

Why?

(EVALUATE)

MIRIAH MIKE (G) and MICHAEL (S)

the doodles are MG's fault

Why ask Why?

because they asked us to!

Stop The Evaluation Arms Race! A Call to Evaluate Visualization Evaluation

Position Paper *

Michael Gleicher
Department of Computer Sciences
University of Wisconsin - Madison
gleicher@cs.wisc.edu

Keywords

Visualization, Evaluation

Position Statement Summary: My position is that improving evaluation for visualization goes beyond simply developing more sophisticated evaluation techniques. It also requires us to improve our sophistication in reporting and assessing these evaluations and to make sure we consider the range of motivations for evaluation.

I would like to raise this issue at the workshop because I feel the latter issues (the motivations and the reporting/assessment) may be neglected as the community focuses on developing new evaluation techniques. While new techniques provide increasing sophistication in order to quantify what is “really important,” these methods seem to serve a more narrow range of purposes: they focus on summative assessment rather than providing insight to improve future tools. Additionally, as the evaluation methods applied become more complex, it becomes harder to assess their results, not only because there are more issues and details to consider, but also because for reporting the details, we lack standards of the sort we have for simpler, psychology-style statistical experiments.

Personal Statement Summary: While I consider myself a relative newcomer to the field of visualization, I have a longer history in other areas in which similar issues arise. My work in visualization has taken up the challenges in translating perception research into actionable design guidelines for visualization systems, which has included performing our own “low-level” evaluations to build models that inform design better than the prior studies in the perception literature. We have also confronted the issues in assessing the systems we have built. I have also needed to consider these issues in evaluation in my work on animation, image and video processing, virtual reality, and bioinformatics.

1. INTRODUCTION

Visualization serves many rich, complex, and open-ended goals. Challenges include helping a scientist make discoveries using a massive amount of data, aiding a mass audience in understanding a complex topic, or helping an analyst detect anomalies in a massively complex network. Evaluation can play an important role in assessing how the visualizations (and the systems that create them) serve these needs. Evaluations can also serve to inform design.

Increasingly, the evaluation methods used in visualization are becoming more sophisticated in order to meet these challenges. For example, sophisticated empirical designs offer to quantify “what really matters” by measuring high-level outcomes (such as “insights” [6] or “learning” [2]), crowd-sourced experiments offer to allow exploration of large parameter spaces, and biometric measurements (eye-tracking, functional near-infrared spectroscopy[7], skin capacitance, ...) offer to quantify viewer response. However, the increasing diversity and frequency of empirical studies raises new questions. As the evaluations become complex, understanding and assessing them becomes complicated. Making complex evaluations interpretable by their audiences (e.g. potential users, reviewers, ...) is an important challenge. It is important to communicate evaluations effectively, but reporting complex evaluations is challenging, especially with novel evaluation techniques.

As evaluation becomes more empirical and more sophisticated, translating it to actionable information to guide the development of visualizations may become more challenging. Quantification of higher-level outcomes (e.g. insight or learning) can provide nice summative assessments that a system performs, but it does not necessarily provide insights into why the system is successful that can be transferred to future work. Indeed, the fact that the measurements are more complex and higher-level often makes them more removed from the design decisions that provide lessons for future systems. At the opposite end of the spectrum, highly controlled experiments (such as perceptual or model-task studies) are often sufficiently de-contextualized from their implications for systems (and the high-level outcomes we desire from them) that their results are challenging to use to inform design.

The challenges of reporting, assessing, and translating evaluations, may grow as the evaluation methods become more

The Four-Level Nested Model Revisited: Blocks and Guidelines

Miriah Meyer
University of Utah
miriah@cs.utah.edu

Michael Sedlmair
University of British Columbia
msedl@cs.ubc.ca

Tamara Munzner
University of British Columbia
tmm@cs.ubc.ca

ABSTRACT

We propose an extension to the four-level nested model of design and validation of visualization system that defines the term “guidelines” in terms of blocks at each level. Blocks are the outcomes of the design process at a specific level, and guidelines discuss relationships between these blocks. Within-level guidelines provide comparisons for blocks within the same level, while between-level guidelines provide mappings between adjacent levels of design. These guidelines help a designer choose which abstractions, techniques, and algorithms are reasonable to combine when building a visualization system. This definition of guideline allows analysis of how the validation efforts in different kinds of papers typically lead to different kinds of guidelines. Analysis through the lens of blocks and guidelines also led us to identify four major needs: a definition of the meaning of block at the problem level; mid-level task taxonomies to fill in the blocks at the abstraction level; refinement of the model itself at the abstraction level; and a more complete set of mappings up from the algorithm level to the technique level. These gaps in visualization knowledge present rich opportunities for future work.

Categories and Subject Descriptors

H.5.2 [Information Systems Application]: User Interfaces—*Evaluation/methodology*

Keywords

Nested model, validation, design studies, visualization

1. INTRODUCTION

In 2009, Munzner proposed the four-level nested design model as a framework for thinking about the design and validation of visualization systems at four cascading levels [22]. This model makes explicit the negative impact of poor design decisions early on in a project and stresses the importance of choosing appropriate validation techniques at each level. The nested model has provided guidance, motivation, framing, and ammunition for a broad range of visualization papers, including problem-driven design studies [7,

11, 24, 27, 29, 32], technique-driven work [18], evaluation [1, 31], models [8, 10, 17, 19, 28, 30, 35], and systems [4, 12].

We use the nested model extensively as a way to guide and reflect about our own work, and propose an extension of the model motivated by our desire to clarify the meaning of the term *guideline*. This term is loosely defined within the visualization literature to describe knowledge that guides how we make design decisions. One of our goals with this work is to clarify the meaning of this term for visualization research in order to assess the impact of both design studies and technique work on guidelines.

The extension proposes *blocks* as a generic term for the outcomes of the design process at the three lower levels: abstractions, techniques, and algorithms. Concrete examples of blocks at each of these levels are that a network is a data abstraction block, a node-link diagram is a visual encoding block, and a specific force-directed layout approach such as GEM [13] is an algorithm block. We can then define *guidelines* as statements about the relationships between blocks, such as a node-link diagram is a good visual encoding of small graphs, or a specific force-directed layout algorithm is faster than another. We consider *guideline* and *characterization* to be synonyms.

Guidelines may pertain to blocks within a single level, and we call these within-level guidelines *comparisons*. Guidelines may also cross between levels; we call these between-level guidelines *mappings* to emphasize their role in moving from one level to the next. Both types of guidelines often arise from reflection after evaluation and validation efforts. Comparison guidelines are the result of pitting one block against others at the same level, and often stem from validation efforts in papers that present a new technique or algorithm. Mapping guidelines provide guidance on how a block at one level is a match or mismatch with a block at an adjacent level. Mappings typically emerge from the validation of design studies. Evaluation papers may result in either kind of guideline, mappings or comparison.

The primary contribution of this paper is our proposed extension to the nested model and the resulting implications, presented in Section 2. Section 3 presents an analysis of open problems in our field illuminated by these extensions: the needs to define *block* at the problem level, to create mid-level task taxonomies at the abstraction level and possibly to refine the model itself at that level, and to establish a more complete set of mappings up from the algorithm level to the technique level. Thus, a secondary contribution is the elucidation of these gaps in our collective knowledge and a call for action to close them.

PREMISE:

CONSIDERING THE MOTIVATIONS

can help us

do GOOD EVALUATION

and good evaluation is really our goal!

Why seek Good evaluations?

because they can:

GUIDE

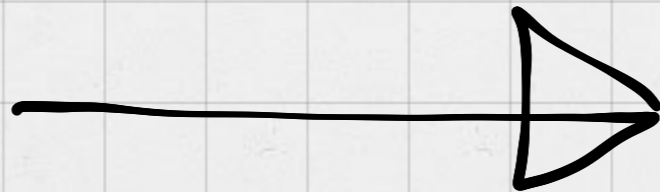
and PERSUADE

and other things we won't talk about now

GUIDE

How well do you inform
the AUDIENCE to do some THING?

non-
prescriptive

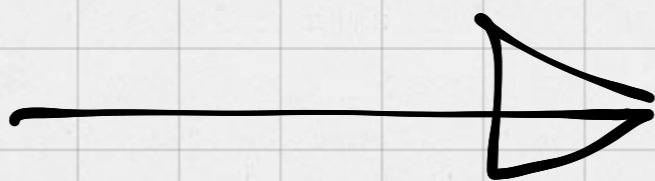


ACTIONABLE

PERSUADE

How well do you convince
the AUDIENCE to believe some THING

vacuous
assertion

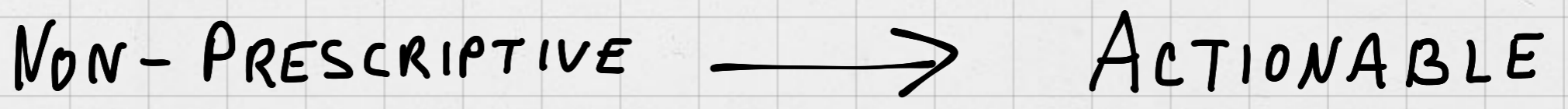


Convincing
argument

PERSUADE



TWO INDEPENDENT AXES



GUINE

PERSUADE

Vacuous assertion → compelling argument

CONVINCE AUDIENCE
AND
TELL THEM WHAT TO DO

AUDIENCE WON'T BELIEVE
WOULDN'T KNOW WHAT TO DO
EVEN IF THEY DID

NON-PRESCRIPTIVE → ACTIONABLE

GUIDE

PERSUADE

Vacuous assertion → compelling argument

CONVINCE AUDIENCE
BUT THEY CAN'T DO
ANYTHING ABOUT IT

TELLS AUDIENCE WHAT TO DO
DOESN'T CONVINCE THEM TO
DO IT

NON-PRESCRIPTIVE → ACTIONABLE

GUIDE

GUIDE

How well do you inform
the audience to do some thing?

PERSUADE

How well do you convince
the audience of some thing

GUIDE

How well do you inform
the audience to do some thing?

PERSUADE

CONTEXT

How well do you convince
the audience of some thing

how ACTIONABLE and how PERSUASIVE
depends on CONTEXT

SCIENTISTS USING TOOL A MAKE MORE DISCOVERIES
THAN USING TOOL B

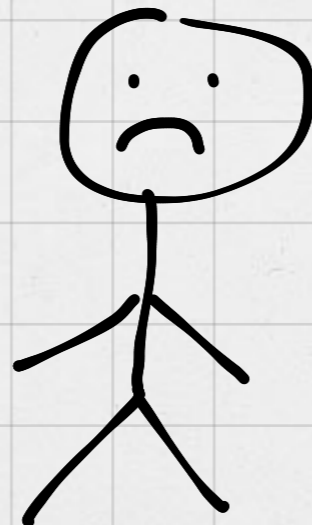
NOT-ACTIONABLE

ACTIONABLE

SCIENTISTS USING TOOL A MAKE MORE DISCOVERIES
THAN USING TOOL B

NOT-ACTIONABLE

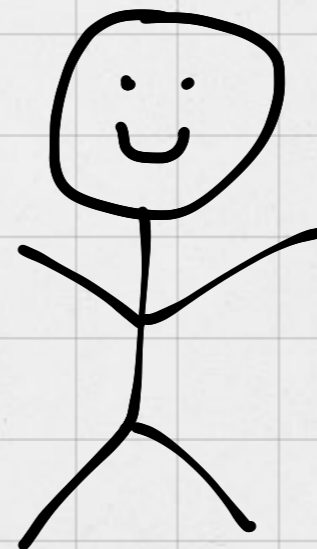
So what?
How does this help me
make better tools?



VIS
RESEARCHER

ACTIONABLE

Cool! I'll buy 10
copies of tool A for
my lab!



BIOLOGY
LAB
DIRECTOR

A PUNDIT ASSERTS MINIMALISM IS GOOD IN A
SELF-PUBLISHED BOOK

NOT PERSUASIVE

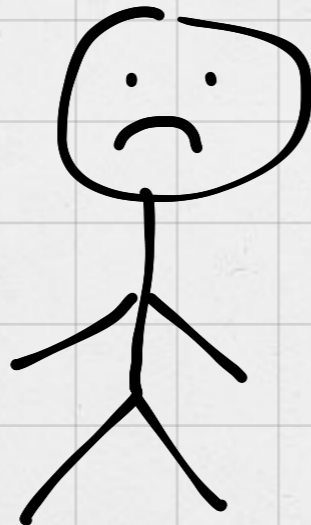
PERSUASIVE

A PUNDIT ASSERTS MINIMALISM IS GOOD IN A SELF-PUBLISHED BOOK

NOT PERSUASIVE

Show me some evidence!
Do a study - give me stats!

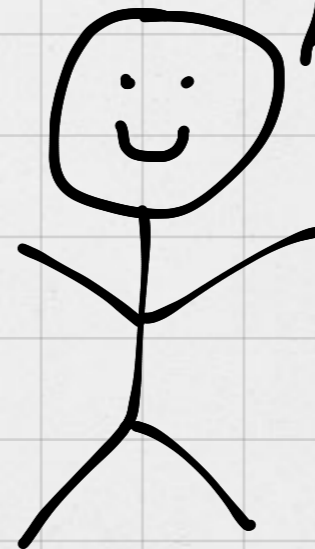
VIS
RESEARCHER



PERSUASIVE

Wow! He's famous - he
must know what he's
talking about
And he's a good writer too!

NORMAL
PERSON



EVALUATING EVALUATIONS

How well do they *guide*?

How well do they *persuade*?

MAKING GOOD EVALUATIONS

MAKE IT ACTIONABLE

MAKE IT PERSUASIVE

TO THE TARGET AUDIENCE

HOW TO MAKE PERSUASIVE EVALUATIONS?

MEASURE THE RIGHT THINGS

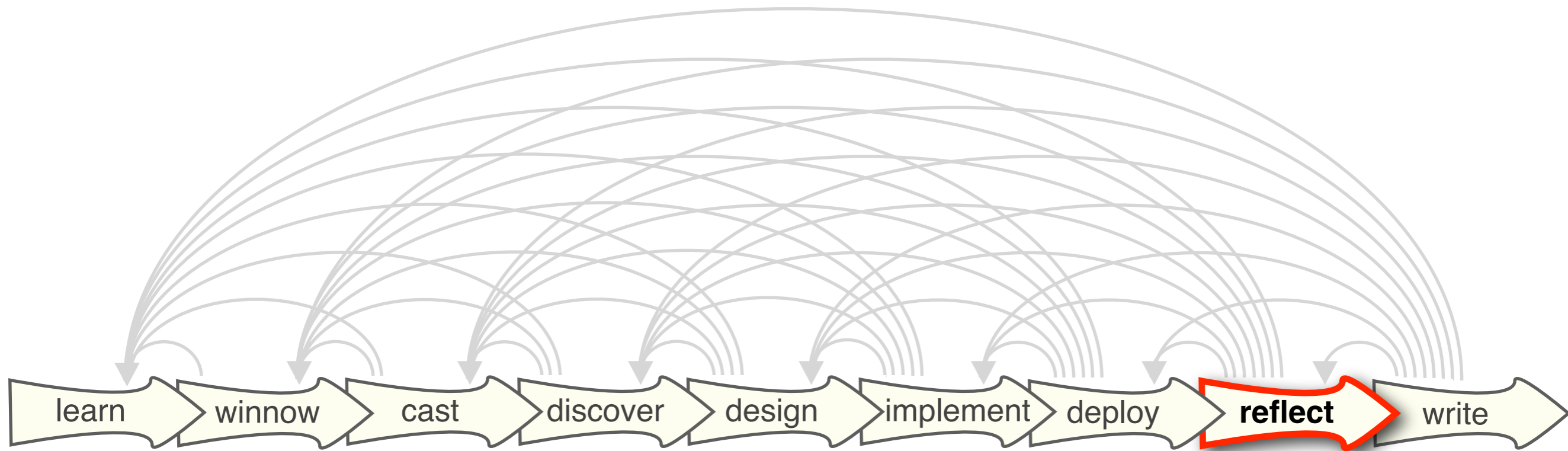
DESIGN GOOD EXPERIMENTS

REPORT IT IN A WAY THAT CONVINCES THE AUDIENCE

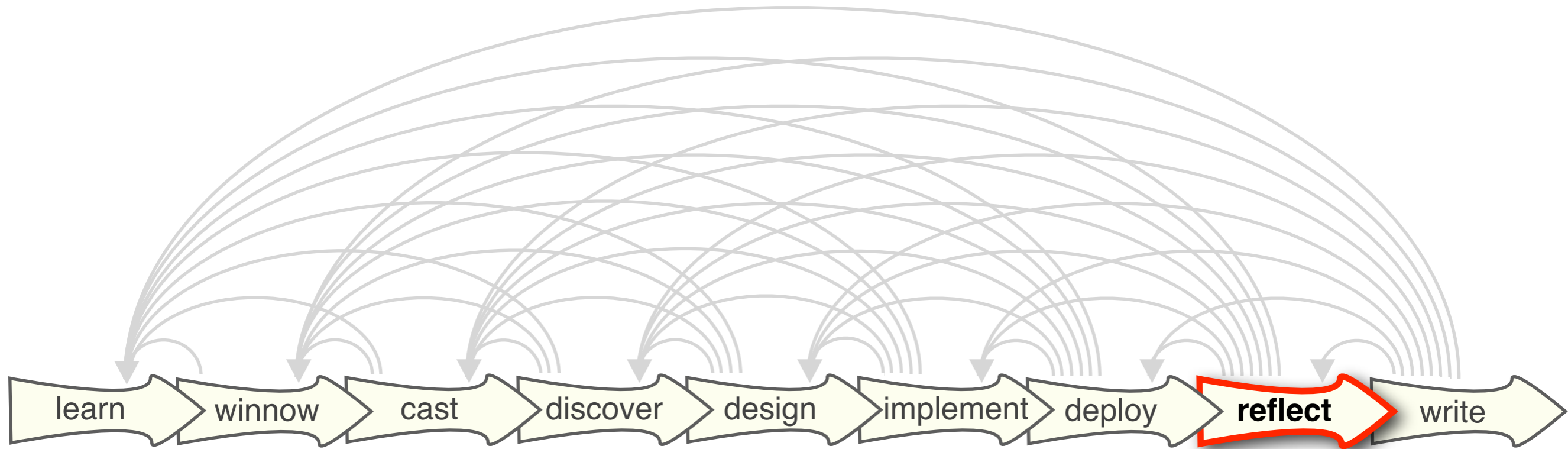
Sophisticated methods may be harder to report

How TO MAKE ACTIONABLE EVALUATIONS ?

?

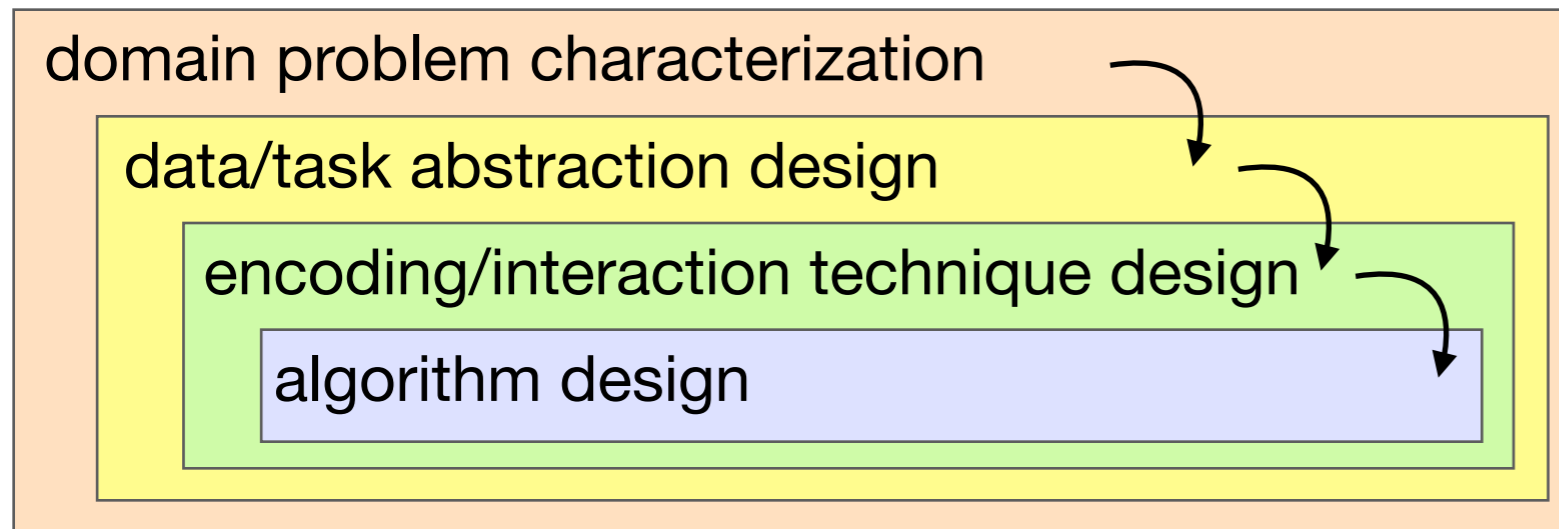


Design Study Methodology: Reflections for the Trenches and the Stacks.
M. Sedlmair, M. Meyer, T. Munzner, IEEE TVCG (Proceedings of InfoVis 2012).



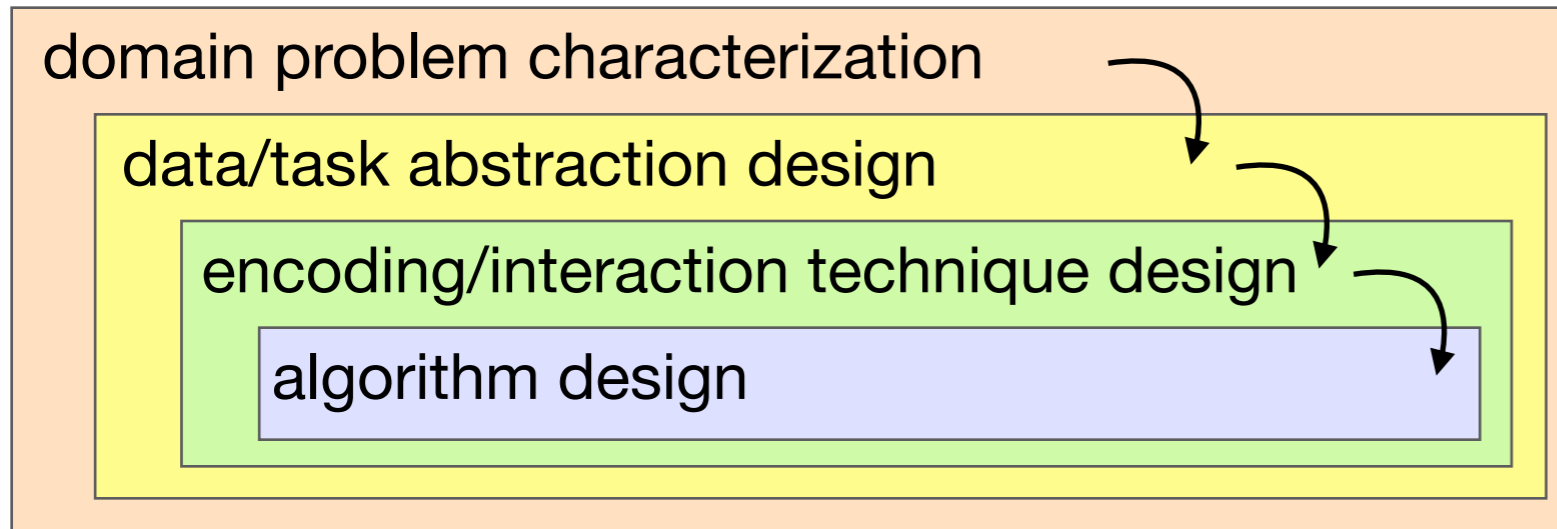
confirm | refine | reject | propose
guidelines

NESTED MODEL



Munzner 2009

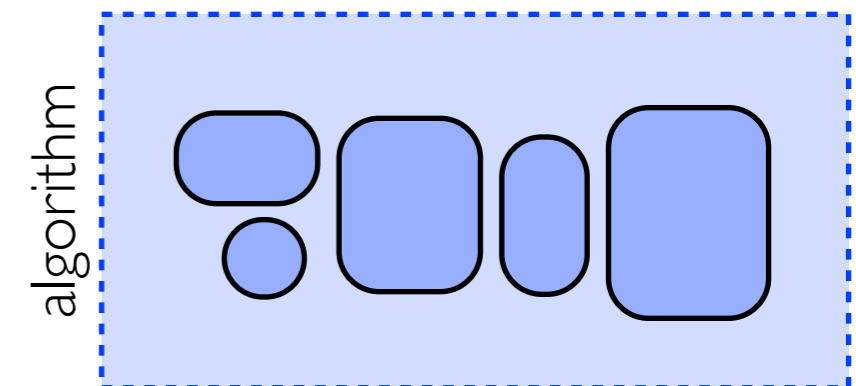
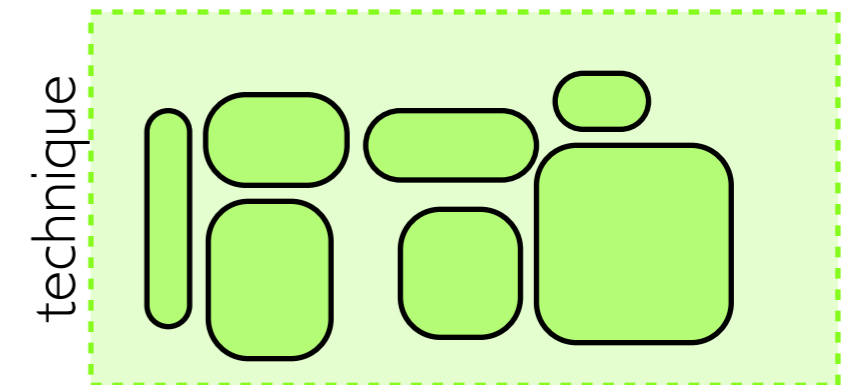
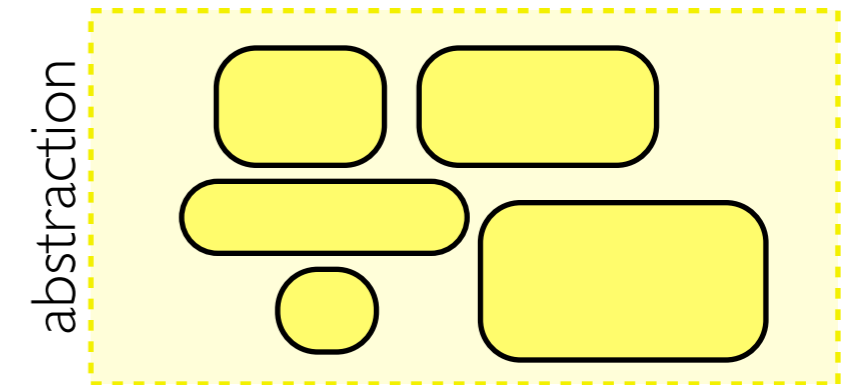
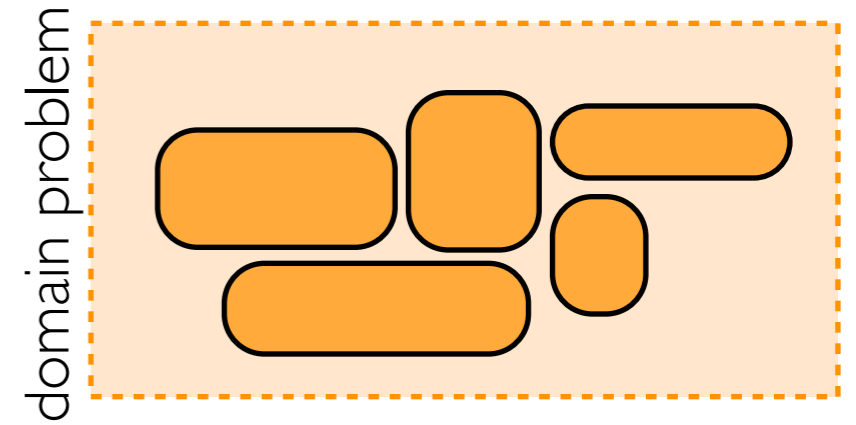
NESTED MODEL



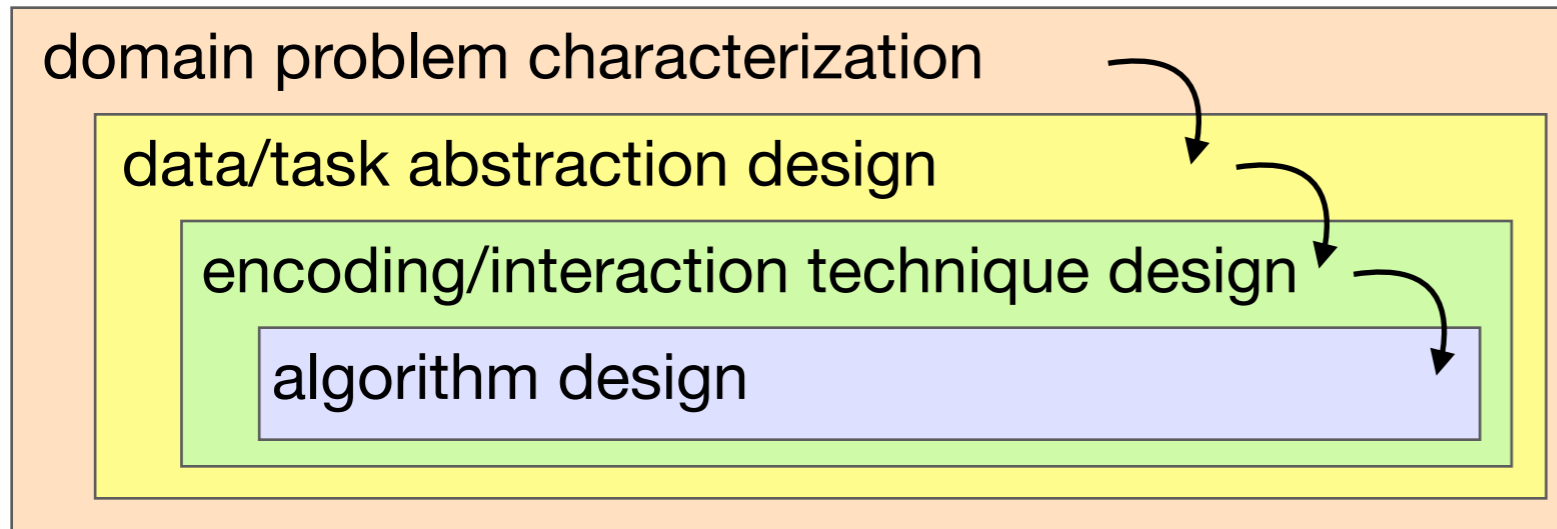
Munzner 2009

blocks
outcome of a design decision

EXTENSION



NESTED MODEL



Munzner 2009

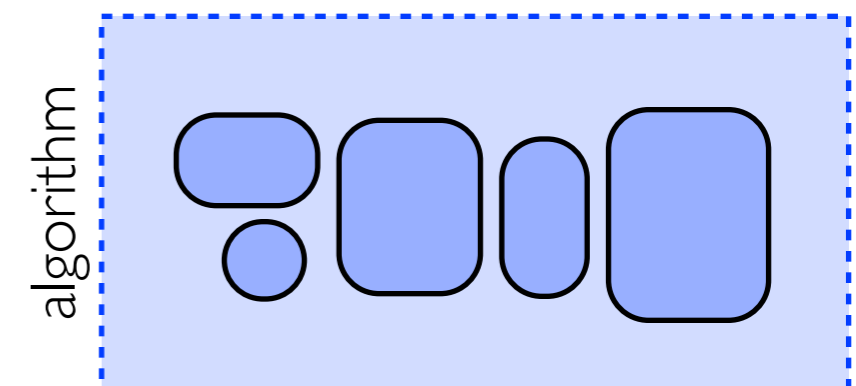
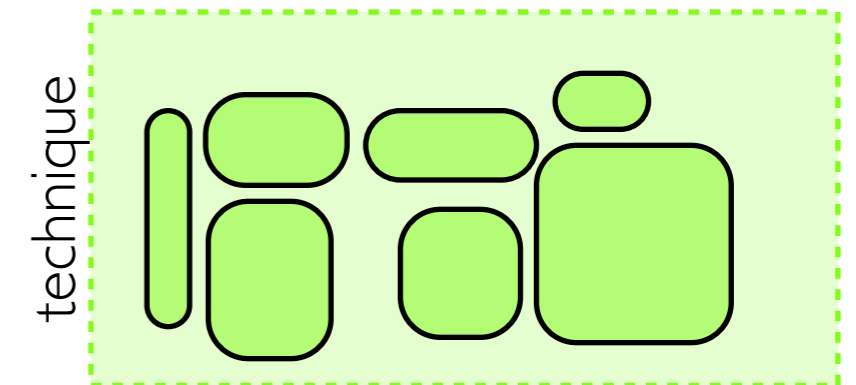
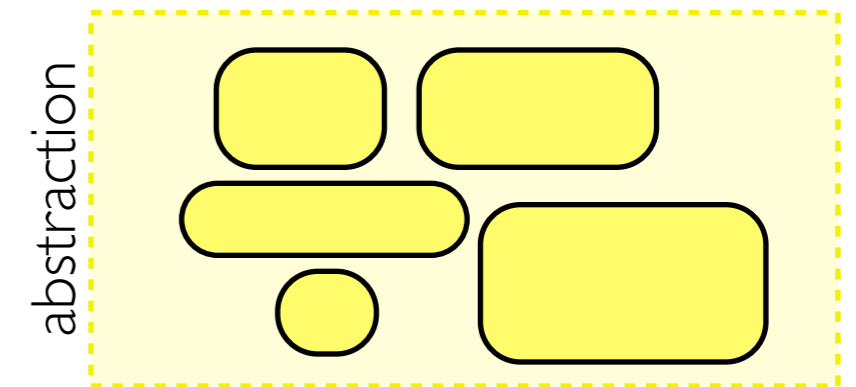
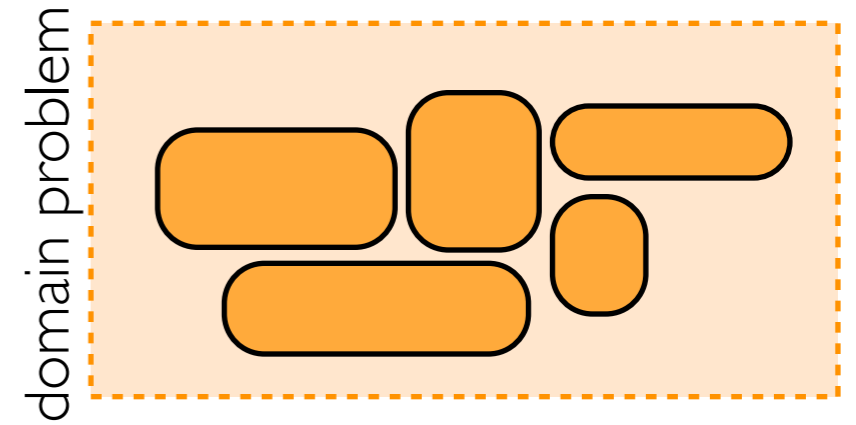
blocks
outcome of a design decision

directed graph

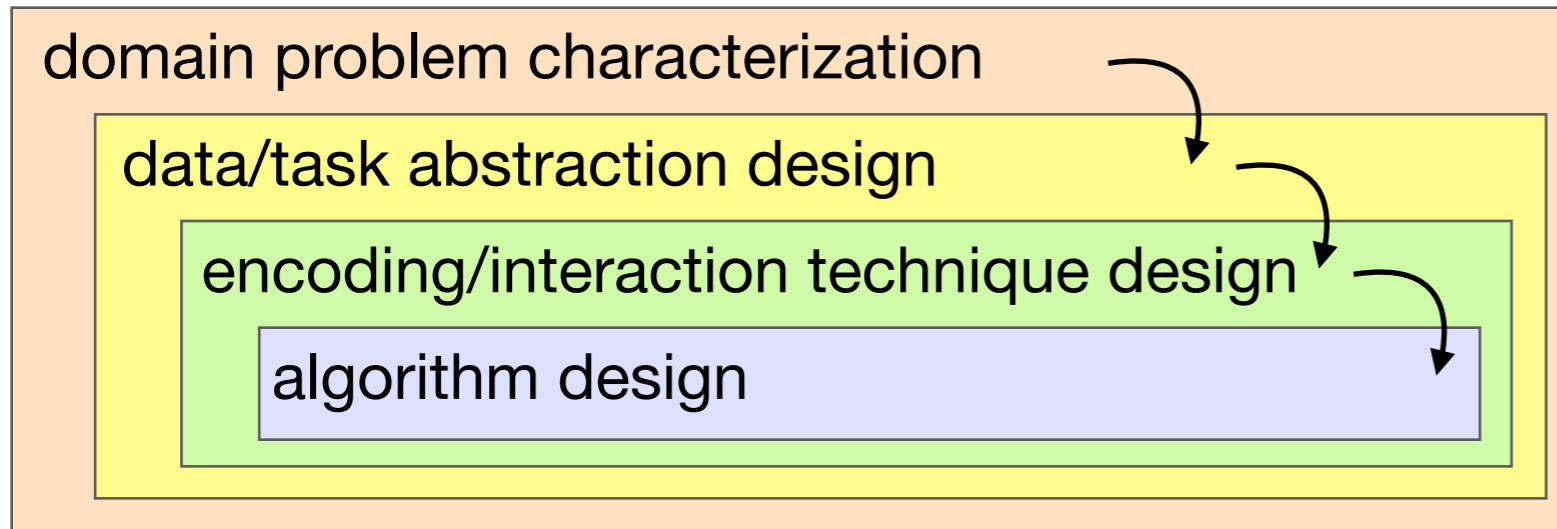
node-link
diagram

force-directed layout

EXTENSION

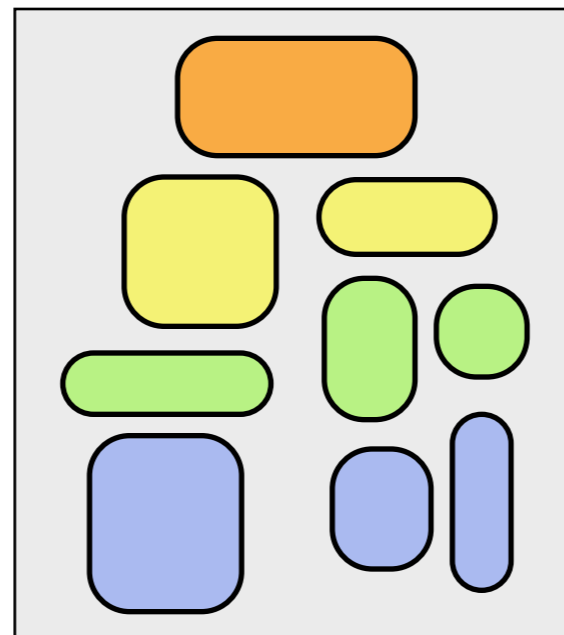


NESTED MODEL

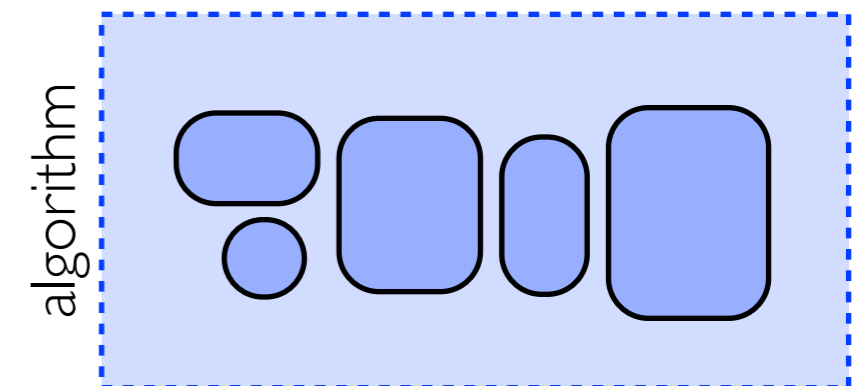
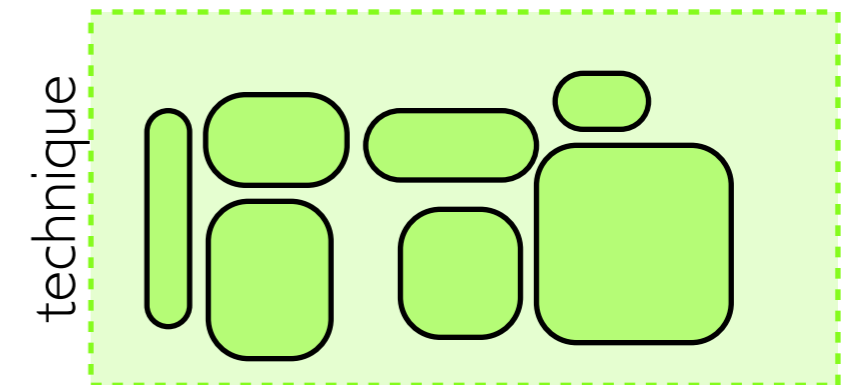
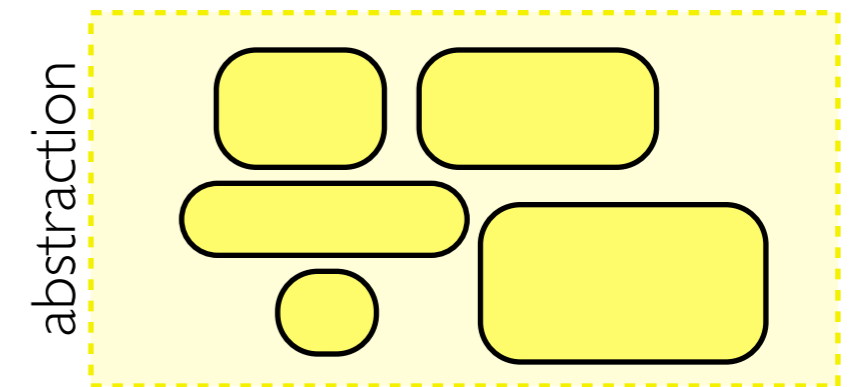
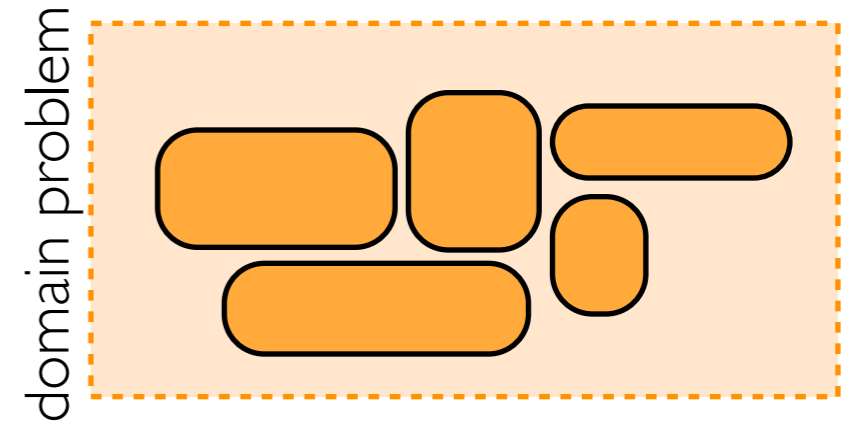


Munzner 2009

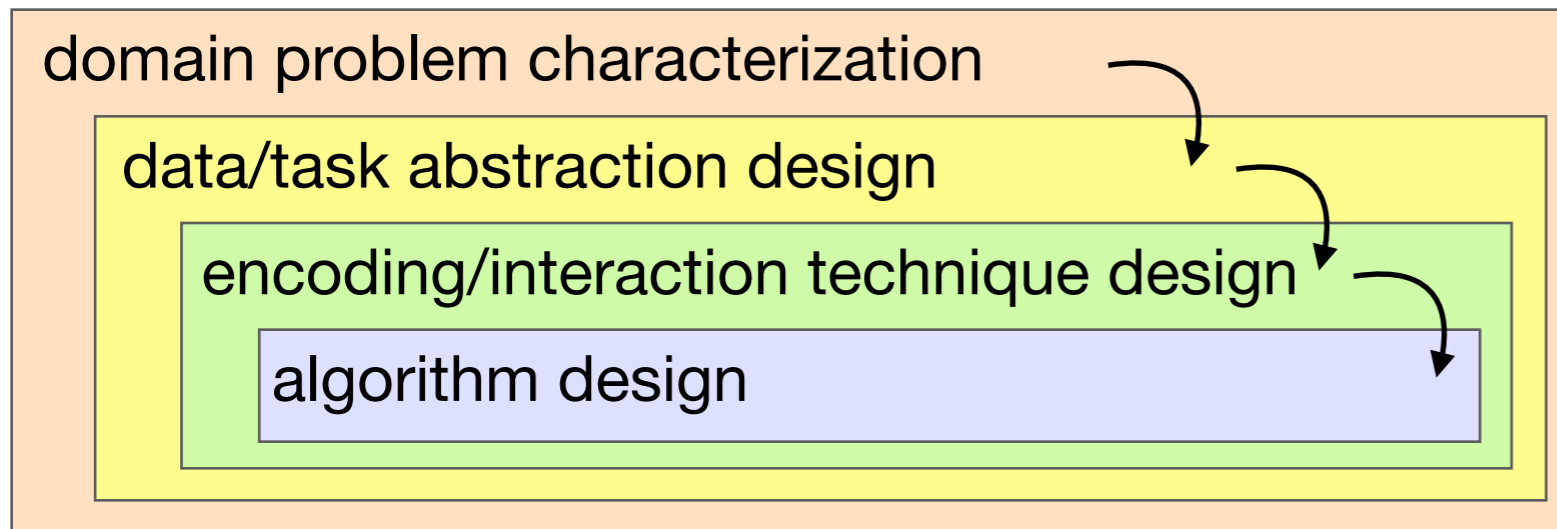
blocks
outcome of a design decision



EXTENSION



NESTED MODEL

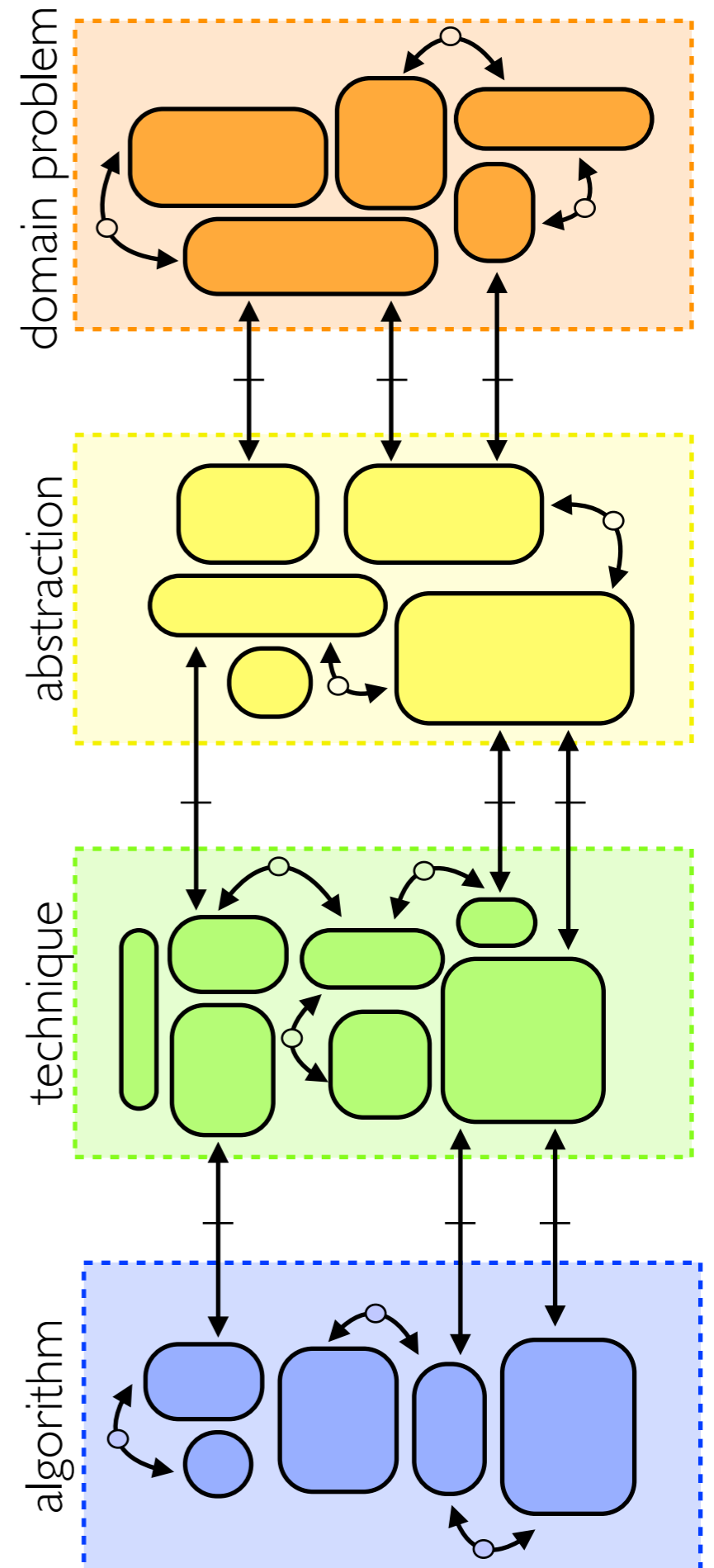


Munzner 2009

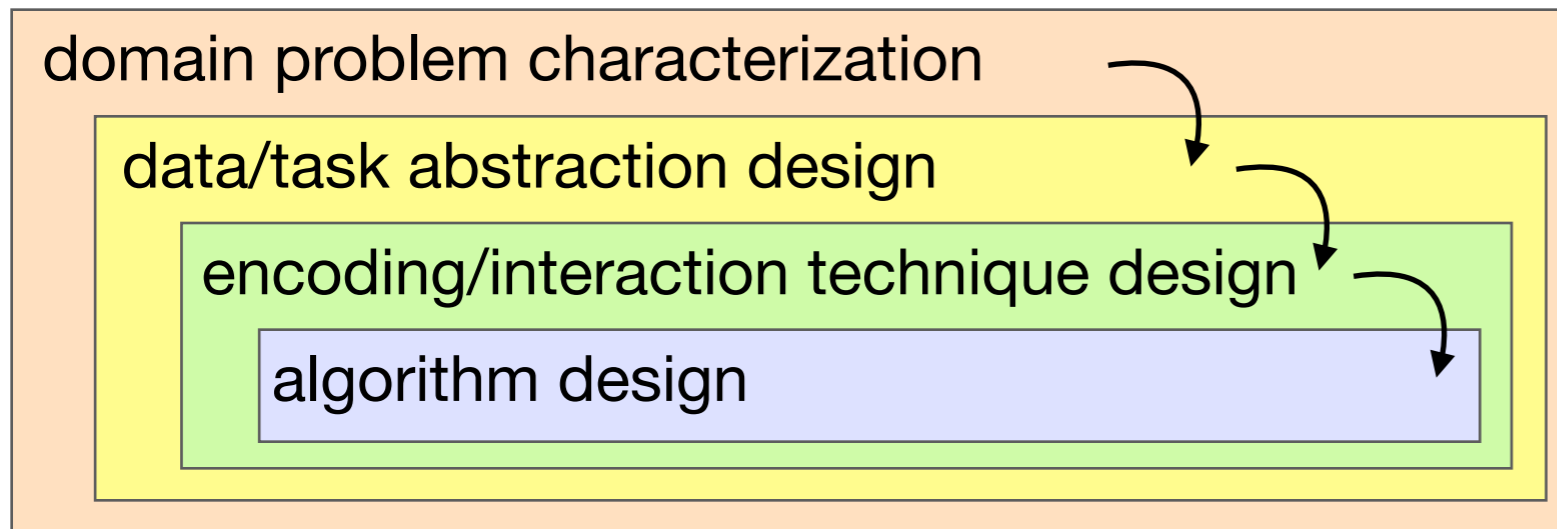
blocks guidelines

statement about relationship
between blocks

EXTENSION

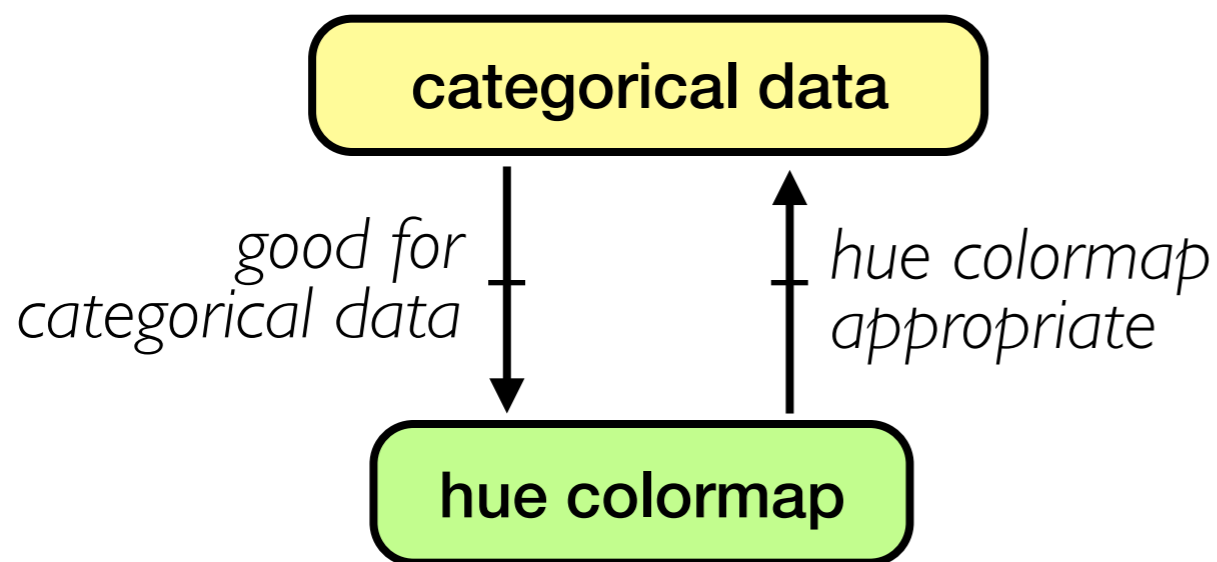


NESTED MODEL

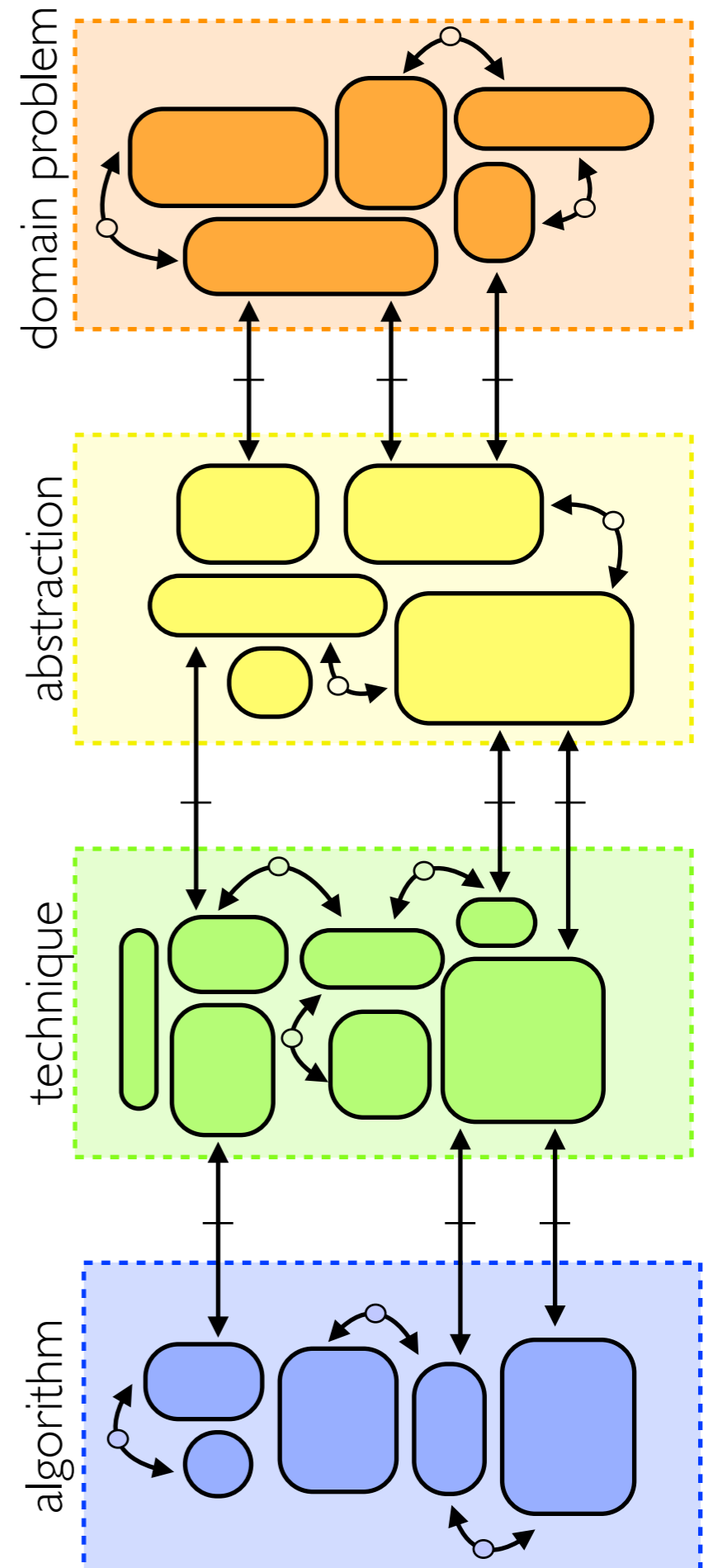


Munzner 2009

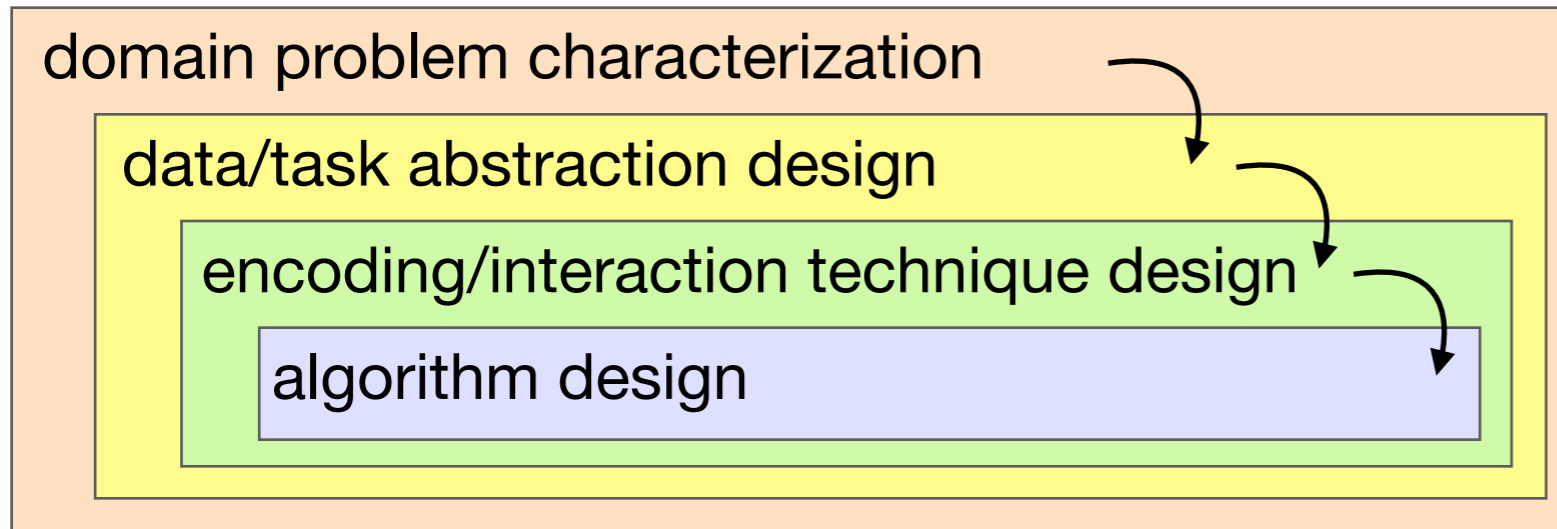
blocks guidelines



EXTENSION



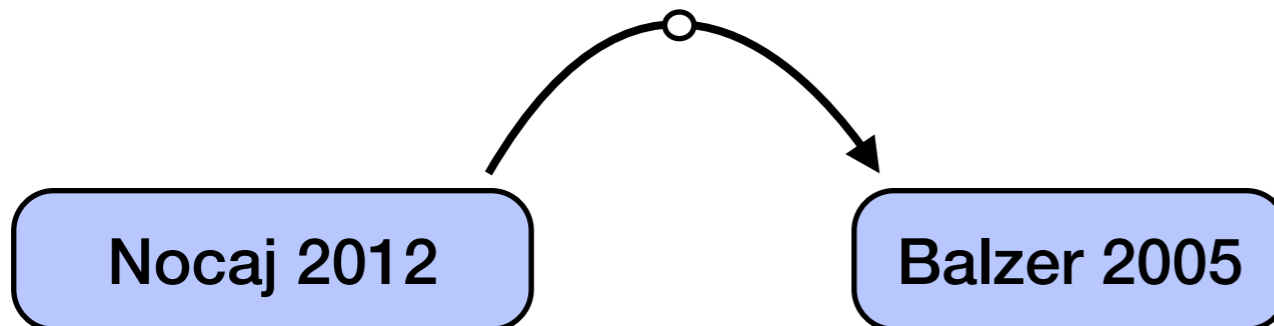
NESTED MODEL



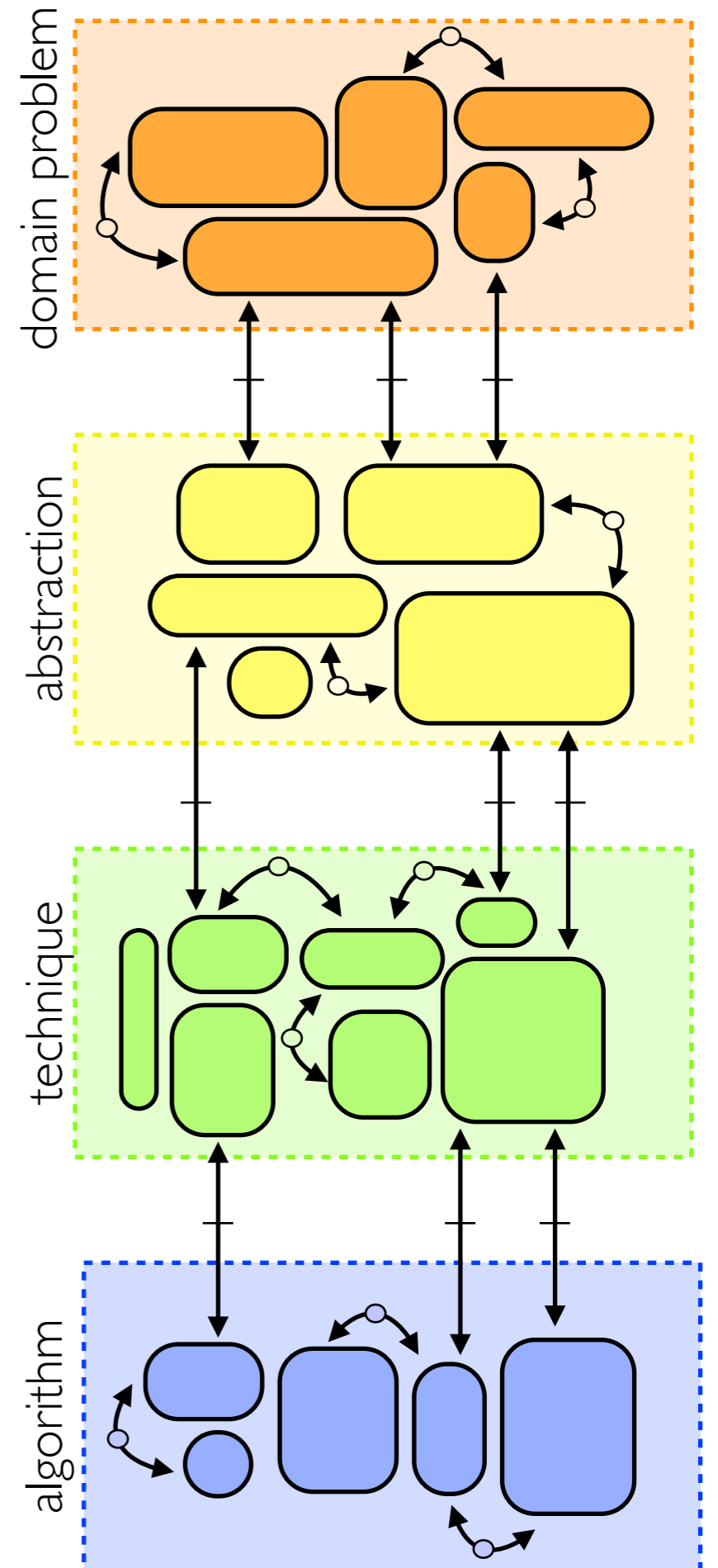
Munzner 2009

blocks guidelines

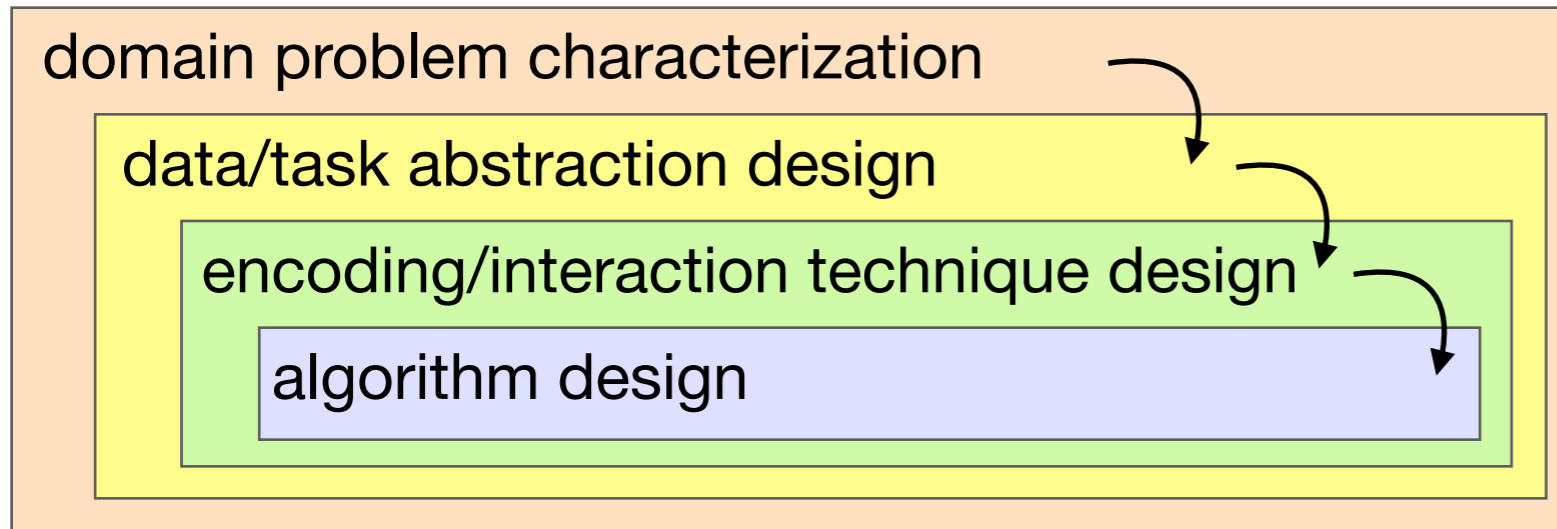
faster Voronoi treemap



EXTENSION



NESTED MODEL



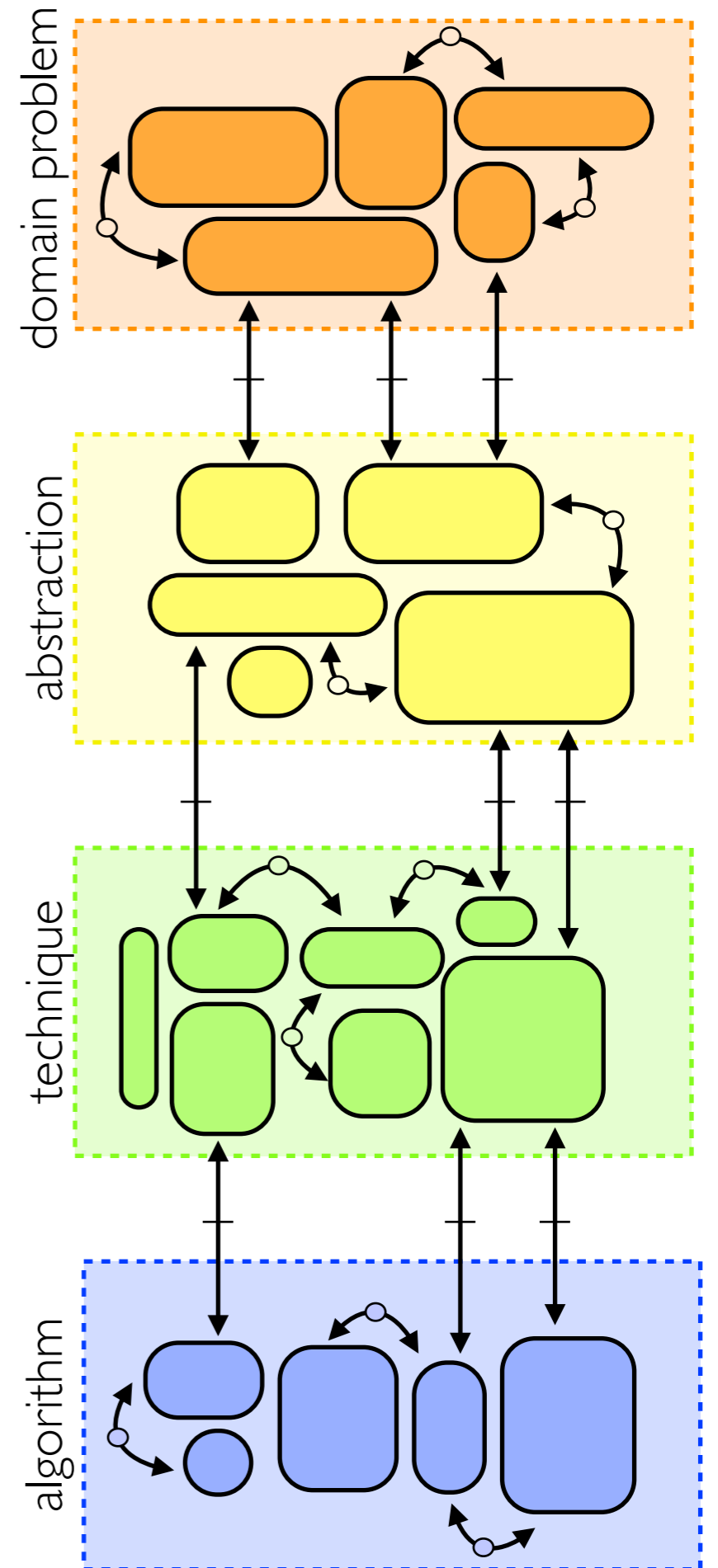
Munzner 2009

blocks guidelines

between-level mapping \updownarrow

within-level comparison \updownarrow



EXTENSION

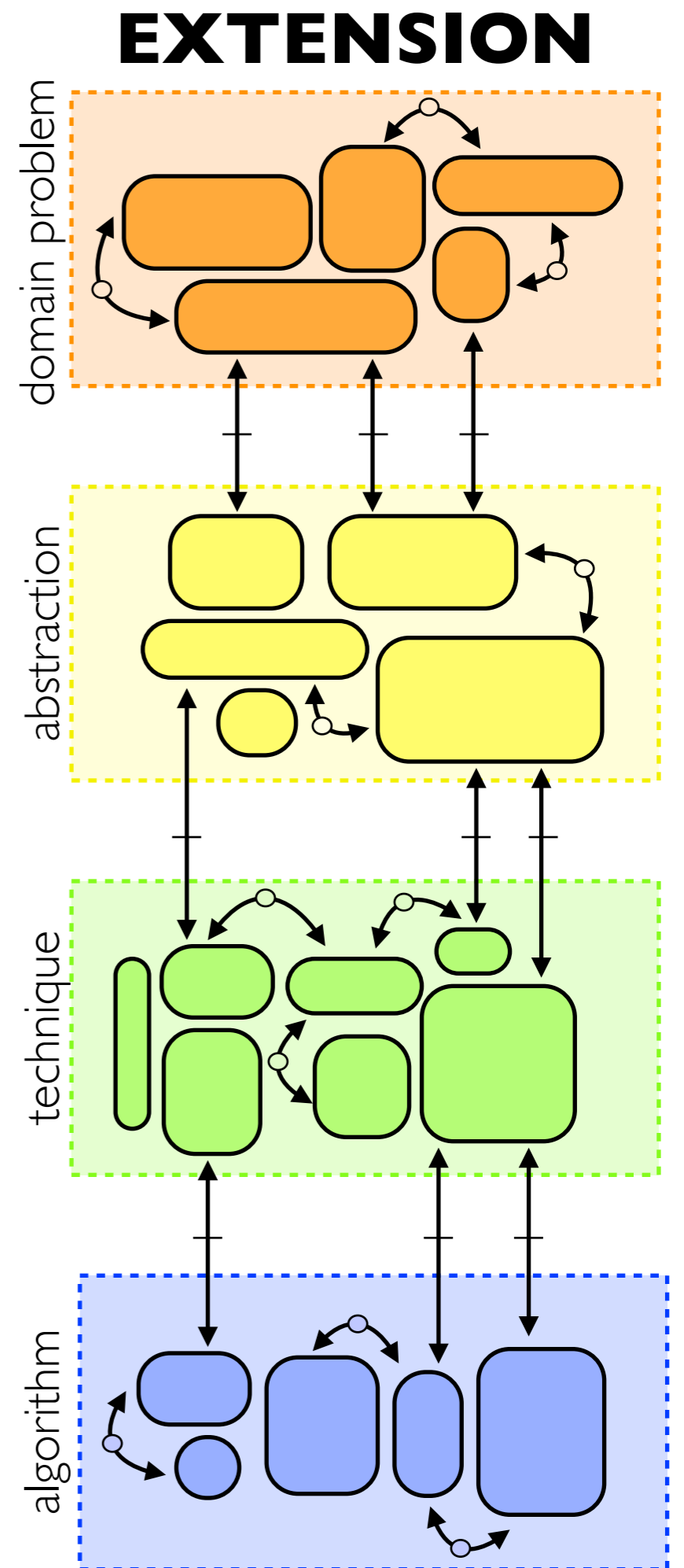


implications

mapping problems to abstractions
mapping algorithms to techniques
reporting algorithm comparisons

blocks guidelines

between-level mapping 
within-level comparison 



persuade & guide

blocks & guidelines

vocabulary

persuade & guide

actionable & convincing

blocks

guidelines

mappings & comparisons

DISCUSSION IDEAS

Are these useful concepts? Are these the right words for them?
Are there other useful axes to describe evaluations?

How do we make evaluations actionable? Is this the right goal?
Are there other types of actionability besides guidelines?

What evaluation methods to use for which contexts?
What does the diversity of contexts mean for method development?
Are all goals and contexts being served?

Is the visualization community too focused on itself as a context?
Do our evaluations help anyone other than us? Should they?