Paper Presentation:
An Analysis of Automated Visual Analysis Classification: Interactive Visualization Task Inference of Cancer Genomics Domain Experts

John-Jose Nunez

CPSC 547

November 19, 2019

Have you ever thought….

- “ They only included 3 users in their user study, do they really speak for all users? (overfitting?)”

- “Is a user study in the lab even applicable to the real world? (observation effect)”

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.

An Analysis of Automated Visual Analysis Classification: Interactive Visualization Task Inference of Cancer Genomics Domain Experts

Connor C. Gramazio, *Student Member, IEEE*, Jeff Huang, and David H. Laidlaw, *Fellow, IEEE*

- https://vimeo.com/364568057

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

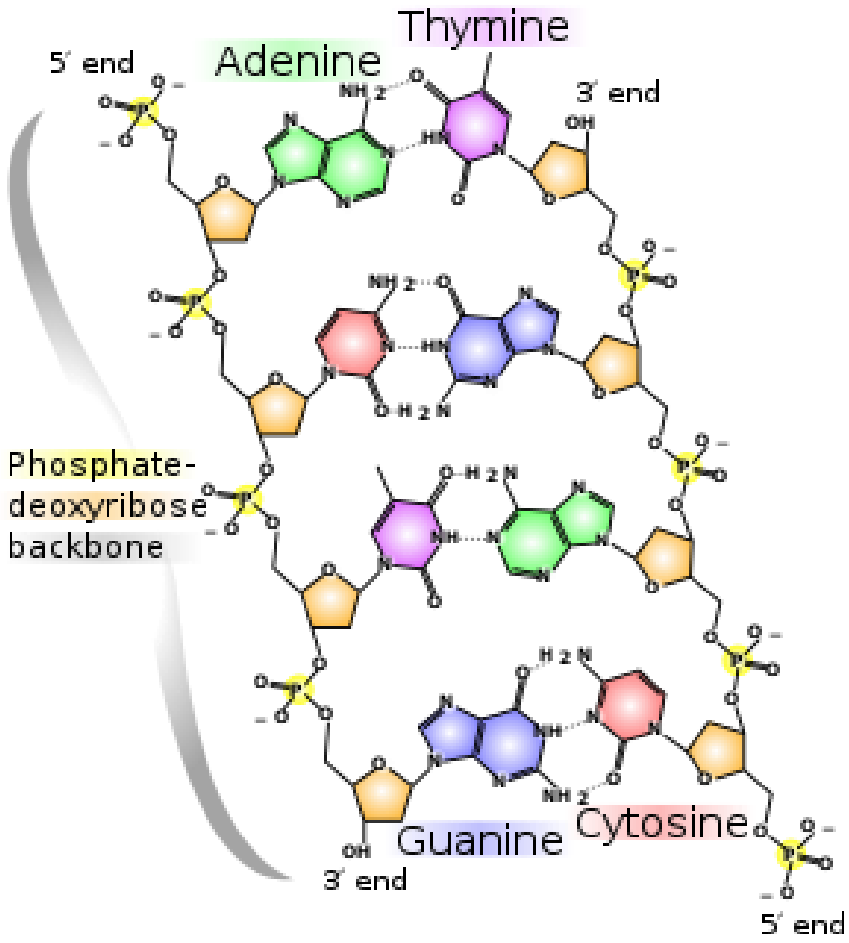
Tool Being Studied: MAGI

- Online visualization tool

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

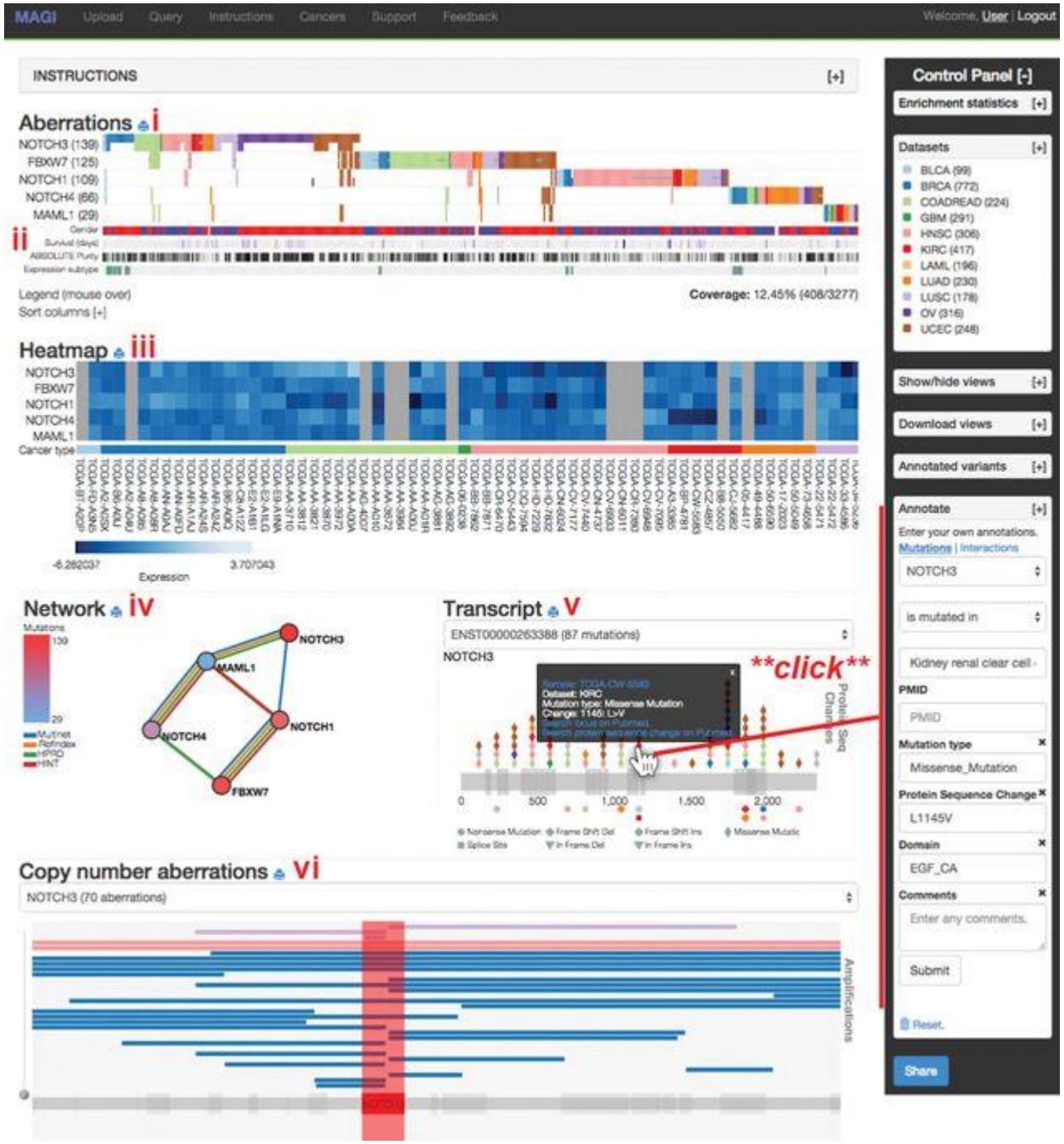
Super Brief Domain Background

- 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



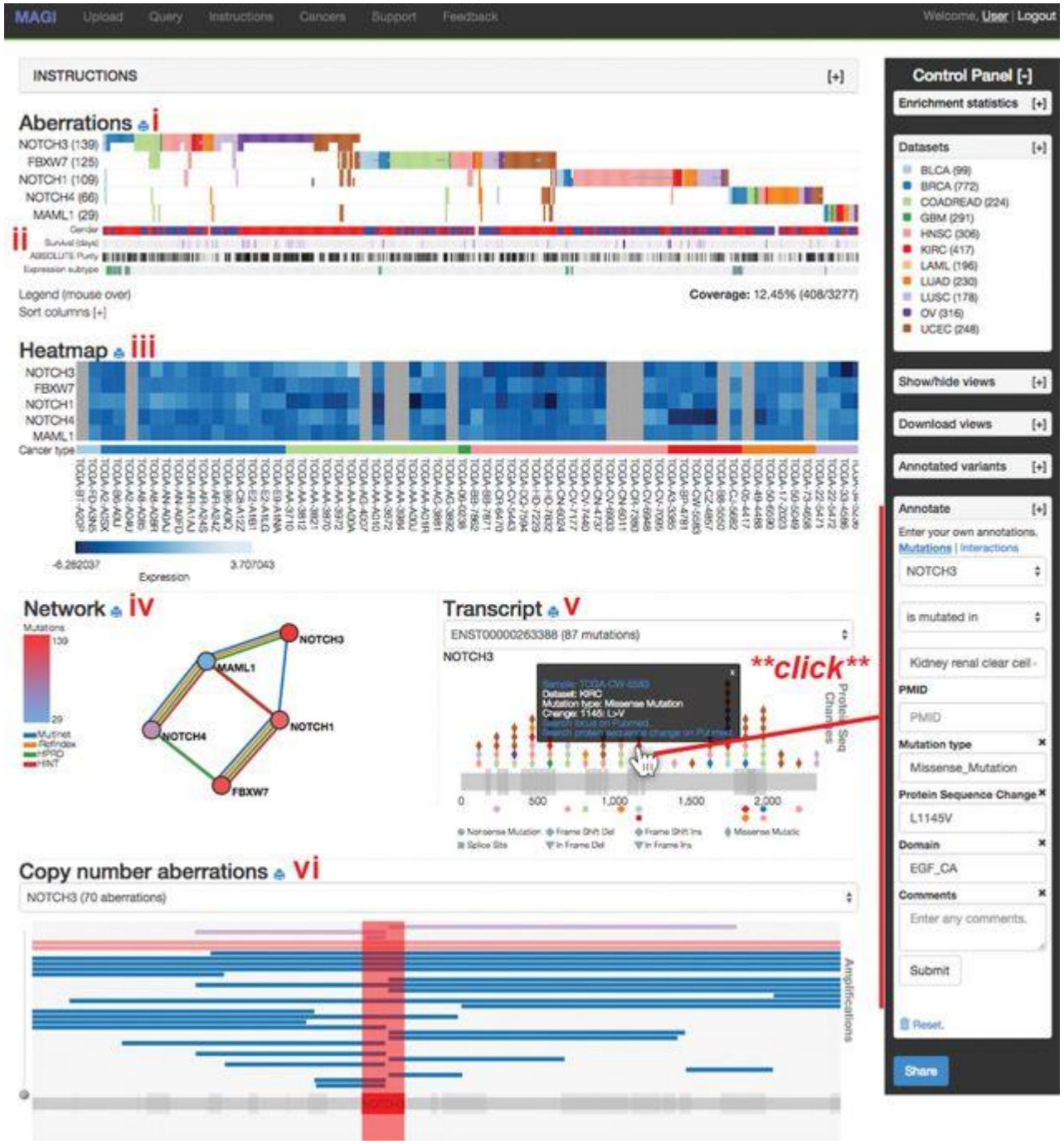
Tool: MAGI

- (i) Aberration view
  - Pattern of mutations in gene sets across tumors
- (ii) Aberration view row/heat maps
  - Show gender, survival, purity



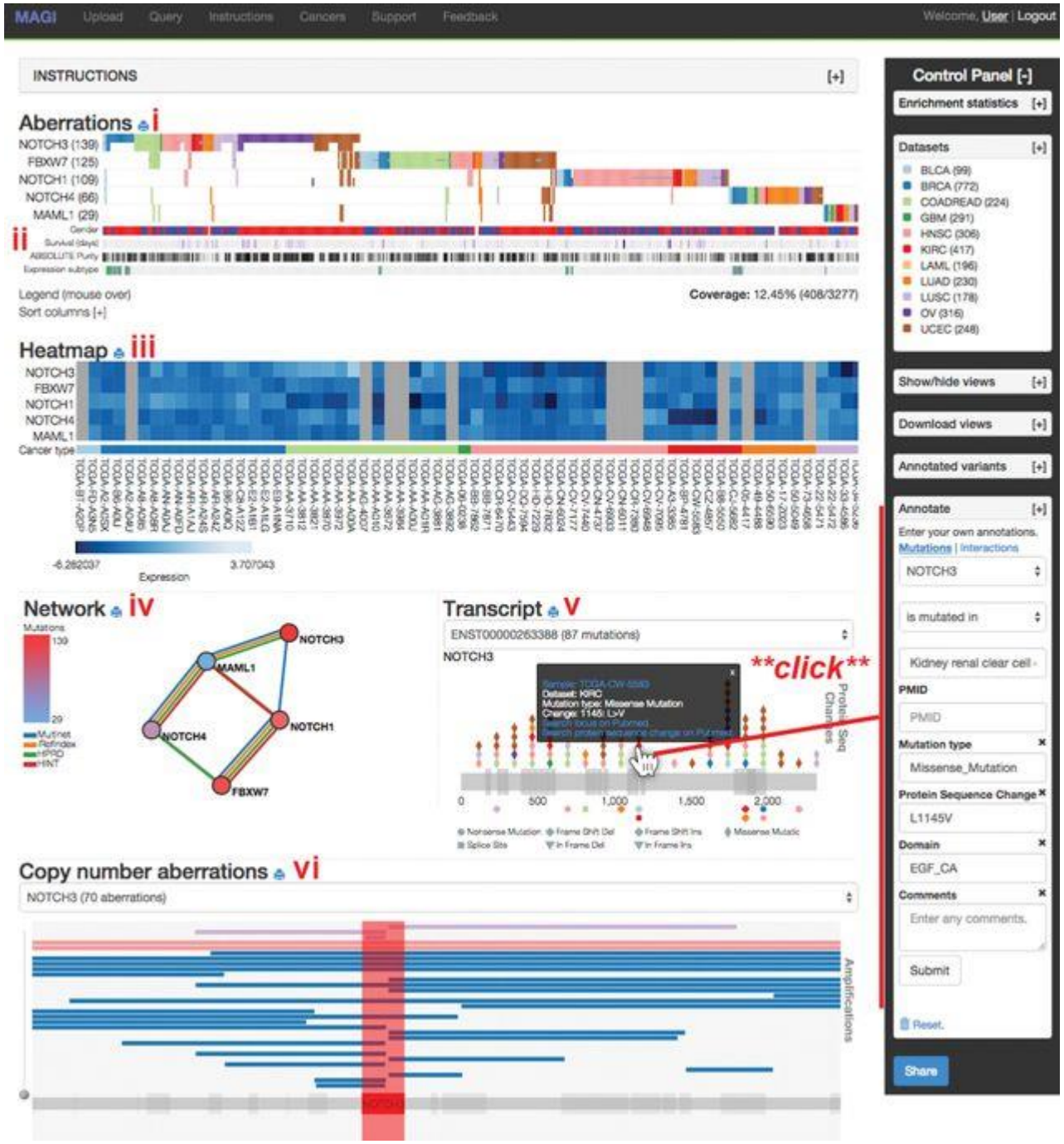
Tool: MAGI

- (iii) Heat map
  - User uploads, e.g. shows methylation for different tumors
- (iv) Network view
  - Interaction between gene



Tool: MAGI

- (v) Transcript view
  - Detail view of subset (one gene) showing mutation types/location
- (vi) Copy-number view
  - Another detail view of one gene
- Clicking activates highlighting to show a linked view



MAGI: Who/What/Why/How

- Who:
  - Cancer researchers in wet/dry labs, industry
- What:
  - DNA mutations present in cancer samples

MAGI

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

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

Step 1: Task Identification with MAGI Creators

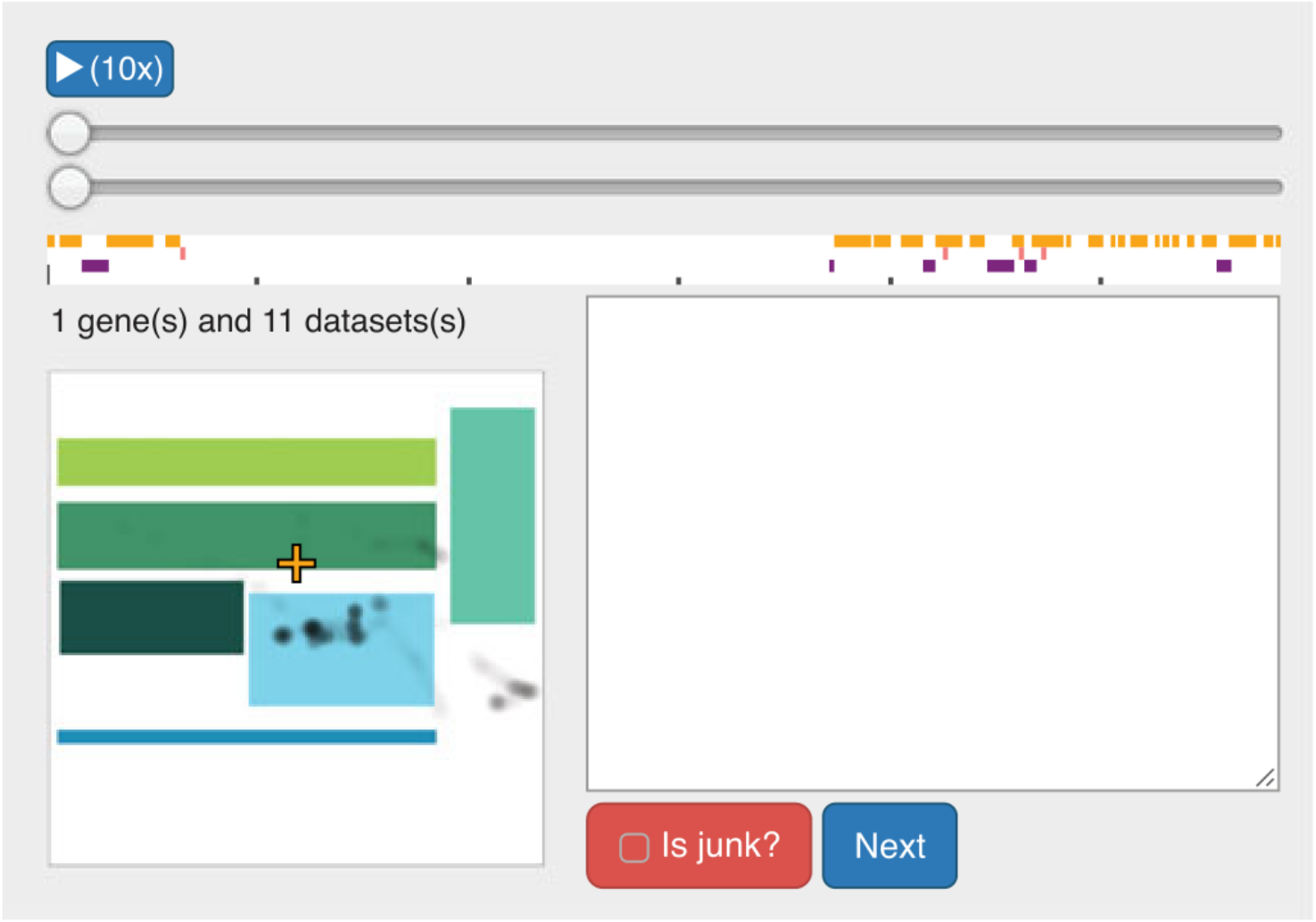
- 2 participants who created MAGI
- Randomly sampled logs from MAGI users
- 25 tasks labelled per participants, with free text
  - But based on a separate vis of the log data

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

TABLE 1
Data Contained in Each MAGI Mouse Trace Interaction Log

Type of information	Attributes
Mouse events	{click, move, scroll}, time, x, y
Tooltip events	x, y, width, height
MAGI components ($\times 6$)	x, y, width, height
Window state	width, height
Query	number of genes and datasets

MAGI components refer to the five visualizations and control panel.

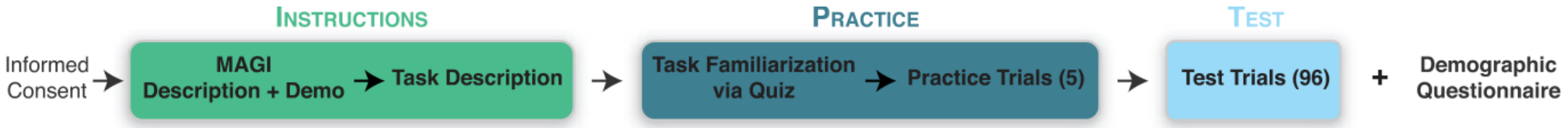


Rectangles = areas of tool.

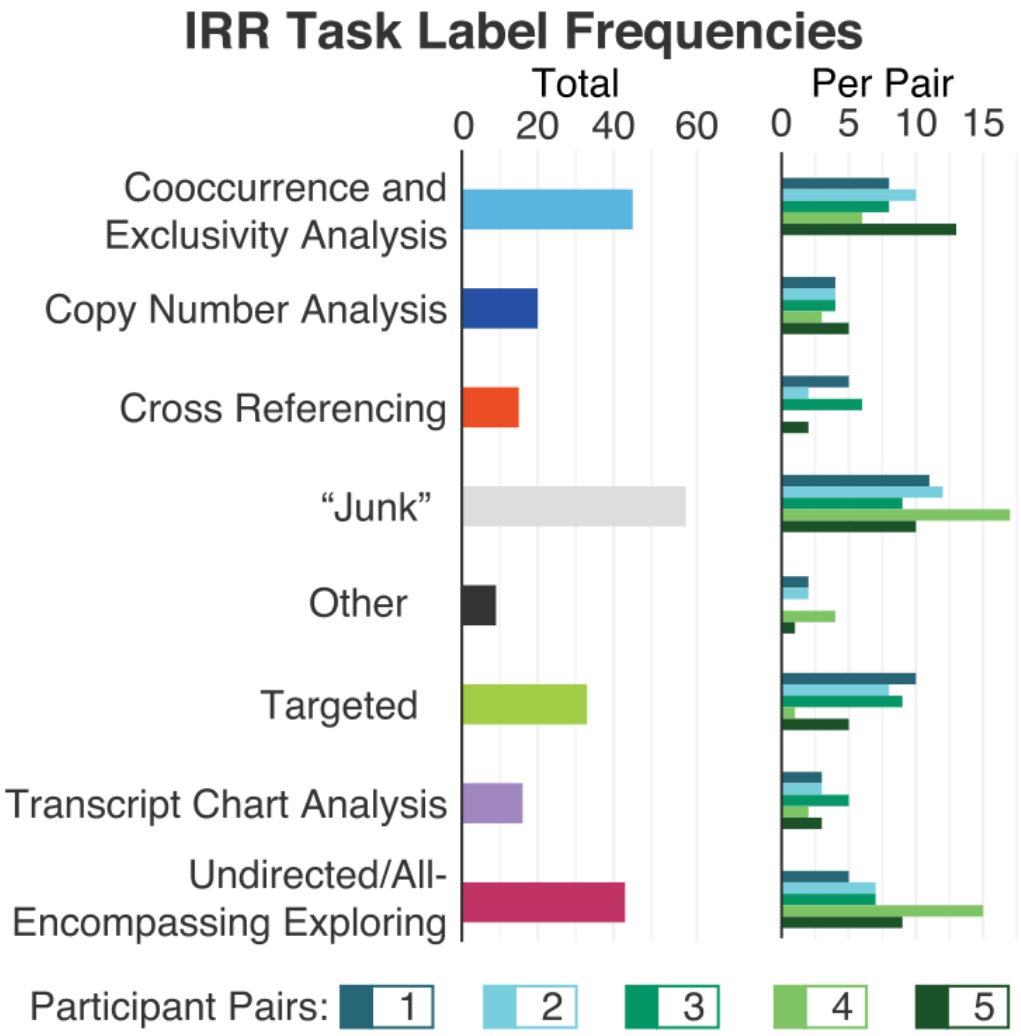
Orange/red/purple = mouse movement/click/scroll

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



- Half of the log trials were repeated
  - Inter-rater reliability measured
  - Fleiss' K 0.405
  - "fair-to-good" reliability
- Accuracies consistent among rater groups
- Group 4 weakest, the CS partner had least experience



Step 3: Task Classification

- Testing set:
  - the 48 trails all groups did used for IRR

- Training set:
  - all remaining trials (48*5), used for training and cross-validation

TABLE 2
An Overview of Three Feature Sets Used in Our Classification
(Not Shown: "All," the Combination of These Sets)

ROI Transition [37]	Dwell [4]	Mouse Tracking [38]
transition count	total time	stationary H
transitioned-to count	μ dwell time	transition H
	σ dwell time	total time $\forall$ ROI
	# datasets	active time $\forall$ ROI
	# genes	dwell time $\forall$ ROI
		μ active time $\forall$ ROI
		μ dwell time $\forall$ ROI

ROI transition count is short-hand for the complete adjacency matrix of transition features between each ROI. Transitioned-to count sums one dimension of the complete matrix. μ: mean, σ: deviation, H: entropy.

ROI = Region of Interest (views of the tool)

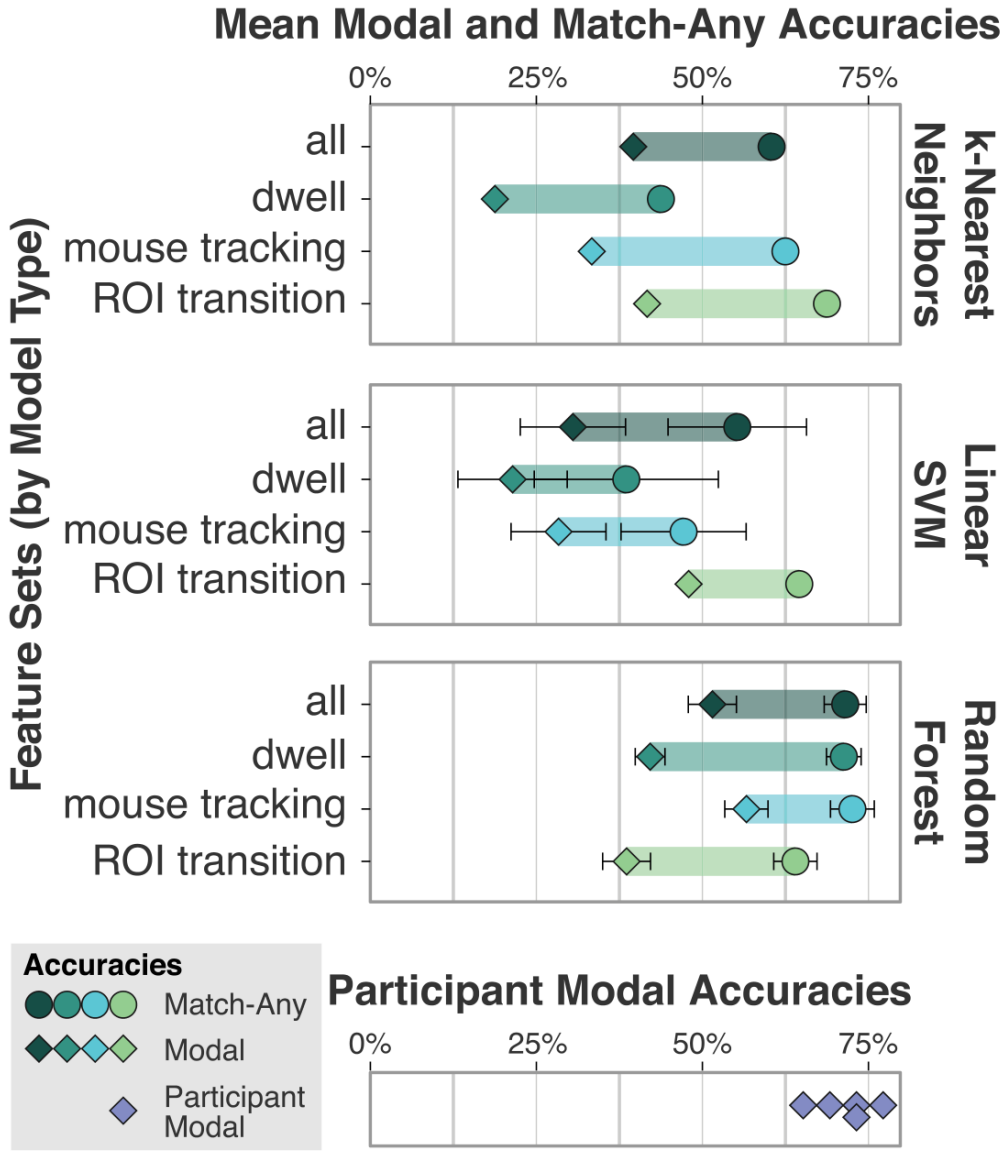
Classifiers Tested

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

TABLE 3
Parameter Selection for Each Tested Classifier

Classifier	Feature Set	Parameters
k-nearest	All	$k = 9$, w=distance
k-nearest	Dwell	$k = 10$, w=uniform
k-nearest	ROI Transition	$k = 5$, w=distance
k-nearest	Mouse Tracking	$k = 7$, w=uniform
Linear SVM	All	$c = 69.519$
Linear SVM	Dwell	$c =< 0.001$
Linear SVM	ROI Transition	$c = 0.001$
Linear SVM	Mouse Tracking	$c = 0.004$
Random Forest	All	estimators=75
Random Forest	Dwell	estimators=40
Random Forest	ROI Transition	estimators=40
Random Forest	Mouse Tracking	estimators=40

w: weight.



Match any = at least 1 group labelled

- 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

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

Broader Generalizability

- Show that mouse interactions may be more deterministic than text-focused interaction logs

- Unsupervised learning's potential an open problem
  - Segmenting logs, however, could be a difficulty

Critique - Strengths

- **Objective user studies better allow “evidence-based” design and reproducible (real) results**

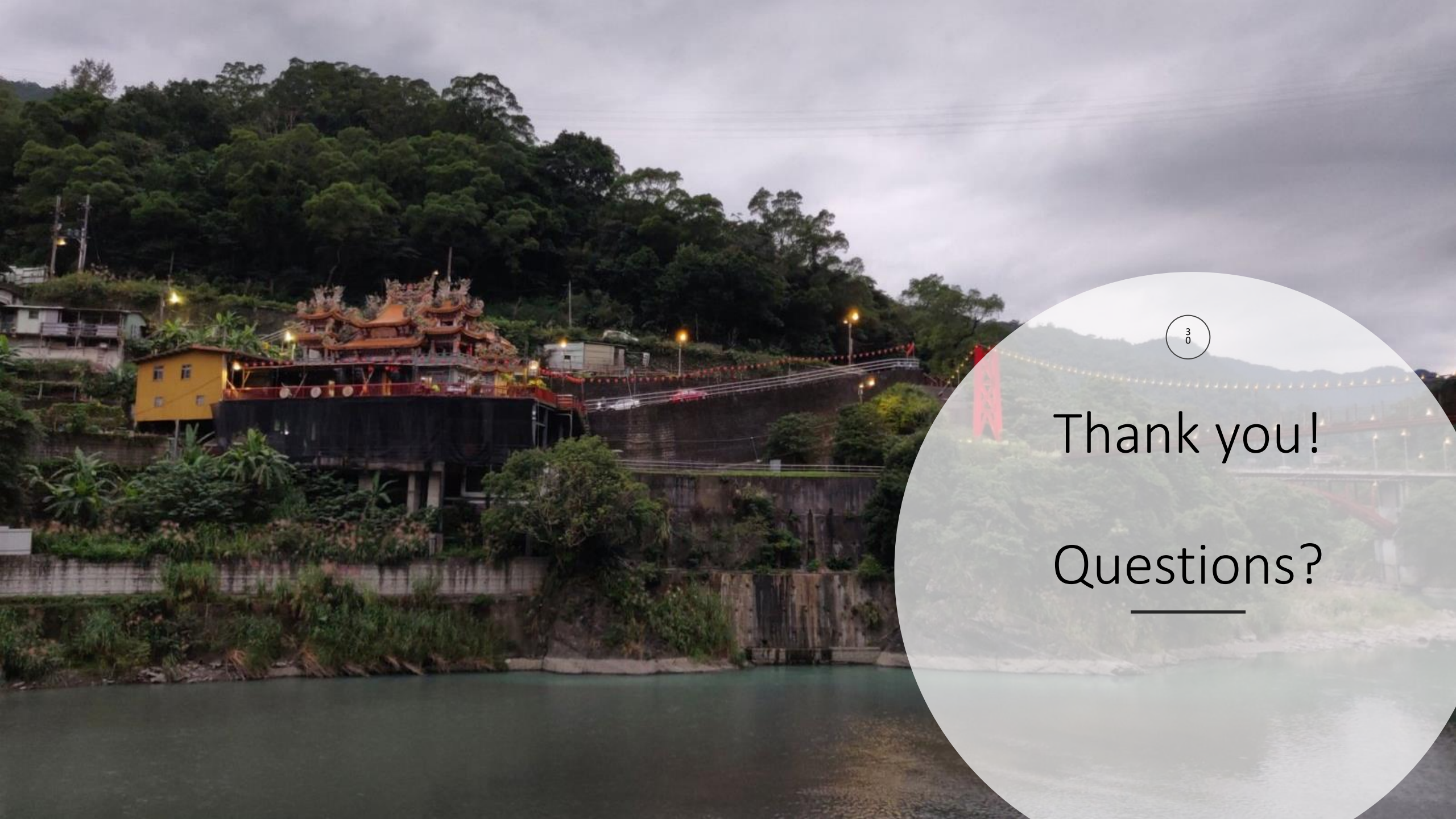
- Machine learning: used a few, popular models (not too many, not too few?)

- 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?

MAGI (published 2015)



Thank you!

Questions?