Understanding models: A challenge for visual analytics

Jack van Wijk
TU/e and JADS

Delft Data Science Seminar
Delft, February 6, 2018

Overview
• Presentation, interaction, visual analytics
• Understanding models: three cases

Anscombe’s quartet

Data
produce and collect
Visualization
People
want insights

Francis Anscombe, 1973

Anscombe’s quartet

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td>y1</td>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td>x2</td>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td>y2</td>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td>x3</td>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td>y3</td>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.75</td>
</tr>
<tr>
<td>x4</td>
<td>6.0</td>
<td>6.74</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td>y4</td>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>3.39</td>
<td>8.0</td>
<td>12.50</td>
</tr>
<tr>
<td>x5</td>
<td>12.0</td>
<td>10.94</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td>y5</td>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td>x6</td>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.84</td>
</tr>
</tbody>
</table>

Four data-sets: same average x and y, same variance, same correlation, …

Visualization
produce and collect

BIG
MESSY

want insights
Standard diagrams don’t scale

- Multivariate visualization: scatterplot
- Tree visualization: tree diagram
- Graph visualization: node link diagram

SequoiaView

Van Wijk and Van de Wetering, 1999

Botanically inspired treevis

Botanically inspired treevis

One picture is worth a lot of pixels

- Focus on visual presentation
- Show everything in one picture?

Limits on perception…

How many disks?

Chris Healey: Perception in Visualization

Limits on perception…

How many disks?

Chris Healey: Perception in Visualization
Limits on perception…

How many red objects?

Chris Healey: Perception in Visualization

Limits on perception…

How many red disks?

Chris Healey: Perception in Visualization

Visualization

Information Visualization

• The use of computer-supported, interactive, visual representations of abstract data to amplify cognition (Card et al., 1999)

Data → Visualization → User

BIG MESSY

representation

interaction
Keshif.me (M. Adil Yalçın et al.) demo

Image categorization

Van der Corput & Van Wijk, IEEE PacificVis, 2016

Multivariate network visualization

Van den Elzen & Van Wijk, IEEE InfoVis, 2014

Antibiotics resistance patterns

Niels Drost, MSc 2017

Visualization

BIG MESSY

Visualization

BIG MESSY

Statistics, machine learning, data mining, …
Visual Analytics

Model: statistics, machine learning, data mining, …

- Steer model
- Control vis.

Computers
- Fast
- Precise
- Computation
- Search and store data
- Graphics

Humans
- Flexible, inventive
- Solve problems
- Handle new situations
- Handle incomplete and/or inconsistent information
- Have domain knowledge and experience
- Can see things that are hard to compute

Visual Analytics: Synergy between computers and humans

Data size

- Business graphics
- Infos
- Visual analytics

Try to move to the left:
- Filter, aggregate, statistics, machine learning, …
- Without loosing essential information.

Should we trust the model?

Automated decisions…?

- You qualify for our special offer
- You are not admitted to our education program
- Your job application is put aside
- Your mortgage request cannot be honored
- Your research proposal is rejected
- You should get vitrectomie
- Your probation request is declined
- You are fired
- You are arrested

Should we let the computer decide?
The challenge

• How to obtain transparency in automated decision making?

• How to present the evidence and reasoning used, such that humans can understand, validate, and judge decisions made?

Who’s at stake?

Data Scientist

Decision Maker

Subject

Train & test cases

Interaction

How to enable people to explore the model with minimal effort?

Presentation

How to show the relations between data and results in an understandable way?

Varying needs

How to support stakeholders with varying needs and capabilities?

Is my model correct?

Does the label match with my domain knowledge?

Case 1: Decision tree visualization

Problem:

• Support construction of decision trees

• Enable domain expert to bring in domain knowledge

White box approach:

• Model explicitly shown

Complex models

Increasing complexity:

• rules

• logistic regression

• decision trees

• support vector machines

• random forests

• neural networks

• deep learning networks

Size matters:

• 1000 rules?

• 100 variables?

• 50 layers?

• 10 dimensions?

• 100 trees?

• 1000’s of nodes?

• millions of nodes?
Case 2: Rationale Visualization for Safety and Security

Approach:
• show strongly simplified model
• for one case

Problem

Context
Case 3: Insurance Fraud detection

MSc project Dennis Collaris

Support fraud detection team in prioritization of cases

Approach:
- show strongly simplified model
- for one case

Start point

Data set:
- 38,138 insurance policies
- 49 attributes per policy
- 129 confirmed fraud

Model:
- Bagging ensemble of
  - 100 Random Forest models, each with
  - 500 CART decision trees

Dennis Collaris, 2018
Finally

- Presentation, interaction, visual analytics
- Transparent models decisions crucial
- Custom cases can be solved, but require much effort and expertise:
  - extraction meaningful explanation
  - understandable visualization
- Many challenges ahead
- Stop for red lights

Thank you!