CSC2537 / STA2555 - INFORMATION VISUALIZATION

EVALUATION AND VALIDATION

Fanny CHEVALIER

UNIVERSITY OF TORONTO

This lecture is heavily inspired from slides by **Christopher Collins, Tobias Isenberg and John Stasko.**
Thanks to all of them for sharing their material.

FOREWORDS

Summary

- you know about **visual variables** and **data**
- you know about several **data representations**
- you know about **interaction** with visualizations

In this lecture you will

- learn about how to **evaluate visualizations**
- learn about **eight evaluation scenarios**
- learn about different **evaluation methods**

Evaluation:
Why, when, what, how?

WHY EVALUATE VISUALIZATIONS?

Creation of new techniques is very important but… it's also important to know that we're getting better

- want to learn what aspects of the visualization or systems "work"
- want to ensure that methods are improving
- want to ensure that technique actually helps people and isn't just "cool"
- NOT: because I need that section in my paper to get it accepted … *sigh*

WHY EVALUATE VISUALIZATIONS?

- find out if visualizations are **effective**
- find out if visualizations are **efficient**
- find out if visualizations are **useful**
- find out if interactive tools are **usable**
- **motivate** the creation of new visualizations
- learn about **tasks** and **representations** to support
- learn about **existing visualization tools** to improve or replace
- learn how people **reason** with visualization
- learn how people **communicate/collaborate** with visualizations

WHEN EVALUATE VISUALIZATIONS?

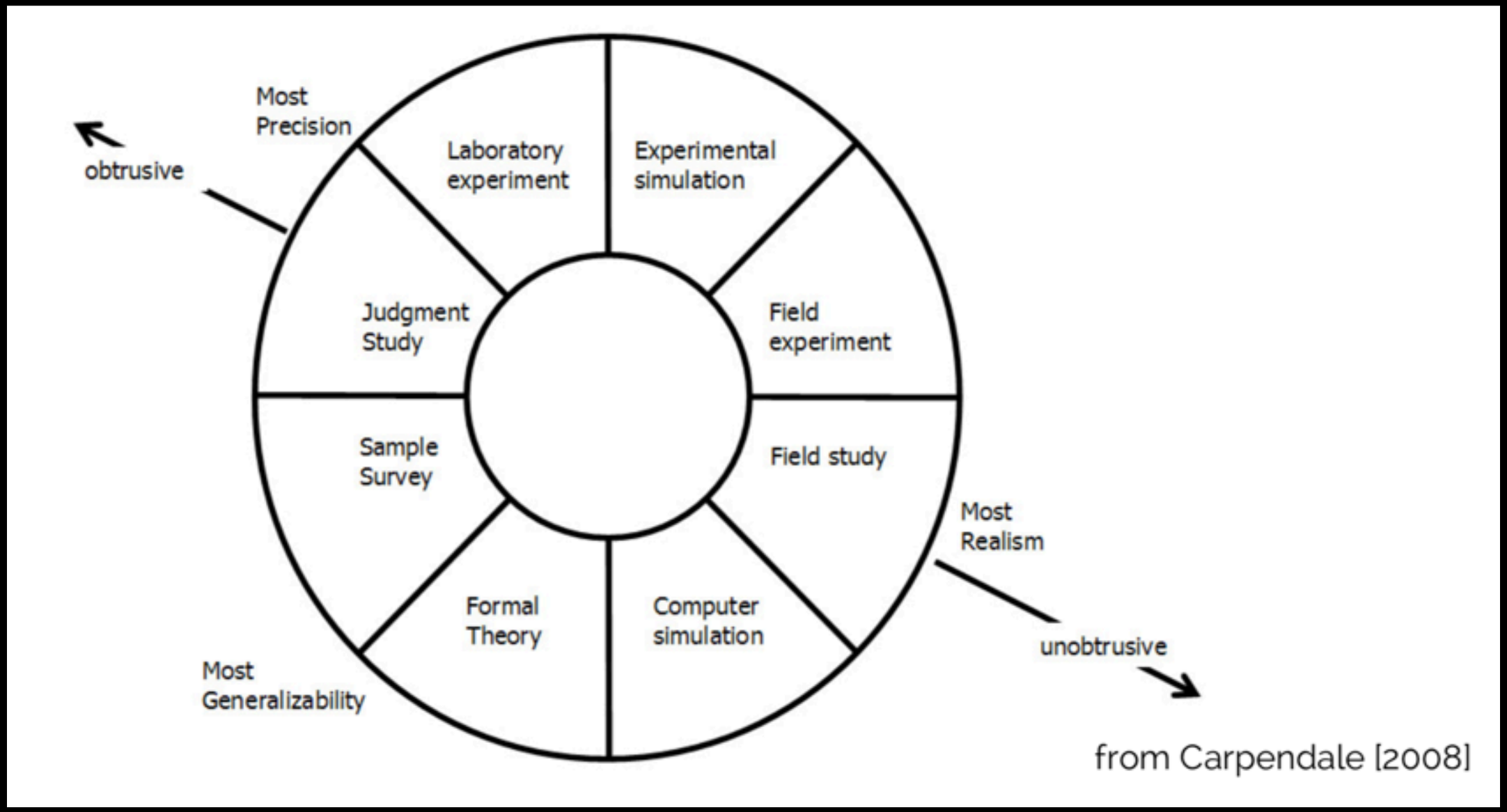
- when the visualization is completed/done…
- throughout the development…
- before even starting…

WHAT TO EVALUATE IN VISUALIZATION?

- does my visualization show what I want it to show?
- can people draw the right conclusions from it?
- can my visualization be used by the target audience?
- does my visualization support realistic tasks?
- does my visualization scale with (realistic) data sizes?
- what do people want/need to see in their data?
- how does my visualization compare to another?
- is my visualization precise enough for a given task?
- …

HOW TO EVALUATE VISUALIZATIONS?

- commonalities to HCI work
- methodologies borrowed from social sciences



from Carpendale [2008]

HOW TO EVALUATE VISUALIZATIONS?

We also care about

- motivations for new visualizations
- visualization implementation

Thus we need to

- talk to people
- observe people
- measure what people do
- measure the performance of tools
- critique visualizations
- compare visualizations

EVALUATION METHODS VS. EVALUATION GOALS

Different goals for evaluating visualizations

- effectiveness, efficiency, scalability, usability

The same evaluation method can be used to learn about different goals

- e.g. quantitative experiment to study both visualization effectiveness and its usability

Goals facilitate focus on what one wants to learn, and then to pick a suitable method

EVALUATION GOALS

EIGHT SCENARIO OF EVALUATION GOALS

1. Understanding environments and work practices
2. Visual data analysis and reasoning
3. Evaluating communication through visualization
4. Evaluating collaborative data analysis
5. User performance
6. User experience
7. Algorithm performance
8. Qualitative result inspection

Derived from coding published vis papers

1. UNDERSTANDING ENVIRONMENTS AND WORK PRACTICES

Derive an understanding of the work, analysis, or information processing practices by a given group of people

With or without software use

Example: evaluation with experts to understand their data analysis needs and requirements for developing a visualization

1. UNDERSTANDING ENVIRONMENTS AND WORK PRACTICES

Derive an understanding of the work, analysis, or information processing practices by a given group of people

With or without software use

Example: evaluation with experts to understand their data analysis needs and requirements for developing a visualization

2. VISUAL DATA ANALYSIS AND REASONING

Assess how a visualization tool supports analysis and reasoning about data

Asses how a visualization tool helps to derive relevant knowledge in a given domain

Examples:

- study experts using an existing too on their data
- analyze how experts can solve domain-specific questions with a new tool

3. EVALUATING COMMUNICATION THROUGH VISUALIZATION

Assess the communicative value of a visualization or visual representation

Taking into account goals such as teaching/learning, idea representation, or casual use

Examples:

- a study that assesses how well a visualization can communicate medical information to a patient

4. EVALUATING COLLABORATIVE DATA ANALYSIS

Understand to what extent a visualization tool supports collaborative data analysis by groups of people

Examples:

- an observational study that looks at how people share resources and analysis findings in a group

5. USER PERFORMANCE

Objectively measure how specific features affect the performance of people with a system

Often called “user study” (inadequate term)

Examples:

- controlled experiments using time and error rate

6. USER EXPERIENCE

Subjective feedback and opinions on a visualization (tool)

Can yield both qualitative descriptions and quantitative measurements

Examples:

- interviews of study participants
- Likert-scale questionnaires

7. ALGORITHM PERFORMANCE

Quantitatively study the performance or quality of visualization algorithms

Objective analysis of performance metrics

Does not involve actual end users or participants

Examples:

- measurement of rendering speed
- measurement of memory performance
- (quantitative) comparison to visualization ground truth

8. QUALITATIVE RESULT INSPECTION

Qualitative discussions and assessments of visualization results
Does not involve actual end users or participants
Typically done by creator of visualization
Asks the viewer of a resulting image to make an assessment for themselves, based on argumentation
Can also be done with external jury

Examples:

- discussion that essential aspects can be seen
- (qualitative) comparison to visualization ground truth

TWO GROUPS OF SCENARIOS

Evaluating the process of data analysis

- understanding environments and work practices
- visual data analysis and reasoning
- evaluating communication through visualization
- evaluating collaborative data analysis

Evaluating visualizations, tools, algorithms

- user performance
- user experience
- algorithm performance
- qualitative result inspection

Visualization Evaluation: 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

Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

Upstream

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*)

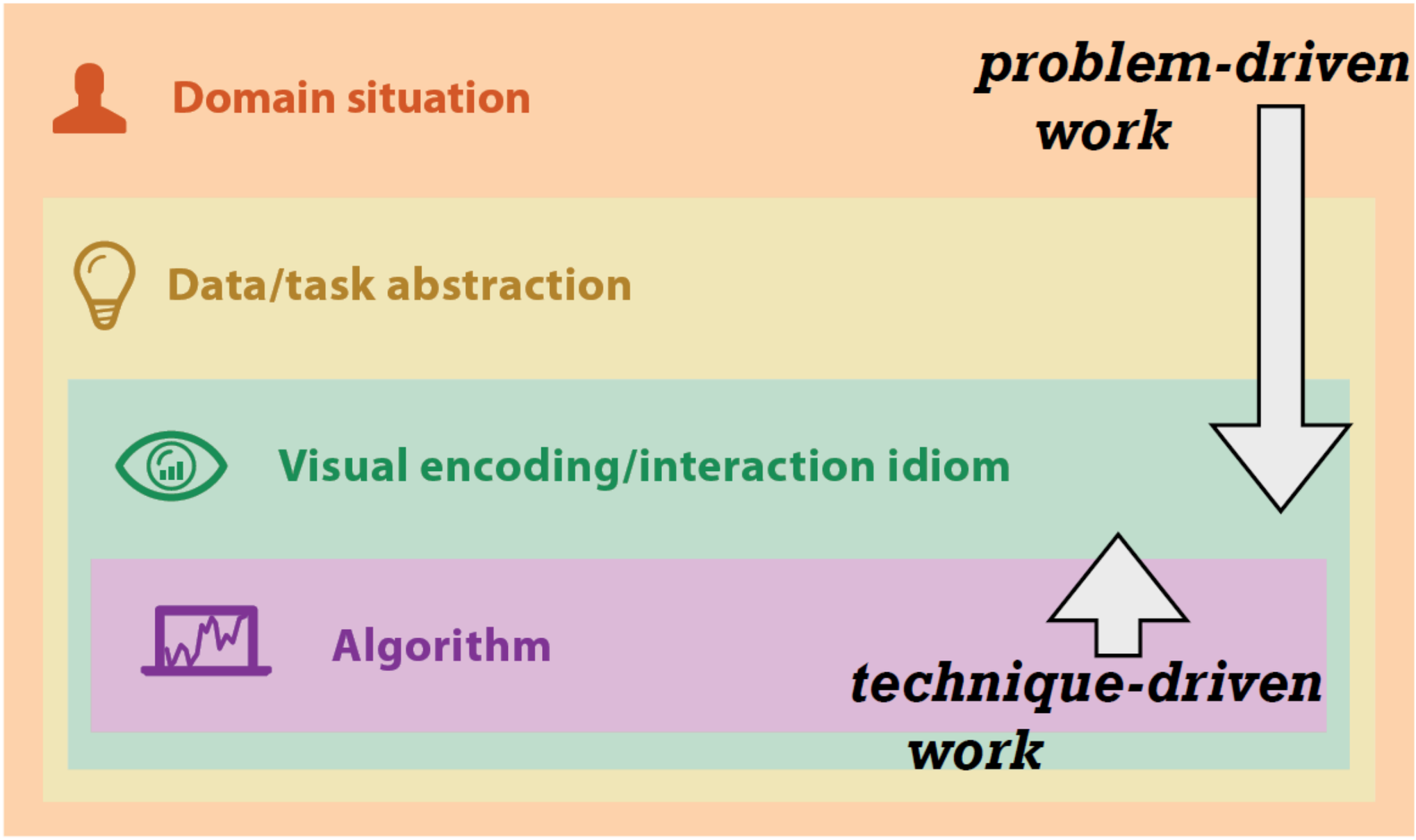
Measure adoption

Downstream

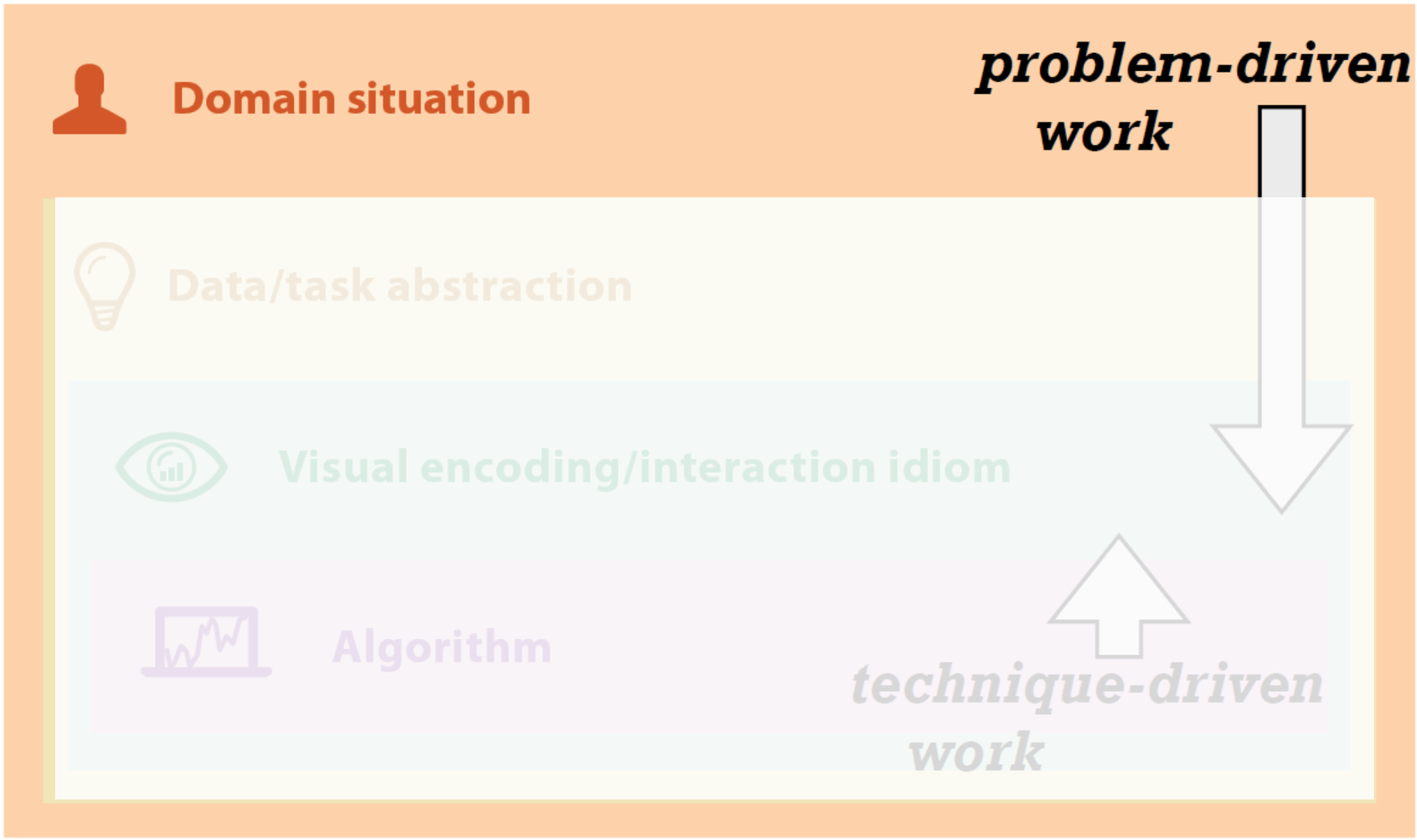
Creation of Vis

Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

DIRECTIONALITY AND SCOPE

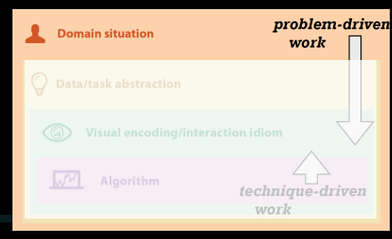


Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.



Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

DOMAIN SITUATION

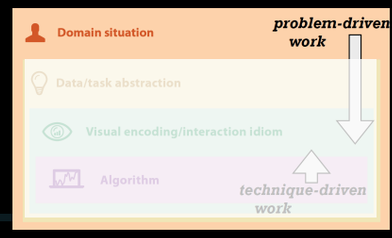


Understand:
target users, their questions, their data, their tasks

- Upstream technique: Observational study
- Downstream technique: Measure adoption, deployment study

- Pitfalls: make assumptions

OBSERVATIONAL STUDY (CSC318 / CSC428)

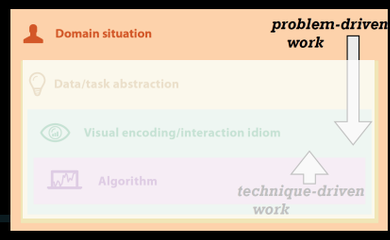


Also called a "grounding study"

Observe people doing their work ***now:***

- what data do they use?
- what tools do they use? do they make sketches?
- what questions do they have?
- where do they get frustrated?

EXAMPLE: MEDSTORY

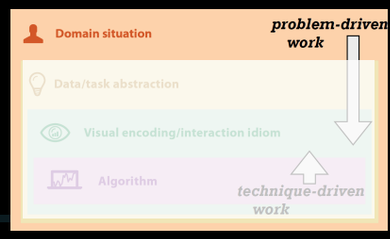


Our study involved 8 physicians, recruited on a volunteer basis from a range of clinical scenarios, including four different health institutions in Canada (spanning outpatient hospital clinics and private clinics), five areas of specialization (*i.e.*, 2 general practitioners (GPs), 3 genetic pediatricians, 1 developmental pediatrician, 1 gastroenterologist and 1 orthopedic surgeon) and spanning 2 to 30 years of medical practice after residence (12 years in average). We conducted semi-structured interviews with all participants, and a full-day observation of the clinical practice for five of them, amounting to a total of 25 hours (average of 4 hours each). We sought to answer the following research questions:

1. *Clinical workflow:* What is the role of text in clinical practice, and what activities does it facilitate?
2. *Structure:* What are the embodiment(s) and format(s) of clinical text?
3. *Tasks:* How do physicians use clinical text?
4. *Challenges:* What obstacles do physicians face with clinical text?

N. Sultanum et al. (2018) More text please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. CHI'18

EXAMPLE: MEDSTORY

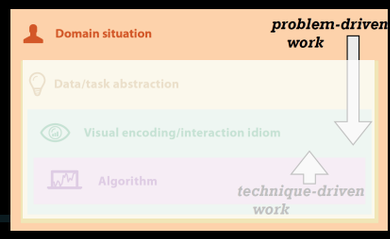


Our study involved 8 physicians, recruited on a volunteer basis from a range of clinical scenarios, including four different health institutions in Canada (spanning outpatient hospital clinics and private clinics), five areas of specialization (*i.e.*, 2 general practitioners (GPs), 3 genetic pediatricians, 1 developmental pediatrician, 1 gastroenterologist and 1 orthopedic surgeon) and spanning 2 to 30 years of medical practice after residence (12 years in average). We conducted semi-structured interviews with all participants, and a full-day observation of the clinical practice for five of them, amounting to a total of 25 hours (average of 4 hours each). We sought to answer the following research questions:

1. *Clinical workflow:* What is the role of text in clinical practice, and what activities does it facilitate?
2. *Structure:* What are the embodiment(s) and format(s) of clinical text?
3. *Tasks:* How do physicians use clinical text?
4. *Challenges:* What obstacles do physicians face with clinical text?

N. Sultanum et al. (2018) More text please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. CHI'18

EXAMPLE: MEDSTORY

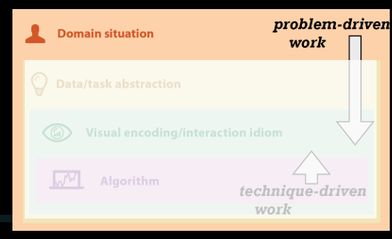


Our study involved 8 physicians, recruited on a volunteer basis from a range of clinical scenarios, including four different health institutions in Canada (spanning outpatient hospital clinics and private clinics), five areas of specialization (*i.e.*, 2 general practitioners (GPs), 3 genetic pediatricians, 1 developmental pediatrician, 1 gastroenterologist and 1 orthopedic surgeon) and spanning 2 to 30 years of medical practice after residence (12 years in average). We conducted semi-structured interviews with all participants, and a full-day observation of the clinical practice for five of them, amounting to a total of 25 hours (average of 4 hours each). We sought to answer the following research questions:

1. *Clinical workflow:* What is the role of text in clinical practice, and what activities does it facilitate?
2. *Structure:* What are the embodiment(s) and format(s) of clinical text?
3. *Tasks:* How do physicians use clinical text?
4. *Challenges:* What obstacles do physicians face with clinical text?

N. Sultanum et al. (2018) More text please! Understanding and Supporting the Use of Visualization for Clinical Text Overview. CHI'18

OBSERVATIONAL STUDY: RESEARCH QUESTIONS



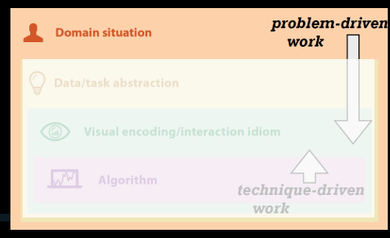
What do you want to know? What questions will your research answer?

Context & practices: How do physicians interact with their peers and their patients?

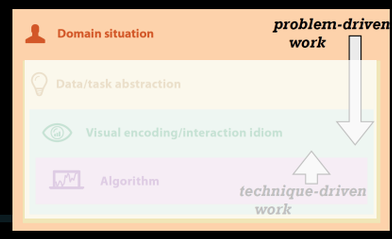
User requirements (and constraints): What kind of health information would users need in an integrated health portal?

Research questions are for you, not for your users.

EXAMPLE:
DESIGNING A MEDICAL DEVICE FOR NURSES

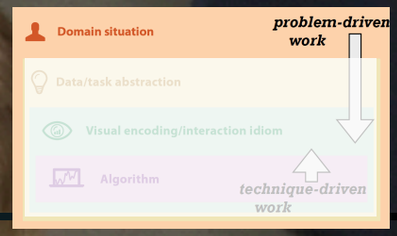


EXAMPLE:
DESIGNING A MEDICAL DEVICE FOR NURSES

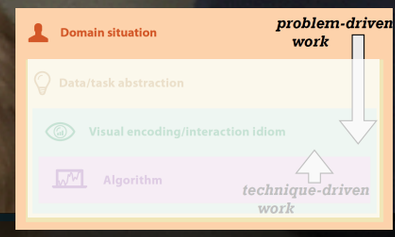


EXAMPLE:
DESIGNING A MEDICAL DEVICE FOR NURSES

PARTICIPATORY DESIGN



PARTICIPATORY DESIGN

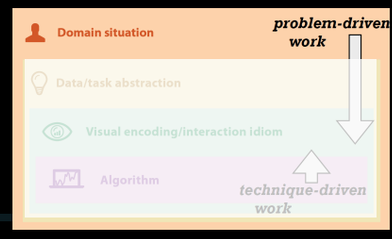


Member of the domain community is part of the design team.

- At early stage only
- Periodically
- Thorough the process

- Pro: insight of domain expert
- Con: different languages, expectations, potential for constrained thinking “we've always done it this way!”

ADOPTION / DEPLOYMENT STUDY



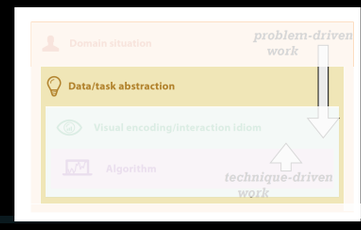
Also called a “longitudinal case study”: deploy the visualization and track its usage.

How?

- Invite participants to provide periodic feedback (interviews, diaries)
- Regular software logging
- Observations

Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

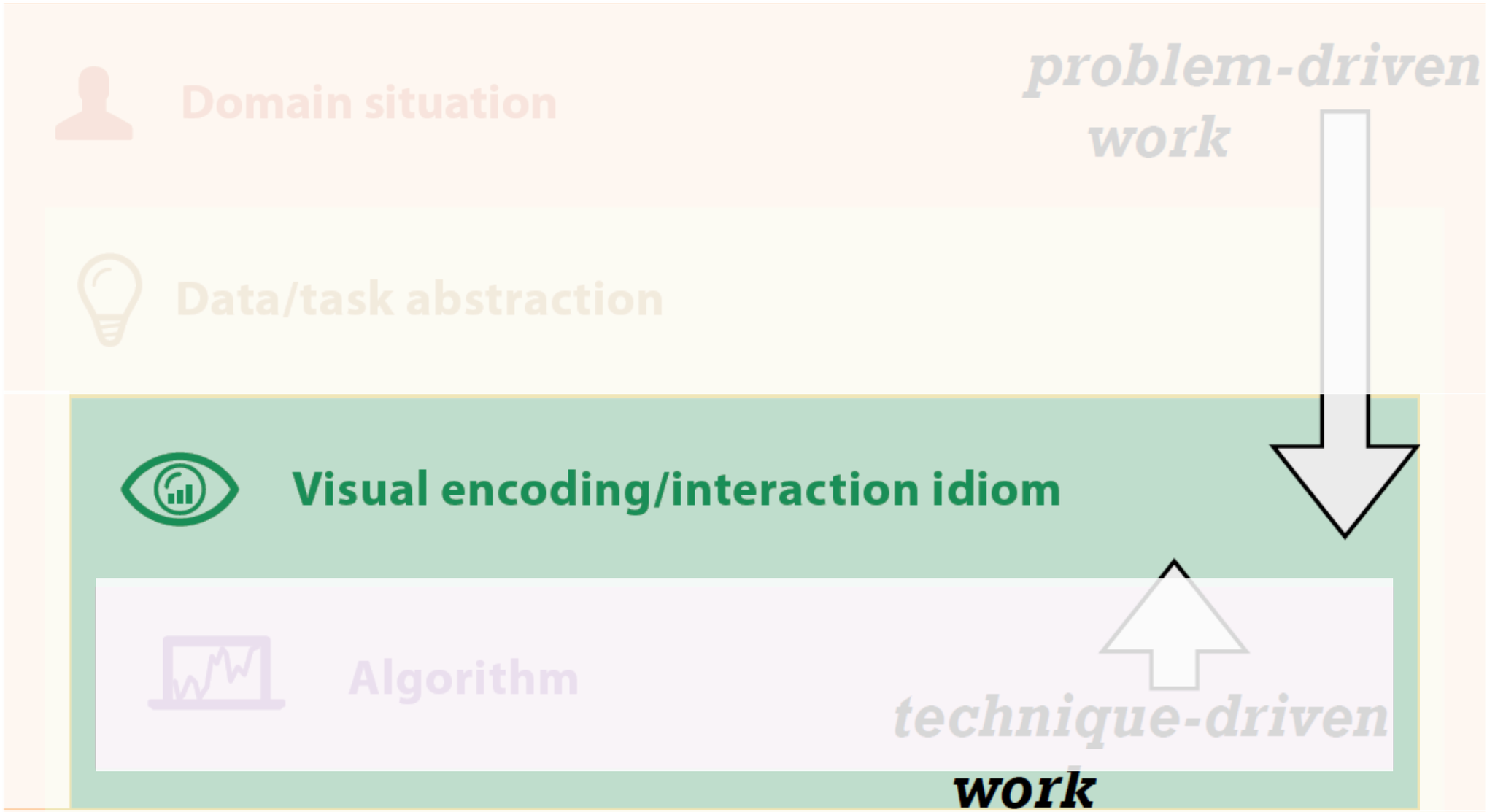
TASK AND DATA ABSTRACTION



Task abstraction — why? what for?
Consider data transformation — make it suit the task. Examples?

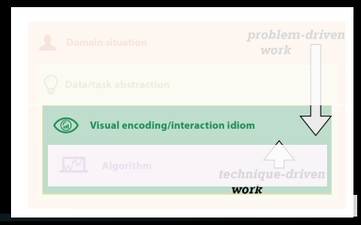
- No common upstream validation techniques
- Downstream: Field study — does the abstraction make sense to target users? Does it address their questions?

- Pitfall: No justification of decisions



Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

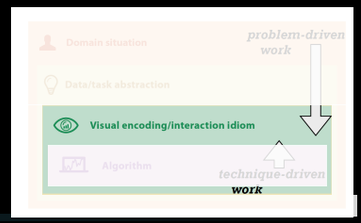
VISUAL ENCODING & INTERACTION IDIOM



- Upstream: Does the encoding match the task? Compare it to alternatives.
- Downstream:
  - User study: Analyze qualitatively (case study) or quantitatively (lab study)
  - Artifact: Analyze result image quantitatively (e.g. edge crossing, clutter) or qualitatively (aesthetic)

- Challenge: Isolating the part to study!
  - Examples: Medstory (difficult to study) vs. Useful Junk (easy to study)

CASE STUDY

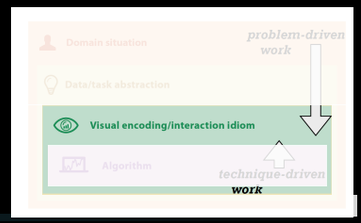


Demonstration of concepts and theories through an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions

Example:

- Learning about the utility of using ontology-based representation of data in the context of: medical experts conducting phenotype comparisons of patients to support diagnosis of rare genetic diseases
- Learning about the utility of representing uncertainty with sketchiness in the context of: communicating environmental trends and predictions in urban planning

CONTROLLED EXPERIMENT



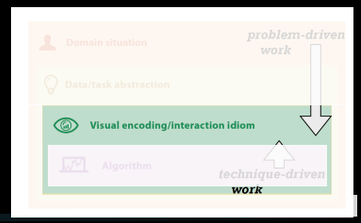
Traditional scientific method.

Collecting performance data while people are using a system (often lots of data!)

- Exploratory data collection
    - hope something interesting shows up
    - difficult to analyze
- Targeted data collection
    - look for specific information, but may miss something
    - e.g. frequency & type of request for assistance
    - e.g. frequency of use of different parts of the system
    - e.g. number of errors and when they occurred
    - e.g. time it takes to complete some operation

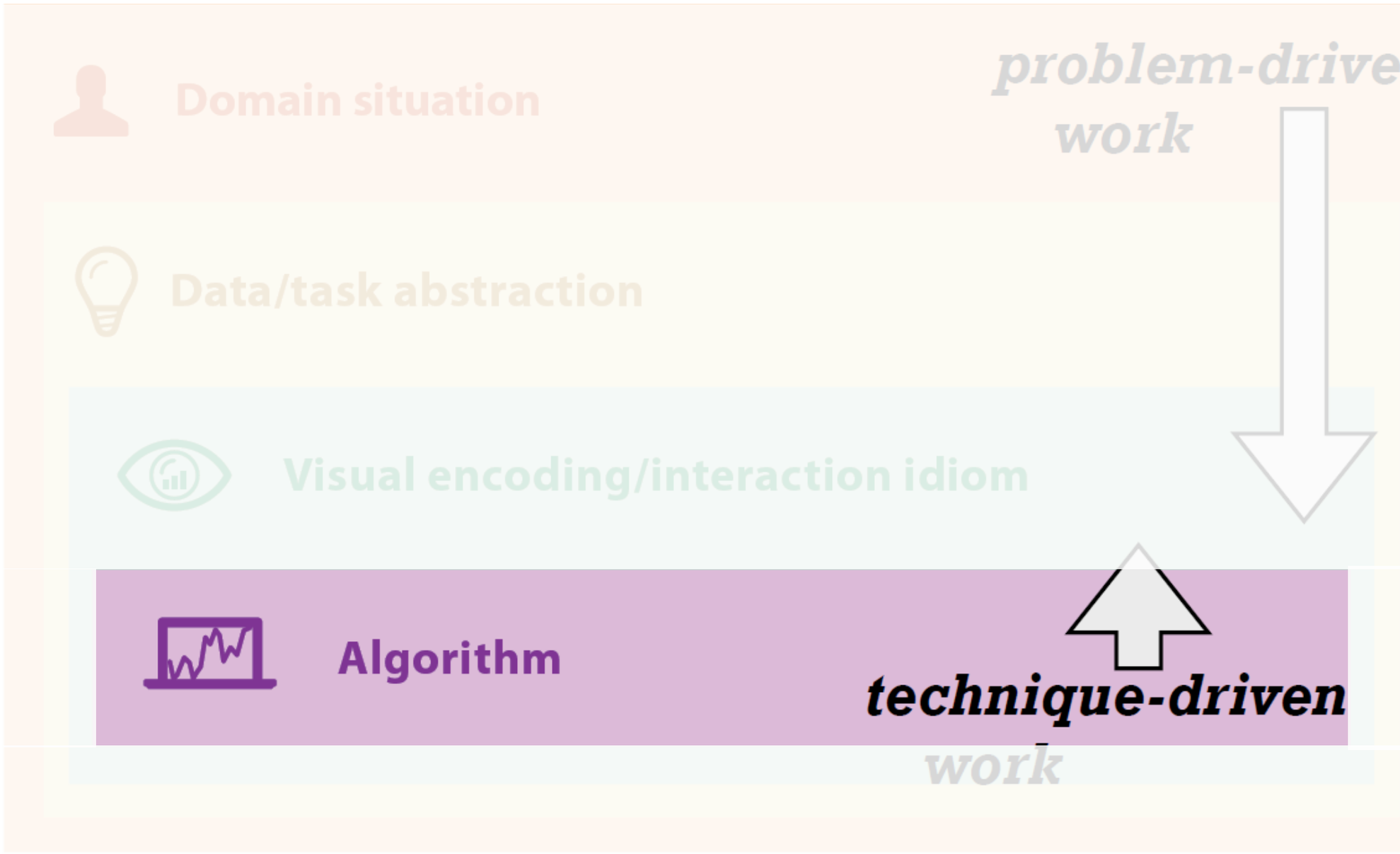
All of these tell you something about the usability of the visualization

CONTROLLED EXPERIMENT



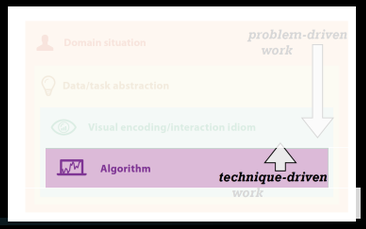
people - data - hypothesis - measure - statistic

- People: sample vs. population
- Data: variables measured from sample (statistic) to learn about the population (parameter)
- Null hypothesis: statement about the world, equality, trying to ***disprove*** it
- Statistics: analysis of results, confidence, hypothesis rejected?
- Beware of the problems with NHST



Source: Tamara Munzner: A Nested Model for Visualization Design and Validation. 2009.

ALGORITHM



- Upstream: proof, measure computational complexity
- Downstream: time/memory efficiency

LOW-COST EVALUATION

LOW-COST EVALUATION

Doesn’t require participants

Can be performed on early prototypes or even sketches

Repeatedly applied without additional cost (money, organisational effort)

HEURISTICS

A heuristic method is used to come to a solution rapidly that is hoped to be close to the best possible answer, or “optimal solution”. A heuristic is a “rule of thumb”, an educated guess, an intuitive judgement or simply common sense. A heuristic is a general way of solving a problem. Heuristics as a noun is another name heuristics methods.

http://en.wikipedia.org/wiki/Heuristic

HEURISTICS FOR INFOVIS

Zuk and Carpendale's *Selection of perceptual and cognitive heuristics [26]*	Ensure visual variable has sufficient length [3][25][26] Don't expect a reading order from color [3][25][26] Color perception varies with size of colored item [25][3][26] Local contrast affects color & gray perception [25][26] Consider people with color blindness [25][26][22] Preattentive benefits increase with field of view [3][25][26][11] Quantitative assessment requires position or size variation [3][26] Preserve data to graphic dimensionality [24][3][26] Put the most data in the least space [24][26] Remove the extraneous (ink) [24][26] Consider Gestalt Laws [25][26] Provide multiple levels of detail [24][25][26] Integrate text wherever relevant [24][25][26]

SUMMARY

VALIDATION

Validation can be carried out at all stages of design, from domain situation to algorithms

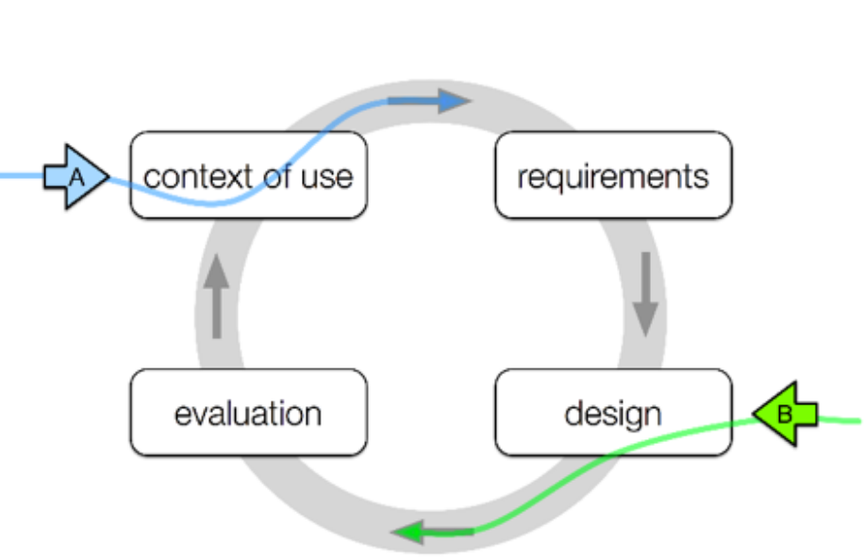
Validation take many forms, including justifications based on design principles, observational (field) studies, expert evaluation (heuristics, cognitive walkthrough), lab studies, and longitudinal case studies (deployment)

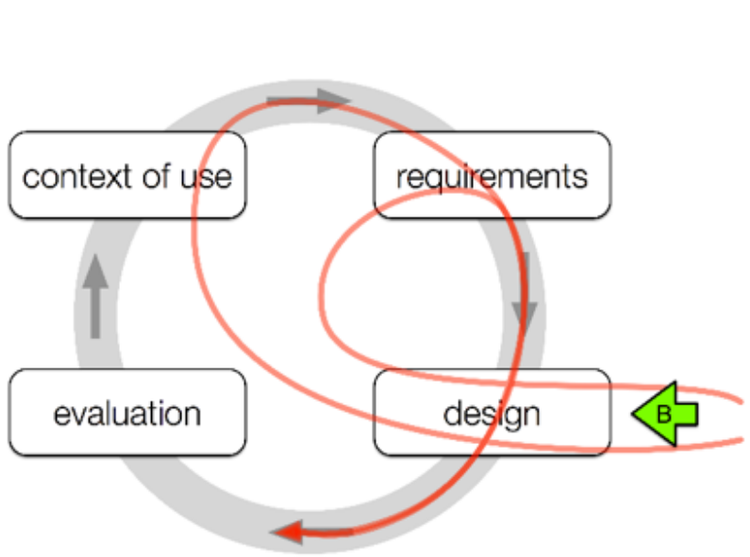
STRATEGIES

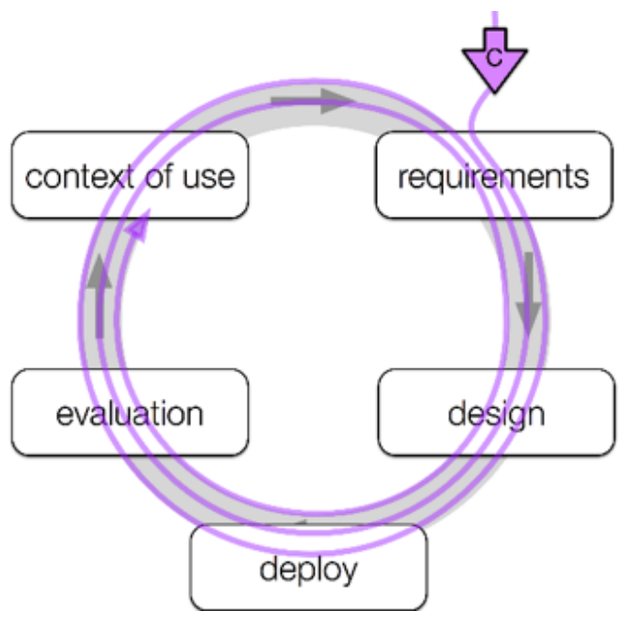
Problem-driven: start at a domain and work inward

Technique-driven: start at an idiom or algorithm, then exemplify usefulness with domain and task

DESIGN PROCESSES







EVALUATING UI VS. INFOVIS

Seems comparable but…

what are some differences?

USABILITY VS. UTILITY

Big difference

- usability is not the same as utility, which seems to be a key factor for Infovis
- can think of visualizations that are very usable but not useful or helpful
- more difficult to measure success of an infovis because more domain knowledge and situated use is required

THREATS TO VALIDITY

Threats to validity refers to when your validation technique isn’t answering the question it is supposed to address

Due to wrong validation technique, wrong question, or errors in the setup of the technique

THREATS TO VALIDITY

Threats to **external** validity: doesn't represent real use:

- study carried out in a lab when tool will be used in an office
- field study disrupted by the presence of the researchers

Threats to **internal** validity: experimental results are not trustworthy due to differences between trials:

- different abilities of participants
- different lighting conditions affecting the screen

THREATS TO VALIDITY

Threat Wrong problem

Validate Observe and interview target users

Threat Wrong task/data abstraction

Threat Ineffective encoding/interaction idiom

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 task

Validate Test on target users, collect anecdotal evidence of utility

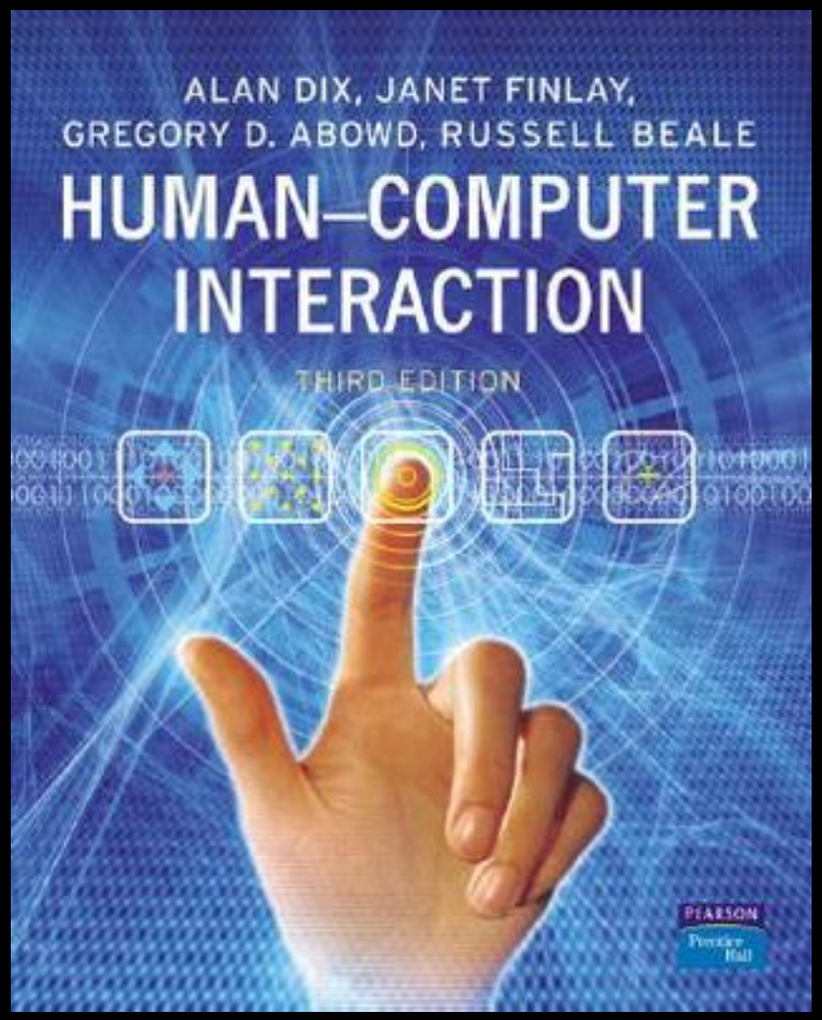
Validate Field study, document human usage of deployed system

Validate Observe adoption rates

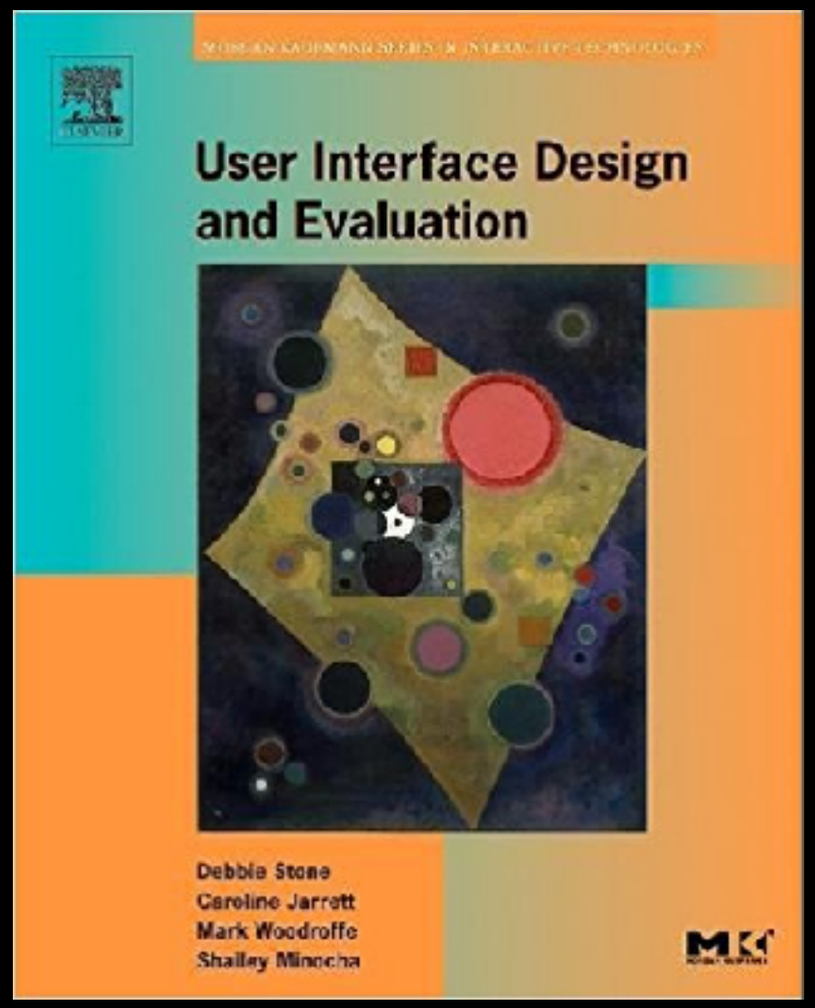
BELIV

Workshop focused on this topic

READINGS


ALAN DIX, JANET FINLAY,
GREGORY D. ABOWD, RUSSELL BEALE
HUMAN–COMPUTER INTERACTION
THIRD EDITION
PEARSON
Prentice Hall


Carol M. Barnum
Usability
Testing
Essentials
Ready, Set...Test!
MK
Foreword by Steve Krug, author
of Don't Make Me Think!


User Interface Design and Evaluation
Debbie Stone
Caroline Jarrett
Mark Woodroffe
Shailey Minocha
MK