Introduction to Data Analytics Methods

Daniel Archambault

Previously in CSCM27...

- Name some types of graphs
- Are there specific algorithms for them?
- What type of algorithm can draw graphs without assumption of structure?
- Is there a relationship with MDS?

Previously in CSCM27... (2)

- Data analytics and visualisation in the form of clustering

Intro. Data Analytics Methods

Data Analytics and Visual Analytics

- Visual analytics involves two things:
  1. visualisation interfaces to support exploration
  2. data analytics and mining to support scale & analysis
- We have done a fair bit on visualisation and visual interfaces
- Now, we begin to look at data analytics processes

The Role of Visualisation is Important

Why is a windscreen important for a car?
Visualisations are the windscreens for data science.

- Nawww, I don't believe you
  - visualisation is not useful
  - the data sets are too large
  - no point in visualising
- Well, perhaps the right visualisation is a table of statistics? Right?

Have you met the Datasaurus...?

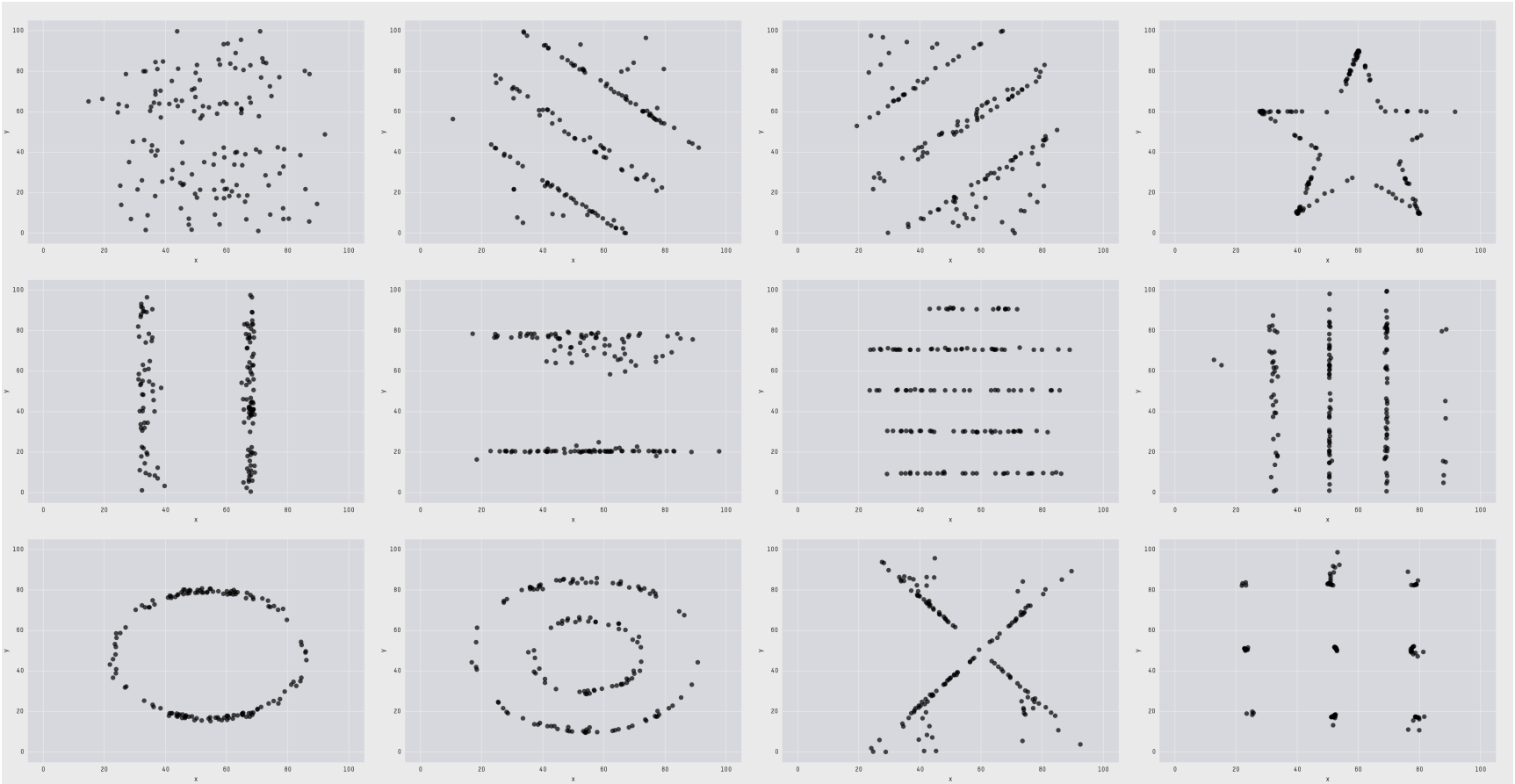
(It's kinda like Anscombe's Quartet...)

Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (ACM CHI '17). 1290-1294.

Same Stats, Different Graphs

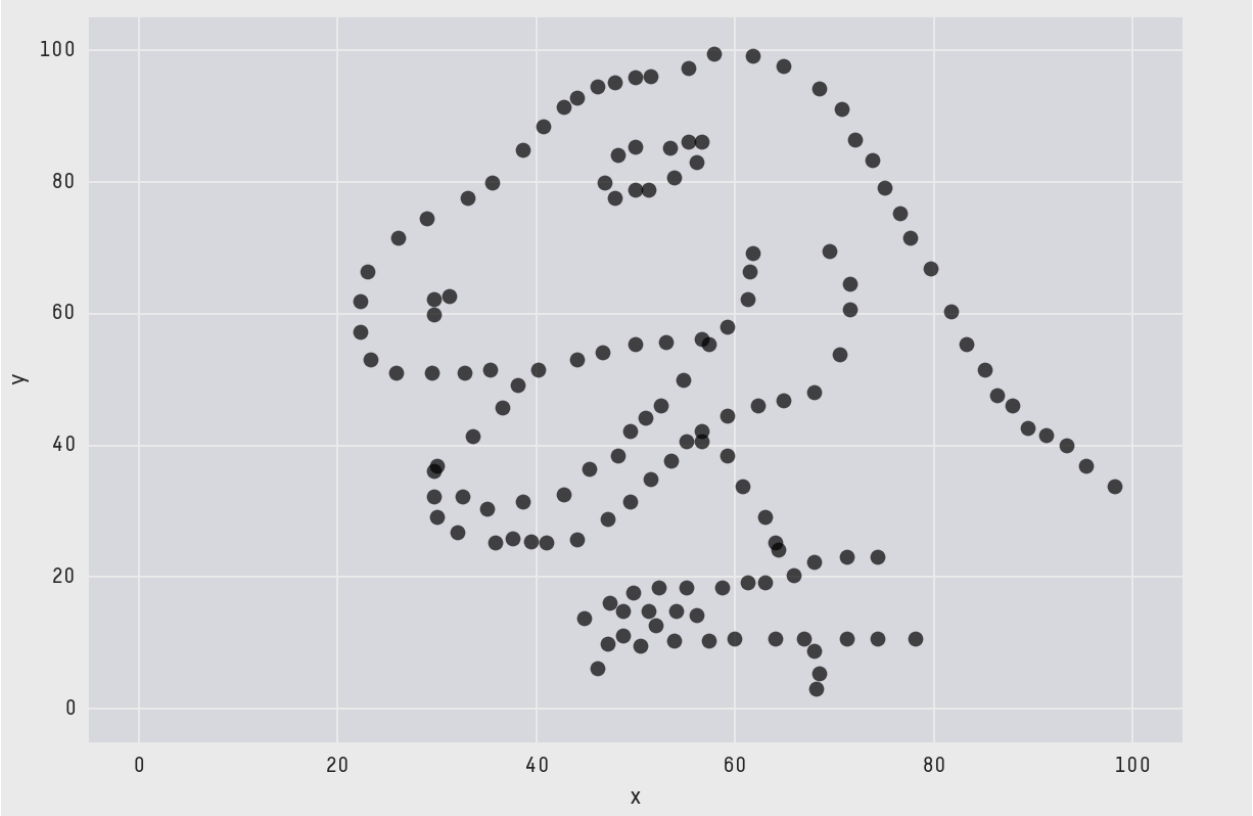
- All of these graphs have the exact same statistics:
  - mean in x: 54.26, y: 47.83
  - standard deviation in x: 16.76, in y: 26.93
  - correlation: -0.06

All these graphs have the same stats...



https://www.autodeskresearch.com/publications/samestats

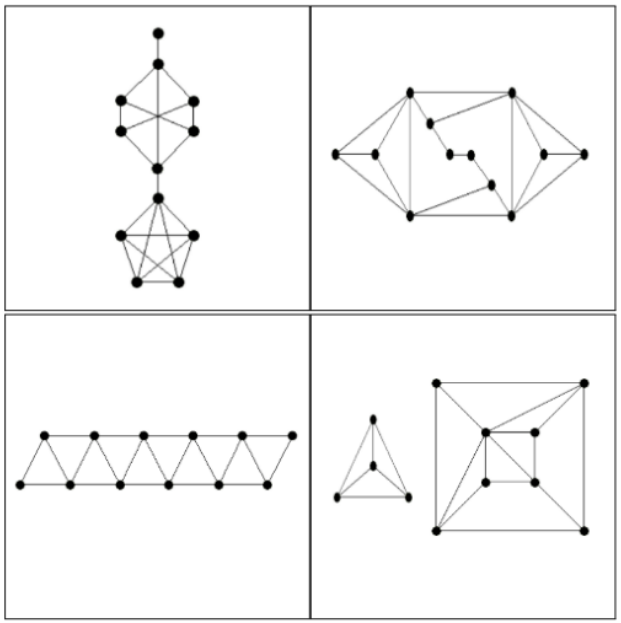
... and so does this one.



https://www.autodeskresearch.com/publications/samestats

- Forget bugs! There could be dinosaurs in your data!

Same Result in Network Analysis



H. Chen, U. Soni, Y. Lu, R. Maciejewski and S. Kobourov, "Same Stats, Different Graphs (Graph Statistics and Why We Need Graph Drawings)," 26th Symposium on Graph Drawing (GD), 2018.

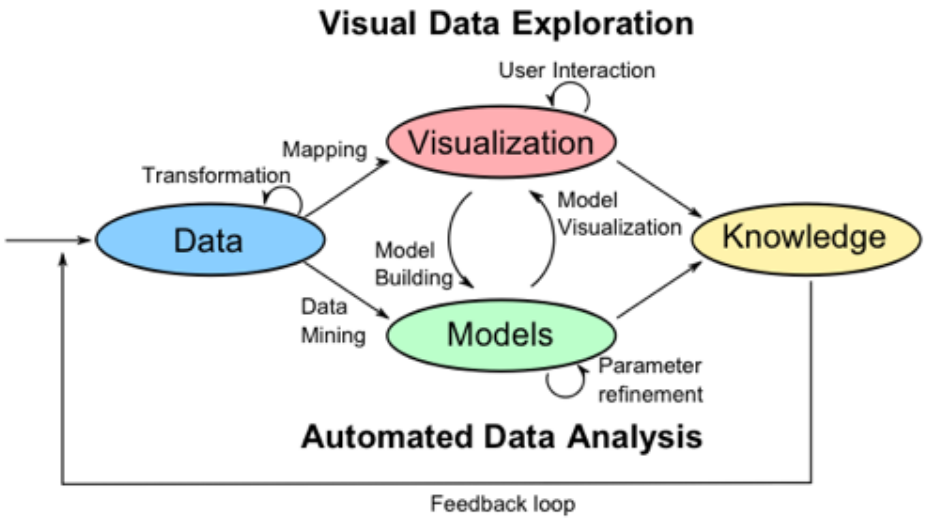
- All of these networks have the same statistics
  - vertices, edges, triangles, girth, clustering coefficient

Moral of the Story...

- *... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.*
  - F. J. Anscombe, 1973
- *Never trust summary statistics alone; always visualize your data*
  - Alberto Cairo
  - `http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html`
- But, surely, we can just apply any old visualisation? They're all the same? Right?
- Did you attend the last few lectures?

Recall: Daniel A. Keim Definition of Visual Analytics

- Processing information transparent for analytic discourse
  - visualisation communicates to user
  - machine learning automates data process
  - leverage strengths of human and machine



- This lecture provides an overview of ways to integrate

Keim D., Andrienko G., Fekete JD., Gorg C., Kohlhammer J., Melancon G. (2008) Visual Analytics: Definition, Process, and Challenges. In: Kerren A., Stasko J.T., Fekete JD., North C. (eds) Information Visualization. Lecture Notes in

Outline

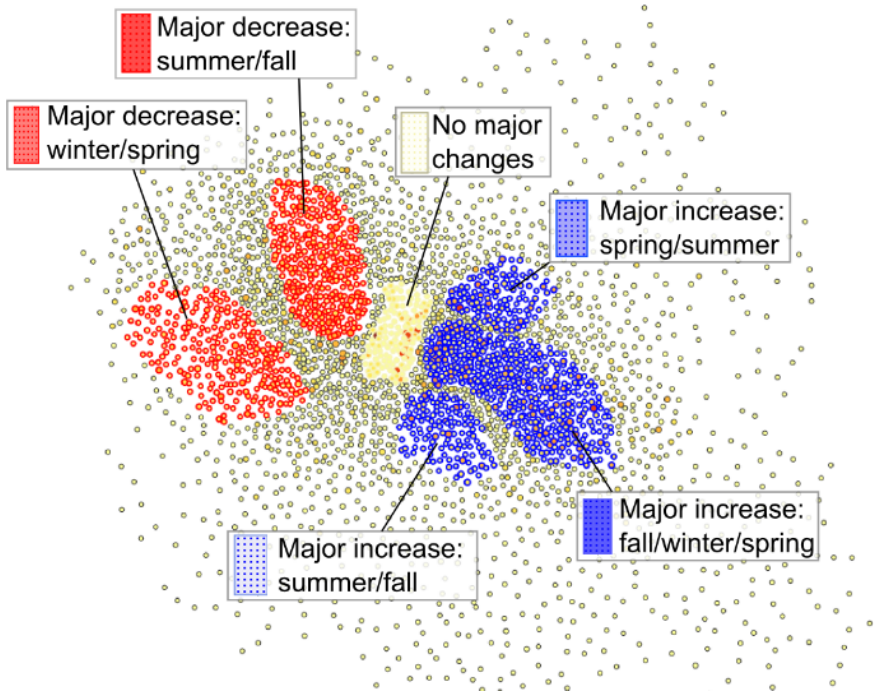
- Lecture provides overview of how to integrate ML methods
  - Dimensionality Reduction
  - Clustering
  - Classification
  - Regression

Endert, A., Ribarsky, W., Turkay, C., Wong, B.L. W., Nabney, I., Blanco, I. D. and Rossi, F. (2017), The State of the Art in Integrating Machine Learning into Visual Analytics. Computer Graphics Forum, 2017

Dimensionality Reduction

- Wait... what? I thought that was a visualisation method?
  - PCA, MDS and those methods are used to display data
- These methods can be used for visualisation
- But, they start out as data analytics methods
- Dimensionality reduction helps
  - reduce information speeding up other computations
  - optimise computational resources

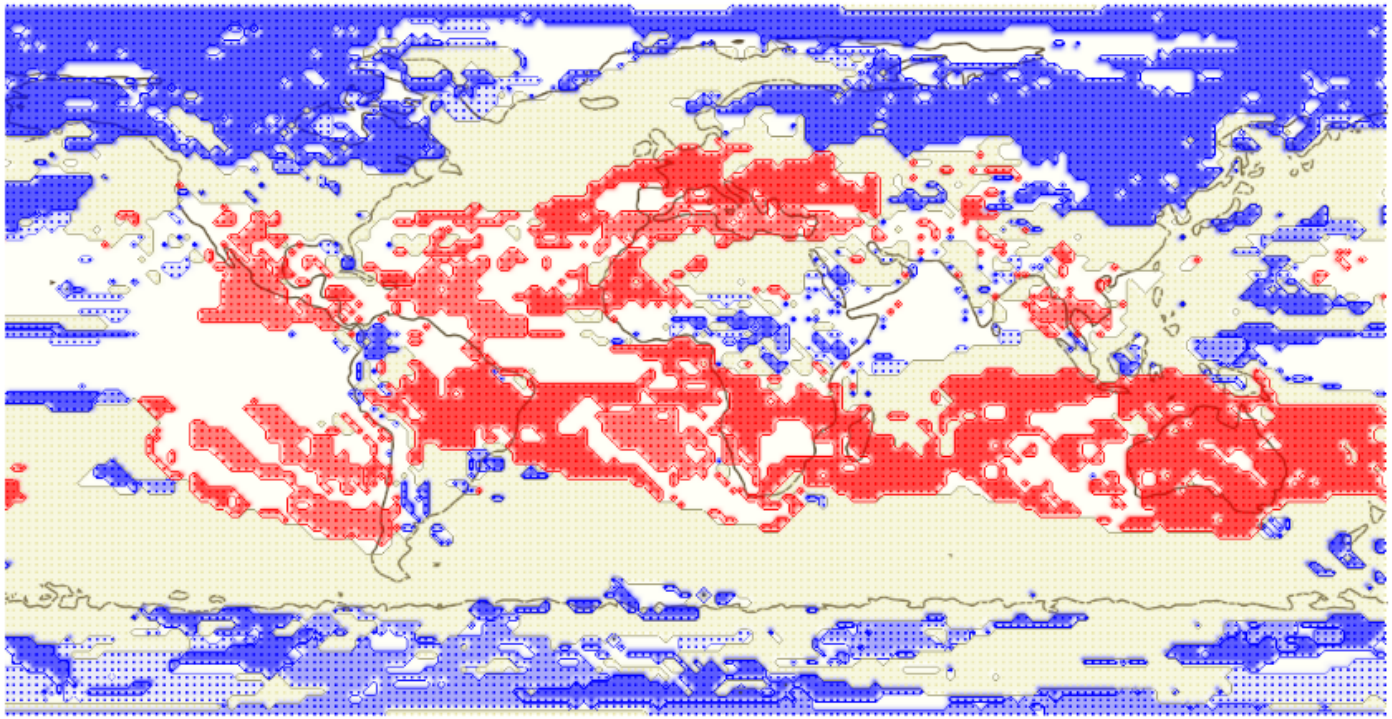
Dimensionality Reduction Detects Features



Heike Janicke, Michael BÃ¶ttinger, and Gerik Scheuermann. 2008. Brushing of Attribute Clouds for the Visualization of Multivariate Data. IEEE Transactions on Visualization and Computer Graphics 14, 6 (November 2008), 1459-1466.

- Dimensionality reduction places similar things nearby
- Clusters emerge that can be detected in the lower dimensional

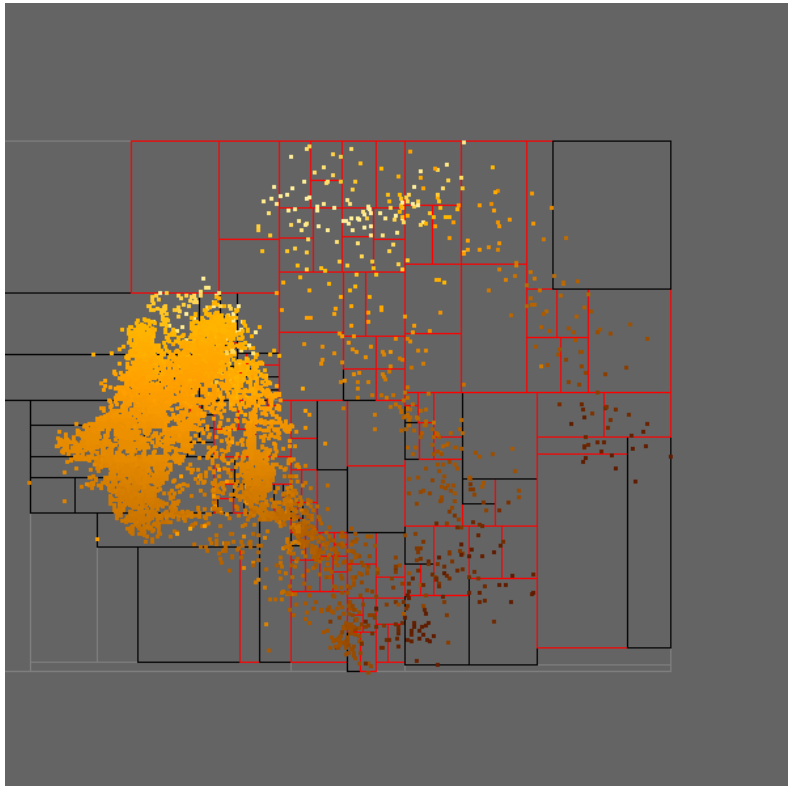
Visualised for Sensemaking



Heike Janicke, Michael BÃ¶ttinger, and Gerik Scheuermann. 2008. Brushing of Attribute Clouds for the Visualization of Multivariate Data. IEEE Transactions on Visualization and Computer Graphics 14, 6 (November 2008), 1459-1466.

- Given analysed data, can visualise it in new ways

Dimensionality Reduction Speed of Algorithms



M. Williams and T. Munzner, "Steerable, Progressive Multidimensional Scaling," IEEE Symposium on Information Visualization, pp. 57-64.

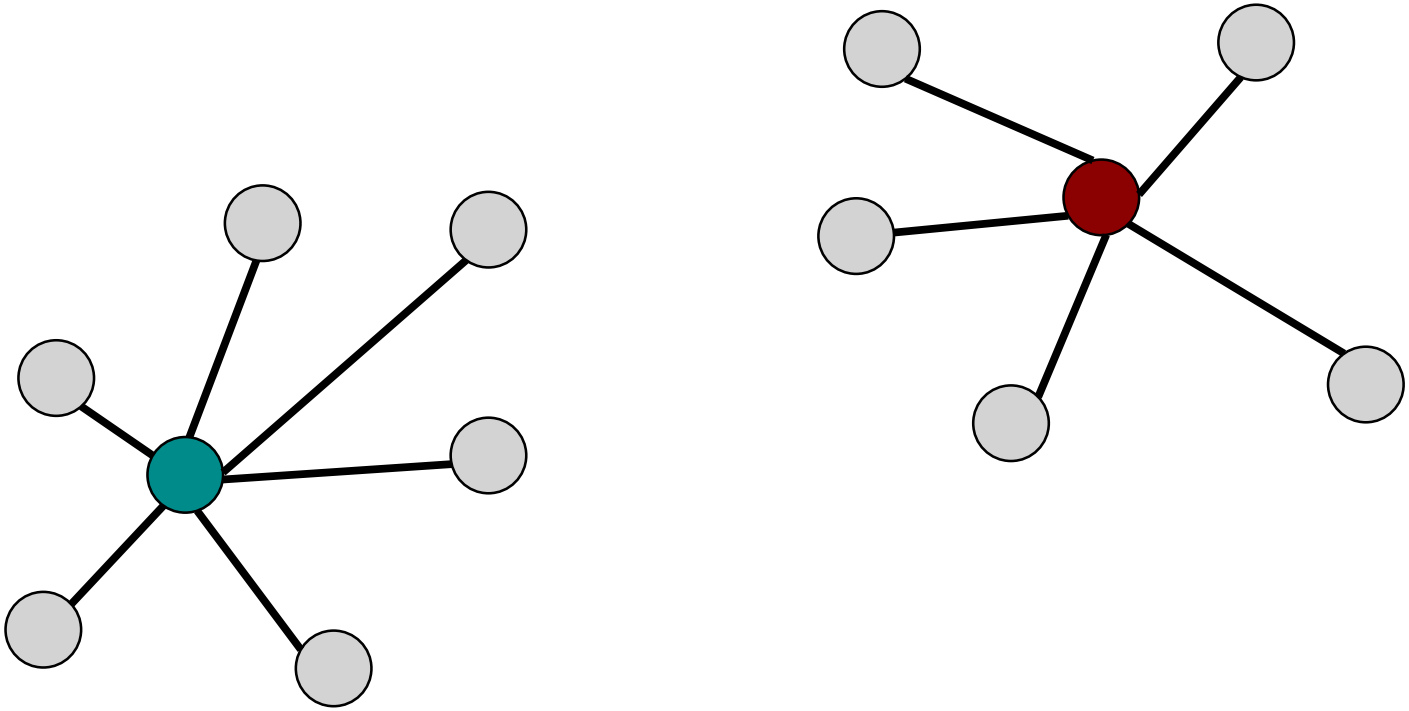
- Detecting and removing dimensions that matter little helps speed up process

Data Clustering

- Find clusters of data points in space
- Many algorithms to do this
- We focus on two
  - k-means clustering
  - self-organising maps

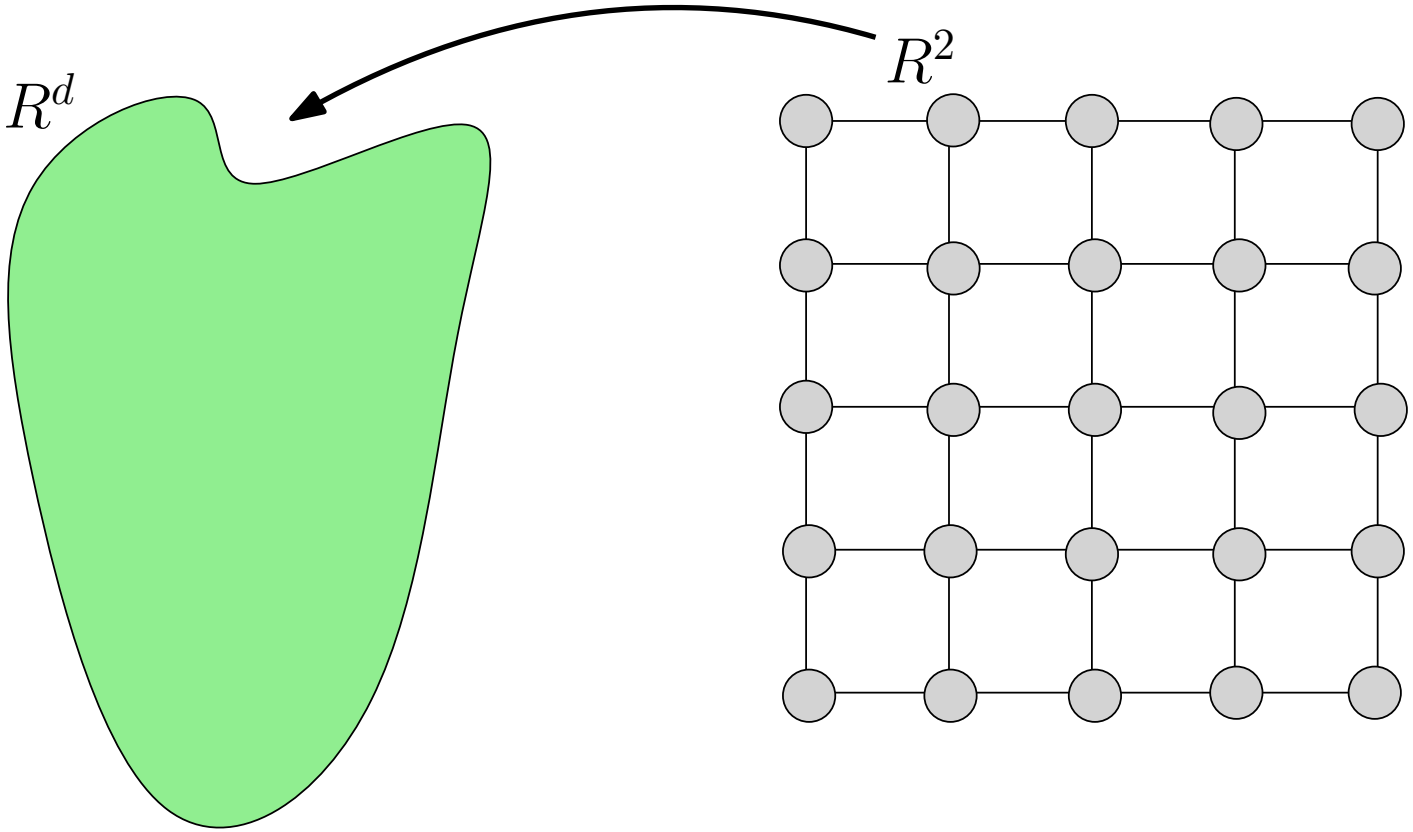
Data Clustering by k-means

$$R^d$$



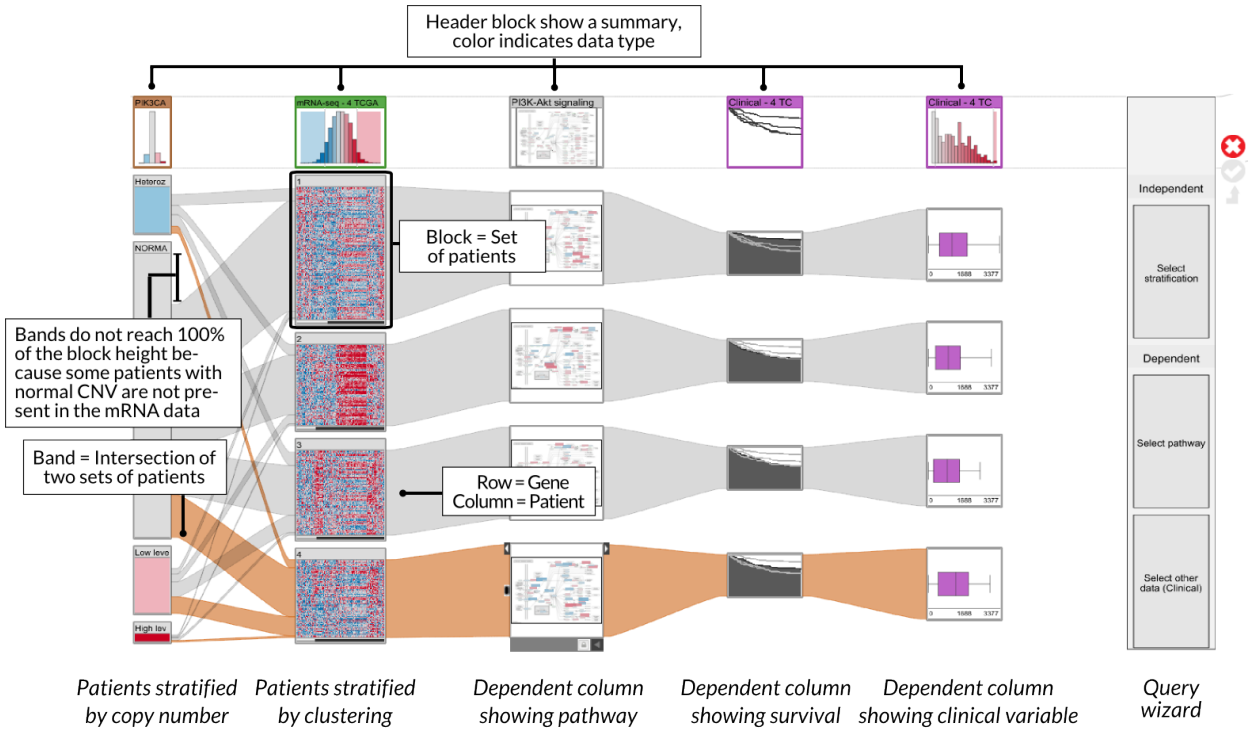
- Randomly place k-centres inside R^d
- Iteratively move centres until distance to all points is minimised

Data Clustering by Self-Organising Maps (SOMs)



- Create sheet of artificial neurons (neural net)
- Crinkle sheet in high dimensional space to “cover” it

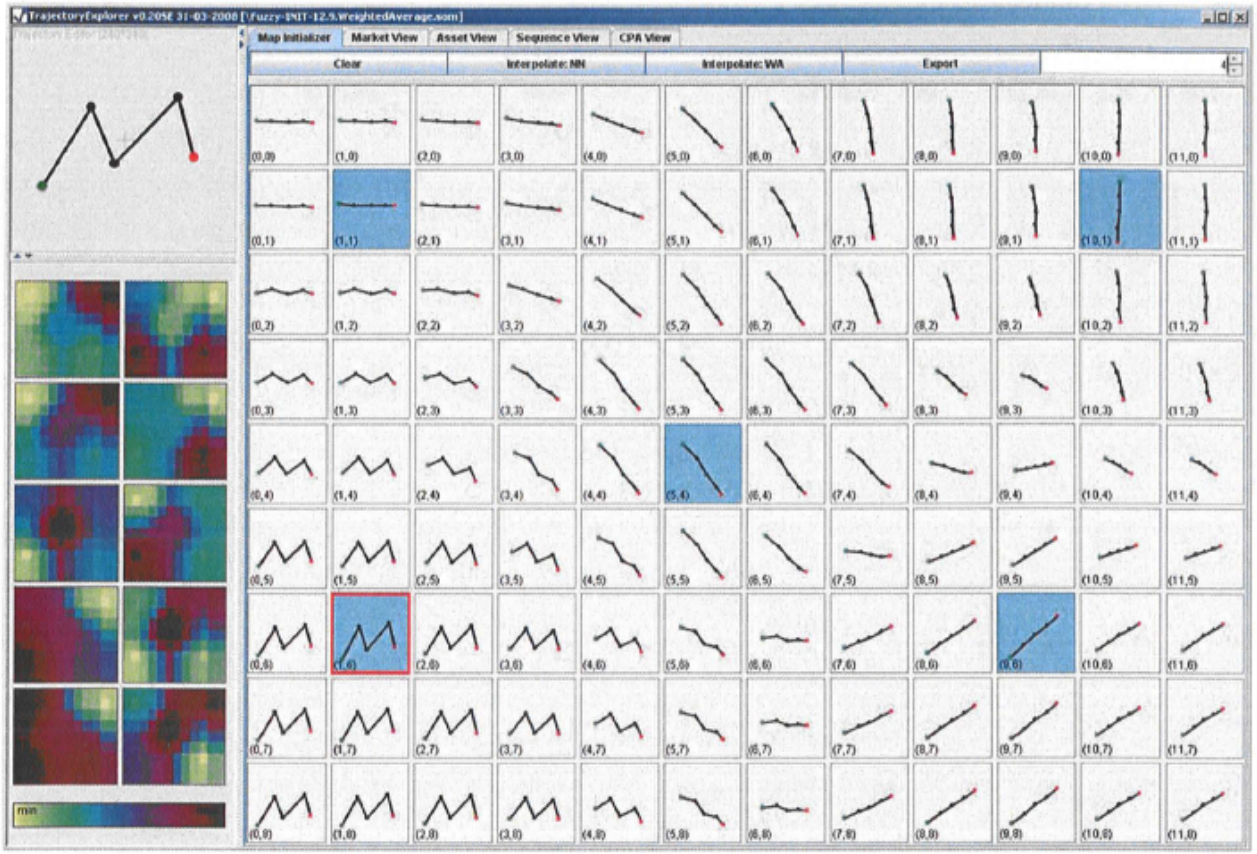
Explore Multiple Clusterings of Points



Lex, A., Streit, M., Schulz, H.-J., Partl, C., Schmalstieg, D., Park, P.J. and Gehlenborg, N. (2012), StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. Computer Graphics Forum, 31: 1175–1184.

- Explore several clusterings of genes for cancer subtypes

Trajectory Analysis: SOMs



T. Schreck, J. Bernard, T. Tekusova and J. Kohlhammer, "Visual cluster analysis of trajectory data with interactive Kohonen Maps," 2008 IEEE Symposium on Visual Analytics Science and Technology, Columbus, OH, 2008, pp. 3-10.

Classification: Random Decision Trees

- Create a random forest of decision trees
  - nodes at each level encode a decision based on an attribute
  - data set is classified by following a path in the tree
- Training on example data modifies how decisions are made on the nodes

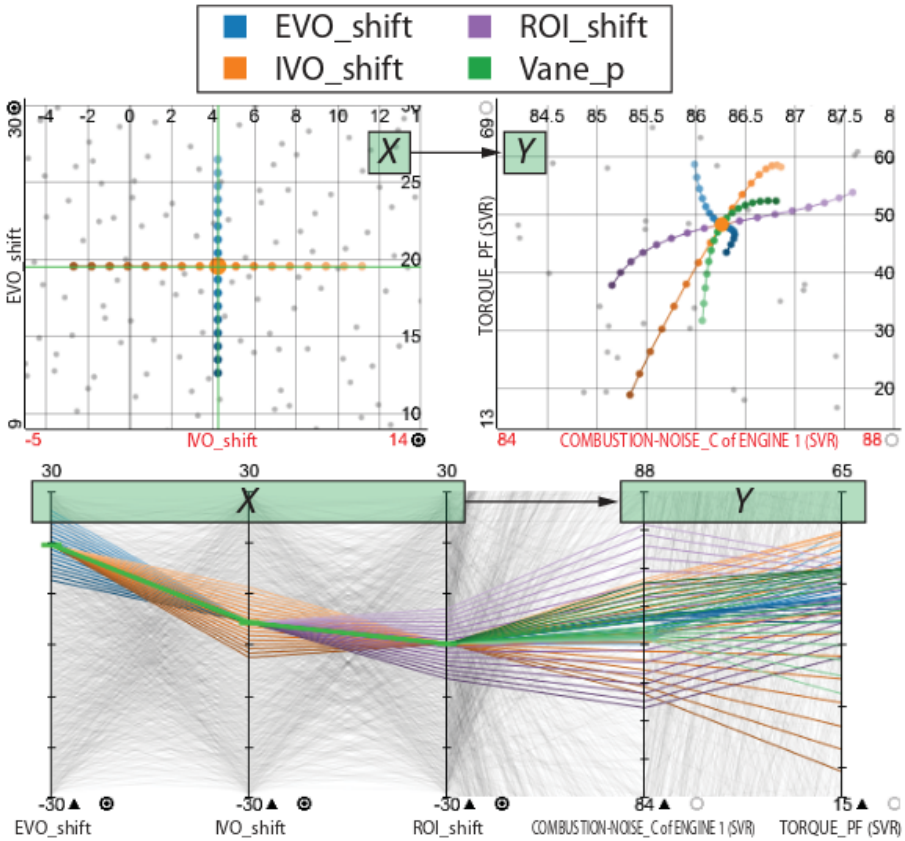
You are what you tweet?

diabetes: +	*Mexican* (mexican, tacos, burrito), *American Diet* (chicken, baked, beans, fried), #food, *After Work* (time, home, after, work), #pdx, my, lol, #fresh, *Delicious* (foodporn, yummy, yum), #fun, morning, special, good, cafe, #nola, fried, bacon, #cooking, all, beans
diabetes: -	#dessert, *Turkish* (turkish, kebab, istanbul), #foodporn, #paleo, #meal, *Paleo Diet* (paleo, chicken, healthy), i, *Giveaway* (win, competition, enter), *I, You* (i, my, you, your), your, new, today, #restaurant, *Japanese* (ramen, japanese, noodles), some, jerk, #tapas, more, *Healthy DIY* (salad, chicken, recipe), *You, We* (you, we, your, us)
Democrat	#vegan, #yum, w, served, #brunch, *Deli* (cheese, sandwich, soup), photo, #rvadine, *Restaurant Advertising* (open, today, come, join), #breakfast, #bacon, delicious, #food, #dinner, 21dayfix, like, #ad, *Giveaway* (win, competition, enter), toast, 1
Republican	my, #lunch, i, *Airport* (airport, lounge, waiting), easy, #meal, tonight, #healthy, #easy, us, sunday, *After Work* (time, home, after, work), #party, #twye, *First-Person Casual* (my, i, lol), your, #snack, join, #delicious, house

D. Fried, M. Surdeanu, S. Kobourov, M. Hingle and D. Bell, "Analyzing the language of food on social media," 2014 IEEE International Conference on Big Data (Big Data), pp. 778-783.

Regression

- Uniform relations between one or more multivariate regions of the data
- Simplest form is linear – try to find a line of best fit
- More complicated versions exist related to machine learning
  - i.e. logistic regression



Berger, W., Piringer, H., Filzmoser, P. and GrÃ¶ller, E. (2011), Uncertainty-Aware Exploration of Continuous Parameter Spaces Using Multivariate Prediction. Computer Graphics Forum, 30: 911–920.

- Non-linear and high dimensional correlations in colour.

Data Analytics Methods and Visualisation

- Name some of the data analytics methods in this lecture
- Give a problem and an appropriate analytics method
- How does that change the visualisation you use?