

# The gap between tools for learning and for doing statistics

Amelia McNamara

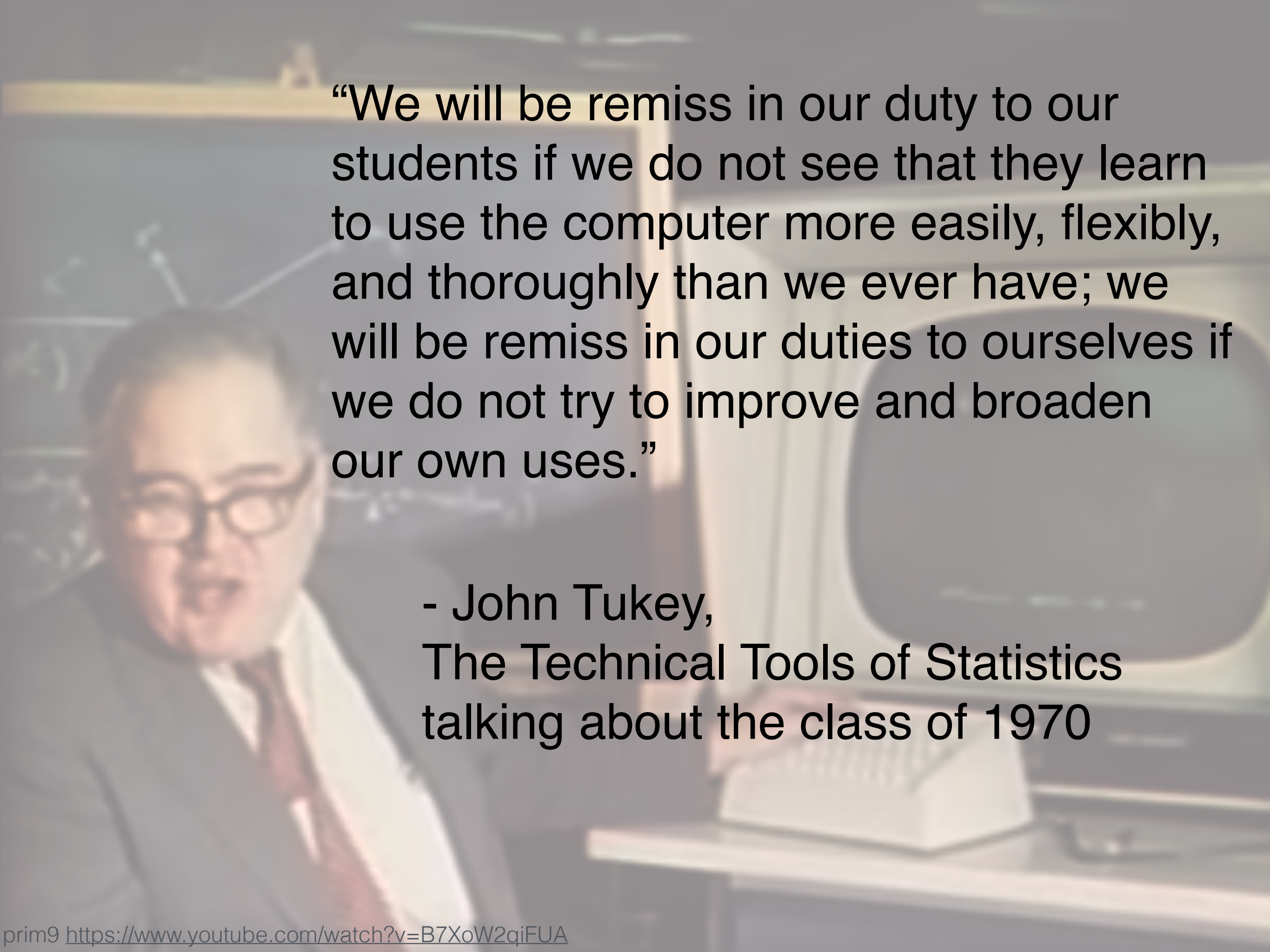
Visiting Assistant Professor of Statistical and Data Sciences

Smith College

[@AmeliaMN](#)



- Tools for teaching and learning statistics
- Tools for doing statistics
- Beginning to bridge the gap
- The future



“We will be remiss in our duty to our students if we do not see that they learn to use the computer more easily, flexibly, and thoroughly than we ever have; we will be remiss in our duties to ourselves if we do not try to improve and broaden our own uses.”

- John Tukey,  
The Technical Tools of Statistics  
talking about the class of 1970

Tools for  
teaching and  
learning  
statistics



### Rossman/Chance Applet Collection

#### Data Analysis

- [Descriptive Statistics](#) (js)
- [Guess the Correlation](#) (js)
- [Least Squares Regression](#) (js)

#### Sampling Distribution Simulations

- [Reeses Pieces](#) (js)
- [Sampling Words](#) (js)
- [Sampling from a Finite Population](#) (js)
- [Sampling from a Probability Model](#) (j)
- [Sampling Regression Lines - Population Model](#) (j)
- [Simulating Confidence Intervals for Population Parameter](#) (js)
- [Improved Batting Averages \(Power\)](#) (js)
- [ANOVA simulation](#) (js)

#### Classics (j)

- [Histogram Bin Width](#)
- [Dotplot Summaries](#)
- [Sampling Pennies](#)
- [Sampling Change](#)
- [Sampling 2005 Senators](#)
- [Friendly Observers](#)
- [Dolphin Study applet](#)
- [Yawning Study applet](#)
- [Two-way Table simulation applet](#)
- [Randomization Test for quantitative response \(two groups\)](#) (f)
- [Simulating Confidence Intervals for Population Parameter](#)
- [Simulating Intervals for different population shapes](#)
- [Random Babies](#)

#### Probability

- [Random Babies](#) (js)
- [Secretary Problem](#) (j)
- [Normal Probability Calculator](#) (js)
- [Randomizing Subjects](#) (js)
- [Random number generator](#) (js)

#### Statistical Inference

- [One proportion inference](#) (js)
- [Analyzing Two-way Tables](#) (js)
- [Matched Pairs](#) (js)
- [Randomization test for quantitative response \(multiple groups\)](#) (js)
  - [two means](#)
- [Randomization test for categorical response \(multiple groups\)](#) (js)
  - [Dolphin Study applet](#)
- [Analyzing Two Quantitative Variables](#) (js)
- [Theory-based Inference](#) (js)

Click [here](#) to access old applets page

j = java applet (click here for help on running java on [macs, pc](#))  
 js = javascript  
 f = flash

### StatKey

to accompany [Statistics: Unlocking the Power of Data](#)  
 by Lock, Lock, Lock, Lock, and Lock

Descriptive Statistics and Graphs	Bootstrap Confidence Intervals	Randomization Hypothesis Tests		
One Quantitative Variable	CI for Single Mean, Median, St.Dev.	Test for Single Mean		
One Categorical Variable	CI for Single Proportion	Test for Single Proportion		
One Quantitative and One Categorical Variable	CI for Difference in Means	Test for Difference in Means		
Two Categorical Variables	CI for Difference in Proportions	Test for Difference in Proportions		
Two Quantitative Variables	CI for Slope, Correlation	Test for Slope, Correlation		
Sampling Distributions	Mean	Proportion		
Theoretical Distributions	Normal	t	$\chi^2$	F
More Advanced Randomization Tests	$\chi^2$ Goodness-of-Fit	$\chi^2$ Test for Association	ANOVA for Difference in Means	ANOVA for Regression

# Rossman/Chance Applet Collection

## Comparing Groups (Quantitative Response)

Sample data:  Unstacked

Treatment	strength
A	74
A	66
A	77
B	67
B	54
B	65
C	64
C	70
C	62

Use Data Clear

Summary Statistics:

Group	n	Mean	SD
A	3	72.33	5.41
B	3	61.67	5.12
C	3	65.33	5.29
grand	9	67.40	4.71

Statistic: MAD Observed MAD=5.267

Show ANOVA Table:

Shuffled Summary Statistics:

Group	n	Mean	SD
A	3	69.33	4.93
B	3	65.33	5.12
C	3	69.33	5.14
grand	9	67.40	5.12

Shuffled MAD=2.33

Count Samples: Greater Than  Count

Page last modified 04/18/2016 22:44:57

### Notes:

- MAD = mean absolute difference
- This applet does not work in IE8 but should work in other browsers.
- Right now pasted data must have variable names (use single words, no symbols)
- When pasting unstacked data, use \* to fill in empty values if the groups have unequal sample sizes
- To default to multiple groups, click [here](#). For two groups, click [here](#)

<http://www.rossmanchance.com/applets>

### StatKey Randomization Test for a Difference in Means

Randomization method: **Resample Groups**

Generate 1 Sample Generate 10 Samples Generate 100 Samples Generate 1000 Samples Reset Plot

Randomization Dotplot of  $\bar{x}_1 - \bar{x}_2$ , Null hypothesis:  $\mu_1 = \mu_2$

Left Tail  Two-Tail  Right Tail

sample = 2170  
mean = -0.0000  
st dev = 0.297

Original Sample  
 $\bar{x}_1 - \bar{x}_2 = 0.70, s_1 = 34, s_2 = 37$

Randomization Sample  
 $\bar{x}_1 - \bar{x}_2 = 0.00, s_1 = 34, s_2 = 37$

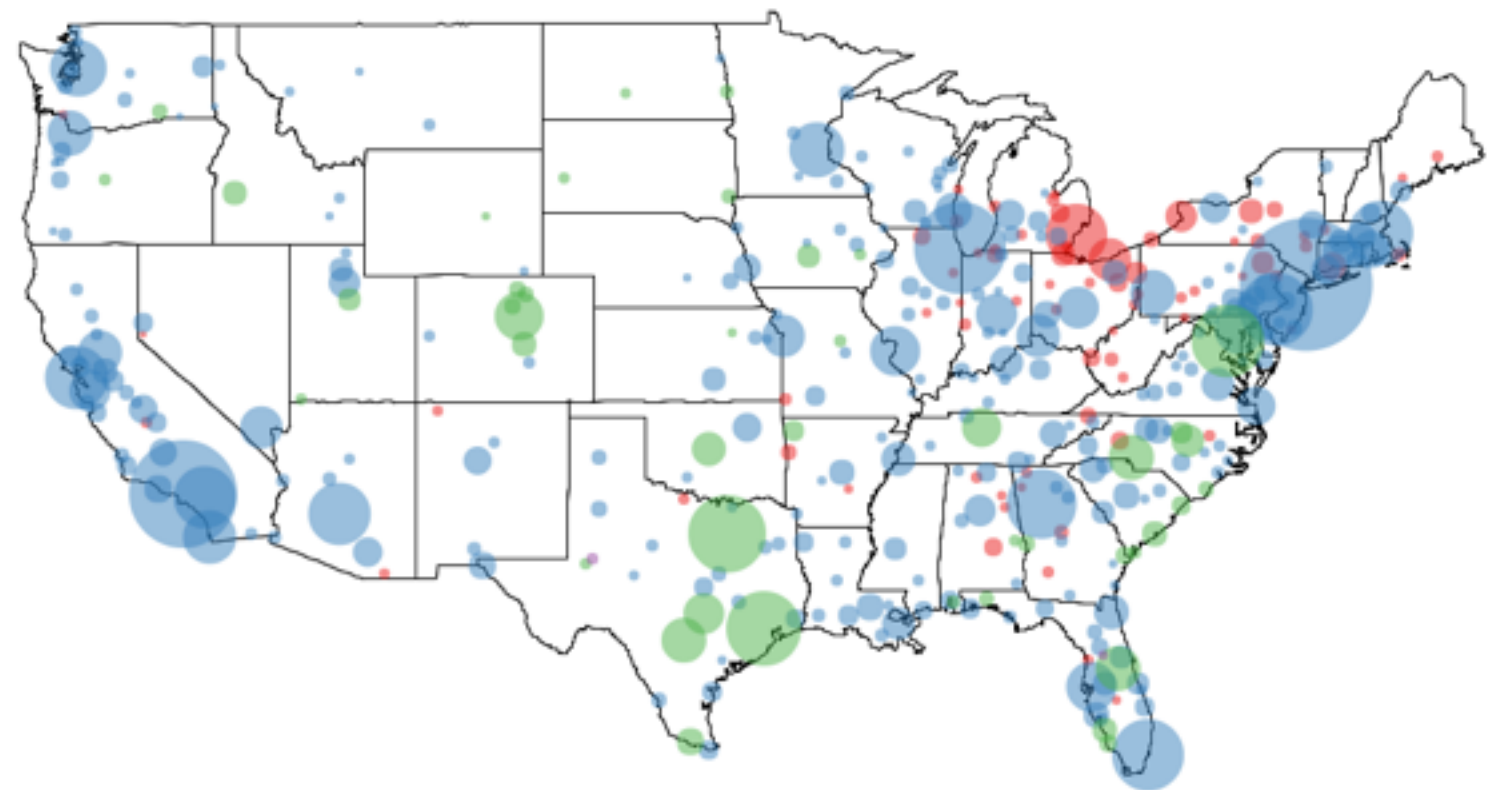
StatKey v. 0.3.12 is written in JavaScript and should work well with any current browser including Chrome, Firefox, Safari, Opera, and IE.  
Comments, feedback, and bug reports can be sent to [lock5stat@gmail.com](mailto:lock5stat@gmail.com).

<http://lock5stat.com>



### Visualizing the population change in US Metro areas

Copy Print Mail Link Embed Twitter Facebook



**Change%**

- -5 to 0
- 0 to 5
- 5 to 10
- 10 to 15

**Tags:**  
map

**HTML link:**  
<A href="http://www.statcrunch.com/5.0/viewresult.php?resid=1802998">Visualizing the population change in US Metro areas</A>

**Comments**

Want to comment? [Subscribe](#)  
 Already a member? [Sign in.](#)

**Result Properties**

Thumbnail:



Owner: websterwest



Created: Sep 21, 2015

Size: 122KB

Share: yes

Views: 1950

**Data set for this result:**

Metropolitan Statistical Areas in the U.S. - Population, Location



By statcrunchhelp  
On Oct 14, 2014

**Recently shared results for this data set:**

The changing population of U.S. Metropolitan Areas



By scsurvey  
On Aug 28, 2015

Percent Change in population from 2010 to 2013 for U.S. Metro Areas



By statcrunchhelp  
On Oct 14, 2014

[View all shared >](#)

**Reports with this result:**

None

The screenshot shows the TinkerPlots software interface. The window title is "TinkerPlots - [US Students]". The menu bar includes File, Edit, Object, Data, Window, and Help. The toolbar contains icons for Cards, Table, Plot, Sampler, and Text.

The main interface is divided into two panes. The left pane, titled "US Students", displays a data table for "case 1 of 82". The table has columns for Attribute, Value, Unit, and Formula. The right pane, also titled "US Students", displays a scatter plot of blue circles representing data points.

Below the data table, there is a description of the dataset: "82 U.S. high school students in Western Massachusetts, 1990".

Below the description, there is an "Attribute Description" section with the following text:

- Gender:** Gender
- School:** High school attended
- BirthYear:** Year of birth
- Height:** Height
- Weight:** Weight
- OldSibs:** Number of older siblings
- YoungSibs:** Number of younger siblings
- Children:** Total number of children in family
- Parents:** Parents deceased, separated, or together
- MoneyOnYou:** Number of dollars currently carrying

Below the attribute description, there is a "Questions" section with two questions:

1. Do students with jobs spend fewer hours per week doing homework?
2. Do the students who do more homework tend to get better grades?

The bottom of the window shows the version information: "TinkerPlots(tm): a data analysis construction set, version 2.0b14".



# Fathom<sup>®</sup>

## Dynamic Data Software

Collection Table Graph Summary Estimate Test Model Slider Meter Text

**Mammals** See collection comments.  
Is there a relationship between how long mammals sleep and how long they live?

**Mammals** Scatter Plot

**Mammals** Percentile Plot

**Mammals** Histogram

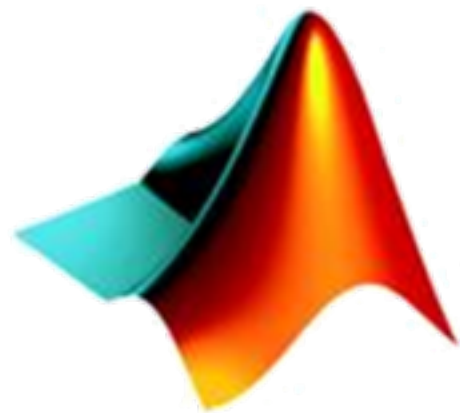
**Mammals** Line Scatter Plot

Tools for teaching and learning statistics are typically

- Inexpensive
- Accessible
- Interactive
- Curated

# Tools for doing statistics





MATLAB®





```
Terminal Shell Edit View Window Help
amelia — R — 80x24
Last login: Wed Apr 16 15:39:29 on ttys000
Amelias-MacBook-Air:~ amelia$ R

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █
```



```

R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

[History restored from /Users/amelia/.Rapp.history]

> |
```





~/Dropbox/Documents/Teaching/101c - RStudio

Discussion7.Rmd x

```
49 mydata=data.frame(Y, X)
50
51 require(leaps)
52 ms2 = regsubsets(Y~poly(X,10), data=mydata, nvmax=10)
53 coef(ms2,4)
54 ms3 = regsubsets(Y~poly(X,10, raw=TRUE), data=mydata, nvmax=10)
55 coef(ms3,4)
56 ...
57
58 **Back to polynomial regression**
59 -----
60 So, the "raw" parameter determines whether you use orthogonal polynomials or raw polynomial. They work
61 out about the same when you do predictions, so it doesn't really matter which one you use.
62
63 Lets plot the fits. First, we need to do some predictions.
64 ```{r}
65 ageLims = range(Wage$age)
66 ageGrid = seq(from=ageLims[1], to=ageLims[2])
67
68 m2 = lm(wage~poly(age, 3), data=Wage)
69 m3 = lm(wage~poly(age, 2), data=Wage)
70 m4 = lm(wage~age, data=Wage)
71
72 predictions1 = predict(m1, newdata=list(age=ageGrid))
73 predictions1 = c(predictions1, predict(m2, newdata=list(age=ageGrid)))
74 predictions1 = c(predictions1, predict(m3, newdata=list(age=ageGrid)))
75 predictions1 = c(predictions1, predict(m4, newdata=list(age=ageGrid)))
76
77 predData = data.frame(ageGrid = rep(ageGrid, 4), preds=predictions1, poly=c(rep(4, length(ageGrid)), rep
78 (3, length(ageGrid)), rep(2, length(ageGrid)), rep(1, length(ageGrid))))
79 predData$poly = factor(predData$poly)
80
```

Environment History

Global Environment

Environment is empty

Files Plots Packages Help Viewer

R: Fitting Linear Models

## Fitting Linear Models

### Description

lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although [aov](#) may provide a more convenient interface for these).

### Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

### Arguments

formula	an object of class " <a href="#">formula</a> " (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
data	an optional data frame, list or environment (or object coercible by <a href="#">as.data.frame</a> to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.
subset	an optional vector specifying a subset of observations to be used in the fitting process.

Console ~/Dropbox/Documents/Teaching/101c/

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale


R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

# Tools for doing statistics are often

- Expensive
- Static
- Flexible
- Extensible
- Reproducible





What's in  
between?



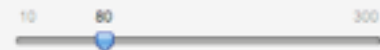
# Shiny

by RStudio

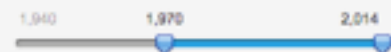
## Movie explorer

### Filter

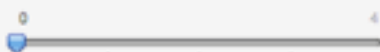
Minimum number of reviews on Rotten Tomatoes



Year released



Minimum number of Oscar wins (all categories)



Dollars at Box Office (millions)



Genre (a movie can have multiple genres)

All

Director name contains (e.g., Miyazaki)

Cast names contains (e.g. Tom Hanks)

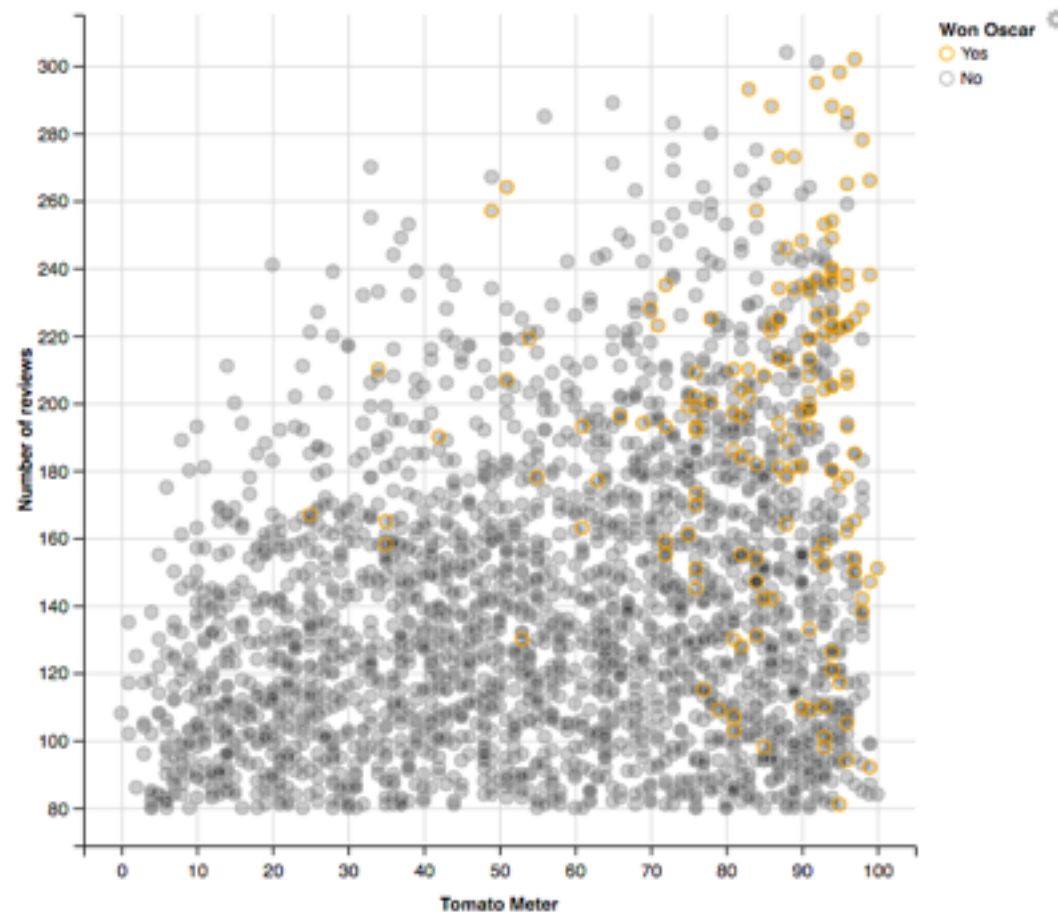
X-axis variable

Tomato Meter

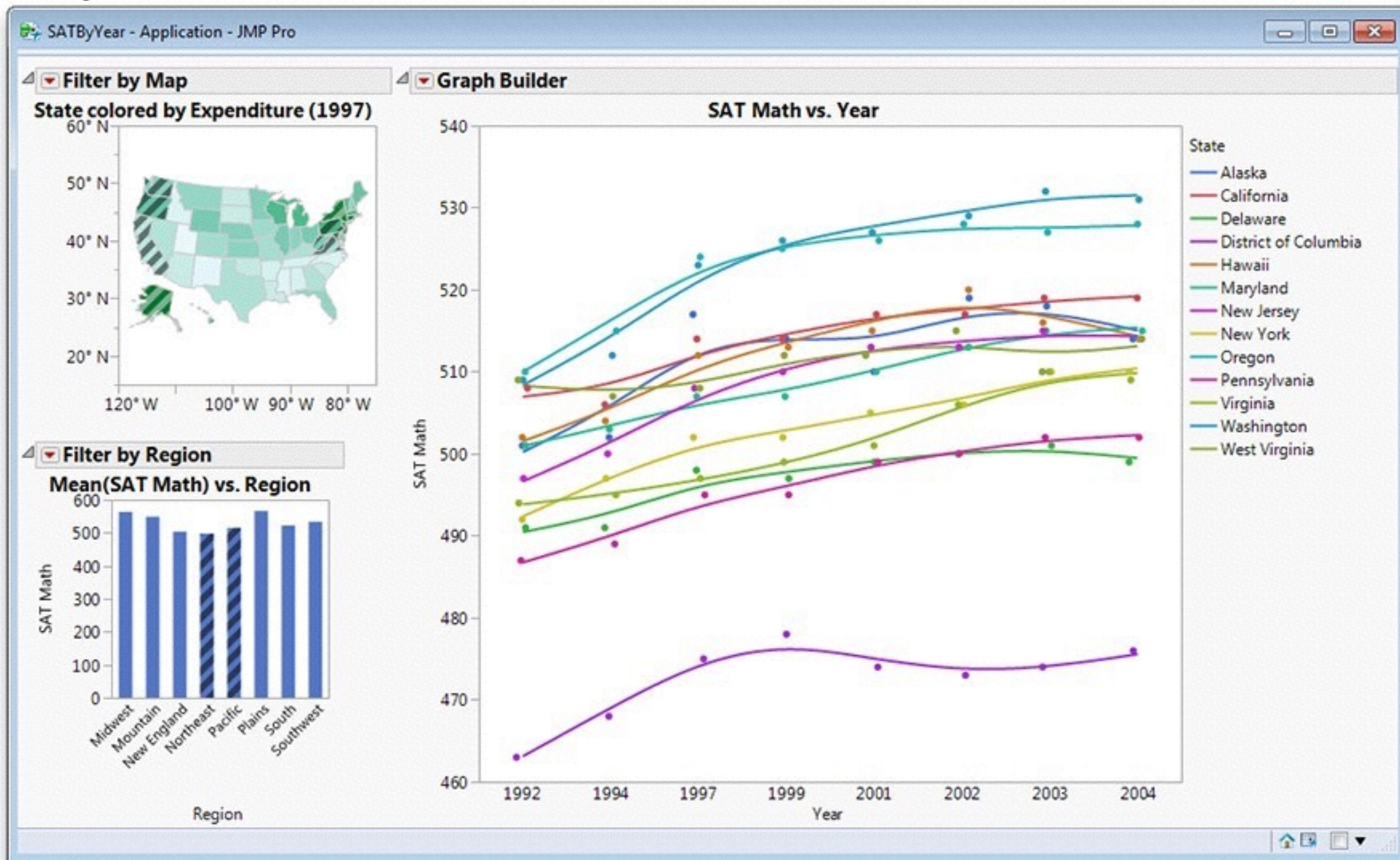
Y-axis variable

Number of reviews

Note: The Tomato Meter is the proportion of positive reviews (as judged by the Rotten Tomatoes staff), and the Numeric rating is a normalized 1-10 score of those reviews which have star ratings (for example, 3 out of 4 stars).



Number of movies selected:  
2557

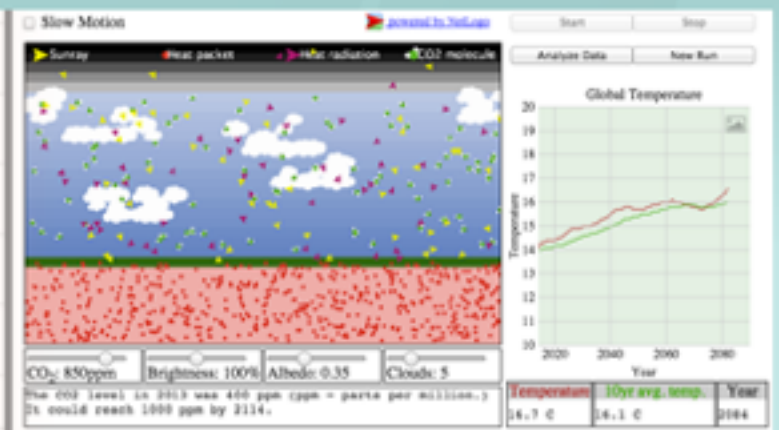
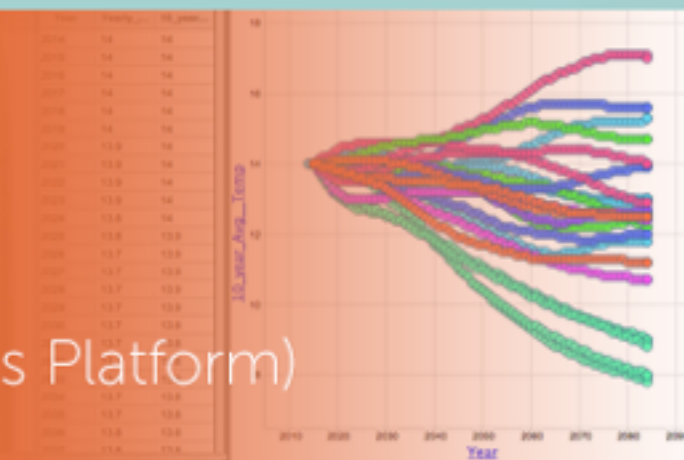




# CODAP

(Common Online Data Analysis Platform)

SHARE PRINT



Untitled Document User: guest Version 1.1 (0283 IS)

File Game Table Graph Map Slider Calc Text Options

Clear Data... Login

### Next Gen MW

Share About

Achieved Terminal Velocity

#### Distance vs. Time

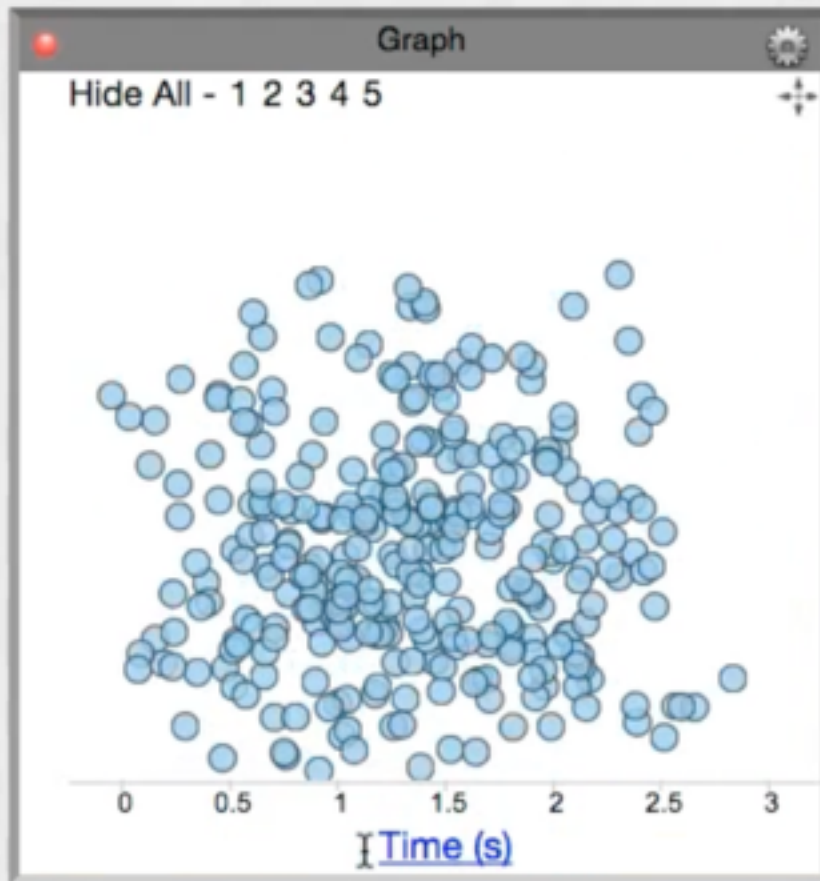
#### Velocity vs. Time

Mass of jumper (g)  Parachute size (cm<sup>2</sup>)

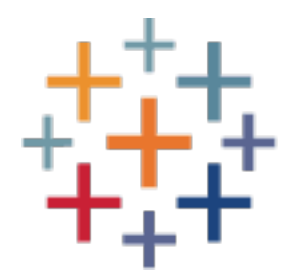
12.5 s Start Stop Analyze Data New Run

### 5 runs/305 measurements Case Table

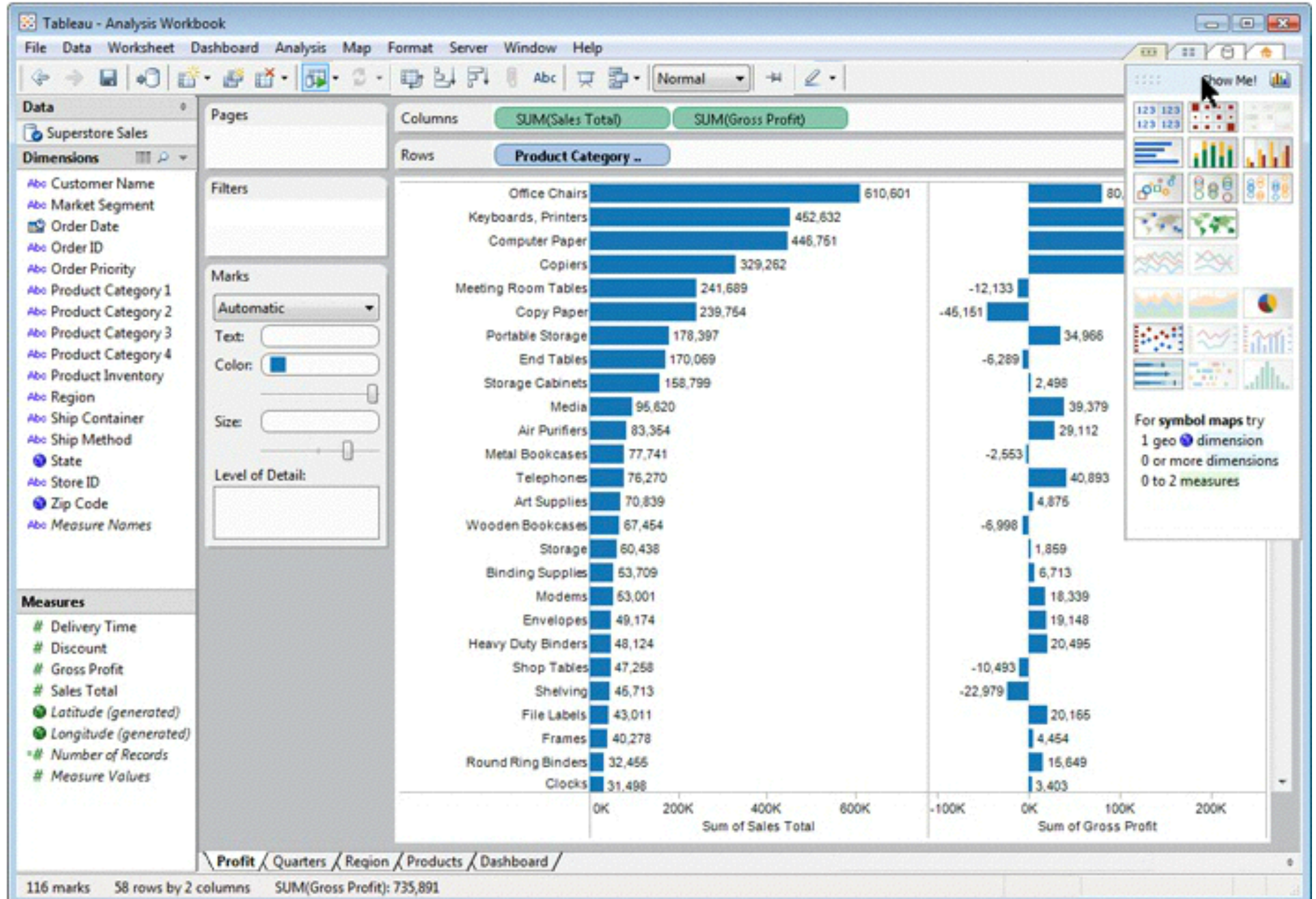
Row	mass_o...	parach...	termina...	Time (s)	Distanc...	Velocit...
1	200	700	2.29	2.33	4.58	1.78
2	200	800	2	2.37	4.51	1.78
3	200	1000	1.6	2.41	4.44	1.78
4	200	1100	1.45	2.45	4.36	1.78
5	200	900	1.78	2.5	4.29	1.78







# tableau®





## Characters

From Source: characters

## Data

name	Myriel	Napoleon	Mlle.Baptistine
group	1	1	1
index	0	1	2

1-20 of 77

+ NEW TRANSFORM

## Scales

Fill Color Size Font Size

+ NEW PIPELINE



## Layer 1

AXES

MARKS

## Labels

## Nodes

Type

## Pipeline

Characters + VISUAL LAYOUTS

## Force-Directed Layout

## LINKS DATA

Data connections  
Source source Target target

## LINKS STROKE

Line Type line  
Color #ccc  
Width 0.5

## CONNECTIONS

Distance 80 Strength 1 Tension 0

## NODES

Charge -220 Friction 0.9 Gravity 0.1

Output x y weight

## Properties

## POSITION

X x Y y

## GEOMETRY

Shape circle Size Size weight

## FILL

Color Fill Color group  
Opacity 0.3

## STROKE

Color Fill Color group  
Width 0.75



24 Columns 345 Rows 2 Data Types

Grid

Filter in grid

##	AZ_Phoenix	##	CA_Los_Angeles	##	CA_San_Diego	##	CA_San_Francisco	##
	65 - 228		59 - 273				47 - 219	
	PHXR-SA		LXXR-SA				SFXR-SA	
			59.43				46.96	
			59.89				47.3	
			60.4				47.84	
			61.32				47.98	
			62.03				48.31	
			62.78				48.61	
			63.46				49.08	
			64.13				49.54	

## The Transformer

1 of 5

This is **the Transformer**, where you can transform your messy data into clean data.

The Transformer is populated with a sample of your dataset. Let's discover what's in your data!

[Learn more](#)

[Don't show me any helpers](#)

Next

expression





Show as: **rows** records    Show: 5 10 25 50 rows

All	V2	norm	typ
☆	1.	4A YARN DYEING	accord
☆	2.	4S PARK STYLE	accord
☆	3.		
☆	4.		
☆	5.		
☆	6.		
☆	7.		
☆	8.	A&A Trousers Ltd	
☆	9.	A & B OUTERWEAR LIMITED	
☆	10.	A.K.M. Knit Wear Limited	

- Facet
- Text filter
- Edit cells
- Edit column**
  - Split into several columns...
  - Add column based on this column...**
  - Add column by fetching URLs...
  - Add columns from Freebase ...
  - Add columns from DBpedia ...
  - Extract entities from text (Zemanta API)
- Transpose
- Column statistics
- Sort...
- View
- Reconcile
- Extract named entities...
- Rename this column
- Remove this column
- Move column to beginning
- Move column to end
- Move column left
- Move column right

### Add column based on column V2

New column name:

On error:  set to blank  store error  copy value from original column

Expression:  Language:  No syntax error.

**Preview**    History    Starred    Help

row	value	cell.recon.match.name
1.	4A YARN DYEING LTD.	Error: Cannot retrieve field from null
2.	4S Park Style Ltd.	Error: Cannot retrieve field from null
3.	4 Knitwear ltd	Error: Cannot retrieve field from null
4.	4 You Clothing Ltd	4 You Clothing Ltd.
5.	A Class Composite Ltd.	Error: Cannot retrieve field from null
6.	A J Super Garments Ltd.	A J Super Garments Ltd.
7.	A Plus Ltd Ltd	Error: Cannot retrieve field from null

flickr: psychemedia

# Reaching across the gap



Tools for learning should aim for

- Reproducibility
- Shareability

Tools for doing should aim for

- Interactivity
- Accessibility

## Selected References

- Biehler, Rolf. (1997). Software for Learning and for Doing Statistics. *International Statistical Review*, 65(2). <http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.1997.tb00399.x/abstract>
- Biehler, Rolf., Dani Ben-Zvi, Arthur Bakker and Katie Makar. (2012). "Technology for enhancing statistical reasoning at the school level." *Third International Handbook of Mathematics Education*. [http://link.springer.com/chapter/10.1007/978-1-4614-4684-2\\_21](http://link.springer.com/chapter/10.1007/978-1-4614-4684-2_21)
- McNamara, Amelia. (2015). Bridging the Gap Between Tools for Learning and For Doing Statistics. <http://escholarship.org/uc/item/1mm9303x>
- Tukey, John. (1965). The Technical Tools of Statistics. *The American Statistician*, 19(2). <http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1965.10479711#.VzOMAxUrJBw>



Questions? Comments? Thoughts?

[amcnamara@smith.edu](mailto:amcnamara@smith.edu)

[@AmeliaMN](#)

