Visual Analytics– Evaluation Techniques

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

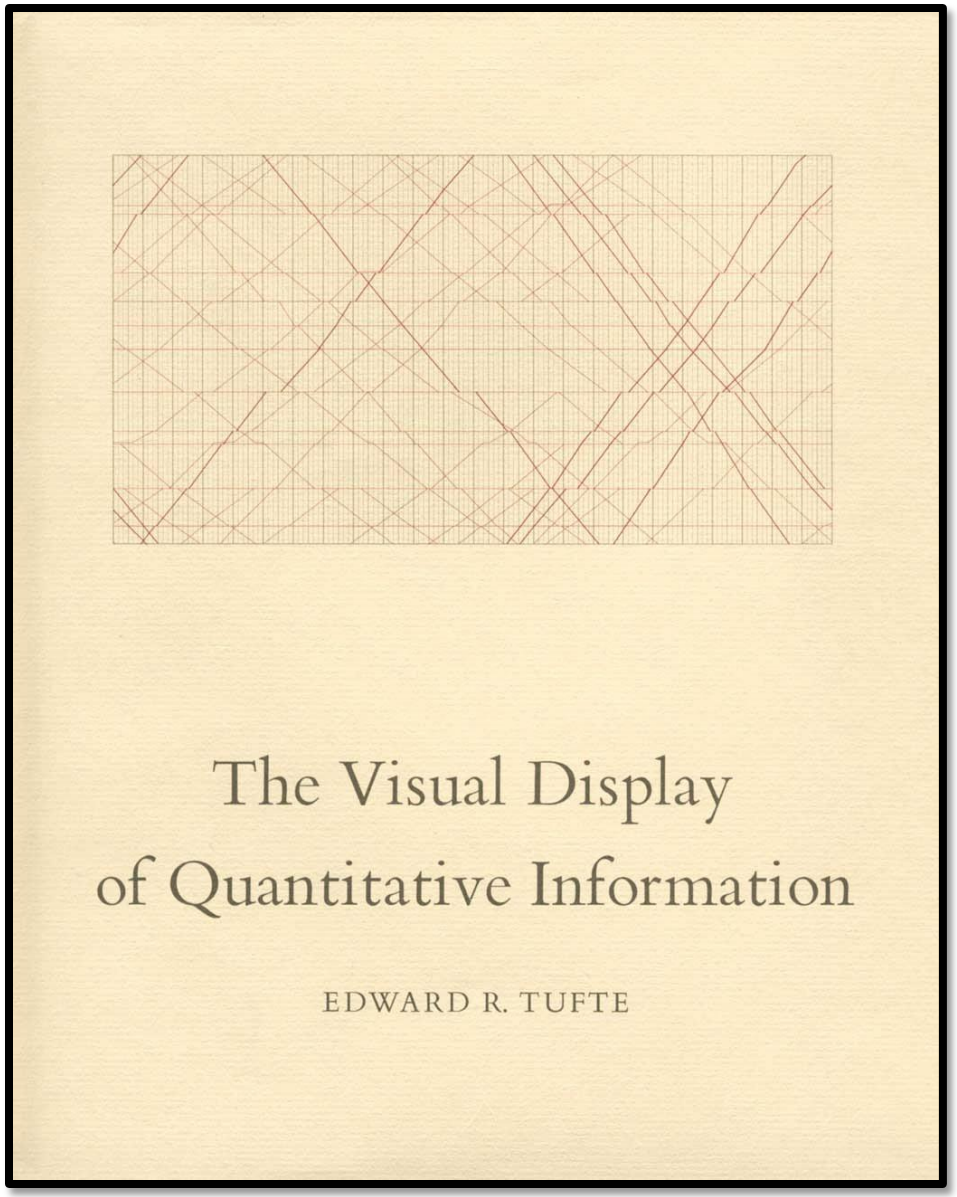
- Evaluation of visual analytic systems

Discussion

How do we measure the **effectiveness** of a visualization system?

Tufte, 1983

- "Above all else, show the data."



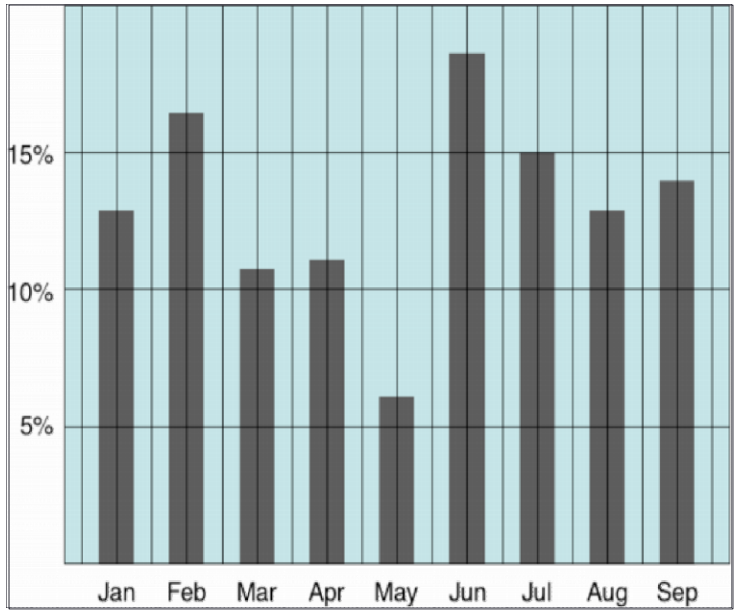
Tufte, 1983

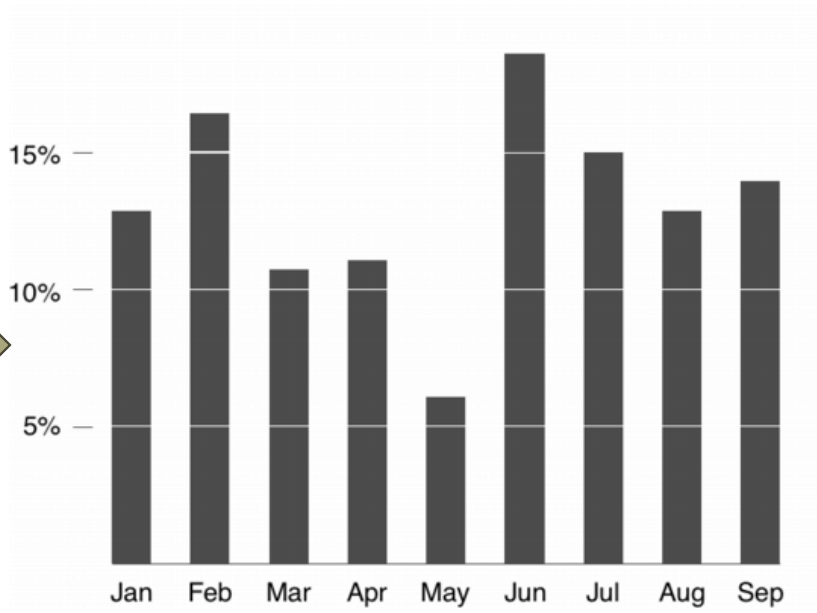
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1 - proportion of a graphic that can be erased

Tufte: maximize the data-ink ratio



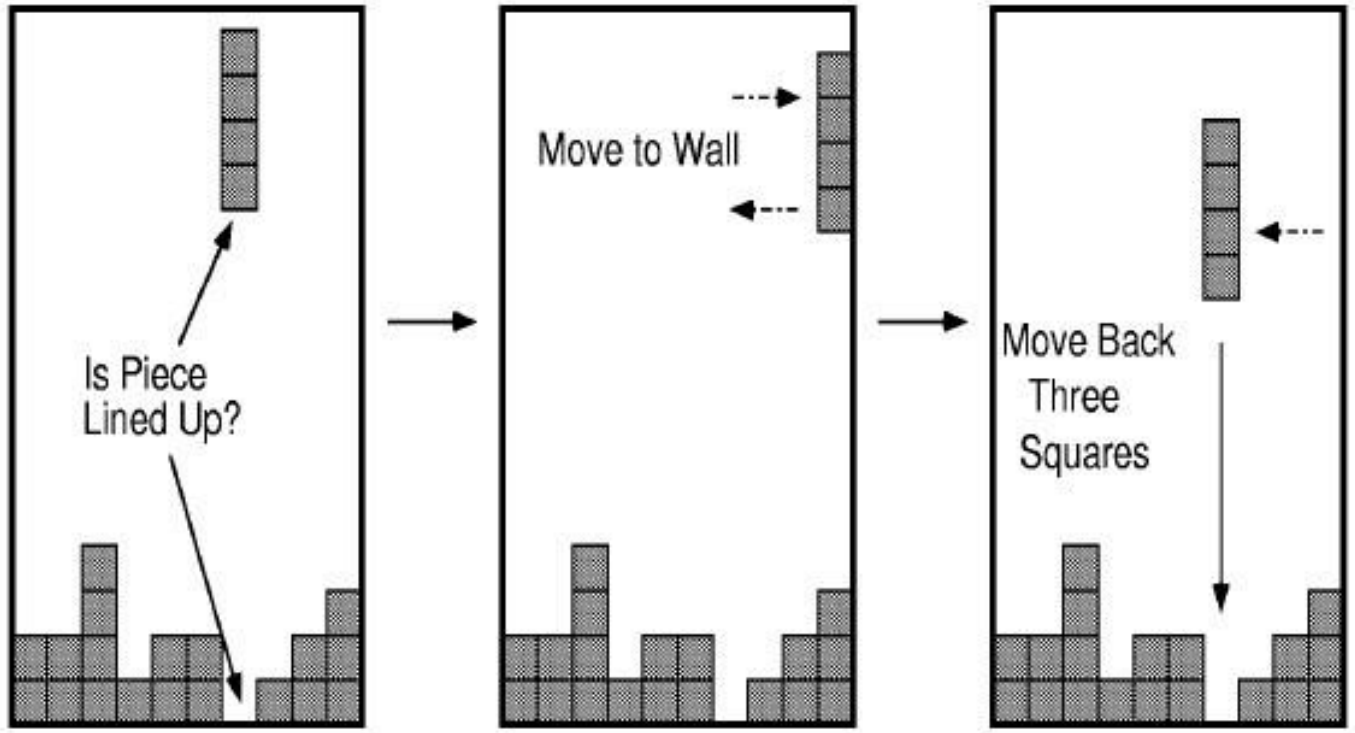


Discussion

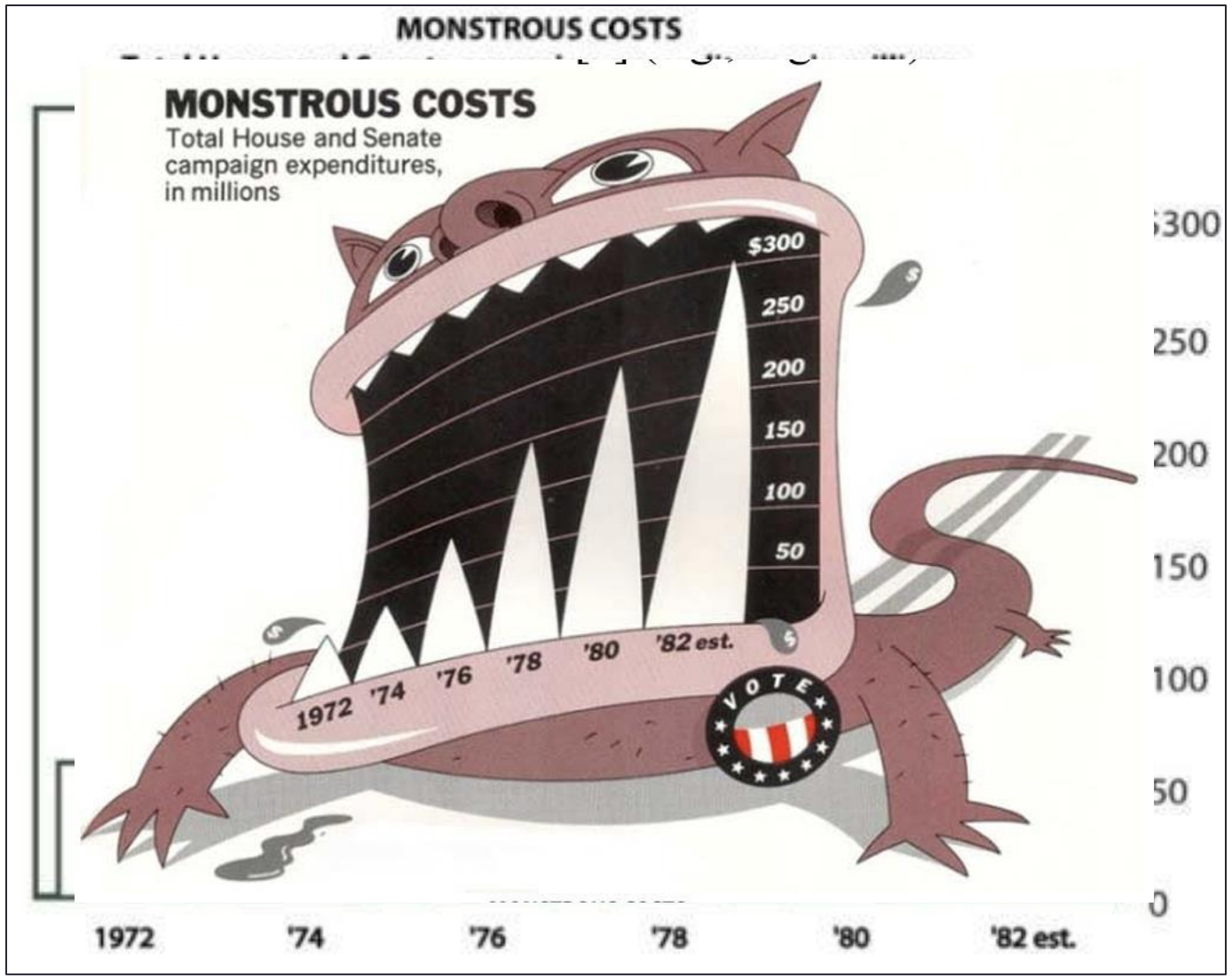
- What are the pros and cons of using data-ink ratio to evaluate visual analytic tools?

Flashback: Epistemic Action

The purpose of some actions is not the effect they have on the environment but **the effect they have on the humans**.

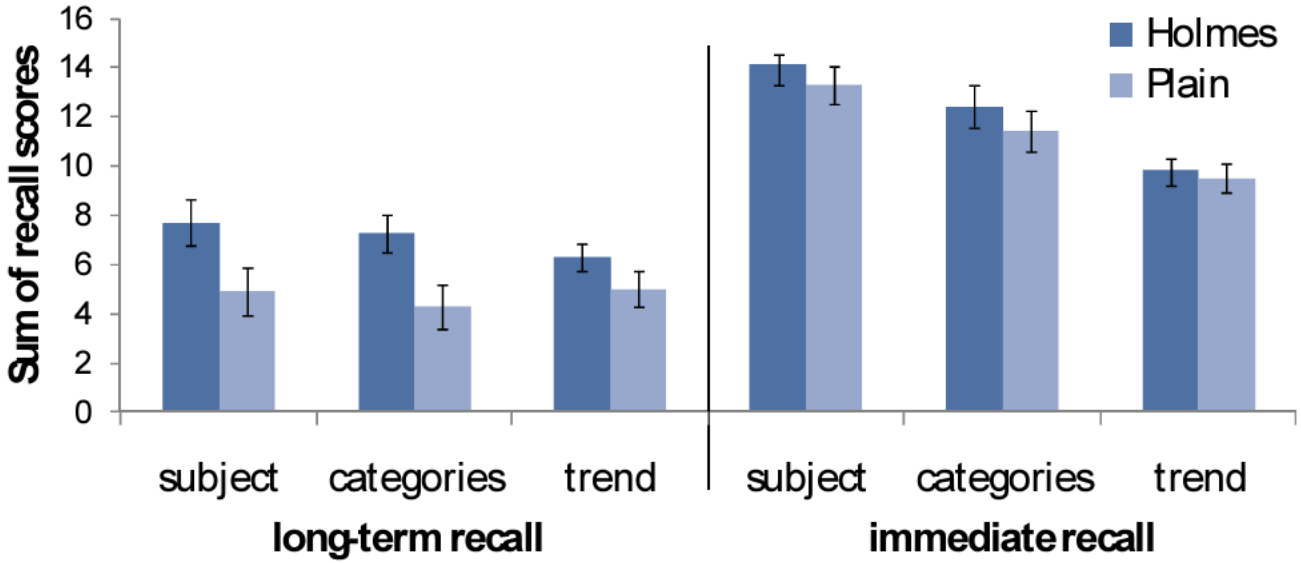


A caveat: "chart junk" and recall



Bateman et al. "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts", CHI 2010

A caveat: "chart junk" and recall



Bateman et al. "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts", CHI 2010

A caveat: "chart junk" and preference

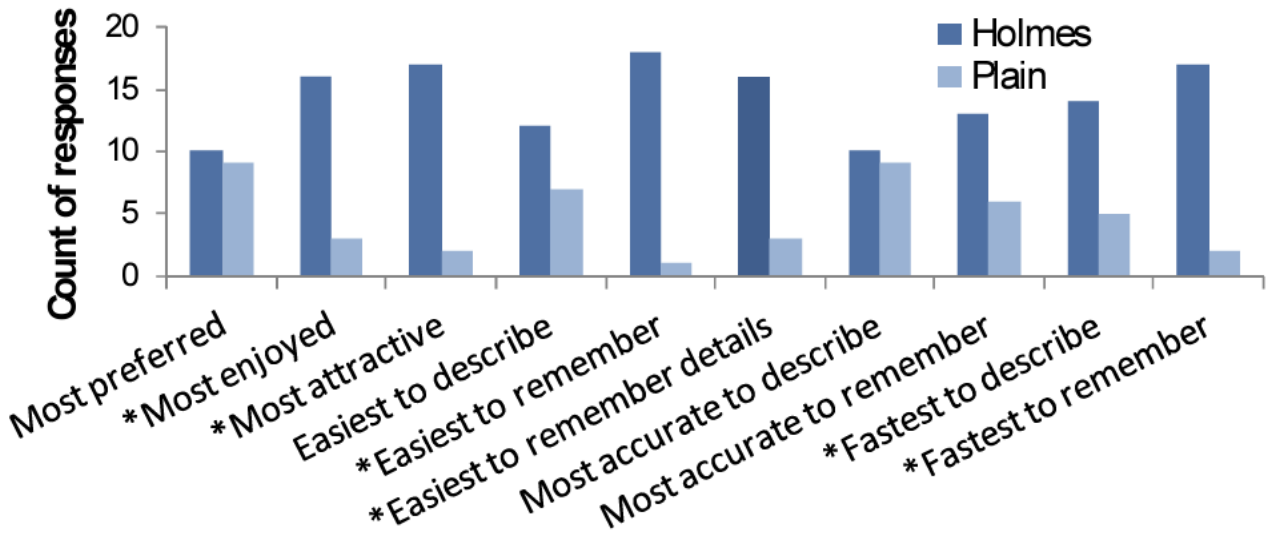
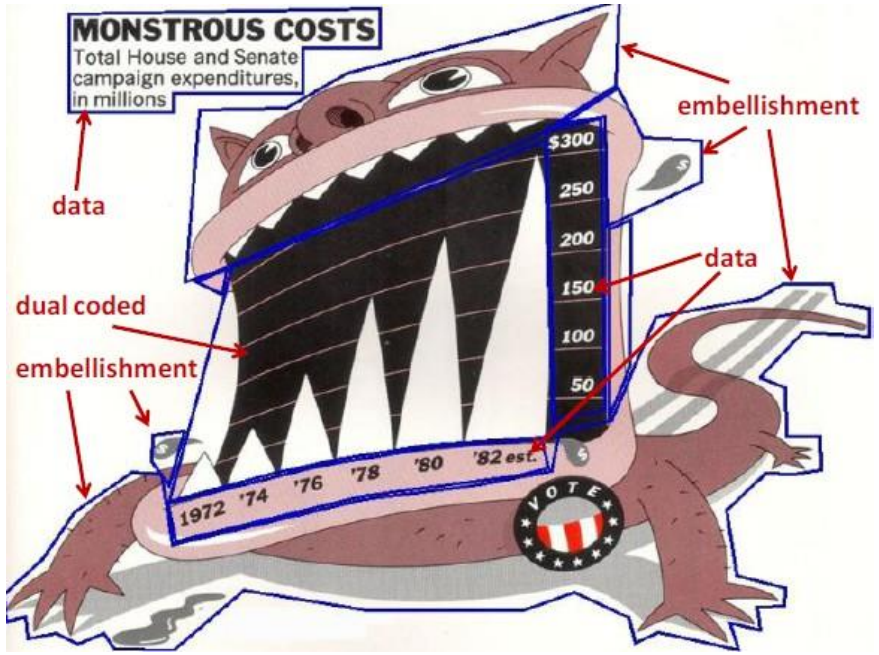


Figure 8. Count of user responses: *indicates significant difference between chart types from chi-squared test at α=0.05

Bateman et al. "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts", CHI 2010

Chart junk and eye gaze



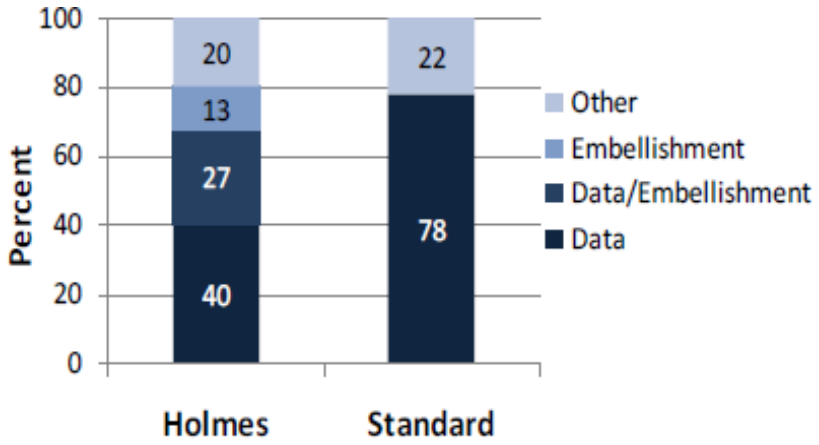
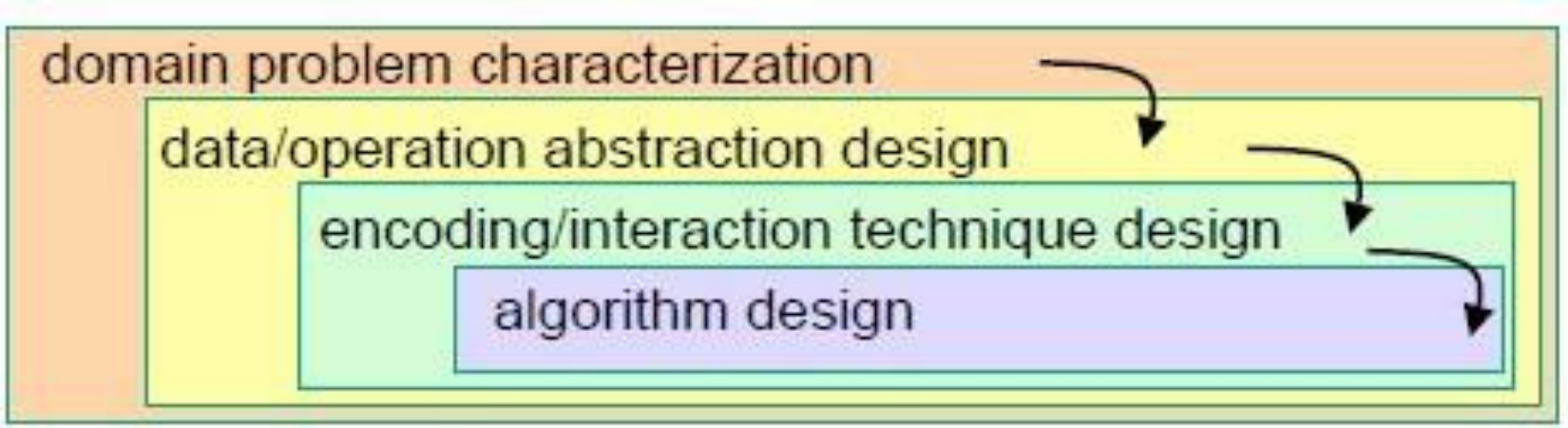


Figure 9. Percentage of on-screen time spent looking at different chart elements for Holmes and Plain charts.

Discussion

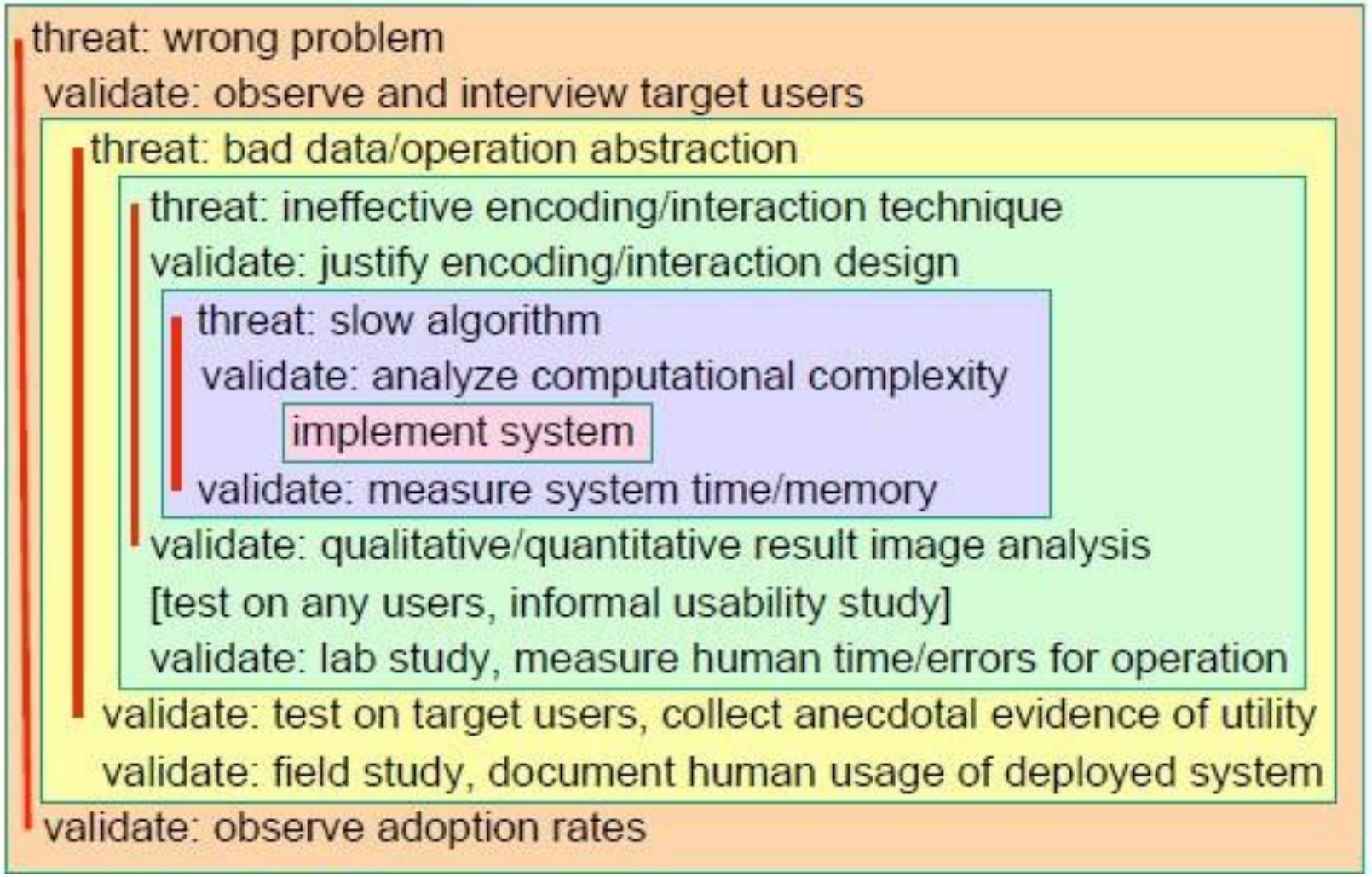
- Know any **compelling** examples of visual embellishment?
- **Tragic** ones?
- What's the right balance between Tufte and ChartJunk?

Nested Model of VIS Design (Munzner, 2009)



Munzner, Tamara. "A nested model for visualization design and validation." Visualization and Computer Graphics, IEEE Transactions on 15.6 (2009): 921-928.

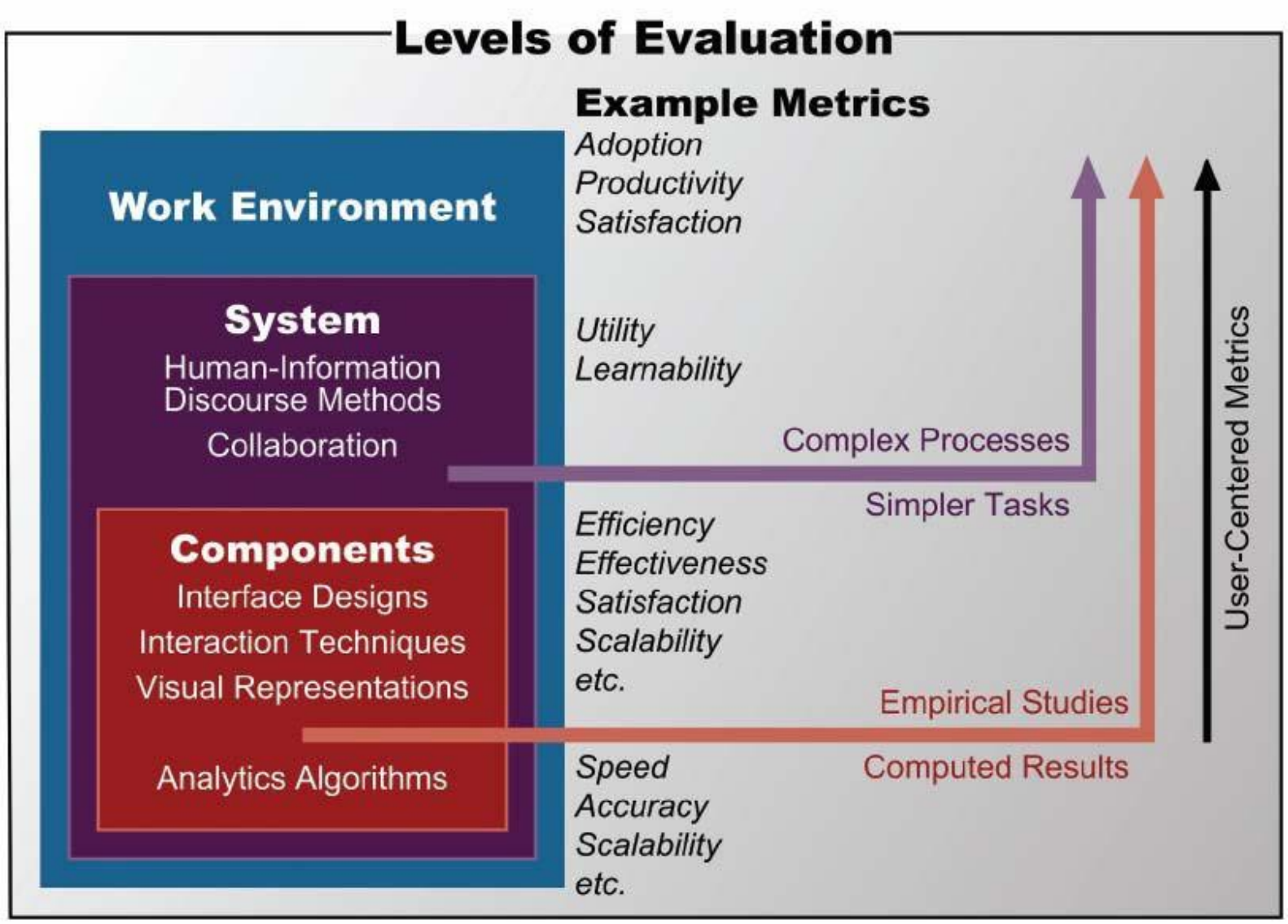
Threats


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

Evaluation "threats"

- **Mismatch**: a common problem in evaluating VIS systems
- Examples:
  - the value a new visual encoding can't be measured using a quantitative timing of the algorithm
  - mischaracterized task can't be addressed in a formal lab study

Matching **methods** and **metrics**



Insight-based evaluation (North et. al, 2005)

Measure the usefulness of a visualization by counting the **number of insights** a person generated while using it

Insight-Based Evaluation Method

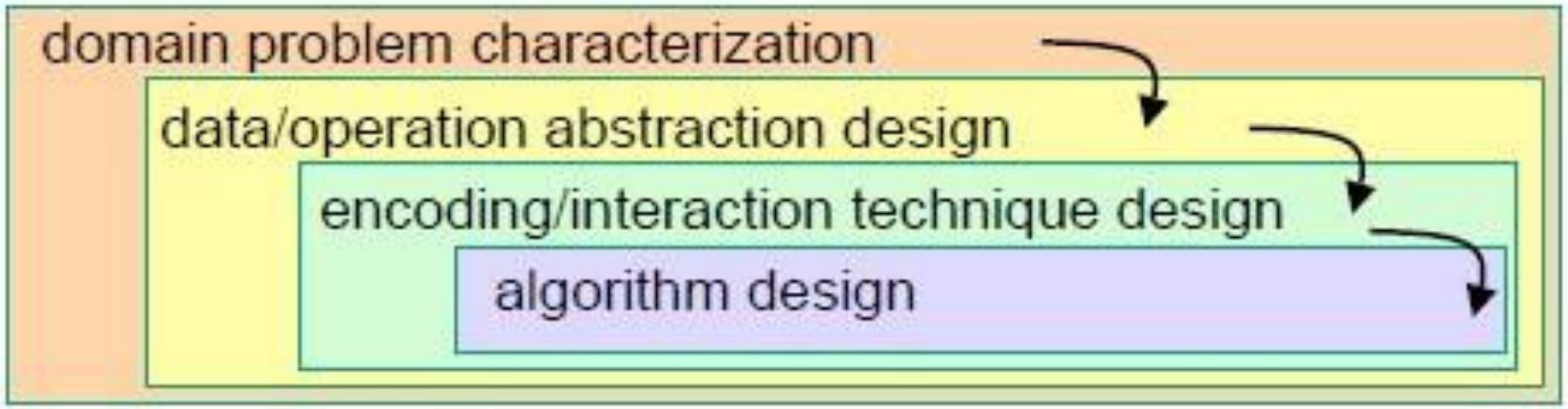
- No "benchmark tasks"
- Training on data and visualization for 15 minutes
- Participants list **questions** that they would like to pursue
- Asked to examine the data for as long as necessary until **no new insights** can be gained
- During analysis, participants are **asked to comment** on their observations, inferences, and conclusions

Evaluating the Results

- Tally up the number of insights:
  Insights: distinct observations about the data
  Baseline: all insights generated by all participants
- Various quantitative statistics on insight generation (time spent, time to first insight, etc.)

Discussion

What does insight-based evaluation address?



Problem: defining "insight"

North's definition:

> "[Insight is] an individual observation about the data by the participant, a **unit of discovery**. It is straightforward to recognize insight occurrences in a think-aloud protocol as any data observation that the user mentions is considered an insight."

Example 1

"Our tool allows the biologists to interactively visualize and explore the whole set of trees, providing **insight** into the overall distribution and possible conflicting hypothesis"

Insight = knowledge about the overall distribution

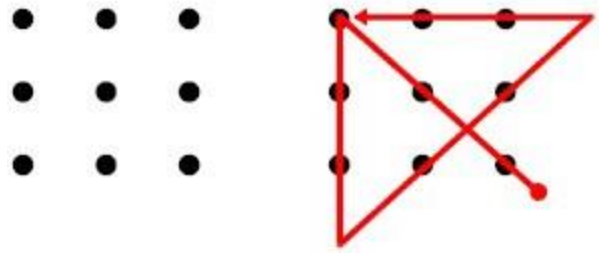
Example 2

"The analyst determined the answers to these questions, but also came up with further **insights** that she shared with people from other administrative units. She used the discovered information to advise other administrators of **certain previously unknown relationships in their data**"

Insight = information about previously unknown relationships

Cognitive science definition

- Something measurable in the frontal and temporal lobes (superior temporal gyrus).
- Spontaneous insight vs. model-building insight



boot	boot camp
summer	summer camp
ground	campground

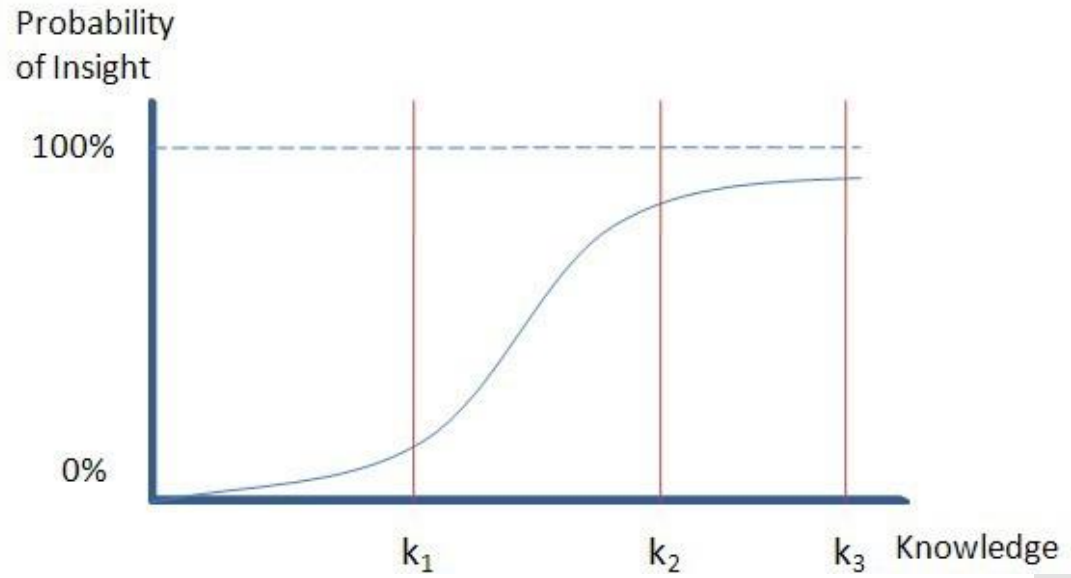
Disambiguating "Insight"

- Knowledge-building insight:
  Discovering insight, gaining insight, and providing insight
  Insight as a substance, that accumulates over time and could be measured/quantified

- Spontaneous insight:
  Experiencing insight, having an insight, or a moment of insight
  Insight as a discrete event, that occurs at a specific moment in time and could be observed

Discussion

- Can we measure knowledge-building insight?
- Can we measure spontaneous insight?
- Are they related?



MILCs – Shneiderman and Plaisant (2006)

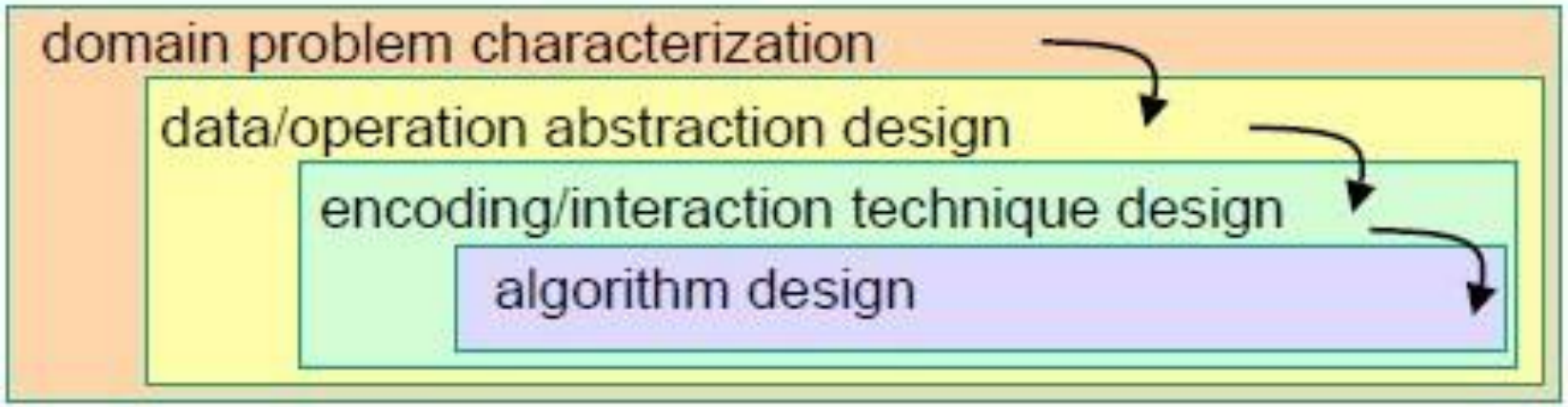
- Multi-dimensional In-depth Long-term Case studies
- Hypothesis: the efficacy of tools can be assessed by documenting:
  Usage (observations, interviews, surveys, logging, etc.)
  How successful the users are in achieving their professional goals

Definition

- Multi-dimensional: using observations, interviews, surveys, and loggers
- In-Depth: intense engagement of the researchers with the expert users to the point of becoming a partner or assistant
- Long-term: longitudinal studies that begin with training in use of a specific tool through proficient usage that leads to strategy changes for the expert users.
- Case studies: detailed reporting about a small number of individuals working on their own problems, in their own environment

Discussion

What do MILCs address?



Motivation

- MILCs have been embraced by a small community of researchers interested in studying creativity support
- Challenges:
  Cannot control for the users
  Cannot control for the tasks
  Toy problems in laboratories are not indicative of real-world problems and environments

Execution issues with MILCs

- Duration is always a problem
- Number of participants has to be small
- Formalities are difficult
  - Understand organization policies and work culture
  - Gain access and permission to observe or interview
  - Observe users in their workplace, and collect subjective and objective quantitative and qualitative data.
  - Compile data of all types in all dimensions
  - Interpret the results
  - Isolate factors
  - Need to repeat the process

Learning-based evaluation (Chang, 2010)

- Working assumption: "the goal of visualization is to gain insight and knowledge"
- Big idea: maybe we should evaluate a visualization based on whether or not the user actually gains insight or knowledge after using a visualization

Much like learning in education…

- How would an instructor choose between two textbooks for a course?
- We could:
  - Ask the students which book they prefer
  - Issue: they might like a book because its cover is pretty
  - Ask colleagues what book they prefer
  - Issue: different students in different environments
  - Ask the students to find some information in the book and measure how quickly they can perform the task
  - Issue: this only demonstrates how well the book is organized

Metaphor for visualization evaluation

- In a best case scenario, we would:
  - Ask half of the class to use book one to learn a subject
  - Ask the other half to use another book to learn the same subject
- Then we give the two groups the same test, and whichever scores higher "wins"

Traditional LBE

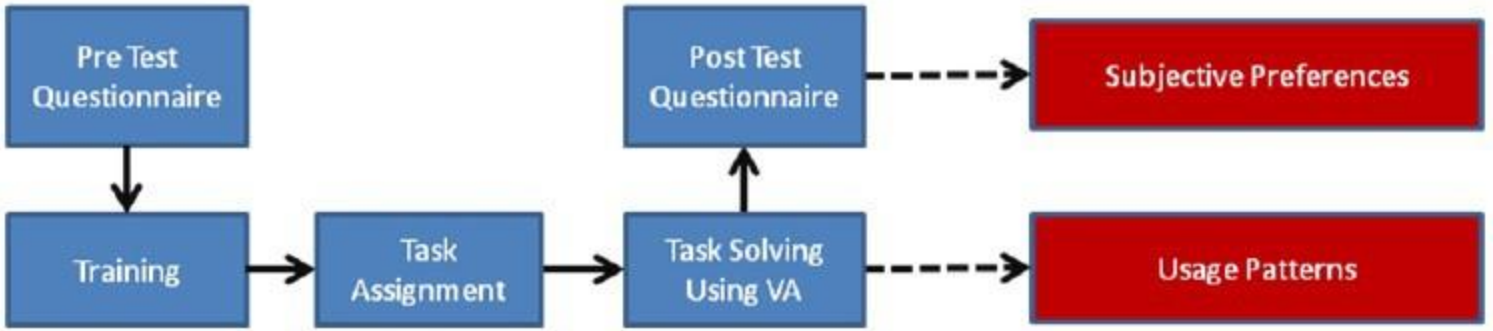


Figure 1. A pipeline for typical visualization evaluations

Discussion

Potential problems with this method?

Single-system LBE

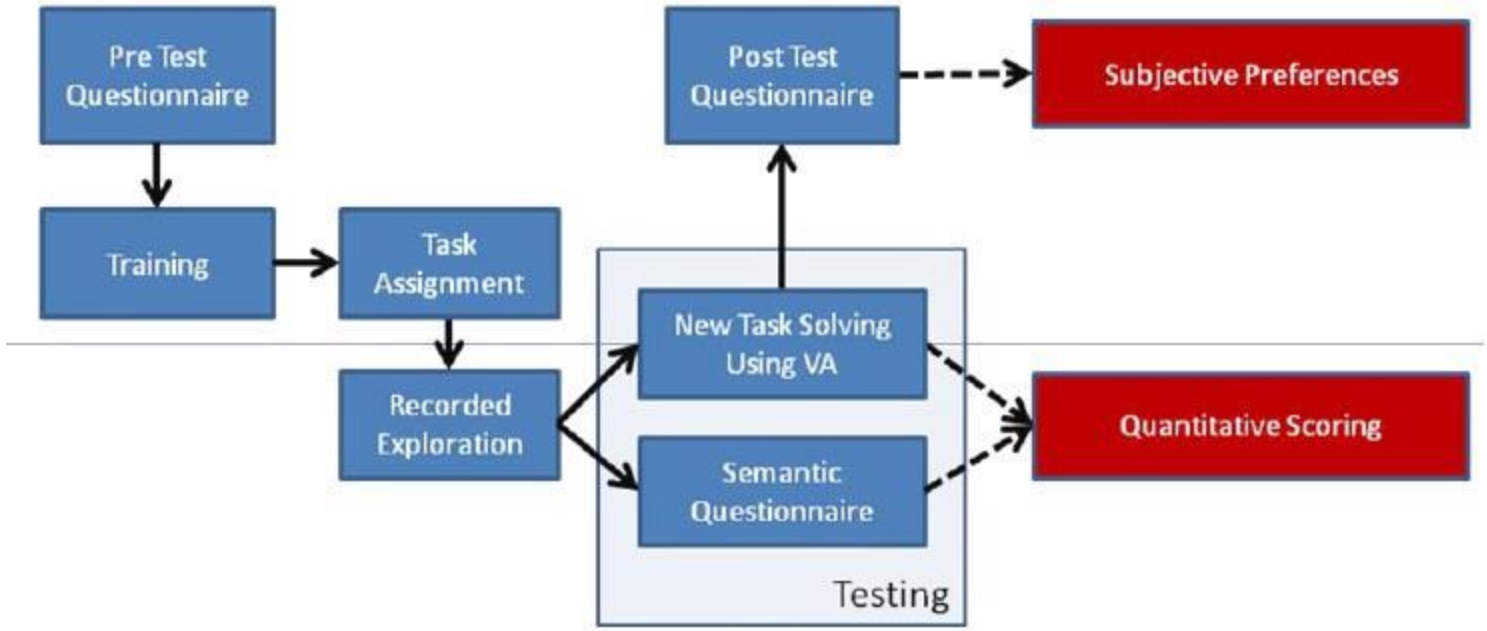


Figure 2. A pipeline for knowledge-based visualization evaluations

Your Turn!

- Play with the Visual Analytics tool datavoyager: https://vega.github.io/voyager/
- Design two different types of evaluations you might do on this tool
- Be prepared to share your ideas and why you chose them

Discussion

How should we evaluate your final projects?