Human-centered Data Science

Michael Sedlmair
Visualization & Data Analysis Group
universität wien

LMU Munich
Diplom in Computer Science, 2001-2007

BMW Group Research & Technology
PhD in Computer Science, 2007-2010

Univ. of British Columbia, Canada
PostDoc, 2010-2012

Univ. of Vienna, Austria
Assistant Professor, 2013-now

Human-Computer Interaction (HCI)

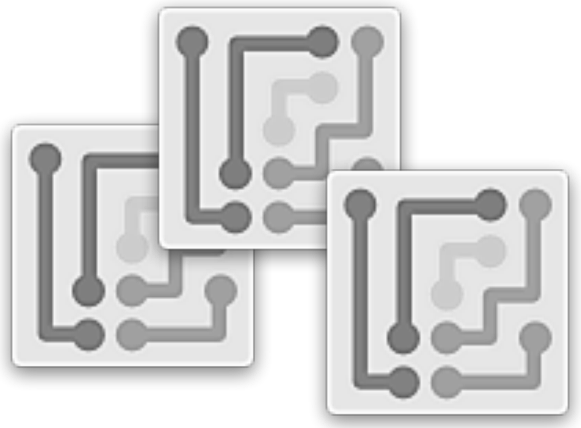
Data Visualization

Data Science (algorithmic analysis)

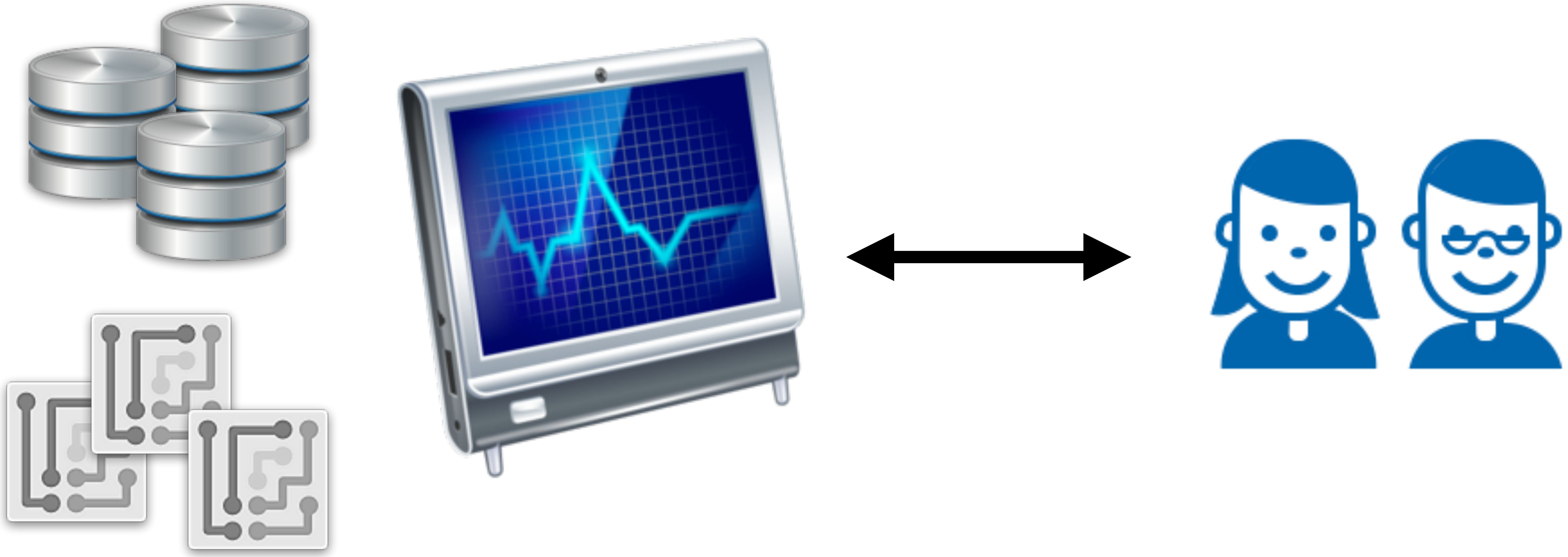
Data Science

Data

Algorithms



My interest: **Human-centered Data Science**



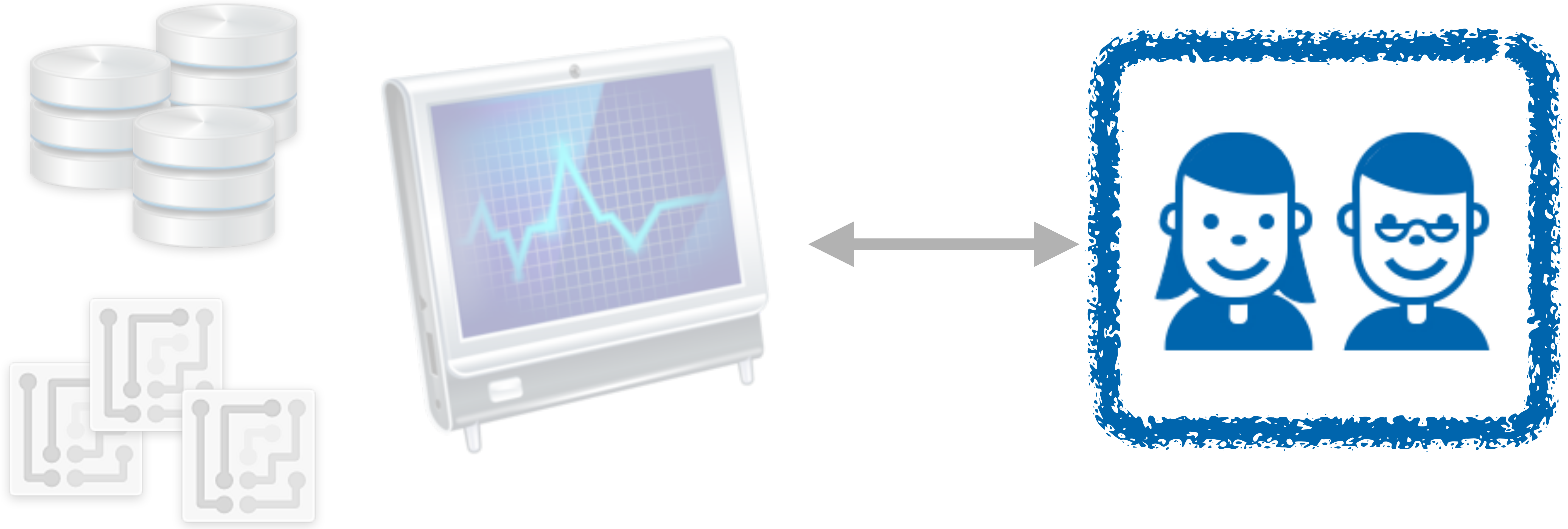
Scientists

Engineers

Journalists

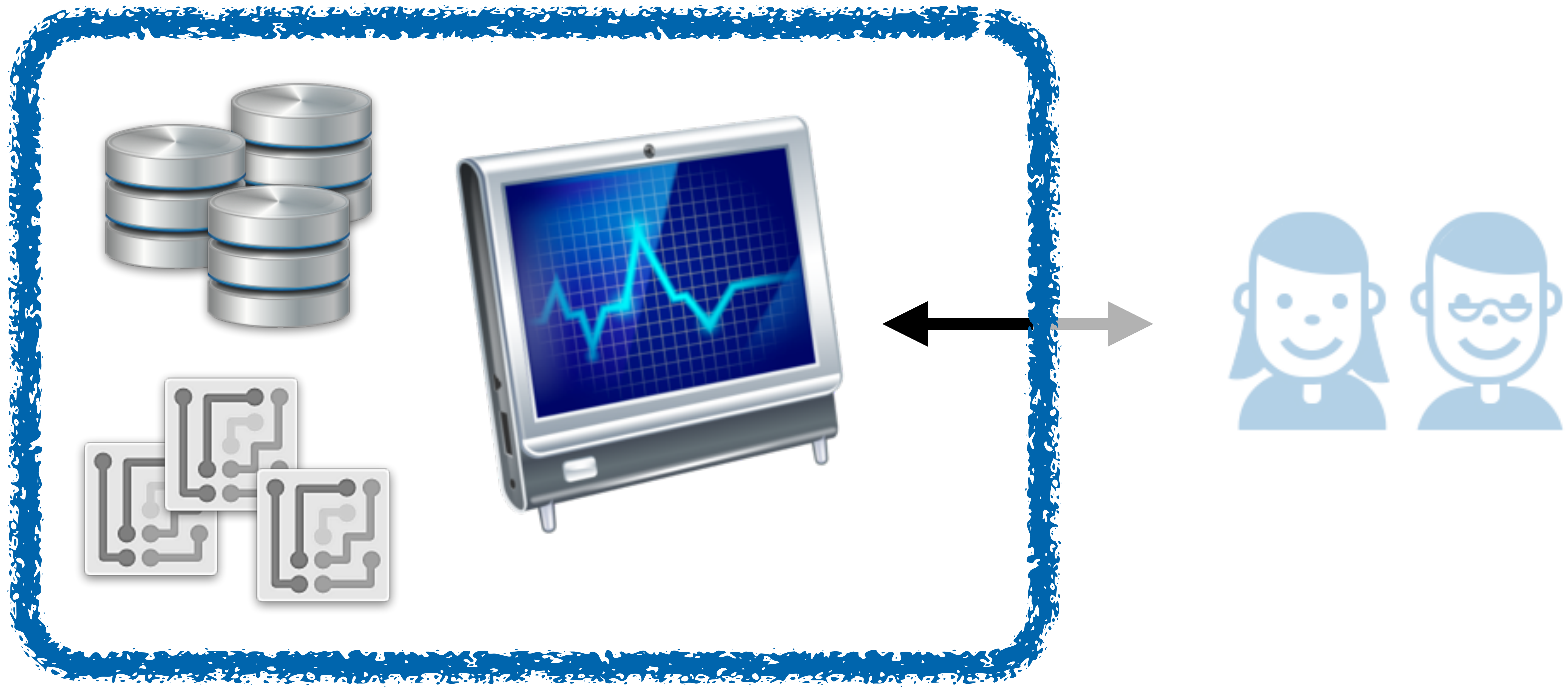
Policy Makers

Goals of my work



1. Human as a first class citizen in Data Science research

Goals of my work



2. Human-centered tools & techniques

➡ interactive visualization

➡ algorithmic analysis

Help to:

Understand patterns

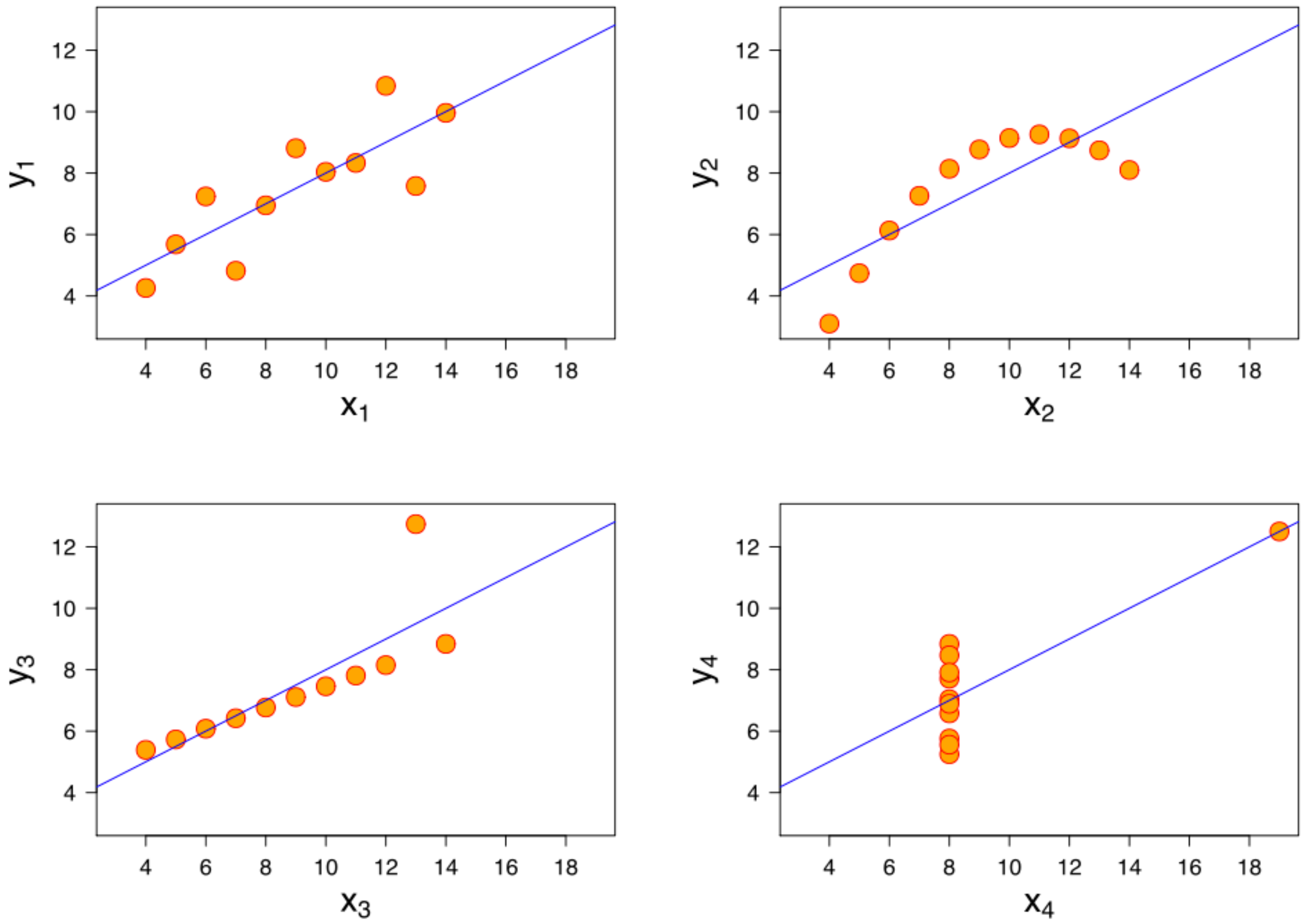
Gain insights

Make decisions

Why interactive visualization?

- **humans** can see patterns that algorithm cannot

Identical statistics	
x mean	**9**
x variance	**10**
y mean	**8**
y variance	**4**
x/y correlation	**1**

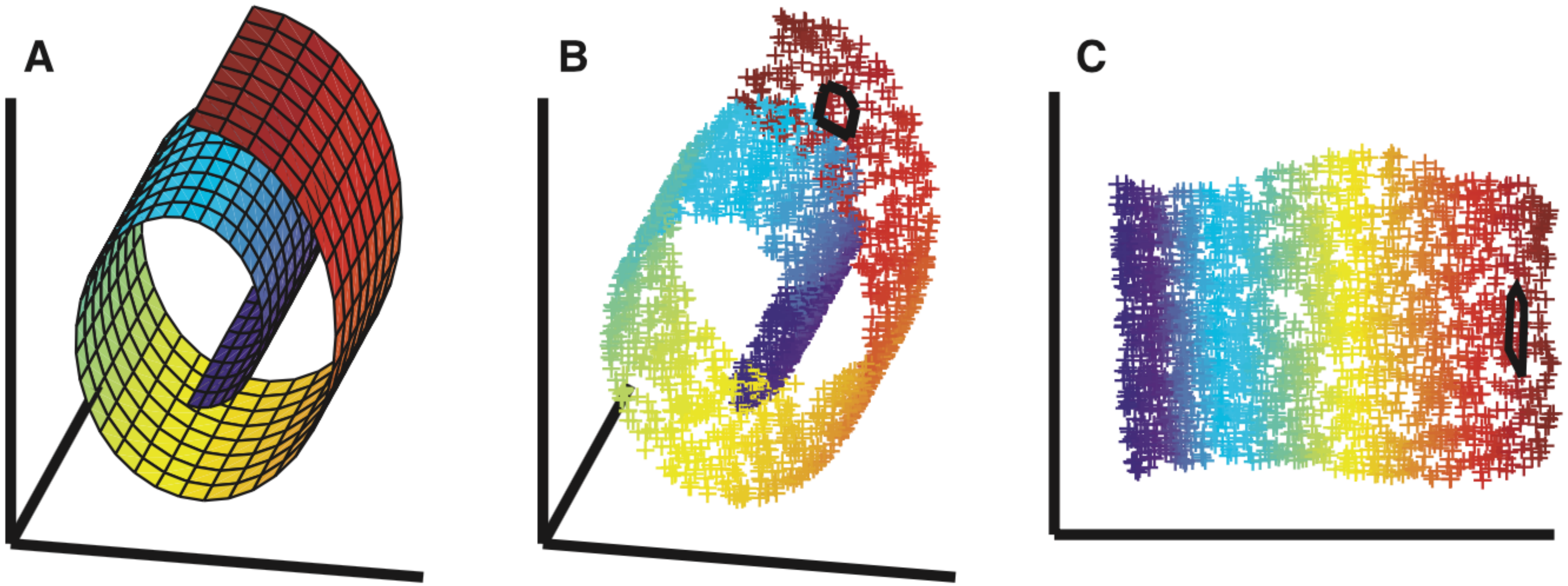


Anscombe's Quartet

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Why algorithmic analysis?

- **algorithms** can reveal patterns that humans cannot see



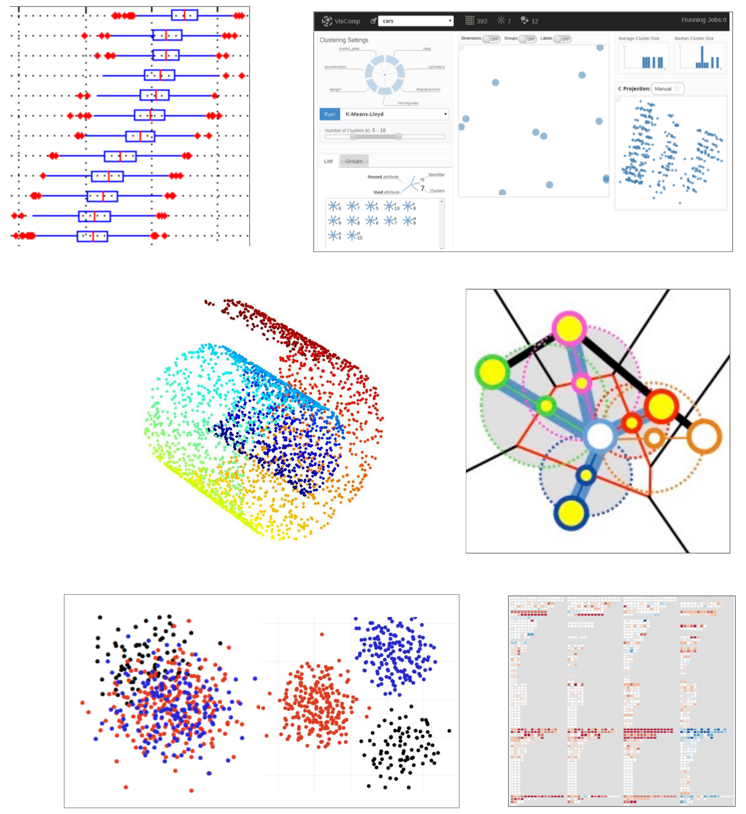
Dimension reduction:
LLE unfolding swiss roll

Roweis, Saul.
Nonlinear dimensionality reduction by locally linear embedding.
Science, p. 2323-2326, 2000.

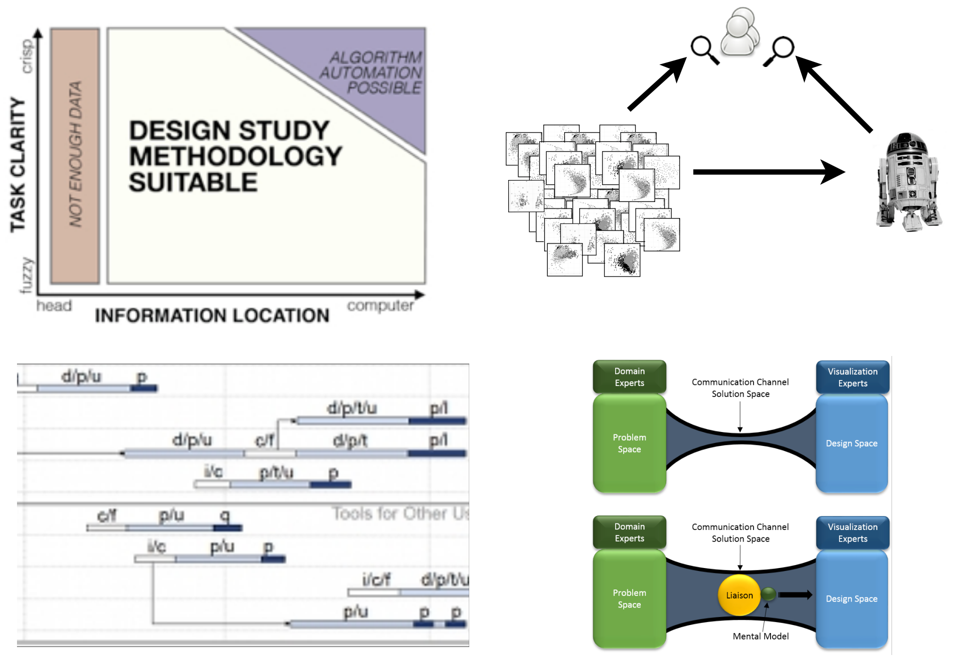
problem-driven

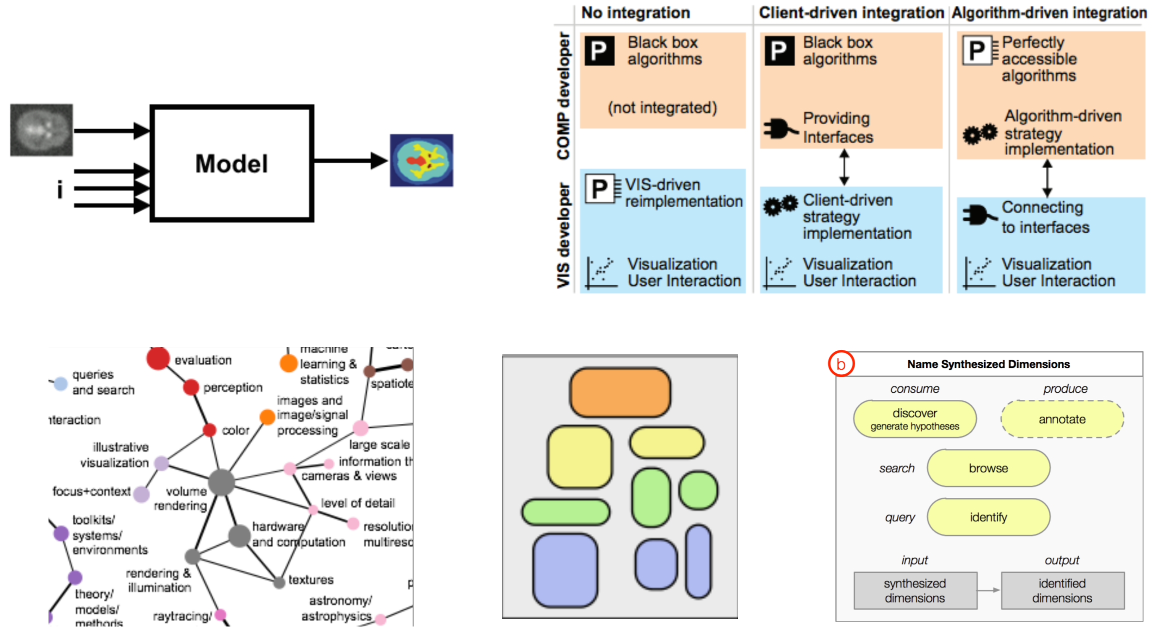
MY WORK

technique-driven



methods & theory

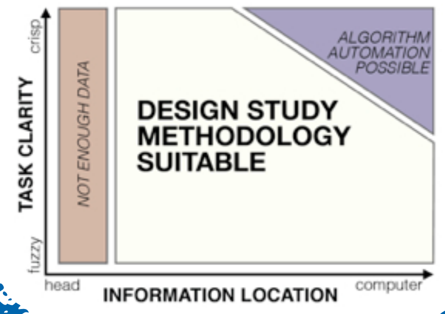




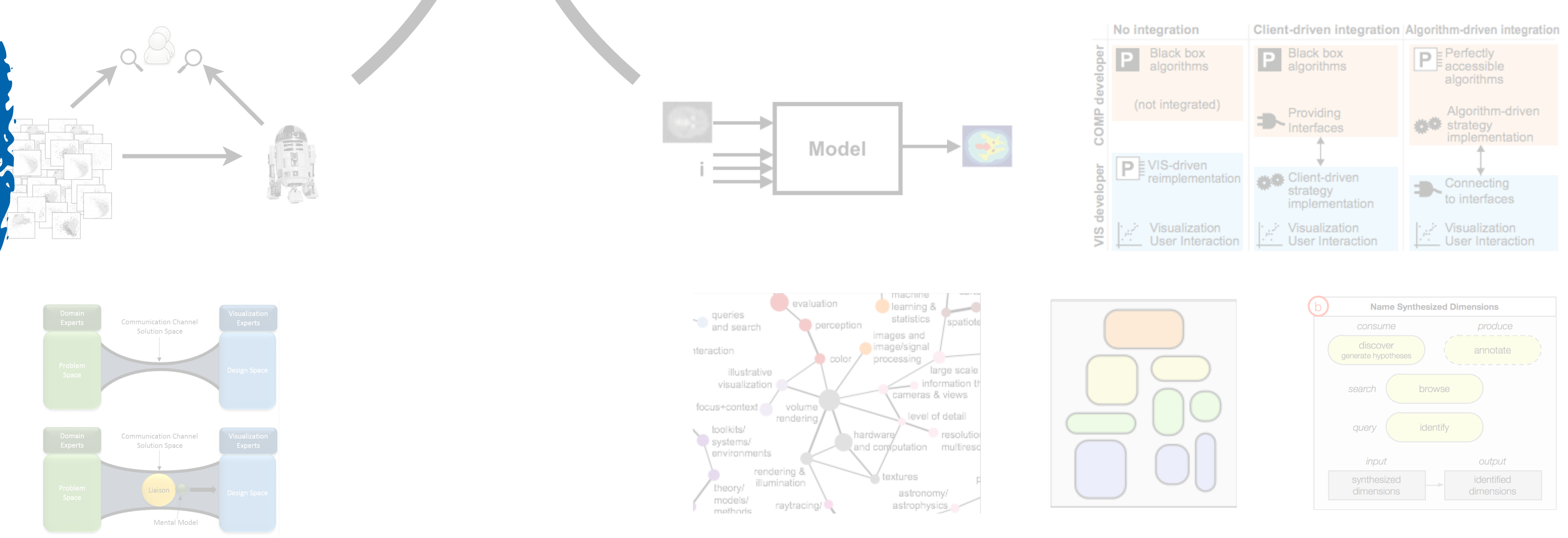
pro

Today’s focus:
Design Studies

- interdisciplinary project
- analysis of domain problem
- design and validation of a visual analysis tool

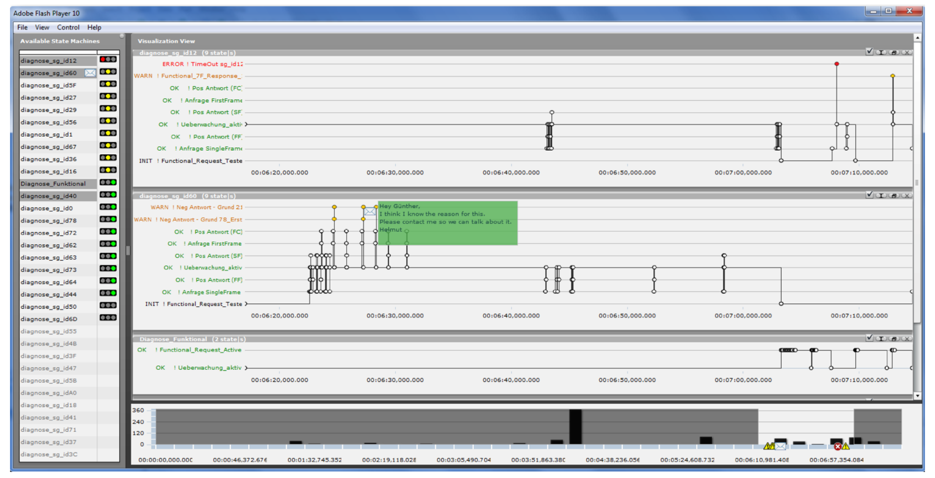


methods & theory



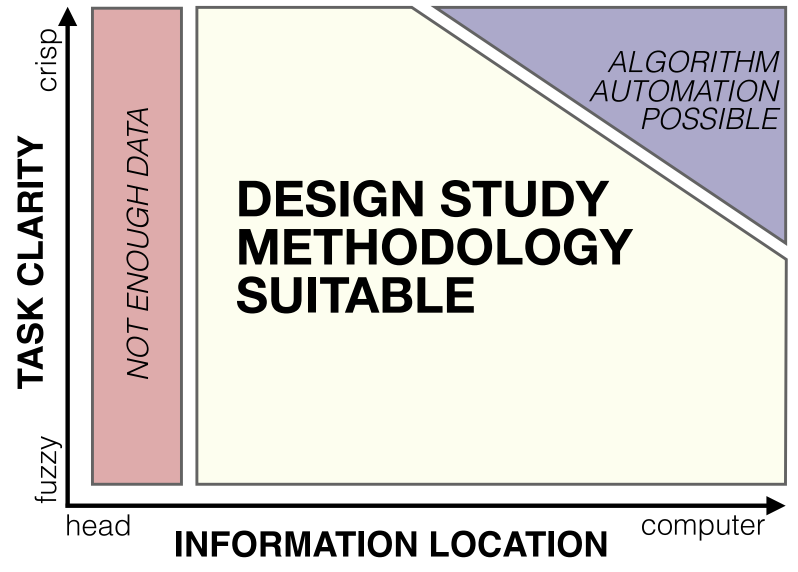
Outline

- Cardiogram
  - example design study with automotive engineers



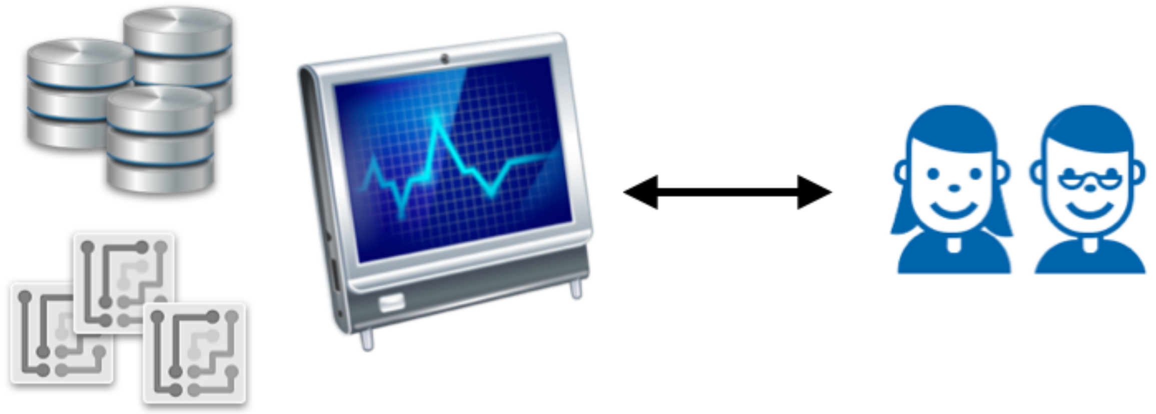
[ACM CHI, 2011]

- Design Study Methodology
  - meta-paper: how to do design studies



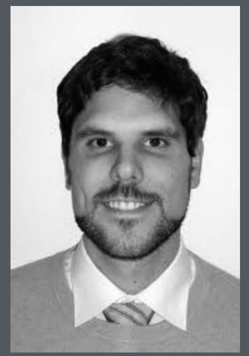
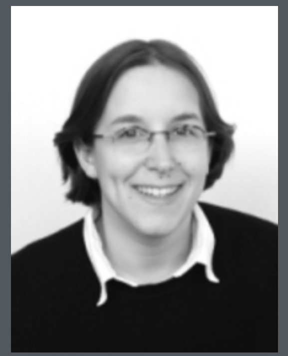
[IEEE TVCG, 2012]

- Future Work
  - 3 major challenges



Cardiogram:
A Design Study with BMW engineers

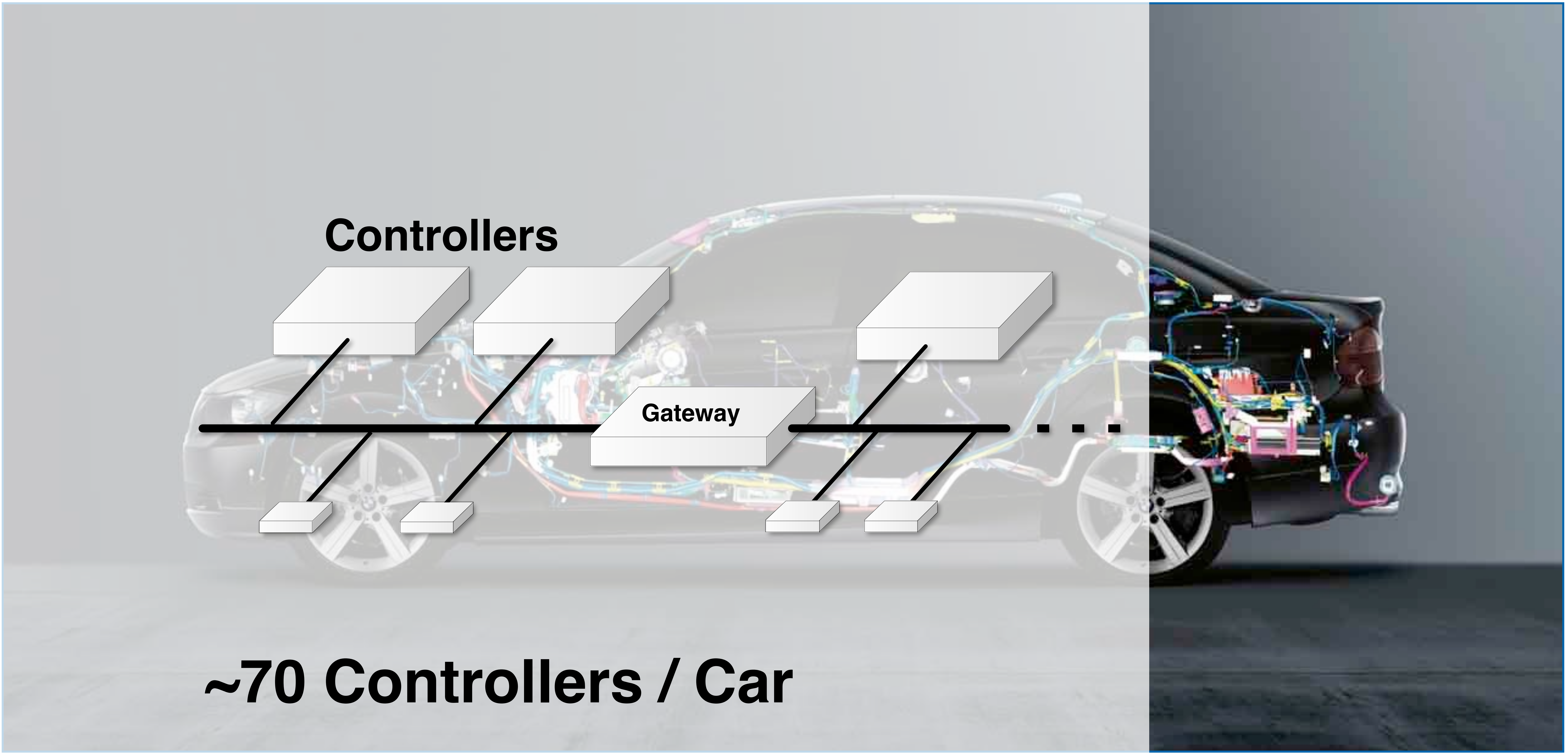
Cardiogram: Visual Analytics for Automotive Engineers
M. Sedlmair, P. Isenberg, D. Baur, M. Mauerer, C. Pigorsch, A. Butz
ACM CHI, p. 1727–1736, 2011.



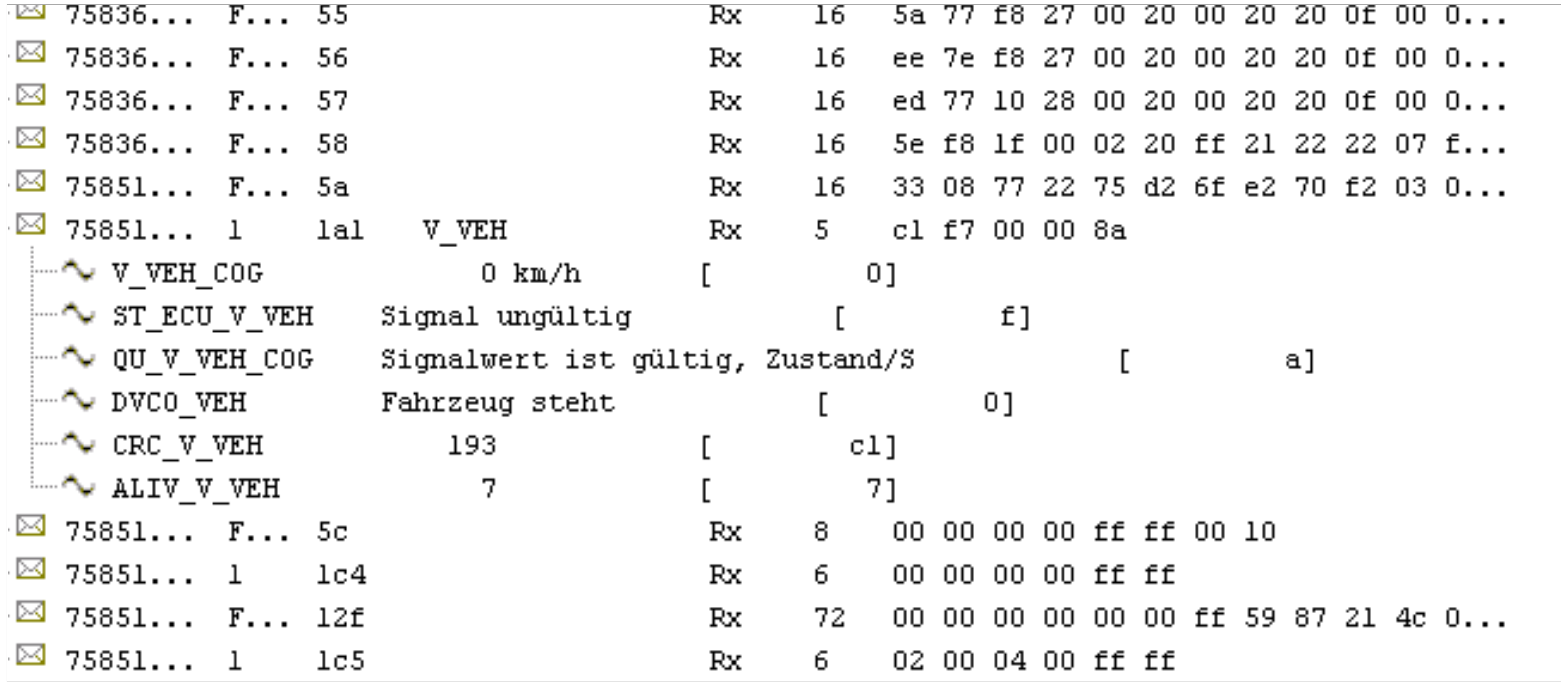
More and more electronics

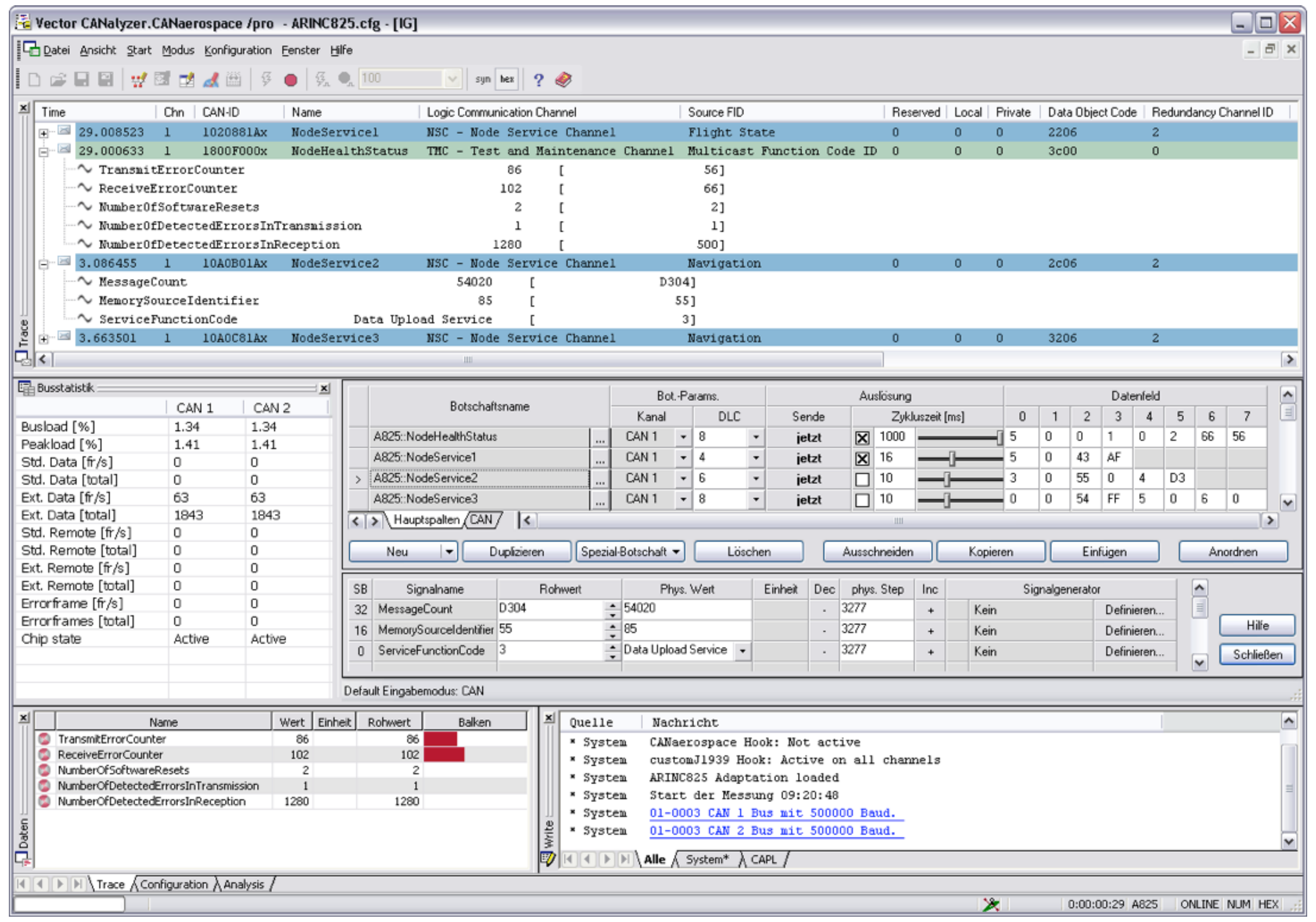
... enabled by in-car communication networks.



Problem characterization

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



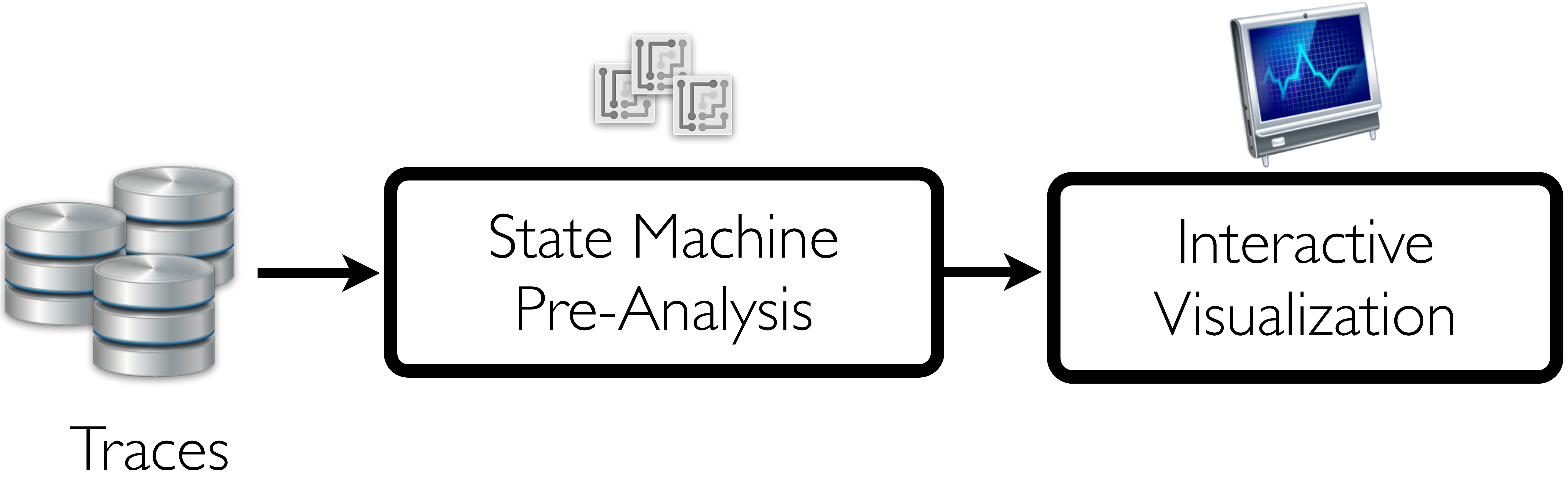


Challenges

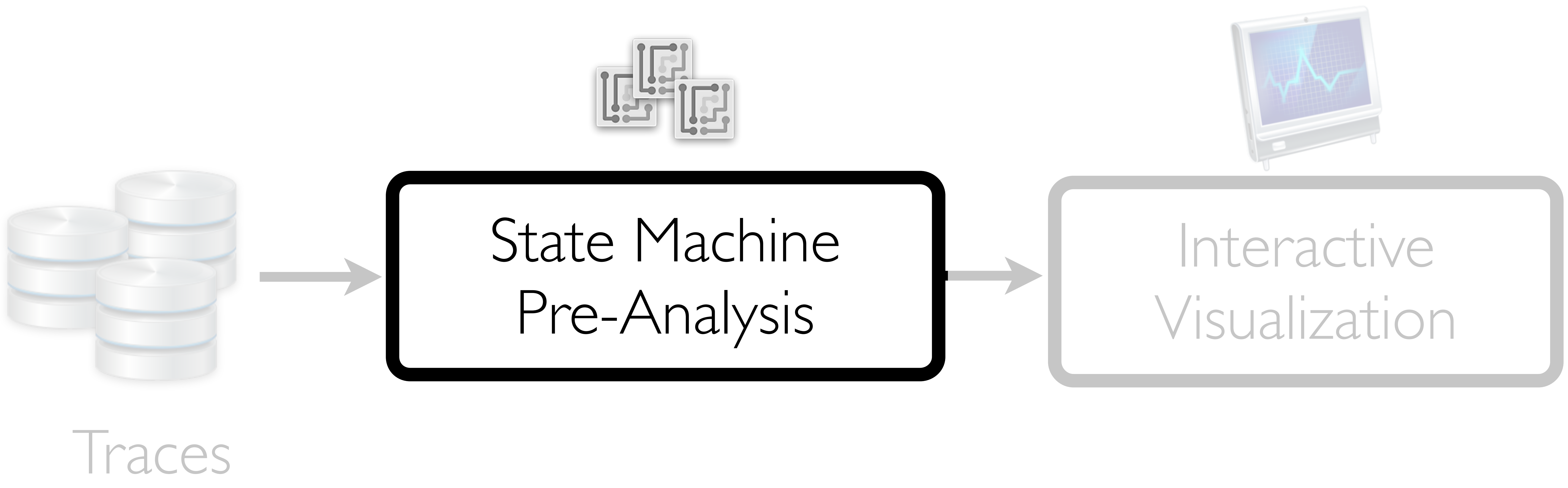
- handling masses of test traces (large and many)
- understanding correlation between trace and car behavior

- *(more in the paper)*

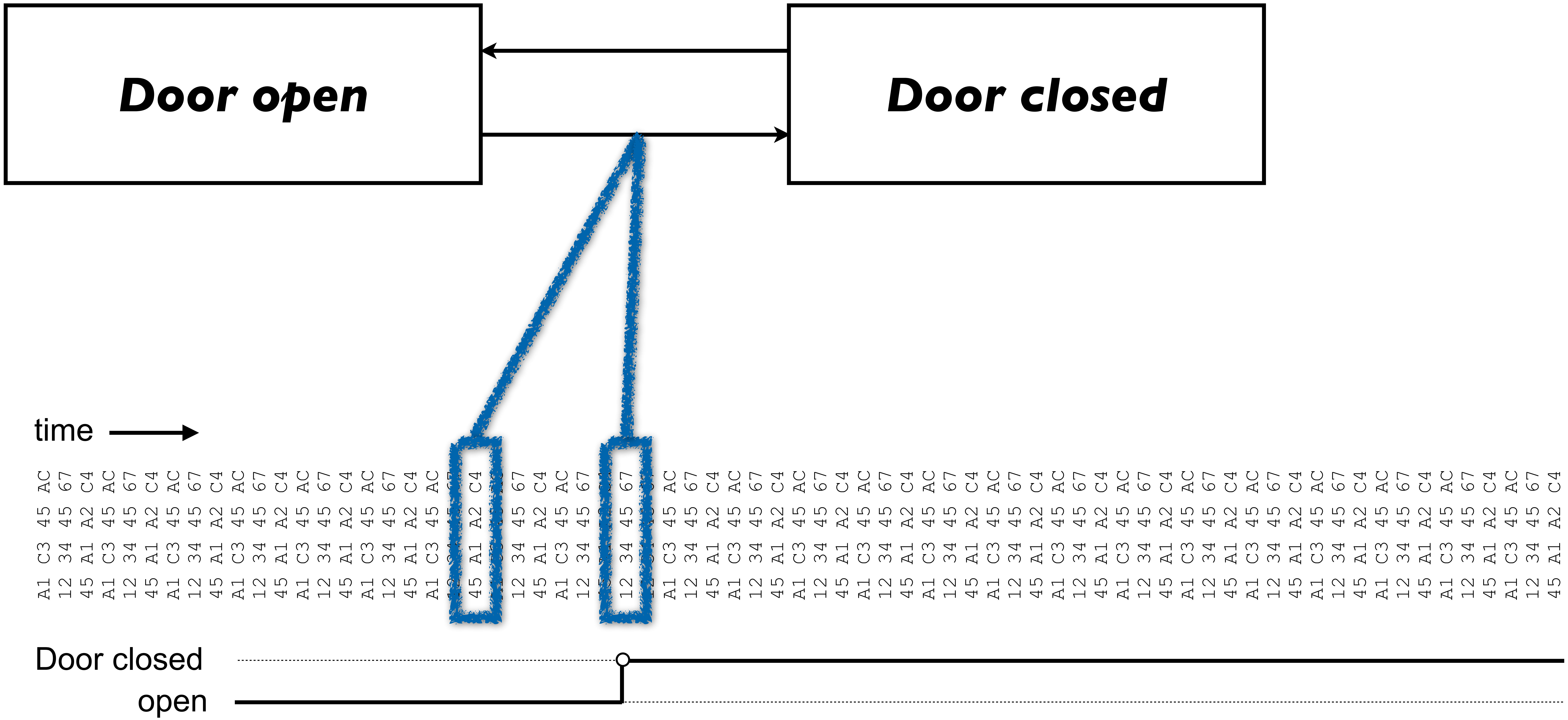
Cardiogram


State Machine
Pre-Analysis
Interactive
Visualization
Traces

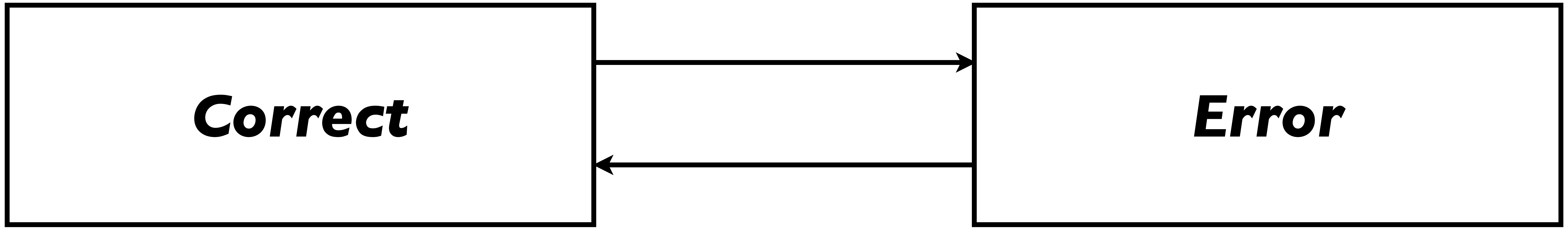
Cardiogram:
state machine pre-analysis



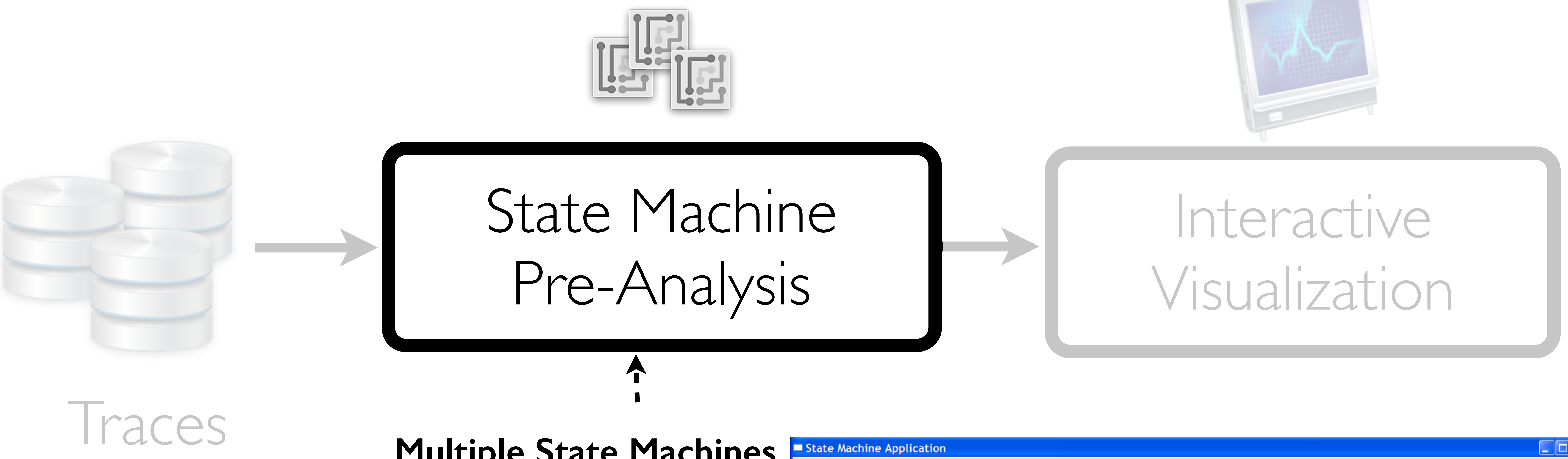
State machine:
behavior abstraction



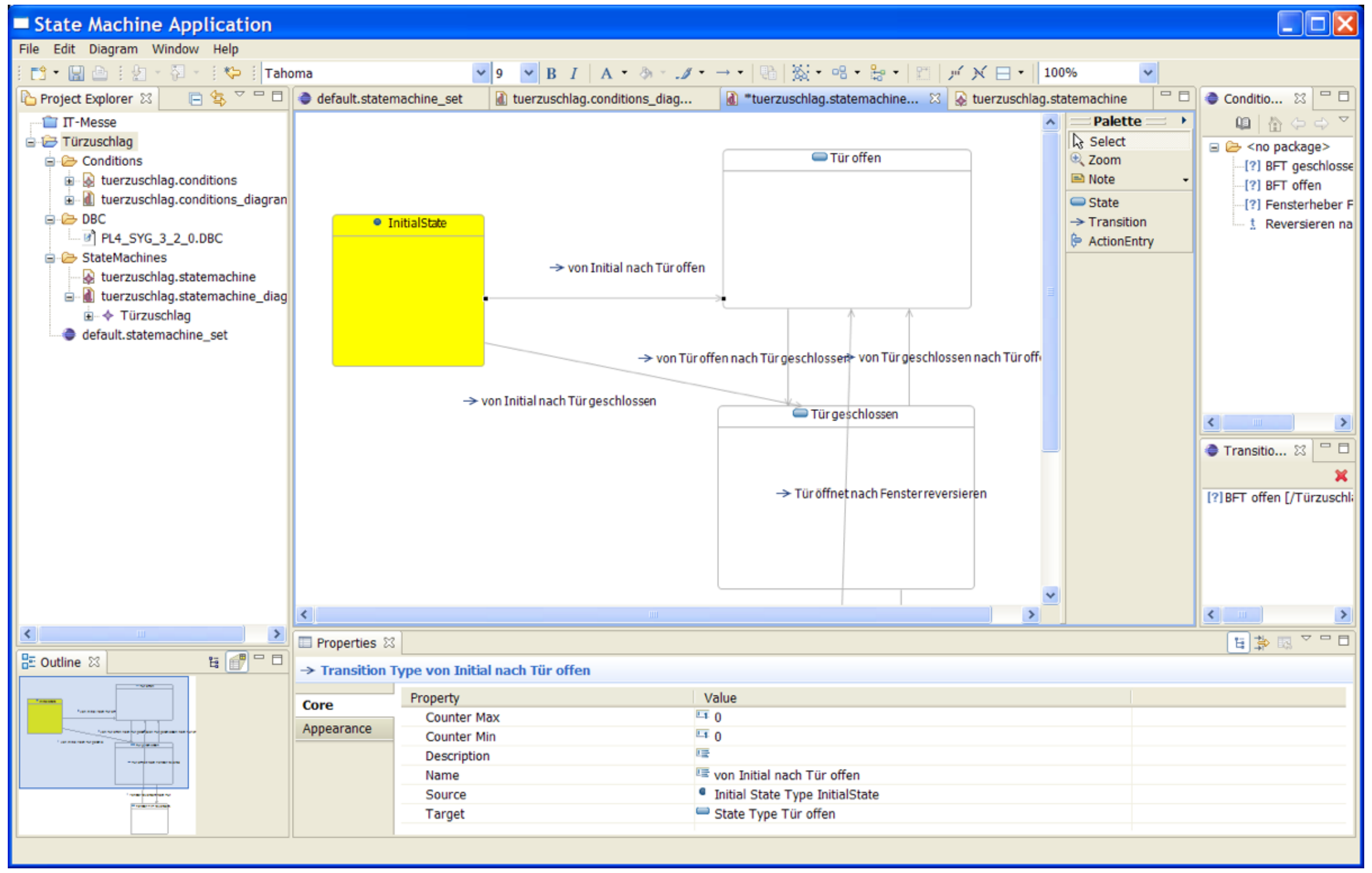
State machine:
automatic error detection



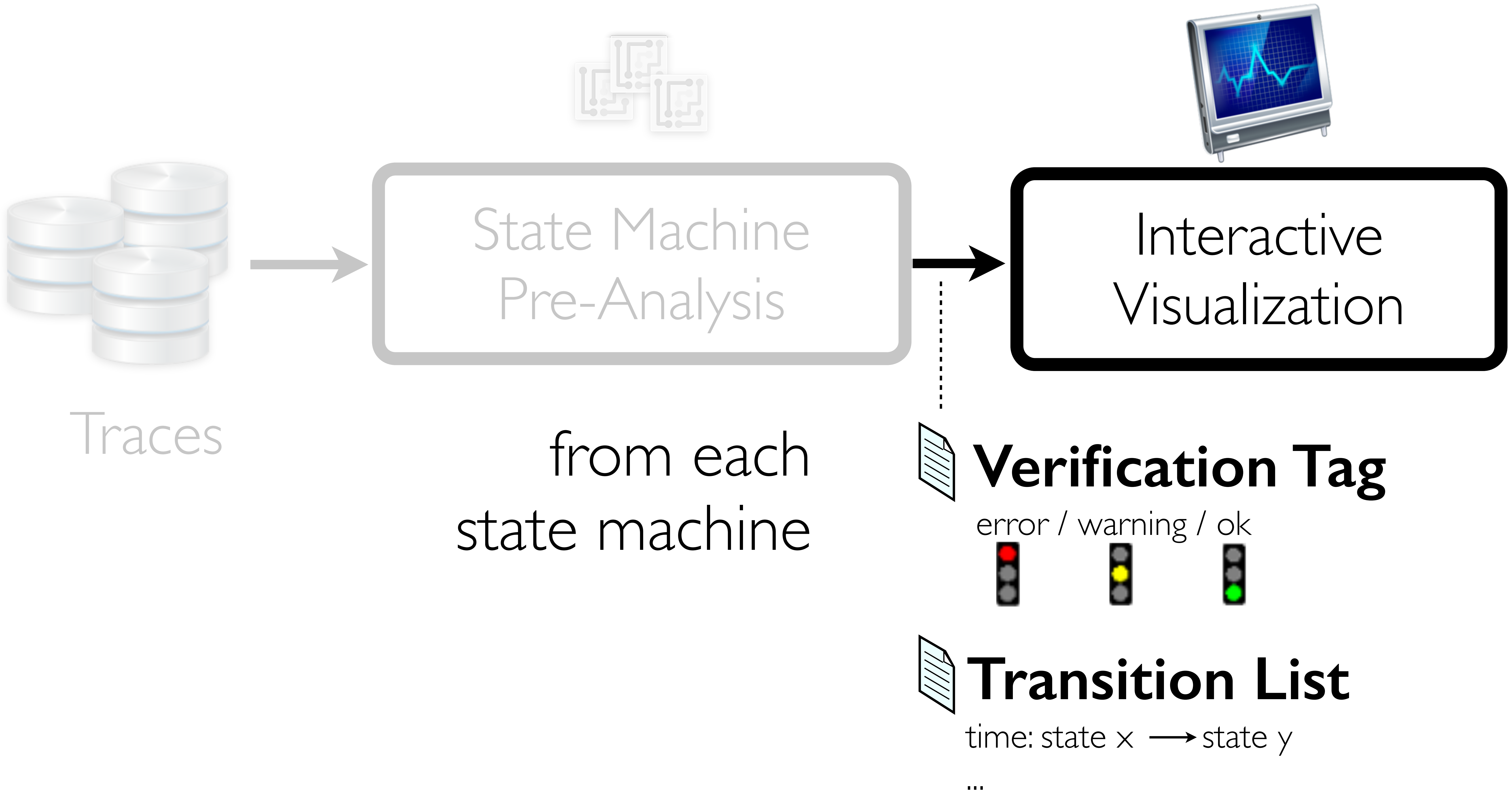
Cardiogram:
create/select state machines

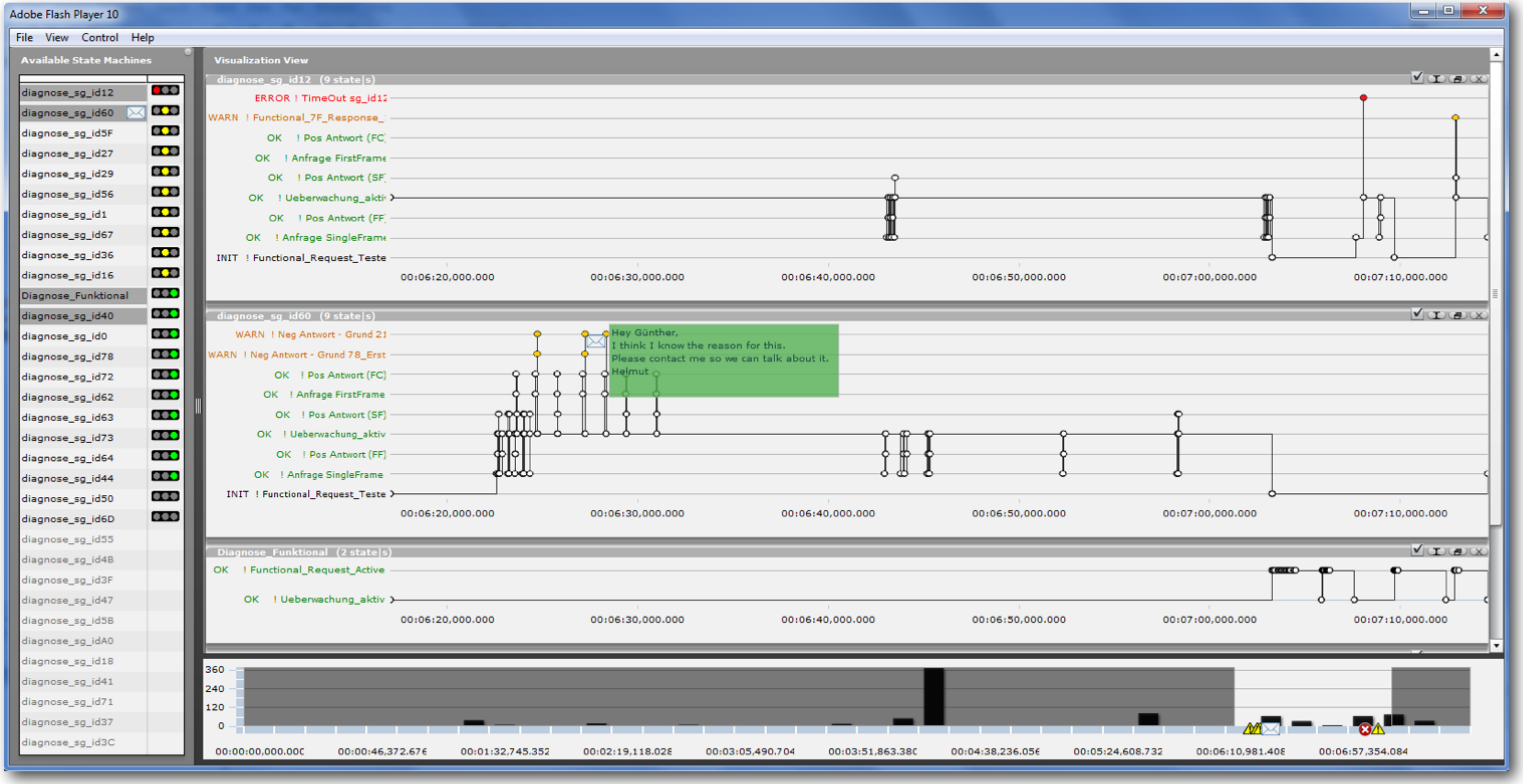


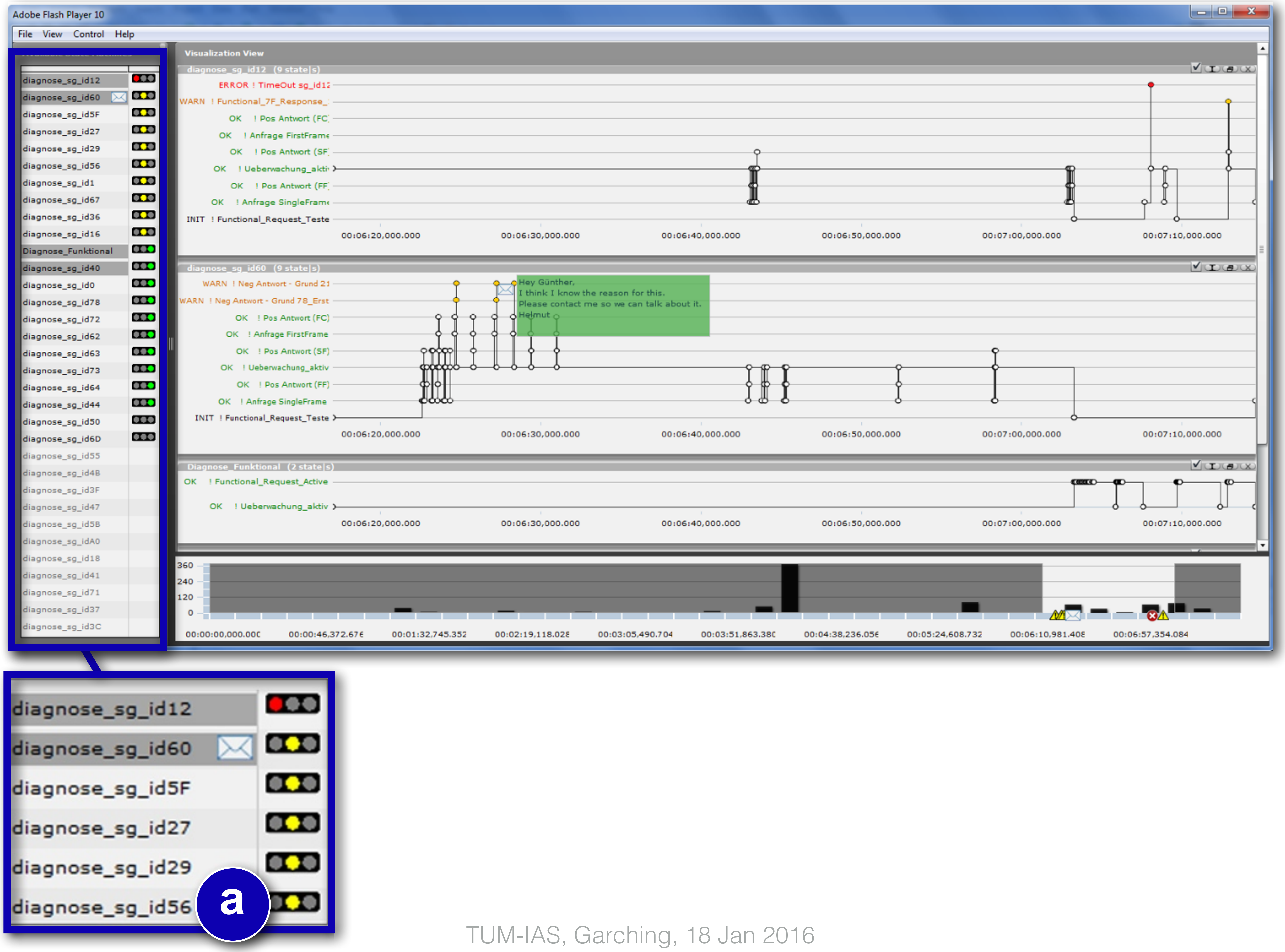
Multiple State Machines
- State Machine 1
- State Machine 2
- State Machine 3
... (dozens)

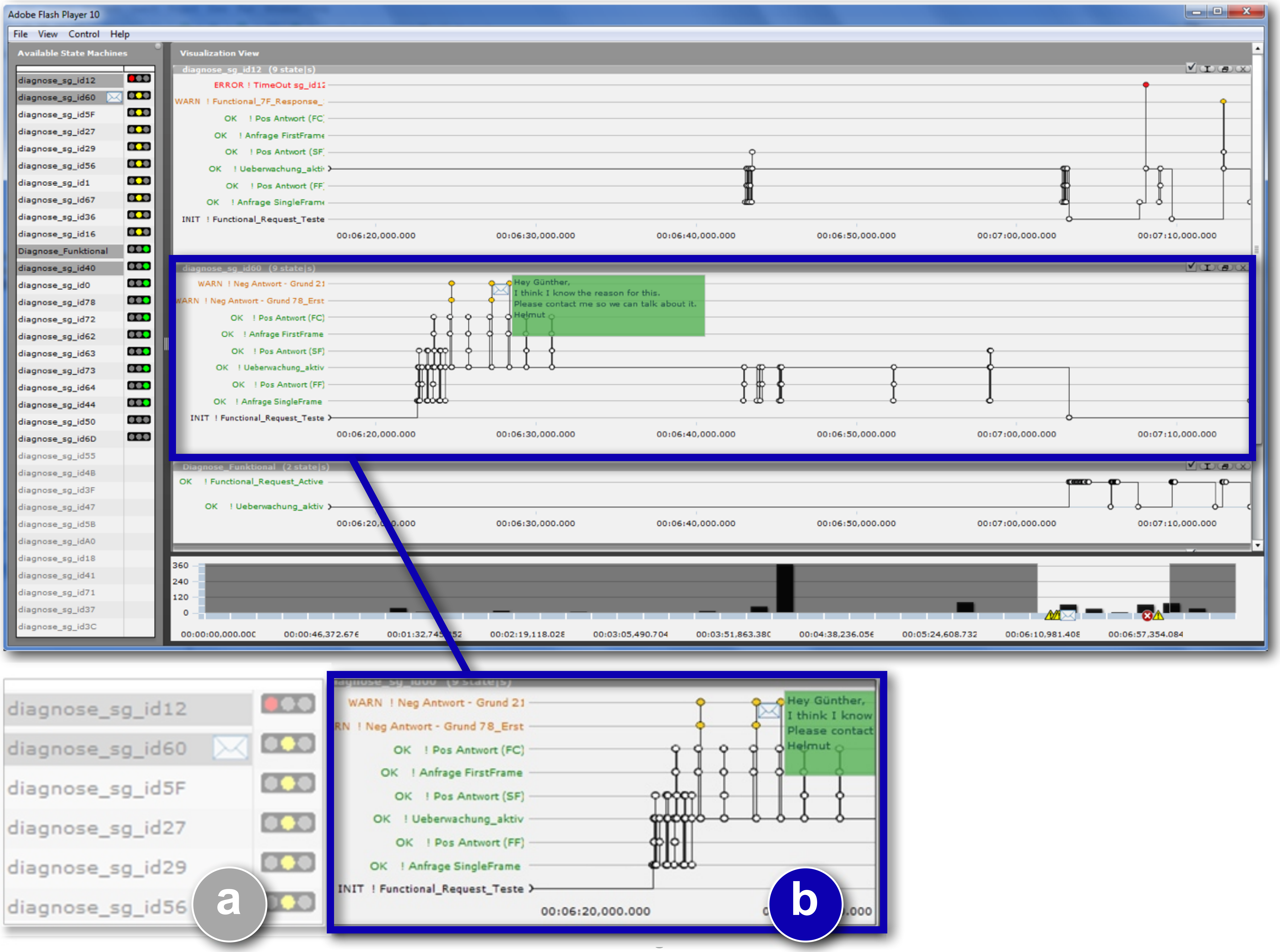


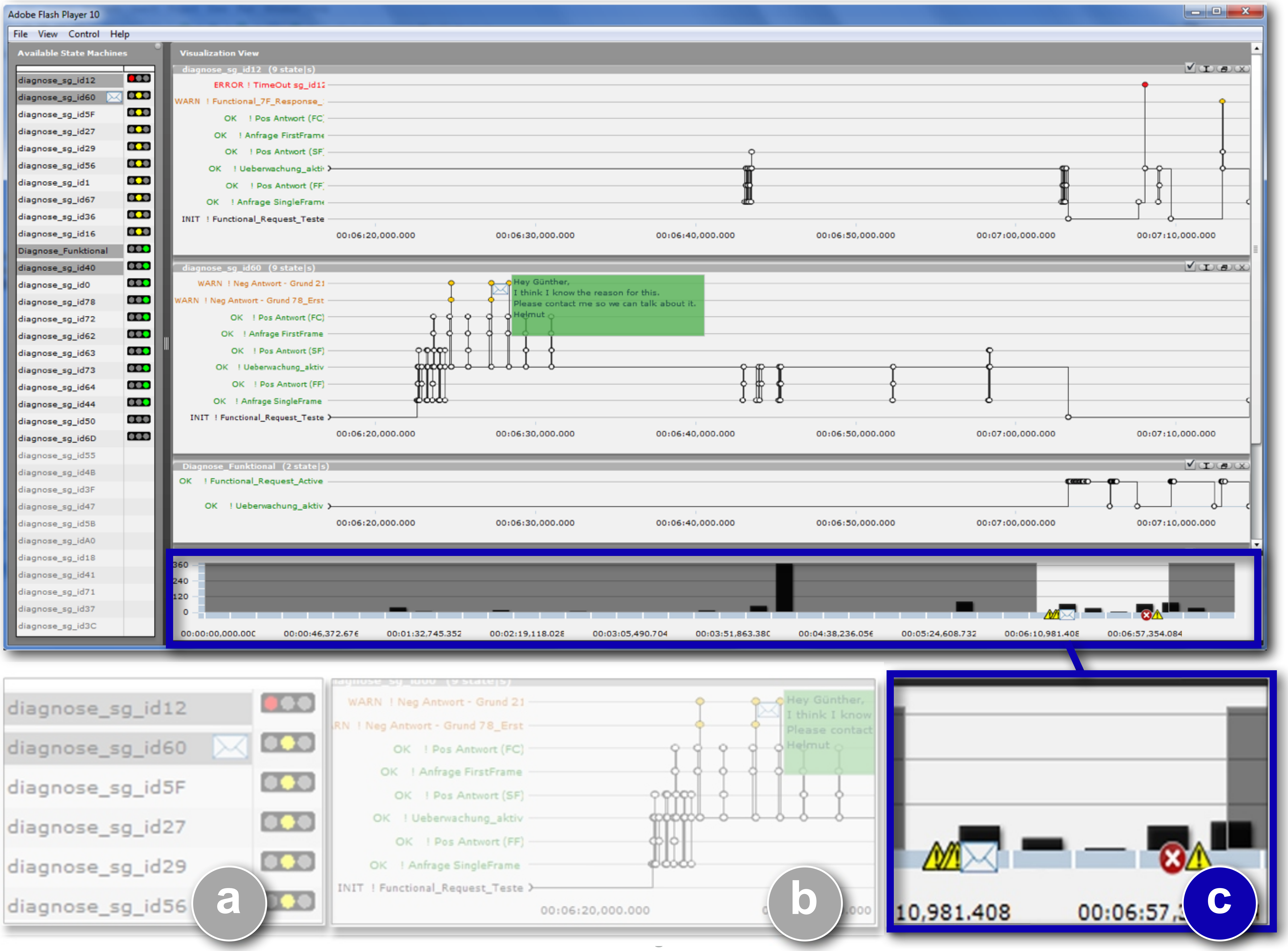
Cardiogram:
data interpretation / data reduction




Adobe Flash Player 10
File View Control Help
Available State Machines
Visualization View
diagnose_sg_id12 (9 state(s))
diagnose_sg_id60 (9 state(s))
Diagnose_Funktional (2 state(s))
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut


Adobe Flash Player 10
File View Control Help
Visualization View
diagnose_sg_id12 (9 state(s)
diagnose_sg_id60 (9 state(s)
Diagnose_Funktional (2 state(s)
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
diagnose_sg_id12
diagnose_sg_id60
diagnose_sg_id5F
diagnose_sg_id27
diagnose_sg_id29
diagnose_sg_id56
a


Adobe Flash Player 10
File View Control Help
Available State Machines
Visualization View
Hey Günther,
I think I know the reason for this.
Please contact me so we can talk about it.
Helmut
a
b



Michael Sedlmair

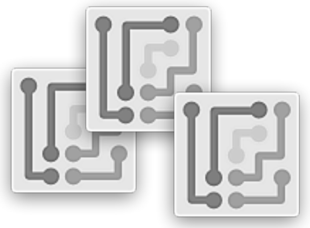
Longitudinal field study:
1 year, 15 engineers

Results:

- externalization and sharing of expert knowledge 

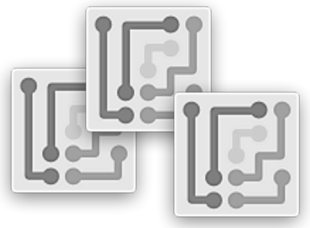
Longitudinal field study:
1 year, 15 engineers

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 
- novel insights: understand behavioral correlations 

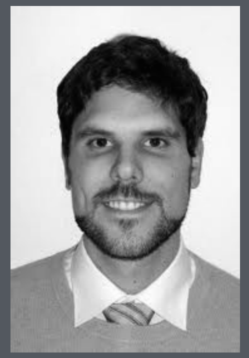
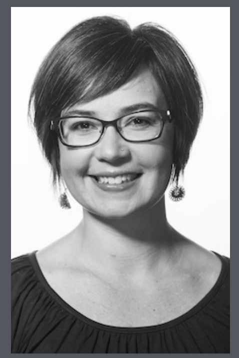
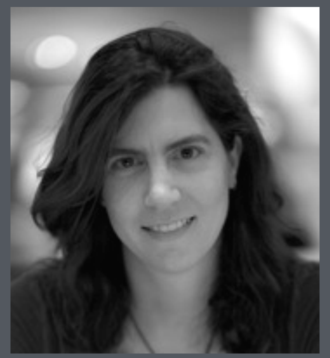
Design Study Methodology

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).

Best paper honorable mention.

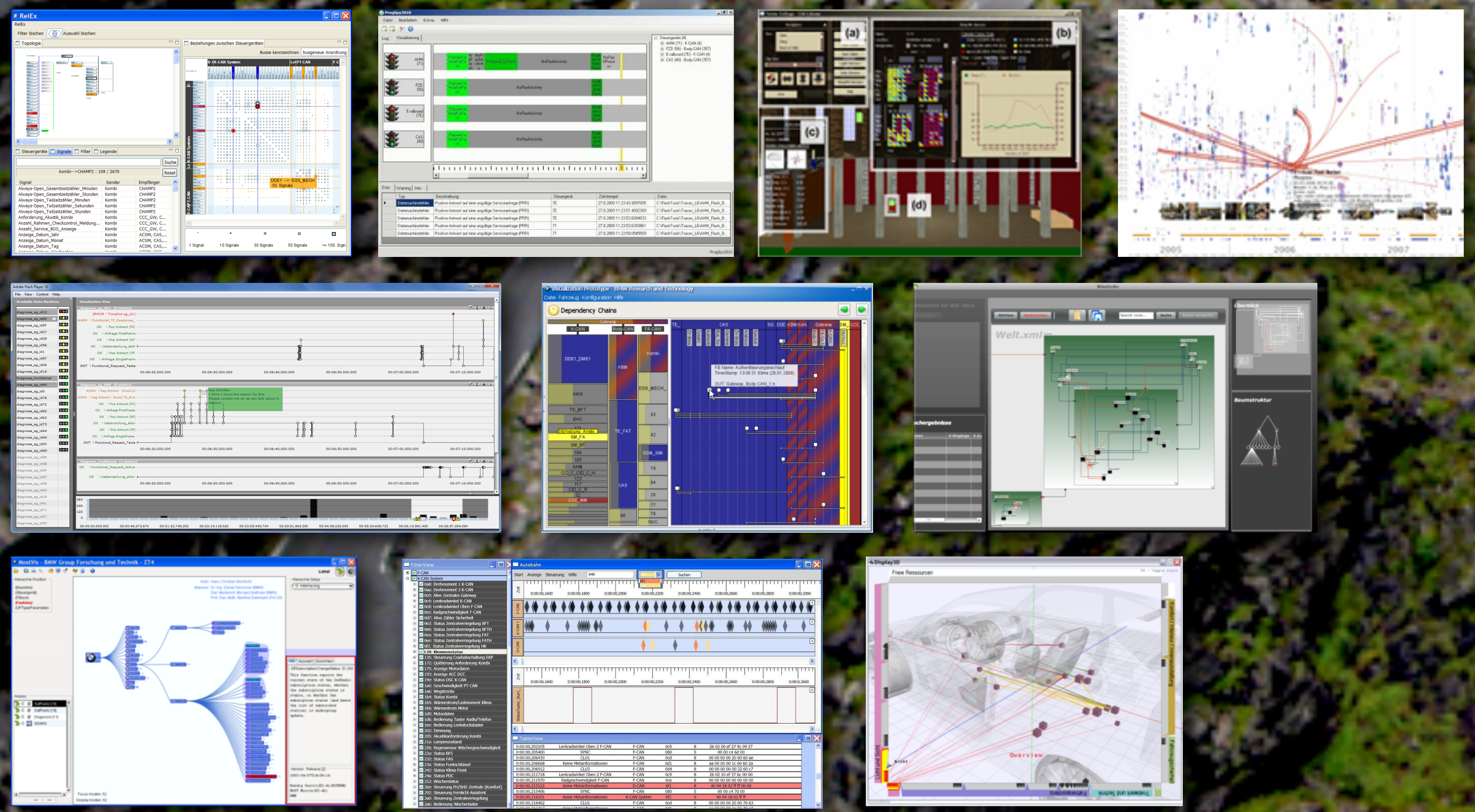




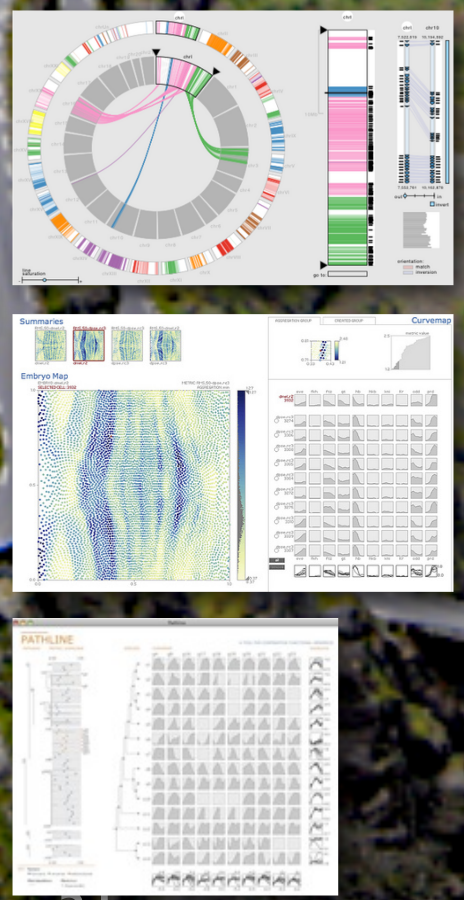
Design studies:
long and winding road with many pitfalls!

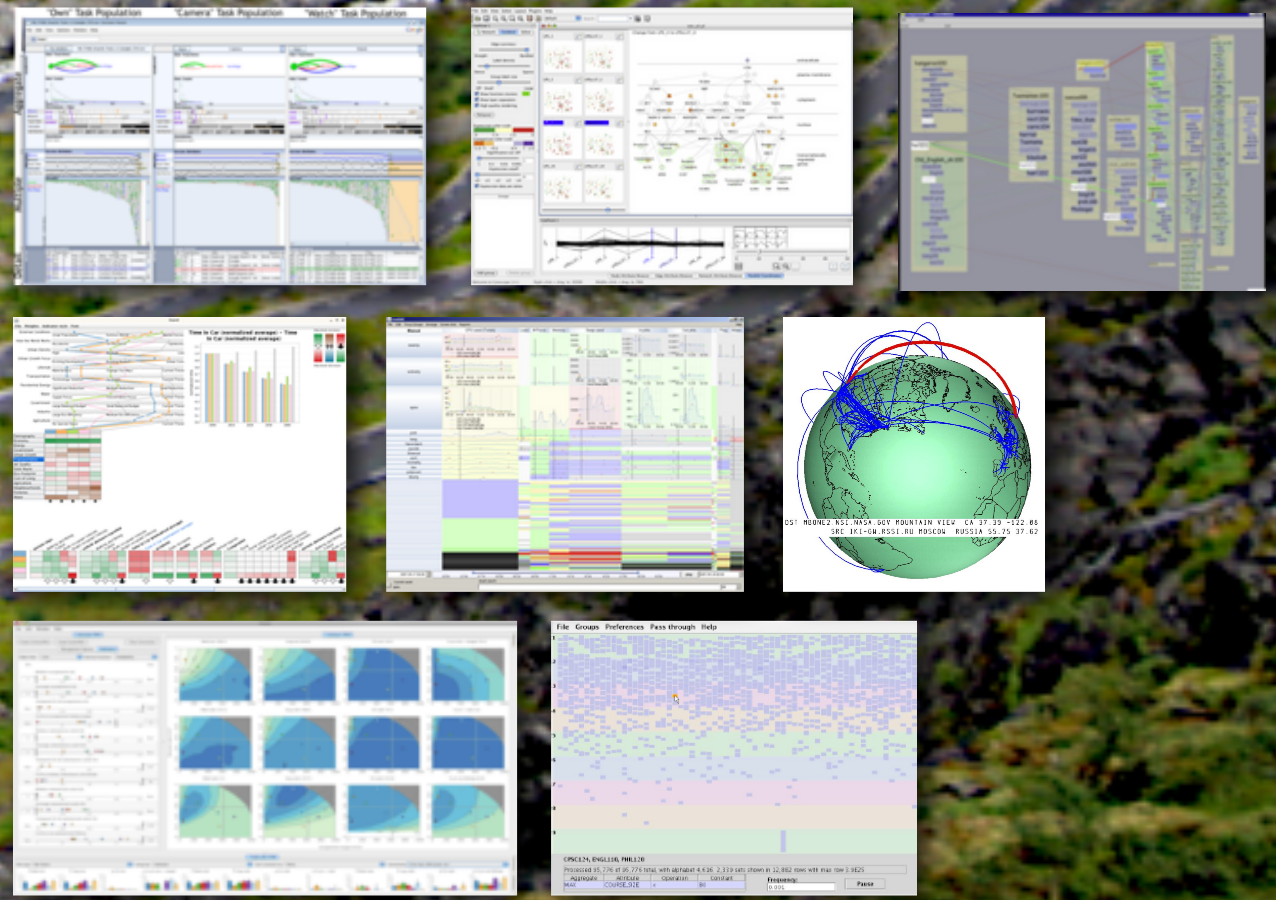
et al.

et al.

et al.

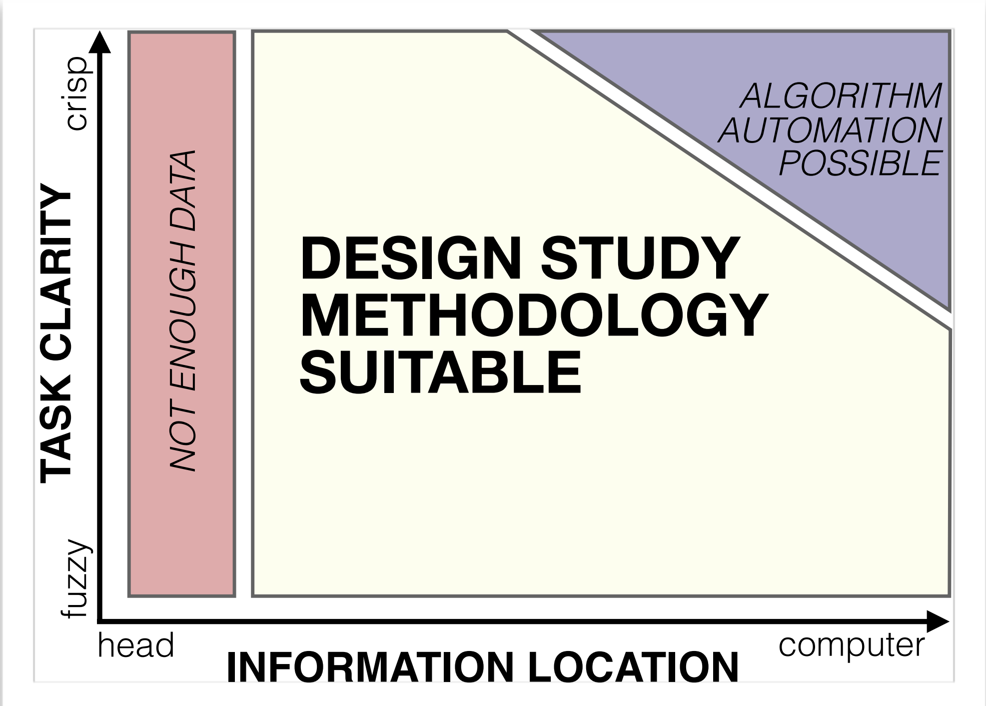


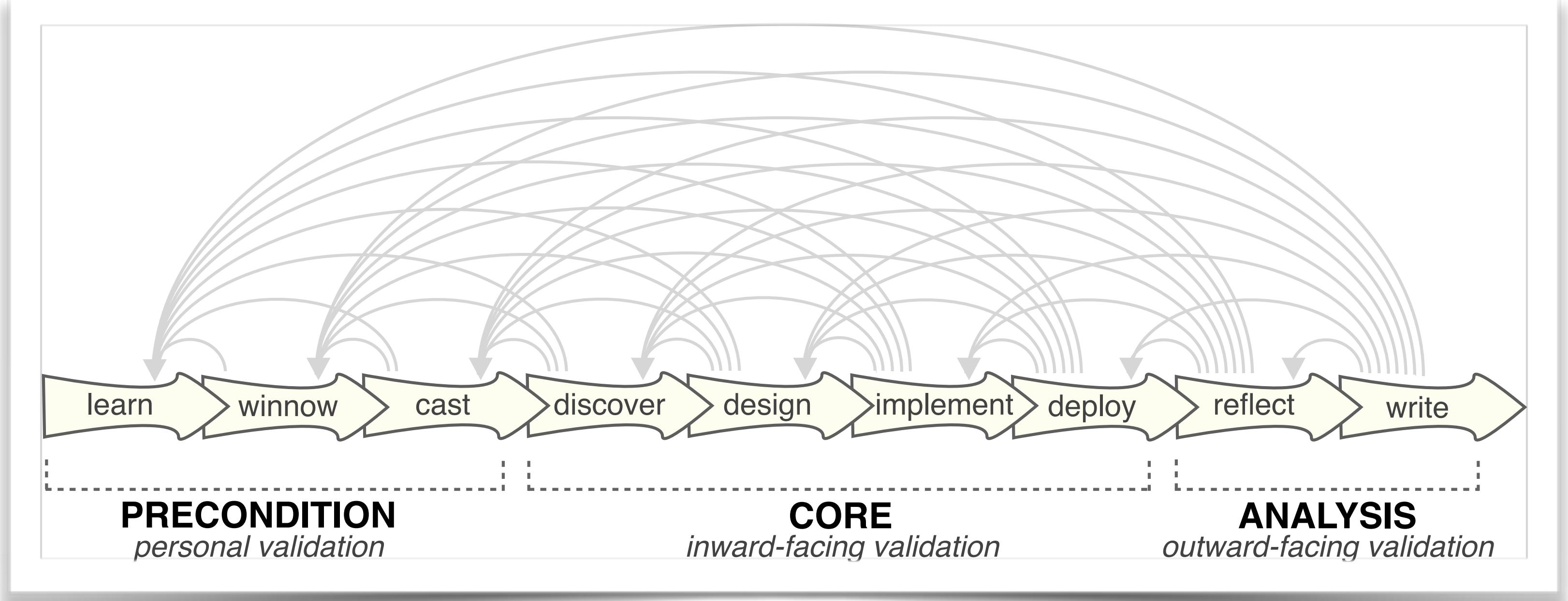




How to do design studies?

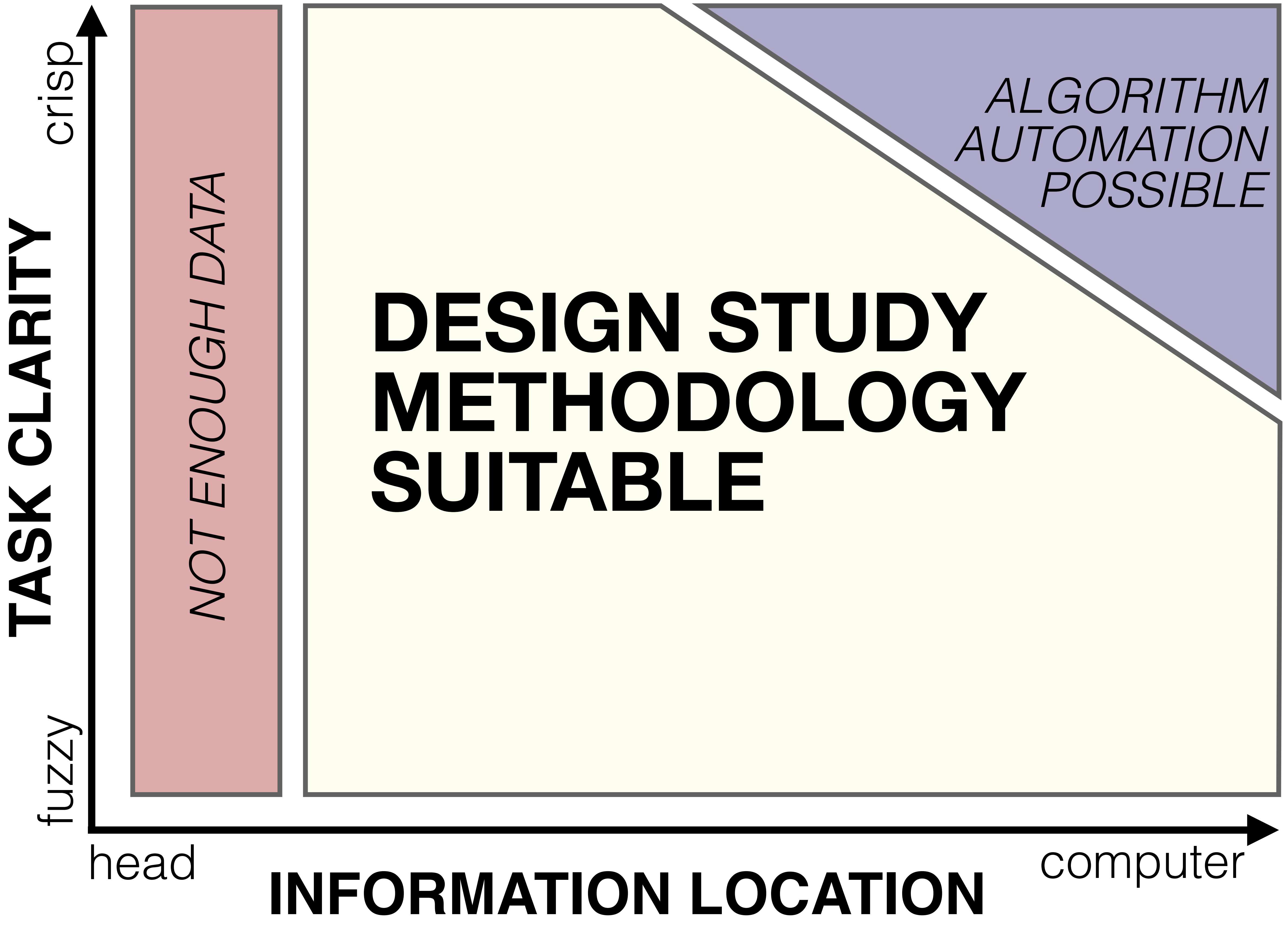
- Definitions
- 9-stage framework
- 32 pitfalls





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

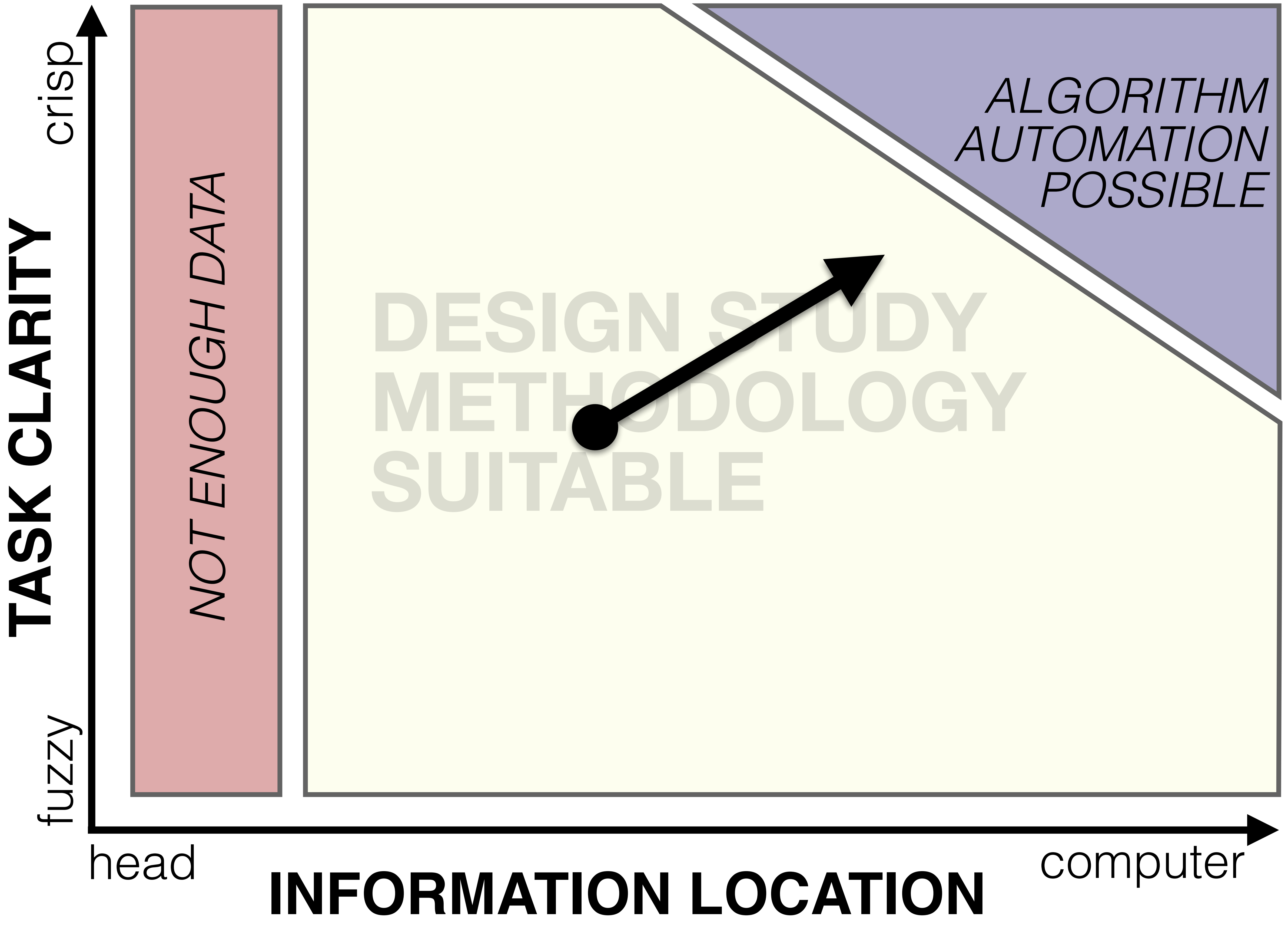
Why do design studies?



Examples:

- finding unknown errors (Cardiogram)
- exploratory analysis of scientific problems
- decision-making problems
- model building & validation

Why do design studies?

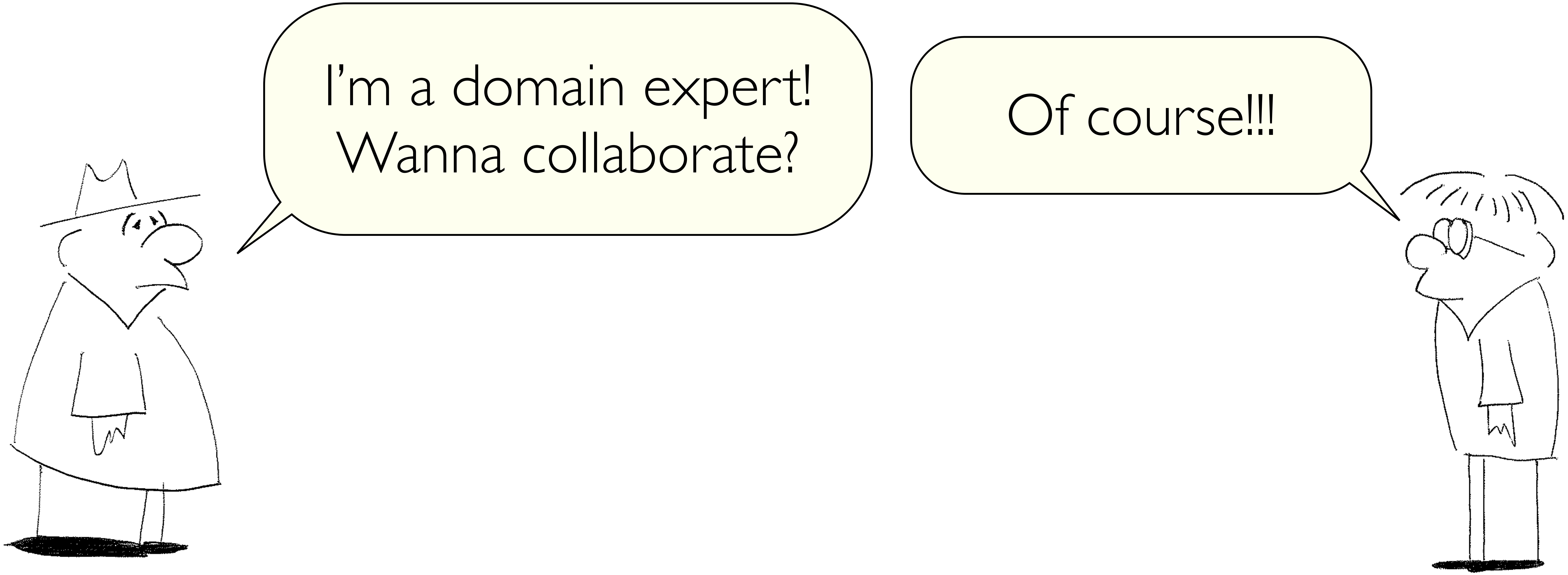


Contributions:

- problem characterization
- visual analysis tool
- reflection

32 pitfalls

example: premature commitment



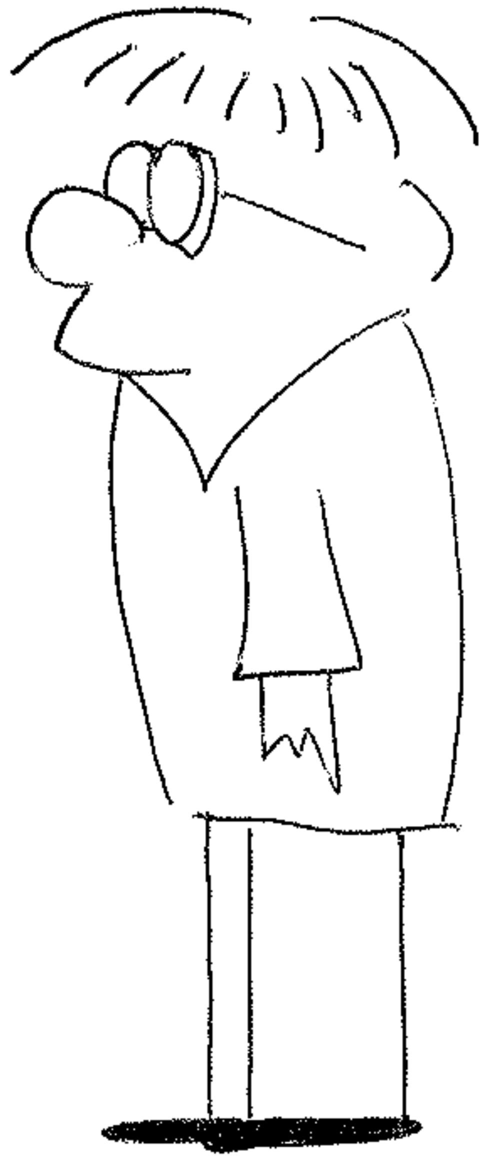
Collaborator

Scientist

Important considerations

Collaborator

Have **data**?
Have **need**?
Does the **problem** fit?
...



Scientist

Future work:

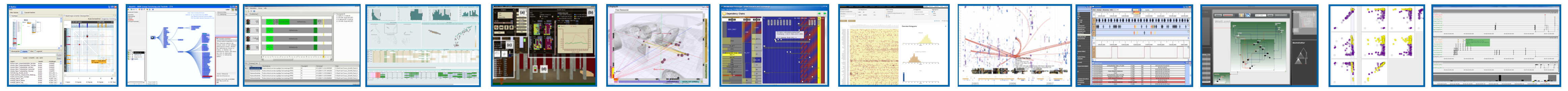
3 major challenges in human-centered data science

Challenge 1:
Understand human factors in Data Science

- **practices, startegies, stumbling blocks**
  - field studies

Sedlmair, Brehmer, Ingram, Munzner.
Dimensionality Reduction in the Wild: Gaps and Guidance.
UBC Computer Science Technical Report TR-2012-03.

- **mapping to solutions**
  - design studies



Sedlmair et al., A Dual-View Visualization of In-Car Communication Processes, *IV 2008.*

Ruhland et al., LibViz: Data Visualisation of the old Library, *VSMM 2008.*

Sedlmair et al., User-centered Development of a Visual Exploration System for In-Car Communication, Smart Graphics, *2009.*

Sedlmair et al., MostVis: An Interactive Visualization Supporting Automotive Engineers in MOST Catalog Exploration, *IV 2009.*

Baur et al., The Streams of Our Lives: Visualizing Listening Histories in Context, *IEEE TVCG (Proc. InfoVis) 2010.*

Sedlmair et al., Cardiogram: Visual Analytics for Automotive Engineers, ACM CHI *2011.*

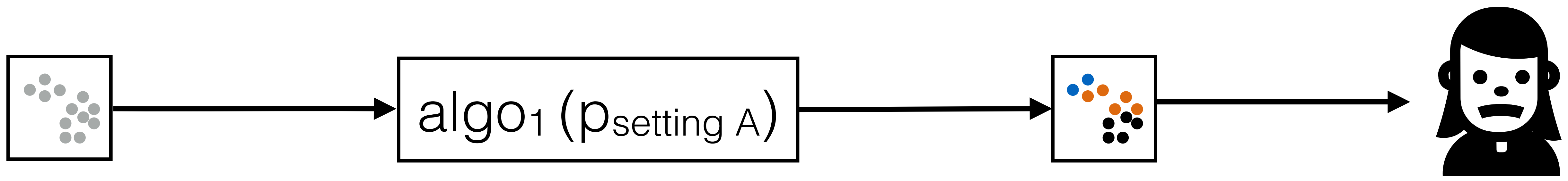
Sedlmair et al., RelEx: Visualization for Actively Changing Overlay Network Specifications, *IEEE TVCG (Proc. InfoVis) 2012.*

Bergner et al., Paraglide: Interactive Parameter Space Partitioning for Computer Simulations, *IEEE TVGC 2013.*

Hund et al., Visual Analytics for Concept Exploration in Subspaces of Patient Groups, *to appear, BRIN 2016.*

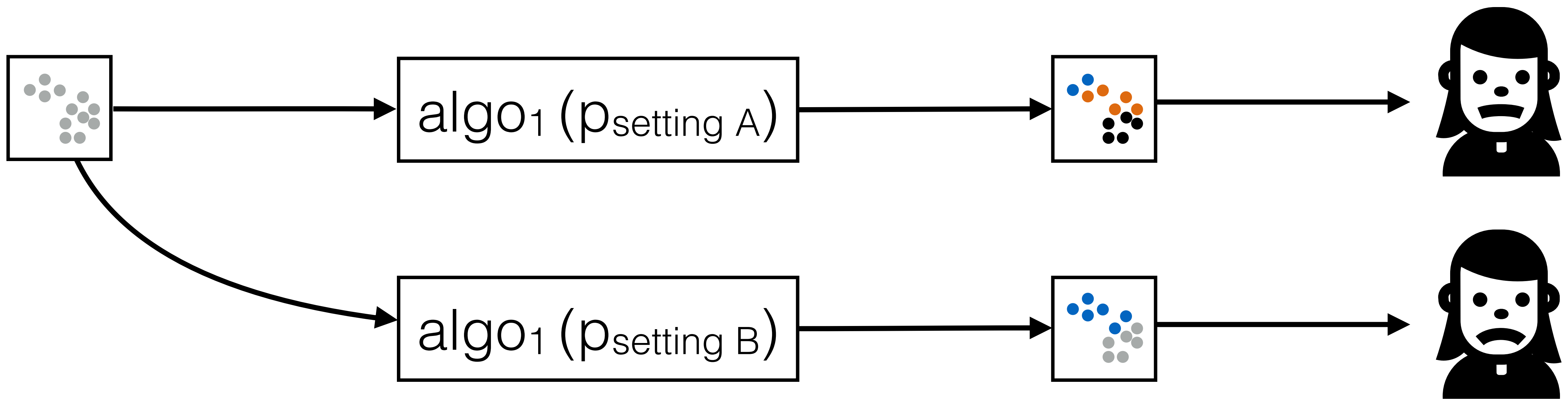
Challenge 2:
Guide human users in algorithmic choices

- traditional approach: trial and error (e.g., clustering)



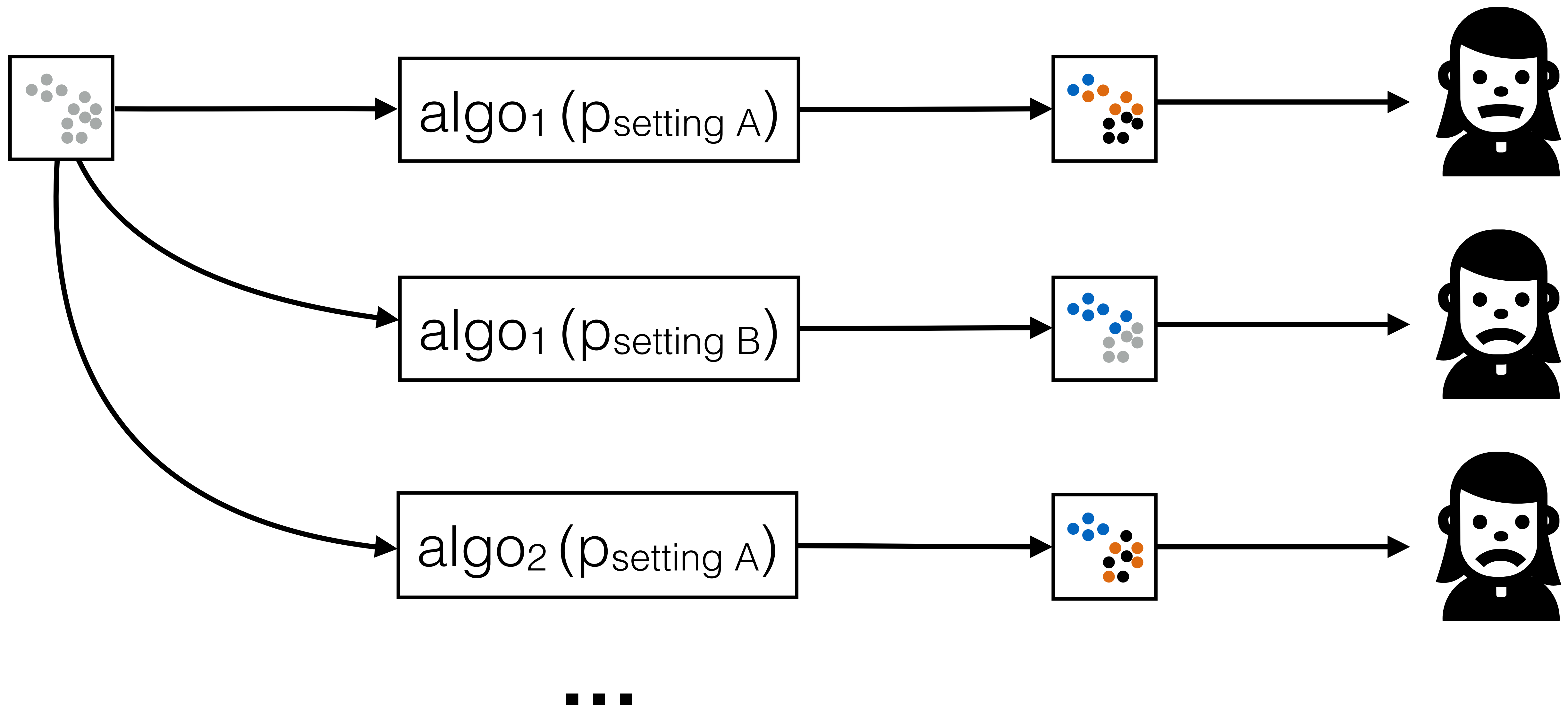
Challenge 2:
Guide human users in algorithmic choices

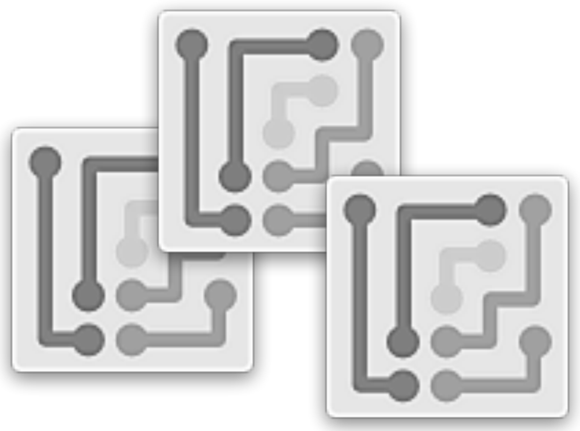
- traditional approach: trial and error (e.g., clustering)



Challenge 2:
Guide human users in algorithmic choices

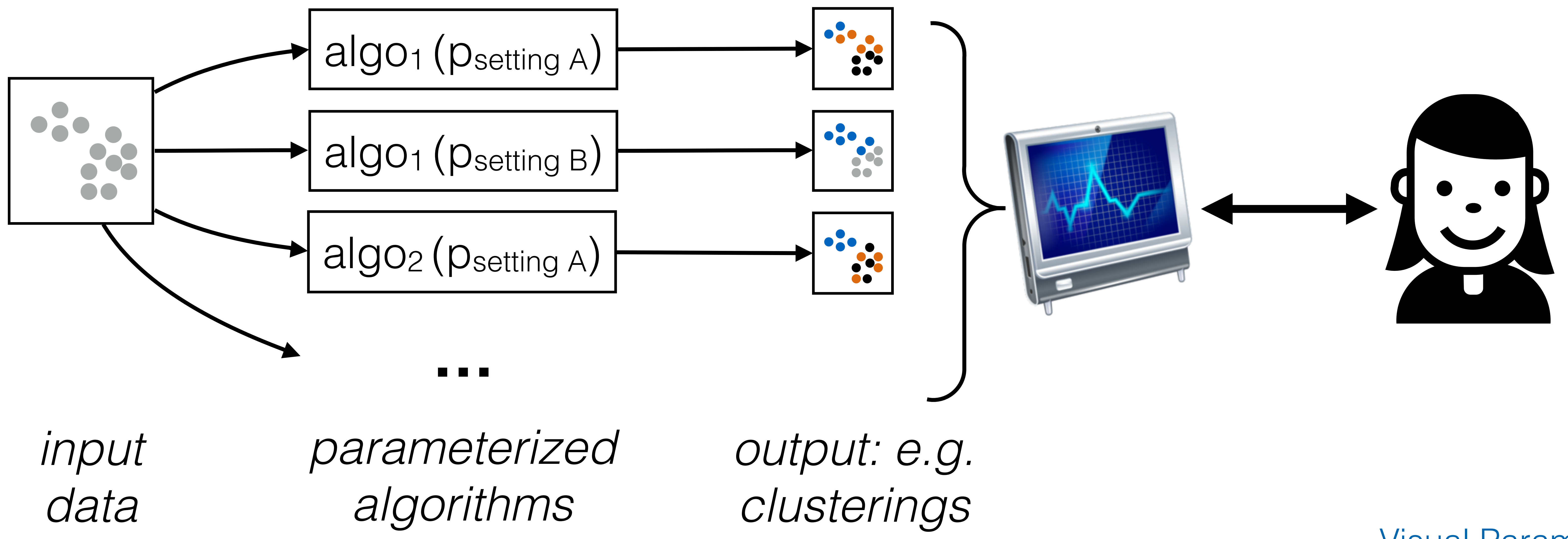
- traditional approach: trial and error (e.g., clustering)





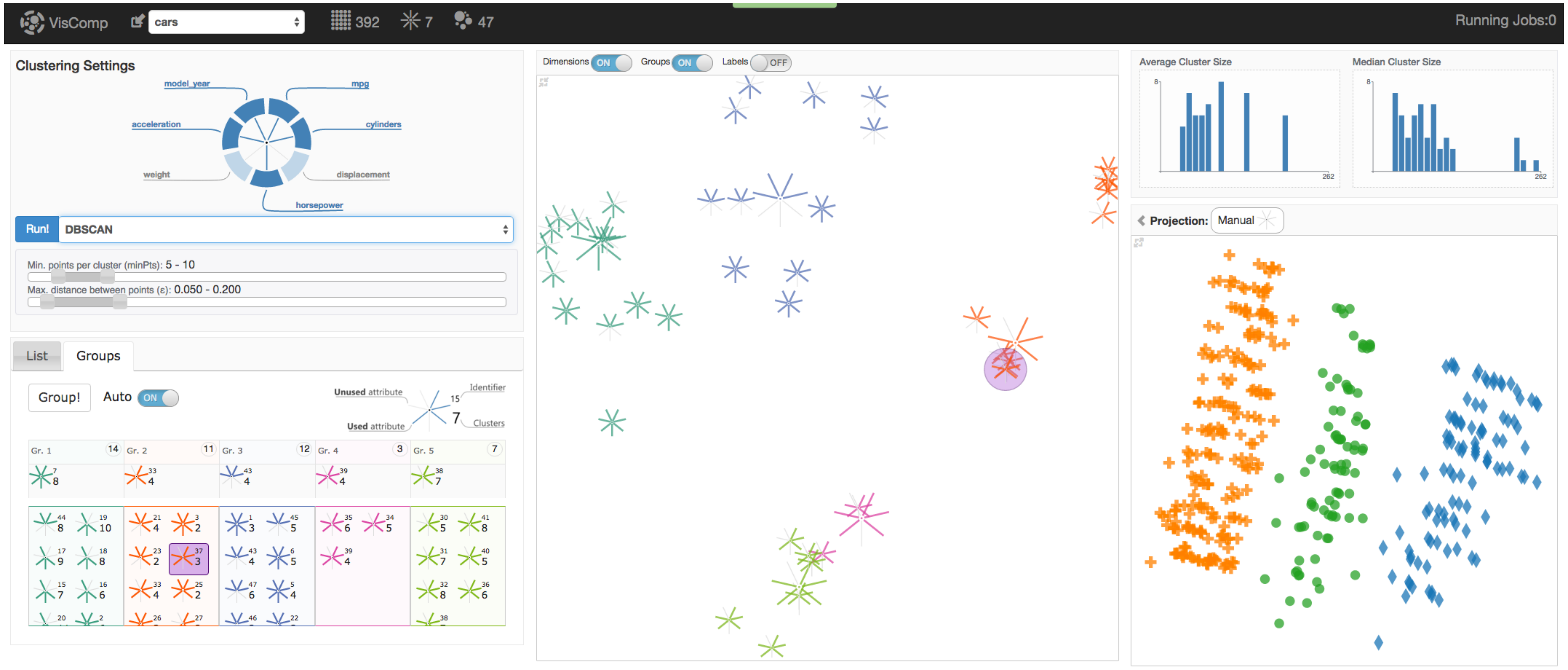
Challenge 2:
Guide human users in algorithmic choices

- traditional approach: trial and error (e.g., clustering)
- instead: visual parameter space analysis



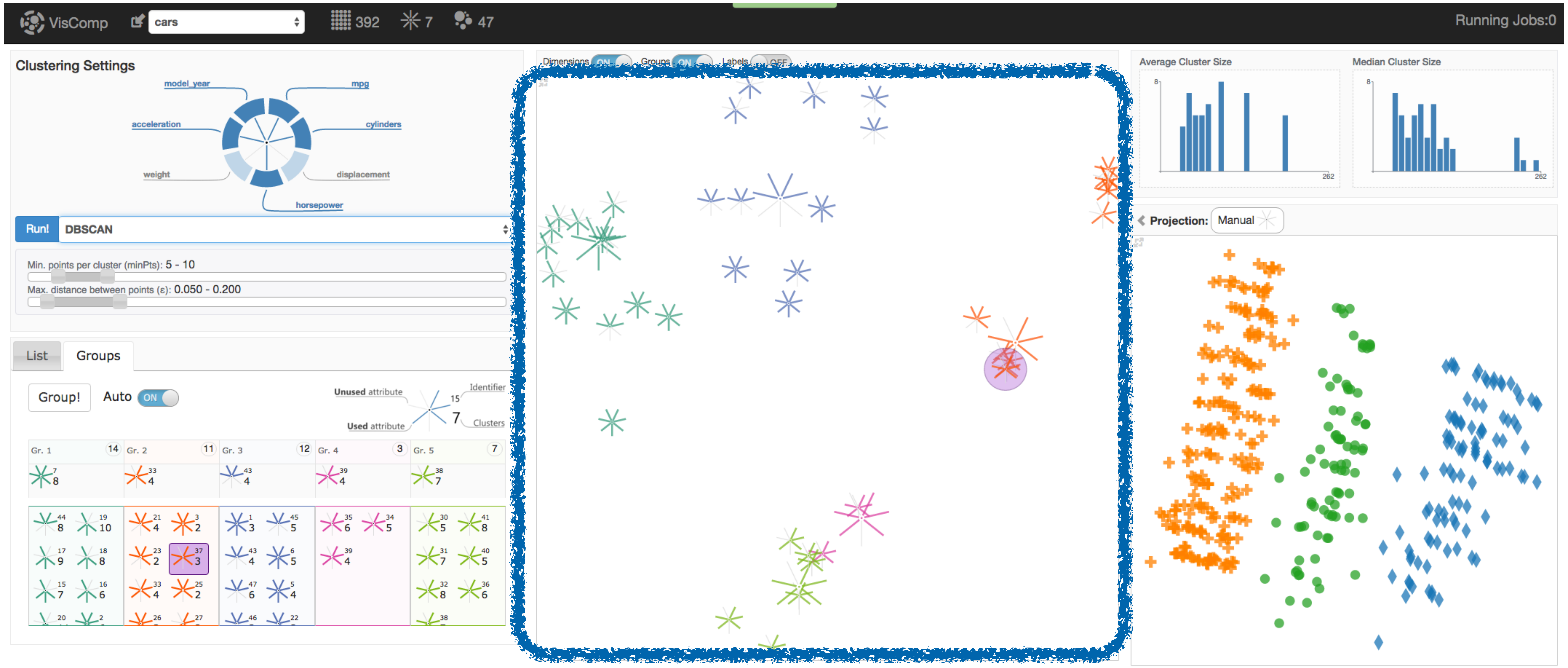
Sedlmair, Heinzl, Bruckner, Piringer, Möller.
Visual Parameter Space Analysis: A Conceptual Framework.
IEEE TVCG (Proc. InfoVis), 20(12): 2161-2170, 2014.

Challenge 2:
Example — ClustComp (ongoing work)



Challenge 2:

Example — ClustComp (ongoing work)

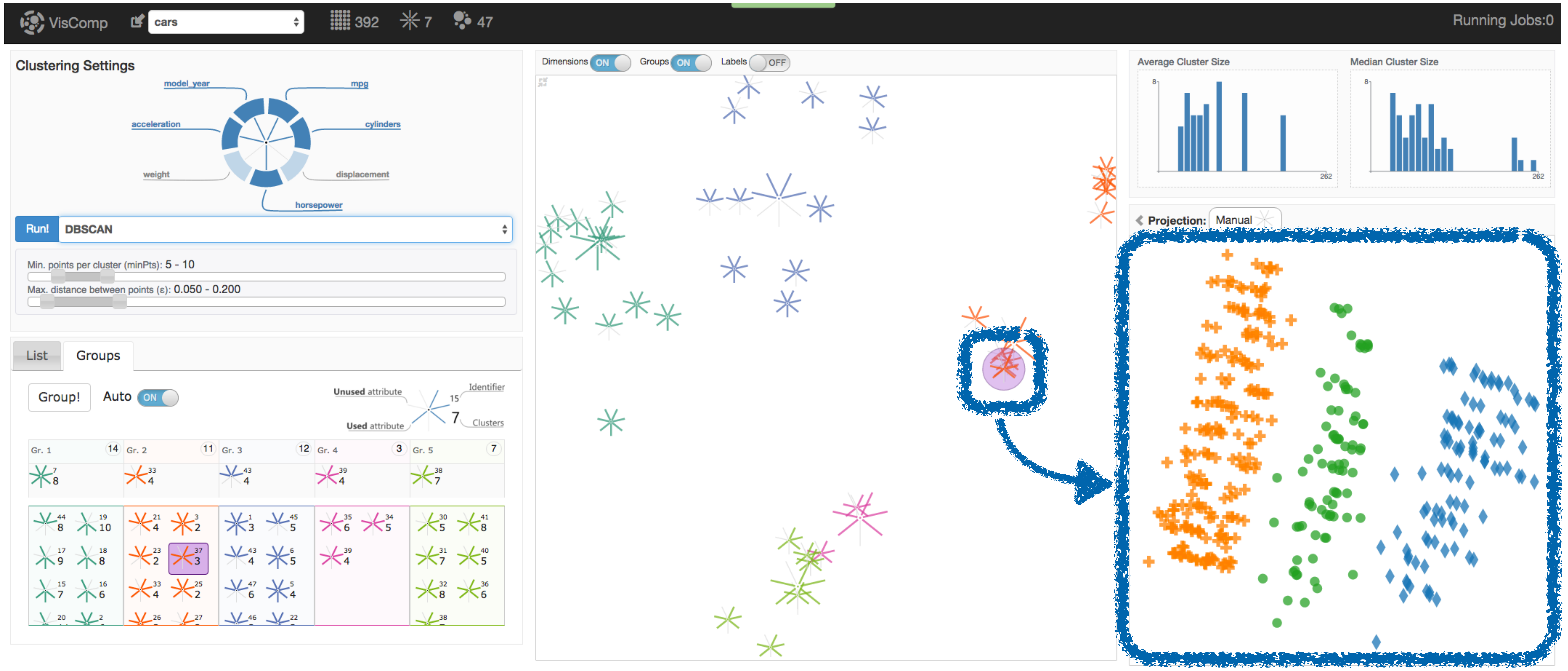


Overview of clusterings

(distance = similarity)

Challenge 2:

Example — ClustComp (ongoing work)

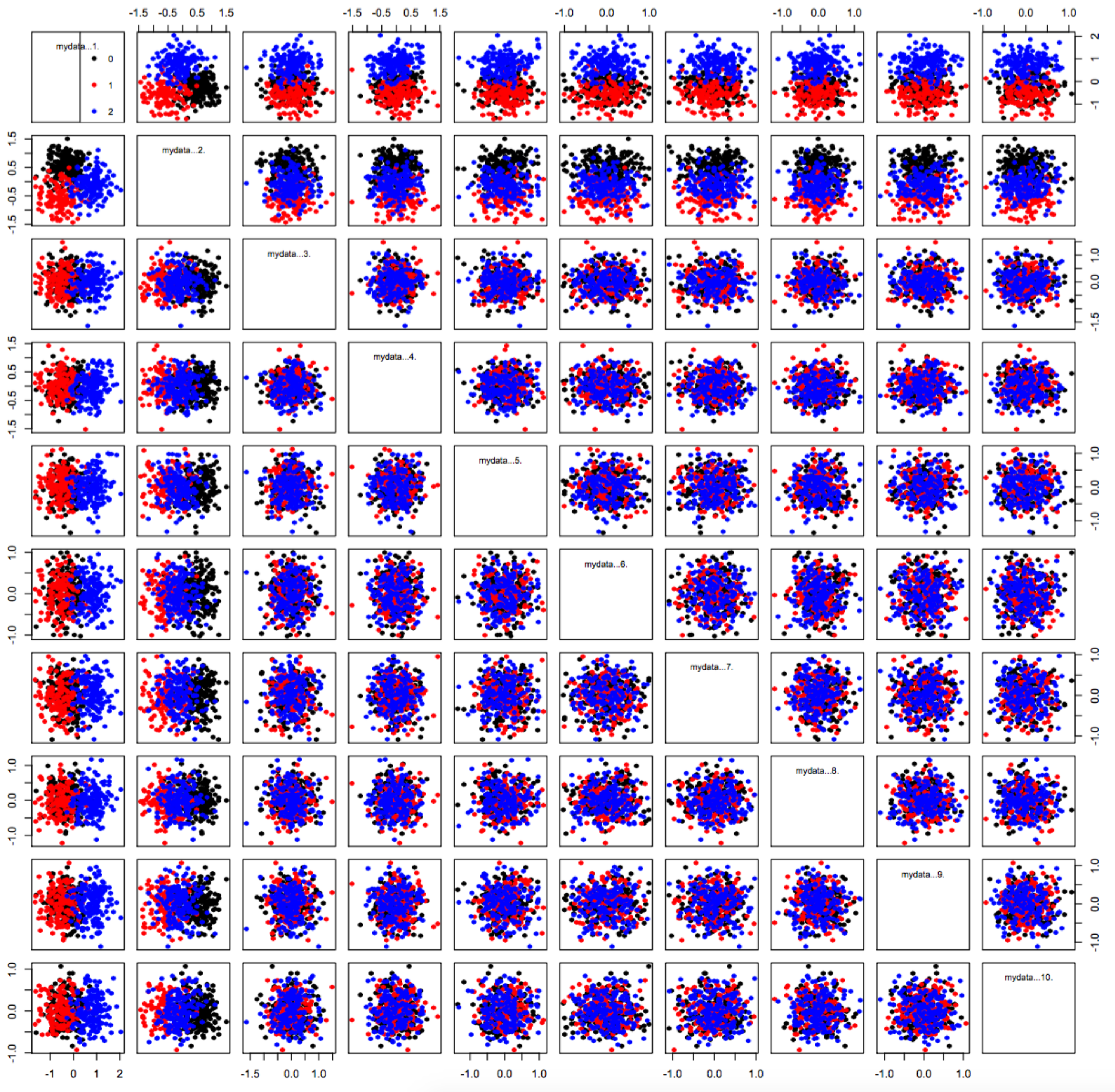


Show cluster results

(upon selecting one)

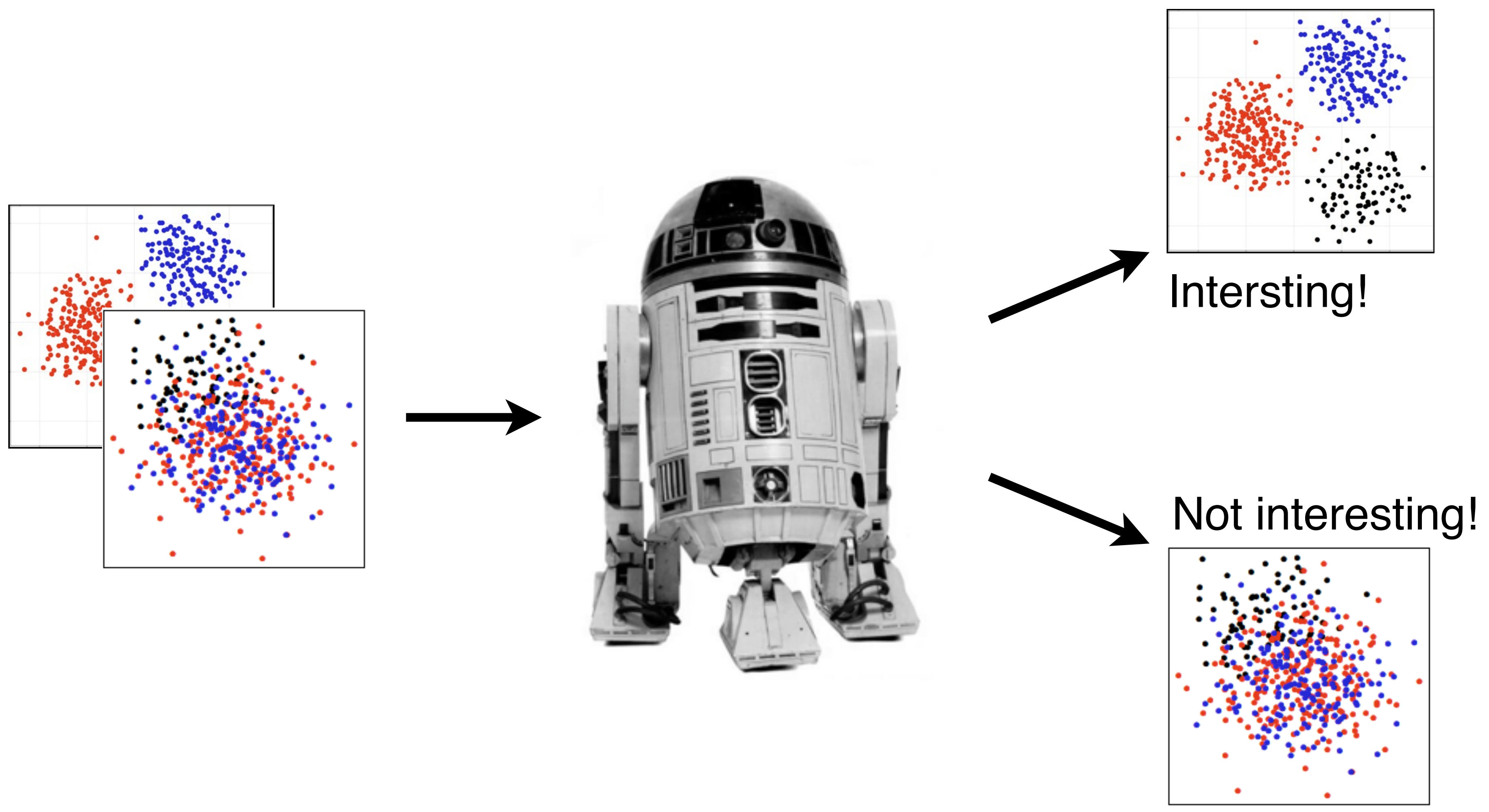
Challenge 3:
Guide human users in visualization choices

- traditional approach: manual



Challenge 3:
Guide human users in visualization choices

- traditional approach: manual
- instead: algorithmically modeling human perception



Sedlmair, Tatu, Munzner, Tory
A Taxonomy of Visual Cluster Separation Factors
Computer Graphics Forum (Proc. EuroVis), 31(3): 1335-1344, 2012.

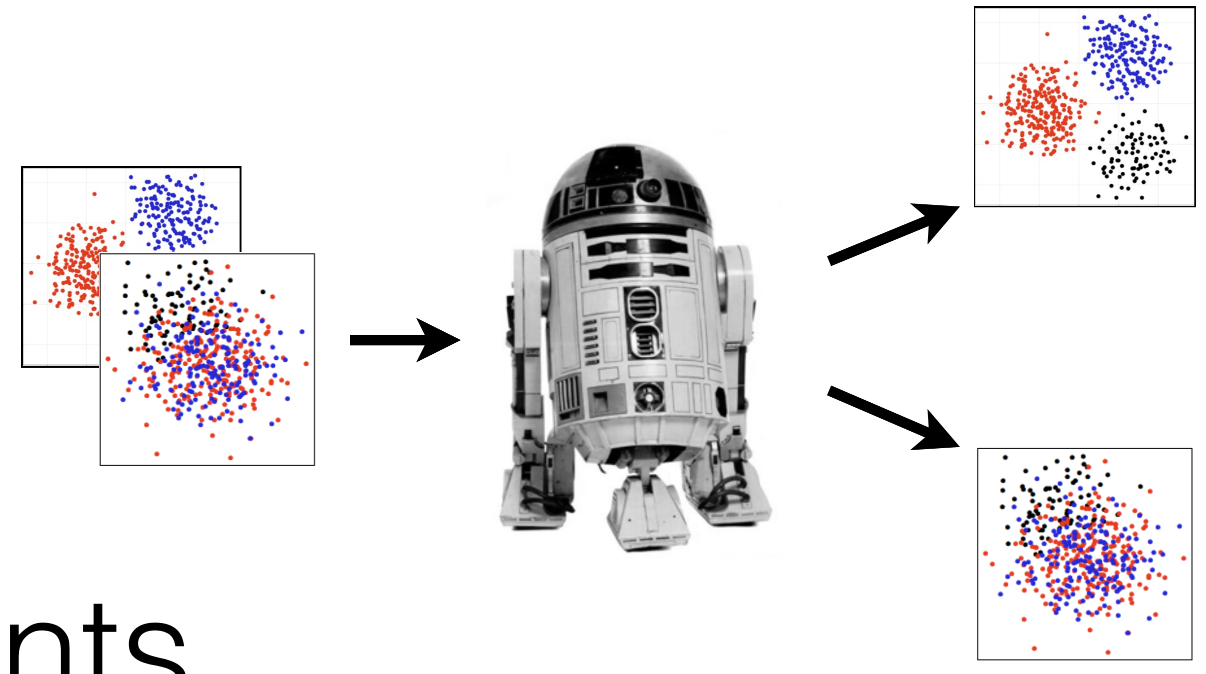
Sedlmair, Munzner, Tory
Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices.
IEEE TVCG (Proc. InfoVis), 19(12): 2634-2643, 2013.

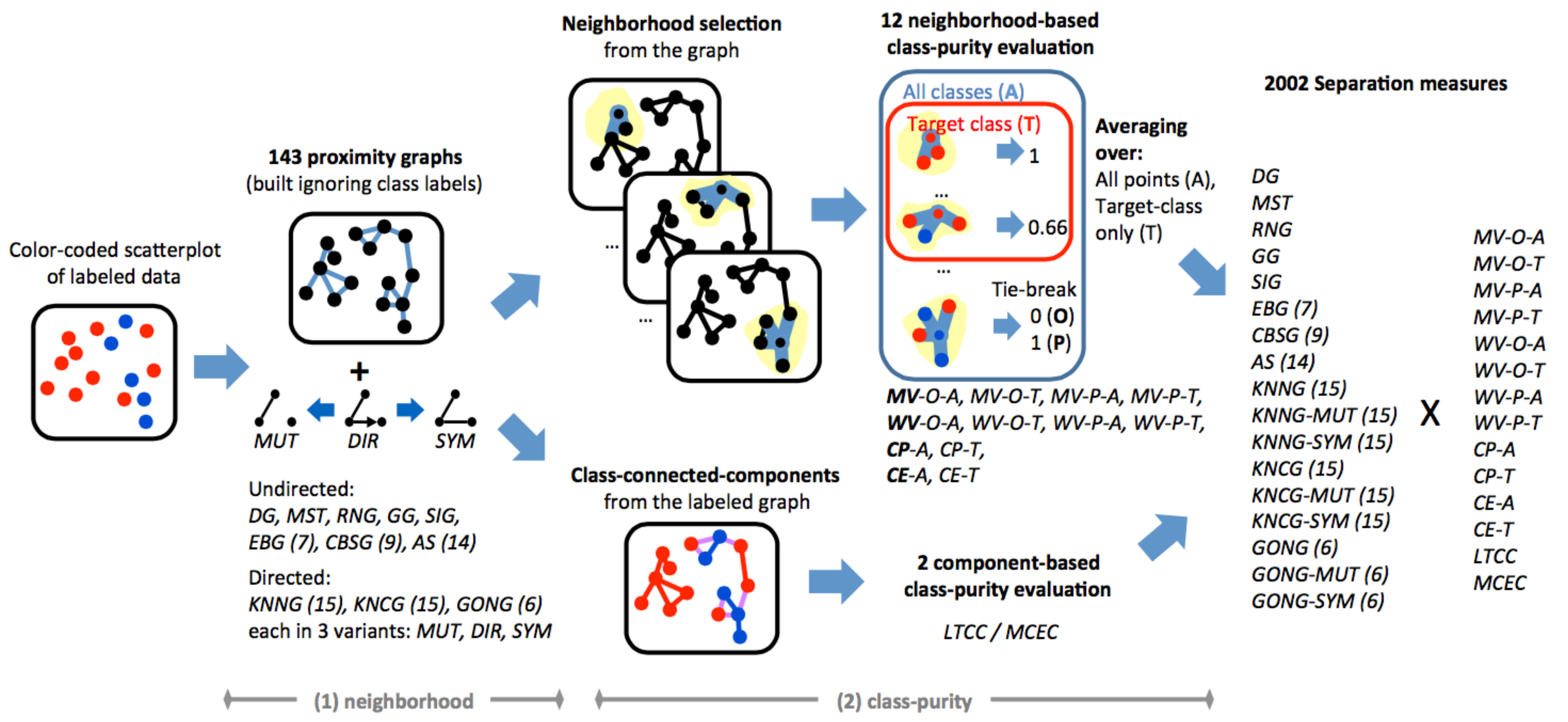
Sedlmair, Aupetit
Data-driven Evaluation of Visual Quality Measures.
Computer Graphics Forum (Proc. EuroVis), 34(3): 201-210, 2015.

Aupetit, **Sedlmair**
SepMe: 2002 New Visual Separation Measures
IEEE Pacific Visualization Symposium, 2016 (in submission).

Challenge 3:
Example — Visual Separation Measures

- Basically, a **computer vision problem**
  - gathered human judgements of 816 scatterplots
  - trained 2002 measures predicting human judgements
  - significantly outperformed state-of-the-art measures!





$$AWTN(C_s) = \frac{\sum_{u\in\{t,o\}} N_u D_{C_u}}{\sum_{u\in\{t,o\}} N_u}$$

$$CS(C_s) = \frac{1}{N_s}\sum_{u\in\{t,o\}}\sum_{x\in C_u}\frac{|ExtMST(x)\cap C_u|}{|ExtMST(x)|}$$

$$DC(C_s) = 1 - \frac{\sum_{z\in\mathcal{I}}|\epsilon NN(z)|H_{\epsilon NN}(z)}{\sum_{z\in\mathcal{I}}|\epsilon NN(z)|}$$

$$CAL(C_s) = \frac{\sum_{u\in\{t,o\}} N_u d^2_{c_u,c_s}}{\sum_{u\in\{t,o\}}\sum_{x\in C_u} d^2_{c_u,x}}$$

$$DUNN(C_s) = \frac{\min_{(x,y)\in S_{btn}} d_{x,y}}{\max_{(x,y)\in S_{wtn}} d_{x,y}}$$

$$DC(C_s) = 1 - \frac{\sum_{z\in\mathcal{I}}|\epsilon NN(z)|H_{\epsilon NN}(z)}{\sum_{z\in\mathcal{I}}|\epsilon NN(z)|}$$

$$CDM(C_s) = \sum_{z\in\mathcal{I}}|\rho_{C_t}(z,K) - \rho_{C_o}(z,K)|$$

$$GAM(C_s) = \frac{d^+ - d^-}{d^+ + d^-}$$

etc. etc.

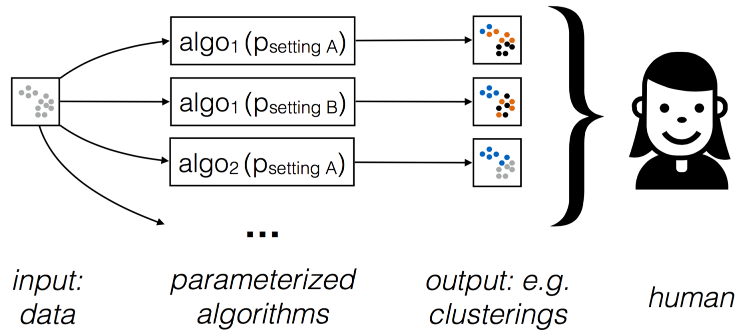
Aupetit, **Sedlmair**
SepMe: 2002 New Visual Separation Measures
IEEE Pacific Visualization Symposium, 2016 (in submission).

How does it fit into TUM?

very nicely, many opportunities …

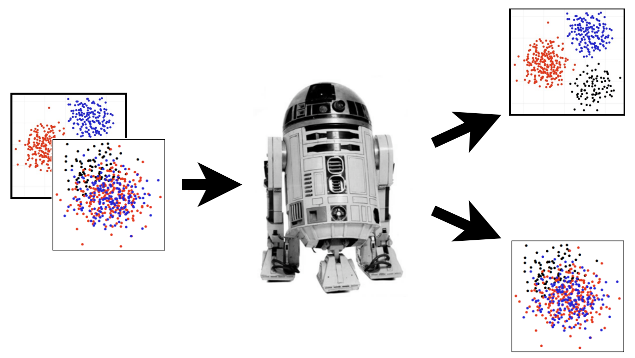
Human-centered approaches

➡ *HCI (e.g., Prof. Klinker, Prof. Bengler)*



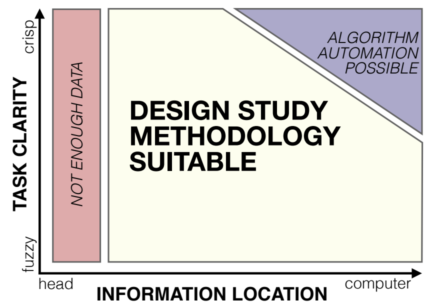
Visual parameter space analysis

➡ *Ensemble/Uncertainty Visualization (e.g., Prof. Westermann)*



Visual quality measures

➡ *Computer Vision (e.g., Prof. Cremers)*

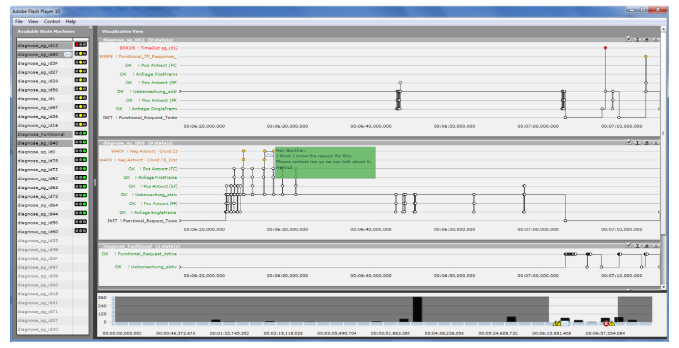


Interdisciplinary design studies

➡ *ill-defined scientific problems (plenty of opportunities at TUM-IAS)*

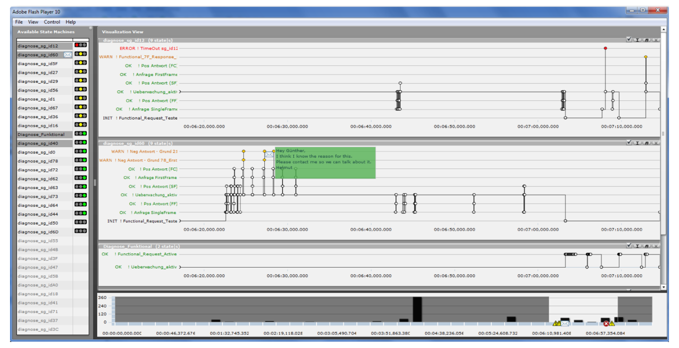
Summary

Human-centered Data Science

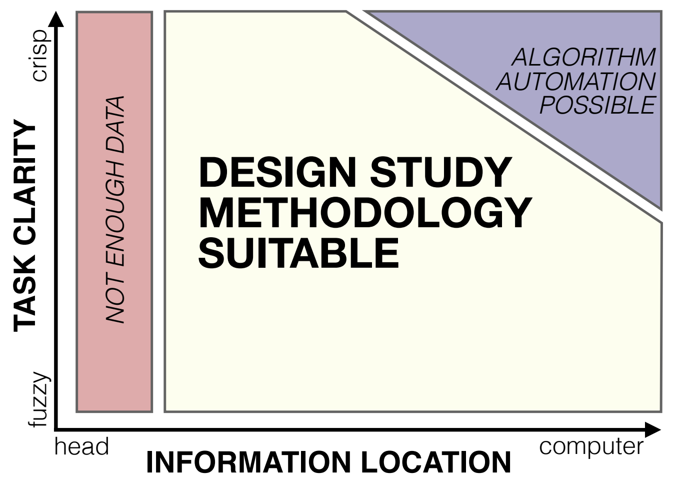


- Cardiogram Design Study
  - impact: adopted by BMW engineers

Human-centered Data Science



- Cardiogram Design Study
    - impact: adopted by BMW engineers



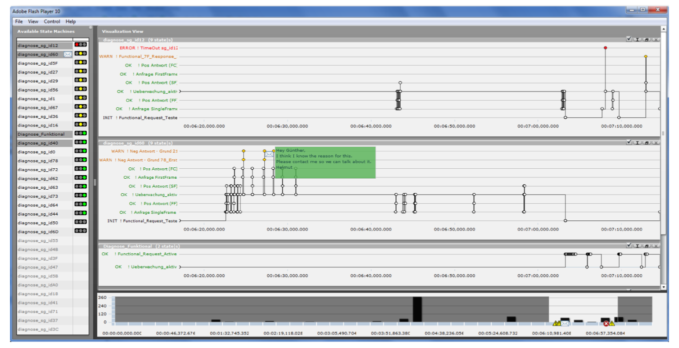
- Design Study Methodology
    - impact: used and well-cited

```
hi folks,

I just read the design study paper by sedlmair (thanks manu for sending
it to me) and it was quite revealing for me. I am so glad that I read it
at the beginning of my phd. If you have any papers that you consider as
a must read for every phd-student, then pls let know.

haichao
```

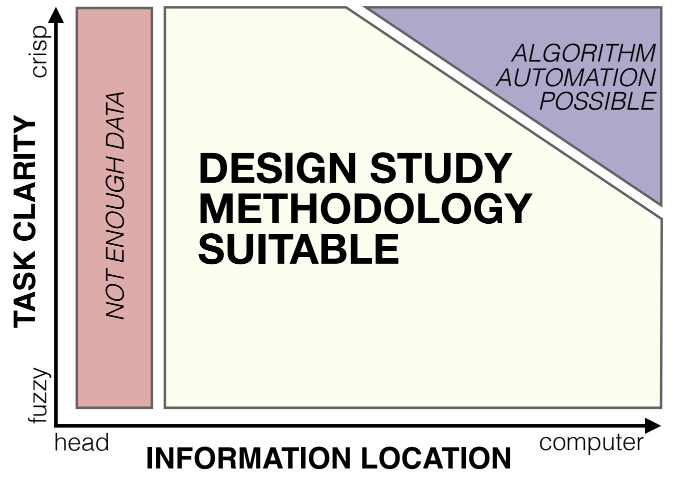
Haichao Miao,
PhD student, TU Vienna
(forwarded by Manu Waldner)

bold and color added

Human-centered Data Science



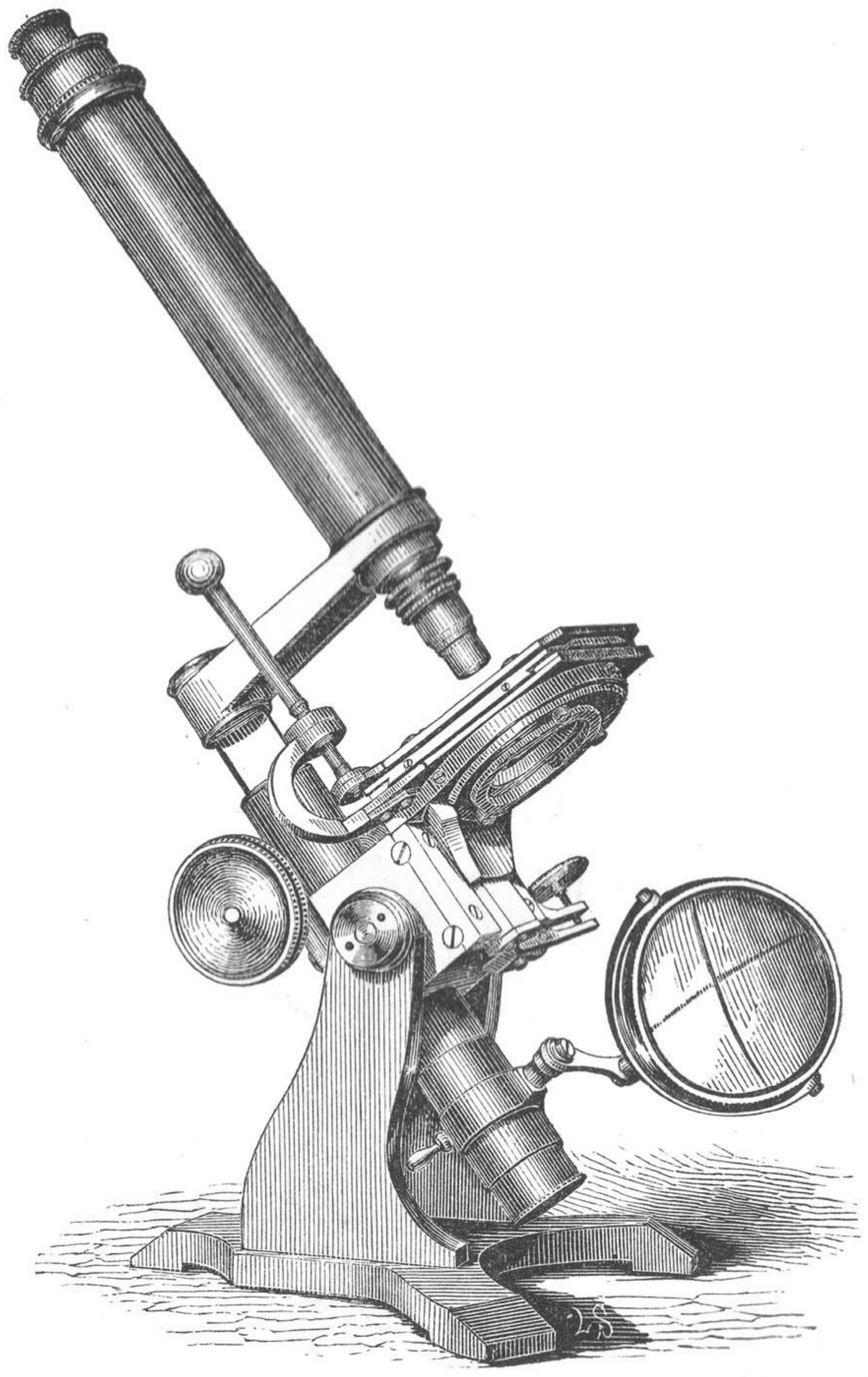
- Cardiogram Design Study
  - impact: adopted by BMW engineers



- Design Study Methodology
  - impact: used and well-cited

- Future Work
  - vision: help democratizing Data Science

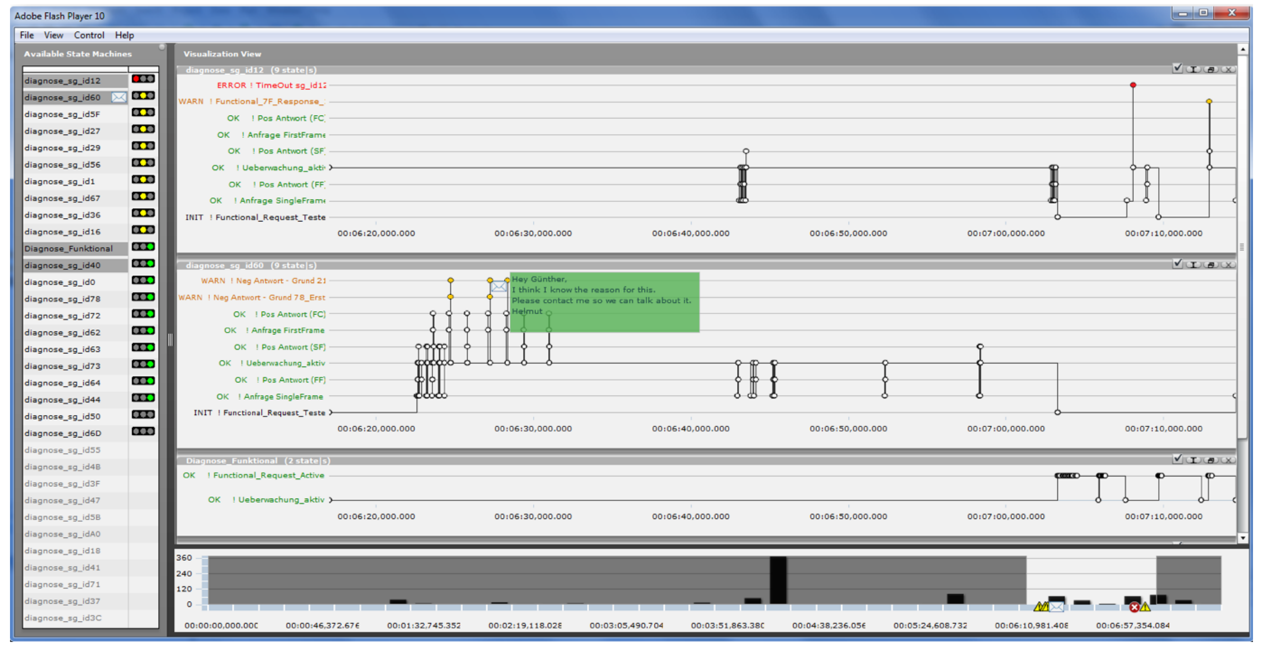
Human-centered Data Science

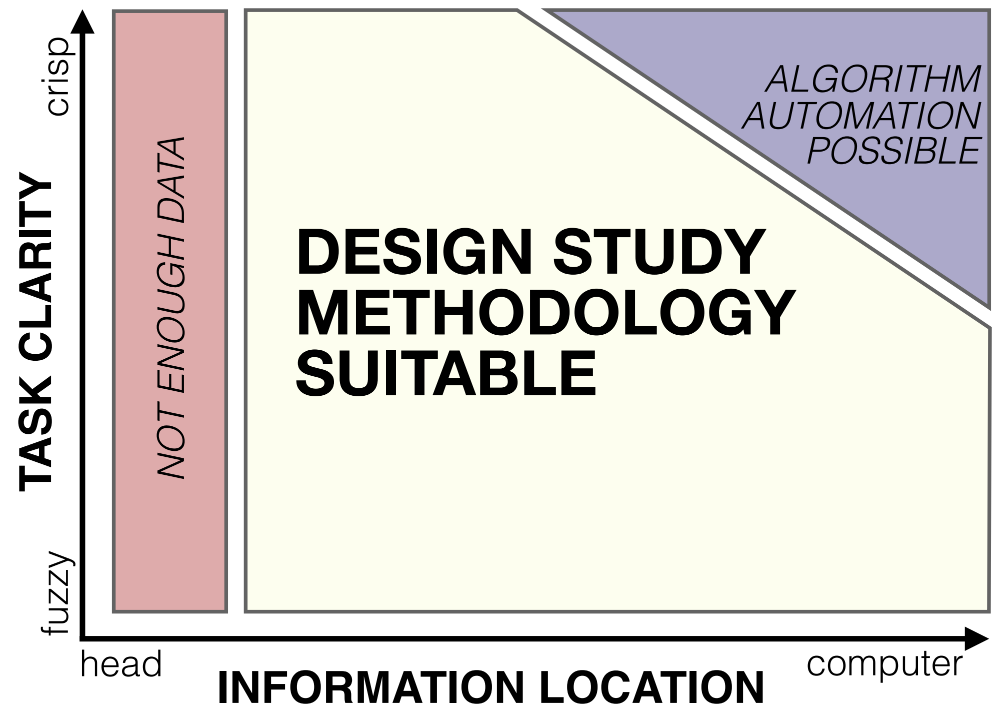


http://www.microscope-antiques.com/grunow.html

http://galileo.rice.edu/sci/instruments/telescope.html

Thank you!





Human-centered Data Science

Michael Sedlmair
Visualization & Data Analysis Group
universität wien