overview: scalability

Big data, where we have millions+ of data points hard to visualise;
Perceptual Scalability: Hard for user to perceive and understand cluttered datapoints on limited space;

Possible Solutions:
- Alpha Blending
- Filtering
- Sampling
- Visualize aggregations (Data binning, Transfer function)

Interactive Scalability: Big data makes interaction slow (Database queries, inefficient data processing, rendering performance); should be below 100ms

Possible Solutions:
- Sampling
- Progressive analytics: always produce partial results
- Precomputation (data cubes, data tiles)
- Caching & prefetching (predict what user will do next; e.g. last step, random step, most common step)
- Efficient rendering (webGL is cool)

detail: clustering, high-dim, low-dim, alternative?

-> high dimensional = wide data

Clustering: Divide data into subsets where objects in the same subset are similar to each other with respect to a given similarity measure;

High-dim: k-means
Low-dim: subspace clustering

Alternatives:
Feature selection: Select subset of existing features
Feature extraction: Transform existing features into low-dim space
Hybrid approach: Mix both

TSE

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

Difference between structured and unstructured data

Structured data: tables, defined names/attributes; easy to search and filter
Unstructured: do not have a predefined schema; music, text, images, videos

What is the Confusion Matrix

For evaluating classifier accuracy; Shows wrongly/corrected classified classes; should all be on diagonal

Dimensionality Reduction, why is it useful, name two different methods

Wide data hard to visualise;

Linear Methods:
- PCA: linear projection with variance maximised and square distance minimised between original and projected data points; operates on normalised data; are the eigenvectors of covariance matrix; then -> similarity preserving scatterplot; Biplot;
- Custom linear projections (NLP: text bias dimension, Austria vs Germany)
- Sequence of linear subspaces (Grand Tour)

Non-linear methods:
- t-SNE
- UMAP (Uniform Manifold Approximation and Projection)

Kde: what's the advantage/disadvantage vs. histograms and a general explaination

Advantage: non-parametric; Histogram etc depend highly on parameters

Weighting distance of observations from point x

Violin plot, Contour plot (Wanderkarte),

3 interpretability methods for AI + categories

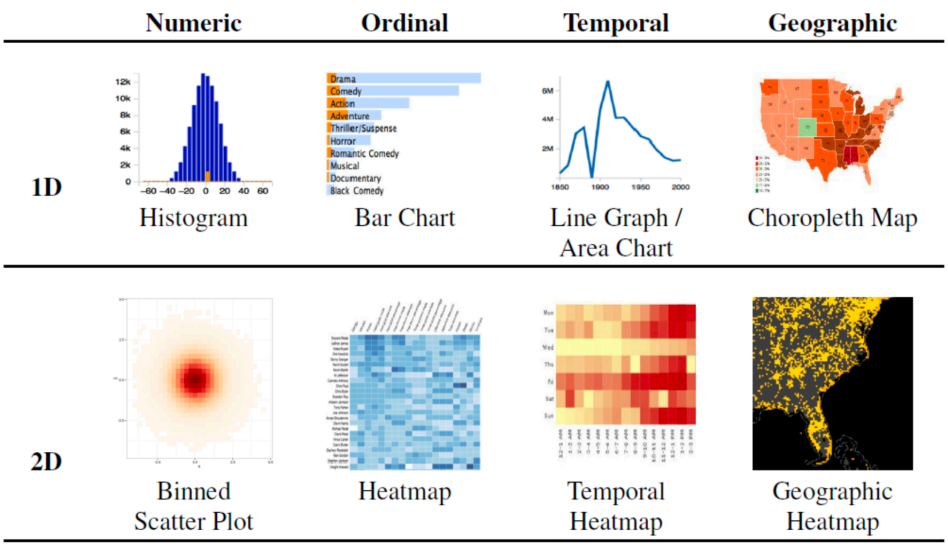
Local (one prediction) vs Global (whole model) & Model-specific (white box) vs Model-agnostic (blackbox)

Methods:
- Weight plot (Bike rental feature influence)
- Effect plot (weights times features)
- Decision Tree
- Local Interpretable Model-Agnostic Explanations (LIME): interpretable local surrogate model for blackbox model; select one sample, create modified copies of it, predict with blackbox model, train surrogate model on blackbox predictions;
- Partial dependence: Set all instances of one feature to same value and see how result changes;
- Individual Conditional Expectation (ICE); partial dependence works on average, ICE shows each sample
- Counterfactual examples: minimal set of changes to change the end result

CNN Methods:
- Features visualisation
- Attributions
- Playground / interactive systems

Binning: what is it? how can it be used with scatter-plots?

Put data into bins and Aggregate visualisations in bins (histogram);



What is Aggregate Visualization? When to use, what are advantages? Which 2 steps (Binning, visual mapping)? Examples for 1D and 2D.

Visualize aggregates instead of individual data values; used when visual clutter on overplotting

Data binning, Transfer function (densities -> visual variables)

Star coordinates. What is it? How to get them? How can you make it interactive?

Curvilinear coordinate system; items represented as points;
Sum of all unit vectors on each coordinate multiplied by value of data element for that coordinate

TODO

aggregate visualization (binned scatterplots and heatmap)

Both 2d;
Binned scatterplot: Group datapoint into bins and make scatterplot; the more points in the bin, the stronger the color

Heat map: same

Family of Curves + Set visualization

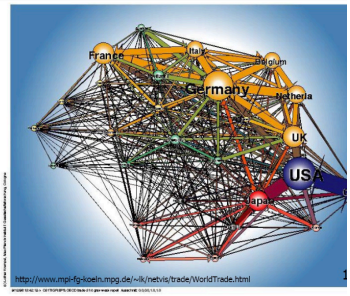
Set: venn diagram;
- set-o-gram: check cardinality of sets; e.g. which set occurs with which other sets (blue boxes hover video)
- Radial sets: same approach, but in radial arrangement; center can be used for other information

What criterion to optimize graph visualization

Aesthetics;

Aesthetics Criteria

- Edge crossings ↓
- Area ↓
- Symmetry ↑
- Edge length ↓
  - Maximal edge length, uniform edge length, total edge length
- Bends of edges ↓
  - Maximal bends, uniform bends, total bends
- Resolution ↑



model interpretablility (accuarcy vs. interpreatablility)

There is a trade-off; more accurate models have lower interpretability (e.g. NN)

Feature Selection: 2 statistical and 2 visual methods

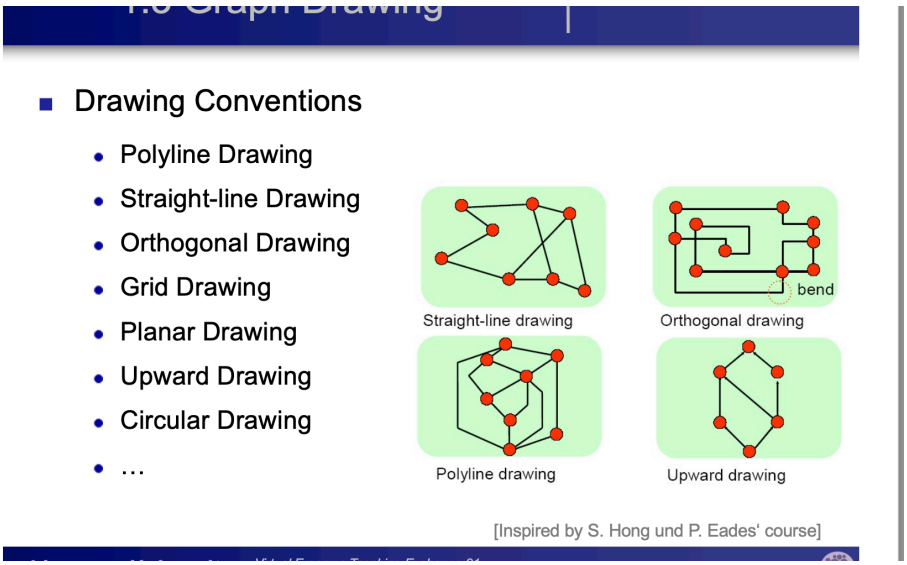
Statistical:
Entropy of distribution, number of potential outliers, number of unique values
Correlation coefficient, number of items in region of interest

Scagnostics

Visual:
Class Consistency: find features that separate class well
Selection and ordering of parallel coordinate axes: every pair is converted to Hough space; good dimension pairs have fewer, well defined clusters in Hough space

overview: graph drawing and boundary to information visualization

Graph drawing: automated drawing of 2d and 3d graphs
Many different ways to draw graphs



Non-linear projections vs linear projections difference

Feature selection (visual, statistical; scagnostics metrics) vs feature extraction
Star coordinates; can points overlap; what to do when they overlap (rotate,scale,remove axis)