Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Fall 2021

The big picture

- **task**: questions, goals, assumptions
- **data**: abstract type (nominal, ordinal, etc.)
- **domain**: metadata, semantics, conceptual model, conventions

- **processing algorithms**
- **mapping**: visual encoding
- **image**: graphical marks, visual channel
Topics

Properties of data
Properties of the image
Mapping data to images

Data
# Data models vs. Conceptual models

**Data models** are formal descriptions
- 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. spatial location

## Taxonomy of Data Models/Types

- 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]
Nominal, ordinal and quantitative

N - Nominal (labels)
Fruits: Apples, oranges, ...
Operations: ≠, ≠

O - Ordered
Quality of meat: Grade A, AA, AAA
Operations: =, ≤, <, >

Q - Interval (location of zero arbitrary)
Dates: Jan, 19, 2016; Loc.: (LAT 33.98, LON -118.45)
Like a geometric point. Cannot compare directly
Only differences (i.e. intervals) may be compared
Operations: =, ≠, <, >, ≈

Q - Ratio (location of zero fixed)
Physical measurement: Length, Mass, ...
Counts and amounts
Like a geometric vector, origin is meaningful
Operations: =, ≠, <, >, ≈, ÷

From data model to N,O,Q

Data model
- 32.5, 54.0, -17.3, ...
- Floating point numbers

Conceptual model
- Temperature (°C)

N,O,Q
- Burned vs. Not burned (N)
- Hot, warm, cold (O)
- Continuous range of values (Q-Int)
Dimensions and measures

Dimensions: (~ independent variables)
Often discrete variables describing data (N, O)
Categories, dates, binned values

Measures: (~ dependent variables)
Data values that can be aggregated (Q)
Numbers to be analyzed
Aggregate as sum, count, average, std. deviation

Distinction is not strict. The same variable may be treated either way depending on the task.

Example: U.S. Census Data

People Count: # of people in group
Year: 1850 – 2000 (every decade)
Age: 0 – 90+
Sex: Male, Female
Marital Status: Single, Married, Divorced, ...
### 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</td>
<td>N</td>
</tr>
<tr>
<td>Marital Status</td>
<td>N</td>
</tr>
</tbody>
</table>

2348 data points

### Census: Dim. or Meas.?

<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</td>
<td>Dimension</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Dimension</td>
</tr>
</tbody>
</table>

2348 data points
Data Tables and Transformations

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 attribute names and data types

A database is a collection of tables (relations)
Relational algebra [Codd 1970] / SQL

Operations on data tables: table(s) in, table out
- Projection (SELECT) – select a set of columns
- Selection (WHERE) – filter rows
- Sorting (ORDER BY) – order rows
- Aggregation (GROUP BY, SUM, MIN, …) partition rows into groups and summarize
- Combination (JOIN, UNION, …) integrate data from multiple tables

Projection (SELECT) – select a set of columns

<table>
<thead>
<tr>
<th>day</th>
<th>stock</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3</td>
<td>AMZN</td>
<td>957.10</td>
</tr>
<tr>
<td>10/3</td>
<td>MSFT</td>
<td>74.26</td>
</tr>
<tr>
<td>10/4</td>
<td>AMZN</td>
<td>965.45</td>
</tr>
<tr>
<td>10/4</td>
<td>MSFT</td>
<td>74.69</td>
</tr>
</tbody>
</table>
### Relational algebra [Codd 1970] / SQL

**Selection (WHERE) – filter rows**

\[
\text{select * where price > 100}
\]

<table>
<thead>
<tr>
<th>day</th>
<th>stock</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3</td>
<td>AMZN</td>
<td>957.10</td>
</tr>
<tr>
<td>10/3</td>
<td>MSFT</td>
<td>74.26</td>
</tr>
<tr>
<td>10/4</td>
<td>AMZN</td>
<td>965.45</td>
</tr>
<tr>
<td>10/4</td>
<td>MSFT</td>
<td>74.69</td>
</tr>
</tbody>
</table>

### Relational algebra [Codd 1970] / SQL

**Sorting (ORDER BY) – order records**

\[
\text{select * order by stock}
\]

<table>
<thead>
<tr>
<th>day</th>
<th>stock</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3</td>
<td>AMZN</td>
<td>957.10</td>
</tr>
<tr>
<td>10/3</td>
<td>MSFT</td>
<td>74.26</td>
</tr>
<tr>
<td>10/4</td>
<td>AMZN</td>
<td>965.45</td>
</tr>
<tr>
<td>10/4</td>
<td>MSFT</td>
<td>74.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>day</th>
<th>stock</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3</td>
<td>AMZN</td>
<td>957.10</td>
</tr>
<tr>
<td>10/4</td>
<td>AMZN</td>
<td>965.45</td>
</tr>
<tr>
<td>10/3</td>
<td>MSFT</td>
<td>74.26</td>
</tr>
<tr>
<td>10/4</td>
<td>MSFT</td>
<td>74.69</td>
</tr>
</tbody>
</table>
Relational algebra [Codd 1970] / SQL

Aggregation (GROUP BY, SUM, MIN, …)

select stock, min(price) group by stock

<table>
<thead>
<tr>
<th>day</th>
<th>stock</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/3</td>
<td>AMZN</td>
<td>957.10</td>
</tr>
<tr>
<td>10/3</td>
<td>MSFT</td>
<td>74.26</td>
</tr>
<tr>
<td>10/4</td>
<td>AMZN</td>
<td>965.45</td>
</tr>
<tr>
<td>10/4</td>
<td>MSFT</td>
<td>74.69</td>
</tr>
</tbody>
</table>

Roll-Up and Drill-Down

Want to examine population by year and age? Roll-up the data (i.e. aggregate) along marst.

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

Want to breakdown by marital status? **Drill-down** into additional dimensions

```sql
SELECT year, age, marst sum(people)
FROM census
GROUP BY year, age, marst
```
Which format might we prefer? Why?
Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Advantage: Flexible starting point for analysis, transformation, and visualization. Our pivoted table variant was not “tidy”!

Common Data Formats

CSV: Comma-Separated Values

```
year, age, marst, sex, people
1850, 0, 0, 1, 1483789
1850, 5, 0, 1, 1411067
...```

Common Data Formats

**CSV: Comma-Separated Values**

```
year,age,marst,sex,people
1850,0,0,1,1483789
1850,5,0,1,1411067
...
```

**JSON: JavaScript Object Notation**

```
[
    {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},
    {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},
    ...
]
```

Announcements

**Class participation requirements**

- Complete readings and notebooks before class
- In-class discussion
- Post at least 1 discussion substantive comment/question per week.
- Due by 7am the following Monday
- 1 pass for the quarter

Class home page

[https://magrawala.github.io/cs448b-fa21/](https://magrawala.github.io/cs448b-fa21/)
Reading/Notebook/Lecture Responses

Good responses typically exhibit one or more

- Critiques of arguments made in the papers/lectures
- Analysis of implications or future directions for ideas in readings/lectures
- Insightful questions about the readings/lectures

Responses should not be summaries

---

Observable Notebooks – Vega-Lite

Vega-Lite is a declarative API for producing visualizations

Make sure to go through a do exercises (fork the notebook)

Monday 9/27 lecture will assume you’ve done 1st three notebooks
Office Hours

**Maneesh:** 2-3pm Wed, Coupa Café Y2E2 and Canvas/Zoom
**Dae Hyun:** 10-11am Thu, CEMEX Aud and Canvas/Zoom
**Shana Hadi:** 7-8:00pm Sun, via Canvas/Zoom

Happy to schedule other OH by appointment
Outside of OH use Slack to connect with us

https://canvas.stanford.edu/courses/144332/external_tools/11232

Assignment 1: Visualization Design

Design a static visualization for a data set.
You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

**Data: Stanford Undergraduate Majors**

The Stanford Daily publishes a variety of datasets through the Stanford Open Data Portal. They have published a data table containing information about the number of Stanford students majoring in 70 different subject areas from 2011-2019. We have filtered and wrangled this data to the top 10 majors over the time period to produce a dataset with the following variables:

- **Number of records:**
- **Variable Names:**
  - **Year:** Academic year between 2011-2012 and 2018-2019.
  - **Subject:** Subject areas in which students majored.
  - **Number of Students:** Number of students majoring in the area.

The extracted dataset is available in csv format: StanfordTopTenMajors2010s.csv

Due by 7am on Mon Sep 27
Assignment 1: Visualization Design

Pick a guiding question, use it to title your visualization
Design a static visualization for that question
You are free to use any tools (including pen & paper)

Deliverables (upload via Canvas; see A1 page)
PDF of your visualization with a short description including design rationale (≤ 4 paragraphs)

Due by 7am on Mon Sep 27
Marks and Visual Variables

Marks: geometric primitives
- points
- lines
- areas

Visual Variables: control mark appearance
- Position (2x)
- Size
- Value
- Texture
- Color
- Orientation
- Shape

Semiology of Graphics
J. Bertin, 1967

Coding information in position

1. A, B, C are distinguishable
2. Three pts colinear: B 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
**Coding info 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”**

<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
Visual Encoding

Encodings: Map Data to Mark Attr.

mark: rect
data \rightarrow size (height)
Encodings: Map Data to Mark Attr.

mark: rect
data $\rightarrow$ size (height)

mark: point
data$_1$ $\rightarrow$ x-pos
data$_2$ $\rightarrow$ y-pos

Encodings: Map Data to Mark Attr.

mark: rect
data $\rightarrow$ size (height)

mark: point
data$_1$ $\rightarrow$ x-pos
data$_2$ $\rightarrow$ y-pos

mark: point
data$_1$ $\rightarrow$ x-pos
data$_2$ $\rightarrow$ y-pos
data$_3$ $\rightarrow$ color
Encodings: Map Data to Mark Attr.

mark: rect
data → size (height)

mark: point
data₁ → x-pos
data₂ → y-pos

mark: point
data₁ → x-pos
data₂ → y-pos
data₃ → color

mark: point
data₁ → x-pos
data₂ → y-pos
data₃ → color
data₄ → size

Deconstructions
Playfair 1786/1801

- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q)
- Exports/Imports → color (N, O)
- Balance for/against → area (maybe length??) (Q)
- Balance for/against → color (N, O)
Map of the Market [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)

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

90

91
Minard 1869: Napoleon’s march

Carte figurative des pertes humaines en la campagne de la Russie, 1812–1813. Made by A. Minard, Ingenieur, Special assistant to Napoleon, in 1869. Example of a single-axis composition based on a slide from Mackinlay.

Single axis composition

[based on slide from Mackinlay]
**Mark composition**

temperature $\rightarrow$ y-position (Q, linear)

+ longitude $\rightarrow$ x-position (Q, linear)

= temp over longitude (Q x Q)

[based on slide from Mackinlay]

---

**Mark composition**

latitude $\rightarrow$ y-position (Q)

longitude $\rightarrow$ x-position (Q)

+ army size $\rightarrow$ width (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?
Automated design

Jock Mackinlay’s APT 86

Combinatorics of encodings

Challenge:
Assume 8 visual encodings and n data fields
Pick the best encoding from the exponential number of possibilities \((n+1)^8\)
Principles

**Challenge:**
Assume 8 visual encodings and n data fields
Pick the best encoding from the exponential number of possibilities $(n+1)^8$

**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

---

Mackinlay’s expressiveness criteria

**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 → 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; ✗ Length of bar says something untrue about N data.

Fig. 11. Incorrect use of a bar chart for the Nation relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the Nation relation.

[Mackinlay, APT, 1986]
**Mackinlay’s effectiveness criteria**

**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.

**Subject of perception lecture**

---

**Mackinlay’s ranking**

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Ordinal</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density</td>
<td>Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Saturation</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Hue</td>
<td>Connection</td>
</tr>
<tr>
<td>Area</td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density</td>
</tr>
<tr>
<td>Density</td>
<td>Containment</td>
<td>Saturation</td>
</tr>
<tr>
<td>Saturation</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>

**Conjectured effectiveness of the encoding**
Mackinlay’s Design Algorithm

User formally specifies data model and type
Input: list of data variables ordered by importance

APT searches over design space
    Tests expressiveness of each visual encoding (rule-based)
    Generates encodings that pass test
    Rank by perceptual effectiveness criteria
Outputs most effective visualization

Encode most important data using highest ranking visual variable for the data type

<table>
<thead>
<tr>
<th>Price</th>
<th>Mileage</th>
<th>Weight</th>
<th>Repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>13,500</td>
<td>22</td>
<td>3000</td>
<td>great</td>
</tr>
<tr>
<td>7,200</td>
<td>31</td>
<td>1500</td>
<td>ok</td>
</tr>
<tr>
<td>11,300</td>
<td>12</td>
<td>4200</td>
<td>terrible</td>
</tr>
</tbody>
</table>

mark: lines
Price \( \rightarrow \) y-pos (Q)
Mileage \( \rightarrow \) x-pos (Q)
Weight \( \rightarrow \) size (Q)
Repair \( \rightarrow \) color (N)

Automating the design of graphical presentation of relational information
J. Mackinlay, 1986
Cars Data
1. Price (Q)
2. Mileage (Q)
3. Weight (Q)
4. Repair (Q)

Limitations

Does not cover many visualization techniques
- Networks, maps, diagrams
- Also, 3D, animation, illustration, ...

Does not consider interaction
Does not consider semantics or conventions
Assumes single visualization as output
Summary

Formal specification
- Data model: relational data, N,O,O types
- Image model: marks, attributes, encodings
- Encodings mapping data to image

Choose expressive and effective encodings
- Rule-based test of expressiveness
- Perceptual effectiveness rankings
Sepal and petal lengths and widths for three species of iris (Fisher 1936).

<table>
<thead>
<tr>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>4.4</td>
<td>1.7</td>
<td>0.4</td>
</tr>
<tr>
<td>5.7</td>
<td>5.6</td>
<td>2.8</td>
<td>3.8</td>
</tr>
<tr>
<td>6.3</td>
<td>3.3</td>
<td>5.5</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Format of the data in Appendix 14, pp. 365-366
Chambers, Cleveland, Kleiner, Tukey, *Graphical Methods for Data Analysis*
Bertin’s specification

STANDARD SCHEMAS

1 COMPONENT  2 COMPONENTS  3 COMPONENTS  MORE THAN 3 COMPONENTS

DIAGRAMS

NETWORKS

MAPS

COMPREHENSIVE IMAGE  PROCESSING GRAPHICS (Comprehensive)  INVENTORIES (Comprehensive)  MESSAGES (Simplified)


Communicate: Exports and Imports

Exports and Imports to and from DENMARK & NORWAY from 1600 to 1780

[Brewer 1827]

[Playfair 1786]
Data cube

Measure

Width

Length

Petal

Sepal

Organ

I. setosa

I. versicolor

I. virginica

Species

Projections summarize data

Multiscale visualization using data cubes [Stolte et al. 02]