Visualization Evaluation

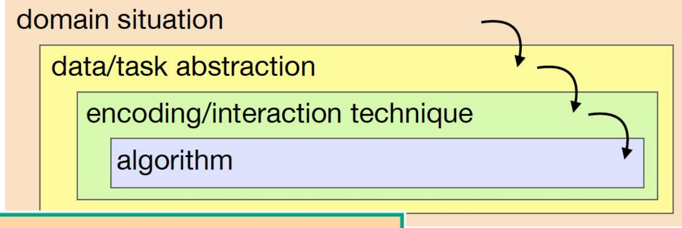
Chen He

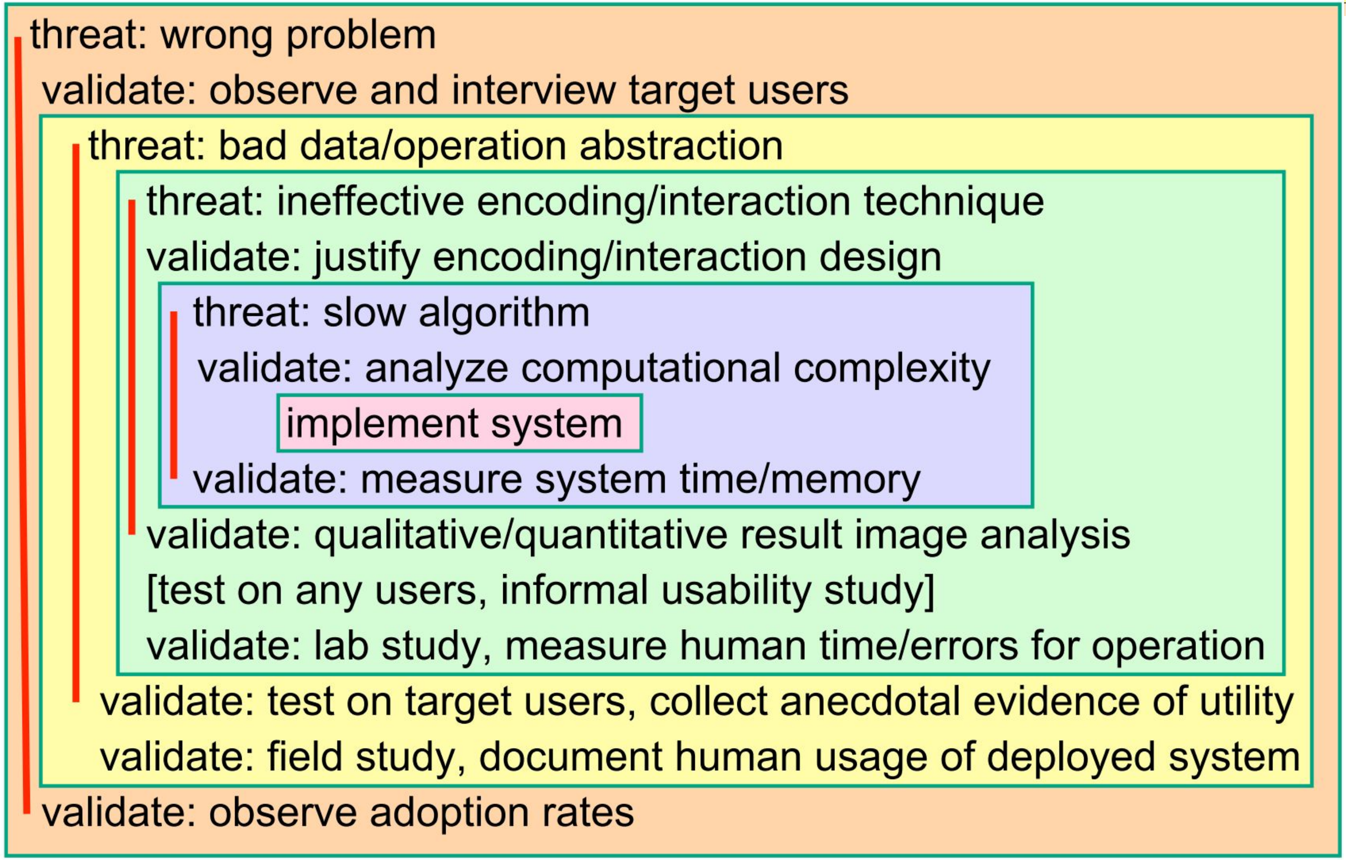
After this lecture, you will be able to

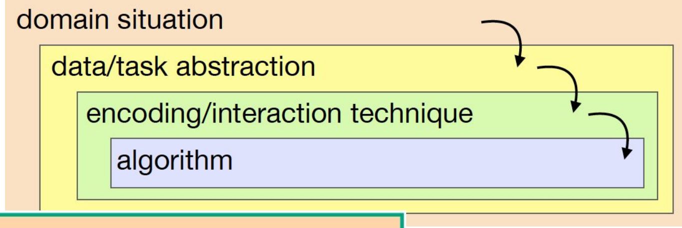
Describe the process of **quantitative** user study;

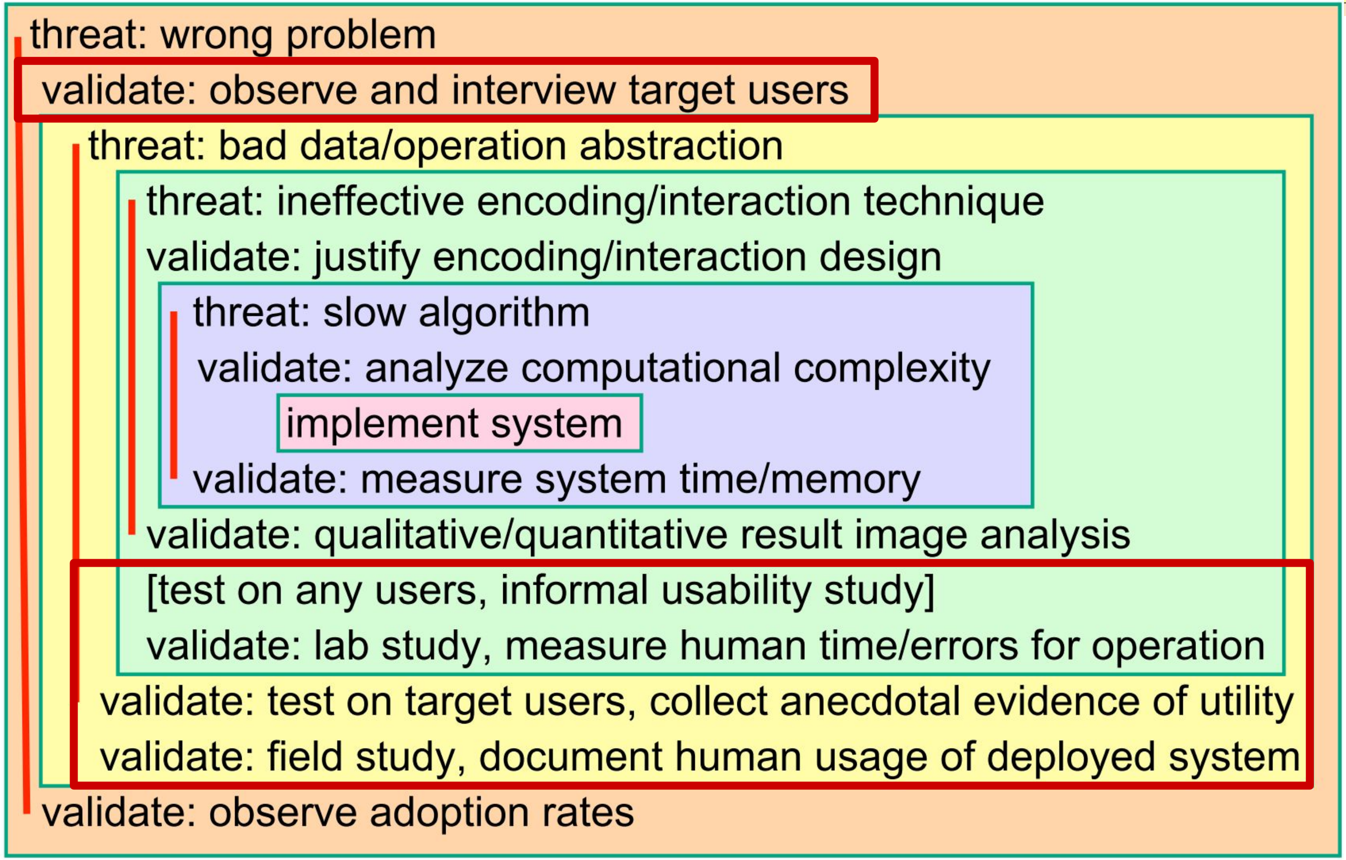
List common methods for **qualitative** user study;

Recognize insight-based user study and crowdsourcing and when to use them.


domain situation
data/task abstraction
encoding/interaction technique
algorithm


threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates


domain situation
data/task abstraction
encoding/interaction technique
algorithm


threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Carpendale, Sheelagh. Evaluating information visualizations. *Information visualization*. Springer, 2008.

Quantitative evaluation

Hypothesis development

Identification of the independent variables
Control of the independent variables
Elimination of complexity

Observation, measurement of
the dependent variables

Application of statistics

Andolina, Salvatore, et al. Intentstreams: smart parallel search streams for branching exploratory search. *IUI*. 2015.

Case study: IntentStreams

Independent variables & Hypothesis

IntentStreams: a system supporting **parallel browsing** and **branching** during search without the need to open new tabs.

Baseline: A traditional Google Search interface.

Compared to the baseline, IntentStreams generates (1) more parallel streams, (2) more revisits, and (3) more branches.

Task & procedure

You have to write an **essay** on recent developments of X where you have to cover **as many subtopics as possible**. You have 20 minutes to collect the material that will provide inspiration for your essay. You have additional 5 minutes to write your essay. Two topics: (1) NASA, and (2) China Mobile.

Before the experiment, participants received detailed instructions on how to use the system and performed a 5-minute training session.

Elimination of complexity

Same dataset used: a news repository with more than 25 million English language editorial news articles.

Within-subjects design: Counterbalanced by changing the order of the topics and the systems.

13 participants screened to have little knowledge about the two topics.

Dependent variables

Number of parallel streams (tabs opened for the baseline);

Number of revisits;

Number of branches.

Application of statistics

IntentStreams on average generated

7.84 more queries ($SD = 7.27$),

6.38 more parallel streams ($SD = 4.03$),

4.54 more revisits ($SD = 4.52$),

3.62 more branches ($SD = 4.01$).

Paired t-tests indicate that all those differences are statistically significant ($p < 0.01$).

Findings

Parallel search supported in IntentStreams.

Branching supported in IntentStreams.

IntentStreams supports **more exploration** seen from the higher number of queries.

https://www.nngroup.com/articles/between-within-subjects/

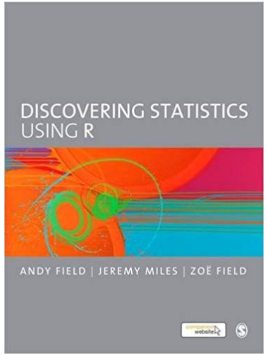
Remark

Within-subjects (the same person tests all the conditions)	Require fewer participants; Minimize the random noise.
Between-subjects (different people test each condition)	Minimizes the learning and transfer across conditions; Shorter sessions, less tiring; Easier to set up.

Remark

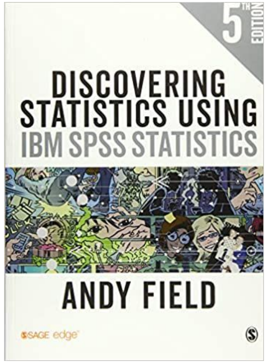
More statistics in Human-computer interaction CSM13401.

Prof. Bart Knijnenburg. https://www.usabart.nl/QRMS/



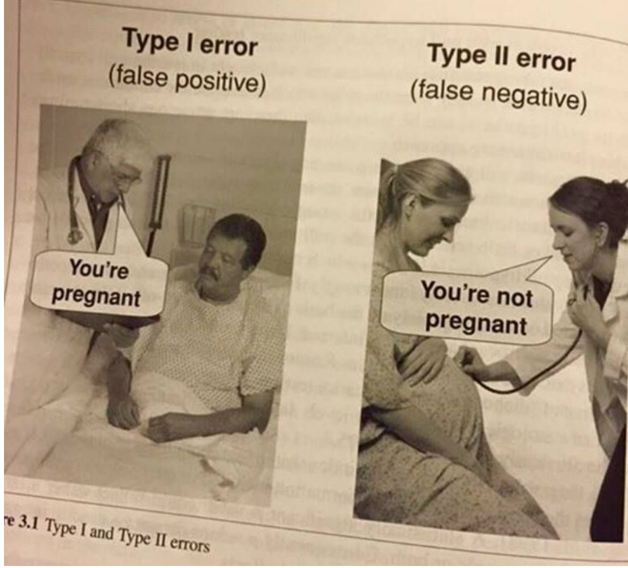
Books:

Andy Field, Jeremy Miles, and Zoë Field. Discovering statistics using R, 1st Edition. SAGE Publications Ltd, 2012.



Andy Field. Discovering Statistics Using IBM SPSS Statistics, 5th Edition. SAGE Publications Ltd, 2018.

Challenges



Conclusion validity
Is there a relationship between the independent and dependent variables?
Type I (false positive) and Type II Errors (false negative)

Internal validity
Is the relationship causal?
That is, are there possible alternate causes for the results seen in the study?

Challenges

Construct validity
Whether the experiment has been designed and run in a manner that answers the intended questions.

External validity
Can we **generalize** the study results to other people/places/times?

Ecological validity
How closely the experimental setting matches the real setting?

https://en.wikipedia.org/wiki/Observer-expectancy_effect
https://en.wikipedia.org/wiki/Demand_characteristics

Potential confounding factors

Experimenter effect

A researcher's cognitive bias causes them to subconsciously influence the participants of an experiment.

Demand characteristic

Participants form an interpretation of the experiment's purpose and subconsciously change their behavior to fit that interpretation.

Qualitative evaluation

Think-aloud protocol

Interview

Questionnaire

Expert review

etc.

Carpendale, Sheelagh. Evaluating information visualizations. *Information visualization*. Springer, 2008.

Think-aloud protocol

Encourage participants to **speak their thoughts** as they progress through the experiment.

+ Provide insights by hearing participants' thoughts, plans, and frustrations;
- Not natural, reducing the realism of the study.

Carpendale, Sheelagh. Evaluating information visualizations. *Information visualization*. Springer, 2008.

Interview

Sort out the right questions to ask;

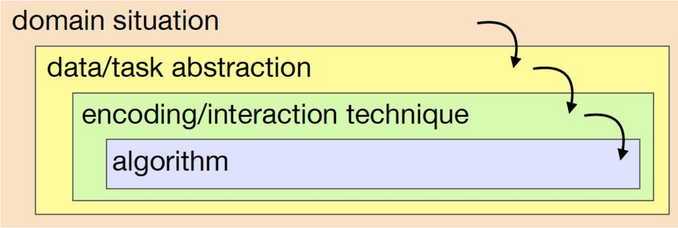
Actively listen to what the participant says.

Alspaugh, Sara, et al. Futzing and moseying: Interviews with professional data analysts on exploration practices. *TVCG* 2018.

Case study: Understand the domain situation

Interview 30 professional data analysts to

- understand "data exploration" in practices;
- guide future tool development.



Case study: Interview questions

What are **typical data exploration scenarios**?

How does data exploration **relate to the other parts** of analysts' workflow?

What are the **tedious** parts of data exploration? What are the most **challenging** parts?

What **tools and techniques** do analysts use to explore data?

Do analysts use (interactive) visualizations? If so, how?

What **automation** have analysts developed for themselves to facilitate exploration?

If advanced automation could be harnessed to help analysts explore data, how would they like this to work, ideally?

Which features do they most **appreciate** about the tools they use, and which are they **lacking**?

Case study: Recommendations for tool development

Combine **direct manipulation** with **command line** tools.

Make it easier to create **reusable modules** to encapsulate common workflows in analysis tools.

Continue to research and develop tools for **recording history and provenance** of both analysis and data.

etc.

Questionnaire

System Usability Scale (SUS)

User Experience Questionnaire (UEQ)

The NASA Task Load Index (NASA-TLX)

User Engagement Scale (UES)

Recommender systems' Quality of user experience (ResQue)
https://hci.epfl.ch/research-projects/resque/

Etc.

https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

System Usability Scale (SUS)

Score the following 10 items in a Five-point Likert scale from Strongly Agree to Strongly disagree.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

System Usability Scale (SUS) - Interpreting scores

Convert the original scores of 0-40 to 0-100

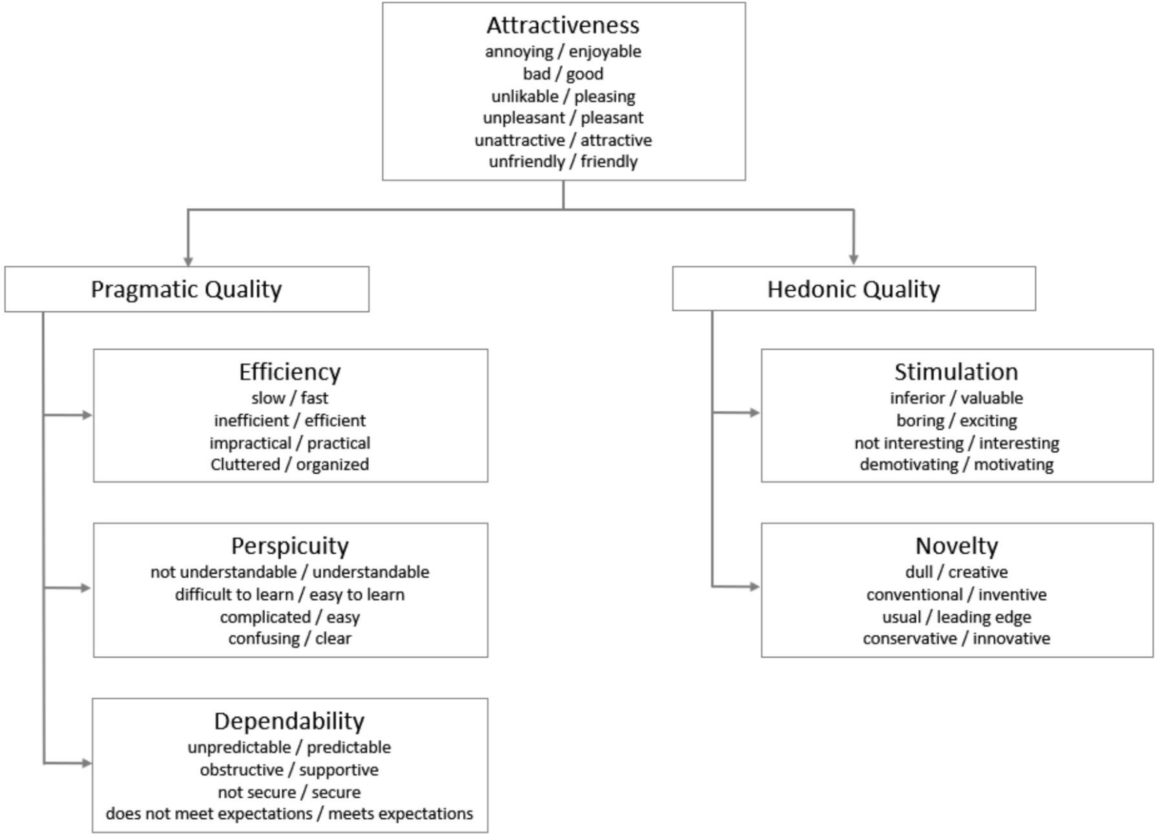
Based on research,

above 68 would be considered above average

below 68 is below average.

https://www.ueq-online.org/

User Experience Questionnaire (UEQ)



The NASA Task Load Index (NASA-TLX)

Mental Demand

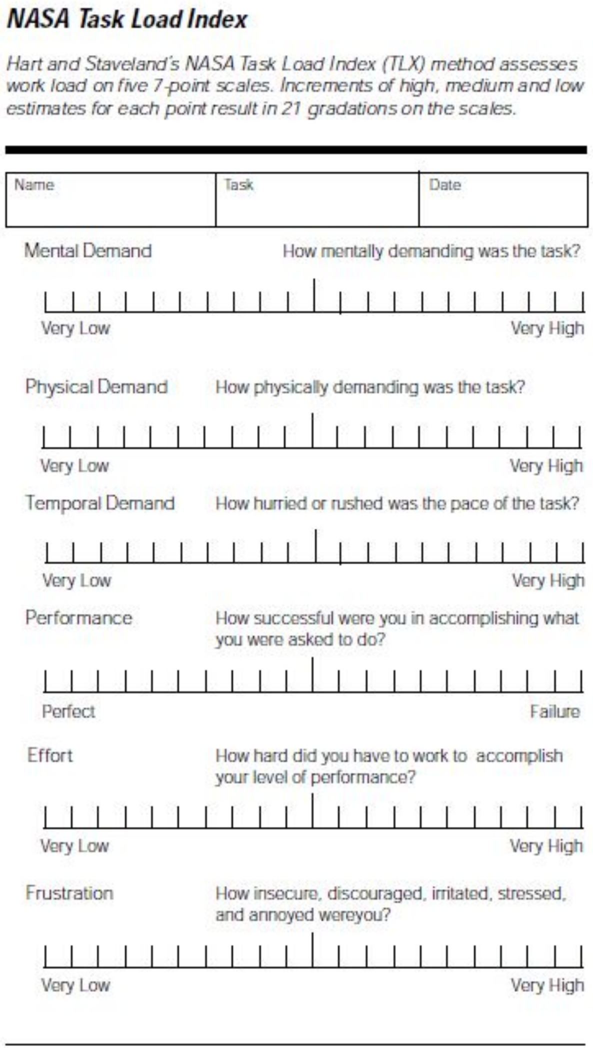
Physical Demand

Temporal Demand

Overall Performance

Effort

Frustration Level



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

O'Brien et al. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies, 2018*.

User Engagement Scale (UES)

Focused attention	FA-S.1	I lost myself in this experience.
	FA-S.2	The time I spent using Application X just slipped away.
	FA-S.3	I was absorbed in this experience.
Perceived usability	PU-S.1	I felt frustrated while using this Application X.
	PU-S.2	I found this Application X confusing to use.
	PU-S.3	Using this Application X was taxing.
Aesthetic appeal	AE-S.1	This Application X was attractive.
	AE-S.2	This Application X was aesthetically appealing.
	AE-S.3	This Application X appealed to my senses.
Reward factor	RW-S.1	Using Application X was worthwhile.
	RW-S.2	My experience was rewarding.
	RW-S.3	I felt interested in this experience.

Expert review

No fixed structure

Guide exploration + interview

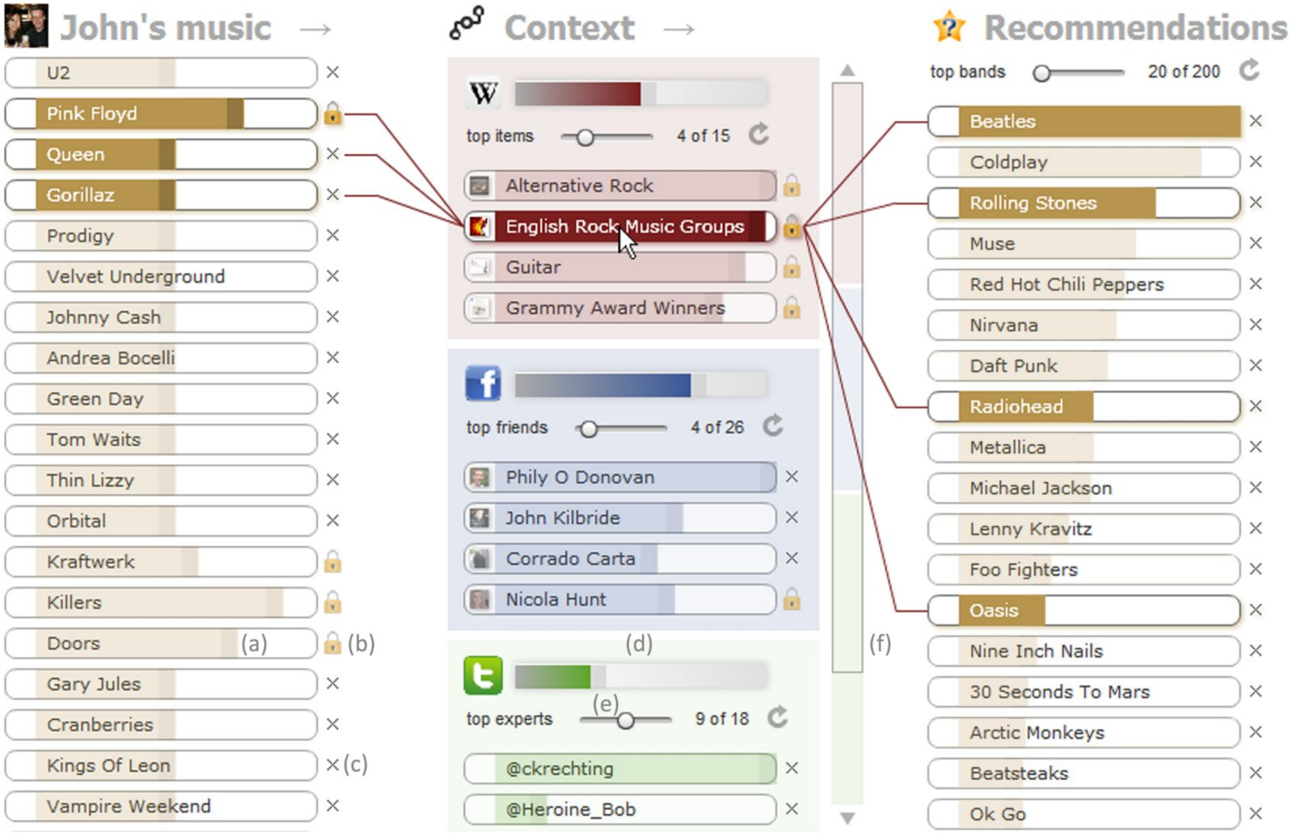
Compare with experts normal workflow

Combining qualitative and quantitative measures

Qualitative methods can help clarify quantitative data by providing missing explanatory details.

Bostandjiev, S., O'Donovan, J. and Höllerer, T., TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems* (pp. 35-42). 2012

Case study: TasteWeights



Research questions

What (if any) is the **benefit** of explaining a hybrid recommendation process through a user interface?

How does **interaction** at recommendation time affect **accuracy and user experience**?

Independent & dependent variables

Condition 1: View and Adjust the left column;

Condition 2: View and adjust the left two columns;

Condition 3: Full version (see the results of the adjustment).

Dependent variable: Recommendation accuracy

Task & procedure

Participants: 32 university students from 10 different majors.

Within-subjects design.

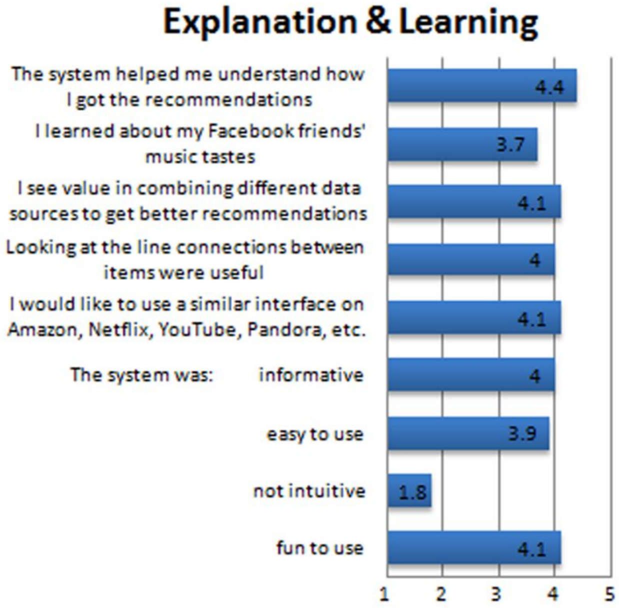
Task: Tweak the system under each of the conditions.

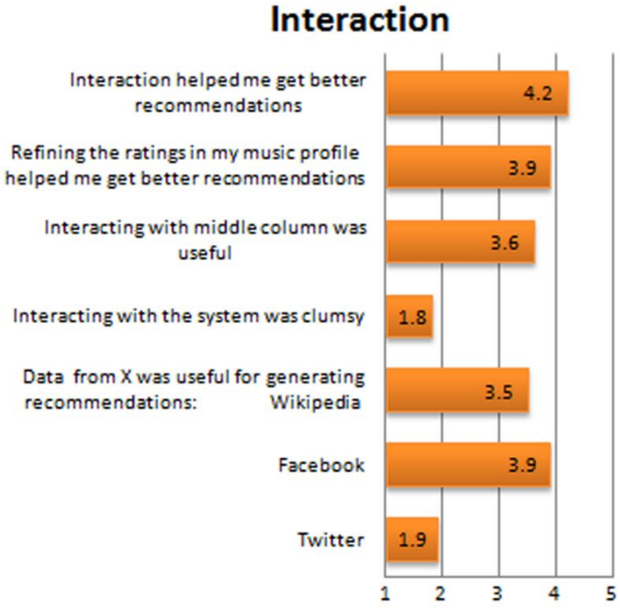
After that, rate a randomized list of recommendations.

Pre-study questionnaire, training, task, post-questionnaire.

Result

The full interaction one achieved the highest accuracy score of 7.54.
Combining qualitative measures:





Limitations

Not a fair scientific comparison.

But it shows interactive feedback can be beneficial.

Answers to research questions

What (if any) is the benefit of explaining a hybrid recommendation process through a user interface?

Explaining a hybrid recommendation process through a user interface can increase user satisfaction.

How does interaction at recommendation time affect accuracy and user experience?

Interaction at recommendation time can improve recommendation accuracy and user experience.

Insight-based evaluation

The purpose of visualization is insight, not pictures. -- Ben Shneiderman

As opposed to task-based evaluation that measures task time and accuracy.

How to evaluation the **exploratory** feature of visualization.

Proposed by Saraiya et al. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations, 2005.

Insight-based evaluation

threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

An Insight-Based Methodology for Evaluating Bioinformatics Visualizations

Insight: an individual observation about the data by the participant, a unit of discovery.

Participant: Think-aloud protocol.

Evaluator: identify and codify all individual occurrences of insights.

“While most genes showed higher expression value for the Lupus group as compared to the Control group, there were other genes that were less expressed for the Lupus group. ”

To quantify insights

Number of insights

Time to first insight

Domain value

Directed versus unexpected

Correctness

Breadth versus depth

Category

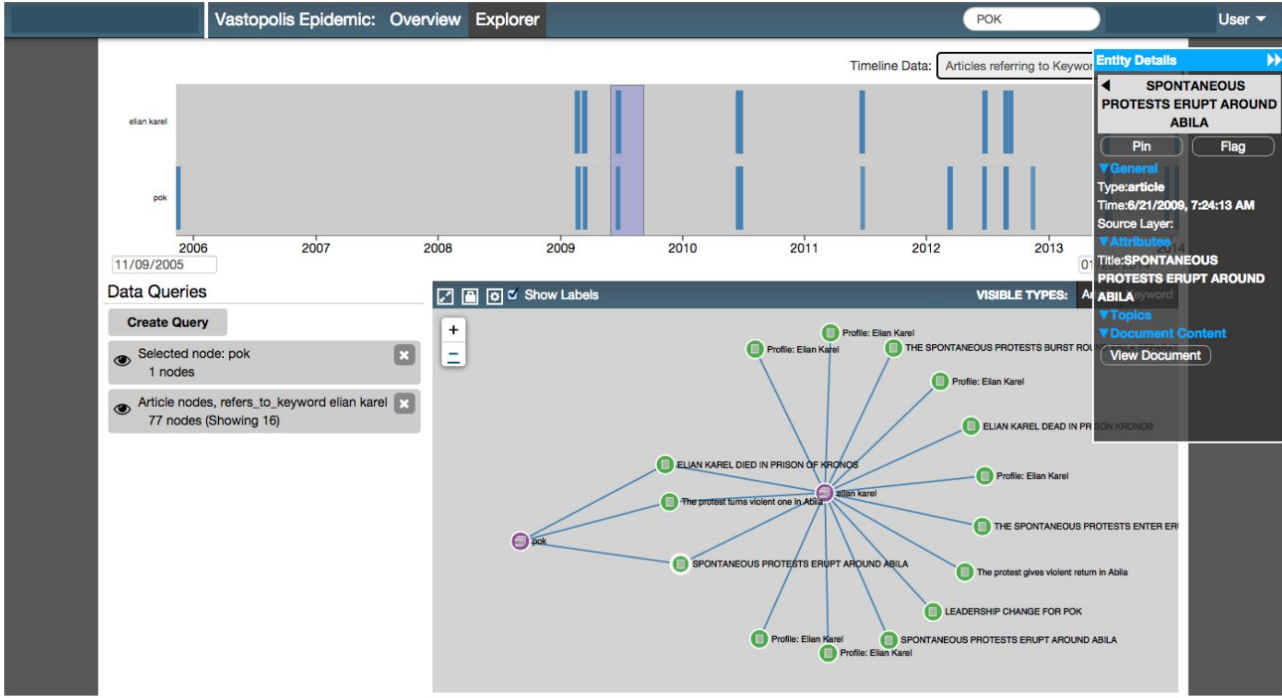
Guo, Hua, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics* 22, no. 1 (2015): 51-60.

Case study: Research goal

To understand how interactions lead to insight generation.

Interactions:

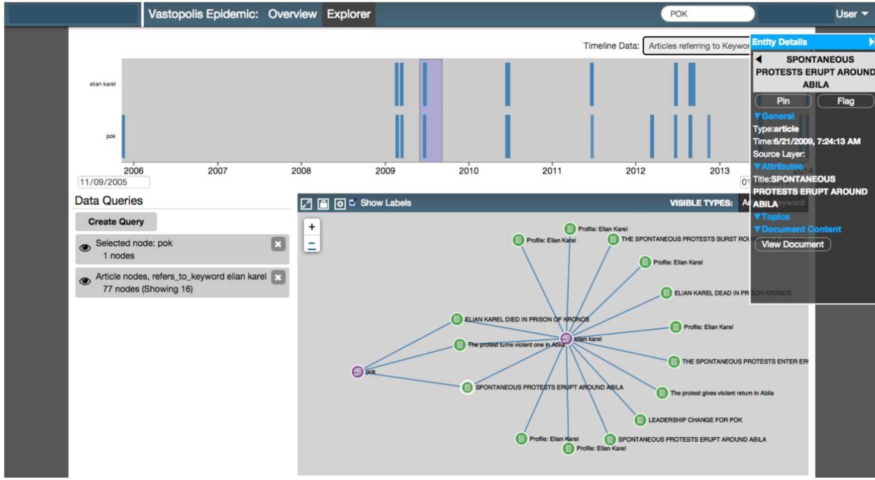
Select
Explore
Elaborate
Reconfigure
Filter
Connect
Retrieve



Task

The dataset contains texts, such as news reports, resumes, and email headers, relevant to **a disappearance case** that happened in a fictional country.

Task: Analyze the dataset using the visual analysis system and **identify possible explanations** behind the disappearance case, with supporting evidence from the dataset.



Procedure

1. 15-20 minutes training;
2. 45 minutes on the analysis task with a think-aloud protocol: explain the analysis processes and report insights as clearly as possible.

Extract interaction patterns

Orienting (Reconfigure – Explore – Elaborate)

Locating (Retrieve – Elaborate – Elaborate, Elaborate – Retrieve – Elaborate)

Sampling (Explore – Elaborate – Elaborate – Elaborate, Explore – Elaborate – Elaborate)

Elaborating (Elaborate – Elaborate – Elaborate)

Code insights

Term	Definition	Example
Fact	A statement that is true given the VAST Challenge 2014 dataset and describes the existence or properties of an event or an entity	"The police questioned a Gastech employee named Elian Karel after the disappearance."
Generalization	A statement that describes connections among entities relevant to the disappearance case	"There's one Gastech employee who shares the same last name with a POK member."
Hypothesis	A hypothetical statement relevant to the disappearance case	"Henk might be motivated to join POK because his wife was sick due to the mess of the environment."

Two coders categorized the insights independently.

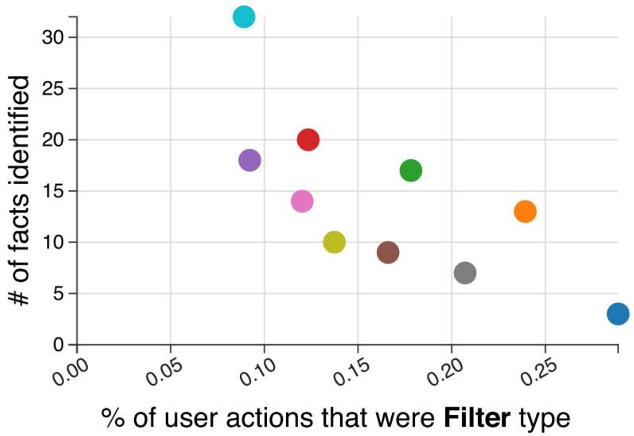
The correlation between the coding results from the two authors is 92.66%.

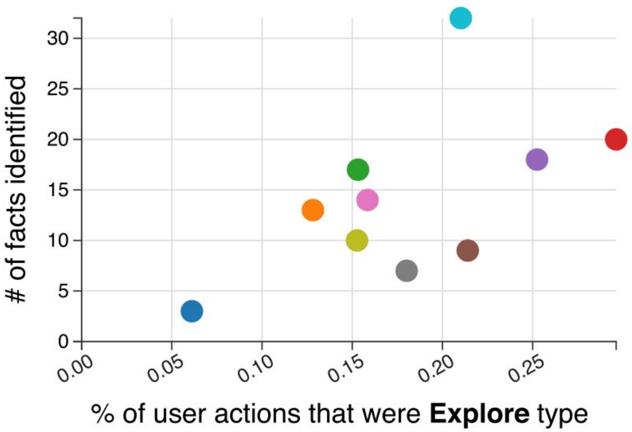
Then discussed the coding results to resolve some of the inconsistencies.

Results

Exploration actions **foster** insights, whereas **filtering** actions **inhibit** insights.

Sampling pattern has a moderate **positive** correlation with the number of **generalizations**; **Elaborating** pattern has a moderate **negative** correlation with the number of **generalizations**.





Crowdsourcing

A new labor market phenomenon where **simple, often monotonous labor tasks** are replaced by open self-managed recruitment of **large groups of people** from the general public.

Prolific https://www.prolific.co/

Amazon Mechanical Turk https://www.mturk.com/

HIT: Human Intelligence Task

Crowdsourcing vs. Lab study

Controlled lab studies:

- Small samples;
- Participants with narrow demographic backgrounds.

Crowdsourcing:

+ Large samples;
+ Diverse samples;
+ Easier and faster data collection;
- Limited data collection methods.

Crowdsourcing

Large samples offset individual differences.

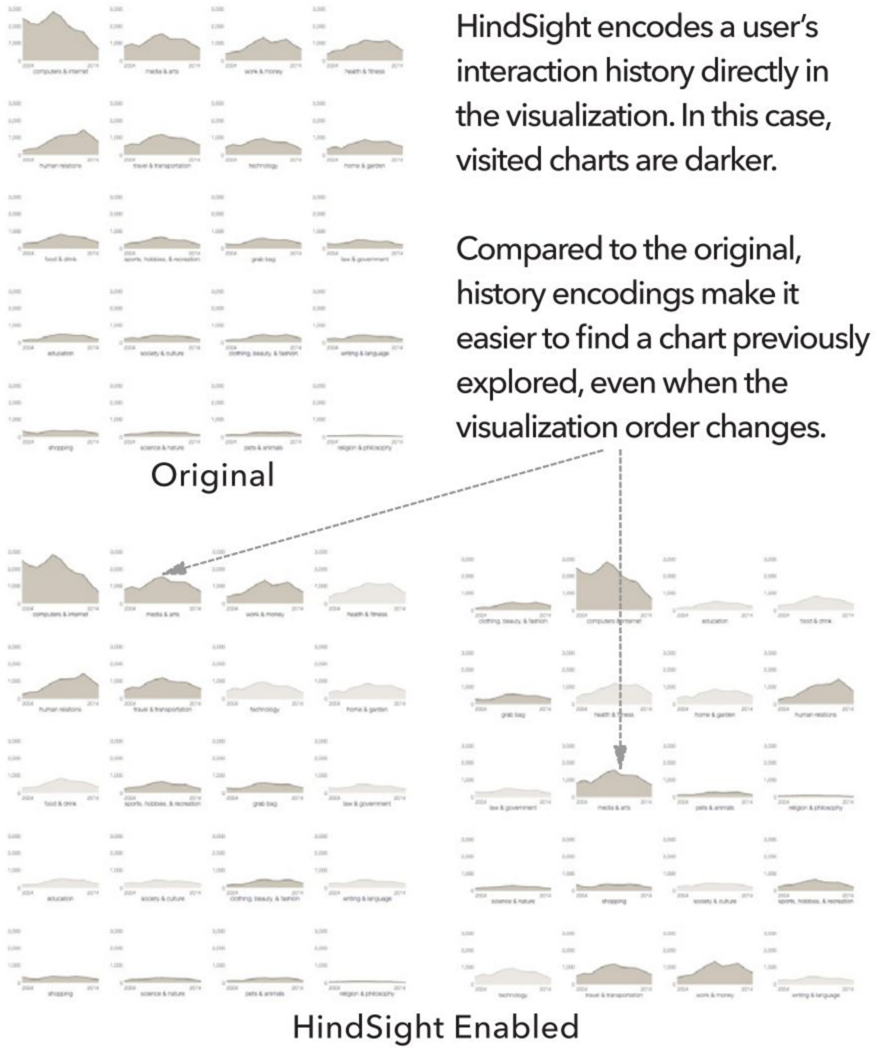
Within-subjects (the same person tests all the conditions)	Require fewer participants; Minimize the random noise.
Between-subjects (different people test each condition)	Minimizes the learning and transfer across conditions; Shorter sessions, less tiring; Easier to set up.

Pilot study is critical.

Automatically validate the collected data to approve or reject the work.

Feng, M., Deng, C., Peck, E. M., & Harrison, L. (2016). Hindsight: Encouraging exploration through direct encoding of personal interaction history. *TVCG*.

Case study: HindSight



Research questions

How does HindSight impact **exploration behavior** such as number of charts visited, total time spent exploring the data, and patterns of exploration?

How does HindSight impact the **insights** that people recall immediately after interacting with a visualization?

Task & procedure

Between-subjects design.

Exploration: without time limit. After they finish, they advanced to the Insight phase through a button press.

Measures

Visited charts

Revisited charts

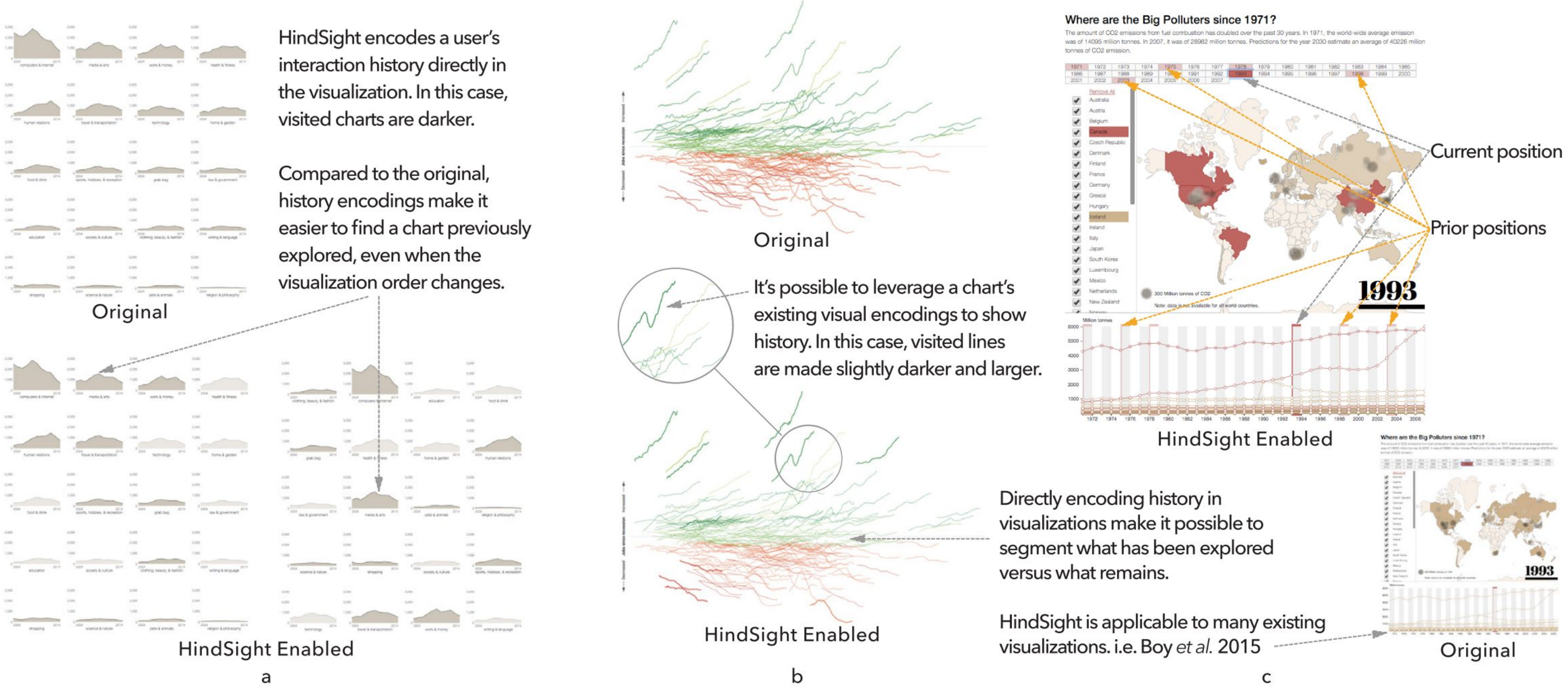
Exploration time

Mentions of charts in insights

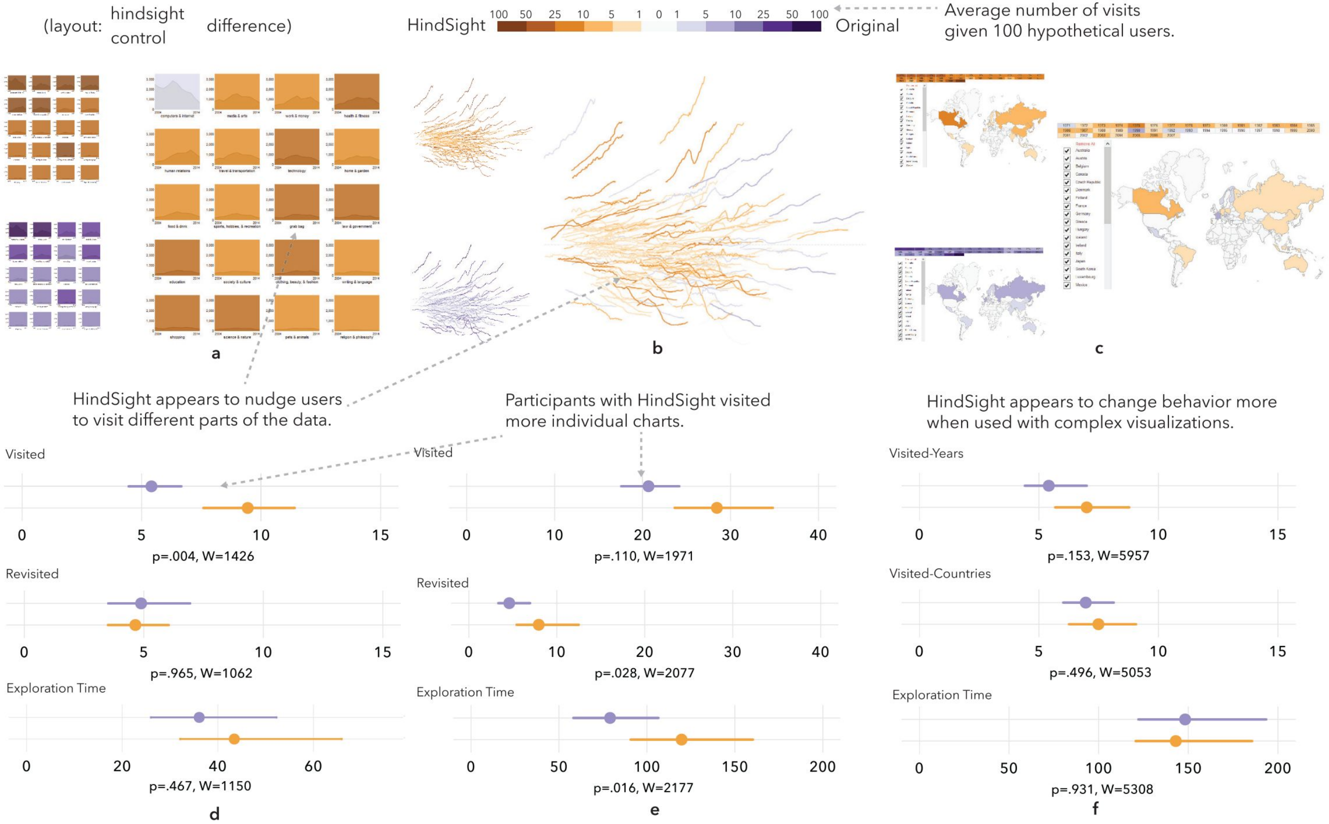
Power analysis deciding on the number of subjects: https://www.statmethods.net/stats/power.html

Data collection

Participants:	92	116	206
Insights:	363	492	831



Findings -- Behavior analysis

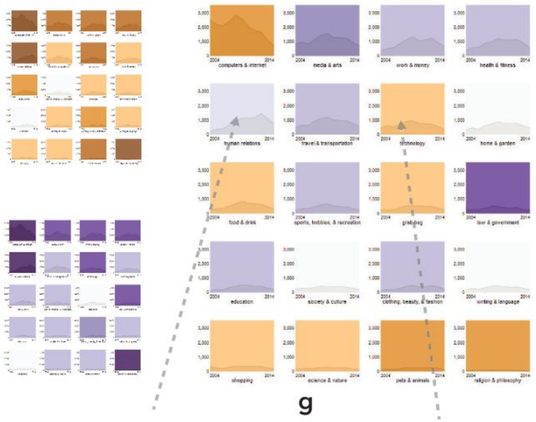


Findings -- Insight analysis

(layout: hindsight control difference)

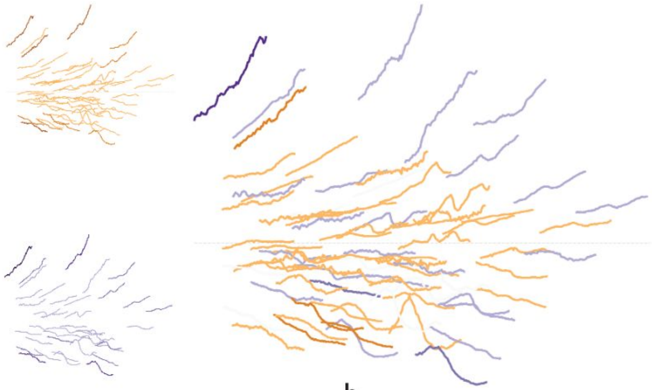
HindSight  Original

Average number of mentions given 100 hypothetical users.



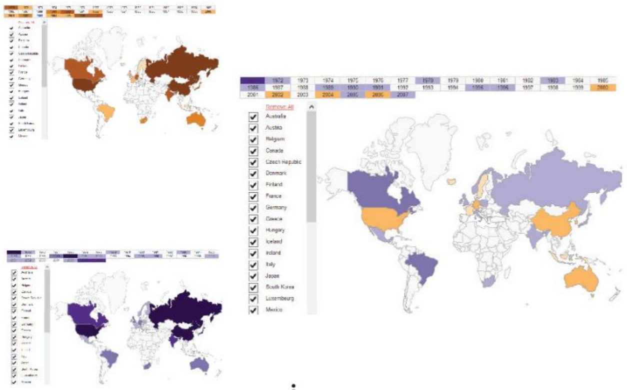
g

HindSight also impacted the insights users generated after exploring datasets.



h

In 255charts, users in the HindSight condition mentioned more charts of the center, while those in the control focused onthe perimeter.



i

Again, in low information datasets HindSight appears to have some effect, but not much.

Recommended readings on crowdsourcing

Borgo, Rita, et al. Information visualization evaluation using crowdsourcing. Computer Graphics Forum. Vol. 37. No. 3. 2018.

Borgo, Rita, et al. Crowdsourcing for information visualization: Promises and pitfalls. Evaluation in the crowd. Crowdsourcing and human-centered experiments. Springer, Cham, 2017. 96-138.

Willett, Wesley, Jeffrey Heer, and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012.

Recap

Quantitative evaluation

Qualitative evaluation

Insight-based evaluation

Crowdsourcing

BELIV Workshop

Evaluation and Beyond - Methodological Approaches for Visualization

A well-known venue that encourages the study of novel evaluation methods.

https://beliv-workshop.github.io/