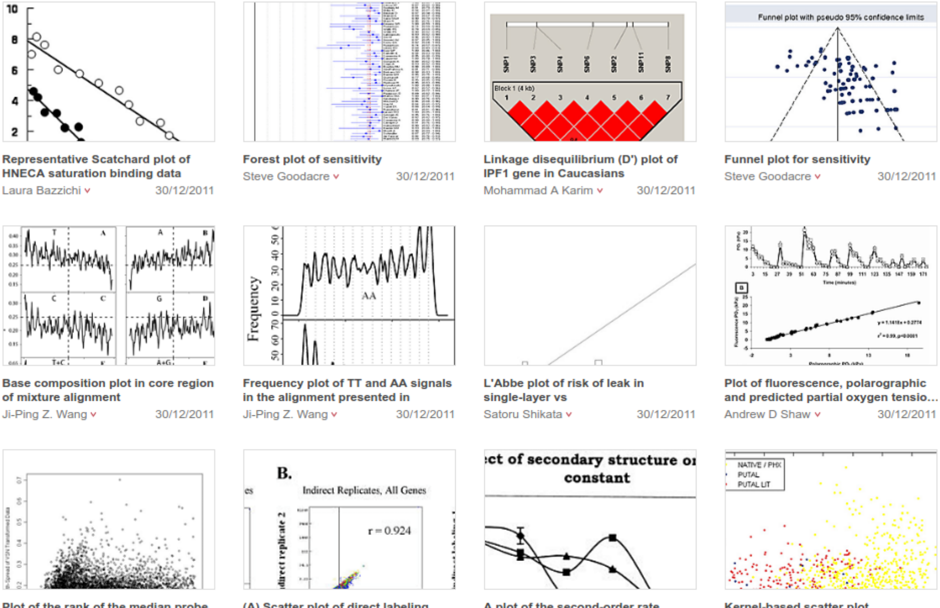
Reverse Engineering Chart Data with WebPlotDigitizer

Ankit Rohatgi <ankitrohatgi@hotmail.com>

May 3, 2017

http://arohatgi.info/WebPlotDigitizer
http://github.com/ankitrohatgi/WebPlotDigitizer

Raw Data?


Funnel plot with pseudo 95% confidence limits
SNP1
SNP3
SNP4
SNP6
SNP2
SNP11
SNP8
Block 1 (4 kb)
Frequency
AA
B.
Indirect Replicates, All Genes
r = 0.924
ect of secondary structure or
constant
NATIVE / PHX
PUTAL
PUTAL LIT

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 IPF1 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 tensio...
Andrew D Shaw 30/12/2011

Plot of the rank of the median probe

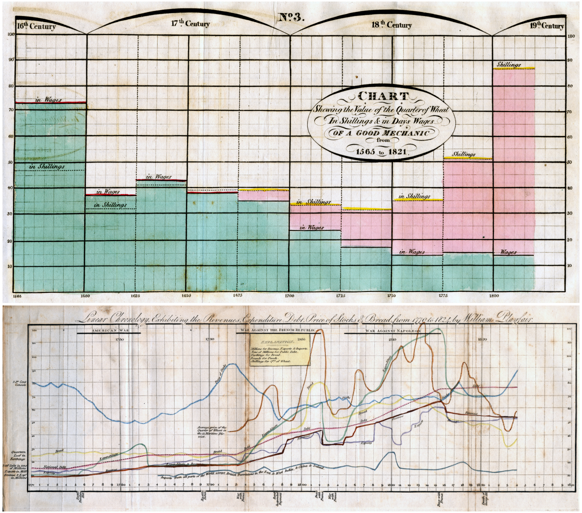
(A) Scatter plot of direct labeling

A plot of the second-order rate

Kernel-based scatter plot

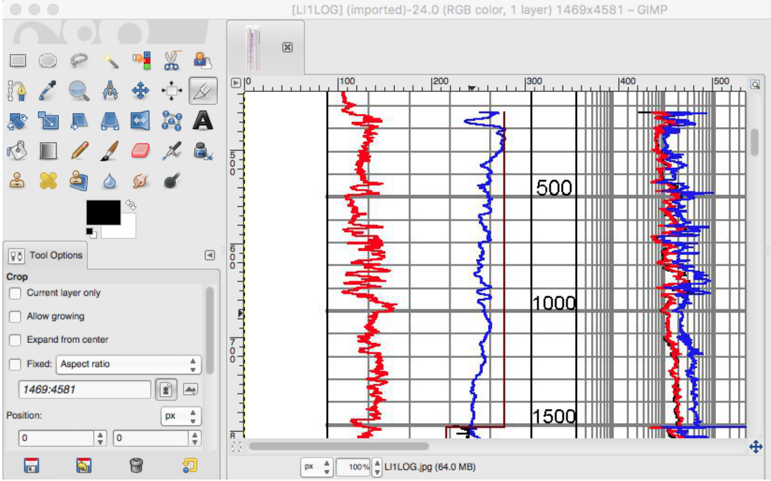
Fetching Raw Data

Contact Authors



William Playfair, 1759-1823

Fetching Raw Data



[LI1LOG] (imported)-24.0 (RGB color, 1 layer) 1469x4581 – GIMP

Tool Options

Crop

□ Current layer only

□ Allow growing

□ Expand from center

□ Fixed: Aspect ratio

1469:4581

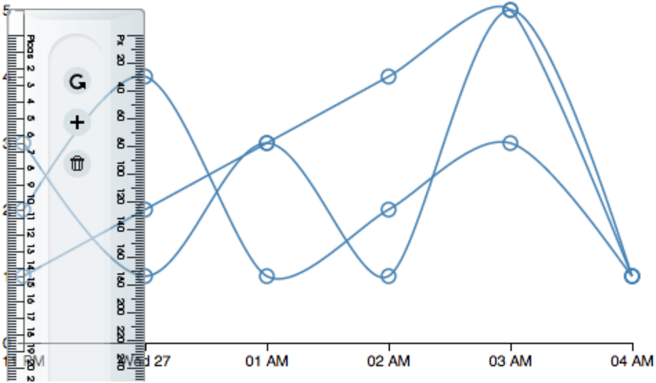
Position: px

0 0

px 100% LI1LOG.jpg (64.0 MB)

Pixel Counting?

Fetching Raw Data



Geometry?

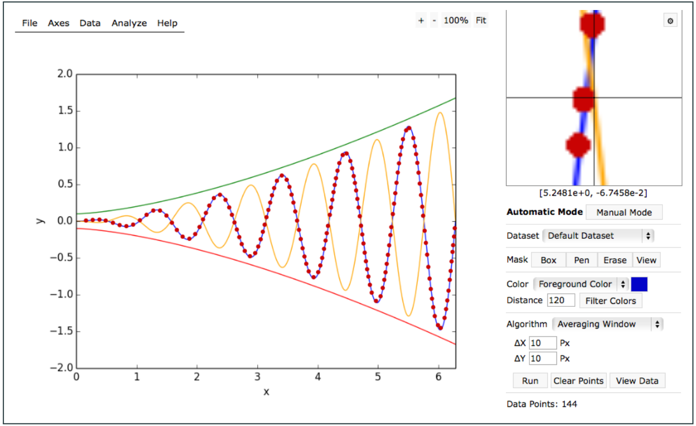
Existing tools

- 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].

WebPlotDigitizer



[5.2481e+0, -6.7458e-2]

Dataset: Default Dataset

Color: Foreground Color

Distance: 120

Algorithm: Averaging Window

ΔX: 10 Px

ΔY: 10 Px

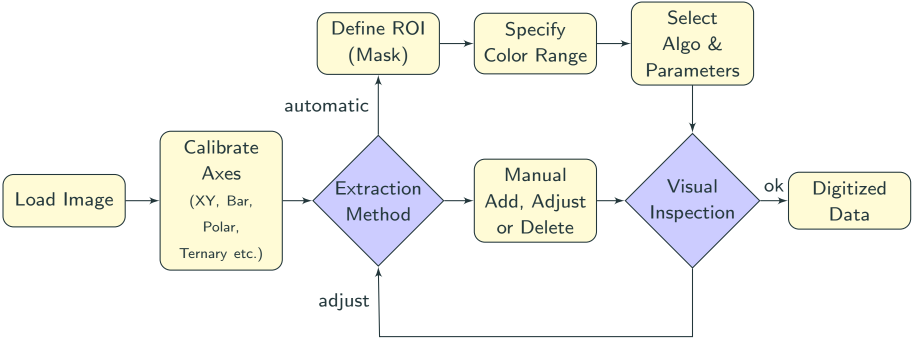
Data Points: 144

- Free, opensource, web based tool
- Works with a wide variety of charts
- Partial automation with sub-pixel resolution algorithms

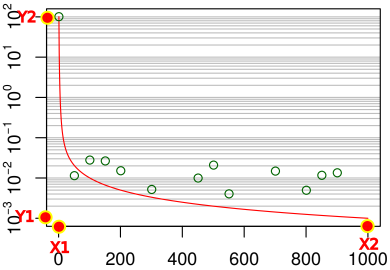
Demonstration

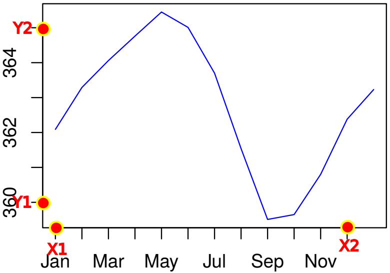
Also available on http://arohatgi.info/WebPlotDigitizer

Workflow

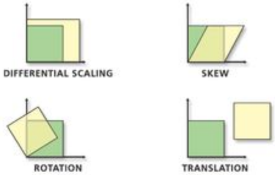

Load Image
Calibrate Axes
(XY, Bar, Polar, Ternary etc.)
Extraction Method
automatic
Define ROI (Mask)
Specify Color Range
Select Algo & Parameters
Manual Add, Adjust or Delete
Visual Inspection
ok
Digitized Data
adjust

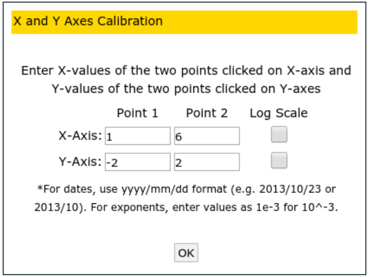
XY Charts





Affine Transformation





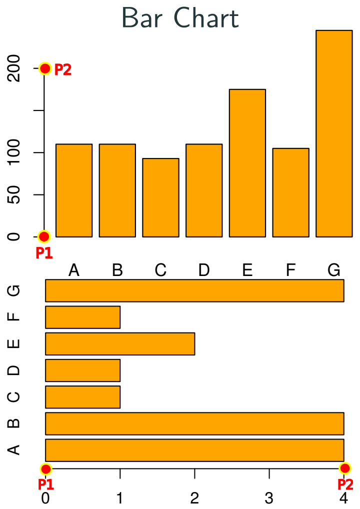
X and Y Axes Calibration

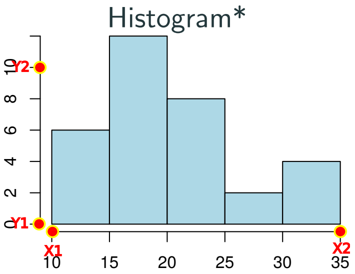
Enter X-values of the two points clicked on X-axis and Y-values of the two points clicked on Y-axes

	Point 1	Point 2	Log Scale
X-Axis:	1	6	□
Y-Axis:	-2	2	□

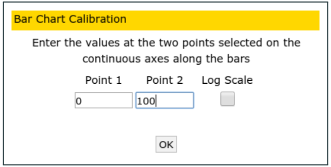
*For dates, use yyyy/mm/dd format (e.g. 2013/10/23 or 2013/10). For exponents, enter values as 1e-3 for 10^-3.

Bar Charts and Histograms





*Calibrate as a 2D XY plot

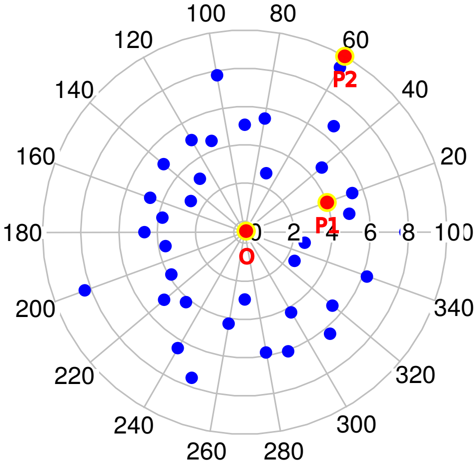


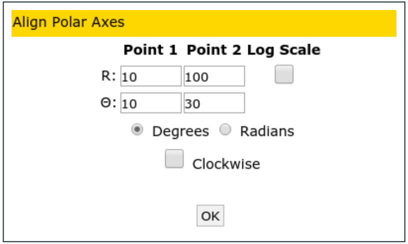
Bar Chart Calibration

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

Point 1	Point 2	Log Scale
0	100	□

Polar Diagrams





Align Polar Axes

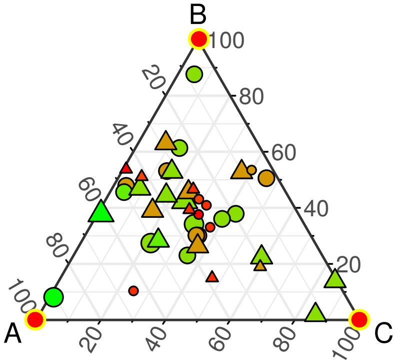
	Point 1	Point 2	Log Scale
R:	10	100	□
Θ:	10	30	

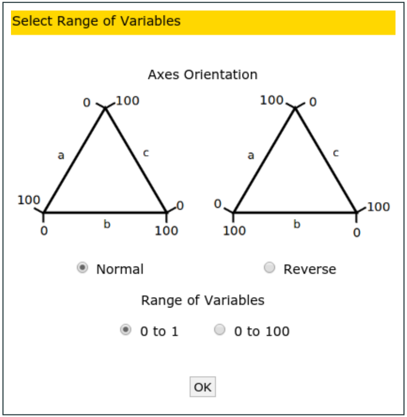
● Degrees ○ Radians

□ Clockwise

OK

Ternary Diagrams




Select Range of Variables

Axes Orientation

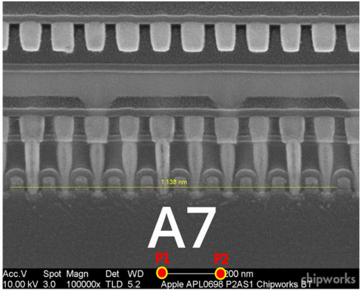
● Normal ○ Reverse

Range of Variables

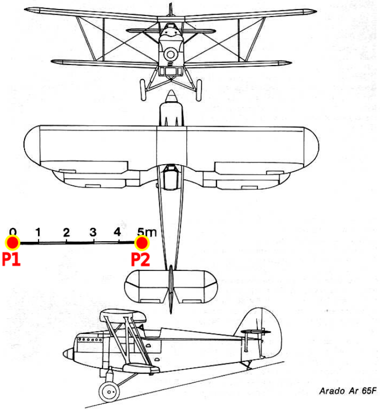
● 0 to 1 ○ 0 to 100

OK

Scaled Images (Maps, Microscope, etc.)

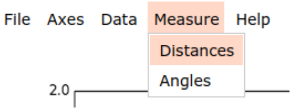

1,138 nm
A7
P1
P2
Acc.V Spot Magn Det WD
10.00 kV 3.0 100000x TLD 5.2
200 nm
Apple APL0698 P2AS1 Chipworks BT
chipworks

Apple A7 (ifixit.com)

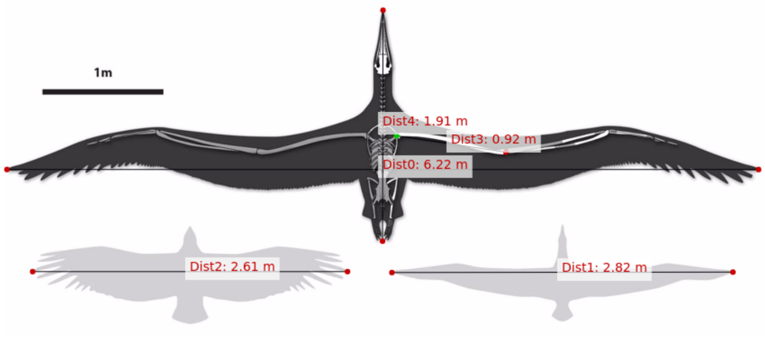


Arado AR 65F (the-blueprints.com)

Measurement Tools



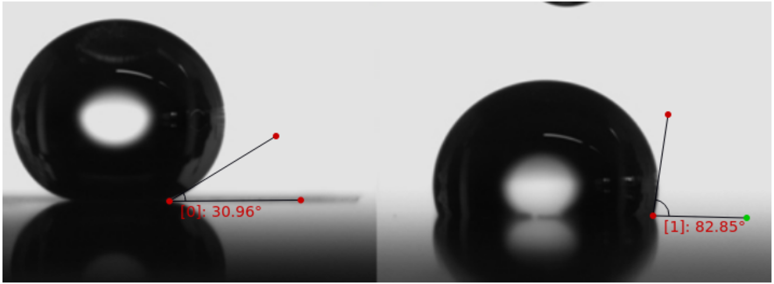
Distance Measurement



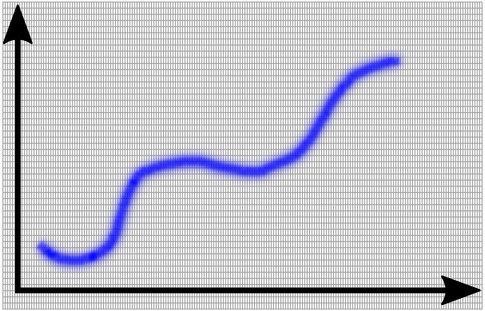
Pelagornis sandersi (National Geographic)

Measurement Tools

Angle Measurement



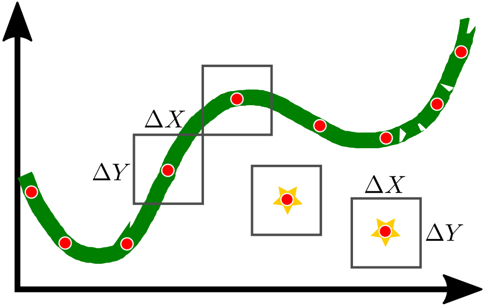
Auto-Extraction

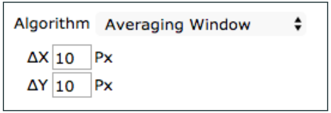


Challenges:

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

Averaging Window Algorithm





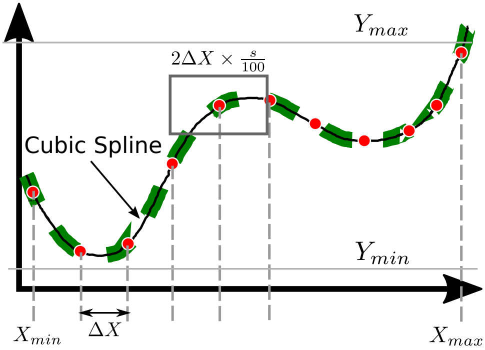
Algorithm: Averaging Window

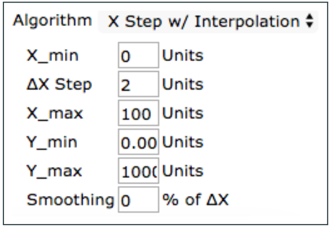
ΔX 10 Px

ΔY 10 Px

Suited for continuous curves and data points

X Step with Interpolation Algorithm





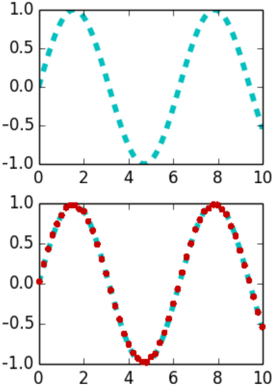
Algorithm X Step w/ Interpolation

X_min	0	Units
ΔX Step	2	Units
X_max	100	Units
Y_min	0.00	Units
Y_max	100(	Units
Smoothing	0	% of ΔX

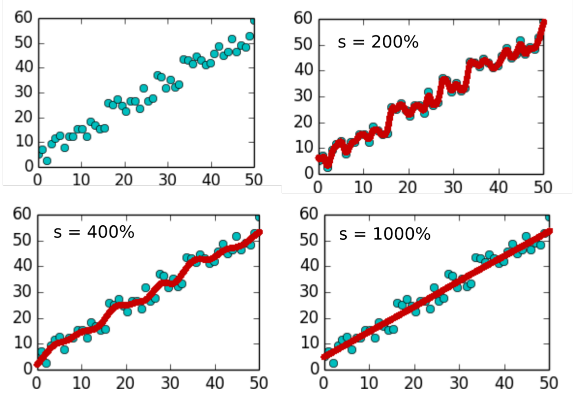
Suited for continuous or discontinuous curves, data points and noisy data

X Step with Interpolation Algorithm

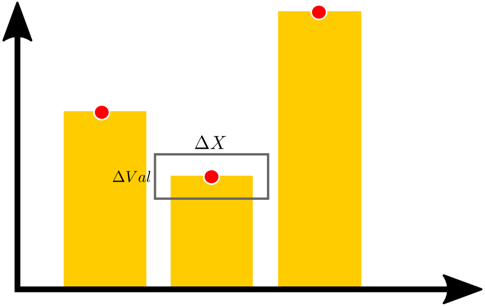
Dashed Lines

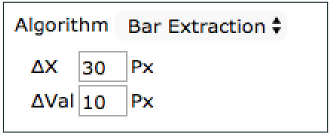


Noisy Data



Bar Charts and Histogram Algorithms





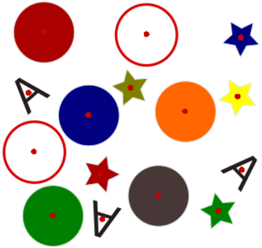
Algorithm: Bar Extraction

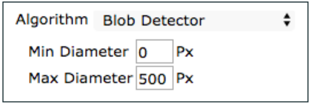
ΔX 30 Px

ΔVal 10 Px

Suited for vertical or horizontal bar charts and histograms

Blob Extraction





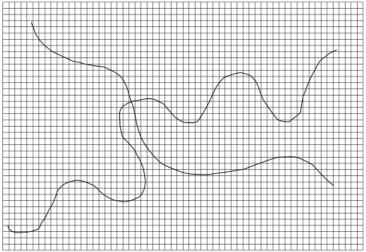
Algorithm: Blob Detector

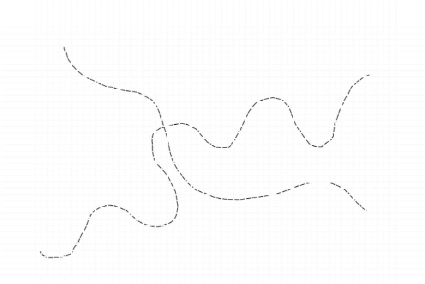
Min Diameter: 0 Px

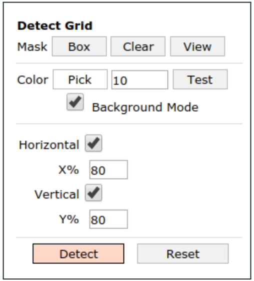
Max Diameter: 500 Px

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

Grid Removal







Detect Grid

Mask Box Clear View

Color Pick 10 Test

☑ Background Mode

Horizontal ☑

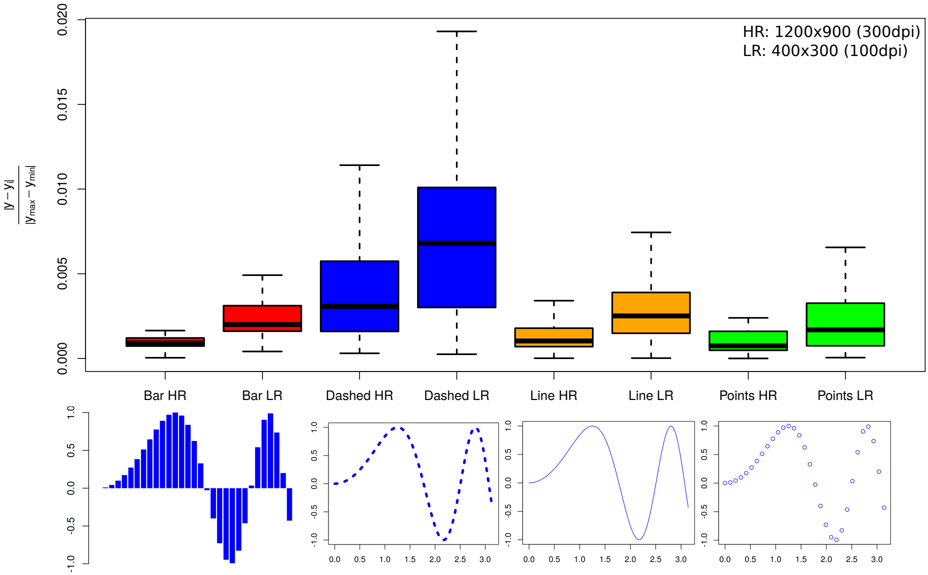
X% 80

Vertical ☑

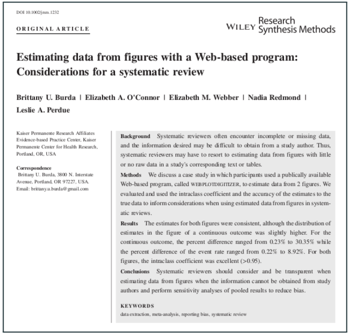
Y% 80

Detect Reset

Accuracy: Summary



Accuracy: Independent Studies



DOI 10.1002/jrsm.1232

ORIGINAL ARTICLE

WILEY Research Synthesis Methods

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

Brittany U. Burda | Elizabeth A. O'Connor | Elizabeth M. Webber | Nadia Redmond | Leslie A. Perdue

Kaiser Permanente Research Affiliates Evidence-based Practice Center, Kaiser Permanente Center for Health Research, Portland, OR, USA

Correspondence
Brittany U. Burda, 3800 N. Interstate Avenue, Portland, OR 97227, USA.
Email: brittany.u.burda@gmail.com

Background Systematic reviewers often encounter incomplete or missing data, and the information desired may be difficult to obtain from a study author. Thus, systematic reviewers may have to resort to estimating data from figures with little or no raw data in a study's corresponding text or tables.

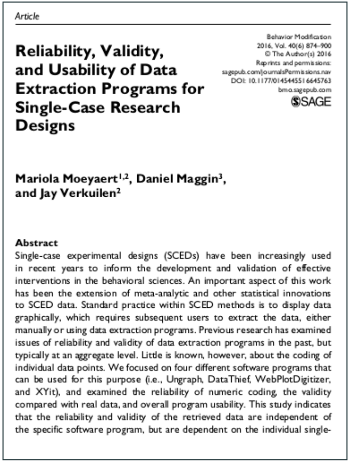
Methods We discuss a case study in which participants used a publically available Web-based program, called WEBPLOTDIGITIZER, to estimate data from 2 figures. We evaluated and used the intraclass coefficient and the accuracy of the estimates to the true data to inform considerations when using estimated data from figures in systematic reviews.

Results The estimates for both figures were consistent, although the distribution of estimates in the figure of a continuous outcome was slightly higher. For the continuous outcome, the percent difference ranged from 0.23% to 30.35% while the percent difference of the event rate ranged from 0.22% to 8.92%. For both figures, the intraclass coefficient was excellent (>0.95).

Conclusions Systematic reviewers should consider and be transparent when estimating data from figures when the information cannot be obtained from study authors and perform sensitivity analyses of pooled results to reduce bias.

KEYWORDS
data extraction, meta-analysis, reporting bias, systematic review

[11]



Article

Behavior Modification
2016, Vol. 40(6) 874–900
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0145445516645763
bmo.sagepub.com
SAGE

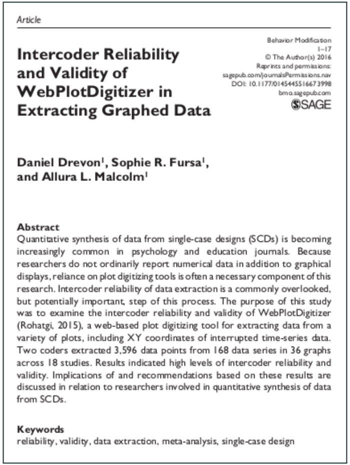
Reliability, Validity, and Usability of Data Extraction Programs for Single-Case Research Designs

Mariola Moeyaert[1,2], Daniel Maggin[3], and Jay Verkuilen[2]

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 SCED data. 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 (i.e., Ungraph, DataThief, WebPlotDigitizer, and XYit), 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-

[12]



Article

Behavior Modification
1–17
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0145445516673998
bmo.sagepub.com
SAGE

Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data

Daniel Drevon[1], Sophie R. Fursa[1], and Allura L. Malcolm[1]

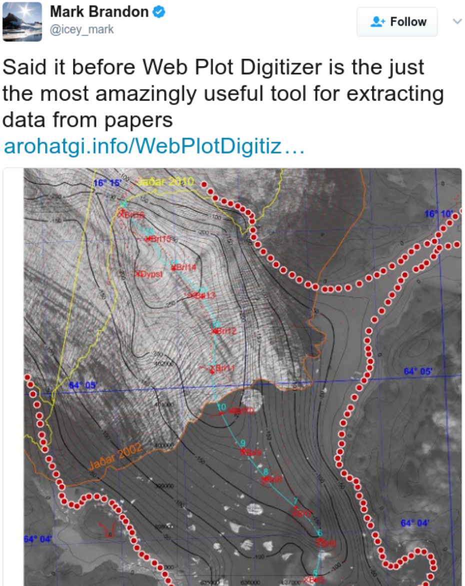
Abstract

Quantitative synthesis of data from single-case designs (SCDs) is becoming increasingly common in psychology and education journals. Because researchers do not ordinarily report numerical data in addition to graphical displays, reliance on plot digitizing tools is often a necessary component of this research. Intercoder reliability of data extraction is a commonly overlooked, but potentially important, step of this process. The purpose of this study was to examine the intercoder reliability and validity of WebPlotDigitizer (Rohatgi, 2015), a web-based plot digitizing tool for extracting data from a variety of plots, including XY coordinates of interrupted time-series data. Two coders extracted 3,596 data points from 168 data series in 36 graphs across 18 studies. Results indicated high levels of intercoder reliability and validity. Implications of and recommendations based on these results are discussed in relation to researchers involved in quantitative synthesis of data from SCDs.

Keywords
reliability, validity, data extraction, meta-analysis, single-case design

[13]

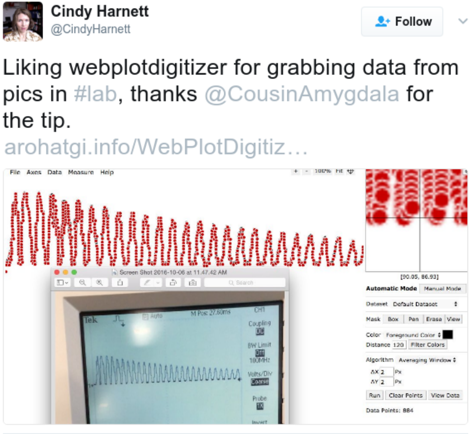
Interesting Use Cases



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…

[90.05, 86.93]

Automatic Mode | Manual Mode

Dataset: Default Dataset

Color: Foreground Color

Distance: 120

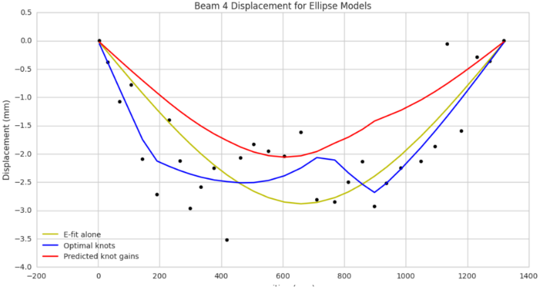
Algorithm: Averaging Window

ΔX: 2 Px

ΔY: 2 Px

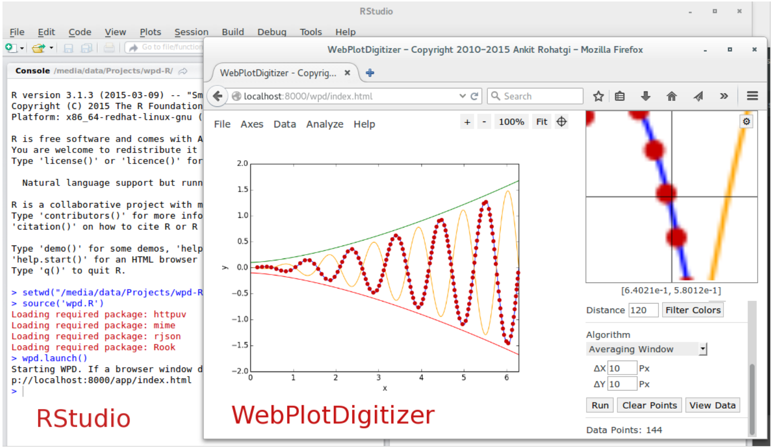
Data Points: 884

Interesting Use Cases



Thurston Sexton. “Optimal Modeling of Knots in Wood”. Arizona State University, 2015

R Package (Under Development)



```
R version 3.1.3 (2015-03-09) -- "Sm
Copyright (C) 2015 The R Foundation
Platform: x86_64-redhat-linux-gnu (

R is free software and comes with A
You are welcome to redistribute it
Type 'license()' or 'licence()' for

  Natural language support but runn

R is a collaborative project with m
Type 'contributors()' for more info
'citation()' on how to cite R or R

Type 'demo()' for some demos, 'help
'help.start()' for an HTML browser
Type 'q()' to quit R.

> setwd("/media/data/Projects/wpd-R
> source('wpd.R')
Loading required package: httpuv
Loading required package: mime
Loading required package: rjson
Loading required package: Rook
> wpd.launch()
Starting WPD. If a browser window d
p://localhost:8000/app/index.html
> |
```

RStudio

[6.4021e-1, 5.8012e-1]

Distance 120

Algorithm

Averaging Window

ΔX 10 Px

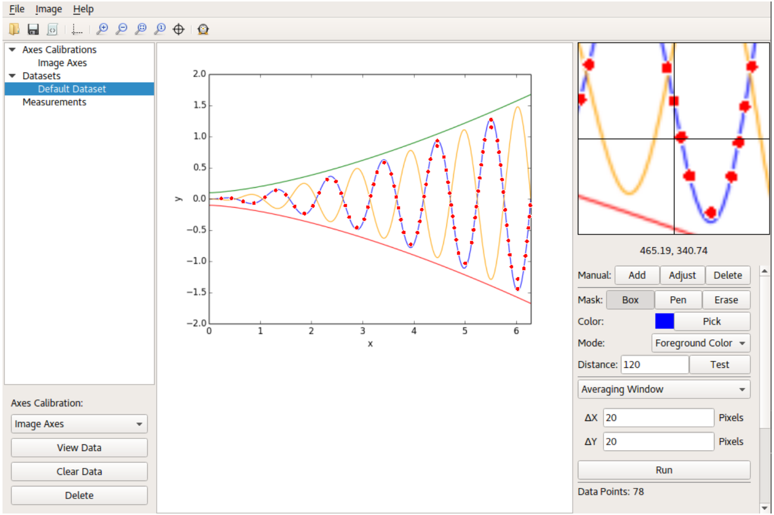
ΔY 10 Px

Data Points: 144

WebPlotDigitizer

https://github.com/ankitrohatgi/digitizeR

Native Desktop App (Under Development)



465.19, 340.74

Color: 🟦

Mode: Foreground Color

Distance: 120 Test

Averaging Window

ΔX 20 Pixels

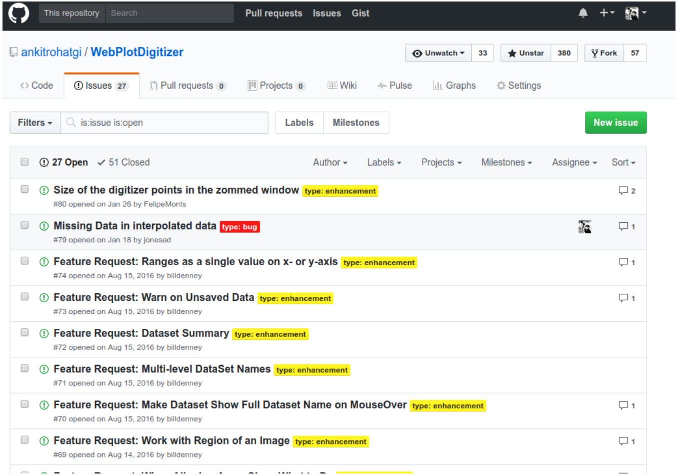
ΔY 20 Pixels

Data Points: 78

Axes Calibration: Image Axes

Qt/C++

Public Issue Tracker



ankitrohatgi / WebPlotDigitizer

27 Open ✓ 51 Closed

- Size of the digitizer points in the zommed window `type: enhancement` — 2
  #80 opened on Jan 26 by FelipeMonts
- Missing Data in interpolated data `type: bug` — 1
  #79 opened on Jan 18 by jonesad
- Feature Request: Ranges as a single value on x- or y-axis `type: enhancement` — 1
  #74 opened on Aug 15, 2016 by billdenney
- Feature Request: Warn on Unsaved Data `type: enhancement` — 1
  #73 opened on Aug 15, 2016 by billdenney
- Feature Request: Dataset Summary `type: enhancement`
  #72 opened on Aug 15, 2016 by billdenney
- Feature Request: Multi-level DataSet Names `type: enhancement`
  #71 opened on Aug 15, 2016 by billdenney
- Feature Request: Make Dataset Show Full Dataset Name on MouseOver `type: enhancement` — 1
  #70 opened on Aug 15, 2016 by billdenney
- Feature Request: Work with Region of an Image `type: enhancement` — 1
  #69 opened on Aug 14, 2016 by billdenney

- http://arohatgi.info/WebPlotDigitizer
- http://www.github.com/ankitrohatgi/WebPlotDigitizer
- ankitrohatgi@hotmail.com
- ankit_rohatgi

References

[1] *PlotDigitizer*. URL: http://plotdigitizer.sourceforge.net/.

[2] *DigitizeIt*. URL: http://www.digitizeit.de/.

[3] *GetData graph digitizer*. URL: http://www.getdata-graph-digitizer.com/.

[4] *Engauge Digitizer*. URL: http://markummitchell.github.io/engauge-digitizer/.

[5] Biosoft. *Ungraph*. URL: http://www.biosoft.com/w/ungraph.htm.

[6] Geomatix. *XYit*. URL: http://www.geomatix.net/xyit/.

[7] Gonzalo Gabriel Méndez, Miguel A. Nacenta, and Sebastien Vandenheste. "iVoLVER". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI 2016*. Association for Computing Machinery (ACM), 2016. DOI: 10.1145/2858036.2858435.

[8] Noah Siegel et al. "FigureSeer: Parsing Result-Figures in Research Papers". In: *Computer Vision – ECCV 2016*. Springer Nature, 2016, pp. 664–680. DOI: 10.1007/978-3-319-46478-7_41.

References

[9] Daekyoung Jung et al. “ChartSense: Interactive Data Extraction from Chart Images”. In: ACM, May 2017. URL: https://www.microsoft.com/en-us/research/publication/chartsense-interactive-data-extraction-chart-images/.

[10] Sagnik Ray Choudhury and Clyde Lee Giles. “An Architecture for Information Extraction from Figures in Digital Libraries”. In: *Proceedings of the 24th International Conference on World Wide Web - WWW 2015 Companion*. Association for Computing Machinery (ACM), 2015. DOI: 10.1145/2740908.2741712.

[11] Brittany U. Burda et al. “Estimating data from figures with a Web-based program: Considerations for a systematic review”. In: *Research Synthesis Methods* (2017). DOI: 10.1002/jrsm.1232.

[12] Daniel Drevon, Sophie R. Fursa, and Allura L. Malcolm. “Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data”. In: *Behav. Modif.* 41.2 (Mar. 2017), pp. 323–339. DOI: 10.1177/0145445516673998.

[13] M. Moeyaert, D. Maggin, and J. Verkuilen. “Reliability, Validity, and Usability of Data Extraction Programs for Single-Case Research Designs”. In: *Behav. Modif.* 40.6 (Apr. 2016), pp. 874–900. DOI: 10.1177/0145445516645763.

[14] Thurston Sexton. “Optimal Modeling of Knots in Wood”. Arizona State University, 2015.