

# **Design of Large Scale Log Analysis Studies**

*A short tutorial...*

Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, Jaime Teevan

HCIC Feb, 2010

# What can we (HCI) learn from logs analysis?

---

- ▶ **Logs are the traces of human behavior**
  - ▶ ... seen through the lenses of whatever sensors we have
- ▶ **Actual behaviors**
  - ▶ As opposed to recalled behavior
  - ▶ As opposed to subjective impressions of behavior



# Benefits

---

- ▶ Portrait of real behavior... warts & all
  - ▶ ... and therefore, a more complete, accurate picture of ALL behaviors, including the ones people don't want to talk about
- ▶ Large sample size / liberation from the tyranny of small N
  - ▶ Coverage (long tail) & Diversity
- ▶ Simple framework for comparative experiments
- ▶ Can see behaviors at a resolution / precision that was previously impossible
- ▶ → **Can inform more focused experiment design**



# Drawbacks

---

- ▶ Not annotated
- ▶ Not controlled
- ▶ No demographics
- ▶ Doesn't tell us the *why*
- ▶ Privacy concerns
  - ▶ AOL / Netflix / Enron / Facebook public
  - ▶ Medical data / other kinds of personally identifiable data



00:32 ...now I know...  
00:35 ... you get a lot of weird things..hold on...  
00:38 “Are Filipinos ready for gay flicks?”  
00:40 How does that have to do with what  
I just....did...?  
00:43 Ummm...  
00:44 So that’s where you can get surprised...  
you’re like, where is this... how does  
this relate...umm...



# What are logs for this discussion?

---

- ▶ User behavior events over time
  - ▶ User activity primarily on web
    - ▶ Edit history
    - ▶ Clickstream
    - ▶ Queries
    - ▶ Annotation / Tagging
    - ▶ PageViews
    - ▶ ... all other instrumentable events (mousetracks, menu events....)
  - ▶ Web crawls (e.g., content changes)
    - ▶ E.g., programmatic changes of content



# Other kinds of large log data sets

---

- ▶ Mechanical Turk (may / may not be truly log-like)
- ▶ Medical data sets
- ▶ Temporal records of many kinds...
  - ▶



# Overview

---

- ▶ **Perspectives on log analysis**
  - ▶ Understanding User Behavior (Teevan)
  - ▶ Design and Analysis of Experiments (Jeffries)
  - ▶ Discussion on appropriate log study design (all)
  
- ▶ **Practical Considerations for log analysis**
  - ▶ Collection & storage (Dumais)
  - ▶ Data Cleaning (Russell)
  - ▶ Discussion of log analysis & HCI community (all)





## Section 2: Understanding User Behavior



Jaime Teevan & Susan Dumais  
Microsoft Research



# Kinds of User Data

---

## **User Studies**

*Controlled interpretation of behavior with detailed instrumentation*



## **User Groups**

*In the wild, real-world tasks, probe for detail*



## **Log Analysis**

*No explicit feedback but lots of implicit feedback*



# Kinds of User Data

---

	Observational
<b>User Studies</b> <i>Controlled interpretation of behavior with detailed instrumentation</i>	In-lab behavior observations
<b>User Groups</b> <i>In the wild, real-world tasks, probe for detail</i>	Ethnography, field studies, case reports
<b>Log Analysis</b> <i>No explicit feedback but lots of implicit feedback</i>	Behavioral log analysis

**Goal: Build an abstract picture of behavior**

---



# Kinds of User Data

---

	Observational	Experimental
<b>User Studies</b> <i>Controlled interpretation of behavior with detailed instrumentation</i>	In-lab behavior observations	Controlled tasks, controlled systems, laboratory studies
<b>User Groups</b> <i>In the wild, real-world tasks, probe for detail</i>	Ethnography, field studies, case reports	Diary studies, critical incident surveys
<b>Log Analysis</b> <i>No explicit feedback but lots of implicit feedback</i>	Behavioral log analysis	A/B testing, interleaved results

**Goal: Build an abstract picture of behavior**

**Goal: Decide if one approach is better than another**

---



# Web Service Logs

## ▶ Example sources

- ▶ Search engine
- ▶ Commerce site

## ▶ Types of information

- ▶ Queries, clicks, edits
- ▶ Results, ads, products

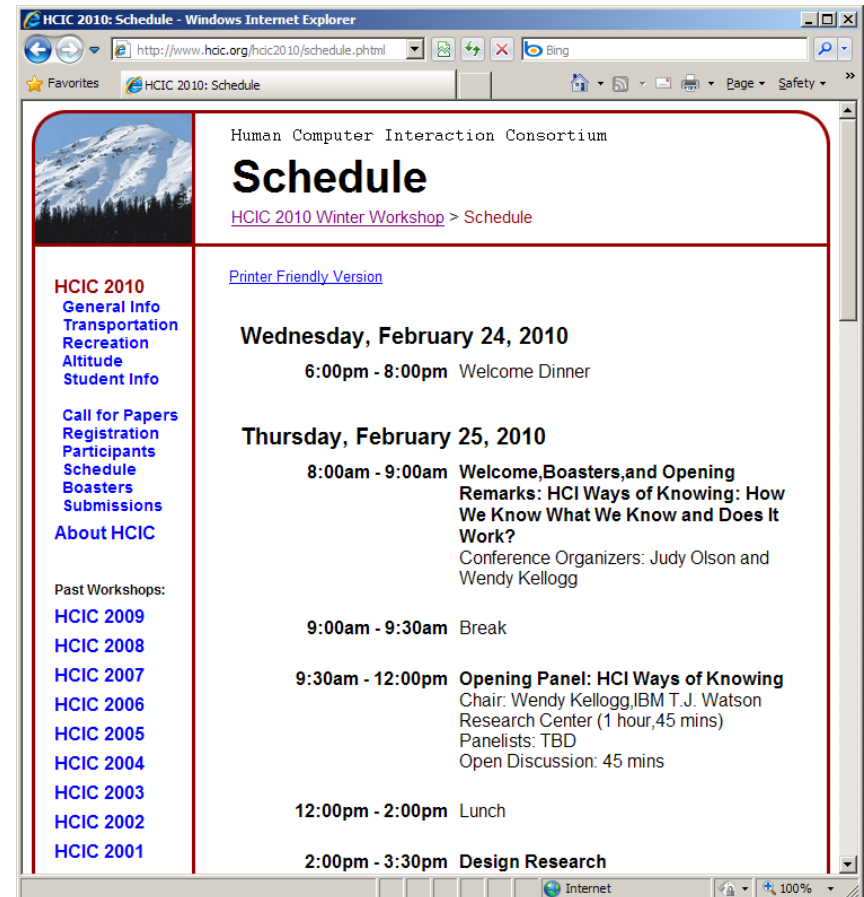
## ▶ Example analysis

- ▶ Click entropy
- ▶ Teevan, Dumais and Liebling. *To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent*. SIGIR 2008

The image shows a screenshot of a Windows Internet Explorer browser displaying a Bing search results page for the query "human computer interaction". The search results are categorized as "ALL RESULTS" and show 1-10 of 8,030,000 results. The top result is a Wikipedia entry titled "Human-computer interaction - Wikipedia, the free encyclopedia". Below it are several other results, including "HCI Bibliography: Human-Computer Interaction Resources" and "Welcome | Human-Computer Interaction Institute". Red arrows point from the text "Academic field" to the Wikipedia result and the HCI Bibliography result. The browser's address bar shows the URL "http://www.bing.com/search?q=human+computer+int". The browser's taskbar at the bottom shows the "Internet" icon and a 100% zoom level.

# Web Browser Logs

- ▶ Example sources
  - ▶ Proxy
  - ▶ Logging tool
- ▶ Types of information
  - ▶ URL visits, paths followed
  - ▶ Content shown, settings
- ▶ Example analysis
  - ▶ Revisitation
  - ▶ Adar, Teevan and Dumais. *Large Scale Analysis of Web Revisitation Patterns*. CHI 2008



# Web Browser Logs

## ▶ Example sources

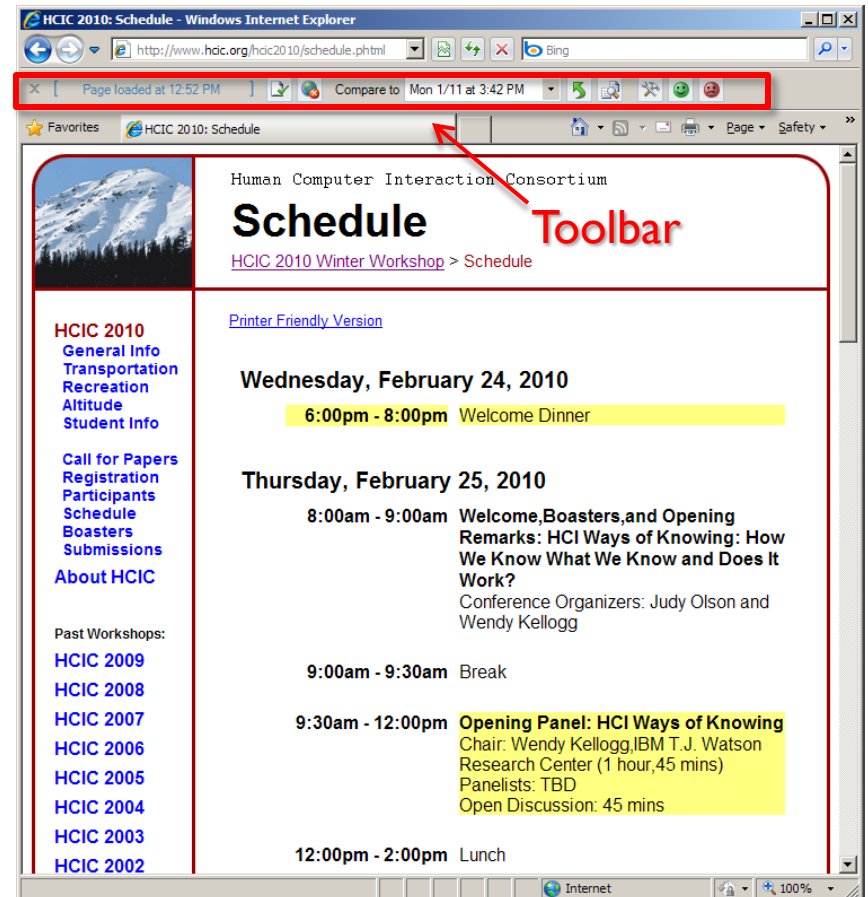
- ▶ Proxy
- ▶ Logging tool

## ▶ Types of information

- ▶ URL visits, paths followed
- ▶ Content shown, settings

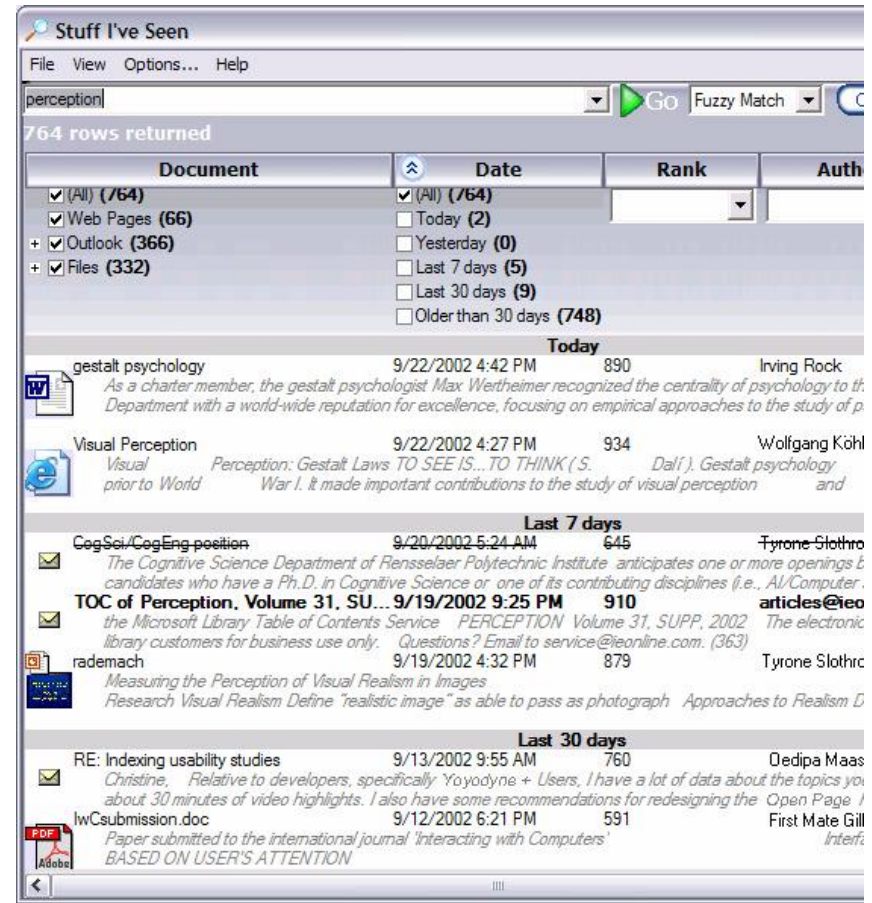
## ▶ Example analysis

- ▶ DiffIE
- ▶ Teevan, Dumais and Liebling. *A Longitudinal Study of How Highlighting Web Content Change Affects People's Web Interactions.* CHI 2010



# Rich Client-Side Logs

- ▶ Example sources
  - ▶ Client application
  - ▶ Operating system
- ▶ Types of information
  - ▶ Web client interactions
  - ▶ Other client interactions
- ▶ Example analysis
  - ▶ Stuff I've Seen
  - ▶ Dumais et al. *Stuff I've Seen: A system for personal information retrieval and re-use*. SIGIR 2003



# Logs Can Be Rich and Varied

---

## Sources of log data

- ▶ **Web service**
  - ▶ Search engine
  - ▶ Commerce site
- ▶ **Web Browser**
  - ▶ Proxy
  - ▶ Toolbar
  - ▶ Browser plug-in
- ▶ **Client application**

## Types of information logged

- ▶ **Interactions**
  - ▶ Queries, clicks
  - ▶ URL visits
  - ▶ System interactions
- ▶ **Context**
  - ▶ Results
  - ▶ Ads
  - ▶ Web pages shown





# Using Log Data

---

- ▶ What can we learn from log analysis?
- ▶ What can't we learn from log analysis?
- ▶ How can we supplement the logs?



# Using Log Data

---

- ▶ **What can we learn from log analysis?**
  - ▶ Now: Observations
  - ▶ Later: Experiments
- ▶ **What can't we learn from log analysis?**
- ▶ **How can we supplement the logs?**



# Generalizing About Behavior

Buttons clicks

Feature use

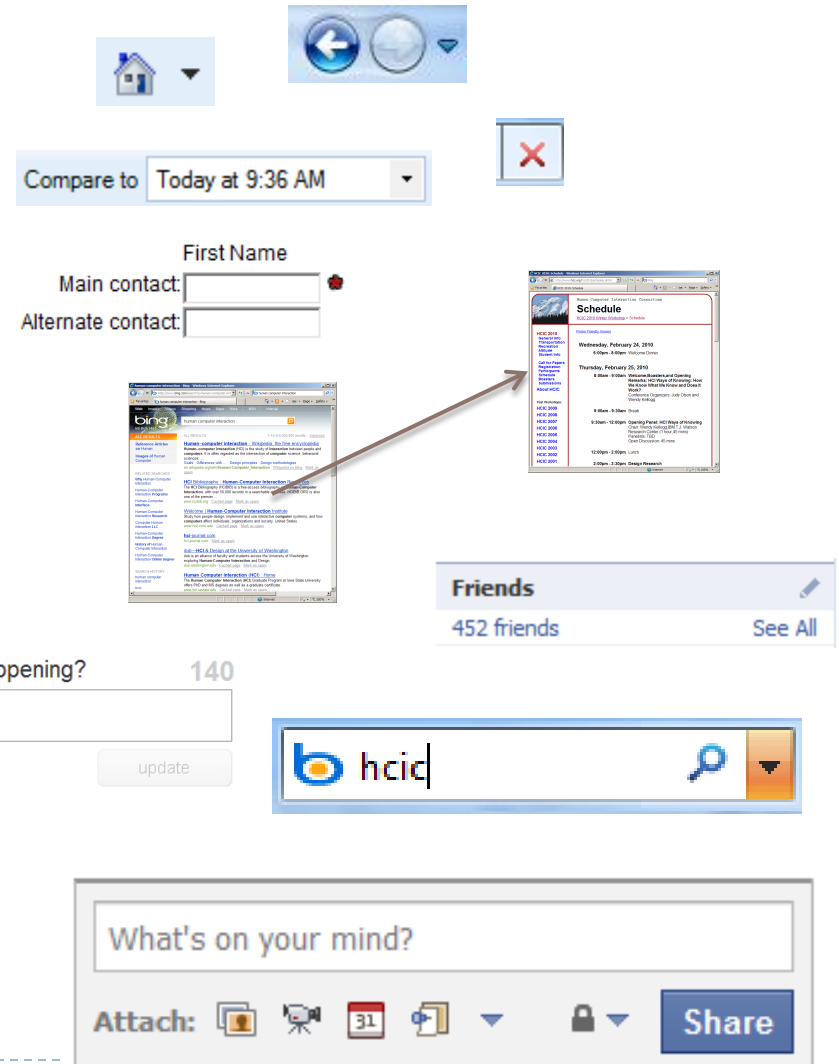
Structured answers

Information use

Information needs

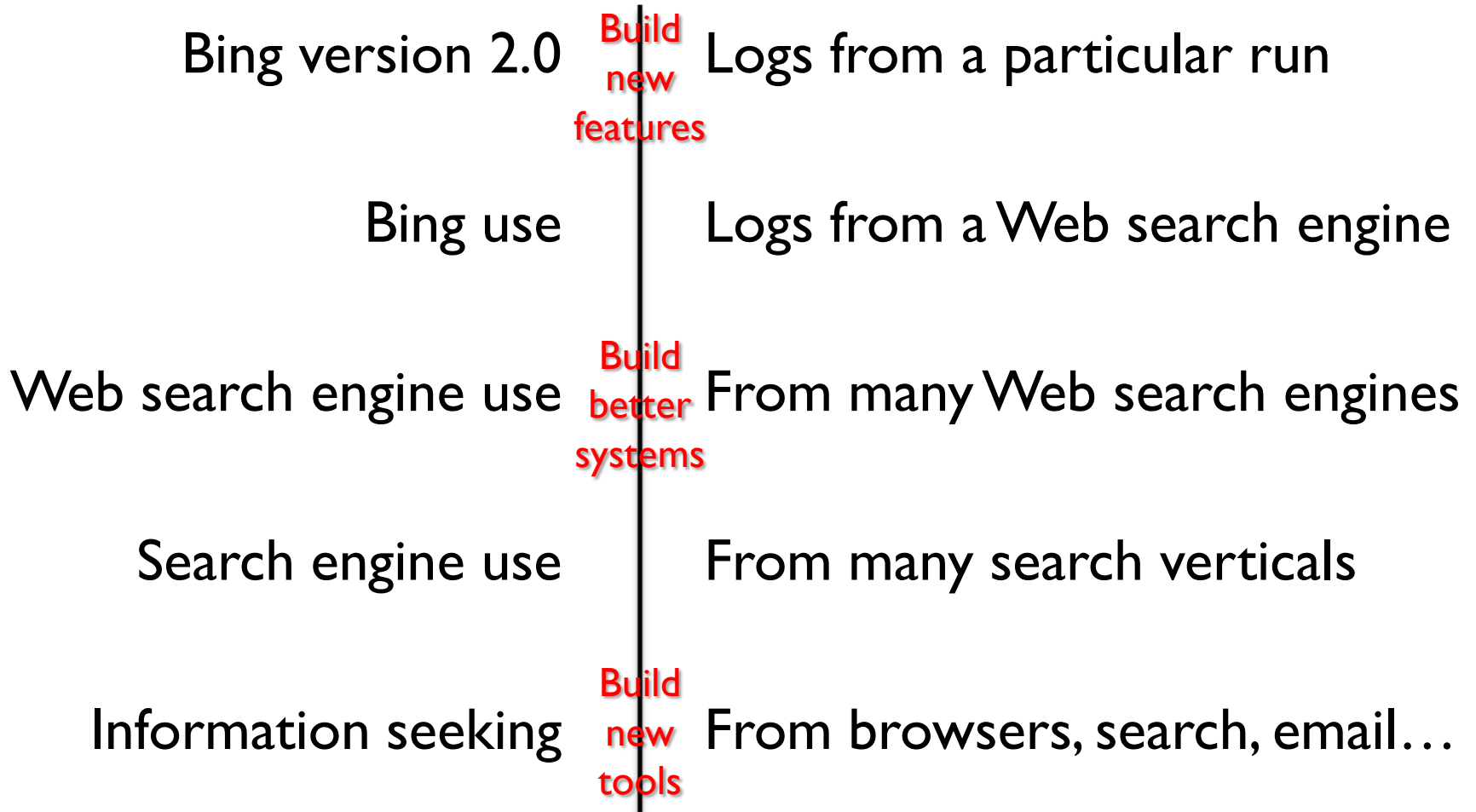
What people think

Human behavior



# Generalizing Across Systems

---



# What We Can Learn from Query Logs

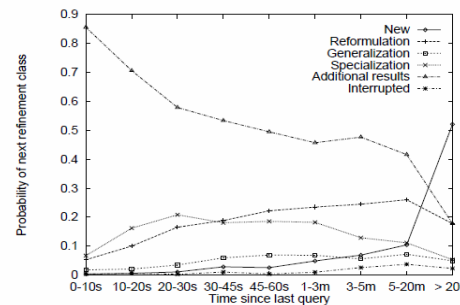
- ▶ **Summary measures**
  - ▶ Query frequency
  - ▶ Query length
- ▶ **Analysis of query intent**
  - ▶ Query types and topics
- ▶ **Temporal features**
  - ▶ Session length
  - ▶ Common re-formulations
- ▶ **Click behavior**
  - ▶ Relevant results for query
  - ▶ Queries that lead to clicks

Queries appear 3.97 times  
[Silverstein et al. 1999]

2.35 terms  
[Jansen et al. 1998]

Navigational,  
Informational,  
Transactional  
[Broder 2002]

Sessions 2.20  
queries long  
[Silverstein et al. 1999]



[Lau and Horvitz, 1999]

	retrieval function		
	bxx	tfc	hand-tuned
avg. clickrank	6.26±1.14	6.18±1.33	6.04± 0.92

[Joachims 2002]



Query	Time	User
hcic	10:41am 2/18/10	142039
snow mountain ranch	10:44am 2/18/10	142039
snow mountain directions	10:56am 2/18/10	142039
hcic	11:21am 2/18/10	659327
restaurants winter park	11:59am 2/18/10	318222
winter park co restaurants	12:01pm 2/18/10	318222
chi conference	12:17pm 2/18/10	318222
hcic	12:18pm 2/18/10	142039
cross country skiing	1:30pm 2/18/10	554320
chi 2010	1:30pm 2/18/10	659327
hcic schedule	1:48pm 2/18/10	142039
hcic.org	2:32pm 2/18/10	435451
mark ackerman	2:42pm 2/18/10	435451
snow mountain directions	4:56pm 2/18/10	142039
hcic	5:02pm 2/18/10	142039

Query	Time	User
* hcic	10:41am 2/18/10	142039
snow mountain ranch	10:44am 2/18/10	142039
snow mountain directions	10:56am 2/18/10	142039
* hcic	11:51am 2/18/10	659327
restaurants winter park	11:59am 2/18/10	318222
winter park co restaurants	12:01pm 2/18/10	318222
* chi conference	12:17pm 2/18/10	318222
* hcic	12:18pm 2/18/10	142039
cross country skiing	1:30pm 2/18/10	554320
* chi 2010	1:30pm 2/18/10	659327
hcic schedule	1:48pm 2/18/10	142039
* hcic.org	2:32pm 2/18/10	435451
mark ackerman	2:42pm 2/18/10	435451
snow mountain directions	4:56pm 2/18/10	142039
* hcic	5:02pm 2/18/10	142039

Query typology



Query	Time	User
* hcic	10:41am 2/18/10	142039
snow mountain ranch	10:44am 2/18/10	142039
snow mountain directions	10:56am 2/18/10	142039
* hcic	11:31am 2/18/10	659327
restaurants winter park	11:59am 2/18/10	318222
winter park co restaurants	12:01pm 2/18/10	318222
chi conference	12:17pm 2/18/10	318222
* hcic	12:31pm 2/18/10	142039
cross country skiing	1:30pm 2/18/10	554320
chi 2010	1:30pm 2/18/10	659327
hcic schedule	1:48pm 2/18/10	142039
hcic.org	2:32pm 2/18/10	435451
mark ackerman	2:42pm 2/18/10	435451
snow mountain directions	4:56pm 2/18/10	142039
* hcic	5:02pm 2/18/10	142039

Query typology

Query behavior





Query	Time
hcic	10:41am 2/18/10
snow mountain ranch	10:44am 2/18/10
* snow mountain directions	10:56am 2/18/10
hcic	11:31am 2/18/10
restaurants winter park	11:59am 2/18/10
winter park co restaurants	12:01pm 2/18/10
chi conference	12:17pm 2/18/10
hcic	12:31pm 2/18/10
cross country skiing	1:30pm 2/18/10
chi 2010	1:30pm 2/18/10
hcic schedule	1:48pm 2/18/10
hcic.org	2:32pm 2/18/10
mark ackerman	2:42pm 2/18/10
* snow mountain directions	4:56pm 2/18/10
hcic	5:02pm 2/18/10

Query typology

Query behavior

Long term trends

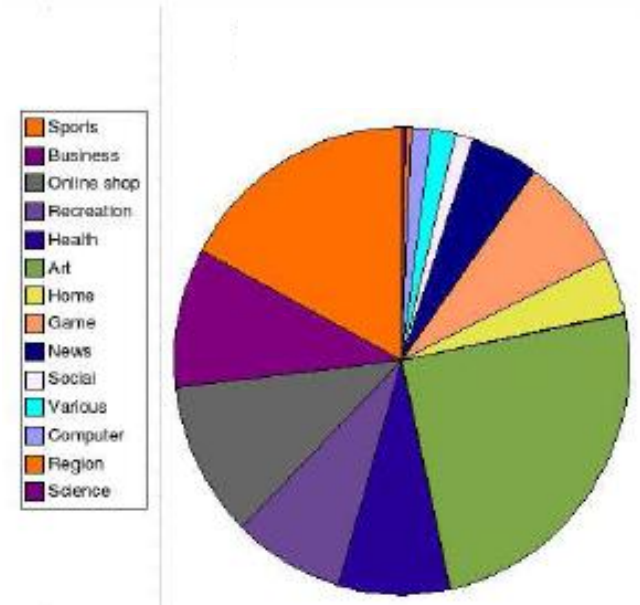
## Uses of Analysis

- ▶ Ranking
  - ▶ E.g., precision
- ▶ System design
  - ▶ E.g., caching
- ▶ User interface
  - ▶ E.g., history
- ▶ Test set development
- ▶ Complementary research

# Partitioning the Data

---

- ▶ Language
- ▶ Location
- ▶ Time
- ▶ User activity
- ▶ Individual
- ▶ Entry point
- ▶ Device
- ▶ System variant



[Baeza Yates et al. 2007]

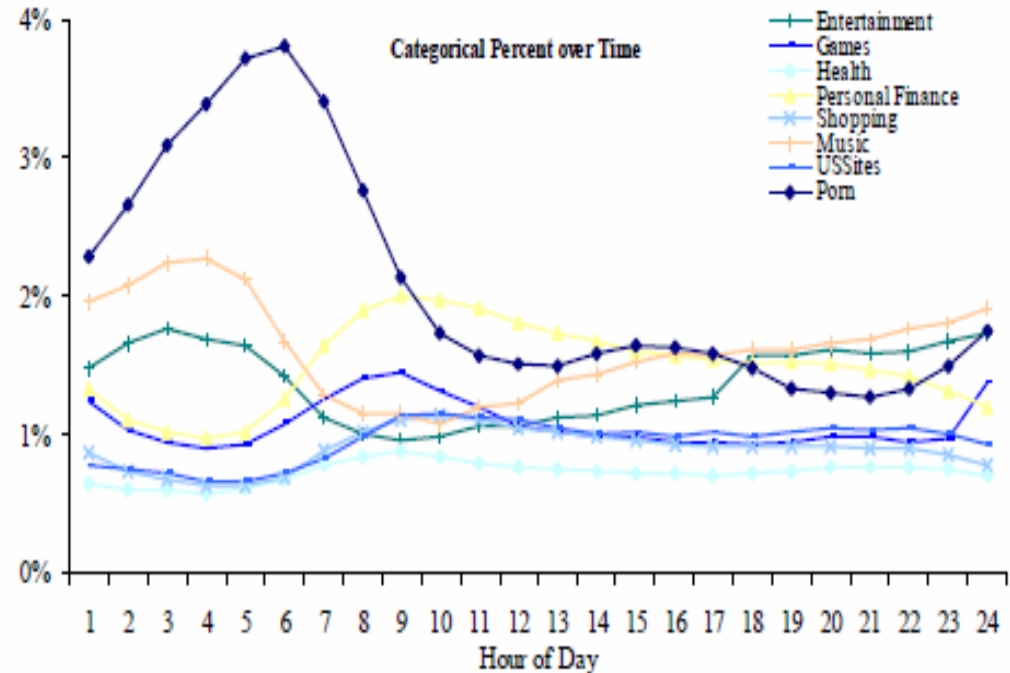
---

▶

# Partition by Time

---

- ▶ Periodicities
- ▶ Spikes
- ▶ Real-time data
  - ▶ New behavior
  - ▶ Immediate feedback
- ▶ Individual
  - ▶ Within session
  - ▶ Across sessions



[Beitzel et al. 2004]

---



# Partition by User

---

All queries: 13,060 queries (100%)	Overlapping Click Queries – 5072 queries (39%)				No Common Clicks 7988 (61%)
	Equal Click Queries – 3777 (29%)		Some Common Clicks 1295 (10%)		
	Single Identical Click 3737 (29%)	Multiple Identical Clicks 40 (< 1%)			
Equal Query Queries 4256 (33%)	Navigational Queries 3100 (24%)	36 (< 1%)	635 (5%)	485 (4%)	
Different Query 8804 (67%)	637 (5%)	4 (< 1%)	660 (5%)	7503 (57%)	

[Teevan et al. 2007]

- ▶ Identification: Temporary ID, user account
  - ▶ Considerations: Coverage v. accuracy, privacy, etc.
- 



# What Logs Cannot Tell Us

---

- ▶ People's intent
- ▶ People's success
- ▶ People's experience
- ▶ People's attention
- ▶ People's beliefs of what's happening
- ▶ Limited to existing interactions
- ▶ Behavior can mean many things



# Example: Click Entropy

▶ Question: How ambiguous is a query?

▶ Answer: Look at variation in clicks.

[Teevan et al. 2008]

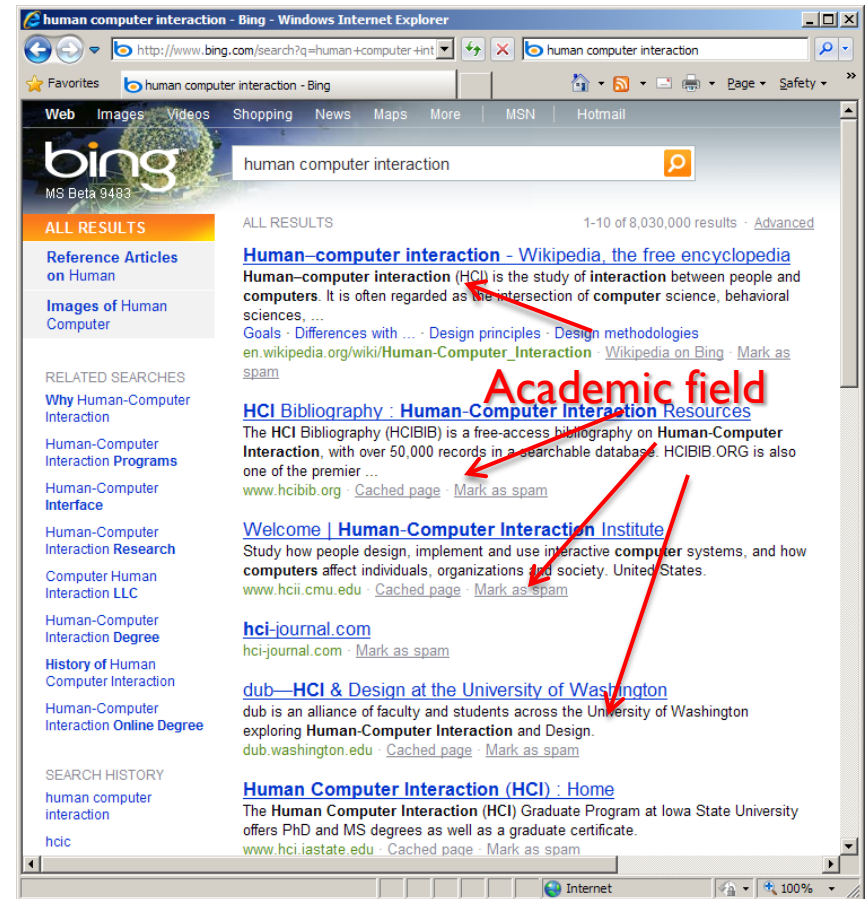
▶ Click entropy

▶ Low if no variation

*human computer interaction*

▶ High if lots of variation

*hci*



# Which Has Lower Click Entropy?

---

▶ [www.usajobs.gov](http://www.usajobs.gov) v. federal government jobs

▶ find phone number v. [msn live search](https://www.msn.com)

▶ [singapore pools](http://singaporepools.com) v. singaporepools.com

↖  
Click entropy = 1.5

Result entropy = 5.7

↖  
Click entropy = 2.0

Result entropy = 10.7

*Results change*



# Which Has Lower Click Entropy?

---

- ▶ www.usajobs.gov v. federal government jobs
- ▶ find phone number v. msn live search
- ▶ singapore pools v. singaporepools.com
- ▶ tiffany v. tiffany's
- ▶ nytimes v. connecticut newspapers

Click entropy = 2.5

Click position = 2.6

Click entropy = 1.0

Click position = 1.6

*Results change*

*Result quality varies*





# Which Has Lower Click Entropy?

---

- ▶ www.usajobs.gov v. federal government jobs
- ▶ find phone number v. msn live search
- ▶ singapore pools v. singaporepools.com
- ▶ tiffany v. tiffany's
- ▶ nytimes v. connecticut newspapers
- ▶ **campbells soup recipes** v. vegetable soup recipe
- ▶ **soccer rules** v. hockey equipment

*Results change*

*Result quality varies*

*Task affects # of clicks*

Click entropy = 1.7

Click /user = 1.1

Click entropy = 2.2

Clicks/user = 2.1



# Dealing with Log Limitations

---

- ▶ Look at data
- ▶ Clean data
- ▶ Supplement the data
  - ▶ Enhance log data
    - ▶ Collect associated information (e.g., what's shown)
    - ▶ Instrumented panels (critical incident, by individual)
  - ▶ Converging methods
    - ▶ Usability studies, eye tracking, field studies, diary studies, surveys



# Example: Re-Finding Intent

---

- ▶ **Large-scale log analysis of re-finding**

[Tyler and Teevan 2010]

- ▶ *Do people know they are re-finding?*
- ▶ *Do they mean to re-find the result they do?*
- ▶ *Why are they returning to the result?*

- ▶ **Small-scale critical incident user study**

- ▶ Browser plug-in that logs queries and clicks
- ▶ Pop up survey on repeat clicks and 1/8 new clicks

= **Insight into intent + Rich, real-world picture**

- ▶ Re-finding often targeted towards a particular URL
- ▶ Not targeted when query changes or in same session



# Section 3: Design and Analysis of Experiments

Robin Jeffries & Diane Tang

# Running Experiments

---

- ▶ **Make a change, compare it to some baseline**
  - ▶ make a visible change to the page. Which performs better - the old or the new?
  - ▶ change the algorithms behind the scenes. Is the new one better?
  - ▶ compare a dozen variants and compute "optimal values" for the variables in play (find a local/global maximum for a treatment value, given a metric to maximize.)



# Experiment design questions

---

- ▶ What is your population
- ▶ How to select your treatments and control
- ▶ What to measure
- ▶ What log-style data is **not** good for



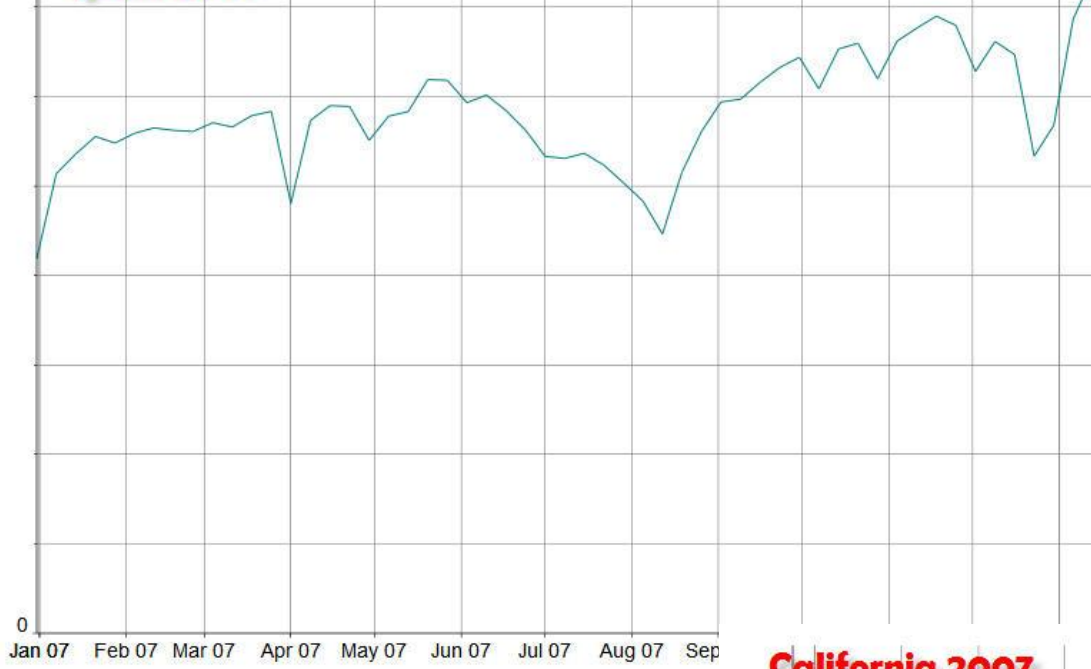
# Selecting a population

---

- **a population** is a set of people
  - in particular location(s)
  - using particular language(s)
  - during a particular time period
  - doing specific activities of interest
- Important to consider how those choices might impact your results
  - Chinese users vs. US users during Golden Week
  - sports related change during Super Bowl week in US vs. UK
  - users in English speaking countries vs. users of English UI vs. users in US



# Spain 2007



# California 2007





# Sampling from your population

---

- **A sample** is a segment of your population
  - e.g., the subset that gets the experimental treatment vs. the control subset
  - important that samples be randomly selected
    - with large datasets, useful to determine that samples are not biased in particular ways (e.g., pre-periods)
  - within-user sampling (all users get all treatments) is very powerful (e.g., studies reordering search results)
- How big a sample do you need?
  - depends on the size of effect you want to detect -- we refer to this as **power**
  - in logs studies, you can trade off number of users vs. time



# Power

---

- **power** is  $1 - \text{prob}(\text{Type II})$  error
  - probability that when there really is a difference, you will statistically detect it
  - most hypothesis testing is all about Type I error
- **power depends on**
  - size of difference you want to be able to detect
  - standard error of the measurement
  - number of observations
- **power can (and should be) pre-calculated**
  - too many studies where there isn't enough power to detect the effect of interest
  - there are standard formulas, e.g., [en.wikipedia.org/wiki/Statistical\\_power](https://en.wikipedia.org/wiki/Statistical_power)



# Power example: variability matters

---

	<b>effect size</b> (% change from control)	<b>standard error</b>	<b>events required</b> (for 90% power at 95% conf. interval)
Metric A	1%	4.4	1,500,000
Metric B	1%	7.0	4,000,000



# Treatments

---

- **treatments:** explicit changes you make to the user experience (directly or indirectly user visible)
- may be compared to other treatments or to the control
  - if multiple aspects change, need multiple comparisons to tease out the different effects
    - you can make sweeping changes, but you often cannot interpret them.
    - a multifactorial experiment is sometimes the answer
  - example: google video universal
    - change in what people see: playable thumbnail of video for video results (left vs. right)
    - change in when they see it: algorithm for which video results show the thumbnail





superbowl commercials

Search

[Advanced Search](#)

Web [+ Show options...](#)

Results 1 - 10 of abo

### News results for **superbowl commercials**



[CBC.ca](#)

[Super Bowl Commercials: Super bowl ads, are they funny?](#) - 19 hours ago  
**Super Bowl Commercials:** Super bowl ads. We will update you on the Super Bowl 44 commercials here. I have been waiting since mid of 2009 to catch up NFL ...

[TV Shows Now! \(blog\)](#) - [19290 related articles »](#)

[Post Super Bowl: Ads, ads and more ads](#) -

[Reuters UK \(blog\)](#) - [1413 related articles »](#)

[Super Bowl Commercials: Oprah, Dave and Jay Make it Easier to Take ...](#) - [The Faster Times](#) - [3 related articles »](#)

### Latest results for **superbowl commercials** - [Pause](#)

[Super Bowl Ad – Oprah Winfrey, Jay Leno & David Letterman starrer ...](#)

[Entertainment and Showbiz!](#) - 2 minutes ago

The **Super Bowl commercial** ads which were selling for US \$3.1 million per 30-seconds slot were another thing to look up to as giant companies unveiled their ...

[Errors of Enchantment » Super Bowl Commercial Hits Close to Home](#)

[Errors of Enchantment](#) - 2 minutes ago

I watched the **Super Bowl** last night and while the game was quite good, the **commercials** were pretty lame overall. However, one **commercial** by ...

[Watch all The Super Bowl Commercials in 2010 and Vote for the Best One](#)

[Super Bowl Commercials SuperBowl-Ads.com - News, Polls, History](#)

**Super Bowl Commercials** - Super Bowl Advertising - Watch all your favorite Super Bowl ads.

[www.superbowl-ads.com/](#) - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)

[2010 Super Bowl Commercials -- NFL FanHouse](#)

Watch the 2010 **Super Bowl commercials** online. Every 2010 Super Bowl ad is here.

[superbowlads.fanhouse.com/](#) - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)

[Superbowl Commercials | 2010 Super Bowl Ads & Commercials](#)

Hundreds of **superbowl commercials** shown from 1960 to 2010. Our Super Bowl ads are carefully selected from dozens of online services, so you don't have to go ...

[www.superbowl-commercials.org/](#) - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)

### Video results for **superbowl commercials**



[Hyundai Super Bowl XLIV Advertising Brett ...](#)

32 sec - 5 days ago

[www.youtube.com](#)



[Hulu's Superbowl Commercial](#)

- 60sec  
1 min 3 sec - Feb 1, 2009

[www.youtube.com](#)

# Example: Video universal

---

- ▶ show a playable thumbnail of a video in web results for highly ranked video results
  - ▶ explore different visual treatments for thumbnails **and** different levels of triggering the thumbnail
  - ▶ treatments
    1. thumbnail on right and conservative triggering
    2. thumbnail on right and aggressive triggering
    3. thumbnail on left and conservative triggering
    4. thumbnail on left and aggressive triggering
    5. control (never show thumbnail; never trigger)
  - ▶ → **note** that this is not a complete factorial experiment (should have 9 conditions)
- 



# Controls

---

- a **control** is the standard user experience that you are comparing a change to
- What is the right control?
  - gold standard:
    - equivalent sample from same population
    - doing similar tasks
    - using either
      - The existing user experience
      - A baseline “minimal” “boring” user experience



# How controls go wrong

---

- ▶ treatment is opt-in
- ▶ treatment or control limited to subset (e.g., treatment only for English, control world-wide)
- ▶ treatment and control at different times
- ▶ control is all the data, treatment is limited to events that showed something novel





# Counter-factuals

---

- ▶ controls are not just who/what you count, but *what you log*
    - ▶ you need to identify the events where users *would have experienced* the treatment (since it is rarely all events)
      - > referred to as **counter-factual**
    - ▶ video universal example: log in the control when either conservative or aggressive triggering would have happened
      - control shows no video universal results
      - log that this page would have shown a video universal instance under (e.g.,) aggressive triggering
      - enables you to compare equivalent subsets of the data in the two samples
- 



# Logging counter-factuals

---

- ▶ needs to be done at expt time
  - ▶ often very hard to reverse-engineer later
  - ▶ gives a true apples-to-apples comparison
  - ▶ not always possible (e.g., if decisions being made "on the fly")



# What should you measure?

---

- ▶ often have dozens or hundreds of possible effects
  - ▶ clickthrough rate, ave. no. of ads shown, next page rate,
  - ▶ some matter almost all the time
    - in search: CTR
  - ▶ some matter to your hypothesis
    - if you put a new widget on the page, do people use it?
    - if you have a task flow, do people complete the task?
  - ▶ some are collaterally interesting
    - increased nextpage rate to measure "didn't find it"
  - ▶ sometimes finding the "right" metrics is hard
    - "good abandonment"



# Remember: log data is NOT good for...

---

- Figuring out *why* people do things
  - need more direct user input
- Tracking a user over time
  - without special tracking software, the best you can do on the web is a cookie
    - a cookie is **not** a user [Sue to discuss more later]
- Measuring satisfaction/feelings directly
  - there are some indirect measures (e.g., how often they return)



# Experiment Analysis

---

- ▶ Common assumptions you can't count on
- ▶ Confidence intervals
- ▶ Managing experiment-wide error
- ▶ Real world challenges
- ▶ Simpson's Paradox
- ▶ Not losing track of the big picture



# Experiment Analysis for large data sets

---

- ▶ Different from Fisherian hypothesis testing
  - ▶ Too many dependent variables
    - > t-test, F-test often don't make sense
  - ▶ don't have factorial designs
  - ▶ Type II error is as important as Type I

	True difference exists	True difference does not exist
Difference observed in expt	Correct positive result	False Alarm (Type I error)
Difference not observed in expt	Miss (Type II error)	Correct negative result

## Many assumptions don't hold:

- > independence of observations
- > normal distributions
- > homoscedasticity



# Invalid assumptions: independent observations

---

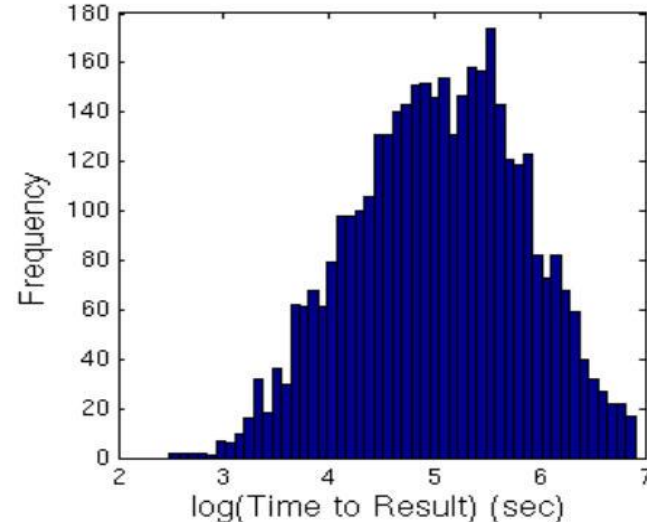
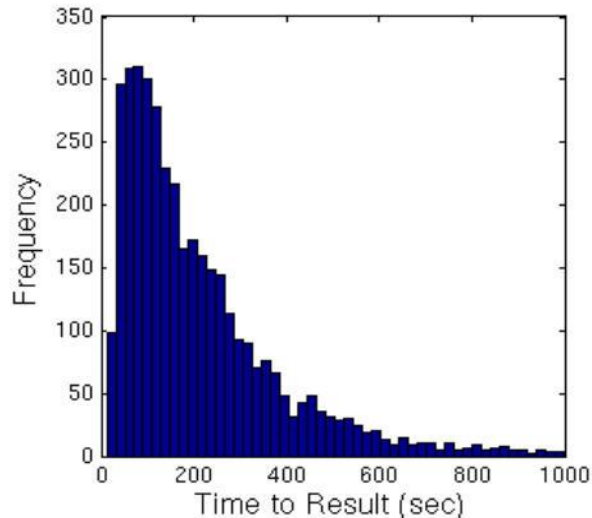
- ▶ if I clicked on a "show more" link before, I'm more likely to do it again
- ▶ if I queried for a topic before, I'm more likely to query for that topic again
- ▶ if I search a lot today, I'm more likely to search a lot tomorrow



# Invalid assumptions: Data is Gaussian

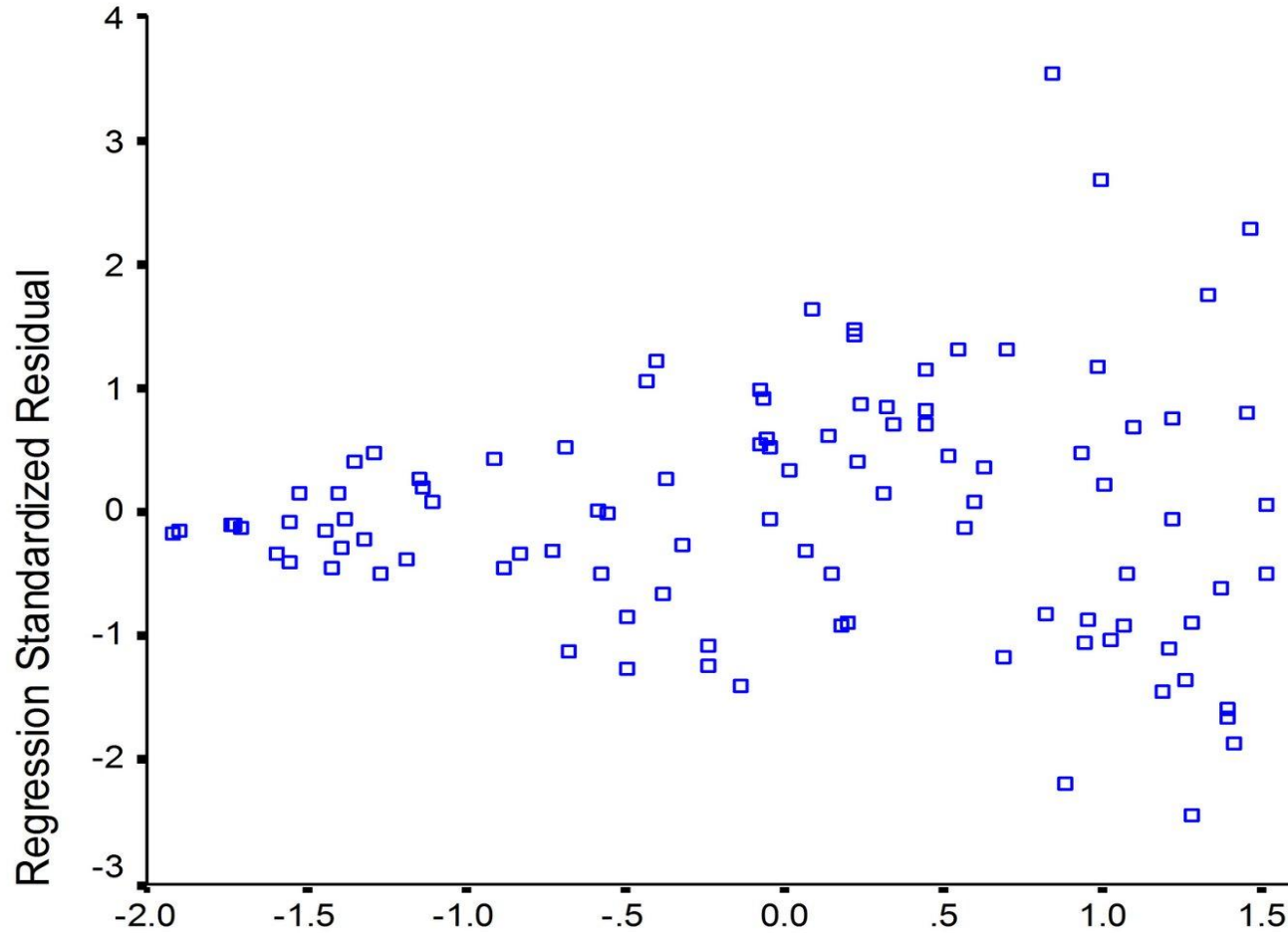
---

- Doesn't the law of large numbers apply?
  - Apparently not
- What to do: transform the data if you can
- Most common for time-based measures (e.g., time to result)
  - log transform can be useful
  - geo-metric mean (multiplicative mean) is an alternative transformation





# Invalid assumptions: Homoscedasticity



Variability (deviation from line of fit) is not uniform

# Confidence intervals

---

- **confidence interval (C.I.):** interval around the treatment mean that contains the true value of the mean  $x\%$  (typically 95%) of the time
- C.I.s that do not contain the control mean are statistically significant
- this is an independent test for each metric
  - thus, you will get 1 in 20 results (for 95% C.I.s) that are spurious -- you just don't know which ones
- C.I.s are not necessarily straightforward to compute.



# Managing experiment wide error

---

- ▶ **Experiment wide error:** overall probability of Type I error.
  - ▶ Each individual result has a 5% chance of being spuriously significant (Type I error)
  - ▶ Close to 1.0 that at least one item is spuriously significant.
- ▶ If you have a set of *a priori* metrics of interest, you can modify the confidence interval size to take into account the number of metrics
- ▶ Instead, you may have many metrics, and not know all of the interesting ones until after you do the analysis.
- ▶ Many of your metrics may be correlated
  - ▶ Lack of a correlation when you expect one is a clue



# Managing real world challenges

---

- Data from all around the world
  - eg: collecting data for a given day (start/end times differ), collecting "daytime" data
- One-of-a-kind events
  - death of Michael Jackson/Anna Nicole Smith
  - problems with data collection server
  - data schema changes
- Multiple languages
  - practical issues in processing many orthographies
    - ex: dividing into words to compare query overlap
  - restricting language:
    - language  $\neq$  country
    - query language  $\neq$  UI language



# Analysis challenges

---

- **Simpson's paradox:** simultaneous mix and metric changes

**Batting averages**

	<b>1995</b>	<b>1996</b>	<b>Combined</b>
Derek Jeter	12/48 .250	183/582 .314	195/630 <b>.310</b>
David Justice	104/411 <b>.253</b>	45/140 <b>.321</b>	149/551 .270

- changes in mix (denominators) make combined metrics (ratios) inconsistent with yearly metrics



# More on Simpson's paradox

---

- ▶ neither the individual data (the yearly metrics) or the combined data is inherently more correct
  - ▶ it depends, of course, on what you want to do
- ▶ once you have mix changes (changes to the denominators across subgroups), all metrics (changes to the ratios) are suspect
  - ▶ **always** compare your denominators across samples
  - ▶ if you wanted to produce a mix change, that's fine
  - ▶ can you restrict analysis to the data not impacted by the mix change (the subset that didn't change)?
  - ▶ minimally, be up front about this in any writeup



# Detailed analyses → Big picture

---

- ▶ not all effects will point the same direction
  - ▶ take a closer look at the items going in the "wrong" direction
    - can you interpret them?
      - > e.g., people are doing fewer next pages because they are finding their answer on the first page
    - could they be artifactual?
    - what if they are real?
      - > what should be the impact on your conclusions? on your decision?
- ▶ significance and impact are not the same thing
  - ▶ Couching things in terms of % change vs. absolute change helps
  - ▶ A substantial effect size depends on what you want to do with the data



# Summing up

---

- Experiment design is not easy, but it will save you a lot of time later
  - population/sample selection
  - power calculation
  - counter-factuals
  - controlling incidental differences
- Analysis has its own pitfalls
  - Type I (false alarms) and Type II (misses) errors
  - Simpson's paradox
  - real world challenges
- Don't lose the big picture in the details





# Section 4: Discussion

All

# Our story to this point...

---

## ▶ Perspectives on log analysis

### 2. Understanding user behavior *Jamie*

- ▶ What you can / cannot learn from logs
- ▶ Observations vs. experiments
- ▶ Different kinds of logs

### 3. How to design / analyze large logs *Robin*

- ▶ Selecting populations
  - ▶ Statistical Power
  - ▶ Treatments
  - ▶ Controls
  - ▶ Experimental error
- 



# Discussion

---

- ▶ How might you use log analysis in your research?
- ▶ What other things might you use large data set analysis to learn?
  - ▶ Time-based data vs. non-time data
- ▶ Large vs. small data sets?
- ▶ How do HCI researchers review log analysis papers?
  - ▶ Isn't this just "large data set" analysis skills?
    - ▶ (A la medical data sets)
    - ▶ Other kinds of data sets:
      - Large survey data
      - Medical logs
      - Library logs





## Section 5: Practical Considerations for Log Analysis



# Overview

---

- ▶ **Data collection and storage** [Susan Dumais]
  - ▶ How to log the data
  - ▶ How to store the data
  - ▶ How to use the data responsibly
- ▶ **Data analysis** [Dan Russell]
  - ▶ How to clean the data
- ▶ **Discussion: Log analysis and the HCI community**





# Section 6: Data Collection, Storage and Use



Susan Dumais and Jaime Teevan  
Microsoft Research

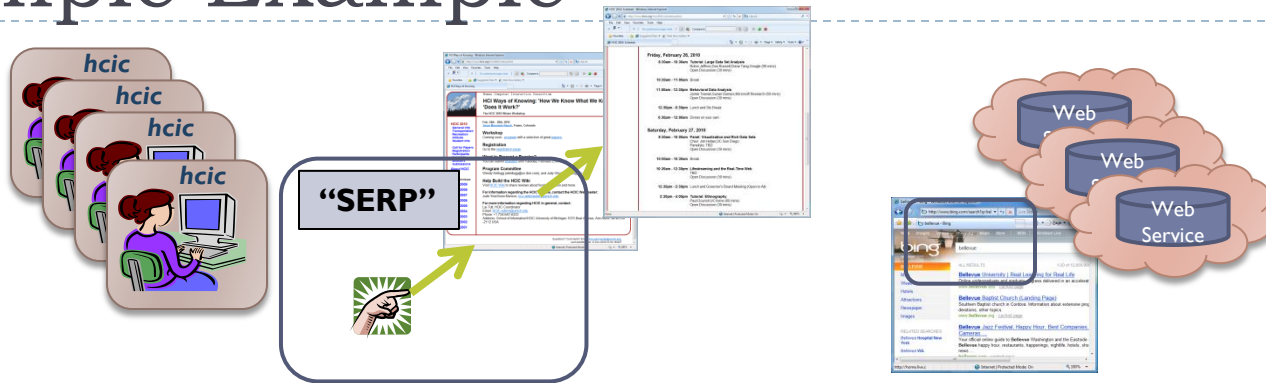
# Overview

---

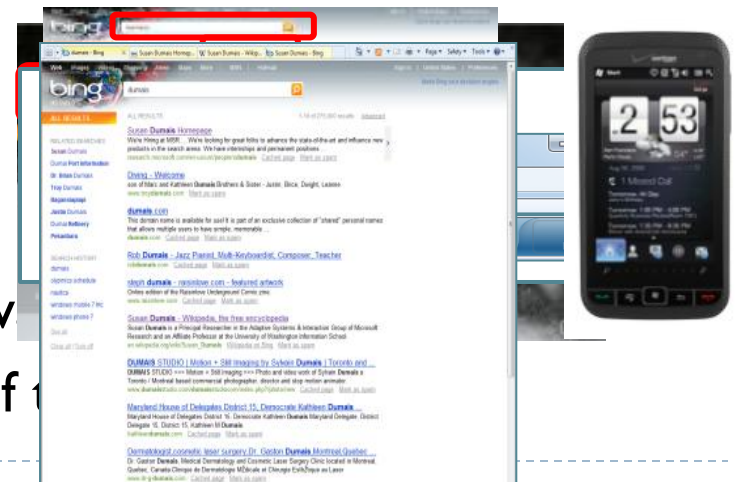
- ▶ How to log the data?
- ▶ How to store the data?
- ▶ How to use the data responsibly?
  
- ▶ Building large-scale systems out-of-scope



# A Simple Example

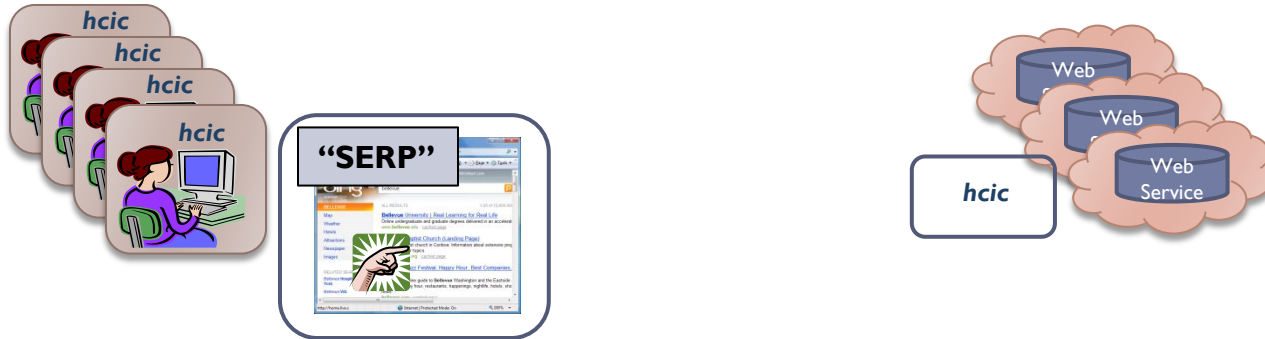


- ▶ Logging search Queries and Clicked Results
- ▶ Logging Queries
  - ▶ Basic data: <query, userID, time>
  - ▶ Additional contextual data:
    - ▶ Where did the query come from?
    - ▶ What results were returned?
    - ▶ What algorithm or presentation was used?
    - ▶ Other metadata about the state of the user's device





# A Simple Example (cont'd)



- ▶ Logging Clicked Results (on the SERP)
  - ▶ How can a Web service know which links are clicked?
    - ▶ Proxy re-direct [adds complexity & latency; may influence user interaction]
    - ▶ Script (e.g., CSJS) [dom and cross-browser challenges]
  - ▶ What happened after the result was clicked?
    - ▶ Going beyond the SERP is difficult
  - ▶ Was the result opened in another browser window or tab?
    - ▶ Browser actions (back, caching, new tab) difficult to capture
    - ▶ Matters for interpreting user actions [next slide]
  - ▶ Need richer client instrumentation to interpret search behavior

# Browsers, Tabs and Time

## ► Interpreting what happens on the SERP

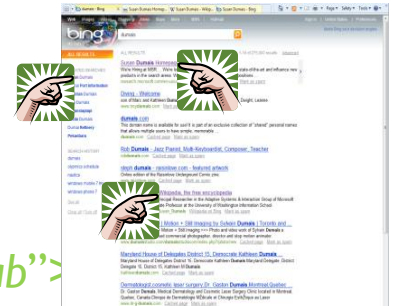
### • Scenario 1:

- 7:12 SERP shown
- 7:13 click R1  
*<“back” to SERP>*
- 7:14 click R5  
*<“back” to SERP>*
- 7:15 click RS1  
*<“back” to SERP>*
- 7:16 go to new search engine



### • Scenario 2

- 7:12 SERP shown
- 7:13 click R1  
*<“open in new tab”>*
- 7:14 click R5  
*<“open in new tab”>*
- 7:15 click RS1  
*<“open in new tab”>*
- 7:16 read R1
- 10:21 read R5
- 13:26 copies links to doc



- Both look the same, if all you capture is clicks on result links
- Important in interpreting user behavior
  - Tabbed browsing accounted for 10.5% of clicks in 2006 study
  - 81% of observed search sequences are ambiguous

# Richer Client Instrumentation

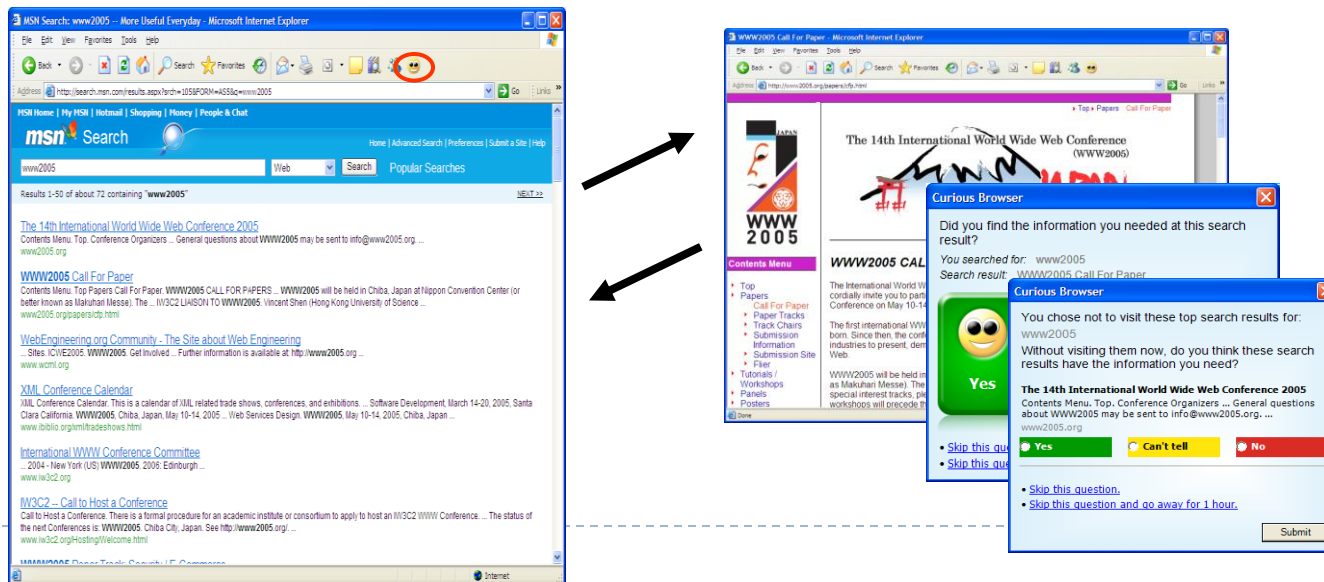
---

- ▶ **Toolbar (or other client code)**
  - ▶ Richer logging (e.g., browser events, mouse/keyboard events, screen capture, eye-tracking, etc.)
  - ▶ Several HCI studies of this type [e.g., Keller et al., Cutrell et al., ...]
  - ▶ Importance of robust software, and data agreements
- ▶ **Instrumented panel**
  - ▶ A group of people who use client code regularly; may also involve subsequent follow-up
  - ▶ Nice mix of *in situ* use (the what) and support for further probing (the why)
  - ▶ E.g., Curious Browser [next slide]
- ▶ **Data recorded on the client**
  - ▶ But still needs to get logged centrally on a server
  - ▶ Consolidation on client possible



# Example: Curious Browser

- ▶ Plug-in to examine relationship between explicit and implicit behavior
  - ▶ Capture lots of implicit actions (e.g., click, click position, dwell time, scroll)
  - ▶ Probe for explicit user judgments of relevance of a page to the Query
- ▶ Deployed to ~4k people in US and Japan
- ▶ Learned models to predict explicit judgments from implicit indicators
  - ▶ 45% accuracy w/ just click; 75% accuracy w/ click + dwell + session
- ▶ Used to learn identify important features, and run model in online evaluation



# Setting Up Server-side Logging

---

- ▶ What to log?
  - ▶ Log as much as possible
  - ▶ But ... make reasonable choices
    - ▶ Richly instrumented client experiments can provide some guidance
    - ▶ Pragmatics about amount of data, storage required will also guide
- ▶ What to do with the data?
  - ▶ The data is a large collection of events, often keyed w/ time
    - ▶ E.g., <time, userID, action, value, context>
  - ▶ Keep as much raw data as possible (and allowable)
  - ▶ Post-process data to put into a more usable form
    - ▶ Integrating across servers to organize the data by time, userID, etc.
    - ▶ Normalizing time, URLs, etc.
    - ▶ Richer data cleaning [Dan, next section]



# Three Important Practical Issues

---

## ▶ Scale

- ▶ Storage requirements
  - ▶ E.g., 1k bytes/record x 10 records/query x 10 mil queries/day = 100 Gb/day
- ▶ Network bandwidth
  - ▶ Client to server
  - ▶ Data center to data center

## ▶ Time

- ▶ **Client time** is closer to the user, but can be wrong or reset
- ▶ **Server time** includes network latencies, but controllable
- ▶ In both cases, need to synchronize time across multiple machines
- ▶ Data integration: Ensure that joins of data are all using the same basis (e.g., UTC vs. local time)
- ▶ Importance: Accurate timing data is critical for understanding sequence of activities, daily temporal patterns, etc.

## ▶ What is a user?

---



# What is a User?

---

- ▶ **Http cookies, IP address, temporary ID**
  - ▶ Provides broad coverage and easy to use, but ...
  - ▶ Multiple people use same machine
  - ▶ Same person uses multiple machines (and browsers)
    - ▶ How many cookies did you use today?
  - ▶ Lots of churn in these IDs
    - ▶ Jupiter Res (39% delete cookies monthly); Comscore (2.5x inflation)
- ▶ **Login, or Download of client code (e.g., browser plug-in)**
  - ▶ Better correspondence to people, but ...
  - ▶ Requires sign-in or download
  - ▶ Results in a smaller and biased sample of people or data (who remember to login, decided to download, etc.)
- ▶ **Either way, loss of data**



# How To Do Log Analysis at Scale?

---

- ▶ MapReduce, Hadoop, Pig ... oh my!
- ▶ What are they?
  - ▶ **MapReduce** is a programming model for expressing distributed computations while hiding details of parallelization, data distribution, load balancing, fault tolerance, etc.
    - ▶ Key idea: partition problem into pieces
    - ▶ Map (input\_key, input\_value) -> list of intermediate values
    - ▶ Reduce (output\_key, intermediate values)
  - ▶ **Hadoop** open-source implementation
  - ▶ **Pig** execution engine on top of Hadoop
- ▶ Why would you want to use them?
  - ▶ Efficient for ad-hoc operations on large-scale data
  - ▶ E.g., Count number words in a large collection of documents
- ▶ How can you use them?
  - ▶ Many universities have compute clusters
  - ▶ Also, Amazon EC3, Microsoft-NSF, and others

```
void map(String name, String document):  
    // name: document name  
    // document: document contents  
    for each word w in document:  
        EmitIntermediate(w, "1");  
  
void reduce(String word, Iterator partialCounts):  
    // word: a word  
    // partialCounts: a list of aggregated partial counts  
    int result = 0;  
    for each pc in partialCounts:  
        result += ParseInt(pc);  
    Emit(AsString(result));
```





# Using the Data Responsibly

---

- ▶ **What data is collected and how it can be used**
  - ▶ User agreements (terms of service)
  - ▶ Emerging industry standards and best practices
- ▶ **Trade-offs**
  - ▶ More data: more intrusive and potential privacy concerns, but also more useful for analysis and system improvement
  - ▶ Less data: less intrusive, but less useful
- ▶ **Risk, benefit, trust**



# Using the Data Responsibly

---

- ▶ **Control access to the data**
  - ▶ Internally: access control; data retention policy
  - ▶ Externally: risky (e.g., AOL, Netflix, Enron, FB public)
- ▶ **Protect user privacy**
  - ▶ Directly identifiable information
    - ▶ Social security, credit card, driver's license numbers
  - ▶ Indirectly identifiable information
    - ▶ Names, locations, phone numbers ... you're so vain (e.g., AOL)
    - ▶ Putting together multiple sources indirectly (e.g., Netflix, hospital records)
      - Linking public and private data
      - *k*-anonymity
- ▶ **Transparency and user control**
  - ▶ Publicly available privacy policy
  - ▶ Giving users control to delete, opt-out, etc.



# Data cleaning for large logs

Dan Russell

# Why clean logs data?

---

- ▶ *The big false assumption:* Isn't logs data intrinsically clean?
  - ▶ A: Nope.



# Typical log format

---

**210.116.18.93** - - [23/Jan/2005:13:37:12 -0800]

"**GET** /modules.php?name=News&file=friend&op=FriendSend&sid=8225 **HTTP/1.1**" **200 2705**  
"http://www.olloo.mn/modules.php?name=News&file=article&catid=25&sid=8225" "**Mozilla/4.0**  
(compatible; MSIE 6.0; **Windows NT 5.1; SV1**)" ...

- **Client IP** - 210.126.19.93
- **Date** - 23/Jan/2005
- **Accessed time** - 13:37:12
- **Method** - **GET** (to request page ), **POST**, **HEAD** (send to server)
- **Protocol** - **HTTP/1.1**
- **Status code** - **200** (Success), **401**, **301**, **500** (error)
- **Size of file** - 2705
- **Agent type** - **Mozilla/4.0**
- **Operating system** - **Windows NT**

<http://www.olloo.mn/modules.php?name=News&file=article&catid=25&sid=8225> →

→ <http://www.olloo.mn/modules.php?name=News&file=friend&op=FriendSend&sid=8225>

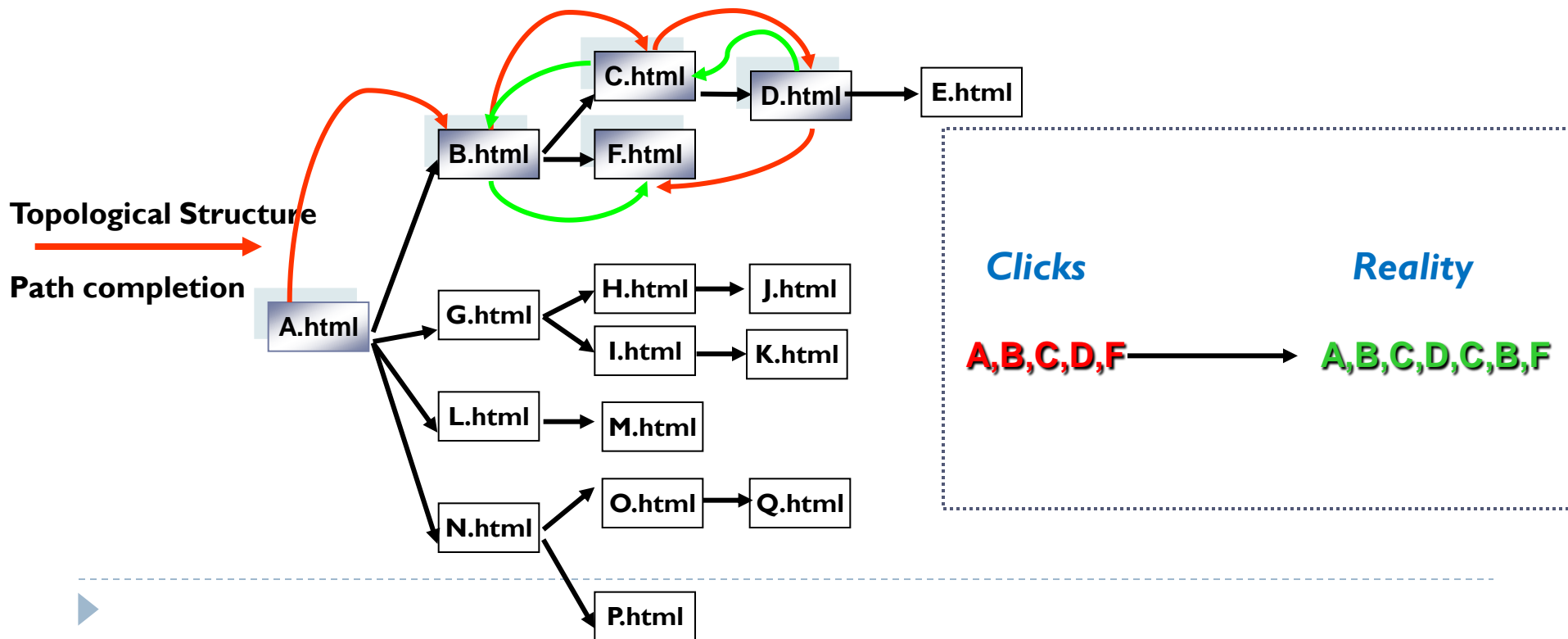
**What this really means...** A visitor (210.126.19.93) viewing the news who sent it to friend.

---



# Sources of noise

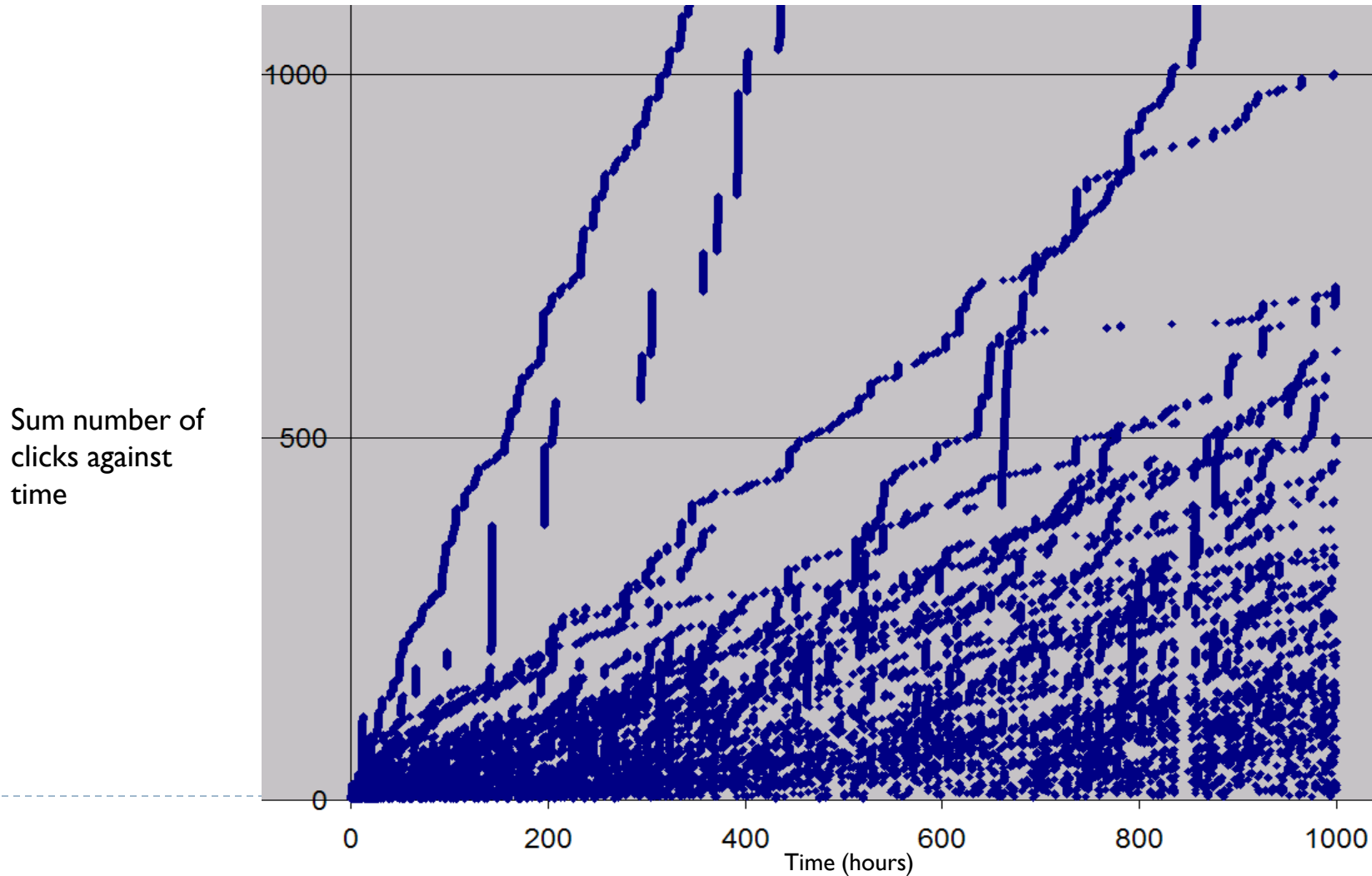
- ▶ Non-completion due to caching (back button)
- ▶ Also... tabs... invisible...
  - ▶ Also – new browser instances.



# A real example

---

- ▶ A previously unknown gap in the data



# What we'll skip...

---

- ▶ Often data cleaning includes
  - (a) input / value validation
  - (b) duplicate detection / removal
    - ▶ We'll assume you know how to do that
  - (c) multiple clocks – syncing time across servers / clients
  
- ▶ But... note that valid data definitions often shift out from under you. (See schema change later)





# When might you NOT need to clean data?

---

## ▶ Examples:

- ▶ When the data is going to be presented in ranks.
  - ▶ Example: counting most popular queries. Then outliers are either really obvious, or don't matter
- ▶ When you need to understand overall behavior for system purposes
  - ▶ Example: traffic modeling for queries—probably don't want to remove outliers because the system needs to accommodate them as well!



# Before cleaning data

---

- ▶ Consider the point of cleaning the data
  - ▶ What analyses are you going to run over the data?
  - ▶ Will the data you're cleaning **damage** or **improve** the analysis?

So...what  
DO I want to  
learn from  
this data?



How about  
we remove  
all the short  
click  
queries?



# Importance of data expertise

---

- ▶ Data expertise is important for understanding the data, the problem and interpreting the results
  - ▶ *Often...background knowledge particular to the data or system:*
    - ▶ “That counter resets to 0 if the number of calls exceeds N”.
    - ▶ “The missing values are represented by 0, but the default amount is 0 too.”
- ▶ Insufficient DE is a common cause of poor data interpretation
- ▶ DE should be documented with the data metadata



# Outliers

---

- ▶ Often indicative either of
  - ▶ measurement error, or that the population has a heavy-tailed distribution.
  - ▶ Beware of distributions with highly non-normal distributions
    - ▶ Be cautious when using tool or intuitions that assume a normal distribution (or, when sub-tools or models make that assumption)
    - ▶ a frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations



# Outliers: Common types from search

---

- ▶ **Quantity:**

- ▶ 10K searches from the same cookie in one day
- ▶ Suspicious whole numbers: *exactly* 10,000 searches from single cookie



# Outliers: Common types from search

---

## ▶ Quantity:

- ▶ 10K searches from the same cookie in
- ▶ Suspicious whole numbers: *exactly* 10,000 searches from the same cookie

Time of day	Query
12:02:01	[ google ]
13:02:01	[ google ]
14:02:01	[ google ]
15:02:01	[ google ]
16:02:01	[ google ]
17:02:01	[ google ]

## ▶ Repeated:

- ▶ The same search repeated over-frequently
- ▶ The same search repeated at the same time (10:01AM)
- ▶ The same search repeated at a repeating interval (every 1000 seconds)



# Treatment of outliers: Many methods

---

- ▶ Remove outliers when you're looking for **average** user behaviors
  - ▶ *Methods:*
    - ▶ Error bounds, tolerance limits – control charts
    - ▶ Model based – regression depth, analysis of residuals
    - ▶ Kernel estimation
    - ▶ Distributional
    - ▶ Time Series outliers
    - ▶ Median and quantiles to measure / identify outliers

---

▶ Sample reference:  
*Exploratory Data Mining  
and Data Quality*, Dasu &  
Johnson (2004)

# Identifying bots & spam

---

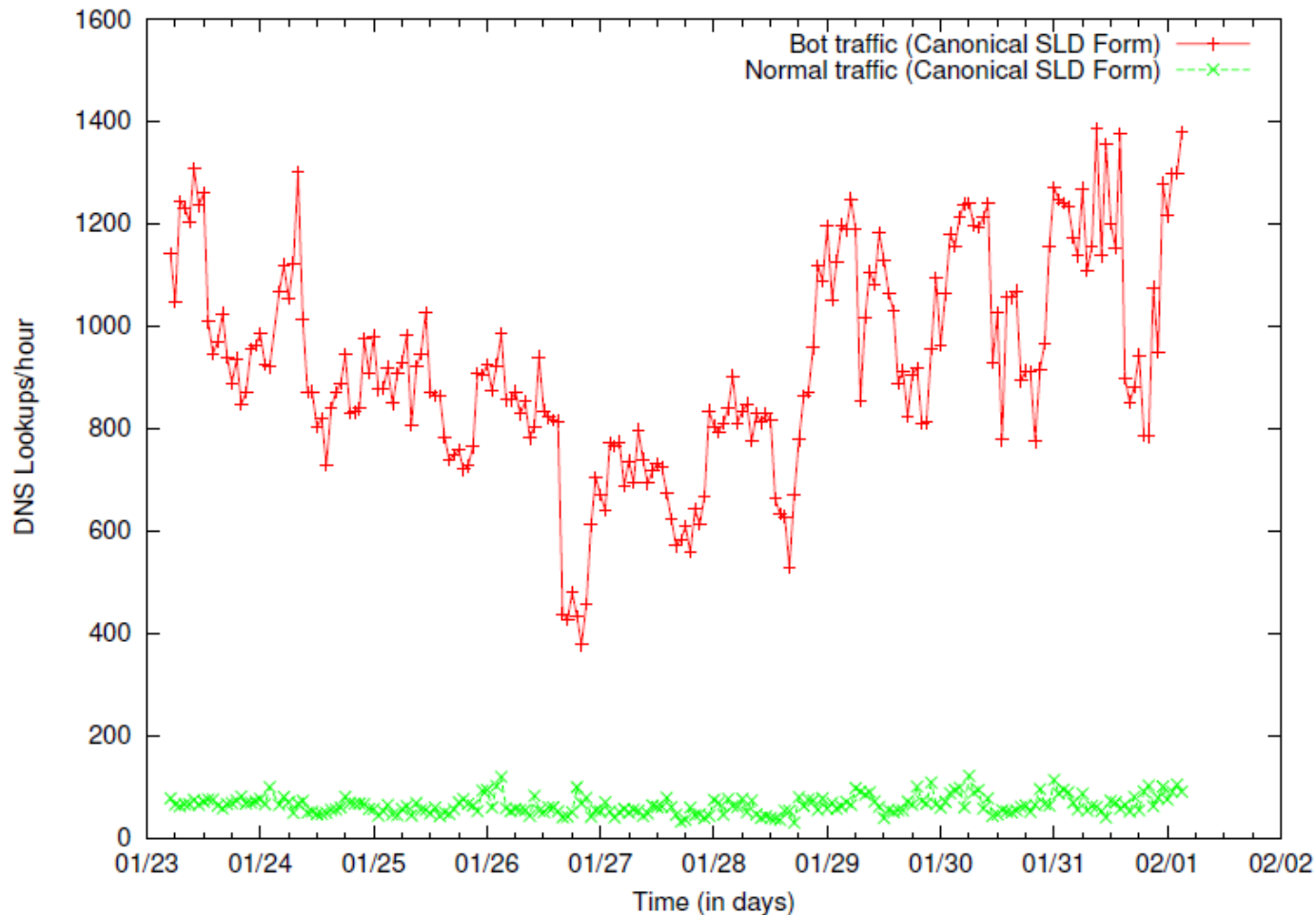
- ▶ Adversarial environment
- ▶ How to ID bots:
  - ▶ Queries too fast to be humanoid-plausible
  - ▶ High query volume for a single query
  - ▶ Queries too specialized (and repeated) to be real
  - ▶ Too many ad clicks by cookie





# Bot traffic tends to have pathological behaviors

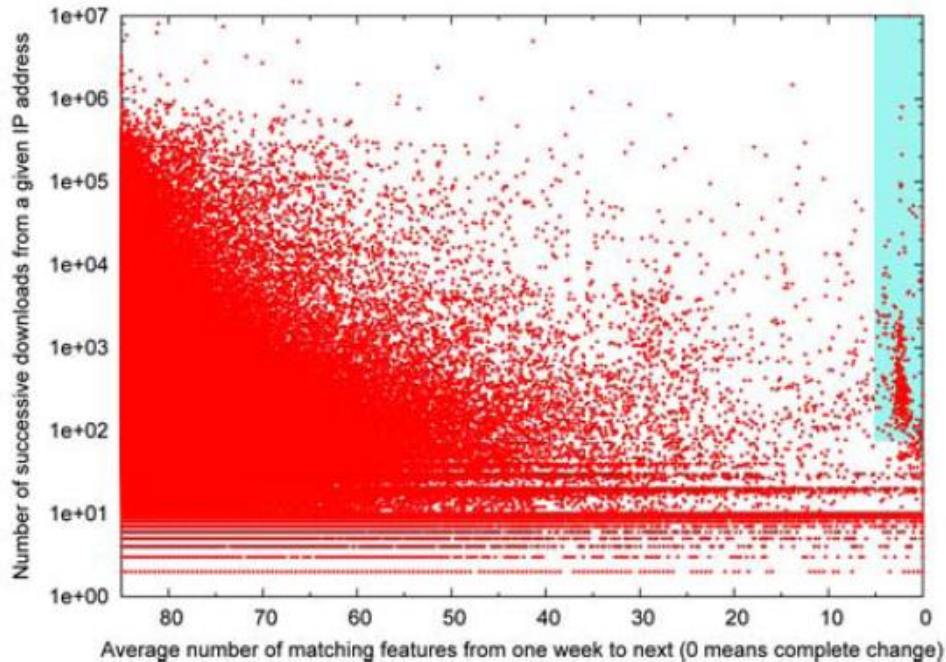
- ▶ Such as abnormally high page-request or DNS lookup rates



# How to ID spam

Spam, Damn Spam, and Statistics: Using statistical analysis to locate spam web pages. D. Fetterly, M. Manasse and M. Najork. *7th Int'l Workshop on the Web and Databases*, June 2004.

- ▶ Look for outliers along different kinds of features
  - ▶ Example: click rapidity, interclick time variability,

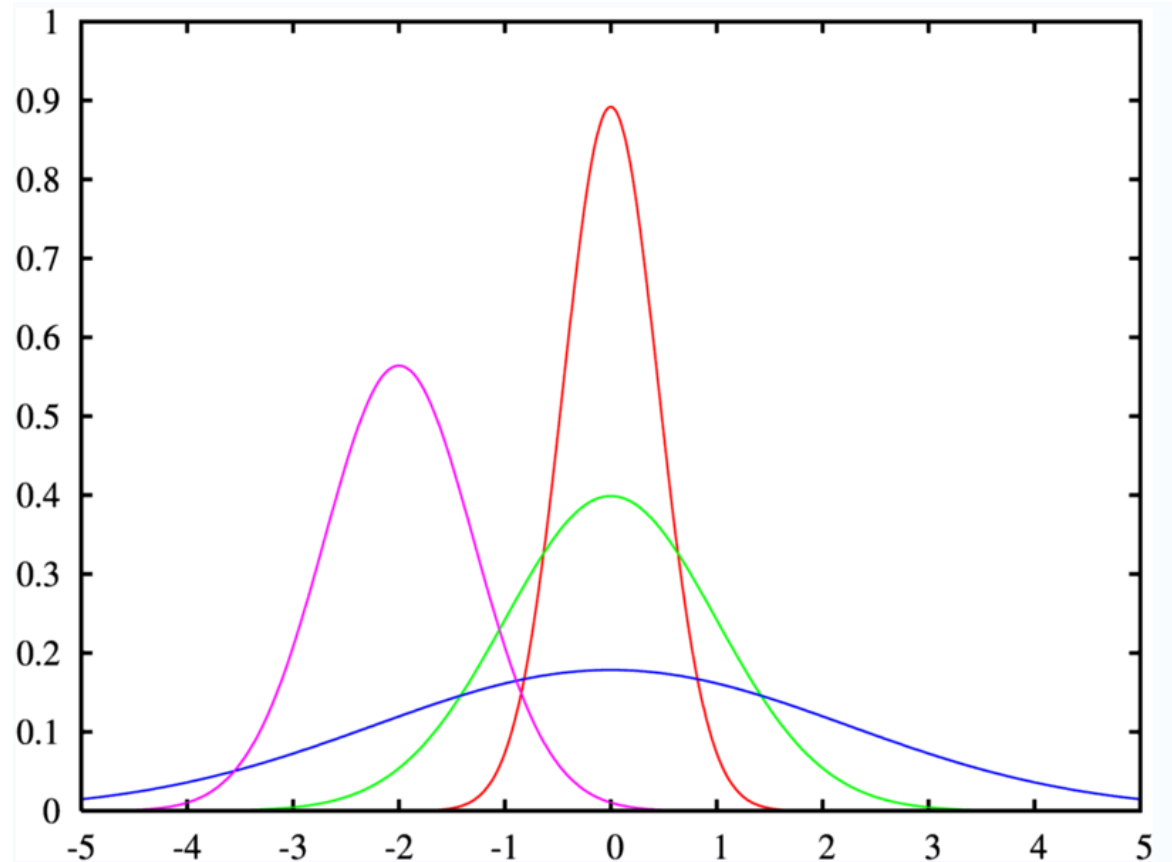


**Spammy sites often change many of their features (page titles, link anchor text, etc.) rapidly week to week**

# Bots / spam clicks look like mixtures

---

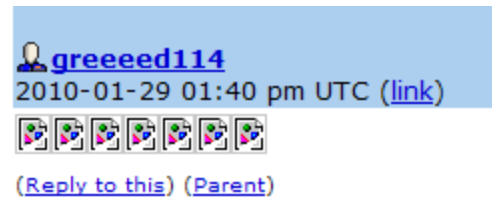
- ▶ Although bots tend to be tightly packed and far from the large mass of data



# Story about spam...

---

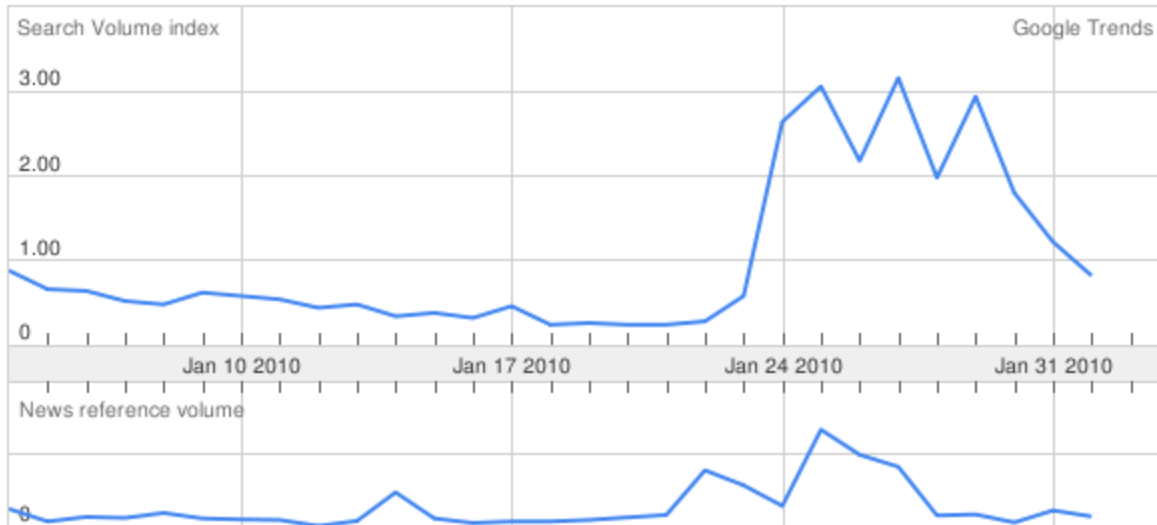
- ▶ 98.3% of queries for [naomi watts] had no click
- ▶ Checking the referers of these queries led us to a cluster of LiveJournal users
- ▶ `img src="http://www.google.ru/search?q=naomi+watts...`
- ▶ What??
- ▶ Comment spam by **greeed114**. No friends, no entries. Apparently trying to boost Naomi Watts on IMDB, Google, and MySpace.



# Did it work?

naomi watts

1.00







No news articles were found

Rank by

## Regions

1. [Russian Federation](#) 
2. [Belarus](#) 
3. [Ukraine](#) 
4. [Latvia](#) 

## Cities

1. Moscow, Russian Federation 
2. St Petersburg, Russian Federation 
3. Novosibirsk, Russian Federation 
4. Yekaterinburg, Russian Federation 



# Cleaning heuristics:

Be sure to account for known errors

---

## ▶ Examples:

### ▶ Known data drops

- ▶ e.g., when a server went down during data collection period – need to account for missing data

### ▶ Known edge cases

- ▶ e.g., when errors occur at boundaries, such as timing cutoffs for behaviors (when do you define a behavior such as a search session as “over”)



# Simple ways to look for outliers

---

- ▶ Simple queries are effective:

```
Select Field, count(*) as Cnt
from Table
Group by Field
Order by Cnt Desc
```

- ▶ Hidden NULL values at the head of the list, typos at the end of the list

- ▶ Visualize your data

- ▶ Often can see data discrepancies that are difficult to note in statistics
- ▶ LOOK at a subsample... **by hand**. (Be willing to spend the time)



## But ultimately...

---

- ▶ Nearly all data cleaning operations are special purpose, one-off kinds of operations

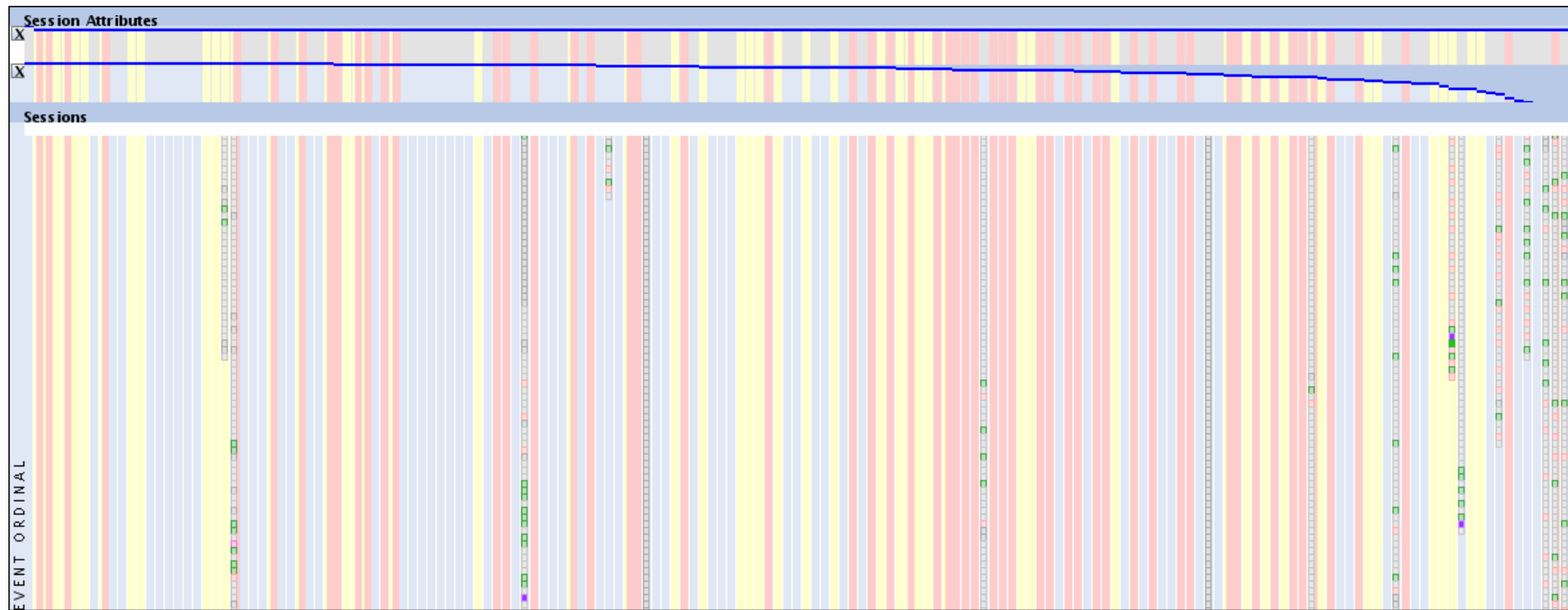




# But ultimately...

---

- ▶ Big hint: Visual representations of the data ROCK!  
Why? Easy to spot all kinds of variations on the data quality that you might not anticipate *a priori*.



## Careful about *skew*, not just outliers

---

- ▶ For example, if an NBA-related query is coming from Wisconsin, search queries are biased by local preferences. Google Trends and Google Insights data shows pretty strong indications of this (look at the Cities entries in either product):
  - ▶ <http://www.google.com/trends?q=Milwaukee+bucks&ctab=0&geo=all&date=all&sort=0>
  - ▶ <http://www.google.com/trends?q=lakers&ctab=0&geo=all&date=all&sort=0>
  - ▶ <http://www.google.com/trends?q=celtics&ctab=0&geo=all&date=all&sort=0>
  - ▶ <http://www.google.com/trends?q=manchester+united&ctab=0&geo=all&date=all>
  - ▶ <http://www.google.com/trends?q=chelsea&ctab=0&geo=all&date=all&sort=0>
  - ▶ <http://www.google.com/insights/search/#q=lakers%2C%20celtics%2Cmilwaukee%20bucks&cmpt=q>
  - ▶ <http://www.google.com/insights/search/#q=arsenal%2Cmanchester%20united%2Cchelsea&cmpt=q>
- ▶ Using this data will generate some interesting correlations. For example, Ghana has a higher interest in Chelsea (because one of the Chelsea players is Ghanaian).
- ▶ Similarly for temporal variations (see Robin's query volume variation over the year)



Tip: Use commas to compare multiple search terms.

Searches Websites

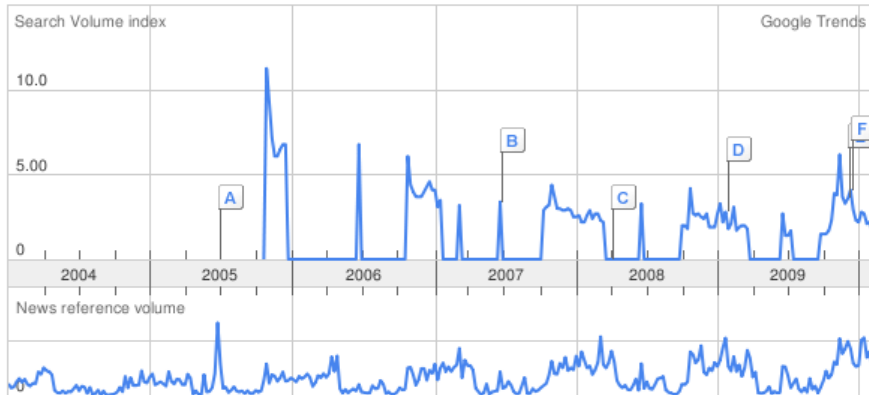
All regions

All years

Scale is based on the average worldwide traffic of **milwaukee bucks** in all years. [Learn more](#)

milwaukee bucks

1.00



- A** [Top pick visits Milwaukee Bucks, says he's glad draft is over](#)  
Duluth News Tribune - Jun 29 2005
- B** [Milwaukee Bucks' image at stake after selecting Yi Jianlian at NBA draft](#)  
CBC News - Jun 29 2007
- C** [New Jersey Nets \(32-47\) at Milwaukee Bucks \(26-53\), 8:30 pm](#)  
TheNewsTribune.com - Apr 12 2008
- D** [Milwaukee Bucks Lose Michael Redd For The Rest Of The Season](#)  
Bleacher Report - Jan 25 2009
- E** [Brandon Jennings scores 22 to lift Milwaukee Bucks over Toronto Raptors](#)  
Appleton Post Crescent - Dec 10 2009
- F** [Kobe Bryant hits another game winner against the Milwaukee Bucks](#)  
Examiner.com - Dec 17 2009

[More news results »](#)

Rank by

Regions	Cities	Languages
1. <a href="#">United States</a>	1. Brookfield, WI, USA	1. English
2. <a href="#">Australia</a>	2. Cedarburg, WI, USA	2. Turkish
3. <a href="#">Canada</a>	3. New Berlin, WI, USA	3. Chinese
4. <a href="#">Spain</a>	4. Waukesha, WI, USA	4. Spanish
5. <a href="#">China</a>	5. Hawkins, WI, USA	5. French
6. <a href="#">United Kingdom</a>	6. Milwaukee, WI, USA	6. German
	7. Fond Du Lac, WI, USA	
	8. Green Bay, WI, USA	
	9. Stevens Point, WI, USA	
	10. Appleton, WI, USA	

# Pragmatics

---

- ▶ **Keep track of what data cleaning you do!**
  - ▶ Add lots of metadata to describe what operations you've run (It's too easy to do the work, then forget which cleaning operations you've already run.)
    - ▶ Example: data cleaning story from ClimateGate –only the cleaned data was available...
  - ▶ Add even more metadata so you can interpret this (clean) data in the future.
    - ▶ **Sad story:** I've lost lots of work because I couldn't remember what this dataset was, how it was extracted, or what it meant... as little as 2 weeks in the past!!



# Pragmatics

---

- ▶ **BEWARE** of truncated data sets!
  - ▶ All too common: you think you're pulling data from Jan 1, 20??
    - Dec 31, 20??, but you only get Jan 1 – Nov 17
- ▶ **BEWARE** of censored / preprocessed data!
  - ▶ Example: Has this data stream been cleaned-for-safe-search before you get it?
    - ▶ **Story:** Looking at queries that have a particular UI treatment. (Image universal triggering) We noticed the porn rate was phenomenally low. Why? Turns out that this UI treatment has a porn-filter BEFORE the UI treatment is applied, therefore, the data from the logs behavior was already implicitly run through a porn filter.



# Pragmatics

---

- ▶ **BEWARE** of capped values
  - ▶ Does your measuring instrument go all the way to 11?
  - ▶ Real problem: time on task (for certain experiments) is measured only out to  $X$  seconds. All instances that are  $> X$  seconds are either recorded as  $X$ , or dropped. (Both are bad, but you need to know which data treatment your system follows.)
    - ▶ This seems especially true for very long user session behaviors, time-on-task measurements, click duration, etc.
  - ▶ Metadata should capture this
  - ▶ **Note:** big spikes in the data often indicate this kind of problem



# Pragmatics

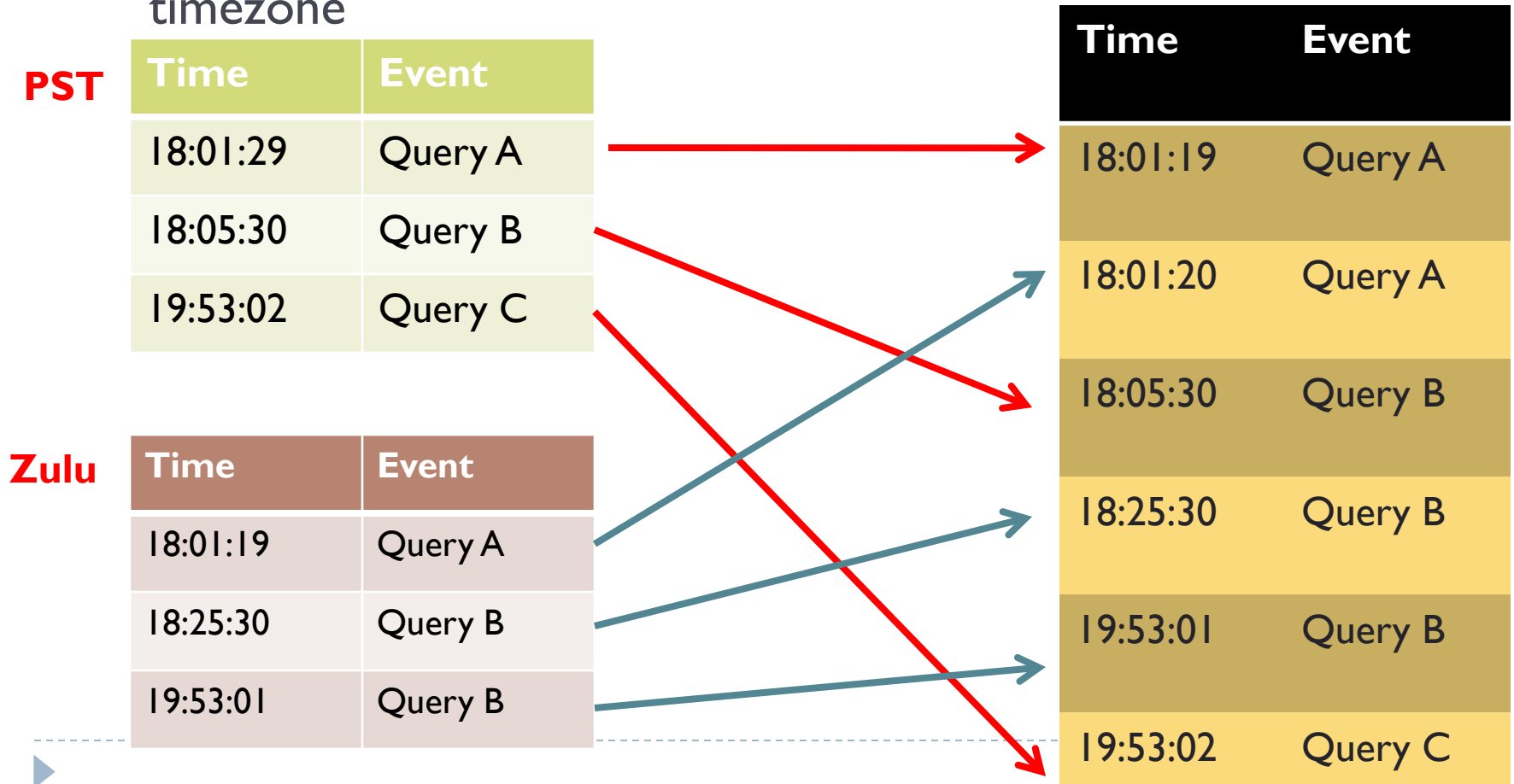
---

- ▶ **Do sanity checks constantly**
  - ▶ Don't underestimate their value.
  - ▶ Right number of files? Roughly the right size? Expected number of records?
  - ▶ Does this data trend look roughly like previous trends?
  - ▶ Check sampling frequency (Are you using downsampled logs, or do you have the complete set?)



# Data integration

- ▶ Be sure that joins of data are all using the same basis
  - ▶ e.g., time values that are measured consistently – UTC vs. local timezone





# Data Cleaning Summary

---

- ▶ **CAUTION:** Many, many potholes to fall into
- ▶ **Know** what the purpose of your data cleaning is for
- ▶ **Maintain** metadata
- ▶ **Beware** of domain expertise failure
- ▶ **Ensure** that the underlying data schema is what you think it is



# Section 8: Log Analysis and the HCI Community

All

# Discussion: Log Analysis and HCI

---

- ▶ Is log analysis relevant to HCI?
- ▶ How to present/review log analysis research
  - ▶ Observational
  - ▶ Experimental
- ▶ How to generate logs
- ▶ Sources of log data



# Is Log Analysis Relevant to HCI?

---

- ▶ “Know thy user”
  - ▶ *In situ* large-scale log provide unique insights
  - ▶ Real behavior
- ▶ What kinds of things can we learn?
  - ▶ Patterns of behavior (e.g., info seeking goals)
  - ▶ Use of systems (e.g., how successful are people in using the current vs. new system)
  - ▶ Experimental comparison of alternatives



# How to Present/Review Log Analysis

---

- ▶ **Examples of successful log analysis papers**
  - ▶ Several published logs analysis of observational type
  - ▶ But fewer published reports of the experimental type
- ▶ **Determining if conclusions are valid**
  - ▶ Significance unlikely to be a problem
  - ▶ Data cleanliness important
  - ▶ Only draw supported claims (careful with intent)



# How to Generate Logs

---

- ▶ **Use existing logged data**
  - ▶ Explore sources in your community (e.g., proxy logs)
  - ▶ Work with a company (e.g., intern, visiting researcher)
  - ▶ Construct targeted questions
- ▶ **Generate your own logs**
  - ▶ Focuses on questions of unique interest to you
- ▶ **Construct community resources**
  - ▶ Shared software and tools
    - ▶ Client side logger (e.g., VIBE logger)
  - ▶ Shared data sets
  - ▶ Shared experimental platform to deploy experiments (and to attract visitors)
  - ▶ Other ideas?



# Interesting Sources of Log Data

---

- ▶ Anyone who runs a Web services
- ▶ Proxy (or library) logs at your institution
- ▶ Publically available social resources
  - ▶ Wikipedia (content, edit history)
  - ▶ Twitter
  - ▶ Delicious, Flickr
  - ▶ Facebook public data?
- ▶ Others?
  - ▶ GPS
  - ▶ Virtual worlds
  - ▶ Cell call logs

