Information Visualization
Reduce: Aggregation & Filtering
Project Peer Reviews

Tamara Munzner
Department of Computer Science
University of British Columbia

Week 12, 16 Nov 2022
https://www.cs.ubc.ca/~tmunzner/S47-13

Today
• first: project peer reviews
  – join your matched teams
  • you’ve already read other student’s written update
  – we’ll begin by private Piazza post (if your counterpart(s) weren’t prepared)
  – record discussion/thoughts in gdoc (freelime)
  – Revs: A critiques B; then B critiques A
• break
• Q&A / mini-lecture this time
  – Ch 13, Reduce

Peers 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, critiques record branchnode (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/why not
  – 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)
  – symposium last week’s readings / discussion (light, 2 readings)
  – Ch 14: End-of-Sem CourseNotes
  – framework: heuristics; local belief (anti-patterns); visual design space; literature review
  – (type: design study)
  – in-class post-exam meetings with Tamara
  – oral feedback on project progress, after I’ve read them

How to handle complexity: 3 previous strategies + 1 more

Visualization Analysis & Design
Reduce: Aggregation & Filtering (Ch 13)

Tamara Munzner
Department of Computer Science
University of British Columbia
@tamaramunzner

Q&A / Backup Slides

Reduce items and attributes
• reduce/increase: inverses
• filter
  – pro: straightforward and intuitive
  – no understanding of compute
  – can out of sight, out of mind
• aggregation
  – pro: informs 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 larger/smaller than x
  – positive/negative
• 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 result of action

Idiom: Crossfilter
• cross filtering
• coordinated views/controls combined
  – all-sorted histogram/box/whisker update when any range change

Aggregate
• a group of elements is represented by a smaller number of derived elements
• derive data
  – new table lies in bins, values are counts
• bin size crucial
  – pattern can change dramatically depending on discretization
  – opportunity for interaction: control bin size on the fly

Idiom: Histogram
• static item aggregation
• task: find distribution
• data: table
• derived data
• aggregate
  – items
  – attributes

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

Derive data
• new table: keys are bins, values are counts
  – opportunity for interaction: control bin size on the fly
  – pattern can change dramatically depending on discretization
  – opportunity for interaction: control bin size on the fly

Embed - Focus+Context
Ch 14: Embed - Focus+Context

Teacher: Remove items, attributes to reduce... 

Student: Reduce items and attributes...
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

Spatial aggregation

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

Gerrymandering: MAUP for political gain

- MAUP: Modifiable Areal Unit Problem
- clusters
- Democratic won 51% of the vote but only 5 out of 18 house seats

Dimensionality vs attribute reduction

- attribute aggregation
- derive low-dimensional target space from high-dimensional measured space
- capture most 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
- loose factors, hidden variables

Dynamic aggregation: Clustering

- clustering: classification of items into similar bins
- based on similarity measure
- hierarchical algorithms produce “similarity tree”: cluster hierarchy
- agglomerative clustering: start with each node as own cluster, then iteratively merge
- cluster hierarchy: derived data used in many dynamic aggregation idioms

Dimensionality reduction for documents

- why do people do DR?
- improve performance of downstream algorithm
- avoid curse of dimensionality
- data analysis
- if look in the output: visual data analysis
- abstract tasks when visualizing DR data
- dimension-oriented tasks
- naming synthesized dims; inspect data represented by lowD points
- cluster-oriented tasks
- verifying, naming, matching to classes

Dimension-oriented tasks

- naming synthesized dims; inspect data represented by lowD points

Cluster-oriented tasks

- verifying, naming, matching to classes
- in discernable clusters
clearly discernable clusters
clear match cluster/class
partial match cluster/class
no match cluster/class

Idiom: boxplot

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

Capturing & using material reflectance

- reflectance measurement: interaction of light with real materials (spheres)
- goal: simulate completely new materials
- need for more concise model
  - 104 materials: 4 M pixels = 402M dims
  - want concise model with meaningful knobs
  - how do they play off each other?
  - DR to the rescue!

Reflectance Model. SIGGRAPH 2003

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

Linear DR

- first try: PCA (linear)
  - results: error falls off sharply after ~45 dimensions
  - screw plot suggests 10-15 dims

New dimensionalities needed for more concise model

- points when simulating new materials
- linear vs nonlinear

- result: error falls off sharply after ~45 dimensions
- cons: lose all ties to original dims/attribs
- many techniques proposed
  - UMAP
  - t-SNE: excellent for clusters
  - but some trickiness remain: http://distill.pub/2016/misread-tsne/
  - https://pair-code.github.io/understanding-umap/
  - scree plot suggests 10-15 dims

Finding semantics for synthetic dimensions

- look for meaning in scatterplots
  - synthetic dims created by algorithm but named by human analysis
  - points represent real-world images (spheres)
  - people inspect images corresponding to points to decide if sets could have meaningful names
  - cross-check meaning
    - arrows show simulated images (aspects) made from model
    - check if those match dimension semantics

Understanding 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

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

Lab Study: Test Human Response Time and Error

- hypotheses
  - points are better than landscapes
  - speed, yes
  - result: better: 2.4x faster. 5-14 more accurate
  - 2D landscapes (color only) better than 3D landscapes (color + height redundantly encoded)
  - result: yes
  - significantly faster, no significant difference is accuracy

Nonlinear dimensionality reduction

- can handle curved rather than linear structure
- cons: lose all ties to original dims/attribs
- mapping synthesized dims to original dims is difficult
- many techniques proposed
  - UMAP
  - t-SNE: excellent for clusters
  - but some trickiness remain: http://distill.pub/2016/misread-tsne/
  - t-SNE confusingly share family of techniques, both linear and nonlinear
  - minimize stress or arrow metrics
  - early formulations equivalent to PCA

Information Landscapes

- 2D or 3D landscape from set of DR points
  - height based on density
  - oddly popular choice in DR
  - despite known autocorrelation problems with 3D
  - assertions: pattern recognition, spatial reasoning familiar