Visual Analytics - Introduction

Eduard Gröller

Institute of Computer Graphics and Algorithms
Vienna University of Technology
Goals of VA  [VisMaster, 2010]

Creation of tools and techniques to enable people to:

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, and understandable assessments
- Communicate these assessment effectively for action
What is Visual Analytics?

“Visual Analytics is the science of analytical reasoning supported by a highly interactive visual interface.” [Wong and Thomas 2004]

“Visual Analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets” [Keim 2010]

Detect the expected and discover the unexpected
Visual Analytics Process

First step: preprocess and transform data
  Data cleaning, normalization, grouping, data fusion

Automated methods
  + Scale well
  - Get stuck in local optima
  - Run in a black box fashion

Visualization
  + Interactive data analysis
  - Scalability

Visual Analytics integrates both
  Tied together by the user
  Alternating between visual and automatic methods

[Keim 2006]
Interdisciplinary!
Challenges

Data
Dealing with very large, diverse, variable quality datasets

Users
Meeting the needs of the users

Design
Assisting designers of visual analytic systems

Technology
Providing the necessary infrastructure
Data Mining Definition

Automatic algorithmic extraction of valuable information from raw data

Diagram:
- Data → Mining → Computational Model → Interpretation and Verification → Hypotheses
Knowledge Discovery and Data Mining (KDD)

Semi or fully automated analysis of massive data sets

Contributions are more about general methodologies

Black-box methods in the hands of end users
   Users need to understand the algorithms for using them
   What attributes to use? What similarity measure? etc.
   Often trial and error
The Ability Matrix

Performance of a Human

Performance of a Computer

Data Storage
Numerical Calculations
Searching/Finding
Logic
Planning
Diagnosis
Prediction
Cognition
Common Knowledge
Creativity

Insight is generated by the human – not the computer!

adapted from Daniel Keim, Uni. Konstanz
Why Graphics?

Figures are richer; provide more information with less clutter and in less space.

Figures provide the 'gestalt' effect: they give an overview; make structure more visible.

Figures are more accessible, easier to understand, faster to grasp, more comprehensible, more memorable, more fun, and less formal.

list adapted from: [Stasko et al. 1998]
Statistics vs. Visualization:
Anscombe’s Quartet

<table>
<thead>
<tr>
<th>Anscombe's quartet</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td></td>
<td></td>
<td></td>
<td></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>
Statistics vs. Visualization: Anscombe’s Quartet

Statistics profile is the same for all!

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x in each case</td>
<td>9 (exact)</td>
</tr>
<tr>
<td>Variance of x in each case</td>
<td>11 (exact)</td>
</tr>
<tr>
<td>Mean of y in each case</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td>Variance of y in each case</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td>Correlation between x and y in each case</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>
Anscombe’s Quartet

Four datasets that have identical simple statistical properties, yet appear very different when graphed.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x in each case</td>
<td>9 (exact)</td>
</tr>
<tr>
<td>Variance of x in each case</td>
<td>11 (exact)</td>
</tr>
<tr>
<td>Mean of y in each case</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td>Variance of y in each case</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td>Correlation between x and y in each case</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>
Visualization Can Be Biased

The same data plotted with different scales is perceived dramatically differently.

(a) Equally (uniformly) large scale in both $x$ and $y$

(b) Large scale in $x$

(c) Large scale in $y$

(d) Scale determined by range of $x$- and $y$-values.

[Ward, Grinstein, Keim 2011]
Diagram vs. Visualization

A **diagram** represents **information**.
A **visualization** represents **data**.

The DIKW-Hierarchy
according to [Bellinger 2004]

From Epistemology:
Data = ground truth
Information = phenomena
Knowledge = causes
Wisdom = possible (inter-)actions

"Above all else, show the data"
- Edward R. Tufte
Mantras

Guide to visually explore data - Shneiderman’s Mantra:
Overview first, zoom/filter, details on demand

[Shneiderman, 1996]

Describes how data should be presented on screen

For massive datasets it is difficult to create overview without losing interesting patterns

Extended Mantra for VA:
Analyze first, show the important, zoom/filter, analyze further, details on demand

[Keim, 2006]
Traditional Data Mining vs. Visual Analysis Processes
KDD Pipeline

Visualization Pipeline

[Fayyad 1996]

[dos Santos and Brodlie 2004]
Uncertainty

- What is not surrounded by uncertainty cannot be the truth  [Richard Feynman]

- True genius resides in the capacity for evaluation of uncertain, hazardous, and conflicting information  [Winston Churchill]

- Doubt is not a pleasant condition, but certainty is absurd  [Voltaire]
Uncertainty

Definition

“Degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations with caution” [Hunter 1993]

Measurement data

- e.g., DNA microarray expression data

Can be handled in data or view space
Data Management Challenges

“Big Data”
Uncertainty
Semantics Management
Data Streaming
Distributed and Collaborative VA
VA for the Masses
What is “Big Data”? 

Moving target

Fields dealing with this kind of data:

- Meteorology
- Genomics
- Connectomics
- Complex physics simulations
- Biological and environmental research
- Business intelligence

<table>
<thead>
<tr>
<th>SI decimal prefixes</th>
<th>Binary usage</th>
<th>IEC binary prefixes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name (Symbol)</td>
<td>Value</td>
<td>Name (Symbol)</td>
</tr>
<tr>
<td>kilobyte (kB)</td>
<td>$10^3$</td>
<td>2$^{10}$</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$10^6$</td>
<td>2$^{20}$</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$10^9$</td>
<td>2$^{30}$</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$10^{12}$</td>
<td>2$^{40}$</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$10^{15}$</td>
<td>2$^{50}$</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$10^{18}$</td>
<td>2$^{60}$</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$10^{21}$</td>
<td>2$^{70}$</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$10^{24}$</td>
<td>2$^{80}$</td>
</tr>
</tbody>
</table>

See also: Multiples of bits · Orders of magnitude of data

Visual Steering to Support Decision Making in Visdom

Jürgen Waser


http://www.visdom.at/
Flood emergency assistance

- New Orleans 2005: 17th canal levee breach

Image courtesy of USACE, US Army Corps of Engineers
Flood emergency assistance

- Testing sandbag configurations in a virtual environment
Solution: World Lines

World Lines

Simulations

Control

Time
Solution: World Lines

World Lines

Control

Knowledge

Simulations

Time

Multi-View

System
Worldlines – Multiple Linked Views
Worldlines – Multiple Linked Views
SimVis: Interactive Visual Analysis of Large & Complex Simulation Data

Dr. Helmut Doleisch
VRVis Research Center

http://www.VRVis.at/
Motivation

- large data sets from simulation
- **goal**: support **exploration** and **analysis** of results
  - analyze n-dim. data **interactively**
  - use **3D visualization**
  - overview, zoom and filter, detail on demand (Shneiderman’s information seeking mantra)
- **challenge**: occlusion, interactive data handling
Interactive Data Handling

sample data set size:

- 540 million data items
- currently working to expand to billions

<table>
<thead>
<tr>
<th>cells</th>
<th>timesteps</th>
<th>attributes</th>
<th>cells * timesteps</th>
<th>cells * timesteps * attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>704.900</td>
<td>20</td>
<td>16</td>
<td>14.098.000</td>
<td>225.568.000</td>
</tr>
<tr>
<td>150.124</td>
<td>600</td>
<td>6</td>
<td>90.074.400</td>
<td>540.446.400</td>
</tr>
<tr>
<td>7.680.000</td>
<td>288</td>
<td>15</td>
<td>2.211.840.000</td>
<td>33.177.600.000</td>
</tr>
</tbody>
</table>

Helmut Doleisch
http://www.simvis.at/
SimVis: Interactive Visual Analysis of Large & Complex Simulation Data
SimVis

- VRVis’ solution for these challenges
- Feature-based visualization framework

SimVis key features:
- Multiple, linked views
- Interactive feature specification
- Focus+Context visualization
- Smooth feature boundaries
- Explicit feature representation
- On-the-fly attribute derivation
SimVis: Multiple Views

- Scatterplots, histogram, 3D(4D) view, etc.

an attribute

another attribute

3D + time + color + opacity

an attribute

cell count

an attribute
Brushing

- Move/alter/extend brush interactively
- Update linked F+C views in real-time
VAICo: Visual Analysis for Image Comparison

Johanna Schmidt\textsuperscript{1}, M. Eduard Gröller\textsuperscript{1}, Stefan Bruckner\textsuperscript{2}

\textsuperscript{1}Vienna University of Technology, Austria
\textsuperscript{2}University of Bergen, Norway
VAICo

Visual Analysis for Image Comparison

Johanna Schmidt
Vienna University of Technology, Austria

M. Eduard Gröller
Vienna University of Technology, Austria

Stefan Bruckner
University of Bergen, Norway
YMCA - Your Mesh Comparison Application

[Image of a 3D model of a gargoyle with a mesh overlay showing different areas highlighted in different colors.]

[Diagrams showing algorithm comparisons labeled Alg. 01, Alg. 02, and Alg. 03.]

[Johanna Schmidt et al.]
Literature on Visual Analytics


Literature on Visualization

ISBN: 1568814739

ISBN: 3540649441

ISBN: 9781568813066
Literature on Information Visualization


Acknowledgements

- For material for this lecture unit
  - Marc Streit, Johannes Kepler University Linz
Praktika, Bachelorarbeiten, Diplomarbeiten

http://www.cg.tuwien.ac.at/courses/projekte/
Christmas Tree Case Study