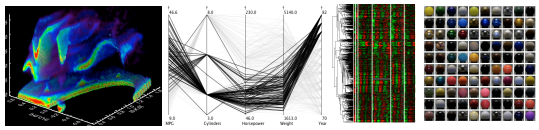
High Dimensional Data & Dimensionality Reduction



Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

Readings for This Week:

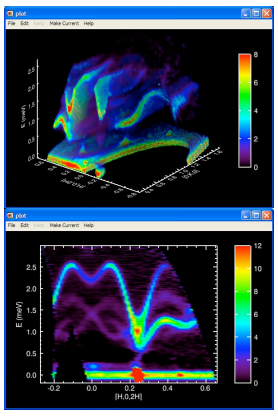
- "Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation", Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala, InfoVis 2008.
- "Software Design Patterns for Information Visualization", Jeffrey Heer, Maneesh Agrawala, TVCG 2006.

Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

Scientific Data

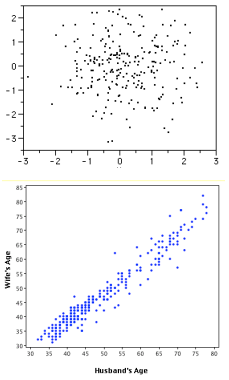
- For many 3D spatial locations during an experiment or simulation
- Time-varying temperature, velocity, pressure, humidity, etc.



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

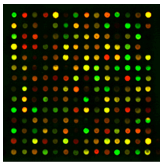
Misc. Personal Data

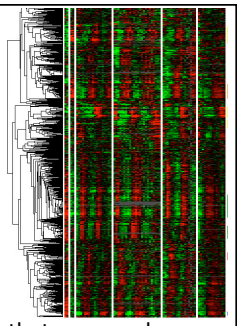
- Height, weight, eye color, phone number, address, IQ, age, cholesterol score, grade in Data Structures, etc.
- Example Hypotheses:
  - There is no correlation between height & phone number
  - There is a positive correlation between the ages of spouses
- Scatter plot: Look at 2 dimensions at a time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

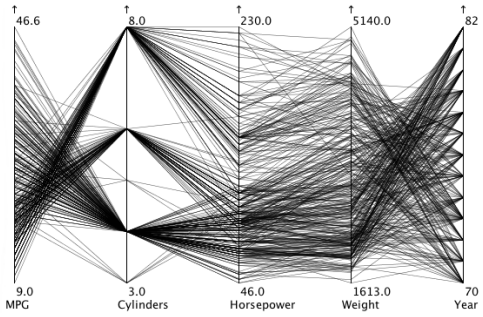
vs.

- 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

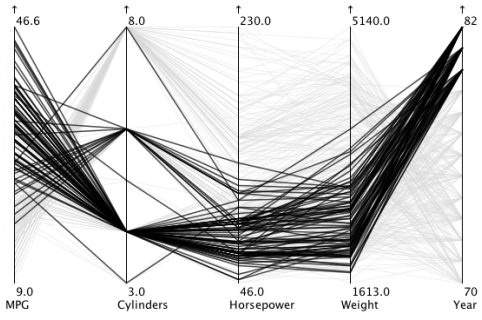
- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

Parallel Coordinates



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

Parallel Coordinates

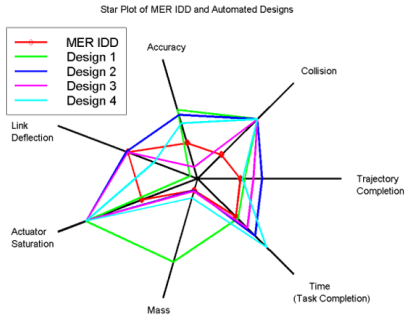


http://eagereyes.org/techniques/parallel-coordinates

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?)
- How many data points (lines)?
- How could color, line thickness, etc. be used to highlight patterns in the data?
- Data exploration or debugging tool? (Iterate)
- Or final visualization?
- Interaction: e.g., selection or filtering

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

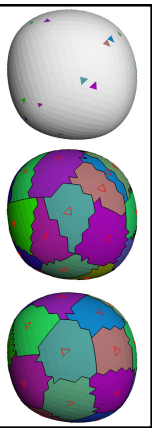
Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

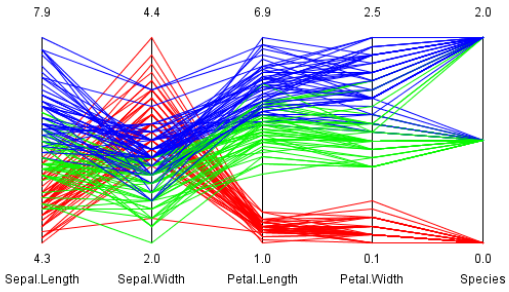
K-Means Clustering

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

1. Choose k, how many clusters you want (oracle)
2. Select k points from your data at random as initial team representative
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 until change < threshold



Clustering & Parallel Coordinates



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

How to do (K-means) Clustering

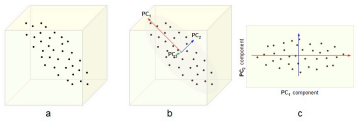
- Determine your *distance function*
  - In spatial datasets, often just be Euclidean distance
    - Maybe also add in surface normal, etc.
  - Relative weighting of different dimensions
    - Especially tricky when units are unrelated convert to % of range
    - Also problematic when values are binary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

Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

Principle Components Analysis (PCA)

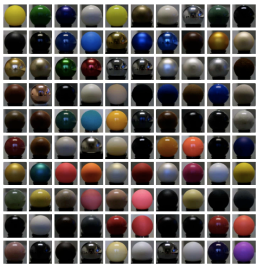
- Takes high dimensional data, where some/many axes are correlated



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



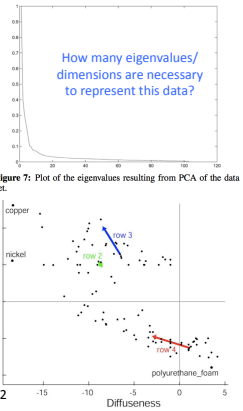
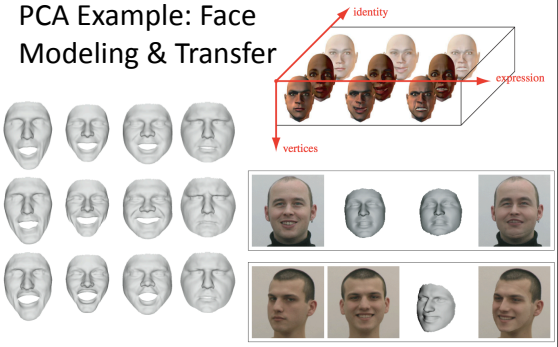


Figure 7: Plot of the eigenvalues resulting from PCA of the data set.

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

PCA Example: Face Modeling & Transfer



"Face Transfer with Multilinear Models", Vlasic, Brand, Pfister, & Popovic, SIGGRAPH 2005

Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

General Massive Data Visualization Tips

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

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

Today's Class

- Readings for this Week
- Examples of High Dimensional Data
- Parallel Coordinates
- Data Clustering
- Principle Components Analysis (PCA)
- General Massive Data Visualization Tips
- Next Week's Readings
- Assignment 5 & Mid-Term Presentation

Readings for Next Week: *(pick one)*

- "Tree visualization with Tree-maps: A 2-d space-filling approach", Ben Shneiderman, 1991

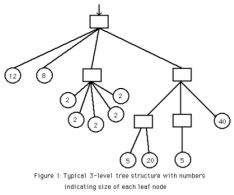


Figure 1: Typical 3-level tree structure with numbers indicating size of each leaf node

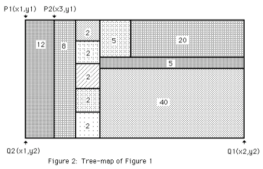
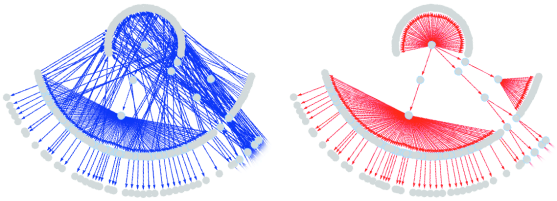


Figure 2: Tree-map of Figure 1

Readings for Next Week: *(pick one)*

- "Heapviz: Interactive Heap Visualization for Program Understanding and Debugging" Aftandilian, Kelley, Gramazio, Ricci, Su, & Guyer, 2010



Homework Assignment 5: due Tuesday @ 11:59pm
Wrangling High Dimensional Data

- Identify a high dimensional dataset
  - *may be the same data you or your teammate used for Assignment #4!*
- *Hypothesize* a pattern/correlation in the data
- Experiment with:
  - different styles of high-dimensional visualization *and/or*
  - different pre-processing (e.g., k-means, pca, etc.)
- *Analyze* the results

- Focus: Analysis & Validation & Visualization Execution

Mid-Term Presentations Wednesday March 7th

- ~ 5 minutes per person (10 mins per team of 2)
- Short & polished "Elevator Pitch" style presentation
  - Laptop projection (e.g., powerpoint or live demo), *or*
  - Poster printed
- Highly encouraged (not required) to revisit & extend a previous assignment
- Teams encouraged (not required)
- Formal peer feedback (incorporated into grade)