Reverse Engineering Chart Data with WebPlotDigitizer

Ankit Rohatgi <ankitrohatgi@hotmail.com>
May 3, 2017

http://arohatgi.info/WebPlotDigitizer
http://github.com/ankitrohatgi/WebPlotDigitizer
Raw Data?

- Representative Scatchard plot of HNECA saturation binding data
  - Laura Bazzichi
  - 30/12/2011

- Forest plot of sensitivity
  - Steve Goodacre
  - 30/12/2011

- Linkage disequilibrium (D') plot of IFT1 gene in Caucasians
  - Mohammad A Karim
  - 30/12/2011

- Funnel plot for sensitivity
  - Steve Goodacre
  - 30/12/2011

- Base composition plot in core region of mixture alignment
  - Ji-Ping Z. Wang
  - 30/12/2011

- Frequency plot of TT and AA signals in the alignment presented in
  - Ji-Ping Z. Wang
  - 30/12/2011

- L'Abbe plot of risk of leak in single-layer vs
  - Satoru Shikata
  - 30/12/2011

- Plot of fluorescence, polarographic and predicted partial oxygen tension...
  - Andrew D Shaw
  - 30/12/2011

- Plot of the rank of the median probe

- Scatter plot of direct labeling
  - (A) Scatter plot of direct labeling

- Effect of secondary structure on constant

- Kernel-based scatter plot
Fetching Raw Data

Pixel Counting?
Fetching Raw Data

Geometry?
Existing tools

• A few tools are available[1–6], but with many issues:
  • Difficult to access or incompatible with the operating system
  • Support only XY charts
  • Complicated interface
  • Accuracy concerns
  • Some are expensive, but not significantly better
  • Minimal automation
Existing tools

- A few tools are available[1–6], but with many issues:
  - Difficult to access or incompatible with the operating system
  - Support only XY charts
  - Complicated interface
  - Accuracy concerns
  - Some are expensive, but not significantly better
  - Minimal automation
- Complete automation is still an area of active research[7–10].
- Free, opensource, web based tool
- Works with a wide variety of charts
- Partial automation with sub-pixel resolution algorithms
Demonstration

Also available on http://arohatgi.info/WebPlotDigitizer
Workflow

1. **Load Image**
   - Calibrate Axes (XY, Bar, Polar, Ternary etc.)

2. **Define ROI (Mask)**
   - Specify Color Range
   - Select Algo & Parameters

3. **Extraction Method**
   - Manual Add, Adjust or Delete

4. **Visual Inspection**
   - adjust
   - ok

5. **Digitized Data**
XY Charts

Affine Transformation

X and Y Axes Calibration

Enter X-values of the two points clicked on X-axis and Y-values of the two points clicked on Y-axes

<table>
<thead>
<tr>
<th>X-Axis:</th>
<th>Point 1</th>
<th>Point 2</th>
<th>Log Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*For dates, use yyyy/mm/dd format (e.g. 2013/10/23 or 2013/10). For exponents, enter values as 1e-3 for 10^-3.
Bar Charts and Histograms

Bar Chart

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
</tr>
</tbody>
</table>

Histogram*

*Calibrate as a 2D XY plot

Enter the values at the two points selected on the continuous axes along the bars

<table>
<thead>
<tr>
<th>Point 1</th>
<th>Point 2</th>
<th>Log Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>
Polar Diagrams

Align Polar Axes

Point 1  Point 2 Log Scale
R:  10   100
Θ:  10   30

- Degrees
- Radians
- Clockwise

OK
Ternary Diagrams
Scaled Images (Maps, Microscope, etc.)

Apple A7 (ifixit.com)

Arado AR 65F (the-blueprints.com)
Measurement Tools

Angle Measurement

[1]: 30.96°
[2]: 82.85°
Auto-Extraction

Challenges:

- Color and shape based image segmentation
- Region of interest identification
- Sub-pixel thinning, centroid estimation
Averaging Window Algorithm

Suites for continuous curves and data points
X Step with Interpolation Algorithm

Suited for continuous or discontinuous curves, data points and noisy data
X Step with Interpolation Algorithm

Dashed Lines

Noisy Data

\( s = 200\% \)

\( s = 400\% \)

\( s = 1000\% \)
Bar Charts and Histogram Algorithms

Suited for vertical or horizontal bar charts and histograms
Blob Extraction

- Connected component labeling
- Computes:
  - Centroid
  - Area
  - Moment Invariant
- Shape based extraction

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Blob Detector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Diameter</td>
<td>0 Px</td>
</tr>
<tr>
<td>Max Diameter</td>
<td>500 Px</td>
</tr>
</tbody>
</table>
Grid Removal
Accuracy: Independent Studies

Estimating data from figures with a Web-based program: Considerations for a systematic review

Brittany C. Barca | Elizabeth A. O'Connor | Elizabeth M. Volber | Natal B Oldmond | Leslie A. Porter

Accredited Reference: accuracy, Independent Research, Center for Research, Portland, OR, USA.

Correspondence: Brittany C. Barca, 360 S, Portland, OR 97232; Email: barca@wusc.edu.

Abstract: Systematic reviews often require extracting raw data, and information derived may be difficult to obtain from a study author. This systematic review examined how to extract data from figures, with tables as a way to systematically improve the accuracy of the estimates.

Keywords: extraction, accuracy, systematic review.

Relying on Web-based programs for data extraction, and Usability of Data Analysis Programs for Single-Case Research Designs

Mariola Moeyaert, Daniel Maggin, and Jay Verkuilen

Abstract: Single-case experimental designs (SCEDs) have been increasingly used in recent years to inform the development and validation of effective interventions in the behavioral sciences. An important aspect of this work has been the extension of meta-analytic and other statistical innovations to SCEDs. Standard practice within SCED methods is to display data graphically, which requires subsequent users to extract the data, either manually or using data extraction programs. Previous research has examined issues of reliability and validity of data extraction programs in the past, but typically at an aggregate level. Little is known, however, about the coding of individual data points. We focused on four different software programs that can be used for this purpose (e.g., Egraph, DataThief, WebPlotDigitizer, and XfA) and examined the reliability of numeric coding, the validity compared with real data, and overall program usability. This study indicates that the reliability and validity of the retrieved data are independent of the specific software program, but are dependent on the individual single-case researcher.

Keywords: reliability, validity, data extraction, meta-analysis, single-case design.
Interesting Use Cases

Mark Brandon
@icey_mark

Said it before Web Plot Digitizer is the just the most amazingly useful tool for extracting data from papers
arohatgi.info/WebPlotDigitiz…
Interesting Use Cases

Cindy Harnett
@CindyHarnett

Liking webplotdigitizer for grabbing data from pics in #lab, thanks @CousinAmygdala for the tip.
arohatgi.info/WebPlotDigitiz…
Interesting Use Cases

R Package (Under Development)

https://github.com/ankitrohatgi/digitizeR
Native Desktop App (Under Development)

Qt/C++
Public Issue Tracker
References


