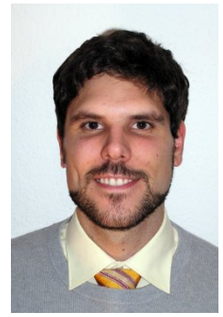
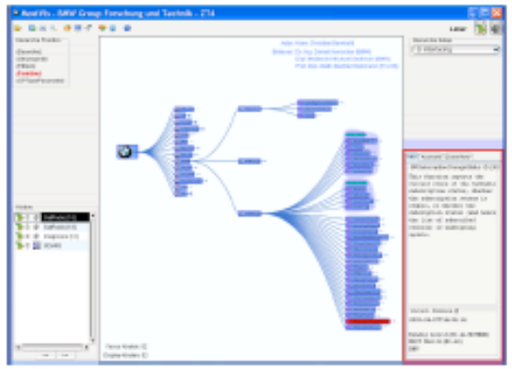
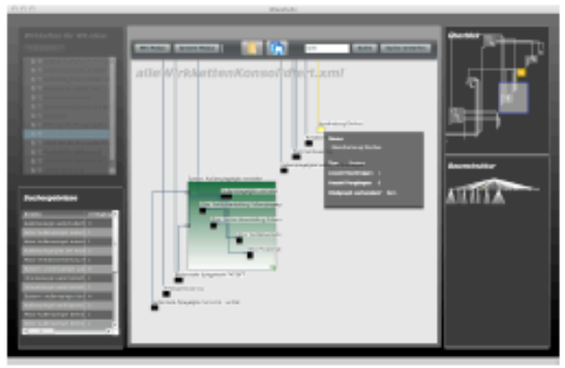
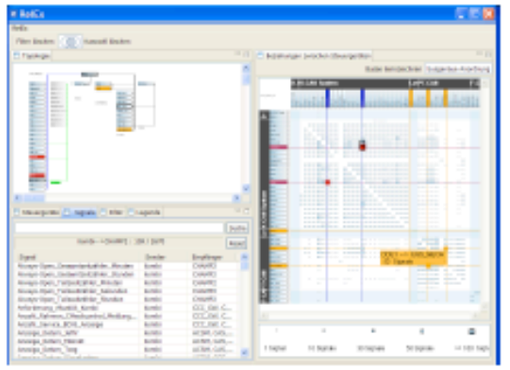
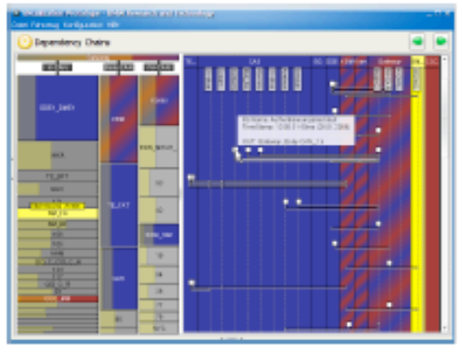
A Taxonomy of Visual Cluster Separation Factors

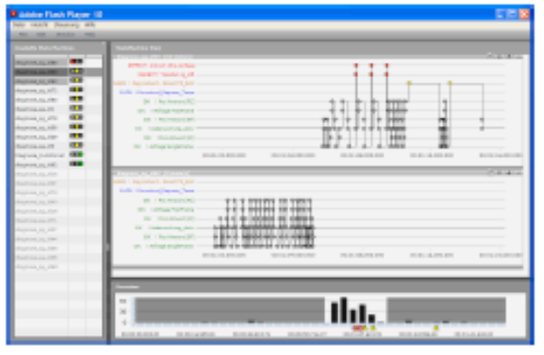
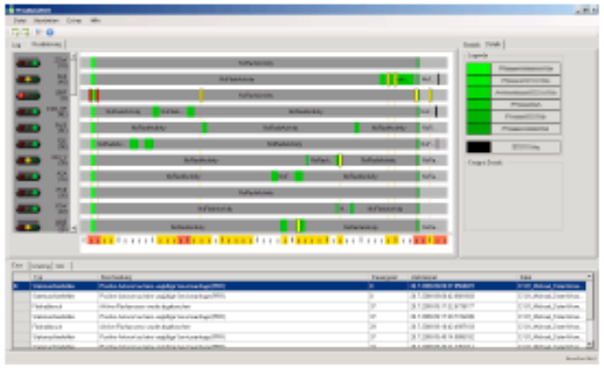
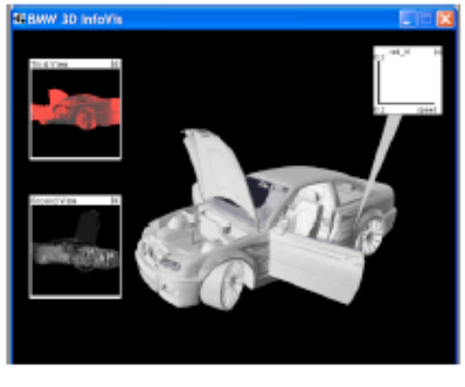
Michael Sedlmair, Andrada Tatu, Tamara Munzner, Melanie Tory
University of British Columbia, Vancouver, CA
http://www.cs.ubc.ca/~msedl

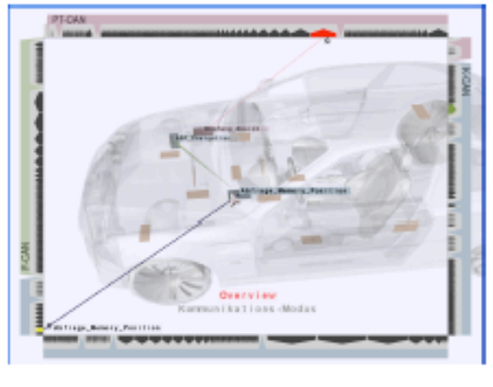


Background

- Phd at LMU, Munich with Andreas Butz: HCI
  - Automotive Design Studies at BMW

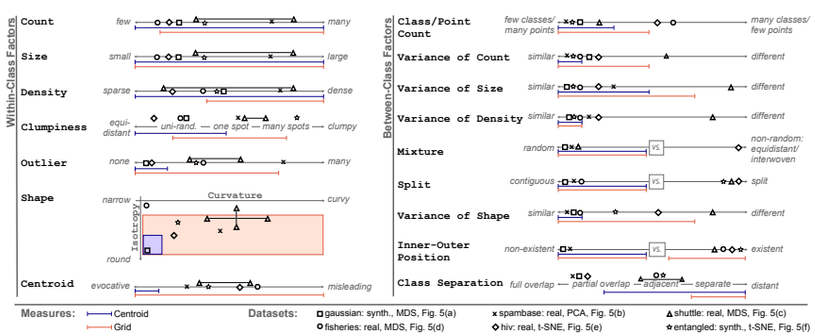






Background

- PostDoc at UBC, Vancouver with Tamara Munzner: InfoVis
  - Highdim projects
    - Ethnographic field study: What people are doing with DR?

- **Data Characterization**



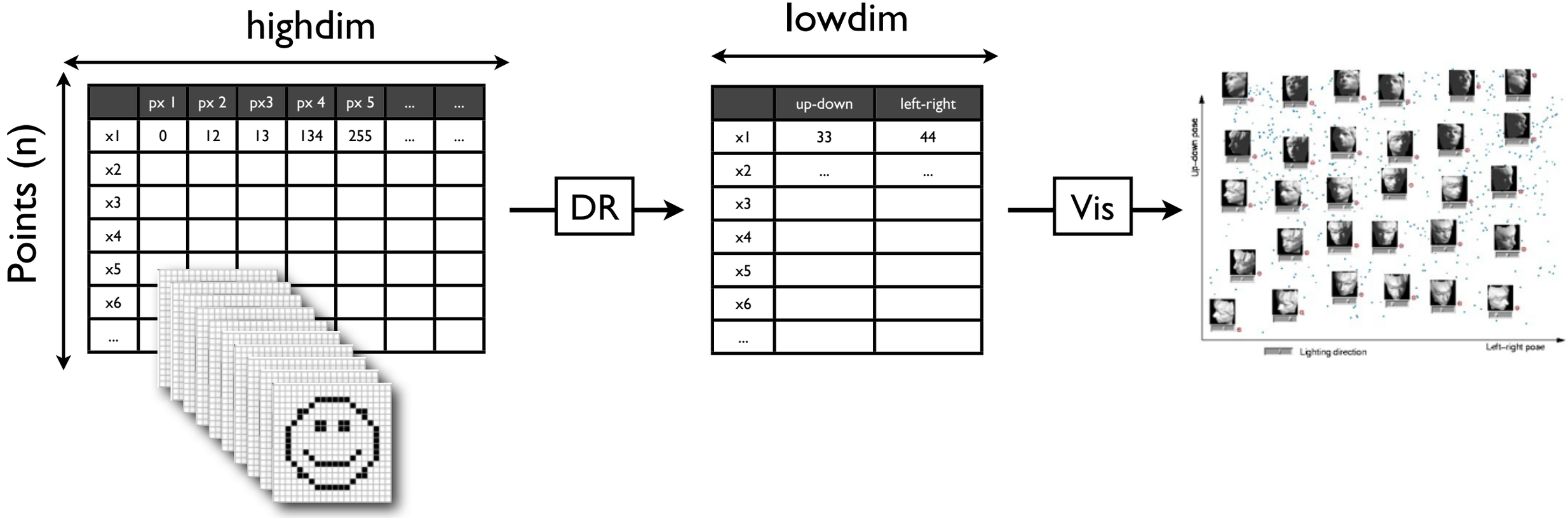
- Design Study Methodology (with Miriah Meyer, University of Utah)

A Taxonomy of Visual Cluster Separation Factors

Motivation

Episode 1:
Once upon a time ...

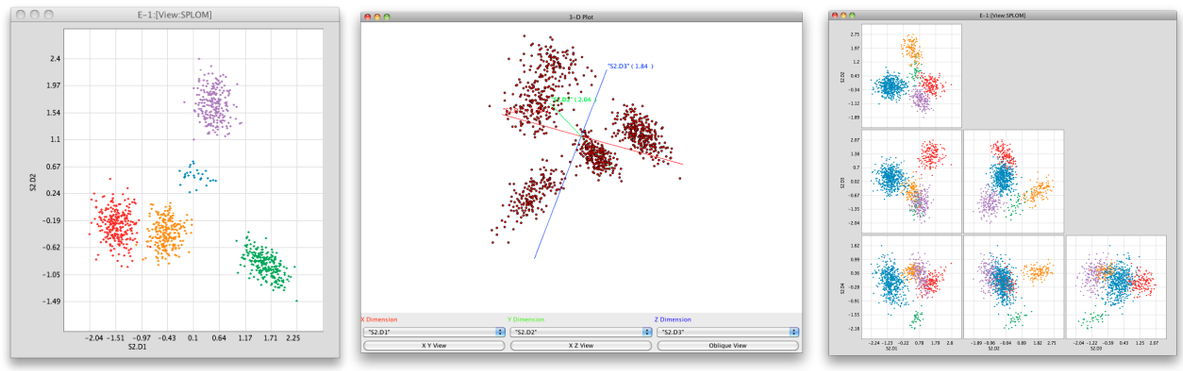
- What *visualization encodings (VisEnc)* are good for visualizing *dimensionally reduced (DR)* data?



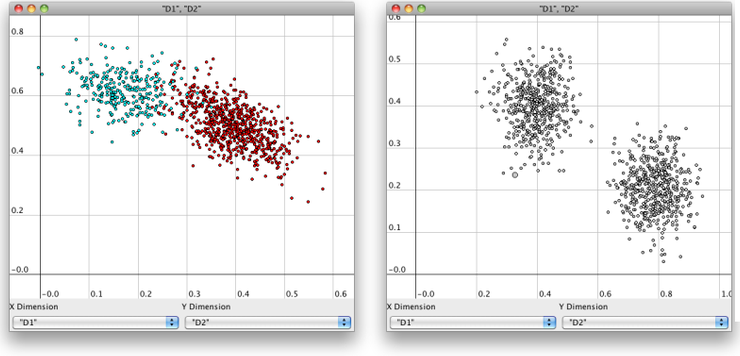
What are people doing?

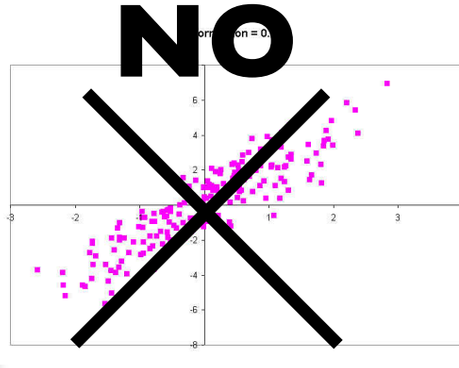
- Informed from ongoing ethnographic field study (not this project!)

- DR techniques
  - PCA, MDS, but also newer ones such as t-SNE



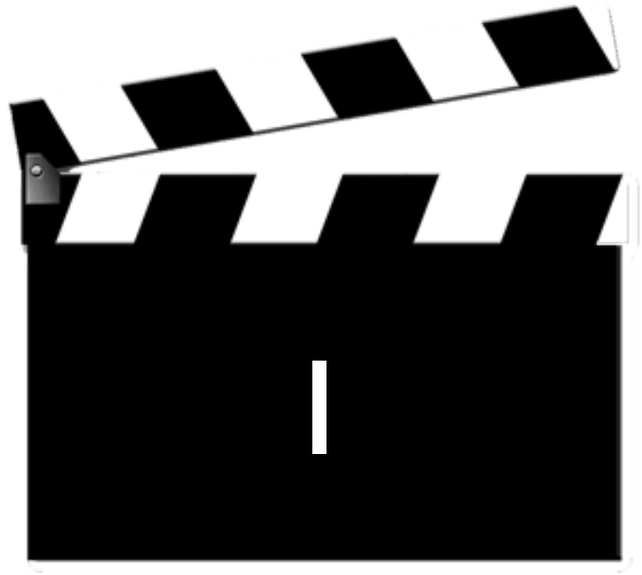
- VisEnc
  - 2d Scatterplot
  - 3d Scatterplot
  - Scatterplot matrices (SPLOMs)





- Tasks:
  - Cluster identification / verification
  - NOT correlation

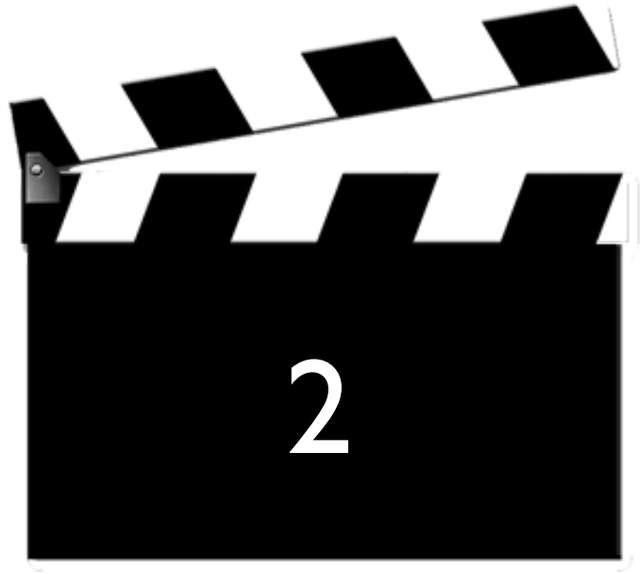
User Study?



- *Idea:*
  - Run a user study
  - Measuring the **human efficiency** (time/error) **in making a judgment** (about clusters)
    - to find out which vis technique works best

- *Started piloting and found ...*
  - What really matters is how the dataset looks like
  - Confounding variable = **dataset characteristic**

Automatic Data Study



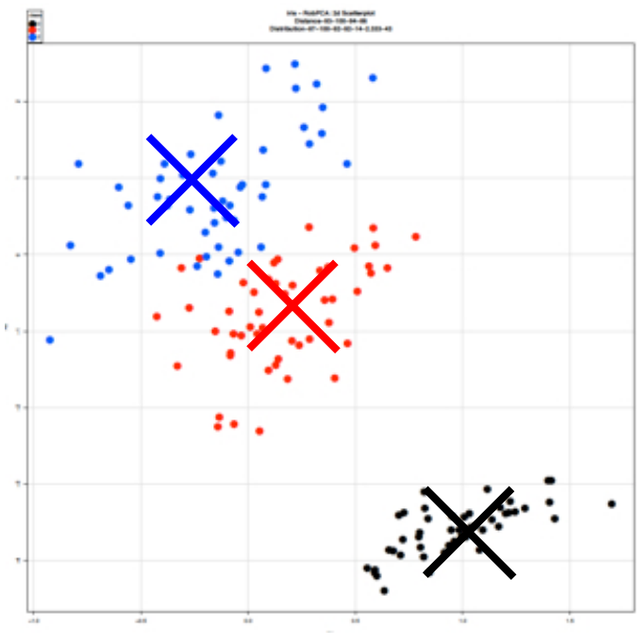
- *Idea:*
  - Take a larger, well-chosen sample of real and synthetic datasets
  - Use recent **cluster separation metrics** to characterize datasets in terms of when which vis technique is best
  - Use **the judgement itself**, where the judgement is done by the metric

Centroid and Grid Metric

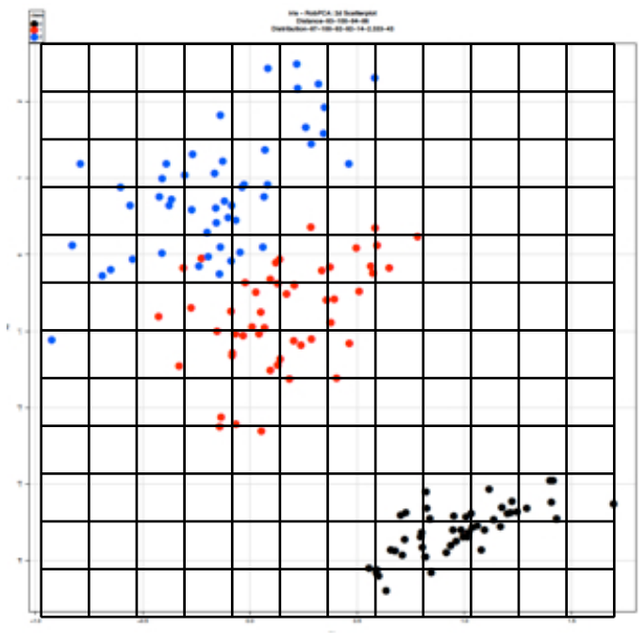
	Centroid Metric	Grid Metric
Sips et al. EuroVis'09	Distance Consistency	Distribution Consistency
Tatu el al. VAST 2009	---	2D Histogram Density Measure (2D-HDM)

- Found to be the current cutting edge [Tatu et al., AVI'10]
- Value between 0 (worst) and 100 (best) for a 2D Scatterplot

iris
RobPCA



Centroid: 93



Grid: 97

But...

- *Again: Started piloting and found ...*
  - The metrics judgment often does not align with human judgement
  - There are hidden assumptions about the data that lead to
    - FN: false negatives (low metric values for sth good)
    - FP: false positives (high metric values for sth poor)
- Reason: **Data characteristics**
- Current metrics cannot be used for a reliable automatic judgment!

Qualitative Data Study

- Changing the route: **Data Characteristics!**

- *Idea:*
  - Do a manual qualitative inspection of datasets
  - Make **human judgement**
  - Compare to metric judgement

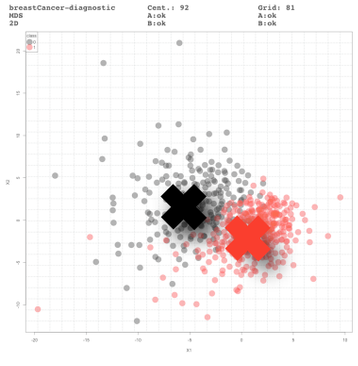
- *Adaption of research question:*
  1. Identifying **data characteristics** that matter for cluster verification
  2. **Evaluating** cluster separation **metrics**

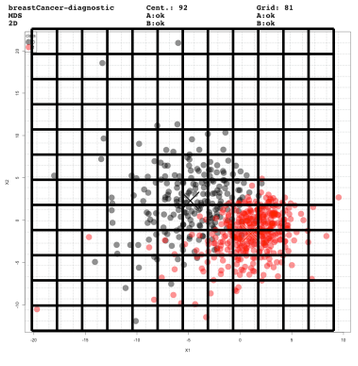
Methods

Methods 1:
Choosing the variables

- **75 datasets**
  - 31 real / 44 synthetic
  - pre-classified / BUT: Keep non-classified in mind!
- **4 DR** techniques: PCA, MDS, RobPCA, t-SNE
- **3 VisEncs:**
  - 2D Scatterplot
  - Interactive 3D Scatterplot
  - SPLOM

- **2 metrics:** centroid and grid
  - Extension: 3D, SPLOM
  - Extension: Classwise





- **816** datasets instances (Scatterplots to look at)

Methods 2:
Open and Axial Coding

- Coding = Method for qualitative data analysis from social science
- Open Coding = Figuring out codes and iteratively refine *codeset*
- Axial Coding = Analyze relation between codes and categorize *codeset*

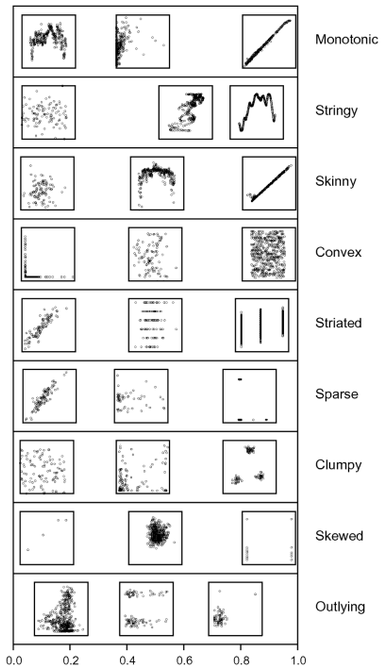
- **Open Coding**
  - Two researchers (Andrada and me): Multiple passes over 816 dataset instances
  - Coding: Factors affecting visual class separability
  - Coding: Failure cases of metrics:
    - ok - dubious - poor / classwise poor
    - if not ok: reason

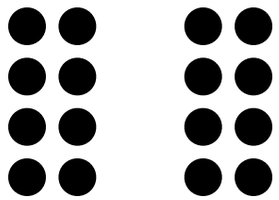
- **Axial Coding**
  - Merging codesets
  - Create and refine taxonomy of factors that matter

Qualitative Data Study:
A side contribution of our project

- Qualitative studies (coding) have been used for
  - User analysis: Audio, video, notes, ...
  - Literature analysis
- We: Data analysis
- In Vis: Qualitative Data Study as an **inverse** to a User Study
  - User Study:
    Many users, few datasets
  - Qualitative Data Study:
    Few users, many datasets

Related Work





They

- **Scagnositcs [Wilkinson 2005]**
  - General factors
  - Informal exploration
  - Mathematical depiction
- **Gestalt principles**
  - Perceptual fundamentals
- **Clustering (ML)**
  - Automation

We

- Cluster separability
- Systematic and rigorous
- Human perception
- Operational guidance
- Human judgment

Taxonomy

A taxonomy of visual cluster separation factors

Influence: Scale > Point Distance > Shape > Position

Variance →

Within-Class Factors

Count
few → many

Size
small → large

Density
sparse → dense

Clumpiness
equidistant | uniformly random | one dense spot | many dense spots | clumpy

Outlier
none ← → many

Shape
narrow → curvy (Curvature)
round (Isotropy)

Centroid
evocative ← → misleading

Between-Class Factors

Class/Point Count
few classes many points → many classes few points

Variance of Count
similar → different

Variance of Size
similar → different

Variance of Density
similar → different

Mixture
random | equidistant | interwoven (VS. VS.)

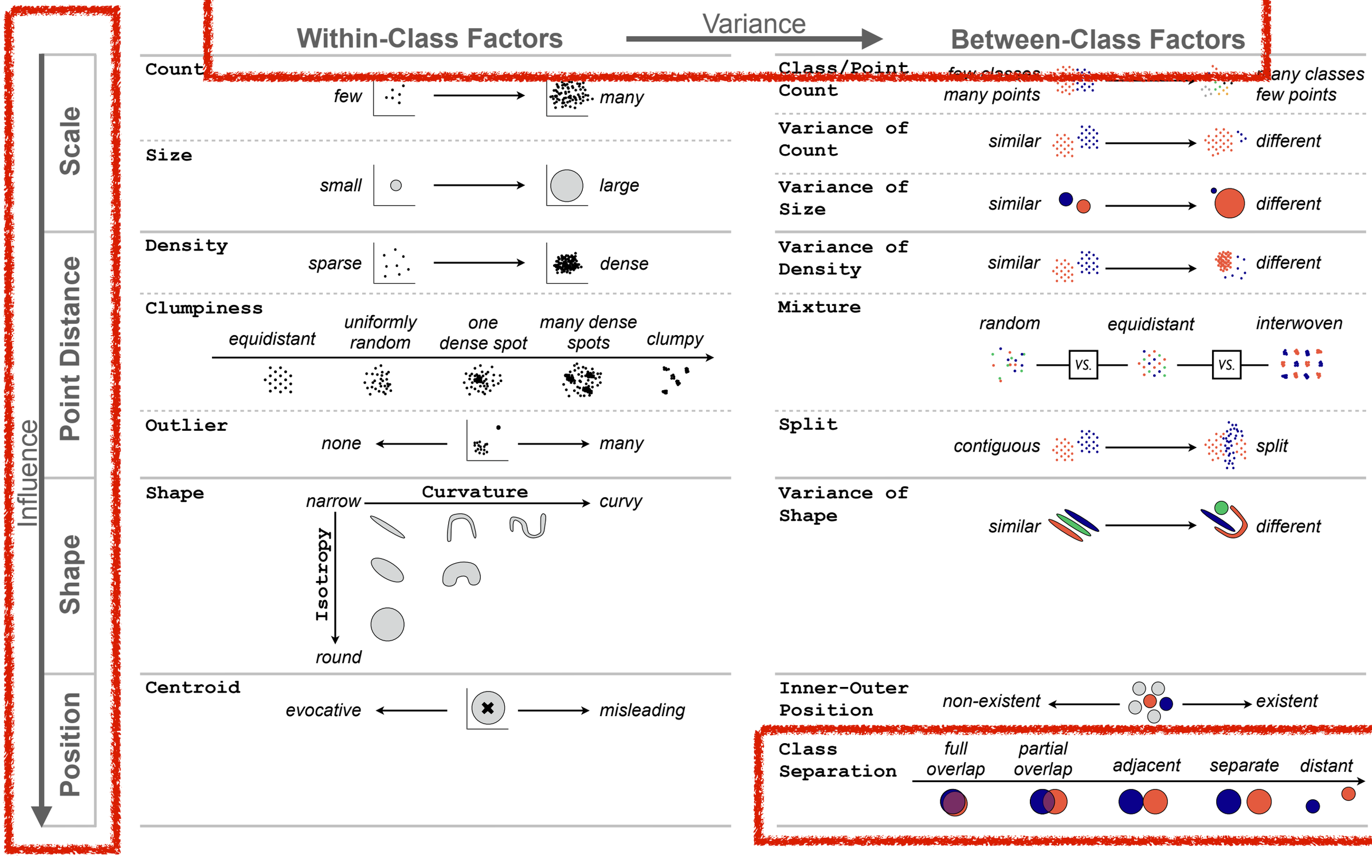
Split
contiguous → split

Variance of Shape
similar → different

Inner-Outer Position
non-existent ← → existent

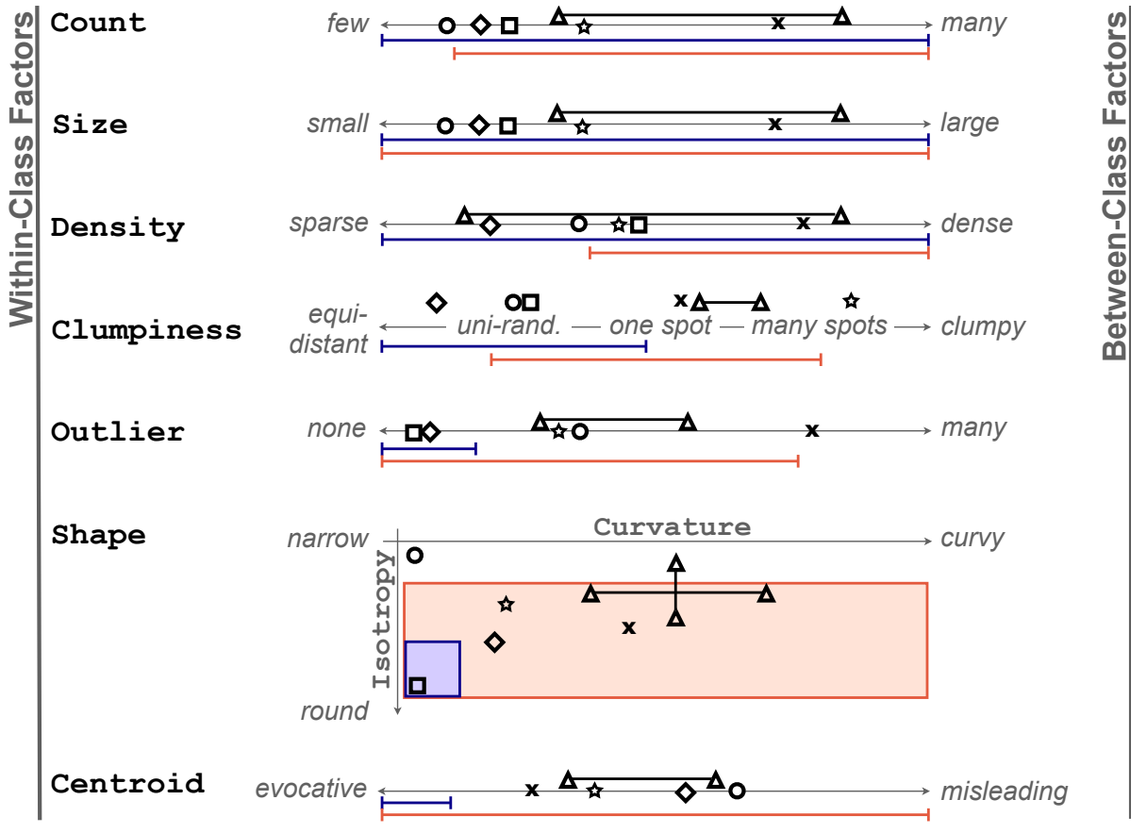
Class Separation
full overlap | partial overlap | adjacent | separate | distant

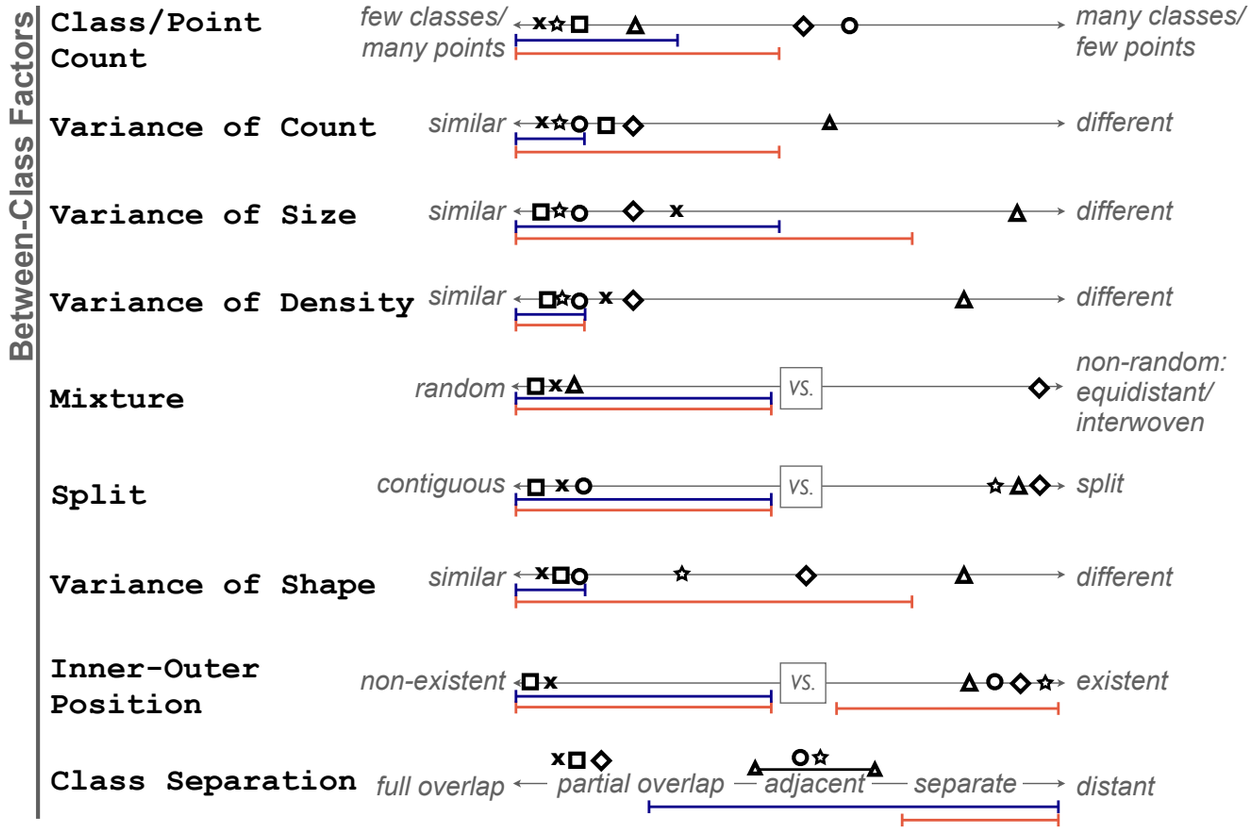
Top-level structure



What can we do with it:

- Mapping **assumptions** of metrics onto taxonomy axes
- Mapping **datasets** onto taxonomy axes



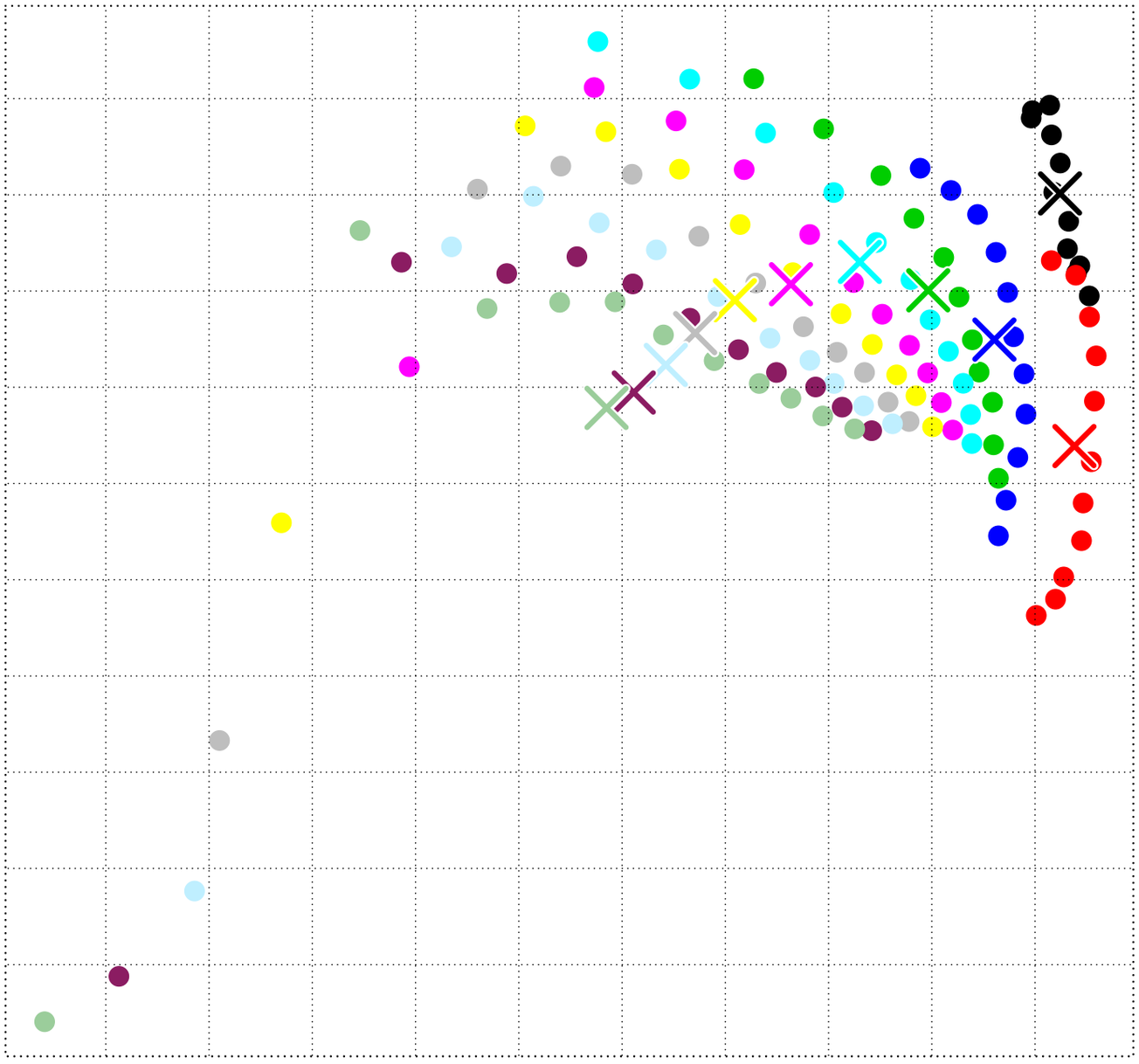


Measures: Centroid, Grid

Datasets: gaussian: synth., MDS, Fig. 5(a); fisheries: real, MDS, Fig. 5(d); spambase: real, PCA, Fig. 5(b); hiv: real, t-SNE, Fig. 5(e); shuttle: real, MDS, Fig. 5(c); entangled: synth., t-SNE, Fig. 5(f)

Examples

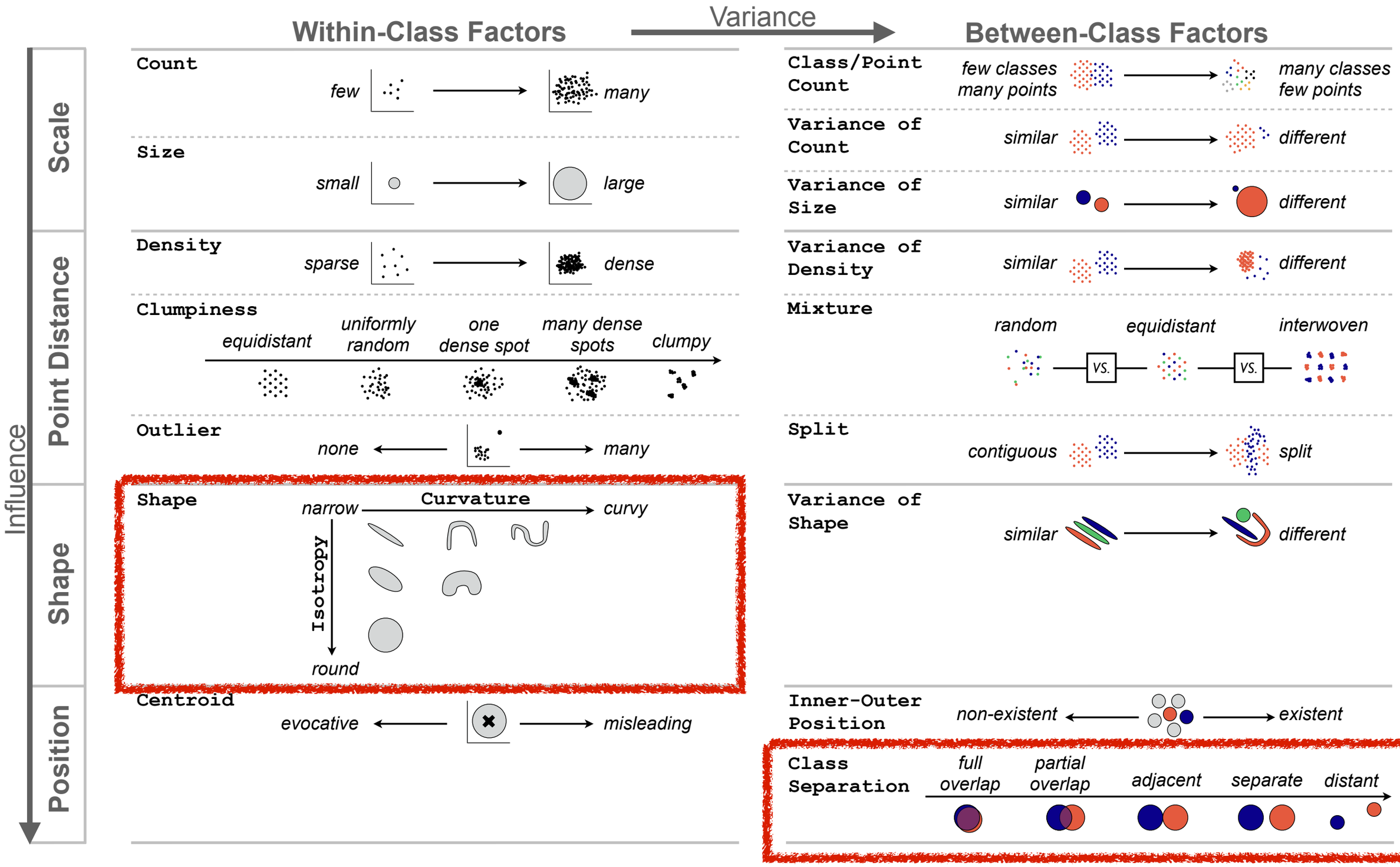
Fisheries, real, MDS



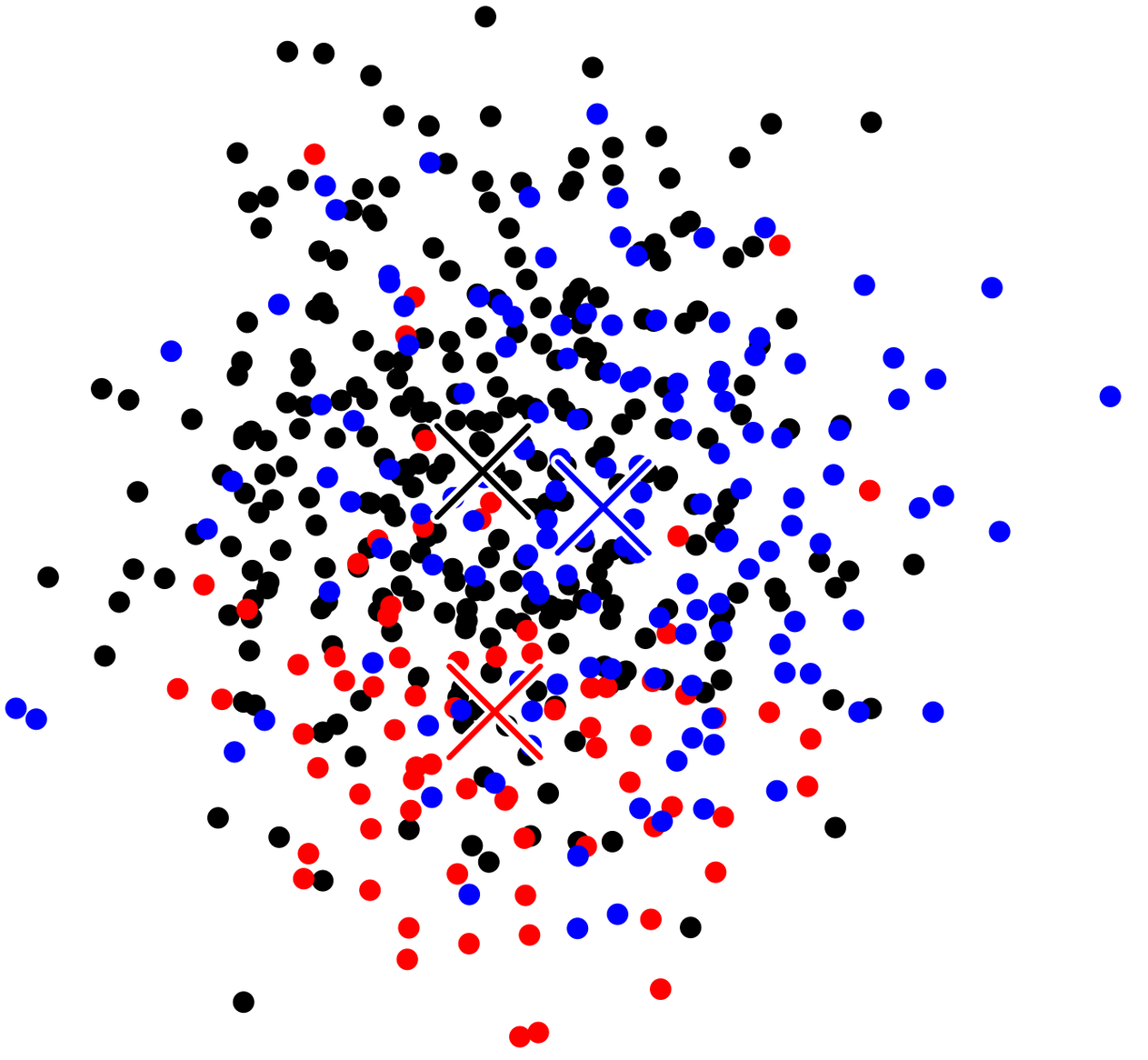
Centroid:
Overall: **29**
Problem: **FN**
Reason: **“Stringy” shapes**

Grid:
Black & Red: **~70-80**
Others: **~40-50**
Problem: **FN**
Reason: **Adjacent strings**

In terms of taxonomy ...



Gaussian, synthetic, MDS



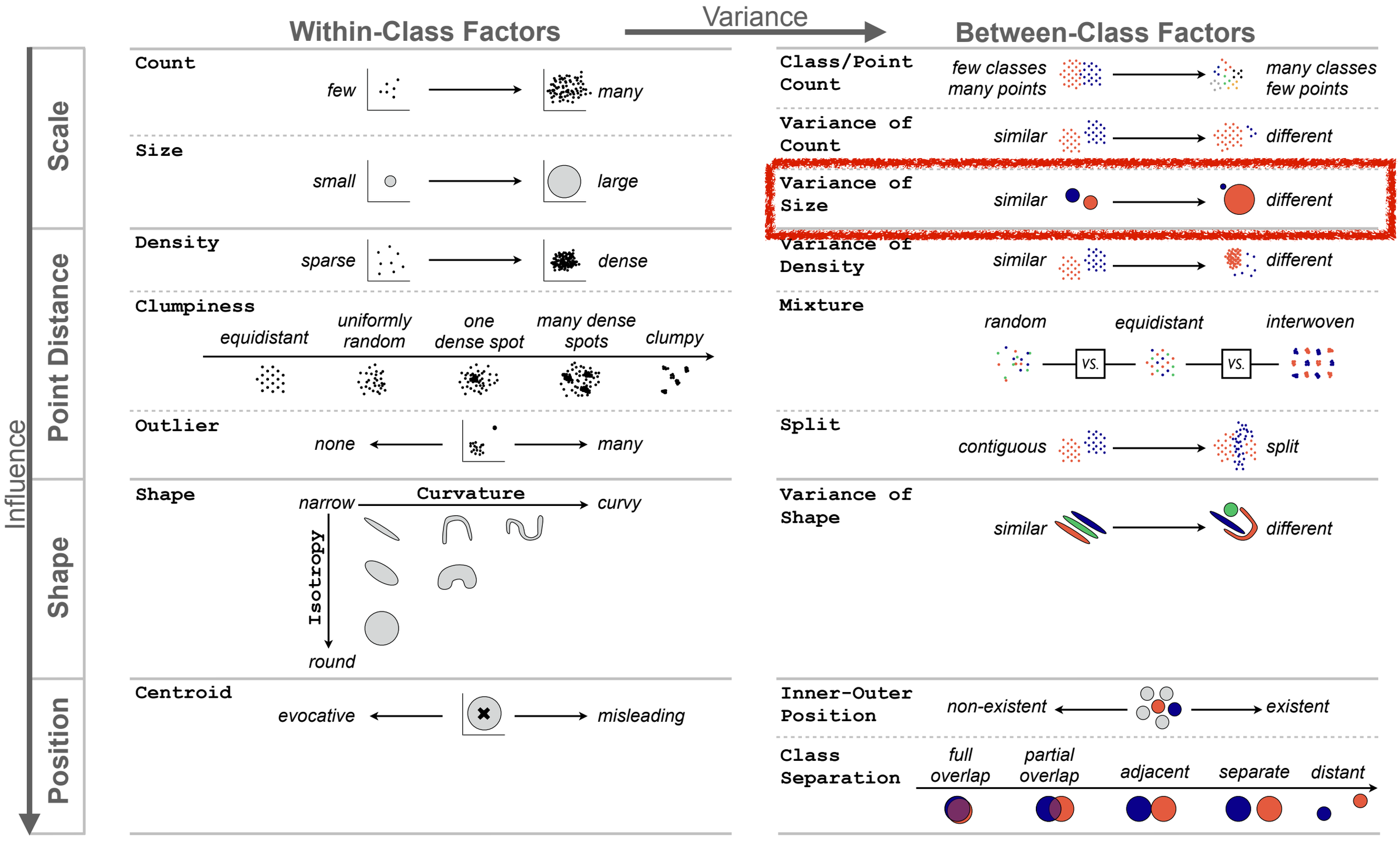
Centroid:

Red: **77**

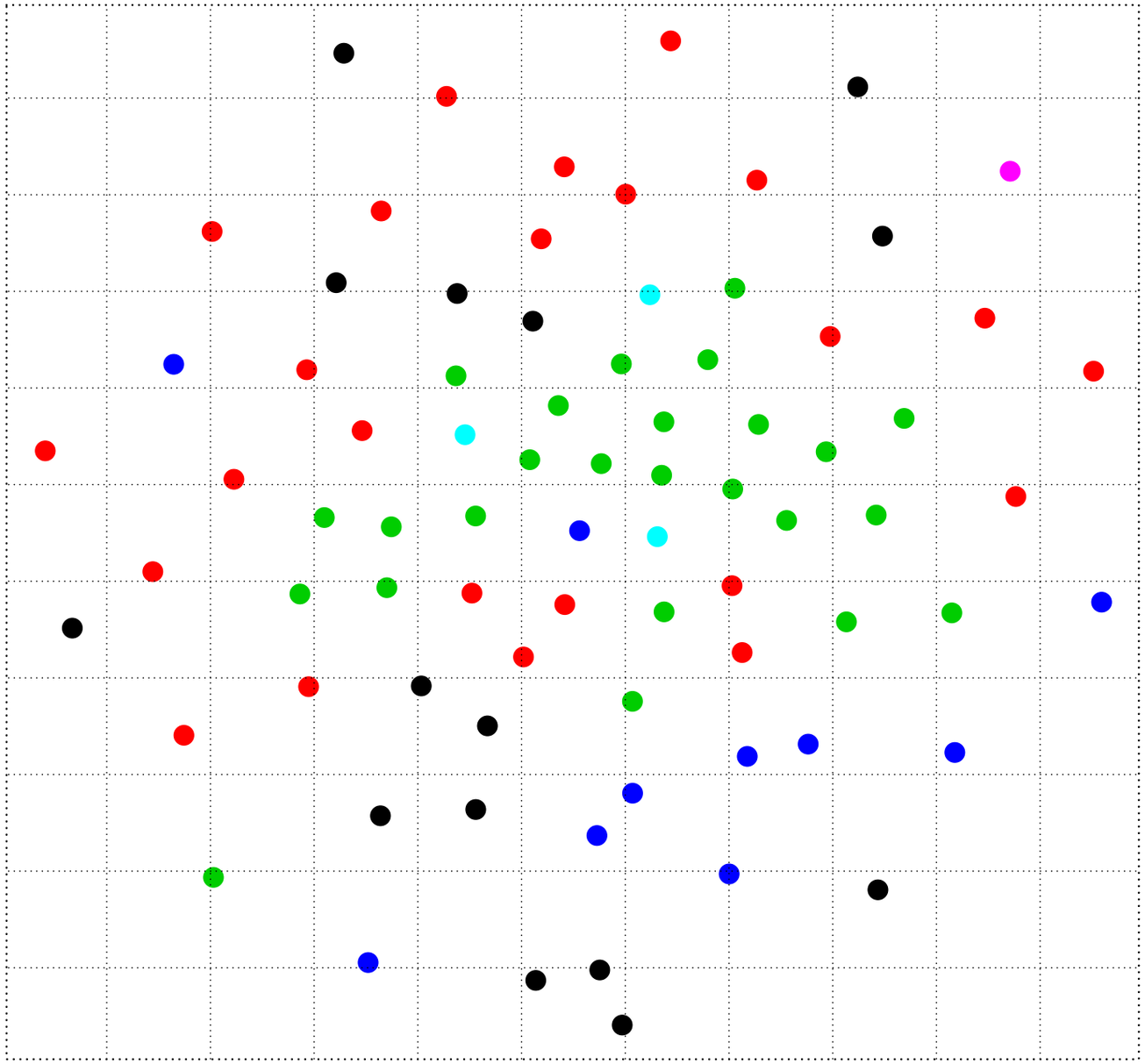
Problem: **FP**

Reason: **Bigger classes overspread smaller one**

In terms of taxonomy ...

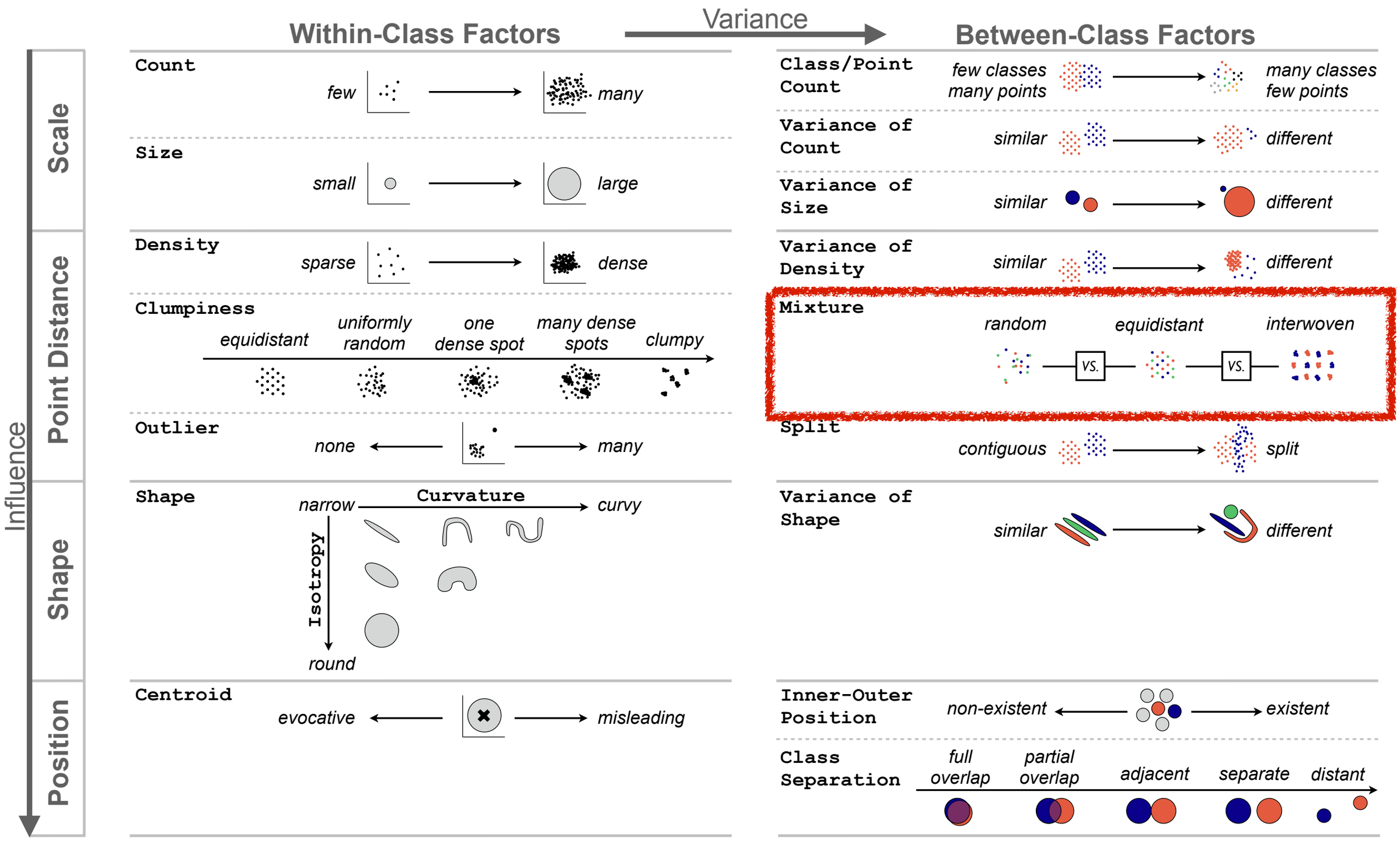


HIV, real, t-SNE



Grid:
Overall: **99**
Problem: **FP**
Reason: **“Equidistant” points**

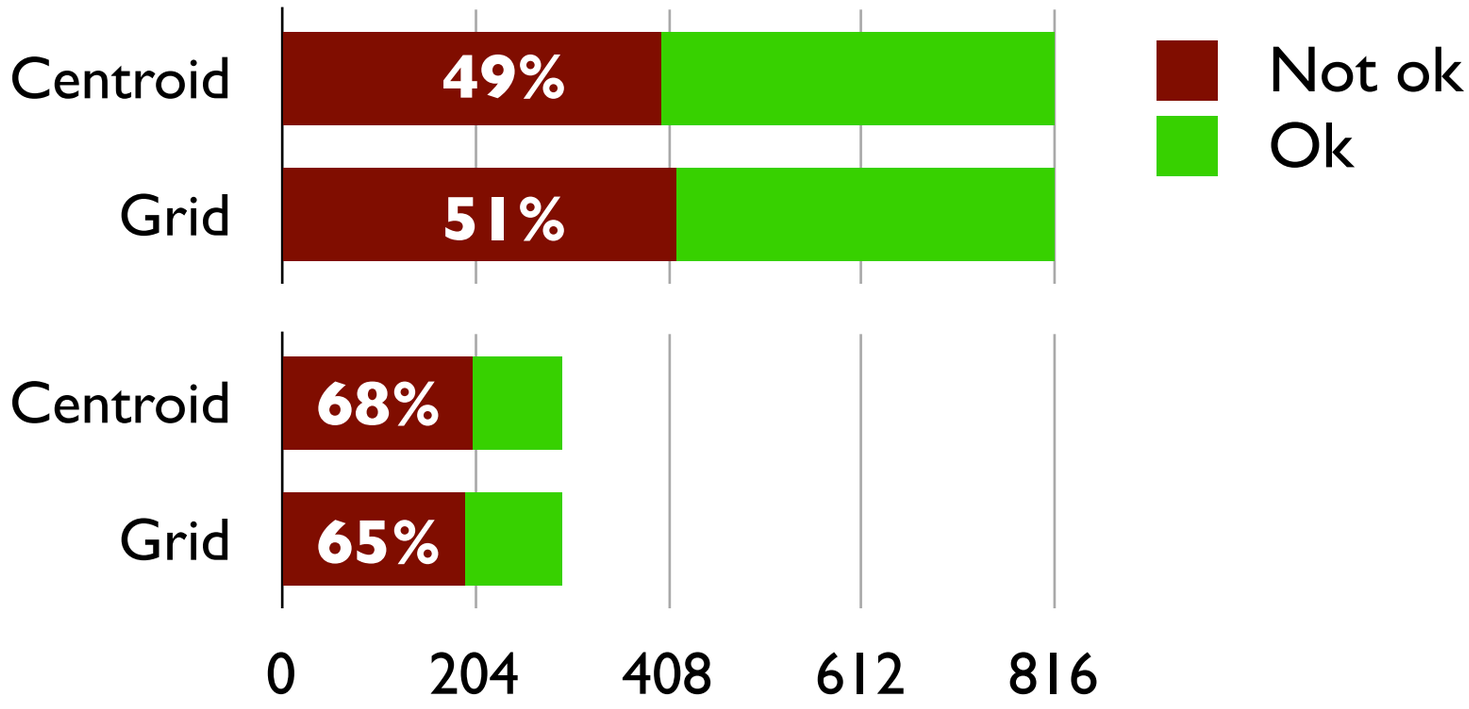
In terms of taxonomy ...



Summary of results

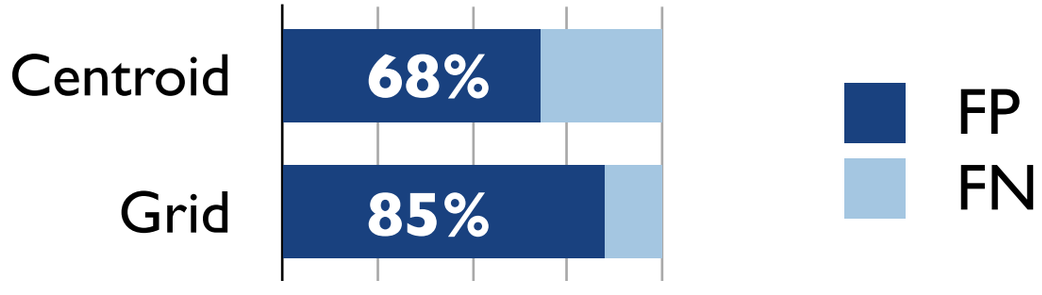
High-level results

All (816)



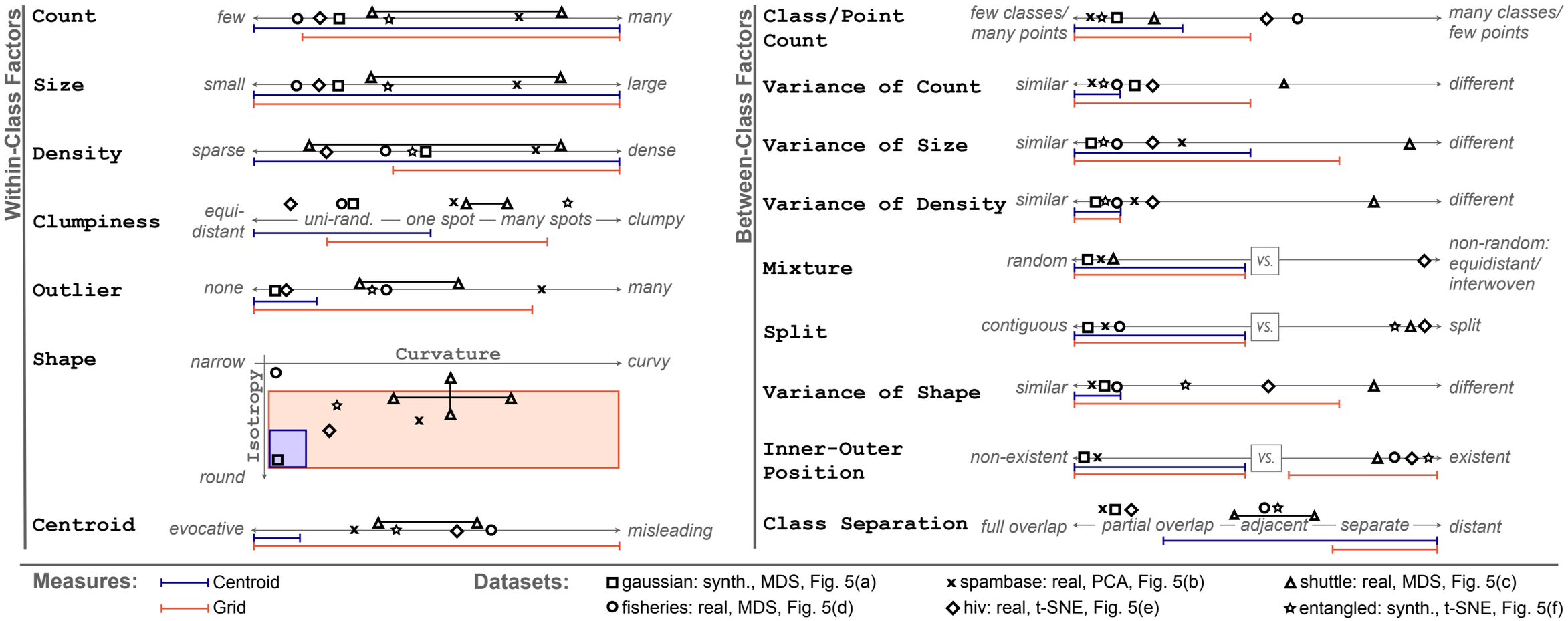
Only real (296)

FP / FN (All failure cases)



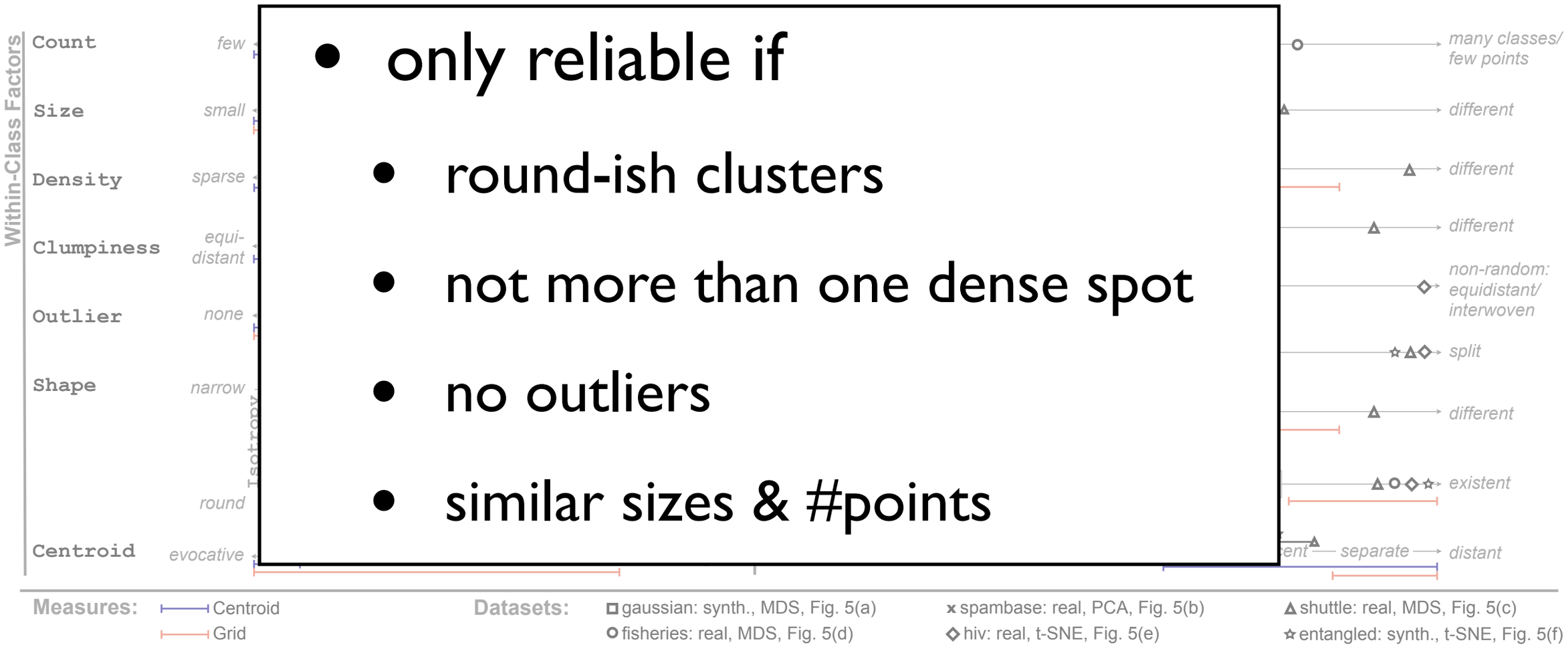
Summary:

Mapping assumptions onto taxonomy axes



Centroid:

Mapping assumptions onto taxonomy axes

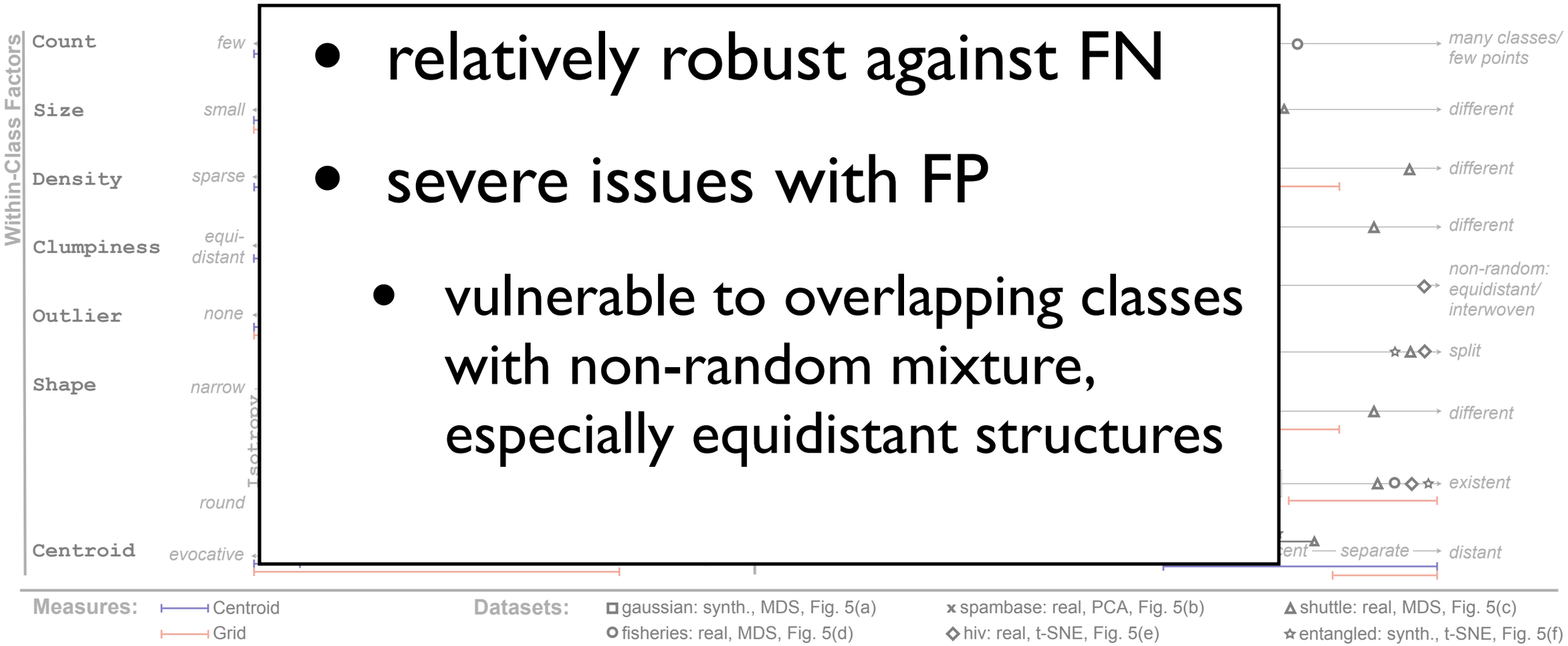


- only reliable if
  - round-ish clusters
  - not more than one dense spot
  - no outliers
  - similar sizes & #points

Grid:

Mapping assumptions onto taxonomy axes

- relatively robust against FN
- severe issues with FP
  - vulnerable to overlapping classes with non-random mixture, especially equidistant structures



Conclusion

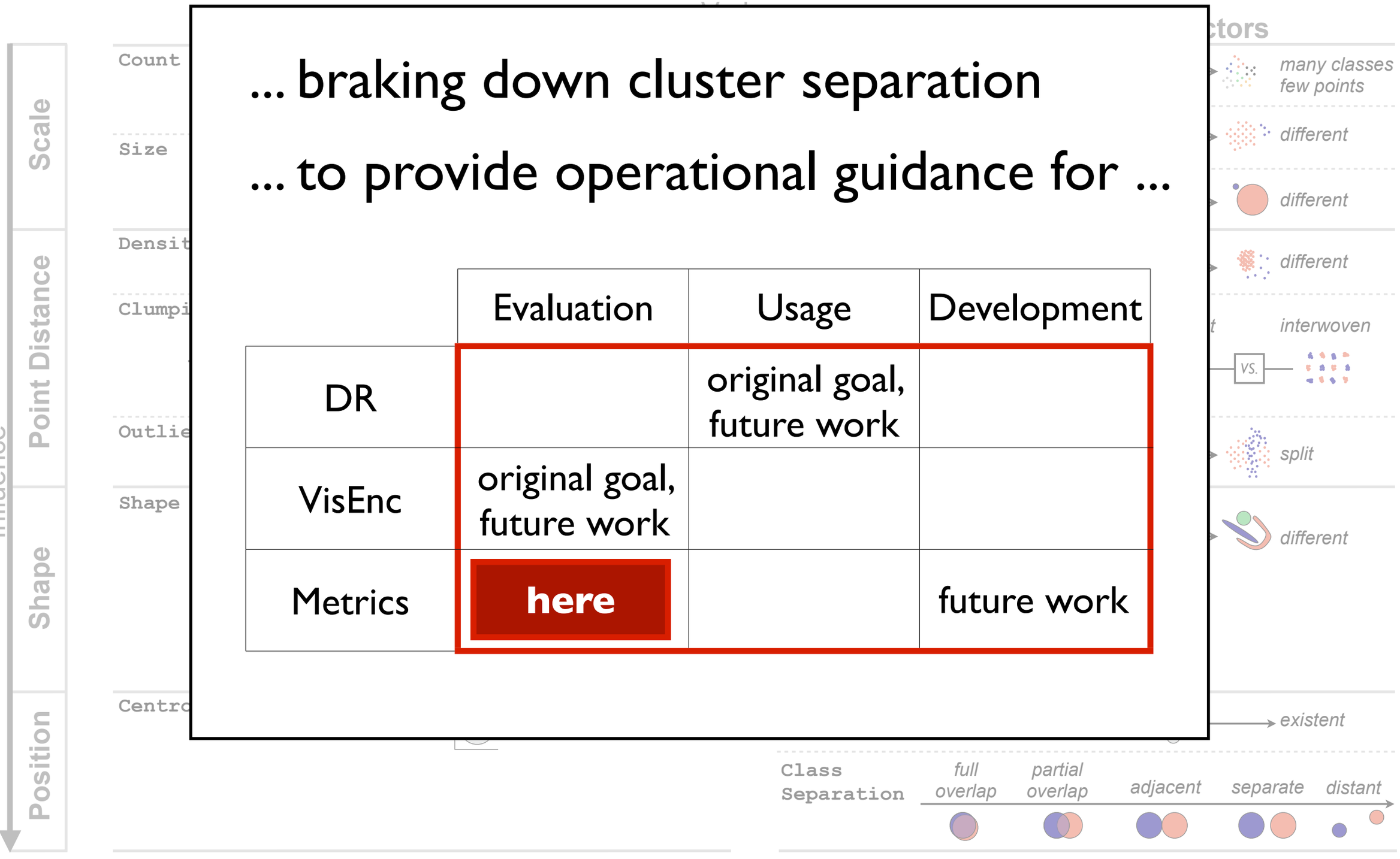
Main Contribution

A Taxonomy of Visual Cluster Separation Factors

... braking down cluster separation

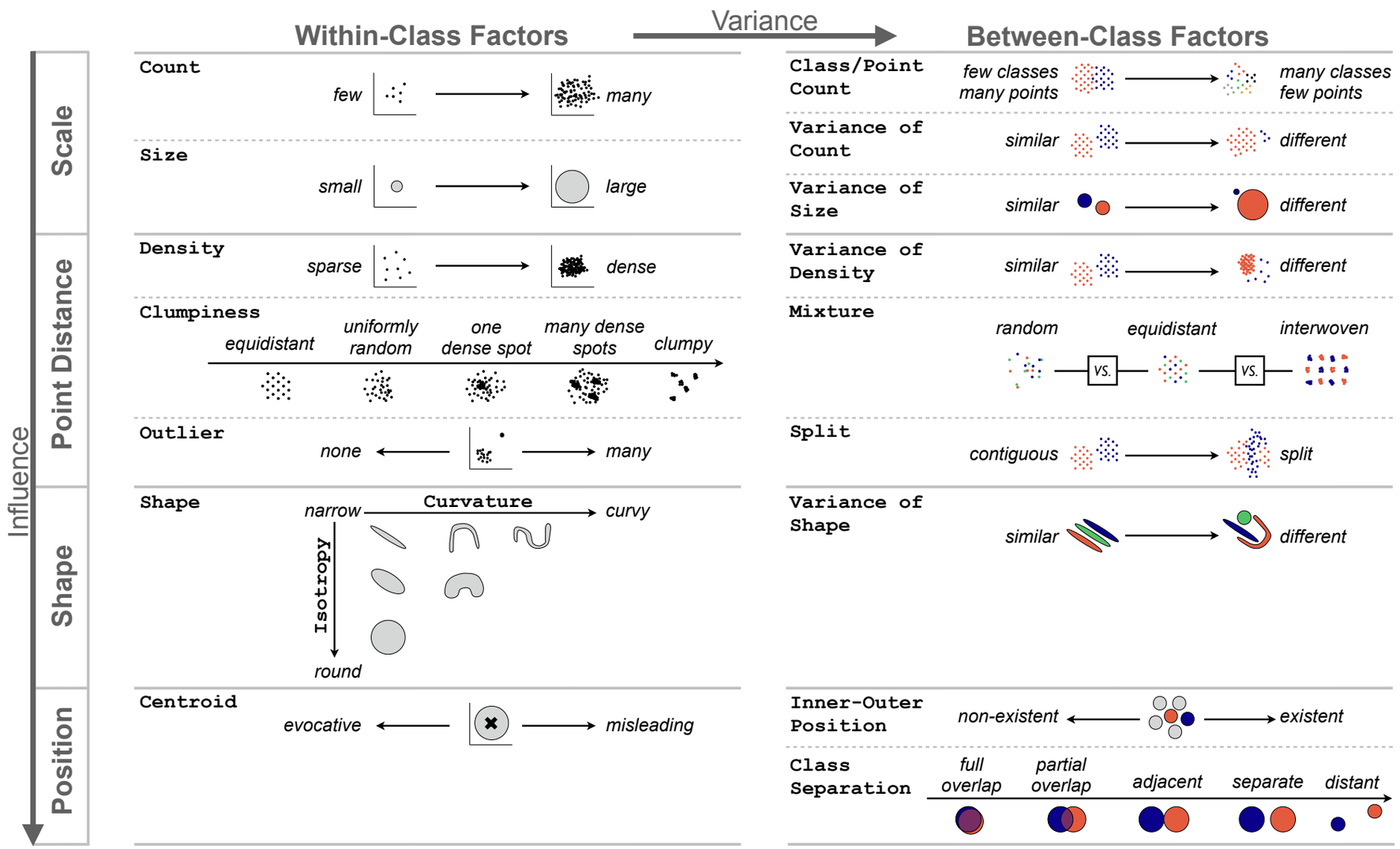
... to provide operational guidance for ...

	Evaluation	Usage	Development
DR		original goal, future work	
VisEnc	original goal, future work		
Metrics	**here**		future work



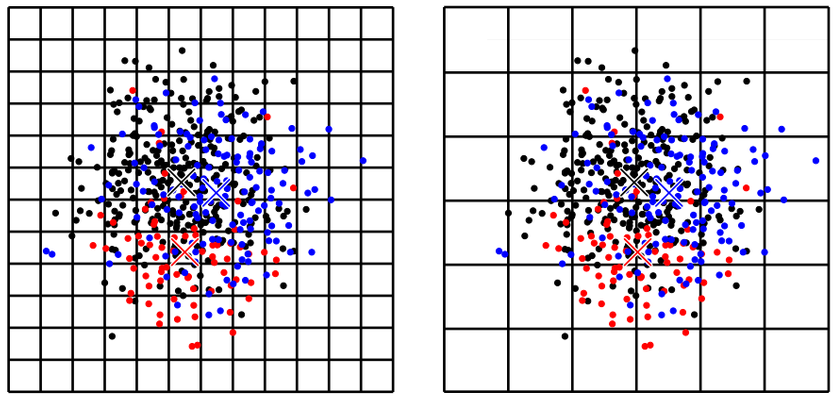
A Taxonomy of Visual Cluster Separation Factors

Michael Sedlmair, Andrada Tatu, Tamara Munzner, Melanie Tory
University of British Columbia, Vancouver, CA
Slides: http://www.cs.ubc.ca/~msedl/talks/konstanz2012.pdf



Appendix

Adopting the Grid Size?



- Developers [Sips'09]: "insensitive to grid size changes"
  - No: Test with 50 instances - insensitive only in 16%
- Static: Does not work well with different #points
- Dynamic: No straight forward automatic way