Reduce Items & Attributes

Visual and Exploratory Data Analysis 23S

06.06.2023

Timothée Schmude

Schedule

Lecture on Apr 25:
Facet into Multiple Views

Today:
Reduce Items & Attributes

Lecture on June 12:
Manipulate Views
Embed: Focus + Context

7	Apr 25/28	Facet into Multiple Views (AC) pdf [Munzner Ch. 12] Improvise	Q&A - D3 Tutorial
	Apr 30	Due: A3 (23:55)	
8	May 02/05	Tasks (FW) pdf [Munzner Ch. 2+3] Task Typology	Q&A Design Assignment (A4)
9	May 09/12	Design Studies (FW) pdf [Munzner Ch. 4] (MizBee) An example of a design study	VIS in Digital Humanities I (FW) pdf
	May 14	Due: A4 (23:55)	
10	May 16/19	Arrange Networks / Trees (CK) pdf [Munzner Ch. 9+10]	VIS in Digital Humanities II (FW) pdf [Collection Visualization Survey] Survey
11	May 23/26	Arrange Networks / Trees (CK) pdf [Munzner Ch. 9+10] Example: GraphDiaries by Bach et al.	No lecture
	May 30 / Jun 02	No Lecture	No Lecture
12	Jun 06/09	Reduce Items & Attributes (TS) pdf [Munzner Ch. 13] DimStiller	No Lecture
13	Jun 13/16	Manipulate View, Embed: Focus+Context (TM) pdf [Munzner Ch. 11+14]	TextVis for DH (AC, FW) pdf
	Jun 18	Due: A5 (23:55)	
14	Jun 20/23	TextVis (TM) pdf	Final
15	Jun 27/30	A5 Presentations	A5 Presentations

Handling visual complexity

Datasets often too large/complex to show everything at once in a single static view → overwhelming visual clutter

Five major approaches:

- Facet into multiple views — Lecture on Apr 25
- Reducing amount of data shown in a view
- Deriving new data and including it in the view — Today
- Change view over time
- Embed: Focus + Context — Lecture on Jun 12

Readings

- **Munzner.** Visualization Analysis and Design:
  - Chapter 11 (Manipulate View)
  - Chapter 13 (Reduce Items and Attributes)
- **Ward, Grinstein and Keim.** Interactive Data Visualization:
  - Chapter 2.3 (Data Preprocessing)

Data Reduction

Reduce amount shown

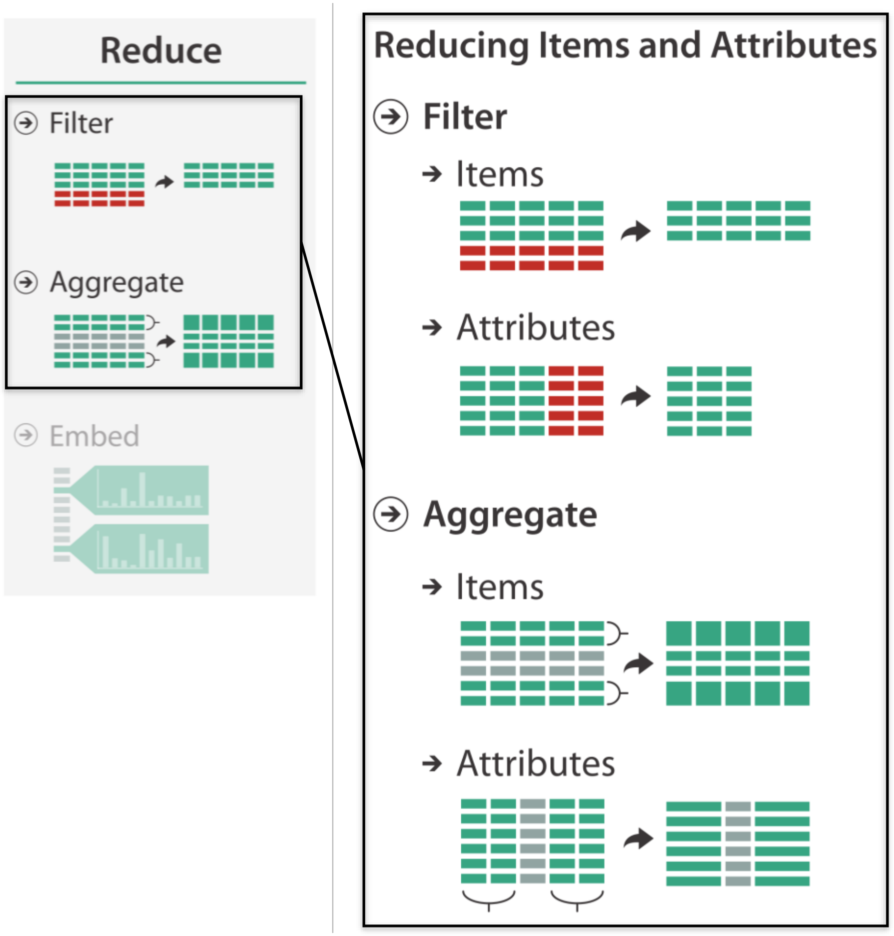
- Filter: Eliminate elements from view
- Aggregate: Create single new element that stands for multiple others

Attributes (columns / dimensions)

Items (rows)

country	happiness_rank	happiness_score	economy	family	health
Norway	1	7,53700	1,61646	1,53352	0,796667
Denmark	2	7,52200	1,48238	1,55112	0,792566
Iceland	3	7,50400	1,48063	1,61057	0,833552
Switzerland	4	7,49400	1,56498	1,51691	0,858131
Finland	5	7,46900	1,44357	1,54025	0,809158
Netherlands	6	7,37700	1,50394	1,42894	0,810696

Element = attribute or item



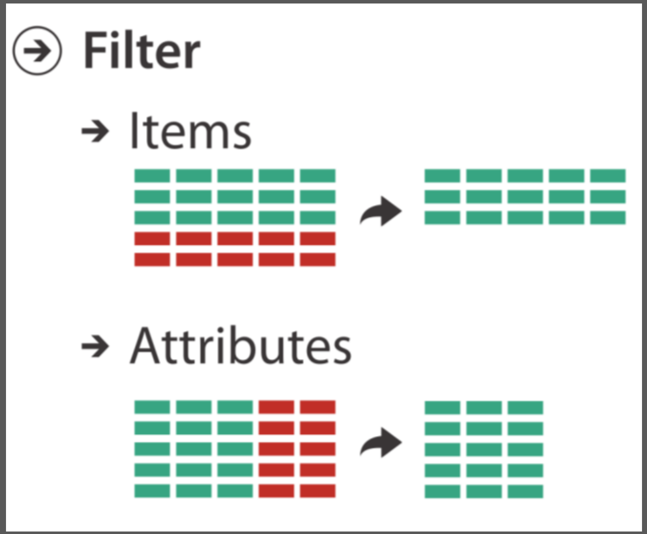
Can be combined

Bidirectional (reduce <-> increase)

Filtering
Items
Attributes

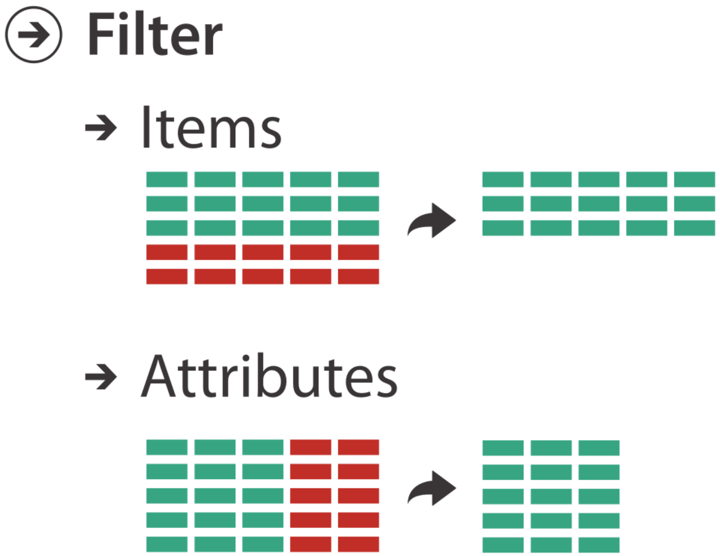
Aggregation
Items
Attributes - Dimensionality Reduction
Linear
Non-linear

Filtering



Filtering

- Pros:
  - Straight-forward & intuitive for user
  - Easy to compute
- Cons:
  - Out of sight, out of mind
  - Challenge to design effectively



Filtering

Items

Attributes

Aggregation

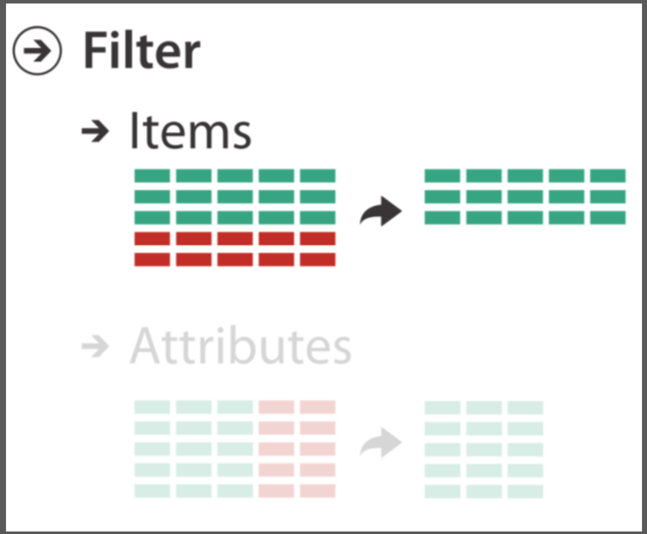
Items

Attributes - Dimensionality Reduction

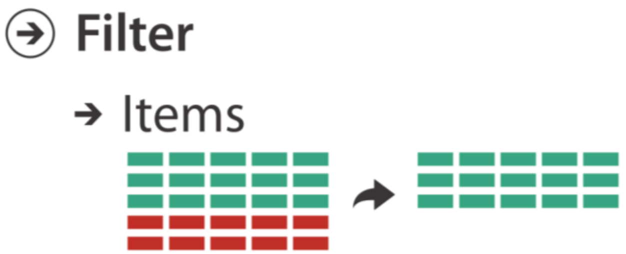
Linear

Non-linear

Item Filtering



Item filtering



Question: How to implement?

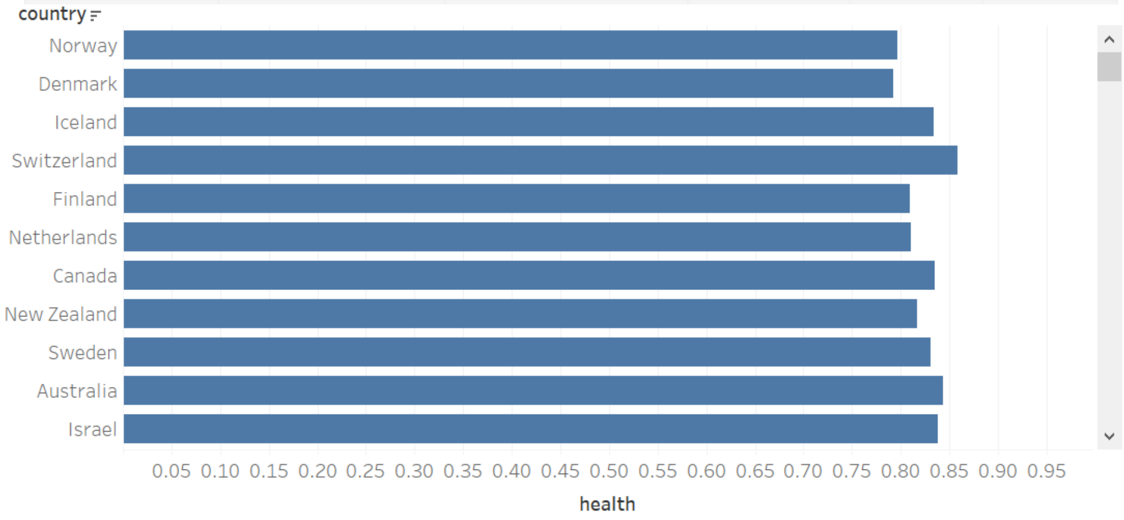
Health

Min 0,80

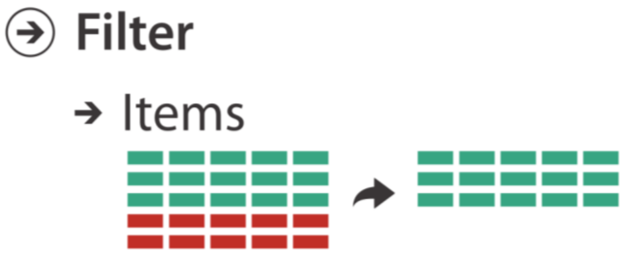
Max 0,81

Filter

country	happiness_rank	happiness_score	economy	family	health
Norway	1	7,53700	1,61646	1,53352	0,796667
Denmark	2	7,52200	1,48238	1,55112	0,792566
Iceland	3	7,50400	1,48063	1,61057	0,833552
Switzerland	4	7,49400	1,56498	1,51691	0,858131
Finland	5	7,46900	1,44357	1,54025	0,809158
Netherlands	6	7,37700	1,50394	1,42894	0,810696
Canada	7	7,31600	1,47920	1,48135	0,834558

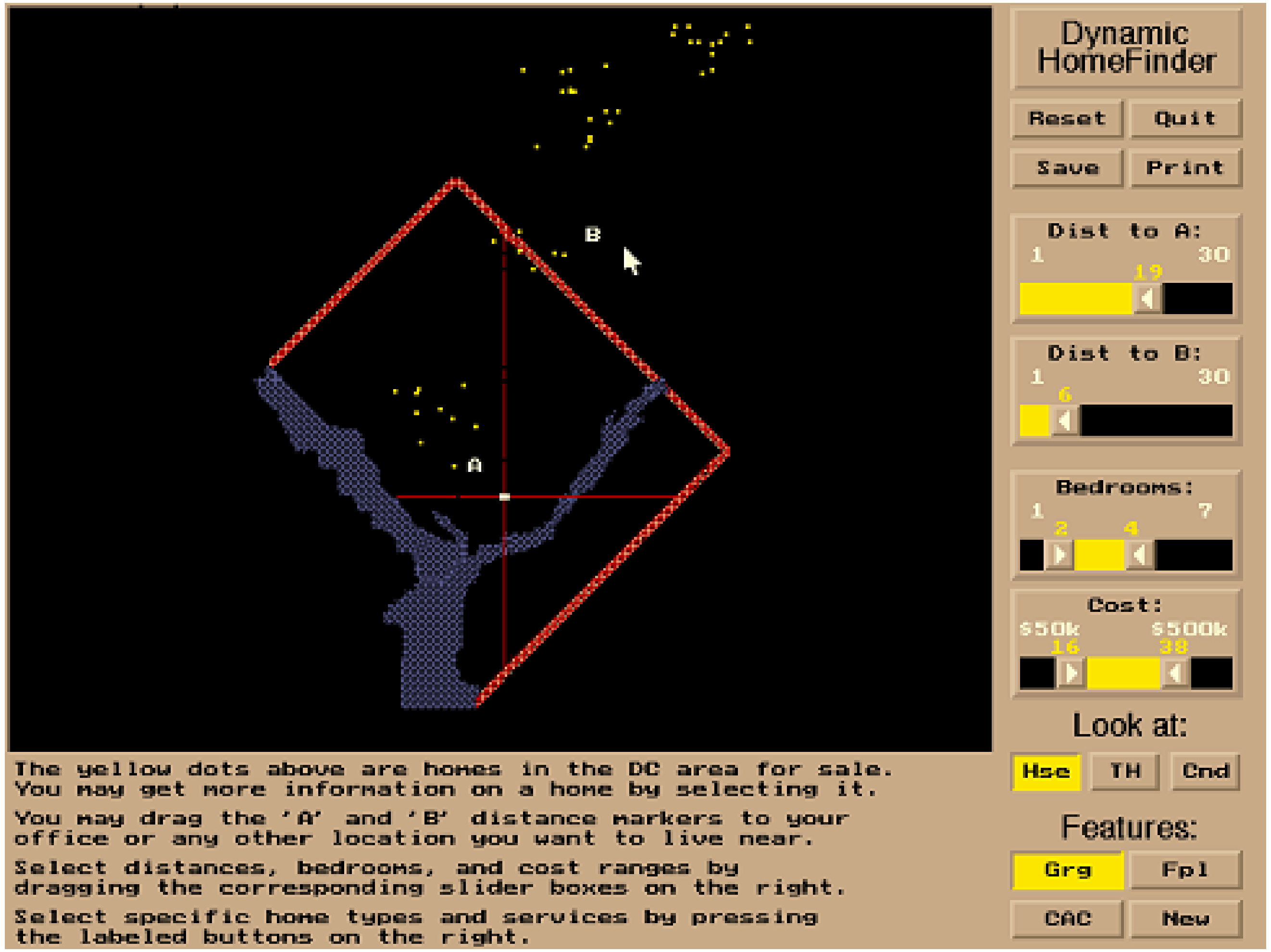


- Dynamic Queries:
  - Interactively change the view
  - Restrict items based on a query



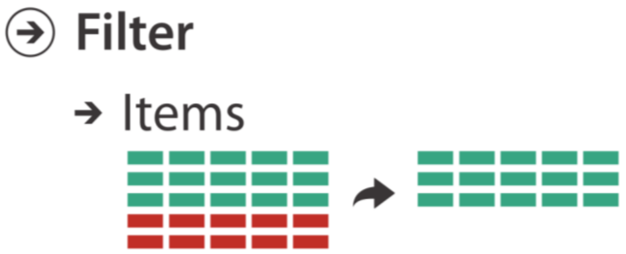
Dynamic Queries: HomeFinder

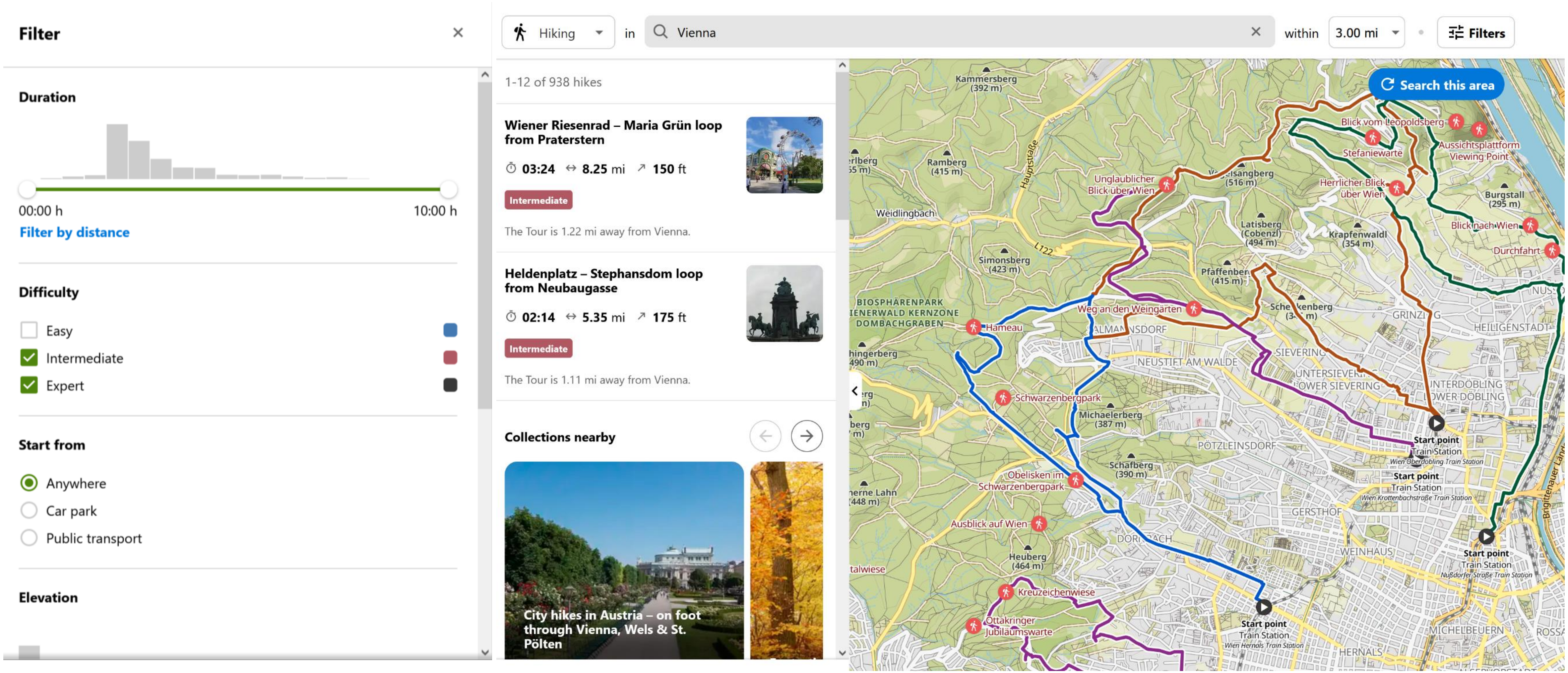
- Filter for each attribute
- Immediate visual feedback



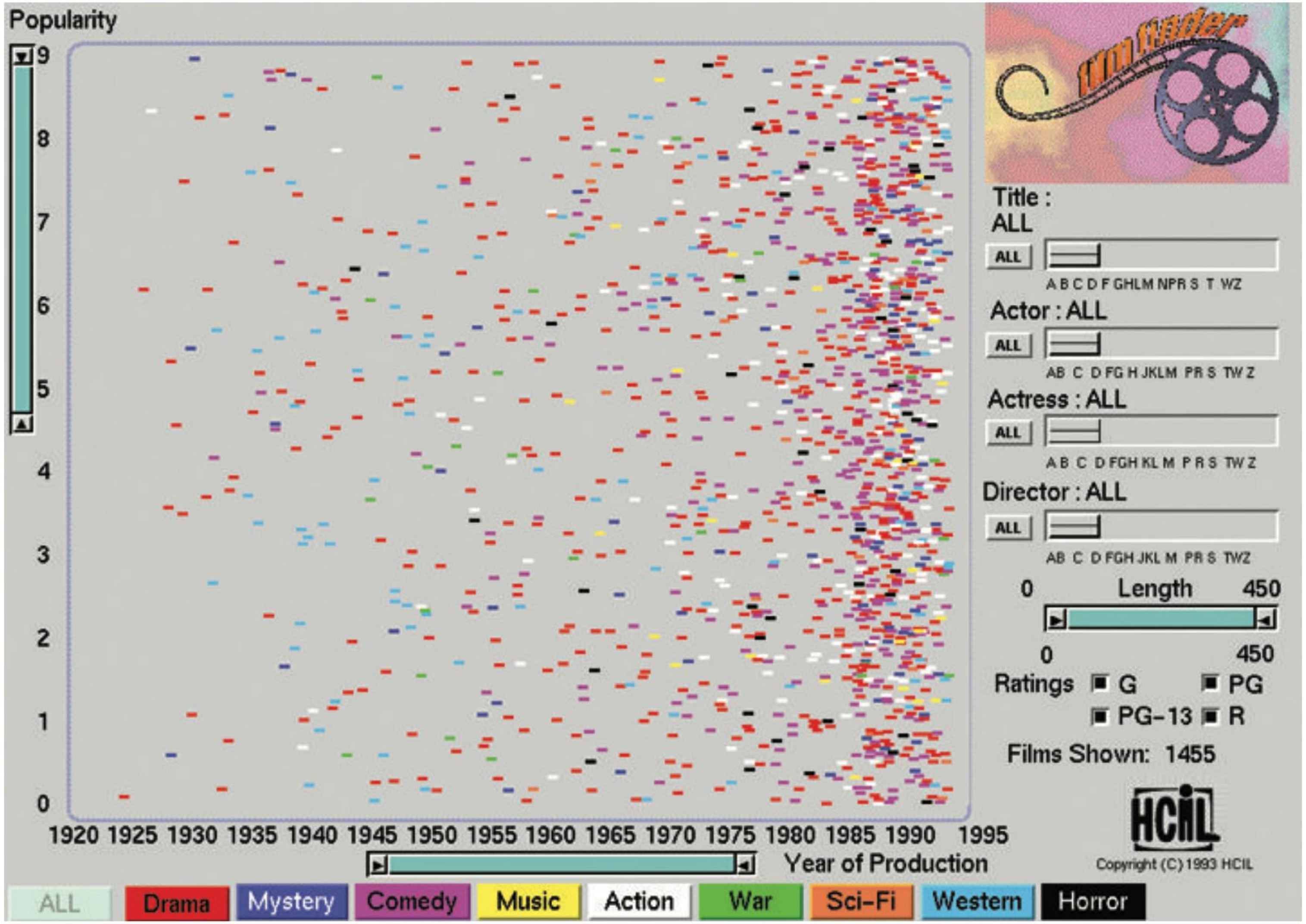
C. Williamson and B. Shneiderman, The dynamic HomeFinder: evaluating dynamic queries in a real-estate information exploration system, 1992, doi: [10.1145/133160.133216](https://doi.org/10.1145/133160.133216).

Dynamic Queries: Komoot





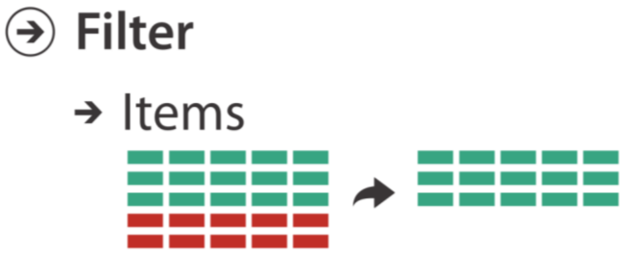
Dynamic Queries: Film Finder

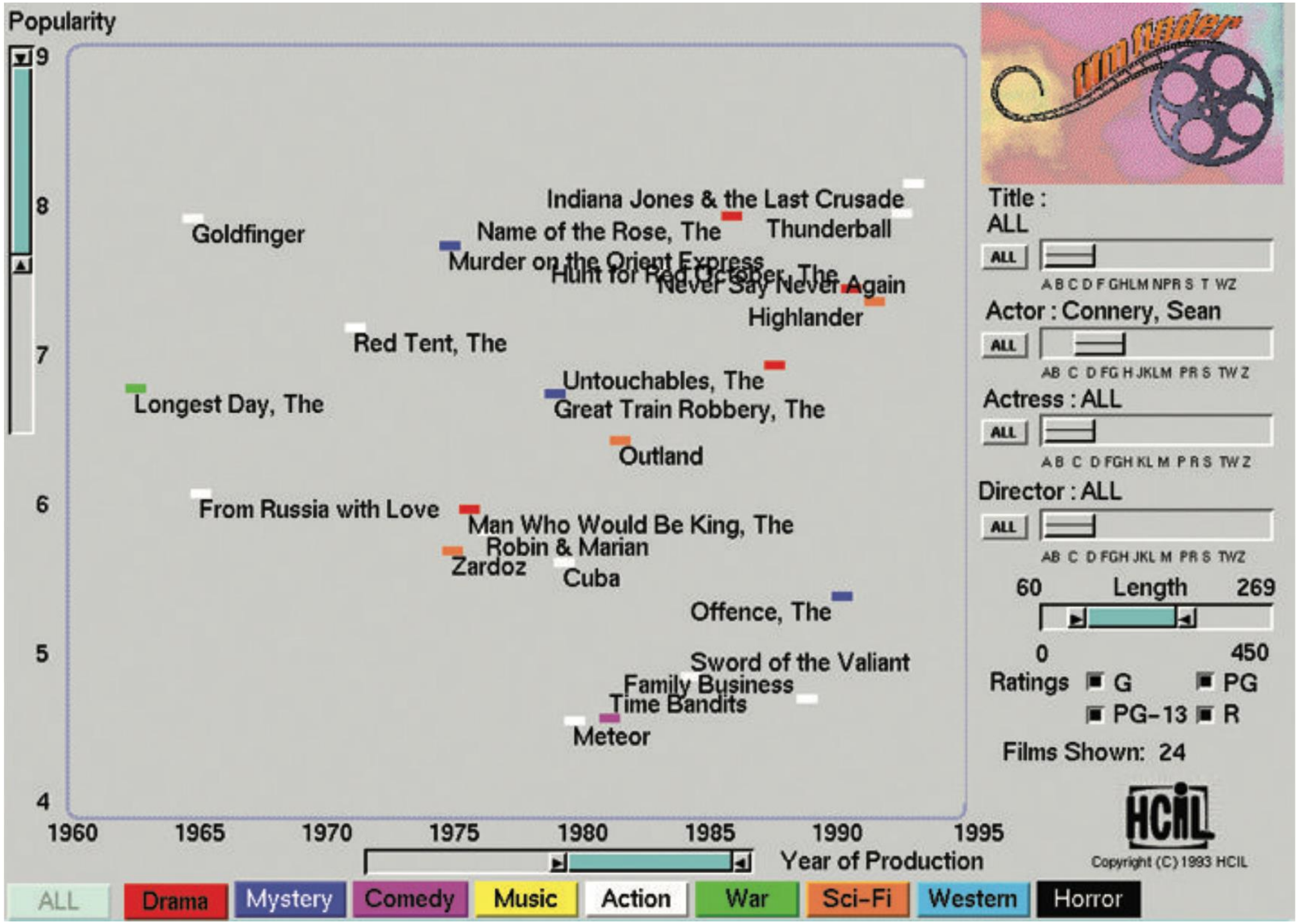


Overview over all films in database, color-coded by genre

C. Ahlberg and B. Shneiderman, Visual information seeking using the FilmFinder, 1994, doi: 10.1145/259963.260431.

Dynamic Queries: Film Finder





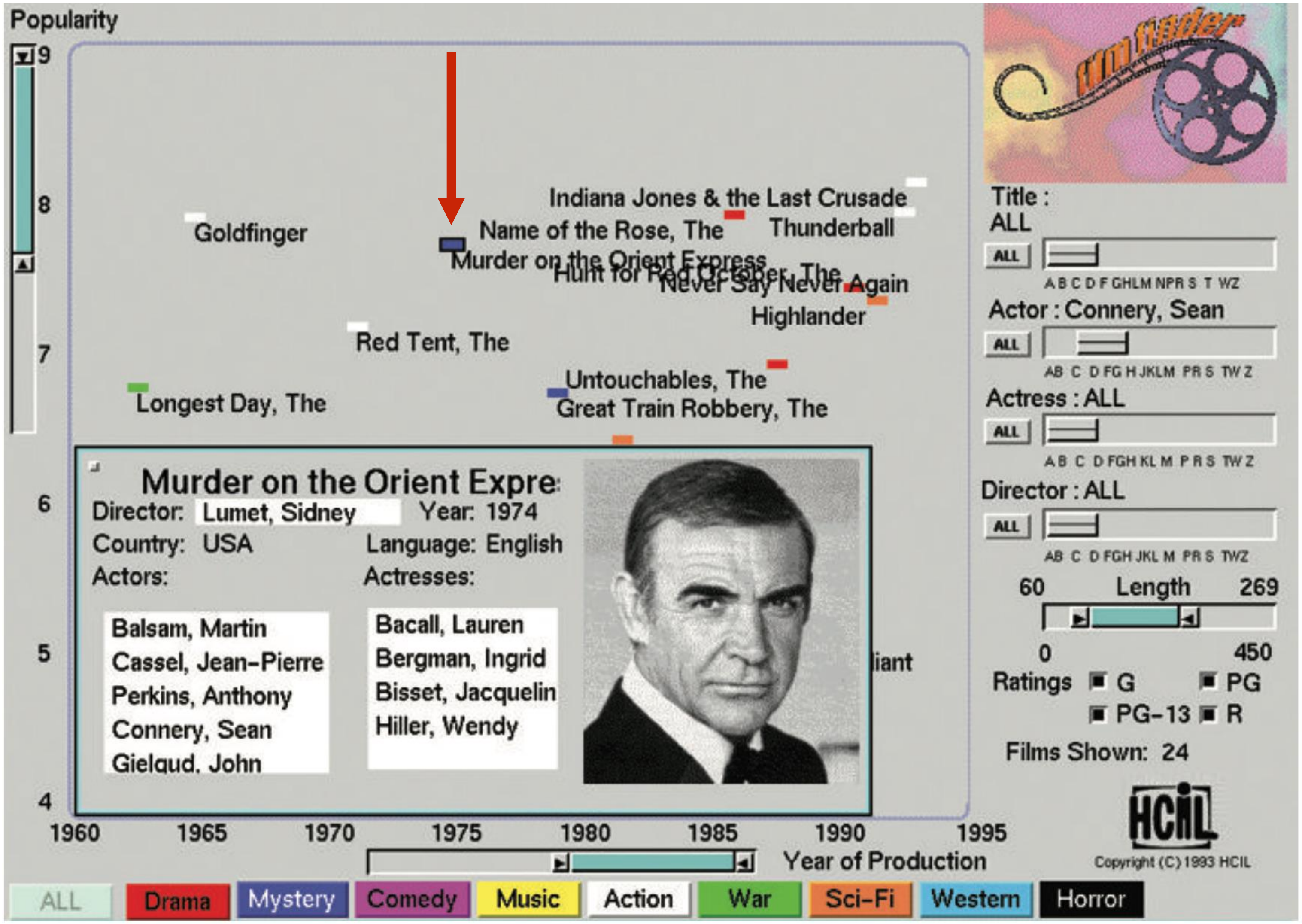
Filter by

Actor

Length

C. Ahlberg and B. Shneiderman, Visual information seeking using the FilmFinder, 1994, doi: [10.1145/259963.260431](https://doi.org/10.1145/259963.260431).

Dynamic Queries: Film Finder



Filter

→ Items

Select single item

C. Ahlberg and B. Shneiderman, Visual information seeking using the FilmFinder, 1994, doi: [10.1145/259963.260431](https://doi.org/10.1145/259963.260431).

Dynamic Queries: Film Finder

→ Filter

→ Items

Shneiderman’s Mantra:

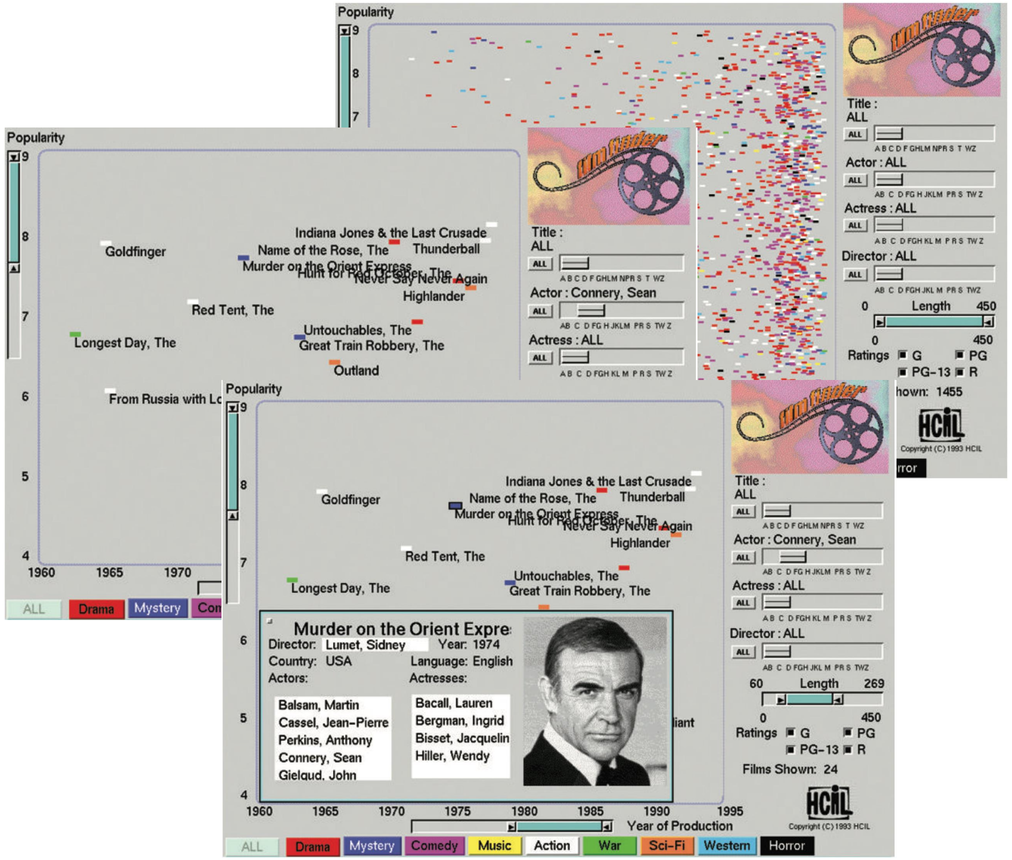
Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, doi: [10.1109/VL.1996.545307](https://doi.org/10.1109/VL.1996.545307).

Overview first,

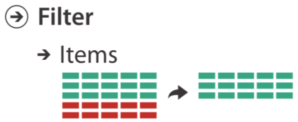
Zoom and filter,

Details on demand

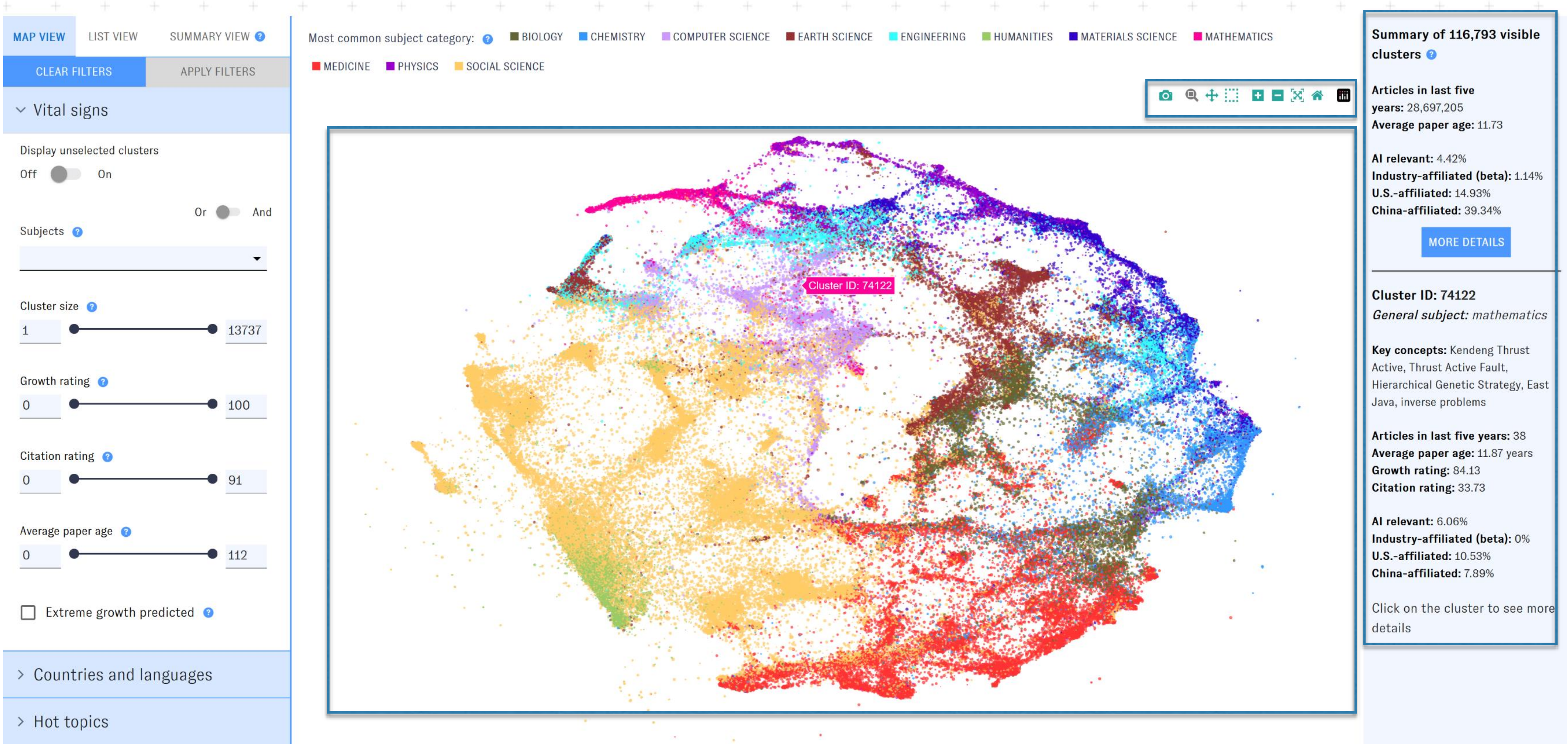
→ Facet into Multiple Views



C. Ahlberg and B. Shneiderman, Visual information seeking using the FilmFinder, 1994, doi: [10.1145/259963.260431](https://doi.org/10.1145/259963.260431).

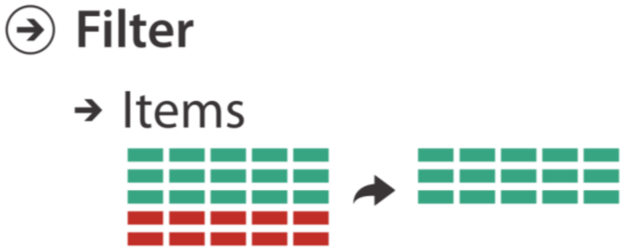


Dynamic Queries: Map of Science

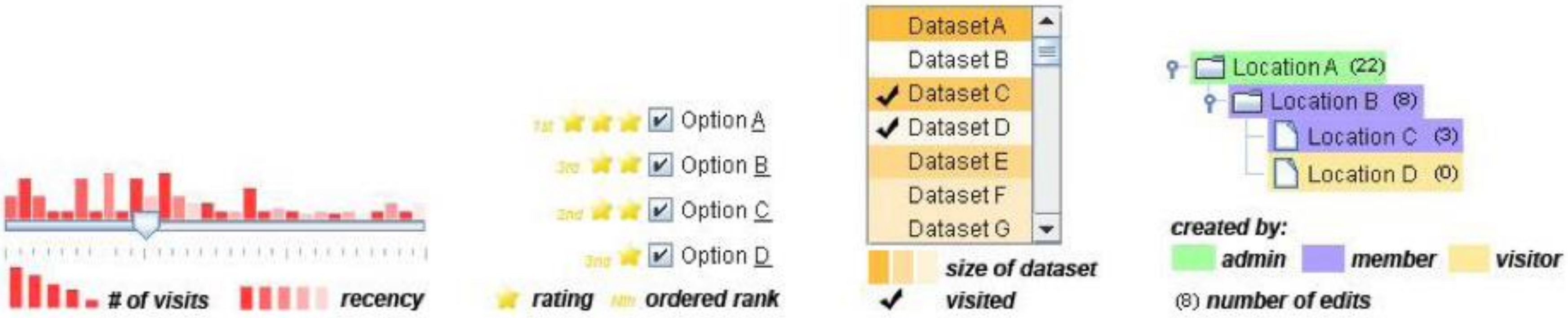


sciencemap.eto.tech

Scented widgets

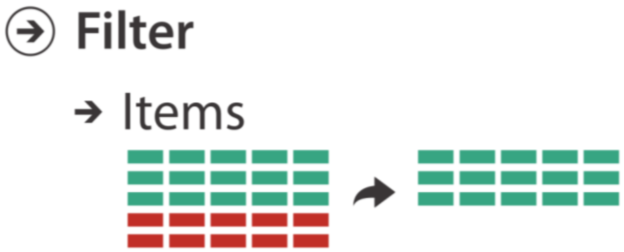


Provide information to user about the impact of filtering:

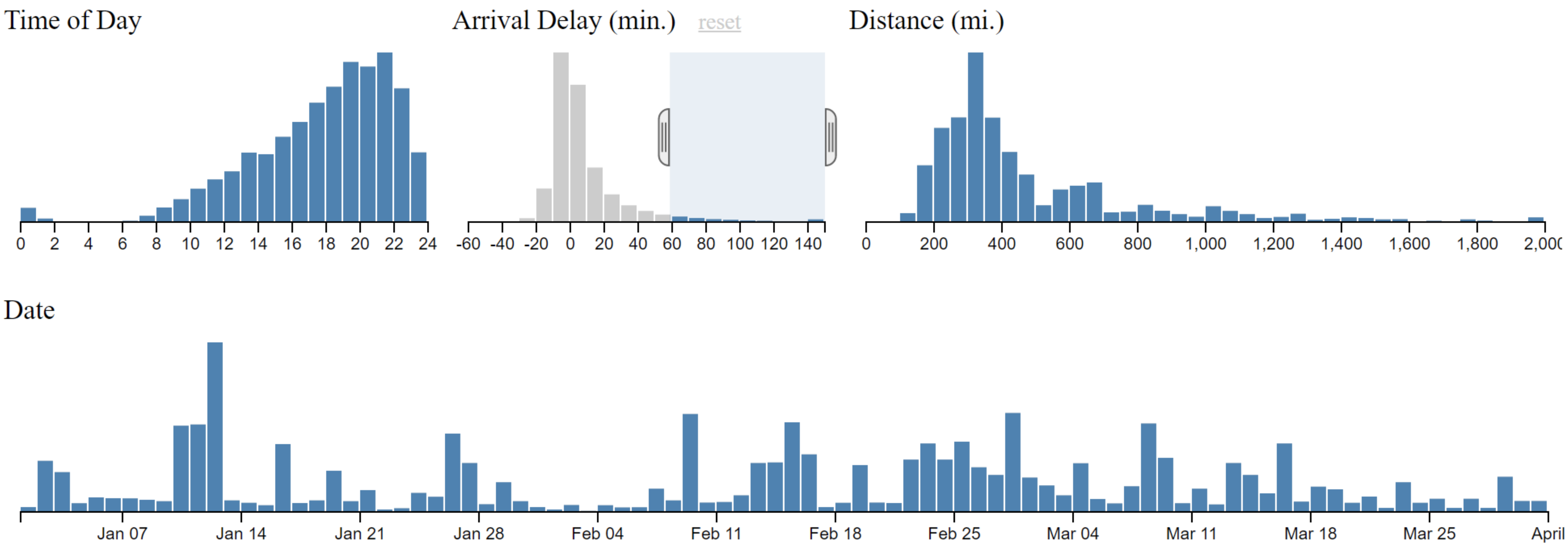


W. Willett et al., Scented Widgets: Improving Navigation Cues with Embedded Visualizations, 2007, doi: [10.1109/TVCG.2007.70589](https://doi.org/10.1109/TVCG.2007.70589).

Cross Filtering



Coordinated views/filter controls combined:



http://square.github.io/crossfilter/

→ Facet into Multiple Views

Filtering

Items

Attributes

Aggregation

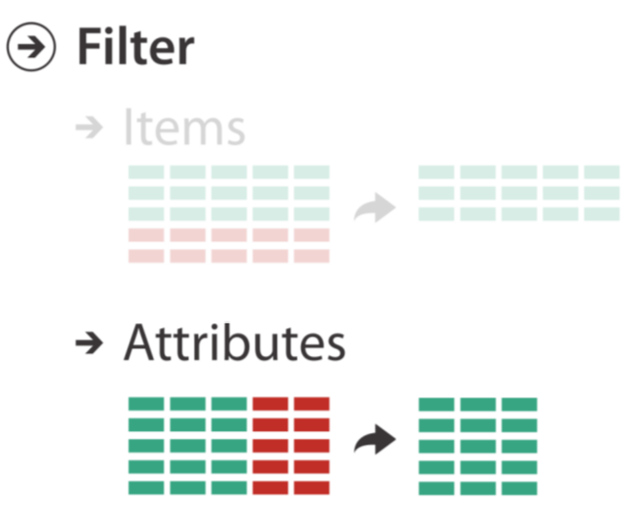
Items

Attributes - Dimensionality Reduction

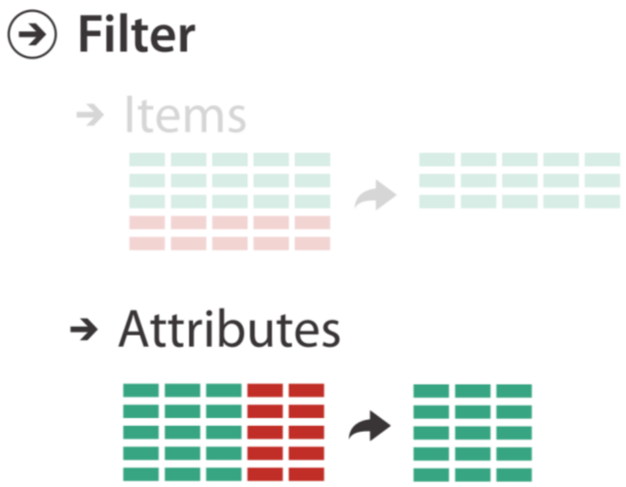
Linear

Non-linear

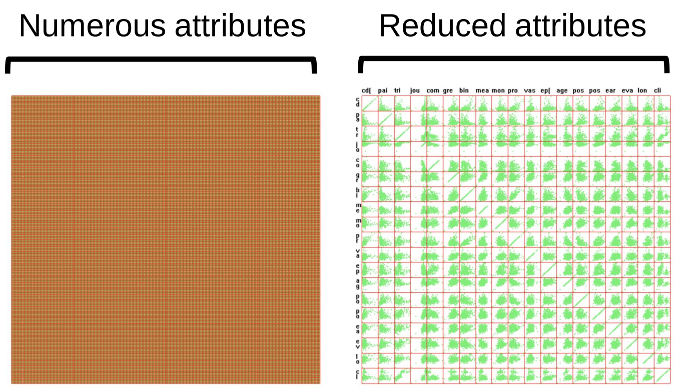
Attribute Filtering



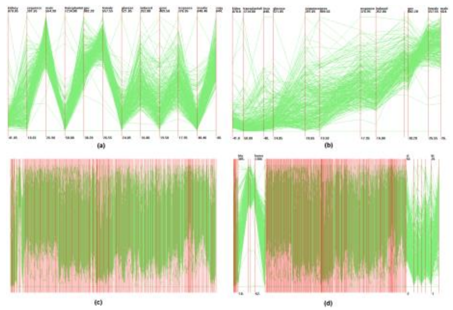
Attribute Filtering



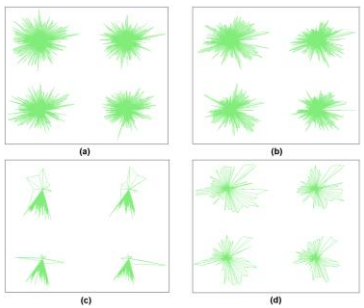
- Eliminate attributes rather than items
- Often combined with attribute ordering
  - Similarity/Importance metric for attribute (e.g. variance)
- Dimensional Ordering, Spacing, and Filtering Approach (DOSFA)



Scatterplot Matrix



Parallel Coordinates



Star Glyphs

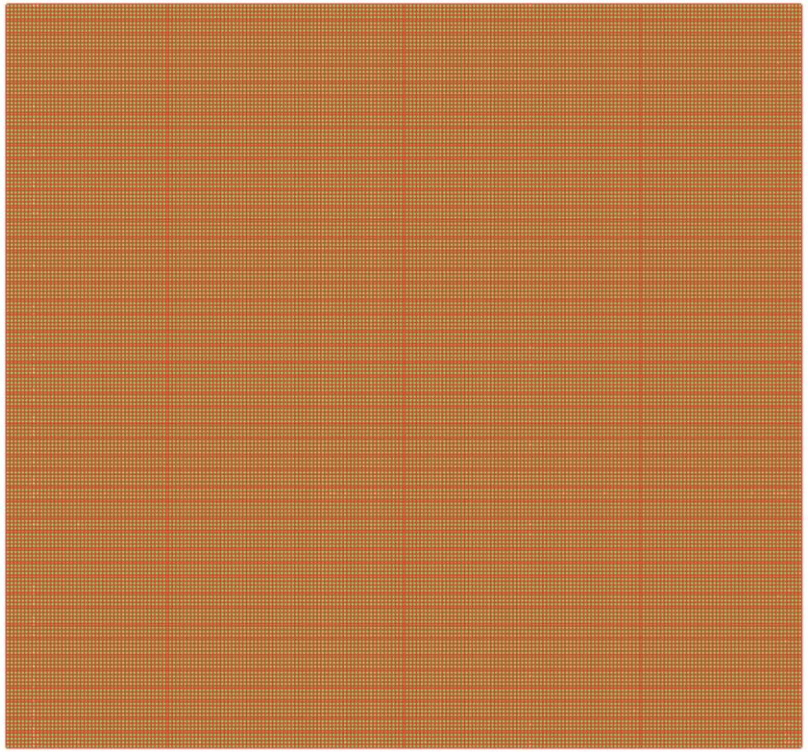
J. Yang et al., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets, 2003, doi: [10.1109/INFVIS.2003.1249015](https://doi.org/10.1109/INFVIS.2003.1249015).

Scatterplot Matrix

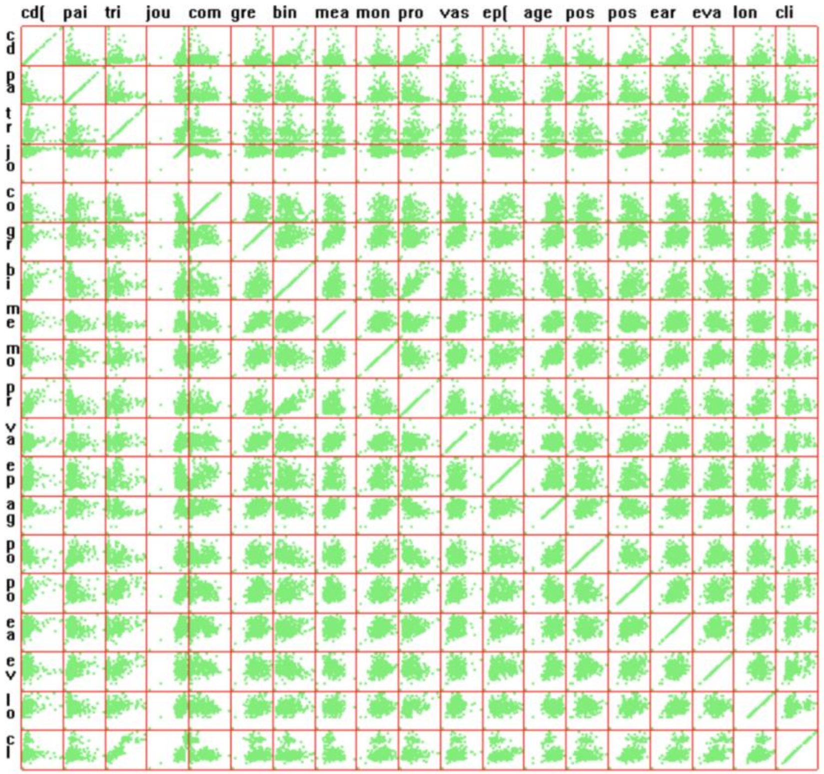
Filter

→ Attributes

Unfiltered



Filtered



J. Yang et al., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets, 2003, doi: [10.1109/INFVIS.2003.1249015](https://doi.org/10.1109/INFVIS.2003.1249015).

Parallel Coordinates

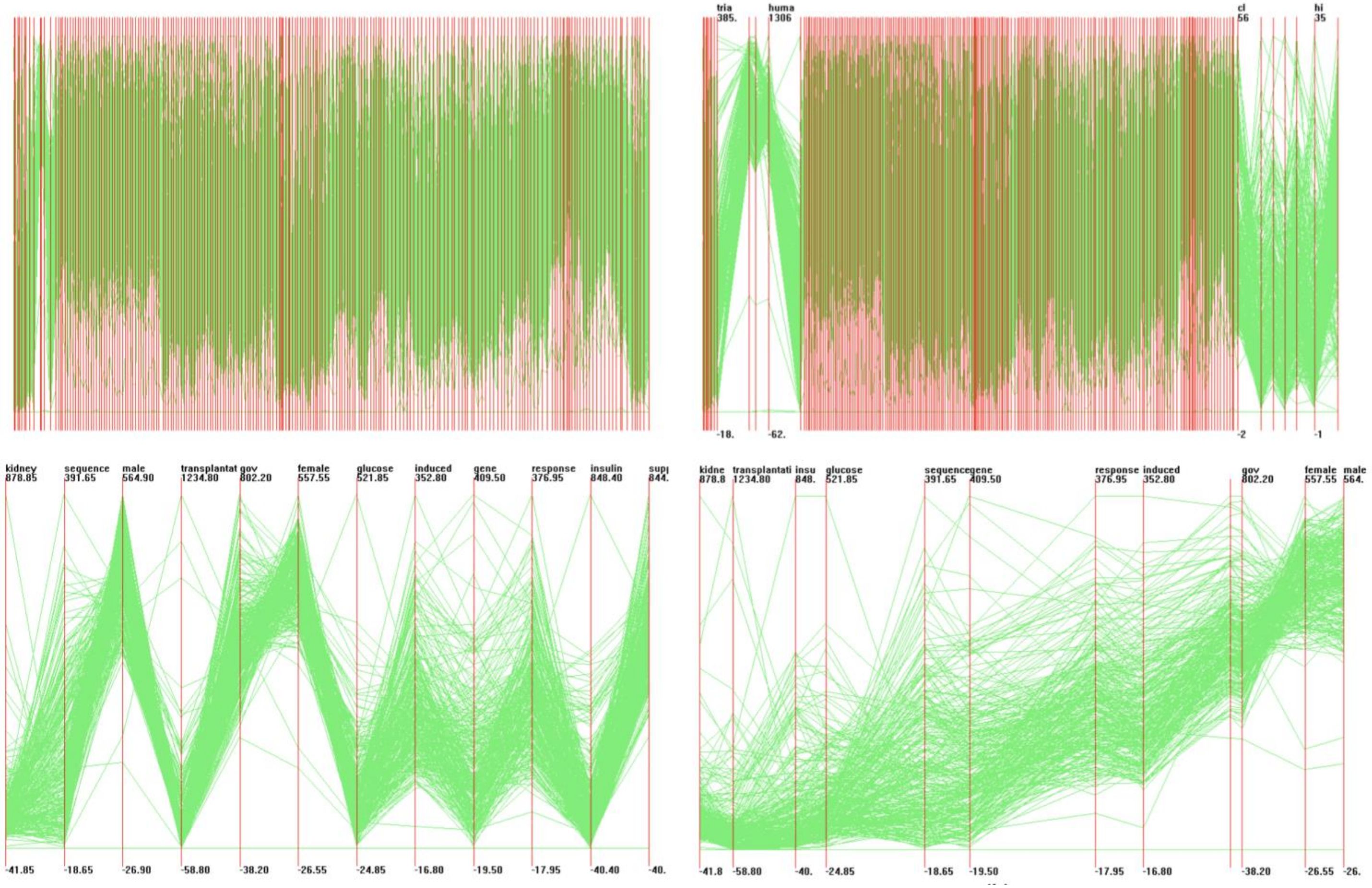
Filter

→ Attributes

Unordered | Ordered + Spaced

Unfiltered

Filtered



J. Yang et al., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets, 2003, doi: [10.1109/INFVIS.2003.1249015](https://doi.org/10.1109/INFVIS.2003.1249015).

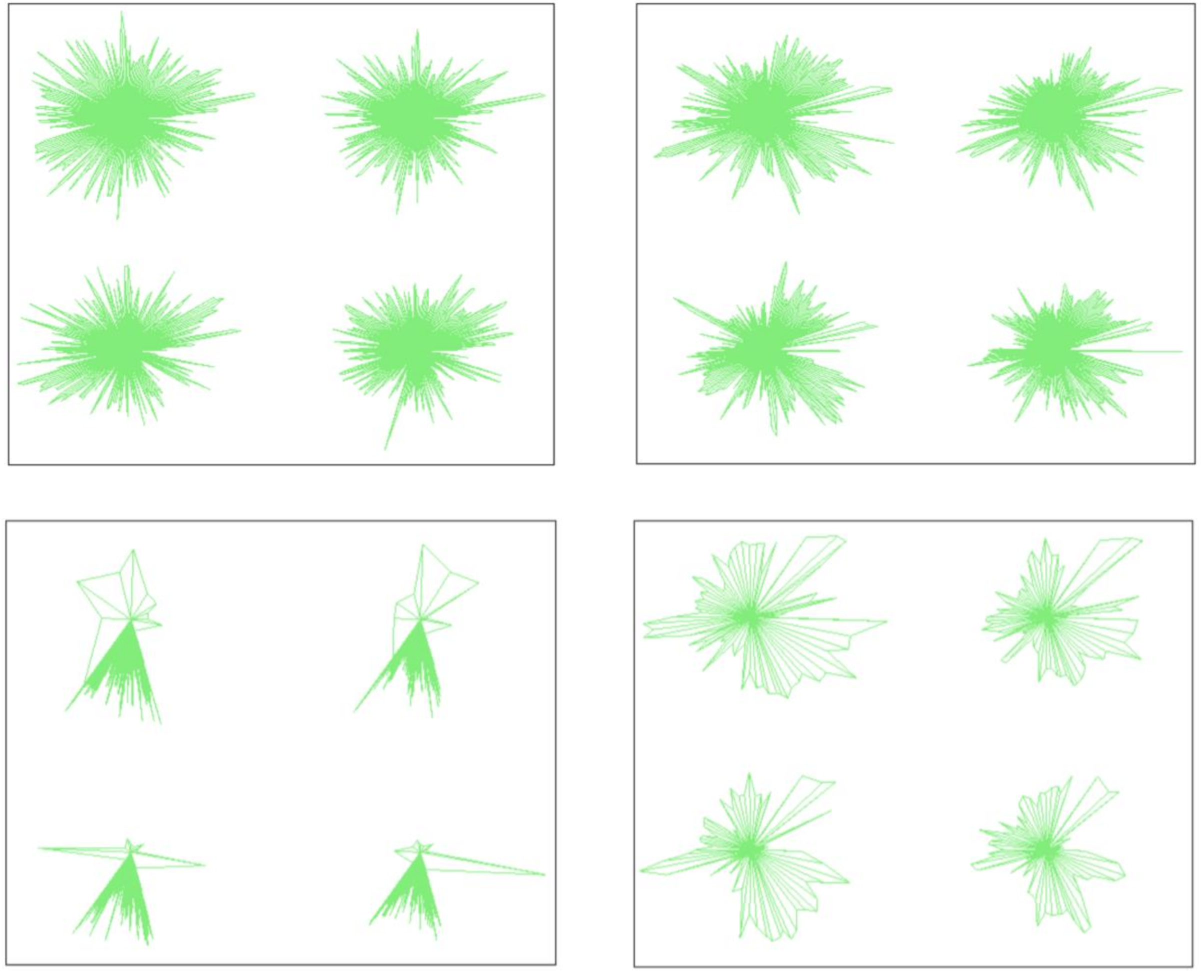
Star Glyphs

Filter
→ Attributes

Unordered | Ordered

Unfiltered

Filtered



→ Arrange Tables
+ Spatial Data

J. Yang et al., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets, 2003, doi: [10.1109/INFVIS.2003.1249015](https://doi.org/10.1109/INFVIS.2003.1249015).

Filtering

Items

Attributes

Aggregation

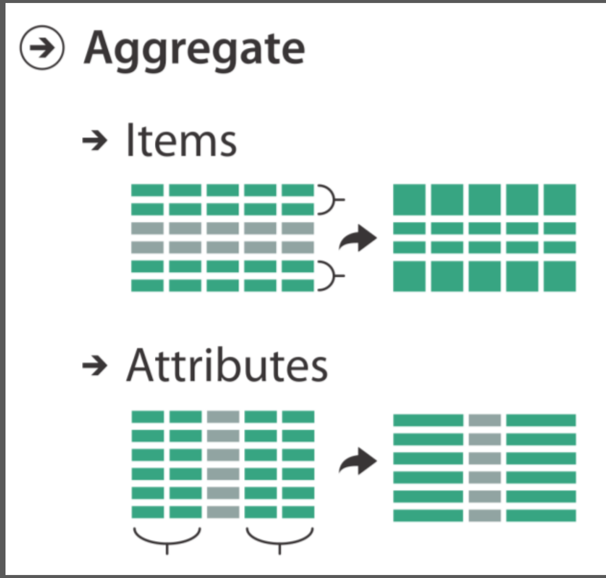
Items

Attributes - Dimensionality Reduction

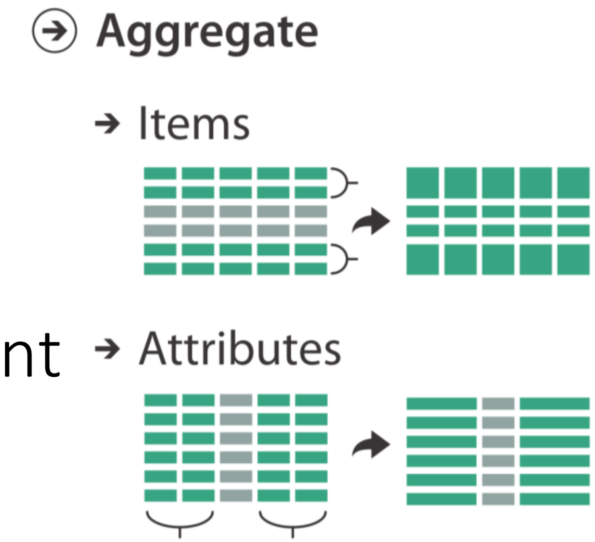
Linear

Non-linear

Aggregation



Aggregation



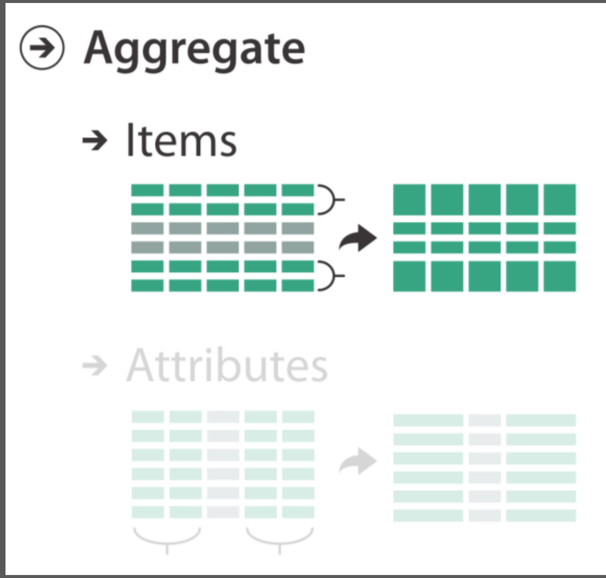
Group of elements represented by new derived element

- Pros:
  - Potentially safer than filter from cognitive point of view, as it…
  - …conveys (some) information about entire set it replaces
- Cons:
  - Cannot convey all information
  - Challenge: Design/Implementation:
    How/What to summarize? Needs to match task!

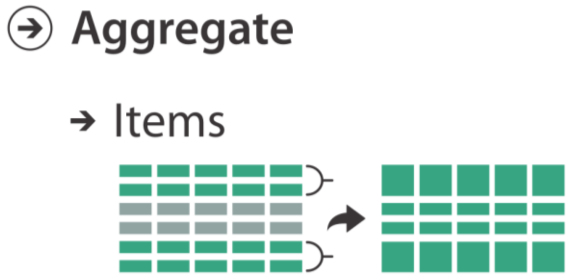
Filtering
- Items
- Attributes

Aggregation
- Items
- Attributes - Dimensionality Reduction
  - Linear
  - Non-linear

Item Aggregation



Item Aggregation



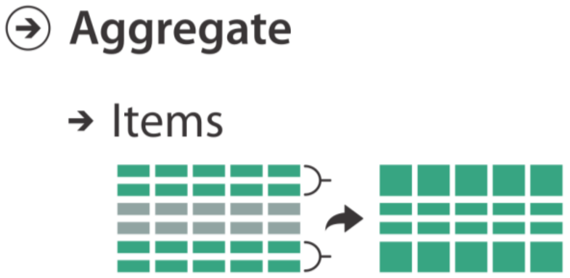
- Basic aggregation operators:
  - Average (arithmetic, geometric, mode, median, …)
  - Count, Sum
  - Minimum, Maximum
- Other grouping operations: binning, hierarchical clustering, …
- Challenge:
  Avoid eliminating interesting signals

Item Aggregation

Question: How to implement?

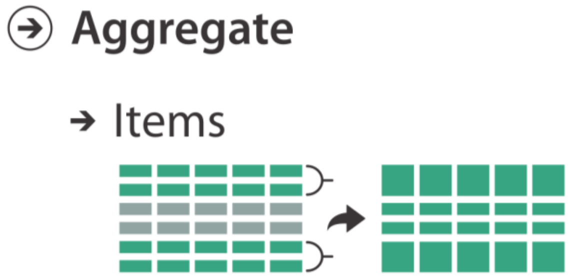
Example Task:

Distribution of happiness_score



country	happiness_rank	happiness_score	economy	family	health
Norway	1	7,53700	1,61646	1,53352	0,796667
Denmark	2	7,52200	1,48238	1,55112	0,792566
Iceland	3	7,50400	1,48063	1,61057	0,833552
Switzerland	4	7,49400	1,56498	1,51691	0,858131
Finland	5	7,46900	1,44357	1,54025	0,809158
Netherlands	6	7,37700	1,50394	1,42894	0,810696
Canada	7	7,31600	1,47920	1,48135	0,834558
New Zealand	8	7,31400	1,40571	1,54820	0,816760
Sweden	9	7,28400	1,49439	1,47816	0,830875
Australia	10	7,28400	1,48441	1,51004	0,843887
Israel	11	7,21300	1,37538	1,37629	0,838404
Costa Rica	12	7,07900	1,10971	1,41640	0,759509
Austria	13	7,00600	1,48710	1,45994	0,815328
United States	14	6,99300	1,54626	1,41992	0,774287
Ireland	15	6,97700	1,53571	1,55823	0,809783
Germany	16	6,95100	1,48792	1,47252	0,798951
Belgium	17	6,89100	1,46378	1,46231	0,818092

Histogram



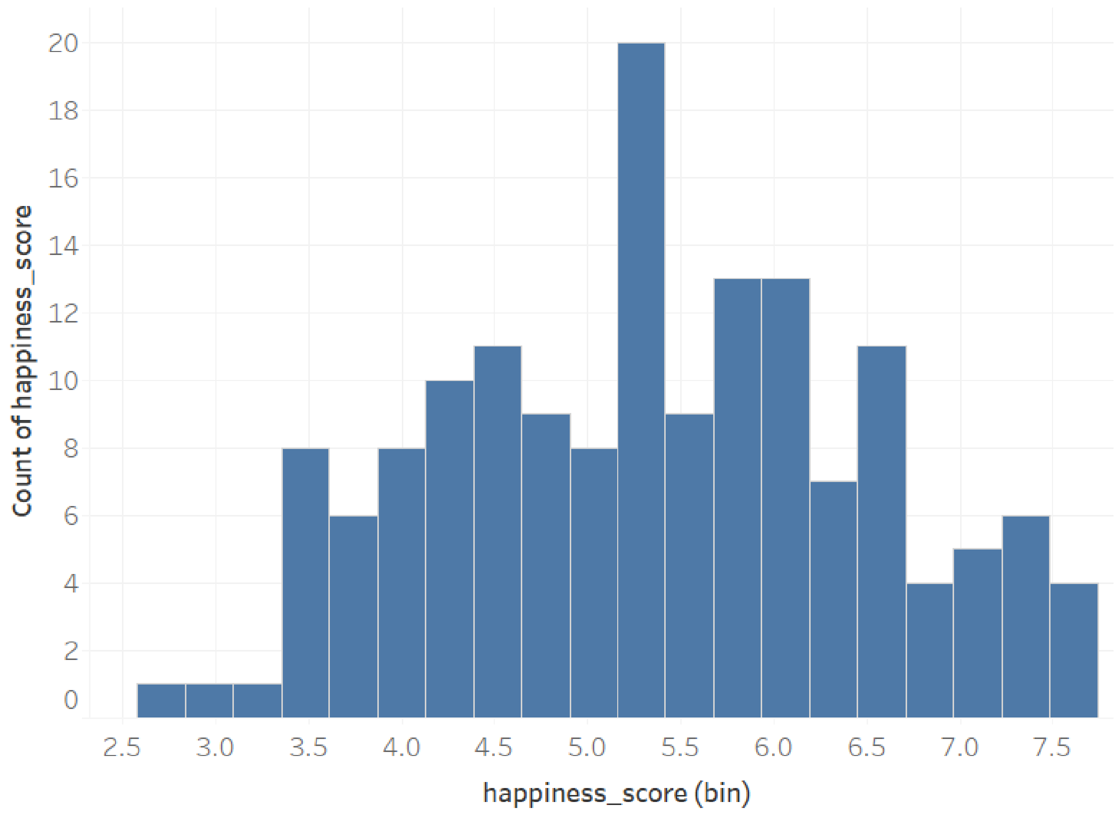
Task: Find distribution

Derived data – new table:

- Keys are bins (ranges)
- Values are counts
  (how many values fall into the respective range)

Note:

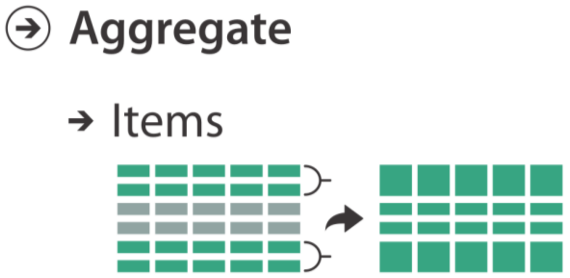
- Does not show data directly, instead shows its distribution
- Bin size crucial!



using suggested bin size – 0.258

Tableau Software

Histogram



Bin size:

- Can affect pattern drastically
- Opportunity for interaction: Control bin size

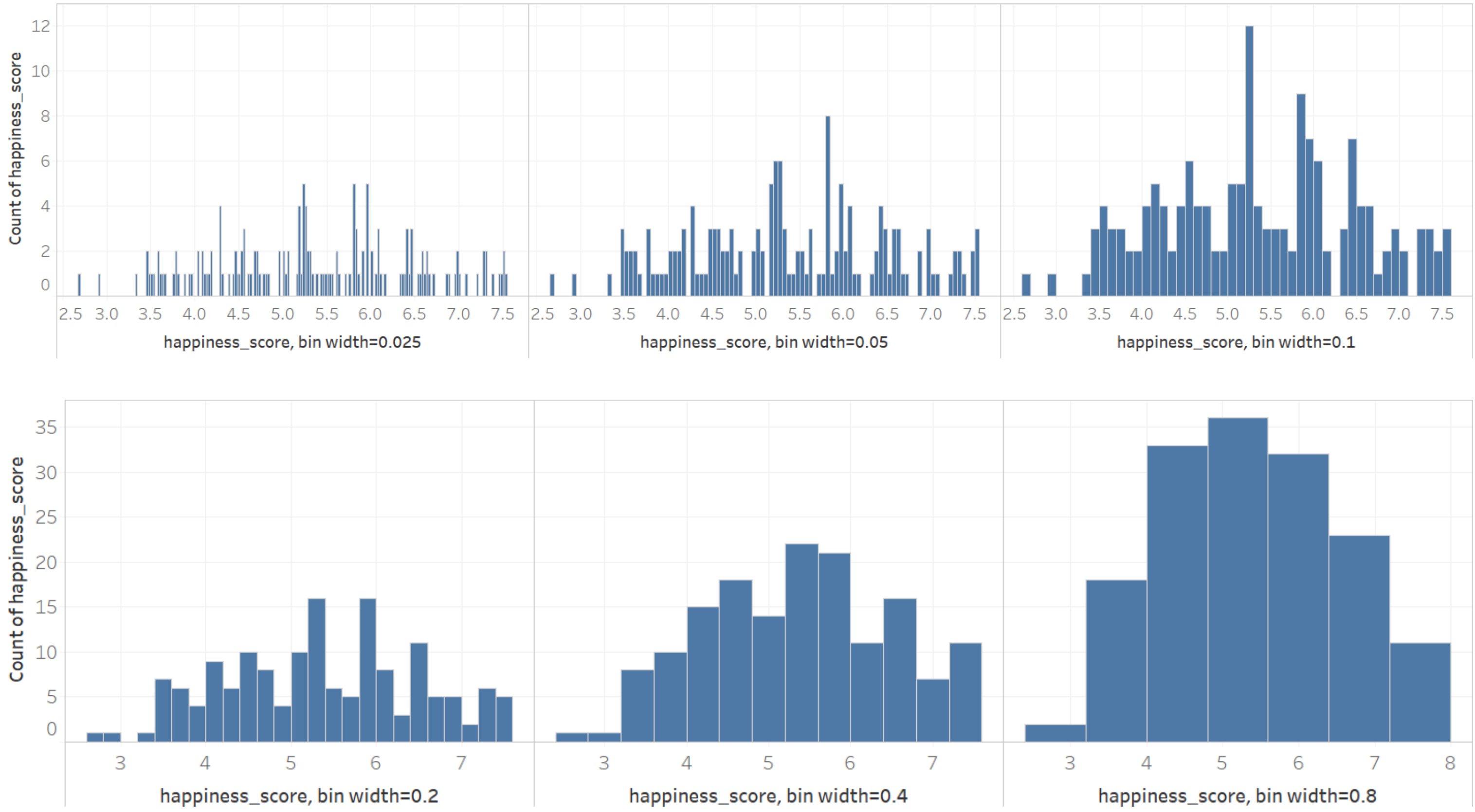
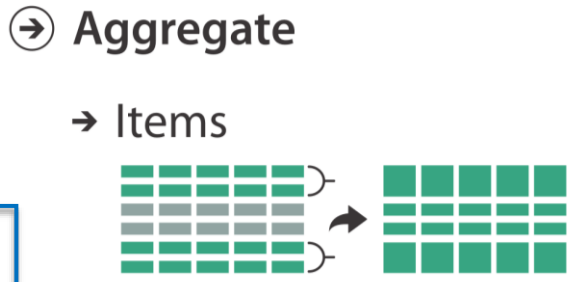
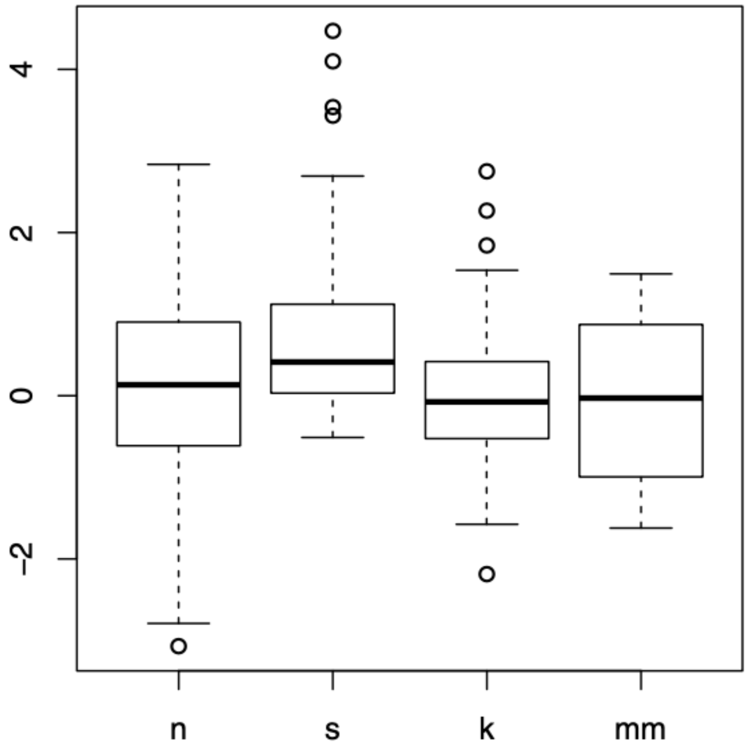
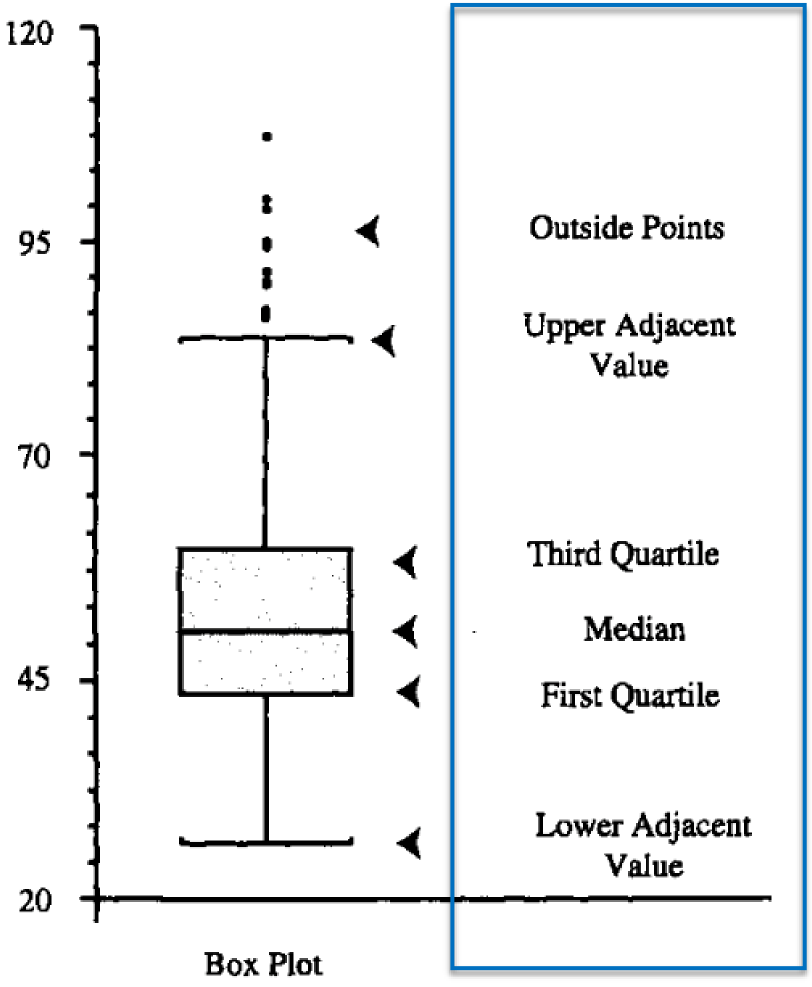


Tableau Software

Box Plots

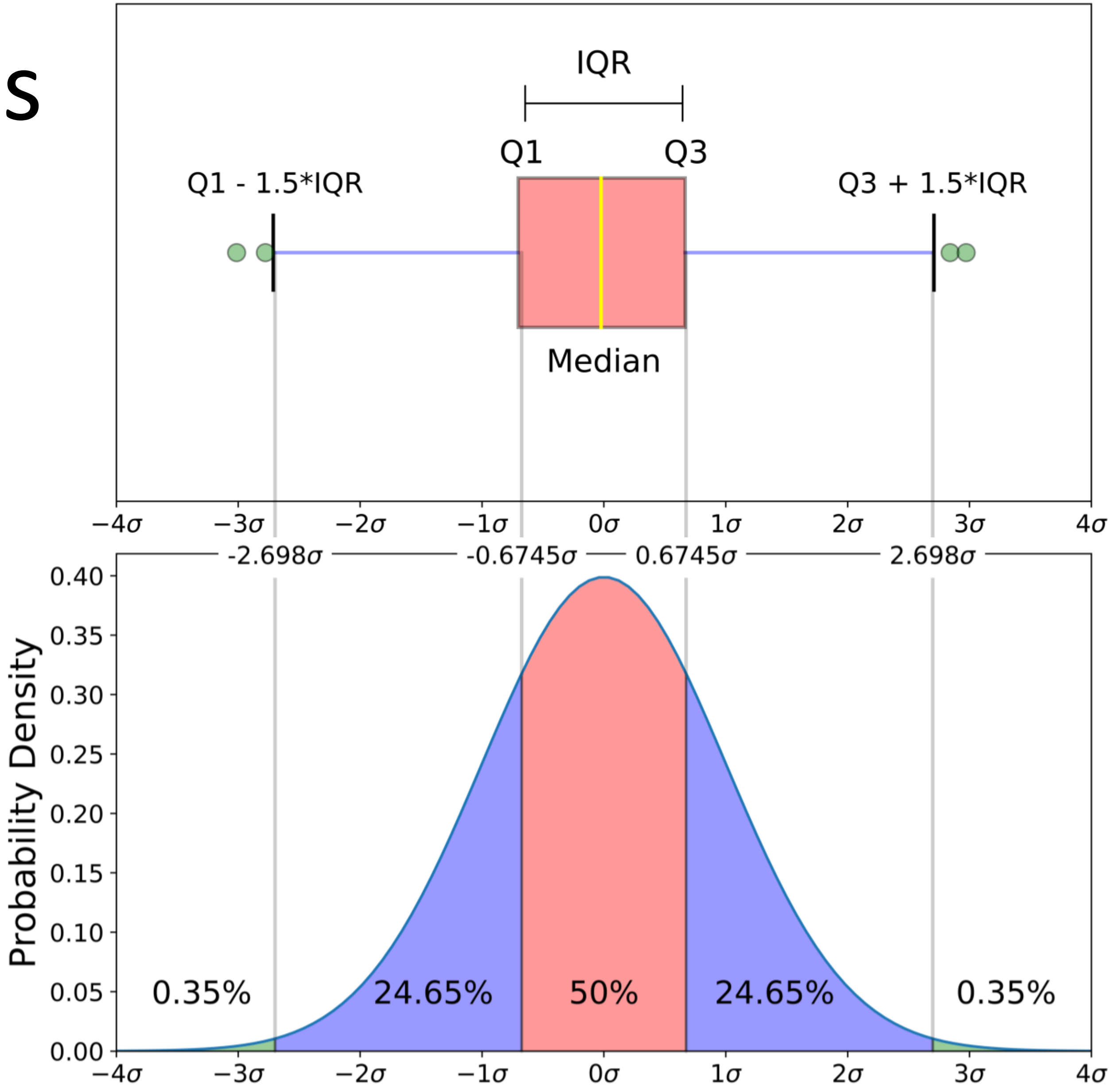




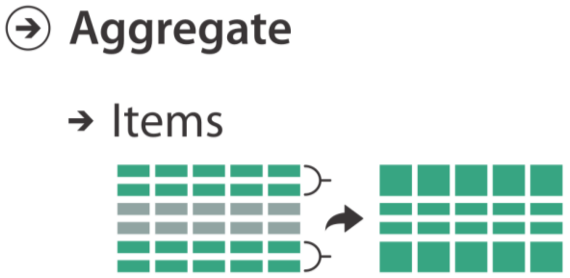


J. L. Hintze and R.D. Nelson, Violin Plots: A Box Plot-Density Trace Synergism, 1998, doi: [10.1080/00031305.1998.10480559](10.1080/00031305.1998.10480559).

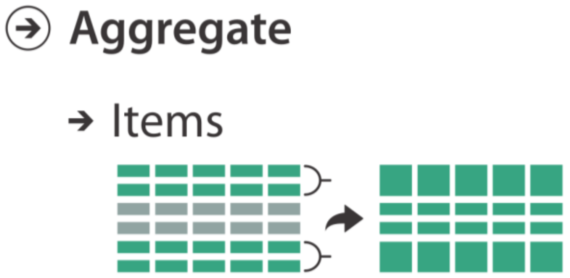
Box Plots



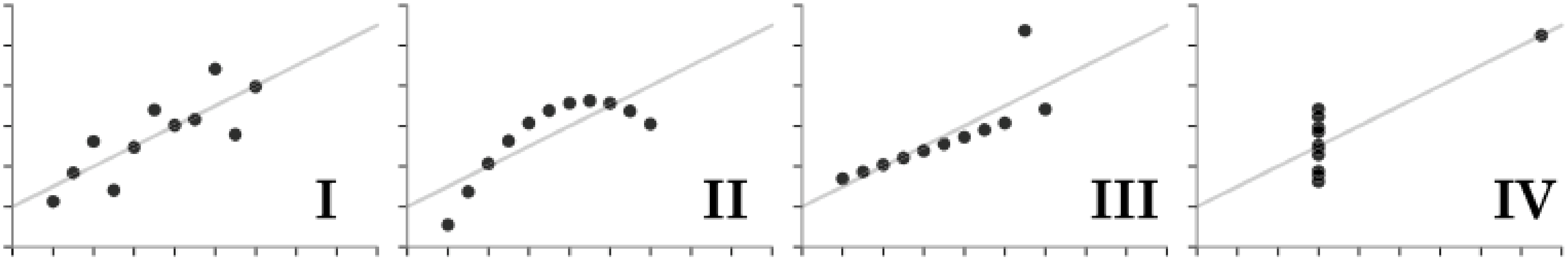
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51



Anscombe’s Quartet



All four datasets have same mean, standard deviation and correlation:

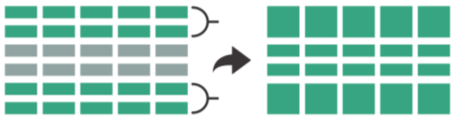


F. J. Anscombe, Graphs in Statistical Analysis, 1973, doi: [10.1080/00031305.1973.10478966](https://doi.org/10.1080/00031305.1973.10478966).

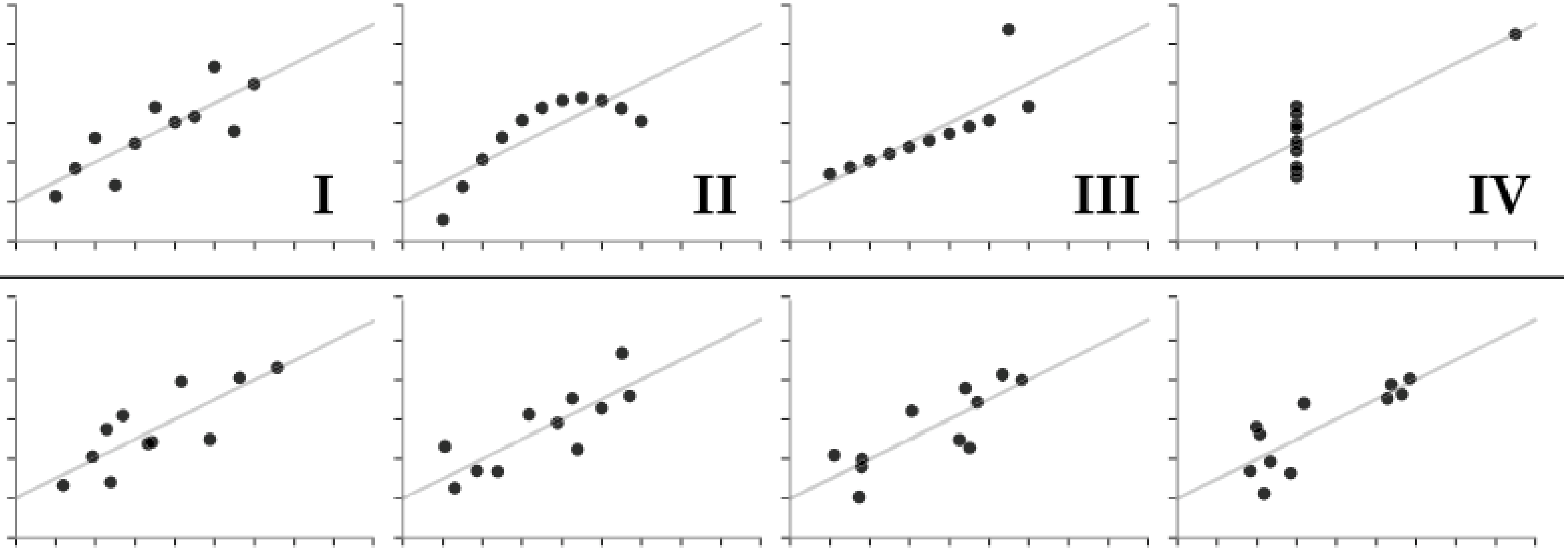
Same Stats, Different Graphs

Aggregate

Items



Algorithm to create any arbitrary dataset with the summary statistics of a given dataset

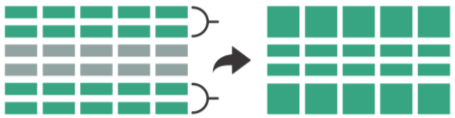


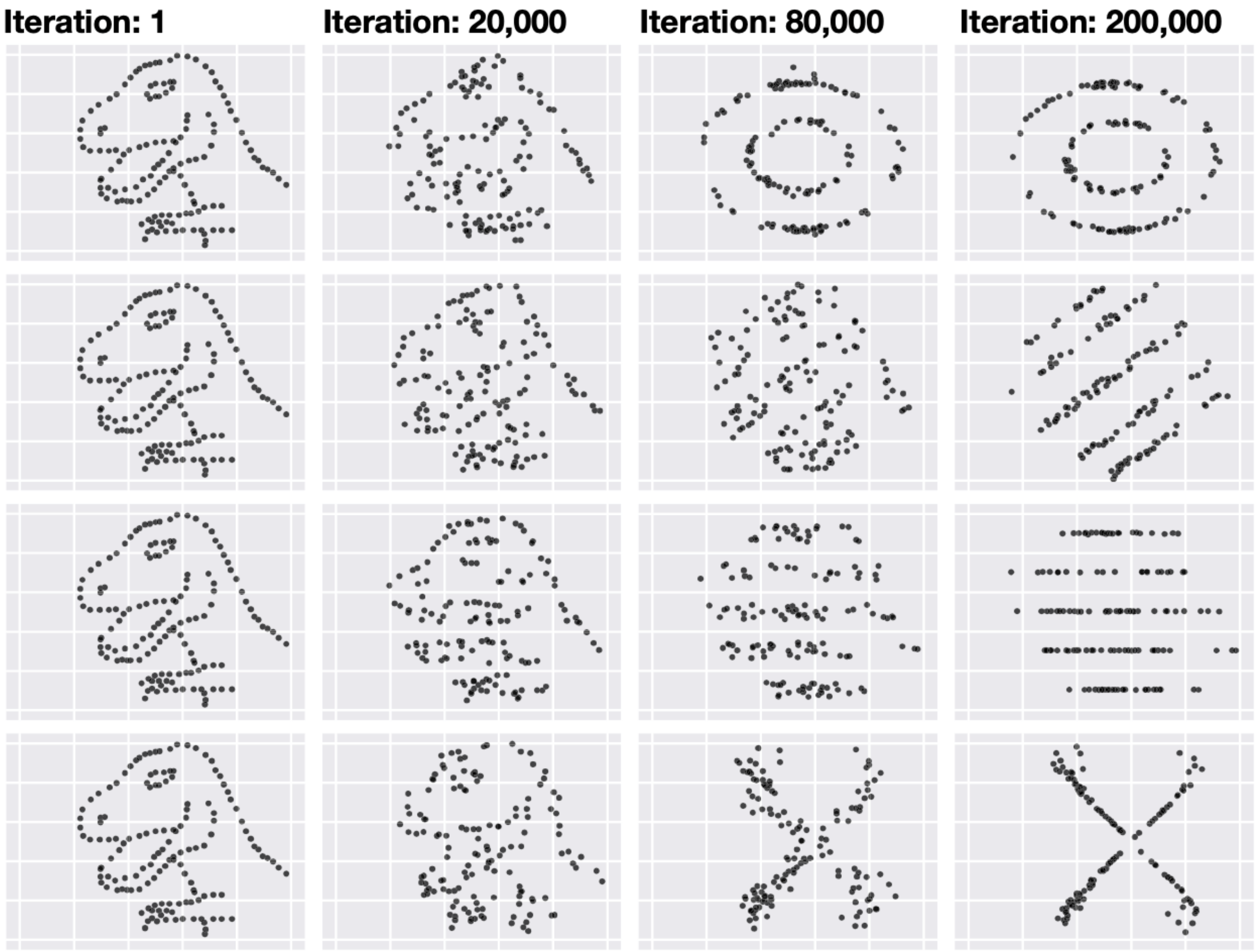
J. Matejka and G. Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, 2017, doi: [10.1145/3025453.3025912](https://doi.org/10.1145/3025453.3025912).

Same Stats, Different Graphs

Aggregate

→ Items

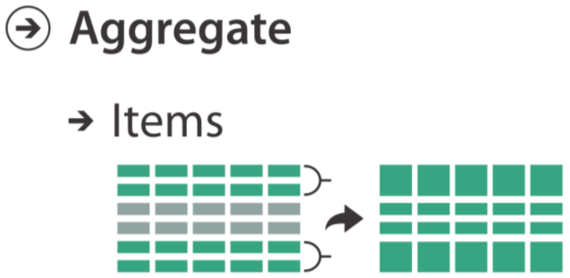


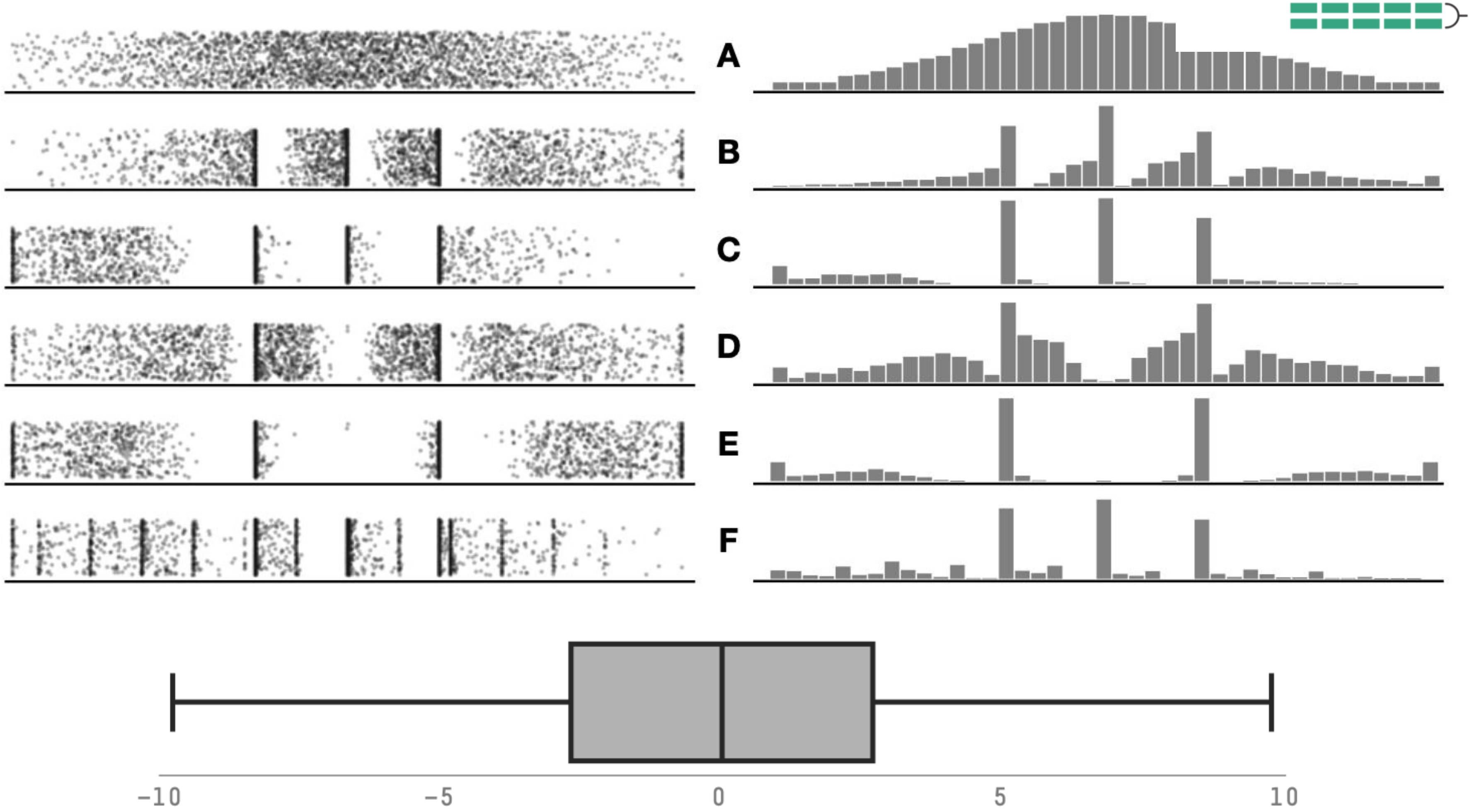



```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

J. Matejka and G. Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, 2017, doi: 10.1145/3025453.3025912.

Same Stats, Different Graphs



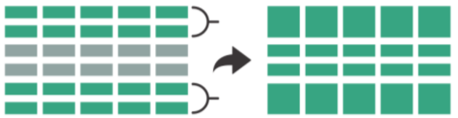


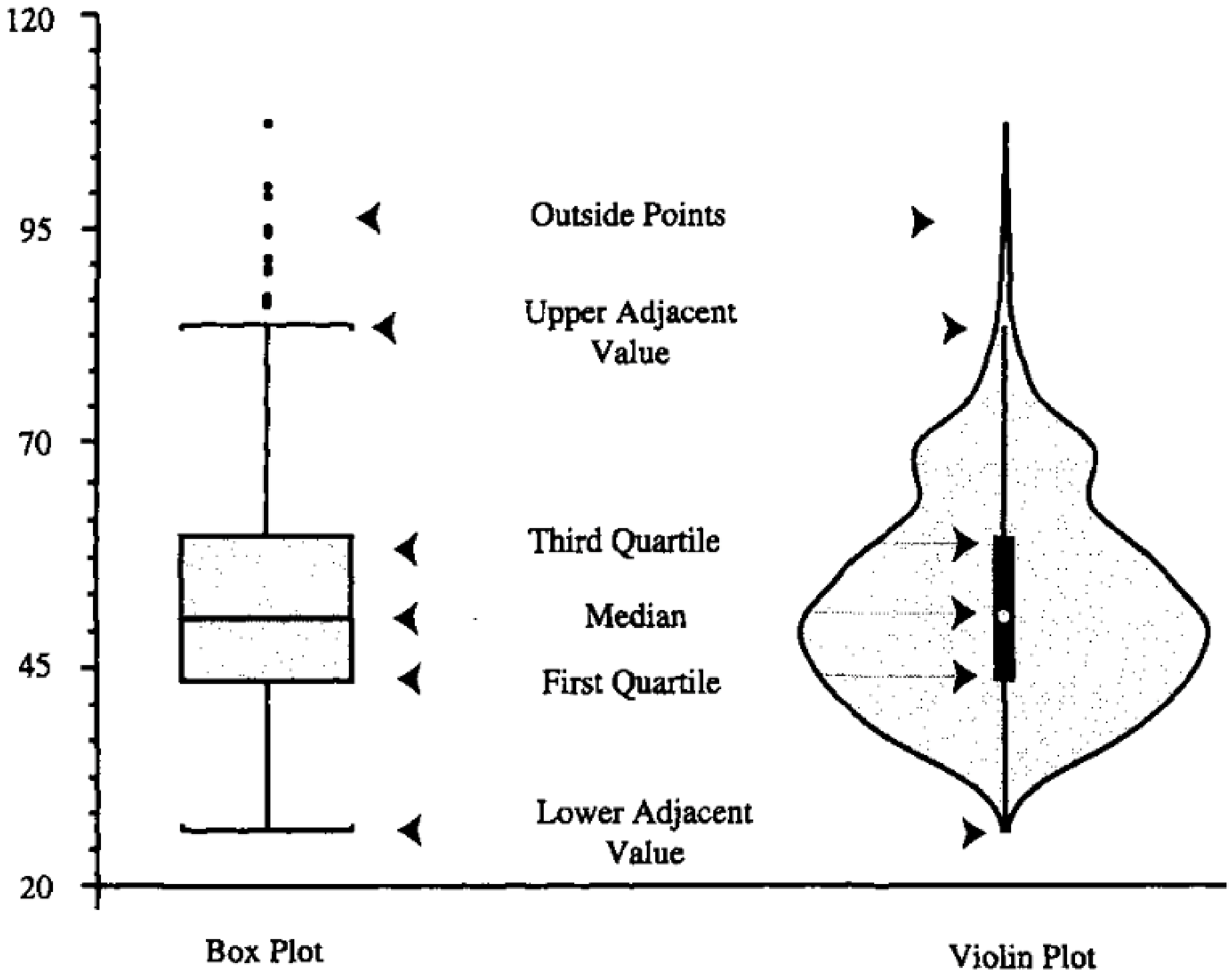
J. Matejka and G. Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, 2017, doi: [10.1145/3025453.3025912](https://doi.org/10.1145/3025453.3025912).

Violin Plot

Aggregate

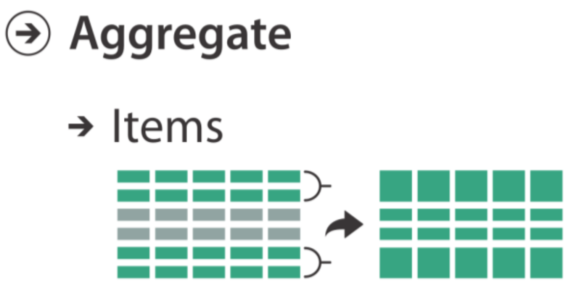
→ Items

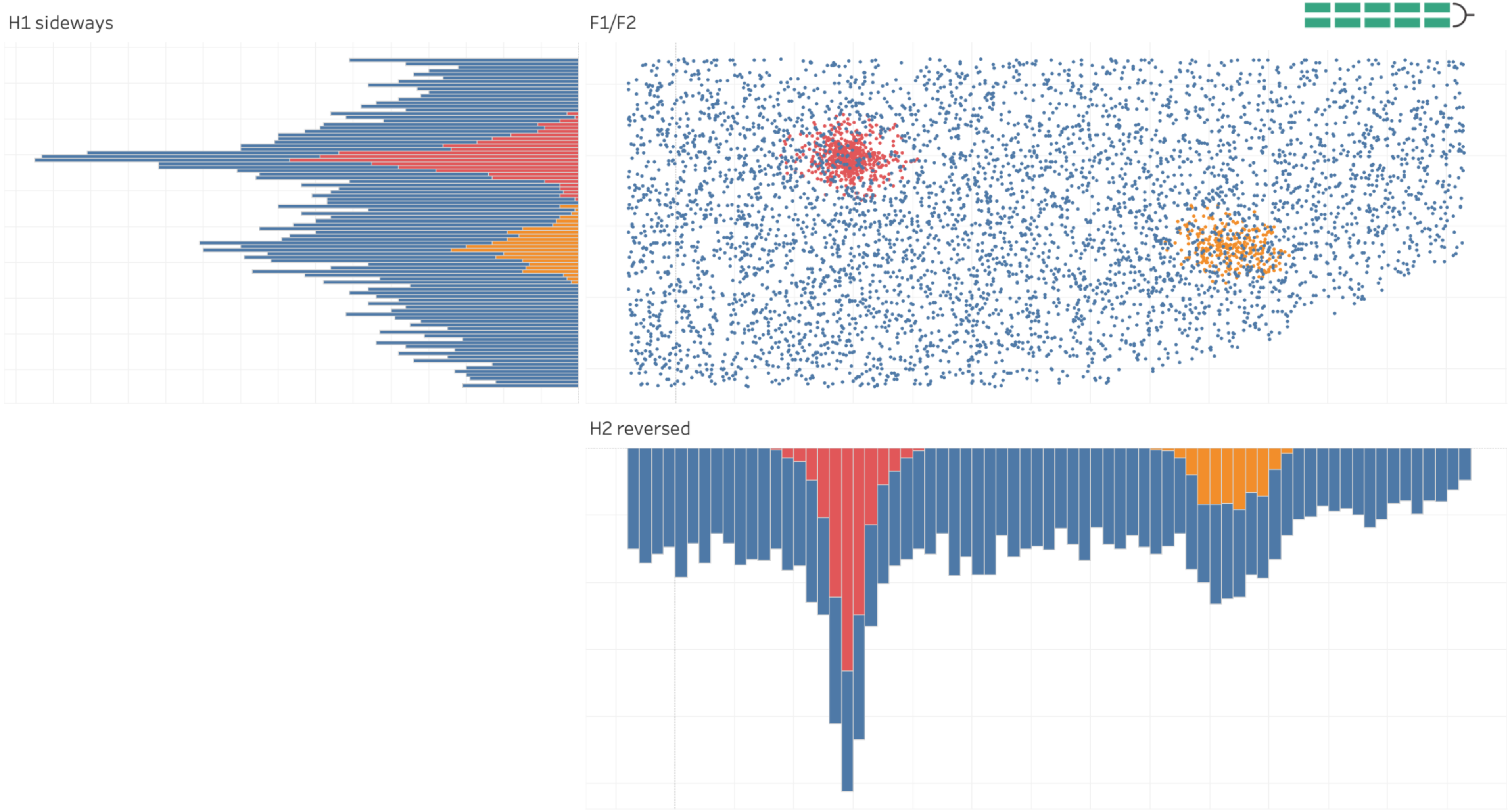




J. L. Hintze and R.D. Nelson, Violin Plots: A Box Plot-Density Trace Synergism, 1998, doi: [10.1080/00031305.1998.10480559](https://doi.org/10.1080/00031305.1998.10480559).

Histogram + Scatterplot

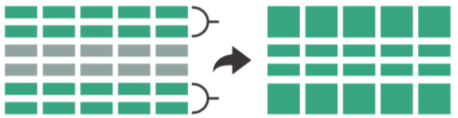


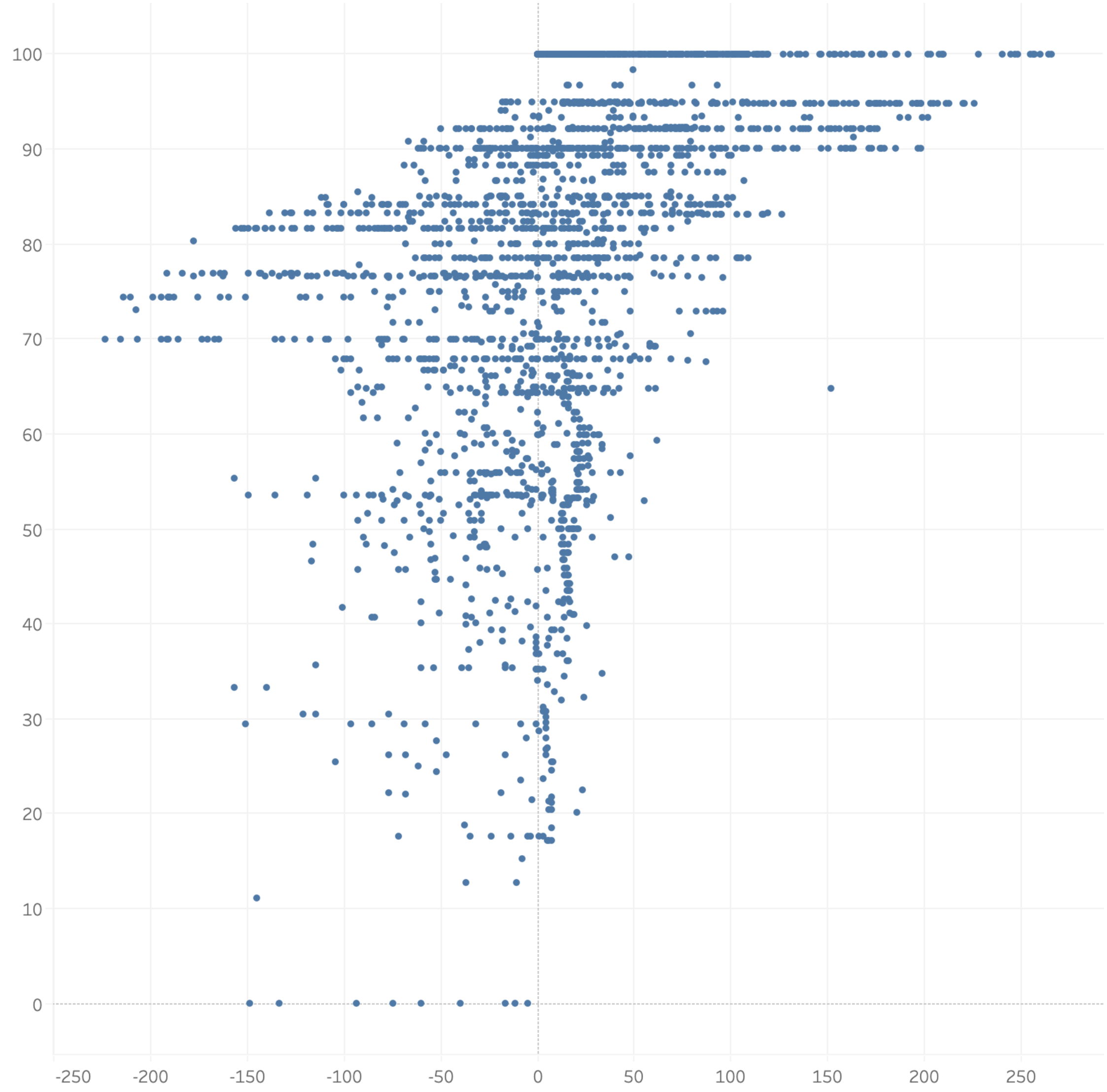


Overplotting in Scatterplots

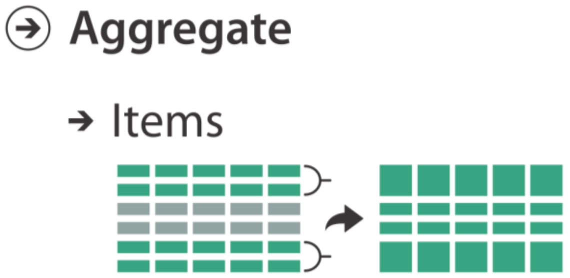
Aggregate

Items

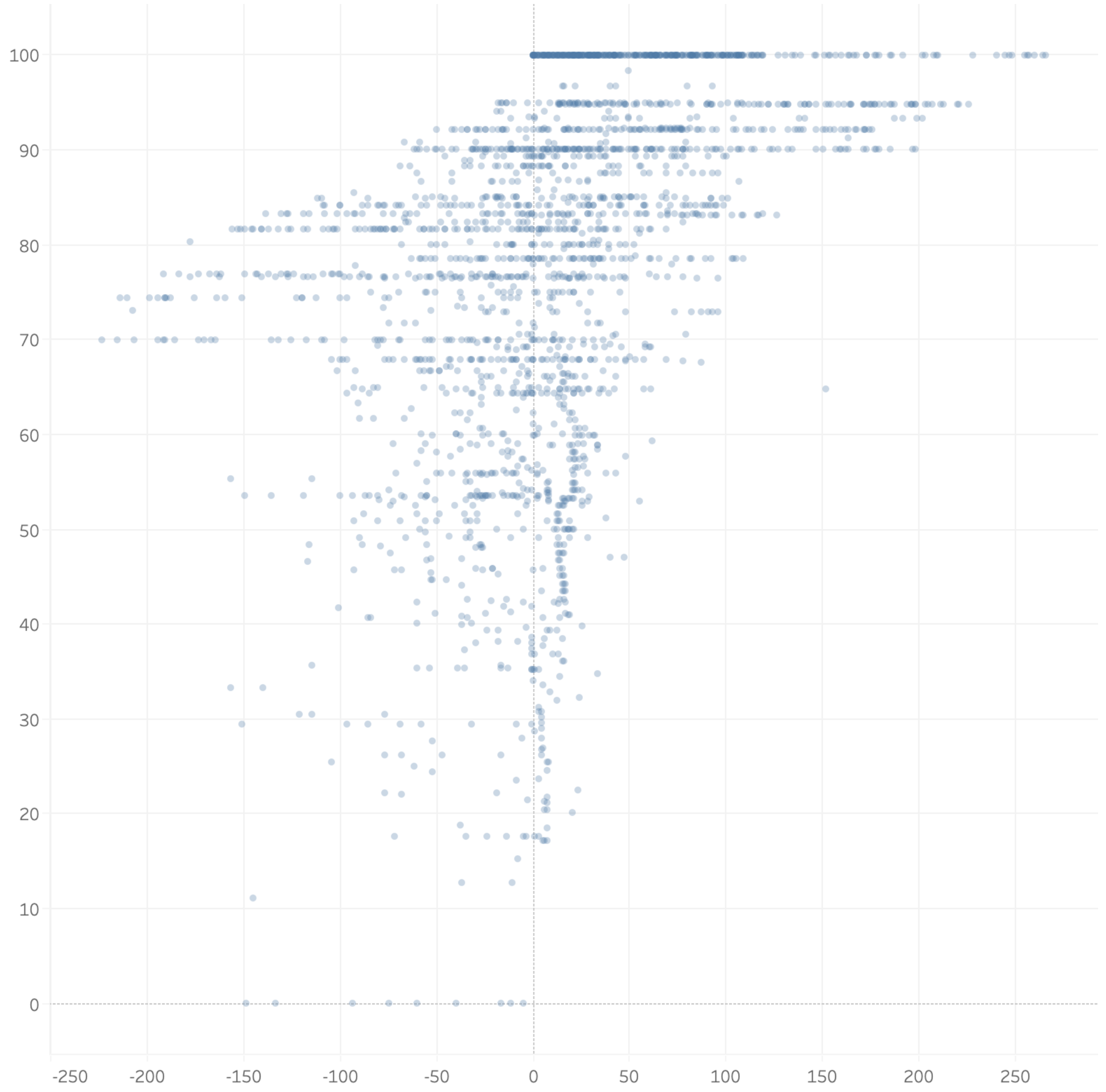




Overplotting in Scatterplots



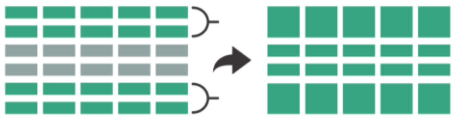
Use lower opacity

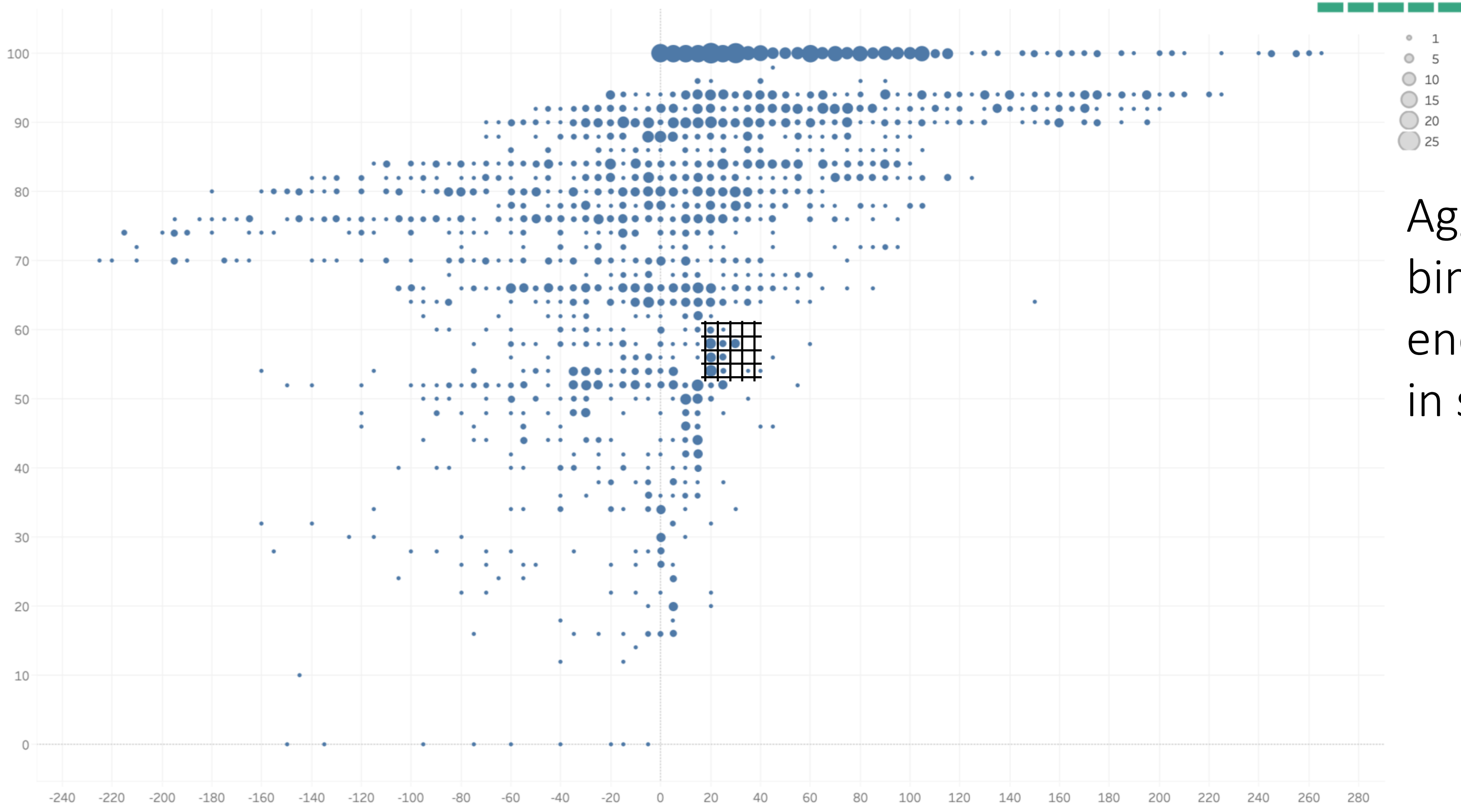


Overplotting in Scatterplots

➔ **Aggregate**

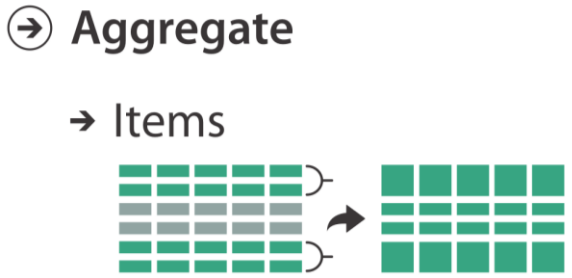
➔ Items

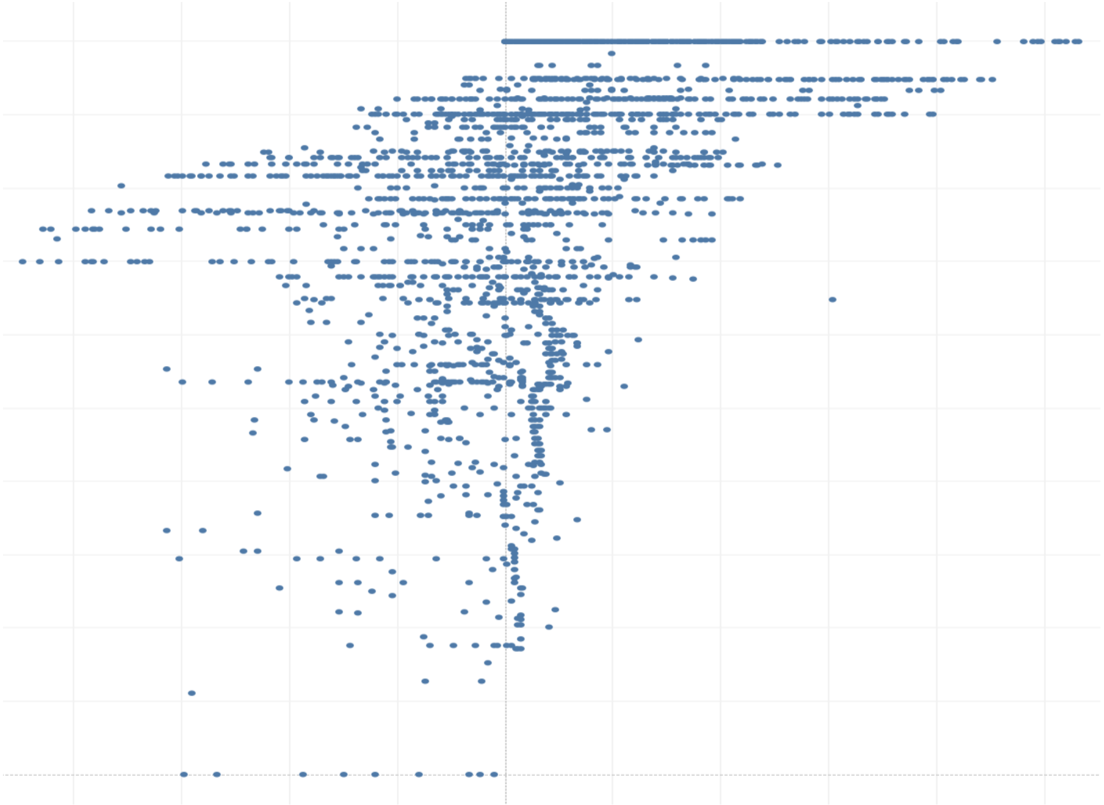


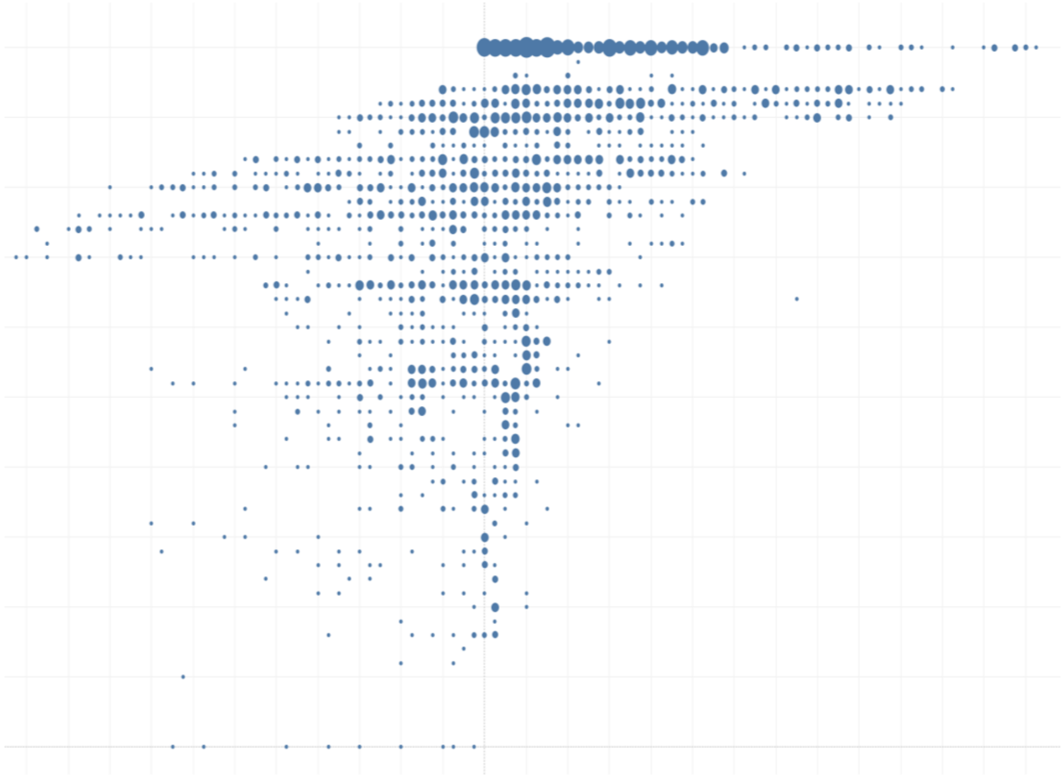


Aggregate into bins, encode count in size

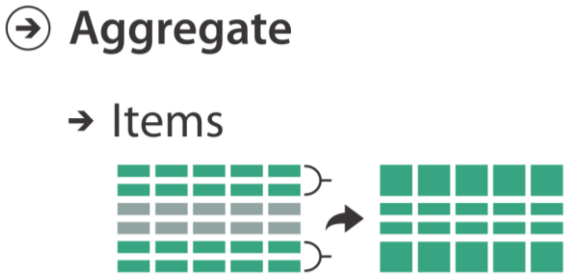
Overplotting in Scatterplots





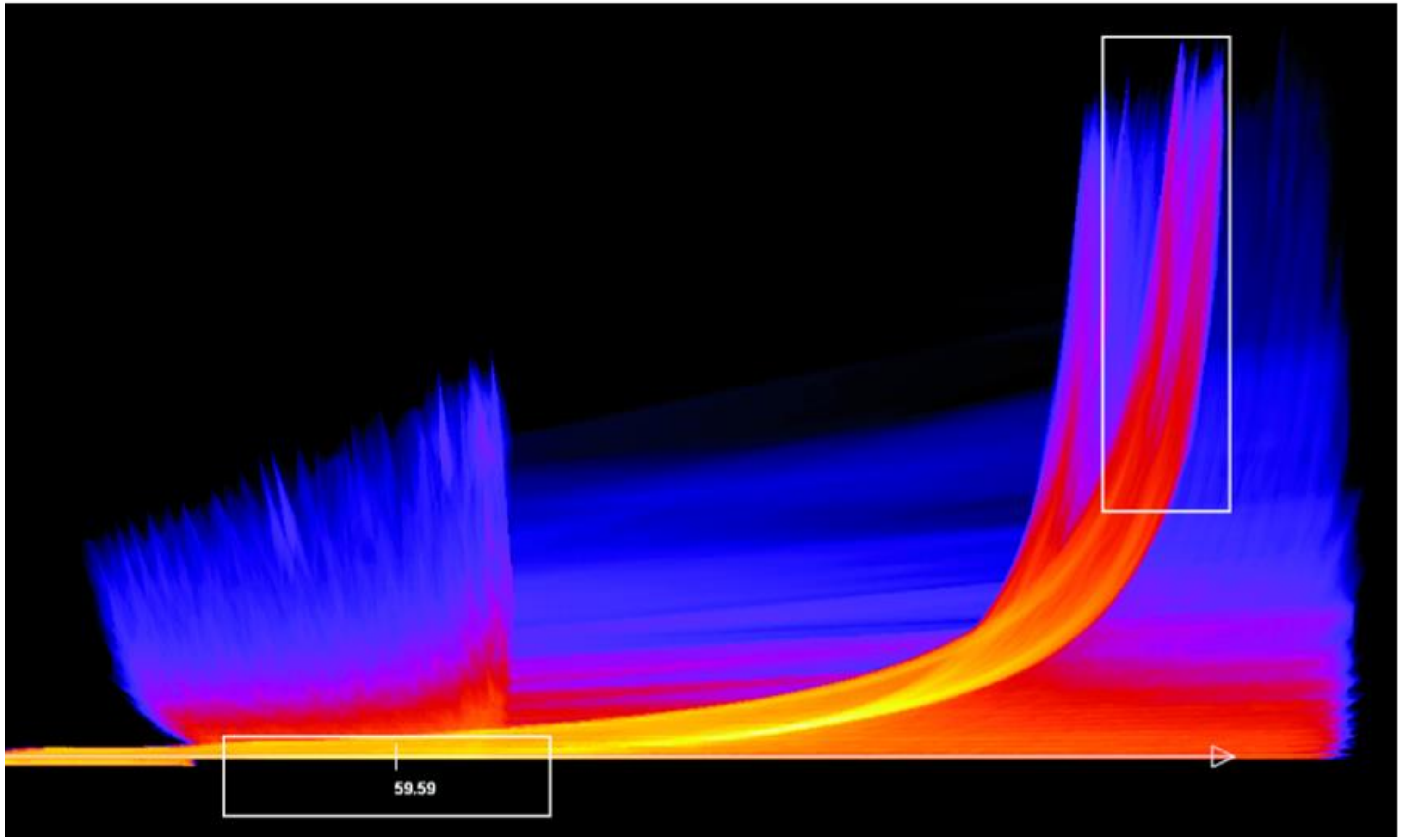


Continuous scatterplots



Derived data - new table:

- Quant. attribute: Overplot density
- Dense, space-filling, 2D matrix
- Color encodes overplot density

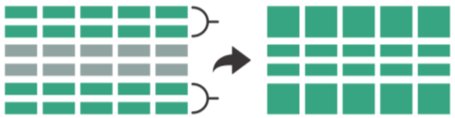


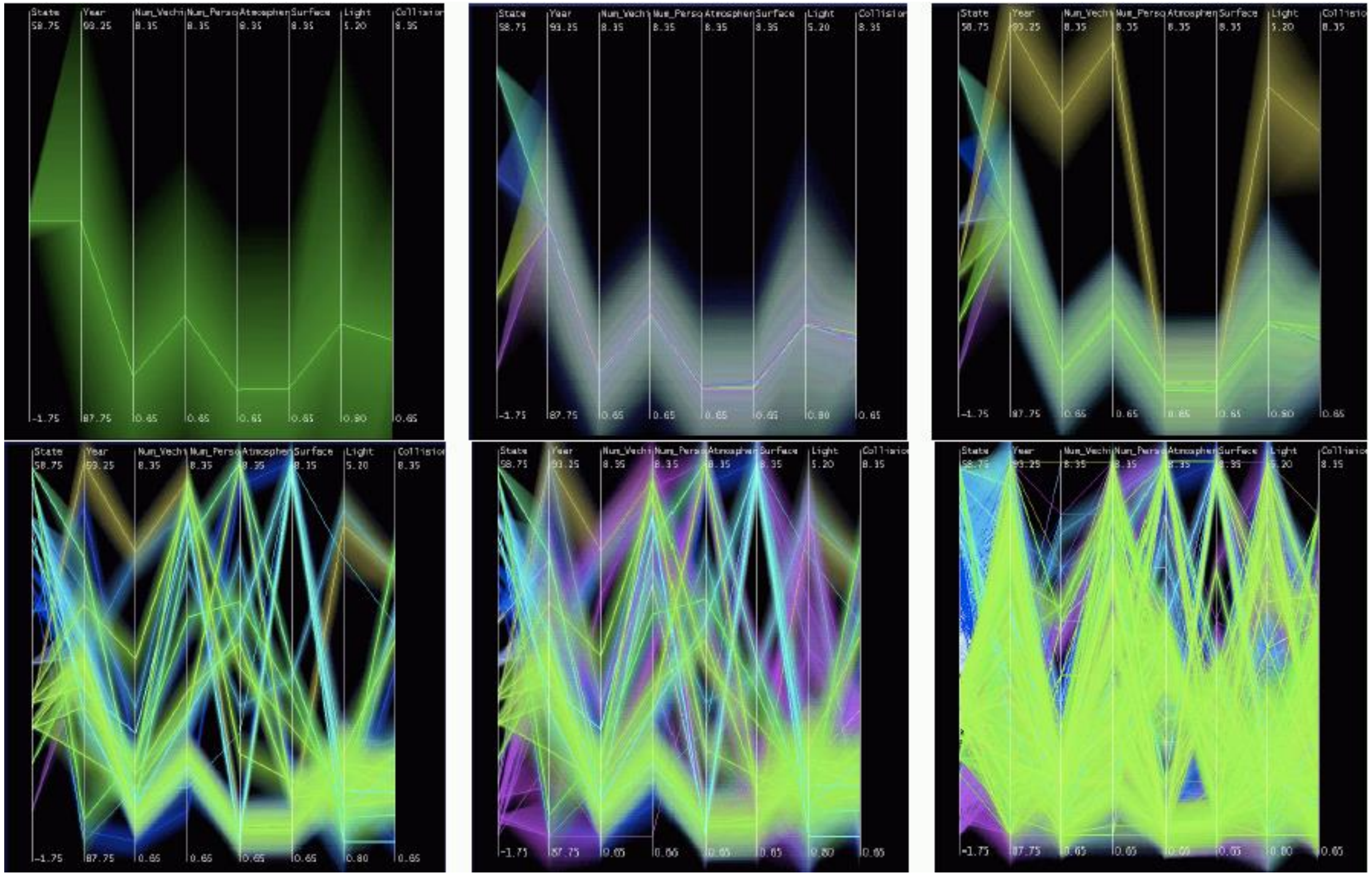
S. Bachthaler and D. Weiskopf, Continuous Scatterplots, 2008, doi: [10.1109/TVCG.2008.119](https://doi.org/10.1109/TVCG.2008.119).

Hierarchical Parallel Coordinates

➔ **Aggregate**

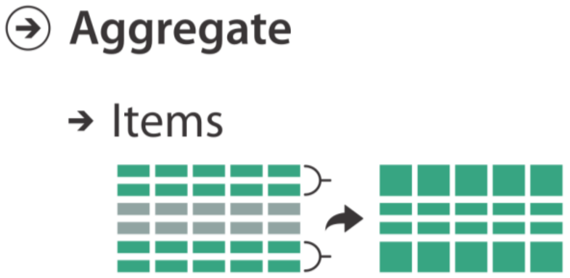
→ Items





Y.-H. Fua et al., Hierarchical Parallel Coordinates for Exploration of Large Datasets, 1999, doi: [10.1109/VISUAL.1999.809866](https://doi.org/10.1109/VISUAL.1999.809866).

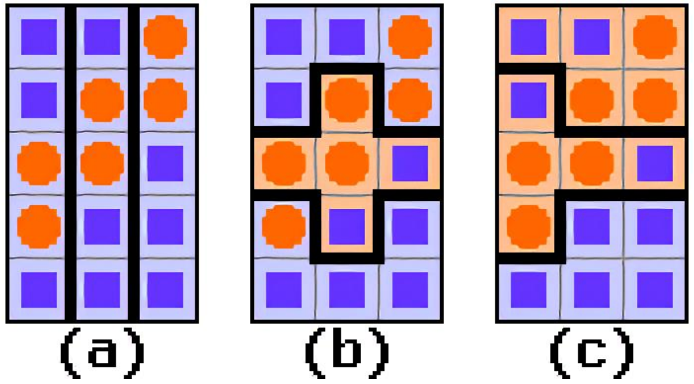
Spatial Aggregation

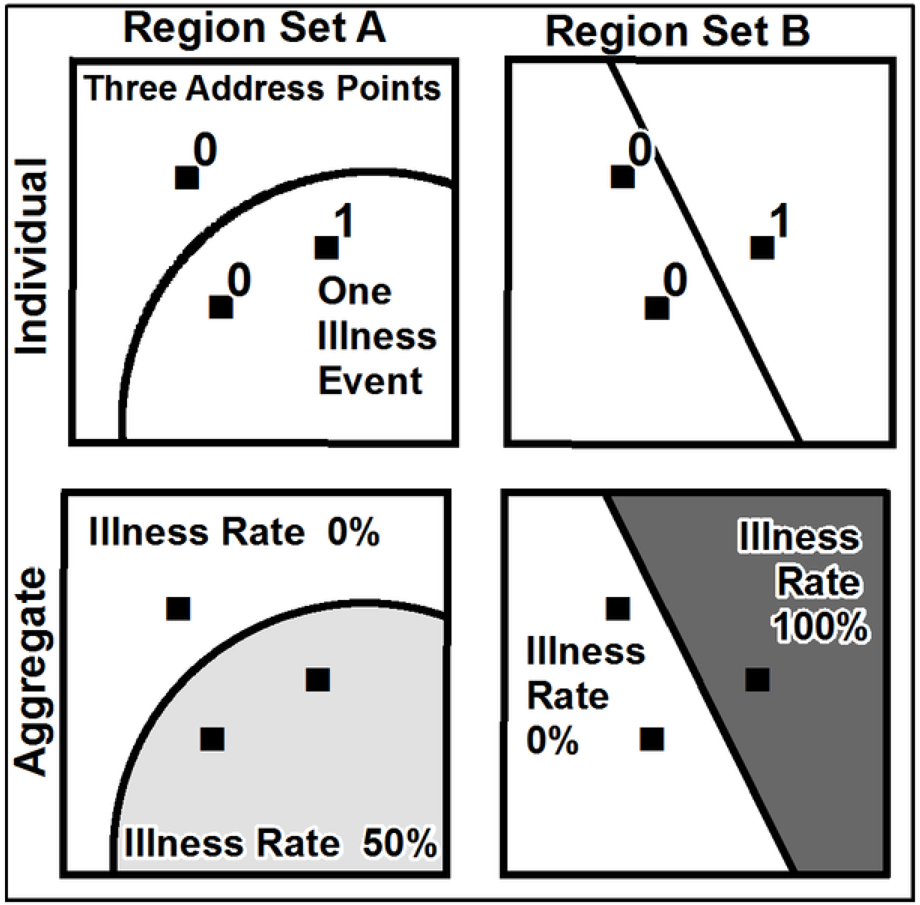


Modifiable Areal Unit Problem

Question: Can you think of implications this might have?

Gerrymandering & its influence on electoral results in a non-proportional system:

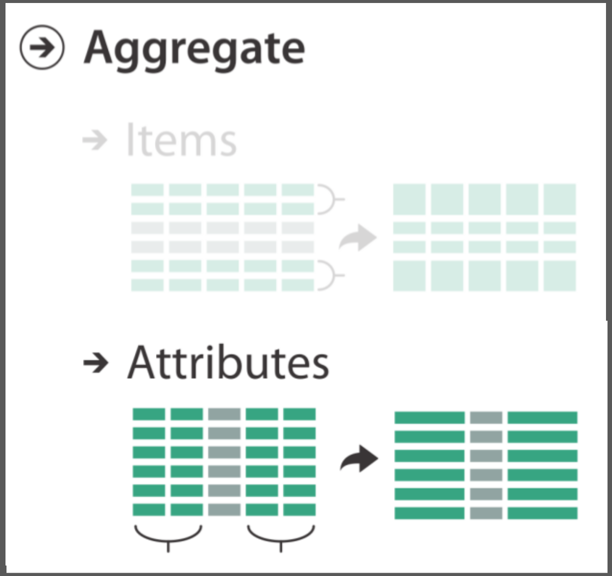




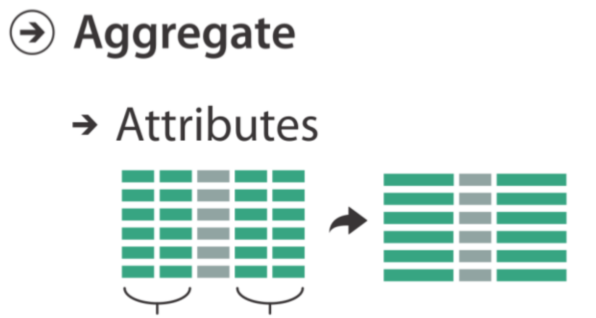
Filtering
Items
Attributes

Aggregation
Items
Attributes - Dimensionality Reduction
Linear
Non-linear

Attribute Aggregation – Dimensionality Reduction



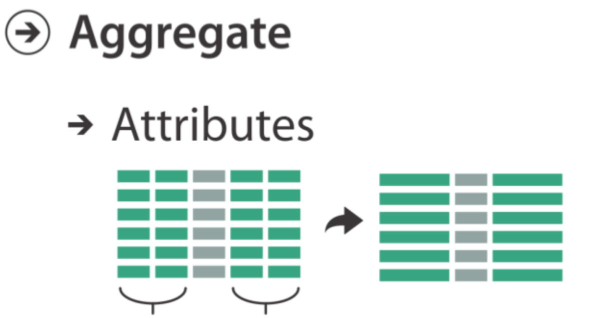
Dimensionality vs. attribute



- Vocabulary used in field not consistent
  - Dimension/Attribute
  - Attribute aggregation ~ attribute synthesis ~ **Dimensionality Reduction (DR)**

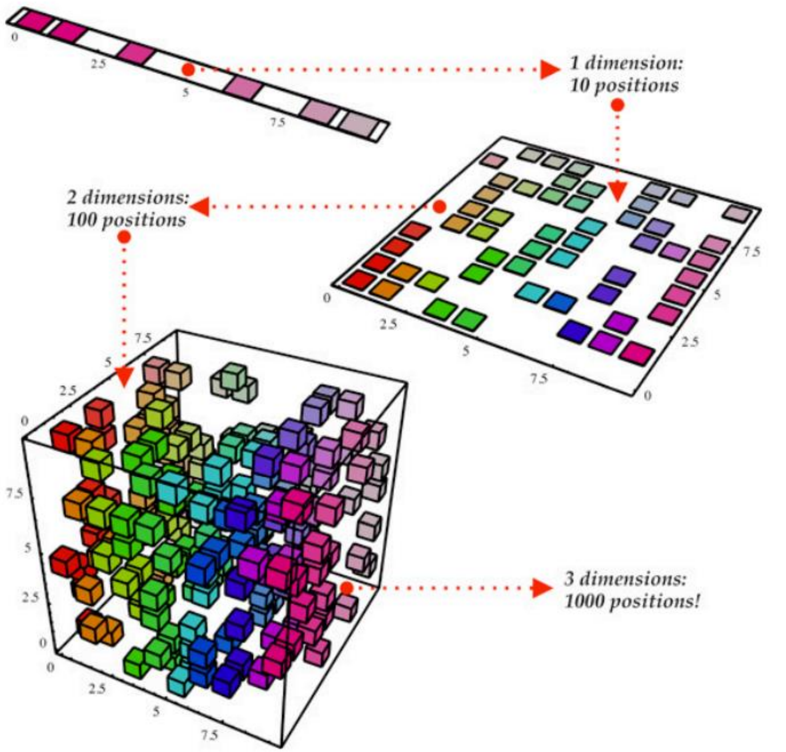
- Create smaller set of **new dimensions:**
  - Set size is smaller than original
  - New dimensions completely synthetic

When to use DR



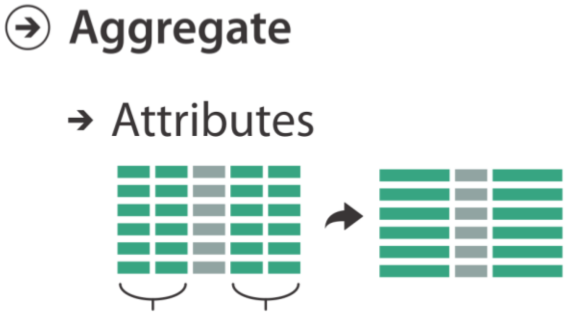
- Uncovering Hidden Structure
- Measurements made in sprawling space
  - Documents, images
- DR only suitable if (almost) all information could be conveyed with fewer dimensions
  - How do you know?

  →Need to estimate true dimensionality to check if different than original

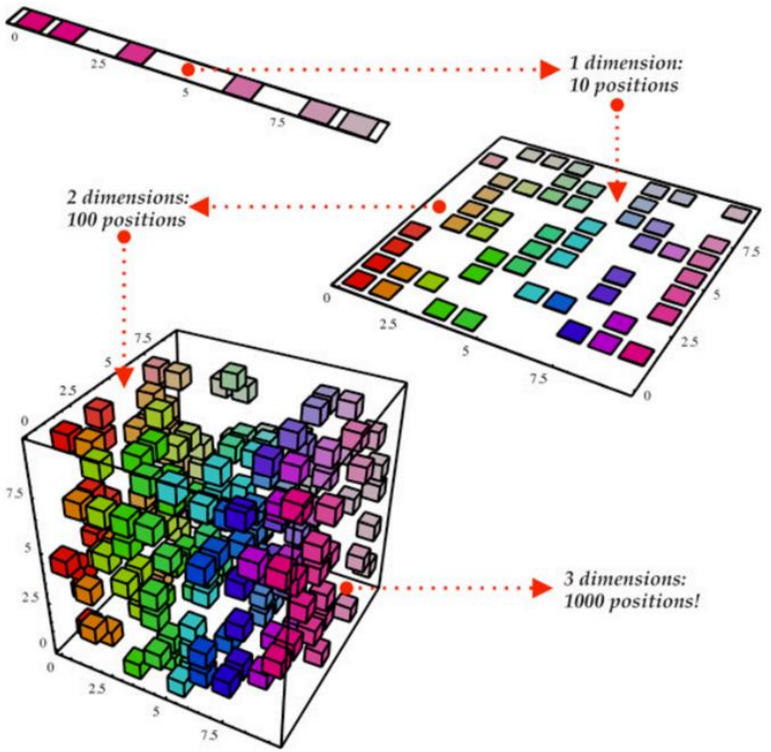


https://www.pinecone.io/learn/dimensionality-reduction/

Dimensionality Reduction (DR)

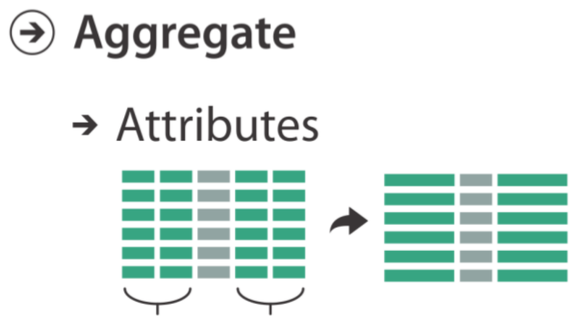


- Mapping multidimensional space into space of fewer
  - Typically 2D for clarity
  - 1D/3D/higher-D possible
  - Keep/explain as much variance as possible
  - Show underlying dataset structure
- Motivation:
  - True/intrinsic dimension of dataset is much lower than measured dim
- Linear vs. non-linear approaches



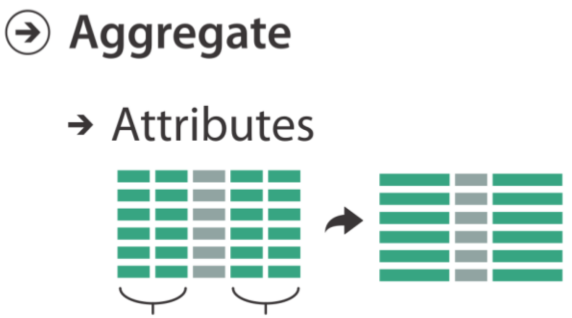
https://www.pinecone.io/learn/dimensionality-reduction/

Estimating True Dimensionality

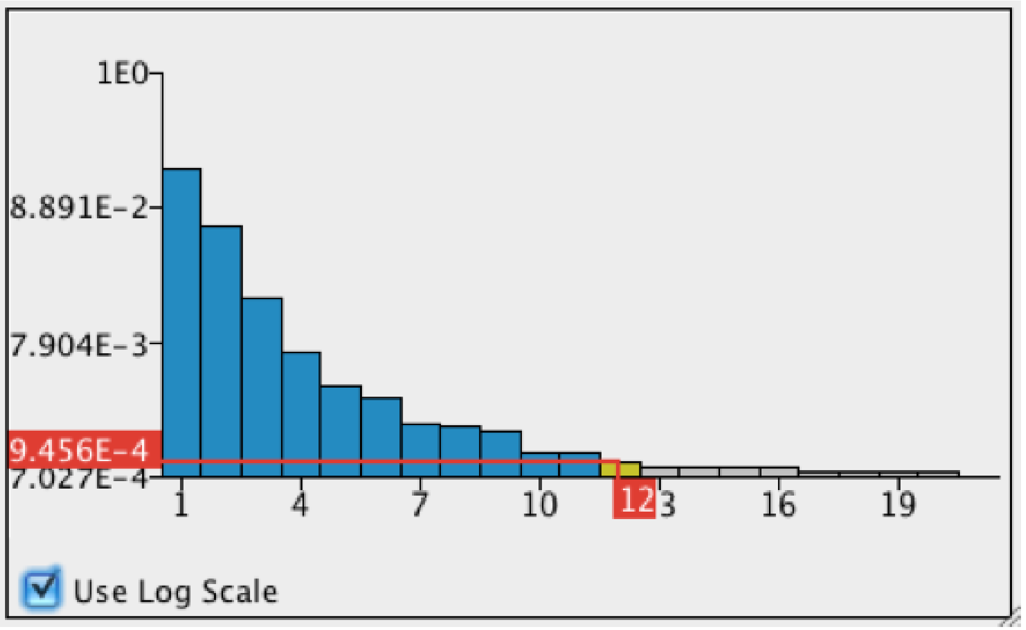


- Error for low-dim projection vs. high-dim original
- No single correct answer; many metrics proposed
  - Cumulative variance that is not accounted for
  - Match variations in distance (vs. actual distance values)
  - Difference between inter-point distances in high and low dimensions

Showing Dimensionality Estimates

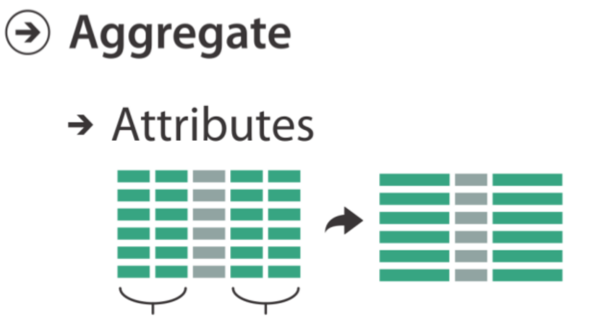


- Scree plots as simple way: error against # dims
  - Original dataset: 294 dims
  - Estimate: almost all variance preserved with < 20 dims



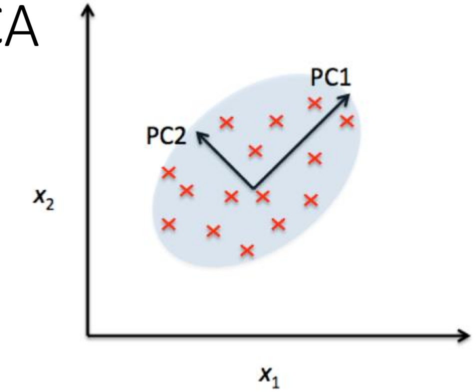
S. Ingram et al., DimStiller: Workflows for dimensional analysis and reduction, 2010, doi: [10.1109/VAST.2010.5652392](https://doi.org/10.1109/VAST.2010.5652392).

Linear Dimensionality Reduction

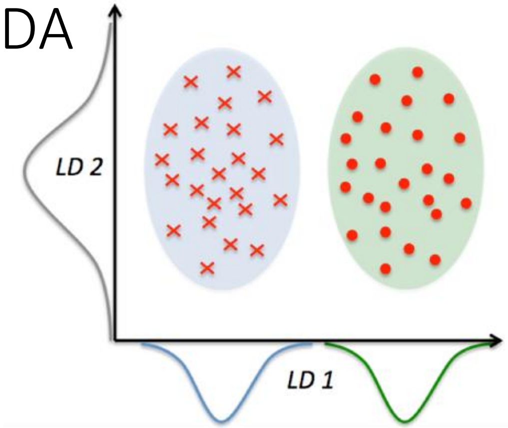


- Based on linear projections
- Given dimensions have a strong meaning
- Want to preserve linearity in layout
- Examples:
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Some variants of Multi-Dimensional Scaling (MDS)
  - etc.

PCA

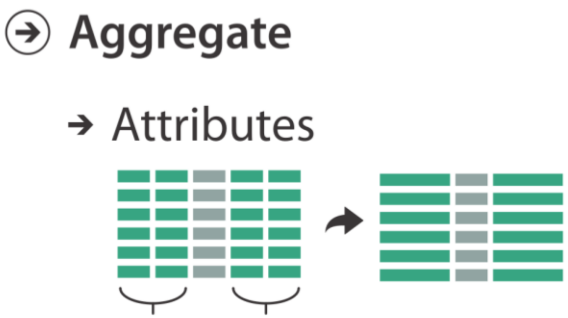


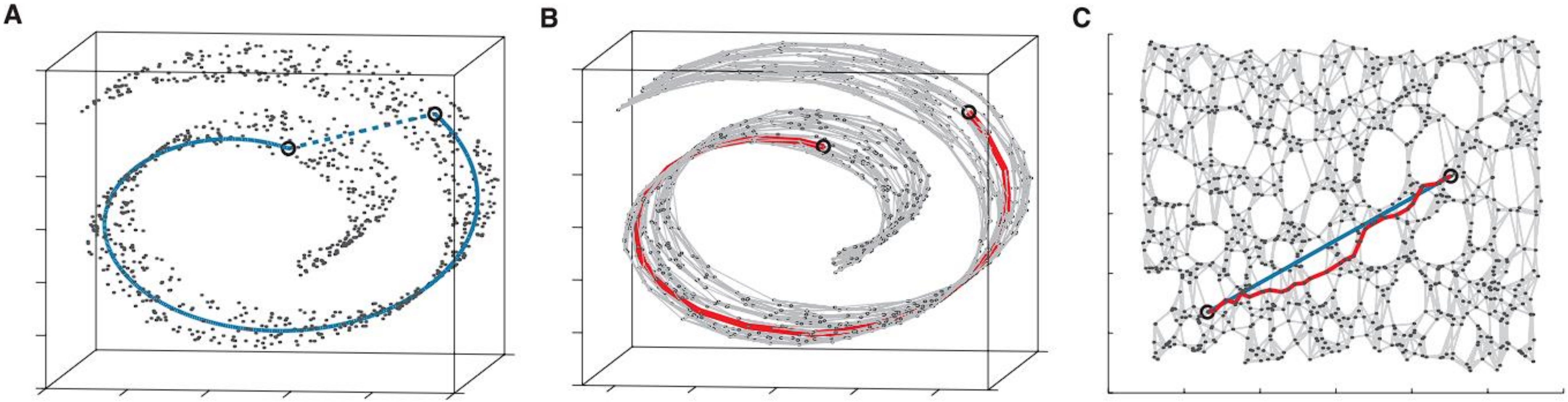
LDA



https://sebastianraschka.com/faq/docs/lda-vs-pca.html

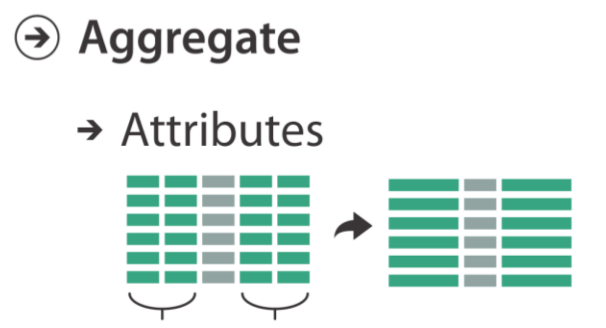
Problem for linear DR



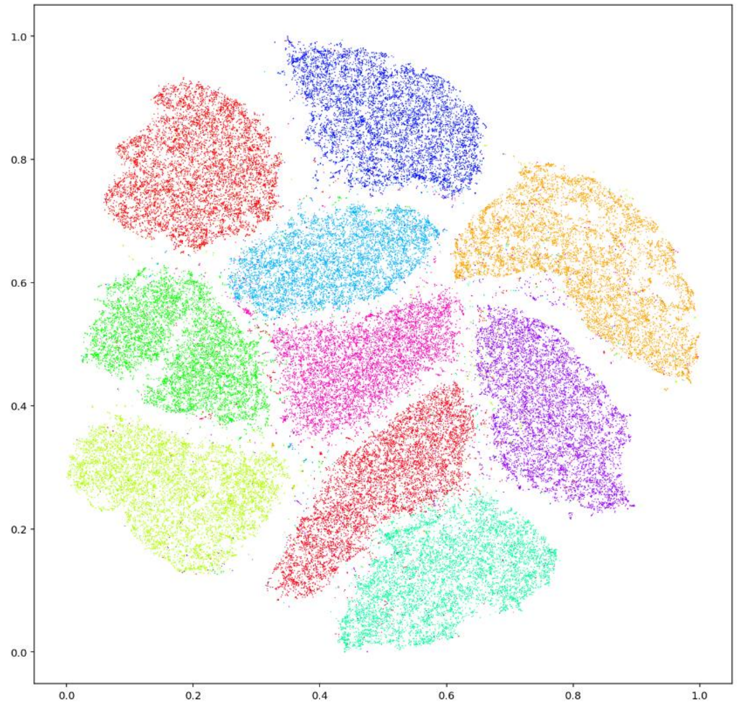


J. B. Tenenbaum et al., A Global Geometric Framework for Nonlinear Dimensionality Reduction, 2000, doi: [10.1126/science.290.5500.2319](https://doi.org/10.1126/science.290.5500.2319).

Nonlinear DR



- No inherent meaning to given dimensions
- Minimize differences between interpoint distances in high and low dimensions
- Examples:
  - Some variants of Multi-Dimensional Scaling (MDS)
  - Isomap
  - Local Linear Embedding (LLE)
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)



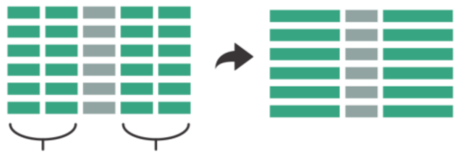
https://tinyurl.com/3kjmbjyn

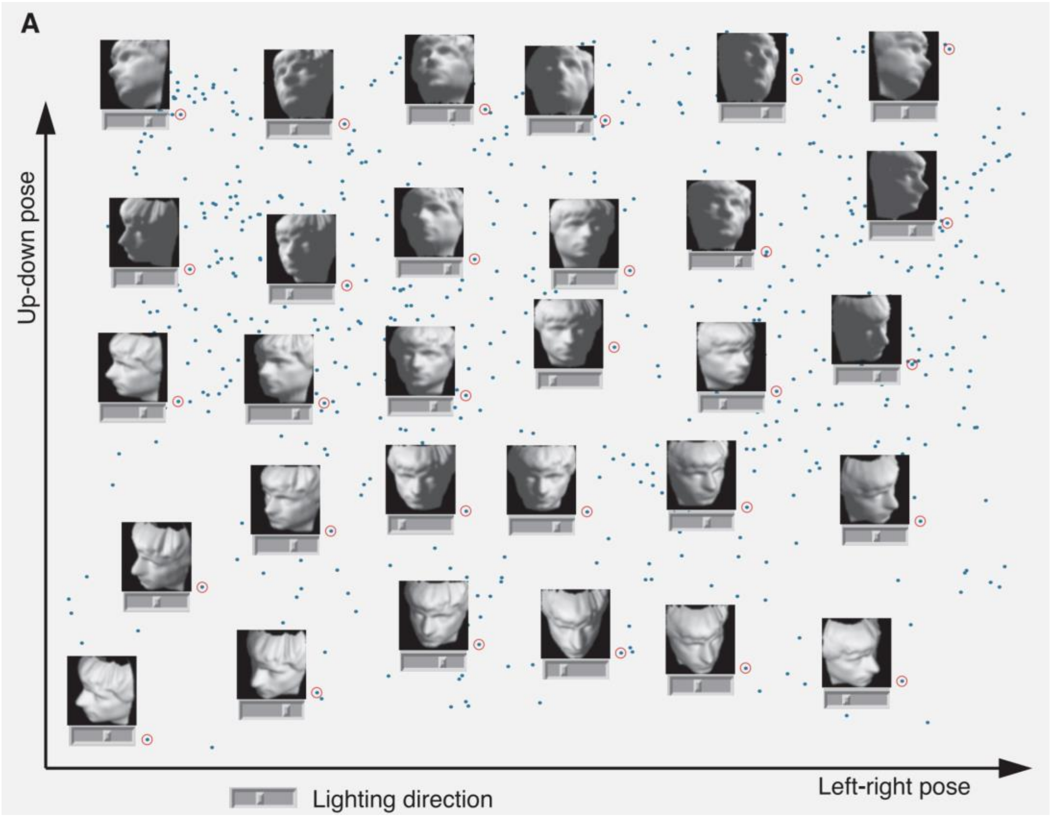
DR: Isomap

- Head poses, 4096 D vector
- Reduced to 2D
- Dimensions correlate highly with one degree of freedom of underlying original data:
  - Up-down pose
  - Left-right pose

⊕ Aggregate

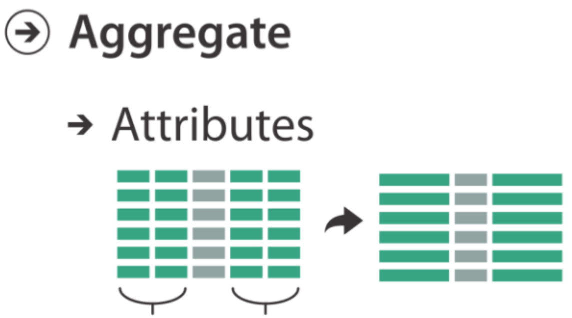
→ Attributes





J. B. Tenenbaum et al., A Global Geometric Framework for Nonlinear Dimensionality Reduction, 2000, doi: 10.1126/science.290.5500.2319.

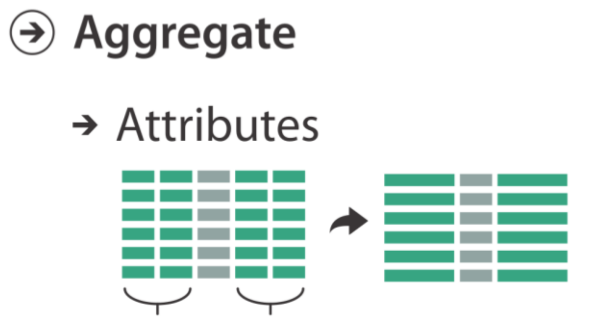
DR: Goals / Tasks



- Interpret new synthetic dimensions
  - Keep/explain as much variance as possible
- Find clusters
  - Or compare / evaluate with previous clustering
- Understand structure
  - Absolute position not reliable
  - Fine grained structure not reliable

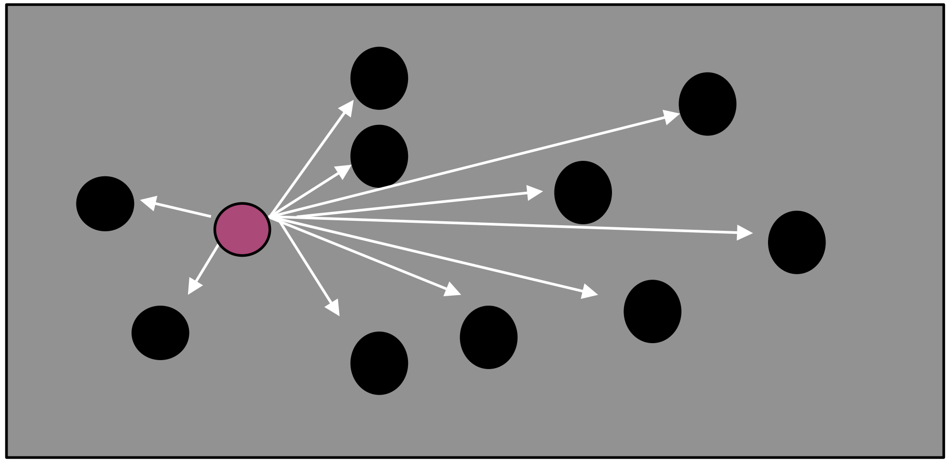
M. Brehmer et al., Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences,2014, doi: 10.1145/2669557.2669559.

DR: MDS – Iterative Spring Model

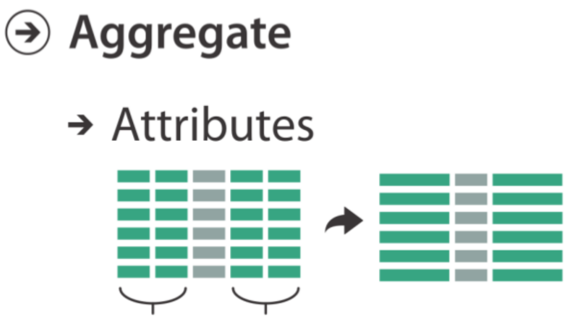


Naive Spring Model

- Repeat for all points
  - Compute spring force to all other points
    - Difference between high dim vs. low dim distance (stress)
  - Move to better location using computed forces
- Compute distances between all points
  - O(n2) iteration, O(n3) algorithm

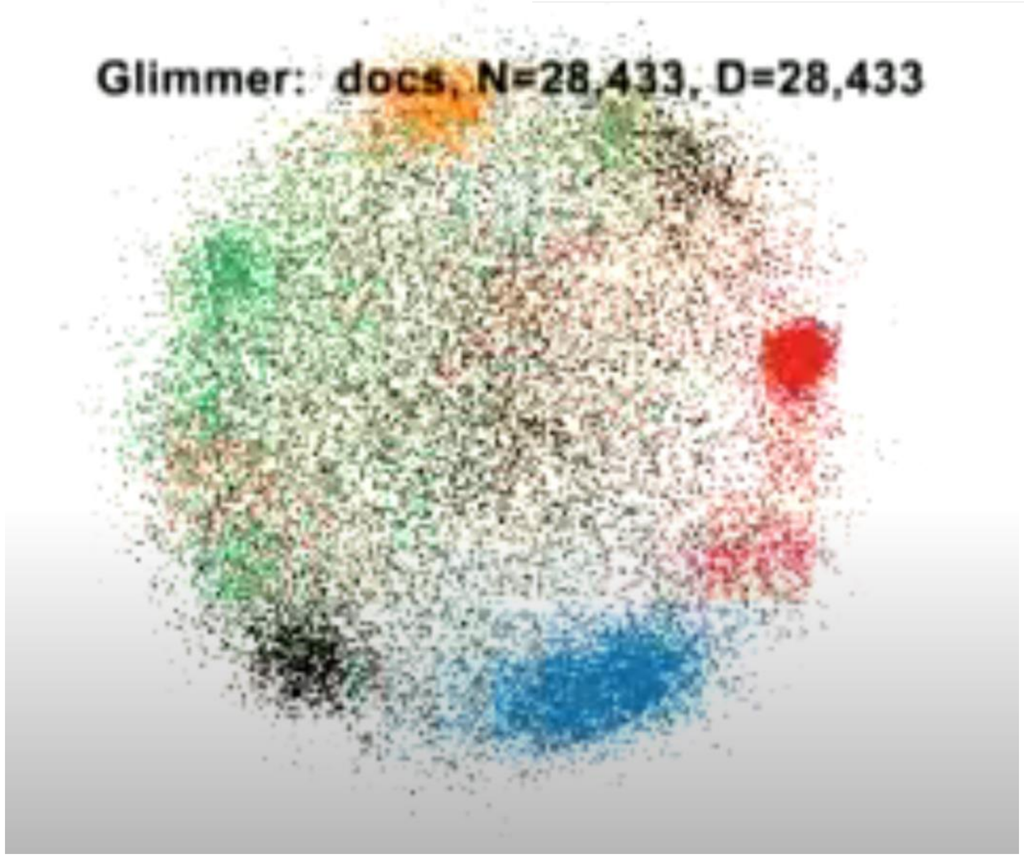


Multi-Dimensional Scaling (MDS)



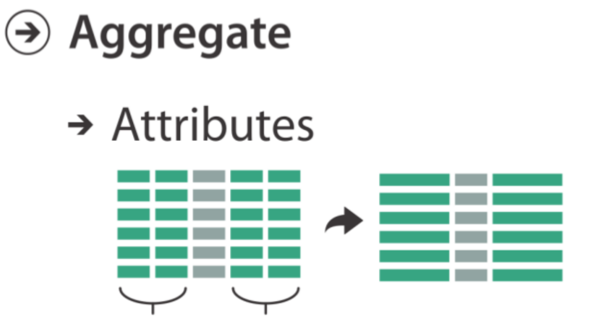
- Many different algorithms exist
- Example implementation:

  Glimmer



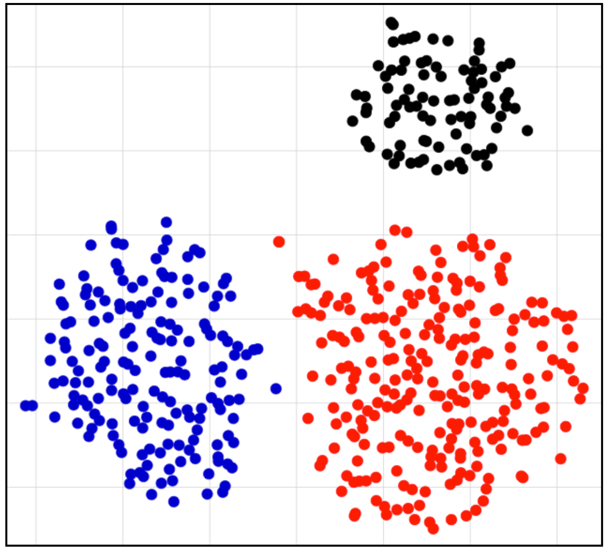
S. Ingram et al., Glimmer: Multilevel MDS on the GPU, 2009, doi: 10.1109/TVCG.2008.85.

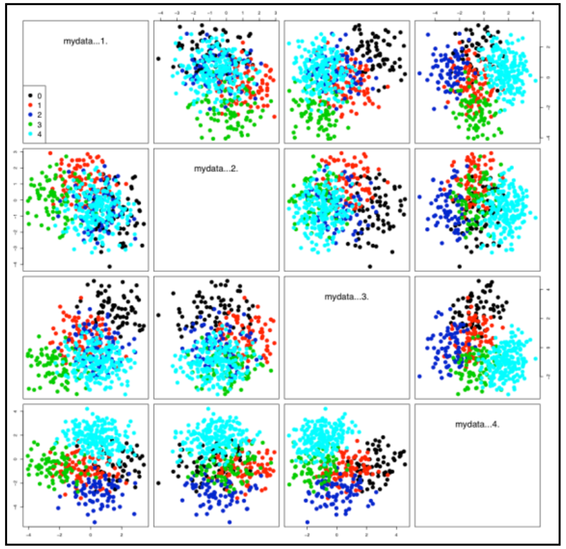
Showing DR data



- Scatterplot showing points
  - Only works if true dimensionality is 2 (... or 3)
  - Need to drill down to see what points represent

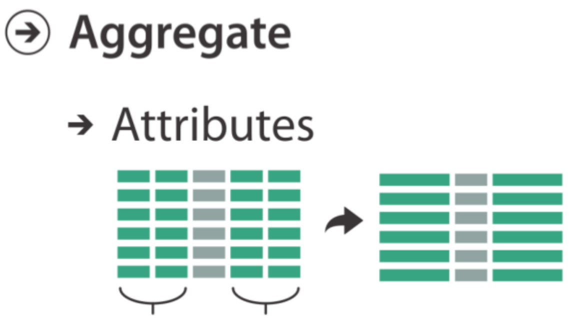
- Scatterplot matrix
  - Safe choice





M. Sedlmair et al., Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices, 2013, doi: [10.1109/TVCG.2013.153](https://doi.org/10.1109/TVCG.2013.153).

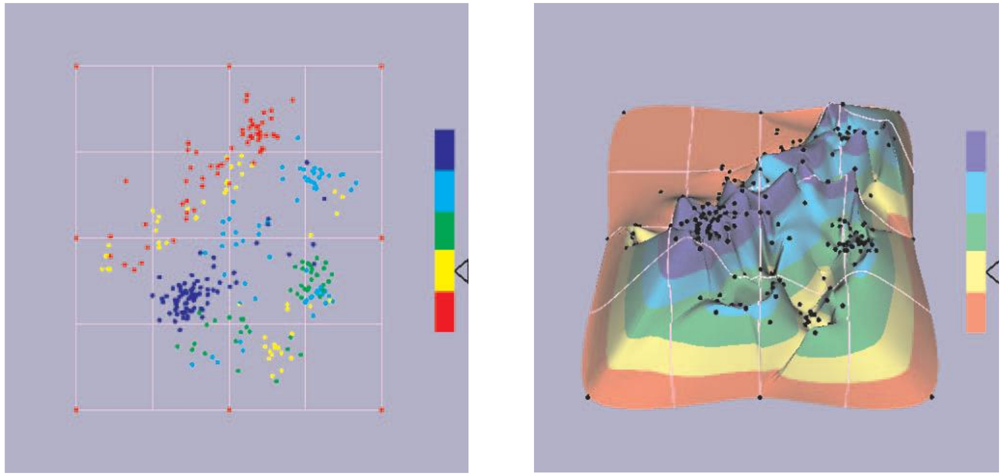
How **not** to show DR data



- 3D scatterplots?
  - Avoid! studies show many perceptual drawbacks, little benefits for DR data (clustering tasks)

M. Sedlmair et al., Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices, 2013, doi: 10.1109/TVCG.2013.153.

- Landscapes?
  - Avoid! studies show worse than just using points

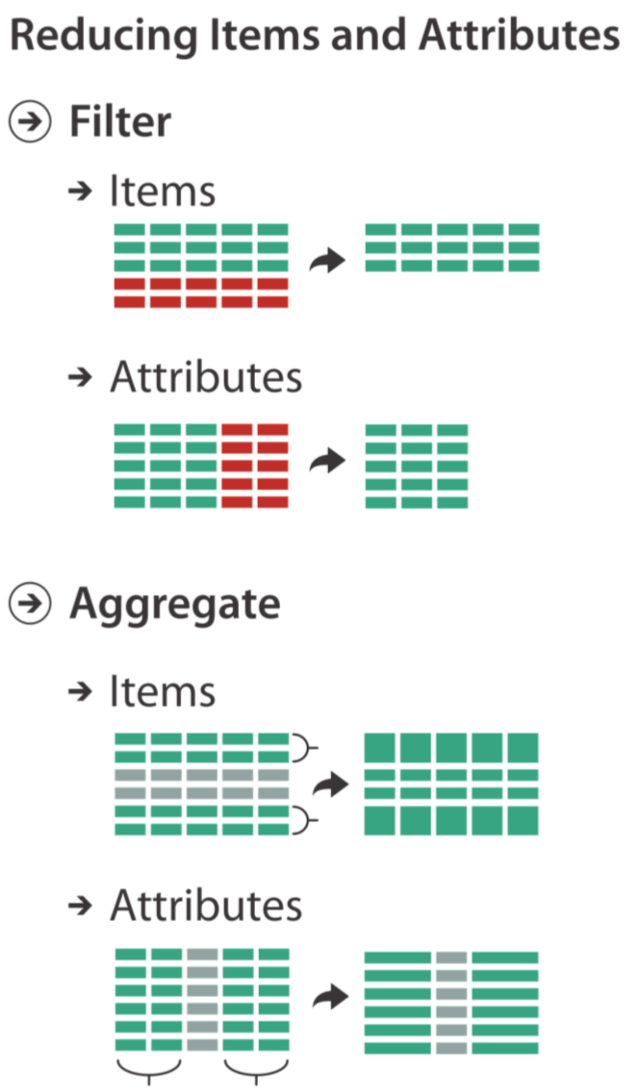


M. Tory et al., Spatialization Design: Comparing Points and Landscapes, 2007, doi: 10.1109/TVCG.2007.70596.

In conclusion

- Reduction can be achieved via **Filtering** or **Aggregation**
- Both filtering and aggregation can be done for
  - Items (Statistical / derived measures), or
  - Attributes (Dimensionality Reduction)
- Consider filter & aggregation pitfalls, tasks and usability!

Questions?



Reduce Items & Attributes

Visual and Exploratory Data Analysis 23S

06.06.2023

Timothée Schmude