

Understanding Your Data Through Plots

Ethan Burns

eaburns at cs.unh.edu



UNIVERSITY *of* NEW HAMPSHIRE

September 13, 2012

Introduction

- Outline
- Why Do We Care About Plots?
- What are Plots For?
- Tables of Data
- Pictures of Data
- More Tables
- More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools

Introduction

Outline

Introduction

■ Outline

■ Why Do We Care About Plots?

■ What are Plots For?

■ Tables of Data

■ Pictures of Data

■ More Tables

■ More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools

- Why use plots?
- Different types of plots
- Tools that I use

Why Do We Care About Plots?

Introduction

■ Outline

■ Why Do We Care About Plots?

■ What are Plots For?

■ Tables of Data

■ Pictures of Data

■ More Tables

■ More Pictures

Distributions of Values

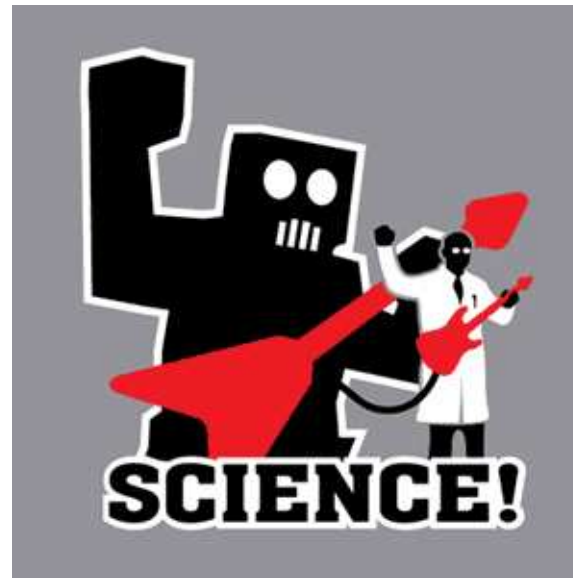
Trends in Data

Simple Plotting Tool

Other Tools

Grad Students do Research:

- Theoretical analysis
- **Experimental analysis**—lots of data



What are Plots For?

Introduction

- Outline
- Why Do We Care About Plots?

■ What are Plots For?

- Tables of Data
- Pictures of Data
- More Tables
- More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools

1. Understand behavior of new techniques

We need to see what the data is telling us

2. Demonstrate (to others) that new approaches work well

We want to convince others using our data

Clear, and obvious display of data

Tables of Data

Experiments generate lots of data

[Introduction](#)

- [Outline](#)
- [Why Do We Care About Plots?](#)
- [What are Plots For?](#)

■ **[Tables of Data](#)**

- [Pictures of Data](#)
- [More Tables](#)
- [More Pictures](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

My data:

	1	2	3	4	5	6	7	8
run 1:	-2.47	0.75	2.96	13.57	16.65	26.18	36.32	50.98
run 2:	-1.46	2.37	7.17	10.68	18.60	26.05	37.46	46.85
run 3:	1.40	1.86	6.00	5.95	15.37	28.78	38.20	47.01
run 4:	1.98	-0.23	1.13	4.70	16.27	25.89	34.31	48.83
run 5:	0.31	-1.90	4.56	5.52	17.41	25.69	33.86	47.33

Tables of Data

Experiments generate lots of data

[Introduction](#)

- [Outline](#)
- [Why Do We Care About Plots?](#)
- [What are Plots For?](#)

■ **[Tables of Data](#)**

- [Pictures of Data](#)
- [More Tables](#)
- [More Pictures](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

My data:

	1	2	3	4	5	6	7	8
run 1:	-2.47	0.75	2.96	13.57	16.65	26.18	36.32	50.98
run 2:	-1.46	2.37	7.17	10.68	18.60	26.05	37.46	46.85
run 3:	1.40	1.86	6.00	5.95	15.37	28.78	38.20	47.01
run 4:	1.98	-0.23	1.13	4.70	16.27	25.89	34.31	48.83
run 5:	0.31	-1.90	4.56	5.52	17.41	25.69	33.86	47.33

What is going on here?

Pictures of Data

Introduction

- Outline
- Why Do We Care About Plots?
- What are Plots For?

- Tables of Data
- **Pictures of Data**

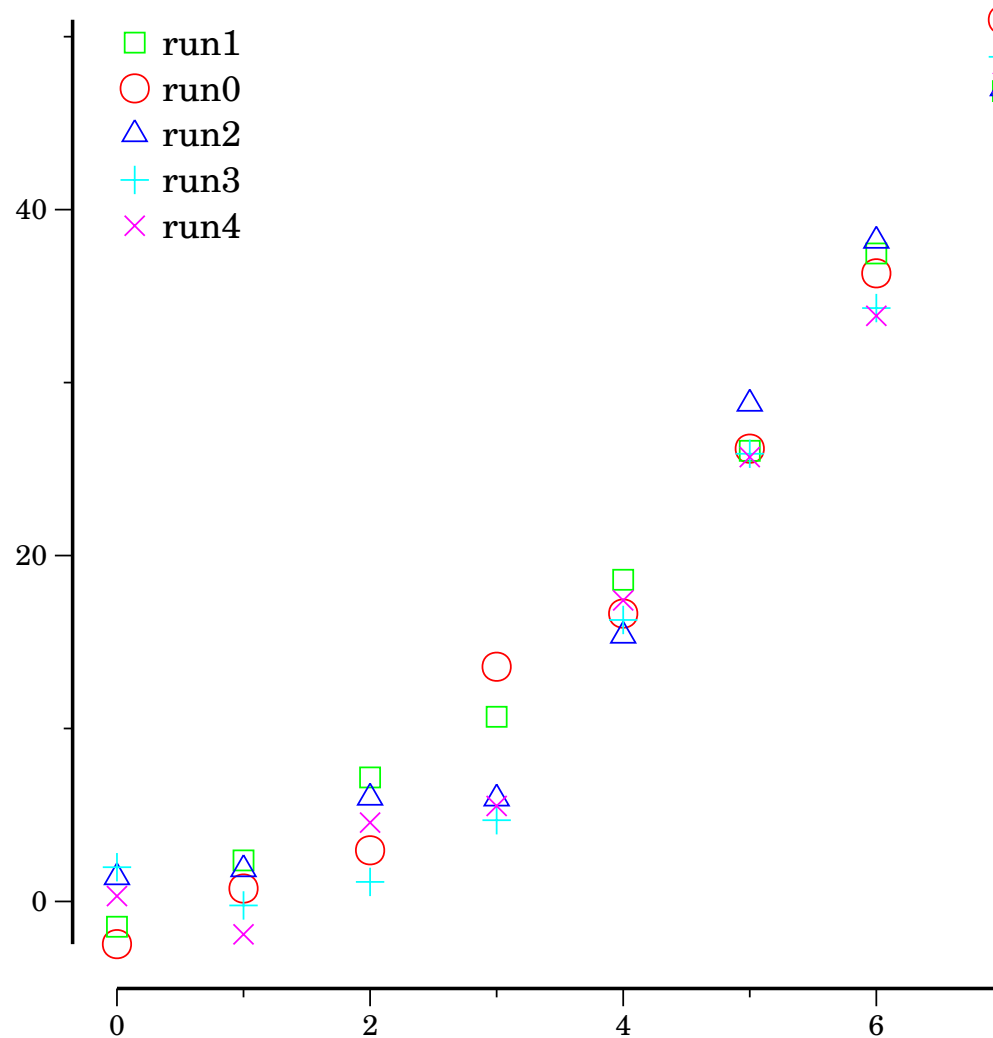
- More Tables
- More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools



Now we can see what our data is telling us

More Tables

Introduction

- Outline
- Why Do We Care About Plots?
- What are Plots For?
- Tables of Data
- Pictures of Data
- **More Tables**
- More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools

	<i>Mystery</i>				<i>Mprime</i>			
	Unpop		FF		Unpop		FF	
task	time	steps	time	steps	time	steps	time	steps
prob-01	0.3	5	0.04	5	0.4	5	0.04	5
prob-02	3.3	8	0.25	10	13.5	8	0.27	10
prob-03	2.1	4	0.08	4	5.9	4	0.09	4
prob-04	-	-	-	-	3.9	9	0.04	10
prob-05	-	-	-	-	19.2	17	-	-
prob-06	-	-	-	-	-	-	-	-
prob-07	-	-	-	-	-	-	-	-
prob-08	-	-	-	-	52.5	10	0.40	10
prob-09	3.3	8	-	-	13.5	8	0.16	10
prob-10	-	-	-	-	79.0	19	-	-
prob-11	1.4	11	0.05	9	2.9	11	0.06	9
prob-12	-	-	-	-	8.0	12	0.20	10
prob-13	370.1	16	-	-	89.3	15	0.16	10
prob-14	162.1	18	-	-	-	-	-	-
prob-15	17.3	6	0.98	8	14.6	6	3.39	8
prob-16	-	-	-	-	25.2	13	0.28	7
prob-17	13.1	5	0.70	4	4.0	5	0.92	4
prob-18	-	-	-	-	-	-	-	-
prob-19	11.8	6	-	-	24.7	6	0.99	9
prob-20	22.5	7	0.41	13	62.8	17	3.11	13
prob-21	-	-	-	-	22.1	11	-	-
prob-22	-	-	-	-	135.7	16	643.19	23
prob-23	-	-	-	-	55.0	18	3.09	14
prob-24	-	-	-	-	24.8	15	2.7	9
prob-25	0.4	4	0.02	4	0.5	4	0.02	4
prob-26	6.0	6	0.85	7	16.4	14	0.16	10
prob-27	3.8	9	0.05	5	2.8	7	0.78	5
prob-28	1.4	9	0.01	7	1.6	11	0.08	5
prob-29	0.9	4	0.06	4	1.5	4	0.30	4
prob-30	20.8	14	0.23	11	17.7	12	1.86	11

More Pictures

Introduction

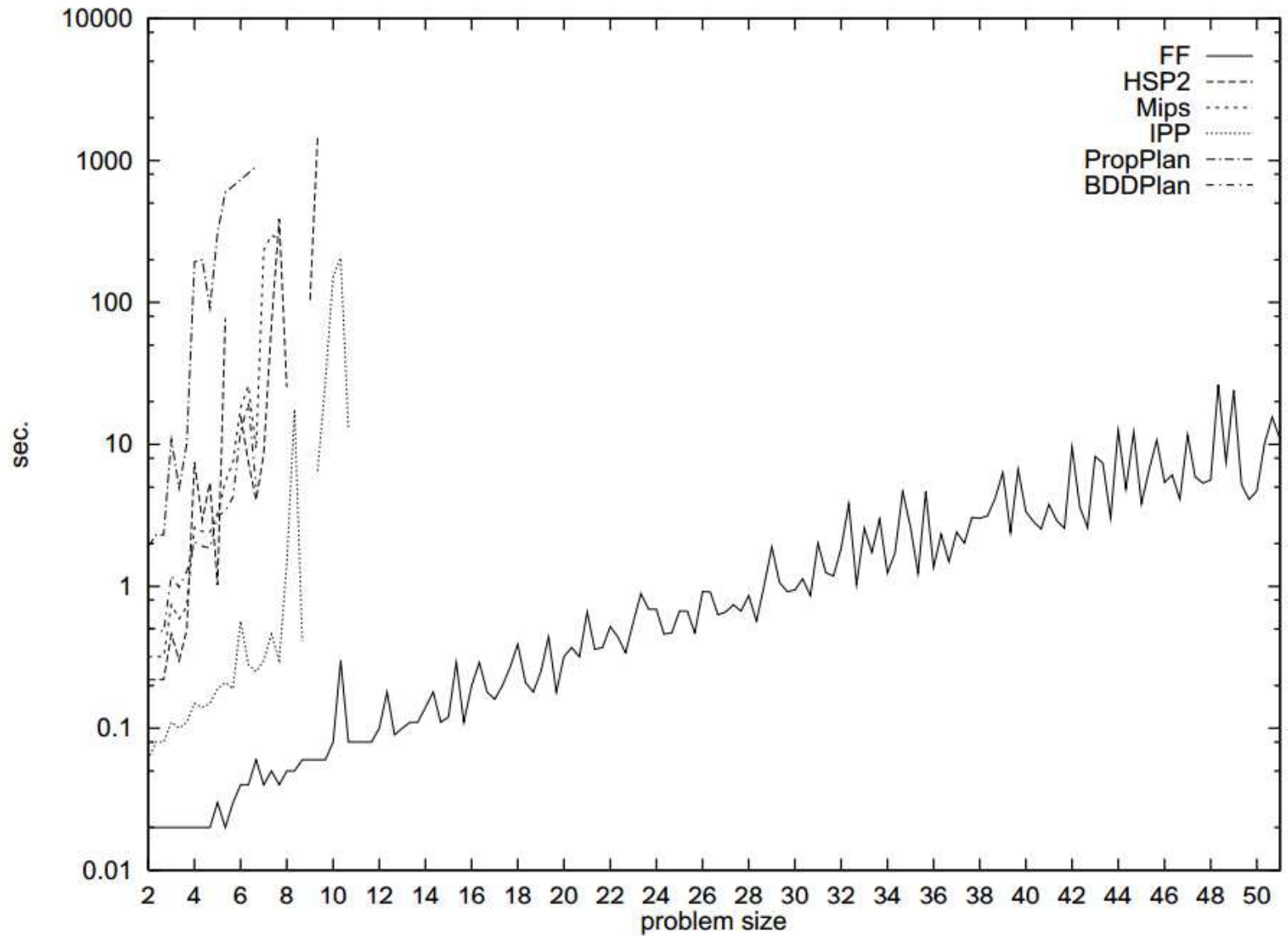
- Outline
- Why Do We Care About Plots?
- What are Plots For?
- Tables of Data
- Pictures of Data
- More Tables
- More Pictures

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools



[Introduction](#)

[Distributions of Values](#)

- Histograms
- Heatmap
- Bin Width
- Comparing
- Box Plots
- Grouped Box Plots
- Log Scales
- Paired Data
- Summary

[Trends in Data](#)

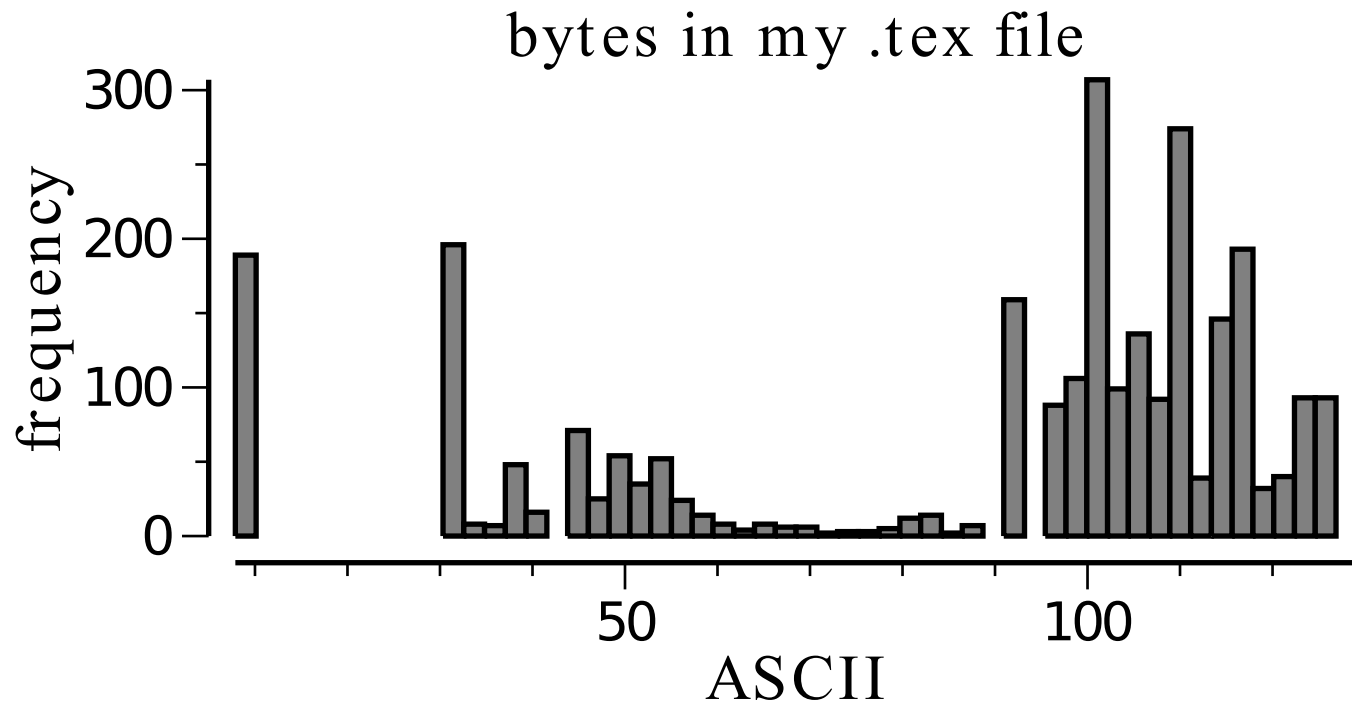
[Simple Plotting Tool](#)

[Other Tools](#)

Distributions of Values

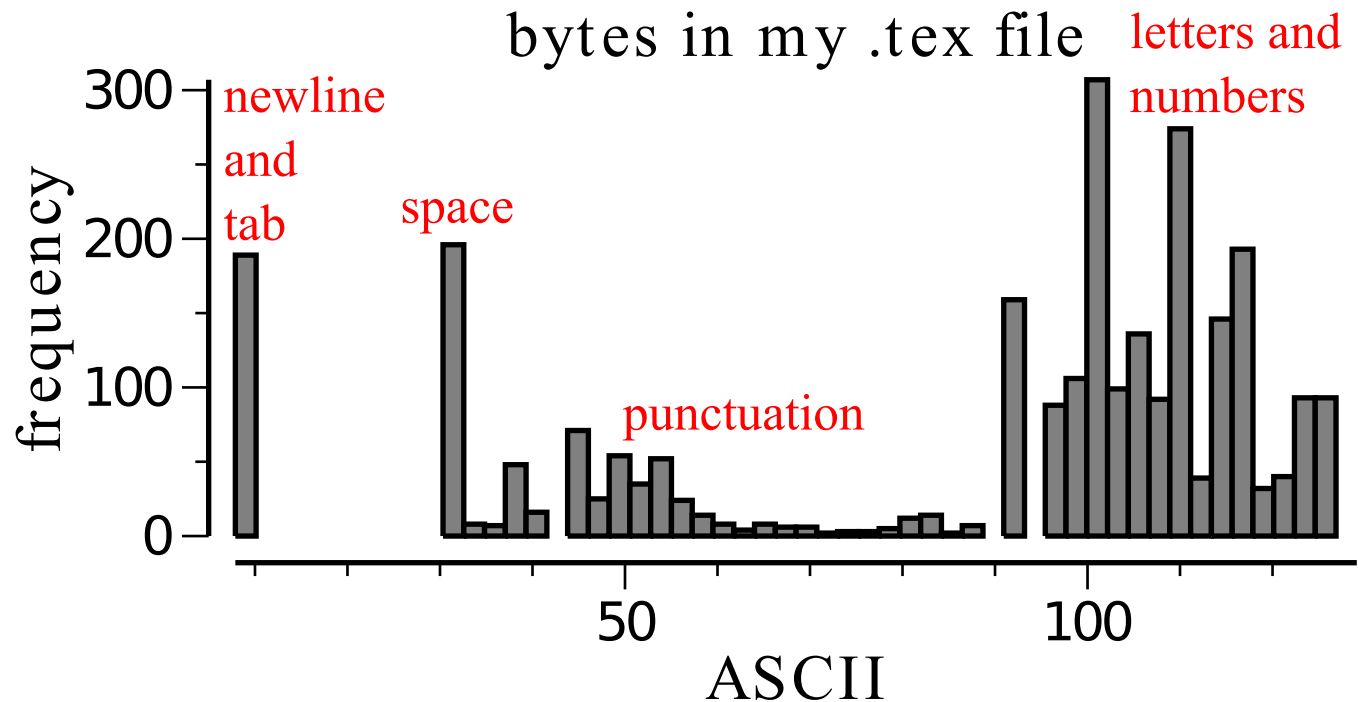
Histograms

Histograms show a distribution of values:



Histograms

Histograms show a distribution of values:



Can quickly show modes—areas of high frequency

2D Histograms: Heatmaps

[Introduction](#)

[Distributions of Values](#)

[Histograms](#)

[Heatmap](#)

[Bin Width](#)

[Comparing](#)

[Box Plots](#)

[Grouped Box Plots](#)

[Log Scales](#)

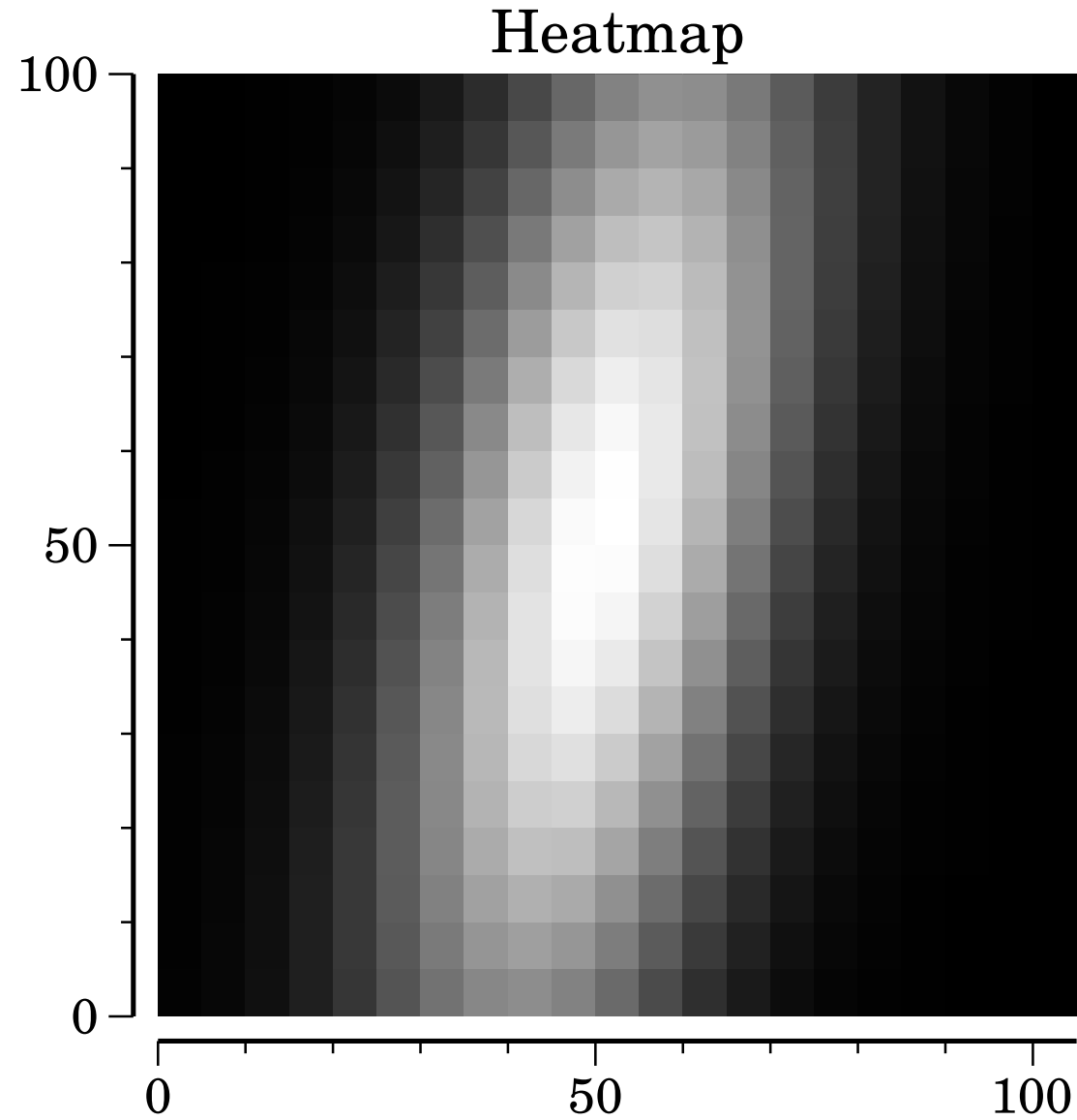
[Paired Data](#)

[Summary](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

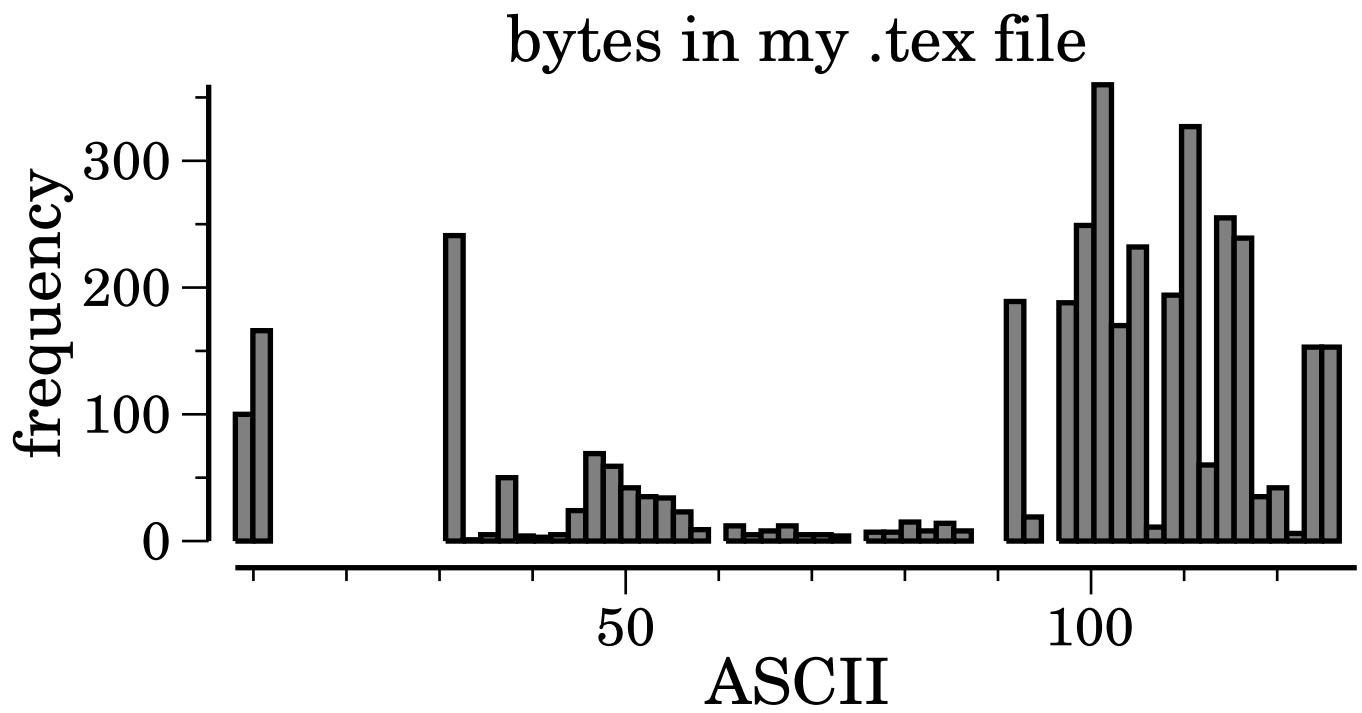
[Other Tools](#)



Bin Width

The width of bins can have a big impact on the histogram

- Introduction
- Distributions of Values
- Histograms
- Heatmap
- Bin Width
- Comparing
- Box Plots
- Grouped Box Plots
- Log Scales
- Paired Data
- Summary
- Trends in Data
- Simple Plotting Tool
- Other Tools

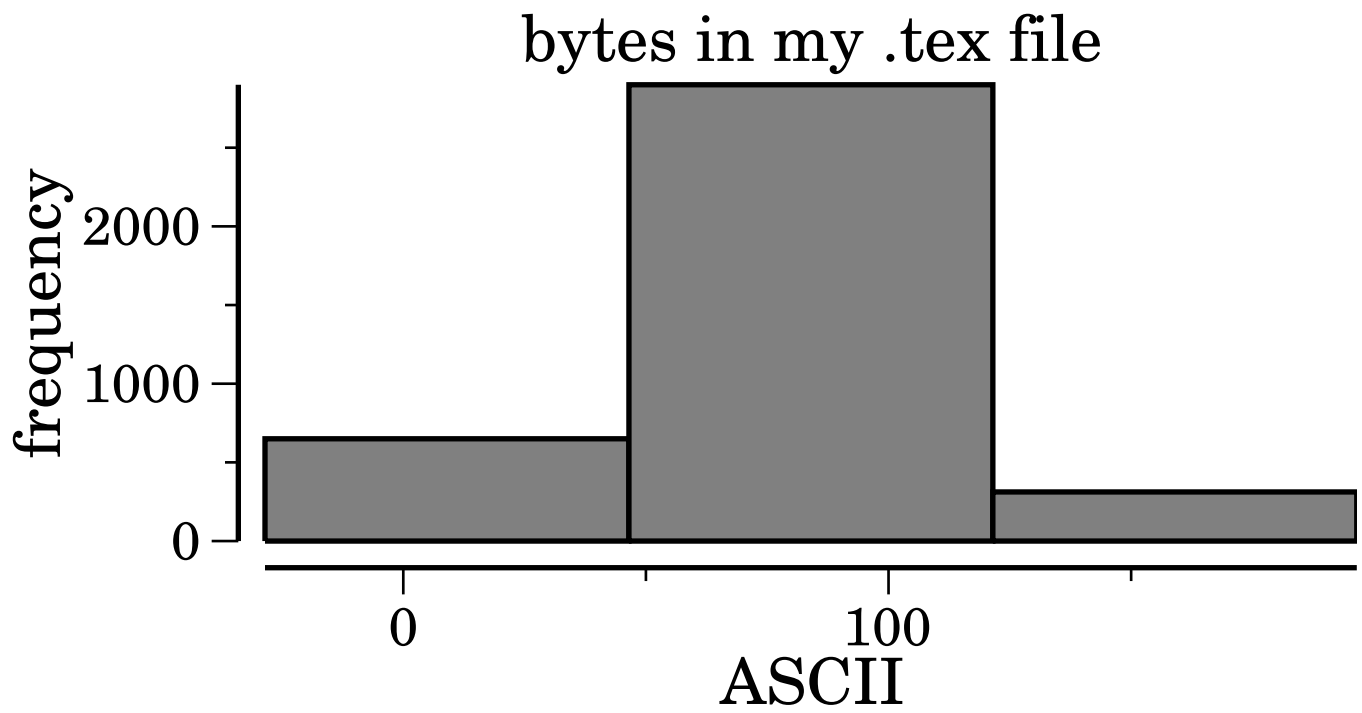


If bins are too big, information can be hidden

Bin Width

The width of bins can have a big impact on the histogram

- Introduction
- Distributions of Values
- Histograms
- Heatmap
- Bin Width
- Comparing
- Box Plots
- Grouped Box Plots
- Log Scales
- Paired Data
- Summary
- Trends in Data
- Simple Plotting Tool
- Other Tools

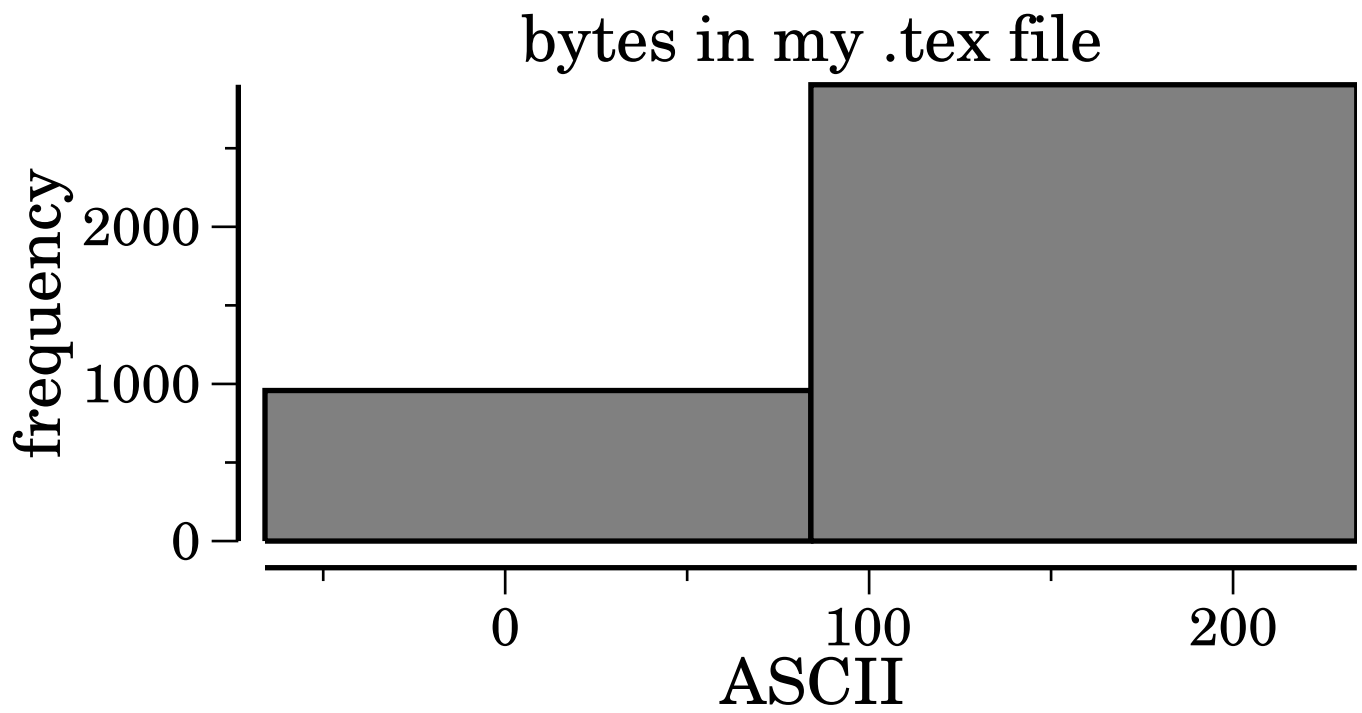


If bins are too big, information can be hidden

Bin Width

The width of bins can have a big impact on the histogram

- Introduction
- Distributions of Values
- Histograms
- Heatmap
- Bin Width
- Comparing
- Box Plots
- Grouped Box Plots
- Log Scales
- Paired Data
- Summary
- Trends in Data
- Simple Plotting Tool
- Other Tools



If bins are too big, information can be hidden

Comparing distributions

[Introduction](#)

[Distributions of Values](#)

■ Histograms

■ Heatmap

■ Bin Width

■ **Comparing**

■ Box Plots

■ Grouped Box Plots

■ Log Scales

■ Paired Data

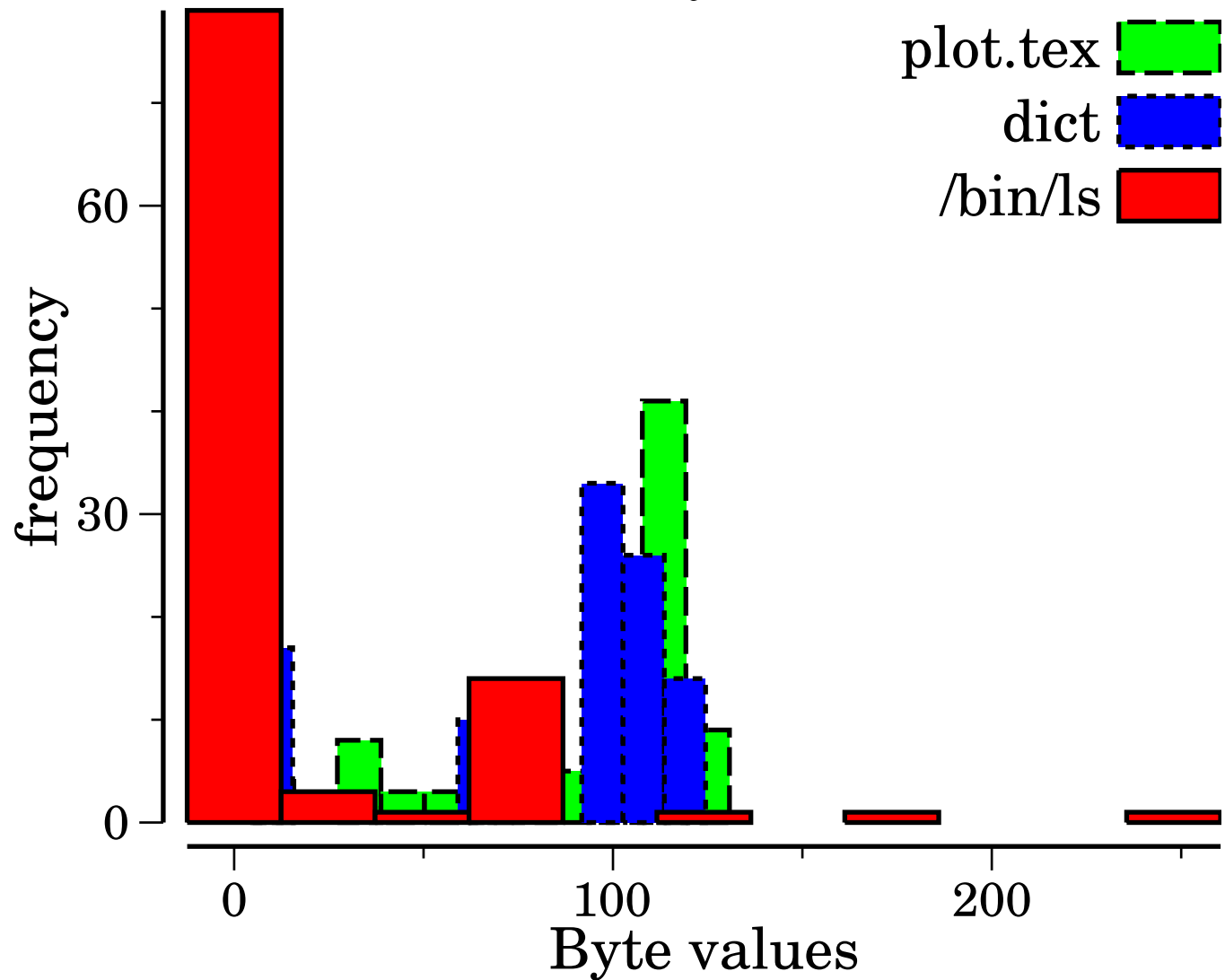
■ Summary

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

first 100 bytes in files



Box Plots

[Introduction](#)

[Distributions of Values](#)

■ Histograms

■ Heatmap

■ Bin Width

■ Comparing

■ **Box Plots**

■ Grouped Box Plots

■ Log Scales

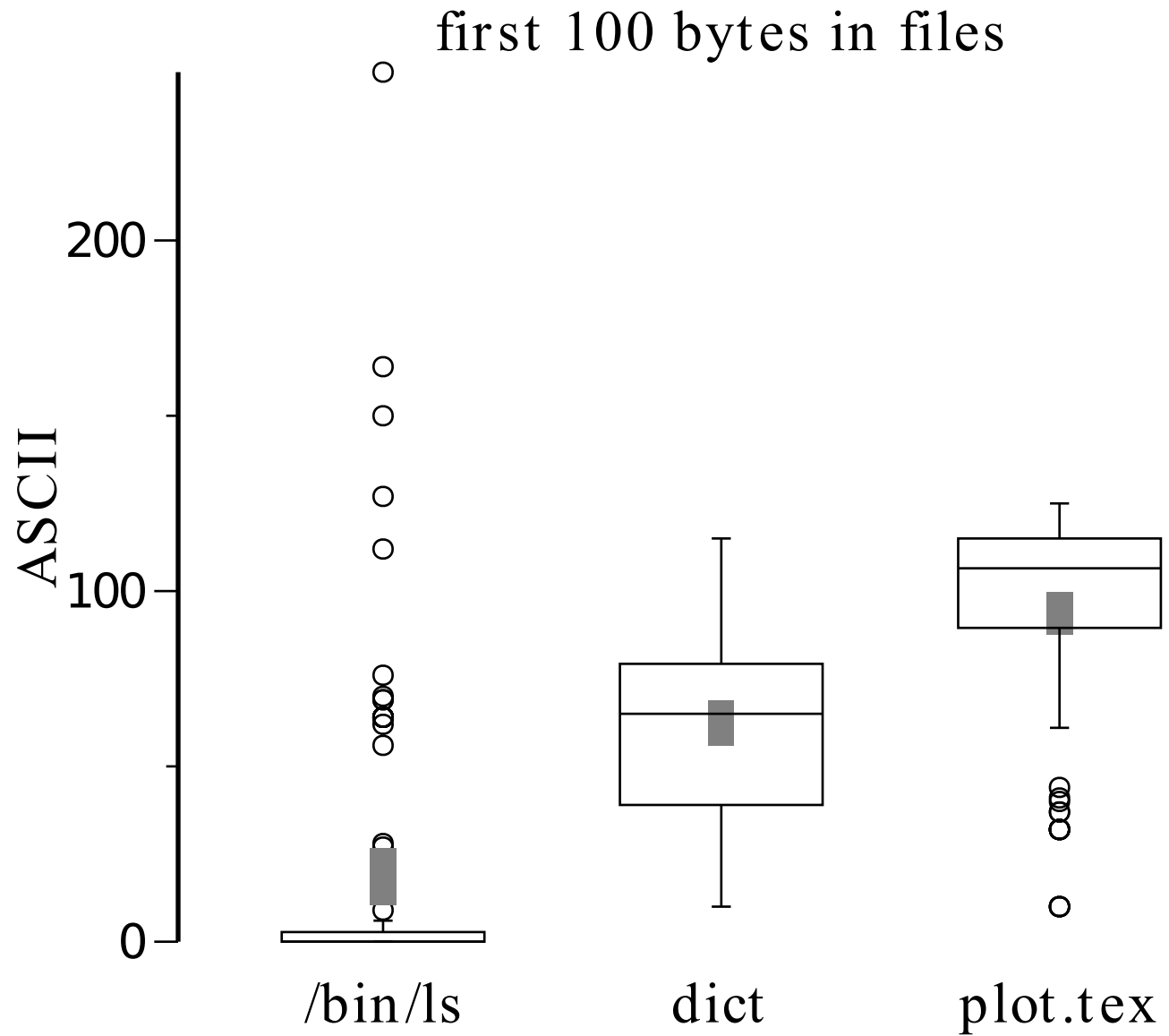
■ Paired Data

■ Summary

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)



Box Plots

Introduction

Distributions of Values

■ Histograms

■ Heatmap

■ Bin Width

■ Comparing

■ **Box Plots**

■ Grouped Box Plots

■ Log Scales

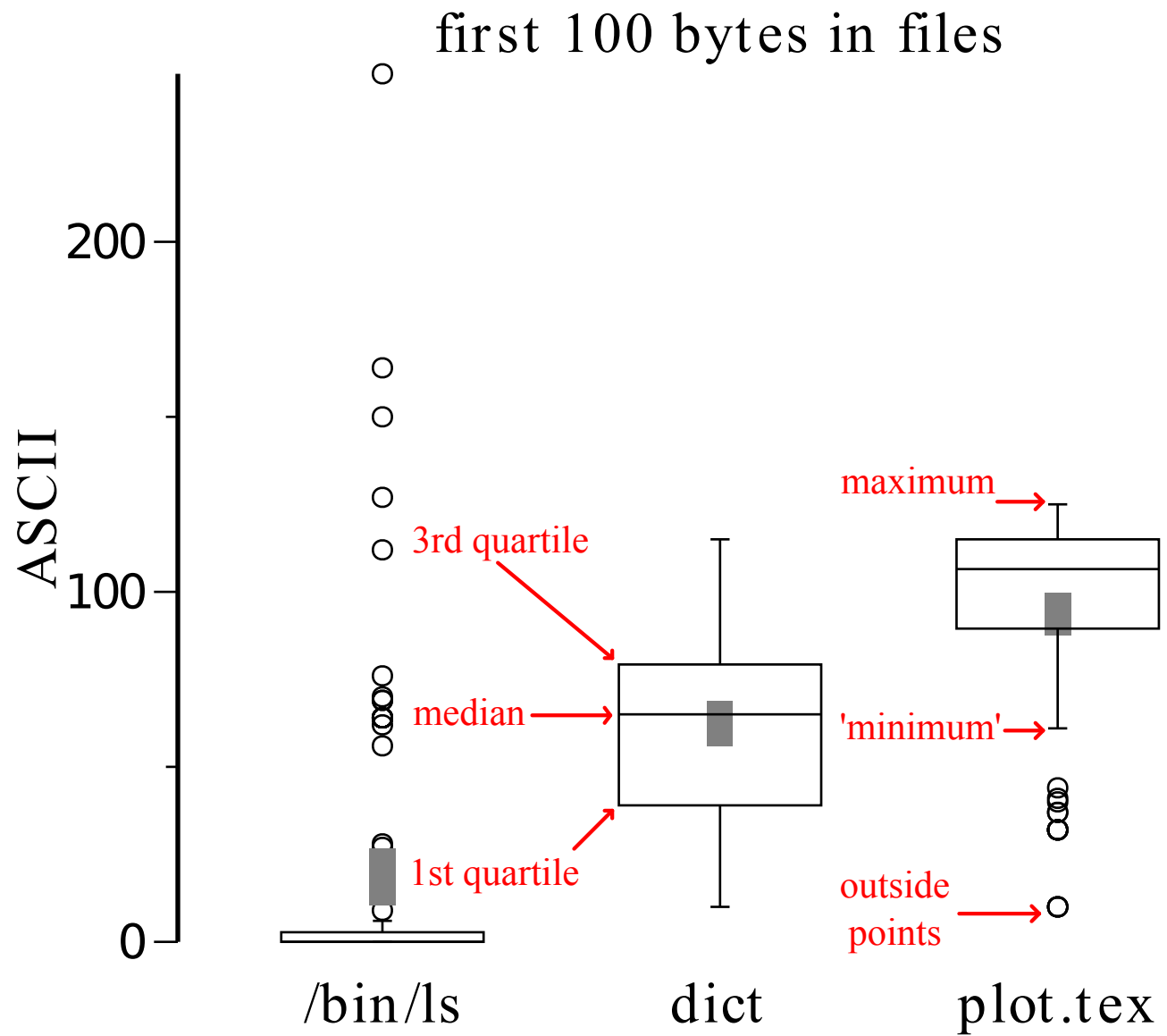
■ Paired Data

■ Summary

Trends in Data

Simple Plotting Tool

Other Tools



Grouped Box Plots

[Introduction](#)

[Distributions of Values](#)

■ Histograms

■ Heatmap

■ Bin Width

■ Comparing

■ Box Plots

■ **Grouped Box Plots**

■ Log Scales

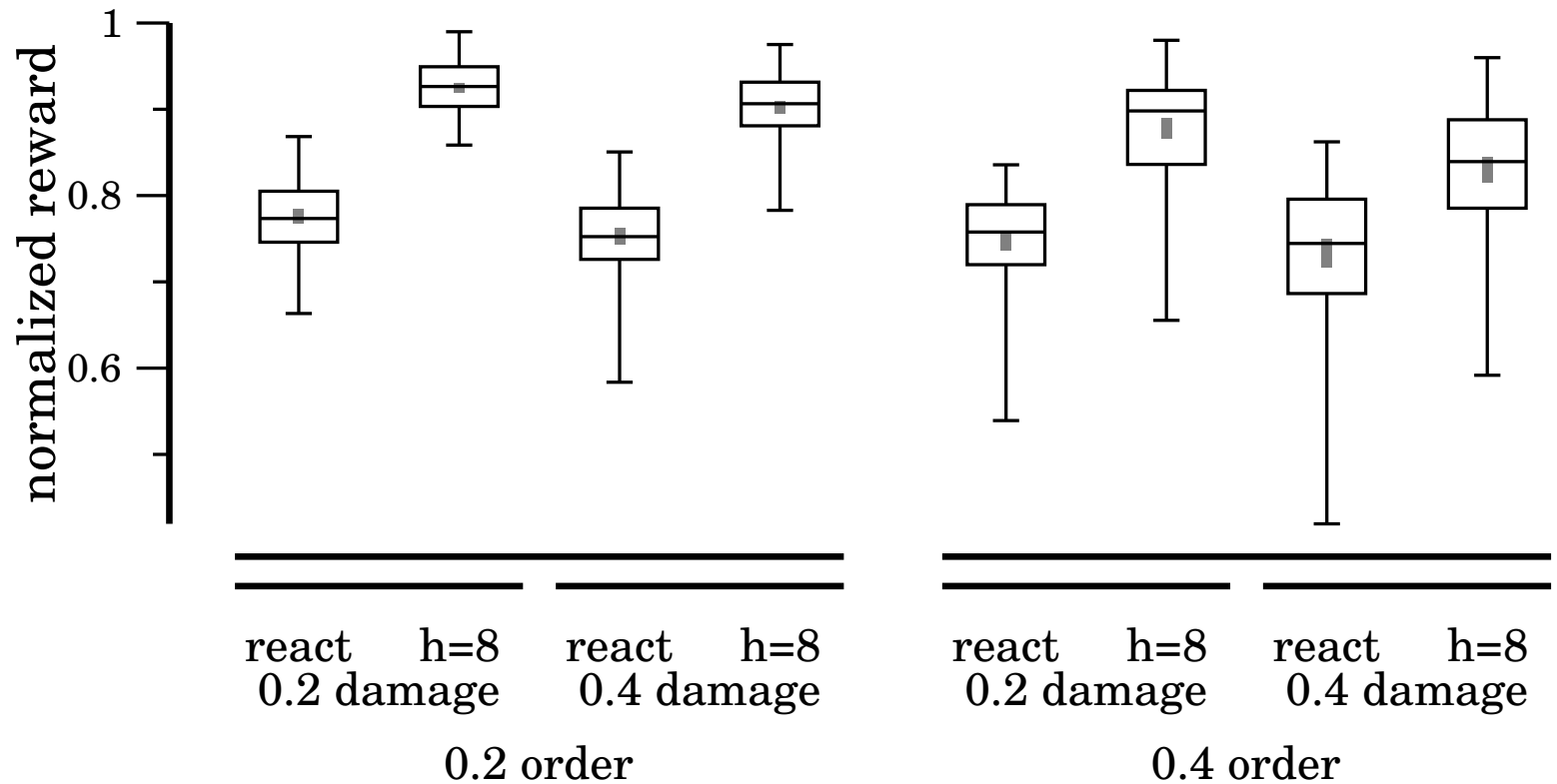
■ Paired Data

■ Summary

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)



Log Scales

Introduction

Distributions of Values

- Histograms
- Heatmap
- Bin Width
- Comparing
- Box Plots
- Grouped Box Plots

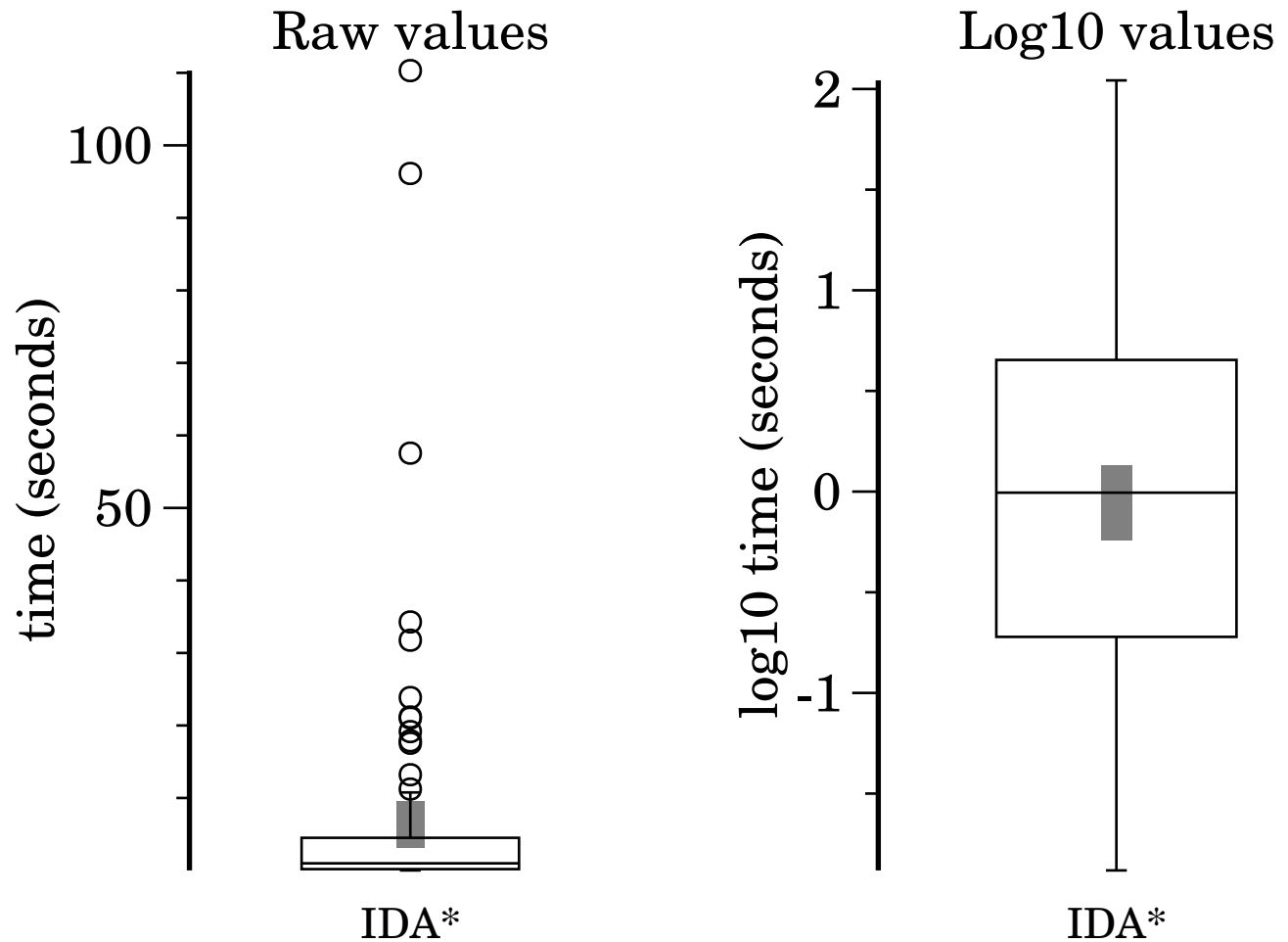
■ Log Scales

- Paired Data
- Summary

Trends in Data

Simple Plotting Tool

Other Tools



log₁₀ values can spread out data for visualization

Paired Data

[Introduction](#)

[Distributions of Values](#)

■ Histograms

■ Heatmap

■ Bin Width

■ Comparing

■ Box Plots

■ Grouped Box Plots

■ Log Scales

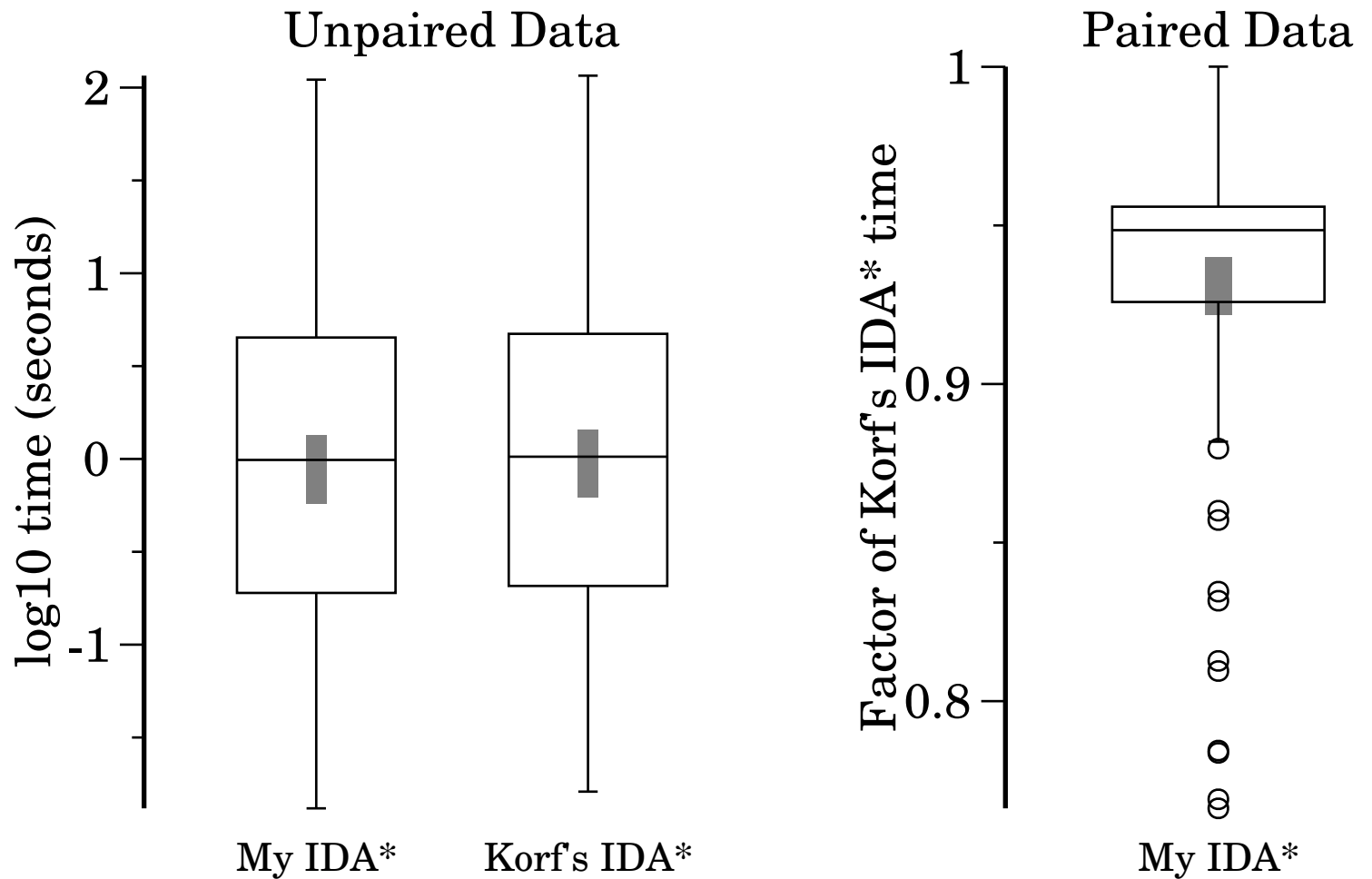
■ Paired Data

■ Summary

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)



Paired data—show instance-by-instance differences

Summary

[Introduction](#)

[Distributions of Values](#)

■ [Histograms](#)

■ [Heatmap](#)

■ [Bin Width](#)

■ [Comparing](#)

■ [Box Plots](#)

■ [Grouped Box Plots](#)

■ [Log Scales](#)

■ [Paired Data](#)

■ [Summary](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

- Histograms easily show data distributions
 - ◆ Careful when choosing bin widths
 - ◆ Histograms are difficult to compare
- Heatmaps are like histograms for 2D data
- Box plots make comparing distributions easy
 - ◆ Grouped box plots can help to show trends
- Log scales can help spread out data for visualization
- Paired data is more powerful

Introduction

Distributions of Values

Trends in Data

- Lines
- Lines and Error Bars
- Scatter Plots
- Confidence Intervals
- More Scatter Plots
- Scatters and Lines
- More Logs and Paired Data
- Summary

Simple Plotting Tool

Other Tools

Trends in Data

Lines

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

Lines

Lines and Error Bars

Scatter Plots

Confidence Intervals

More Scatter Plots

Scatters and Lines

More Logs and Paired Data

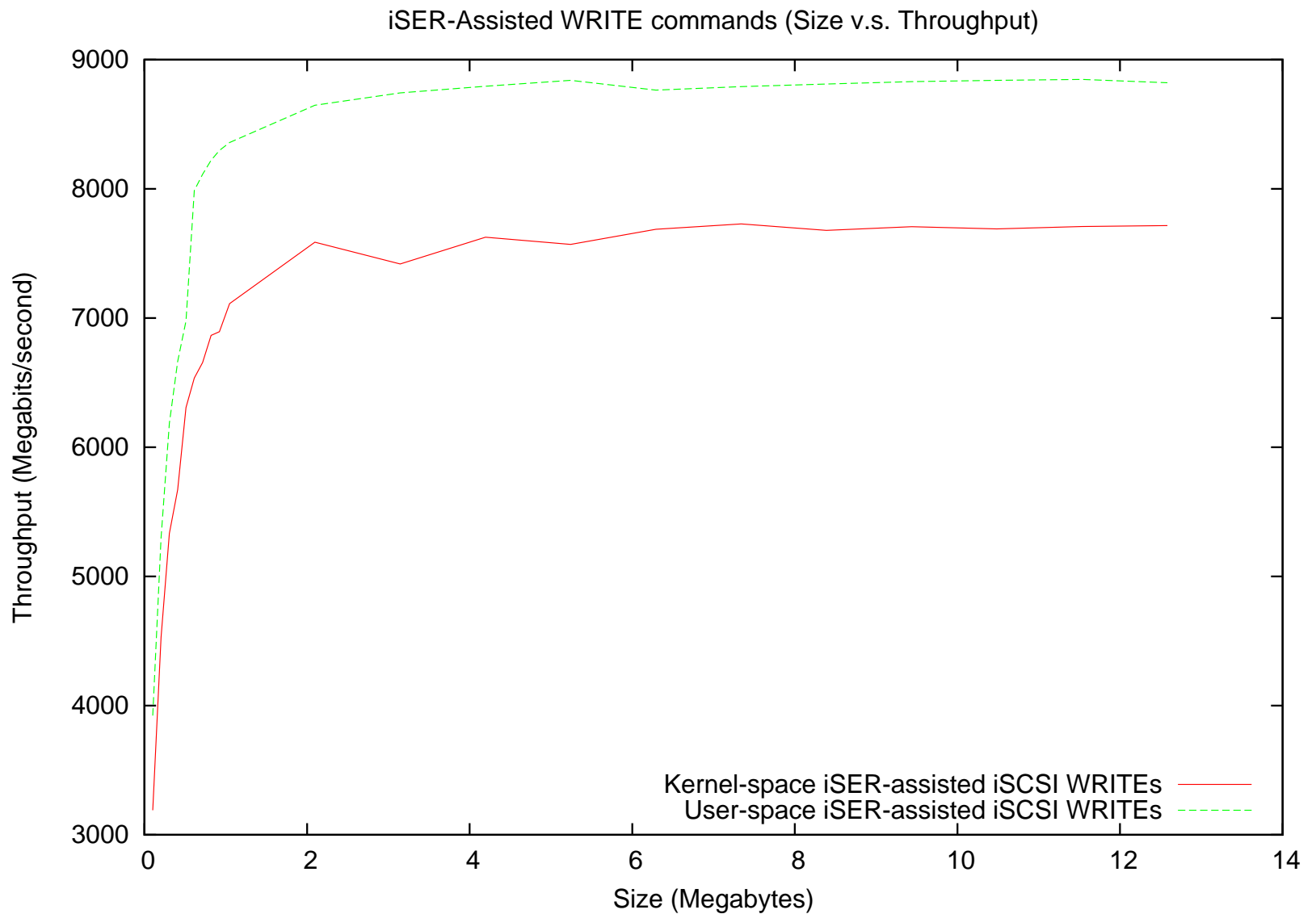
Summary

[Simple Plotting Tool](#)

[Other Tools](#)

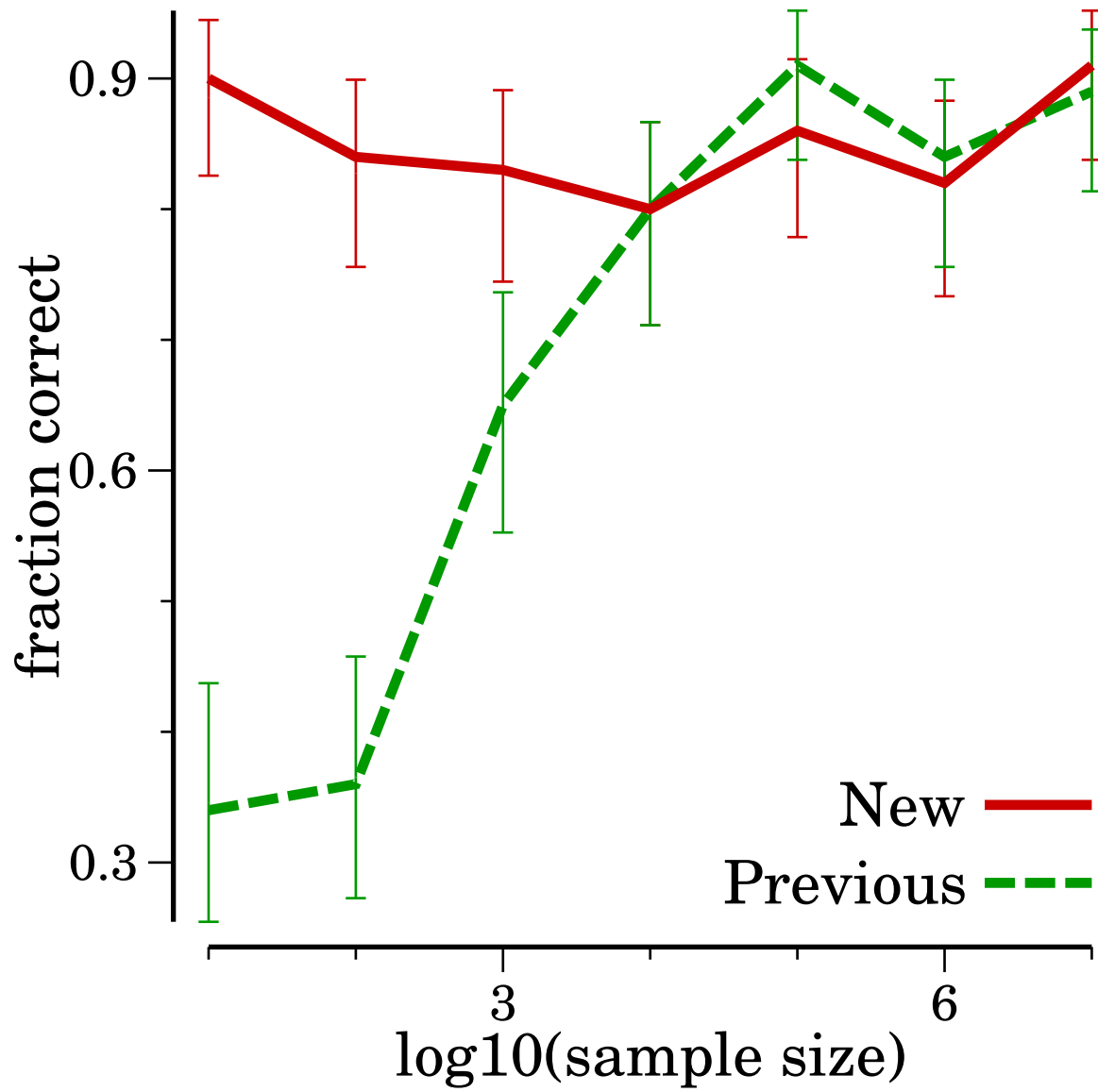
[Simple Plotting Tool](#)

[Other Tools](#)



Lines and Error Bars

- Introduction
- Distributions of Values
- Trends in Data
 - Lines
 - Lines and Error Bars
 - Scatter Plots
 - Confidence Intervals
 - More Scatter Plots
 - Scatters and Lines
 - More Logs and Paired Data
 - Summary
- Simple Plotting Tool
- Other Tools



Scatter Plots

Introduction

Distributions of Values

Trends in Data

■ Lines
■ Lines and Error Bars

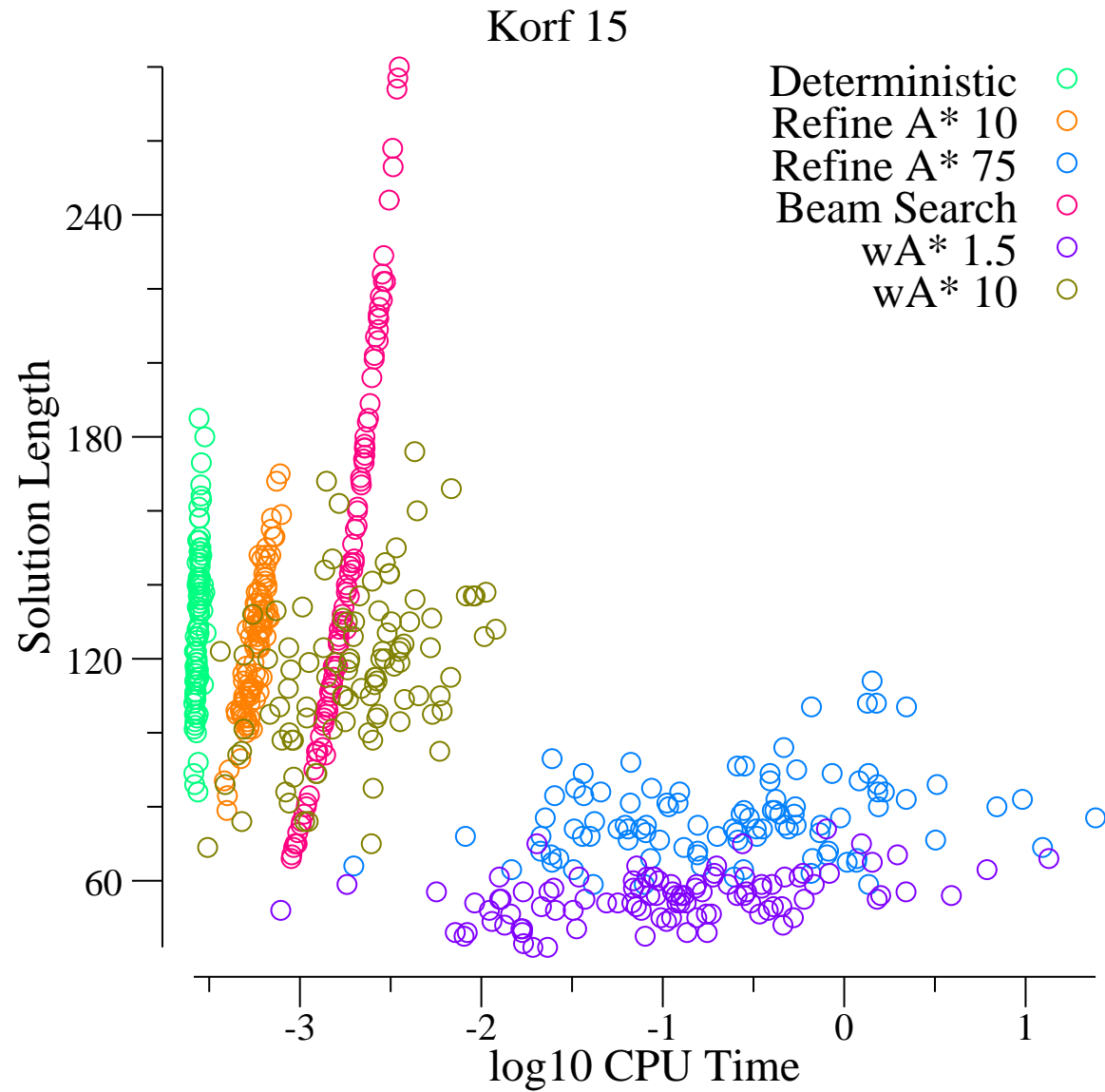
■ Scatter Plots

■ Confidence Intervals
■ More Scatter Plots

■ Scatters and Lines
■ More Logs and Paired Data
■ Summary

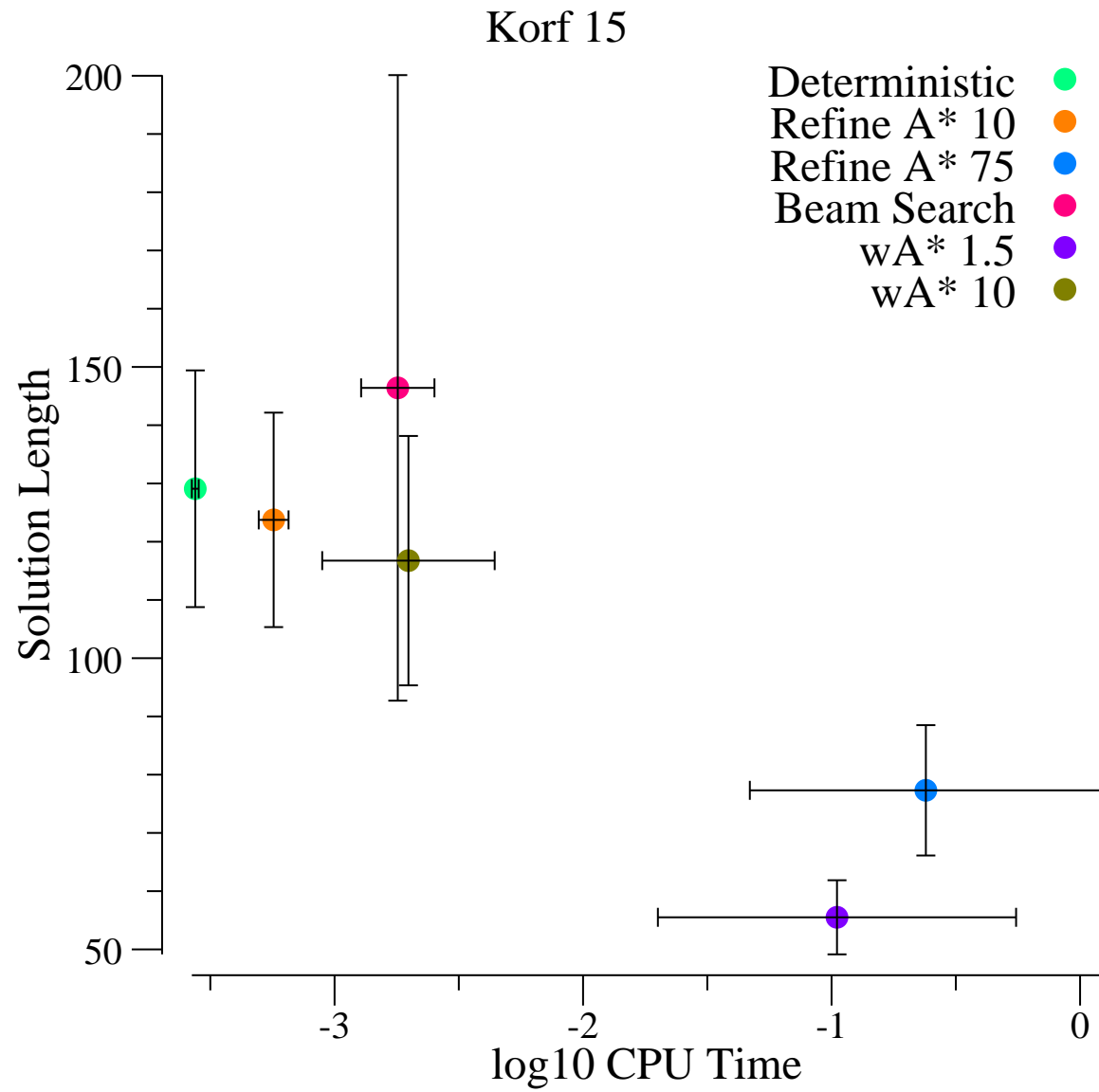
Simple Plotting Tool

Other Tools



Confidence Intervals

- Introduction
- Distributions of Values
- Trends in Data
- Lines
- Lines and Error Bars
- Scatter Plots
- Confidence Intervals
- More Scatter Plots
- Scatters and Lines
- More Logs and Paired Data
- Summary
- Simple Plotting Tool
- Other Tools



More Scatter Plots

Introduction

Distributions of Values

Trends in Data

■ Lines
■ Lines and Error Bars

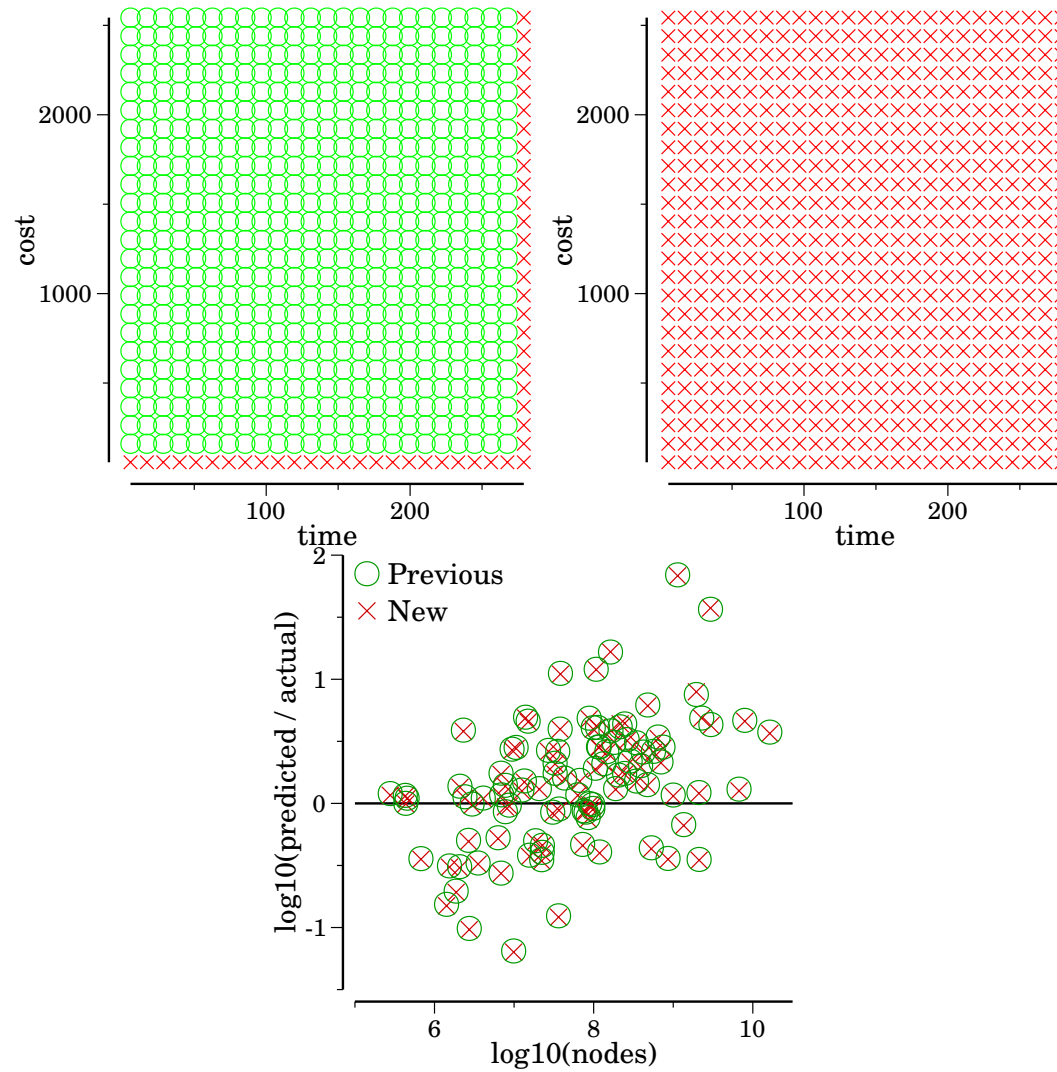
■ Scatter Plots
■ Confidence Intervals

■ More Scatter Plots

■ Scatters and Lines
■ More Logs and Paired Data
■ Summary

Simple Plotting Tool

Other Tools



Scatters and Lines

Introduction

Distributions of Values

Trends in Data

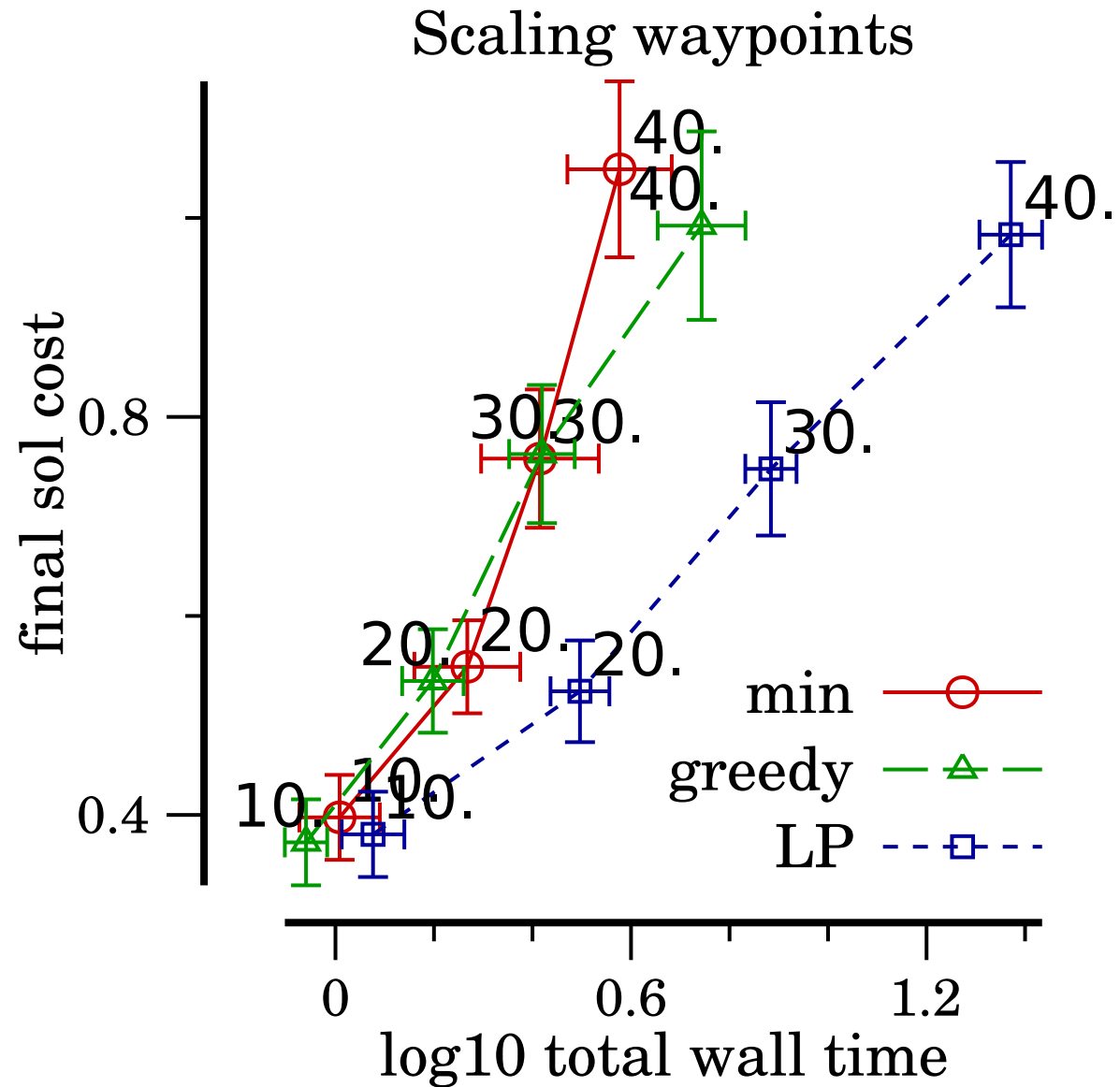
- Lines
- Lines and Error Bars
- Scatter Plots
- Confidence Intervals
- More Scatter Plots

■ Scatters and Lines

- More Logs and Paired Data
- Summary

Simple Plotting Tool

Other Tools



More Logs and Paired Data

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

■ Lines
■ Lines and Error Bars

■ Scatter Plots
■ Confidence Intervals

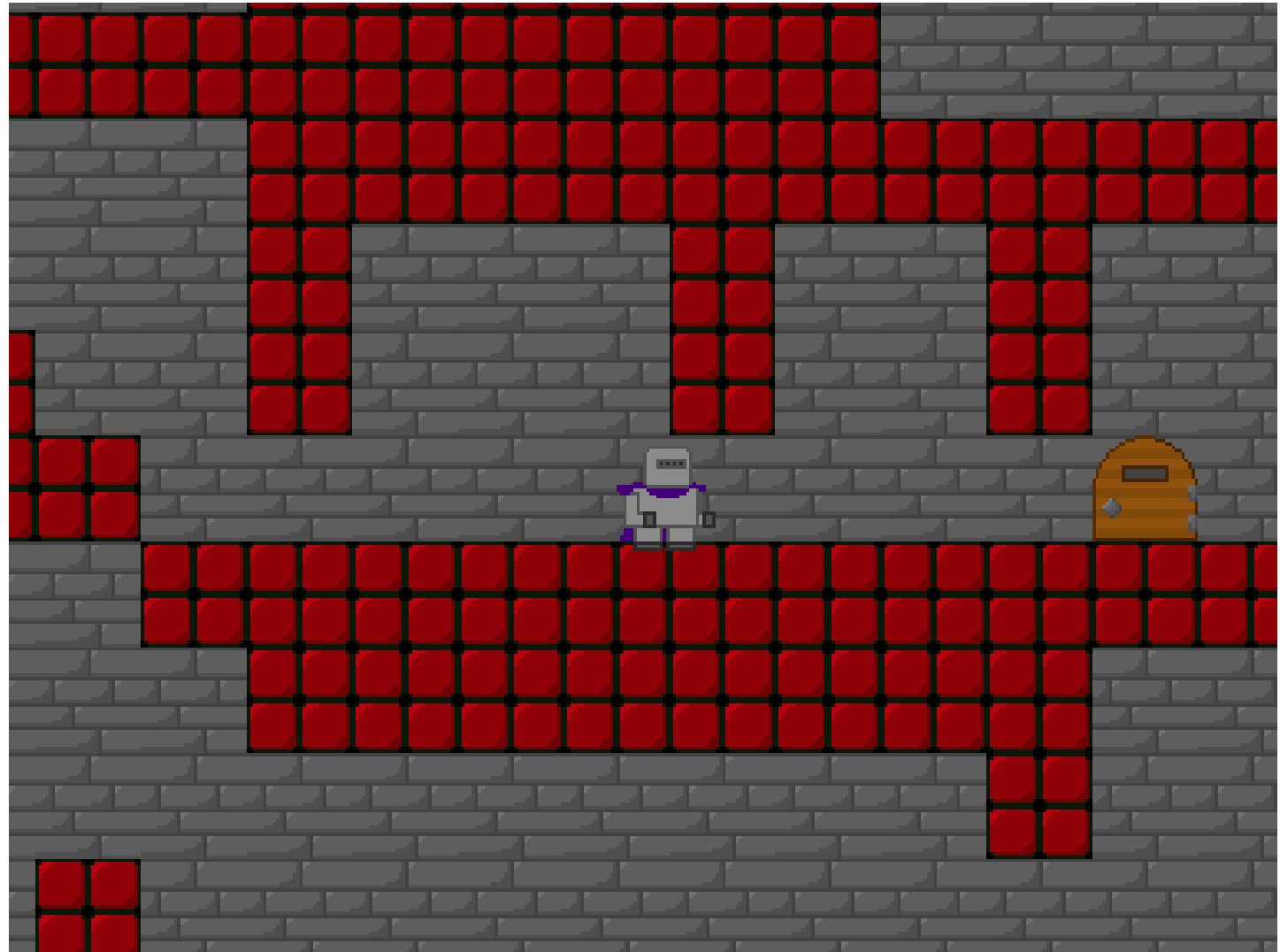
■ More Scatter Plots

■ Scatters and Lines
■ **More Logs and Paired Data**

■ Summary

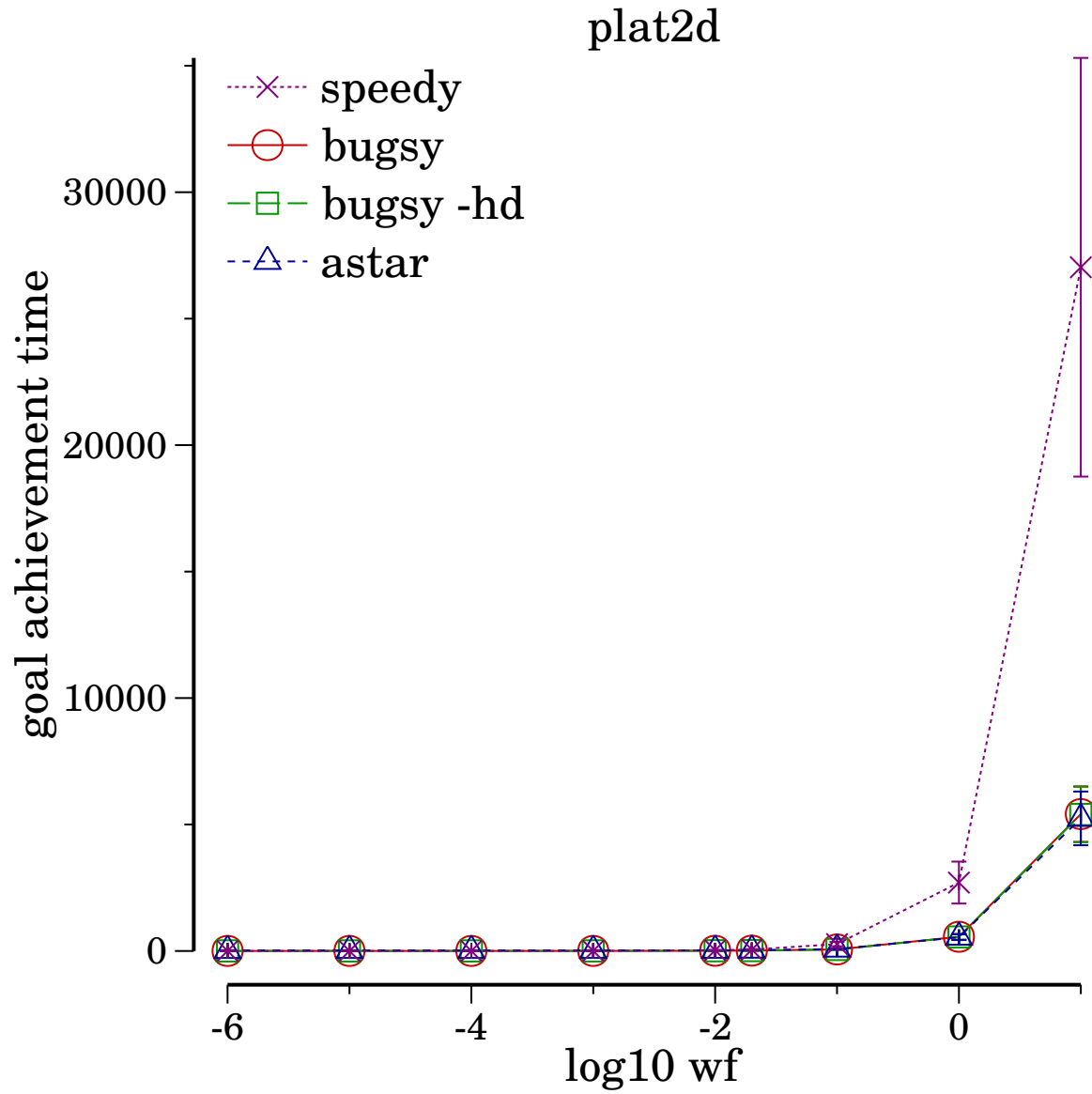
[Simple Plotting Tool](#)

[Other Tools](#)



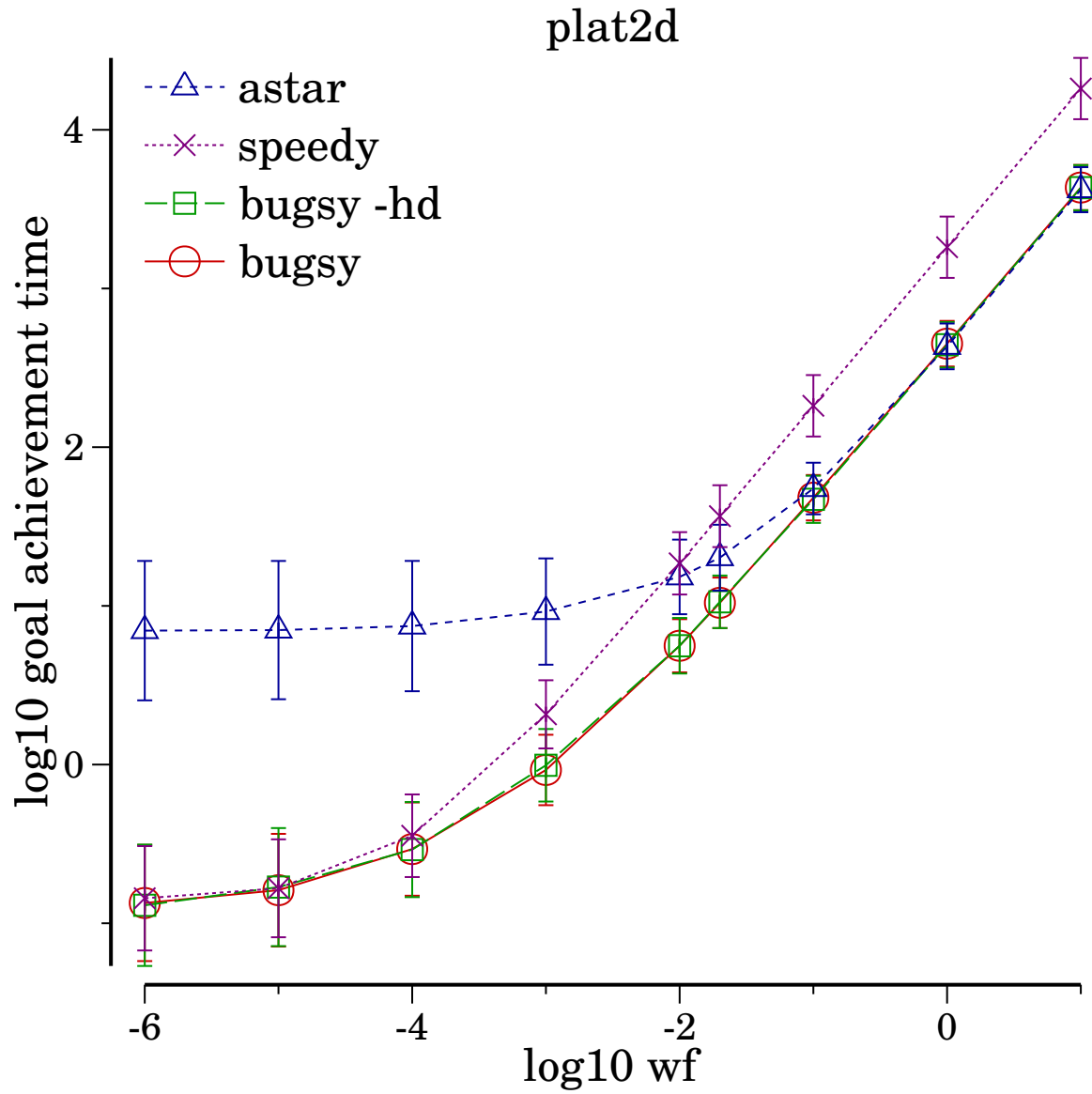
More Logs and Paired Data

- Introduction
- Distributions of Values
- Trends in Data
 - Lines
 - Lines and Error Bars
 - Scatter Plots
 - Confidence Intervals
 - More Scatter Plots
 - Scatters and Lines
 - More Logs and Paired Data**
 - Summary
- Simple Plotting Tool
- Other Tools



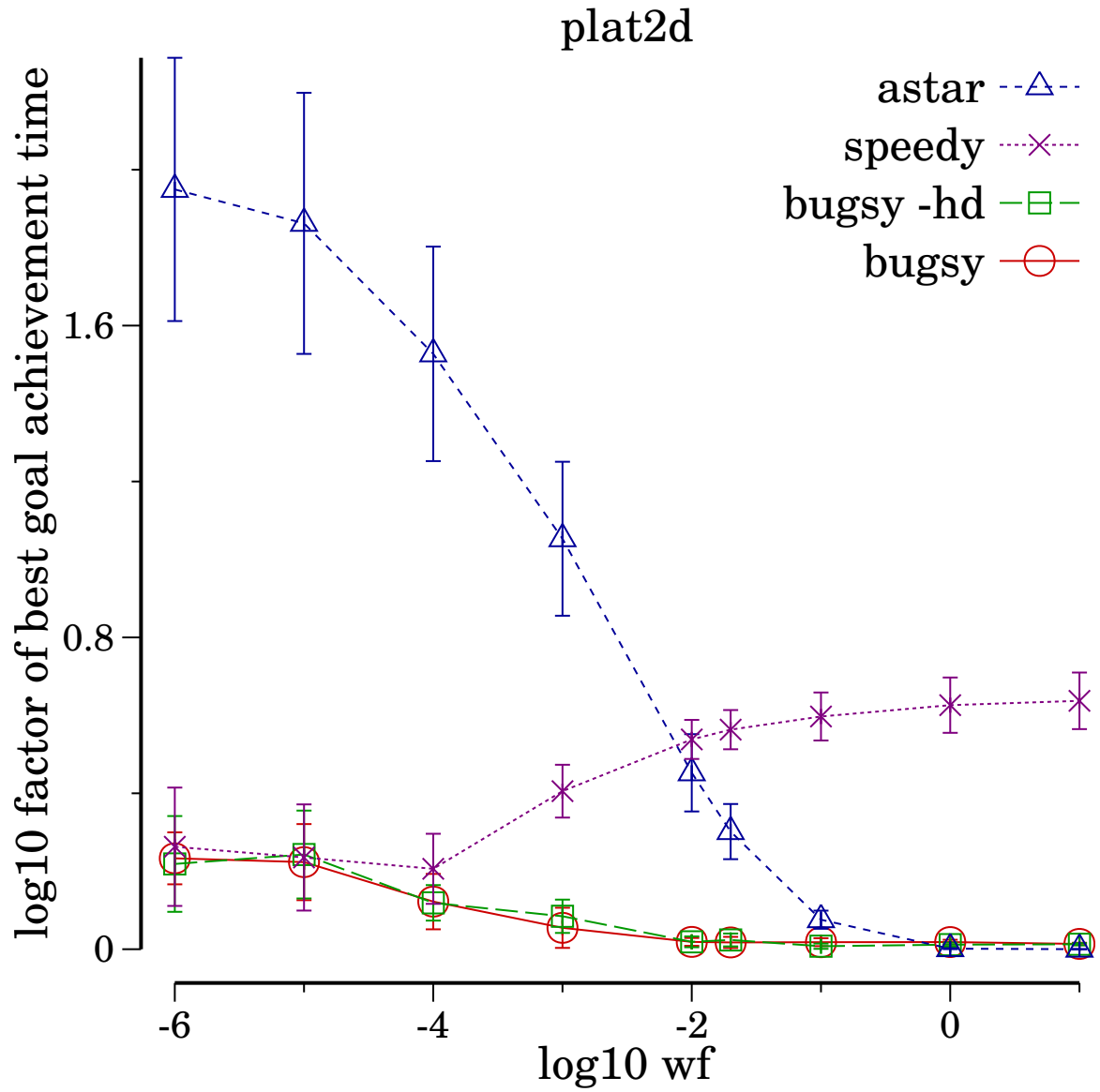
More Logs and Paired Data

- Introduction
- Distributions of Values
- Trends in Data
 - Lines
 - Lines and Error Bars
 - Scatter Plots
 - Confidence Intervals
 - More Scatter Plots
 - Scatters and Lines
 - More Logs and Paired Data**
 - Summary
- Simple Plotting Tool
- Other Tools



More Logs and Paired Data

- Introduction
- Distributions of Values
- Trends in Data
 - Lines
 - Lines and Error Bars
 - Scatter Plots
 - Confidence Intervals
 - More Scatter Plots
 - Scatters and Lines
 - More Logs and Paired Data**
 - Summary
- Simple Plotting Tool
- Other Tools



Summary

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

■ Lines

■ Lines and Error Bars

■ Scatter Plots

■ Confidence Intervals

■ More Scatter Plots

■ Scatters and Lines

■ More Logs and Paired Data

■ Summary

[Simple Plotting Tool](#)

[Other Tools](#)

- Lines easily show trends in data
- Scatter plots can show trends in points
- Use confidence intervals—or some measure of variance
- Logs can be helpful here too
- Paired data is always better

Introduction

Distributions of Values

Trends in Data

Simple Plotting Tool

- What is It?
- Why a New Tool?
- Spread Sheets
- Benefits of SPT

Other Tools

Simple Plotting Tool

What is It?

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

■ [What is It?](#)

■ [Why a New Tool?](#)


■ [Spread Sheets](#)

■ [Benefits of SPT](#)

[Other Tools](#)

Simple Plotting Tool—SPT

<http://www.cs.unh.edu/~eaburns/spt>

- An(other) open source plotting tool
- Created by the UNH artificial intelligence group
- Easy to create many useful types of plots
- An Objective Caml  API
- A simple lisp-like input language

Why Make a New Plotting Tool?

Introduction

Distributions of Values

Trends in Data

Simple Plotting Tool

■ What is It?

■ Why a New Tool?

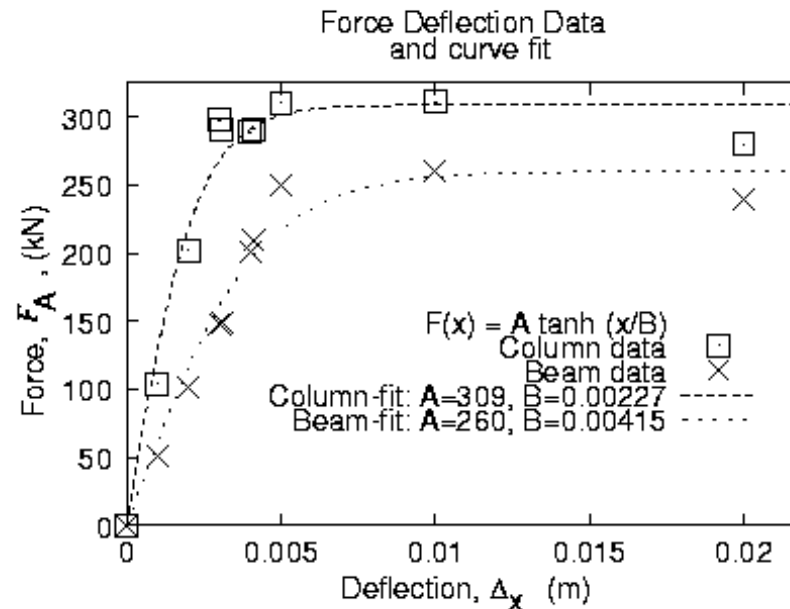
■ Spread Sheets

■ Benefits of SPT

Other Tools

Spread sheets a lot of **manual work**

GNU plot is ugly (in my opinion, and Wheeler's too)



Matplotlib better—still draws ticks inside and data on the axes

R not too bad!

Spread Sheets

Introduction

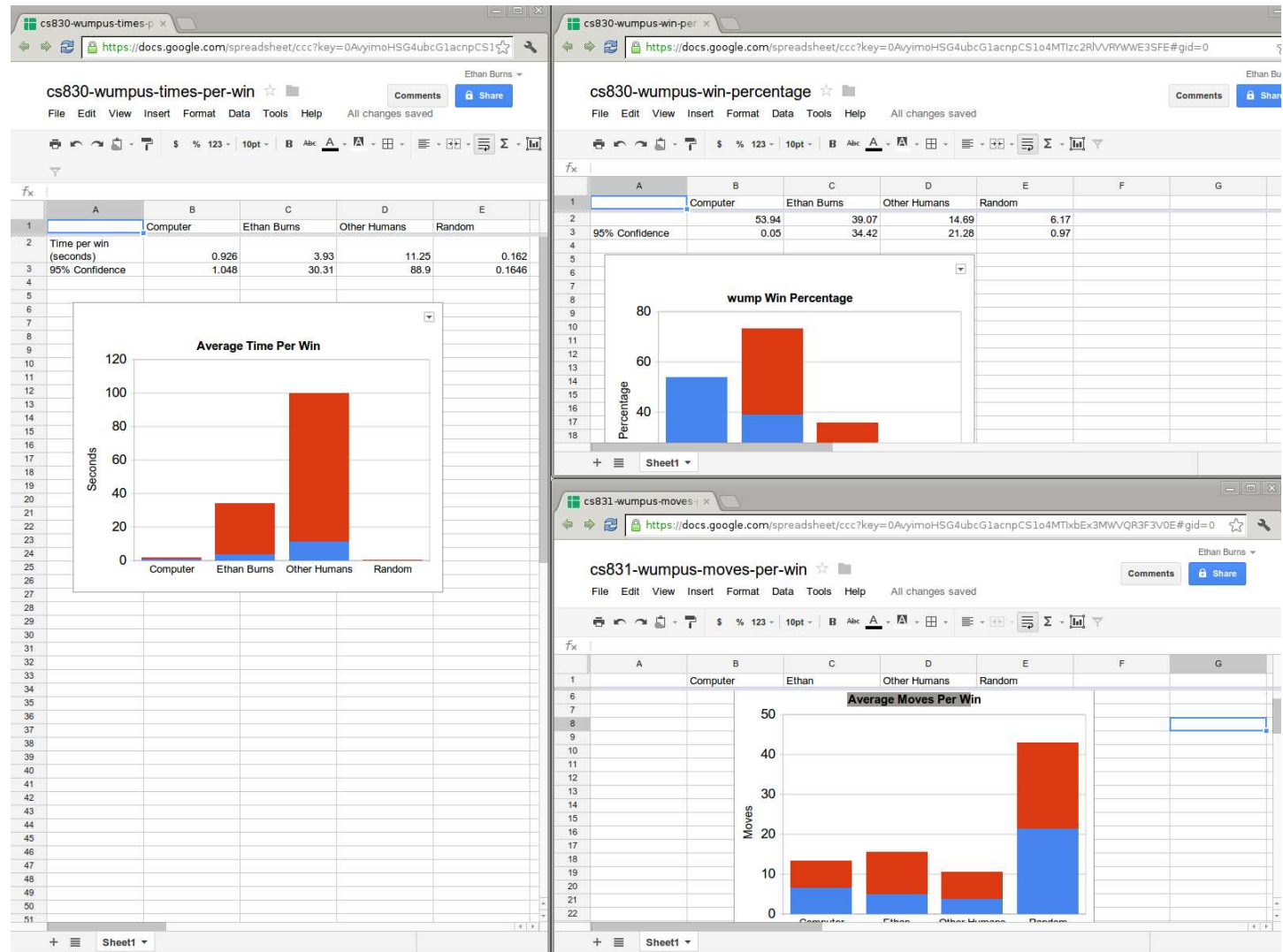
Distributions of Values

Trends in Data

Simple Plotting Tool

- What is It?
- Why a New Tool?
- Spread Sheets
- Benefits of SPT

Other Tools



Benefits of SPT

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

■ What is It?

■ Why a New Tool?

■ Spread Sheets

■ **Benefits of SPT**

[Other Tools](#)

- Professional quality plots (not cartoony)
 - ◆ Greater data-ink ratio (Edward Tufte)
 - Axes do not box in the data
 - Not too many tick marks
 - ◆ Axes are not drawn over the data
- Very easy to make box plots
 - With confidence intervals
 - Grouped box plots too
- Lines and scatters with 95% confidence intervals
- Histograms and heatmaps from $x,y(,z)$ tuples

Introduction

Distributions of Values

Trends in Data

Simple Plotting Tool

Other Tools

- Results Database
- Alternatives
- Plotinum

Other Tools

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

[Results Database](#)

[Alternatives](#)

[Plotinum](#)

I store my results in a simple database called RDB—Results DataBase (or is it Ruml DataBase?)

- A simple file-system-based database
- Easy to find data files given a set of key=value pairs
- Has an OCaml API, a C++ API, and shell scripting support
- Simple data files: key=value, or key=multi-value pairs
- **Framework** connecting RDB → OCaml → SPT

Other More Standard Alternatives

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

[Other Tools](#)

■ Results Database

■ Alternatives

■ Plotinum

■ MongoDB

■ CouchDB

■ SQLite

Plotinum: My Latest Plotting Tool

[Introduction](#)

[Distributions of Values](#)

[Trends in Data](#)

[Simple Plotting Tool](#)

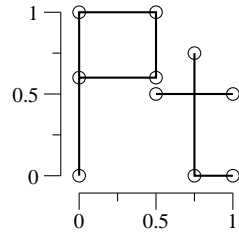
[Other Tools](#)

■ [Results Database](#)

■ [Alternatives](#)

■ [Plotinum](#)

- Another-nother open source plotting tool



<http://code.google.com/p/plotinum>



- Written in Go

golang.org, check it out!

- Simpler, more flexible, and more extendable than SPT
...but, a little less complete at the moment