Visual Data Analysis:
real users, real data,
right methods

Michael Sedlmair

Junior Group Leader
University of Vienna, Austria
Jul 22, 2013

data

becomes ubiquitous

visualization & data analysis

real users/real data:

- understand real needs, problems and practices & provide visual and computational solutions

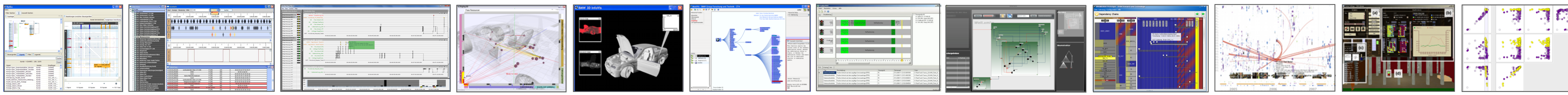
- characterize how proposed data analysis techniques perform under real circumstances

right methods:

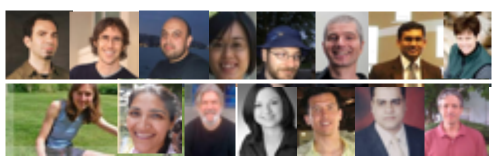
- innovate and refine research methods for visual data analysis

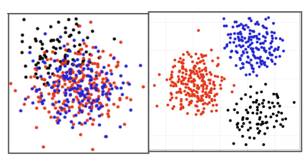
real users/real data:

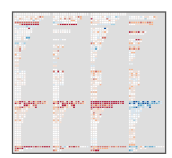
Applied visualization projects (9 BMW, 3 others)



Studies on high-dimensional data analysis techniques

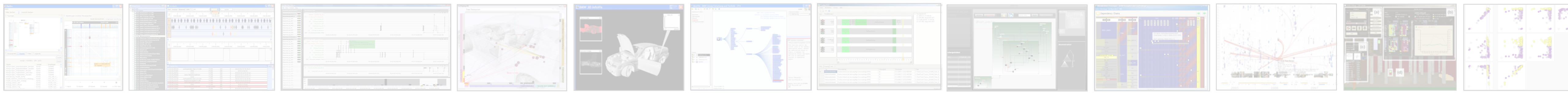




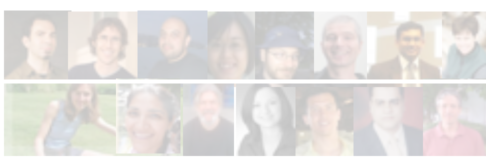


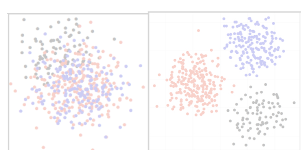
real users/real data:

Applied visualization projects (9 BMW, 3 others)



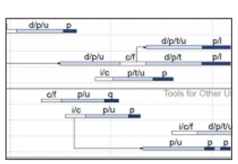
Studies on high-dimensional data analysis techniques

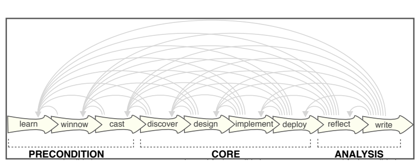


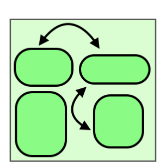


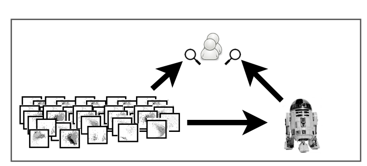
right methods:

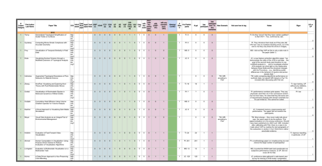
Novel and refined research methods/methodologies





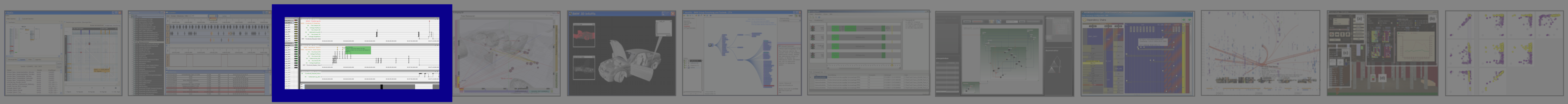






real users/real data:

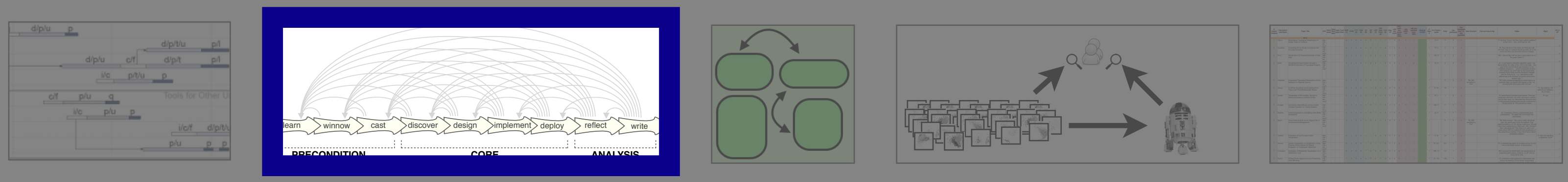
Applied visualization projects (9 BMW, 3 others)



Studies on high-dimensional data analysis techniques

right methods:

Novel and refined research methods/methodologies



real users/real data:

a project with automotive engineers

Cardiogram: Visual Analytics for Automotive Engineers

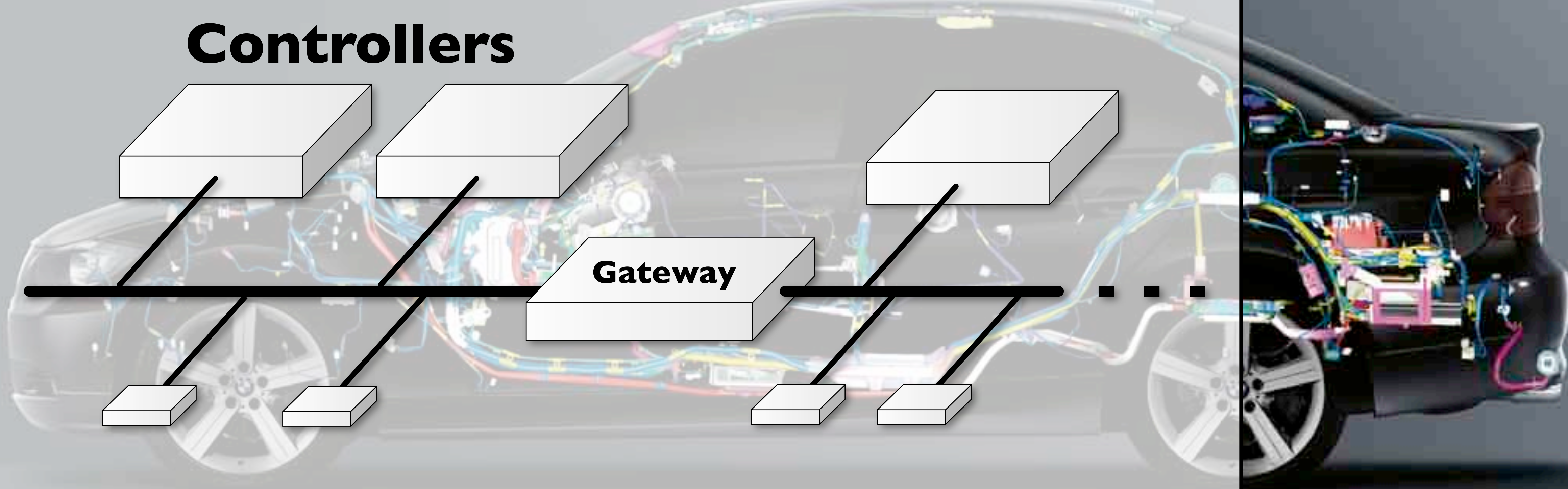
M. Sedlmair, P. Isenberg, D. Baur, M. Mauerer, C. Pigorsch, A. Butz
In Proc. ACM Conference on Human Factors in Computing Systems, p. 1727–1736, 2011, (CHI 2011).






more and more electronics...

... enabled by in-car communication networks



~70 Controllers / Car

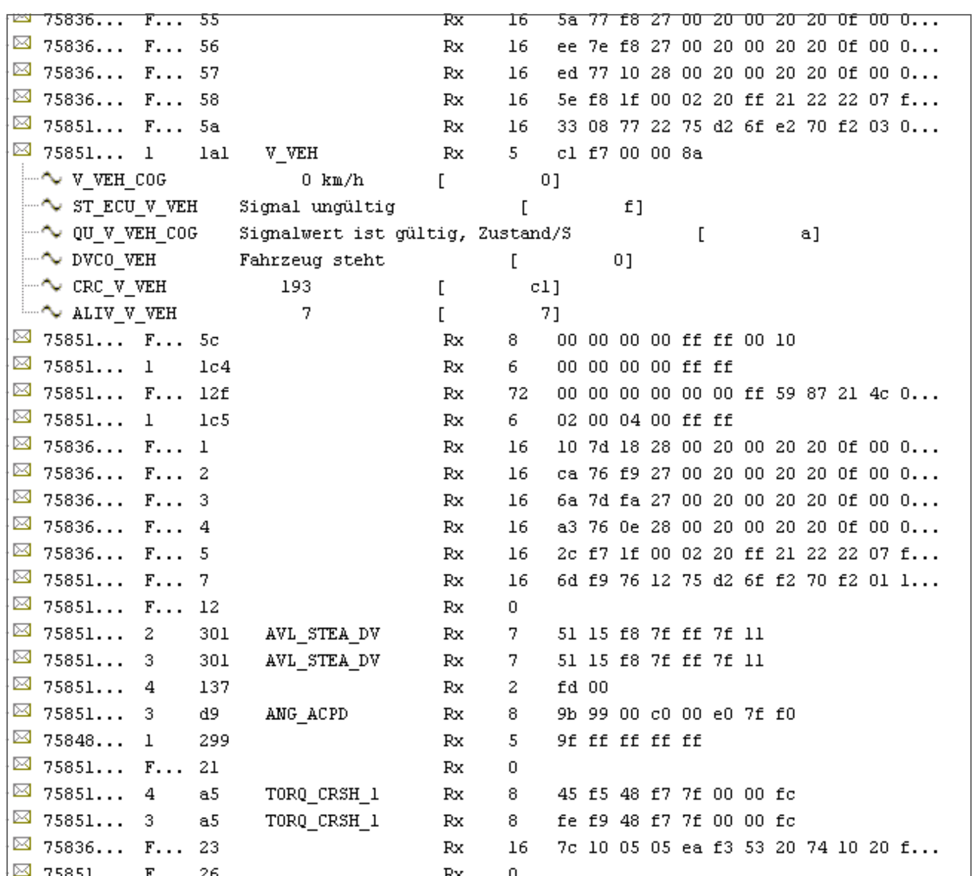
problem characterization

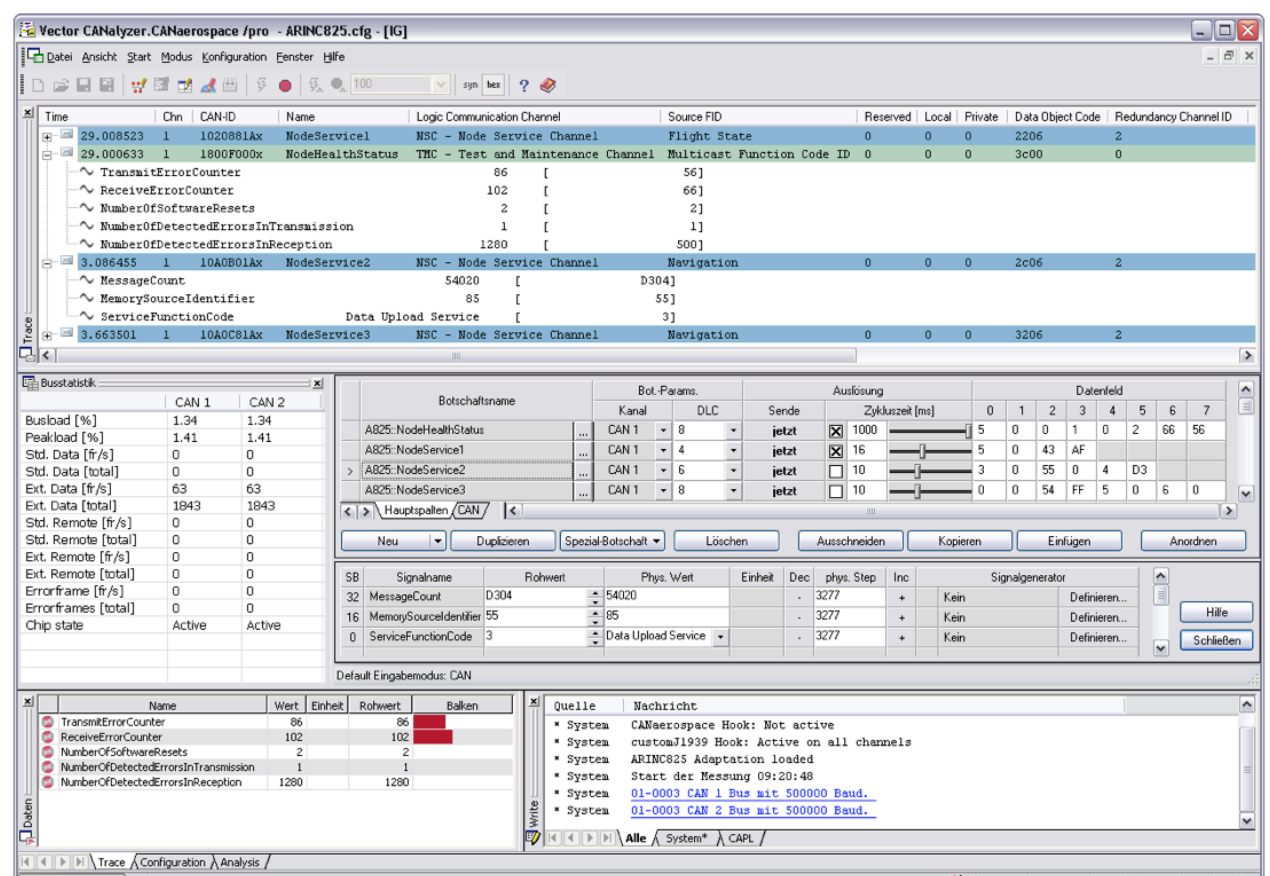
task: finding errors

process: test drives

data: recorded traces (15k messages/sec)

current practices: mainly textual lists





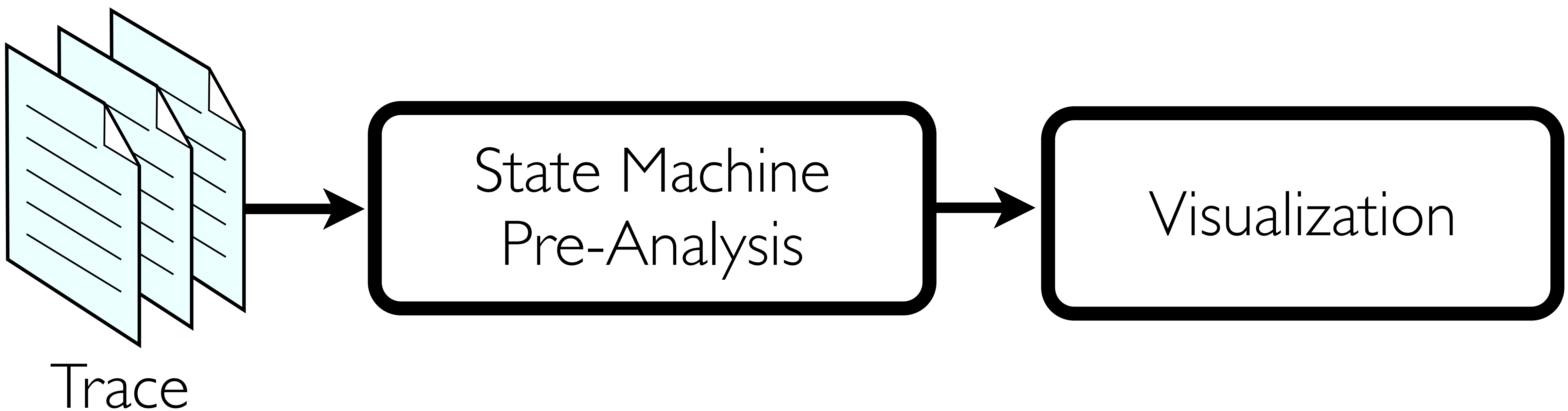
problem characterization

task: finding errors
process: test drives
data: recorded traces (15k messages/sec)
current practices: mainly textual lists

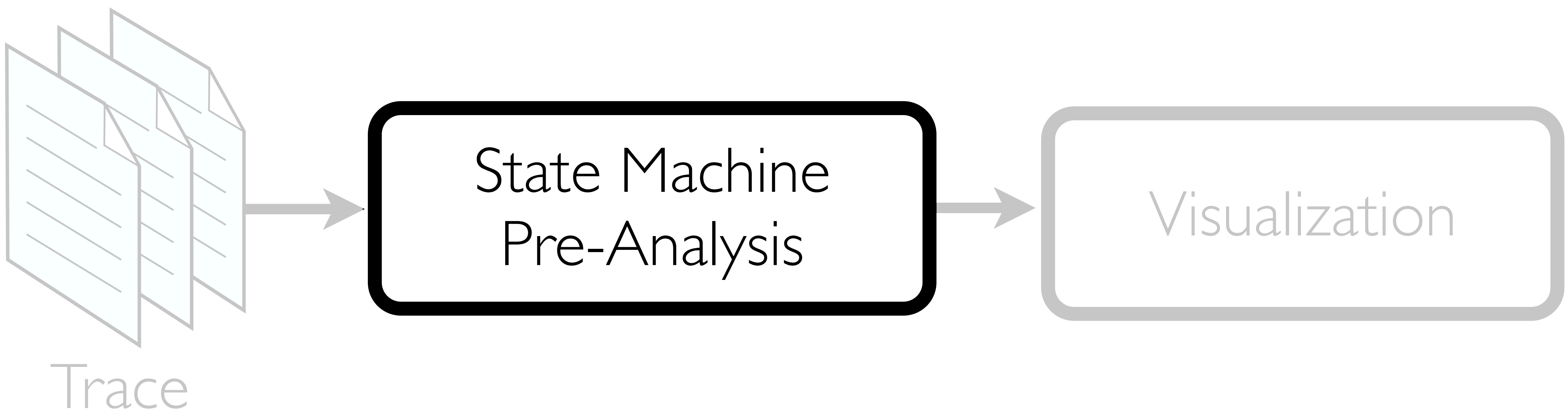
problems:

- handling masses of test traces (large and many)
- understanding correlation between trace and car behaviour
- ...

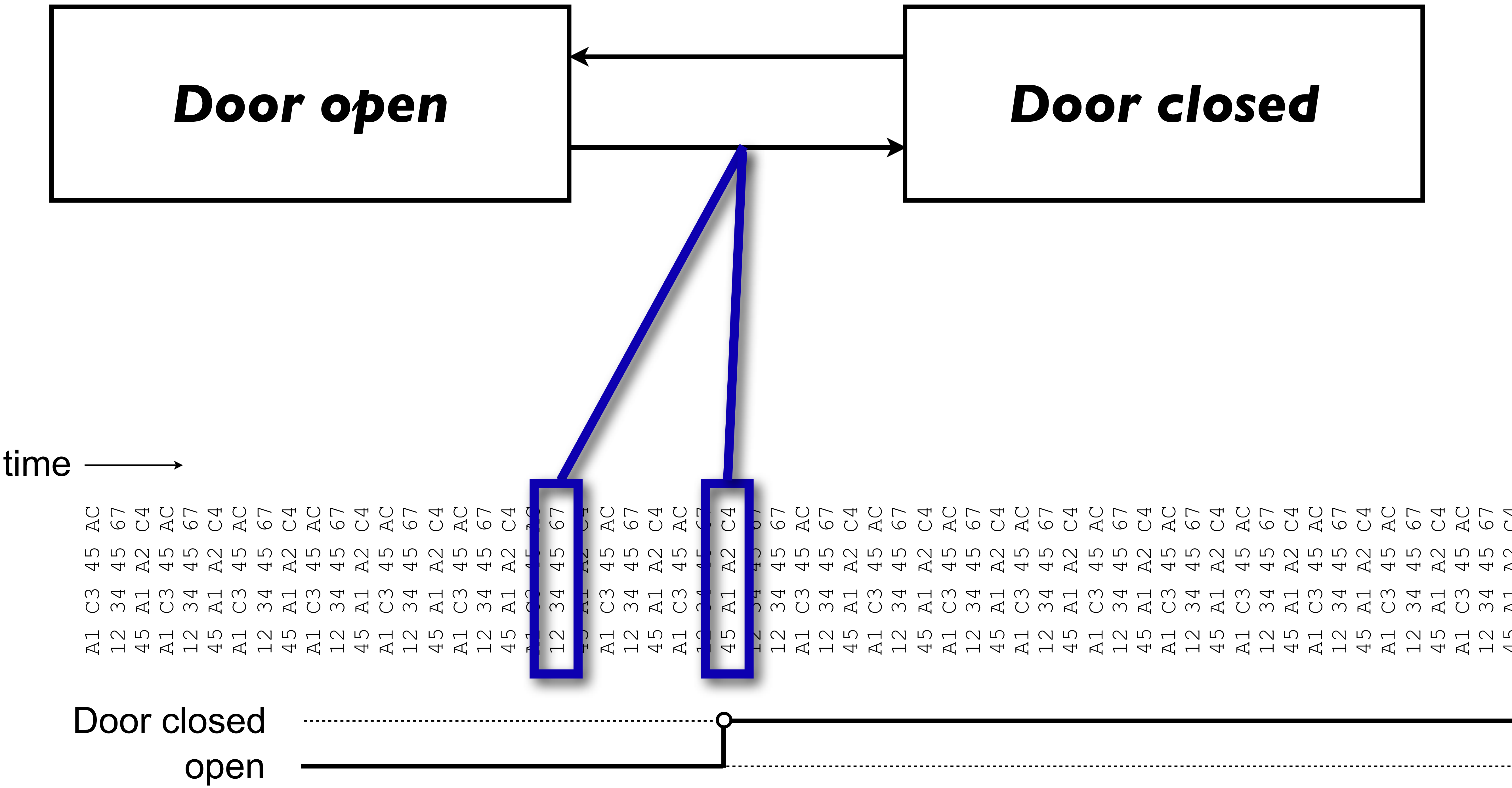
cardiogram:
overview



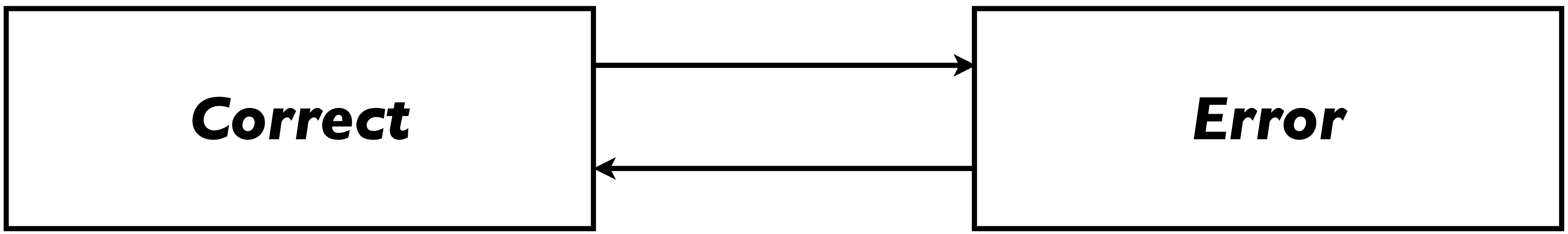
cardiogram:
state machine pre-analysis



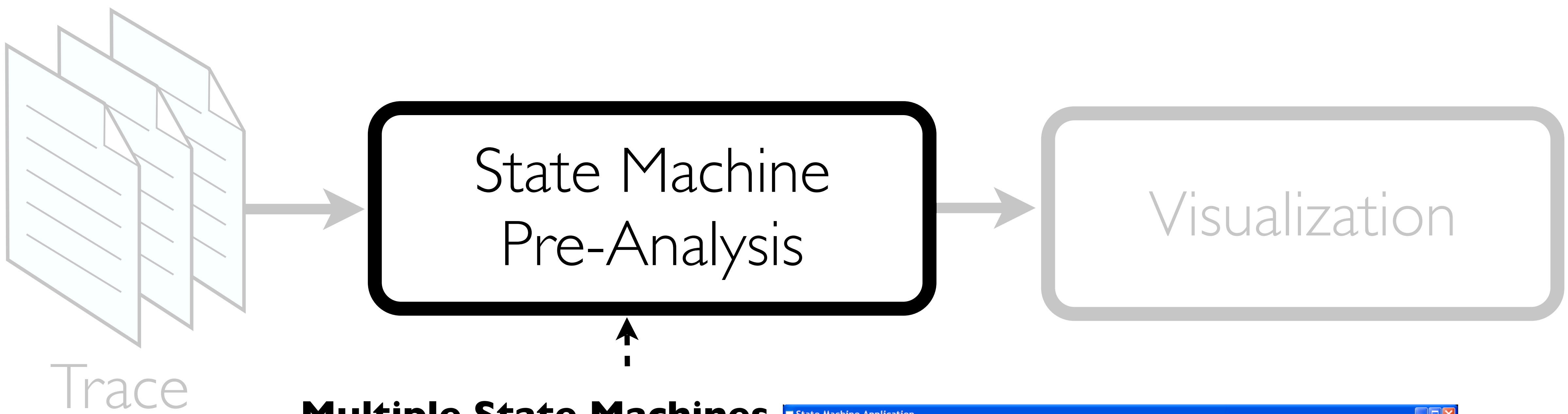
state machine:
behaviour abstraction



state machine:
automatic error detection



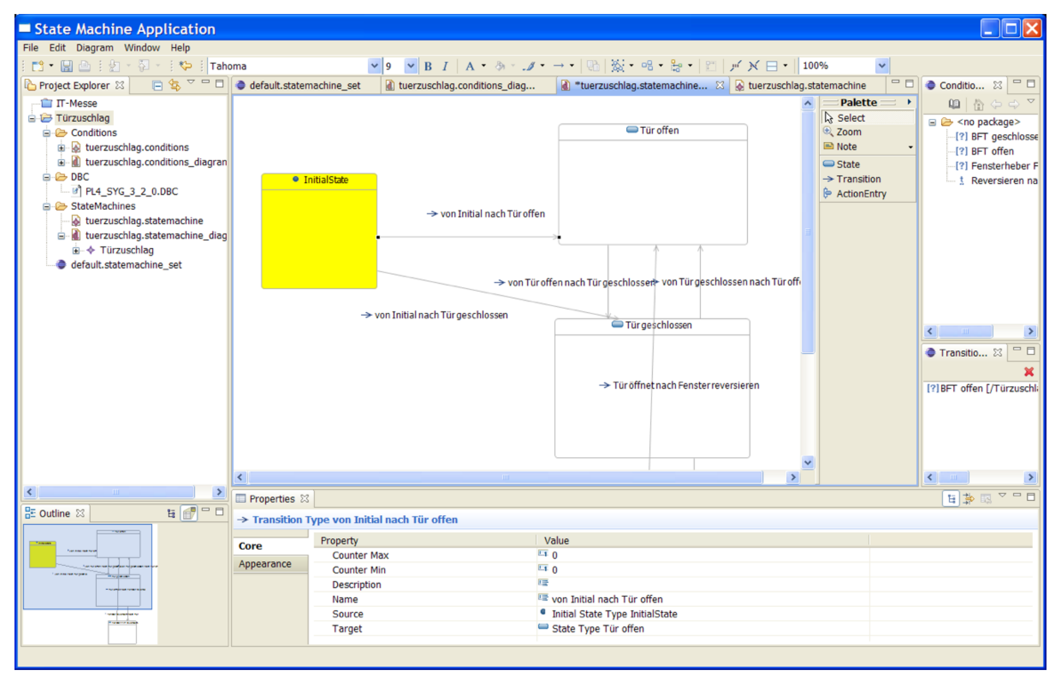
cardiogram:
create/select state machines



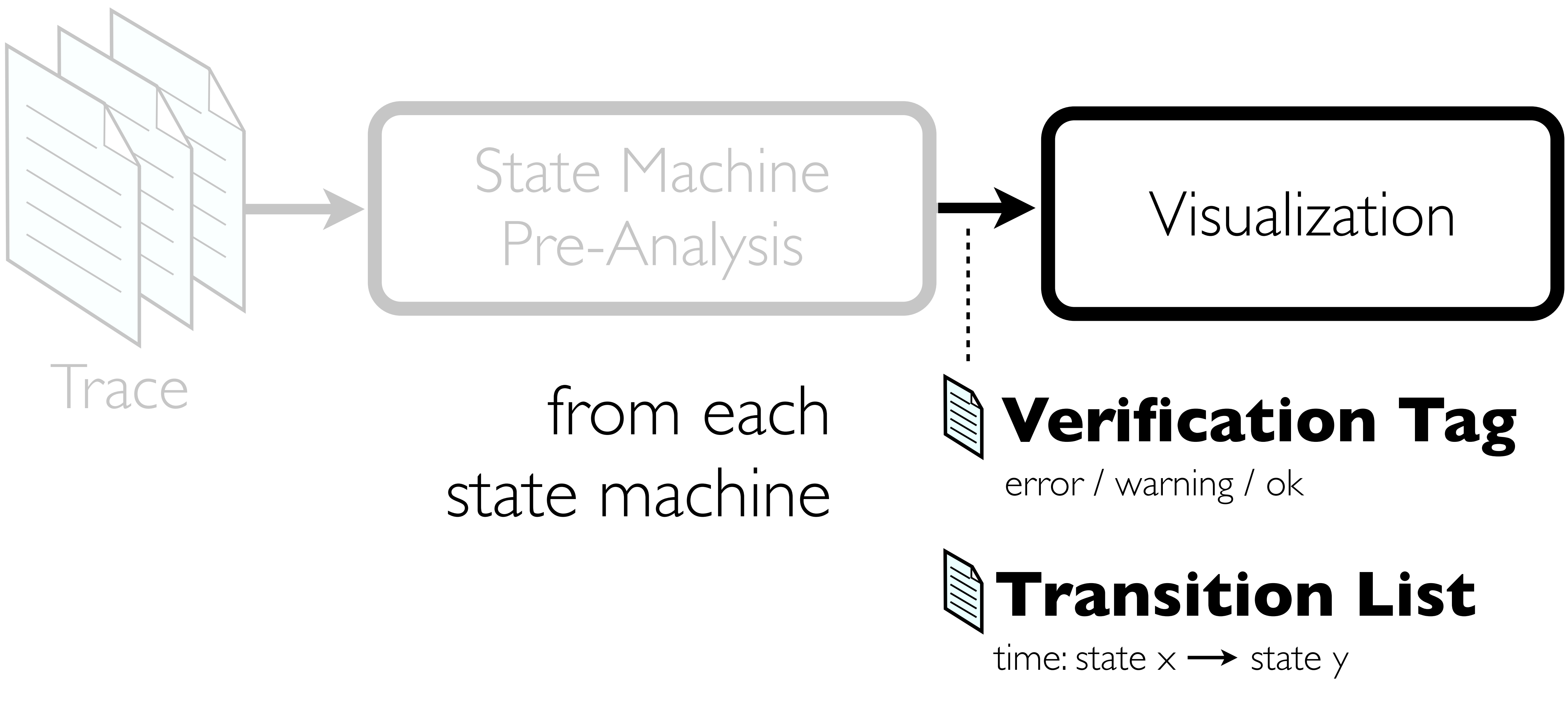
Multiple State Machines

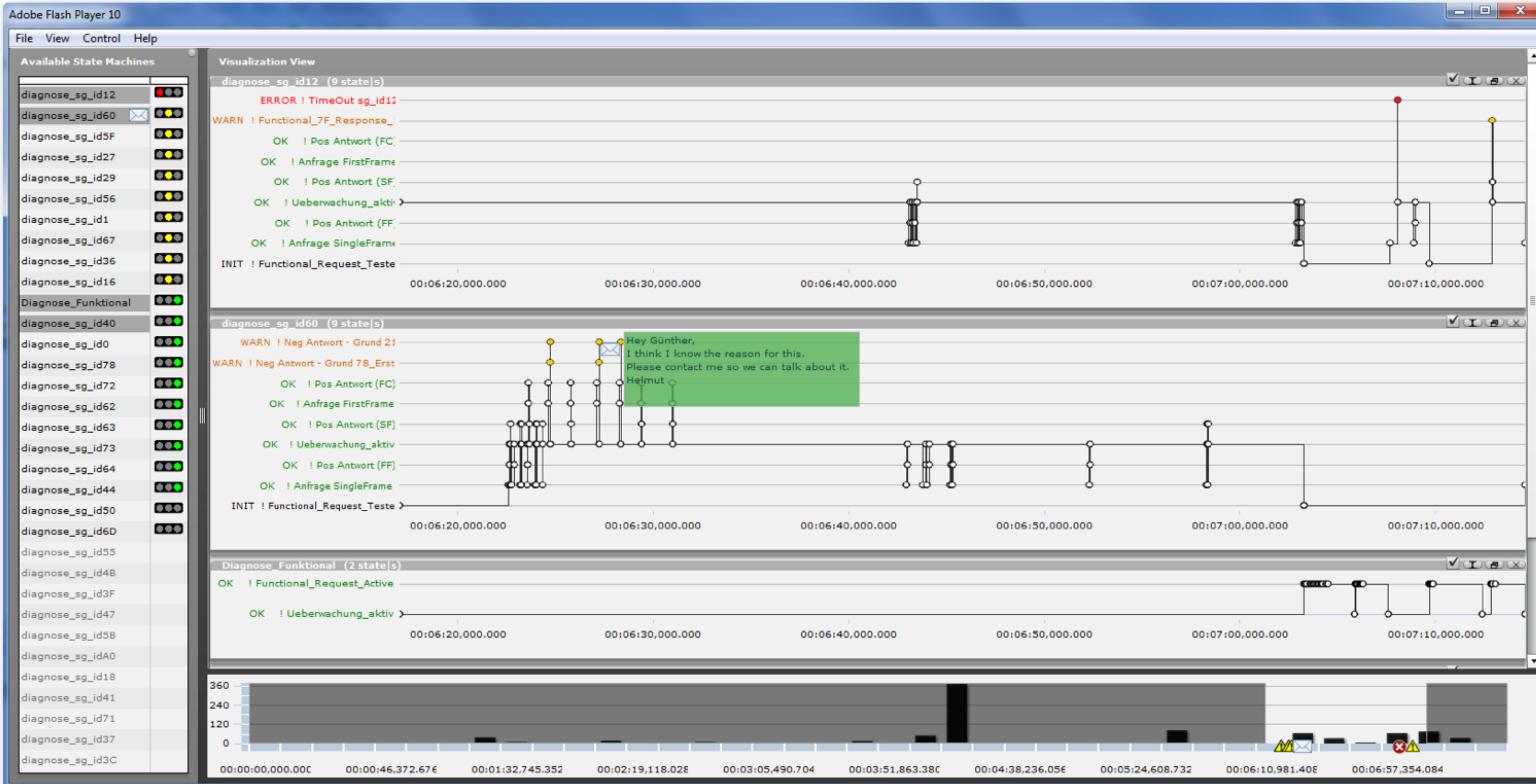
- State Machine 1
- State Machine 2
- State Machine 3

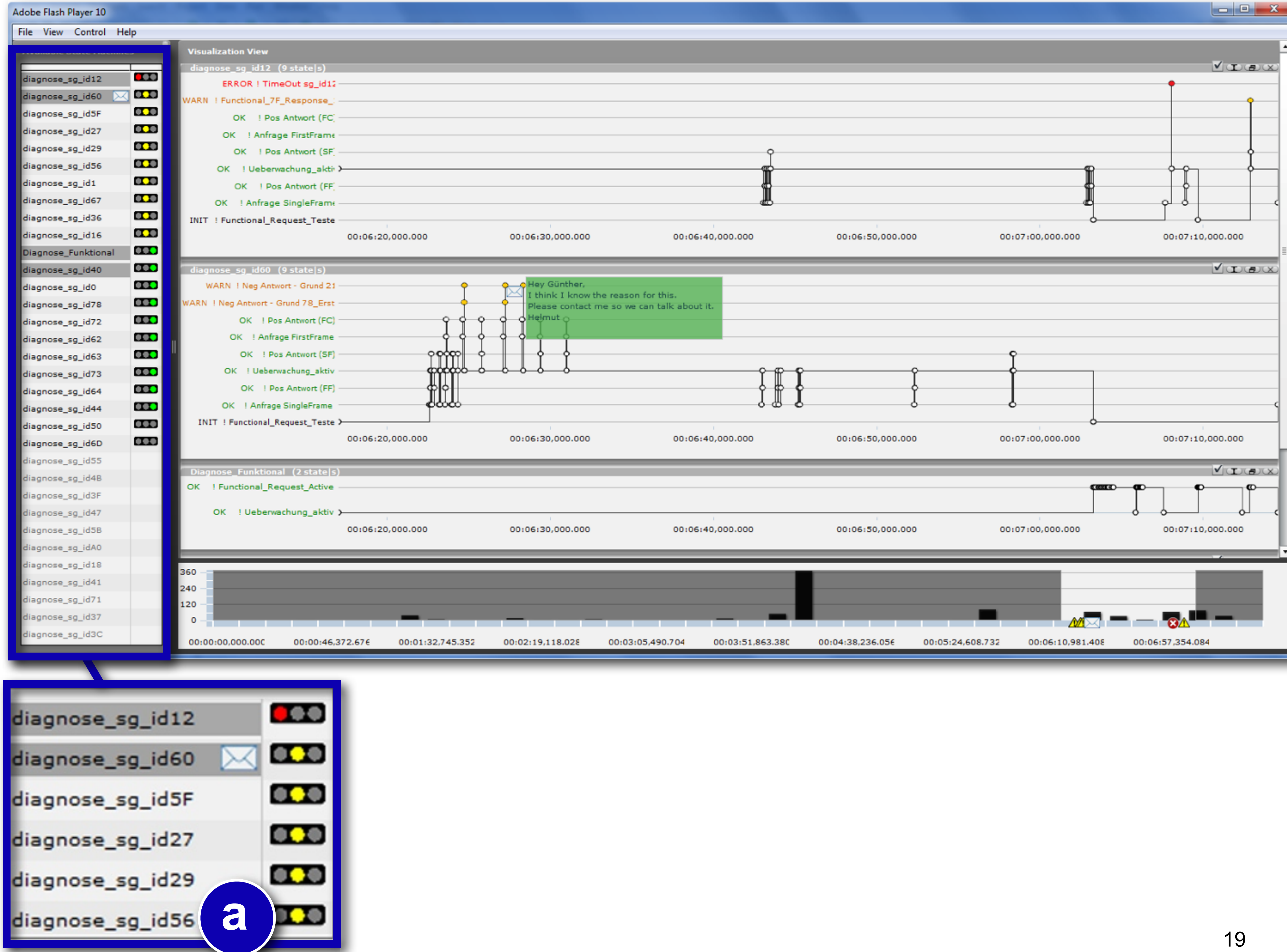
... (dozens)



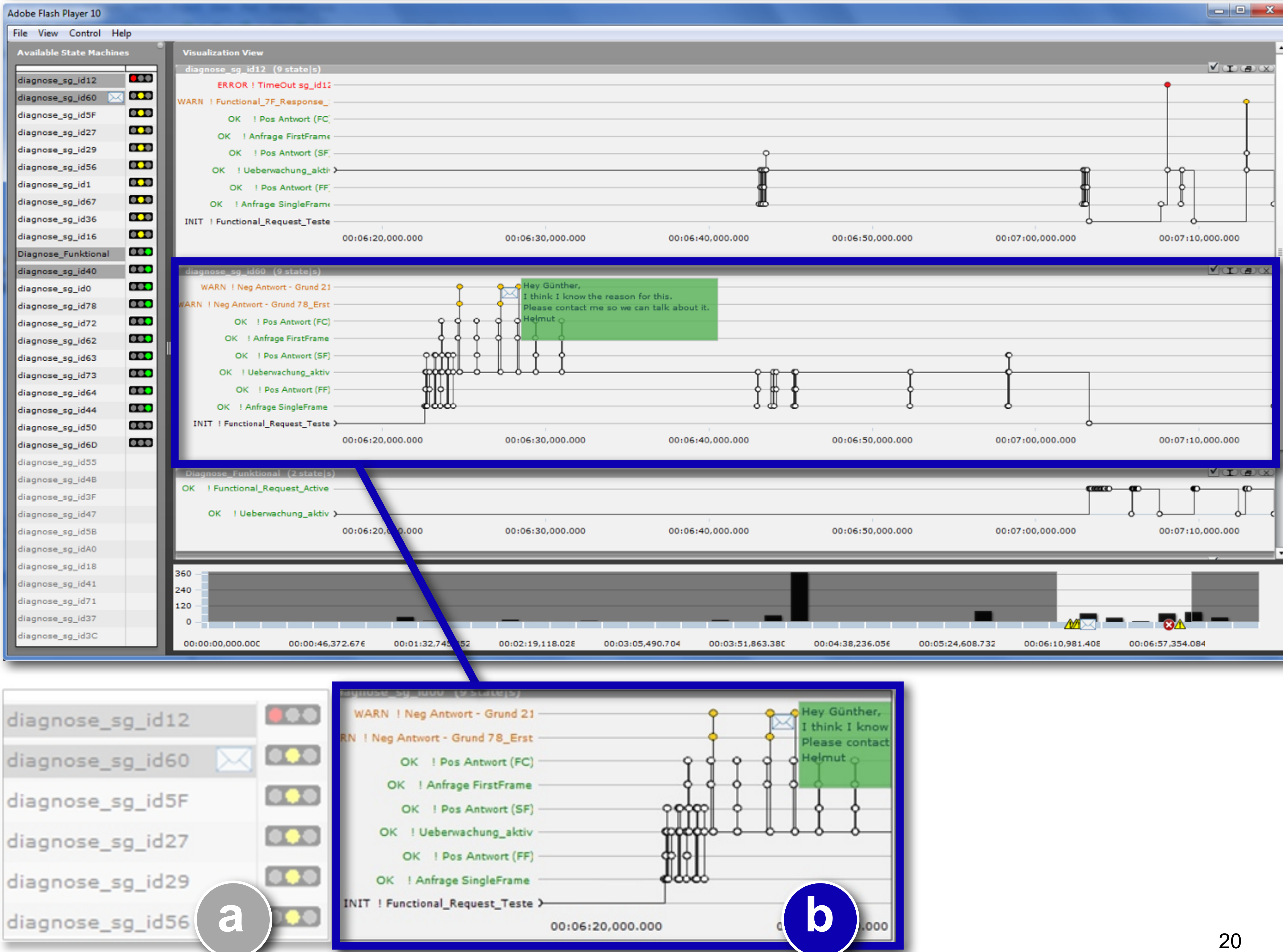
cardiogram:
- data interpretation
- data reduction

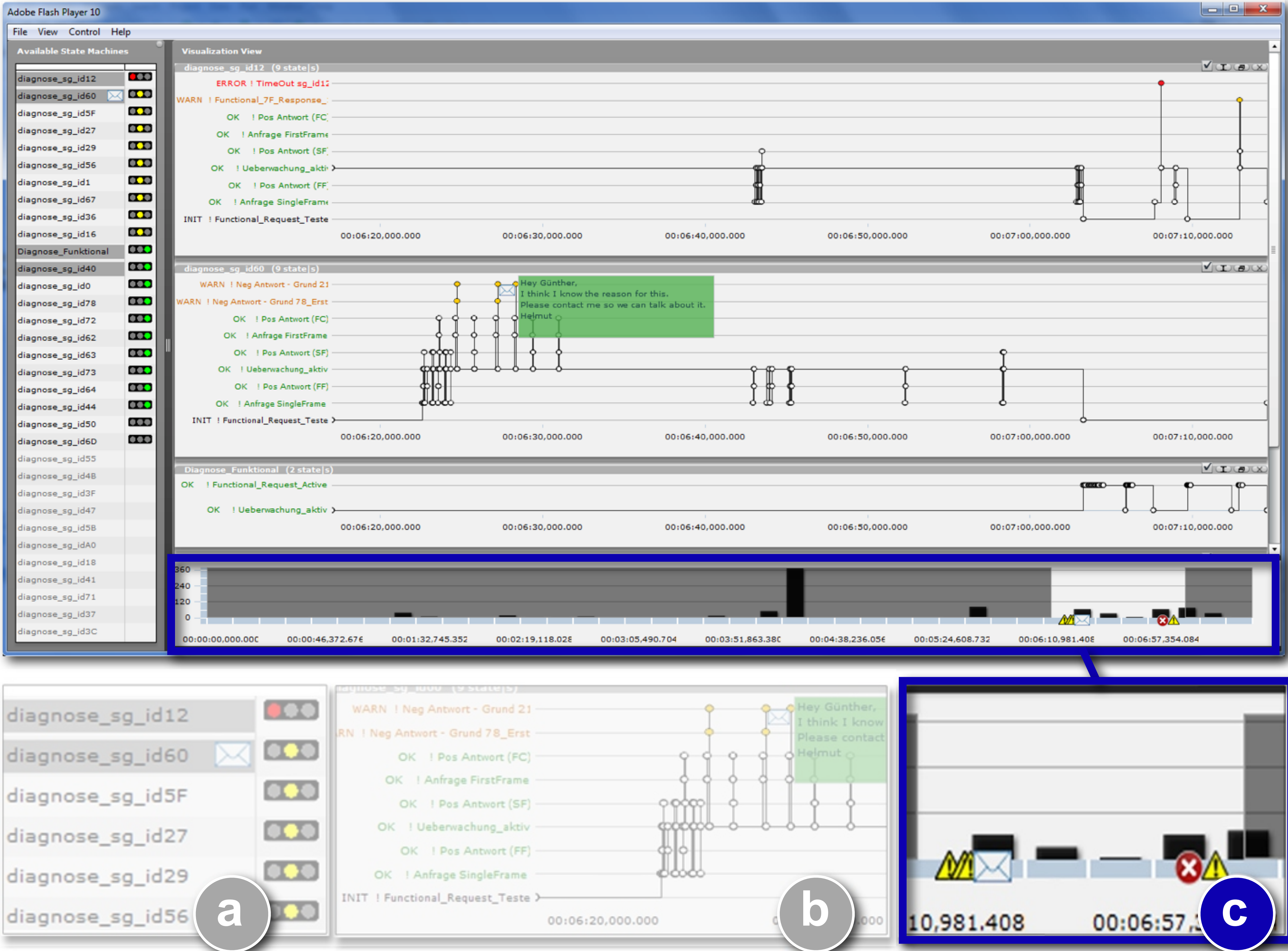



Adobe Flash Player 10
File View Control Help
Available State Machines
diagnose_sg_id12
diagnose_sg_id60
diagnose_sg_id5F
diagnose_sg_id27
diagnose_sg_id29
diagnose_sg_id56
diagnose_sg_id1
diagnose_sg_id67
diagnose_sg_id36
diagnose_sg_id16
Diagnose_Funktional
diagnose_sg_id40
diagnose_sg_id0
diagnose_sg_id78
diagnose_sg_id72
diagnose_sg_id62
diagnose_sg_id63
diagnose_sg_id73
diagnose_sg_id64
diagnose_sg_id44
diagnose_sg_id50
diagnose_sg_id6D
diagnose_sg_id55
diagnose_sg_id4B
diagnose_sg_id3F
diagnose_sg_id47
diagnose_sg_id5B
diagnose_sg_idA0
diagnose_sg_id18
diagnose_sg_id41
diagnose_sg_id71
diagnose_sg_id37
diagnose_sg_id3C
Visualization View
diagnose_sg_id12 (9 state(s))
ERROR ! TimeOut sg_id12
WARN ! Functional_7F_Response_
OK ! Pos Antwort (FC)
OK ! Anfrage FirstFrame
OK ! Pos Antwort (SF)
OK ! Ueberwachung_aktiv
OK ! Pos Antwort (FF)
OK ! Anfrage SingleFrame
INIT ! Functional_Request_Teste
00:06:20,000.000
00:06:30,000.000
00:06:40,000.000
00:06:50,000.000
00:07:00,000.000
00:07:10,000.000
diagnose_sg_id60 (9 state(s))
WARN ! Neg Antwort - Grund 21
WARN ! Neg Antwort - Grund 78_Erst
OK ! Pos Antwort (FC)
OK ! Anfrage FirstFrame
OK ! Pos Antwort (SF)
OK ! Ueberwachung_aktiv
OK ! Pos Antwort (FF)
OK ! Anfrage SingleFrame
INIT ! Functional_Request_Teste
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
Diagnose_Funktional (2 state(s))
OK ! Functional_Request_Active
OK ! Ueberwachung_aktiv
360
240
120
0
00:00:00,000.000
00:00:46,372.676
00:01:32,745.352
00:02:19,118.028
00:03:05,490.704
00:03:51,863.380
00:04:38,236.056
00:05:24,608.732
00:06:10,981.408
00:06:57,354.084



Adobe Flash Player 10
File View Control Help
Visualization View
diagnose_sg_id12 (9 state(s))
ERROR ! TimeOut sg_id12
WARN ! Functional_7F_Response_
OK ! Pos Antwort (FC)
OK ! Anfrage FirstFrame
OK ! Pos Antwort (SF)
OK ! Ueberwachung_aktiv
OK ! Pos Antwort (FF)
OK ! Anfrage SingleFrame
INIT ! Functional_Request_Teste
00:06:20,000.000
00:06:30,000.000
00:06:40,000.000
00:06:50,000.000
00:07:00,000.000
00:07:10,000.000
diagnose_sg_id60 (9 state(s))
WARN ! Neg Antwort - Grund 21
WARN ! Neg Antwort - Grund 78_Erst
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
Diagnose_Funktional (2 state(s))
OK ! Functional_Request_Active
OK ! Ueberwachung_aktiv
360
240
120
0
00:00:00,000.000
00:00:46,372.676
00:01:32,745.352
00:02:19,118.028
00:03:05,490.704
00:03:51,863.380
00:04:38,236.056
00:05:24,608.732
00:06:10,981.408
00:06:57,354.084
diagnose_sg_id12
diagnose_sg_id60
diagnose_sg_id5F
diagnose_sg_id27
diagnose_sg_id29
diagnose_sg_id56
diagnose_sg_id1
diagnose_sg_id67
diagnose_sg_id36
diagnose_sg_id16
Diagnose_Funktional
diagnose_sg_id40
diagnose_sg_id0
diagnose_sg_id78
diagnose_sg_id72
diagnose_sg_id62
diagnose_sg_id63
diagnose_sg_id73
diagnose_sg_id64
diagnose_sg_id44
diagnose_sg_id50
diagnose_sg_id6D
diagnose_sg_id55
diagnose_sg_id4B
diagnose_sg_id3F
diagnose_sg_id47
diagnose_sg_id5B
diagnose_sg_idA0
diagnose_sg_id18
diagnose_sg_id41
diagnose_sg_id71
diagnose_sg_id37
diagnose_sg_id3C
a


Adobe Flash Player 10
File View Control Help
Available State Machines
Visualization View
diagnose_sg_id12
diagnose_sg_id60
diagnose_sg_id5F
diagnose_sg_id27
diagnose_sg_id29
diagnose_sg_id56
diagnose_sg_id1
diagnose_sg_id67
diagnose_sg_id36
diagnose_sg_id16
Diagnose_Funktional
diagnose_sg_id40
diagnose_sg_id0
diagnose_sg_id78
diagnose_sg_id72
diagnose_sg_id62
diagnose_sg_id63
diagnose_sg_id73
diagnose_sg_id64
diagnose_sg_id44
diagnose_sg_id50
diagnose_sg_id6D
diagnose_sg_id55
diagnose_sg_id4B
diagnose_sg_id3F
diagnose_sg_id47
diagnose_sg_id5B
diagnose_sg_idA0
diagnose_sg_id18
diagnose_sg_id41
diagnose_sg_id71
diagnose_sg_id37
diagnose_sg_id3C
diagnose_sg_id12 (9 state(s))
ERROR ! TimeOut sg_id12
WARN ! Functional_7F_Response_
OK ! Pos Antwort (FC)
OK ! Anfrage FirstFrame
OK ! Pos Antwort (SF)
OK ! Ueberwachung_aktiv
OK ! Pos Antwort (FF)
OK ! Anfrage SingleFrame
INIT ! Functional_Request_Teste
00:06:20,000.000
00:06:30,000.000
00:06:40,000.000
00:06:50,000.000
00:07:00,000.000
00:07:10,000.000
diagnose_sg_id60 (9 state(s))
WARN ! Neg Antwort - Grund 21
WARN ! Neg Antwort - Grund 78_Erst
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
Diagnose_Funktional (2 state(s))
OK ! Functional_Request_Active
OK ! Ueberwachung_aktiv
360
240
120
0
00:00:00,000.000
00:00:46,372.676
00:01:32,745.352
00:02:19,118.028
00:03:05,490.704
00:03:51,863.380
00:04:38,236.056
00:05:24,608.732
00:06:10,981.408
00:06:57,354.084
a
b


Adobe Flash Player 10
File View Control Help
Available State Machines
Visualization View
diagnose_sg_id12 (9 state|s)
ERROR ! TimeOut sg_id12
WARN ! Functional_7F_Response_
OK ! Pos Antwort (FC)
OK ! Anfrage FirstFrame
OK ! Pos Antwort (SF)
OK ! Ueberwachung_aktiv
OK ! Pos Antwort (FF)
OK ! Anfrage SingleFrame
INIT ! Functional_Request_Teste
diagnose_sg_id60 (9 state|s)
WARN ! Neg Antwort - Grund 21
WARN ! Neg Antwort - Grund 78_Erst
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
Diagnose_Funktional (2 state|s)
OK ! Functional_Request_Active
OK ! Ueberwachung_aktiv
00:06:20,000.000
00:06:30,000.000
00:06:40,000.000
00:06:50,000.000
00:07:00,000.000
00:07:10,000.000
a
b
c

longitudinal field study:
1 year, 15 engineers

results

- externalization and sharing of expert knowledge

longitudinal field study:
1 year, 15 engineers

results

- externalization and sharing of expert knowledge
- complete coverage of traces vs. sparse sampling

longitudinal field study:
1 year, 15 engineers

results
- externalization and sharing of expert knowledge
- complete coverage of traces vs. sparse sampling
- understand behavioral correlations

longitudinal field study:

1 year, 15 engineers

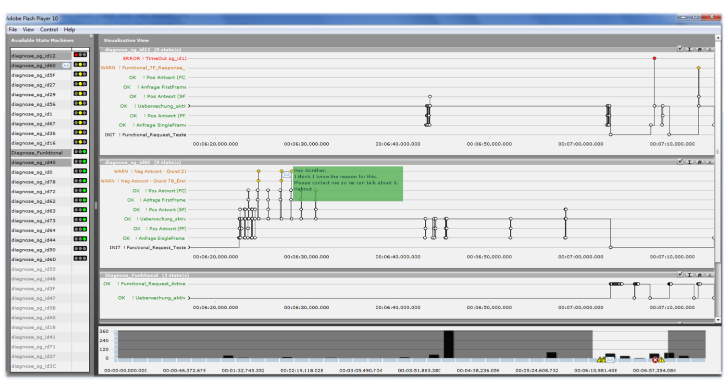
results

- externalization and sharing of expert knowledge
- complete coverage of traces vs. sparse sampling
- understand behavioral correlations

deployed and adopted

contributions

Problem and requirements analysis



Cardiogram: computational and visual analysis approach

Longitudinal evaluation and adoption

how to?

*"A **design study** is a project in which visualization researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualization design guidelines."*

[Sedlmair et al.: Design Study Methodology: Reflections from the Trenches and the Stacks, InfoVis 2012]

design studies:
long and winding road with many pitfalls!

right methods:
design study methodology

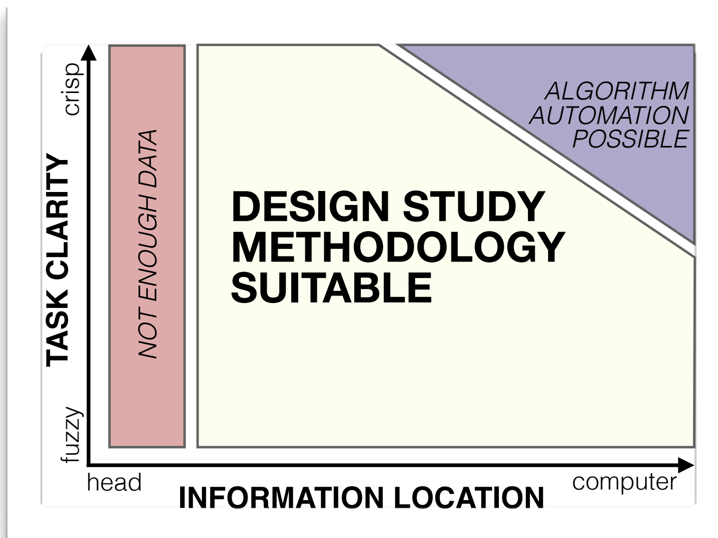
Design Study Methodology: Reflection from the Trenches and the Stacks

M. Sedlmair, M. Meyer, T. Munzner

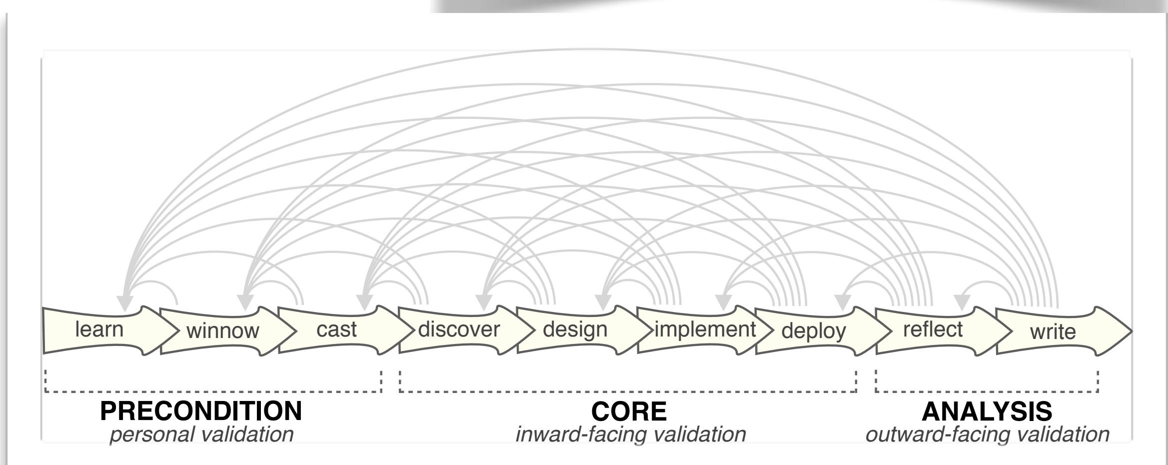
IEEE TVCG 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).
Best paper honorable mention.

main contributions
guiding other researchers with ...



Definitions



9-stage framework

PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

32 pitfalls

future work

short-term

1. New design study projects

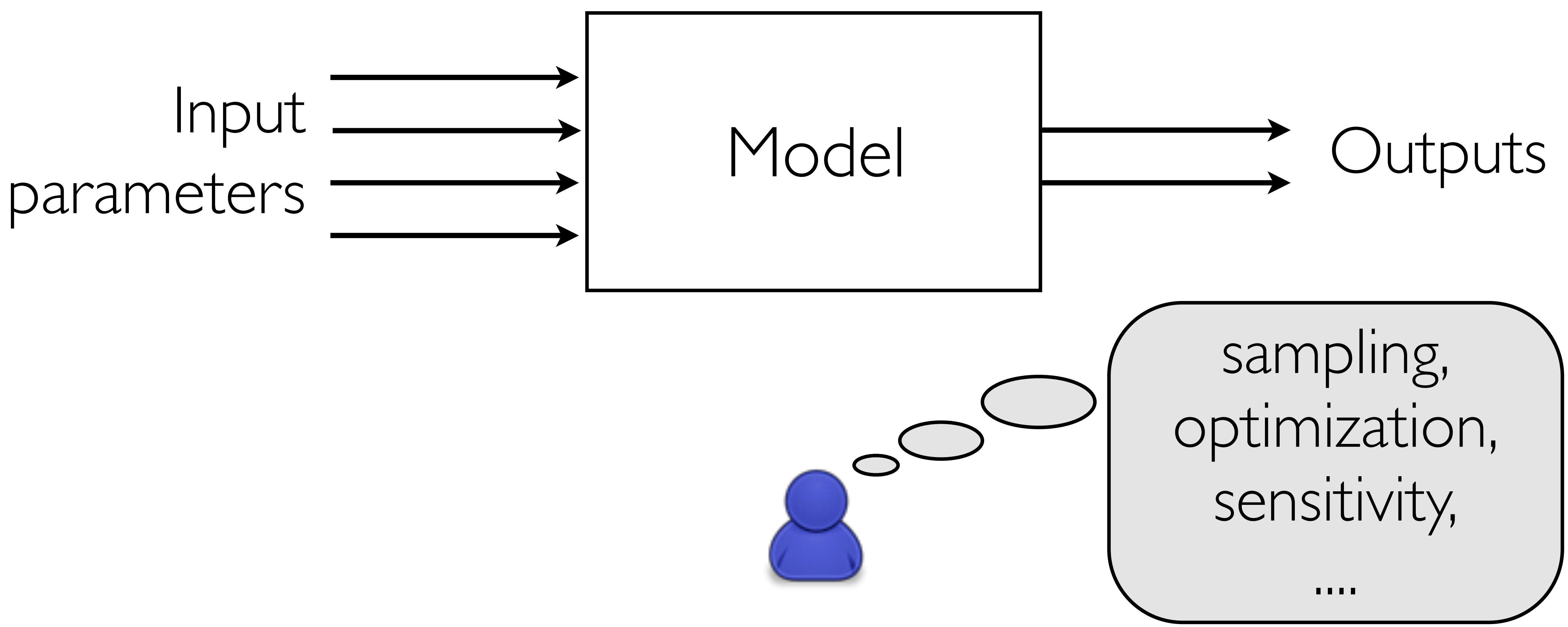
real users/real data:
- contribute to their fields

right methods:
- validate, improve, extend design study methodology

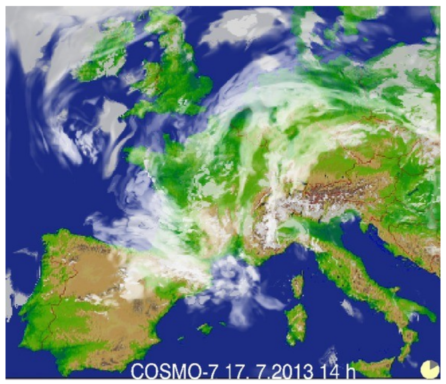
1. New design study projects

Model building, validation, usage

*[Bergner, **Sedlmair**, Möller, Abdolyousef, Saad: ParaGlide: Interactive Parameter Space Partitioning for Computer Simulations, IEEE TVCG, 2013]*

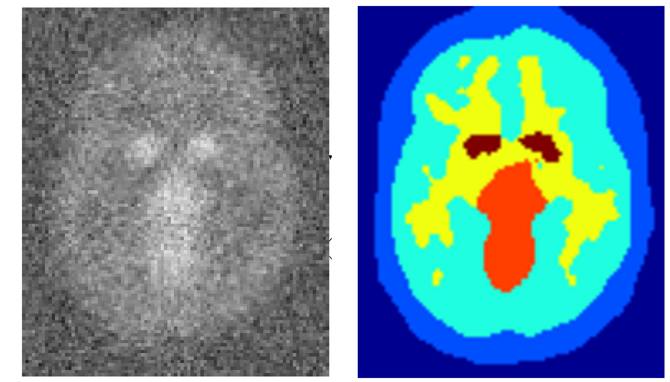


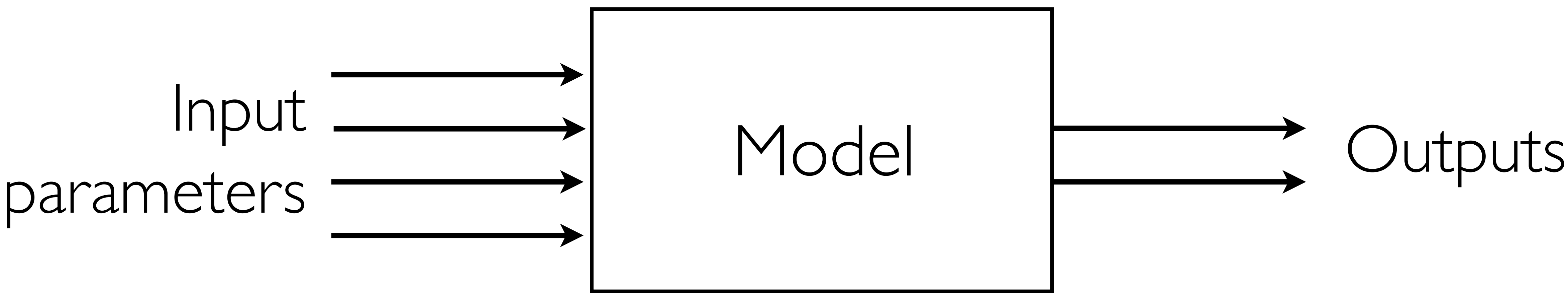
Weather prediction



Market prediction

Image Segmentation





2. High-dimensional data analysis techniques

problem-driven

technique-driven

2. High-dimensional data analysis techniques

real users, real data:

- understand techniques in real circumstances

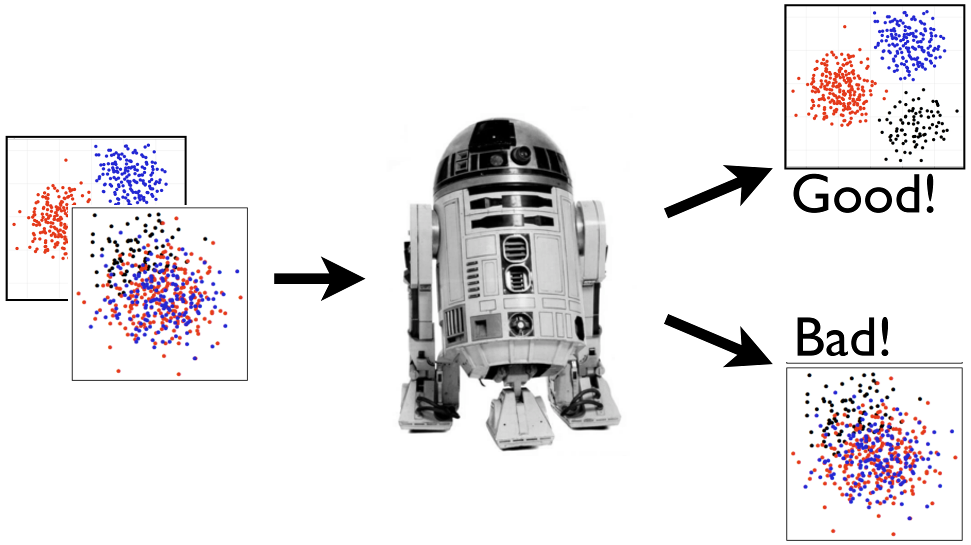
2. High-dimensional data analysis techniques

real users, real data:

- understand techniques in real circumstances

cluster separation measures

*[**Sedlmair**, Tatu, Munzner, Tory: A Taxonomy of Visual Cluster Separation Factors, EuroVis 2012]*



dimension reduction

*in submission: [**Sedlmair**, Brehmer, Ingram, Munzner: Dimension reduction in the wild, IEEE TVCG]*

clustering

...

2. High-dimensional data analysis techniques

real users, real data:
- understand techniques in real circumstances

new techniques:
- overcoming identified problems

right methods:
- adapt methods from psychology/social science/HCI

future work

long-term

Vision:

“The visualization and data analysis cookbook”

abstract data & task

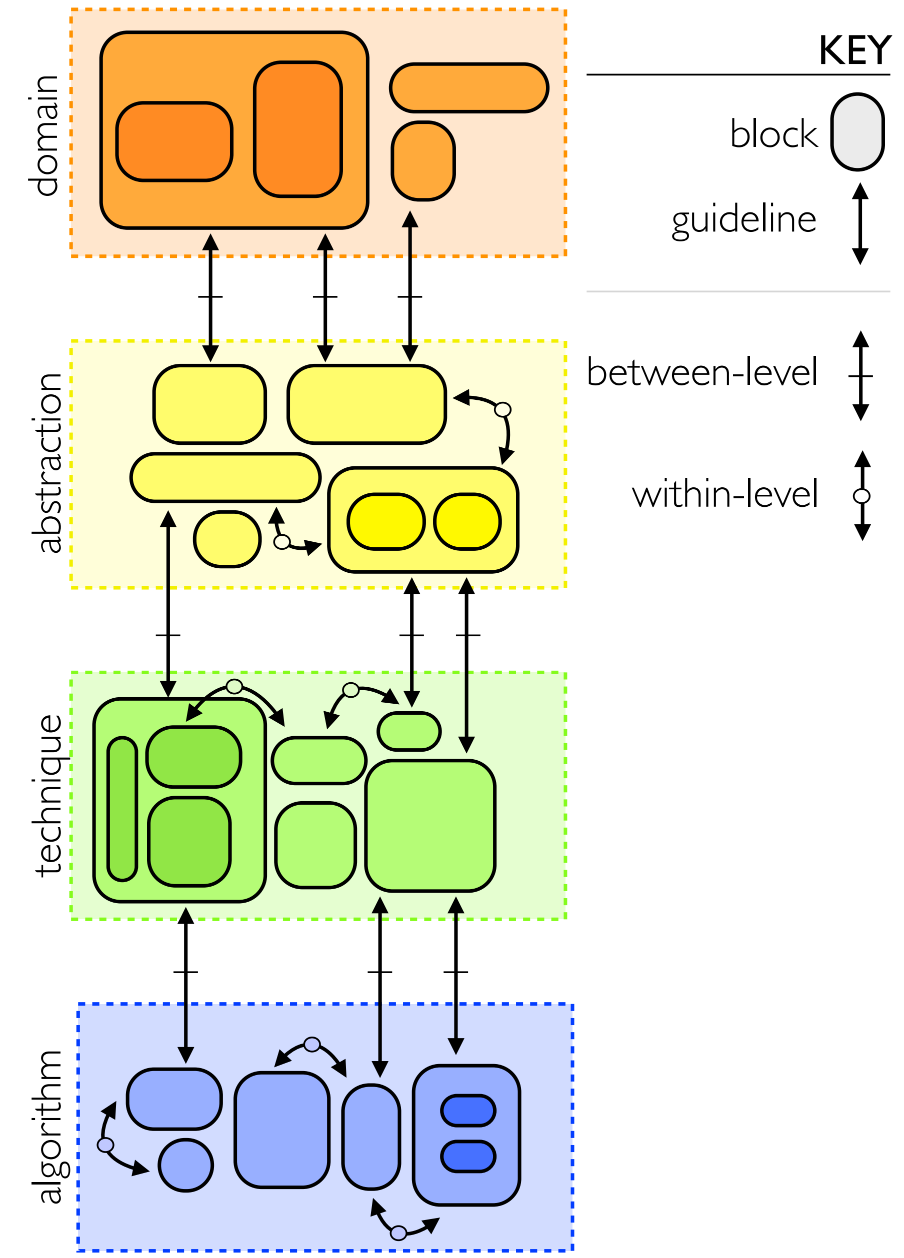
↓

data analysis technique

“The visualization and data analysis cookbook”

blocks and guidelines

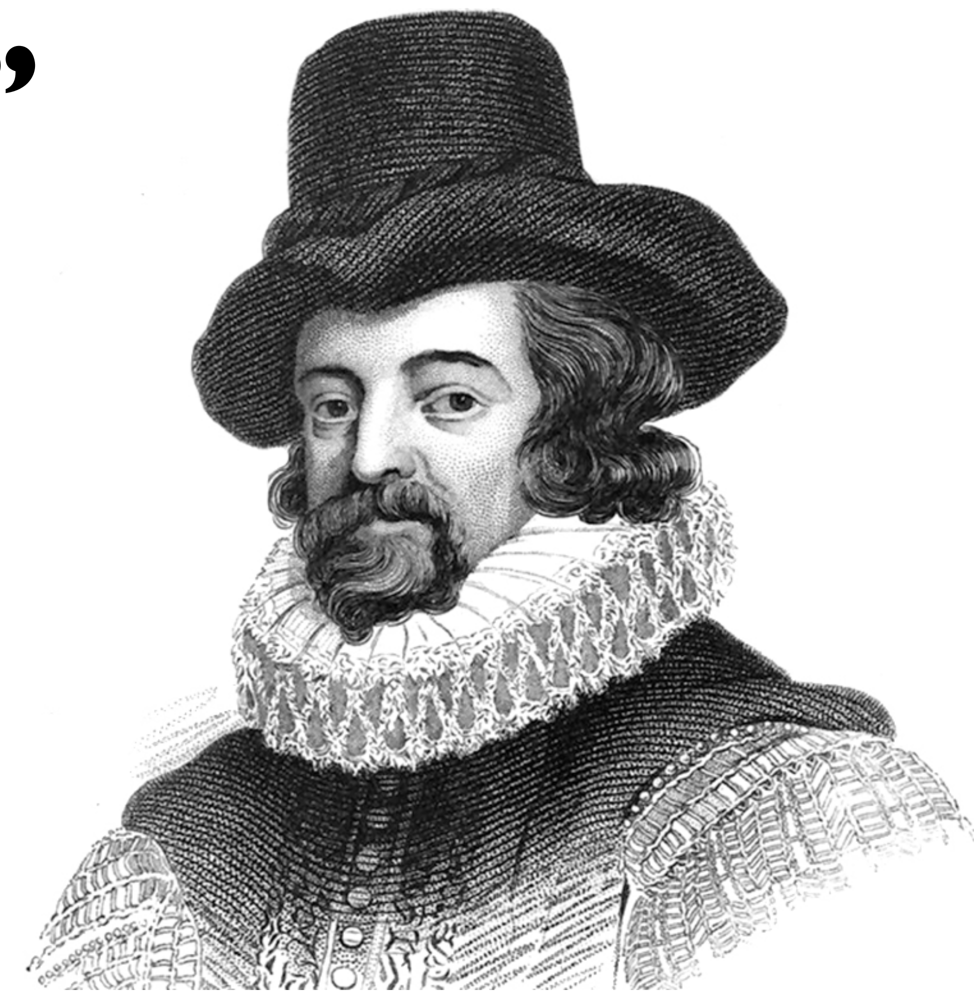
*in submission: [Meyer, **Sedlmair**, Munzner: The Nested Blocks and Guidelines Model, IVS Journal]*



“The visualization and data analysis cookbook”

vast amounts of real users/
real data work necessary

natural science
vs. design research



[Sir Francis Bacon 1561-1626]

City University

How does it fit?

very nicely ...

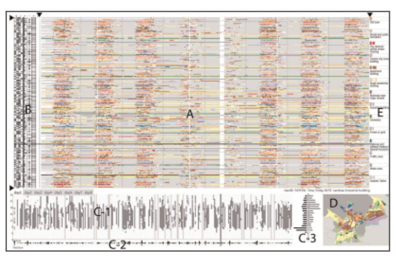
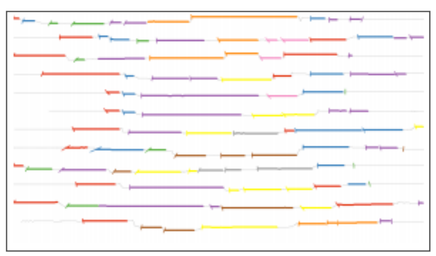
common interests

complementing expertise

common interests

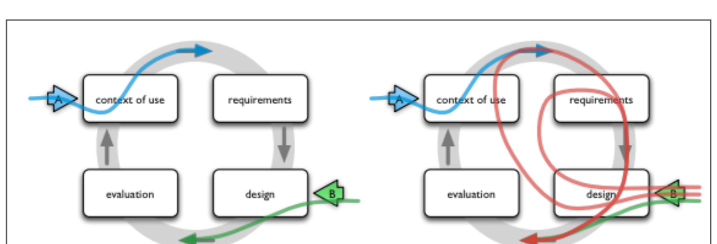
real users/real data:

- Urban planning *[Zhang et al., 2013]*
- Bird migration *[Kölzsch et al., 2013]*
- Voting data *[Wood et al., 2011]*


right methods:

- Human-centered approaches *[Lloyd & Dykes et al., 2011]*
- Designer=User *[Dillingham et al., 2012]*



complementing expertise

City / giCentre	me
background	
Geoscience/ Cartography	Computer Science/ HCI
data	
spatial temporal	network / traces high-dimensional
partners	
Government, Willis, ...	BMW, Google, AT&T, ...

very nicely ...

common interests

complementing expertise

⇨ ample collaboration opportunities

⇨ strengthening City's world-leading role in problem-driven visualization and data analysis research

Teaching

(a) struggle points

(b) approaches

based on ...

Teaching @university:
- HCI undergrad course, University of Vienna
- various guest lectures in Visualization

Supervision:
12 grad / 6 undergrad

Tutoring:
- math *(private, 5 years, ~15 students, 1-3h/week)*
- guitar *(school, 7 years, ~50 students, 15h/week)*

CS student

Literature

general motivation

Diversity in previous knowledge of students
- different education
- spare-time activities
- diverse cultural backgrounds

general motivation

Diversity in previous knowledge of students
- different education
- spare-time activities
- diverse cultural backgrounds

⇨ **different struggle points**

Design Thinking / Creativity

general struggle point:
How to go about designing a tool for real users?

Math / Algorithms

general struggle point:
Understanding and using math for data analysis?

why these two aspects?

Design Thinking / Creativity

Math / Algorithms

crucial combination to be successful as visual data scientist

traditionally, either good at one or the other (my experience)

approach for design thinking

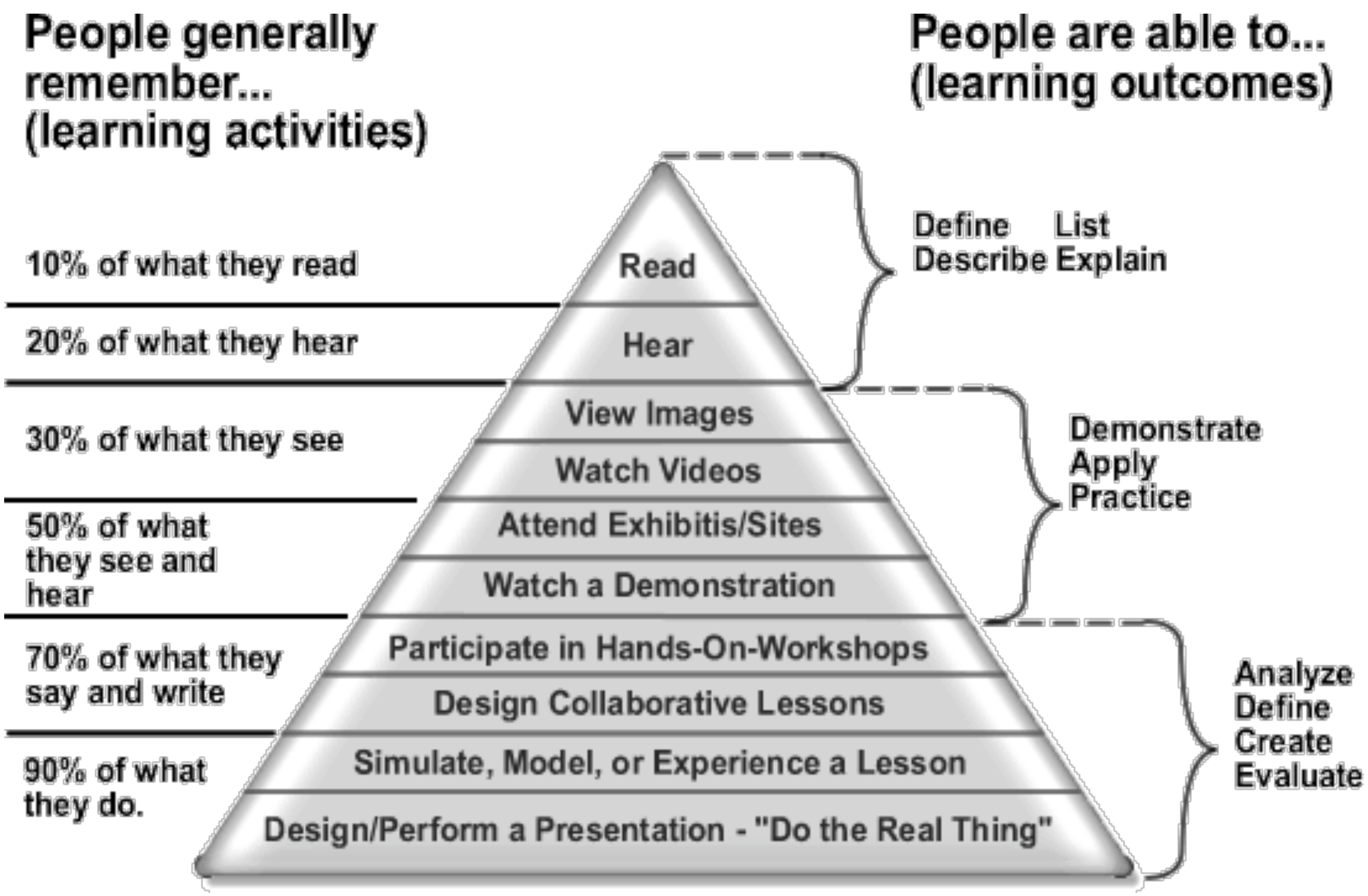
Design Thinking / Creativity

Math / Algorithms

How to go about designing a tool for real users?

Interactive Classroom Experience

interactive classroom experience



Dale's cone of learning

http://en.wikipedia.org/wiki/File:Edgar_Dale%27s_cone_of_learning.gif

multi-faceted mode
teaching and assessment

- team projects with milestones
- single exercises
- student presentations in class
- teacher presentations in class
- rich interactions
- in-class hands-on activities

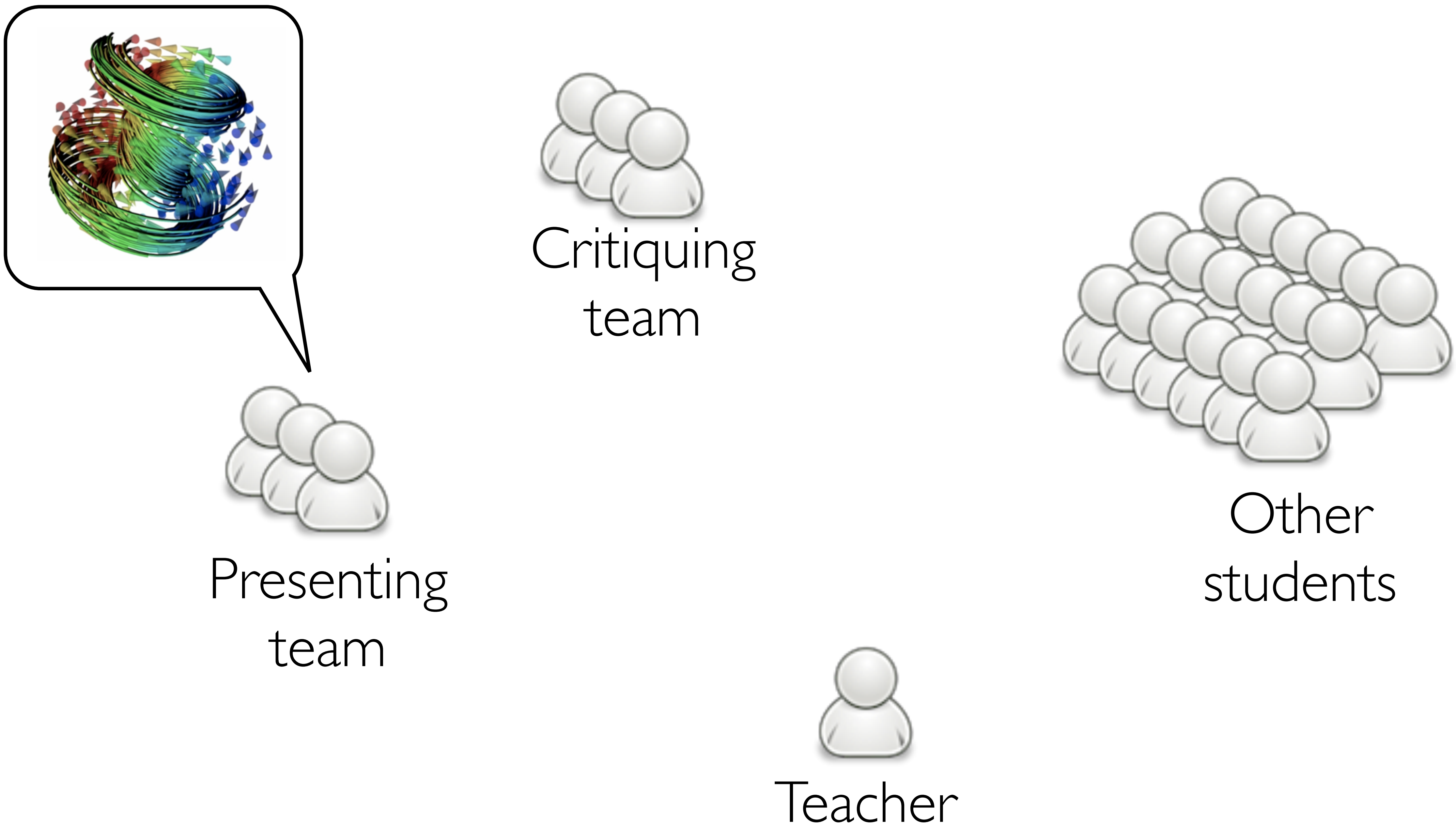
example:
two specific struggle points

Design critiques:
How to evaluate early design ideas?

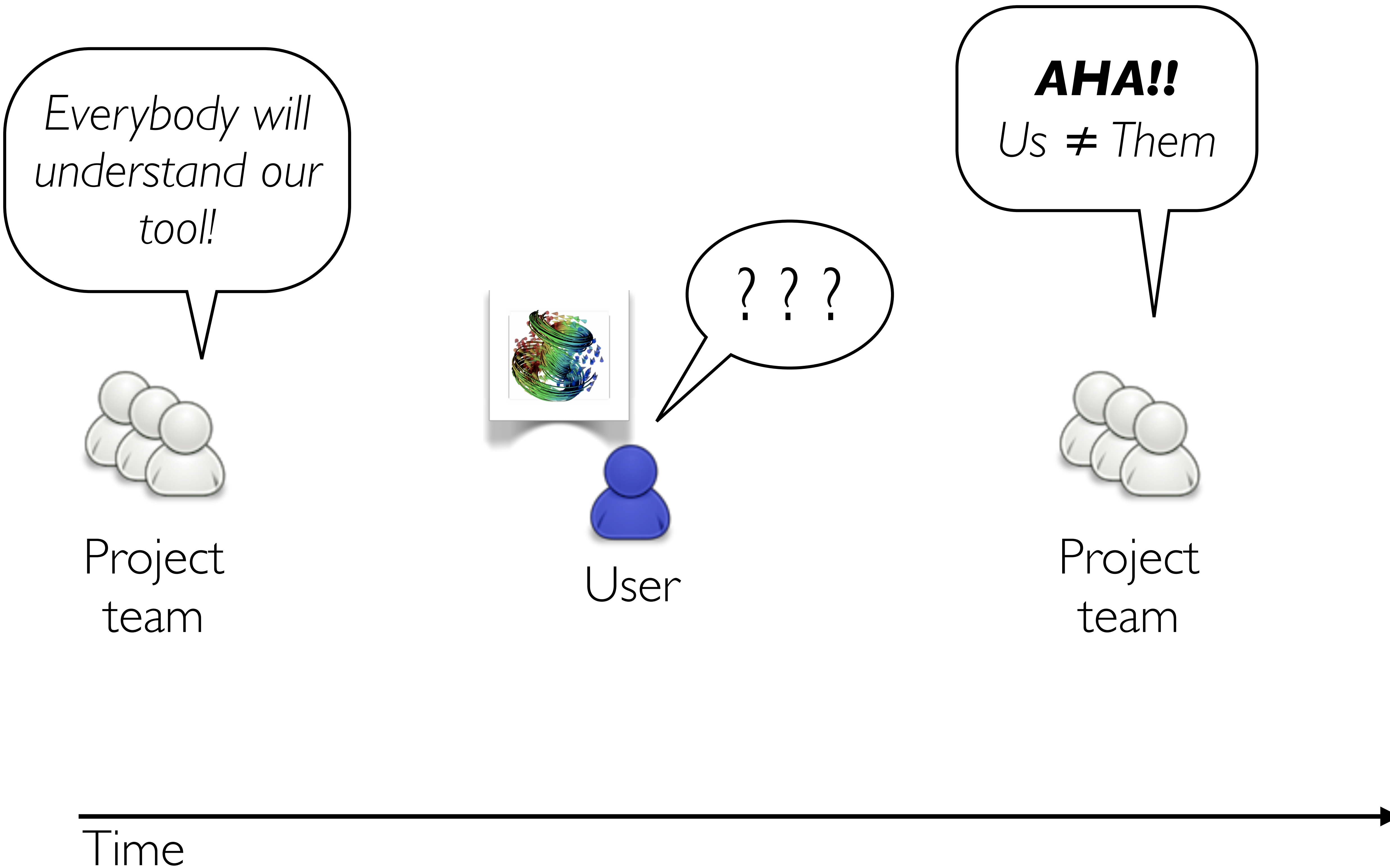
User Testing:
Understand the importance of user testing!

example:
in-class design critiques

[Hundhausen et al.: An Empirical Study of the "Prototype Walkthrough": A Studio-Based Activity for HCI Education, ACM CHI 2013]



example:
user testing



challenges:
team work

e.g., weak students "piggybacking" on strong students

Potential solutions:
- student presentations
- additional single exercises
- concluding team colloquia

approach for math/algorithms

Design Thinking / Creativity

Math / Algorithms

How to go about designing a tool for real users?

Understanding and using math for data analysis?

Interactive Classroom Experience

Providing Mental Models

providing mental models

... clear, memorable ways to think about complex phenomena.

For instance, by using **visual representations** for explaining data analysis algorithms.

Specific struggle point:
Understanding, using, and programming the centroid class separation metric!

example:

centroid metric

Hard to understand from math alone!

"Given a data space $X \subseteq R_n$ and a class structure $C(X)$ defining m classes. Let c_i be a class and $centr(c_i)$ its centroid, and let x be $x \in X$ with $clabel(x) = i$. **CD** describes the property of class members that the distance $d(x, centr(ci))$ to its class centroid should be always minimal in comparison to the distance to all other centroids, thus

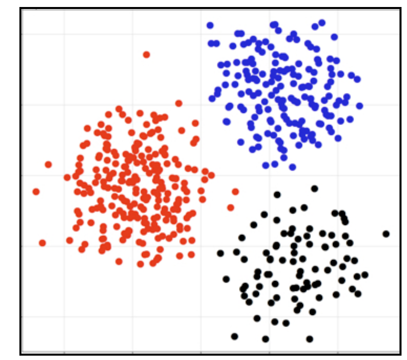
$$d(x, centr(c_i)) < d(x, centr(c_j)) \forall j : 1 \leq j \leq m; j \neq i \qquad (1)$$

and d denotes a metric defined in X. $\mathbf{CD}(x, centr(c_i)) = true$ denotes that the centroid property for x and its centroid $centr(c_i)$ is fulfilled." [SNLH09]

"Let $X \subseteq R_n$ be a n-D data set with k data points. Let $C(X)$ be a class structure of X defining m classes $C(X) = \{c_1, ..., c_m\}$. Let c_i be a class and $centr(c_i)$ its centroid in $C(X)$. Let $clabel(x)$ be the class label of a point $x \in X$. Let $v(X)$ be a 2-D view of X, then distance consistency $DSC(v(C))$ is defined as the classification error

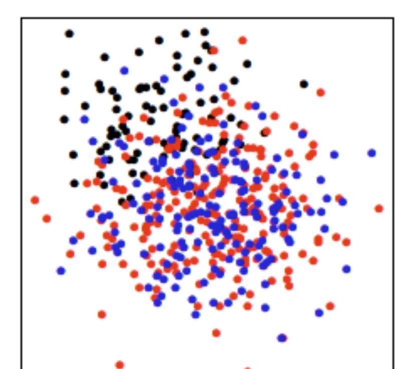
$$\mathbf{DSC} = \frac{|x' \in v(X) : \mathbf{CD}(x', centr'(c_{clabel(x)})) \neq true|}{k} \qquad (2)$$

with x' is the 2-D projection of the data point x and $centr'(c_i)$ is the 2-D projection of the centroid of class c_i." [SNLH09]

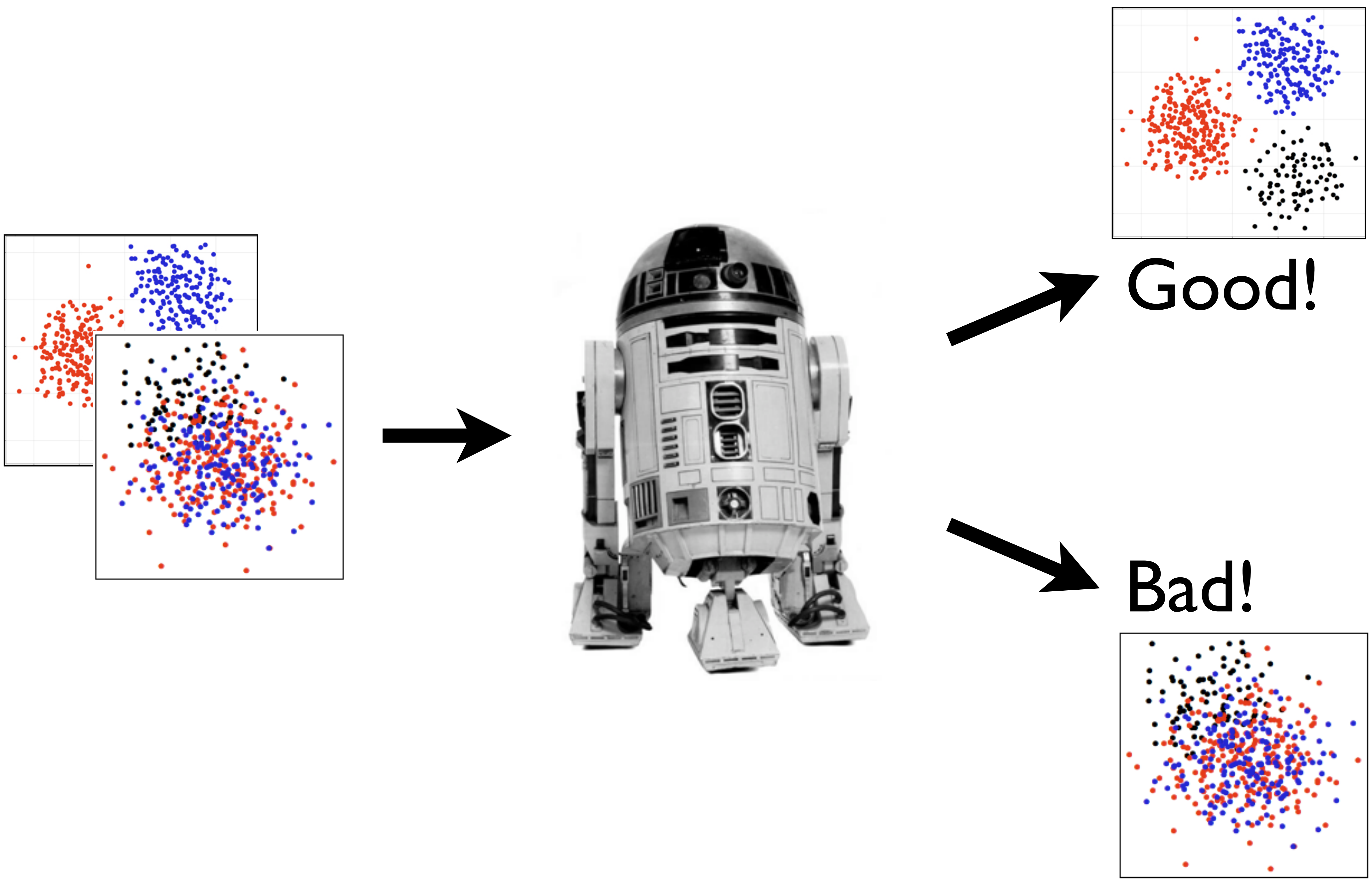


Good!

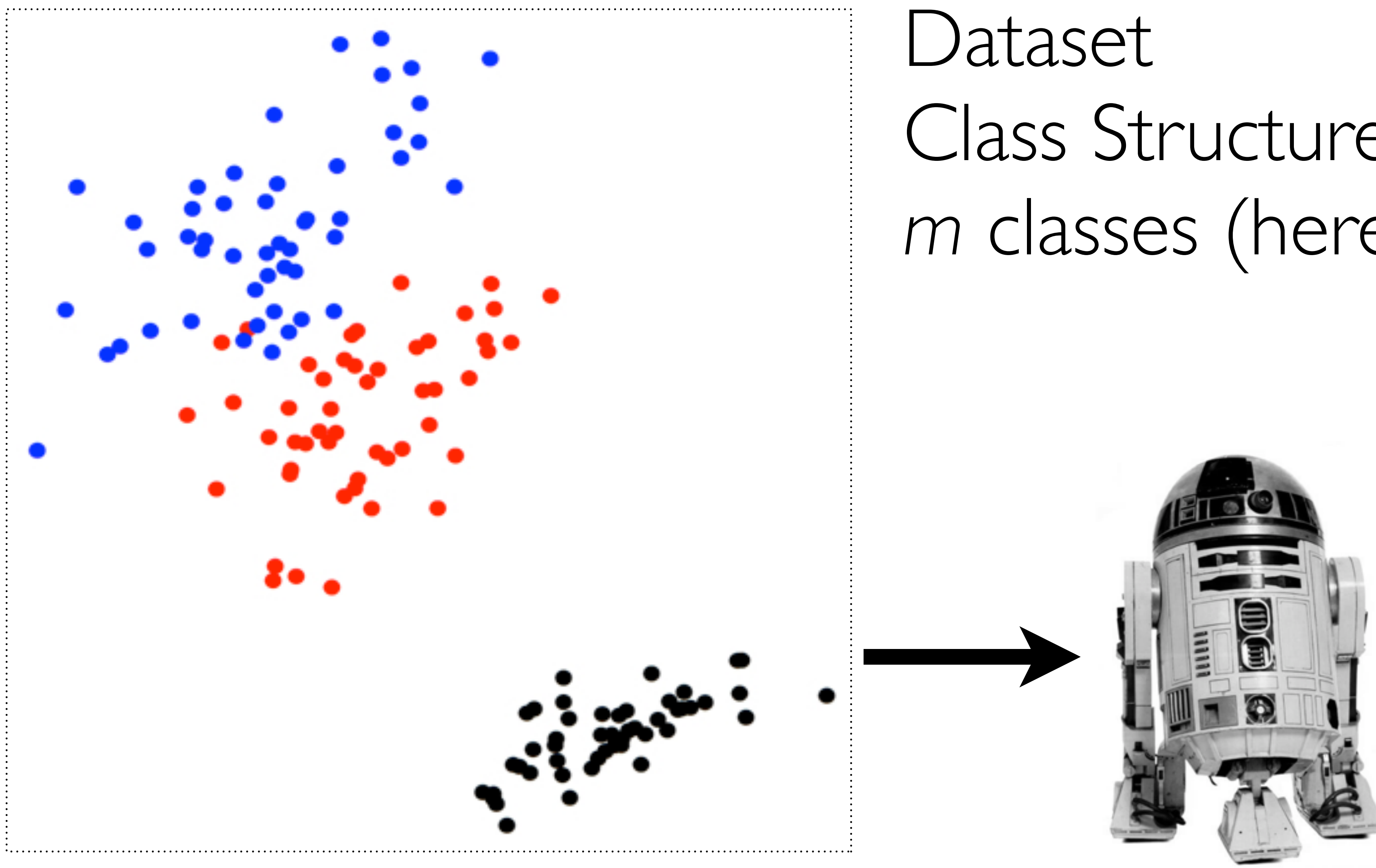
Bad!



‘the algorithm’

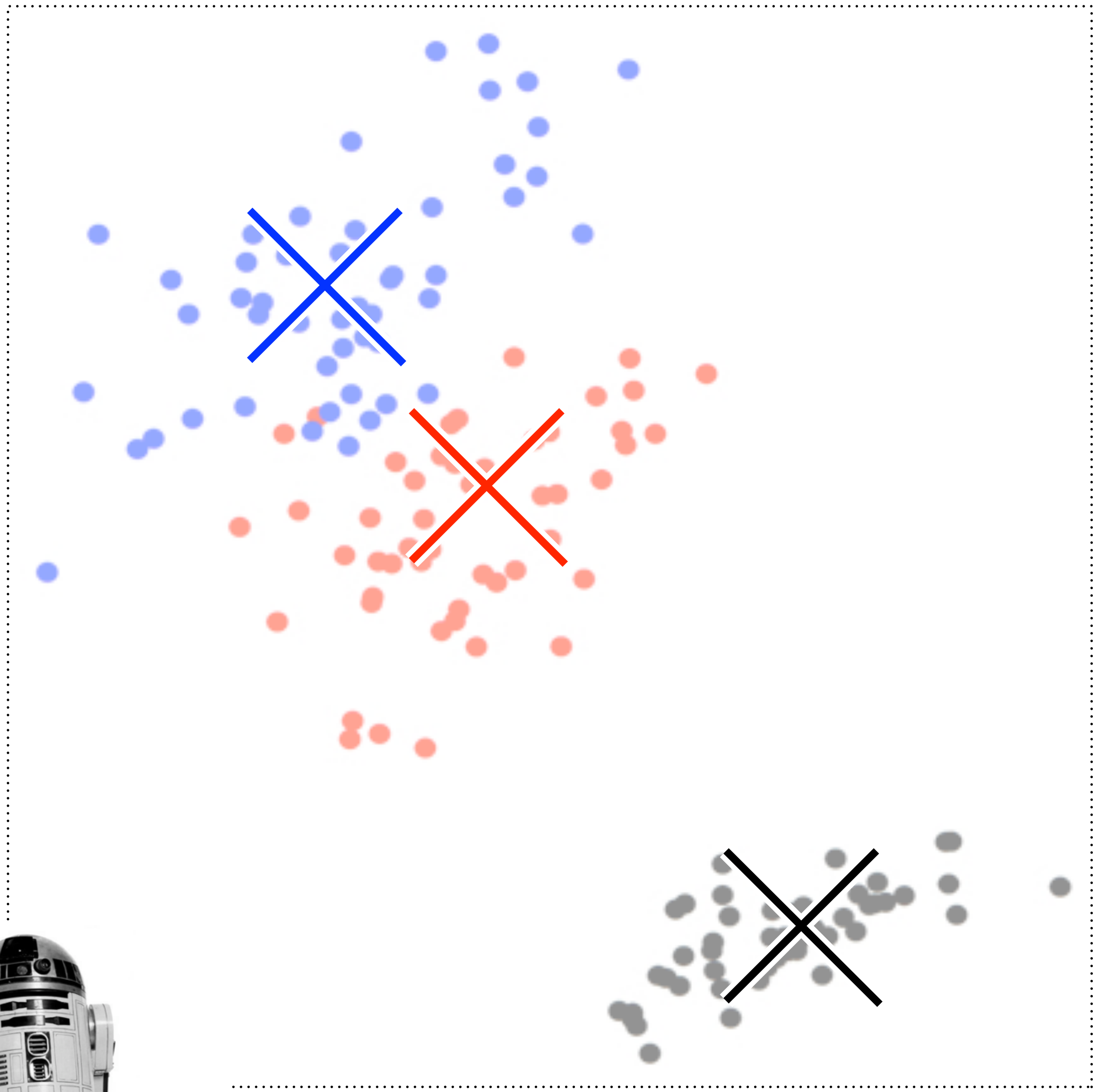


what goes in?



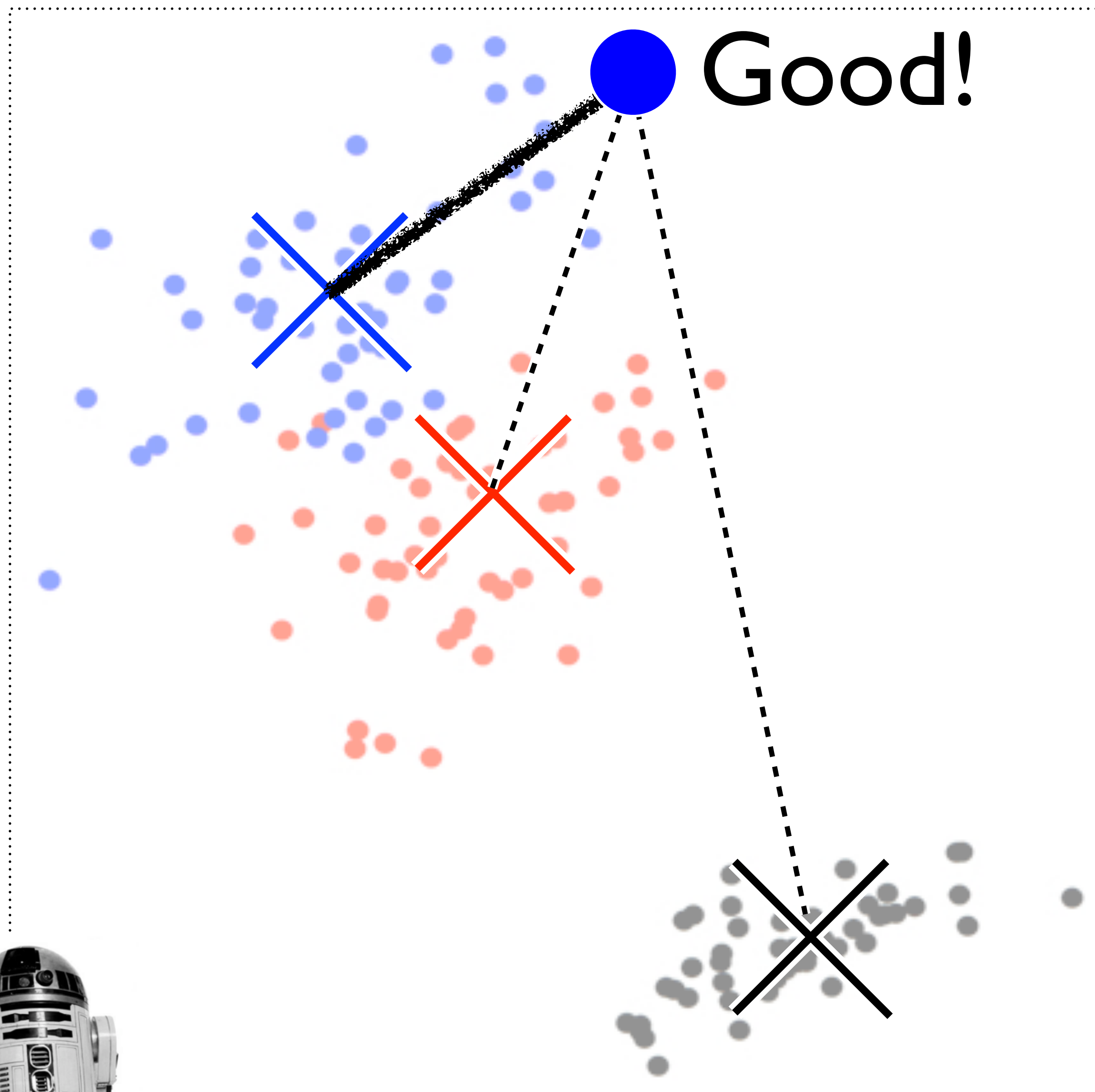
Dataset
Class Structure
m classes (here $m = 3$)

compute centroids



$centr(c_i)$

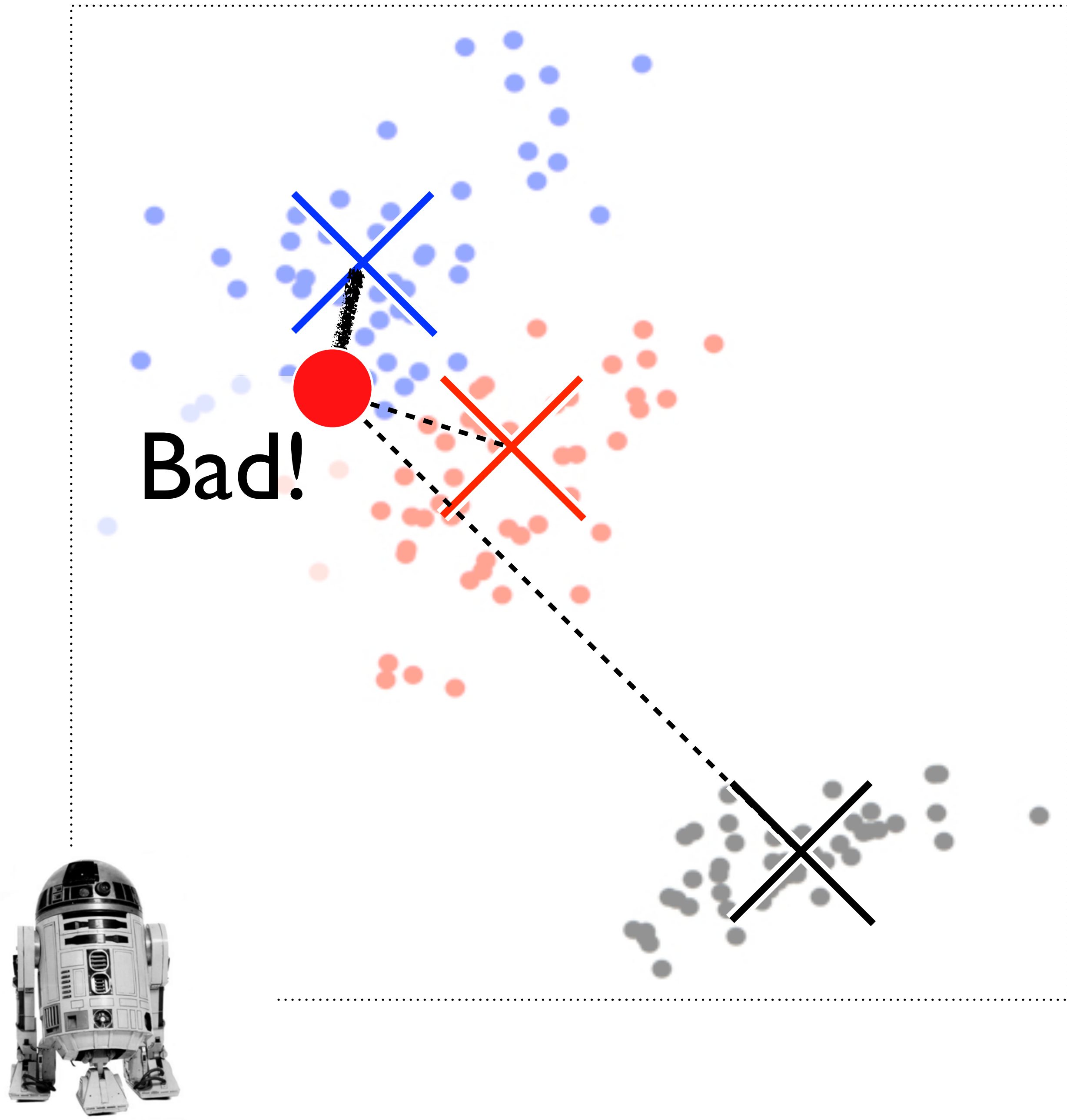
point: “own centroid closest”?



Compare distances:
d(x, centr(c_i))

point ***x***
to all centroids ***centr(c_i)***

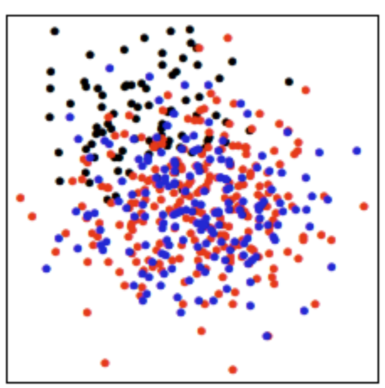
iterate over all points

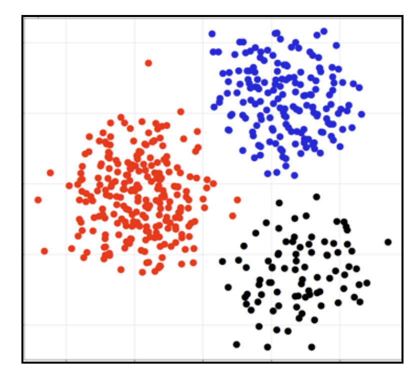


what comes out?

$$\frac{\#\textbf{good}\ \text{points}}{\#\textbf{all}\ \text{points}} * 100$$

(Centroid Metric)





0 .... **100**

(Bad View) *(Good View)*

teaching goals

e.g., program such algorithms

Different angles of attack:
- mental model
- mathematical depiction
- code (learning by doing)

challenges:
mental models

- For some algorithms easier than for others
- visual vs. verbal thinkers
- time to produce good mental models

Potential solutions:
- metaphors
- evocative terminology

Design Thinking / Creativity

Math / Algorithms

general course layout

Interactive Classroom Experience

specific activity

Providing Mental Models

Design Thinking / Creativity

Math / Algorithms

Interactive Classroom Experience → ?

✓ ← Providing Mental Models

Design Thinking /
Creativity

Math /
Algorithms

Interactive
Classroom
Experience

Combined Classes

Providing
Mental
Models

Summary

visual data analysis research

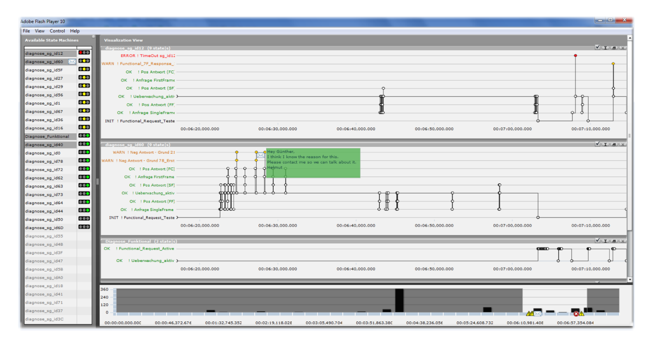
real users/real data

right methods

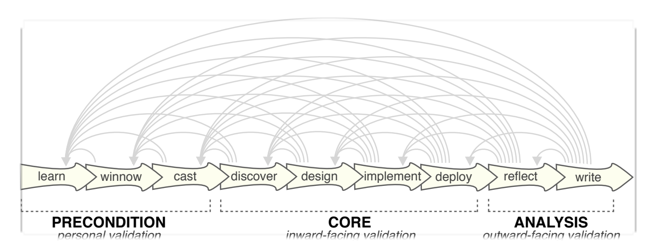
1. applied design studies

- important contributions to application domains
- important steps towards the "visualization and data analysis cookbook"
- novel and refined methods/methodologies

Cardiogram

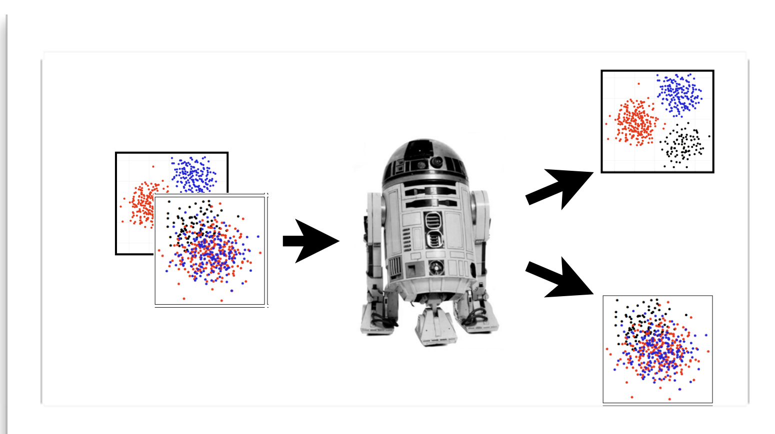


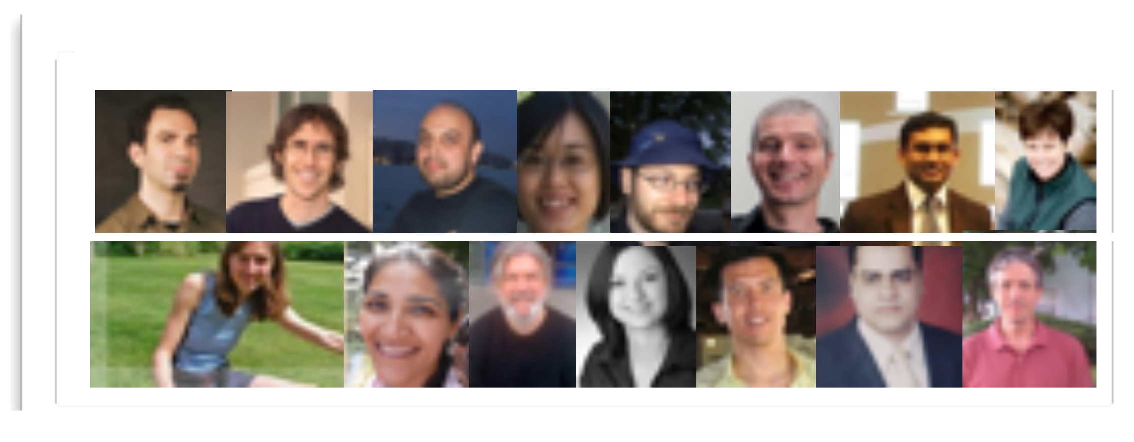
Design Study Methodology



2. high-dimensional data analysis techniques

- understanding in real situations
- imperative for guiding future technical research





teaching

visual data scientist
needs design and technical skills!

struggle points
- design thinking
- algorithms/math

approaches
- interactive classroom experience
- providing mental models

thanks...

phd (2007-2010)
BMW Group, Munich, Germany
University of Munich, Germany

A. Butz

postdoc (2010-2012)
University of British Columbia,
Vancouver, Canada

T. Munzner

junior group leader (2013-now)
University of Vienna, Austria

T. Mölller

Visual Data Analysis:
real users, real data, right methods

Thank you!

Michael Sedlmair

https://homepage.univie.ac.at/michael.sedlmair
michael.sedlmair@univie.ac.at