CSCI 4550/6550 Interactive Visualization

https://www.cs.rpi.edu/~cutler/classes/visualization/S24/

Lecture 5: Parallel Coordinates & High Dimensional Data

Today

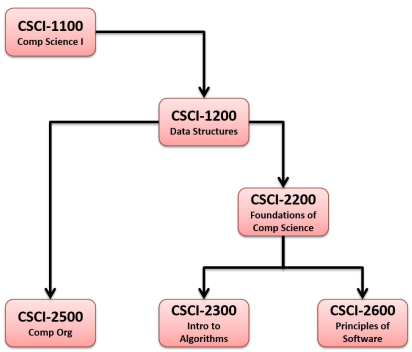
- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Worksheet ~~Grades~~ Feedback

- Worksheets are usually *Design Exercises*
  - Use what you've learned from the readings
  - Exact details usually aren't important
  - Sketch your hypothesis for what the data/visualization will look like (fake the data)
- Yes, it will be part of the grade:
  - "Use Color"
  - "Team of 2" (or 3)
  - *Read the instructions carefully!*
- There will be a curve on everything
  - Curve will be shown for each item in Rainbow Grades
  - I expect all students with *good, consistent attendance and effort to get an A or a B* (I use grade modifiers +/-)

HW 3: Graphviz & RPI CSCI Course Data

- Learn GraphViz: Open-source software for automated graph drawing
- Parse JSON data from QuACS




```
{
  "crse": 4550,
  "id": "CSCI-4550",
  "sections": [
    {
      "act": 39,
      "attribute": "Communication Intensive",
      "cap": 45,
      "credMax": 4.0,
      "credMin": 4.0,
      "crn": 93902,
      "crse": 4550,
      "rem": 6,
      "sec": "01",
      "subj": "CSCI",
      "timeslots": [
        {
          "dateEnd": "04/24",
          "dateStart": "01/08",
          "days": [
            "T",
            "F"
          ],
          "instructor": "Barbara Cutler",
          "location": "Troy Building 2012",
          "timeEnd": 1550,
```

Graphviz - Installation & Documentation

- File format
- Bipartite - not obvious how to do
- Building & linking pretty straightforward
- Include html to do crazy rendering… web-embeddable?
- Documentation for attributes is pretty good
- Guide pdf is good level of detail
- Gallery of examples with *dot* files is nice

Pre-Processing the CSCI Course Data

- Cleanup / Fix obvious typos
- Suggested simplifications:
  - Only include courses commonly taken by CSCI majors
  - Only include catalog courses with permanent numbers (if 2nd digit is a ‘9’ it’s not a permanent number)
  - Only include primary course #, no cross listings
  - Simplify prereqs (omit/ignore/replace co-reqs, “or” options)
  - Only include faculty who have taught recently
  -

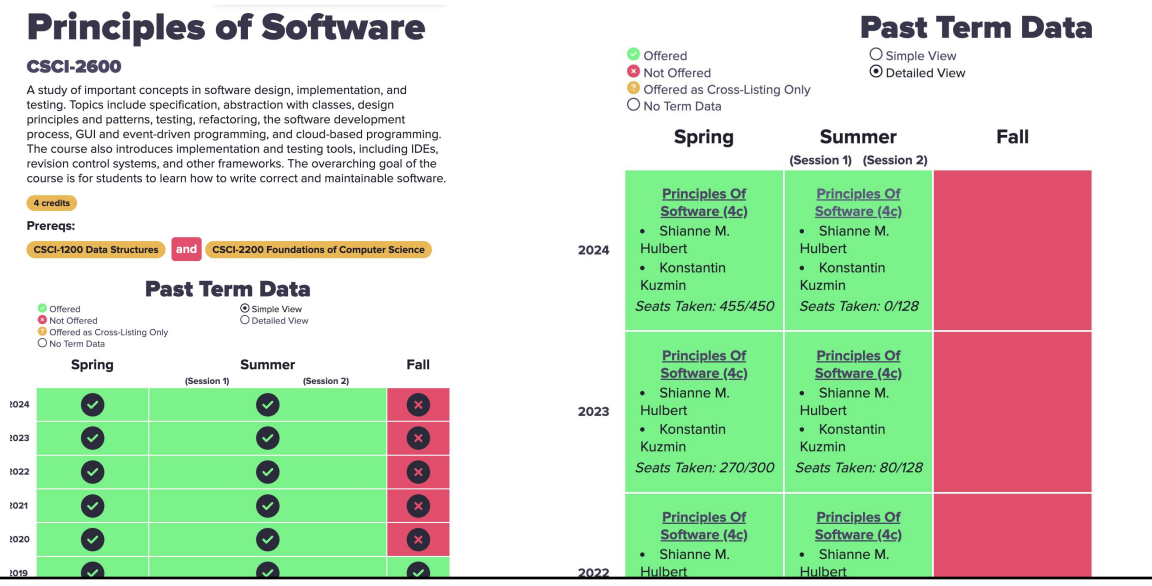
Prerequisites Graph & Data Pre-Processing

- What challenges do you expect in working with the QuACS data?

Prerequisites Graph

- Rooted tree w/ consistent orientation for edge direction? Left to right vs. top to bottom?
  - How to handle courses with multiple prereqs at different “depth”?
  - Can / should redundant prereq edges be simplified / removed?
- Max depth (# of semesters)? Max width (choices per semester)?
- How to visualize which semester is a course offered?
- Optimization of Layout
  - We hope related courses / concentrations naturally cluster.
  - What if typical senior class display too “early”?
  - What about disconnected nodes (w/ no common prereqs)? Automatic layout might fail.
- Future work:
  - Make it dynamic/interactive!
  - How to handle AP / transfer credit? How to handle dual majors?
  - How to handle schedule conflicts? How to handle requirements that change by entry year?
  - How to visualize recommended workload balance, informal/suggested prereqs, informal course recommendations from peers & advisors

Historical Course Data from QuACS / Quatalog



Principles of Software

CSCI-2600

A study of important concepts in software design, implementation, and testing. Topics include specification, abstraction with classes, design principles and patterns, testing, refactoring, the software development process, GUI and event-driven programming, and cloud-based programming. The course also introduces implementation and testing tools, including IDEs, revision control systems, and other frameworks. The overarching goal of the course is for students to learn how to write correct and maintainable software.

4 credits

Prereqs:

CSCI-1200 Data Structures and CSCI-2200 Foundations of Computer Science

Past Term Data

Offered
Not Offered
Offered as Cross-Listing Only
No Term Data

Simple View
Detailed View

	Spring	Summer (Session 1)	Summer (Session 2)	Fall
2024	✓	✓	✓	✗
2023	✓	✓	✓	✗
2022	✓	✓	✓	✗
2021	✓	✓	✓	✗
2020	✓	✓	✓	✗
2019	✓	✓	✓	✓

Past Term Data

Offered
Not Offered
Offered as Cross-Listing Only
No Term Data

Simple View
Detailed View

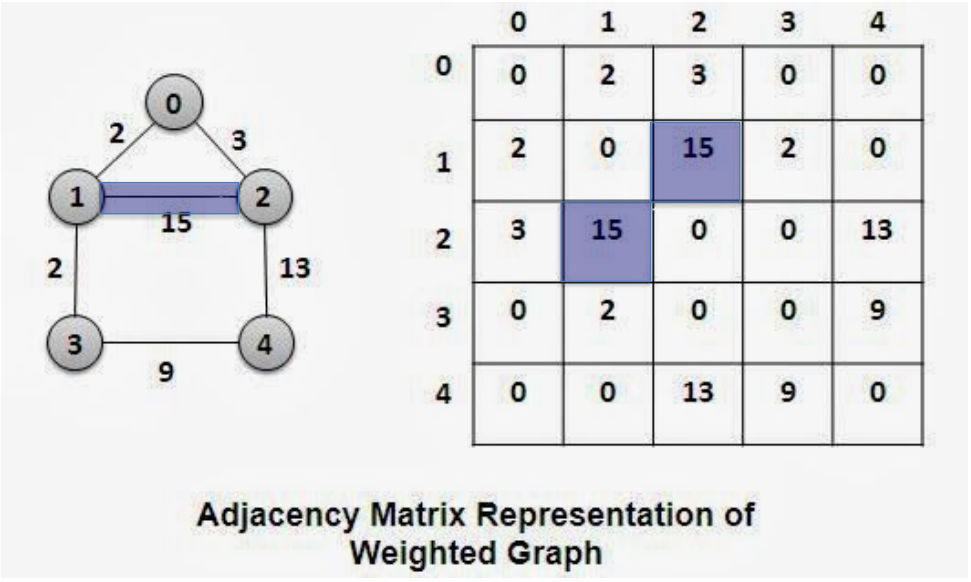
	Spring	Summer (Session 1)	Summer (Session 2)	Fall
2024	Principles Of Software (4c) • Shianne M. Hulbert • Konstantin Kuzmin *Seats Taken: 455/450*	Principles Of Software (4c) • Shianne M. Hulbert • Konstantin Kuzmin *Seats Taken: 0/128*		
2023	Principles Of Software (4c) • Shianne M. Hulbert • Konstantin Kuzmin *Seats Taken: 270/300*	Principles Of Software (4c) • Shianne M. Hulbert • Konstantin Kuzmin *Seats Taken: 80/128*		
2022	Principles Of Software (4c) • Shianne M. Hulbert	Principles Of Software (4c) • Shianne M. Hulbert		

“Who Teaches What” Graph

- What questions do you have about the data?

“Who Teaches What” Graph

- What questions do you have about the data?
  - How many classes has Goldschmidt taught?
  - What teachers have a small # of different classes?
  - What teachers “own” specific classes?
  - “6 degrees of Goldschmidt”
  - Which classes are juggled between many teachers?
  - Are the concentrations/areas naturally grouped by professor/research area?
- How can automatic layout help?
- What do we hope automatic layout will do?



	0	1	2	3	4
0	0	2	3	0	0
1	2	0	15	2	0
2	3	15	0	0	13
3	0	2	0	0	9
4	0	0	13	9	0

Adjacency Matrix Representation of Weighted Graph

http://www.thecrazyprogrammer.com/2014/03/representation-of-graphs-adjacency-matrix-and-adjacency-list.html

Revising the Visualization

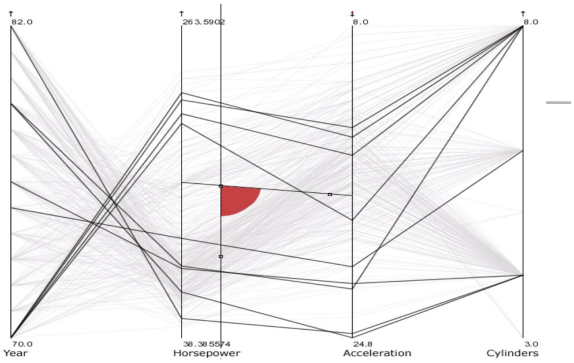
- Have modest goals for your initial visualization
- As time allows, think about how you can revise & improve:
  - Overall node/edge layout
  - Use of Color
  - Node shape
  - Directed/Undirected Edges
  - Edge weight/thickness
  - Labels on Nodes / Edges
  - Annotations on the visualization as a whole
  - Legends
  - …

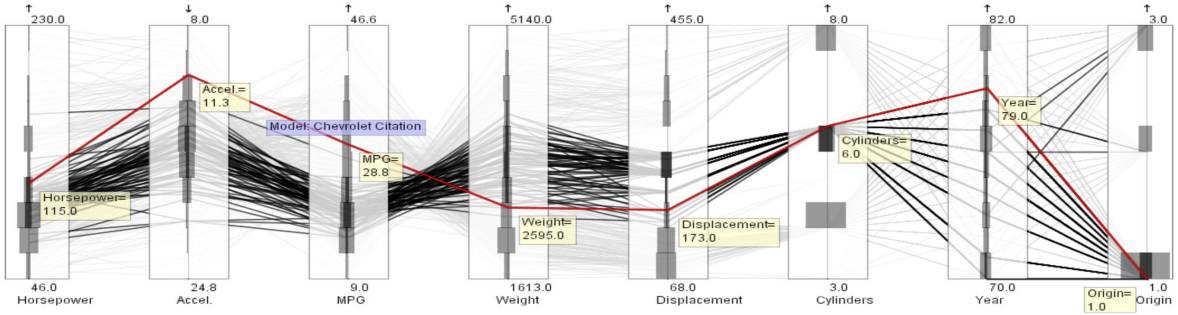
Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Reading for Today

"Angular Brushing of Extended Parallel Coordinates", Hauser, Ledermann, and Doleisch, InfoVis 2002





Parallel Coordinates Overview

- Polyline for every data point
- Real-time reordering of axes
- Interactive brushing of a single axis to mark subset of data to display
- Histogram on axes – very effective!
- Composite brushes (and/or on multiple axes)
  - Important for handing complex user queries of the data
- Hierarchical clustering for really big datasets
- 3D extruded wings
- Higher order splines rather than polylines

"Angular Brushing of Extended Parallel Coordinates", Hauser, Ledermann, and Doleisch, InfoVis 2002

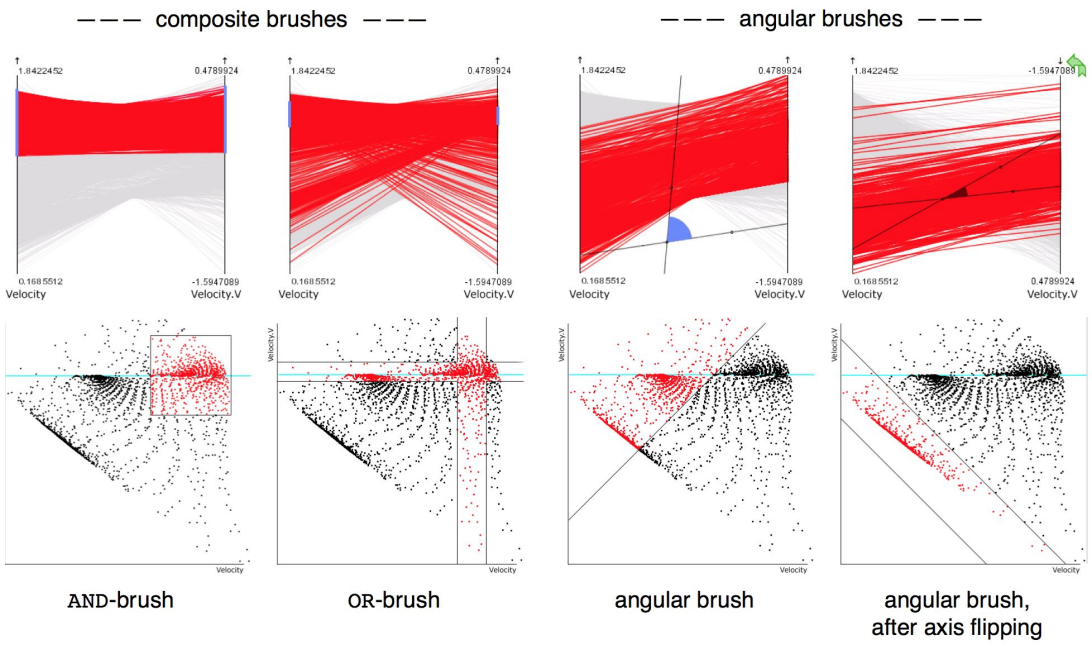
Contributions

- Angle Brushing
  - Emphasizes parallel coordinates strength at expressing relationships between coordinate axes
  - Slope indicates positive or negative correlation
  - Outliers really stand out
  - Really desirable tool!
  - Obvious, once someone else thought of it!
- Smooth Brushing (continuous, not binary in/out, fuzzy logic)
- Multiple brushes, automatically named composite brushes, linking with scatterplot visualization
  - Especially neat to use this to explain how angle brushing works!
- Flexible layout – reorder/add/delete/flip axes, scale & pan (helps with brushing actions)
- Graphics tricks for reasonable performance with 10,000-1,000,000 polylines

"Angular Brushing of Extended Parallel Coordinates",
Hauser, Ledermann, and Doleisch, InfoVis 2002

- Brush ≠ thicken the line (not explained in paper)
- Focus + context: applies across all visualization types
- Well written, intuitive diagrams illustrated text
- Would like to see user study on this work
  - After adding these new features… Will it still be intuitive?
- Using a single dataset is good (for consistency and to focus on the contributions) and bad (what about datasets with binary data or outliers?)
- I hadn't seen this before (and I was convinced!)
- Visualization is a young field!
- Is this too "incremental"??
- What are the current (easy-to-use) toolkits to create new parallel coordinates visualization?

"Angular Brushing of Extended Parallel Coordinates",
Hauser, Ledermann, and Doleisch, InfoVis 2002



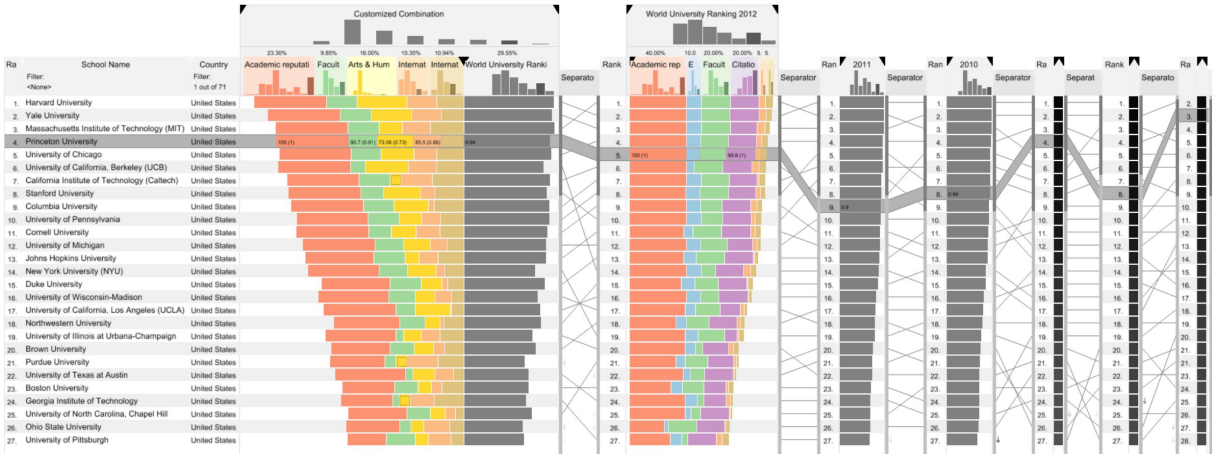
"Angular Brushing of Extended Parallel Coordinates",
Hauser, Ledermann, and Doleisch, InfoVis 2002

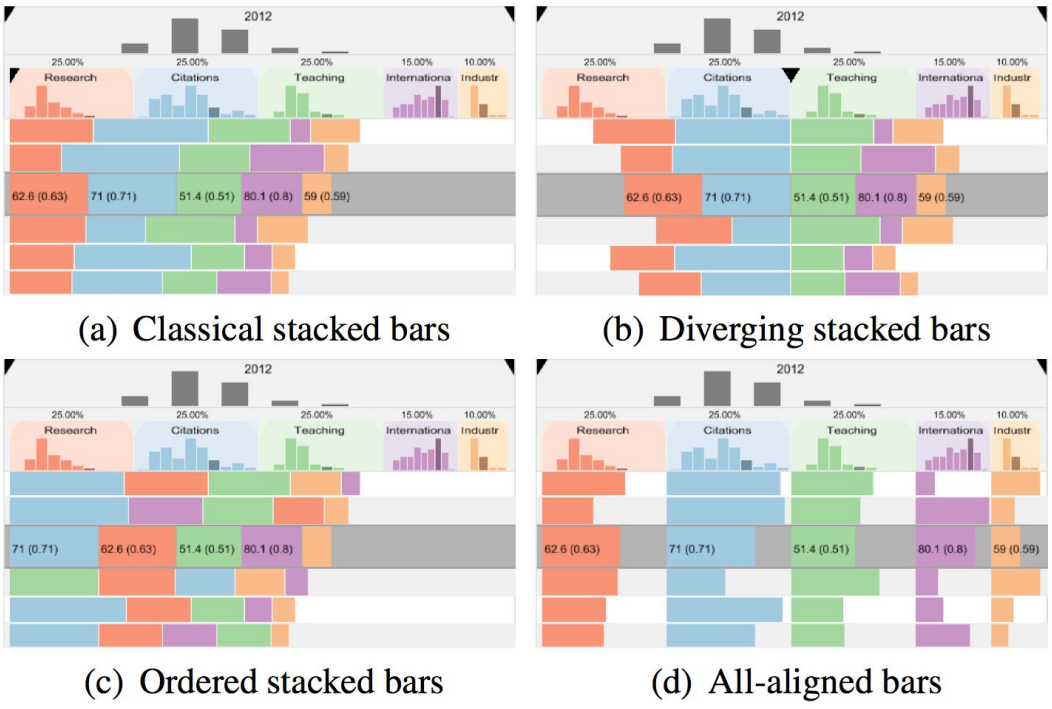
Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Reading for Today

"LineUp: Visual Analysis of Multi-Attribute Rankings", Gratzl, Lex, Gehlenborg, Pfister and Streit, IEEE INFOVIS 2013





(a) Classical stacked bars

(b) Diverging stacked bars

(c) Ordered stacked bars

(d) All-aligned bars

“LineUp: Visual Analysis of Multi-Attribute Rankings”

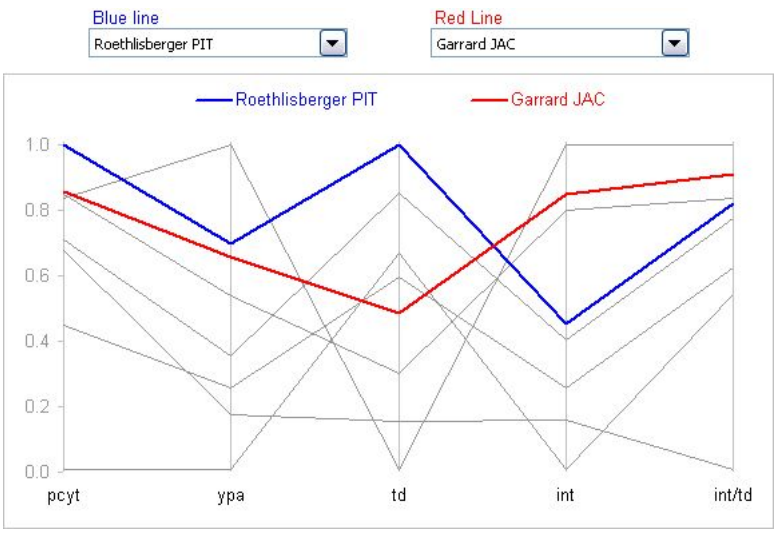
- For this type of data, static visualization is not sufficient – interactivity is essential
- Leverages and combines a wealth of prior visualization designs
- Major contribution: User/Viewer Customization
  - Too bad the tool is not (currently?) available on the web and the work hasn’t continued(?)
- Doesn’t seem difficult to implement, the UI design is the contribution
- Works well for clean data where all entries have similarly good scores. What about datasets where entries have much more varied data per column?
- Well written: laid out 10 goals/requirements of the tool, Thoughtful & detailed related work discussion

Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Parallel Coordinates

Use Color!



https://peltiertech.com/Excel/Charts/ParallelCoord.html

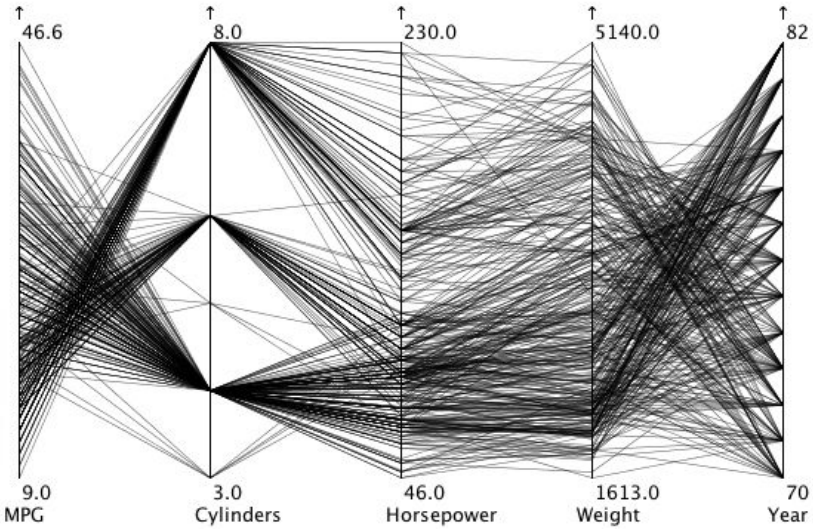
Designing Visualizations using Parallel Coordinates

- How many dimensions (vertical axes)?
- In what order should the axes appear?
- Which direction should each axis run (up or down?)
- Should the axis scale be linear or log scale?
- How many data points (lines)?
- How could color, line thickness, etc. be used to highlight patterns in the data?
- Use as data exploration or debugging tool (iterate on the design/analysis)? Or as final visualization?
- How to use interaction? e.g., selection or filtering.

Designing Visualizations using Parallel Coordinates

- How many dimensions (vertical axes)? *About 10, maximum*
- In what order should the axes appear? *It depends.*
  Are there highly correlated axes? What do you want to compare?
- Which direction should each axis run (up or down?)
  Be intuitive, but flip one axis if there is an inverse correlation.
- Should the axis scale be linear or log scale? *It depends.*
- How many data points (lines)? *If you have alot of data,*
  you will need interactive tools for selection & highlighting.
- How could color, line thickness, etc. be used to highlight patterns in the data?
- Use as data exploration or debugging tool (iterate on the design/analysis)?
  Or as final visualization? *Both!*
- How to use interaction? e.g., selection or filtering.

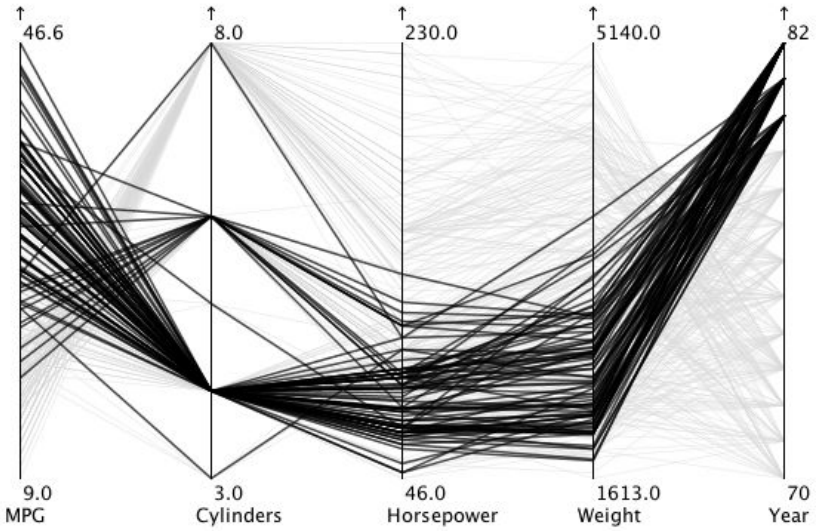
Parallel Coordinates



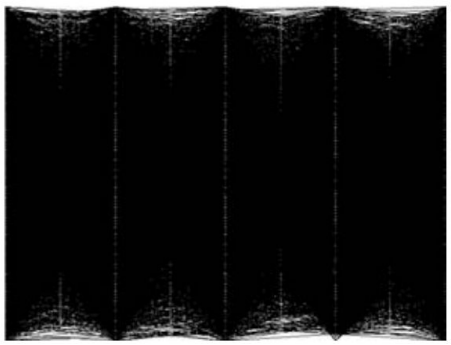
http://eagereyes.org/techniques/parallel-coordinates

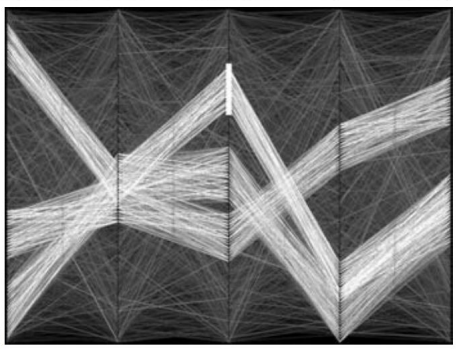
Parallel Coordinates

Use interaction & selection

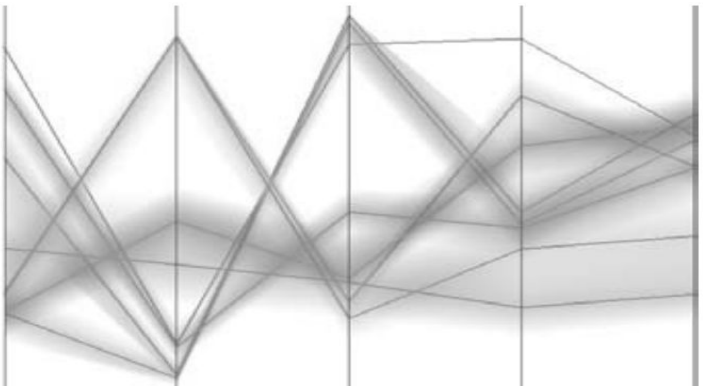


http://eagereyes.org/techniques/parallel-coordinates



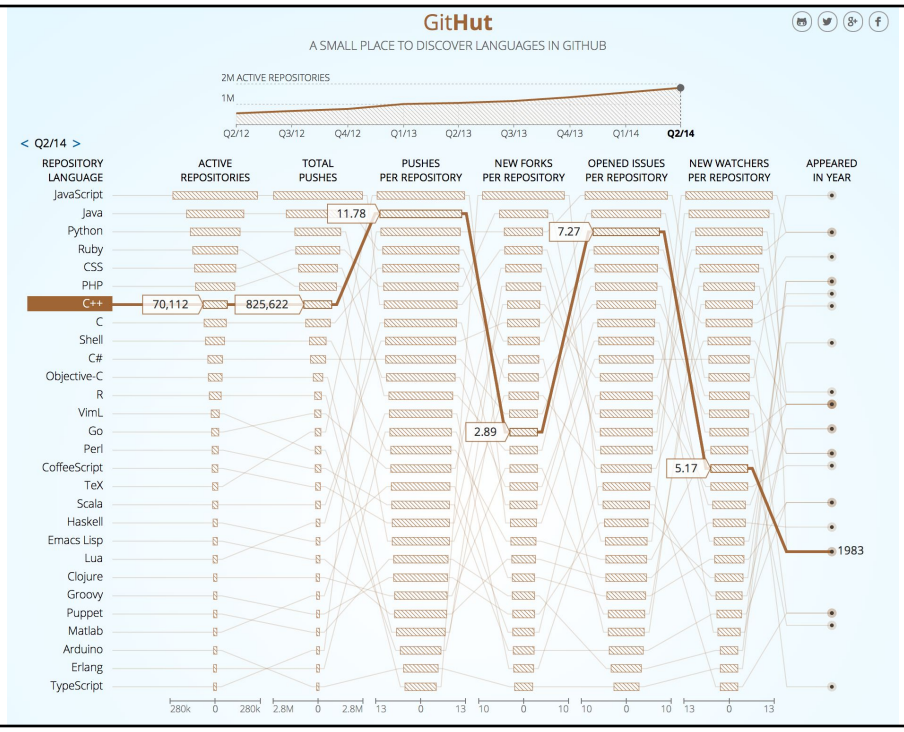


"Uncovering Clusters in Crowded Parallel Coordinates Visualizations", Artero, Ferreira de Oliveira, & Levkowitz, InfoVis 2004



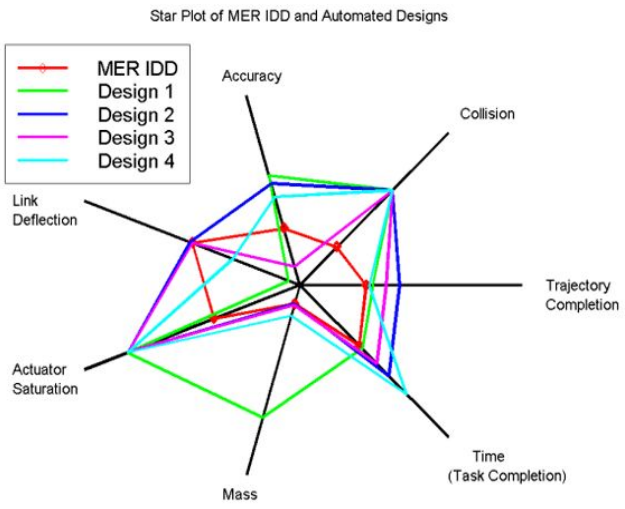
- Parallel coordinates becomes less effective when data is high dimensional and/or quantity of data points becomes huge
  - Visual crowding is not unique to parallel coordinates
- Synthetic test dataset with 5 clusters (varying sizes) embedded in random noise
  - Goal is to filter signal from this noise!
  - Keep all data for context, but de-emphasize what is not in focus (focus + context theme of visualization)
- Good use of illustrations and comparison to previous Parallel Coordinates methods
  - Nice that same dataset was used for all examples
- Paper only used B&W, adding color would be helpful too
- Can these clusters be identified from a scatterplot matrix? Or by the interactive brushing methods?
- Does re-ordering the axes make these IPC Frequency Plot or IPC Density Plot more or less effective?

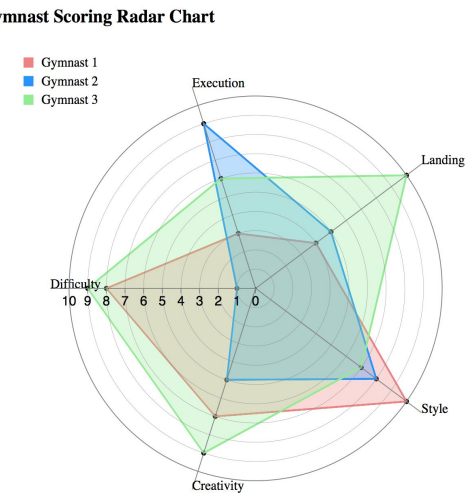
Uncovering Clusters in Crowded Parallel Coordinates Visualizations",
Artero, Ferreira de Oliveira, & Levkowitz, InfoVis 2004



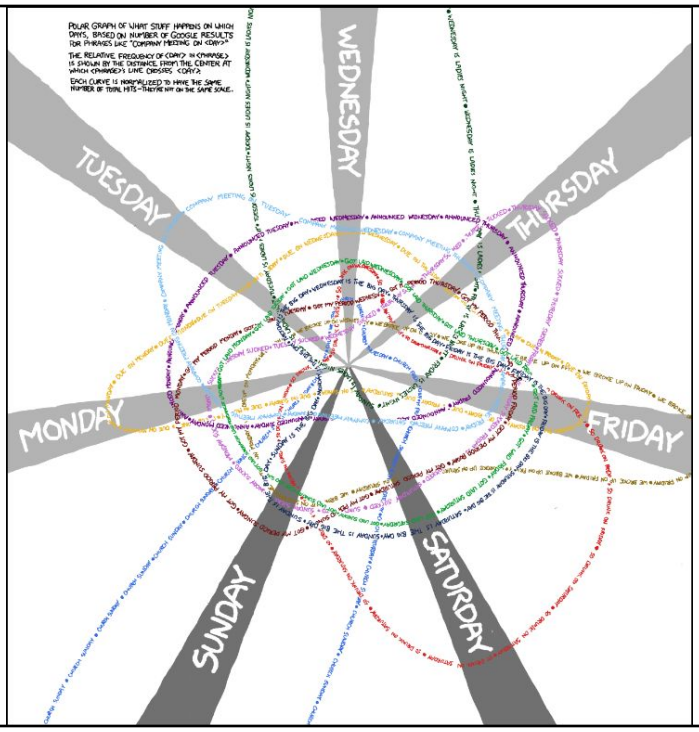
http://githut.info/

Radar Chart *(a.k.a. web chart spider chart, star chart, star plot, cobweb chart, irregular polygon, polar chart, kiviat diagram)*





From NASA: http://en.wikipedia.org/wiki/File:MER_Star_Plot.gif http://www.cs.middlebury.edu/~candrews/showcase/infovis_techniques_s16/radar_chart/



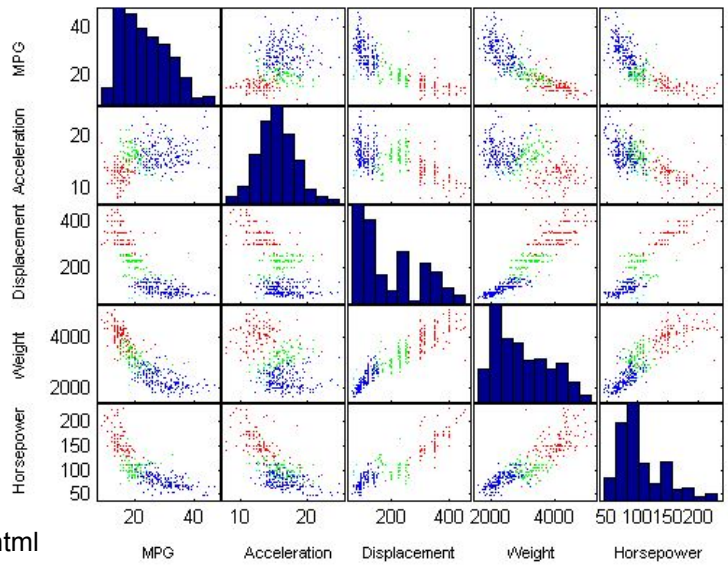
http://xkcd.com/930/

Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Scatterplot Matrix

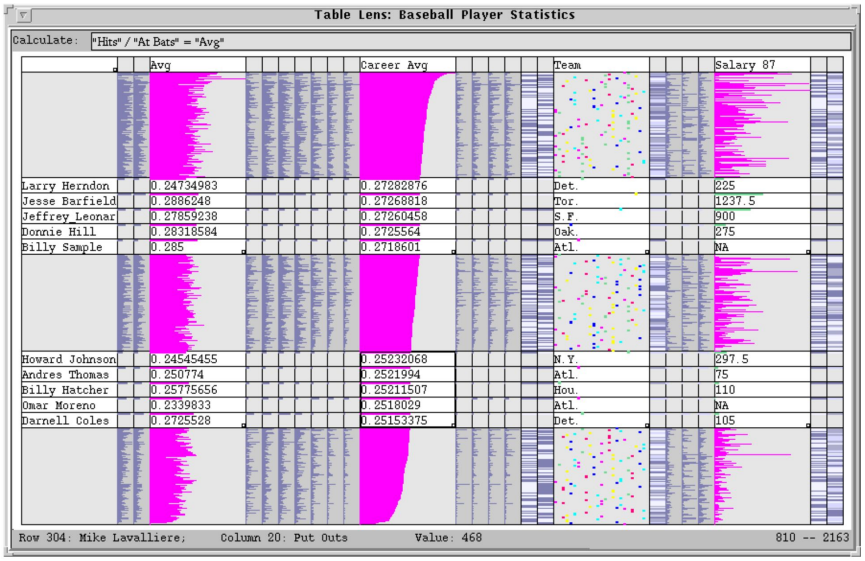
Use color to sneak in one more dimension (in this example… # of cylinders)



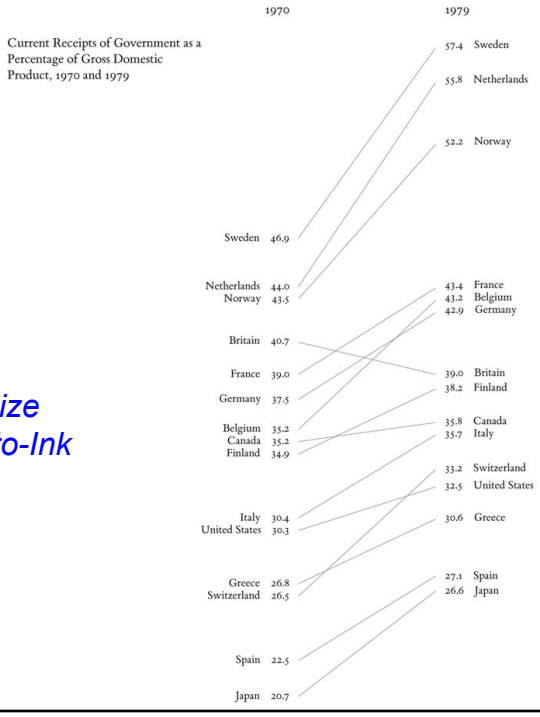
http://www.mathworks.com/help/stats/examples/visualizing-multivariate-data.html

“The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information” Ramana Rao, SIGCHI Conference on Human Factors in Computing Systems 1994

Focus+context

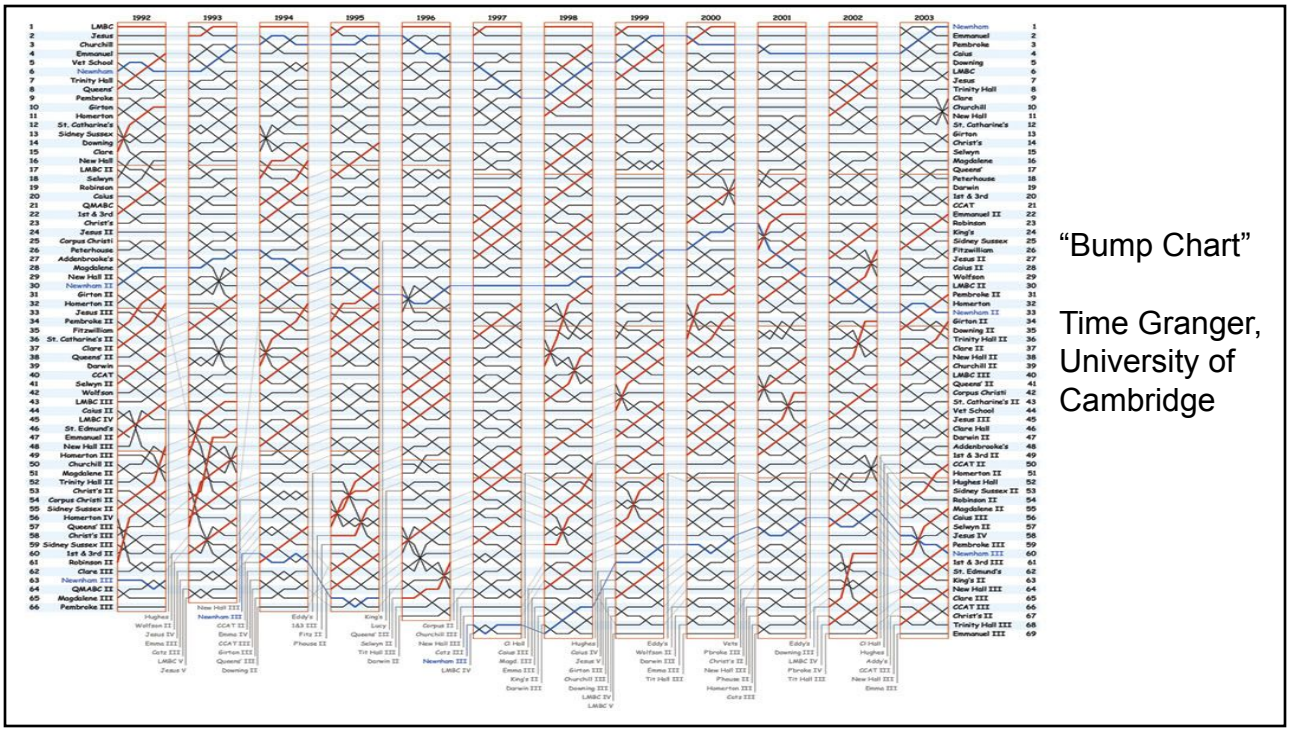


Optimize Data-to-Ink ratio!



“Slope graph”

from:
Edward Tufte
The Visual Display of Quantitative Information 1983



"Bump Chart"

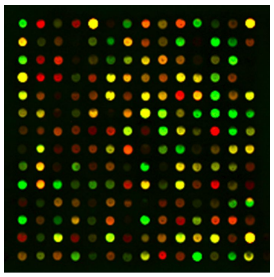
Time Granger, University of Cambridge

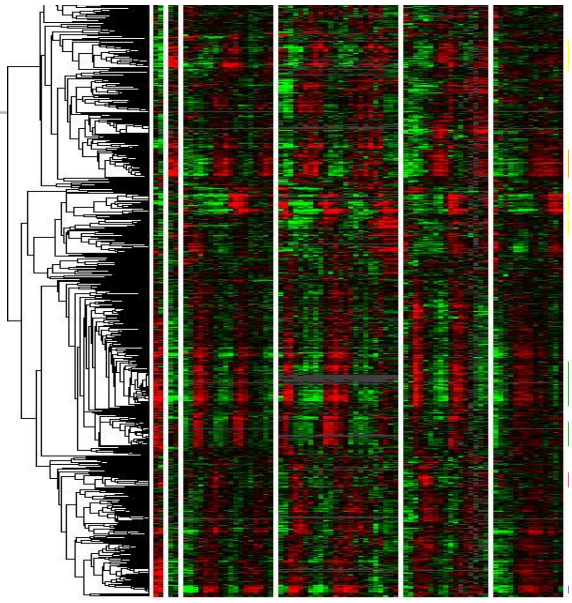
General Massive Data Visualization Tips

- Use your spatial real estate effectively
  - sort, organize
  - cluster, separate
  - layout, relative distances
  - high resolution, large format media
- Color, Contrast, Intensity, Transparency
  - layering, overlapping
- Use interactivity wisely

http://www.auscillate.com/itp/listview/

Gene Expression





- Expression level for hundreds of genes
- For many trials, different individuals/conditions)
- *Automatically discover correlations: genes that commonly work together or in opposition… How?*

http://www.imbb.forth.gr/people/poirazi/researchEP.html

http://www.bioss.ac.uk/~dirk/essays/GeneExpression/bayes_net.html

RankExplorer: Visualization of Ranking Changes in Large Time Series Data

Conglei Shi, Weiwei Cui, Shixia Liu, *Member, IEEE*, Panpan Xu, Wei Chen, and Huamin Qu, *Member, IEEE*

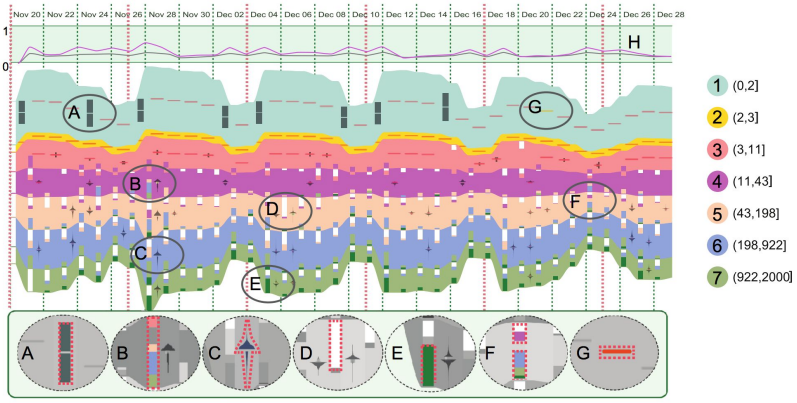
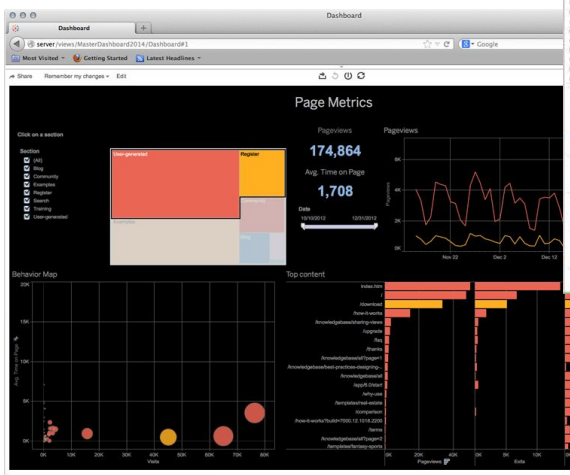


Fig. 1. RankExplorer visualization of the top 2000 Bing search queries from Nov. 20 to Dec. 29 in 2011. All queries are divided into seven categories. The width of each layer at a time point encodes the total query count at that time. The color bar and glyphs encode the content changes in each ranking category. From the color bar, we can observe: 1) the change between layers (the bar segments with the colors of other layers in B and F); 2) new queries coming in (the white segment in D); 3) recurring queries (the dark green segment in E). From the changing glyphs, we can see: 1) a non-change pattern (only red line in G); 2) a swap pattern (the two equal-height segments in A represent that the two queries swap their rankings); 3) a shift pattern (the increasing part is significantly larger than the decreasing part in C). From the trend curve (H), we can see the degree of ranking change over time.

Tableau

http://www.tableausoftware.com/



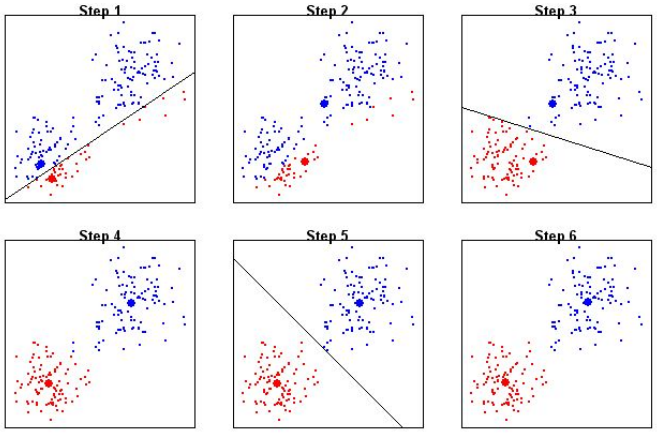
Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

K-means Clustering

For a set of 2D/3D/*n*D points:

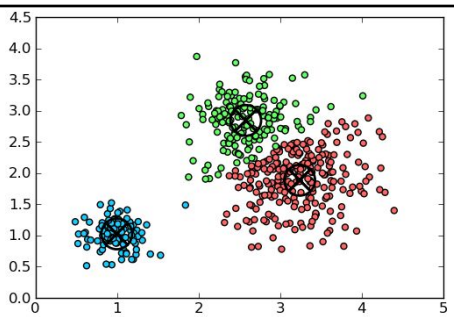
- Choose *k*, how many clusters
- Select k points from your data at random as initial "team" representatives
- Every other point determines which team representative it is closest to and joins that team
- Average the positions of all team members, this is the team's new representative
- Repeat 3-5 times until change < threshold



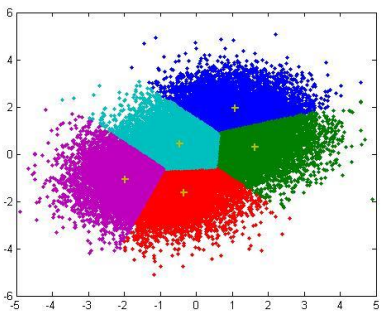
http://astrostatistics.psu.edu/su09/lecturenotes/clus2.html

K-means Clustering

- *Assumption: We know what number to select for k (k = # of clusters)*
- Works quite well, when the data can be meaningfully classified (and *k* was well selected)
- With dense data, output is visually similar to a Voronoi diagram (k-Means chooses the data points that define the cells)



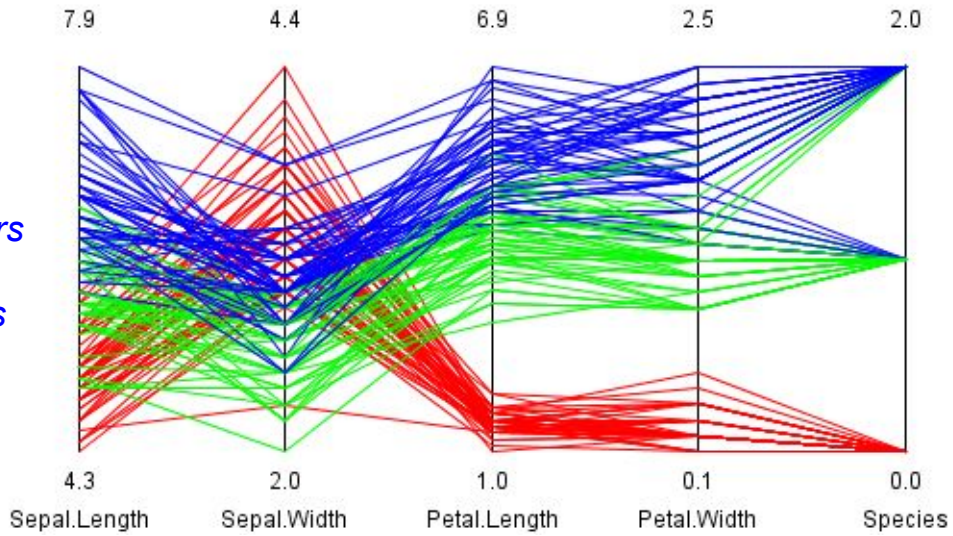
http://blog.mpacula.com/2011/04/27/k-means-clustering-example-python/



"Efficient K-Means Clustering using JIT" Yi Cao

Clustering & Parallel Coordinates

Color applied to data clusters discovered by k-means



http://wanderinformatiker.at/unipages/ParCoord/clustering_en.html

K- means Implementation *Efficiency & Efficacy*

Determine your *distance function*

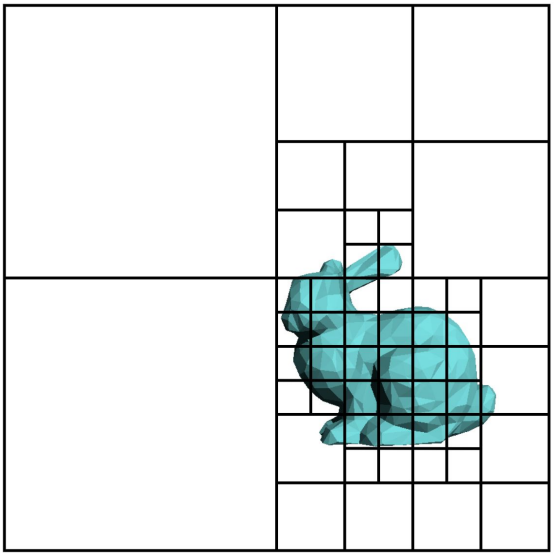
- In spatial datasets, often just be Euclidean distance
  - Maybe also add in surface normal, etc.
- *Determine relative weighting / importance of different dimensions*
  - How to combine unrelated units - can convert to % of range
  - Problematic when values are binary

Finding nearest neighbors can be expensive

- Last lecture: Computational Geometry Closest Pair of Points
- Another option: *Use a spatial data structure*!

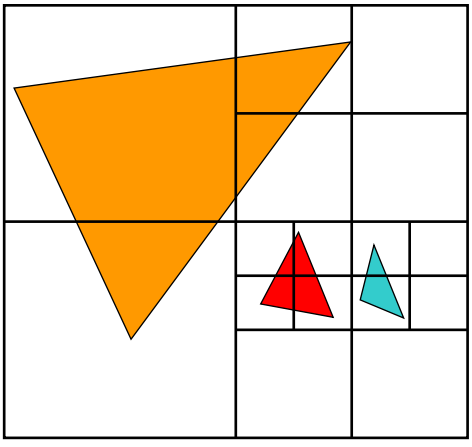
Quadtree / Octree: Bisect every dimension

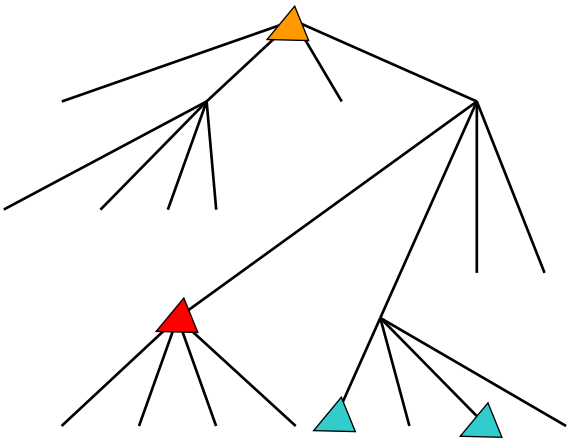
- Subdivide until each cell contains no more than *n* elements, or maximum depth *d* is reached



Quadtree / Octree: Bisect every dimension

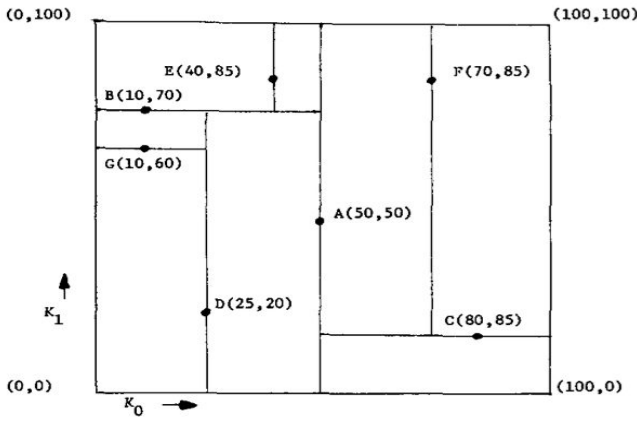
- Primitives can live at intermediate levels, or be pushed to lowest level of grid

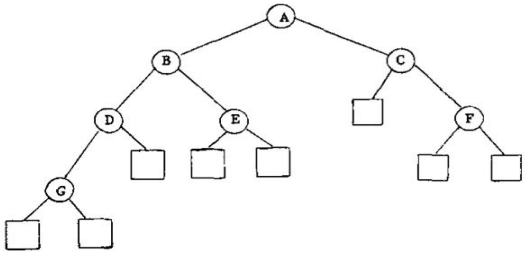




k-D Tree *split one dimension, aim for equal # on each side*

- "Multidimensional Binary Search Trees Used for Associative Searching", Bentley, Communications of the ACM, 1975





Today

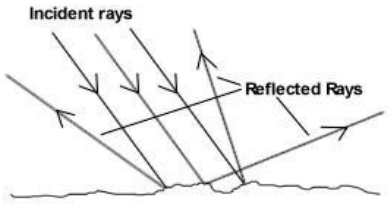
- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

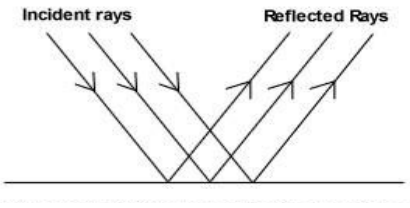
Physically-Based Ray Tracing of Materials

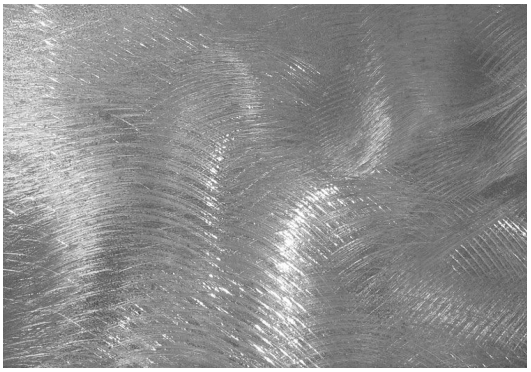
Ideal diffuse (Lambertian)
e.g. chalk, matte paint

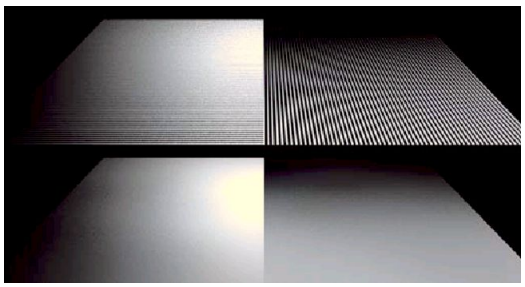
Non-ideal reflectors
"glossy"

Ideal specular
(mirror)







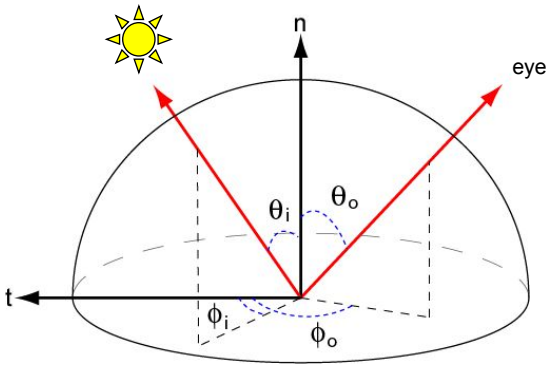


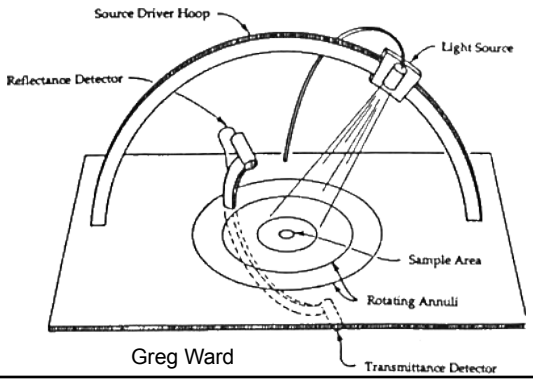
"Predicting reflectance functions from complex surfaces", Westin et al. SIGGRAPH 1992

BRDF: Bidirectional Reflectance Distribution Function

- Ratio of light coming from one direction that gets reflected in another direction
- 4D function: incoming θ_i, φ_i outgoing θ_o, φ_o

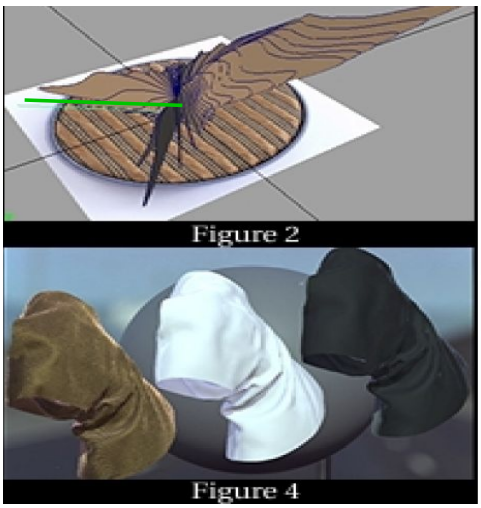
Complete material data capture: Gonioreflectometer



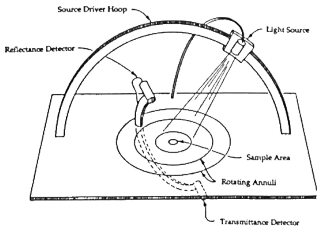


Greg Ward

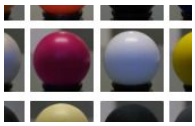
BRDFs in the Movie Industry



Measured BRDF in film production: realistic cloth appearance for "The Matrix Reloaded"
Borshukov, SIGGRAPH 2003 Sketches & Applications



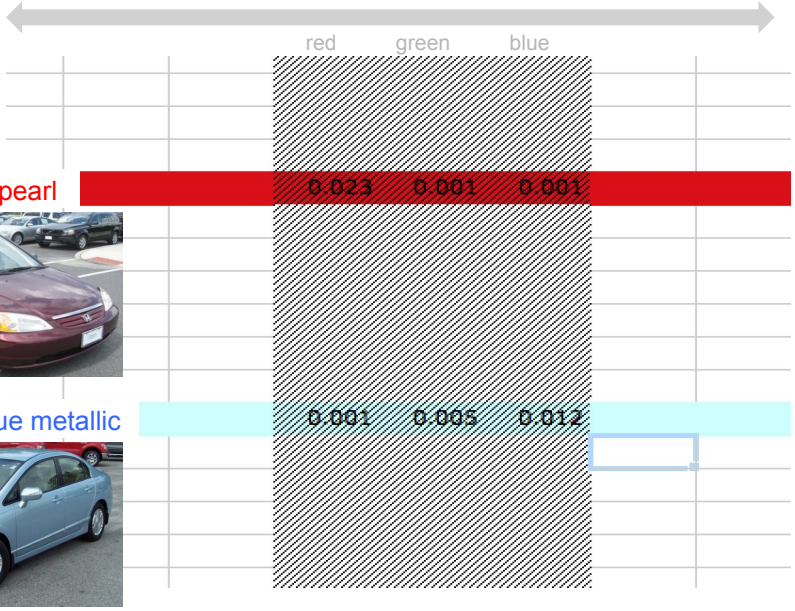
4,374,000 different measurements
(incoming θ_i ,φ_i outgoing θ_o, φ_o) + RGB per material



104 different materials

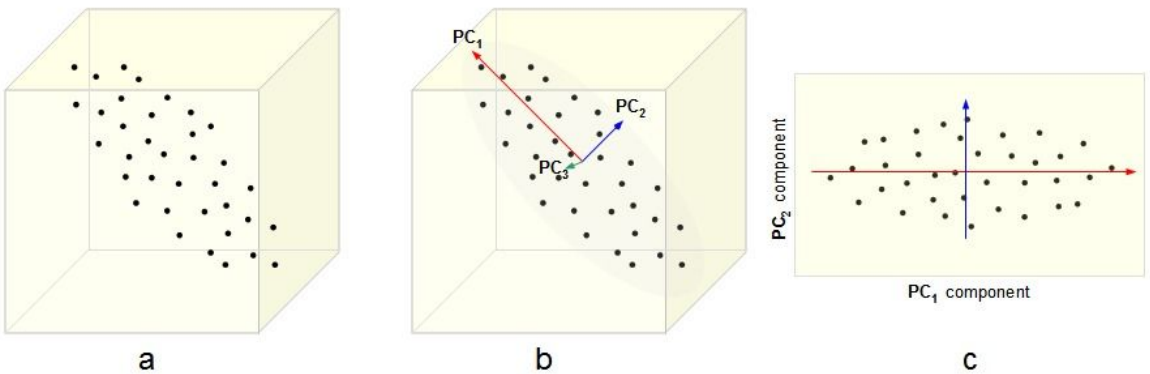
Radiant ruby pearl

Opal silver blue metallic



Principal Components Analysis (PCA)

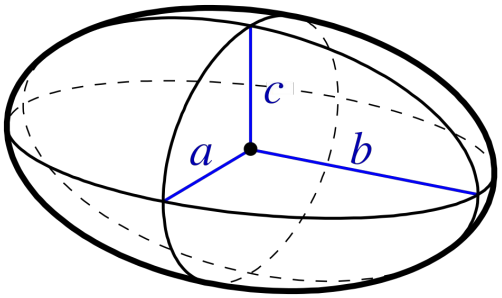
- Takes high dimensional data, where some/many axes are correlated
- Goal: Reduce to a smaller set of dimensions that are *not correlated*



http://cnx.org/content/m11461/latest/

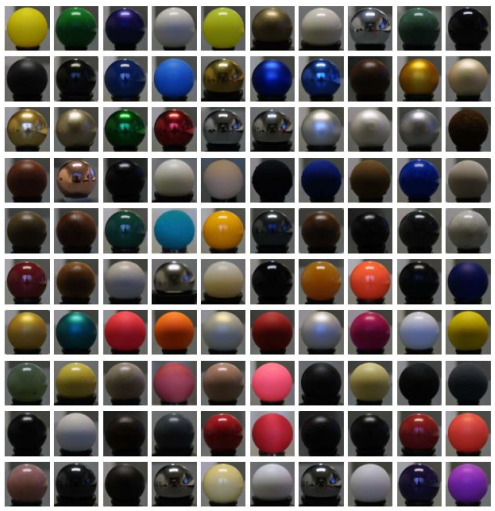
Principal Components Analysis (PCA)

- Best fit an ellipse (for 2D) or ellipsoid (for 3D+) to the data
  - Ellipsoid axes (eigenvectors) are perpendicular
- Dimensions/axes form a new basis/coordinate system
  - Each example from the original data can be defined as a linear combination of the new axes
- Simplify by omitting dimensions with the shortest ellipsoid axes (the axes with least variance, smallest eigenvalues)



https://commons.wikimedia.org/wiki/File:Ellipsoid-1-tab.svg

PCA Example: Material Reflectance Model



Matusik, Pfister, Brand, & McMillan,
"A Data-Driven Reflectance Model" SIGGRAPH 2002

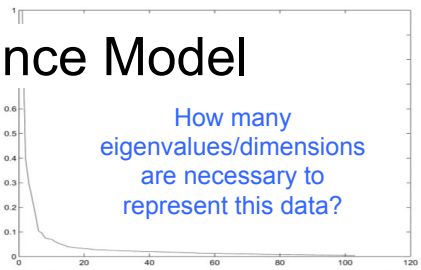
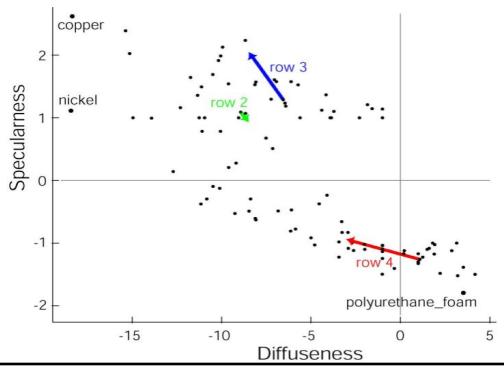


Figure 7: Plot of the eigenvalues resulting from PCA of the data set.

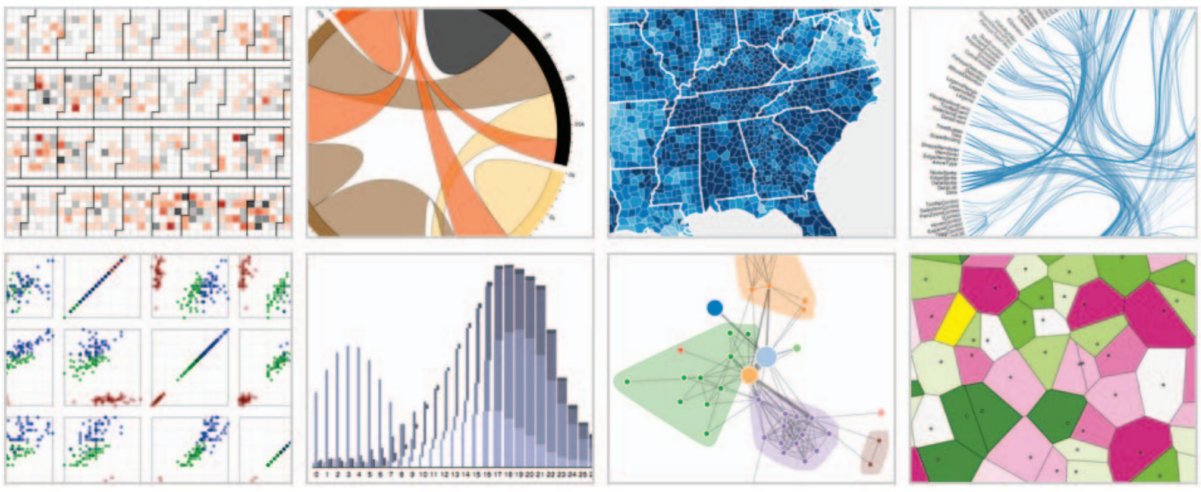


Today

- Discussion about Graphviz & Homework 3
- Readings for Today
  - “Angular Brushing of Extended Parallel Coordinates”
  - "LineUp: Visual Analysis of Multi-Attribute Rankings"
- Designing Parallel Coordinates
- Related Visualizations
- Data Simplification / Organization: k-Means Clustering
- Dimensionality Reduction: Principal Components Analysis (PCA)
- Readings for Friday

Reading for Friday *pick one*

"D3: Data-Driven Documents", Bostock, Ogievetsky, & Heer, TVCG 2011



Reading for Friday *pick one*

"Interaction Techniques for Selecting and Manipulating Subgraphs in Network Visualizations",
McGuffin & Jurisica,
IEEE TVCG 2009