UWE Bristol | University of the West of England

Presentation by

Dr. Phil Legg

Associate Professor in Cyber Security

Security Data Analytics and Visualisation

5: Visualisation

Date: Autumn 2019

Recap

- Last week we looked at Machine Learning techniques
  - We will reflect more on this in our "pause week" – week 6

- Learn by examples – examples of 3 common techniques available on Blackboard – *Clustering, Regression, Neural Network*

Data analytics pipeline

How do we visualize data?

- 2-dimensional charts and plots
- 3-dimensional data representations
- Focus-and-Context
- Interaction

UWE Bristol University of the West of England

Visualisation

- What do we mean by visualisation?
- What is the purpose, or benefit, of visualisation?
- Types of visualisation
- Visual channels and their appropriate uses
- Visualisation for Cyber Security

UWE Bristol | University of the West of England

Benefits of Visualisation?

UWE Bristol | University of the West of England

Types of Visualisation

Bar Chart

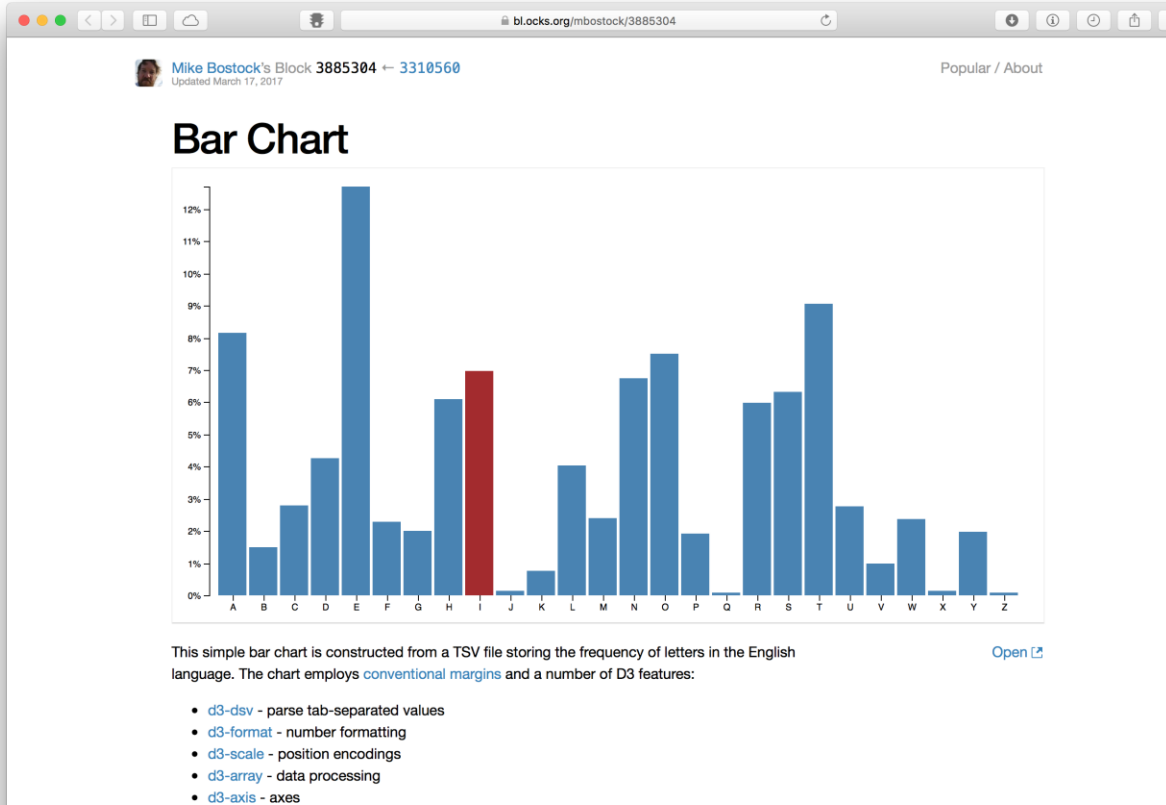
Visual Channels:

Height (Size)

Colour (Selection)

Useful for showing categorical count data

Similar to histogram



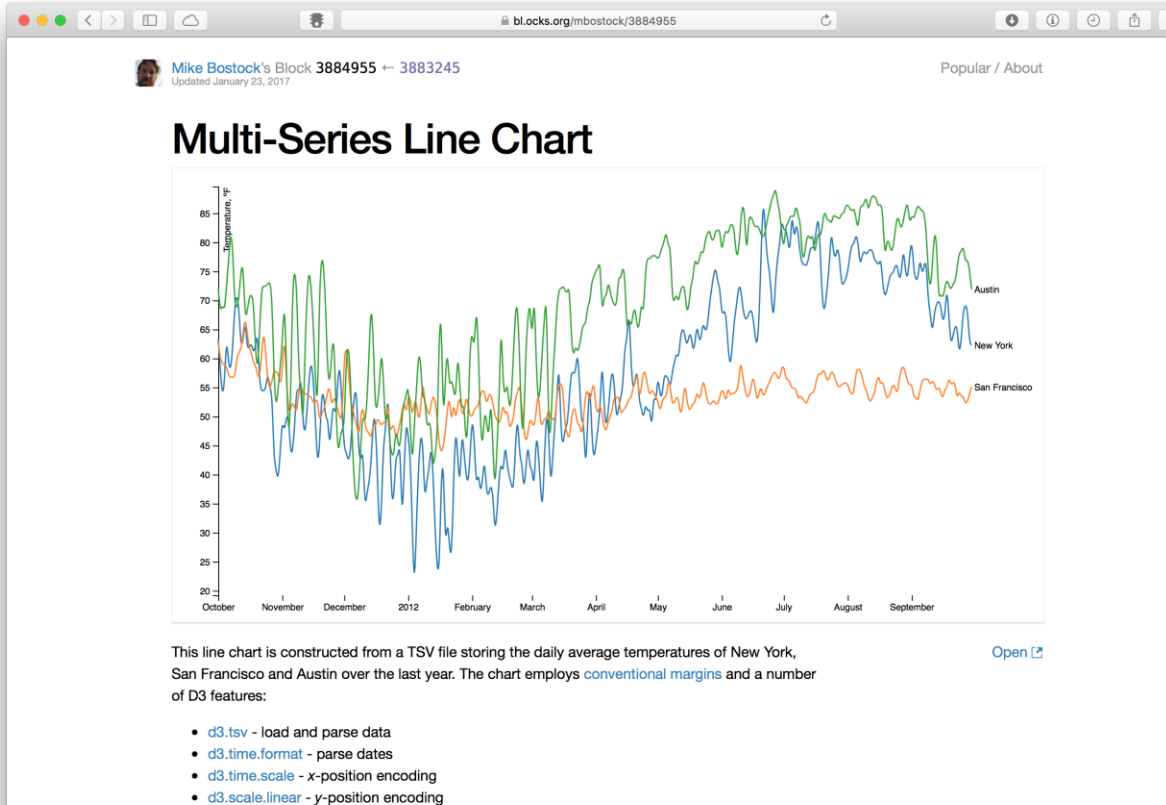
Line Chart

Visual Channels:

Height (Size)

Colour (Selection)

Useful for showing time-series data, where time is on the X axis.



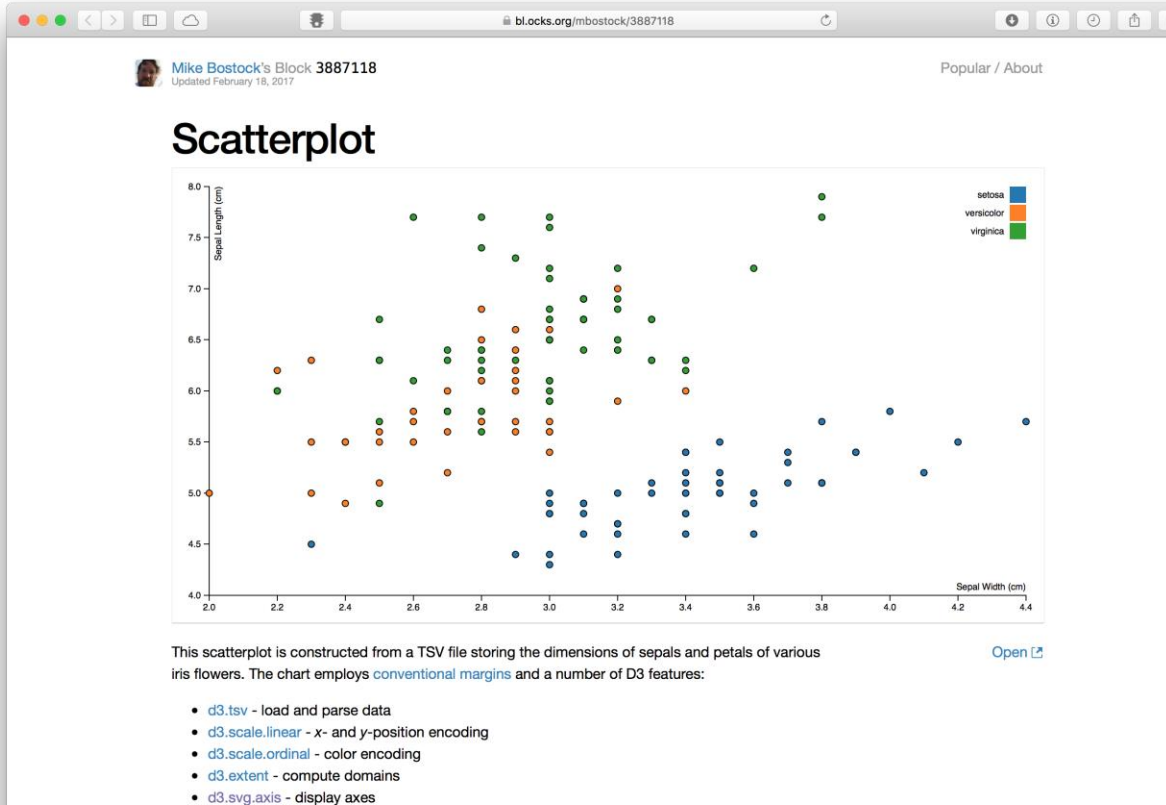
Scatter Plot

Visual Channels:

Height (Size)

Colour (Selection)

Useful for showing correlation between two variables

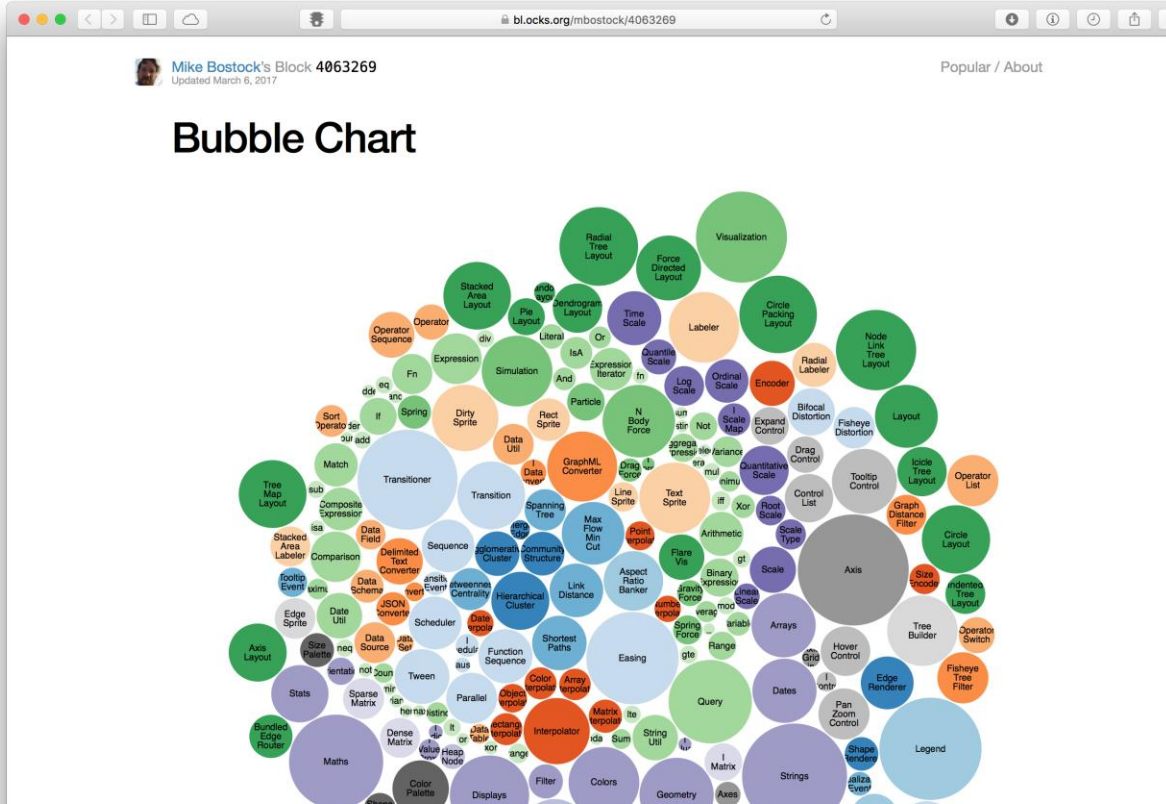


Bubble Chart

Visual Channels:

Colour and Size

Shows count data for categories – so similar to bar chart – however removes the ordering problem of bar chart

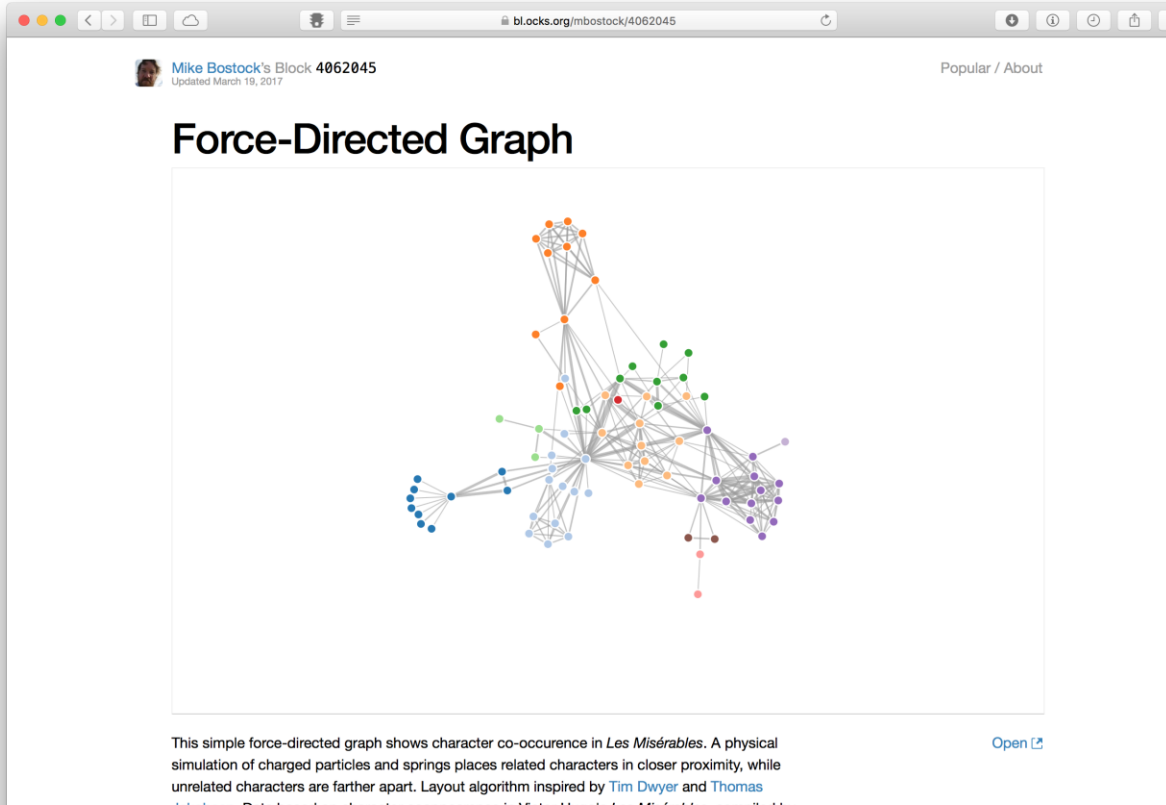


Force-Directed Graph

Visual Channels:

Colour and Size

Useful for showing the relationship between different entities (nodes)



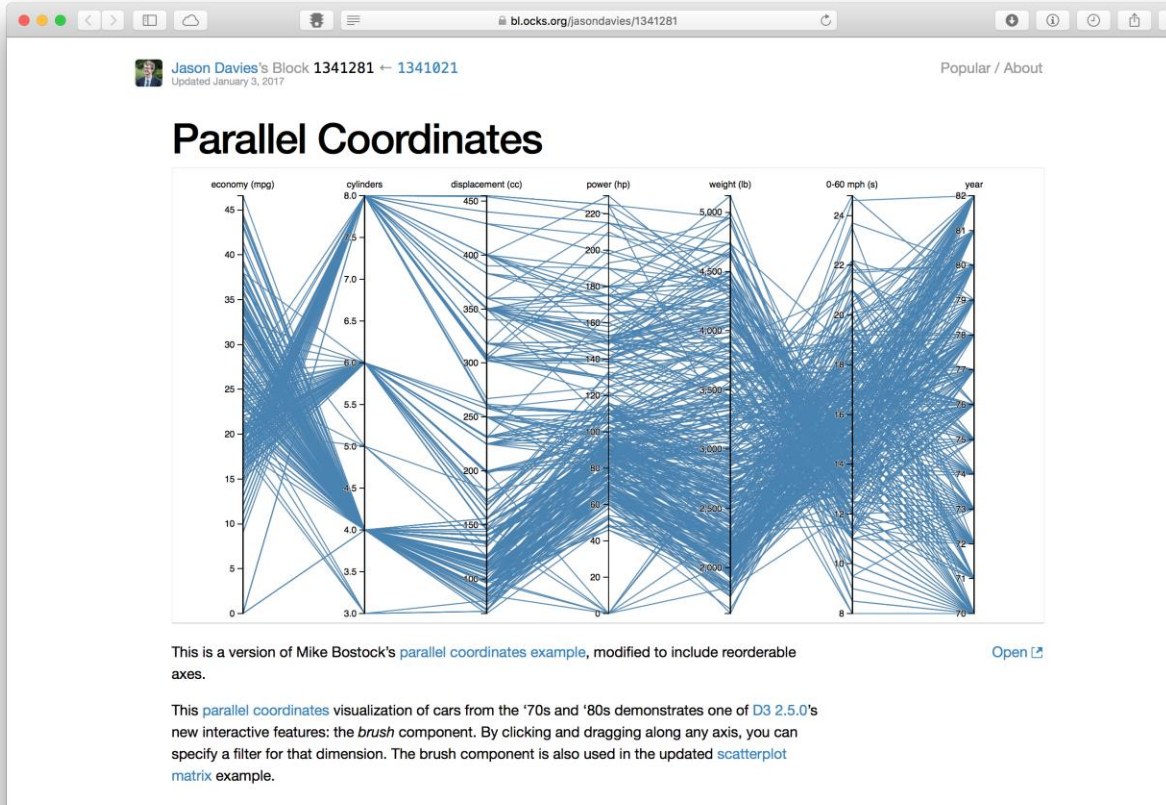
Parallel Coordinates

Visual Channels:

Colour and Size

Useful for showing multiple attributes of data – can also show correlation between variables (when axis are positioned next to each other)

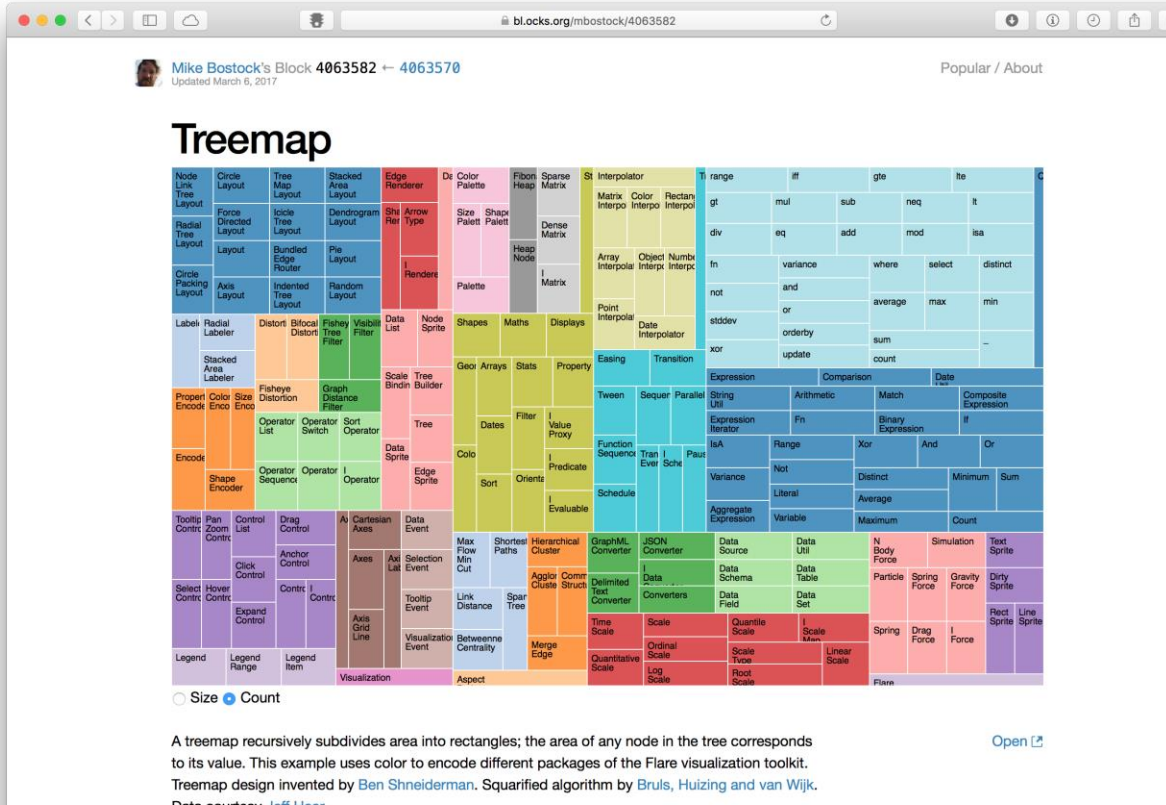
Axis ordering can change representation



Treemap

Similar to a treemap but uses a radial layout

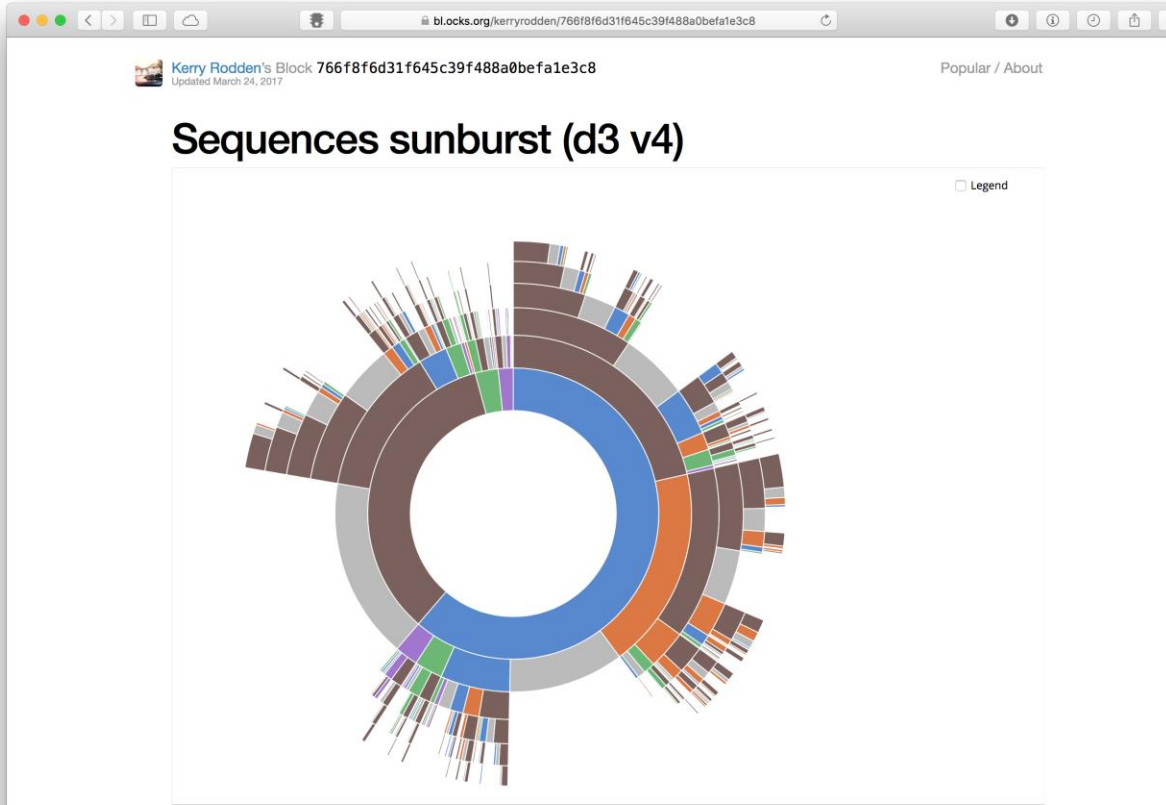
Useful for showing hierarchy in data, and relative count data associated with entities



Sunburst

Similar to a treemap but uses a radial layout

Useful again for hierarchy – possibly makes hierarchy more apparent than treemaps – but essentially the same data

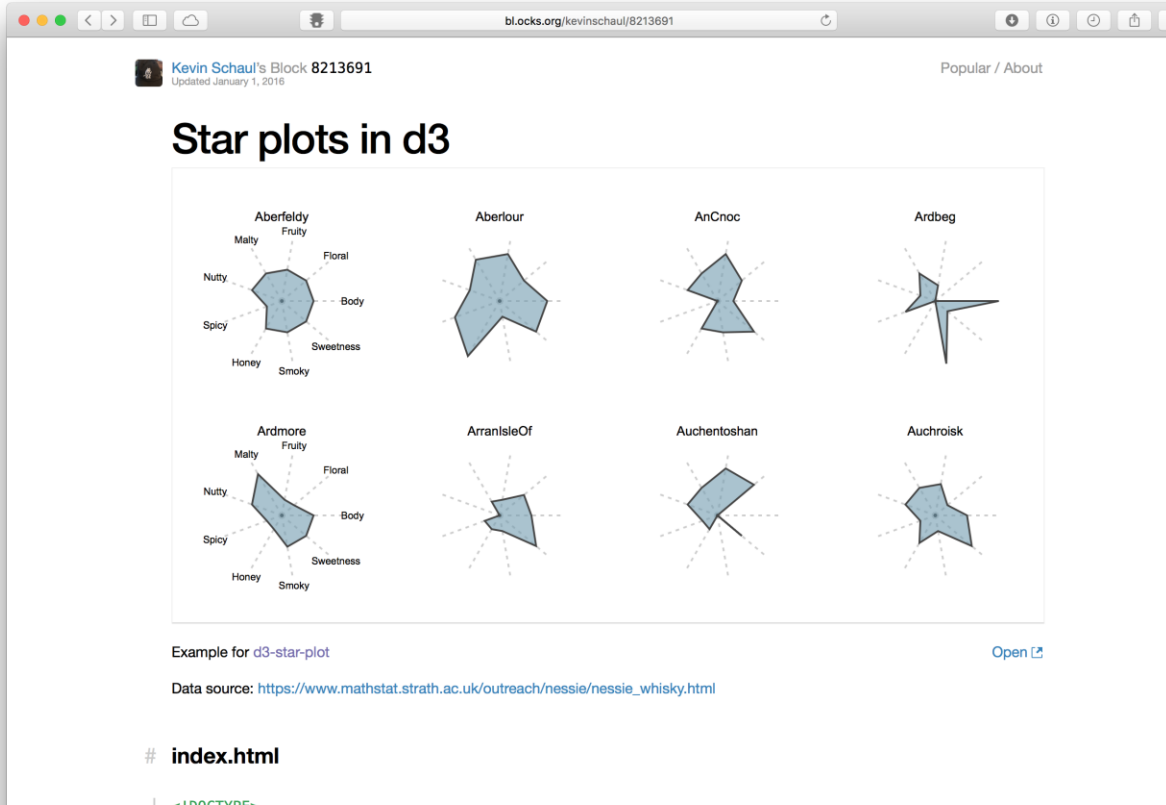


Star Plots

Multi-variate plots – essentially small radial parallel coordinates plot.

Can be referred to as a "Glyph".

Also similar to parallel coordinates (except that coordinates are in radial layout).



Glyph Visualisation

Small icon that illustrates multi-variate data attributes.

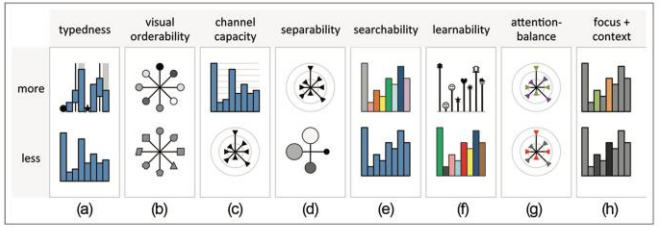


Figure 2. (a)–(h) Variations of glyph design in accordance to the design principles of sortable glyph. For each principle, the top row depicts a glyph with greater emphasis and the bottom row depicts a glyph with less emphasis.

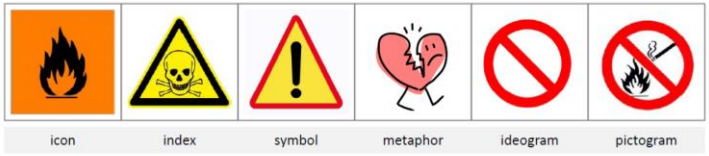


Figure 1: In philosophy, language studies and psychology, signs may take one of the three forms, icon, index and symbol. In many contexts, terms such as visual metaphor, ideogram and pictogram are also used to denote subclasses of signs.

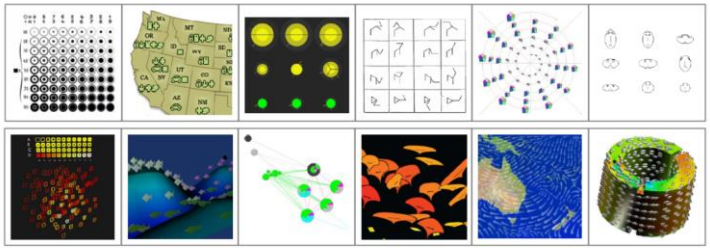


Figure 4: Examples of many different glyphs, which make use of different visual channels (e.g., size, line thickness, symbol, color, proportion, aspect-ratio, orientation, curvature and so on) to convey the variation of data records in different attribute dimensions.

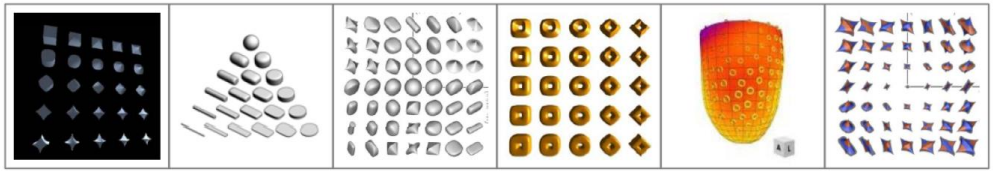


Figure 5: A collection of glyphs that are designed for visualizing multivariate volumetric data, including diffusion tensor imaging data.

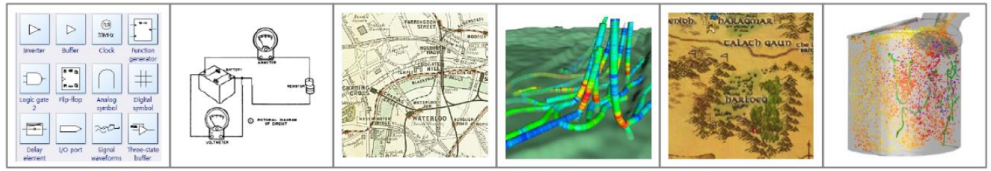


Figure 6: In a broad interpretation, glyphs can be connected together to form a schematic diagram, or to convey the common properties of a line, an area or a volume. For example, an abstract icon of a railway track can be considered as a glyph. When many of them are connected together, they denote a railway line on a map.

UWE Bristol University of the West of England

What visualisations are most suitable for security investigations?

Parallel Co-ordinates

- Parallel Co-ordinates convey multi-variate data
- Each axis is an attribute of the data, and a connected line shows one instance from the data
- Many have looked at how parallel co-ordinates can be used for understanding network activity (Visualizing Network Activity using Parallel Coordinates)

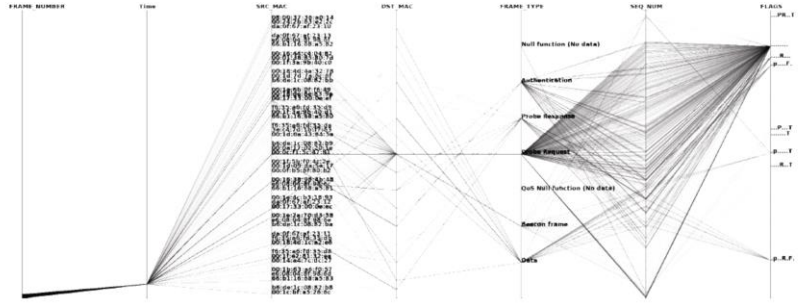


Figure 2: Regular WLAN traffic

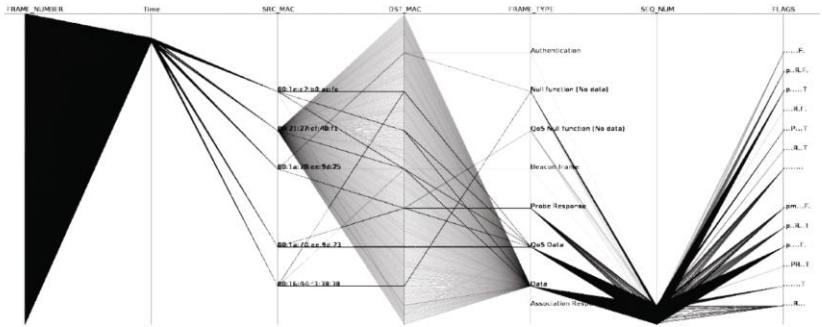


Figure 3: WEP cracking attack

UWE Bristol University of the West of England

Parallel Co-ordinates

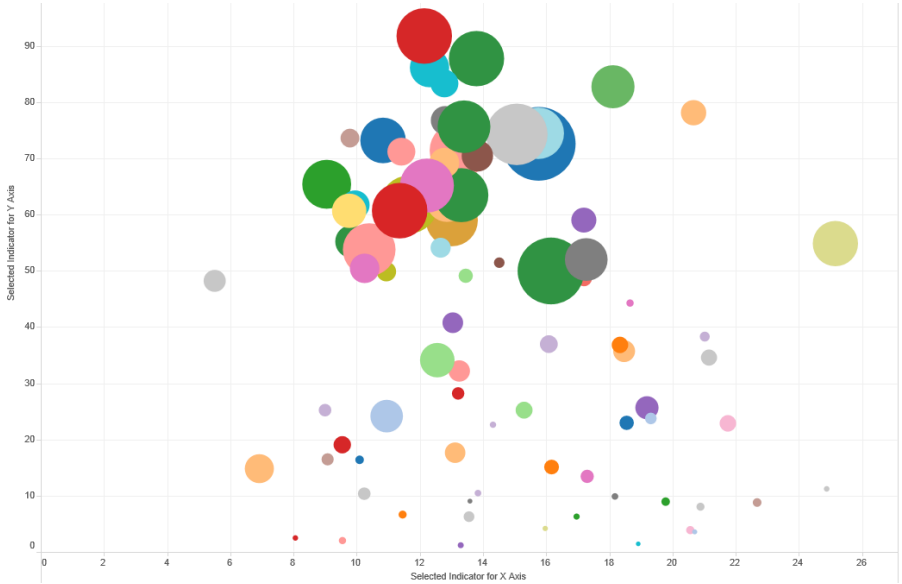
- Parallel Co-ordinates convey multi-variate data
- Suppose we have:
  Source IP, Source Port, Dest IP, Dest Port as our 4 columns
- Visual "signatures" can be **extremely effective** for identifying different network attacks
  (Fast detection and visualization of network attacks on parallel coordinates)

Table 1 – Graphical signatures of nine attacks.

Implied Attack	Signature	Divergences
Portscan		1:1:*m*:1
Hostscan		1:*m*:1:1
Worm		1:*m*:1:1
Source-spoofed DoS (port fixed)		*m*:1:1:1
Backscatter		1:*m*:*m*:1
Source-spoofed DoS (port varied)		*m*:1:*m*:1
Distributed hostscan		*m*:*m*:1:1
Network-directed DoS		*m*:*m*:*m*:1
Single-source DoS		1:1:1:1

Scatter Plot

- Provides spatial relation between observations
- Imagine if each points represents a single user of a organisation network
- What does it mean for a point to be far from others?
- How do we decide on a 2-dimensional representation of our data?



Examples – Packet Visualisation

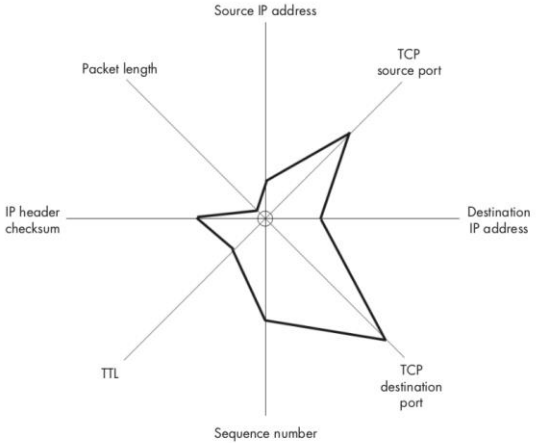


Figure 1-10: Example of a starplot visualization. Starplots are used to display multivariate data by plotting values on axes that extend from a central point and connect the data points. This figure depicts eight values, one per axis.

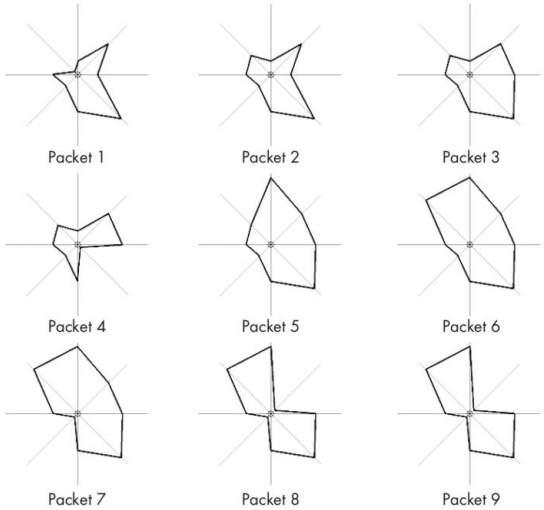


Figure 1-11: Using a three-by-three starplot matrix to illustrate small multiples. Note how you can easily compare and contrast the images and how each set of values takes on a distinct shape. For example, it's easy to see that packets 8 and 9 are strikingly similar.

Examples – File Visualisation

Photo by Daniel Conti (http://www.contidesign.com). Used with permission.

Figure 2-4: Photo of the George Washington Bridge in New York City. I'll create a small smart book that shows this image in a variety of formats and use the results t observe the security behavior of a Microsoft Word document.

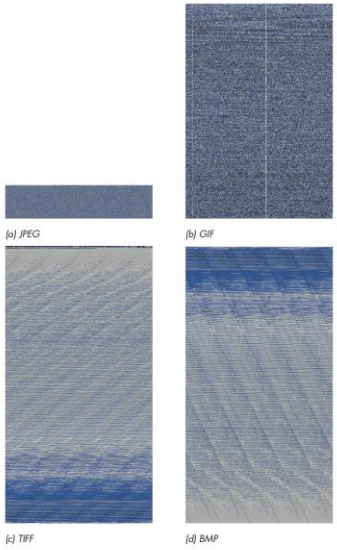
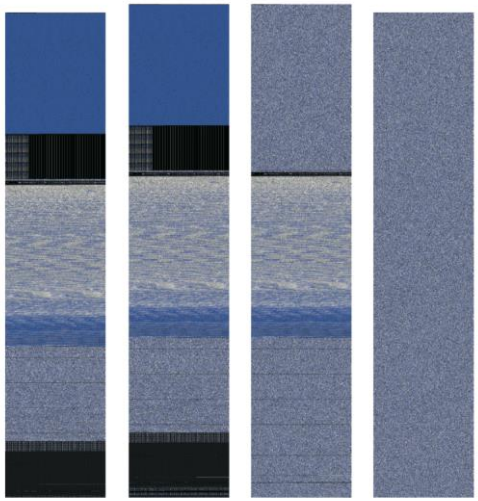


Figure 2-5: Binary visualizations of the George Washington Bridge photo in four different file formats



(a) Original Word document

(b) Word document with modify password protection

(c) Word document with open password protection

(d) Word document protected with the Advanced Encryption Standard algorithm

Figure 2-8: Binary visualization of a Microsoft Word document with various degrees of protection. Note that the original document (a) and the modify password–protected document (b) are visually identical. Neither the text nor the image is protected. When an open password is used (c), the blue region of text is replaced with encrypted text (seen as white noise), but the embedded image has not been protected. When the file is encrypted with a third-party encryption program (d) the entire document appears as white noise.

Examples – Parallel Coordinates

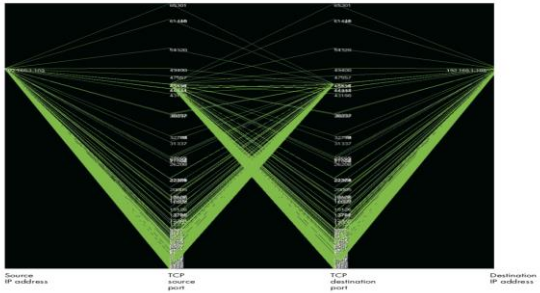


Figure 3-9: Using the parallel coordinate plot technique to view an Nmap port scan. This image depicts both inbound and outbound packets, which form the two overlapping Vs of network traffic.

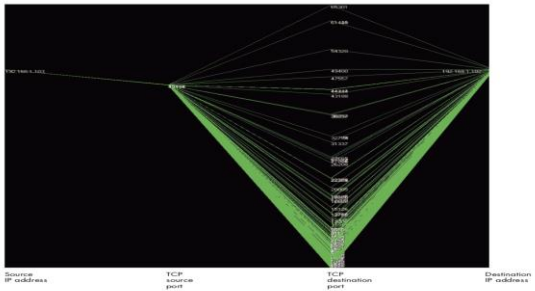


Figure 3-10: Filtering responses from the scan. By filtering the data to remove responses from the target computer, we more clearly see the probe packets sent by Nmap. Note that the tight clusters of TCP source and destination port values make exact values difficult to determine.

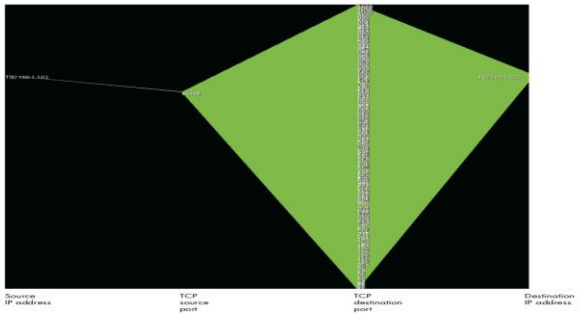


Figure 3-12: Zooming in on TCP destination ports. By zooming in on the well-known ports between 0 and 1023, it appears as if Nmap probed each port.

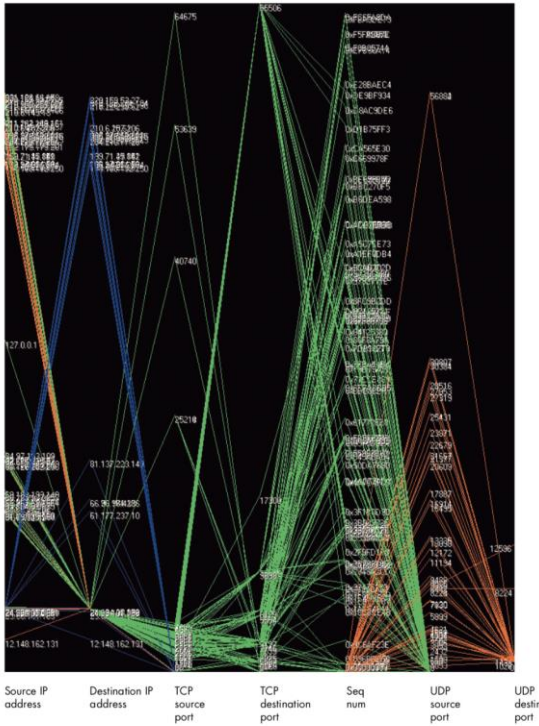


Figure 5-8: IP addressing and port information

Examples – Node Link

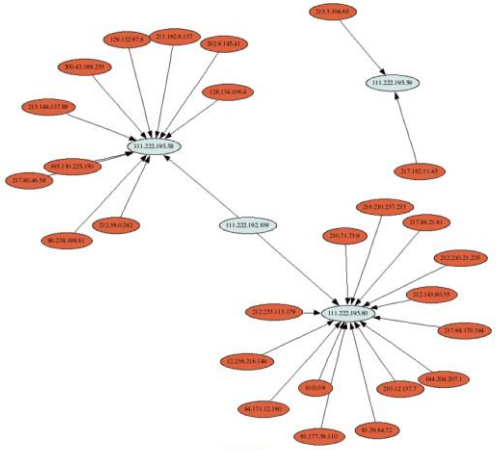


Figure 7-3: A link graph using source and destination addresses as nodes. An edge is drawn if a connection was detected between the addresses. Color provides additional information about the graph nodes; in this case the color signifies the IP address range: Machines on the internal network are blue and external machines are orange.

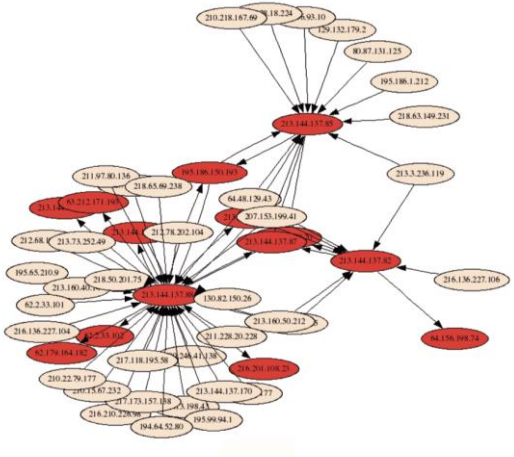


Figure 7-7: Graph showing source and destination nodes. The color assignment is such that a third, invisible field is used to determine the color of the nodes. If the source machine utilized a common spyware port to access the destination machine, the source node is colored red.

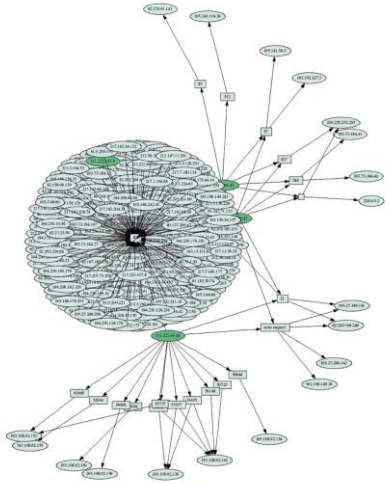
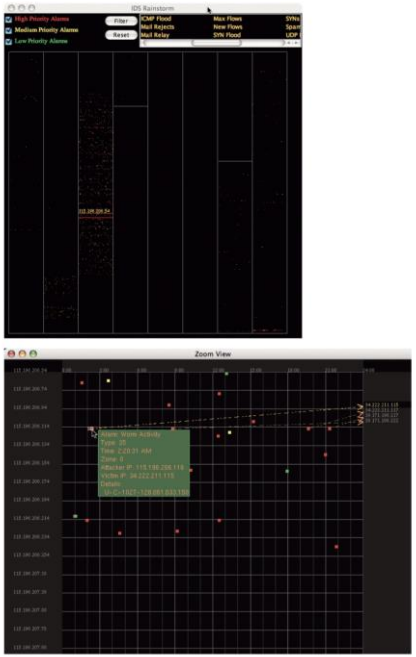


Figure 7-8: Blocked firewall outbound traffic. This graph helps to identify misconfigurations of either the firewall ruleset or the machine trying to connect to the outside. This view is only a first attempt and is cluttered by the large amount of web activity surrounding port 80. I'll improve the image by going through all six graph generation steps.

Examples – Treemaps and Scatters



Images by Kulsoom Abdullah. Used with permission.

Figure 6-12: Visualization of intrusion detection alerts using the IDS RainStorm system. Using a very long axis that wraps from the bottom of one column to the top of the next to plot IP addresses (vertical) and time (horizontal), a system administrator can monitor enterprise class networks (top). By selecting a region of interest, the user can zoom in to a more detailed view (bottom).

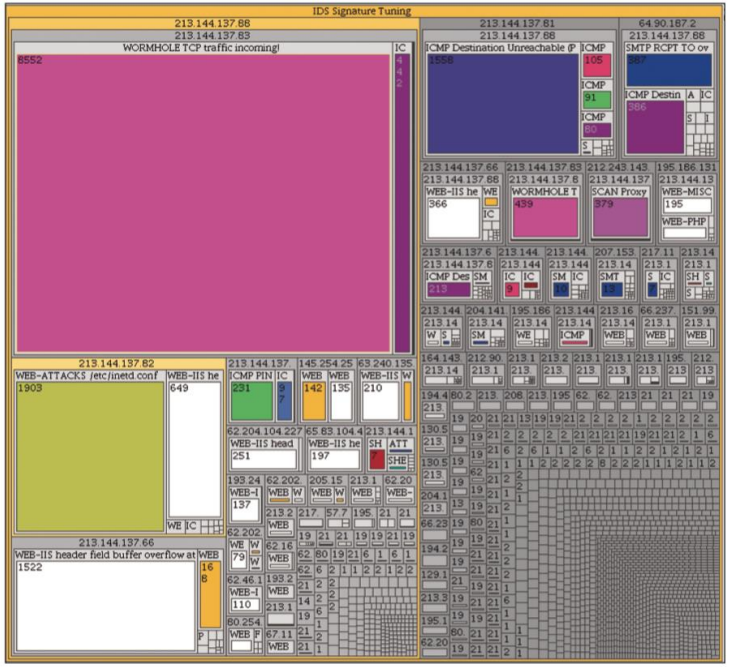


Figure 8-5: This TreeMap shows the Snort alert log from Figure 8-4 after eliminating the false positive, which took up most of the space in the initial TreeMap.

Visual Channels

Colour Size Orientation

Shape Texture Opacity

Visual Channels

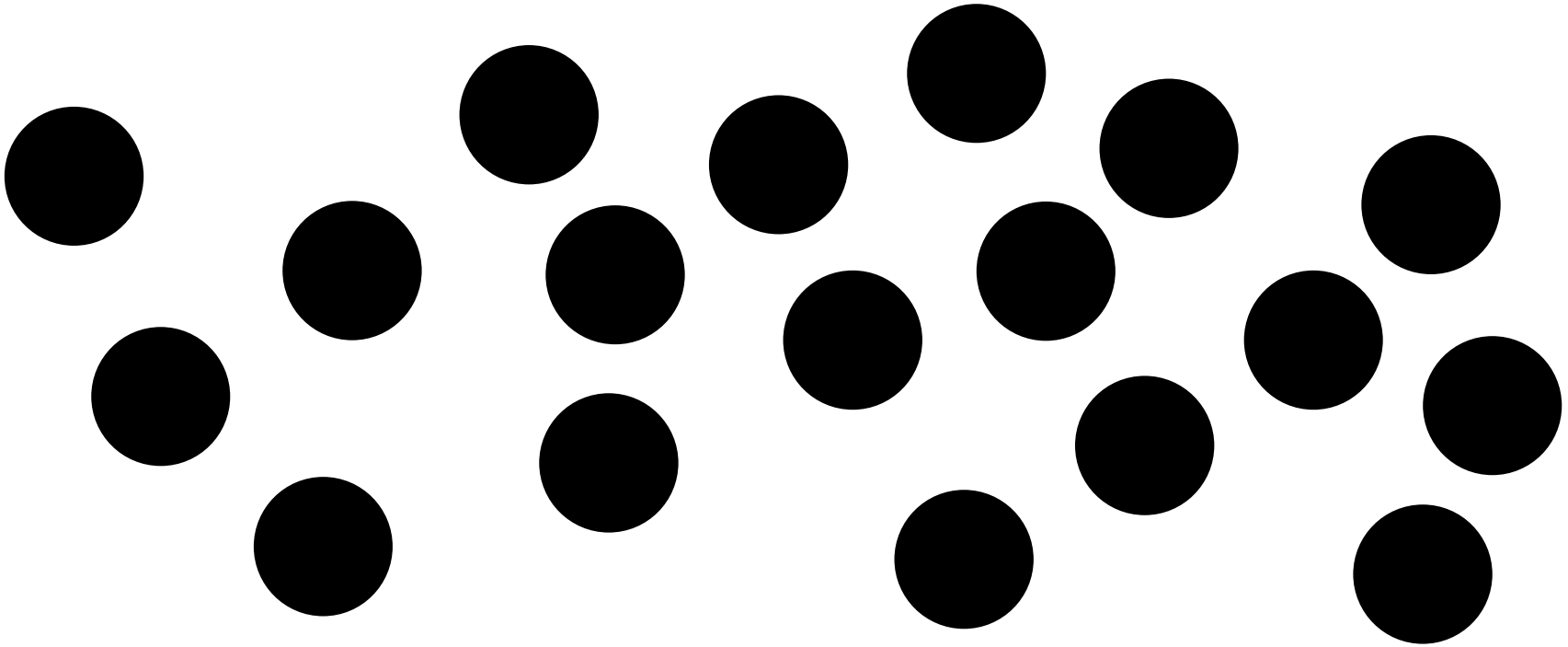
- Each data attribute is mapped to a visual cue or channel
  - Spatial positioning is also a visual channel (e.g., mapping data to axis)

- How do we know which channels should map to which data attribute?
  - No fixed rule
  - Think about the types of data:
    - Nominal: text labels (e.g., name) / categories (e.g., car type)
    - Ordinal: Data is ordered however difference is unknown (likert scale, threat level)
    - Interval: data is ordered and interval is known (e.g., temperature)
    - Ratio: As interval, but with an absolute zero (e.g., packet size)

http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

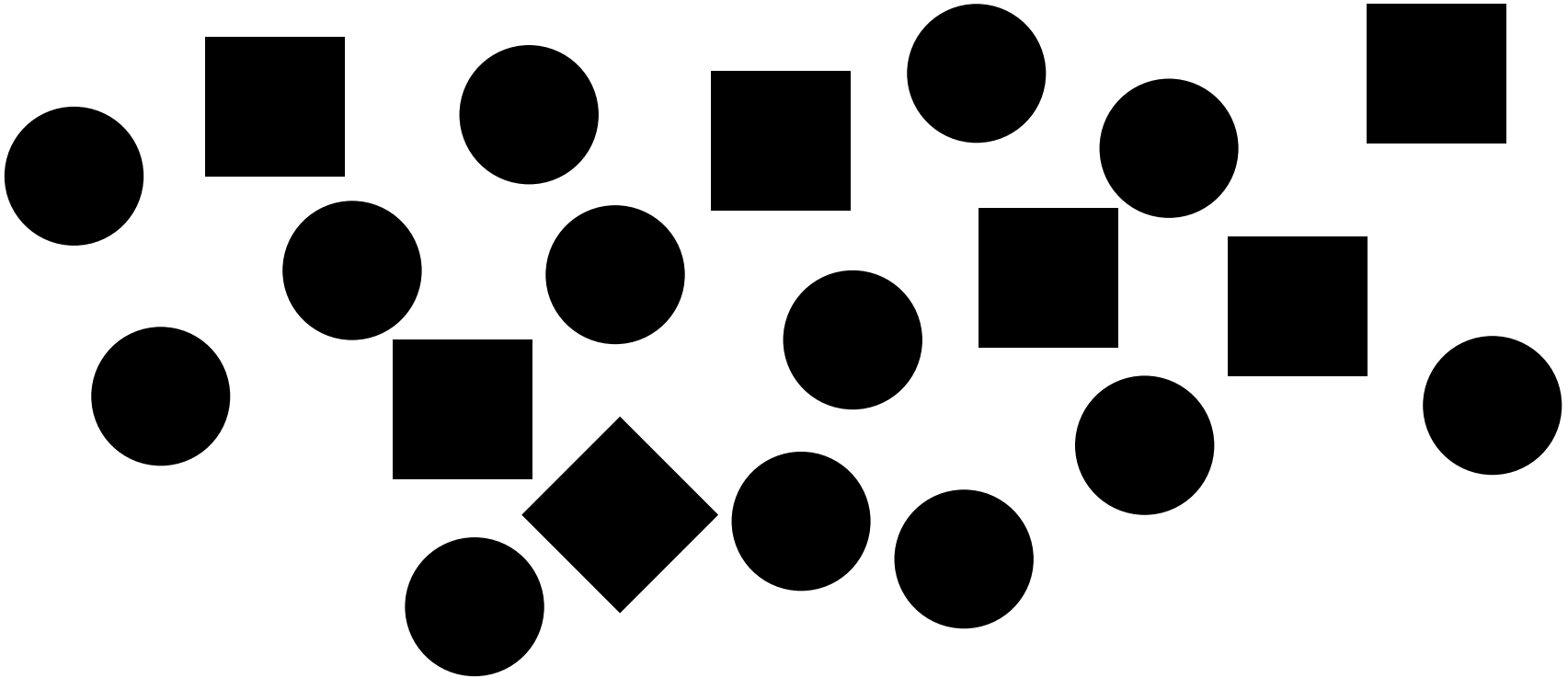
UWE Bristol University of the West of England

Visual Channels

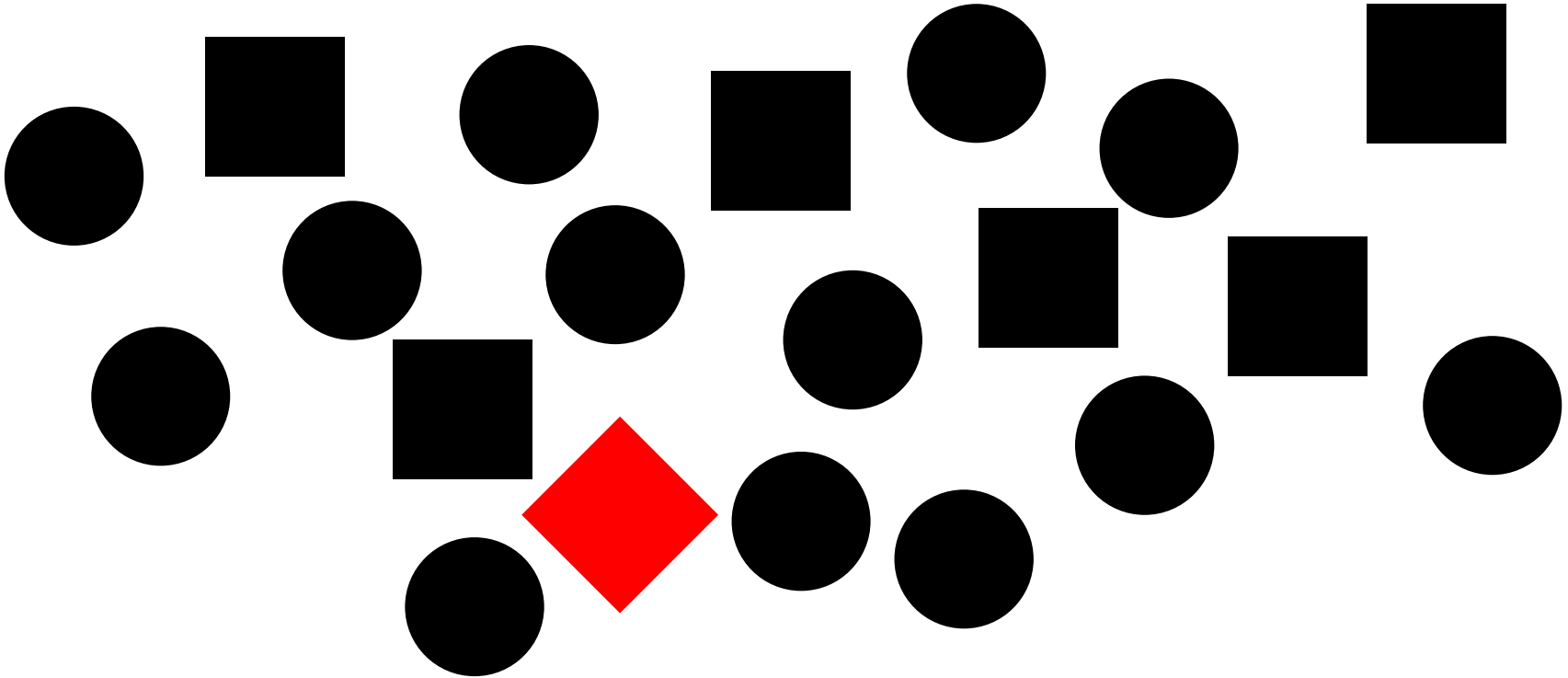


Visual Channels

Which attribute is the anomaly?

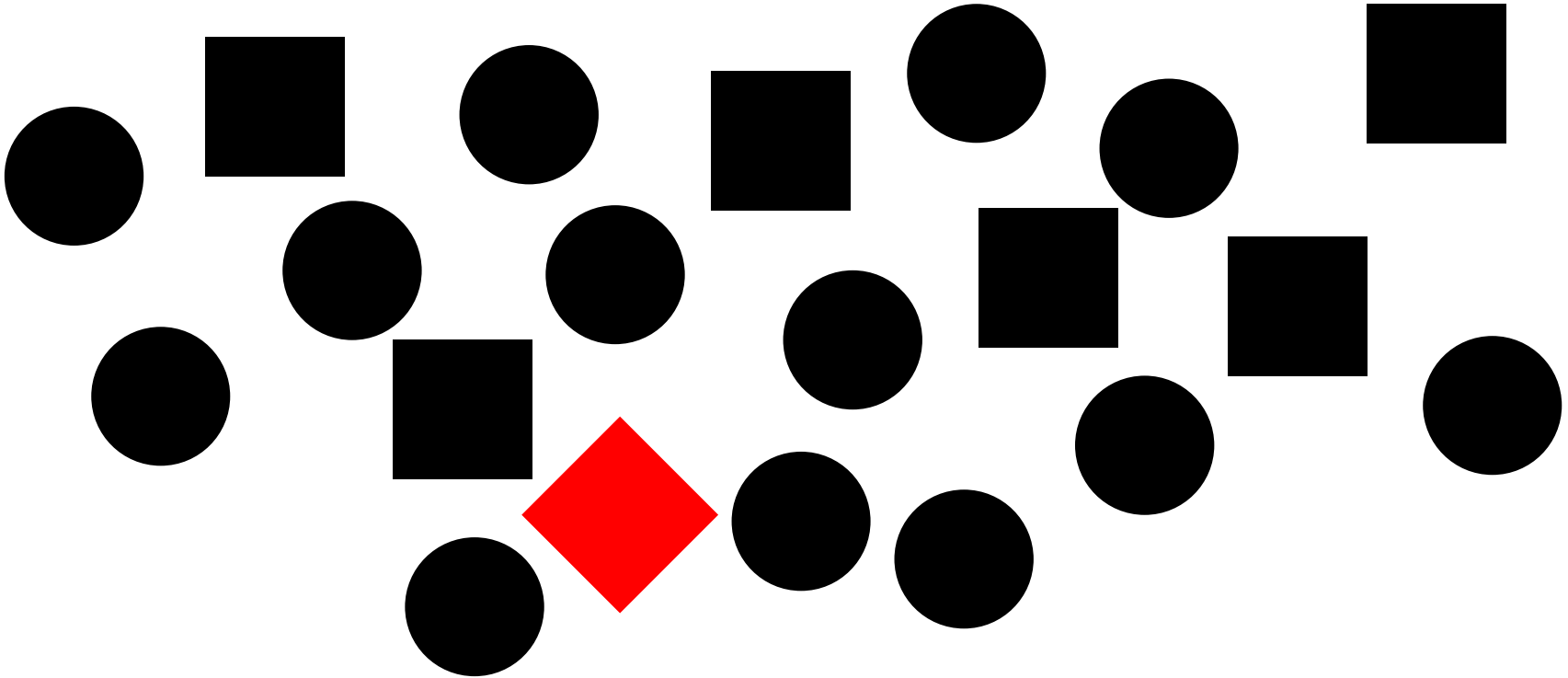


Visual Channels



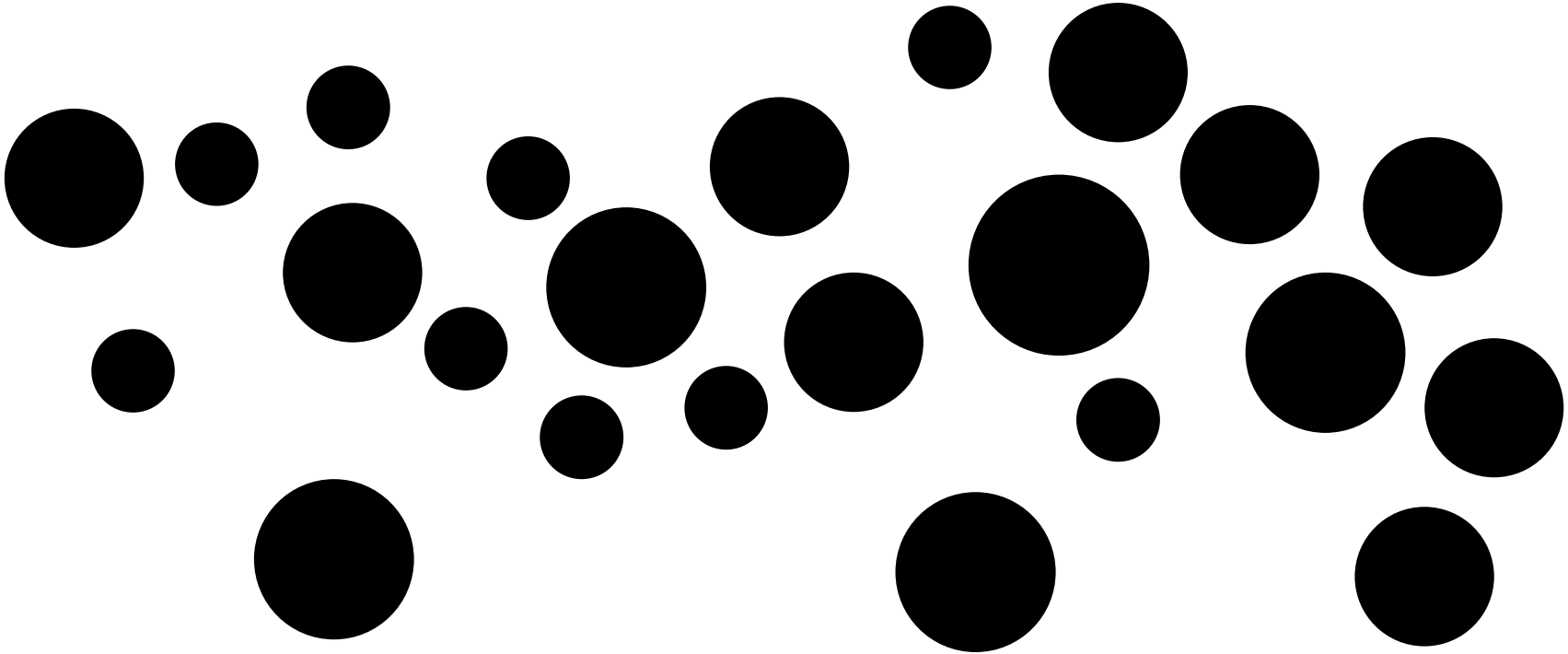
Visual Channels

When shape differs this is easier to find

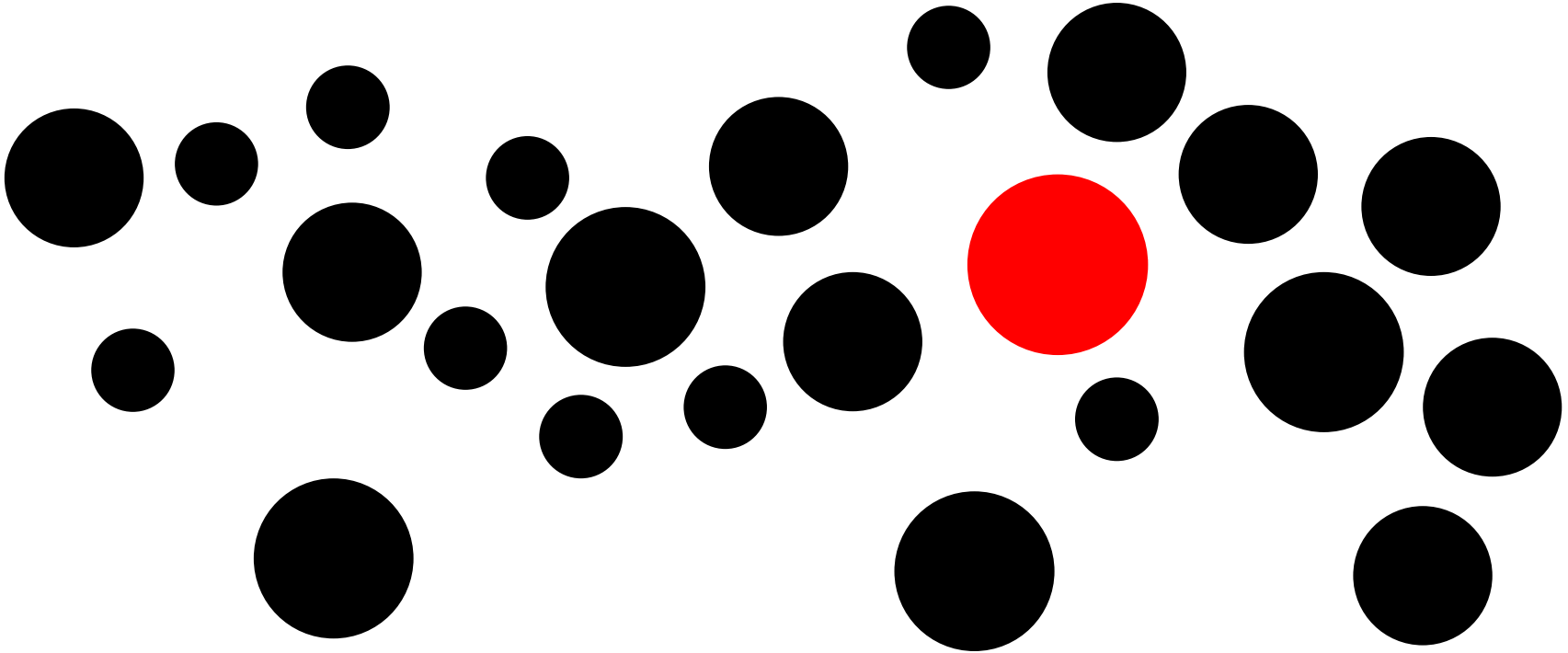


Visual Channels

Which attribute is the anomaly?

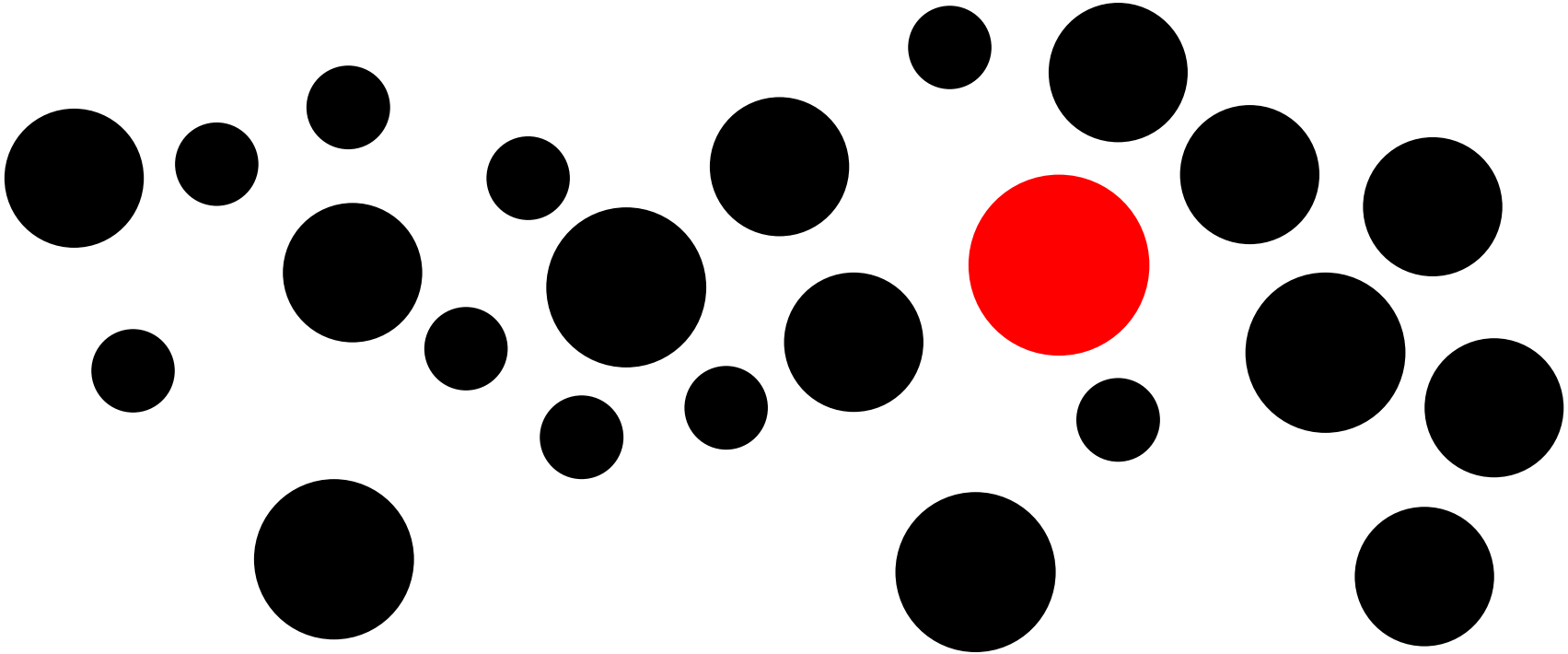


Visual Channels



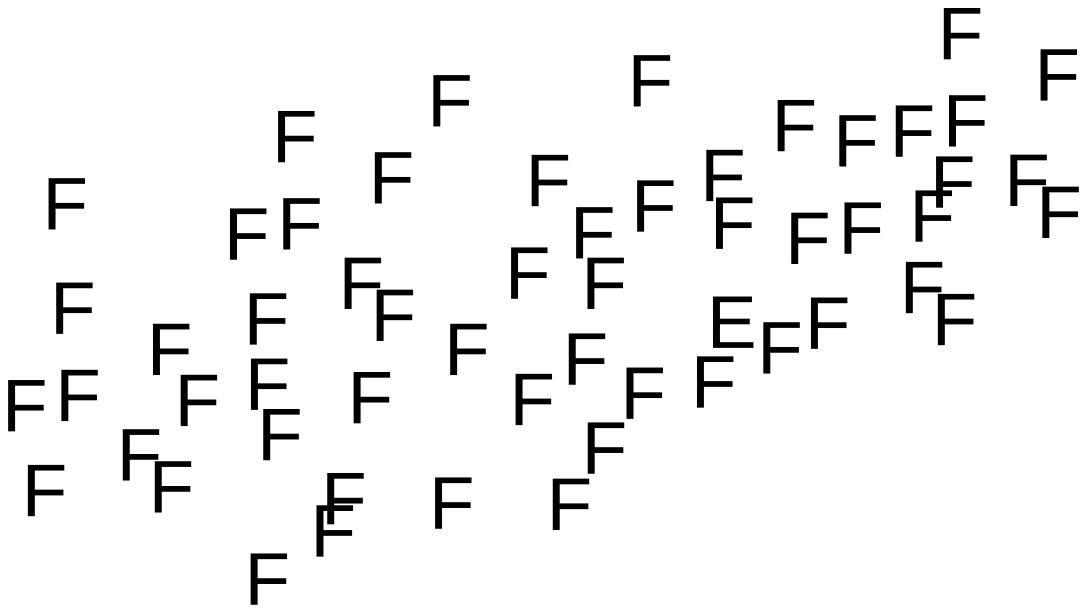
Visual Channels

Different to find when size is similar

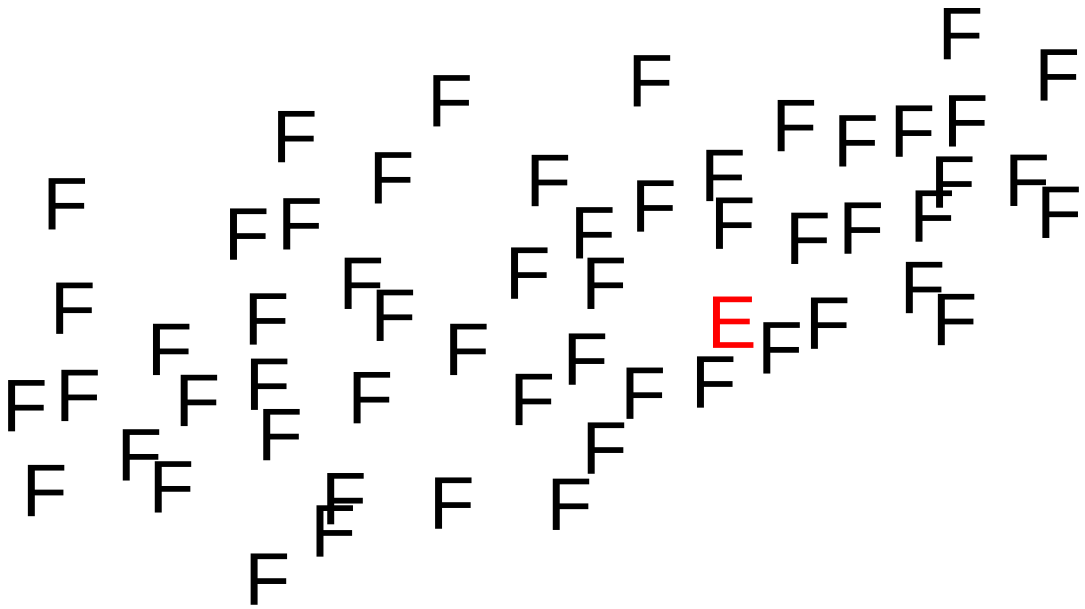


Visual Channels

Which attribute is the anomaly?

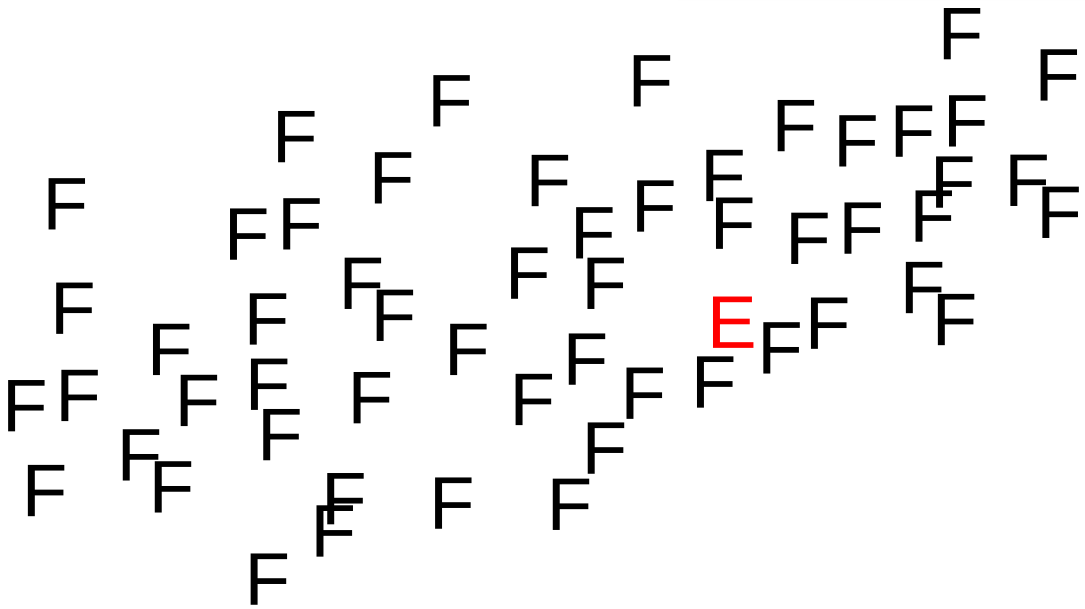


Visual Channels



Visual Channels

When similar shape, difficult to find

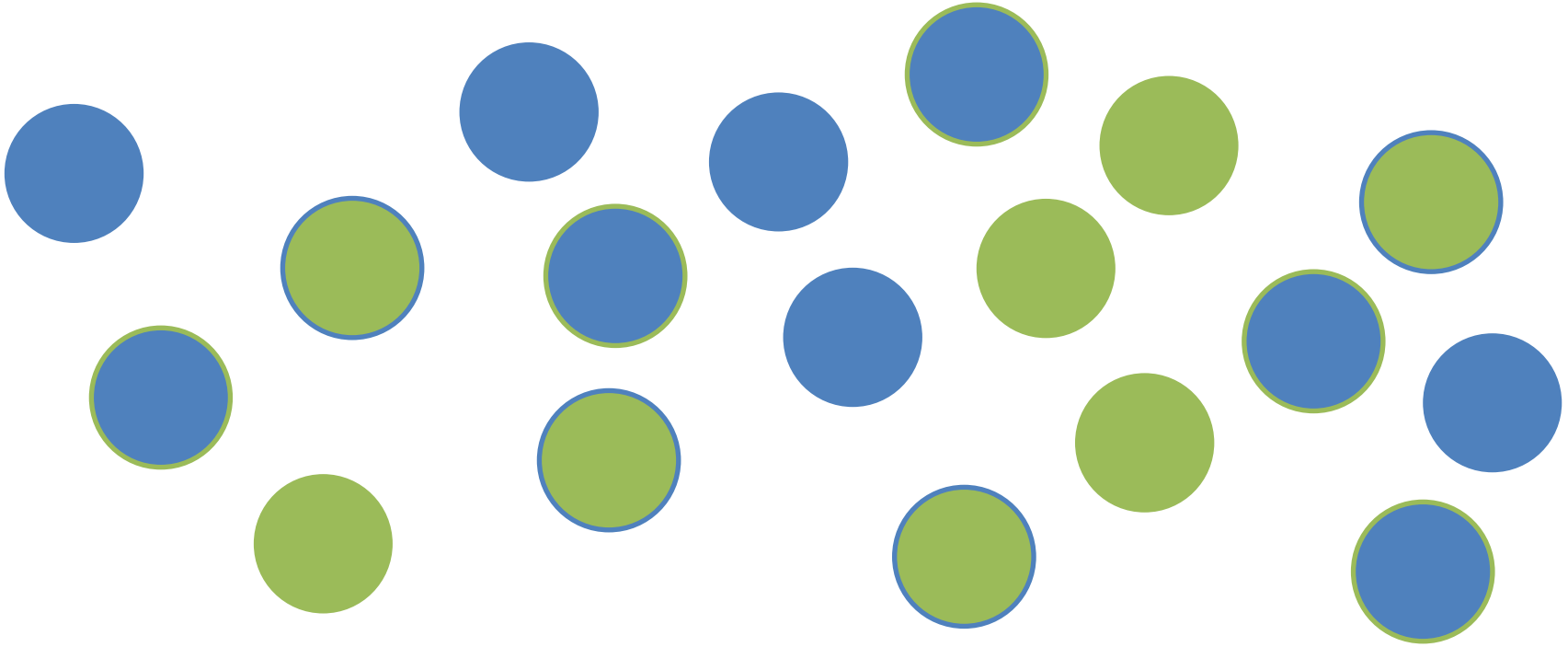


Visual Channels

5 categorical values mapped to colour

Visual Channels

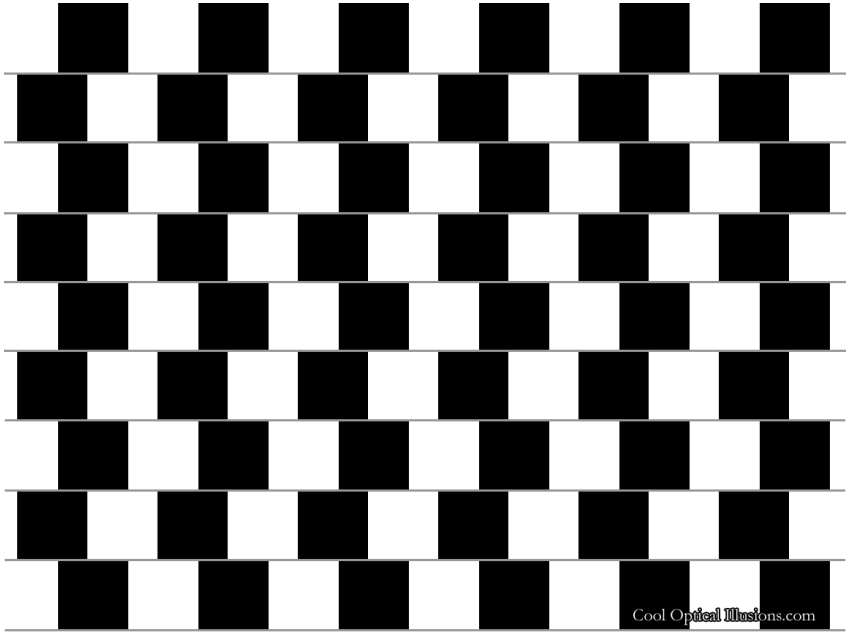
2 attributes for fill, 2 attributes for stroke



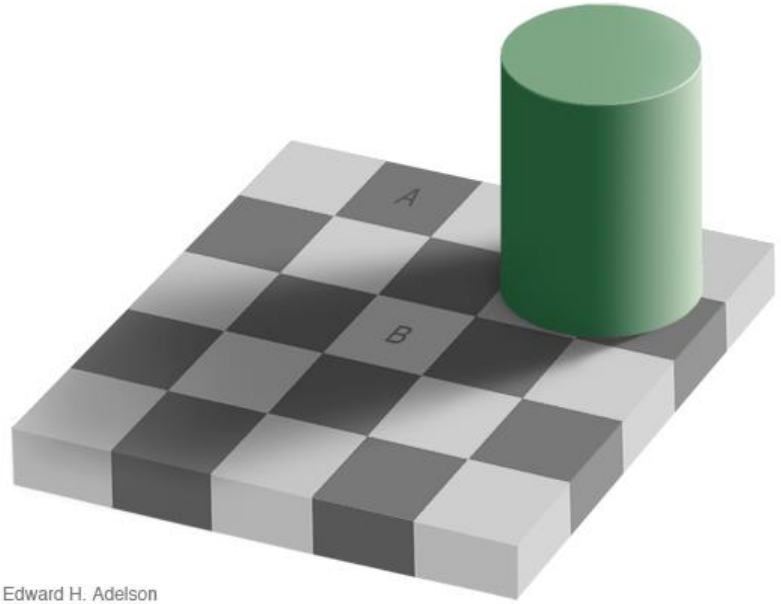
Always consider how humans will perceive visual representation...

... many cases where humans fail

Optical Illusions

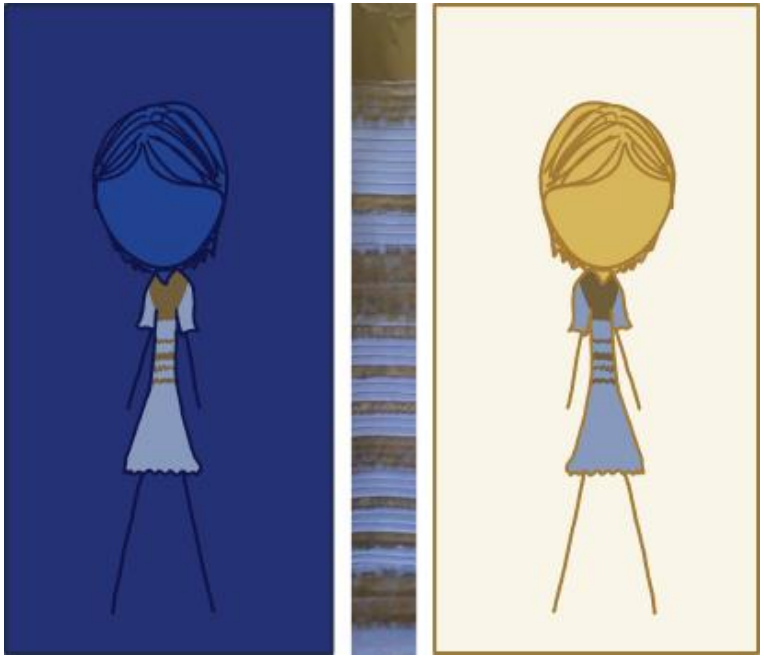


Optical Illusions



Edward H. Adelson

Optical Illusions



[12 optical illusions that show how colour can trick the eye]

UWE Bristol | University of the West of England

We need to be careful when visually encoding data…

… make sure it is not lost in translation

Takeaway

- What are the different techniques for visualization, when are they appropriate, and how can they help in security?
  - E.g., node-link, parallel co-ordinates, treemaps
- What are the different visual channels used to depict data in a visualization, and how are they best utilised?
  - E.g., colour, shape, size, texture, opacity, orientation…….
- How can we validate design choices to avoid unintended artefacts in our visualization?
  - E.g., avoid the ‘symptoms’ of an optical illusions