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 … Outline

Who we are

What is user engagement

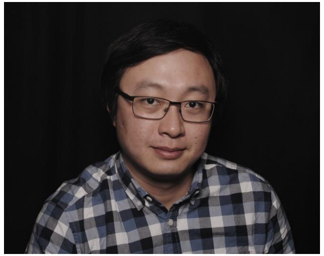
Approaches to measure user engagement

The focus of this tutorial

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).

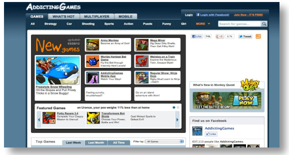
All the above can translate into the "emotional, cognitive and behavioural **connection** that exists, at any point in time **and** over time, between a user and a technological resource" (O'Brien, Lalmas & Yom-Tov, 2013)

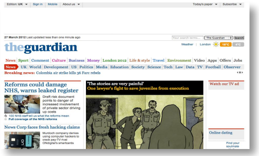
Why is it important to engage users?

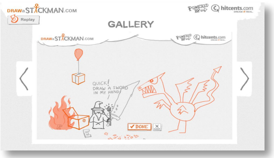
Users have increasingly enhanced expectations about their interactions with technology

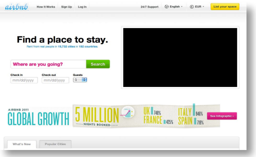
... resulting in increased competition amongst the providers of (online) services.

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

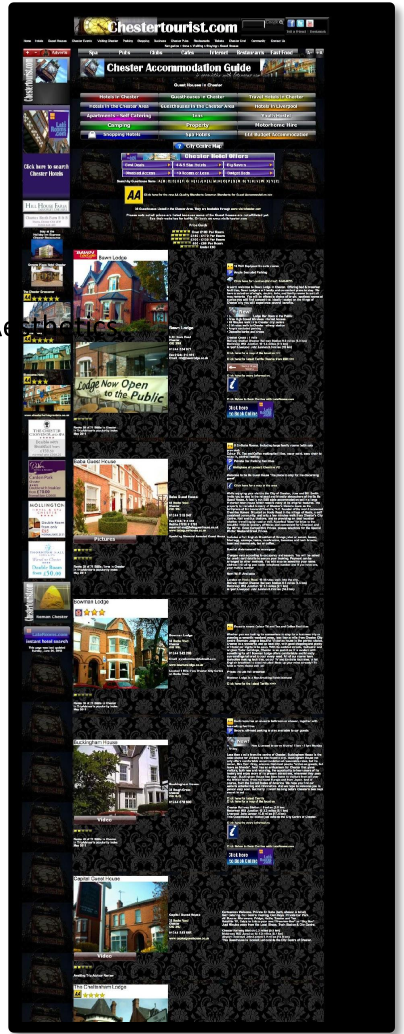






(O’Brien, Lalmas & Yom-Tov, 2013)



-Aesthetics

Is this site engaging?

aesthetics

leisure

Is this site engaging?

shopping

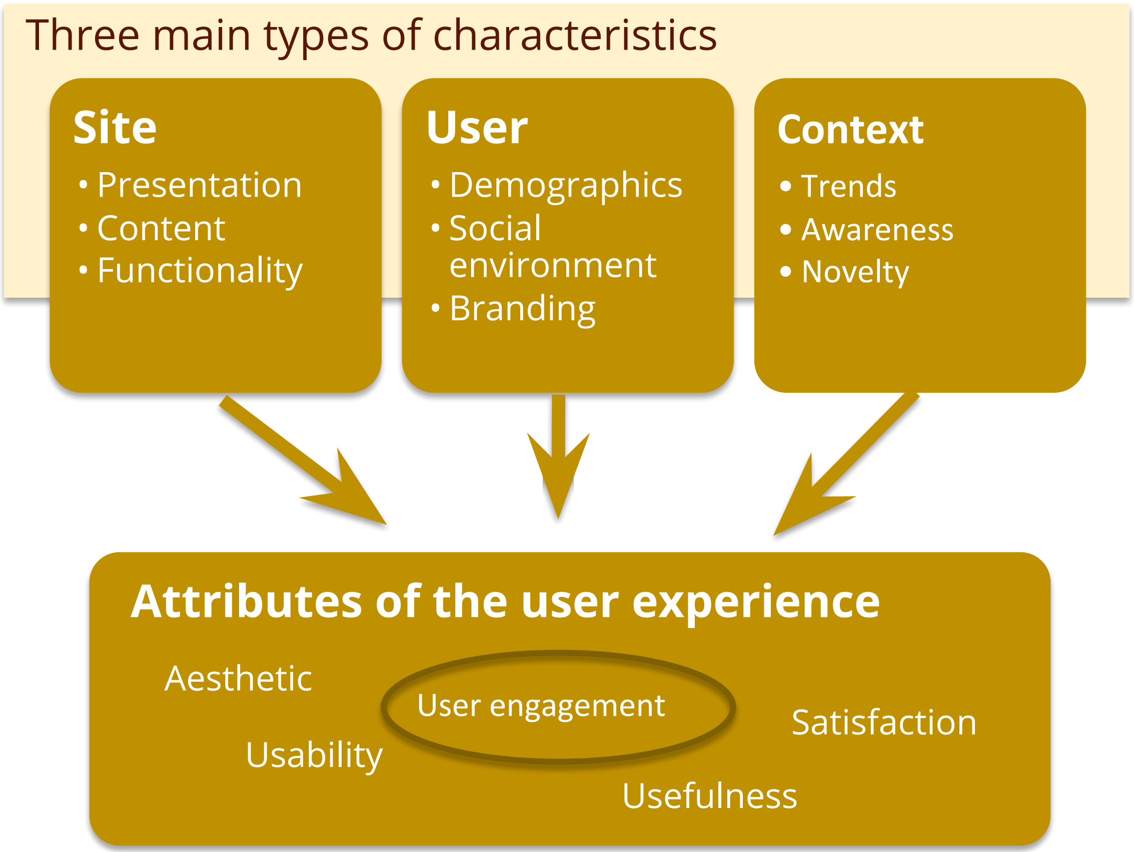
usability

Is this site engaging?

trust

news

What influences user engagement?



Many connections

Considerations in measuring user engagement

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

(O’Brien, Lalmas & Yom-Tov, 2013)

Methods to measuring user engagement

Self-reported engagement
subjective

Questionnaire, interview, report, product reaction cards

User study (lab/online)

mostly qualitative

Cognitive engagement
objective

Task-based methods

Neurological

Physiological

User study (lab/online)

mostly quantitative, scalability an issue

Interaction engagement
objective

Analytics
Data science

Data study

quantitative, large scale

Metrics … Our focus

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 ... Outline

Terminology, context & consideration

Facets of user engagement

Sessions and metrics

Intra-session metrics

Inter-session metrics

Other metrics

Proposed taxonomy

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

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

Three levels of metrics

our focus in this section

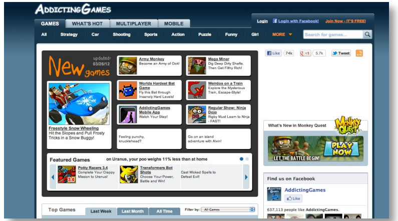
Business metrics -- KPIs

Behavioral metrics -- online metrics, analytics

Optimisation metrics -- metrics used to train machine learning algorithms

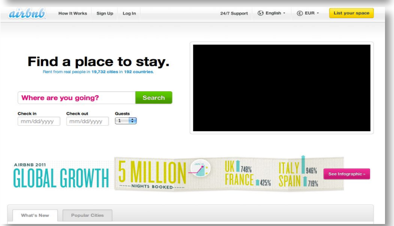
These three levels are connected

Why do we need several metrics of online behaviour?

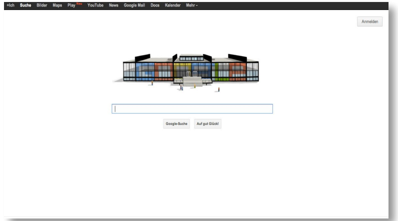


Games
Users spend much time per visit

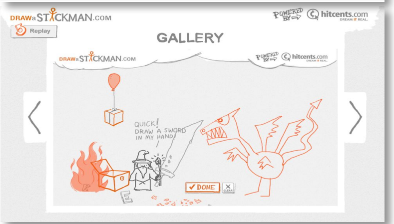
Social media
Users come frequently and stay long



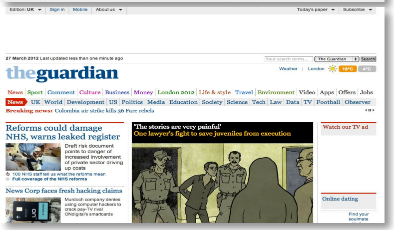
Service
Users visit site, when needed, e.g. to renew subscription



Search
Users come frequently and do not stay long



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



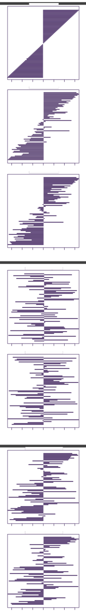
News
Users come periodically, e.g. morning and evening

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



$\tau_{intra} = 0.61$ $\tau_{inter} = 0.23$

(Lehmann et al., 2012)

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)

(Lehmann et al., 2012)

What may impact user engagement?

Why?	**Who?**	**When?**	**Where?**	**What?**
Task	Demographics	Temporality	View	Function
	Recency	Usage level	Platform	

Segmentation

Temporality … When?

work-related

daily news

	popularity	activity	loyalty
model m_{t5}	**{wd}++**		
model m_{t4}	**{wd}++**	**{wd}++**	
model m_{t3}		**{we}++**	**{we}++**
model m_{t2}	**{we}++**		**{wd}++**
model m_{t1}	**{we}++**		**{we}++**

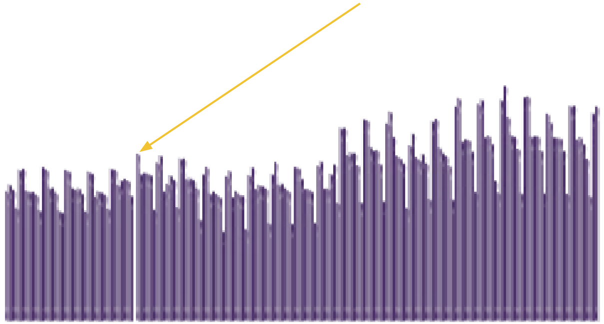
average ++ *high* **{wd}** *weekdays* **{we}** *weekends*

hobbies,
interest-specific
weather

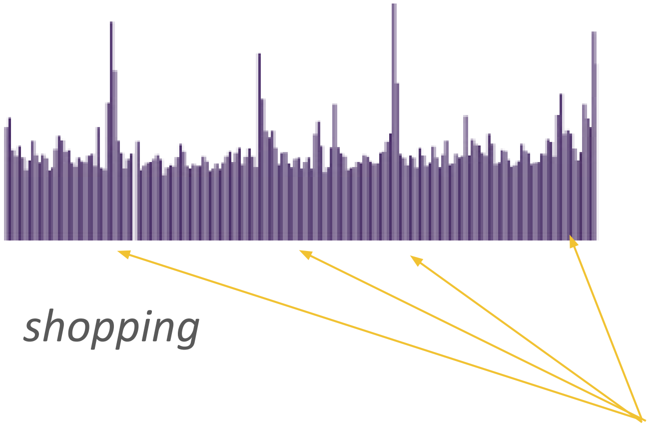
Engagement varies from weekdays to weekends (Lehmann et al., 2012)

Temporality … When?

site crash



Social network



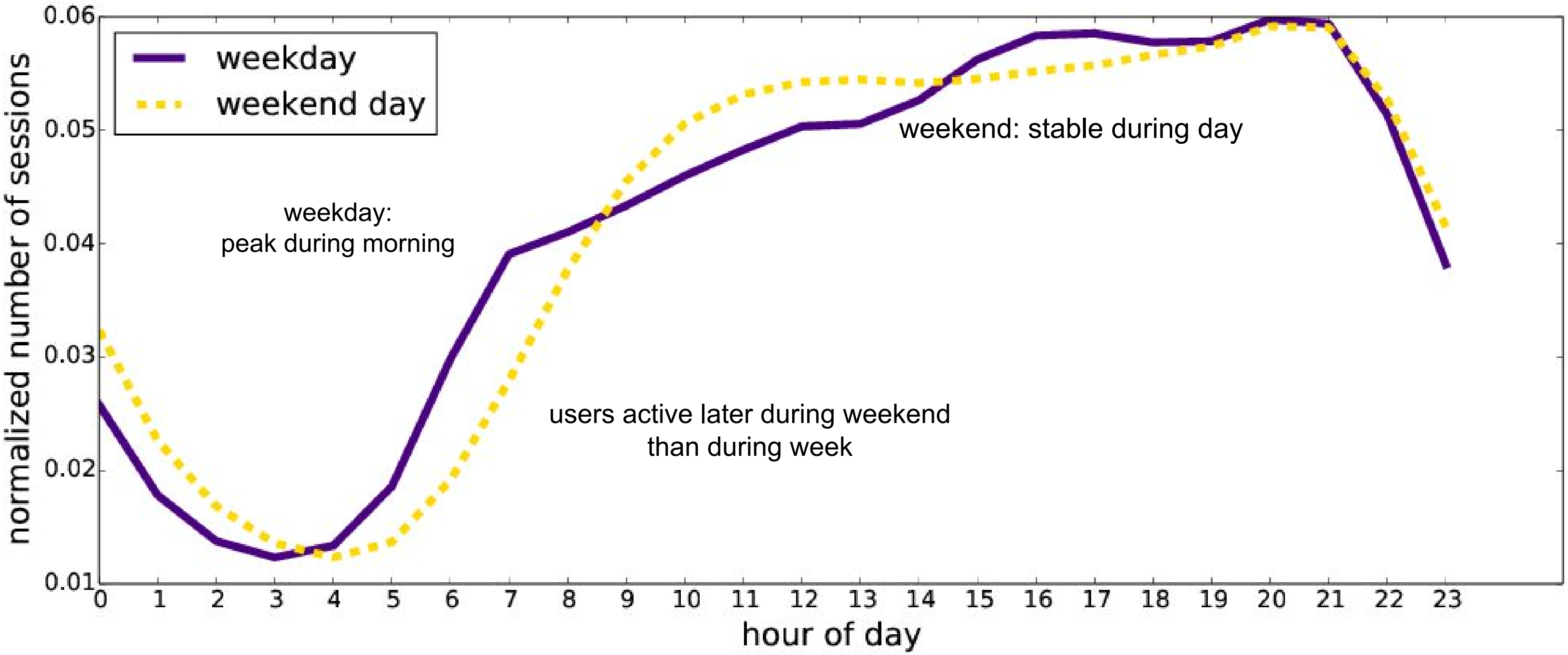
shopping

payday

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

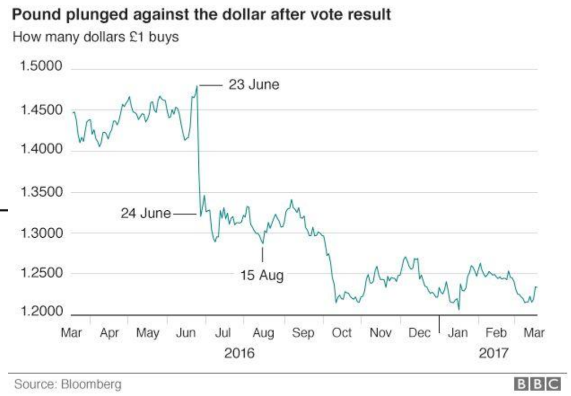
Engagement is influenced by internal (e.g. crash) and **external** factors (e.g. major events)

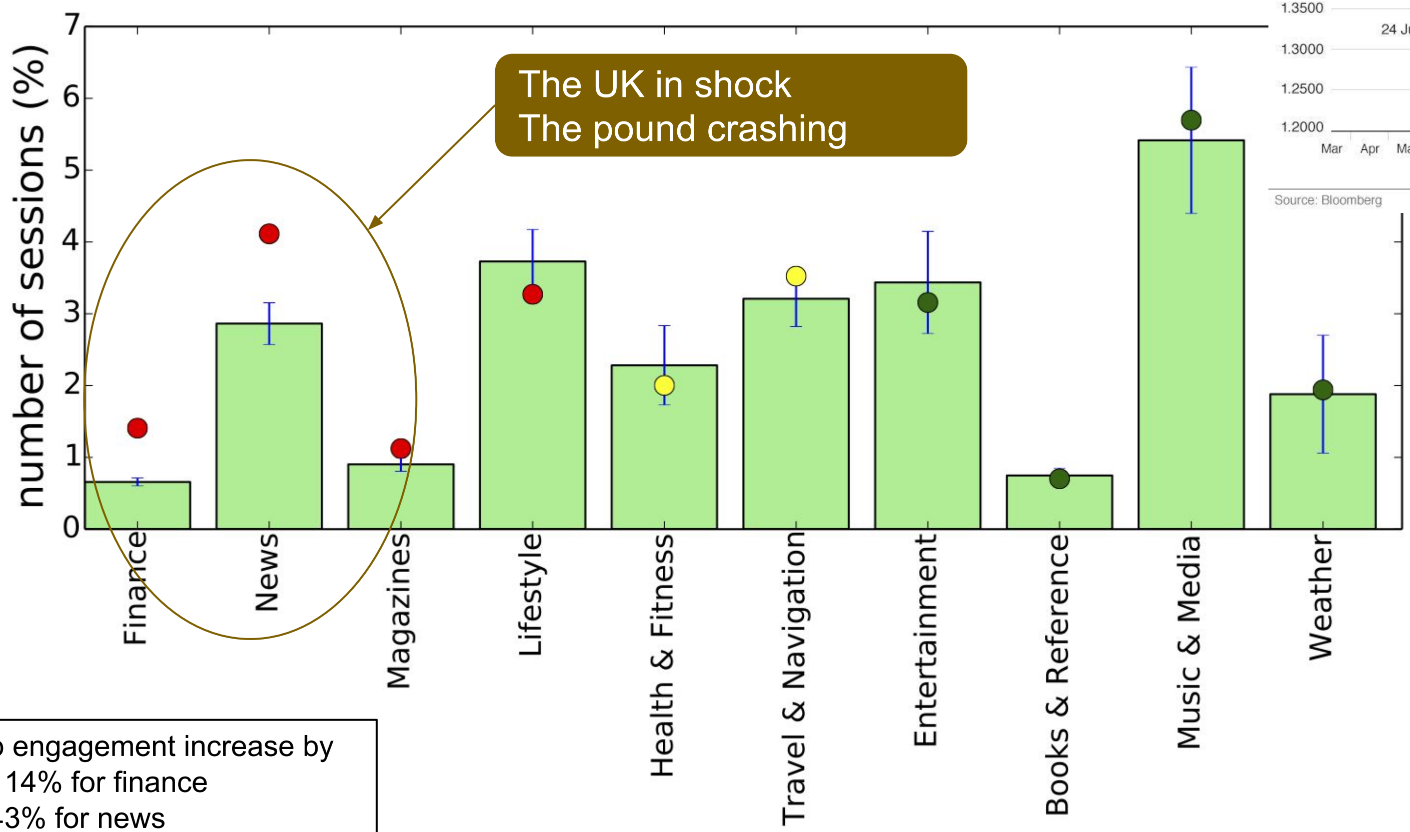
Periodicity (day) … When?



230K mobile apps, 600M daily unique users, 1 month sample (Van Canneyt etal, 2017)

External factors (news) … When?





The UK in shock
The pound crashing

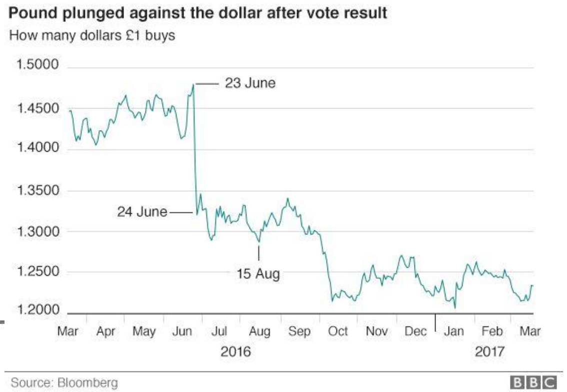
(Van Canneyt etal, 2017)

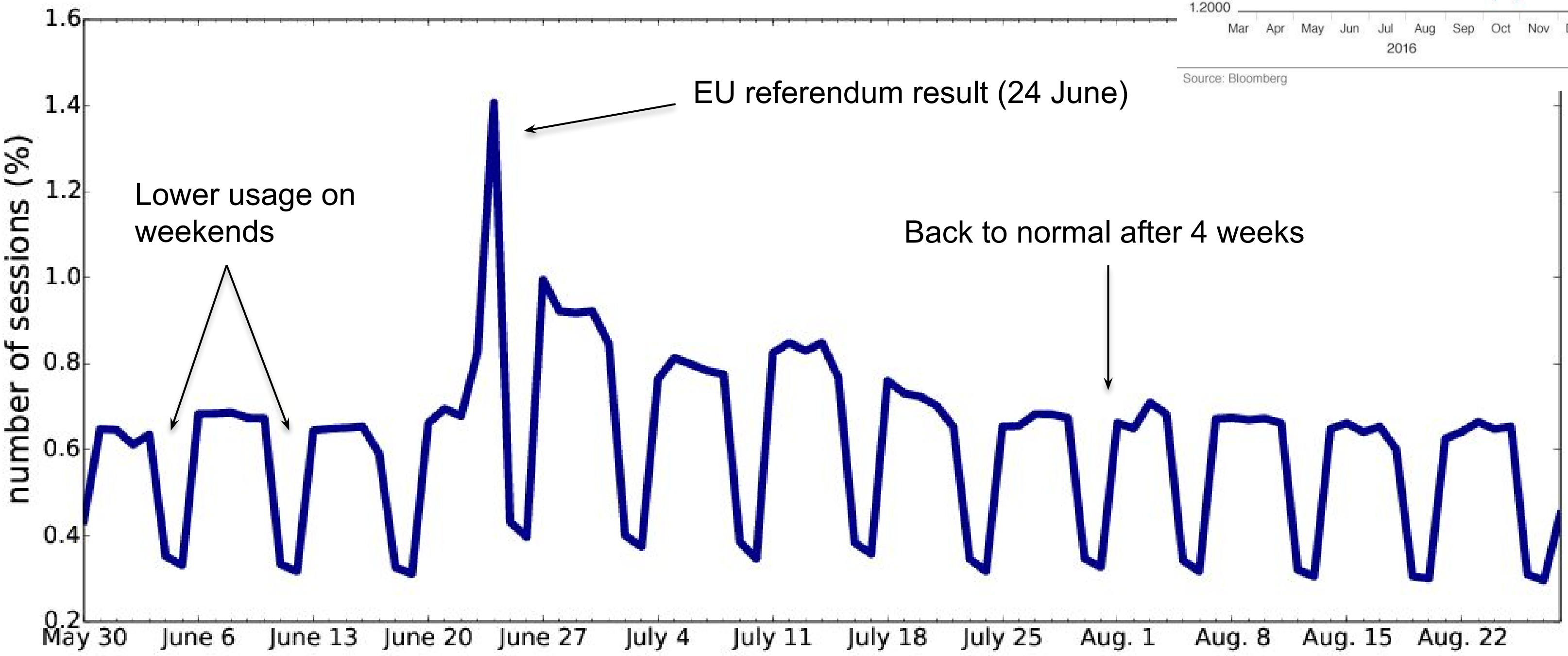
Day of the referendum result
24 June, 2016 (UK)

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

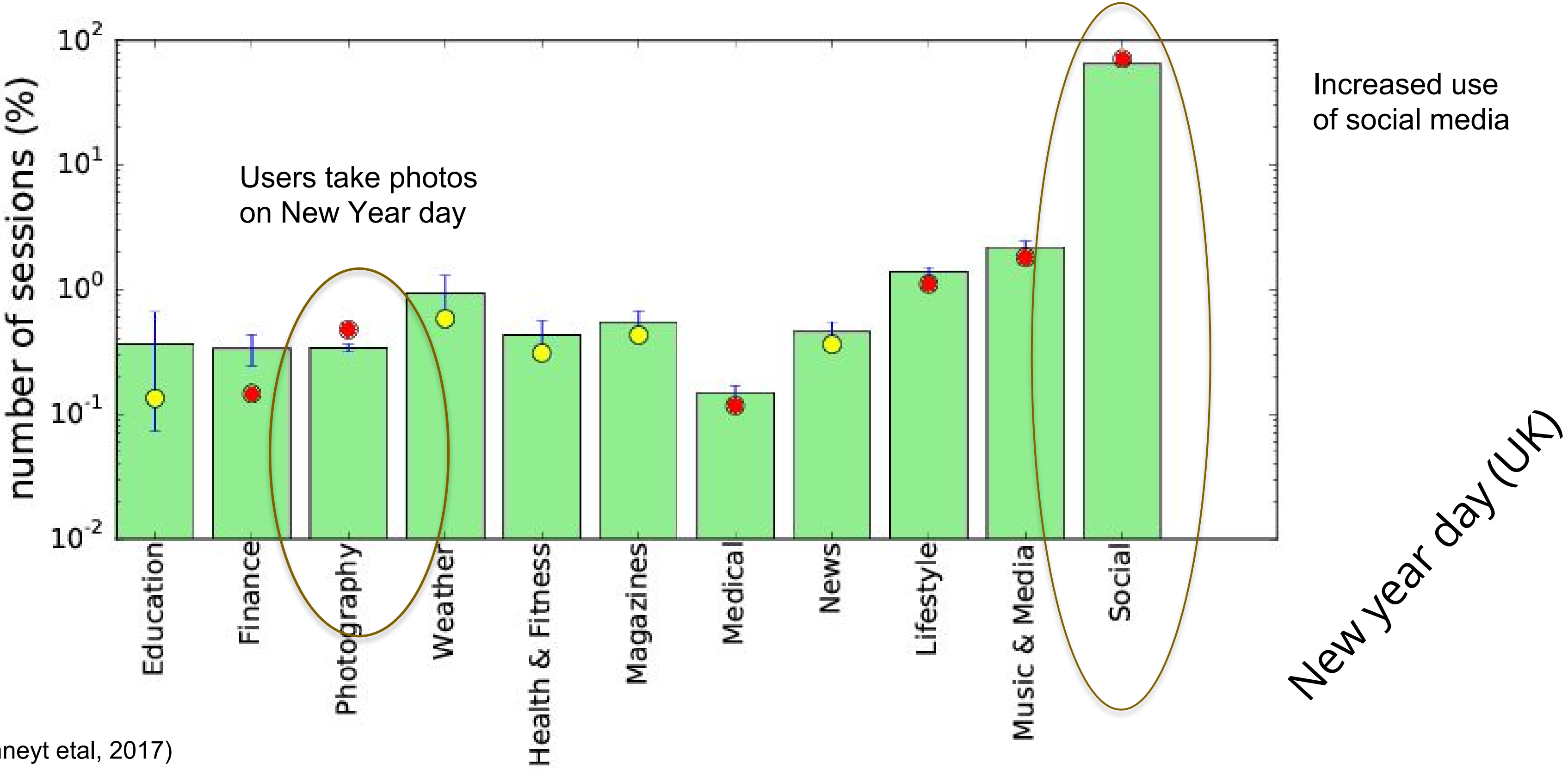
External factors (news) … When?

Finance apps (Van Canneyt etal, 2017)





External factors (social) … When?



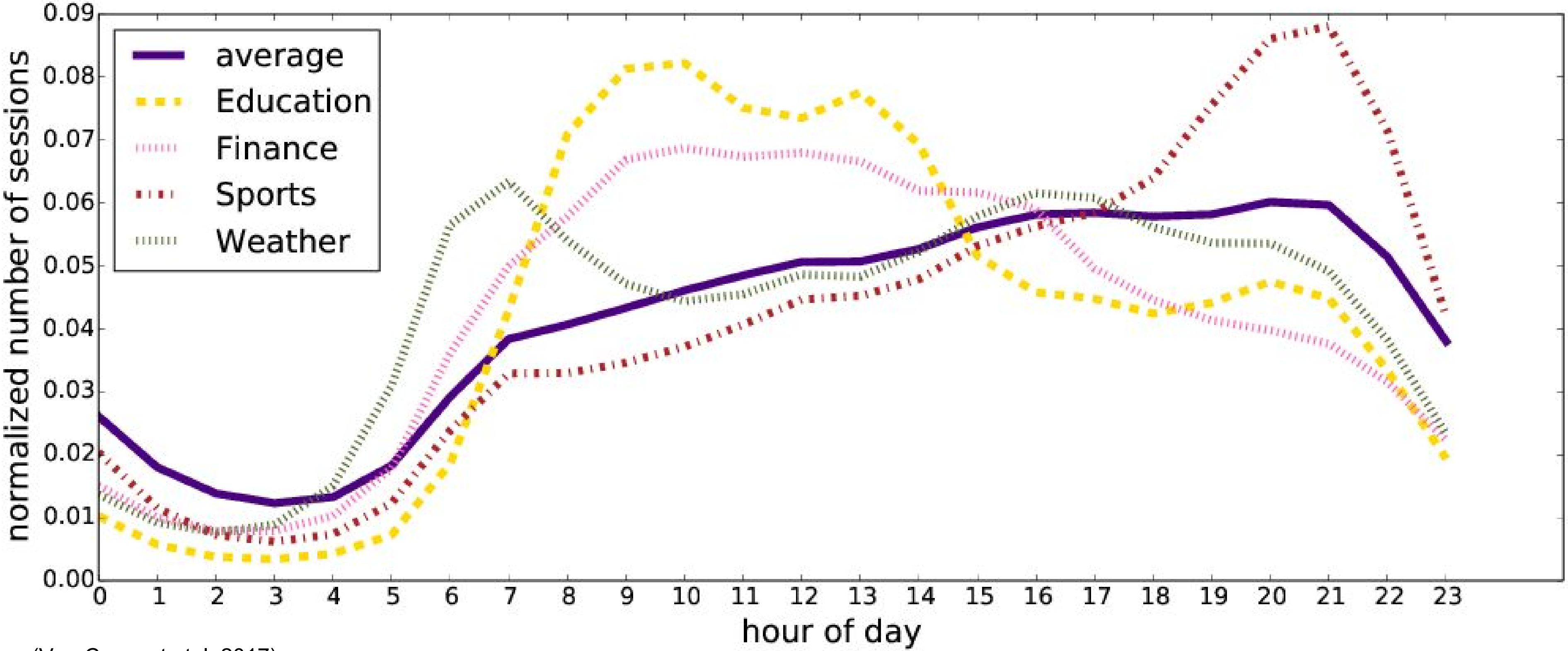
(Van Canneyt etal, 2017)

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?

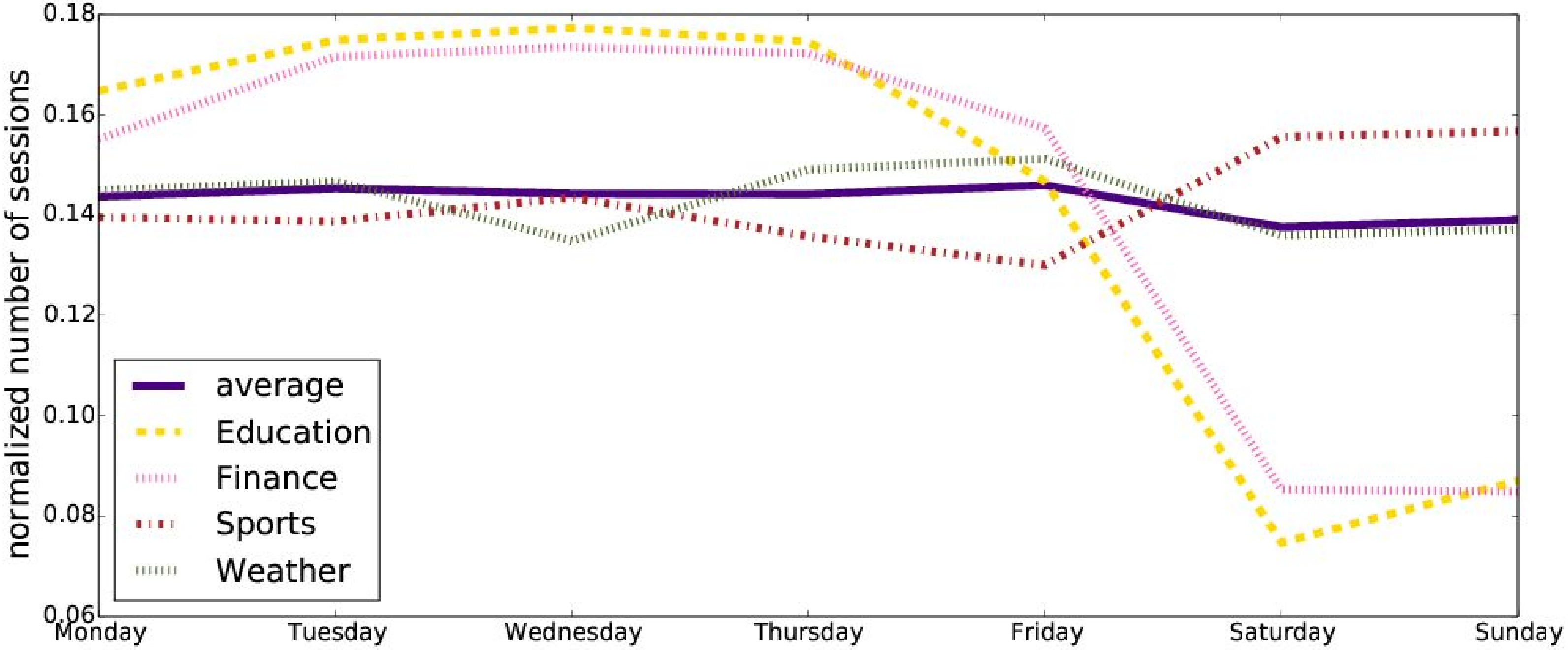


(Van Canneyt etal, 2017)

Task (week) … Why?

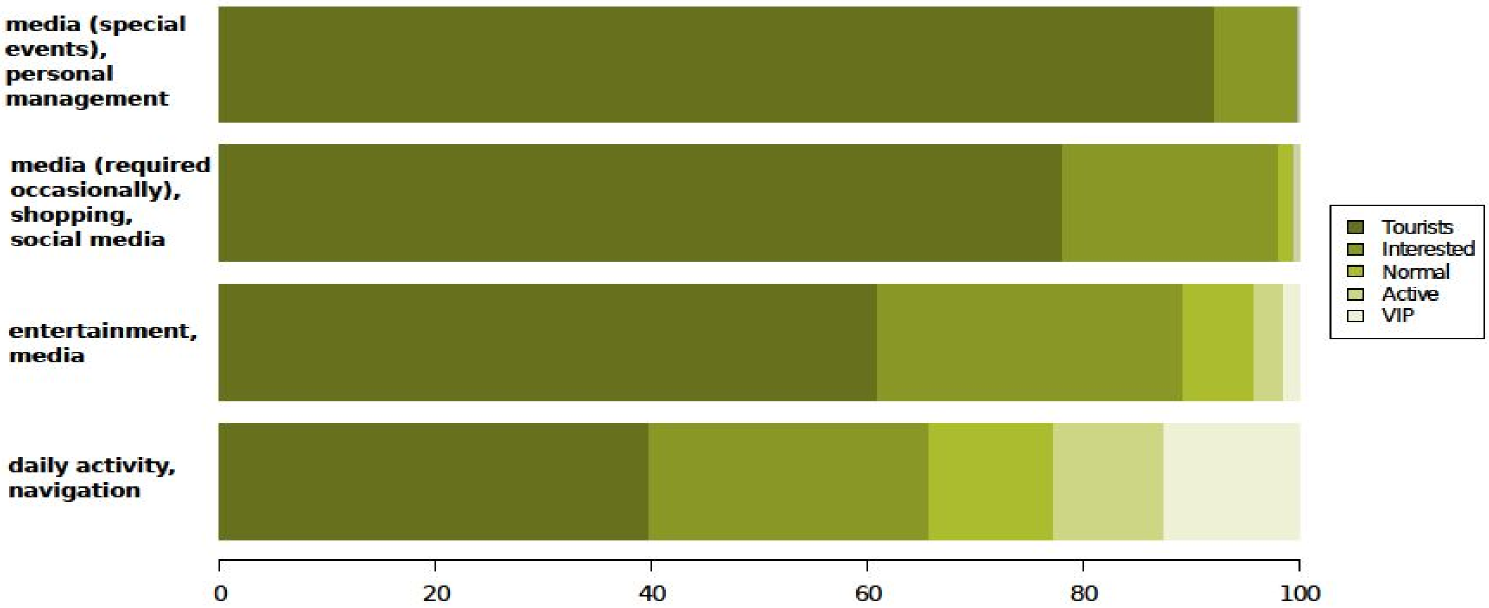
(Van Canneyt etal, 2017)

Week: productivity
Weekend: sports



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

(Lehmann etal, 2012)



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; endurability; 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)

(O’Brien, Lalmas & Yom-Tov, 2013)

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)

(O’Brien, Lalmas & Yom-Tov, 2013)

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)

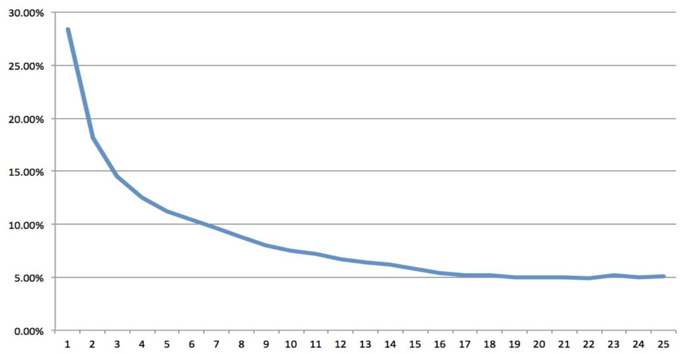
From day 1 → focus on new users

- 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

...

https://www.wordstream.com/blog/ws/2016/05/13/targeted-email-marketing-strategies

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)

(Peterson et al., 2008)

Time … From visit to session

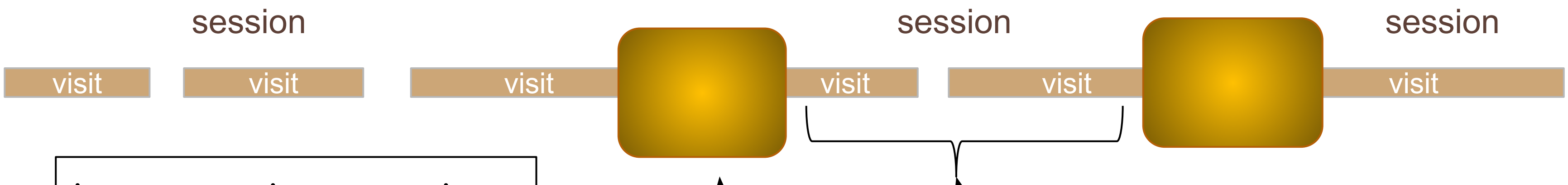
Dwell time is time spent on site (page) during a visit

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-session metrics

- page level or less
- visit level
- session level

- return soon
- remain engaged later on

inter-session metrics

long-term value (LTV) metrics

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 • Dwell time • Session duration • Page view (click depth) • Revisit rate • Bounce rate **Interaction** • Click through rate (CTR) • Number of shares & likes (social & digital media) • Conversion rate (e-commerce) • Streamed for more that x seconds **Contribution** • Number of replies • Number of blog posts • Number of uploads	**Granularity** Module ↓ Viewport ↓ Page ↓ Visit ↓ Session	**From one session to the next session (return soon)** • Time between sessions (absence time) **From one session to a next time period such next week, or in 2 weeks time (remain engaged later on)** • Number of active days • Number of sessions • Total usage time • Number of clicks • Number of shares • Number of thumb ups • ...

Intra- vs inter-sessions metrics … Granularity

Intra-session metrics

Module → Viewport → Page → Visit → Session

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

Inter-session metrics

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

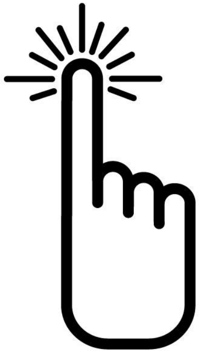
Examples of intra-session metrics

- 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



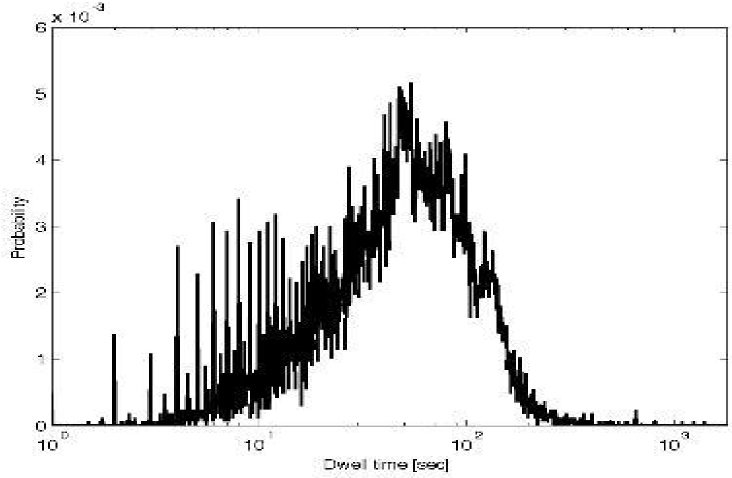
- Relate to abandonment rate
- Issues include clickbait, site design

Dwell time ... Involvement

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



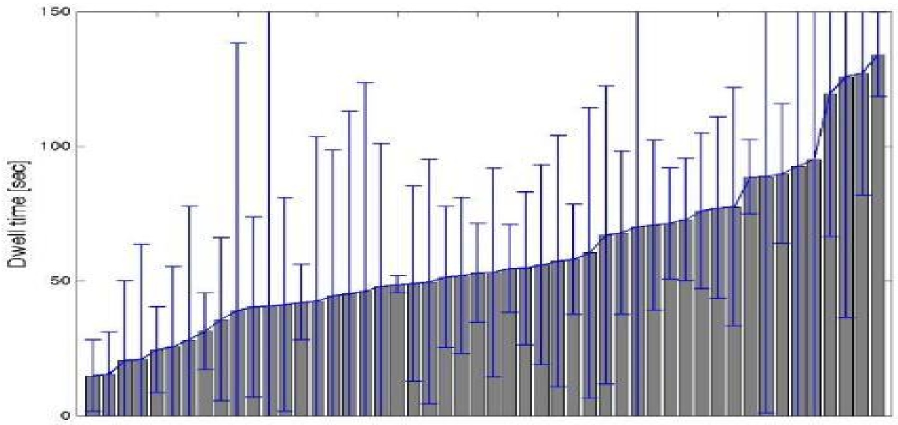
Distribution of dwell times on 50 websites

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

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

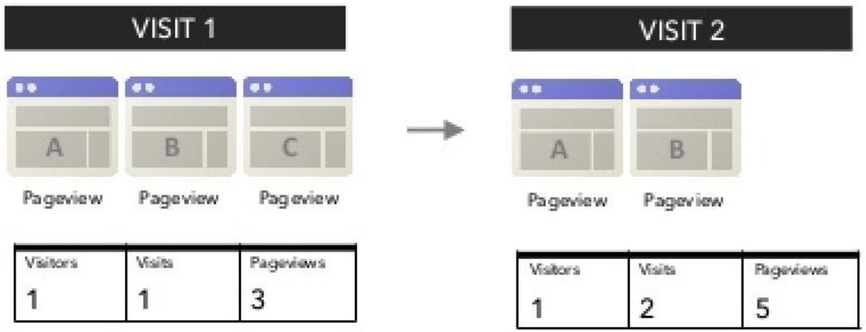
(O'Brien, Lalmas & Yom-Tov, 2013)

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



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

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

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.

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

Conversion rate ... Interaction

Fraction of sessions which end in a desired user action

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

Online advertising using conversion as cost model to charge advertisers

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

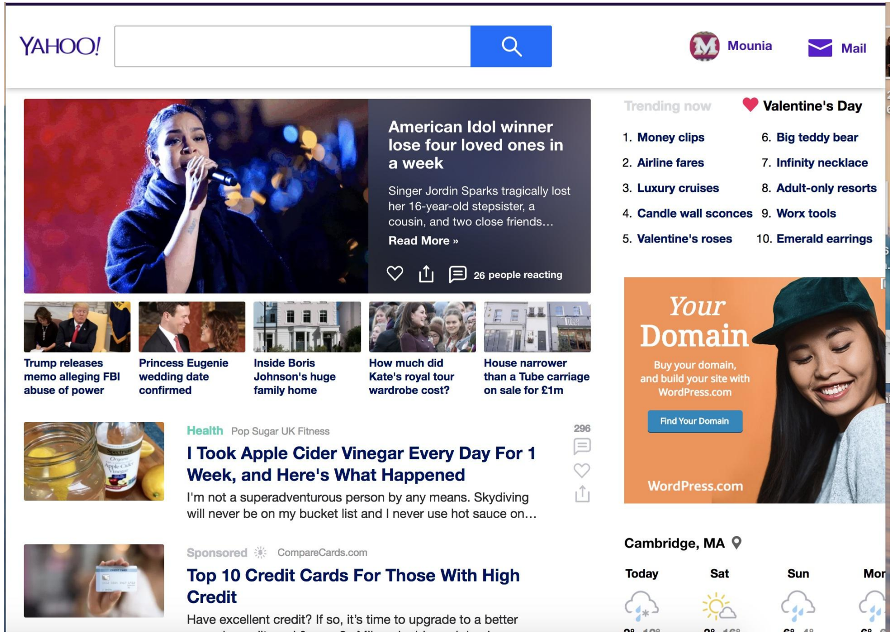
dwell time often used as proxy of satisfactory experience as may reflect affinity with the brand

Revisit rates … Involvement

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.

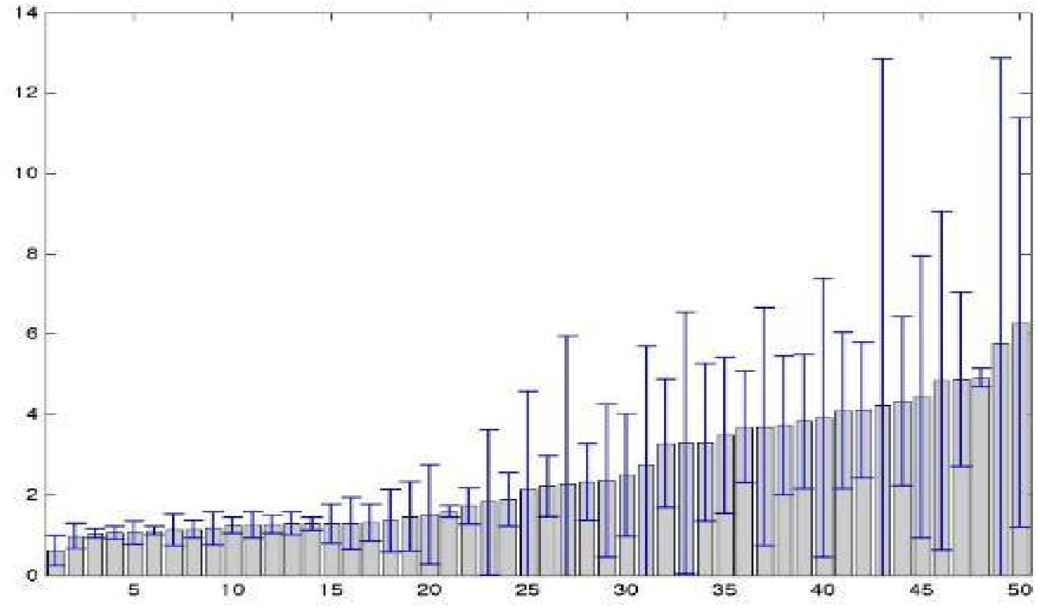
Useful for sites such as news aggregators, where returns indicate that user believes there may be more information to glean from the site



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

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



(O’Brien, Lalmas & Yom-Tov, 2013)

Revisit rate … Session length

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

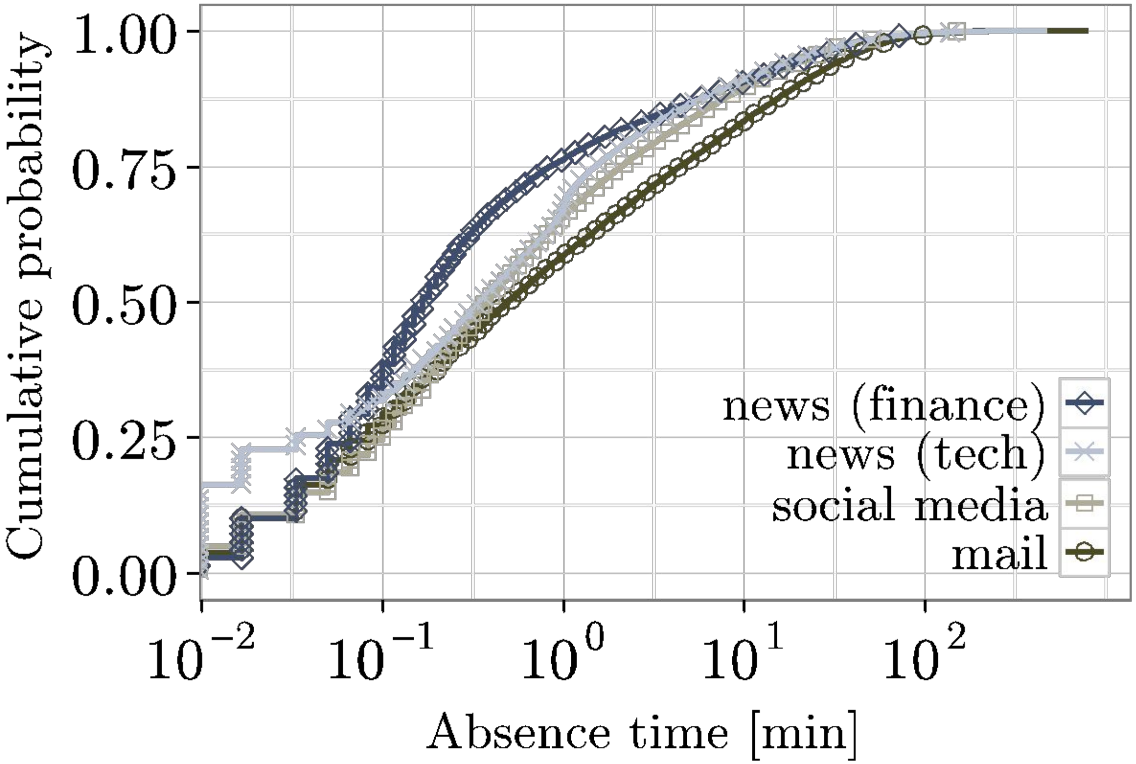
Categorization of the most frequent accessed sites

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

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

Time between each revisit



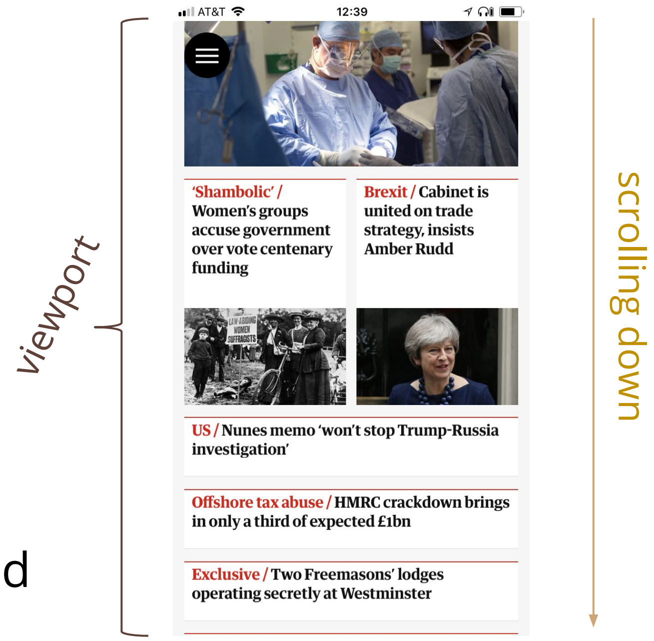
50% of sites are revisited after less than 1 minute

(Lehmann etal, 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

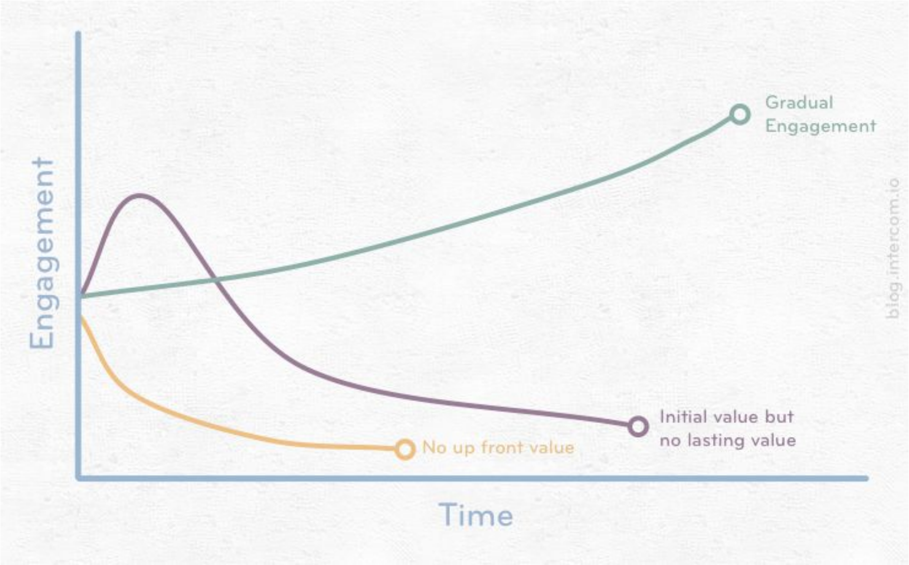
Why inter-session metrics?

(Kohavi et al, 2012)

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

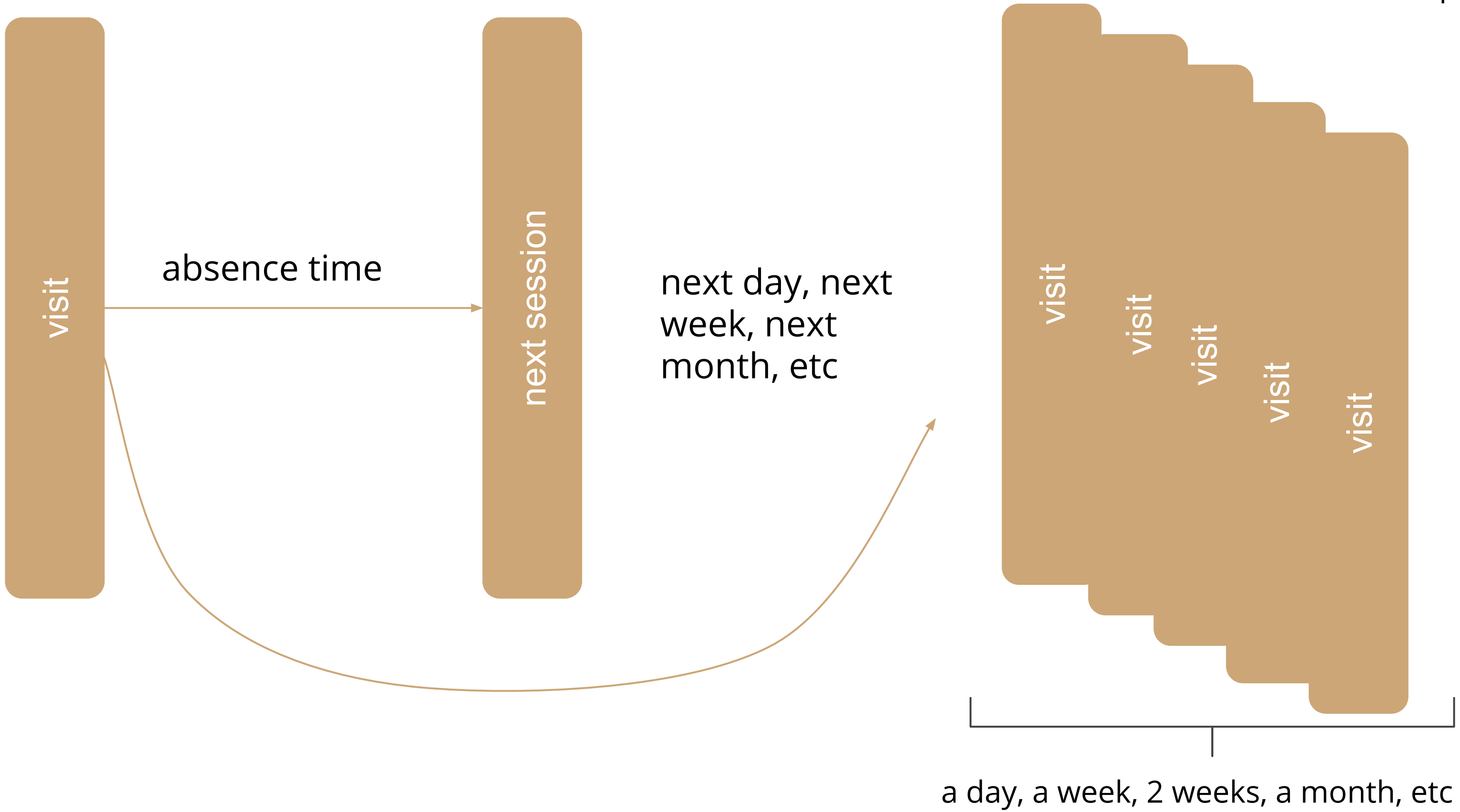
- Consider a very poor ranking function introduced into a search engine by mistake

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

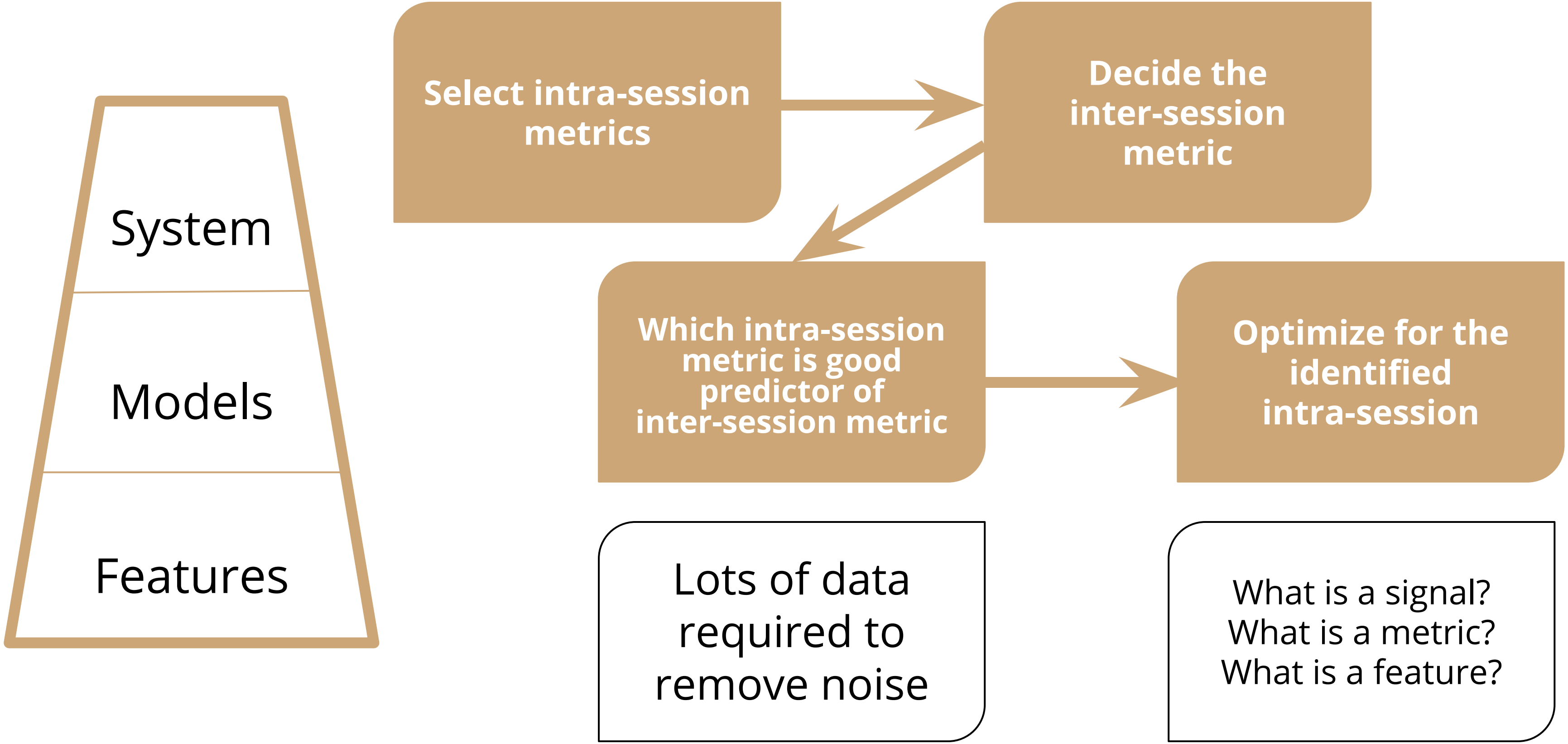


Inter-session metrics

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

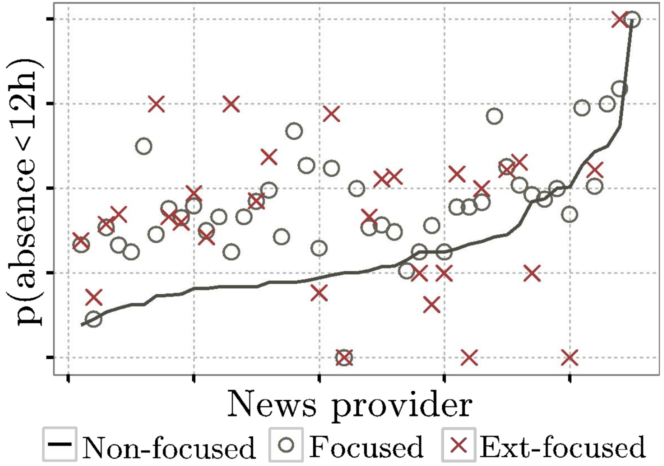


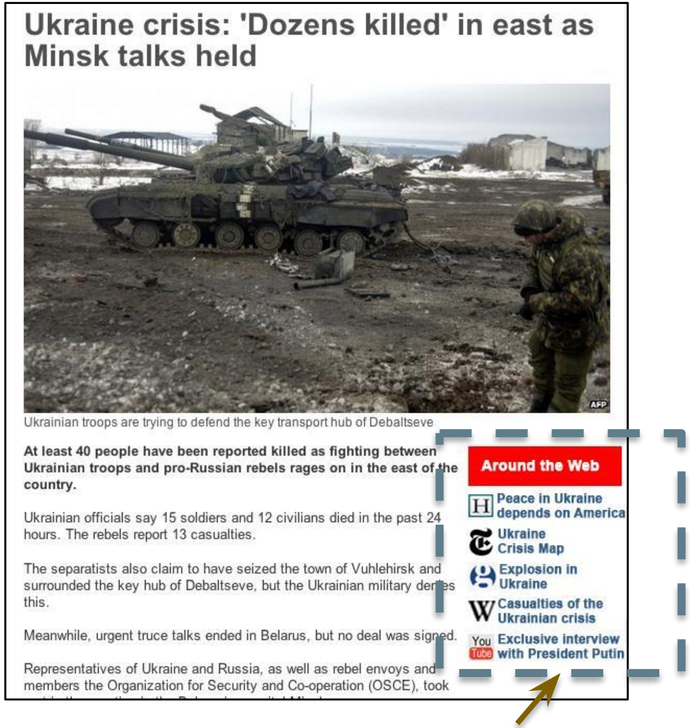
Intra- vs inter-sessions metrics … Optimization



Example I: Focused news reading

(Lehmann etal., 2016)



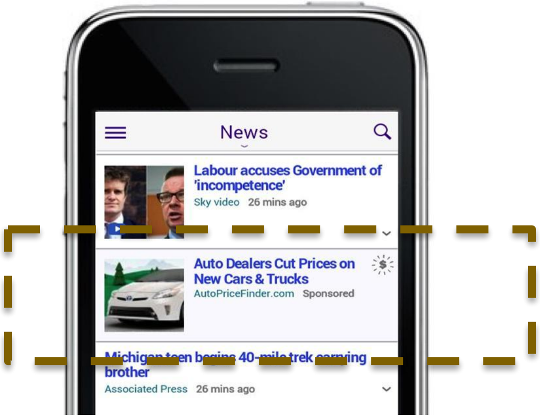


Related off-site content

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%)

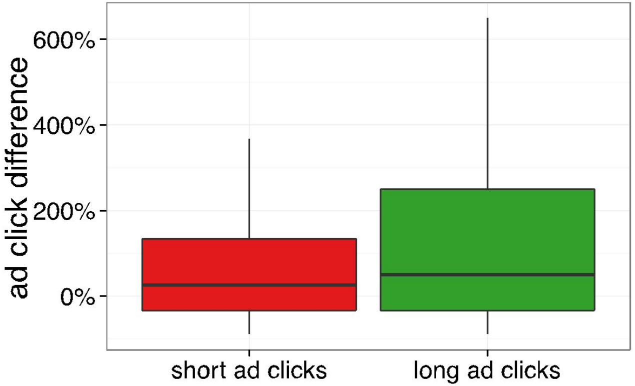
Example II: Ad landing page quality



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)



(Lalmas etal, 2015)

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

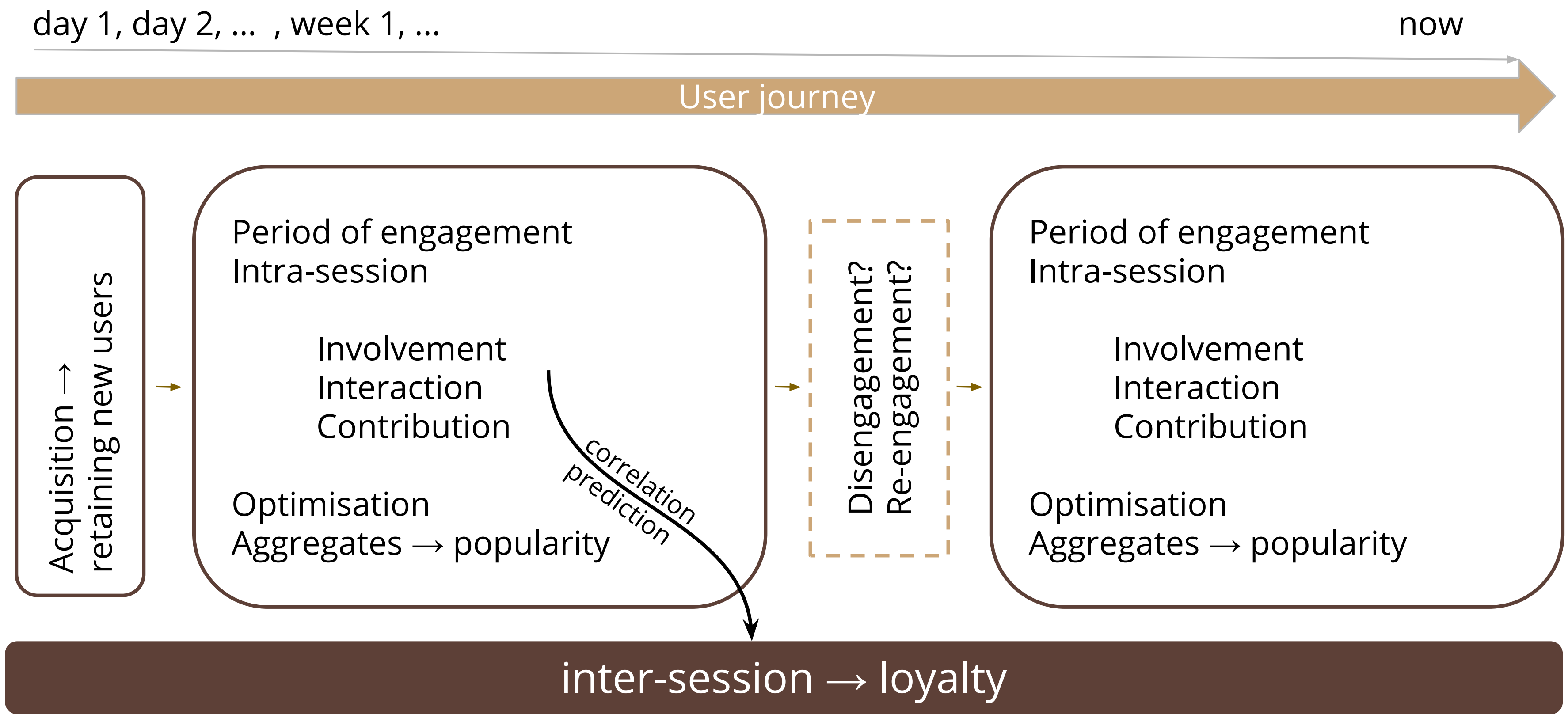
How valuable different users are based on lifetime performance → value that a user is expected to generate over a given period time, e.g. such as 12 months

- 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

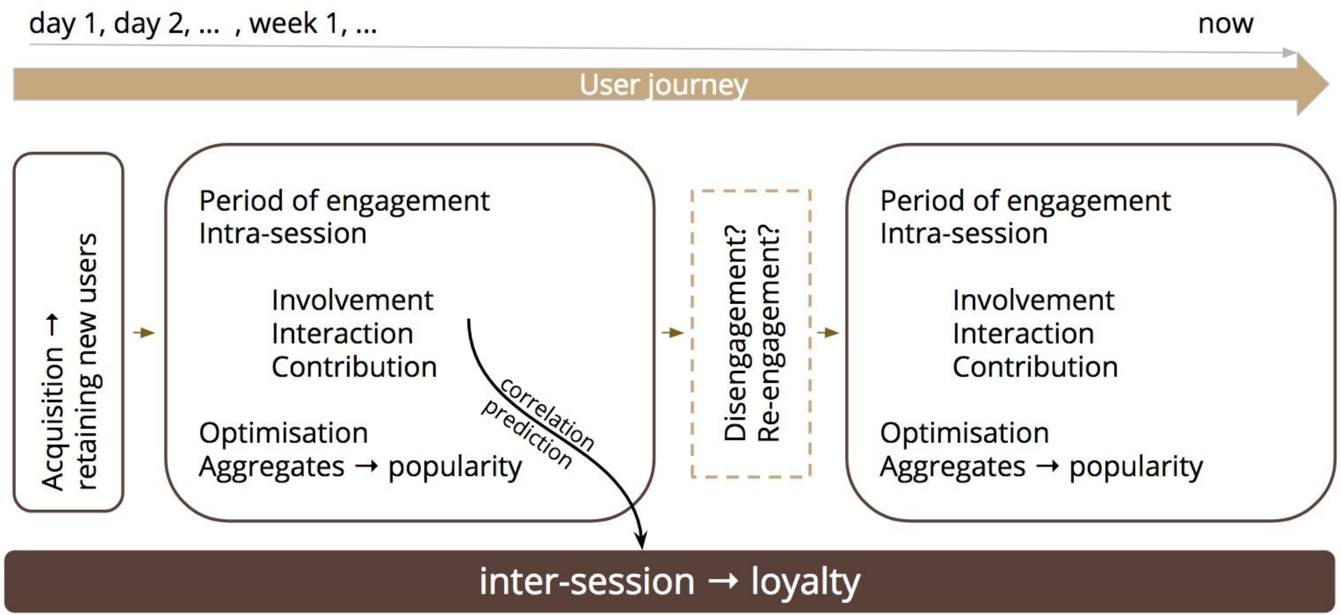
Help analyzing acquisition strategy (customer acquisition cost) and estimate further marketing costs

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

Taxonomy of metrics .... in two slides



Taxonomy of metrics .... in two slides



- 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

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 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 **<u>compared</u>**.

Why experiments

Main benefits of having experiments

- Metrics can be **measured**, **tracked** and <u>**compared**</u>.
- We can **learn**, **improve** and <u>**optimize**</u>.

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

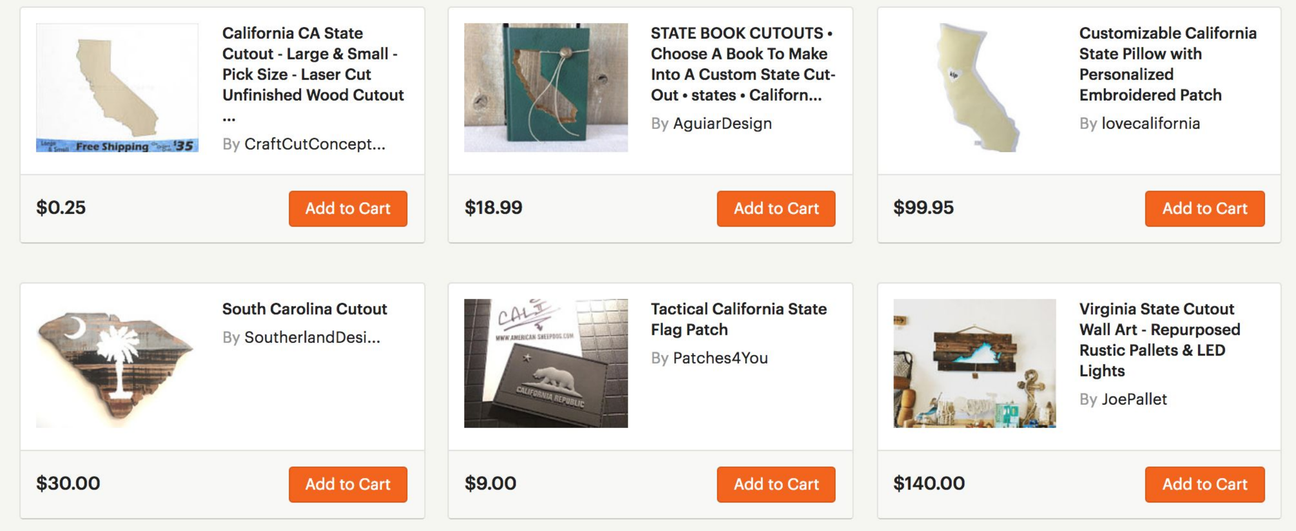
Experiments and Non-Experiments

Sometimes, experiments may not be feasible or practical.

Experiments and Non-Experiments

Sometimes, experiments may not be feasible or practical.

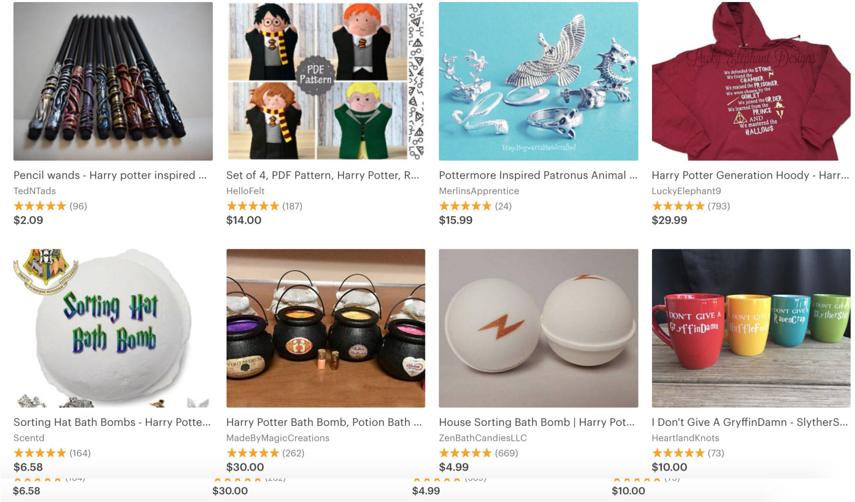
- **Example 1**:
  We want to test which "Add to Cart" button may lead to more Monthly-Active-Users (MAUs).

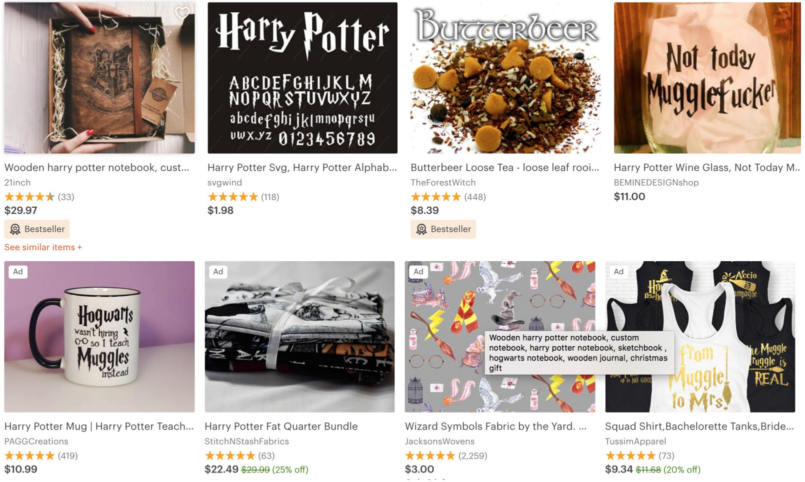


Experiments and Non-Experiments

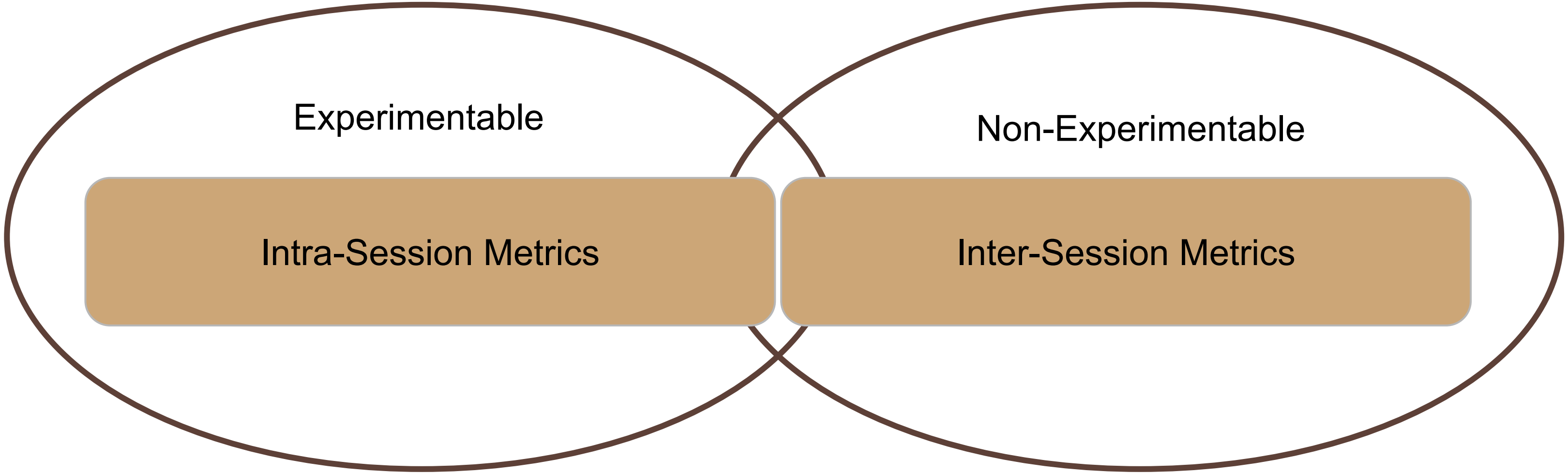
Sometimes, experiments may not be feasible or practical.

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





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.

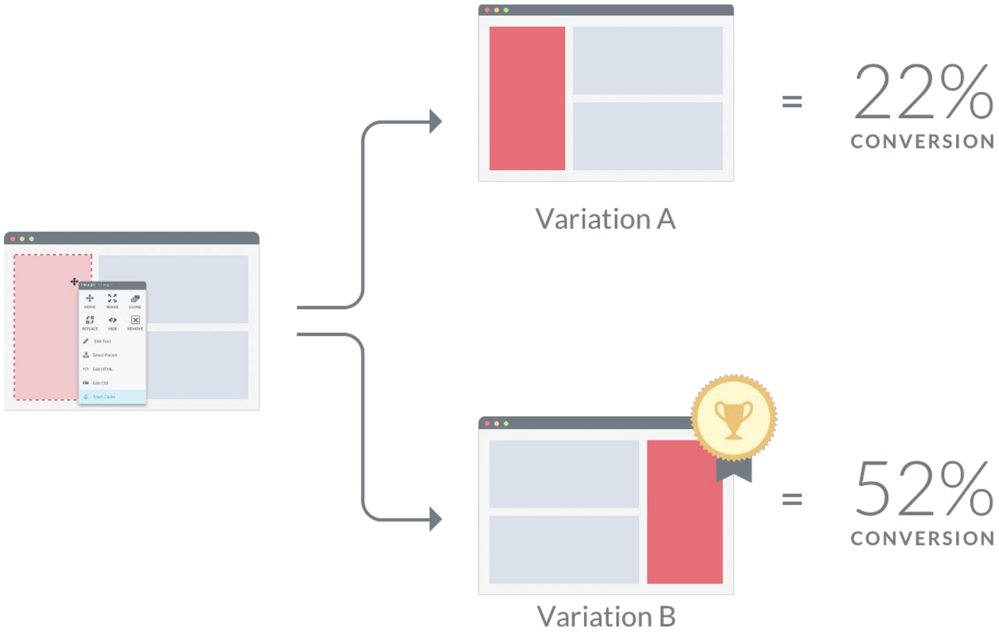
Types of experiments

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

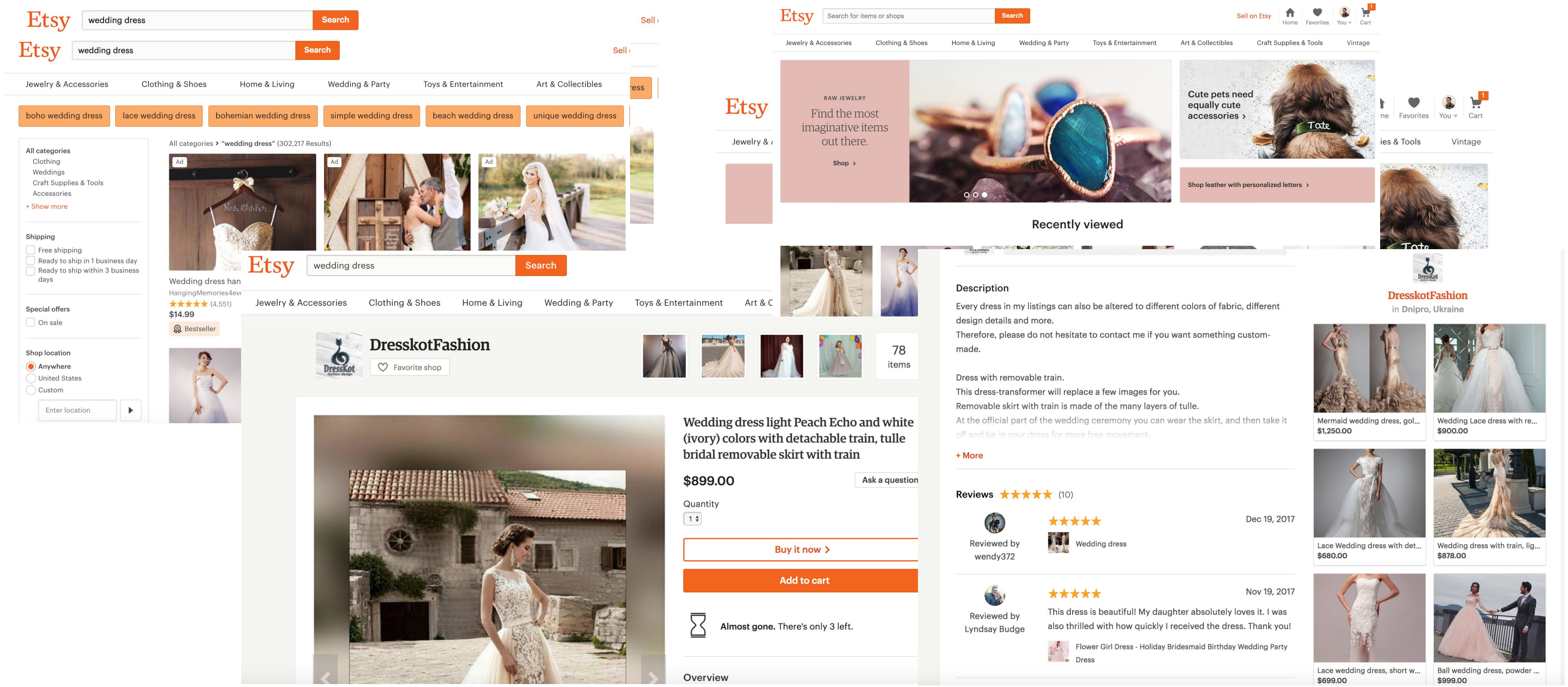
Online experiment

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

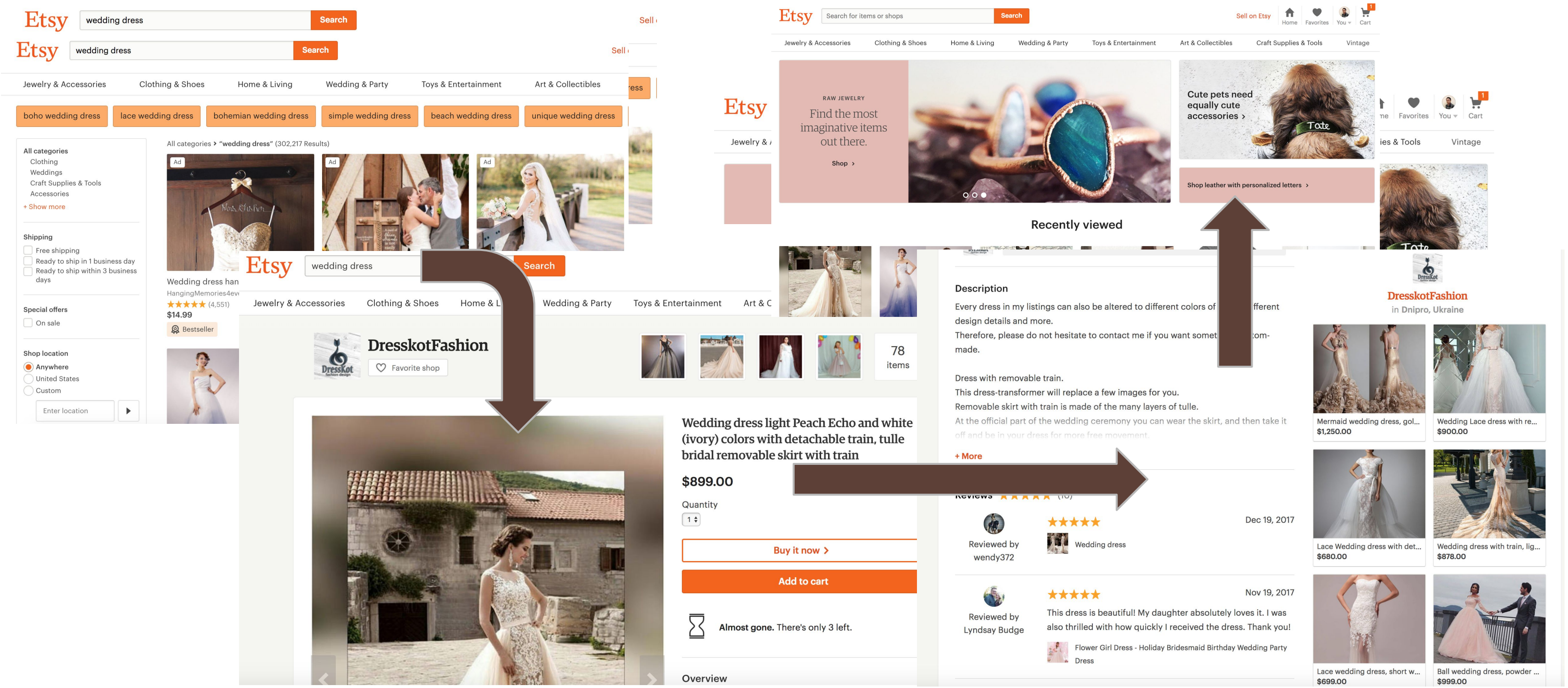
Online experiment

A/B Tests or Bucket Tests or Online Controlled Experiments



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.

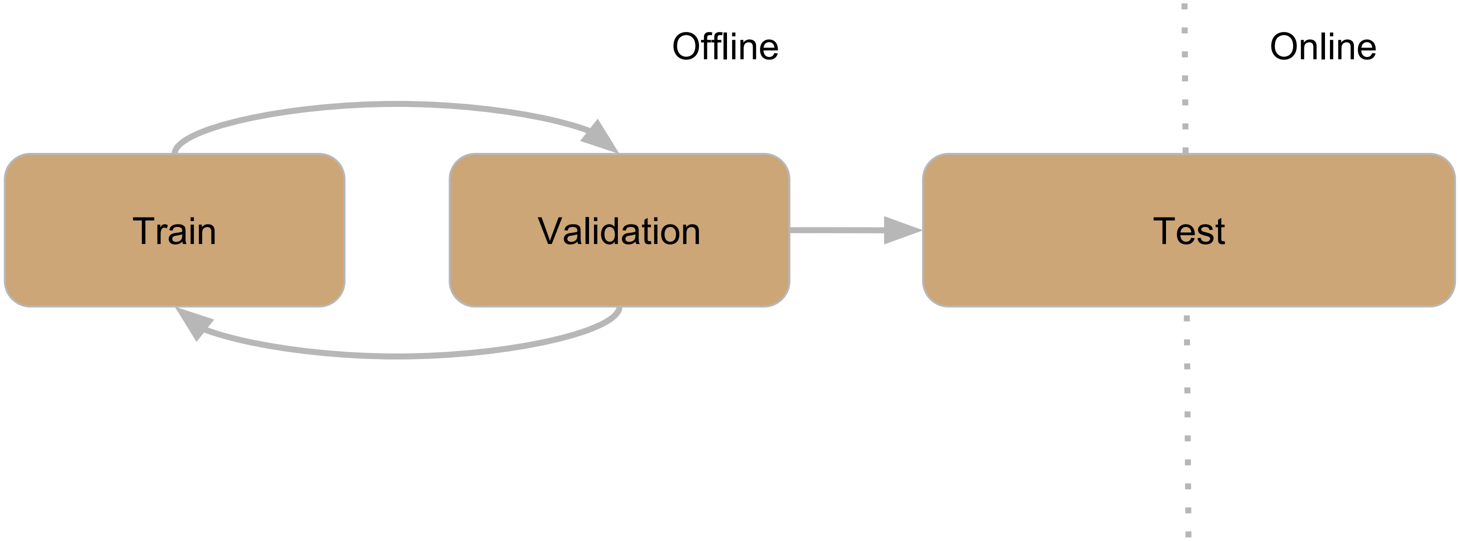
Offline experiment

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

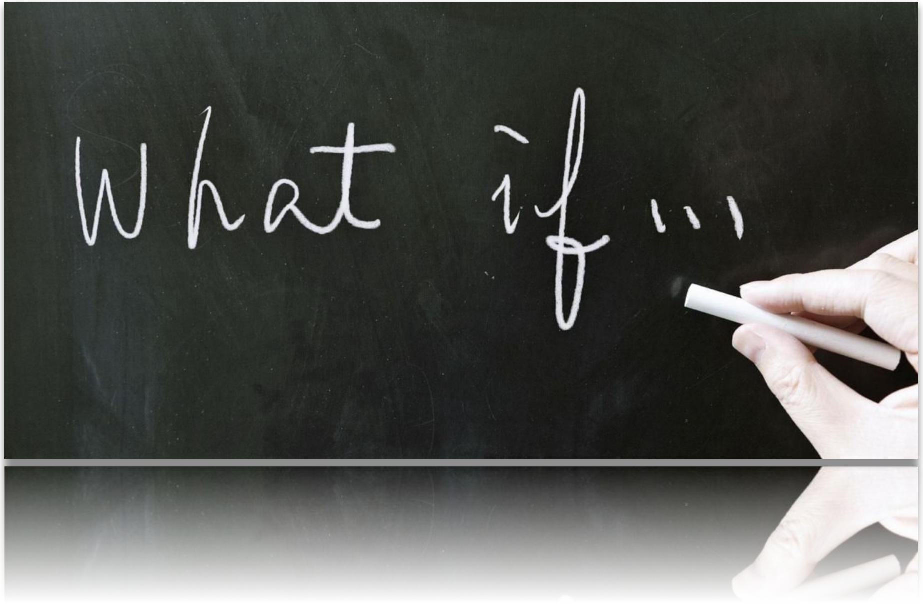
- Indirect and can be reused
- Good machine learning theories
- Correlation
- Static

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

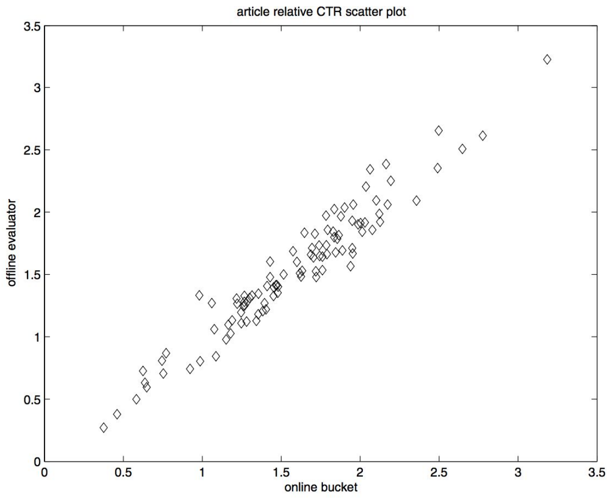


Figure 2: Articles' CTRs in the online bucket versus offline estimates.

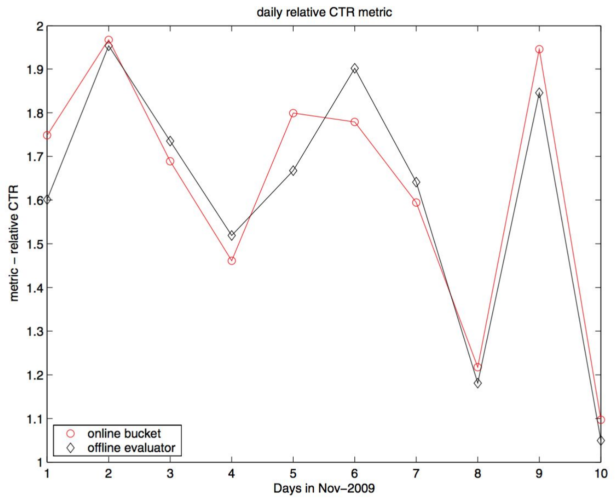


Figure 3: Daily overall CTRs in the online bucket versus offline estimates.

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

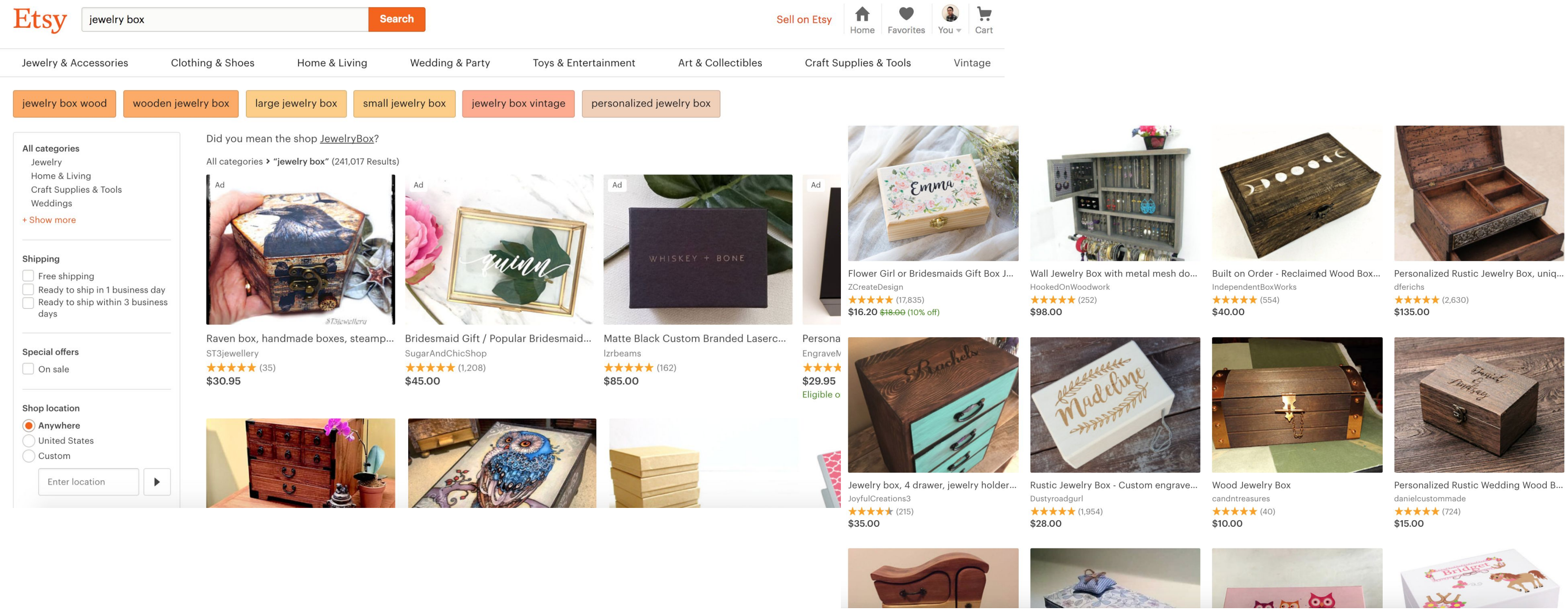
Counterfactual Offline Reasoning/Experiment

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

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

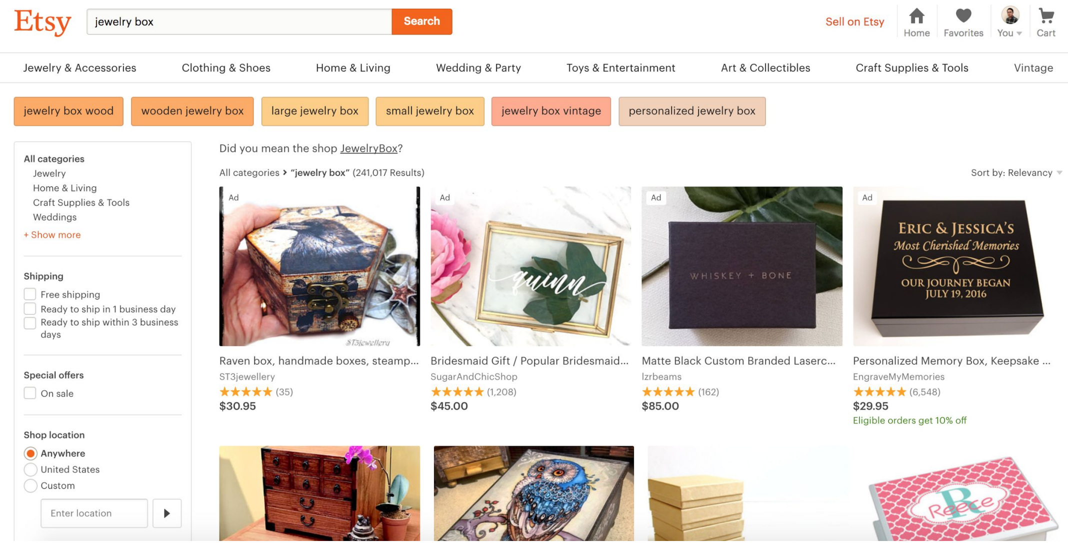


Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

- Generalization to Non-uniform Logging/Exploration

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



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 $t = 1, \cdots, T$ **do**
  for $i = 1, \cdots, K$ **do**
    Draw θ_i according to Beta$(S_i + \alpha, F_i + \beta)$.
  end for
  Compute p such that $p_k = \frac{\theta_k}{\sum \theta_k}$.
  Sample N **items from Mult.**$(\mathbf{p})$.
  Observe N rewards $\mathbf{r}_t$.
  Update S and F for those N items according to $\mathbf{r}_t$.
  Logging N items, $\mathbf{p}$ and $\mathbf{r}_t$.
end for

Offline A/B Experiment

Counterfactual Offline Reasoning/Experiment

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

Algorithm 3 Thompson Sampling for Bernoulli Ranked-list Bandit

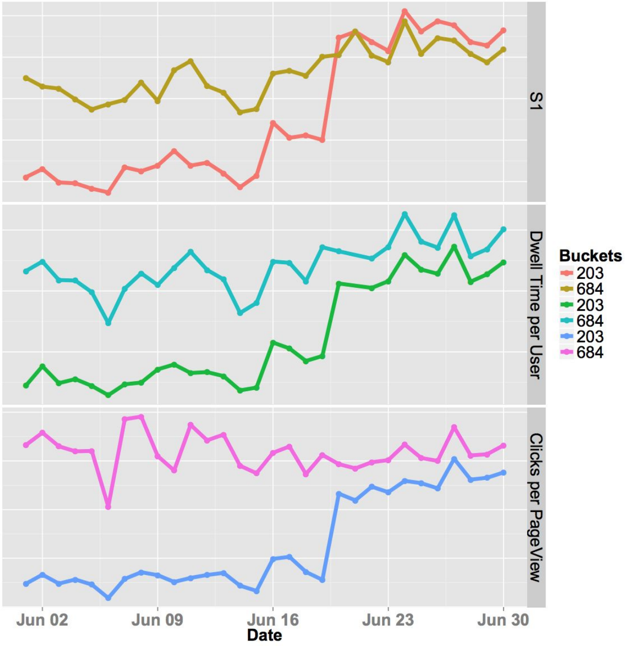
---

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

our focus in this section

Behavioral metrics -- online metrics, analytics

Optimisation metrics -- metrics used to train machine learning algorithms

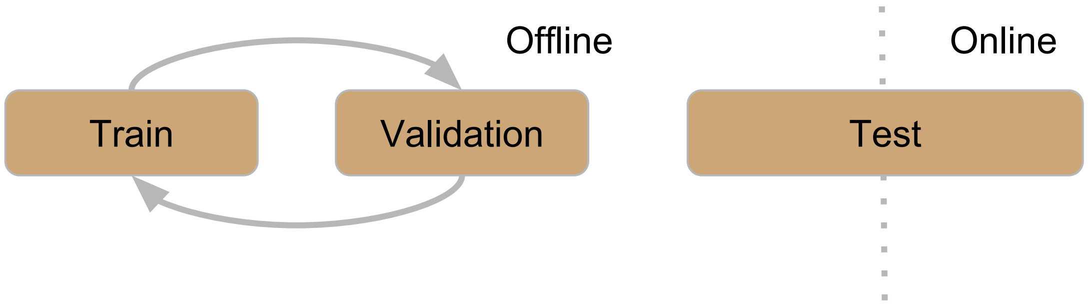
These three levels are connected

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)

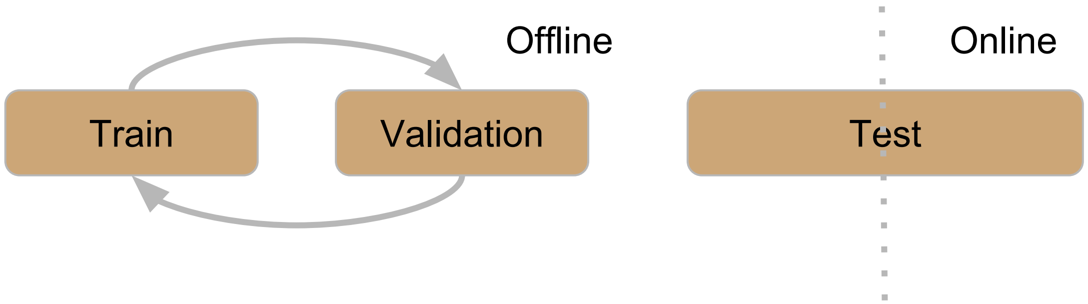


Offline Optimization

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

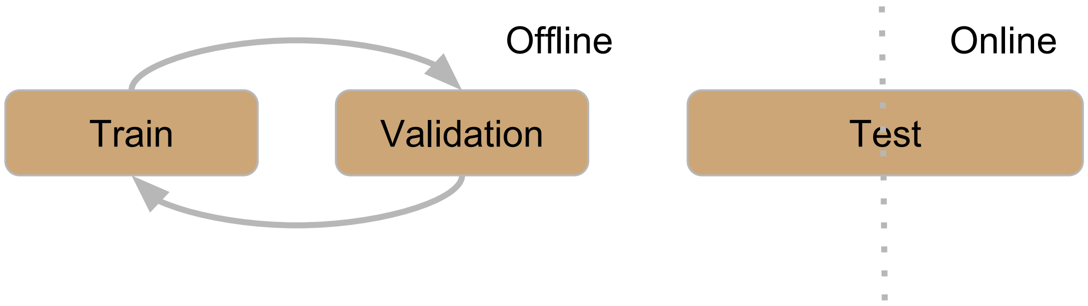
Offline Optimization

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



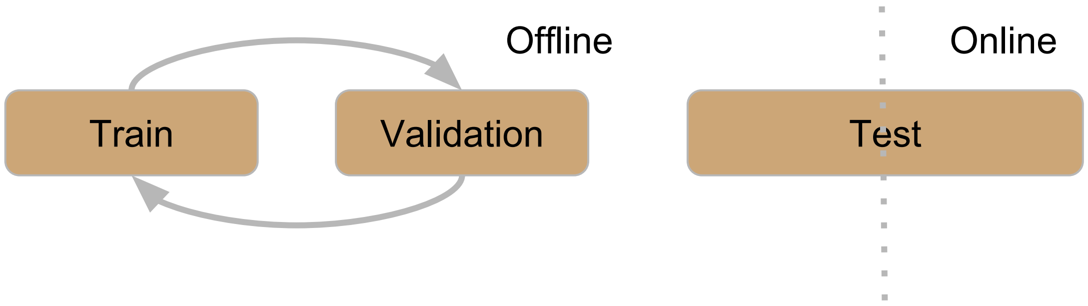
Offline Optimization

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



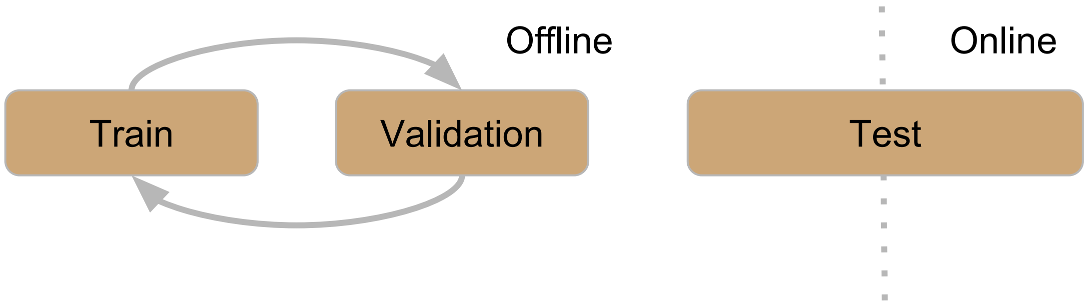
Offline Optimization

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



Offline Optimization

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



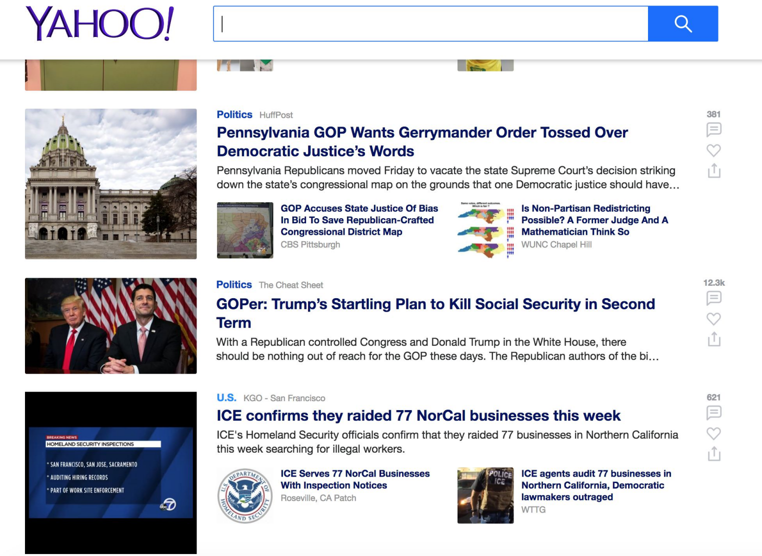
Online Optimization

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

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 **<u>Generalized Linear Model (Bernoulli)</u>** to model how likely a user *u* would click on an item a_i at iteration *i*.
2. Use **<u>Moving Average</u>** 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^2Bandit

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

Online Optimization

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

Algorithm 1 r^2Bandit

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

Online Optimization

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

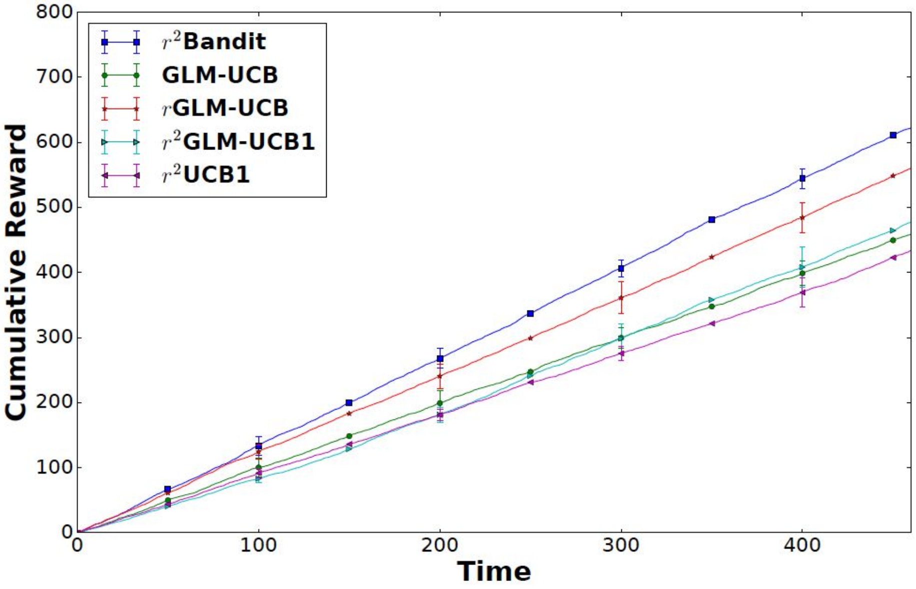
Simulations

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

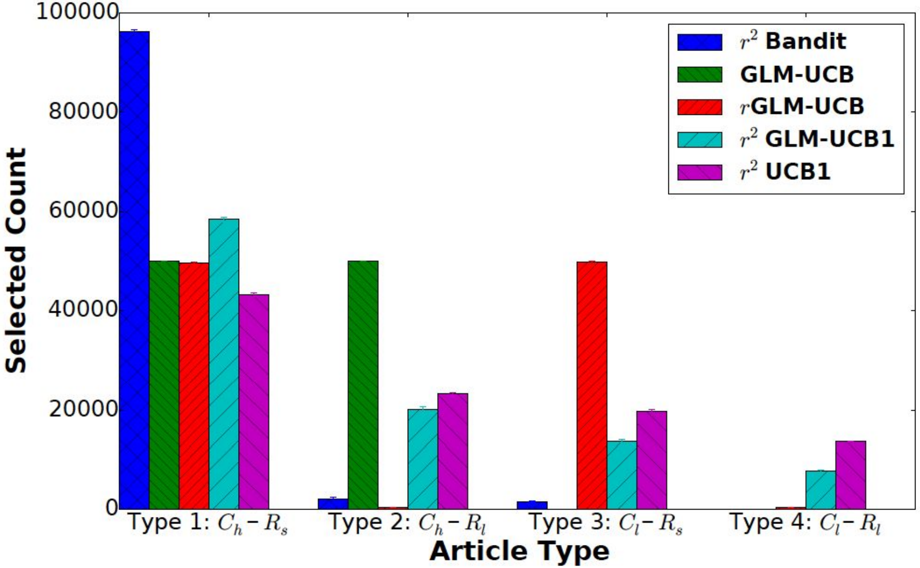


(a) Cumulative clicks over time

Online Optimization

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

Simulations

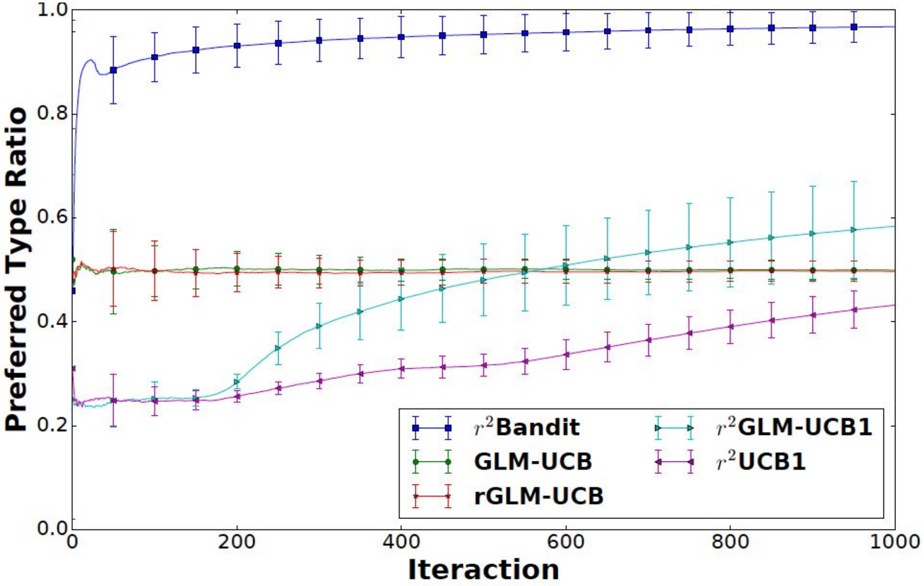


(b) Distribution of selected item types

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

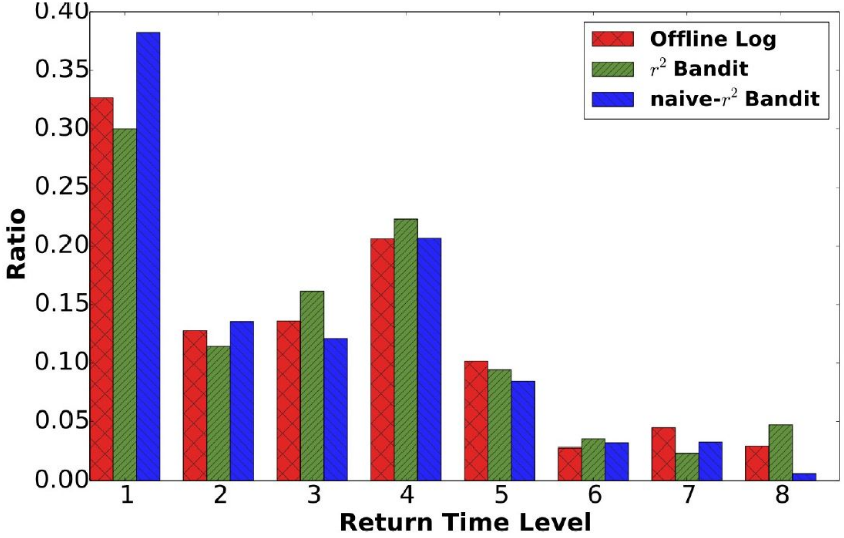


Figure 2: Discretized user return time distribution.

Online Optimization

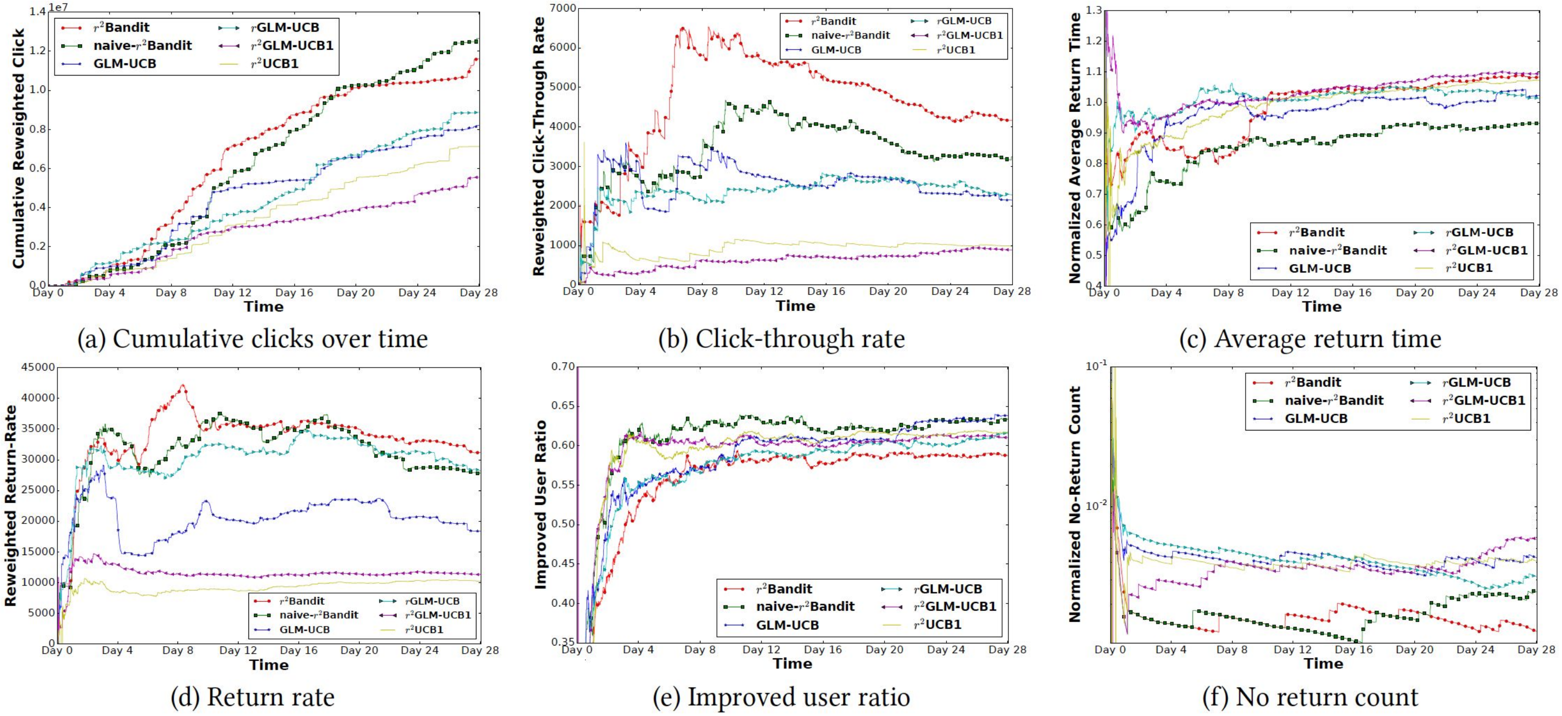
How to Online Optimize User Intra-Session Metrics and Inter-Session Metrics

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



(a) Cumulative clicks over time
(b) Click-through rate
(c) Average return time
(d) Return rate
(e) Improved user ratio
(f) No return count

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

Online Optimization

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

Real-World Dataset: Word Cloud

(a) Top clicked articles

(b) Top returning articles

Figure 4: Word cloud of algorithm selected article content.

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

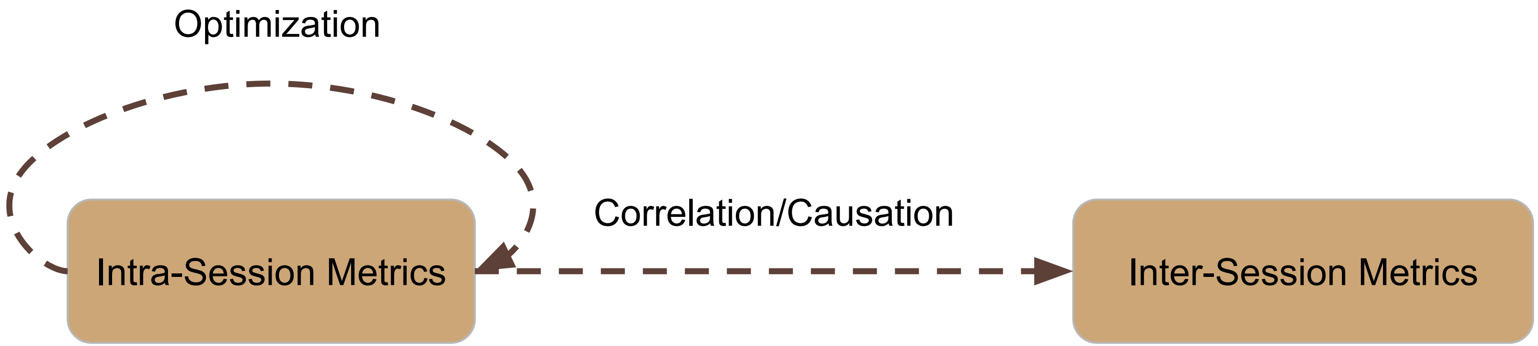
If inter-session metrics can be **<u>explicitly modeled</u>** or write them down in their **<u>clear form</u>**, you can use online optimization tools to **<u>directly optimize</u>** them.

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

…

Optimization Inter-Session Metrics

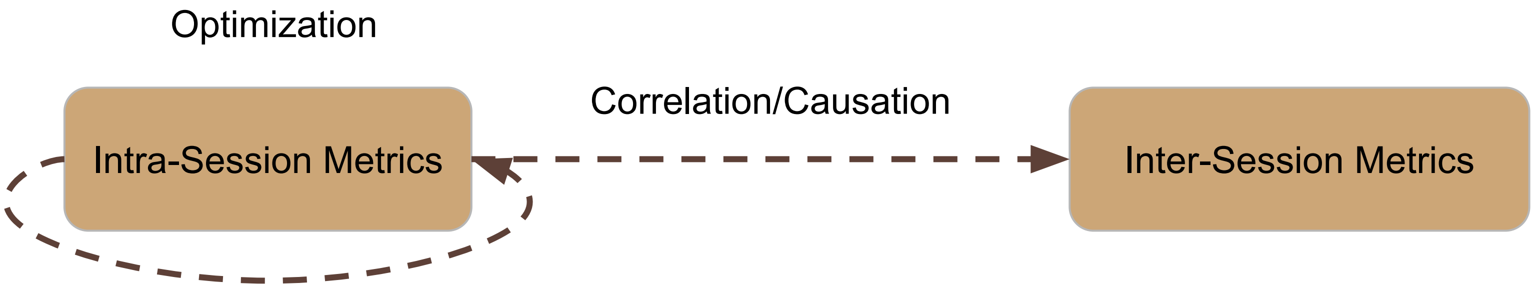
Approach II



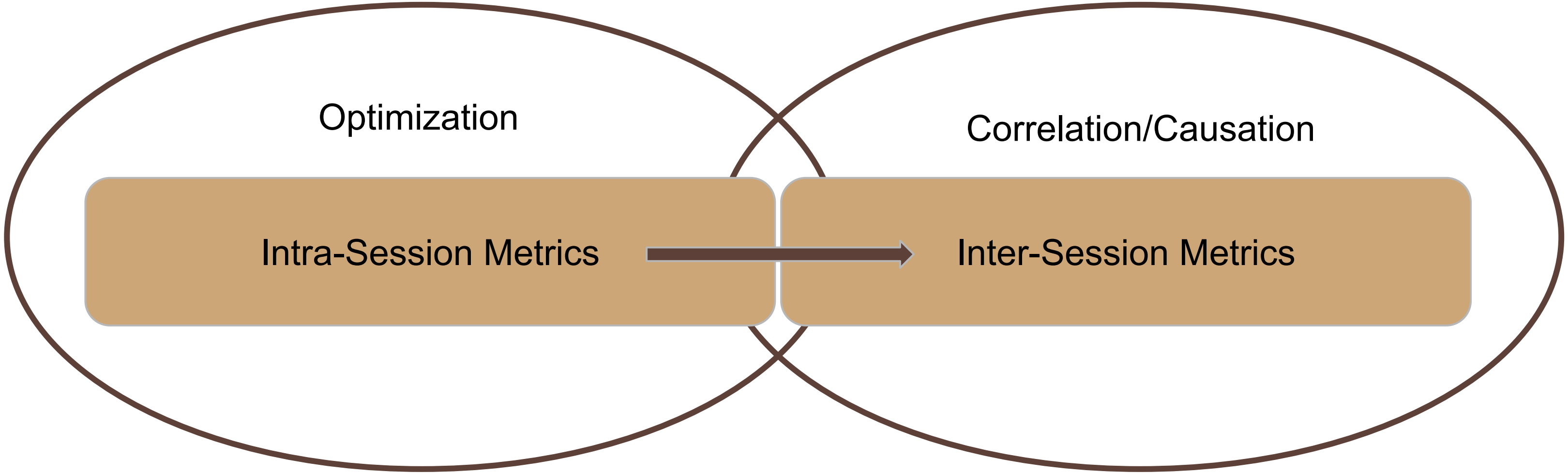
Optimization Inter-Session Metrics

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

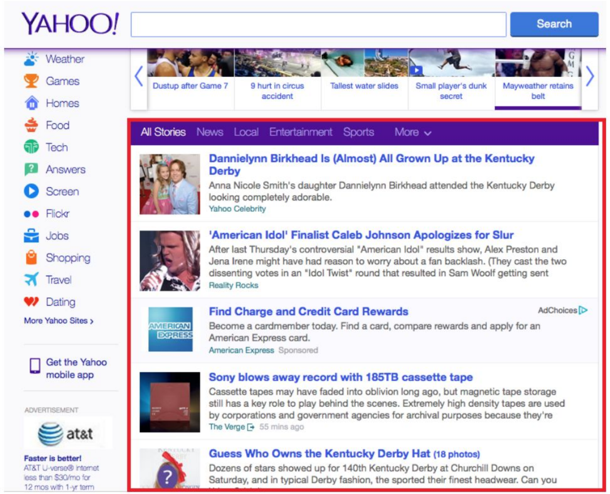


Figure 1: A snapshot of Yahoo's homepage in U.S. where the content stream is highlighted in red.

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

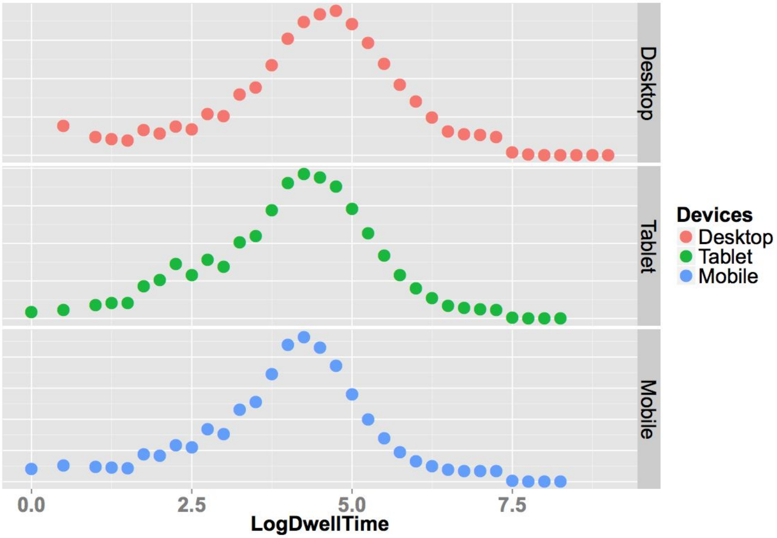


Figure 2: The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

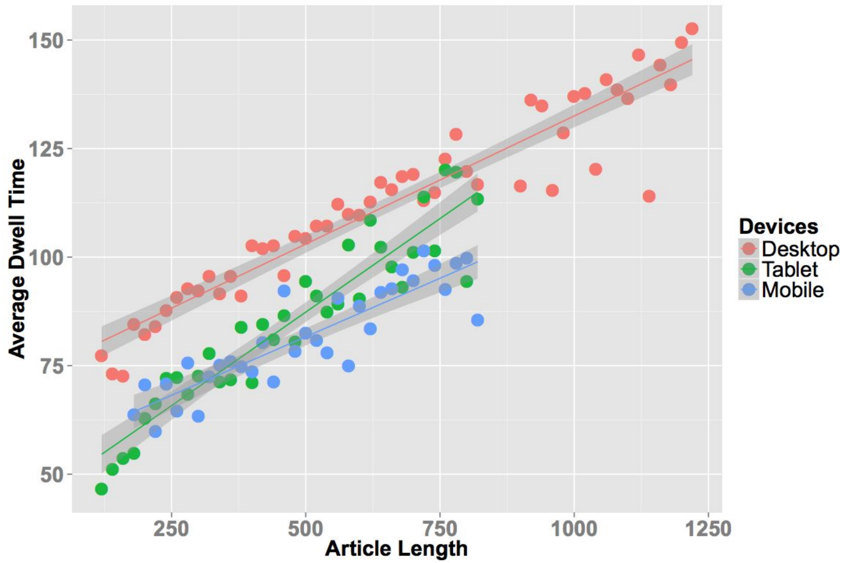


Figure 3: The relationship between the average dwell time and the article length where X-axis is the binned article length and the Y-axis is binned average dwell time.

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

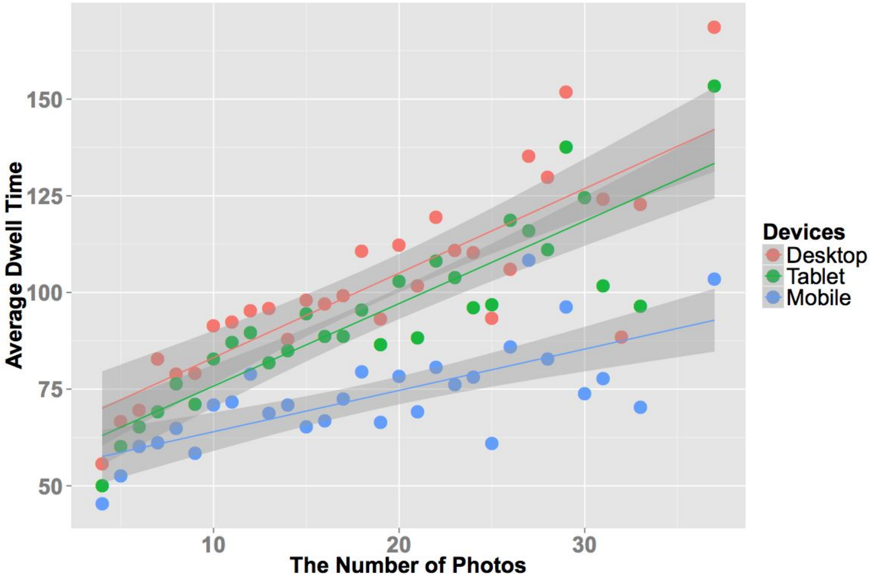


Figure 4: The relationship between the average dwell time and the number of photos on a slideshow where X-axis is the binned number of photos and the Y-axis is binned average dwell time.

Optimization Inter-Session Metrics

Beyond Clicks: Dwell Time in Personalization

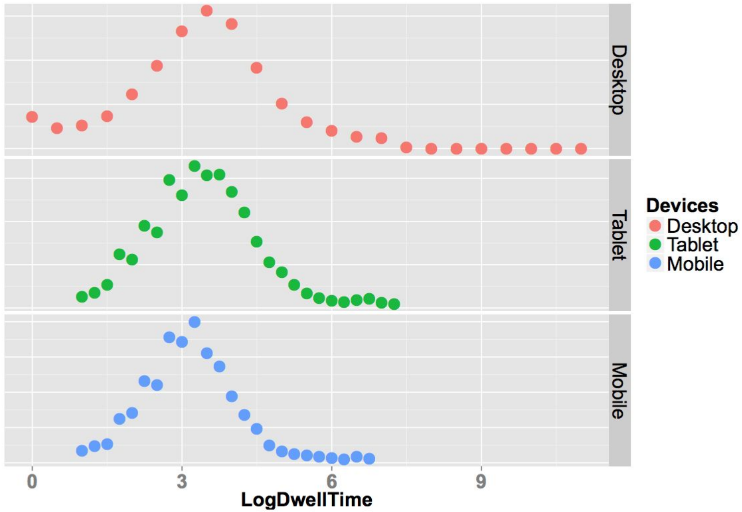


Figure 5: The (un)normalized distribution of log of dwell time for slideshows across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

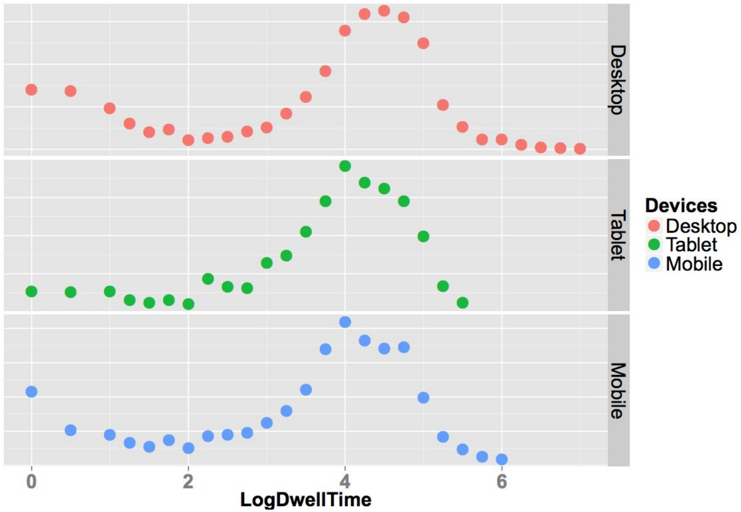


Figure 6: The (un)normalized distribution of log of dwell time for videos across different devices. The X-axis is the log of dwell time and the Y-axis is the counts.

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

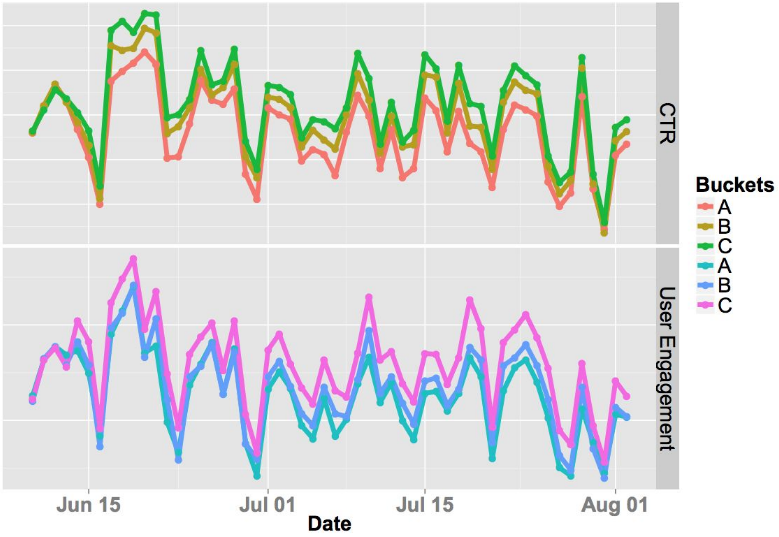


Figure 7: The relative performance comparison between three buckets. The top figure shows the relative CTR difference and the bottom figure shows the relative user engagement difference.

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.

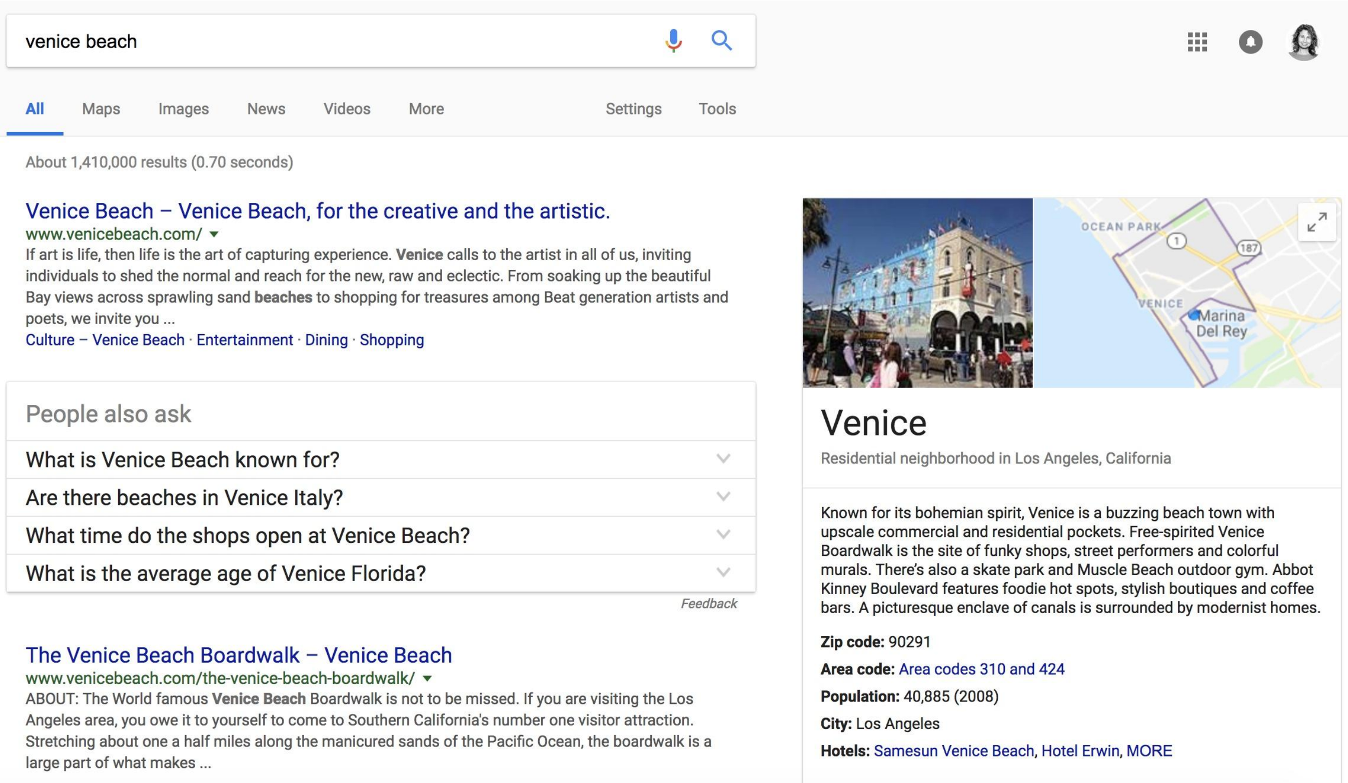
Optimization Inter-Session Metrics

Summary

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

Application: Search

Is this a good search engine?



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

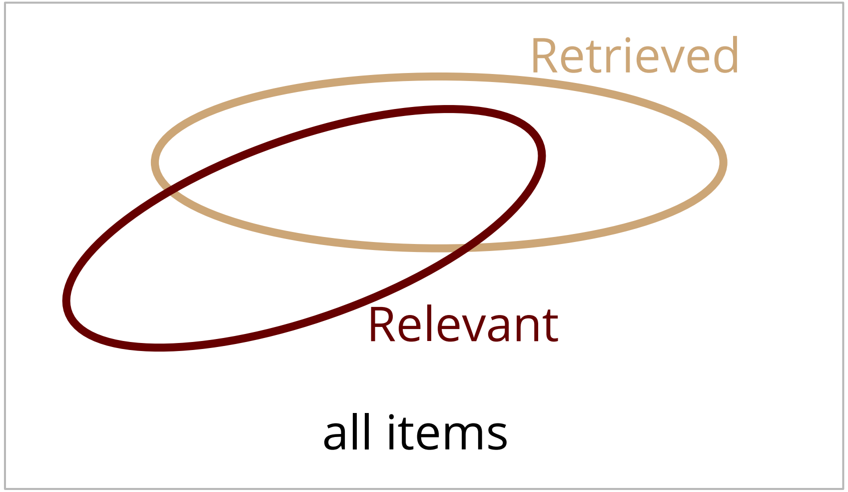
- Users find what they want and return to the search engine for their next information need → **user engagement**

But let us remember:

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

(Manning, Raghavan & Schütze, 2008; Baeza-Yates & Ribeiro-Neto, 2011)

Evaluating the relevance of a search engine result



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

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

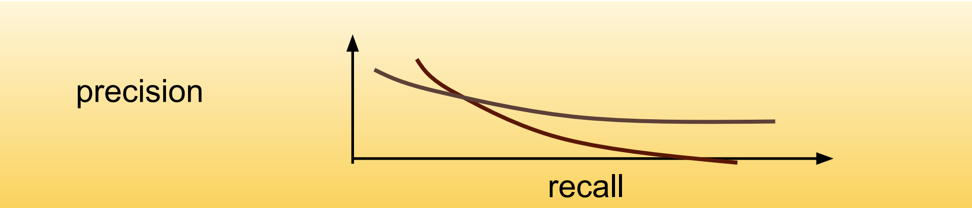
Example:

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

<u>Query</u>: ***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

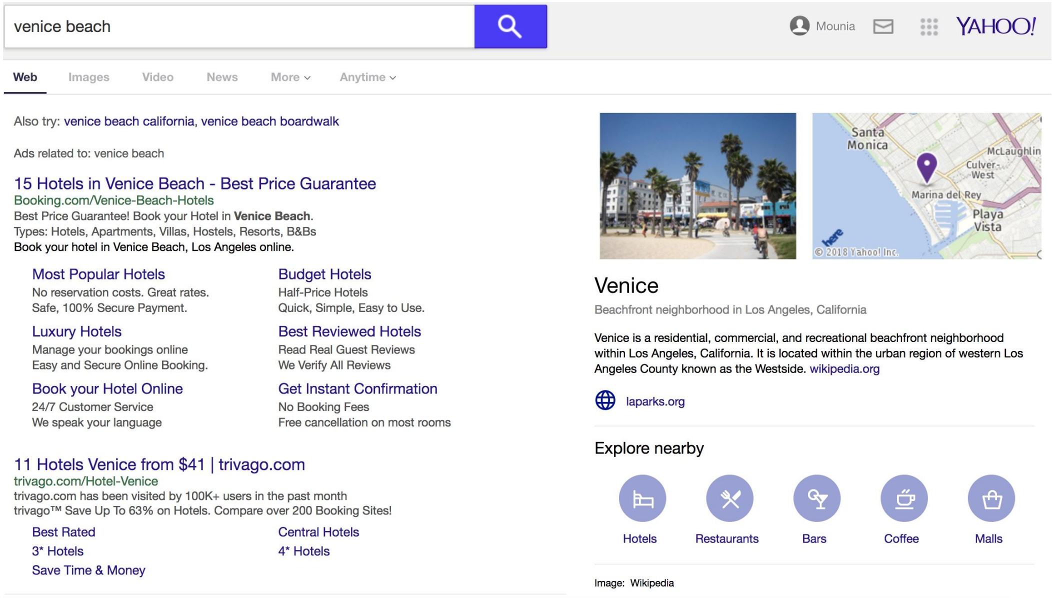
- User behavior in online settings (clicks, skips, dwell time)

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



An important question:

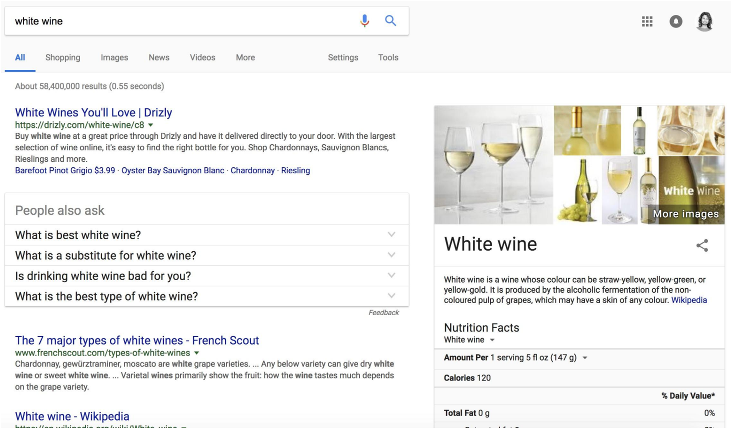
when is signal a metric and when is it a feature of the ranking (machine learning) algorithm?

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



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

MAP … User satisfaction

Precision-based search

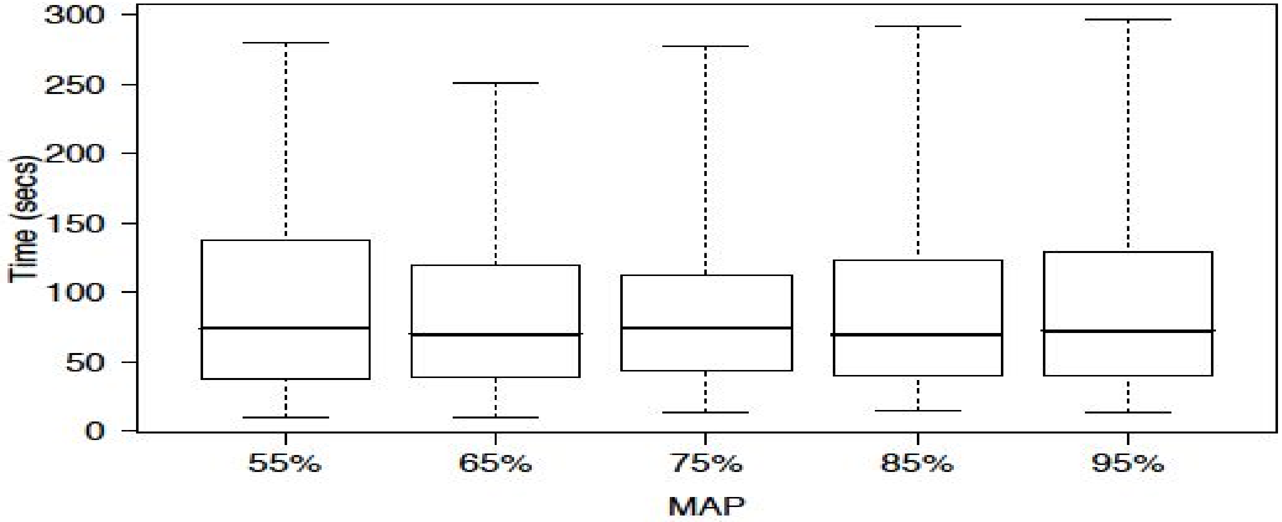


Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

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

(Turpin & Scholer, 2006)

MAP ... User satisfaction

Recall-based search

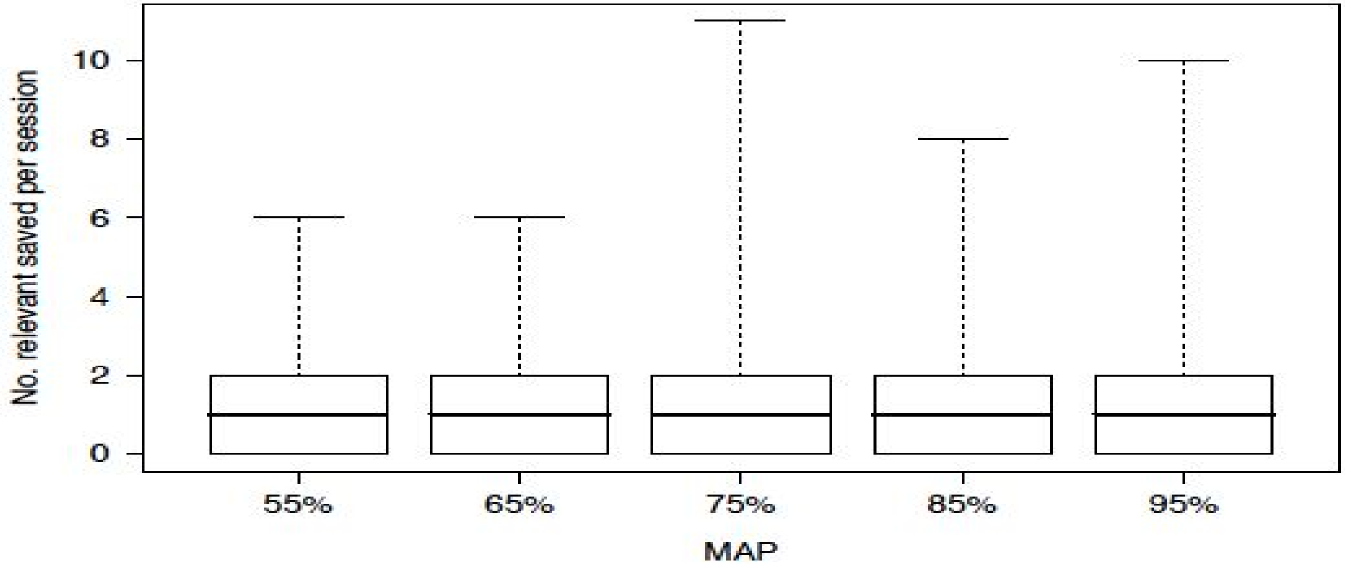
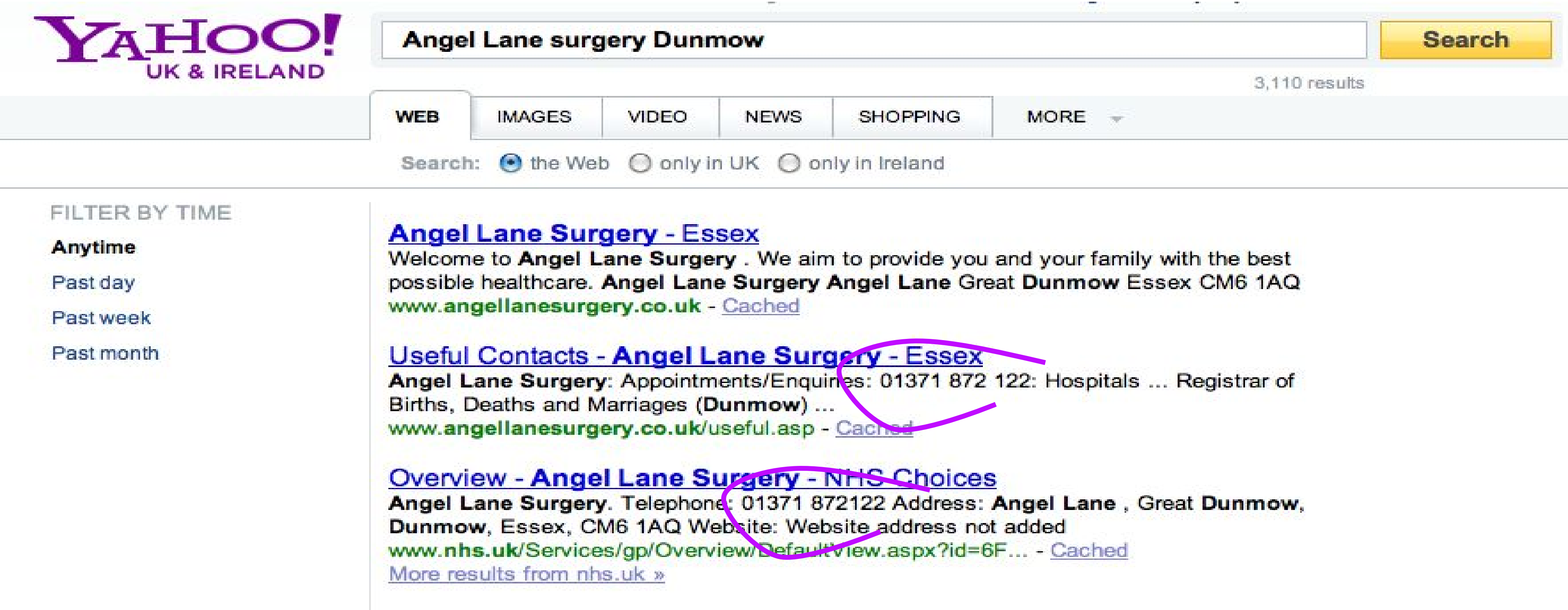


Figure 7: Number of relevant documents found by users within five minutes for systems with differing MAP.

(Turpin & Scholer, 2006)

No click … User satisfaction



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**

(Huang et al, 2011)

Cickthrough 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)

(Huang et al, 2011)

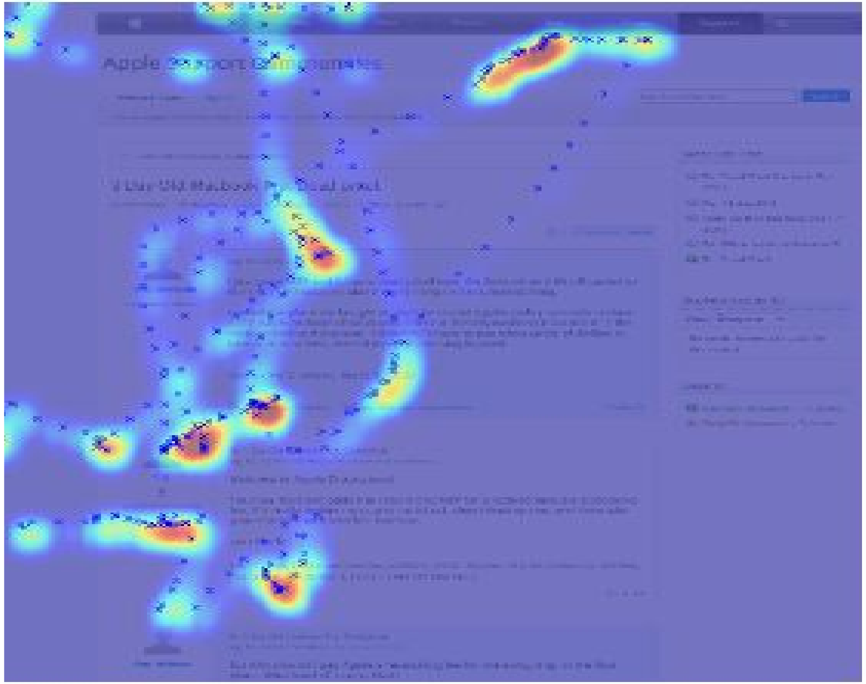
Dwell time … User satisfaction

(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



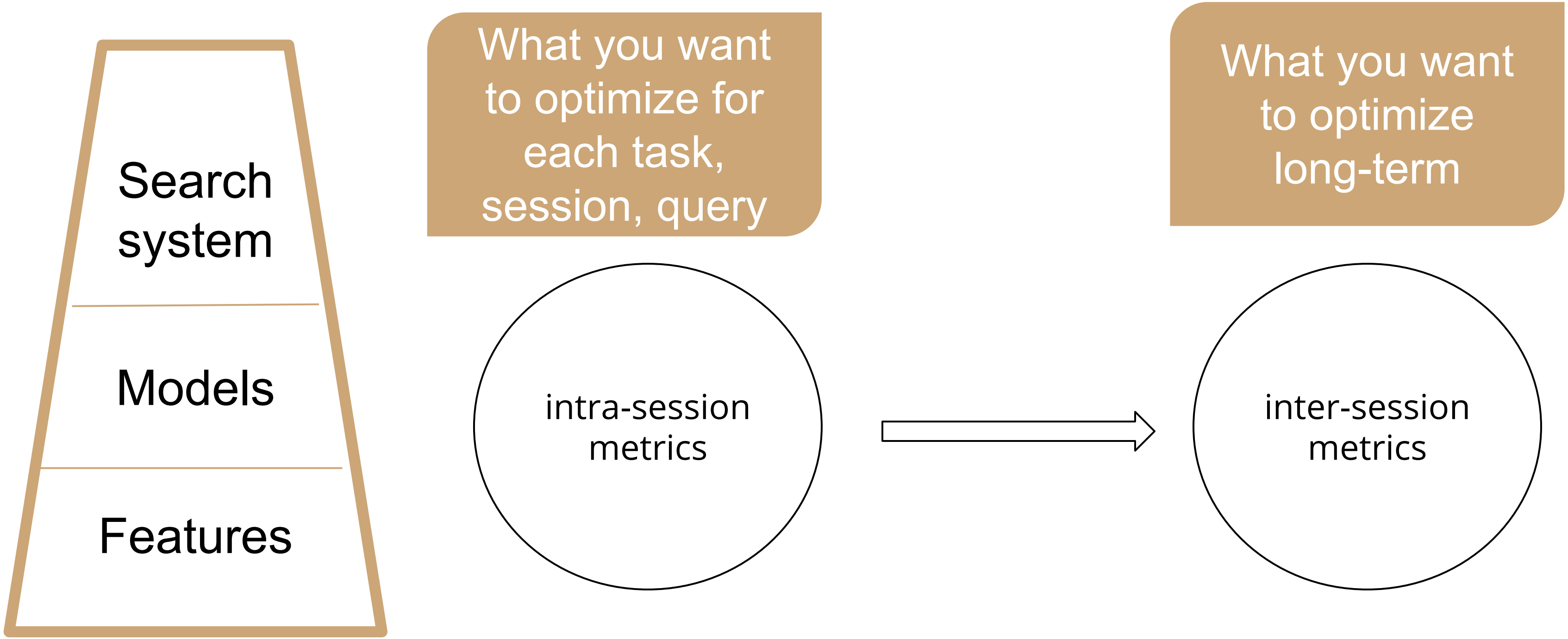
(a) relevant (dwell time: 70s)

(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
- ...

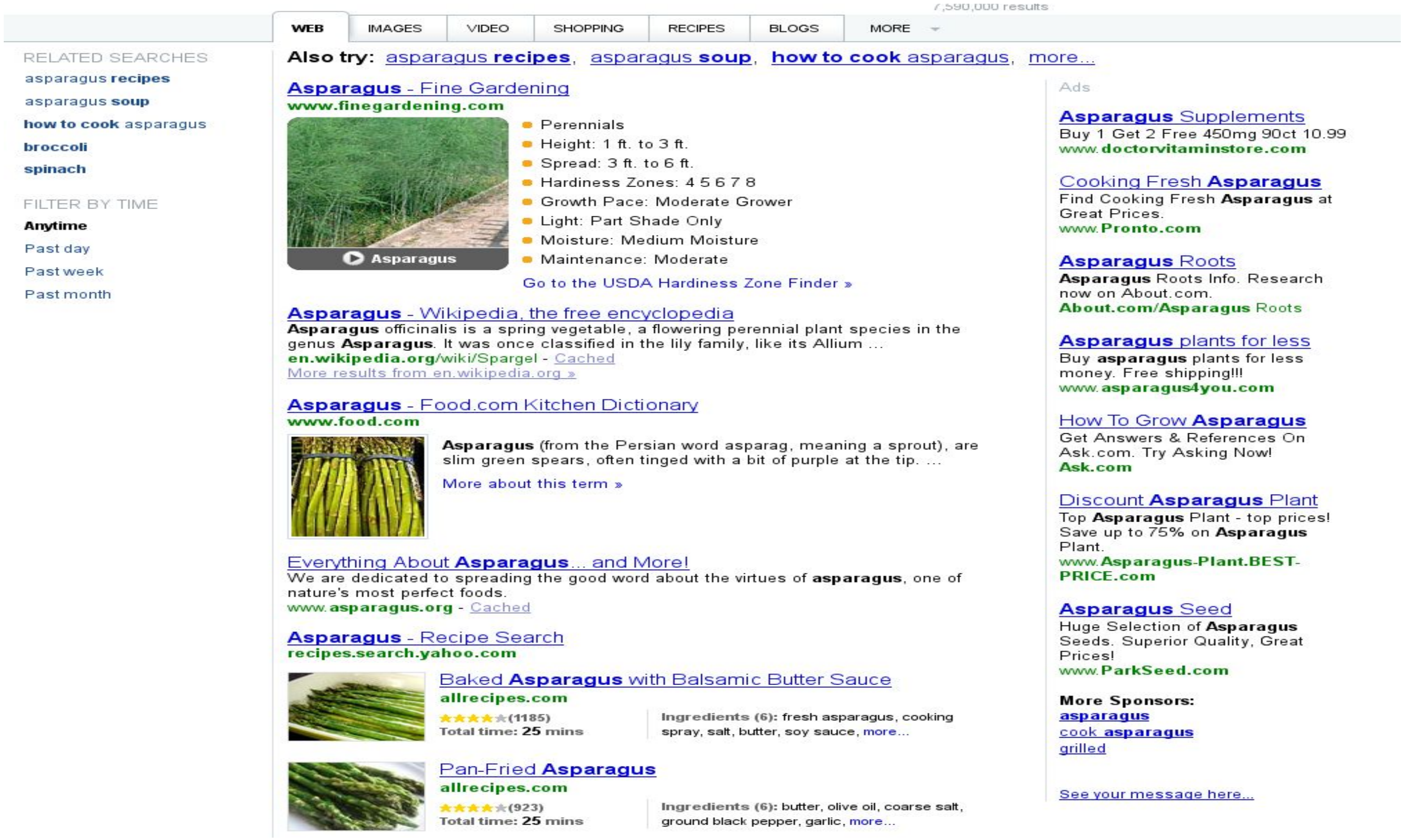
users satisfied with the search session are likely to return sooner and frequently to the search engine

Inter-session metrics for search

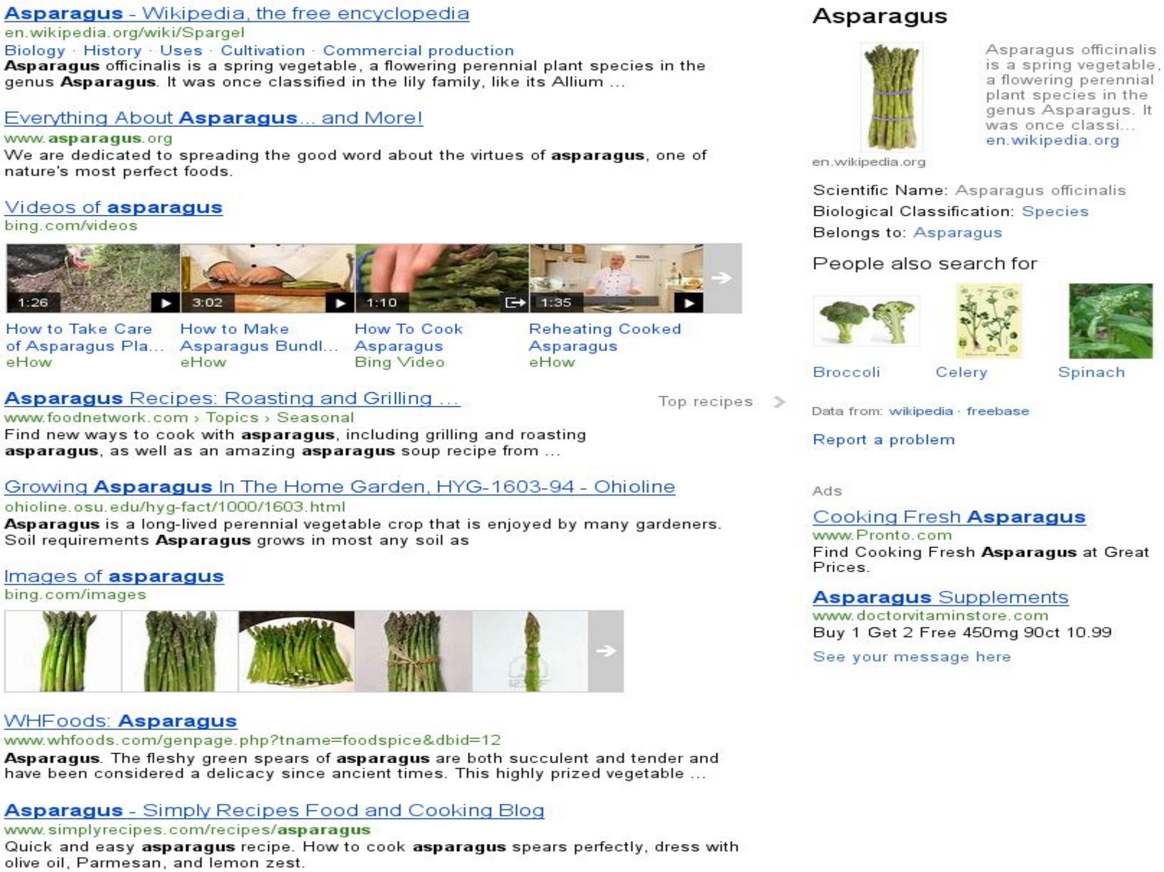
- Absence time
- Number of search sessions in next 2 weeks
- Number of queries next day
- ...

Absence time on Yahoo Japan (Dupret & Lalmas, 2013)
Absence time on Bing (Chakraborty etal, 2014)
Dwell time & search engine re-use (Hu etal, 2011)

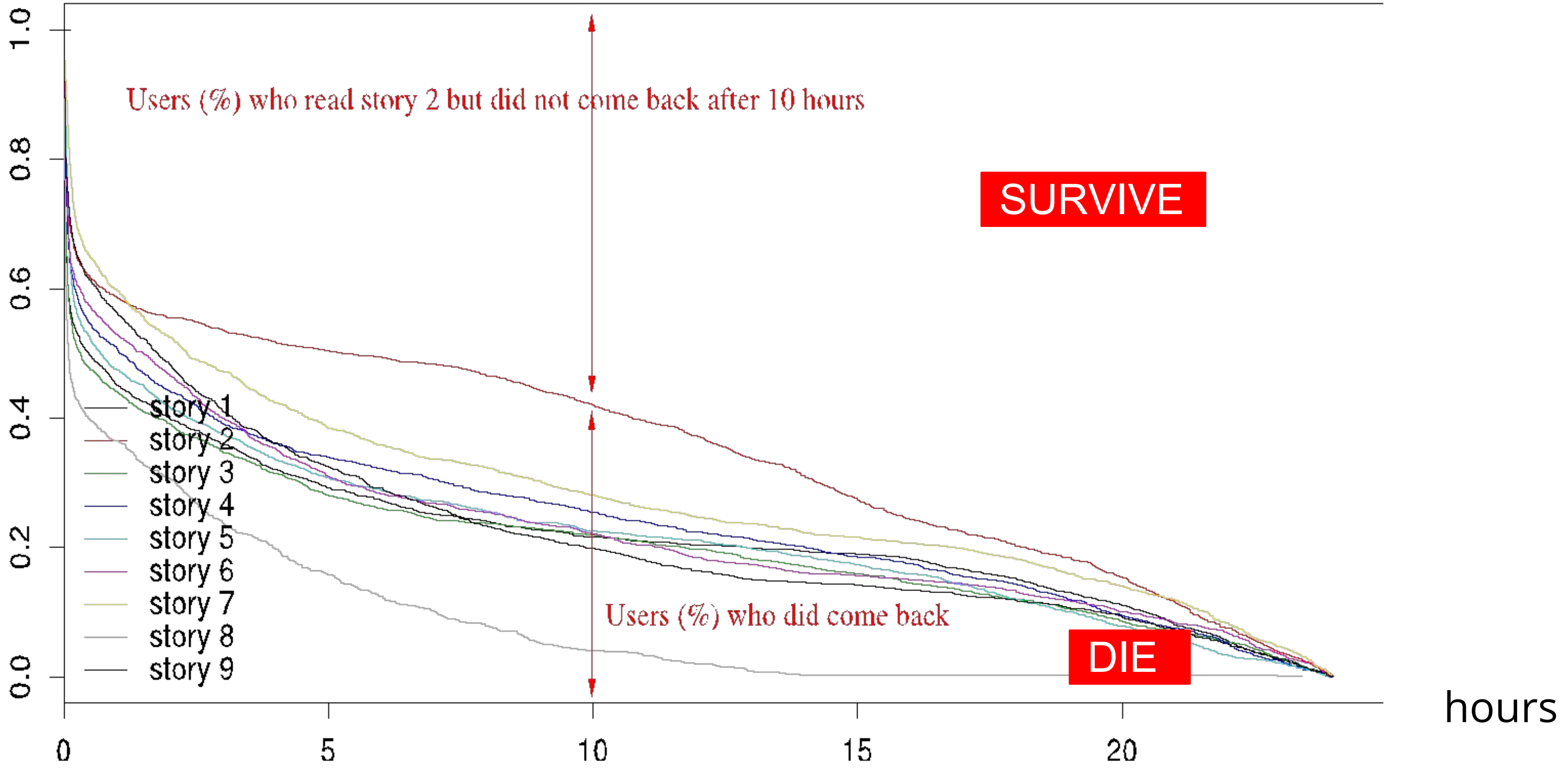
Search result page for “asparagus” … Study I



Another search result page for “asparagus”



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

Session boundary: 30 minutes of inactivity

Two-weeks click data on Yahoo Answer Japan search

One millions users

Six ranking functions

(Dupret & Lalmas, 2013)

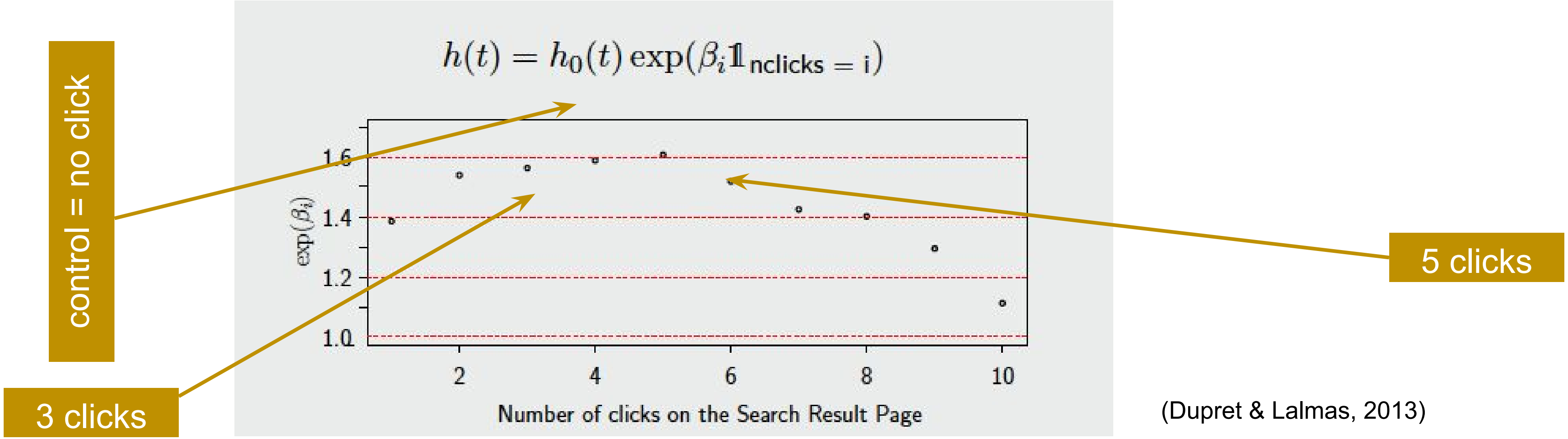
DCG versus absence to evaluate five ranking functions

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

(Dupret & Lalmas, 2013)

Absence time and number of clicks

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



(Dupret & Lalmas, 2013)

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

(Dupret & Lalmas, 2013)

Absence time and search experience … Study II

intra-session search metrics → absence time

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

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

Negative
Abandoned session
Quick-back clicks

(Chakraborty et al., 2014)

Absence time and search experience … Studies I & II

intra-session search metrics → absence time

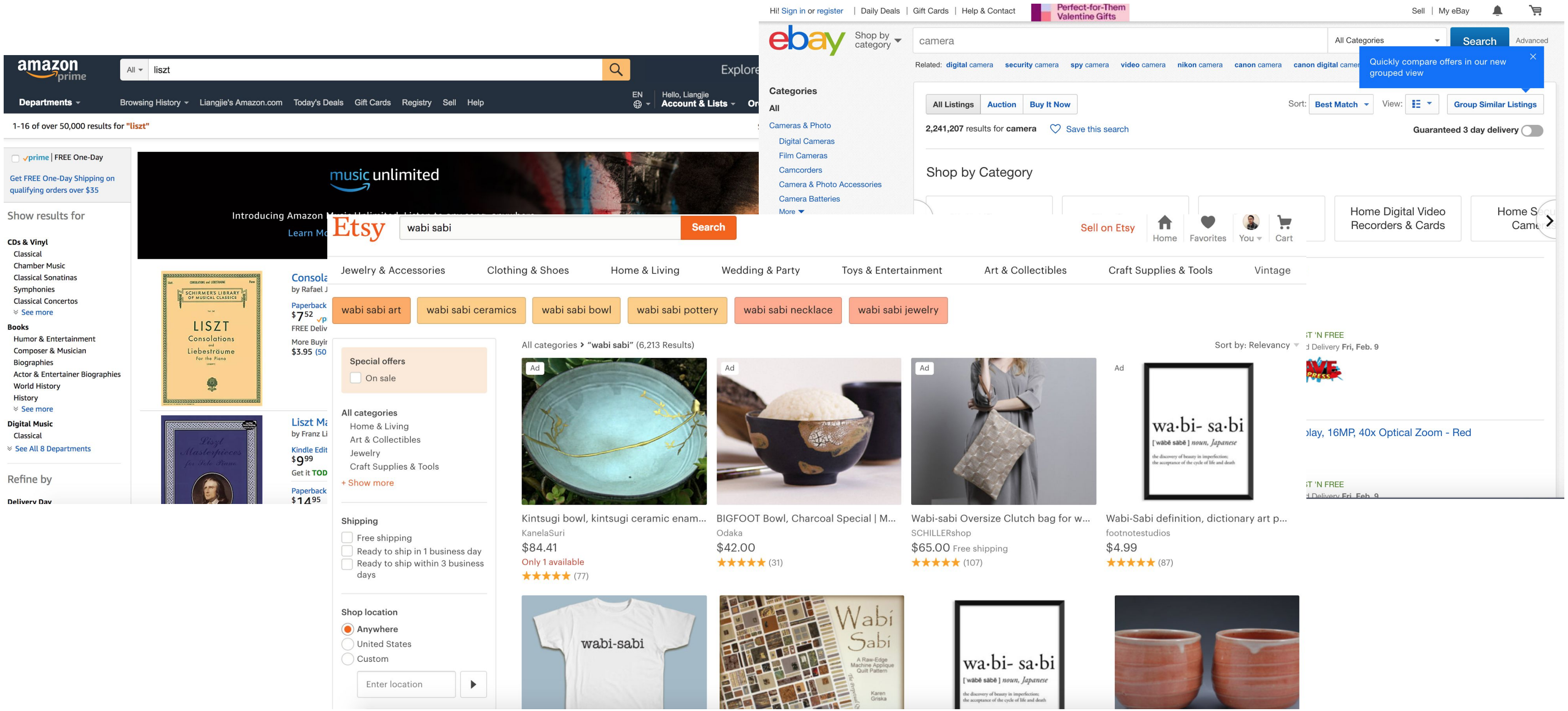
Demonstrated that absence time is an appropriate inter-session metric for search because of the correlation & predictive power of known indicators of a positive search experience

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



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

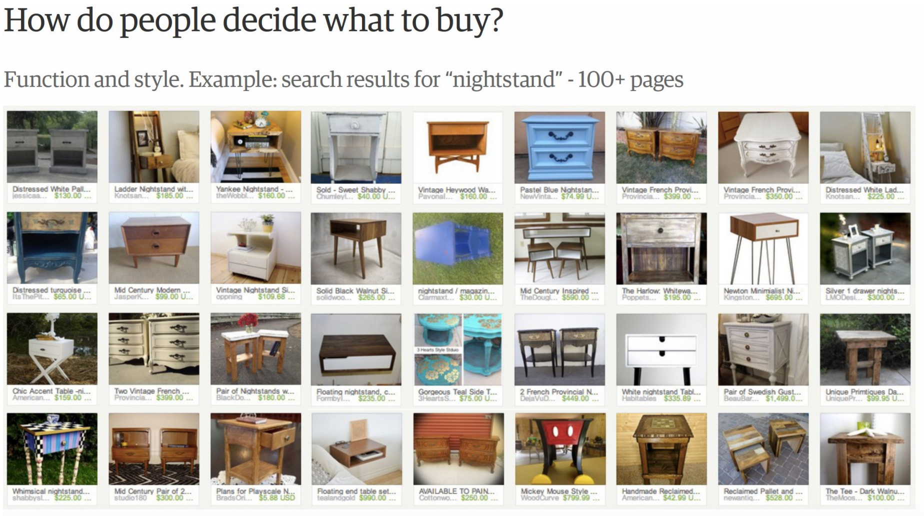
Application: E-commerce

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

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

Application: E-commerce

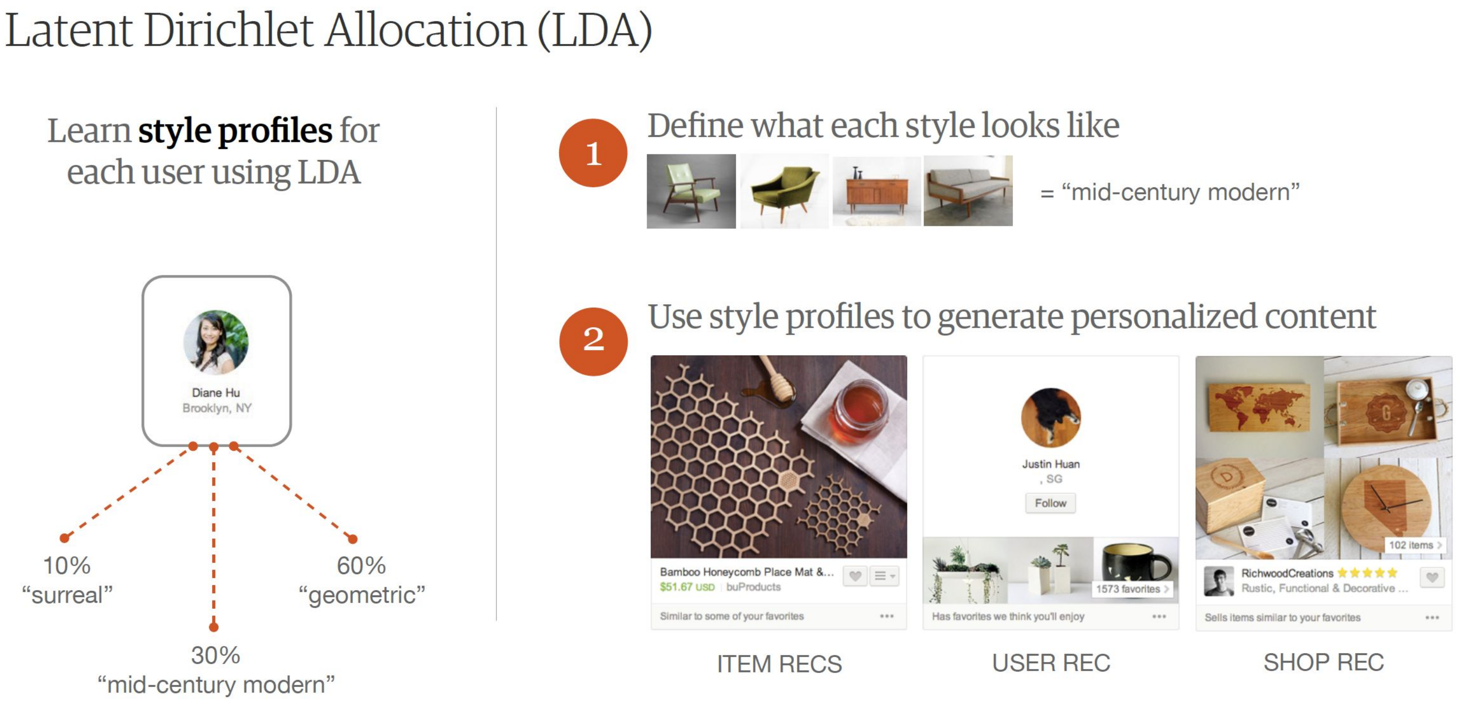
- **Discovering Styles for Recommendation in E-Commerce**



- **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.

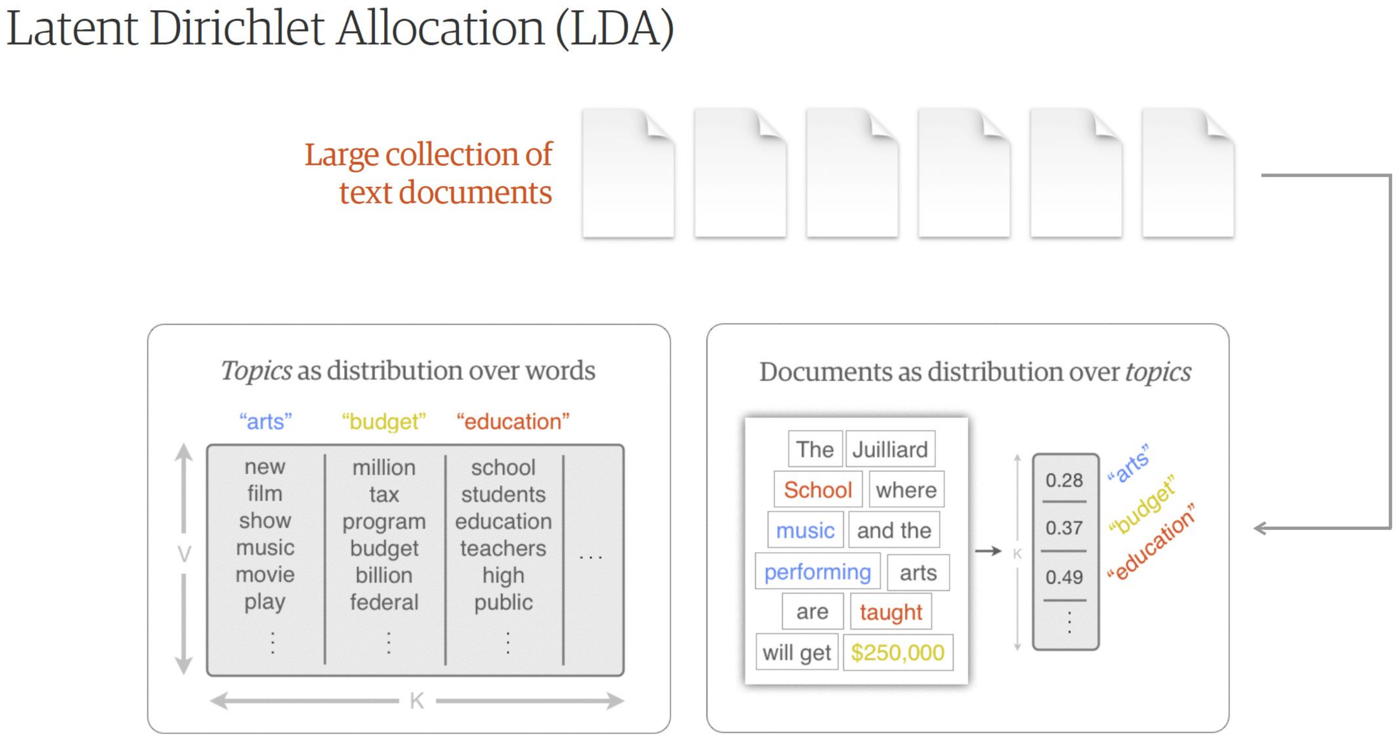
Application: E-commerce

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



Application: E-commerce

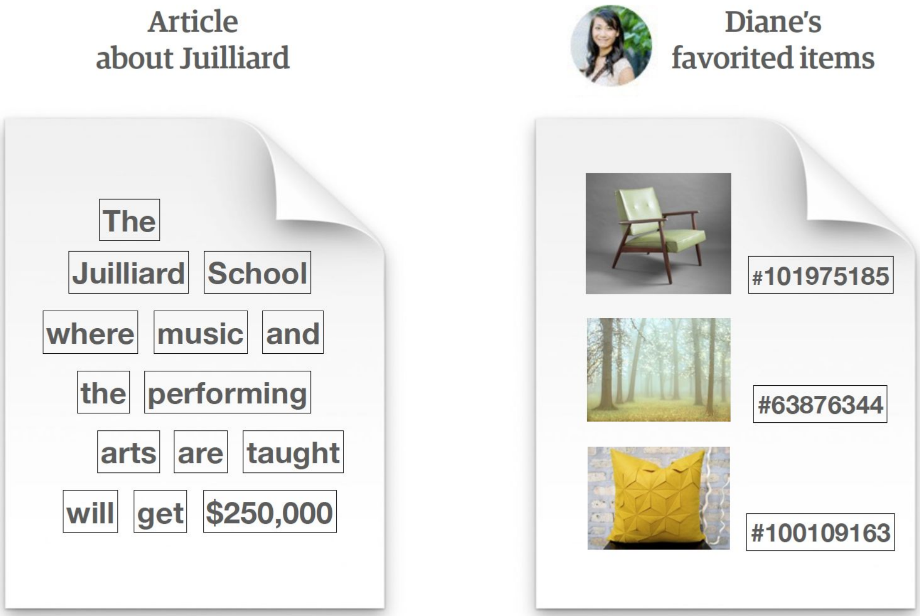
- **Discovering Styles for Recommendation in E-Commerce**



Application: E-commerce

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

Latent Dirichlet Allocation (LDA)



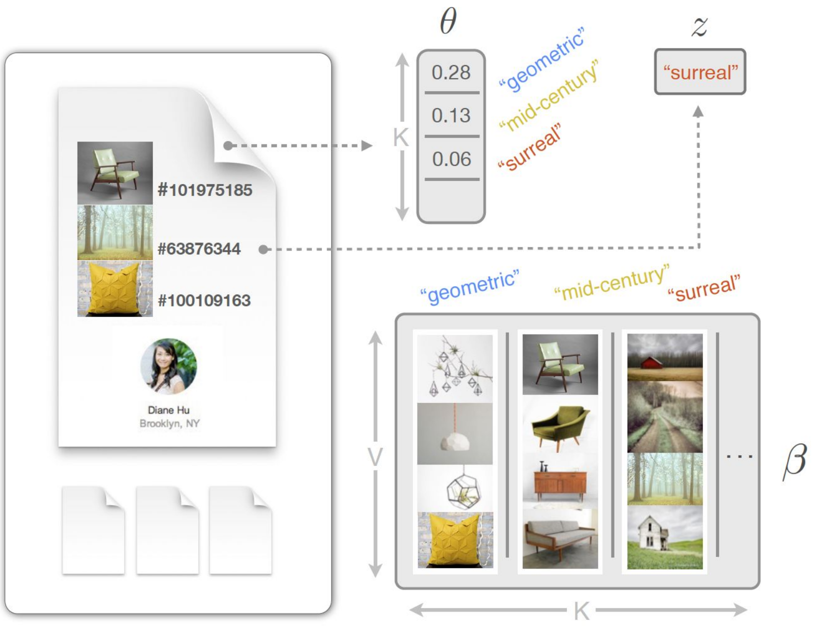
Application: E-commerce

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

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

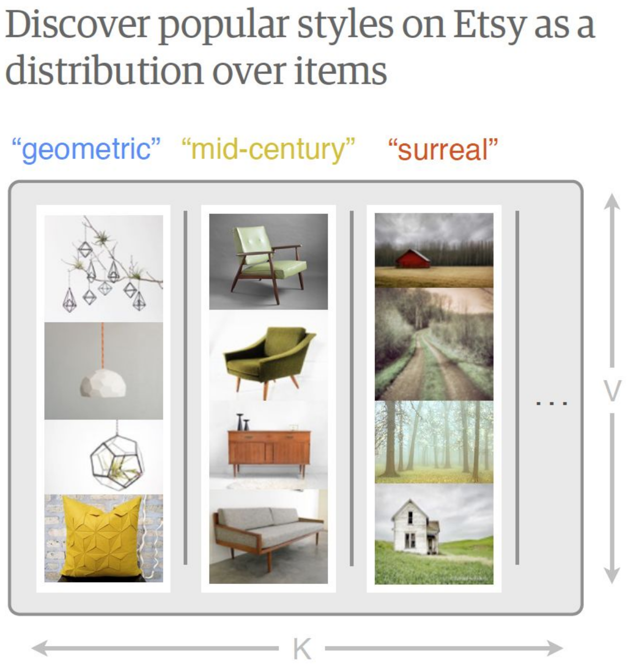
For each user u,

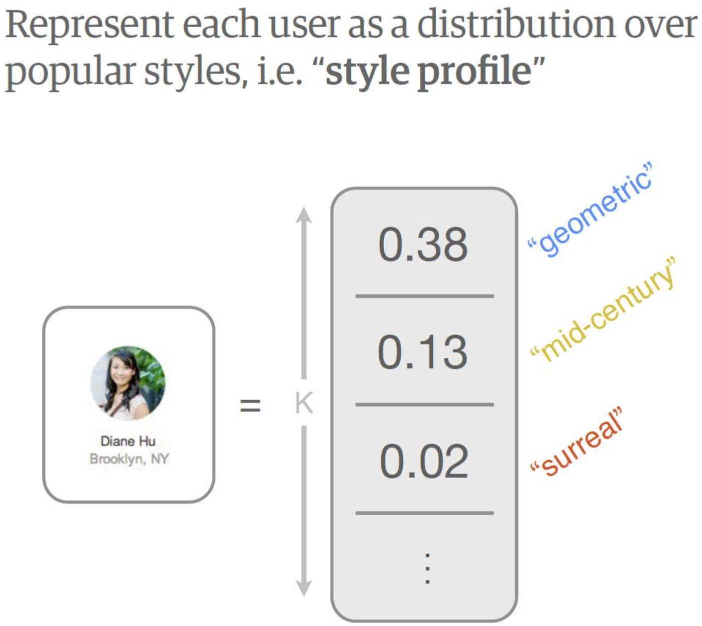
1. Draw a style profile:
   $\theta \sim Dirichlet(\alpha)$
2. For each item, x_n that user u has favorited,
   (a) Draw a style:
   $z_n \sim Multinomial(\theta)$
   (b) Draw an item:
   $x_n \sim Multinomial(\beta_{z_n})$



Application: E-commerce

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

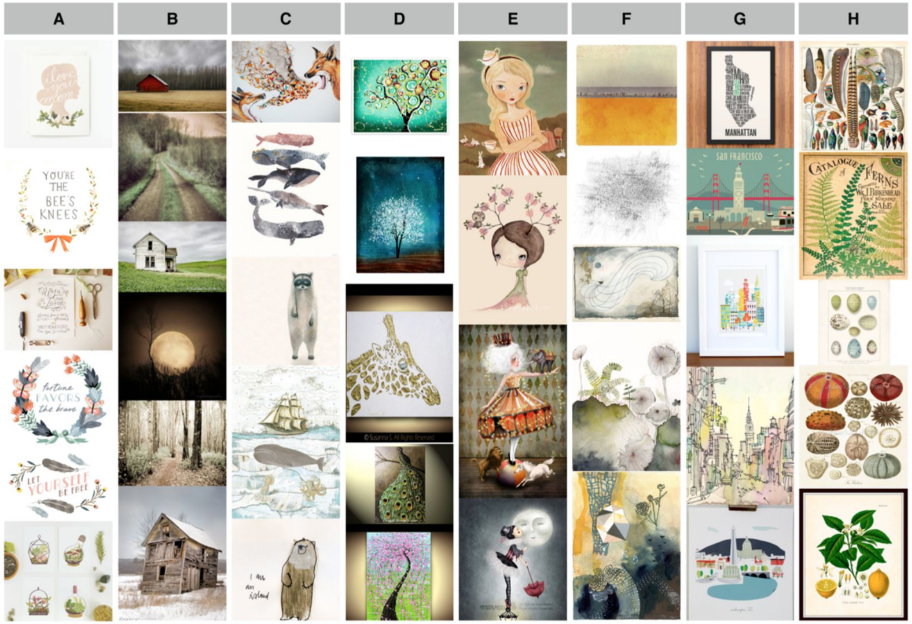




Application: E-commerce

- **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

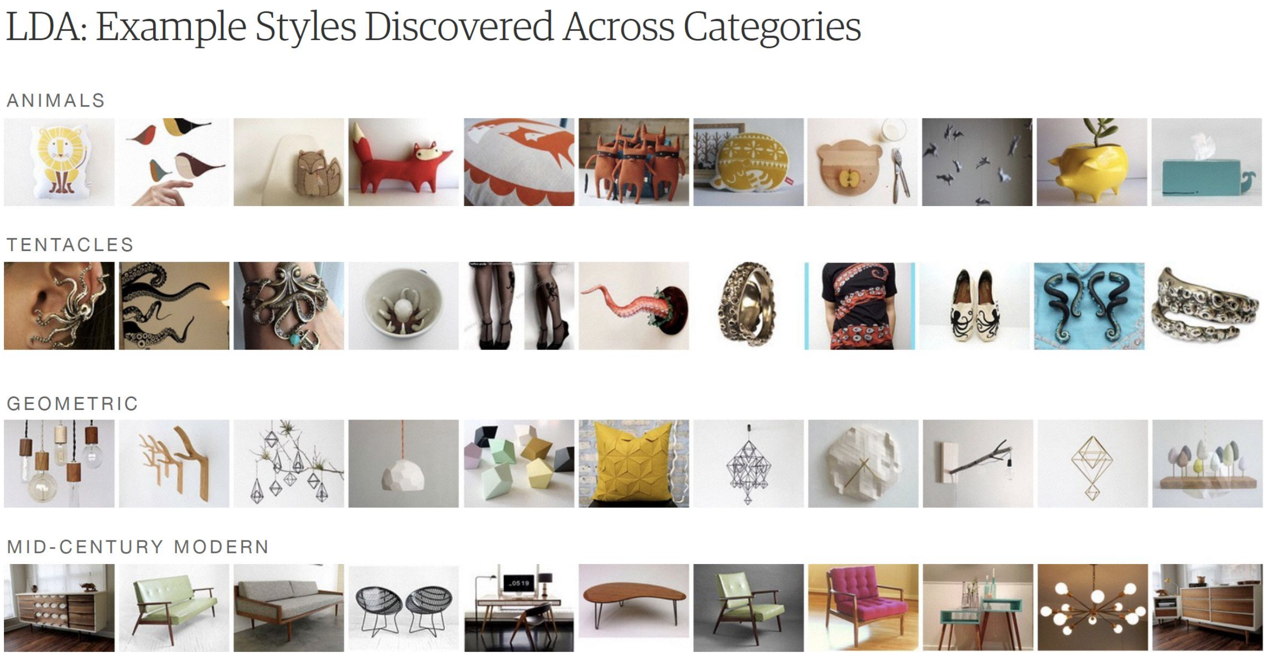
F = Whimsical/Abstract

G = Cities

H = Vintage

Application: E-commerce

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

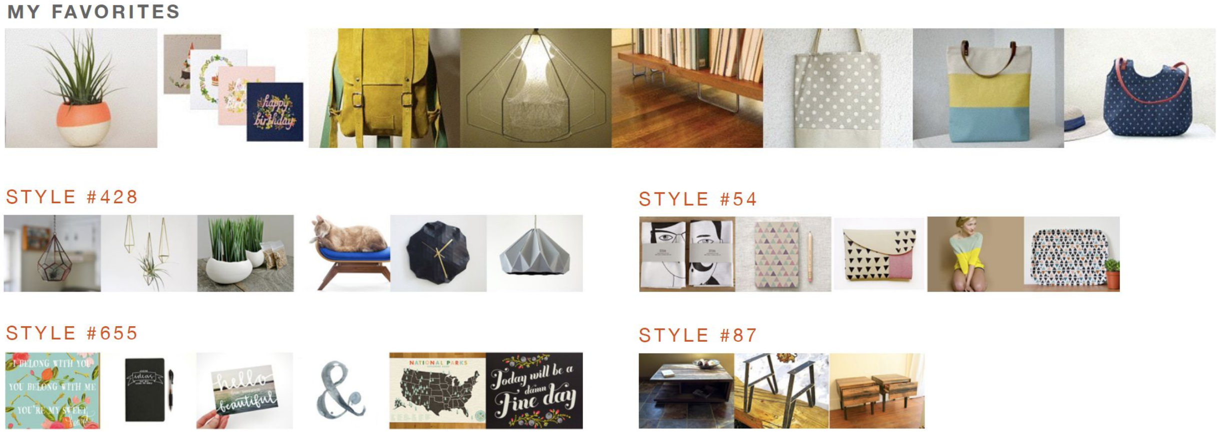


Application: E-commerce

- **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



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**

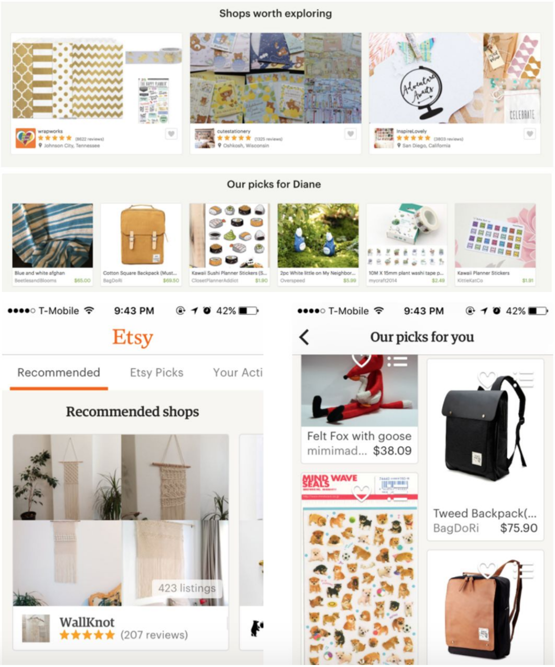
(1) Personalized Recommendations

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

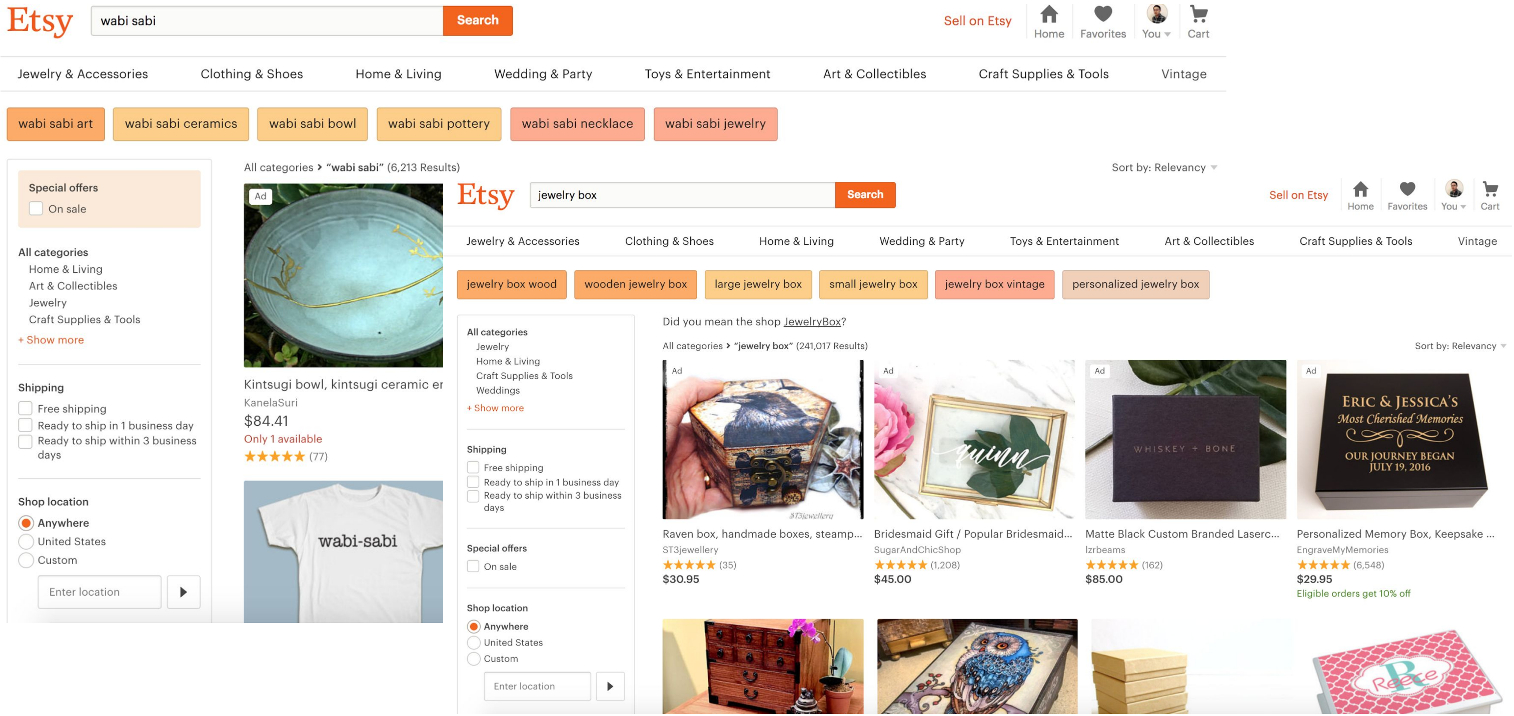
Application: E-commerce

- 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

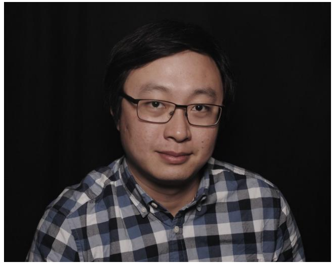


Application: E-commerce

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

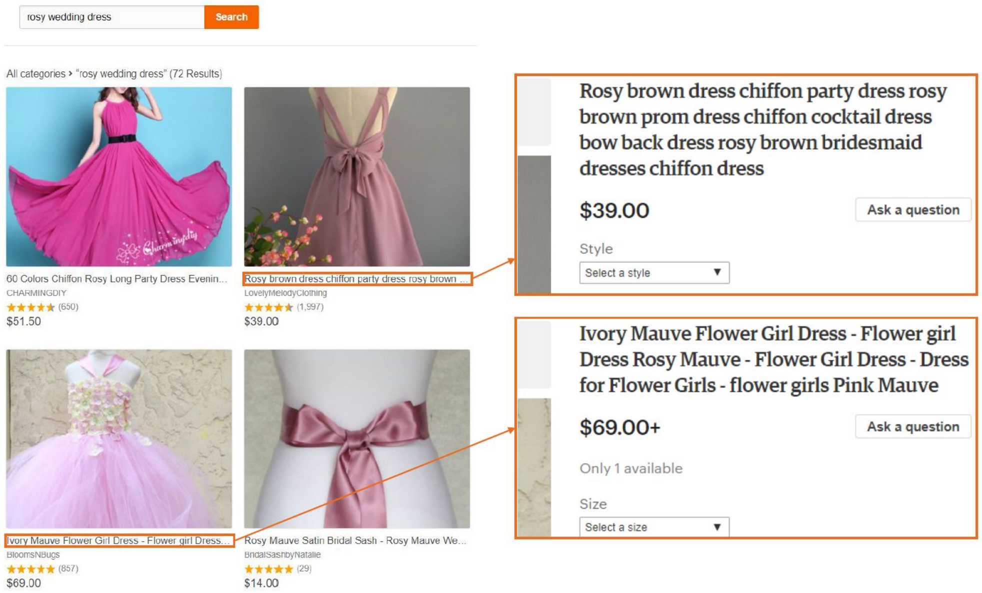
- **Expected GMV**

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

Application: E-commerce

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

- **Purchase Decision Process**



Search Page

Product Page

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)**

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

$$\downarrow$$

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

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

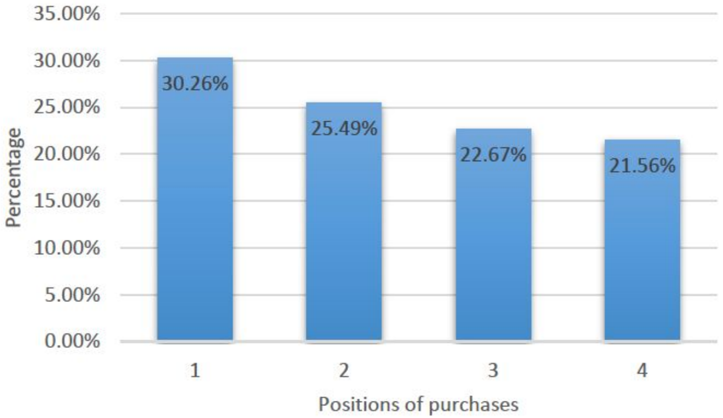


Figure 2: Position distribution of items being purchased in the top 4 spots of a search result page. The first position achieves the most purchases, while nearly 70% of purchases are in the lower positions.

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	$BM25$
		Log $BM25$
		LM_{DIR}
		LM_{JM}
		LM_{ABS}
Revenue		$Price$
		$Price - Cat.Mean$
		$(Price - Cat.Mean)/Cat.Mean$

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

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