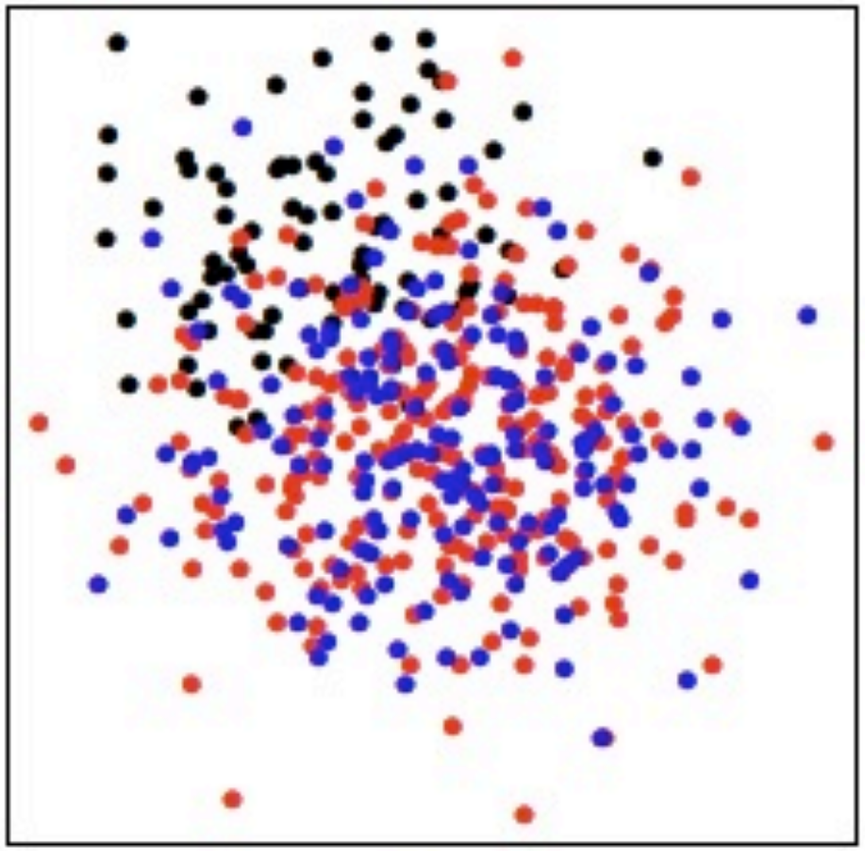
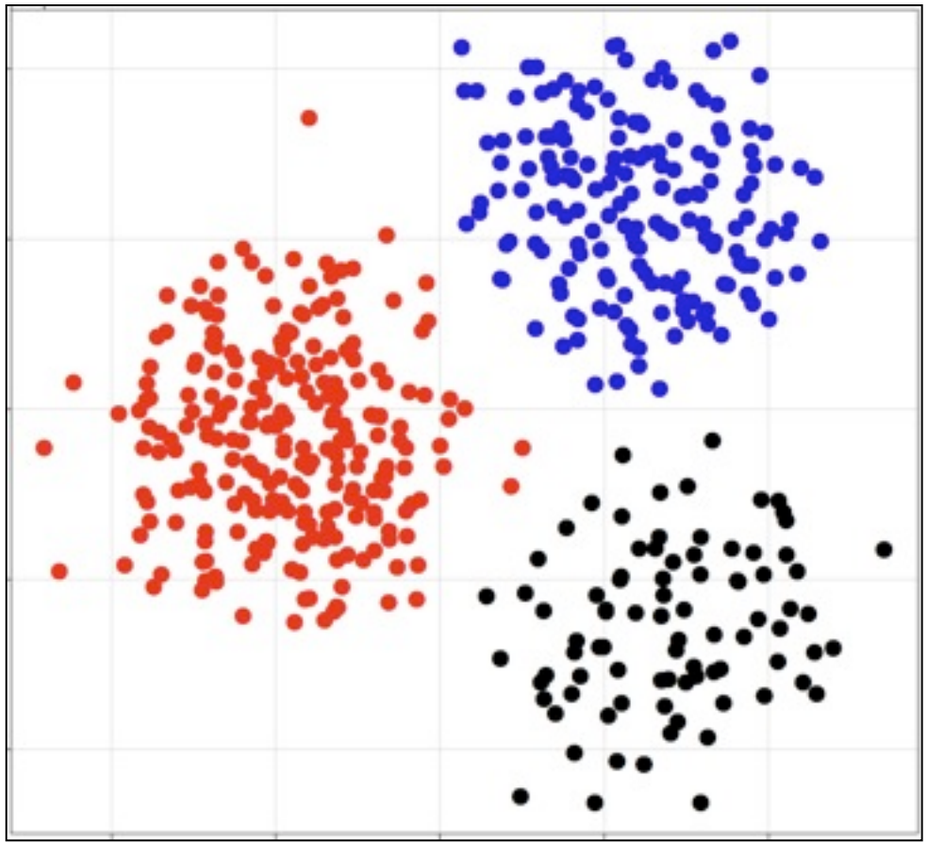
A Taxonomy of Visual Cluster Separation Factors

Michael Sedlmair,[1] Andrada Tatu[2], Tamara Munzner[1], Melanie Tory[3]
[1] Univ. of British Columbia, [2] Univ. of Konstanz, [3] Univ. of Victoria

Visual Cluster Separation



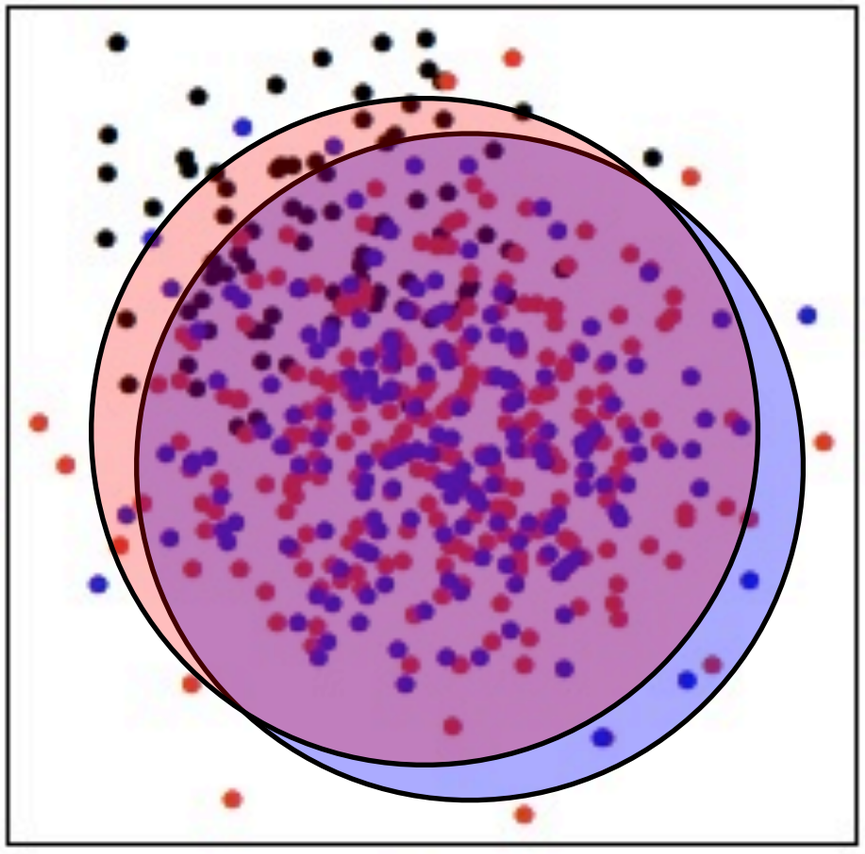
Bad!

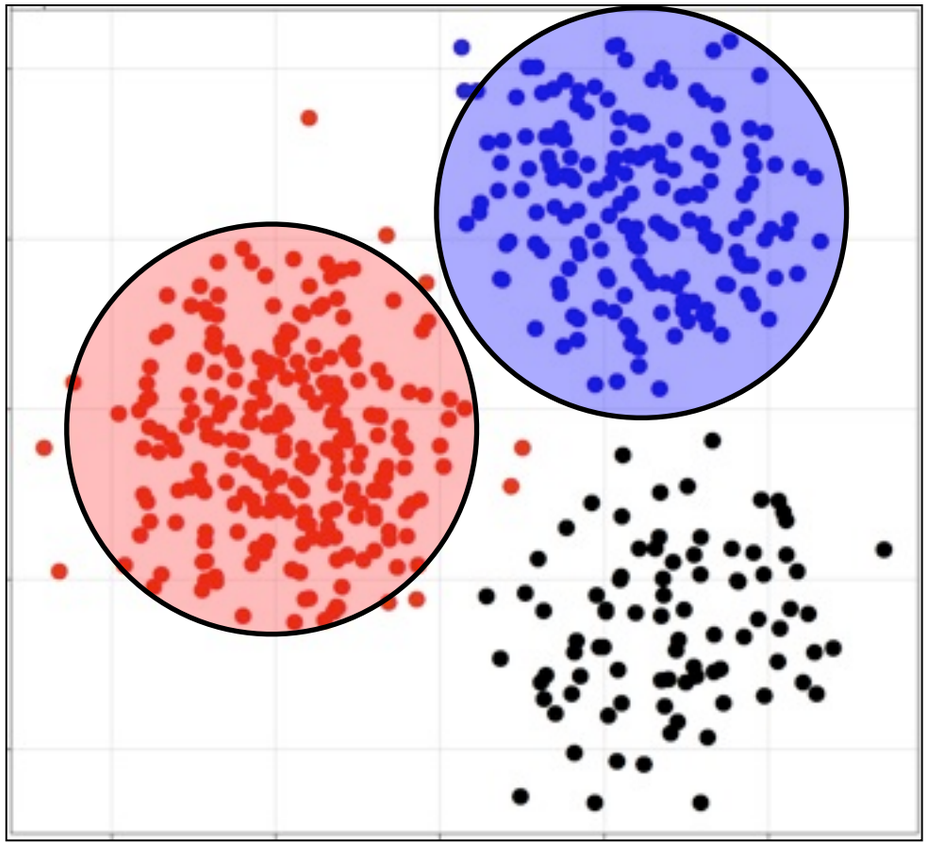


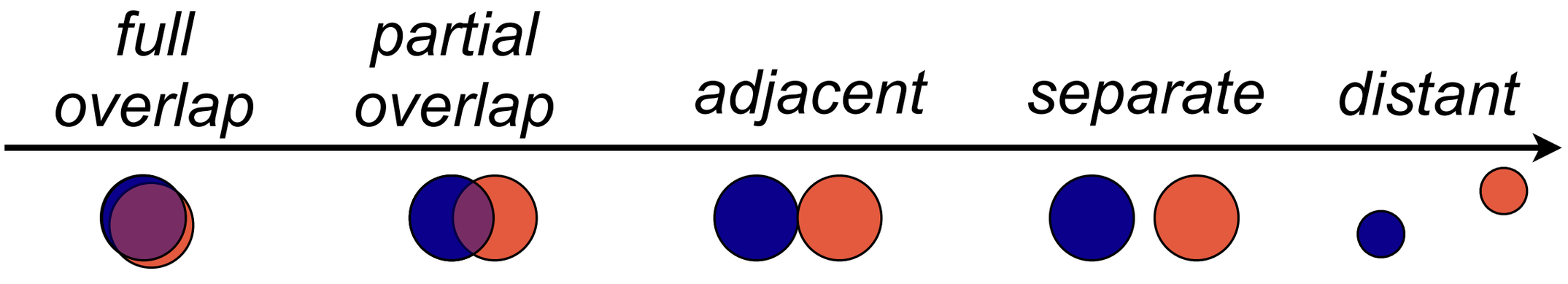
Good!

Cluster Separation:

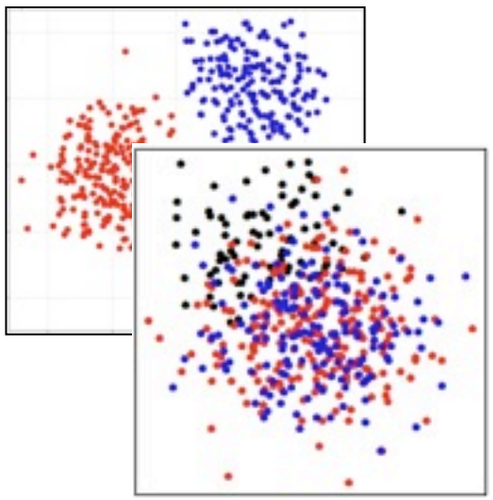
Simple Idea

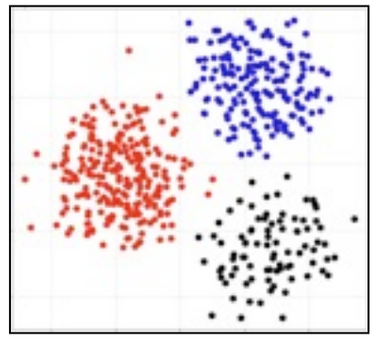






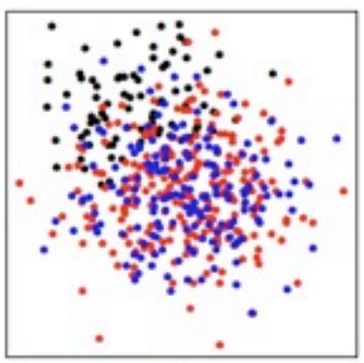
Automatic Cluster Separation





Good!

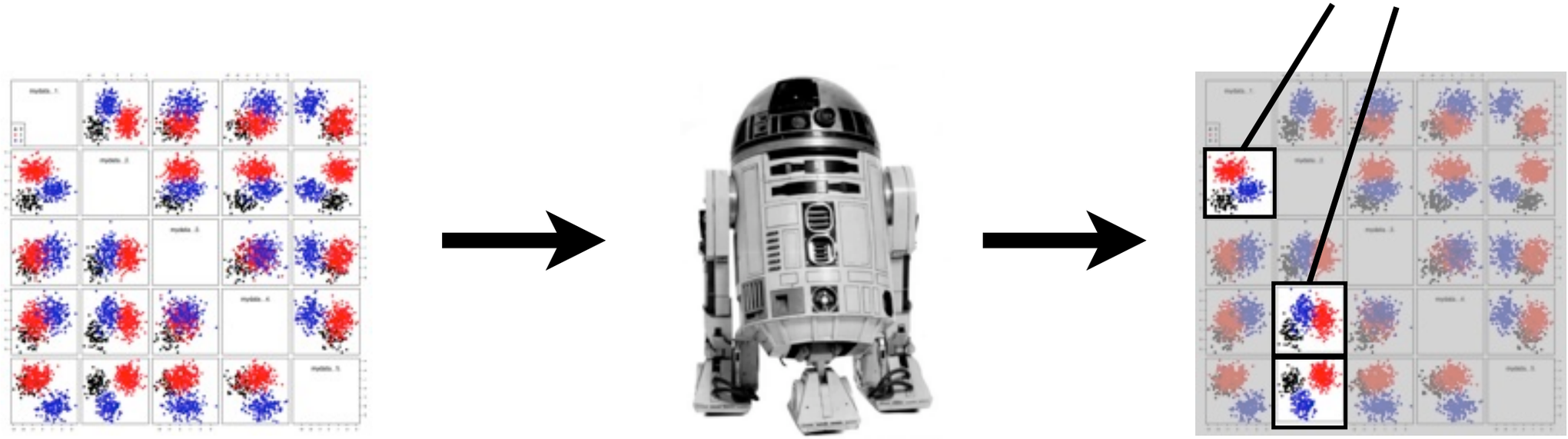
Bad!



Automatic Cluster Separation

- Many **cluster separation measures** proposed recently*
- For semi-automatic guidance in high-dim data analysis

Good!

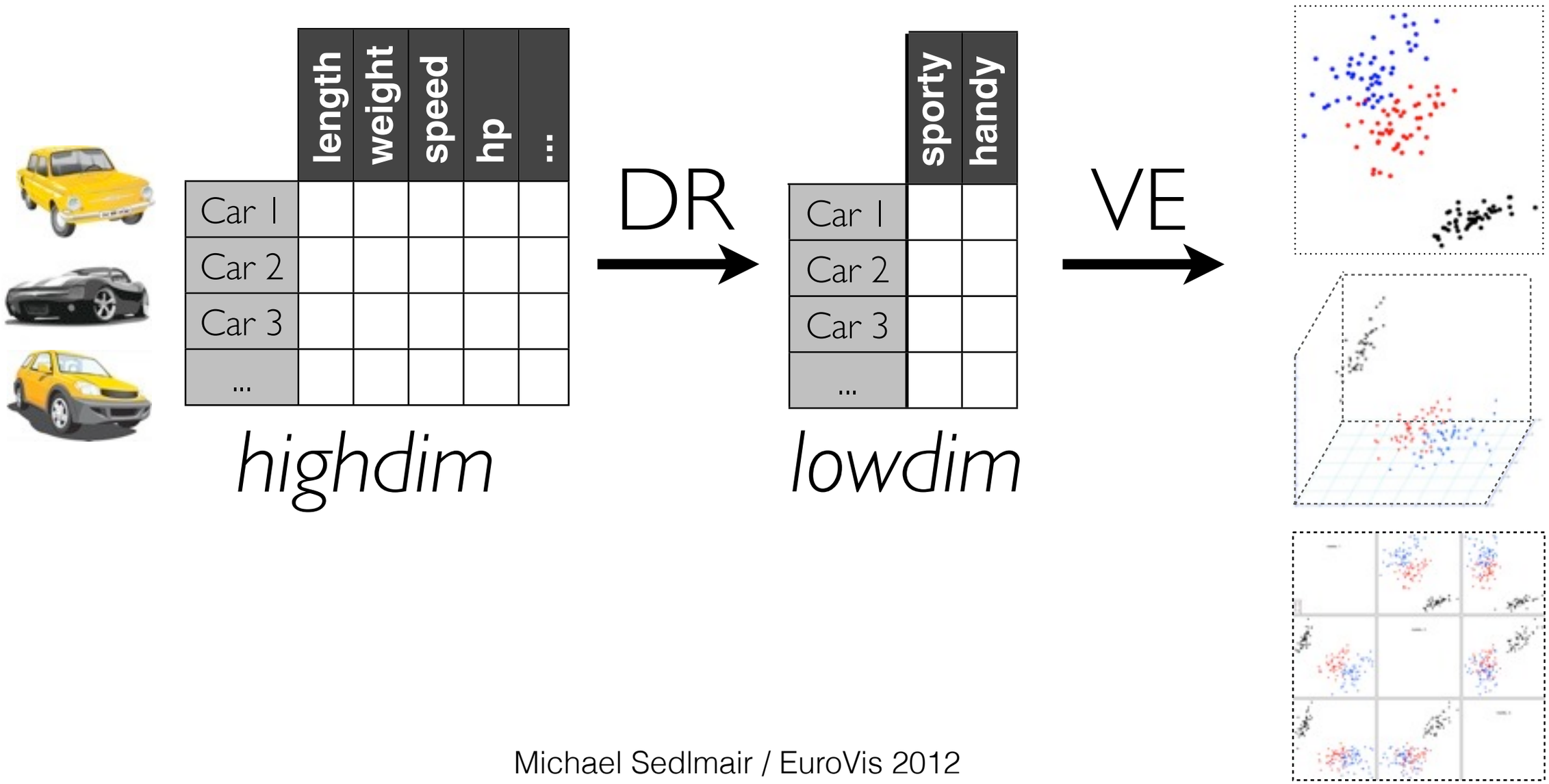


* Sips et al.: Selecting good views of high-dimensional data using class consistency [EuroVis 2009]
* Tatu et al.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data [VAST 2009]

Our original intention:

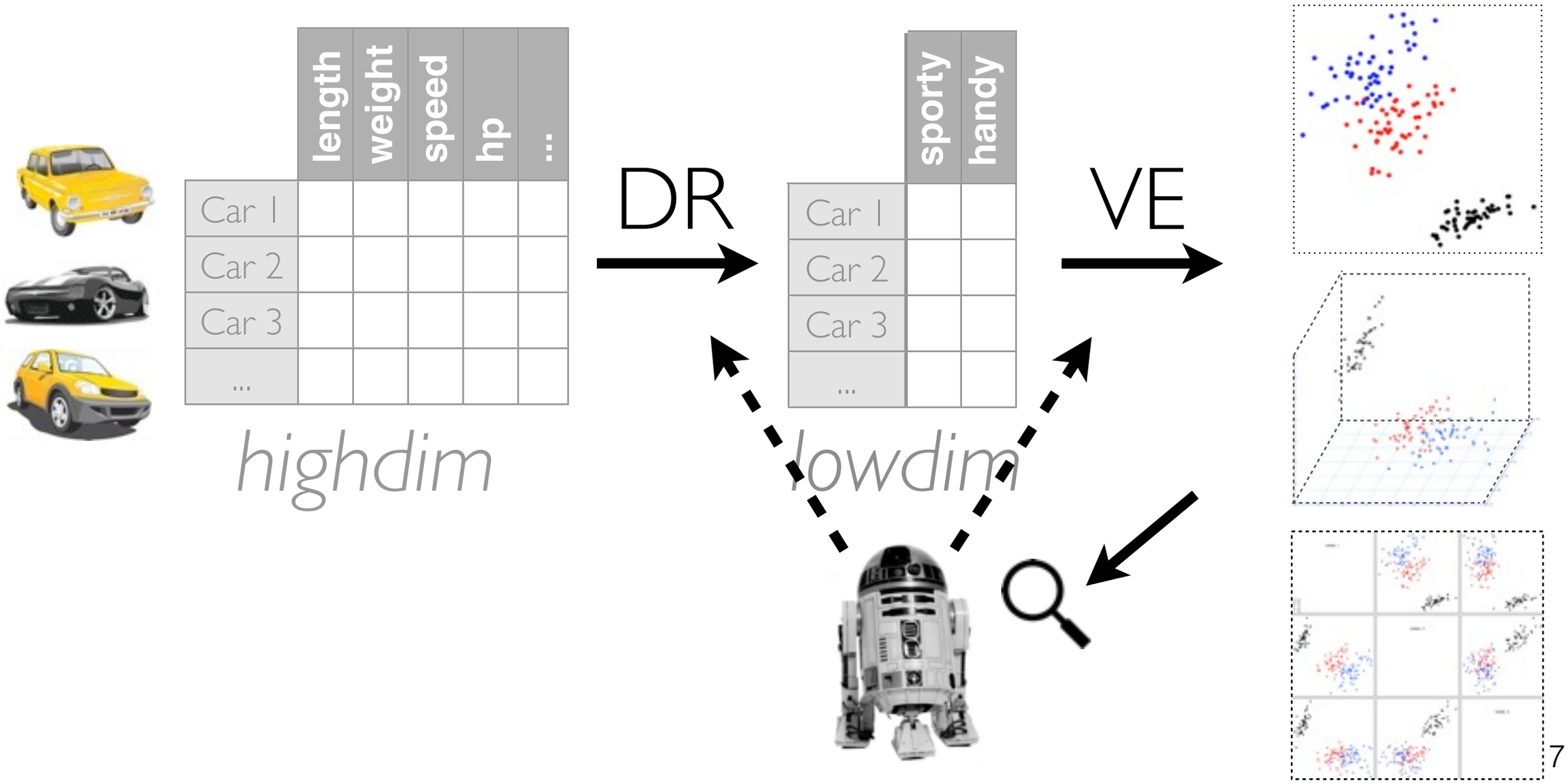
DR and VE guidance

- *DR = Dimension Reduction*: PCA, MDS, ...
- *VE = Visual Encoding:* Scatterplots (2D, 3D, SPLOM)

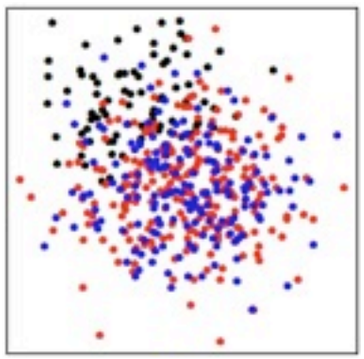


Our original intention:
DR and VE guidance

- *DR = Dimension Reduction*: PCA, MDS, ...
- *VE = Visual Encoding:* Scatterplots (2D, 3D, SPLOM)



Automatic vs. Human?

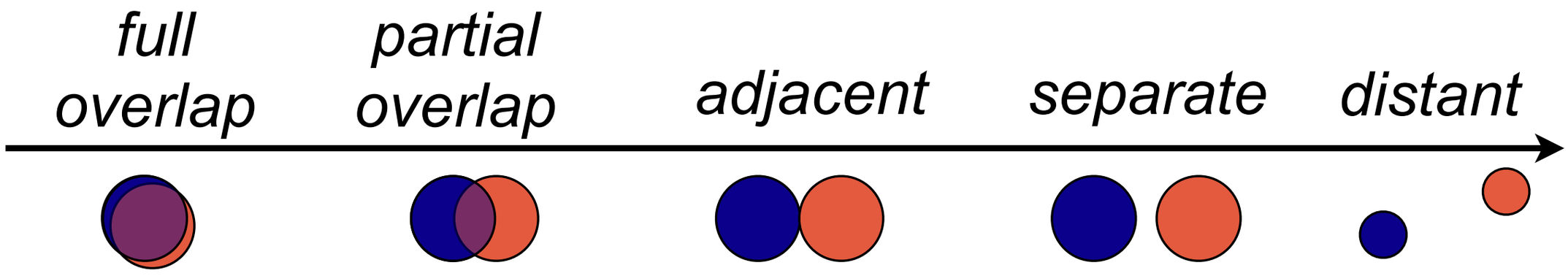


Good!

No! Huh?

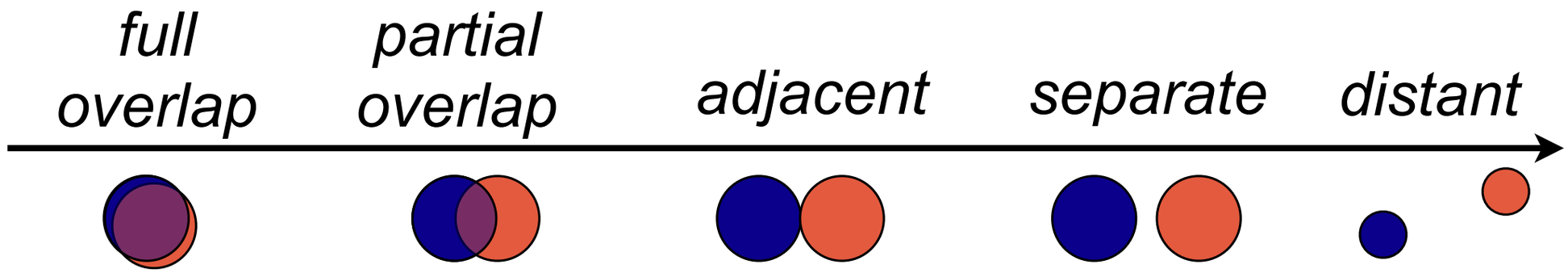
Cluster Separation:

Simple Idea - Is this enough?



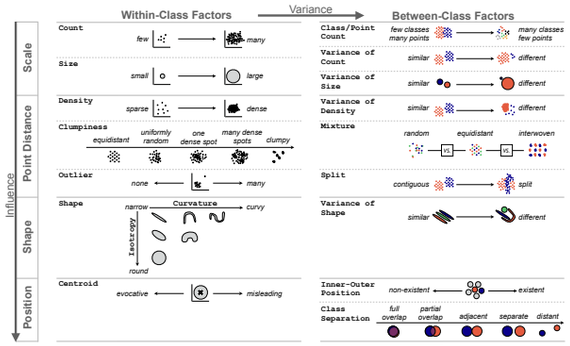
Our Goals

- What factors matter in human cluster perception?

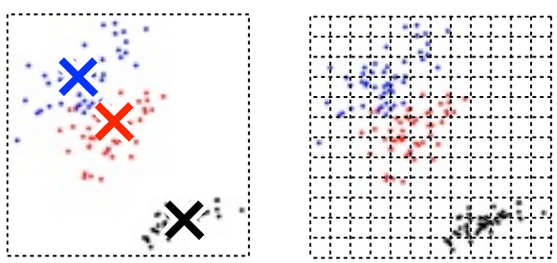


- How reliable are current separation measures on a diverse group of datasets?

Main Contributions



Taxonomy of visual cluster separation factors

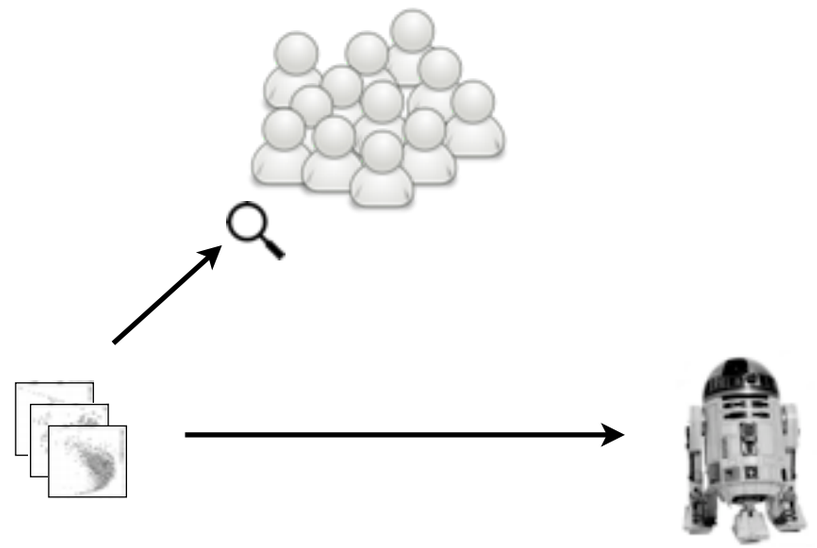


In-depth evaluation of 2 state-of-the-art separation measures

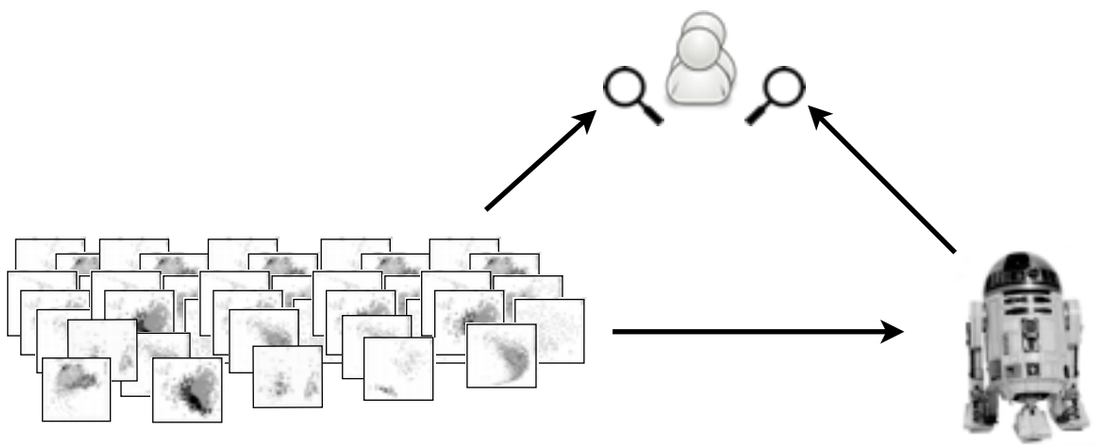
Qualitative Data Study

User vs. Data Study

- Previous work on measure evaluation: User studies
  - few datasets - many users

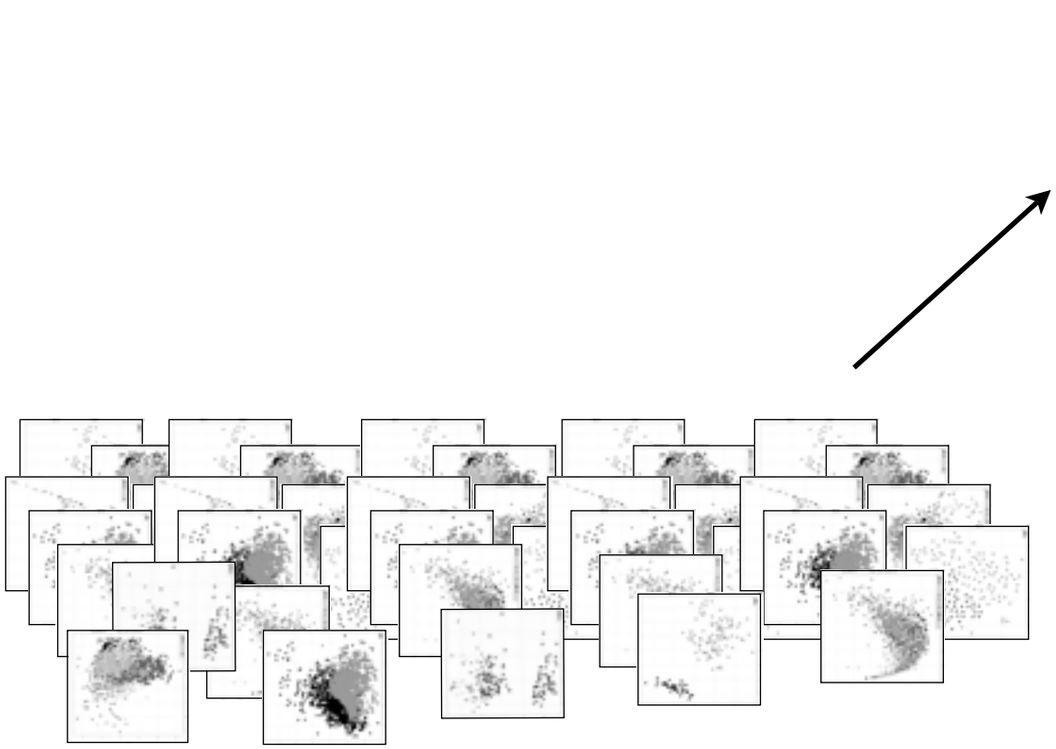


- *Missing: Dataset variety*
- Us: Data study
  - many datasets - few users

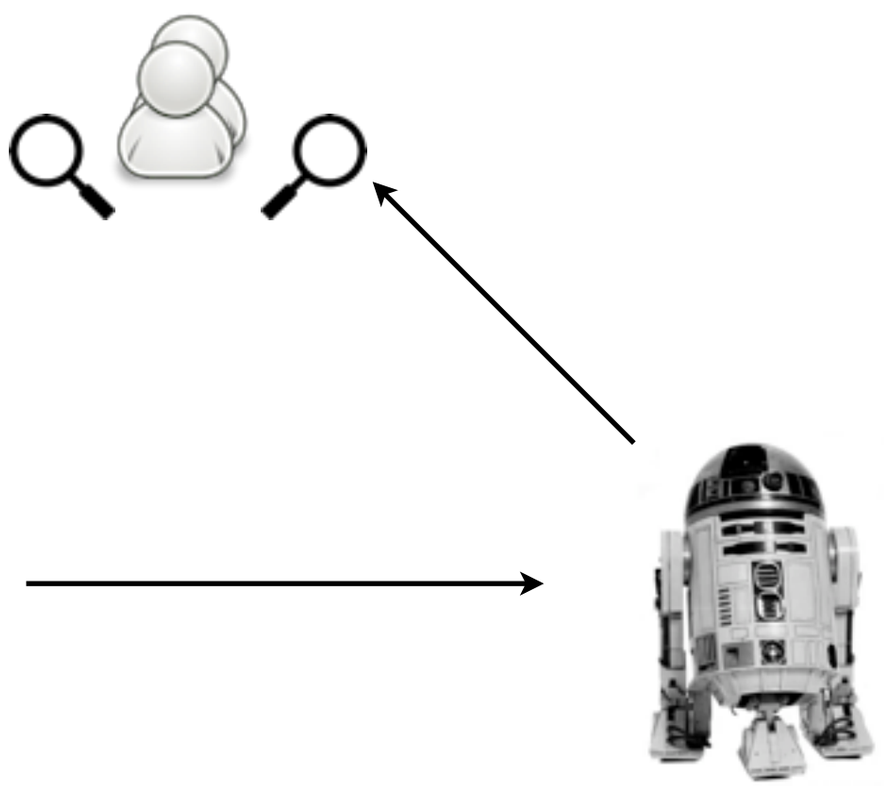


Qualitative Data Study

2 x



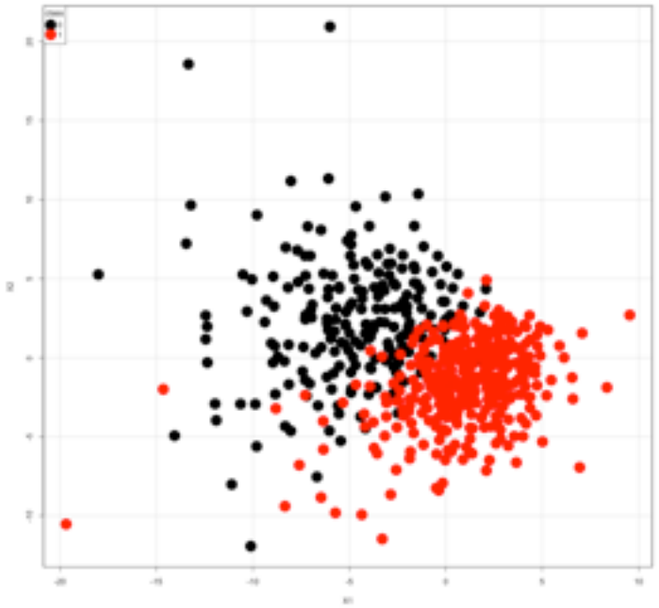
816 x

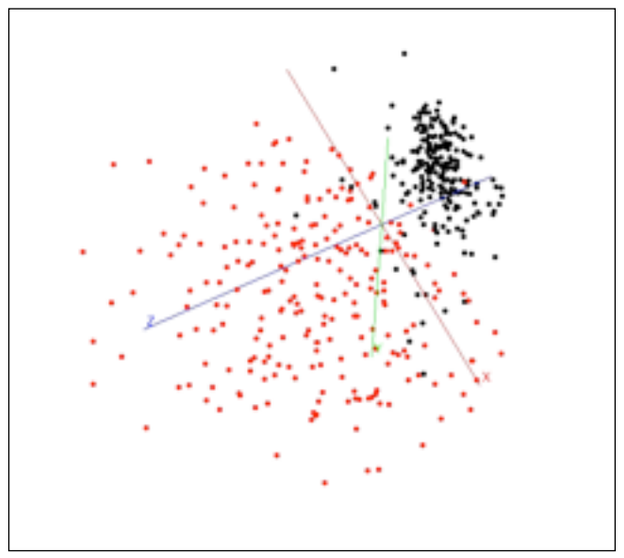


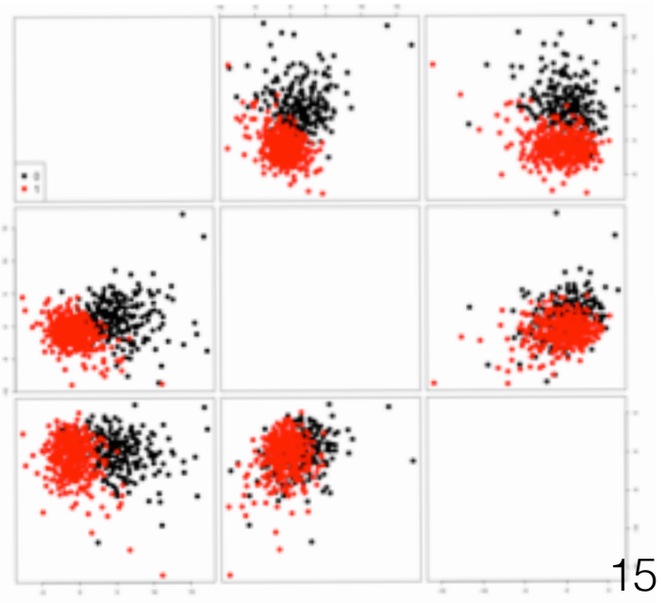
2 x

816 dataset instances

- 75 datasets
  - 31 real / 44 synthetic
  - pre-classified
- 4 DR techniques:
  - PCA, RobPCA, Glimmer MDS, t-SNE
- **3 Visual Encodings:**
  - 2D Scatterplot
  - Interactive 3D Scatterplot
  - SPLOM

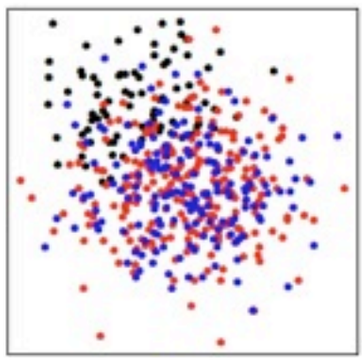






2 Measures

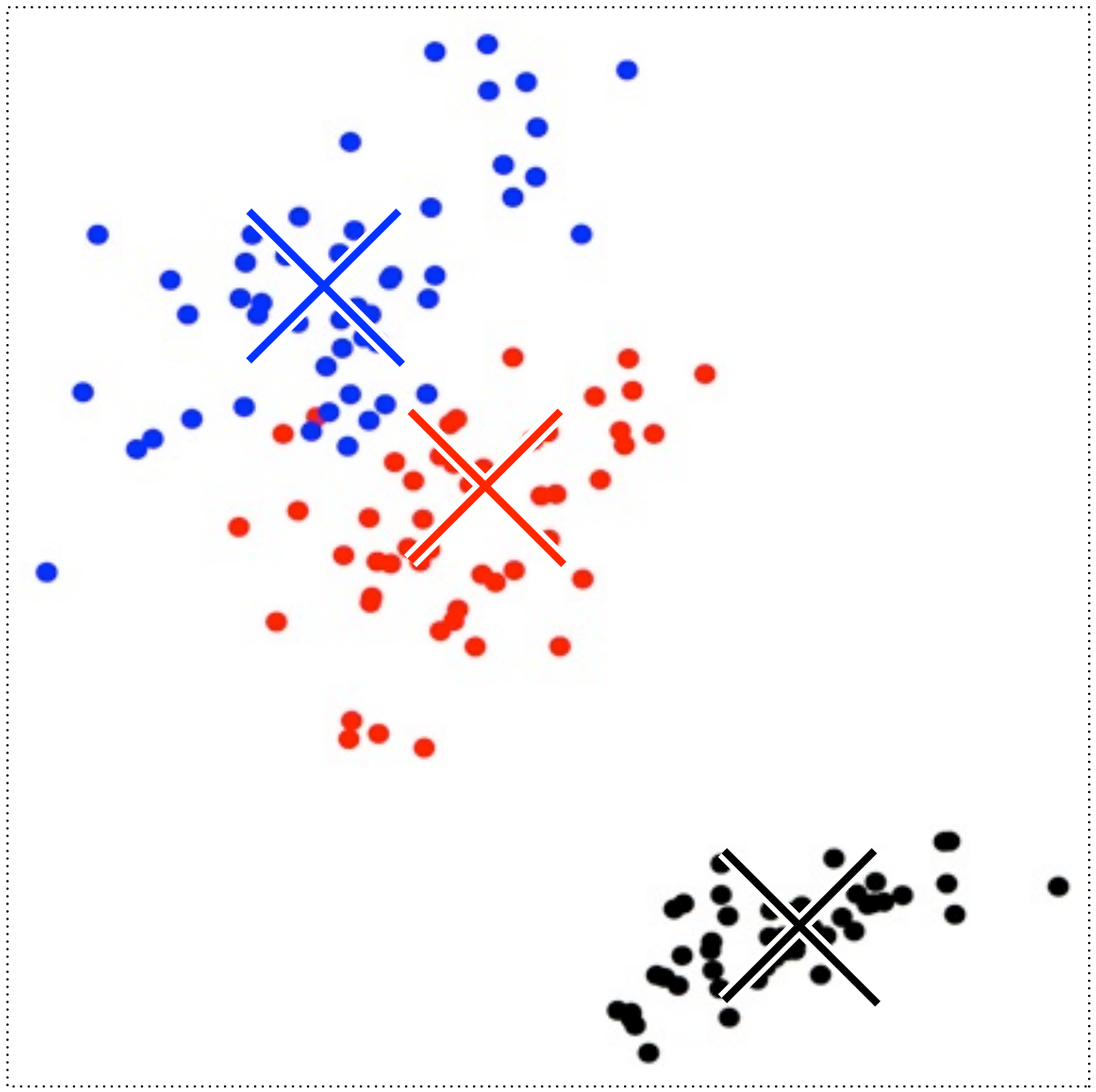
- Centroid[1] and Grid[1,2] Measures for 2D Scatterplots *(names adapted)*
- Found to be the current cutting edge[3]



0 (worst) - 100 (best)

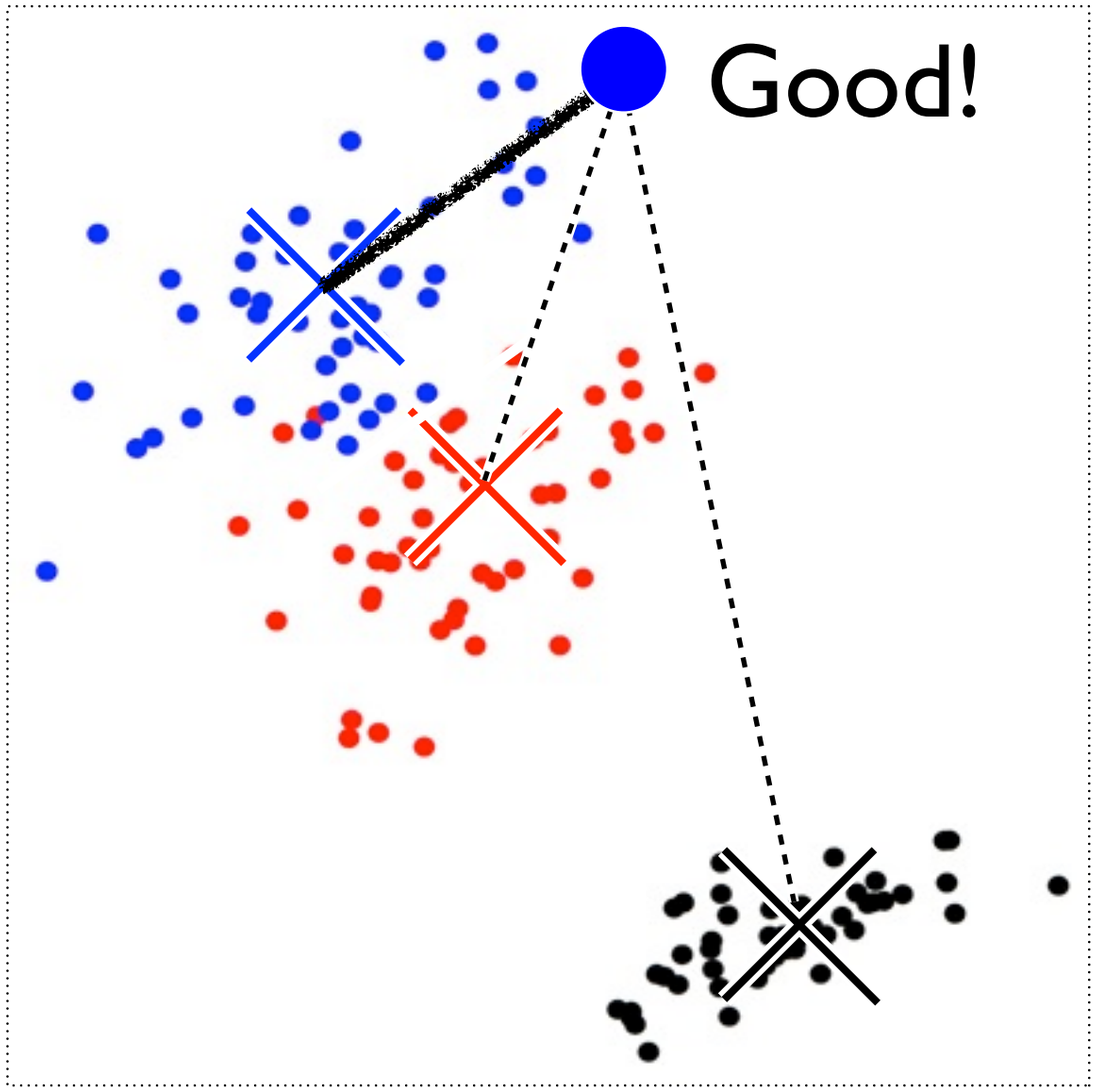
1. Sips et al.: Selecting good views of high-dimensional data using class consistency [EuroVis 2009]
2. Tatu et al.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data [VAST 2009]
3. Tatu et al.: Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data [AVI 2010]

Centroid Measure



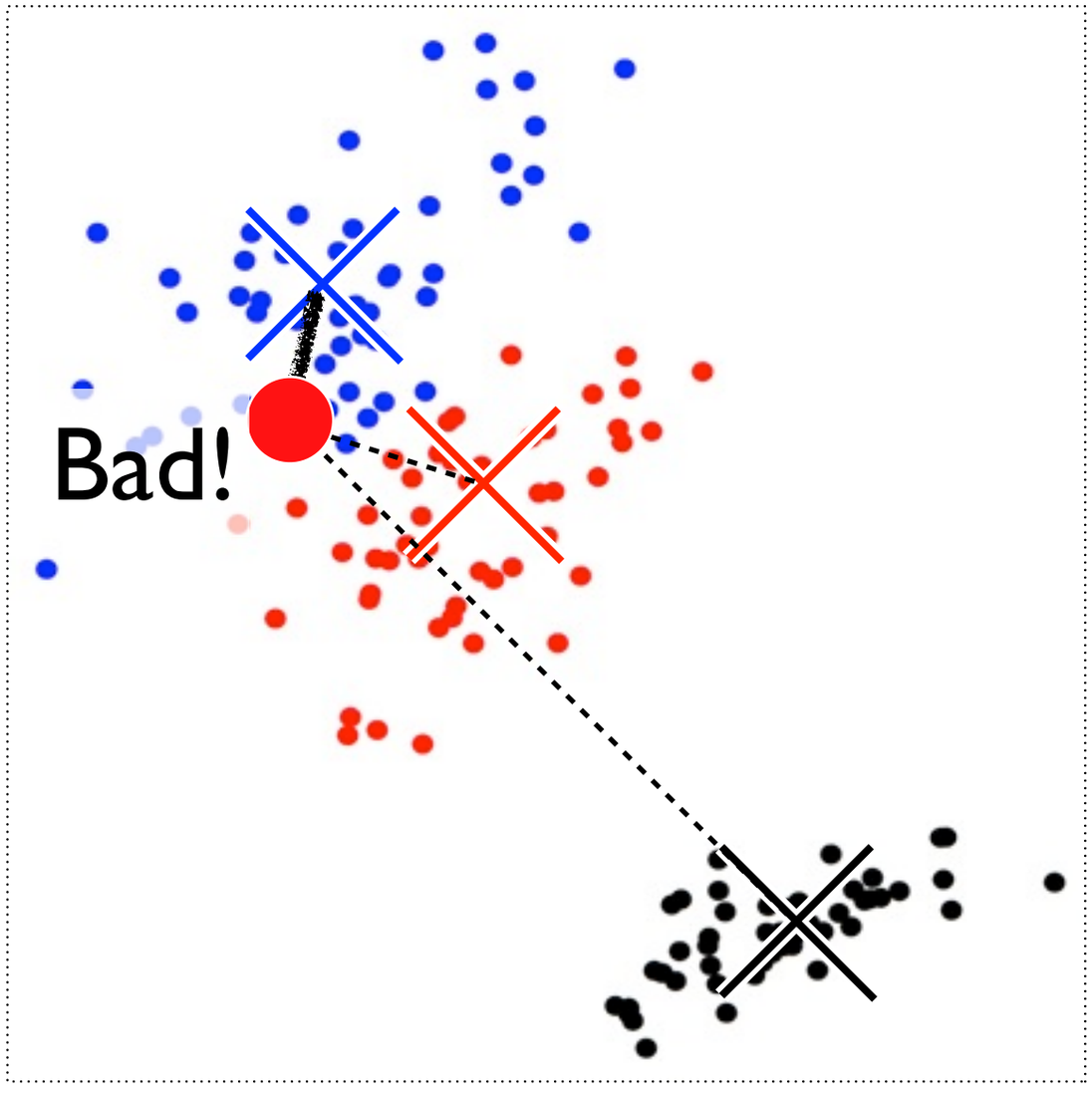
Centroid: 93

Centroid Measure



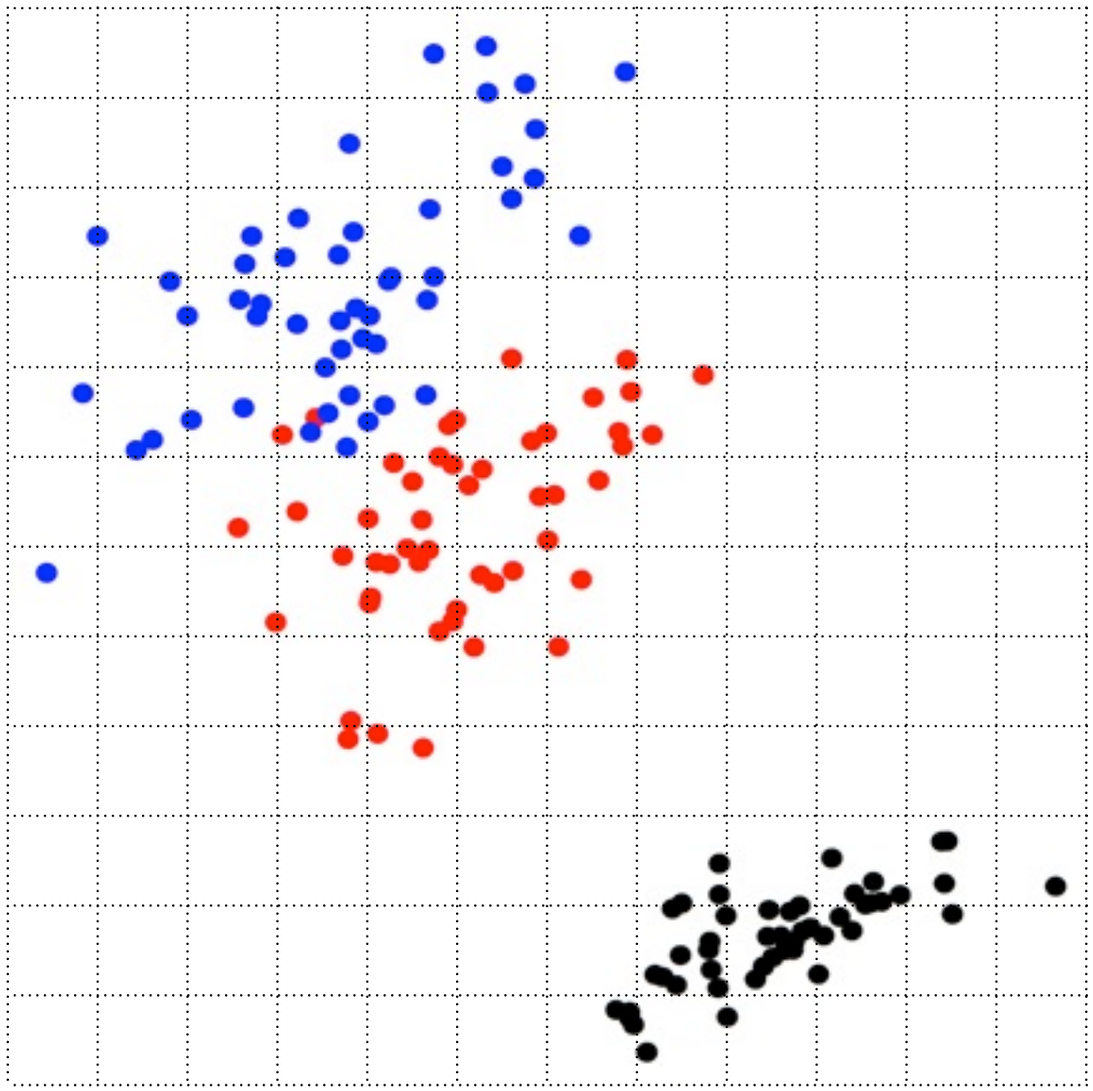
Centroid: 93

Centroid Measure



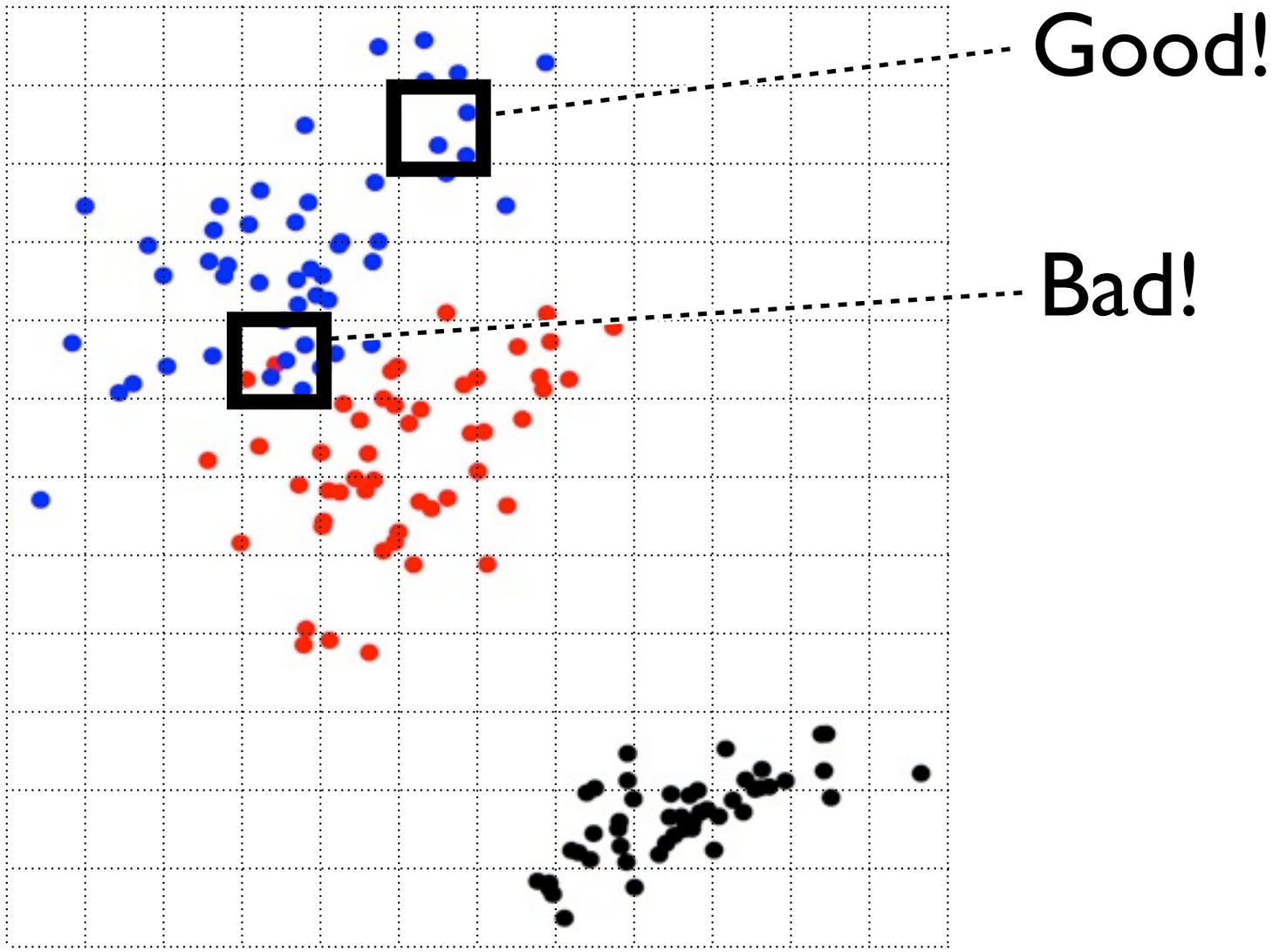
Centroid: 93

Grid Measure



Grid: 97

Grid Measure

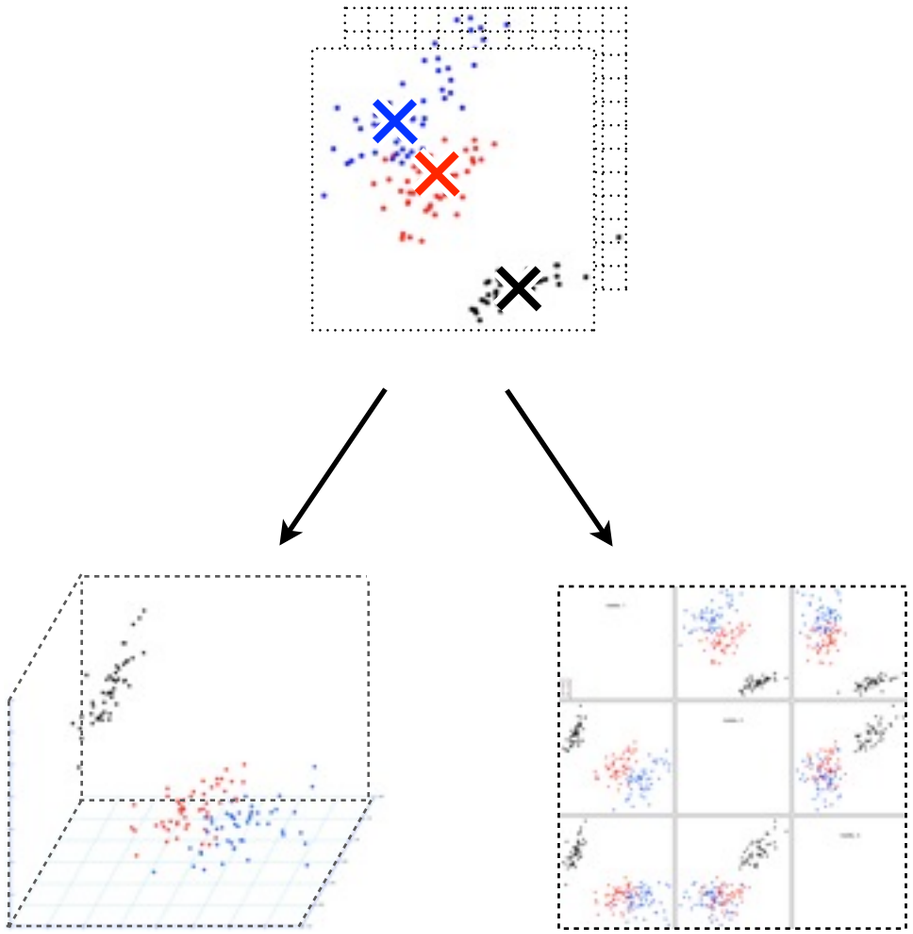


Grid: 97

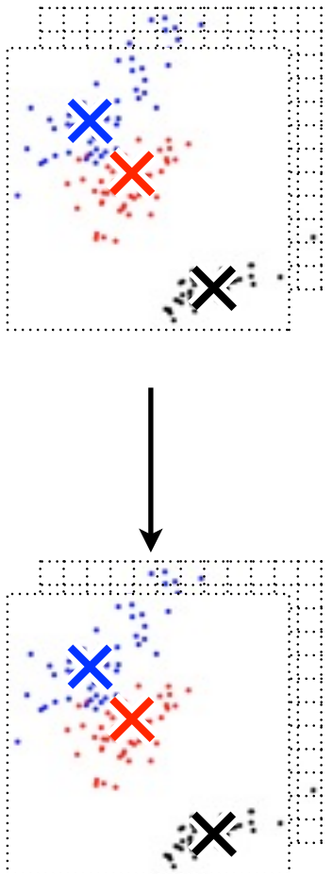
Extensions of Measures

Straight forward

3D / SPLOM



Classwise



Overall: 93

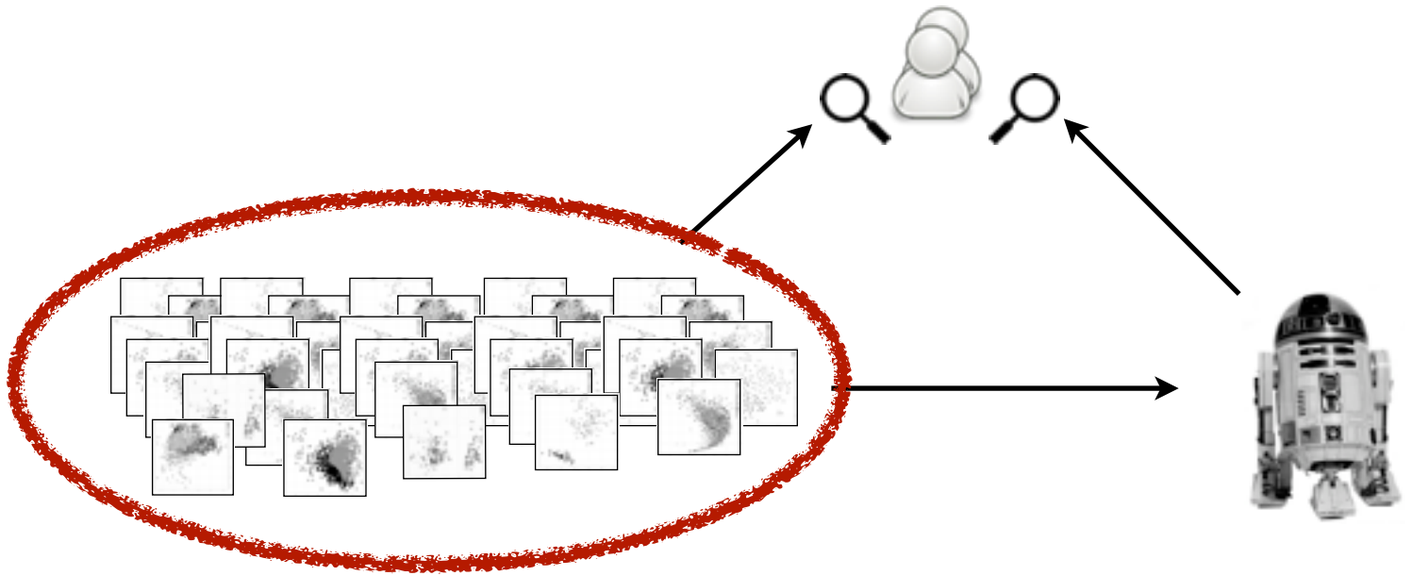
blue: 86

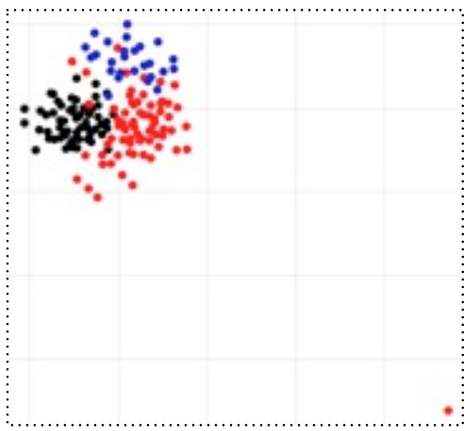
red: 94

black: 100

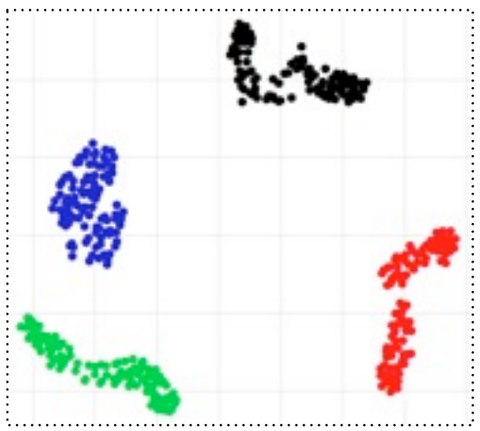
Data Analysis
and
High-Level Results

Data analysis (part 1):
Qualitative analysis of cluster separation factors

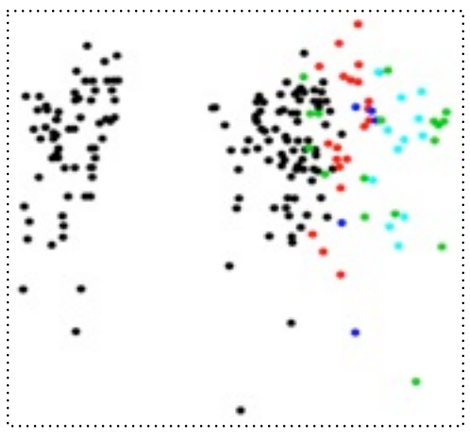




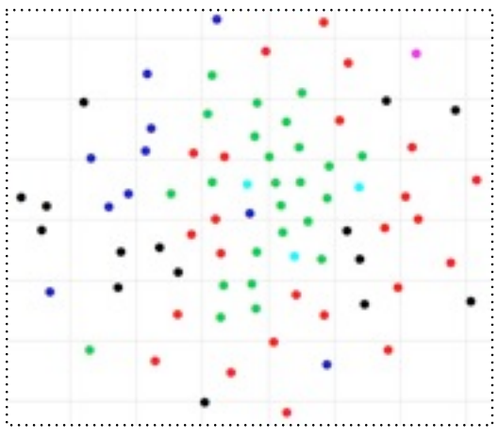
outlier



shape



split



equidistant points

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

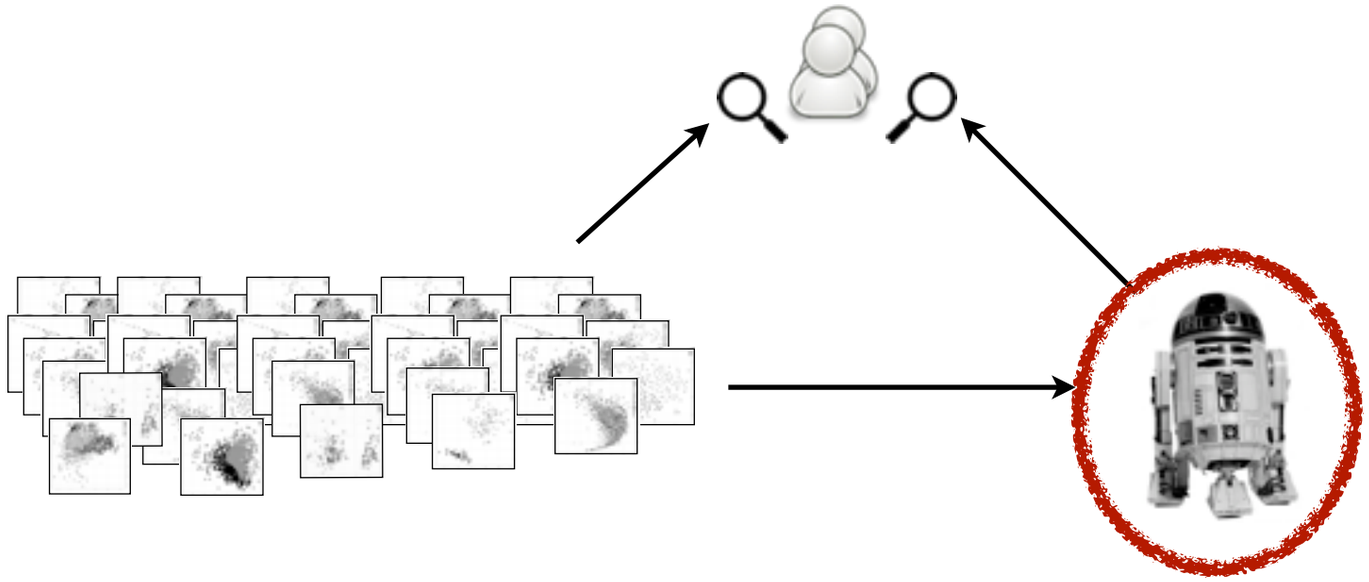
Split
contiguous → split

Variance of Shape
similar → different

Inner-Outer Position
non-existent ← → existent

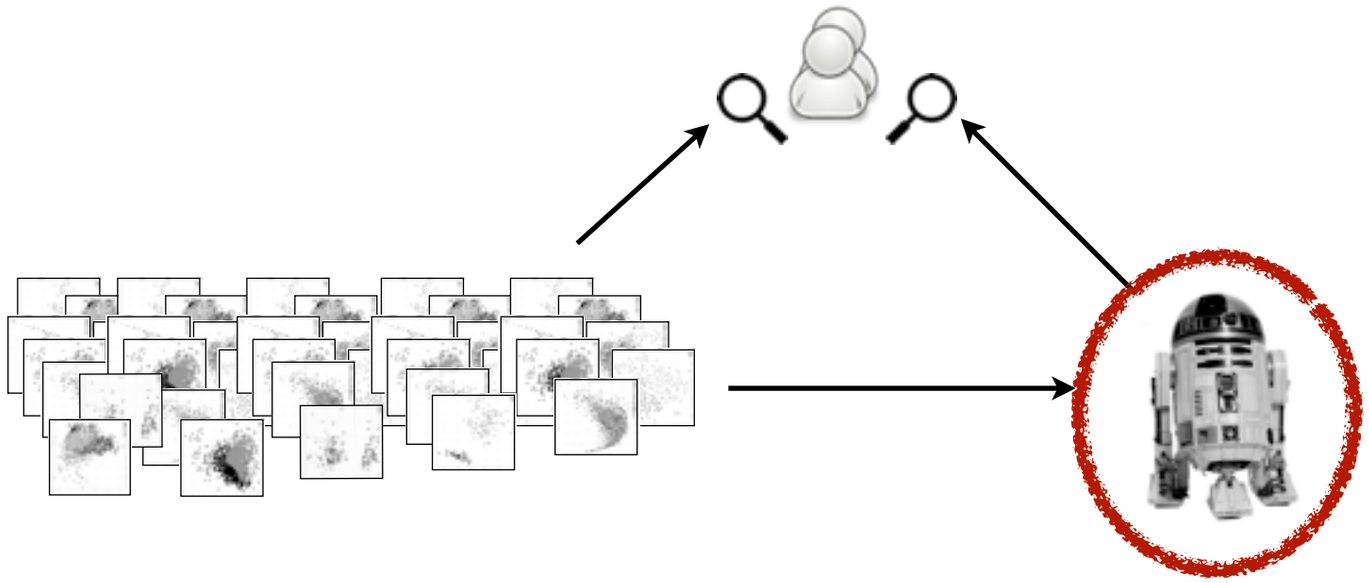
Class Separation
full overlap | partial overlap | adjacent | separate | distant

Data analysis (part 2):
Evaluating the measures

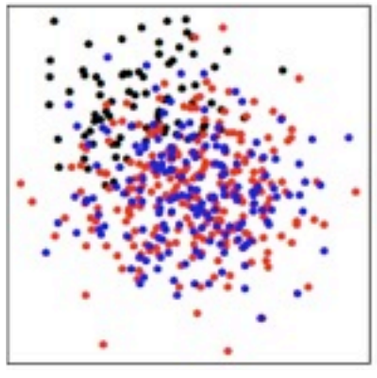


Measure aligns with human judgement?

Data analysis (part 2):
Evaluating the measures

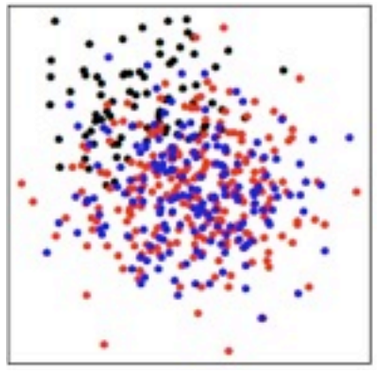


OK



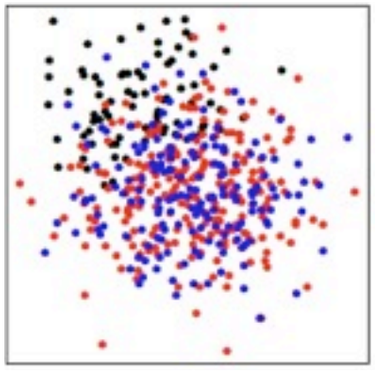
Measure: **10**

dubious



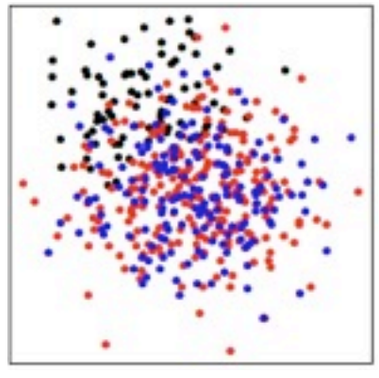
Measure: **50**

poor



Measure: **90**

classwise-poor

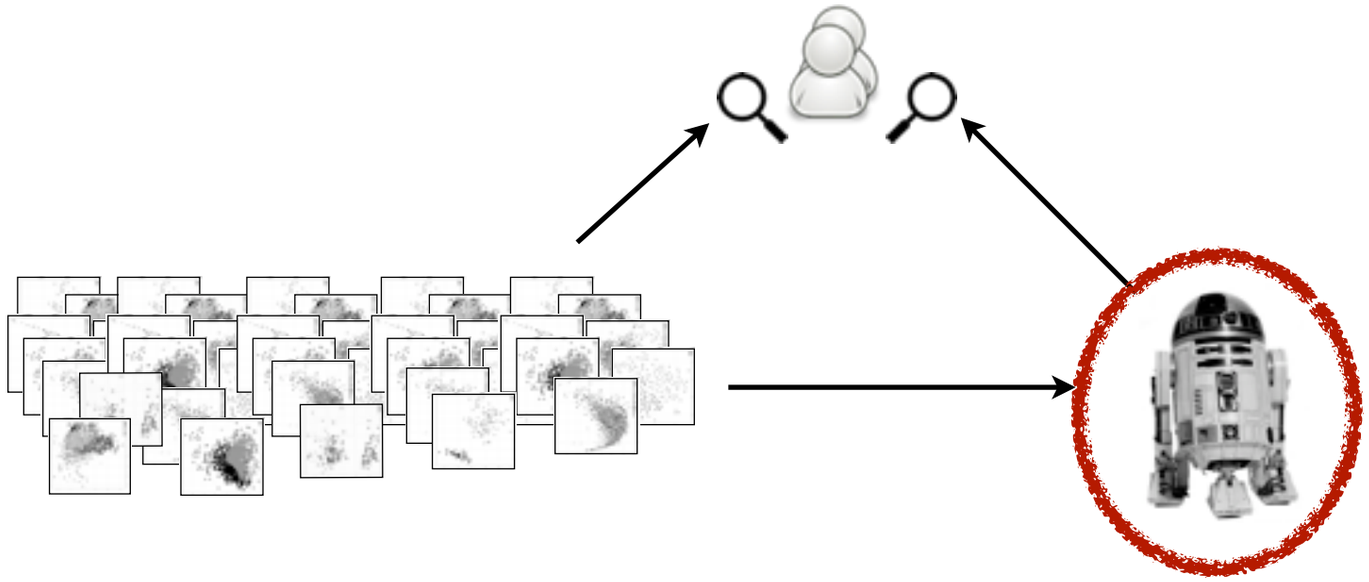


blue: 100

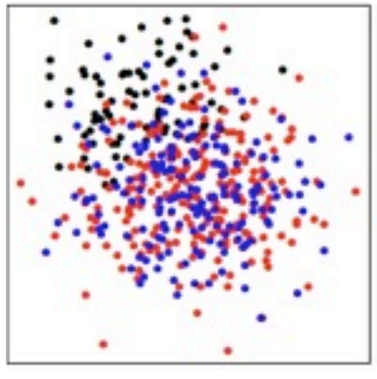
red: 7

black: 20

Data analysis (part 2):
Evaluating the measures

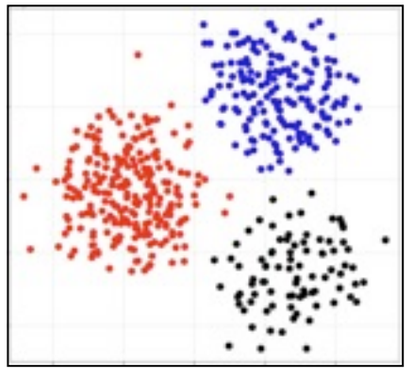


False Positive



Measure: **90**

False Negative

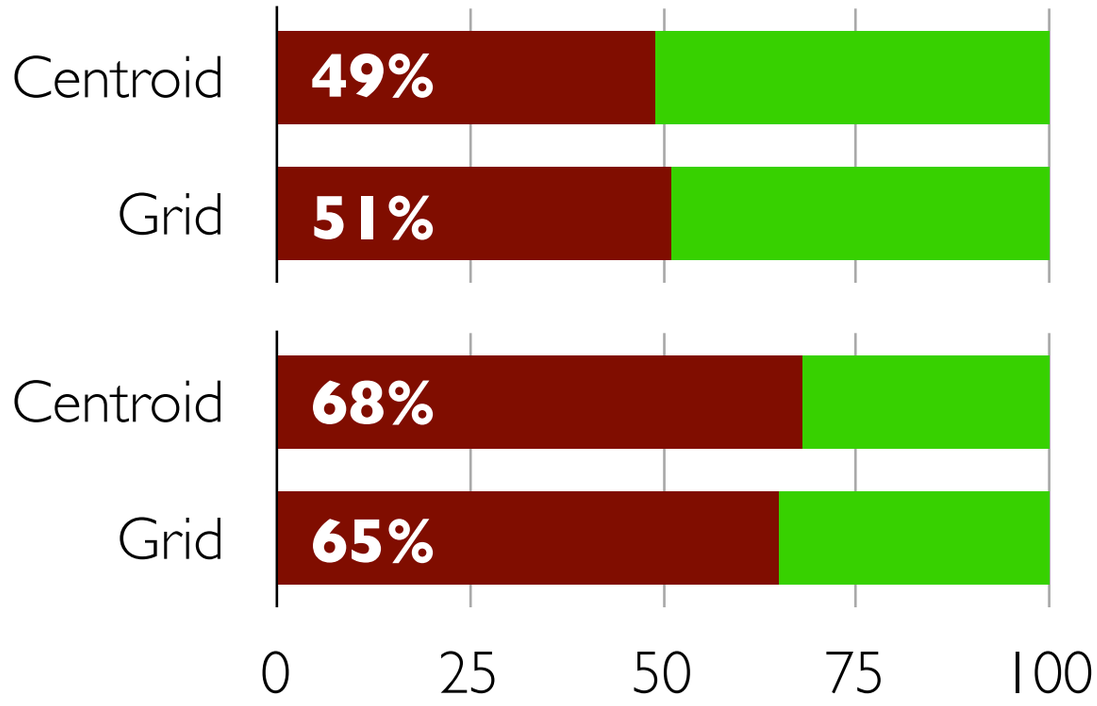


Measure: **10**

High-level results

Poor / Dubious / Classwise-Poor (Failure cases)
Ok

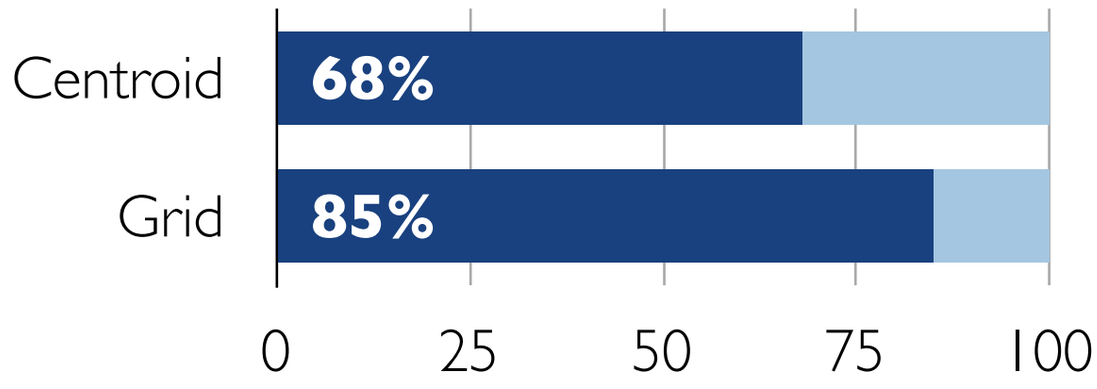
All (816)



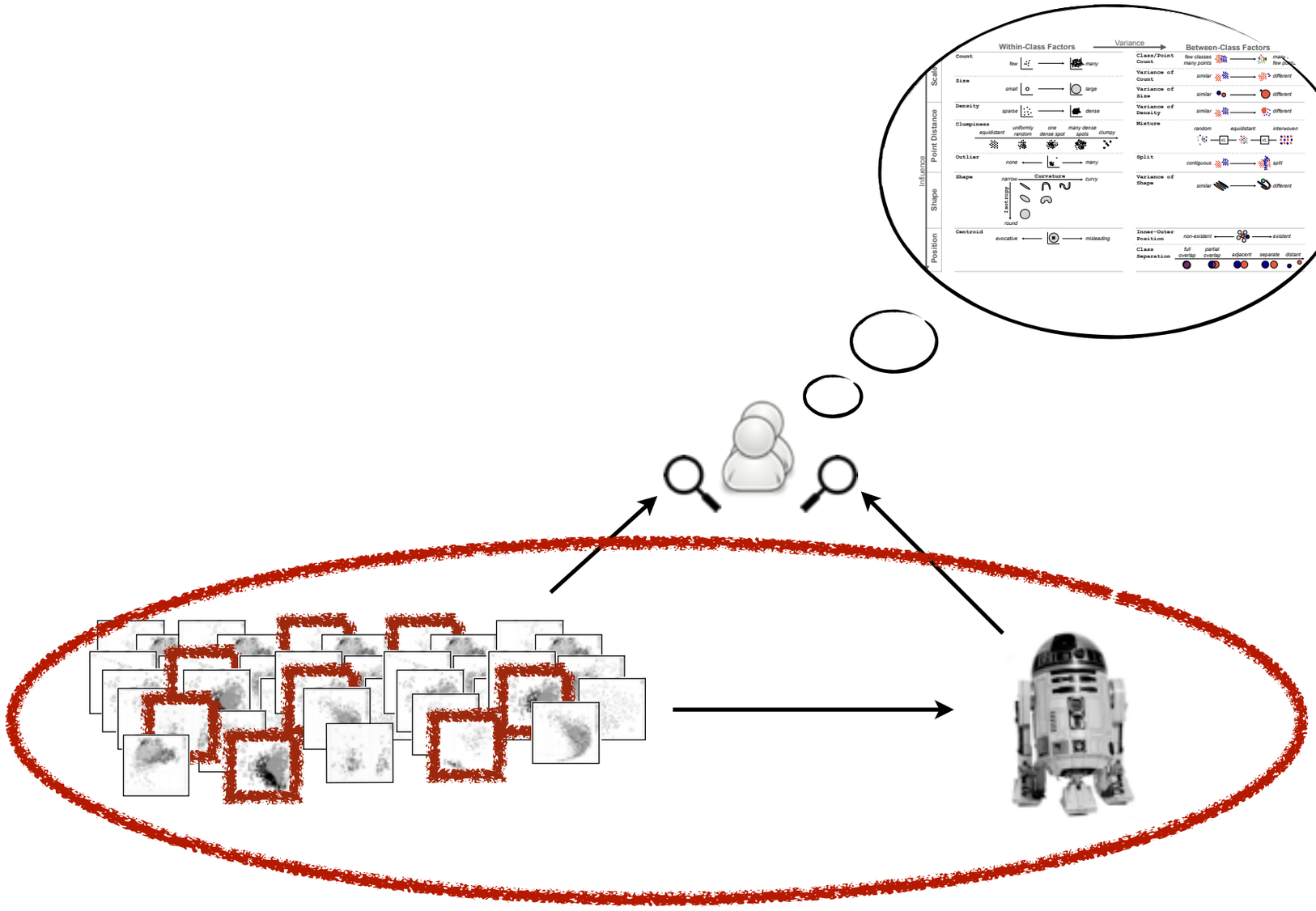
Only real (296)

False Positives
False Negatives

All failure cases

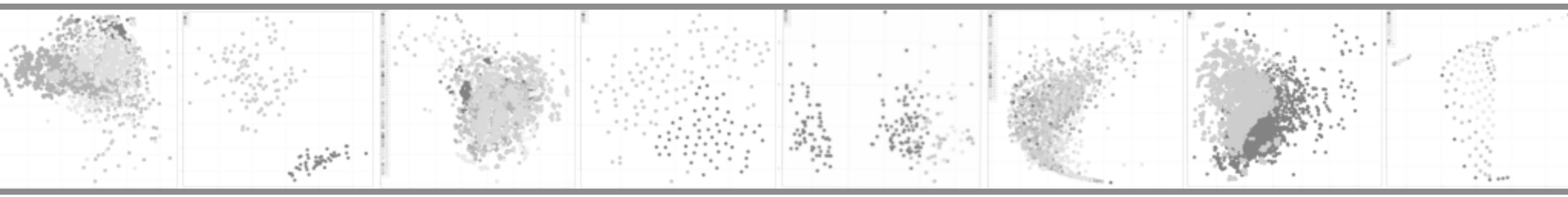


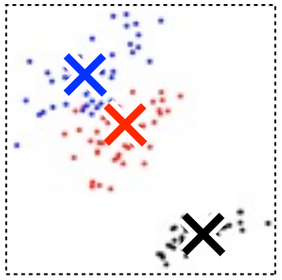
Data analysis (part 3):
Qualitative analysis of failure reasons



Using the factors we found in part I
to explain the reasons why measures failed!

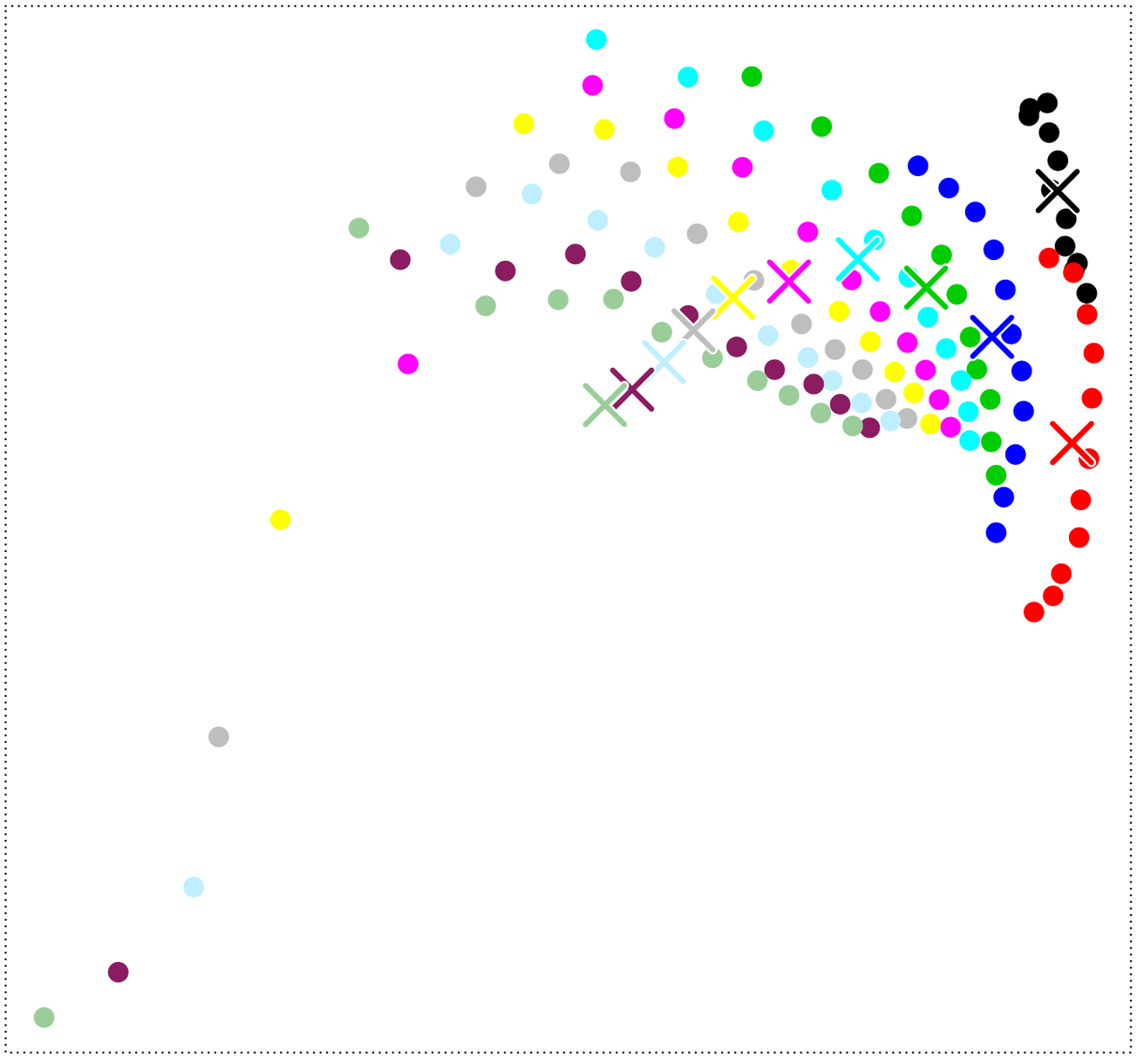
Walkthrough





Centroid:

Stringy / outliers



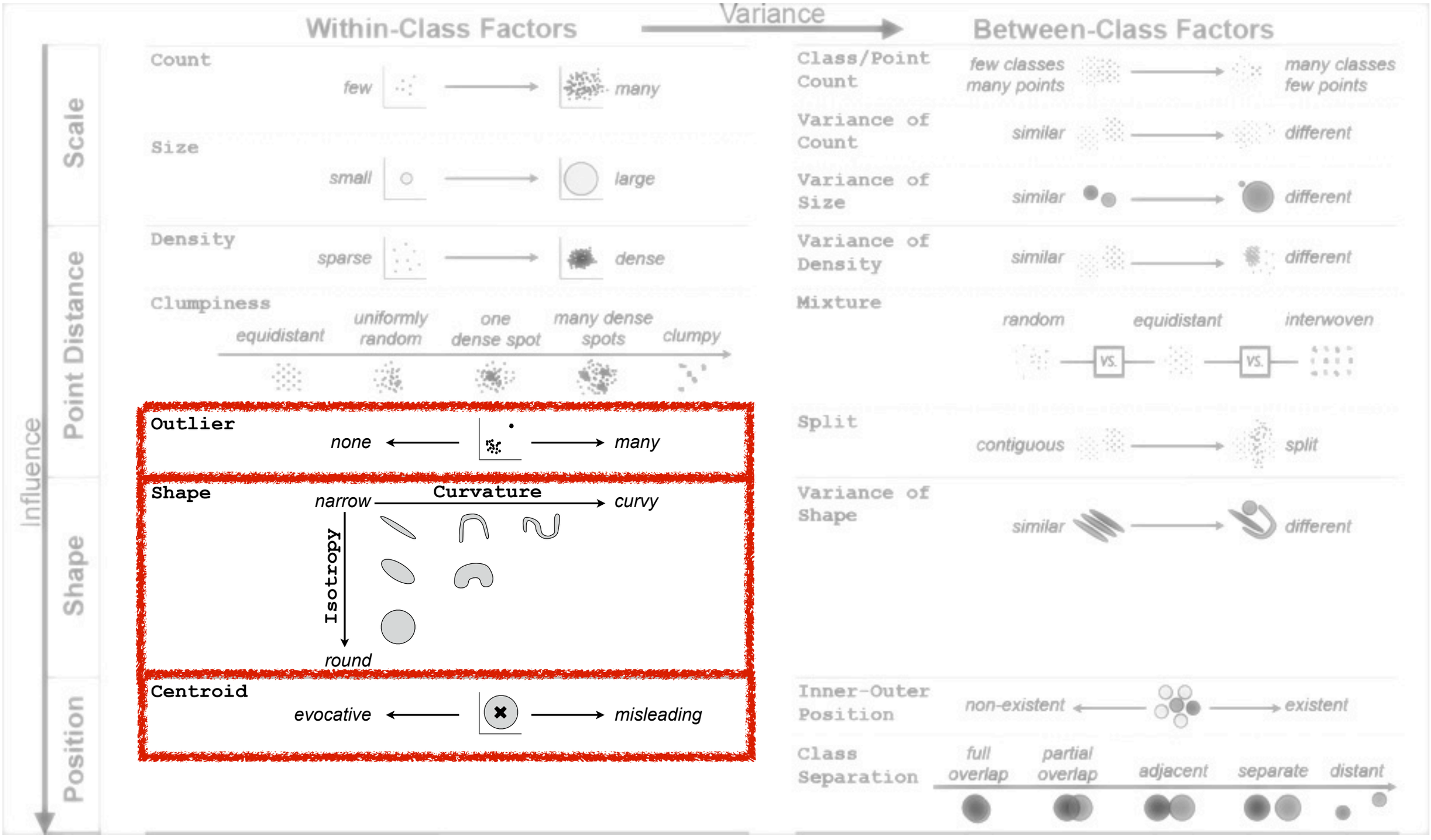
Overall: **29 (Bad)**

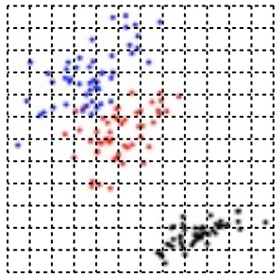
Problem: **FN**

Data: Fisheries, real

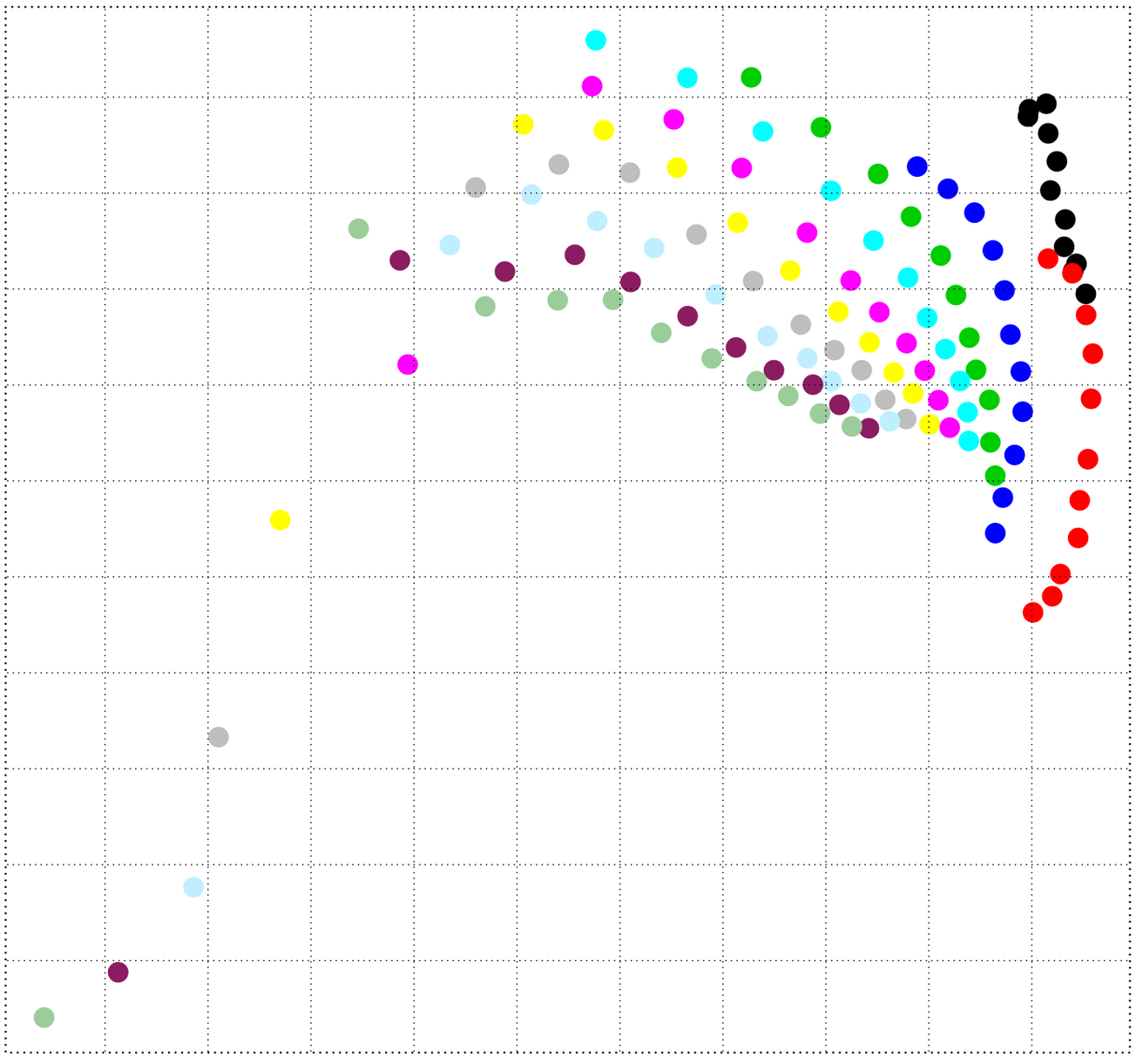
DR: MDS

In terms of taxonomy ...





Grid:
Adjacent Strings



Black & Red: ~70-80

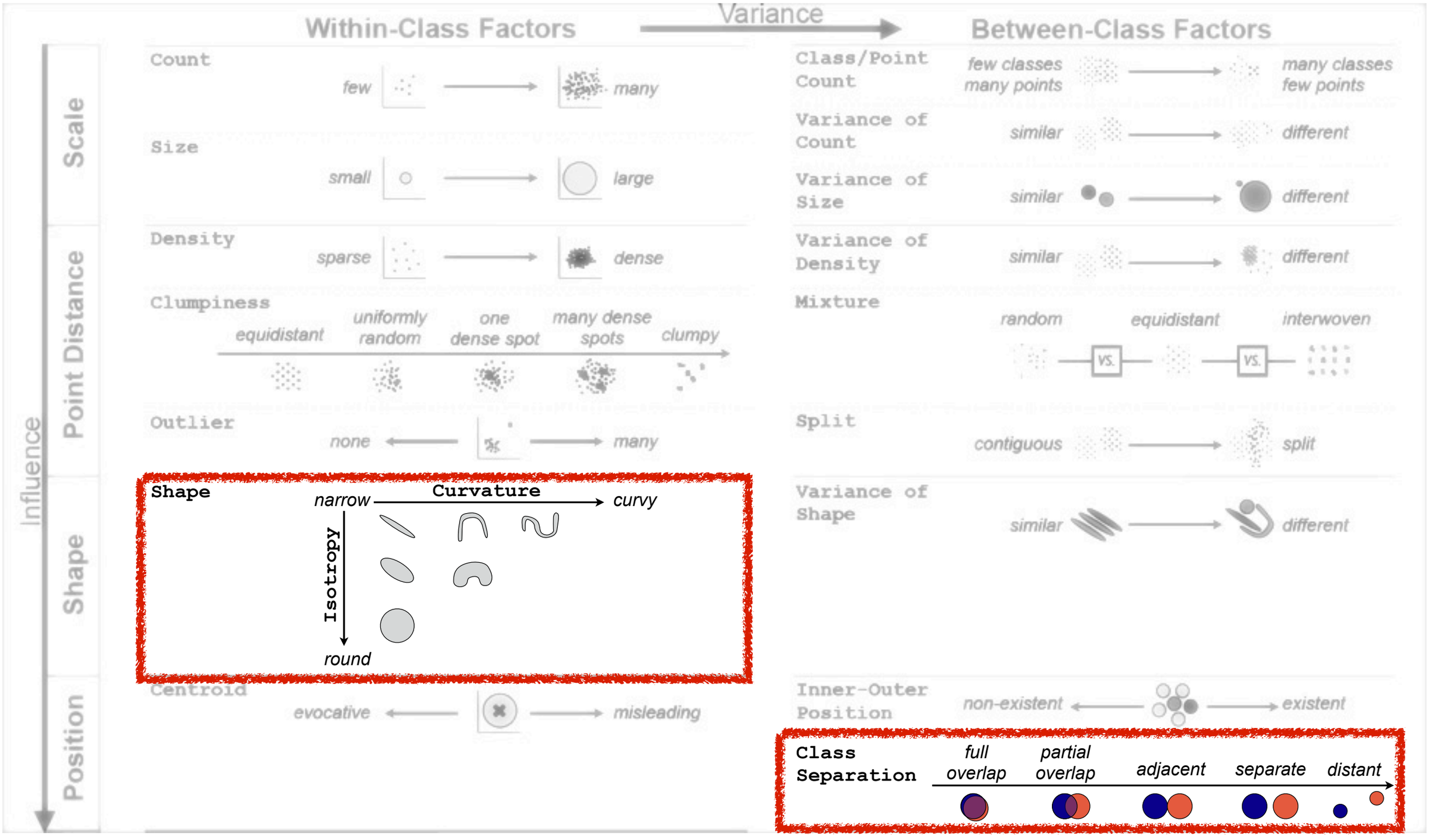
Others: **~40-50 (Bad)**

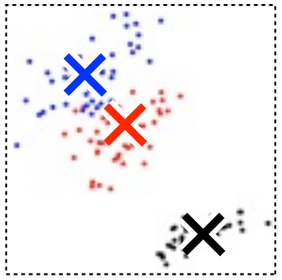
Problem: **FN**

Data: Fisheries, real

DR: MDS

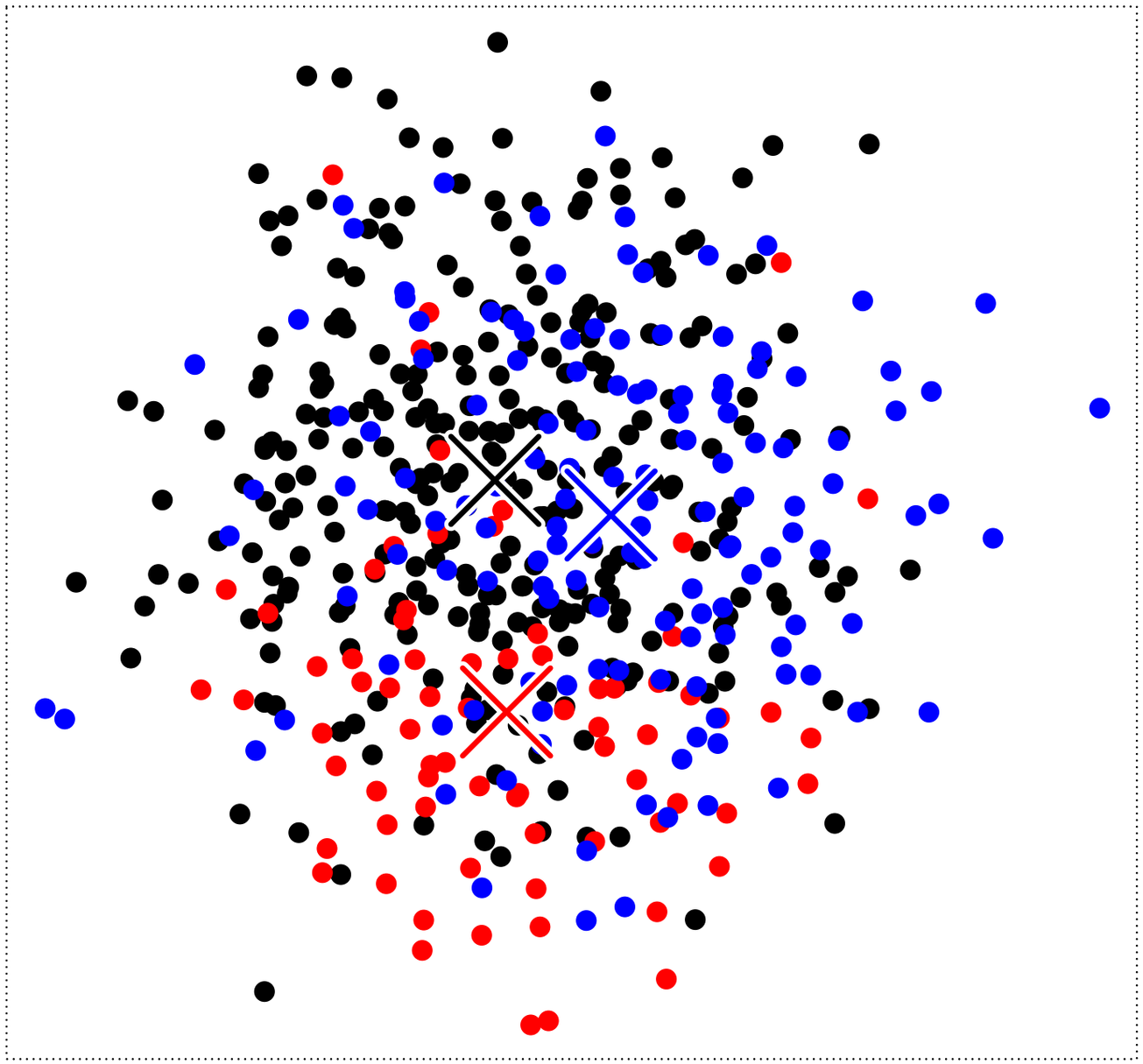
In terms of taxonomy ...





Centroid:

Big Classes Overspread Small

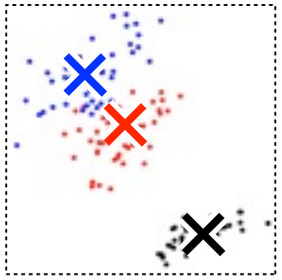


Red: **77 (Good)**

Problem: **FP**

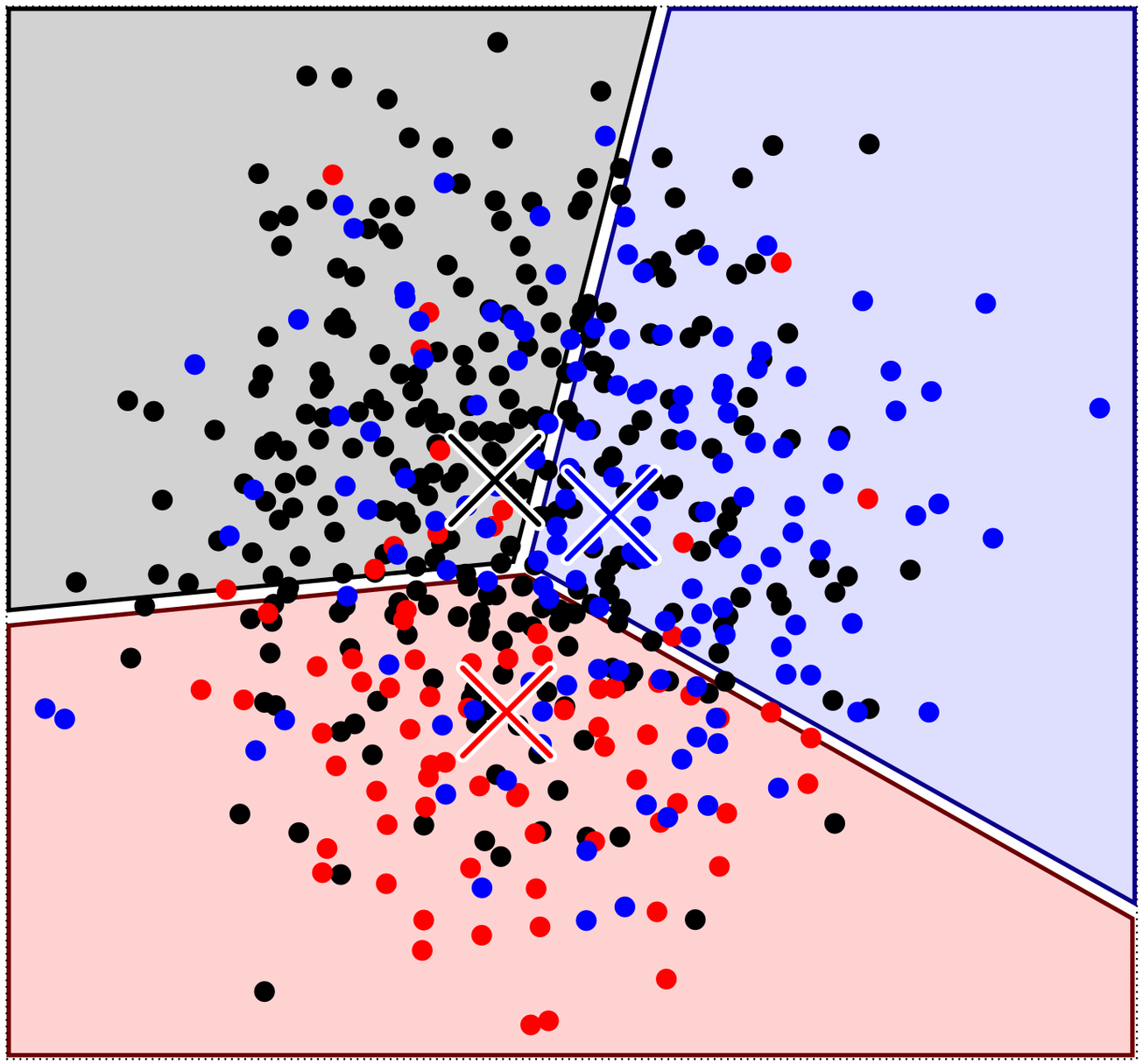
Data: Gaussian, synthetic

DR: MDS



Centroid:

Big Classes Overspread Small



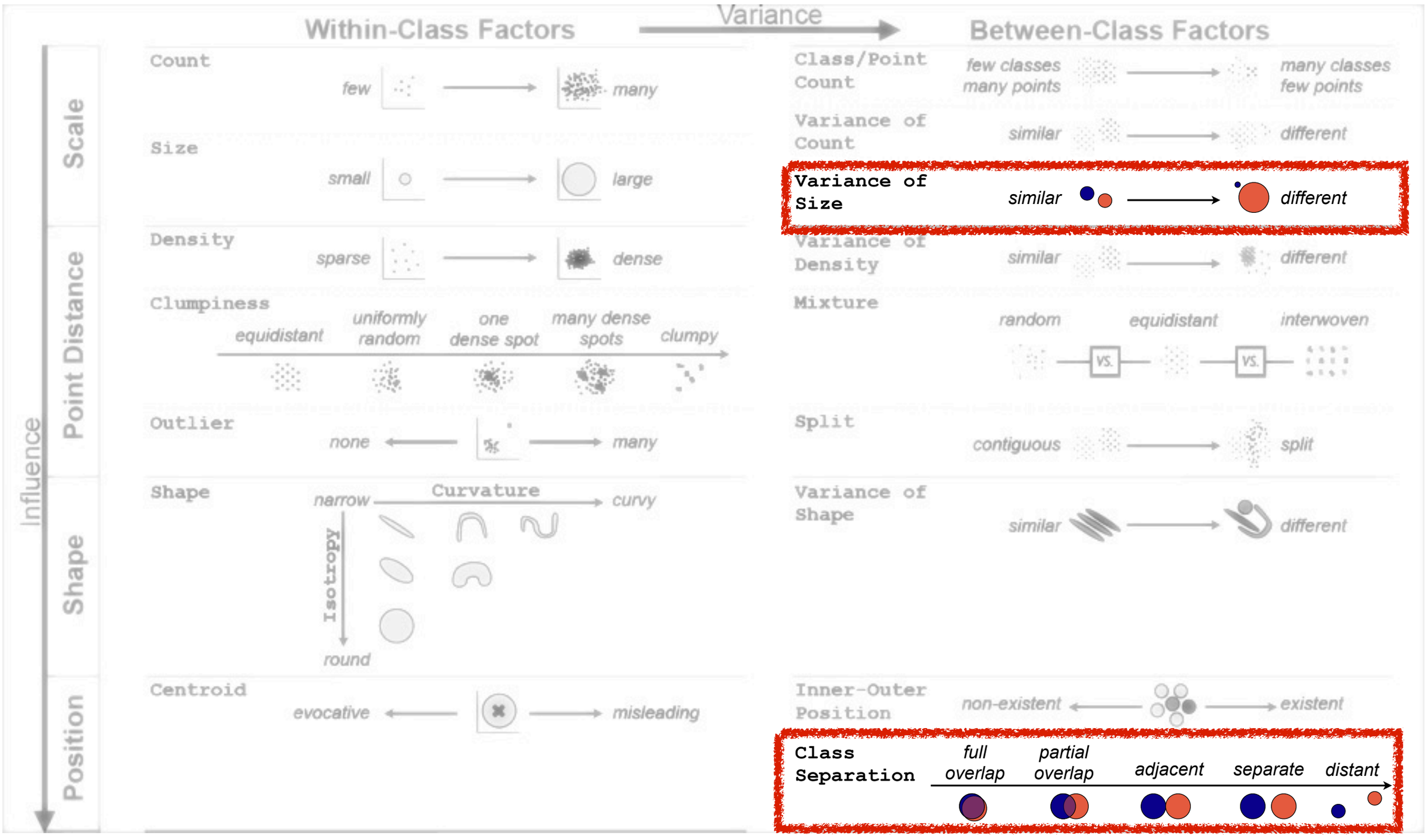
Red: **77 (Good)**

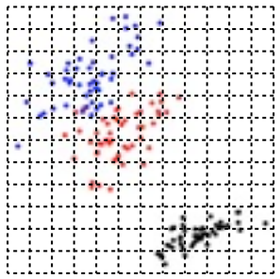
Problem: **FP**

Data: Gaussian, synthetic

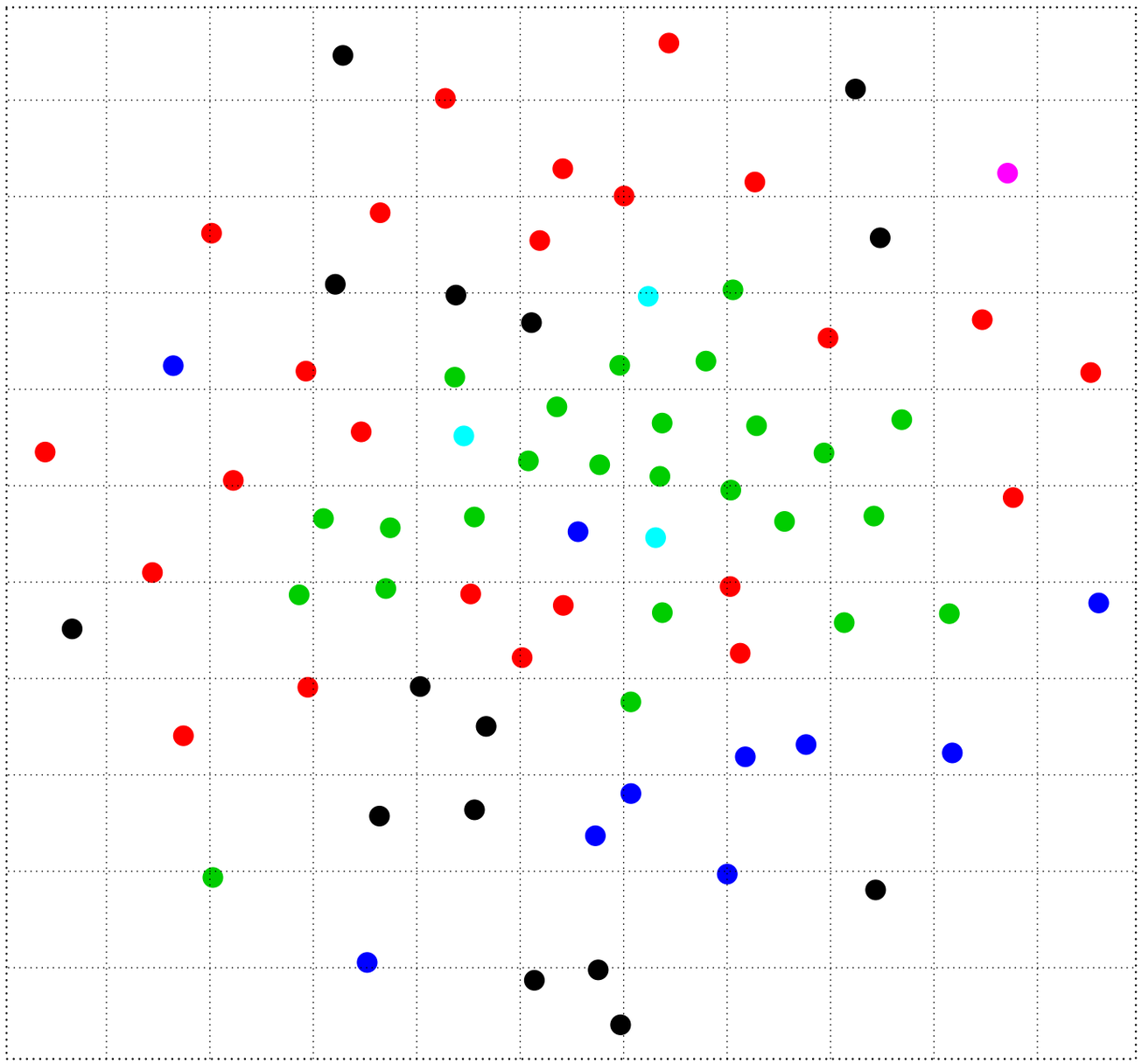
DR: MDS

In terms of taxonomy ...





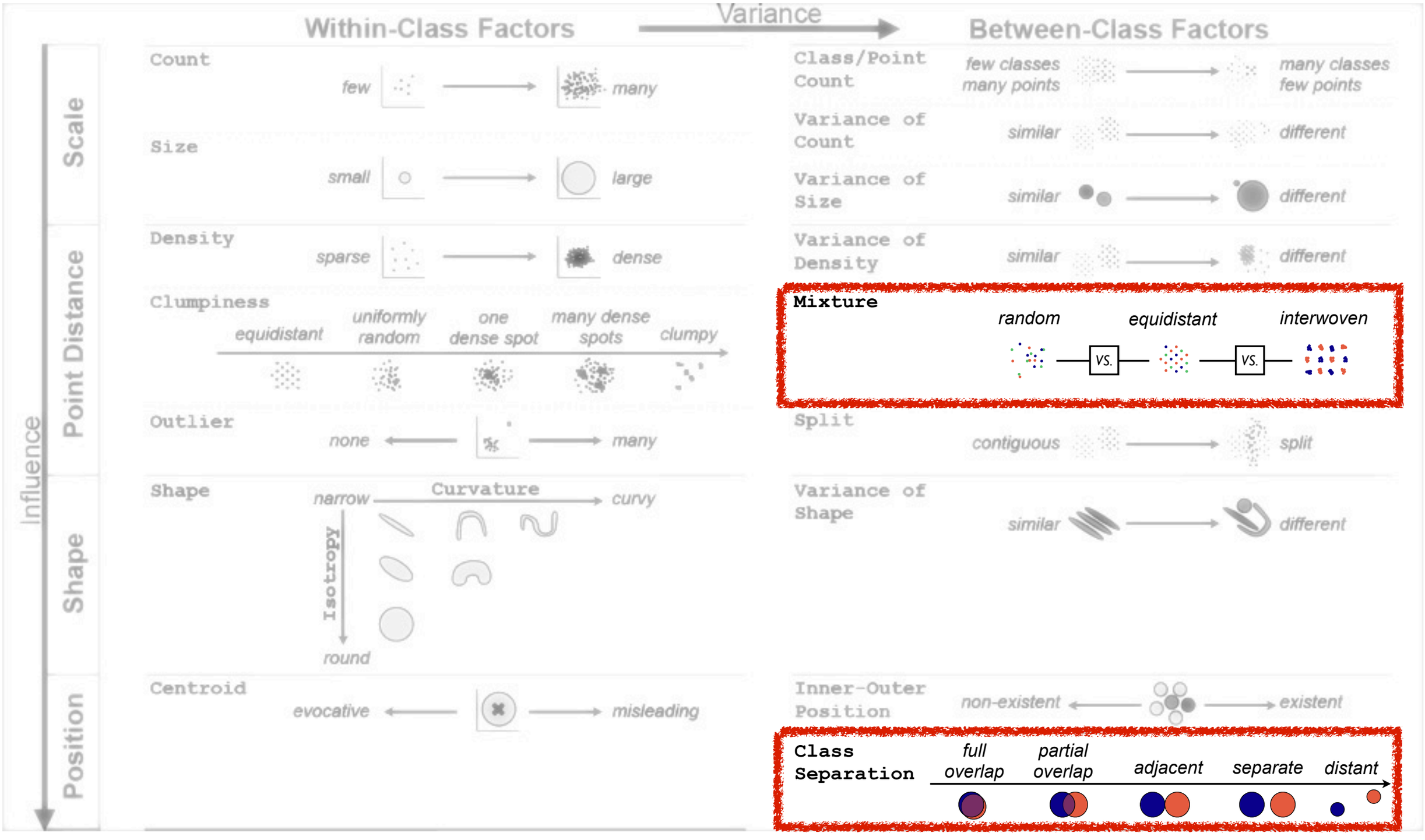
Grid:
Equidistant Points



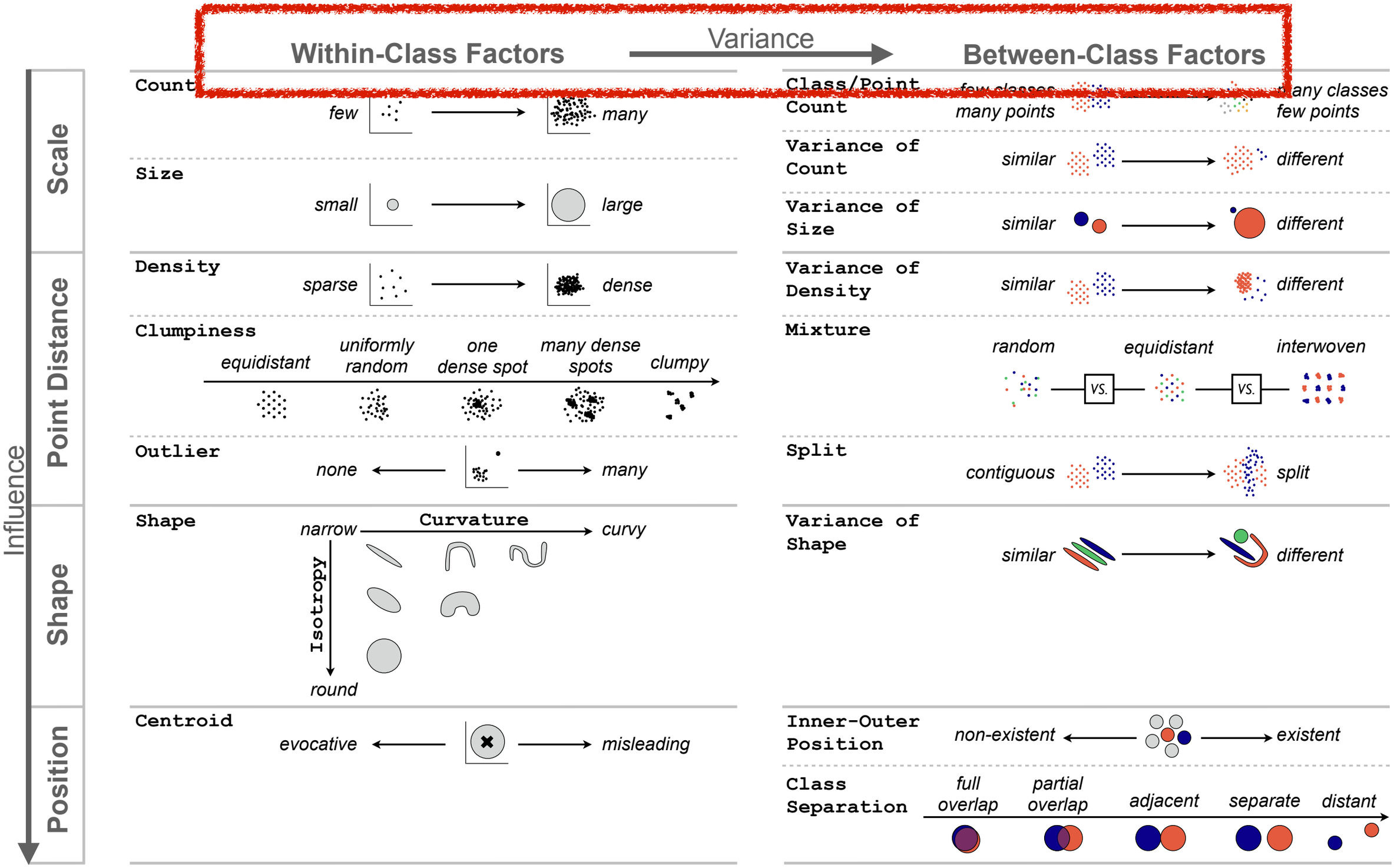
Overall: **99 (Good)**
Problem: **FP**

Data: HIV, real
DR: t-SNE

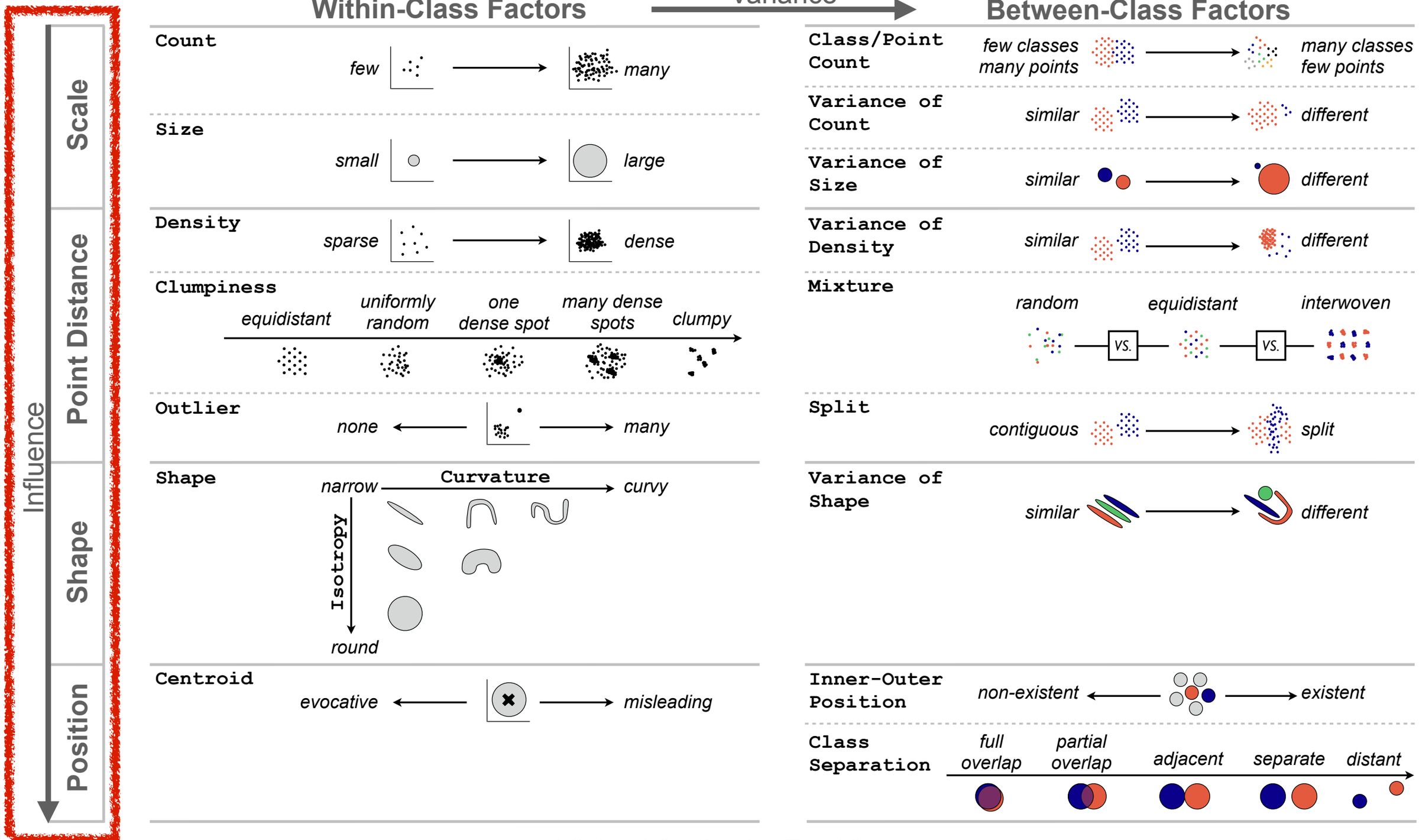
In terms of taxonomy ...



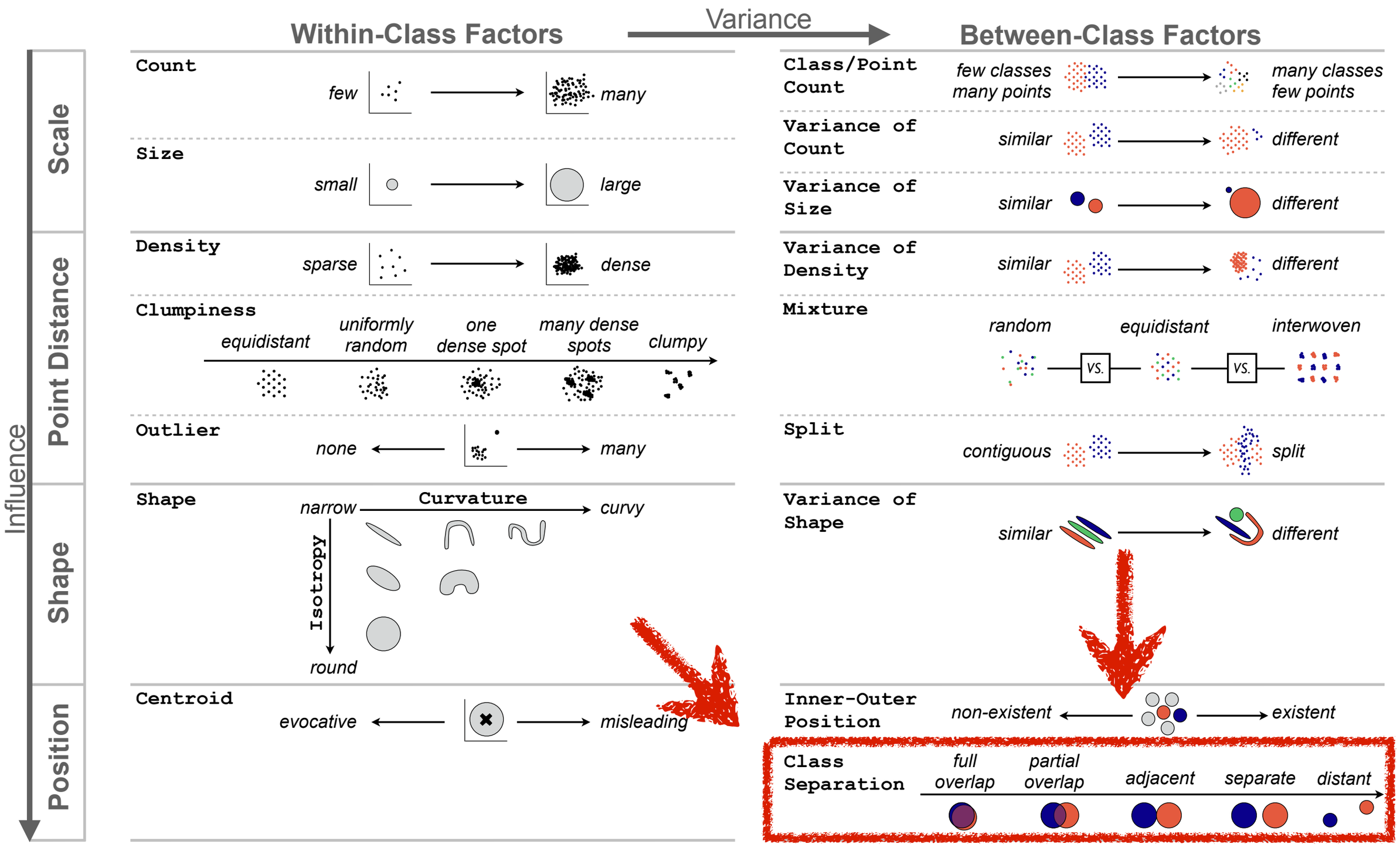
A taxonomy of visual cluster separation factors



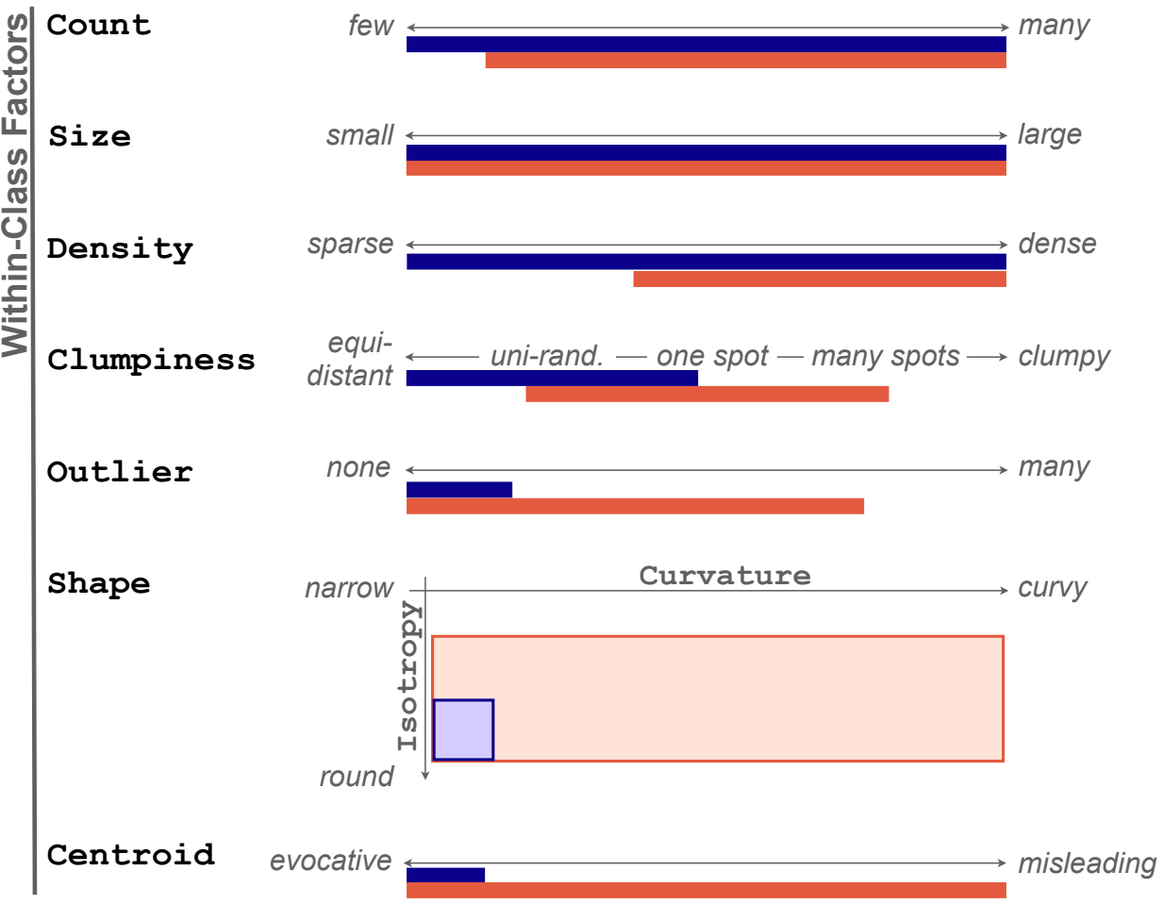
A taxonomy of visual cluster separation factors

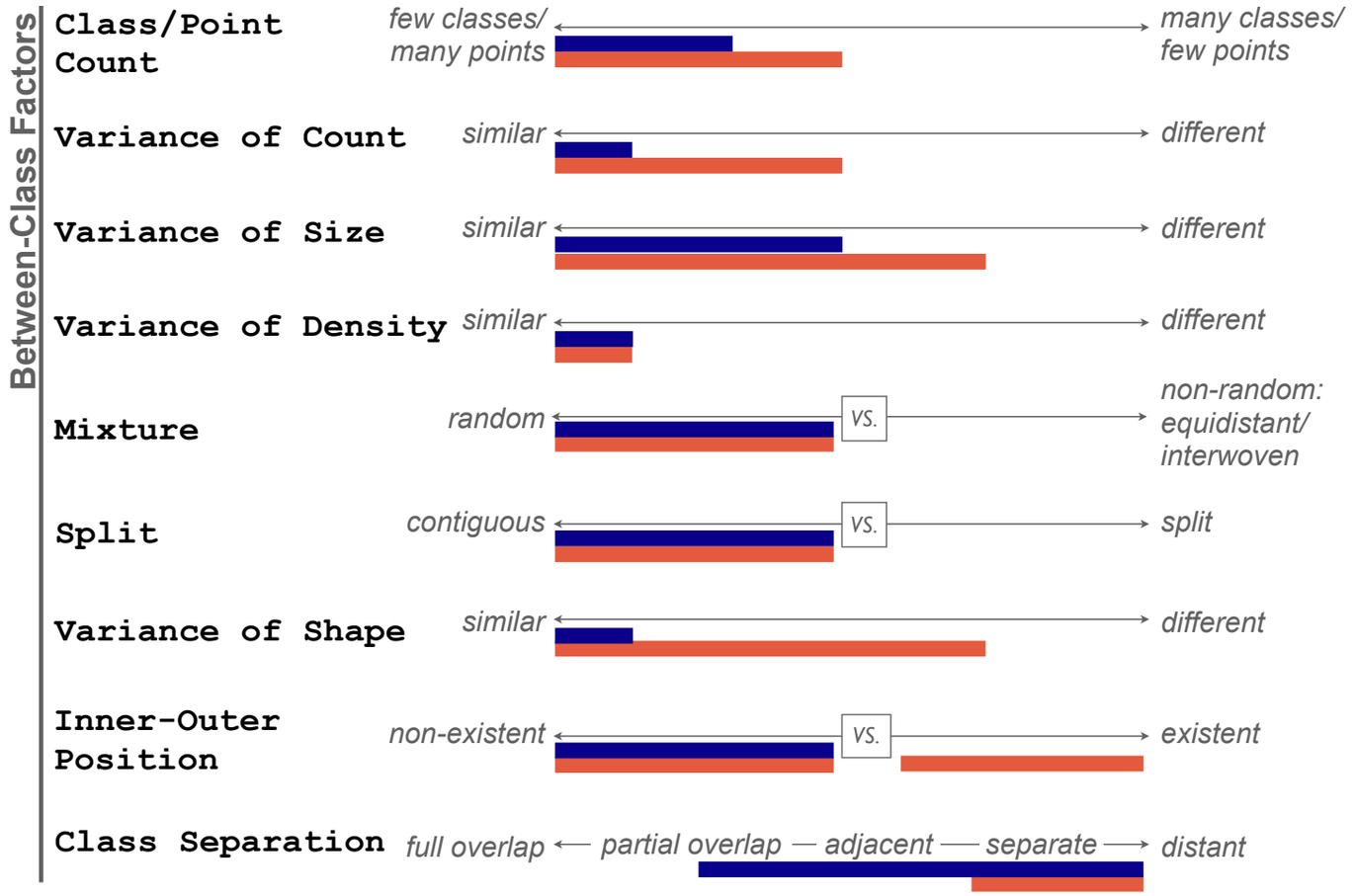


A taxonomy of visual cluster separation factors



Mapping measure assumptions onto taxonomy



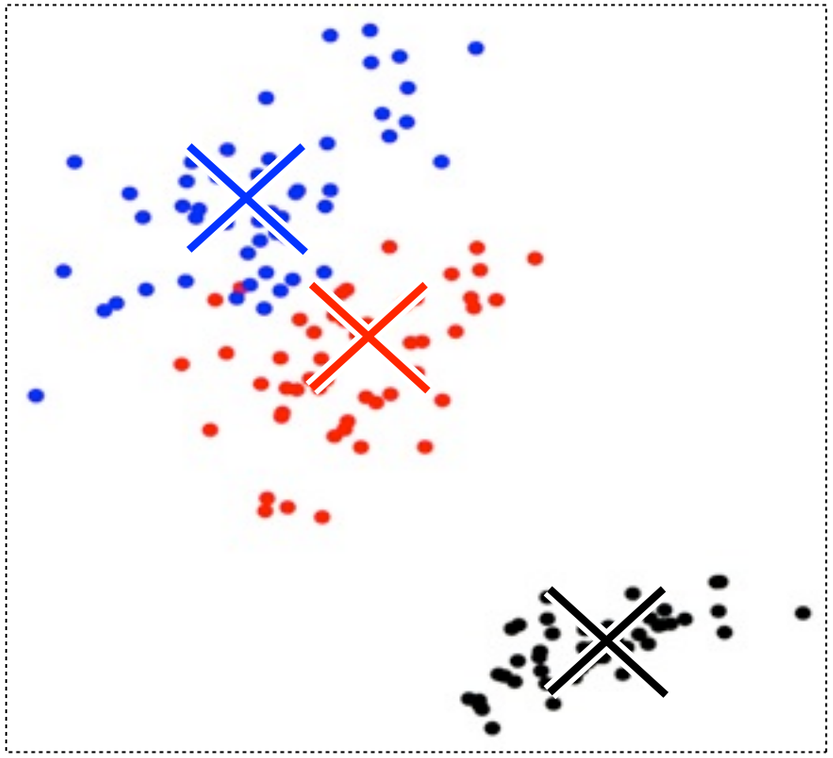


Centroid

Grid

Centroid:

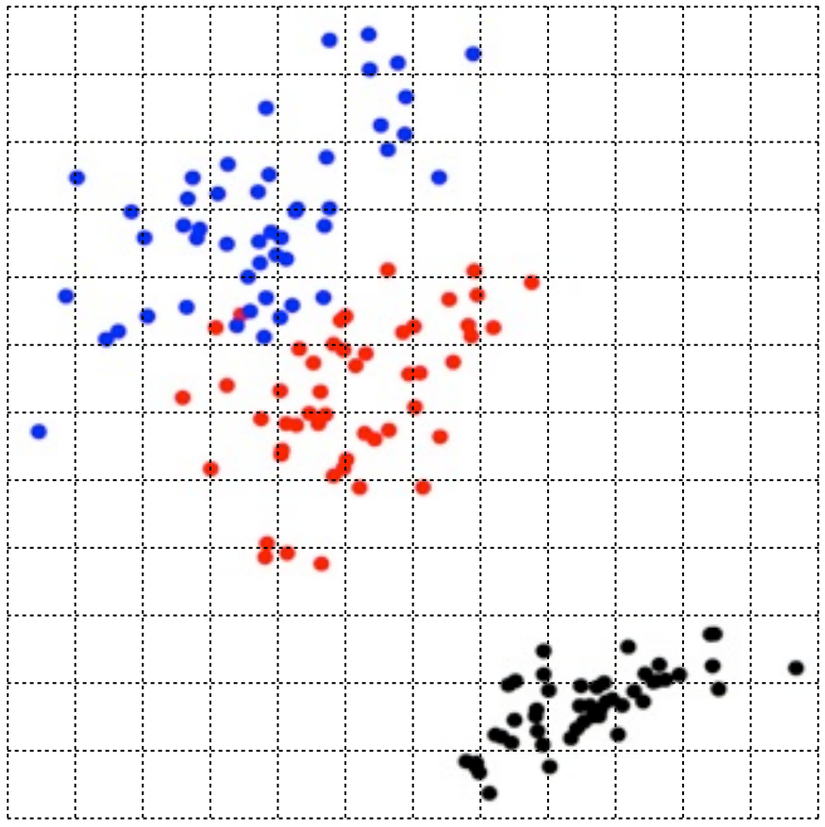
Mapping assumptions onto taxonomy axes



- 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

Related Work

They

Us

Scagnositcs

[Wilkinson 2005]

Mathematical depiction

Human perception

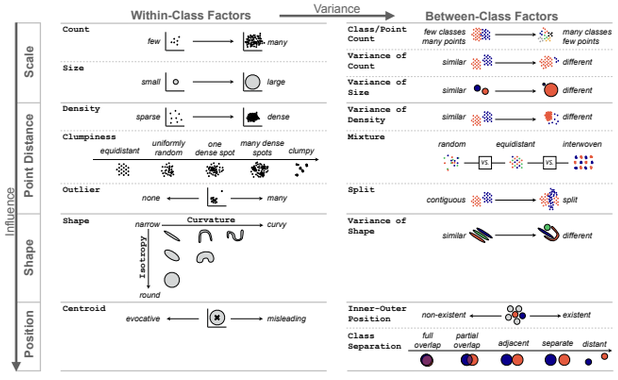
Gestalt principles

General capability

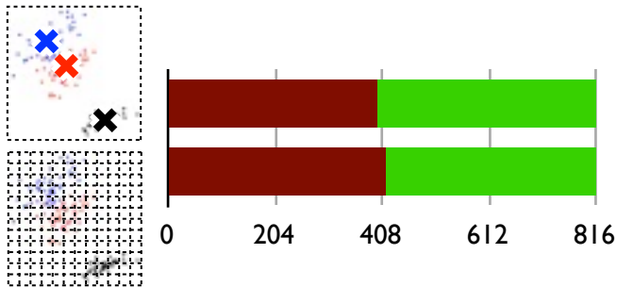
Specific guidance

Conclusion

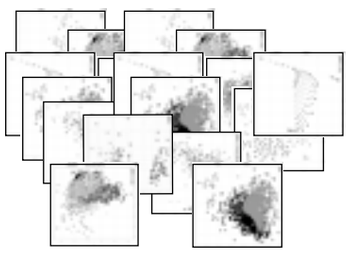
Contributions



Taxonomy of visual cluster separation factors



In-depth evaluation of 2 state-of-the-art separation measures



Qualitative data study

A Taxonomy of Visual Cluster Separation Factors

Michael Sedlmair,[1] Andrada Tatu[2], Tamara Munzner[1], Melanie Tory[3]
[1] Univ. of British Columbia, [2] Univ. of Konstanz, [3] Univ. of Victoria
http://www.cs.ubc.ca/labs/imager/tr/2012/VisClusterSep/