Collecting Data

Where Does All This Data Come From?

As we're discussing processing and visualizing all this data, it's important to know where it comes from! There are a couple of questions we need to answer:

- How can computers help us collect data?
- How can we best collect and store data so that we can analyze it with computers?

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
- Sensors
- Transactional data from credit cards
- Websites storing information about you
- Crowdsourcing data

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
  - If we want to get information from people about people, we can send out surveys!
  - We have to be mindful that not everyone wants to take surveys, so we could potentially get some skewed results!
- Sensors
- Transactional data from credit cards
- Websites storing information about you
- Crowdsourcing data

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
- Sensors
  - We can use all kinds of different sensors to gather information about the world around us!
    - Thermometers, barometers, etc.
  - Usually these'll interface directly with our computers, which lets us store the data we collect directly!
- Transactional data from credit cards
- Websites storing information about you
- Crowdsourcing data

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
- Sensors
- Transactional data from credit cards
    - Anytime you make a transaction using a credit/debit card, it gets tracked!
    - This allows us to see our own spending history
    - This also allows credit card companies to see how we spend our money!
- Websites storing information about you
- Crowdsourcing data

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
- Sensors
- Transactional data from credit cards
- Websites storing information about you
  - Anytime you visit a website, you're likely to have information collected about your visit.
  - Clicks, time spent on different pages, browser being used, etc.
  - This can allow the website to tailor itself to your visit!
- Crowdsourcing data

How Data is Collected

There are many different vectors by which data can be collected, but here are a few examples!

- Surveys
- Sensors
- Transactional data from credit cards
- Websites storing information about you
- Crowdsourcing data
    - Pose a question to a large number of Internet users, or potentially track the behavior of a large group of people on the Internet

How Data is Stored

The first place that data is stored is on your computer's local disk memory.

The data can then be uploaded to a **database** - a computing system designed specifically for storing large amounts of information.

If the data is sensitive, it can be encrypted! We might want to encrypt login information, medical records, or any number of other types of information!

How Data is Stored

Data is most often stored in **data centers** - warehouse-sized buildings full of computers used just to store large quantities of information!

Accessing and Using the Data

As you might expect, accessing all of this data in an easy and efficient way can be challenging!

Database query languages and Data APIs allow us to ask a database for the data we want to use in our programs.

- SQL (Structured Query Language)
- API provided by the owner (Data.gov, YouTube, Google)

Making Data Usable

Something to keep in mind while working with data is that some kinds of data are much easier to process and visualize than others! **Numbers** are almost always going to be easier to process than **text**.

Making Data Usable

Also important is ensuring that the data you're working with has been **sanitized**. This process involves throwing out data that hasn't been formatted correctly!

For example, let's say we had a survey where we asked the user to enter their age in a text box and someone entered `"$$"` instead of a number. Since we can't visualize that and it doesn't fit with the rest of our data, we can discard that data point!

Making Data Usable

Another factor to consider in data collection is the method by which you're actually collecting data! If you want to do a survey of people in your class, you could have them:

write down their answers on a piece of paper

or

enter their answers into a spreadsheet

Making Data Usable

Let's say you want to find out the eye color of everyone in your class. There's a few options you could pick to do this:

Have everyone give you a physical picture of their eyes

Have everyone send you a digital picture of their eyes

Have everyone fill in a survey where they write in the color of their eyes

Have everyone fill in a survey where the colors can be chosen from a dropdown menu

Moral of the Story

Computers automate data collection, which makes it easier than ever for us to gain all kinds of information about the world around us.

Storing and accessing this data is challenging, though technologies have been developed and put in place to alleviate the challenge.

There is a LOT of data collected and stored every second!

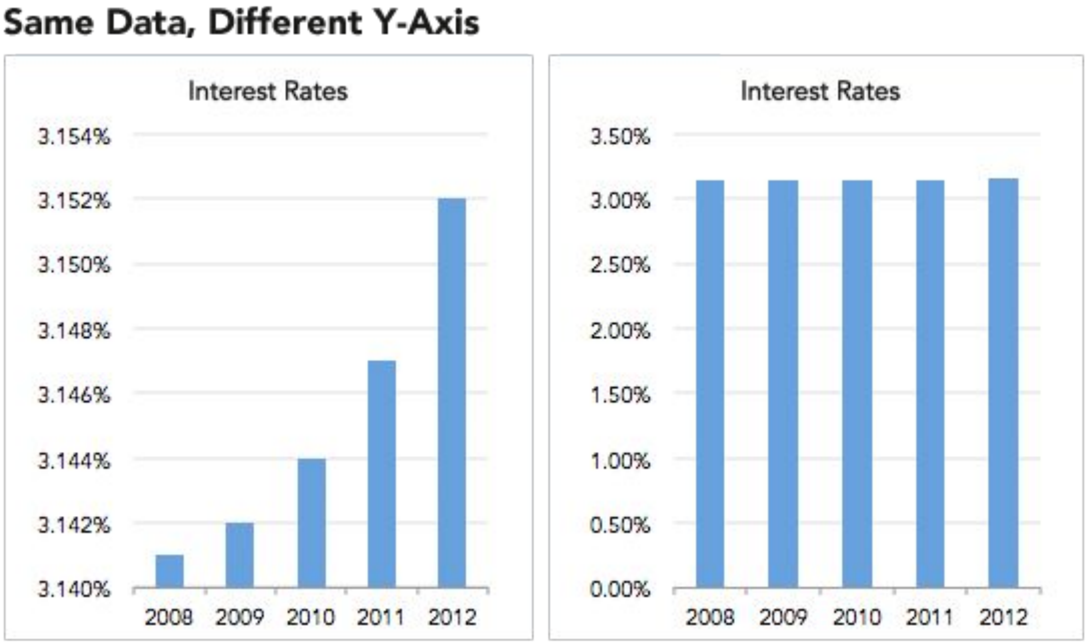
Data Limitations

Data Limitations

While visualizations can make make reading and interpreting our data much easier, there are still limitations to what we can conclude from examining them!

Visualizations can be misleading by skewing the axes, messing with the labels, or leaving out relevant data points.

Misleading Visualizations - Truncation

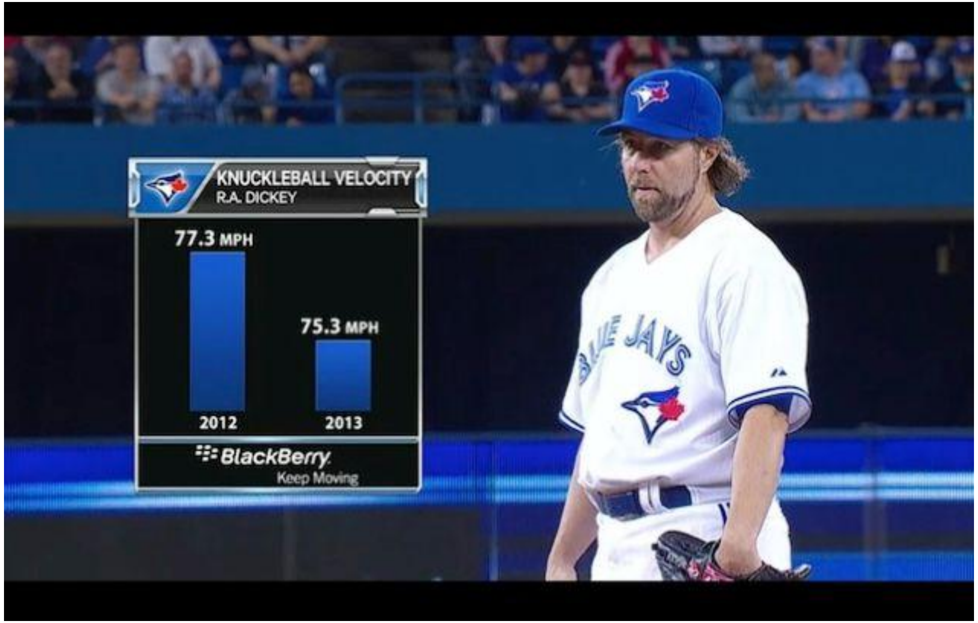


One way we can make our data misleading is by **truncating** the axes.

The difference between these graphs is only the y axis - on the left, it has a range of 0.014%, where on the right it has a range of 3.5%!

Source: Agata Kwapien http://www.datapine.com/blog/misleading-data-visualization-examples/

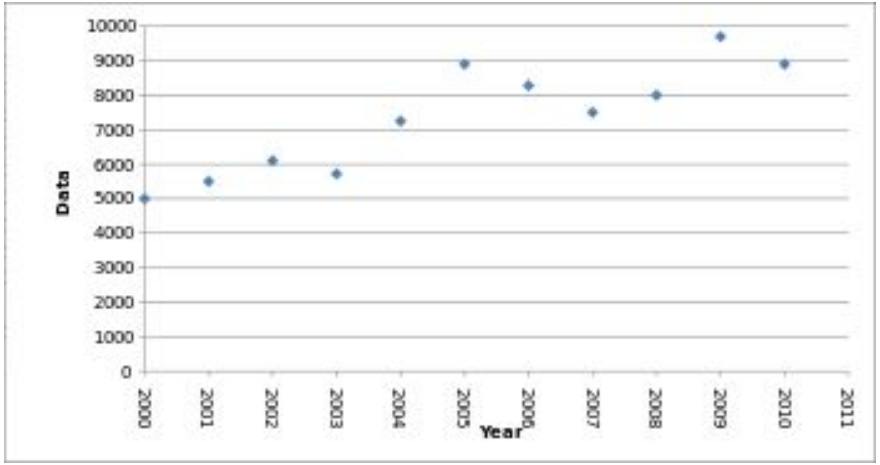
Misleading Visualizations - Truncation

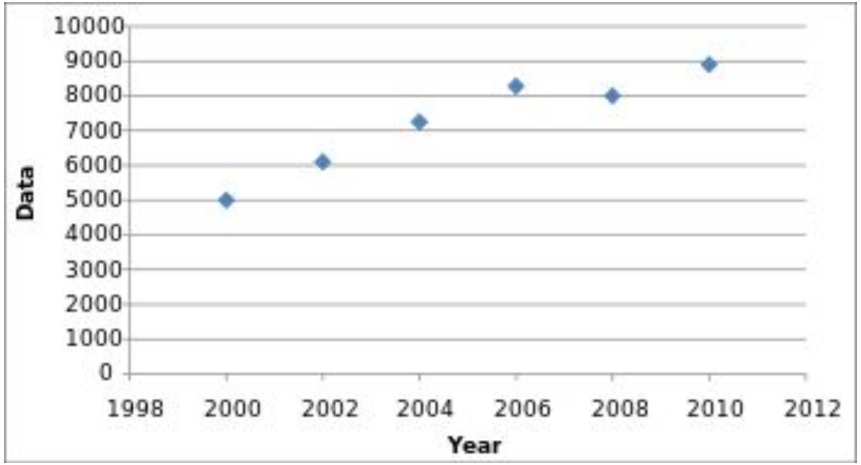


Here's another example of y-axis truncation! We can see that the difference between the two numbers is only 2 mph, but the bar on the right looks half as tall as the bar on the left!

Source: Ravi Parikh http://gizmodo.com/how-to-lie-with-data-visualization-1563576606

Misleading Visualizations - Omit Data





Another way we can make our data misleading is by **omitting** data points to make a fake trend!

If we leave out the data points on odd years from the graph on the left, we see a mostly upward graph on the right!

Misleading Visualizations - Break Convention

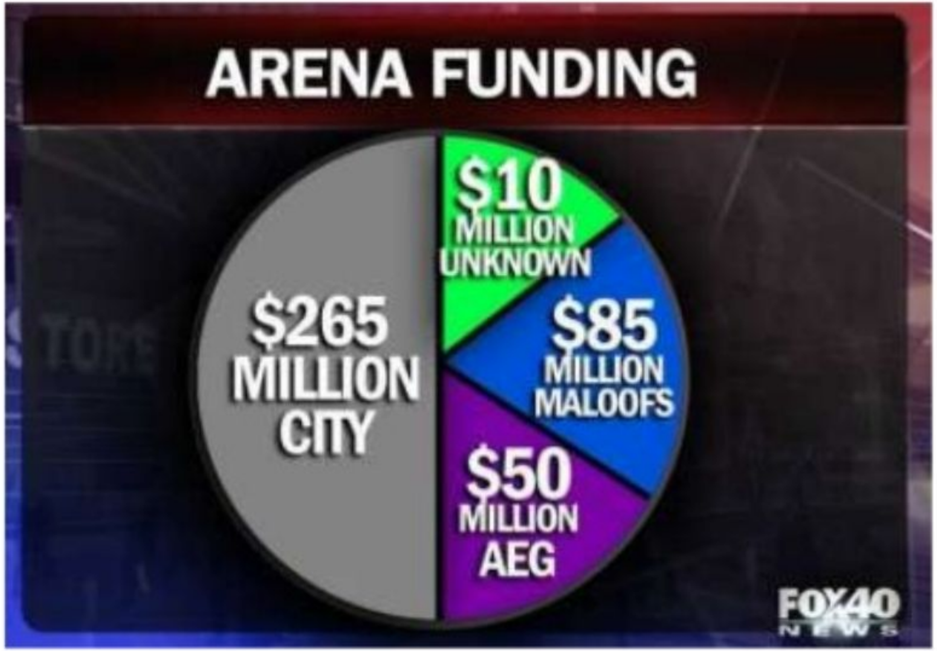


Image Source: http://www.visionjar.com

If we break convention, we can trick viewers into thinking the data says something it doesn't.

Pie charts SHOULD always add up to 100%, but if we look at this chart, it shows that the city is contributing 50% of the funds.

Misleading Visualizations - Break Convention

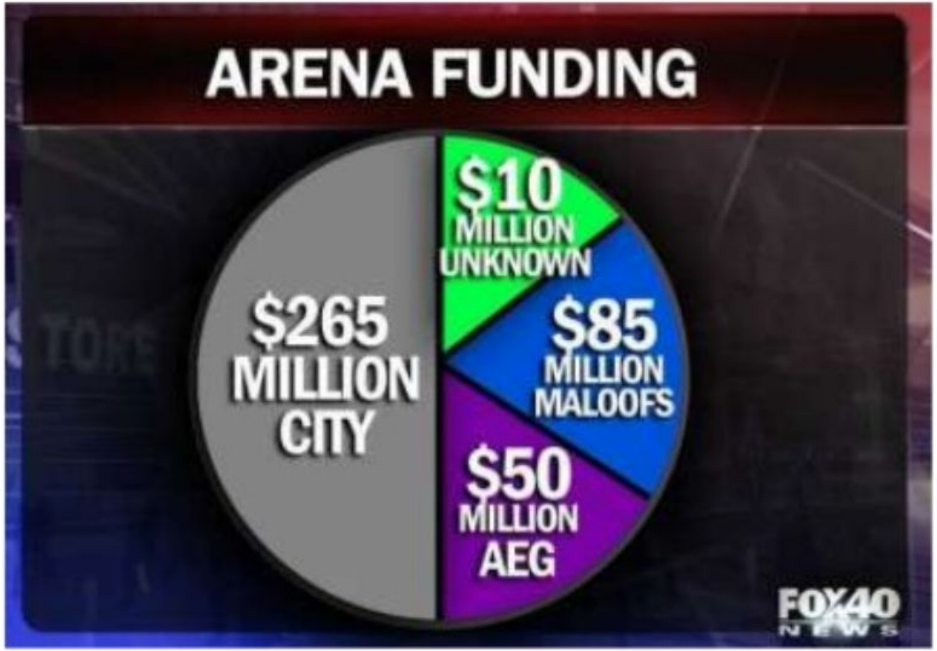
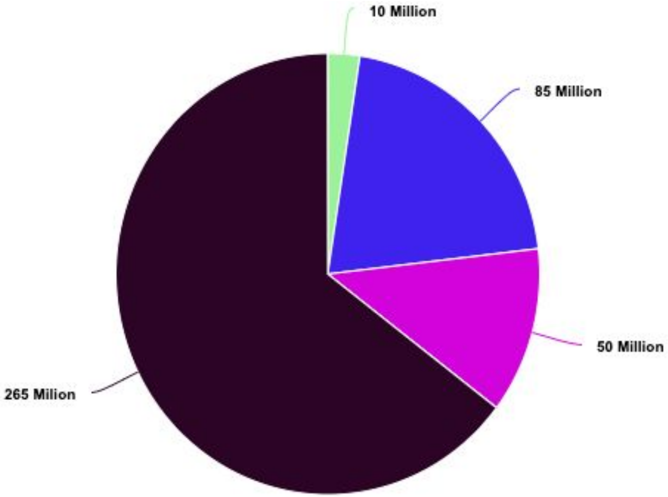
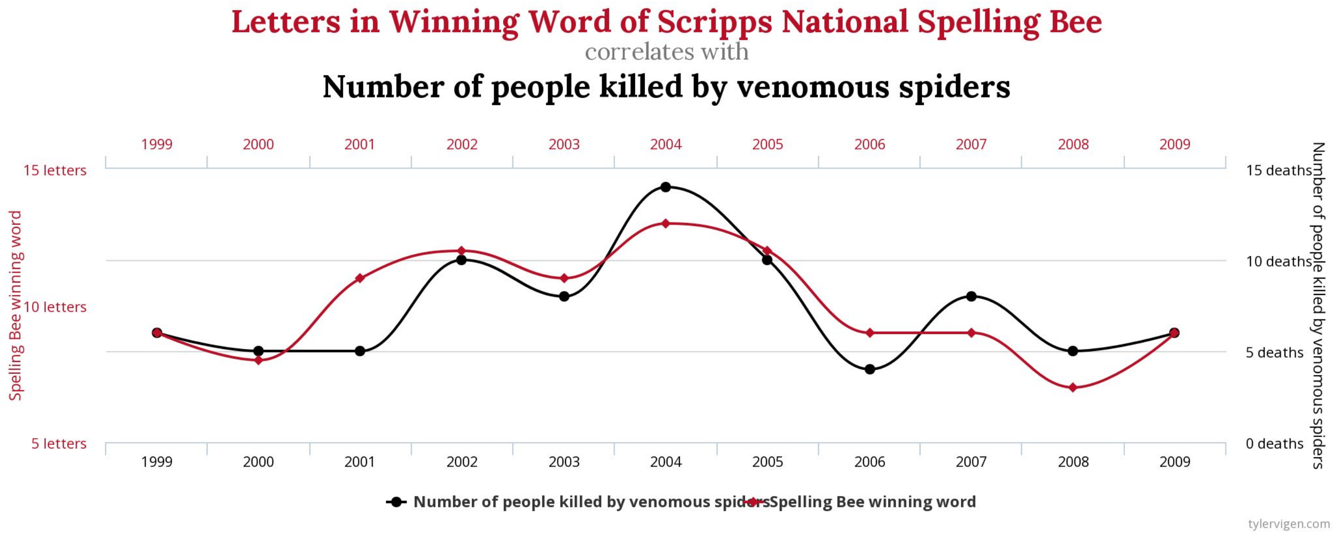


Image Source: http://www.visionjar.com



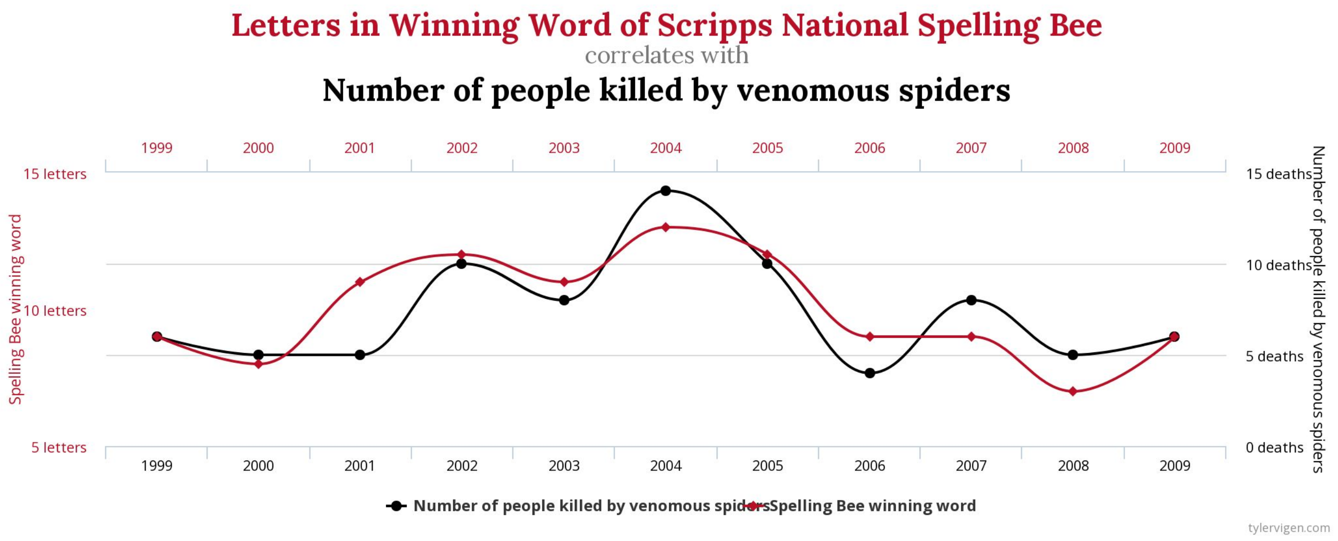
The actual slice sizes are significantly different than in that graphic, though!

Misleading Visualizations - Correlation



One last thing to remember is that **correlation** is NOT the same as **causation** - even if two things are happening at similar rates, that doesn't mean they're tied together!

Misleading Visualizations - Correlation



A website that documents many of these silly correlations is https://www.tylervigen.com/spurious-correlations.
You can even find your own correlations!

Metadata

One very important factor when we're working with data is that data's **metadata**. Just like in the last unit, **metadata** refers to the information *about* the data! We need to be careful to examine this information to see if there's any bias in the data being analyzed. Here's some questions we should ask:

- Where was this data collected?
- Who collected this data?
- How long ago was this data collected?
- How large is this data set? Is it accurate?

Limitations

Some questions can’t be answered if we don’t have the correct data!

For example, you can’t plot annual trends over time if you only have data from 1 year. Similarly, you can't determine worldwide opinions if you only interview Americans.

Being aware of the limitations inherent in the data you've collected is important for ensuring that you do a valid and meaningful analysis.