Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, *University of Vienna*
Tamara Munzner, *University of British Columbia*
Melanie Tory, *University of Victoria*

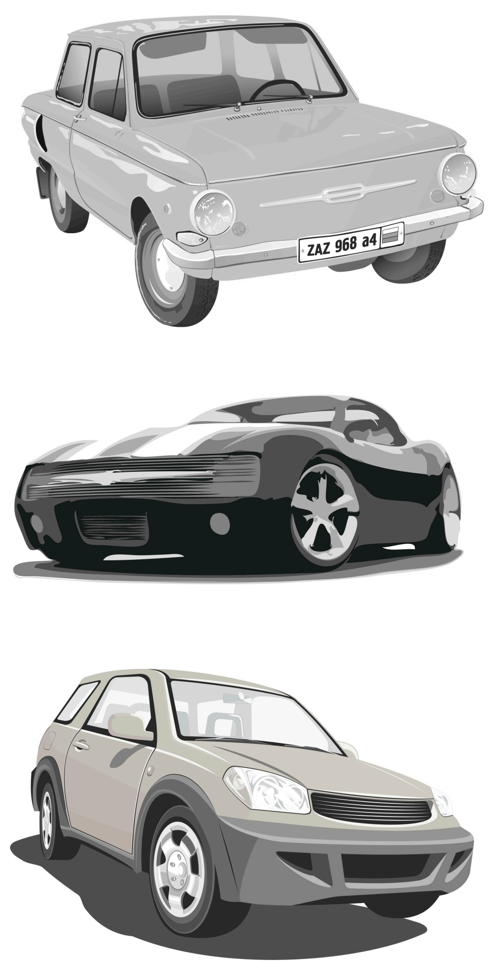


High-dimensional Data

	length	weight	speed	hp	...
Car 1					
Car 2					
Car 3					
...					

highdim

Dimension Reduction (DR)



	length	weight	speed	hp	...
Car 1					
Car 2					
Car 3					
...					

highdim

DR

e.g., using PCA

	sporty	handling
Car 1		
Car 2		
Car 3		
...		

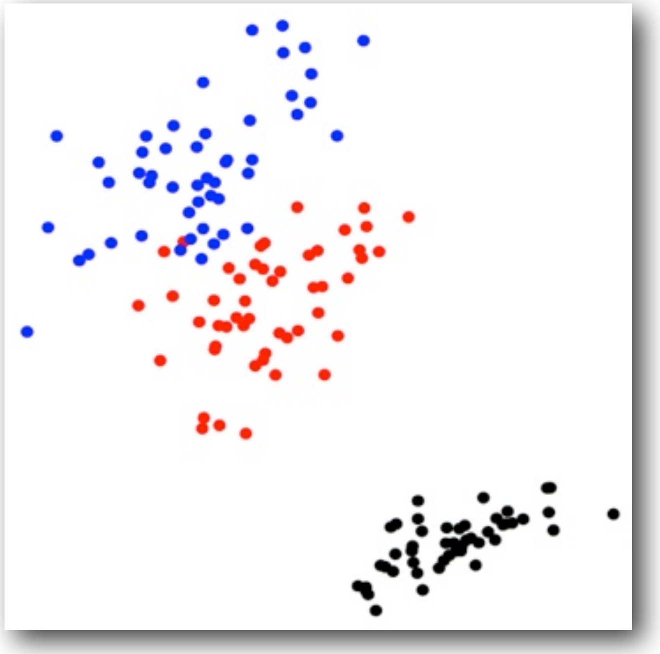
lowdim

Visualizing DR Data

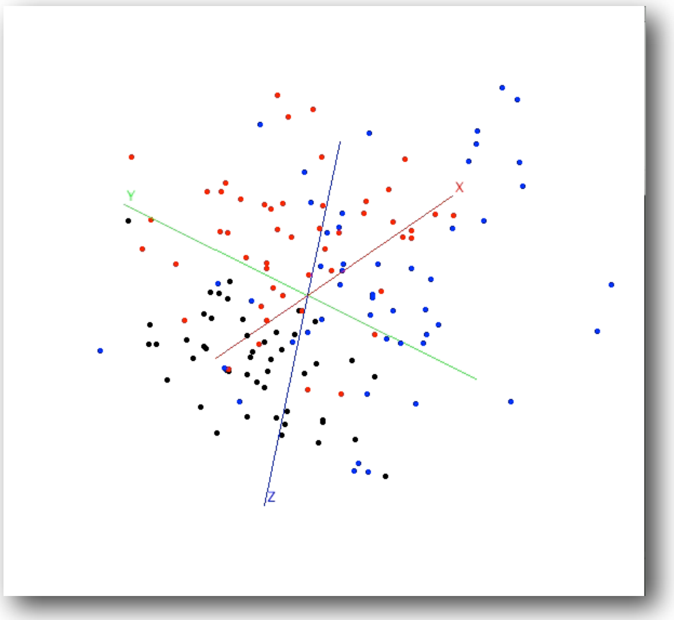
	sporty	handling
Car 1		
Car 2		
Car 3		
...		

lowdim

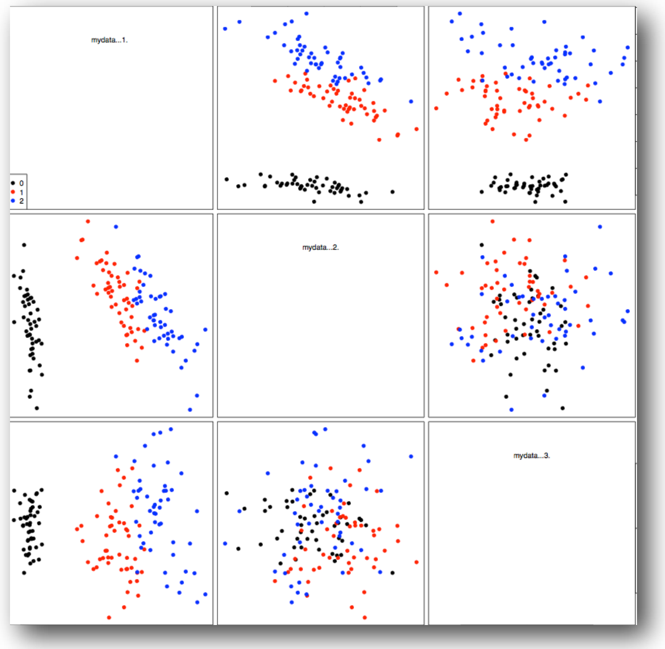
Visualization →



2D Scatterplot



interactive 3D Scatterplot



Scatterplot Matrix (SPLOM)

Which visual encoding technique to use for visualizing DR data?

2D, 3D, SPLOM?

Related Work

General abstract data

- 3D often inappropriate

  Chalmers: Using a landscape metaphor to represent a corpus of documents [COSIT'93]
  Cockburn and McKenzie: An evaluation of cone trees [British Conf. on HCI'00]
  Cockburn and McKenzie: Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments [CHI'02]
  Newby: Empirical study of a 3D visualization for information retrieval tasks [Intelligent Information Systems'02]
  Tory et al.: Spatialization design: comparing points and landscapes [InfoVis'07]
  Tory et al.: Comparing dot and landscape spatializations for visual memory differences [InfoVis'09]
  Westerman and Cribbin: Mapping semantic information in virtual space: dimensions, variance and individual differences [IJHCS'00]

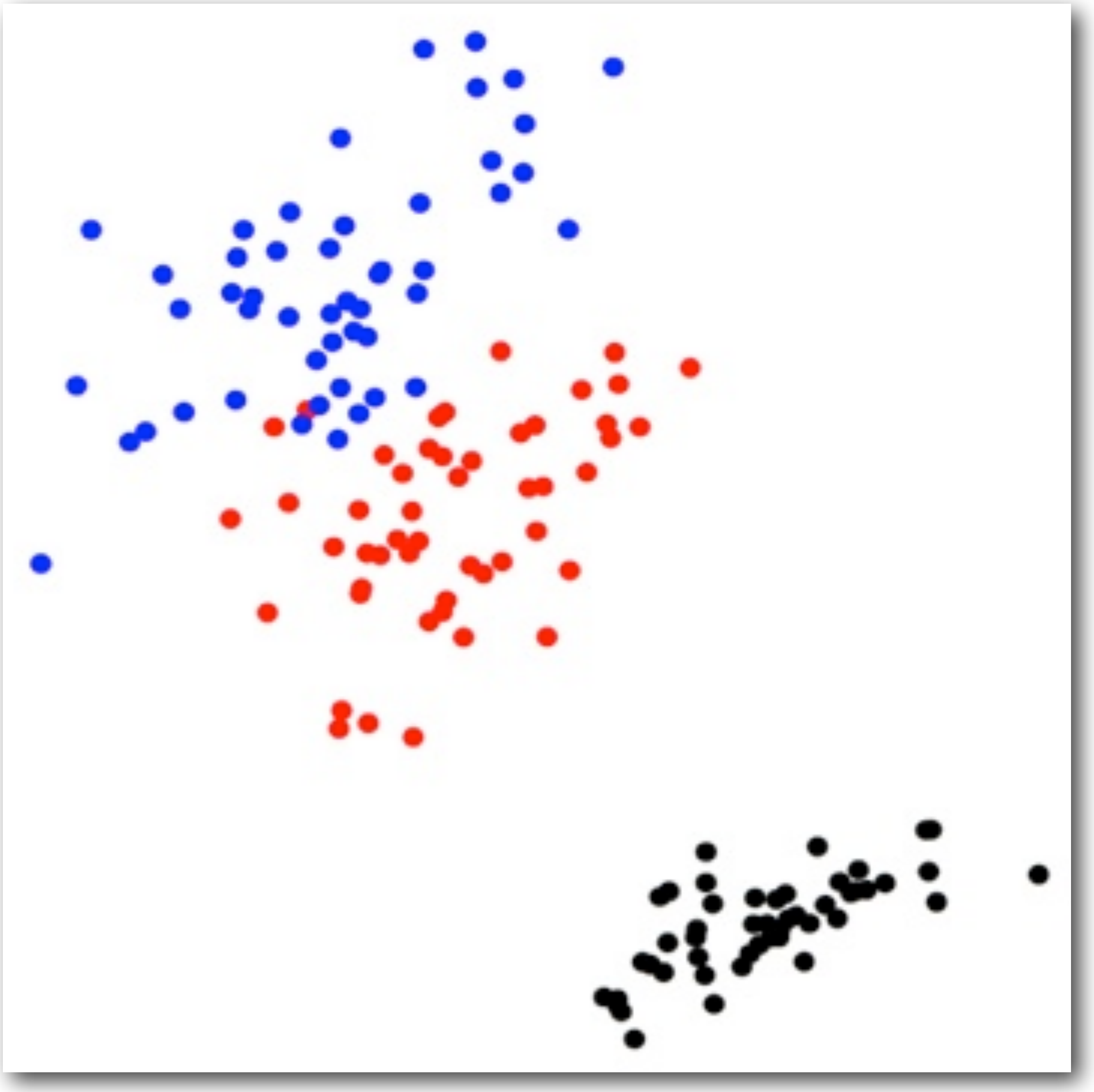
DR data

- 3D **is** used in certain domains
- No studies on scatterplot choices for DR data

Contributions

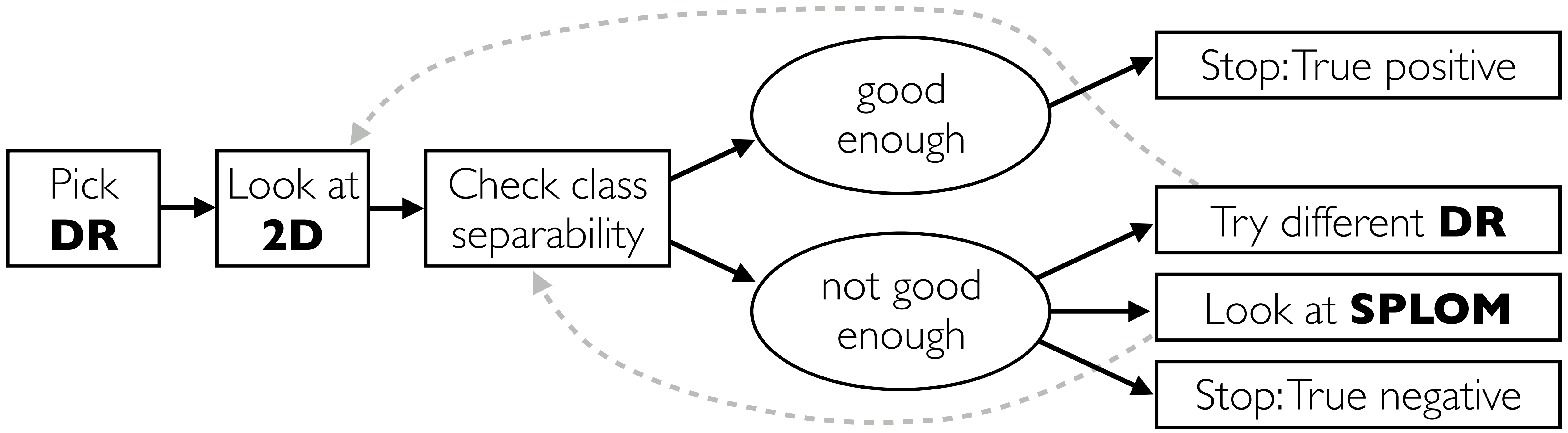
1. Data Study

- in-depth analysis of 816 scatterplots
- task: visual cluster verification

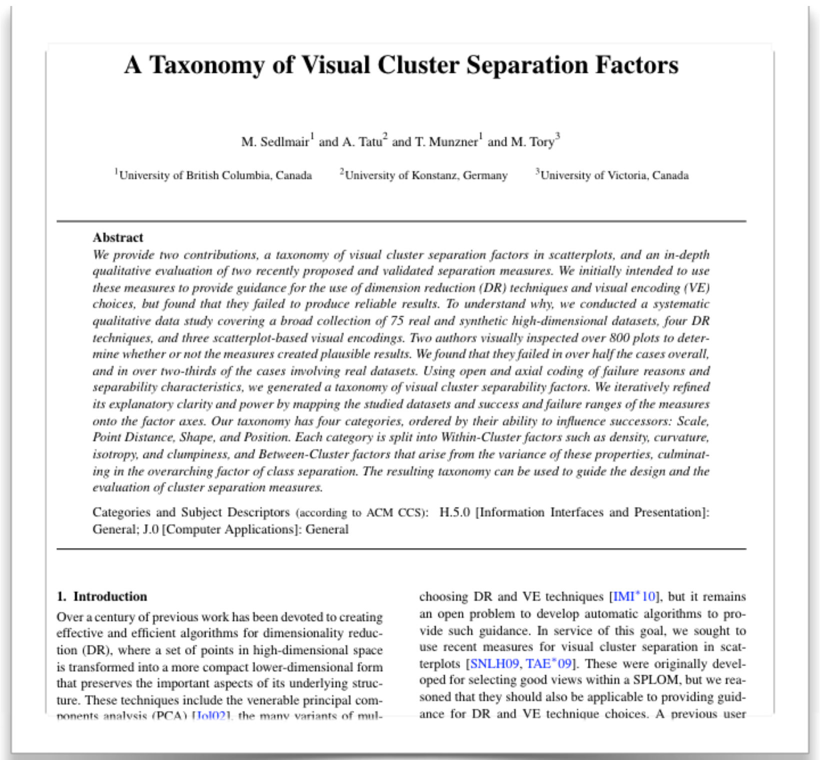


Contributions

1. Data Study
   - qualitative analysis of 816 scatterplots
   - task: visual cluster verification
2. Workflow Model



(see paper)



A Taxonomy of Visual Cluster Separation Factors

M. Sedlmair[1] and A. Tatu[2] and T. Munzner[1] and M. Tory[3]

[1]University of British Columbia, Canada [2]University of Konstanz, Germany [3]University of Victoria, Canada

Abstract

We provide two contributions, a taxonomy of visual cluster separation factors in scatterplots, and an in-depth qualitative evaluation of two recently proposed and validated separation measures. We initially intended to use these measures to provide guidance for the use of dimension reduction (DR) techniques and visual encoding (VE) choices, but found that they failed to produce reliable results. To understand why, we conducted a systematic qualitative data study covering a broad collection of 75 real and synthetic high-dimensional datasets, four DR techniques, and three scatterplot-based visual encodings. Two authors visually inspected over 800 plots to determine whether or not the measures created plausible results. We found that they failed in over half the cases overall, and in over two-thirds of the cases involving real datasets. Using open and axial coding of failure reasons and separability characteristics, we generated a taxonomy of visual cluster separability factors. We iteratively refined its explanatory clarity and power by mapping the studied datasets and success and failure ranges of the measures onto the factor axes. Our taxonomy has four categories, ordered by their ability to influence successors: Scale, Point Distance, Shape, and Position. Each category is split into Within-Cluster factors such as density, curvature, isotropy, and clumpiness, and Between-Cluster factors that arise from the variance of these properties, culminating in the overarching factor of class separation. The resulting taxonomy can be used to guide the design and the evaluation of cluster separation measures.

Categories and Subject Descriptors (according to ACM CCS): H.5.0 [Information Interfaces and Presentation]: General; J.0 [Computer Applications]: General

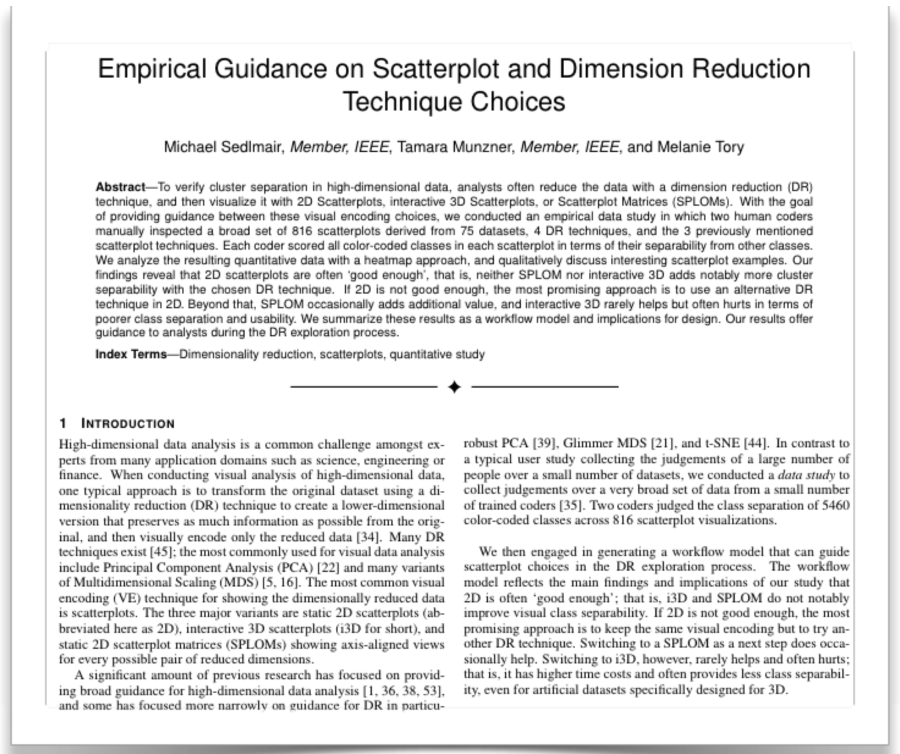
1. Introduction

Over a century of previous work has been devoted to creating effective and efficient algorithms for dimensionality reduction (DR), where a set of points in high-dimensional space is transformed into a more compact lower-dimensional form that preserves the important aspects of its underlying structure. These techniques include the venerable principal components analysis (PCA) [Jol02], the many variants of mul-

choosing DR and VE techniques [IMI*10], but it remains an open problem to develop automatic algorithms to provide such guidance. In service of this goal, we sought to use recent measures for visual cluster separation in scatterplots [SNLH09, TAE*09]. These were originally developed for selecting good views within a SPLOM, but we reasoned that they should also be applicable to providing guidance for DR and VE technique choices. A previous user

EuroVis'12 | InfoVis'13

Sedlmair et al.: A taxonomy of visual cluster separation factors [EuroVis'12] | (today)



Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, *Member, IEEE*, Tamara Munzner, *Member, IEEE*, and Melanie Tory

Abstract—To verify cluster separation in high-dimensional data, analysts often reduce the data with a dimension reduction (DR) technique, and then visualize it with 2D Scatterplots, interactive 3D Scatterplots, or Scatterplot Matrices (SPLOMs). With the goal of providing guidance between these visual encoding choices, we conducted an empirical data study in which two human coders manually inspected a broad set of 816 scatterplots derived from 75 datasets, 4 DR techniques, and the 3 previously mentioned scatterplot techniques. Each coder scored all color-coded classes in each scatterplot in terms of their separability from other classes. We analyze the resulting quantitative data with a heatmap approach, and qualitatively discuss interesting scatterplot examples. Our findings reveal that 2D scatterplots are often 'good enough', that is, neither SPLOM nor interactive 3D adds notably more cluster separability with the chosen DR technique. If 2D is not good enough, the most promising approach is to use an alternative DR technique in 2D. Beyond that, SPLOM occasionally adds additional value, and interactive 3D rarely helps but often hurts in terms of poorer class separation and usability. We summarize these results as a workflow model and implications for design. Our results offer guidance to analysts during the DR exploration process.

Index Terms—Dimensionality reduction, scatterplots, quantitative study

1 INTRODUCTION

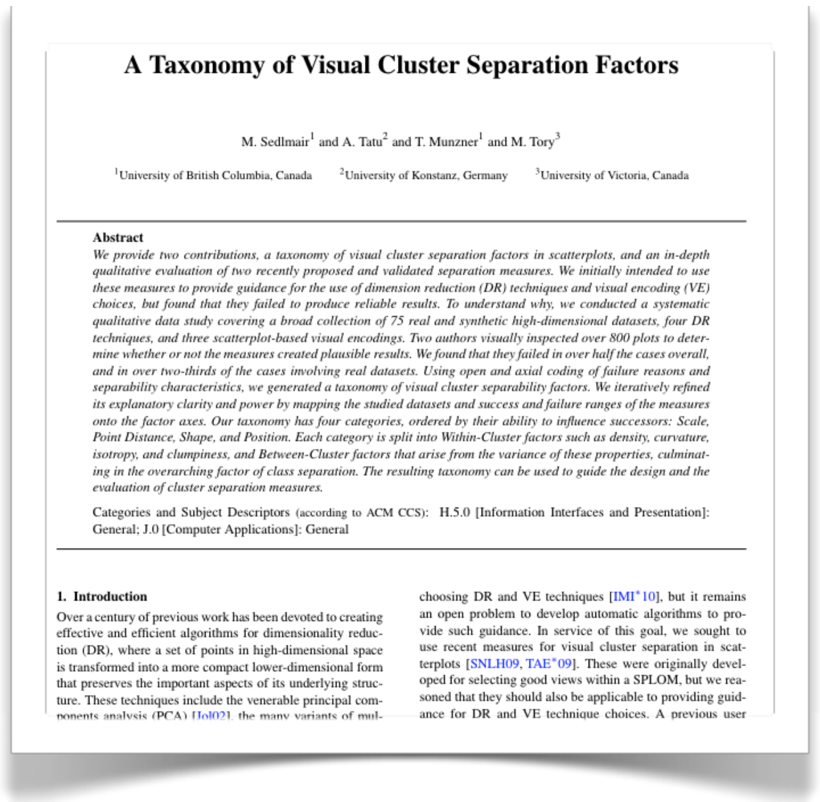
High-dimensional data analysis is a common challenge amongst experts from many application domains such as science, engineering or finance. When conducting visual analysis of high-dimensional data, one typical approach is to transform the original dataset using a dimensionality reduction (DR) technique to create a lower-dimensional version that preserves as much information as possible from the original, and then visually encode only the reduced data [34]. Many DR techniques exist [45]; the most commonly used for visual data analysis include Principal Component Analysis (PCA) [22] and many variants of Multidimensional Scaling (MDS) [5, 16]. The most common visual encoding (VE) technique for showing the dimensionally reduced data is scatterplots. The three major variants are static 2D scatterplots (abbreviated here as 2D), interactive 3D scatterplots (i3D for short), and static 2D scatterplot matrices (SPLOMs) showing axis-aligned views for every possible pair of reduced dimensions.

A significant amount of previous research has focused on providing broad guidance for high-dimensional data analysis [1, 36, 38, 53], and some has focused more narrowly on guidance for DR in particu-

robust PCA [39], Glimmer MDS [21], and t-SNE [44]. In contrast to a typical user study collecting the judgements of a large number of people over a small number of datasets, we conducted a *data study* to collect judgements over a very broad set of data from a small number of trained coders [35]. Two coders judged the class separation of 5460 color-coded classes across 816 scatterplot visualizations.

We then engaged in generating a workflow model that can guide scatterplot choices in the DR exploration process. The workflow model reflects the main findings and implications of our study that 2D is often 'good enough'; that is, i3D and SPLOM do not notably improve visual class separability. If 2D is not good enough, the most promising approach is to keep the same visual encoding but to try another DR technique. Switching to a SPLOM as a next step does occasionally help. Switching to i3D, however, rarely helps and often hurts; that is, it has higher time costs and often provides less class separability, even for artificial datasets specifically designed for 3D.

2 part project


A Taxonomy of Visual Cluster Separation Factors

M. Sedlmair[1] and A. Tatu[2] and T. Munzner[1] and M. Tory[3]

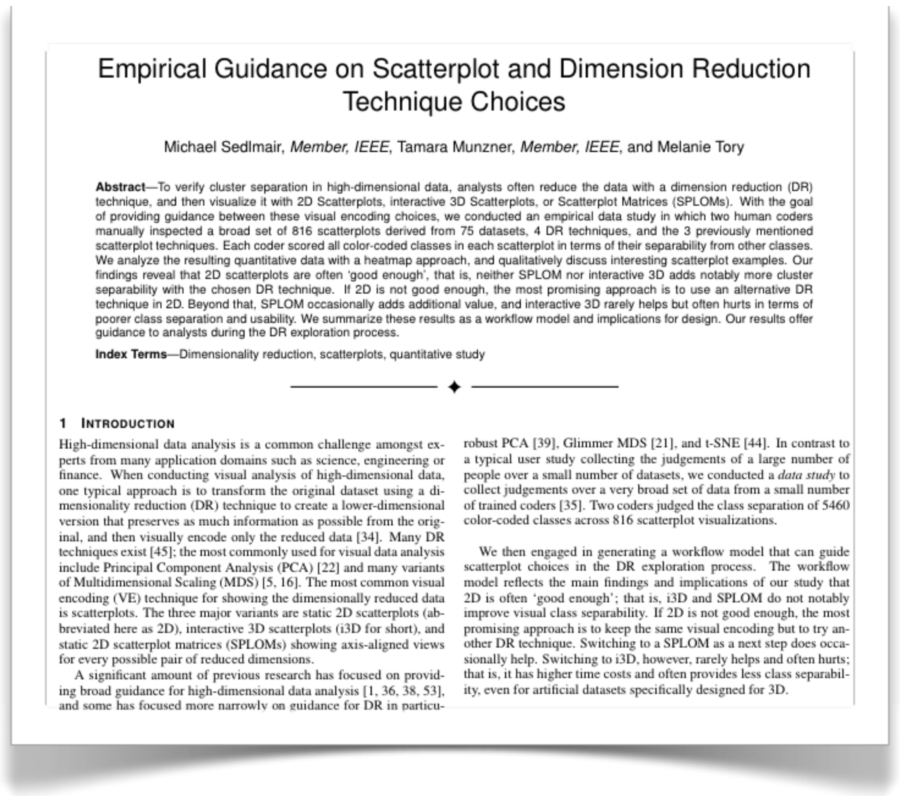
[1]University of British Columbia, Canada [2]University of Konstanz, Germany [3]University of Victoria, Canada

Abstract

1. Introduction

EuroVis'12 | InfoVis'13

Sedlmair et al.: A taxonomy of visual cluster separation factors [EuroVis'12] | (today)

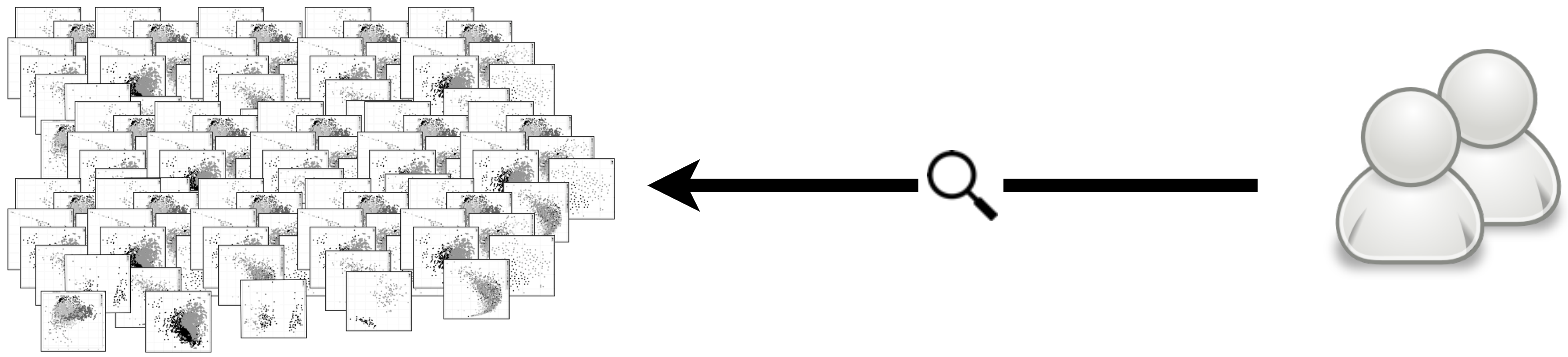

Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

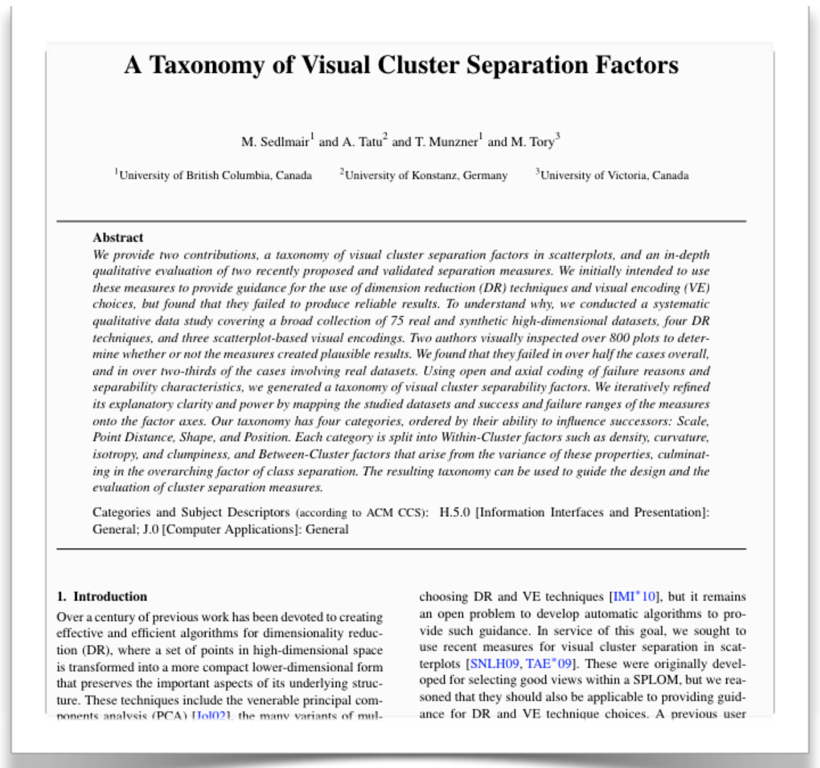
Michael Sedlmair, *Member, IEEE*, Tamara Munzner, *Member, IEEE*, and Melanie Tory

1 Introduction

Same method/base data:

data study with same 816 scatterplots

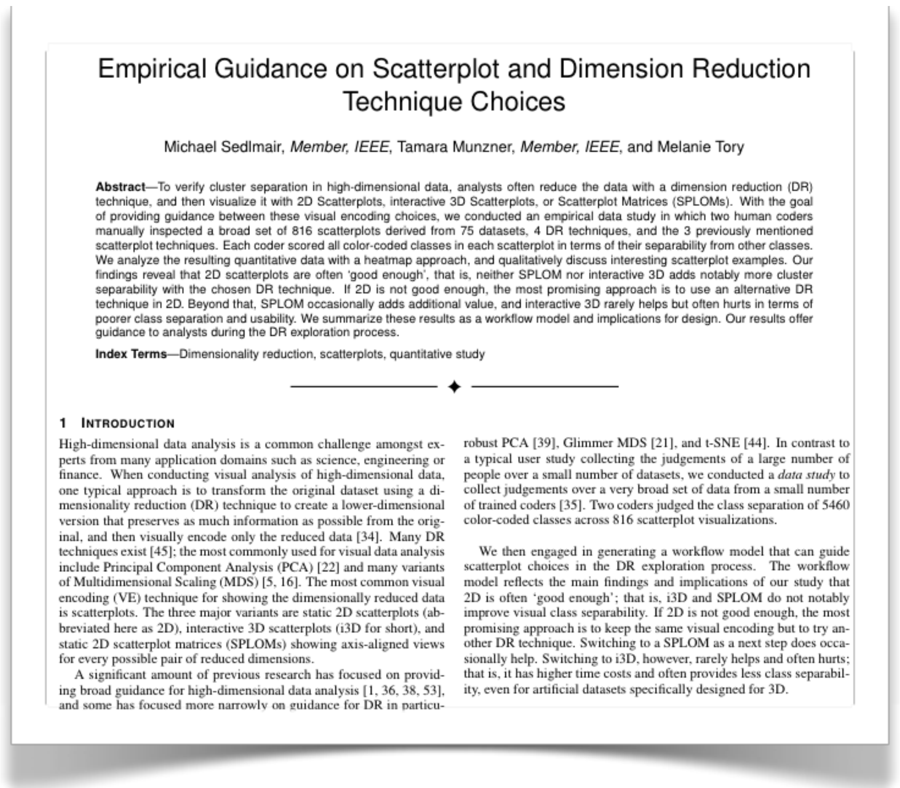



A Taxonomy of Visual Cluster Separation Factors

M. Sedlmair[1] and A. Tatu[2] and T. Munzner[1] and M. Tory[3]

EuroVis'12 | InfoVis'13

Sedlmair et al.: A taxonomy of visual cluster separation factors [EuroVis'12] | (today)


Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, Member, IEEE, Tamara Munzner, Member, IEEE, and Melanie Tory

Same method/base data:

data study with same 816 scatterplots

Different data gathering/analysis:

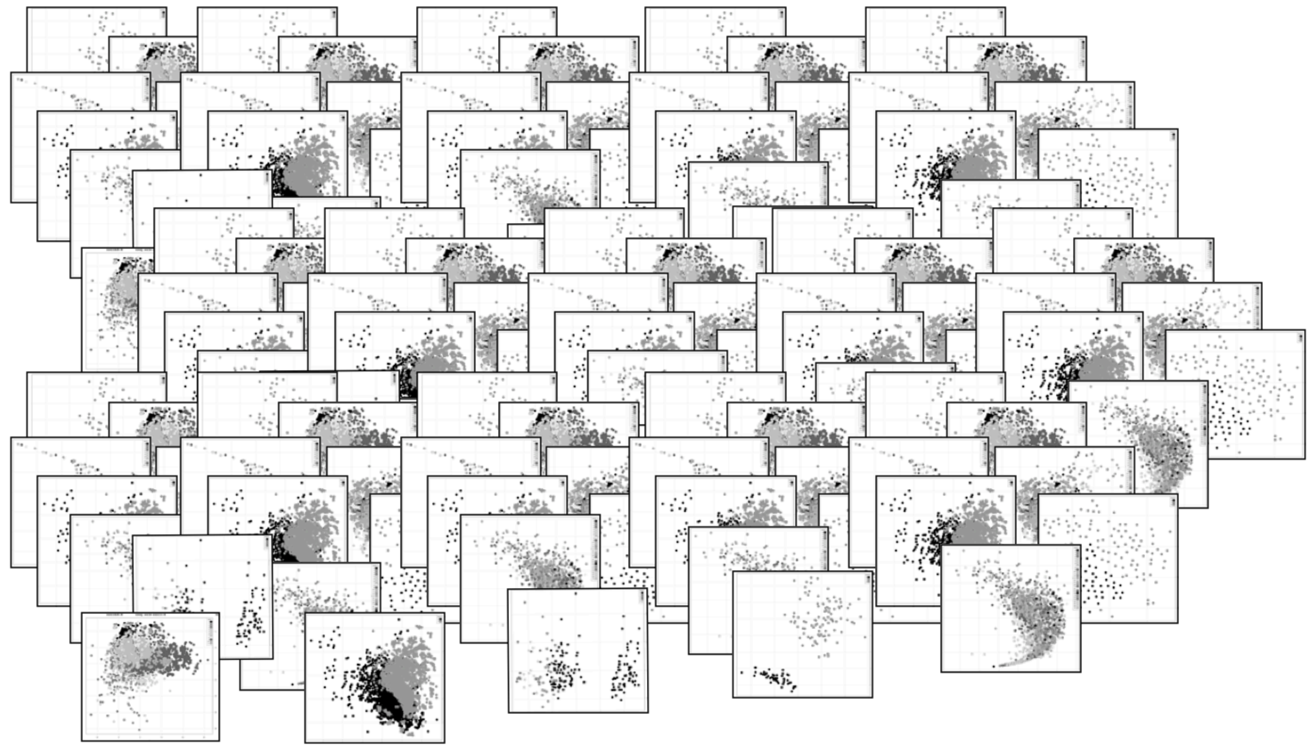
EuroVis'12	InfoVis'13
qualitative coding	quantitative data

Different goals/contributions:

EuroVis'12	InfoVis'13
taxonomy of visual cluster separation factors evaluation of automatic class separation measures	Comparing visual encoding choices: 2D, 3D, and SPLOM

Method

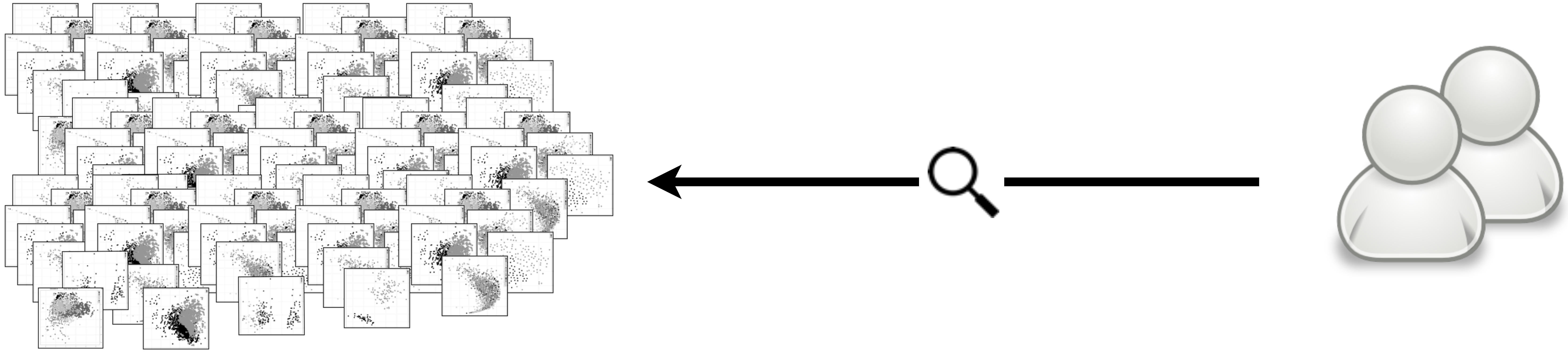
Data Study



Many
Scatterplots

2 human
expert coders

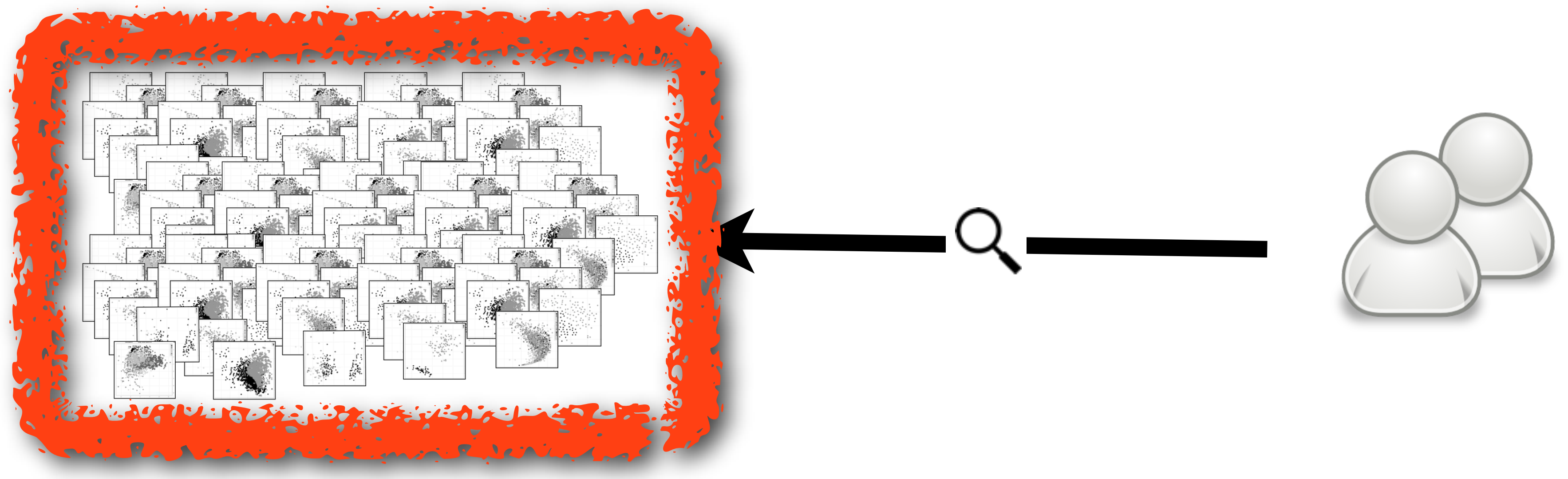
Data Study



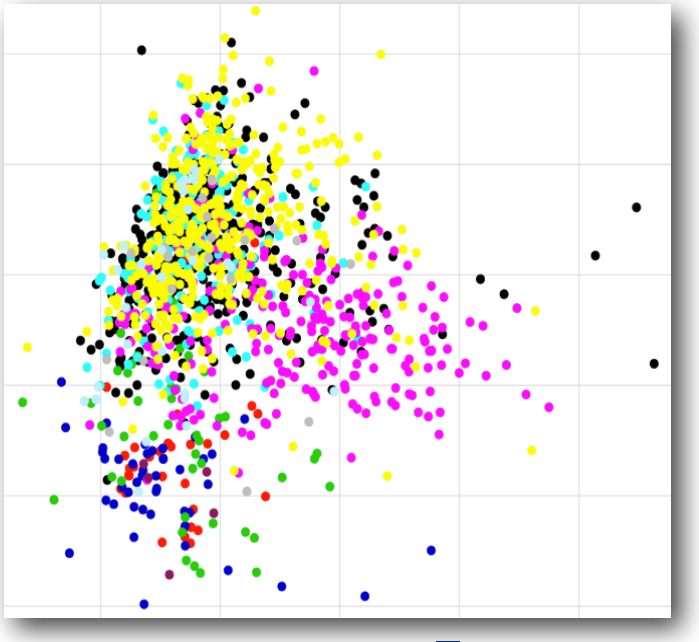
Reasons:

- data characteristics outweigh user differences
- need for reliable cluster separation judgement

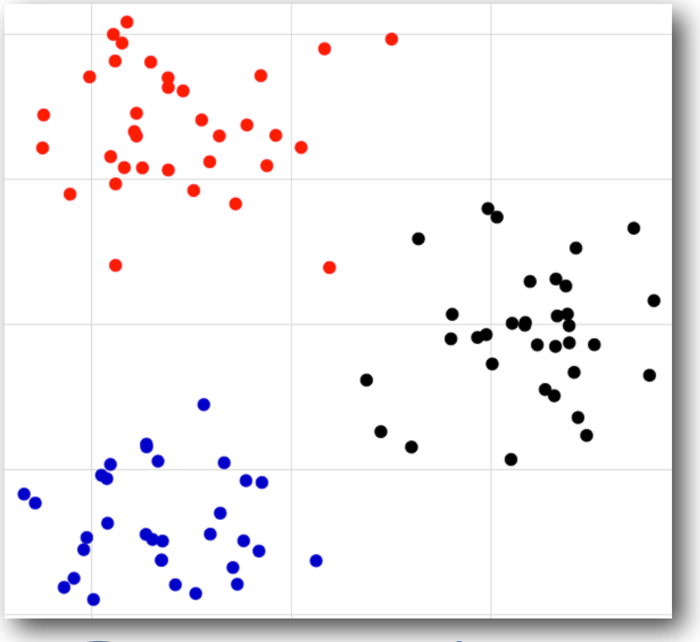
  Sedlmair et al.: A taxonomy of visual cluster separation factors [EuroVis'12]



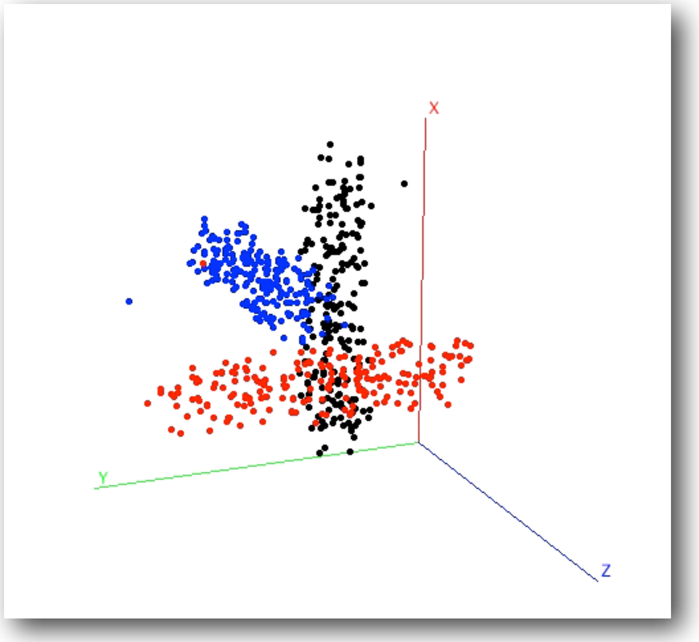
75 pre-classified datasets



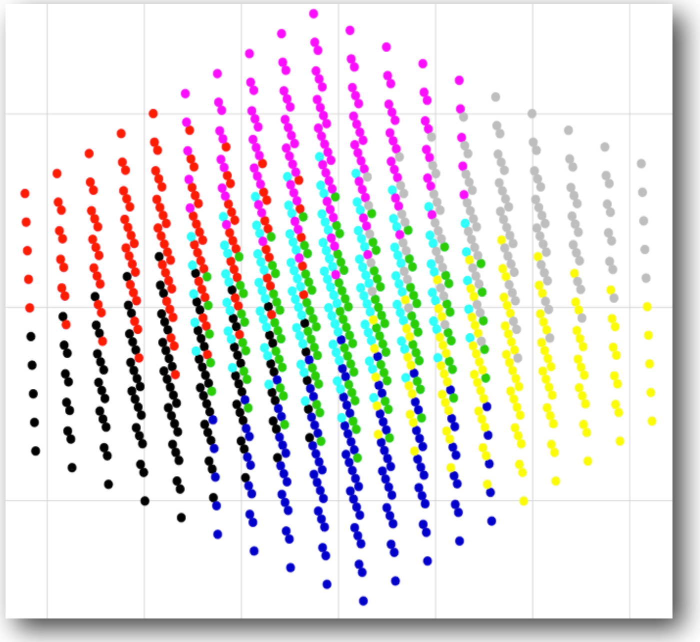
real
(31)



Gaussian
(16)

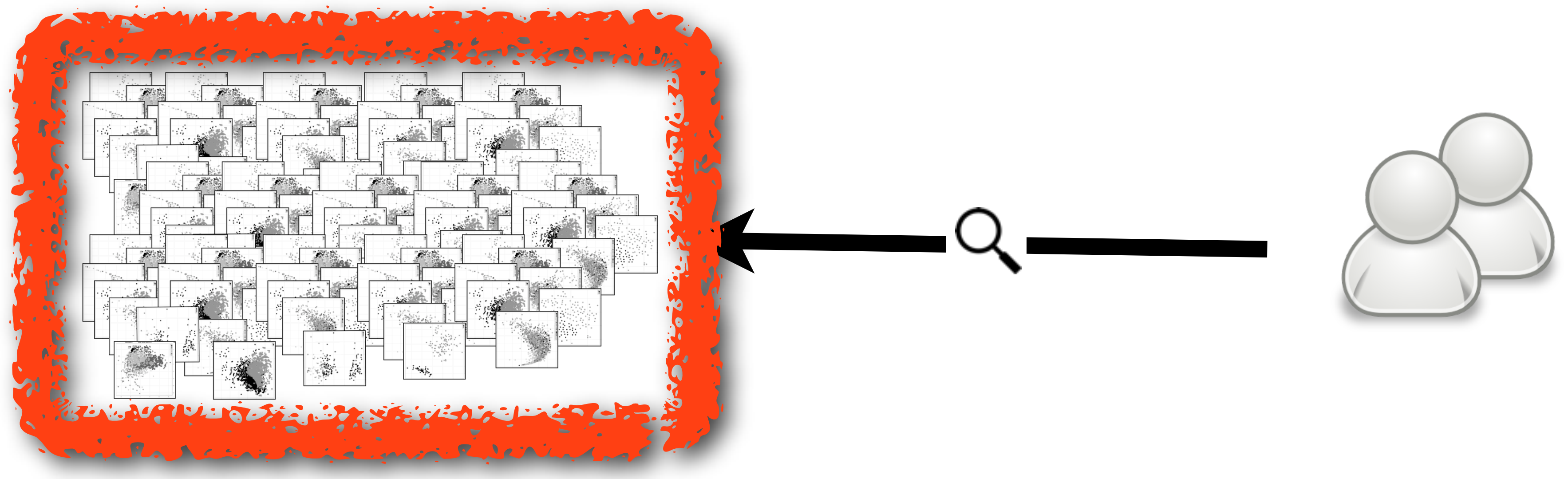


entangled
(24)



grid
(4)

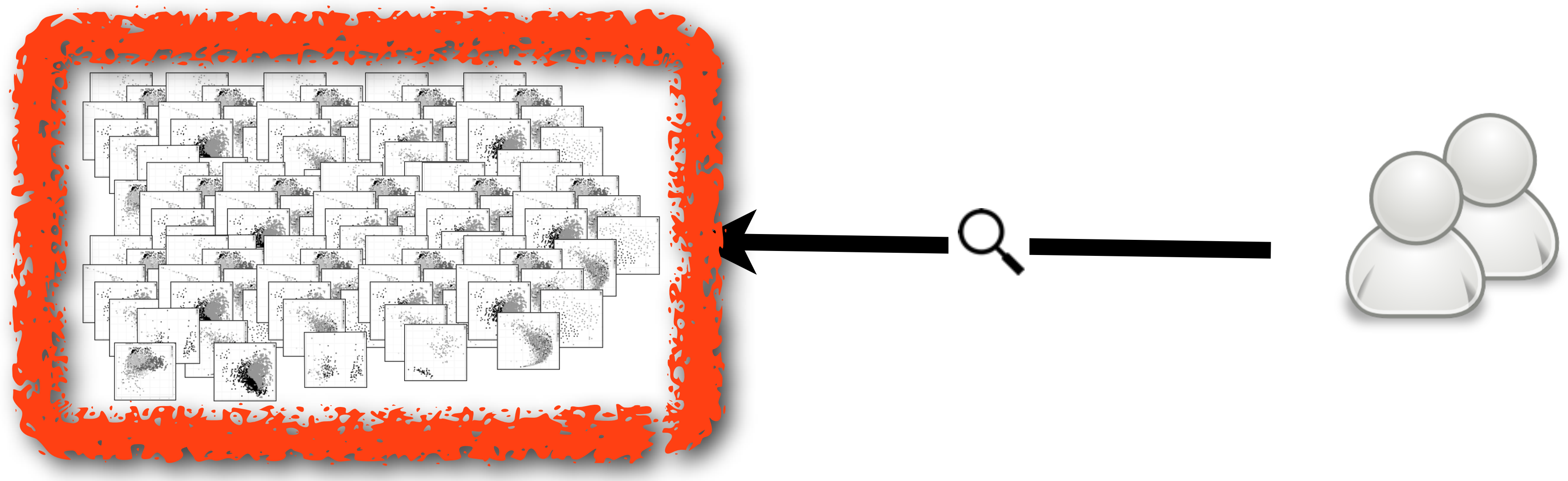
├──────── synthetic ────────┤



75 pre-classified datasets

4 DR techniques

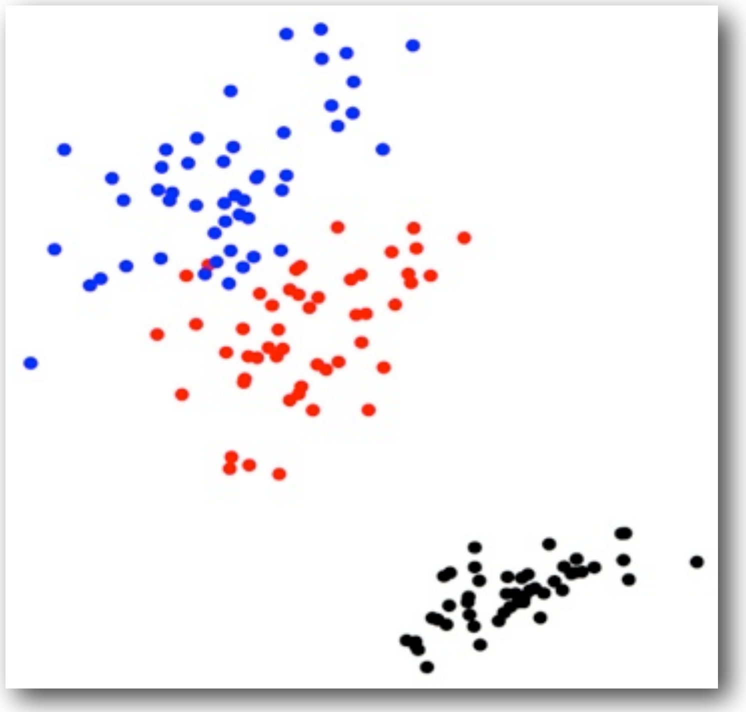
- PCA (linear)
- Robust PCA (linear)
- Glimmer MDS (non-linear)
- t-SNE (non-linear)

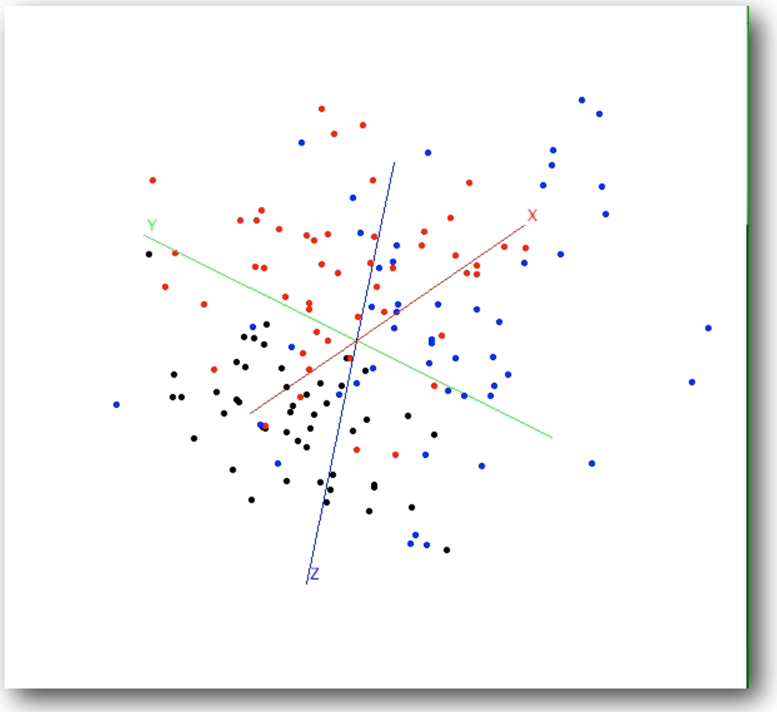


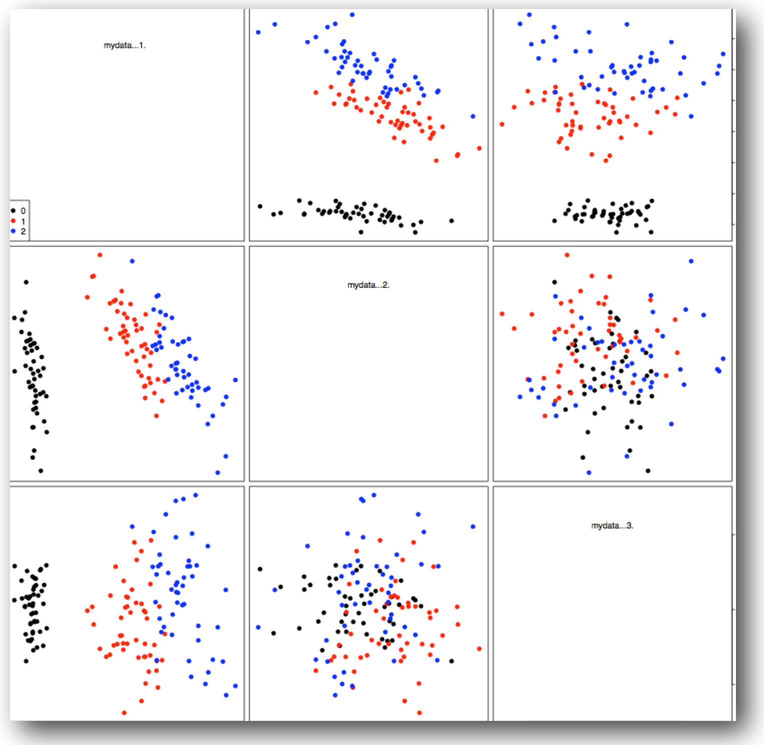
75 pre-classified datasets

4 DR techniques

3 visual encodings

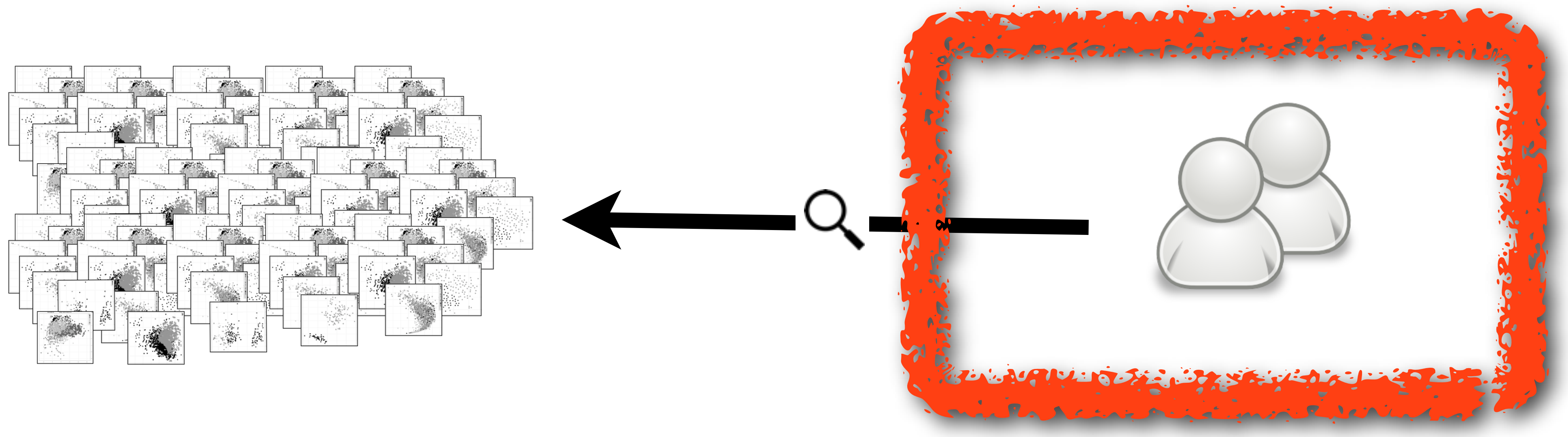






SPLOM:
3 - 7 dim.

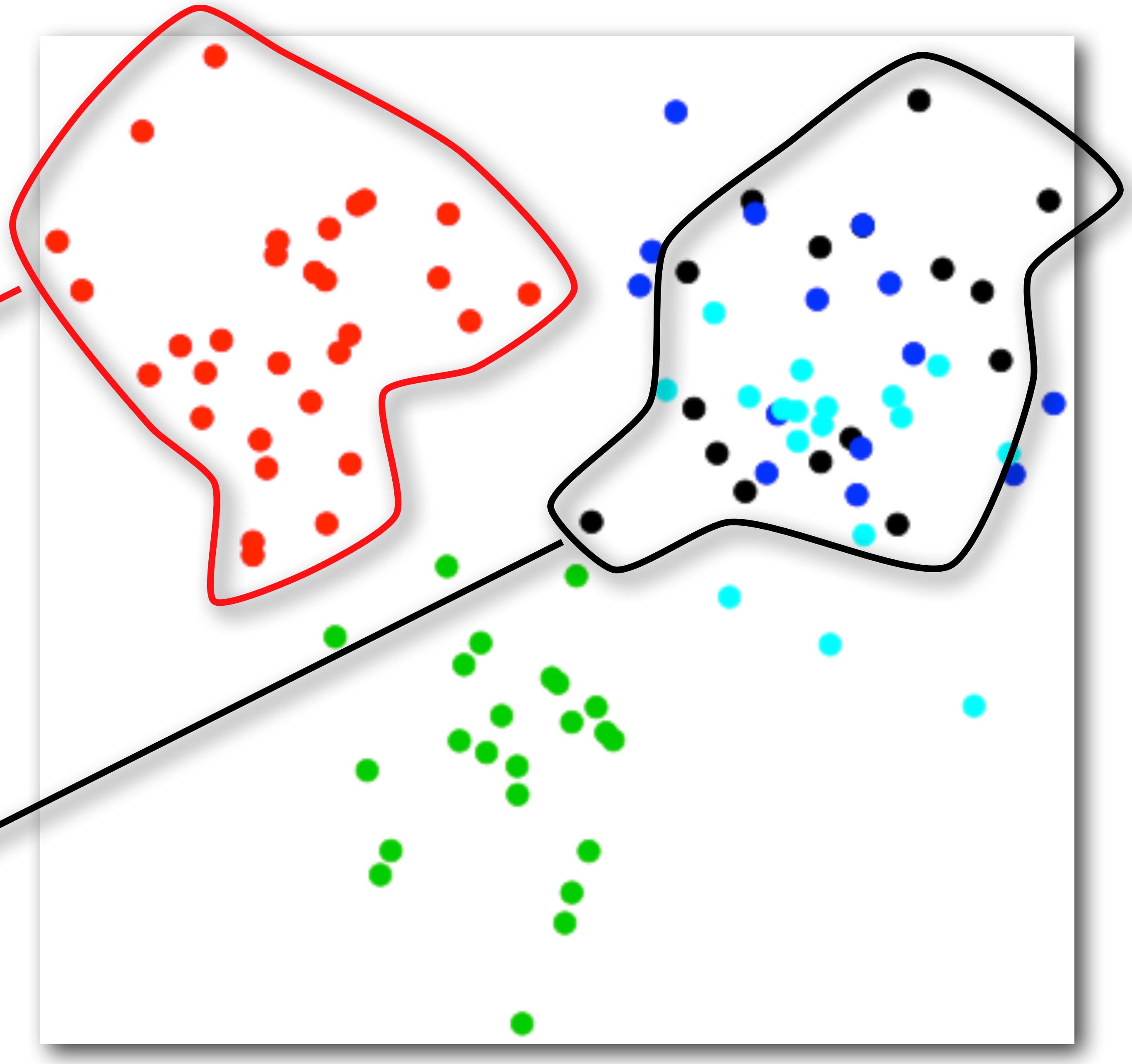
→ 816 Plots

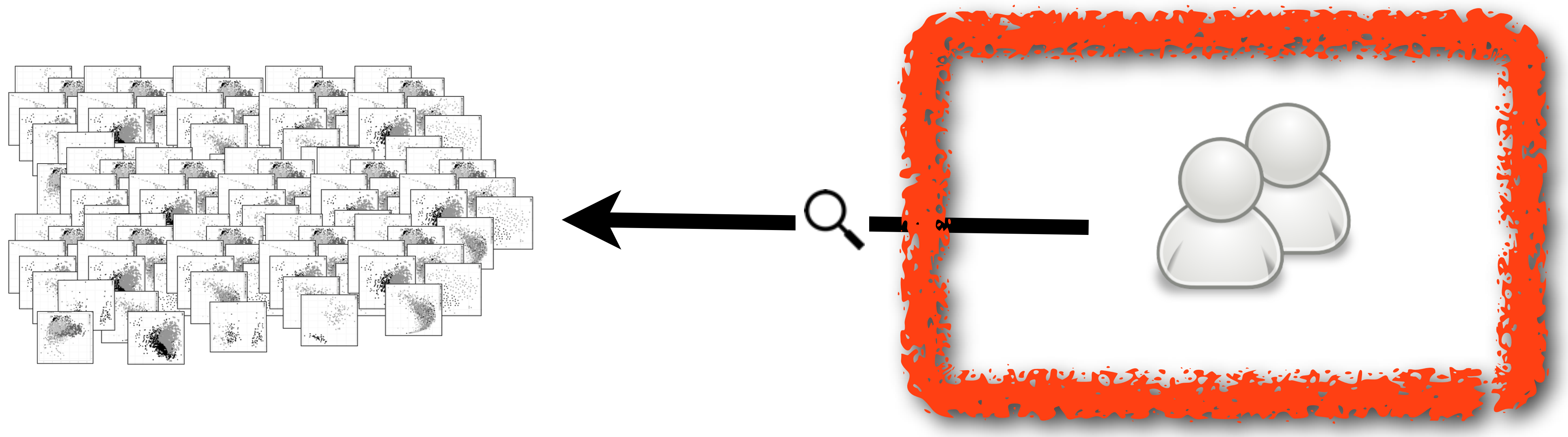


2 human expert coders

- inspect all 816 Plots
- judge all clusters:

5 = *nicely separated*
4 ...
3 ...
2 ...
1 = *not separated*





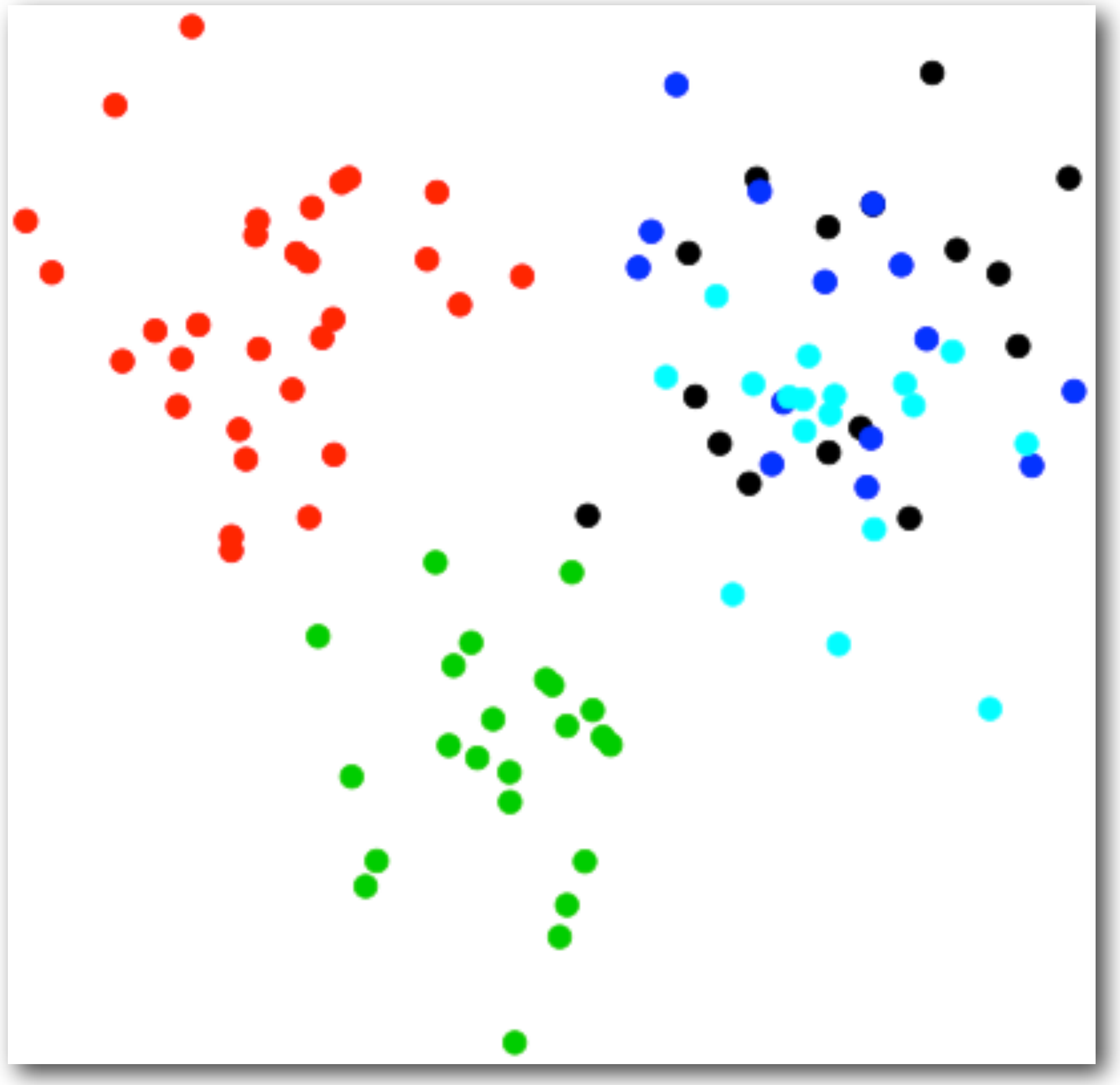
5460

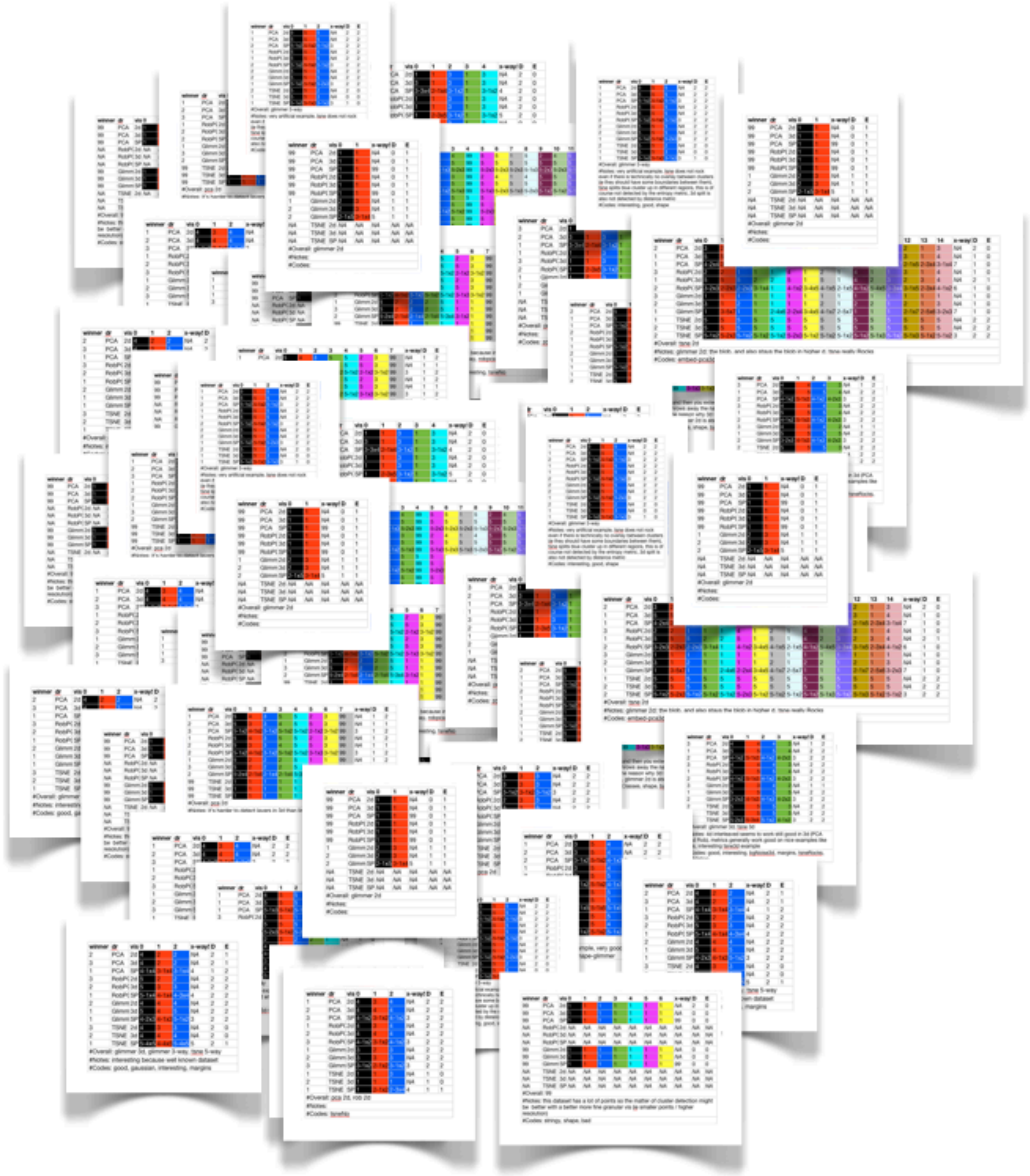
Class judgments / coder
~80 hours coding / coder

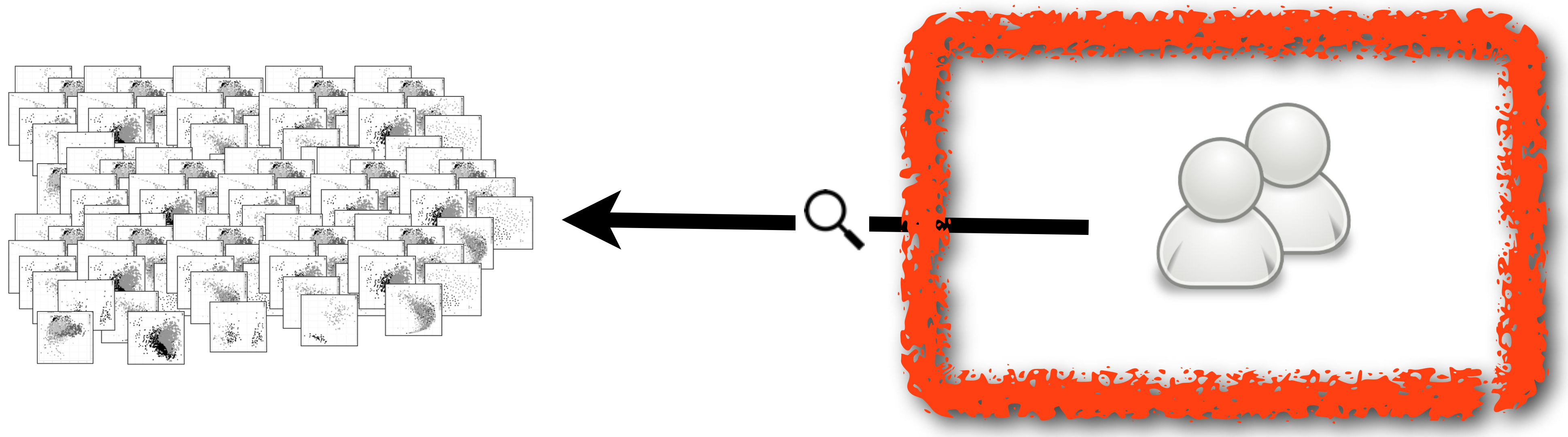
2 human expert coders

- inspect all 816 Plots
- judge all clusters:

5 = *nicely separated*
4 …
3 …
2 …
1 = *not separated*







Judging Reliability

- high inter-coder reliability (*Krippendorff's alpha = 0.86*)
- echoing previous findings

  Lewis et al.: Human cluster evaluation and formal quality measures: a comparative study [CogSci'12]

Data Analysis
&
Results

Cost Assumption

2D < SPLOM < 3D

- Based on rich body of previous work*

*** previous work:**

Drawbacks of 3D

Chalmers: Using a landscape metaphor to represent a corpus of documents [COSIT'93]

Cockburn and McKenzie: An evaluation of cone trees [British Conf. on HCI'00]

Cockburn and McKenzie: Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments [CHI'02]

Newby: Empirical study of a 3D visualization for information retrieval tasks. J. Intelligent Information Systems, 18(1):31–53, 2002.

Tory et al.: Spatialization design: comparing points and landscapes [InfoVis'07]

Tory et al.: Comparing dot and landscape spatializations for visual memory differences [InfoVis'09]

Westerman and Cribbin: Mapping semantic information in virtual space: dimensions, variance and individual differences [IJHCS'00]

Interaction Costs

Lam: A framework of interaction costs in information visualization [InfoVis'08]

Van Wijk: Views on visualization [TVCG'06]

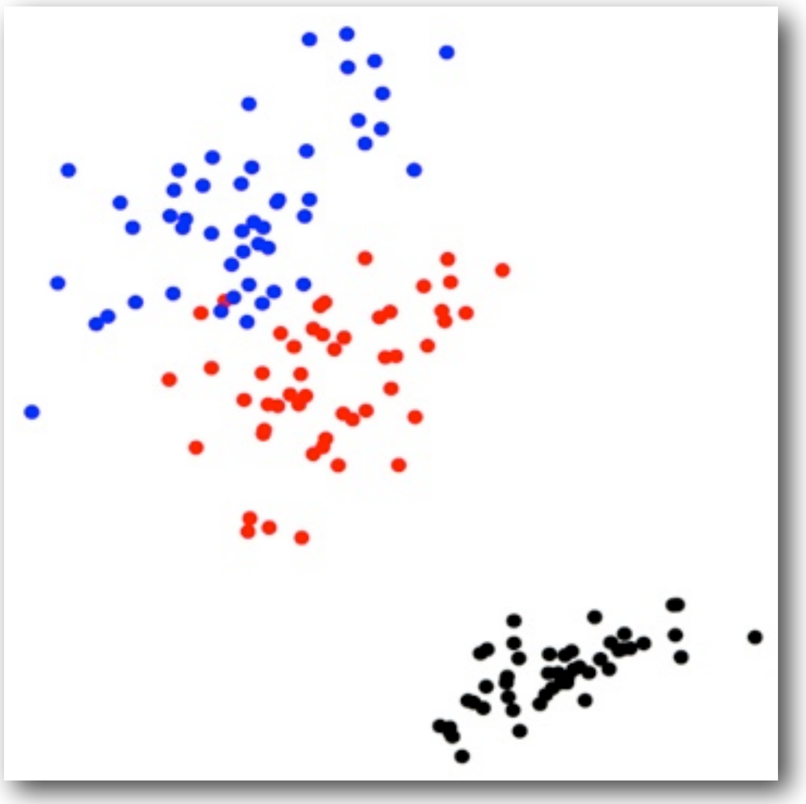
Cost Assumption

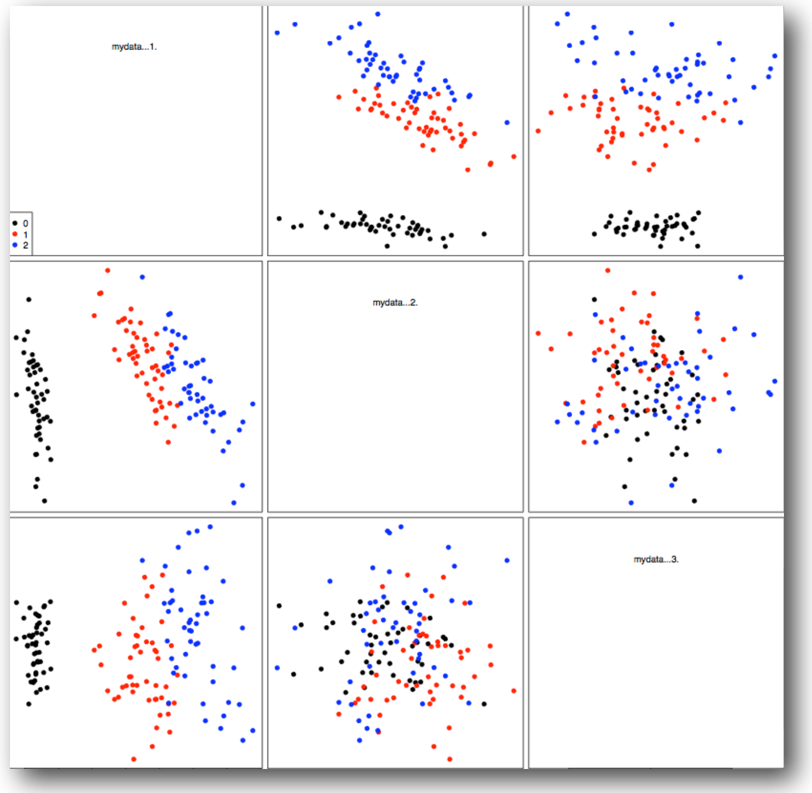
2D < SPLOM < 3D

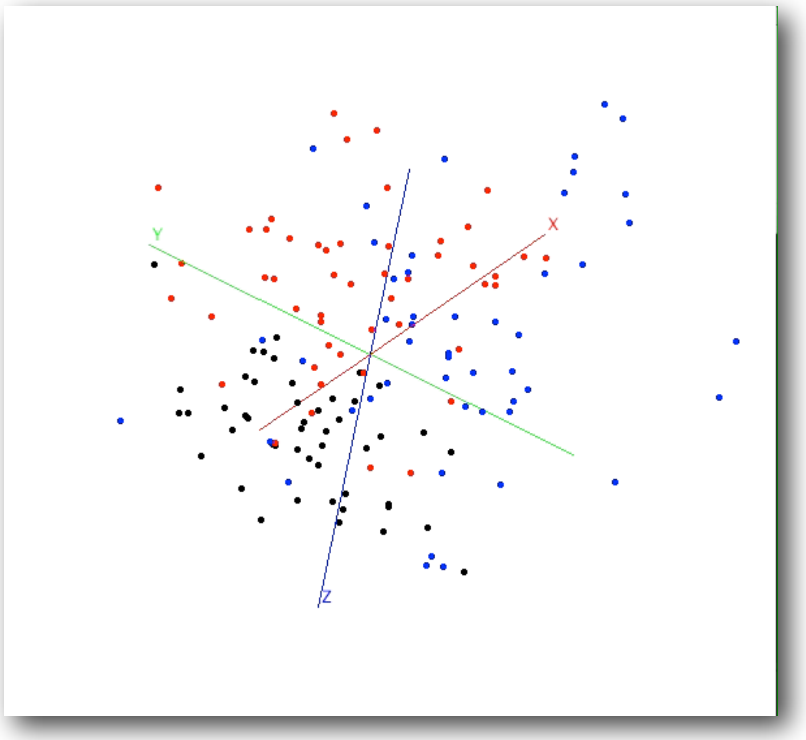
- Based on rich body of previous work

Reasons:

- 2D (low): static, directly visible
- SPLOM (medium): switching attention between views
- 3D (high): interaction to resolve occlusions







Cost Assumption

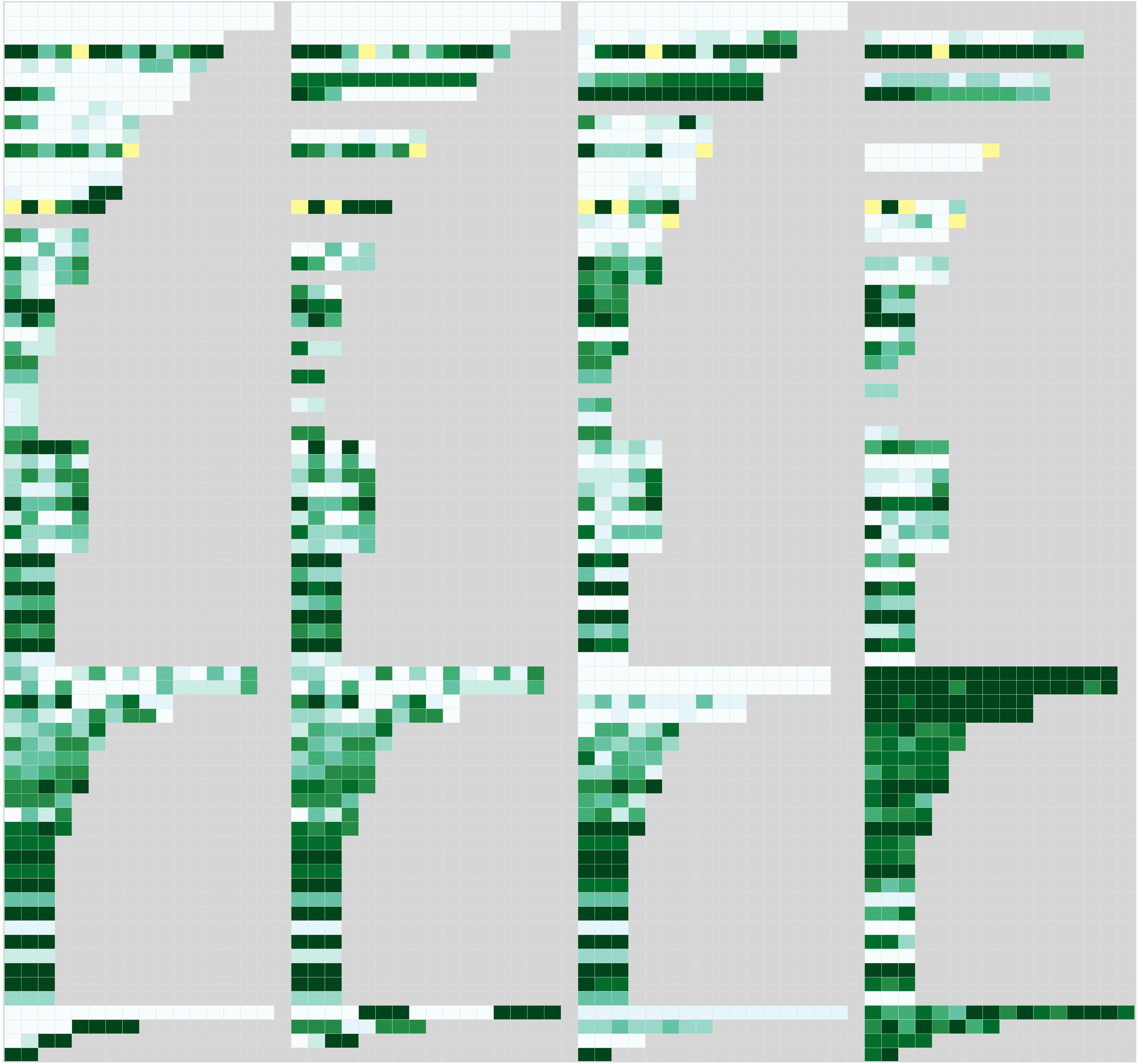
- Use a higher cost visual encoding **only** if it provides notably better class separation

- Use 2D if "good enough", if not then SPLOM, then 3D

Data Analysis

1. Heatmaps Approach
   - reveals a lot of the details

2. Statistical Analysis
   - confirms heatmap analysis
   - **see paper**



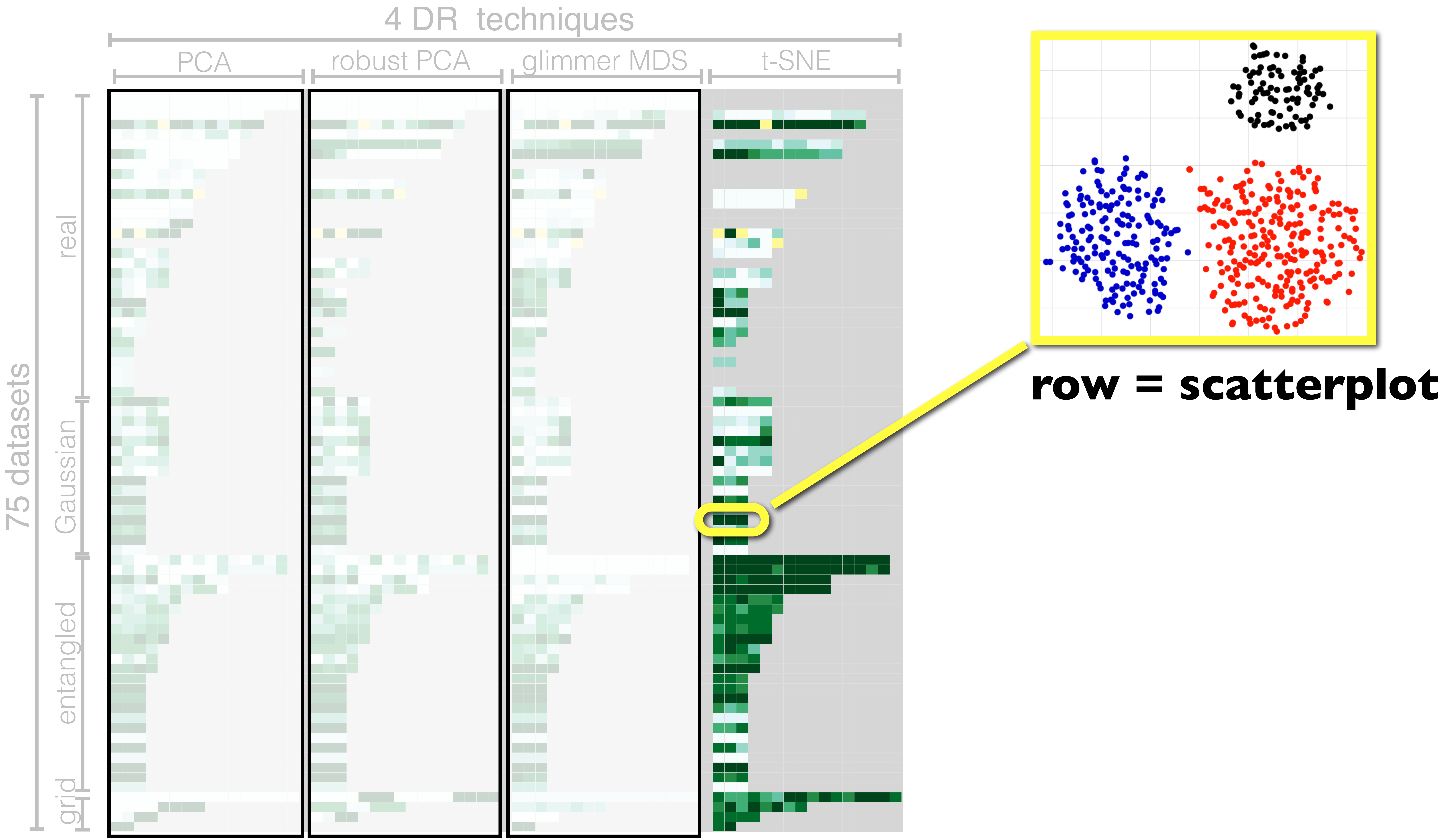
Base heatmaps

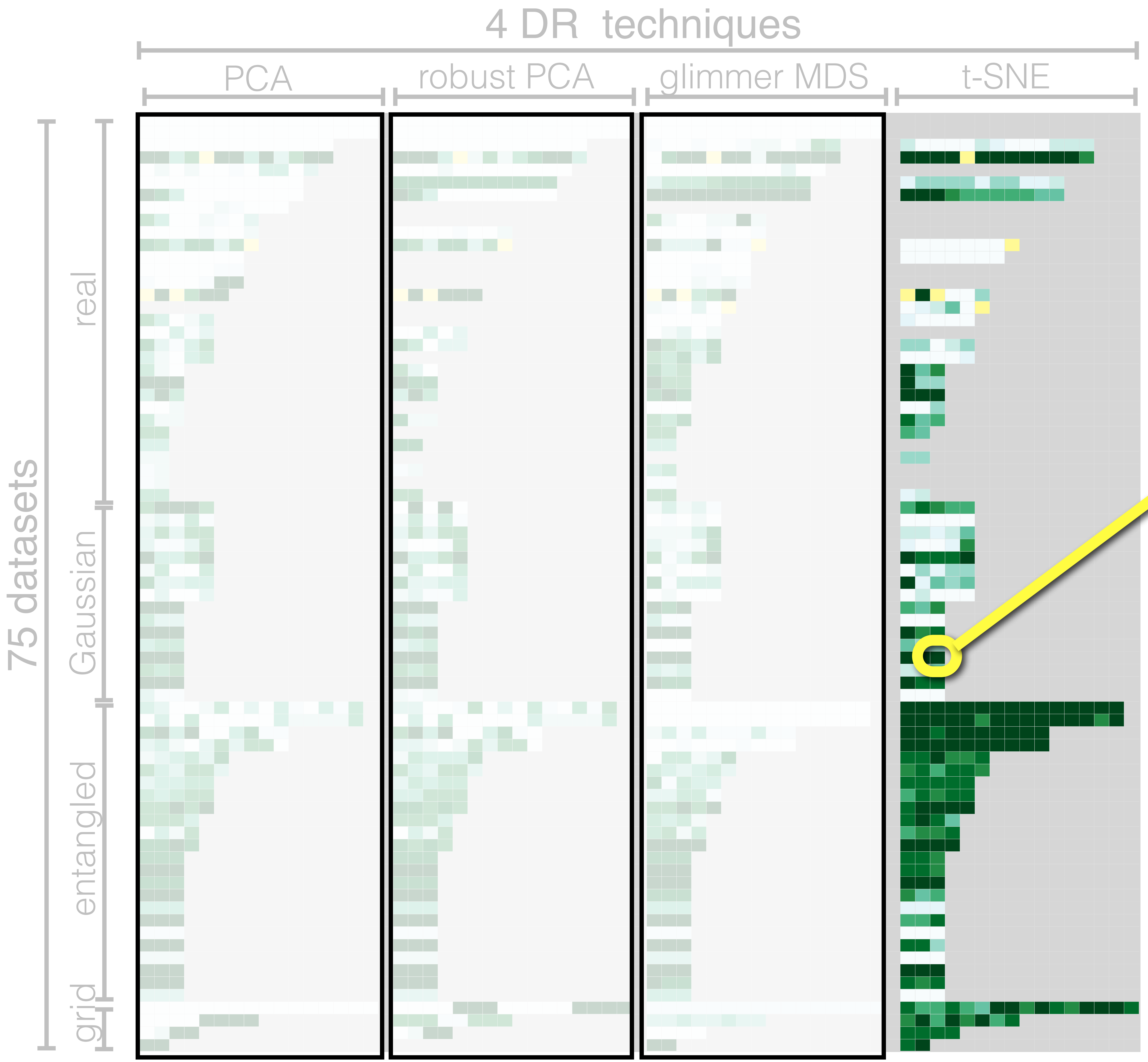
Showing averaged scores of two coders

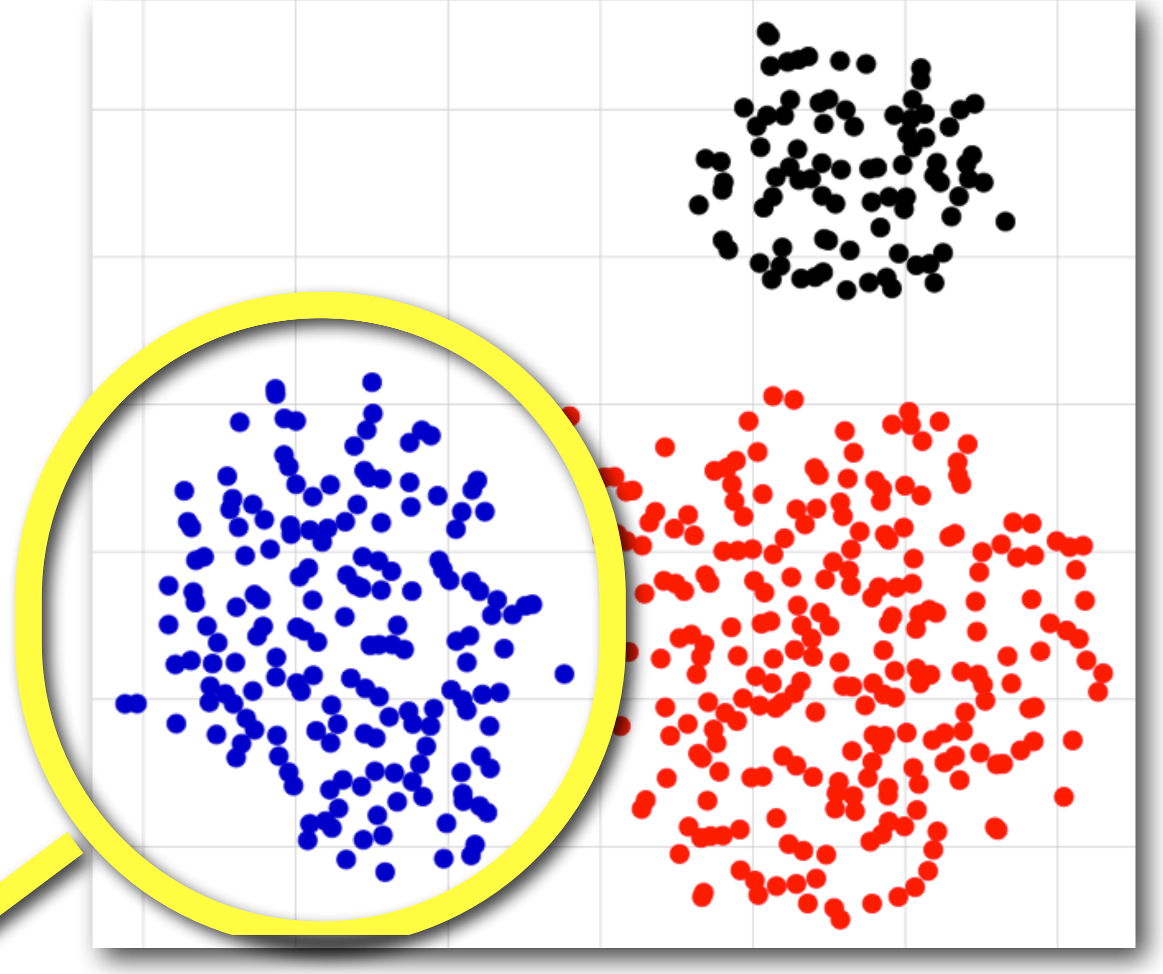
real

highly synthetic


4 DR techniques
PCA
robust PCA
glimmer MDS
t-SNE
75 datasets
real
Gaussian
entangled
grid

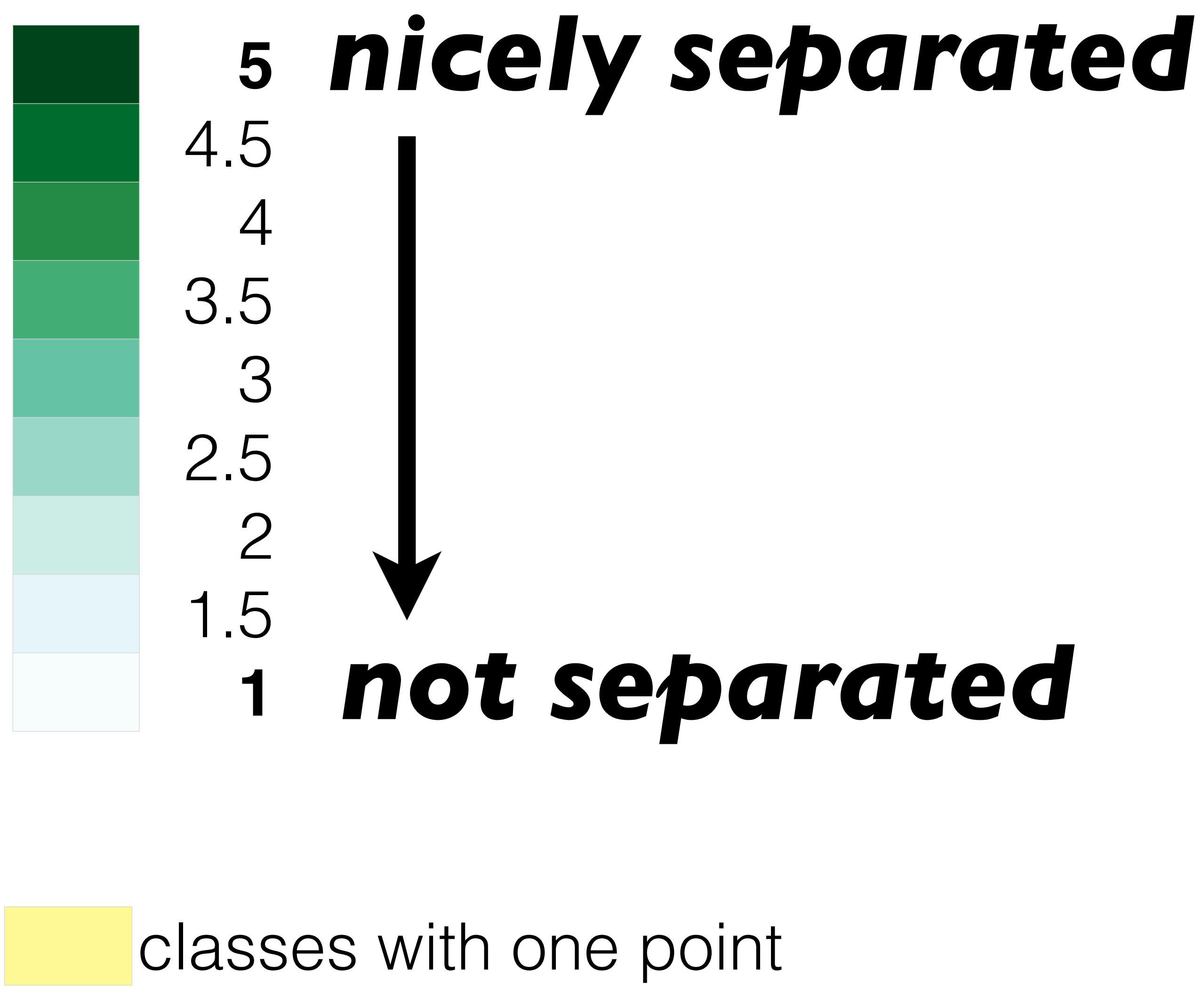

4 DR techniques
PCA
robust PCA
glimmer MDS
t-SNE
75 datasets
real
Gaussian
entangled
grid
row = scatterplot

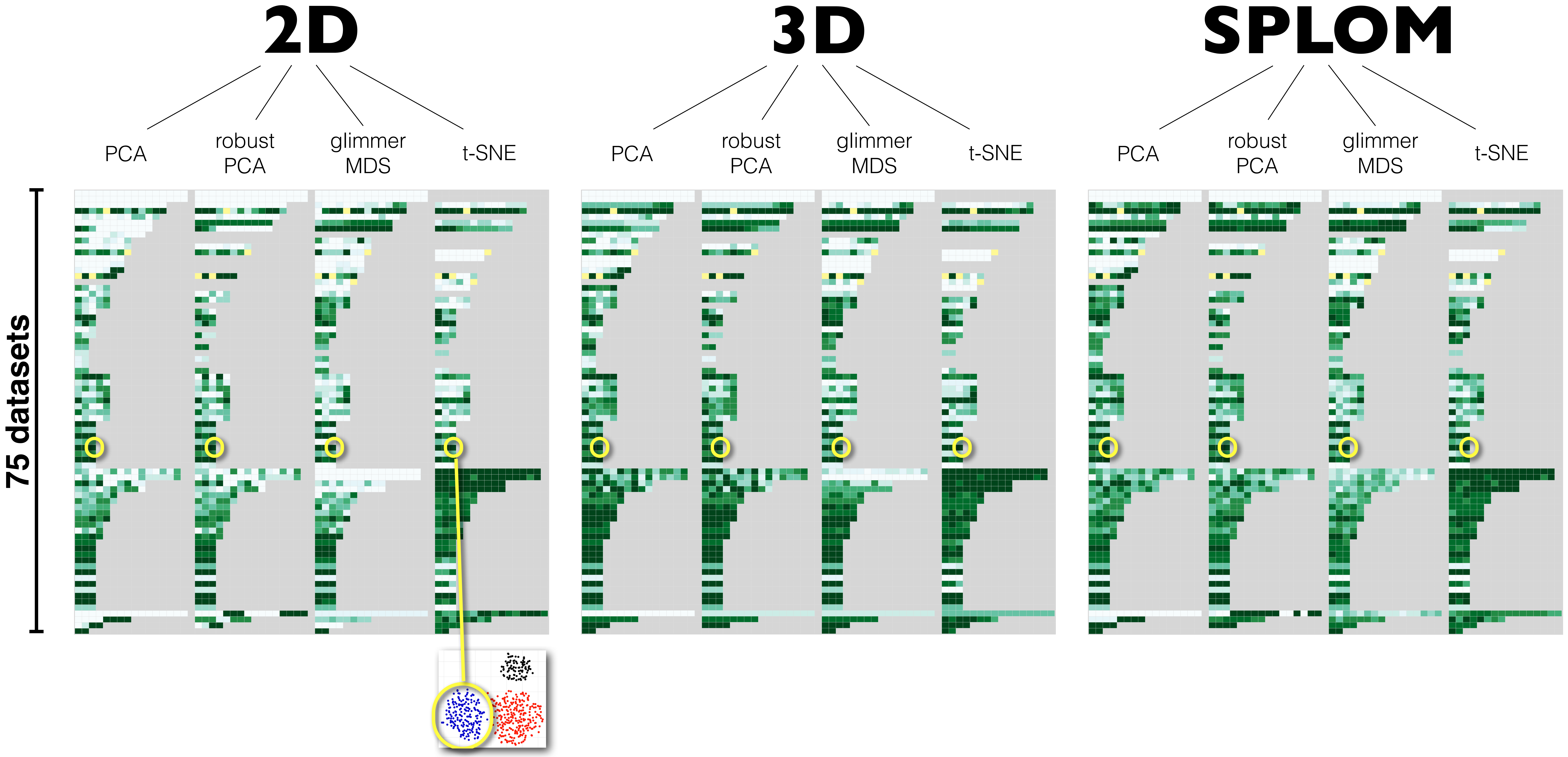




cell = class

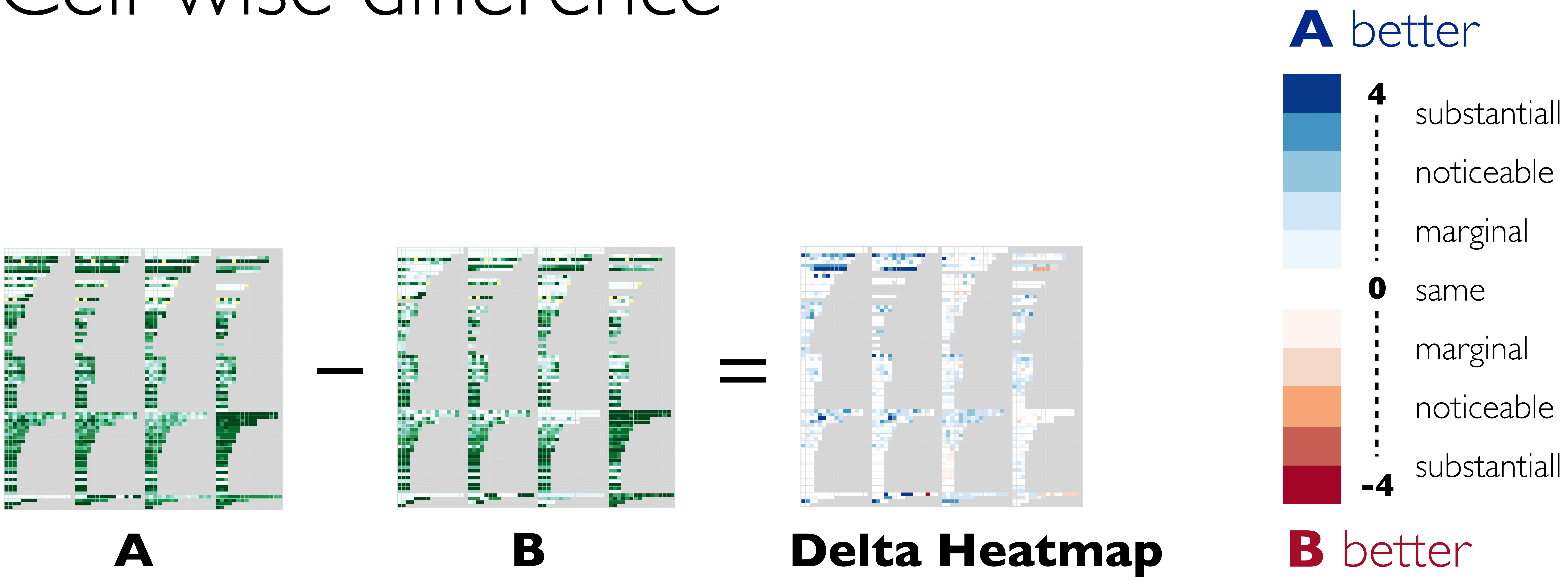
averaged scores





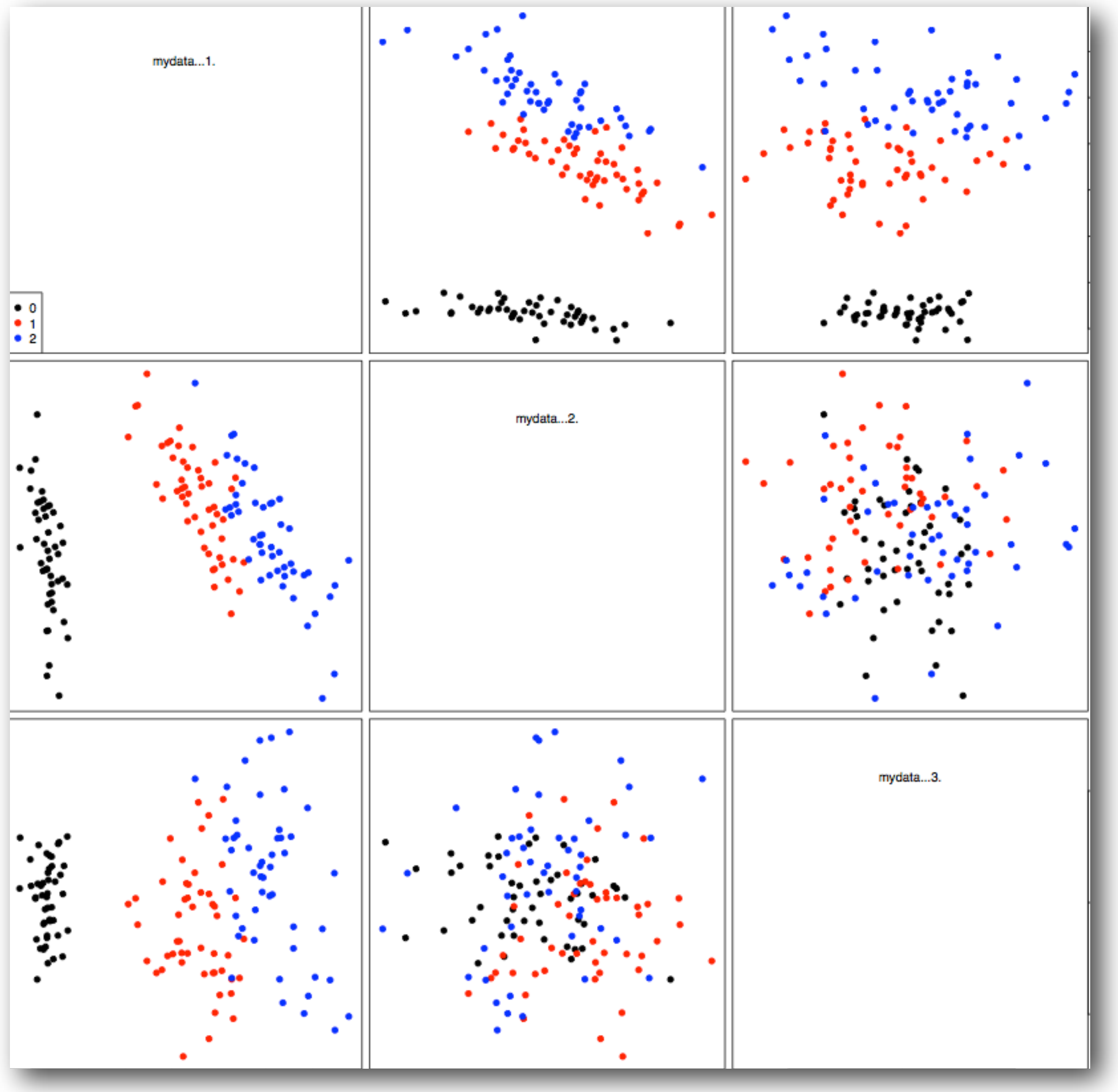
2D
3D
SPLOM
PCA
robust PCA
glimmer MDS
t-SNE
PCA
robust PCA
glimmer MDS
t-SNE
PCA
robust PCA
glimmer MDS
t-SNE
75 datasets

Delta Heatmaps:
Cell-wise difference



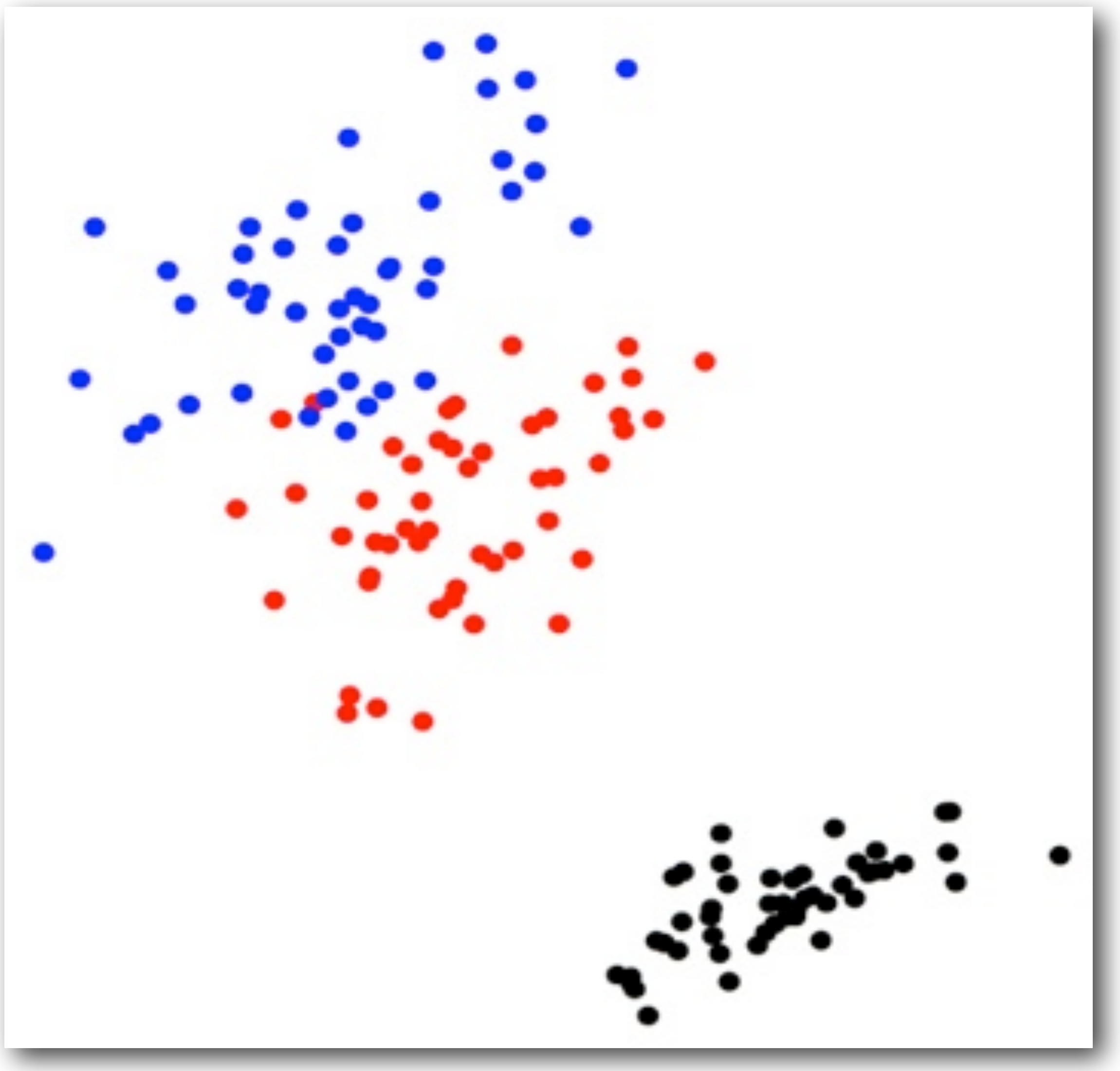
Within-DR

SPLOM vs. **2D**



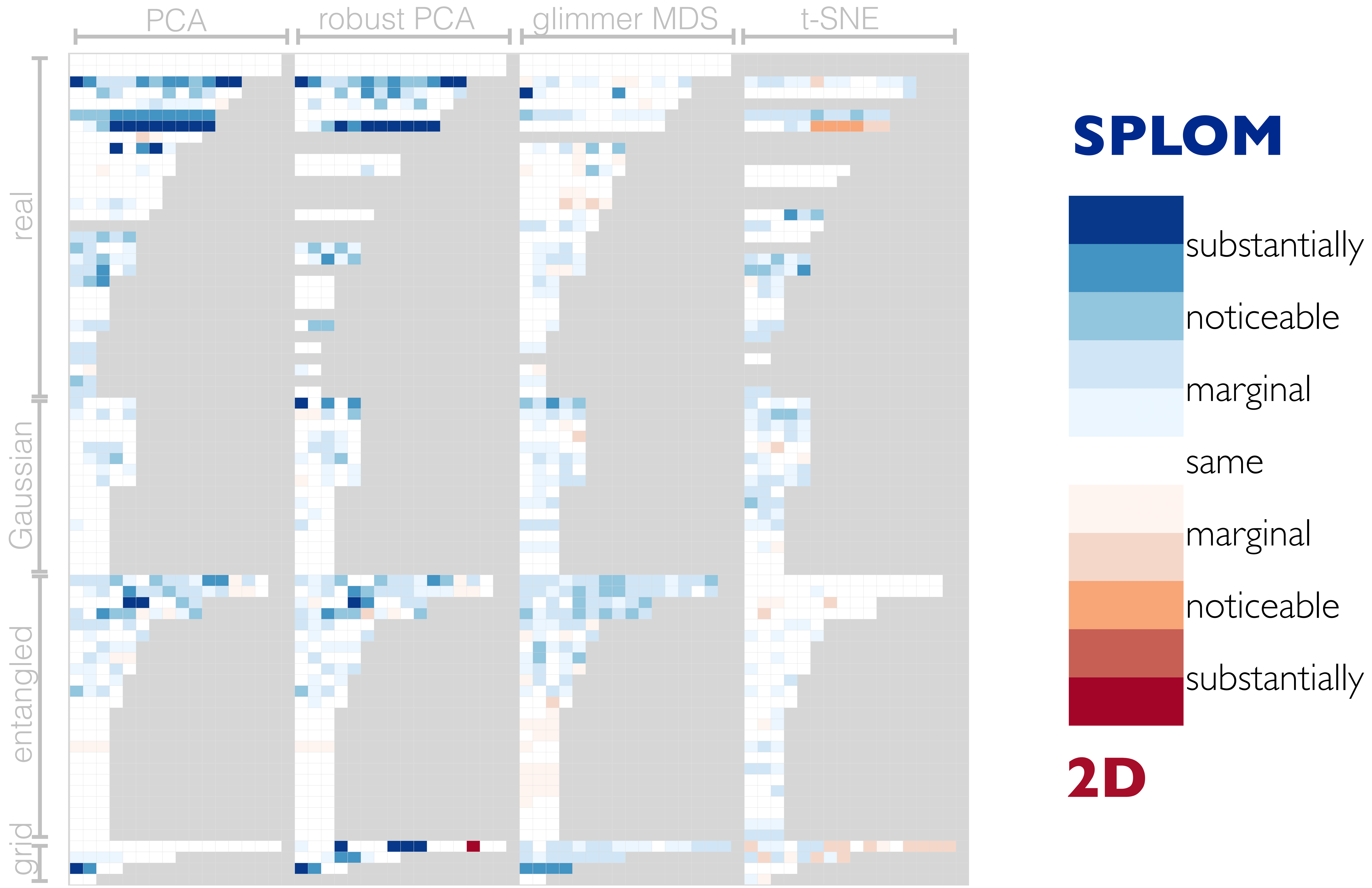
SPLOM_{PCA}

which is better?



2D_{PCA}

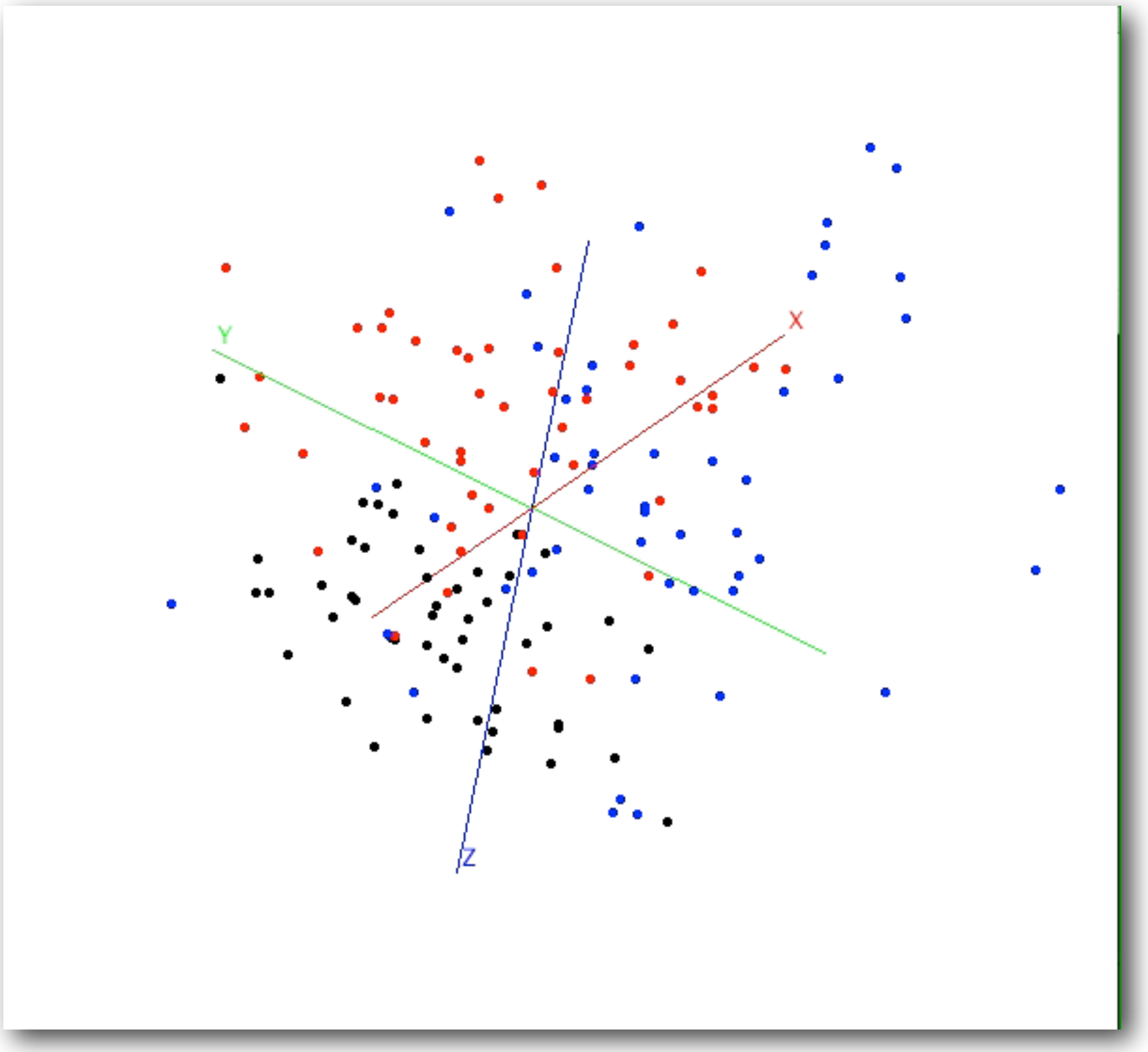
SPLOM vs. 2D


PCA
robust PCA
glimmer MDS
t-SNE
real
Gaussian
entangled
grid
SPLOM
substantially
noticeable
marginal
same
marginal
noticeable
substantially
2D

SPLOM vs. 2D

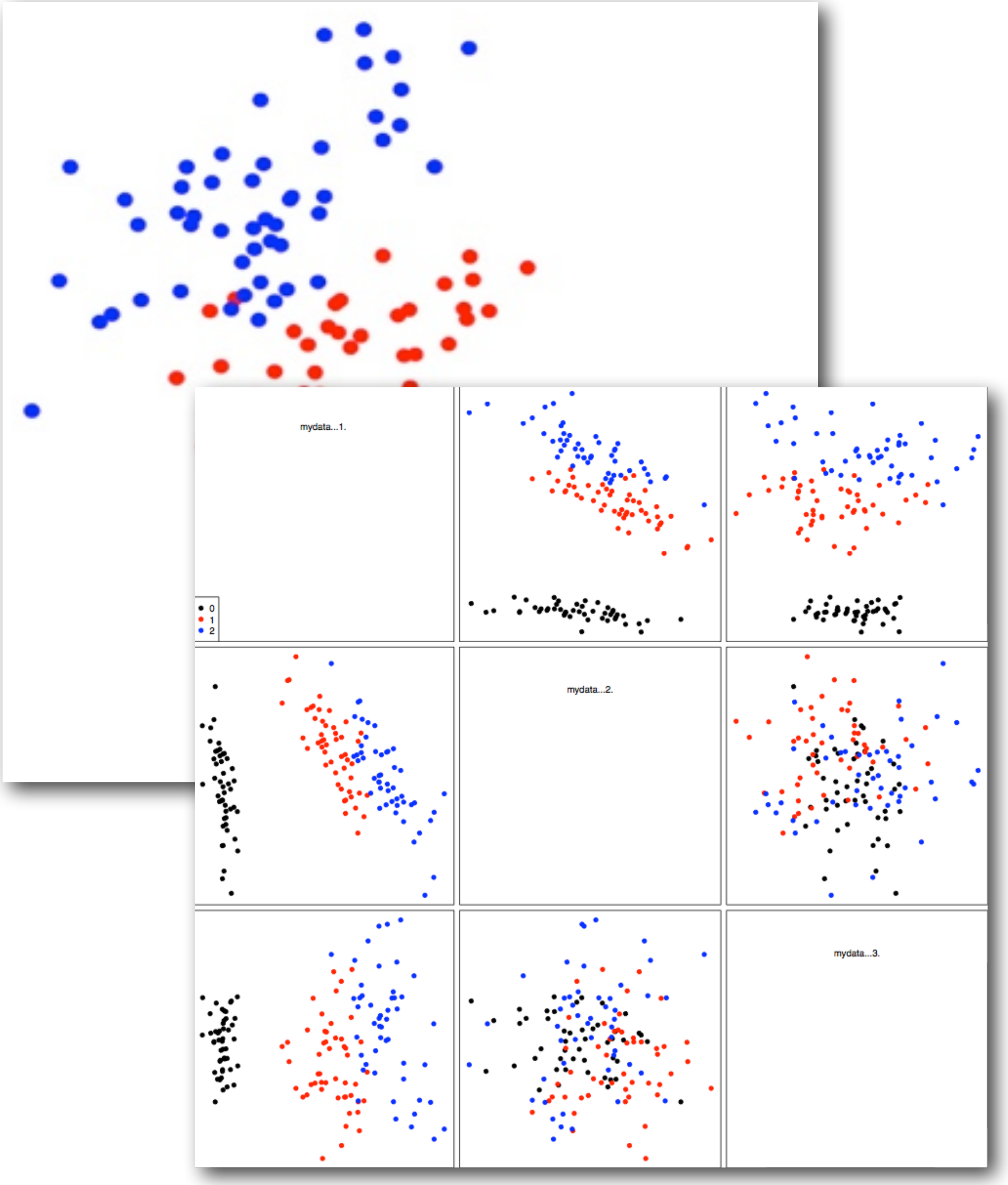
data: gauss-n100-10d-5smallCl, synthetic-Gaussian, robust PCA

3D vs.
best of (2D, SPLOM)



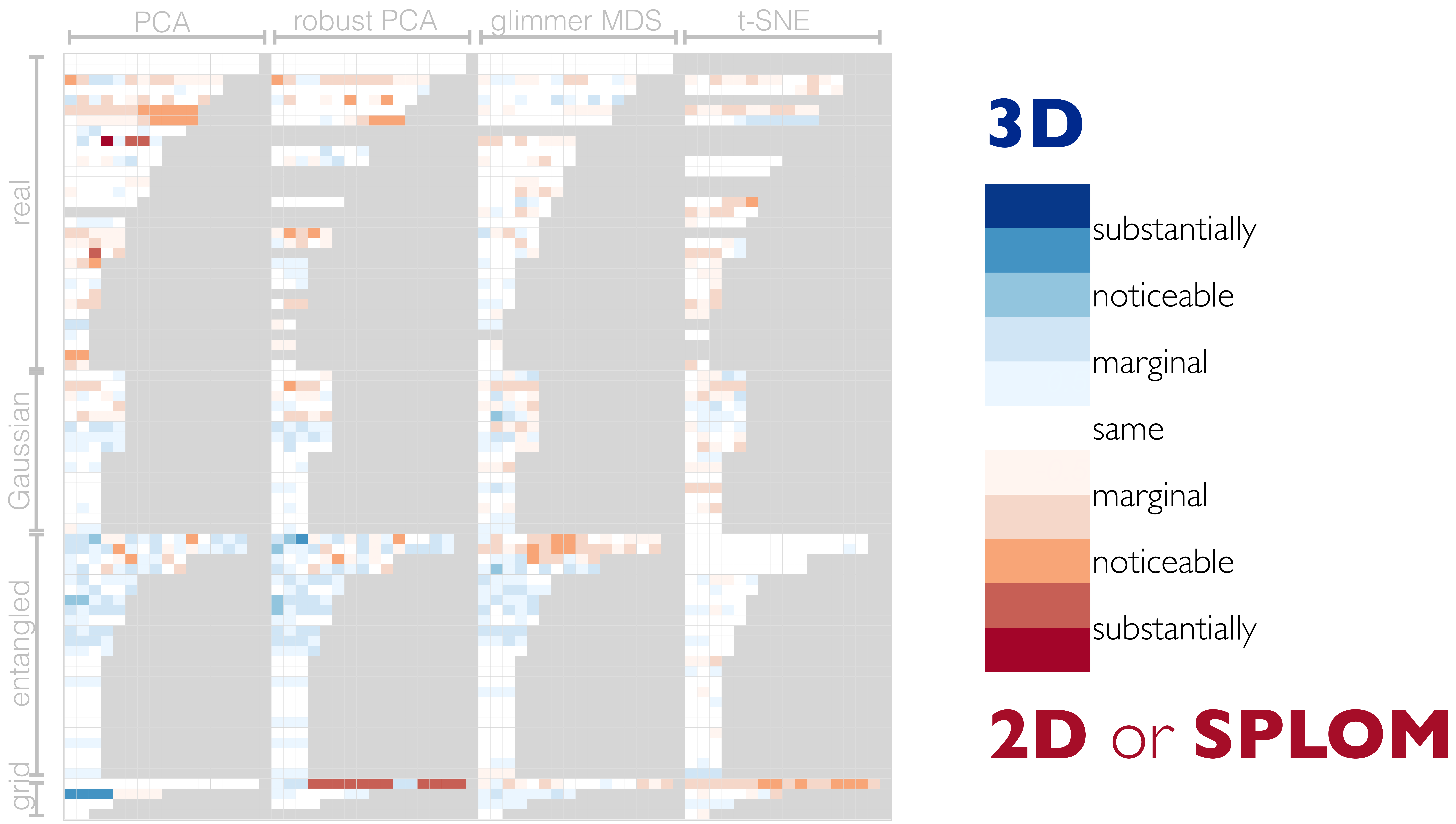
3D$_{PCA}$

which is better?

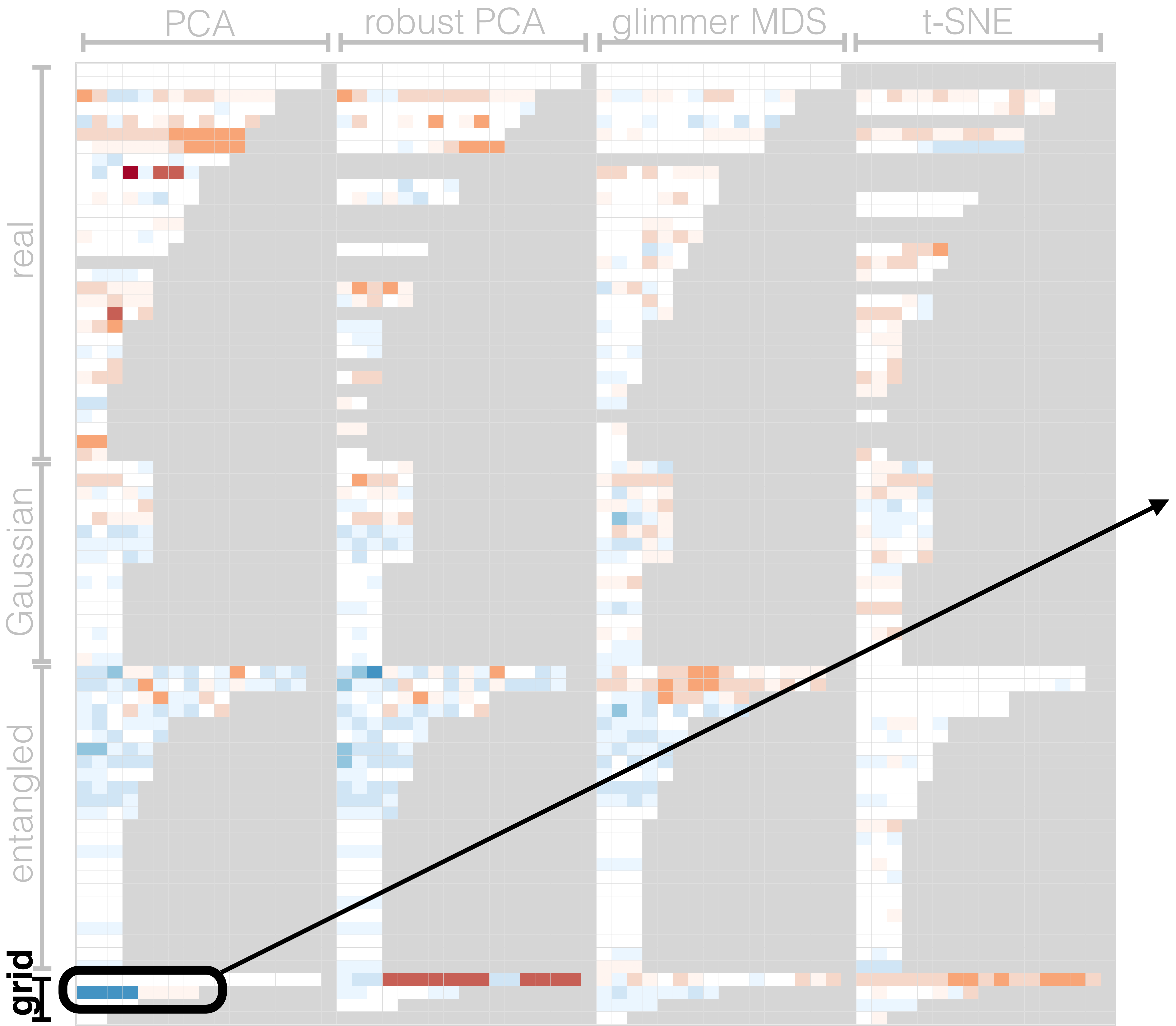


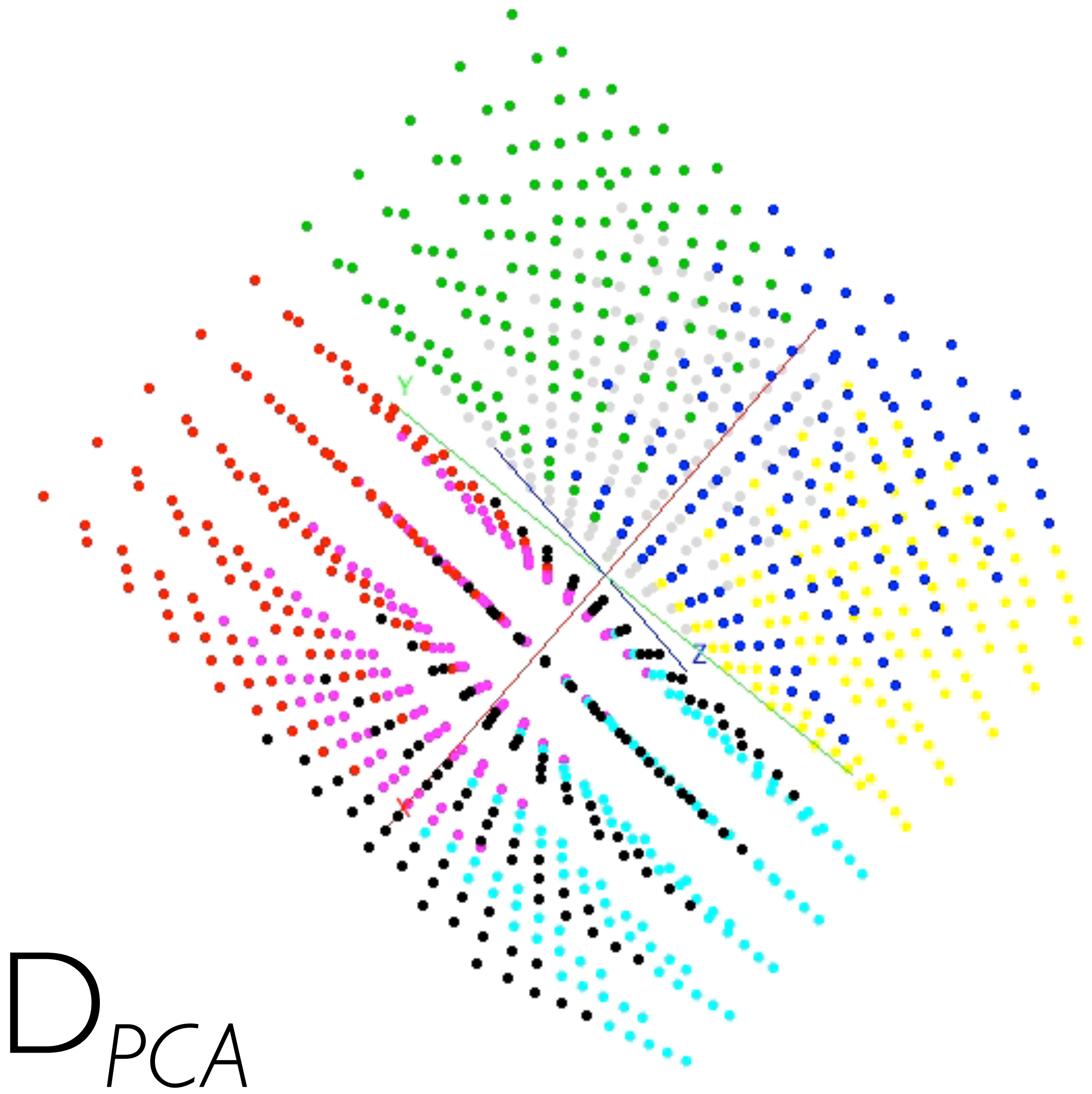
2D$_{PCA}$ or SPLOM$_{PCA}$

3D vs. (2D, SPLOM)



3D vs. (2D, SPLOM)





$3D_{PCA}$

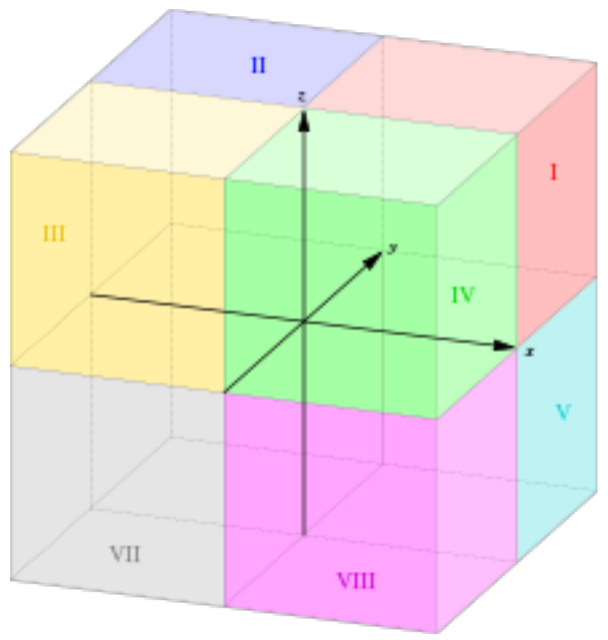
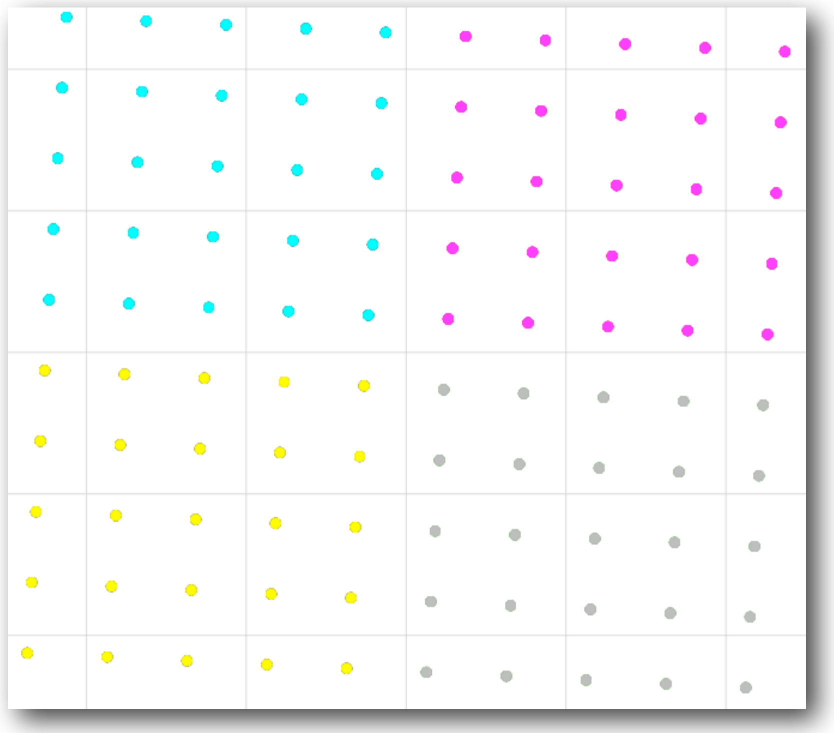


Image source: http://upload.wikimedia.org/wikipedia/commons/thumb/6/60/Octant_numbers.svg/220px-Octant_numbers.svg.png

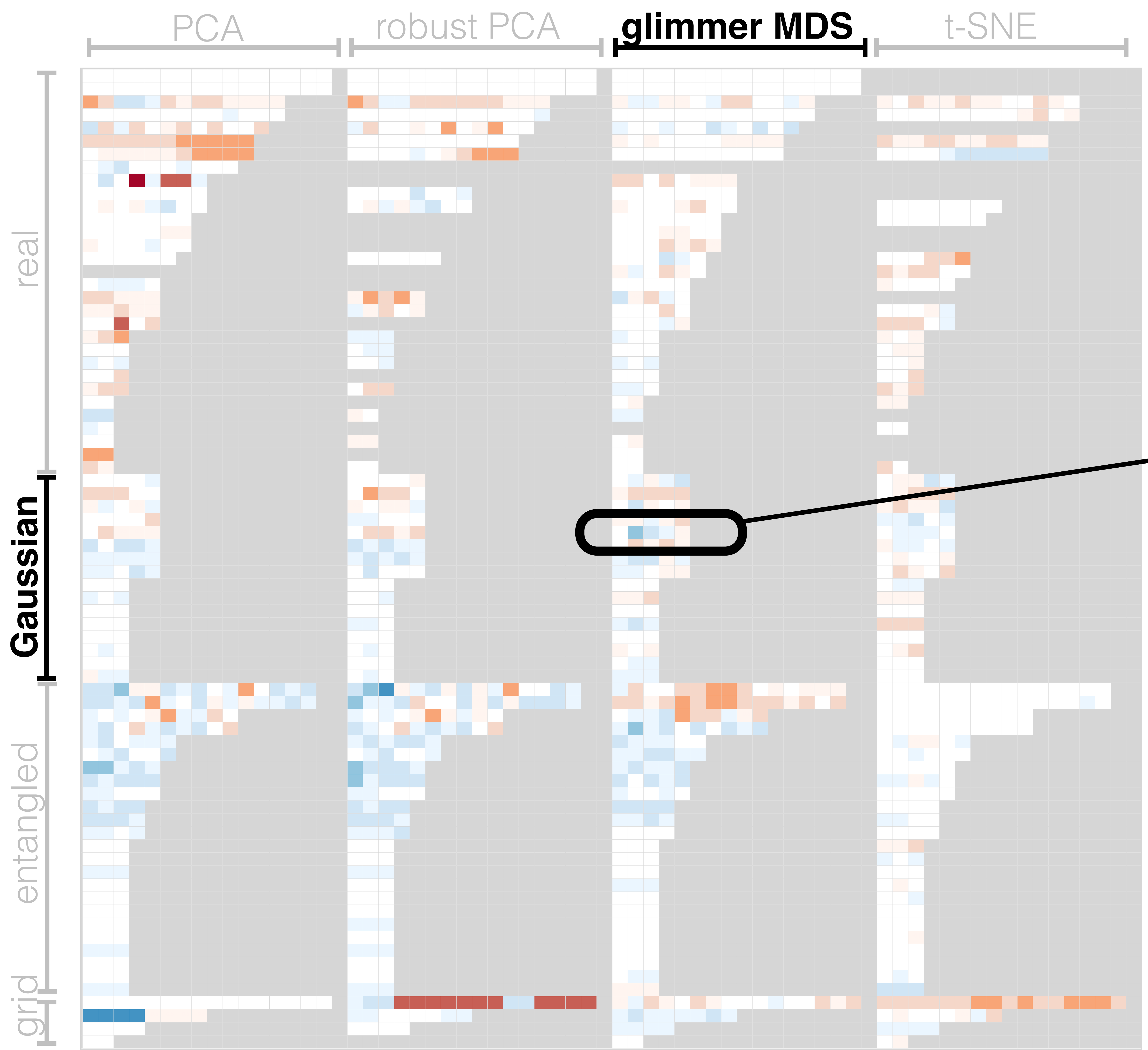
vs.

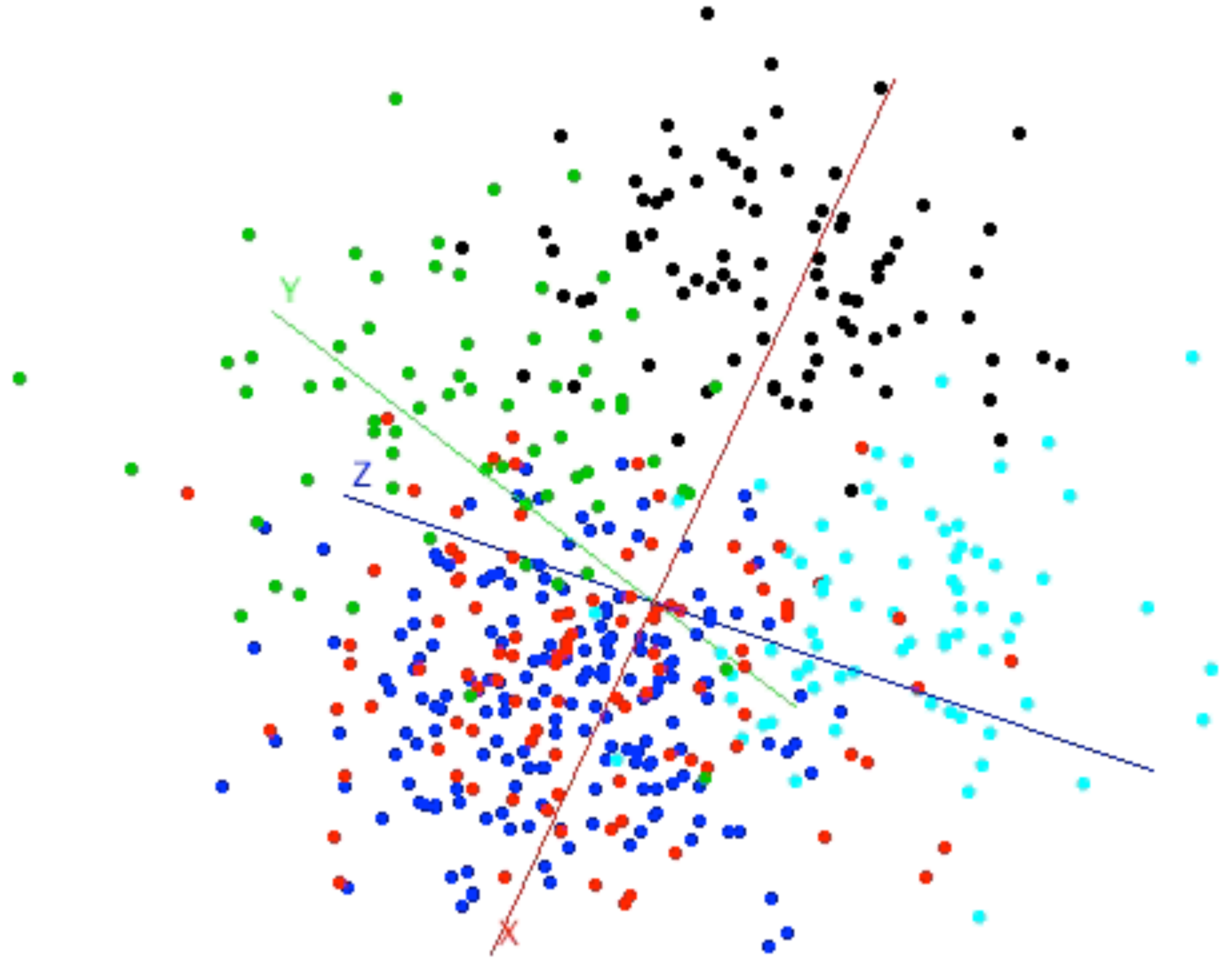


$2D_{PCA}$

$SPLOM_{PCA}$

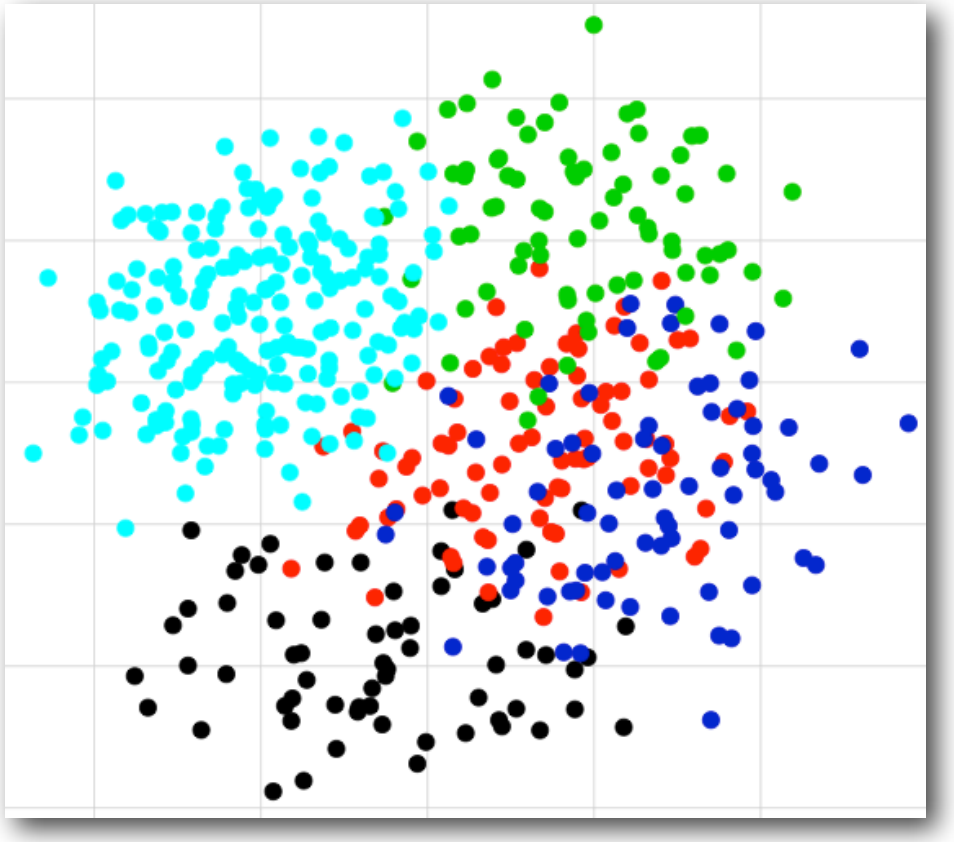
3D vs. (2D, SPLOM)



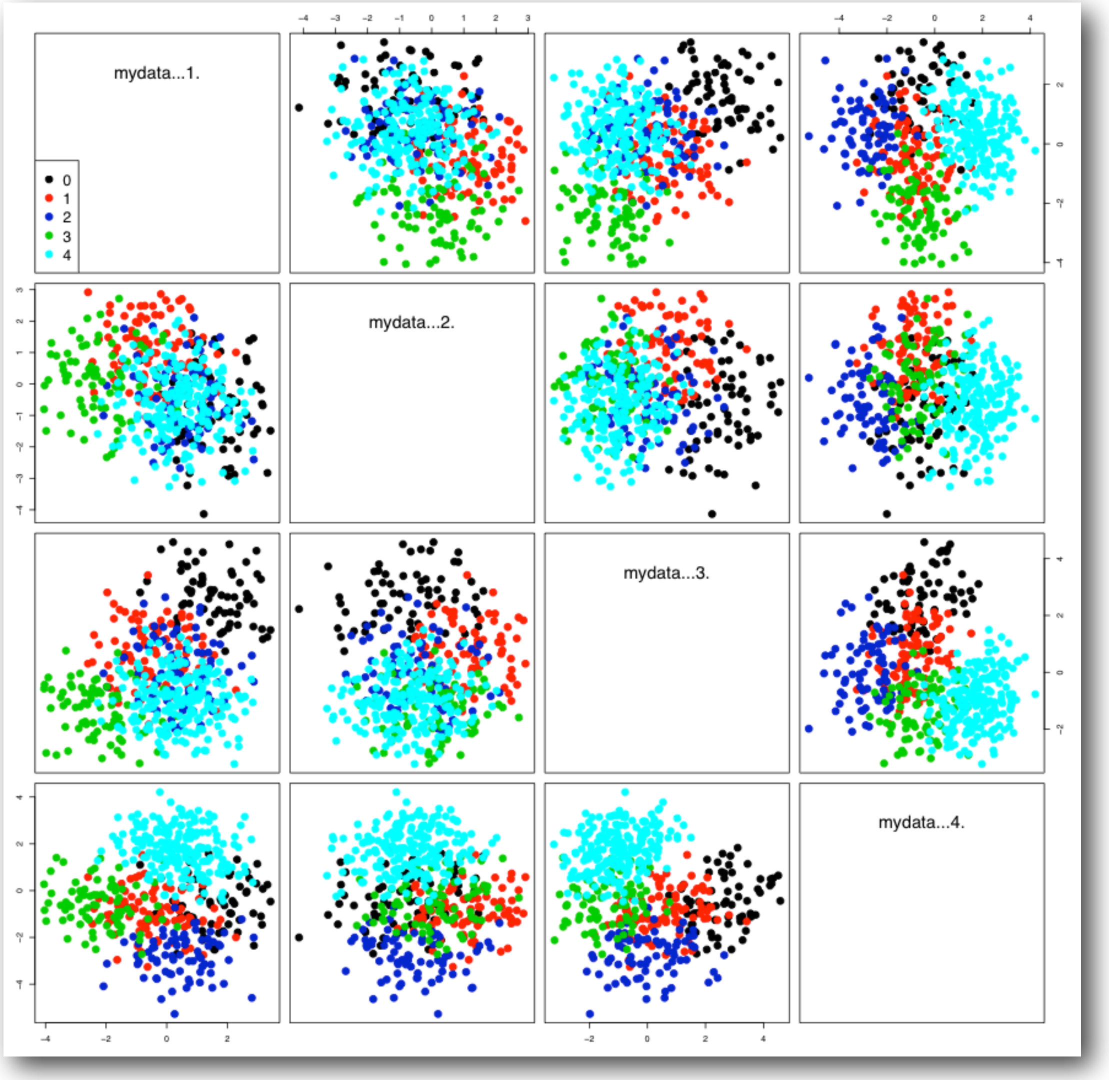


3D$_{MDS}$

vs.



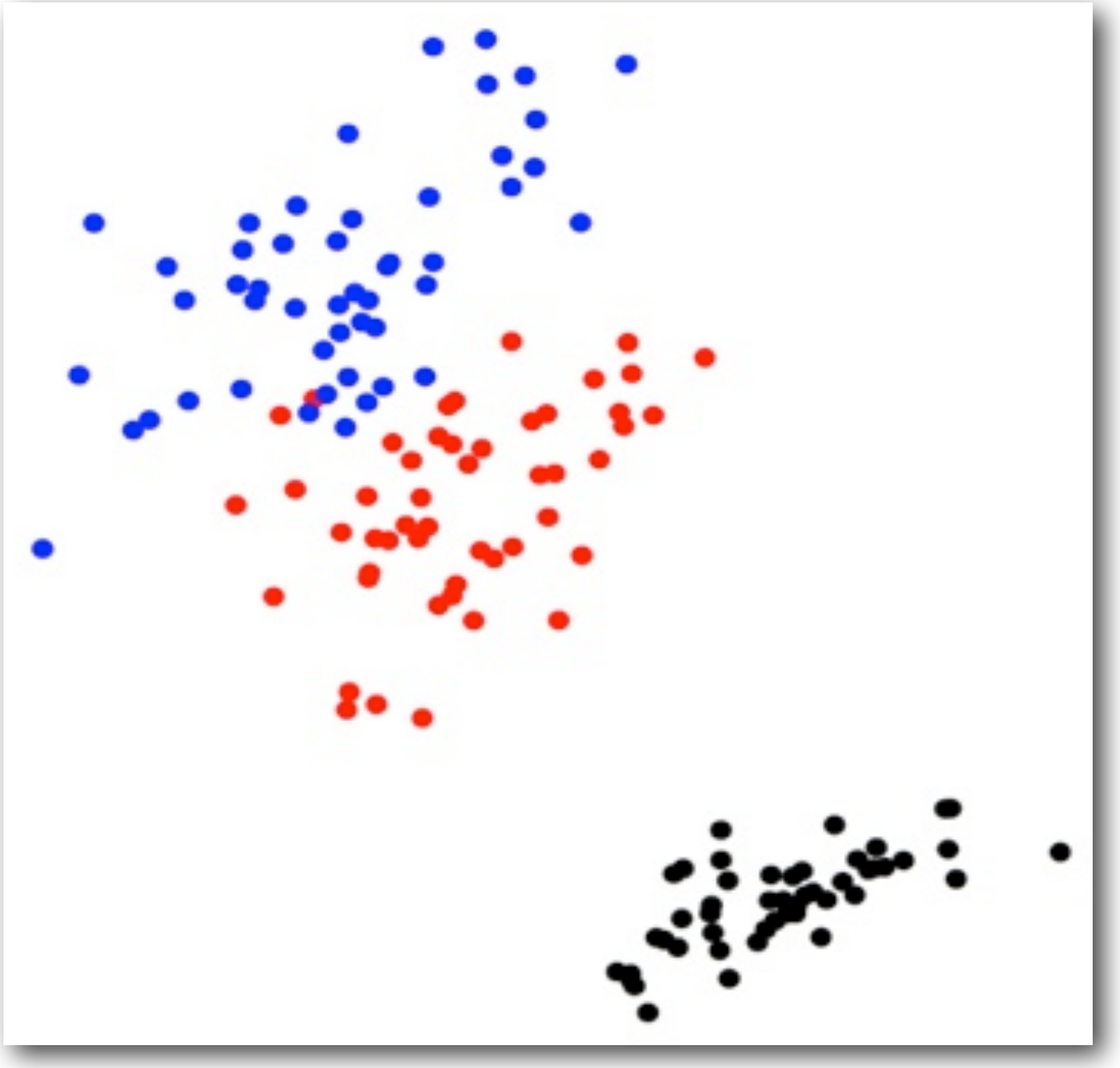
2D$_{MDS}$



SPLOM$_{MDS}$

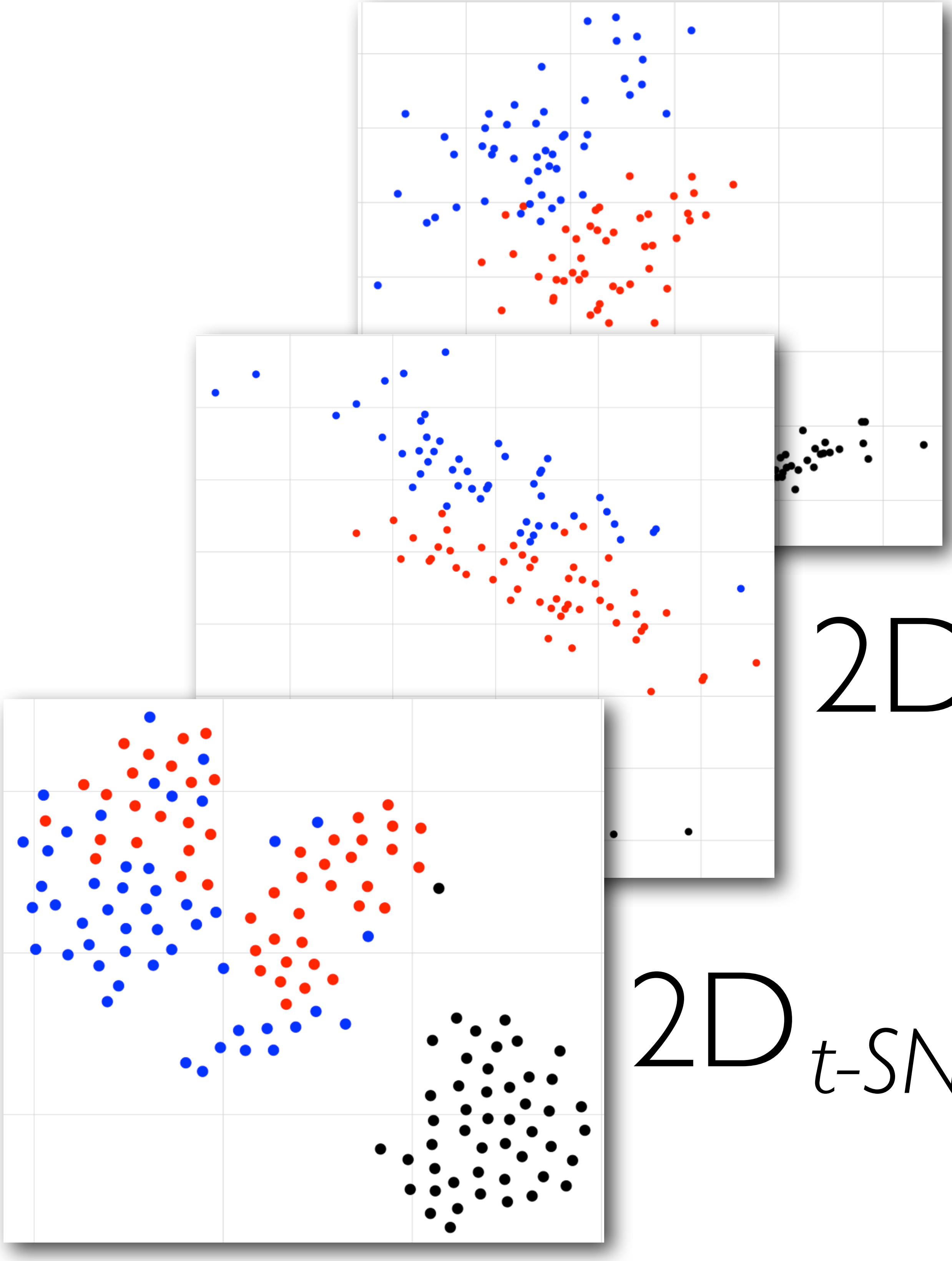
Between-DR

2D vs.
best of (2D*from other DRs*)



2D *PCA*

which is better?

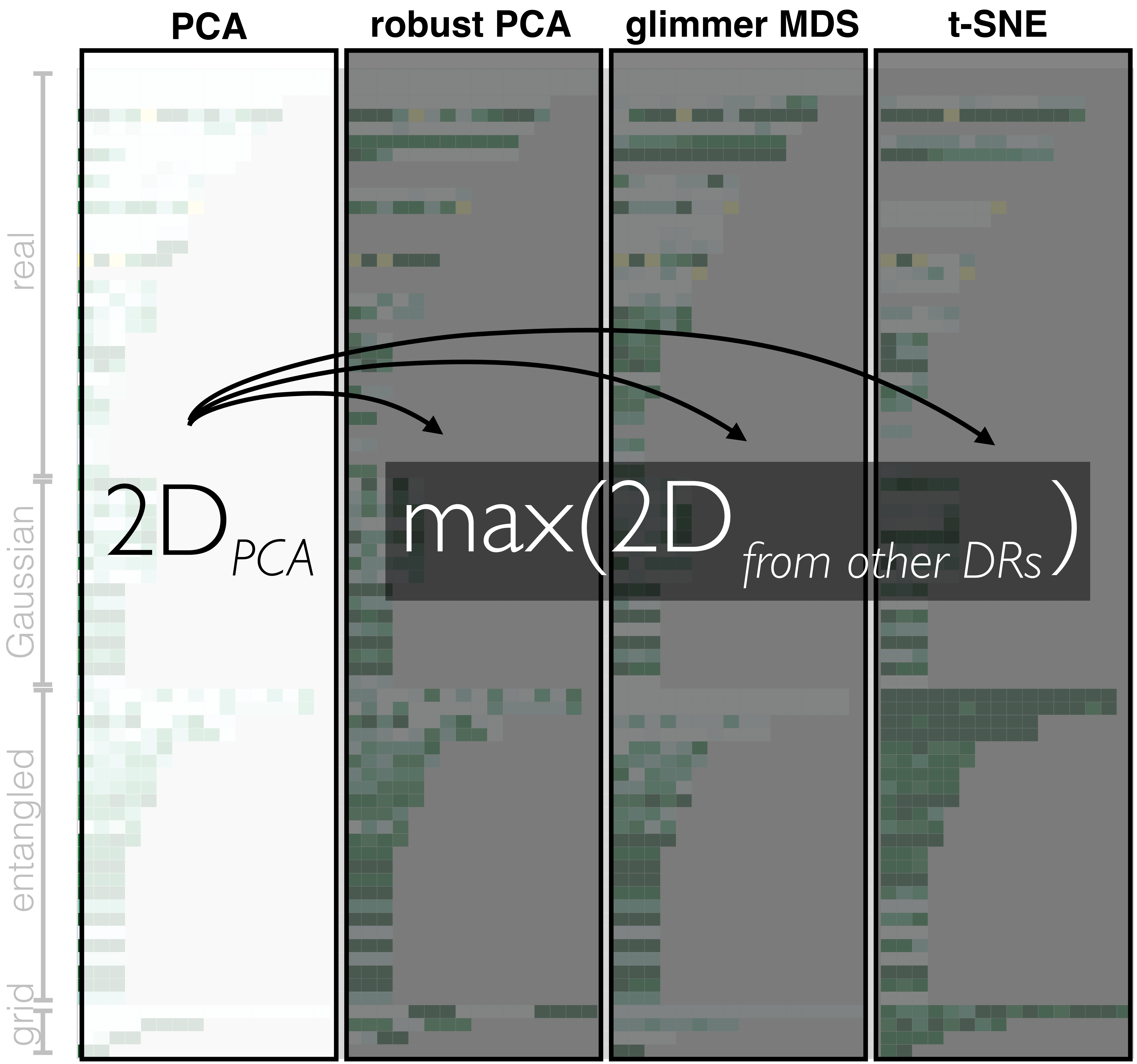


2D *robust PCA*

2D *glimmer MDS*

2D *t-SNE*

2D vs. ($2D_{from\ other\ DRs}$)

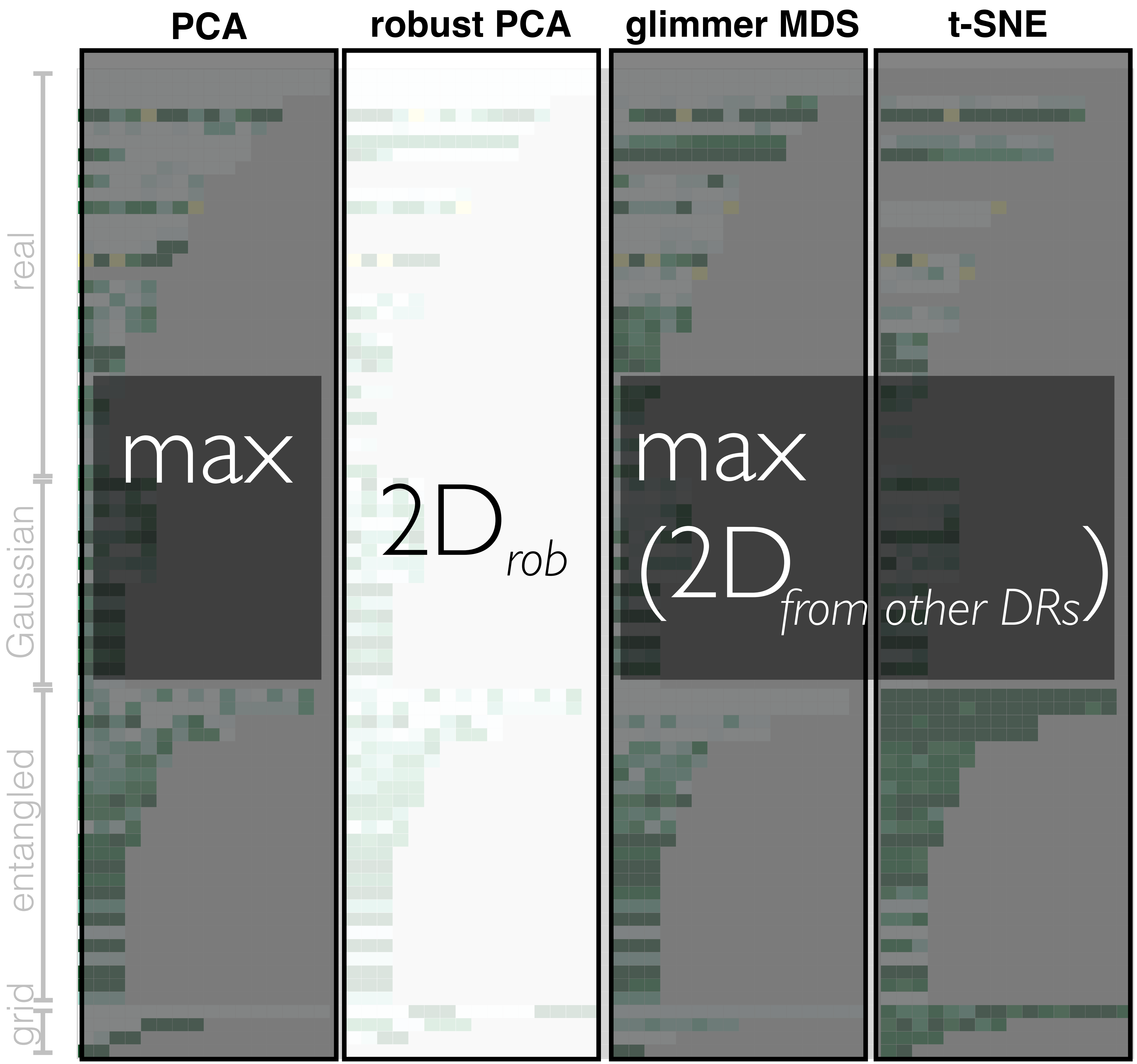


Cross-column

differences in

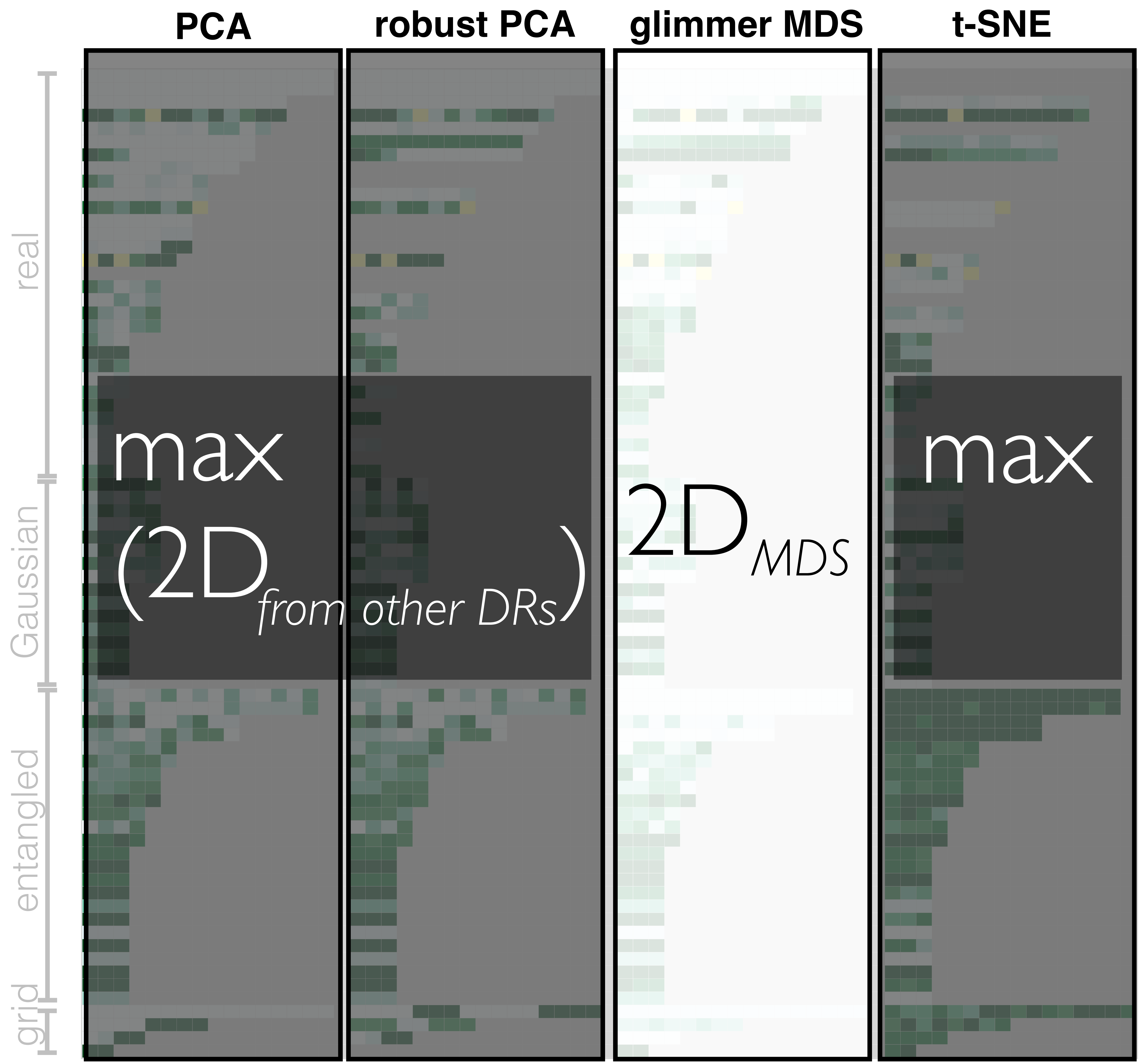
2D base heatmap

2D vs. ($2D_{from\ other\ DRs}$)



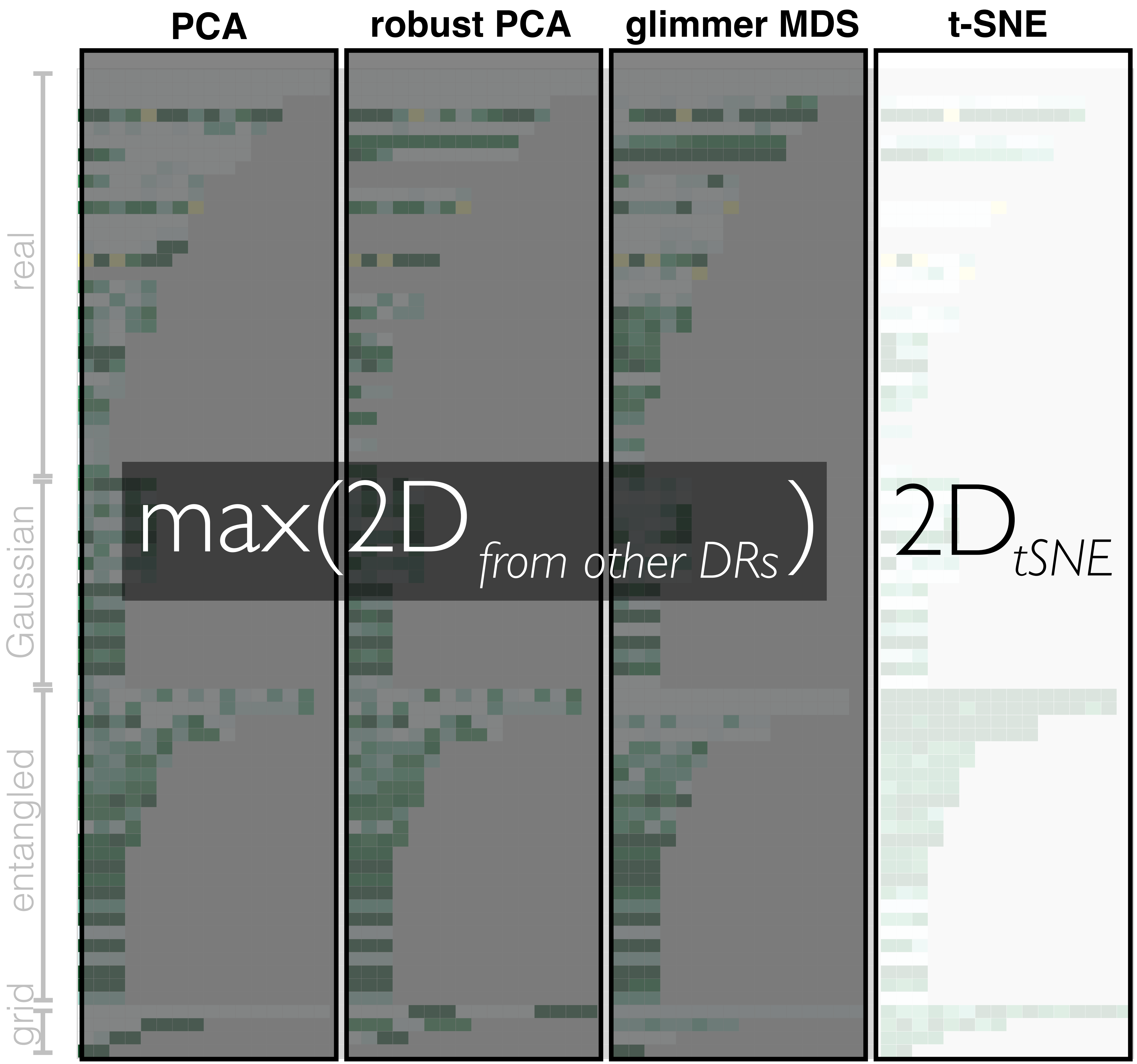
Cross-column differences in 2D base heatmap

2D vs. $(2D_{from\ other\ DRs})$



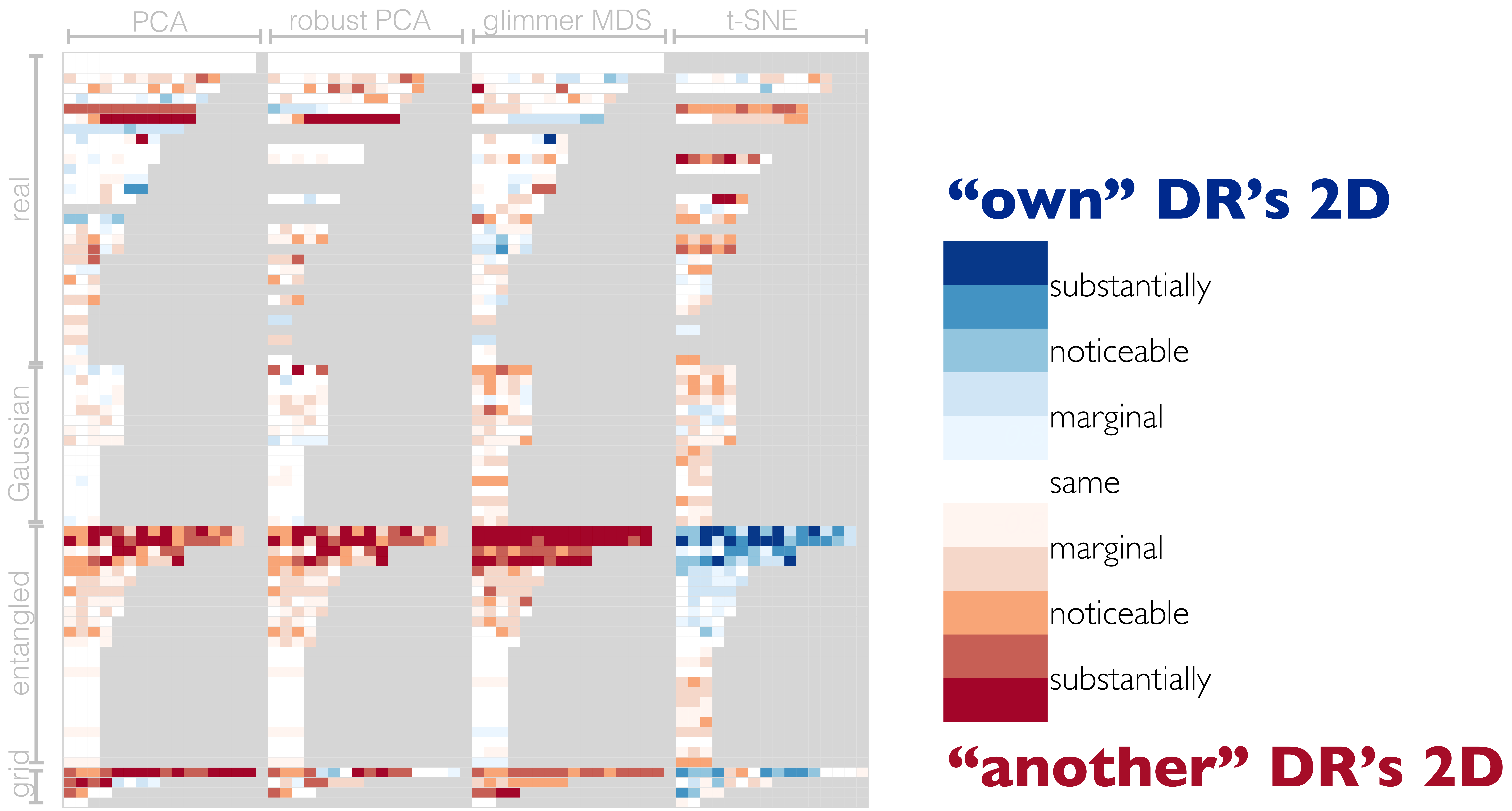
Cross-column
differences in
2D base heatmap

2D vs. ($2D_{\text{from other DRs}}$)

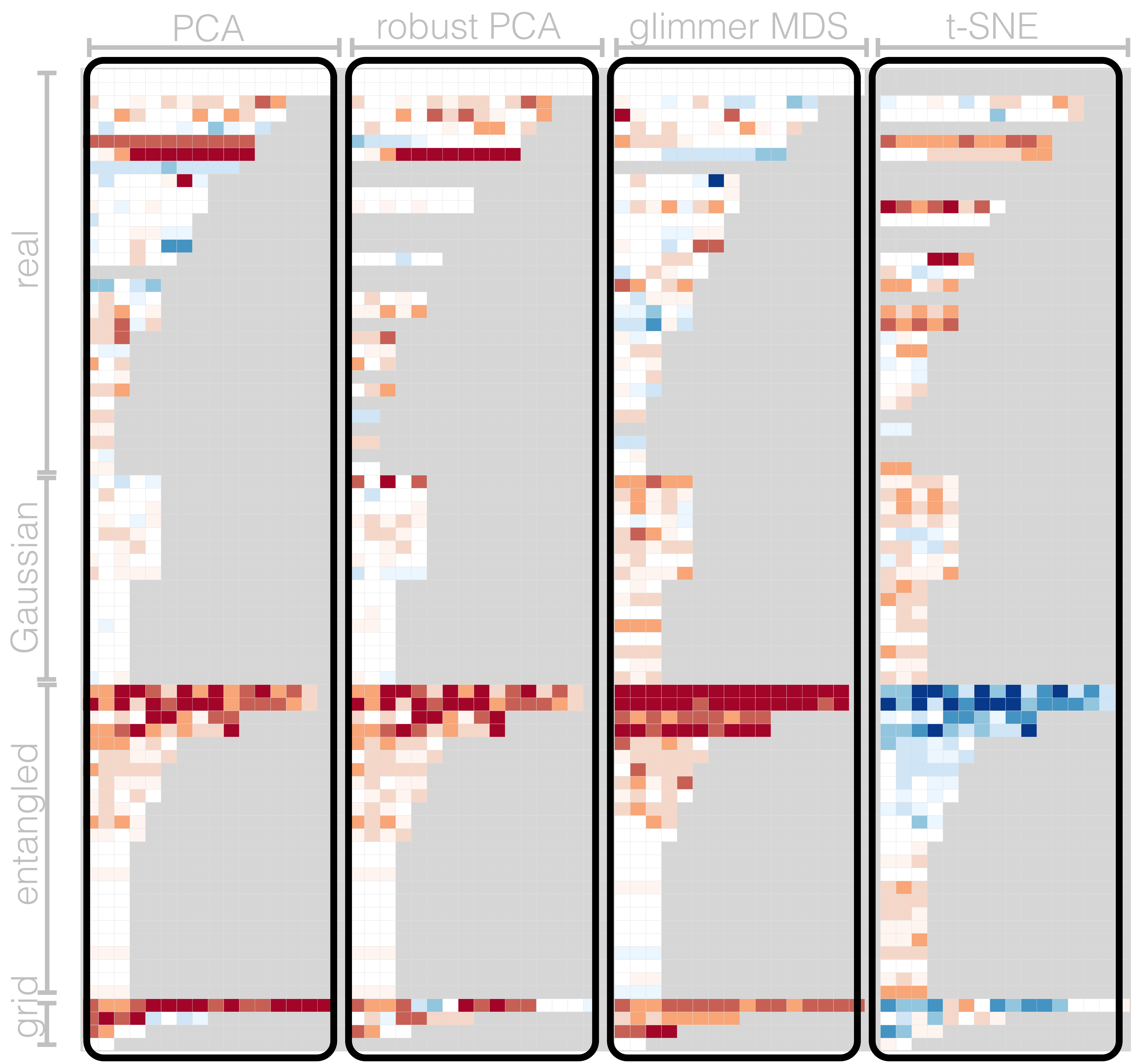


Cross-column
differences in
2D base heatmap

2D vs. $(2D_{\text{from other DRs}})$



2D vs. (2D*from other DRs*)



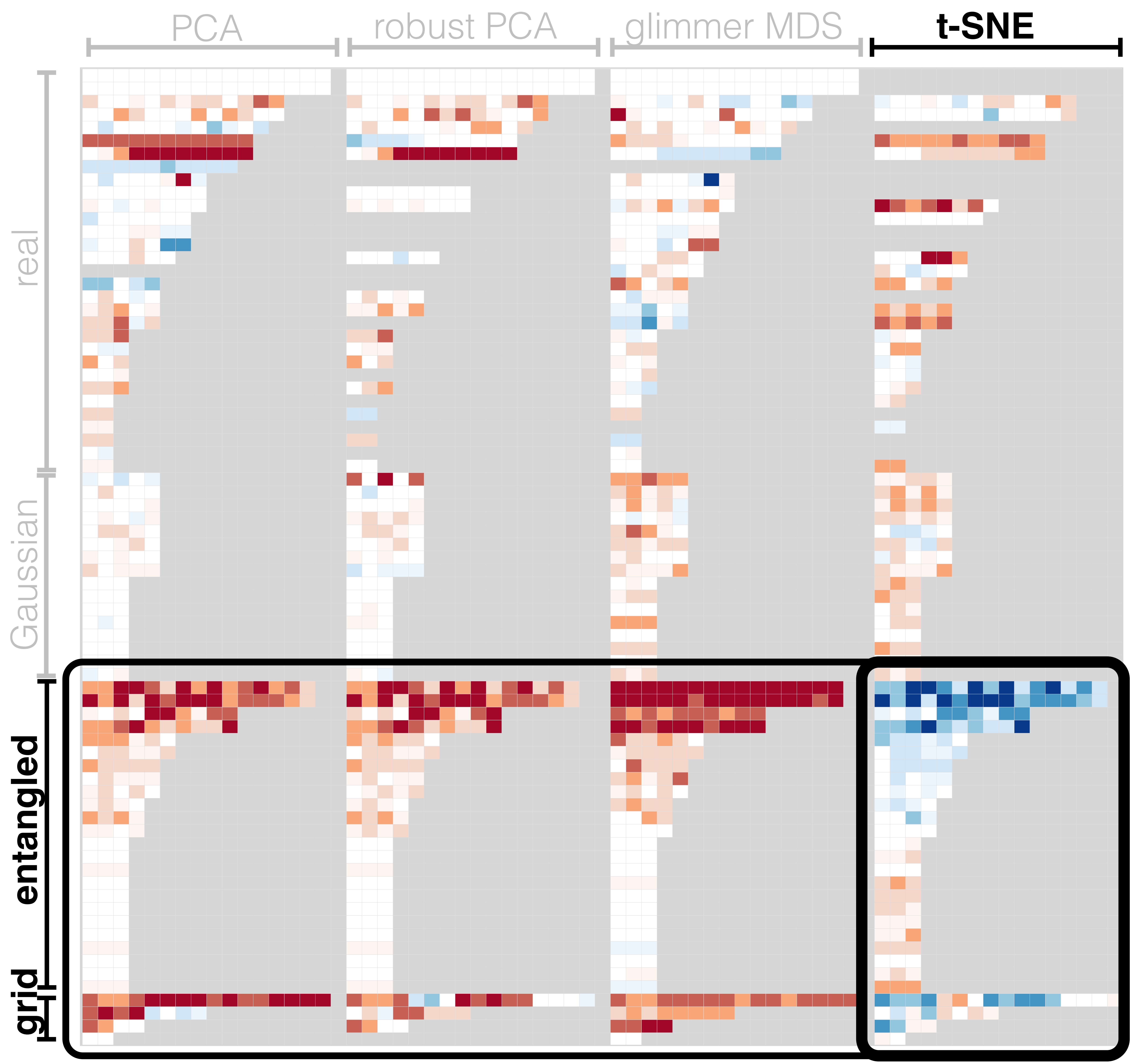
no one and only DR

"own" DR's 2D

- substantially
- noticeable
- marginal
- same
- marginal
- noticeable
- substantially

"another" DR's 2D

2D vs. (2D$_{\textit{from other DRs}}$)



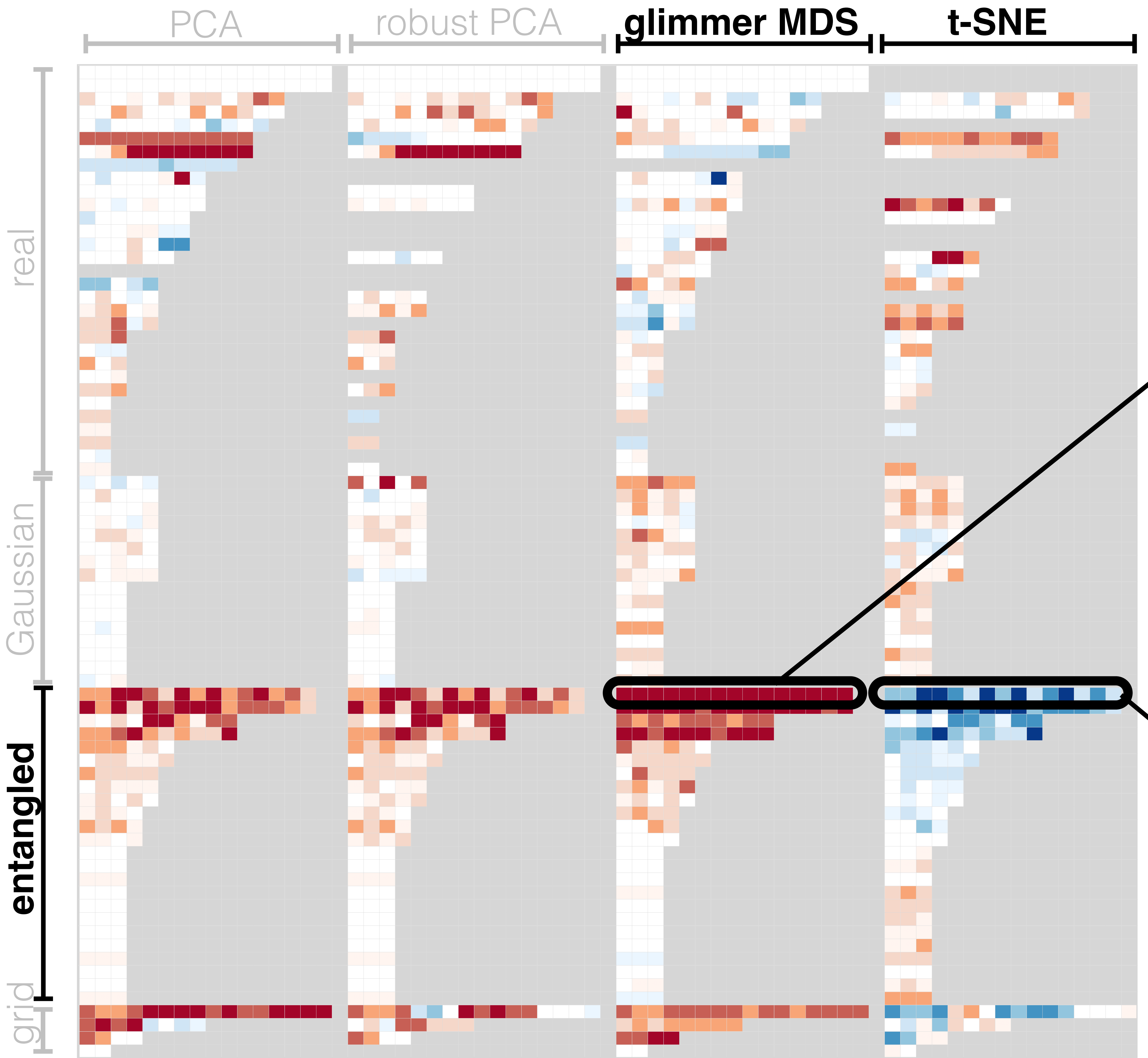
t-SNE good for highly synthetic datasets:

entangled
(intended to benefit 3D)

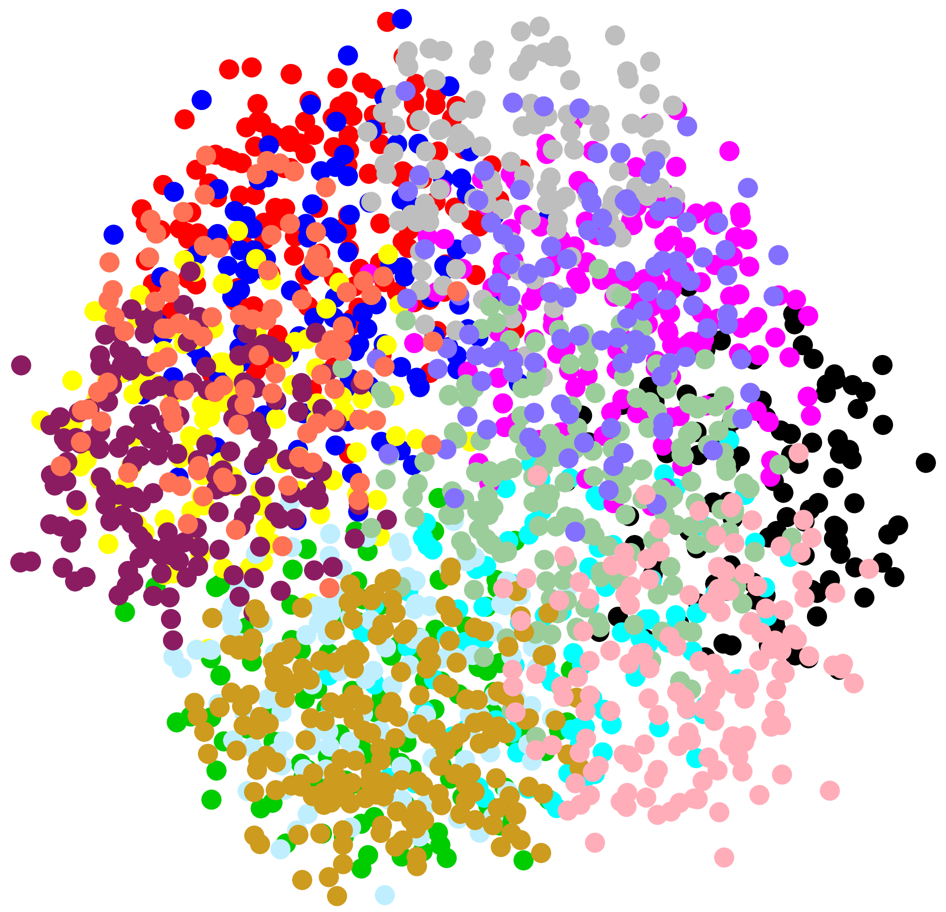
grid

2D vs. (2D$_{from\ other\ DRs}$)

same dataset, different DR

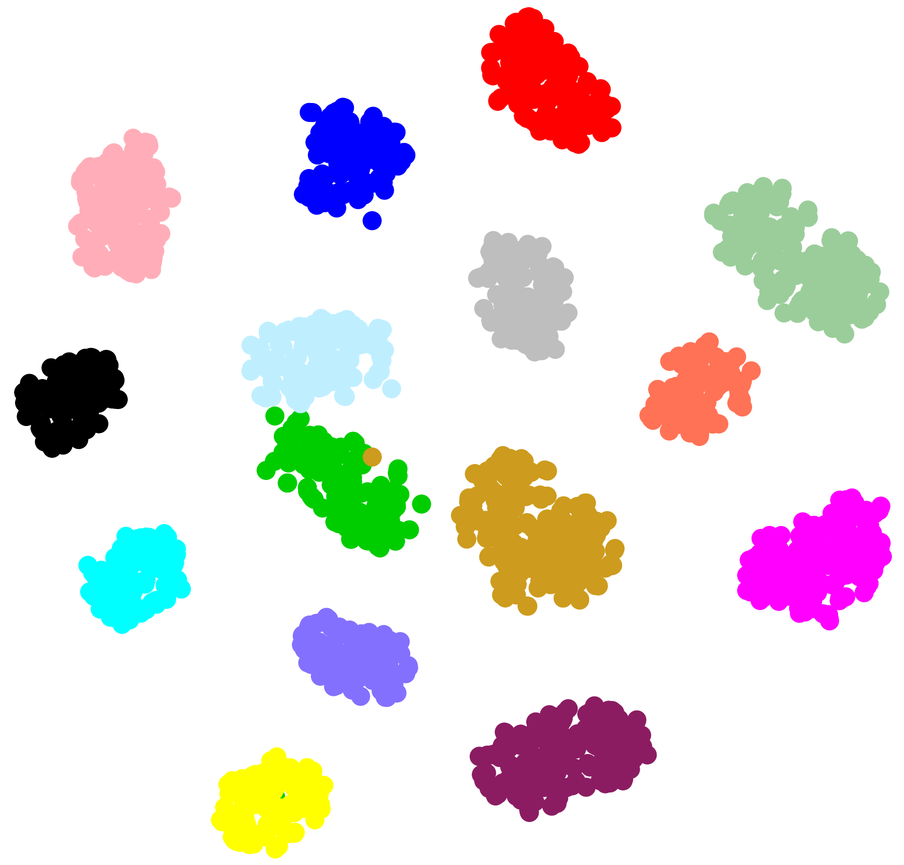


$2D_{MDS}$



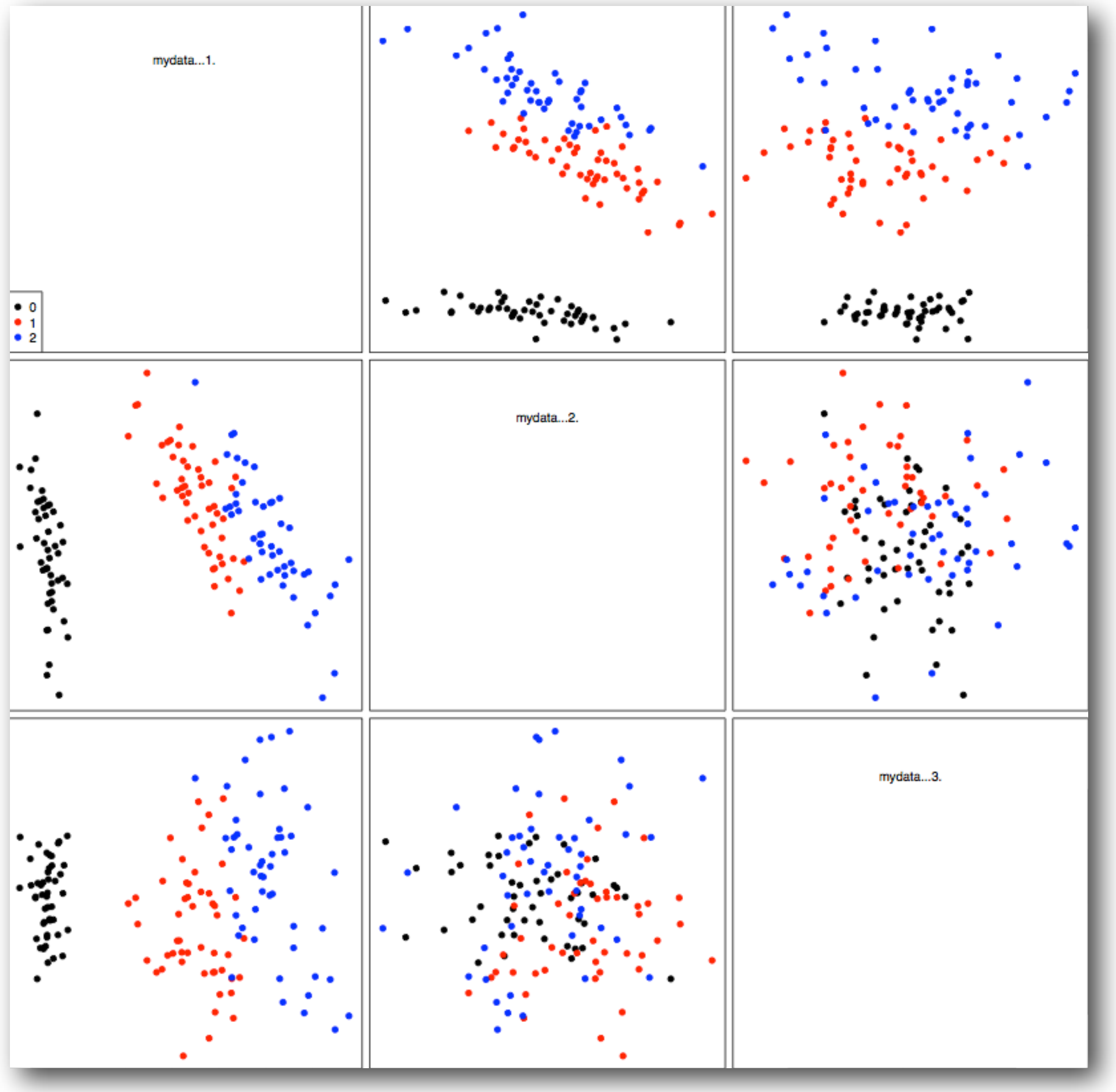
vs.

$2D_{t\text{-}SNE}$



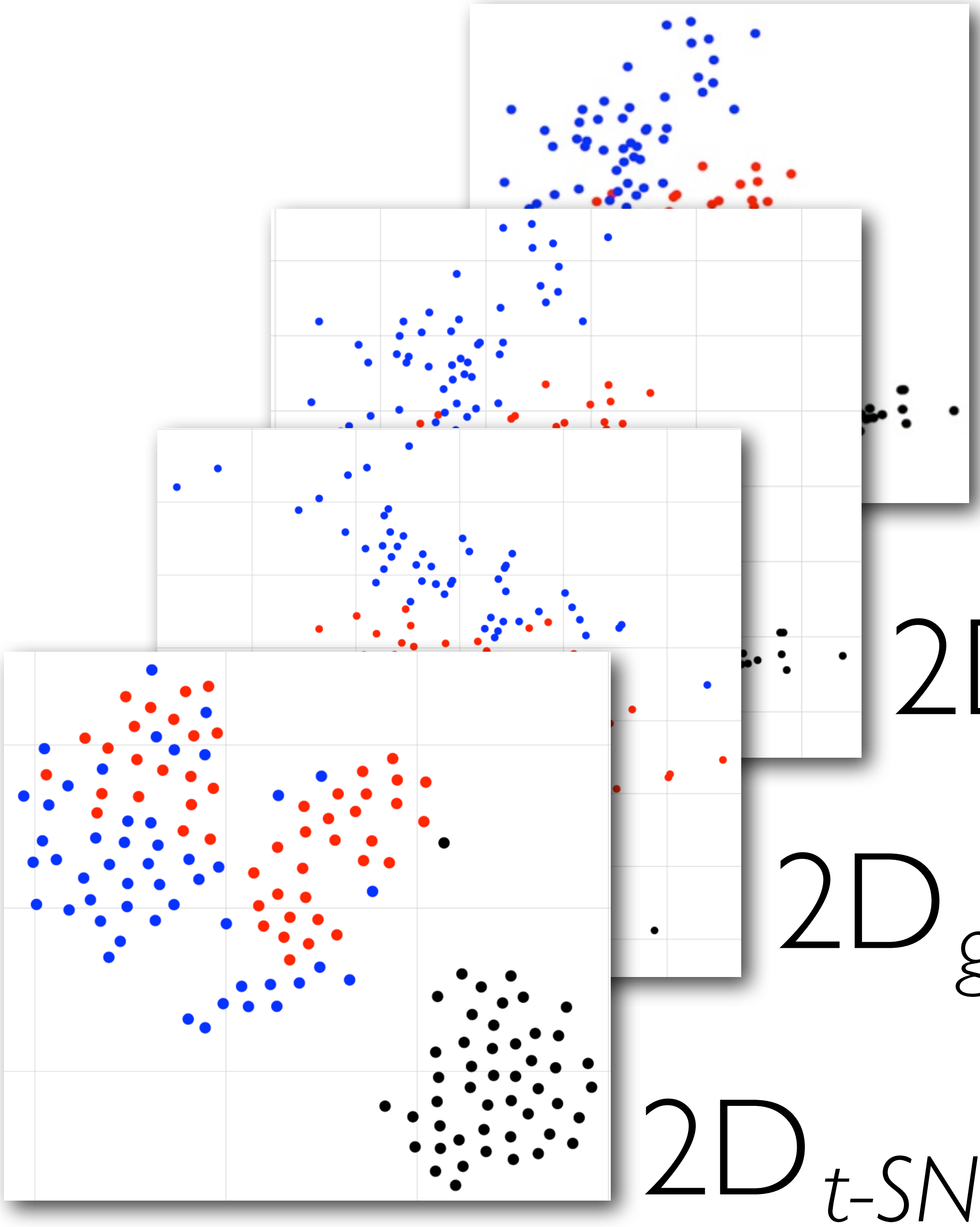
data: entangled2-15d-adjacent, synthetic-entangled,glimmer MDS / t-SNE

SPLOM vs. **best of (2D*from all DRs*)**



SPLOM *PCA*

which is better?



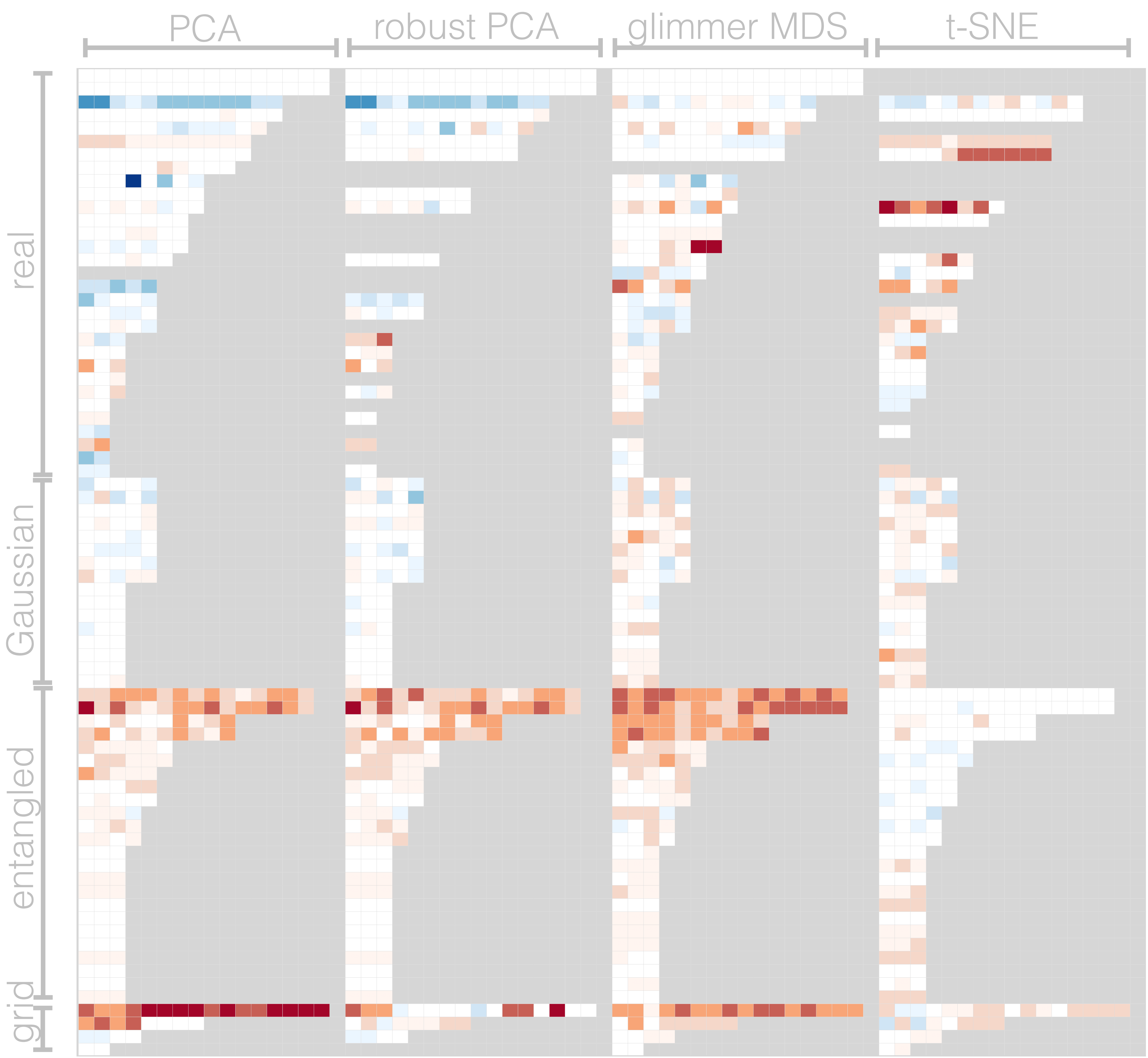
2D *PCA*

2D *robust PCA*

2D *glimmer MDS*

2D *t-SNE*

SPLOM vs. (2D$_{\textit{from all DRs}}$)



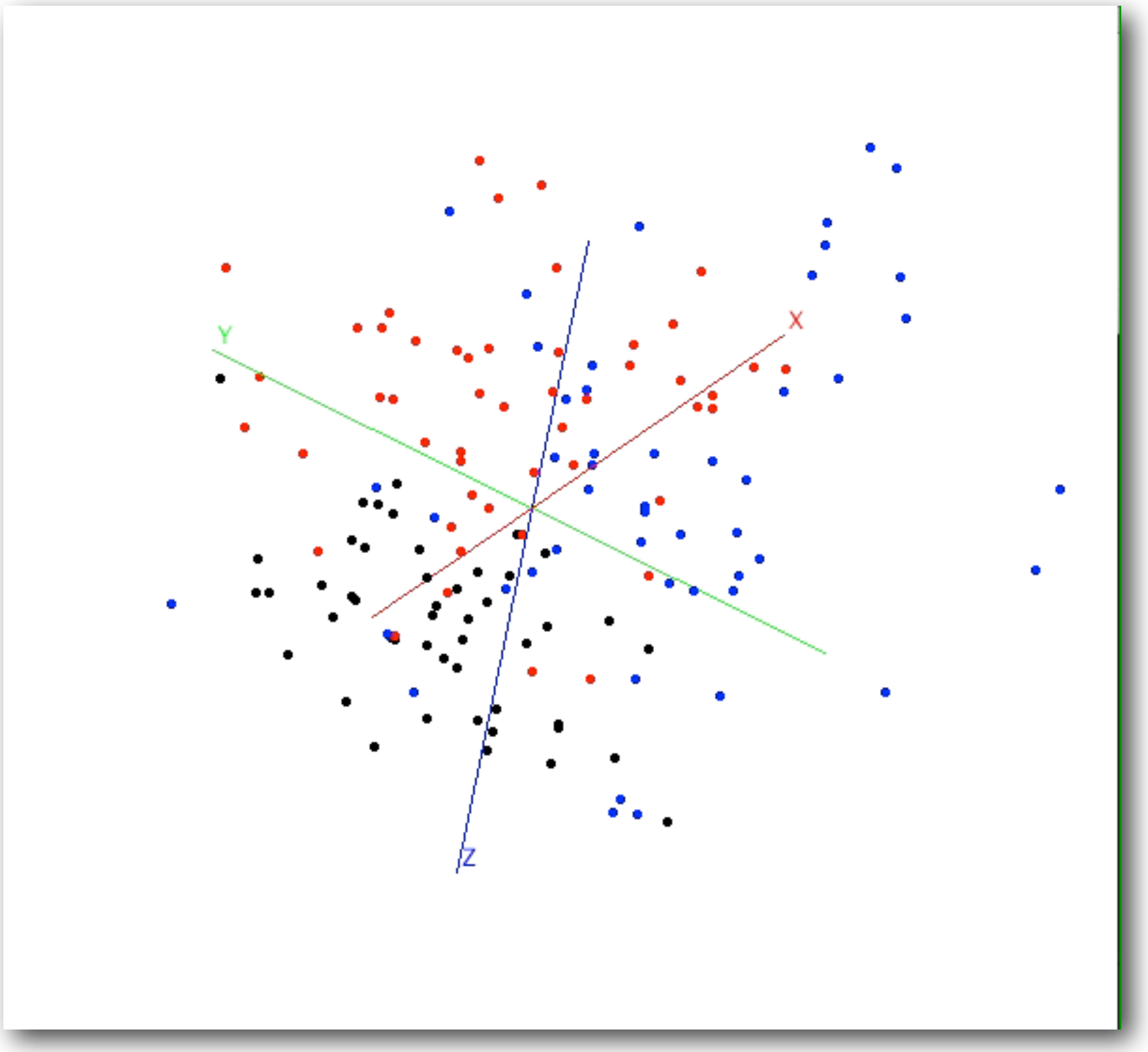
SPLOM

substantially
noticeable
marginal
same
marginal
noticeable
substantially

one of DR's 2D

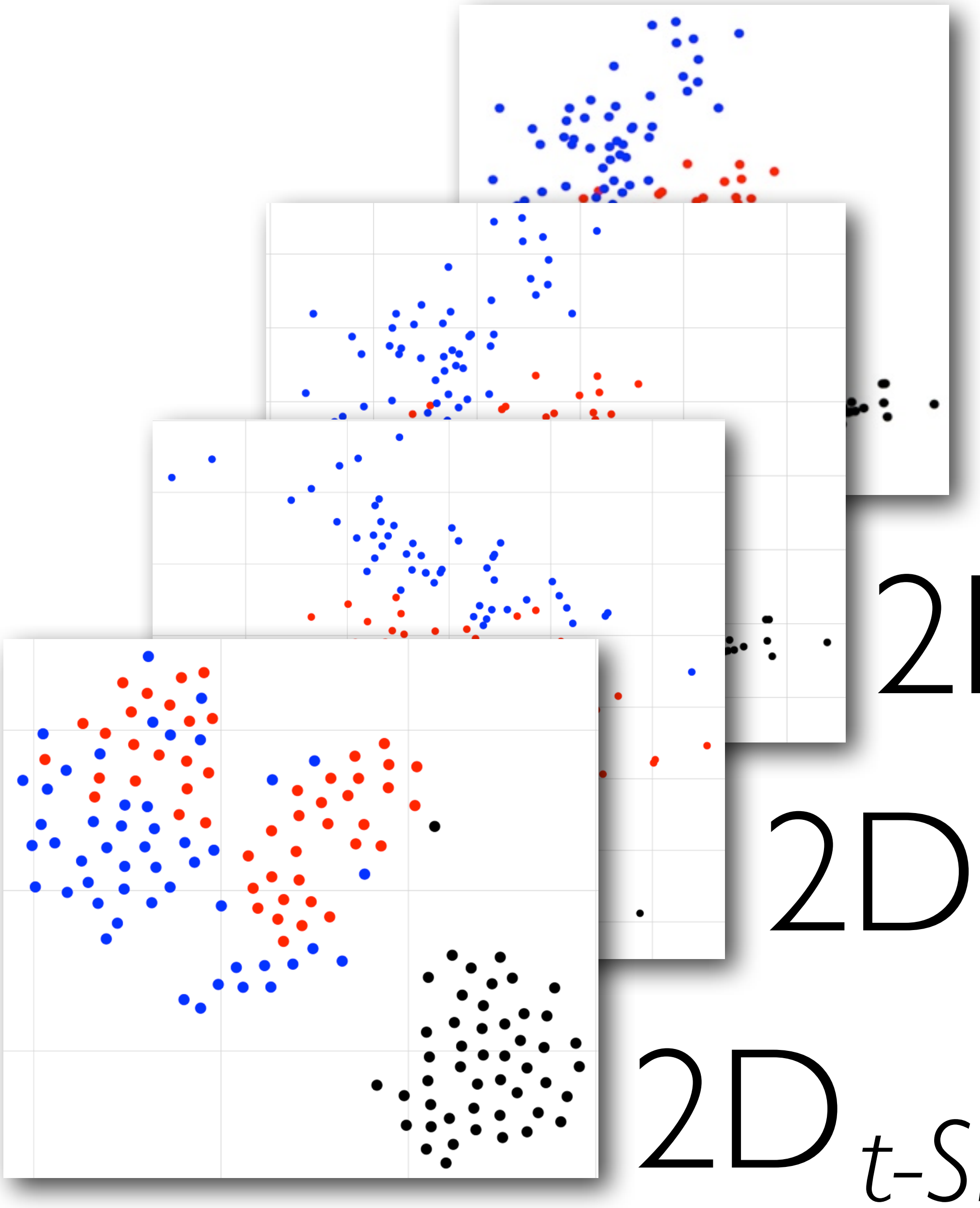
SPLOM vs. (2D$_{from\ all\ DRs}$)

3D vs.
best of (2D$_{\textit{from all DRs}}$, SPLOM)



3D$_{PCA}$

which is better?

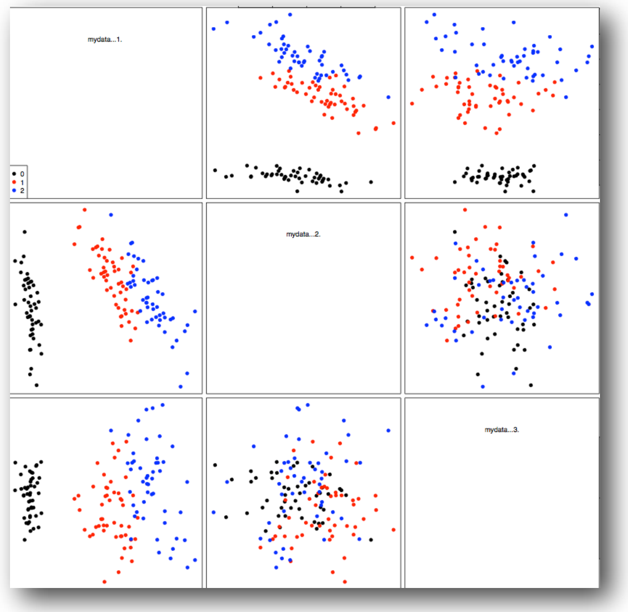


2D$_{PCA}$

2D$_{robust\ PCA}$

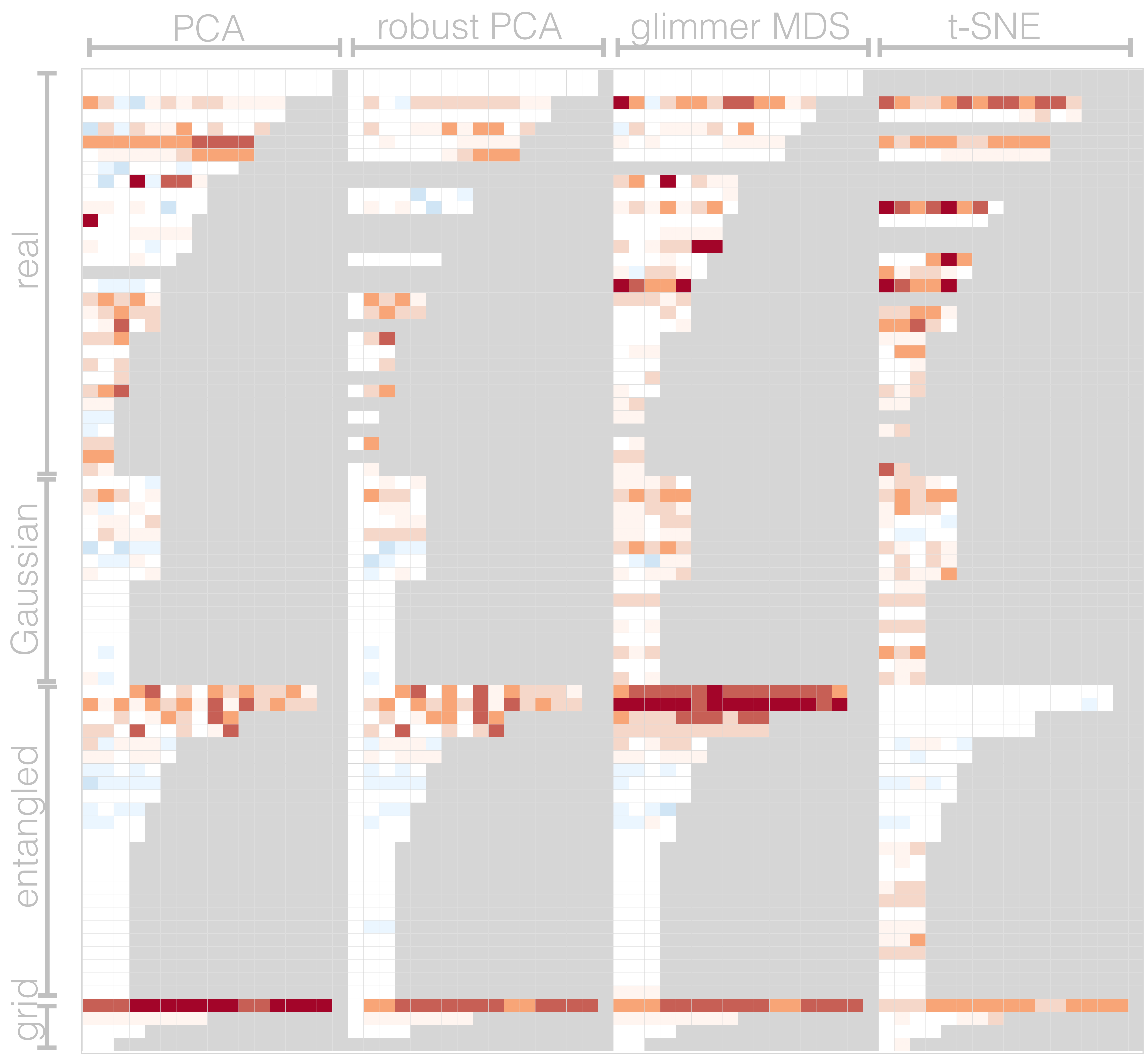
2D$_{glimmer\ MDS}$

2D$_{t\text{-}SNE}$



SPLOM$_{PCA}$

3D vs. (SPLOM$_{own}$, 2D$_{from\ all\ DRs}$)



no noticeably better class in 3D

3D

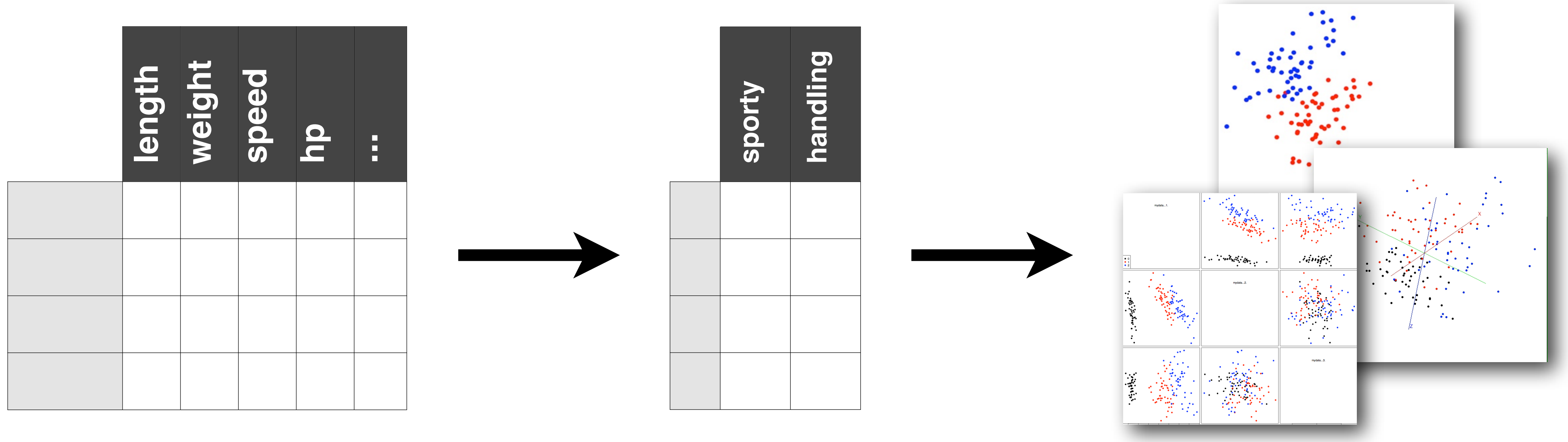
SPLOM or **one of DR's 2D**

Summary

Summary

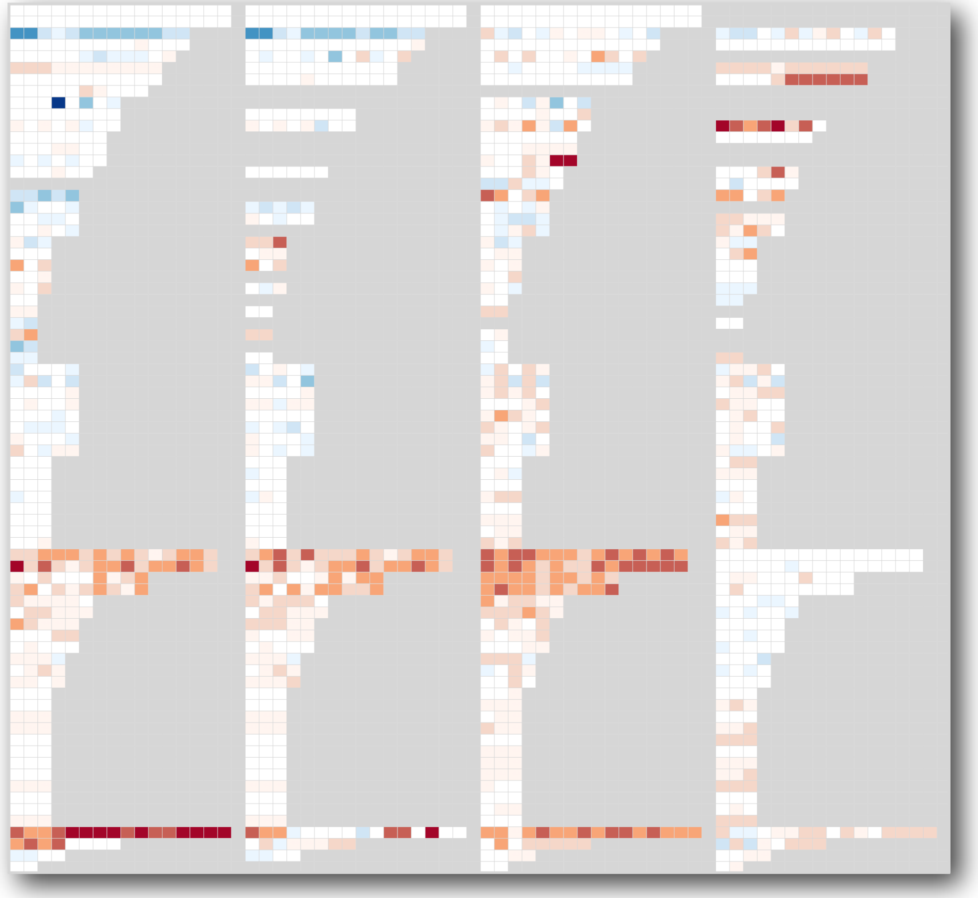
Which visual encoding to use for dimensionally reduced data?

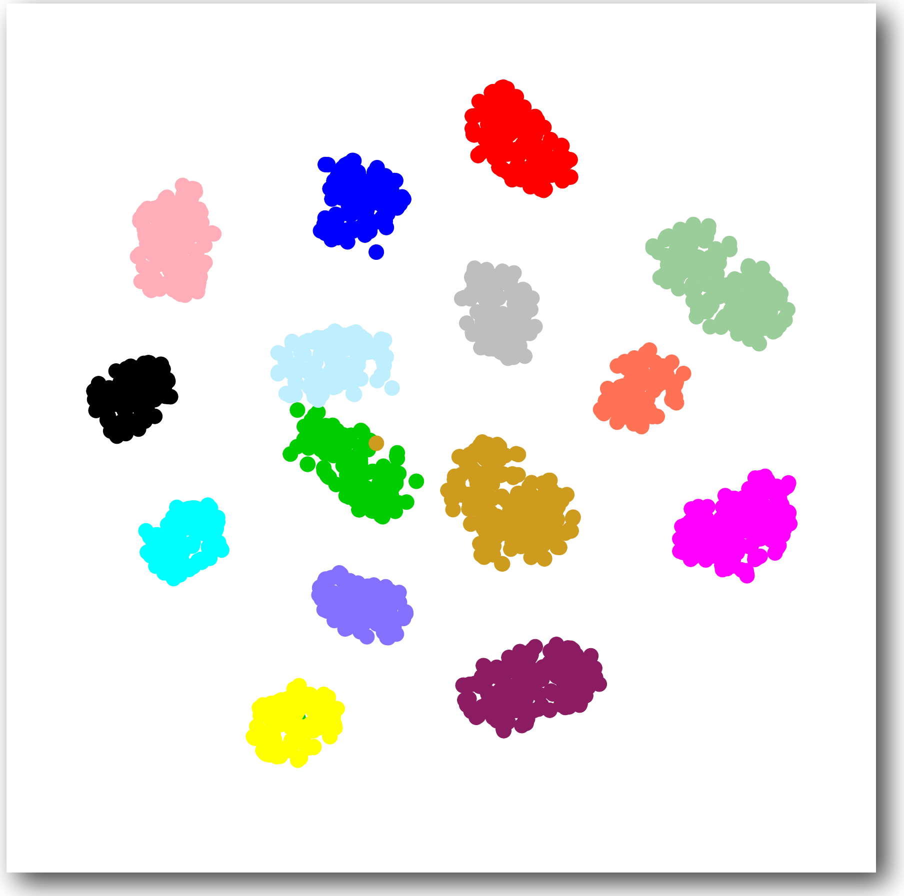
- 2D, interactive 3D, SPLOM?



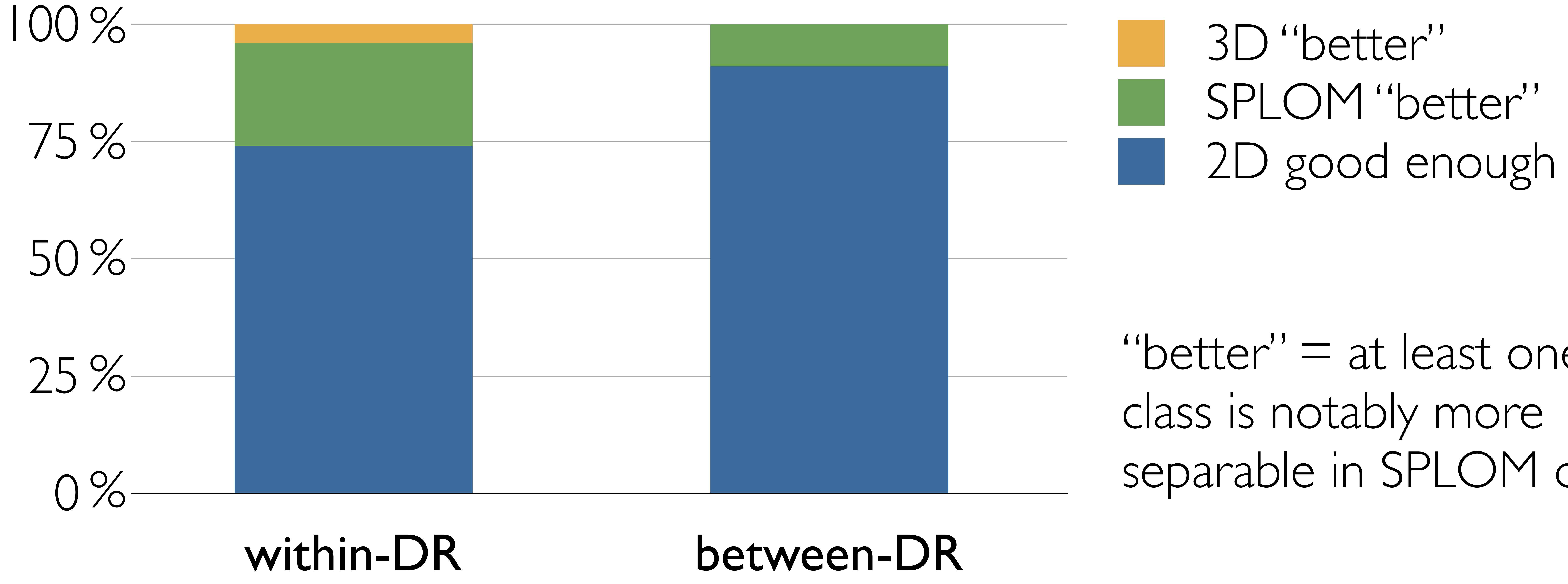
Data study

- Heatmap analysis
- Examples





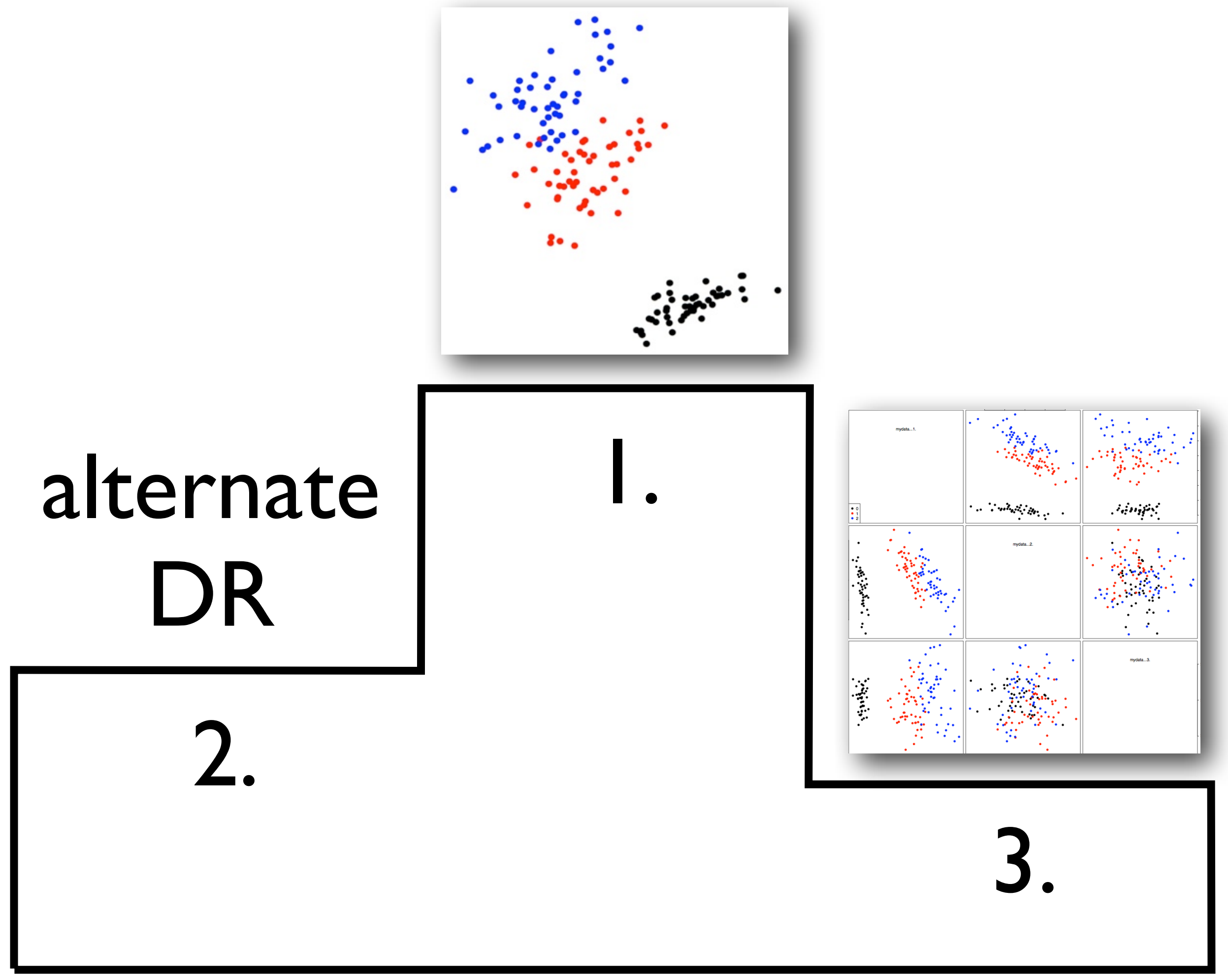
Results



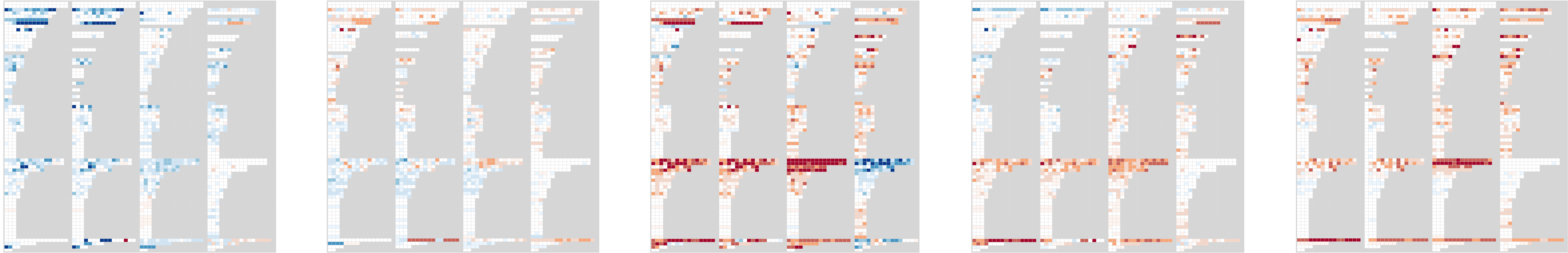
"better" = at least one class is notably more separable in SPLOM or 3D

Implications

- **Use 2D:** 2D often good enough
- **Change DR:** if not, change DR technique
- **Then SPLOM:** SPLOM occasionally helps
- **No 3D:** 3D rarely helps and often hurts



Thanks!



Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, Tamara Munzner, Melanie Tory

contact: michael.sedlmair@univie.ac.at

project page: http://www.cs.ubc.ca/labs/imager/tr/2013/ScatterplotEval/