Cody Dunne
Northeastern University

Validation & Evaluation

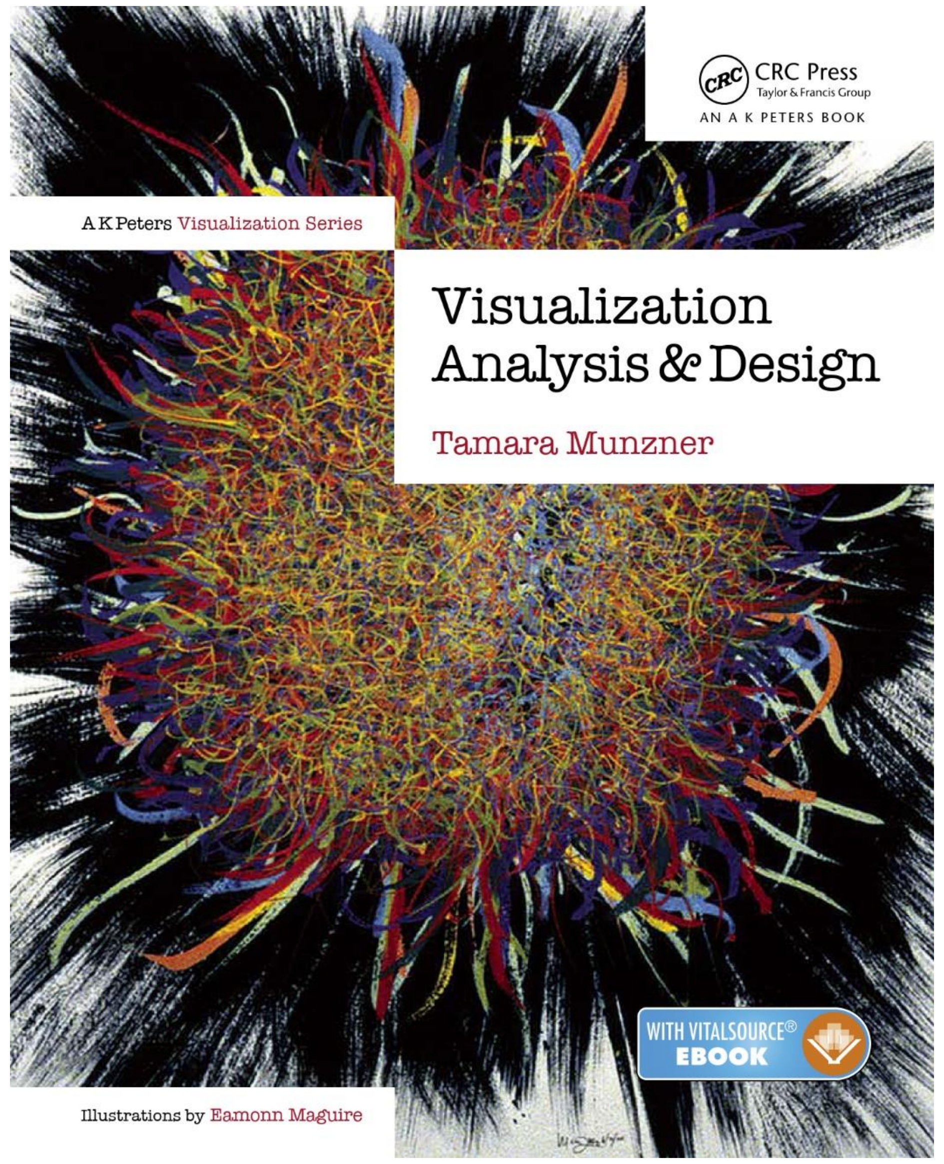
Reading Quiz

Q4—Validation & Evaluation

Friday: Marvin Zelen Symposium

~5 min

The Nested Model for Visualization Validation



"Nested Model"

Domain situation
Observe target users using existing tools

Data/task abstraction

Visual encoding/interaction idiom
Justify design with respect to alternatives

Algorithm
Measure system time/memory
Analyze computational complexity

Analyze results qualitatively

Measure human time with lab experiment (*lab study*)

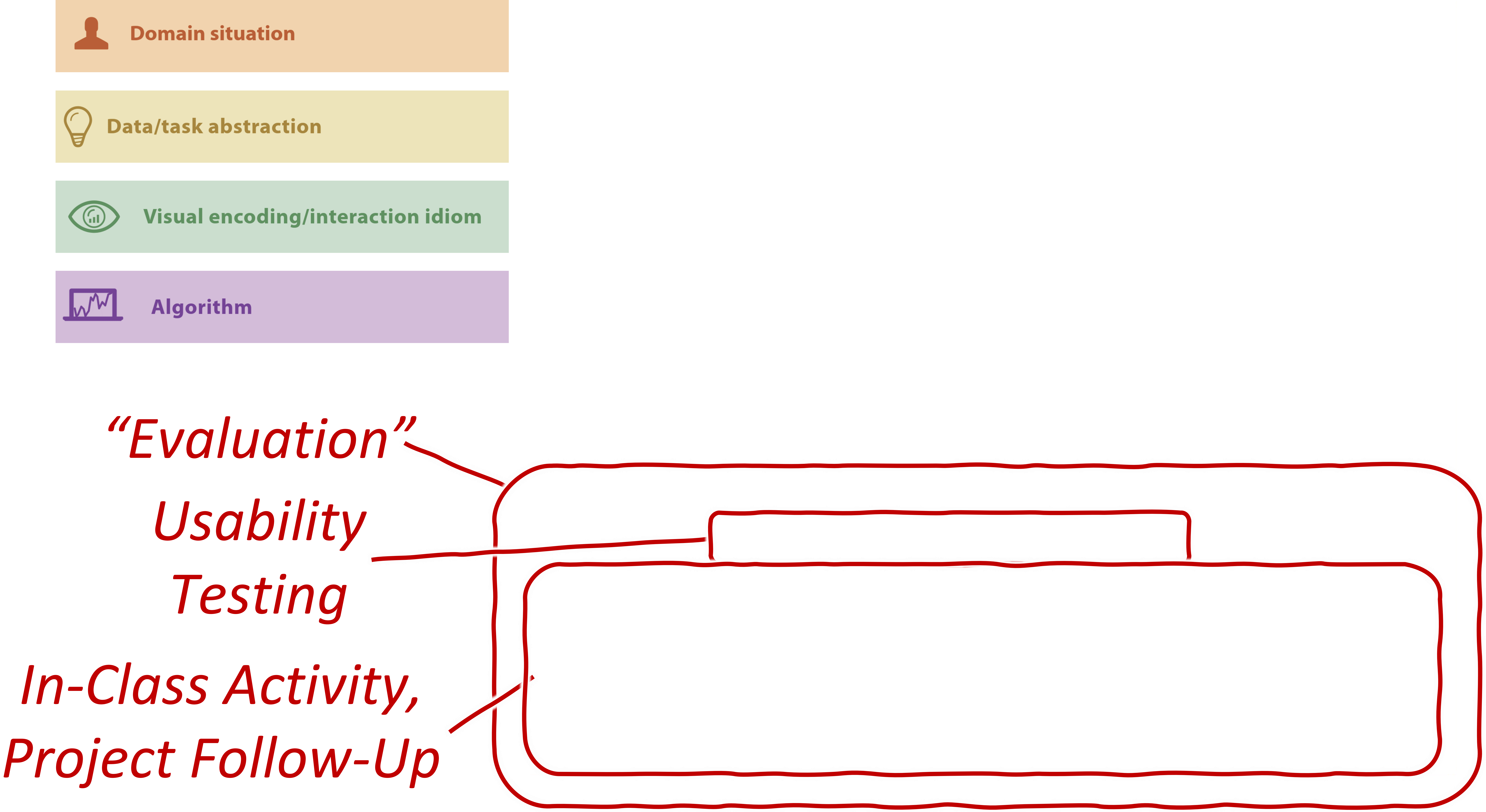
Observe target users after deployment (*field study*)

Measure adoption

[Tamara Munzner]

Threats to Validity

✓ Final Project validation



Empirical Studies in Information Visualization: Seven Scenarios

Empirical Studies in Information Visualization: Seven Scenarios

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

Abstract—We take a new, scenario-based look at evaluation in information visualization. Our seven scenarios, evaluating visual data analysis and reasoning, evaluating user performance, evaluating user experience, evaluating environments and work practices, evaluating communication through visualization, evaluating visualization algorithms, and evaluating collaborative data analysis were derived through an extensive literature review of over 800 visualization publications. These scenarios distinguish different study goals and types of research questions and are illustrated through example studies. Through this broad survey and the distillation of these scenarios, we make two contributions. One, we encapsulate the current practices in the information visualization research community and, two, we provide a different approach to reaching decisions about what might be the most effective evaluation of a given information visualization. Scenarios can be used to choose appropriate research questions and goals and the provided examples can be consulted for guidance on how to design one's own study.

Index Terms—Information visualization, evaluation.

1 Introduction

Evaluation in information visualization is complex since, for a thorough understanding of a tool, it not only involves assessing the visualizations themselves, but also the complex processes that a tool is meant to support. Examples of such processes are exploratory data analysis and reasoning, communication through visualization, or collaborative data analysis. Researchers and practitioners in the field have long identified many of the challenges faced when planning, conducting, and executing an evaluation of a visualization tool or system [10], [41], [54], [63]. It can be daunting for evaluators to identify the right evaluation questions to ask, to choose the right variables to evaluate, to pick the right tasks, users, or data sets to test, and to pick appropriate evaluation methods. Literature guidelines exists that can help with these problems but they are almost exclusively focused on methods—"structured as an enumeration of methods with focus on *how* to carry them out, without prescriptive advice for *when* to choose between them." ([54, p.1], author's own emphasis).

This paper takes a different approach: instead of focusing on evaluation methods, we provide an in-depth discussion of evaluation scenarios, categorized into those for understanding data analysis processes and those which evaluate visualizations themselves.

The scenarios for understanding data analysis are

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

The scenarios for understanding visualizations are

- Evaluating user performance (UP),
- evaluating user experience (UE), and
- evaluating visualization algorithms (VA).

Our goal is to provide an overview of different types of evaluation scenarios and to help practitioners in setting the right evaluation goals, picking the right questions to ask, and to consider a variety of methodological alternatives to evaluation for the chosen goals and questions. Our scenarios were derived from a systematic analysis of 850 papers (361 with evaluation) from the information visualization research literature (Section 5). For each evaluation scenario, we list the most common evaluation goals and outputs, evaluation questions, and common approaches in Section 6. We illustrate each scenario with representative published evaluation examples from the information visualization community. In cases where there are gaps in our community's evaluation approaches, we suggest examples from other fields. We strive to provide a wide coverage of the methodology space in our scenarios to offer a diverse set of evaluation options. Yet, the "Methods and Examples" lists in this paper are not meant to be comprehensive as our focus is on choosing among evaluation scenarios. Instead, we direct the interested reader

- *H. Lam is with Google, Inc, Mountain View, CA. E-mail: heidi.lam@gmail.com.*
- *E. Bertini is with the Department of Computer and Information Science, University of Konstanz, Box 78, Konstanz 78457, Germany. E-mail: enrico.bertini@uni-konstanz.de.*
- *P. Isenberg is with INRIA, Université Paris-Sud, Team Aviz, Bat 650, Saclay, Orsay Cedex 91405, France. E-mail: petra.isenberg@inria.fr.*
- *C. Plaisant is with the University of Maryland, 2117C Hornbake South Wing, College Park, MD 20742. E-mail: plaisant@cs.umd.edu.*
- *S. Carpendale is with the Department of Computer Science, University of Calgary, 2500 University Dr. NW, Calgary, AB T2N 1N4, Canada. E-mail: sheelagh@ucalgary.ca.*

Manuscript received 8 Sept. 2010; revised 6 Nov. 2011; accepted 9 Nov. 2011; published online 30 Nov. 2011.
Recommended for acceptance by C. North.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2010-09-0224.
Digital Object Identifier no. 10.1109/TVCG.2011.279.

Lam et al., 2012

Empirical Studies in Information Visualization: Seven Scenarios

Lam et al., 2012

7 Evaluation Scenarios

How to understand your data:

- Understanding Environments and Work Practices — Domain situation
- Evaluating Visual Data Analysis and Reasoning — Data/task abstraction
- Evaluating Communication Through Visualization — Visual encoding/interaction idiom
- Evaluating Collaborative Data Analysis — Data/task abstraction

How to understand your visualization:

- Evaluating User Performance — Visual encoding/interaction idiom — Data/task abstraction
- Evaluating User Experience — Visual encoding/interaction idiom — Data/task abstraction
- Evaluating Visualization Algorithms — Algorithm

Understanding environments and work practices

Domain situation

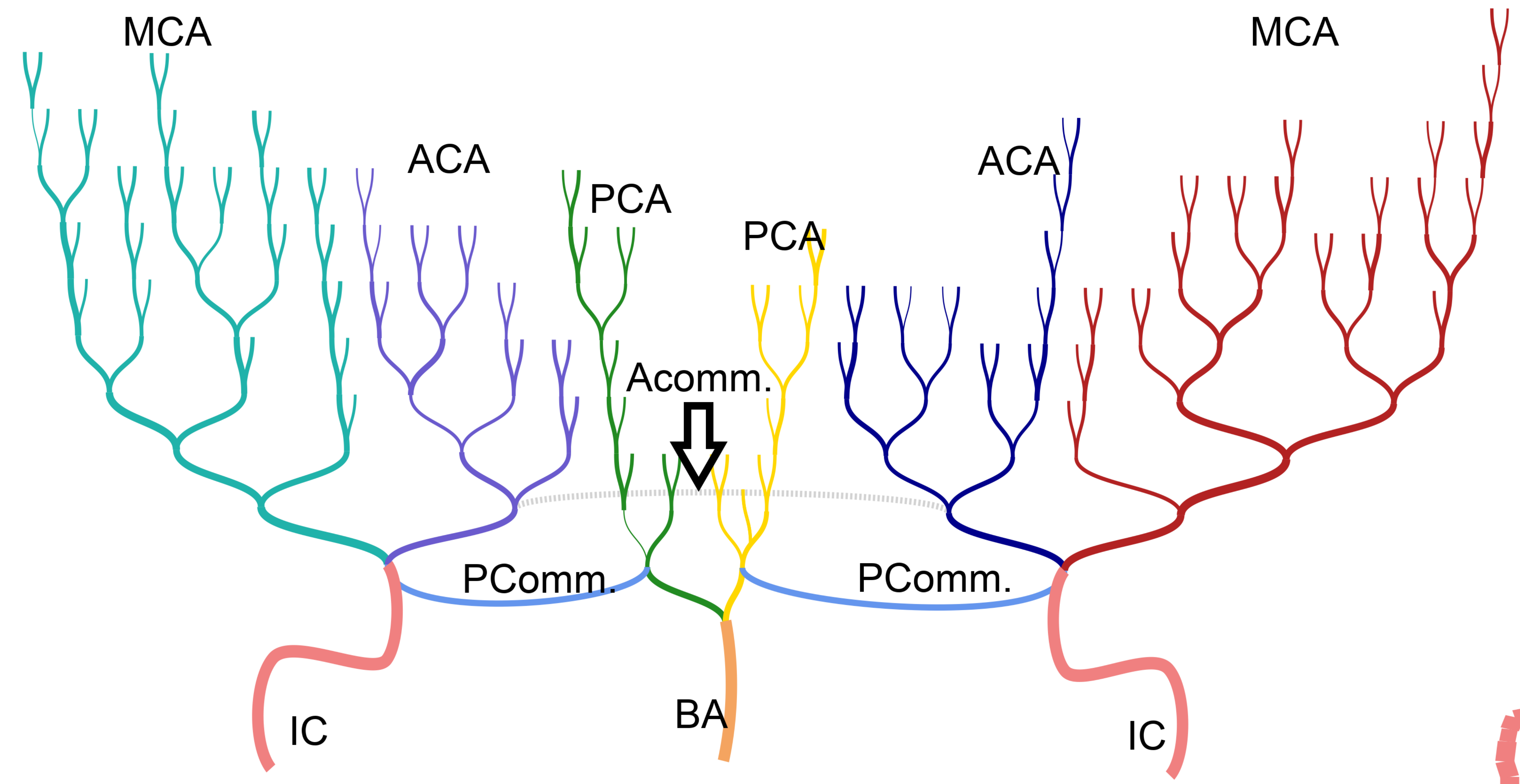
- Goals & outputs
  - Understand work, analysis, or information processing practices of people
  - Without software in use: inform design
  - With software in use: assess factors for adoption, how appropriated for future design
- Evaluation Questions
  - Context of use?
  - Integrate into which daily activities?
  - Supported analyses?
  - Characteristics of user group and environment?
  - What data & tasks?
  - What visualizations/tools used?
  - How current tools solve tasks?
  - Challenges and usage barrier?

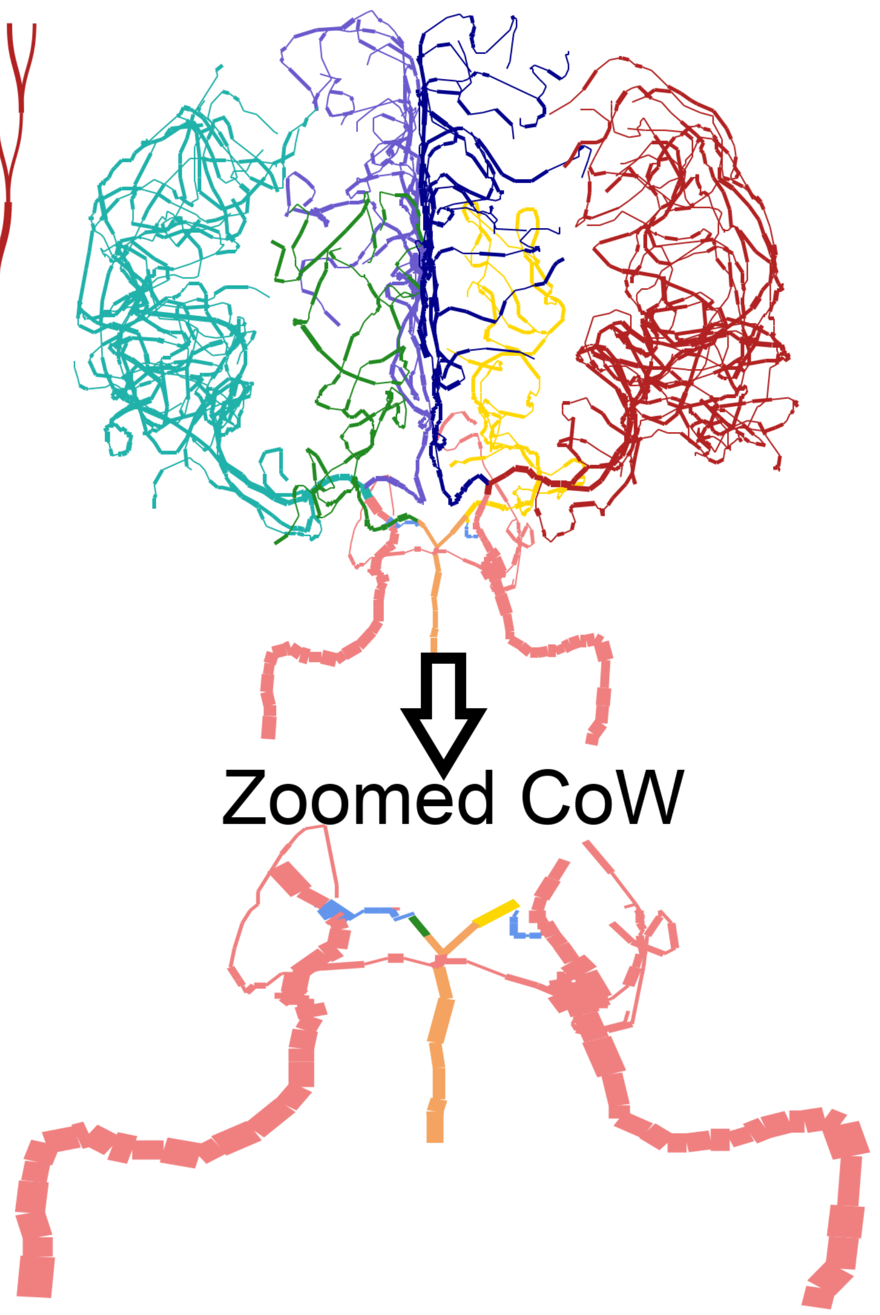
Understanding environments and work practices

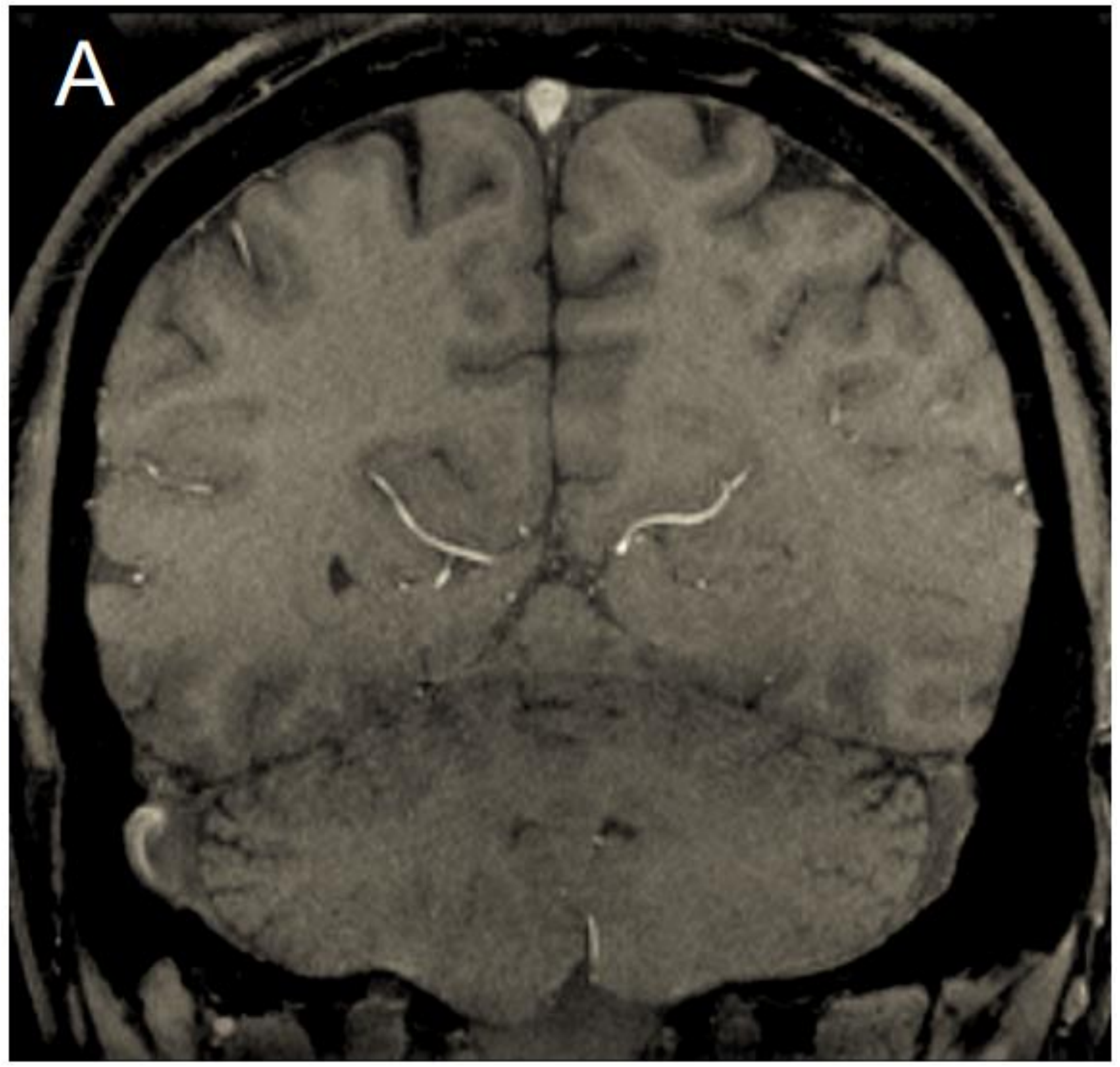
Domain situation

- Methods
  - Field Observation
    - Real world, free use of tool
    - Derive requirements
  - Interviews
    - Contextual inquiry: interview then observe in routines, with little interference
    - Pick the right person
    - Laboratory context w/domain expert
  - Laboratory Observation
    - How people interact with each other, tools
    - More control of situation

Understanding environments and work practices: Example







Pandey, Dunne, et al., 2019

Evaluating visual data analysis and reasoning

Data/task abstraction

- Goals & outputs
  - Assess visualization tool's ability to support visual analysis and reasoning
    - As a whole! Not just a technique
  - Quantifiable metrics or subjective feedback
- Evaluation Questions: Does it support…
  - Data exploration?
  - Knowledge discovery?
  - Hypothesis generation?
  - Decision making?

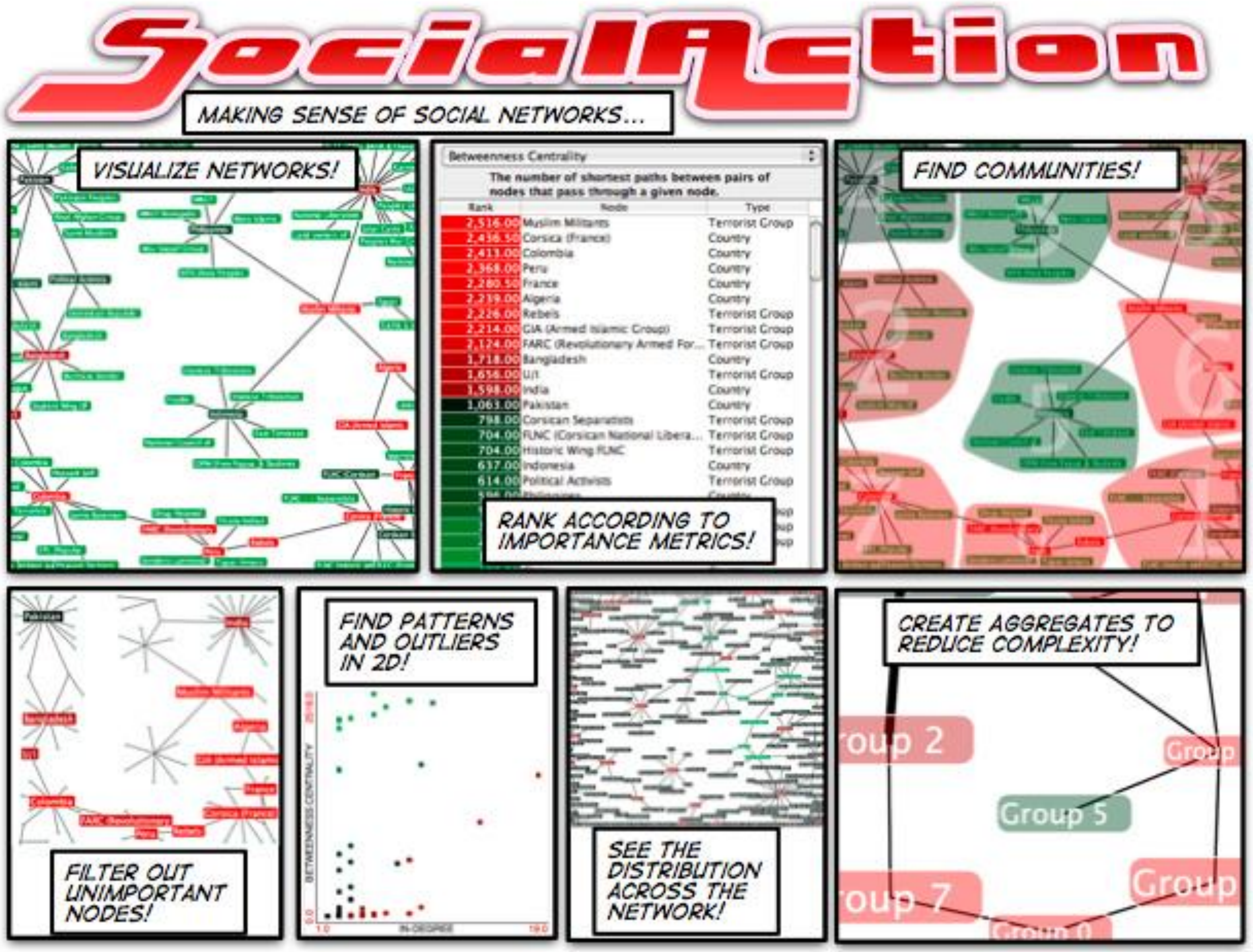
Evaluating visual data analysis and reasoning

Data/task abstraction

- Methods
  - Case studies
    - Motivated experts with own data in own environment
    - Can be longitudinal
    - Insight-Based (Saraiya et al., 2004)
      - Unguided, diary, debriefing meetings
    - MILCS: Multidimensional In-depth Long-term Case studies (Shneiderman & Plaisant, 2006)
      - Guided, observations, interviews, surveys, automated logging
      - Assess interface efficacy, user performance, interface utility
      - Improve system during
  - Lab observations and interviews
    - Code results
    - Think aloud
  - Controlled Experiment
    - Isolate important factors

Lam et al., 2012

Evaluating visual data analysis and reasoning



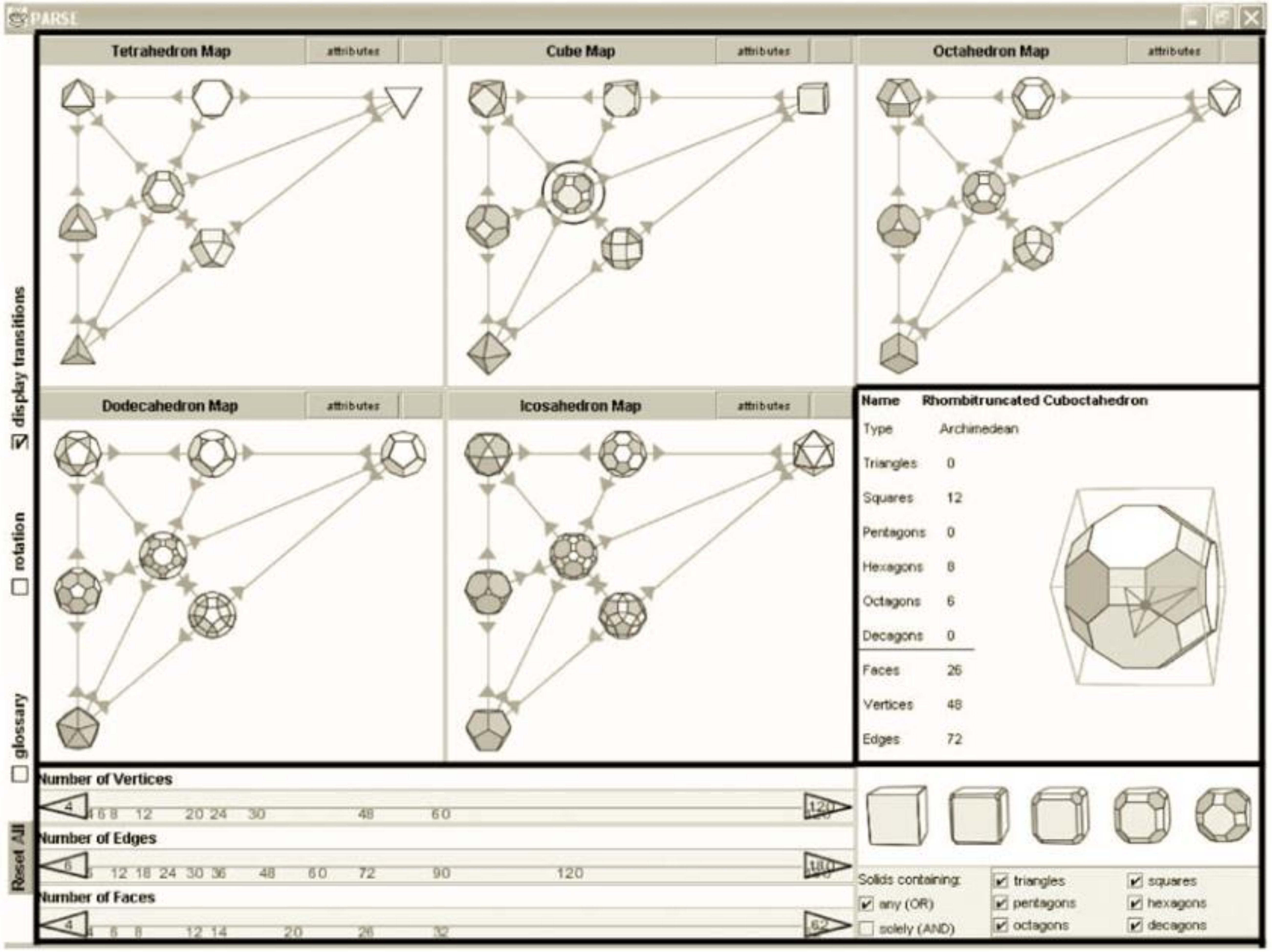
Perer et al., 2006

Evaluating communication through visualization

Visual encoding/interaction idiom

- Goals & outputs
  - How effectively is a message delivered and acquired
- Evaluation Questions
  - Quantitative: learning rate, information retention and accuracy
  - Qualitative: interaction patterns
- Methods
  - Controlled experiments
  - Field observation & interviews

Evaluating communication through visualization: Example



Evaluating Collaborative Data Analysis

Data/task abstraction

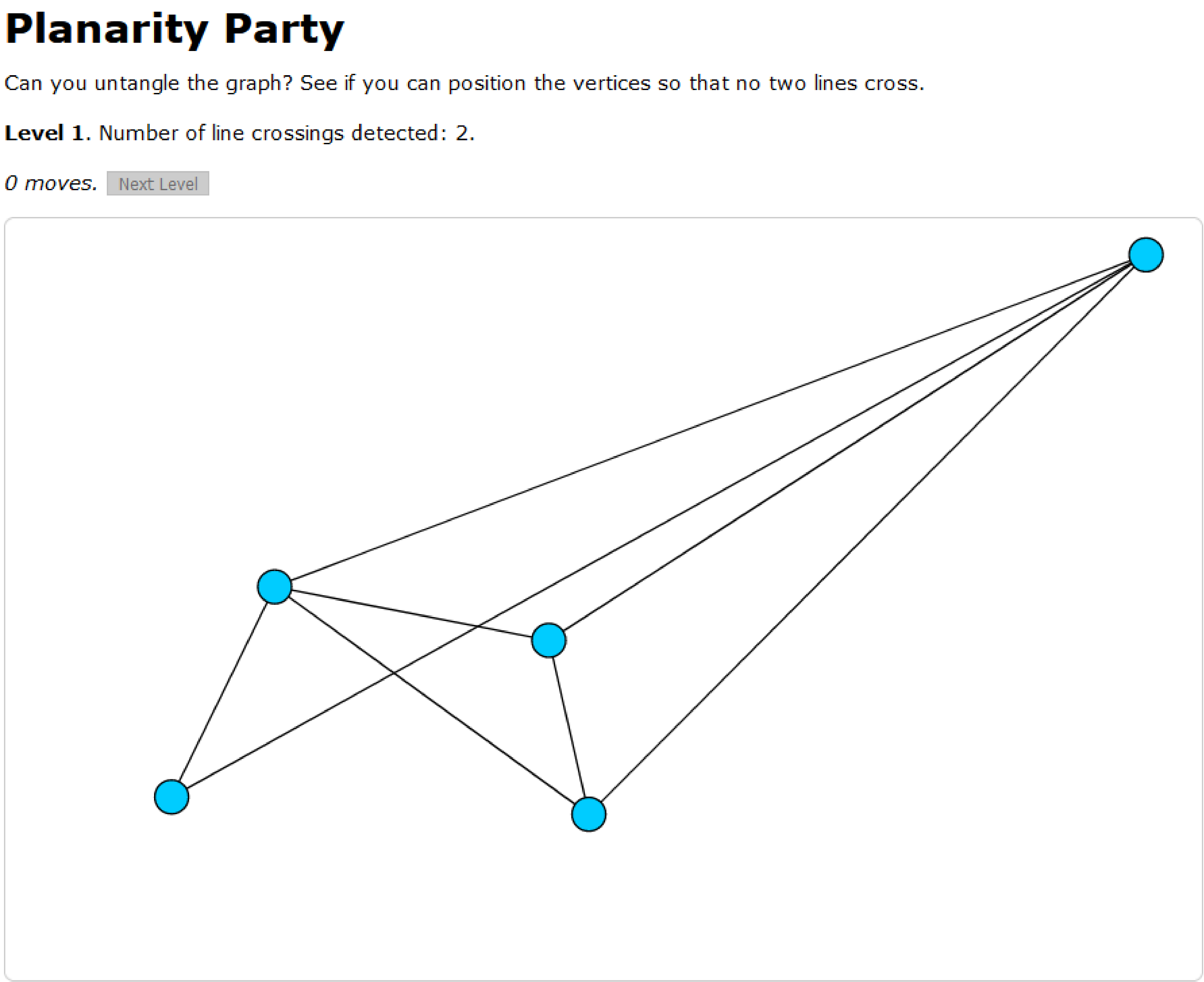
- Goals & outputs
  - Evaluate support for taskwork and teamwork
  - Holistic understanding of group work processes or tool use
  - Derive design implications
- Evaluation Questions
  - Effective and efficient?
  - Satisfactorily support or stimulate group sensemaking?
  - Support group insight?
  - Is social exchange and communication facilitated?
  - How is the tool used? Features, patterns…
  - What is the process? User requirements?

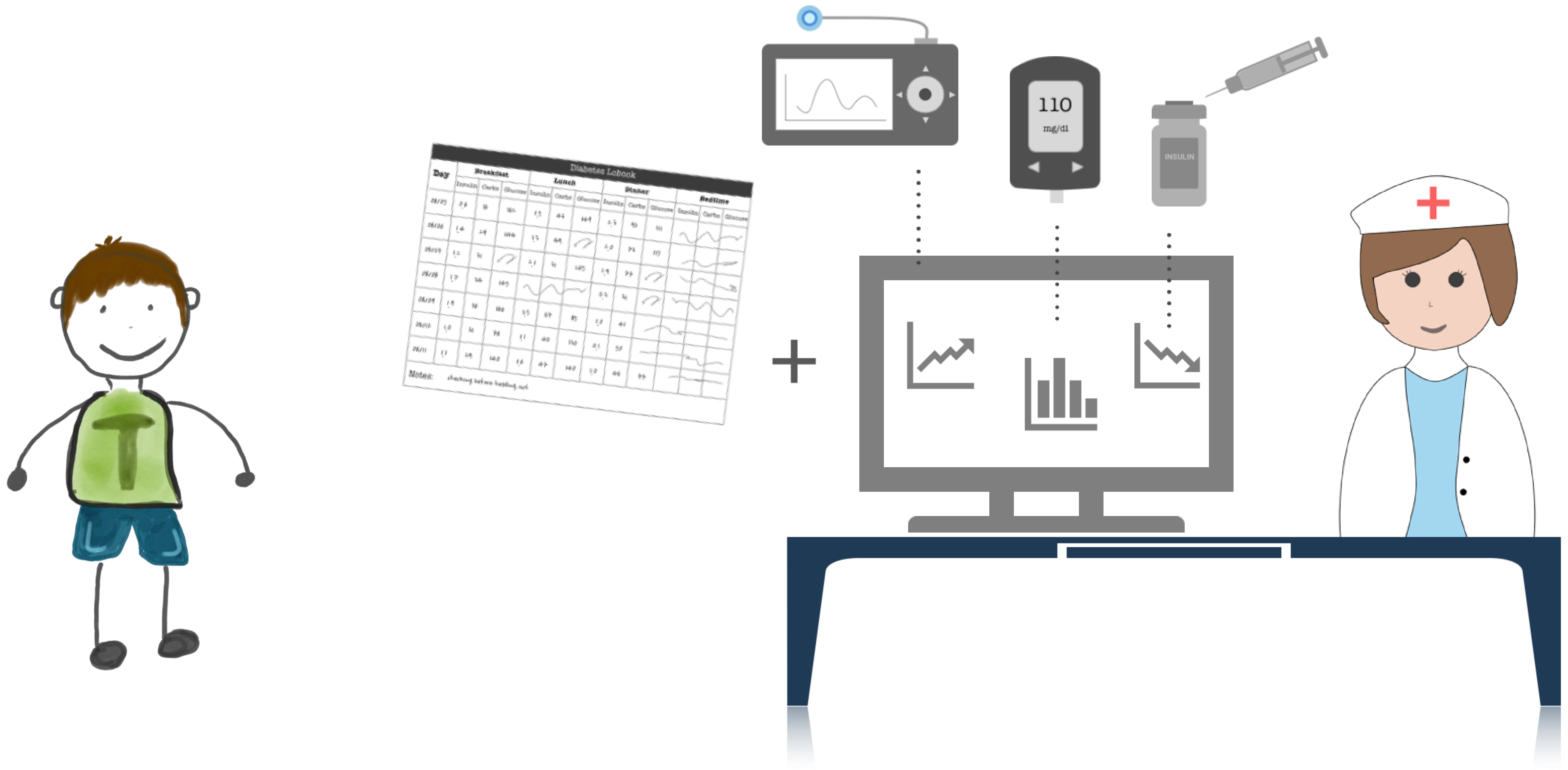
Evaluating Collaborative Data Analysis

Data/task abstraction

- Methods
  - Context critical, but early formative studies less dependant
  - Heuristic evaluation
    - Heuristics: actions, mechanics, interactions, locales needed
  - Log analysis
    - Distributed or web-based tools
    - Combine with questionnaire or interview
    - Hard to evaluate unlogged & qualitative aspects
  - Field or laboratory observation
    - Involve group interactions and harmony/disharmony
    - Combine with insight-based?

Evaluating Collaborative Data Analysis: Examples

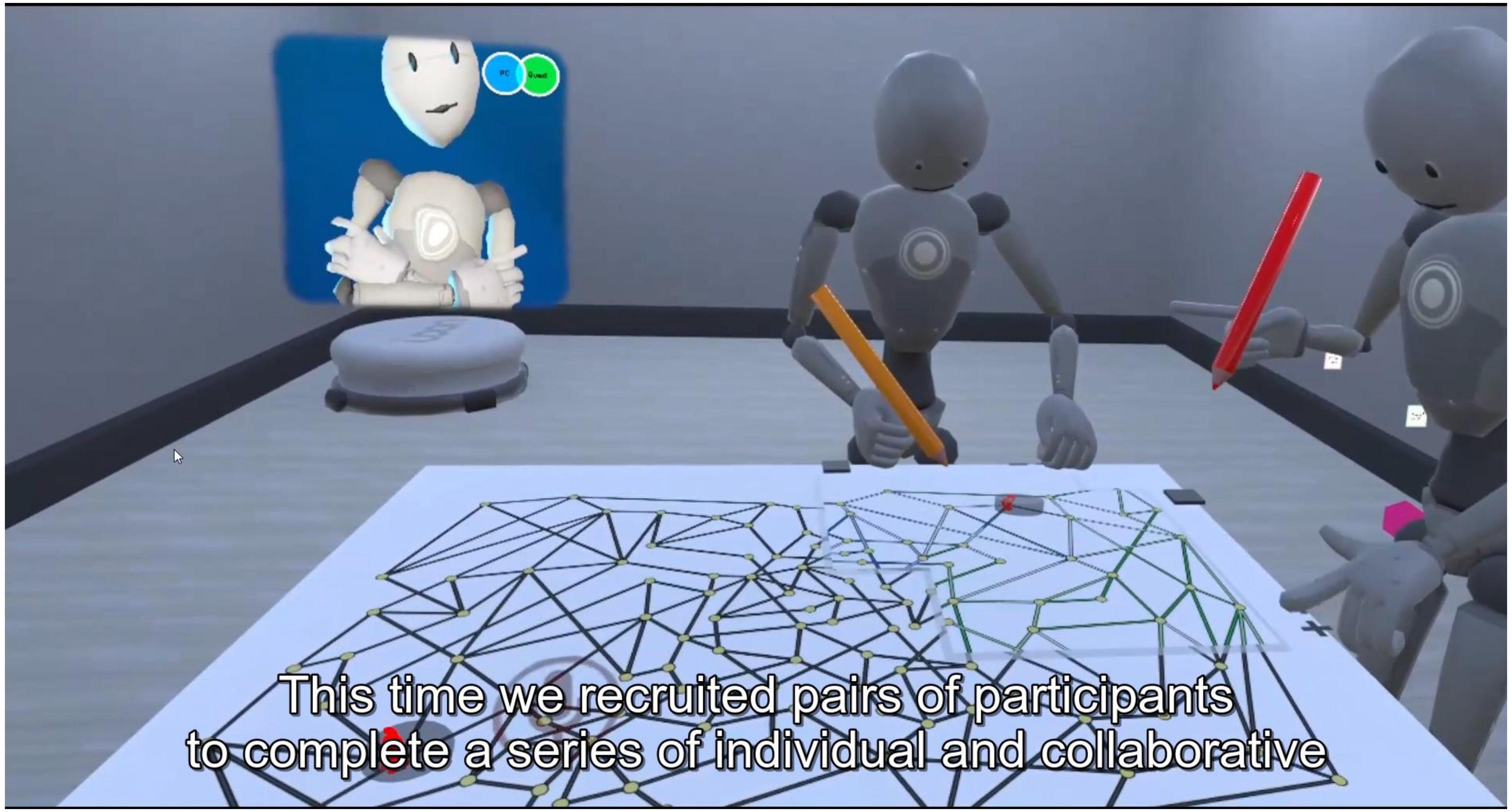




[Schwab, … Dunne, … et al., 2020]

[Zhang, … Dunne, … et al., 2018]

Evaluating Collaborative Data Analysis



[Saffo et al., 2021](#)

Model	Parameters	Estimate (original)	Estimate (our result)	SE (original)	SE (our result)	sr (original)	sr (our result)	p (original)	p (our result)	R^2 (original)	R^2 (our result)	SE (ms) (original)	SE (ms) (our result)	Adj. R^2 (our result)
Hoffmann (Fitts') [14]	Intercept	804.287	461.674	123.000	21.963	-	-	-	-	0.501	0.780	427.450	83.113	0.778
	ID	411.940	123.752	42.440	6.920	0.708	0.883	<0.001	<0.001	-	-	-	-	-
Murata & Iwase [22]	Intercept	805.404	426.211	123.907	21.691	-	-	-	-	0.501	0.788	429.697	82.077	0.783
	ID	411.760	123.173	42.678	6.842	0.707	0.878	<0.001	<0.001	-	-	-	-	-
	Sin(φ)	8.506	-23.678	61.130	13.065	0.010	-0.088	0.890	0.073	-	-	-	-	-
Cha & Myung [7]	Intercept	819.971	458.177	152.789	20.898	-	-	-	-	0.501	0.806	431.963	78.897	0.800
	ID	411.846	124.293	42.060	6.588	0.707	0.885	<0.001	<0.001	-	-	-	-	-
	Sin(φ)	6.429	-21.950	62.735	12.573	0.008	-0.082	0.919	0.084	-	-	-	-	-
	θ	-0.290	0.557	1.759	0.193	-0.012	0.089	0.135	0.005	-	-	-	-	-
Machuca & Stuerzlinger [3]	Intercept	795.413	458.381	118.073	21.190	-	-	-	-	0.545	0.798	410.202	80.070	0.794
	ID	415.547	124.640	40.742	6.674	0.713	0.889	<0.001	<0.001	-	-	-	-	-
	CTD	3.505	0.651	1.164	0.230	0.211	0.134	0.003	0.006	-	-	-	-	-
Clark et al. [8]	Intercept	732.369	454.952	105.456	21.286	-	-	-	-	0.645	0.802	364.162	79.790	0.795
	ID	433.127	125.566	36.319	6.690	0.741	0.890	<0.001	<0.001	-	-	-	-	-
	θ	-2.681	0.373	1.170	0.277	-0.142	0.064	0.024	0.181	-	-	-	-	-
	S x θ	0.336	0.012	0.060	0.012	0.346	0.048	<0.001	0.313	-	-	-	-	-

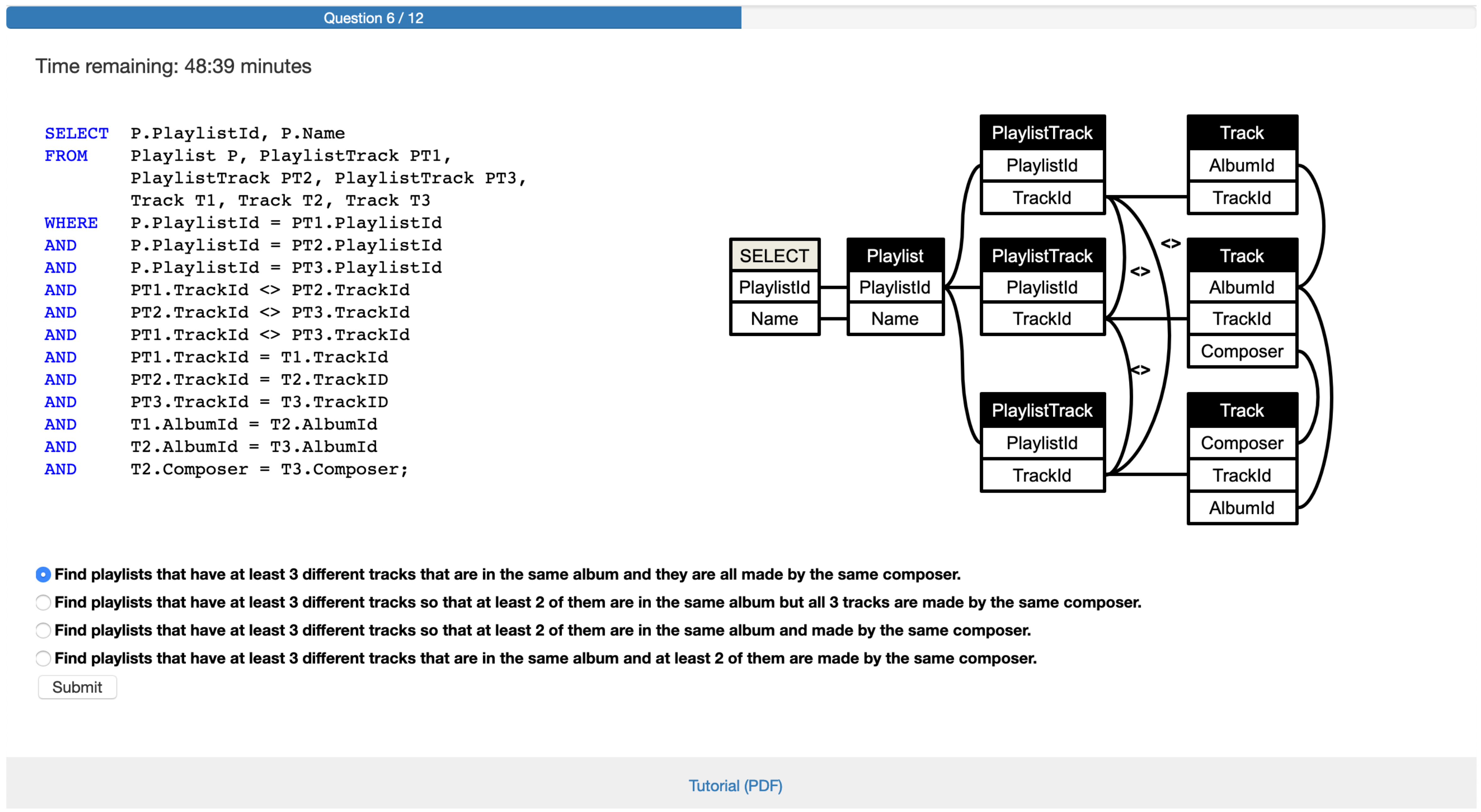
Evaluating User Performance

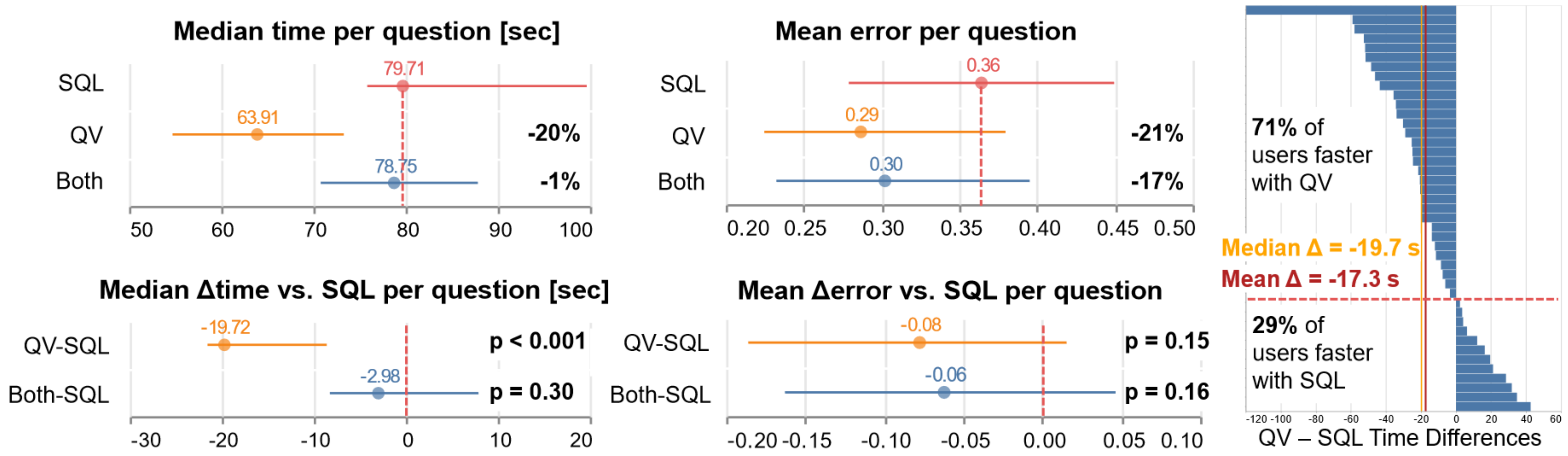
Data/task abstraction

Visual encoding/interaction idiom

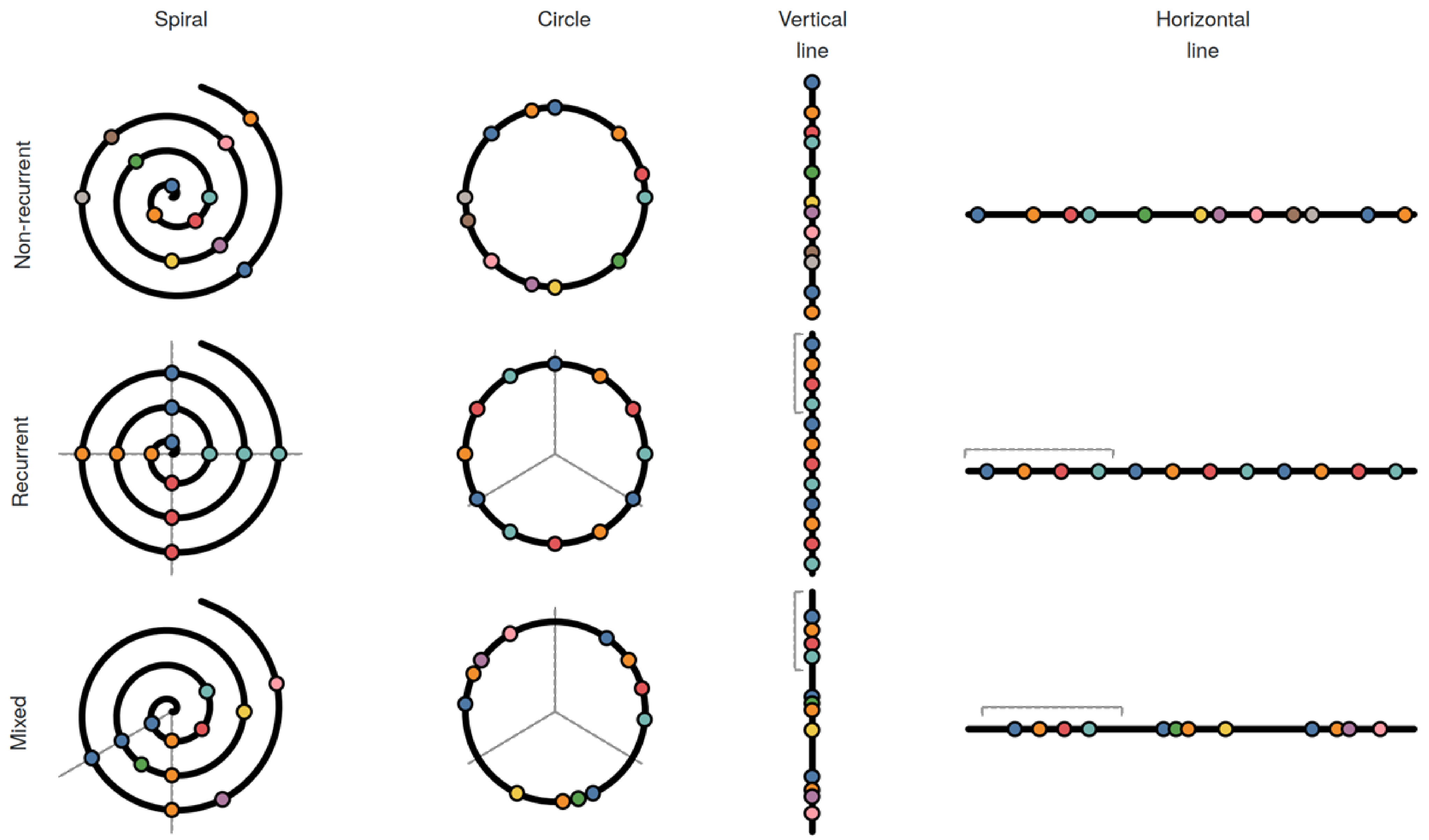
- Goals & outputs
  - Measure specific features
  - Time, accuracy, and error; work quality (if quantifiable); memorability
  - Descriptive statistics results
- Evaluation Questions
  - What are the limits of human perception and cognition?
  - How do techniques compare?
- Methods
  - Controlled experiment → design guideline, model, head-to-head
    - Few variables
    - Simple tasks
    - Individual differences matter
  - Field logs
    - Suggest improvements, recommendation systems

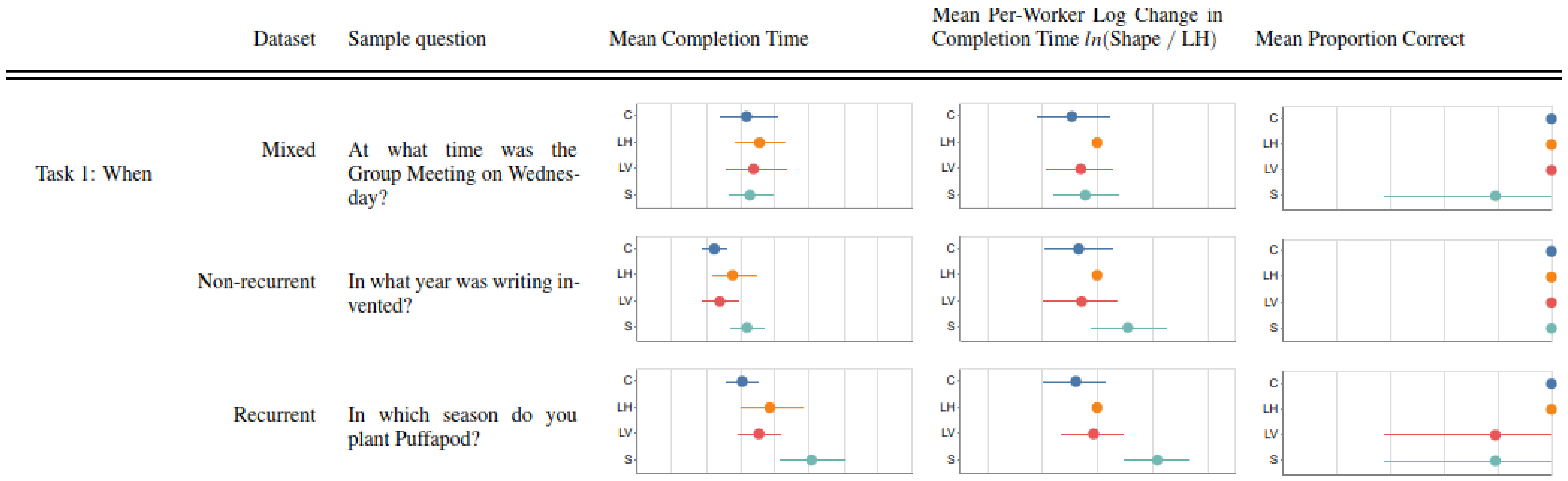
Evaluating User Performance: Examples





Leventidis, Dunne, et al., 2020

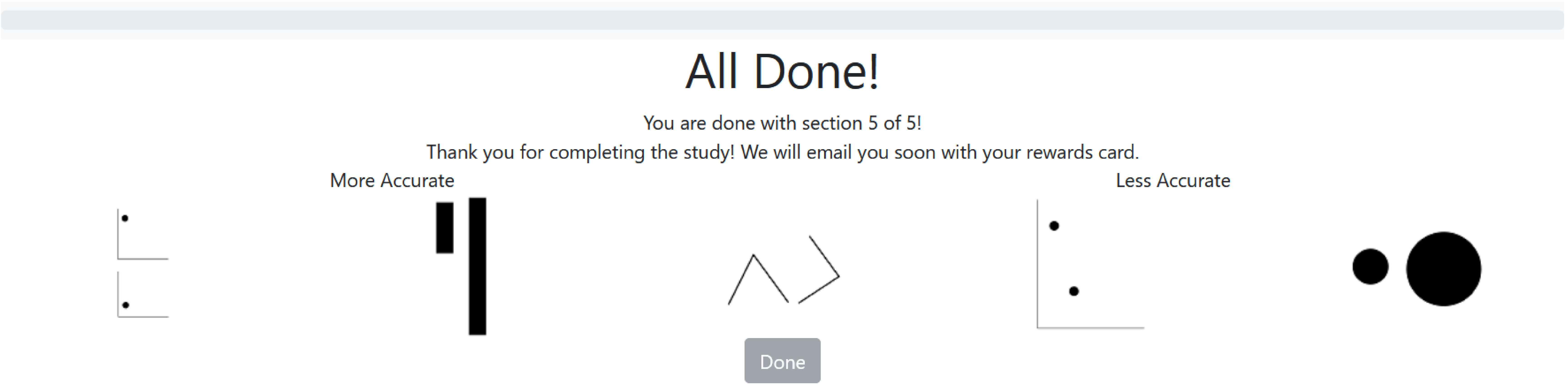




Di Bartolomeo, Dunne, et al., 2020

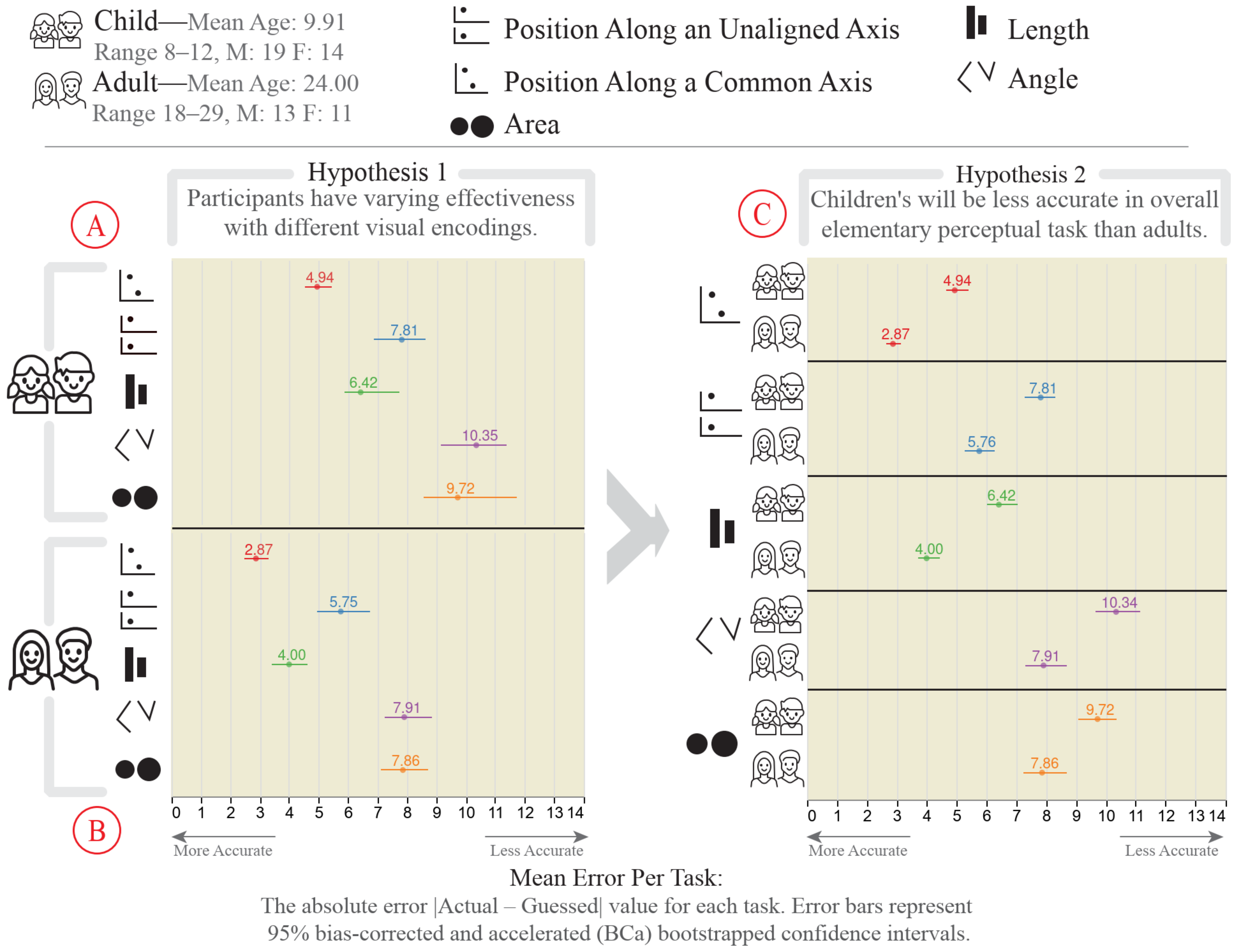
In-Class Study—Graphical Perception

~16 min



Panavas et al., 2022

Paper results



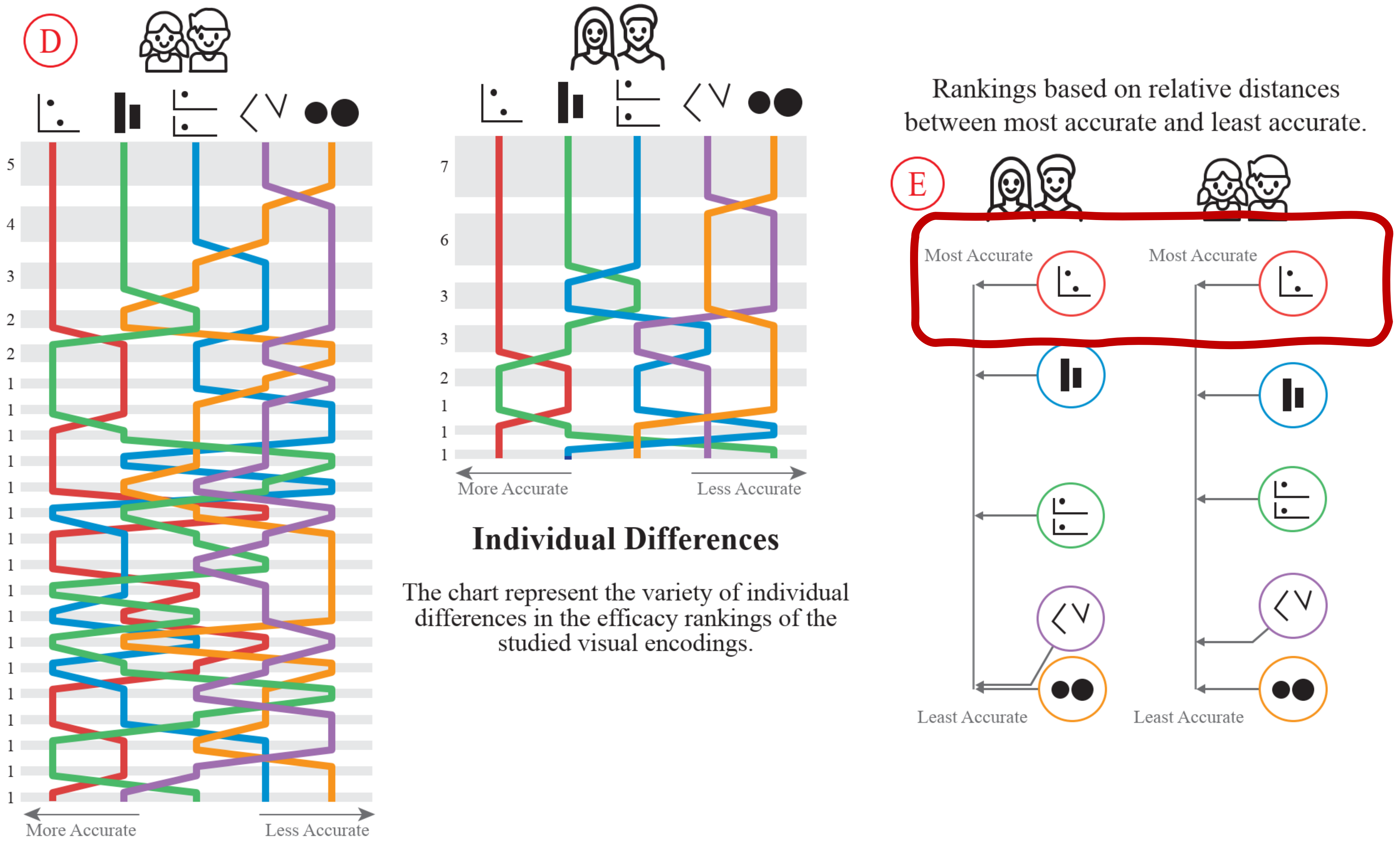


Fig. 5. Summative results for Hypothesis 1 and 2 and an exploratory analysis of individual differences in rankings. In (A), (B), and (C) the error bars show 95% bias-corrected and accelerated (BCa) bootstrapped confidence intervals [23]. (A *rough* rule of thumb for reading 95% CIs is that if two intervals overlap by less than 1/4 of their average length, then the comparison will have p < .05 [22].) The mean absolute error for each encoding is shown in (A) for children and (B) for adults. In (C), the previous two charts are rearranged to compare children with adults. Children are clearly less accurate when using each of the encodings. The exploratory analysis included, (D), shows the variation in encoding rankings among individual children (left) and adults (right). Each line represents an encoding, ranked left-to-right in increasing mean absolute error for each task. The grey rows are sized to represent the count of individuals with a shared ranking. E.g,. the top row shows that 5 children ranked Position Along a Common Axis as most accurate, followed by Length, Position Along an Unaligned Axis, Angle, and lastly Area. The line-row intersections show the encoding ranking for that row. Children displayed a larger variety of individual differences in encoding rankings than adults. Finally, (E) shows more simply the *overall* rankings we found for adults and children.

Panavas et al., 2022

For Next Time

neu-ds-4200-s22.github.io/schedule

Look at the upcoming assignments and deadlines

- Textbook, Readings, & Reading Quizzes—Variable days
- In-Class Activities—If due, they are due 11:59pm the same day as class

Everyday Required Supplies:
- 5+ colors of pen/pencil
- White paper
- Laptop and charger

Use Canvas Discussions for general questions, email codydunne-and-tas@ccs.neu.edu for questions specific to you.

Week	Topics	Assignments
#1: Jan 17–21	What is visualization Design rules of thumb	A1—Setting up
#2: Jan 24–28	JS development, projects Marks & channels	A2—Encodings & xenographics
#3: Jan 31–Feb 04	Data types and tasks, Tableau D3 tutorial 1/2	P1—Pitches★
#4: Feb 07–11	In-class group formation D3 tutorial 2/2	A3—Tableau analysis P2—Proposal★
#5: Feb 14–18	Altair and JupyterLab Practice Design Study	A4—D3 basic charts
#6: Feb 21–25	Arrange Tables Color, pop-out, illusions	A5—Altair basic charts P3—Interview & tasks
#7: Feb 28–Mar 04	Interaction & animation In-class project meetings 1/2	A6—D3 event handling P4—Data and sketches
#8: Mar 07–11	In-class project meetings 2/2 Trees & networks	P5—Final sketches & plan★
Mar 14–18	*Spring Break*	
#9: Mar 21–25	Spatial, 3D, and scientific vis. Office hours	A7—D3 brushing⧗ P6—Implementation 1⧗
#10: Mar 28–Apr 01	Validation & evaluation Marvin Zelen Symposium	
#11: Apr 04–08	How to give a talk, storytelling Project usability testing	A8—Brushing & linking⧗
#12: Apr 11–15	Project presentations 1/2 Project presentations 2/2	P7—Presentations★⧗
#13: Apr 18–22	Flex day	P8—Presentation peer review
#14: Apr 25–29	Reflecting & project work	
May 02–06		P9—Video & Final Deliverables★⧗