EuroVis 2018
Machine Learning Methods in Visualization for Big Data 2018
4 June 2018, Brno, Czech Republic

Machine Learning Methods in Visualisation for Big Data 2018

Daniel Archambault[1] Ian Nabney[2]
Jaakko Peltonen[3]

1 Swansea University

2 University of Bristol

3 University of Tampere, Aalto University

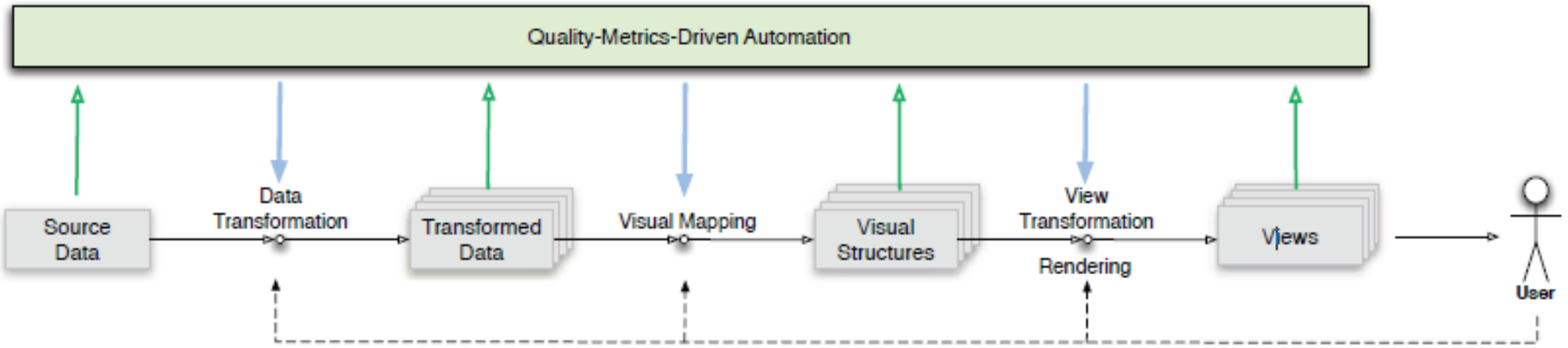
Evaluating Visualisation Techniques

- Why is evaluation difficult but important?
- User-based evaluation: perceptual/subjective evaluation, study design
- Metric-based evaluation
  - Model-based metrics and their limitations
  - Unsupervised learning metrics
  - Task-based metrics

Why is evaluation important?

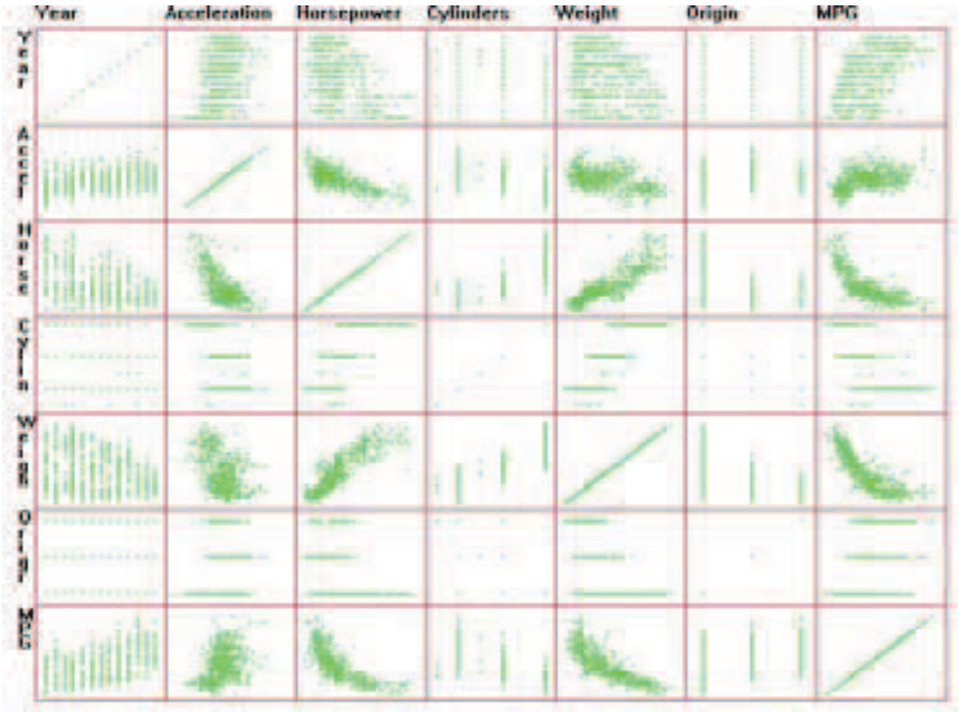
- There is a need to compare the ’quality’ of different visualisations
- This matters to algorithm developers to demonstrate the value (or lack of it) of new techniques
- This matters to practitioners since they are likely to generate multiple visualisations (parameter settings, different visualisation methods), sometimes in the thousands, and need to choose between them or guide the process of visualisation.
- This relates to automation or semi-automation of analysis process.
- We focus on the use of evaluation in high-dimensional data analysis

Visualization pipeline

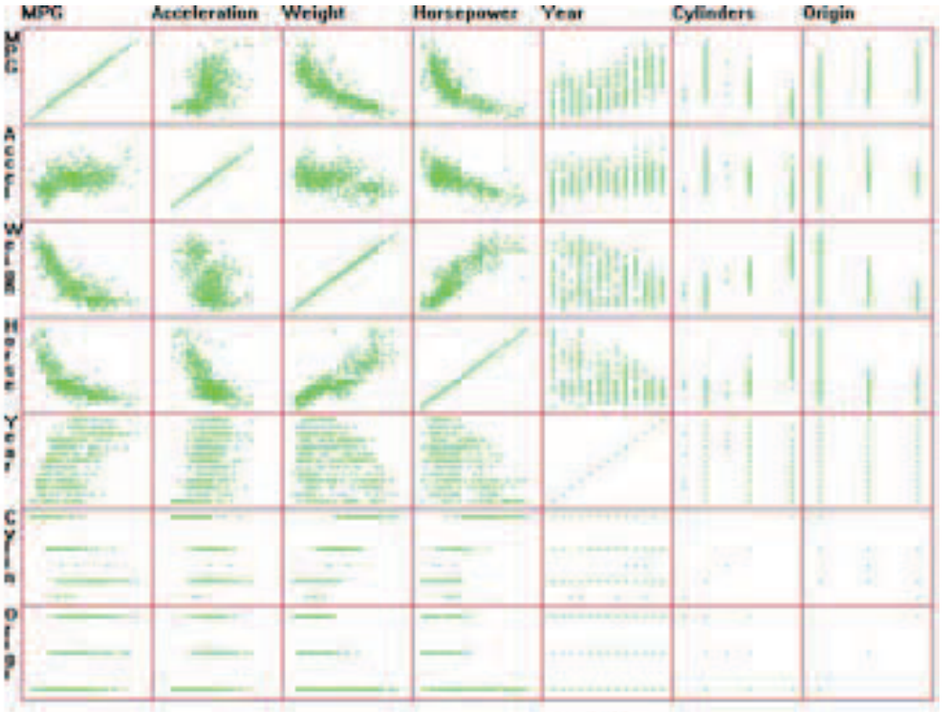


- Quality metrics can be calculated in the data space, image space or a combination of the two. Metrics calculated at the *View* stage draw information from the rendered image, whereas the others draw information from the data space and elements of the visual structures
- Quality metrics generate help evaluate alternatives.
- Metrics do not replace the user. Indeed, the whole purpose of data visualization is to aid a human user in data analysis.

How evaluation can guide visualisation



(a)



(b)

- Clutter reduction through axes reordering in a scatterplot matrix. (Peng et al. 2004).

Categories of usability

- Expressiveness and semantic quality of visualization: usability of visual representation
- Interface usability: interaction mechanisms
- Data usability: quality of data supporting users' task

Why is evaluation difficult?

- Dimensionality reduction is an inherently *unsupervised* task – there is no ’ground truth’ or ’gold standard’.
- There is no single globally correct definition of what quality means.
- There is a wide variety of dimensionality reduction methods with different assumptions and modelling approaches

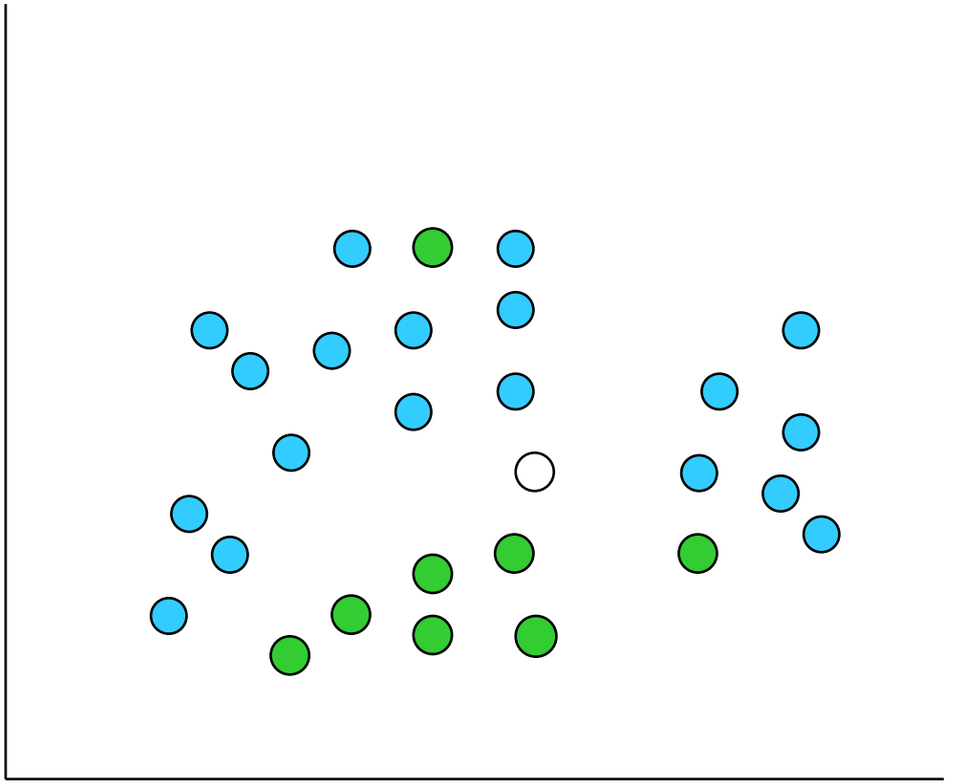
- How can we tell? Not many papers on the subject!

Semi-Supervised Models

- In a *supervised* task we know the outcome for each example (e.g. a class or continuous value) and we try to develop a model that can predict that outcome. Classification or regression
- In an *unsupervised* task we have data, but no variable represents a single outcome for each example and we try to develop a model that looks for groups in the data. Clustering or visualisation
- In some unsupervised tasks we want a target variable to influence the output: *semi-supervised* or *relative supervision*.

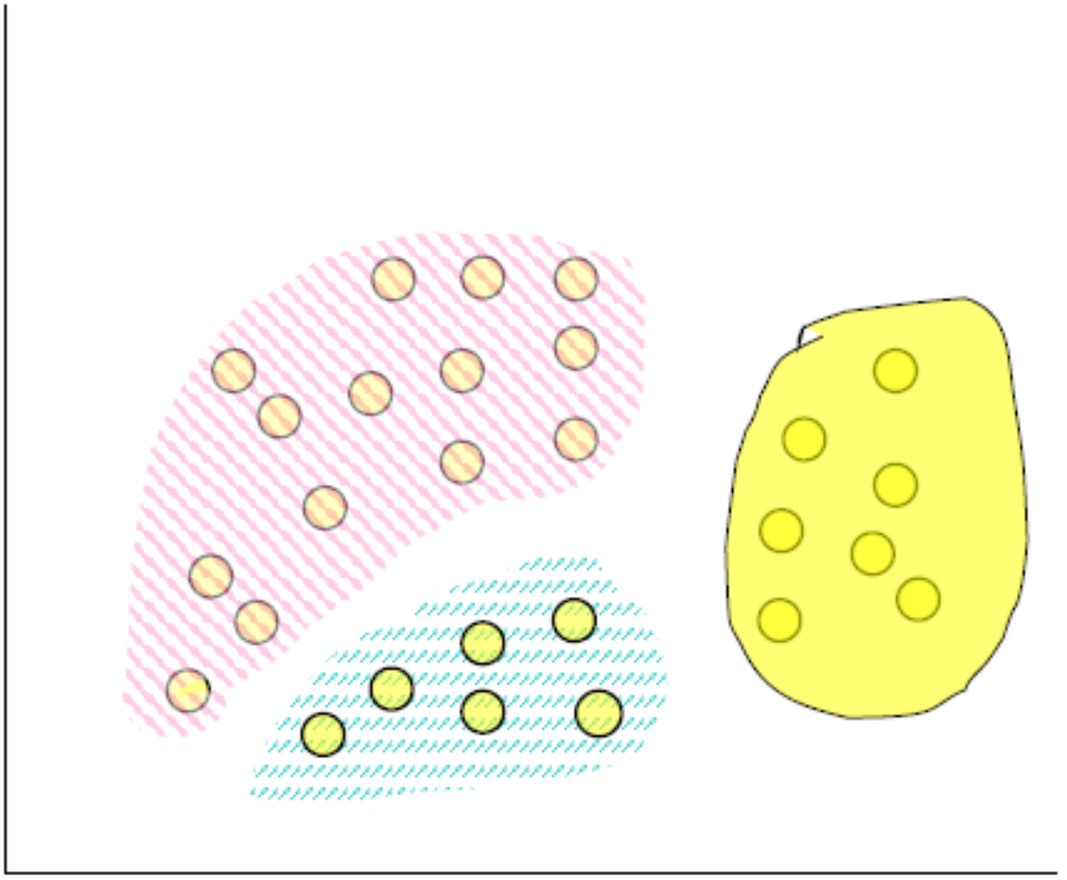
Supervised Task: Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



Unsupervised Models

- Find natural grouping of instances given unlabelled data



Models

- GTM
- GPLVM
- Neuroscale

Metric learning

- Many statistical methods rely on distances as much or more than they do on feature values:
  - nearest neighbor regression/classification uses distances to find the nearest neighbors
  - many clustering approaches such as k-means use distances as part of the algorithm to optimize the clustering
  - in information retrieval, “best” results are often the ones most similar to the query according to some distance
- Dimensionality reduction methods such as multidimensional scaling, Sammon mapping, Self-organizing maps, Stochastic Neigbor Embedding, Neighbor Retrieval Visualizer, and others are distance-based
- In many cases distances from a new distance function can be just plugged in to dimensionality reduction methods. (In some cases more is needed.)

Topographic Mappings

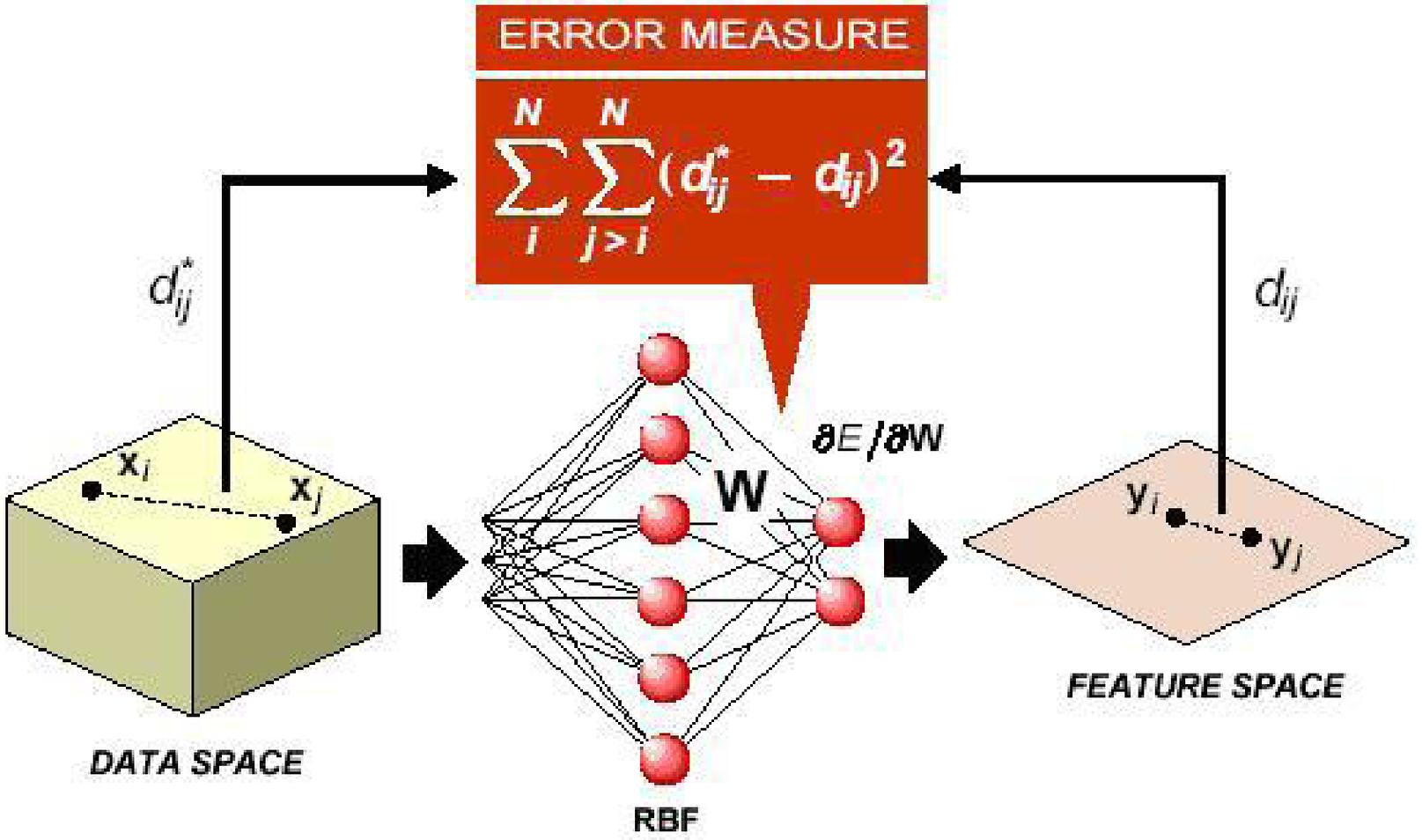
- Basic aim is that distances in the visualisation space are as close a possible to those in original data space.
- Given a dissimilarity matrix d_{ij}| we want to map data points x_i to points y _i in a feature space such that their dissimilarities $\tilde{d}_{ij}$ are as close as possible to d_{ij}|
- The map is said to preserve similarities. The stress measure is used as objective function.

$$E = \frac{1}{\sum_{ij} d_{ij}} \sum_{i<j} \frac{\left(d_{ij} - \tilde{d}_{ij}\right)^2}{d_{ij}}$$

Multi-dimensional Scaling

- Given distances or dissimilarities between every pair of observations try to preserve these as far as possible in lower-dimensional space.
- In classical scaling, the distance between the objects is assumed to be Euclidean. A linear projection then corresponds to PCA.
- The Sammon mapping is a non-linear multidimensional scaling technique more general (and more widely used) than classical scaling.
- Neuroscale is a neural network based scaling technique that has the advantage of actually giving a map that generalises!

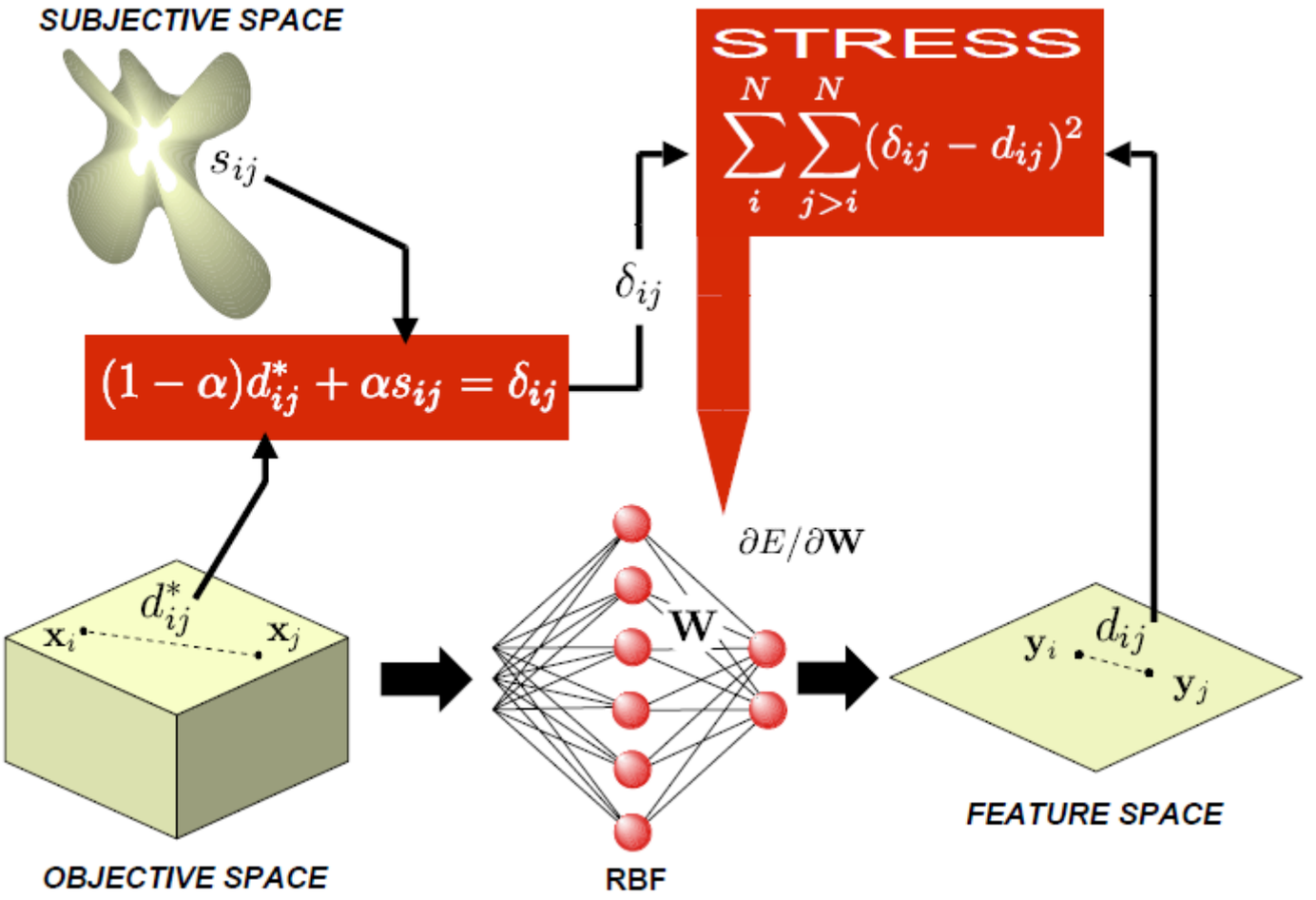
Neuroscale


ERROR MEASURE
$\sum_{i}^{N}\sum_{j>i}^{N}(d_{ij}^{*} - d_{ij})^2$
d_{ij}^{*}
d_{ij}
x_i
x_j
y_i
y_j
W
$\partial E / \partial W$
DATA SPACE
RBF
FEATURE SPACE

Subjective metrics

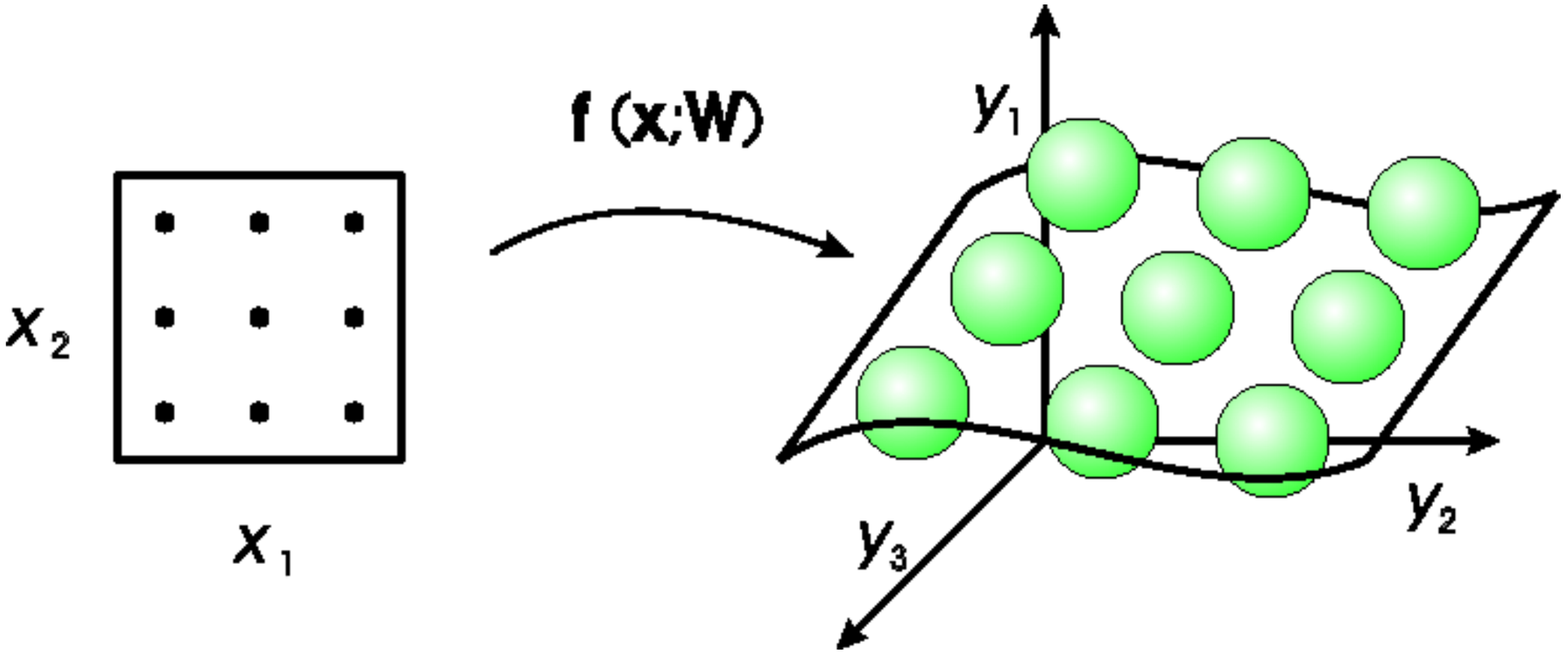
- Modify the stress measure:
- .
- $$E' = \sum_{i}^{N} \sum_{j}^{N} (\delta_{ij} - \| \mathbf{y}_i - \mathbf{y}_j \|)^2$$
- Inter-point distances for pairs of points in different classes are modified by the addition of some constant term k, such that their separation should be exaggerated in the resultant map.
- $$\delta_{ij} = \begin{cases} d_{ij}^* & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class,} \\ d_{ij}^* + k & \text{otherwise.} \end{cases}$$
- Other formulations are possible – can use a dissimilarity matrix for classes or distance for an auxiliary continuous variable. The relative weight of objective and subjective elements can be controlled by a parameter.

Neuroscale Operation



Generative Topographic Mapping

- Mapping from latent space to data space
- A thick rubber sheet studded with tennis balls. GTM defines p(y|x;W); use Bayes' theorem to compute p(x|y*;W) for a given point y* in data space.



Algorithm

- EM algorithm involves expectation (E-step): extend this to missing values as well as missing kernel.
- $$\hat{\mathbf{t}}_{nj}^{m} = \mathrm{E}(\mathbf{t}_n^m | z_{nj} = 1, \mathbf{t}_n^o, \theta_j) = (\mathbf{y}_j^m)^{old} + \Sigma_j^{mo} \Sigma_j^{oo^{-1}} (\mathbf{t}_n^o - (\mathbf{y}_j^o)^{old})$$
- GTM model uses spherical covariance, hence this inference is quite uninformative
- $$\hat{\mathbf{t}}_{nj}^{m} = (\mathbf{y}_j^m)^{old}$$
- Class membership (if available) can provide more accurate inference.
- $$r_{njc} = P(\mathbf{x}_j | \mathbf{t}_n^o, c_n) = \frac{P(\mathbf{x}_j, c_n | \mathbf{t}_n^o)}{\sum_{k=1}^{K} P(\mathbf{x}_k, c_n | \mathbf{t}_n^o)}$$

Evaluation depends on purpose

- Use visualization both for specific tasks but also data-driven hypothesis formation.
- Metric should measure what the user requires the visualization for:
  - Accurate representation of data point relationships (local/global)
  - Good class separation: ’clustering’ (many types)
  - Identification of outliers
  - Reduction of noise
  - ’Understanding’ data – perception of data characteristics
  - Choosing how to represent data

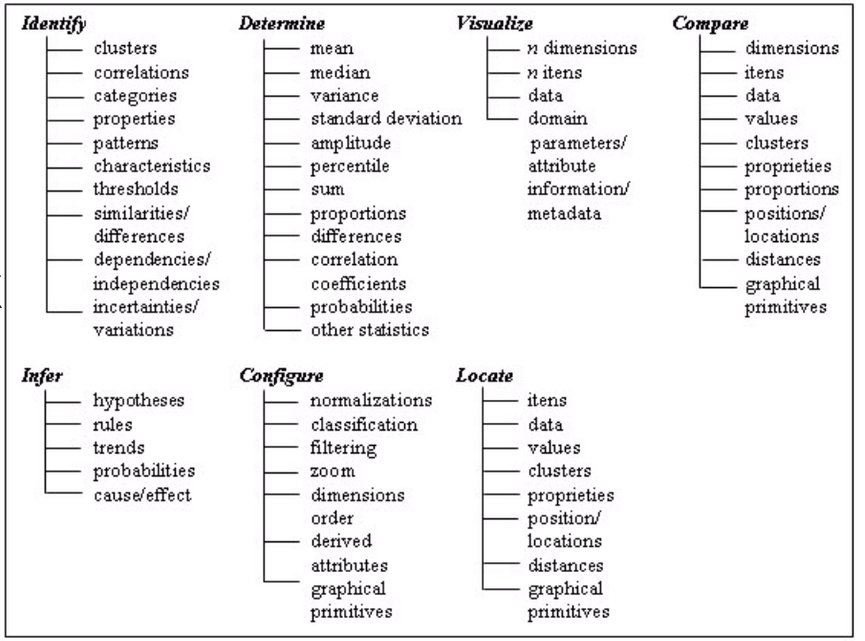
User-based evaluation

- Evaluating user performance
- Evaluating user experience

- How large is the search space?
- How many users do you need?
- Richer qualitative data (e.g. reaction cards, choosing cards/words to reflect UX)
- Humans not good at quantifying what they see (e.g. is one plot ’more structured’ than another?).

User study (Pillat, 2005)

- Compare parallel coordinates and Radviz
- Taxonomy of user tasks
- Car dataset from Statlib:
  - 392 records; 7 attributes
- Tasks: outliers, clusters, class (origin)
- Four questions and qualitative feedback
- Five graduate students in study
-



Study Set-up

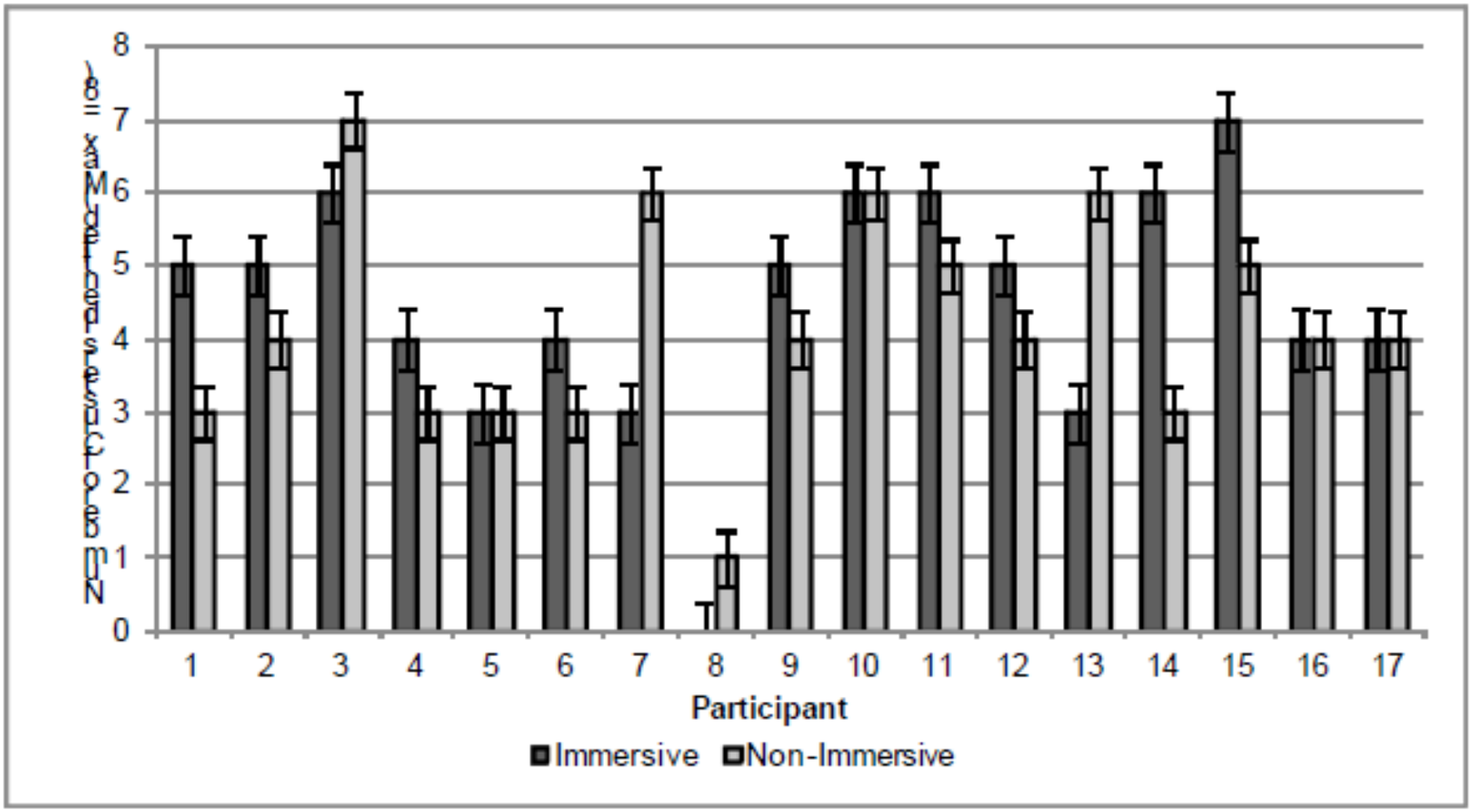
- A within-groups, counter-balanced study protocol in which each participant used both our immersive and our non-immersive visualisation environment to complete a set of prescribed tasks.
- The order in which participants were exposed to the different visualisation environments was counter-balanced to mitigate against the effects of learning – i.e., half of the participants used the immersive environment first and then the non-immersive environment, with the other half using the environments in the opposite order.
- Use of different seeds enabled us to generate non-identical datasets of similar properties.

Task Definition

- Participants were shown the first dataset and asked to identify – by entering the reference for the centre point – as many clusters as they believed were present. Described, in free text, the clustering within the dataset.
- Also asked to identify – again by reference – as many outliers as they believed were present in the data.
- It was left up to the participants to define/interpret what constituted a cluster and an outlier.
- Repeated for a second dataset.
- Participants also asked to identify changes between two datasets from four choices.
- Completed a paper-based NASA TLX1 questionnaire to reflect on the workload associated with the visualisation environment.
- 24 participants: 17 valid records.
- No difference in outlier detection.

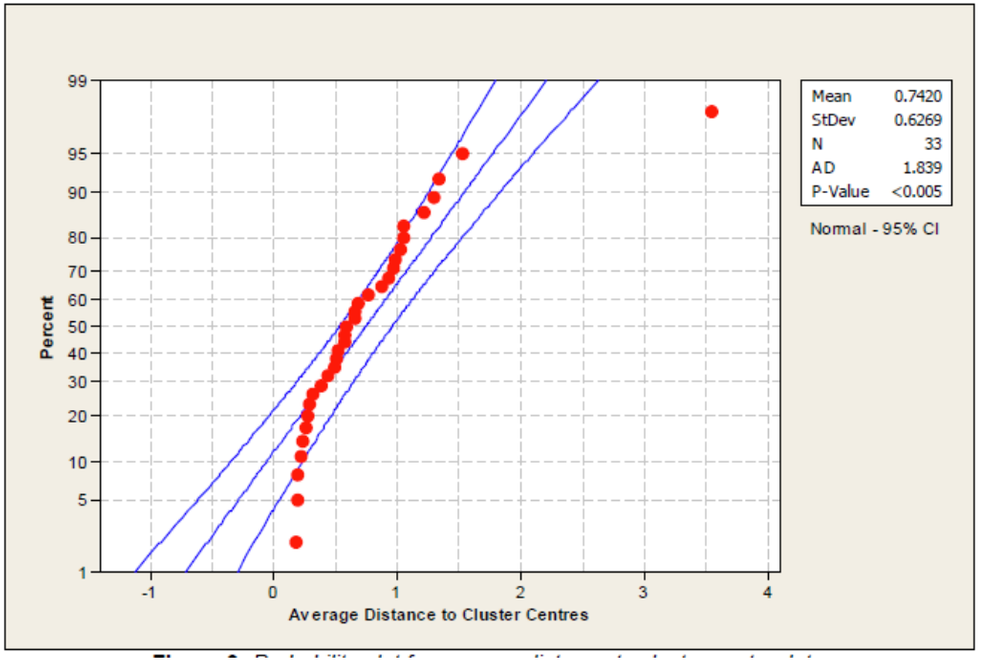
Cluster Identification

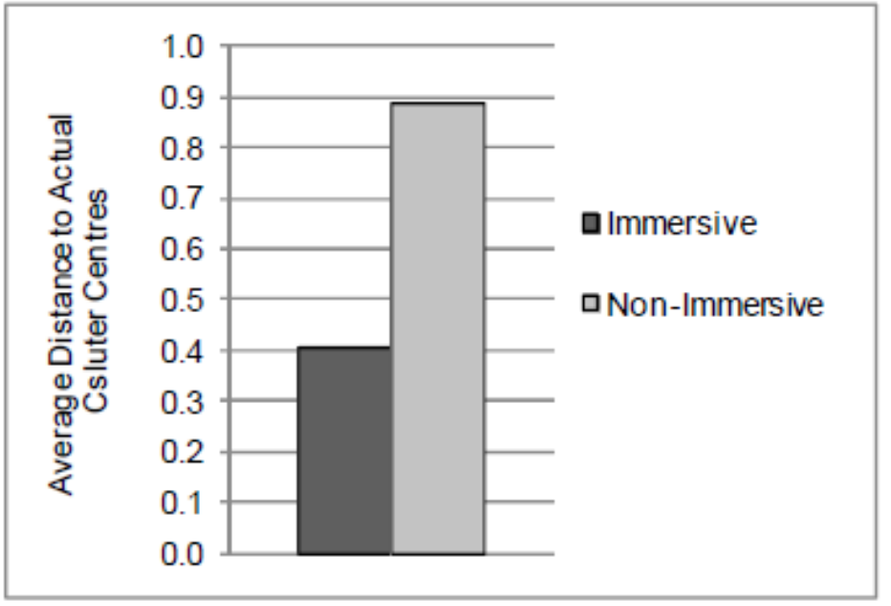
- 8 clusters: average identified in immersive was 4.5; non-immersive was 4.2.



Cluster Identification Accuracy

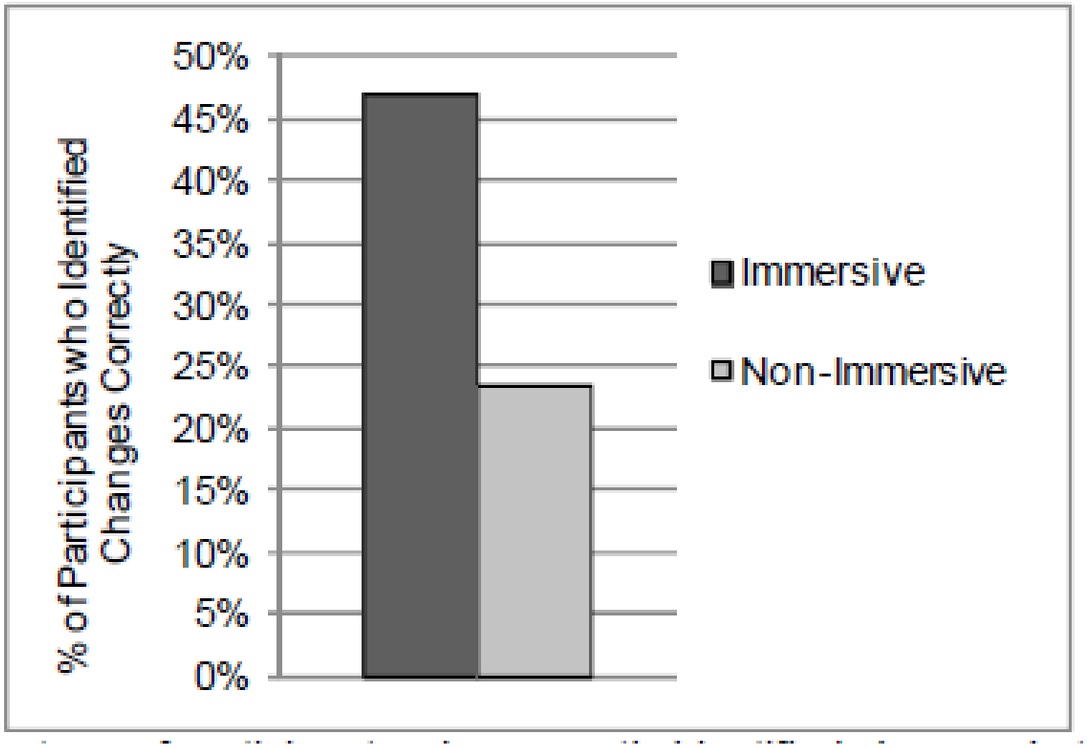
- Next considered average distance to cluster centre.
- Excluded outlier – rest has a normal distribution.
- Two-sample t-test was significant ($p<10^{-4}$).
- Being able to immerse oneself within a complex dataset increases one's ability to accurately identify data points relative to the spatial location of clusters.





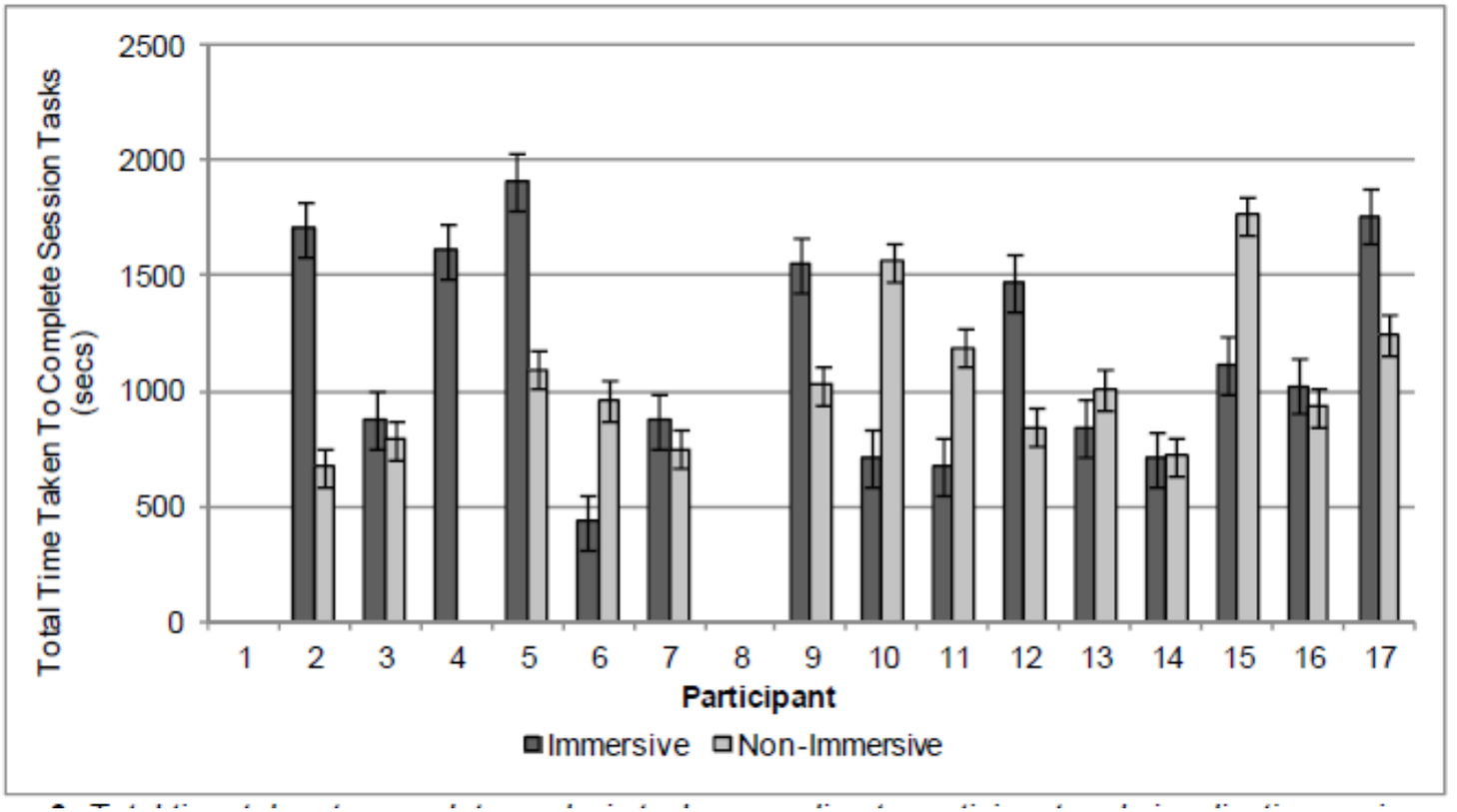
Dataset Changes

- Almost half (47.1%) of the participants using the immersive environment could accurately detect the changes in a dataset compared to less than one quarter (23.5%) of participants using the non-immersive environment.



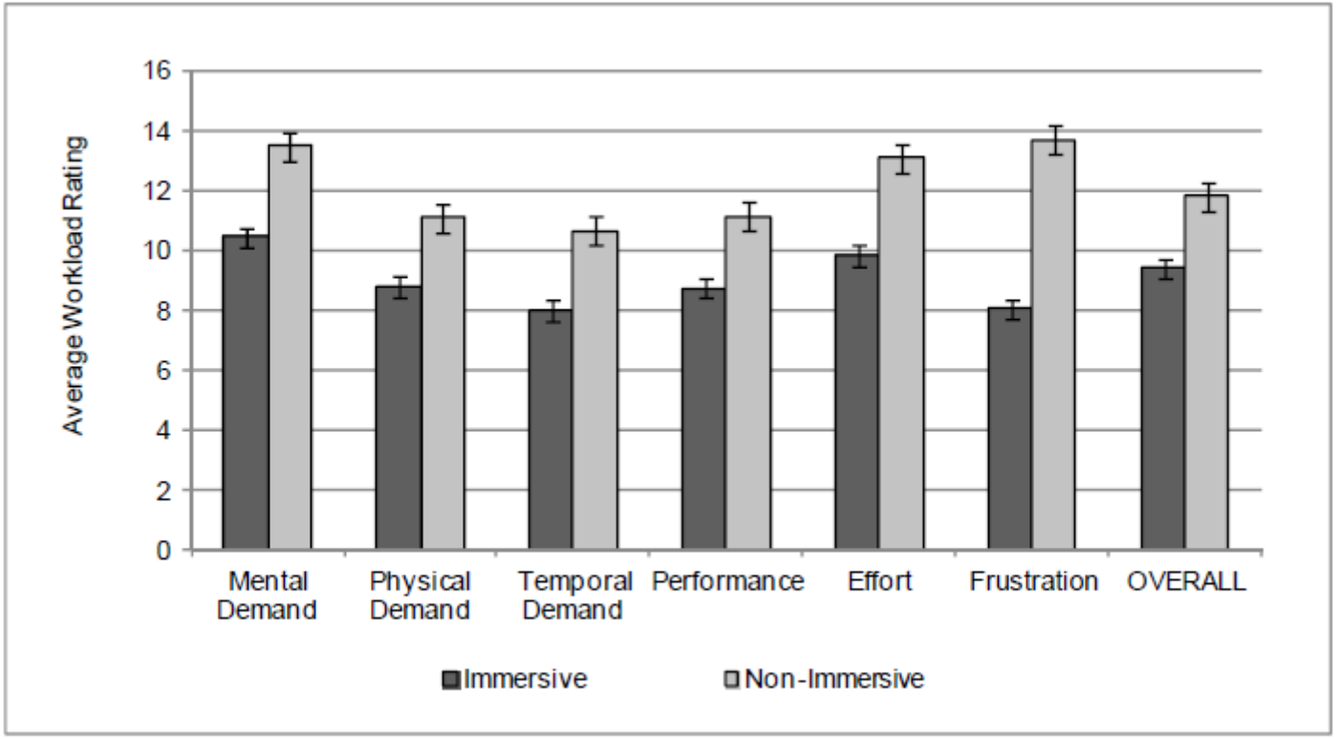
Speed of Analysis

- Took slightly longer to complete tasks using the immersive environment (average of 1117s compared to an average of 1039s), but this was not statistically significant (p=0.622).



Subjective Response

- Of those that responded, 82.6% stated a preference for the immersive environment,
- 73.9% stated that the immersive environment was more enjoyable,
- 90% stated that the immersive environment was more effective than the non-immersive environment.



Conclusions

- Immersive environment does not reduce the time taken to analyse complex sets of 3D data.
- It does show potential for supporting users to
  - achieve increased accuracy in data point identification/selection
  - increased ability to visually record and retain dataset patterns and to then accurately identify changes in the data.
- It also shows potential for
  - reducing the workload associated with complex 3D data analysis activities
  - eliciting a better subjective response from users – an important factor in attaining user acceptance and adoption of a technology.

Metric-based evaluation

- Model-based evaluation
- Unsupervised learning metrics
- Task-based metrics

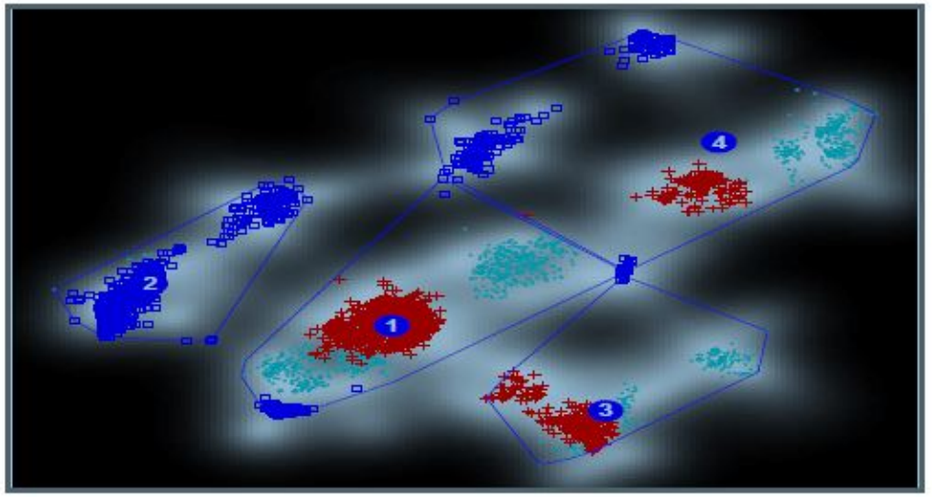
Model-based evaluation

- Most models have an associated cost function
  - Neuroscale, MDS: stress
  - PCA: variance
  - GTM, GPLVM, probabilistic PCA: log likelihood
- But:
  - Some models don’t have a cost function (SOM)
  - Cost functions are incompatible
  - Need some form of regularisation to compare different architectures/parameter spaces
- So, only enough when comparing a relatively narrow set of possible models

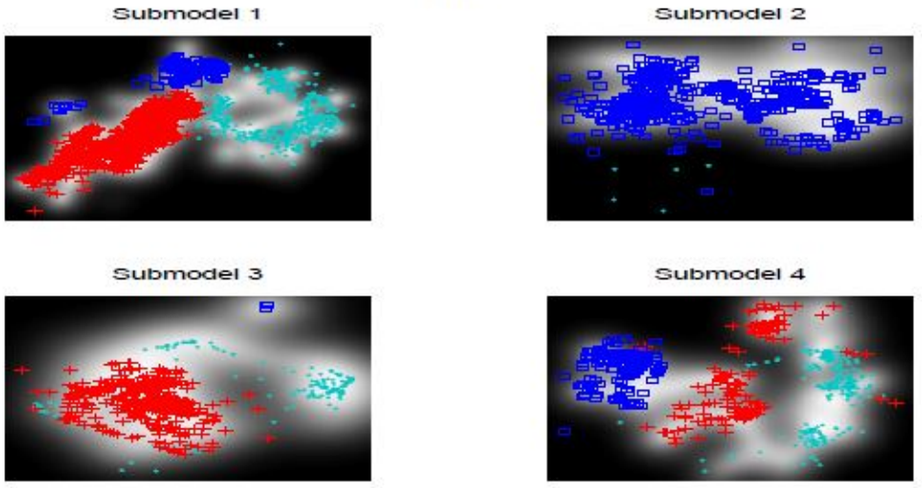
Unsupervised learning metrics

- Stress can always be calculated – but is it always relevant? (Alternative is to measure correlation of data and visualization space distances).
- Metrics that take account of local neighbourhood preservation
- Visualization Distance Distortion: compute *k* nearest neighbours for each point and measure the relative stress (or similar) for these points
- Trustworthiness measures the fraction of data points distant in the data space that become neighbours in the projection space
- Continuity measures the fraction of neighbouring data points in the data space become distant in the projection space.
- Mean relative ranks in data and latent space are similar, but consider weighted rankings of neighbourhood points.

Example: hierarchical visualization



(a)



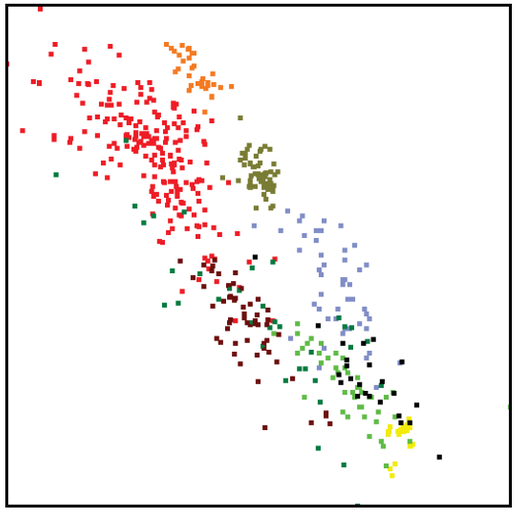
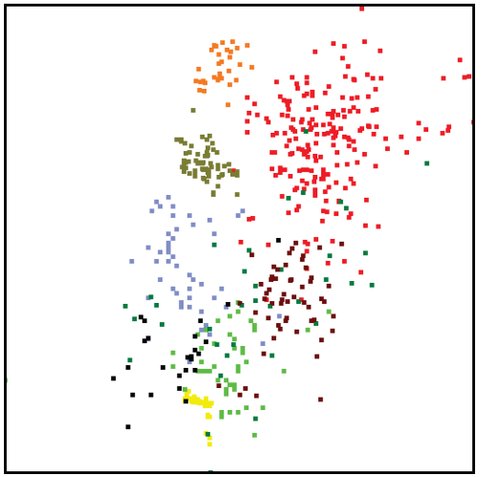
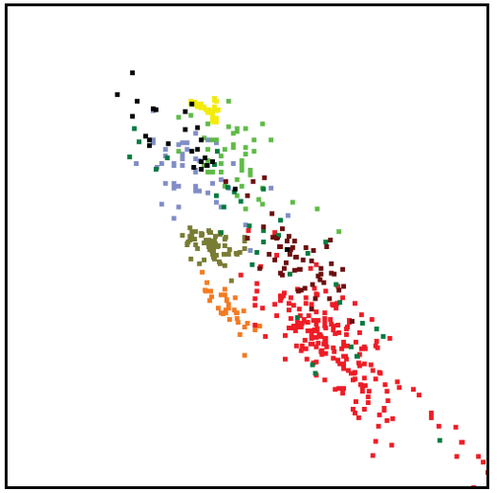
(b)

Clustering		Trustworthiness		Continuity	
		L1	L2	L1	L2
K-means	C1	0.8136	**0.8532**	0.8347	**0.8656**
	C2	0.6731	**0.7329**	0.7107	**0.7185**
	C3	0.7740	**0.7824**	0.7818	**0.7569**
	C4	0.8972	**0.9079**	0.8817	**0.8943**
GMM	C1	0.8132	**0.8605**	0.8361	**0.8519**
	C2	0.6738	**0.7212**	0.7111	**0.7375**
	C3	0.7737	**0.8190**	0.7819	**0.8478**
	C4	0.8979	**0.9132**	0.8799	**0.9252**
Interactive	C1	0.8136	**0.8610**	0.8362	**0.8856**
	C2	0.6731	**0.7329**	0.7107	**0.7185**
	C3	0.7828	**0.7938**	0.7891	**0.8367**
	C4	0.9034	**0.9095**	0.8858	**0.8953**

Task-based metrics

- Objective metrics based on task to be performed, but without the use of trials or subjective experiments
- These metrics often work best in a semi-supervised way: with additional class information
- For example – if we want the visualization to tell us about class separation, need a measure of class separation in the visualization space
- If we want the visualization to preserve class information, use the nearest-neighbour classification error (in visualization space) normalised by value in data space

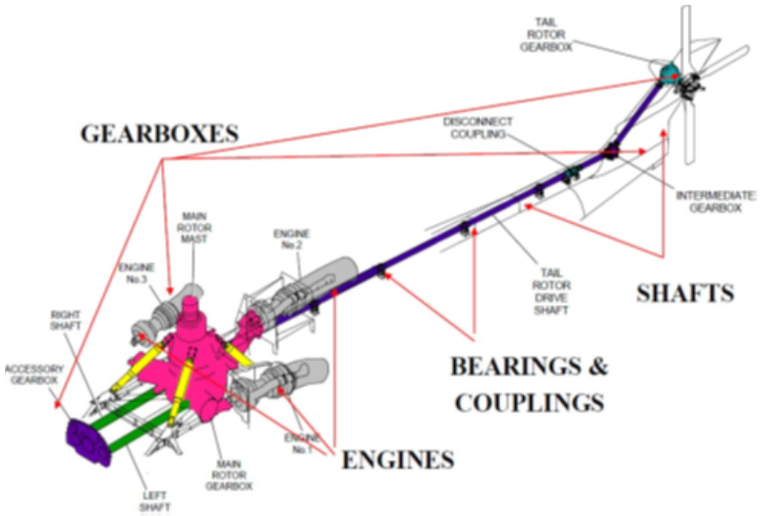
Cluster quality

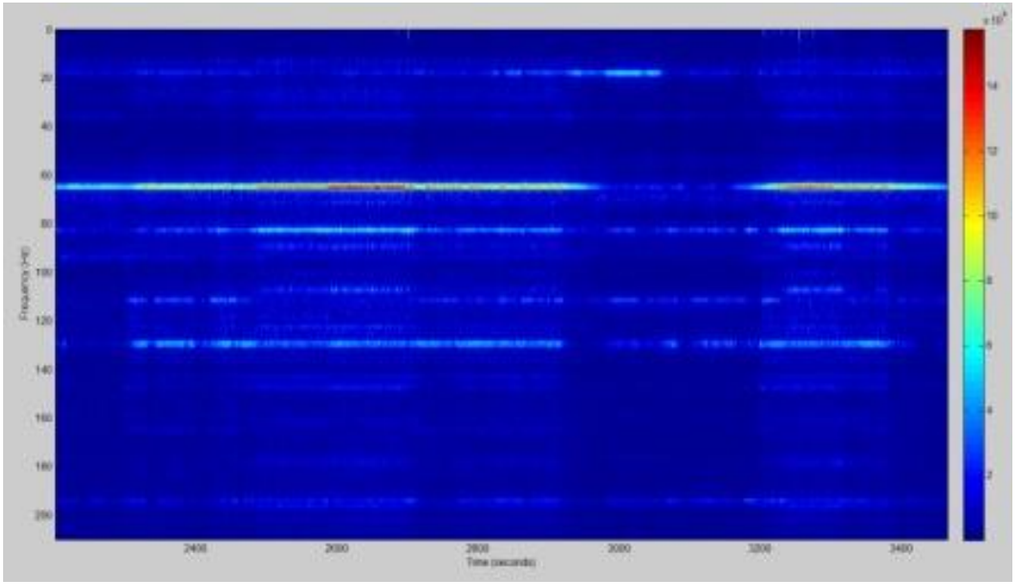




- Ranking projections according to class density measure favouring projections with minimal overlap between classes: image processing algorithm to detect clusters. (Tatu et al. 2009)
- Detection of clusters using image processing is highly non-trivial

Agusta Westland: airframe monitoring

- 8 sensors measuring vibration; 108 frequency bands

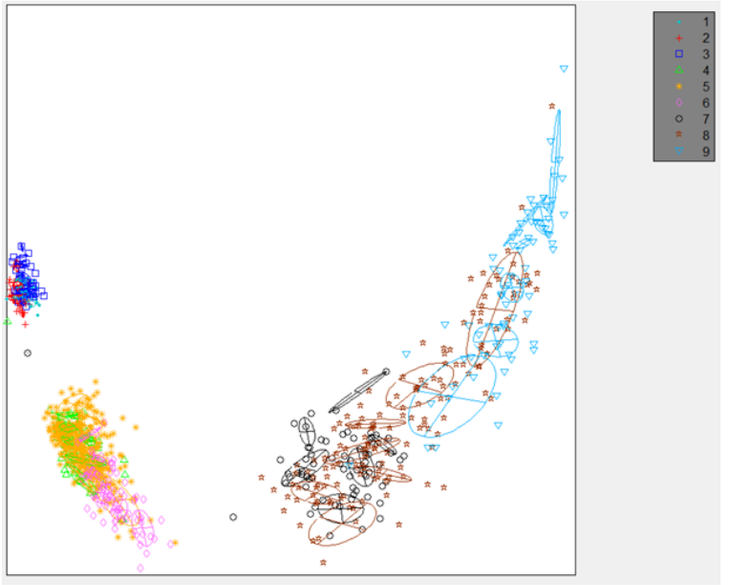




Metrics

- Known classes corresponding to flight modes
- Fit a Gaussian mixture model (GMM) to each class in visualization space. Use a variational Bayesian to automate model complexity
- Compute Kullback-Leibler distance between all possible class pairs GMMs as a measure of overall class separation.

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$



Conclusions

- Challenging because no ground truth
- Task-based metrics acknowledge the purpose of visualization, but tend to be harder to make objective
- Important to study the correlation between task performance metrics (as carried out by humans) and quantitative task-based metrics
- Common question asked by practitioners, so important to make progress

References

- E. Bertini, A. Tatu, and D. A. Keim, Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, *IEEE Trans. on Visualization and Computer Graphics,* **17**, 2203-2212, 2011.
- D. A. Keim, H-P. Kriegel, Visualization Techniques for Mining Large Databases: A Comparison, *IEEE Trans. on Knowledge and Data Engineering*, **8**, 923-938, 1996.
- R. Etemadpour, R. Motta, J. G. S. Paiva, R. Minghim, M. C. F. De Oliveira, L. Linsen, Perception-Based Evaluation of Projection Methods for Multidimensional Data Visualization, *IEEE Trans. on Visualization and Computer Graphics*, **21**, 81-94 , 2015.