数据可视化与可解释性人工智能 (I)

Huamin Qu (屈华民)
HKUST (香港科技大学)
Hong Kong University of Science and Technology
HKUST VisLab

• Awards
  • 11 best paper/honorable mention awards
  • Distinguished Collaborator Award 2016 from Huawei Noah's Ark Lab
  • IBM Faculty Award 2009
  • 2014 Higher Education Scientific and Technological Progress Award presented by the Ministry of Education of China
  • HKICT Award and APICTA Award; IEEE VAST Challenge Award
  • Yelp Dataset Challenge Round 10 Grand Prize Award

• Research
  • Top 5 visualization group in the world based on the output in the top journal of the field
  • The largest visualization group in Asia and one of the largest in the world (30 members including 23 Ph.D)
  • Technologies adopted by Microsoft, IBM, Huawei, Tencent, Bosch, etc.

• Alumni
  • More than 70 (15 PhDs+17 Mphils+…)
  • Working in industry: Microsoft Research Asia, IBM Watson, Bosch Research USA, Siemens, Google, Airbnb, Visa Research, etc.
  • Working in academia: Zhejiang University, Tongji University, University of Electronic Science and Technology of China, Shenzhen University
  • Working in government: Office of the Government Chief Information Officer (OGCIO) Hong Kong
Urban Informatics/Smart Cities, Social Media, Text Analytics, Explainable AI, E-Learning, Social Network, AR/VR
什么是数据可视化？

• Input: data  Output: visual form  Goal: insight
Anscombe’s Quartet: Four datasets

<table>
<thead>
<tr>
<th>Anscomb's quartet</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</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>

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

- **$Y_1$** vs. **$X_1$**: Linear relationship.
- **$Y_2$** vs. **$X_2$**: Non-linear relationship.
- **$Y_3$** vs. **$X_3$**: Non-linear relationship with an outlier.
- **$Y_4$** vs. **$X_4$**: Linear relationship with a different scaling of the axes.
数据可视化简史

Original map by John Snow showing the clusters of cholera cases (indicated by stacked rectangles) in the London epidemic of 1854. The contaminated pump is located at the intersection of Broad Street and Cambridge Street (now Lexington Street), running into Little Windmill Street.
Subfields of Data Visualization

- Scientific Visualization (SciVis) – Spatial data

- Information Visualization (InfoVis) – Abstract data

- Visual Analytics (VAST) – Analytical reasoning
奥巴马的大数据计划

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.
Strategy 2: Developing Effective Methods for AI-human Collaboration

Better visualization and user interfaces are additional areas that need much greater development to help humans understand large-volume modern datasets and information coming from a variety of sources.
数据可视化的流程：90年代

Engineering part  Art part
数据可视化的流程：00年代

Courtesy of Denial Keim et al.
数据可视化与人工智能

Figure 1: XAI Concept
Figure 2: XAI Emphasis
GapMinder
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
Make analytics easy.
For analysts, executives, IT, everyone.
新闻联播携手百度"数据说春节"
可视化标准

- 信: Accuracy
- 达: Intuitiveness, Efficiency, and Effectiveness
- 雅: Aesthetic
The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented.

“Lie Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}

where

\text{size of effect} = \frac{|\text{second value} - \text{first value}|}{\text{first value}}

[Tufte, 1991]
Cognitive Science

• Gestalt law
• Popout effect
• Visual encoding principles:
  – Effectiveness principles
  – Expressiveness principles
Which one is more effective?
**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**

- Spatial region
- Color hue
- Motion
- Shape

- **effectiveness principle**
  - encode most important attributes with highest ranked channels

- **expressiveness principle**
  - match channel and data characteristics
可视化的标准

• 信：Accuracy

• 达：Intuitiveness, Efficiency, and Effectiveness

• 雅：Aesthetic
INFOLITICALGRAPHIC

Data visualization is a popular new way of sharing research. Here is a look at some of the visual devices, informational elements, and general trends found in the modern day infographic.

**DESIGN**

**Chart Style**
Percentage of infographics with the following charts:

- Pie Chart: 22%
- Pictorial Chart: 24%
- Line Chart: 24%
- Bar Chart: 32%

**Font**
- Sans Serif: 85%
- Condensed Sans Serif: 15%

**Content**

**Countries Featured**

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>88%</td>
</tr>
<tr>
<td>China</td>
<td>22%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>12%</td>
</tr>
<tr>
<td>Australia</td>
<td>12%</td>
</tr>
<tr>
<td>Canada</td>
<td>10%</td>
</tr>
<tr>
<td>India</td>
<td>10%</td>
</tr>
<tr>
<td>France</td>
<td>10%</td>
</tr>
<tr>
<td>Mexico</td>
<td>8%</td>
</tr>
</tbody>
</table>

**Theme**
Relative popularity of different infographic themes:

- Business: 25%
- Technology: 20%
- Culture: 18%
- Health: 15%
- Politics: 10%
- Environment: 8%

**Key Info**
Percentage of infographics with key:

- 0%: 33%
- 1%: 13%
- 2%: 18%
- 3%: 18%
- 4%: 13%
- 5%: 7%
- 7%: 4%
- 10%: 2%

Average number of symbols per key: 5.1

**Base Color**

- Red: 18%
- Blue: 29%
- Green: 13%
- Yellow: 29%
- Orange: 18%

**Navigational Iconography**
Frequency of arrows & connecting lines in infographics:
- Arrows: 13%
- Lines: 38%
- Both: 13%

**Sections**
Average number of sources per infographic: 2.29

**Credited Sources**

- Average number of words per infographic title: 4.36

**Title**
“RICHEST AND POOREST AMERICAN NEIGHBORHOODS”

**Concept & Design**: Iran Cohl

**Source**: 49 infographics collected at random from www.visualcapitalist.com/infographics
2013: The Year in Interactive Storytelling

Effective tax rate 2007-12

Overall 29.1%

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 $95 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, A.I.G. lost $83 billion while paying $8 billion in taxes. These companies are still included in overall tax rate calculations.

CHART KEY

Color shows effective rate

10 20 30 40 50%

Size shows market capitalization

$1B $10B $50B $100B
LoyalTracker: Visualizing Loyalty Dynamics in Search Engines

Conglei Shi, Yingcai Wu, Member, IEEE, Shixia Liu, Senior Member, IEEE, Hong Zhou and Huamin Qu, Member, IEEE

Fig. 1: LoyalTracker illustrates loyalty dynamics of the users using search engine A. Top and bottom show the same flow view that highlights two different flowing patterns of the users (in orange) selected from a layer flow (top) and a branch flow (bottom) across multiple loyalty categories (layers) over time. The switching histogram on the top shows a visual summary of switching behavior.
The Tea Party wave of 2010 helped House Republicans regain all the ground they had lost in 2006 and 2008 and then some.

While the size of the Republican majority didn’t change very much during this period, its internal makeup did. The centrist bloc nearly vanished, and every election from 1984 to 2004, the far right bloc grew.

Newt Gingrich’s Republican Revolution of 1994 gave the GOP control of the House for the first time in 40 years.

Democratic losses in 1994 came almost exclusively among centrists, leaving the left and far-left blocs relatively untouched—a pattern which would continue in 2010.

The Democratic Party’s gains in 2006 and 2008 came largely from centrist candidates in swing districts—the same ones who would be voted out in 2010.
ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories

Panpan Xu, Honghui Mei, Liu Ren, and Wei Chen

To Real-Time Tracking

Real-time Tracking Panel

Fault Codes
- Unknown status code
- Weld Position (1-4) Velocity Lower Limit Exceeded
- Weld Position (1-4) Velocity Upper Limit Exceeded
- Part Detection Lower Limit Exceeded Lead (1-14)
- Part Detection Upper Limit Exceeded Lead (1-10)
- Part Detection Failure (3/54)
- DESTRUCTIVE TEST PART
- Manual Selected During Cycle
- Light Curtain Entered During Cycle

39
train stops

STAR WARS
(ORIGINAL TRILOGY)
Data-Driven Guides: Supporting Expressive Design for Information Graphics

Nam Wook Kim, Eston Schweickart, Zhicheng Liu, Mira Dontcheva, Wilmot Li, Jovan Popovic, and Hanspeter Pfister

Fig. 1: Nigel Holmes’ *Monstrous Costs* chart, recreated by importing a monster graphic (left) and retargeting the teeth of the monster with DDG (middle). Taking advantage of the data-binding capability of DDG, small multiples are easily created by copying the chart and changing the data for each cloned chart (right).
可视化了的标准

• 信：Accuracy

• 达：Intuitiveness, Efficiency, and Effectiveness

• 雅：Aesthetic
Deep Learning Models

Visualization

Visualization techniques to facilitate
- Debugging
- Understanding
- Performance improvement of deep learning models

Keep Human in the AI loop
Understanding Hidden Memories of Recurrent Neural Networks

Category: Research

Paper Type: application/design study
Urban Informatics, Social Media, Text Visualization, VisMOOCs, Immersive Analytics, Social Network Research Focuses
微信小秘密: 2016 年那些 10w+ 文章是怎么刷爆朋友圈的？

告诉大家一个悲伤的消息，2016 的进度条即将告罄。在 2016 年最后一个工作日，我们还是踏实地点东西吧，和小编一起回头看看这一年发生的大事，又有哪些热点曾经刷爆了我们的朋友圈？

http://mp.weixin.qq.com/s?__biz=MzI5MDAwOTIzOQ==&mid=2650901442&idx=1&sn=ae2b21b
【盘点】2016上半年最全的热点网络事件，你关注了几个？

2016年已经过去一半了，下半年的第一天也快华丽丽的过去了，回顾过去，上半年发生了好多事情，今天我们就来盘点一下上半年发生的“大事情”，排名不分先后~

1、“2016中国第一网红”——papi酱

[图片]

先生们，这将是新媒体营销史上的第一大事件

2016-03-21 罗辑宇 罗辑思维
热点一：AlphaGo 大战李世石
热点二：iPhone7 发布会
热点三：川普当选
Some hot topics

• VIS + Machine Learning: Interactive/Explainable Artificial Intelligence

• VIS + Storytelling: Narrative Visualization

• VIS + Augmented reality: Immersive Analytics
Thank You!