CLrowd Logger: a platform for conducting remote web interaction studies

Henry Feild
Endicott College

November 15, 2013

James Allan

ENDICOTT COLLEGE

UMASS AMHERST

Things we like to do in IR

- Observe and model user behavior

Modeling and Measuring the Impact of Short and Long-Term Behavior on Search Personalization

Personalization of Search Results Using Interaction Behaviors in Search Sessions

Improving Searcher Models Using Mouse Cursor Activity

Search, Interrupted: Understanding and Predicting Search Task Continuation

User Evaluation of Query Quality

- Compare search algorithms / interfaces
  - which do users prefer?
  - time to completion
  - which result in more/fewer clicks, etc.

Absence time and user engagement: Evaluating Ranking Functions

Optimized Interleaving for Online Retrieval Evaluation

What's currently done client-side

1. build toolbar

2. run study
   - recruit participants via fliers, classes, etc.
     - lab studies
     - in situ (install at home)
   - install on campus computers
     - free recruitment, but library-biased

This is slow, expensive, and generally a lot of effort

What we want

- a common, open source platform that deals with the basics
  - interaction data collection
  - data storage
  - privacy
- a common user base
  - can recruit some new users, but already have a significant pool of participants
- an interface for implementing novel studies

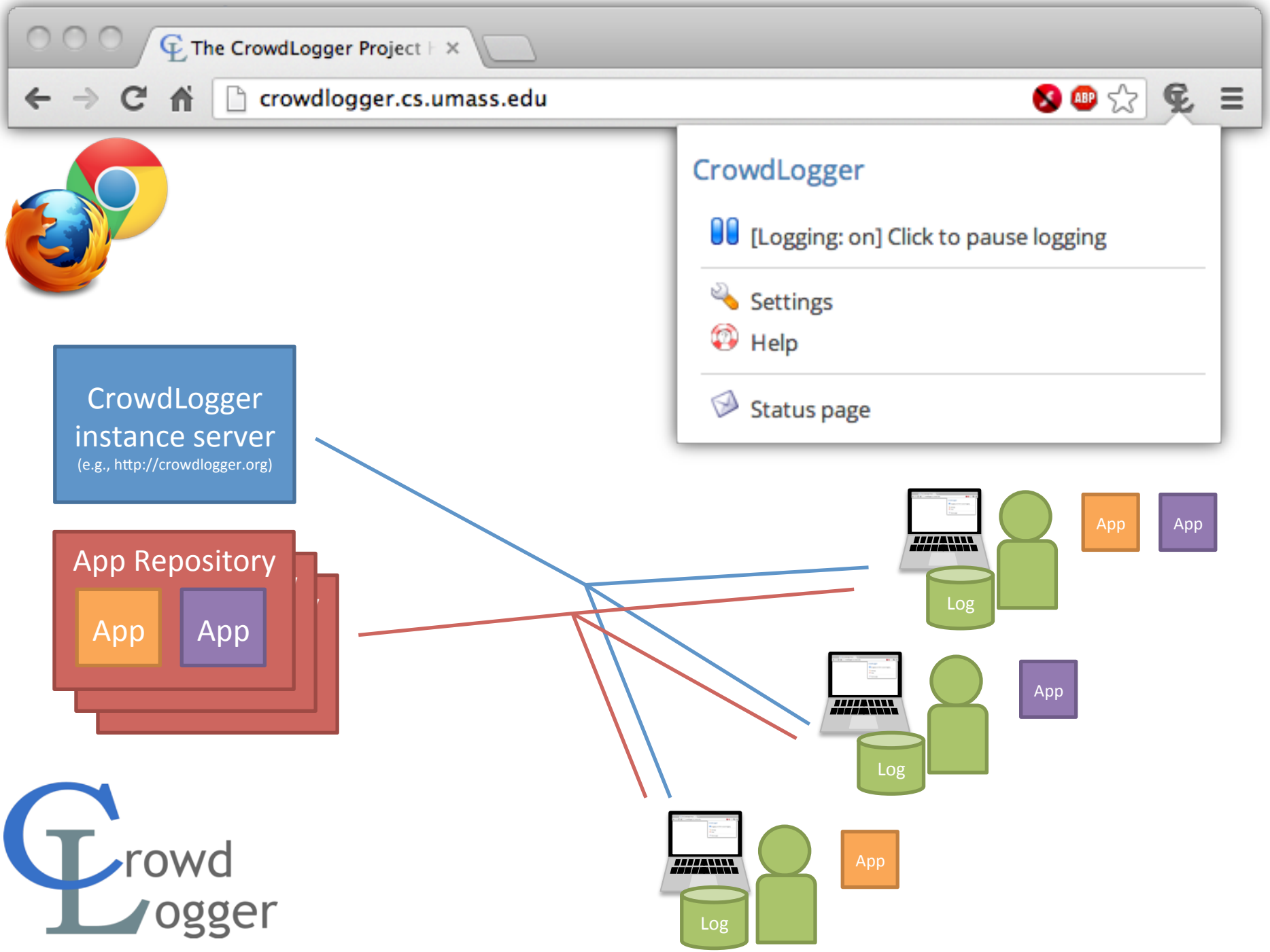
What is CrowdLogger?

CrowdLogger in action

Cross your fingers!

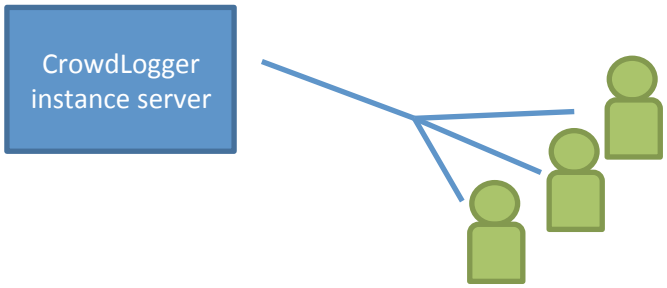
…don't worry, I have screenshots just in case…☺

Issues / Next steps

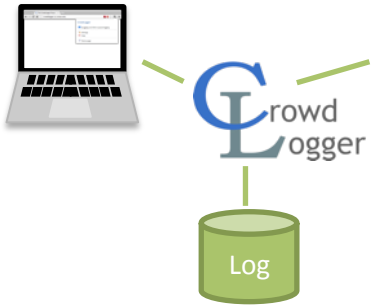

The CrowdLogger Project
crowdlogger.cs.umass.edu
CrowdLogger
[Logging: on] Click to pause logging
Settings
Help
Status page
CrowdLogger instance server
(e.g., http://crowdlogger.org)
App Repository
App
App
Log
App
App
Log
App
Log
App
Crowd Logger

Advantages

User Base = Participant Pool

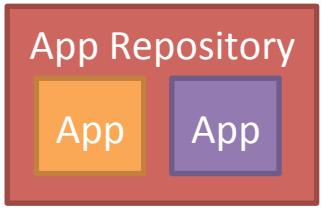


Rich API



Allows apps to access current and historical browsing behavior, store data, interact with the user, upload data privately, and more!

Apps can be developed by anyone

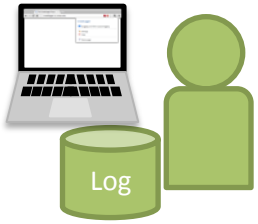


And distributed from a private repository, or a CrowdLogger server instance.

Open Source

The entire code base is available from our Google Project page:
https://code.google.com/p/crowdlogger/

Users' data logged locally



If an app wants to upload data somewhere else, they need permission from the user.

Multiple apps at the same time

CrowdLogger is somewhat akin to a smart phone or tablet—it's an extensible, general-purpose platform with a convenient API.

API Categories

User Data

Historical data

- *get all clicks*
- *get all searches*

Real time data

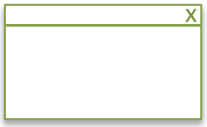
- *on new search, do …*

Aggregate User Data*

Already collected data

- *get all query rewrites*
- *get all query-click pairs*

User Interface



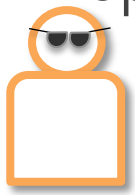
Modify web pages

- *inject JavaScript into pages**

Stand-alone windows/pages

- *present dialog when user searches*
- *modify search page ranking*

Uploading/Privacy



Encryption

Anonymization & aggregation

- *upload via anonymizers*
- *privately aggregate data*

Client-server communication

Request server-side computation

- *run retrieval algorithm for query*

Access server-side data

- *send me synonyms for …*

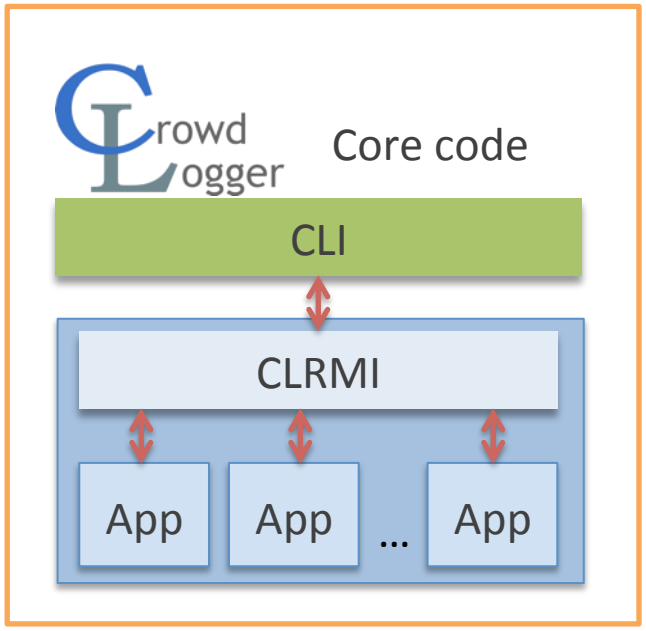
Local data storage

Save data locally

- *settings*
- *models*

CrowdLogger Remote Modules

Also called: **CLRMs** or **Apps**



Parts of an App:

Core files

Set of JavaScript files that are run continuously in the background

Resources

HTML, JavaScript, CSS, images*, etc.

App Example—JSON

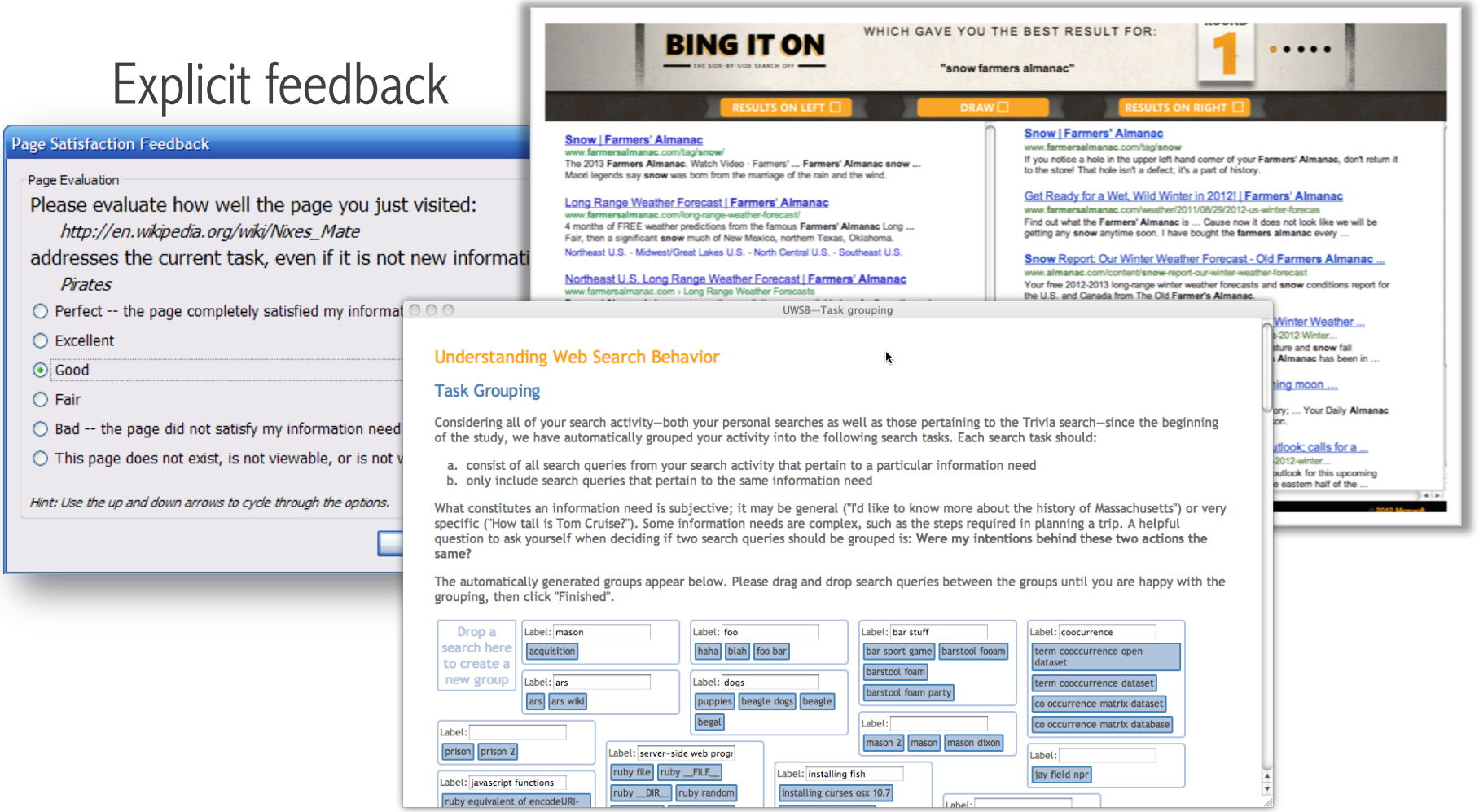

```
{
    "metadata": {
        "clrmid": "demo1",
        "name": "Demo 1",
        "version": "1.6",
        "categories": [
            "app"
        ],
        "description": "A simple CrowdLogger Remote Module demonstration. It shows off the ability to
        open windows, access user data, and use the storage API.",
        "packageURL": "http://localhost/~hfeild/test-clrms/demo/demo.json",
        "permissions": [
            "userdata",
            "ui",
            "storage"
        ],
        "logoURL": http://localhost/~hfeild/test-clrms/demo/demo.jpg
    },
    "module": "\nvar RemoteModule = function( clrmPackage, clrmAPI ){\n    var demo,\n        that =
    this;\n\n    this.init = function(){\n        demo = new that.Demo( clrmPackage, clrmAPI );…",
    "html": {
        "configure-demo.html": "<html>\n<head>\n<title>CLRM Configure Demo</title>…
    },
    "js": {
        "init.js": "\n// This gets set by our opener. It's our connection to the core CLRM…
    }
    "css": {
        "demo.css": "\nbody {\n    background: white;\n    color: #555;\n}\n\n#interactions, #searches,
        #user-input {\n    width: 100%;\n    height: 20%;\n    overflow: auto;\n    background-color:
        #ddd;\n    border: 1px solid #888;\n}\n\n.notice {\n    display: none;\n}\n\n.success {\n
        color: green;\n}\n\n.error {\n    color: red;\n}"
    },
    "misc": {}
}
```

CrowdLogger

Interactive experiments

System comparisons

Explicit feedback

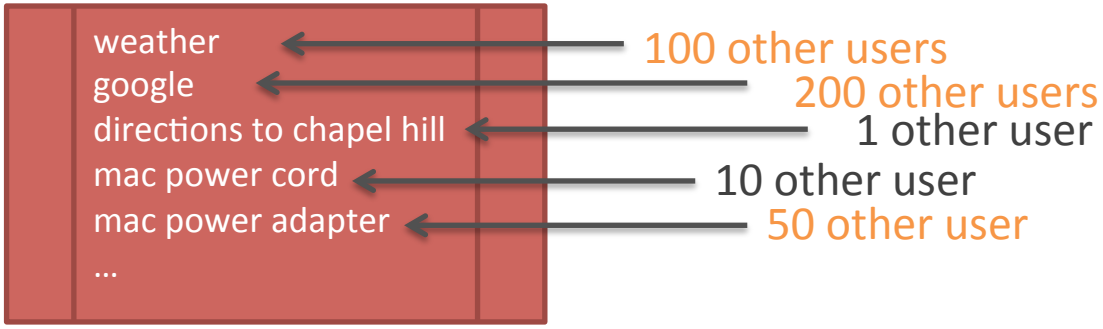


Labeling

Privacy

- Uploading data (study-specific)
- Collecting aggregated data (via the API)

k-Anonymity



If mining queries and k=20, only the ones in orange are revealed

Only in aggregated, no two pieces of information revealed separately are ever tied together

Differential Privacy

$$Pr[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \cdot Pr[\mathcal{A}(D') \in S] + \delta$$

Says: we shouldn't be able to tell if a user's data was or was not part of the dataset based on what is released

What is CrowdLogger?

CrowdLogger in action

Cross your fingers!

…don't worry, I have screenshots just in case…☺

Issues / Next steps

version 2.0.2
CrowdLogger

Status page

Recent Messages

There are no new messages of 1 total messages.
The most recent message is from April 2 ("Welcome to CrowdLogger!").
Click here to see all messages.

Tools

Settings
Search histogram
Clear entire log
Register
View browsing log
Install apps & studies

Apps

Demo 1

Version 1.5
This is just a demonstration message, please ignore.

Open | Configure | Disable | Remove

Search Task Assistant

Enable | Remove

Studies

A questionnaire is ready for you to fill out.
Click the "Open" button to fill it out.
Thanks!

Open | Configure | Disable | Remove

Feedback

If you have any comments or suggestions about the CrowdLogger system that you would like to let us know, you can email us at info@crowdlogger.org or leave an anonymous message at this web page.

Copyright © 2013 The CrowdLogger Project. « University of Massachusetts « All rights reserved. « BSD Liscense

version 2.0.2
CrowdLogger

Settings

There are not many settings just yet. More will show up as we further develop CrowdLogger.

App/Study repositories

Below are listed the repositories currently loaded.

```
http://localhost/~hfeild/cl-test/html/clrms/clrms.php
http://localhost/~hfeild/test-clrms/clrms.php
```

Add

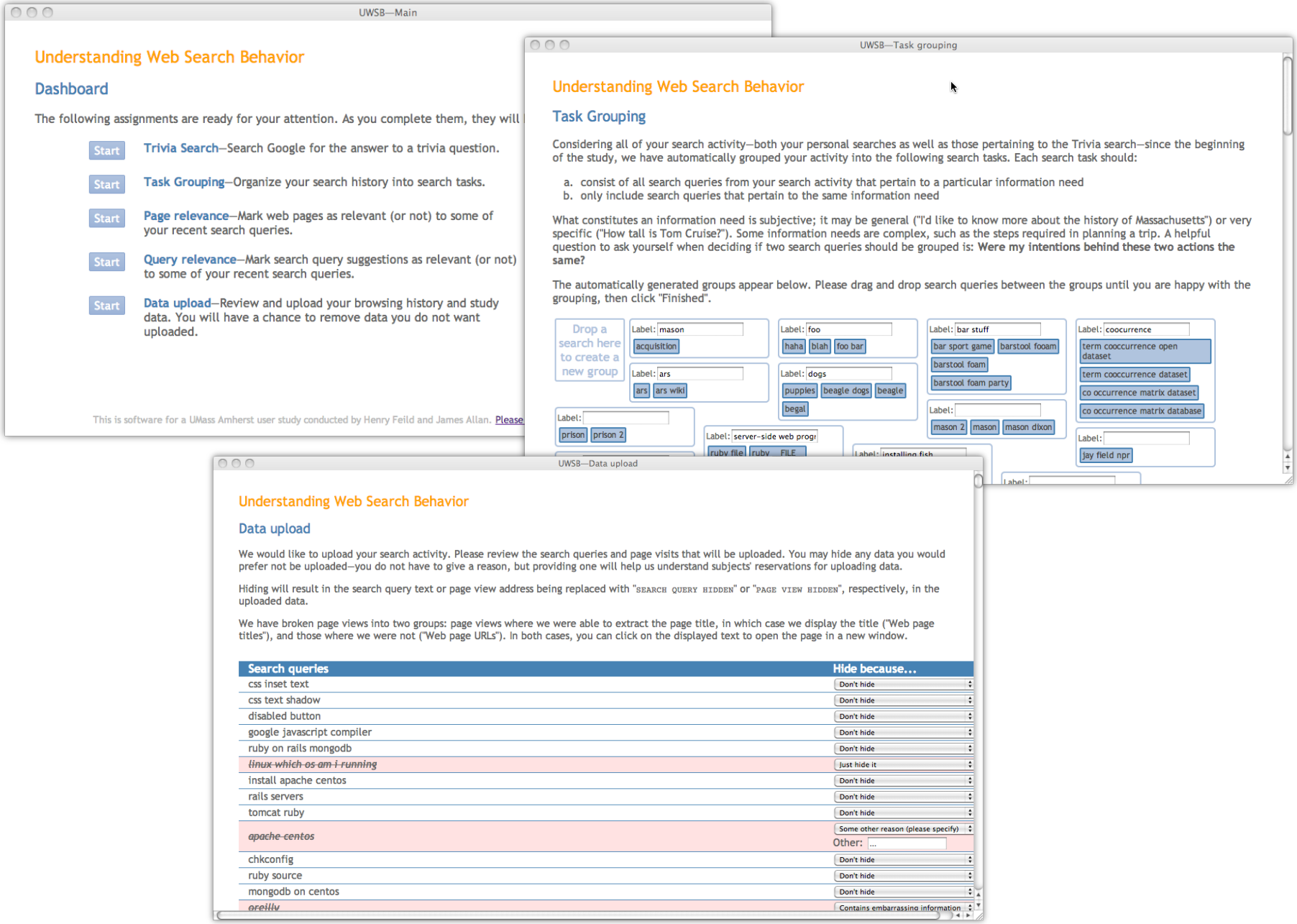
Developer mode

Adds a link to the Welcome Wizard in the main menu and a link to developer tools in the "Tools" section of the status page.

On
Off

Save | Save and close

Copyright © 2013 The CrowdLogger Project. « University of Massachusetts « All rights reserved. « BSD Liscense


UWSB—Main
Understanding Web Search Behavior
Dashboard
The following assignments are ready for your attention. As you complete them, they will
Start
Trivia Search—Search Google for the answer to a trivia question.
Start
Task Grouping—Organize your search history into search tasks.
Start
Page relevance—Mark web pages as relevant (or not) to some of your recent search queries.
Start
Query relevance—Mark search query suggestions as relevant (or not) to some of your recent search queries.
Start
Data upload—Review and upload your browsing history and study data. You will have a chance to remove data you do not want uploaded.
This is software for a UMass Amherst user study conducted by Henry Feild and James Allan. Please
UWSB—Task grouping
Understanding Web Search Behavior
Task Grouping
Considering all of your search activity—both your personal searches as well as those pertaining to the Trivia search—since the beginning of the study, we have automatically grouped your activity into the following search tasks. Each search task should:
a. consist of all search queries from your search activity that pertain to a particular information need
b. only include search queries that pertain to the same information need
What constitutes an information need is subjective; it may be general ("I'd like to know more about the history of Massachusetts") or very specific ("How tall is Tom Cruise?"). Some information needs are complex, such as the steps required in planning a trip. A helpful question to ask yourself when deciding if two search queries should be grouped is: Were my intentions behind these two actions the same?
The automatically generated groups appear below. Please drag and drop search queries between the groups until you are happy with the grouping, then click "Finished".
Drop a search here to create a new group
Label: mason
acquisition
Label: foo
haha
blah
foo bar
Label: bar stuff
bar sport game
barstool fooam
barstool foam
barstool foam party
Label: coocurrence
term cooccurrence open dataset
term cooccurrence dataset
co occurrence matrix dataset
co occurrence matrix database
Label: ars
ars
ars wiki
Label: dogs
puppies
beagle dogs
beagle
begal
Label:
prison
prison 2
Label: server-side web progr
Label:
mason 2
mason
mason dixon
Label:
jay field npr
UWSB—Data upload
Understanding Web Search Behavior
Data upload
We would like to upload your search activity. Please review the search queries and page visits that will be uploaded. You may hide any data you would prefer not be uploaded—you do not have to give a reason, but providing one will help us understand subjects' reservations for uploading data.
Hiding will result in the search query text or page view address being replaced with "SEARCH QUERY HIDDEN" or "PAGE VIEW HIDDEN", respectively, in the uploaded data.
We have broken page views into two groups: page views where we were able to extract the page title, in which case we display the title ("Web page titles"), and those where we were not ("Web page URLs"). In both cases, you can click on the displayed text to open the page in a new window.
Search queries
Hide because...
css inset text
Don't hide
css text shadow
Don't hide
disabled button
Don't hide
google javascript compiler
Don't hide
ruby on rails mongodb
Don't hide
linux which os am i running
Just hide it
install apache centos
Don't hide
rails servers
Don't hide
tomcat ruby
Don't hide
apache centos
Some other reason (please specify)
Other: ...
chkconfig
Don't hide
ruby source
Don't hide
mongodb on centos
Don't hide
oreilly
Contains embarrassing information

What is CrowdLogger?

CrowdLogger in action

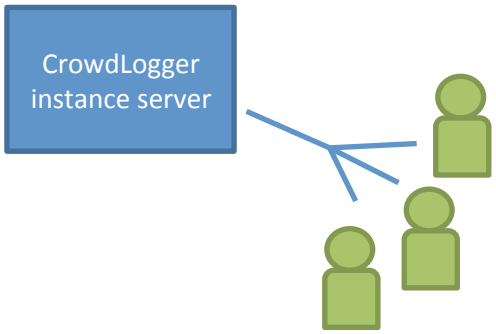
Cross your fingers!

…don't worry, I have screenshots just in case…☺

Issues / Next steps

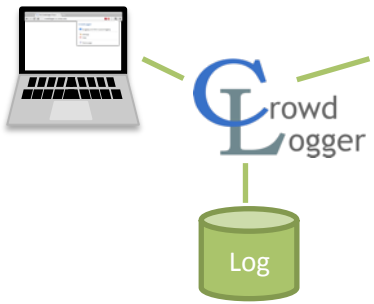
Challenges/Future work

Amassing a large user base



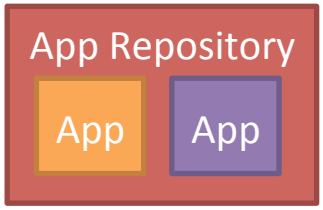
How do we attract and retain users?

Complete/Extend API



There are still a few API classes that we have yet to implement, such as global aggregation of data. There are also many things we can and probably should add to improve utility.

Simplifying app development



We'd like to make it easy for research groups with minimal programming skills to build and deploy apps

Attracting developers

It's good for the development process to be overseen by more than one pair of eyes. This will make the code more maintainable and also more secure.

Logging across more browsers

Right now we only support two of the leading browsers. It would be nice to extend CrowdLogger to IE, Safari, Opera, and others.

Handling multi-apps environments

What happens if two studies are running concurrently, and both modify the browsing UI?

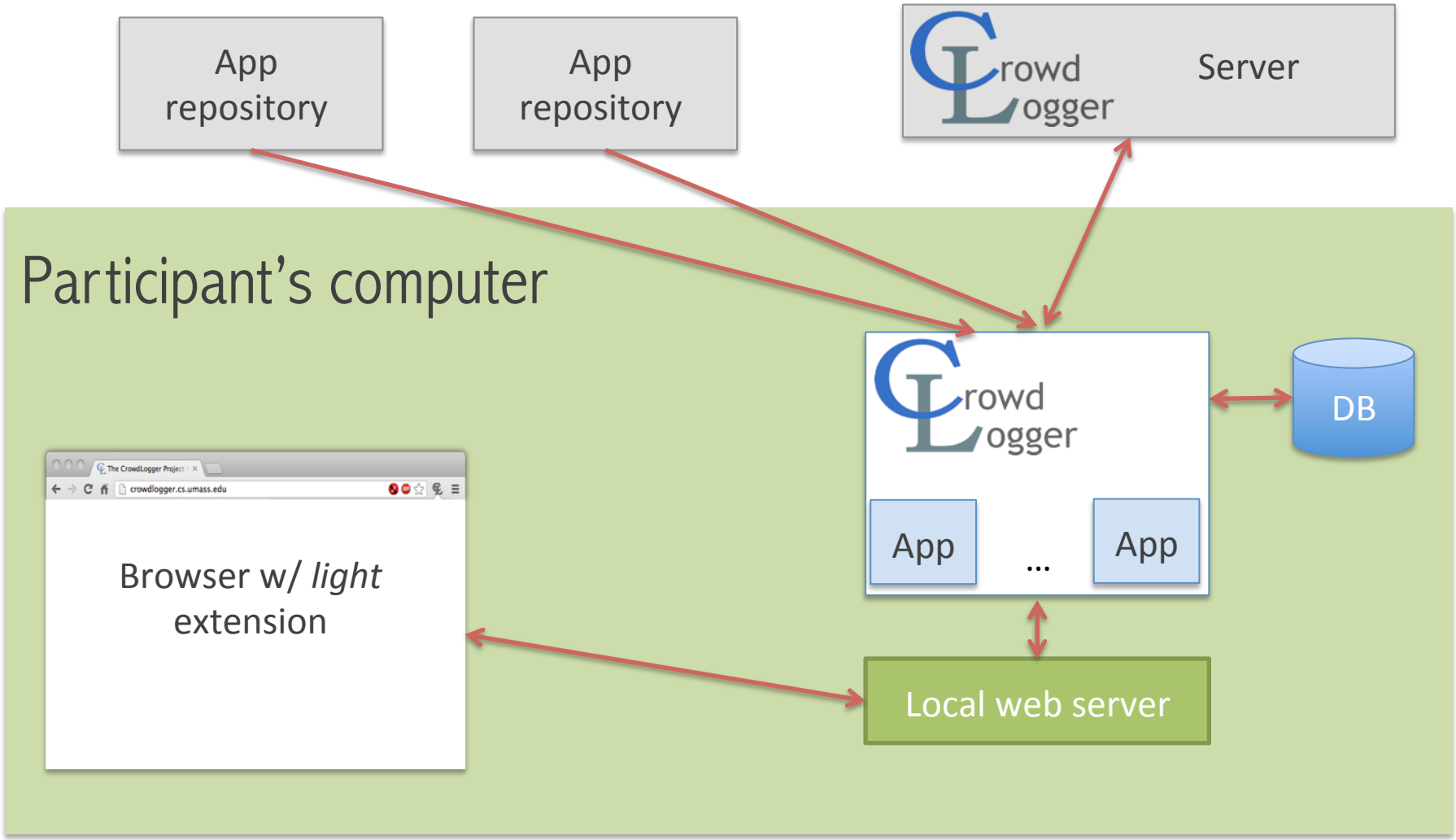
App builder

[SEE DEMO]

Good for:

- Starting from existing apps
- Rapid development
- Less messing around with the nitty gritty
- Research groups without technical support/ programming skills

Local web server



Chrome extension installation

"Google to block local Chrome extensions on Windows starting in January, limit installs to the Chrome Web Store"

…uh oh…

Solutions:

- get CrowdLogger approved for inclusion in Google Web Store
- implement local server model
  - get light extension approved for inclusion in Google Web Store
- release a modified version of Chromium (open source Chrome) for our Chrome users

http://thenextweb.com/google/2013/11/07/google-block-local-chrome-extensions-windows-starting-january-limit-installs-chrome-web-store/

Diverse privacy controls

What will be collected:

All search reformulations. For example, if you search for "blueberry pie" and then "blueberry pie recipes", the pair:

"blueberry pie", "blueberry pie recipes"

will be collected.

How the collected data will be used:

Reformulations will be anonymized and made **publically accessible** and used to, for example, generate search suggestions for you and other users.

Privacy settings:

For each search reformulation collected from you, select the anonymization level: the number of other users that must also share the same reformulation for it to be included in the final data set: 4

☐ I have read the consent form and agree to participate in this study.

Cancel Continue

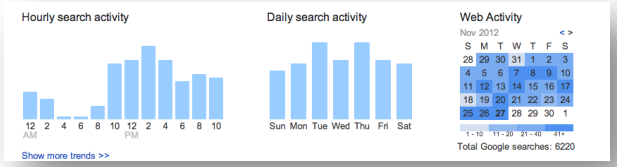
Diverse privacy controls

- what data get's shared with researchers?
- under what conditions?

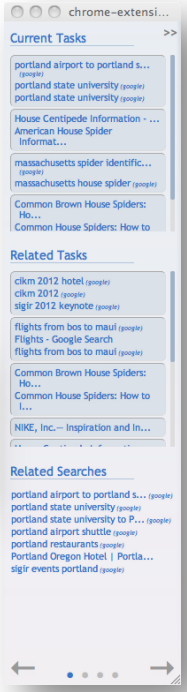
What data is being collected and how it will be used	What is minimally useful to researchers	What users are comfortable with
Query rewrites for public release	Whatever users are comfortable with	**User 1:** only if shared by 9+ other users (k=10) **User 2:** k=1 rewrites
Feedback on retrieval system preference for researcher use only	k=1 anonymized feedback from users	~~**User 1:** k=5 feedback~~ **User 2:** k=1 feedback

Incentivization

Provide a service
- research prototypes
- visualizations
- re-finding tools
- citizen scientist



Google Search History

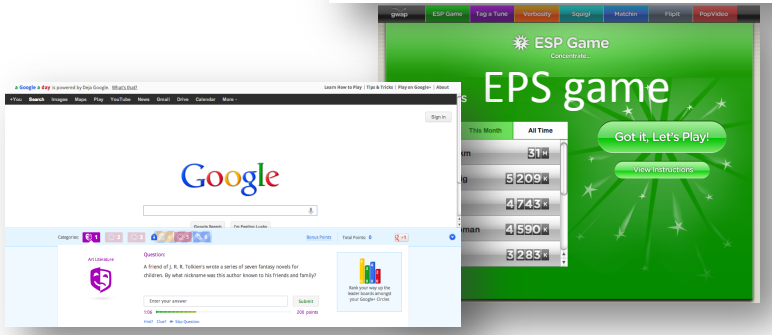


Search Task Assistant

Financial incentives
- gift cards
- virtual currency to 'buy' research apps

Gamification
- study-specific
- could also be a service

YAHOO! ANSWERS



Google-a-day

Thanks!

- If you'd like to help develop, let me know!
- Hank Feild: hfeild@endicott.edu
- CrowdLogger:
  - Instance server: http://crowdlogger.org
  - Git repo: https://code.google.com/p/crowdlogger/
  - Google group: https://groups.google.com/forum/#!forum/crowdlogger-project-news