Topic Notes

InfoVis Evaluation

CS 7450 - Information Visualization
December 3, 2012
John Stasko

Area Focus

- Most of the research in InfoVis that we've learned about this semester has been the introduction of a new visualization technique or tool
  - Fisheyes, cone trees, hyperbolic displays, tilebars, themescapes, sunburst, jazz, …
  - "Isn't my new visualization cool?…"

Evaluation – Why?

- Reasons?

Evaluation – Why?

- Want to learn what aspects of visualizations or systems "works"
- Want to ensure that methods are improving
- Want to insure that technique actually helps people and isn't just "cool"
- NOT: Because I need that section in my paper to get it accepted … sigh

Evaluation – How?

- What do we measure?
  - What data do we gather?
  - What metrics do we use?

- What evaluation techniques should we use?

- (Channel your HCI knowledge)

Evaluation in HCI

- Takes many different forms
  - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, …

- So, which ones are best for evaluating InfoVis systems?

Controlled Experiments

- Good for measuring performance or comparing multiple techniques
- Often quantitative in nature
- What do we measure?
  - Performance, time, errors, …
- Strengths, weaknesses?

Subjective Assessments

- Often observational with interview
- Learn people’s subjective views on tool
  - Was it enjoyable, confusing, fun, difficult, …?
- This kind of personal judgment strongly influence use and adoption, sometimes even overcoming performance deficits
- Strengths, weaknesses?

Running Studies

- Beyond our scope here
- You should learn more about this in CS 6750 or 6455

Evaluating UI vs. InfoVis

- Seems comparable but…
- What are some differences?

Usability vs. Utility

- Big difference
- Usability is not the same as utility, which seems to be a key factor for InfoVis
- Can think of visualizations that are very usable but not useful or helpful
- More difficult to measure success of an infovis because more domain knowledge and situated use is required

Evaluating InfoVis in General

- Very difficult in InfoVis to compare "apples to apples"
  - Hard to compare System A to System B
  - Different tools were built to address different user tasks
- UI can heavily influence utility and value of visualization technique

Evaluating Research

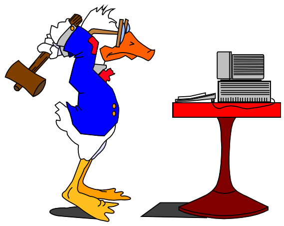
- How does one judge the quality of work in Information Visualization?

Research Evaluation

- Different possible ways
  - Impact on community as a whole, influential ideas
  - Assistance to people in the tasks they care about

Strong View

- Unless a new technique or tool helps people in some kind of problem or task, it doesn't have any value



Broaden Thinking

- Sometimes the chain of influence can be long and drawn out
  - System X influences System Y influences System Z which is incorporated into a practical tool that is of true value to people
- This is what research is all about (typically)

OK, what has research community done?

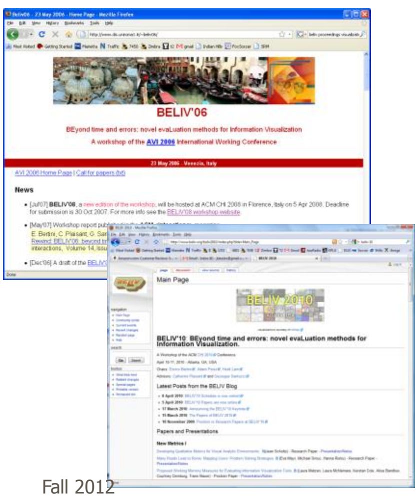
Past Review

- Old journal issue whose special topic focus was Empirical Studies of Information Visualizations
  - *International Journal of Human-Computer Studies*, Nov. 2000, Vol. 53, No. 5
- A bit dated now

BELIV

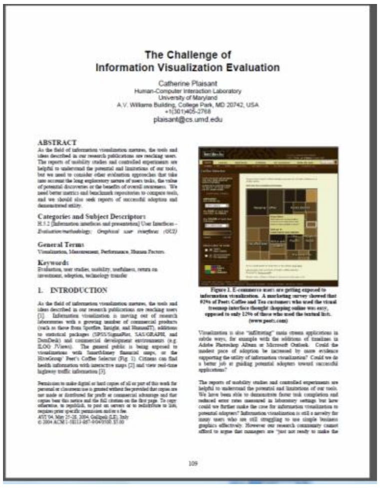
Nice locations!

Workshop focused on this topic



Plaisant '04

- Discusses challenges, possible next steps, and gives examples from work at Maryland


The Challenge of
Information Visualization Evaluation

Catherine Plaisant
Human-Computer Interaction Laboratory
University of Maryland
A.V. Williams Building, College Park, MD 20742, USA
+1(301)405-2768
plaisant@cs.umd.edu

ABSTRACT

Categories and Subject Descriptors

Evaluation Challenges

- Matching tools with users, tasks, and real problems
- Improving user testing
  - Looking at the same data from different perspectives, over a long time
  - Answering questions you didn't know you had
  - Factoring in the chances of discovery and the benefits of awareness
- Addressing universal usability

Possible Next Steps

- Repositories of data and tasks
- Case studies and success stories
- The role of toolkits and development tools

Carpendale '08

- Challenges in infovis evaluation
- Choosing an evaluation approach

Evaluating Information Visualizations

Sheelagh Carpendale

Department of Computer Science, University of Calgary,
2500 University Dr. NW, Calgary, AB, Canada T2N 1N4
sheelagh@ucalgary.ca

1 Introduction

Information visualization research is becoming more established, and as a result, it is becoming increasingly important that research in this field is validated. With the general increase in information visualization research there has also been an increase, albeit disproportionately small, in the amount of empirical work directly focused on information visualization. The purpose of this paper is to increase awareness of empirical research in general, of its relationship to information visualization in particular; to emphasize its importance; and to encourage thoughtful application of a greater variety of evaluative research methodologies in information visualization.

One reason that it may be important to discuss the evaluation of information visualization, in general, is that it has been suggested that current evaluations are not convincing enough to encourage widespread adoption of information visualization tools [57]. Reasons given include that information visualizations are often evaluated using small datasets, with university student participants, and using simple tasks. To encourage interest by potential adopters, information visualizations need to be tested with real users, real tasks, and also with large and complex datasets. For instance, it is not sufficient to know that an information visualization is usable with 100 data items if 20,000 is more likely to be the real-world case. Running evaluations with full data sets, domain specific tasks, and domain experts as participants will help develop much more concrete and realistic evidence of the effectiveness of a given information visualization. However, choosing such a realistic setting will make it difficult to get a large enough participant sample, to control for extraneous variables, or to get precise measurements. This makes it difficult to make definite statements or generalize from the results. Rather than looking to a single methodology to provide an answer, it will probably will take a variety of evaluative methodologies that together may start to approach the kind of answers sought.

The paper is organized as follows. Section 2 discusses the challenges in evaluating information visualizations. Section 3 outlines different types of evaluations and discusses the advantages and disadvantages of different empirical methodologies and the trade-offs among them. Section 4 focuses on empirical laboratory experiments and the generation of quantitative results. Section 5 discusses qualitative approaches and the different kinds of advantages offered by pursuing this type of empirical research. Section 6 concludes the paper.

A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 19–45, 2008.
© Springer-Verlag Berlin Heidelberg 2008

Evaluation Approaches

- Desirable features
  - Generalizability
  - Precision
  - Realism

Quantitative Methods

- Laboratory experiments & studies
- Traditional empirical scientific experimental approach
- Steps
  - Hypothesis development
  - Identification of independent variables
  - Control of Independent variables
  - Elimination of complexity
  - Measurement of dependent variables
  - Application of statistics

Quantitative Challenges

- Conclusion Validity
  - Is there a relationship?
- Internal Validity
  - Is the relationship causal?
- Construct Validity
  - Can we generalize to the constructs (ideas) the study is based on?
- External Validity
  - Can we generalize the study results to other people/places/times?
- Ecological Validity
  - Does the experimental situation reflect the type of environment in which the results will be applied?

Qualitative Methods

- Types
  - Nested methods

    Experimenter observation, think-aloud protocol, collecting participant opinions
  - Inspection evaluation methods

    Heuristics to judge
- Observational context
  - In situ, laboratory, participatory
  - Contextual interviews important

Qualitative Challenges

- Sample sizes
- Subjectivity
- Analyzing qualitative data

Lam, et al '12

- Meta-review: analysis of 850 infovis papers (361 with evaluation)
- Focus on evaluation scenarios

Empirical Studies in Information Visualization: Seven Scenarios

Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale

1 INTRODUCTION

Evaluation Taxonomies

TABLE 1
Taxonomies of Evaluation Methods and Methodologies Based on the Type of Categorization, the Main Categories Themselves, and the Corresponding References

Type	Categories	Refs
Evaluation goals	Summative *(to summarize the effectiveness of an interface)*, formative *(to inform design)*	Andrews [2], Ellis and Dix [22]
Evaluation goals	Predictive *(e. g., to compare design alternatives and compute usability metrics)*, observational *(e. g., to understand user behaviour and performance)*, participative *(e. g., to understand user behaviour, performance, thoughts, and experience)*	Hilbert and Redmiles [34]
Evaluation challenges	Quantitative *(e. g., types validity: conclusion (types I & II errors), construct, external/internal, ecological)*, qualitative *(e. g., subjectivity, sample size, analysis approaches)*	Carpendale [10]
Research strategies	Axes *(generalizability, precision, realism, concreteness, obtrusiveness)* and research strategies *(field, experimental, respondent, theoretical)*	McGrath [53]
Research methods	Class *(e. g., testing, inspection)*, type *(e. g., log file analysis, guideline reviews)*, automation type *(e. g., none, capture)*, effort level *(e. g., minimal effort, model development)*	Ivory and Hearst [42]
Design stages	Nested Process Model with four stages *(domain problem characterization, data/operation abstraction, encoding/interaction technique design, algorithm design)*, each with potential threats to validity and methods of validation	Munzner [54]
Design stages	Design/development cycle stage associated with evaluation goals *("exploratory" with "before design", "predictive" with "before implementation", "formative" with "during implementation", and "summative" with "after implementation")*. Methods are further classified as inspection *(by usability specialists)* or testing *(by test users)*.	Andrews [2]
Design stages	Planning & feasibility *(e. g., competitor analysis)*, requirements *(e. g., user surveys)*, design *(e. g., heuristic evaluation)*, implementation *(e. g., style guide)*, test & measure *(e. g., diagnostic evaluation)*, and post release *(e. g., remote evaluation)*	Usability.net [88]
Design stages	Concept design, detailed design, implementation, analysis	Kulyk et al. [46]
Data and method	Data collected (qualitative, quantitative), collection method (empirical, analytical)	Barkhuus and Rode [5]
Data	Data collected (qualitative, quantitative, mixed-methods)	Creswell [17]
Evaluation scope	Work environment, system, components	Thomas and Cook [82]

Evaluation Scenarios

- Understanding data analysis
  - Understanding environments and work practices (UWP)
  - Evaluating visual data analysis and reasoning (VDAR)
  - Evaluating communication through visualization (CTV)
  - Evaluating collaborative data analysis (CDA)

Evaluation Scenarios

- Understanding visualizations
  - Evaluating user performance (UP)
  - Evaluating user experience (UE)
  - Evaluating visualization algorithms (VA)

Methods

- Coded each paper with tags

TABLE 3
Original Coding Tags, the Number of Papers Classified, and the Final Scenario to Which They Were Assigned

Paper Tags	EuroVis	InfoVis	IVS	VAST	Total	Scenario
Process						
1. People's workflow, work practices	3	1	3	0	7	UWP
2. Data analysis	0	5	3	5	13	VDAR
3. Decision making	0	2	1	4	7	VDAR
4. Knowledge management	0	1	0	2	3	VDAR
5. Knowledge discovery	1	1	0	1	3	VDAR
6. Communication, learning, teaching, publishing	0	0	4	1	5	CTV
7. Casual information acquisition	0	4	0	0	4	CTV
8. Collaboration	0	3	2	4	9	CDA
Visualization						
9. Visualization-analytical operation	0	12	1	0	13	UP
10. Perception and cognition	17	24	15	3	62	UP
11. Usability/effectiveness	25	84	31	18	158	UP&UE
12. Potential usage	7	1	5	9	22	UE
13. Adoption	0	1	3	1	5	UE
14. Algorithm performance	17	37	15	0	69	VA
15. Algorithm quality	1	10	12	5	28	VA
Not included in scenarios						
16. Proposed evaluation methodologies	0	3	0	2	5	-
17. Evaluation metric development	2	6	1	1	10	-

Methods

- For each category the authors describe
  - Goals and outputs
  - Evaluation questions
  - Methods and examples

Examples

- Let’s examine a few example studies utilizing different goals and styles

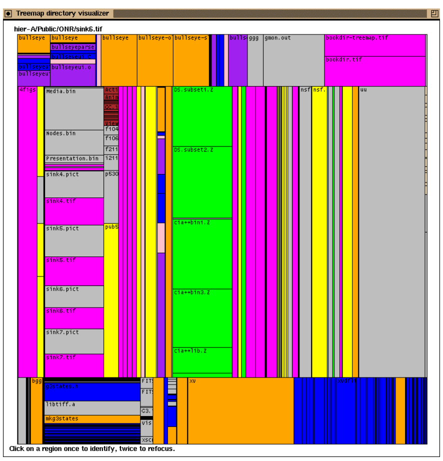
Start

Which Technique is Best?

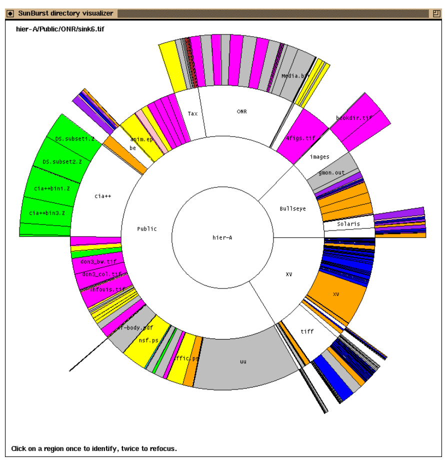
- Space-filling hierarchical views
- Compare Treemap and Sunburst with users performing typical file/directory-related tasks
- Evaluate task performance on both correctness and time

Stasko et al
IJHCS '00

Tools Compared



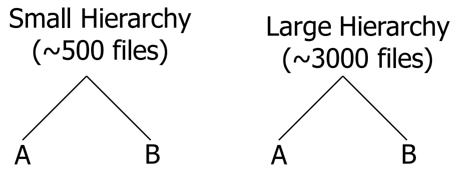
Treemap



SunBurst

Hierarchies Used

- Four in total



- Used sample files and directories from our own systems (better than random)

Methodology

- 60 participants
- Participant only works with a small or large hierarchy in a session
- Training at start to learn tool
- Vary order across participants

SB A, TM B
TM A, SB B
SB B, TM A
TM B, SB A

32 on small hierarchies
28 on large hierarchies

Tasks

- Identification (naming or pointing out) of a file based on size, specifically, the largest and second largest files (Questions 1-2)
- Identification of a directory based on size, specifically, the largest (Q3)
- Location (pointing out) of a file, given the entire path and name (Q4-7)
- Location of a file, given only the file name (Q8-9)
- Identification of the deepest subdirectory (Q10)
- Identification of a directory containing files of a particular type (Q11)
- Identification of a file based on type and size, specifically, the largest file of a particular type (Q12)
- Comparison of two files by size (Q13)
- Location of two duplicated directory structures (Q14)
- Comparison of two directories by size (Q15)
- Comparison of two directories by number of files contained (Q16)

Hypothesis

- Treemap will be better for comparing file sizes
  - Uses more of the area
- Sunburst would be better for searching files and understanding the structure
  - More explicit depiction of structure
- Sunburst would be preferred overall

Small Hierarchy

Hierarchy A					Hierarchy B			
Tool	Phase	Correct			Tool	Phase	Correct	
TM ($n = 8$)	1	9.88	(3.23)		TM ($n = 8$)	1	11.50	(2.14)
SB ($n = 8$)	1	12.88	(1.96)		SB ($n = 8$)	1	10.38	(1.69)
TM ($n = 8$)	2	12.25	(1.75)		TM ($n = 8$)	2	10.75	(2.77)
SB ($n = 8$)	2	12.63	(2.00)		SB ($n = 8$)	2	11.50	(2.00)
TM (collapsed across phase)		11.06	(2.79)		TM (collapsed across phase)		11.13	(2.42)
SB (collapsed across phase)		12.75	(1.91)		SB (collapsed across phase)		10.94	(1.88)

Correct task completions (out of 16 possible)

Large Hierarchy

Hierarchy A					Hierarchy B			
Tool	Phase	Correct			Tool	Phase	Correct	
TM ($n = 7$)	1	8.71	(1.60)		TM ($n = 7$)	1	8.29	(2.14)
SB ($n = 7$)	1	11.43	(1.27)		SB ($n = 7$)	1	11.14	(2.67)
TM ($n = 7$)	2	11.57	(1.27)		TM ($n = 7$)	2	10.86	(1.57)
SB ($n = 7$)	2	11.00	(2.16)		SB ($n = 7$)	2	11.00	(2.00)
TM (collapsed across phase)		10.14	(2.03)		TM (collapsed across phase)		9.57	(2.24)
SB (collapsed across phase)		11.21	(1.72)		SB (collapsed across phase)		11.07	(2.27)

Correct task completions (out of 16 possible)

Performance Results

- Ordering effect for Treemap on large hierarchies
  - Participants did better after seeing SB first
- Performance was relatively mixed, trends favored Sunburst, but not clear-cut
  - Oodles of data!

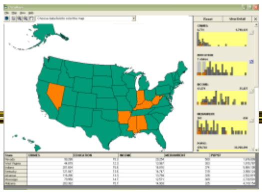
Subjective Preferences

- Subjective preference:
  SB (51), TM (9), unsure (1)
- People felt that TM was better for size tasks (not borne out by data)
- People felt that SB better for determining which directories inside others
  - Identified it as being better for structure

Strategies

- How a person searched for files etc. mattered
  - Jump out to total view, start looking
  - Go level by level

DQ vs. BH



Start

- Empirical Study
  - Use DataMaps, a geographic (US states) data visualization tool
  - Have participants do different tasks with both methods

    How many states have pop between x and y in 1970?
    Given 3 states, which has the lowest median income?
    What's the relationship between education and income?
    List states with pops. 0->x and y->z.
    What kind of a state is Florida?

We saw this earlier in term

Li & North
InfoVis '03

Findings

Functioned more as its own infovis tool

- Brushing histograms better and more highly rated for more complex discovery tasks
  - Attribute correlation, compare, and trend evaluation
- Dynamic queries better for more simple range specification tasks
  - Single range, multiple ranges, multiple criteria

Functioned more as auxiliary control for other vizs

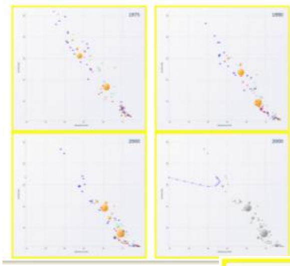
Animation Helpful?

Start

- Examine whether animated bubble charts (a la Rosling and GapMinder) are beneficial for analysis and presentation
- Run an experiment to evaluate the effects of animation

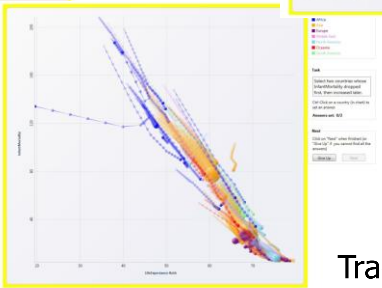
Robertson et al
TVCG (InfoVis) '08

Visualizations Studied



Animation

Small multiples



Traces

Experiment Design

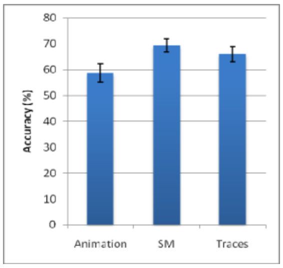
- 3 (animation types) x 2 (data size: small & large) x 2 (presentation vs. analysis)
- Data
  - UN data about countries
- Tasks
  - 24 tasks, 1-3 requires answers per

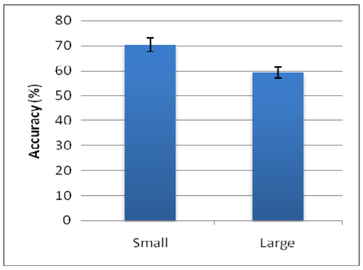
    Example: Select 2 countries with significant decreases in energy consumption

Results

- Accuracy

Measured as percentage correct
65% overall (pretty tough)





Significant:
SM better than animation
Small data size more accurate than large

Results

- Speed
  - Presentation

    Animation faster than small multiples & traces
    15.8 secs vs. 25.3 secs vs. 27.8 secs.
  - Analysis

    Animation slower than small multiples & traces
    83.1 secs. vs. 45.69 secs. vs. 55.0 secs.

Discussion

More on this study coming up on Animation day

- People rated animation more fun, but small multiples was more effective
- As data grows, accuracy becomes an issue
  - Traces & animation get cluttered
  - Small multiple gets tiny
- Animation:
  - "fun", "exciting", "emotionally touching"
  - Confusing, "the dots flew everywhere"

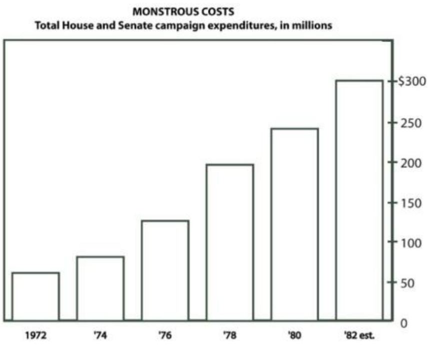
Useful Junk?

Start

- Tufte claimed that graphs loaded with chartjunk are no good
- Is that really so?
- How could you test this?

Comparing

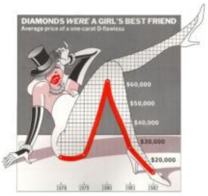
VS.



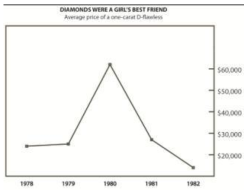
Bateman et al
CHI '10

Methodology

- Two versions of each chart
- Participant sees one
  - Asked immediate interpretation accuracy questions
  - Asked similar questions again 5 minutes or 2-3 weeks later



VS.



Results

- No significant difference in immediate interpretation accuracy, or after 5 minute gap
- After 2-3 week gap, recall of chart topic and details was significantly better for chartjunk graphs
- Participants found the chartjunk graphs more attractive, enjoyed them more, and found them easiest and fastest to remember

Caveats

- Small datasets
- "Normal" charts were really plain
- No interaction
- How about other added interpretations from the flowery visuals?

- Be careful reading too much into this

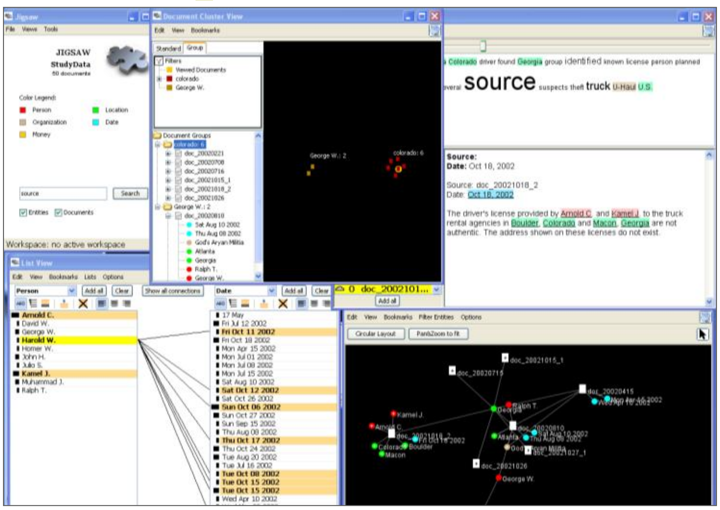
More Complex Task Eval

Start

- Consider investigative analysis tasks involving sensemaking, awareness, and understanding
- Research questions
  – How do people use systems?
  – What characteristics matter?
  – What should we measure/observe?
- Exploring methods for utility evaluation

Kang et al
VAST '08 & *TVCG* '11

System Examined - Jigsaw



Study Design

- Task and dataset   Your HW 8
  - 50 simulated intelligence case reports
    Each a few sentences long
    23 were relevant to plot
  - Identify the threat & describe it in 90 minutes

Source: doc017
Date: Oct 22, 2002

Abu H., who was released from custody after the September 11 incidents and whose fingerprints were found in the U-Haul truck rented by Arnold C. [see doc033] holds an Egyptian passport. He is now known to have spent six months in Afghanistan in the summer of 1999.

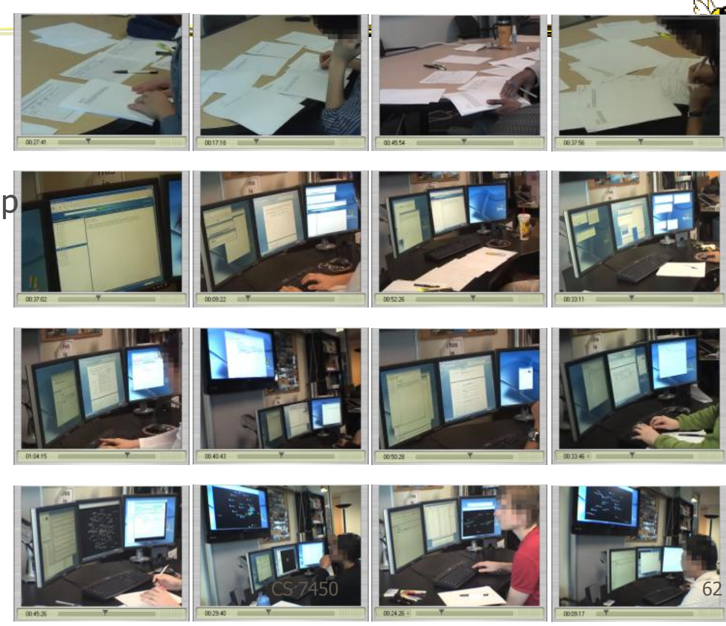
Study Design - Settings

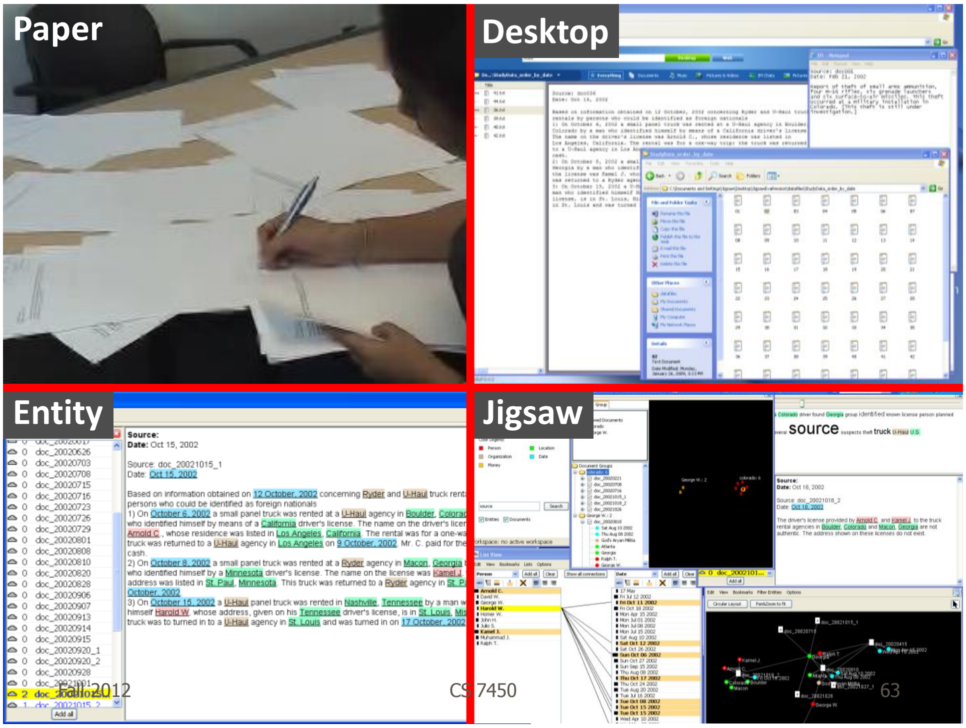
1: Paper

2: Desktop

3: Entity

4: Jigsaw





Performance Measures

- Task sheets (like VAST Contest)
  - Three components (relevant people, events, locations)
  - +1 for correct items, -1 for a misidentified items
- Summary narrative
  - Subjective grading from 1 (low) to 7 (high)
- Two external raters
- Normalized, each part equal, mapped to 100-point scale

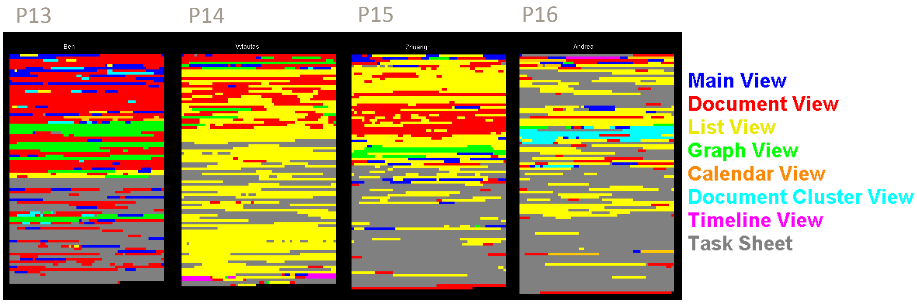
Results

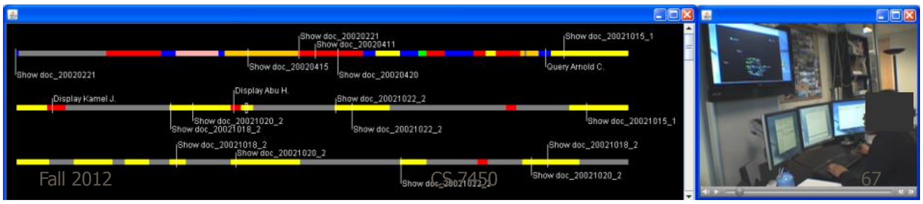
	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Final Score	22.87	65.00	24.26	87.08	62.08	67.13	42.13	29.41	52.23	15.00	29.26	81.19	95.05	58.07	75.20	90.00
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Average Score	49.80				50.19				44.42				79.59			
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23
# of Queries					19	18	48	8	23	61	59	91	44	4	26	8
First Query					40:49	19:55	2:47	12:41	1:31	0:29	0:59	3:12	0:18	5:35	25:37	4:18
Amount of Notes	Many	None	Many	Some	Many	Some	Few	Some	Some	None	None	Few	Some	Few	Few	Few
First Note Taking	0:07		0:05	0:16	1:53	19:57	2:47	8:20	2:37			3:14	0:48	0:32	5:15	78:45
First Task Sheet	43:20	32:53	70:13	3:25	61:35	20:26	7:33	64:11	28:09	0:52	2:55	7:20	48:26	41:48	43:00	5:33

Results

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Final Score	22.87	65.00	24.26	87.08	62.08	67.13	42.13	29.41	52.23	15.00	29.26	81.19	95.05	58.07	75.20	90.00
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Average Score	49.80				50.19				44.42				79.59			
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23
# of Queries					19	18	48	8	23	61	59	91	44	4	26	8
First Query					40:49	19:55	2:47	12:41	1:31	0:29	0:59	3:12	0:18	5:35	25:37	4:18
Amount of Notes	Many	None	Many	Some	Many	Some	Few	Some	Some	None	None	Few	Some	Few	Few	Few
First Note Taking	0:07		0:05	0:16	1:53	19:57	2:47	8:20	2:37			3:14	0:48	0:32	5:15	78:45
First Task Sheet	43:20	32:53	70:13	3:25	61:35	20:26	7:33	64:11	28:09	0:52	2:55	7:20	48:26	41:48	43:00	5:33

Jigsaw Usage Patterns





Investigative Strategies

1. Overview, filter and detail (OFD)
2. Build from detail (BFD)
3. Hit the keyword (HTK)
4. Find a clue, follow the trail (FCFT)

P16: "*I like this people-first approach. Once I identify key people, then things that are potentially important come up, too. I'm an impatient person and don't want to read all documents chronologically.*"

Results by Strategy

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23

Results by Strategy

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23

Results by Strategy

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23

Results by Strategy

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23

Results by Strategy

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Strategy Used	OFD	OFD	BFD	OFD	OFD	OFD	FCFT	BFD	BFD	HTK	HTK	FCFT	FCFT	HTK	OFD	FCFT
Performance	Fair	Very good	Fair	Excel-lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel-lent	Excel-lent	Good	Very good	Excel-lent
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23

Tool Design Implications

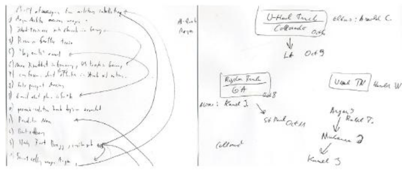
- Support finding starting points/clues
- Guide the analyst to follow the right trail
- Support different strategies of SM process
- Support smooth transition between SM stages
- Provide a workspace
- Allow flexibility in organizing
- Support to find next steps when dead-end
- Facilitate further exploration

Jigsaw's Influence

- Supporting different strategies
- Showing connections between entities

- Helping users find the right clue
- Helping users focus on essential information

- Reviewing hypotheses
- Increasing motivation



Evaluation Recommendations

- Compare system usage to traditional methods
- Collect qualitative data, support with quantitative data
- Consider questions to be answered
- Possible metrics
  - Number of documents viewed
  - When note-taking initiated
  - The quantity of representations created
  - Amount of time and effort in organizing
  - Time spent in reading/processing relevant information

How to Evaluate Many Eyes?

Start

- Two main evaluation papers written about system
- Studied use of system, visualizations being created, discussions about system, etc.

Paper 1

- Case study of early use
- System uses
  - Visual analytics
  - Sociability
  - Generating personal and collective mirrors
  - Sending a message

Viégas et al
HICSS '08

Use Characteristics

Data Topic/Area	Percentage
Society	14.0
Economics	12.7
Obscured/Anon	12.4
Art & culture	10.8
Web & new media	10.3
Science	10.0
Test data	9.5
Politics	7.4
Technology	6.6
…	

Comment Type	Percentage
Observation	46.3
Question	15.8
Affirmation	13.7
Hypothesis	11.6
Socializing	11.6
System design	11.6
Data integrity	9.5
Testing	4.2
Tips	4.2
To do	4.2

Paper 2

- Interview-based study
- Individual phone interviews with 20 users
  - Lots of quotes in paper
- Bloggers vs. regular users
- Also includes stats from usage logs
  - 3069 users
  - 1472 users who uploaded data
  - 5347 datasets
  - 972 users who created visualizations
  - 3449 visualizations
  - 222 users who commented
  - 1268 comments

Danis et al
CHI '08

Findings

- User motivations
  - Analyzing data
  - Broadening the audience, sharing data
- Lots of collaborative discussion
  - Much off the ManyEyes site
- Concerns about data and other eyes

Specific to Infovis?

- How about evaluation techniques specifically focused on infovis?

Insight

- Isn't one of the key ideas about InfoVis that it helps generate insights?
- OK, well let's count/measure insights

- What challenges do you see in this?

Start

Problem Domain

- Microarray experiments: Gain insight into the extremely complex and dynamic functioning of living cells
- Systems-level exploratory analysis of thousands of variables simultaneously
- Big data sets

Saraiya, North, Duca
TVCG '05

Insight

- Insight: An individual observation about the data by the participant, a unit of discovery
- Characteristics
  - Observation
  - Time
  - Domain Value
  - Hypotheses
  - Directed vs Unexpected
  - Category

Insight Characteristics

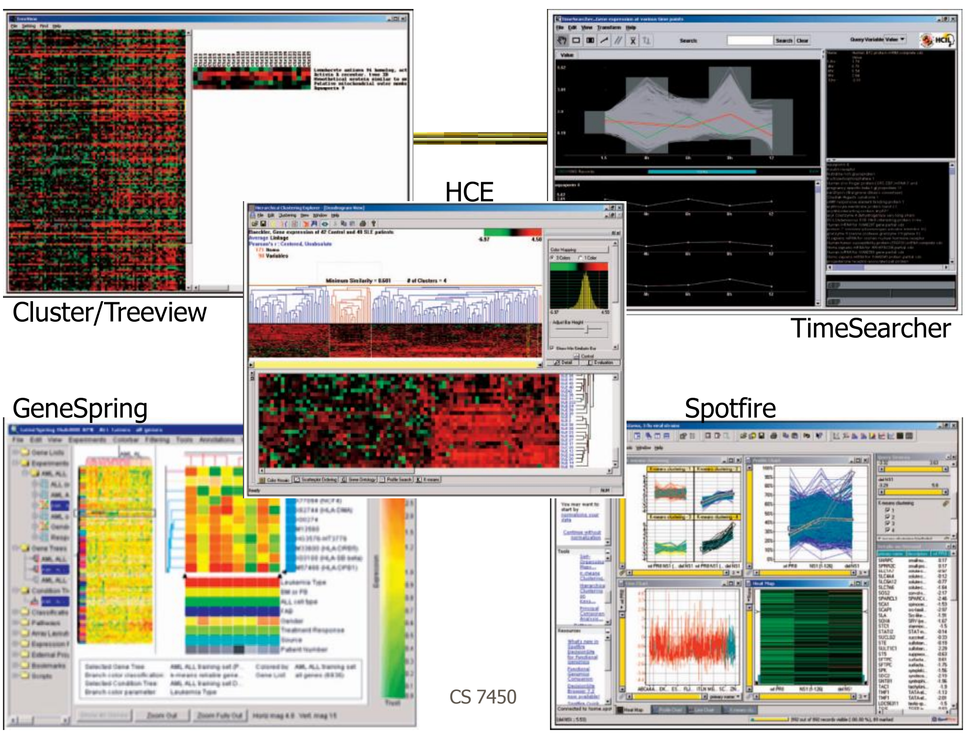
- Complex
  - Involving large amounts of data in a synergistic way
- Deep
  - Builds over time, generates further questions
- Qualitative
  - Can be uncertain and subjective
- Unexpected
  - Often unpredictable, serendipitous
- Relevant
  - Deeply embedded in data domain, connecting to existing domain knowledge

North
CG&A `06

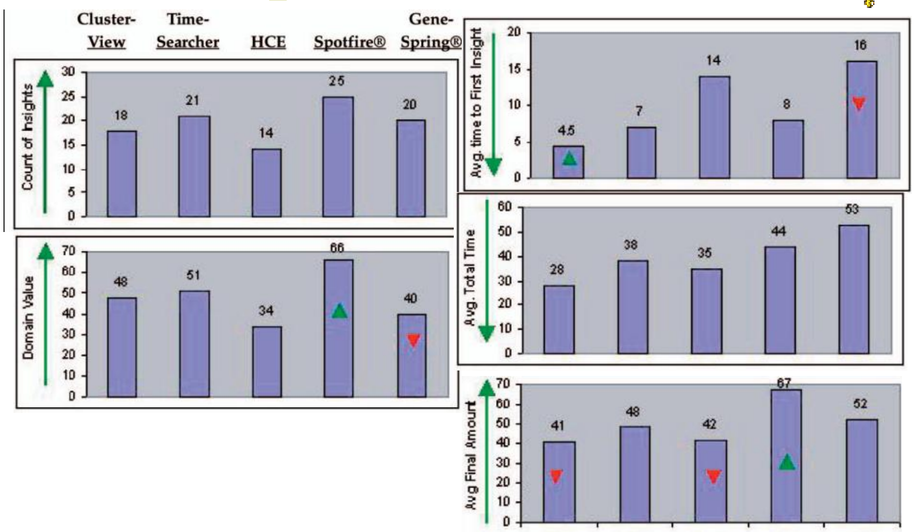
Experiment Design

- Data: Timeseries, Virus, Lupus

Tool	Visual Representations	Interactions
Cluster/ Treeview	Heat-map, Clustered heat-map	O+D
Time-Searcher	Parallel coordinates, line graph	Brushing, O+D, DQ
HCE	Cluster dendrogram, parallel coordinates, heat-map, scatterplot, histogram	Brushing, Zooming, O+D, DQ
Spotfire® 7.2 Functional Genomics	Parallel coordinates, heat-map, scatterplots (2D/3D), histogram, bar/pie chart, tree view, spreadsheet view, Clustered parallel coordinates	Brushing, Zooming, O+D, DQ
GeneSpring ® 5.0	Parallel coordinate, heat-map, scatter-plots (2D/3D), histogram, bar chart, block view, physical position view, array layout view, pathway view, spreadsheet view, compare gene to gene, Clusterested parallel coordinates	Brushing, Zooming



Results



Discussion

- Methodology difficulties
  - Labor intensive
  - Requires domain expert
  - Requires motivated subjects
  - Training and trial time

- Weakness: Short session time (2 hours) when long-term use more desirable

Reconsidering Insight

- Insight with visualization
  - Is not spontaneous "aha!" moments (eg, in cognitive science)
  - Is knowledge-building and model-confirmation
    - Like a substance that people acquire with the aid of systems

Chang et al
CG&A '09

Rethinking Methodology

- Do controlled lab experiments really tell us very much in information visualization?

MILC Technique

- **M**ulti-dimensional
  - observations, interviews, surveys, logging
- **I**n-depth
  - intense engagement of researchers with domain experts so as to almost become a partner
- **L**ong-term
  - longitudinal use leading to strategy changes
- **C**ase Study
  - detailed reporting about small number of people working on their own problems in their own domain

Shneiderman & Plaisant
BELIV '06

Influences

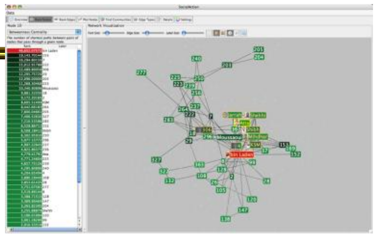
- Ethnography
  - Preparation
  - Field study
  - Analysis
  - Reporting

Guidelines

- Specify focused research questions & goals
- Identify 3-5 users
- Document current method/tool
- Determine what would constitute professional success for users
- Establish schedule of observation & interviews
- Instrument tool to record usage data
- Provide attractive log book for comments, problems, and insights
- Provide training
- Conduct visits & interviews
- Encourage users to continue using best tool for task
- Modify tool as needed
- Document successes and failures

SocialAction

Start



- Evaluation inspired by MILC ideas goals
  - Interview (1 hour)
  - Training (2 hours)
  - Early use (2-4 weeks)
  - Mature use (2-4 weeks)
  - Outcome (1 hour)

Perer & Shneiderman
CHI '08

Methodology

- Four case studies
  - Senatorial voting patterns
  - Medical research knowledge discovery
  - Hospital trustee networks
  - Group dynamics in terrorist networks
- Named names
  - I like it!
- Tell what they did with system

My Reflections

- Nice paper
- Stark contrast to comparative, controlled experiments
- We likely need more of this in InfoVis

Summary

- Why do evaluation of InfoVis systems?
  - We need to be sure that new techniques are really better than old ones
  - We need to know the strengths and weaknesses of each tool; know when to use which tool

Challenges

- There are no standard benchmark tests or methodologies to help guide researchers
  - Moreover, there's simply no one correct way to evaluate
- Defining the tasks is crucial
  - Would be nice to have a good task taxonomy
  - Data sets used might influence results
- What about individual differences?
  - Can you measure abilities (cognitive, visual, etc.) of participants?

Challenges

- Insight is important
  - Great idea, but difficult to measure
- Utility is a real key
  - Usability matters, but some powerful systems may be difficult to learn and use
- Exploration
  - InfoVis most useful in exploratory scenarios when you don't know what task or goal is

    So how to measure that?!

Project

- Demos on Thursday
- 15 minutes, plan appropriately, leave open-ended exploration time
- Bring page or two describing project

- Video due next Monday

Upcoming

- Review & recap
  - Reading
    - Few chapter 13
    - Heer et al ’10