The Value of Visualization

**Record** information
- Blueprints, photographs, seismographs, ...

**Analyze** data to support reasoning
- Develop and assess hypotheses
- Discover errors in data
- Expand memory
- Find patterns

**Communicate** information to others
- Share and persuade
- Collaborate and revise
Make a decision: Challenger

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

Info-Vis vs. Sci-Vis?

Visualization Reference Model
Data and Image Models

The Big Picture

Topics

Properties of data or information
Properties of the image
Mapping data to images
Data models vs. Conceptual models

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

Conceptual models are 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)
4D (relational)
Trees (hierarchies)
Networks (graphs)
Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 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 10, 2006; Location: (LAT 33°9.6', LONG -118°45.6')
- 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
Nominal, Ordinal and Quantitative

N - Nominal (labels)
  - Operations: ≠

O - Ordered
  - Operations: ≠, <, >

Q - Interval (Location of zero arbitrary)
  - Operations: ≠, <, >, -
  - Can measure distances or spans

Q - Ratio (zero fixed)
  - Operations: ≠, <, >, -, ÷
  - Can measure ratios or proportions

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 (°C)

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

Sepal and petal lengths and widths for three species of iris [Fisher, 1936].
**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)
- Projection (SELECT)
- Selection (WHERE)
- 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

---

**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

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

People
Year
Age
Sex
Marital Status

2348 data points

Census: N, O, Q?

**People Count**
**Year**
**Age**
**Sex (M/F)**
**Marital Status**

Q-Ratio
Q-Interval (O)
Q-Ratio (O)
N
N
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>

Roll-Up and Drill-Down

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

```
SELECT year, marst, sum(people)
FROM census
GROUP BY year, marst;
```

Dimensions Measure

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;
```
Which format might we prefer?

Row vs. Column-Oriented Databases

Relational Data Organizations

<table>
<thead>
<tr>
<th>Transactions</th>
<th>vs.</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row-oriented</td>
<td>Column-oriented</td>
<td></td>
</tr>
</tbody>
</table>
Relational Data Organizations

Row-oriented

Column-oriented

Speed-up Analysis
Reduce data transfer
Improved locality
Better data compression

Announcements

Auditors
- Requirements: Come to class and participate (online as well)

Class participation requirements
- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question on wiki within a day of each lecture

Class wiki: [http://cs448b.stanford.edu](http://cs448b.stanford.edu)
**Assignment 1: Visualization Design**

Design a static visualization for a given data set.

**Deliverables** (post to the course wiki)

- Image of your visualization
- Short description and design rationale (≤ 4 para.)

Due by **7:00am on Tuesday 10/2**.

Questions?
**Visual language is a sign system**

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

Jacques Bertin
Sémiologie Graphique, 1967

---

**Bertin’s Semiology of Graphics**

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 proportion are the three signifieds in graphics.” - Bertin

---

**Visual encoding variables**

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape

---
Visual encoding variables

Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape
Transparency
Blur / Focus ...

Information in color and value

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

∵ Encode continuous variables (Q) [not as well]

Hue is normally perceived as unordered
∴ Encode nominal variables (N) using color

Bertin’s “Levels of Organization”

Design Space of Visual Encodings

Note: Bertin actually breaks visual variables down into differentiating (∉) and associating (∈)
Univariate data

Mean

Tukey box plot

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

Playfair 1786

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

http://www.smartmoney.com/marketmap/
Wattenberg 1998

rectangle size: market cap (Q)
rectangle position: market sector (N), market cap (Q)
color hue: loss vs. gain (N, O)
color value: magnitude of loss or gain (Q)

Minard 1869: Napoleon’s march

Single axis composition

Mark composition

y-axis: temperature (Q)
x-axis: longitude (Q) / time (O)

= temp over space/time (Q x Q)

[based on slide from Mackinlay]
Mark composition

y-axis: longitude (Q)

x-axis: latitude (Q)

width: army size (Q)

= army position (Q x Q) and army size (Q)

[based on slide from Mackinlay]

Minard 1869: Napoleon’s march

Depicts at least 5 quantitative variables. Any others?

Formalizing Design
(Mackinlay 1986)
Choosing Visual Encodings

Challenge:
Assume 8 visual encodings and n data attributes. We would like to pick the “best” encoding among a combinatorial set of possibilities with size \((n+1)^8\)

Principle of Consistency:
The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering:
Encode the most important information in the most effective way.

Design Criteria (Mackinlay)

Expressiveness
A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Cannot express the facts
A one-to-many \((1 \rightarrow N)\) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position

Expresses facts not in the data
A length is interpreted as a quantitative value;
\[\therefore\text{Length of bar says something untrue about } N \text{ data} \]

[Mackinlay, APT, 1986]
Design Criteria (Mackinlay)

Expressiveness
A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness
A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

(Effectiveness subject of the Graphical Perception lecture)

Mackinlay’s Design Algorithm

User formally specifies data model and type
• Additional input: ordered list of data variables to show

APT searches over design space
• Tests expressiveness of each visual encoding
• Generates specification for encodings that pass test
• Tests perceptual effectiveness of resulting image

Outputs the “most effective” visualization

Mackinlay’s Ranking

Quantitative
• Position
• Length
• Angle
• Slope
• Area
• Volume
• Density
• Saturation
• Hue
• Texture
• Connection
• Containment

Ordinal
• Position
• Density
• Saturation
• Hue
• Texture
• Connection
• Containment

Nominal
• Position
• Hue
• Texture
• Connection
• Containment
• Length
• Angle
• Slope
• Area
• Volume

Conjectured effectiveness of the encoding

Limitations

Does not cover many visualization techniques
• Bertin and others discuss networks, maps, diagrams
• Does not consider 3D, animation, illustration, photography, ...

Does not model interaction
Summary

- Formal specification
  - Data model
  - Image model
  - Encodings mapping data to image

- Choose expressive and effective encodings
  - Formal test of expressiveness
  - Experimental tests of perceptual effectiveness

Assignment 1: Visualization Design

Design a static visualization for a given data set.

Deliverables (post to the course wiki)
- Image of your visualization
- Short description and design rationale (≤ 4 para.)

Due by 7:00am on Tuesday 10/2.