Big Data in Pictures: Data Visualization

Huamin Qu

Hong Kong University of Science and Technology
What is data visualization?

- “Data visualization is the creation and study of the visual representation of data” - wiki
- Input: data  Output: visual form  Goal: insight
Why visualization?
## Anscombe’s Quartet: Four datasets

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>7.46</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>6.77</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>12.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>7.11</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>7.81</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>8.84</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.08</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>5.39</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>8.15</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>6.42</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.73</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>6.58</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>7.71</td>
<td>8.0</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>8.84</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>7.04</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>5.25</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td>19.0</td>
<td>12.50</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>7.91</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>6.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Anscombe’s quartet: four different datasets.
Anscombe’s Quartet: Statistics

<table>
<thead>
<tr>
<th>Property (in each set)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x</td>
<td>9.0</td>
</tr>
<tr>
<td>Variance of x</td>
<td>10.0</td>
</tr>
<tr>
<td>Mean of y</td>
<td>7.50</td>
</tr>
<tr>
<td>Variance of y</td>
<td>3.75</td>
</tr>
<tr>
<td>Correlation between x and y</td>
<td>0.898</td>
</tr>
<tr>
<td>Linear regression line</td>
<td>$y = 0.5x + 3.0$</td>
</tr>
</tbody>
</table>

Table 1.2: Same statistics in Anscombe’s quartet.
Anscombe’s Quartet: Visualizations
Bible
GapMinder
Data Visualization

• Vis is hot & cool
• Vis is young
• What is vis research?
• What are we doing at HKUST?
Pictures Make Sense of Big Data

Visualization technology can turn data into pictures that are far more comprehensible

By DEBORAH GAGE

Updated Sept. 15, 2013 5:18 p.m. ET

Most people have trouble recalling strings of numbers that are longer than their phone numbers. So how do we begin to comprehend a hundred rows of data, let alone a thousand or a million or a billion rows?
How P&G Presents Data to Decision-Makers

by Tom Davenport | 3:00 PM April 4, 2013
To Go from Big Data to Big Insight, Start with a Visual  
BLOG POSTS | Sinan Aral. Visualization by Nikolaos Hanselmann | Aug 27, 2013

Although data visualization has produced some of the most captivating artistic displays in recent memory, some of which have found...  

2 Things Your Data Visualization Needs  
TIPS | Aug 7, 2013

Data is easier to understand when it’s presented visually. But sometimes designers get wrapped up in the exercise of making something beautiful and forget to make it functional. Here are two thi...  

Get The Picture: Gaining Insights With Data Visualization  
BLOG POSTS | Angelia Herrin | May 27, 2013

As the volume of data flowing into our organization increases, we have a growing need to find ways to identify...  

Don't Read Infographics When You're Feeling Anxious  
BLOG POSTS | Nicholas Diakopoulos | May 23, 2013

Ever had to look at a data visualization while you were in a lousy mood? Chances are you were more...  

How GE Uses Data Visualization to Tell Complex Stories  
BLOG POSTS | Gretchen Gavett | May 9, 2013

GE, perhaps more than any other major company, is dedicated to the use of data visualization as a key part...
Does Your Company Actually Need Data Visualization?

by Bill Shander | 2:00 PM November 14, 2013
Visualization: The Simple Way to Simplify Big Data

BY CHRIS TAYLOR, TIBCO 08.26.13 12:53 PM
What Does Big Data Look Like? Visualization Is Key for Humans

BY DAVID HOFER, DECLARA 01.10.14  4:00 PM
2013: The Year in Interactive Storytelling

Effective tax rate 2007-12

<table>
<thead>
<tr>
<th>Rate</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>&gt; 60%</td>
<td></td>
</tr>
<tr>
<td>N.A.</td>
<td></td>
</tr>
</tbody>
</table>

S&P 500 companies

About one of every seven companies had an effective tax rate lower than 10 percent, including Amazon at 6 percent and Verizon at 9 percent. Nine companies paid no taxes at all.

Each circle represents a company, sized by its market capitalization. The largest is Apple, at more than $400 billion, with an effective tax rate of 14 percent.

Combining earnings and taxes for all S&P 500 companies gives an effective tax rate of 29.1 percent. But rates vary widely by industry.

Three big energy firms paid the most taxes in absolute terms: Exxon $146 billion; Chevron $85 billion; and ConocoPhillips $58 billion.

Effective tax rates cannot be computed for several dozen companies because they lost money over the six-year period. For example, AIG lost $83 billion while paying $6 billion in taxes. These companies are still included in overall tax rate calculations.

Visual and Interactive Features

The Overall Picture | The View by Industry

Find a company or industry...
Obama’s big data plan

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

• Issuing a $2 million award for a research training group to support training for undergraduates to use graphical and visualization techniques for complex data.
Tableau Software Plots Latest Big Data IPO
Make analytics easy.
For analysts, executives, IT, everyone.
新闻联播携手百度"数据说春节"
Data Visualization

• Vis is hot & cool
• Vis is young
• What is vis research?
• What are we doing at HKUST?
Subfields

- Scientific Visualization (SciVis) – Spatial data
- Information Visualization (InfoVis) – Abstract data
- Visual Analytics (VAST) – Analytical reasoning
VIST (Visual Analytics Science and Technology)

InfoVis (Information Visualization)

SciVis (Scientific Visualization)
Welcome!

Vis 2007 and InfoVis 2007 are the premier forums for data and information visualization advances for academia, government, and industry. These events bring together researchers and practitioners with a shared interest in tools, techniques, and technology. The conferences will include an exciting and
Welcome

October 14-19, 2012
The Sheraton Seattle Hotel
1400 Sixth Avenue
Seattle, WA, USA
Paper Submission Guidelines

All conferences at IEEE VIS 2013 use similar submission and review processes. Please read the individual Call for Papers for each of the conferences here: VAST, InfoVis, and SciVis.
Welcome

The Call for Papers for VAST, InfoVis and SciVis are now available. Check them out! Abstracts deadline is Friday, March 21, 2014

IEEE VIS 2014 is the premier forum for advances in visualization. The event-packed week brings together researchers and practitioners from academia, government, and industry to explore their shared interests in tools, techniques, and technology.
VIS - Infographics

• Infographics is static
• Visualization is interactive
StoryLine
VIS - HCI

- Visualization deals with data
- HCI deals with everything involving human & computer interaction
VIS - Data Mining

• Data mining focuses more on automatic algorithms
• Visualization keeps human in the loop and focuses more on interactive analysis
VA and Data Mining

- Data Analysis
  - let computers do what computers are good at

Data mining

- let humans do what they're good at

Visual analytics
Daniel A. Keim’s perspective on visual analytics
Data Visualization

• Vis is hot & cool
• Vis is young
• What is vis research?
• What are we doing at HKUST?
What is vis research?

- Techniques/algorithms
- Applications
- Systems
- Evaluations
- Theory/models
Visualization Process

• Visualization Pipeline

Engineering part          Visual design part
Technique Papers

Novel techniques or algorithms
Application/Design Study Papers

Applying visualization and visual analytics techniques in an application area.
System Papers

A blend of algorithms, technical requirements, user requirements, and design that solves a major problem.
Evaluation Papers

Animations

Traces

Fig. 2. Traces Visualization shows all trace lines simultaneously.
Theory/Model papers

New interpretations of the foundational theory of visualization and visual analytics.

Fig. 5. Barriers in InfoVis Novices' Visual Data Exploration Process. Barriers are indicated with lightning bolts. 1: selection barrier; 2: visual mapping barrier; 3: interpretation barrier.
A Typical Visual Analytics Problem

• IEEE VAST Challenge 2009

An employee is leaking important information to the outside world; hypotheses about his identity and network need to be made or confirmed.

There are three datasets:

– Badge and computer network traffic
– Social network (with a very small geospatial component)
– Video
VAST Challenge 2013

Challenge Description

In this VAST Challenge 2013 Mini-Challenge, we are looking for innovative graphical designs to support situation awareness for large-scale computer networks.

Objective

A company called Big Enterprise has hired you to create an innovative visual design that will become the foundation for their future situation awareness display that shows the state of their entire computer network.
BIG MARKETING
VAST 2013 Mini-Challenge 3

http://www.cs.umd.edu/hcil/varepository/
Role of Visualization in Big Data Analytics

Scalability
Challenges for Visualization Research

Scalability
Better Scalability

- Better visualization
- More people
Better Visualization

• Better overview & summarization
• Better data reduction
• Better visual encoding
• Better user interaction
• ....
Top Ten Interaction Challenges in Extreme-Scale Visual Analytics

P.C. Wong, H.W. Shen and C. Chen 2012

1. In Situ Interactive Analysis
2. User-Driven Data Reduction
3. Scalability and Multi-level Hierarchy
4. Representation of Evidence and Uncertainty
5. Heterogeneous Data Fusion
Top Ten Interaction Challenges in Extreme-Scale Visual Analytics

6. Data Summarization and Triage for Interactive Query
7. Analytics of Temporally Evolving Features
8. The Human Bottleneck
9. Design and Engineering Development
10. The Renaissance of Conventional Wisdom
How to Get More People?

• Lower the bar for visualizations
• Bring visualizations to masses
Lower the Bar for Visualizations

• Google Visualization API
• IBM’s RAVE and Many Eyes
• D3.js

• Taobao’s datav.org
D3.js
Taobao
百度 Echarts

零编程 玩转图表

大数据 彰显魅力

开始制作图表 ➔
Paste in a bunch of text:

OR

Enter the URL of any blog, blog feed, or any other web page that has an Atom or RSS feed.
Data Visualization

- Vis is hot & cool
- Vis is young
- What is vis research?
- What are we doing at HKUST?
Visualization@HKUST
Medical Data Visualization

TVCG'07
TVCG'08 (Vis’08)
TVCG'09 1 (Vis’09)
TVCG’09 2 (Vis’09)
TVCG 2011
High Dimensional and Relational Data

TVCG’08 (InfoVis’08)

tvCG’09
(InfoVis’09)

cGF’08
(EuroVis’08)

tvCG’08
(InfoVis’08)

cGF’09
(EuroVis’09)
Text Data

Dynamic Tag Clouds

OpinionSeer

FacetAtlas

IEEE CG&A’11

TVCG’10
(InfoVis’10)

TVCG’10
(InfoVis’10)

TVCG’11
(InfoVis’11)

ACM TIST’12

TextFlow

TextWheel

(a) Flow/Vector

Tensor/Fiber

Volume/Rendering

Structure/Layout

Exploration/Analytics

(b) Document/Temporal

MCI
Urban Informatics

Air pollution analysis
TVCG’07 (Vis’07)

Trajectory Analysis
VAST’11

Route Zooming
TVCG’09 (Vis’09)

Mobile Phone Data
Three Projects

• Twitter Data Visualization
• Trajectory Data Visualization
• Coursera Data Visualization
Social Flow
Whisper

When and where is a story dispersed?
How information is spread?

Query earthquake on Twitter
Information diffusion: When, where and how the ideas are spread?

The temporal trend, social-spatial extent and community response to a topic of interest.
Challenges

• How can we represent the dynamic information diffusion processes for diffusion pattern detection and comparison?
  – How can we deal with a huge amount of micro blog data in real time?
  – How can we summarize the diffusion processes by considering temporal, spatial, topic and community information?
Visualization design

- Color hues: sentiments
- Color opacity: activeness of tweets / users
- Size: importance (number of followers)
- Shape: types (media / common users / retweeting activity)
Visualization design

• Diffusion series
  – When mouse over on an active tweet, its diffusion series will be shown to illustrate ``when did who reweet whom”.

Layout

Topic Disc

Diffusion Pathway

User Group

Assembling
Sunflower Packing
- **Diffusion Field**
  - Based on electronic fields
  - Topic positive charged
  - User group negative charged

\[
\bar{E}(x,y) = \sum_i E_i = \sum_i k_e \frac{q_i}{r_i^2} \bar{r}
\]
• Flux line layout

\[
\frac{\partial \rho(t)}{\partial t} = \bar{E}(\rho(t)) = \bar{E}(x,y)
\]

\[
\rho(t + \Delta t) = \rho(t) + \int_{t}^{t+\Delta t} \bar{E}(\rho(t))dt
\]
Users are grouped either by geo-locations / their shared interests
  – Use voronoi icons to facilitate comparison
  – Use sunflower packing for efficient layout in case of online monitoring

Layout by longitude or evenly separate circle space
Layout

- Visual clutter reduction
  - Reorder the user groups in match with the center active tweets to reduce crossings
  - Rotate the topic disc to minimize the line lengths
Video
Stories (1) : tracing diffusion events

Japan Earthquake
Temporal diffusion pattern (Medias)
Interview with Domain Experts

• Three experts:
  – political scientist, design expert, communication phD

• Interview methodology:
  – warm up; semi-structure interview with 7 questions; free discussions

• Results

• Discussions
A Complete VA System

• Design goals
• Underlying techniques (data mining; visualization, interaction, etc.)
• Case studies
• User study (questionnaire-based; task-based)
• Expert Interview
Visual Analytics of Trajectory Data

- Fix data errors
- Reconstruct continuous trajectories
- Route diversity analysis

Data Error

Data Uncertainty
Navigation System

• Microsoft T-drive [Yuan et al., 2010]
Encoding Scheme

- **Time**
  - Outer circle
- **Number of trajectories**
  - The height of the bars
- **Duration**
  - Length of the arc
- **Speed**
  - Color

H. Liu et al. IEEE VAST 2011
User Interactions

• Brushing by locations
• Brushing by time span
Traffic Monitoring
Route Suggestion
Route Suggestion
Embed Temporal Displays onto the Roads
Interchange patterns

08:15---08:45

15:15---15:45
Analysis of Clickstreams in the Coursera
Analysis on clickstream pattern-stacked graphs
Analysis on clickstream pattern-stacked graphs

Total number is highest at the beginning; error/stalled often happen at the beginning.

Several Peaks after the beginning part: mainly caused by pause+play (probably difficult/important parts or there are in-video questions: pause to answer questions).
Analysis on clickstream pattern - animation
Analysis on clickstream pattern - animation
Analysis on user behavior
User classification

• Current Classification

- Users who watched only one video – “random” users 28.1%
- Users who watched all videos – “persistent” users 9.8%
- Users who watched more than one video and only those from the first seven videos – “impersistent” users 44.5%
- Other types of users – “others” 17.6%

• Comparisons between different types of users

“seek” action – most frequent and meaningful
  → seek arc comparisons
Blue represents seek operation when users watch the video for the first time, Red represents seek operation when user reviews videos;

The bar chart represents the times users watched on different video periods.

Seek arc visual encoding scheme
Comparison of seek behavior on the first course video for different groups of users (500 users for each group)
Comparison among videos from different weeks for the user group who have watched all the 17 videos (persistent users)
*Specific Parts in Videos Corresponding to Histogram peaks
Some patterns from peaks: 1) “Skip the opening”
Specific Parts in Videos Corresponding to Histogram peaks
Some patterns from peaks: 1) “Skip the opening”
Specific Parts in Videos Corresponding to Histogram peaks

Some patterns from peaks: 2) “New term & Sensitive topic”
*Specific Parts in Videos Corresponding to Histogram peaks
Some patterns from peaks: 2) “New term & Sensitive topic”

In the USA Students from the Top Half of the Income Pool Comprise 80 to 85 Percent of the Elite University Student Population

Source: Pryor, Hurtado, DeAngelo, et al., 2009 with minor changes.
*Specific Parts in Videos Corresponding to Histogram peaks
Some patterns from peaks: 3) “Findings vs. Conclusions”
*Specific Parts in Videos Corresponding to Histogram peaks
Some patterns from peaks: 3) “Findings vs. Conclusions”
Conclusions

• A primer on data visualization
• Examples to showcase HKUST vis projects
• Data visualization is hot, cool, and young
IELM 4320 Design Thinking

A 3-credit, one-month summer course on Design Thinking for engineering students in collaboration with the China Academy of Art (CAA), one of the most influential academies of fine arts in China.

For
All UG & PG students.
Application deadline: 3 March 2014
Interview will be arranged in the early March

Instructors
(CSE) Prof. Huamin Qu
(IELM) Prof. Ravindra Goonetilleke and Prof. Emily Au

Sponsorship
Airfare to Hangzhou and accommodation sponsored by HKUST
Q&A