# Relevant & Engaging

# Statistics & Data Handling

## Data Visualisation

Royal Statistical Society
FOUNDED 1834
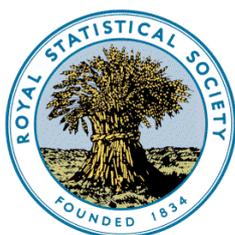
**Centre for
Statistical Education**

## Chapter 6

This booklet is aimed at all secondary level teachers: there are hints and tips that we hope will be useful to support the teaching and learning of statistics and data handling. We hope that you find the material useful. Please email or write to us with suggestions for improvements. We will try to respond to all communications.
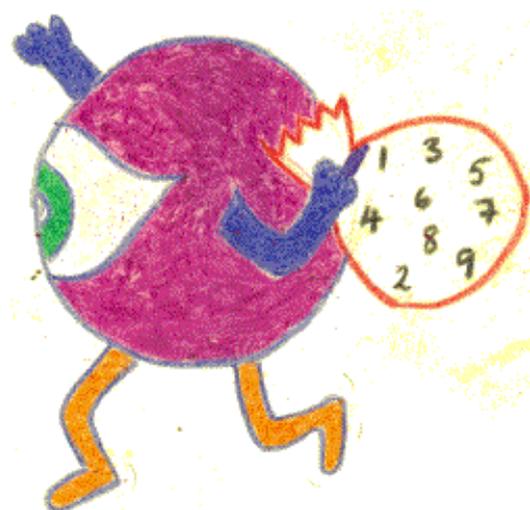
**Doreen Connor**
The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
2009

email info@censusatschool.org.uk

www.censusatschool.org.uk

www.rsscse.org.uk

**Chapter 6**

# Data Visualisation

In this chapter we look at some ways to visualise data. The use of visualisation to present and understand information is certainly nothing new. It has been used in maps, scientific drawings and data plots for well over a thousand years. For example, Ptolemy's Geographia (2$^{nd}$ Century AD) and Minard's map (1861) of Napoleon's invasion of Russia. In fact visualisation covers any technique for creating images, diagrams or animations in order to convey a message. Examples of common visualisations include: tables, matrices, charts, graphs, maps, Venn diagrams, and Chernoff faces to name but a few. Modern technology and computer power now enable us to use ICT to explore and interrogate data in a visual way in order to find patterns and gain some information from the data. We look first at an American application called Tinkerplots and then our own web-based Database Interrogation Tool to show how using data visualization in the classroom can help to enhance the experience our learners get by allowing them to interact with and explore the data themselves. In all our illustrations we will use data from CensusAtSchool.
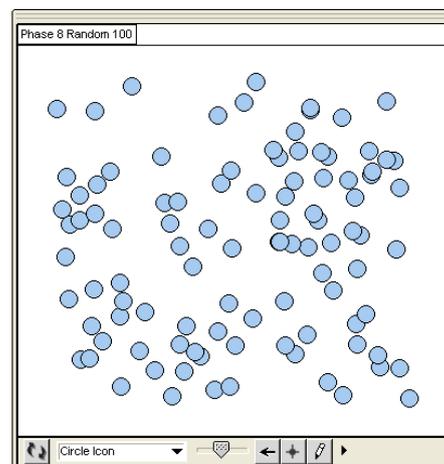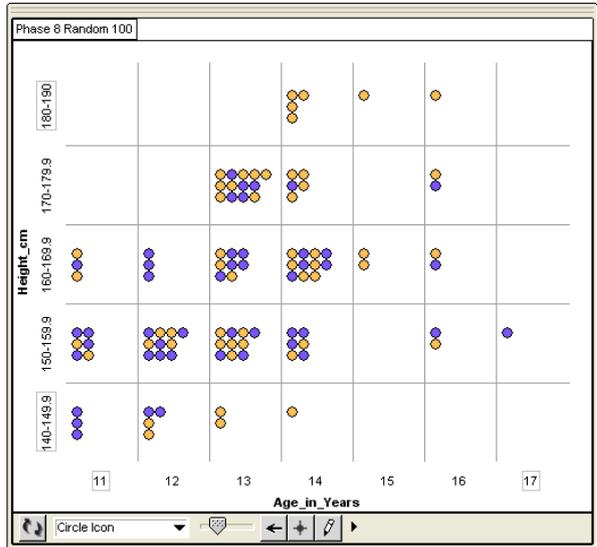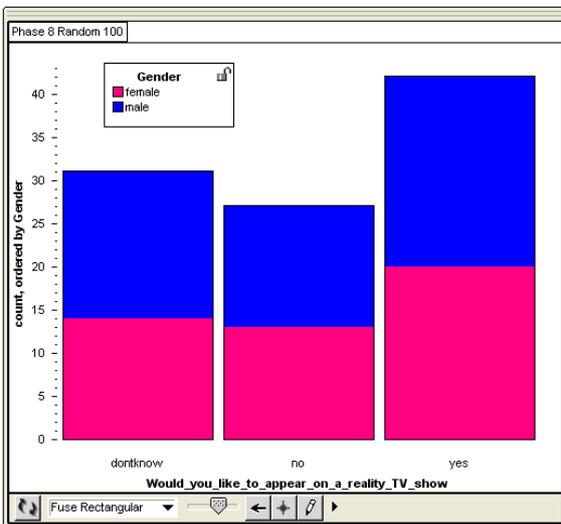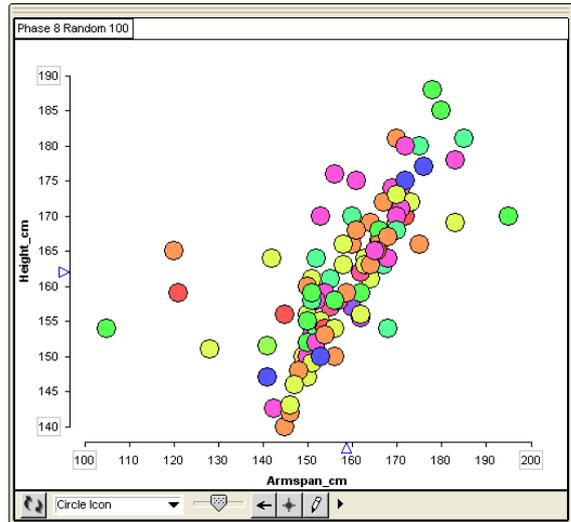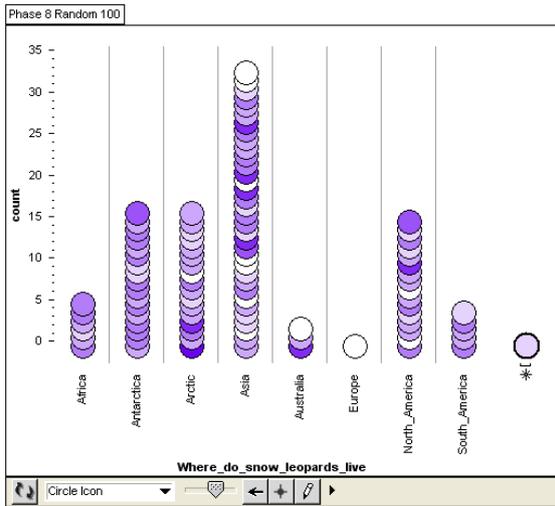
**Data Visualisation using Tinkerplots.**
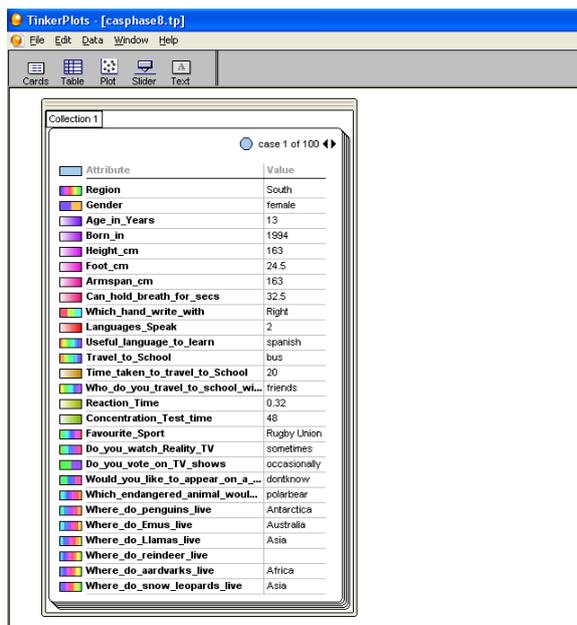


It is very straightforward to put data into Tinkerplots and we will use a random sample of 100 Phase 8 responses to illustrate here. To input the data you simply copy and paste from the CSV or Excel file into a new tinkerplot file. It will generate a new card for each response showing the attributes as below. The colours beside each show the colours the data will be represented on the graphs as. In order to start to visualise the data you simply select a new plot. Each dot represents one case. They are randomly distributed across the plot to start with.

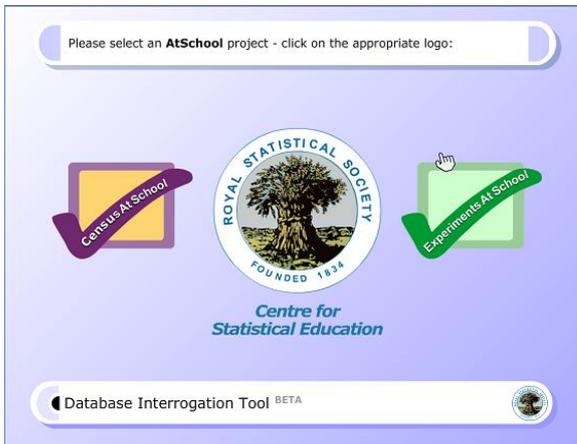By now selecting any attribute from the list the dots become coloured and by dragging attributes to the axes and playing around dragging them you can generate visualizations of the data.

Because Tinkerplots is designed for learners to play with rather than be given a precise set of instructions we will simply show you a few of the data visualizations that can be generated to give you the idea of how the software works.

Phase 8 Random 100

count

35
30
25
20
15
10
5
0

Africa
Antarctica
Arctic
Asia
Australia
Europe
North_America
South_America

Where_do_snow_leopards_live

Circle Icon

---

Phase 8 Random 100

Height_cm

190
185
180
175
170
165
160
155
150
145
140

100  110  120  130  140  150  160  170  180  190  200

Armspan_cm

Circle Icon

---

Phase 8 Random 100

Gender
female
male

40
35
30
25
20
15
10
5
0

dontknow          no          yes

Would_you_like_to_appear_on_a_reality_TV_show

count, ordered by Gender

Fuse Rectangular

---

Phase 8 Random 100

Height_cm

180-190
170-179.9
160-169.9
150-159.9
140-149.9

11   12   13   14   15   16   17

Age_in_Years

Circle Icon

**Data Visualisation using the Database Interrogation tool.**



Screen 1: Starting screen

The Database Interrogation tool has been designed to allow quick and easy access to the large amount of data collected by the *AtSchool* projects such as CensusAtSchool and Experiments AtSchool

The tool works through a web browser such as Internet Explorer and requires a live internet connection so that data can be requested and sent from the *AtSchool* databases. You will also need Flash v8 or later installed.

**Selecting a dataset**

After a particular AtSchool project has been selected, click on the **Next** button and the related selection of experiments or results will appear. Once you have chosen which set of results you want to investigate, the names of variables contained in the selected results database will appear.
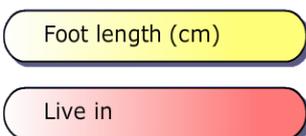
For example, by selecting the **CensusAtSchool** logo and then **UK Secondary 2005**, you will be shown a list of 35 variable names in the database.



Screen 2: List of Databases
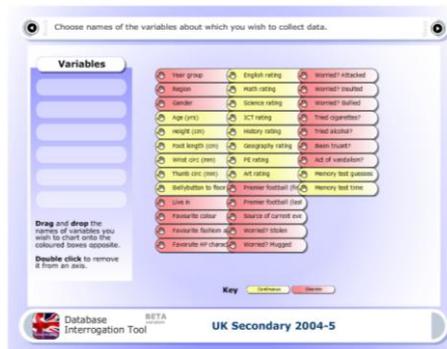


Screen 3: List of names of Variables



The list of variable names is colour-coded. Yellow items, or **tabs**, represent continuous variables. Discrete variables names appear as red tabs.

The coloured tabs can be dragged around the screen and dropped on to the list of variables to be investigated (collected variables). You can investigate a maximum of five **collected variables** in one go.

You may mix and match continuous and discrete variables in your selection.

### Selecting a random sample

You can then select the type of **random** sample you wish to collect. This screen allows you to change the size of the sample and to collect data for a particular category/group.



The size of the sample, that can be collected, can be varied from 50 to 250. The default sample size is 100.

(Larger samples may take longer to collect and analyse depending on the speed of your internet connection.)

Screen 4: Type of random sample

### Selecting a sub-sample from the database

As a possible investigation, you might decide to collect data for a particular year group and gender. To do this, select the **Part of database** tab.





A new screen will appear showing a full list of the variable names from the database. The new tabs work as a clickable buttons rather than as dragable items.

Each coloured tab can be used to help select particular features for your random sample – in other words each tab acts as a filter.

A tab with a star on it indicates that it is one of your **collected variables**. You should use the starred tabs when you only want the collected variables to satisfy particular conditions, for example when you only to collect heights in the range of 100 to 200cm.

Suppose you wish to obtain a random of sample of heights and foot length data for boys aged 14. First click on the **Gender** tab. This will show a list of options on the left hand side of screen. Select the **Male** option. As you do so, the Gender tab will be ticked to show that you are using it as a filter.
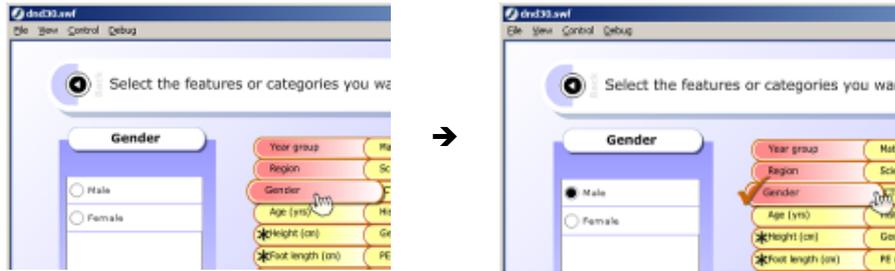
After clicking on the **Age** tab, the maximum and minimum values will appear in the option list. Alter the maximum and minimum values to be 14 and then click on the **Save this option** button.



The interrogation tool will now select a random sample of boys aged 14 and make these available for processing.

**Selecting two samples**

In an investigation, you might need to collect and compare data for two groups for example Girls and Boys. To do this, select the **Part of database** option as described on the previous pages.

On the right hand side of the screen, select the **Two samples** option. You will now be able collect two samples; in the interrogation tool these are referred to as sample **A** and **B**. Once you have chosen the features you want for sample **A**, clicking on the **Next** button allows you to choose features for sample **B**.
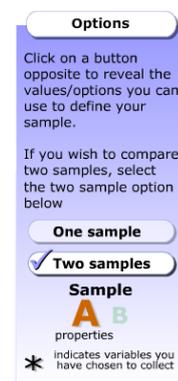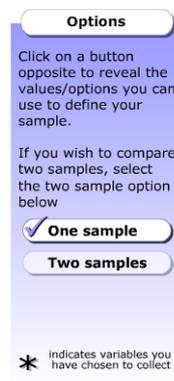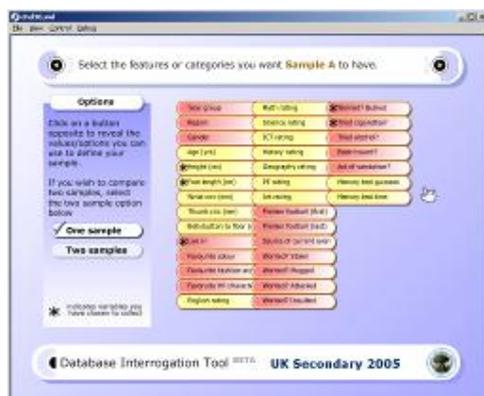
## Displaying the data
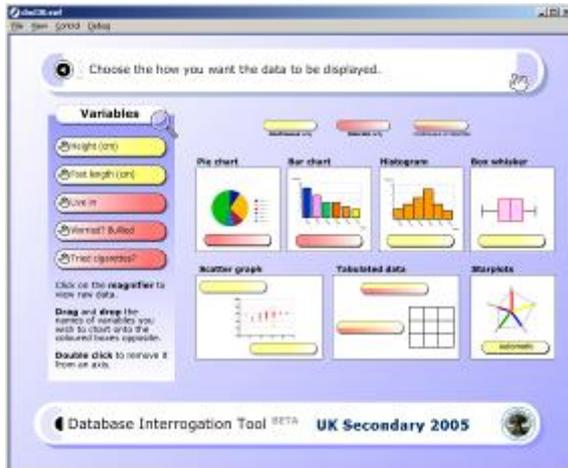


Screen 6: Displaying data

After the data has successfully returned from the online database, it can be displayed in tables and charts.

To view the raw data, select the magnifying glass icon at the top of the **Variables** list.

Each tab in the list of **collected variables**, on the left hand side of the screen, can be dragged and dropped on to the chart icons.

| Histograms, Box-whisker and Scatter charts can only be plotted for continuous variables: | Pie charts and Bar charts can only be plotted for discrete variables: |
|---|---|
|  |  |
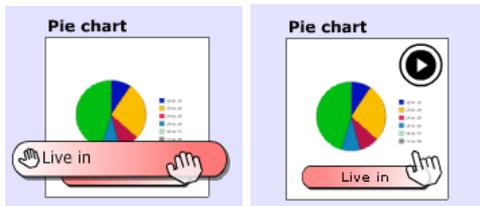


Once an appropriate type of variable has been dropped onto a chart, an arrow button will appear. Click on this arrow to view the chosen chart.

If you have collected two samples, several of the charts and tables allow both sets of results to be plotted simultaneously. Where this option is not available you can swap between samples by selecting one of the tabs shown opposite.



If you have used a sample filter, an information button ⓘ will be added underneath the chart. Clicking on this button will reveal a pop-up window showing the particular sub-samples you used.

## Charting Options



All charts have some additional options. Click on the **Show options** button on the top menu bar. **Bar charts** (and **Pie charts**) can be arranged in descending order (by percentage) or arranged in alphabetical/default order.
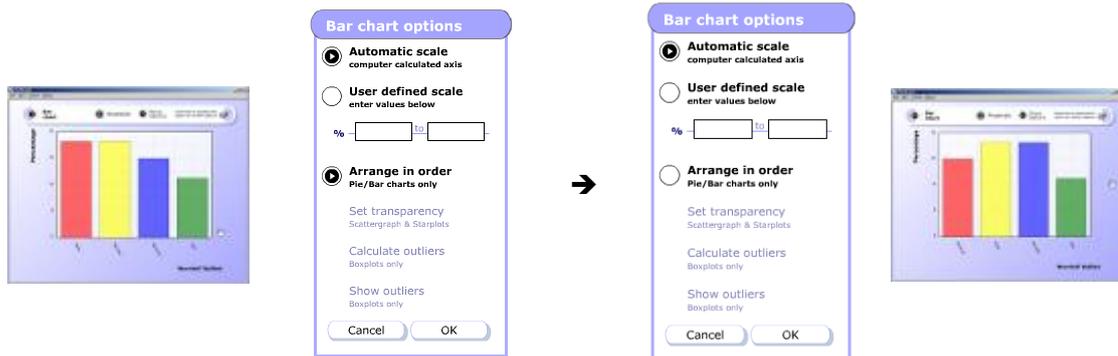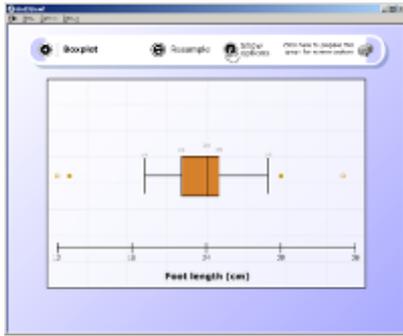


As with Bar charts and Pie charts, options for the layout of the other types of chart can be accessed using **Show Options** button on the top menu bar.



For **Histograms**, you can change the range of values plotted and the number of bins - these alterations will only take place if you select **User defined scale** and click on **OK**.

When you plot a histogram for two samples, each will be drawn back to back. Altering the option values will alter both charts simultaneously.

For **Box and whisker** plots, you can choose to show or ignore outliers.

If the **Calculate outliers** option is not selected, the whiskers will extend from the ends of the box to the maximum and minimum values.

As the mouse moves over parts of particular charts, values will be displayed next to the cursor. For example values of outliers are identified in this way on Box and whisker plots
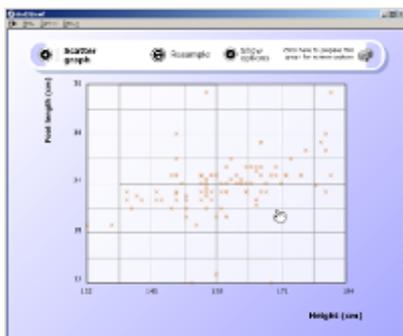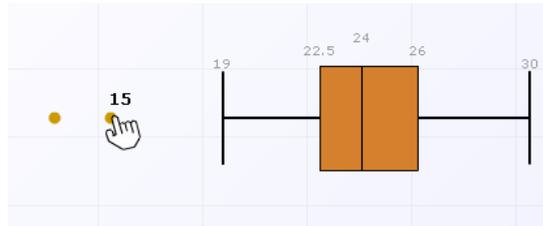




For **scatter graphs** You can choose to alter the range covered by either axis and to alter the transparency of the plots. When using a high transparency most data points will appear faint except in cases where several points, with the same co-ordinates, have been plotted on top of each other.

When you plot a scatter graph for two samples, each will be plotted in different colours. The transparency of each can be adjusted independently.
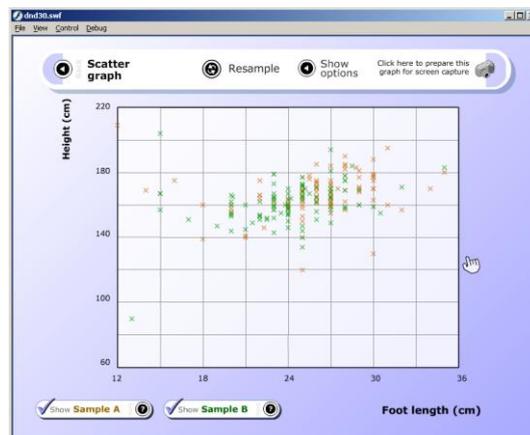
### Tabulated data

Tables can be created by mixing and matching both continuous and discrete variables.



Continuous or Discrete
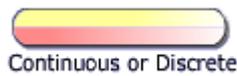
In the example shown opposite, a two-way table has been created for data from the CensusAtSchool project using returns for **Live In** and **Height**.
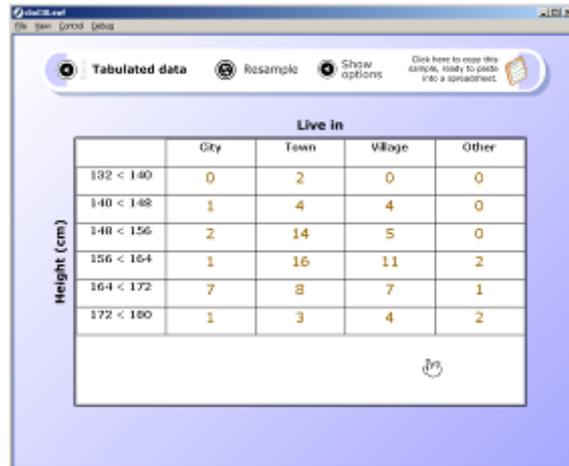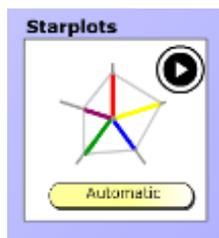


Chart option: Tabulated data

For the discrete variable, **Live In**, four columns are created to match the four possible values (City, Town, Village, Other).

For the continuous variable, **Height**, the number of rows can be varied using **Show options** on the top menu bar. The user can alter the step size and range covered.



### Starplots

Unlike the other chart types you <u>cannot</u> drag and drop the names variables on to the Starplot graphic.

The arrow button for this particular graphic will automatically appear when three or more continuous variables have been collected

Each Starplot consists of equi-angular spokes, called radii, with each spoke representing one of the continuous variables. The length of an individual radius is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points.
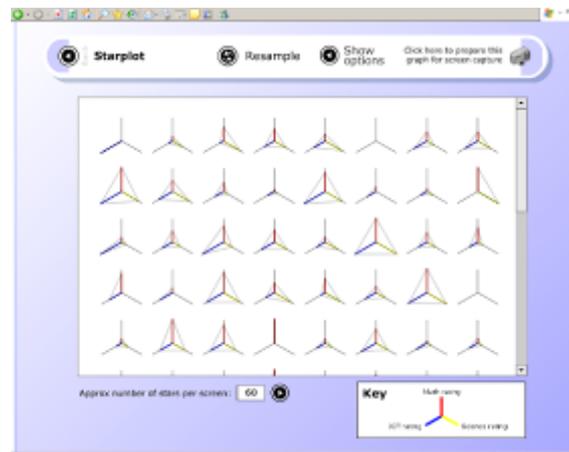


Chart option: Starplots