E6893 Big Data Analytics Lecture 7:

Big Data Visualization — I

Ching-Yung Lin, Ph.D.
Adjunct Professor, Dept. of Electrical Engineering and Computer Science

October 21st, 2022
Outline

Part I : Introduction
  What is visualization?
  Why do we create visualisation?
  Existing Visualisation Techniques

Part II : Big Data Visualization
  Challenges
  Techniques

Part III : How can we visualize big data
  Key techniques
  Open source tools
  Examples

Part IV : Visual Analysis of Big Data

Thanks to Dr. Nan Cao  http://nancao.org
Part I : Introduction
What is visualization?
How can we acquire information?

Listen

Taste & Smell

Touch

Look
Do they effective?

Chemical Signal

Taste & Smell

Bandwidth: ??? kb/s

Physical Signal

Touch

Bandwidth: ??? kb/s

Can not be estimated by information theory
Do they effective?

Sound Signal
Listen

Electronic / Light Signal
Look

Bandwidth: about 0.1 KB / s

> 100 MB/s

“Information Visualization, Perception for Design” 3rd Edition, by Colin Ware
A picture worth 1000 Words
Why efficient?

The Visual Thinking Pipeline

Parallel Processing to Extract Low-Level Visual Properties such as color, shape, etc.

Sequential Goal-Oriented Processing

2 billion neurons
Parallel, automatic
Ching-Yung Lin, Nan Cao, Shixia Liu, Spiros Papadimitriou, Jimeng Sun, and Xifeng Yan. SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence. ICDE 2009

Record of human activities to help find data patterns
Example

Visualization is used for help reasoning and decision making
Example

Summarization of Airlines in United States

What is Information Visualization?

“The action or fact of visualizing; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed.”

-- Oxford English Dictionary

“... finding the artificial memory that best supports our natural means of perception.”

-- Bertin, 1983

The use of computer-supported, interactive, visual representations of abstract data to amplify cognition

-- Cart, Mackinlay, Shneiderman, 1999
Introduction

Why do we create visualization?
Why do we create visualization?

Counting the number of 3s in the following Text:

1235693234870452973467
0378937043679709102539
Counting the number of 3s in the following Text:

1235693234870452973467
0378937043679709102539
Find Patterns

Can you identify the groups of dots in the following figures?

Law of Proximity

we tend to group elements that are closest to each other
Find Patterns: Pre-Attentive Visual Channels

Grey value

Elongation

Curvature

Added surround box

Shape

Added surround color

Filled

Sharpness

Cast shadow

Convex and concave

Sharp vertex

Joined lines

Misalignment

Blinking

Direction of motion

Phase of motion
Why do we create Visualization?

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th></th>
<th>Set B</th>
<th></th>
<th>Set C</th>
<th></th>
<th>Set D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.10</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.10</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.50</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
</tr>
<tr>
<td>mean</td>
<td>9.00</td>
<td>7.50</td>
<td>9.00</td>
<td>7.50</td>
<td>9.00</td>
<td>7.50</td>
<td>9.00</td>
<td>7.50</td>
</tr>
<tr>
<td>std</td>
<td>3.32</td>
<td>2.03</td>
<td>3.32</td>
<td>2.03</td>
<td>3.32</td>
<td>2.03</td>
<td>3.32</td>
<td>2.03</td>
</tr>
<tr>
<td>corr</td>
<td>0.82</td>
<td></td>
<td>0.82</td>
<td></td>
<td>0.82</td>
<td></td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>lin. reg.</td>
<td>y = 3.00 + 0.500x</td>
<td>y = 3.00 + 0.500x</td>
<td>y = 3.00 + 0.500x</td>
<td>y = 3.00 + 0.500x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Seeing data in context
Why do we create visualization?

A picture is worth a thousand words
Some other reasons

See data in context
Find patterns
Telling a story
Attract attentions
Communicate information with others
Summarization and interpretation
Graphical calculation
Expend memory
Inspire people
Existing Visualisation Techniques
Visualization & Visual Analysis Reference Model

- **Raw Data**
- **Abstract Data**
- **Visual Form**
- **View**

**Steps:**
1. **Data Mining**
2. **Encoding and Layout**
3. **Rendering**

**Processes:**
- Filtering
- Rendering
Taxonomy by data types

1D

2D

3D

Multi-D

Temporal

Tree

Graph
Examples: Visualizing 1D Numerical Data

- Line chart with multiple value axes

- Vertical Bar:
  - Year: 2000, Sales per year: 3, 4, 5, 3.8, 4.1, 4.8
  - Year: 2003, Sales per year: 3.4, 4.3, 9.0
  - Year: 2006, Sales per year: 7.4, 7.3
  - Year: 2009, Sales per year: 9.9, 9.0

- Horizontal Bar:
  - Year: 2000, Sales per year: 3.0, 3.8, 4.3, 9.0
  - Year: 2003, Sales per year: 3.4, 4.1, 9.5, 7.4
  - Year: 2006, Sales per year: 7.3, 9.0

- Stacked Vertical Bar:
  - Year: 2000, Sales per year: 6.9, 5.0, 5.5
  - Year: 2001, Sales per year: 3.4, 5.0, 4.7
  - Year: 2002, Sales per year: 3.0, 4.7, 1.9, 5.5

- Stacked Horizontal Bar:
  - Year: 2000, Sales per year: 3.0, 3.4, 1.2, 6.9
  - Year: 2001, Sales per year: 5.0, 5.0, 5.0, 5.0
  - Year: 2002, Sales per year: 3.4, 4.7, 1.9, 5.5
Example: Visualizing 1D Ordinal Data

Examples: 2D Data

Size of each Cell: Stock Market Value
Color: Stock Change
Example: Multi-Dimensional Data

Cylinders \(\downarrow\) Displacement \(\downarrow\) Horsepower \(\downarrow\) Weight \(\downarrow\) Year

MPG

Origin: USA \(\rightarrow\) Japan \(\rightarrow\) Europe
Examples: Visualizing Structured Data
Examples: Visualizing Unstructured Data

Visualization of Text Documents
Examples: Geospatial

Larger cinema markets support stronger domestic film industries.

Countries sized by relative share of worldwide box office revenue, 2009
Examples: Visualizing Spatial Temporal Data

Pulse of the Nation:
U.S. Mood Throughout the Day inferred from Twitter

Less Happy  More Happy

http://www.ccs.neu.edu/home/amislove/twittermood
Examples: Visualizing Spatial Temporal Data

wind map

Dec. 3, 2014
11:35 am EST
(time of forecast download)

top speed: 31.5 mph
average: 8.2 mph

http://hint.fm/wind/
Visualization is not just a beautiful picture

The purpose of visualization is to reveal the insight of the data
InfoVis vs Computer Graphics

Realism

Information
InfoVis v.s. Scientific Visualization

Physical Data

Artificial Data
Three Sub-areas

Scientific Visualization (SCIVIS)

Information Visualization (InfoVis)

Visual Analysis (VA)

Physical Data

Abstract Data

InfoVis

Data Mining

Machine Learning

Traditional

Research Trend

Modern Technique
Major Conferences and Journals

- Scientific Visualization (SCIVIS)
- Information Visualization (InfoVis)
- Visual Analysis (VA)

VisWeek: IEEE SCIVIS, INFOVIS, VAST

- IEEE Transactions on Visualization and Computer Graphics
- Computer Graphics Forum
- IEEE Computer Graphics and Application
Recommended Books

- The Visual Display of Quantitative Information
  - Edward R. Tufte

- Envisioning Information
  - Edward R. Tufte

- Visualization Analysis & Design
  - Tamara Munzner

- Information Visualization: Perception for Design
  - Colin Ware

Illustrations by Eamonn Maguire
Recommend Must-Learn

• SVG
• D3.js
• Bootstrap
• (Of course — HTML, Javascript, and CSS)
Part II: Visualising Big Data
Are you ready to Big Data?
Data, Data, and Data

340 million tweets a day!

4 billion messages a day!
Information Overload
difficult caused by too much information
Challenge

How can we acquire useful information from the overwhelming data
3Vs of Big Data

- **Volume**
  - Terabytes
  - Records
  - Transactions
  - Tables, files

- **Velocity**
  - Batch
  - Near time
  - Real time
  - Streams

- **Variety**
  - Structured
  - Unstructured
  - Semistructured
  - All the above
Big Data Visualization

76425 species
Tree of Life by Dr. Yifan Hu

14.8 million tweets
The information diffusion graph of the death of Osama bin Laden by Gilad Lotan

500 million users
Facebook friendship graph by Paul Butler

Challenging Task:

Squeezing millions and even billions of records into million pixels (1600 X 1200 = 2 million pixels)
Challenges

Visual clutter
How can we avoid visual clutters like overlaps and crossings?

Performance issues
How can we render the huge datasets in real time with rich interactions?

Limited cognition
How can users understand the visual representation when the information is overwhelming?
Techniques (1): Pixel Oriented Visualization

A multidimensional data item contains 6 attributes
Technique(1) : Pixel Oriented Visualization

Database visualization (10,000 items, 6 dimensions)

Order by degree of interests max

(Keim & Kriegel, 1994; 1996) min
Techniques (1): Pixel Oriented Visualization

Database Visualization (10,000 items, 6 dimensions)

Order by attribute similarities
Techniques (1) : Pixel Oriented Visualization

Different Ways for splitting the display region

Values above represent monthly concentrations adjusted to represent 2400 hours on the 15th day of each month. Units are parts per million by volume (ppmv) expressed in the 2008/58 monotonic mixing ratio scale. The "annual average" is the arithmetic mean of the twelve monthly values where no monthly values are missing.
Building a tree for aggregating data items in either a bottom-up or top-down approach.
Technique (2): Aggregation & LOD

- Histogram (Pearson, 1895)
- Heatmap (Wilkinson & Friendly, 2009)
- InfoCube (Stolte et al., 2003)

(Lin et al., 2010)
Techniques (2) : Aggregation

Scatter Plots

Parallel Coordinates

Star Plots

(Elmqvist & Fekete, 2010)

(Yang et al., 2003b)

(Kosara et al., 2006)

(Fua et al. 1999)

(Fua et al. 1999)
Technique (3): Distortion

The diagram illustrates a visual representation of data, with different techniques applied to demonstrate distortion. The visual data is overlaid on a transportation map, showing how data visualization can be applied to real-world scenarios.
Techniques (3) : Distortion

(a) Original Graph and Significance Map

(b) Resizing by Uniform Scaling

(c) Resizing with Significance-aware Grid

(d) Resizing with Adaptive Grid

(e) Resizing by Uniform Scaling

(f) Resizing with Significance-aware Grid

(g) Resizing with Adaptive Grid
Technique (4): Clutter Reduction

- **Sampling**
- **Reordering**
Technique (4): Clutter Reduction

Edge Bundling
Technique (4): Clutter Reduction
Technique (4): Query based Visualization

- Abstract Data
- Filtering
- Visual Form
- Rendering
- View
- User
- Encoding and Layout
- Display
- Data Query
- Rest API(s)
- Query Engine
- Database
A Survey of GPU-Based Large-Scale Volume Visualization, EuroVis, 2014
Part III:
How can we visualize big data?
Visualization & Visual Analysis Reference Model

Encoding : Visual Design
Technique : Layout Algorithm
Using existing tools are easy

D3.js

Data-Driven Documents

Tableau

ManyEyes
Open Source Tools

Python:
  iGraph:  http://igraph.org/redirect.html
  Networkx:  https://networkx.github.io/

JavaScript:
  D3.js (2D, SVG):  http://d3js.org/
  Tree.js (3D, WebGL):  http://threejs.org/

Java:
  prefuse:  http://prefuse.org/
  InofVis Toolkit:  http://ivtk.sourceforge.net/
Developing new ones require knowledge from different areas:

- Computer Graphics
- Data Mining
- Human Computer Interaction
- Industrial Design
- Information Visualization
Example 1: Visualising Streaming Data

Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time
IEEE InfoVis 2012
PART III: How can we visualize big data?

Example 2: Visualizing Large Text Corpus

- ContexTour
  SDM 2010
  Visualizing Heterogeneous Clusters

- FacetAtlas
  TVCG (InfoVis 2010)
  Visualizing Multi-relational Clusters

- SolarMap
  ICDM 2011
  Cluster Interpretation
Part IV: Visual Analysis of Big Data
Visual Analysis v.s. Data Mining

Computational Power

Data Mining

+ 

Human Intelligence

Visual Analysis
Visualization & Visual Analysis Reference Model

Analysis + Visualisation + Interaction
Example 3: Detect Anomalous Users in Twitter

Example 4: Visualizing Large Graphs

g-Miner: Interactive Visual Group Mining on Multivariate Graphs, ACM CHI 2015
Reading Assignment:

The Top 10 Challenges in Extreme-Scale Visual Analytics

Pak Chung Wong  
Pacific Northwest National Laboratory

Han-Wei Shen  
Ohio State University

Christopher R. Johnson  
University of Utah

Chaomei Chen  
Drexel University

Robert B. Ross  
Argonne National Laboratory

Thank You!

Thanks to Dr. Nan Cao  [http://nancao.org](http://nancao.org)
Homework #3 (Due 11/04/2022, 5pm)

Please see the assignment slides
Final Project Proposal Scoring

Proposal  — preparing about 5 pages (each item 1/5 of the proposal score):
  Goal  — novel? challenging?
  Data — 3Vs? New dataset? Existing dataset?
  Methods — planning of methodologies and algorithms? Feasible?
  System — an overview of system. What will be implemented?
  Schedule  — what to achieve by what time, and by whom?
Final Project Report Scoring

- **Title, Author(s)**
- **Abstract**: Briefly describe your problem, approach, and key results.
- **Introduction (5%)**: Describe and define the problem you are working on. Why is it important? Include an overview of your methods and results.
- **Related Work (5%)**: Discuss published works or approaches that are related to your project. What’s the benefit or drawback of the previous works? What kind of problems have they solved? How is your approach similar or different from others?
- **Data (10%)**: Describe the data you are working with for your project. What type of data is it? Where did it come from? How much data are you working with? Did you have to do any preprocessing, filtering, feature engineering, or other special treatment to use this data in your project?
- **Methods (25%)**: Discuss your approach for solving the problems that you set up in the introduction. Why is your approach the right thing to do? Did you consider alternative approaches? Have your tried some methods that didn’t work out? It may be helpful to include figures, diagrams, or tables to describe your method or compare it with other methods.
- **Experiments (20%)**: Discuss the experiments that you performed to demonstrate your approach solves the problem. The experiments will vary depending on the project, but you might compare with previously methods, determine the impact of the components of your system, experiment with different hyper-parameters, architectures, or algorithms, use visualization techniques to gain insight of how your model works, etc. Graphs, tables, and figures are highly recommended to be included to illustrate your experimental results.
- **System Overview (25%)**: Describe the software architecture and tech stacks of your application. Discuss potential bottlenecks and improvements that could be made. Mention the software packages that you used. Mention how to use your application. You could provide screenshots to your application.
- **Conclusion (5%)**: Summarize your key results. What have you learned? What problems have you discovered and solved? Suggest ideas for future extensions or new applications.
- **Writing / Formatting (5%)**: Is your paper clearly written and nicely formatted?