Reduction

Data Visualization

Amit Chourasia

Why reduce?

Manage visual complexity

1. Derive data
2. View change
3. Facet
4. **Reduce**
    1. Filter : reduce items/attributes (beware “out of sight, out of mind notion”)
    2. Aggregation: stand in summarization (precludes details, how and what should be aggregated)
5. Dimensionality reduction

Reducing Items and Attributes

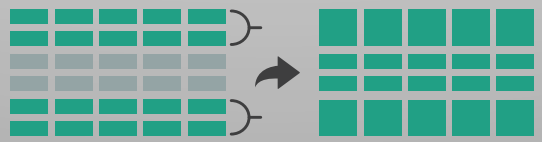
Filter

- Items

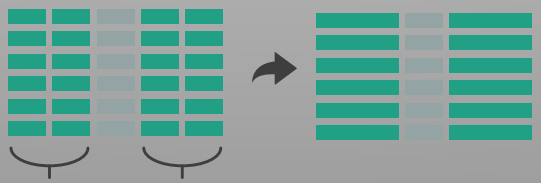
- Attributes

Aggregate

- Items



- Attributes



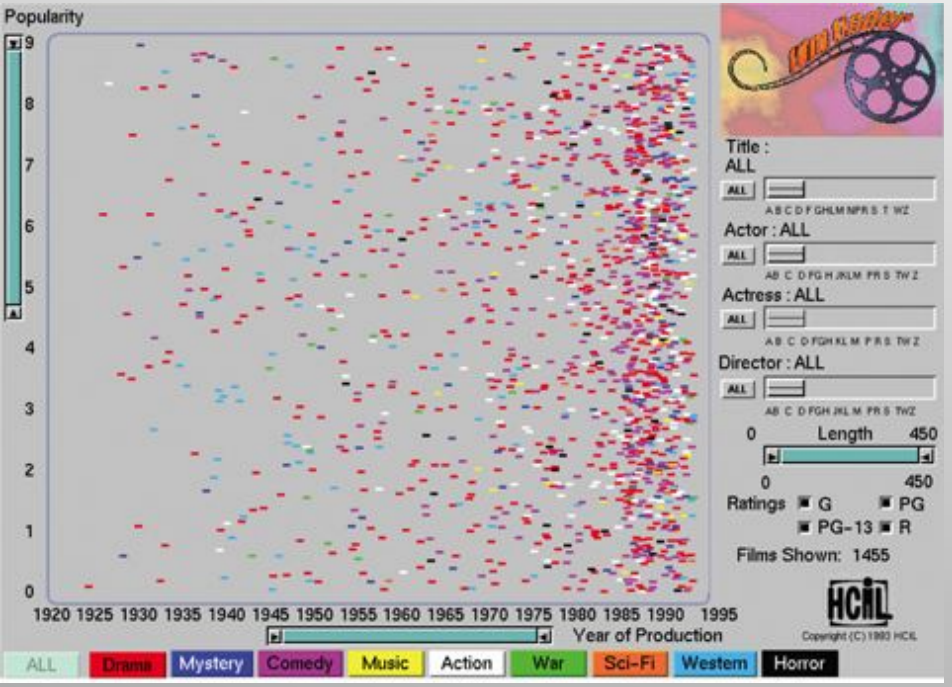
Figures by Eamonn Maguire

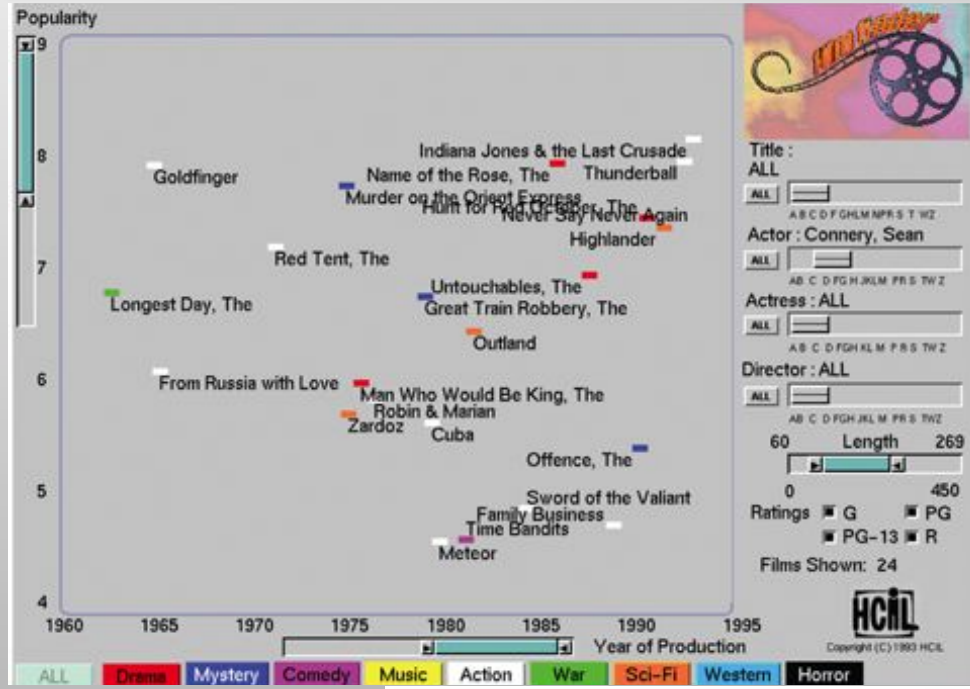
Reduce: Design Choice – Attribute Filtering

Eliminate number of elements

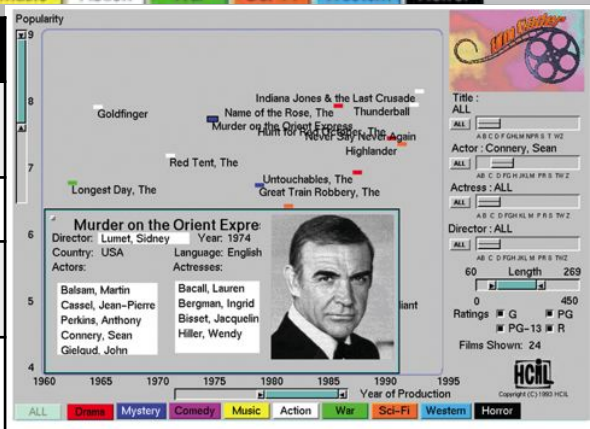
- Item filtering - Range selection

Filter author: Sean Connery
Filter title length: 60 - 269





System	FilmFinder
What: Data	Table: nine value attributes.
How: Encode	Scatterplot; detail view with text/images.
How: Facet	Multiform, overview–detail.
How: Reduce	Item filtering

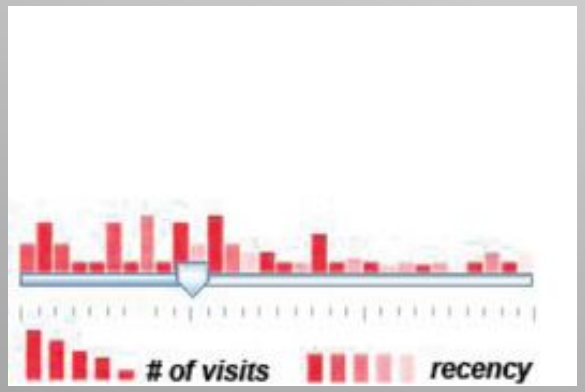


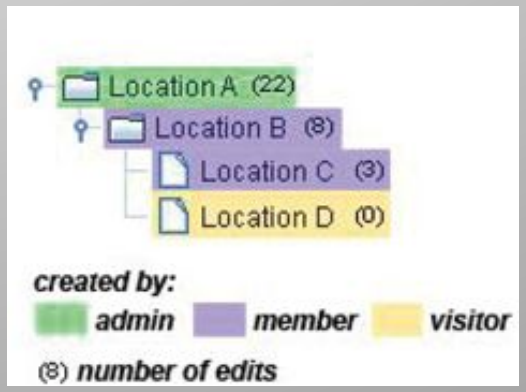
Figures by Ahlberg and Shneiderman. “FilmFinder”

Scented widgets

Augment standard widgets with additional cues to help user decide whether there is value in further drill down

- Add concise graph
- Add icons/text
- Encode more info in the widget



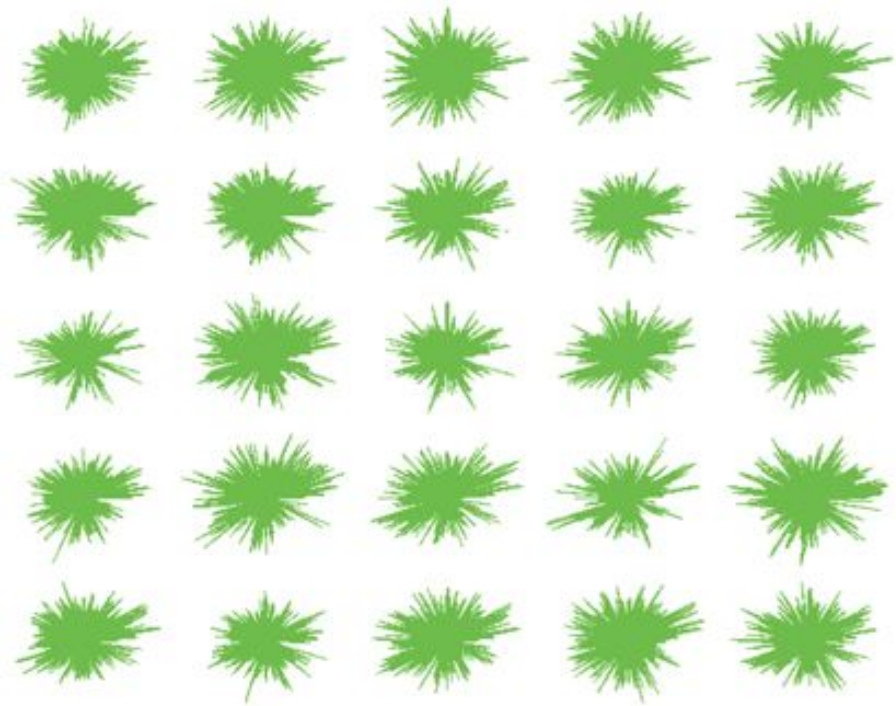


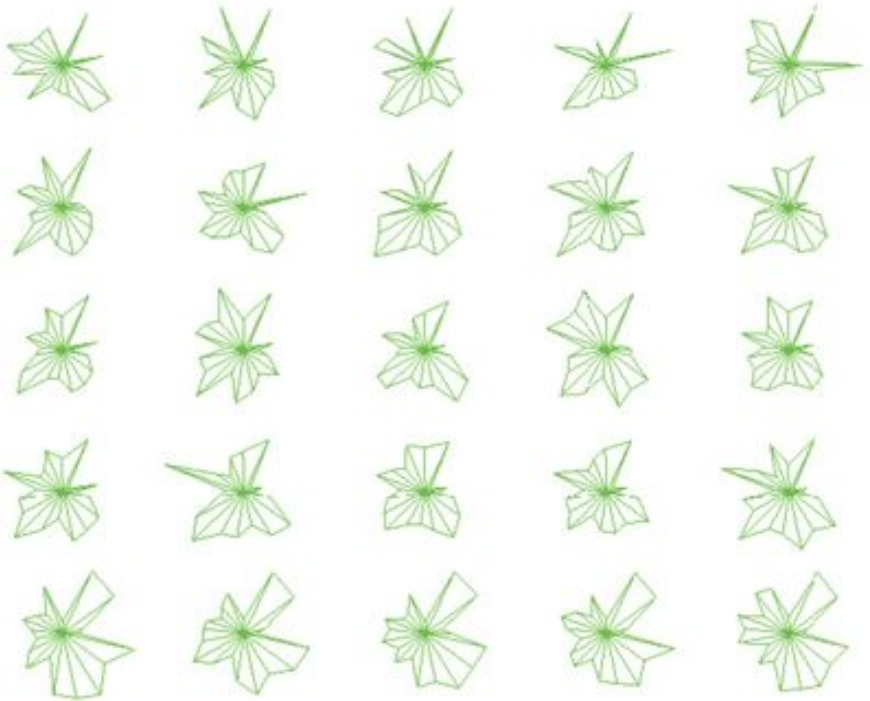
Figures by Willett et al. “Scented Widgets: Improving Navigation Cues with Embedded Visualizations””

Reduce: Design Choice – Item Filtering

Eliminate number of attributes as opposed to items.

Often used with attribute ordering and similarity measures.





System	Dimensional Ordering, Spacing, and Filtering Approach (DOFSA)
What: Data	Table: many value attributes.
How: Encode	Star plots.
How: Facet	Small multiples with matrix alignment.
How: Reduce	Attribute filtering.

Figures by Yang et al. “InterRing: An Interactive Tool for Visually Navigating and Manipulating Hierarchical Structures.”

Reduce: Design Choice – Aggregate

Merge elements to form a new derived element.

Common aggregation methods. However they are rarely sufficient e.g. Anscombe's quartet

- Average
- Minimum
- Maximum
- Count
- Sum

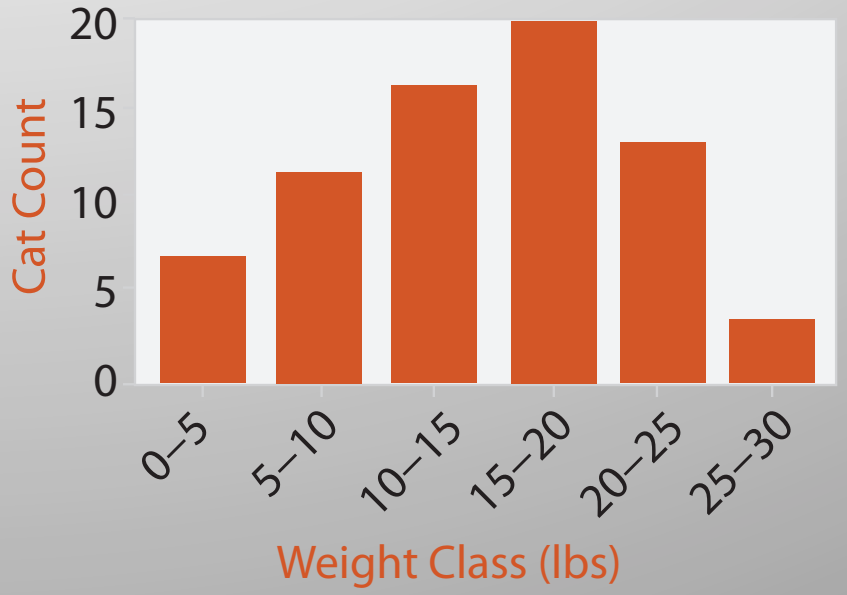
Figures by Yang et al. "InterRing: An Interactive Tool for Visually Navigating and Manipulating Hierarchical Structures."

Reduce: Design Choice – Item Aggregation (Histogram)

Histogram shows distribution of items within an original attribute.

Choice of bin size is critical and tricky.

- Use data characteristics
- Extend the choice to user



Idiom	Histogram
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.

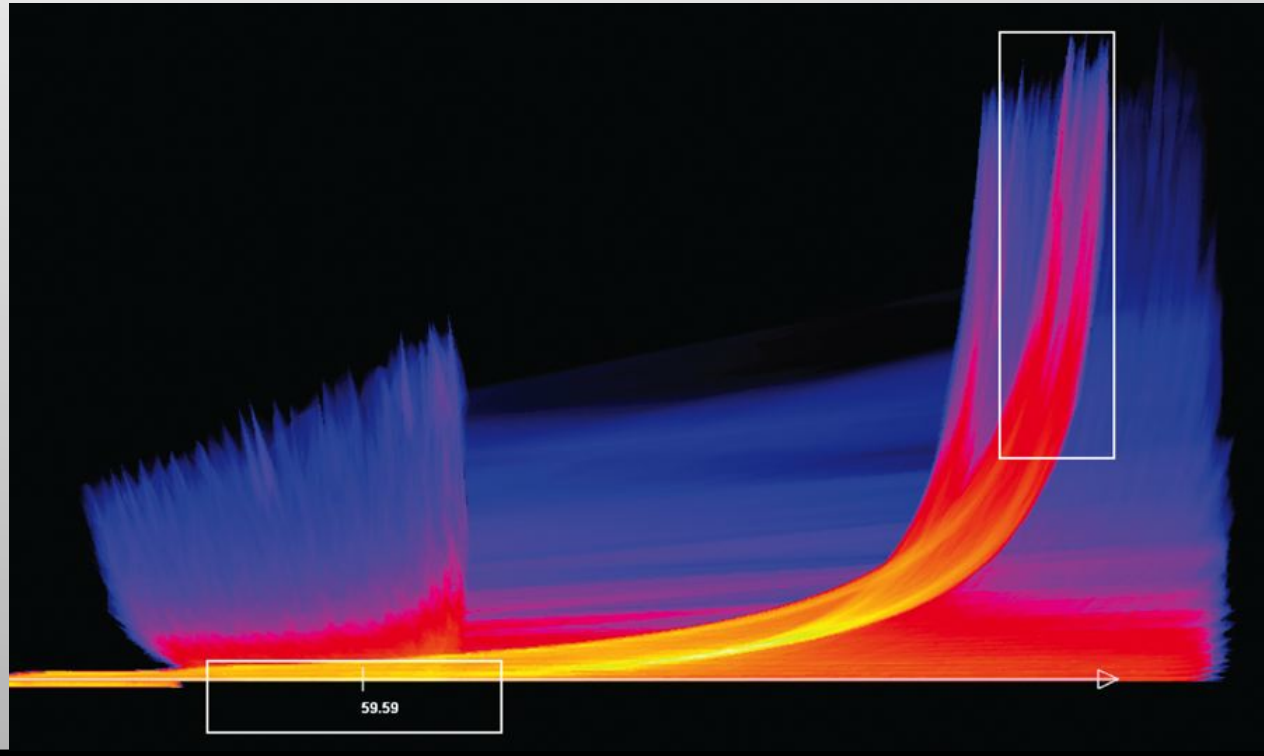
Figures by Eamonn Maguire

Reduce: Design Choice – Item Aggregation (Continuous Scatterplots)

Continuous scatterplots use a dense, space-filling 2D matrix alignment, where each pixel is given a different color.

Scatterplot problems

- Occlusion
- Size coding, labels worsens



Idiom	Continuous scatterplot
What: Data	Table: two quantitative value attributes.
What: Derived	Derived table: two ordered key attributes (x, y pixel locations), one quantitative attribute (overplot density).
How: Encode	Dense space-filling 2D matrix alignment, sequential categorical hue + ordered luminance colormap.
How: Reduce	Item aggregation

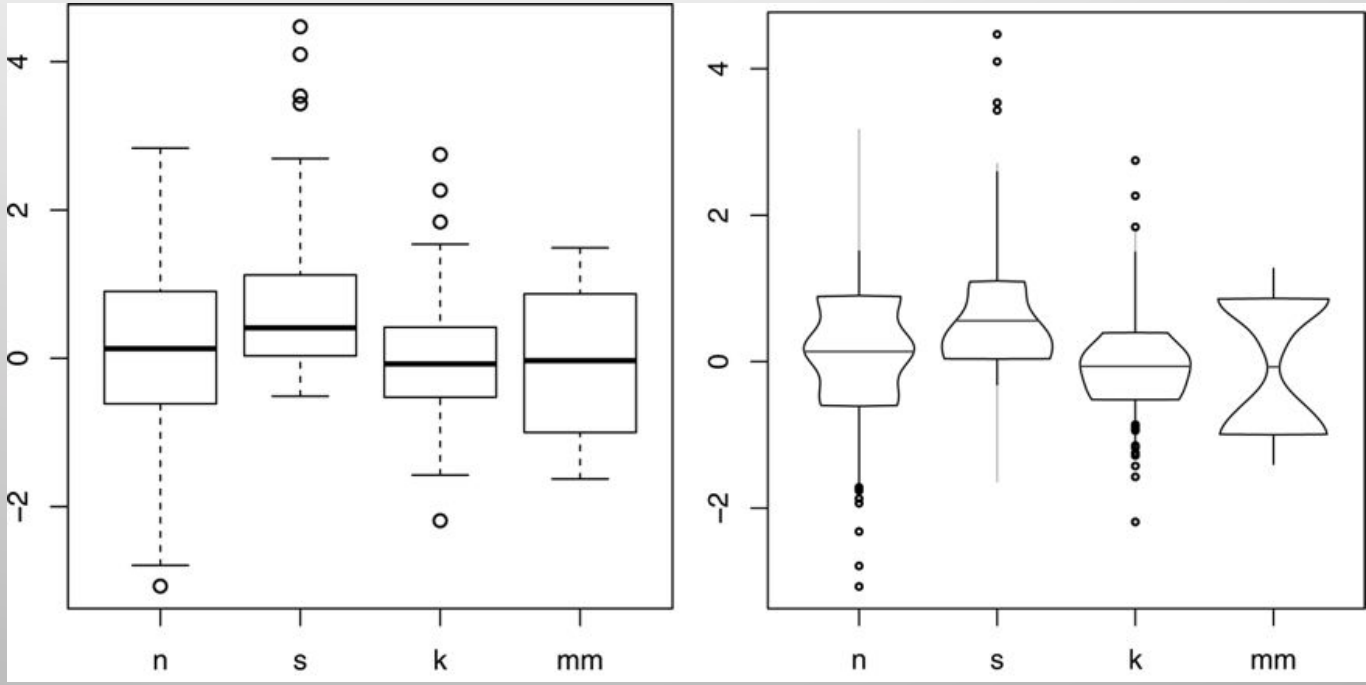
Figures by Bachthaler and Weiskopf. "Continuous Scatterplots."

Reduce: Design Choice – Item Aggregation (Boxplot)

Boxplot shows aggregate statistical summary for one quantitative attribute.

Five items

- Median (50%)
- Quartiles (25% & 75%)
- Upper, Lower fences
- Outliers*



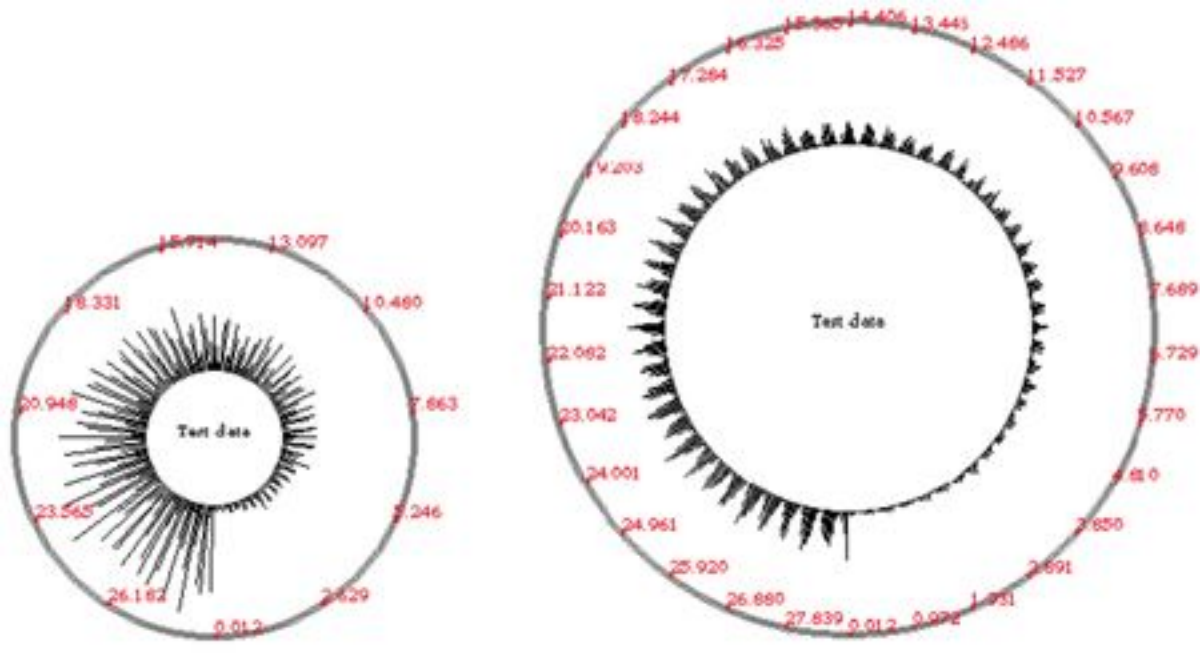
Idiom	Boxplot and vaseplot
What: Data	Table: many quantitative value attributes.
What: Derived	Five quantitative attributes for each original attribute, representing its distribution.
Why: Tasks	Characterize distribution; find outliers, extremes, averages; identify skew.
How: Encode	One glyph per original attribute expressing derived attribute values using vertical spatial position, with 1D list alignment of glyphs into separated with horizontal spatial position.
How: Reduce	Item aggregation.
Scale	Items: unlimited. Attributes: dozens.

Figures by Wickham and Stryjewski. “40 Years of Boxplots.”

Reduce: Design Choice – Item Aggregation (Solarplot)

Solarplot shows a radial histogram.

Base circle radius controls number of bins



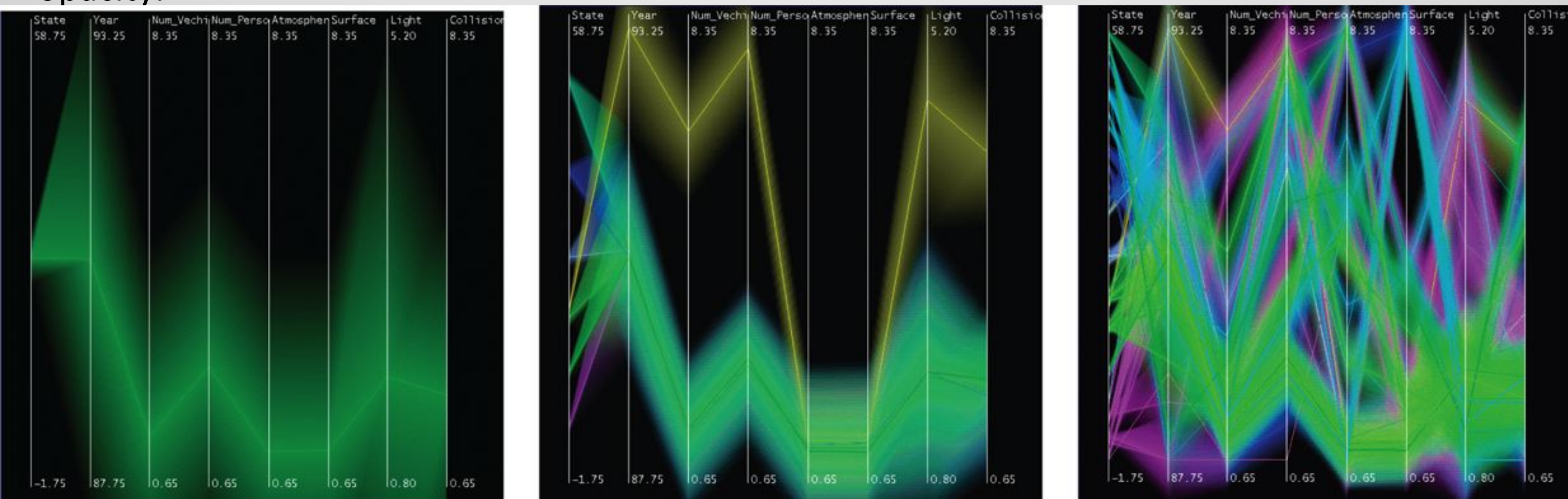
Idiom	Solarplot
What: Data	Table: one quantitative attribute.
What: Derived	Derived table: one derived ordered key attribute (bin), one derived quantitative value attribute (item count per bin). Number of bins interactively controlled.
Why: Tasks	Characterize distribution; find outliers, extremes, averages; identify skew.
How: Encode	Radial layout, line marks. Line length: express derived value attribute; angle: key attribute.
How: Reduce	Item aggregation.
Scale	Original items: unlimited. Derived bins: proportional to screen space allocated.

Figures by Wickham and Stryjewski. "40 Years of Boxplots."

Reduce: Design Choice – Item Aggregation (Hierarchical Parallel Coordinates)

Hierarchical parallel coordinates uses clustering aggregation for scalability.

Cluster – Mean, minimum and maximum; depth and shown with band of varying width and opacity.



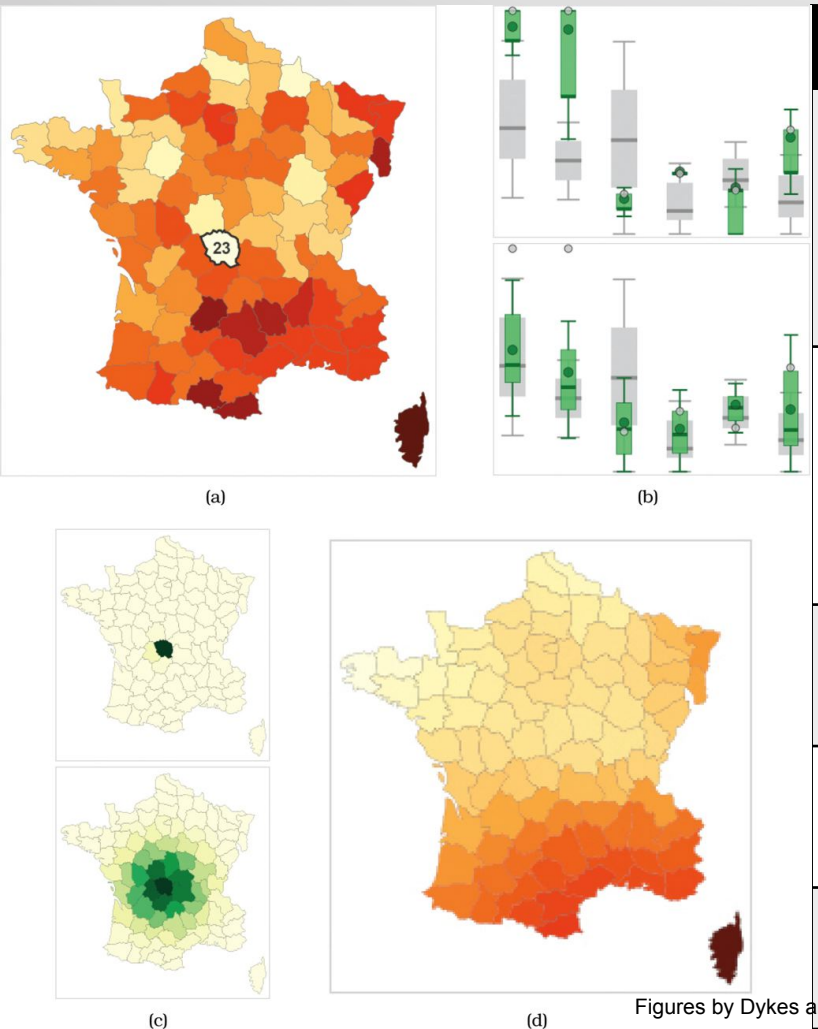
Idiom	Hierarchical Parallel Coordinates
What: Data	Table
What: Derived	Cluster hierarchy atop original table of items. Five per-cluster attributes: count, mean, min, max, depth.
Why: Tasks	Characterize distribution; find outliers, extremes, averages; identify skew.
How: Encode	Parallel coordinates. Color clusters by proximity in hierarchy.
How: Reduce	Interactive item aggregation to change level of detail.
Scale	Items: 10,000–100,000. Clusters: one dozen.

Figures by Wickham and Stryjewski. "40 Years of Boxplots."

Reduce: Design Choice – Item Aggregation (Geowigs)

Geographically Weighted Boxplots

Sophisticated support for spatial aggregation using geographically weighted regression and geographically weighted summary statistics



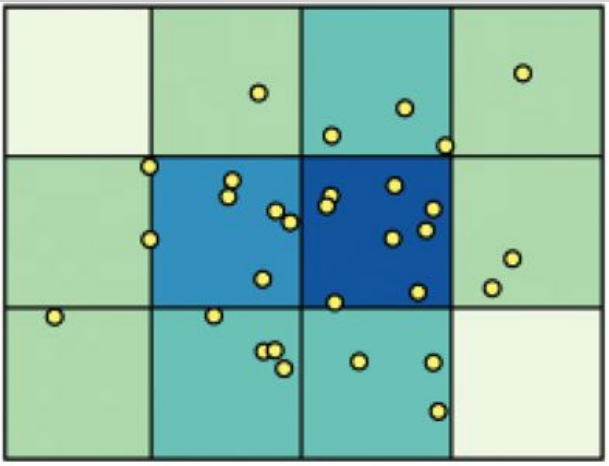
Idiom	Geographically Weighted Boxplots
What: Data	Geographic geometry with area boundaries. Table: Key attribute (area), several quantitative value attributes. Table: Five-number statistical summary distributions for each original attribute.
What: Derived	Multidimensional table: key attribute (area), key attribute (scale), quantitative value attributes (geographically weighted statistical summaries for each area at multiple scales).
How: Encode	Boxplot
How: Facet	Superimposed layers: global boxplot as gray background, current-scale boxplot as green foreground.
How: Reduce	Spatial aggregation.

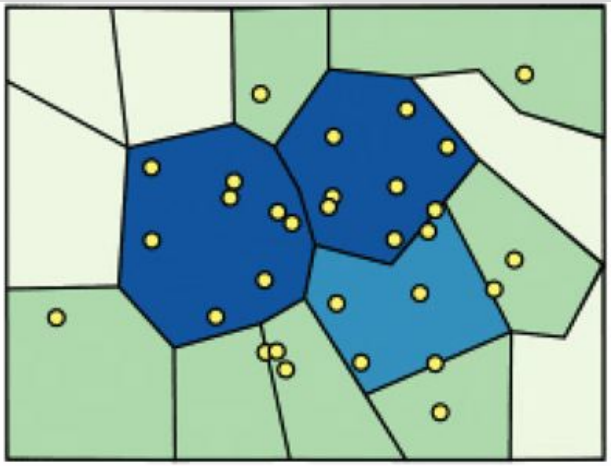
Figures by Dykes and Brunsdon. "Geographically Weighted Visualization: Interactive Graphics for Scale-Varying Exploratory Analysis."

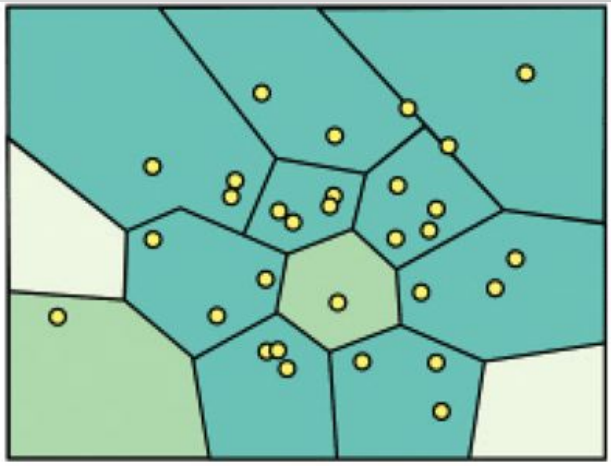
Reduce: Design Choice – Item Aggregation (Spatial)

Modifiable areal unit problem (MAUP)

Issue: Changing the boundaries of the regions used to analyze data can yield dramatically different results.







Figures by https://www.e-education.psu.edu/geog486/node/1864

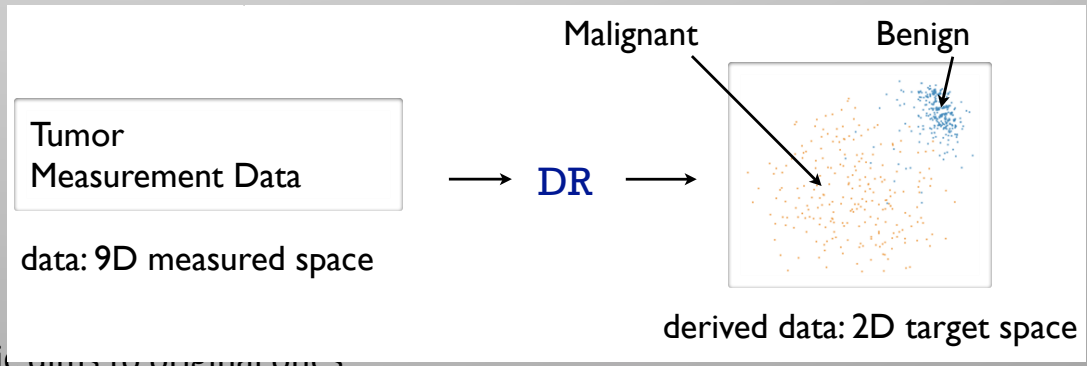
Reduce: Design Choice – Attribute Aggregation (Dimensionality reduction)

Dimensionality reduction (DR) is an aggregation method where the goal is to preserve the meaningful structure of a dataset while using fewer attributes to represent the items.

- Derive low-dimensional target from actual high dimensional data
- Assumptions/Use when
  - Hidden structure in data
  - Significant redundancy in data

Why DR?

- improve performance of downstream algorithm
  - Avoid curse of dimensionality
- data analysis



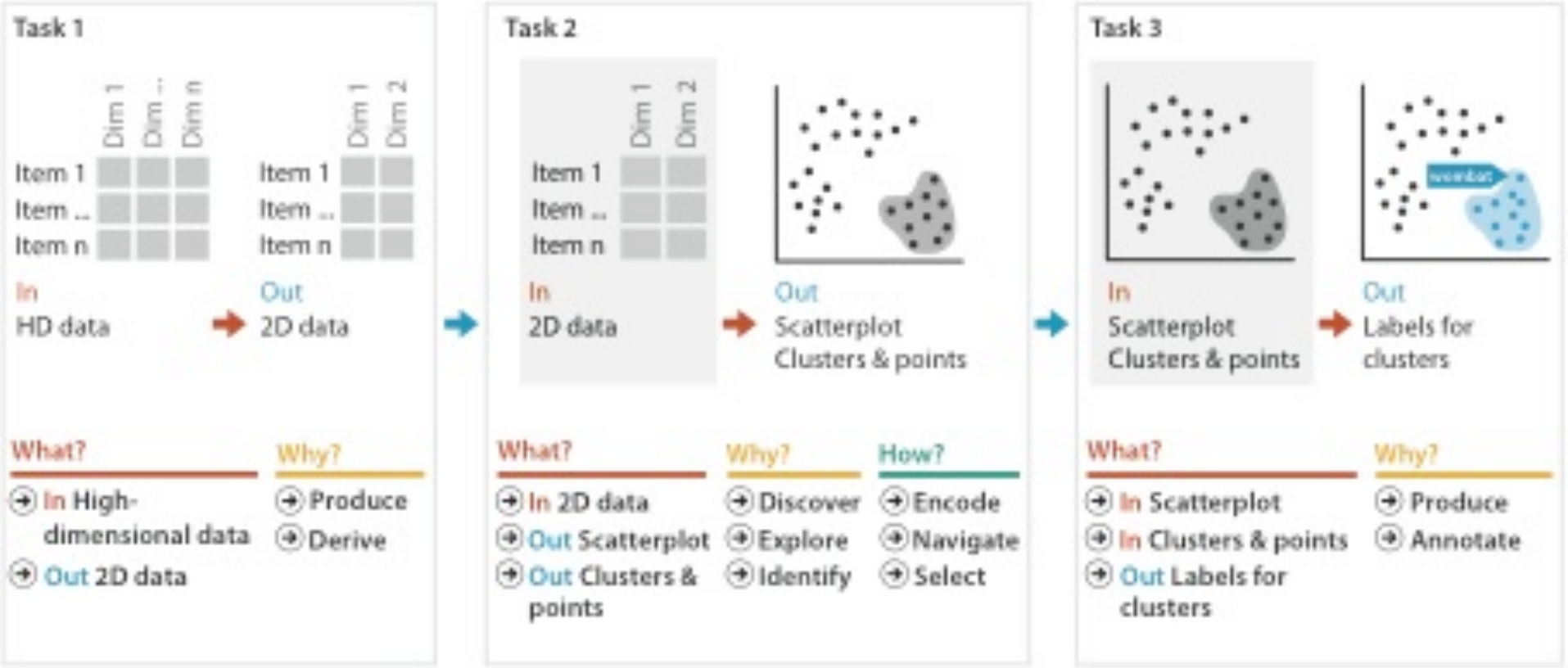
DR tasks

- Dimension-oriented task sequences
  - Name synthetic dimensions, map synthetic dims to original ones
- Cluster-oriented task sequences
  - Verify clusters, name clusters, match clusters and classes

To be continued in Guest Lecture …

Slide: Tamara Munzner,

Dimensionality Reduction for Document Collections



Idiom	Dimensionality Reduction for Document Collections
What: Data	Table with 10,000 attributes.
What: Derived	Table with two attributes.
Why: Tasks	Characterize distribution; find outliers, extremes, averages; identify skew.
How: Encode	Scatterplot, colored by conjectured clustering.
How: Reduce	Attribute aggregation (dimensionality reduction) with MDS
Scale	Original attributes: 10,000. Derived attributes: two. Items: 100,000.

Figures by Eamonn Maguire

Reduce: Design Choice – Attribute Aggregation (Dimensionality reduction)

Bag of Words (DR)

- Analyze the differential distribution of words between the documents
- Find clusters of related documents based on common word use between them

E.g. Ngram

Reduce: Design Choice – Attribute Aggregation (Dimensionality reduction)

Process

Use N attributes to synthesize 2 or more attributes

Estimating true dimensionality

- How do you know when you would benefit from DR?
  - Consider error for low-dim projection vs high-dim projection
- No single correct answer; many metrics proposed
  - Cumulative variance that is not accounted for
  - Strain: match variations in distance (vs actual distance values)
  - Stress: difference between interpoint distances in high and low dims

$$stress(D, \Delta) = \sqrt{\frac{\sum_{ij}(d_{ij} - \delta_{ij})^2}{\sum_{ij} \delta_{ij}^2}}$$

- D: matrix of lowD distances
- Δ: matrix of hiD distances δ_{ij}

Design Choices

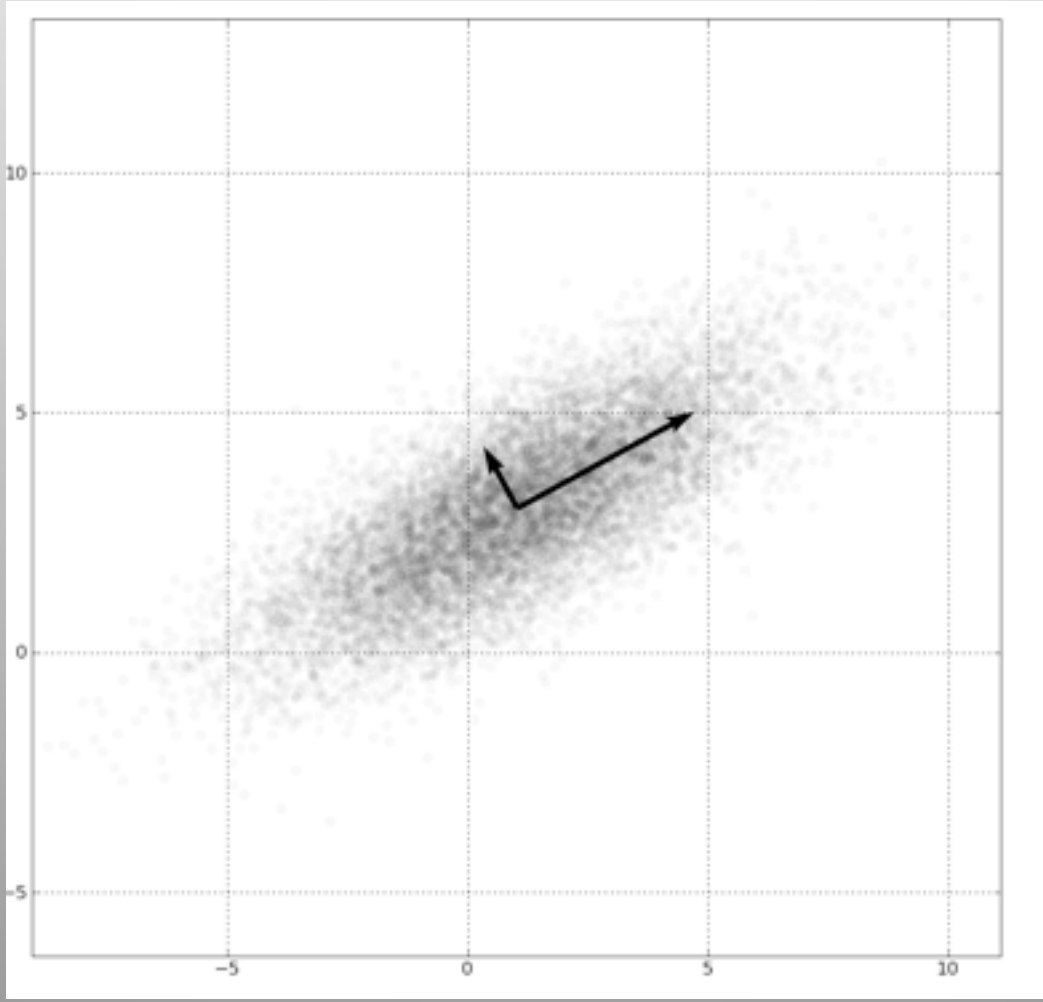
- Scatter plot (2 synthetic attributes)
- Scatter plot matrix (> 2 synthetic attributes)

Slide: Tamara Munzner,

Reduce: Design Choice – Attribute Aggregation (Linear dimensionality reduction: PCA)

Principal components analysis (PCA)

- Describe location of each point as linear combination of weights for each axis
- Finding axes: first with most variance, second with next most, ...



Slide: Tamara Munzner, Figures by hhttp://en.wikipedia.org/wiki/File:GaussianScatterPCA.png]

Reduce: Design Choice – Attribute Aggregation (Non linear dimensionality reduction)

Nonlinear dimensionality reduction

Many techniques proposed

- MDS, charting, isomap, LLE,T-SNE
- Many literatures: visualization, machine learning, optimization, psychology, ...

Pros

Can handle curved rather than linear structure

Cons

- Lose all ties to original dims/attribs
- New dimensions cannot be easily related to originals

Slide: Tamara Munzner,

Dimensionality reduction: Multidimensional Scaling (MDS)

MDS: family of methods, linear and nonlinear!

Classical scaling: minimize strain

- Early formulation equivalent to PCA (linear)
- Nystrom/spectral methods approximate eigenvectors: O(N)
  - Landmark MDS [de Silva 2004], PivotMDS [Brandes & Pich 2006]
- limitations: quality for very high dimensional sparse data

Distance scaling: minimize stress

- Nonlinear optimization: O(N2)
  - SMACOF [de Leeuw 1977]
- Force-directed placement: O(N2)
  - Stochastic Force [Chalmers 1996]
  - limitations: quality problems from local minima

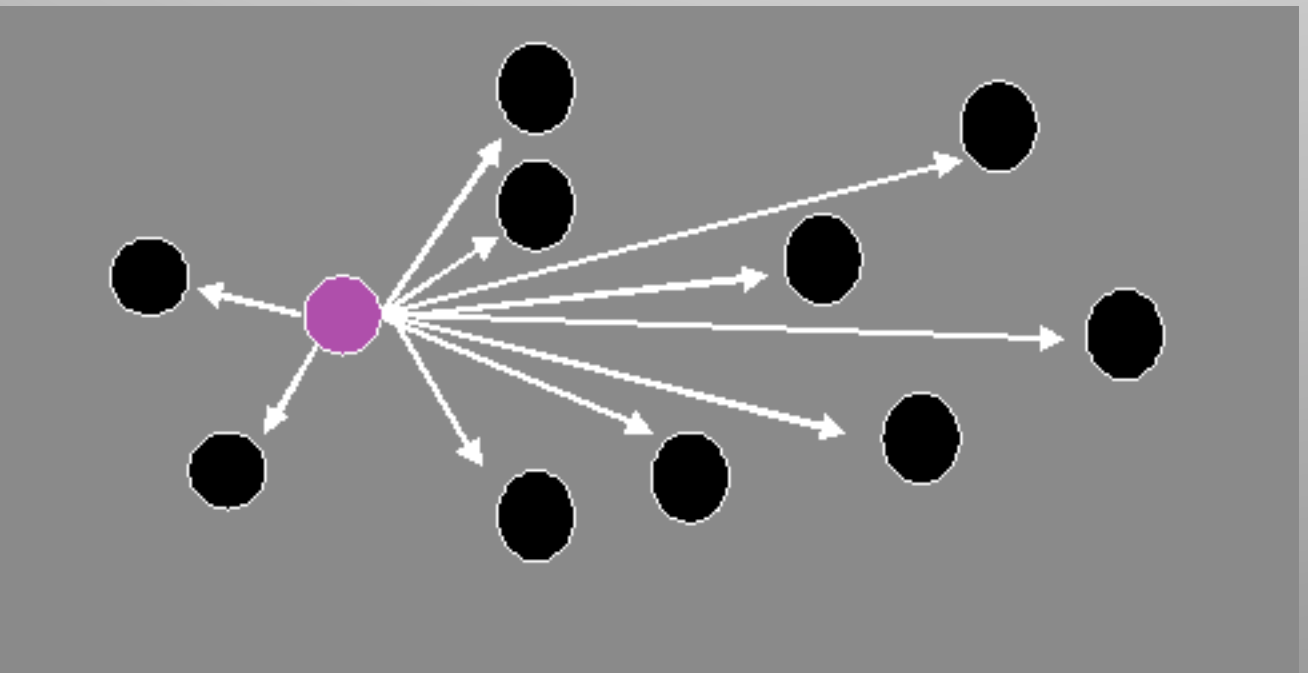
Glimmer goal: O(N) speed and high quality

Slide: Tamara Munzner,

Dimensionality reduction: Multidimensional Scaling (MDS)

Spring-based MDS: naive

- Repeat for all points
  – Compute spring force to all other points
  – Difference between high dim, low dim distance – move to better location using computed forces
- Compute distances between all points – O(N2) iteration, O(N3) algorithm

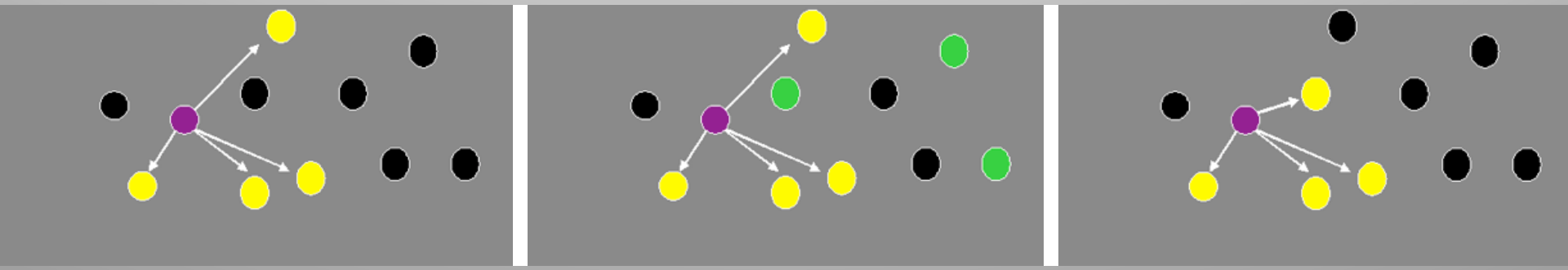


Slide: Tamara Munzner,

Dimensionality reduction: Multidimensional Scaling (MDS)

Faster spring model: Stochastic

- Compare distances only with a few points
  - maintain small local neighborhood set
  - each time pick some randoms, swap in if closer
- Small constant: 6 locals, 3 randoms (typically)
  - O(N) iteration, O(N2) algorithm



Slide: Tamara Munzner,

Readings

- Visual Analysis and Design – Chapter 13, 15