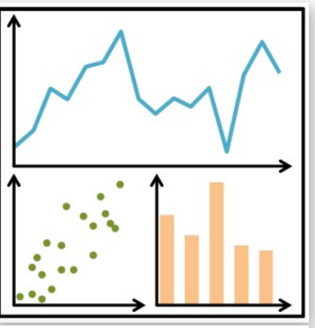
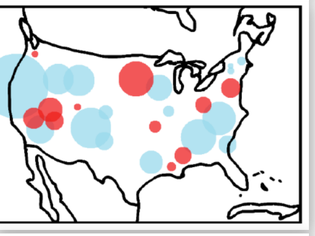
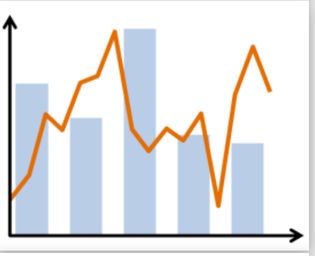
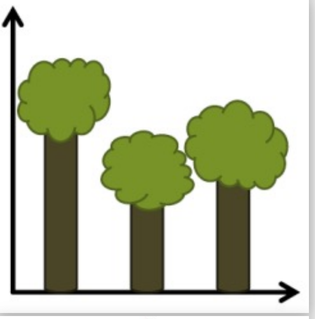
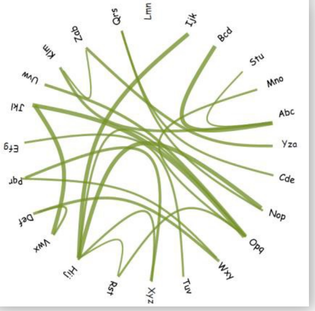






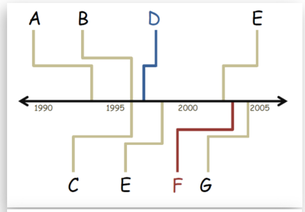
Lecture 13: Validation and Evaluation

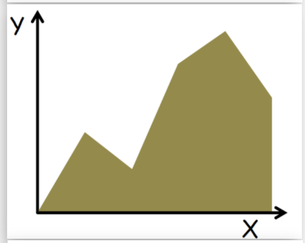
DS 4200

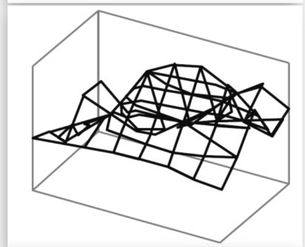
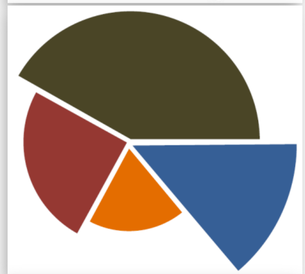
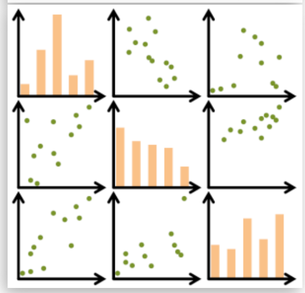
FALL 2022

Prof. Ab Mosca (they/them)

NORTHEASTERN UNIVERSITY





Slides and inspiration from Cody Dunne, Michelle Borkin, Remco Chang, Dylan Cashman, Krzysztof Gajos, Hanspeter Pfister, Miriah Meyer, Jonathan Schwabish, and David Sprague

Last Class

We:

- Reviewed Filtering and Aggregation
- Reviewed Focus+Context Designs

Any Questions?

Note on Schedule

Upcoming schedule:

https://amosca01.github.io/DS4200-F22/#schedule

tl;dr

- Next pm is short and due EOD Wednesday. The following is long and due in 10 days instead of a week.
- This avoids having an assignment due on a holiday.

Validation and Evaluation

Why Evaluate and Validate?

Based off what you know about the visualization design process, why would we need to evaluate and validate our visualization designs?

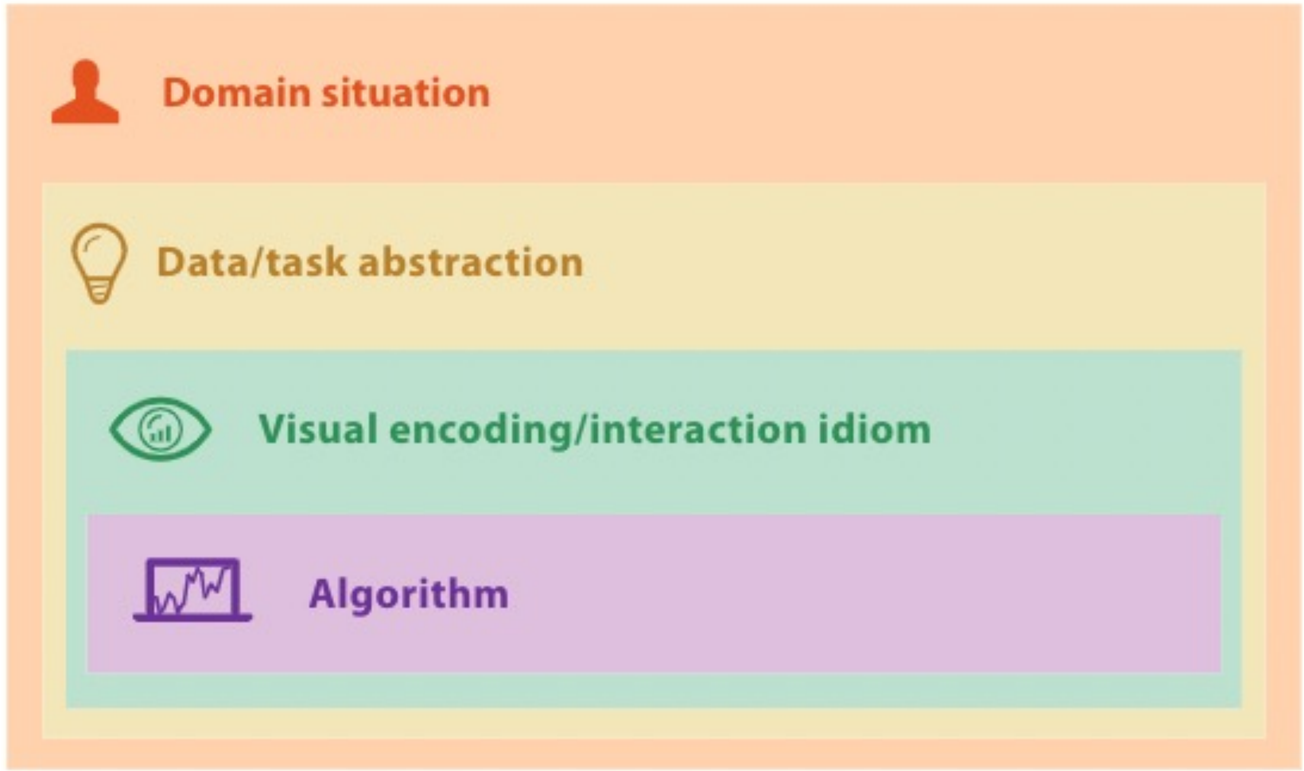
Why Evaluate and Validate?

Based off what you know about the visualization design process, why would we need to evaluate and validate our visualization designs?

- Do my target users understand how to use the visualization?
- Does this support the task(s) I intended to support?
- Are there extra features that get in the way?
- Are there missing features I need to add?
- Did I pick an effective visual encoding?
- Are the interactions responsive enough?

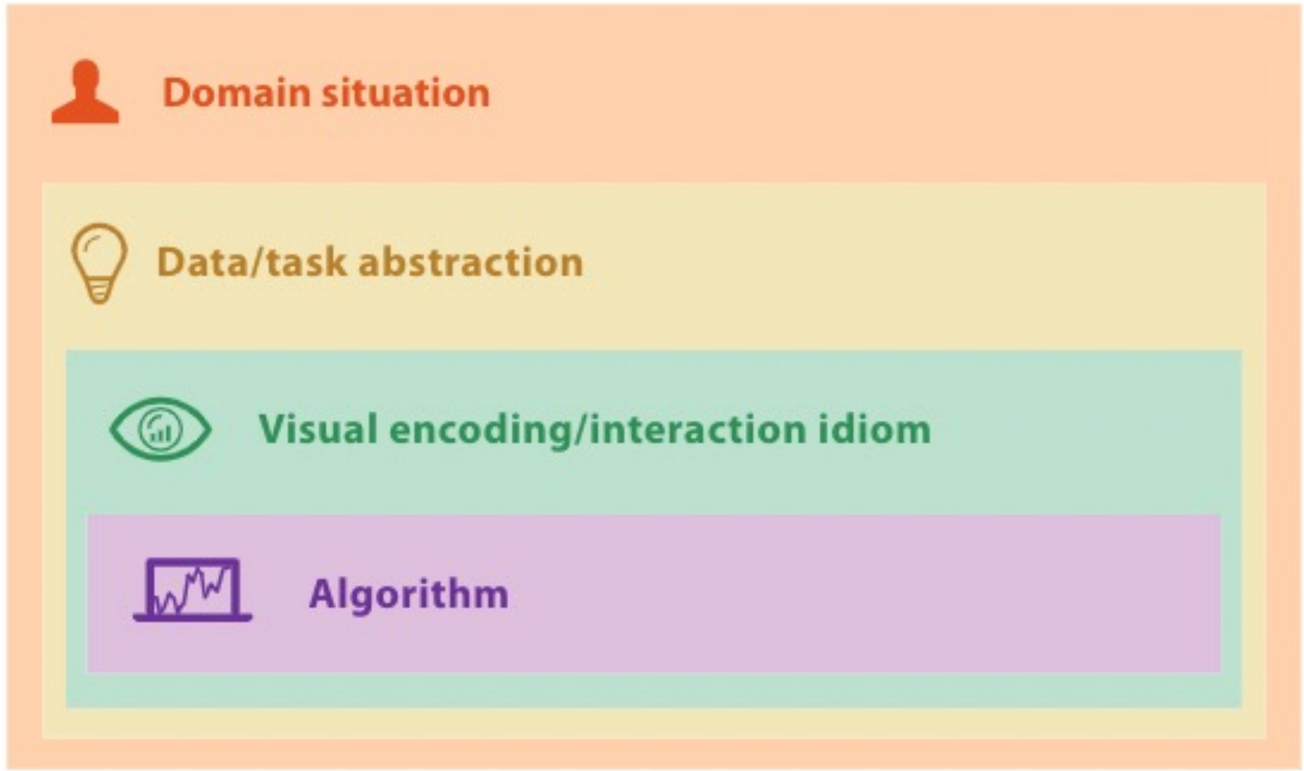
Munzner's Nested Model

- Do my target users understand how to use the visualization?
- Does this support the task(s) I intended to support?
- Are there extra features that get in the way?
- Are there missing features I need to add?
- Did I pick an effective visual encoding?
- Are the interactions responsive enough?



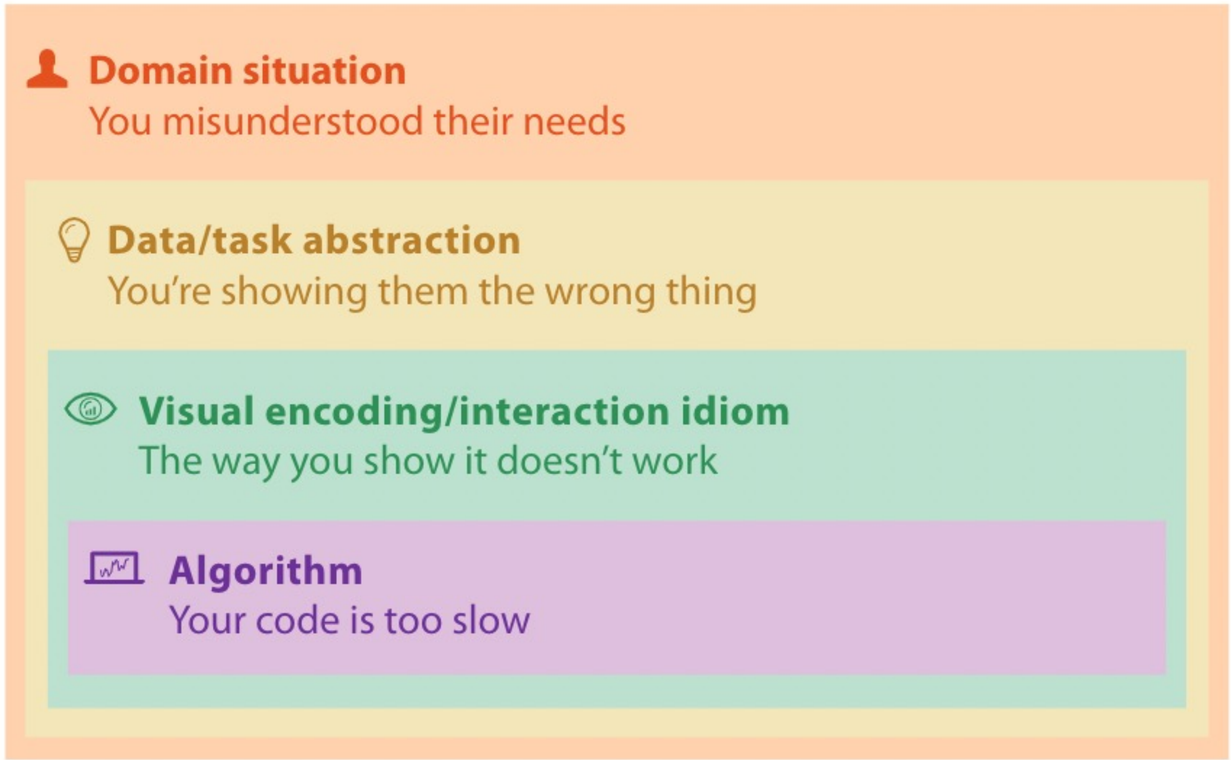
Munzner's Nested Model

→ Threats to validity exist at each stage of visualization design



Munzner's Nested Model

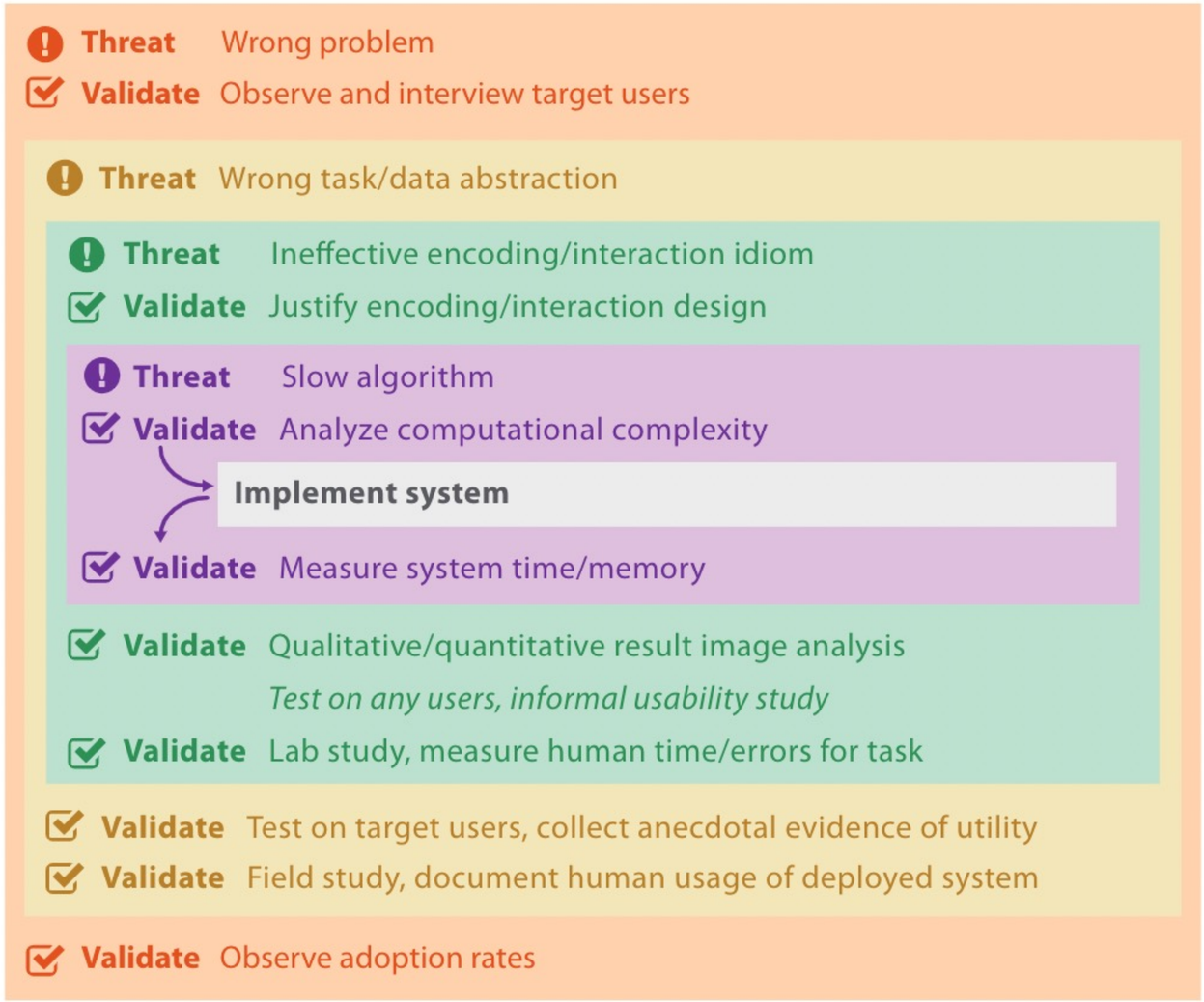
→ Threats to validity exist at each stage of visualization design



Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique

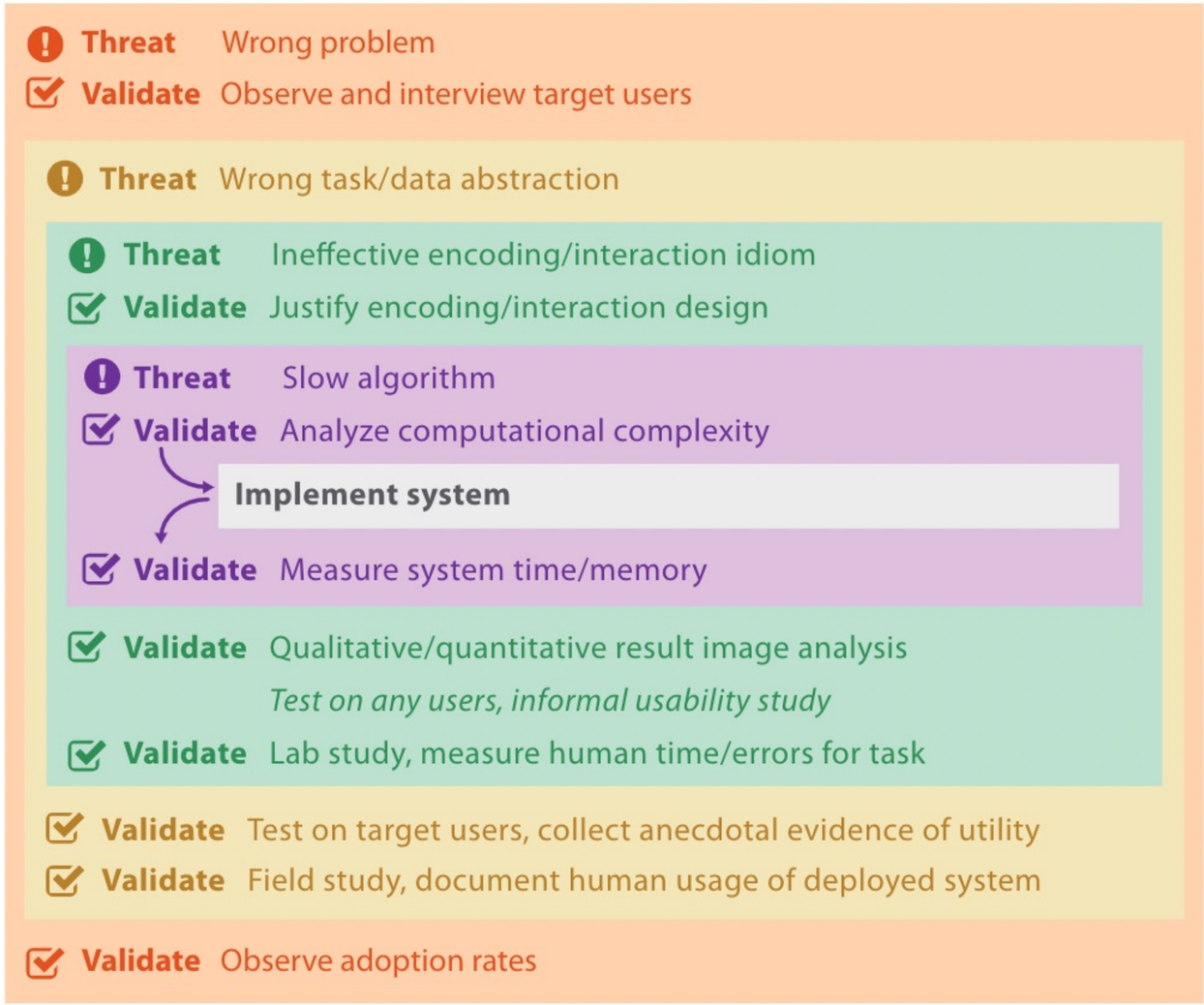


Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique

→**Domain situation**
- →Typically use qualitative evaluation techniques
- →Ex. Ethnographic field studies, semi-structured interviews

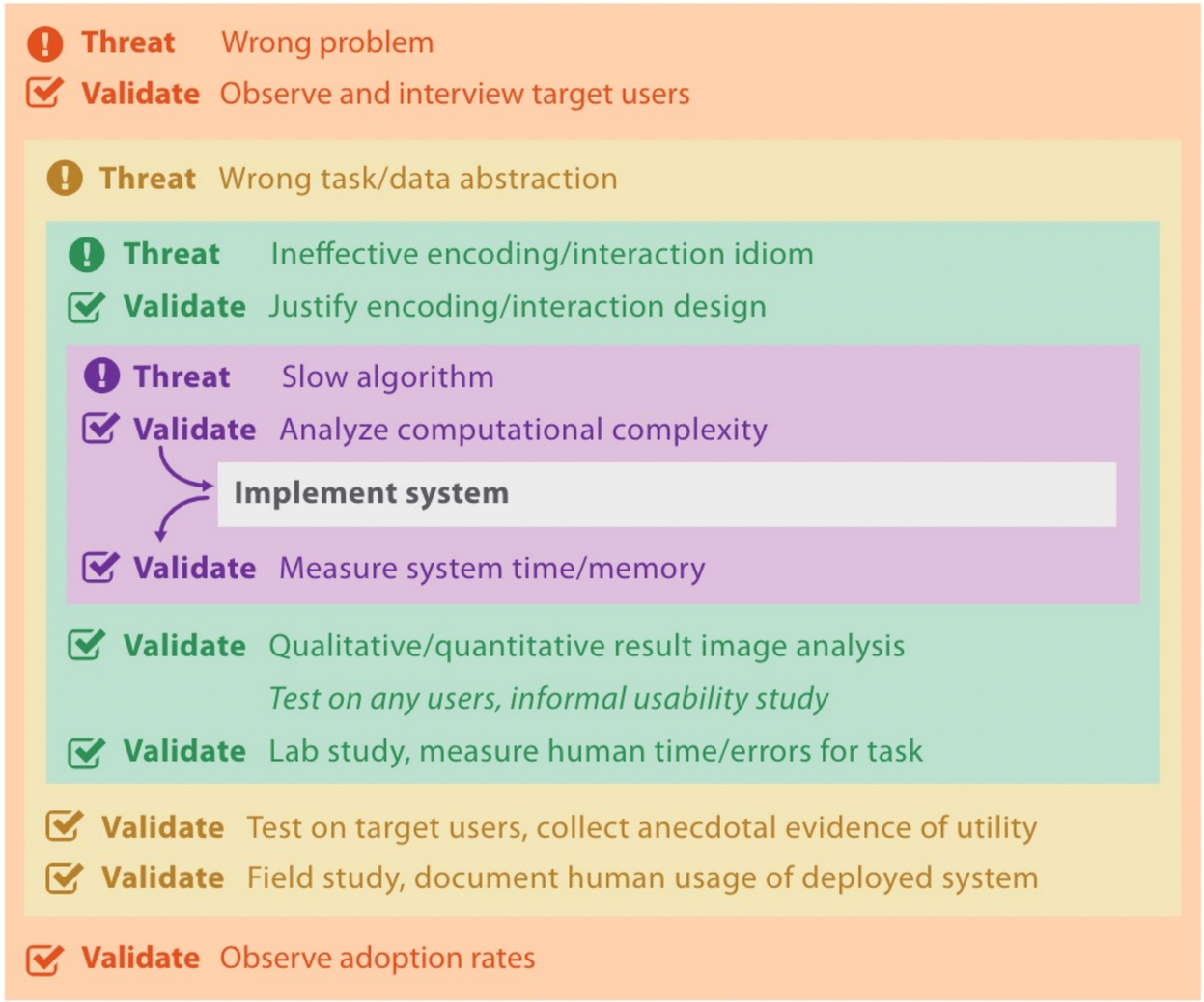


Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique

→**Data / task abstraction**
- →Typically use qualitative evaluation techniques
- →Ex. Short or long term observational studies

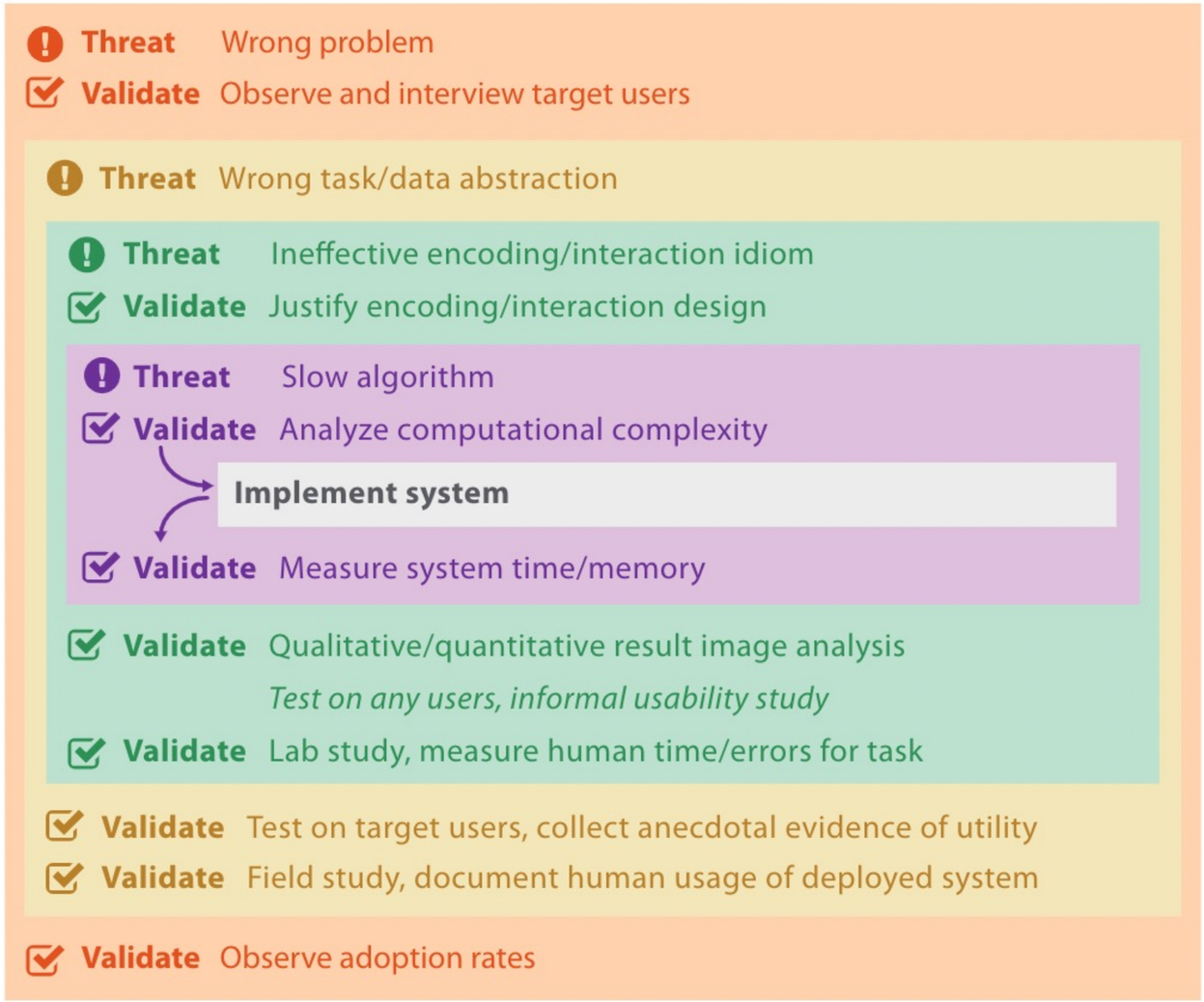


Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique

→**Visual encoding / interaction**
- →Can use qualitative or quantitative evaluation techniques
- →Ex. Informal usability study, formal lab study

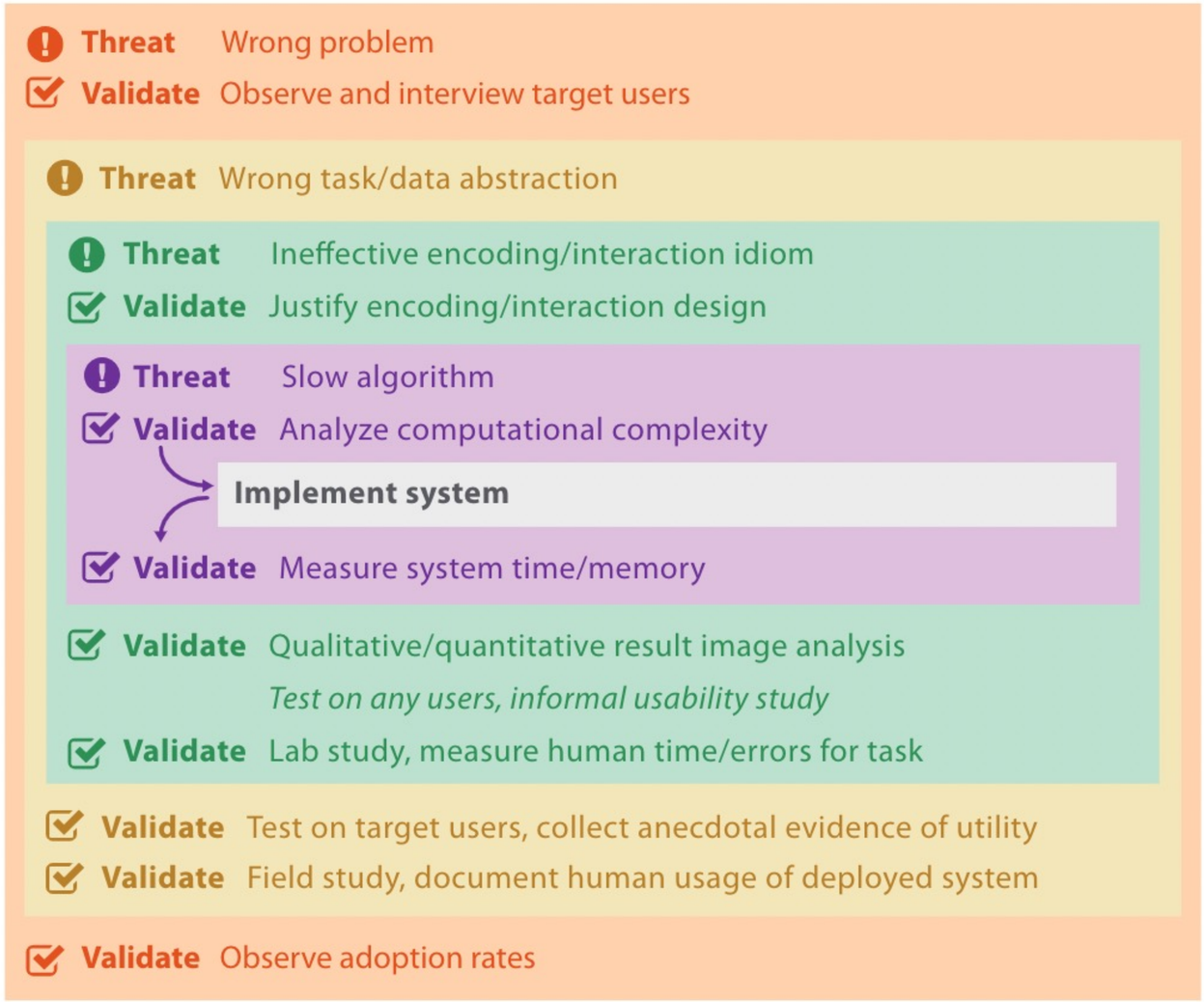


Munzner's Nested Model

→Threats to validity exist at each stage of visualization design

→Validate every step using appropriate technique

→**Algorithm**
- →Quantitative evaluation
- →Analyze computational complexity



Evaluation Methods

Theoretical / Simulation Based

→ Inspection or Principled Rationale

→ Theoretical Analysis

→ Benchmarks

Human Subject Experiments

→ Controlled Experiment

→ Observational Study

→ Field Deployment or Case Studies

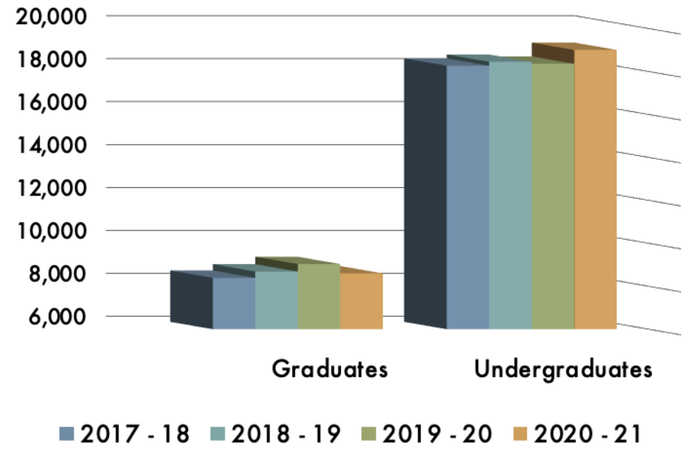
Evaluation Methods

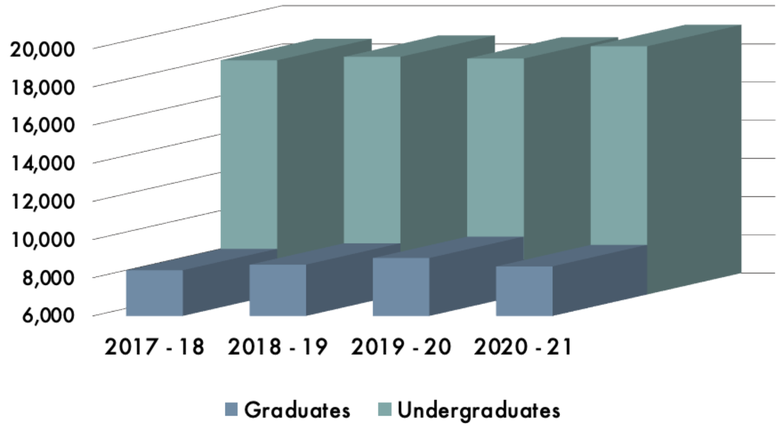
Theoretical / Simulation Based

→ **Inspection or Principled Rationale**

→ Apply design heuristics, perceptual principles

https://www.vislies.org/2021/gallery/





Evaluation Methods

Theoretical / Simulation Based

→ **Theoretical Analysis**

→ Algorithm time and space complexity

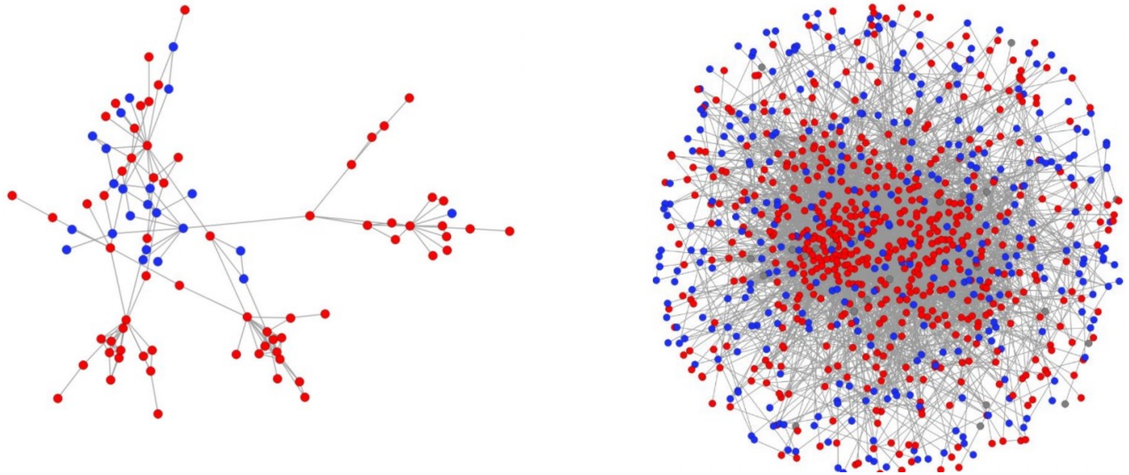
$$O(n) = n^2$$

Evaluation Methods

Theoretical / Simulation Based

→ **Benchmarks**

- → Scalability to larger data sets
- → Performance (e.g., interactive frame rates)



https://www.cs.ubc.ca/~tmm/courses/547-19/slides/javier-persisthomol.pdf

Time Constant	Value (in seconds)
perceptual processing	0.1
immediate response	1
brief tasks	10

Evaluation Methods

Human Subject Experiments

→ **Controlled Experiments**

→ **Observational Studies**

→ **Field Deployment or Case Studies**

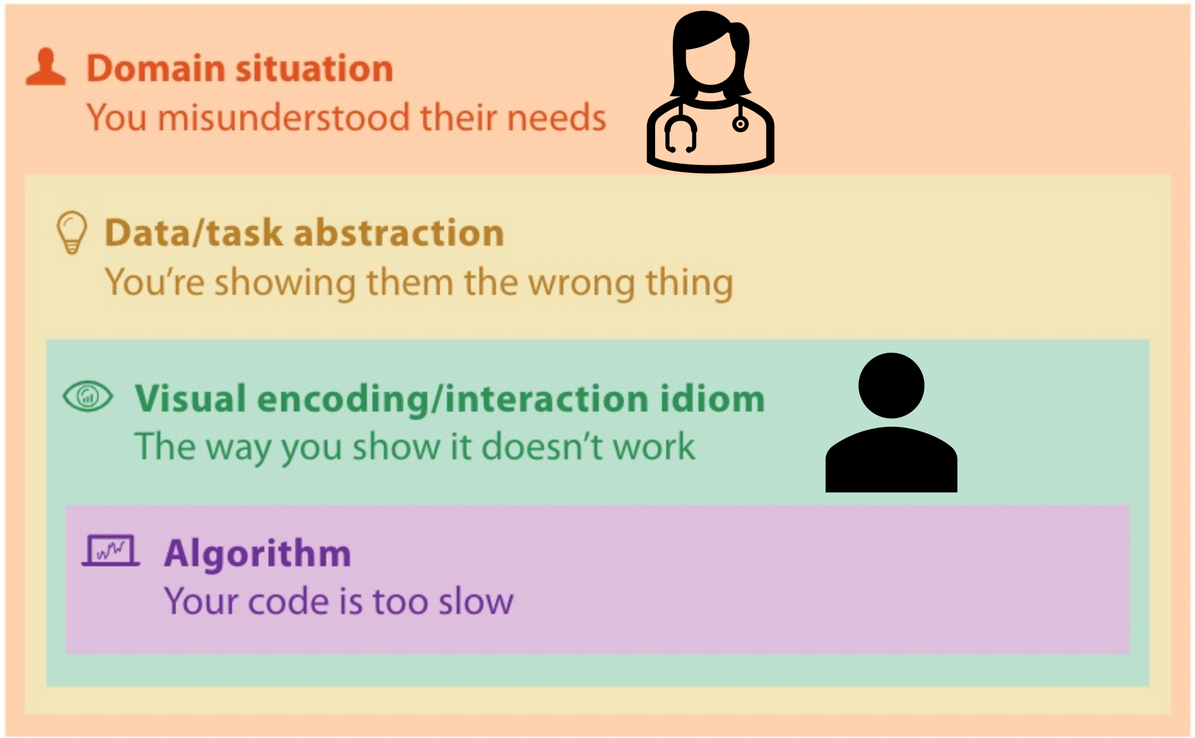
Evaluation Methods

Human Subject Experiments

→ **Controlled Experiments**

→ **Observational Studies**

→ **Field Deployment or Case Studies**



Evaluation Methods

Human Subject Experiments

→ **Controlled Experiment**
- → Specific repeated task across multiple conditions
- → Collect quantitative data (ex. time, accuracy)

A
New Vis

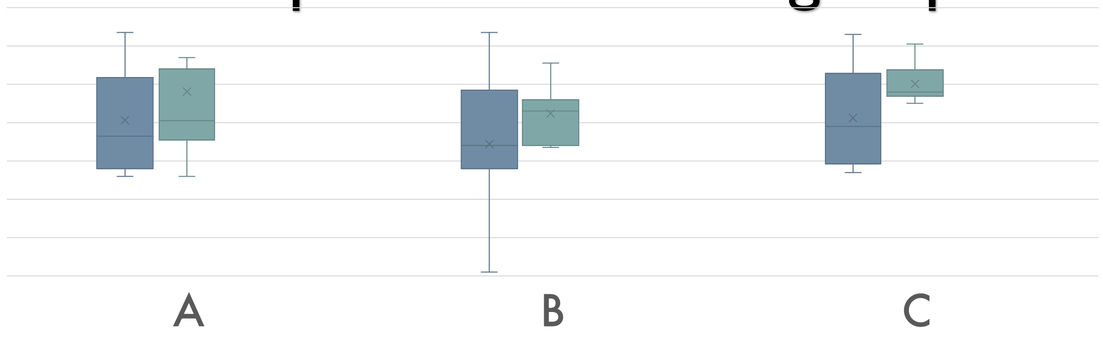
B
Comparison Vis 1

C
Comparison Vis 2

Task: Find all outliers in the data

Collect: Time, Accuracy

Compare data across groups



Evaluation Methods

Human Subject Experiments

→ **Controlled Experiment**
- → Specific repeated task across multiple conditions
- → Collect quantitative data (ex. time, accuracy)

→ There are many ways to design a controlled experiment (i.e. many experimental designs). We'll touch on some more later.

A
New Vis

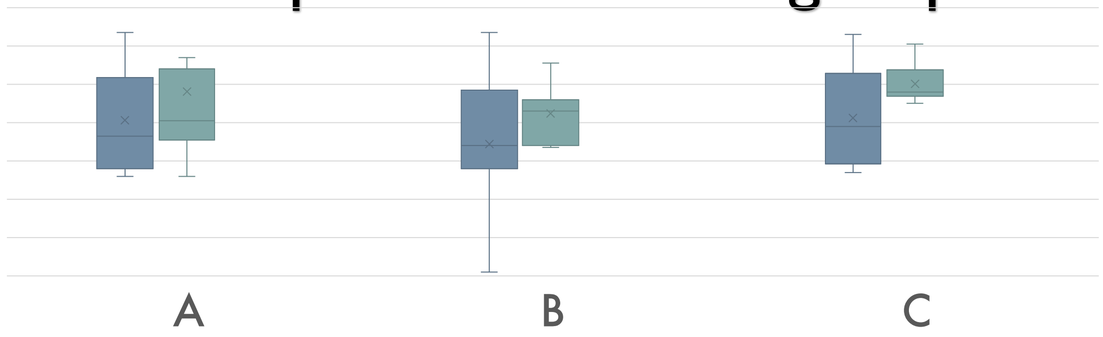
B
Comparison Vis 1

C
Comparison Vis 2

Task: Find all outliers in the data

Collect: Time, Accuracy

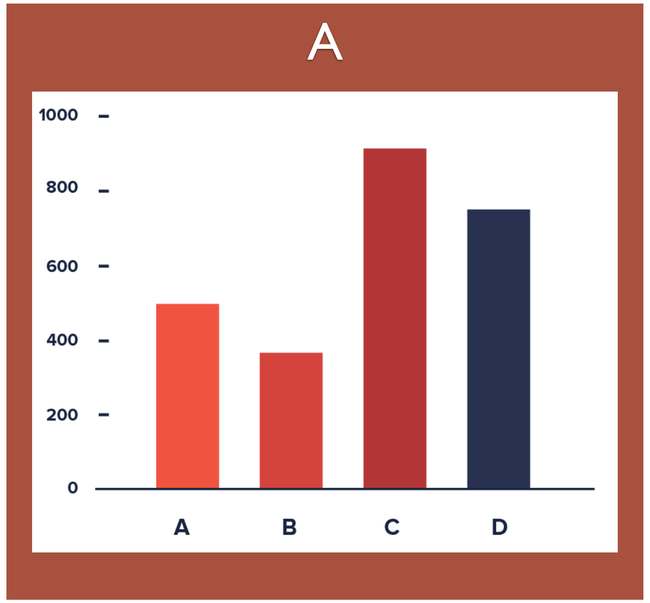
Compare data across groups

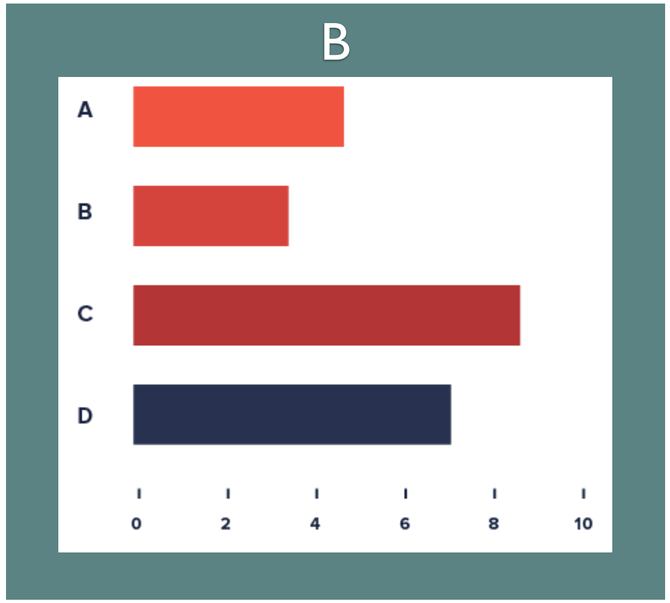


Evaluation Methods

Human Subject Experiments

→ **Controlled Experiment**
- → Specific repeated task across multiple conditions
- → Collect quantitative data (ex. time, accuracy)





Task: ?

Collect: ?

Compare data across groups

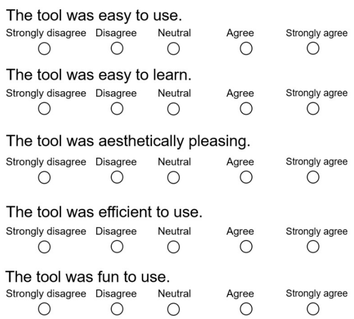
https://datavizproject.com/

Evaluation Methods

Human Subject Experiments

→ **Observational Study**

- → Specific repeated task across multiple conditions (or not)
- → Collect qualitative data (ex. user findings, think-out-loud data, satisfaction) and quantitative



https://neu.co1.qualtrics.com/jfe/form/SV_8kbi3XRZ628CoPH

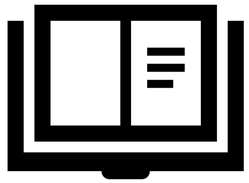
A New Vis	B Comparison Vis 1	C Comparison Vis 2

Task: Find all outliers in the data

Collect: Think out loud, Likert scales

Compare data across groups

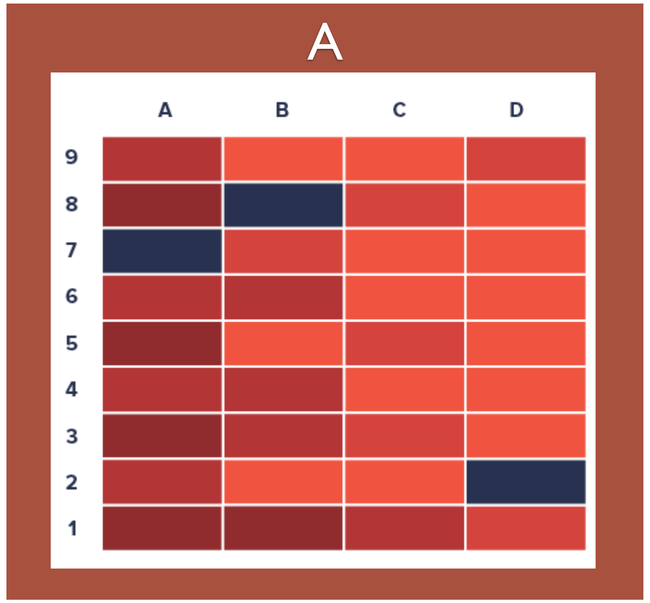
- Code data & compare themes
- Compare # insights
- Compare Likert scale evaluations on satisfaction and ease of use

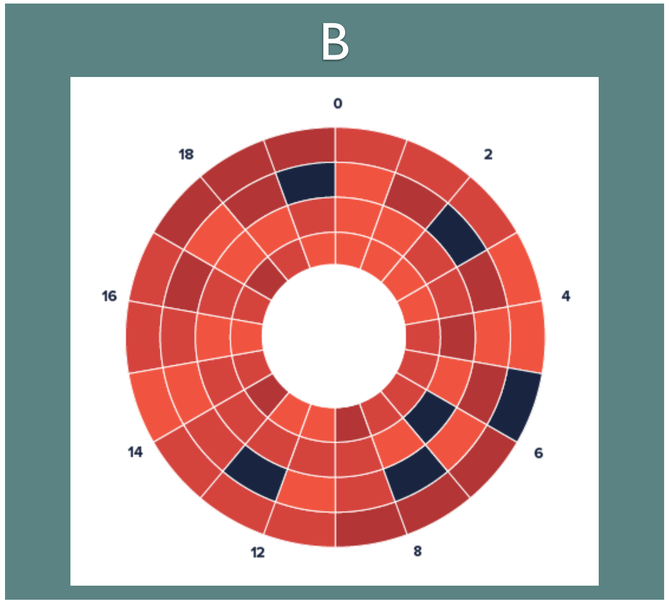


Evaluation Methods

Human Subject Experiments

→ **Observational Study**
- → Specific repeated task across multiple conditions (or not)
- → Collect qualitative data (ex. user findings, think-out-loud data, satisfaction) and quantitative





Task: ?

Collect: ?

Compare data across groups

https://datavizproject.com/

Evaluation Methods

Human Subject Experiments

→ **Observational Study**

Usability Studies

→ Observational study specifically focused on usability of a visualization

→ Range from low to high fidelity testing

Evaluation Methods

Human Subject Experiments

→ **Observational Study**

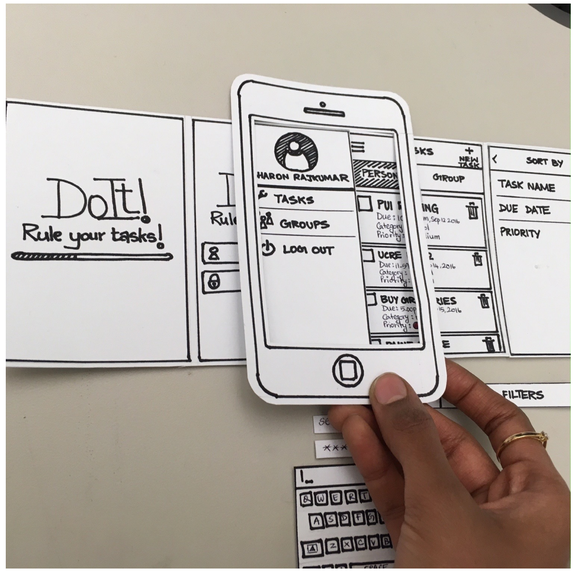
Usability Studies

→ Observational study specifically focused on usability of a visualization

→ Range from low to high fidelity testing

Low Fidelity

→ Paper prototype



Evaluation Methods

Human Subject Experiments

→ **Observational Study**

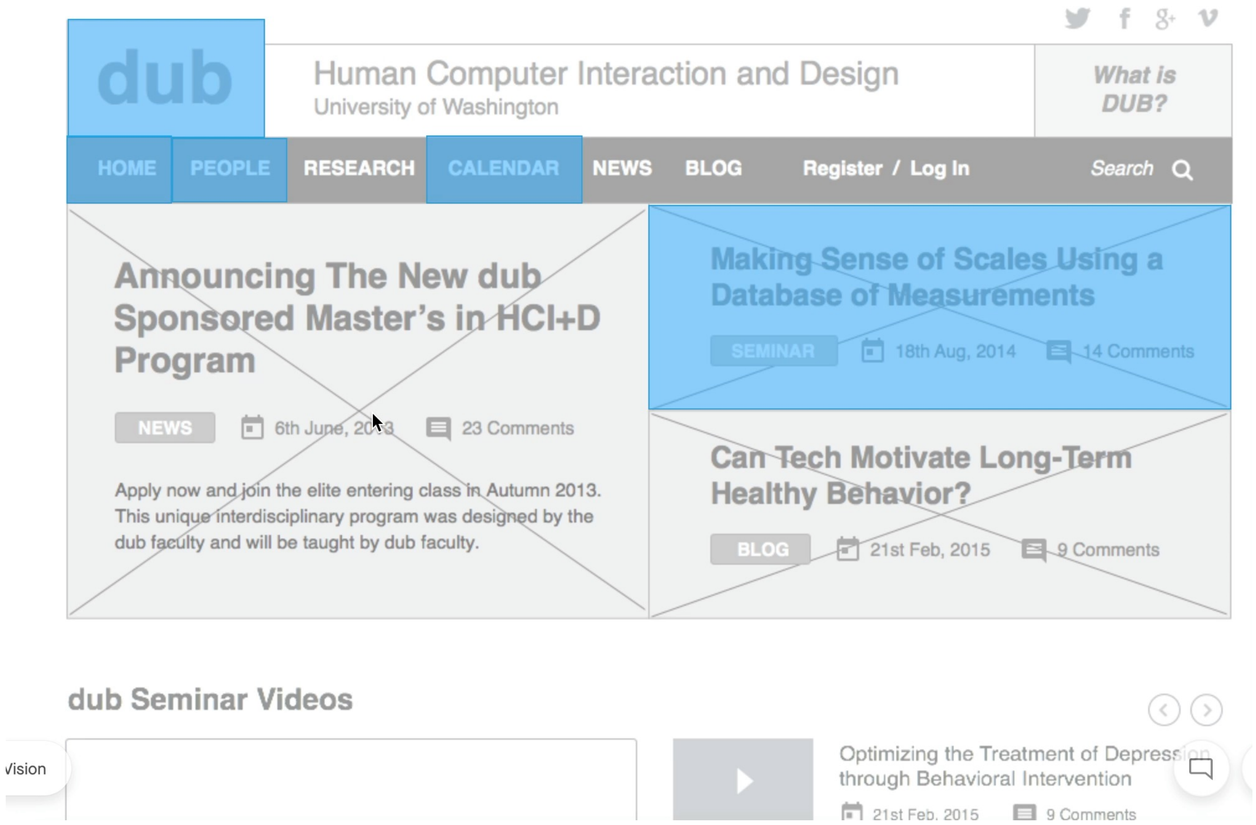
Usability Studies

→ Observational study specifically focused on usability of a visualization

→ Range from low to high fidelity testing

Medium Fidelity

→ Wireframe



Evaluation Methods

Human Subject Experiments

→ **Observational Study**

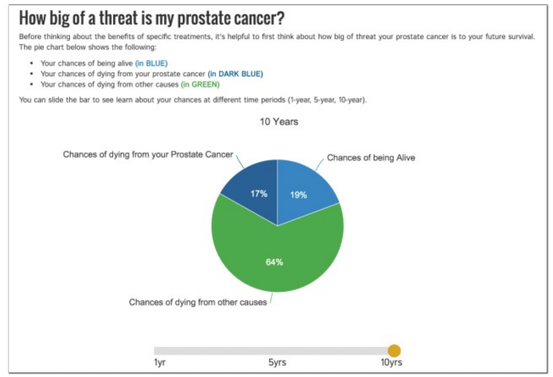
Usability Studies

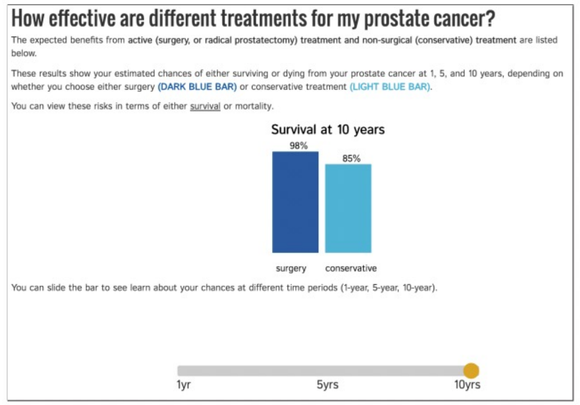
→ Observational study specifically focused on usability of a visualization

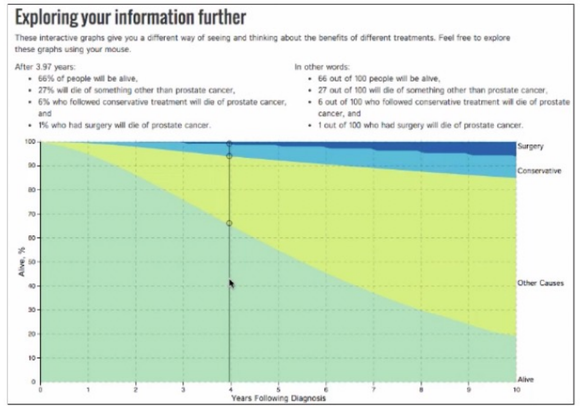
→ Range from low to high fidelity testing

High Fidelity

→ Working prototype





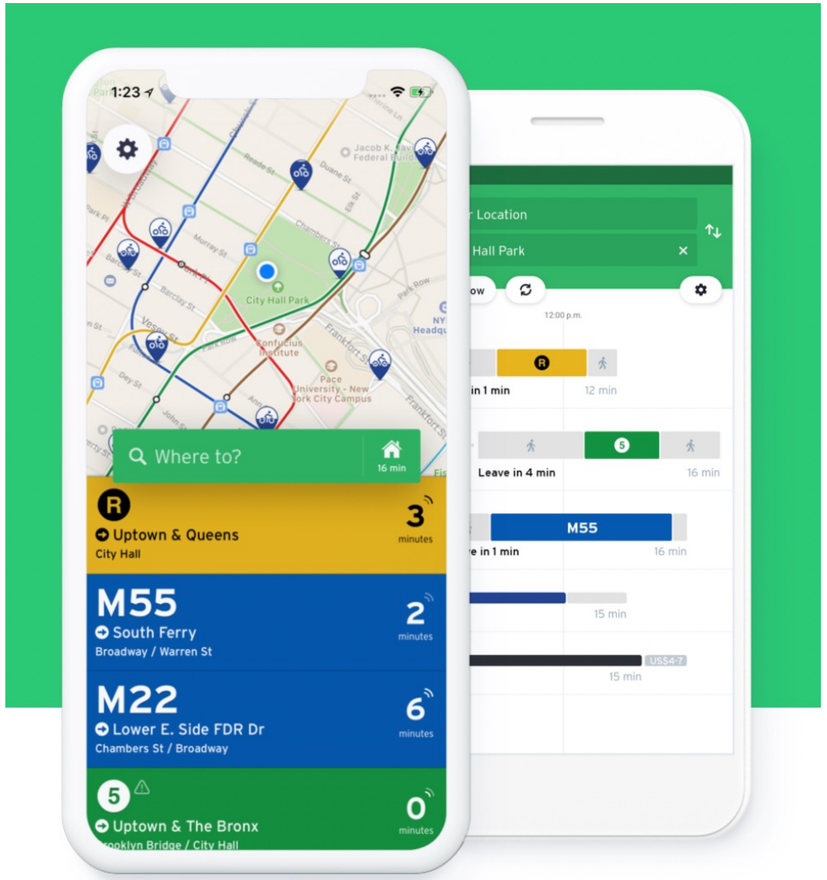


https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7539638

Evaluation Methods

Human Subject Experiments

→ **Field Deployment or Case Studies**
- → Deploy tool in realistic settings
- → Document effects on work practices
- → Observe usage in the "real world"
- → Collect qualitative or quantitative data (ex. satisfaction, number of times used in a week, etc.)



https://transitapp.com/

Evaluation Design

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

6) Analysis

→ How will you analyze collected data

Evaluation Design

1) Goal

→ Munzner's nested model

2) Method

→ Select an evaluation technique

3) Who

→ Who are your participants (and how many)

4) Tasks

→ What will the participant do

5) Metrics

→ Decide what data to collect

6) Analysis

→ How will you analyze collected data

Data Collection

Tasks and Metrics depend on goal

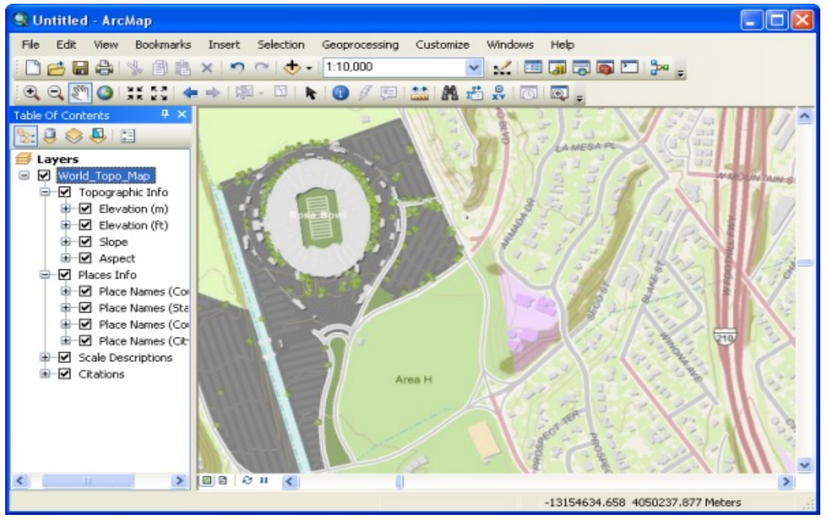
→ **Common metric choices include:**

- Speed
- Accuracy
- Quality of finding
- Satisfaction
- Recall/Memorability
- Requiring help
- Completeness and Thoroughness of Analysis
- False Positives and False Negatives
- Immersiveness
- Deviation from optimality
- Information accessed
- Use frequency
  - → How often do users use this tool?
- Effort / Ease of Use
  - → Cognitive
  - → Physical (how many clicks did it take)
  - → NASA-TLX

Data Collection

ArcGIS vs. Google Maps – Which results in better spatial understanding of an area?

- Researchers compared **speed** and **accuracy** and found no difference



- But now we know users in Google Maps tend to interact a lot more
  - As a result, they don't perform faster or better, but they have a better spatial understanding of the space

- What other data could the researchers have collected?

Data Collection

Insight Based Evaluation (a type of observational study)

- Proposed by Chris North et al. in 2005

- Method: determine the utility of a visualization system by the number of insights generated by the user

- Evaluation:
  - Tally number of insights (distinct observations about the data by a participant)
  - Various quantitative statistics collected on insight generation (time spent, time to first insight, etc.)

Data Collection

Insight Based Evaluation (a type of observational study)

- Method: determine utility based on number of insights generated
  - No specific tasks
  - Participants get training on data and visualization for 15 minutes
  - Participants list some analysis questions that they would like to pursue
  - Participants asked to examine the data for as long as necessary until no new insights can be gained
  - Participants asked to comment on their observations, inferences, and conclusions

- Best for evaluating how well a visualization supports exploratory data analysis

Data Collection

What metrics would you collect?

- **Scenario 1**: We recently released a new interface for our visual analytics system. Can you help determine if our customers like the new design?

- **Scenario 2**: My company currently uses Tableau to compare sales of different items over time, but we've recently heard about Looker and we're wondering if it might be a better choice. Can you help determine if we should we switch?

Let’s take a break! Stretch, go for a walk, be social ☺

Be back here in 10 mins.

Summary

Today we:

- Reviewed evaluation and validation

ic-13 is DUE today.
pm-04 is OUT today and DUE Wed.