Information Visualization
Reduce: Aggregation & Filtering
Project Peer Reviews

Tamara Munzner
Department of Computer Science
University of British Columbia

Week 11, 17 Nov 2021

https://www.cs.ubc.ca/~tmm/courses/547-21
Today

• first: project peer reviews
  – join your matched teams
    • you've already read other team's written update
      – let me know by private Piazza post if your counterpart(s) weren't prepared
    – record discussion/thoughts in gdoc (freeform)
  – first A critiques B; then B critiques A

• break

• Q&A overflow from before
  – Ch 11, Interact, cont
  – Ch 12, Multiple Views

• Q&A / mini-lecture this time
  – Ch 13, Reduce
Peer reviews

• rough structure (adapt as you like, aim for ~45-60 min)
  – talk through initial thoughts when read updates
  – ask clarifying questions
  – get demo to see look/feel & any interaction
  – discuss tradeoffs, design choices, suggestions
  – when conversation winds down, critiquers record braindump (if not done as you go)
  – write DONE at top of your gdoc section & switch!

• tips on giving feedback
  – state what you think is good about the work, and why you think so
  – state what you think needs improvement, including why/rationale
  – offer specific suggestions on how to improve it, as followup
  – keep your feedback focused on the work, not the person who did it
Upcoming

• next week (W12)
  – async: last week of readings / discussion (light, 2 readings)
    • Ch 14: Embed - Focus+Context
      [type: design study]
    – in class: post-update meetings with Tamara
      • oral feedback on project progress, after I've read them

• last week of classes (W13)
  – async: no readings/discussion
  – in class: evals
  – in class: Q&A wrapup (W12)
  – in class: lecture on research process and final writeup expectations
Q&A / Backup Slides
Visualization Analysis & Design

Reduce: Aggregation & Filtering (Ch 13)

Tamara Munzner
Department of Computer Science
University of British Columbia
@tamaramunzner
How to handle complexity: 3 previous strategies

Derive

- derive new data to show within view
- change view over time
- facet across multiple views

Manipulate

- Change
- Select
- Navigate

Facet

- Juxtapose
- Partition
- Superimpose

Actions

- derive new data to show within view
- change view over time
- facet across multiple views
How to handle complexity: 3 previous strategies + 1 more

- Derive
  - derive new data to show within view
  - change view over time
  - facet across multiple views
  - reduce items/attributes within single view

- Manipulate
  - Change
  - Select
  - Navigate

- Facet
  - Juxtapose
  - Partition
  - Superimpose

- Reduce
  - Filter
  - Aggregate
  - Embed
Reduce items and attributes

• reduce/increase: inverses
• filter
  – pro: straightforward and intuitive
    • to understand and compute
  – con: out of sight, out of mind
Reduce items and attributes

• reduce/increase: inverses

• filter
  – pro: straightforward and intuitive
    • to understand and compute
  – con: out of sight, out of mind

• aggregation
  – pro: inform about whole set
  – con: difficult to avoid losing signal

• not mutually exclusive
  – combine filter, aggregate
  – combine reduce, change, facet
Filter

• eliminate some elements
  – either items or attributes

• according to what?
  – any possible function that partitions dataset into two sets
    • attribute values bigger/smaller than x
    • noise/signal

• filters vs queries
  – query: start with nothing, add in elements
  – filters: start with everything, remove elements
  – best approach depends on dataset size
Idiom: FilmFinder

- dynamic queries/filters for items
  - tightly coupled interaction and visual encoding idioms, so user can immediately see results of action

Idiom: **cross filtering**

- item filtering
- coordinated views/controls combined
  - all scented histogram bisliders update when any ranges change

**System: Crossfilter**

- [http://square.github.io/crossfilter/](http://square.github.io/crossfilter/)
- [https://observablehq.com/@uwdata/interaction](https://observablehq.com/@uwdata/interaction)
Aggregate

• a group of elements is represented by a smaller number of derived elements

- Aggregate
  - Items
  - Attributes
Idiom: histogram

- static item aggregation
- task: find distribution
- data: table
- derived data
  - new table: keys are bins, values are counts
- bin size crucial
  - pattern can change dramatically depending on discretization
  - opportunity for interaction: control bin size on the fly
Idiom: **scented widgets**

- augmented widgets show *information scent*
  - better cues for *information foraging*: show whether value in drilling down further vs looking elsewhere
- concise use of space: histogram on slider
Idiom: scented widgets

• augmented widgets show information scent
  –better cues for information foraging: show whether value in drilling down further vs looking elsewhere
• concise use of space: histogram on slider


Idiom: scented widgets

- augmented widgets show *information scent*
  – better cues for *information foraging*: show whether value in drilling down further vs looking elsewhere

- concise use of space: histogram on slider
Scented histogram bisliders: detailed

Idiom: **boxplot**

- static item aggregation
- task: find distribution
- data: table
- derived data
  - 5 quant attributes
    - median: central line
    - lower and upper quartile: boxes
    - lower upper fences: whiskers
      - values beyond which items are outliers
  - outliers beyond fence cutoffs explicitly shown
- scalability
  - unlimited number of items!

[40 years of boxplots, Wickham and Stryjewski, 2012]
Idiom: **Continuous scatterplot**

- static item aggregation
- data: table
- derived data: table
  - key attribs x,y for pixels
  - quant attrib: overplot density
- dense space-filling 2D matrix
- color:
  sequential categorical hue + ordered luminance colormap
- scalability
  - no limits on overplotting: millions of items

Spatial aggregation

• MAUP: Modifiable Areal Unit Problem
  – changing boundaries of cartographic regions can yield dramatically different results
  – zone effects

[http://www.e-education.psu.edu/geog486/l4_p7.html, Fig 4.cg.6]

– scale effects

Gerrymandering: MAUP for political gain

A real district in Pennsylvania:
Democrats won 51% of the vote but only 5 out of 18 house seats

https://www.washingtonpost.com/news/wonk/wp/2015/03/01/this-is-the-best-explanation-of-gerrymandering-you-will-ever-see/
Dynamic aggregation: Clustering

• clustering: classification of items into similar bins
  – based on similarity measure
  – hierarchical algorithms produce "similarity tree": cluster hierarchy
    • agglomerative clustering: start w/ each node as own cluster, then iteratively merge

• cluster hierarchy: derived data used w/ many dynamic aggregation idioms
  – cluster more homogeneous than whole dataset
    • statistical measures & distribution more meaningful
Idiom: **Hierarchical parallel coordinates**

- dynamic item aggregation
- derived data: **cluster hierarchy**
- encoding:
  - cluster band with variable transparency, line at mean, width by min/max values
  - color by proximity in hierarchy

Dimensionality Reduction
Attribute aggregation: Dimensionality reduction

- attribute aggregation
  - derive low-dimensional target space from high-dimensional measured space
    - capture most of variance with minimal error
  - use when you can’t directly measure what you care about
    - true dimensionality of dataset conjectured to be smaller than dimensionality of measurements
    - latent factors, hidden variables

```
tumor measurement data
```

```
data: 9D measured space
```

```
DR
```

```
benign
```

```
malignant
```

```
derived data: 2D target space
```

```
tumor measurement data
```

```
data: 9D measured space
```

```
DR
```

```
benign
```

```
malignant
```

```
derived data: 2D target space
```
Dimensionality vs attribute reduction

• vocab use in field not consistent
  – dimension/attribute

• attribute reduction: reduce set with filtering
  – includes orthographic projection

• dimensionality reduction: create smaller set of new dims/attrs
  – typically implies dimensional aggregation, not just filtering
  – vocab: projection/mapping
Dimensionality reduction & visualization

• why do people do DR?
  – improve performance of downstream algorithm
    • avoid curse of dimensionality
  – data analysis
    • if look at the output: visual data analysis

• abstract tasks when visualizing DR data
  – dimension-oriented tasks
    • naming synthesized dims, mapping synthesized dims to original dims
  – cluster-oriented tasks
    • verifying clusters, naming clusters, matching clusters and classes

Dimension-oriented tasks

- naming synthesized dims: inspect data represented by lowD points

Cluster-oriented tasks

- verifying, naming, matching to classes

Idiom: Dimensionality reduction for documents

Task 1

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

Why?  ➤ Produce  ➤ Derive

Task 2

What?  ➤ In 2D data  ➤ Out Scatterplot

Why?  ➤ Discover  ➤ Explore  ➤ Identify

How?  ➤ Encode  ➤ Navigate  ➤ Select

Task 3

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

Why?  ➤ Produce  ➤ Annotate
Latest algorithms: t-SNE, UMAP

- t-SNE [https://distill.pub/2016/misread-tsne/](https://distill.pub/2016/misread-tsne/)
- UMAP [https://pair-code.github.io/understanding-umap/](https://pair-code.github.io/understanding-umap/)
Interacting with dimensionally reduced data

[https://uclab.fh-potsdam.de/projects/probing-projections/]

Linear dimensionality reduction

• principal components analysis (PCA)
  – finding axes: first with most variance, second with next most, …
  – describe location of each point as linear combination of weights for each axis
    • mapping synthesized dims to original dims
Nonlinear dimensionality reduction

- **pro**: can handle curved rather than linear structure
- **cons**: lose all ties to original dims/attrs
  - new dimensions often cannot be easily related to originals
  - mapping synthesized dims to original dims task is difficult

- **many techniques proposed**
  - many literatures: visualization, machine learning, optimization, psychology, ...
  - techniques: t-SNE, MDS (multidimensional scaling), charting, isomap, LLE,…
    - t-SNE: excellent for clusters
      - but some trickiness remains: http://distill.pub/2016/misread-tsne/
    - MDS: confusingly, entire family of techniques, both linear and nonlinear
      - minimize stress or strain metrics
      - early formulations equivalent to PCA
VDA with DR example: nonlinear vs linear

• DR for computer graphics reflectance model
  – goal: simulate how light bounces off materials to make realistic pictures
    • computer graphics: BRDF (reflectance)
  – idea: measure what light does with real materials

[Fig 2. Matusik, Pfister, Brand, and McMillan. A Data-Driven Reflectance Model. SIGGRAPH 2003]
Capturing & using material reflectance

• reflectance measurement: interaction of light with real materials (spheres)
  – result: 104 high-res images of material
    – each image 4M pixels

• goal: image synthesis
  – simulate completely new materials

• need for more concise model
  – 104 materials * 4M pixels = 400M dims
  – want concise model with meaningful knobs
    • how shiny/greasy/metallic
    • DR to the rescue!

[Figs 5/6. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]
Linear DR

• first try: PCA (linear)
• result: error falls off sharply after ~45 dimensions
  – scree plots: error vs number of dimensions in lowD projection
• problem: physically impossible intermediate points when simulating new materials
  – specular highlights cannot have holes!

[Fig 6/7. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]
Nonlinear DR

- second try: charting (nonlinear DR technique)
  - scree plot suggests 10-15 dims
  - note: dim estimate depends on technique used!

[Fig 10/11. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]
Finding semantics for synthetic dimensions

• look for meaning in scatterplots
  – synthetic dims created by algorithm but named by human analysts
  – points represent real-world images (spheres)
  – people inspect images corresponding to points to decide if axis could have meaningful name

• cross-check meaning
  – arrows show simulated images (teapots) made from model
  – check if those match dimension semantics

[Fig 12/16. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]
Understanding synthetic dimensions

Specular-Metallic

Diffuseness-Glossiness

[Fig 13/14/16. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]
# How?

## Encode

<table>
<thead>
<tr>
<th>Arrange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Express</td>
</tr>
<tr>
<td>Separate</td>
</tr>
<tr>
<td>Order</td>
</tr>
<tr>
<td>Align</td>
</tr>
<tr>
<td>Use</td>
</tr>
</tbody>
</table>

- **Map**
  - from categorical and ordered attributes
  - Color
    - Hue
    - Saturation
    - Luminance
  - Size, Angle, Curvature, ...
  - Shape
    - 
  - Motion
    - Direction, Rate, Frequency, ...

## Manipulate

<table>
<thead>
<tr>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juxtapose</td>
</tr>
<tr>
<td>Select</td>
</tr>
<tr>
<td>Partition</td>
</tr>
<tr>
<td>Navigate</td>
</tr>
<tr>
<td>Superimpose</td>
</tr>
</tbody>
</table>

## Facet

## Reduce

- **Filter**
- **Aggregate**
- **Embed**