Information Visualization
Aggregate & Filter 2

Tamara Munzner
Department of Computer Science
University of British Columbia

Lect 19, 17 Mar 2020

https://www.cs.ubc.ca/~tmm/courses/436V-20

News

- Online lectures and office hours start today, using Zoom: https://zoom.us/j/9016202871
- Lecture mode
  - Plan: I livestream with video + audio + screenshare, will also try recording.
  - You'll be able to just join the session
  - Please connect audio-only, no video, to avoid congestion
  - You'll be auto-muted. If you have a question use the Show Hand (click on Participants, button is at the bottom of the popup window), I'll unmute you myself
- Office hours mode
  - Please do connect with video if possible, in addition to audio
  - I'll use the Waiting Room feature, where I will individually allow you in
    - If I'm already talking to somebody else I'll briefly let you know, then put you back in WR until it's your turn.

News

- Labs will be Zoom + Canvas scheduling
  - different Zoom URL for each TA, stay tuned
  - you can sign up for reserved slots in advance, or check for availability on the fly
  - more details soon

- Final exam plan still TBD
  - but will **not** be in person
  - you are free to leave campus when you want (but are not required to do so)

Schedule shift

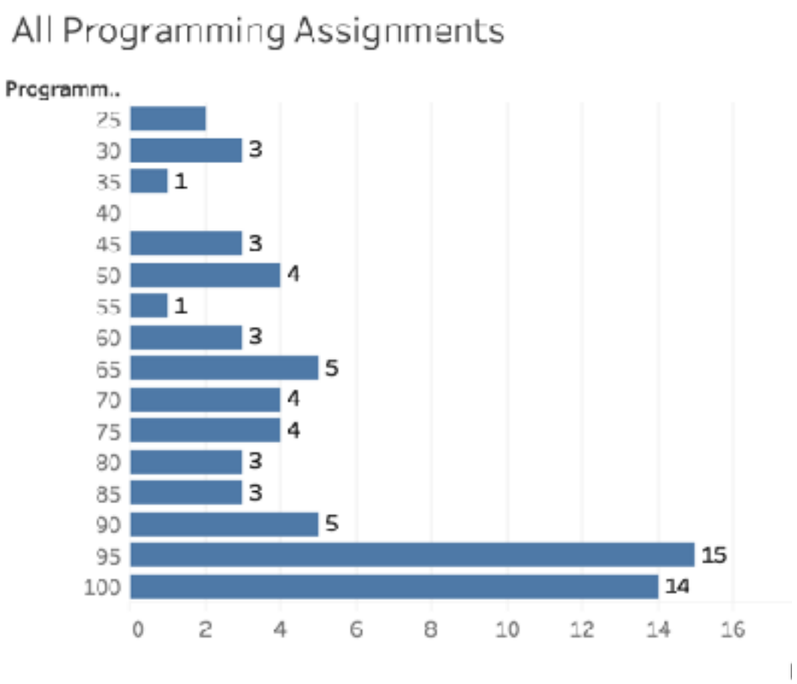
- Nothing due this Wed
- M2 & M3 on schedule
  - M2 due Wed Mar 25
  - M3 due Wed Apr 8
- Combined F5/6
  - will go out Thu Mar 26, due Wed Apr 1

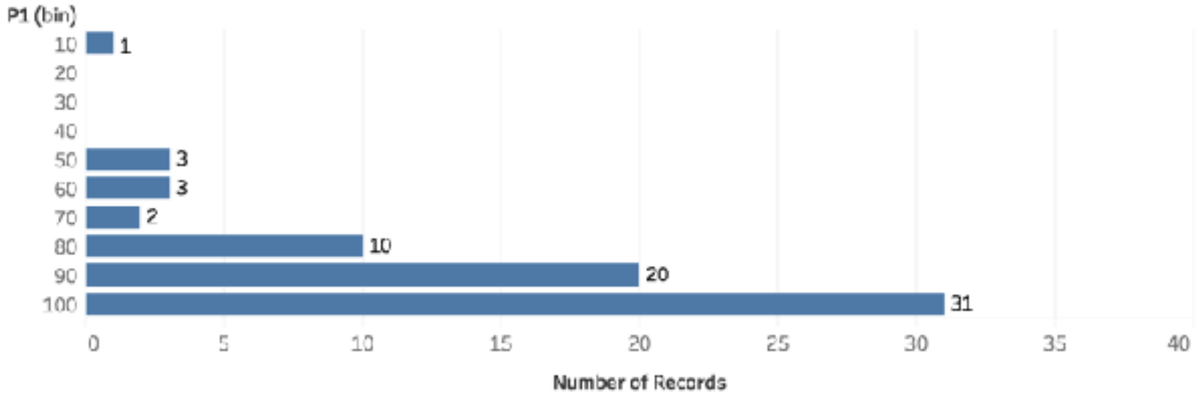
News

- Midterm marks and solutions released
  - Gradescope has detailed breakdown, note stats are wrt total of 75
  - Canvas has percentages, mean was 79%
  - solutions have detailed rubric w/ answer alternatives & explanations
- M1 marks released
  - we specifically suggest meet to discuss during labs or office hrs to several teams
- P3 marks released
  - bimodal distribution

P1-P3 marks

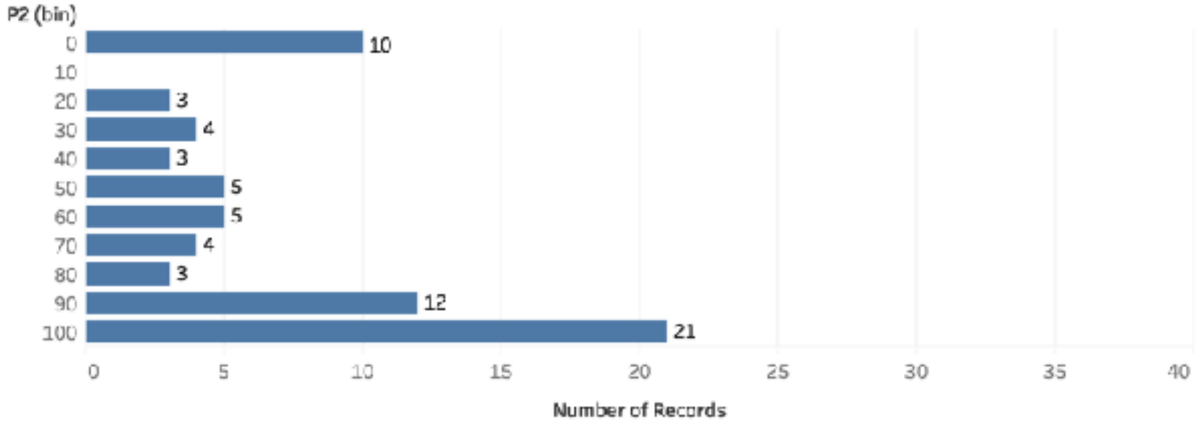
- increasingly bimodal

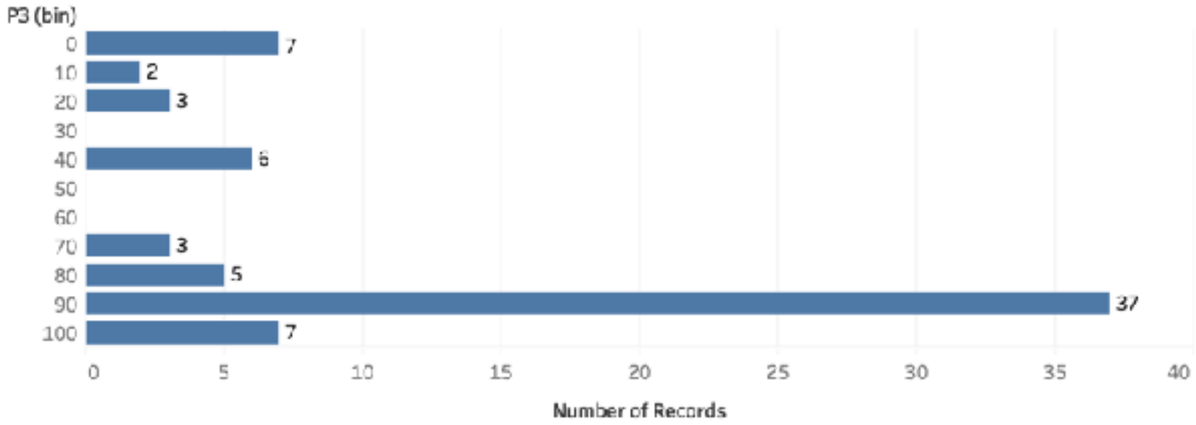




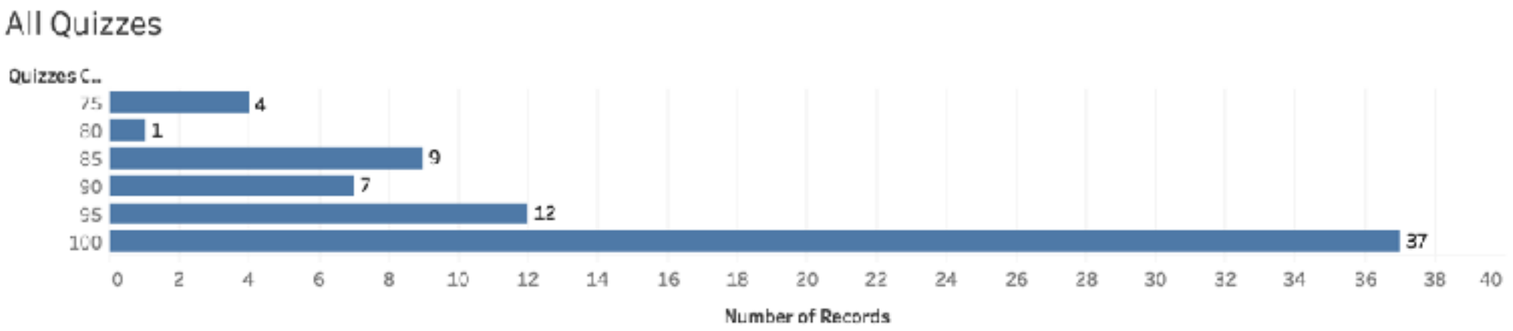
Stats

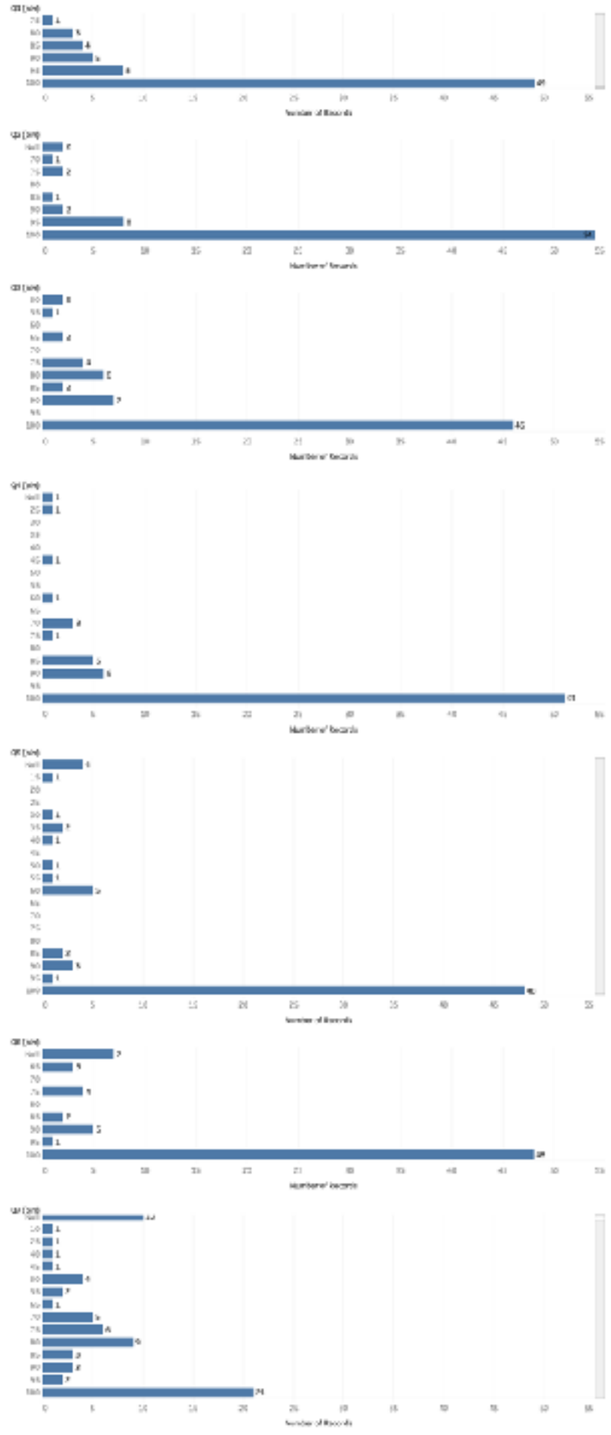
Avg. P1	90.71
P1: 0 or no submission	0.00
Avg. P2	68.03
P2: 0 or no submission	9.00
Avg. P3	75.14
P3: 0 or no submission	6.00



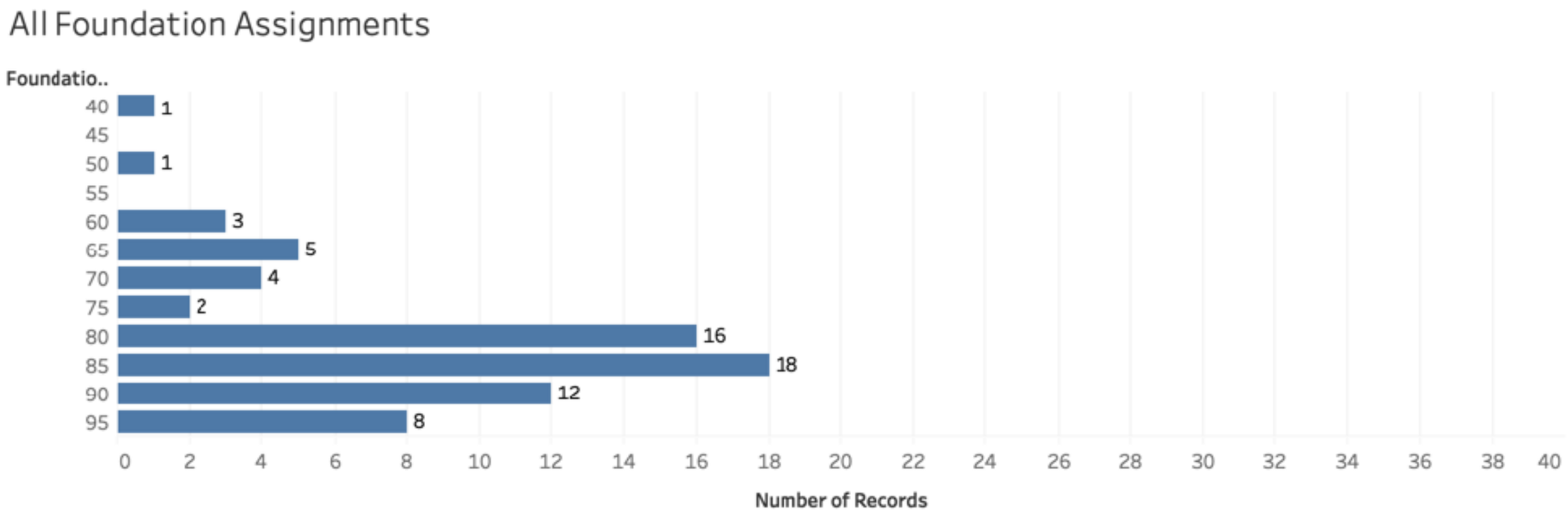


Q1-Q7 marks


All Quizzes
Quizzes C..
75
80
85
90
95
100
4
1
9
7
12
37
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
Number of Records

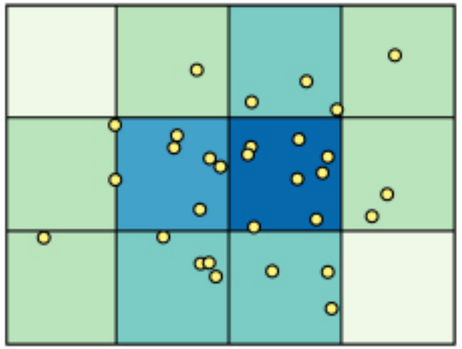
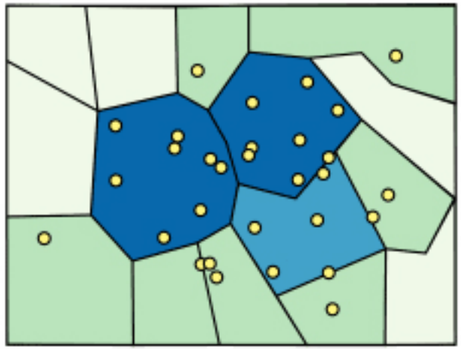
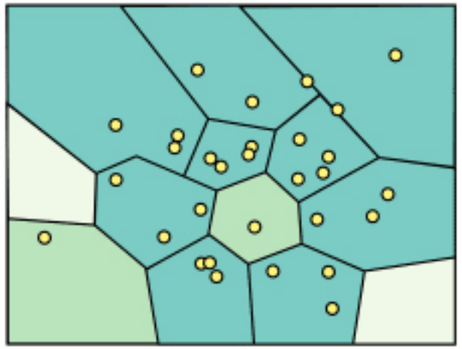


Foundations F1-F4



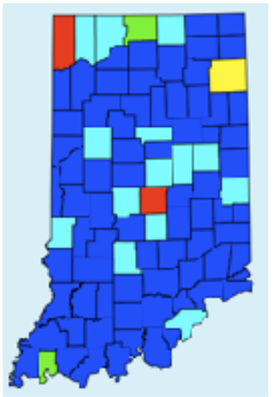
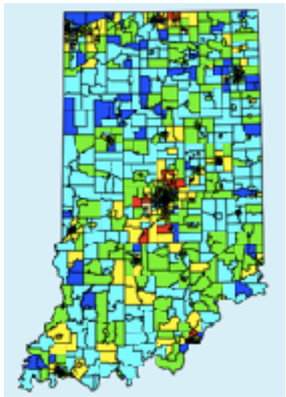
Spatial aggregation

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

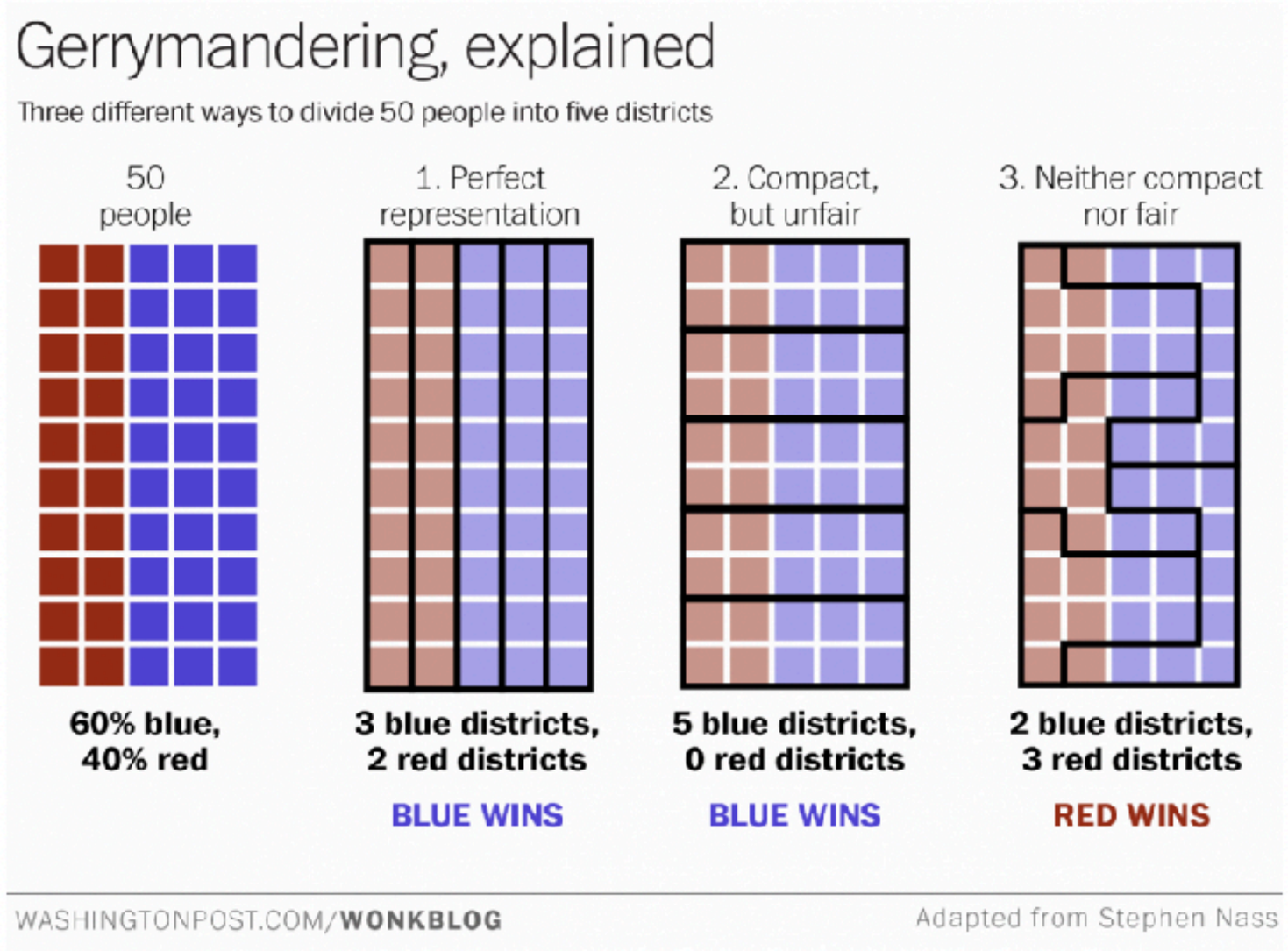
[http://www.e-education.psu/edu/geog486/l4_p7.html, Fig 4.cg.6]

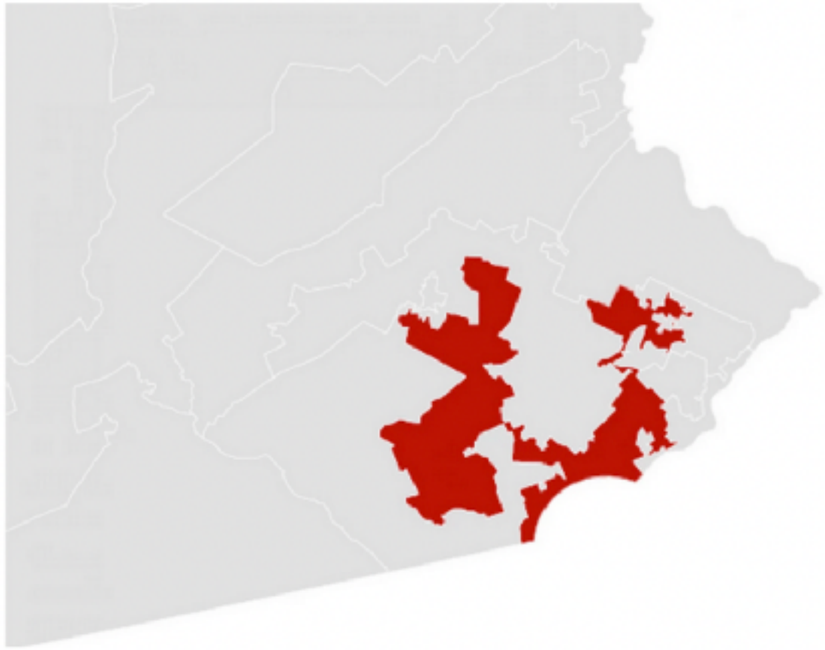
  - scale effects

https://blog.cartographica.com/blog/2011/5/19/the-modifiable-areal-unit-problem-in-gis.html

Gerrymandering: MAUP for political gain



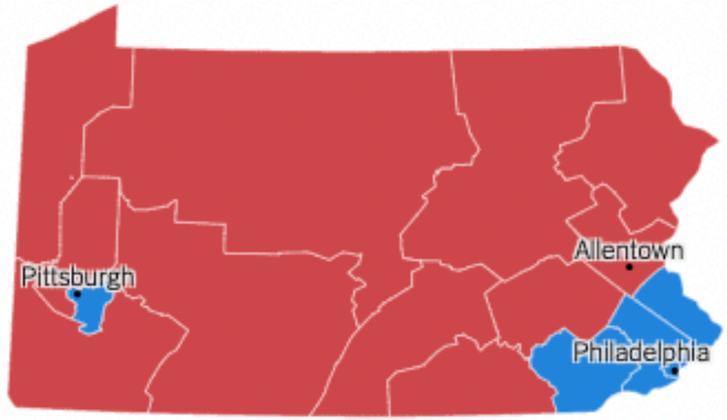


A real district in Pennsylvania:
Democrats won 51% of the vote but only 5 out of 18 house seats

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

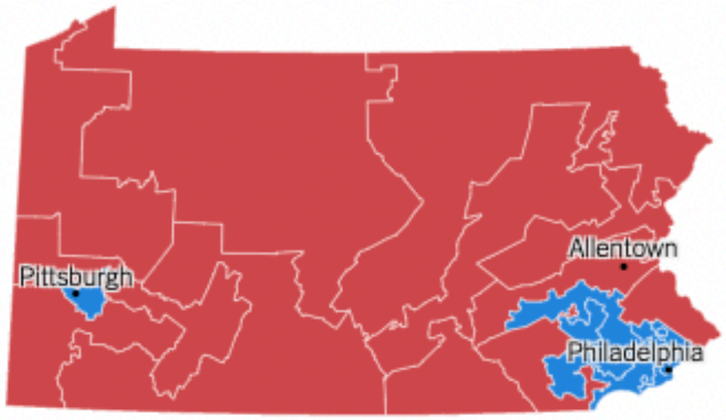
Example: Gerrymandering in PA

Possible nonpartisan map



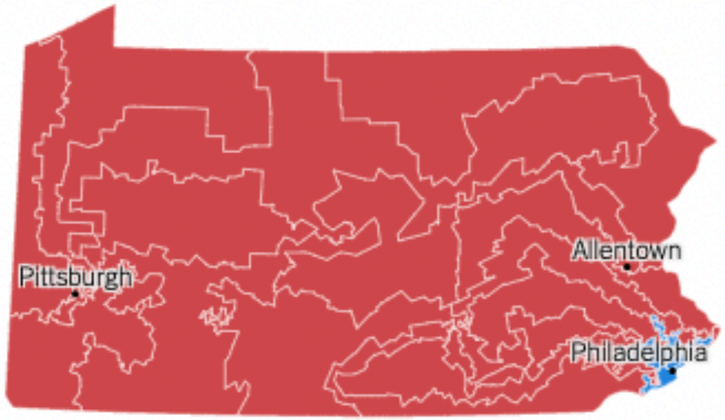
7 Clinton Districts 11 Trump Districts

Current map



6 Clinton Districts 12 Trump Districts

Possible extreme gerrymander

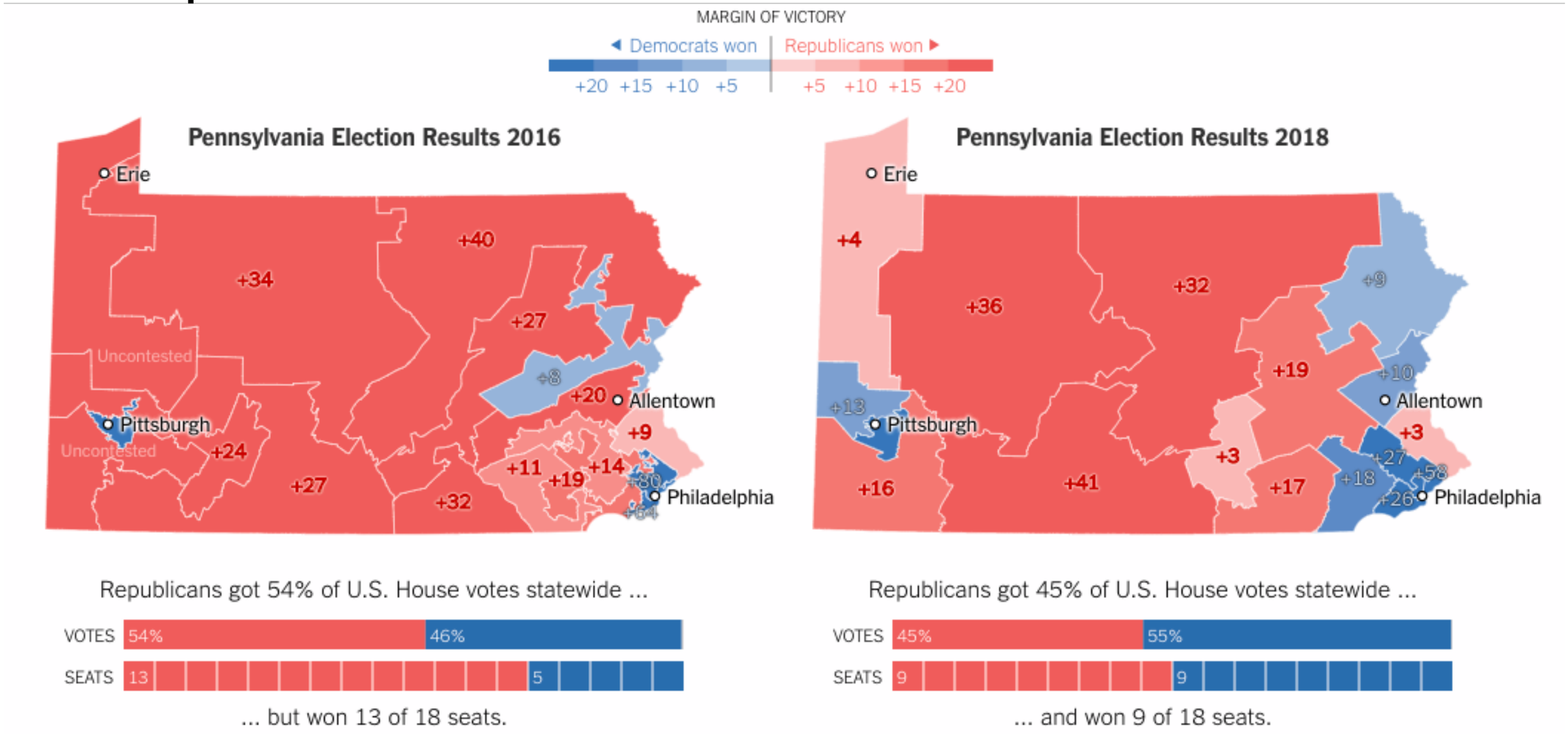


3 Clinton Districts 15 Trump Districts

https://www.nytimes.com/interactive/2018/01/17/upshot/pennsylvania-gerrymandering.html

Example: Gerrymandering in PA

- updated map after court decision



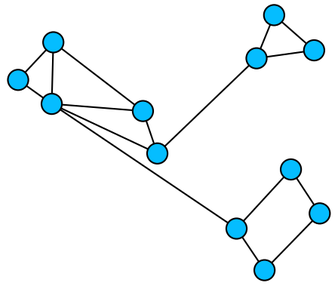
https://www.nytimes.com/interactive/2018/11/29/us/politics/north-carolina-gerrymandering.html?action=click&module=Top%20Stories&pgtype=Homepage

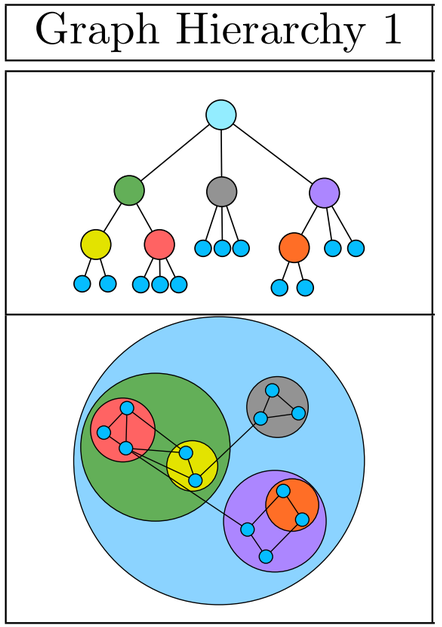
Clustering

- classification of items into similar bins
  - based on similiarity measure
    - Euclidean distance, Pearson correlation
  - partitioning algorithms
    - divide data into set of bins
    - # bins (k) set manually or automatically
  - hierarchical algorithms
    - produce "similarity tree" (dendrograms): 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: **GrouseFlocks**

- data: compound graphs
  - network
  - cluster hierarchy atop it
    - derived or interactively chosen
- visual encoding
  - connection marks for network links
  - containment marks for hierarchy
  - point marks for nodes
- dynamic interaction
  - select individual metanodes in hierarchy to expand/ contract



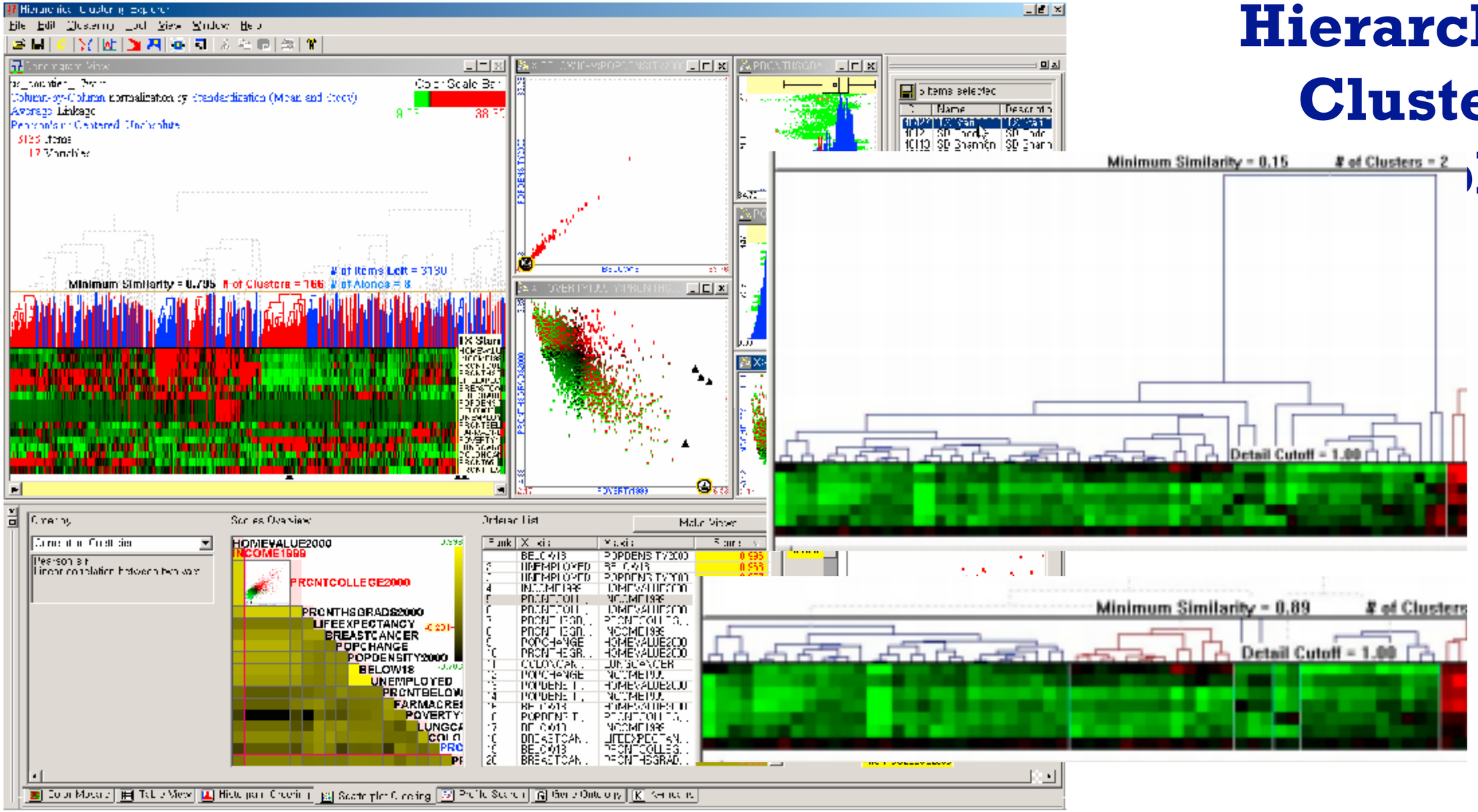


[GrouseFlocks: Steerable Exploration of Graph Hierarchy Space. Archambault, Munzner, and Auber. IEEE TVCG 14(4): 900-913, 2008.]

Idiom: aggregation via hierarchical clustering (visible)

System:

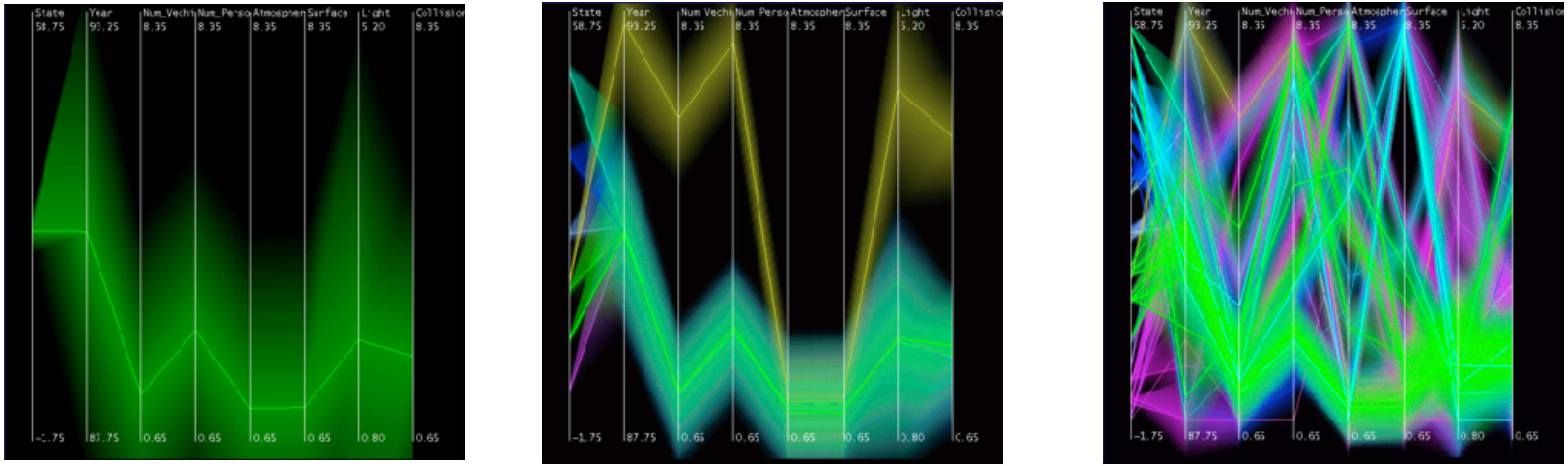
Hierarchical Clustering Explorer



[http://www.cs.umd.edu/hcil/hce/]

Idiom: **Hierarchical parallel coordinates**

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



[Hierarchical Parallel Coordinates for Exploration of Large Datasets. Fua, Ward, and Rundensteiner. Proc. IEEE Visualization Conference (Vis '99), pp. 43– 50, 1999.]

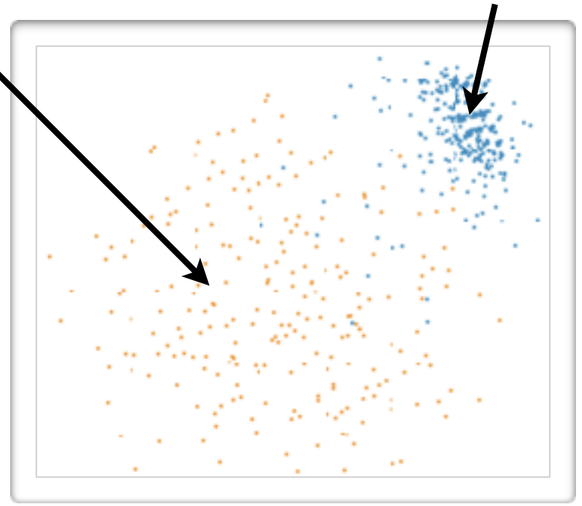
Dimensionality Reduction

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

→ DR →



data: 9D measured space

derived data: 2D target space

Idiom: **Dimensionality reduction for documents**

Task 1

In
HD data

Out
2D data

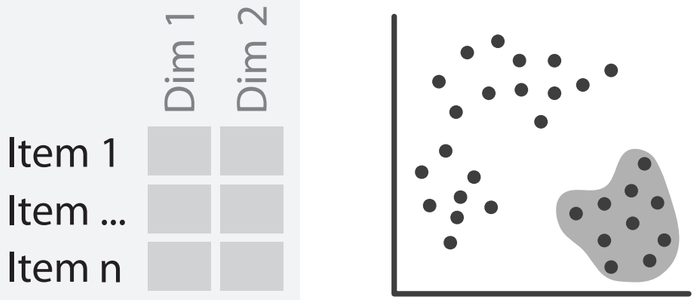
What?

- 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

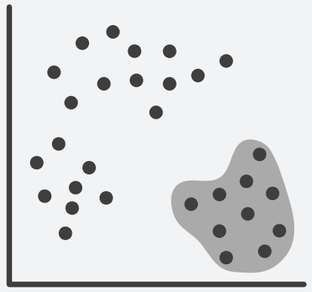
Why?

- Discover
- Explore
- Identify

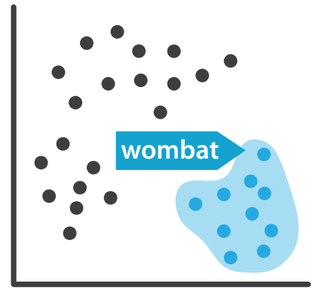
How?

- 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

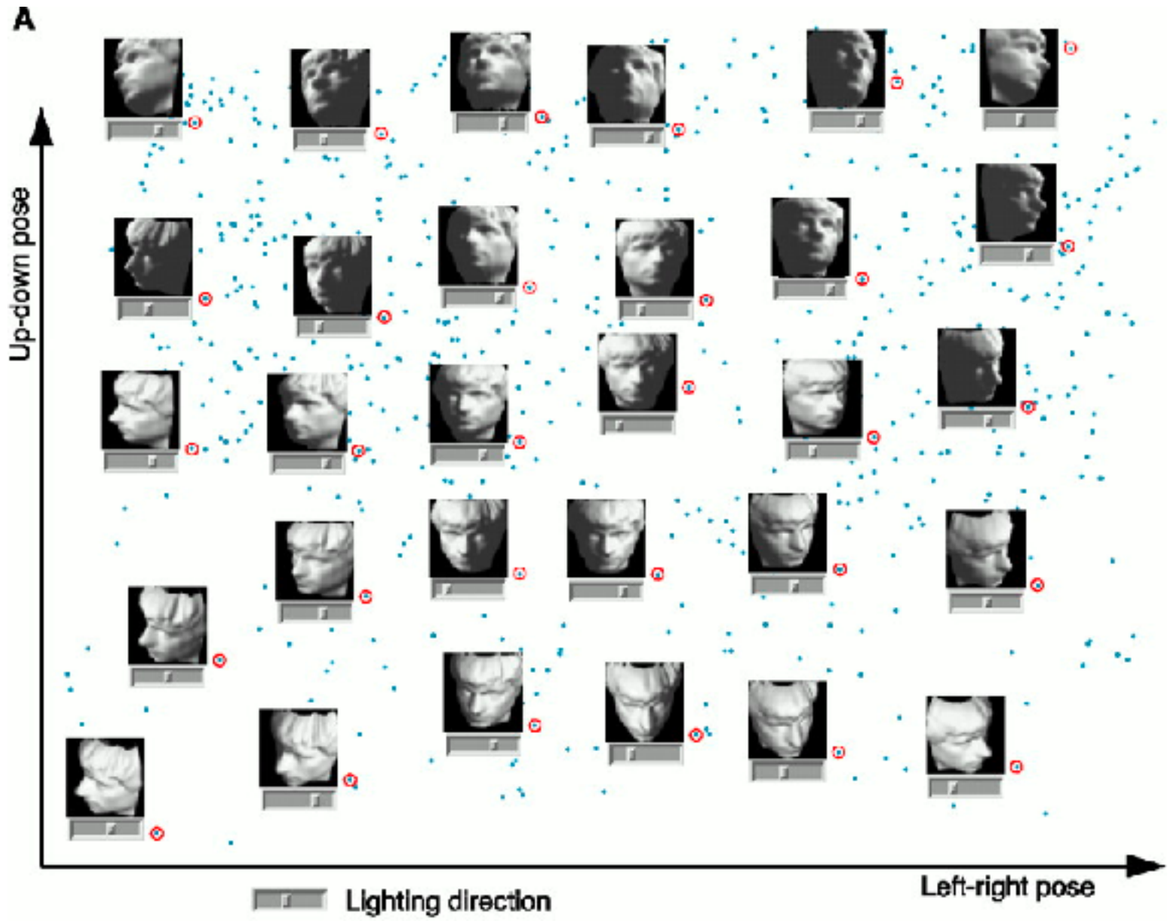
Dimensionality reduction & visualization

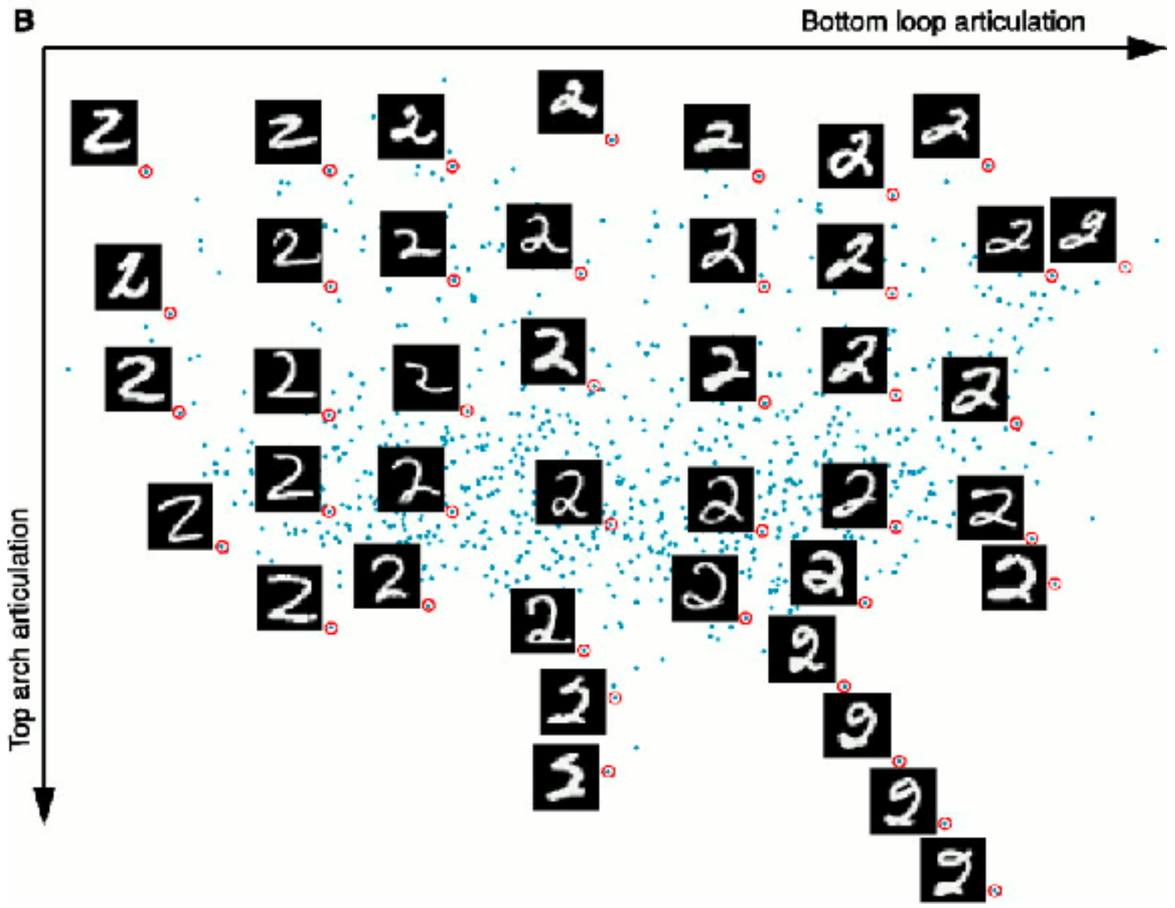
- why do people do DR?
  - improve performance of downstream algorithm
    - avoid curse of dimensionality
  - data analysis
    - if look at the output: visual data analysis
- abstract tasks when visualizing DR data
  - dimension-oriented tasks
    - naming synthesized dims, mapping synthesized dims to original dims
  - cluster-oriented tasks
    - verifying clusters, naming clusters, matching clusters and classes

[Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences. Brehmer, Sedlmair, Ingram, and Munzner. Proc. BELIV 2014.]

Dimension-oriented tasks

- naming synthesized dims: inspect data represented by lowD points

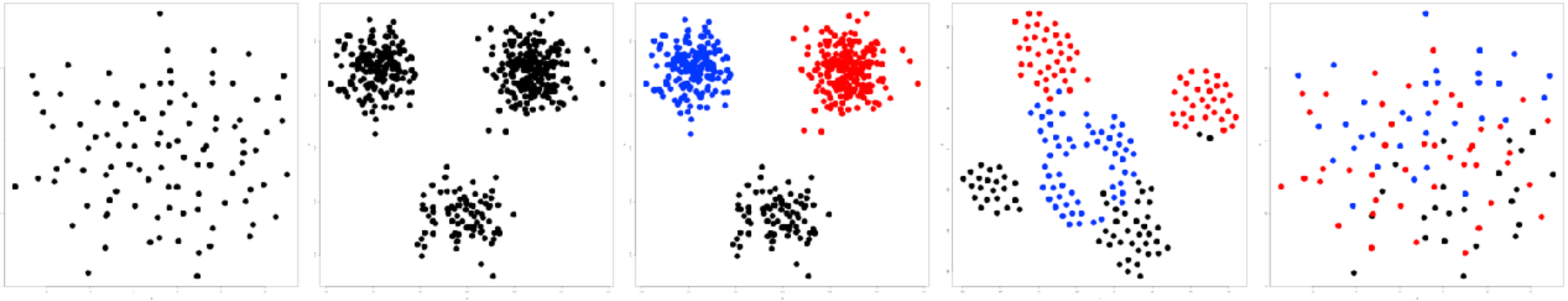




[A global geometric framework for nonlinear dimensionality reduction. Tenenbaum, de Silva, and Langford. Science, 290(5500):2319–2323, 2000.]

Cluster-oriented tasks

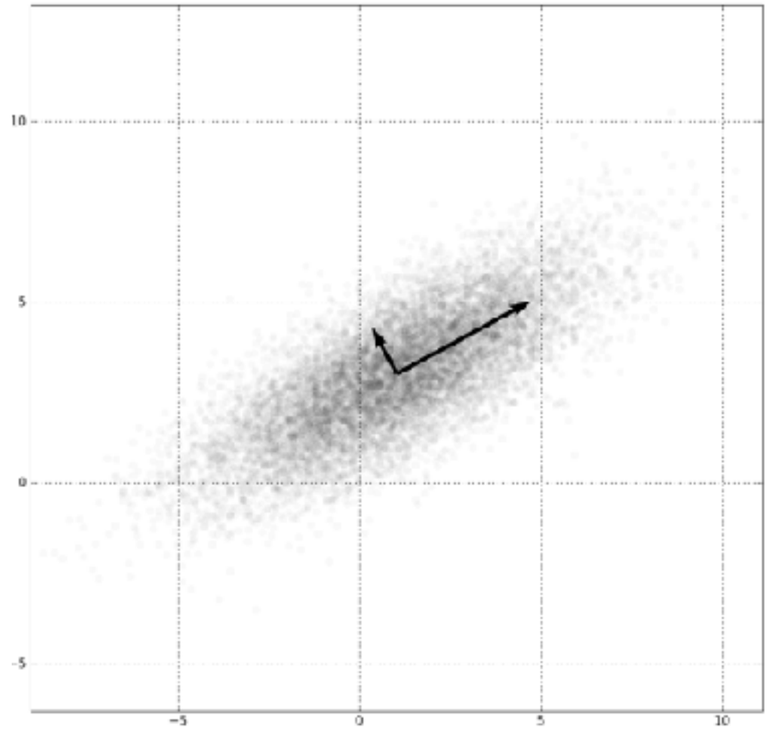
- verifying, naming, matching to classes



[Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences. Brehmer, Sedlmair, Ingram, and Munzner. Proc. BELIV 2014.]

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



[http://en.wikipedia.org/wiki/File:GaussianScatterPCA.png]

Nonlinear dimensionality reduction

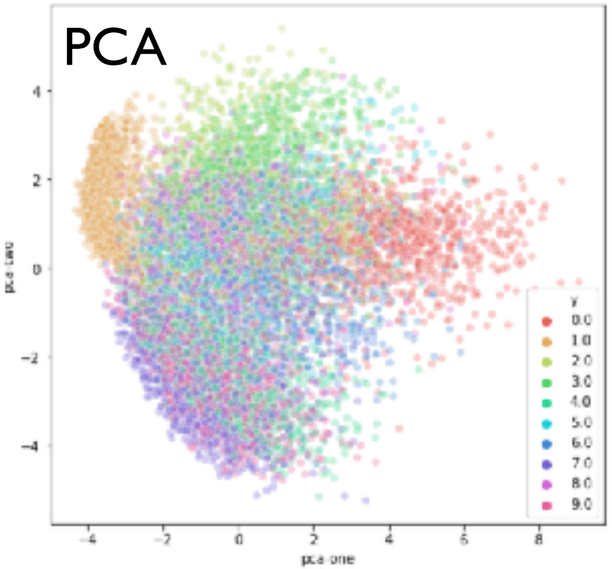
- pro: can handle curved rather than linear structure
- cons: lose all ties to original dims/attribs
  - new dimensions often cannot be easily related to originals
    - mapping synthesized dims to original dims task is difficult
- many techniques proposed
  - many literatures: visualization, machine 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

Nonlinear DR: Many options

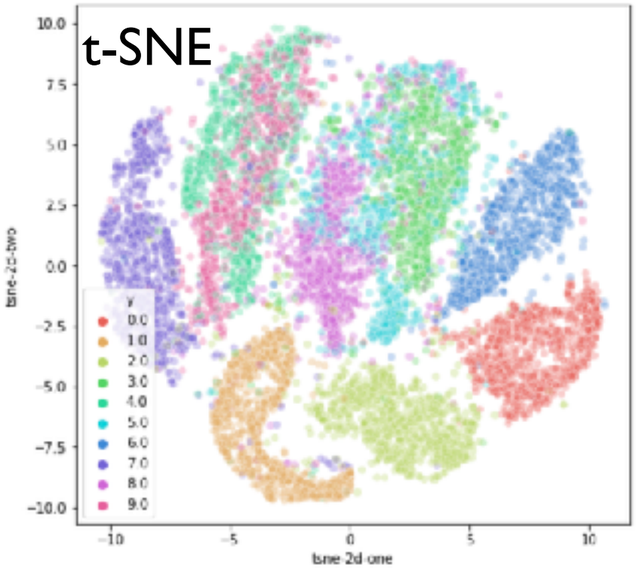
- MDS: multidimensional scaling (treat as optimization problem)
- t-SNE: t-distributed stochastic neighbor embedding
- UMAP: uniform manifold approximation and projection
  - both emphasize cluster structure

MDS

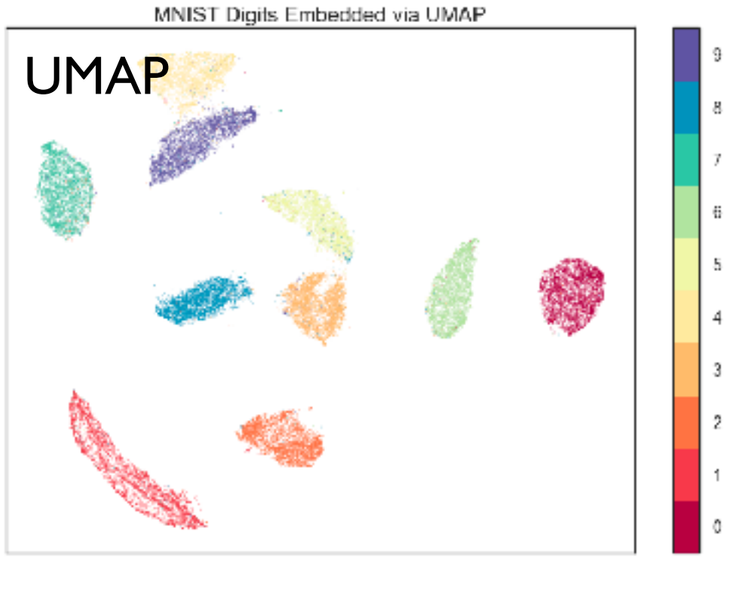
PCA



t-SNE



UMAP



https://colah.github.io/posts/2014-10-Visualizing-MNIST/

https://distill.pub/2016/misread-tsne/

https://pair-code.github.io/understanding-umap/

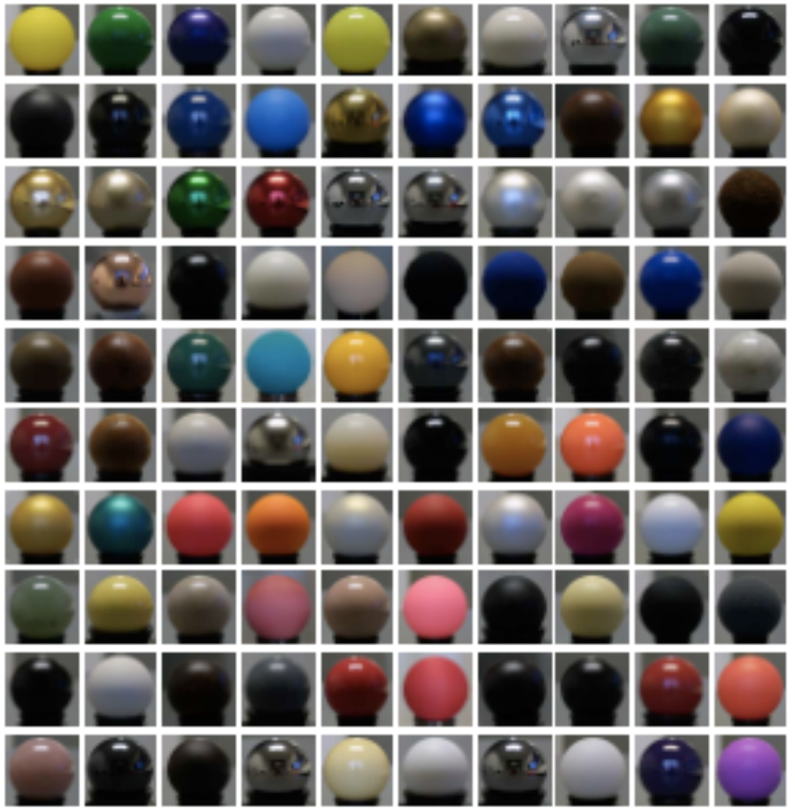
VDA with DR example: nonlinear vs linear

- 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

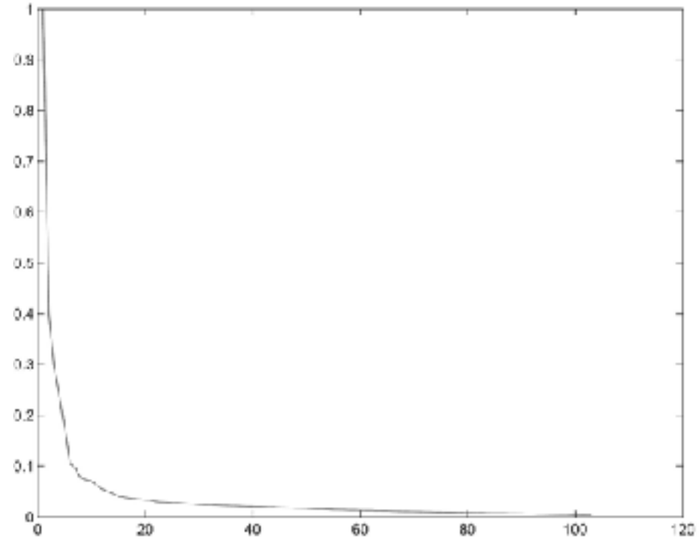
- reflectance measurement: interaction of light with real materials (spheres)
- result: 104 high-res images of material
  - each image 4M pixels
- goal: image synthesis
  - simulate completely new materials
- need for more concise model
  - 104 materials * 4M pixels = 400M dims
  - want concise model with meaningful knobs
    - how shiny/greasy/metallic
    - DR to the rescue!

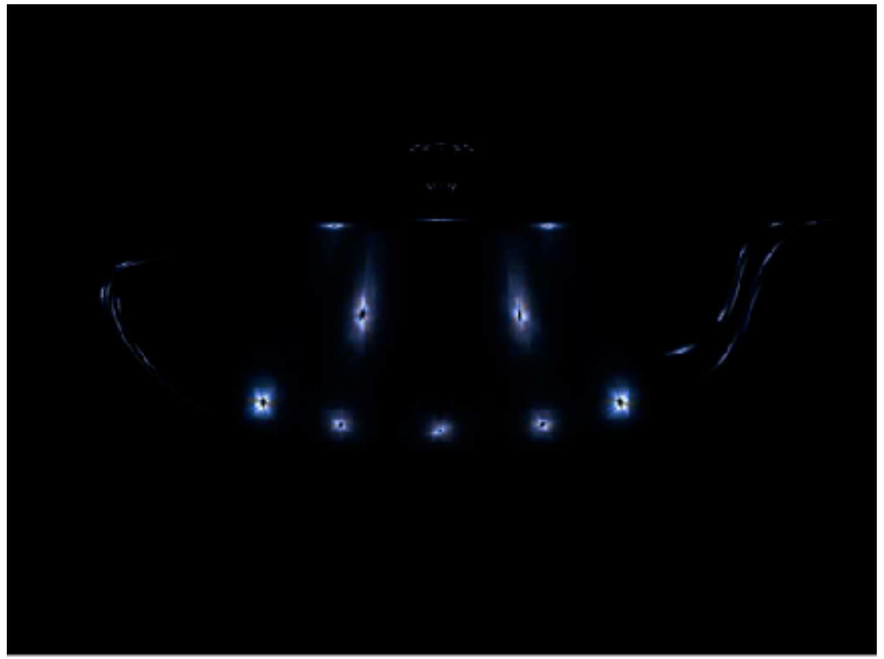


[Figs 5/6. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]

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!

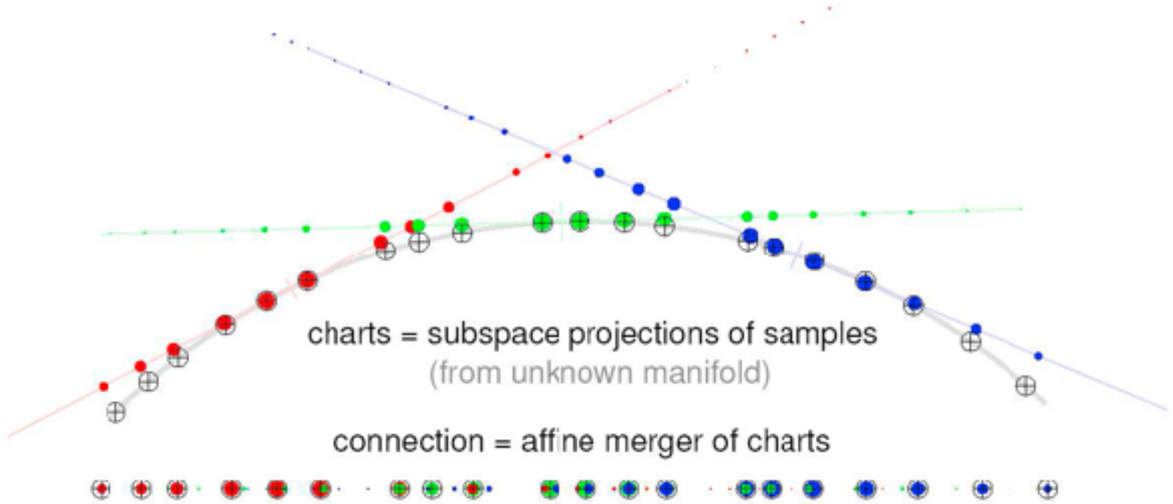


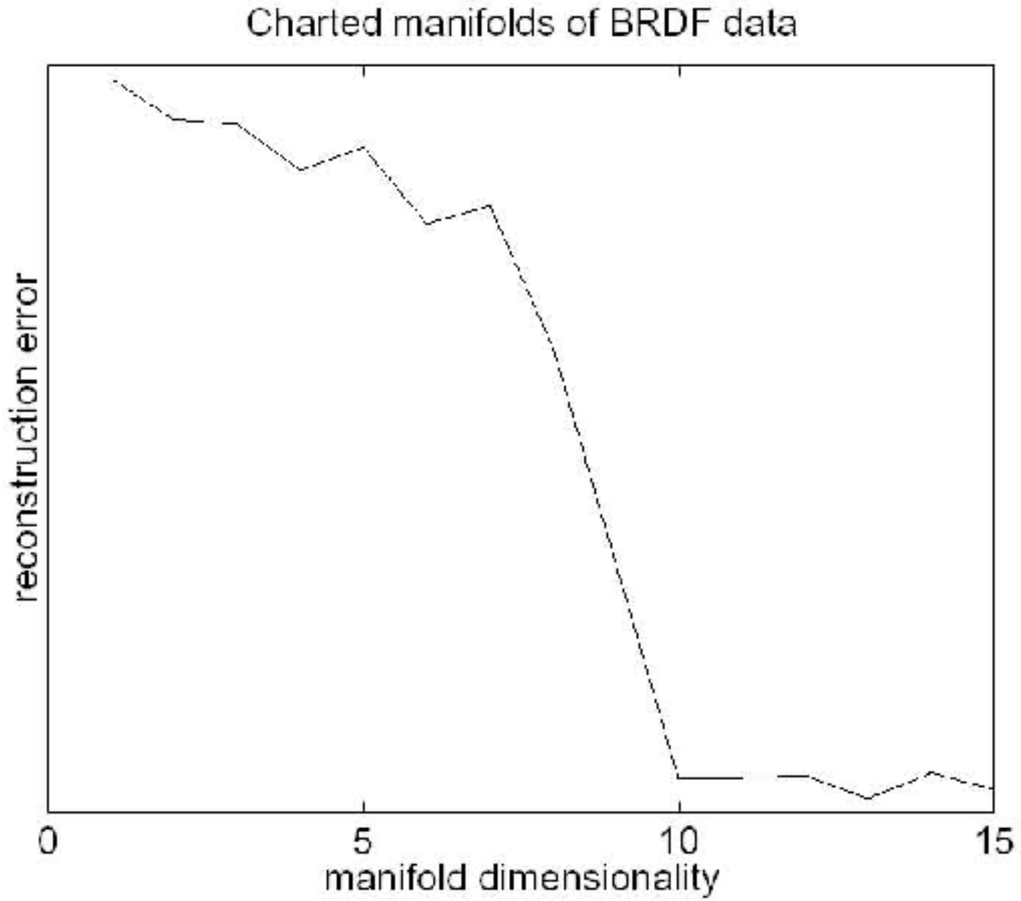


[Figs 6/7. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]

Nonlinear DR

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

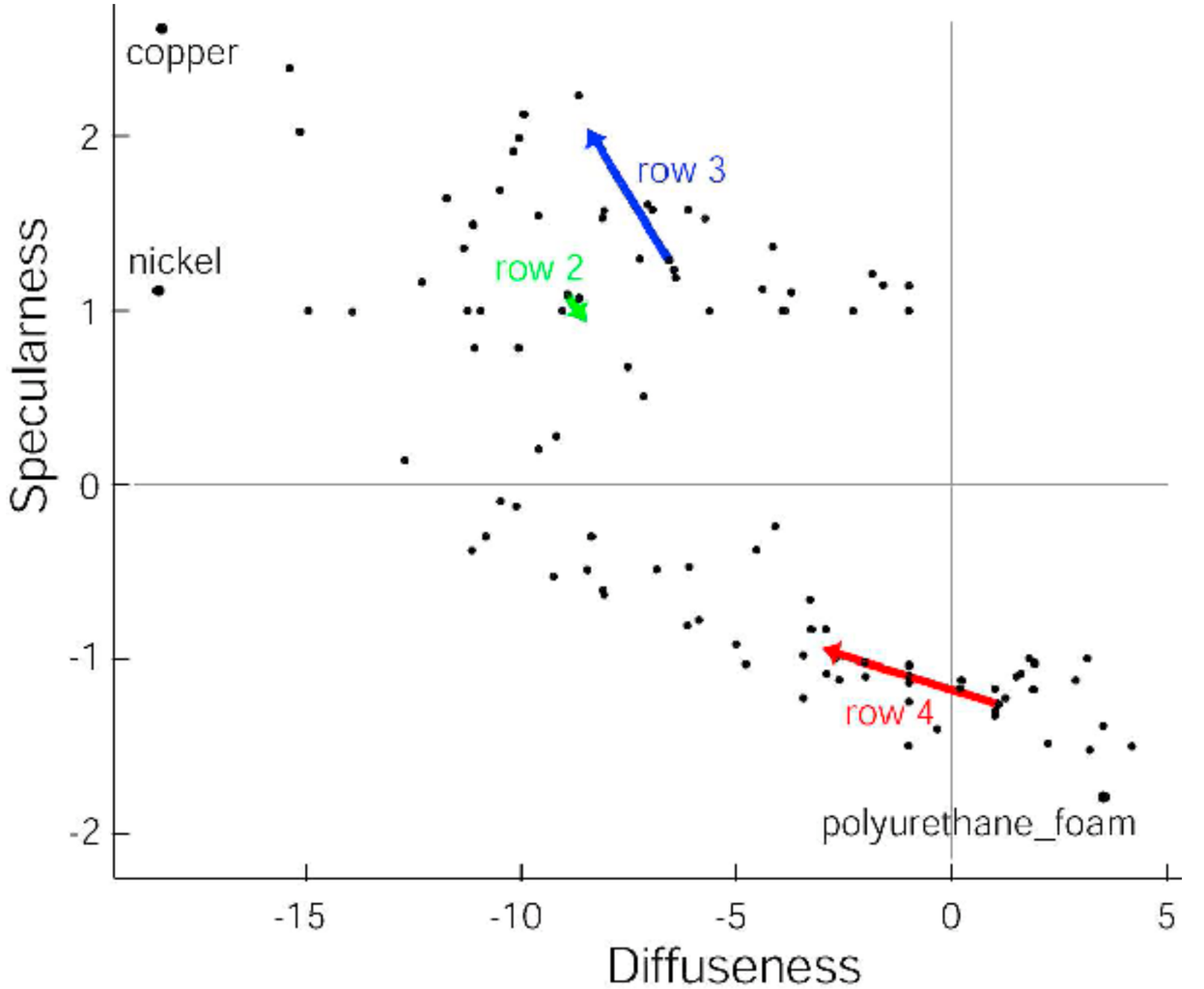




[Fig 10/11. Matusik et al. A Data-Driven Reflectance Model. SIGGRAPH 2003]

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

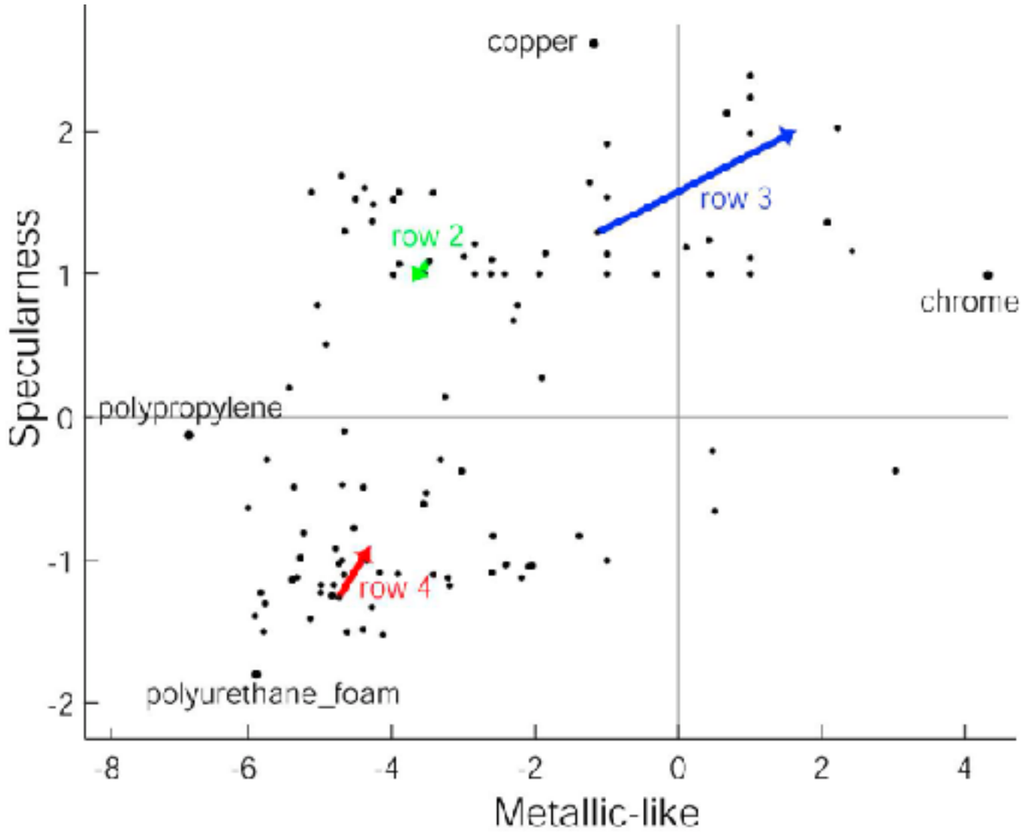


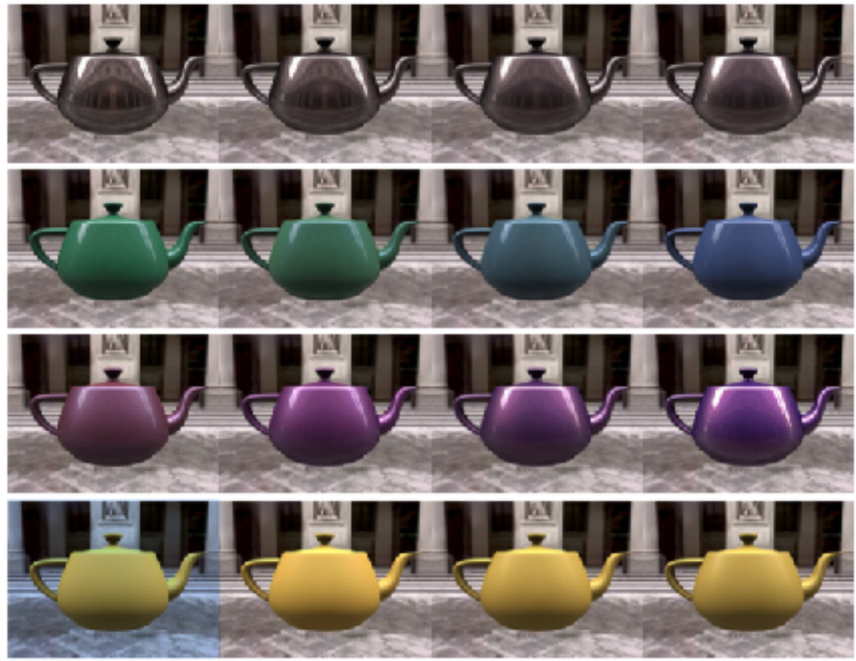
row 4

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

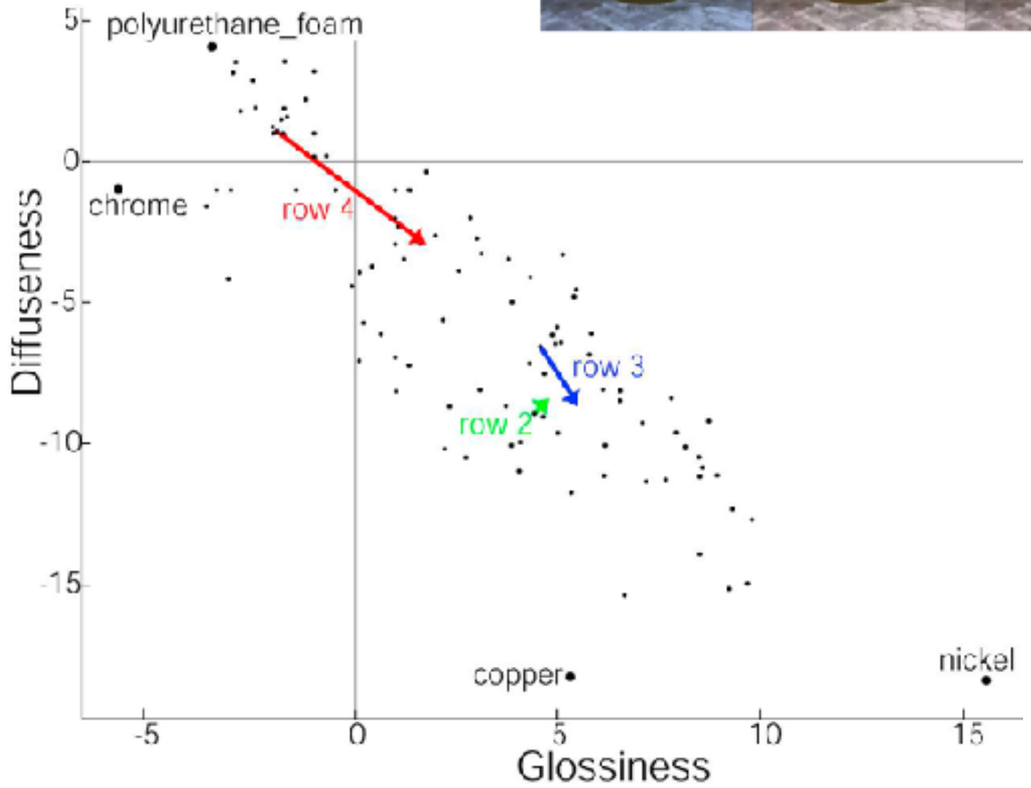
Understanding synthetic dimensions

Specular-Metallic





Diffuseness-Glossiness



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

Embed

Embed: Focus+Context

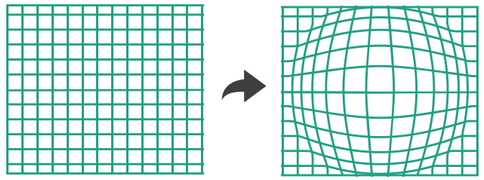
- combine information within single view
- elide
  - selectively filter and aggregate
- superimpose layer
  - local lens
- distortion design choices
  - region shape: radial, rectilinear, complex
  - how many regions: one, many
  - region extent: local, global
  - interaction metaphor

Embed

Elide Data

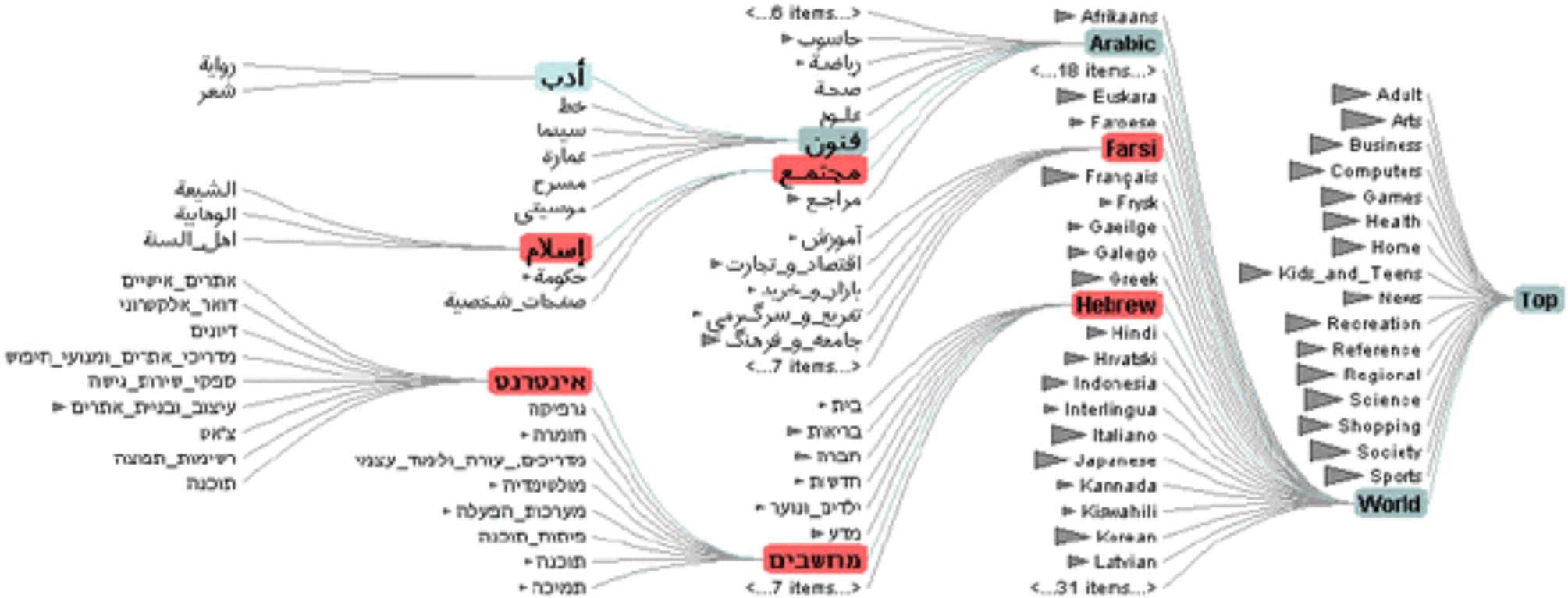
Superimpose Layer

Distort Geometry



Idiom: **DOITrees Revisited**

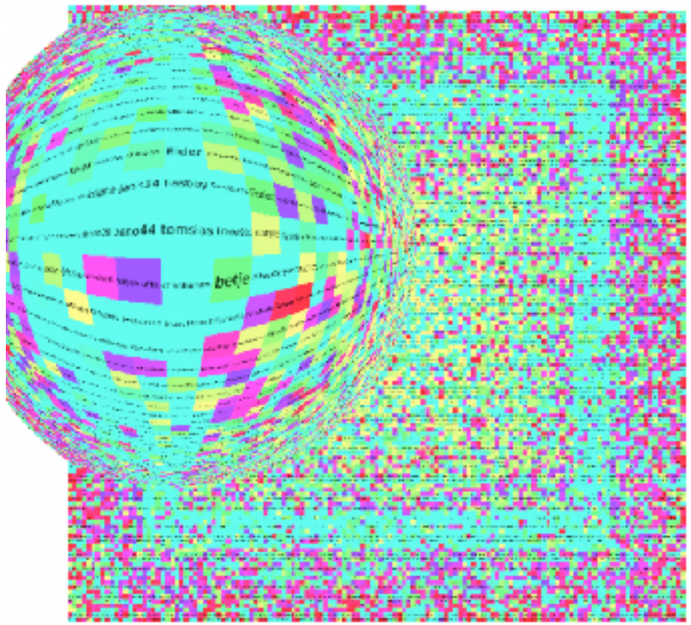
- elide
  - some items dynamically filtered out
  - some items dynamically aggregated together
  - some items shown in detail

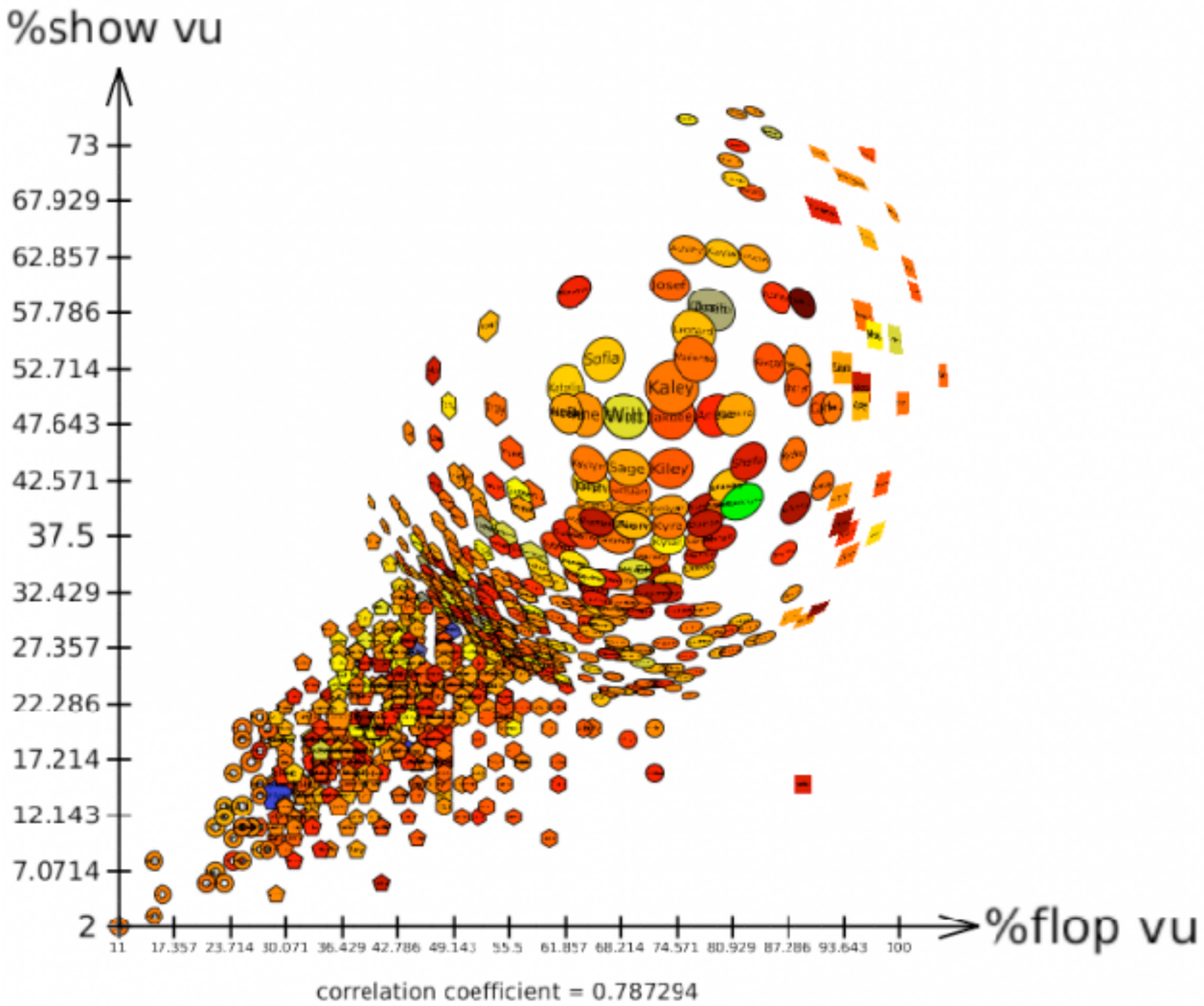


[DOITrees Revisited: Scalable, Space-Constrained Visualization of Hierarchical Data. Heer and Card. Proc. Advanced Visual Interfaces (AVI), pp. 421–424, 2004.]

Idiom: **Fisheye Lens**

- distort geometry
  - shape: radial
  - focus: single extent
  - extent: local
  - metaphor: draggable lens

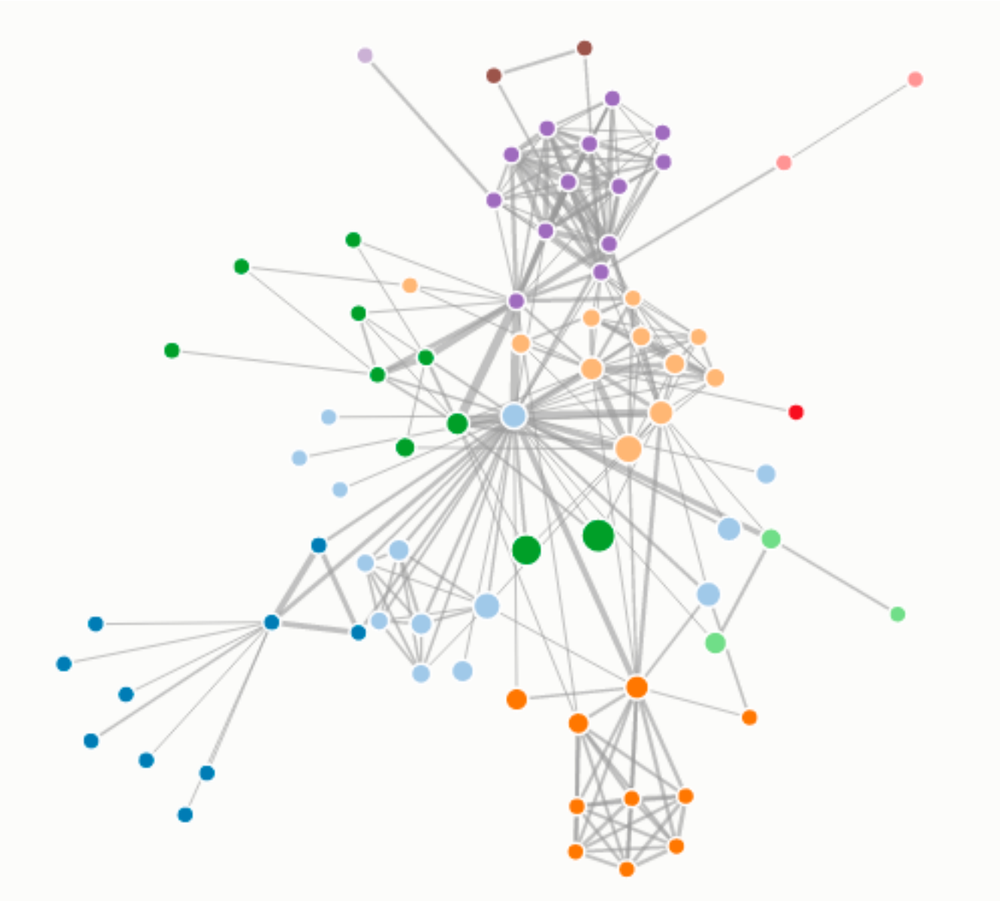




http://tulip.labri.fr/TulipDrupal/?q=node/351
http://tulip.labri.fr/TulipDrupal/?q=node/371

Idiom: **Fisheye Lens**

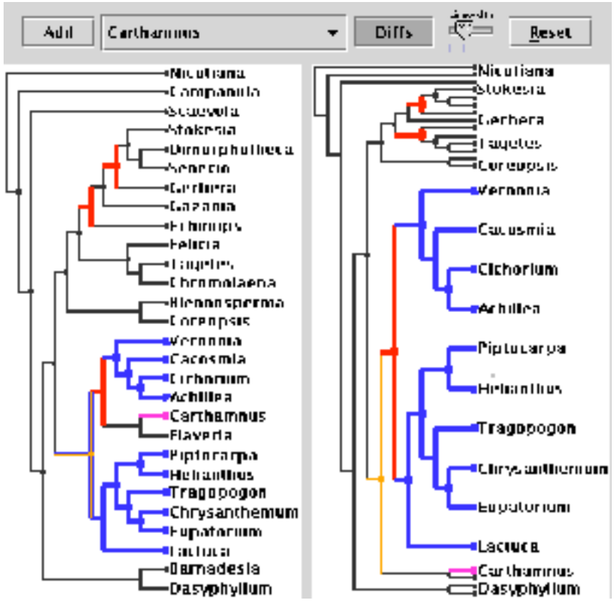
System: **D3**



[D3 Fisheye Lens](https://bost.ocks.org/mike/fisheye/)

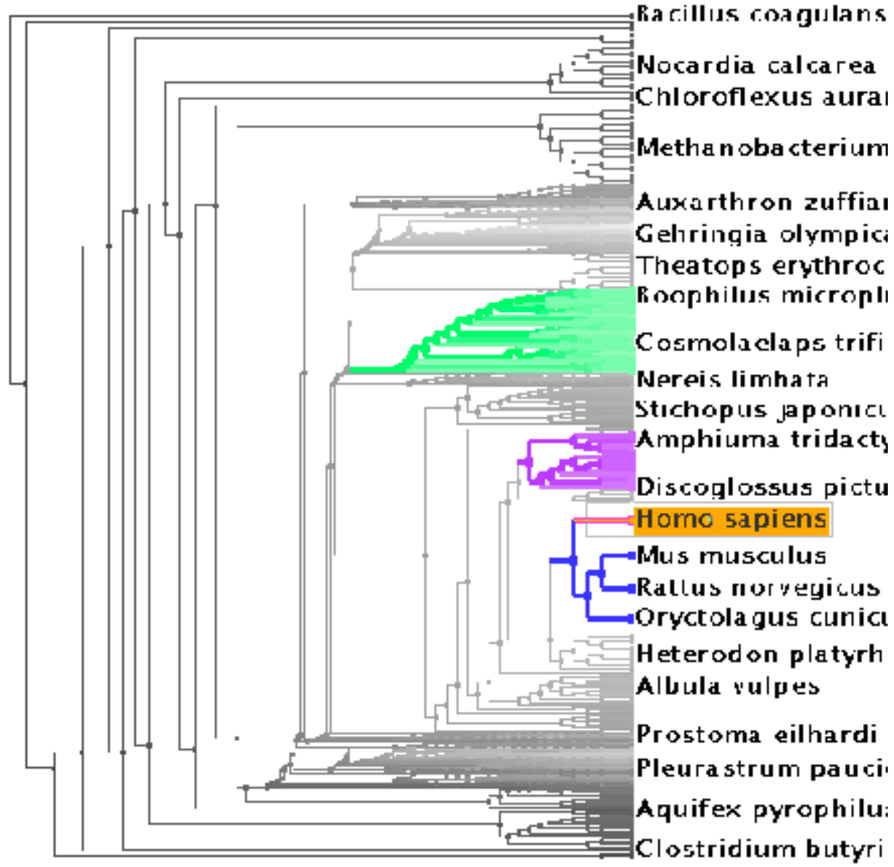
Idiom: **Stretch and Squish Navigation**

- distort geometry
  - shape: rectilinear
  - foci: multiple
  - impact: global
  - metaphor: stretch and squish, borders fixed



https://youtu.be/GdaPj8a9QEo

System: **TreeJuxtaposer**

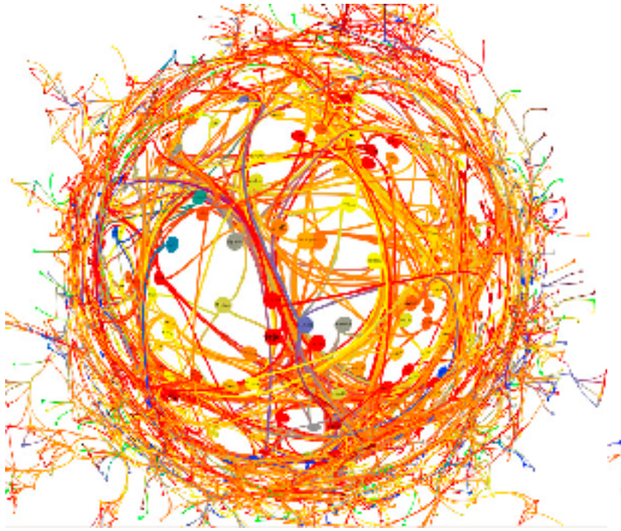


[TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context With Guaranteed Visibility. Munzner, Guimbretiere, Tasiran, Zhang, and Zhou. ACM Transactions on Graphics (Proc. SIGGRAPH) 22:3 (2003), 453– 462.]

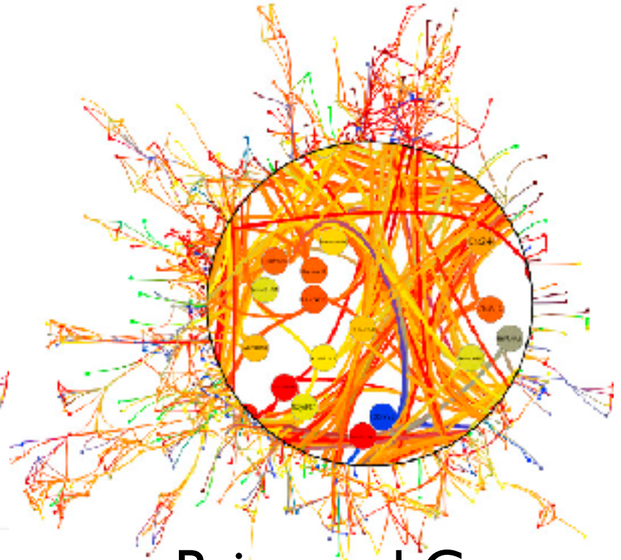
Distortion costs and benefits

- benefits
  - combine focus and context information in single view
- costs
  - length comparisons impaired
    - network/tree topology comparisons unaffected: connection, containment
  - effects of distortion unclear if original structure unfamiliar
  - object constancy/tracking maybe impaired

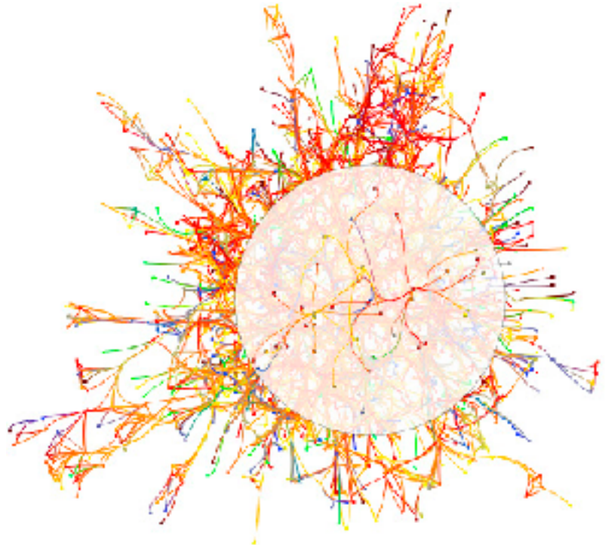
fisheye lens



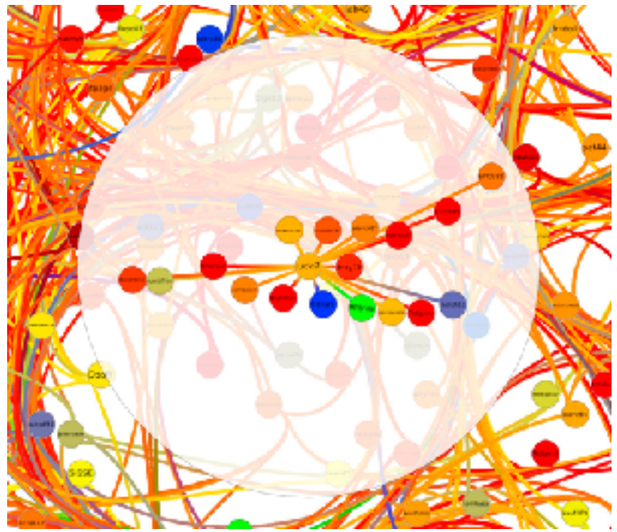
magnifying lens



neighborhood layering



Bring and Go



[Living Flows: Enhanced Exploration of Edge-Bundled Graphs Based on GPU-Intensive Edge Rendering. Lambert, Auber, and Melançon. Proc. Intl. Conf. Information Visualisation (IV), pp. 523–530, 2010.]

Credits

- Visualization Analysis and Design (Ch 13, 14)
- Alex Lex & Miriah Meyer, http://dataviscourse.net/