Metrics, Engagement & “Recommenders”

Mounia Lalmas

About the workshop

The State-based User Modelling Workshop at WSDM 2020 (SUM '20) welcomes participation of researchers and stakeholders to explore techniques for state-based user modelling, and their applications, such as:

- Conversational systems
- User representation and recommender systems
- Task understanding and supporting user tasks
- Task-based information retrieval
- State-aware evaluation
- Human in the loop
- User-aware systems
- Personalisation ML algorithms
- Cognitive/contextual user understanding
- State-aware ML algorithms

Through the (my) lens on user engagement

Outline

About

How

About

User engagement

Metrics

Interpretations

About

User engagement

Metrics

Interpretations

What is user engagement?

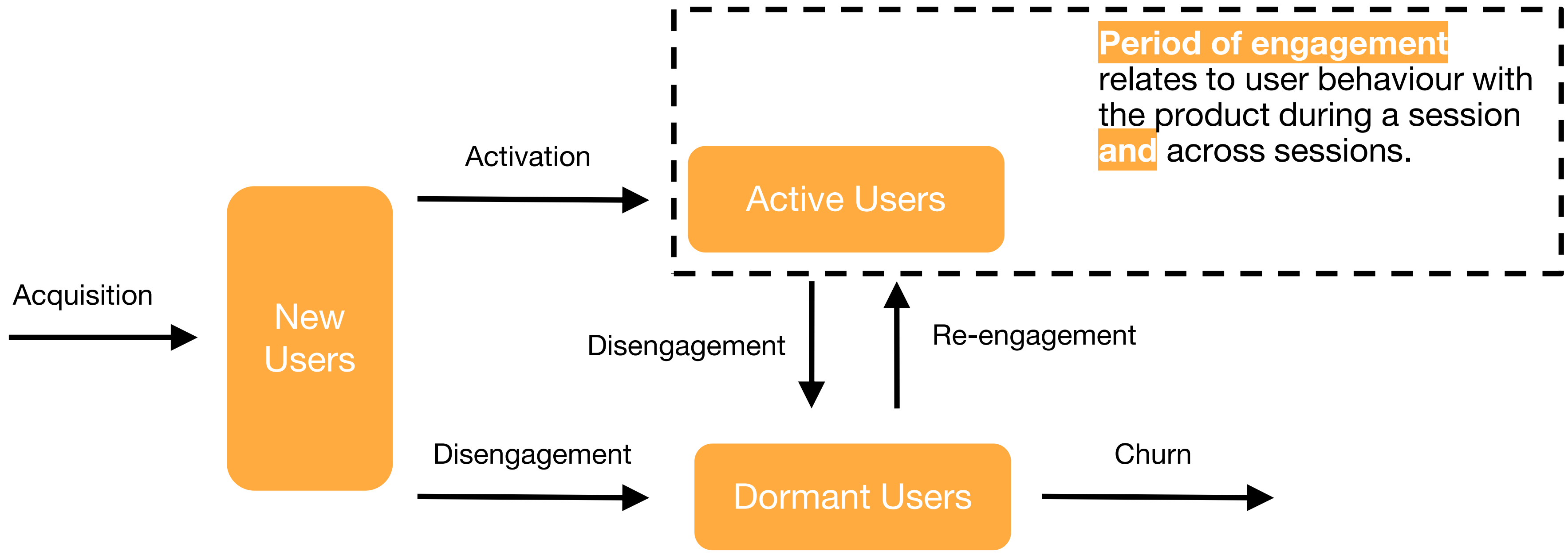
User engagement is the quality of the user experience that emphasizes the positive aspects of interaction – in particular the fact of wanting to use the technology longer and often.

S Attfield, G Kazai, M Lalmas & B Piwowarski. **Towards a science of user engagement (Position Paper)**. WSDM Workshop on User Modelling for Web Applications, 2011.

The engagement life cycle

Point of engagement	**How engagement starts (acquisition & activation)** 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 recommenders.
Disengagement	**Loss of interests leads to passive usage & even stopping usage** Identifying users that are likely to churn often undertaken.
Re-engagement	**Engage again after becoming disengaged** Triggered by relevance, novelty, convenience, remembering past positive experience sometimes as result of campaign strategy.

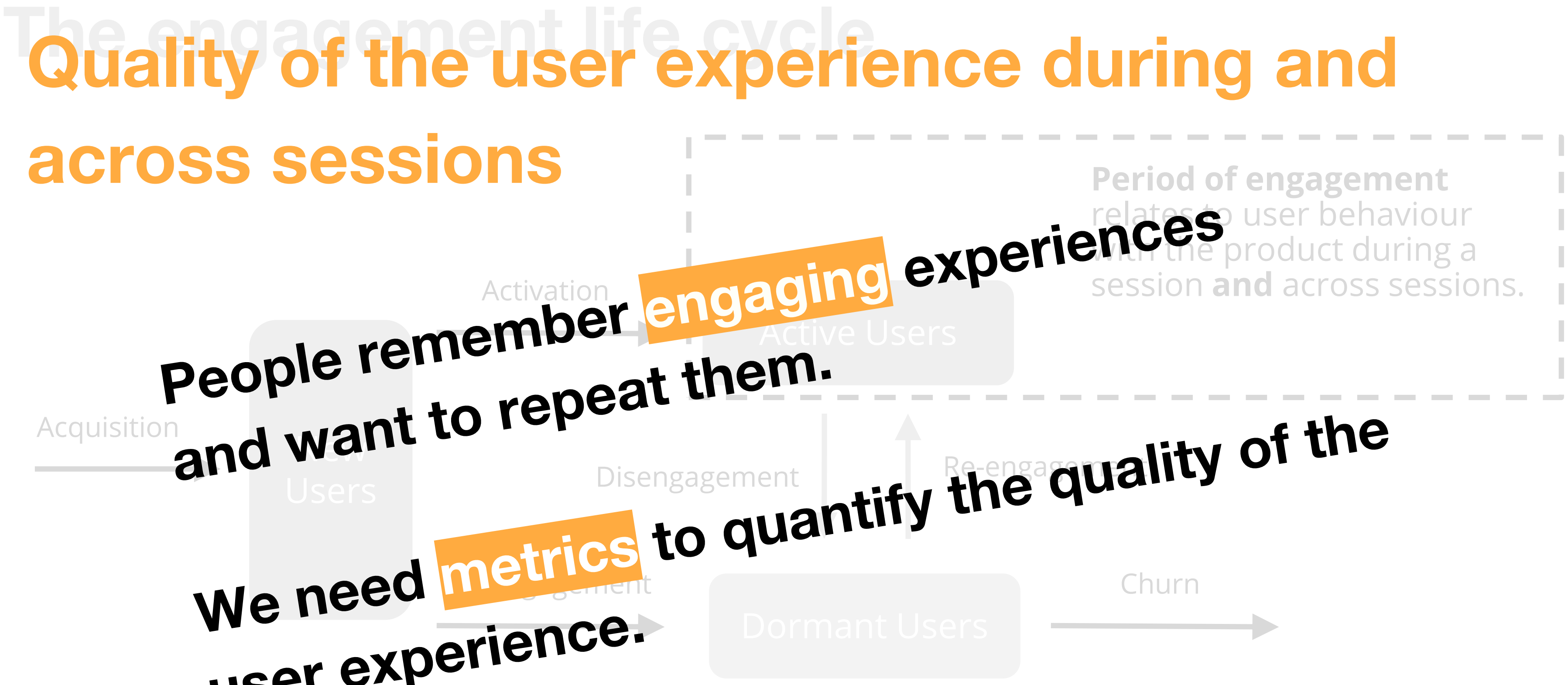
The engagement life cycle



Quality of the user experience during and across sessions

People remember engaging experiences and want to repeat them.

We need metrics to quantify the quality of the user experience.

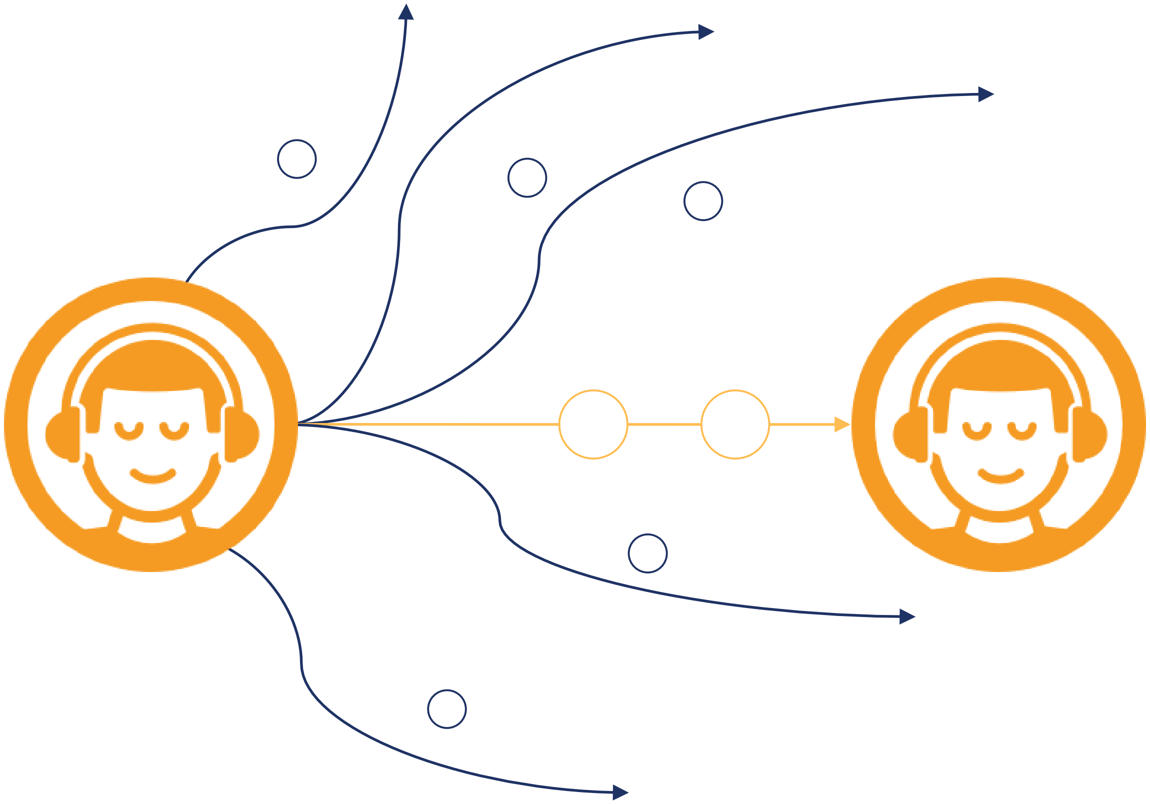


State-aware ML algorithms

It is about the
user engagement
journey.

Within a session.

Across sessions.



About

User engagement

Metrics

Interpretations

Three levels of metrics

1. Business metrics	KPIs
2. Behavioral metrics	Online metrics
3. Optimization metrics	Objective metrics to train recommenders

Optimization metrics mostly quantify how users engage within a session and act as proxy of engagement.

post

save

long click

percentage completion

follow

impression to click

click to stream

abandonment rate

dwell time

State-aware ML algorithms

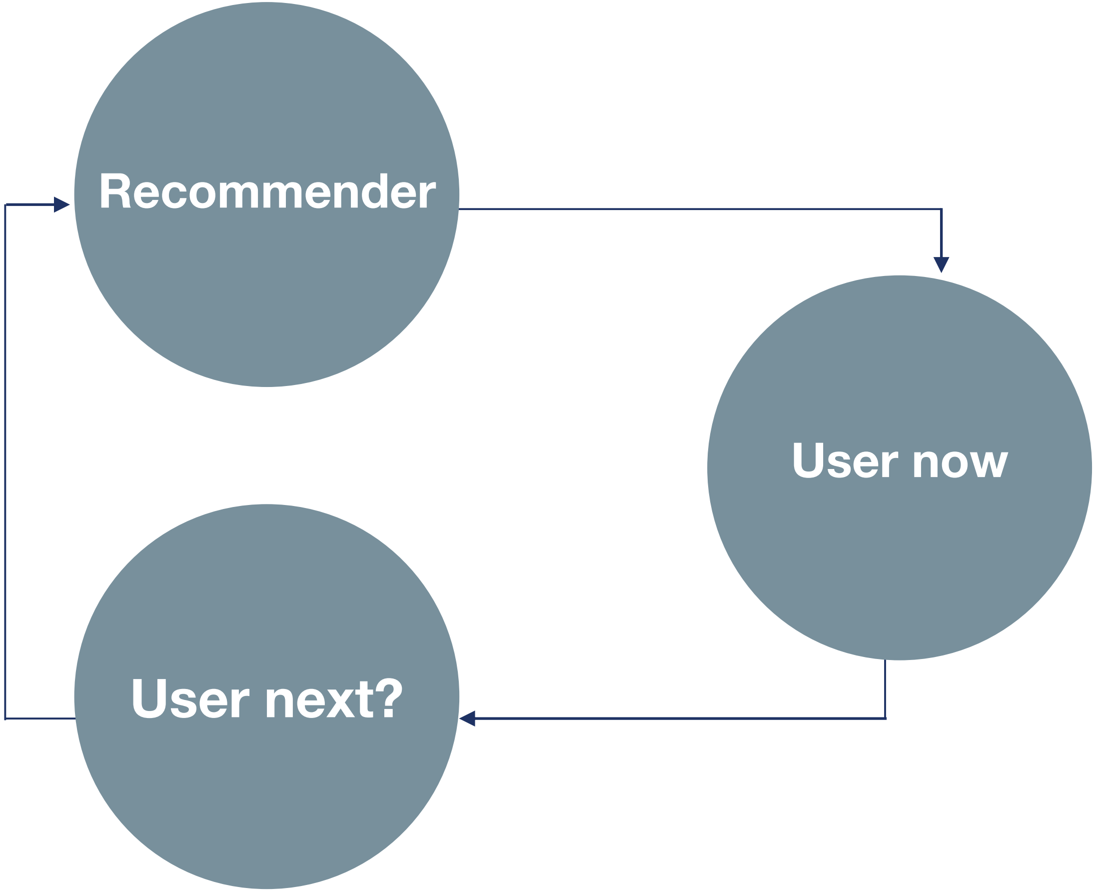
Optimization metric

Within session?

Within this and next sessions?

…

Across sessions?



About

User engagement

Metrics

Interpretations

Click

What is the **value** of a click?

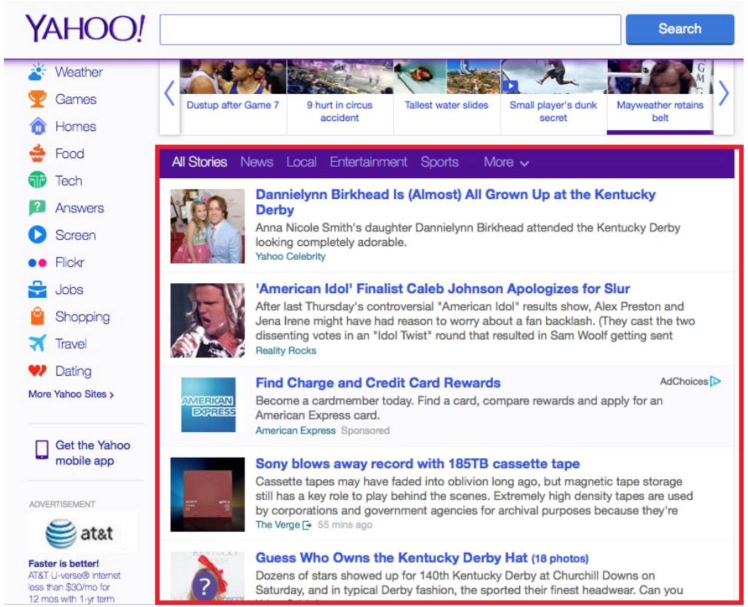
Click-through rate = ratio of users who click on a specific link to the number of total users who view a page, email, advertisement, …

Most used optimization metric.

Dwell time is a better proxy for user interest on a news article than click.

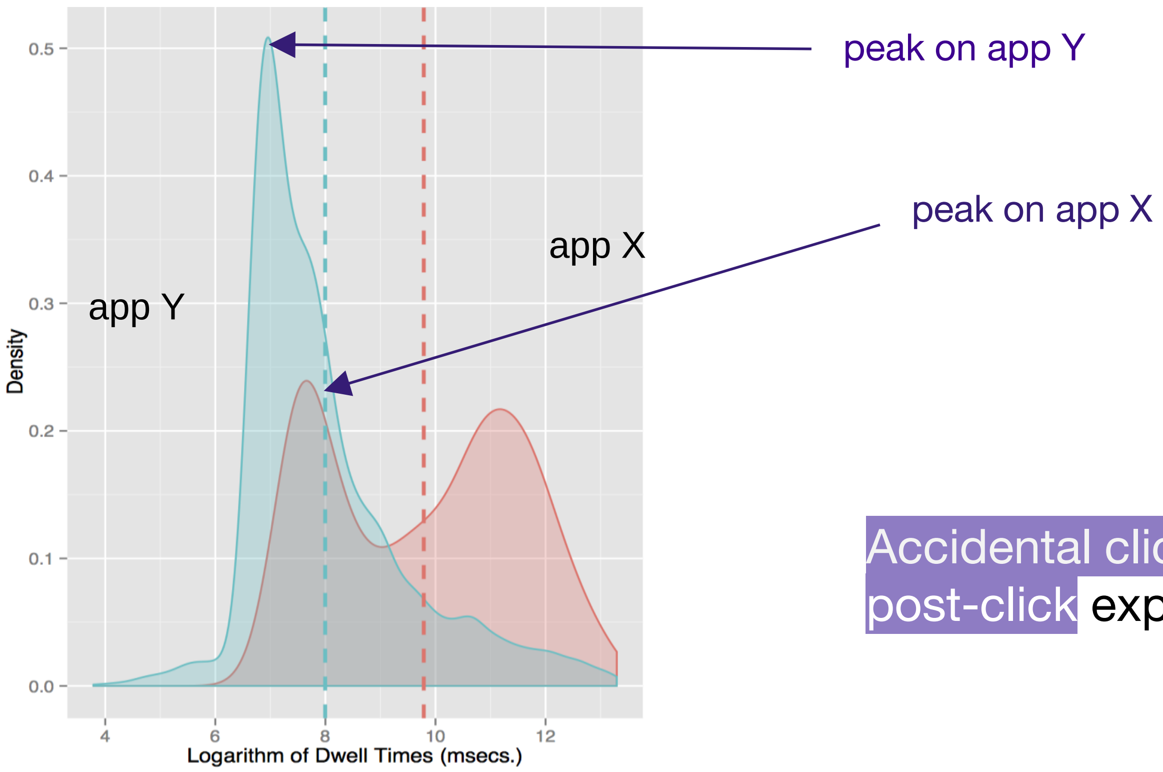
An efficient way to reduce click-baits.

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



X Yi, L Hong, E Zhong, N Nan Liu & S Rajan. **Beyond Clicks: Dwell Time for Personalization**. RecSys 2014.
H Lu, M Zhang, W Ma, Y Shao, Y Liu & S Ma. **Quality Effects on User Preferences and Behaviors in Mobile News Streaming User Modeling**. WWW 2019.

dwell time distribution of apps X and Y for given ad



Accidental clicks do not reflect post-click experience.

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

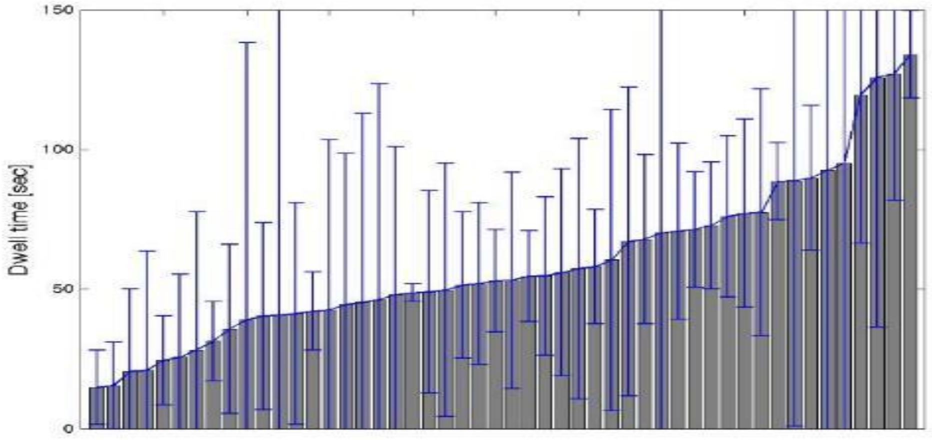
Dwell time

What does spending time really means?

Dwell time = contiguous time spent on a site or web page.

Dwell time varies by site type.

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



average and variance of dwell time of 50 sites

E Yom-Tov, M Lalmas, R Baeza-Yates, G Dupret, J Lehmann & P Donmez. **Measuring Inter-Site Engagement.** BigData 2013.

Reading cursor heatmap of relevant document vs scanning cursor heatmap of non-relevant document

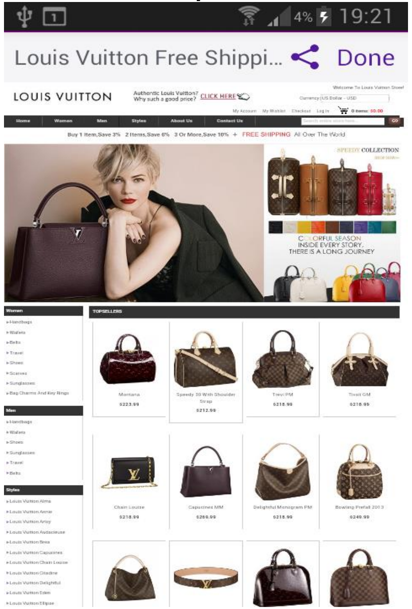
(a) relevant (dwell time: 30s)   *(b) non-relevant (dwell time: 30s)*

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

Dwell time on non-optimized landing pages comparable and even higher than on mobile-optimized ones.

non-mobile optimized

mobile optimized



Dwell time used as proxy of landing page quality.

M Lalmas, J Lehmann, G Shaked, F Silvestri & G Tolomei. **Promoting Positive Post-click Experience for In-Stream Yahoo Gemini Users.** KDD Industry track 2015.

How

Understanding intents

Optimizing for the right metric

Acting on segmentation

Thinking about diversity

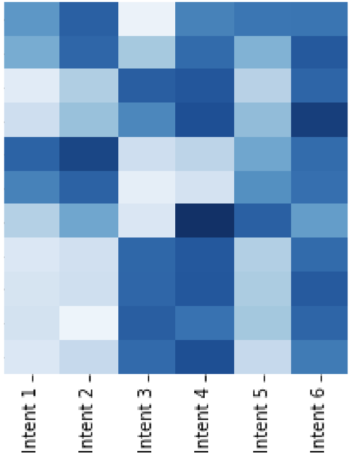
Understanding intents

Home

Considering intent and learning across intents improves ability to infer user satisfaction by 20%.

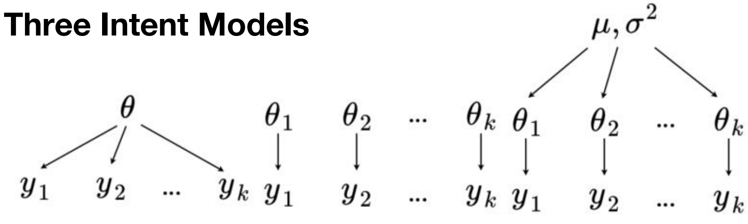
Passively listening
- quickly access playlists or saved m
- play music matching mood or acti
- find music to play in background

Actively engaging
- discover new music to listen to nov
- save new music or follow new play
- explore artists or albums more dee



intent important to
interpret
user interaction

Three Intent Models



(a) Global Model (b) Per-Intent Model (c) Multi-level Model

R Mehrotra, M Lalmas, D Kenney, T Lim-Meng & G Hashemian. **Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations.** WWW 2019.

Understanding intent is hard

Important to consider user intent to predict satisfaction, define optimization metric or interpret a metric.

user research

quantitative research

common intent

intent model

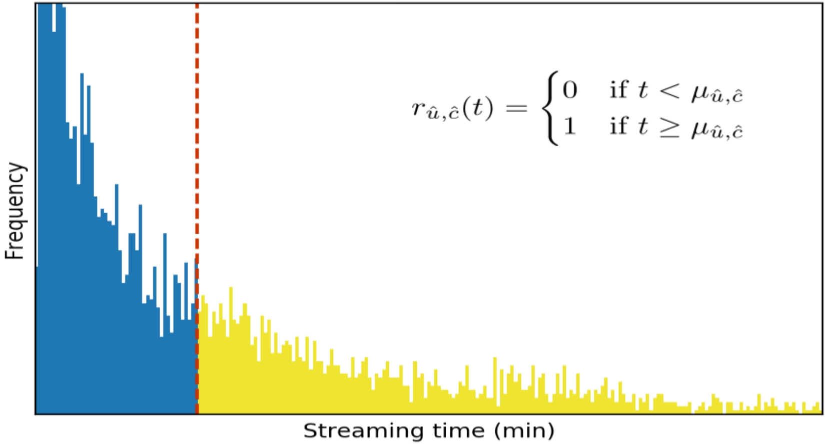
intent-aware optimisation

N Su, J He, Y Liu, M Zhang & S Ma. **User Intent, Behaviour, and Perceived Satisfaction in Product Search.** WSDM 2018.
J Cheng, C Lo & J Leskovec. **Predicting Intent Using Activity Logs: How Goal Specificity and Temporal Range Affect User Behavior.** WWW 2017.

Optimizing for the right metric

Home

Using playlist consumption time to inform metric to optimise for Spotify Home reward function



mean of consumption time

co-clustering user group x playlist type

Optimizing for mean consumption time led to +22.24% in predicted stream rate. Defining per user x playlist cluster led to further +13%.

P Dragone, R Mehrotra & M Lalmas. **Deriving User- and Content-specific Rewards for Contextual Bandits.** WWW 2019.

Choosing metric is important

Recommenders will be very good at optimizing for the chosen metric.

qualitative research

correlation vs causation

involvement

interaction

contribution

X Yi, L Hong, E Zhong, N Nan Liu & S Rajan. **Beyond clicks: dwell time for personalization**. RecSys 2014.
M Lalmas, H O'Brien & E Yom-Tov. **Measuring user engagement**. Morgan & Claypool Publishers, 2014.
J Lehmann, M Lalmas, E Yom-Tov and G Dupret. **Models of User Engagement.** UMAP 2012.

Acting on segmentation

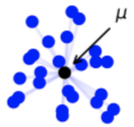
Listening diversity

A Anderson, L Maystre, R Mehrotra, I Anderson and MLalmas. **Algorithmic Effects on the Diversity of Consumption on Spotify**. WWW 2020.

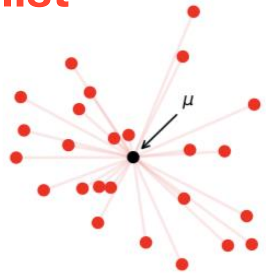
By segmenting users into specialist vs generalists, we observed different retention and conversion behaviors.

Measure of user listening diversity $$f(C) = \frac{1}{|C|} \sum_{c \in C} w_c cos(c, \mu)$$

specialist



generalist



Listening diversity = number of genres liked in past x months
Like a genre = have affinity for at least y artists in that genre

Optimizing for segmentation

Segmentation helps recommenders to perform for users and contents across the spectrum.

who?

what?

when?

where?

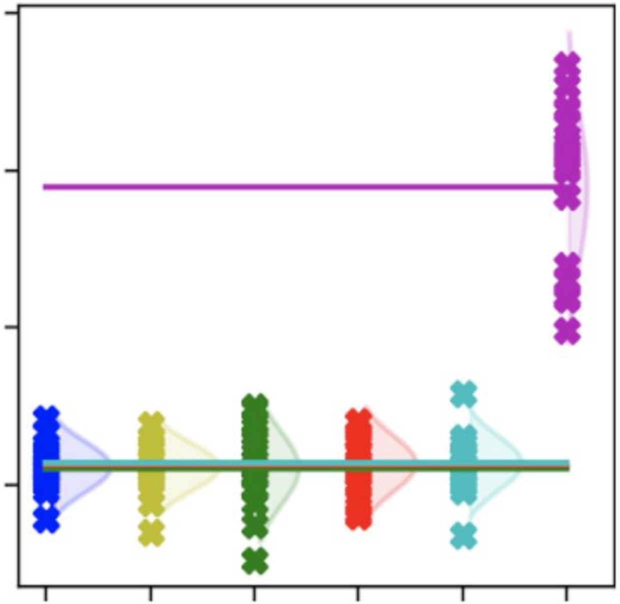
why?

Y Jinyun, W Chu & R White. **Cohort modeling for enhanced personalized search**. SIGIR 2014
S Goel, A Broder, E Gabrilovich & B Pang. **Anatomy of the long tail: ordinary people with extraordinary tastes**. WSDM 2010.
R White, S Dumais & J Teevan. **Characterizing the influence of domain expertise on web search behavior**. WSDM 2009.

Thinking about diversity

Satisfaction

Optimizing for multiple satisfaction objectives together performs better than single metric optimization.



Satisfaction metrics include clicks, stream time, number of song played, etc.

Model is learning more relevant patterns of user satisfaction with more optimization metrics.

Paper under review.

Understanding diversity

When thinking diversity, recommenders become informed about what and who they serve.

algorithmic bias

data bias

under-index

over-index

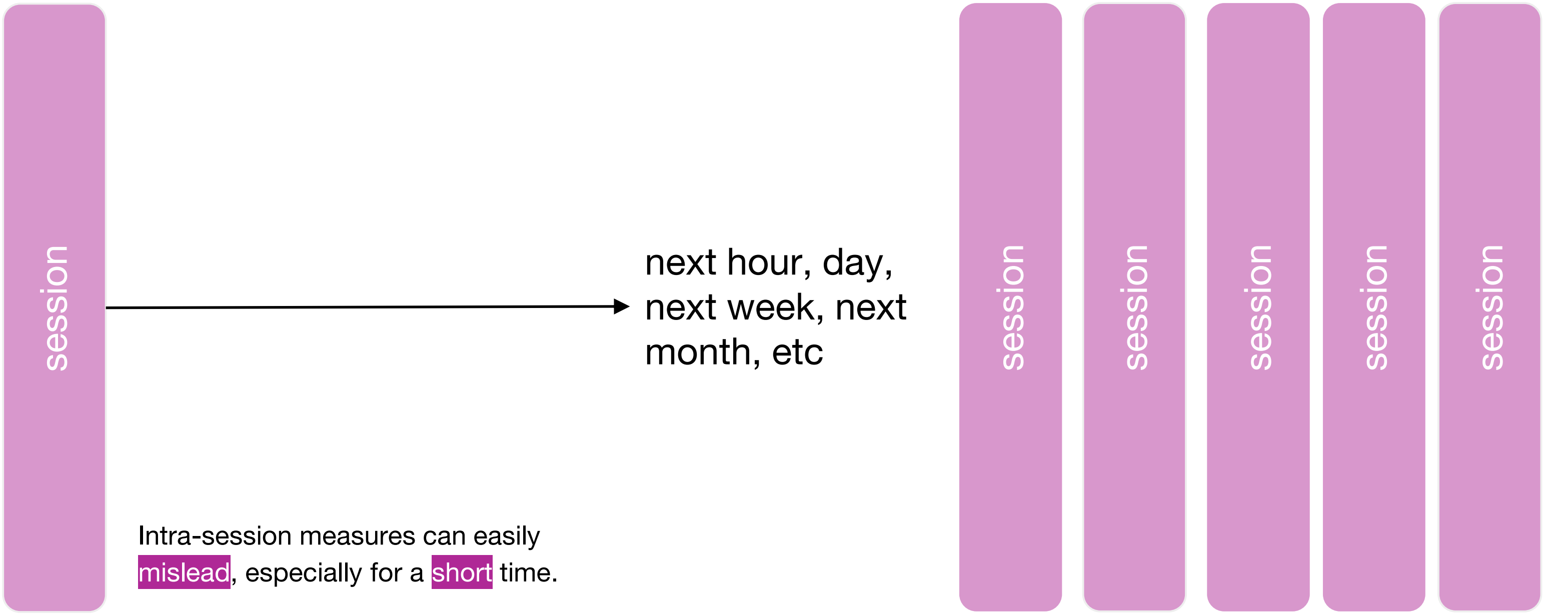
awareness

H Cramer, J Wortman-Vaughan, K Holstein, H Wallach, H Daume, M Dudík, S Reddy & J Garcia-Gathright. **Algorithmic bias in practice**. FAT* Industry Translation Tutorial, 2019.
P Shah, A Soni & T Chevalier. **Online Ranking with Constraints: A Primal-Dual Algorithm and Applications to Web Traffic-Shaping**. KDD 2017.
D Agarwal & S Chatterjee. **Constrained optimization for homepage relevance**. WWW 2015.

One more thing

Intra-session engagement measures user engagement during the session.

Inter-session engagement measures user engagement across sessions and relates to KPIs and business metrics.



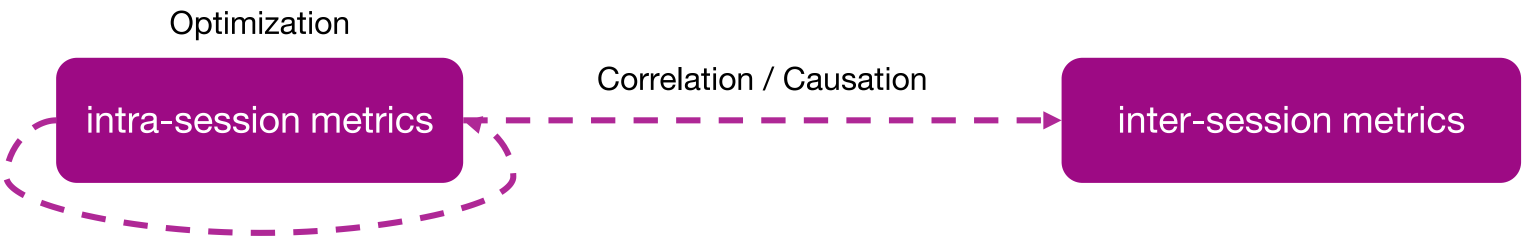
Intra-session measures can easily mislead, especially for a short time.

R Kohavi, A Deng, B Frasca, R Longbotham, T Walker & Y Xu. **Trustworthy online controlled experiments: Five puzzling outcomes explained.** KDD 2012.

We do not optimize well (yet) for inter-session metrics

- Do not capture how users engage during a session
- May not deliver much, if any, in terms of improving recommenders
- Not clear yet how recommenders can learn from using inter-session metrics

We optimize (and monitor) intra-session metrics … but those that **move** inter-session metrics.



Hong & M Lalmas. **Tutorial on Online User Engagement: Metrics and Optimization**. WWW 2019.
H Hohnhold, D O’Brien & D Tang. **Focusing on the Long-term: It’s Good for Users and Business**. KDD 2015.
G Dupret & M Lalmas. **Absence time and user engagement: Evaluating Ranking Functions**. WSDM 2013.

We do not optimize well (yet) for inter-session metrics

- Do not capture how users engage during a session
- May not deliver much, if any, in terms of improving recommenders
- Not clear yet how recommenders learn from using inter-session metrics

We optimize (and monitor) intra-session metrics… but those that move inter-session metrics.

Can research in state-based user modelling help here?

Can we learn metrics using state-aware ML?

Hong & M Lalmas. **Tutorial on Online User Engagement: Metrics and Optimization**. WWW 2019.
H Hohnhold, D O'Brien & D Tang. **Focusing on the Long-term: It's Good for Users and Business**. KDD 2015.
G Dupret & M Lalmas. **Absence time and user engagement: Evaluating Ranking Functions**. WSDM 2013.

Let us recap

.... to inform state-based user modeling

User intents help informing metric optimization & metric interpretation.

Segmentation helps adapting recommender models.

Diversity, intents, segmentation help understanding the value of our recommendations.

Understanding intents

Optimizing for the right metric

Acting on segmentation

Thinking about diversity

Thank you