CSE 332
Introduction to Visualization
High-Dimensional Data

Klaus Mueller
Computer Science Department
Stony Brook University
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Basic tasks and data types</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data sources and preparation</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>4</td>
<td>Notion of similarity and distance</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Data and dimension reduction</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Visual bias</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Introduction to D3</td>
<td>Project 2 out</td>
</tr>
<tr>
<td>8</td>
<td>Visual perception and cognition</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Visual design and aesthetic</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>High-dimensional data – projective methods</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>High-dimensional data – optimizing methods</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Visualization of spatial data: volume visualization intro</td>
<td>Project 3 out</td>
</tr>
<tr>
<td>14</td>
<td>Introduction to GPU programming</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Visualization of spatial data: raycasting, transfer functions</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Illumination and isosurface rendering</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Scientific visualization</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Non-photorealistic and illustrative rendering</td>
<td>Project 4 out</td>
</tr>
<tr>
<td>20</td>
<td>Midterm discussion</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Principles of interaction</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Visual analytics and the visual sense making process</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Visualization of graphs and hierarchies</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Visualization of time-varying and streaming data</td>
<td>Project 5 out</td>
</tr>
<tr>
<td>25</td>
<td>Maps</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Memorable visualizations, visual embellishments</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Evaluation and user studies</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Narrative visualization, storytelling, data journalism, XAI</td>
<td></td>
</tr>
</tbody>
</table>
Feature vectors are typically high dimensional
- this means, they have many elements
- high dimensional space is tricky
- most people do not understand it
- why is that?
- well, because you don’t learn to see high-D when your vision system develops

Object permanence (Jean Piaget)
- the ability to create mental pictures or remember objects and people you have previously seen
- thought to be a vital precursor to creativity and abstract thinking
The curse of dimensionality

As $n \to \infty$
- Cube: side length $l$, diagonal $d$, volume $V$
- $V \to \infty$ for $l > 1$
- $V \to 0$ for $l < 1$
- $V = 1$ for $l = 1$
- $d \to \infty$

and very sparse

most points are here

and not here
High-D Space is Tricky

Essentially hypercube is like a “hedgehog”
Points are all at about the same distance from one another

- concentration of distances
- fundamental equation (Bellman, ’61)

\[
\lim_{n \to \infty} \frac{\text{Dist}_{\text{max}} - \text{Dist}_{\text{min}}}{\text{Dist}_{\text{min}}} \to 0
\]

- so as \( n \) increases, it is impossible to distinguish two points by (Euclidian) distance
  - unless these points are in the same cluster of points
Sparseness Demonstration

Space gets extremely sparse
- with every extra dimension points get pulled apart further
- distances become meaningless
Space gets extremely sparse

- with every extra dimension points get pulled apart further
- distances become meaningless

1D – points are very close

2D – points spread apart

3D – getting even sparser

4D, 5D, ... – sparseness grows further
Indexing (and storage) also gets very expensive

- exponential growth in the number of dimensions

- 4D: 65k cells  
- 5D: 1M cells  
- 6D: 16M cells  
- 7D: 268M cells  

- keep a keen eye on storage complexity
Invented by Al Inselberg in the early 1990s

Good way to see raw high-dimensional data

- but there are shortcomings
- we will see
The N=7 data axes are arranged side by side
- in parallel
Hard to see the individual cars?

- what can we do?
Grouping the cars into sub-populations

- we perform clustering
- can be automated or interactive (put the user in charge)
Computes the mean and superimposes it onto the lines

- allows one to see trends
PC With Illustrative Abstraction

individual polylines
PC With Illustrative Abstraction

completely abstracted away
blended partially
PC With Illustrative Abstraction

all put together – three clusters

[McDonnell and Mueller, 2008]
Interaction is Key

Interaction in Parallel Coordinate
Patterns in Parallel Coordinates

correlation: $r=-1.0$  $r=0$  $r=1.0$
Patterns in Parallel Coordinates

# points

Fisher-z (corresponding to $\rho = 0, \pm 0.462, \pm 0.762, \pm 0.905$)
Li et al. found that twice as many correlation levels can be distinguished with scatterplots.

Information Visualization Vol. 9, 1, 13 – 30
There are $n!$ ways to order the $n$ dimensions

- how many orderings for 7 dimensions?
- 5,040
- but since can see relationships across 3 axes a better estimate is $n!/((n-3)! 3!)$ = 35
- still a lot of axes orderings to try out $\rightarrow$ we need help
Correlation

- a statistical measure that indicates the extent to which two or more variables fluctuate together

\[
r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]
Create a correlation matrix
Run a mass-spring model
Run Traveling Salesman on the correlation nodes
Use it to order your parallel coordinate axes via TSP

[Zhang and Mueller, 2012]
**Interaction with the Correlation Network**

- Vertices are attributes, edges are correlations:
  - vertex: size determined by $\sum_{j=0}^{D} \frac{|\text{correlation}(i,j)|}{D-1}$ for $j \neq i$
  - edge: color/intensity $\rightarrow$ sign/strength of correlation

![Graphs showing all edges and filtered by strength](attachment:graphs.png)

- Attribute centric subset of attributes

[Zhang and Mueller, 2014]
3 subspaces are well separated.
Correlation strength can often be improved by constraining a variable’s value range

- this limits the derived relationships to this value range
- such limits are commonplace in targeted marketing, etc.

[Zhang and Mueller, 2014]
Developed by [Kosara et al. TVCG, 2006]

Parallel coordinates for categorical data
- for example, census and survey data, inventory, etc.
- data that can be summed up in a cross-tabulation

Example
- Titanic dataset
- what can we see here?
Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice.
Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice. Figure 8, for example, at first sight appeared to be showing a change over time, but in fact further inspection shows that the different x-coordinates are almost entirely unrelated to one another and in no particular order.
Multiple visualizations appear to present categorical data as line graphs, which seems a strange choice. Figure 8, for example, at first sight appeared to be showing a change over time, but in fact further inspection shows that the different x-coordinates are almost entirely unrelated to one another and in no particular order. This is such an unusual choice that I'm not sure that I am understanding the role of the graphs correctly.
How to Teach Mainstream Users
Encode user responses based on task complexities

- none (0): cannot report any findings
- low (1): understand representation visual encoding
- medium (2): identify groups and outliers
- high (3): recognize correlations and trends
Visual understanding:
(1) The MPG of the orange-highlighted car is ~40% of its range
(2) There is just one line at the top of the acceleration scale
(3) Heavier cars are faster

Data Understanding:
(1) The number of cylinders of the orange-highlighted car is 4, one fifth between 3 and 8.
(2) Many cars have the same numbers of cylinders, mostly even numbers particularly 4 and 8.
(3) Heavier cars have more cylinders and hence more horsepower and speed.
## Results

<table>
<thead>
<tr>
<th>Participants</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
<th>V10</th>
<th>V11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>After</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Diff.</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
<th>D8</th>
<th>D9</th>
<th>D10</th>
<th>D11</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Fused dataset of 50 US colleges
US News: academic rankings
College Prowler: survey on campus life attributes
ANATOMY OF A SALES PIPELINE

lead generator

lead pool

lead

qualified lead

opportunity

opportunity+

opportunity++

opportunity+++ 
customer

responds

responds++

requests info (RFI)

requests pricing (RFP)

shapes deal

signs/buys

account manager

potential

action

prospect
Scene:
- a meeting of sales executives of a large corporation, Vandelay Industries

Mission:
- review the strategies of their various sales teams

Evidence:
- data of three sales teams with a couple of hundred sales people in each team
Meet Kate, a sales analyst in the meeting room:

“OK...let’s see, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads”
“Let’s go and make a revealing route!”

- she uses the mouse and designs the route shown
- she starts explaining the data like a story ...
“Let’s go and make a revealing route!”

- she uses the mouse and designs the route shown
- she starts explaining the data like a story ...
Kate notices something else:

- now looking at the red team
- there seems to be a spread in effectiveness among the team
- the team splits into three distinct groups

She recommends: “Maybe fire the least effective group or at least retrain them”
Projection of the data items into a bivariate basis of axes
How does 2D projection work in practice?

- N-dimensional point \( x = \{x_1, x_2, x_3, \ldots, x_N\} \)
- a basis of two orthogonal axis vectors defined in N-D space
  \[
  a = \{a_1, a_2, a_3, \ldots, a_N\}
  \]
  \[
  b = \{b_1, b_2, b_3, \ldots, b_N\}
  \]
- a projection \( \{x_a, x_b\} \) of \( x \) into the 2D basis spanned by \( \{a, b\} \) is:
  \[
  x_a = a \cdot x^T
  \]
  \[
  x_b = b \cdot x^T
  \]
where \( \cdot \) is the dot product, \( T \) is the transpose.
Projection causes inaccuracies

- close neighbors in the projections may not be close neighbors in the original higher-dimensional space
- this is called projection ambiguity
SCATTERPLOT FOR TWO ATTRIBUTES

Appropriate for the display of bivariate relationships
What to do when there are more than two variables?

- arrange multivariate relationships into scatterplot matrices
- not overly intuitive to perceive multivariate relationships
### Scatterplot Matrix (SPLOM)

Climatic predictors:

<table>
<thead>
<tr>
<th></th>
<th>WetDays</th>
<th>TempJuly</th>
<th>TempJan</th>
<th>TempAnn</th>
<th>RHJuly</th>
</tr>
</thead>
<tbody>
<tr>
<td>WetDays</td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td>TempJuly</td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td>TempJan</td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td>TempAnn</td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
</tr>
<tr>
<td>RHJuly</td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
<td><img src="image" alt="Plot" /></td>
</tr>
</tbody>
</table>
Scatterplot version of parallel coordinates

- distributes $n(n-1)$ bivariate relationships over a set of tiles
- for $n=4$ get 16 tiles
- can use $n(n-1)/2$ tiles

For even moderately large $n$:
- there will be too many tiles

Which plots to select?
- plots that show correlations well
- plots that separate clusters well
Several metrics, a good one is Distance Consistency (DSC)

\[ \text{DSC} = \frac{\left| x' \in v(X) : \text{CD}(x', \text{centr'}(c_{\text{clabel}(x)}) = \text{true} \right|}{k} \]

- measures how “pure” a cluster is
- pick the views with highest normalized DSC

Dunn Index

Favors clusters that are compact and are well isolated

\[
DI_m = \frac{\min_{1 \leq i < j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k}
\]

\[
\Delta_i = \frac{\sum_{x \in C_i} d(x, \mu)}{|C_i|}, \quad \mu = \frac{\sum_{x \in C_i} x}{|C_i|},
\]

calculates distance of all the points from the mean.

\[
\delta(C_i, C_j)
\]

be this intercluster distance metric, between clusters \(C_i\) and \(C_j\).
Biplots

Plots data points and dimension axes into a single visualization

- uses first two PCA vectors as the basis to project into
- find plot coordinates \([x] [y]\)
  for data points: \([\text{PCA}_1 \cdot \text{data vector}] [\text{PCA}_2 \cdot \text{data vector}]\)
  for dimension axes: \([\text{PCA}_1[\text{dimension}]] [\text{PCA}_2[\text{dimension}]]\)
See data distributions into the context of their attributes

Cape Bounty and Sanagak Lake
Correspondence Analysis with Ecological Classes
Biplots in Practice

See data points into the context of their attributes
Biplots – A Word of Caution

Do be aware that the projections may not be fully accurate

- you are projecting N-D into 2D by a linear transformation
- if there are more than 2 significant PCA vectors then some variability will be lost and won’t be visualized
- remote data points might project into nearby plot locations suggesting false relationships → projection ambiguity
- always check out the PCA scree plot to gauge accuracy
Also called multivariate scatterplot

- biplot-axes length vis replaced by graphical design
- less cluttered view
- but there’s more to this .....
Meet the **Subspace Voyager**

Decomposes high-D data spaces into lower-D subspaces by
- clustering
- classification
- reducing clusters to intrinsic dimensionality via local PCA

Allows users to interactively explore these lower-D subspaces
- explore them as a chain of 3D subspaces
- transition seamlessly to adjacent 3D subspaces on demand
- save observations as you go (and return to them just as well)
VISUALIZE RAW DATA w/ THE SUBSPACE VOYAGER

Interactive Scatterplot

Subspace Trail Map
Trackball-Based Cluster Exploration
Uses genetic-algorithm driven projection pursuit
Several view quality metrics are available
Generate many views and score them (one per ant)

- poor scoring ants die and well-scoring ants survive
- sub paths of high scoring receive pheromone
- pheromone entices ants to take this path again
- each path variation is a parameter choice
- best view corresponds to the path that is converged on
CHASE INTERESTING CLUSTERS – TRANSITION TO ADJACENT 3D SUBSPACES
EDIT AND ANNOTATE CLUSTERS
The Subspace Trail Map
Walk The Subspace Trail Map
Clarify Spatial Relationships
Clarify Spatial Relationships
STAR COORDINATES

Coordinate system based on axes positioned in a star

- a point $P$ is vector sum of all axis coordinates

Interactions

- axis rescaling, rotation
- reveal correlations
- resolve plotting ambiguities

\[ P = O + \sum_{i=1}^{m} d_i \vec{c}_i \]

[E. Kandogan SIGKDD 2001]
Star Coordinates

Operations defined on Star Coords

- scaling changes contribution to resulting visualization
- axis rotation can visualize correlations
- also used to reduce projection ambiguities
Similar to Star Coordinates

- uses a spring model
- difference is normalization by sum of values

\[ P = \frac{\sum_{i=1}^{m} d_i \tilde{c}_i}{\sum_{i=1}^{m} d_i} \]

Figure by: Rubio-Sanchez et al. TVCG 2015
Optimizing the RadViz Layout

Optimize

- correlation-based attribute placement on circle using TSP
- samples placed iteratively into circle using similarity constraints

[Cheng and Mueller, Pacific Vis 2015]
Radar Chart

Equivalent to a parallel coordinates plot, with the axes arranged radially
- each star represents a single observation
- can show outliers and commonalities nicely

Disadvantages
- hard to make trade-off decisions
- distorts data to some extents when lines are filled in
All of these scatterplot displays share the following characteristics

- allow users to see the data points in the context of the variables
- but can suffer from projection ambiguity
- some offer interaction to resolve some of these shortcomings
- but interaction can be tedious

Are there visualization paradigms that can overcome these problems?

- yes, algorithms that optimize the layout to preserve distances or similarities in high-dimensional space
- what is this algorithm?
- yes, MDS (Multi-Dimensional Scaling)
- we have discussed MDS before (so we will skip further discussion)
Describe scatterplot features by graph theoretic measures

- mostly built on minimum spanning tree
- can be used to summarize large sets of scatterplots
Use scagnostics to quickly survey 1,000s of scatterplots

- compute scagnostics measures
- create scatterplot matrix of these measures
- each scatterplot is a point