1. Mention and explain three uses of clustering in data visualisation.

ABSOLUTE NONSENSE, FIX IF DESIRED

a) To assess and measure the preservation of grouping, and separation of groups, when we project labelled high-dimensional data
b) Identify previously unknown categorisations and groupings in our data - colourizing these allows us to identify grouping much more easily in our visualisations
c) To more easily identify outliers and misclassified values - if their cluster membership is very different to their placement in a visualisation, we may wish to investigate the labelling system

2. Mention and explain three uses of force-directed algorithms in data visualisation.

a) To reduce the number of overlapping edges in a graph visualisation (typically only when the graph is either small or sparse)
b) To spread out potential small, isolated groupings in a network based visualisation, that may have all been collapsed to a single point
c) Potential patterns, including chains, loops, diverging branches and subtrees etc may become much more visible on a force directed graph

3. What are stack (or stacked) graphs? What is their main use? How would you apply them for text visualisations?

A form of graph where instead of measuring an attribute for every point against the y axis from 0, we group points by some criteria - typically a time period, and place the y axis measurements one atop the other.

a) Stacked bar chart - we effectively take bars of different colours, place them one atop the other
b) Stacked line graph - (*topic river*) the distance from one line to the line below it represents it's value at that time - not the distance to the x axis

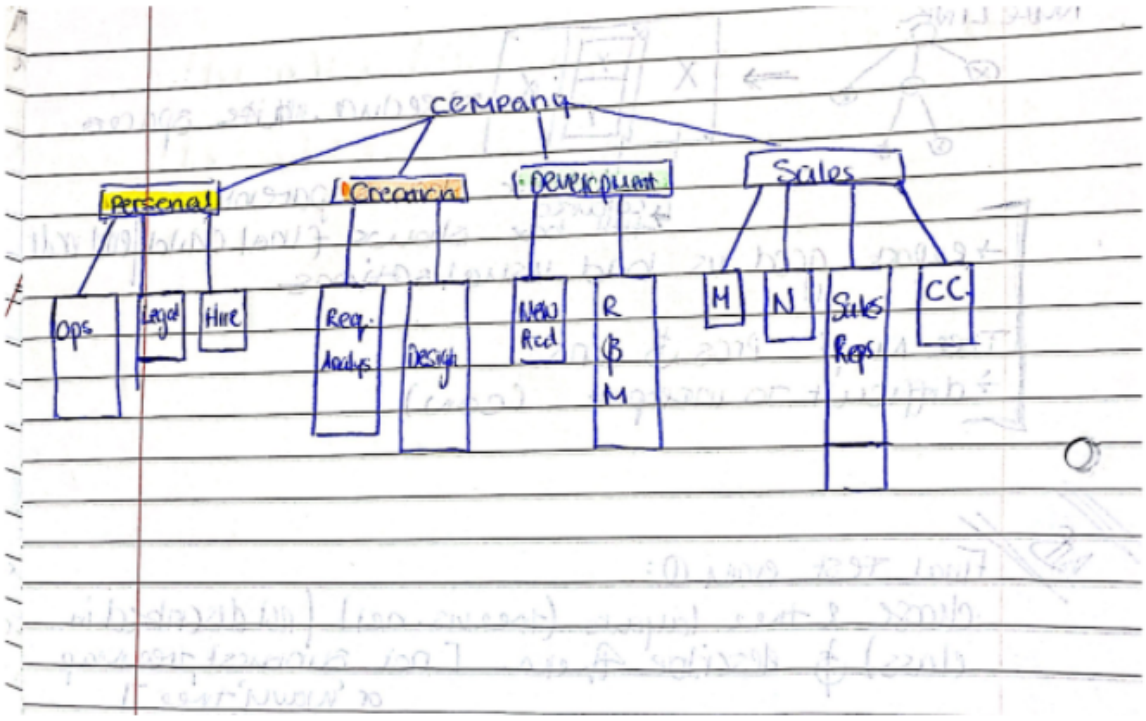
This is used most often when we want to compare *proportions* of different attributes over time - it becomes almost impossible to isolate trends for a particular attribute, **especially** for a stacked line/time series plot.

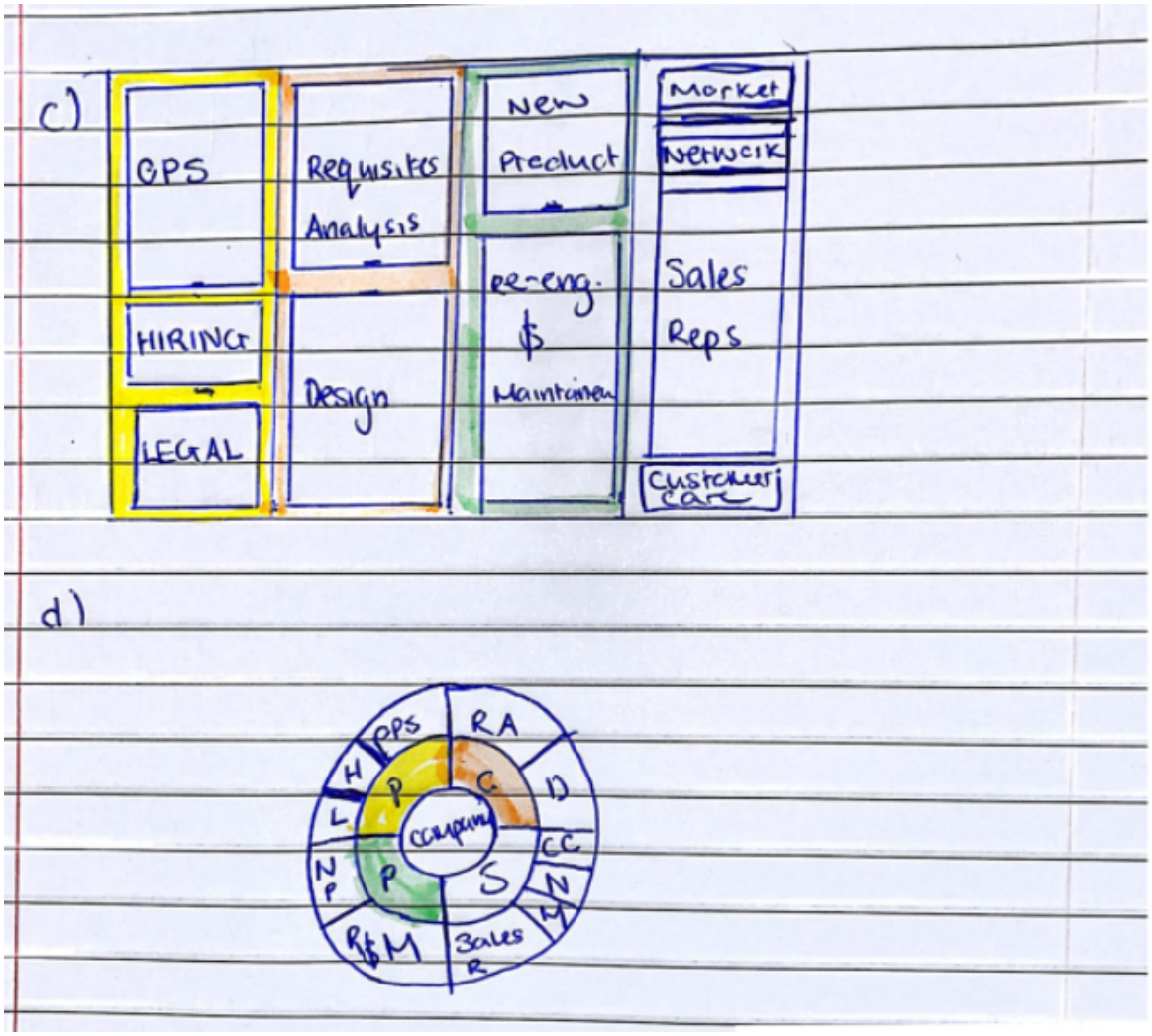
For text visualisations, we can either identify particular key words, or use projection (possible word/sentence/BERT-paragraph vector encoding) to identify key phrases or semantics in different corpora of interest. We then construct a stacked line/topic river plot, that compares the proportion of the popularity of these topics over time - this works particularly well for text analysis, as we rarely care about the absolute value of the usage of a particular term, or the exact number of points that are categorised as a certain semantic. What we are interested in, is the chance of popularity, relevance or importance of these aspects over time.

4. An analyst is organising a data set for managing the numbers of employees of the various departments of the company. The following text describes the data: ***stuff in the document***. Represent that data:

(a) as a table/spreadsheet
(b) as a nodelink tree
(c) as a treemap
(d) as a ring tree (sunburst)

Directorates	Department	Number of employees
Personal	Operations	8
Personal	Hiring	3
Personal	legal	4
Sales	marketing	3
Sales	networking	4
Sales	Sales representations	20
Sales	Customer care	5
Development	New product	10
Development	Re-engineering and maintenance	25
Creation	Requisites analysis	10
Creation	Design	12





[^ I didn't include no. of employees per dept. But potentially, that information might be needed]

5. An analyst is designing a visualisation to support real estate agents to keep control of contents of houses they rent. She has built a fake data set to test alternatives. The following text describes the data: The house has two sections: upstairs and downstairs. Upstairs has two bedrooms, one office and one bathroom. Downstairs has a kitchen, living room, laundry room, bathroom and lounge. Bedroom one has a bed and chair. Bedroom two has a bed, chair and waste bin. Office has a chair, couch, waste bin and lamp. Upstairs bathroom has a waste bin and a small cupboard. Kitchen has a table, 5 chairs, and a cupboard. Living room has a cupboard, big couch, small couch, coffee table, chair and lamp. Laundry room has a washing machine and dryer. Downstairs bathroom has a waste bin and cleaning brush.

Represent that data:

(a) as a table/spreadsheet 1
(b) as a nodelink tree
(c) as a treemap
(d) as a ring tree (sunburst)

6. Draw a comparative table of 5 multidimensional visualisation techniques. What are they useful for? What type of observations do they provide? What is their input? What are the advantages and disadvantages? What applications could they support?

TECHNIQUE 1: LSP

USES:

OBSERVATIONS:

INPUT:

ADVANTAGES/DISADVANTAGES:

APPLICATIONS:

TECHNIQUE 2: PCA

USES:

OBSERVATIONS:

INPUT:

ADVANTAGES/DISADVANTAGES:

APPLICATIONS:

TECHNIQUE 3: K-MEANS/MEDIODS CLUSTERING

Uses: used to find groups which have not been explicitly labeled in the data.

Observations: [this part is a guess] The bigger is the K you choose, the lower will be the variance within the groups in the clustering. K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster or cluster liner), serving as a prototype of the cluster.

Input: the number of clusters K and the data set.

Adv: Relatively simple to implement. Scales to large data sets. Guarantees convergence. Easily adapts to new examples.

Dis: **Clustering data of varying sizes and density.** (k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means). **Clustering outliers.** (Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.) **Scaling with number of dimensions.** (As the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. Reduce dimensionality by using PCA.)

Medoid: Representative objects of a data set or a cluster within a data set whose sum of dissimilarities to all the objects in the cluster is minimal.

Applications: Recommender systems (grouping together users with similar viewing patterns on Netflix, in order to recommend similar content).Anomaly detection (fraud detection, detecting defective mechanical parts). Genetics (clustering DNA patterns to analyze evolutionary biology)

TECHNIQUE 4: FORCE BASED PLACEMENT

USES: 1) Creating graphical representations of networks of high-dimensional points - where these are implicitly inferred from similarity measures of collections of features, rather than hardcoded edge weights. 2) reducing the number of overlapping edges, and improving separation of groups in a network visualisation or projection.

OBSERVATIONS:

INPUT: Rows containing numerous features, and a similarity measure that can be applied over two sets of features

ADVANTAGES/DISADVANTAGES: Not reliant on hard-coded network definitions, can produce excellently communicable and interpretable plots for non-complex graphs, and improve these aspects of multidimensional projections. Is quite slow, and ineffective on complex networks.

APPLICATIONS:

TECHNIQUE 5: MATRICES OF SCATTER PLOTS

USES: Can be very effective in identifying relationships between all pairs of variables (and the direction of the relationships) in the dataset

OBSERVATIONS: Correlation between variables, nature of relationships - negative linear, positive quadratic, negative exponential etc etc

INPUT: Data with a number (greater than two) of numeric, ordinal features.

ADVANTAGES/DISADVANTAGES: It is of unparalleled effectiveness in visually identifying relationships between pairs of features. When we have transitive dependencies, this can be harder to interpret (when variables A and C are both correlated with variable B, both can appear to correlate with each other). Further, interactive/complex dependencies can be hard to identify - when a variable is dependant on the interaction between more than one other variable.

APPLICATIONS: Multi-dimensional regression

7. Visualisation Task: Suppose you want to produce a visualisation that summarises and helps as a guide to a web-based course. The course is composed of a series of linked web pages plus external links and references. Pages can be classes, tasks or additional information. The user must understand the general structure of the course, verify what pages he/she visited and for how long, and quickly find references given for each class. Describe the elements of your visualisation and discuss the possible interactive functionalities the user would have in this context. Draw a schema for your visualisation. Justify your choices of visual elements. What would change if the visualisation were designed to support the teacher in improving the course?

8. Visualisation Task: Given the Data set in Table 1, draw the parallel coordinates visualisation for it. What can you conclude about the data from the visualisation? SEE SHEET FOR TABLE

9. What is the distinction you make between Information Visualisation (InfoVis) and Scientific Visualisation (SciVis)? Mention two common points and five distinct features between them.

COMMON POINTS

a) In both, we are visually exploring our data. Both are usually employed in applications where large volumes of complex data are involved - we cannot describe our data in full with simple summary statistics, and there is too great a volume of information to comprehend by simply reading figures
b) In both cases, we are often entering the visualisation phase with one of two tasks in mind - to explore the data for unexpected behaviours, discrepancies or patterns that we may not have expected, or to assess if some hypothesis or assumption about our data likely holds.

DISTINCTIONS

a) **Domain of application -** in general, info vis is the term for any use of visual representations of some sort of data - in research, business, PR, etc. SciVis is a subset of this, referring to the use of visual representations of some sort of physical phenomena under experimental conditions.
b) **Expectations of input data -** the term InfoVis makes no suggestion as to what form the daat involved with take - SciVis makes a heavy implication that the data being visualised is spatial (in a well-defined either 1 2 or 3 dimensional space), and also likely temporal
c) **Level of abstraction -** InfoV is a very abstract term used to encapsulate the notion of visualising abstract information, whereas SciVis is the term for visualising either observed or simulated measurements of physical processes
d) **Transformation** In infoVis, it is generally necessary to perform all manner of transformations on our data - projections, mappings, various styles of visualisations that summarise the data in different ways. In SciVis, we are usually constrained by the geometry of the space in which the data were measured, and cannot actually transform the data in any drastic way from its raw form - (beyong taking logs of observation measures and similar). For this reason, SciVis is much more concerned with taking raw data, and presenting it in visual form, which contrasts with InfoVis and the elements of data mining and transformation it involves

10. What differs InfoVis and SciVis regarding data types?

a) Info vis involves collecting large quantities of data, often with many different attributes, for exploratory purposes. This data varies massively in type - numerical, ordinal, categorical, etc. often, we are not sure what underlying processes are taking place, what attributes are unnecessary or useless to us, and how we should present the data - there is no concrete, obvious definition of how our data should be treated or presented, it has often no literal physical correspondence. To transform info vis data to be visualised, we have to transform the data to map it to some convenient representation in time and space. These may already exist in the data, but even then, are optional in the visualisation of it.
b) In scientific visualisation, our data is a specific (either measured or simulated) description of a temporal or physical phenomena - our attributes will certainly have either a physical (3D, 2D or 1D) location in space, and almost always a specific point in time at which they are observed/simulated to occur. Visualisations that disregard, or otherwise distort these critical attributes rarely assist us in the interpretation of the data - e.g. modelling an animal's metabolic activity over time could very likely become less meaningful if we project or distort the time attribute of this data, or in aerodynamics, temperature and pressure measured at critical points over a wing cannot be meaningfully visualised without including some representation of the wing, and the space around it

11. What is the distinctions between Direct Volume Rendering and Surface based Rendering?

Both used to visualise 3D datasets.
Surface based rendering is used where we know the underlying geometry of objects in the space we have measurements for, and renders data points based on this knowledge of the

geometry - e.g. in aerodynamics where we measure pressure at different points on an aerofoil, we know where on this pre-existing surface each point was measured - we can render a static 3D model of the known surface, and visualise the measurements atop this
Where geometric features in the scene are small, complex and ill-defined by the experimenter, it is more useful to make no assumption about the underlying geometry of the scene and employ direct volume rendering to reconstruct and render the geometry of the scene around the observed points

12. What are the distinctions of 2D reconstruction and 3D reconstructions and in what circumstance each one would be applied?
~~WHEN IN THE NAME OF FUCK WAS THIS FEATURED ON THE COURSE?? INTERESTING NONETHELESS.~~

Writing this in orange by reason of its being spoofed to 376%. Do not internalise any of this with accepting this it is: over-convoluted, likely incorrect, and irrelevant to the question

In many (a colossal proportion) scientific visualisations, we are observing the value of some measurements in different locations in 3D space - e.g hydrofoil, brain activity, atmospheric pressure (E.G. SIMULATING THE HIGH PRESSURE SYSTEM THAT WAS INFLUENCING THE OBSERVED WEATHER CONDITIONS *SOUTH-WEST* OF UCC AROUND 7PM ON MARCH 3RD). We can visualise this data with colorised (in 3D, this requires also assigning opacity) points in a 3D interactive scene. Static 2D renders can be taken of this scene, but these are still ultimately the result of a 3D reconstruction. This has an obvious glaring flaw - internal details are lost. In the case of visualising dynamics over an aerofoil or hydroil, all of the action takes place over the surface - effectively, we are rendering **onto** a surface in 3D. For a brain activity measurement, we are left with data distributed throughout the internal space of the cranial cavity. A 3D reconstruction will give us a sense of the activity on the outer surface portions of the brain. Whatever 2D render we make on this however, will not reveal the internal measurements.

~~The solution here is to effectively project the 3D space into 2D - but in very specific ways that do not distort space. A common method is to take parallel, thin cross-sectional slices of space, and treating them as a 2D plane, render the measurements on this (along with some reference geometry that indicates where in space the slice is taken). Making these slices interactive (allowing the user to slice up and down, while displaying the renders on-the-fly) can give an excellent illusion of exploring 3D space, without a loss of insight into internal observations.~~

2D reconstruction involves generating contours from the measurements, and rendering surfaces that have 3D geometry (but unlike 3D reconstructions, have no internal information/density etc)

13. What are the main strategies to visualise hierarchies? What distinguishes them?
1. TreeMap
The first level of the hierarchy is shown in rectangles, the size determined by some measure. Then, each rectangle is further subdivided into smaller rectangles for the next level of the

hierarchy and so on. Colour is typically used to highlight one of the hierarchical levels. Due to its rectangular nature, treemaps are able to remain very compact and use the space on screen to its fullest. Theoretically, there is no limit to the number of hierarchical levels that can be visualised with a treemap, but in practice, the data becomes very difficult to see the more levels that you add.

2. Sunburst
Sunbursts are a series of rings, which represent the different hierarchical levels. The innermost ring is the first level, followed by the second level which shows a breakdown of the components of the first, and so on. Like the rectangular treemap, the size of the arc represents the magnitude of a metric and colour is often used to distinguish some level of the hierarchy.

3. Circle Packing
This type of visualisation is similar to packed bubbles, but it addresses the inability to see the relationships between the hierarchical levels, as well as the problem of showing the magnitude of parent levels. Unfortunately, circle packing charts suffer from other problems. The use of circles means that there is a lot of wasted space. Labels can be quite difficult to read, particularly as the number of levels increase.

(prob should also add node link oops)

14. Give an example of an application and a task or question for which a hierarchy is a natural application. Give an example of an application and a task or question for which a hierarchy can be used, but is non-native.

Natural application of Hierarchy:

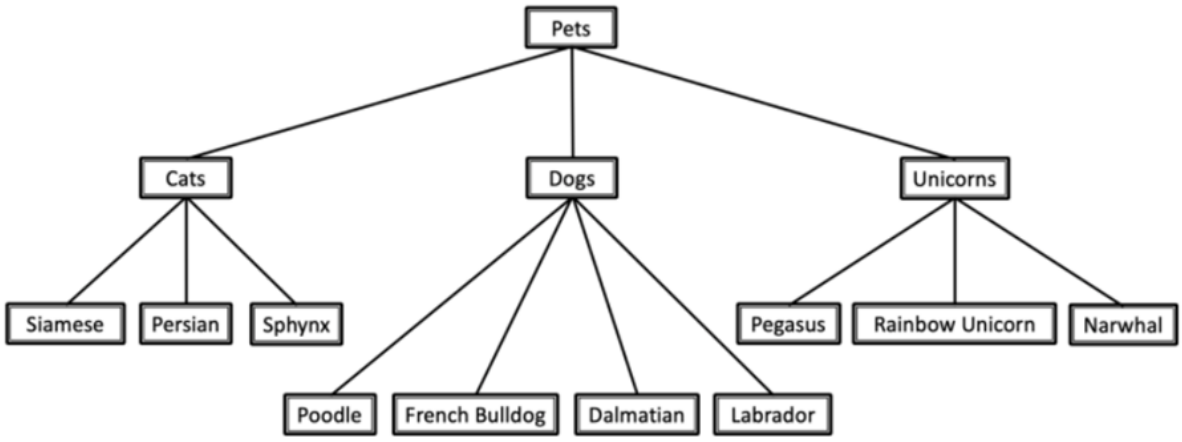
Application: file and folder system on a PC, using an indentation tree. Family Trees, using a node-link tree. Departments within a company, using a TreeMap. (see q.4)

Task/Question: sort the above folder, sub-folders and files. Create a family tree beginning at your grandfather. Create a representation of the different departments and sub departments in this company.

Non-native application of Hierarchy:

Application: classification (taxonomic data), classification of animals for example.

[https://towardsdatascience.com/https-medium-com-noa-weiss-the-hitchhikers-guide-to-hierarchical-classification-f8428ea1e076]



Task/Question: separate and classify the follow animals. <<insert imaginary list of animals>>

[non-native is a pure guess, use at own risk]

15. Mention and explain three different types of layout for a tree, as well as their advantages and disadvantages.

Balloon tree Layout: is a tree layouter that positions the subtrees rooted at a node in a radial fashion around that node. It is ideally suited for huge trees (say, 10,000 nodes) since it computes fast layouts that are quite compact. Advantages are that it is good for grouping related objects together and suitable for diagrams that contain a lot of objects

Hierarchic : The hierarchic layout organises the objects in a hierarchical structure, where objects are placed in tiers above the objects they reference.

Orthogonal : The orthogonal layout places objects so that the lines between objects are as close to vertical and horizontal as possible.

16. What are the main strategies for visualizing networks? What are the main applications? What would be the role of centrality measures in these applications? What can be done to improve clutter problems in network visualization? What can be done to support visual analysis of large networks?

Strategies and applications

We have to decide what sort of graph we want to employ - for small numbers of nodes, with high numbers of edges, where only pairwise relationships matter we'd consider a ring/chord. For small numbers of edges, arcs may be appropriate. More generally, if a relationship between notes is hard-coded in the data, it is best to employ a spectral graph, and if it is inferred from featuresets, it may be better to employ force based placement.

Centrality

Centrality measures can serve multiple purposes, depending greatly on the topography of the specific networks in question. Generally, they serve to identify the most interesting, or important nodes in the network (based on volume of connections). E.g, nodes with highest degree are clearly important, as are nodes with high *"betweenness"* (nodes which when removed, increase the average path lengths between many other nodes drastically). This assumes that in any network there will be a large number of participants, but often a great deal of important behaviour can be understood by studying a few highly influential nodes (by some measure of influence)

Clutter

Definitely the simplest, and most effective way to reduce clutter is to find a mechanism to reduce our number of nodes and edges in a network. To remove nodes, in some cases, we can employ necessary clustering algorithms, or take a K-nearest neighbours approach. Both of these will group similar nodes, which we can then conglomerate into individual super-nodes. Note that edges having weights is necessary for this to be really effective, as otherwise these algorithms, not originally designed to be applied to graph data, may find it difficult to select centroids/mediods and thus implicitly define clear borders between clusters

A similar idea can be used, by identifying the strongly connected components of the network, and representing these as single nodes can greatly simplify the graph.

Removing edges is generally accomplished as a free side effect of reducing the number of nodes in a graph, but can also be accomplished separately. The simplest way is to set some threshold below, where edges with weights below this value are not displayed - including this as an interactive element of the visualisation is particularly useful, as it allows the user to get a sense of the topography of the network at a variety of resolutions, by balancing clutter and simplicity with information conveyed

Force based placement - this is sometimes effective, when there is a small number of nodes or edges that are compressed into a tiny area. For very dense graphs, with huge amounts of overlapping edges, this is a futile endeavour.

RING/CHORD graph - use bundling

Supporting work on large networks

The aforementioned simplification measures can be used to ease analysis. Something like XHipp combines the above ideas with projection.

Simplification by filtering on raw data, reduction on the geometry level when creating the graph and interaction on the image level during rendering,

17. Mention and explain three different types of layouts for a tree, as well as their main advantage.

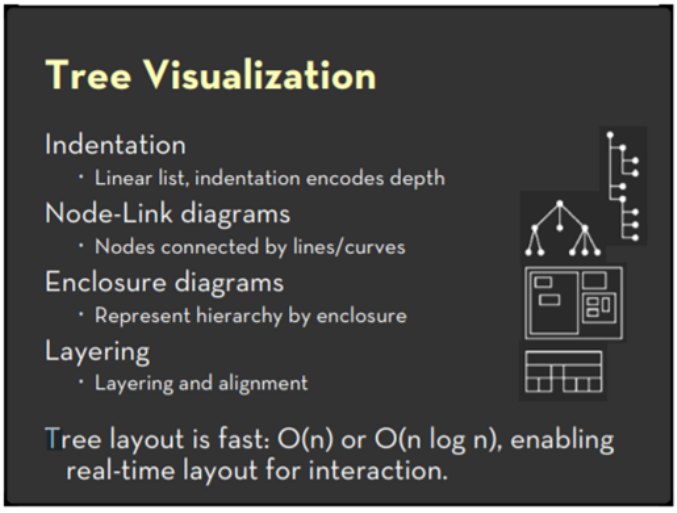
Indentation: places items along vertically spaced rows. **ADV**: indentation clearly shows parent/child relationships. **DIS**: Breadth and depth content for space, requires large amount of scrolling

Node-Link: Nodes are distributed in space, connected by straight or curved lines. Typical approach is to use 2D space to break apart breadth and depth. Often space is used to communicate hierarchical orientation. **ADV**: can easily and repeated divide for subtrees, shows hierarchy **DIS**: Tree breadth often grows exponentially, even with tidier layout, we quickly run out of space

Enclosure: encodes structure using spatial enclosure. **ADV**: Provides a single view of an entire tree. Easier to spot large/small nodes. **DIS**: Difficult to accurately read depth

Layering: Signify tree structure using layering, adjacency, alignment. **ADV**: Involves recursive sub-division of space

[^answer from first question sheet 1,maybe need to redo with new types like sunburst that she mentioned in class (https://treevis.net/)]



18. What insights can be drawn when visualizing text using: force-directed graphs, word trees and similarity trees? What is the purpose of word clouds?

FORCE-DIRECTED GRAPHS -

WORD TREES - word trees can preserve hierarchy in text - either by grammar (sentence order) or specification (e.g. life -> kingdom -> family -> species).

SIMILARITY TREES - can be used to gain insight as to which words and phrases appear together often. Undirected tree

WORD CLOUDS - the quickest and dirtiest way of performing sentiment analysis - but this also makes it the easiest to visualise, communicate, and understand. We can see what words and phrases overall characterise a document or corpus. We see these words directly (rather than needing to examine abstract points on a place/cubic space), and people can intuitively interpret the cloud without any explanation or background information necessary.

19. What is the role of data summarization in InfoVis? Exemplify data summarization for two different types of data.

Summarisation takes some large volume of information, and by discarding some quantity of resolution, condenses it to a smaller volume - > this is most useful for communicating some phenomenon, but can also allow us to reduce dimensionality, variability and scale,making patterns easier to identify in visualisations.

For point based data that can be observed to cluster very effectively (e.g. high silhouette coefficient), we can represent points in our dataset by the centroids of their respective clusters - reducing potentially thousands or millions of points to a representative handful.

Taking simple means of numerical data, especially when grouped by a categorical attribute of interest can often be extremely effective at noticing trends between groups. Boxplots extend this notion by compactly representing interquartile ranges along with the median by category.

For text data, using a vector encoding scheme (e.g. word2vec/sentence2vec/BERT), or even something simpler and more case specific, we can generate a friendly, fixed length (relative small compared to the dimensionality of a full document, but too large to directly visualise) representation that is easier to work further with

20.

(a) For the NJ-tree below, which of the Distance matrices CANNOT be the one that generated it? Why?

(b) Draw an NJ-tree for the answer in 20a