Lecture 34 of 41

Visualization, Part 1 of 3: Data (Quantities & Evidence)

William H. Hsu
Department of Computing and Information Sciences, KSU

Public mirror web site: http://www.kddresearch.org/Courses/CIS636
Instructor home page: http://www.cis.ksu.edu/~bhsu

Readings:
Last class: Chapter 15, Eberly 2nd see http://bit.ly/ieUq45; Ray Tracing Handout
Today: Tufte Handout 1
Next class: Ray Tracing Handout

Lecture Outline

- Reading for Last Class: Chapter 15, Eberly 2nd; Ray Tracing Handout
- Reading for Today: Tufte Handout
- Reading for Next Class: Ray Tracing Handout
- Last Time: Ray Tracing 2 of 2
  - Stochastic & distributed RT
    - Stochastic (local) vs. distributed (nonlocal) randomization
    - “Softening” shadows, reflection, transparency
  - Hybrid global illumination: RT with progressive refinement radiosity
- Today: Visualization Part 1 of 3 – Scientific, Data, Information Vis
  - What is visualization?
  - Tufte 1: The Visual Display of Quantitative Information, 2nd
    - Basic statistical & scientific visualization techniques
    - Graphical integrity vs. lie factor (“How to lie with statisticsvis”)
    - Graphical excellence vs. chartjunk
    - Data-ink, data-ink ratio (& “data-pixels”)
Where We Are

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Lab 4a: Animation Basics</td>
<td>Flash animation handout</td>
</tr>
<tr>
<td>22</td>
<td>Animation 2: Rotations, Dynamics, Kinematics</td>
<td>Chapter 17, esp. §17.1 – 17.2</td>
</tr>
<tr>
<td>23</td>
<td>Demos 4: Modeling &amp; Simulation: Rotations</td>
<td>Chapter 10, 11, §7.3 – 17.2</td>
</tr>
<tr>
<td>24</td>
<td>Collisions 1: axes, OBBs, Lab 4b</td>
<td>§2.4, 3, 8.1, GL handout</td>
</tr>
<tr>
<td>25</td>
<td>Spatial Sorting: Binary Space Partitioning</td>
<td>Chapter 6, esp. 6.1</td>
</tr>
<tr>
<td>26</td>
<td>Demos &amp; More CGA: Picking, HW/Exam</td>
<td>Chapter 7, § 8.4</td>
</tr>
<tr>
<td>27</td>
<td>Lab 5a: Interaction Handling</td>
<td>§ 8.3 – 8.4, 12, 5.0, 5.6, 9.1</td>
</tr>
<tr>
<td>28</td>
<td>Collisions 2: Dynamic, Particle Systems</td>
<td>§ 9.1, particle system handout</td>
</tr>
<tr>
<td>29</td>
<td>Exam 2: Review, Hour Exam 2 (evening)</td>
<td>Chapters 5 – 6, 7 – 8, 12, 17</td>
</tr>
<tr>
<td>30</td>
<td>Lab 5b: Particle Systems</td>
<td>Particle system handout</td>
</tr>
<tr>
<td>31</td>
<td>Animation 3: Control &amp; IK</td>
<td>§ 5.3, CGA handout</td>
</tr>
<tr>
<td>32</td>
<td>Ray Tracing 1: Intersections, ray trees</td>
<td>Chapter 14</td>
</tr>
<tr>
<td>33</td>
<td>Lab 6a: Ray Tracing Basics with POV-Ray</td>
<td>NT handout</td>
</tr>
<tr>
<td>34</td>
<td>Ray Tracing 2: Advanced low survey</td>
<td>Chapter 15, RT handout</td>
</tr>
<tr>
<td>35</td>
<td>Visualization 1: Data (Quantities &amp; Evidence)</td>
<td>Tufte handout (1)</td>
</tr>
<tr>
<td>36</td>
<td>Visualization 2: Objects</td>
<td>Tufte handout (2 &amp; 4)</td>
</tr>
<tr>
<td>37</td>
<td>Color Basics, Term Project Prep</td>
<td>Color handout</td>
</tr>
<tr>
<td>38</td>
<td>Lab 7: Fractals &amp; Terrain Generation</td>
<td>Fractals/Terrain handout</td>
</tr>
<tr>
<td>39</td>
<td>Visualization 2: Processes; Final Review 1</td>
<td>Tufte handout (3)</td>
</tr>
<tr>
<td>40</td>
<td>Project Presentations 1; Final Review 2</td>
<td>–</td>
</tr>
<tr>
<td>41</td>
<td>Project Presentations 2</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Final Exam</td>
<td>Ch. 1 – 8, 10 – 15, 17, 20</td>
</tr>
</tbody>
</table>

Lighly-shaded cells denote the due date of a written problem set, heavily-shaded entries that of a machine problem (programming assignment), blue-shaded entries, that of a paper review, and the green-shaded entry, that of the term project.

Green, blue and red letters denote exam review, exam, and exam solution review dates.

Acknowledgements:
Statistical & Data Visualization

Edward R. Tufte
Professor Emeritus of Political Science, Statistics, & Computer Science
Yale University
http://bit.ly/gKhM0G
http://www.edwardtufte.com

Leland Wilkinson
Executive Vice President, SYSTAT Software
Adjunct Professor of Computer Science, University of Illinois at Chicago
Adjunct Professor of Statistics, Northwestern University
http://www.cs.uic.edu/~wilkinson/
Oops!

every time you make a powerpoint
edward tufted kills a kitten

http://www.edwardtufte.com/tufte/books_vdqi
Apologia Pro PowerPoint Sua

- Tufte’s Criticisms of Microsoft PowerPoint (Summarized in Wikipedia)
  - Used to guide, reassure presenters, rather than enlighten audience
  - Unhelpfully simplistic tables, charts (due in part to low-res displays)
  - Outliners may arrange ideas in necessarily deep hierarchy
    - Not visually retained: must be repeated on each slide
    - Artifact of “outline”, “overview” format
  - Enforcement of linear progression
  - Poor design: typography, chart layout, use of templates, defaults
  - Simplistic thinking due to ideas being squashed into bulleted lists
    - Discontinuity of stories: beginning, middle, end → points
    - Cognitive load on reader: illusion of objectivity, neutrality

- Some (Though Not All) Problems with PowerPoint Avoidable by Design

Definition: Visualization

  - Images: illustrations; photographs, especially modified photos
  - Diagrams: structural diagrams, blueprints, plots & charts
  - Animations: based on simulation or other specifications
- Includes, But Not Limited to, Statistical Graphics
- Kinds of Visualization (Often Abbreviated “Vis” cf. IEEE InfoVis)
  - Scientific: transformation, representation of data for exploration
  - Data: schematic form
    - e.g., relational database form (tuples of attribute values)
    - “Data vis” often synonymous with “statistical vis”
  - Information: spectrum from “raw data” to “info”, “knowledge”
    - Premise: info more structured, organized, abstract than data
    - Emphasis on computational tools
    - Working with (especially analyzing) large data sets

Adapted from book material © 2001 E. R. Tufte, Yale University
Definition: Graphical Excellence

- Complex Ideas
- Communicated with
  - Clarity
  - Precision
  - Efficiency

Graphical Excellence:
Graphical Displays Should... (“DO’s”)

- 1. Show the data
- 2. Induce the reader to think about the substance rather than about
  - Methodology
  - Graphic design
  - Technology of graphic production
  - Something else
- 3. Avoid distorting what the data have to say
- 4. Present many numbers in a small space
- 5. Make large data sets coherent
- 6. Encourage the eye to compare different pieces of data
- 7. Reveal the data at different levels of detail, broad to fine
- 8. Serve a clear purpose: description, evaluation, tabulation, decoration
- 9. Be closely integrated with statistical and verbal descriptions of data

Rules were made to be broken (“What’s different on this slide?”)
Graphics Reveal Data [1]: Limitations of Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: \( Y = 3 + 0.5X \)
standard error of estimate of slope = 0.118
r = 4.24
sum of squares \( X - \bar{X} \) = 110.0
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
\( r^2 = .67 \)

© 2001 E. R. Tufte, Yale University

---

Graphics Reveal Data [2]: Differences in Data Shown by Vis


© 2001 E. R. Tufte, Yale University
Earliest Grid Map:
Song Dynasty, 960 – 1279 CE

This grid map, compiled c. 1100 CE (carved in stone c. 1137 CE) uses a grid of ~100 里 (li) to the square, ~42km in Han dynasty standard units (415.8m per li)

Shown:
Major rivers and tributaries

Thematic Maps & Other Data Maps

- Data Maps: Visual Presentation of Variables over Region (e.g., Spatial)
- Thematic Map: Shows Topic (Theme) Referenced by Geographic Area
- Example: 2010 UNAIDS Report on Global HIV Infection Rates

© 2010, UNAIDS
http://bit.ly/16Se1
Example – Cancer Rates by County [1]:
White Females, All Types of Cancer

All types of cancer, white females; age-adjusted rate by county, 1950–1969

© 2001 E. R. Tufte, Yale University

Example – Cancer Rates by County [2]:
White Males, All Types of Cancer

All types of cancer, white males; age-adjusted rate by county, 1950–1969

What’s different?
Compared to what?
Why?
Example – Cancer Rates by County [3]:
White Females, Respiratory Cancers

Example – Cancer Rates by County [4]:
White Males, Respiratory Cancers

Adapted from book material © 2001 E. R. Tufte, Yale University
Example – Cancer Rates by County [5]:
White Females, GI Cancers

Stomach cancer, white females;
age-adjusted rate by county, 1950–1969

Adapted from book material © 2001 E. R. Tufte, Yale University

Example – Cancer Rates by County [6]:
White Males, GI Cancers

Stomach cancer, white males;
age-adjusted rate by county, 1950–1969

Adapted from book material © 2001 E. R. Tufte, Yale University
Flow Maps [1]:
Minard, 1865 – French Wine Exports


Wikipedia:


Inspired by book material © 2001 E. R. Tufte, Yale University

Flow Maps [2]:
Minard, 1869 – Napoleon in Russia


Adapted from book material © 2001 E. R. Tufte, Yale University
Flow Maps [3]:
Minard’s Map of Napoleon’s March

- 2-D Map
- 6 Scalar Dimensions
  - Size of army
  - Location on 2-D surface by date (compare: latitude & longitude)
  - Direction of movement
  - Date
  - Temperature (referenced by position & date)

How to Represent 6 Dimensions in 2
- Size – width of line and written besides army (main camp) position
- Location – (x, y) coordinate on map; align with timeline on bottom
- Direction – color, arrow
- Date – timeline on bottom
- Temperature – next to date on timeline (today: brush-over tooltip)

Preview:
Small Multiple – Air Pollution Map

© 2001 E. R. Tufte, Yale University
Graphical Integrity [1]: Consistency in Labeling, Baselines

- Operate Revenues
- Net Income (Loss)
- Exploration & Development Expenditures


© 2001 E. R. Tufte, Yale University

Graphical Integrity [2]: Consistency in Time (Independent Axis)

- Commission Payments to Travel Agents

Time frame: one year
Time frame: half year


© 2001 E. R. Tufte, Yale University
Graphical Integrity [3]:
Dangers of Partial Annual Data

Nobel Prizes Awarded in Science, for Selected Countries, 1901-1974
(Number of Prizes)

Nobel Prizes Awarded in Science, for Selected Countries, 1901-1980
(Number of Prizes)

© 2001 E. R. Tufte, Yale University

Lie Factor [1]:
Definition & Example

- **Definition**
  
  \[
  \text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}
  \]

- **Goal:** Keep Close to 1 (e.g., in interval [0.95, 1.05])

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1995, is 5.3 inches long.

\[
\text{Lie Factor} = \frac{5.3 \cdot 0.6}{0.6} = \frac{5.3 \cdot 0.6}{0.6} = 14.8
\]

© 2001 E. R. Tufte, Yale University
Lie Factor [2]: Examples – Excess Dimensions (2- & 3-D)

Lie Factor: ~1 based on height, 2.8 based on area

Graphical Integrity [4]: Need for Data Normalization

Playfair (1786): Nominal Money

Playfair (1786): Real (Standardized, i.e., Inflation-Adjusted aka Deflated) Money
Graphical Integrity [5]: Context – “Compared to What?”

What If…?

Graphical Integrity [6]: Pravda School of Ordinal Graphics

© 2001 E. R. Tufte, Yale University
Graphical Integrity: Tufte’s Six Principles (“DO’s”)

- See also How to Lie With Statistics (Huff, 1984): http://bit.ly/3wAgS0

1. Make Representation of Numbers Proportional to Quantities
   - Ratio of size to numerical value should be close to 1
   - As physically measured on surface of graphic

2. Use Clear, Detailed, Thorough Labeling
   - Don’t introduce or propagate graphical distortion, ambiguity
   - Write out explanations of the data on the graphic itself
   - Label important events in the data

3. Show Data Variation, Not Design Variation

4. Use Standardized (e.g., Inflation-Adjusted) Units, Not Nominal

5. Depict N Data Dimensions with \( \leq N \) Variable Dimensions
   - Don’t use more than \( N \) information-carrying dimensions for \( N \)-D data
   - When graphing data in \( N \)-D, use \( N \)-D ratio (see #1 above)

6. Quote Data in Full Context (Don’t Quote Out of Context)

Data-Ink

Above all else show the data.
- Edward Tufte

- Data-Ink: Non-Erasable Core of A Graphic
  - Non-redundant ink
  - Arranged in response to variation in numbers represented

- Data Density
  - Amount of usable information per unit (space, ink, time, etc.)
  - Want: higher data density as function of resource
  - Example: data-ink ratio

- Non-Data-Ink Can Be Erased to Improve Data Density
Data-Ink Ratio

Data – ink ratio = \frac{\text{data - ink}}{\text{total ink used to print the graphic}}

= \text{proportion of a graphic’s ink devoted to the non-redundant display of data information}

= 1.0 – \text{proportion of a graphic that can be erased without loss of data information}

Low Data-Ink Ratio (Playfair, 1785): Charts with Many Grid Lines & Detailed Labels

Intermediate Data-Ink Ratio (Playfair, 1786): Conventional Charts – Grid Thinned

High Data-Ink Ratio: Electroencephalogram (EEG) Signals

Chartjunk

This chart uses five colors, three dimensions, and two parts to show only five data points!

Redundancy in

Symmetry
Superfluous use of perspective and color

Forget chartjunk, including moiré vibration, the grid, and the duck.

Edward Tufte

© 2001 E. R. Tufte, Yale University


© 1973 H. S. Shyrock & J. S. Siegel

“Duck” here refers to self-promoting decorative graphics.
Statistical Graphics [1]:
Box plots (Spear, 1958; Tukey, 1977)

Mary Eleanor Spear’s “range bar”

and John Tukey’s “box plot”

can be mostly erased without loss of information:

Wikipedia, Box Plot: 
http://bit.ly/AxVBN

Statistical Graphics [2]:
Dot-Dash Scatter Plots (Tufte, 1983)

In a dot-dash plot, sometimes called a Tufte scatterplot, the axes are replaced with marginal distributions (the projection of the bivariate scatter down to one variable)

Wikipedia, Scatter Plot: 

© 2001 E. R. Tufte, Yale University

© 2008 University of Michigan

- Construct Distribution of Variable Using Numbers Themselves
- Construction Algorithm
  - Sort data in ascending order – example:
    44 46 47 49 63 64 66 68 72 75 76 81 84 88 106
  - Designate meaning of stems, leaves
    - Leaves: suffixes, to right of vertical line – usually last digit
    - Stems: prefixes, to left of vertical line – usually all other digits
  - Group data by common stem (prefix) – example:
    - 4 | 4 6 7 9
    - 5 |
    - 6 | 3 4 6 8 8
    - 7 | 2 2 5 6
    - 8 | 4 8
    - 9 |
    - 10 |
    - 6 key: 6|3=63, leaf unit: 1.0, stem unit: 10.0

Statistical Graphics [3]:
Stemplots (Bowley, c. 1900)

Statistical Graphics [4]:
Range-Frame Plot

The min and max values for each variable are made implicit by erasing the axes below the min and above the max value. This is an example of how erasing non-data ink can lead to greater data density and data-ink ratio.
Graphical Excellence – Synopsis [1]

- Well-Designed Presentation of Interesting Data
  - **Substance** – accurate, precise data, labels, other original content
  - **Statistics** – analytical content, summative & descriptive
  - **Design** – non-redundant, concise presentation of multiple variables

- Complex Ideas Communicated with
  - **Clarity** – ease of interpretation and understanding
  - **Precision** – ability to reconstruct original data, process
  - **Efficiency** – rendering time vs. reading & comprehension time

- Gives to Viewer
  - Greatest number of ideas – data
  - In shortest time – “ink ratio” really rate per time (cognitive effort)
  - With least ink – filled space, pixels, primitives, rendered objects
  - In smallest space – total size of graphic, page, viewport, window

Adapted from book material © 2001 E. R. Tufte, Yale University

Graphical Excellence – Synopsis [2]

- Nearly Always Multivariate
- Requires Telling Truth about The Data (Graphical Integrity)

p. 51 (Tufte 1 aka VDQI, 1st & 2nd)

Adapted from book material © 2001 E. R. Tufte, Yale University
Summary

- Reading for Last Class: Chapter 15, Eberly 2e; Ray Tracing Handout
- Reading for Today: Tufte Handout
- Reading for Next Class: Ray Tracing Handout
- Last Time: Ray Tracing 2 of 2
  - Stochastic & distributed RT
  - Hybrid RT (for specular reflectance) & radiosity (for diffuse)
- Today: Visualization Part 1 of 3 – Statistical, Scientific, Data/Info Vis
  - Tufte 1: The Visual Display of Quantitative Information, 2e
  - Graphical integrity
    - Lack: lie factor (“How to lie with statistics visualization”)
    - Desiderata: transparency; labeled axes, clear comparisons
    - “Show variation in data, not presentation”
  - Graphical excellence
    - Lack: chartjunk
    - Desiderata: data-ink, data-ink ratio (& “data-pixels”)

Terminology

- **Visualization**: Using Images, Diagrams, Animations to Communicate
  - **Scientific**: transformation, representation of data for exploration
  - **Statistical / data**: info in schematic form (attributes, variables)
  - **Information**: computational tools; analyzing large, abstract data sets
- **Statistical Visualization Techniques**
  - **Boxplot aka range bar, box-and-whisker diagram**: mean, quartiles
  - **Dot-dash plot aka Tufte scatterplot aka scatter plot with Tufte axes**
  - **Stemplots aka stem-and-leaf display**: prefix (stem), suffix (leaf)
  - **Range-frame plot**: erase axes outside range (min/max x, y)
- **Tufte 1: The Visual Display of Quantitative Information, 2e**
  - **Graphical integrity**: accurate, truthful visual communication
  - Example of lapse in graphical integrity: lie factor (distortion ratio)
  - **Data-ink ratio**: quantity of usableaccessible info per unit of “ink”
  - **Graphical excellence**: high data-ink ratio, no wasted axes
  - Antithesis of graphical excellence: chartjunk (visual clutter)