... 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



Is this site engaging?



aesthetics

leisure

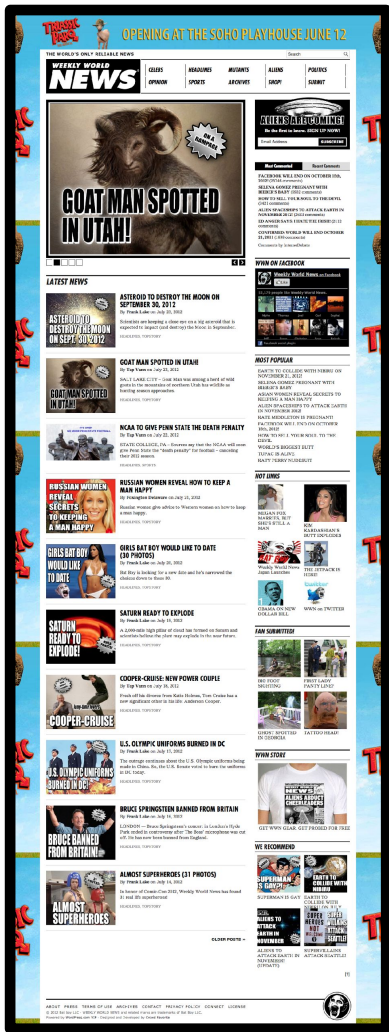
Is this site engaging?



usability

shopping

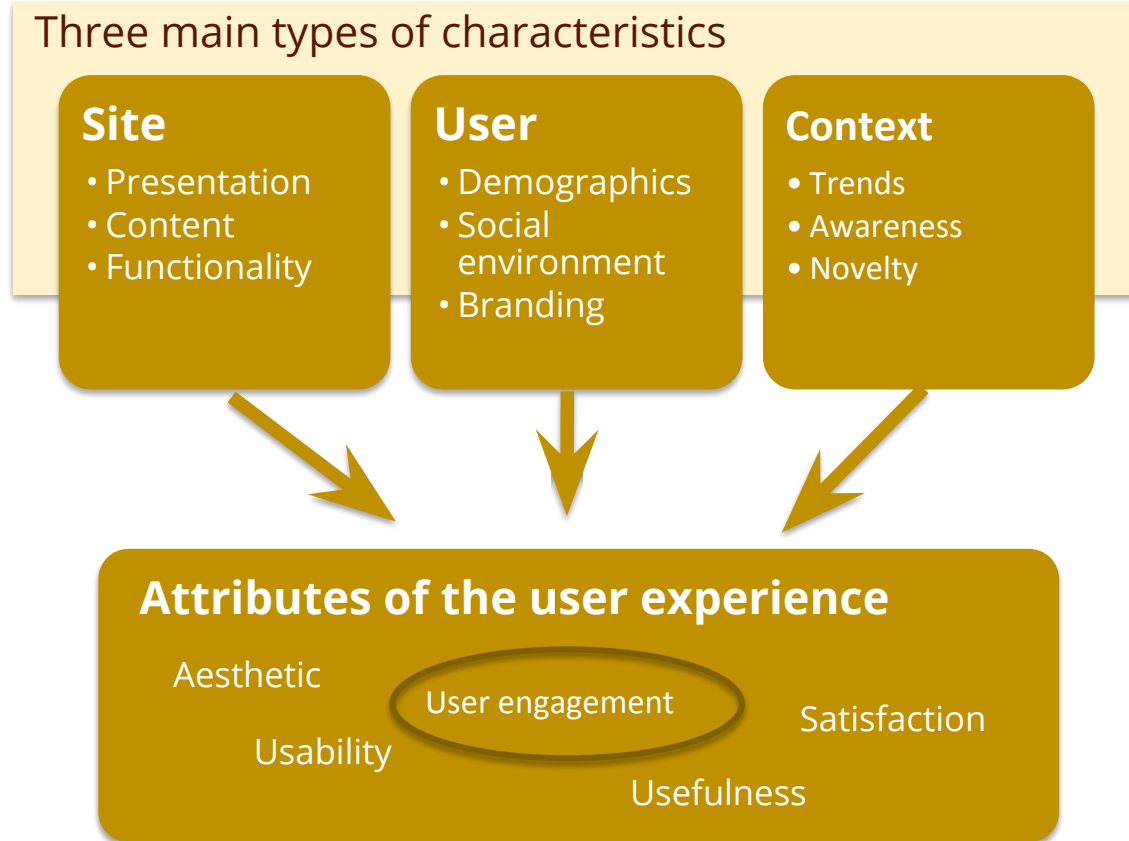
Is this site engaging?



trust

news

What influences user engagement?



Many connections

Considerations in measuring user engagement

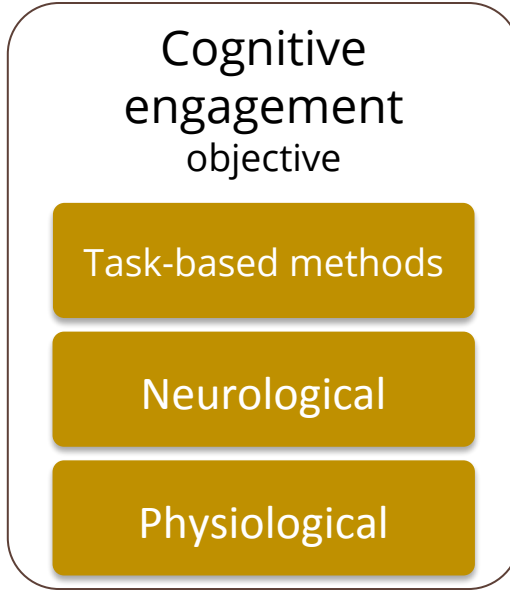
- short term \longleftrightarrow long term
- laboratory \longleftrightarrow “in the wild”
- subjective \longleftrightarrow objective
- qualitative \longleftrightarrow quantitative
- large scale \longleftrightarrow small scale



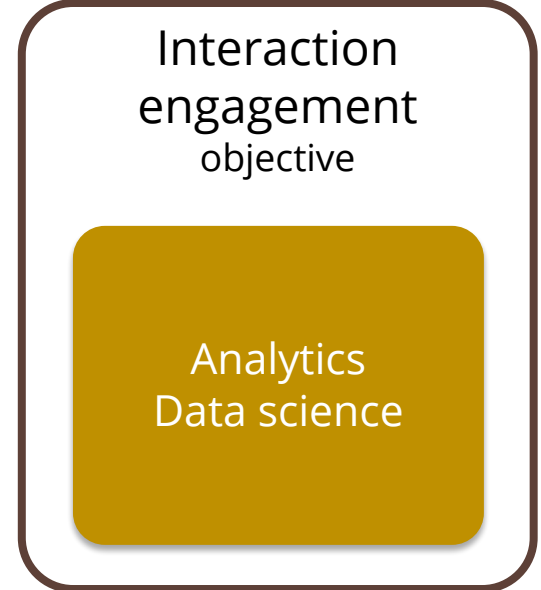
Methods to measuring user engagement



User study (lab/online)
mostly qualitative



User study (lab/online)
*mostly quantitative,
scalability an issue*



Data study
quantitative, large scale



Scope of this tutorial

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 all existing works.



Towards a taxonomy of metrics

Towards a taxonomy of metrics

Terminology, context & consideration

Facets of user engagement

Sessions and metrics

Intra-session metrics

Inter-session metrics

Other metrics

Proposed taxonomy

... Outline

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 on a site

Metric:

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

e.g. click-through rate

Key performance indicator (KPI):

quantifiable measure demonstrating how effectively key business objectives are being achieved

e.g. conversion rate

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

Three levels of metrics

Business metrics

-- KPIs

our focus in this section



Behavioral metrics

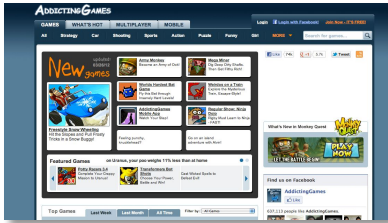
-- online metrics, analytics

Optimisation metrics

-- metrics used to train machine learning algorithms

These three levels are connected

Why do we need several metrics of online behaviour?



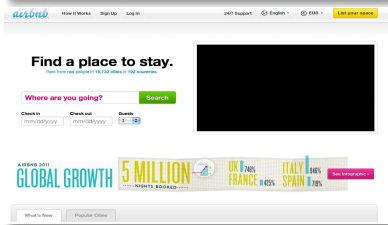
Games

Users spend much time per visit



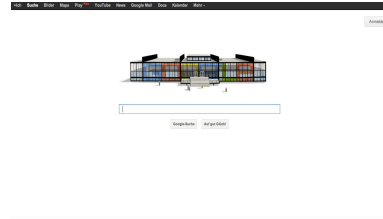
Social media

Users come frequently and stay long



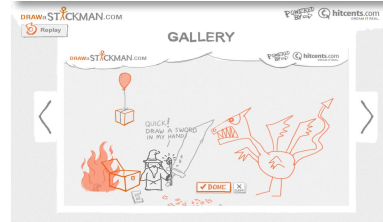
Service

Users visit site, when needed, e.g. to renew subscription



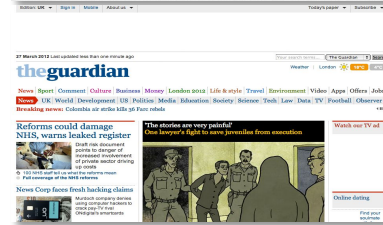
Search

Users come frequently and do not stay long



Niche

Users come on average once a week e.g. weekly post



News








Users come periodically, e.g. morning and evening

Sites differ in their patterns of engagement

A basic taxonomy of metrics

... A starting point

Capture various facets of engagement

Popularity	#Users	Number of distinct users	
	#Visits	Number of visits	
	#Clicks	Number of clicks	
Activity	Click Depth	Average number of page views per visit.	
	Dwell Time	Average time per visit	
Loyalty	#Active Days	Number of days a user visited the site	
	Return Rate	Number of times a user visited the site	

$$\tau_{intra} = 0.61$$

$$\tau_{inter} = 0.23$$

Sites differ in their patterns of engagement ... Indeed

80 sites, 2M users, 1 month sample

interest-specific

media (daily)

media (periodic)

e-commerce

search

	popularity	activity [ClickDepth]	activity [DwellTime]	loyalty
model m_{g6}	--			
model m_{g5}		--	--	++
model m_{g4}		--	++	--
model m_{g3}		++		
model m_{g2}	++			
model m_{g1}				

Some observations made as part of this study (nothing unexpected but metrics aligned well):

Activity depends on the structure and freshness of the site

Loyalty influenced by external and internal factors (e.g. freshness, current interests, bugs, events)

What may impact user engagement?

Why?

Task

Who?

Demographics

Recency

When?

Temporality

Usage level

Where?

View

Platform

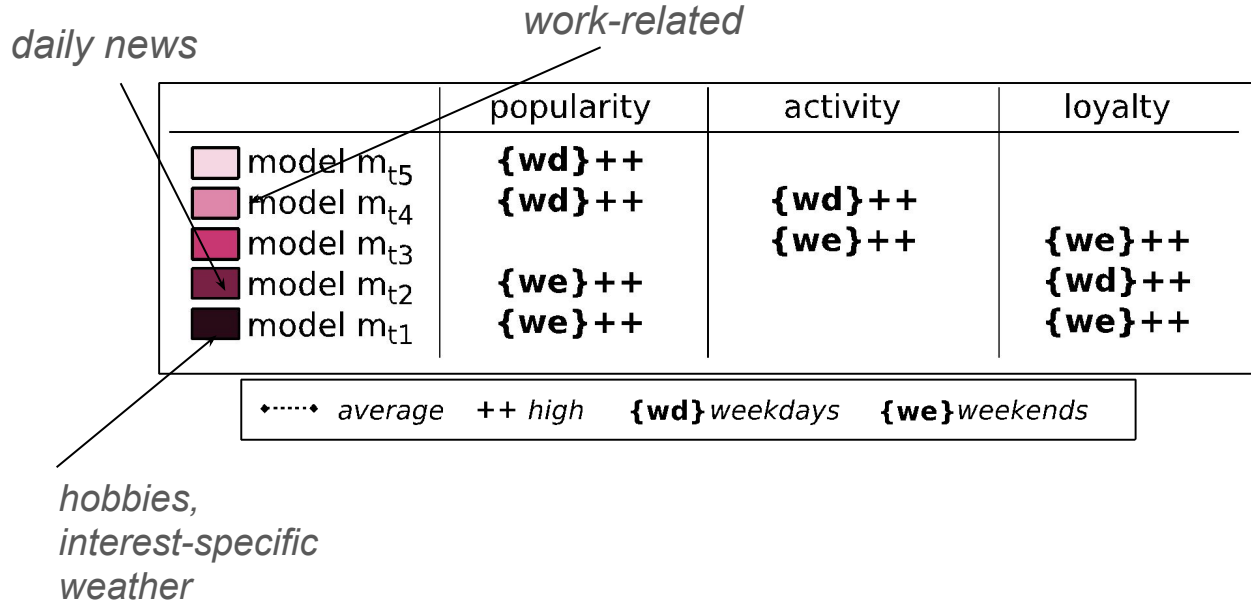
What?

Function

Segmentation

Temporality

... When?

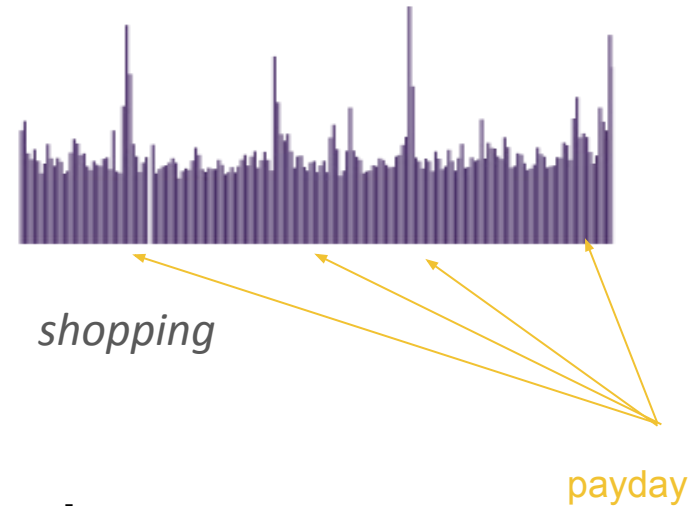
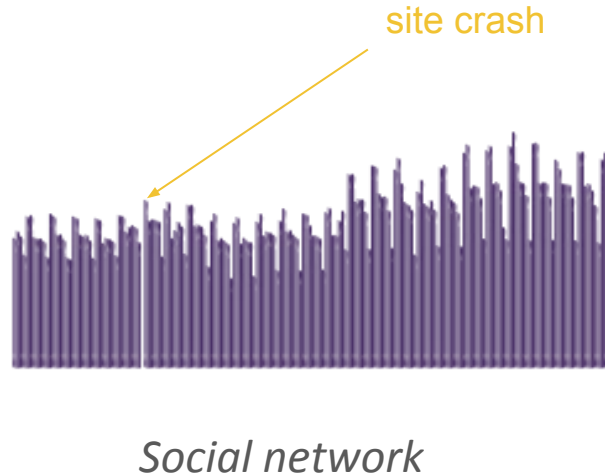


Engagement varies from weekdays to weekends

(Lehmann et al., 2012)

Temporality

... When?

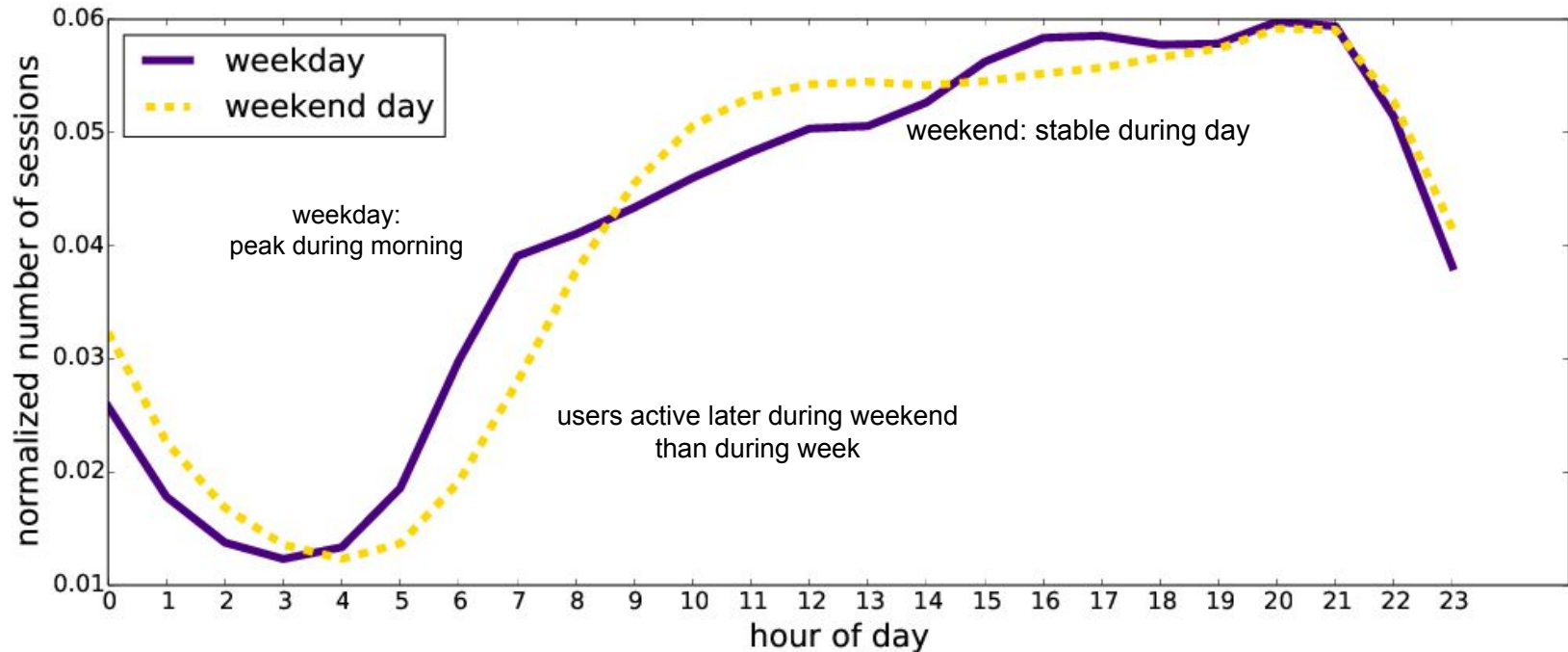


Engagement is **periodic** or may contain **peaks**

Engagement is influenced by internal (e.g. crash) and **external** factors (e.g. major events)

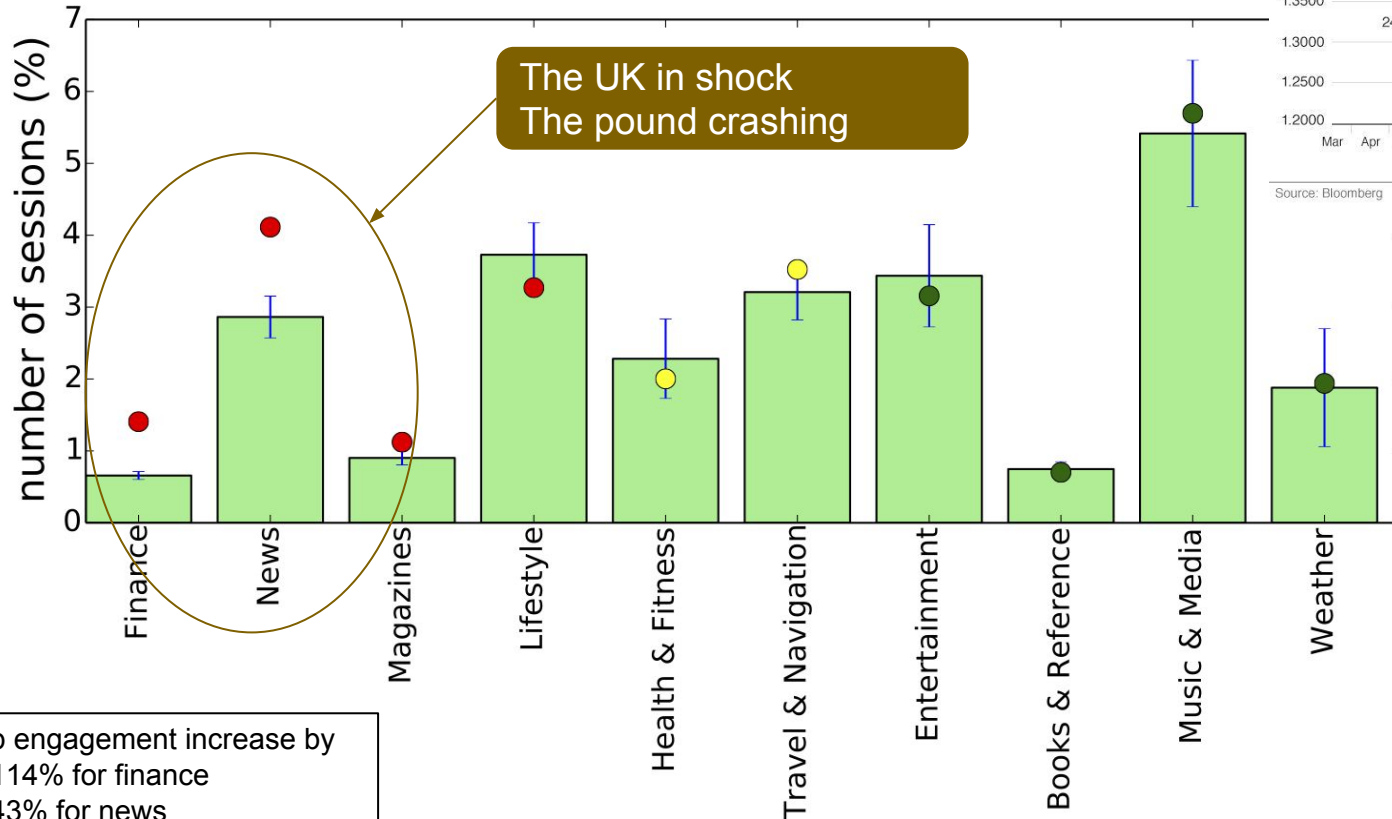
Periodicity (day)

... When?



External factors (news)

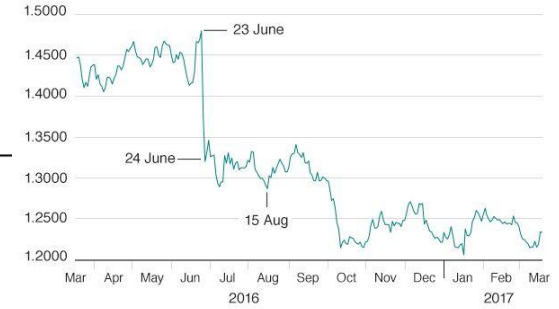
... When?



app engagement increase by
114% for finance
43% for news

Pound plunged against the dollar after vote result

How many dollars £1 buys



Source: Bloomberg



(Van Canneyt et al, 2017)

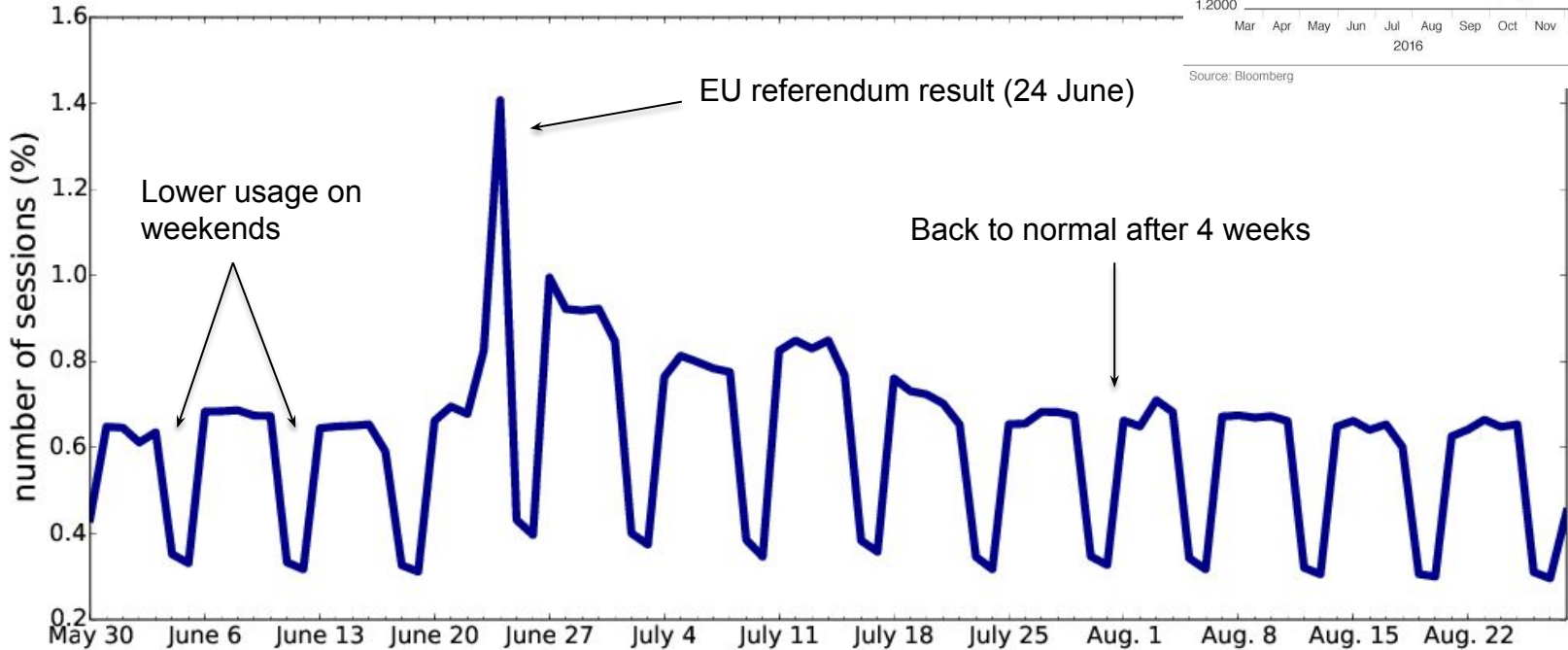
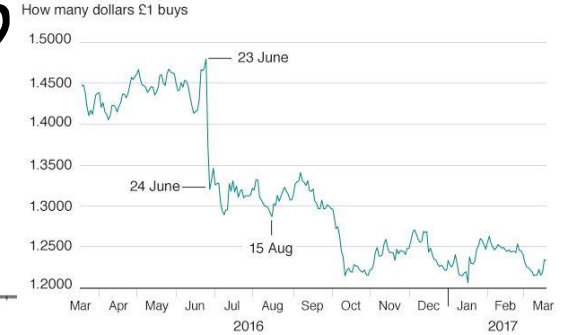
Day of the referendum result
24 June, 2016 (UK)

External factors (news)

... When?

Finance apps (Van Canneyt et al, 2017)

Pound plunged against the dollar after vote result

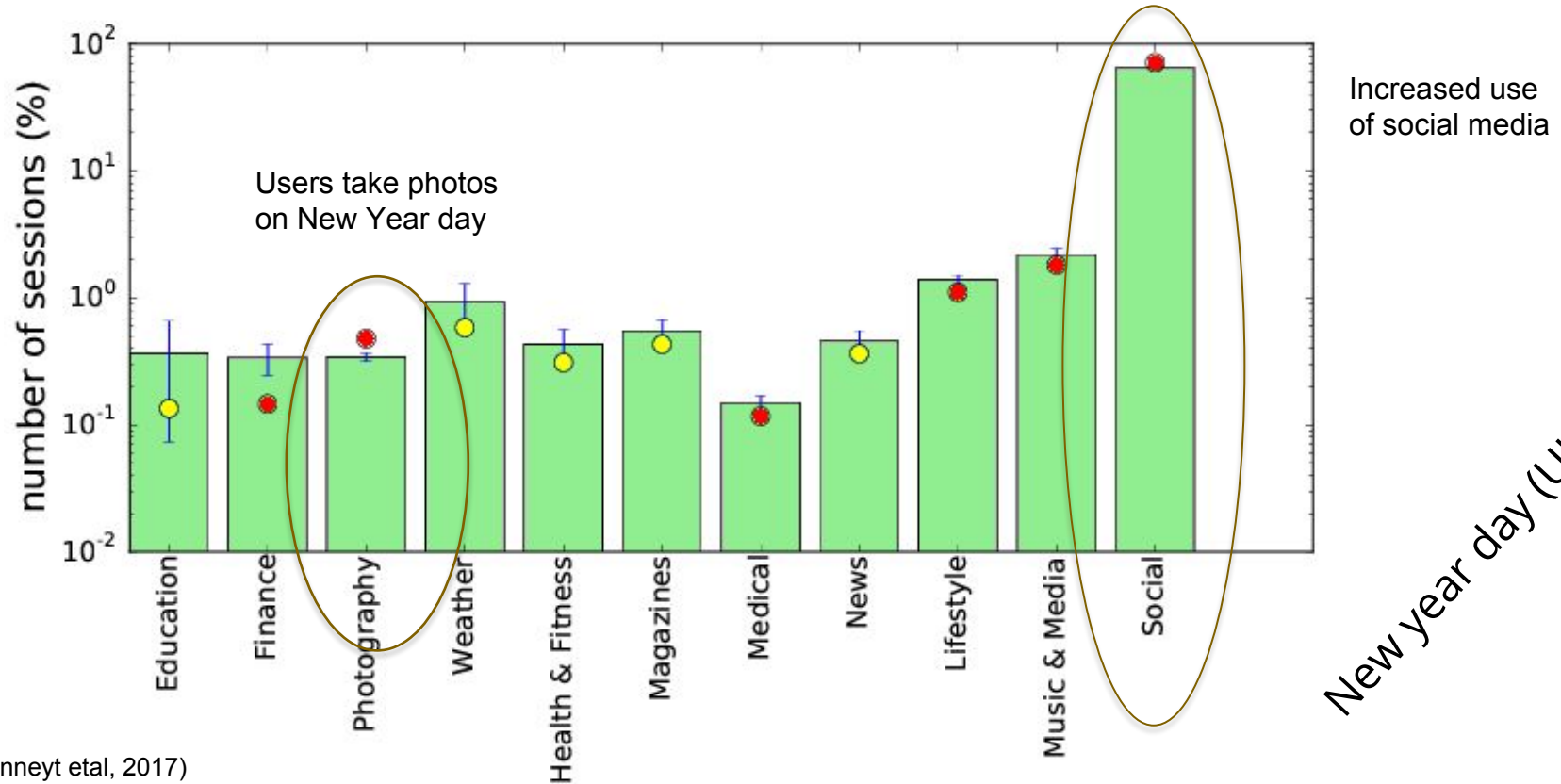


Source: Bloomberg

BBC

External factors (social)

... When?



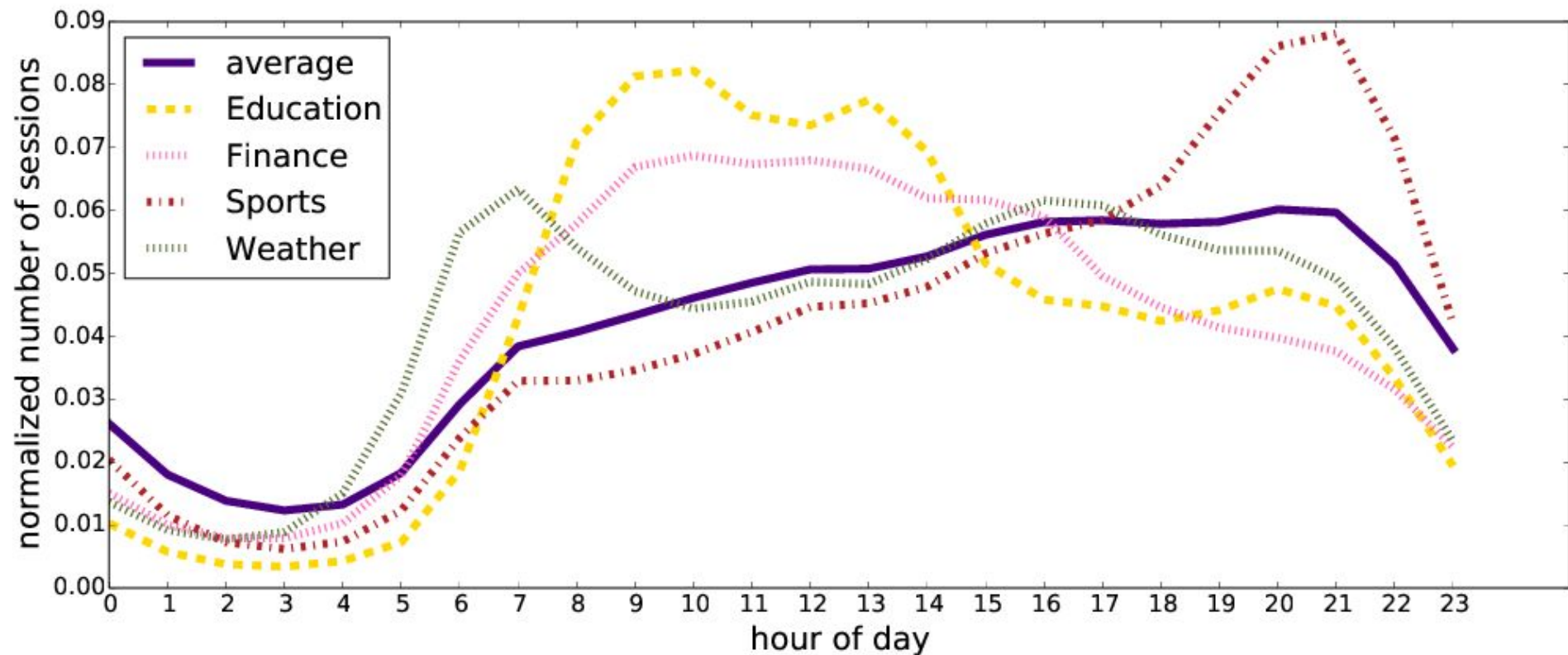
Task

... Why?

- Engagement varies by task
 - A user who accesses a site to check for emails (goal-specific task) has different engagement patterns from one browsing for leisure.
 - Task has an effect on periodicity
- In one study (Yom-Tov et al, 2013), sessions in which 50% or more of the visited sites belonged to the five most common sites (for each user) were classified as goal-specific.
 - Goal-specific sessions accounted for 38% of sessions
 - 92% of users have both goal-specific and non-goal-specific sessions
 - Average “downstream engagement” in goal-specific sessions was lower compared to non-goal-specific ones

Task (day)

... Why?

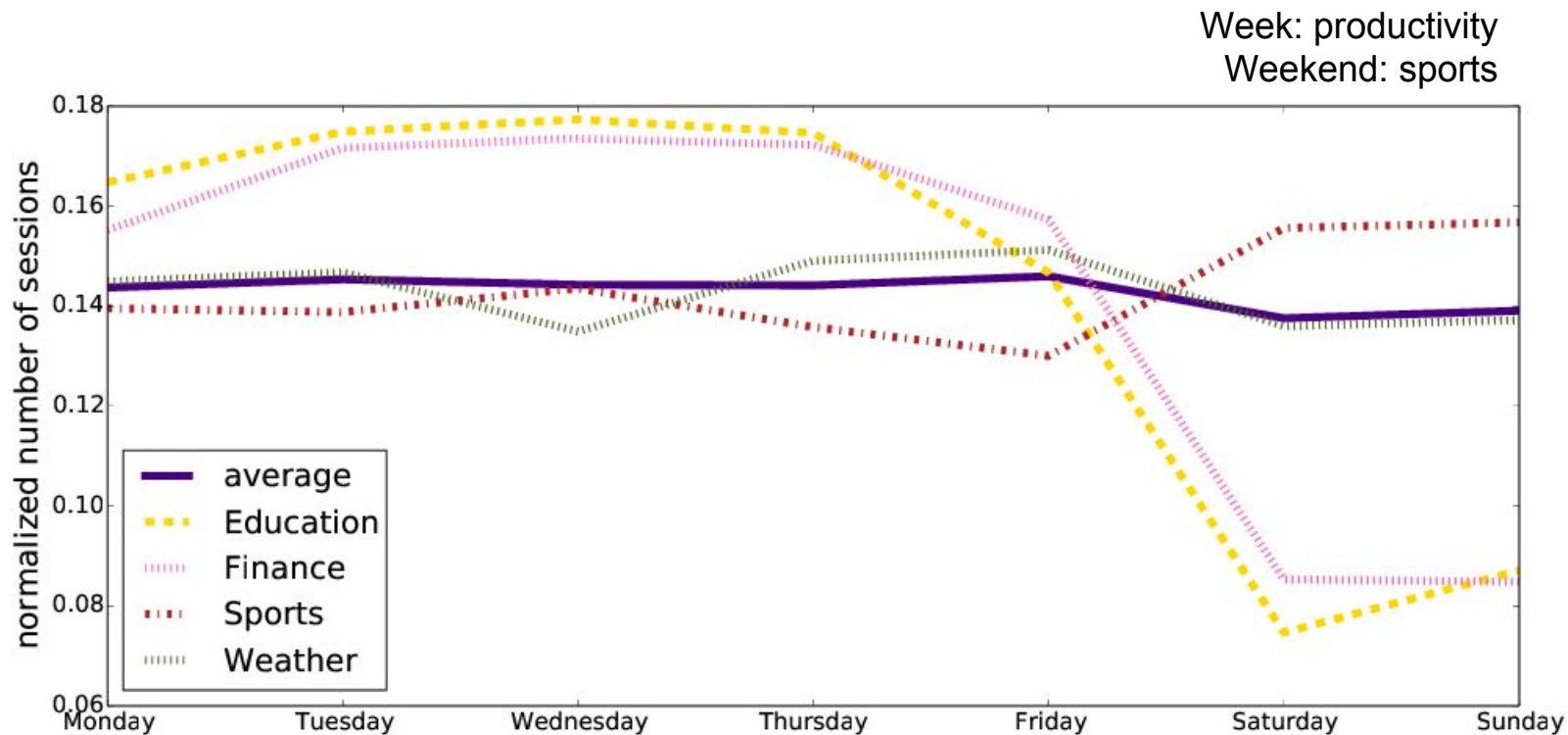


(Van Canneyt et al, 2017)

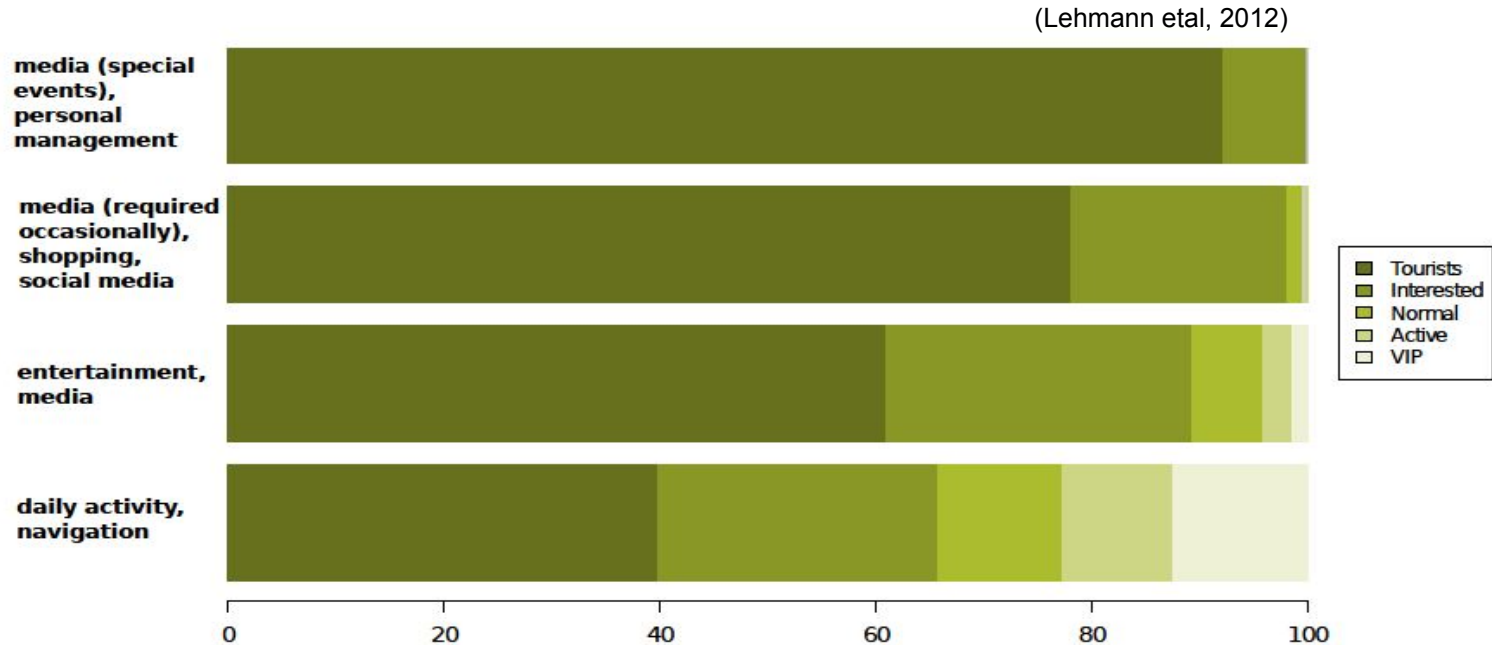
Task (week)

... Why?

(Van Canneyt et al, 2017)



Usage level (how often we see a user) ... When?



Various definitions of usage level (e.g. #days per month/week, #sessions per week)
Discard tourist users in analysis → unless the focus is on new users

Facets of engagement

... Several proposals

Factor

focus attention; positive affect; aesthetics; durability; novelty; richness & control; reputation, trust & expectation; motivation, interests, incentives & benefits

Degree

involvement, interaction, intimacy, influence

Process

point of engagement, period of engagement, disengagement, re-engagement

Index

click depth, duration, recency, loyalty, brand, feedback, interaction

Factor of user engagement (I)

Focused attention (Webster & Ho, 1997; O'Brien, 2008)

- Users must be focused to be engaged
- Distortions in subjective perception of time used to measure it

Positive Affect

(O'Brien & Toms, 2008)

- Emotions experienced by user are intrinsically motivating
- Initial affective "hook" can induce a desire for exploration, active discovery or participation

Aesthetics

(Jacques et al, 1995; O'Brien, 2008)

- Sensory, visual appeal of interface stimulates user and promotes focused attention; perceived usability
- Linked to design principles (e.g. symmetry, balance, saliency)

Endurability

(Read, MacFarlane, & Casey, 2002; O'Brien, 2008)

- People remember enjoyable, useful, engaging experiences and want to repeat them
- Repetition of use, recommendation, interactivity, utility

Factor of user engagement (II)

Novelty

(Webster & Ho, 1997; O'Brien, 2008)

- Novelty, surprise, unfamiliarity and the unexpected; updates & innovation
- Appeal to user curiosity; encourages inquisitive behavior and promotes repeated engagement

Richness and control

(Jacques et al, 1995; Webster & Ho, 1997)

- Richness captures the growth potential of an activity
- Control captures the extent to which a person is able to achieve this growth potential

Reputation, trust and expectation

(Attfield et al, 2011)

- Trust is a necessary condition for user engagement
- Implicit contract among people and entities which is more than technological

Motivation, interests, incentives, and benefits

(Jacques et al., 1995; O'Brien & Toms, 2008)

- Why should users engage?
- Friends using it

Degree of engagement

(Forrester Research, June 2008)

Involvement

- Presence of a user on the site
- Measured by e.g. number of visitors, time spent, revisit rate

Interaction

- Action of a user on the site
- Measured by e.g. CTR, online transaction, uploaded photos

Intimacy

- Affection or aversion of a user
- Measured by e.g. satisfaction rating, sentiment analysis on social media &, comments, surveys, questionnaires

Influence

- Likelihood that a user advocates
- Measured by e.g. forwarding & sharing, invitation to join

Process of user engagement

Point of engagement

(O'Brien & Toms, 2008)

- 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

(Webster & Ahuja, 2006; O'Brien & Toms, 2008)

- Engage again after becoming disengaged
- Triggered by relevance, novelty, convenience, remember past positive experience sometimes as result of campaign strategy

Point of engagement

... Process

Point of engagement relates to acquisition → how users arrive at a site

Which channels users are originating from:

organic search, direct targeting, paid search, referral,
social media, advertising campaign



- is about attracting & acquiring new users
- understand acquisition cost (e.g. important for marketing)

Period of engagement

... Process

Relates to user behavior with site → per page, per visit, per session

Involvement

pageview, dwell time, playtime (e.g. video)

Interaction

click-through rate, #shares, #likes, conversion rate, #save, bounce rate

Contribution

#blog posts, #comments, #create (e.g. playlist), #replies, #uploads (e.g. photo)

Note that
Interaction
(e.g. share) &
Contribution
(e.g. post) may
have an effect
on Influence

some metrics (e.g. #clicks) are aggregated across visits & sessions → popularity

some metrics (e.g. dwell time) are used as optimisation metrics → optimise the page/visit/session

Disengagement

... Process

Churn rate measures the percentage of users not returning to site (the opposite is retention rate)

From day 1 → focus on new users

- Calculated over day (d7, d14), week (w1, w2), and month (d30, d60, d180)
- Apps on average have retention rate of 40% after a month, which can → use as benchmark
- Retaining users over acquiring new ones

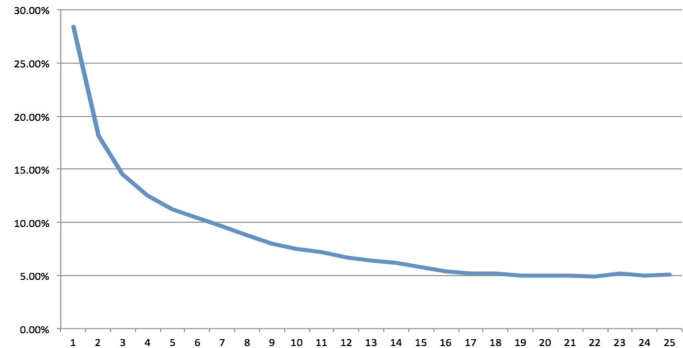
Over units of time → all users

- WoW, MoM, YoY

Churn prediction

Treatment (e.g. reduce ads)

Notification & alerts (e.g. email)



Re-engagement

... Process



Notification

Email

Offer

Marketing

Advertising

...

Index of user engagement

Click Depth Index: page views

Duration Index: time spent on site

Interaction Index: user interaction with site or product (click, upload, transaction)

Recency Index: rate at which users return (frequency)

Loyalty Index: level of long-term interaction the user has with the site or product

Brand Index: apparent awareness of the user of the brand, site, or product (e.g. search terms, social media posts)

Feedback Index: qualitative information including propensity to solicit additional information, supply direct feedback (e.g. rating, review)

Time

... From visit to session



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

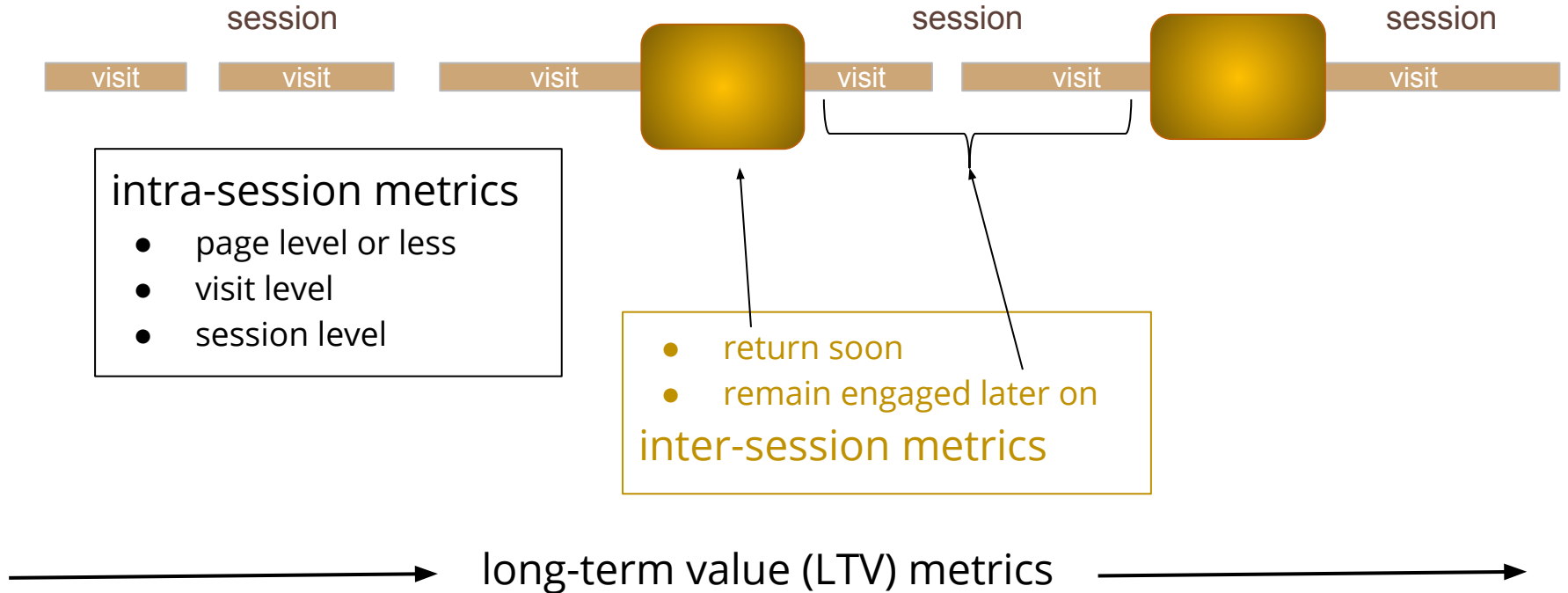
Session length is amount of time user spent on site within the session

Session frequency captures time between two consecutive sessions

session length shows how engaged users are while session frequency shows how often users are coming back (loyalty)

often 30mn is used as threshold for session boundary

Metrics and their relation to sessions



Intra- vs inter-sessions metrics

- intra-session engagement measures user activity on the site during the session
- inter-session engagement measures user loyalty with the site

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

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.

Examples of intra-session metrics

- Measures success in keeping user engaged during the session
 - clicking, spending time, adding content, making a purchase
 - user may leave the site but return within the same session
- **Involvement, Interaction, Contribution**



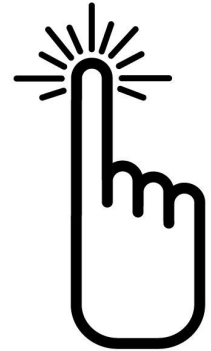
Click-through rates (CTR)

... 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 at least x seconds for music/video sites/formats

- Relate to abandonment rate
- Issues include clickbait, site design



Dwell time

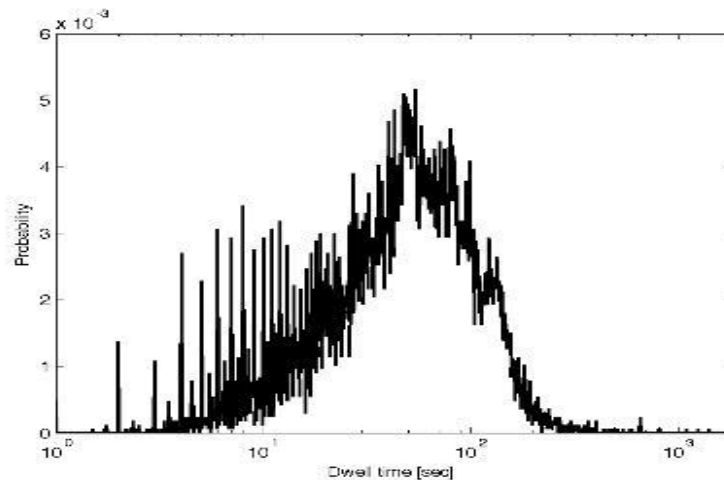
The contiguous time spent on a site or web page

Similar measure is play time for video and music sites

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

(O'Brien, Lalmas & Yom-Tov, 2013)

... Involvement

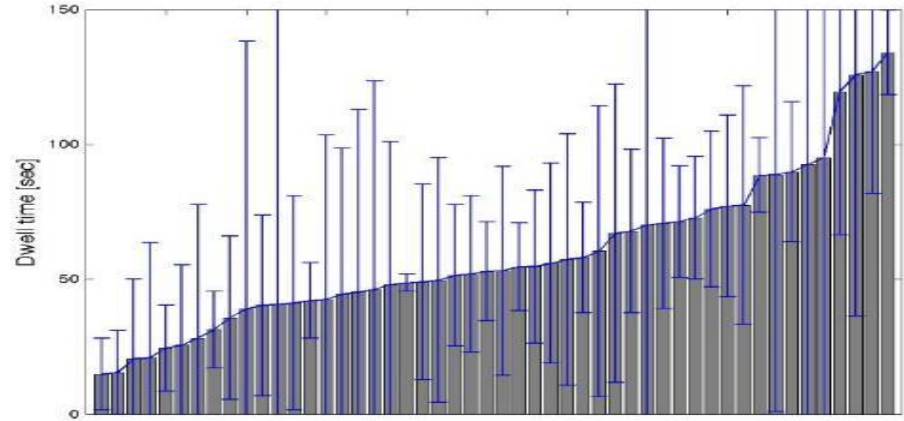


Distribution of dwell times on 50 websites

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

Pageview

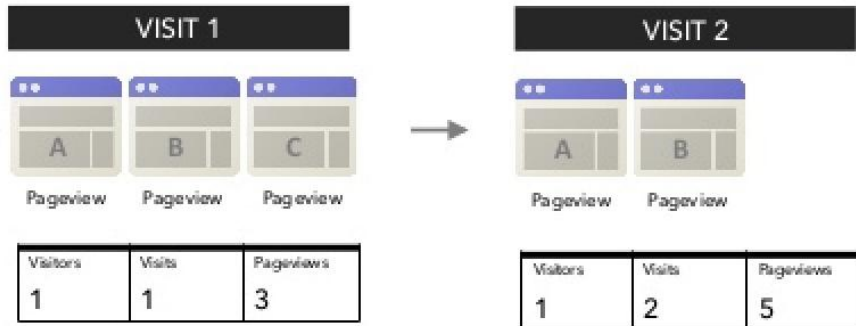
... Involvement

Page view is request to load a single page

Number of pages viewed (**click depth**): average number of contiguous pages viewed during a visit

Reload after reaching the page → counted as additional pageview

If same page viewed more than once → a single unique pageview



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

Social media metrics



... interaction

Applause

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

... interaction

Amplification

#share, #mail

... contribution

Conversations

#comments, #posts,
#replies, #edits

Metrics specific to user generated content sites such as social platforms, including social networks, blogs, wiki, etc.

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

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

Revisit rates

Number of returns to the website **within** 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

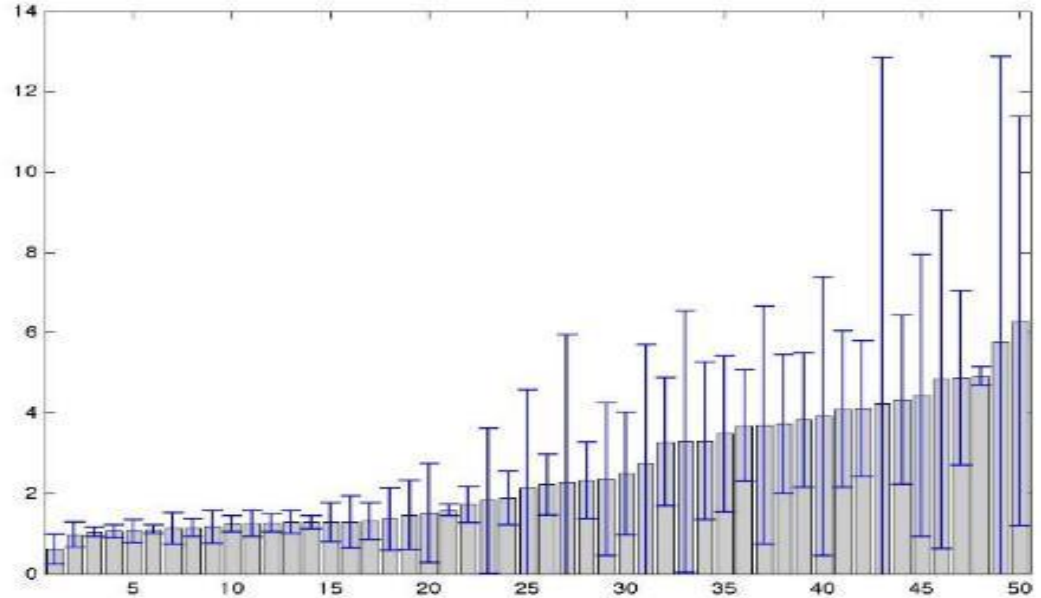
... Involvement

The screenshot shows the Yahoo! homepage interface. At the top left is the 'YAHOO!' logo. To its right is a search bar with a magnifying glass icon. Further right are icons for 'Mounia' and 'Mail'. The main content area features a large article about American Idol winner Jordin Sparks, with a sub-headline 'American Idol winner lose four loved ones in a week' and a brief summary. Below this are five smaller news thumbnails with titles like 'Trump releases memo alleging FBI abuse of power', 'Princess Eugenie wedding date confirmed', 'Inside Boris Johnson's huge family home', 'How much did Kate's royal tour wardrobe cost?', and 'House narrower than a Tube carriage on sale for £1m'. A 'Health' section follows with an article titled 'I Took Apple Cider Vinegar Every Day For 1 Week, and Here's What Happened'. Below that is a sponsored section for 'CompareCards.com' with the headline 'Top 10 Credit Cards For Those With High Credit'. On the right side, there is a 'Trending now' list with 10 items, including 'Valentine's Day', 'Money clips', 'Airline fares', 'Luxury cruises', 'Candle wall sconces', 'Valentine's roses', 'Big teddy bear', 'Infinity necklace', 'Adult-only resorts', 'Worx tools', and 'Emerald earrings'. At the bottom right, there is a 'Your Domain' advertisement and a weather forecast for Cambridge, MA, showing icons for Today, Sat, Sun, and Mon.

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



Revisit rate ... Session length

2.5M users, 785M page views, 1 month sample (Lehmann et al, 2013)

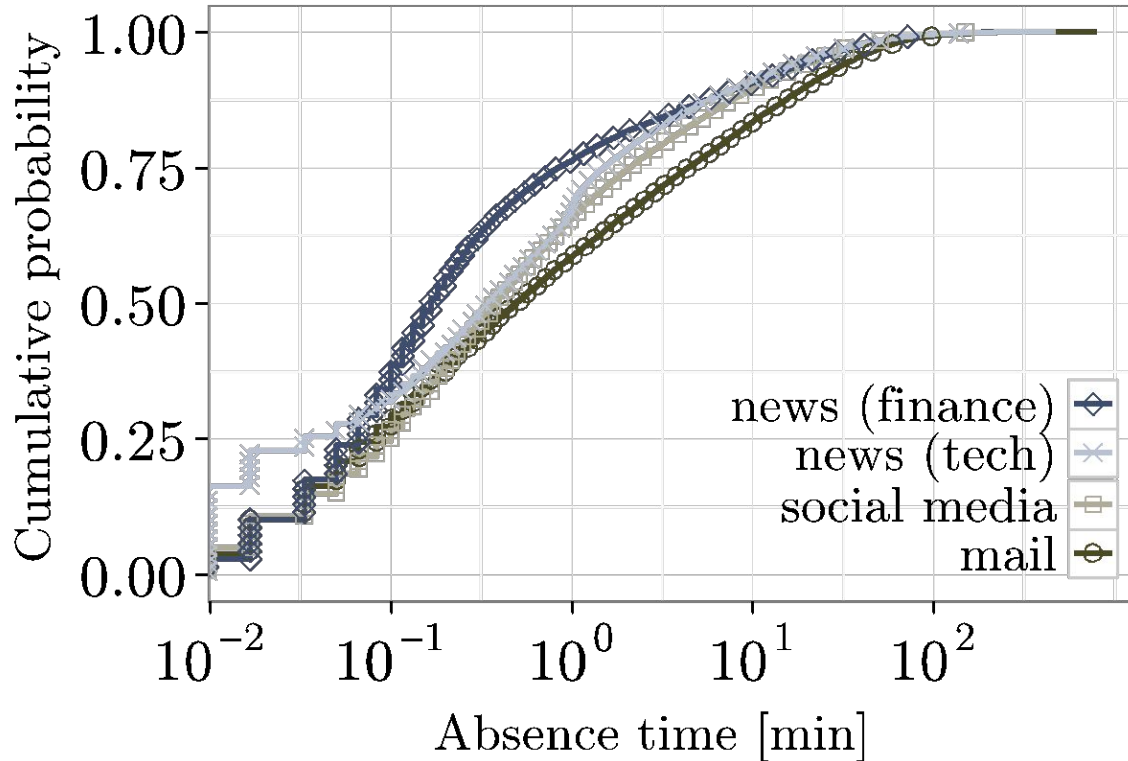
Categorization of the most frequent accessed sites

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

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

Time between each revisit



50% of sites are revisited after less than 1 minute

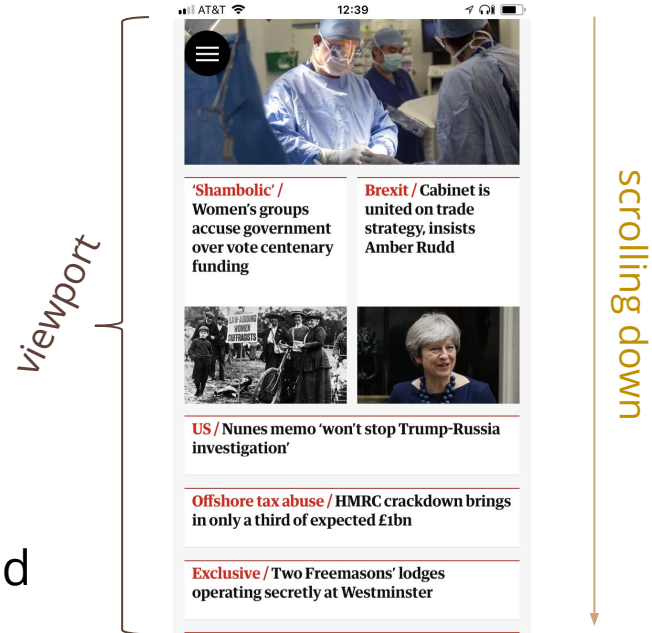
(Lehmann et al, 2013)

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 (e.g. sharing may mean very different things on social media vs news sites)



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.

Examples of inter-session metrics

Loyalty is about having users return to the site again and again, and to perceive the site as beneficial to them

- Return soon
- Remain engaged later on

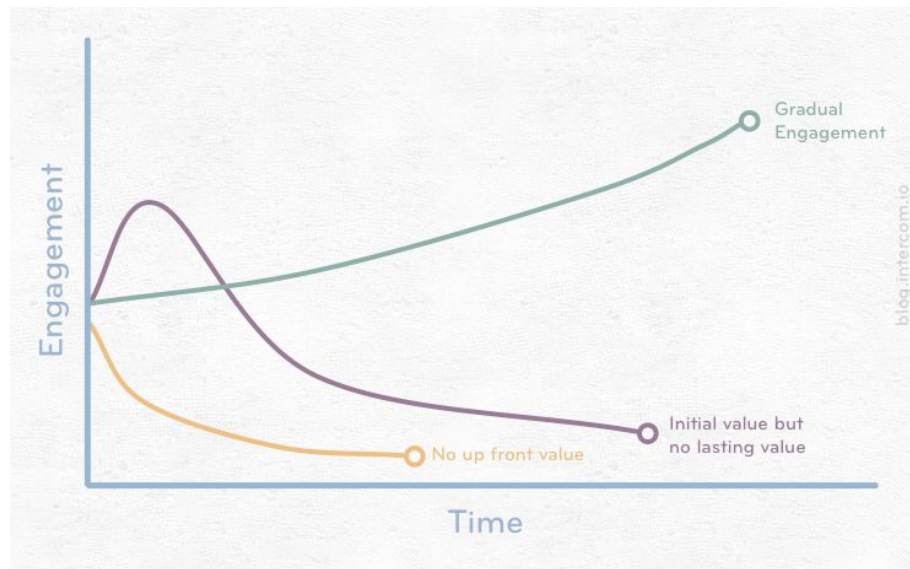


Why inter-session metrics?

(Kohavi et al, 2012)

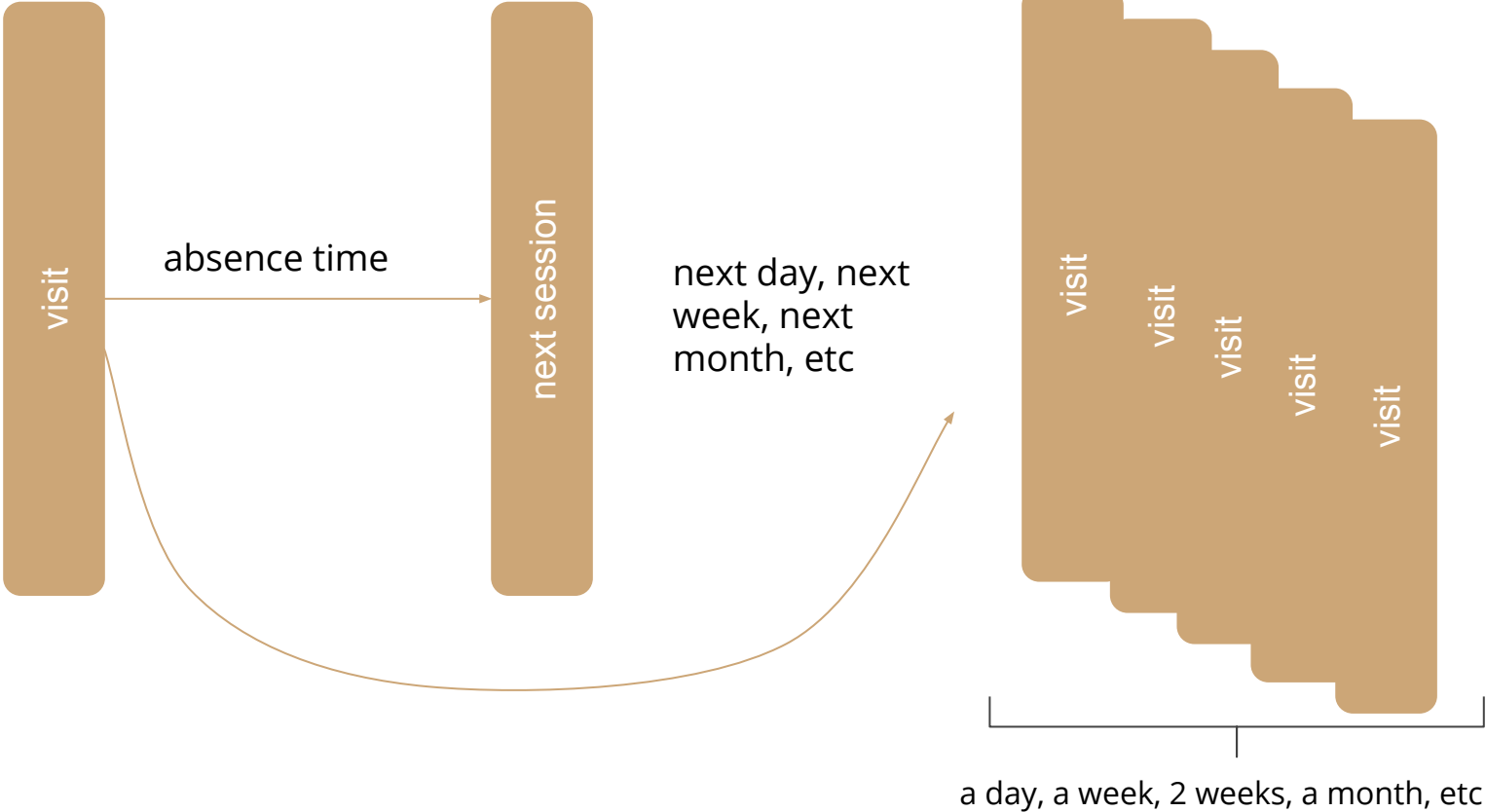
Intra-session measures can easily mislead, especially for a short time

- 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



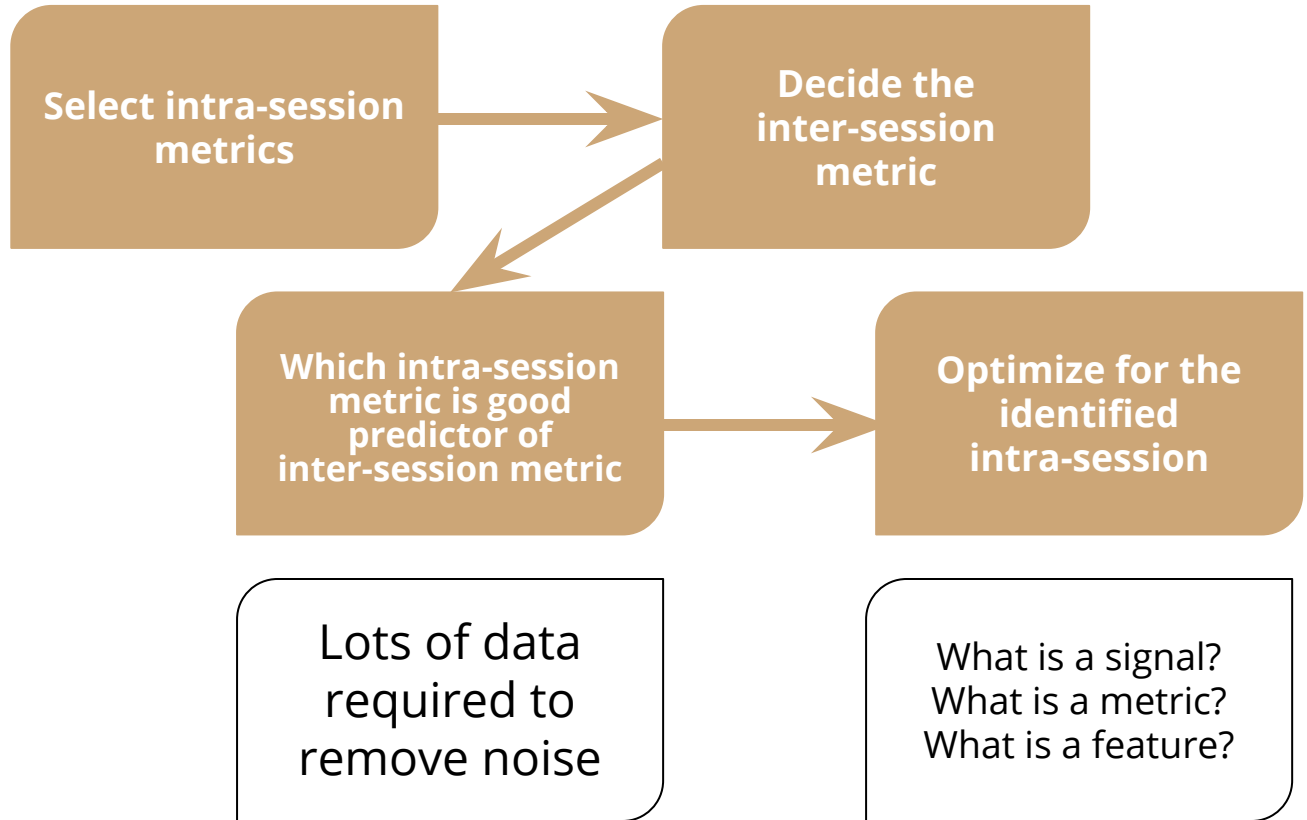
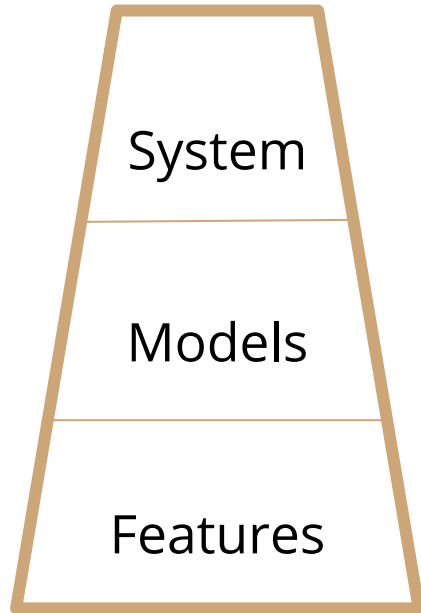
Inter-session metrics

- Total number of visits or sessions
- Total number of days active
- Total number of clicks
- Total amount of time spent ...



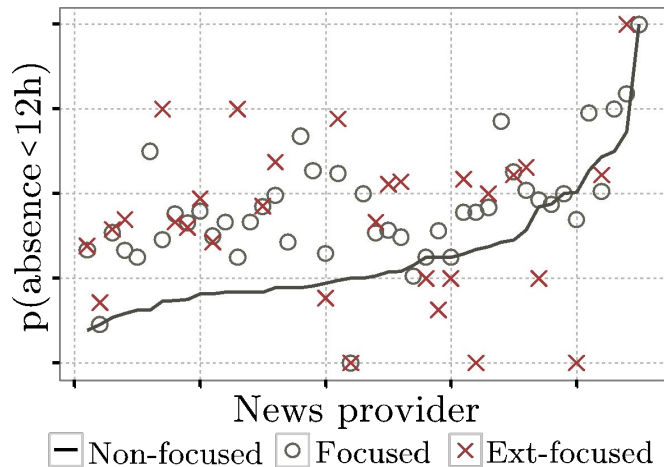
Intra- vs inter-sessions metrics

... Optimization



Example I: Focused news reading

(Lehmann et al., 2016)



Off-site link → absence time

Providing links to related off-site content has a positive long-term effect (for 70% of news sites, probability that users return within 12 hours increases by 76%)

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

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

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

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

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

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

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

Around the Web

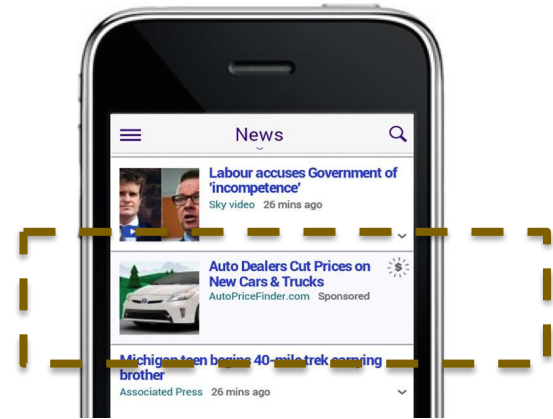
- Peace in Ukraine depends on America
- Ukraine Crisis Map
- Explosion in Ukraine
- Casualties of the Ukrainian crisis
- Exclusive interview with President Putin

Related off-site content

Example II: Ad landing page quality

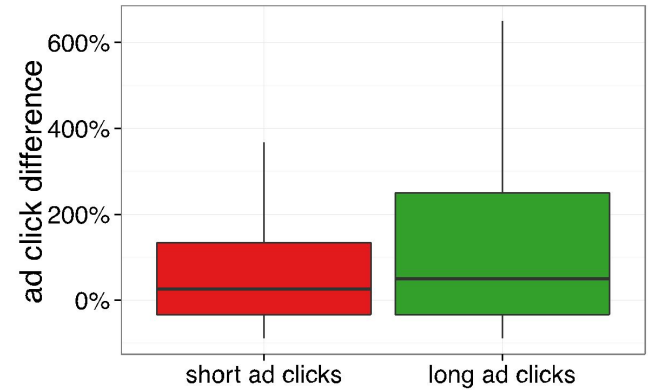
User click on an ad → ad landing page

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)



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 (CLV):
 - based on total amount of purchases

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

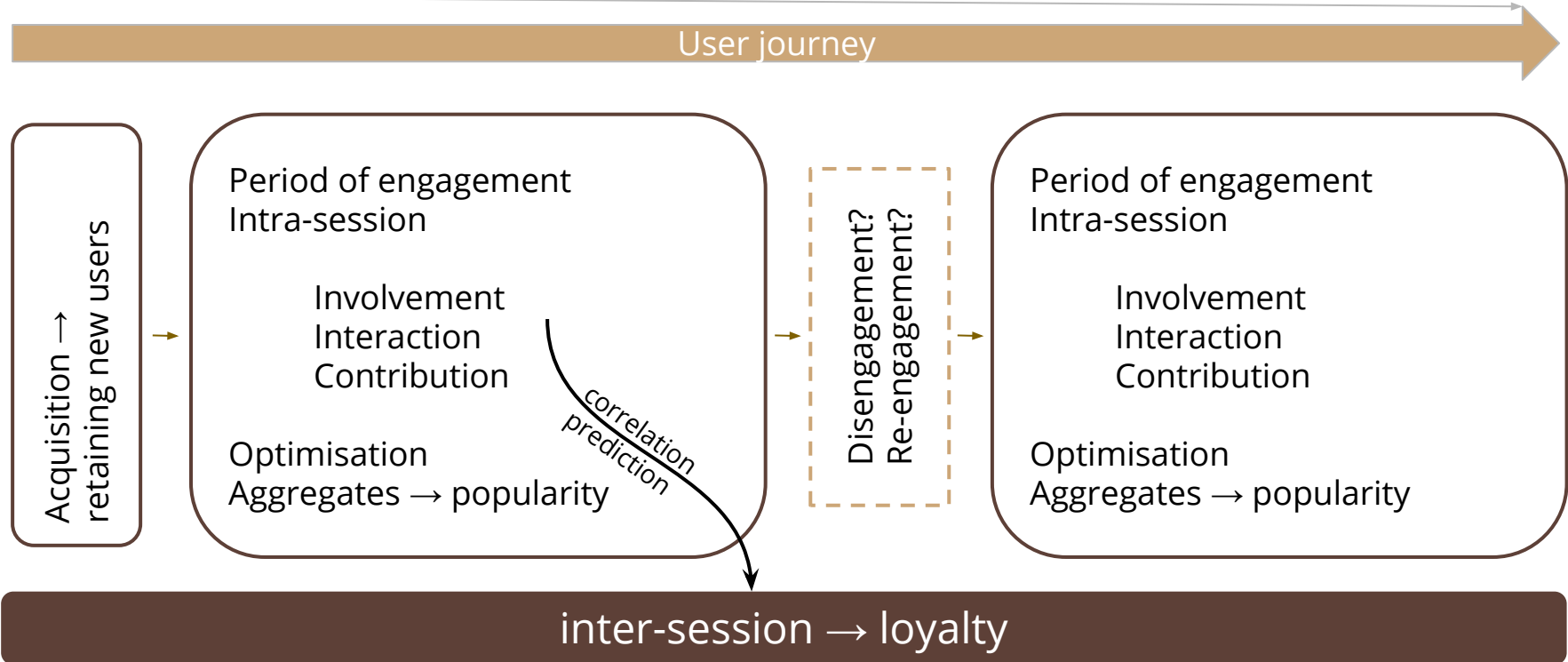
$$\begin{aligned} \text{LTV} > \text{CAC} &= \text{😊} \\ \text{CAC} > \text{LTV} &= \text{😞} \end{aligned}$$

Taxonomy of metrics

... in two slides

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

now

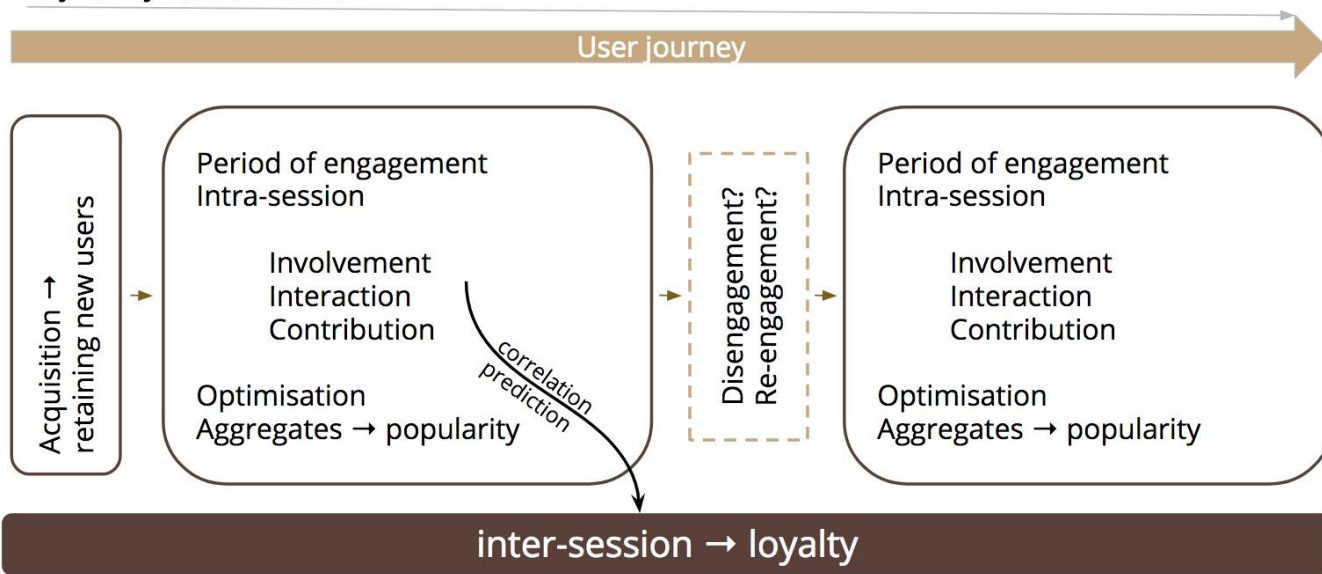


Taxonomy of metrics

... in two slides

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

now



Popularity metrics

Metrics to use to optimize machine learning algorithms

Key performance indicators (KPIs)

Long-term value (LTV) metrics



Experimentation and Evaluation of Metrics

Three levels of metrics

Business metrics -- KPIs

Behavioral metrics -- online metrics, analytics

our focus in this section

Optimisation metrics -- metrics used to train machine learning algorithms

These three levels are connected

Why experiments

Why experiments

Common reasons of not having experiments

- **Let's launch and see what happens and compare metrics before & after.**
Usually in the context of all kinds of product innovations, aiming fast iterations.

Why experiments

Common reasons of not having experiments

- **Let's launch and see what happens and compare metrics before & after.**
Usually in the context of all kinds of product innovations, aiming fast iterations.
- **Too risky.**
Usually in the context of ads, exploration & exploitation and etc.

Why experiments

Common reasons of not having experiments

- **Let's launch and see what happens and compare metrics before & after.**
Usually in the context of all kinds of product innovations, aiming fast iterations.
- **Too risky.**
Usually in the context of ads, exploration & exploitation and etc.
- **Historical data can't represent future.**
Usually in the context of offline experiments

...

Why experiments

Main benefits of having experiments

- Metrics can be **measured, tracked** and **compared**.

Why experiments

Main benefits of having experiments

- Metrics can be **measured, tracked** and **compared**.
- We can **learn, improve** and **optimize**.

Why experiments

Main benefits of having experiments

- Metrics can be **measured, tracked** and **compared**.
- We can **learn, improve** and **optimize**.
- **Save time** and **faster** iterations.

...

Experiments and Non-Experiments







Experiments and Non-Experiments

Sometimes, experiments may not be feasible or practical.

Experiments and Non-Experiments

Sometimes, experiments may not be feasible or practical.

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

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

Experiments and Non-Experiments

Sometimes, experiments may not be feasible or practical.

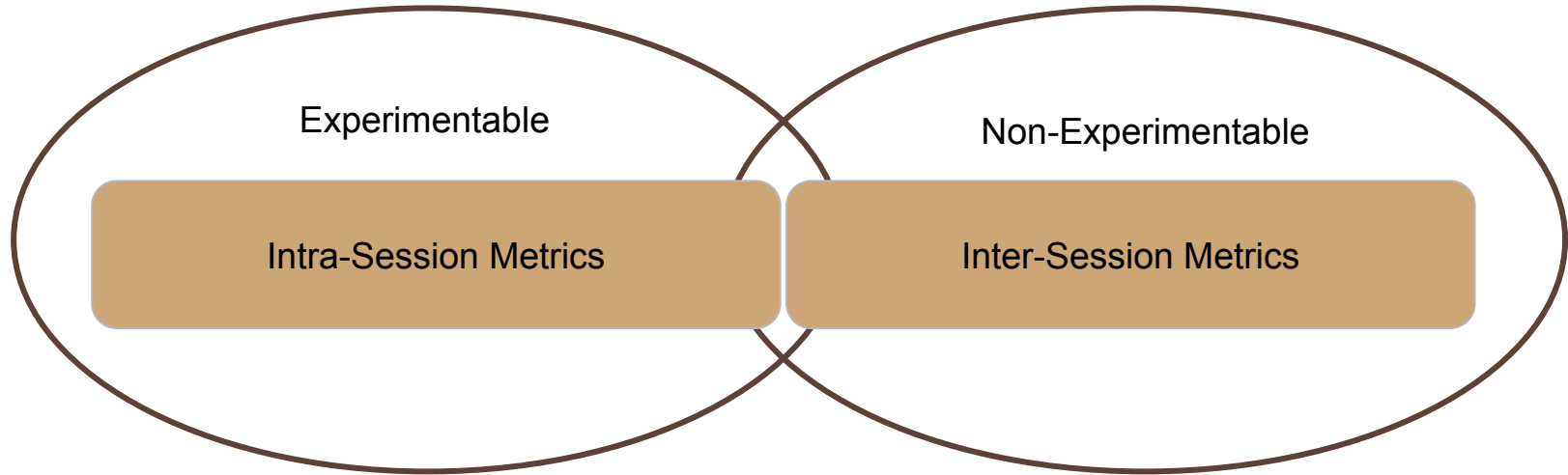
- **Example 2:**

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

The image displays a grid of 20 e-commerce product listings for Harry Potter merchandise. Each listing includes a product image, a title, a seller name, a star rating, the number of reviews, and the price. The products include:

- Pencil wands - Harry potter inspired ... (TedNTads, \$2.09)
- Set of 4, PDF Pattern, Harry Potter, R... (HelloFelt, \$14.00)
- Pottermore Inspired Patronus Animal ... (MerlinsApprentice, \$15.99)
- Harry Potter Generation Hoody - Harr... (LuckyElephant9, \$29.99)
- Wooden harry potter notebook, cust... (Zlitch, \$29.97)
- Harry Potter Svg, Harry Potter Alphab... (svgwind, \$1.98)
- Butterbeer Loose Tea - loose leaf roo... (TheForestWitch, \$8.39)
- Harry Potter Wine Glass, Not Today M... (BEMINDESIGNshop, \$11.00)
- Sorting Hat Bath Bombs - Harry Potte... (Scented, \$6.58)
- Harry Potter Bath Bomb, Potion Bath ... (MadeByMagicCreations, \$30.00)
- House Sorting Bath Bomb | Harry Pot... (ZenBathCandlesLLC, \$4.99)
- I Don't Give A GryffinDamn - SlytherS... (HeartlandKnots, \$10.00)
- Harry Potter Mug | Harry Potter Teach... (PAISGCreations, \$10.99)
- Harry Potter Fat Quarter Bundle (SitenNStashFabrics, \$22.49)
- Wizard Symbols Fabric by the Yard... (JacksonsHovens, \$10.99)
- Squad Shirt, Bachelorette Tanks, Bride... (TussimApparel, \$9.34)

Experiments and Non-Experiments



Experiments

Summary

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

Experiments

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.

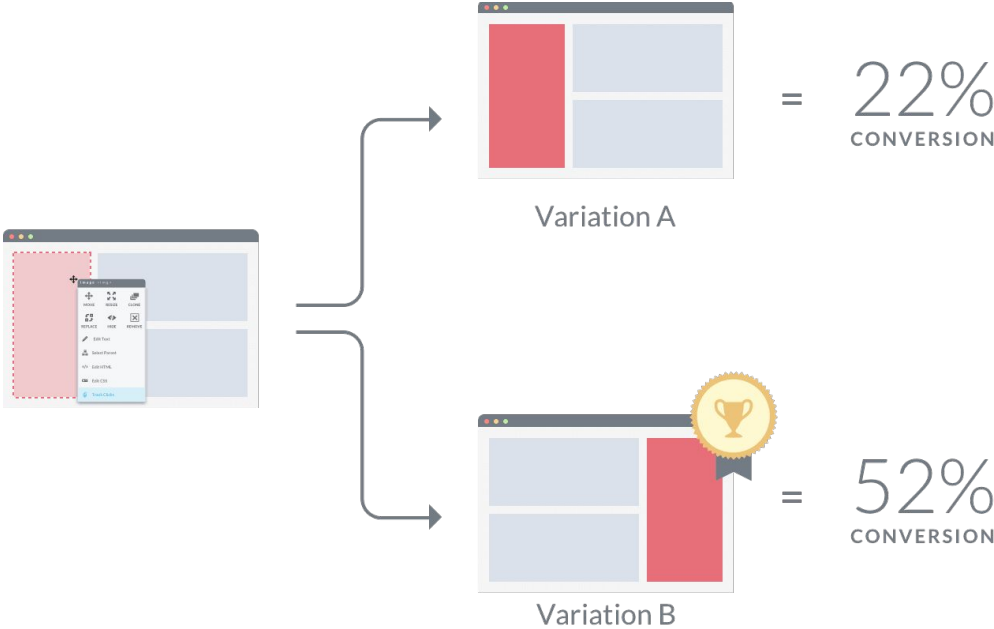
Types of experiments

Types of experiments

- **Online Experiment**
- **Offline Experiment**
- **Offline A/B Experiment**

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments



Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments

- **Have deep roots in classic statistics, with new challenges.**
e.g., “always need more traffic”

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments

- **Have deep roots in classic statistics, with new challenges.**
e.g., “always need more traffic”
- **Can derive causal relationships easier.**
e.g., complex user behaviors

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments

- **Have deep roots in classic statistics, with new challenges.**
e.g., “always need more traffic”
- **Can derive causal relationships easier.**
e.g., complex user behaviors
- **Have direct impact on users.**
e.g., users may decide not to come back

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments

- **Have deep roots in classic statistics, with new challenges.**
e.g., “always need more traffic”
- **Can derive causal relationships easier.**
e.g., complex user behaviors
- **Have direct impact on users.**
e.g., users may decide not to come back
- **Cannot easily be reused.**
e.g., need to re-launch the experiment

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments

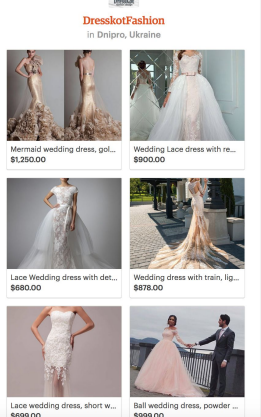
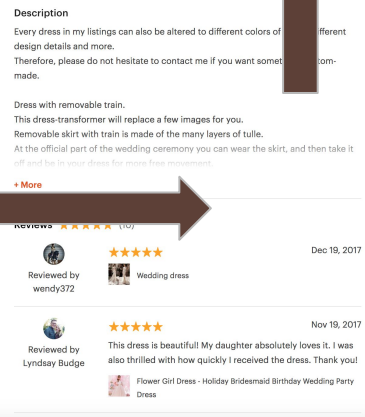
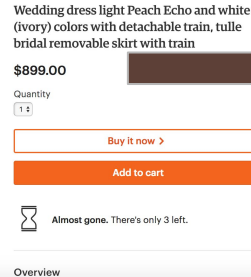
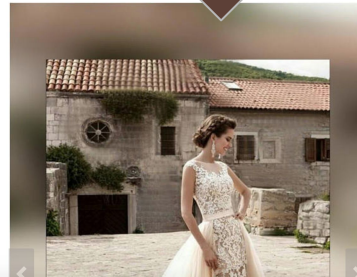
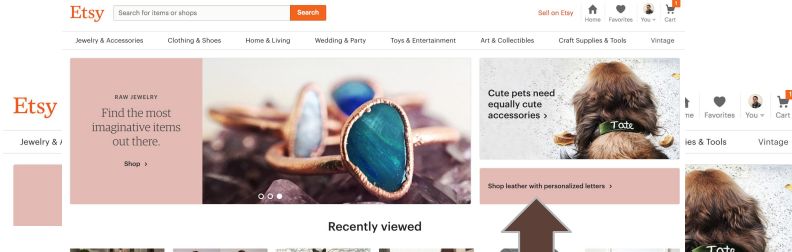
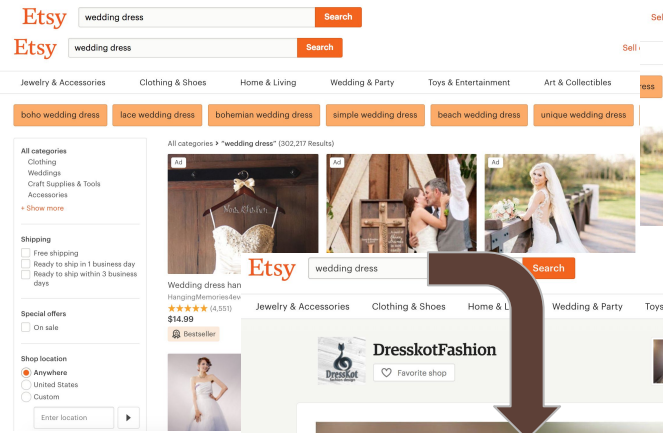
This screenshot shows the Etsy search results for 'wedding dress'. The search bar at the top contains 'wedding dress' and a search button. Below the search bar, there are navigation tabs for various categories: boho wedding dress, lace wedding dress, bohemian wedding dress, simple wedding dress, beach wedding dress, and unique wedding dress. On the left, there is a sidebar with filters for categories, shipping options, special offers, and shop location. The main content area displays a grid of wedding dress listings, with the top listing being 'Wedding dress han' by DresskotFashion, priced at \$14.99.

This screenshot shows the Etsy homepage. The search bar at the top contains 'Search for items or shops'. Below the search bar, there are navigation tabs for various categories: Jewelry & Accessories, Clothing & Shoes, Home & Living, Wedding & Party, Toys & Entertainment, Art & Collectibles, Craft Supplies & Tools, and Vintage. The main content area features a large banner for 'RAW JEWELRY' with the text 'Find the most imaginative items out there.' and a 'Shop' button. To the right, there are several promotional banners, including one for 'Cute pets need equally cute accessories' and another for 'Shop leather with personalized letters'.

This screenshot shows the product page for a wedding dress on the DresskotFashion shop. The product title is 'Wedding dress light Peach Echo and white (ivory) colors with detachable train, tulle bridal removable skirt with train'. The price is \$899.00. The page includes a 'Buy now' button, an 'Add to cart' button, and a note that the item is 'Almost gone. There's only 3 left.' The product description states: 'Every dress in my listings can also be altered to different colors of fabric, different design details and more. Therefore, please do not hesitate to contact me if you want something custom-made. Dress with removable train. This dress-transformer will replace a few images for you. Removable skirt with train is made of the many layers of tulle. At the official part of the wedding ceremony you can wear the skirt, and then take it off and be in your dress for more free movement.' The page also features a 'Reviews' section with 10 reviews, each with a star rating and a date. The reviews are: 'Reviewed by wendy372' (5 stars, Dec 19, 2017), 'Reviewed by Lyndsay Budge' (5 stars, Nov 19, 2017), and 'Reviewed by Flower Girl Dress - Holiday Bridesmaid Birthday Wedding Party Dress' (5 stars, Nov 19, 2017). The product is also featured in a grid of similar items on the right side of the page.

Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments



Online experiment

Metrics for Online Experiments

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

Online experiment

Metrics for Online Experiments

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

Online experiment

Summary

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

References:

- [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

Traditional Offline Dataset/Collection Experiment

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

Offline experiment

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.

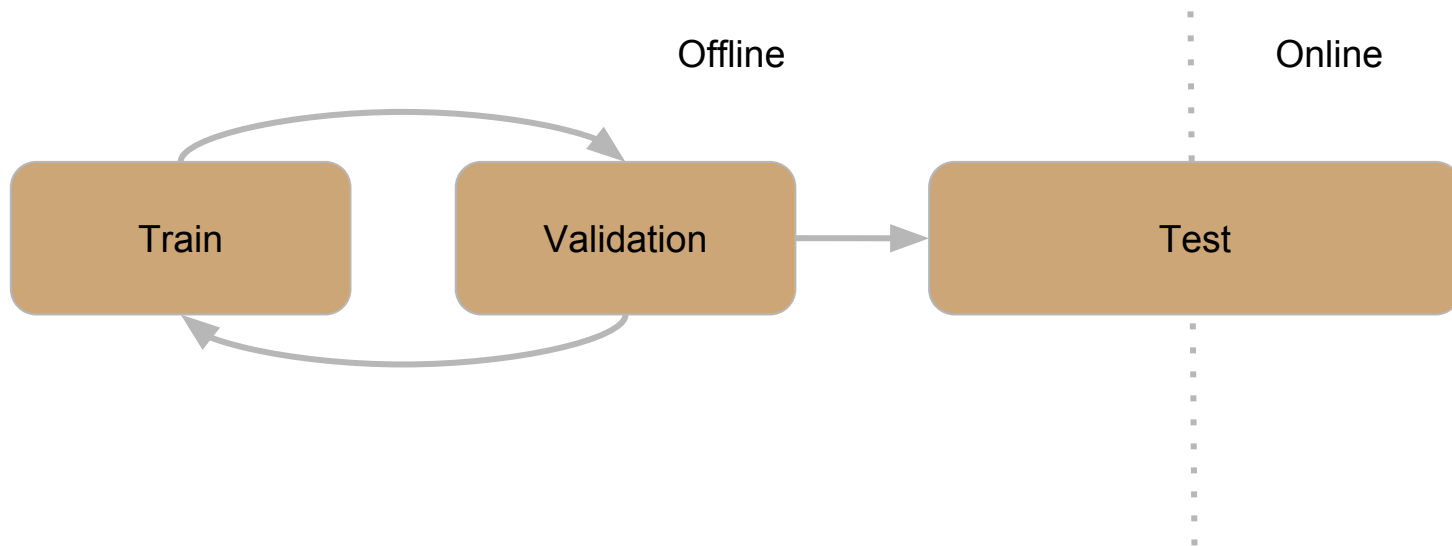
Offline experiment

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

Traditional Offline Dataset/Collection Experiment



Offline experiment

- **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

Summary

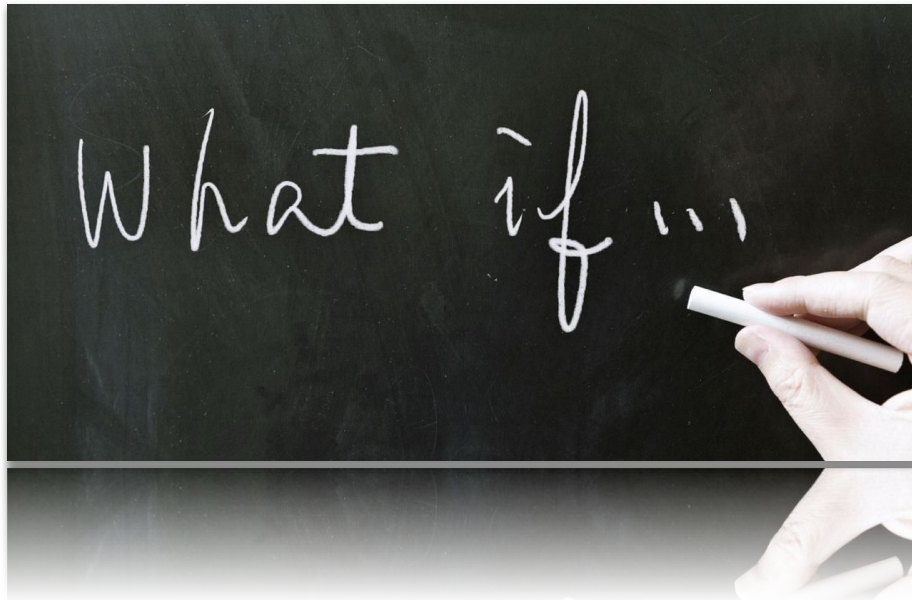
- Indirect and can be reused
- Good machine learning theories
- Correlation
- Static

References:

- [1] Mark Sanderson (2010). **Test Collection Based Evaluation of Information Retrieval Systems**. Foundations and Trends® in Information Retrieval: Vol. 4: No. 4.
- [2] Donna Harman (2011). **Information Retrieval Evaluation**. Synthesis Lectures on Information Concepts, Retrieval, and Services 3:2.

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment



Offline A/B Experiment

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.

References:

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

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

Offline A/B Experiment

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.

References:

- [1] 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

Counterfactual Offline Reasoning/Experiment

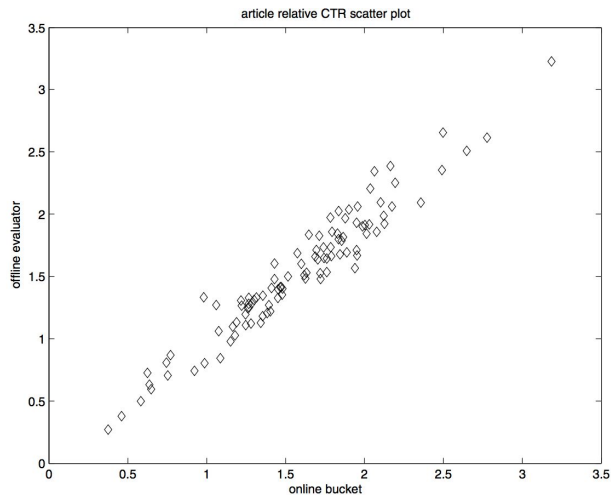


Figure 2: Articles' CTRs in the online bucket versus offline estimates.

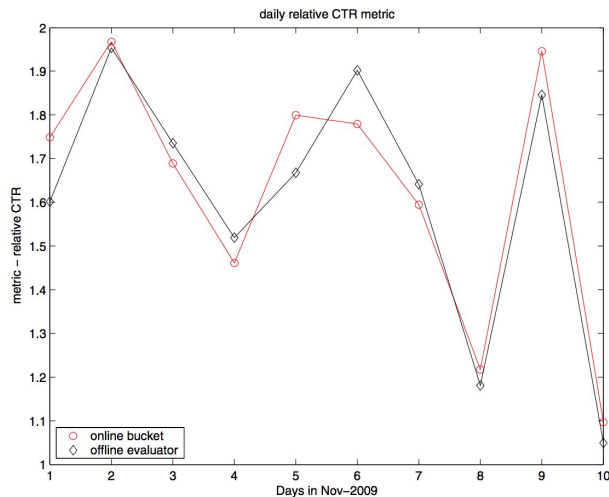


Figure 3: Daily overall CTRs in the online bucket versus offline estimates.

References:

[1] 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

Counterfactual Offline Reasoning/Experiment

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

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

The screenshot shows an Etsy search results page for 'jewelry box'. The search bar at the top contains 'jewelry box' and a 'Search' button. Below the search bar are navigation links for 'Home', 'Favorites', 'You', and 'Cart'. A horizontal menu lists various categories: Jewelry & Accessories, Clothing & Shoes, Home & Living, Wedding & Party, Toys & Entertainment, Art & Collectibles, Craft Supplies & Tools, and Vintage. Below this are filter buttons for 'jewelry box wood', 'wooden jewelry box', 'large jewelry box', 'small jewelry box', 'jewelry box vintage', and 'personalized jewelry box'. On the left side, there is a sidebar with filters for 'All categories', 'Shipping', 'Special offers', and 'Shop location'. The main content area displays a grid of jewelry boxes with their respective images, titles, prices, and ratings. The first item is a 'Raven box, handmade boxes, steamp...' by ST3jewellery for \$30.95. The second is a 'Bridesmaid Gift / Popular Bridesmaid...' by SugarAndChicShop for \$45.00. The third is a 'Matte Black Custom Branded Laser...' by Izbeams for \$85.00. The fourth is a 'Flower Girl or Bridesmaids Gift Box...' by JCreateDesign for \$16.20 (10% off). The fifth is a 'Wall Jewelry Box with metal mesh do...' by HoodedOnWoodwork for \$98.00. The sixth is a 'Built on Order - Reclaimed Wood Box...' by IndependentBoxWorks for \$40.00. The seventh is a 'Personalized Rustic Jewelry Box, uniq...' by dfericha for \$135.00. The eighth is a 'Jewelry box, 4 drawer, jewelry holder...' by JoyfulCreations3 for \$35.00. The ninth is a 'Rustic Jewelry Box - Custom engrave...' by Dustyroadgirl for \$35.00. The tenth is a 'Wood Jewelry Box candelintreasures' by JoyfulCreations3 for \$10.00. The eleventh is a 'Personalized Rustic Wedding Wood B...' by danielcustommade for \$15.00. The twelfth is a 'Personalized Rustic Jewelry Box, uniq...' by danielcustommade for \$15.00. The thirteenth is a 'Personalized Rustic Jewelry Box, uniq...' by danielcustommade for \$15.00. The fourteenth is a 'Personalized Rustic Jewelry Box, uniq...' by danielcustommade for \$15.00.

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

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

The screenshot shows an Etsy search results page for the query "jewelry box". The page features a navigation bar with the Etsy logo, a search bar containing "jewelry box", and a "Search" button. Below the search bar, there are several filter tabs: "jewelry box wood", "wooden jewelry box", "large jewelry box", "small jewelry box", "jewelry box vintage", and "personalized jewelry box". The main content area displays a grid of product listings. Each listing includes a product image, a title, a shop name, a star rating, and a price. For example, the first listing is a "Raven box, handmade boxes, steampunk" by ST3jewellery, priced at \$30.95. Other listings include a "Bridesmaid Gift / Popular Bridesmaid..." by SugarAndChicShop for \$45.00, a "Matte Black Custom Branded Laser..." by lzbreams for \$85.00, and a "Personalized Memory Box, Keepsake ..." by EngraveMyMemories for \$29.95. The page also includes a sidebar with filters for categories, shipping options, special offers, and shop location.

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

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

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

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

Reference:

- [1] Liangjie Hong, 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 T. Dumais, and Thorsten Joachims. **Short-Term Satisfaction and Long-Term Coverage: Understanding How Users Tolerate Algorithmic Exploration**. WSDM 2018.

Offline A/B Experiment

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

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

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

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

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

end for

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

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

 Observe N rewards \mathbf{r}_t .

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

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

end for

Offline A/B Experiment

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

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

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

end for

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

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

 Observe N rewards \mathbf{r}_t .

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

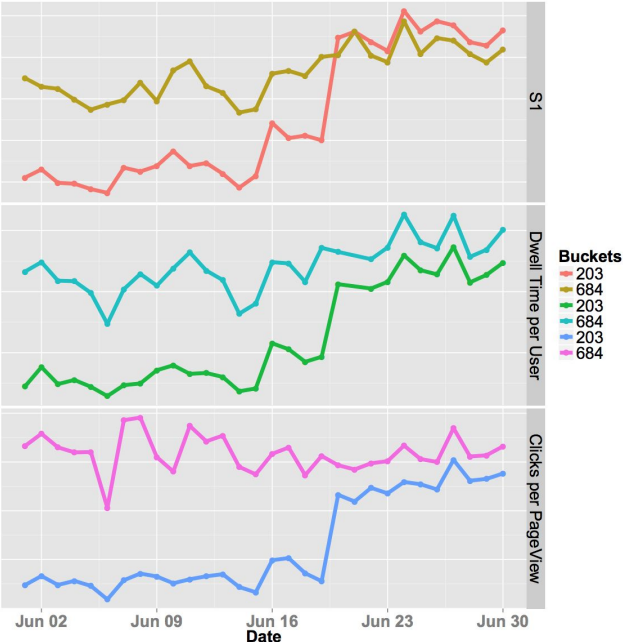
 Logging N items, \mathbf{p} and \mathbf{r}_t .

end for

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

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



Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

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

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

Offline A/B Experiment

Summary

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

References:

- [1] Lihong Li, Jinyoung Kim, Imed Zitouni: **Toward Predicting the Outcome of an A/B Experiment for Search Relevance**. WSDM 2015.
- [2] Adith Swaminathan et al. **Off-policy evaluation for slate recommendation**. NIPS 2017.
- [3] Tobias Schnabel, Adith Swaminathan, Peter I. Frazier, and Thorsten Joachims. 2016. **Unbiased Comparative Evaluation of Ranking Functions**. ICTIR 2016.
- [4] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, Simon Dollé. **Offline A/B testing for Recommender Systems**. WSDM 2018.

Evaluation of Metrics

- Hypothesis Testing
- Causal Inference

Hypothesis Testing

Statistical Comparison

- Well grounded theory for classic cases
- Not well studied in a lot of online settings
- Gold standard for statistical difference
- Weak for practical difference

References:

[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 (March 2012), 34 pages.

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

References:

- [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.

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,...



Optimisations for Metrics

Three levels of metrics

Business metrics -- KPIs

Behavioral metrics -- online metrics, analytics

our focus in this section

Optimisation metrics -- metrics used to train machine learning algorithms

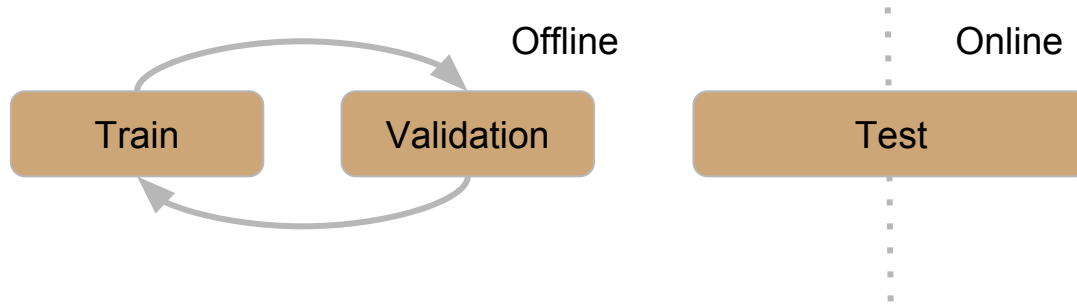
These three levels are connected

Optimisations for Metrics

- Offline Experiments → Offline Optimization
- Online Experiments → Online Optimization
- Offline A/B Experiments → Counterfactual Optimization
- From Intra-Session to Inter-Session Metrics Optimization

Offline Optimization

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



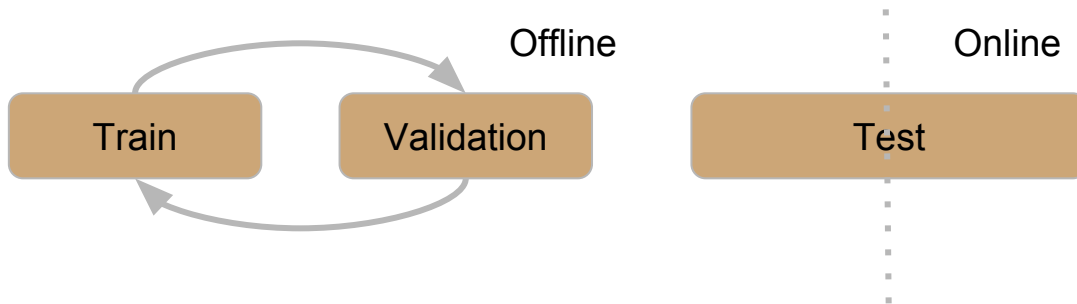
Offline Optimization

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

Offline Optimization

- **Bias**

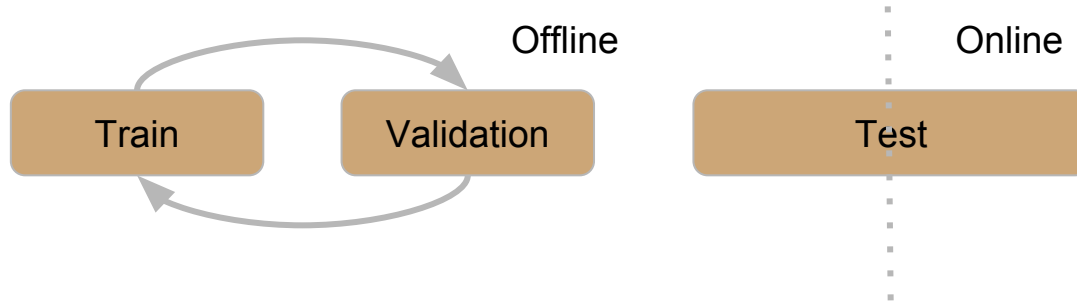
Examples: presentation bias, system bias...



Offline Optimization

- **Concept Drifts**

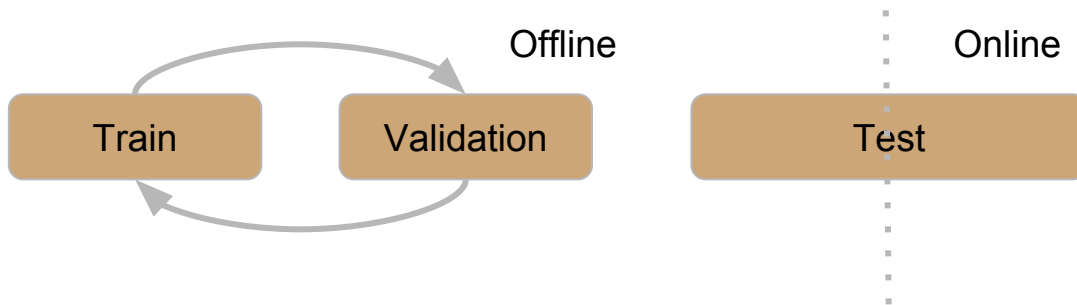
Examples: seasonal, interest shift...



Offline Optimization

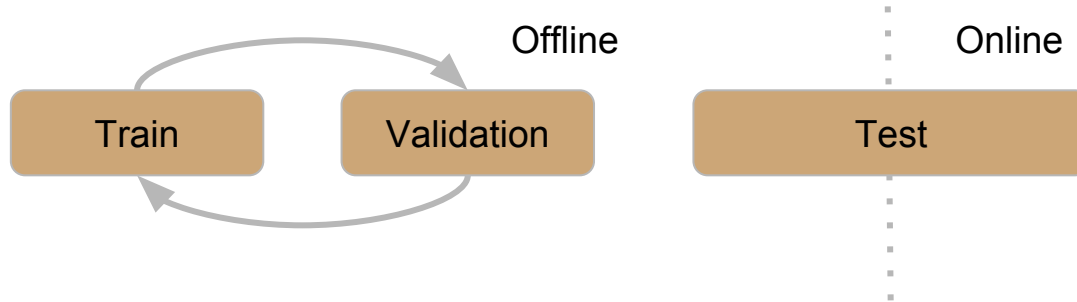
- **Different of offline metrics and online metrics**

Examples: AUC/nDCG versus DAU...



Offline Optimization

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



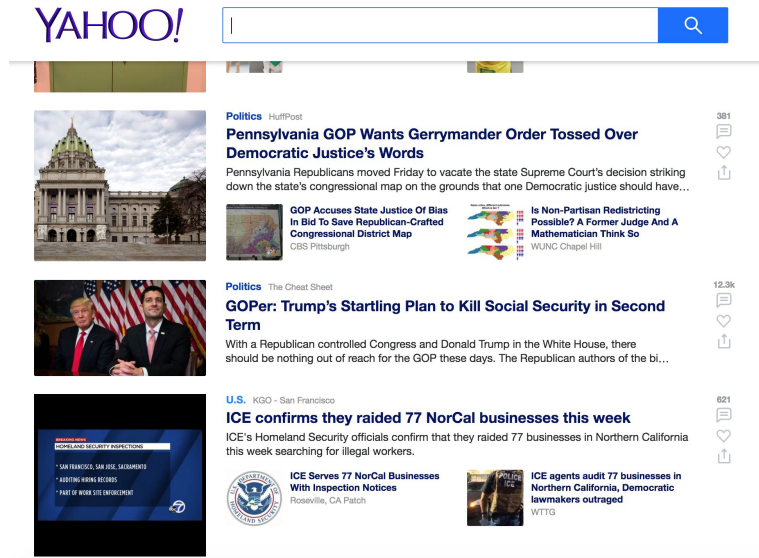
Online Optimization

Online Optimization

- **Online Learning**
- **Contextual Multi-armed Bandit**
- **Reinforcement Learning**

Online Optimization

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



Reference:

[1] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. **Returning is Believing: Optimizing Long-term User Engagement in Recommender Systems.** In CIKM 2017.

Online Optimization

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.

[2] Y. Koren, R. Bell, and C. Volinsky. **Matrix Factorization Techniques for Recommender Systems**. Computer 42, 8 (2009), 30–37.

Online Optimization

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.

[2] Y. Koren, R. Bell, and C. Volinsky. **Matrix Factorization Techniques for Recommender Systems**. Computer 42, 8 (2009), 30–37.

- Users may leave because of boredom from popular items.

[1] Komal Kapoor, Karthik Subbian, Jaideep Srivastava, and Paul Schrater. **Just in Time Recommendations: Modeling the Dynamics of Boredom in Activity Streams**. In WSDM 2015.

Online Optimization

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

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

Online Optimization

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**. In WWW 2016. 761–770.

[2] Mounia Lalmas, Jane.e Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. **Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users**. In KDD 2015.

[3] Nan Du, Yichen Wang, Niao He, Jimeng Sun, and Le Song. **Time-Sensitive Recommendation From Recurrent User Activities**. In NIPS 2015.

[4] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. **A Hazard Based Approach to User Return Time Prediction**. In KDD 2014.

Online Optimization

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

Balance between

- 1. Maximize immediate reward of the recommendation**

Online Optimization

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.**

Online Optimization

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**.

Online Optimization

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

Some more related work about *multi-armed bandit*:

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

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

Online Optimization

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

Main Idea

Online Optimization

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 .

Online Optimization

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...

Online Optimization

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.

Online Optimization

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.

Online Optimization

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.

Online Optimization

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

Approximations

Online Optimization

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

Approximations

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

Online Optimization

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.

Online Optimization

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.

Online Optimization

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

Online Optimization

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 .

Online Optimization

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.

Online Optimization

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.

Online Optimization

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].

Online Optimization

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.

Online Optimization

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

Algorithm 1 r^2 Bandit

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

Online Optimization

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

Algorithm 1 r^2 Bandit

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

Online Optimization

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

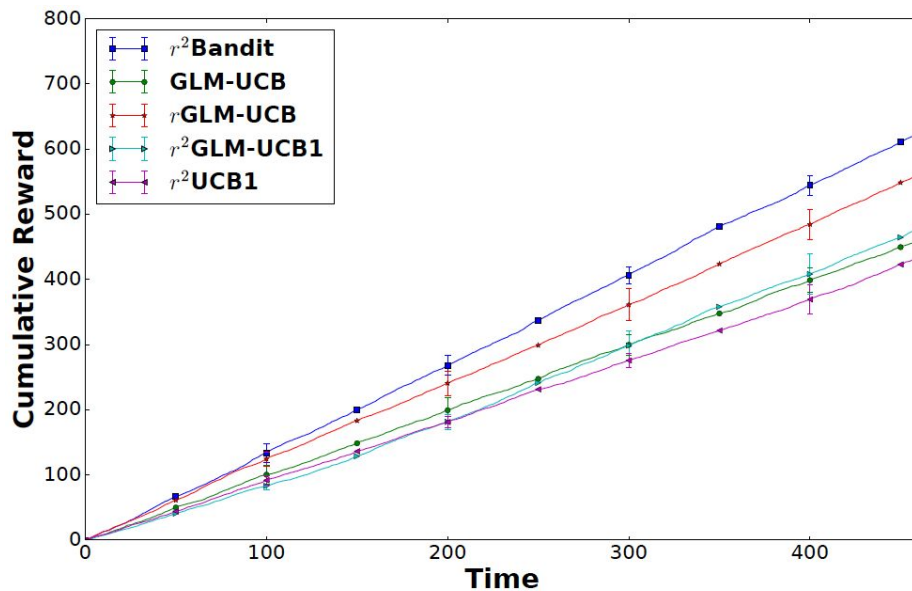
Simulations

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.

Online Optimization

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

Simulations

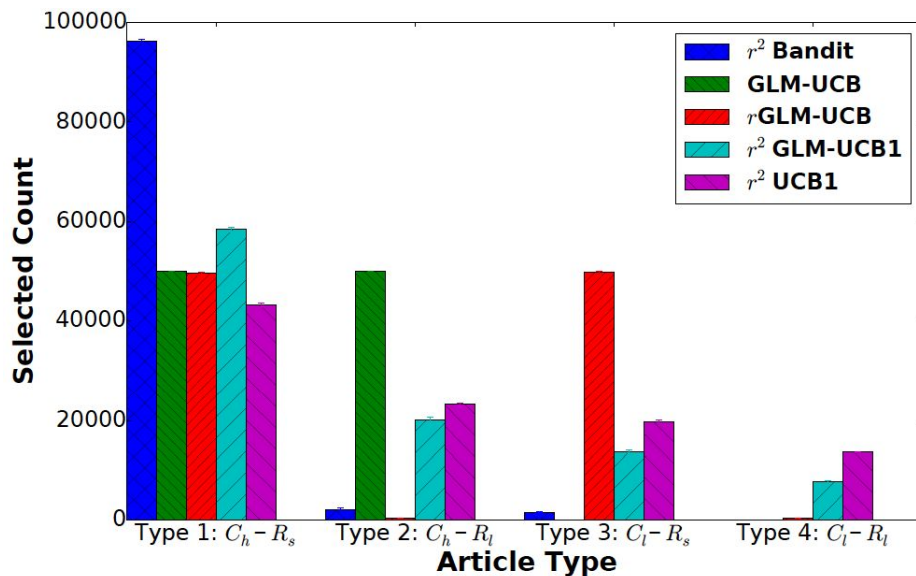


(a) Cumulative clicks over time

Online Optimization

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

Simulations

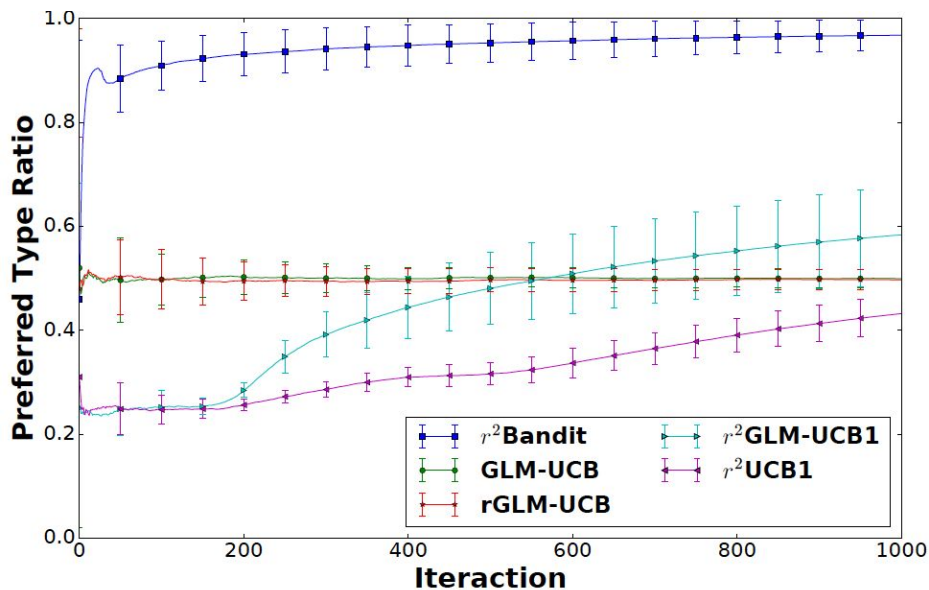


(b) Distribution of selected item types

Online Optimization

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

Simulations



(c) Evolution of preferred item type ratio

Online Optimization

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.

Online Optimization

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

Real-World Dataset

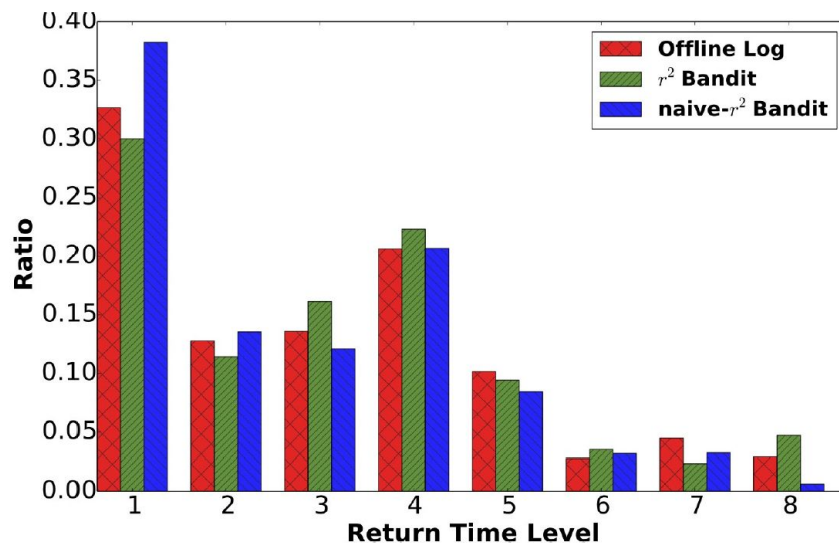


Figure 2: Discretized user return time distribution.

Online Optimization

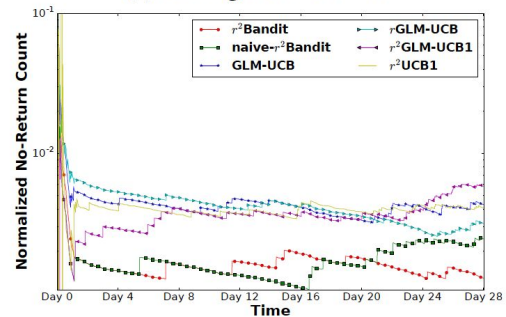
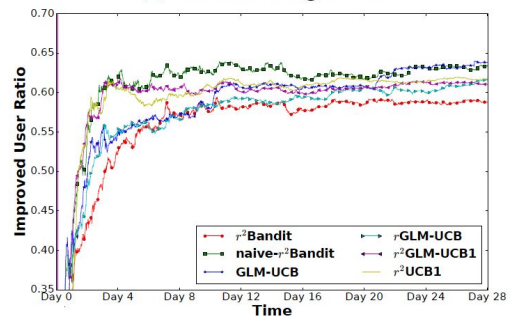
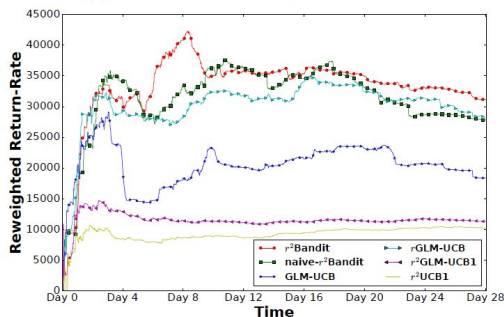
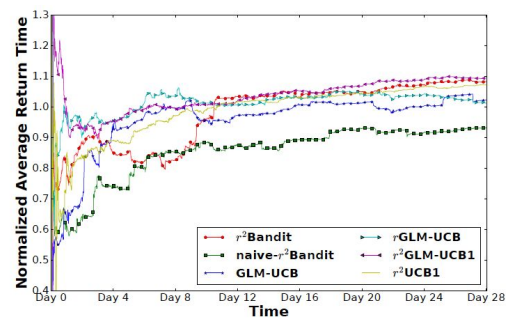
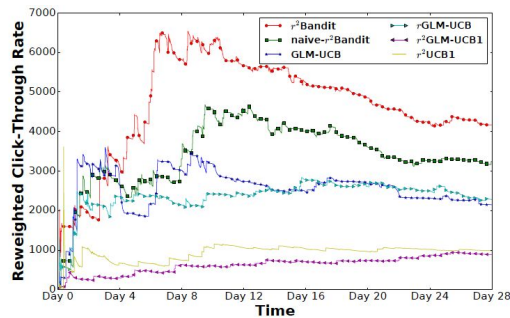
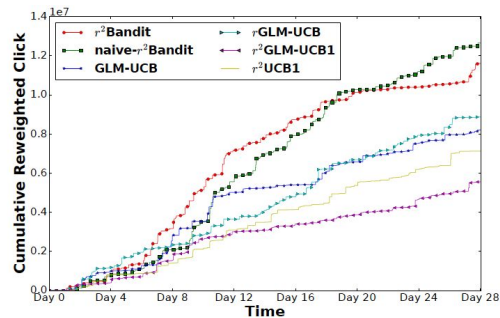
How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

Real-World Dataset: Evaluation

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

Online Optimization

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



(d) Return rate

(e) Improved user ratio

(f) No return count

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

Counterfactual Optimization

- Emerging topics
- Optimization under counterfactual setting, simulating A/B testing

References:

- [1] Xuanhui Wang, Michael Bendersky, Donald Metzler, 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, Adith Swaminathan. **Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement**. SIGIR 2016 Tutorial.
- [4] Adith Swaminathan, Thorsten Joachims. **Counterfactual Risk Minimization: Learning from Logged Bandit Feedback**. ICML 2015.

Counterfactual Optimization

Generic Idea:

1. Rewrite the objective function with inverse propensity scoring.
2. Try to optimize or approximate the new objective.

References:

- [1] Xuanhui Wang, Michael Bendersky, Donald Metzler, 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, Adith Swaminathan. **Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement**. SIGIR 2016 Tutorial.
- [4] Adith Swaminathan, Thorsten Joachims. **Counterfactual Risk Minimization: Learning from Logged Bandit Feedback**. ICML 2015.

Optimization Inter-Session Metrics

Optimization 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.

Optimization Inter-Session Metrics

Approach I

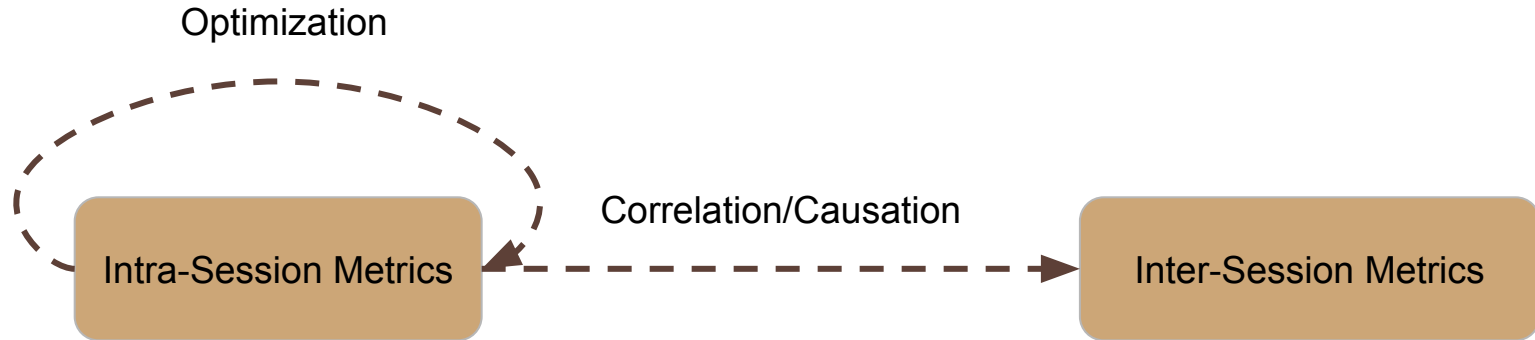
If inter-session metrics can be explicitly modeled or write them down in their clear form, you can use online optimization tools to directly optimize them.

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

...

Optimization Inter-Session Metrics

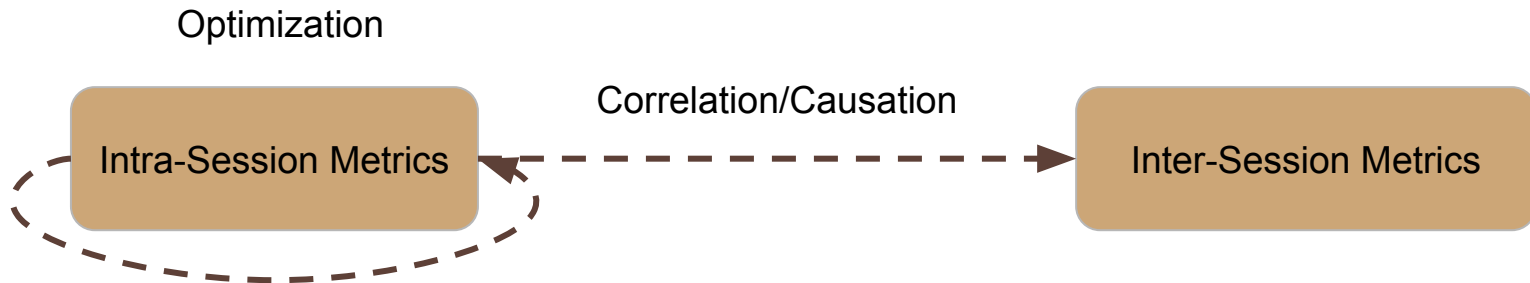
Approach II



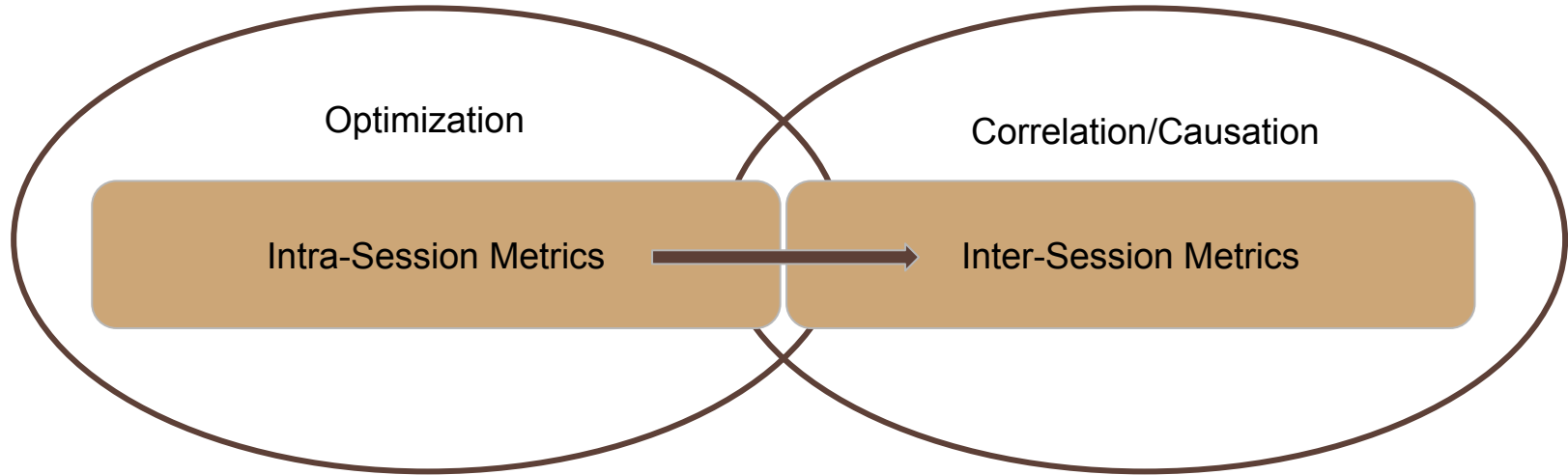
Optimization Inter-Session Metrics

Approach II

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



Optimization Inter-Session Metrics



Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

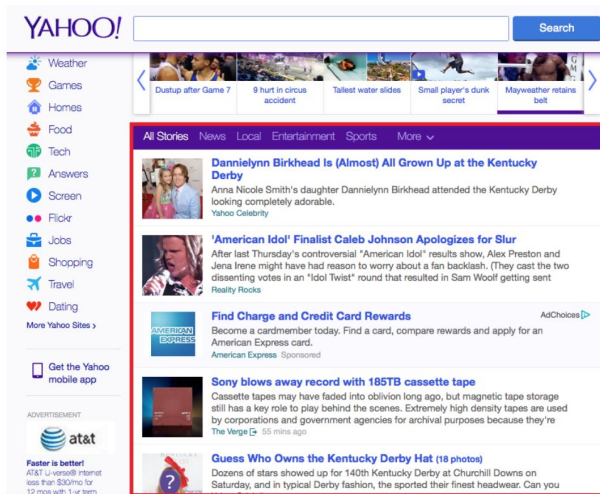


Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.

Reference:

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

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

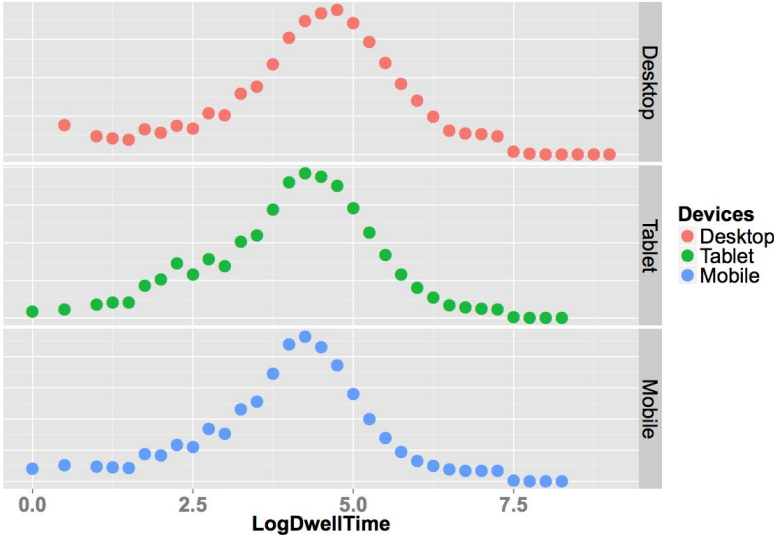


Figure 2: The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

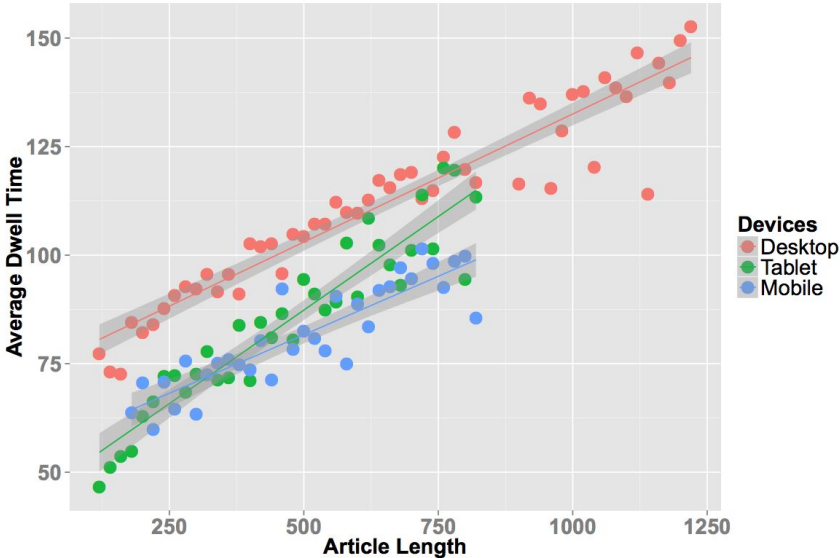


Figure 3: The relationship between the average dwell time and the article length where X-axis is the binned article length and the Y-axis is binned average dwell time.

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

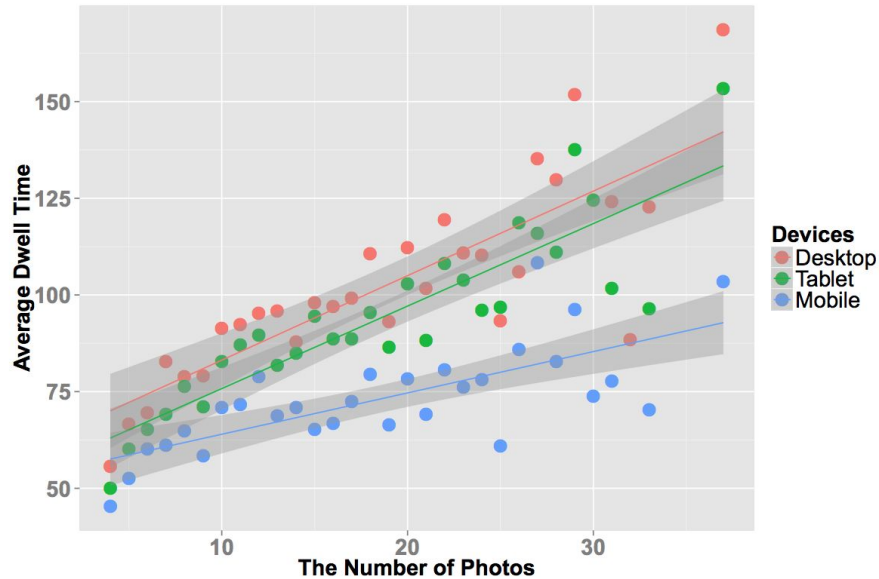


Figure 4: The relationship between the average dwell time and the number of photos on a slideshow where X-axis is the binned number of photos and the Y-axis is binned average dwell time.

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

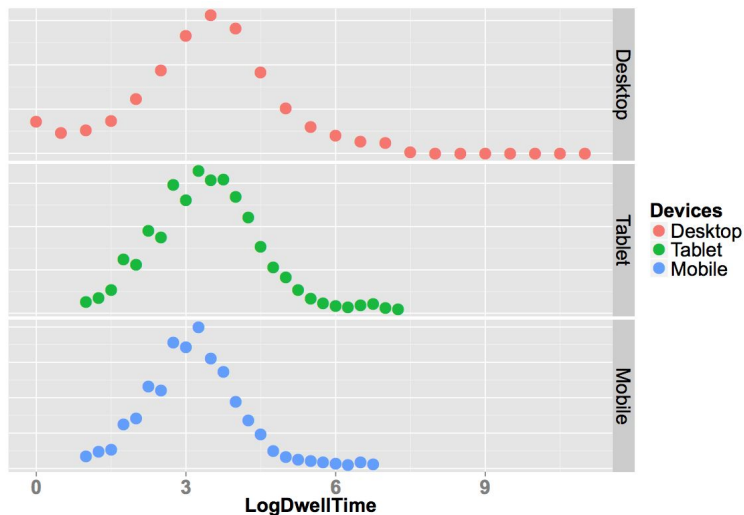


Figure 5: The (un)normalized distribution of log of dwell time for slideshows across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

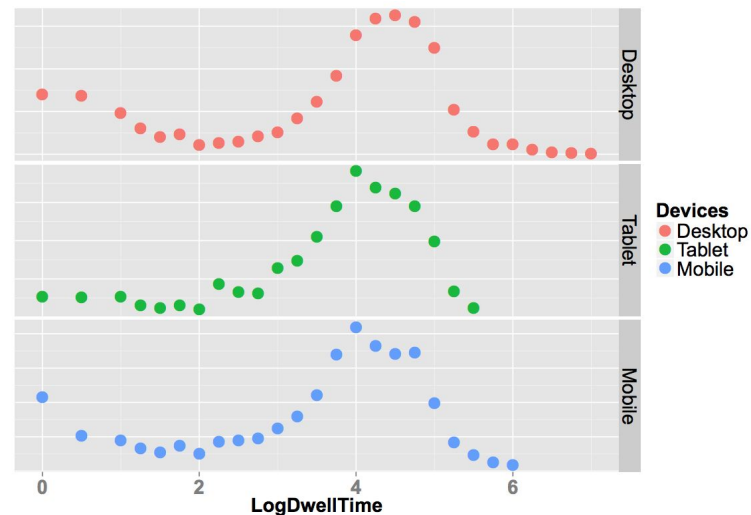


Figure 6: The (un)normalized distribution of log of dwell time for videos across different devices. The X-axis is the log of dwell time and the Y-axis is the counts.

Optimization Inter-Session Metrics

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

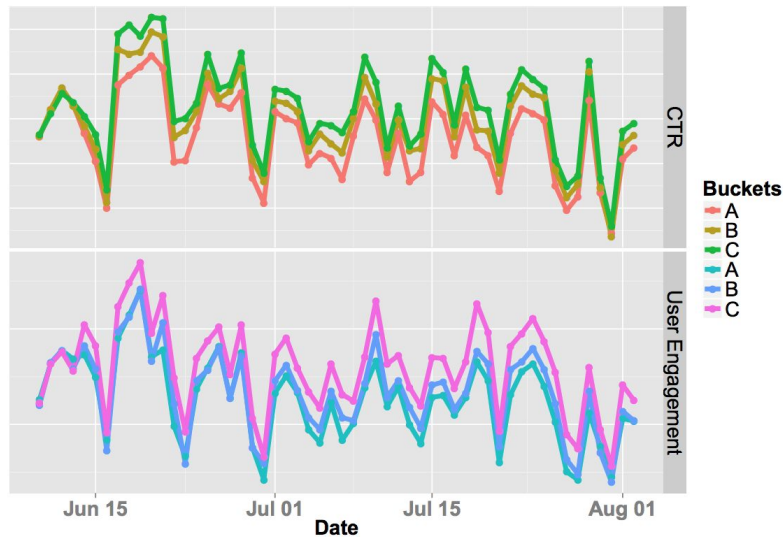


Figure 7: The relative performance comparison between three buckets. The top figure shows the relative CTR difference and the bottom figure shows the relative user engagement difference.

Optimization Inter-Session Metrics

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.

Optimization Inter-Session Metrics



Summary

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



Application: Search

Is this a good search engine?

[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 ...



Venice

Residential neighborhood in Los Angeles, California

Known for its bohemian spirit, Venice is a buzzing beach town with upscale commercial and residential pockets. Free-spirited Venice Boardwalk is the site of funky shops, street performers and colorful murals. There's also a skate park and Muscle Beach outdoor gym. Abbot Kinney Boulevard features foodie hot spots, stylish boutiques and coffee bars. A picturesque enclave of canals is surrounded by modernist homes.

Zip code: 90291
Area code: [Area codes 310 and 424](#)
Population: 40,885 (2008)
City: Los Angeles
Hotels: [Samesun Venice Beach](#), [Hotel Erwin](#), [MORE](#)

There is a rich history in evaluating ranking algorithms in information retrieval and web search

How to evaluate a search engine

Coverage
Speed
Query language
User interface



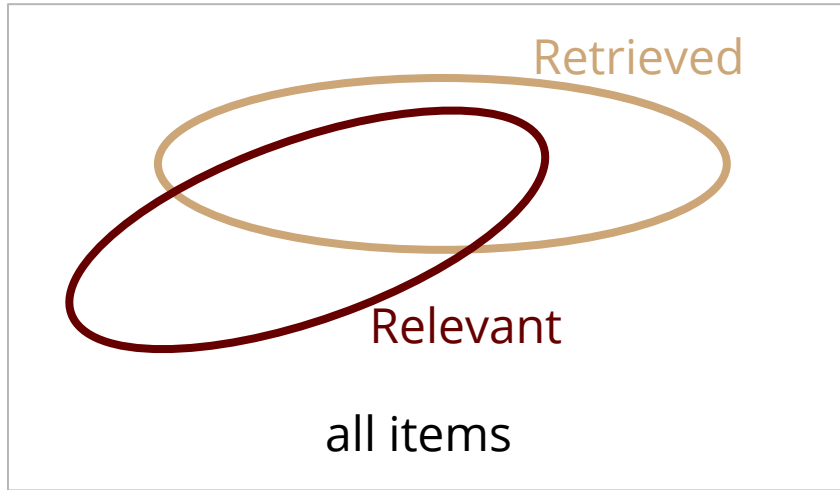
User happiness

- 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

Evaluating the relevance of a search engine result



User **information need** translated into a **query**

Relevance assessed relative to **information need** *not* the **query**

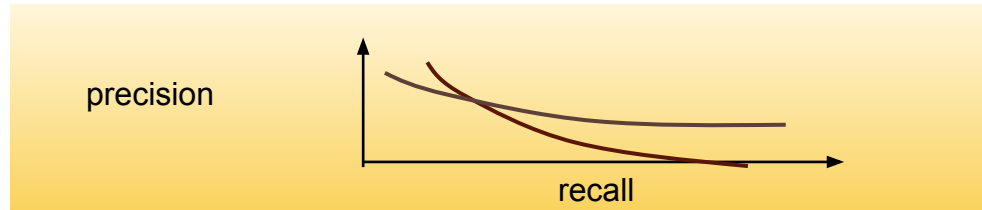
Example:

Information need: *I am looking for tennis holiday in a beach resort with lots of places to eat seafood*

Query: ***tennis academy beach seafood***

Evaluation measures:

- precision, recall, R-precision; precision@n; MAP; F-measure; ...
- bpref; nDCG; rank-biased precision, expected reciprocal rank,, ...



Evaluating the relevance of a search engine result

Explicit signals

- Test collection methodology (TREC, CLEF, NCTIR, ...)
- Human labeled corpora
- Crowdsourcing

Implicit signals

- User behavior in online settings (clicks, skips, dwell time)

Explicit and implicit signals can be used together

An important question:

when is signal a metric and when is it a feature of the ranking (machine learning) algorithm?

Examples of implicit signals ... measures ... metrics

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

Hover rate

The screenshot shows a Yahoo! search results page for the query "venice beach". The search bar at the top contains the text "venice beach" and a magnifying glass icon. To the right of the search bar, there is a user profile icon labeled "Mounia", an email icon, and the "YAHOO!" logo. Below the search bar, there are tabs for "Web", "Images", "Video", "News", "More", and "Anytime". The main content area displays search results for "venice beach". It includes a section for "Also try: venice beach california, venice beach boardwalk" and "Ads related to: venice beach". The primary result is "15 Hotels in Venice Beach - Best Price Guarantee" from Booking.com, with a sub-headline "Best Price Guarantee! Book your Hotel in Venice Beach." and a description: "Types: Hotels, Apartments, Villas, Hostels, Resorts, B&Bs. Book your hotel in Venice Beach, Los Angeles online." Below this are two columns of hotel categories: "Most Popular Hotels" (No reservation costs, Great rates, Safe, 100% Secure Payment), "Luxury Hotels" (Manage your bookings online, Easy and Secure Online Booking), "Book your Hotel Online" (24/7 Customer Service, We speak your language), "Budget Hotels" (Half-Price Hotels, Quick, Simple, Easy to Use), "Best Reviewed Hotels" (Read Real Guest Reviews, We Verify All Reviews), and "Get Instant Confirmation" (No Booking Fees, Free cancellation on most rooms). At the bottom, there is a section for "11 Hotels Venice from \$41 | trivago.com" with a sub-headline "trivago.com/Hotel-Venice" and a description: "trivago.com has been visited by 100K+ users in the past month. trivago™ Save Up To 63% on Hotels. Compare over 200 Booking Sites!". Below this are two columns of hotel categories: "Best Rated" (3* Hotels, Save Time & Money) and "Central Hotels" (4* Hotels). To the right of the search results, there is a map of Venice Beach, California, showing the coastline and surrounding areas like Santa Monica, Culver City, and Playa Vista. Below the map is a section for "Venice" with a description: "Beachfront neighborhood in Los Angeles, California. Venice is a residential, commercial, and recreational beachfront neighborhood within Los Angeles, California. It is located within the urban region of western Los Angeles County known as the Westside. wikipedia.org". Below the description is a link to "laparks.org" and a section for "Explore nearby" with icons for Hotels, Restaurants, Bars, Coffee, and Malls. At the bottom of the page, there is a small text "Image: Wikipedia".

An important question:
when is signal a metric and when is it a feature of the ranking (machine learning) algorithm?

What is a happy user in search?

1. The user information need is satisfied
2. The user has learned about a topic and even about other topics
3. The system was inviting and even fun to use



Intra-session

The actual search session

Inter-session

Users come back soon and frequently

The screenshot shows a Google search for "white wine". The search bar contains "white wine" and the search button is visible. Below the search bar, there are tabs for "All", "Shopping", "Images", "News", "Videos", "More", "Settings", and "Tools". The search results show "About 58,400,000 results (0.55 seconds)". The first result is "White Wines You'll Love | Drizly" with a link to "https://drizly.com/white-wine/c8". Below this, there is a "People also ask" section with four questions: "What is best white wine?", "What is a substitute for white wine?", "Is drinking white wine bad for you?", and "What is the best type of white wine?". Below this, there is a result for "The 7 major types of white wines - French Scout" with a link to "www.frenchscout.com/types-of-white-wines". At the bottom, there is a result for "White wine - Wikipedia". On the right side of the search results, there is a "White wine" card with a grid of images and a "More Images" link. Below the images, there is a "White wine" title and a description: "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". Below the description, there is a "Nutrition Facts" section for "White wine" with a table showing "Amount Per 1 serving 5 fl oz (147 g)" and "Calories 120".

Evaluating the actual search session

... Metrics

Mean average precision (MAP)

Number of clicks or CTR

Dwell time

Well established metrics of engagement with search results
Used as metrics to optimize in ranking algorithms
Also can be used as features in ranking algorithms

But how do they relate to user engagement?

→ inter-session consideration

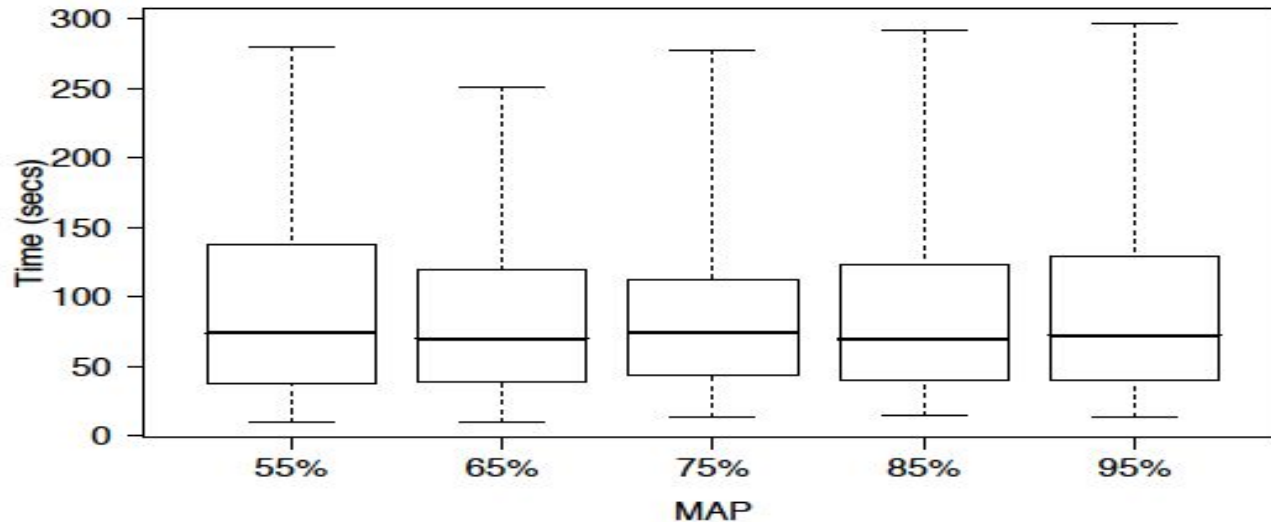


Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

Similar results obtained with P@2, P@3, P@4 and P@10

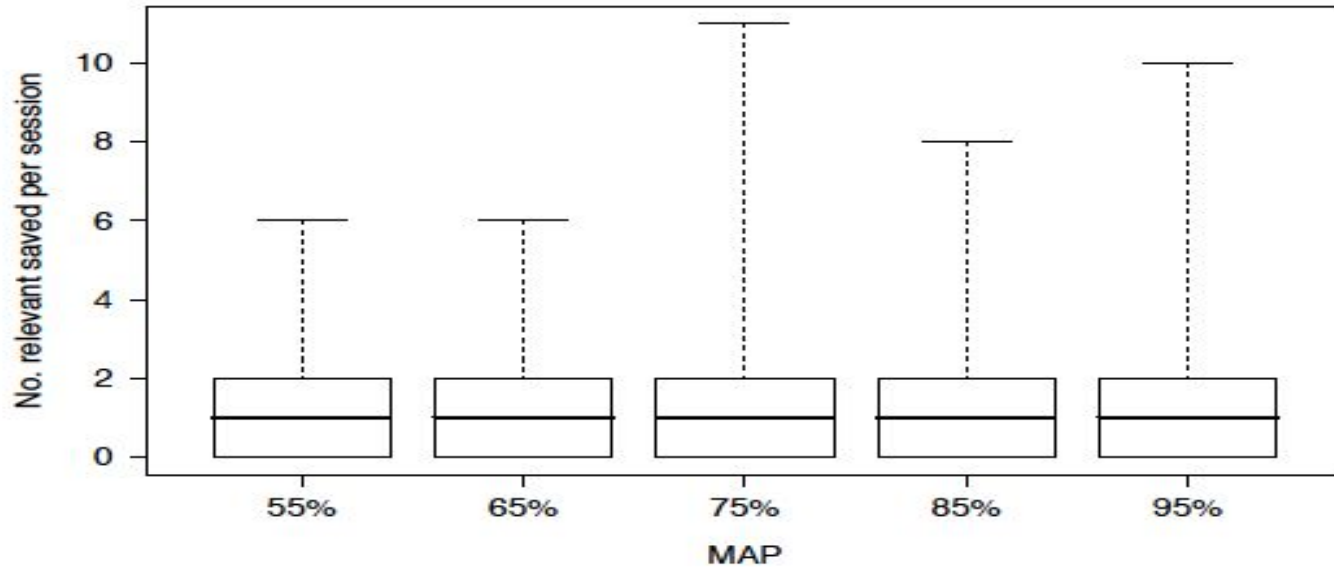


Figure 7: Number of relevant documents found by users within five minutes for systems with differing MAP.

No click

... User satisfaction



Angel Lane surgery Dunmow

Search

3,110 results

WEB

IMAGES

VIDEO

NEWS

SHOPPING

MORE

Search: the Web only in UK only in Ireland

FILTER BY TIME

Anytime

Past day

Past week

Past month

[Angel Lane Surgery - Essex](#)

Welcome to **Angel Lane Surgery** . We aim to provide you and your family with the best possible healthcare. **Angel Lane Surgery Angel Lane Great Dunmow Essex CM6 1AQ**
www.angellanesurgery.co.uk - [Cached](#)

[Useful Contacts - Angel Lane Surgery - Essex](#)

Angel Lane Surgery: Appointments/Enquiries: 01371 872 122: Hospitals ... Registrar of Births, Deaths and Marriages (**Dunmow**) ...
www.angellanesurgery.co.uk/useful.asp - [Cached](#)

[Overview - Angel Lane Surgery - NHS Choices](#)

Angel Lane Surgery. Telephone: 01371 872122 Address: **Angel Lane** , Great **Dunmow**, **Dunmow**, Essex, CM6 1AQ Website: Website address not added
www.nhs.uk/Services/gp/Overview/DefaultView.aspx?id=6F... - [Cached](#)
[More results from nhs.uk »](#)

I just wanted the phone number ... I am totally happy

No click

... User satisfaction

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

Clickthrough rate:

% of clicks when URL shown (per query)

Hover rate:

% hover over URL (per query)



Unclicked hover:

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

Max hover time:

Maximum time user hovers over a result (per SERP)

No click

... User satisfaction

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)

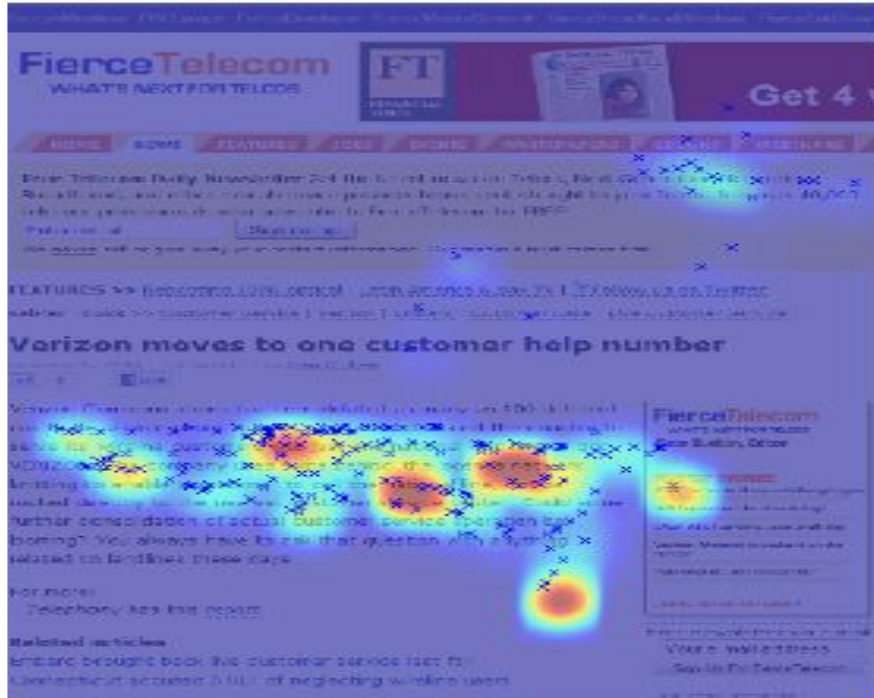
Cursor speed

Average cursor speed (pixel/second)

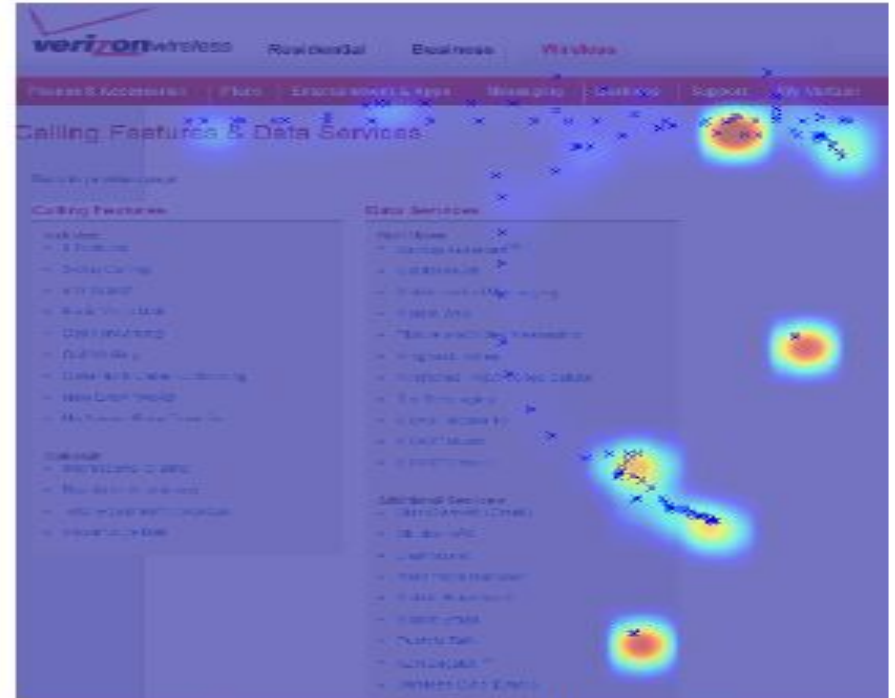
Slower when answers in snippet (good abandonment)

Dwell time

... User satisfaction



(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 (both dwell time of 30s)

(Guo & Agichtein, 2012)

Dwell time

... User satisfaction



(a) relevant (dwell time: 70s)



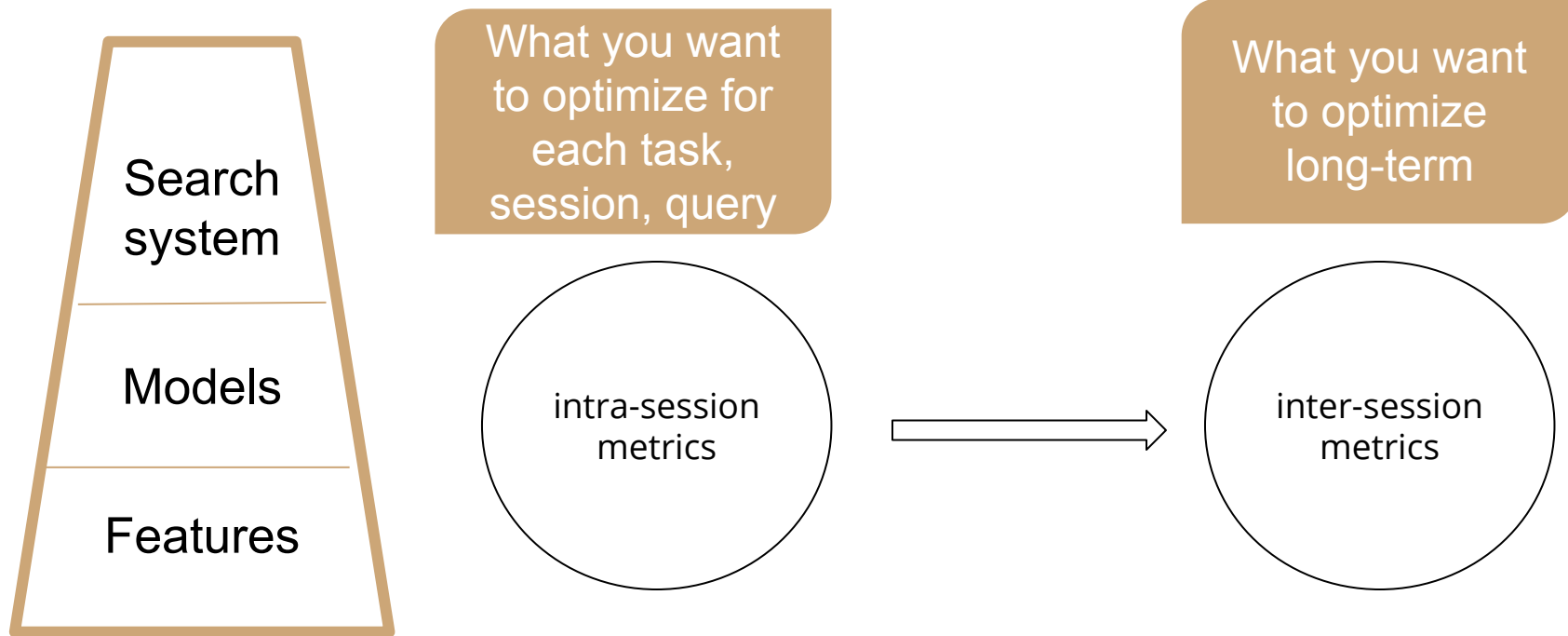
(b) non-relevant (dwell time: 80s)

“reading” a relevant long document vs “scanning” a long non-relevant document

(Guo & Agichtein, 2012)

From intra- to inter-session metrics

... We recall



From intra- to inter-session metrics

Intra-session metrics for search

(Proxy: relevance of search results)

- Number of clicks
- Time to 1st click
- Skipping
- Dwell time
- Click through rate
- Abandonment rate
- Number of query reformulations
- Hover rate
- ...

users satisfied with the search session are likely to return sooner and frequently to the search engine

Inter-session metrics for search

- Absence time
- Number of search sessions in next 2 weeks
- Number of queries next day
- ...

Absence time on Yahoo Japan (Dupret & Lalmas, 2013)

Absence time on Bing (Chakraborty et al, 2014)

Dwell time & search engine re-use (Hu et al, 2011)

Search result page for "asparagus" ... Study I

7,590,000 results

WEB IMAGES VIDEO SHOPPING RECIPES BLOGS MORE

RELATED SEARCHES

- asparagus **recipes**
- asparagus **soup**
- how to cook** asparagus
- broccoli**
- spinach**

FILTER BY TIME

- Anytime**
- Past day
- Past week
- Past month

Also try: [asparagus recipes](#), [asparagus soup](#), [how to cook asparagus](#), [more...](#)

Asparagus - Fine Gardening
www.finegardening.com



- Perennials
- Height: 1 ft. to 3 ft.
- Spread: 3 ft. to 6 ft.
- Hardiness Zones: 4 5 6 7 8
- Growth Pace: Moderate Grower
- Light: Part Shade Only
- Moisture: Medium Moisture
- Maintenance: Moderate

[Go to the USDA Hardiness Zone Finder >](#)

Asparagus - Wikipedia, the free encyclopedia

Asparagus officinalis is a spring vegetable, a flowering perennial plant species in the genus **Asparagus**. It was once classified in the lily family, like its Allium ...
en.wikipedia.org/wiki/Spargel - [Cached](#)
[More results from en.wikipedia.org >](#)

Asparagus - Food.com Kitchen Dictionary
www.food.com



Asparagus (from the Persian word asparag, meaning a sprout), are slim green spears, often tinged with a bit of purple at the tip. ...

[More about this term >](#)

Everything About **Asparagus** ... and More!

We are dedicated to spreading the good word about the virtues of **asparagus**, one of nature's most perfect foods.
www.asparagus.org - [Cached](#)

Asparagus - Recipe Search
recipes.search.yahoo.com



Baked Asparagus with Balsamic Butter Sauce
allrecipes.com
★★★★☆ (1185)
Total time: **25 mins**

Ingredients (6): fresh asparagus, cooking spray, salt, butter, soy sauce, more...



Pan-Fried Asparagus
allrecipes.com
★★★★☆ (923)
Total time: **25 mins**

Ingredients (6): butter, olive oil, coarse salt, ground black pepper, garlic, more...

Ads

Asparagus Supplements
Buy 1 Get 2 Free 450mg 90ct 10.99
www.doctorvitaminstore.com

Cooking Fresh Asparagus
Find Cooking Fresh **Asparagus** at Great Prices.
www.Pronto.com

Asparagus Roots
Asparagus Roots Info. Research now on About.com.
About.com/Asparagus_Roots

Asparagus plants for less
Buy **asparagus** plants for less money. Free shipping!!!
www.asparagus4you.com

How To Grow Asparagus
Get Answers & References On Ask.com. Try Asking Now!
Ask.com

Discount Asparagus Plant
Top **Asparagus** Plant - top prices! Save up to 75% on **Asparagus** Plant.
www.Asparagus-Plant.BEST-PRICE.com

Asparagus Seed
Huge Selection of **Asparagus** Seeds. Superior Quality, Great Prices!
www.ParkSeed.com

More Sponsors:
[asparagus](#)
[cook asparagus](#)
[grilled](#)

[See your message here...](#)

Another search result page for “asparagus”

[Asparagus - Wikipedia, the free encyclopedia](#)

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

[Biology](#) · [History](#) · [Uses](#) · [Cultivation](#) · [Commercial production](#)

Asparagus officinalis is a spring vegetable, a flowering perennial plant species in the genus **Asparagus**. It was once classified in the lily family, like its *Allium* ...

[Everything About Asparagus... and More!](#)

[www.asparagus.org](#)

We are dedicated to spreading the good word about the virtues of **asparagus**, one of nature's most perfect foods.

[Videos of asparagus](#)

[bing.com/videos](#)



[How to Take Care of Asparagus Plants](#)
eHow

[How to Make Asparagus Bundles](#)
eHow

[How To Cook Asparagus](#)
Bing Video

[Reheating Cooked Asparagus](#)
eHow

[Asparagus Recipes: Roasting and Grilling ...](#)

[www.foodnetwork.com](#) > [Topics](#) > [Seasonal](#)

Find new ways to cook with **asparagus**, including grilling and roasting **asparagus**, as well as an amazing **asparagus** soup recipe from ...

[Top recipes](#) >

[Growing Asparagus In The Home Garden, HYG-1603-94 - Ohioline](#)

[ohioline.osu.edu/hyg-fact/1000/1603.html](#)

Asparagus is a long-lived perennial vegetable crop that is enjoyed by many gardeners. Soil requirements **Asparagus** grows in most any soil as

[Images of asparagus](#)

[bing.com/images](#)



[WHFoods: Asparagus](#)

[www.whfoods.com/genpage.php?tname=foodspice&dbid=12](#)

Asparagus. The fleshy green spears of **asparagus** are both succulent and tender and have been considered a delicacy since ancient times. This highly prized vegetable ...

[Asparagus - Simply Recipes Food and Cooking Blog](#)

[www.simplyrecipes.com/recipes/asparagus](#)

Quick and easy **asparagus** recipe. How to cook **asparagus** spears perfectly, dress with olive oil, Parmesan, and lemon zest.

Asparagus



[en.wikipedia.org](#)

Asparagus officinalis is a spring vegetable, a flowering perennial plant species in the genus *Asparagus*. It was once classi...
[en.wikipedia.org](#)

Scientific Name: *Asparagus officinalis*

Biological Classification: [Species](#)

Belongs to: [Asparagus](#)

People also search for



[Broccoli](#)



[Celery](#)



[Spinach](#)

Data from: [wikipedia](#) · [freebase](#)

[Report a problem](#)

Ads

[Cooking Fresh Asparagus](#)

[www.Pronto.com](#)

Find Cooking Fresh **Asparagus** at Great Prices.

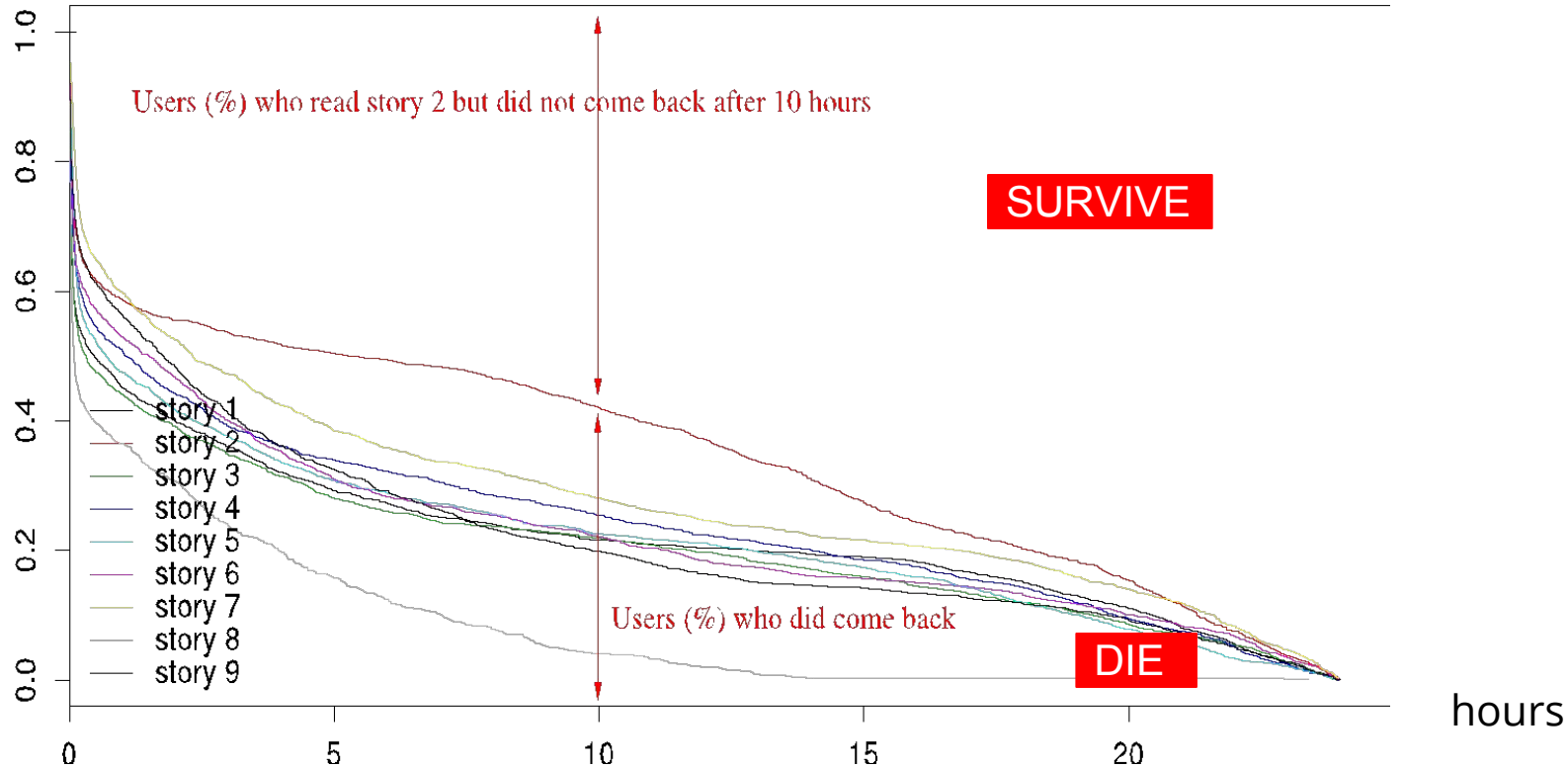
[Asparagus Supplements](#)

[www.doctorvitaminstore.com](#)

Buy 1 Get 2 Free 450mg 90ct 10.99

[See your message here](#)

Absence time and survival analysis



DIE = RETURN TO SITE → SHORT ABSENCE TIME

Absence time applied to search

... Study I

Ranking functions on Yahoo Answer Japan

The screenshot shows the Yahoo! Japan Q&A search interface. The search term is 'best sushi'. The results are ranked by relevance. The top result is 'What's your best sushi experience?' with 46 answers. The second result is a CNN article about 'The best sushi restaurants in Tokyo' with 23 answers. The third result is a question about favorite sushi with 175 answers. On the right side, there are sponsored search results for 'すしランキング' (Sushi Ranking) and '寿司 ランキング' (Sushi Ranking).

Session boundary:
30 minutes of
inactivity

Two-weeks click data on Yahoo Answer Japan search
One millions users

Six ranking functions

DCG versus absence to evaluate five ranking functions



DCG@1

Ranking Alg 1

Ranking Alg 2

Ranking Alg 3

Ranking Alg 4

DCG@5

Ranking Alg 1

Ranking Alg 3

Ranking Alg 2

Ranking Alg 4

Absence time

Ranking Alg 1

Ranking Alg 2

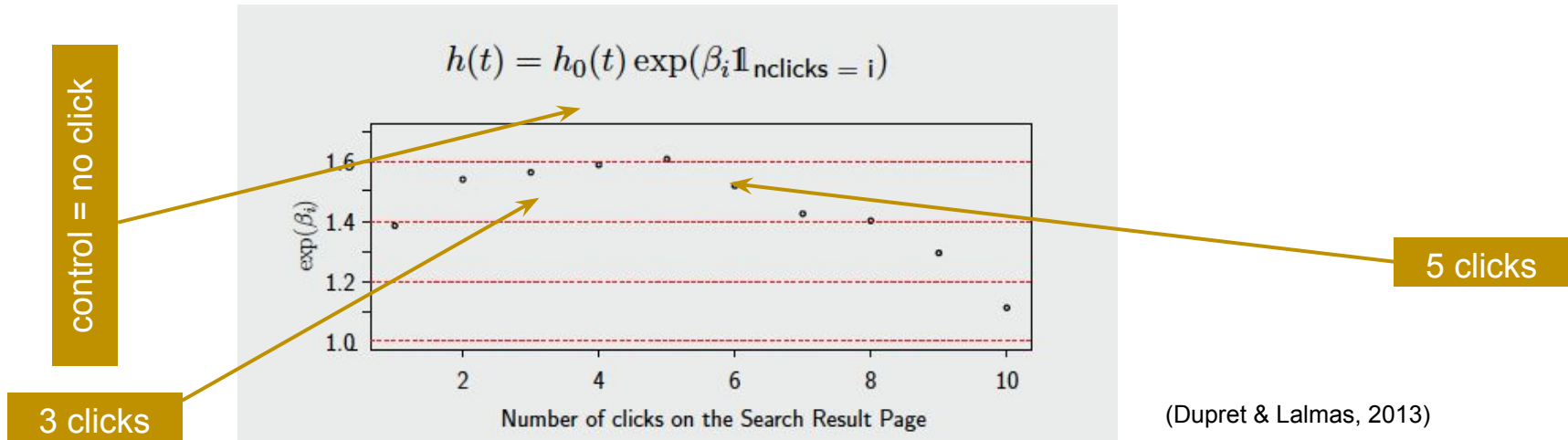
Ranking Alg 5

Ranking Alg 3

Ranking Alg 4

Absence time and number of clicks

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



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

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

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

Absence time and search session

... What else?

intra-session search metrics → absence time



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

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

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

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



Application: E-commerce

Application: E-commerce

The image displays two overlapping screenshots of e-commerce websites. The background screenshot is Amazon.com, showing search results for "liszt". The foreground screenshot is eBay.com, showing search results for "wabi sabi".

Amazon Screenshot (Background):

- Search bar: "liszt"
- Results: "1-16 of over 50,000 results for 'liszt'"
- Featured item: "LISZT Consolations for Violin" by Franz Liszt, priced at \$7.52 (Paperback) or \$3.95 (Kindle).
- Navigation: "Departments", "Browsing History", "Today's Deals", "Gift Cards", "Registry", "Sell", "Help".

eBay Screenshot (Foreground):

- Search bar: "wabi sabi"
- Results: "All categories > 'wabi sabi' (6,213 Results)"
- Filters: "wabi sabi art", "wabi sabi ceramics", "wabi sabi bowl", "wabi sabi pottery", "wabi sabi necklace", "wabi sabi jewelry".
- Shipping options: "Free shipping", "Ready to ship in 1 business day", "Ready to ship within 3 business days".
- Shop location: "Anywhere" (selected), "United States", "Custom".
- Featured items:
 - Kintsugi bowl, kintsugi ceramic enam... KanelaSuri, \$84.41, 5 stars (77).
 - BIGFOOT Bowl, Charcoal Special | M... Odaka, \$42.00, 5 stars (31).
 - Wabi-sabi Oversize Clutch bag for w... SCHILLERahop, \$65.00, 5 stars (107).
 - Wabi-Sabi definition, dictionary art p... footnotestudios, \$4.99, 5 stars (87).
- Navigation: "Hi Sign in or register", "Daily Deals", "Gift Cards", "Help & Contact", "Perfect-for-Them Valentine Gifts".

Application: E-commerce

- **Search**
- **Recommendation**
- **Advertising**

Application: E-commerce

- **Search**
 - **Recommendation**
 - **Advertising**
-

- **Shopping**
- **Discovery**

...

Application: E-commerce



Application: E-commerce

- **Search**
 - Generic search v.s. E-commerce search
 - Relevance
 - Revenue
 - Diversity
 - Discovery
- **Recommendation**
 - Rating/favorite prediction
 - Clicks and purchase funnel
 - Revenue
 - Seasonal
 - Occasion
 - Inventory
- **Advertising**
 - Two-sided marketplace

Application: E-commerce

- **Search**
 - **Recommendation**
 - **Advertising**
-

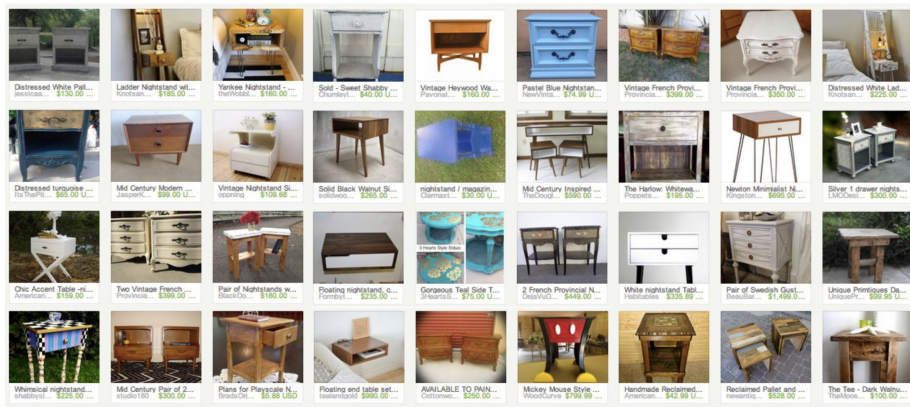
- **How to measure**
- **How to optimize**

Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

How do people decide what to buy?

Function and style. Example: search results for “nightstand” - 100+ pages



- **Reference:**

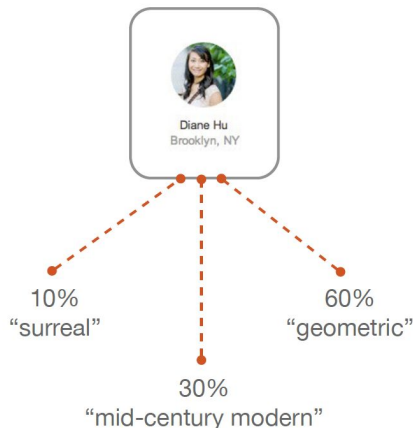
[1] Diane J. Hu, Rob Hall, and Josh Attenberg. **Style in the Long Tail: Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce.** In KDD 2014.

Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

Latent Dirichlet Allocation (LDA)

Learn **style profiles** for each user using LDA



1 Define what each style looks like

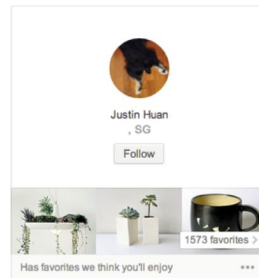


= "mid-century modern"

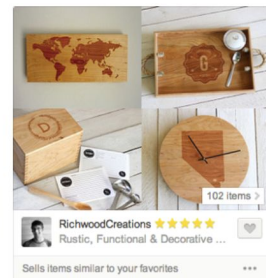
2 Use style profiles to generate personalized content



ITEM RECS



USER REC

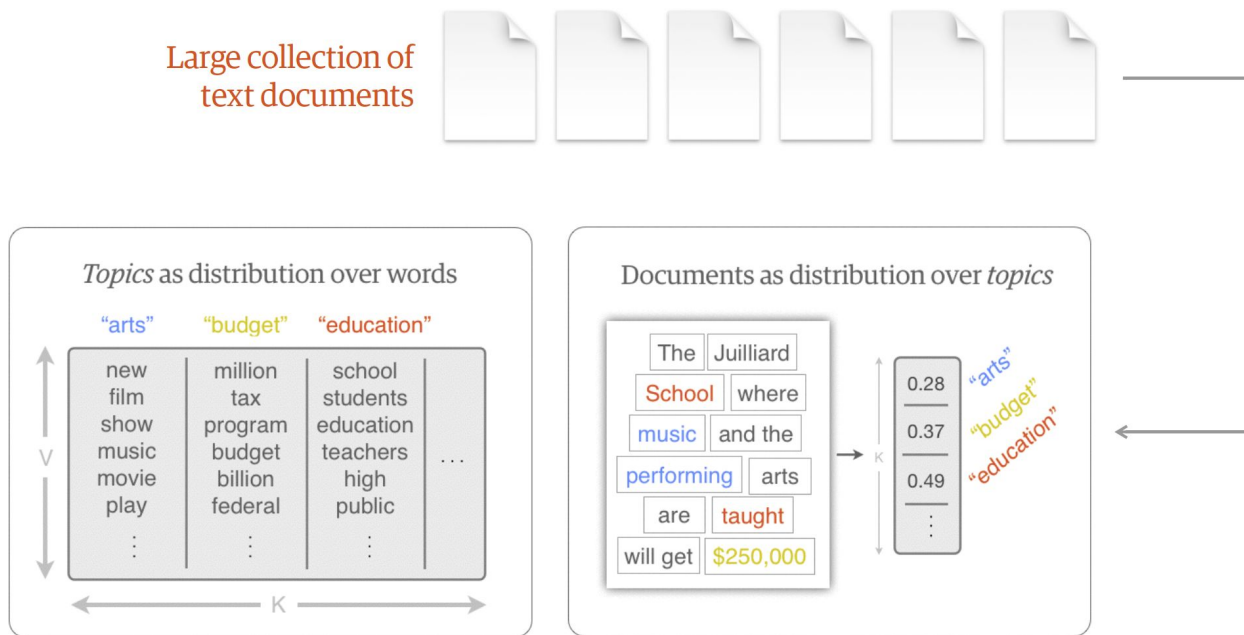


SHOP REC

Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

Latent Dirichlet Allocation (LDA)

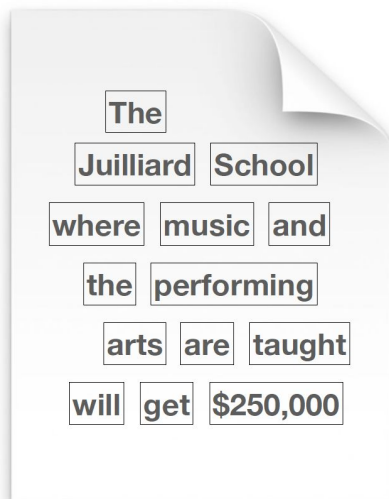


Application: E-commerce

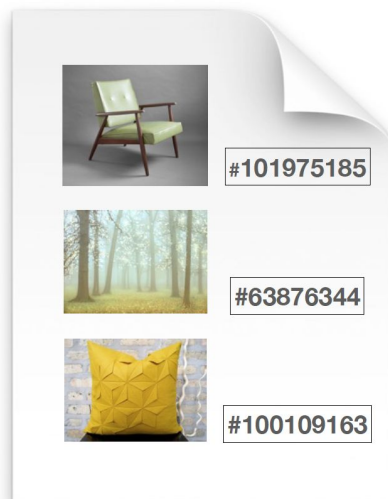
- **Discovering Styles for Recommendation in E-Commerce**

Latent Dirichlet Allocation (LDA)

Article
about Juilliard



Diane's
favorited items



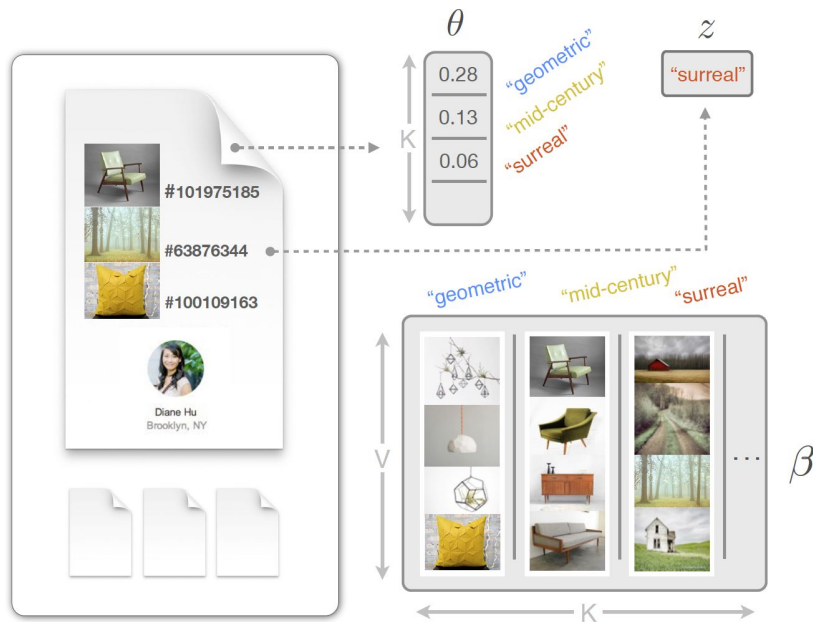
Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

Assume: Each user's favored items are generated by this process:

For each user u ,

1. Draw a style profile:
 $\theta \sim \text{Dirichlet}(\alpha)$
2. For each item, x_n
that user u has favorited,
 - (a) Draw a style:
 $z_n \sim \text{Multinomial}(\theta)$
 - (b) Draw an item:
 $x_n \sim \text{Multinomial}(\beta_{z_n})$

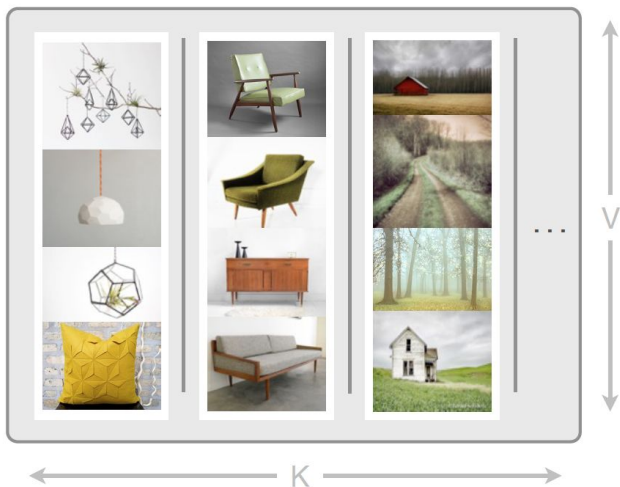


Application: E-commerce

- Discovering Styles for Recommendation in E-Commerce

Discover popular styles on Etsy as a distribution over items

“geometric” “mid-century” “surreal”



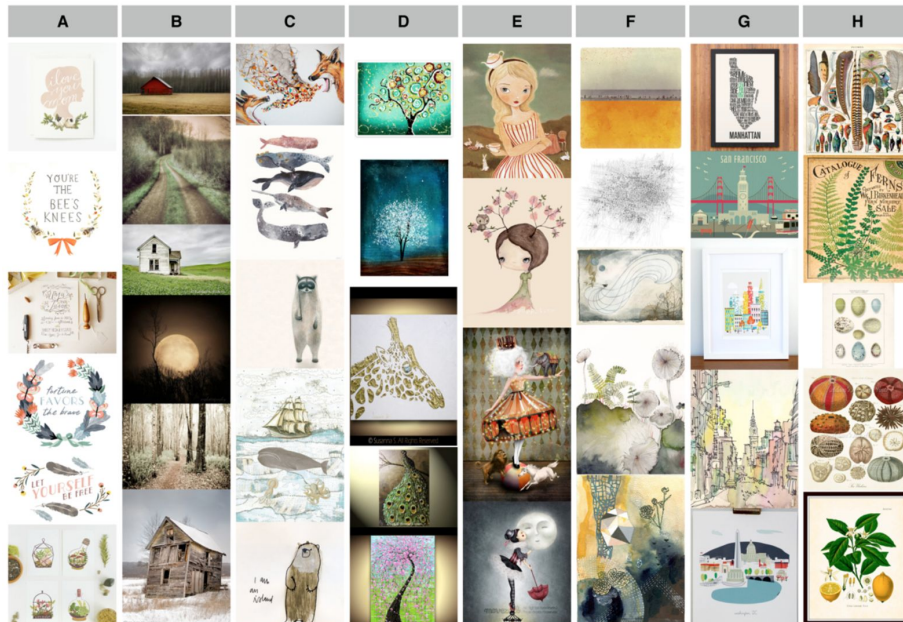
Represent each user as a distribution over popular styles, i.e. “style profile”



Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

LDA: Example Styles Discovered Within Category



Example of learned styles that contain art prints:

A = Botanical

B = Surreal landscapes

C = Whimsical

D = Acrylic/Abstract

E = French Dolls

F = Whimsical/Abstract

G = Cities

H = Vintage

Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

LDA: Example Styles Discovered Across Categories

ANIMALS



TENTACLES



GEOMETRIC



MID-CENTURY MODERN



Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

LDA: Generating Listing Recommendations

Given that each user has an style profile:

Recommend N listings from most highly weighted styles

MY FAVORITES



STYLE #428



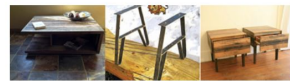
STYLE #54



STYLE #655



STYLE #87



Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

Metric	Control (95%)	On (Diff) (5%)
Conversion Rate	–	+ 0.32%
Pages Viewed Rate	–	+ 1.18%
Activity Feed Visit Rate	–	+ 7.51%
User Follow Rate	–	+ 13.43%
Item Favorite Rate	–	+ 2.81%
Shop Favorite Rate	–	+ 2.44%

Table 3: Stage 2 of user recommendation experiments with live A/B user testing. Bolded numbers in the *Diff* column indicate statistical significance.

Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

(1) Personalized Recommendations

Our Picks For You | Homepage & App

MaxMF + Item-based on Views/Faves/Purchases

Shop Recommendations | Homepage & App

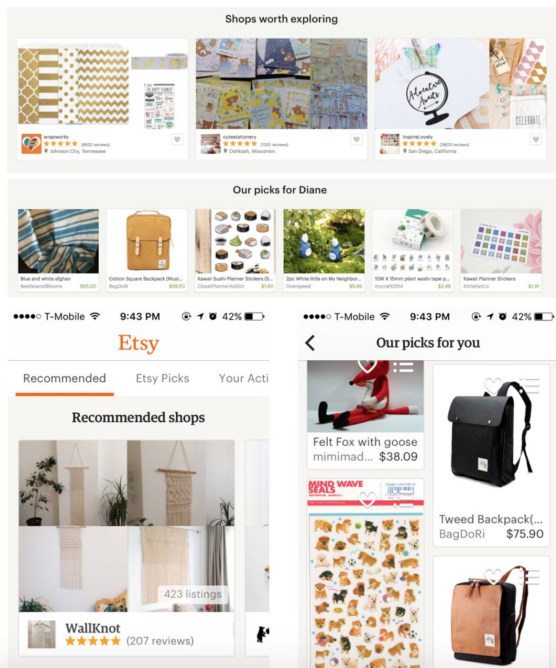
Latent Dirichlet Allocation on Favorites

Similar to Recently Viewed | App

Item-based on Views/Faves/Purchases

Personalized Etsy Finds | Email

MaxMF + Item-based on Views/Faves/Purchases



Application: E-commerce

- **Discovering Styles for Recommendation in E-Commerce**

(2) Substitute Recommendations

Find most similar listings based on TFIDF and Image Features

Products: Sold-out Listings, GPLA Listings, Mobile Listings, Leo Listings Page, Non-empty Cart Page

(3) Complementary Recommendations

From co-purchase data, find complementary taxonomy paths and suggest most similar listing in complementary taxonomy

Products: Leo Complementary Listings

(4) Trending Recommendations

Hubs & Authorities (HITS) finds influential users, and recommending listings/shops they favorite; Also, heuristics based on listings and shops that are dwelled/favorited frequently

Products: Local Shop Recs on Homepage

Application: E-commerce

- How to measure the success of recommender systems in E-commerce?
- How to construct unified framework to optimize recommendation in different modules/pages?
- How to measure *style, quality, artistic...?*

...

Application: E-commerce

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

The image shows two screenshots of the Etsy website. The top screenshot displays search results for 'wabi sabi', showing various product categories like 'wabi sabi art', 'wabi sabi ceramics', etc. The bottom screenshot shows search results for 'jewelry box', featuring a grid of product listings with details like price, ratings, and shipping options.

Top Screenshot: Search for 'wabi sabi'

- Search bar: wabi sabi
- Navigation: Sell on Etsy, Home, Favorites, You, Cart
- Categories: Jewelry & Accessories, Clothing & Shoes, Home & Living, Wedding & Party, Toys & Entertainment, Art & Collectibles, Craft Supplies & Tools, Vintage
- Filters: wabi sabi art, wabi sabi ceramics, wabi sabi bowl, wabi sabi pottery, wabi sabi necklace, wabi sabi jewelry
- Left Sidebar:
 - Special offers: On sale
 - All categories: Home & Living, Art & Collectibles, Jewelry, Craft Supplies & Tools
 - Shipping: Free shipping, Ready to ship in 1 business day, Ready to ship within 3 business days
 - Shop location: Anywhere, United States, Custom
- Results: All categories > "wabi sabi" (6,213 Results). Featured product: Kintsugi bowl, kintsugi ceramic er, KanelaSuri, \$84.41, Only 1 available, 5 stars (77).

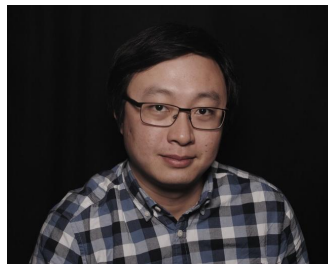
Bottom Screenshot: Search for 'jewelry box'

- Search bar: jewelry box
- Navigation: Sell on Etsy, Home, Favorites, You, Cart
- Categories: Jewelry & Accessories, Clothing & Shoes, Home & Living, Wedding & Party, Toys & Entertainment, Art & Collectibles, Craft Supplies & Tools, Vintage
- Filters: jewelry box wood, wooden jewelry box, large jewelry box, small jewelry box, jewelry box vintage, personalized jewelry box
- Left Sidebar:
 - All categories: Jewelry, Home & Living, Craft Supplies & Tools, Weddings
 - Shipping: Free shipping, Ready to ship in 1 business day, Ready to ship within 3 business days
 - Special offers: On sale
 - Shop location: Anywhere, United States, Custom
- Results: All categories > "jewelry box" (241,017 Results). Did you mean the shop [JewelryBox?](#)
 - Product 1: Raven box, handmade boxes, steamp... ST3jewellery, 5 stars (35), \$30.95
 - Product 2: Bridesmaid Gift / Popular Bridesmaid... SugarAndChicShop, 5 stars (1,208), \$45.00
 - Product 3: Matte Black Custom Branded Laser... Izrbeams, 5 stars (162), \$85.00
 - Product 4: Personalized Memory Box, Keepsake ... EngraveMyMemories, 5 stars (6,548), \$29.95, Eligible orders get 10% off

Application: E-commerce

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

- **Liang Wu**, PhD Student from Arizona State University
- **Diane Hu**, Staff Data Scientist at Etsy
- **Liangjie Hong**, Head of Data Science at Etsy



Application: E-commerce

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

- **Expected GMV**

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

Application: E-commerce

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

- Purchase Decision Process

rosy wedding dress

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

90 Colors Chiffon Rosy Long Party Dress Evenin...
CHARMINGDYY
★★★★★ (650)
\$51.50

Rosy brown dress chiffon party dress rosy brown...
LovelyMelodyClothing
★★★★★ (1,597)
\$39.00

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

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

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

\$39.00

Style

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

\$69.00+

Only 1 available

Size

Application: E-commerce

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

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

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

Application: E-commerce

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

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

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



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

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

Application: E-commerce

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

Application: E-commerce

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

Application: E-commerce

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

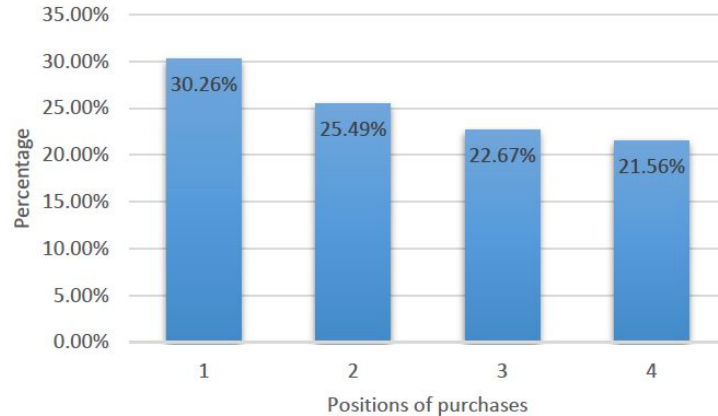


Figure 2: Position distribution of items being purchased in the top 4 spots of a search result page. The first position achieves the most purchases, while nearly 70% of purchases are in the lower positions.

Application: E-commerce

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	BM_{25}
			Log BM_{25}
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]	LNNet	
	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	

Application: E-commerce

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.

Application: E-commerce

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.

Application: E-commerce

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.



Recap and open challenges

Recap

- Introduction and Scope
- Towards a Taxonomy of Metrics
- Experimentation and Evaluation of Metrics
- Optimisation for Metrics
- Applications
 - Search
 - E-commerce

Challenges

- How to systematically discover new metrics?
- How to measure metrics (metrics of metrics)?
- How to quantify users' holistic feelings?
- Can we *learn* metrics?
- Advance methodologies to optimize intra- and inter-session metrics.



References

... to come