Introduction

- Interaction log analysis can circumvent these problems
  - Can study larger populations so wider range of uses
  - "Ecological validity", no interference from direct observation

- Specifically look at mouse interactions
  - Substitute for eye-tracking
  - More information than what software features used

Related Work

- "Clickstream interactive research"
  - What users click to navigate webpages

- Action log analysis
  - Sequences of basic software interactions eg filter, sort, select
  - Hand-coding interactions
    - Applied to a similar tool in this paper

Tool: MAGI

- (i) Aberration view
  - Pattern of mutations in gene sets across tumors

- (ii) Aberration view row/heat maps
  - Show gender, survival, purity

- (iv) Network view
  - Interaction between gene

- (iii) Heat map
  - User uploads, e.g. shows methylation for different tumors

Step 1: Task Identification with MAGI Creators

- 2 participants who created MAGI
- Randomly sampled logs from MAGI users
- 25 tasks labelled per participants, with free text
- Based on a separate vis of the log data

Then grouped these descriptions into 8 separate task categories (in a few slides)

Solution

- Let's collect logs from many real world users
- Let's then use machine learning to automatically classify those logs to understand use patterns etc.

Super Brief Domain Background

- DNA is the code of our cells
- Cancer results from bugs in code (mutation)
- Cancer to mutation == many to many
- Mutations relevant for diagnosis, treatment

Tool Being Studied: MAGI

- Online visualization tool
- Cancer genomics
  - Investigate DNA mutations associated with cancers
  - Users: from wet lab biologists to pharmaceutical researchers

MAGI: Who/What/Why/How

- Who:
  - Cancer researchers in wet/dry labs, industry

- What:
  - DNA mutations present in cancer samples

- Why:
  - Mostly discover
  - Browse/explore
  - Some identify, mostly compare

- How:
  - Multiform, overview/detail views
  - Linked views
  - Multiple idioms (heatmap, network graph, bar charts...)

TABLE 1

<table>
<thead>
<tr>
<th>Data Contained in Each MAGI Mouse Trace Interaction Log</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of information</td>
<td>Attributes</td>
</tr>
<tr>
<td>Mouse events</td>
<td>(click, move, scroll, time, x, y)</td>
</tr>
<tr>
<td>Tooltip events</td>
<td>(x, y, width, height)</td>
</tr>
<tr>
<td>MAGI components (=0)</td>
<td>(number of genes and datasets)</td>
</tr>
<tr>
<td>Window state</td>
<td>(width, height)</td>
</tr>
<tr>
<td>Query</td>
<td>(Mutation type/location)</td>
</tr>
</tbody>
</table>

Rectangles = areas of tool
Orange/red/purple = mouse movement/click/scroll
Step 2: Generate Task Labels with Users
- 5 grad student pairings, containing 1 genomics expert and 1 vis expert
- Labelled logs with 1 of the 8 defined tasks
- 96-random order trails
- 48-trials unique
- 48-trials repeated between subjects

Step 3: Task Classification
- Training set:
  - all remaining trials (48*5), used for training and cross-validation
- Testing set:
  - the 48 trails all groups did used for IRR

Critique
- Random Forests significantly the best
- Mouse tracking the best (across all), including better than all
- Dwell, ROI transitions may have worsened performance
- Best combo RF + mouse tracking

Results/Author’s Discussions
- Benefits from these results to tool design:
  - Which parts of the tool are used most often
  - Proximity to most used parts matter
  - “Top-down” vs “bottom-up” strategies for exploration
- Some contradicted prior user studies e.g. what tasks used

Critique – Strengths
- Objective user studies better allow “evidence-based” design and reproducible (real) results
- Machine learning: used a few, popular models (not too many, not too few?)
- Thought out design, e.g. quantifying inter-rater reliability

Critique – Weaknesses
- Labelling, should that have been done with the logs vs screen capture?
- Better gold standard? But perhaps could not get data
- Were “match any” results a bit deceptive?
- At least in the main result figure?
- The tool lent itself well to the study, but was it popular/representative/used?

Classifiers Tested
- Used random forests, SVMs, and k-nearest neighbours
- Justified as using machine learning models that are widely familiar
- Tested different sets of features

Broader Generalizability
- Show that mouse interactions may be more deterministic than text-focused interaction logs
- Unsupervised learning’s potential an open problem
  - Segmentation logs, however, could be a difficulty

Results/Author’s Discussions
- As classification results different than previous user studies, authors suggest utilizing a combination (user study + automated classification)
- Make detailed predictions with in lab-observations
- Identify bias using logs

MAGI (published 2015)