User Behaviors on Web

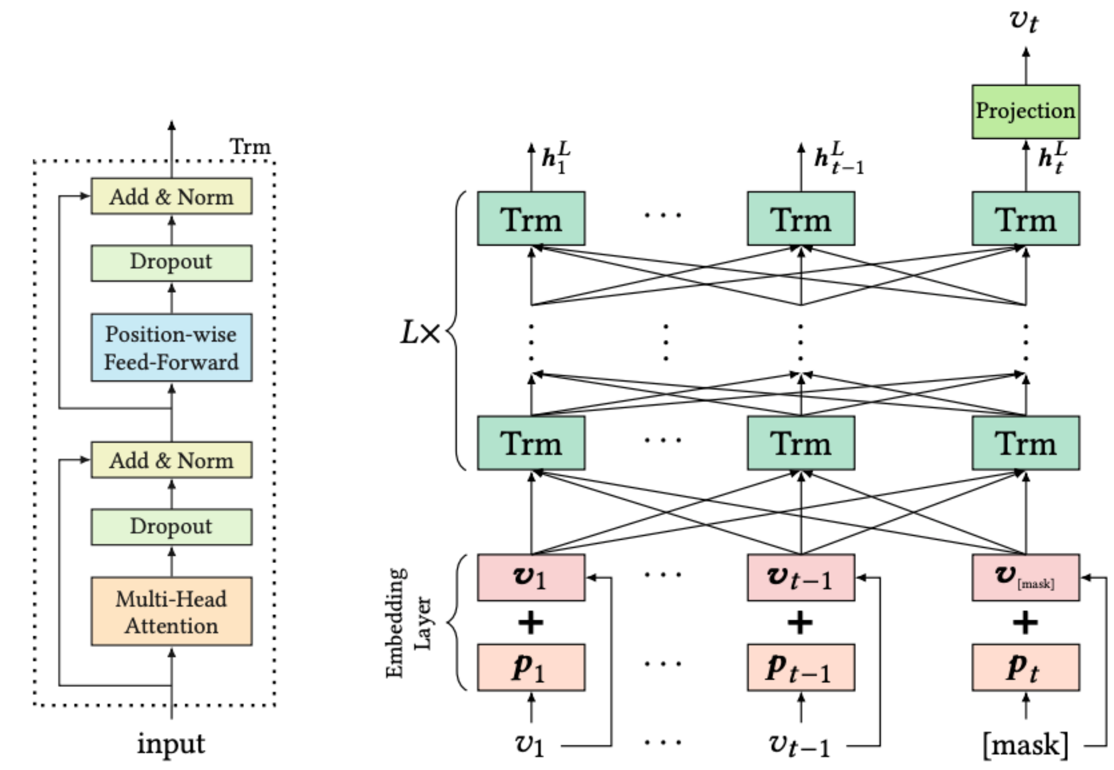
SUBTITLE/TAGLINE

User Behaviors on Web

- The better you understand me, the better you can serve me
- Basic task: infer the information need, intents, interests of users from their past behaviors, and then predict their future behavior (e.g., click given a query)
- Sample problems:
  - Identifying sessions in query logs
  - Predicting accesses to a given page (e.g., for caching)
  - Recognizing human vs. automated queries
  - Recommending alternative queries, landing pages, …

User Behaviors on Web

- Current state-of-the-art recommendation system



(a) Transformer Layer. (b) BERT4Rec model architecture.

Picture from Sun et al. 19

Query Log analysis

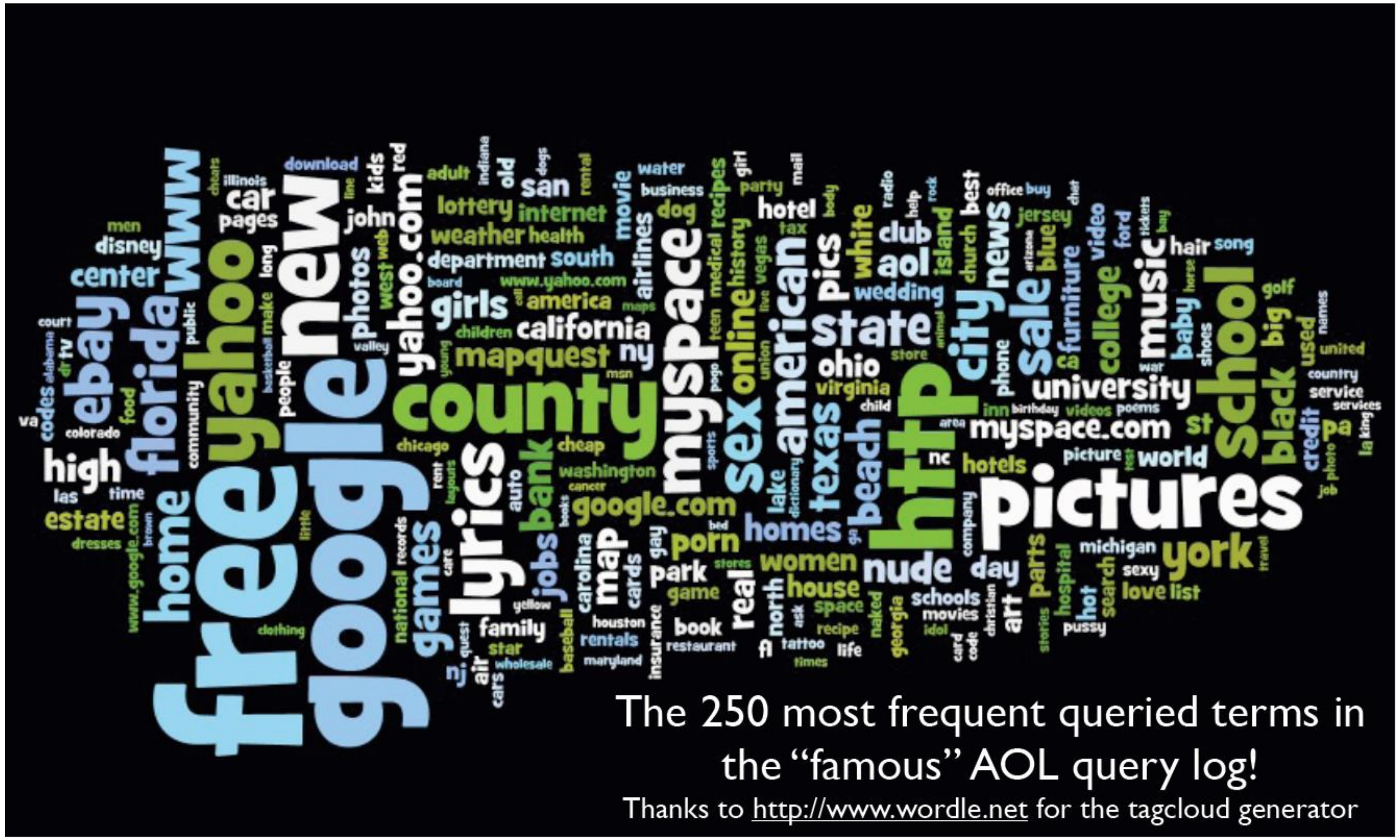
- Main idea: log the user behaviors/actions in web search
- Analyze the log to better understand the users

Query Log analysis

- Ma
- An

AnonID	Query	QueryTime	ItemRank	ClickURL
100218	tennessee department of transportation	2006-03-01 11:08:30	1	http://www.tdot.state.tn.us
100218	tennessee federal court	2006-03-01 11:53:44	1	http://www.constructionweblinks.com
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	2	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	dixie youth softball	2006-03-02 10:36:48	2	http://www.dixie.org
100218	cdwg	2006-03-03 14:29:07	1	http://www.cdwg.com
100218	cdwg scam cdwge	2006-03-03 14:30:11		
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	2	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	pensacola police department	2006-03-07 09:34:28	1	http://www.pensacolapolice.com
100218	memphis pd	2006-03-07 09:42:33	1	http://www.memphispolice.org
100218	nashville metro pd	2006-03-07 09:44:43	1	http://www.police.nashville.org
100218	florida highway patrol	2006-03-07 09:48:35	1	http://www.fhp.state.fl.us
100218	tennessee highway patrol	2006-03-07 09:49:52	1	http://www.state.tn.us
100218	florida bureau of investigations	2006-03-07 09:51:08	2	http://www.flsbi.com
100218	florida bureau of investigations	2006-03-07 09:51:08	1	http://www.fhp.state.fl.us
100218	government finance officers asssociation	2006-03-07 21:16:11		
100218	state of tennessee controllers manual	2006-03-07 21:17:12		
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	3	http://www.comptroller.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	4	http://www.tbr.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	9	http://audit.tennessee.edu
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	1	http://www.nysscpa.org
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	4	http://www.massdor.com
100218	municipality fraud detection techniques	2006-03-07 21:41:40		
100218	municipal fraud audit detection internal controls	2006-03-07 21:43:15		
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	1	http://www.whitehouse.gov
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	4	http://www.nhlgc.org
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	7	http://www.sao.state.ut.us
100218	evaluating internal controls a local government managers guide	2006-03-07 21:51:18	5	http://www.allbusiness.com

Query Log analysis



The 250 most frequent queried terms in the "famous" AOL query log!

Thanks to http://www.wordle.net for the tagcloud generator

- Slide from Ricardo Baeza-Yates

Query Log Analysis in Literature

- Enhance ranking – retrieval, advertisement
- Query suggestion; refinement; expansion; substitution, …
- Spelling check
- Other tasks …

Query Log Analysis in Literature

Query log name	Public	Period	# Queries	# Sessions	# Users
Excite '97	Y	Sep '97	1,025,908	211,063	$\sim 410,360$
Excite '97 (small)	Y	Sep '97	51,473	N.D.	$\sim 18,113$
Altavista	N	Aug 2^{nd} - Sep 13^{th} '98	993,208,159	285,474,117	N.D.
Excite '99	Y	Dec '99	1,025,910	325,711	$\sim 540,000$
Excite '01	Y	May '01	1,025,910	262,025	$\sim 446,000$
Altavista (public)	Y	Sep '01	7,175,648	N.D.	N.D.
Tiscali	N	Apr '02	3,278,211	N.D.	N.D.
TodoBR	Y	Jan - Oct '03	22,589,568	N.D.	N.D.
TodoCL	N	May – Nov '03	N.D.	N.D.	N.D.
AOL (big)	N	Dec 26^{th} '03 – Jan 1^{st} '04	$\sim 100,000,000$	N.D.	$\sim 50,000,000$
Yahoo!	N	Nov '05 – Nov '06	N.D.	N.D.	N.D.
AOL (small)	Y	Mar 1^{st} - May 31^{st} '06	36,389,567	N.D.	N.D.

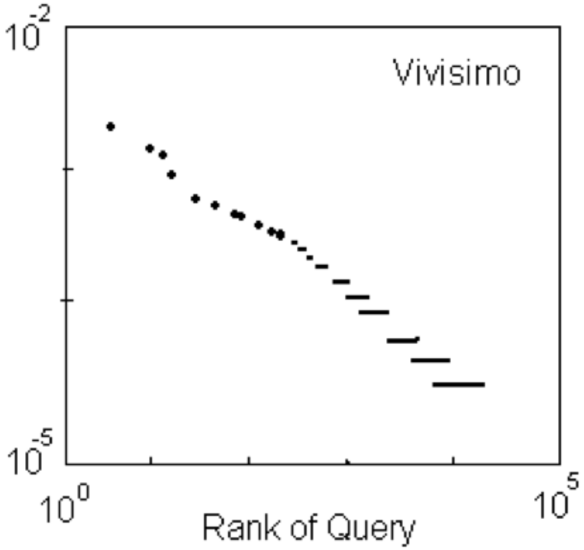
- Mei and Church 08: MSN Search – 18 months, 637 million unique queries, 585 million unique urls, 193 million unique IP addresses

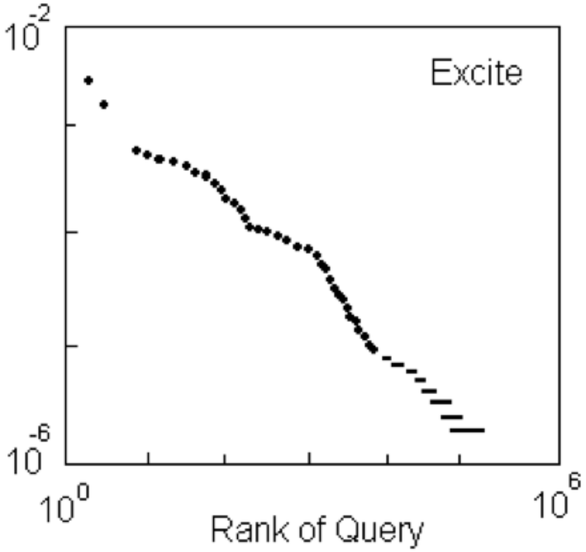
Main Results of Query Log Analysis

- Average number of terms in a query is ranging from a low of 2.2 to a high of 2.6
- The most common number of terms in a query is 2
- 45% (2001) of queries are about Commerce, Travel, Economy, People (was 20%1997)
  - The queries about adult content or entertainment decreased from 20% (1997) to around 7% (2001)
- The majority of users don’t refine their query
  - The number of users who viewed only a single page increase 29% (1997) to 51% (2001) (Excite)
  - 85% of users viewed only first page of search results (AltaVista)

This slide is from Pierre Baldi

Power-law Characteristics





Power-Law in log-log space

$$f(r) = c\, r^{-a}$$
$$\log(f(r)) = \log(c) - a \log(r)$$

- Frequency f(r) of Queries with Rank r
  - 110000 queries from Vivisimo
  - 1.9 Million queries from Excite
- There are strong regularities in terms of patterns of behavior in how we search the Web

This slide is from Pierre Baldi

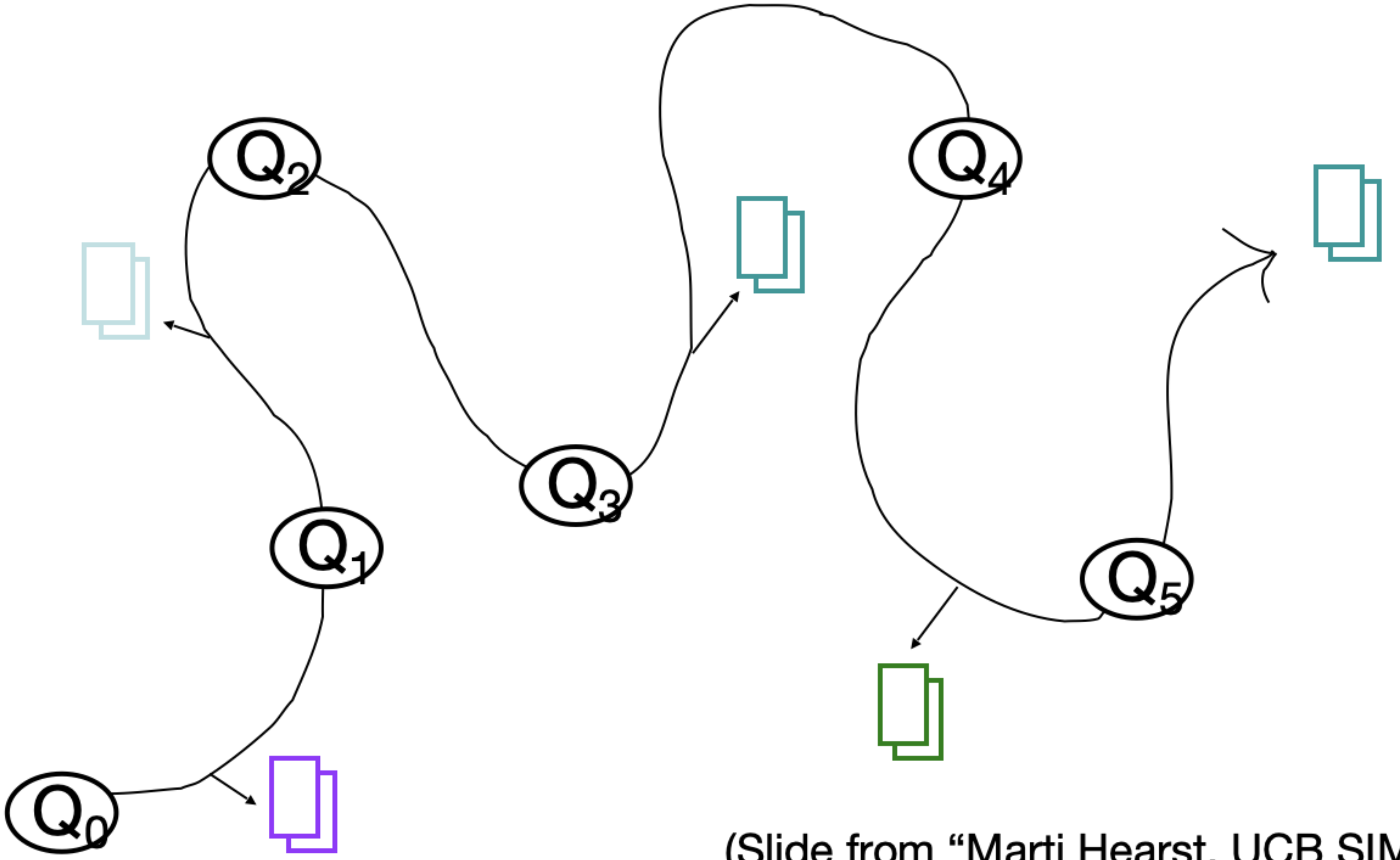
Entropy of Search Logs - How Hard is Search? With Personalization? With Backoff? (Mei and Church, 2008)

- Traditional Search
  - H(URL | Query)
  - 2.8 (= 23.9 – 21.1)
- Personalized Search
  - H(URL | Query, IP)
  - 1.2 (= 27.2 – 26.0)

Personalization cuts H in Half!

	Entropy (H)
Query	21.1
URL	22.1
IP	22.1
All But IP	23.9
All But URL	26.0
All But Query	27.1
All Three	27.2

A sketch of a searcher… “moving through many actions towards a general goal of satisfactory completion of research related to an information need.” (after Bates 90)



(Slide from “Marti Hearst, UCB SIMS, Fall 98) mustang

www.fordvehicles.com/cars/mustang

www.mustang.com

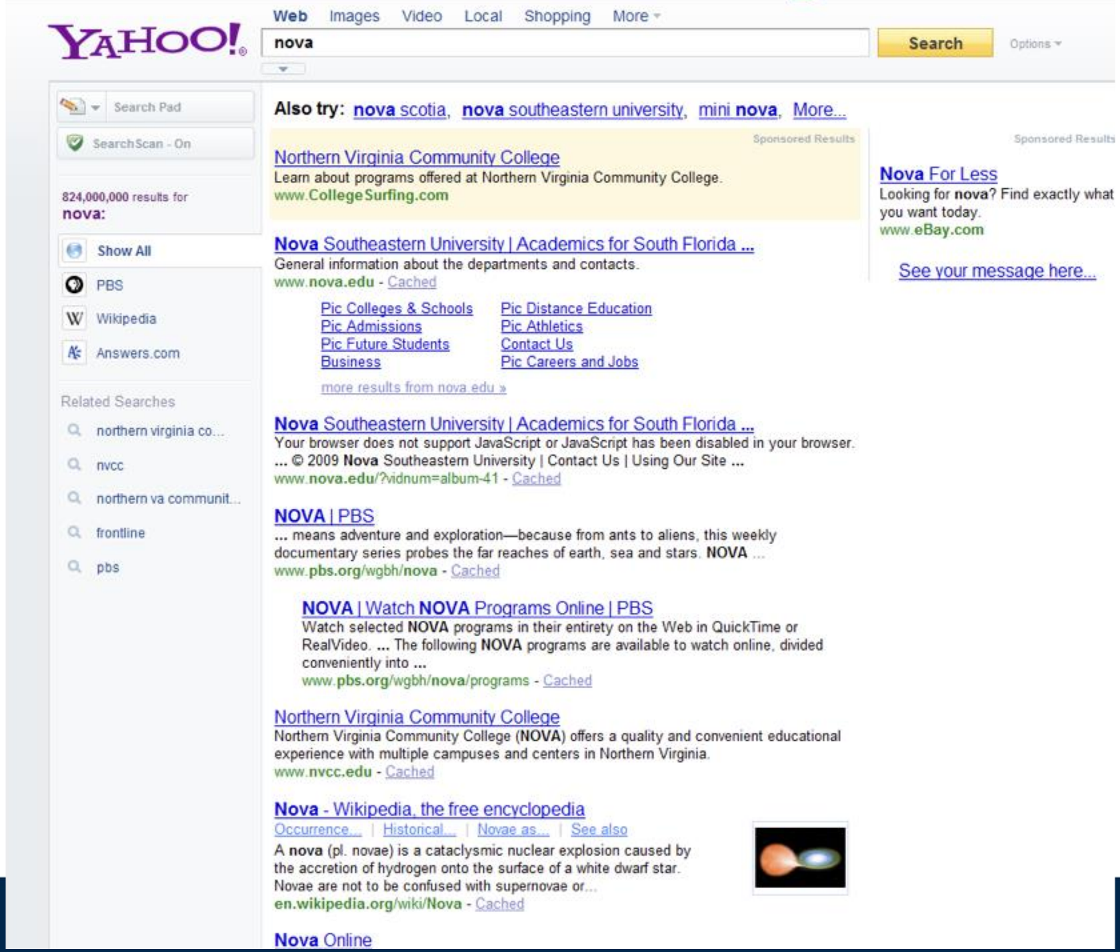
AlsoTry

ford mustang

en.wikipedia.org/wiki/Ford_Mustang

Nov

...



Search sequence

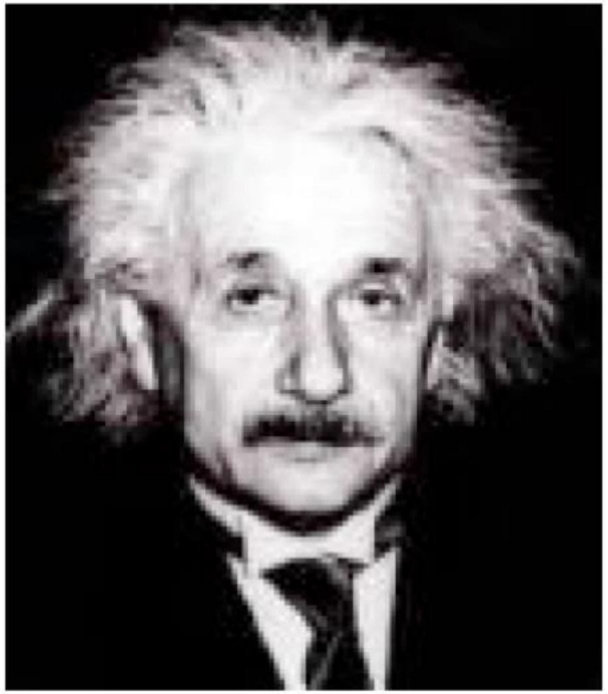
Query Session Detection

- Roughly defined as queries that are submitted by the same user in a short period of time
- Hypothesis:
  - Queries in the same session are related
  - Queries in the same session reflect the same mission/task, etc.
  - Queries in the same session reflect the "modification" relationship
- How to segment query sequence into sessions?
- Heuristic methods; Machine learning methods (hidden Markov models, conditional random fields, etc)

Example – A Poet's Corner

- AOL User 23187425 typed the following queries within a 10 minutes time-span:
    - you come forward 2006-05-07 03:05:19
    - start to stay off 2006-05-07 03:06:04
    - i have had trouble 2006-05-07 03:06:41
    - time to move on 2006-05-07 03:07:16
    - all over with 2006-05-07 03:07:59
    - joe stop that 2006-05-07 03:08:36
    - i can move on 2006-05-07 03:09:32
    - give you my time in person 2006-05-07 03:10:07
    - never find a gain 2006-05-07 03:10:47
    - i want change 2006-05-07 03:11:15
    - know who iam 2006-05-07 03:11:55
    - curse have been broken 2006-05-07 03:12:30
    - told shawn lawn mow burn up 2006-05-07 03:13:50
    - burn up 2006-05-07 03:14:14
    - was his i deal 2006-05-07 03:15:13
    - i would have told him 2006-05-07 03:15:46
    - to kill him too 2006-05-07 03:16:18

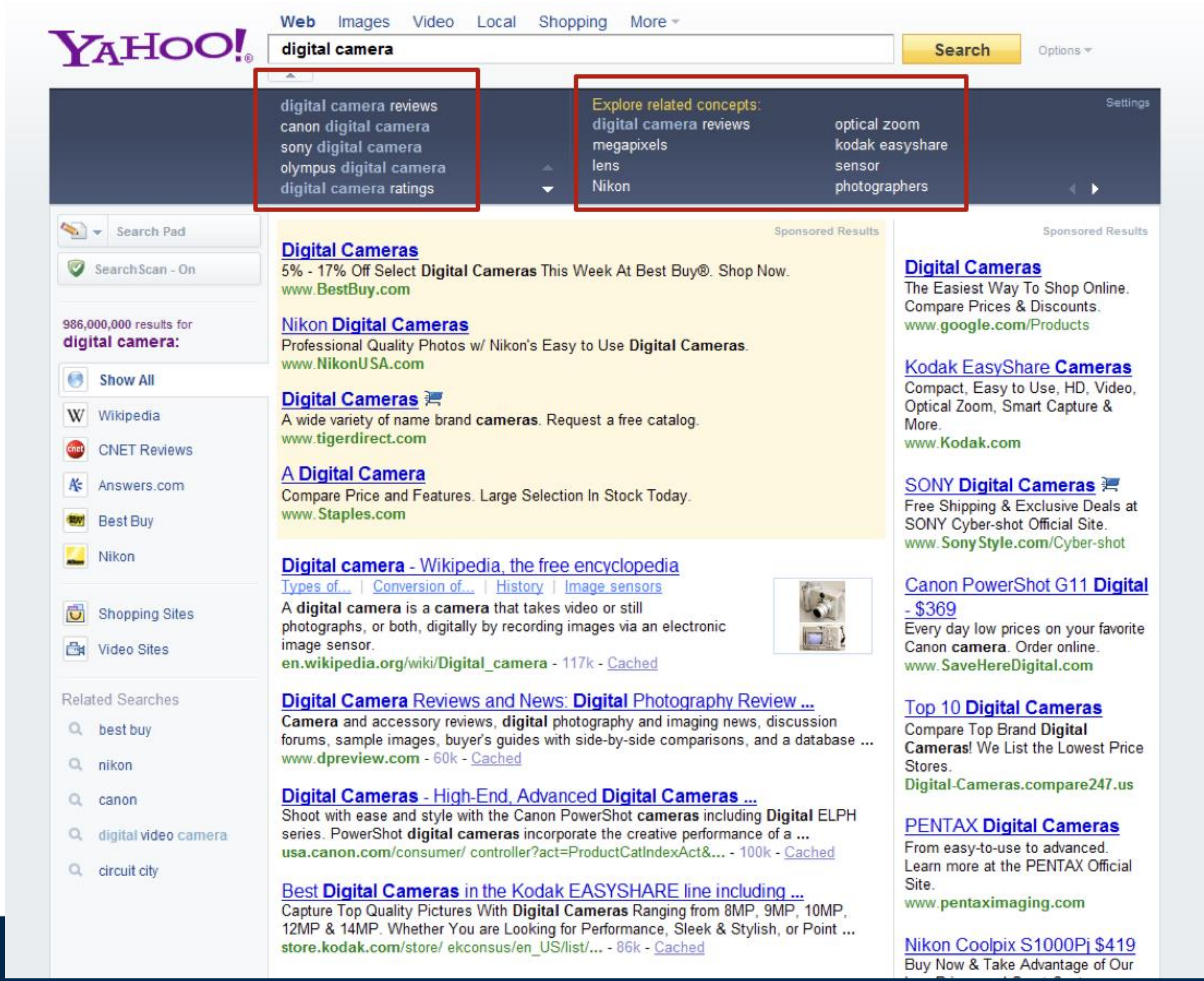
Query Reformulation – Spelling Correction



[Cucerzan and Brill, 2004]

albert einstein	4834
albert einstien	525
albert einstine	149
albert einsten	27
albert einsteins	25
albert einstain	11
albert einstin	10
albert eintein	9
albeart einstein	6
aolbert einstein	6
alber einstein	4
albert einseint	3
albert einsteirn	3
albert einsterin	3
albert eintien	3
alberto einstein	3
albrecht einstein	3
alvert einstein	3

Query Suggestions (Expension)



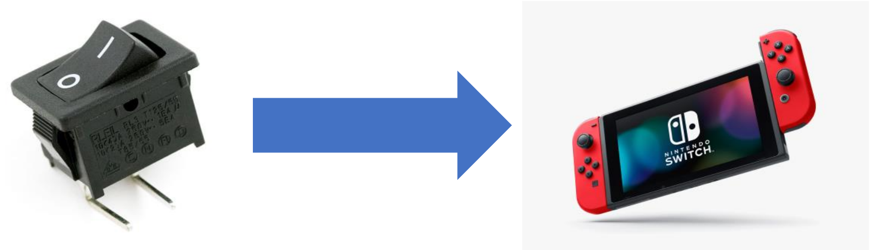
Query Semantic Selection

- The meanings of words are changing over time.
  - E.g., switch

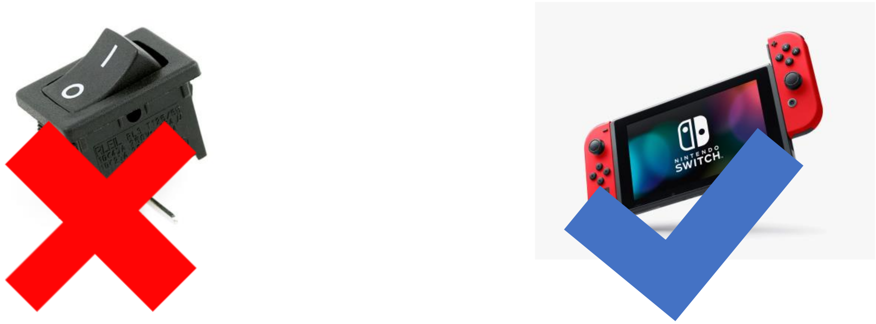
This slide is from Zhuofeng Wu

Query Semantic Selection

- The meanings of words are changing over time.
  - E.g., switch



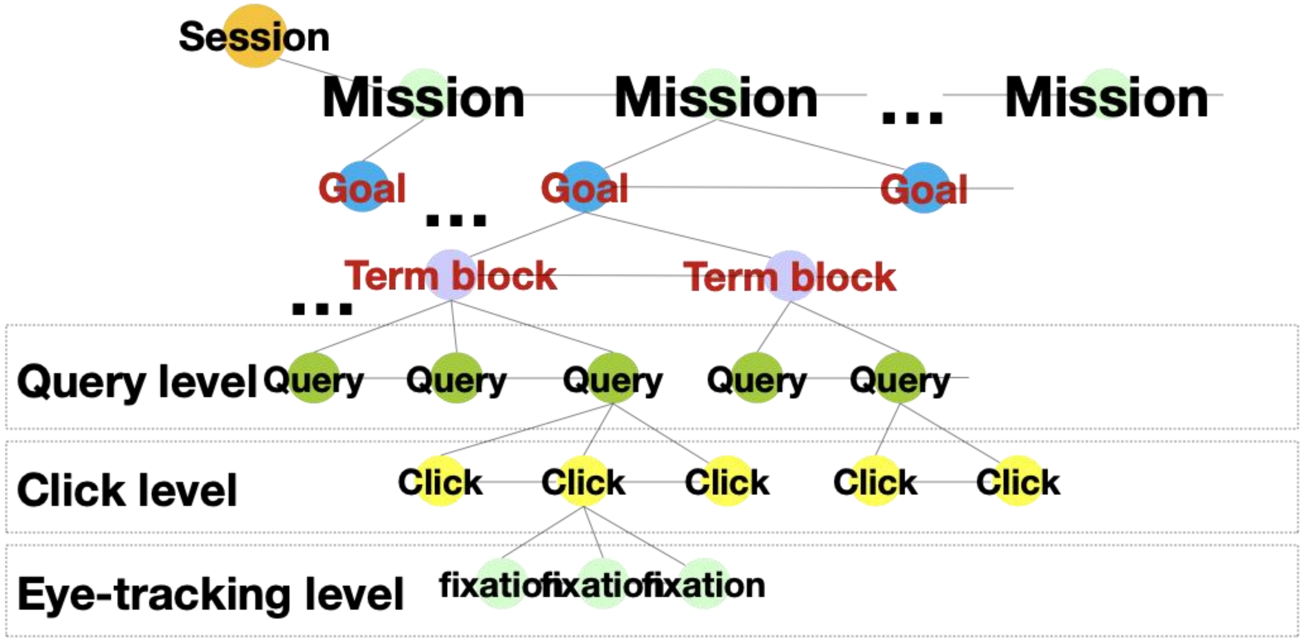
  - Hypothesis:
    - Users prefer to know the new meaning of the word



This slide is from Zhuofeng Wu

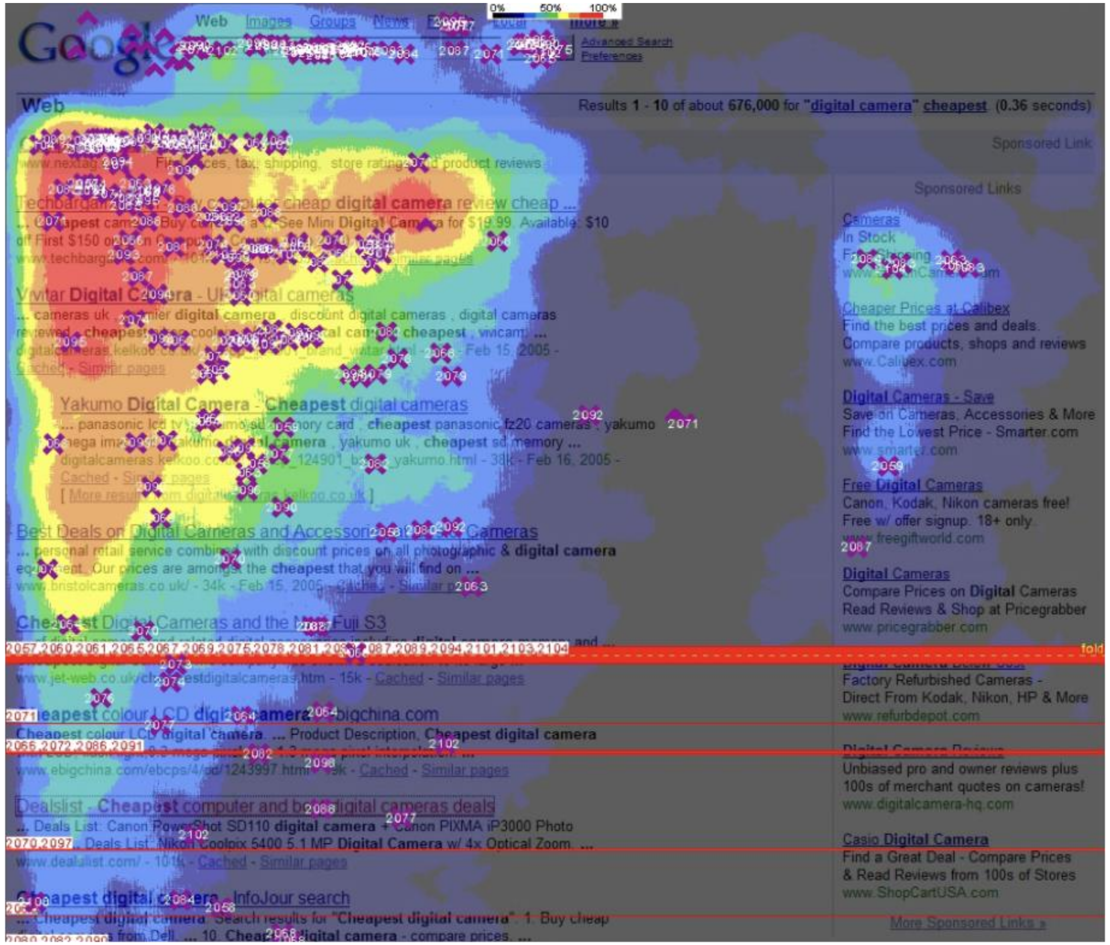
Beyond Query Logs?

- Browsing logs
- Eye-tracking logs
- Social bookmarks?



Nested search sequences – Mei et al. 09

Eye Tracking (Golden Triangle)



- Google Eye Tracking Heat Map, Eyetools Eyetracking Research

Understanding the individual

- Gather information beyond the query
- Explicit v. implicit
- Client-side v. server-side

This slide is from Jaime Teevan

Learning More Explicitly v. Implicitly

- Explicit
  - User shares more about query intent
  - User shares more about interests
  - Hard to express interests explicitly

Query Words

berkeley restaurants

California or Illinois?
Fancy or low key?

This slide is from Jaime Teevan

Learning More Explicitly v. Implicitly

- Explicit
  - User shares more about query intent
  - User shares more about interests
  - Hard to express interests explicitly
- Implicit
  - Query context inferred
  - Profile inferred about the user
  - Less accurate, needs lots of data

This slide is from Jaime Teevan

Personalized search

Personalized search: Basic Idea

Lies at the intersection of Information retrieval and Recommender systems.

Why do we need personalization?

Understanding queries in isolation is hard For e.g. Query – “MSR”

- Microsoft Research
- Mountain Safety Research

- 46% of people found improvement with core ranking
- 70% of people found improvement with personalization
- More of these stats at: https://www.forbes.com/sites/blakemorgan/2020/02/18/50-stats-showing-the-power-of-personalization/

Personalized search: Solutionizing

Solution:

We need to personalize the results based on each user information

What exactly do we mean by user information?

- Who is asking? A programmer vs a carpenter
- What have they done in the past? Visited URLs?
- Where they are? In Michigan vs in California
- When is it? Is it winter or summer?

How to approach personalization

Almost all the approaches tackles the problem by figuring out the actual intent of search

Query logs give an ample amount of information for giving these answers. We use the information available to figure out the intent of query

- E.g., if I type "map" → "Google maps", "Bing maps", "Apple maps" vs it could be *area map*, *Europe map*, etc.

But not all queries have a potential of personalization

- For e.g., "New York Times" → 95% people go to nytimes.com, hence less scope of personalization

It's important to learn when to personalize!

When to personalize?

Goal: to define a score that can determine a personalizability of a Query

How can we do it?

- Use a Machine learning based model i.e., we can model this problem as a **classification problem**

  P(personalizability | Query) → will give us a probability between 0 – 1 that defines if the query is personalizable

- The above model runs for each user.
- A lot of these models takes both local and global information
- We have a global model for all the users that tells us the personalizability of a query, and then, a user-specific model that's specific for a user. → example later

How to personalize?

There are broadly two ways in which we can personalize search output for the user:

- User-Interface based personalization: Changing the layout of results based on user
- Algorithmic-based personalization: Changing the search results based on user

User-interface personalization

Basic goal of User-interface based personalization:

- Reduce working memory load
- Provide alternative interfaces for novice and expert users
- Reorder content based on search history
- Basic elements required:

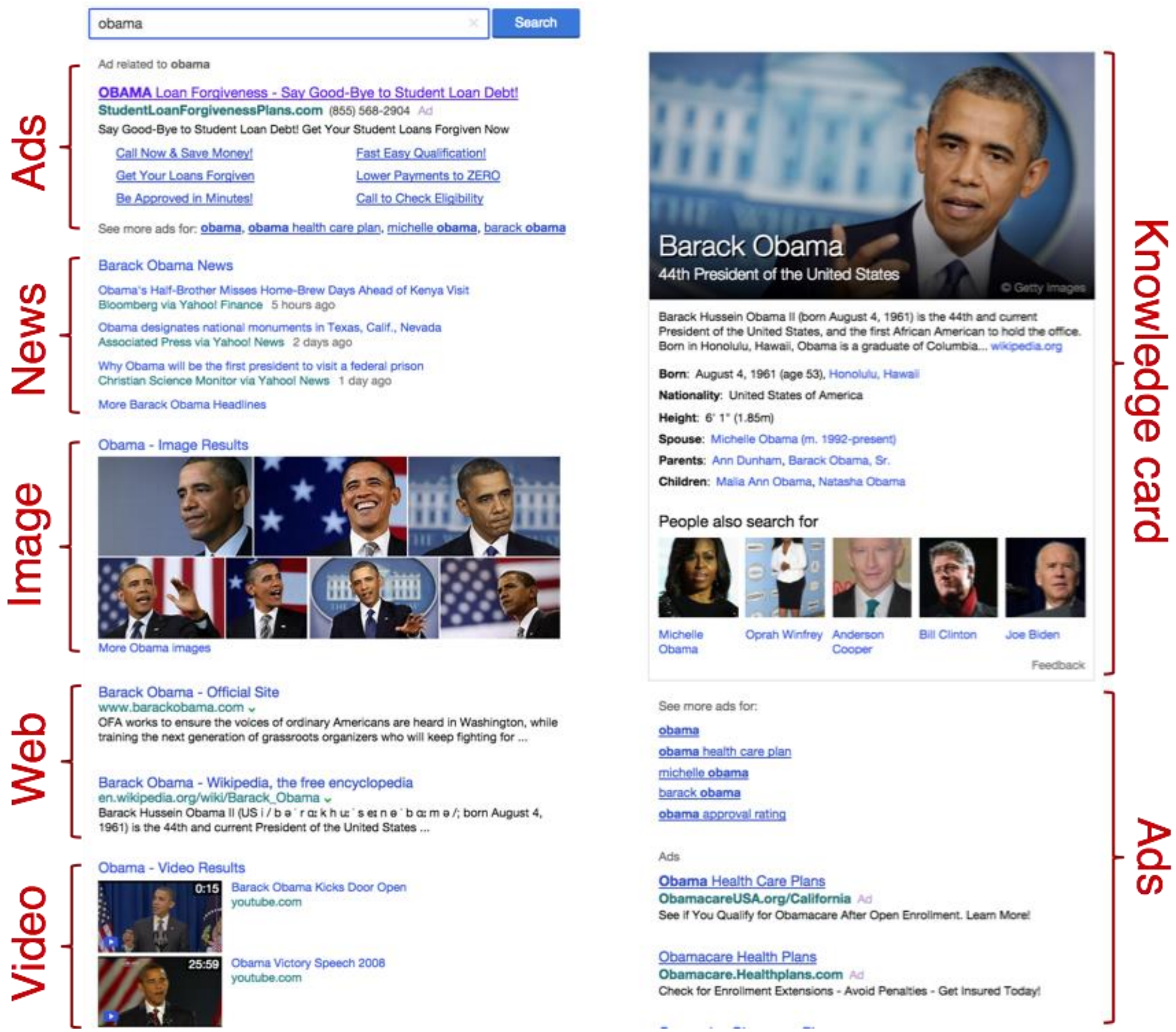
  - Document set selection: What documents to show where?

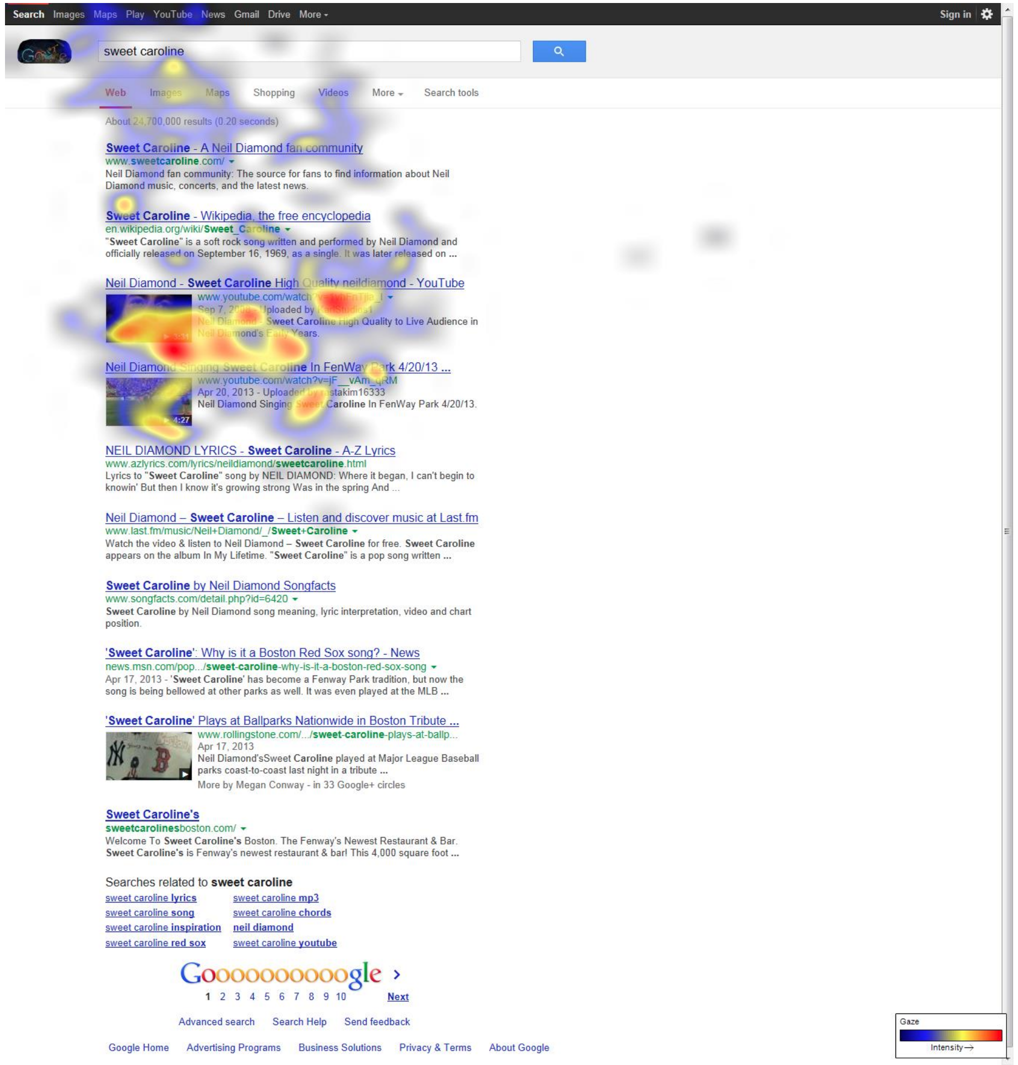
  - Query specification: What exact "intent" are the results for?

  - Result examination: Easy to examine results

  - Interaction support (feedback): Can give back feedback if its not pleasing

Beyond Ranking: Optimizing Whole-Page presentation(Wang et al, WSDM 2016 best paper

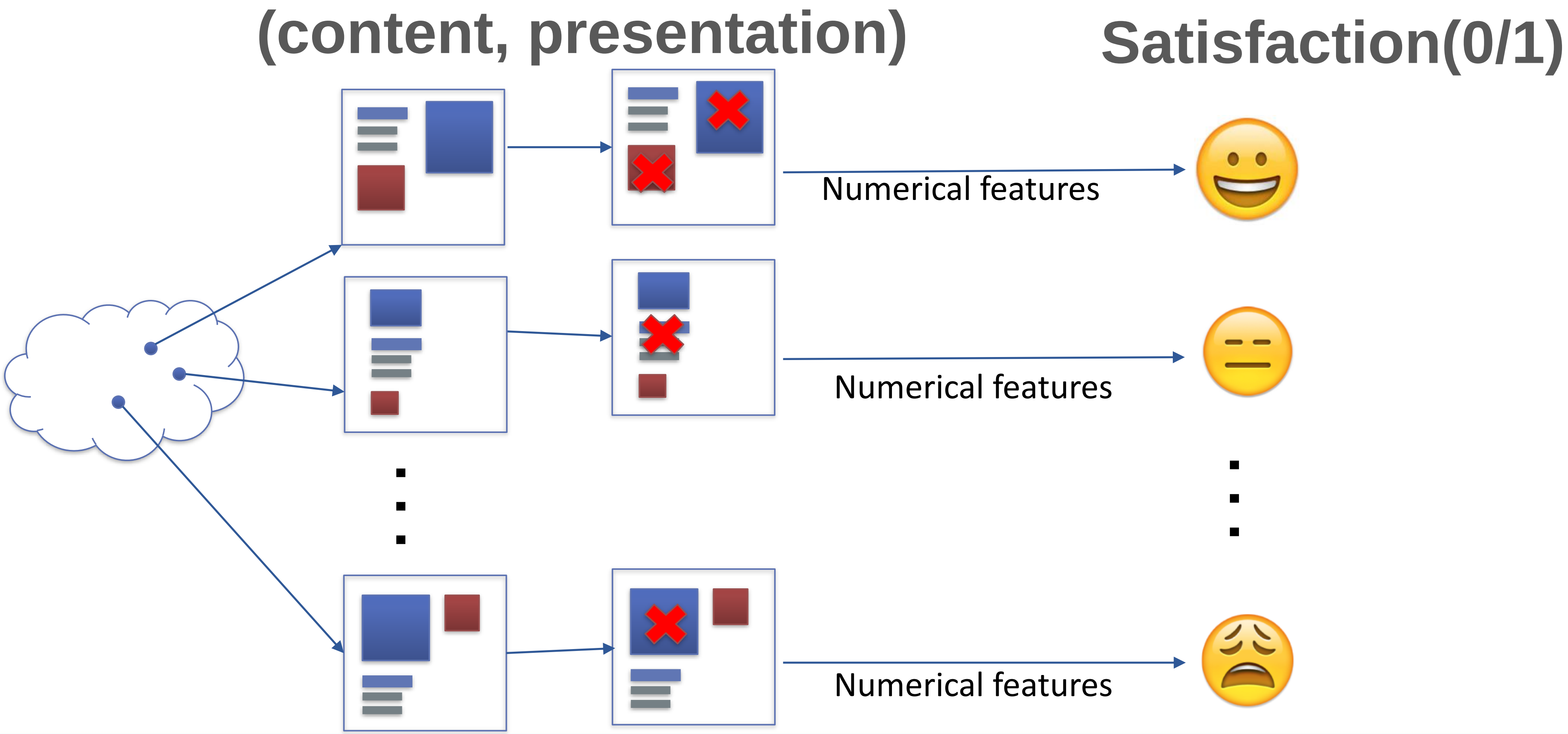




Interface as a Machine learning problem

- Goal: To optimize the page layout given a user and a Query
- How to do it?
  - For any ML problem, we first need a loss function to optimize
  - We will treat this as a classification problem and use its loss function

Interface as a Machine learning problem



Algorithmic-based personalization: model building

Basic idea: Simple → Use the user history to figure out the intent of the given query

How to do it?

Step 1 : Feature extraction

- We can extract different kinds of features from user history:
  - User Content: Queries, desktop index, explicit profile etc.
  - User Behavior: visited web pages, feedback(explicit & implicit)
  - Context features: Location, time of the day/week, etc.
- Factors impacting this feature generation:
  - Short-term history vs long-term history
  - Is it for an individual vs for a group

Algorithmic-based personalization: model usage

- There could also be other factors that influences the model building. These are:

  - Where doe model reside: Server, Client → *Compute power*

  - How used: Ranking, Suggestions, etc.

  - When used: How often are they used?

Step2: Using features to get intent

- We can treat it as a classification problem where we classify the intent of user
- The output are tokenized words(application of BERT-based model)
- OR we can simply use a simple conditional statements to get a rough idea about intent

Algorithmic-based personalization: Case studies

- Based on the usage patterns, and the features that are generated, lets look at 4 case studies where we can use personalization.

  - Navigational search: Search done for navigations

  - Client-side personalized search: Use more User based features

  - Using Long term and Short-term contexts

  - Temporal contexts: Personalization based on time and space

1 Navigational search: Search done for navigations

- Re-finding a web page is common in Web search

  - 33% of queries are repeated queries

  - 39% of clicks are repeated clicks
- Many of these are "navigational" queries

  - e.g., new York times → nytimes.com

  - Shows consistent intent across individuals

  - Identified via low click entropy
- A different version of these are "personal navigation" queries

  -Different intent across individuals but consistent for an individual

1 Navigational search: Search done for navigations

- Navigational queries are low hanging fruit for search engines
- These queries comprise of ~12% of total queries
- They have a high prediction accuracy of ~95%
- In short, high coverage, low risk prediction!

2 Client-side personalized search

- "Client-side" → Simply means that the model is sitting on your device. In the form of cache, cookies, history, bookmarks, etc.
- Re-ranking of web results using user specific information

2 Client-side personalized search

- Personalized ranking model output:

  - Final score = weighted average of web score and personal score

  - score = $\underbrace{\text{alpha x (web score)}}_{\text{Global score}} + \underbrace{\text{(1 – alpha) x personal score}}_{\text{local score}}$

  - alpha → lies between 0 - 1

  - web score = global scores assigned by the search engine

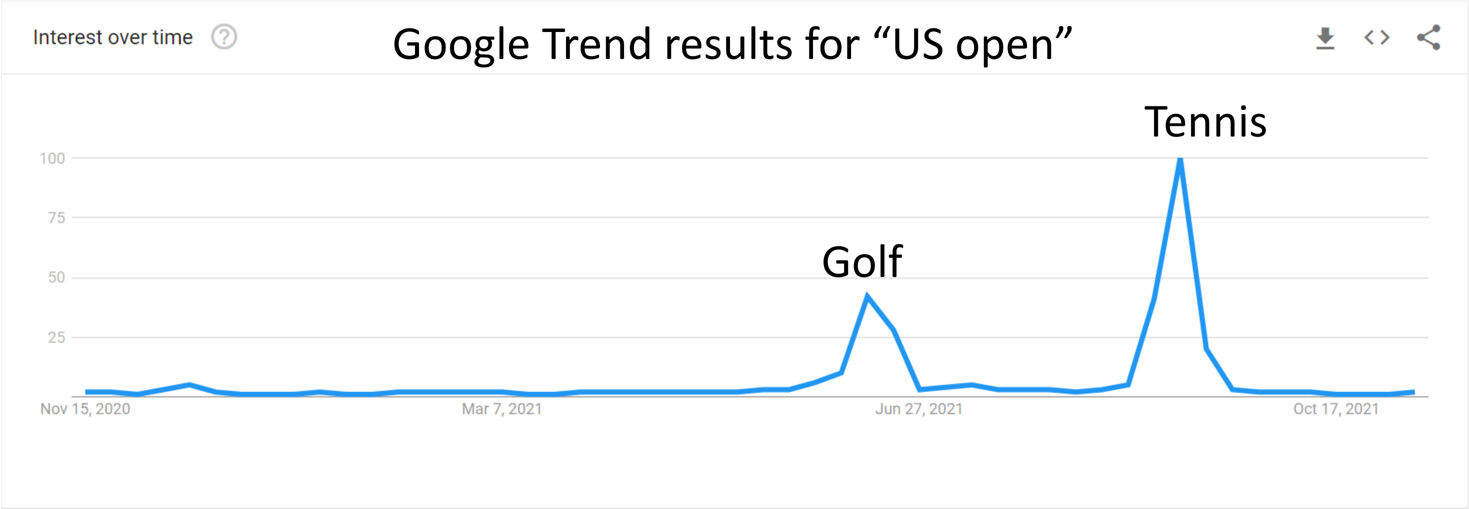
  - personal score → depends on Content and interaction history of user

3 Using Long term and Short-term contexts

- Long term preferences:
  - Content: could use language models, topic models, etc.
  - Analyze behavior: Specific queries, visited URLs
- Short-term tasks
  - Analyze queries within a current session
  - 60% of search has multiple queries in a session
  - We try to predict the intent of current query, given the immediate previous query in the session

4 Temporal contexts: Personalization based on time

- Queries are not uniformly distributed over time
- For the same query, the intent can be different depending on the time



- For e.g., if I type “US open” before the event, I am looking for the tickets and schedule but if I search for “US open” after the event has occurred, then it’s more about the outcome of matches

4 Temporal contexts: Personalization based on time

- Solution:

  -Use time-aware retrieval models → An easy way to do this is add time dependent variables like date, week, etc. to features

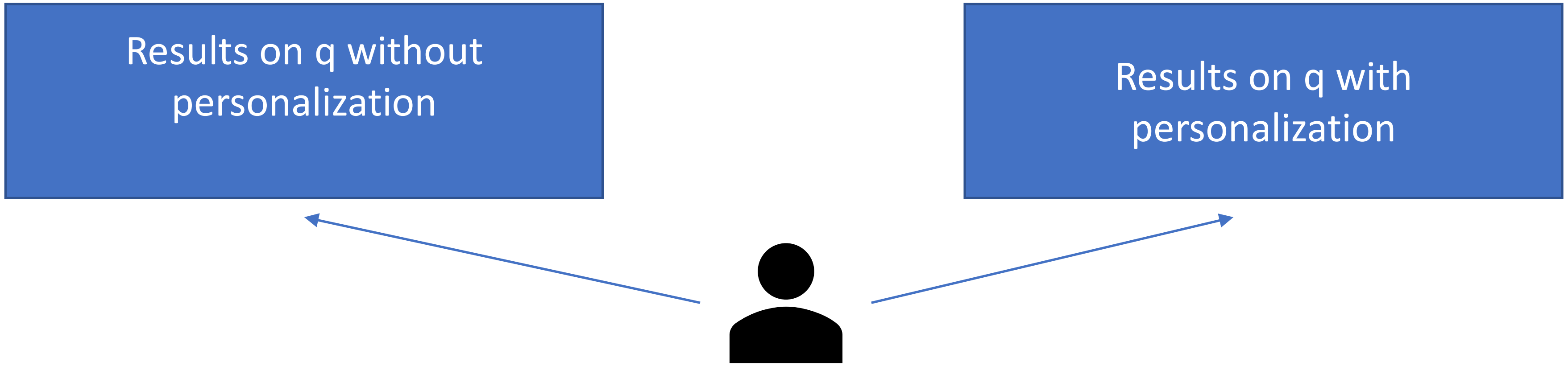
  - Output here is again user intent

Evaluate Personalization

- Recently, personalization has led to Filter Bubble effects → where certain users are simply unable to access information that the search engines' algorithm decides is irrelevant
- Basic strategy for evaluation of personalization in search

  - Use a defined set of queries q

  - perform A/B testing for these queries among different groups



Challenges in personalization

- User centric challenges:

  - Privacy

  - Serendipity and novelty : exploration vs exploitation of content

  - Control and transparency

- System-centric challenges

  -Optimization: Storage, run-time, caching

  - Evaluation: measurement, experimentation

Thanks!