How to Evaluate (your) Visualizations
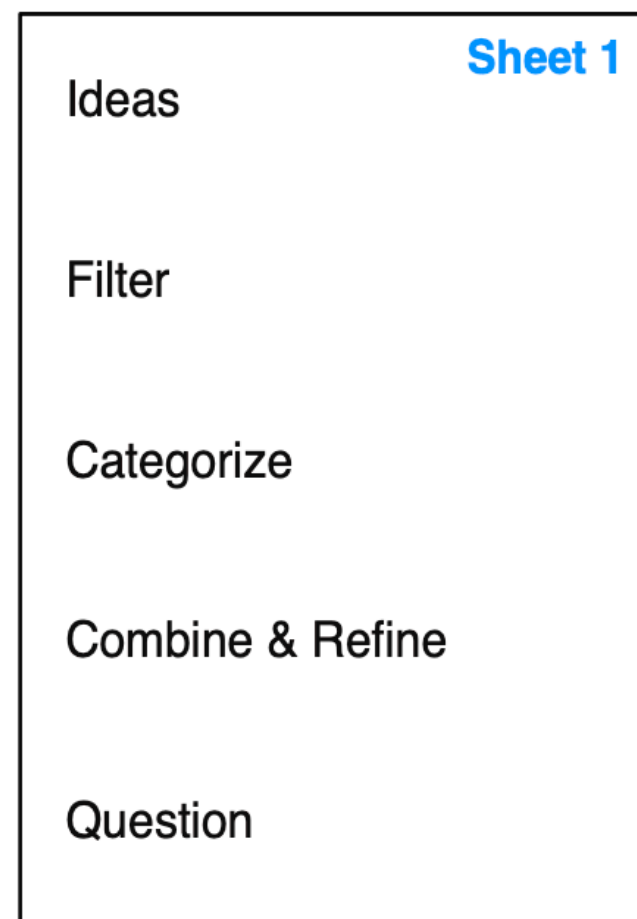
PA214 — Visualization II
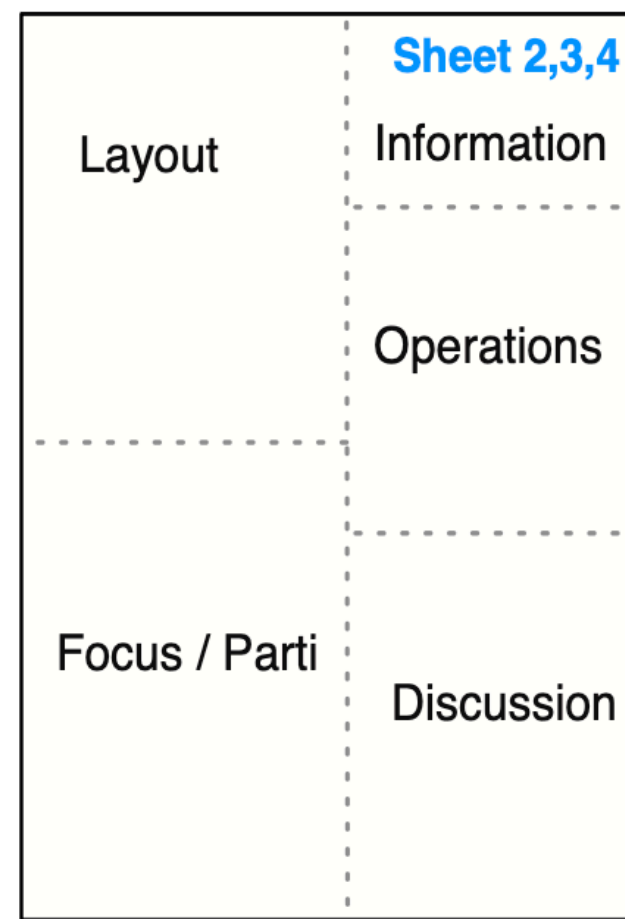
Vít Rusňák

Talk Outline

- Methodologies for visualization design

- Evaluation Categories

  - Understanding the tool vs. understanding the processes

  - Evaluation without users vs. with users

- Some tips and tricks for doing the evaluation
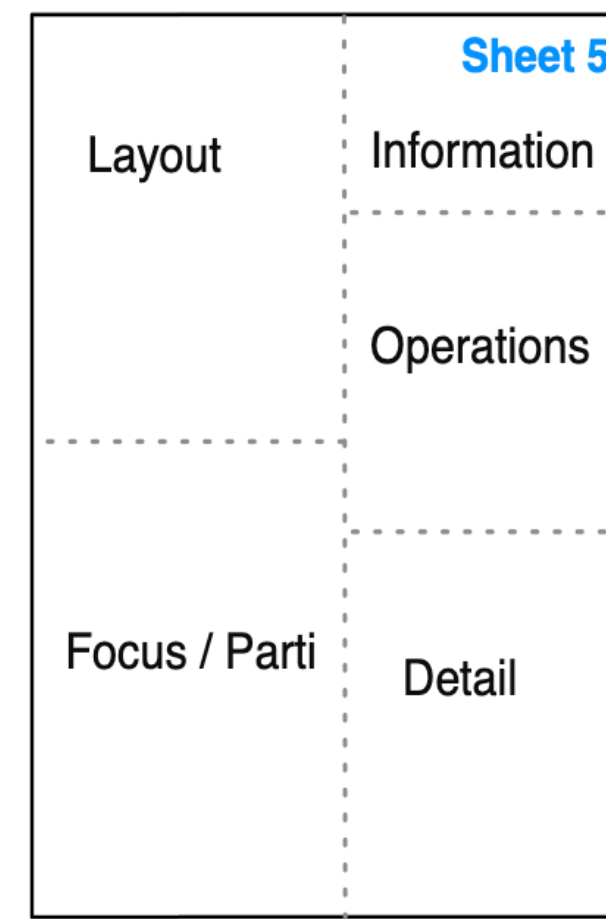
Why do we evaluate the visualizations?
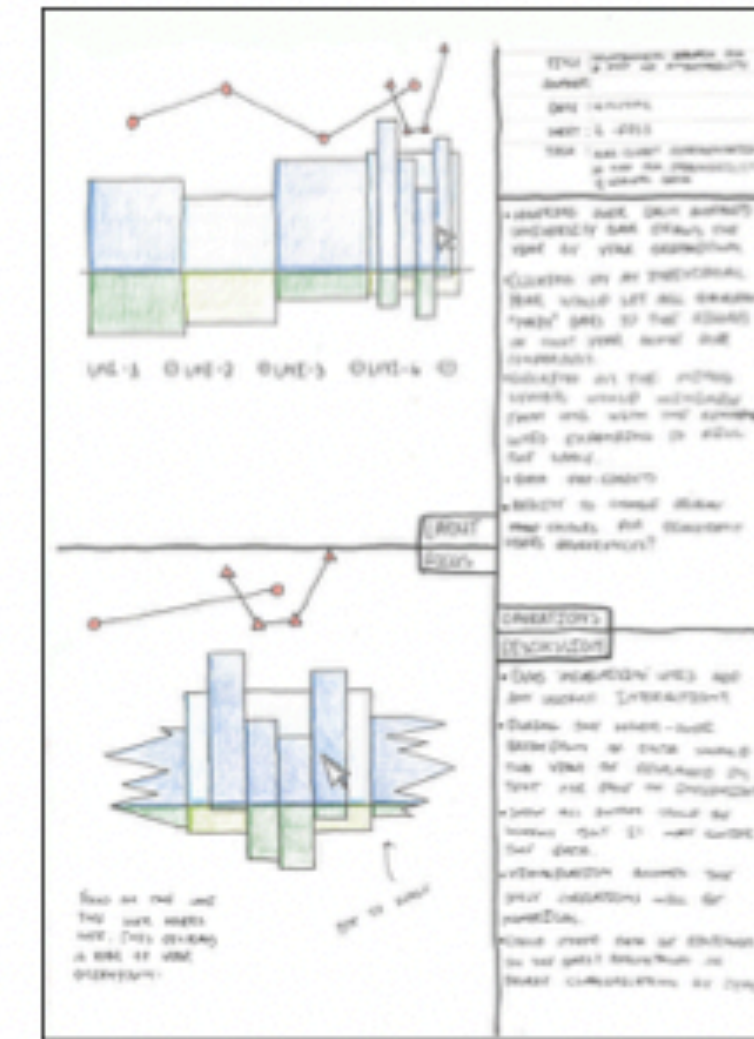
Five Design Sheets
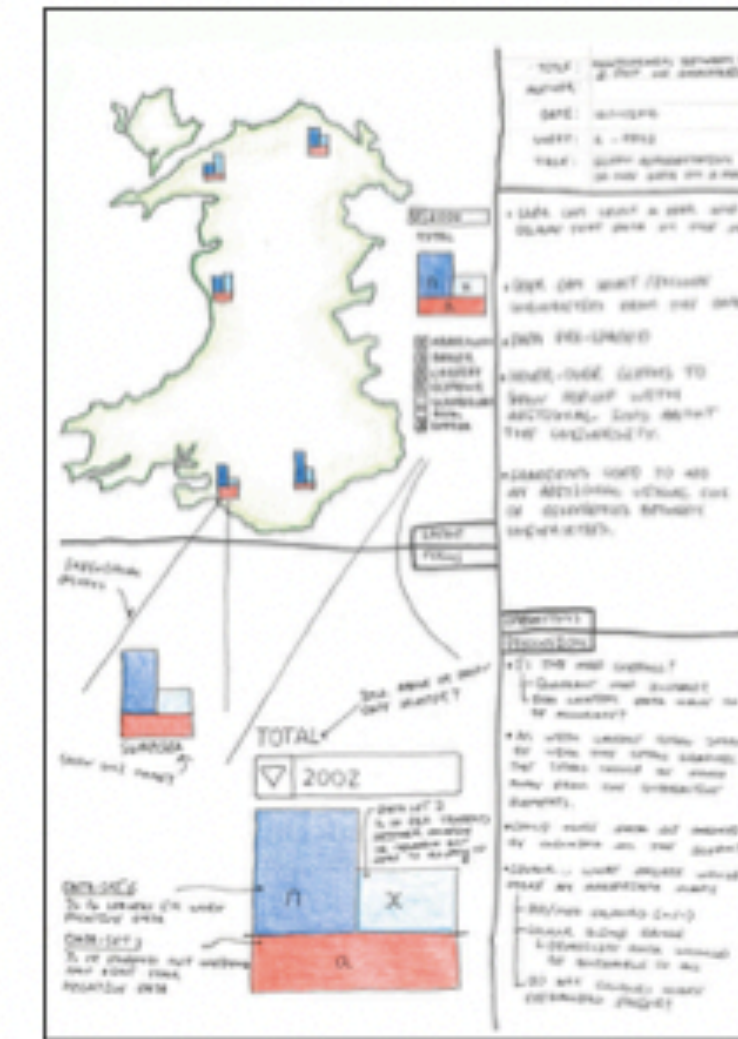
Fig. 3: The FdS sheets. (a) Sheet 1: Generate Ideas, filter, categorize, combine & refine then question. (b) Sheets 2,3,4 with the five sections in the 2-row 3-row format; (c) Sheet 5, the realization sheet where *Detail* is included instead of Discussion.

https://pdritsos.com/projects/FDS/
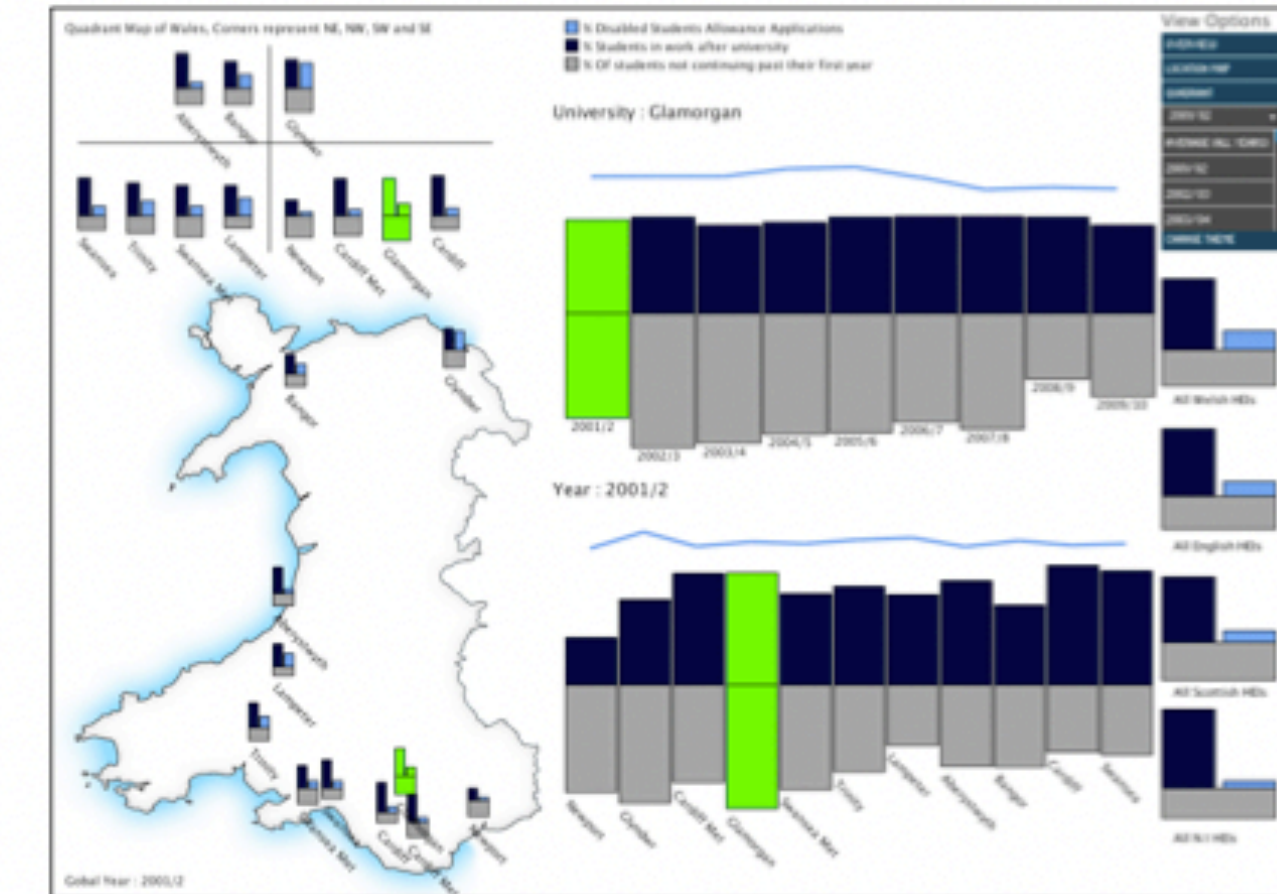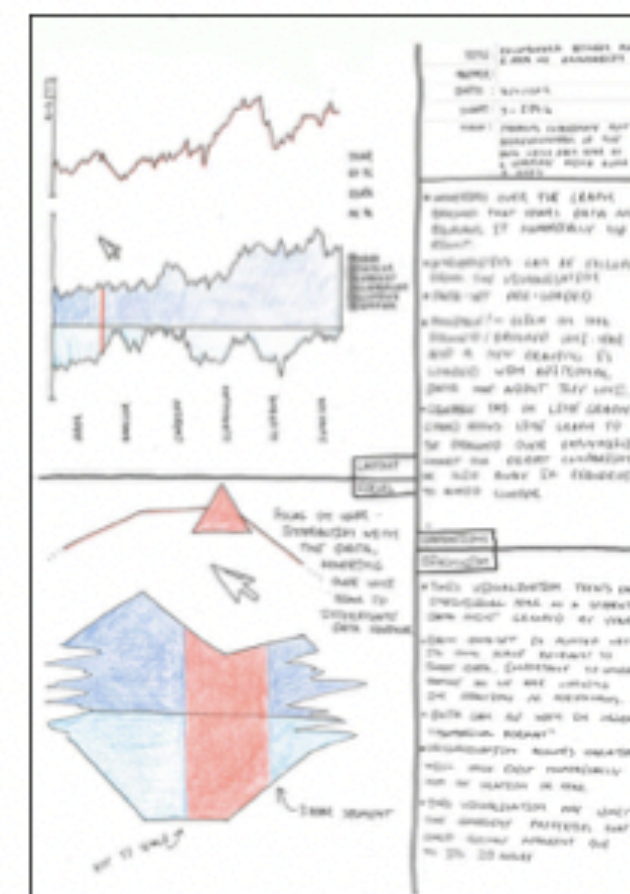
Design Study Methodology
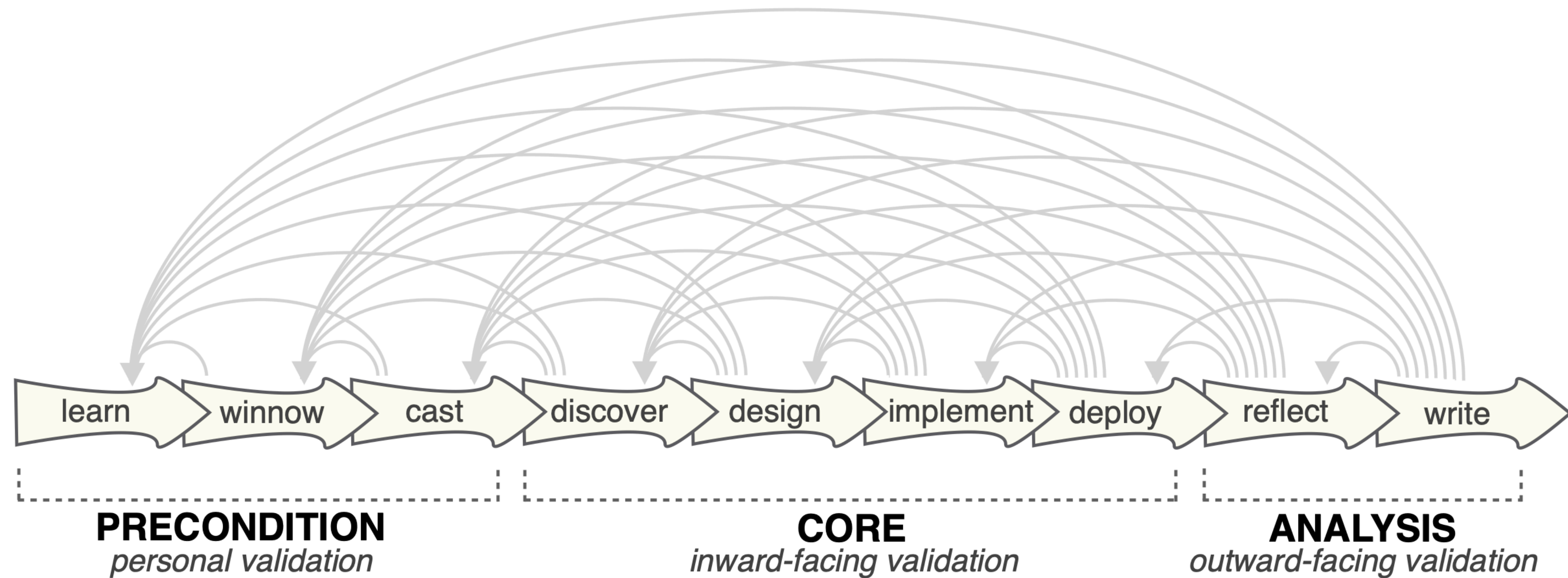
Fig. 2. Nine-stage design study methodology framework classified into three top-level categories. While outlined as a linear process, the overlapping stages and gray arrows imply the iterative dynamics of this process.

Nested Model Methodology

User-Centered Design

Adapted from: B. Shneiderman et al. Designing the User Interface, Fig. 4.1
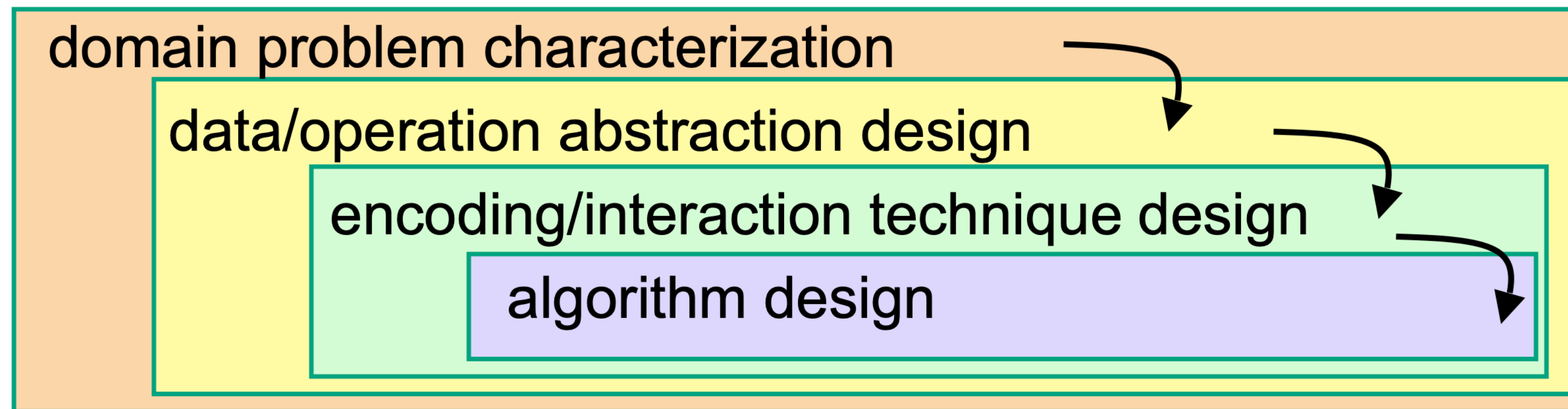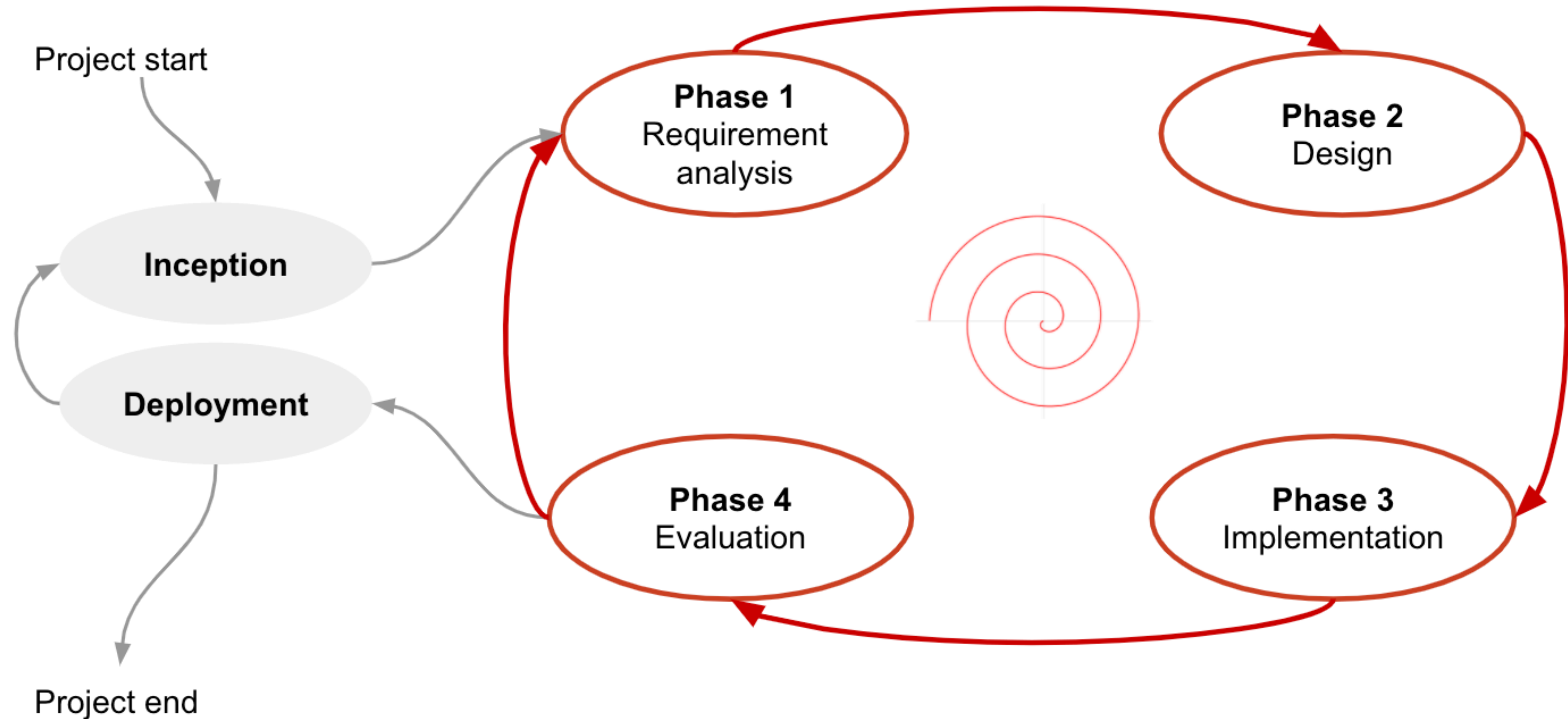
Which evaluation methods do you know?

Evaluation Categories

Most Precision

obtrusive

Laboratory experiment

Experimental simulation

Judgment Study

Field experiment

Sample Survey

Field study

Formal Theory

Computer simulation

Most Generalizability

Most Realism

unobtrusive

Abstract

Tasks

Concrete
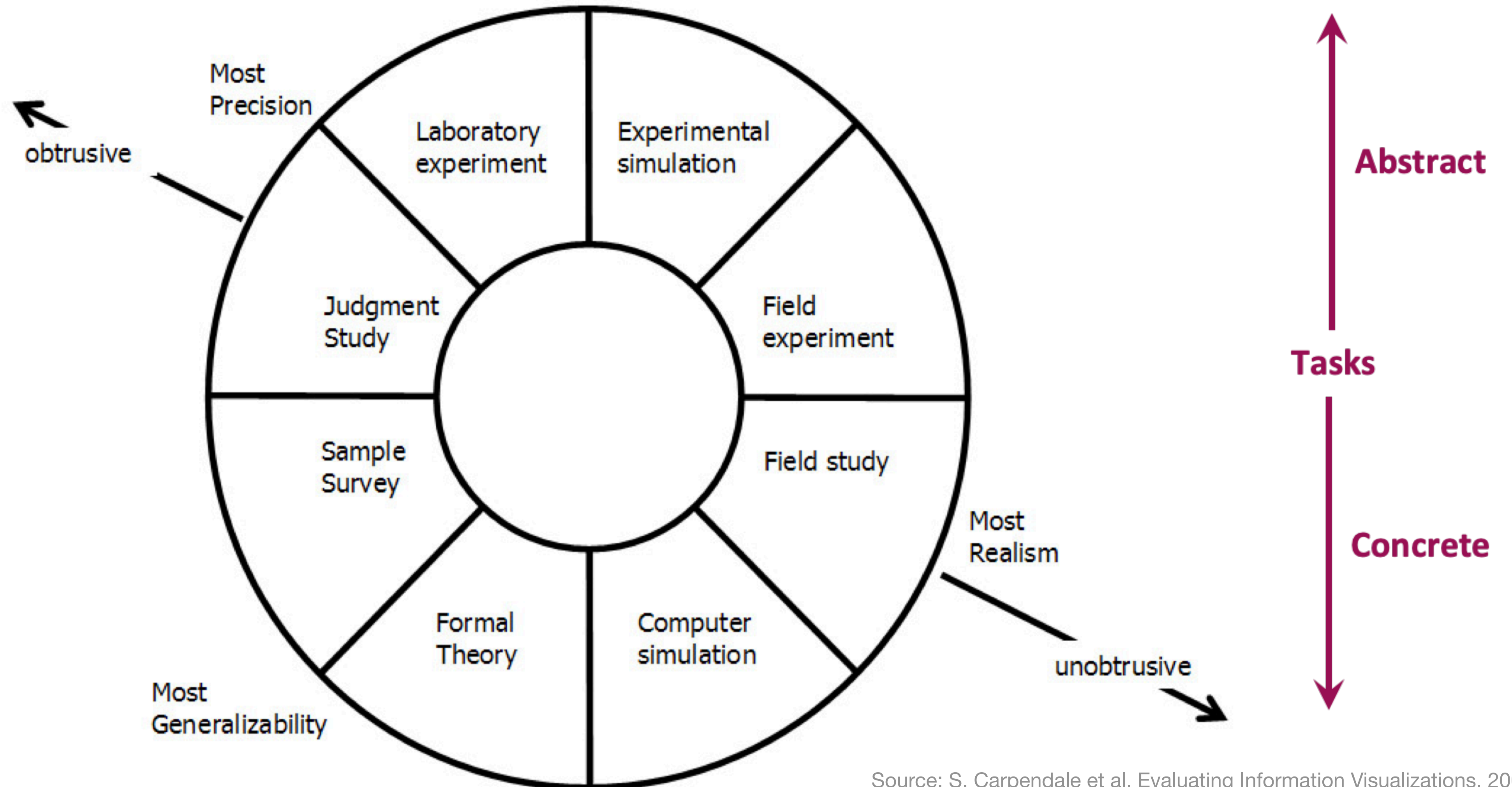
The aim

Understanding the tool

- Algorithm/Technique performance — benchmarking

- Qualitative result inspection — expert evaluation, heuristics*

Understanding the processes (with users)
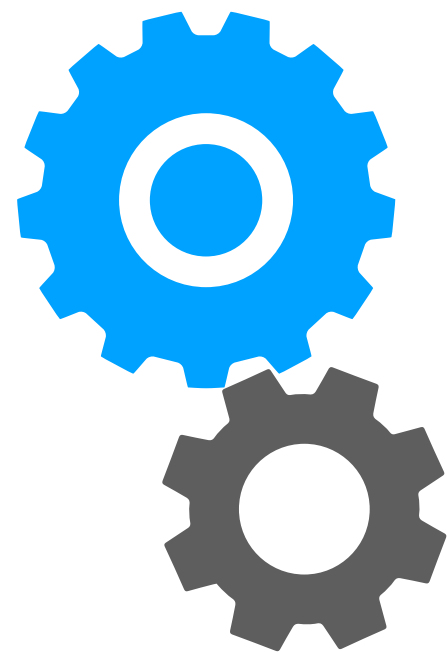
- Understanding environments and work practices — questionnaires, surveys, …*

- User evaluation methods - user experience and performance

- Reasoning — case studies

*later on, talk by Simone Kriglstein

Benchmarking

- Quantitative

- Performance comparison of a (novel) algorithm or technique

- Computation time, rendering speed (fps), memory footprint, …

- The importance of test datasets and their availability

- Reported using descriptive statistics

Fig. 1. Visualizations of 2D data (as pseudocolored height fields) and 3D data (volume rendered) used in our experiments.

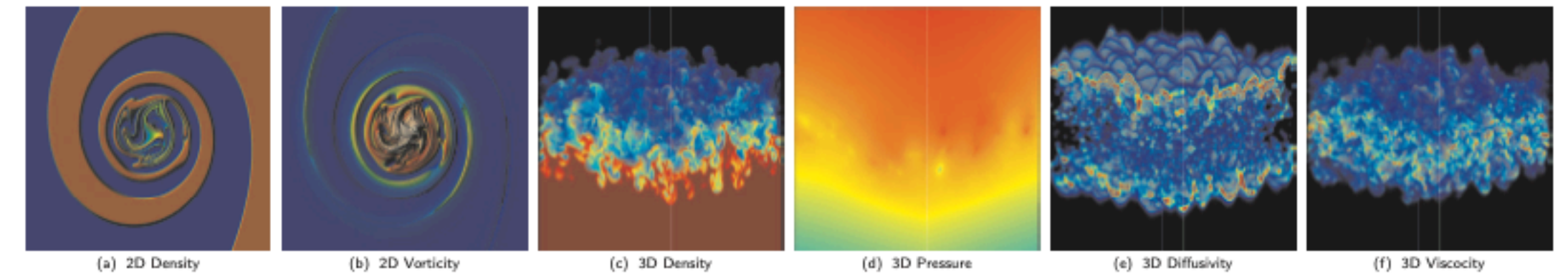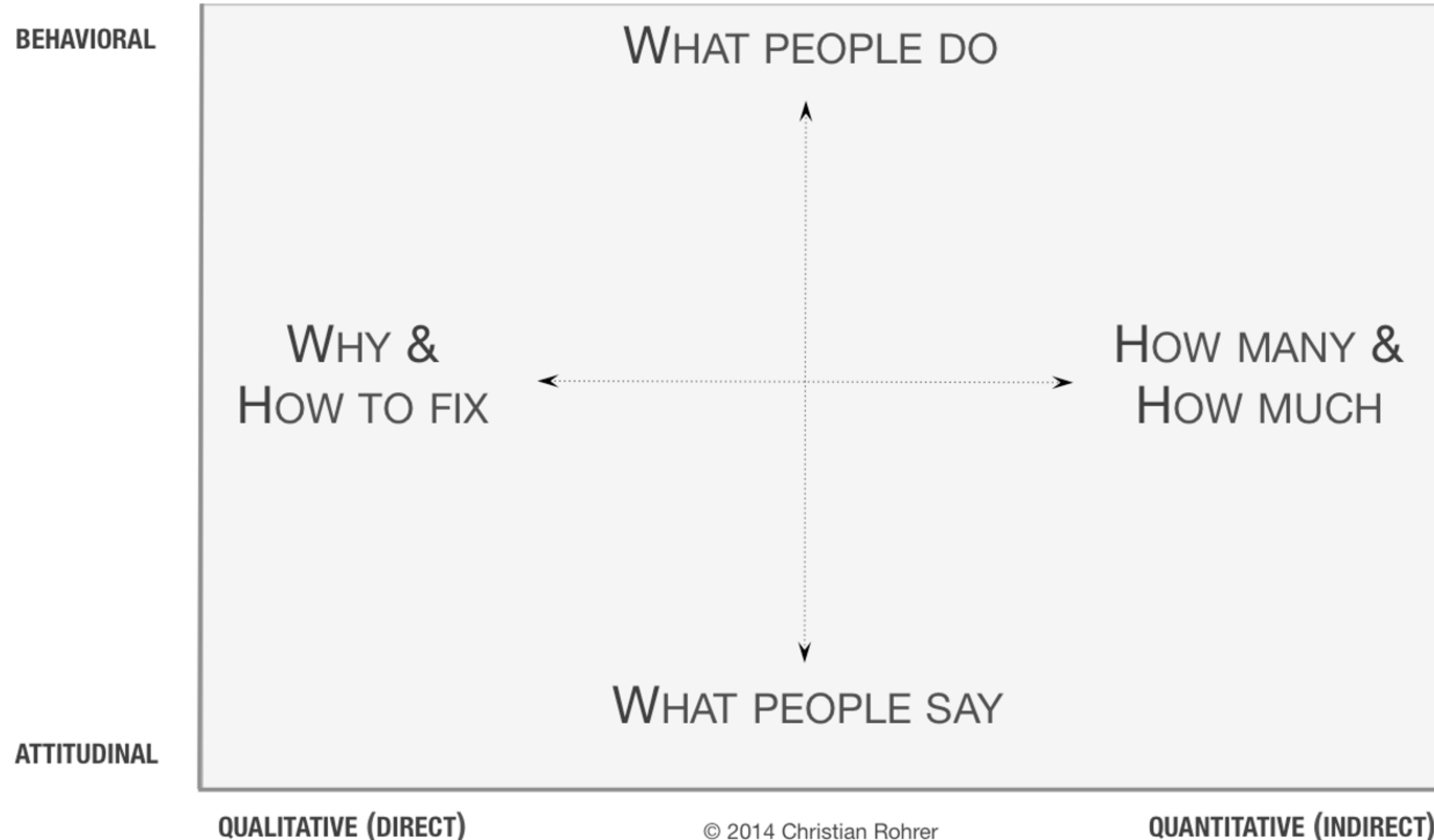|	data set							compressed size (MB) and compression time (seconds)									
name	unique (%)	entropy (bits)	range (bits)	min	max	size (MB)	time (sec)	zlib		[RKB2006]		[EFF2000]		[ILS2005]		new scheme	
m2d density	3.89	3.49	21.83	8.7E−01	1.2E+00	19.6	0.71	1.6	0.86	4.3	**0.49**	4.4	0.56	**1.3**	1.08	1.3	0.56
m2d vorticity	99.20	22.25	31.05	-1.4E+02	2.5E+01	19.6	0.71	18.4	2.14	**11.8**	**1.21**	15.5	1.29	12.9	2.22	13.8	1.49
m3d density	7.67	5.16	23.60	1.0E+00	3.0E+00	364.5	12.81	50.4	17.55	100.5	9.06	96.3	**8.48**	35.7	19.03	**35.5**	9.25
m3d pressure	27.29	23.91	31.06	-3.7E+00	2.3E+03	364.5	12.80	229.2	99.76	95.6	9.31	87.9	**8.87**	**40.1**	18.79	40.4	9.96
m3d diffusivity	36.87	23.19	30.02	0.0E+00	6.8E+00	364.5	12.68	297.6	42.90	250.8	19.09	239.3	**15.02**	**198.8**	31.92	203.0	18.47
m3d viscocity	50.07	24.86	28.59	8.6E−15	2.9E+00	364.5	12.62	314.0	46.09	249.4	18.95	246.1	**14.68**	209.2	32.66	**207.5**	19.45
h3d temp	65.70	23.54	31.56	-7.7E+01	1.0E+35	95.4	3.77	75.8	14.56	59.3	4.64	53.0	**4.27**	**44.1**	8.04	44.1	5.06
h3d pressure	81.82	24.13	31.58	-3.4E+03	1.0E+35	95.4	3.78	82.3	12.00	64.3	5.14	52.9	**4.87**	**45.0**	7.78	45.2	5.34
h3d x velocity	84.18	24.18	31.55	-5.3E+01	1.0E+35	95.4	3.89	86.1	11.27	67.4	6.22	63.3	**4.59**	**54.5**	8.86	55.4	5.44
h3d y velocity	84.32	24.18	31.55	-4.6E+01	1.0E+35	95.4	3.83	84.5	11.42	67.1	5.74	62.3	**5.04**	**53.5**	8.64	53.8	5.53
h3d z velocity	86.82	24.24	31.54	-3.2E+01	1.0E+35	95.4	3.87	88.4	10.76	85.6	8.50	76.9	**5.29**	**68.9**	9.83	69.1	6.65
M3d density	40.14	18.84	52.59	1.0E+00	3.0E+00	288.0	11.28	136.8	41.91	160.3	11.63	121.6	**10.94**	-		**105.2**	11.63
M3d pressure	100.00	25.17	63.00	-2.2E+00	2.2E+00	288.0	11.20	272.6	35.18	237.3	**14.91**	225.1	16.59	-		**208.4**	17.20
M3d x velocity	100.00	25.17	63.00	-2.2E+00	2.3E+00	288.0	10.83	275.6	32.30	230.4	**14.73**	215.1	15.91	-		**197.7**	16.84
M3d y velocity	100.00	25.17	63.00	-2.1E+00	2.3E+00	288.0	10.54	275.1	32.19	223.1	**14.27**	215.2	15.16	-		**197.7**	16.65
M3d z velocity	100.00	25.17	63.00	-5.2E+00	9.0E+00	288.0	10.32	275.5	32.62	226.6	**14.74**	213.7	16.05	-		**196.8**	16.14
atom x position	61.10	23.82	31.01	-4.8E−02	4.6E+02	107.7	7.07	84.3	21.18	76.0	7.88	78.8	**7.61**	67.3	12.88	68.6	9.07
atom y position	45.90	23.32	26.99	3.7E−02	2.1E+03	107.7	7.08	65.9	30.76	60.4	6.97	56.4	**6.31**	47.0	10.49	**46.9**	7.73
atom z position	61.68	23.84	27.48	9.1E−05	4.6E+02	107.7	7.46	94.6	19.86	82.6	9.00	86.1	**8.25**	**75.7**	13.80	78.2	9.93
atom y velocity	64.65	23.87	30.96	-1.5E−01	1.4E−01	107.7	7.30	95.7	19.88	93.8	10.07	99.1	**9.65**	84.3	14.93	87.6	9.92
atom temp	64.91	23.94	27.41	3.0E−03	7.1E+03	107.7	6.69	95.7	19.76	91.6	10.27	95.9	**8.34**	84.6	15.02	84.6	10.31
atom energy	3.45	18.57	21.79	-3.6E+00	-2.7E+00	107.7	7.15	77.9	38.59	74.1	7.98	71.2	**7.01**	60.8	12.66	**60.5**	8.30
lucy	61.39	24.38	31.09	-6.1E+02	1.2E+03	160.5	-	137.8	-	99.5	-	90.0	-	**73.6**	-	77.8	-
david$_{1mm}$	25.23	17.08	31.11	-4.4E+03	1.8E+03	322.5	-	144.9	-	155.7	-	163.4	-	**108.6**	-	131.9	-
torso	84.72	18.48	31.08	-2.7E+02	5.8E+02	1.9	-	1.7	-	1.5	-	1.5	-	**1.3**	-	1.3	-
rbl	71.90	20.14	25.99	1.5E+00	3.6E+02	8.4	-	7.1	-	5.8	-	5.6	-	**4.7**	-	4.8	-

Table 1. Compression results for the Miranda (m2d, m3d, M3d) and hurricane (h3d) structured grids, the atom point set, the lucy and david triangle meshes, and the torso and rbl tetrahedral meshes. All data but M3d is represented in single precision. The [ILS2005] scheme operates on single precision only, hence the missing values. For the meshes we report only the compressed size of vertex coordinates; timings are dominated by connectivity coding, and are hence excluded. The range measures (the logarithm of) the number of floating-point values between min and max. Note that the first-order entropy is limited by the number of samples in a data set.

Source: P. Lindstrom and M. Isenburg, "Fast and Efficient Compression of Floating-Point Data," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1245-1250, Sept.-Oct. 2006.

User Experience and Performance

BEHAVIORAL

WHAT PEOPLE DO

WHY &
HOW TO FIX

HOW MANY &
HOW MUCH

WHAT PEOPLE SAY

ATTITUDINAL

QUALITATIVE (DIRECT)

© 2014 Christian Rohrer

QUANTITATIVE (INDIRECT)

Evaluating User Performance

- Gathering **evidence**, not proving things (mathematicians do)

- Focus on users' effectiveness while using the system

- Used to obtain quantitative data about test participants' performance when they perform the tasks during usability tests

- Results are compared against *baseline*, *automatic* or *competing techniques*

- **Hypotheses testing**

- Reported using both inferential and descriptive statistics

Hypothesis
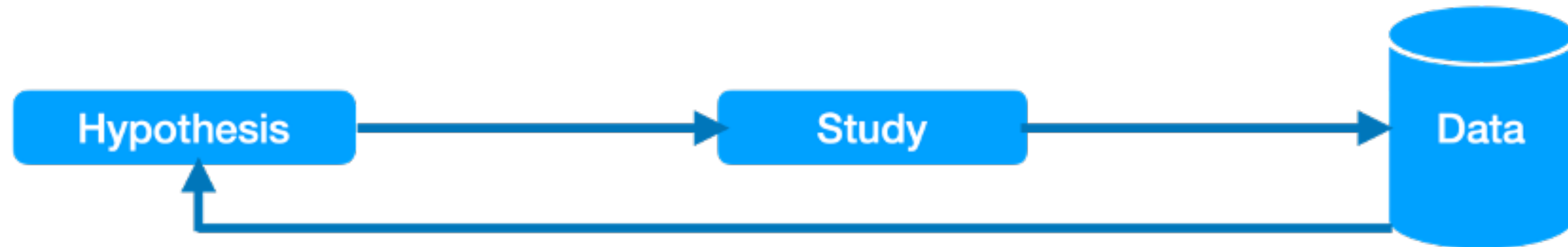
- A precise problem statement that can be directly tested through an empirical investigation

- **More focused statement** that can be **examined by an experiment**

- The goal is to find statistical evidence to **confirm** or **reject** null hypothesis in a reliable fashion

- *Null hypothesis (H_0)*: There is <u>no difference</u> between the OLD and NEW methods.

- *Alternative hypothesis (H_1)*: the NEW method will <u>perform better</u> than the OLD method.

Tasks

- **Low-level tasks**

  - *Example: Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.*

  - Usually described by some task taxonomy (check <u>Task Taxonomy for Graph Visualization</u>)

- **High-level tasks**

  - *Example: Due to errors in the data, several nodes may represent the same entity. For example, the co-authorship graphs often have duplicate author nodes. Identify whether two or more nodes represent the same person.*
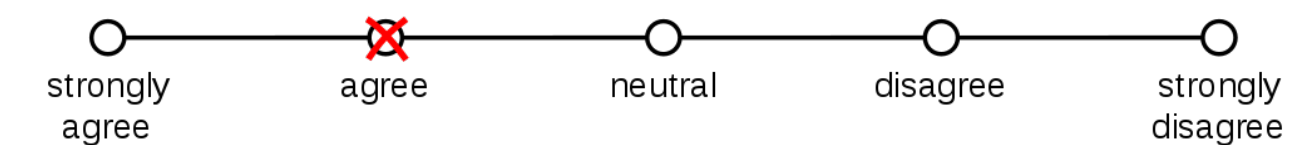
- **Explorative tasks**

  - More general assignment, tasks implicitly include sub-tasks and higher cognitive load (creativity)

  - *E.g., replicability study: Re-create this example visualization using our (shiny) tool.*

But what if you want to…

… **evaluate users' experience** in terms of user satisfaction, system usability, learnability or others?
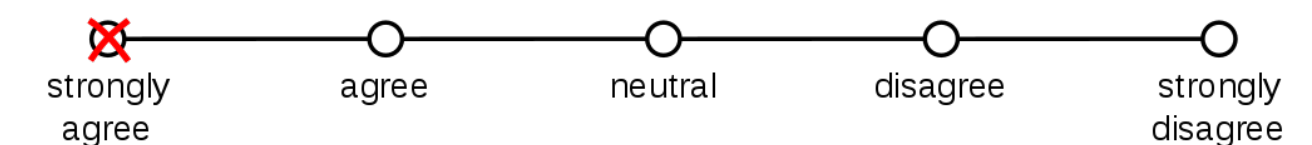
Likert Scales

- Statement soliciting level of agreement

- Gradations between responses are (more or less) equal

- Ordinal data => Be careful with averaging (median is often better)

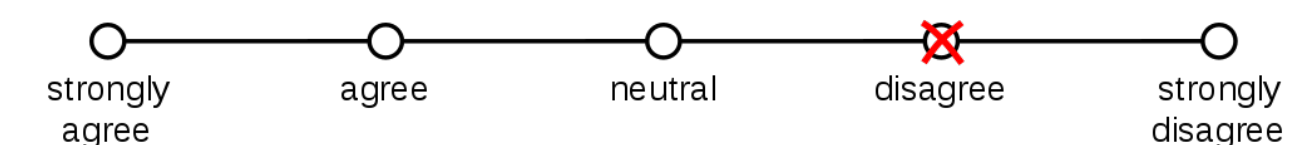- Even vs. odd number of options

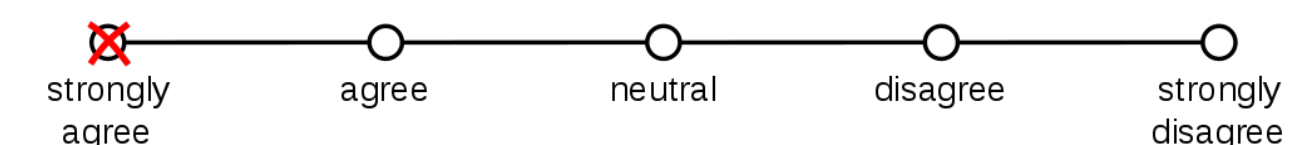**Website User Survey**

1. The website has a user friendly interface.

strongly agree — agree — neutral — disagree — strongly disagree

2. The website is easy to navigate.

strongly agree — agree — neutral — disagree — strongly disagree

3. The website's pages generally have good images.

strongly agree — agree — neutral — disagree — strongly disagree

4. The website allows users to upload pictures easily.

strongly agree — agree — neutral — disagree — strongly disagree

5. The website has a pleasing color scheme.

strongly agree — agree — neutral — disagree — strongly disagree

Source: https://en.wikipedia.org/wiki/Likert_scale
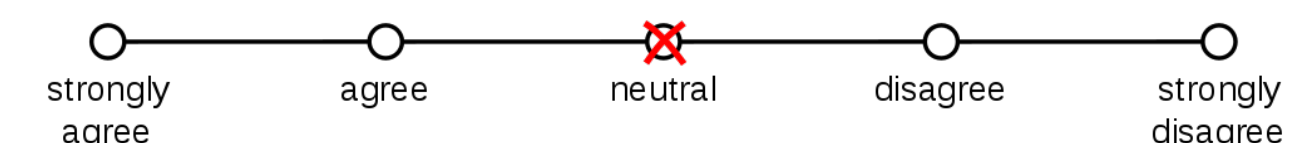
Standardized Usability Questionnaires

Questionnaires designed for the assessment of perceived usability, typically with a specific set of questions presented in a specified order using a specified format with specific rules for producing scores based on the answers of respondents.

J. Sauro, J. R. Lewis, Quantifying the User Experience, 2016

- Post-task: SEQ, SMEQ, ER, NASA-TLX, …

- Post-study: SUS, UMUX, SUMI, PSSUQ, …

- Benefits:
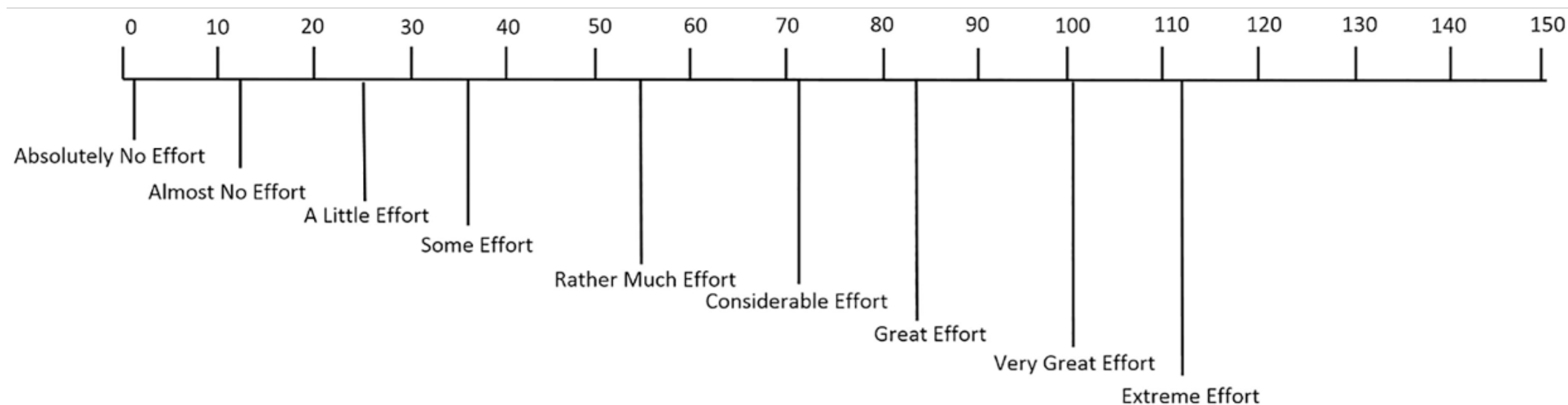 - objectivity, replicability, quantification, economy, generalization, communication

Post-task: Examples

- Single Ease Question (SEQ)

Overall, this task was :

Very Easy ○ ○ ○ ○ ○ ○ ○ Very Difficult
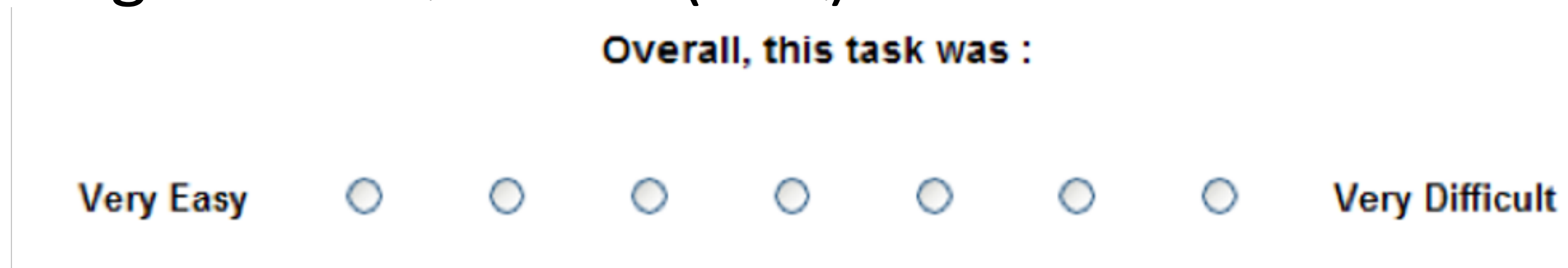
Source: Sauro J. and Dumas J. S. Comparison of Three One-Question, Post-Task Usability Questionnaires.

- Subjective Mental Effort Question (SMEQ)

| 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 | 130 | 140 | 150 |

Absolutely No Effort

Almost No Effort

A Little Effort

Some Effort

Rather Much Effort

Considerable Effort

Great Effort

Very Great Effort

Extreme Effort

Source: So, et al. Subjective mental effort questionnaire.

Post-study: Examples

System Usability Scale (SUS)

Questions:

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.

Strongly Disagree 1	2	3	4	Strongly Agree 5
○	○	○	○	○

Usability Metric for User Experience (UMUX) and UMUX-Lite

1. [This system's] capabilities meet my requirements.
   1 2 3 4 5 6 7
   Strongly Disagree — Strongly Agree

2. Using [this system] is a frustrating experience.
   1 2 3 4 5 6 7
   Strongly Disagree — Strongly Agree

3. [This system] is easy to use.
   1 2 3 4 5 6 7
   Strongly Disagree — Strongly Agree

4. I have to spend too much time correcting things with [this system].
   1 2 3 4 5 6 7
   Strongly Disagree — Strongly Agree

Source: Finstad K. The Usability Metric for User Experience

UMUX-Lite – same 7-point likert scale, only two questions

- *This system's* capabilities meet my requirements.
- *This system* is easy to use.

Source: Sauro J. Measuring Usability: From the SUS to the UMUX-Lite. MeasuringU

Case Studies

- "A detailed reporting about a small number of individuals working on their own problems in their normal environment"★

  - *Case study != Usage scenario*

- **Four key aspects**:

  - in-depth investigation of a small number of cases (often up to 5)

  - examination in context (how the participant use the tool in his/her natural setting, not a lab-study)

  - multiple data sources

  - emphasis on qualitative data an analysis (results in validity and reliability concerns)

- Summarized feedback (feature requests, opinion of participants on the tool functions and limits and its applicability in their work)

★ B. Shneiderman and C. Plaisant. 2006. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. BELIV '06
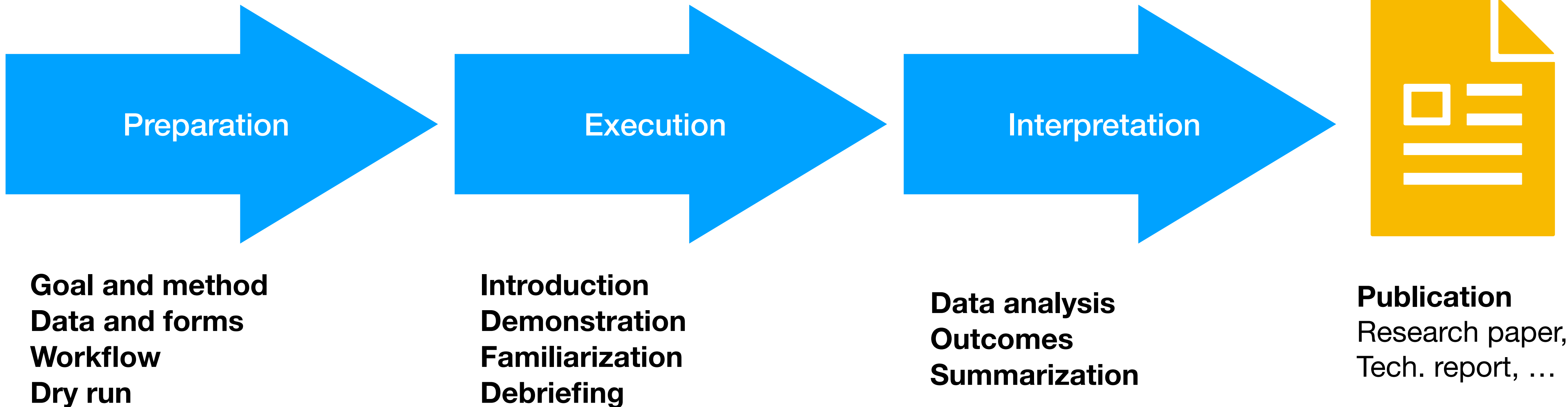
Goals of Case Studies

- Exploration — understanding novel problems or situations

- Explanation — developing models that can be used to understand a context of use

- Description — documenting a system, technology use (in context) or the process

- Demonstration — showing how the tool was successfully used

Case Study Design

- There are four main components of a case study design:

  - (research) questions — What are you interested in?

  - hypotheses or propositions — What you expect to find?

  - units of analysis — What are you focusing on?

  - data analysis plan — Which data we collect and how to process them?

Evaluation Workflow

Preparation

Goal and method
Data and forms
Workflow
Dry run

Execution

Introduction
Demonstration
Familiarization
Debriefing

Interpretation

Data analysis
Outcomes
Summarization

Publication
Research paper,
Tech. report, …

Preparation

- Set the goal, then choose the method (with/without participants)

- Prepare data and related documents, datasets, consent forms, questionnaires (pre-, post-)

- Always do the pilot or dry run => identification of unexpected problems

- Make a checklist — always follow the same steps

- Get the participants

Participants

- People ***participating*** in the experiment (don't use ~~subjects~~)

- How many?

  - Short answer: use the same number as used in similar research

  - Too many: unnecessary work

  - Too few: fail to get statistically significant results => paper reject

Execution

- Follow the checklist

- Do not change experiment design or conditions after starting it

- Use different dataset for practice trials and main experiment

- With participants:

  - Get consent first, debrief participants afterwards

  - Record: audio/video, mouse traces, make notes

Evaluator's Toolbox

Within vs. Between Subjects

Within-subjects design
The same participant tests all conditions corresponding to a variable.

Between-subjects design
Different participants are assigned to different conditions corresponding to a variable.

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

Advantages & Limitations

+ Smaller sample size
+ Effective isolation of individual differences
+ More powerful tests
- Hard to control learning effect
- Large impact of fatigue

+ Avoids learning effect
+ Better control of confounding factors (e.g., fatigue)
- Requires more people
- Harder to get statistically significant results
- Large impact of individual differences

Counterbalancing

- The effect of one condition "carries over" into the next one

- Common in within-subjects designs, e.g., learning effect

- Counterbalancing = compensation of carryover effects

- The order of tasks or datasets used in the experiment

- (Pseudo)Randomized order — one for each participant

- **Latin Square**: n×n array filled with n different symbols, occurring **exactly once** in **each row** and **column** (=Sudoku).

  - Problem with the odd ones (from order 3)

  - Solution (for even-ordered only) is **Balanced Latin Square**

    - Online generator
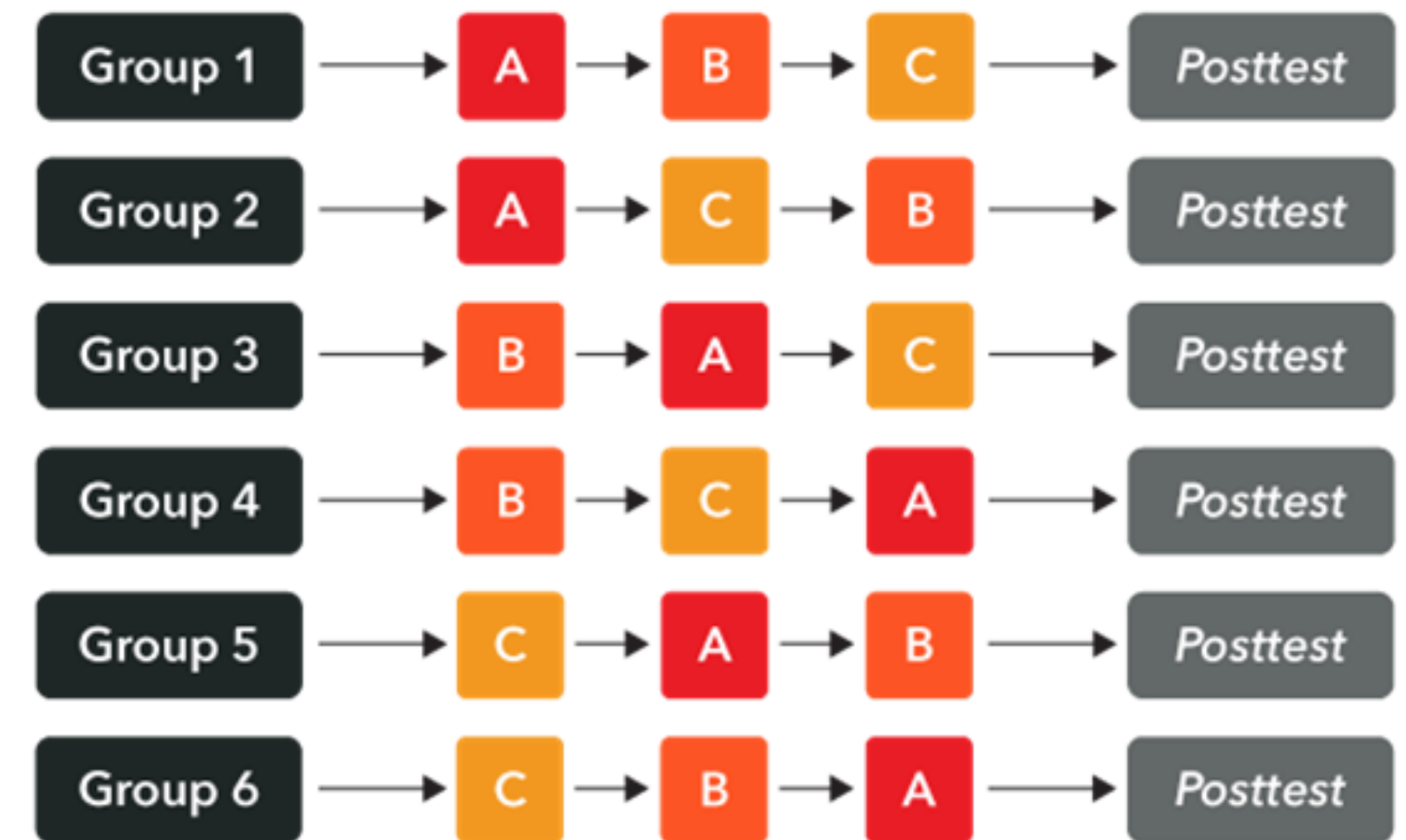
Consent Form

- Who you are

- What you are asking the participants to do

- What kind of data you will be collecting and how it will be used

- What rights the participant has

- If they will be compensated

- The participant must explicitly say "yes" to the consent form

SIMON FRASER UNIVERSITY

INFORMED CONSENT BY SUBJECTS TO PARTICIPATE IN EVALUATION OF AN INTERACTIVE COMPUTER SYSTEM FOR DATA VISUALIZATION

The University and those conducting this project subscribe to the ethical conduct of research and to the protection at all times of the interests, comfort, and safety of subjects. This form and the information it contains are given to you for your own protection and full understanding of the procedures. Your signature on this form will signify that you have received a document which describes the procedures, possible risks, and benefits of this research project, that you have received an adequate opportunity to consider the information in the document, and that you voluntarily agree to participate in the project.

Knowledge of your identity is not required. You will not be required to write your name or any other identifying information on the research questionnaires. An audio recording of your voice and a video recording of the computer screen only will be made during the session. The video and audio recordings of the session will be reviewed only by the Principal Investigator. All research materials will be held confidential by the Principal Investigator and kept in a secure location. These research materials will be destroyed after the completion of the study.

Having been asked by Daryl H. Hepting of the School of Computing Science of Simon Fraser University to participate in a research project study, I have read the procedures specified in the accompanying information sheet. I understand the procedures to be used in this study and the personal risks and benefits to me in taking part. I understand that I may withdraw my participation in this study at any time.

I understand that my decision to participate in this study, and my subsequent involvement in it, will have absolutely no bearing on any other dealings I have with Mr. Hepting. This includes, but is not limited to, the case that I am a student in the CMPT 361 course taught by Mr. Hepting, offered at SFU during the 99-2 semester.

I understand that I may register any complaint I might have about the study with the researcher named above or with Dr. Jim Delgrande, Director, School of Computing Science of Simon Fraser University, Burnaby, BC, V5A 1S6, telephone 604-291-4277.

I may obtain copies of the results of this study, upon its completion, by contacting Mr. Daryl Hepting, in care of the School of Computing Science at Simon Fraser University.

I understand that my supervisor or employer may require me to obtain his or her permission prior to my participation in a study such as this.

I agree to participate by completing: a pre-task questionnaire; a training session on the prototype software system; a task with the prototype software system; and a post-task questionnaire. I understand that these activities will require approximately one hour at a time scheduled with Mr. Hepting. I understand that the experiment will be conducted in Room 9836 in the Applied Science Building of Simon Fraser University.

NAME (please type or print legibly): _____

ADDRESS: _____
_____

SIGNATURE: _____

WITNESS: _____

DATE: _____

A COPY OF THIS SIGNED **CONSENT FORM** AND A **SUBJECT FEEDBACK FORM** WILL BE PROVIDED TO YOU AT YOUR EXPERIMENT SESSION.

Source: D. Hepting. "A New Paradigm for Exploration in Computer-Aided Visualization". Dissertation thesis. Simon Fraser University. 1999.

Color Perception Test

- Shinobu Ishihara, 1917

- Ishihara plates

- Diagnostic test for color perception deficiencies

- 38 plates (full set)

- Variants with 10, 12 or 24
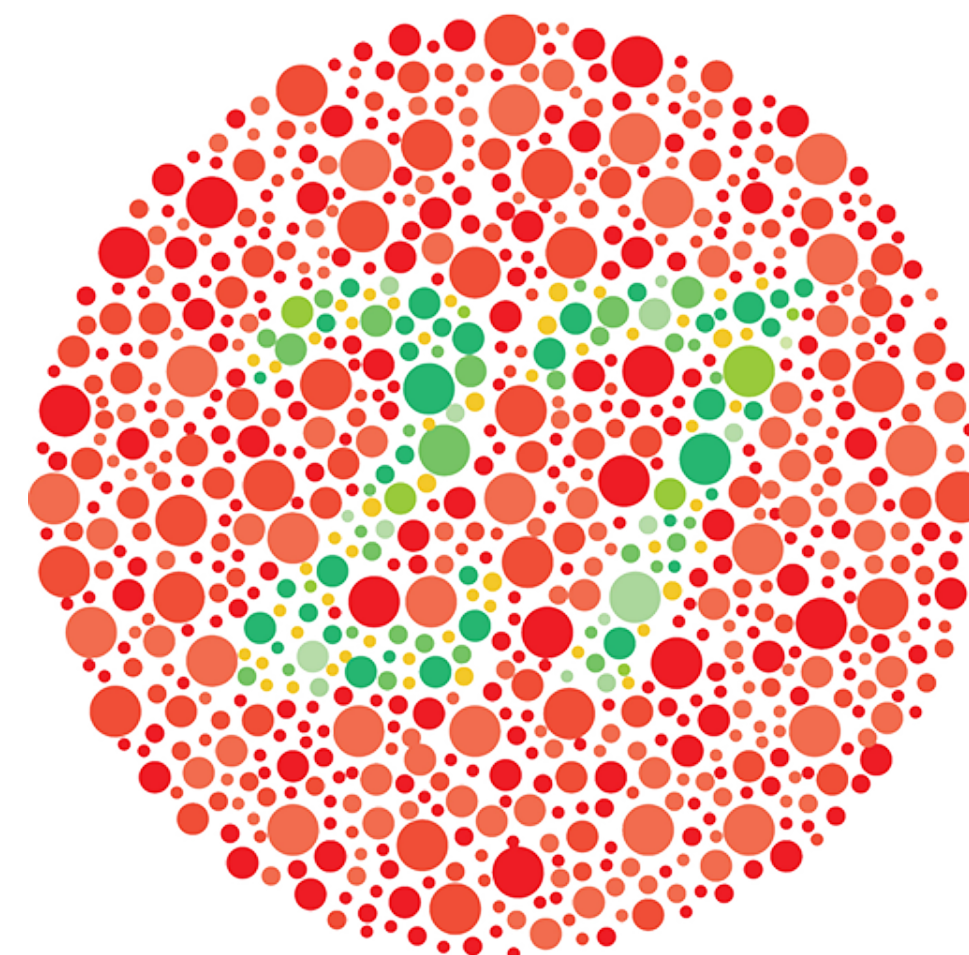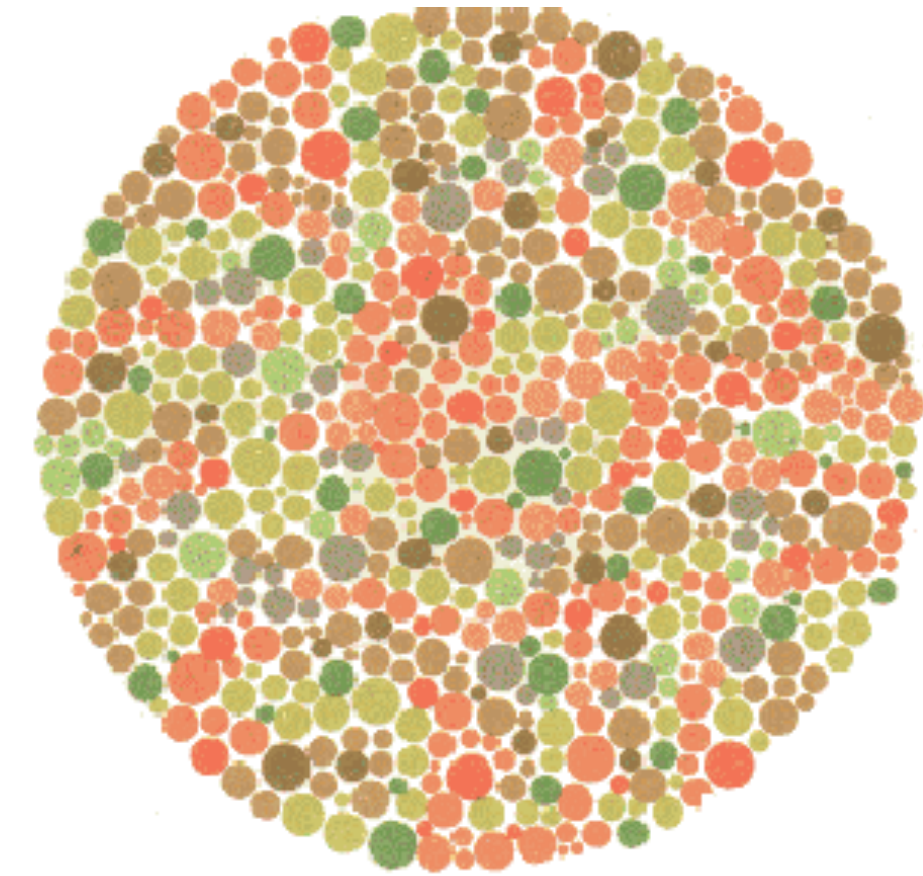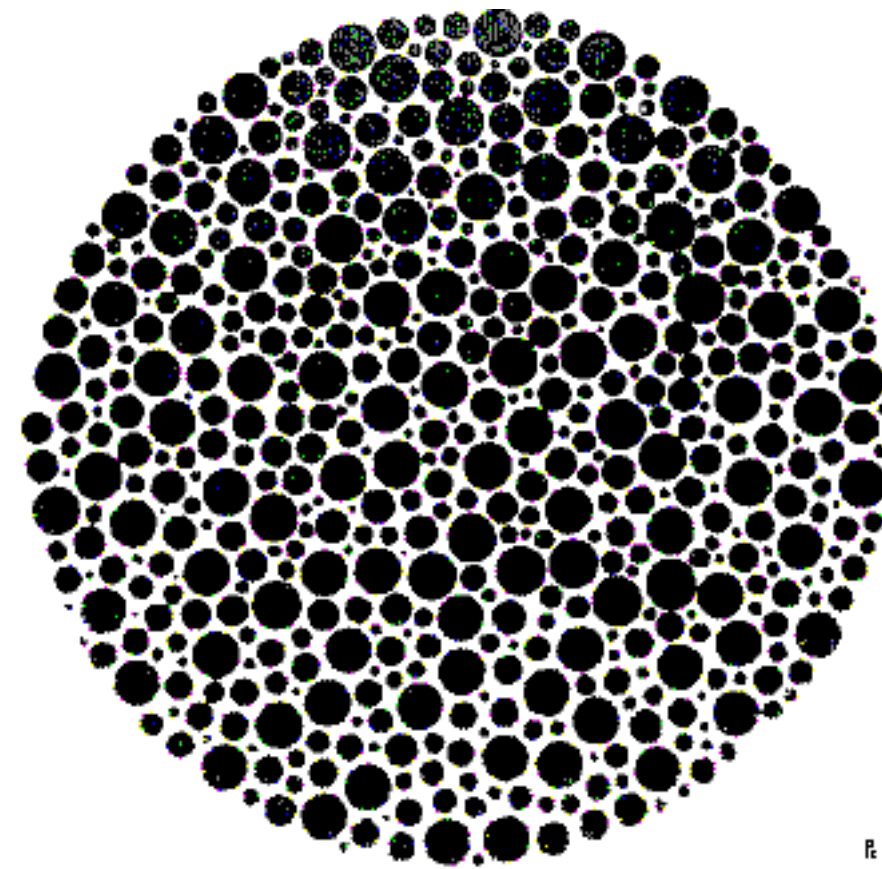
Statistical Evaluation

Descriptive statistics
- Summary of a data set characteristics
- Mean, median, mode, standard deviation, spread, central tendency, …

Inferential statistics
- Infers properties of a population based on a sample data
- Testing hypotheses and deriving estimates
- Parametric (t-test, ANOVA) and non-parametric tests

- Concrete methods are out of scope of this talk
  - Further reading: Statistical Methods for HCI Research

Take away…

- In SciVis, InfoVis, VAST, we mostly do:

  - algorithm benchmarking,
    user performance (quantitative)

  - case studies, qualitative inspection,
    user experience (qualitative)

IEEE Vis/SciVis: total numbers & percent of evaluation scenarios
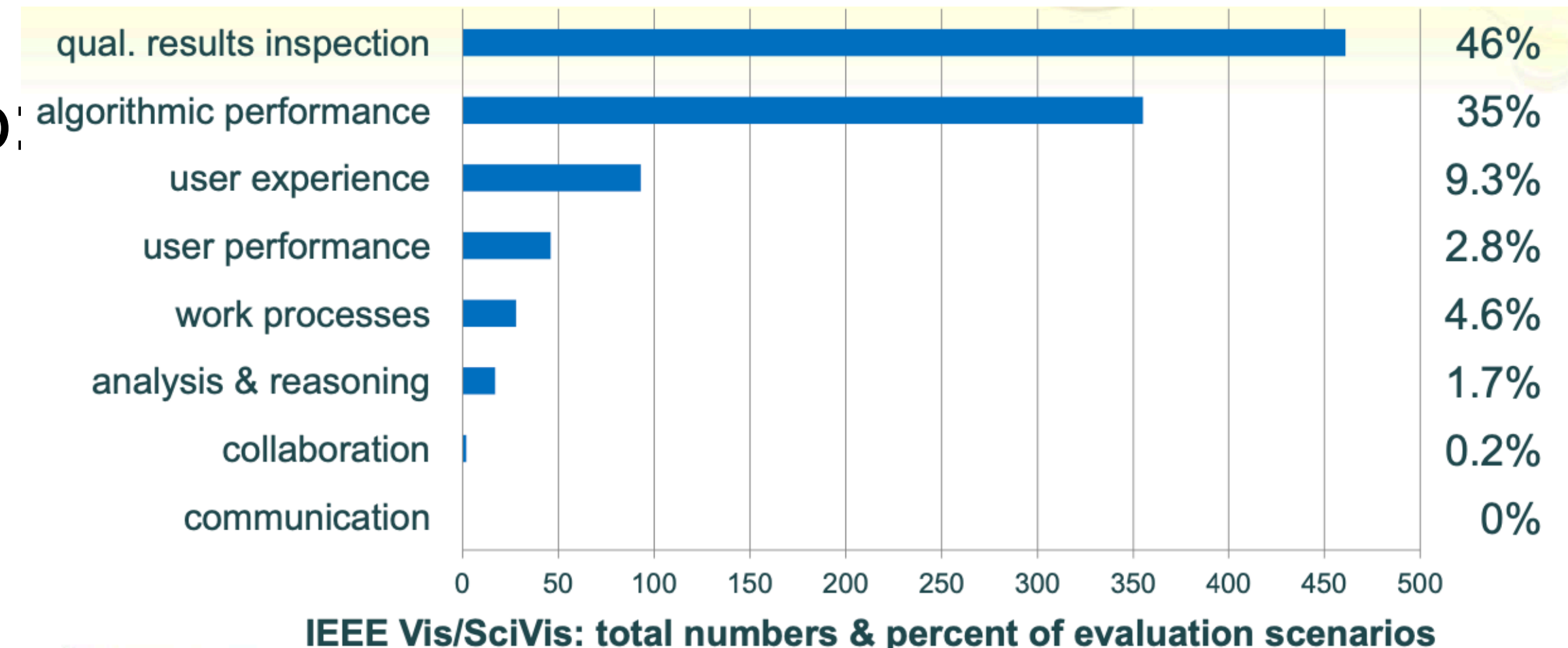
- Contribution of **real users** is invaluable but also painful (involve them ASAP)

- Use **methodologies** and **best practices** from the field (learn from papers)

- Evaluation methods are similar (same) to those in HCI

Chart source: http://tobias.isenberg.cc/personal/papers/Isenberg_2013_SRP_Slides.pdf

References

1. J. Lazar, J. H. Feng, and H. Hochheiser. 2010. Research Methods in Human-Computer Interaction. Wiley Publishing.

2. I. S. MacKenzie. 2013. Human-Computer Interaction: An Empirical Research Perspective (1st. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

3. J. Sauro and J. R. Lewis. 2016. "Quantifying the User Experience, Second Edition: Practical Statistics for User Research". 2nd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

4. J. W. Creswell, Ch. N. Poth, "Qualitative Inquiry and Research Design: Choosing Among Five Approaches", 4th Ed. SAGE Publishing, 2018.

5. T. Isenberg et al., "A Systematic Review on the Practice of Evaluating Visualization," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2818-2827, Dec. 2013.

6. B. Shneiderman, et al. "Designing the User Interface: Strategies for Effective Human-Computer Interaction". 6th ed. Pearson Global Edition. 2018.