Modeling Long-Term Search Engine Usage

Ryen White, Ashish Kapoor & Susan Dumais
Microsoft Research

Key Problem

- What are key trends in search engine usage?
  - Identify long-term patterns of usage
  - Understand key variables that affect behavior
- Can we predict long-term search engine usage?
  - Determine indicators that are predictive of trends

Prior Work

- Short-term Usage:
  - Predict Switch within Sessions
    (Heath & White 2008, Laxman et al. 2008, White & Dumais 2009)
  - Predict good search engines for a query
    (White et al. 2008)

- Economic / Conceptual Models
  - Identify factors influencing search engine choice
    (Capraro et al. 2003)
  - Models of satisfaction
    (Keaveney et al. 2001, Mittal et al. 1998)

Long-Term Search Logs

- Six months of toolbar data (26 weeks)
  - Sep 2008 through February 2009
- Three search engines
  - Bing, Google and Yahoo
- Users with at least 10 queries every week
  - 10K users for our analysis
  - English speaking, located in US

Long-Term Search Logs (summarized for each week)

fractionEngine	Fraction of queries issued to search engine
queryCountEngine	Number of queries issued to search engine
avgEngineQueryLength	Average length (in words) of queries to search engine
fractionEngineSAT	Fraction of search engine queries that are satisfied
fractionNavEngine	Fraction search engine queries defined as navigational
fractionNavEngineSAT	Fraction of queries in fractionNavEngine that are satisfied

SAT score: Dwell time greater than equal to 30 seconds (Fox et al. 2005)

Outline

- Identifying Key Trends
- Indicators of User Behavior
- Predicting Search Engine Usage
- Conclusion and Future Work

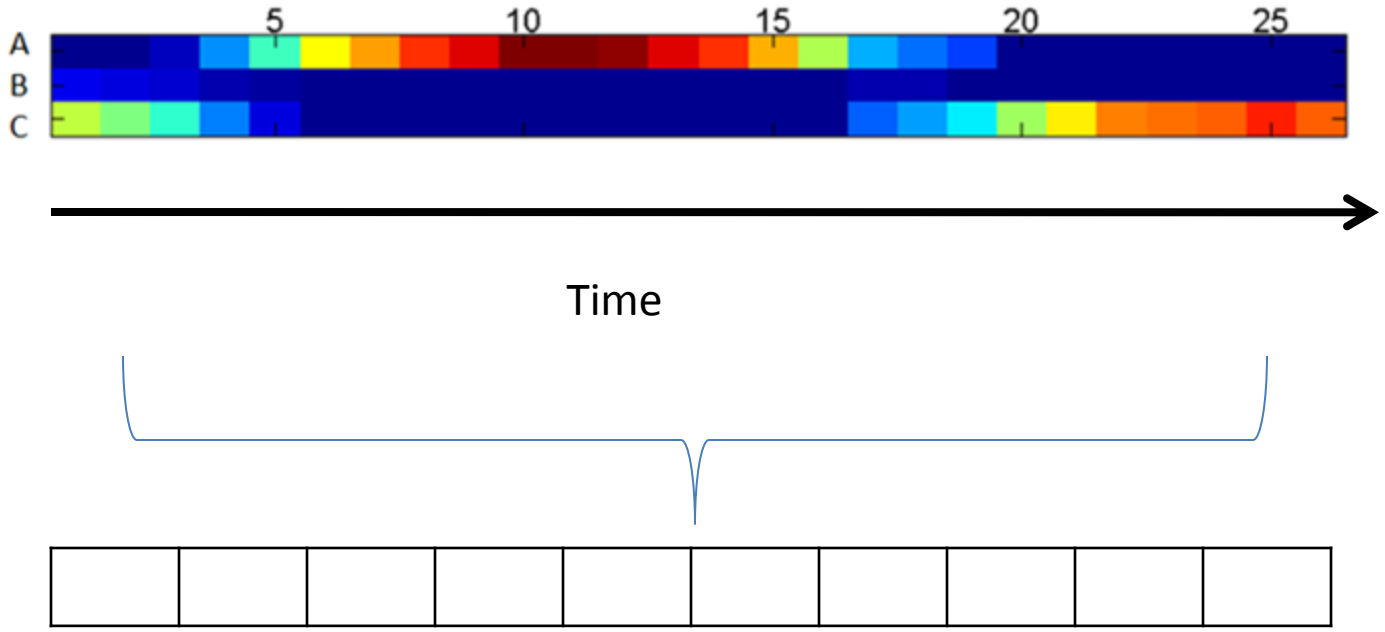
Outline

- Identifying Key Trends
- Indicators of User Behavior
- Predicting Search Engine Usage
- Conclusion and Future Work

Identifying Basis Behaviors

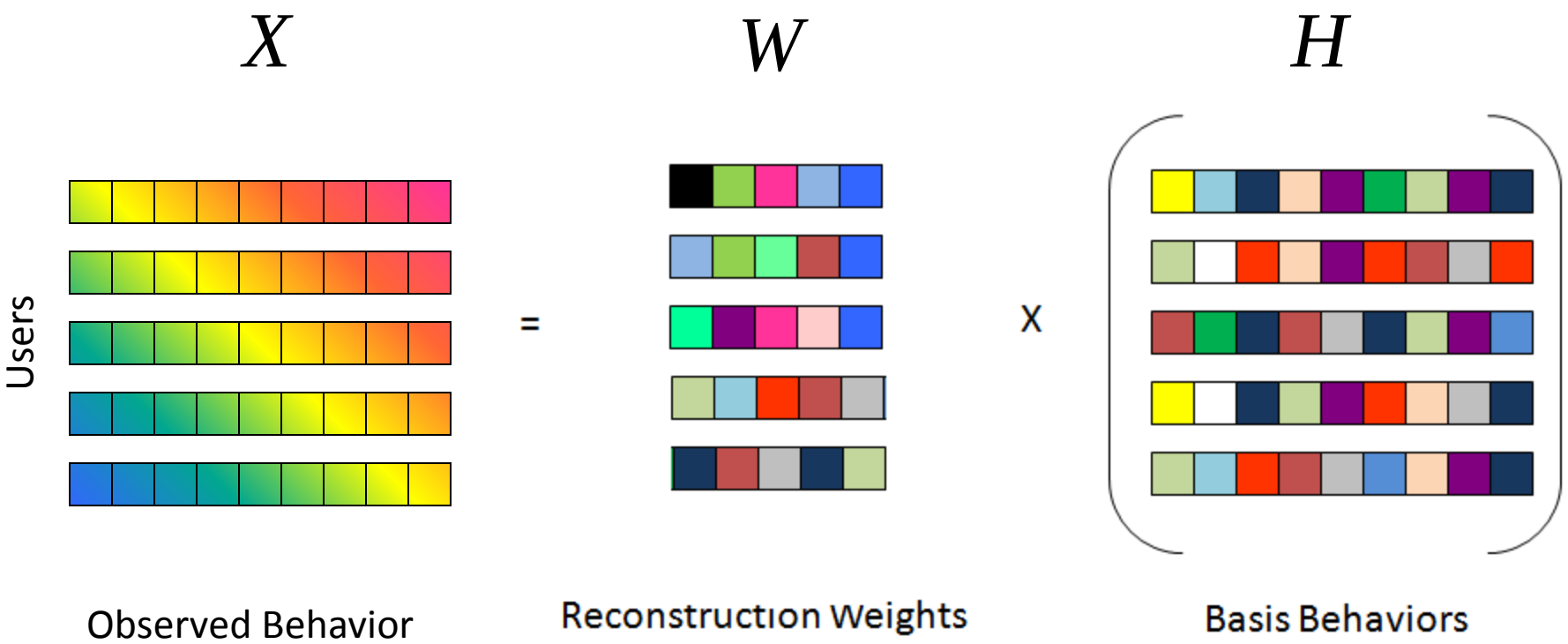
Primary Behavior Indicator: fractionEngine

Search engine

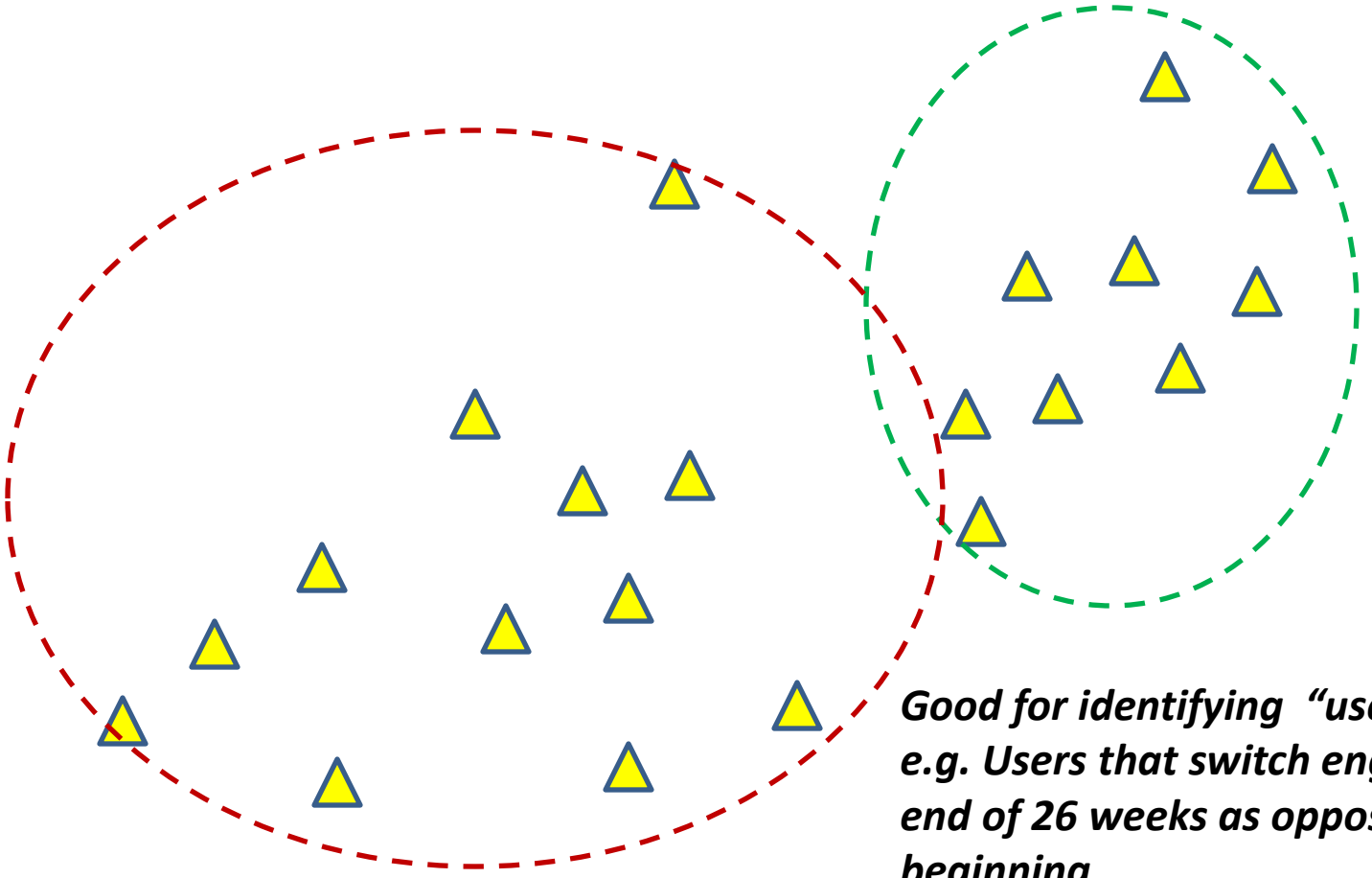


26 X 3 dimensional behavior vector (per user)

Identifying Basis Behaviors



Option 1: Clustering

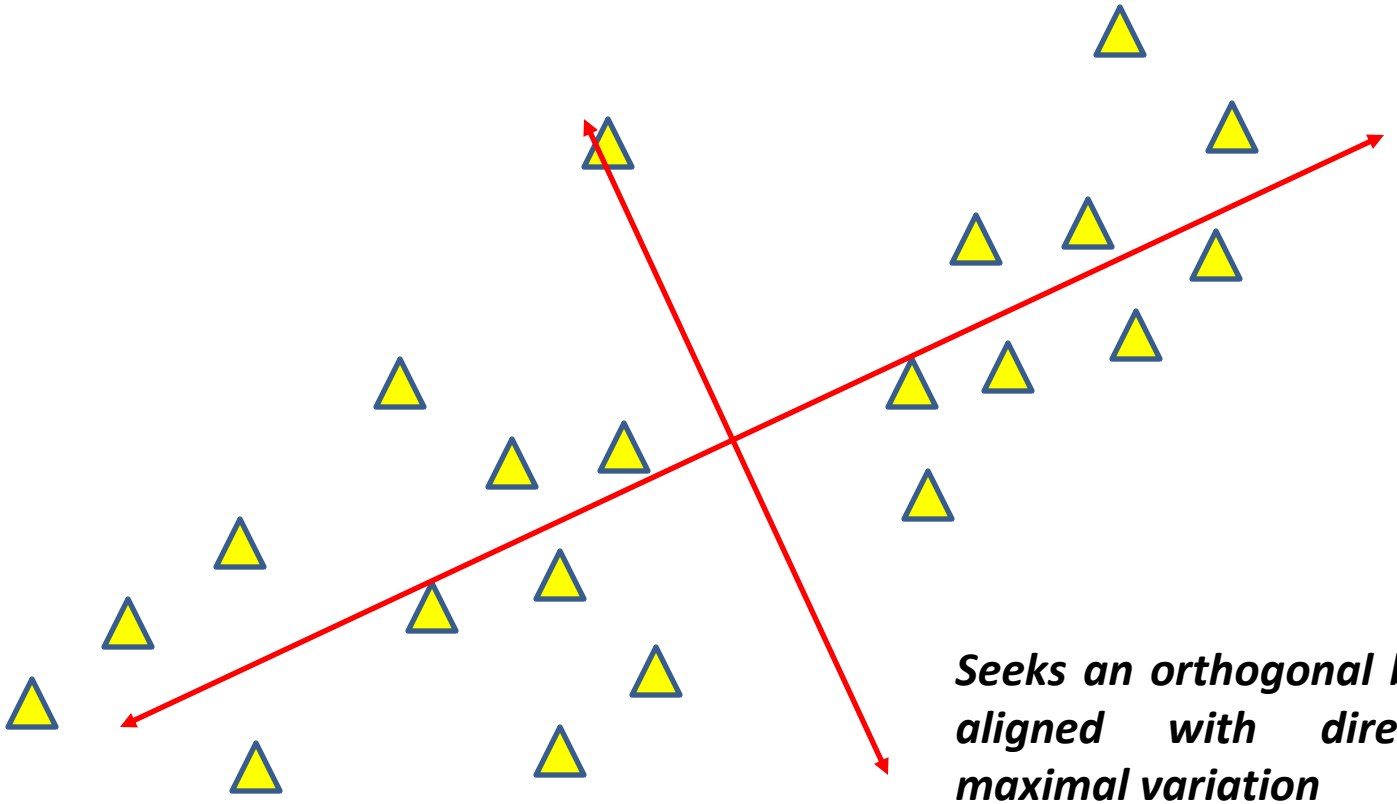


Good for identifying "user prototypes" e.g. Users that switch engines towards end of 26 weeks as opposed to the beginning

Might not recover basis behavior

corresponds to one user

Option 2: PCA a.k.a. Eigen Analysis

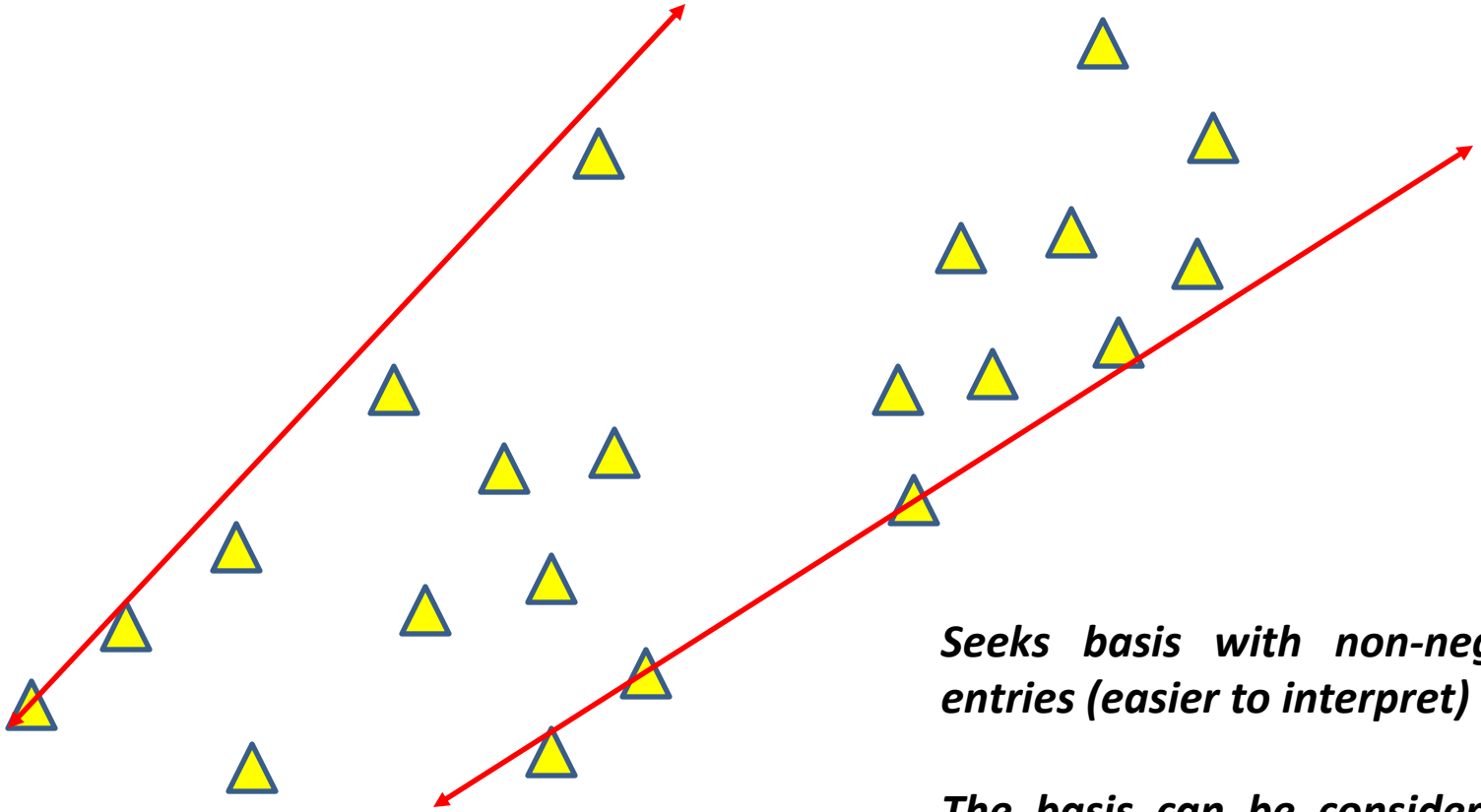


Seeks an orthogonal basis that's aligned with directions of maximal variation

Basis vectors are hard to interpret as the basis vectors will have negative values

corresponds to one user

Option 3: Non-negative matrix factorization



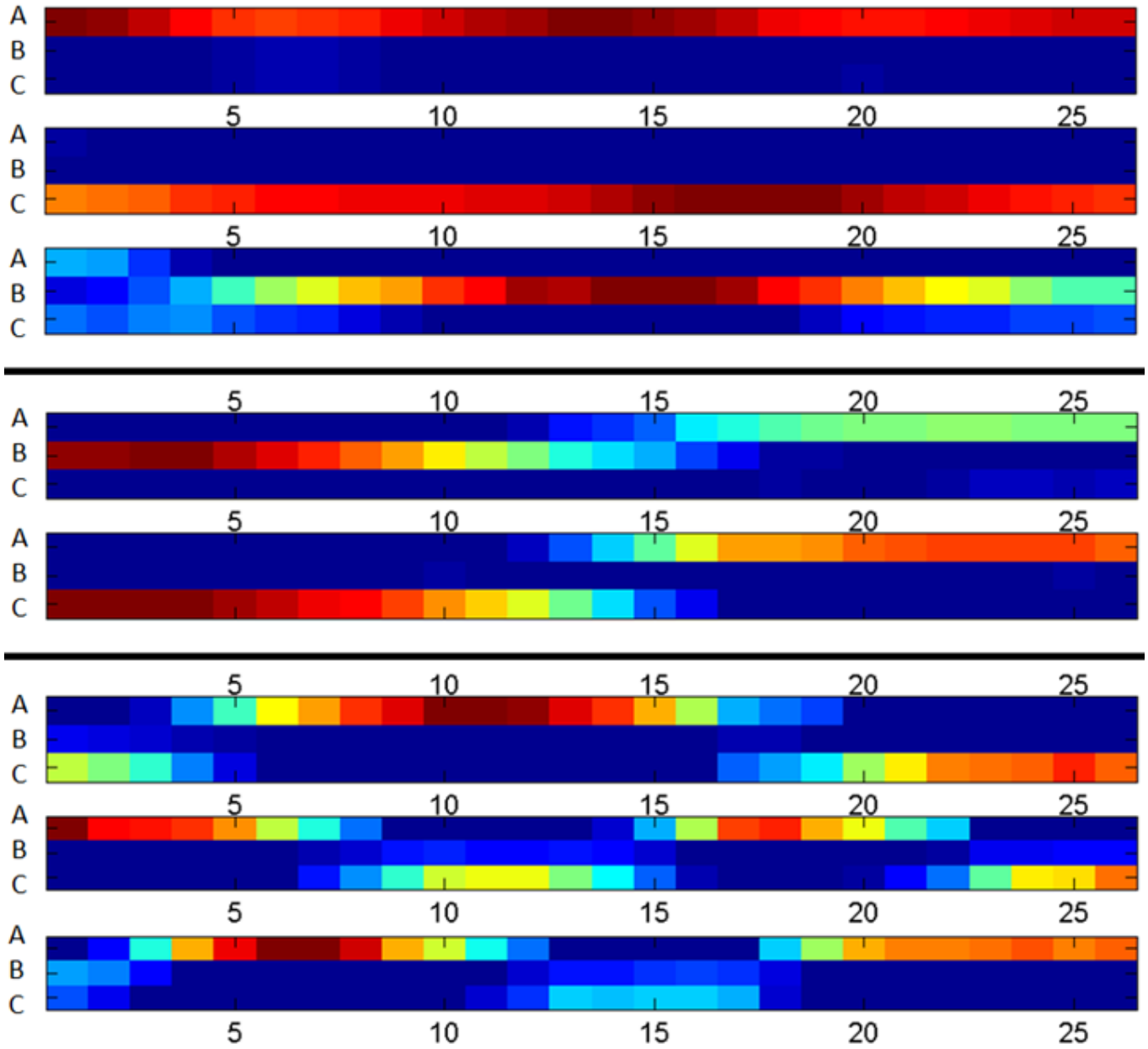
Seeks basis with non-negative entries (easier to interpret)

The basis can be considered as parts / building blocks

Numerically harder problem

△ corresponds to one user

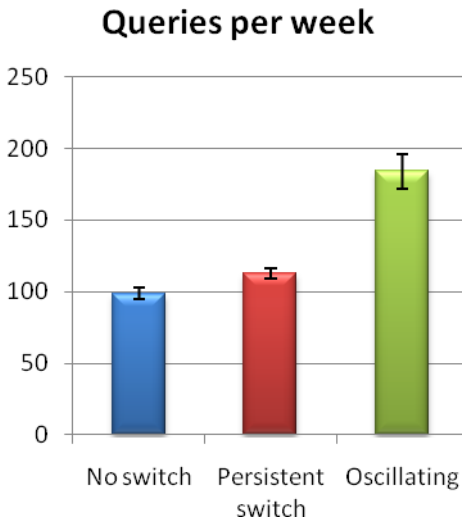
Key Trends in Long-Term Search Engine Usage



Outline

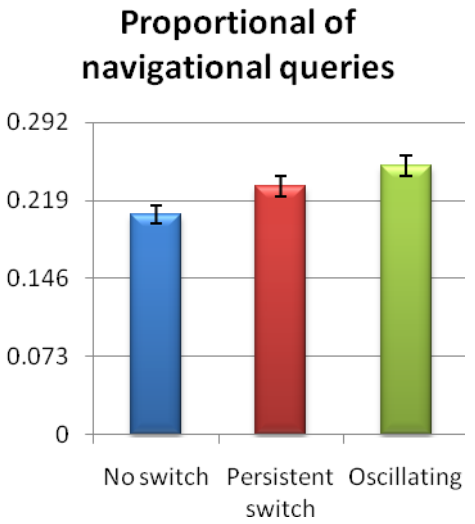
- Identifying Key Trends
- Indicators of User Behavior
- Predicting Search Engine Usage
- Conclusion and Future Work

What are key differentiating factors across the three groups?

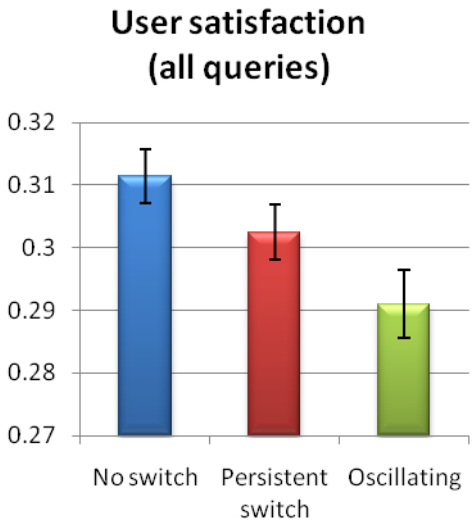


Users in oscillating group issue a significantly higher number of queries than the others

Oscillating == Skilled, aware of multiple search engines

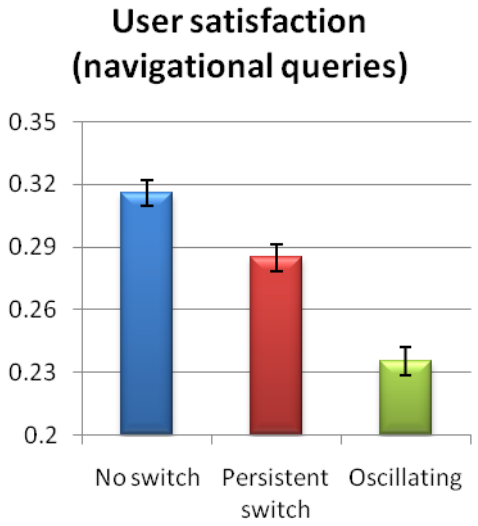


What are key differentiating factors across the three groups?

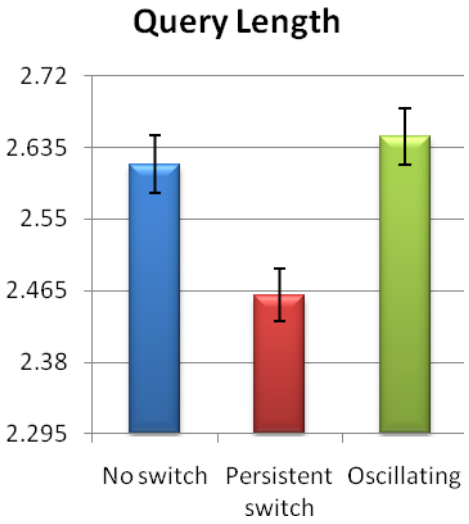


Users in oscillating group are hardest to please!

Low user satisfaction == Hard queries, more demanding in terms of required information



What are key differentiating factors across the three groups?



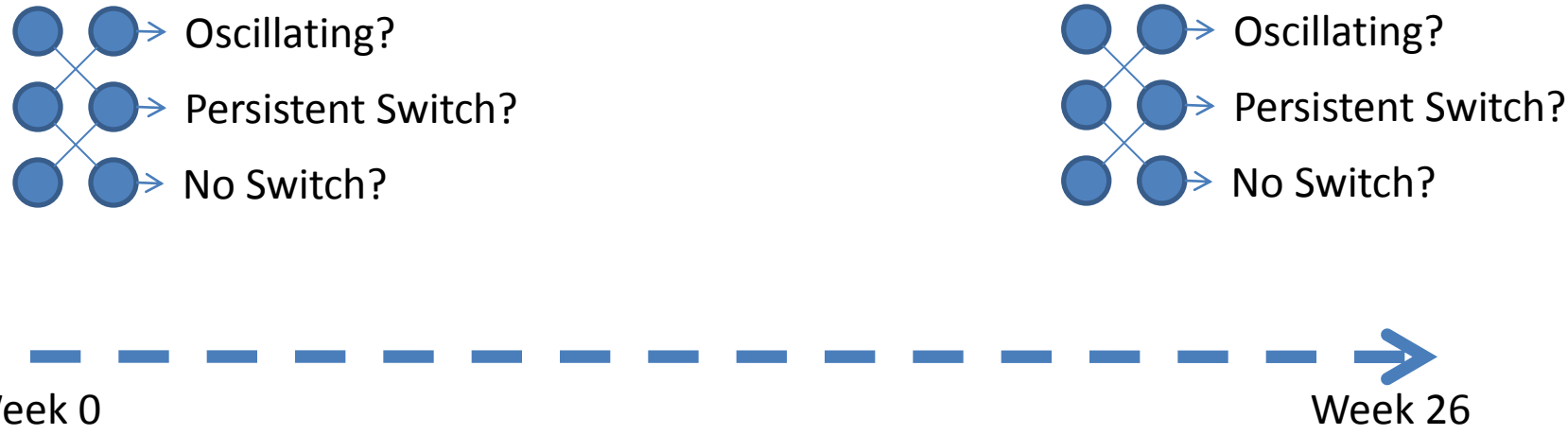
Users that make the persistent switch issue shortest (possibly simpler) queries.

Shorter / simpler queries == Non-expert population, less familiar with search engines

Outline

- Identifying Key Trends
- Indicators of User Behavior
- Predicting Search Engine Usage
- Conclusion and Future Work

Prediction Goal



Feature Extraction

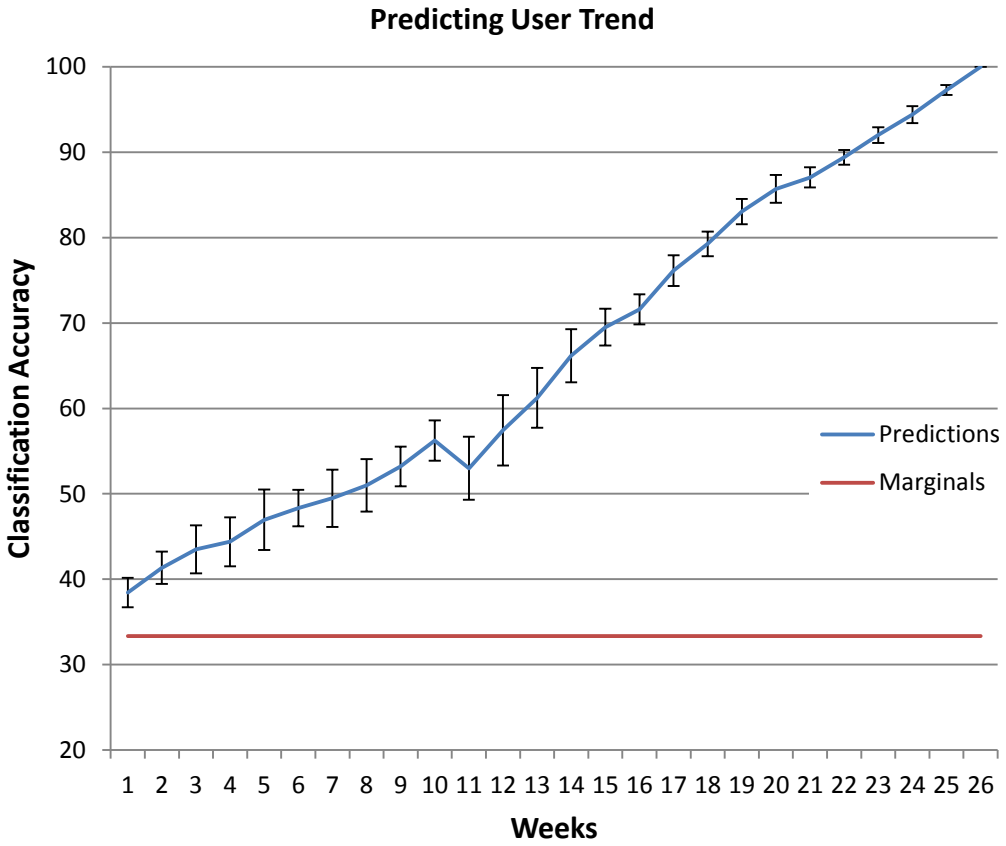
fractionEngine
queryCountEngine
avgEngineQueryLength
fractionEngineSAT
fractionNavEngine
fractionNavEngineSAT

F_1
F_2
F_3
F_4
.
.
.
F_K

Experimental Protocol

- Dataset
  - 500 user from each class (1500 total)
  - 50-50 train-test split
  - Results averaged over 10 random train-test splits
- Classifier
  - Gaussian process regression
  - Linear kernel
  - Classify users as number of weeks observed is varied

Can We Predict Search Engine Usage?



Gaussian Process Regression (Linear Kernel)

Most Informative Features

$$y = w^T \cdot x$$

No Switch vs. Rest	Pers Switch vs. Rest	Oscillate vs. Rest
isOneEngineDominant	min fractionEngine A	min fractionEngine C
min fractionEngine A	min fractionEngine C	isOneEngineDominant
ObservedPersistSwitch	min fractionEngine B	ObservedPersistSwitch
max fractionEngine A	max fractionEngine A	min fractionEngineSAT C
min fractionEngine B	max fractionEngine C	mean fractionEngineSAT A
mean fractionEngineSAT A	isOneEngineDominant	min fractionEngine B
mean fractionEngineA	max queryCountEngine C < 50	mean fractionEngineSAT B
min fractionNavEngine A	min fractionEngineSAT C	mean fractionEngineSAT C
mean fractionNavEngine A	mean fractionNavEngine A	max queryCountEngine B < 50
max fractionEngine C	ObservedPersistSwitch	min fractionEngineSAT B

Conclusion and Future Work

- Discovered 3 key trends in long term search engine usage
  - No Switch, Persistent Switch, Oscillating
- Possible to predict usage behaviors
  - Extract features about user satisfaction, past usage behavior

- In future:
  - Additional data / features (e.g. demographics?)
  - Can we dissuade users from making a persistent switch from our engine (if we detect it in advance)?

Questions?

{ryenw, akapoor, sdumais}@microsoft.com