Visual Data Analysis of Unstructured and Big Data
João Comba
Outline

- Basic Visual Data Analysis Concepts
- 12 Examples of Interesting Problems
  - High-level problem description
    - Data
    - Important questions to answer
    - Video illustrating solution
Data is everywhere

- **Social**: every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.  
  source: [http://www.internetlivestats.com](http://www.internetlivestats.com)

- **Urban**: NYC Taxi, Metro and Bike data, San Francisco open data, etc.

- **Sensor**: various types
  - fitness (Garmin, Nike, Polar, Fitbit, GPS trackers, etc)
  - health (Withings, phone apps, etc)

- **Scientific**: simulation, medicine, etc
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools.
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools

Diagram:
- Data
  - Transformation
  - Data Mining
  - Parameter Refinement
- Knowledge
  - Models
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools

Diagram:
- Data
- Visualization
- Knowledge
- Models
- Transformation
- Mapping
- Data Mining
- Parameter Refinement
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining, and analytical tools.
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools.
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools

Visualization Reveals Data

- show the data
- induce the viewer to think about the substance rather about methodology, graphic design …
- avoid distorting what the data have to say
- present many numbers in small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from broad view to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set

Examples of Interesting Problems

- Published Work

1. Visualization of Running Races
2. Environmental Phenological Analysis (Biology)
3. Traffic Analysis in Urban Environments
4. Visual Analysis of Bike Sharing Systems
5. Real-Time Visual Exploration of Big Data
6. Visual Exploration of Software Repositories
Published Work
Visualizing Running Races Through the Multivariate Time-Series of Multiple Runners

Guilherme Oliveira, João Comba, Rafael Torchelsen, Claudio Silva, Maristela Padilha

GPS and Heart-Rate Monitors
GPS and Heart-Rate Monitors

<table>
<thead>
<tr>
<th>Activity Name</th>
<th>Activity Type</th>
<th>Course</th>
<th>Start</th>
<th>Time</th>
<th>Distance</th>
<th>Elevation Gain</th>
<th>Avg Speed(Avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 26 Sep 16:19</td>
<td>1:44:19</td>
<td>12.09</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>Running</td>
<td>Running</td>
<td>--</td>
<td>Tue, 20 Sep 17:10</td>
<td>23:50</td>
<td>0.04</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 12 Sep 18:02</td>
<td>1:11:51</td>
<td>6.58</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Sun, 11 Sep 16:55</td>
<td>1:20:15</td>
<td>5.05</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Fri, 9 Sep 17:39</td>
<td>1:24:49</td>
<td>8.04</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 5 Sep 18:13</td>
<td>5:35</td>
<td>5.84</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Fri, 2 Sep 17:35</td>
<td>1:35:16</td>
<td>12.01</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 29 Aug 18:16</td>
<td>1:18:28</td>
<td>10.03</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 22 Aug 18:08</td>
<td>1:29:16</td>
<td>11.13</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Fri, 12 Aug 16:13</td>
<td>1:07:34</td>
<td>5.92</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Wed, 10 Aug 18:17</td>
<td>1:35:48</td>
<td>12.01</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 8 Aug 19:33</td>
<td>7:30</td>
<td>0.82</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Porto Alegre Running</td>
<td>Running</td>
<td>--</td>
<td>Mon, 8 Aug 18:11</td>
<td>1:20:18</td>
<td>8.52</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Vila Velha Running</td>
<td>Running</td>
<td>--</td>
<td>Thu, 28 Jul 17:33</td>
<td>1:24:35</td>
<td>10.01</td>
<td>143</td>
<td></td>
</tr>
</tbody>
</table>
GPS and Heart-Rate Monitors
PhenoVis – A Tool for Visual Phenological Analysis of Digital Camera Images Using Chronological Percentage Maps
Guilherme Oliveira, Lucas Schnorr, Jurandy Almeida, Bruna Alberton, Leonor Patricia Morellato, Ricardo Torres, João Comba
Information Sciences 372 (2016) 181–195
Phenology

- Study of *periodic* plant and animal life cycle events and how these are influenced by seasonal and inter-annual variations in *climate*
DATA: Continuous, Long Term and Multi-ecosystem

Phenological Eyes Network (PEN)
http://pen.agbi.tsukuba.ac.jp/index.html
Phenological phases

- Average used to evaluate phenological variations in a year

\[
\%r = \frac{r}{r+g+b} \\
\%g = \frac{g}{r+g+b} \\
\%b = \frac{b}{r+g+b}
\]

Time: 1 year
The problem with the average

%g_{cc} = \frac{G}{(R+G+B)} = 0.3905
Percentage Maps

Given a pre-defined color map

Scale it proportionally to the percentage distribution

Percentage Map

Compute for each pixel a value (e.g. gcc) and generate a percentage distribution.
Chronological Percentage Maps (CPMs)

Stacks of percentage maps in chronological order

030
085
117
128
137
142
187
284
288
300
336
349

Figure 4: Percentage maps for 12 days: original data for the respective days (row 1 and 4), recoloring of pixels using $gcc$ and a brown-green-blue-purple color mapping (rows 2 and 5), and recoloring using the hue index and hue mapping (rows 3 and 6). The percentage maps for the $gcc$ (left) and hue (right) indexes are shown in the last row.
Figure 4: Percentage maps for 12 days: original data for the respective days (row 1 and 4), recoloring of pixels using \( g \)cc and a brown-green-blue-purple color mapping (rows 2 and 5), and recoloring using the hue index and hue mapping (rows 3 and 6). The percentage maps for the \( g \)cc (left) and hue (right) indexes are shown in the last row.
Chronological Percentage Maps (CPMs)

Stacks of percentage maps in chronological order

[Image of stacked percentage maps]

Legend:
- Orange: 1
- Red: 2
- Green: 3
- Blue: 4
- Yellow: 5
- Pink: 6
- Brown: 7
- Purple: 8
- Gray: 9
- Black: 10
- Light Blue: 11
- White: 12
Using Probabilistic Vector-Valued Functions to Explore Traffic Dynamics in Urban Environments

Jorge Poco, Harish Doraiswamy, Huy Vo, João Comba, Juliana Freire, Cláudio Silva

Computer Graphics Forum, Volume 34 (2015), Number 3
Visualizing the Dynamics of Bike-Sharing Systems

Guilherme Oliveira, Jose L. Sotomayor, Rafael Torchelsen, João Comba

Computers&Graphics, 2016 (available online)
Video

- Bike-Sharing Systems
Video

- Bike-Sharing Systems

  Full

  [Image of a full bike-sharing station with many bicycles available]

  Empty

  [Image of an empty bike-sharing station with few bicycles available]
Video

- Bike-Sharing Systems

  Full

  Empty

  Rebalancing
Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data

Cicero Pahins, Sean, Carlos Scheidegger, João Comba

Hashedcubes Query

/region/0/0/0/0/1/1
Spatial Constraint

/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23

/where/day_of_week=0|1|3|4
Hashedcubes Query

• `/brightkite`
  • Dataset

• `/region`
  • Query type

• `/region/0/0/0/0/1/1`
  • Spatial Constraint

• `/where/day_of_week=0|1|3|4`
  • Categorical Constraint

• `/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23`
  • Categorical Constraint
Heatmap

- No constraint

- /region/0/0/0/0/1/1
  - Spatial Constraint

- /where/day_of_week=0|1|3|4
  - Categorical Constraint

- /where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23
  - Categorical Constraint
Histogram

- No constraint

- /region/0/0/0/0/1/1
  - Spatial Constraint

- /where/day_of_week=0|1|3|4
  - Categorical Constraint

- /where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23
  - Categorical Constraint
Histogram

- No constraint

- /region/0/0/0/0/1/1
  - Spatial Constraint

- /where/day_of_week=0|1|3|4
  - Categorical Constraint

- /where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23
  - Categorical Constraint
Line Chart

- No constraint

- /region/0/0/0/1/1
  - Spatial Constraint

- /where/day_of_week=0|1|3|4
  - Categorical Constraint

- /where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23
  - Categorical Constraint
Pivot Concept and Hierarchy
Pivot Concept and Hierarchy

0 1 2 3 4 5 6 7 8
3 3 2 3 2 2 3 2 1 2 3 2 2 2 2 2 2 1 1 3 2 1 2 2 1 2 1

6 7 8 3 4 5 0 1 2
1 3 2 1 2 2 1 2 1 2 3 2 2 2 2 2 2 1 3 3 2 3 2 2 3 2 1

[0-2] [3-5] [6-8]
Pivot Concept and Hierarchy

João Comba
Pivot Concept and Hierarchy

1 2 3 4 5 6 7 8
3 3 2 3 2 2 2 2 2 2 2 2 1 1 3 2 1 2 2 1 2 1
6 7 8 3 4 5 0 1 2
1 3 2 1 2 2 1 2 1 2 3 2 2 2 2 2 2 2 1 3 3 2 3 2 2 3 2 1

[0-2] [3-5] [6-8]

7 8 6 4 5 3 1 2 0
1 2 2 1 2 1 1 3 2 2 2 2 2 2 2 2 1 2 3 2 3 2 2 3 2 1 3 3 2

[0-1] [2-2] [3-4] [5-5] [6-7] [8-8]

8 7 6 5 4 3 2 1 0
1 2 1 1 2 2 1 3 2 2 2 2 1 2 2 2 2 3 2 3 2 1 3 2 2 3 3 2

[0-0] [1-2] [3-3] [4-5] [6-6] [7-8]
Pivot Concept and Hierarchy
Metric Evolution Maps: Multidimensional Attribute-driven Exploration of Software Repositories

Renato Silva, Eduardo Vernier, P. Rauber, João Comba, Rosane Minghim, Alexandru Telea
VMV 2016
Multidimensional Projection
Group classes with similar quality metrics in the same visual neighborhood
Conclusions

- Visual Data analysis has several interesting problems
  - Data Mining Algorithms
  - Visualization Techniques
  - High-Performance Computing
  - Spatial Data Structures and Geometric Algorithms
  - Machine Learning Algorithms
  - Mathematical and Statistical Analysis
  - Software implementation (web, prototypes)
Questions ?

**Short Bio**

I am an associate professor in the Computer Graphics Group at the "Instituto de Informática" of the Federal University at Rio Grande do Sul (UFRGS), Brazil. I received a Ph.D. in Computer Science from Stanford University under the supervision of Leonidas J. Guibas. Before that, I received a masters degree in Computer Science from the Federal University of Rio de Janeiro (UFRJ), Brazil, working with Ronaldo Marinho Persiano. My bachelor’s degree in Computer Science was given by the Federal University of Rio Grande do Sul, Brazil.

**Publications by type**

- Journals
- Conferences
- Books
- Book Chapters
- Patents
- Technical Reports
- Conferences/TRs (Portuguese)

**Publications by topic**

- Visualization
- Geometric Algorithms/Data Structures
- High Performance and Parallel Computing
- Graphics Hardware and Games
- Supervised Ph.D. and M.Sc.
- Ph.D. Advisor
- M.Sc. Advisor

**Publications by year**

- 2016
- 2015
- 2014
- 2013
- 2012
- 2011
- 2010
- 2009
- 2008
- 2007
- 2006
- 2005
- 2004
- 2003
- 2002
- 2001
- 2000
- 1999
- 1998
- 1997
- 1996
- 1995
- 1994
- 1993
- 1992
- 1991
- 1990
- 1989
- 1988
- 1987
- All Years

**CV and other statistics**

- Curriculum Lattes (CNPq-Brazil)
- Google Scholar
- DBLP

**Funding**

- CNPq

**International Collaborations**

- USA: CNPq-NSE
- Germany: CAPES-DAAD
- Netherlands: CAPES-BRANETEC

**Professional Activities**

- Conference Chair
- Program Committee
- Organizing Committee
- Co-Chair of CISE VizCorner

**e-mail:** joao.comba@gmail.com

**web:** http://www.inf.ufrgs.br/~comba