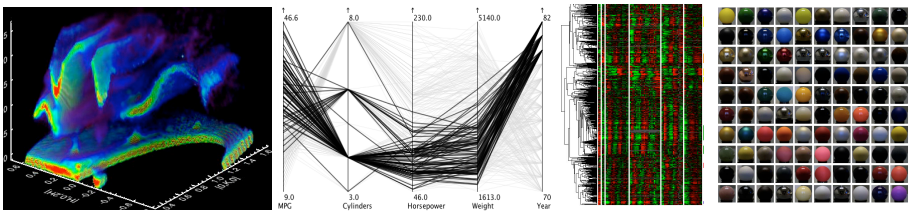
High Dimensional Data & Dimensionality Reduction

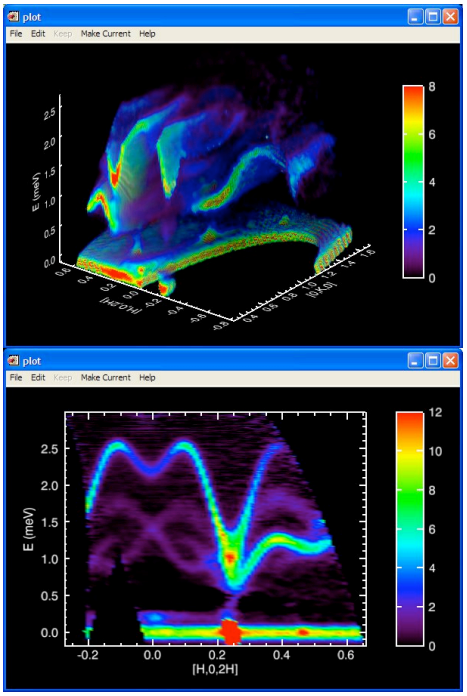


Today's Class

- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

Scientific Data

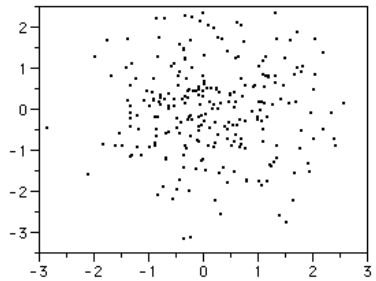
- For many 3D(or 2D) spatial locations during an experiment or simulation…
- Collect time-varying temperature, velocity, pressure, humidity, etc.

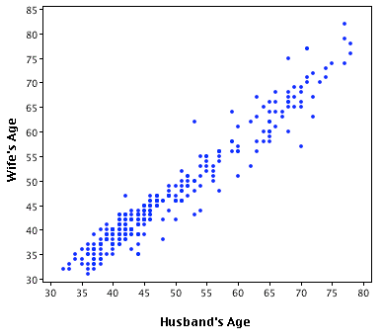


http://www.ncnr.nist.gov/dave/screenshots.html

Misc. Personal Data

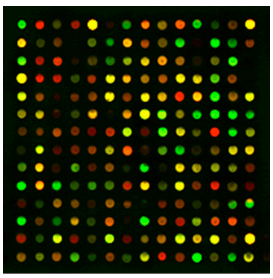
- Height, weight, eye color, phone number, address, IQ, age, cholesterol score, grade in Data Structures, etc.
- Example Hypotheses:
  - There is no correlation between height & phone number
  - There is a positive correlation between the ages of spouses
- Scatter plot: Look at 2 dimensions at a time
- Scale/units on each axis are often very different!

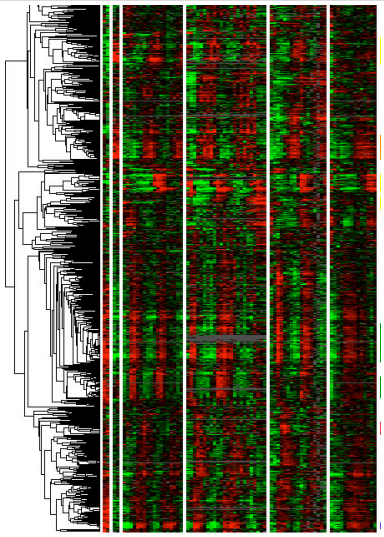




http://www.mzandee.net/~zandee/statistiek/stat-online/chapter4/pearson.html

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

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

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

>3 Dimensions vs. "Really High Dimensions"

- Obvious/intuitive dimensions
  - Position, Orientation, Time, Temperature, Color, etc.

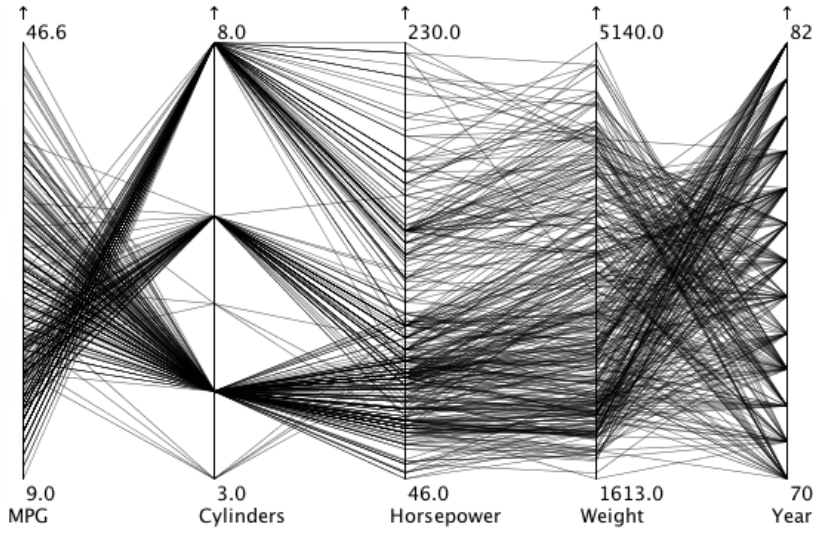
vs.

- Hundreds or thousands of attributes
  - May be floating point values or binary values
  - Stored as a "feature vectors" for each data point
  - Nearest Neighbor calculations become very expensive
  - *… visualization seems impossible*

Today's Class

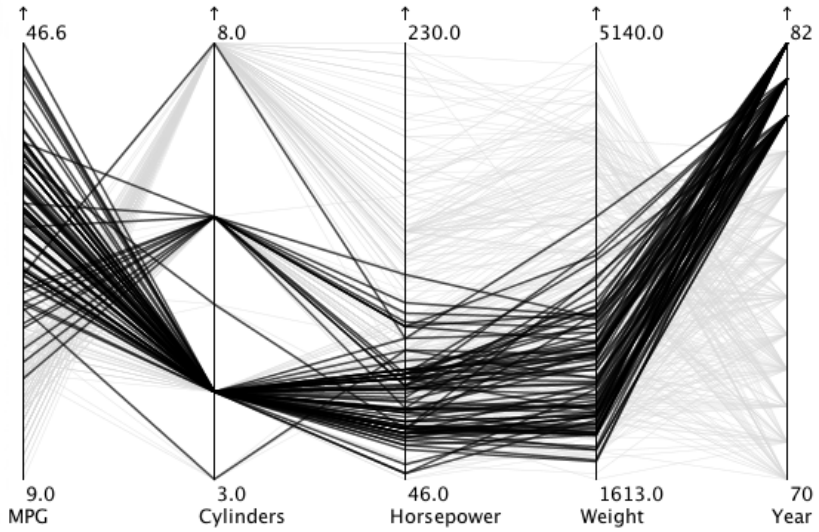
- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

Parallel Coordinates



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

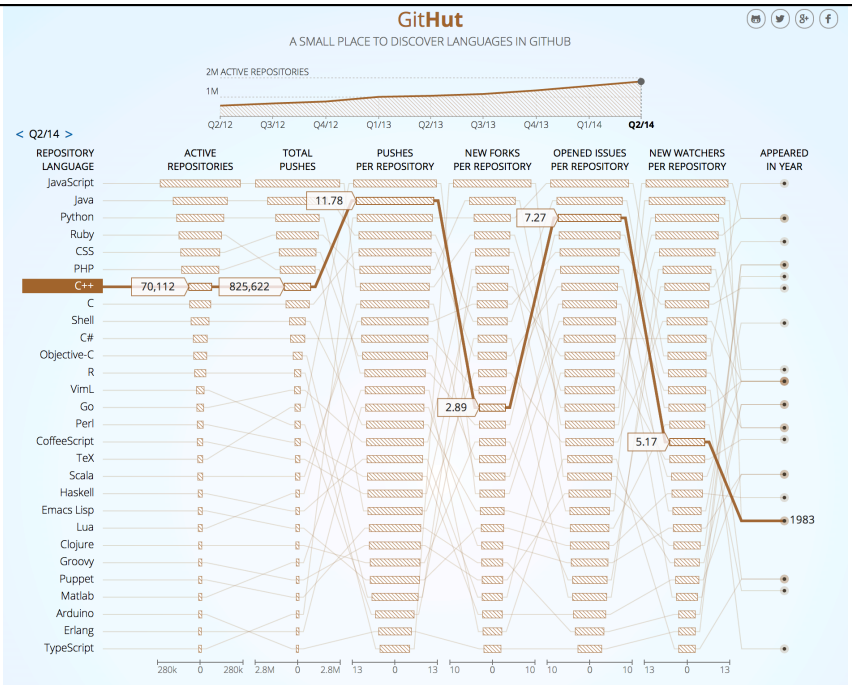
Parallel Coordinates



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

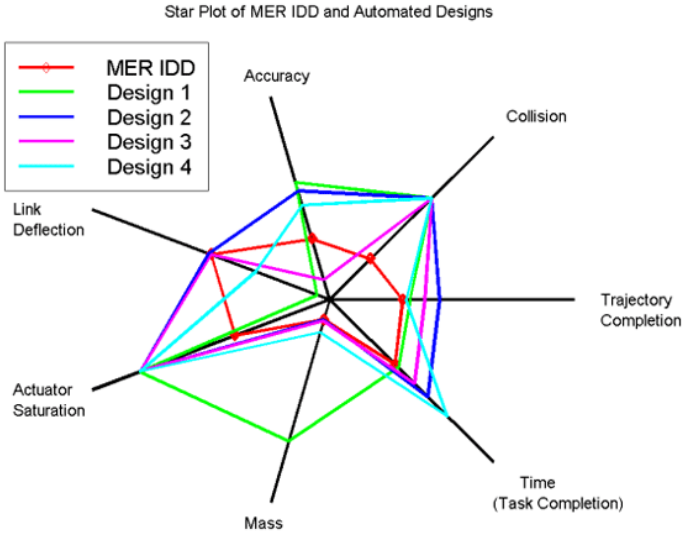
Designing Visualizations using Parallel Coordinates

- 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

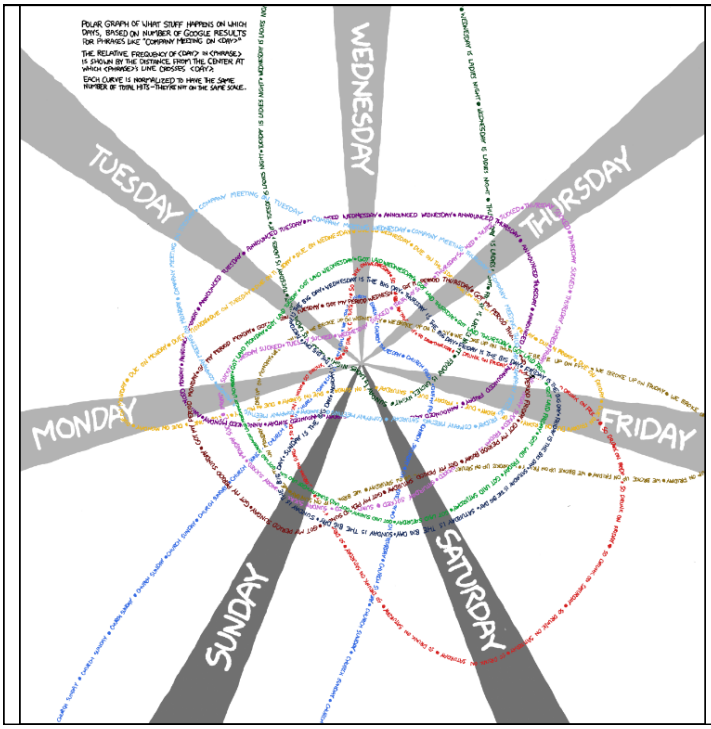


http://githut.info/

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://xkcd.com/930/

Today's Class

- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

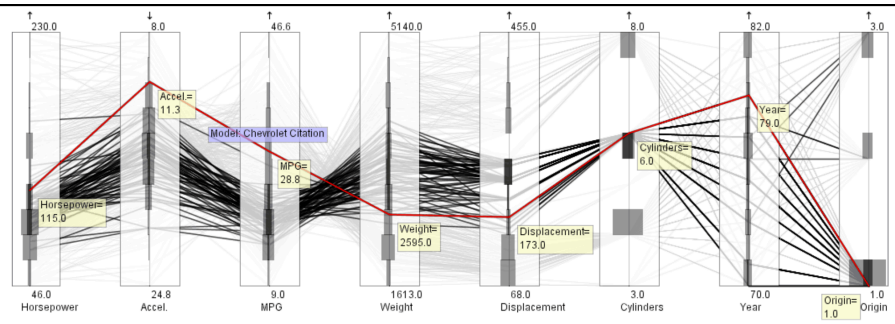


Figure 2. Extended parallel coordinates, a sample view of the *cars* data-set: cars with six cylinders were emphasized through brushing, histograms are laid over axes, and one data-point is shown with all details.

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

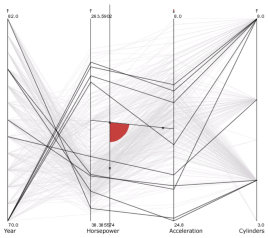


Figure 4. Reading between the lines: whereas most line-segments go up in-between the 2nd and the 3rd axis (visualizing a positive correlation of values there), just a few go down – those have been emphasized through angular brushing.

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

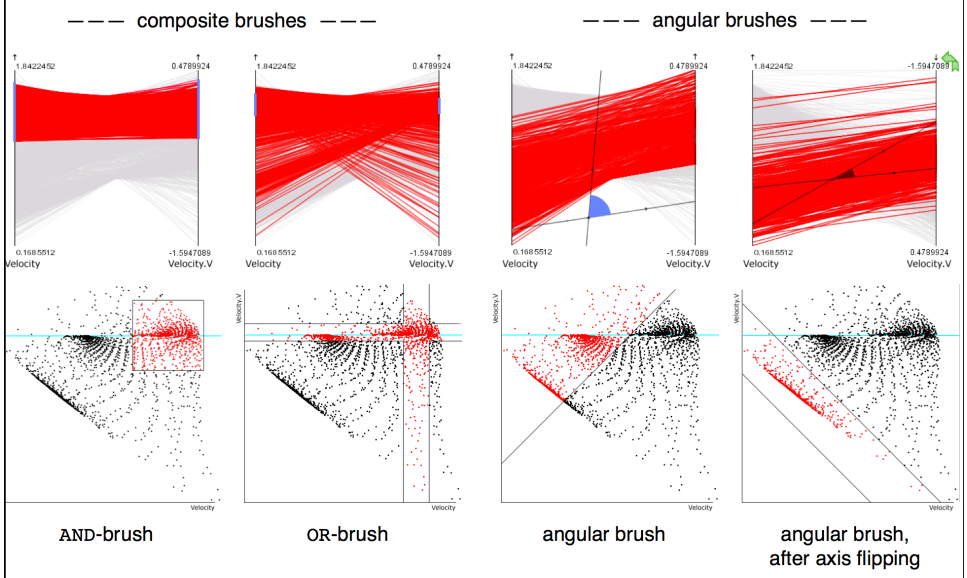
Contributions

- 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

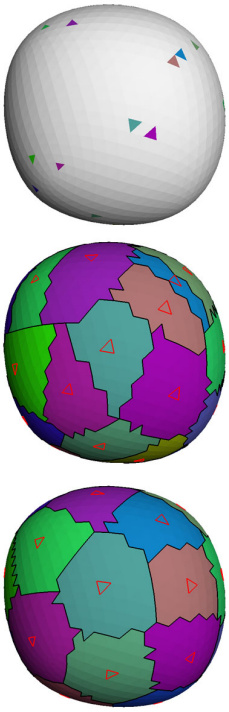


"Angular Brushing of Extended Parallel Coordinates",
Hauser, Ledermann, and Doleisch, InfoVis 2002

Today's Class

- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

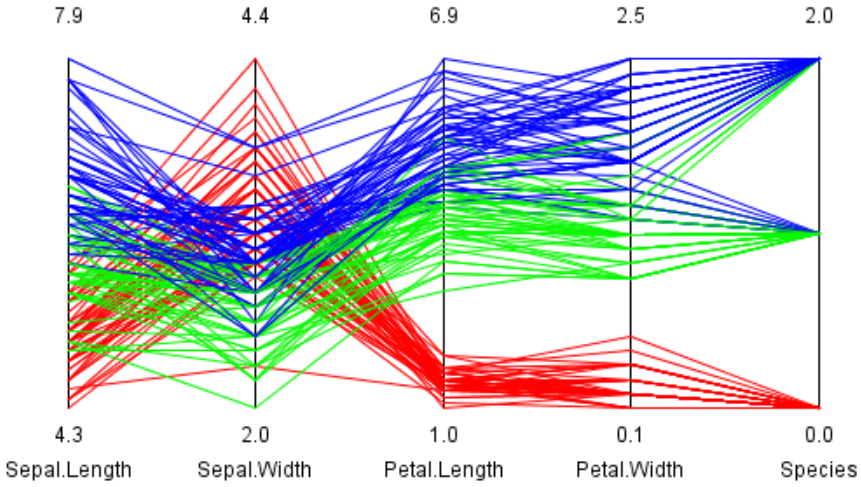
K-Means Clustering



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

1. Choose k, how many clusters you want ("oracle" tells us...)
2. Select k points from your data at random as initial team representatives
3. Every other point determines which team representative it is closest to and joins that team
4. The team averages the positions of all members, this is the team's new representative
5. Repeat 3-5 times until change < threshold

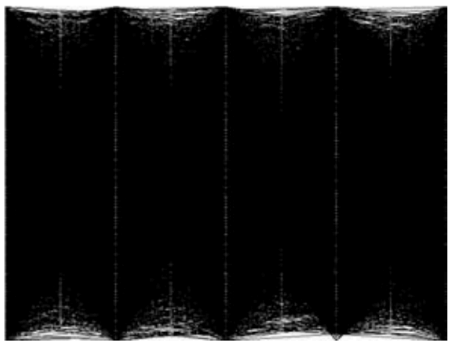
Clustering & Parallel Coordinates



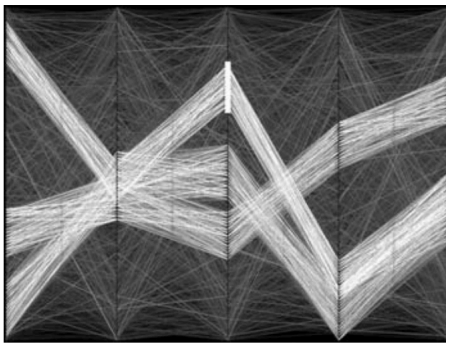
http://wanderinformatiker.at/unipages/ParCoord/clustering_en.html

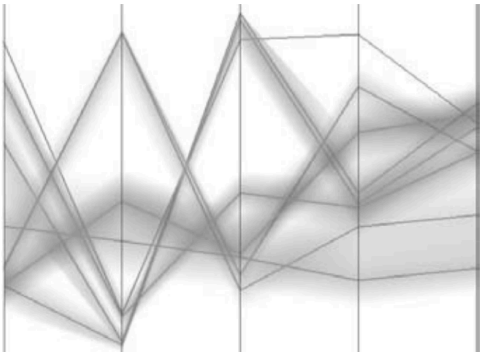
How to do (K-means) Clustering

- Determine your *distance function*
  - In spatial datasets, often just be Euclidean distance
    - Maybe also add in surface normal, etc.
  - Relative weighting of different dimensions
    - Especially tricky when units are unrelated convert to % of range
    - Also problematic when values are binary
- Finding nearest neighbors can be expensive
  - Use a spatial data structure



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





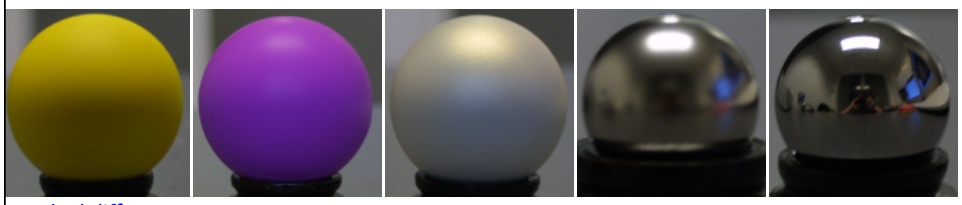
- Parallel coordinates becomes less effective when data is high dimensional and/or quantity of data points becomes huge
  - Visual crowding is not unique to parallel coordinates
- Synthetic test dataset with 5 clusters (varying sizes) embedded in random noise
  - Goal is to filter signal from this noise!
  - 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

Today's Class

- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

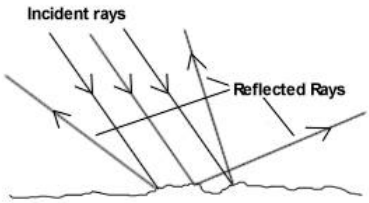
Real World Materials

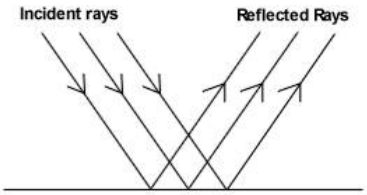


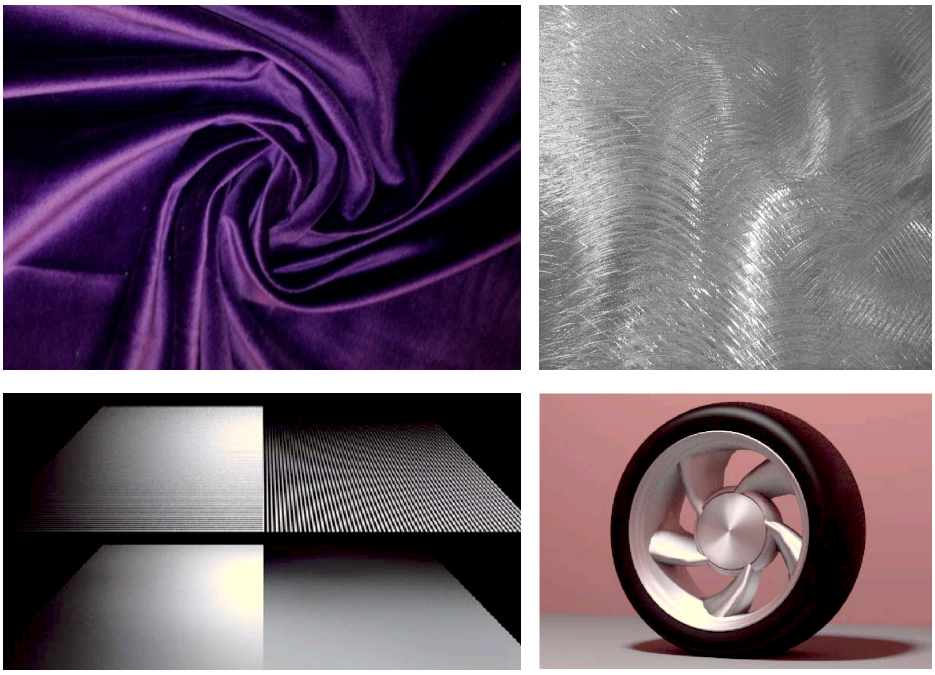
Ideal diffuse (Lambertian) e.g. chalk, matte paint

Non-ideal reflectors "glossy"

Ideal specular (mirror)



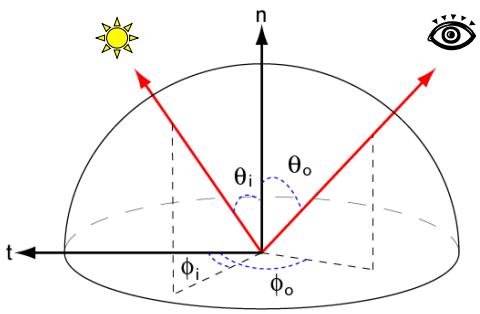




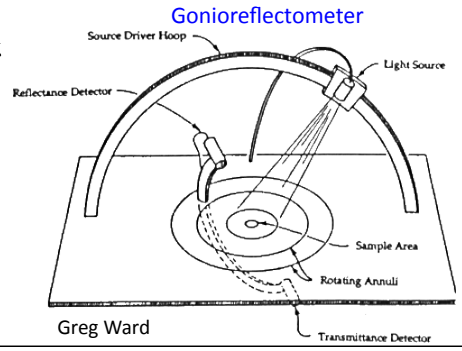
"Predicting reflectance functions from complex surfaces", Westin et al. SIGGRAPH 1992

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

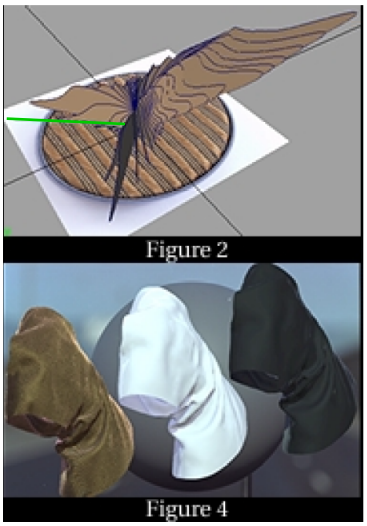


Complete material data capture: Gonioreflectometer



Greg Ward

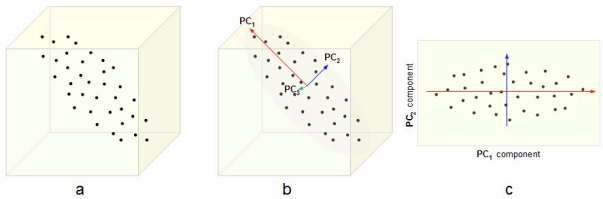
BRDFs in the Movie Industry



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

Principle Components Analysis (PCA)

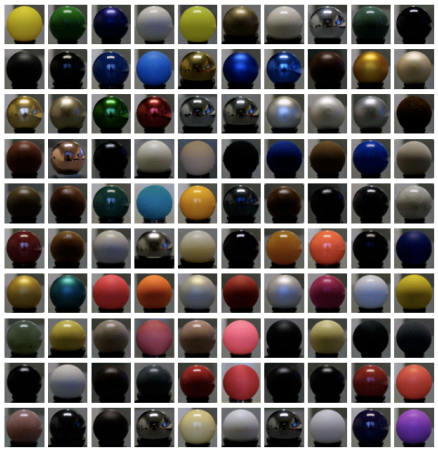
- Takes high dimensional data, where some/many axes are correlated



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

- Reduce to a smaller set of dimensions that are *not correlated*
- 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
- Essentially we want to find the internal structure that best explains the variance in the data

PCA Example: Material Reflectance Model



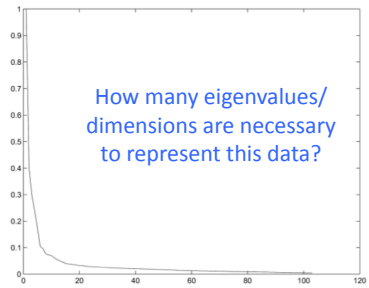
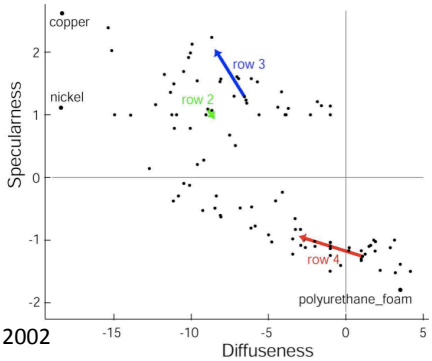
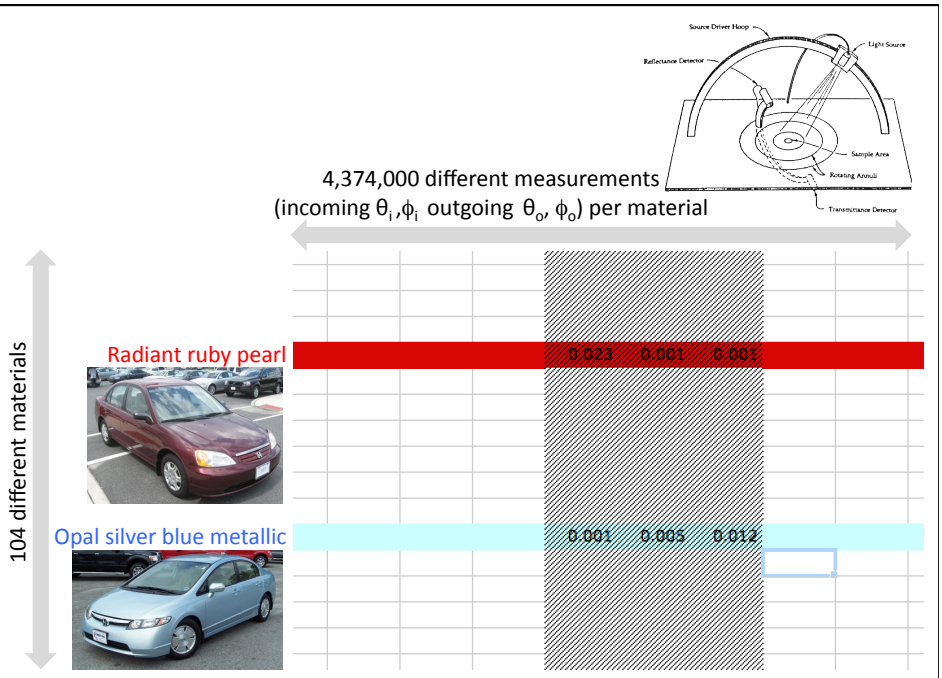


Figure 7: Plot of the eigenvalues resulting from PCA of the data set.



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



Today's Class

- Examples of High Dimensional Data
- Parallel Coordinates
  - Reading for Today: "Angular Brushing of Extended Parallel Coordinates"
- Data Clustering
  - "Uncovering Clusters in Crowded Parallel Coordinates Visualizations"
- Principle Components Analysis (PCA)
  - Example from Graphics: Materials! "A Data-Driven Reflectance Model"
- Summary: General Massive Data Visualization Tips
- Assignment 5 & Reading for Tuesday

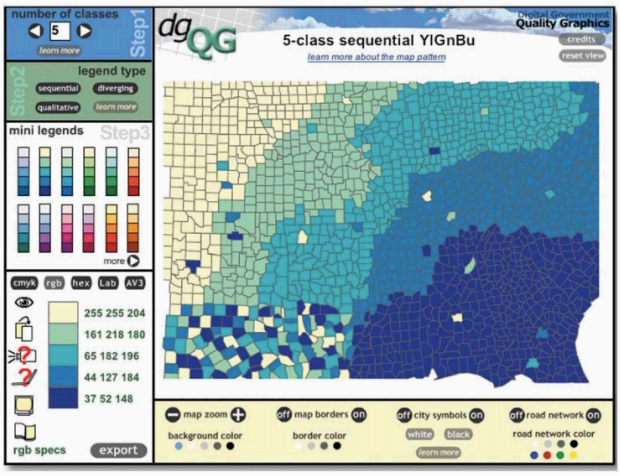
General Massive Data Visualization Tips

- Use your spatial real estate effectively
  - sort, organize
  - cluster, separate
  - layout, relative distances
- Color, Contrast, Intensity, Transparency
  - layering, overlapping

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

Reading for Tuesday

- "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps", Harrower & Brewer, The Cartographic Journal, 2003.



Homework Assignment 5: Experimenting with Color

- Revisit an earlier assignment/data/toolkit
  - Make a non-color-related improvement to this visualization
- Prepare many versions of the same visualization experimenting with different color palettes, e.g.:
  - Shades of grey
  - Black & white
  - Cool vs. warm tones
  - Bold/saturated vs. pastel
  - Colorblind aware
  - Light vs dark background and/or color negation
  - Etc.
- Analyze the effectiveness of the color scheme for each visualization.
  - How well does it convey the message? Or mislead the viewer?
- Compare the visualizations to each other.