LIFTING THE CURSE OF DIMENSIONALITY

A DASHBOARD FOR HIGH DIMENSIONAL DATA VISUALIZATION

MACHINE LEARNING IN PRACTICE

Analysis Pipeline:

- Collect, clean, explore data
- Process data to develop features and select model
- Choose a model that uses features to answer a question

The Problem with Big Data:

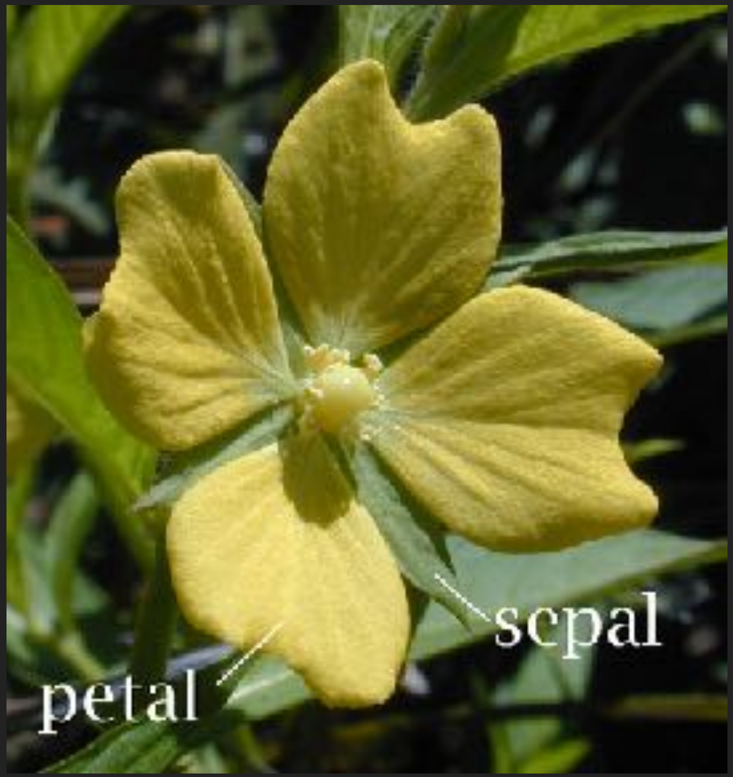
- Haphazard collection. Interferes with data exploration stage: what is relevant? What questions can the dataset answer?
- Classifier performance: as features/dimensions are added, similarity become harder to measure mathematically (the distance between features becomes HUGE)

GOOD PREDICTION, HARD INTERPRETATION?

Just Use “Black Box” Methods?

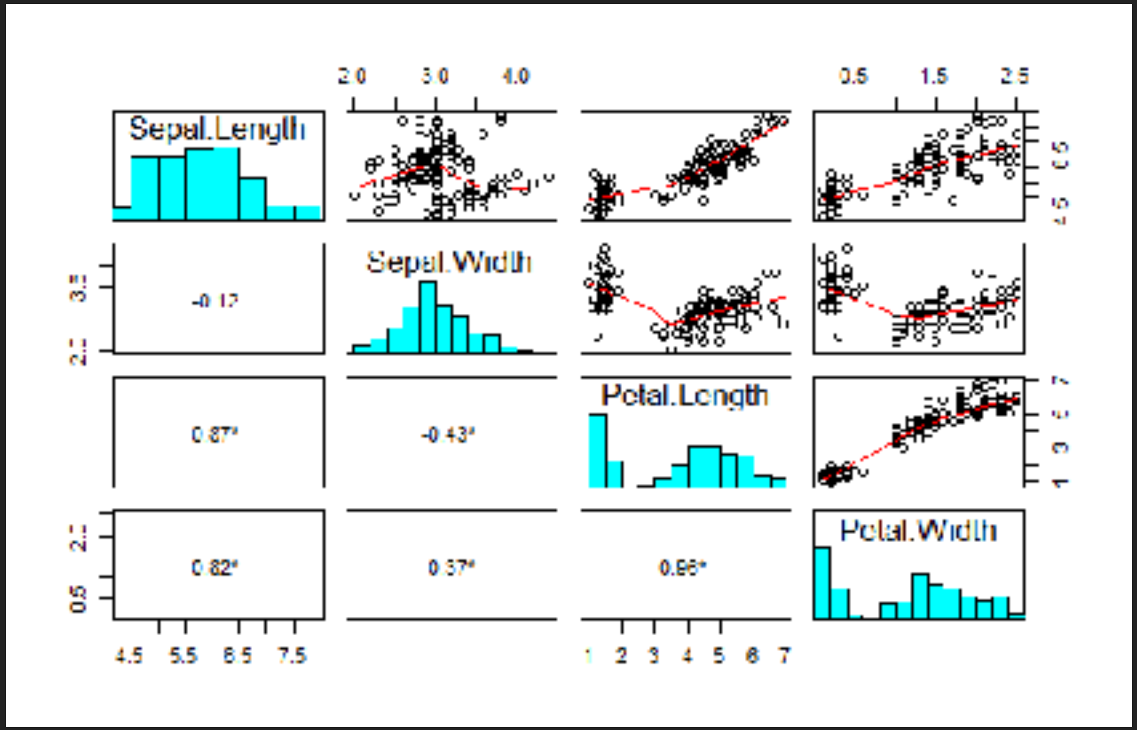
- Like: Neural Networks, Random Forests
- Great for classifying/predicting
- Not good enough for science: we lose the ability to interpret the model
- Not good enough for small data: don’t work as well as hand-crafted models
- e.x.: predict patient visual stimulus using fMRI data. Good prediction with opaque model doesn’t answer questions of what part of brain is most active

PREVIOUS APPROACHES: CLASSICAL STATISTICS

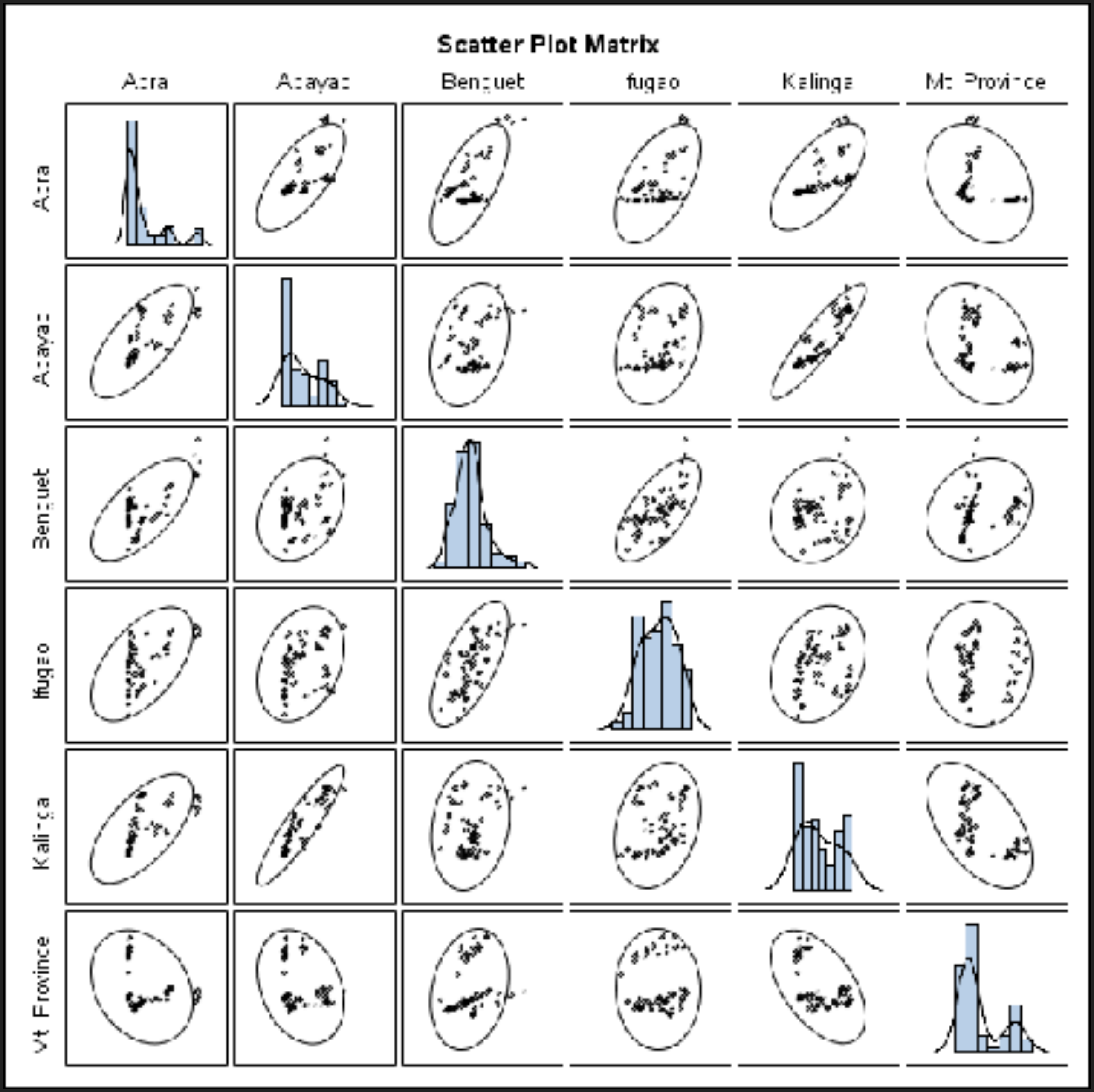


Statisticians have been dealing with the data exploration problem for at least a century. Their tools:

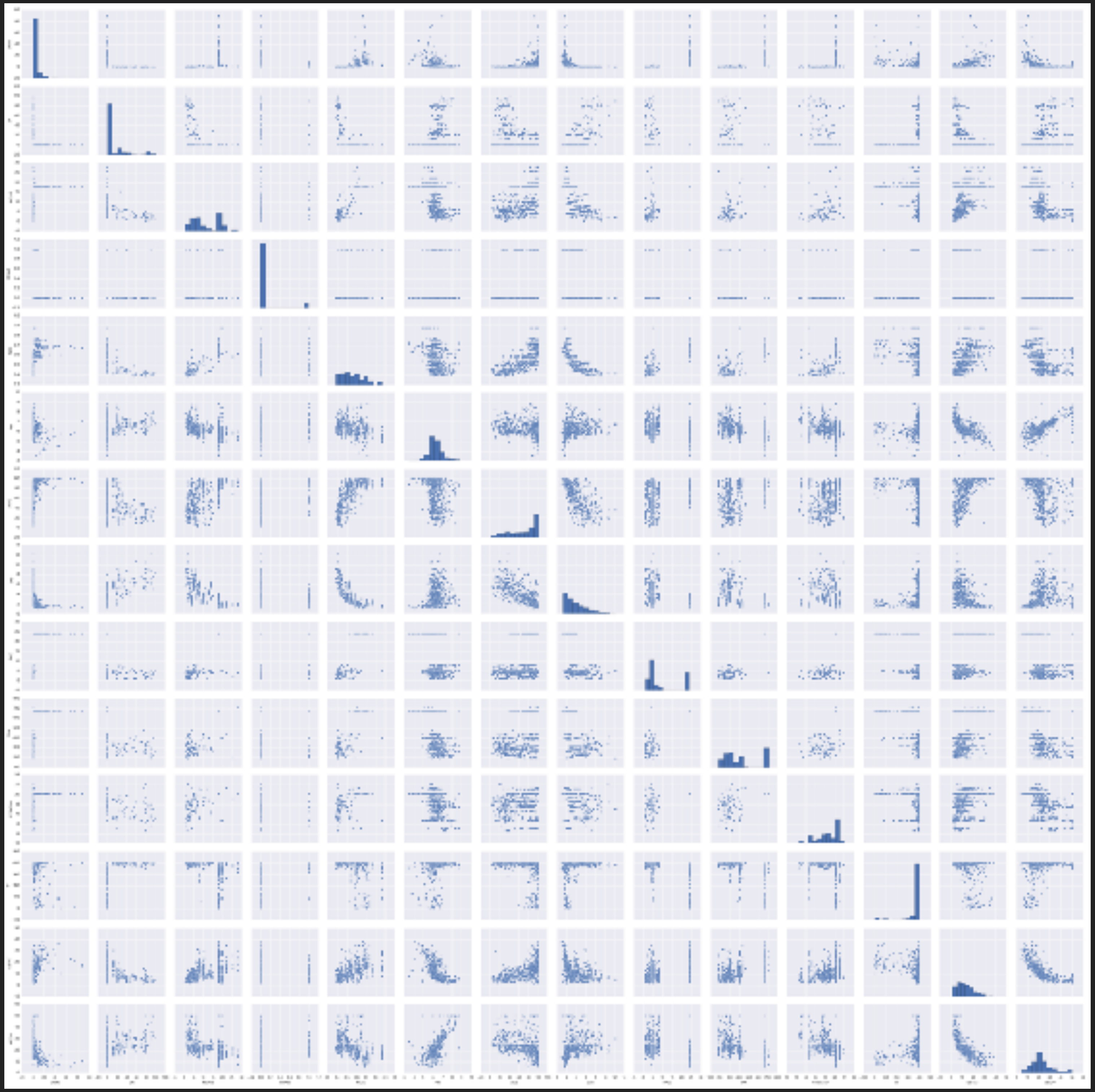
- Scatterplot matrices
  - Indispensable for finding pairwise relationships
- Histograms
  - Expose distribution of a particular feature
- The problem: suitable only for low dimensions



UP TO 9 DIMENSIONS, STILL USEFUL GIVEN MODIFICATIONS:

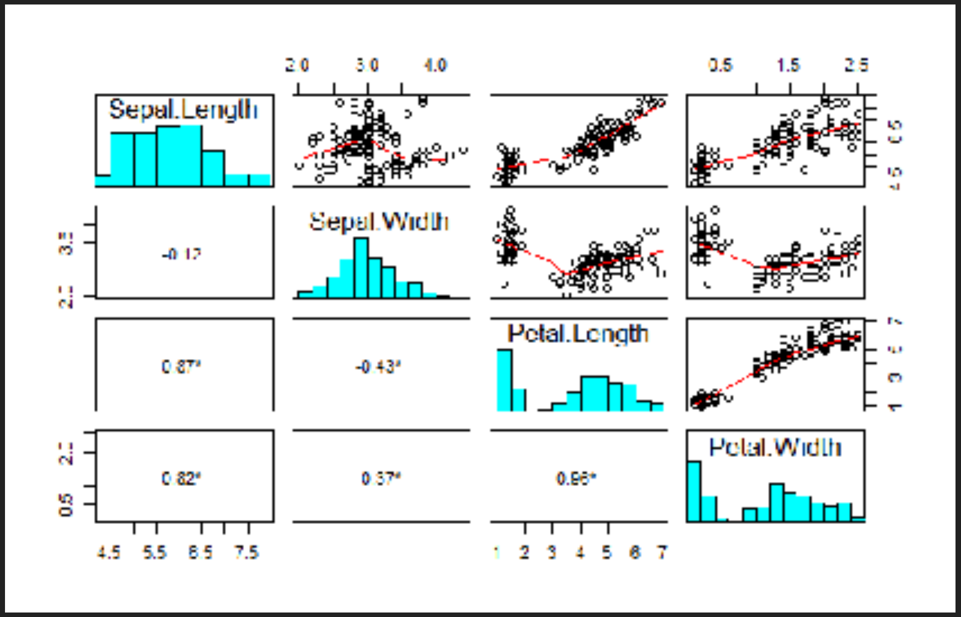


MORE THAN 9 DIMENSION, USELESS FOR EXPLORATION:



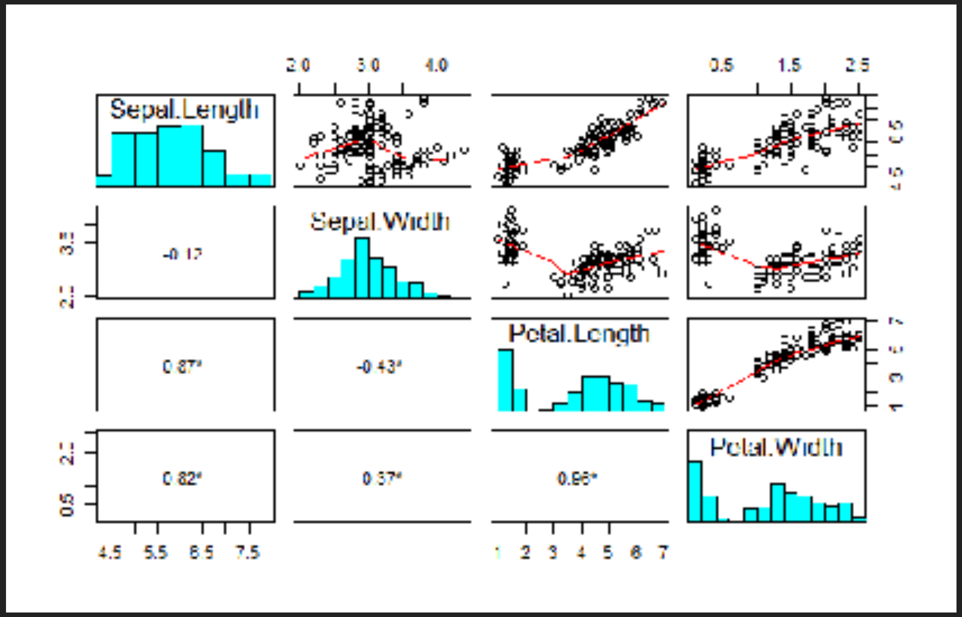
MACHINE LEARNING NEEDS NEW VISUALIZATION TOOLS

- No Free Lunch theorem
  - Model exploration is a larger concern
  - Tools like d3.js allow dynamic exploration of complex data
  - But: high dimensionality of data interferes with naive dynamic visualizations
  - My contribution: a visualization dashboard that combines algorithms for dimensionality reduction with visualization



MACHINE LEARNING NEEDS NEW VISUALIZATION TOOLS

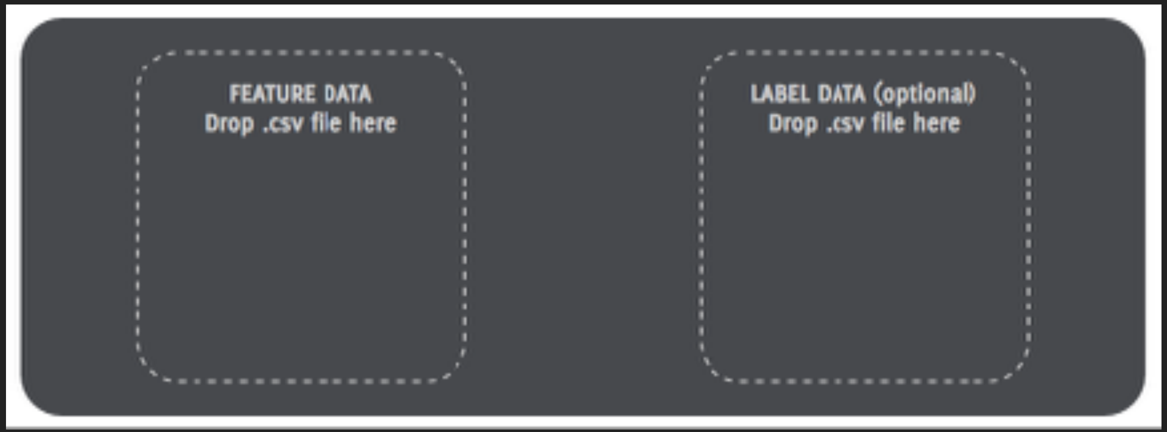
- No Free Lunch theorem
  - Model exploration is a larger concern
  - Tools like d3.js allow dynamic exploration of complex data
  - But: high dimensionality of data interferes with naive dynamic visualizations
  - My contribution: a visualization dashboard that combines algorithms for dimensionality reduction with visualization

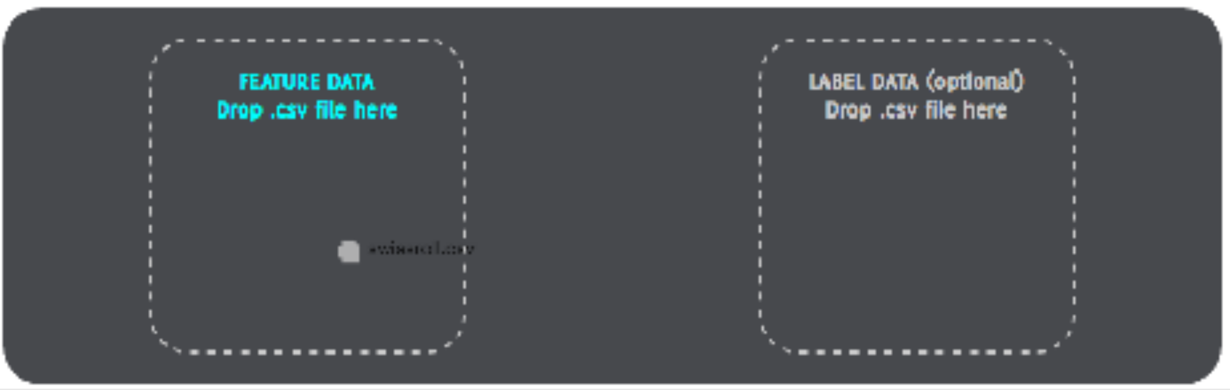


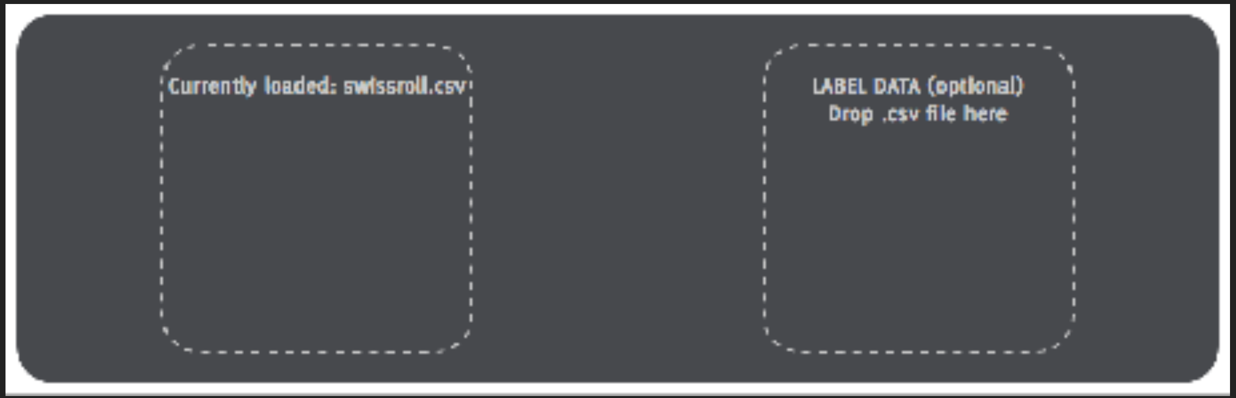
VISUALIZATION DASHBOARD: DATA DROP WIDGET

DRAG-AND-DROP DATA FILE INTO APPLET

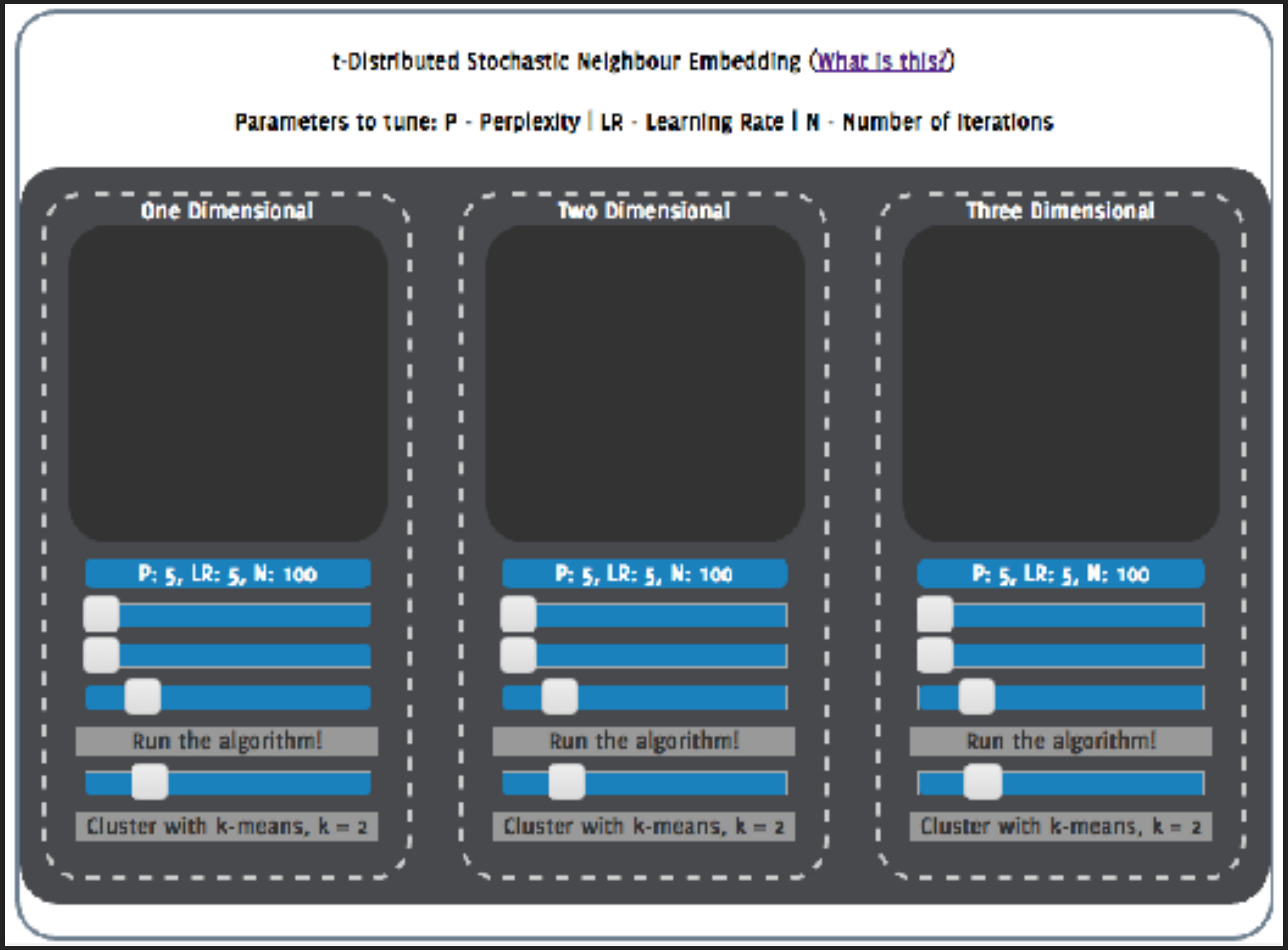
- Each time new comma-separated file is dragged into interface, the plots are cleared
- Optional labels file for pre-labelled data
- Highlight animation when file is dragged over top
- Shows file name when loaded






Currently loaded: swissroll.csv

VISUALIZATION WIDGETS: REDUCE AND CLUSTER



t-Distributed Stochastic Neighbour Embedding (What is this?)

Parameters to tune: P - Perplexity | LR - Learning Rate | N - Number of Iterations

One Dimensional

P: 5, LR: 5, N: 100

Cluster with k-means, k = 2

Two Dimensional

P: 5, LR: 5, N: 100

Cluster with k-means, k = 2

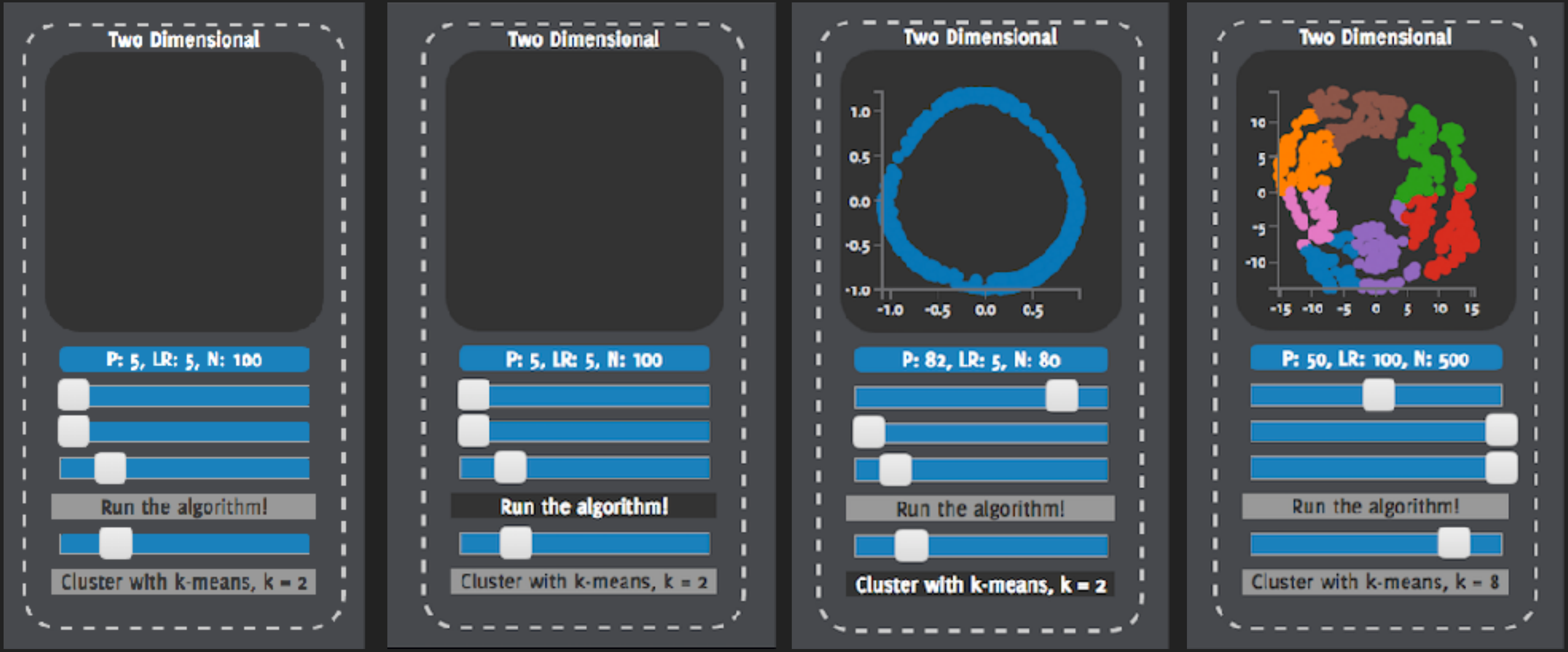
Three Dimensional

P: 5, LR: 5, N: 100

Cluster with k-means, k = 2

- Each dimensionality reduction algorithm has its own interface widget
- 3 divisions per algorithm for each of 3 possible visualization for reduction
- Each reduction can be tuned individually and compared against other three

VISUALIZATION WIDGET: ACTIONS



P: 5, LR: 5, N: 100

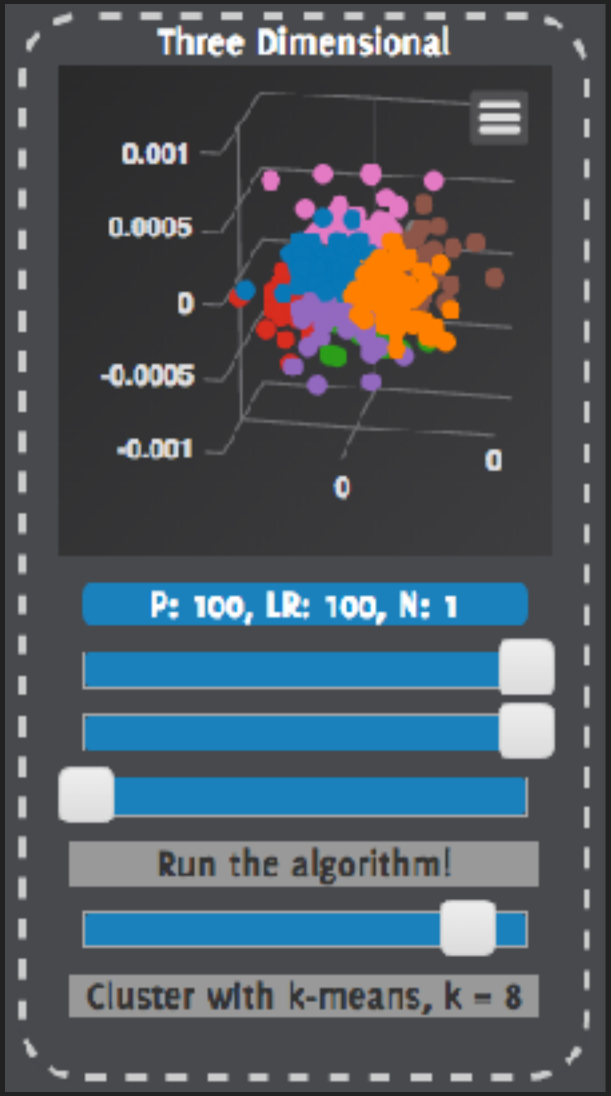
P: 5, LR: 5, N: 100

P: 82, LR: 5, N: 80

P: 50, LR: 100, N: 500

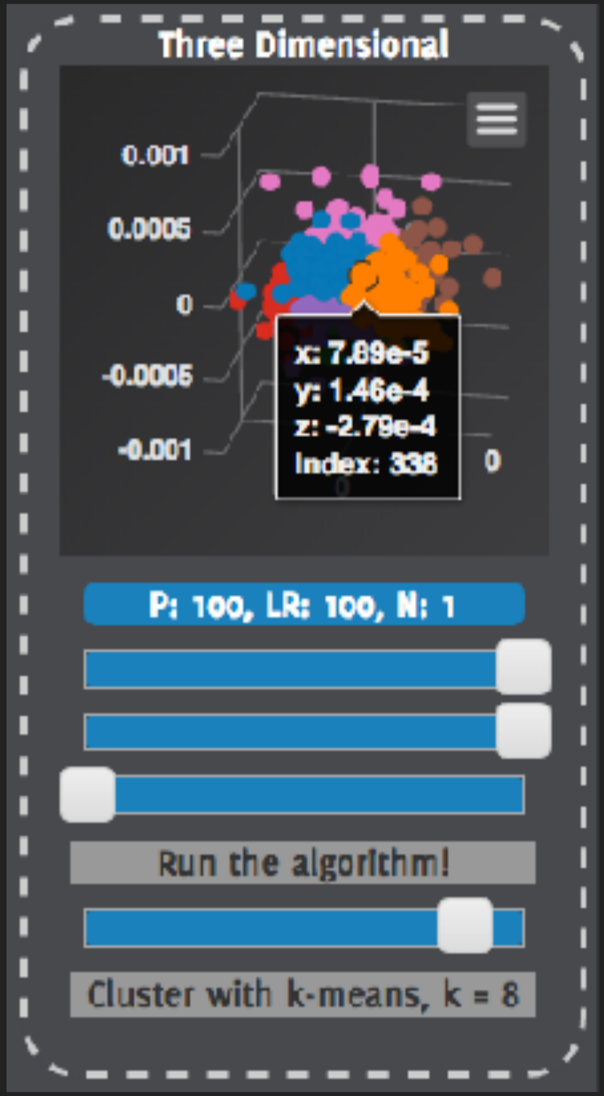
- Tunable parameters are depicted below the plotting region
- Individual sliders are used to select the values of the parameters
- Buttons invert color to show the possibility of actions
- k-means clustering is available at the bottom to find relationships using similarity by distance

VISUALIZATION WIDGET: 3D DATA EXPLORATION



P: 100, LR: 100, N: 1

Cluster with k-means, k = 8



x: 7.89e-5
y: 1.46e-4
z: -2.79e-4
Index: 338

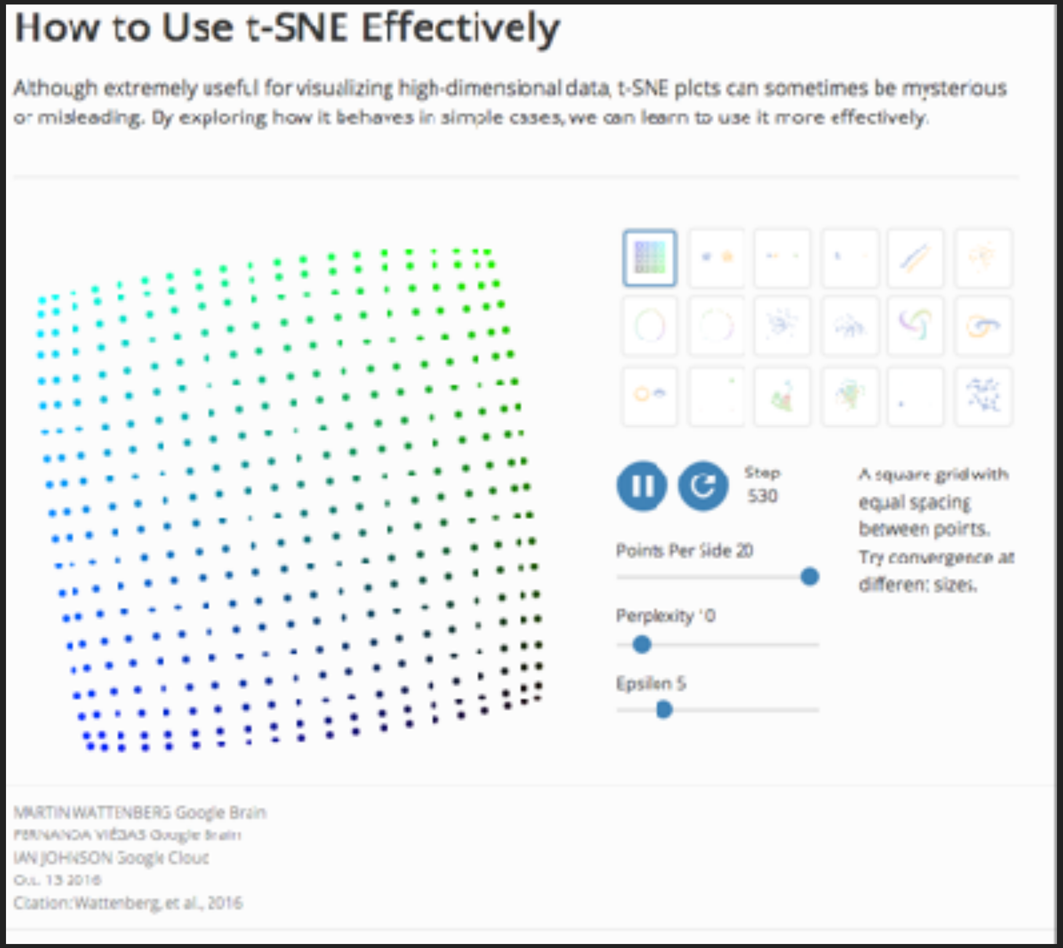
P: 100, LR: 100, N: 1

Cluster with k-means, k = 8

- Users can cluster using k-means in 3 dimensions, then pan and rotate around the data points
- Tooltips appear with information about the point location and its original index in the dataset
- This allows users to identify points that are similar in high dimensions, as candidates for further exploration during feature selection

VISUALIZATION DASHBOARD: DATA DROP WIDGET

VISUALIZATION WIDGET: HYPERLINKS TO MORE INFORMATION



How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.

Step 530

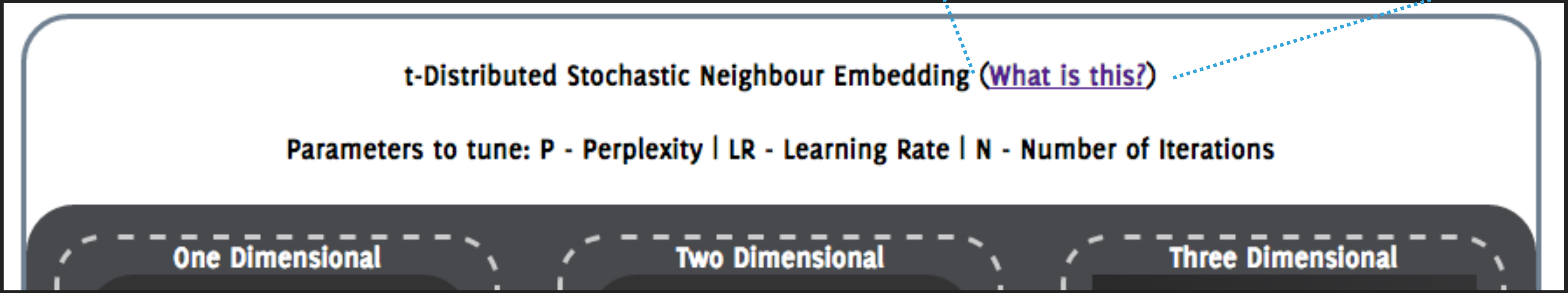
Points Per Side 20

Perplexity 10

Epsilon 5

A square grid with equal spacing between points. Try convergence at different sizes.

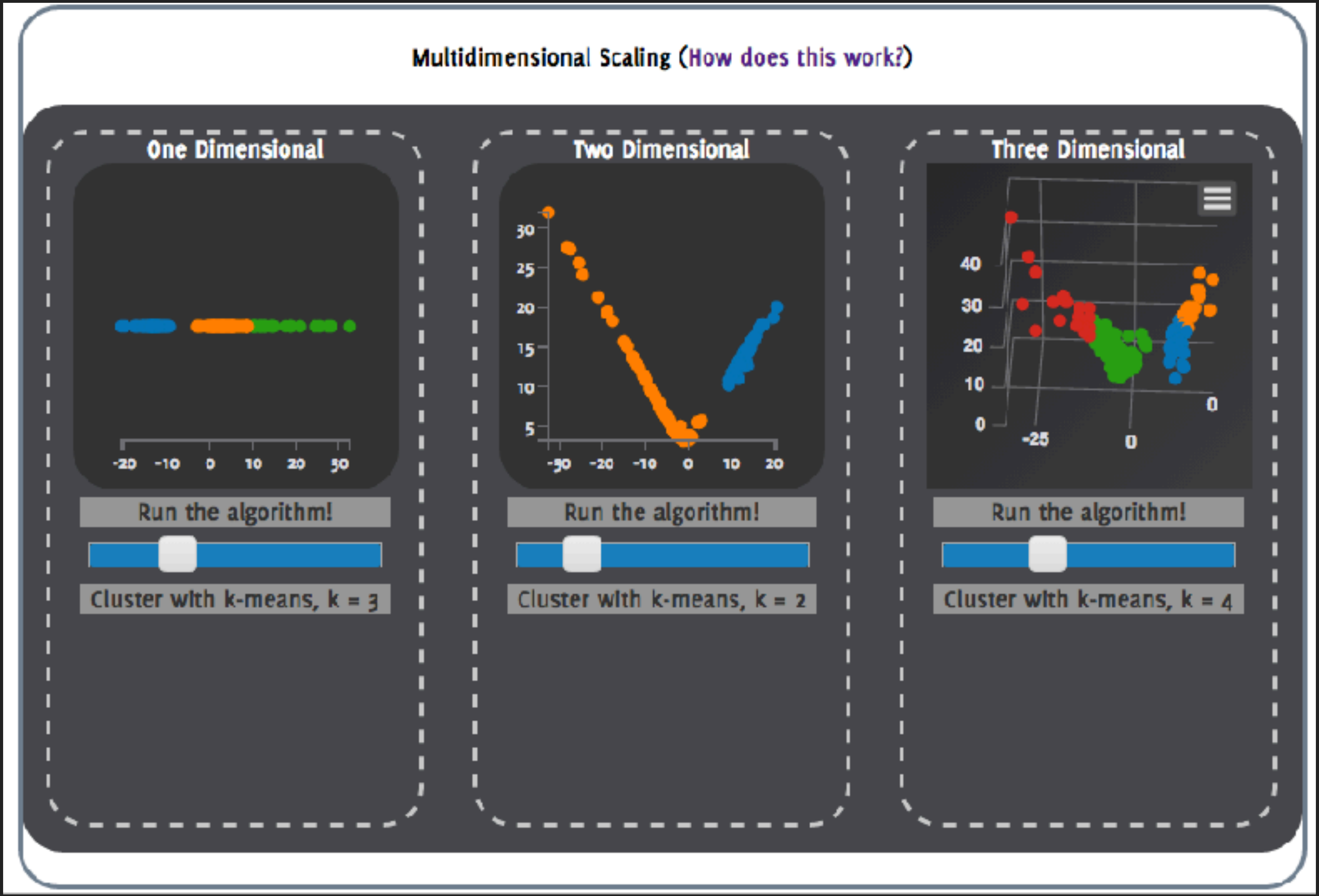
MARTIN WATTENBERG Google Brain
FERNANDA VIÉGAS Google Brain
IAN JOHNSON Google Cloud
Oct. 13 2016
Citation: Wattenberg, et al., 2016



t-Distributed Stochastic Neighbour Embedding (What is this?)

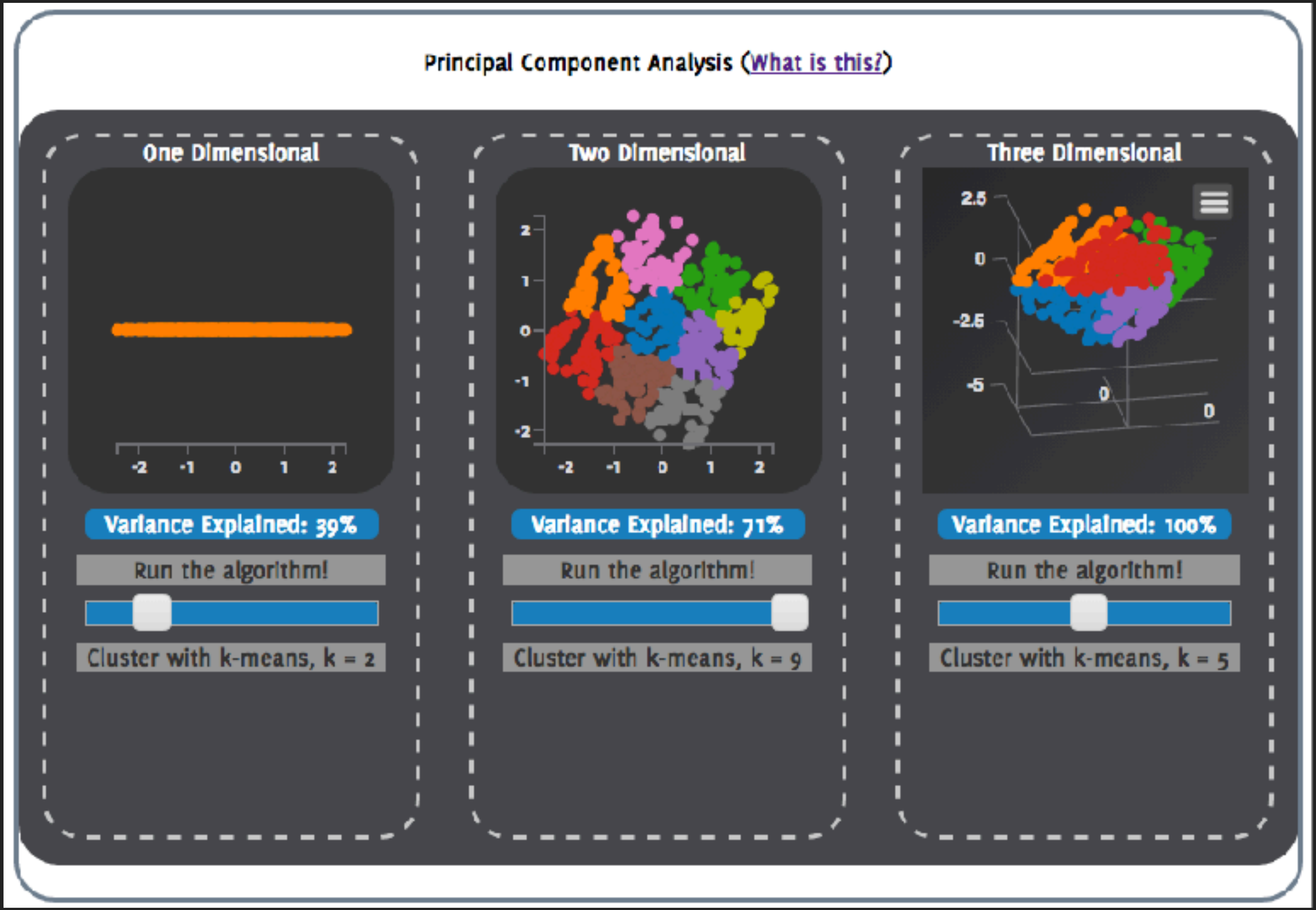
Parameters to tune: P - Perplexity | LR - Learning Rate | N - Number of Iterations

OTHER WIDGETS CURRENTLY IMPLEMENTED: MDS



Multidimensional Scaling (How does this work?)

OTHER WIDGETS CURRENTLY IMPLEMENTED: PCA



Principal Component Analysis (What is this?)

One Dimensional

Variance Explained: 39%

Cluster with k-means, k = 2

Two Dimensional

Variance Explained: 71%

Cluster with k-means, k = 9

Three Dimensional

Variance Explained: 100%

Cluster with k-means, k = 5

VISUALIZATION DASHBOARD: DATA DROP WIDGET

LIVE DEMO: FINDING LATENT STRUCTURE IN DATASET

VISUALIZATION DASHBOARD: DATA DROP WIDGET

LIVE DEMO: DISCOVERING CLUSTERS IN FLOWERS DATASET

FORMATIVE STUDY: DESCRIPTION

- One pilot participant, graduate student in Machine Learning group at University of Toronto
- Unstructured interview format
- Given no description about interface and asked to determine its function for studying data stored in a comma-separated file
- Participant discovered data-loading and dimension reduction features by playing with interface

FORMATIVE STUDY: ACTIONABLE FEEDBACK

- Participant requested dynamic data table to edit individual rows and get immediate feedback in visualizations
- Confused by technical terminology of parameters in t-SNE dimensionality reduction technique
- After discovering 3D rotation in visualization widget, participant repeatedly tried to pan 2D visualization by clicking and dragging
- Takeaways: useful to allow “drilling down” by zooming into particular groups data points and panning: create a focus area. Dynamic visualization for adding/removing of data is priority for next iteration

CONCLUSION

- Demonstrated a high-dimensional data visualization dashboard that leverages advanced machine learning to reduce data to 1, 2, or 3 dimensions of variation
- Shown how this can augment data exploration by enhancing latent structure and natural grouping discovery
- Identified areas for future improvement through formative study with pilot participant