Reduce Items and Attributes

Han-Wei Shen

Five Major Strategies for Big Data

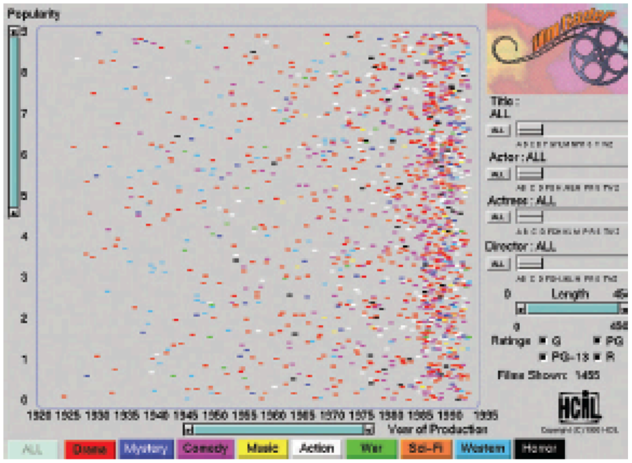
- Derive new (and more compact) data (Tamara Chapter 3)
- Change a view over time (Chapter 11)
- Facet data into multiple views (Chapter 12)
- Reduce items and attributes (Chapter 13)
- Focus+Context viewing (Chapter 14)

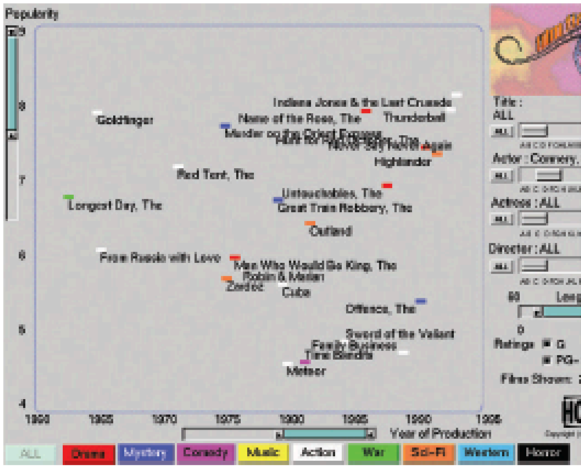
Filtering

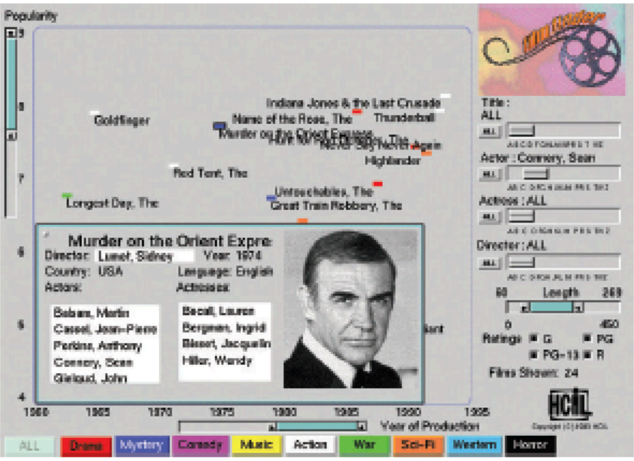
- Eliminate or select some items and/or attributes to make visual exploration more effective
- Challenges:
  - Without information losses
  - Support effective dynamic queries – tightly coupled with visual encoding and interaction
  - Do it efficiently

Filter Items

- Example: FilmFinder
  - Use sliders to control what items (films) to show in a scatter plot
  - The marks automatically adapt to the number of items shown (enlarged and labeled)
  - Detail information pops out with mouse over







Murder on the Orient Expre

Director: Lumet, Sidney Year: 1974

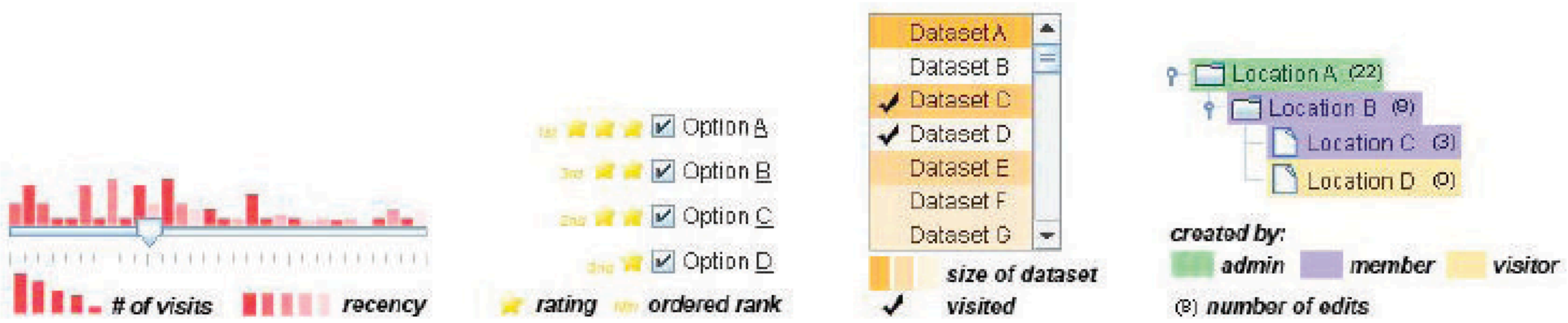
Country: USA Language: English

Actors:
Balsam, Martin
Cassel, Jean-Pierre
Perkins, Anthony
Connery, Sean
Gielgud, John

Actresses:
Bacall, Lauren
Bergman, Ingrid
Bisset, Jacqueline
Hiller, Wendy

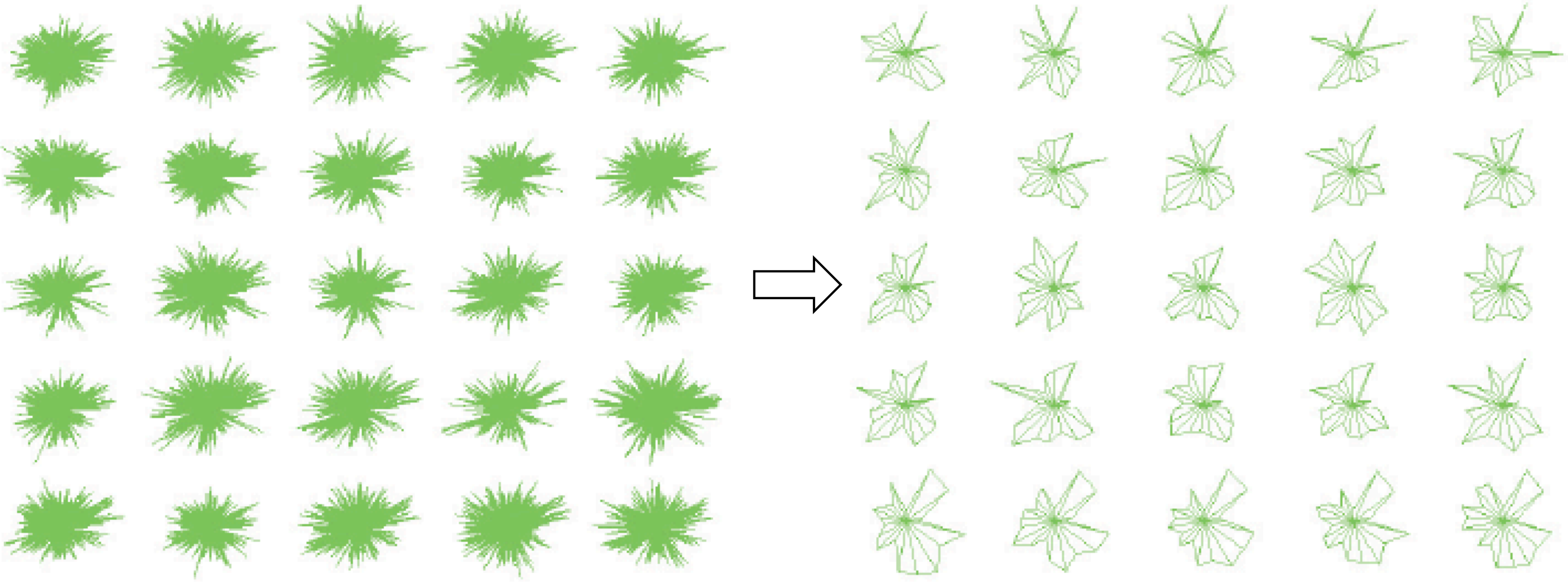
Scented Widget

- Augment the selection widget with concise visual information about the data



Attribute Filtering

- To show the same number of items, but fewer attributes
- Can be combined with item filtering
- Can also benefit from attribute ordering (or clustering) based on their similarity, and then only show the unique ones

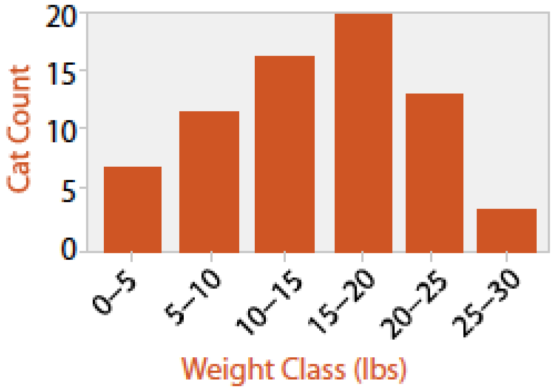


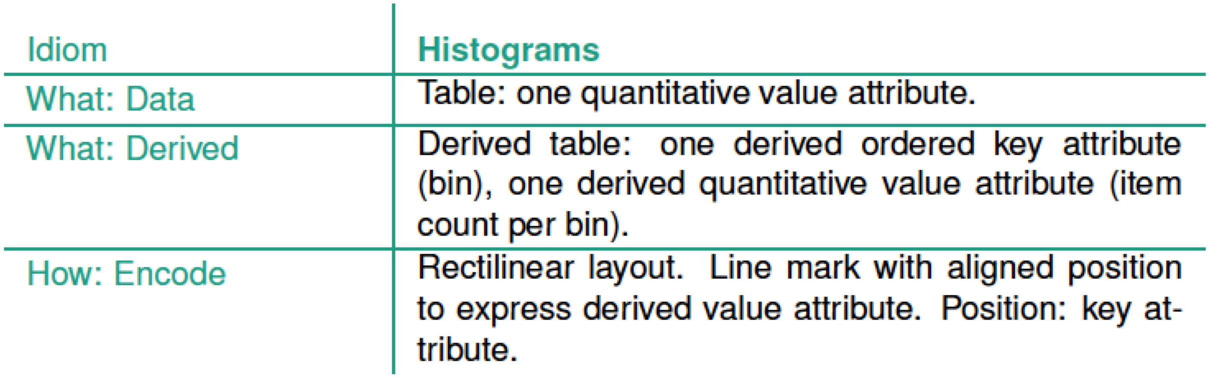
Aggregation

- A group of elements is represented by a new derived elements, e.g., average values
  - Elements are merged with aggregation, as opposed to be filtered/eliminated
- Basic aggregation: average, minimum, maximum, count, and sum
- Challenge: avoid eliminating interesting information

Example: Histogram

- Bin the data into different ranges, or different categorical types, and then count the number of items in each bin

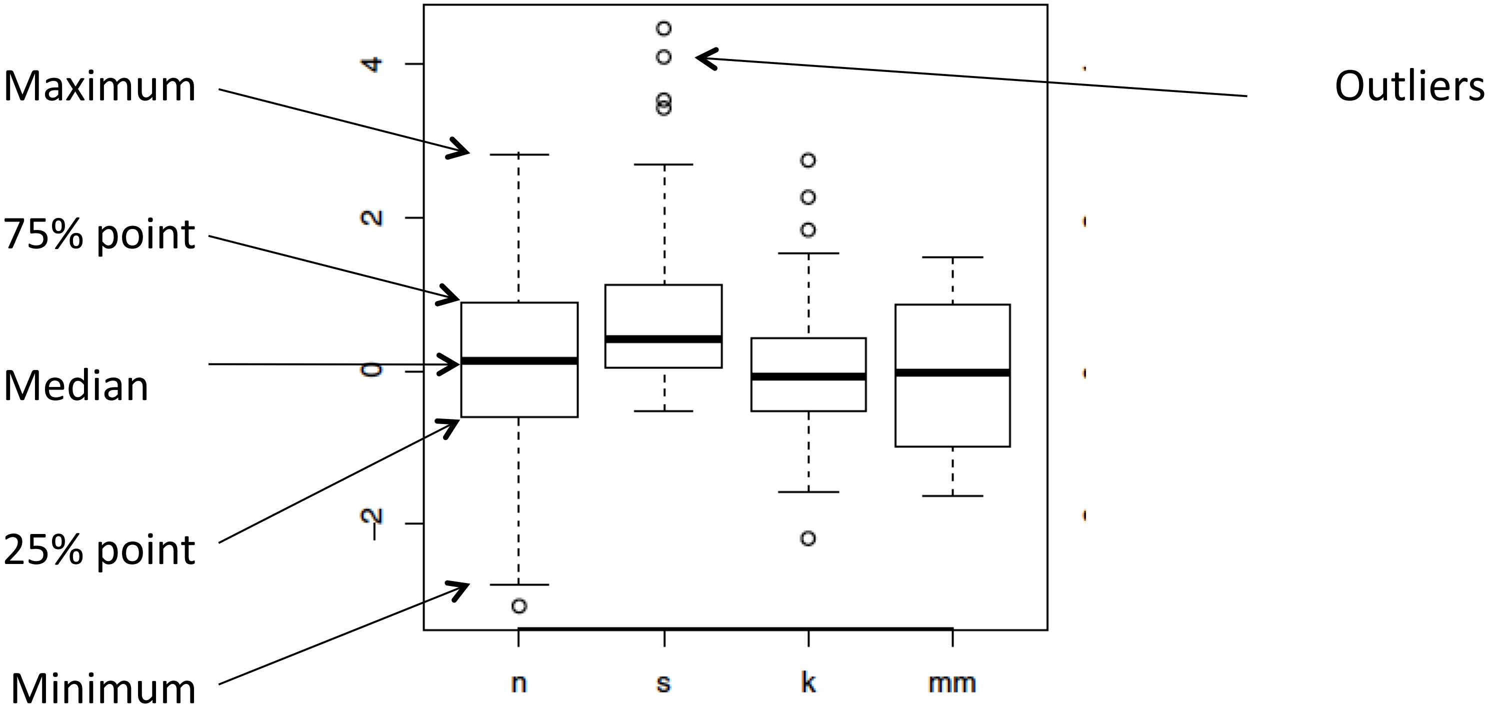




Idiom	**Histograms**
What: Data	Table: one quantitative value attribute.
What: Derived	Derived table: one derived ordered key attribute (bin), one derived quantitative value attribute (item count per bin).
How: Encode	Rectilinear layout. Line mark with aligned position to express derived value attribute. Position: key attribute.

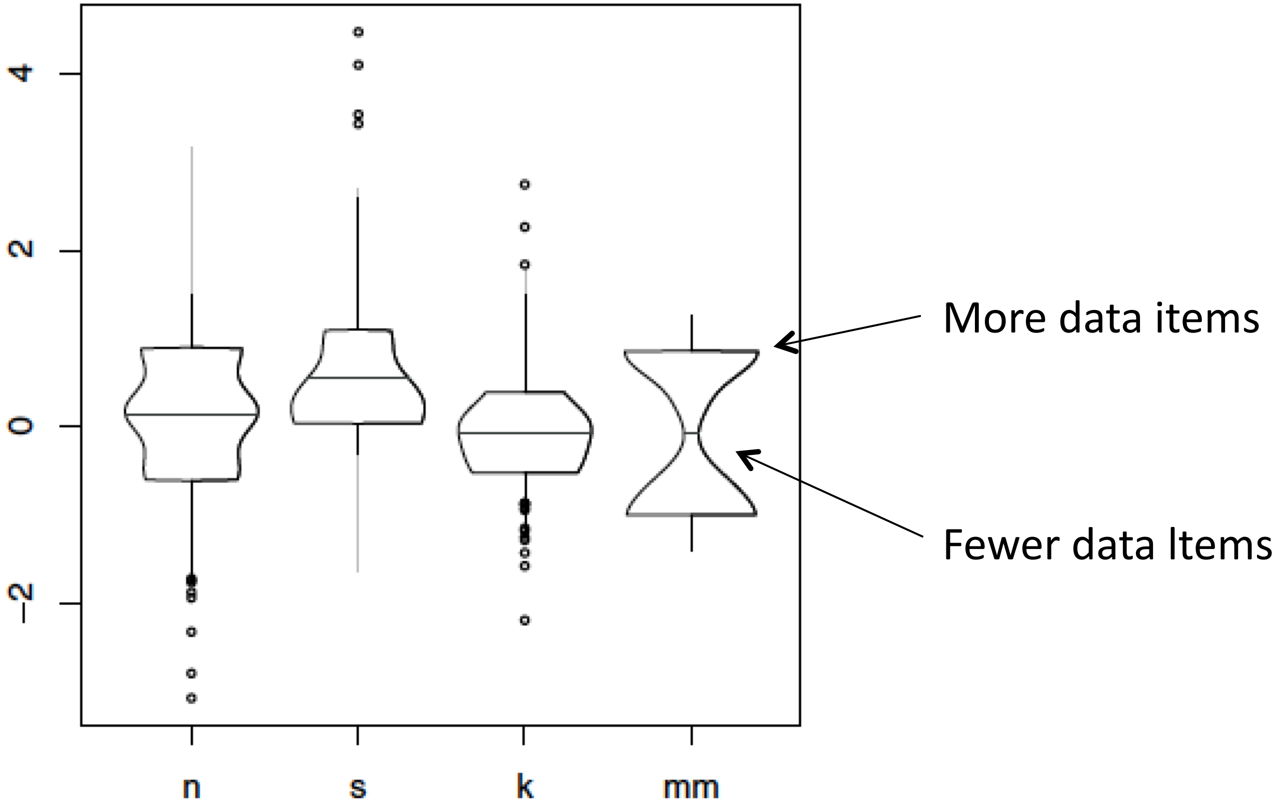
Example: Boxplots

- Compute five basic quantities: median (50% point), first quartile (25% point), third quartile (75% point), and two extremes (minimum and maximum)



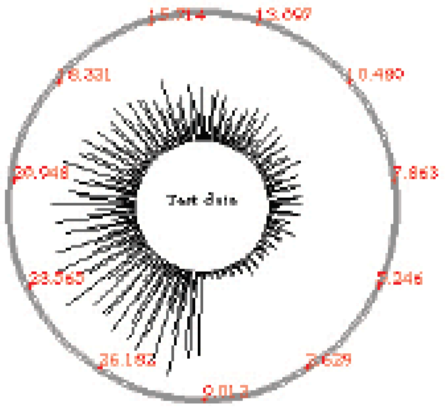
Example: Vaseplots

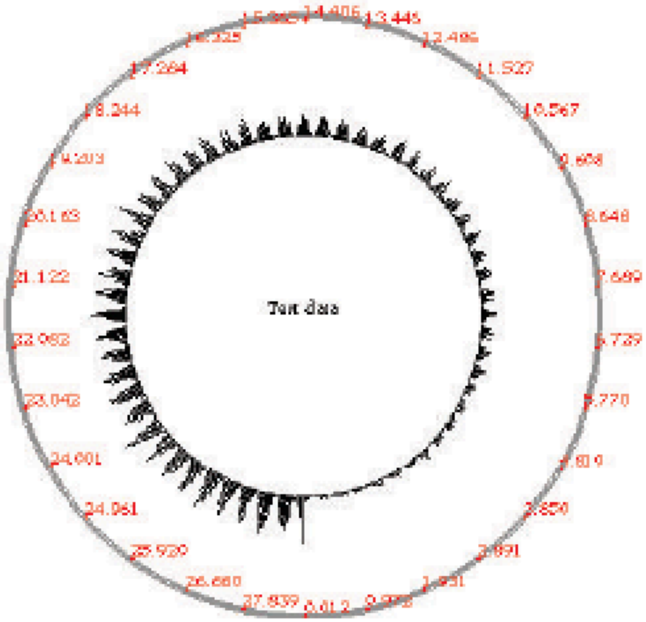
- Augment boxplots with width to depict data density



Example: Solarplots

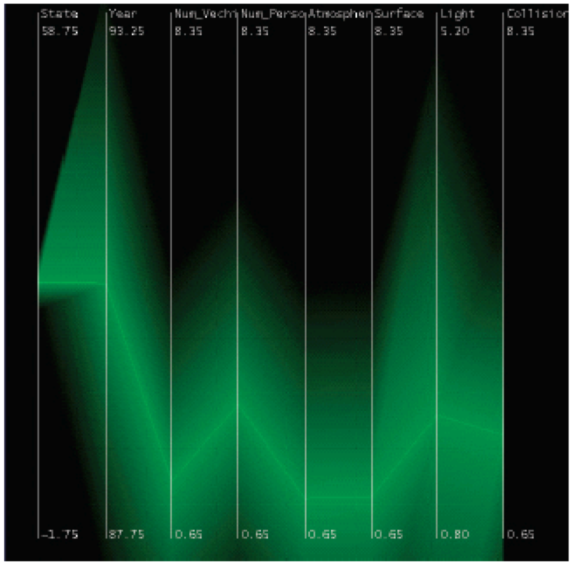
- Different rings indicate histograms of different aggregation levels, inner most being the coarsest

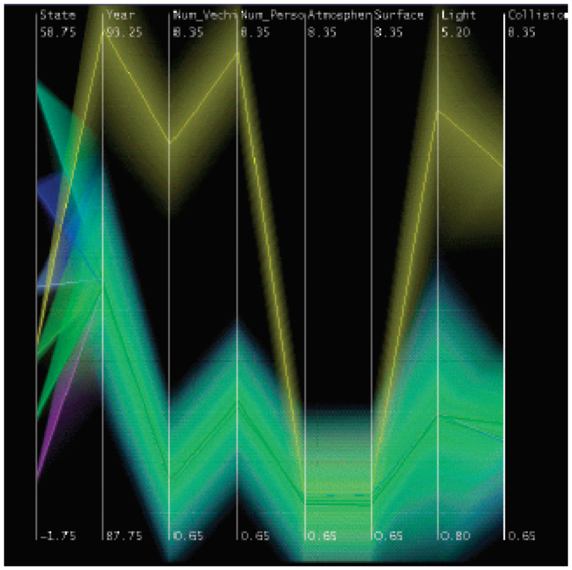


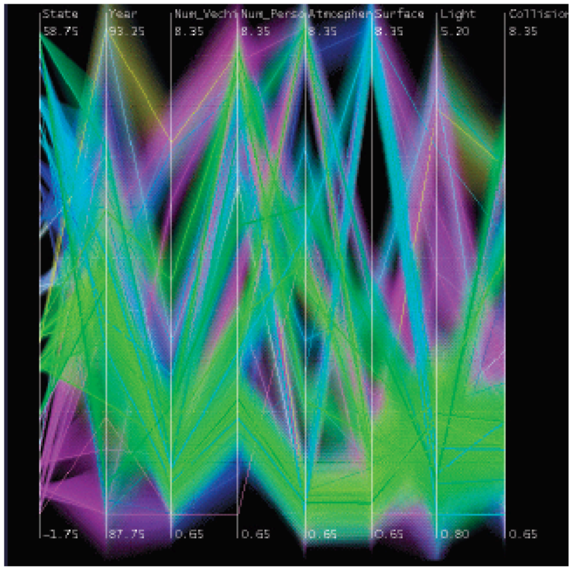


Example: Hierarchical Parallel Coordinates Plots

- Cluster the data items into different number of groups, and show the groups (mean, min, max) instead of the raw data items in PCP
- Inspect the clusters/data at different levels







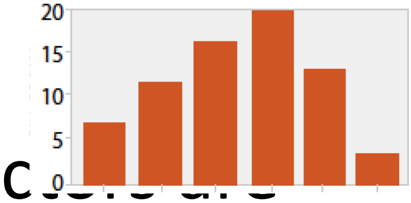
Attribute Aggregation: Dimensionality Reduction

- Use a fewer derived attributes to represent the original data attributes
  - Dimensions: number of attributes
- Goal:
  - preserve the meaningful structure in the data even with the new dimensions
  - This often means preserve the distance between the raw data points
- Common techniques
  - Multidimensional Scaling (MDS)
  - Principle Component Analysis (PCA)

Example: Document Collection

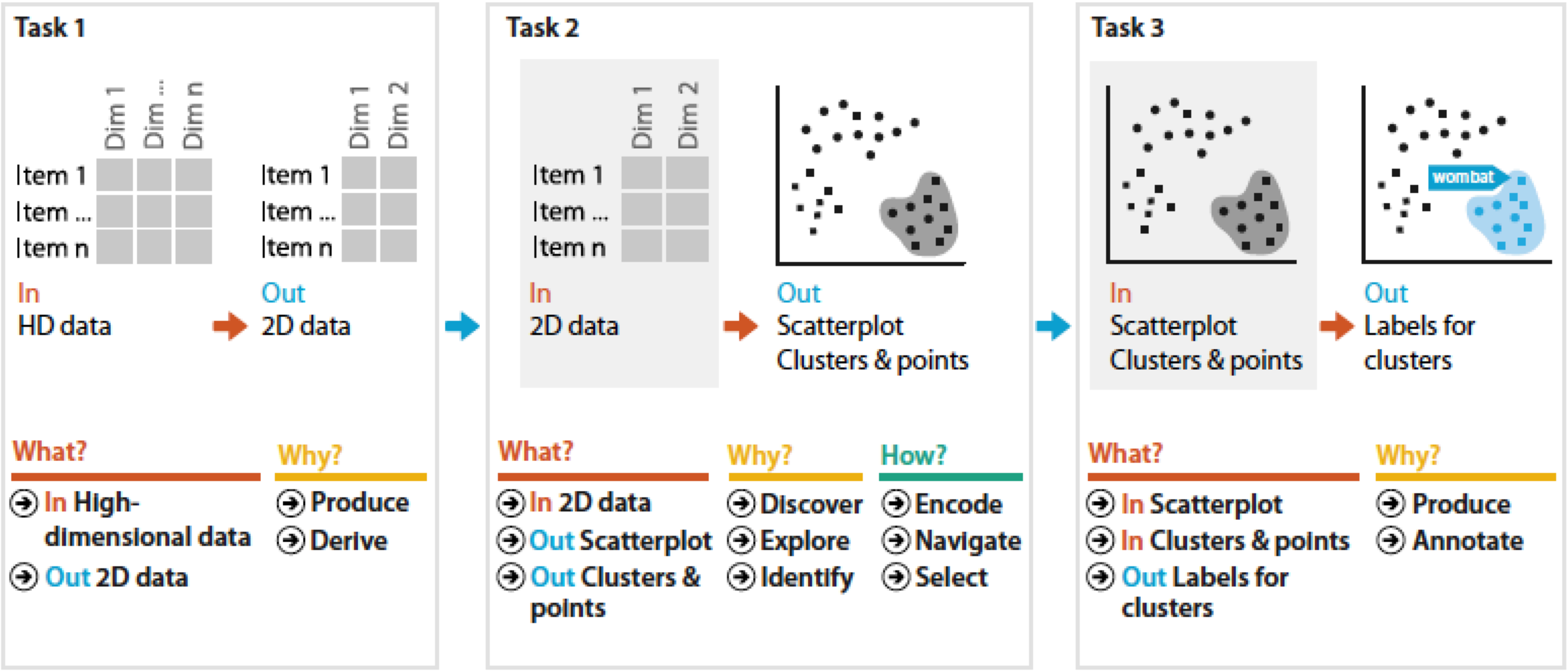
- Transform a document into a bag of words, and count the frequency of each words
  - (vis, tool, filter, aggregate, channel, …)

    (75, 10, 25, 34, 50, …..)
  - This is called a feature vector
- The dimensionality of the feature vectors are typically very high, so need to be reduced



- Apply dimensionality reduction so that each document can be represented as a visualizable mark (a point for example)

Example: Document Collection



Task 1

In
HD data

Out
2D data

What?
- In High-dimensional data
- Out 2D data

Why?
- Produce
- Derive

Task 2

In
2D data

Out
Scatterplot
Clusters & points

What?
- In 2D data
- Out Scatterplot
- Out Clusters & points

Why?
- Discover
- Explore
- Identify

How?
- Encode
- Navigate
- Select

Task 3

In
Scatterplot
Clusters & points

Out
Labels for clusters

What?
- In Scatterplot
- In Clusters & points
- Out Labels for clusters

Why?
- Produce
- Annotate

Display Dimensionality Reduction Results

- Two dimensions (e.g. output from Multidimensional Scaling, MDS) can be displayed as a scatter plot
- More than two dimensions can use scatterplot matrix (SPLOM)
- Need to allow the user to inspect the original high dimensional data by selecting the low dimensional derived attributes