Quality of Visualization
Origins

Based on the CHI 2003 Tutorial by Marti Hearst and a Tutorial by Tamara Munzner
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Introduction

Goals of Information Visualization
Case Study: The Journey of the TreeMap
Key Questions
What is Information Visualization?

Visualize: to form a mental image or vision of …

Visualize: to imagine or remember as if actually seeing.

American Heritage dictionary, Concise Oxford dictionary
What is Information Visualization?

“Transformation of the symbolic into the geometric”
(McCormick et al., 1987)

“... finding the artificial memory that best supports our natural means of perception.”
(Bertin, 1983)

The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skill by making use of the visual system.
Information Visualization

Problem:

HUGE Datasets: How to understand them?

Solution

Take better advantage of human perceptual system
Convert information into a graphical representation.

Issues

How to convert abstract information into graphical form?
Do visualizations do a better job than other methods?
6 Quality of Visualization

Visualization

Success Stories
The Power of Visualization

1. Start out going Southwest on ELLSWORTH AVE
   Towards BROADWAY by turning right.
2. Turn RIGHT onto BROADWAY.
3. Turn RIGHT onto QUINCY ST.
4. Turn LEFT onto CAMBRIDGE ST.
5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
6. Turn RIGHT onto RUSSELL ST.
The Power of Visualization

The estimated travel time is 5 minutes for 2.16 miles of travel, total of 6 steps.

<table>
<thead>
<tr>
<th>Directions</th>
<th>Elapsed Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Begin at <strong>17 Ellsworth Ave</strong> on Ellsworth Ave and go Southwest for 0.1 miles to 17 Ellsworth Ave</td>
<td>0.1</td>
</tr>
<tr>
<td>2 Turn right on <strong>Broadway</strong> and go Northwest for 0.5 miles</td>
<td>0.5</td>
</tr>
<tr>
<td>3 Turn right on <strong>Quincy St</strong> and go North for 0.5 miles</td>
<td>0.5</td>
</tr>
<tr>
<td>4 Turn left on <strong>Cambridge St</strong> and go West for 0.8 miles</td>
<td>0.8</td>
</tr>
<tr>
<td>5 Bear right on <strong>Massachusetts Ave,Mass Ave,RT-2A</strong> and go North for 2.0 miles</td>
<td>2.0</td>
</tr>
<tr>
<td>6 Turn right on <strong>Russell St</strong> and go Northeast for 2.2 miles to 77 Russell St</td>
<td>2.2</td>
</tr>
</tbody>
</table>
Visualization Success Story
Illustration of John Snow’s deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

Horizontal lines indicate location of deaths.

Illustration of John Snow’s deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

Horizontal lines indicate location of deaths.

Purposes of Information Visualization

To help:
  Explore
  Calculate
  Communicate
  Decorate
Two Different Primary Goals:
Two Different Types of Vis

Explore/Calculate
  Analyze
  Reason about Information

Communicate
  Explain
  Make Decisions
  Reason about Information
Goals of Information Visualization

More specifically, visualization should:

- Make large datasets coherent
  (Present huge amounts of information compactly)
- Present information from various viewpoints
- Present information at several levels of detail
  (from overviews to fine structure)
- Support visual comparisons
- Tell stories about the data
Why Visualization?

Use the eye for pattern recognition; people are good at
  scanning
  recognizing
  remembering images

Graphical elements facilitate comparisons via
  length
  shape
  orientation
  texture

Animation shows changes across time

Color helps make distinctions

Aesthetics make the process appealing
The Need for Critical Analysis

We see many creative ideas, but they often fail in practice

The hard part: how to apply it judiciously
  Inventors usually do not accurately predict how their invention will be used

This tutorial will emphasize
  Getting past the coolness factor
  Examining usability studies
Case Study:
The Journey of the TreeMap

The TreeMap (Johnson & Shneiderman ‘91)

Idea:

Show a hierarchy as a 2D layout
Fill up the space with rectangles representing objects
Size on screen indicates relative size of underlying objects.
Early Treemap Applied to File System
Treemap Problems

Too disorderly
  What does adjacency mean?
  Aspect ratios uncontrolled leads to lots of skinny boxes that clutter

Color not used appropriately
  In fact, is meaningless here

Wrong application
  Don’t need all this to just see the largest files in the OS
Successful Application of Treemaps

Think more about the use
  Break into meaningful groups
  Fix these into a useful aspect ratio

Use visual properties properly
  Use color to distinguish meaningfully
    Use only two colors:
      Can then distinguish one thing from another
    When exact numbers aren’t very important

Provide excellent interactivity
  Access to the real data
  Makes it into a useful tool
TreeMaps in Action
A Good Use of TreeMaps and Interactivity
Treemaps in Peet’s site

COFFEE SELECTOR

CLICK HERE FOR HELP.
Analysis vs. Communication

MarketMap’s use of TreeMaps allows for sophisticated analysis

Peets’ use of TreeMaps is more for presentation and communication

This is a key contrast
Open Issues

Does visualization help?

The jury is still out

Still supplemental at best for text collections

A correlation with spatial ability

Learning effects: with practice ability on visual display begins to equal that of text

Does visualization sell?

Jury is still out on this one too!

This is a *hot* area! More ideas will appear!
Key Questions to Ask about a Viz

1. What does it teach/show/elucidate?
2. What is the key contribution?
3. What are some compelling, *useful* examples?
4. Could it have been done more simply?
5. Have there been usability studies done? What do they show?
What we are *not* covering

- Scientific visualization
- Statistics
- Cartography (maps)
- Education
- Games
- Computer graphics in general
- Computational geometry
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Visual Principles
Visual Principles

Types of Graphs
Pre-attentive Properties
Relative Expressiveness of Visual Cues
Visual Illusions
Tufte’s notions
  Graphical Excellence
  Data-Ink Ratio Maximization
  How to Lie with Visualization
References for Visual Principles

Kosslyn: Types of Visual Representations
Lohse et al: How do people perceive common graphic displays
Bertin, MacKinlay: Perceptual properties and visual features
Tufte/Wainer: How to mislead with graphs
A Graph is: (Kosslyn)

A visual display that illustrates one or more relationships among entities

A shorthand way to present information

Allows a trend, pattern, or comparison to be easily apprehended
6 Quality of Visualization

Types of Symbolic Displays (Kosslyn 89)

- Graphs
- Charts
- Maps
- Diagrams
Types of Symbolic Displays

Graphs

at least two scales required
values associated by a symmetric “paired with” relation

Examples: scatter-plot, bar-chart, layer-graph
Types of Symbolic Displays

Charts

discrete relations among discrete entities
structure relates entities to one another
lines and relative position serve as links

Examples:
family tree
flow chart
network diagram
Types of Symbolic Displays

Maps

- internal relations determined (in part) by the spatial relations of what is pictured
- labels paired with locations

Examples:
- map of census data
- topographic maps
Types of Symbolic Displays

Diagrams

- schematic pictures of objects or entities
- parts are symbolic (unlike photographs)
- how-to illustrations
- figures in a manual

Anatomy of a Graph (Kosslyn 89)

Framework
  sets the stage
  kinds of measurements, scale, ...

Content
  marks
  point symbols, lines, areas, bars, ...

Labels
  title, axes, tic marks, ...
Basic Types of Data

Nominal (qualitative)
(no inherent order)
 city names, types of diseases, ...

Ordinal (qualitative)
(ordered, but not at measurable intervals)
 first, second, third, …
 cold, warm, hot

Interval (quantitative)
 list of integers or reals
Common Graph Types
Combining Data Types in Graphs

Examples?

<table>
<thead>
<tr>
<th>Nominal</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Nominal</td>
<td>Interval</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Interval</td>
</tr>
<tr>
<td>Interval</td>
<td>Interval</td>
</tr>
</tbody>
</table>
Scatter Plots

Qualitatively determine if variables

are highly correlated
linear mapping between horizontal & vertical axes

have low correlation
spherical, rectangular, or irregular distributions

have a nonlinear relationship
a curvature in the pattern of plotted points

Place points of interest in context

color representing special entities
When to use which type?

Line graph
  x-axis requires quantitative variable
  Variables have contiguous values
  familiar/conventional ordering among ordinals

Bar graph
  comparison of relative point values

Scatter plot
  convey overall impression of relationship between two variables

Pie Chart?
  Emphasizing differences in proportion among a few numbers
Classifying Visual Representations

Lohse, G L; Biolsi, K; Walker, N and H H Rueter,
A Classification of Visual Representations
CACM, Vol. 37, No. 12, pp 36-49, 1994

Participants sorted 60 items into categories

Other participants assigned labels from Likert scales

Experimenters clustered the results various ways.
Subset of Example Visual Representations
From Lohse et al. 94

1. soil triangle
2. missile crisis
3. USA tornadoes
4. 
5. microscope
6. circular tree diagram
7. auto repair records
Department of Computer Science and Engineering

Quality of Visualization

Subset of Example Visual Representations

From Lohse et al.

10. gibberellins
11. heart
12. microbiology
14. Gantt chart
15. Pittsburgh map
16. list of integrals
18. data model
19. response surface
20. wheelbarrow
Likert Scales
(and percentage of variance explained)
16.0  emphasizes whole – parts
11.3  spatial – nonspatial
10.6  static structure – dynamic structure
10.5  continuous – discrete
10.3  attractive – unattractive
10.1  nontemporal – temporal
  9.9  concrete – abstract
  9.6  hard to understand – easy
  9.5  nonnumeric – numeric
  2.2  conveys a lot of info – conveys little
Experimentally Motivated Classification
(Lohse et al. 94)

- Graphs
- Tables (numerical)
- Tables (graphical)
- Charts (time)
- Charts (network)
- Diagrams (structure)
- Diagrams (network)
- Maps
- Cartograms
- Icons
- Pictures
Interesting Findings

Lohse et al. 94

Photorealistic images were least informative
   Echos results in icon studies – better to use less complex, more schematic images

Graphs and tables are the most self-similar categories
   Results in the literature comparing these are inconclusive

Cartograms were hard to understand
   Echos other results – better to put points into a framed rectangle to aid spatial perception

Temporal data more difficult to show than cyclic data
   Recommend using animation for temporal data
Visual Properties

Preattentive Processing

Accuracy of Interpretation of Visual Properties

Illusions and the Relation to Graphical Integrity

All Preattentive Processing figures from Healey 97

Department of Computer Science and Engineering
Preattentive Processing

A limited set of visual properties are processed preattentively
(without need for focusing attention).
This is important for design of visualizations
what can be perceived immediately
what properties are good discriminators
what can mislead viewers
Example: Color Selection

Viewer can rapidly and accurately determine whether the target (red circle) is present or absent. Difference detected in color.
Example: Shape Selection

Viewer can rapidly and accurately determine whether the target (red circle) is present or absent. Difference detected in form (curvature)
Pre-attentive Processing

< 200 - 250ms qualifies as pre-attentive

- eye movements take at least 200ms
- yet certain processing can be done very quickly, implying low-level processing in parallel

If a decision takes a fixed amount of time regardless of the number of distractors, it is considered to be preattentive.
Example: Conjunction of Features

Viewer *cannot* rapidly and accurately determine whether the target (red circle) is present or absent when target has two or more features, each of which are present in the distractors. Viewer must search sequentially.

All Pre-attentive Processing figures from Healey 97

Department of Computer Science and Engineering
Example: Emergent Features

Target has a unique feature with respect to distractors (open sides) and so the group can be detected preattentively.
Example: Emergent Features

Target does not have a unique feature with respect to distractors and so the group cannot be detected preattentively.
Asymmetric and Graded Preattentive Properties

Some properties are asymmetric
  a sloped line among vertical lines is preattentive
  a vertical line among sloped ones is not

Some properties have a gradation
  some more easily discriminated among than others
Use Grouping of Well-Chosen Shapes for Displaying Multivariate Data
SUBJECT PUNCHED QUICKLY OXIDIZED  TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC SUBJECT PUNCHED QUICKLY OXIDIZED  TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC
Text NOT Preattentive

SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCR SNMULOC
### Preattentive Visual Properties

(Healey 97)

<table>
<thead>
<tr>
<th>Property</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>width</td>
<td>Julesz [1985]</td>
</tr>
<tr>
<td>size</td>
<td>Triesman &amp; Gelade [1980]</td>
</tr>
<tr>
<td>curvature</td>
<td>Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>number</td>
<td>Julesz [1985]; Trick &amp; Pylyshyn [1994]</td>
</tr>
<tr>
<td>terminators</td>
<td>Julesz &amp; Bergen [1983]</td>
</tr>
<tr>
<td>intersection</td>
<td>Julesz &amp; Bergen [1983]</td>
</tr>
<tr>
<td>closure</td>
<td>Enns [1986]; Triesman &amp; Souther [1985]</td>
</tr>
<tr>
<td></td>
<td>Kawai et al. [1995]; Bauer et al. [1996]</td>
</tr>
<tr>
<td>intensity</td>
<td>Beck et al. [1983]; Triesman &amp; Gormican [1988]</td>
</tr>
<tr>
<td>flicker</td>
<td>Julesz [1971]</td>
</tr>
<tr>
<td>direction of motion</td>
<td>Nakayama &amp; Silverman [1986]; Driver &amp; McLeod [1992]</td>
</tr>
<tr>
<td>binocular lustre</td>
<td>Wolfe &amp; Franzel [1988]</td>
</tr>
<tr>
<td>stereoscopic depth</td>
<td>Nakayama &amp; Silverman [1986]</td>
</tr>
<tr>
<td>3-D depth cues</td>
<td>Enns [1990]</td>
</tr>
<tr>
<td>lighting direction</td>
<td>Enns [1990]</td>
</tr>
</tbody>
</table>
Gestalt Properties

*Gestalt:* form or configuration

Idea: forms or patterns transcend the stimuli used to create them.

  Why do patterns emerge?
  Under what circumstances?

---

Why perceive pairs vs. triplets?
Figure and Ground

Escher illustrations are good examples
Vase/Face contrast

Subjective Contour
More Gestalt Laws

Law of Proximity
Stimulus elements that are close together will be perceived as a group

Law of Similarity
like the preattentive processing examples

Law of Common Fate
like preattentive motion property
move a subset of objects among similar ones and they will be perceived as a group
Which Properties are Appropriate for Which Information Types?
Accuracy Ranking of Quantitative Perceptual Tasks
Estimated; only pairwise comparisons have been validated
(Mackinlay 88 from Cleveland & McGill)
Interpretations of Visual Properties

Some properties can be discriminated more accurately but don’t have intrinsic meaning

(Senay & Ingatious 97, Kosslyn, others)

Density (Greyscale)
  Darker -> More

Size / Length / Area
  Larger -> More

Position
  Leftmost -> first, Topmost -> first

Hue
  ??? no intrinsic meaning

Slope
  ??? no intrinsic meaning
# Ranking of Applicability of Properties for Different Data Types

(Mackinlay 88, Not Empirically Verified)

<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>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Saturation</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color 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>Color Saturation</td>
</tr>
<tr>
<td>Color Saturation</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
</tbody>
</table>
Color Schemes

Order these (low->hi)

- Purple
- Green
- Orange
- Yellow
- Red
Color Schemes

- Gray scale
- Full spectral scale
- Single sequence part spectral scale
- Single sequence single hue scale
- Double-ended multiple hue scale
Color Purposes

Call attention to specific items
Distinguish between classes of items
  Increases the number of dimensions for encoding
Increase the appeal of the visualization
Using Color

Proceed with caution

- Less is more
- Representing magnitude is tricky

Examples

- **Red-orange-yellow-white**
  - Works for costs
  - Maybe because people are very experienced at reasoning shrewdly according to cost

- **Green-light green-light brown-dark brown-grey-white** works for atlases

**Grayscale** is unambiguous but has limited range
Visual Illusions

People don’t perceive length, area, angle, brightness they way they “should”.

Some illusions have been reclassified as systematic perceptual errors

  e.g., brightness contrasts (grey square on white background vs. on black background)

  partly due to increase in our understanding of the relevant parts of the visual system

Nevertheless, the visual system does some really unexpected things.
Illusions of Linear Extent

Mueller-Lyon (off by 25-30%)

Horizontal-Vertical
Illusions of Area

Delboeuf Illusion

![Illustration of Delboeuf Illusion]

Height of 4-story building overestimated by approximately 25%
The two filled black circles are exactly the same size; however, the one on the left may *seem* larger or smaller.
What are good guidelines for Infoviz?

- Use graphics appropriately
  - Don’t use images gratuitously
  - Don’t lie with graphics!
    - Link to original data
  - Don’t conflate area with other information
    - E.g., use area in map to imply amount

- Make it interactive (feedback)
  - Brushing and linking
  - Multiple views
  - Overview + details

- Match mental models
## Channels: Rankings

<table>
<thead>
<tr>
<th>Category</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position on common scale</td>
<td><img src="image" alt="Position on common scale" /></td>
</tr>
<tr>
<td>Position on unaligned scale</td>
<td><img src="image" alt="Position on unaligned scale" /></td>
</tr>
<tr>
<td>Length (1D size)</td>
<td><img src="image" alt="Length (1D size)" /></td>
</tr>
<tr>
<td>Tilt/angle</td>
<td><img src="image" alt="Tilt/angle" /></td>
</tr>
<tr>
<td>Area (2D size)</td>
<td><img src="image" alt="Area (2D size)" /></td>
</tr>
<tr>
<td>Depth (3D position)</td>
<td><img src="image" alt="Depth (3D position)" /></td>
</tr>
<tr>
<td>Color luminance</td>
<td><img src="image" alt="Color luminance" /></td>
</tr>
<tr>
<td>Color saturation</td>
<td><img src="image" alt="Color saturation" /></td>
</tr>
<tr>
<td>Curvature</td>
<td><img src="image" alt="Curvature" /></td>
</tr>
<tr>
<td>Volume (3D size)</td>
<td><img src="image" alt="Volume (3D size)" /></td>
</tr>
<tr>
<td>Spatial region</td>
<td><img src="image" alt="Spatial region" /></td>
</tr>
<tr>
<td>Color hue</td>
<td><img src="image" alt="Color hue" /></td>
</tr>
<tr>
<td>Motion</td>
<td><img src="image" alt="Motion" /></td>
</tr>
<tr>
<td>Shape</td>
<td><img src="image" alt="Shape" /></td>
</tr>
</tbody>
</table>
Channels: Rankings

**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**

- Spatial region
- Color hue
- Motion
- Shape

Department of Computer Science and Engineering
Channels: Rankings

**Magnitude Channels: Ordered Attributes**
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**
- Spatial region
- Color hue
- Motion
- Shape

**Attribute Types**
- Categorical
- Ordered
- Ordinal
- Quantitative
Channels: Rankings

Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

6 Quality of Visualization
Channels: Rankings

Magnitude Channels: Ordered Attributes
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

Identity Channels: Categorical Attributes
- Spatial region
- Color hue
- Motion
- Shape

Expressiveness - match channel and data characteristics
Effectiveness - channels differ in accuracy of perception
- Spatial position ranks high for both
Grouping

Marks as Links

- Containment
- Connection

containment
connection

proximity
- same spatial region

similarity
- same values as other categorical channels

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape
Principles of Graphical Excellence

Graphical excellence is

the well-designed presentation of interesting data – a matter of substance, of statistics, and of design

consists of complex ideas communicated with clarity, precision and efficiency

is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space

requires telling the truth about the data.
Tufte’s Notion of Data Ink Maximization

What is the main idea?

- draw viewers attention to the substance of the graphic
- the role of redundancy
- principles of editing and redesign

What’s wrong with this? What is he really getting at?
Tufte Principle

Maximize the data-ink ratio:

\[
\text{Data-ink ratio} = \frac{\text{data ink}}{\text{total ink used in graphic}}
\]

Avoid “chart junk”
Tufte Principles

Use multifunctioning graphical elements
Use small multiples
Show mechanism, process, dynamics, and causality
High data density
  Number of items/area of graphic
  This is controversial
    White space thought to contribute to good visual design
    Tufte’s book itself has lots of white space
Tufte’s Graphical Integrity

Some lapses intentional, some not

\[
\text{Lie Factor} = \frac{\text{size of effect in graph}}{\text{size of effect in data}}
\]

Misleading uses of area
Misleading uses of perspective
Leaving out important context
Lack of taste and aesthetics
A common example of a high lie factor occurs when both dimensions of a two-dimensional figure are made proportional to the same data, so that the size of the figure is proportional to the square of the data; for instance,

<table>
<thead>
<tr>
<th>Year</th>
<th>Books circulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>100</td>
</tr>
<tr>
<td>2002</td>
<td>141</td>
</tr>
<tr>
<td>2003</td>
<td>200</td>
</tr>
</tbody>
</table>

An example of a low lie factor can be seen in the "Cones" custom chart format in Microsoft Excel.

The heights of the (truncated) cones are proportional to the data, but their areas on the screen and their apparent volumes make the larger data values seem relatively small.
How to Exaggerate with Graphs
from Tufte ’83

THE SHRINKING FAMILY DOCTOR
In California
Percentage of Doctors Devoted Solely to Family Practice

<table>
<thead>
<tr>
<th>Year</th>
<th>1964</th>
<th>1975</th>
<th>1990</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27%</td>
<td>16.0%</td>
<td>12.0%</td>
</tr>
</tbody>
</table>

1: 2,247 RATIO TO POPULATION
8.023 Doctors

Lie factor = 2.8

Los Angeles Times, August 5, 1979, p. 3
How to Exaggerate with Graphs from Tufte ’83
Howard Wainer
How to Display Data Badly (Video)

http://www.dartmouth.edu/~chance/ChanceLecture<AudioVideo.html
Rules of Thumb

Guidelines and considerations, not absolute rules

- when to use 3D? when to use 2D?
- when to use eyes instead of memory?
- when does immersion help?
- when to use overviews?
- how long is too long?
- which comes first, form or function?
Unjustified 3D all too common, in the news and elsewhere

http://viz.wtf/post/137826497077/eye-popping-3d-triangles
http://viz.wtf/post/139002022202/designer-drugs-ht-ducqn
Depth vs power of the plane

high-ranked spatial position channels: **planar** spatial position
– not depth!

**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
Depth vs power of the plane

high-ranked spatial position channels: **planar** spatial position
– not depth!

**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

Depth (3D position)
Depth vs power of the plane

high-ranked spatial position channels: **planar** spatial position
– not depth!

⚙️ **Magnitude Channels: Ordered Attributes**

<table>
<thead>
<tr>
<th>Position on common scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position on unaligned scale</td>
</tr>
<tr>
<td>Length (1D size)</td>
</tr>
<tr>
<td>Tilt/angle</td>
</tr>
<tr>
<td>Area (2D size)</td>
</tr>
<tr>
<td>Depth (3D position)</td>
</tr>
</tbody>
</table>
Depth vs power of the plane

high-ranked spatial position channels: **planar** spatial position
– not depth!

Steven’s Psychophysical Power Law: $S = I^N$

**Magnitude Channels: Ordered Attributes**
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
No unjustified 3D: Danger of depth

we don’t really live in 3D: we see in 2.05D

– acquire more info on image plane quickly from eye movements
– acquire more info for depth slower, from head/body motion

We can only see the outside shell of the world
Occlusion hides information

occlusion
interaction can resolve, but at cost of time and cognitive load
Perspective distortion loses information

- interferes with all size channel encodings
- power of the plane is lost!

[Visualizing the Results of Multimedia Web Search Engines. Mukherjea, Hirata, and Hara. InfoVis 96]
3D vs 2D bar charts

3D bars:
very difficult to justify!
– perspective distortion
– occlusion
faceting into 2D almost always better choice

[http://perceptualedge.com/files/GraphDesignIQ.html](http://perceptualedge.com/files/GraphDesignIQ.html)
Tilted text isn’t legible

text legibility
– far worse when tilted from image plane

further reading


No unjustified 3D example: Time-series data
extruded curves: detailed comparisons impossible
No unjustified 3D example: Transform for new data abstraction

derived data: cluster hierarchy

juxtapose multiple views: calendar, superimposed 2D curves
**Justified 3D: shape perception**

benefits outweigh costs when task is shape perception for 3D spatial data

- interactive navigation supports synthesis across many viewpoints

---

TARGETS

Spatial Data

Shape

---

Justified 3D: Economic growth curve

constrained navigation steps through carefully designed viewpoints

No unjustified 3D

3D legitimate for true 3D spatial data

3D needs very careful justification for abstract data
- enthusiasm in 1990s, but now skepticism
- be especially careful with 3D for point clouds or networks
No unjustified 2D

consider whether network data requires 2D spatial layout
- especially if reading text is central to task!
- arranging as network means lower information density and harder label lookup compared to text lists

benefits outweigh costs when topological structure/context important for task
- be especially careful for search results, document collections, ontologies
Eyes beat memory

principle: external cognition vs. internal memory
- easy to compare by moving eyes between side-by-side views
- harder to compare visible item to memory of what you saw

implications for animation
- great for choreographed storytelling
- great for transitions between two states
- poor for many states with changes everywhere
  consider small multiples instead

literal                       abstract

animation
show time with time

small multiples
show time with space
Resolution beats immersion

immersion typically not helpful for abstract data

– do not need sense of presence or stereoscopic 3D

– desktop also better for workflow integration

data difficult to justify

second wave: AR/MR
Overview first, zoom and filter, details on demand

influential mantra from Shneiderman


overview = summary

– microcosm of full vis design problem

_query_ ⇒ Identify ⇒ Compare ⇒ Summarise
Rule of thumb: **Responsiveness is required**

**visual feedback: three rough categories**

- **0.1 seconds: perceptual processing**
  - subsecond response for mouseover highlighting - ballistic motion
- **1 second: immediate response**
  - fast response after mouseclick, button press - Fitts’ Law limits on motor control
- **10 seconds: brief tasks**
  - bounded response after dialog box - mental model of heavyweight operation (file load)

**scalability considerations**

- highlight selection without complete redraw of view (graphics frontbuffer)
- show hourglass for multi-second operations (check for cancel/undo)
- show progress bar for long operations (process in background thread)
- rendering speed when item count is large (guaranteed frame rate)
Function first, form next

dangerous to start with aesthetics
  – usually impossible to add function retroactively

start with focus on functionality
  – possible to improve aesthetics later on, as refinement
  – if no expertise in-house, find good graphic designer to work with
  – aesthetics do matter! another level of function
    visual hierarchy, alignment, flow
Gestalt principles in action
Form: Basic graphic design ideas

proximity
  - do group related items together
  - avoid equal whitespace between unrelated

alignment
  - do find/make strong line, stick to it
  - avoid automatic centering

repetition
  - do unify by pushing existing consistencies

contrast
  - if not identical, then very different
  - avoid not quite the same

  - fast read, very practical to work through whole thing
Best practices: Labelling

make visualizations as self-documenting as possible
  – meaningful & useful title, labels, legends
    axes and panes/subwindows should have labels
      – and axes should have good min/max boundary tick marks
    everything that’s plotted should have a legend
      – and own header/labels if not redundant with main title
  use reasonable numerical format
    – avoid scientific notation in most cases

https://xkcd.com/833/
Rules of Thumb Summary

No unjustified 3D
- Power of the plane
- Disparity of depth
- Occlusion hides information
- Perspective distortion dangers
- Tilted text isn’t legible

No unjustified 2D
Eyes beat memory
Resolution over immersion
Overview first, zoom and filter, details on demand
Responsiveness is required
Function first, form next
Arranging Tables
Focus on Tables

Dataset Types

- **Tables**
  - Items (rows)
  - Attributes (columns)
  - Cell containing value

- **Multidimensional Table**
  - Key 1
  - Key 2
  - Attributes
  - Value in cell

- **Networks**
  - Link
  - Node (item)

- **Trees**

- **Fields (Continuous)**
  - Grid of positions
  - Cell
  - Attributes (columns)
  - Value in cell

- **Geometry (Spatial)**
  - Position

Department of Computer Science and Engineering
Keys and values

key
- independent attribute
- used as unique index to look up items
- simple tables: 1 key
- multidimensional tables: multiple keys

value
- dependent attribute, value of cell
Keys and values

key
  – independent attribute
  – used as unique index to look up items
  – simple tables: 1 key
  – multidimensional tables: multiple keys

value
  – dependent attribute, value of cell

classify arrangements by keys used
  – 0, 1, 2, ...

→ Tables
  Attributes (columns)
  Items (rows)
  Cell containing value

→ Multidimensional Table
  Key 1
  Key 2
  Attributes
  Value in cell

→ Keys
→ Express Values
→ 1 Key List
→ 2 Keys Matrix
Idiom: **scatterplot**

*express* values (magnitudes)

- quantitative attributes

no keys, only values

Idiom: **scatterplot**

**express** values (magnitudes)
- quantitative attributes
- no keys, only values
  - data
    - 2 quant attribs
  - mark: points
  - channels
    - horiz + vert position

**Idiom: scatterplot**

express values (magnitudes)
- quantitative attributes

no keys, only values
- data
  2 quant attribs
- mark: points
- channels
  horiz + vert position
- tasks
  find trends, outliers, distr
- scalability
  hundreds of items

Scatterplots: Encoding more channels

additional channels viable since using point marks

– color
– size (1 quant attribute, used to control 2D area)
  note radius would mislead, take square root since area grows quadratically
– shape

https://www.d3-graph-gallery.com/graph/bubble_basic.html

Scatterplot tasks

correlation

https://www.mathsisfun.com/data/scatter-pv-clots.html
Scatterplot tasks

correlation

clusters/groups, and clusters vs classes

https://www.cs.ubc.ca/labs/imager/ir/2014/DRVisTasks/
Some keys

- Keys
- Express Values

- 1 Key
- List

- 2 Keys
- Matrix
Some keys: Categorical regions

- Separate
- Order
- Align
Regions: Separate, order, align

- Separate
- Order
- Align

Regions: contiguous bounded areas distinct from each other
  - separate into spatial regions: one mark per region (for now)

Use categorical or ordered attribute to separate into regions
  - no conflict with expressiveness principle for categorical attributes

Use ordered attribute to order and align regions

- 1 Key List
- 2 Keys Matrix
Separated and aligned and ordered

best case
Separated and aligned but not ordered

limitation: hard to know rank. what's 4th? what's 7th?
Separated but not aligned or ordered

limitation: hard to make comparisons with size (vs aligned position)
Idiom: **bar chart**

one key, one value

- **data**
  - 1 categ attrib, 1 quant attrib
- **mark**: lines
- **channels**
  length to express quant value
  spatial regions: one per mark
  - separated horizontally, aligned vertically
  - ordered by quant attrib
    » by label (alphabetical), by length attrib (data-driven)
- **task**
  compare, lookup values
- **scalability**
  dozens to hundreds of levels for key attrib [bars], hundreds for values
Idiom: **stacked bar chart**

one more key

- data
  - 2 categ attrib, 1 quant attrib
- mark: vertical stack of line marks
  - **glyph**: composite object, internal structure from multiple marks
- channels
  - length and color hue
  - spatial regions: one per glyph
    - aligned: full glyph, lowest bar component
    - unaligned: other bar components
- task
  - part-to-whole relationship
- scalability: asymmetric
  - for **stacked** key attrib, 10-12 levels [segments]
  - for **main** key attrib, dozens to hundreds of levels [bars]

[Image of stacked bar chart]

https://www.d3-graph-
gallery.com/graph/barplot_stacked_basicWide.html
Idiom: **streamgraph**

generalized stacked graph

- emphasizing horizontal continuity
  vs vertical items

- data
  1 categ key attrib (movies)
  1 ordered key attrib (time)
  1 quant value attrib (counts)

- derived data
  geometry: layers, where height
  encodes counts
  1 quant attrib (layer ordering)

**Idiom: streamgraph**

generalized stacked graph

- emphasizing horizontal continuity vs vertical items
- data
  - 1 categ key attrib (movies)
  - 1 ordered key attrib (time)
  - 1 quant value attrib (counts)
- derived data
  - geometry: layers, where height encodes counts
  - 1 quant attrib (layer ordering)
- scalability
  - hundreds of time keys
  - dozens to hundreds of movies keys
    - more than stacked bars: most layers don’t extend across whole chart


Idiom: **dot / line chart**

one key, one value

- data
  - 2 quant attribs
- mark: points
  - AND line connection marks between them
- channels
  - aligned lengths to express quant value
  - separated and ordered by key attrib into horizontal regions
- task
**Idiom: dot / line chart**

- **one key, one value**
  - data
    - 2 quant attrs
  - mark: points
    - AND line connection marks between them
  - channels
    - aligned lengths to express quant value
    - separated and ordered by key attrib into horizontal regions
  - task
    - find trend
      - connection marks emphasize ordering of items along key axis by explicitly showing relationship between one item and the next
  - scalability
    - hundreds of key levels, hundreds of value levels
Choosing bar vs line charts

depends on type of key attrib
- bar charts if categorical
- line charts if ordered

do not use line charts for categorical key attribs
- violates expressiveness principle
  implication of trend so strong that it overrides semantics!
  - “The more male a person is, the taller he/she is”
Chart axes: label them!

best practice to label
– few exceptions: individual small multiple views could share axis label

https://xkcd.com/833/
Chart axes: avoid cropping y axis

include 0 at bottom left or slope misleads

[Truncating the Y-Axis: Threat or Menace? Correll, Bertini, & Franconeri, CHI 2020.]
Chart axes: avoid cropping y axis

include 0 at bottom left or slope misleads
–some exceptions (arbitrary 0, small change matters)

[Truncating the Y-Axis: Threat or Menace? Correll, Bertini, & Franconeri, CHI 2020.]
Idiom: Indexed line charts

data: 2 quant attribs
  – 1 key + 1 value
derived data: new quant value attrib
  – index
  – plot instead of original value
task: show change over time
  – principle: normalized, not absolute
scalability
  – same as standard line chart

https://public.tableau.com/profile/ben.jones#!/vizhome/CAStateRevenues/Revenues
Idiom: Gantt charts

one key, two (related) values

– data
  1 categ attrib, 2 quant attribs

– mark: line
  length: duration

– channels
  horiz position: start time
  (+end from duration)

– task
  emphasize temporal overlaps & start/end dependencies between items

– scalability
  dozens of key levels [bars]
  hundreds of value levels [durations]
Idiom: **Slopegraphs**

- **two values**
  - data
    - 2 quant value attribs
      (1 derived attrib: change magnitude)
  - mark: point + line
    - line connecting mark between pts
- **channels**
  - 2 vertical pos: express attrib value
    (linewidth/size, color)
- **task**
  - emphasize changes in rank/value
- **scalability**
  - hundreds of value levels
  - dozens of items

[Visit the Tableau Slopegraphs visualization](https://public.tableau.com/profile/ben.jones#!/vizhome/Slopegraphs/Slopegraphs)
2 Keys

Express Values

1 Key
List

2 Keys
Matrix
Idiom: heatmap

two keys, one value

– data
  2 categ attrs (gene, experimental condition)
  1 quant attrb (expression levels)

– marks: point
  separate and align in 2D matrix
    – indexed by 2 categorical attributes

– channels
  color by quant attrb
    – (ordered diverging colormap)

– task
  find clusters, outliers

– scalability
  1M items, 100s of categ levels, ~10 quant attrb levels
Heatmap reordering

https://blogs.sas.com/content/iml/2018/05/02/reorder-variables-correlation-heat-map.html
Idiom: cluster heatmap

in addition

- derived data
  2 cluster hierarchies

- dendrogram
  parent-child relationships in tree with connection line marks
  leaves aligned so interior branch heights easy to compare

- heatmap
  marks (re-)ordered by cluster hierarchy traversal
  task: assess quality of clusters found by automatic methods
Tables
Axis Orientation

- Rectilinear
- Parallel
- Radial
Idioms: radial bar chart, star plot

star plot
  - line mark, radial axes meet at central point

radial bar chart
  - line mark, radial axes meet at central ring
  - channels: length, angle/orientation

bar chart
  - rectilinear axes, aligned vertically

accuracy
  - length not aligned with radial layouts
    less accurately perceived than rectilinear aligned
Idiom: **radar plot**

radial line chart
- point marks, radial layout
- connecting line marks

avoid unless data is cyclic
“Radar graphs: Avoid them (99.9% of the time)”

Idioms: **pie chart, coxcomb chart**

**pie chart**
- **interlocking area** marks with angle channel: **2D area varies**
  - separated & ordered radially, uniform height
- accuracy: area less accurate than rectilinear aligned line length
- task: part-to-whole judgements

**coxcomb chart**
- line marks with length channel: **1D length varies**
  - separated & ordered radially, uniform width
- direct analog to radial bar charts

**data**
- 1 categ key attrib, 1 quant value attrib
Coxcomb / nightingale rose / polar area chart

invented by Florence Nightingale:
Diagram of the Causes of Mortality in the Army in the East
Coxcomb: perception

encode: 1D length
decode/perceive: 2D area

nonuniform line/sector width as length increases
– so area variation is nonlinear wrt line mark length!

bar chart safer: uniform width, so area is linear with line mark length
– both radial & rectilinear cases
Pie charts: perception

some empirical evidence that people respond to arc length
– decode/perceive: not angles
– maybe also areas?…
donut charts no worse than pie charts


Pie charts: best practices

not so bad for two (or few) levels, for part-to-whole task

https://eagereyes.org/pie-charts
Pie charts: best practices

not so bad for two (or few) levels, for part-to-whole task
dubious for several levels if details matter

https://eagereyes.org/pie-charts
Pie charts: best practices

not so bad for two (or few) levels, for part-to-whole task
dubious for several levels if details matter
terrible for many levels

https://eagereyes.org/pie-charts
Idioms: **normalized stacked bar chart**

- **task**
  - part-to-whole judgements

- **normalized stacked bar chart**
  - stacked bar chart, normalized to full vert height
  - single stacked bar equivalent to full pie
    - high information density: requires narrow rectangle

- **pie chart**
  - information density: requires large circle

http://bl.ocks.org/mbostock/3886208
http://bl.ocks.org/mbostock/3887235
http://bl.ocks.org/mbostock/3886394
Idiom: **glyphmaps**

rectilinear good for linear vs nonlinear trends

radial good for cyclic patterns – evaluating periodicity
Axis Orientation

- Rectilinear
- Parallel
- Radial
**Idiom: SPLOM**

- scatterplot matrix (SPLOM)
  - rectilinear axes, point mark
  - all possible pairs of axes
  - scalability
    - one dozen attrs
    - dozens to hundreds of items

---

Wilkinson et al., 2005
Idioms: parallel coordinates

scatterplot limitation
– visual representation with orthogonal axes
– can show only two attributes with spatial position channel

Scatterplot Matrix

Table

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>Physics</th>
<th>Dance</th>
<th>Drama</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>95</td>
<td>70</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>80</td>
<td>60</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>90</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>95</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>80</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

http://www.michaelmcguffin.com/courses/vis/
**Idioms: parallel coordinates**

**scatterplot limitation**
- visual representation with orthogonal axes
- can show only two attributes with spatial position channel

**alternative: line up axes in parallel to show many attributes with position**
- item encoded with a line with $n$ segments
- $n$ is the number of attributes shown

**parallel coordinates**
- parallel axes, jagged line for item
- rectilinear axes, item as point
  - axis ordering is major challenge
- scalability
  - dozens of attrs
  - hundreds of items

---

**Table**

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>Physics</th>
<th>Dance</th>
<th>Drama</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>95</td>
<td>70</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>80</td>
<td>60</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>90</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>95</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>80</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>
Task: Correlation

scatterplot matrix
- positive correlation
  diagonal low-to-high
- negative correlation
  diagonal high-to-low
- uncorrelated: spread out

parallel coordinates
- positive correlation
  parallel line segments
- negative correlation
  all segments cross at halfway point
- uncorrelated
  scattered crossings

Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $p = 1, .8, .2, 0, -.2, -.8, \text{ and } -1$. 

https://www.mathsisfun.com/data/scatter-xy-plots.html
Parallel coordinates, limitations
visible patterns only between neighboring axis pairs
how to pick axis order?
– usual solution: reorderable axes, interactive exploration
– same weakness as many other techniques
downside of interaction: human-powered search
– some algorithms proposed, none fully solve
Orientation limitations

rectilinear: scalability wrt #axes

- 2 axes best
- 3 problematic
- 4+ impossible

Axis Orientation
- Rectilinear
**Orientation limitations**

rectilinear: scalability wrt #axes
- 2 axes best, 3 problematic, 4+ impossible

parallel: unfamiliarity, training time
Orientation limitations

rectilinear: scalability wrt #axes
   2 axes best, 3 problematic, 4+ impossible

parallel: unfamiliarity, training time

radial: perceptual limits
   – polar coordinate asymmetry
     angles lower precision than length
     nonuniform sector width/size depending on radial distance
   – frequently problematic
     but sometimes can be deliberately exploited!
     – for 2 attribs of very unequal importance

Layout density

/Layout Density

- Dense
- Space-Filling
Idiom: Dense software overviews

- data: text
  - text + 1 quant attrib per line

- derived data:
  - one pixel high line
  - length according to original

- color line by attrib

- scalability

  ➔ Layout Density

  ➔ Dense

Arrange tables

- Express Values
- Separate, Order, Align Regions
  - Separate
  - Order
  - Align
- Axis Orientation
  - Rectilinear
  - Parallel
  - Radial
- Layout Density
  - Dense

1 Key
List

2 Keys
Matrix
How?

<table>
<thead>
<tr>
<th>Encode</th>
<th>Manipulate</th>
<th>Facet</th>
<th>Reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map from <strong>categorical</strong> and <strong>ordered</strong> attributes</td>
<td><strong>Change</strong></td>
<td><strong>Juxtapose</strong></td>
<td><strong>Filter</strong></td>
</tr>
<tr>
<td>▶ Color</td>
<td>▶ Hue</td>
<td>▶ Select</td>
<td>▶ Partition</td>
</tr>
<tr>
<td>▶ Size, Angle, Curvature, ...</td>
<td>▶ Saturation</td>
<td>▶ Navigate</td>
<td>▶ Superimpose</td>
</tr>
<tr>
<td>▶ Shape</td>
<td>▶ Luminance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ ● ■ △</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶ Motion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direction, Rate, Frequency, ...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**What?**

**Why?**

**How?**
6 Quality of Visualization

How?

Encode

- Arrange
  - Express
  - Separate
- Order
  - Align
- Use

Manipulate

- Map from categorical and ordered attributes
  - Color
    - Hue
    - Saturation
    - Luminance
  - Size, Angle, Curvature, ...
  - Shape
    - + ○ ■ △
  - Motion
    - Direction, Rate, Frequency, ...

Facet

- Change

Reduce

- Juxtapose

- Select

- Partition

- Aggregate

- Filter

- Embed
Chart axes

labelled axis is critical

avoid cropping y-axis

– include 0 at bottom left

– or slope misleads

http://www.thefunctionalart.com/2015/10/if-you-see-bullshit-say-bullshit.html
Idiom: dual-axis line charts

controversial

– acceptable if commensurate
– beware, very easy to mislead!

Idiom: **connected scatterplots**

scatterplot with line connection marks
- popular in journalism
- horiz + vert axes: value attril
- line connection marks: temporal order
- alternative to dual-axis charts
  horiz: time
  vert: two value attrilbs

empirical study
- engaging, but correlation unclear

http://steveharoz.com/research/connected_scatterplot/
Choosing line chart aspect ratios

1: banking to 45 (1980s)
   - Cleveland perceptual argument: most accurate angle judgement at 45

Fig 7.1 Sunspot Data: Aspect Ratio 1

Fig 7.2 Annual Report: Aspect Ratio 2
Choosing line chart aspect ratios

2: multi scale banking to 45 (2006)
   – frequency domain analysis to find ratios
     FFT the data, convolve with Gaussian to smooth
   – find interesting spikes/ranges in power spectrum
     cull nearby regions if similar, ensure overview
   – create trend curves (red) for each aspect ratio

[Multi-Scale Banking to 45 Degrees.
Heer and Agrawala, Proc InfoVis 2006]
Choosing line chart aspect ratios

3: arc length based aspect ratio (2011)
   – minimize the arc length of curve while keeping the area of the plot constant
   – parametrization and scale invariant
   – symmetry preserving
   – robust & fast to compute

meta-points from this progression
   – young field; prescriptive advice changes rapidly
   – reasonable defaults required deep dive into perception meets math

[Arc Length-Based Aspect Ratio Selection. Talbot, Gerth, and Hanrahan. Proc InfoVis 2011]
Breaking conventions

presentation vs exploration

- engaging/evocative
- inverted y axis

blood drips down on Poe

https://public.tableau.com/profile/ben.jones#/

https://en.wikipedia.org/wiki/Edgar_Alan_Poe_bibliography

Ben Jones, 7 October 2015
Network Data
Network data

networks
- model relationships between things
  aka graphs
- two kinds of items, both can have attributes
  nodes
  links

tree
- special case
- no cycles
  one parent per node
Network tasks: topology-based and attribute-based

topology based tasks
- find paths
- find (topological) neighbors
- compare centrality/importance measures
- identify clusters / communities

attribute based tasks (similar to table data)
- find distributions, ...

combination tasks, incorporating both
- example: find friends-of-friends who like cats
topology: find all adjacent nodes of given node
attributes: check if has-pet (node attribute) == cat
Node-link diagrams

nodes: point marks
links: line marks
  – straight lines or arcs
  – connections between nodes
intuitive & familiar
  – most common
  – many, many variants
Criteria for good node-link layouts

minimize

- edge crossings, node overlaps
- distances between topological neighbor nodes
- total drawing area
- edge bends

maximize

- angular distance between different edges
- aspect ratio disparities

emphasize symmetry

- similar graph structures should look similar in layout
Criteria conflict

most criteria NP-hard individually
many criteria directly conflict with each other

Minimum number of edge crossings vs. Uniform edge length
Space utilization vs. Symmetry

Schulz 2004
Optimization-based layouts

formulate layout problem as optimization problem

convert criteria into weighted cost function

\[ F(\text{layout}) = a \times [\text{crossing counts}] + b \times [\text{drawing space used}] + \ldots \]

use known optimization techniques to find layout at minimal cost

- energy-based physics models
- force-directed placement
- spring embedders
Force-directed placement

physics model
- links = springs pull together
- nodes = magnets repulse apart

algorithm
- place vertices in random locations
- while not equilibrium
  calculate force on vertex
  - sum of
    » pairwise repulsion of all nodes
    » attraction between connected nodes
  move vertex by c * vertex_force

Force-directed placement properties

strengths
- reasonable layout for small, sparse graphs
- clusters typically visible
- edge length uniformity

weaknesses
- nondeterministic
- computationally expensive: $O(n^3)$ for $n$ nodes
  each step is $n^2$, takes $\sim n$ cycles to reach equilibrium
- naive FD doesn't scale well beyond 1K nodes
- iterative progress: engaging but distracting

https://bl.ocks.org/steveharoz/8c3e2524079a8c440df60c12b72b5d03
Idiom: **force-directed placement**

**visual encoding**
- link connection marks, node point marks

**considerations**
- spatial position: no meaning directly encoded
  - left free to minimize crossings
- proximity semantics?
  - sometimes meaningful
  - sometimes arbitrary, artifact of layout algorithm
- tension with length
  - long edges more visually salient than short

**tasks**
- explore topology; locate paths, clusters

**scalability**
- node/edge density $E < 4N$

---

Department of Computer Science and Engineering
Idiom: **circular layouts / arc diagrams (node-link)**

restricted node-link layouts: lay out nodes around circle or along line

- original: network
- derived: node ordering attribute (global computation)

considerations: node ordering crucial to avoid excessive clutter from edge crossings
- examples: before & after barycentric ordering

[Image of circular layouts and arc diagrams]

[Links]
Adjacency matrix representations

derive adjacency matrix from network

Adjacency Matrix
Derived Table

Department of Computer Science and Engineering
Adjacency matrix examples
Node order is crucial: Reordering

https://bost.ocks.org/mike/miserables/
Adjacency matrix

An adjacency matrix is a way to represent a graph. Each row and column corresponds to a node in the graph, and the entries in the matrix indicate whether there is an edge between the corresponding nodes. In this example, nodes A through H are shown in the top row and column headings. The matrix entries are shaded, with white indicating no edge and gray indicating an edge. The graph on the right shows the relationships suggested by the matrix:

- Node A is connected to nodes B, C, and D.
- Node B is connected to nodes C, D, and E.
- Node C is connected to nodes A and D.
- Node D is connected to nodes A, C, and E.
- Node E is connected to nodes A and D.
- Node F is not connected to any other nodes.
- Node G is not connected to any other nodes.
- Node H is not connected to any other nodes.

This matrix is good for topology tasks related to neighborhoods (e.g., node 1-hop neighbors) but is bad for topology tasks related to paths.
Structures visible in both

http://www.michaelmcguffin.com/courses/vis/patternsInAdjacencyMatrix.png
Idiom: **adjacency matrix view**

**data: network**
- transform into same data/encoding as heatmap

**derived data: table from network**
- 1 quant attrib
  - weighted edge between nodes
- 2 categ attribs: node list x 2

**visual encoding**
- cell shows presence/absence of edge

**scalability**
- 1K nodes, 1M edges
Node-link vs. matrix comparison

node-link diagram strengths
- topology understanding, path tracing
- intuitive, flexible, no training needed

adjacency matrix strengths
- focus on edges rather than nodes
- layout straightforward (reordering needed)
- predictability, scalability
- some topology tasks trainable

empirical study
- node-link best for small networks
- matrix best for large networks
  if tasks don’t involve path tracing!

Idiom: **NodeTrix**

hybrid nodelink/matrix
capture strengths of both

Trees
Node-link trees

Reingold-Tilford

- tidy drawings of trees
  - exploit parent/child structure
- allocate space: compact but without overlap
  - rectilinear and radial variants

- nice algorithm writeup

http://bl.ocks.org/mbostock/4339184
http://bl.ocks.org/mbostock/4063550
http://billmill.org/pymag/trees/

Idiom: **radial node-link tree**

**Data**
- tree

**Encoding**
- link connection marks
- point node marks
- radial axis orientation
  - angular proximity: siblings
  - distance from center: depth in tree

**Tasks**
- understanding topology, following paths

**Scalability**
- 1K - 10K nodes (with/without labels)

Link marks: Connection and containment

marks as links (vs. nodes)

– common case in network drawing
– 1D case: connection
  ex: all node-link diagrams
  emphasizes topology, path tracing
  networks and trees
– 2D case: containment
  ex: all treemap variants
  emphasizes attribute values at leaves
  (size coding)
  only trees

Idiom: \textbf{treemap}

data
  \begin{itemize}
  \item tree
  \item 1 quant attrib at leaf nodes
  \end{itemize}

encoding
  \begin{itemize}
  \item area containment marks for hierarchical structure
  \item rectilinear orientation
  \item size encodes quant attrib
  \end{itemize}

tasks
  \begin{itemize}
  \item query attribute at leaf nodes
  \item ex: disk space usage within filesystem
  \end{itemize}

scalability
  \begin{itemize}
  \item 1M leaf nodes
  \end{itemize}

\textbf{Enclosure}

\begin{itemize}
\item Containment Marks
\end{itemize}
Idiom: implicit tree layouts (sunburst, icicle plot)

alternative to connection and containment: position

– show parent-child relationships only through relative positions

Treemap  Sunburst  Icicle Plot
Idiom: implicit tree layouts (sunburst, icicle plot)

alternative to connection and containment: position
– show parent-child relationships only through relative positions

Treemap    Sunburst    Icicle Plot
Idiom: implicit tree layouts (sunburst, icicle plot)

alternative to connection and containment: position
– show parent-child relationships only through relative positions

Treemap  Sunburst  Icicle Plot

Spatial Position

[Images of visualization examples]
Tree drawing idioms comparison

Comparison: tree drawing idioms

- link relationships
- tree depth
- sibling order

Comparison: tree drawing idioms

data shown
– link relationships
– tree depth
– sibling order

design choices
– connection vs containment link marks
– rectilinear vs radial layout
– spatial position channels
Comparison: tree drawing idioms

data shown
  – link relationships
  – tree depth
  – sibling order

design choices
  – connection vs containment link marks
  – rectilinear vs radial layout
  – spatial position channels

considerations
  – redundant? arbitrary?
  – information density?
    avoid wasting space
    consider where to fit labels!

treevis.net: Many, many options!

https://treevis.net/
Arrange networks and trees

Node–Link Diagrams
Connection Marks

Spatial Position

Adjacency Matrix
Derived Table

Enclosure
Containment Marks
Network Data with Trees
Multilevel networks

derive cluster hierarchy of metanodes on top of original graph nodes

---

[Schulz 2004]
Idiom: **GrouseFlocks**

data: compound network
- network
- cluster hierarchy atop it
derived or interactively chosen

visual encoding
- connection marks for network links
- containment marks for hierarchy
- point marks for nodes

dynamic interaction
- select individual metanodes in hierarchy to expand/contract

*GrouseFlocks: Steerable Exploration of Graph Hierarchy Space*, Archambault, Münzner, Auber.
Idiom: **sfdp** (multi-level force-directed placement)

data: compound graph

– original: network
– derived: cluster hierarchy atop it

considerations

– better algorithm for same encoding technique
  
  same: fundamental use of space
  hierarchy used for algorithm speed/quality
  but not shown explicitly

scalability

– nodes, edges: 1K-10K
– hairball problem eventually hits

**Idiom:** hierarchical edge bundling

**Data**

- any layout of compound network
  
  network: software classes (nodes), import/export between classes (links)
  
  cluster hierarchy: class package structure

- derived: bundles of edges with same source/destination (multi-level)

**Idiom:** curve edge routes according to bundles

**Task:** edge clutter reduction
Hierarchical edge bundling

works for any layout: treemap vs radial
Spatial Data
Focus on Spatial

Dataset Types

- Tables
  - Attributes (columns)
  - Items (rows)
  - Cell containing value

- Multidimensional Table

- Networks
  - Node (item)
  - Link

- Trees

- Fields (Continuous)
  - Attributes (columns)
  - Value in cell

- Geometry (Spatial)
  - Grid of positions
  - Cell
  - Position

Department of Computer Science and Engineering
How?

Encode

- **Arrange**
  - Express
  - Separate

- **Order**
  - Align

- **Use**

Map from **categorical** and **ordered** attributes

- **Color**
  - Hue
  - Saturation
  - Luminance

- **Size, Angle, Curvature, ...**

- **Shape**
  - ● □ △

- **Motion**
  - Direction, Rate, Frequency, ...

Manipulate

- **Change**

Facet

- **Juxtapose**

Reduce

- **Filter**

- **Aggregate**

- **Embed**

Department of Computer Science and Engineering
6 Quality of Visualization

How?

Encode

- **Arrange**
  - Express
  - Separate
- **Order**
  - Align
- **Use**

Map from *categorical* and *ordered* attributes

- **Color**
  - Hue
  - Saturation
  - Luminance
- **Size, Angle, Curvature, ...**
- **Shape**
  - + ● ■ ▲
- **Motion**
  - Direction, Rate, Frequency, ...

Manipulate

- **Change**
- **Select**
- **Navigate**

Facet

- **Juxtapose**
- **Partition**
- **Superimpose**

Reduce

- **Filter**
- **Aggregate**
- **Embed**

What?

Why?

How?
Spatial data

use given spatial position
when?

– dataset contains spatial attributes and they have primary importance
– central tasks revolve around understanding spatial relationships

examples

– geographical/cartographic data
– sensor/simulation data
Geographic Maps
Geographic Map

Interlocking marks
shape coded
area coded
position coded

- cannot encode another attribute with these channels, they're "taken"
Thematic maps

show spatial variability of attribute ("theme")
- combine geographic / reference map with (simple, flat) tabular data
- join together
  - region: interlocking area marks (provinces, countries with outline shapes)
    - also could have point marks (cities, locations with 2D lat/lon coords)
  - region: categorical key attribute in table
    - use to look up value attributes

major idioms
- choropleth
- symbol maps
- cartograms
- dot density maps
Idiom: **choropleth map**

use given spatial data

- when central task is understanding spatial relationships

data

- geographic geometry
- table with 1 quant attribute per region

encoding

- position:
  use given geometry for area mark boundaries
- color:
  sequential segmented colormap

[http://bl.ocks.org/mbostock/4060606](http://bl.ocks.org/mbostock/4060606)
Beware: Population maps trickiness!

[https://xkcd.com/1138]
Beware: Population maps trickiness!

spurious correlations: most attributes just show where people live

[https://xkcd.com/1138]
Beware: Population maps trickiness!

spurious correlations: most attributes just show where people live
consider when to normalize by population density
  encode raw data values
    – tied to underlying population
but should use normalized values
    – unemployed people per 100 citizens, mean family income

[https://xkcd.com/1138]
Beware: Population maps trickiness!

Spurious correlations: most attributes just show where people live.

Consider when to normalize by population density:
- Encode raw data values
  - Tied to underlying population
- But should use normalized values
  - Unemployed people per 100 citizens, mean family income

General issue:
- Absolute counts vs relative/normalized data
- Failure to normalize is common error
Choropleth maps: Recommendations

only use when central task is understanding spatial relationships
show only one variable at a time
normalize when appropriate
be careful when choosing colors & bins
best case: regions are roughly equal sized
Choropleth map: Pros & cons

pros
- easy to read and understand
- well established visualization (no learning curve)
- data is often collected and aggregated by geographical regions

cons
- most effective visual variable used for geographic location
- visual salience depends on region size, not true importance wrt attribute value
  - large regions appear more important than small ones
- color palette choice has a huge influence on the result
Idiom: **Symbol maps**

Symbol is used to represent aggregated data (mark or glyph)
- allows use of size and shape and color channels
  aka proportional symbol maps, graduated symbol maps
keep original spatial geometry in the background
often a good alternative to choropleth maps
Symbol maps with glyphs
Symbol map: Pros & cons

pros

– somewhat intuitive to read and understand
– mitigate problems with region size vs data salience
  marks: symbol size follows attribute value
  glyphs: symbol size can be uniform

cons

– possible occlusion / overlap
  symbols could overlap each other
  symbols could occlude region boundaries
– complex glyphs may require explanation / training
Idiom: **Contiguous cartogram**

interlocking marks:
- shape, area, and position coded

derive new interlocking marks
- based on combination of original interlocking marks and new quantitative attribute

algorithm to create new marks
- input: target size
- goal: shape as close to the original as possible
- requirement: maintain constraints

relative position
- contiguous boundaries with their neighbors
Idiom: **Grid Cartogram**

Uniform-sized shapes arranged in rectilinear grid maintain approximate spatial position and arrangement.
Cartogram: Pros & cons

pros
- can be intriguing and engaging
- best case: strong and surprising size disparities
- non-contiguous cartograms often easier to understand

cons
- require substantial familiarity with original dataset & use of memory
  compare distorted marks to memory of original marks
  mitigation strategies: transitions or side by side views
- major distortion is problematic
  may be aesthetically displeasing
  may result in unrecognizable marks
- difficult to extract exact quantities
Idiom: **Dot density maps**

visualize distribution of a phenomenon by placing dots

one symbol represents a constant number of items

- dots have uniform size & shape
- allows use of color channel

task:
show spatial patterns, clusters
Dot density maps: Pros and cons

pros
- straightforward to understand
- avoids choropleth non-uniform region size problems

cons
- challenge: normalization, just like choropleths
  show population density (correlated with attribute), not effect of interest
- perceptual disadvantage:
  difficult to extract quantities
- performance disadvantage:
  rendering many dots can be slow
Map Projections

mathematical functions that map 3D surface geometry of the Earth to 2D maps
all projections of sphere on plane necessarily distort surface in some way
interactive: philogb.github.io/page/myriahedral/ and jasondavies.com/maps/
Mercator Projection

» Heavily distorts country sizes; particularly close to the poles.
Spatial Fields
Idiom: **topographic map**

**data**
- geographic geometry
- scalar spatial field
  - 1 quant attribute per grid cell

**derived data**
- isoline geometry
  - isocontours computed for specific levels of scalar values

**task**
- understanding terrain shape
  - densely lined regions = steep

**pros**
- use only 2D position, avoid 3D challenges
- color channel available for other attributes

**cons**
- significant clutter from additional lines
Idioms: \textit{isosurfaces, direct volume rendering}

data
  \begin{itemize}
    \item scalar spatial field (3D volume)
      \begin{itemize}
        \item 1 quant attribute per grid cell
      \end{itemize}
  \end{itemize}

task
  \begin{itemize}
    \item shape understanding, spatial relationships
  \end{itemize}
Idioms: **isosurfaces, direct volume rendering**

**data**
- scalar spatial field (3D volume)
  - 1 quant attribute per grid cell

**task**
- shape understanding, spatial relationships

**isosurface**
- derived data: isocontours computed for specific levels of scalar values

Idioms: isosurfaces, direct volume rendering

data
- scalar spatial field (3D volume)
  1 quant attribute per grid cell

task
- shape understanding, spatial relationships

isosurface
- derived data: isocontours computed for specific levels of scalar values

direct volume rendering
- transfer function maps scalar values to color, opacity
  no derived geometry

Color
Idiom design choices: Visual encoding

**Encode**

- **Arrange**
  - Express
  - Order
  - Use

- **Separate**

- **Align**

- **Map**
  - from **categorical** and **ordered** attributes
    - Color
      - Hue
      - Saturation
      - Luminance
  - Size, Angle, Curvature, ...
  - Shape
  - Motion
    - Direction, Rate, Frequency, ...

*Department of Computer Science and Engineering*
Idiom design choices: Beyond spatial arrangement

Encode

- **Arrange**
  - Express
  - Order
  - Use

- **Separate**
  - Align

- **Map**
  - from *categorical* and *ordered* attributes
  - Color
    - Hue
    - Saturation
    - Luminance
  - Size, Angle, Curvature, ...
  - Shape
    - Direction, Rate, Frequency, ...

What?

Why?

How?
Channels: What's up with color?

**Magnitude Channels:** Ordered Attributes
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels:** Categorical Attributes
- Spatial region
- Color hue
- Motion
- Shape

Effectiveness scale from Best to Least.
Decomposing color

first rule of color: do not (just) talk about color!
- color is confusing if treated as monolithic

decompose into three channels
- ordered can show magnitude
  - luminance: how bright (B/W)
  - saturation: how colourful
- categorical can show identity
  - hue: what color

channels have different properties
- what they convey directly to perceptual system
- how much they can convey
  how many discriminable bins can we use?
Color Channels in Visualization
Categorical vs ordered color
Categorical color: limited number of discriminable bins

human perception built on relative comparisons
- great if color contiguous
- surprisingly bad for absolute comparisons

noncontiguous small regions of color
- fewer bins than you want
- rule of thumb: 6-12 bins, including background and highlights

6 Quality of Visualization

Categorical color: limited number of discriminable bins

Ordered color: limited number of discriminable bins

Gregor Aisch, vis4.net/blog/posts/choropleth-maps/
Ordered color: Rainbow is poor default

problems

- perceptually unordered
- perceptually nonlinear
Ordered color: Rainbow is poor default

problems

– perceptually unordered
– perceptually nonlinear
Ordered color: Rainbow is poor default

problems
– perceptually unordered
– perceptually nonlinear

benefits
– fine-grained structure visible and nameable


http://www.research.ibm.com/people/l/lloydt/color/color.HTM]
Ordered color: Rainbow is poor default

problems
  – perceptually unordered
  – perceptually nonlinear

benefits
  – fine-grained structure visible and nameable

alternatives
  – large-scale structure: fewer hues


Ordered color: Rainbow is poor default

problems
– perceptually unordered
– perceptually nonlinear

benefits
– fine-grained structure visible and nameable

alternatives
– large-scale structure: fewer hues
– fine structure: multiple hues with monotonically increasing luminance [eg viridis]
Viridis / Magma: sequential colormaps

monotonically increasing luminance, perceptually uniform colorful, colorblind-safe

– R, python, D3

https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html
Ordered color: Rainbow is poor default

problems
- perceptually unordered
- perceptually nonlinear

benefits
- fine-grained structure visible and nameable

alternatives
- large-scale structure: fewer hues
- fine structure: multiple hues with monotonically increasing luminance [eg viridis]

legit for categorical
- segmented saturated rainbow is good!


Interaction between channels: Not fully separable

color channel interaction

- size heavily affects saliency
- small regions need high saturation
- large regions need low saturation

http://colorbrewer2.org/
Interaction between channels: Not fully separable

color channel interactions
- size heavily affects salience
- small regions need high saturation
- large regions need low saturation

saturation & luminance:
- not separable from each other!
- also not separable from transparency

http://colorbrewer2.org/
Interaction between channels: Not fully separable

color channel interactions
- size heavily affects salience
- small regions need high saturation
- large regions need low saturation

saturation & luminance:
- not separable from each other!
- also not separable from transparency
- small separated regions: 2 bins safest (use only one of these channels), 3-4 bins max
- contiguous regions: many bins (use only one of these channels)

http://colorbrewer2.org/
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Promising Techniques
Promising Techniques & Approaches

Perceptual Techniques

- Animation
- Grouping / Gestalt principles
- Using size to indicate quantity
- Color for Accent, Distinction, Selection
  NOT FOR QUANTITY!!!!

General Approaches

- Standard Techniques
  - Graphs, bar charts, tables
- Brushing and Linking
- Providing Multiple Views and Models
- Aesthetics!
Standard Techniques

It’s often hard to beat:
- Line graphs, bar charts
- Scatterplots (or Scatterplot Matrix)
- Tables

A Darwinian view of visualizations:
- Only the fittest survive
- We are in a period of great experimentation; eventually it will be clear what works and what dies out.

A bright spot:
- Enhancing the old techniques with interactivity
  - Example: Spotfire
    - Adds interactivity, color highlighting, zooming to scatterplots
  - Example: TableLens / Eureka
    - Adds interactivity and length cues to tables
Spotfire: Integrating Interaction with Scatterplots

Ahlberg & Shneiderman, Color plate 1. The FilmFinder.
Spotfire/IVEE: Integrating Interaction with Scatterplots

Ahberg & Shneiderman, Color plate 2. Categories have been selected, the displayed is zoomed into 1960-95 and popularity 4-9, and Sean Connery has been selected.
Brushing and Linking

Interactive technique
  - Highlighting
  - Brushing and Linking

At least two things must be linked together to allow for brushing
  - select a subset of points
  - see the role played by this subset of points in one or more other views

Example systems
  - Graham Will’s EDV system
  - Ahlberg & Sheiderman’s IVEE (Spotfire)
Linking types of assist behavior to position played (from Eick & Wills 95)
Baseball data: Scatterplots and histograms and bars (from Eick & Wills 95)
What was learned from interaction with this baseball data?

Seems impossible to earn a high salary in the first three years

High salaried players have a bimodal distribution (peaking around 7 & 13 yrs)

Hits/Year a better indicator of salary than HR/Year

High paid outlier with low HR and medium hits/year. Reason: person is player-coach

There seem to be two differentiated groups in the put-outs/assists category (but not correlated with salary) Why?
Animation

“The quality or condition of being alive, active, spirited, or vigorous” (dictionary.com)

“A dynamic visual statement that evolves through movement or change in the display”

“… creating the illusion of change by rapidly displaying a series of single frames” (Roncarelli 1988).
We Use Animation to...

Tell stories / scenarios: cartoons
Illustrate dynamic process / simulation
Create a character / an agent
Navigate through virtual spaces
Draw attention
Delight
Cartoon Animation Principles

Chang & Unger ‘93

Solidity (squash and stretch)
  Solid drawing
  Motion blur
  Dissolves

Exaggeration
  Anticipation
  Follow through

Reinforcement
  Slow in and slow out
  Arcs
  Follow through
Why Cartoon-Style Animation?

Cartoons’ theatricality is powerful in communicating to the user.
Cartoons can make UI engage the user into its world.
The medium of cartoon animation is like that of graphic computers.
Application using Animation: Gnutellavision

Visualization of Peer-to-Peer Network

- Hosts (with color for status and size for number of files)
- Nodes with closer network distance from focus on inner rings
- Queries shown; can trace queries

Gnutellavision as exploratory tool

- Very few hosts share many files
- Uneven propagation of queries
- Qualitative assessment of queries (simple)
Each node is placed at the center of the angular sector allocated to it defined by the width of that subtree.
Animation in Gnutellavision

Goal of animation is to help maintain context of nodes and general orientation of user during refocus

Transition Paths

Linear interpolation of polar coordinates
Node moves in arc not straight line
Moves along circle if not changing levels (like great circles on earth)
Spirals in or out to next ring
Animation (continued)

Transition constraints

Orientation of transition to minimize rotational travel
(Move former parent away from new focus in same orientation)

Avoid cross-over of edges
(to allow users to keep track of which is which)

Animation timing

Slow in Slow out timing (allows users to better track movement)
Transition Constraint - Orientation

https://www.youtube.com/watch?v=AvfIIC2xZw
Transition Constraint - Order

https://www.youtube.com/watch?v=AvfIlcE2xZw
Usability Testing

In general, users appreciated the subtleties added to the general method when the number of nodes increased.

Perhaps the most interesting result is that most people preferred rectangular movement for the small graph and polar coordinate movement for the large one.

<table>
<thead>
<tr>
<th>Overall Preference of Users</th>
<th>No Features</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Graph</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Large Graph</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>
Hyperbolic Tree


Also uses animation

Tree-based layout; leaves stretch to infinity

Only a few labels can be seen at a time

https://www.youtube.com/watch?v=ZUwpgzDNhME
Department of Computer Science and Engineering
Issues

Displaying text

The size of the text
  Works good for small things like directories
  Not so good for URLs

Only a portion of the data can be seen in the focus at one time

Only works for certain types of data - Hierarchical
Not clear if it is actually useful for anything.
Animating Algorithms

Kehoe, Stasko, and Taylor, “Rethinking Evaluation of Algorithm Animations as Learning Aids”

Why previous studies present no benefits:
- No or limited benefits from particular animations
- Benefits are not captured in measurements
- Design of experiments hides the benefits

Methods for this study:
- Combination of qualitative & quantitative
- More flexible setting
- Metrics: score for each type of questions, time used, usage of materials, qualitative data from observations & interviews
Department of Computer Science and Engineering
Findings

Value of animation is more apparent in interactive situations
Most useful to learn procedural operations
Makes subject more accessible & less intimidating \(\rightarrow\) increase motivation
What Isn’t Working?

The existing studies indicate that we don’t yet know how to make the following work well for every-day tasks:

- Pan-and-Zoom
- 3D Navigation
- Node-and-link representations of concept spaces
Zoom, Overview + Detail

An exception, possibly:

Benjamin B. Bederson: PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. UIST 2001: 71-80
Overview + Detail

Overview + Detail


A study on integrating Overview + Detail on a Map search task
  Incorporating panning & zooming as well.
  They note that panning & zooming does not do well in most studies.

Results seem to be
  Subjectively, users prefer to have a linked overview
  But they aren’t necessarily faster or more effective using it
  Well-constructed representation of the underlying data may be more important.

More research needed as each study seems to turn up different results, sensitive to underlying test set.
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Problem Solving
Problem Solving

A Detective Tool for Multidimensional Data
  Inselberg on using Parallel Coordinates

Analyzing Web Clickstream Data
  Brainerd & Becker, Waterson et al.

Information Visualization for Pattern Detection
  Carlis & Konstan on Periodic Data

Visualization vs. Analysis
  Comments by Wesley Johnson of Chevron
Multidimensional Detective


Figure 1: The full dataset consisting of 473 batches

Inselberg’s Principles for analysis using visualizations:
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!
A Detective Story


The Dataset:

Production data for 473 batches of a VLSI chip
16 process parameters
The yield: % of produced chips that are useful
  X1
The quality of the produced chips (speed)
  X2
10 types of defects (zero defects shown at top)
  X3 ... X12
4 physical parameters
  X13 ... X16

The Objective:

Raise the yield (X1) and maintain high quality (X2)
Multidimensional Detective


Do Not Let the Picture Scare You!!

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

Each line represents the values for one batch of chips.

This figure shows what happens when only those batches with both high X1 and high X2 are chosen.

Notice the separation in values at X15.

Also, some batches with few X3 defects are not in this high-yield/high-quality group.

Figure 2: The batches high in Yield, X1, and Quality, X2.
Multidimensional Detective

Now look for batches which have *nearly* zero defects.
   For 9 out of 10 defect categories
Most of these have low yields
Surprising because we know from first diagram that some defects are ok.
Go back to first diagram, looking at defect categories
Notice that X6 behaves differently than the rest
Allow two defects, where one defect in X6
This results in the very best batch appearing
Multidimensional Detective

Fig 5 and 6 show that high yield batches don’t have non-zero values for defects of type X3 and X6

Don’t believe your assumptions …

Looking now at X15 we see the separation is important

Lower values of this property end up in the better yield batches
Automated Analysis

A. Inselberg, Automated Knowledge Discovery using Parallel Coordinates, INFOVIS '99
Case Study: E-Commerce Clickstream Visualization

Brainerd & Becker, IEEE Infovis 2001

Aggregate nodes using an icon (e.g. all the checkout pages)

Edges represent transitions
Wider means more transitions
Customer Segments

Collect

- Clickstream
- Purchase history
- Demographic data

Associates customer data with their clickstream

Different color for each customer segment
Layout

Aggregation based on file system path
Initial Findings

Gender shopping differences
Initial Findings (cont)

Checkout process analysis
Newsletter hurting sales
WebQuilt

Interactive, zoomable directed graph
Nodes = web pages
Edges = aggregate traffic between pages

Directed graph

Nodes: visited pages
  Color marks entry and exit nodes

Arrows: traversed links
  Thicker: more heavily traversed
  Color
    Red/yellow: Time spend before clicking
    Blue: optimal path chosen by designer
http://pda.edmunds.com

Where Smart Car Buyers Start

Edmunds2Go!

- Vehicle Prices & Reviews
- Dealer Locator
- Auto Tools

About Us | Help

© 2000-2001 Edmunds.com Inc.
Pilot Usability Study

Edmunds.com PDA web site
Visor Handspring equipped with a OmniSky wireless modem
10 users asked to find…
   Anti-lock brake information on the latest Nissan Sentra model
   The Nissan dealer closest to them.
In the Lab vs. Out in the Wild

Comparing in-lab usability testing with WebQuilt remote usability testing

5 users were tested in the lab

5 were given the device and asked to perform the task at their convenience

All task directions, demographic data, and follow up questionnaire data was presented and collected in web forms as part of the WebQuilt testing framework.
Entry pages green
Exit pages cyan
## Findings

<table>
<thead>
<tr>
<th>Browser</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Interact before load (3)</td>
<td>• Difficulty with input in questionnaire (3)</td>
</tr>
<tr>
<td>• No forward button (2)</td>
<td>• <strong>Difficulty scrolling</strong> (2)</td>
</tr>
<tr>
<td></td>
<td>• Device errors unrelated to testing (1)</td>
</tr>
<tr>
<td></td>
<td>• Tried writing on screen (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site Design</th>
<th>Test Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Falsely completed task (4)</td>
<td>• <strong>Falsely completed task</strong> (4)</td>
</tr>
<tr>
<td>• Long download times (4)</td>
<td>• <strong>Difficulty remembering task description</strong> (3)</td>
</tr>
<tr>
<td>• Ping-pong behavior (3)</td>
<td>• Difficulty with input in questionnaire (3)</td>
</tr>
<tr>
<td>• Interact before load (3)</td>
<td>• Questionnaire wording problems (3)</td>
</tr>
<tr>
<td>• <strong>Too much scrolling</strong> (2)</td>
<td>• Forgot how to end task (1)</td>
</tr>
<tr>
<td>• Save address functionality not clear (1)</td>
<td>• Confusing task description (1)</td>
</tr>
<tr>
<td>• <strong>Back button navigation</strong> (0)</td>
<td></td>
</tr>
<tr>
<td>• Would like more features (0)</td>
<td></td>
</tr>
<tr>
<td>• Finds site useful (0)</td>
<td></td>
</tr>
</tbody>
</table>
Findings

WebQuilt methodology is promising for uncovering site design related issues.

1/3 of the issues were device or browser related.

- Browser and device issues can not be captured automatically with WebQuilt unless they cause an interaction with the server
- Can be revealed via the questionnaire data.
Visualization for Analysis

Carlis & Konstan, UIST 1998

Problem: data that is both periodic and serial
- Time students spend on different activities
- Tree growth patterns
  - Time: which year
  - Period: yearly
- Multi-day races such as the Tour de France
- Calendars arbitrarily wrap around at end of month
- Octaves in music

How to find patterns along both dimensions?
Analyzing Complex Periodic Data

Baphia Cappardifolia:
tropical plant

Figure 1. A Spiral of Archimedes. (for color figures see the electronic proceedings or www.cs.umn.edu/~carlis)

Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Cappardifolia during the period 1980 – 1988.

Carlis & Konstan, UIST 1998.
Analyzing Complex Periodic Data

- Consumption values for each month appear as spikes
- Each food has its own color
- Boundary line (in black) shows when season begins/ends

Figure 4. A spiral display of year-month consumption percentages for 12 highly consumed foods during the period 1980 – 1988. Rotated and zoomed in to show one season and boundary lines.

Carlis & Konstan, UIST 1998.
Figure 3. A spiral display of monthly consumption percentages for all 112 foods during the period 1980 – 1988.
Visualization vs. Analysis?

Applications to data mining and data discovery.

Wesley Johnson ’02:

Visualization tools are helpful for exploring hunches and presenting results

Examples: scatterplots

They are the WRONG primary tool when the goal is to find a good classifier model in a complex situation.

Need:

Solid insight into the domain and problem
Tools that visualize several alternative models.
Emphasize “model visualization” rather than “data visualization”
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Visualizing Documents and Search
Documents and Search

Why Visualize Text?
Why Text is Tough
Visualizing Concept Spaces
  Clusters
  Category Hierarchies
Visualizing Retrieval Results
Usability Study Meta-Analysis
Why Visualize Text?

To help with Information Retrieval
  give an overview of a collection
  show user what aspects of their interests are present in a collection
  help user understand why documents retrieved as a result of a query

Text Data Mining
  Mainly clustering & nodes-and-links

Software Engineering
  not really text, but has some similar properties
Why Text is Tough

Text is not pre-attentive

Text consists of abstract concepts
  which are difficult to visualize

Text represents similar concepts in many different ways
  space ship, flying saucer, UFO, figment of imagination

Text has very high dimensionality
  Tens or hundreds of thousands of features
  Many subsets can be combined together
As the man walks the cavorting dog, thoughts arrive unbidden of the previous spring, so unlike this one, in which walking was marching and dogs were baleful sentinals outside unjust halls.
Why Text is Tough

Abstract concepts are difficult to visualize

Combinations of abstract concepts are even more difficult to visualize

- time
- shades of meaning
- social and psychological concepts
- causal relationships
Why Text is Tough

Language only hints at meaning

Most meaning of text lies within our minds and common understanding

“How much is that doggy in the window?”

how much: social system of barter and trade (not the size of the dog)

“doggy” implies childlike, plaintive, probably cannot do the purchasing on their own

“in the window” implies behind a store window, not really inside a window, requires notion of window shopping
Why Text is Easy

Text is highly redundant

When you have lots of it

Pretty much any simple technique can pull out phrases that seem to characterize a document

Instant summary:

Extract the most frequent words from a text
Remove the most common English words

People are very good at attributing meaning to lists of otherwise unrelated words
Guess the Text:

10 PEOPLE  4 LARGE
10 ALL      4 INDEPENDENT
9 STATES    4 FREE
9 LAWS      4 DECLARATION
8 NEW       4 ASSENT
7 RIGHT     3 WORLD
7 GEORGE    3 WAR
6 WILLIAM   3 USURPATIONS
6 THOMAS    3 UNITED
6 JOHN      3 SEAS
6 GOVERNMENT 3 RIGHTS
5 TIME
5 POWERS
5 COLONIES
Visualization of Text Collections

How to summarize the contents of hundreds, thousands, tens of thousands of texts?

Many have proposed clustering the words and showing points of light in a 2D or 3D space.

Examples

  Showing docs/collections as a word space
  Showing retrieval results as points in word space
TextArc.org (Bradford Paley)
Department of Computer Science and Engineering

Quality of Visualization

Galaxy of News

Rennison 95

Galaxy of News

Rennison 95
Winter storm dumps more snow on weary region

Galaxy of News

Winter storm dumps more snow on weary region, slows traffic

Quality of Visualization

Reynison 95
Themescapes (Wise et al. 95)
ScatterPlot of Clusters
(Chen et al. 97)
Clustering for Collection Overviews

Two main steps

- cluster the documents according to the words they have in common
- map the cluster representation onto a (interactive) 2D or 3D representation

Since text has tens of thousands of features

- the mapping to 2D loses a tremendous amount of information
- only very coarse themes are detected
<table>
<thead>
<tr>
<th>Cluster 1 Size: 8</th>
<th>key army war francis spangle banner air song scott word poem british</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star–Spangled Banner, The</td>
<td></td>
</tr>
<tr>
<td>Key, Francis Scott</td>
<td></td>
</tr>
<tr>
<td>Fort McHenry</td>
<td></td>
</tr>
<tr>
<td>Arnold, Henry Harley</td>
<td></td>
</tr>
<tr>
<td>Briton, Anthus</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2 Size: 68</th>
<th>film play career win television role record award york popular stage p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burstyn, Ellen</td>
<td></td>
</tr>
<tr>
<td>Stanwyck, Barbara</td>
<td></td>
</tr>
<tr>
<td>Berle, Milton</td>
<td></td>
</tr>
<tr>
<td>Zukor, Adolph</td>
<td></td>
</tr>
<tr>
<td>Broadway, Talk of the Town</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 3 Size: 97</th>
<th>bright magnitude cluster constellation line type contain period spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>star</td>
<td></td>
</tr>
<tr>
<td>Galaxy, The</td>
<td></td>
</tr>
<tr>
<td>extragalactic systems</td>
<td></td>
</tr>
<tr>
<td>interstellar matter</td>
<td></td>
</tr>
<tr>
<td>constellation, star</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 4 Size: 67</th>
<th>astronomer observatory astronomy position measure celestial telescope</th>
</tr>
</thead>
<tbody>
<tr>
<td>astronomy and astrophysics</td>
<td></td>
</tr>
<tr>
<td>astrometry</td>
<td></td>
</tr>
<tr>
<td>Agena</td>
<td></td>
</tr>
<tr>
<td>astronomical catalogs and atlases</td>
<td></td>
</tr>
<tr>
<td>Harmakal, Sir William</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 5 Size: 10</th>
<th>family specie flower animal arm plant shape leaf brittle tube foot horn</th>
</tr>
</thead>
<tbody>
<tr>
<td>blazing star</td>
<td></td>
</tr>
<tr>
<td>brittle star</td>
<td></td>
</tr>
<tr>
<td>bishop’s–cap</td>
<td></td>
</tr>
<tr>
<td>feather star</td>
<td></td>
</tr>
</tbody>
</table>
How Useful is Collection Cluster Visualization for Search?

Three studies find negative results
Study 1


This study compared

a system with 2D graphical clusters
a system with 3D graphical clusters
a system that shows textual clusters

Novice users

Only textual clusters were helpful (and they were difficult to use well)
Study 2: Kohonen Feature Maps


Comparison: Kohonen Map and Yahoo

i.e. self organizing map for entertainment sub-category of Yahoo

Task:

“Window shop” for interesting home page
Repeat with other interface

Results:

Starting with map could repeat in Yahoo (8/11)
Starting with Yahoo unable to repeat in map (2/14)

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.128&rep=rep1&type=pdf
Study 2 (cont.)

Participants liked:

- Correspondence of region size to # documents
- Overview (but also wanted zoom)
- Ease of jumping from one topic to another
- Multiple routes to topics
- Use of category and subcategory labels
Study 2 (cont.)

Participants wanted:

- hierarchical organization
- other ordering of concepts (alphabetical)
- integration of browsing and search
- correspondence of color to meaning
- more meaningful labels
- labels at same level of abstraction
- fit more labels in the given space
- combined keyword and category search
- multiple category assignment (sports+entertain)
NIRVE Interface by Cugini et al. 96. Each rectangle is a cluster. Larger clusters closer to the “pole”. Similar clusters near one another. Opening a cluster causes a projection that shows the titles.

Study 3


This study compared:
- 3D graphical clusters
- 2D graphical clusters
- textual clusters

15 participants, between-subject design

Tasks
- Locate a particular document
- Locate and mark a particular document
- Locate a previously marked document
- Locate all clusters that discuss some topic
- List more frequently represented topics
Study 3

Results (time to locate targets)
Text clusters fastest
2D next
3D last
With practice (6 sessions) 2D neared text results; 3D still slower
Computer experts were just as fast with 3D

Certain tasks equally fast with 2D & text
Find particular cluster
Find an already-marked document

But anything involving text (e.g., find title) much faster with text.
Spatial location rotated, so users lost context

Helpful viz features
Color coding (helped text too)
Relative vertical locations
Summary: Visualizing Clusters

Huge 2D maps may be inappropriate focus for information retrieval

- cannot see what the documents are about
- space is difficult to browse for IR purposes
- (tough to visualize abstract concepts)

Perhaps more suited for pattern discovery and gist-like overviews
IR Infovis Meta-Analysis

(Empirical studies of information visualization: a meta-analysis, Chen & Yu IJHCS 53(5),2000)

Goal

Find invariant underlying relations suggested collectively by empirical findings from many different studies

Procedure

Examine the literature of empirical infoviz studies
35 studies between 1991 and 2000
27 focused on information retrieval tasks
But due to wide differences in the conduct of the studies and the reporting of statistics, could use only 6 studies
IR Infovis Meta-Analysis

(Empirical studies of information visualization: a meta-analysis, Chen & Yu IJHCS 53(5),2000)

Conclusions:

IR Infoviz studies not reported in a standard format

Individual cognitive differences had the largest effect
   Especially on accuracy
   Somewhat on efficiency

Holding cognitive abilities constant, users did better with simpler visual-spatial interfaces

The combined effect of visualization is not statistically significant
So What Works?


Color highlighting of query terms in results listings
Sorting of search results according to important criteria (date, author)
Grouping of results according to well-organized category labels.
  Cha-cha
  Flamenco

Only if highly accurate:
  Spelling correction/suggestions
  Simple relevance feedback (more-like-this)
  Certain types of term expansion

Note: most don’t benefit from visualization!
Cha-Cha

Teoma: appears to combine categories and clusters

(this version before it was bought by askjeeves)
Teoma: Now in prime time

Showing 1-10 of about 291,000

TEMIS - Text Mining Solutions
Text Mining Technology and Consulting company offering software components ("Insight Discoverer") for the efficient analysis of large document...
www.temis-group.com/

Untangling Text Data Mining
Defines data mining, information access, and corpus-based computational linguistics, and then discusses the relationship of these to text data...
www.sims.berkeley.edu/~hearst/papers/acl99...
[Related Pages]

Hearst, Marti
Untangling Text Data Mining, ACL'99.
www.sims.berkeley.edu/~hearst/
[Related Pages]
[More results from www.sims.berkeley.edu]

Text Mining at Waikato
The Text Mining group at the University of Waikato in New Zealand. With a focus on Viterbi search and entropy-based methods the group has a...
www.cs.waikato.ac.nz/~nzdl/textmining/
[Related Pages]
[More results from www.cs.waikato.ac.nz]

KRDL Text Mining Home Page
Text Mining Research Group at Kent Ridge Digital Labs in Singapore; charset=iso-8859-1...
textmining.krdl.org.sg/
[Related Pages]
Better to reduce the viz

Flamenco – allows users to steer through the category space

Uses
- Dynamically-generated hypertext
- Color for distinguishing and grouping
- Careful layout and font choices

Focused first on the users’ needs
### Media
- Aquatint (2025)
- Basketry (44)
- Book (666)
- Ceramic (1008)
- Costume (660)
- Decorative box (163)
- Domestic object (176)
- Drawing (2624)
- Drypoint (2143)
- Etching (9507)
- Furnishing (127)
- Glass (651)
- More...

### Nature
- Animal material (515)
- Birds (1437)
- Bodies of water (3604)
- Creatures (801)
- Fish (219)
- Flowers (1220)
- Geological formations (2122)
- Heavens (2353)
- Hoofed mammals (2480)
- Invertebrates and arthropods (330)
- Mammals (2116)
- Plant material (788)
- More...

### Location
- Africa (463)
- Asia (1325)
- Australia (21)
- Central America (134)
- Europe (23331)
- Middle East (78)
- North America (11111)
- Oceania (111)
- Roman Empire (4)
- South America (453)

### Places and Spaces
- Bridges (592)
- Building parts (3088)
- Buildings (2393)
- Dwellings (1709)
- Lawn (20)
- Open spaces (1732)
- Roads (1480)
- Workplaces (753)

### Date
- 1 - 1000 A.D. (138)
- 12th century (3)
- 13th century (1)
- 14th century (3)
- 15th century (76)
- 16th century (1225)
- 17th century (3058)
- 18th century (2287)
- 19th century (7552)
- 20th - 21st century (18)
- 20th century (14295)
- 21st century (12)
- More...

### People
- Aristocrats (974)
- Children (2501)
- Men (7372)
- Occupations (715)
- Women (5906)

### Shapes, Colors, and Materials
- Colors (5861)
- Decorations (1441)
- Fabrics (345)
- Metal (273)
- Paper (457)
- Shapes (2752)
- Visual framing (5911)
Using Thumbnails to Search the Web


Design Goals

- Enhance features that help the user decide whether document is relevant to their query
  - Emphasize text that is relevant to query
    - Text callouts
  - Enlarge (make readable) text that might be helpful in assessing page
    - Enlarge headers
Text and Image Summaries

Text summaries
Lots of abstract, semantic information

Image summaries (plain thumbnails)
Layout, genre information
Gist extraction faster than with text

Benefits are complementary
Create textually-enhanced thumbnails that leverage the advantages of both text summaries and plain thumbnails
Putting Callouts in a Separate Visual Layer

Transparency
Occlusion

Junctions indicate the occurrence of these events.
Design Issues:

Color Management

Problems: Callouts need to be both readable and draw attention
Solution: Desaturate the background image, and use a visual search model to choose appropriate colors
Colors look like those in highlighter pens

Resizing of Text

Problem: We want to make certain text elements readable, but not necessarily draw attention to them
Solution: Modify the HTML before rendering the thumbnail
Examples

- Hybrid car mileage
- MiniDisc Player

---

Department of Computer Science and Engineering
Tasks

Criteria: tasks that…

- Are representative of common queries
- Have result sets with different characteristics
- Vary in the number of correct answers

4 types of tasks

- **Picture:** “Find a picture of a giraffe in the wild.”
- **Homepage:** “Find Kern Holoman’s homepage.”
- **Side-effects:** “Find at least three side effects of halcion.”
- **E-commerce:** “Find an e-commerce site where you can buy a DVD player. Identify the price in dollars.”
Conditions

Text summary

The Lycaenum -- GHB
......seems to be without serious side effects." His almost off-hand...
...recovery with no long-term side effects is universal." They...
www.algonet.se/~spot/arch/texts/ghbfaq.html

Department of Computer Science and Engineering
Collections of Summaries

100 results in random order

Approximately same number of each summary type on a page
Method

Procedure

18 questions, with 100 query results each
Entire process took about 75 minutes

6 practice tasks

3 questions for each of the 4 task types

e.g., each participant would do one E-commerce question using text, one E-commerce question using plain thumbnails, and one E-commerce question using enhanced thumbnails

Questions blocked by type of summary
WebLogger recorded user actions during browsing
Semi-structured interview

Participants

12 members of the PARC community
Results

Average total search times, by task:

- Picture: 61 secs
- Homepage: 80 secs
- E-commerce: 64 secs
- Side effects: 128 secs

Results pooled across all tasks:

- Subjects searched 20 seconds faster with enhanced thumbnails than with plain
- Subjects searched 30 seconds faster with enhanced thumbnails than with text summaries
- Mean search time overall was 83 seconds
Results

Normalized total search time (s)
Results: User Responses

Participants preferred enhanced thumbnails
  7/12 preferred overall
  5/12 preferred for certain task types

Enhanced thumbnails are intuitive and less work than text or plain thumbnails
  One subject said searching for information with text summaries did not seem hard until he used the enhanced thumbnails.

Many participants reported using genre information, cues from the callouts, the relationship between search terms, etc.
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Comparing Approaches
Comparing 3 Commercial Systems

Alfred Kobsa, An Empirical Comparison of Three Commercial Information Visualization Systems, INFOVIS'01.

Figure 1a. A screenshot from Eureka that shows how a user might solve the question “Did males cheat more on their girlfriends than females on their boyfriends?” (an actual question used in the experiment). After grouping the attribute “Gender” and sorting the column “Did you cheat?,” a user can compare the number of “Yes” entries and thus find that more females than males indicated having cheated. (One male and one female gave no answers.)

Figure 1b. This screenshot shows one possible way for solving the same problem in InfoZoom, specifically in its overview mode. After clicking at, and thereby zooming into, the “Yes” entries in the attribute “Did you cheat?,” users can see from the length of the bars in the Gender category that females indicated more frequently having cheated than males.

Figure 2. Spotfire's geographical representation of heavy metal concentrations through a scatterplot diagram.
Comparing 3 Commercial Systems

Eureka (InXight)

Figure 1a. A screenshot from Eureka that shows how a user might solve the question “Did males cheat more on their girlfriends than females on their boyfriends?” (an actual question used in the experiment). After grouping the attribute “Gender” and sorting the column “Did you cheat?”, a user can compare the number of “Yes” entries and thus find that more females than males indicated having cheated. (One male and one female gave no answers.)
Comparing 3 Commercial Systems

InfoZoom (HumanIT)

Figure 1b. This screenshot shows one possible way for solving the same problem in InfoZoom, specifically in its overview mode. After clicking at, and thereby zooming into, the “Yes” entries in the attribute “Did you cheat?”, users can see from the length of the bars in the Gender category that females indicated more frequently having cheated than males.
Comparing 3 Commercial Systems

SpotFire
Infozoom Overview

• Presents data in three different views.

  • Wide view shows data set in a table format.

  • Compressed view packs the data set horizontally to fit the window width.

  • Overview mode has all attributes in ascending or descending order and independent of each other.
## InfoZoom Overview View

The actors with the most films

<table>
<thead>
<tr>
<th>Film</th>
<th>Title</th>
<th>Year</th>
<th>Length</th>
<th>Subject</th>
<th>Actor</th>
<th>Actress</th>
<th>Director</th>
<th>Popularity</th>
<th>Awards</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>90</td>
<td>91</td>
<td>94</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Action</td>
<td>Comedy</td>
<td>Drama</td>
<td>Horror</td>
<td>Music</td>
<td>Mystery</td>
<td>SciFi</td>
<td>Westerns</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of objects: 1740

The number of attributes: 10

10 attributes differ

Status: Ready to Visualize
InfoZoom Overview View
### InfoZoom Compressed Table View

#### 6 Quality of Visualization

![InfoZoom Compressed Table View Image]

220 of 220 Objects
42 Attributes differ

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company Short Name</td>
<td>Company Name, Location, Structure of Corporation, CEO, CEO (Photo),</td>
</tr>
<tr>
<td>Company Name</td>
<td>Upper Management, Director of Staff, Director of Staff (Photo), Division Head</td>
</tr>
<tr>
<td>Location</td>
<td>Type of Company Relation, Markets, Lines of Business, Parent Company</td>
</tr>
</tbody>
</table>
# InfoZoom Wide Table View

## Quality of Visualization

### The actors with the most films

<table>
<thead>
<tr>
<th>Queries</th>
<th>Perform</th>
</tr>
</thead>
</table>

### 1740 of 1740 Objects

<table>
<thead>
<tr>
<th>10 Attributes differ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

### Film Title

- Wild at Heart
- Goodbye Again
- Hunt for Red October, The
- Terminator, The
- Terminator 2
- John Cleese on How to Irritate People
- Au Revoir les Enfants
- The Ballad of Narayama
- Cyrano De Bergerac

### Year

- 1990
- 1961
- 1990
- 1984
- 1991
- 1993
- 1987
- 1983
- 1999

### Length

- 125
- 120
- 135
- 108
- 136
- 65
- 193
- 128
- 13

### Subject

- Drama
- Drama
- Drama
- Action
- Action
- Comedy
- Drama
- Drama
- Drama

### Actor

- Cage, Nicolas
- Perkins, Anthony
- Connery, Sean
- Schwarzenegger A.
- Schwarzenegger A.
- Cleese, John
- Manesse, Gaspard
- Depardieu, Gerard

### Actress

- Dern, Laura
- Bergman, Ingrid
- Hamilton, Linda
- Hamilton, Linda
- Booth, Connie
- Racette, Francine
- Missing
- Brochet

### Director

- Lynch, David
- Litvak, Anatole
- McTernan, J.
- Cameron, J.
- Cameron, J.
- Malle, Louis
- Imamura, Shohei
- Rappeneau, Jean-Paul

### Popularity

- 6
- 6
- 8
- 17
- 8
- 62
- 35
- 15
- 86

### Awards

- No
- No
- No
- No
- No
- No
- No
- No
- No

Show table in uncompressed mode

NUM

---
Datasets

• Multidimensional data: three databases were used
  • Anonymized data from a web based dating service (60 records, 27 variables)
  • Technical data of cars sold in 1970 – 82 (406 records, 10 variables)
  • Data on the concentration of heavy metals in Sweden (2298 records, 14 variables)
Sample Questions

Do more women than men want their partners to have a higher education?

What proportion of the men live in California?

Do all people who think the bar is a good place to meet a mate also believe in love at first site?

Do heavier cars have more horsepower?

Which manufacturer produced the most cars in 1980?

Is there a relationship between the displacement and acceleration of a vehicle?
Experiment Design

• The experimenters generated 26 tasks from all three data sets.
• 83 participants. Between-subjects design.
• Each was given one visualization system and all three data sets.
• Type of visualization system was the independent variable between them.
• 30 mins were given to solve the tasks of each data set i.e 26 tasks in 90 mins.
Overall Results

- Mean task completion times:
  - Infozoom users: 80 secs
  - Spotfire users: 107 secs
  - Eureka users: 110 secs

- Answer correctness:
  - Infozoom users: 68%
  - Spotfire users: 75%
  - Eureka users: 71%

- Not a time-error tradeoff
- Spotfire more accurate only 6 questions
Eureka - problems

• Hidden labels: Labels are vertically aligned, max 20 dimensions

• 3+ Attributes: Problems with queries involving three or more attributes

• Correlation problems: Some participants had trouble answering questions correctly that involved correlations between two attributes.
Spotfire - problems

• Cognitive setup costs: Takes participants considerable time to decide on the right representation and to correctly set the coordinates and parameters.

• Biased by scatterplot default: Though powerful, many problems cannot be solved (well) with it.
Infozoom - problems

• Erroneous Correlations
  • Overview mode has all attributes sorted independent of each other

• Narrow row height in compressed view

• Participants did not use row expansion and scatterplot charting function which shows correlations more accurately
Geographic Questions

• Spotfire should have done better on these
  • Which part of the country has the most copper
  • Is there a relationship between the concentration of vanadium and that of zinc?
  • Is there a low-level chrome area that is high in vanadium
• Spotfire was only better only for the last question (out of 6 geographic ones)
Discussion

• Many studies of this kind use relatively simple tasks that mirror the strengths of the system
  • Find the one object with the maximum value for a property
  • Count how many of certain attributes there are
• This study looked at more complex, realistic, and varied questions.
Discussion

• Success of a visualization system depends on many factors:
  • Properties supplied
    • Spotfire doesn’t visualize as many dimensions simultaneously
  • Operations
    • Zooming easy in InfoZoom; allows for drill-down as well
    • Zooming in Eureka causes context to be lost
    • Column view in Eureka makes labels hard to see
Information Exploration “Shootout”
http://ivpr.cs.uml.edu/shootout/about.html

Data Mining Applications

One component focuses on visualization
Comparing Tree Views

T. Barlow and P. Neville, Comparison of 2D Visualizations of Hierarchies, INFOVIS’01.

Problem

Organization Chart is de facto standard for visualizing decision trees. Is there a better compact view of the tree for the overview window?

Solution

Two usability studies to determine which tree works best.
Goal: Compact View of Tools

Figure 1. Example of compact view in data mining
Decision Trees

Each split constitutes a rule or variable in predictive model

Begin Splitting into nodes

Often hundreds of leaves
Decision Trees – What makes a good visualization

Uses
- For novice-helps them understand models
- Experts-initial evaluation of decisions tree without looking at models

Criteria for usability in study
- Ease of Interpretation of Topology (Parent Child Sibling relations)
- Comparison of Node Size
- User preference
Different views examined in study

Org Chart     Tree Ring     Icicle Plot     TreeMap

Figure 2. Different views of the same tree
Usability Test 1:

Users:
15 colleagues familiar with org chart but not others

Tasks
Is the tree binary or n-ary?
Is the tree balanced or unbalanced?
Find deepest common ancestor of two nodes
Number of levels?
Find three largest leaves (excluding org chart)

Data: Created 8 trees for analysis

Study Design
Randomized order of tasks
4X5 design (almost)
Timed task from appearance on screen until spacebar tap
Results

Response Time

TreeMap slowest; no statistical difference between others

Response Accuracy

No significant difference

User Preference

Prefer icicle map and org chart (faster)
Dislike tree map
Discussion

Org chart served as benchmark

Icicle plot favored amongst others

Hypothesis: Same left to right / top to bottom structure

TreeRing did well

TreeMap suffered from poor accuracy

Offset of rectangles required because of off (which is needed for selection)
Usability Test II: Tree implementation

Three views:

- TreeMap eliminated from this round

Tasks

Node Description

- Four versions – select those nodes or leaves that meet certain criteria

Node Analysis:

- Memorize a highlighted node – find again after tree redrawn in different position
Results

Tree rings slower for description but fast and accurate for memory tasks

Perhaps due to unique geometric forms / spatial clues
Conclusions

TreeMap not useful for this type of task
Org Chart/Icicle seem to be best overall
TreeRing has merits for certain tasks

Icicle chosen for implementation

  Best design considering Org Chart could not be used for node size tasks

However:

  Didn’t seem to actually do tests on trees as large as the ones they describe as typical of datamining
Visualizing Conversations
Text-Based Chat

Katesmiles1 enters
You tell Horse_99 me too
Horse_99 says Real.
Speci_Man_98 says Where you from Kim?
Soapbox_7 leaves, heading for the Gen-X Love #19
Horse_99 says On here/
Horse_99 says Lets go private and find out.
Muta4 leaves heading for another room
Muta4 leaves
Horse_99 says Sure.
Speci_Man_98 says Ever been to new York?

Katesmiles1 enters
You tell Horse_99 me too
Horse_99 says Real.
Speci_Man_98 says Where you from Kim?
Soapbox_7 leaves, heading for the Gen-X Love #19
Soapbox_7 says
GODESS_OFLOVE leaves, heading for the Gen-X Love #19
Kim_24_98 says connecticut
Horse_99 How old are you Mika?
Sycam leaves
Kim_24_98 says where are you from
Muta4 says 16
You tell Horse_99 are you talking to me when you say Mika?
Speci_Man_98 says Paducah, ky
Horse_99 says 24/m/a
Horse_99 says Yes
You say 26
Kim_24_98 says cool, I'm 25f
Horse_99 says On here/
Horse_99 says Lets go private and find out.
Muta4 leaves heading for another room
Muta4 leaves
Horse_99 says Sure.
Speci_Man_98 says Ever been to new York?
Chat Circles
Chat Circles

“Chat Circles is a graphical interface for synchronous communication that uses abstract shapes to convey identity and activity.”

Each participant appears as a colored circle, which is accompanied by the user name.

Location of circles will also identify participants (important for many users having similar colors associated).

Participants’ circles become larger when posting occurs (circle adapts to text length).

Circle appears bright when posting occurs.

Circles of inactive users fade in the background.
Chat Circles – Conversational Groupings

There is only ONE room in Chat Circles
Groupings are achieved by moving closer to other participants
At any time, a participant can view all other participants
A participant can also detect interesting conversations in different areas of the room by looking at how many circles are gathered and how often circles become larger
Overview panel in Chat Circles II nice example of focus + context
Chat Circles History

I also thought that was very weird... Do you remember the last time it happened?
History Log Patterns

+ Easy to see “lurkers”
+ Sequence and size of messages quickly visible

- Not very scalable
History Log Patterns

+/- User-centric: only 1 point of view represented

- Impossible to see all the text at once – requires individual mouse rollovers

- Easy to see “out of range” conversations – *but why would you want to?*
Agenda

Introduction
Visual Principles
What Works?
Visualization in Analysis & Problem Solving
Visualizing Documents & Search
Comparing Visualization Techniques
Design Exercise
Wrap-Up
Design Exercise
Design Exercise

BreakingStory

(Reffel, Fitzpatrick, Ayedelott SIMS final project, at CHI 2003)

Create an application that supplies a visualization for trends over time in web-based news. The primary purpose is to provide an overview, but it should also be possible to view text from individual news sources on specific days. Its goal is to inform, inspire, and enlighten, and also to make people want to look more deeply at the news.
What is BreakingStory?

BreakingStory is a tool to help you explore online news. Curious about when a phrase was first mentioned? Wondering if references to a current event vary by geographic region? Here you can find out!

How do you use it?

You can use this site to search for words or phrases that have appeared in online news. The results will be displayed in one or more charts that show the history of references.

Single Chart View allows you to examine references from one geographic area or news site.

Multiple Chart View shows a number of small charts next to each other, allowing you to compare references from different geographic areas or news sites.

You can also view the full text of the front pages that matched your search.

Don't know where to begin? Try looking at an example: World Aids Day

Want More?

- Help - suggestions, advice for using BreakingStory
- Site List - the news sites we collect and their locations
- FAQ - the who what when where how of BreakingStory
- Project Page - detailed information on the project
- E-mail us - we want to hear your comments, really we do
Front page hits:

- oil: 0.26 (1794 total references in 1305 pages)
- peace: 0.22 (1494 total references in 1163 pages)

Selected date range:

17 August 2002 to 04 April 2003 (231 days total)
Another Approach: ThemeRiver

Wrap-up: Guidelines for Success
Key Questions to Ask about a Viz

1. Is it for analysis or presentation?
2. What does it teach/show/elucidate?
3. What is the key contribution?
4. What are some compelling, *useful* examples?
5. Could it have been done more simply?
6. Have there been usability studies done? What do they show?
Holistic Design Goals for Information Visualization

- Tailor to the application and the domain
- Create highly interactive and integrated systems
- Embed the visualization within a larger application
- Provide alternative views
Visualization with a Light Touch: Orbitz.com
Visualization with a Light Touch: Orbitz.com
Visualization with a Light Touch: Orbitz.com
Visualization with a Light Touch: Orbitz.com

Fri, Apr 11  anytime  Oakland (All Locations), CA (OAK)
     New York (All Locations), NY (NYC)

Fri, Apr 18  anytime  New York (All Locations), NY (NYC)
     Oakland (All Locations), CA (OAK)

prices above are per person and may not be purchased on Orbitz without applicable service fees

We checked for better fares at nearby airports, and your search found our lowest prices.
Lowest fare below beats this trip’s 30-day average by $81
Visualization with a Light Touch: Orbitz.com

**Quality of Visualization**

**Orbitz.com**

**Search Location:** Fort Lauderdale (Fort Lauderdale Beach), Florida, United States

### Showing All Hotels

<table>
<thead>
<tr>
<th>Distance</th>
<th>Price Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2 miles</td>
<td>from $63 to $59</td>
</tr>
<tr>
<td>2-5 miles</td>
<td>from $45 to $99</td>
</tr>
<tr>
<td>5-15 miles</td>
<td>from $43 to $270</td>
</tr>
</tbody>
</table>

### List Hotels by:

- Price
- Distance

### From 119

**Holiday Inn Ft Lauderdale Airport**

- 999 Fort Lauderdale Boulevard
- 0.4 Miles East of Fort Lauderdale International Airport

Each of our 240 guestrooms has a safe and refrigerator, irons, data port, TV and in room movies and voice mail.

**Holiday Inn**

- avg. nightly rate
- Orbitz Saver
- 118.80 USD
- more rates and rooms

### From 209

**Marriott Fort Lauderdale Marina**

- 1881 Southeast 17th Street, Fort Lauderdale, FL 33316
- 2.7 Miles South of Fort Lauderdale Beach

Located two miles south of downtown, the Fort Lauderdale Marriott Marina is nestled in an upscale marina on Florida's beautiful Intracoastal Waterway: a series of residential canals leading to the ocean. The Broward County Convention Center and Port...

<table>
<thead>
<tr>
<th>Hotel and Room Details, Photos and Maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectacular view of our tropically landscaped...</td>
</tr>
</tbody>
</table>

**Marriott**

- avg. nightly rate
- 209.00 USD
- Orbitz Saver
- 219.00 USD
- more rates and rooms

**Doubletree Hotel Ft Lauderdale-Oceanfront**

- 440 Seabreeze Boulevard, Ft Lauderdale, FL 33316
- 1.4 Miles South of Fort Lauderdale Beach

Ideally located on beautiful Ft Lauderdale Beach, bordered by the sparkling Atlantic and the Intracoastal Waterway, the Doubletree Ft. Lauderdale Oceanfront is your perfect...

- avg. nightly rate
- Orbitz Saver
- 118.80 USD
- more rates and rooms
For more information

My course:
http://www.sims.berkeley.edu/courses/is247/s02/Lectures.html

Atlas of Cyberspaces:
http://www.geog.ucl.ac.uk/casa/martin/atlas/atlas.html

Gallery of Data Visualization; The Best and Worst of Statistical Graphics
http://www.math.yorku.ca/SCS/Gallery/

Tamara Munzner’s collection:
http://graphics.stanford.edu/courses/cs348c-96-fall/resources.html