E6893 Big Data Analytics Lecture 8:

Big Data Visualization — I

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

October 25th, 2019
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**
- **Touch**
- **Look**
- **Taste & Smell**
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
  - Bandwidth: *about 0.1 KB / s*

- **Electronic / Light Signal**
  - Look
  - Bandwidth: > *100 MB/s*

---

“Information Visualization, Perception for Design” 3rd Edition, by Colin Ware
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
Example
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
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
Why do we create visualization?
Why do we create visualization?

Counting the number of 3s in the following Text:

1235693234870452973467
0378937043679709102539
Find Patterns

Counting the number of 3s in the following Text:

1235693234870452973467
0378937043679709102539
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>Set B</th>
<th>Set C</th>
<th>Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
<tr>
<td>mean</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>
</tr>
<tr>
<td>corr</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</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>
</tr>
</tbody>
</table>
Seeing data in context

Four scatter plots are shown, each with a linear trend line. The plots are labeled as follows:

- **Y1** vs. **X1**
- **Y2** vs. **X2**
- **Y3** vs. **X3**
- **Y4** vs. **X4**

Each plot contains scattered data points, with **X** values ranging from 4 to 18 and **Y** values ranging from 4 to 12.
Why do we create visualization?

A picture is worth a thousand words

A better communication method
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

Quality

Raw Data → Abstract Data → Visual Form → View

Data Mining

Encoding and Layout

Rendering
Taxonomy by data types

1D

Temporal

2D

Tree

3D

Graph

Multi-D
Examples: Visualizing 1D Numerical Data

- Line chart with multiple value axes
- Pie chart
- Vertical bar chart
- Horizontal bar chart
- Stacked vertical bar chart
- Stacked horizontal bar chart

- Financial chart

© 2019 CY Lin, Columbia University
Example: Visualizing 1D Ordinal Data

Examples: 2D Data

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

Cylinders \downarrow \text{Displacement} \downarrow \text{Horsepower} \downarrow \text{Weight} \downarrow \text{Acceleration} \downarrow \text{Year}

MPG

Origin: USA, Japan, 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

Map showing the market share of domestic filmmakers with different colors representing different percentage ranges.
Examples: Visualizing Spatial Temporal Data

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

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* by Edward R. Tufte
- *Envisioning Information* by Edward R. Tufte
- *Information Visualization: Perception for Design* by Colin Ware
- *Visualization Analysis & Design* by Tamara Munzner
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

3 Vs of Big Data
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?
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 500-meter manometric friction 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**
  - (Elmqvist & Fekete, 2010)
  - (Yang et al., 2003b)

- **Parallel Coordinates**
  - (Kosara et al., 2006)
  - (Fua et al. 1999)

- **Star Plots**
  - (Fua et al. 1999)
Technique (3) : Distortion
Techniques (3) : Distortion

Original Graph and Significance Map

Resizing by Uniform Scaling

Resizing with Significance-aware Grid

Resizing with Adaptive Grid

(a) Resizing by Uniform Scaling

(b) Resizing with Significance-aware Grid

(c) Resizing with Adaptive Grid

(d) Resizing with Adaptive Grid

(e) Resizing with Adaptive Grid

(f) Resizing with Adaptive 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

Diagram:

- Abstract Data filtering
- Visual Form rendering
- View
- Data
- 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?
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
PART III: How can we visualize big data?

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

FacetAtlas
TVCG (InfoVis 2010)

SolarMap
ICDM 2011

Visualizing Heterogeneous Clusters

Visualizing Multi-relational Clusters

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:

Thank You!

Thanks to Dr. Nan Cao  http://nancao.org
Homework #4 (Due 11/15/2019, 5pm)

Please see the assignment slides