CSE 512 - Data Visualization

Multidimensional Vis

Jeffrey Heer  University of Washington
Last Time:
Exploratory Data Analysis
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.
# Antibiotic Effectiveness

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Streptomycin</th>
<th>Neomycin</th>
<th>Gram Staining</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aerobacter aerogenes</em></td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>negative</td>
</tr>
<tr>
<td><em>Brucella abortus</em></td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>negative</td>
</tr>
<tr>
<td><em>Brucella anthracis</em></td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>positive</td>
</tr>
<tr>
<td><em>Diplococcus pneumoniae</em></td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em></td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>negative</td>
</tr>
<tr>
<td><em>Proteus vulgaris</em></td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>negative</td>
</tr>
<tr>
<td><em>Salmonella (Eberthella) typhosa</em></td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>negative</td>
</tr>
<tr>
<td><em>Salmonella schottmuelleri</em></td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>negative</td>
</tr>
<tr>
<td><em>Staphylococcus albus</em></td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus fecalis</em></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus hemolyticus</em></td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus viridans</em></td>
<td>0.005</td>
<td>10</td>
<td>40</td>
<td>positive</td>
</tr>
</tbody>
</table>
How do the drugs compare?

Original graphic by Will Burtin, 1951
How do the drugs compare?

- **Penicillin**
  - *Brucella anthracis* + 0.001
  - *Streptococcus hemolyticus* + 0.001
  - *Diplococcus pneumoniae* + 0.005
  - *Streptococcus viridans* + 0.005
  - *Staphylococcus albus* + 0.007
  - *Staphylococcus aureus* + 0.03
  - *Streptococcus faecalis* + 1

- **Streptomycin**
  - *Brucella abortus* 1
  - *Salmonella (Eberthella) typhosa* 1
  - *Proteus vulgaris* 3
  - *Salmonella schottmuelleri* 10
  - *Escherichia coli* 100
  - *Mycobacterium tuberculosis* 800
  - *Klebsiella pneumoniae* 350
  - *Pseudomonas aeruginosa* 350
  - *Aerobacter aerogenes* 370

- **Neomycin**
  - *Brucella abortus* 0.007
  - *Salmonella (Eberthella) typhosa* 0.008
  - *Proteus vulgaris* 0.1
  - *Salmonella schottmuelleri* 0.09
  - *Escherichia coli* 0.1
  - *Mycobacterium tuberculosis* 2
  - *Klebsiella pneumoniae* 2
  - *Pseudomonas aeruginosa* 1.6
  - *Aerobacter aerogenes* 1
Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009
[The Elements of Graphing Data. Cleveland 94]
Linear regression ...

[The Elements of Graphing Data. Cleveland 94]
Linear regression w/out outlier ...

[The Elements of Graphing Data, Cleveland 94]
Transforming Data

How well does the curve fit the data?

[Cleveland 85]
Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

[Cleveland 85]
Multiple Plotting Options

Plot model in data space

Plot data in model space

[Cleveland 85]
A2: Exploratory Data Analysis

Use visualization software to form & answer questions

First steps:
Step 1: Pick domain & data
Step 2: Pose questions
Step 3: Profile the data
Iterate as needed

Create visualizations
Interact with data
Refine your questions

Make a notebook
Keep record of your analysis
Prepare a final graphic and caption

Due by 5:00pm Friday, April 15
Tutorials!

Visualization Tools
Tue 4/12, 3:00-4:20pm PAA 114A
Introduction to Tableau, plus a few others.

d3.js: Data-Driven Documents
Tue 4/19, 3:00-4:20pm PAA 114A
Focus on D3, touches on HTML/CSS/JS
The Design Space of Visual Encodings
Univariate Data

Factors

Variable

A
B
C

1

A
B
C
D
Univariate Data

Tukey box plot

Factors

A  B  C  D  E

Middle 50%  Median  low  high

1  20

A  B  C

Variables

A  B  C  D  E
Bivariate Data

Scatter plot is common
Trivariate Data

3D scatter plot is possible
Three Variables

Two variables \([x, y]\) can map to points
Scatterplots, maps, ...

Third variable \([z]\) must use
Color, size, shape, ...

\[
\begin{align*}
1 & \\
2 & \\
\end{align*}
\]
Large Design Space

Multidimensional Data
Visual Encoding Variables

- Position (X)
- Position (Y)
- Size
- Value
- Texture
- Color
- Orientation
- Shape

~8 dimensions?
Example: Coffee Sales

Sales figures for a fictional coffee chain

Sales | Q-Ratio
---|---
Profit | Q-Ratio
Marketing | Q-Ratio
Product Type | N {Coffee, Espresso, Herbal Tea, Tea}
Market | N {Central, East, South, West}
Encode “Sales” (Q) and “Profit” (Q) using Position
Encode “Product Type” (N) using Hue
Encode “Market” (N) using Shape.
Encode “Marketing” (Q) using Size
A *trellis plot* subdivides space to enable comparison across multiple plots. Typically nominal or ordinal variables are used as dimensions for subdivision.
Small Multiples

[MacEachren 95, Figure 2.11, p. 38]
Small Multiples

[MacEachren 95, Figure 2.11, p. 38]
Scatterplot Matrix (SPLOM)

Scatter plots for pairwise comparison of each data dimension.
Multiple Coordinated Views

- How long in majors
- Select high salaries
- Average assists vs average putouts (fielding ability)
- Distribution of positions played
- Average career HRs vs average career hits (batting ability)
Linking Assists to Position
Observation: We have evolved a sophisticated ability to interpret faces.

Idea: Map data variables to facial features.

Question: Do we process facial features in an uncorrelated way? (i.e., are they separable?)

This is just one example of nD “glyphs”
Visualizing Multiple Dimensions

**Strategies:**
Avoid “over-encoding”
Use space and small multiples intelligently
Reduce the problem space
Use interaction to generate *relevant* views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is key.
Parallel Coordinates
Parallel Coordinates [Inselberg]
Parallel Coordinates [Inselberg]

Figure 1: The full dataset consisting of 473 batches
The Multidimensional Detective

Production data for 473 batches of a VLSI chip
16 process parameters
X1: The yield: % of produced chips that are useful
X2: The quality of the produced chips (speed)
X3-12: 10 types of defects (0 defects shown at top)
X13-16: 4 physical parameters

Objective:
Raise the yield (X1) and maintain high quality (X2)

Parallel Coordinates [Inselberg]

Figure 1: The full dataset consisting of 473 batches
Inselberg’s Principles

1. Do not let the picture scare you.
2. Understand your objectives. Use them to obtain visual cues.
3. Carefully scrutinize the picture.
4. Test your assumptions, especially the “I am really sure of’s”.
5. You can’t be unlucky all the time!
Each line represents a tuple (e.g., VLSI batch) Filtered below for high values of $X_1$ and $X_2$

Figure 2: The batches high in Yield, $X_1$, and Quality, $X_2$. 
Look for batches with *nearly* zero defects (9/10). Most of these have low yields -> defects OK.
Figure 5: The best batch. Highest in Yield, $X_1$, and very high in Quality, $X_2$.

Figure 7: Upper range of split in $X_{15}$.
Notice that X6 behaves differently.
Allow 2 defects, including X6 -> best batches
Radar Plot / Star Graph

"Parallel" dimensions in polar coordinate space
Best if same units apply to each axis
Dimensionality Reduction
Dimensionality Reduction

http://www.ggobi.org/
Principal Components Analysis

1. Mean-center the data.
2. Find \( \perp \) basis vectors that maximize the data variance.
3. Plot the data using the top vectors.
PCA of Genomes [Demiralp et al. '13]
Time Curves [Bach et al. ’16]

Timeline:
1 2 3 4 5 6 7
Circles are data cases with a time stamp. Similar colors indicate similar data cases.

Folding:
1 2 3 4 5 6 7

Time curve:
1 2 3 4 5 6 7
The temporal ordering of data cases is preserved. Spatial proximity now indicates similarity.

(a) Folding time

Wikipedia “Chocolate” Article

U.S. Precipitation over 1 Year
Many Reduction Techniques!

- Principal Components Analysis (PCA)
- Multidimensional Scaling (MDS)
- Locally Linear Embedding (LLE)
- t-Dist. Stochastic Neighbor Embedding (t-SNE)
- Isomap
- Auto-Encoder Neural Networks
- Topological methods
...
Tableau / Polaris
Polaris [Stolte et al.]
Tableau

Data Display

Data Model

Encodings
Tableau Demo

The dataset:
Federal Elections Commission Receipts
Every Congressional Candidate from 1996 to 2002
4 Election Cycles
9216 Candidacies
Dataset Schema

Year (Qi)
Candidate Code (N)
Candidate Name (N)
Incumbent / Challenger / Open-Seat (N)
Party Code (N) [1=Dem, 2=Rep, 3=Other]
Party Name (N)
Total Receipts (Qr)
State (N)
District (N)

This is a subset of the larger data set available from the FEC.
Hypotheses?

What might we learn from this data?
Hypotheses?

What might we learn from this data?
Correlation between receipts and winners?
Do receipts increase over time?
Which states spend the most?
Which party spends the most?
Margin of victory vs. amount spent?
Amount spent between competitors?
Tableau Demo
Tableau/Polaris Approach

Insight: can simultaneously specify both database queries and visualization
Choose data, then visualization, not vice versa
Use smart defaults for visual encodings
More recently: automate visualization design
Specifying Table Configurations

Operands are the database fields
Each operand interpreted as a set {...}
Quantitative and Ordinal fields treated differently

Three operators:
concatenation (+)
cross product (x)
nest (/)
<table>
<thead>
<tr>
<th>Region</th>
<th>Segment</th>
<th>Technology</th>
<th>Office Supplies</th>
<th>Furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>Consumer</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>Central</td>
<td>Corporate</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>Central</td>
<td>Home Office</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>East</td>
<td>Consumer</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>East</td>
<td>Corporate</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>East</td>
<td>Home Office</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>South</td>
<td>Consumer</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>South</td>
<td>Corporate</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>South</td>
<td>Home Office</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>West</td>
<td>Consumer</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>West</td>
<td>Corporate</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>West</td>
<td>Home Office</td>
<td>Blue</td>
<td>Orange</td>
<td>Green</td>
</tr>
</tbody>
</table>

SUM(Sales) = $2,297,201
GROUP BY Category, Region, Segment
Table Algebra: Operands

**Ordinal fields**: interpret domain as a set that partitions table into rows and columns.
Quarter = \{\{(Qtr1),(Qtr2),(Qtr3),(Qtr4)\}\} ->

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>95892</td>
<td>101760</td>
<td>105282</td>
<td>98225</td>
</tr>
</tbody>
</table>

**Quantitative fields**: treat domain as single element set and encode spatially as axes.
Profit = \{\{(Profit[-410,650])\}\} ->

![Profit Scale]
Concatenation (+) Operator

Ordered union of set interpretations

Quarter + Product Type
= \{\text{(Qtr1),(Qtr2),(Qtr3),(Qtr4)}\} + \{\text{(Coffee), (Espresso)}\}
= \{\text{(Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso)}\}

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
<th>Coffee</th>
<th>Espresso</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>59</td>
<td>57</td>
<td>53</td>
<td>151</td>
<td>21</td>
</tr>
</tbody>
</table>

Profit + Sales = \{\text{(Profit[-310,620]),(Sales[0,1000])}\}
Cross (x) Operator

Cross-product of set interpretations

Quarter x Product Type =
{(Qtr1,Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4, Tea)}

Product Type x Profit =
Nest (/) Operator

Cross-product filtered by existing records

Quarter x Month ->
creates twelve entries for each quarter. i.e.,
(Qtr1, December)

Quarter / Month ->
creates three entries per quarter based on
tuples in database (not semantics)
Table Algebra

The operators (+, x, /) and operands (O, Q) provide an algebra for tabular visualization.

Algebraic statements are then mapped to:
- **Visualizations** - trellis plot partitions, visual encodings
- **Queries** - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop. Note that this specifies operands *NOT* operators! Operators are inferred by data type (O, Q).
<table>
<thead>
<tr>
<th>State</th>
<th>Coffee</th>
<th>Espresso</th>
<th>Herbal Tea</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Connecticut</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Florida</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Illinois</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Iowa</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Louisiana</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Missouri</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Nevada</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>New Mexico</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>New York</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Ohio</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Oregon</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Texas</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Utah</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Washington</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
</tbody>
</table>
Quantitative-Quantitative
Ordinal-Quantitative
Querying the Database

1. Select records from the database, filtering by user-defined criteria.

2. Partition the records into layers and panes. The same record may appear in multiple partitions.

3. Group, sort, and aggregate the relations within each pane.

4. Render and compose layers.
Visualizing Multiple Dimensions

Strategies:
Avoid “over-encoding”
Use space and small multiples intelligently
Reduce the problem space
Use interaction to generate relevant views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is key.