Data and Image Models

Maneesh Agrawala

CS 294-10: Visualization
Fall 2013

Last Time: The Purpose of Visualization
Three functions of visualizations

Record information
- Photographs, blueprints, …

Support reasoning about information (analyze)
- Process and calculate
- Reason about data
- Feedback and interaction

Convey information to others (present)
- Share and persuade
- Collaborate and revise
- Emphasize important aspects of data

Record information

Gallop, Bay Horse “Daisy” [Muybridge 1884-86]
Analysis: Challenger

Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

Communicate: War Deaths

Crimean War Deaths [Nightingale 1858]
Announcements

Auditors, please enroll in the class (1 unit, P/NP)
- Requirements: Come to class and participate (online as well)
- Requirements: Assignment 1

Class participation requirements
- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question by 3pm on day of lecture

All, add yourself to participants page on the wiki

Class wiki
http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/
Assignment 1: Visualization Design

Due by 9am on Sep 11

Data and Image Models
The big picture

- **task**
- **data**
  - physical type: int, float, etc.
  - abstract type: nominal, ordinal, etc.
- **domain**
  - metadata
  - semantics
  - conceptual model

**processing algorithms**

**mapping**
- visual encoding
- visual metaphor

**image**
- visual channel
- retinal variables

---

Topics

- Properties of data or information
- Properties of the image
- Mapping data to images
### Data

#### Data models vs. Conceptual models

**Data models:** low level descriptions of the data
- Math: Sets with operations on them
- Example: integers with + and × operators

**Conceptual models:** mental constructions
- Include semantics and support reasoning

**Examples (data vs. conceptual)**
- (1D floats) vs. Temperature
- (3D vector of floats) vs. Space
Taxonomy

- 1D (sets and sequences)
- Temporal
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Trees (hierarchies)
- Networks (graphs)

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Schneiderman 96]

Types of variables

Physical types
- Characterized by storage format
- Characterized by machine operations

Example:
  - bool, short, int32, float, double, string, ...

Abstract types
- Provide descriptions of the data
- May be characterized by methods/attributes
- May be organized into a hierarchy

Example:
  - plants, animals, metazoans, ...

Nominal, ordinal and quantitative

N - Nominal (labels)
- Fruits: Apples, oranges, …

O - Ordered
- Quality of meat: Grade A, AA, AAA

Q - Interval (Location of zero arbitrary)
- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Like a geometric point. Cannot compare directly
- Only differences (i.e. intervals) may be compared

Q - Ratio (zero fixed)
- Physical measurement: Length, Mass, Temp, …
- Counts and amounts
- Like a geometric vector, origin is meaningful

S. S. Stevens, On the theory of scales of measurements,  1946
From data model to N,O,Q data type

Data model
- 32.5, 54.0, -17.3, …
- floats

Conceptual model
- Temperature

Data type
- Burned vs. Not burned (N)
- Hot, warm, cold (O)
- Continuous range of values (Q)

Iris Setosa
Iris Versicolor
Iris Virginica
Sepal and petal lengths and widths for three species of iris [Fisher 1936].

<table>
<thead>
<tr>
<th>ID</th>
<th>Case</th>
<th>Species_No</th>
<th>Species</th>
<th>Organ</th>
<th>Width</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Setosa</td>
<td>Petal</td>
<td>2</td>
<td>4.4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Virginica</td>
<td>Petal</td>
<td>2.6</td>
<td>6.9</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Setosa</td>
<td>Petal</td>
<td>4.4</td>
<td>6.9</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>Virginica</td>
<td>Petal</td>
<td>4.4</td>
<td>6.9</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>Setosa</td>
<td>Petal</td>
<td>4.4</td>
<td>6.9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>Virginica</td>
<td>Petal</td>
<td>4.4</td>
<td>6.9</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>Setosa</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>Virginica</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>Setosa</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>Virginica</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>Setosa</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
<td>Virginica</td>
<td>Sepal</td>
<td>3.4</td>
<td>5.1</td>
</tr>
</tbody>
</table>

[Excel spreadsheet image]
Relational data model

Represent data as a **table** *(relation)*

Each **row** *(tuple)* represents a single record
  
  Each record is a fixed-length tuple

Each **column** *(attribute)* represents a single **variable**
  
  Each attribute has a **name** and a **data type**

A table’s **schema** is the set of names and data types

A **database** is a collection of tables (relations)

---

Relational algebra [Codd]

Data transformations (SQL)

- **Selection** *(WHERE)* – restrict values
- **Projection** *(SELECT)* – choose subset of attributes
- **Sorting** *(ORDER BY)*
- **Aggregation** *(GROUP BY, SUM, MIN, …)*
- **Set operations** *(UNION, …)*
- **Combine** *(INNER JOIN, OUTER JOIN, …)*
Statistical data model

Variables or measurements
Categories or factors or dimensions
Observations or cases

<table>
<thead>
<tr>
<th>Month</th>
<th>Control</th>
<th>Placebo</th>
<th>300 mg</th>
<th>450 mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>March</td>
<td>165</td>
<td>163</td>
<td>166</td>
<td>168</td>
</tr>
<tr>
<td>April</td>
<td>162</td>
<td>159</td>
<td>161</td>
<td>163</td>
</tr>
<tr>
<td>May</td>
<td>164</td>
<td>158</td>
<td>161</td>
<td>153</td>
</tr>
<tr>
<td>June</td>
<td>162</td>
<td>161</td>
<td>158</td>
<td>160</td>
</tr>
<tr>
<td>July</td>
<td>166</td>
<td>158</td>
<td>160</td>
<td>148</td>
</tr>
<tr>
<td>August</td>
<td>163</td>
<td>158</td>
<td>157</td>
<td>150</td>
</tr>
</tbody>
</table>

Blood Pressure Study (4 treatments, 6 months)
Dimensions and measures

**Dimensions:** Discrete variables describing data
- Dates, categories of values (independent vars)

**Measures:** Data values that can be aggregated
- Numbers to be analyzed (dependent vars)
- Aggregate as sum, count, average, std. deviation

Dimensions and measures

**Independent vs. dependent variables**
- Example: \( y = f(x,a) \)
- **Dimensions:** Domain(\(x\)) \(\times\) Domain(\(a\))
- **Measures:** Range(\(y\))
Example: U.S. Census Data

People: # of people in group
Year: 1850 – 2000 (every decade)
Age: 0 – 90+
Sex: Male, Female
Marital Status: Single, Married, Divorced, ...

Example: U.S. Census

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>Sex</th>
<th>Marital Status</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>1850</td>
<td>0</td>
<td>0</td>
<td>Single</td>
<td>2348</td>
</tr>
<tr>
<td>1860</td>
<td>0</td>
<td>0</td>
<td>Married</td>
<td>2348</td>
</tr>
<tr>
<td>1870</td>
<td>0</td>
<td>0</td>
<td>Divorced</td>
<td>2348</td>
</tr>
</tbody>
</table>

2348 data points
### Census: N, O, Q?

<table>
<thead>
<tr>
<th>People Count</th>
<th>Q-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Q-Interval (O)</td>
</tr>
<tr>
<td>Age</td>
<td>Q-Ratio (O)</td>
</tr>
<tr>
<td>Sex (M/F)</td>
<td>N</td>
</tr>
<tr>
<td>Marital Status</td>
<td>N</td>
</tr>
</tbody>
</table>

### Census: Dimension or Measure?

<table>
<thead>
<tr>
<th>People Count</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Dimension</td>
</tr>
<tr>
<td>Age</td>
<td>Depends!</td>
</tr>
<tr>
<td>Sex (M/F)</td>
<td>Dimension</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Dimension</td>
</tr>
</tbody>
</table>
## Census: N, O, Q?

<table>
<thead>
<tr>
<th>People</th>
<th>Q-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Q-Interval (O)</td>
</tr>
<tr>
<td>Age</td>
<td>Q-Ratio (O)</td>
</tr>
<tr>
<td>Sex</td>
<td>N</td>
</tr>
<tr>
<td>Marital Status</td>
<td>N</td>
</tr>
</tbody>
</table>

## Roll-Up and Drill-Down

Want to examine marital status in each decade? Roll-up the data along the desired dimensions.

```sql
SELECT year, marst, sum(people) 
FROM census 
GROUP BY year, marst;
```
Roll-Up and Drill-Down

Need more detailed information?
Drill-down into additional dimensions

SELECT year, age, marst, sum(people)
FROM census
GROUP BY year, age, marst;
Image
Visual language is a sign system

Images perceived as a set of signs
Sender encodes information in signs
Receiver decodes information from signs

Semiology of Graphics, 1967

Jacques Bertin
Information in position

1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables

"Resemblance, order and proportional are the three signfields in graphics." - Bertin

Visual variables

- Position (x 2)
- Size
- Value
- Texture
- Color
- Orientation
- Shape

Note: Bertin does not consider 3D or time
Note: Card and Mackinlay extend the number of vars.

Information in color and value

Value is perceived as ordered
- Encode ordinal variables (O)

Hue is normally perceived as unordered
- Encode nominal variables (N) using color
Bertins’ “Levels of Organization”

<table>
<thead>
<tr>
<th>Position</th>
<th>N</th>
<th>O</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>N</td>
<td>O</td>
<td>Q</td>
</tr>
<tr>
<td>Value</td>
<td>N</td>
<td>O</td>
<td>Q</td>
</tr>
<tr>
<td>Texture</td>
<td>N</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shape</td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- N Nominal
- O Ordered
- Q Quantitative

Note: Q < O < N

Note: Bertin actually breaks visual variables down into differentiating (≠) and associating (≡)

Encoding rules
Univariate data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean

Middle 50%

Tukey box plot

Factors

Measure

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, …

Large design space (visual metaphors)

### Multidimensional data

How many variables can be depicted in an image?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“With up to three rows, a data table can be constructed directly as a single image ... However, an image has only three dimensions. And this barrier is impassible.”

Bertin
Deconstructions

Stock chart from the late 90s
Stock chart from the late 90s

- x-axis: time (Q)
- y-axis: price (Q)

---

Playfair 1786

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.
Playfair 1786

- x-axis: year (Q)
- y-axis: currency (Q)
- color: imports/exports (N, O)