Evaluation

Jessica Dawson

533C Topic Presentation

November 9, 2011

Outline

- Human-centered design for geovis
  - David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

- Evaluation through insight
  - Purvi Saraiya, Chris North, Karen Duca. *An Insight-Based Methodology for Evaluating Bioinformatics Visualizations.* IEEE Trans. Vis. Comput. Graph. 11(4):443-456 (2005)

- Crowdsourced perception experiments
  - Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Theme

Proposing and evaluating *methods of evaluation* for the development of infovis applications.

HUMAN-CENTERED APPROACHES IN GEOVISUALIZATION DESIGN: INVESTIGATING MULTIPLE METHODS THROUGH A LONG-TERM CASE STUDY

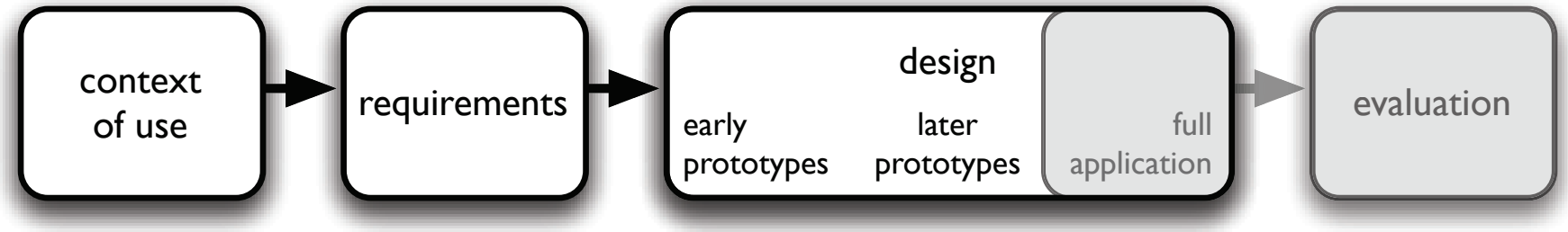
David Lloyd and Jason Dykes. (Proc. InfoVis 2011).

Overview

- Problem
  - How to apply human-centered (HC) design processes to the early stages of *geovis* design?
- Method + Evaluation
  - In depth, 3-year case study with 3 domain specialists
  - Follow HC design process to design a geovisualization
- Paper summary of the whole process
  - Published details of the study at each stage in separate papers

Case-Study Method

- Use stages in ISO Standard 13407 on human centered-design



- Focusing specifically on early stages (in white)

- Employ multiple HC methods at each stage
  - Assess effectiveness of each method for the goals of the stage

David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

Stage 1: Understanding Context of Use

- Briefly . . .
- Goal
  - Understand "users, tasks and the organizational an physical environment"
- Methods:
  - Field research methods, contextual inquiry
  - Lots of data collection methods
    - interviews, observation, questionnaires, content analysis, card sorting.
- Results
  - Mostly inline with expected results from other domains
  - Specifically interesting for vis: realize need to *understand data in context*

Stage 2: Establishing Requirements

- Looking for approaches that encourage participatory, collaborative engagement of users

- Methods
  - Standard Volere method
    - structured template of generic questions
  - Alternatives:
    - Lectures and elicitation of ideas through card sorting, interviews, sketching
    - Expert interviews with geovis design experts

Stage 2 Results

- Volere Method: Ineffective
- Lecture: overwhelmed specialists
  - Sketching somewhat effective
  - But difficulty determining priority/suitability of tools
- Expert Interviews
  - Effective, but missing domain knowledge
- Expert Interviews and sketching similar

Stage 3: Early Prototype Designs

- Are wireframe style prototypes useful for geovis design?

- Paper wireframe prototypes
  - Application states as multiples on a single sheet
  - Interactions conveyed verbally

- Method
  - Generated designs from stage 1 and 2 output; fake data
  - Specialists used a think-aloud protocol

David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

Stage 3 Results

- Wireframes successful for communicating design

- Real data important
  - Tradeoff of 'quick' prototyping

	Specialist	Wireframe 1	Wireframe 2	Specialist total
approval	1	3	4	7
	2	3	4	7
	3	1	4	5
idea	1	5	5	10
	2	2	6	8
	3	2	5	7
limitation	1	0	5	5
	2	2	3	5
	3	2	2	4
opinion	1	1	5	6
	2	2	8	10
	3	4	5	9
query	1	0	0	0
	2	6	4	10
	3	4	3	7
	Total	**37**	**63**	**100**

David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

Stage 4: Later Prototype Designs

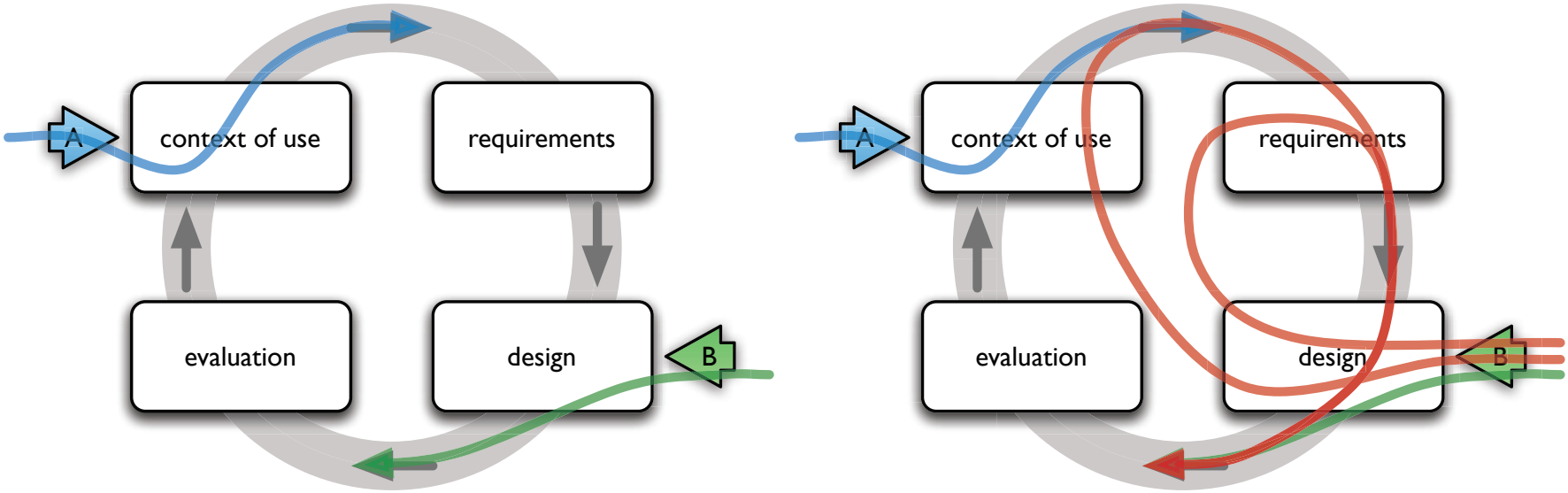
- Goals:
  - Do prototypes provoke feedback? Do prototypes elicit exploratory behavior?
- Prototypes
  - Paper and digital versions
- Method
  - User testing with intervention
  - Real domain data, simple tasks
  - Counts of suggestions/behaviour recorded

David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

Stage 4: Results

- Exploration Behavior
  - Similar amounts of task driven exploration for both paper and digital
- Feedback and Improvements
  - Paper prototype yielded more suggestions (except interface-related)
- Sketchiness communicated ‘suggestive’ rather than ‘definitive’
- In short: prototyping works
  - the quicker and sketchier the better

Conclusions



- HC design methods can be effectively employed for geovis
  - With vis specific limitations

David Lloyd and Jason Dykes. *Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study.* Proc. InfoVis 2011.

Critique

- Tried lot of different methods at each stage
  - What works/what doesn't work for vis
- Lots of different data collection methods
  - qualitative analysis when possible
- Prototyping works!
  - Good evaluation of prototyping effectiveness
- 3 years is a long time!

Questions?

AN INSIGHT-BASED METHODOLOGY FOR EVALUATING BIOINFORMATICS VISUALIZATIONS.

Purvi Saraiya, Chris North, Karen Duca. IEEE Trans. Vis. Comput. Graph. 11(4):443-456 (2005)

Overview

- Problem:
  - How to evaluate infovis tools for biologists when tasks are *exploratory* and *open ended*?
- Proposed Solution:
  - Measure *insight* instead of performance
  - But can insight be measured in a controlled experimental setting?
- Evaluation + Method:
  1. Development of *Insight-based* methodology
  2. Evaluation of popular bioinformatics tools with respect to insight

Characterizing Insight

- Pilot Study
  - Think aloud observation with 5 participants
  - Exploratory, no protocol or task

- Results
  - An insight = an individual observation
  - Recognized as *any data observation* the user mentions aloud

- Characteristics
  - The actual observation made
  - Time to reach insight
  - Domain value of insight
  - Generated hypothesis?
  - Expected vs. unexpected insight
  - Correctness
  - Breadth vs. Depth of insight
  - Category (overview? pattern groups? details?)

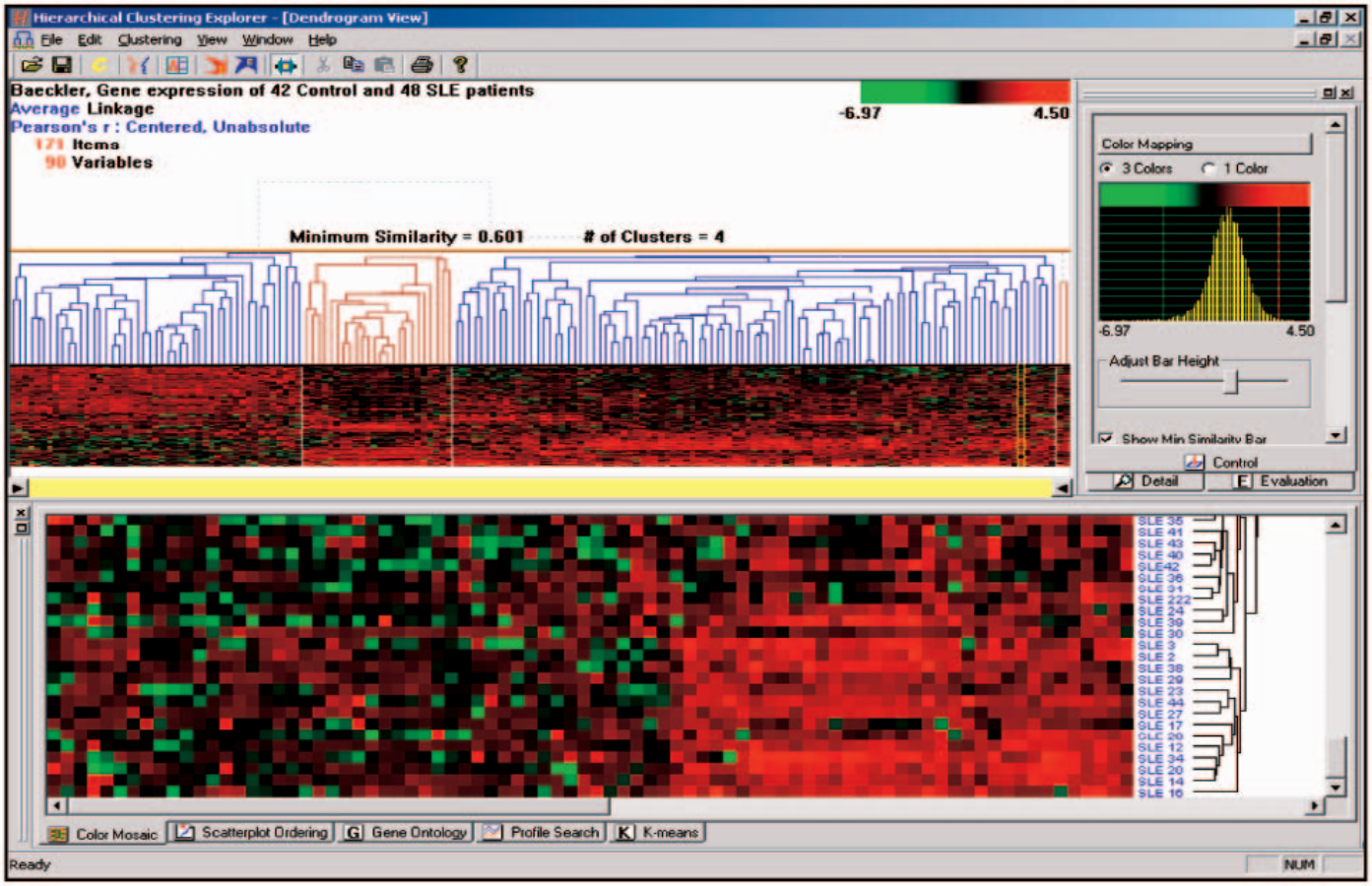
Experiment

- Evaluation of 5 popular bioinformatics tools in terms of *insight*
- Protocol:
  - Mix of controlled experiment and usability testing
  - Think aloud observation
- Design:
  - 3 multi-dimensional microarray data sets, between-subjects
  - 5 microarray visualization tools, between-subjects
    - Clusterview
    - TimeSearcher
    - HCE
    - Spotfire
    - GeneSpring

Microarray Tools

- Broad selection of techniques and capabilities
  - Heatmaps, parallel coordinates, clustering, etc.
  - Some support multiple visualization techniques, some support only one;
- In depth discussion of tools out scope
  - See paper for details

Tool Example: HCE



Purvi Saraiya, Chris North, Karen Duca. *An Insight-Based Methodology for Evaluating Bioinformatics Visualizations*. IEEE Trans. Vis. Comput. Graph. 11(4):443-456 (2005)

Experiment

- Design continued. . . :
  - 30 participants
    - Biology background; mix of experts, novices
    - 2 per dataset, per tool
  - Exploratory task
    - Examining interactions among genes and conditions.

- Analysis
  - Insights identified and coded by experimenters from video

Results

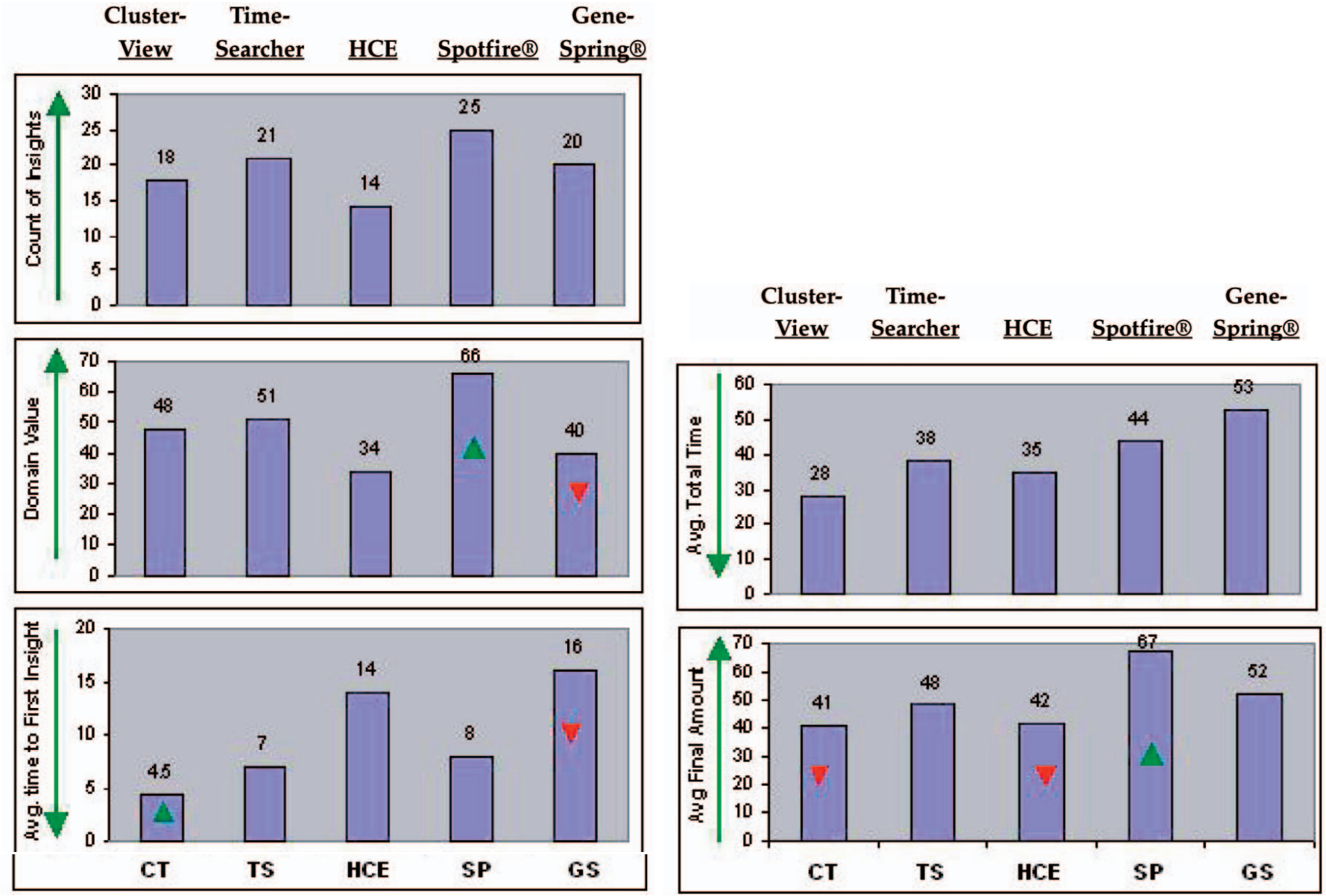
- Lots of results
  - Mainly qualitative

What we won't discuss

- Paper has great details for:
  - General tendencies across dataset and tools with respect to insight
  - The pros/cons of specific tools

What we will discuss

- How effective was the insight-based methodology?



Purvi Saraiya, Chris North, Karen Duca. *An Insight-Based Methodology for Evaluating Bioinformatics Visualizations*. IEEE Trans. Vis. Comput. Graph. 11(4):443-456 (2005)

Effectiveness?

- By using insight characteristics as a measure, the authors came to some strong conclusions

- Also novel high-level observations
  - Domain experts performed on par with novices
  - More breadth insights than depth insights
  - Multiple views affects confidence

Limitations

- Coding of insights labour intensive
- Without tasks, it can be difficult to motivate users
- Domain experts are required for *deep*, *meaningful* insights

Critique

- New method based on insights
  - Applicable to a wide range of vis-domain
  - Not just for summative design

- Experiment was only between subjects
  - What about difference in insight for one user with multiple tools?

Questions?

CROWDSOURCING GRAPHICAL PERCEPTION: USING MECHANICAL TURK TO ASSESS VISUALIZATION DESIGN.

Jeffrey Heer and Michael Bostock. (Proc. CHI 2010)

Overview

- Problem:
  - Are web-based evaluations through Amazon's Mechanical Turk (MTurk) a viable method for graphical perception experiments?
- Evaluations:
  1. Replicate prior laboratory studies;
  2. Generate of new graphical perception results
- Provide cost/benefit analysis

Web-Based Evaluations

- Increasing use of web-based platforms to perform experiments and conduct user research

- Benefits
  - Substantial reductions in cost/time to result
  - Ecological validity

- Possible Limitations
  - Vis perspective
    - Lack of control over display configurations, viewing environment, etc

Mechanical Turk (MTurk)

- Popular *micro-task* market

- Requesters post jobs, called *HITs* (*Human Intelligence Tasks*)
  - *HITs* come with a small reward, e.g. $0.01 -$0.10,
  - a maximum number of assignments that can be performed
- A pool of workers, called *Turkers*, select *HITs* to perform
  - Requesters pay Turkers for completed *HITs*

- Considerations for experimentation
  - Qualification tasks can be introduced
  - Flexibility through embedding your own web pages

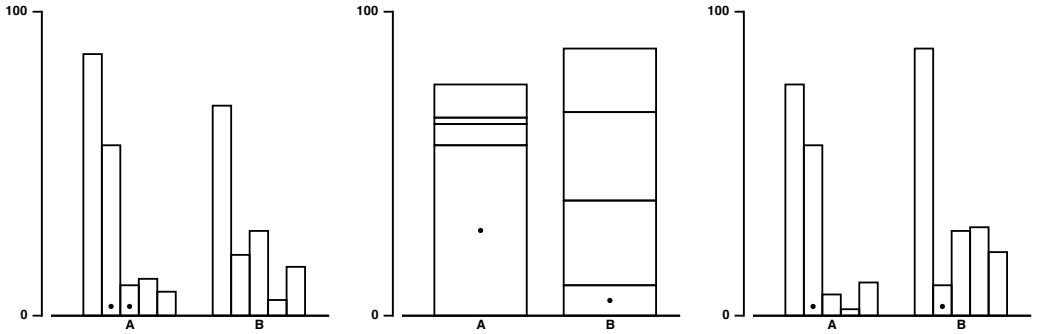
Experiment 1A

- Replication of Cleveland and McGill study
  - W.S. Cleveland and R. McGill. *Graphical Perception: Theory, experimentation and application to the development of graphical methods*. J. Am. Statistical Assoc. 79:531-544 (1984).

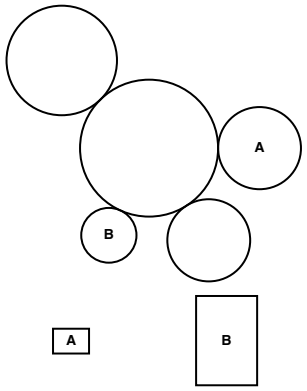
- Study ranked visual variables by their effectiveness
  - For each visual encoding, users asked to "identify the smaller of two marked values" and then,
  - "make a quick visual judgement" to estimate what percentage the smaller is of the larger.

Experiment 1A Design

- Design
  - 7 judgment types
  - 10 charts
    - 70 trials (individual *HITs*)
  - Subjects paid $0.05/judgment



- Judgment task encodings
  - Looked at position and length (original study)
  - Angle and circular encoding (modified to match study format)

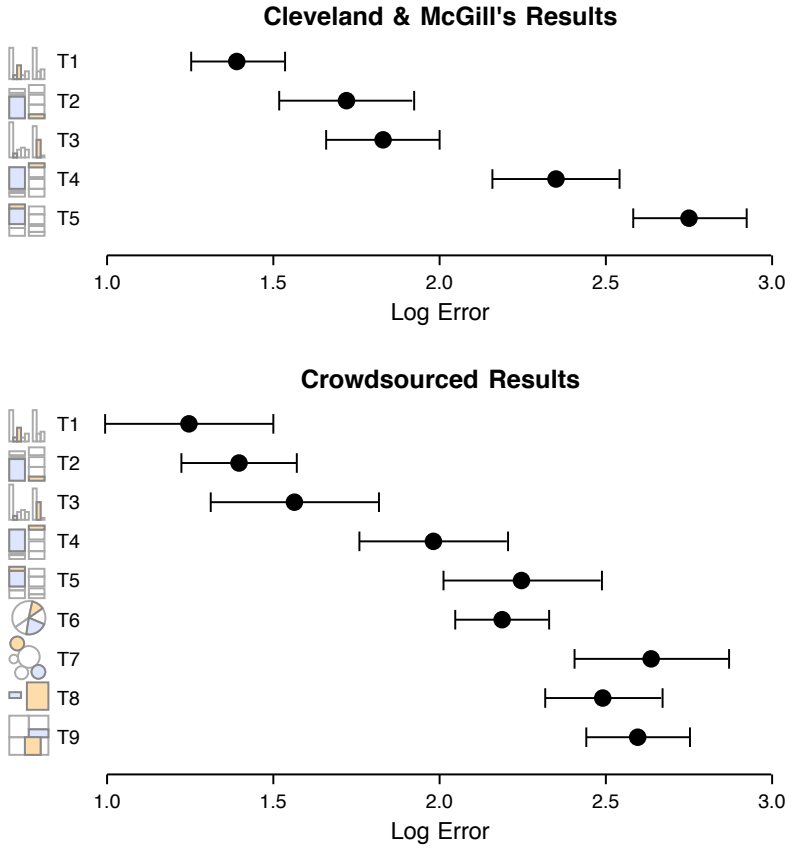


Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Results

- Analysis
  - 50 subjects, 3481 responses
- Replicated data exploration
  - Absolute error measure of accuracy

  $Log_2(|judged\ percent - true\ percent| + 1/8)$
- Results not identical, but similar
  - Rankings preserved, success!
- Additional Experiment 1B
  - Novel experiment, tasks 8/9 in chart
  - For details see paper



Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Experiment 2 (Briefly. . .)

- Successful replication of Stone and Bartram study,
  - M. Stone and L. Bartram. *Alpha, contrast, and the perception of visual metadata*. Proc. Color Imaging Conf. 2009.

- Subjects configure transparency (alpha value) across varying backgrounds and densities
- Additional measure of screen configurations was recorded and analyzed

- See paper for details

Experiment 3

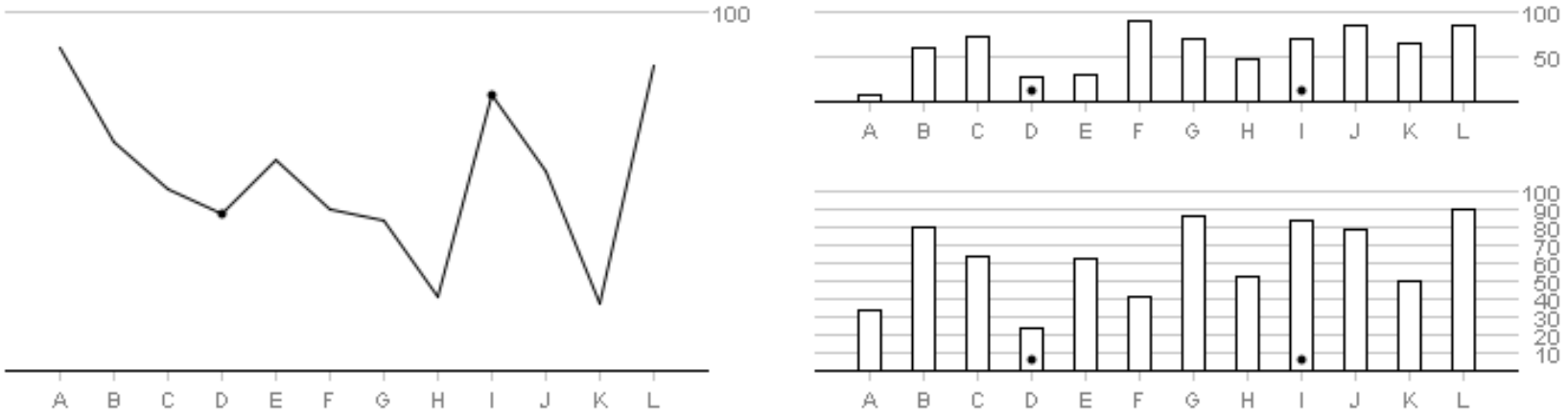
- Novel experiment to assess crowdsourcing for experiments looking at chart size variations

- Examined effects of *chart size* and *gridline spacing* on the accuracy of value comparisons in charts

Experiment 3 Design

- 2 chart x 3 height x 4 gridline spacing
- 72 trials (individual *HITs*)
- Subjects paid $0.02/*HIT*

- Task
  - Participants asked to identify the smaller marked element, and then estimate the difference between the two

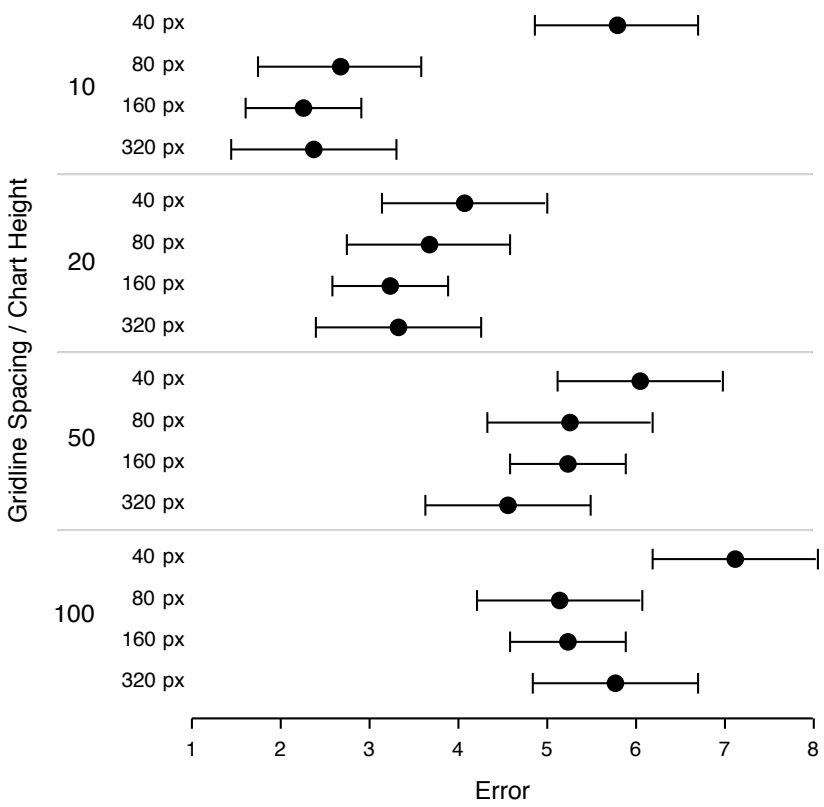


Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Results

- Analysis
  - Estimation error

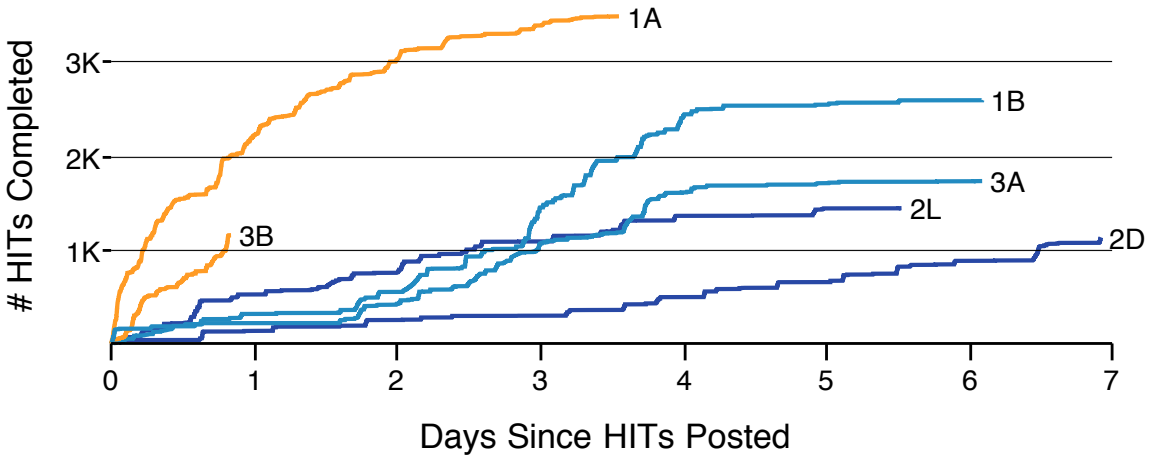
    |judged difference – true difference|
  - Response time could not be analyzed because of unreliability
- Significant effect of chart size
- and gridlines
  - See paper for details



Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Performance and Cost

- Cost-saving
  - Total expenditure, $367.77; a lab experiment would be $2190
- Time-saving
  - Days instead of weeks to complete experiment



Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Limitations and Considerations

- Turkers overlap across studies
- *HIT* Completion rates vary
- Reward level has effects
  - Raising $ decreases time to results, but Turkers seem to be less accurate
- Lots more in paper
- The good news
  - Turkers provide high-quality results (most of the time)

Jeffrey Heer and Michael Bostock. *Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design.* Proc. CHI 2010.

Overall Results

- MTurk is a viable option for perception experiments
  - Successfully replicated 2 experiments
  - Conducted 2 novel experiments with interesting results
- However, it comes with a lot of limitations
  - May be best used in combination with other evaluation techniques

Critique

- Replication of results and novel experiments convincing
- Gathered data about the process of running an Mturk experiment
  - Able to create guidelines for running studies based off experience

Synthesis

- Emphasis on new methods for evaluation for a variety of infovis domains
  - Geovis, bioinformatics, graphical perception
- Evaluating the effectiveness of the evaluation methods through different methodological approaches
  - Case studies and field work
  - Web-based controlled experiments
- All three tackle evaluations targets a different design stages
  - Pre-pre-design
  - Pre-design to prototyping
  - Summative design

Questions?