Multi-Dimensional Vis
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.
Graph Viewer

Roll-up by:
- All

Visualization:
- Matrix

Sort by:
- Linkage

Edge centrality filters:
# Antibiotic Effectiveness

## Table 1: Burtin’s data.

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

Will Burtin, 1951

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Antibiotic Streptomycin</th>
<th>Neomycin</th>
<th>Gram stain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>-</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>+</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>+</td>
</tr>
<tr>
<td>Klebsiella pneumoniae</td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>-</td>
</tr>
<tr>
<td>Salmonella (Eberthella) typhosa</td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>-</td>
</tr>
<tr>
<td>Salmonella schottmuelleri</td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>-</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus fecalis</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus hemolyticus</td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus viridans</td>
<td>0.005</td>
<td>10</td>
<td>40</td>
<td>+</td>
</tr>
</tbody>
</table>
How do the bacteria group w.r.t. resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009
Transforming data

How well does the curve fit 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]
Choropleth maps of cancer deaths in Texas.

One plot shows a real data set. The others are simulated under the null hypothesis of spatial independence.

Can you spot the real data? If so, you have some evidence of spatial dependence in the data.
Multidimensional Visualization
Visual Encoding Variables

- Position
- Length
- Area
- Volume
- Value
- Texture
- Color
- Orientation
- Shape

~8 dimensions?
## Example: Coffee Sales

Sales figures for a fictional coffee chain:

<table>
<thead>
<tr>
<th>Product Type</th>
<th>N</th>
<th>{Coffee, Espresso, Herbal Tea, Tea}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>N</td>
<td>{Central, East, South, West}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sales</th>
<th>Q-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profit</td>
<td>Q-Ratio</td>
</tr>
<tr>
<td>Marketing</td>
<td>Q-Ratio</td>
</tr>
</tbody>
</table>
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.
Separation: Small Multiples

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

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

Scatter plots enabling pair-wise comparison of each data dimension.
Multiple Coordinated Views

- How long in majors
- Select high salaries
- Avg assists vs avg putouts (fielding ability)
- Avg career HRs vs avg career hits (batting ability)
- Distribution of positions played
Linking Assists to Positions
Dimensionality Reduction

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

1. Mean-center the data.
2. Find $\perp$ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.
Chernoff Faces (1973)

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

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key.
Parallel Coordinates
Parallel Coordinates [Inselberg]
Figure 1: The full dataset consisting of 473 batches
The Multidimensional Detective

The Dataset:

- 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 ... X12: 10 types of defects (zero defects shown at top)
  - X13 ... X16: 4 physical parameters

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

Parallel Coordinates

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, X1, and very high in Quality, X2.

Figure 7: Upper range of split in X15
Notice that $X_6$ behaves differently.
Allow 2 defects, including $X_6 \rightarrow$ best batches

Figure 1: The full dataset consisting of 473 batches
Radar Plot / Star Graph

“Parallel” dimensions in polar coordinate space
Best if same units apply to each axis
Tableau / Polaris
Polaris

Research at Stanford by Stolte, Tang, and Hanrahan.
Tableau

Data Model

Encodings

Data Display
Tableau Demo

The dataset:
Federal Elections Commission Receipts
Every Congressional Candidate from 1996 to 2002
4 Election Cycles
9216 Candidacies
Data Set 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
Assignment 2: 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 wiki notebook
- Keep record of your analysis
- Prepare a final graphic and caption

Due by 5:00pm
Monday, Jan 27
Polaris/Tableau 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 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 Chart]
Concatenation (+) Operator

Ordered union of set interpretations

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

Profit + Sales = \{(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 =

![Graph showing the cross-product of Quarter x Product Type and Product Type x Profit](image-url)
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 \textit{algebra} for tabular visualization.

Algebraic statements are then mapped to:
- \textbf{Visualizations} - trellis plot partitions, visual encodings
- \textbf{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)
## Ordinal - Ordinal

<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

- Start by visualizing individual dimensions
- Avoid "over-encoding"
- Use space and small multiples intelligently
- Use interaction to generate *relevant* views

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key.