From Devices to People: Attribution of Search Activity in Multi-User Settings

Ryen White, Ahmed Hassan, Adish Singla, Eric Horvitz
Microsoft Research, USA; ETHZ, Switzerland

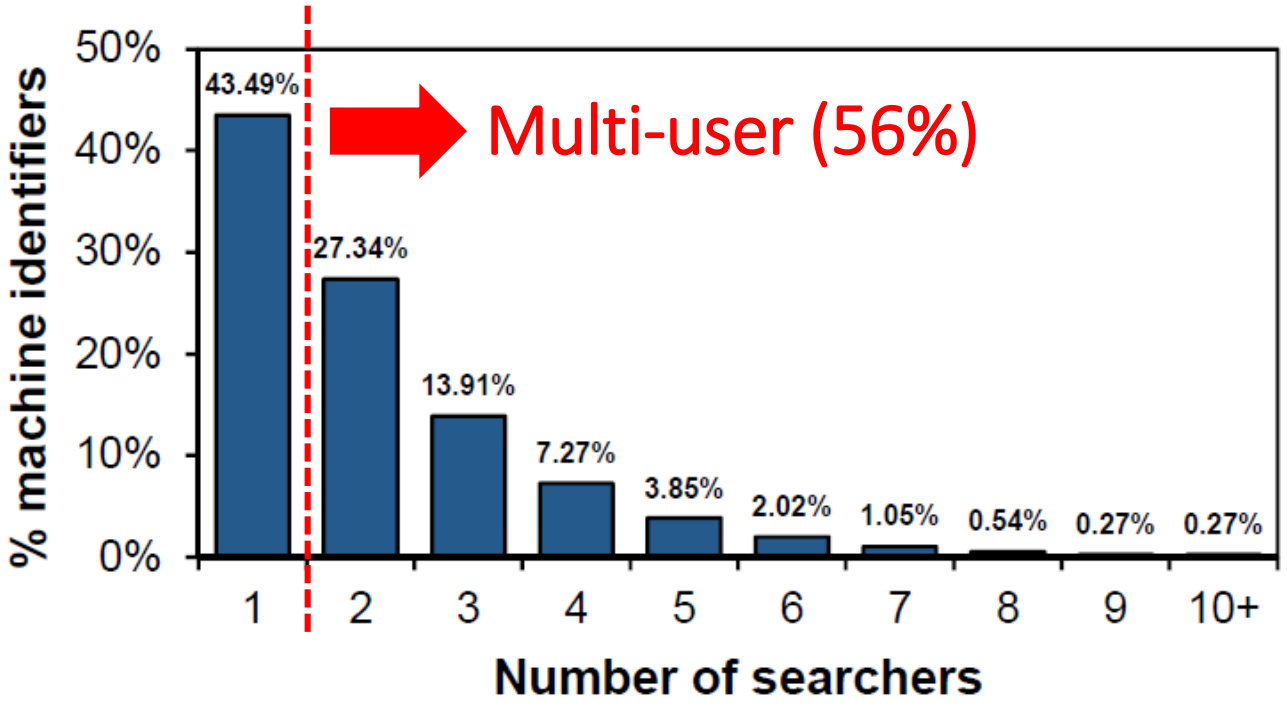
Contact: ryenw@microsoft.com

IDs in Behavioral Analysis @ Scale

- Search engines use machine ids based on cookies etc.
- **Assume 1:1 mapping** from ids (e.g., FDED432F901D) to people
- However, multi-user computer usage is common
- 2011 Census data: 75% of U.S. households have a computer
  - In most homes that machine is shared between multiple people

Multi-User Web Search

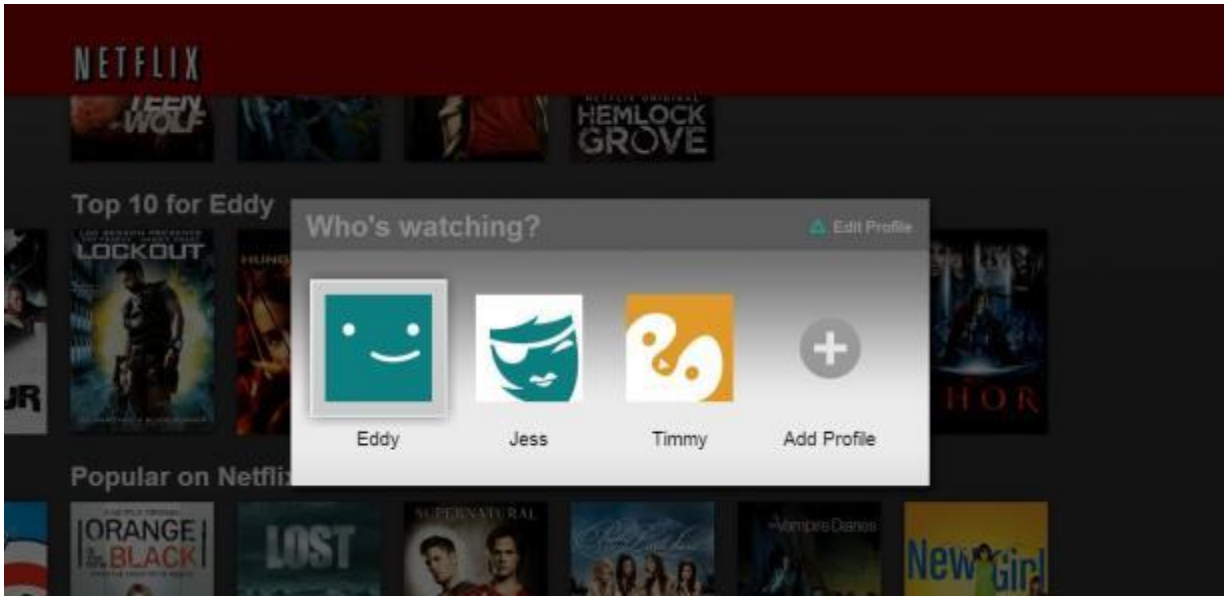
- Analyzed two years' of comScore search data (all engines, en-US)
- Both machine identifiers *and* person identifiers (users self-identify)



- **Takeaway:** 56% of machine ids comprise multi-user behavior

Handling Multiple Users

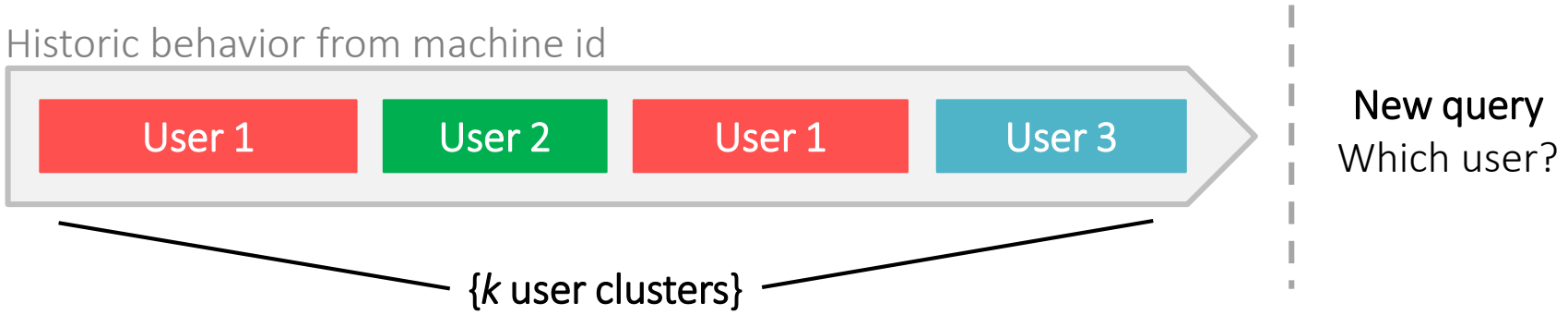
- Limited current solutions in search engines (users can sign-in)
- Some solutions in other domains, e.g., streaming media
- Users can be asked to confirm identify (cumbersome), e.g.,



Our Focus: Can we do this automatically? (in context of search)

Activity Attribution Challenge

Given a stream of data from a machine identifier, attribute observed **historic** and **new** behavior to the correct person



Applications for: personalization, advertising, privacy protection

Related work in signal processing and fraud detection—hardly any related work in user behavior analysis

Research on “individual differences” in search activity is relevant

Three parts to analysis

1. **Characterizing** differences in behavior from a machine given the presence of one user versus multiple users

2. **Predicting**:
- Presence of multiple users (1 vs. *N* problem) (Classification task)
- Estimating the number of users on a machine (Regression task)

3. **Associating** behavior to the correct user (via clustering in our case)
- e.g., New query arrives, which user issued that query?

Focus on characteristics and prediction in this presentation

comScore Search Log Data

- Two years of data (2011-2013)
- Purchased data from comScore (non proprietary)
- Summary statistics:

Statistic	*Value*
Total number of queries	576,470,390
Total number of machines	1,748,425
Total number of searchers	3,836,037
Average queries / machine	328.89 (stdev=1279.80)
Average duration (in days) / machine	126.07 (stdev=171.29)

- Person information per machine is ground truth

Characterization

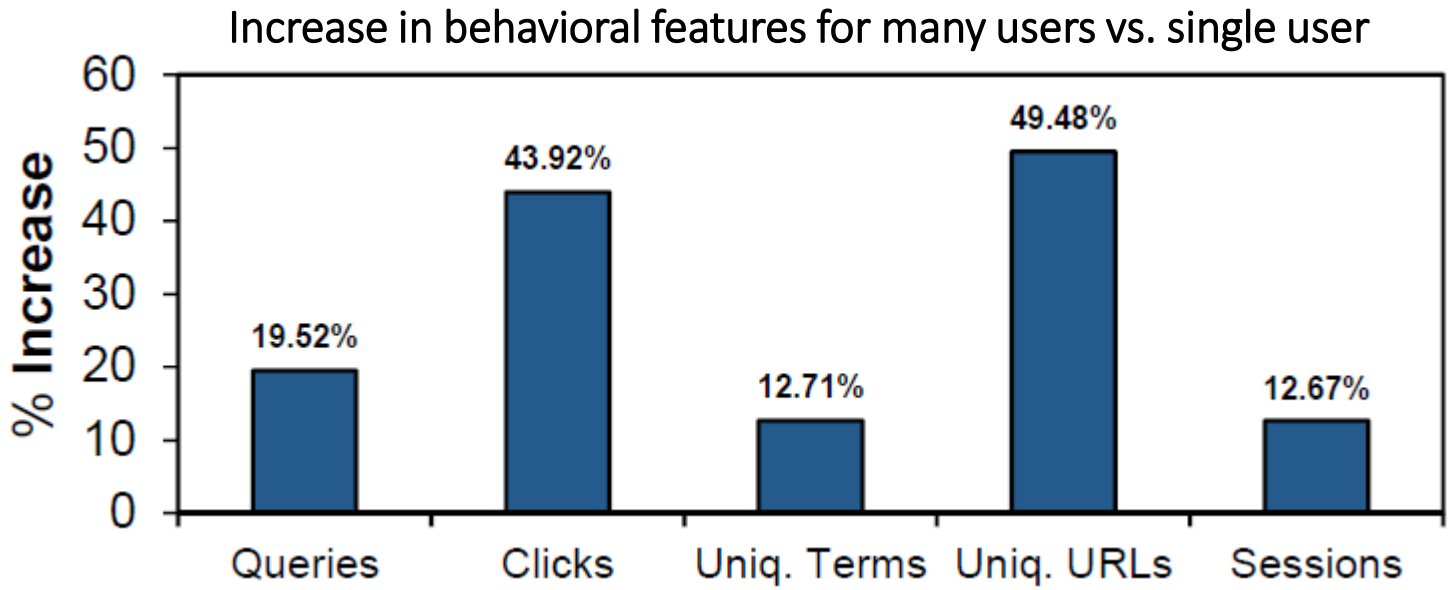
Are there within-id behavioral differences for one user vs. many?

Characterization

Characterize behavior observed from a machine identifier along a number of different dimensions:

- **Behavioral:** # Queries, # Clicks, # Unique Query Terms, etc.
- **Temporal:** Times machine used, Variations in time (hour, day)
- **Topical:** Types of topics, Variation in topics of queries/clicks
- **Content:** Nature of results viewed, inc. readability level

Behavioral Features



- Much more search behavior when there are multiple users
- More searching and clicking, and diversity in queries/clicks
- **BUT** some of this also applies to active searchers …

Temporal Features

Variance in time at which searches are issued, specifically:

- Day of week entropy
- Time of day entropy

Large differences with varying numbers of searchers associated with searching on the machine

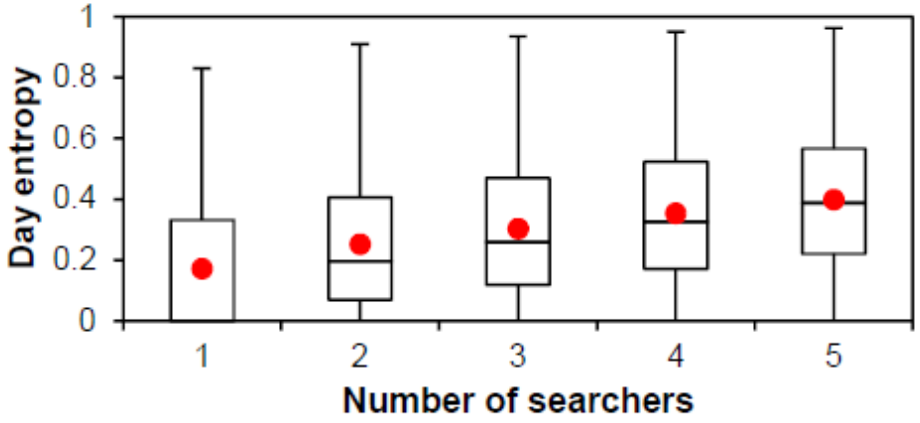


Figure 3. Box-and-whisker plot for day entropy for machines with diff. # searchers. Mean is dot. Median is horizontal line.

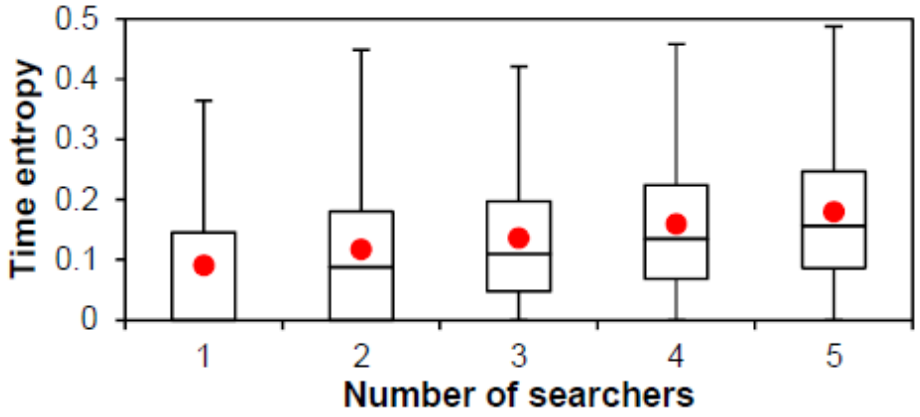


Figure 4. Box-and-whisker plot for time entropy for machines with diff. # searchers. Mean is dot. Median is horizontal line.

Topical/Content Features

Observed similar variations in entropy for topics and the readability of content

Topic pair (Ti, Tj) in 4-hr bucket

Topic association:

$$NPMI(T_i, T_j) = -\log \frac{p(T_i, T_j)}{p(T_i)p(T_j)} \Big/ -\log p(T_i, T_j)$$

Multi-searcher machines overestimate topic associations for 90% of pairs

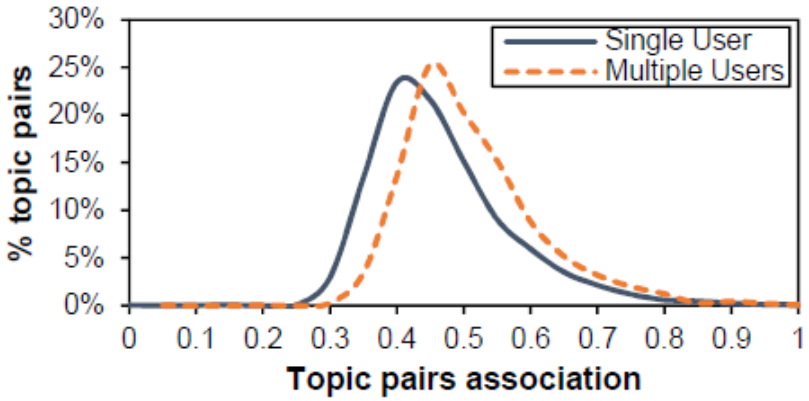


Figure 7. Distribution of topic pairs association from single-searcher machines (true dist.) and multi-searcher machines.

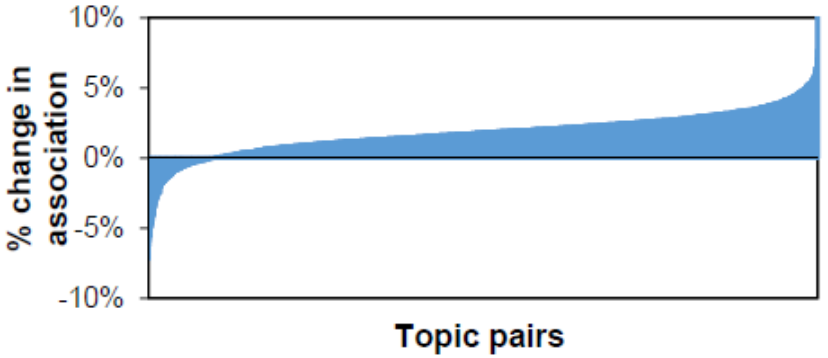


Figure 8. % change (error) in topic association for multi-searcher vs. single-searcher (truth). Positive change shows multi-searcher machines overestimate truth for 90% of pairs.

Prediction

Can we predict multi-user ids?

Prediction

- Two prediction tasks:

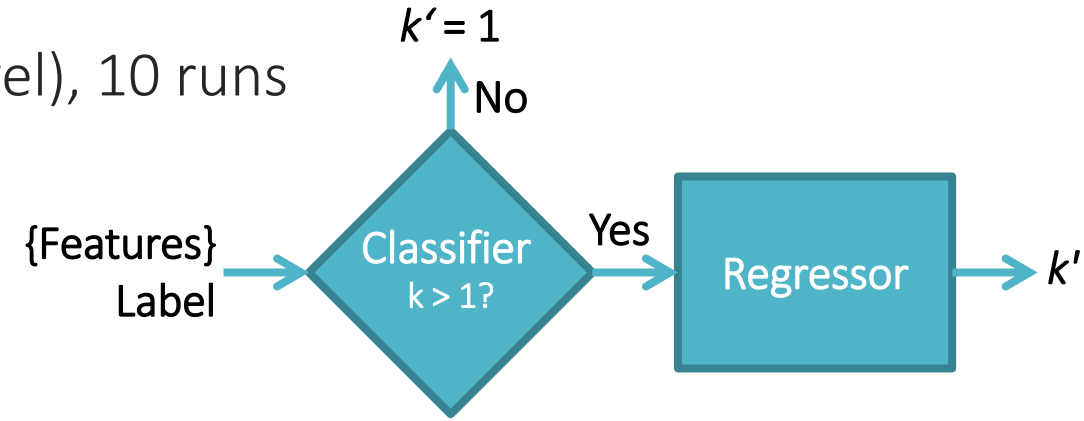
1. **Classification task**

Question: *Is a machine identifier composed of multiple people?*

2. **Regression task**

Question: *If multiple users behind machine identifier, how many?*

- MART classification and regression
- Use all features described so far
- 10-fold CV (at user level), 10 runs
- Can chain models:



Features and Labels

- Features from the characterization:
  - Behavioral, Temporal, Topical, and Content
- Plus **Referential**
  - Indications that there is likely to be another member of household, e.g., reference to spouse, child, roommate, etc. in queries

- Labels:
  - **Classification**: Multi-user (1) vs. single user (0)
  - **Regression**: Number of users associated with machine id

Classification: Results

Table 3. Classification performance for each classifier, ordered by classification accuracy. All differences significant using *t*-tests at $p < 0.001$ for accuracy and AUC for each classifier versus marginal and versus *All*.

Features	*Accuracy*	*Pos. Prec.*	*Pos. Recall*	*Neg. Prec.*	*Neg. Recall*	*AUC*
All	0.8635	0.8662	0.8973	0.8597	0.8196	0.9366
Temporal	0.8552	0.8531	0.8986	0.8582	0.7986	0.9267
Topical	0.8324	0.8399	0.8694	0.8218	0.7824	0.9105
Content	0.8271	0.8351	0.8651	0.8157	0.7776	0.9055
Behavioral	0.8096	0.8027	0.8795	0.8208	0.7185	0.8827
Referential	0.6450	0.8751	0.4342	0.5552	0.9193	0.6871
Marginal	0.5651	0.5651	1.0000	0.0000	0.0000	0.5000

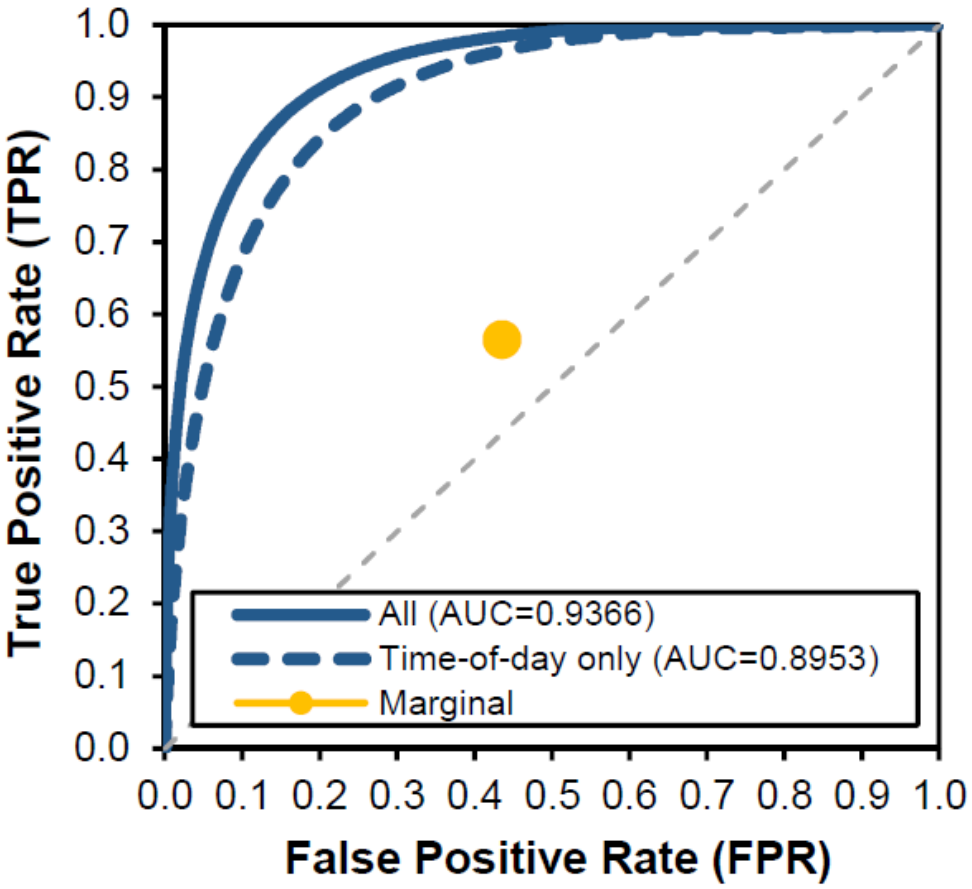
Temporal features appear important for this task

Classification: Time of Day ONLY

Variant that uses eight features that are only associated with the time of day only (e.g., hour bucket, bucket entropy)

Perf. similar to full model

Simpler to implement than the complete range of features highlighted earlier – **8 features vs 80!**



Prediction: Regression

- Same features as for classification, different label

Table 4. Regression performance for each of the feature classes, ordered by MAE. All differences significant using *t*-tests at $p < 0.001$ for MAE and RMSE for each classifier versus baselines and versus *All*.

Features	*MAE*	*NRMSE*
All	0.6377	0.0917
Topical	0.6906	0.1055
Temporal	0.7232	0.1146
Content	0.7490	0.1150
Behavioral	0.8054	0.1204
Referential	0.9784	0.1325
Marginal	1.4799	0.2150
Random	3.8078	0.4652

$NRMSE = RMSE/(k_{max} - k_{min})$

- Time-of-day features not as useful here (*NRMSE*=0.1300)

Top Features

Table 5. Top five features by evidential weight for the classification and the regression tasks. Feature weights and correlation coefficients for features vs. labels are also shown. Weights are normalized w.r.t. the highest-weighted feature.

	Feature	*Class*	*Weight*	*r*
Classifier	NumTimeBuckets	Temporal	1.0000	+0.180
	FractionWeekday	Temporal	0.6353	+0.444
	FractionQueries_KidsAndTeens	Topical	0.6031	+0.159
	TimeEntropy	Temporal	0.4306	+0.233
	FractionReferenceOtherPerson	Referential	0.3412	+0.149
Regressor	FractionQueries_KidsAndTeens	Topical	1.0000	+0.271
	PageReadability	Content	0.6108	+0.395
	FractionQueries_ShoppingChildren	Topical	0.5550	+0.209
	FractionReferenceOtherPerson	Referential	0.5496	+0.199
	TopicEntropy	Topical	0.3797	+0.143

- Need the additional features for regression task
  - Features linked to children's interests are important
  - Where there is a child, there is at least one adult (=> N ≥ 2)

Assignment

Can we assign to correct user?

Assignment

- Given the *k'* from the regressor, run *k*-means clustering on the history from each machine identifier
- Real-time assignment – given a user session:
  - Compare 1st query in session to cluster(s), assign to most similar
  - Compute similarity between session/cluster representative

Accuracy = proportion of assignments correct

Purity = proportion of assigned cluster to correct user

Baseline = one user

Table 9. Average accuracy and purity of assignment. Baseline assumes unique mapping of machine identifier to searcher. All differences significant with *t*-tests at $p < 0.001$.

k	*Accuracy*	*Purity of Assignment*	*Baseline*
All (2–10)	0.742 (+56%)	0.659 (+39%)	0.475
2	0.771 (+51%)	0.700 (+37%)	0.512
3	0.649 (+89%)	0.512 (+50%)	0.343
4	0.531 (+102%)	0.395 (+50%)	0.263
5	0.451 (+116%)	0.333 (+60%)	0.208
6–10	0.361 (+96%)	0.289 (+57%)	0.184

Discussion

- comScore data based on self-identification (Errors? Not apparent)

- Need to explore:

  - Utility of sign-in to search engines as proxy for person identifier
    - *e.g.,* 

  - Different analysis timeframes (e.g., one month vs. two years)

Conclusions and Future Work

- Introduced activity attribution challenge
- Clear differences in logged behavior for one user vs. many
- Possible to accurately:
  1. Predict if multiple users are behind a machine id (AUC = 0.94)
  2. Estimate number of users behind machine id (NRMSE = 0.092)
  3. Assign queries to people (75% accuracy, 56% gain over baseline)
- **Future work**: Apply methods to personalization, advertising, etc.