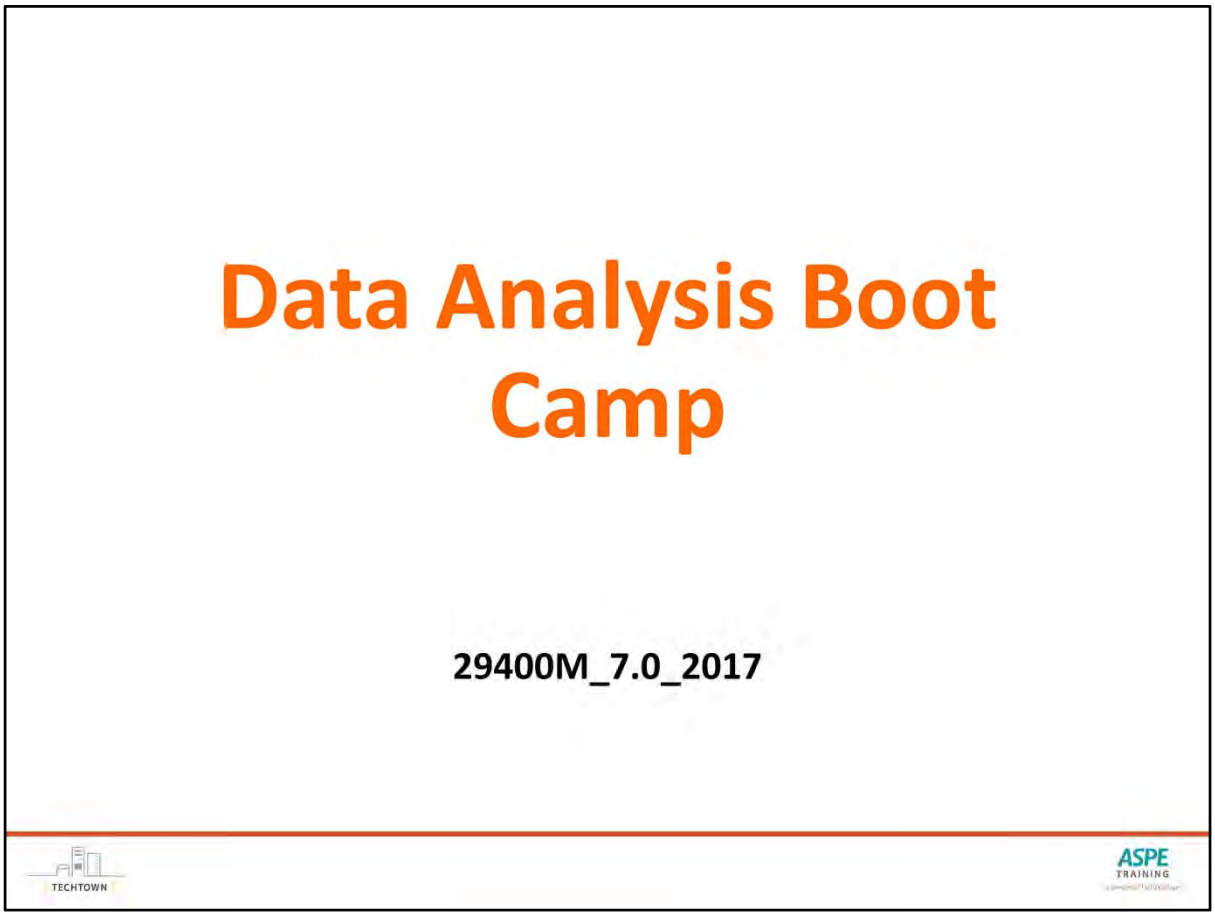
Data Analysis Boot Camp

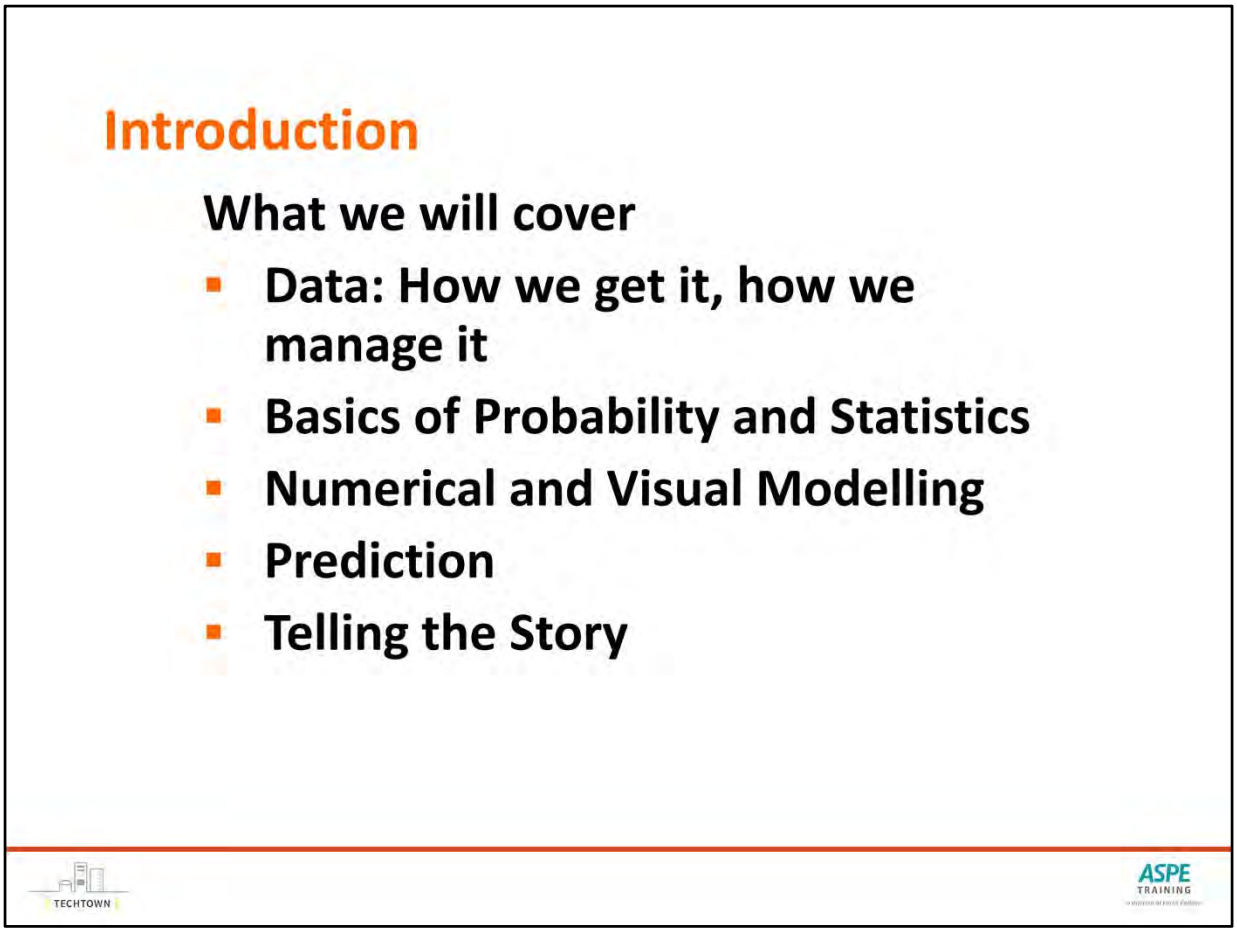
Classroom Manual

Learn practical, hands-on data analysis skills and tools to maximize and leverage your data.

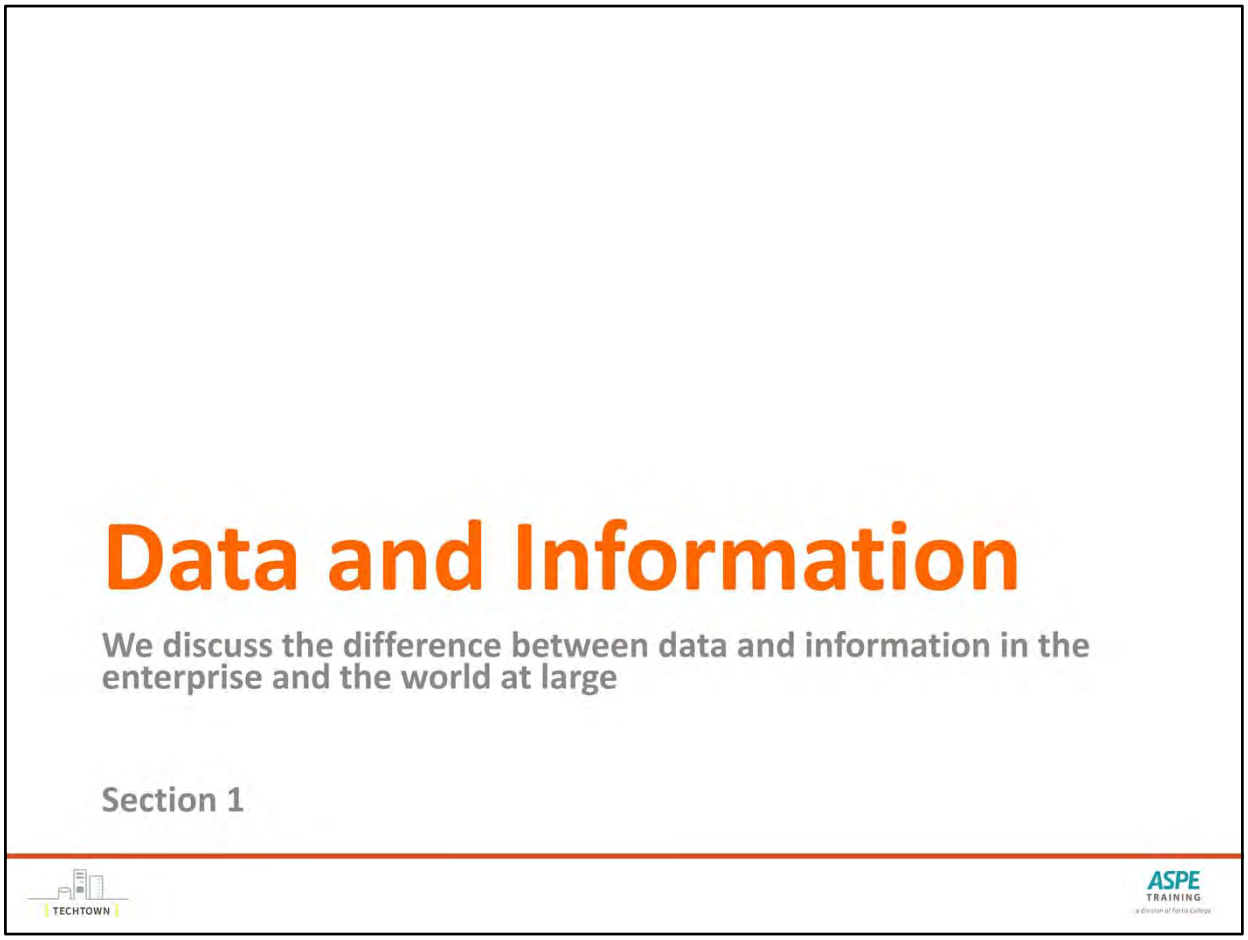
29400PG_7.0_2017



NOTES:



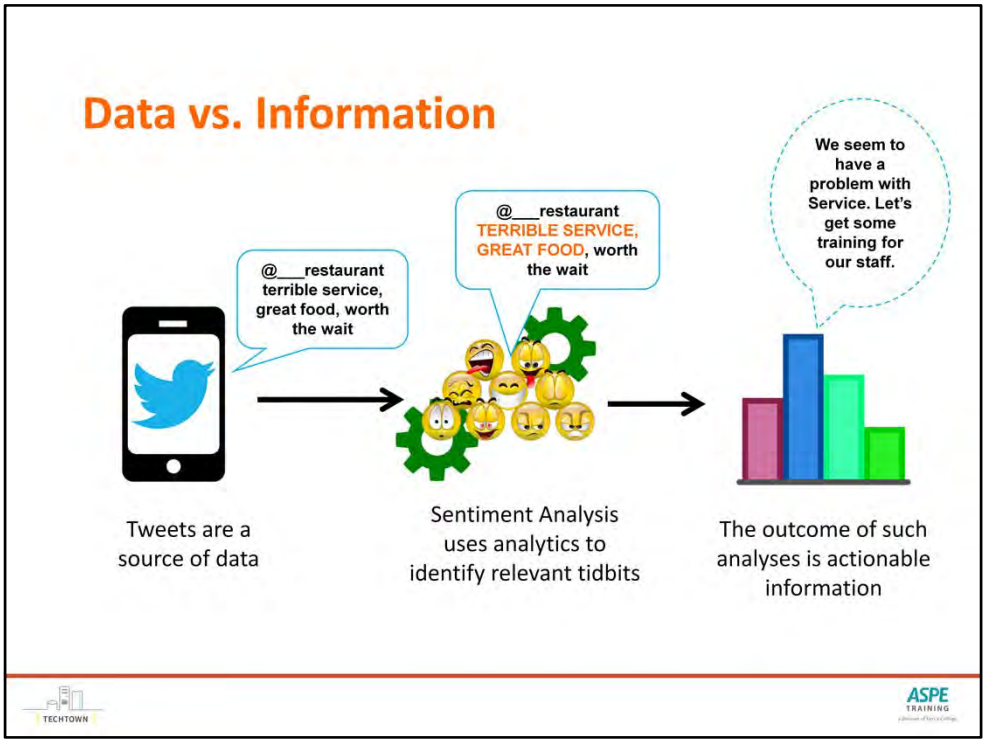
This class discusses data as it is in the real world. When analyzing any data, the first step is to get data. Very often these come from external sources and will require extensive methods to evaluate, clean, store and manage. Once that data is "in house" a variety of techniques can be applied to extract information from the data you have managed to collect. All of this starts with a basic understanding of probability and statistics. Analysis techniques applied to data in the real world are very often statistical in nature for the simple reason that the world is a complex place. We don't know many "textbook" equations that we can use to determine a business cycle, an economic cycle, or social phenomenon. We have to gather data and start to model it. Then we can work from the models to create predictions, but these are inherently hazy. They follow probability and statistical rules. Starting from these, you can gain insight into what complex software algorithms are telling you. After you gain confidence in your predictions, you can begin to tell the story of your conclusions.



NOTES:

Data is all of the little bits and pieces of digital "things" we generate every minute of every hour. It is not of itself information that can be immediately used for productive conclusions. Data must be gathered, probed and prodded, and understood before it can be brought into the organization for useful analysis. This is called profiling. After the data is profiled, it must be cleansed of all sorts of inconsistencies that is inherent in the way all of these bits are captured. Not all data is numerical in its nature, so some sort of scheme must be applied to quantify anything we want to analyze. The scheme must be numerical in nature because we are dealing, in the end, with numbers. The numbers are how we model, how we predict, how we determine data that is not where we might expect it to be (outliers).

Data in the real world is growing rapidly, and so are ways to gather it and process it. How we do so is really the scope of many classes of programming and data processing, but they are growing almost as fast as the data itself.

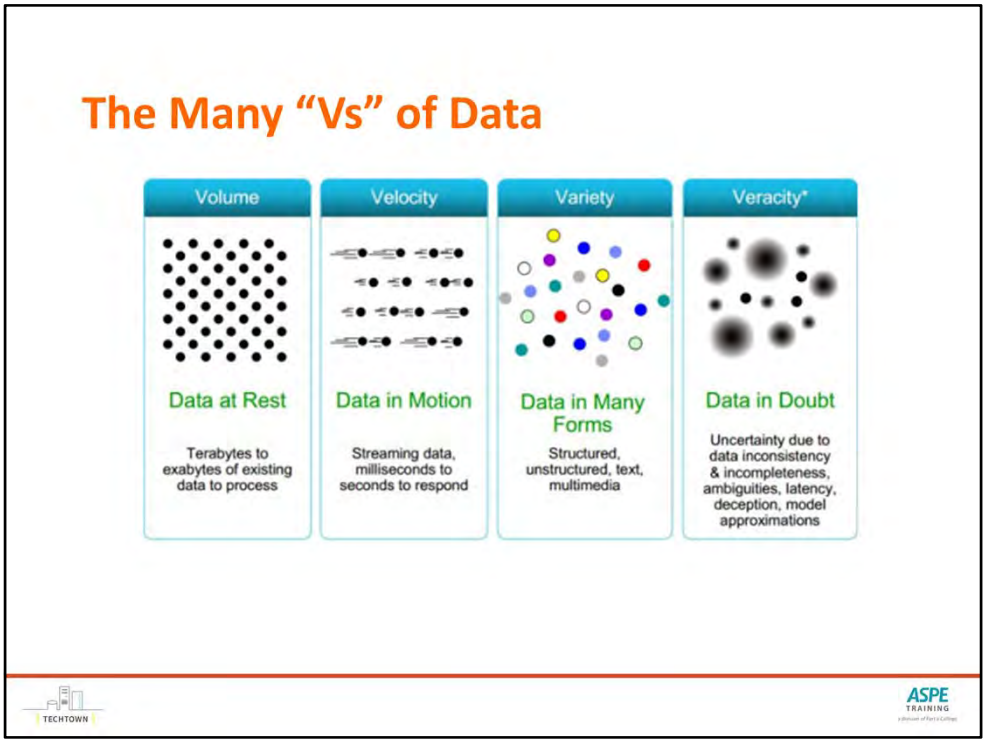


This is the path of data into the enterprise from random bits to actionable information from which conclusions can be drawn. In general, this is done by identifying the type of data you want to collect and perhaps the sources from which the data should come. Then, programmatically, the sources are polled for their data and the software brings the data into the enterprise. For example, the day's exchange rates for all currencies might be pulled from a web service at your favorite bank.

When the data is gathered, it must be stored and managed. There are many ways to do this, the most common in the enterprise being the relational database (Oracle, MS Sql Server, MySql). At this point, database professionals work with the analysts to clean the data to make it usable…. but more on all this later.

Consider Twitter and its effect on customer service…
Twitter is often where people will go to express their opinion of a company's customer service. The person's tweet is issued from the Twitter software service and captured by any client software listening for it. How do we gauge the customer's opinion? The tweet is natural language with all possibilities of interpretation as one would expect. Moving from natural language to quantifiable aspects of the mood those tweets are trying to convey is known as Sentiment Analysis and involves complex algorithms to determine the mood the tweets are trying to convey. Once the mood is read and quantified, analysis can begin and actionable information developed.

Now in the real world, we cannot develop a model of customer service progress from one person's tweet(s). We need to identify all customers who are on Twitter talking about the company. When we get enough, the sentiment analysis of that scope of data can turn into real information.



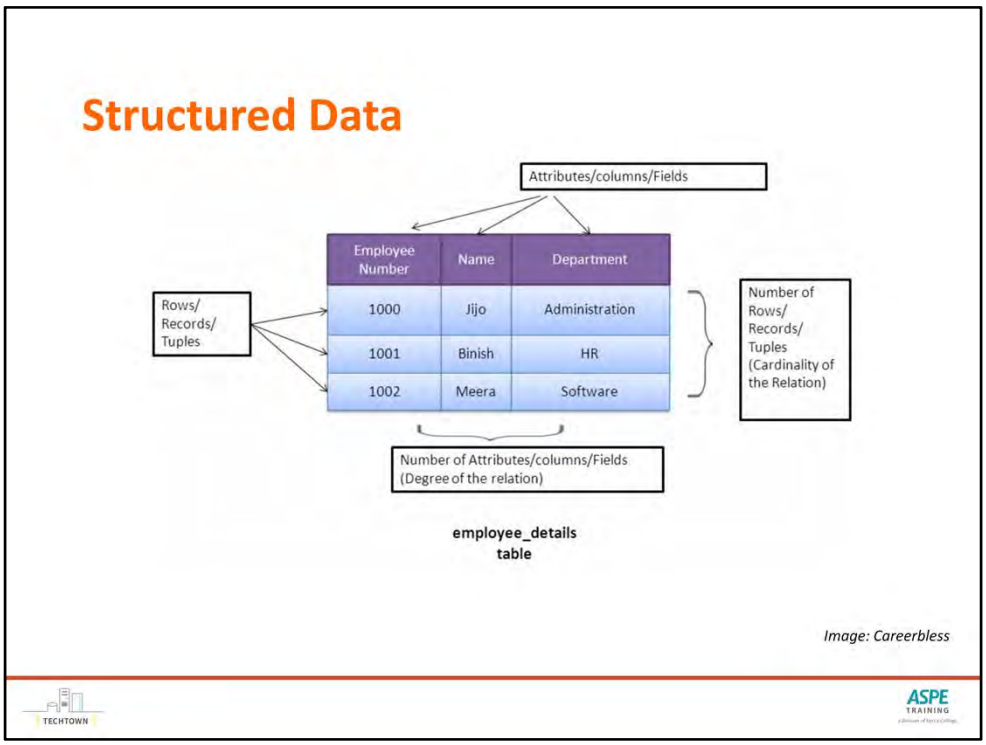
The process of bringing data into the enterprise requires the consideration of the Vs of data

Volume – what is required, what is available, and what can be stored internally all force one to think about the amount of data (Mega, Tera, Peta) being considered for a given project

Velocity – issues of availability of data due to the physical limitations of moving the data through the networks have to be considered when planning an analysis project

Variety – bringing data from the many different external sources (tweets, Yelp, Facebook, Government centers, etc) to be used internally requires transformation from many different forms

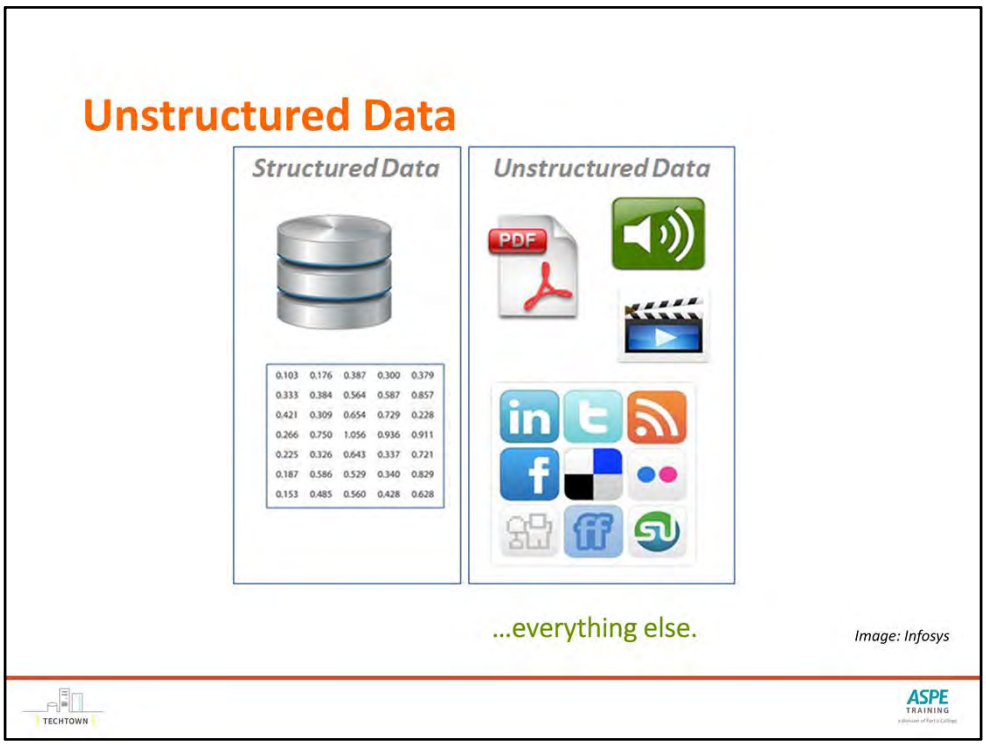
Veracity – when using external data, what is the quality of that data? Is it good enough to lead to actionable information with a high degree of confidence?



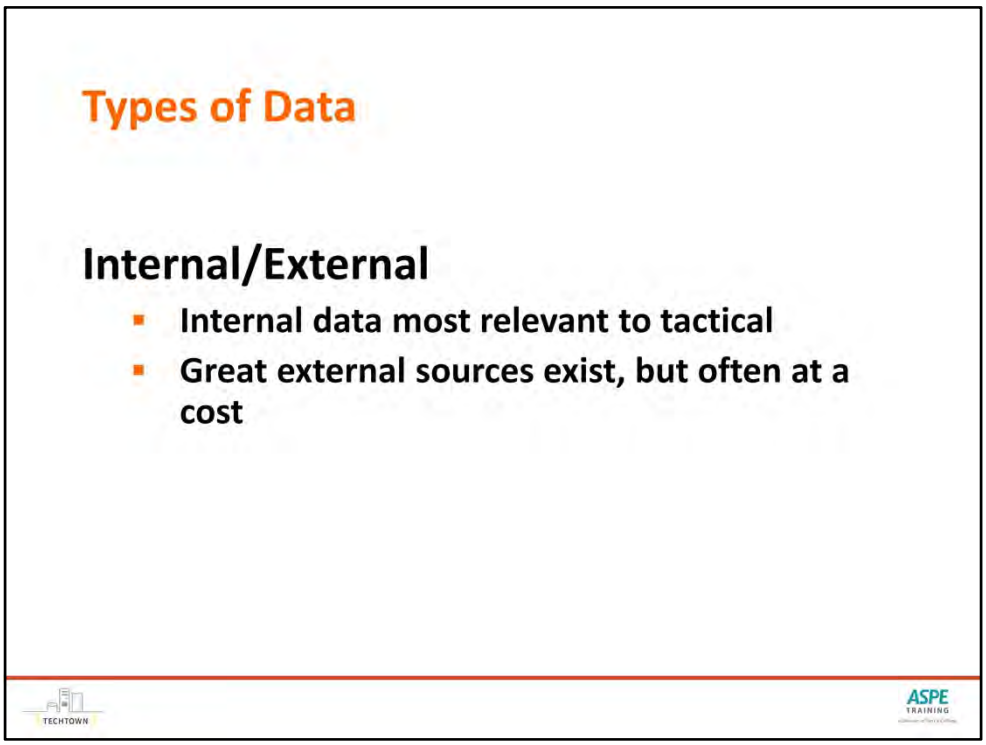
The internal enterprise usually makes use of highly structured data. Relational databases are the norm inside of most major organizations. These implement rules around highly structured collections of information enforcing:

- Data types
- Relationships between tables
- Constraints on what data can be entered (beyond just data type)

When data is managed by a RDBMS, a lot of the dirtier aspects of data are automatically handled. This is because the data types and relationships will “kick out” anything that violates these when the data comes into the enterprise. Of course, this can also mean that some valuable data is lost. Why is it valuable if it is dirty? Depends on the situation. For example, some survey responses may be rejected because the respondent did not enter their city in the form. It all depends on what you decide is necessary when you frame the problem. We will see this many times throughout this course. In the end, the analyst is indispensable. The automated processes and analytical tools are not foolproof means to an end.



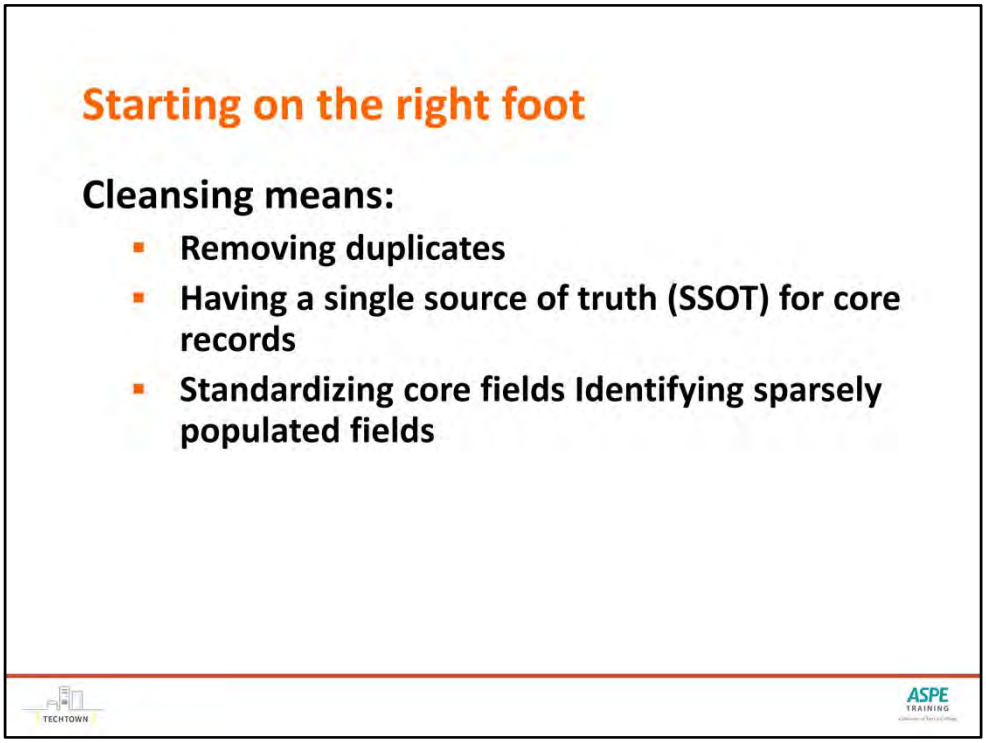
Even though different applications like PDF and MPEG have a structure, the data placed in them can be unstructured. Consider the title of a document in a PDF. PDF rules may be enforced to center and bold the title, but the words in the title itself may be anything from simple text to text with numbers and symbols, of a variety of lengths and case. This way, PDF readers know how to typeset the document, but the data they are typesetting is not in any structure in the RDBMS sense.



Internal data is data that is within the enterprise. The data is created by the enterprise, stored and managed by the enterprise. It spends its entire lifetime under your control. External data comes from somewhere else. It is gathered by you (or someone with your team) but its initial integrity is under the control of an outside group. Generally, if you want high quality data, that data comes with a fee. However, it may be that the fee is reasonable compared what it would take for the internal team to develop it.

Consider demographic data for an advertisement you want to run. There are companies that specialize in maintaining demographic databases that you can purchase. They may be expensive, but to reproduce this data on your own requires a considerable effort and expense.

In the end, bad data is worthless and potentially very damaging to the organization. The damage can be in many forms, but, ruling out actual malicious code, the damage is in the form of erroneous conclusions for the enterprise. A great deal of money can be lost if conclusions are faulty and the first place to start to prevent that is to ensure that the data is clean. What does it mean to be clean? We will cover these in the next slides.

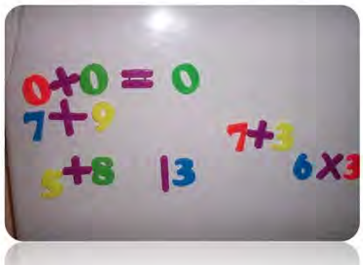


One aspect of dirty data is data duplication. Perhaps you are tracking a group of people for a medical study over a period of one year. Imagine if you do not account for spellings of a person's name when you take the second or third survey – Jane Doe submits the second survey as J. Doe and becomes two different people in the database. Now the data is not tracking Jane's first and second surveys, but you are tracking two different people. One simple way to prevent this is the Single Source of Truth where every participant is tracked by a unique ID and the name is keyed into the DB once and only once. This prevents multiple copies of a name in multiple places in the data that are seemingly different but actually the same.

Starting on the right foot

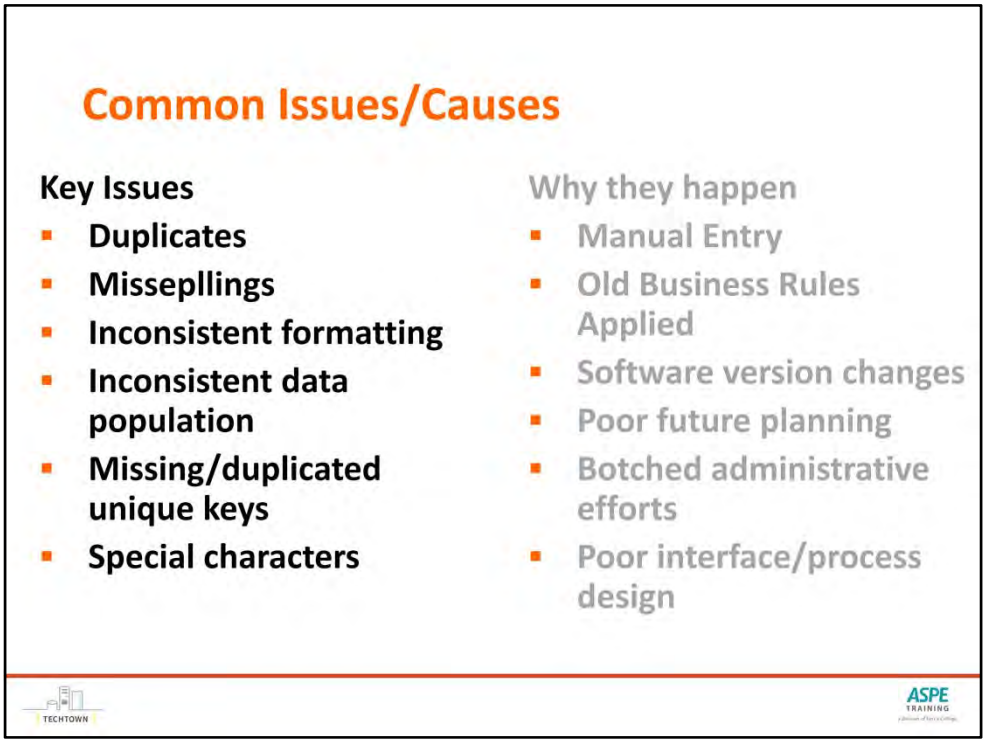
The risks of skipping or neglecting data cleansing:

- **Inaccurate analysis**
- **Wasted time/money**
- **Inconsistent results**



There are many ways that dirty data leads to these results. Some more ways,

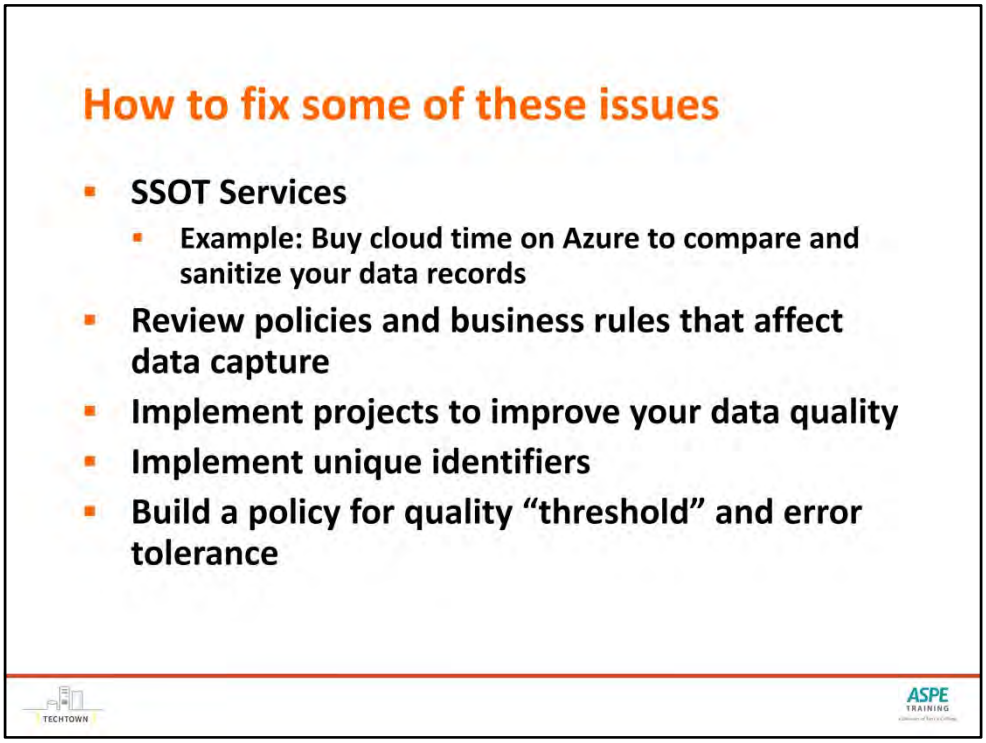
- Duplicate records can lead to improper sample sizes when performing simple statistics
- Data sources which provide continuous data updates (say from stock information) can generate inconsistent trending information when the data is not treated properly on import.



These are a list of some common causes of dirty data. Many of these are caused simply by working with manual entry. For instance, hand typing a department in a text field instead of offering a dropdown list of available departments the way most state dropdowns are available. The Y2K problem was an example of poor future planning as only two digit years were required for a very long time.

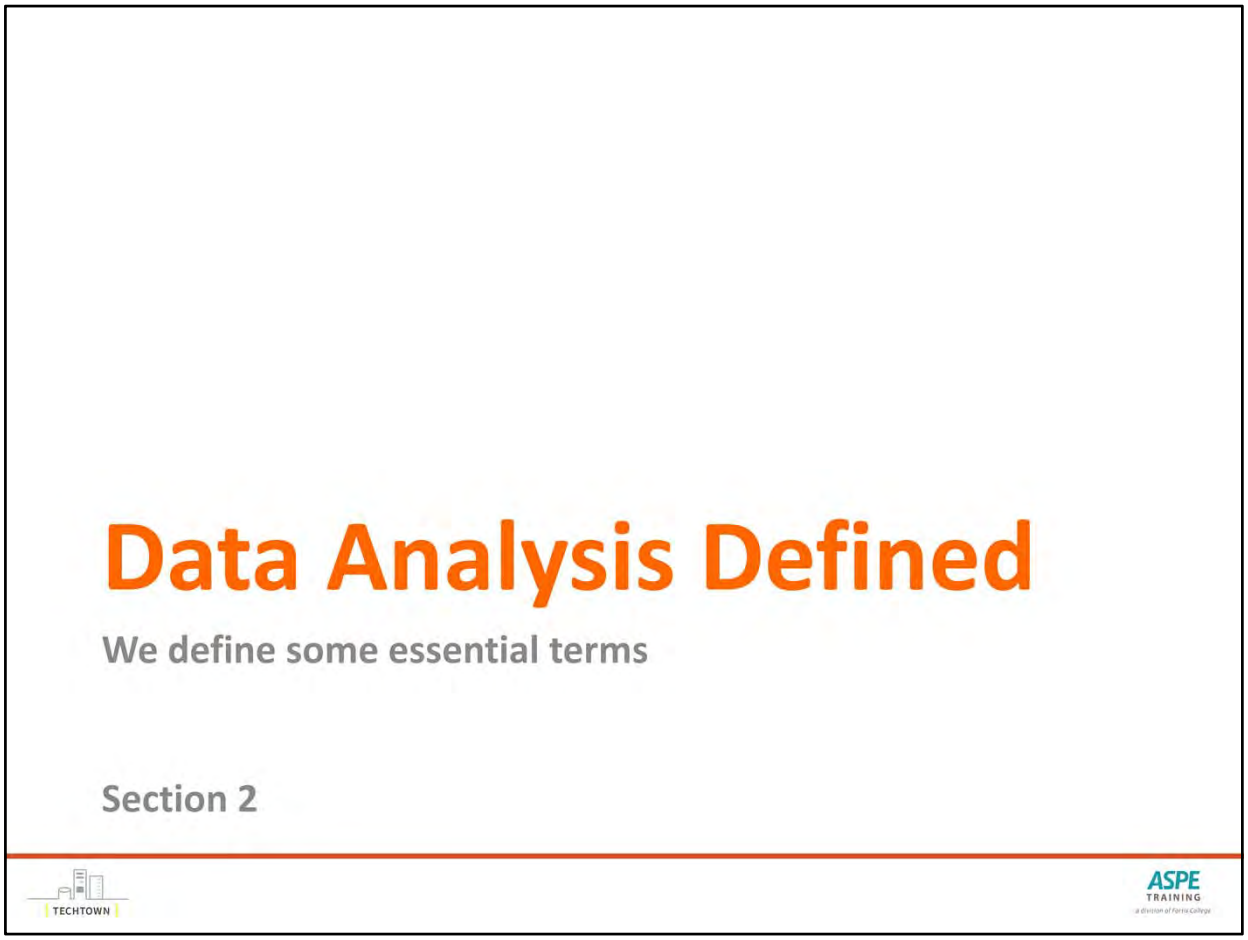
To look at just a few…

Inconsistent formatting is often found in dates.
Misspellings in names of people and places (from registration forms or surveys)
Inconsistent data population is often found when downloading an Excel file from an external data source. This is because data was not rigorously collected.

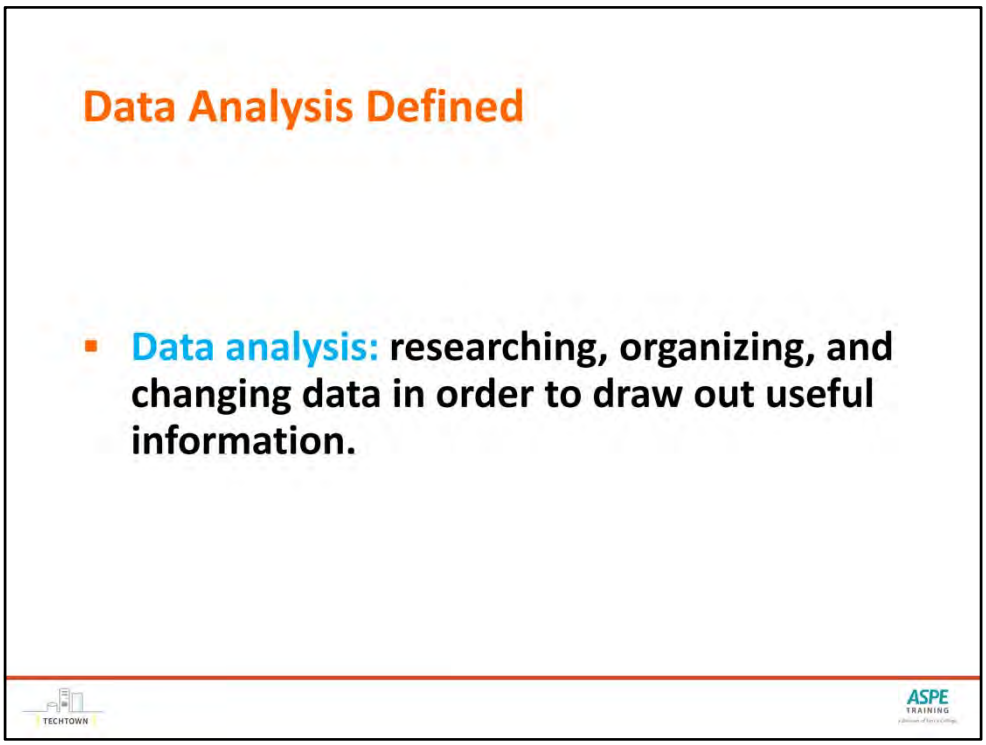


We already spoke about SSOT. This is very important. Consider the state dropdown mentioned a few slides ago. The SSOT for this is usually a database table with all of the state codes in it. Perhaps each one has an integer ID in the table as well. When there is any record stored for a user filling out this form (whatever form it may be), the state code is referenced by the ID in the table. The presentation on the screen comes from the SSOT table so that the spelling is always the same. When you need to generate a report, or view the state data for a given person in any way, you resolve the keys linking to the SSOT table, so that you always get consistent spelling.

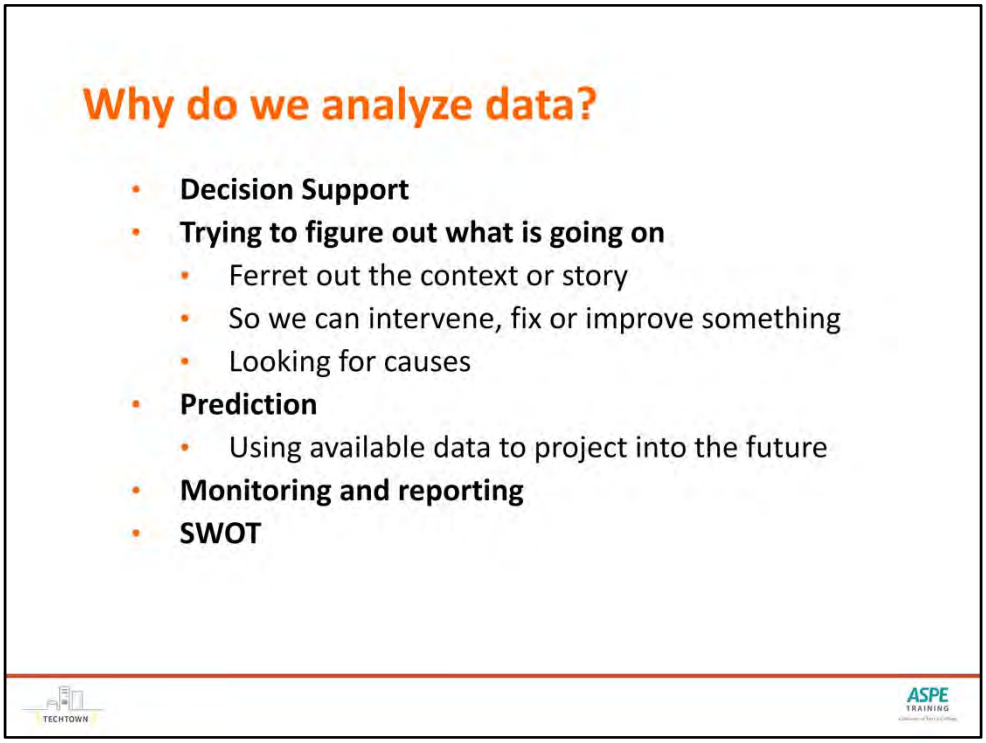
If you have older projects, you can implement projects to review the older data and determine how it may be fixed. Even with new projects, review policies are important. Having a policy to kick data out when you need to preserve data integrity is usually referring to the minimum ”atomic unit” of data. The analytics team has to decide the minimum effective amount of data in any record. If the data integrity dips below the minimum, then it should just be rejected and noted in order to maintain quality of any conclusions drawn from the data.



NOTES:



Data is everywhere, but of itself is not useful. Information comes from analyzing data, but what does it mean to analyze data. There are, of course, specific analyses that everyone is familiar with – averaging the semester grades, percentage of people who shopped at Wal-Mart in the last 10 days, etc. To analyze data is to perform research in a field, organize data that is relevant in that field, and shape it in ways that tell us useful things. Changing data here does not mean that we change the data in any way we like to draw our conclusions, it means to refactor or reshape the data so that different dimensions are compared together, or to rescale so that two different variables are measured on the same graduation…things like that.

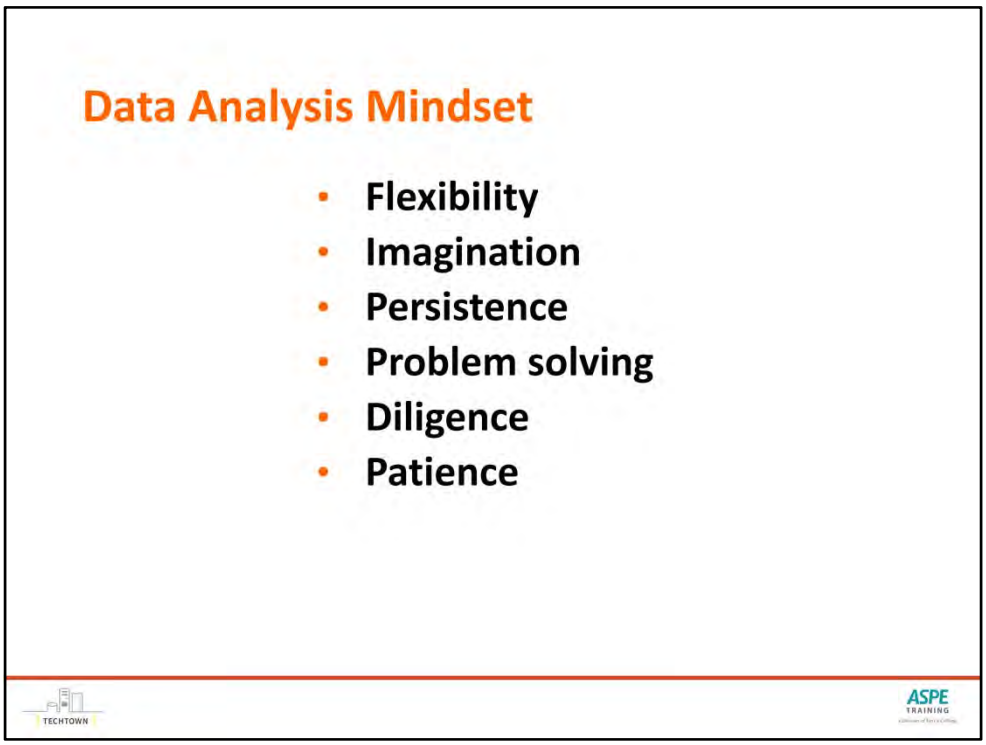


We analyze data because we are looking for some useful information. The information can be to make a prediction, or to see where we have been and understand what happened in the past. Of course, the past is the past, but analysis of data about a past event can help us determine what is to come. Economics is a good example of analyzing past data. A recession is generally defined as two consecutive quarters of negative GDP growth. Basically, by definition, you cannot tell if the economy is in a recession until past data has been analyzed.

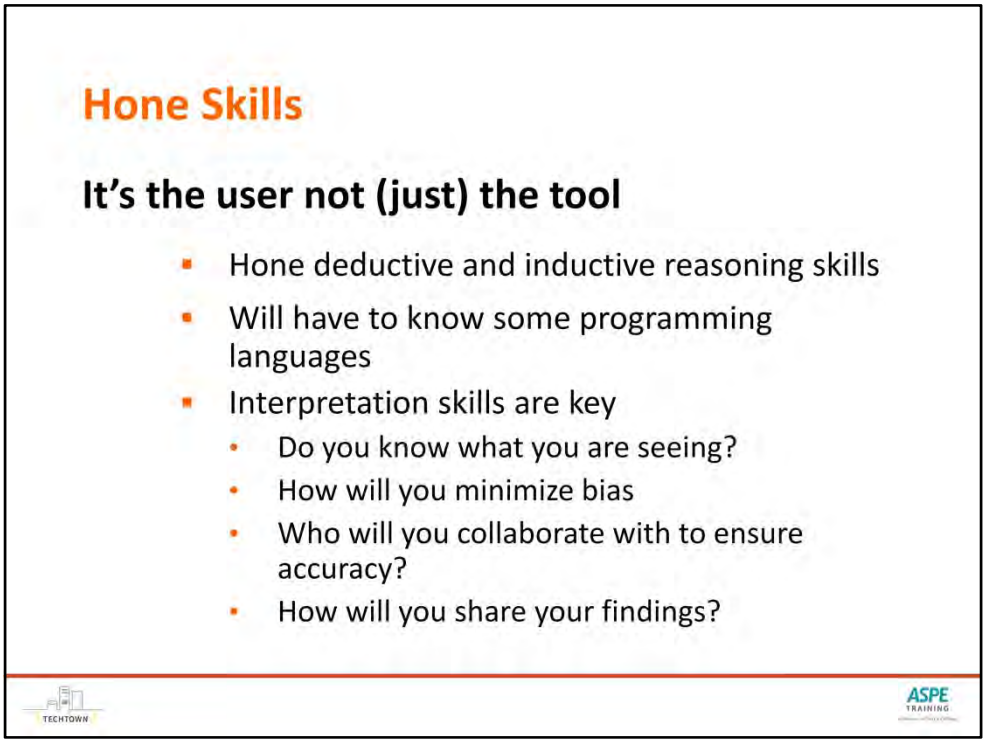
In order to make a prediction, we generally have to have a model. In order to make a model, we gather data and analyze it. Later in the course, we will work with moving data by analysis to a model. From the model, you can push outward and try to make predictions about the future. The accuracy of the predictions depends on many things, no the least of which is the quality of the data and analysis.

Data can also be use for monitoring and reporting problems. Why does it take data to determine if something is a problem ? A simple error in a computer network can just be reported as is when it occurs, but what about reporting credit card fraud? Credit card fraud, or at least potential credit card fraud, is found from mining millions of records and developing a model for what may be considered fraud. Then all new transactions are monitored and compared against data for what is considered normal. If the new transaction fails the “normal” test, then the monitoring system can notify the company about the potential fraud.

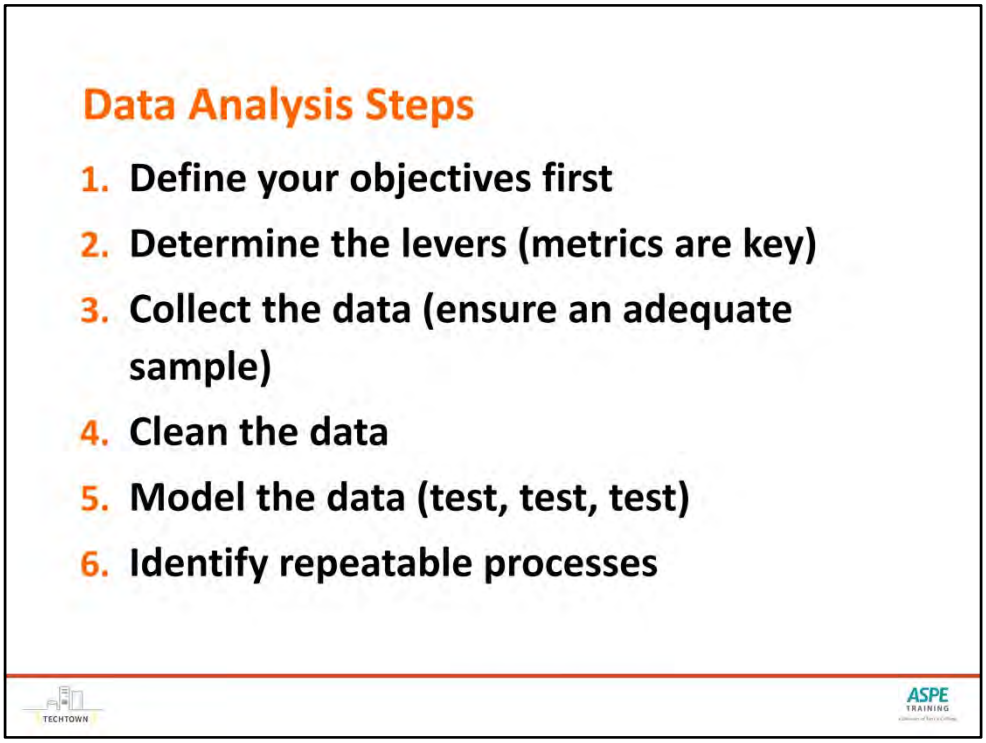
And then there is SWOT – strengths, weaknesses, opportunities and threats. You can read more about SWOT at Wikipedia (https://en.wikipedia.org/wiki/SWOT_analysis) but basically it is the structured approach to business planning. You can answer questions like “who may be moving into our market?” and “what markets may we be missing with our product?” by data analysis.



The mindset of an analyst is one of curiosity and tenacity. Finding the story embedded in numerical data requires a lot of thought and persistence. One must also be comfortable with the lack of clear results. Very often the results of analysis are suggestive, but not definitive. Simply put, there is not an answer, only solutions which are better or worse than other solutions.

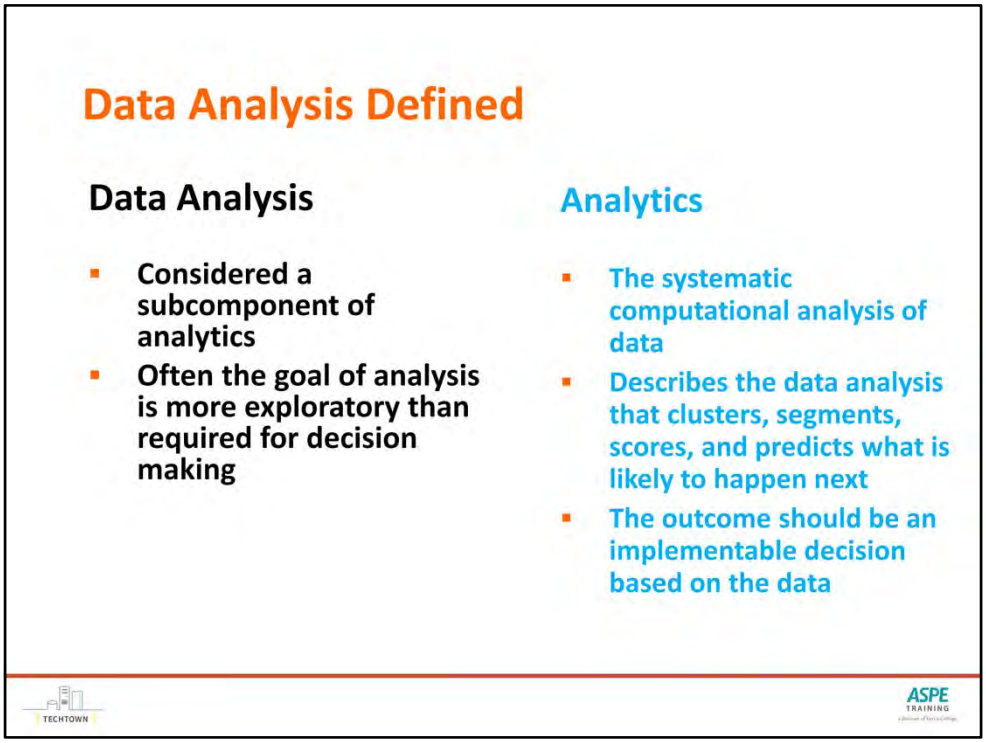


In the end, computers calculate. This is all that they do. It takes a person to discern the quality of the result. That quality relies on the proper framing of the problem, the best data sources, the proper analytical techniques for the field, and good communication of the results.



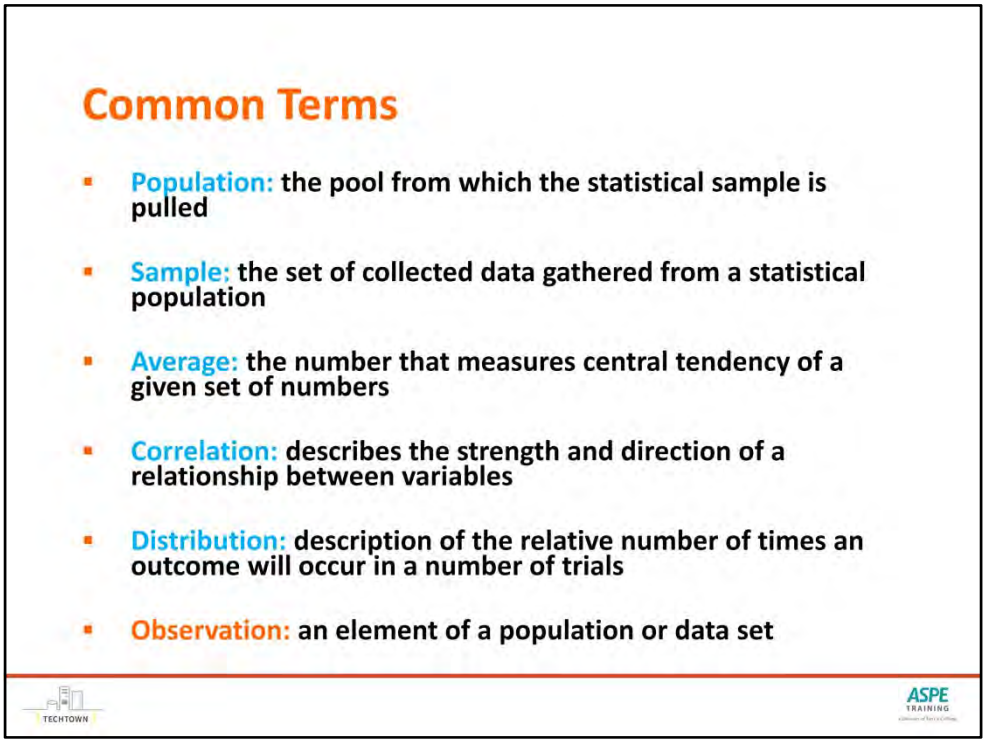
This set of steps is a guideline for the data analyst. Define your problem and what you want to accomplish, determine what metrics to measure and what constitutes success or failure (when applicable), gather the data and cleanse it, analyze and build your model and go over the cycle again.

When building a model, remember to test, test, test. Use the original data and make sure you get out of your model the actual result that the data suggests. Find other data and test against it. Make a short term prediction and test against the actual result. The process is a cycle because each new piece of relevant data improves the model.



The difference between analysis and analytics is one of degree. Analysis is usually reserved for using mathematics to look at data and turn it into information. Analytics often refers to a prebuilt or purchased package of analysis techniques commonly applied to a particular business or science problem.

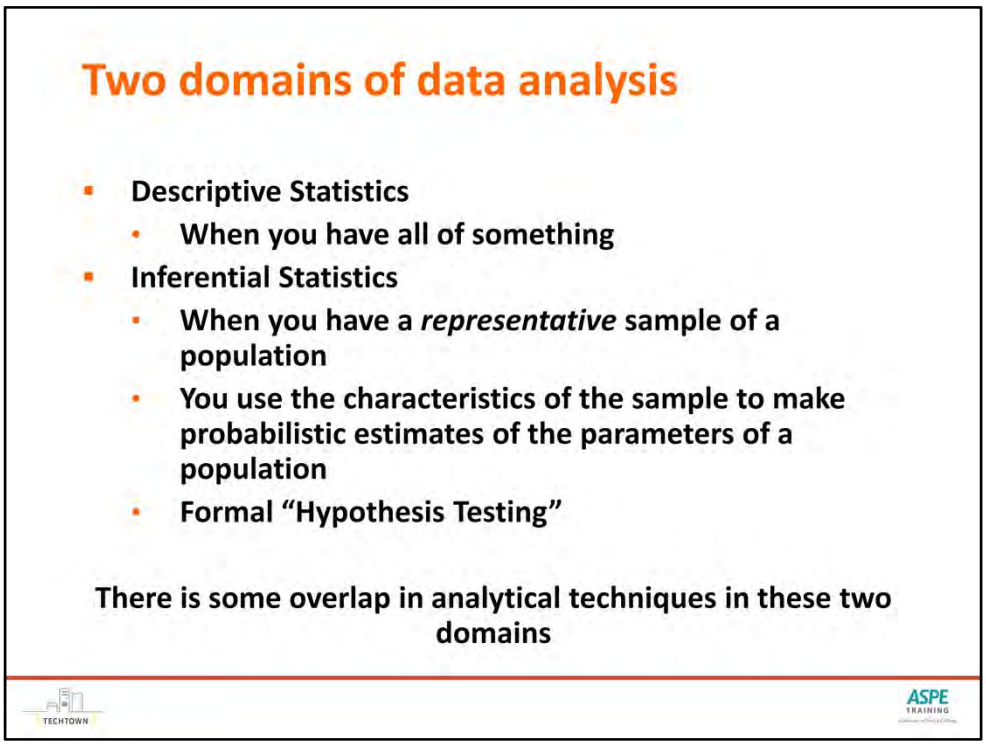
For example, finding averages and percentiles is analysis. An analytics package will run the average and percentiles automatically. It obviously is not that simple, but that is the idea. Analytics is packaging a set of analysis techniques which we know to be relevant to a particular field.



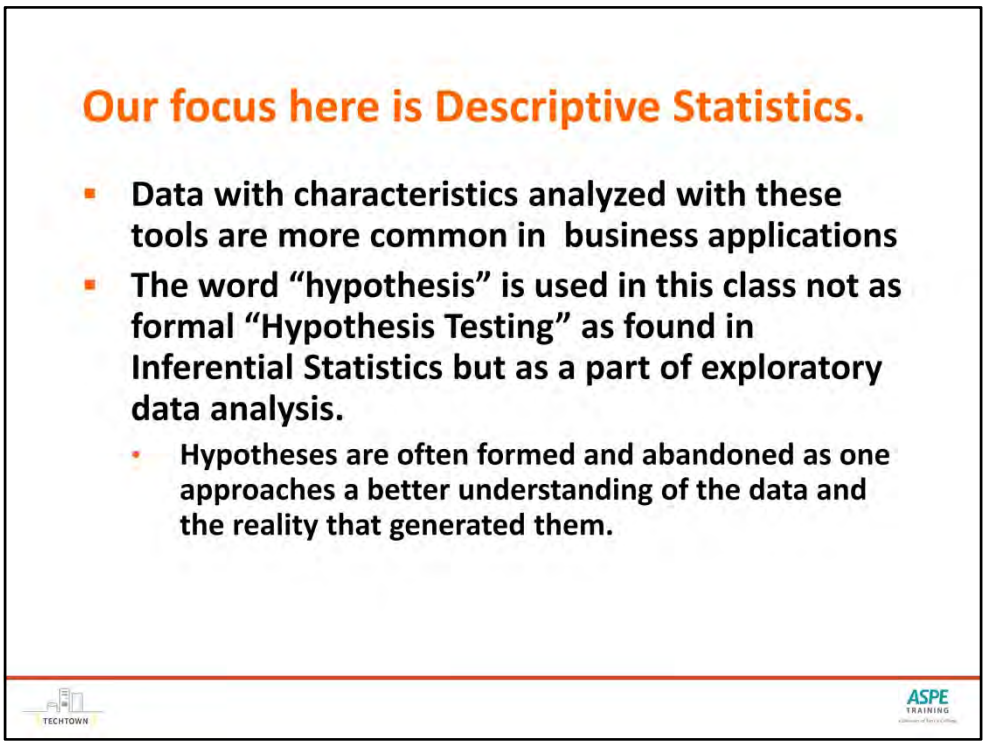
These are some common terms in analysis. Notice that a sample is a subset of a population, but the sample can be the entire population. This is a distinction which leads to two formulas for something called the standard deviation which will be seen later.

The average is never just the average. There are at least three commonly used averages which we will cover in this class. The average measures the central tendency of a set, which means that it describes where the numbers tend to be. It stands to reason that three averages probably measure slightly different things about that tendency. There will be more discussion in coming slides.

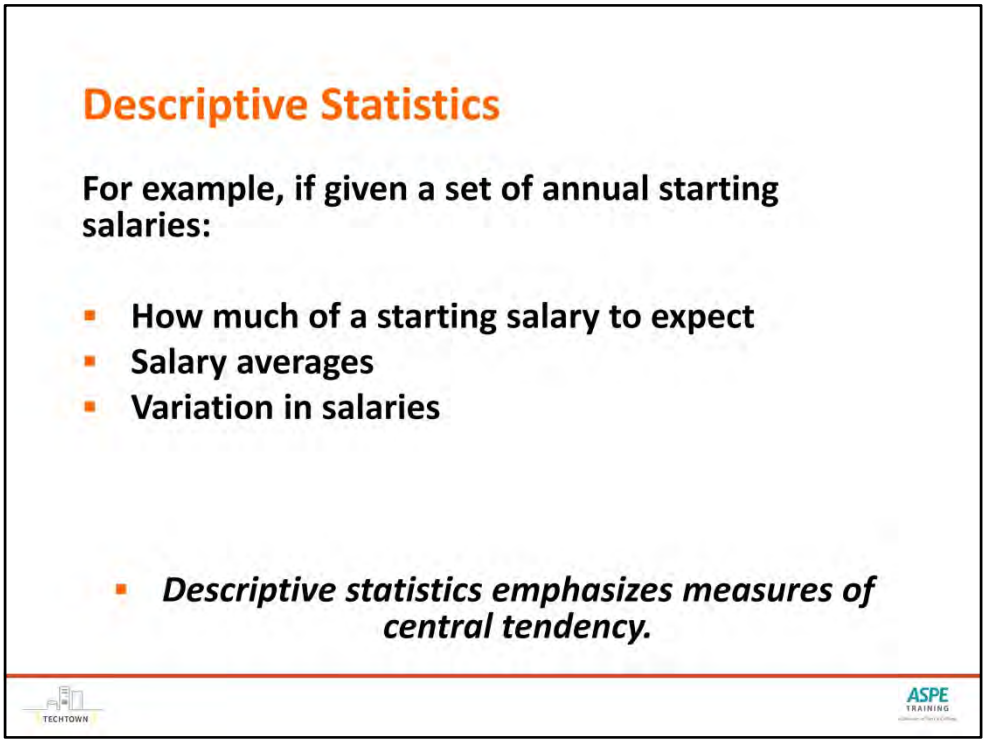
The distribution is an important concept and critical in developing models. It describes how outcomes are distributed. Of course, this uses almost the exact word it is meant to define so lets dig a little deeper. When you have outcomes of an event – coin tosses, games, catastrophic economic collapse on a global scale – the outcomes can be noted and, when performing the same test over and over again, the number of each outcome can be counted. When the outcomes are all noted and counted, they for a distribution.



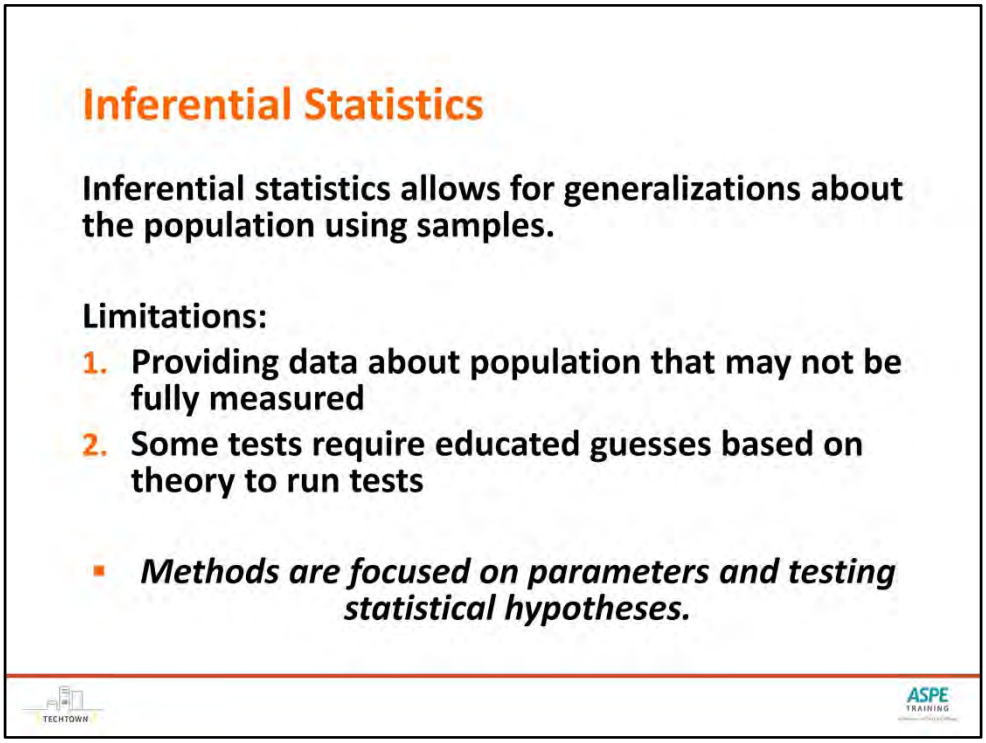
NOTES:



NOTES:



Here is that phrase, central tendency, again. When we have a set of numbers, in this case starting salaries as reported by various companies, the central tendency describes how those number cluster together. That is the point of descriptive statistics – describing how things fit together.



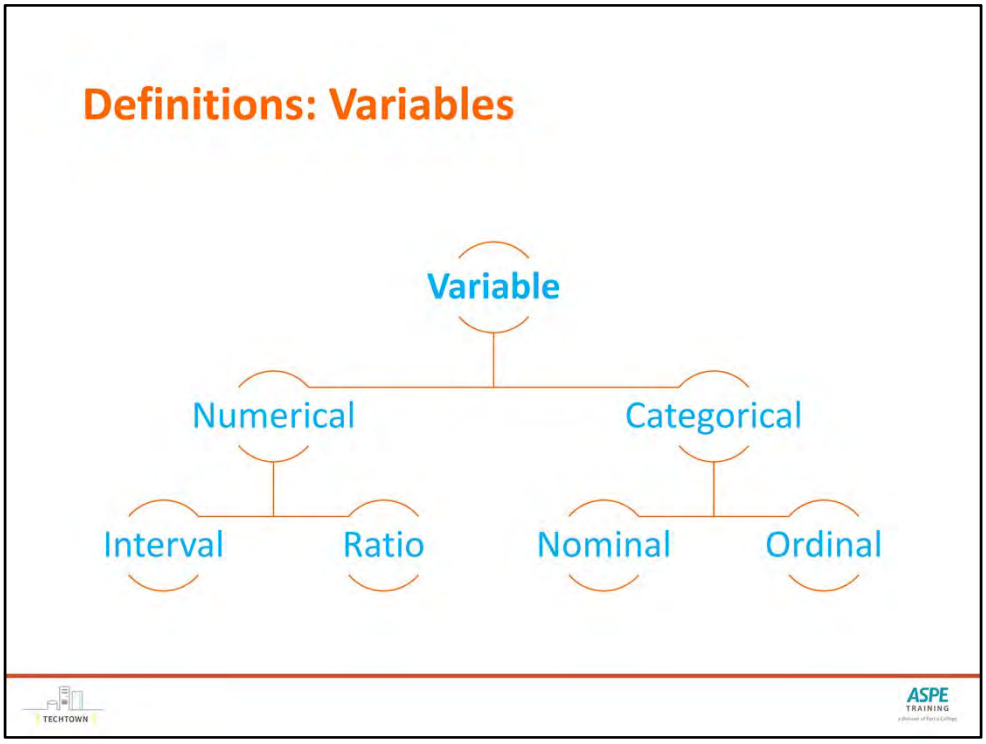
This is much more complex than simple descriptive statistics. The bulk of this course will ignore it, but it is important to be aware of the distinction. Inferential statistics has a broad array of tools and techniques used to come to conclusions about small samples of much larger populations. This is realm of medical studies for pharmaceuticals etc.

Types of Variables

We look at what kinds of variables are involved in data analysis

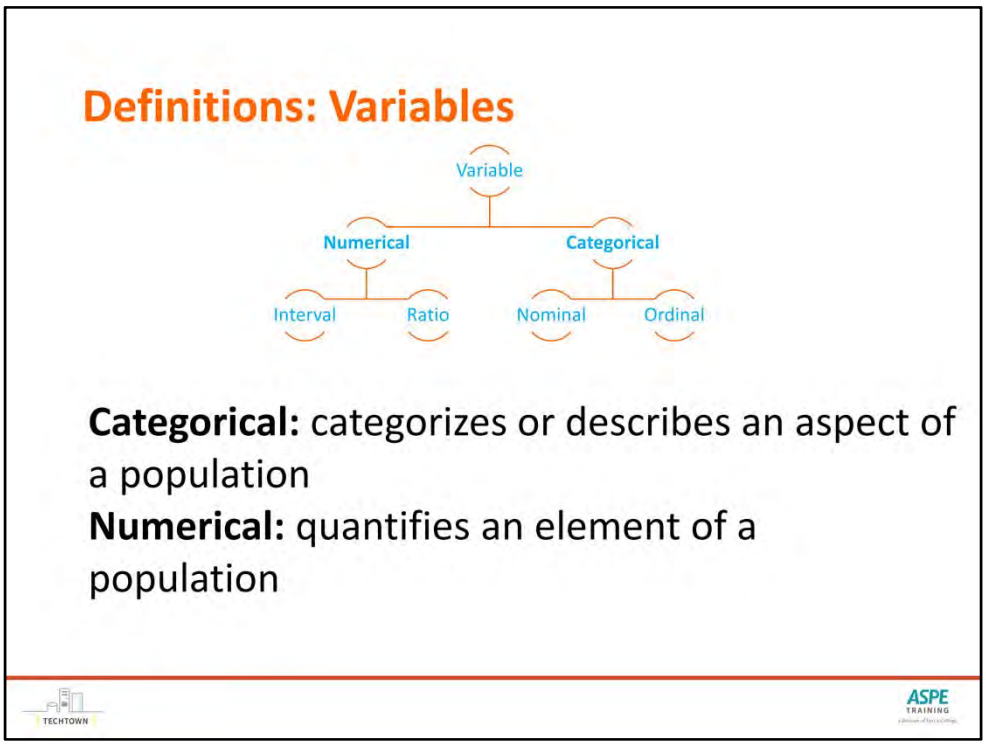
Section 3

NOTES:



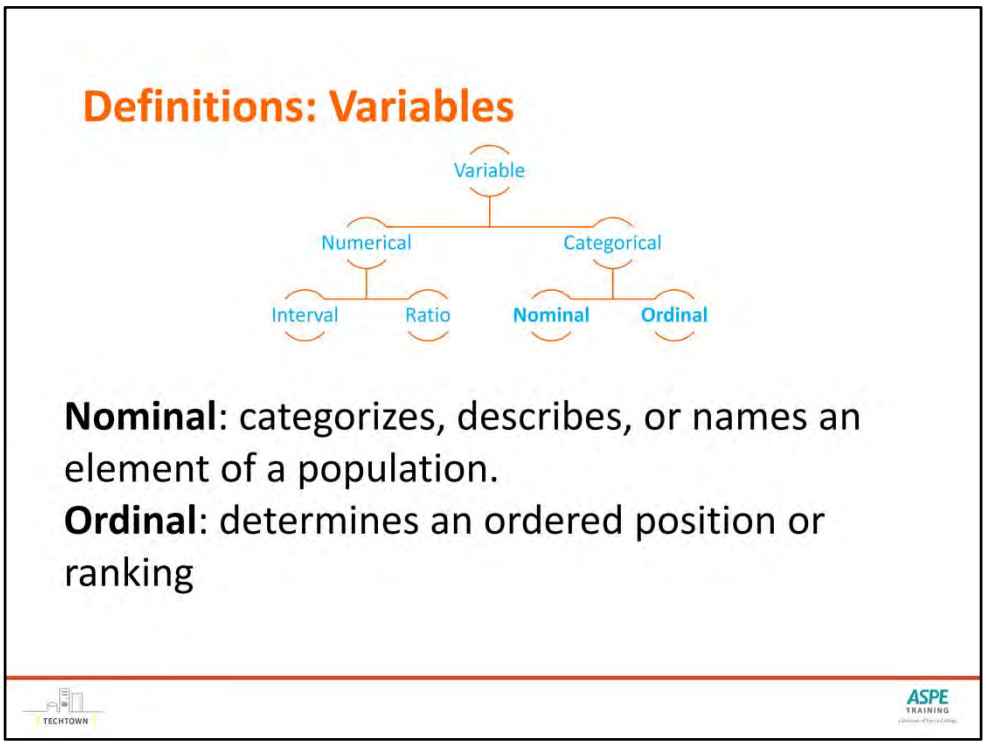
Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

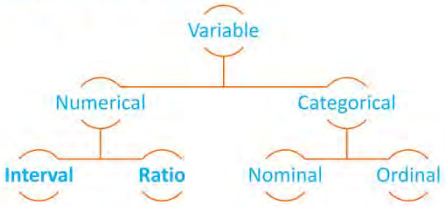
NOTES:



Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:

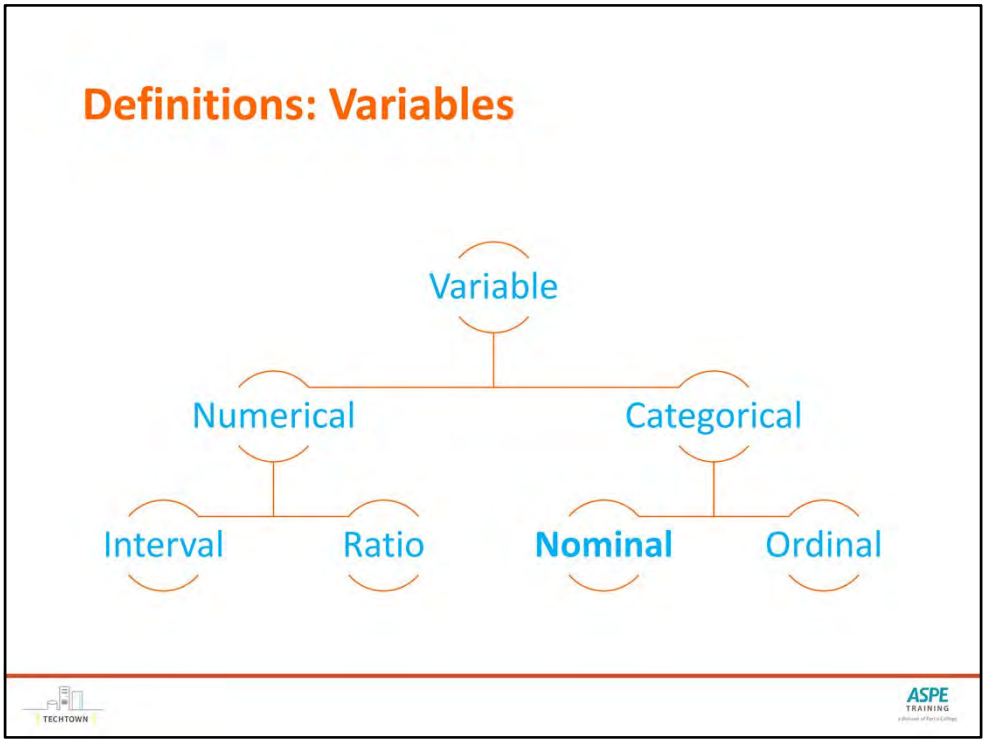
Definitions: Variables



Interval: the distance between two values is meaningful (Celsius on the thermometer)
Ratio: possesses a unique, meaningful and non-arbitrary zero value (Kelvin on the thermometer)

Variables are classified according to the diagram on this slide. The classifications dictate how we can (and should) display data for these variables.

NOTES:



Let's go through these in some details with examples – starting with Nominal…

NOTES:

Nominal Variables

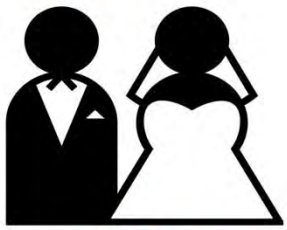
- Nominal= “Name ONLY”
- Nominal variables contain mere codes assigned to objects as labels, they are not measurements.
- Not a measure of quantity. Measures identity and difference. People either belong to a group or they do not.

Eye color: blue, brown, green, etc.; Biological sex (male or female)

NOTES:

Nominal Variables: More Examples

- **Marital Status: Married, Single, Divorced, Widowed**



- **Country of Origin:**
  - **1 = United States**
  - **2 = Canada**
  - **3= Mexico**
  - **4 = Other**

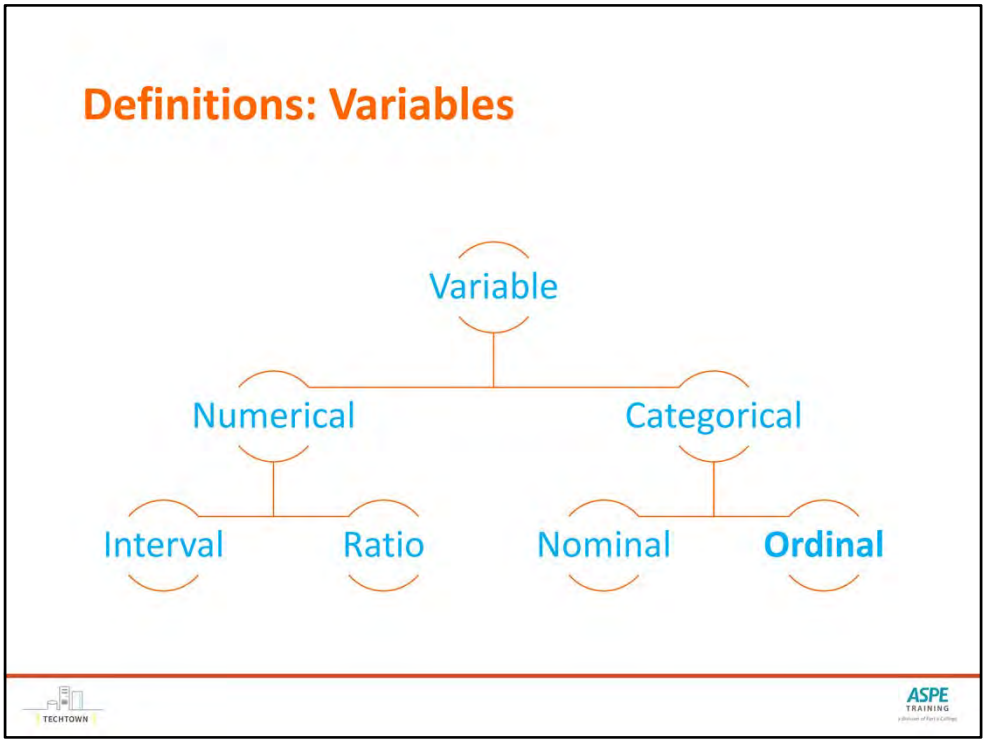
(Notice: Numbers are only labels!!!)

All of these are categories. Yes, the second one of these has numbers, but they are really just short hand for the labels. They are not really ordered numbers.
Notice that there is nothing here that is ranked in order of importance or quality or anything else. They are just labels.

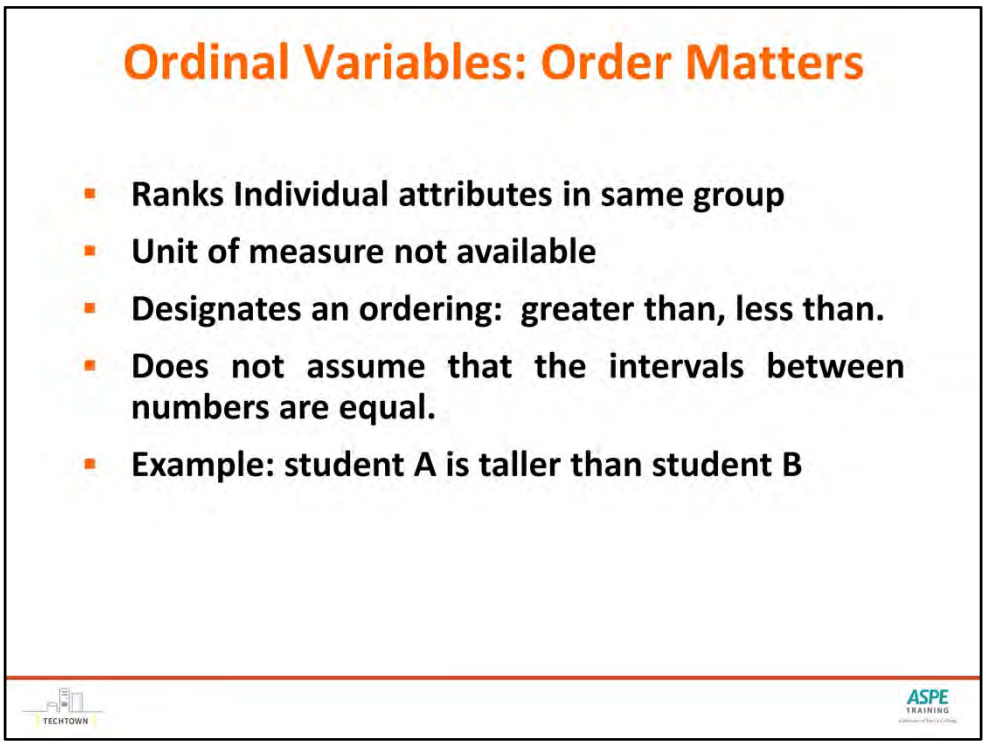
Nominal Variable: What Statistics/Operations Can I Compute?

Statistics/Operation	OK to Compute?
Frequency Distribution and mode	Yes
Median And Percentiles.	No
Addition Or Subtraction.	No
Mean, Standard Deviation, Standard Error of The Mean.	No
Ratio, Or Coefficient Of Variation.	No

This slide describes what calculations we can perform on nominal variables. Notice that it is not much. We can only describe the frequency distribution. If you do not already know what that is, don't worry, we will cover it later. For now, just notice that nominal variables are limited in what they can be used for.



NOTES:



As discussed, order matters here. Consider data that is collected from Jan 2016 through May 2017 every month. When displaying in a chart, January 2016 comes before January 2017, so any outcomes in 1/2016 should be displayed before 1/2017. We should not necessarily group all January data together regardless of year. Notice that January is still just a label. The labels do have a time ordering to them which ranks the outcomes in the labels.

Ordinal Variable: Examples

- **Rank your food preference where**

  1 = favorite food and 4 = least favorite:

  ____ sushi

  ____ chocolate

  ____ hamburger

  ____ papaya

- **Final position of horses in a thoroughbred race is an ordinal variable. The horses finish first, second, third, fourth, and so on.**

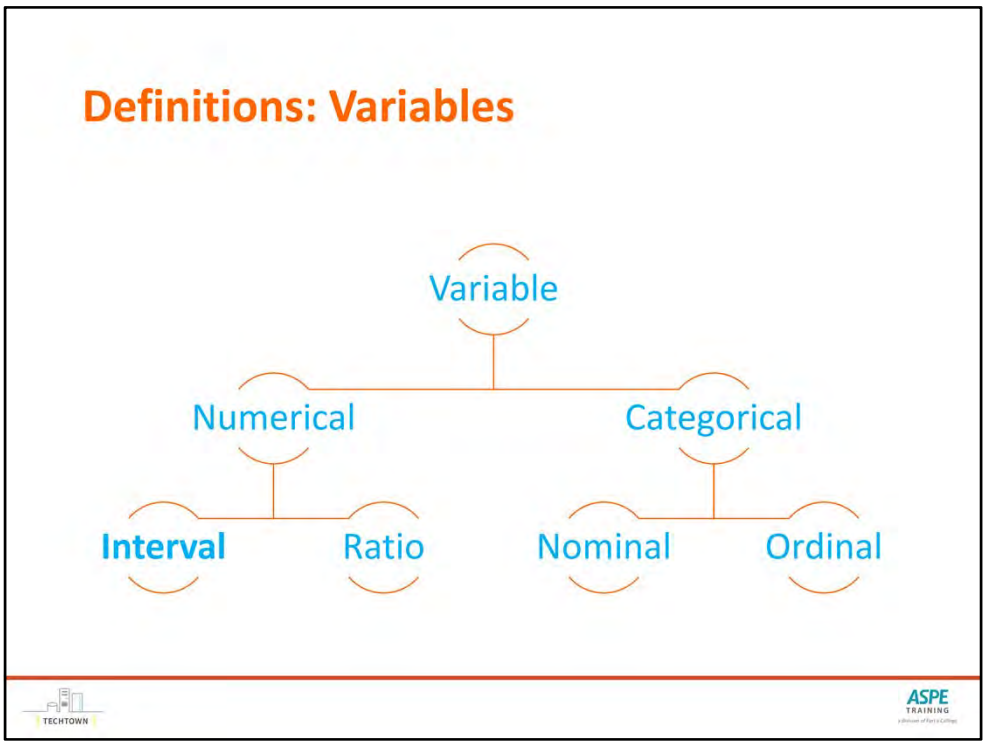
Also notice that the difference between first and second is not necessarily equivalent to the difference between second and third, or between third and fourth. You may love hamburgers slightly more than chocolate, but both way more than sushi and papaya.

Ordinal Variable: What Statistics/Operations Can I Compute?

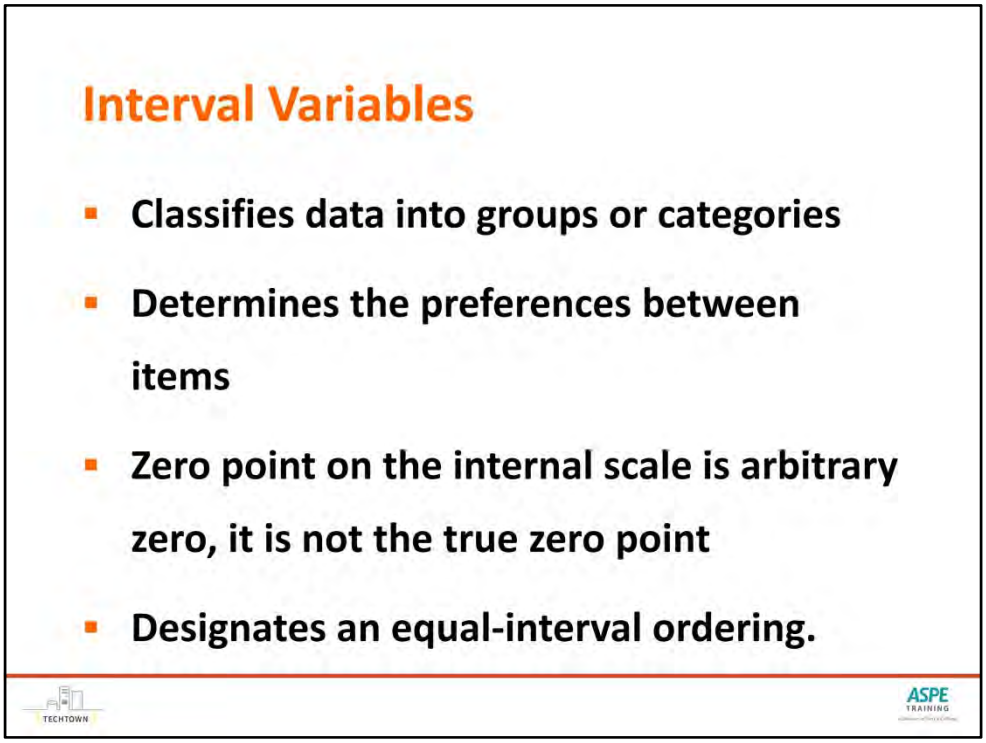
Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	No
Mean, Standard Deviation, Standard Error Of The Mean.	No
Ratio, Or Coefficient Of Variation.	No

What can we compute? Well, since we have not covered the techniques yet, just notice that we have added one more group from our list by rank ordering the set.

NOTES:



NOTES:



Now we are in the variables that work with numbers. You can imagine that this will open a new set of calculations we can perform on our set. These are some general statements about interval variables. Note that with numbers, we can define explicit intervals and make them equal in size (1-10, 11-20, 21-30, etc)

Interval Variables: Examples

Some Temperature Scales

- **Fahrenheit**
- **Celsius**

TECHTOWN

ASPE TRAINING

It is meaningful to say that 25 degrees Celsius is 3 degrees hotter than 22 degrees Celsius, and that 17 degrees Celsius is the same amount hotter (3 degrees) than 14 degrees Celsius. Notice, however, that 0 degrees Celsius does not have a natural meaning. That is, 0 degrees Celsius does not mean the absence of heat!

Interval Variables: Examples

Likert scale:

How do you feel about "Intro to Data Analysis"?

1 = I'm totally dreading this class!

2 = I'd rather not take this class.

3 = I feel neutral about this class.

4 = I'm interested in this class.

5 = I'm SO excited to take this class!

TECHTOWN

ASPE TRAINING

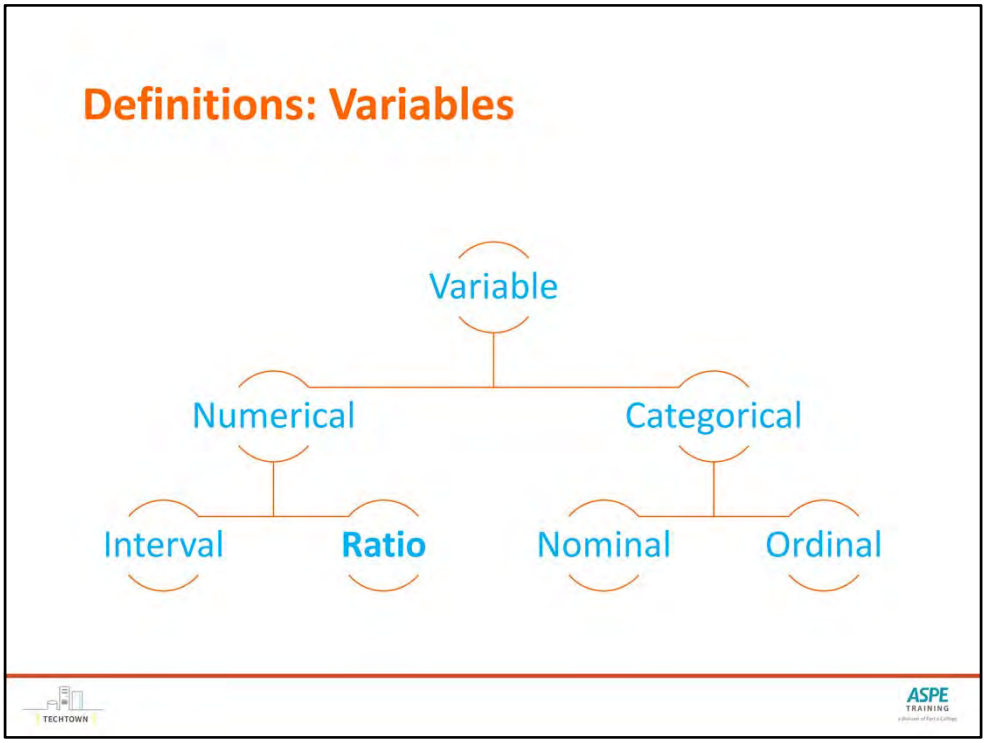
A Likert scale is common in surveys. This is an example of a survey question with a numerical values mapped to statements. They are often balanced around the middle point. Here you see 3 is the midpoint and there are two options higher and two options lower.

Interval Variables: Which Statistic/ Operation Can I Compute?

Statistic/Operation	Compute?
Frequency Distribution.	Yes
Median And Percentiles.	Yes
Add Or Subtract.	Yes
Mean, Standard Deviation, Correlation, Regression, Analysis Of Variance	Yes
Ratio, Or Coefficient Of Variation.	No

Now we have greatly expanded what we can do for analysis. Numbers give us a lot of power.

NOTES:



Lastly…

NOTES:

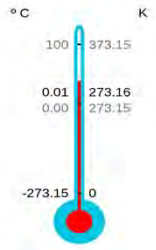
Ratio Variables

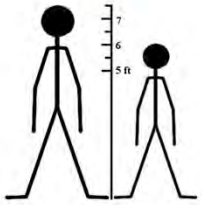
- This is the highest level of measurement and has the properties of an interval scale; coupled with fixed origin or zero point.
- It clearly defines the magnitude or value of difference between two individual items or intervals in same group.

NOTES:

Ratio Variables

- Temperature in Kelvin (zero is the absence of heat. Can't get colder).
- Measurements of heights of students in this class (zero means complete lack of height).
- True Multiples: Someone 6 ft tall is *twice* as tall as someone 3 feet tall.
- Heart beats per minute has a very natural zero point. Zero means no heart beats.



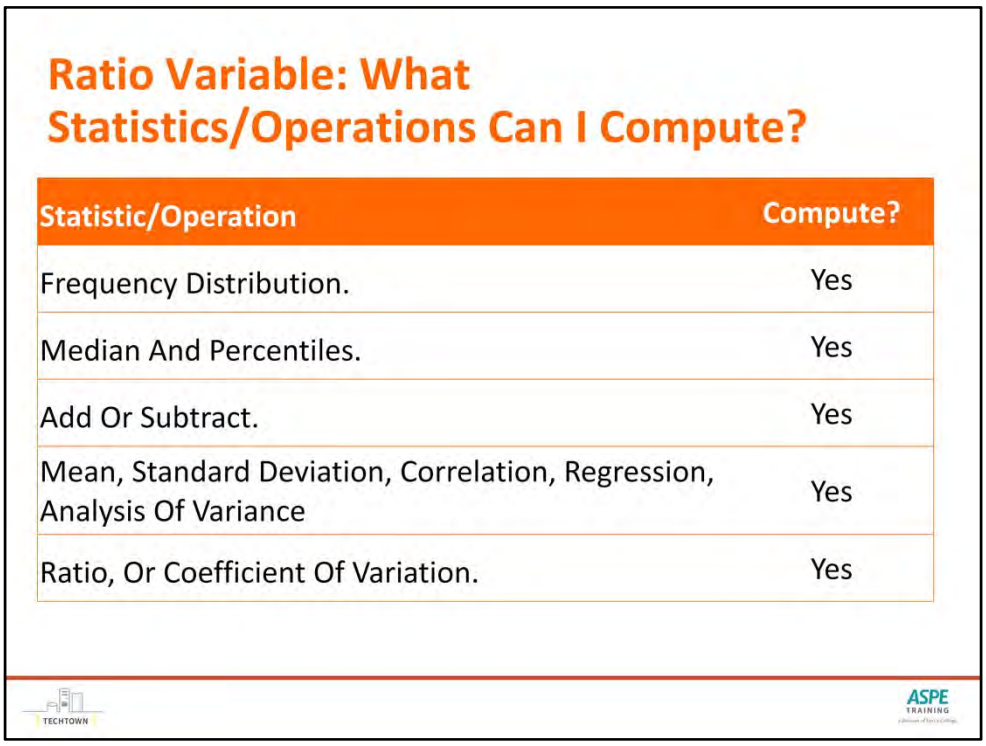


TECHTOWN

ASPE TRAINING

We already saw Kelvin, now we add height (no one is negative inches tall), and heart beats (no one has negative heart beat).

NOTES:



And with ratio variables, we now have the spectrum.

NOTES:

Summary of Variable Types

Level of measurement: (Variables)	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

TECHTOWN

ASPE TRAINING

The summary of everything covered to this section about the utility of variable types.

Central Tendency of Data

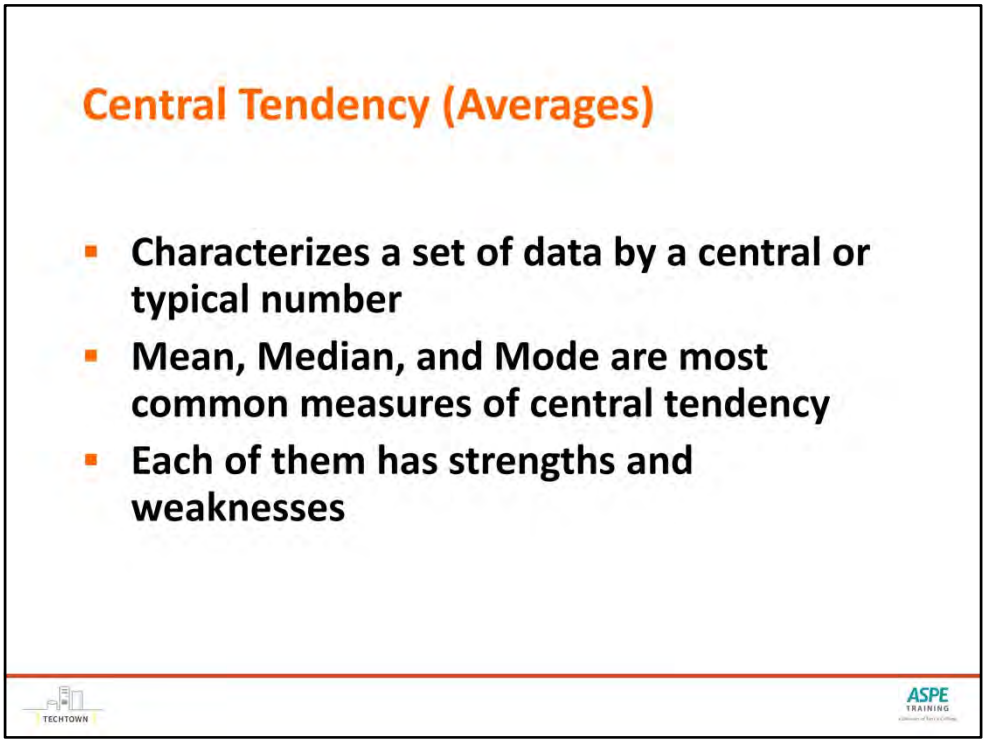
We look at the basic description of the typical values of data sets

Section 4

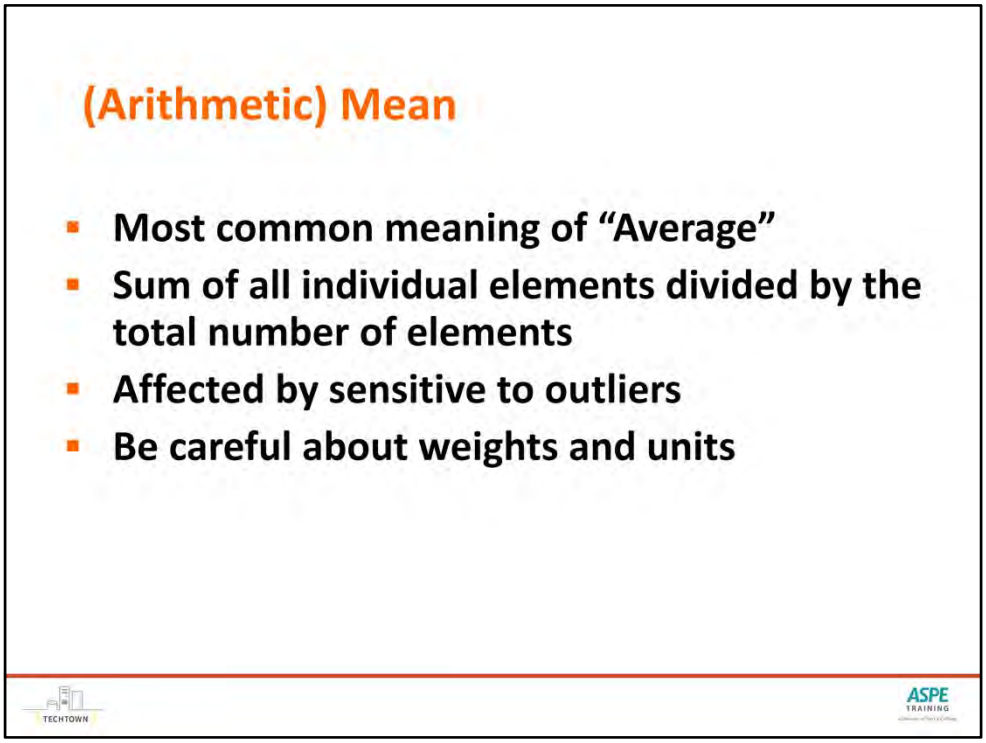
TECHTOWN

ASPE TRAINING

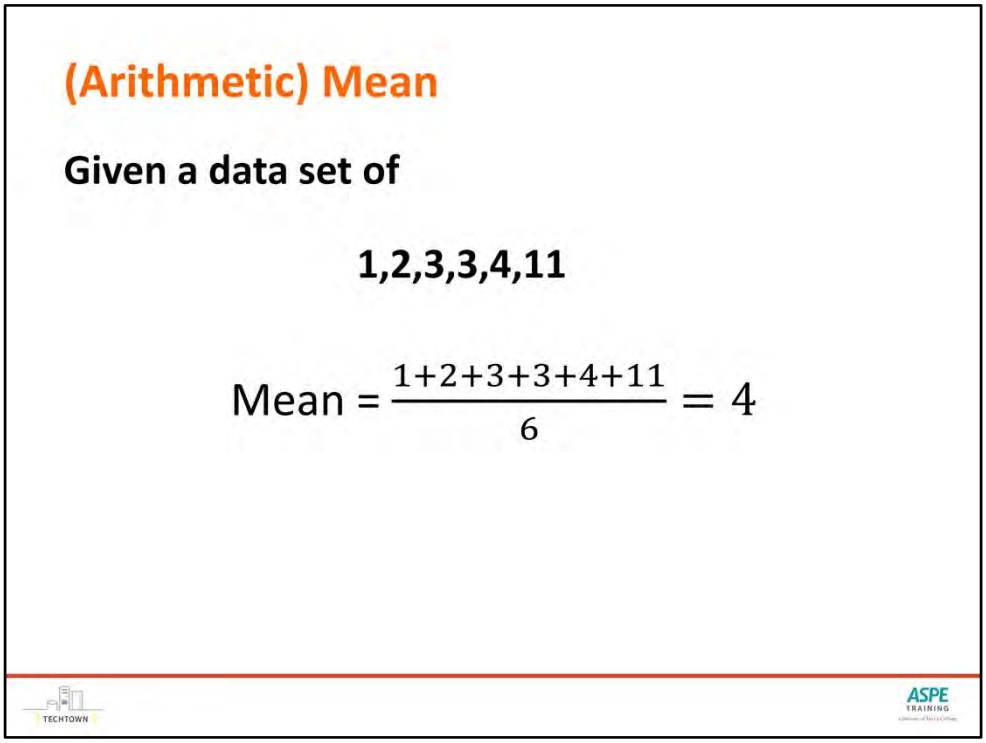
NOTES:



When confronted with most sets of data, we want to know how it tends to cluster together. Where are the most points? Now we are speaking of this independent what the data actually measures, so it is an abstract concept, which is what makes it powerful. By powerful, we mean that applying the same math to vastly different sets allows you to draw conclusions about those sets. Remember, for the information to be useful, you have to know something about the field the sets are from. Knowing the average of the number of particles emitted from a radioactive isotope and the average of the population of North American cities doesn't really lead you to any conclusions without knowing the problem.

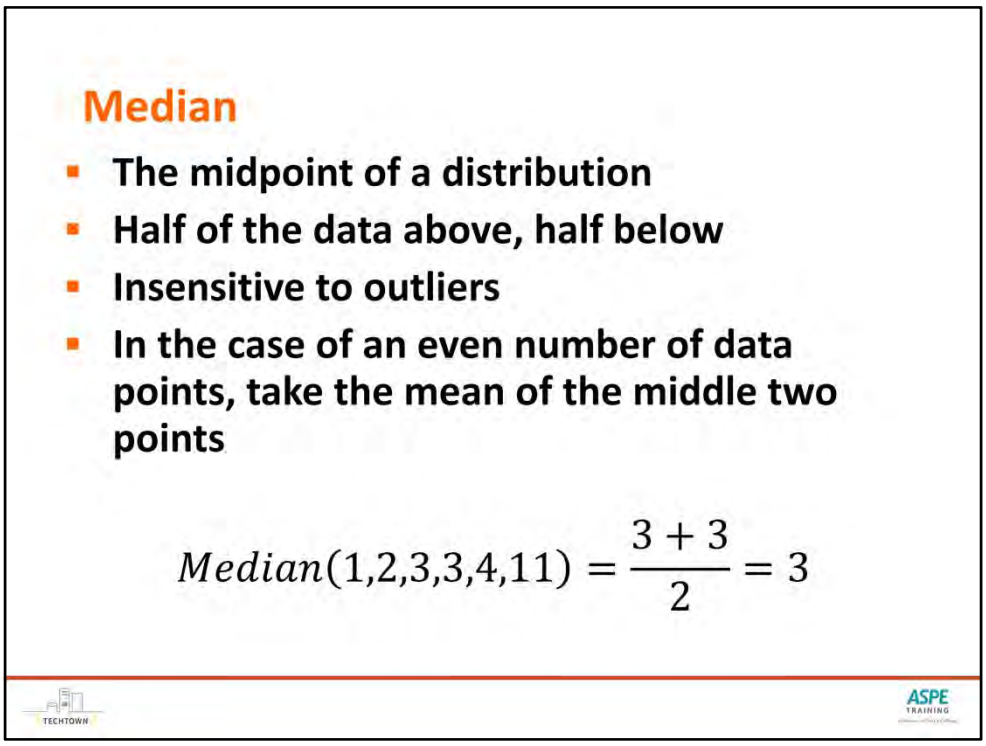


This is what most people mean when they say average. The mathematical definition is next.

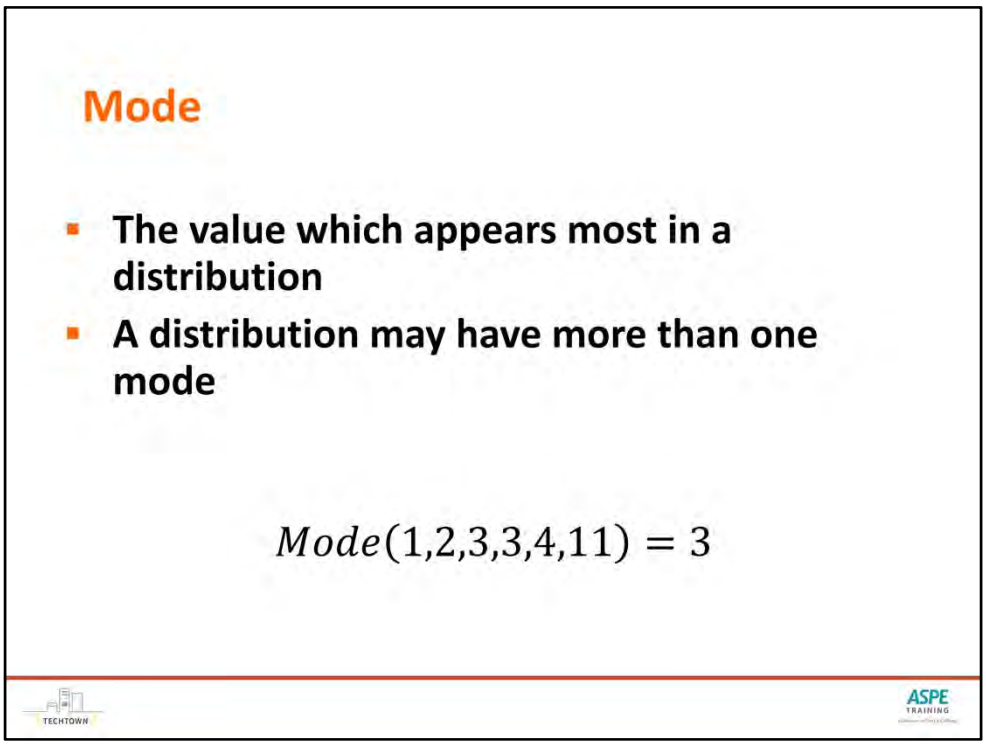


We just add the numbers in the set together and divide by the number of elements in the set. The result is a single number that we set clusters around. Note that the numerical result may not actually actually be in the set (the average of 2 and 3 is 2.5).

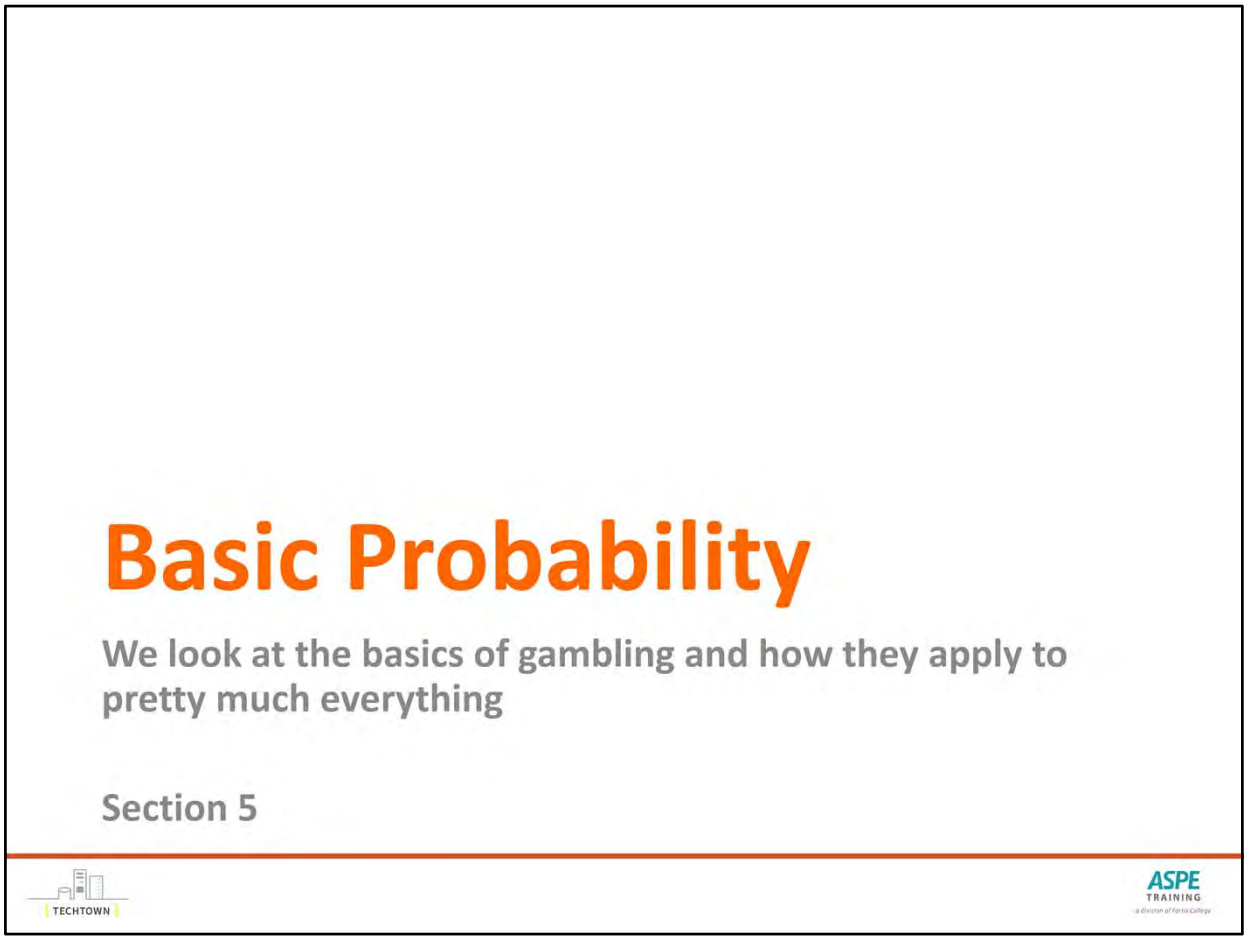
Notice what happens if I add another number, say 215, to the set. What happens to the average?



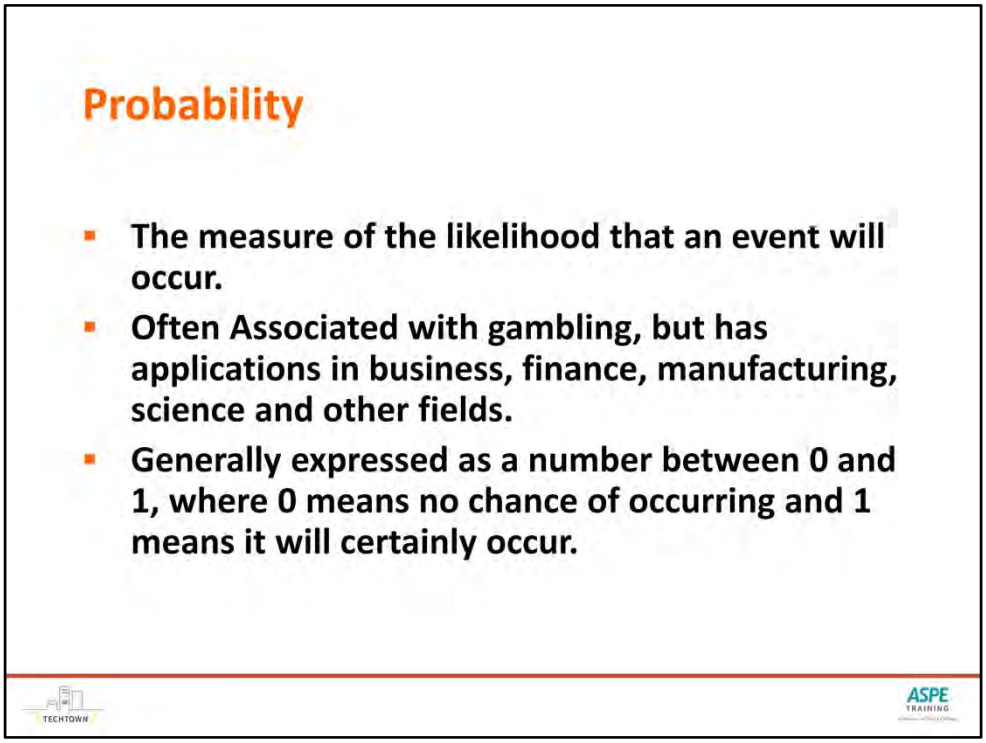
This is another way of clustering around a single number in a set, but here, the number is in the set. This average is insensitive to outliers because adding another number that is really large or really small compared to the existing set doesn't significantly change the actual count of the data by much, so the "half above and half below" value is not affected much.



This clustering description also generates a number that is in the set. It is also insensitive to outliers because outliers don't usually occur multiple times.



NOTES:

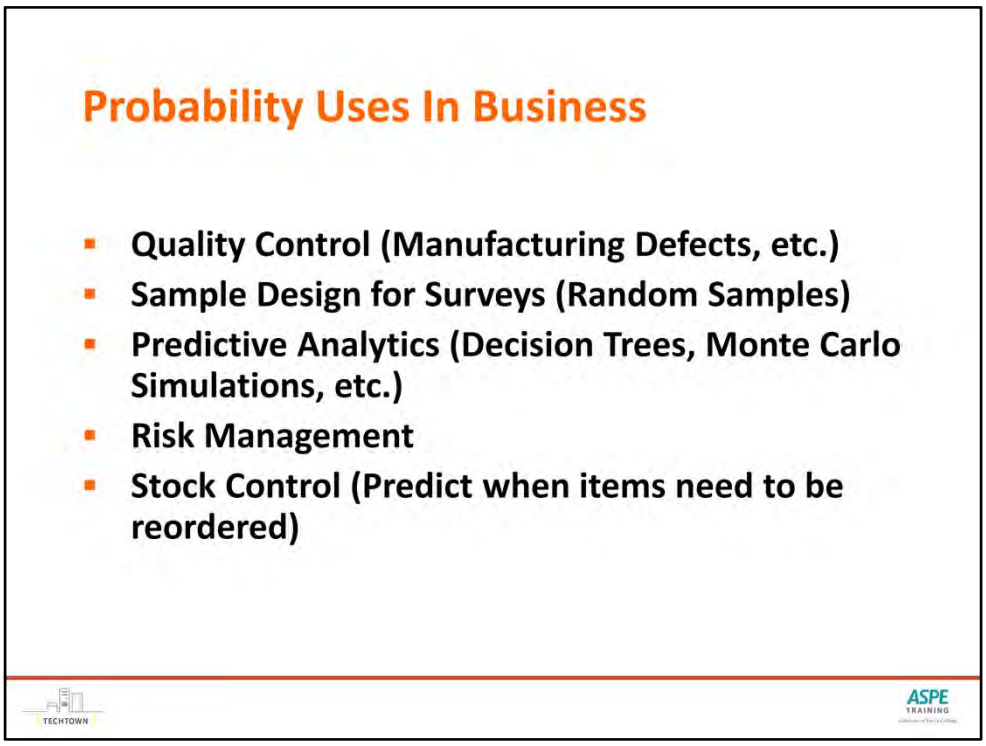


Probability is just a game of counting. It describes such things as

Heads vs Tails
Number of times you will draw an ace out of a deck
Likelihood you will roll a 7 on two dice
How likely are you to get a red 32 in roulette

Are you sensing a theme here? There is a reason. Gambling was a primary motivator behind the study of probability.

You often here things like a 10% chance of rain, but it is important to remember that probability is really measure between 0 and 1. We just tend to multiply by 100 when we present results to non-analysts.



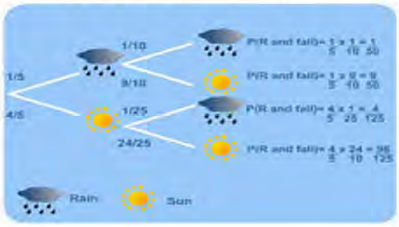
All of these are commonplace applications of probability (and statistics).

Probability: Several Ways We Can Calculate It

- **Simple probability calculated from frequencies**
- **Conditional probability from contingency tables**
- **Conditional probability from tree diagrams**

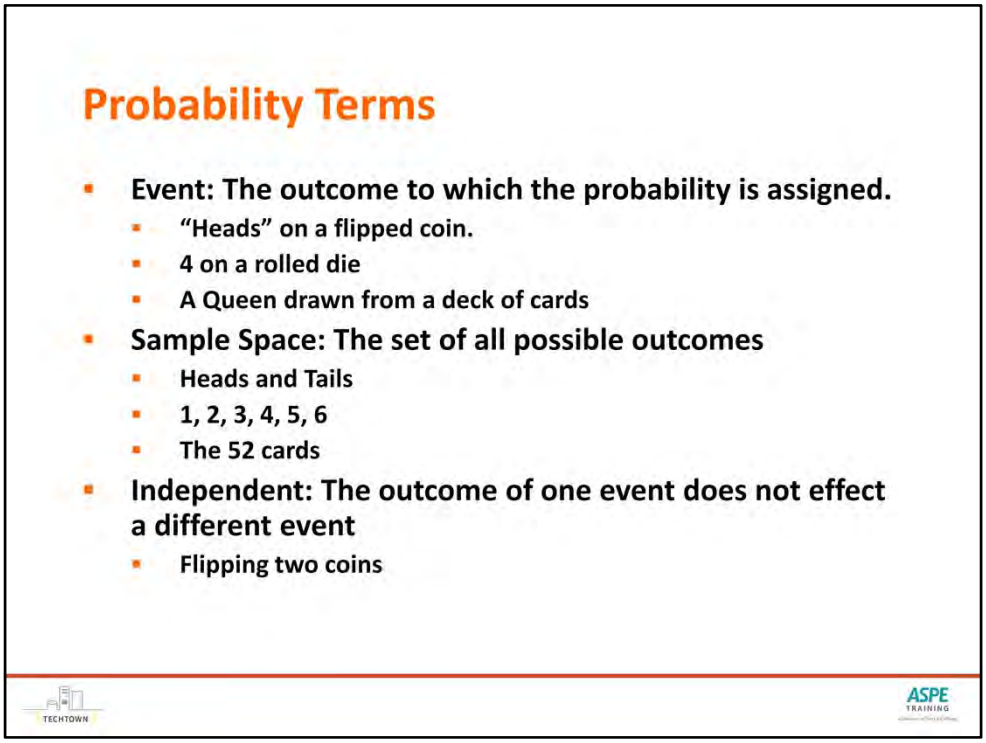
Color	Frequency	Relative Frequency
Purple	7	7/20=35%
Blue	3	3/20=15%
Pink	5	5/20=25%
Orange	5	5/20=25%
Total	20	20/20=100%

Geographic Location	Years of Experience			
	0-5	6-9	10 or more	Managers
Midwest	4	3	2	9
East Coast	7	5	3	15
West Coast	12	10	7	29
Total	23	18	12	53



The charts pictured here show 3 common ways of calculating probability

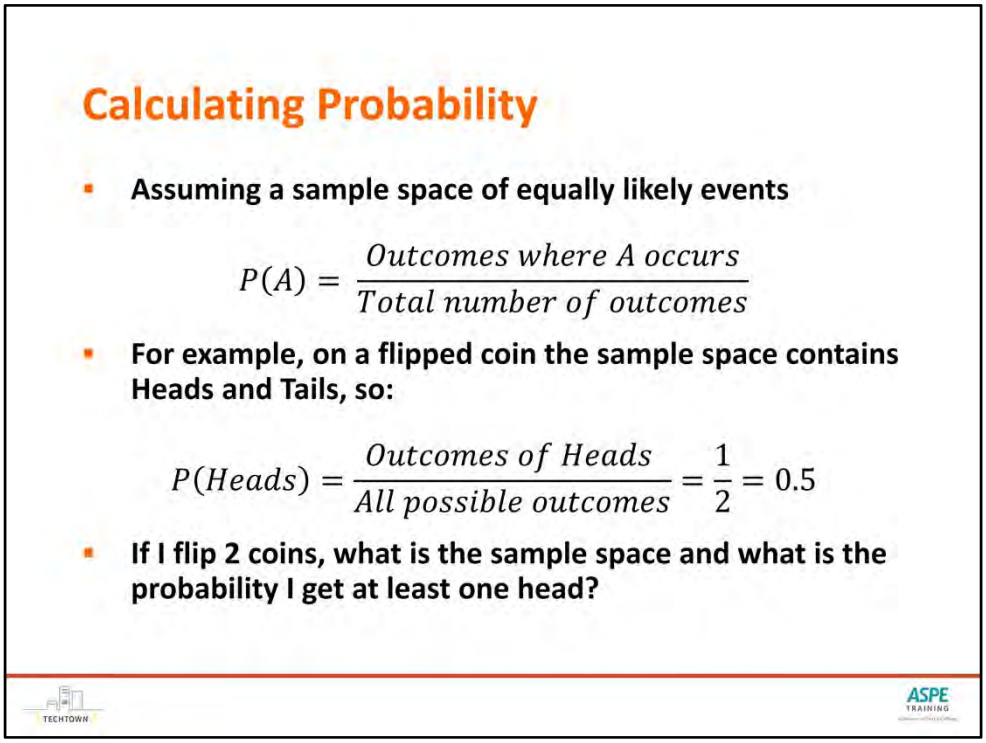
- Calculated from Frequencies – the chart shows how often something of a particular color is observed (i.e. a marble in a bag of marbles) and the total count of marbles
- Contingency Table – the chart tracks two factors in a matrix. The probabilities are read from the totals
- Tree Diagram – tracking the different options through the branches of the tree



The event is the thing that we are performing (flipping a coin) and the outcome is the result of the event when measured (got heads). We assign the probability of getting the outcome of heads when we have the event of flipping a coin (in this case, it is P=0.5).

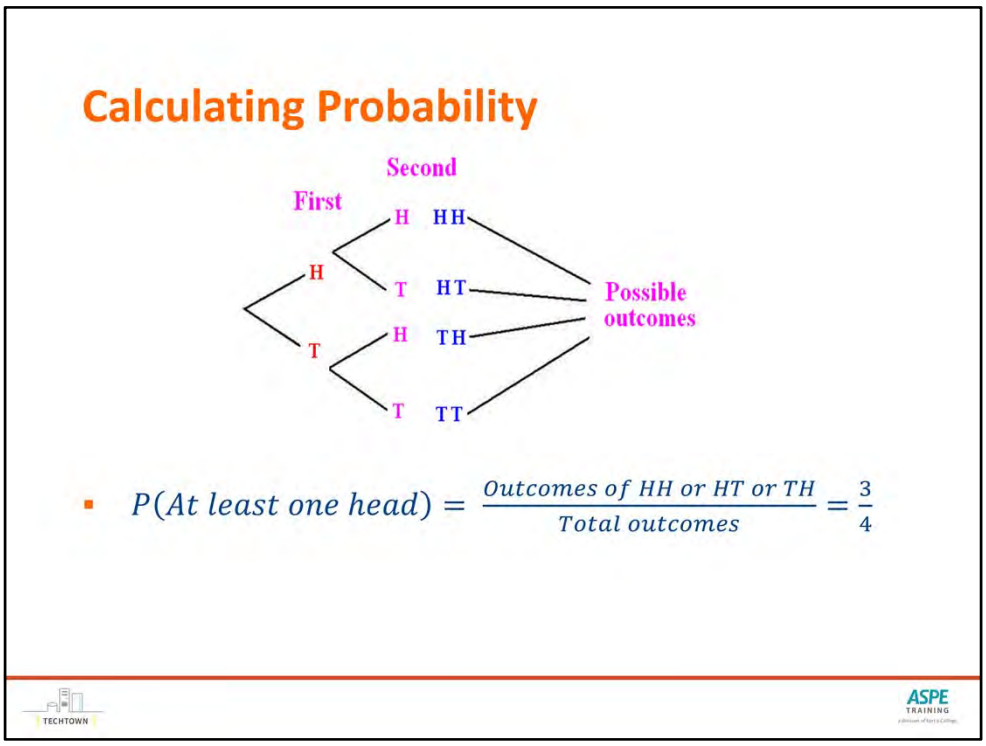
The sample space describes all of the possible outcomes that an event can have.

"Independent" will be an important term for us. The notion that one coin flip does not affect the outcome of another coin flip, or that a person's height does not depend on another randomly selected person's height, is very powerful and leads to insightful mathematics.



The math described here speaks for itself, but it is one of the first mathematical descriptions, so spend some time understanding the notation. P(A) is the probability of an outcome that we label as A. The probability is calculated by dividing the number of ways that A occurs by the total number of possible outcomes.

When flipping a coin, heads can only occur one way and there are two possible outcomes for the coin toss. Therefore, P(H) = 1/2



This one is harder. Here we are flipping 2 coins and we want to know the probability of getting at least one head. Using this tree diagram makes it easy to visualize the space of outcomes.

The way we read this is:

Follow the line to H for the first toss, then we can follow the line to H for the second toss and get a possible outcome of HH
Follow the line to H for the first toss, then follow the line to T for the second toss to get a possible outcome of HT
Repeat for T as the first toss…. TH and TT

There are four possible outcomes and 3 of them have an H in them.

Calculating Probability from a Contingency Table

A supermarket did a survey to investigate customers' drink spending in relation to their mode of travel.

	Amount spent on drink			
Mode of travel	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

$$P(Customer\ spends < \$20) = \frac{95 + 88}{250} = 0.732$$

TECHTOWN ASPE TRAINING

The contingency table allows the analyst to read the count of the outcomes directly from wither the row or column. In this example, to determine the probability that a customer spends less $20 for refreshments is found by reading the last row (Total) for all columns where the money spent is less than $20. Adding up the counts in the two columns is the outcome count for this outcome. Then divide by the total count of outcomes in the last column.

Can you repeat this to answer the question "What is the likelihood that a person will spend less than $20 given that they took the bus to the store?"

Conditional Probability

Mode of travel	Amount spent on drink			
	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

Probability that a customer will spend more than $20, GIVEN that they arrived on foot?

$$P = \frac{10}{70} = 0.143$$

A little more practice but this involves the challenge in the last slide. The part about the ”given that they took the bus”. Here, there are 70 people total who came on foot and 10 of those spent at least $20. This is the conditional part… the “given that”. Now try the challenge question from the previous slide.

Conditional Probability

Mode of travel	Amount spent on drink			
	None	$1 and under $20	At least $20	Total
On foot	40	20	10	70
By bus	30	35	15	80
By car	25	33	42	100
Total	95	88	67	250

More Formally (A = (>$20), B=(On foot))

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{10}{70}$$

Expected Value

- **Given the probabilities and their values, what value should I expect**
- **Similar to a weighted average**

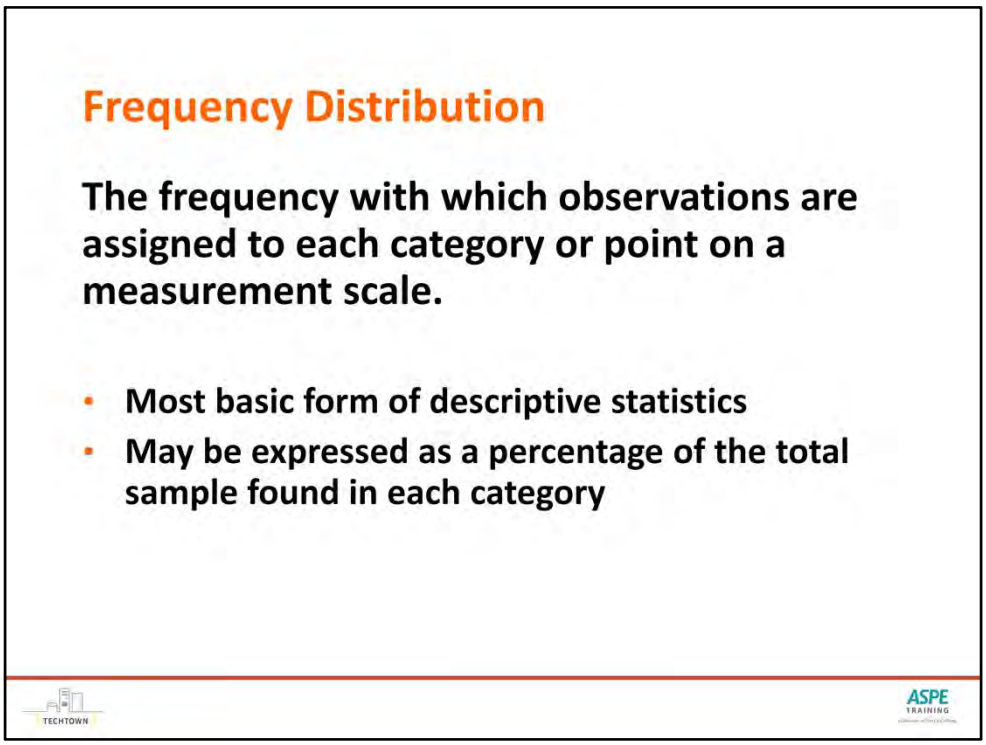
Possibility (n)	Probability P(n)	Return x(n)
1	0.27	105
2	0.30	100
3	0.23	156
4	0.20	130

- $E(x) = \sum_n P(n)x(n) = 0.27 \times 105 + 0.30 \times 100 + 0.23 \times 156 + 0.20 \times 130 = 120.23$

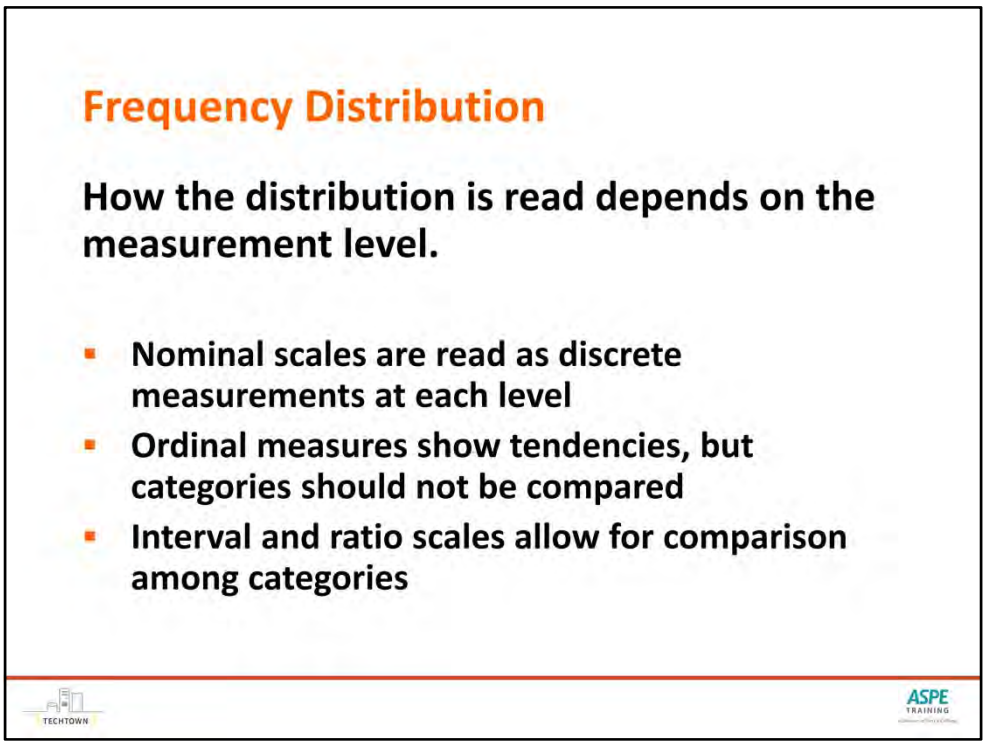
TECHTOWN

ASPE TRAINING

The ”Expected Value” of this set of numbers is defined by the probability with which they occur. Given this set of numbers and the probability that each number occurs in the table, the value one would expect on any observation is defined by the sum of the products of x(n) and the P(n). In this set of observations, the expected value is calculated to be 120.



This the definition of the frequency distribution. It is the workhorse of descriptive statistics. First we define categories, then we observe something. We assign each observation to a category and then count how many are there. That's all there is too it!



Be careful about how you interpret though, the type of variable limits the analysis that can be done.

Frequency Distribution

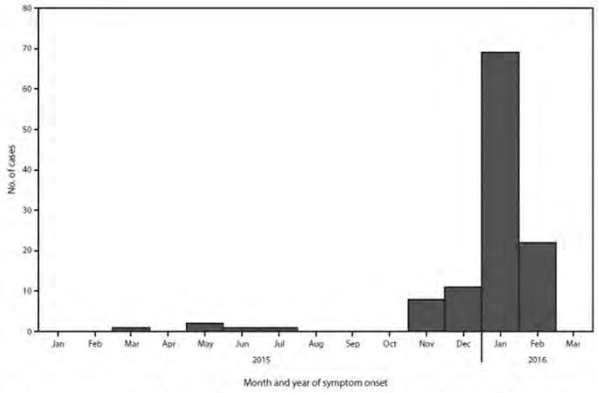


Figure. Month of illness onset for 115 patients with laboratory evidence of Zika virus infection among residents of U.S. states and the District of Columbia — January 1, 2015–February 26, 2016*

The distribution you see here has two important points to note:

- The buckets (bins) are time ordered, or ordinal, values
- The sum of all bins is 115 patients

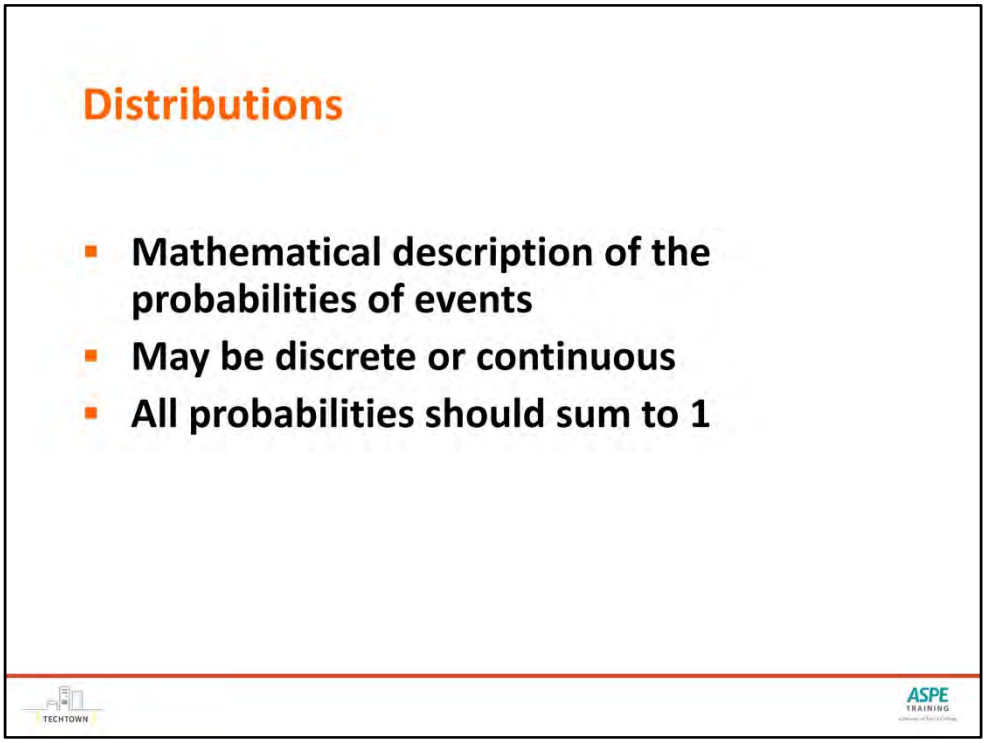
Additionally, see how the x axis is laid out over two years. This is an example of where it is important that we do not combine both Januarys together because the time ordering tells us something important.

Distributions

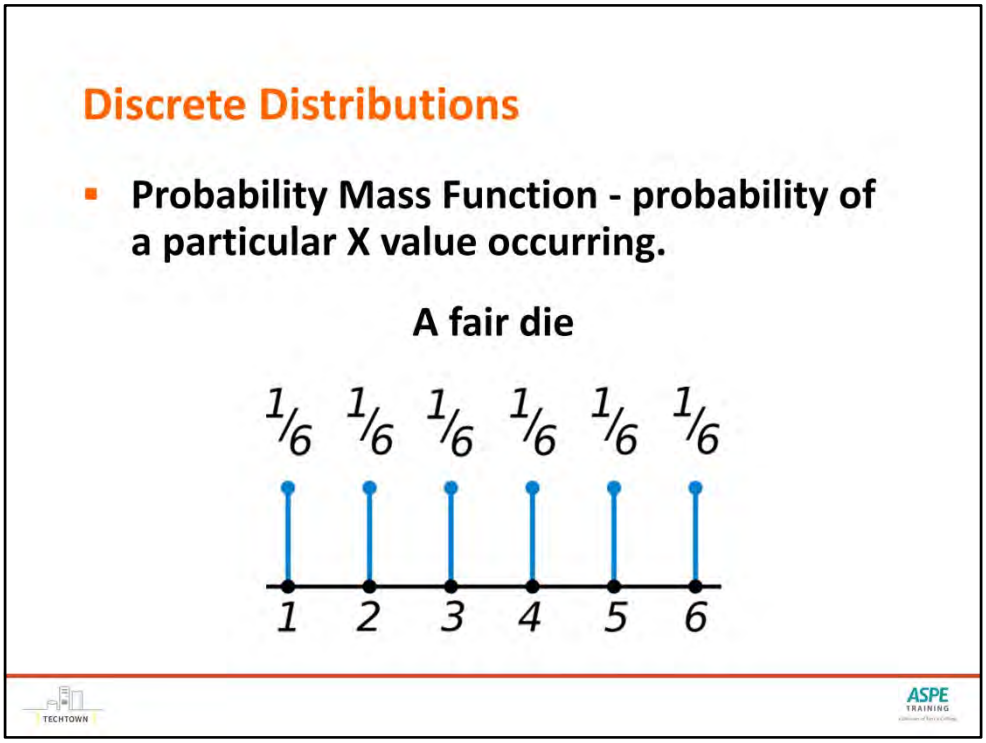
We look at at how probabilities of events lead to distributions, how to describe these distributions, and what they tell us

Section 6

NOTES:

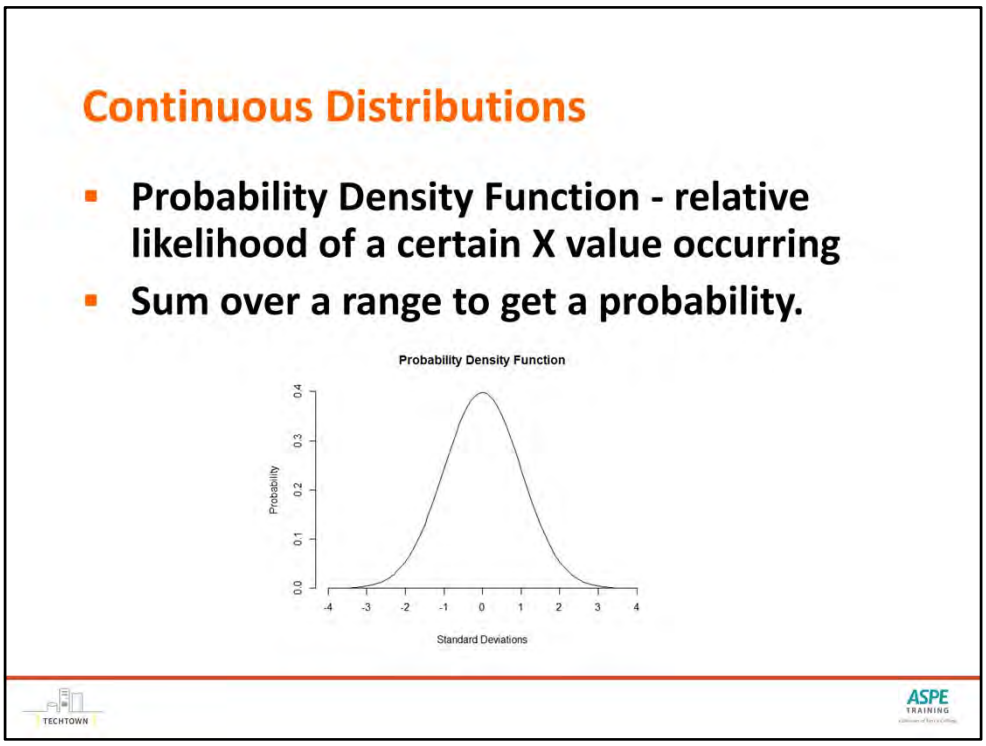


When frequency plots categorize all of the population, they lead us to distributions. Distributions show how all of the population is distributed among the space of outcomes. The can be discrete, where the outcomes are small and countable, or continuous, like the real number line. When looking at a distribution, the frequencies of each outcome added together should add to 1.



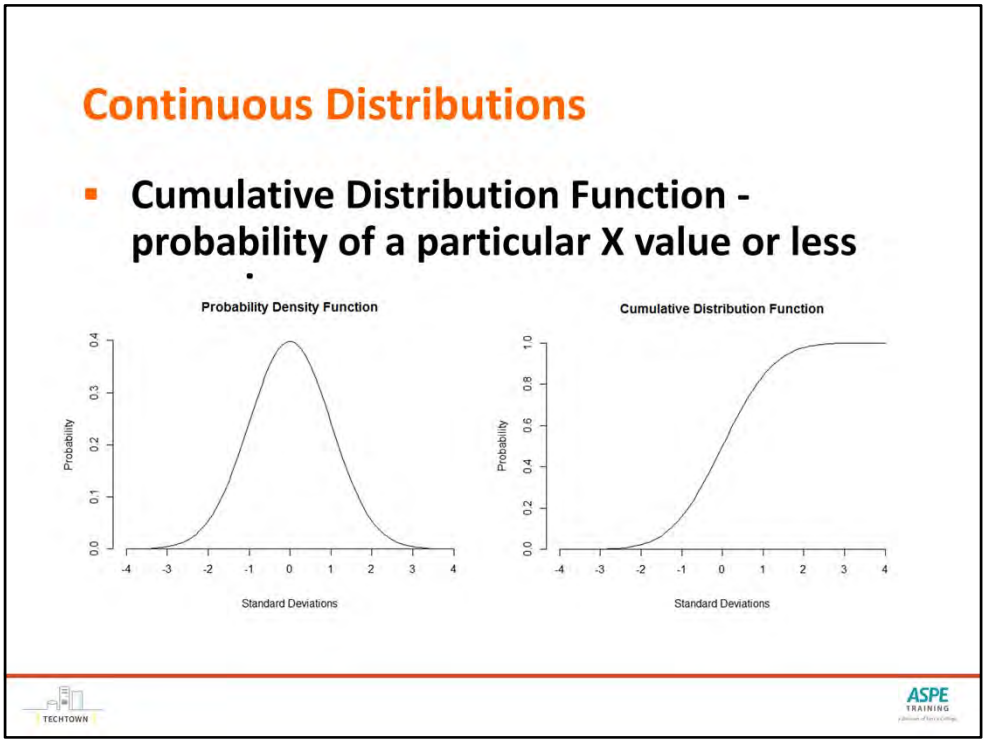
You can determine the probability that a discrete variable will be exactly equal to a particular value.

In this case, we are looking at a single six sided die. The probability of rolling any number on that die is shown in the PMF pictured. When we add all of the probabilities, we get 6*(1/6) to get 1.

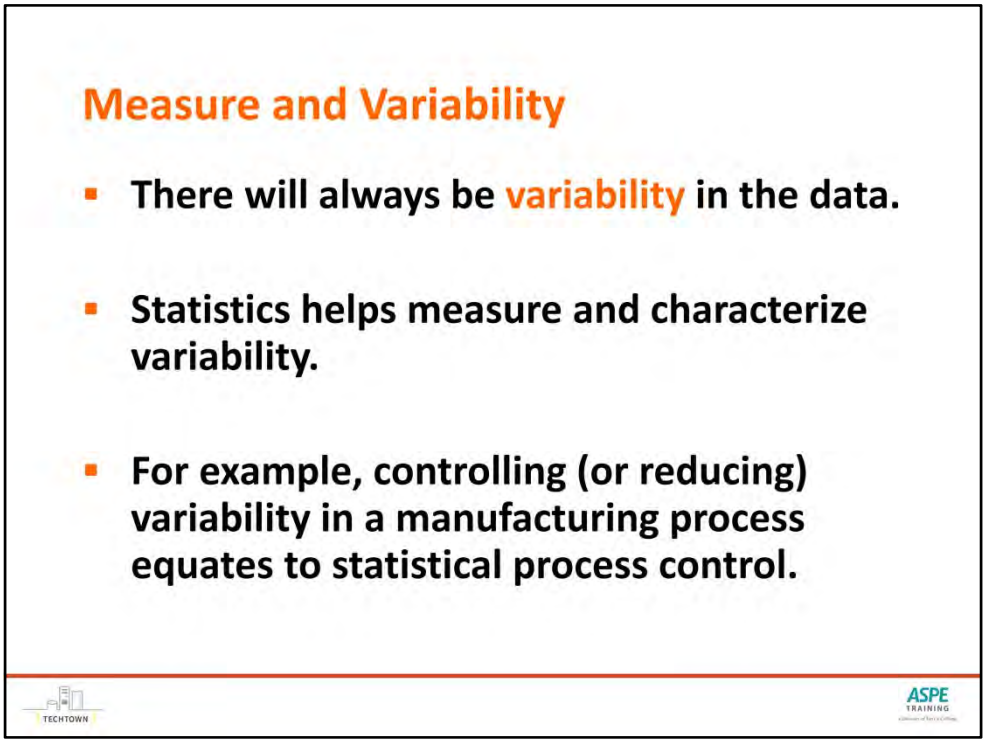


You can determine the probability that a continuous variable will be equal to a particular value by summing (integrating) over a range. In this case, we are looking at a what's called a normal distribution.

We will not be integrating in this class, we will let the tools do it for us, but it is important to understand what it means. In this distribution, we must integrate from -4 to 4 to obtain 1. Notice that there are an infinite number of points between any two numbers on the scale.



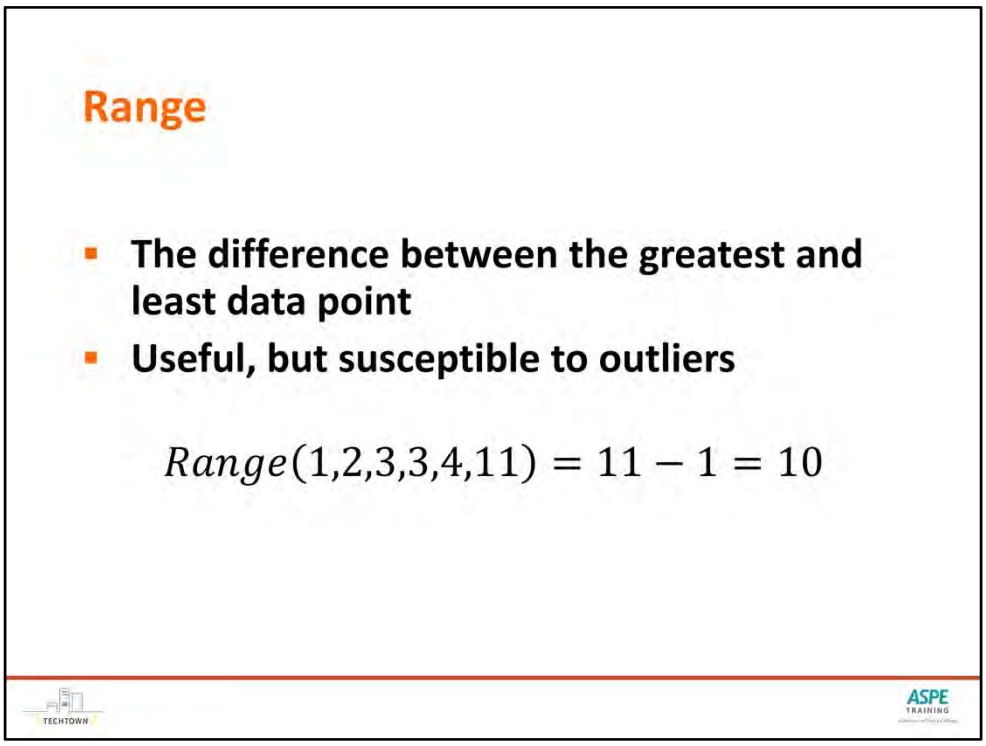
The continuous distribution on the left is the probability density function that we saw in the last slide. The function on the right is what happens when we integrate the one on the left at each stage. It is cumulative, or it keeps adding, as we move from -4 to 4. When you use Excel functions to sample distribution functions, the cumulative function is often used.



We are building up the basic tools for analysis. So far, we have learned the basics of probability and how they lead to distributions (of which, we will cover a number of them later). Now we begin to describe those distributions.

Since distributions are the plots of outcomes, they are made from data. Real data always has variability because the world is a messy place. Statistics helps us characterize the variability. By doing this we can determine such things as quality in a manufacturing process, or the likelihood that some even will happen in the future.

What follows are some definitions of concepts that define the variability of data.



The Range includes all outliers. The definition of an outlier will come later, but it is important to remember when characterizing a data set that the range, when reported, includes the entire set.

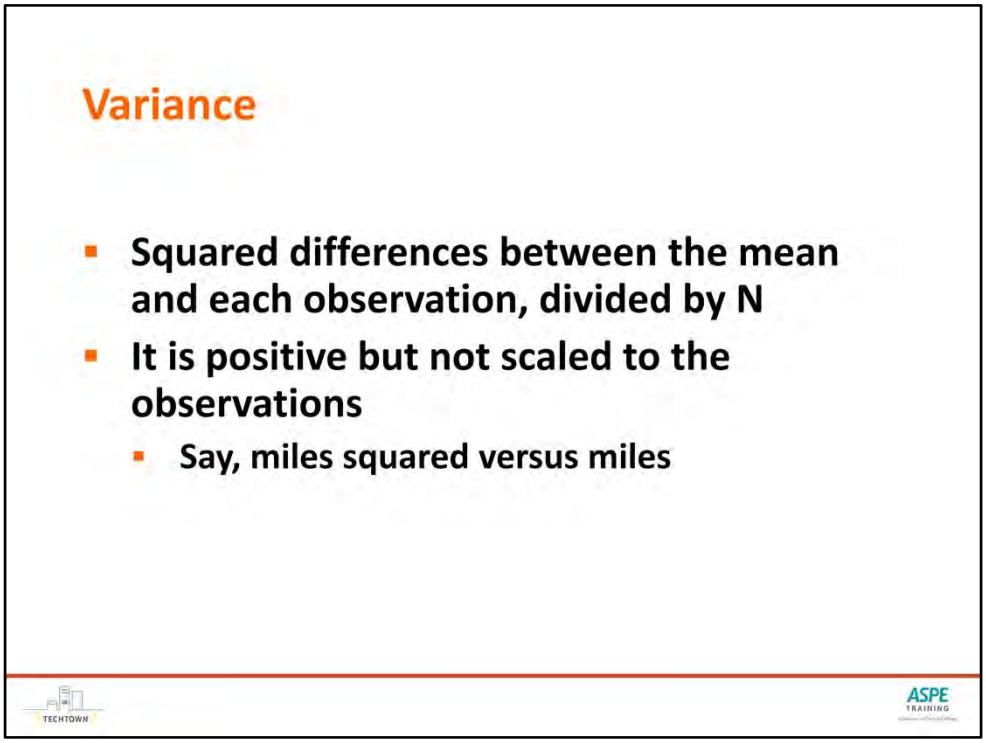
Quartiles

- Divide distribution into 4 sets with equal numbers of points
- 25% of the data is less than the First (or Lower) Quartile
- The Median is the Second Quartile
- 25% of the data is greater than the Third (or Upper) Quartile
- The Interquartile Range is the difference between the Third and First Quartiles

TECHTOWN

ASPE TRAINING

NOTES:



The Variance, defined in this slide, is the Expectation Value of the squared deviation from the mean. Compare the definition above with the definition of the Expectation Value and note how the values of P(n) = 1/N for each observation in the data set. Then x(n) is the squared difference between the mean and the value of each observation.

Variance

$Variance(1,2,3,3,4,11)?$

- **Take the Mean** $(24/6 = 4)$
- **Sum the squared differences with the mean**

$$(1-4)^2 + (2-4)^2 + (3-4)^2 + \cdots = 64$$

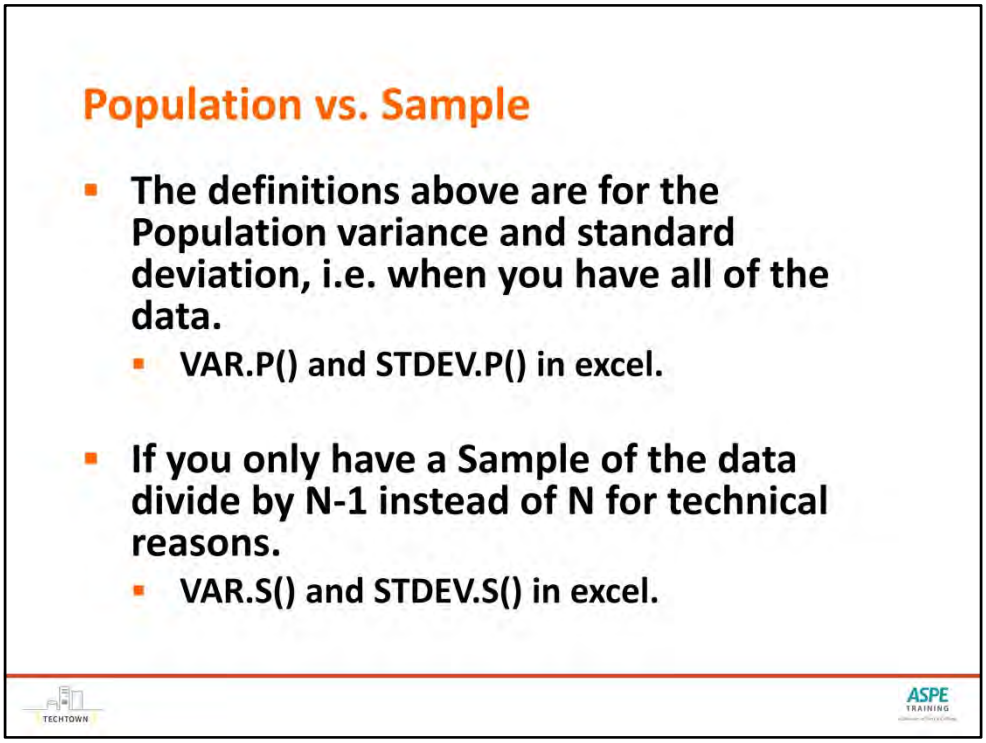
- **Divide by the count (6)**

$$64/6 = 10.667$$

Standard Deviation

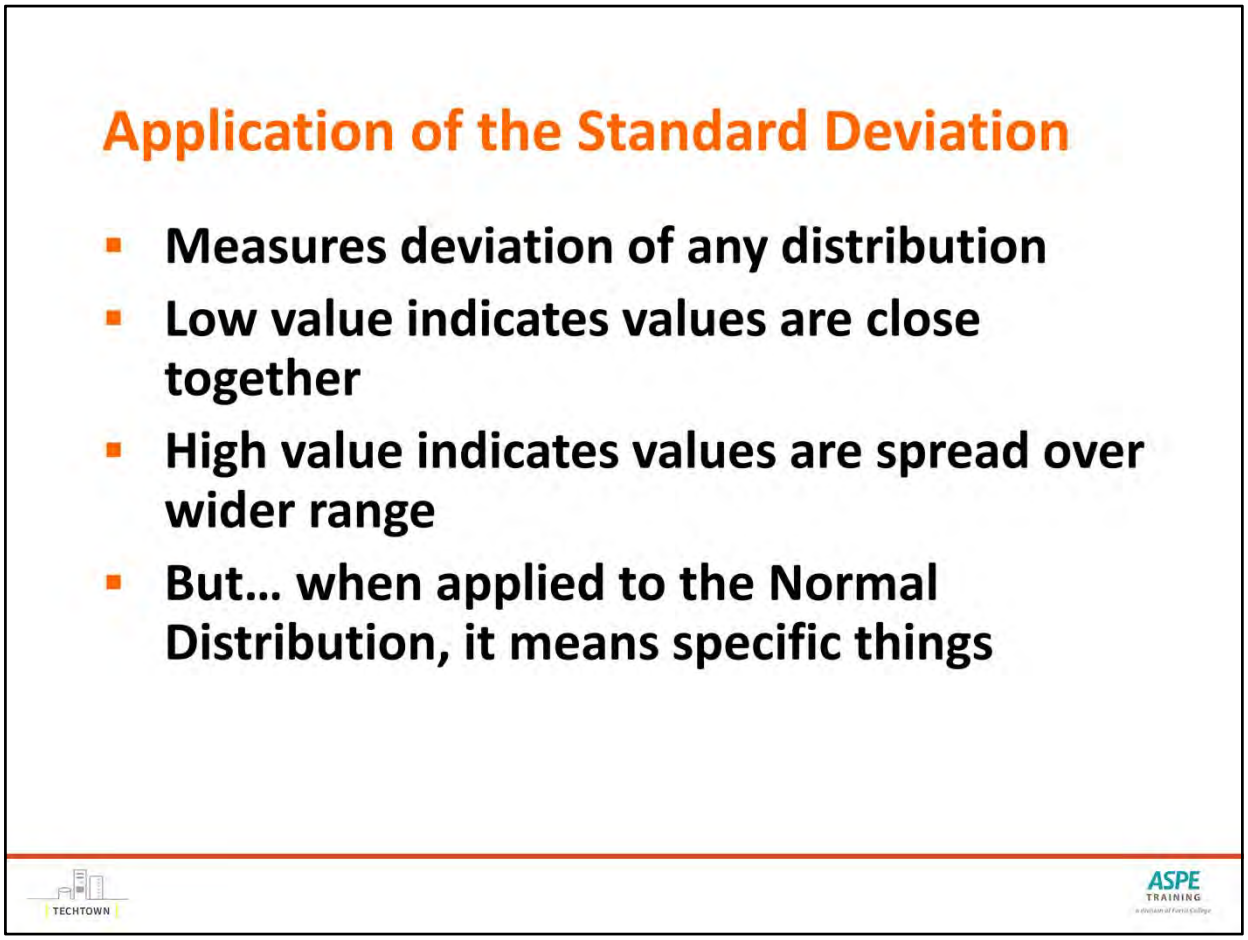
- The square root of the variance
- It is scaled to the observations and is also positive
- Standard Deviation of $(1,2,3,3,4,11)$?

$$\sqrt{Variance(1,2,3,3,4,11)} = \sqrt{10.667} = 3.266$$



Remember earlier in the course when we talked about population and sample? We consider analysis when we have knowledge of the full population for most of this course. Often, the real world considers samples of a population. This leads to a subtle difference functions for the standard deviation and the variance. The difference is in the use of Bessel's Correction.

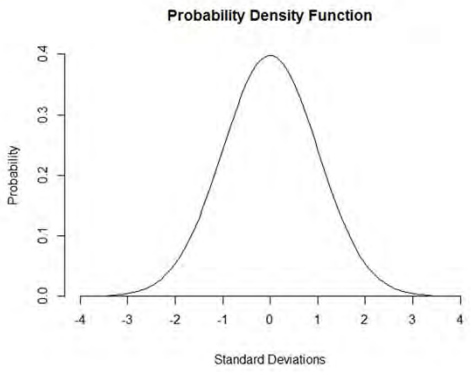
The explanation of the N-1 (Bessel's Correction) is complex, but is related to the fact that you are estimating the mean of the population from the sample mean. This introduces bias into the calculation that is corrected by the N-1.



We talked about the variance and the standard deviation. They apply everywhere, but when applied to a specific distribution, the std dev tells us many things. So let's talk about the normal distribution.

Normal Distribution

- Also called Gaussian
- Appears when events are independent
- Useful because of the Central Limit Theorem



Central Limit Theorem say, basically, that the outcome of a large number of measurements of independent random variables will be distributed along the normal curve seen here. The details are unimportant to the scope of the class, but because of this theorem, the normal distribution show up in a lot of places in nature, technology, business, etc. Most popular appearance is in the Bell Curve of IQ in the last couple of decades.

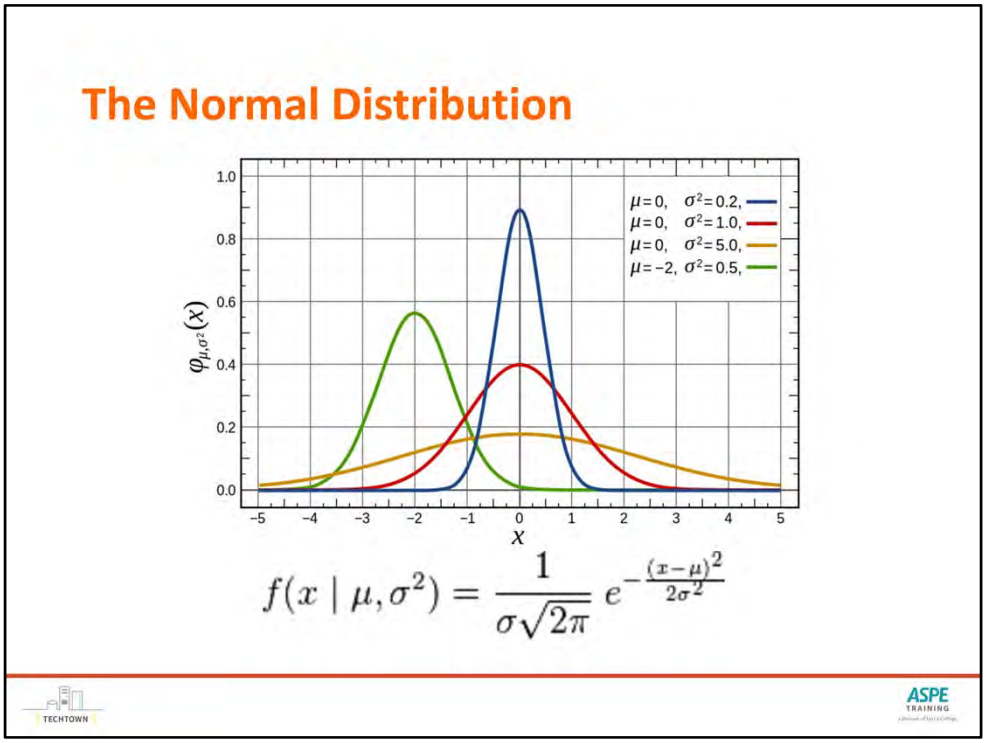


Image is from Wikipedia https://en.wikipedia.org/wiki/Normal_distribution#Definition

This plot shows the normal curve and how it changes with the value of the mean (moves left or right) and the standard deviation (tall and thin, or short and fat).
The mathematical equation is not important for this class, but sometimes, people like to see what is happening behind the scenes. Notice that the funny little u character, called mu, is the arithmetic mean, and the o with loop on it, sigma, is the std dev. When we square the std dev, we get the variance. So this formula has the average, the std dev and the variance all in the same equation.

Standard Deviation and the Normal Distribution

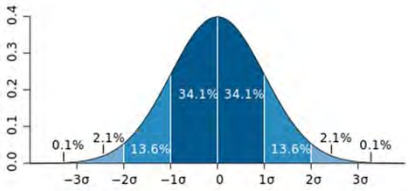
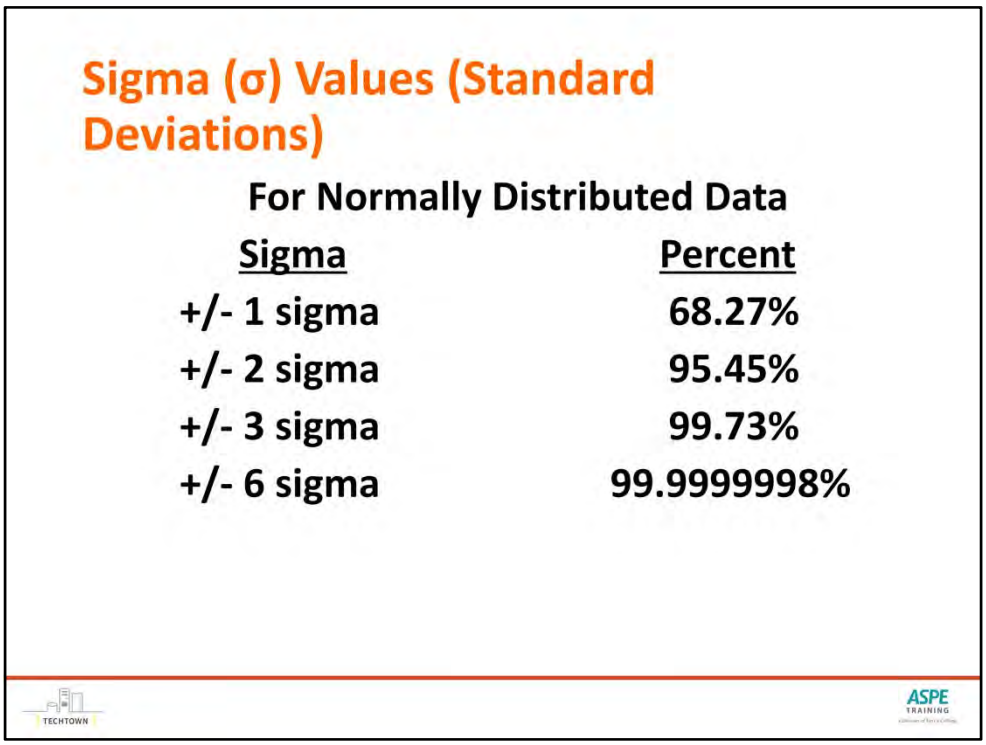
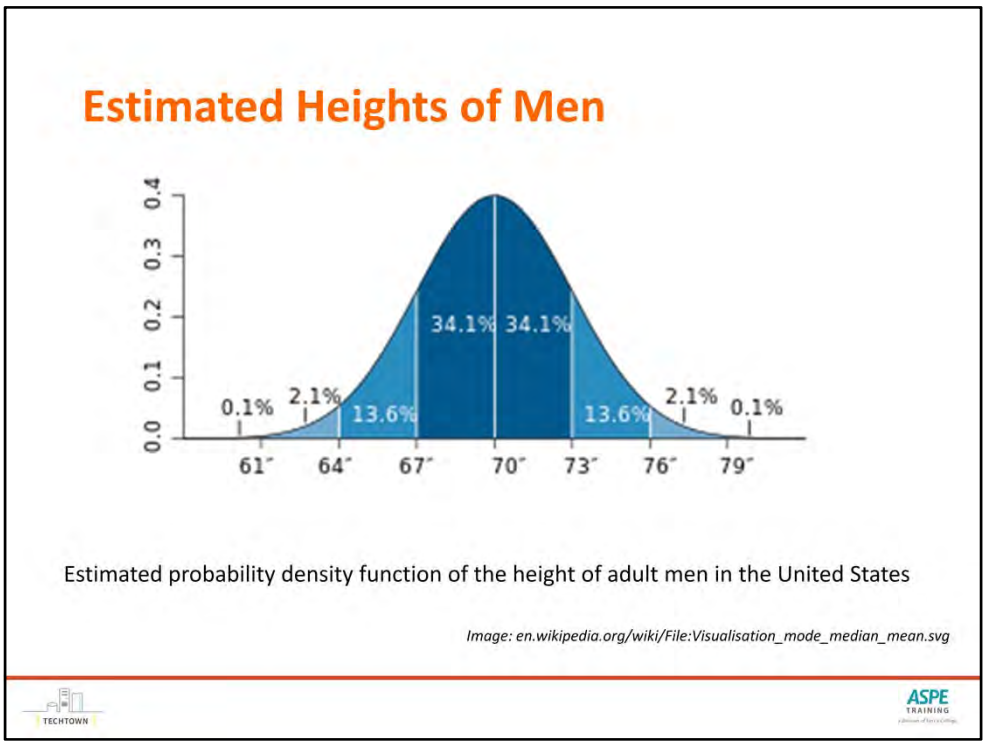


Image: https://en.wikipedia.org/wiki/Standard_deviation

As promised, the application of the standard deviation to the normal distribution shows some interesting information. What this shows is the percentage of the population of a series of observations that lie between the + and – values of the standard deviation. Those values are on the next page.

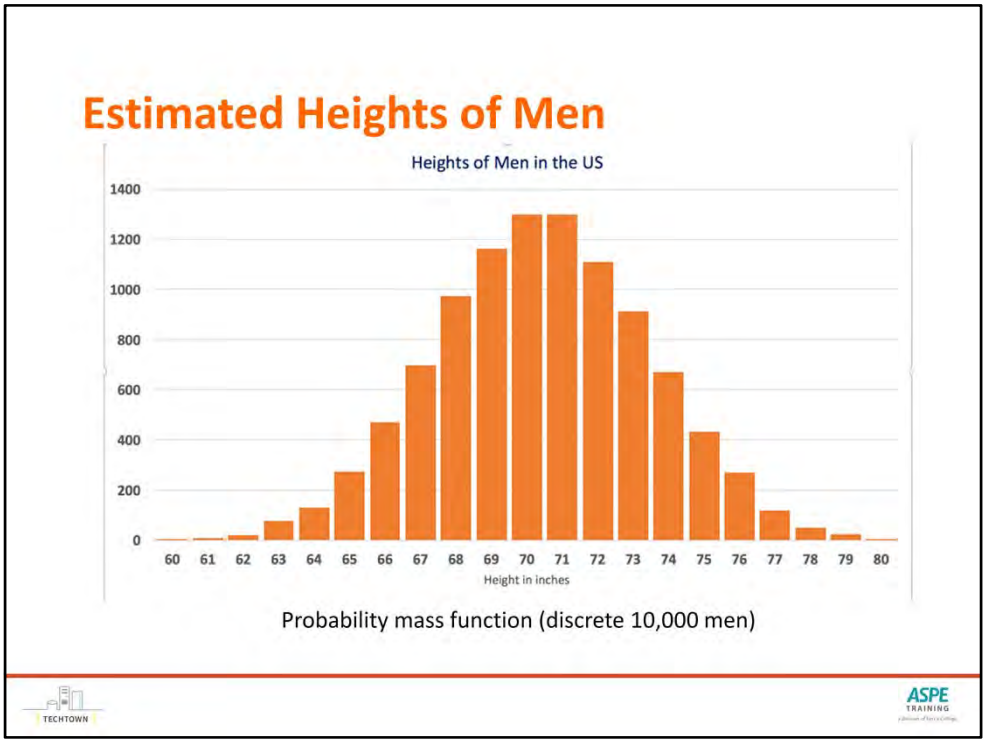


That last number is why the organization that developed the process improvement tools for business calls the suite Six Sigma.

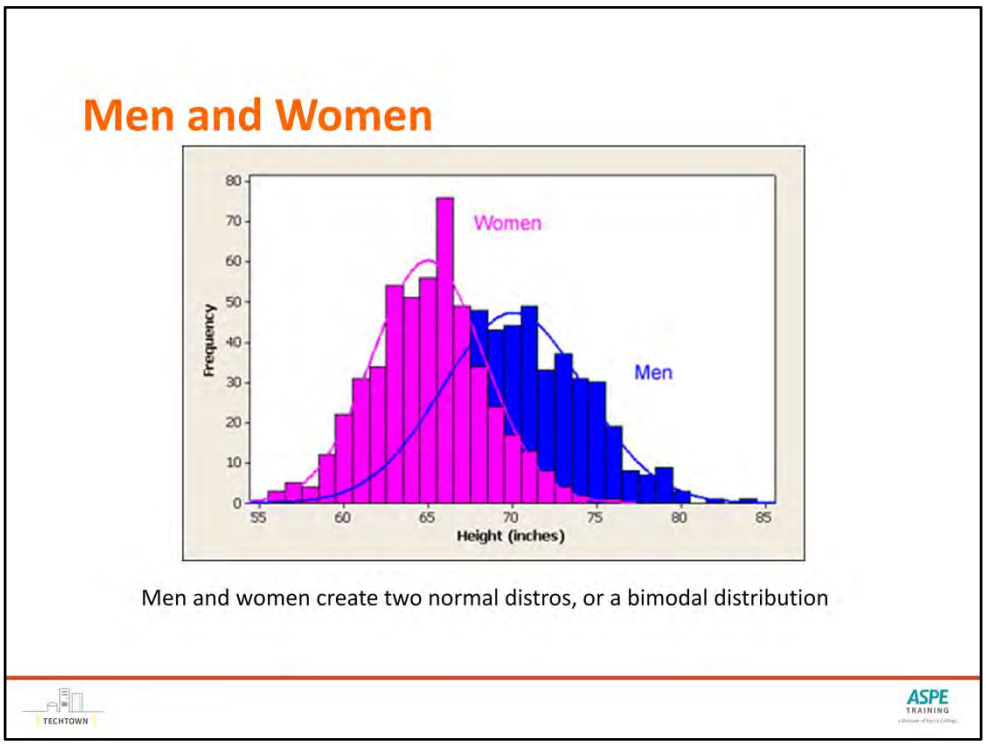


When we measure the heights of randomly selected men in the US, we find the normal distribution.

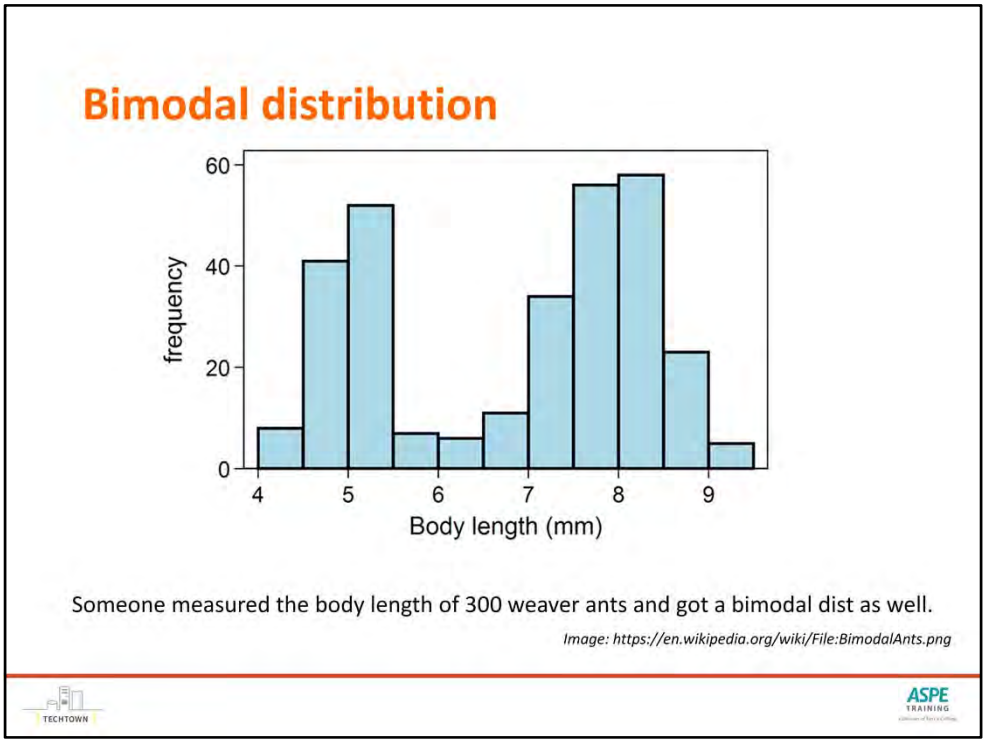
The standard deviation of this data set is 3 inches. Just over 68% of the men measured lie between 70 – 3 inches and 70 + 3 inches



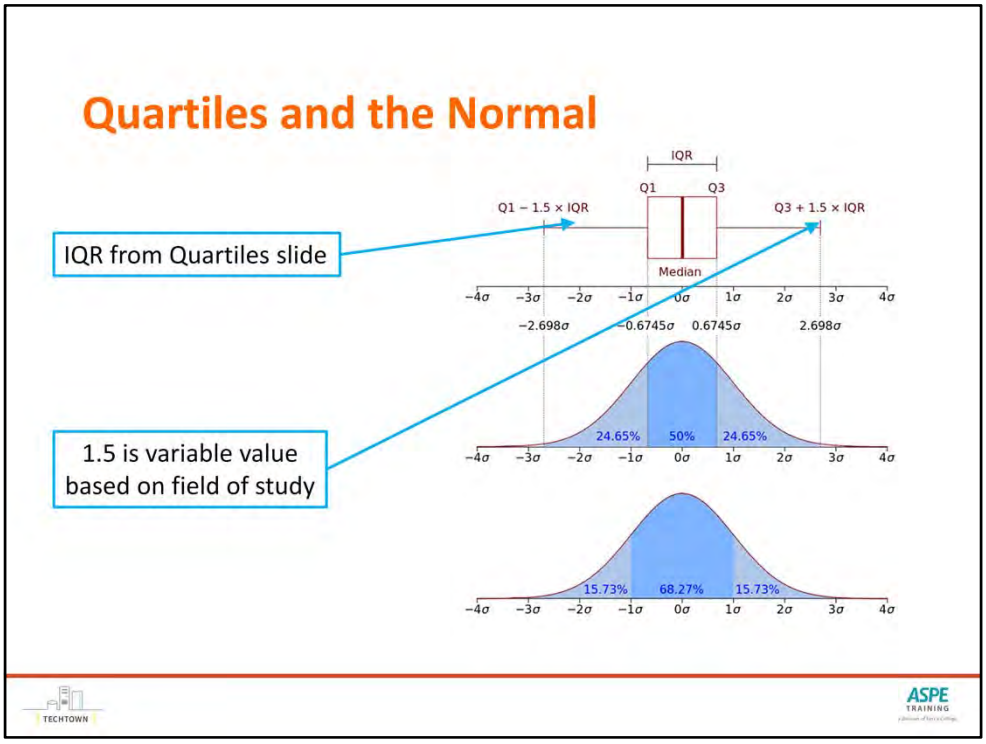
Here is a plot of the measurement of 10,000 men. It is discrete (countable to 10,000) but notice the shape it makes. When you measure another 10,000, you get more accurate distribution and if you take the buckets and subdivide them into decimal intervals, you will get a more continuous looking curve.



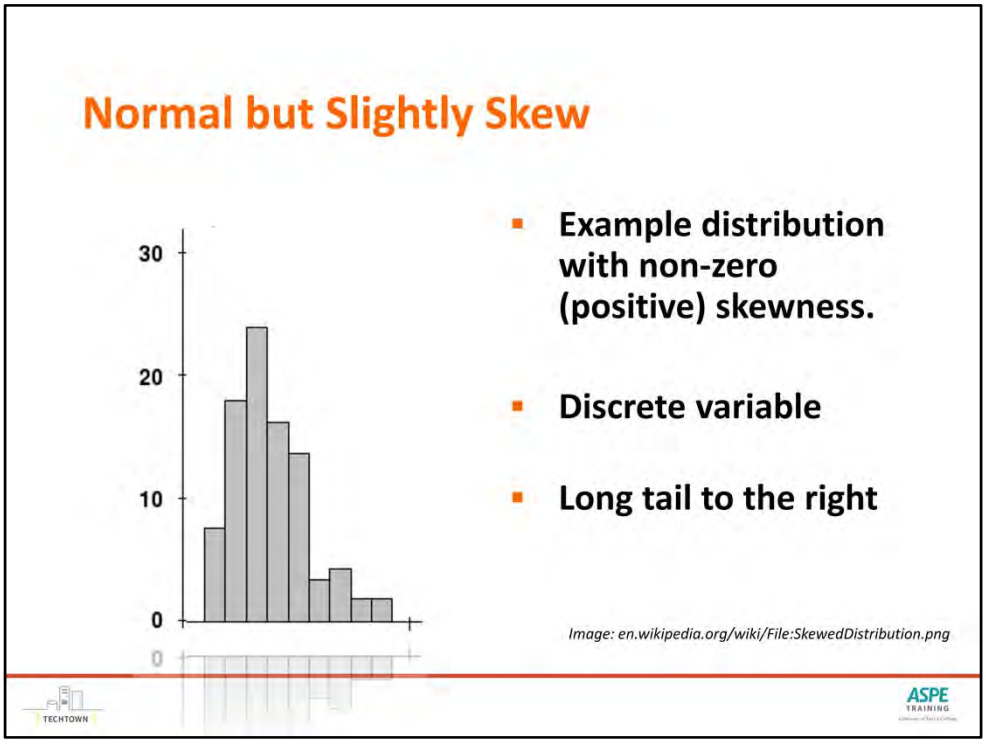
This image shows measurements made of men and women. The bars in the histogram show the actual counts of the distribution (the discrete data set) and the lines show the fit to the distribution. This is very often the case. The discrete counts match more closely the fits when the number of people measured increase. It is unknown how many were counted in this sample.



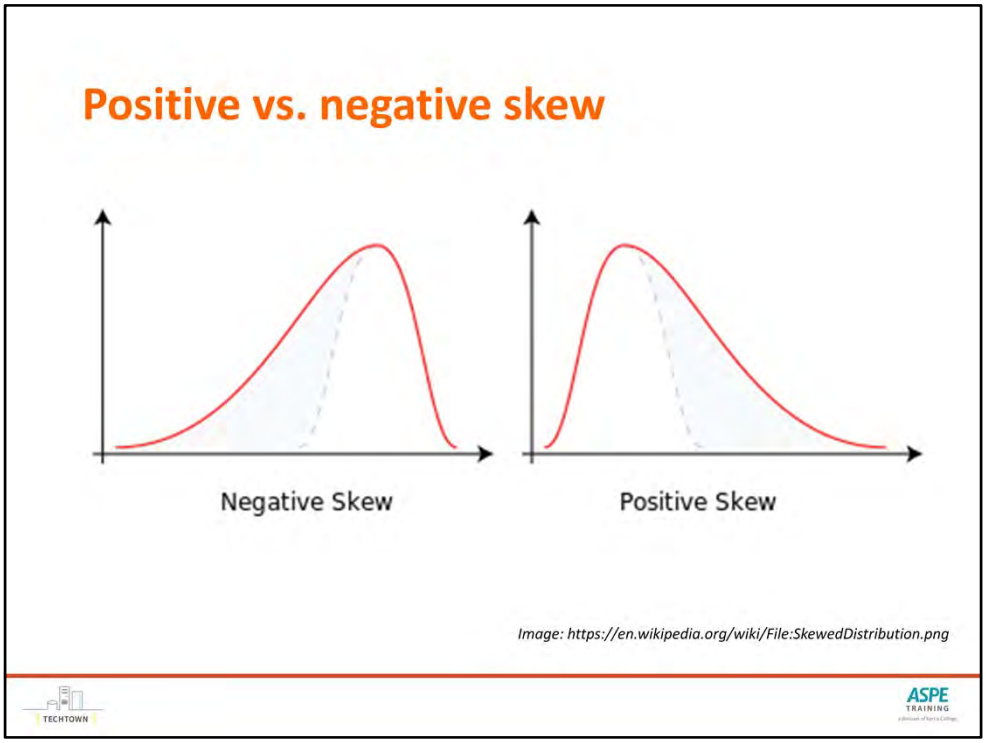
This is a bimodal distribution. They are built from two different normal distributions with two different means laid on top of each other.



The box plot is a typical way of describing the entire distribution of a data set graphically. It shows the relationship between the quartiles, the IQR, whiskers (the outer vertical lines) and any points that may lay outside these whiskers – commonly defined as outliers.

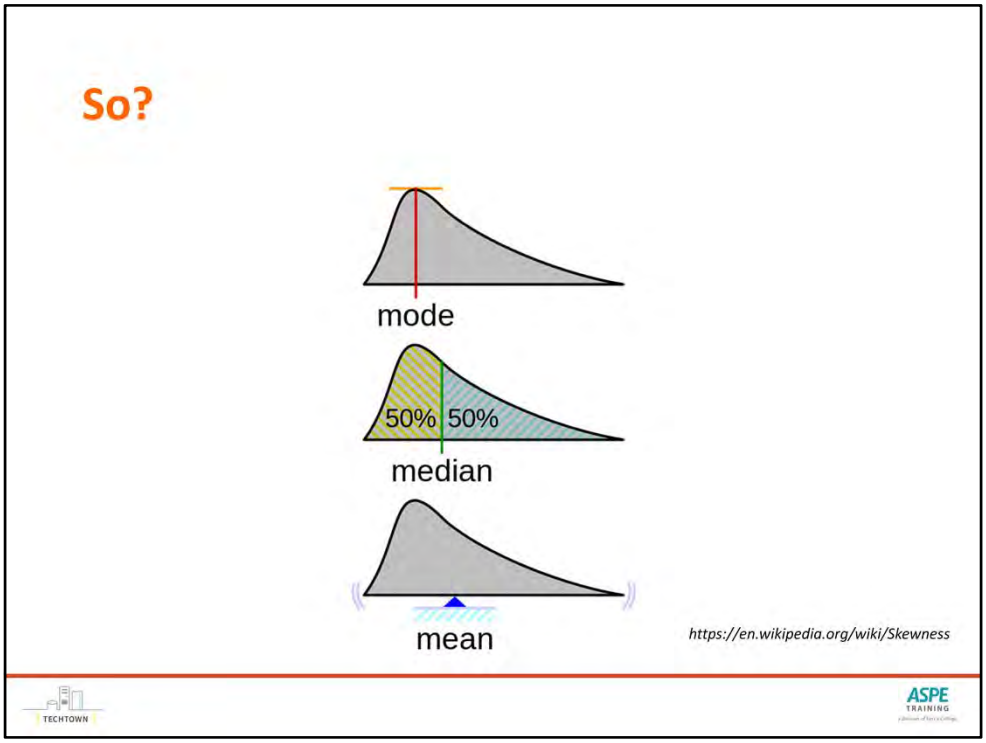


When a normal distribution is weighted more heavily on one side, we say it has skew.



Be careful what we call positive and negative skew, it follows the tail.

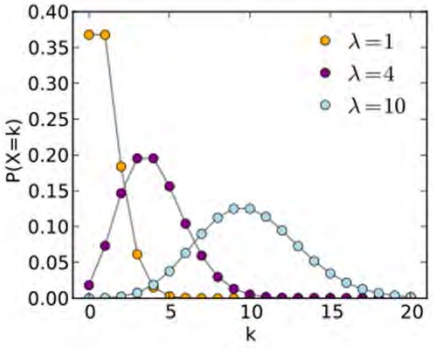
NOTES:



When we have a skewed distribution, the mean, median and mode tell us something about the skew. Here we see that the mode is at the peak of the curve. The median tells us the 50/50 point and the mean is the “balance” point. The more radically skewed the distribution becomes, the more these three numbers will deviate from each other.

Poisson Distribution

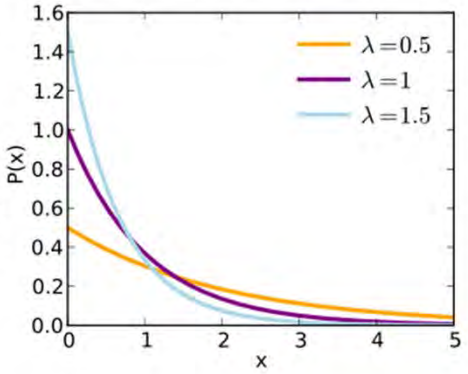
- Independent Events with a known average rate
- Calls between 11AM and Noon at a call center



Another type of distribution is formed from two constraints on the normal distribution. First, have positive only events (0 or above). Second, have few events. With these we see something called the Poisson distribution. The example here is a call center where we count the number of calls between 11am and noon. The average rate is low and the events are independent and we cannot have negative number of calls.

Exponential Distribution

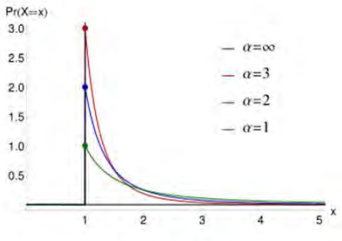
- Shows decaying probability
- Time until the next call at a call center



This is a different type of distribution altogether. It shows that fall off, or decay, as the x variable gets larger.

NOTES:

Pareto Distribution ("80/20")

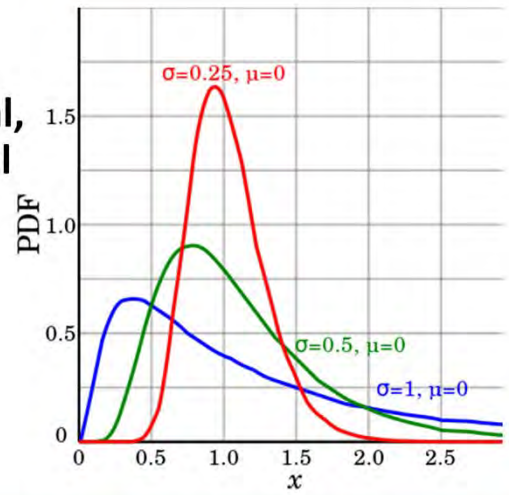


- Special type of Power Law distribution
- Wealth distribution
- Size of meteorites

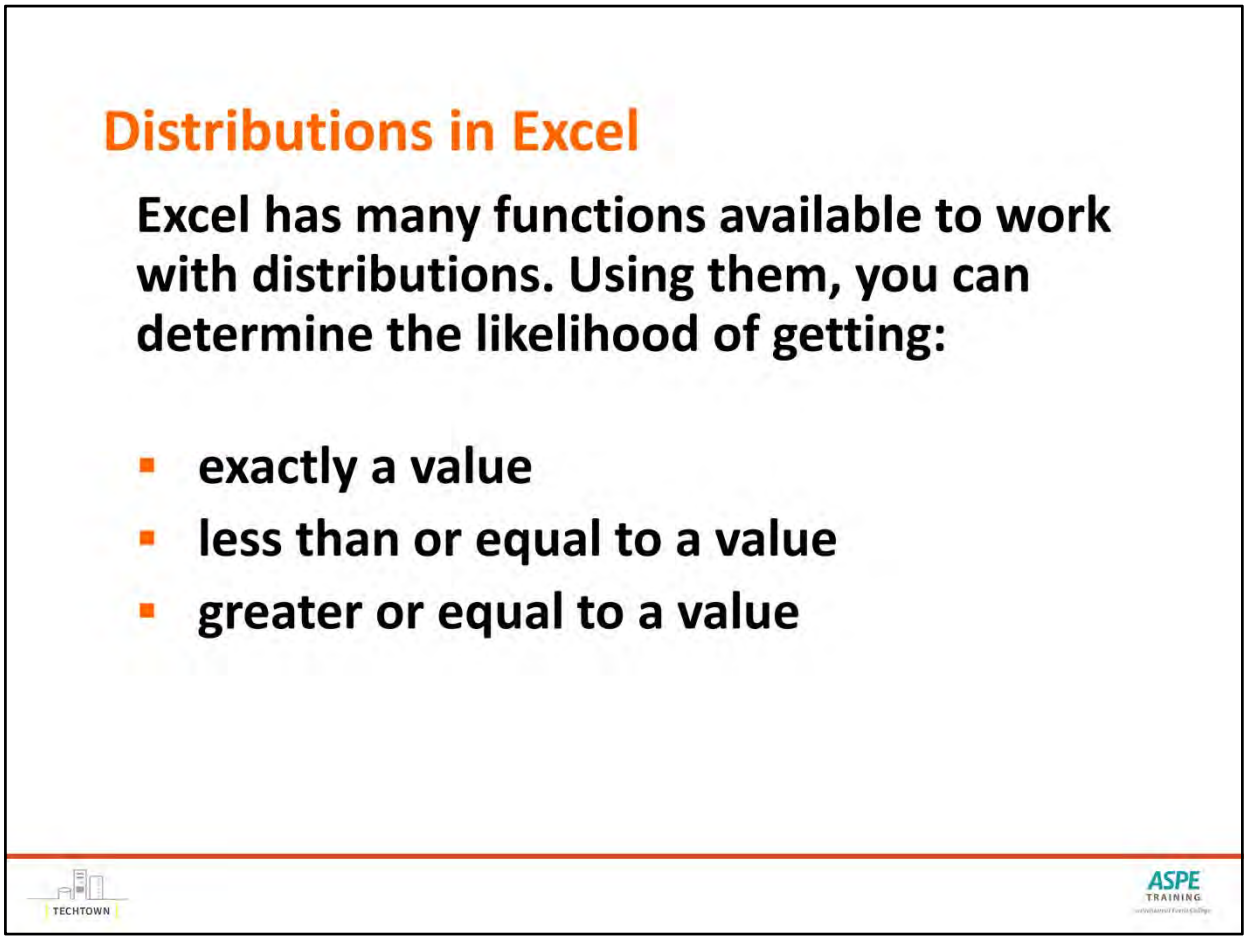
The Pareto comes from a trend seen in manufacturing but applies to many things. We know it as "80% of the work takes 20% of the time and the remaining 20% of the work takes 80% of the time".

Log Normal Distribution

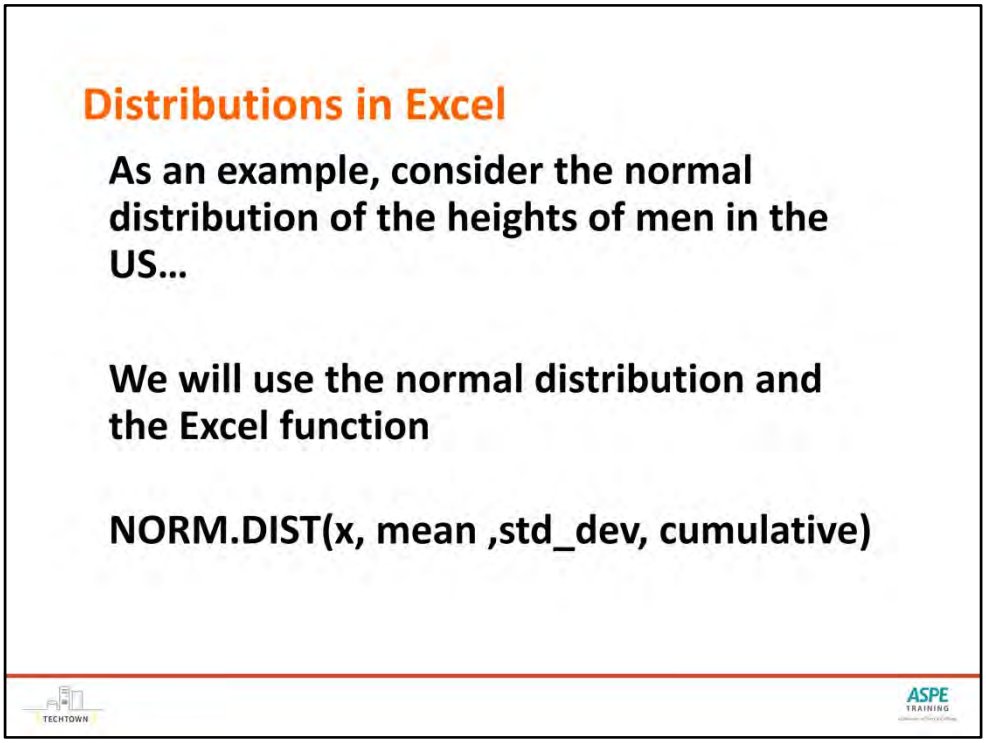
- If X is Log Normal, Y=ln(X) is Normal
- Dwell time on online articles
- Repair times



NOTES:



Now that we have seen a number of distributions, let's look at how we can use Excel to extract information out of them. What follows is a series of examples of how to use Excel formulas for the distributions.



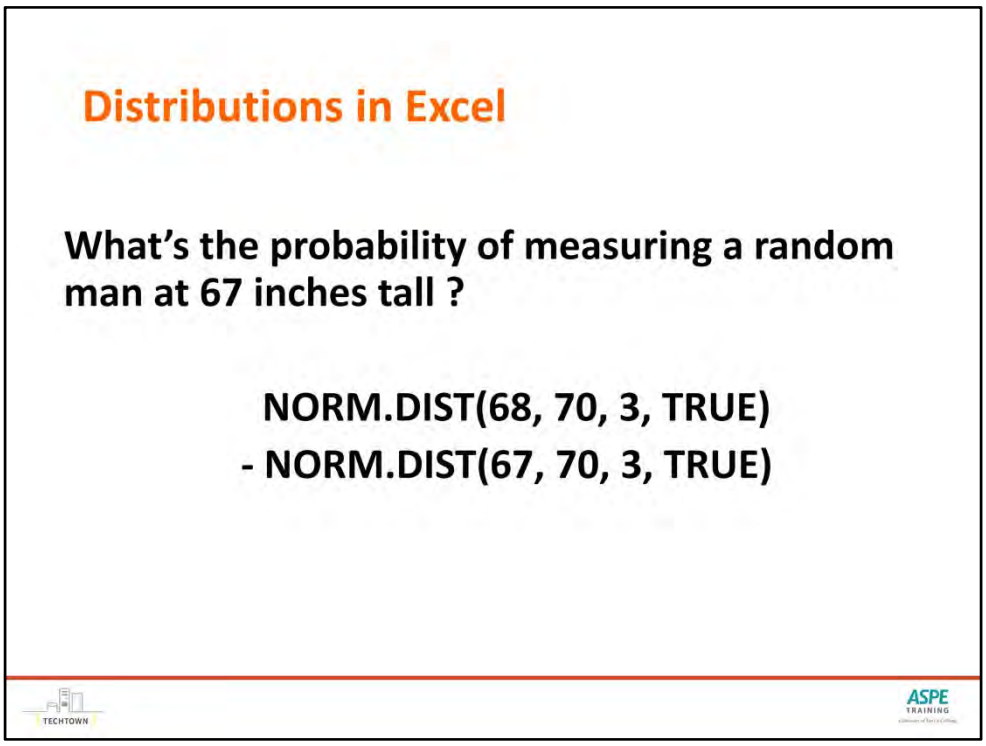
This Excel function allows us to pass the number of interest (x) and define a normal distribution with the mean and std_dev of our choosing. The last argument tells us whether we are using less than and equal to (TRUE) or just the probability density at that point (FALSE). FALSE requires more advanced understanding of the PDF, we use TRUE throughout.

Distributions in Excel

What's the probability of measuring a random man at less than or equal to 67 inches tall ?

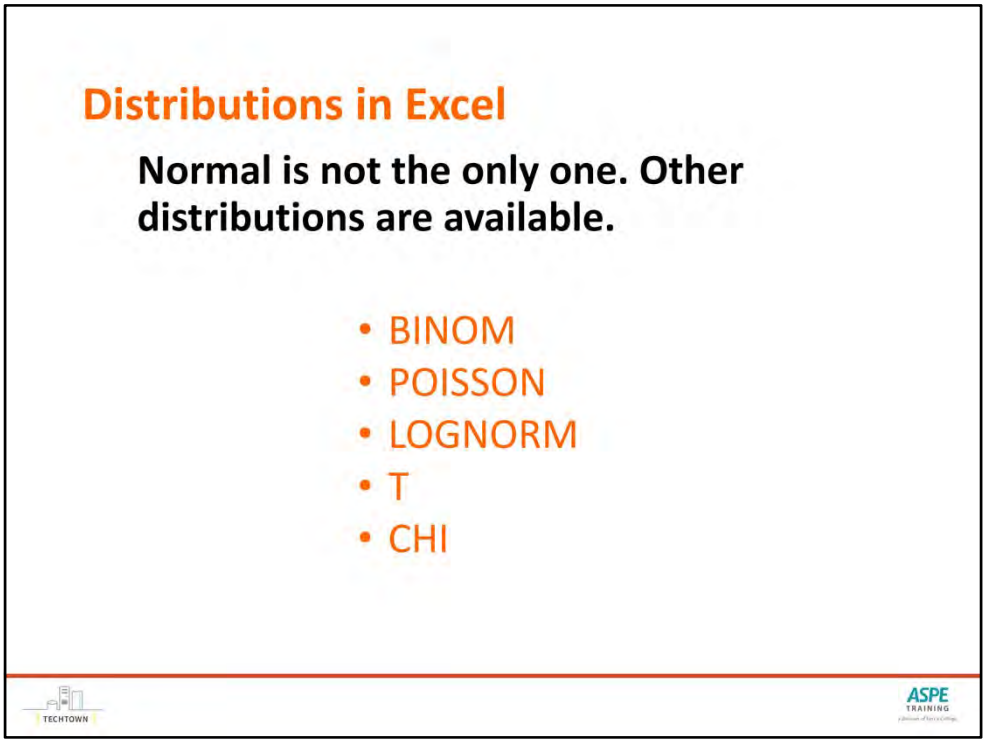
NORM.DIST(67, 70, 3, TRUE)

NOTES:



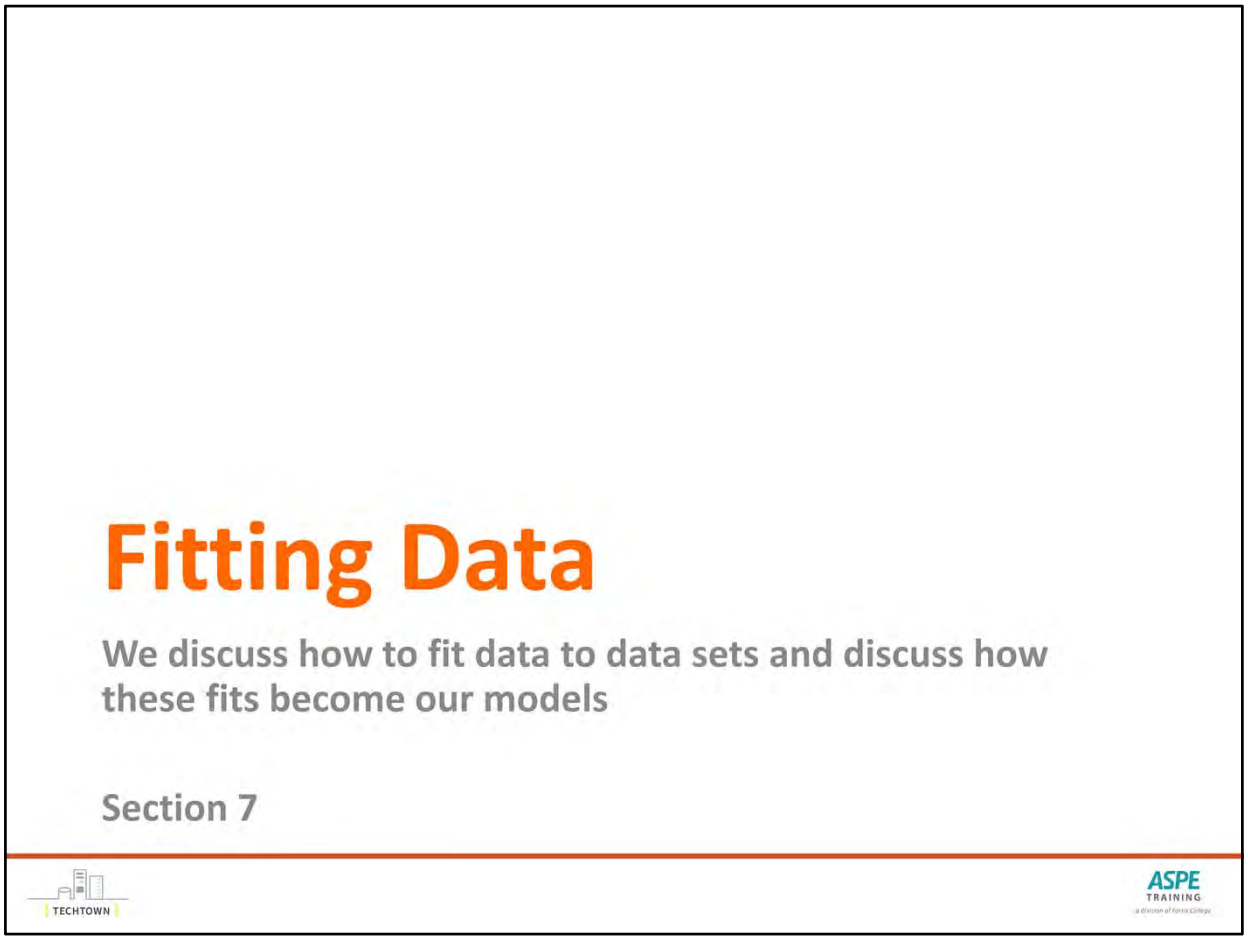
We find the value at 68 inches and subtract the value at 67 inches because we are only interested in whole number inch values (not 67.2 or 67.6, etc).

NOTES:

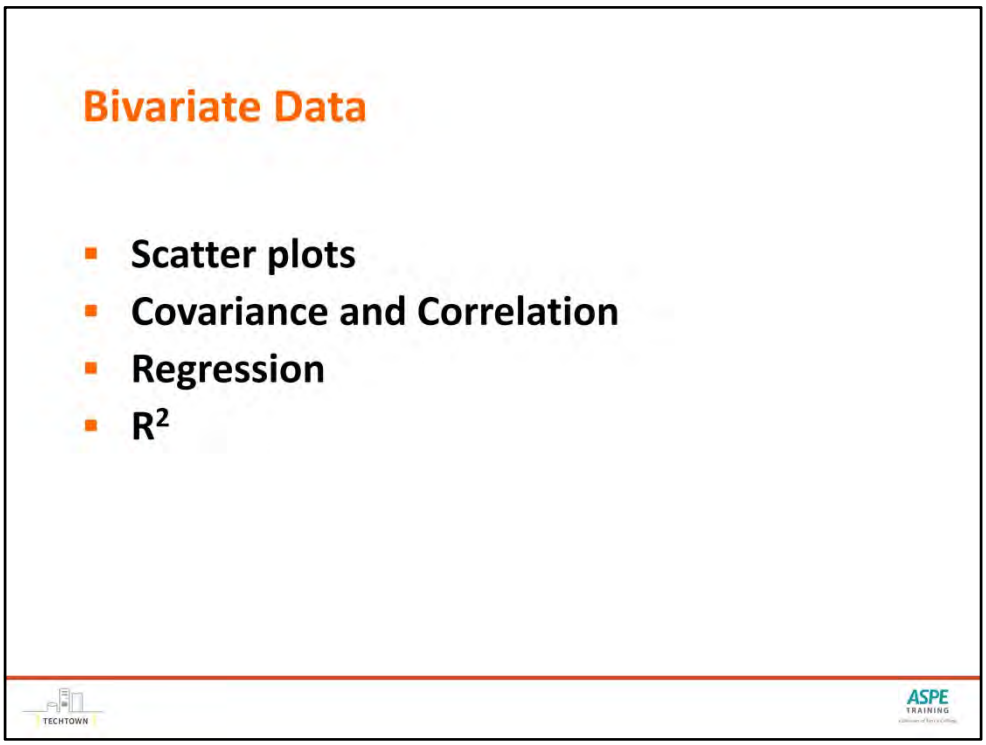


The ”1 – “ is because

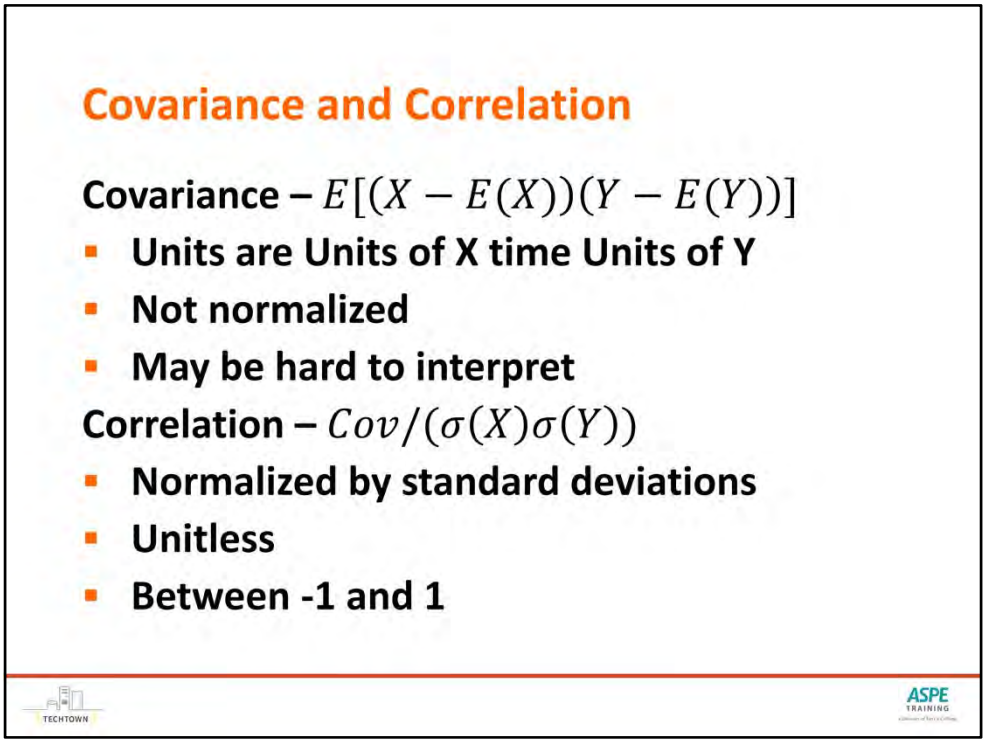
- All distributions sum to a probability of 1
- The formula with cumulative set to true gives us “less than or equal to”. The ”greater than” part is 1 – “less than or equal to”



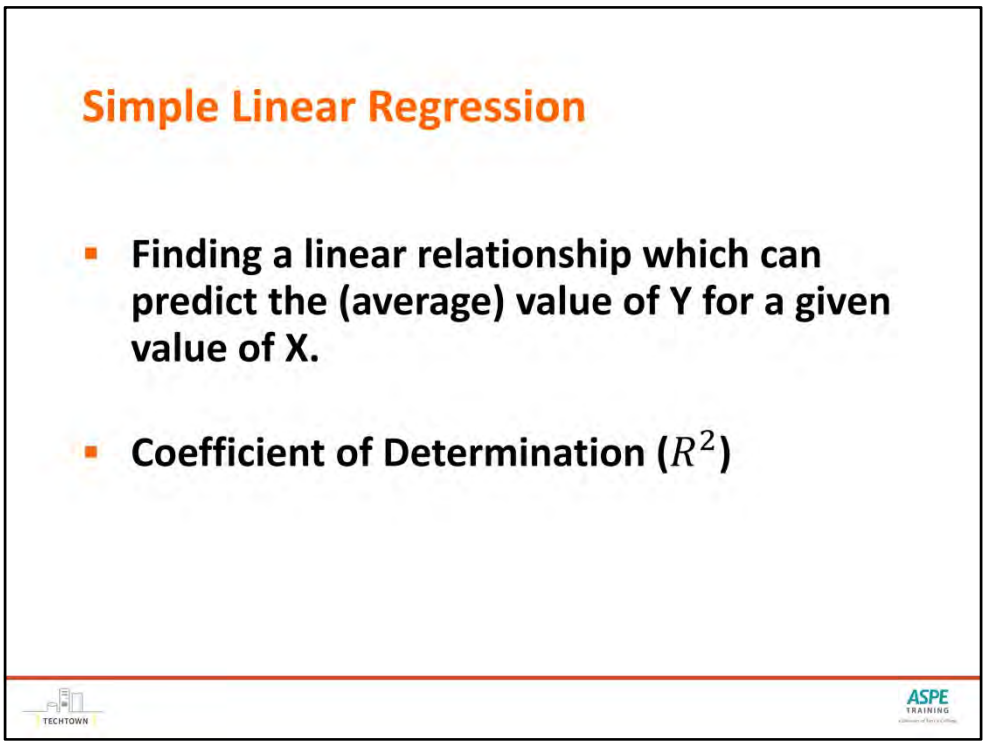
From here we are moving into basic modelling. We will look to fit math equations to data and notice what these fits can tell us. Once we know the math equations, we can make predictions about what is to come. This power is limited for a number of reasons which we will discuss along the way.



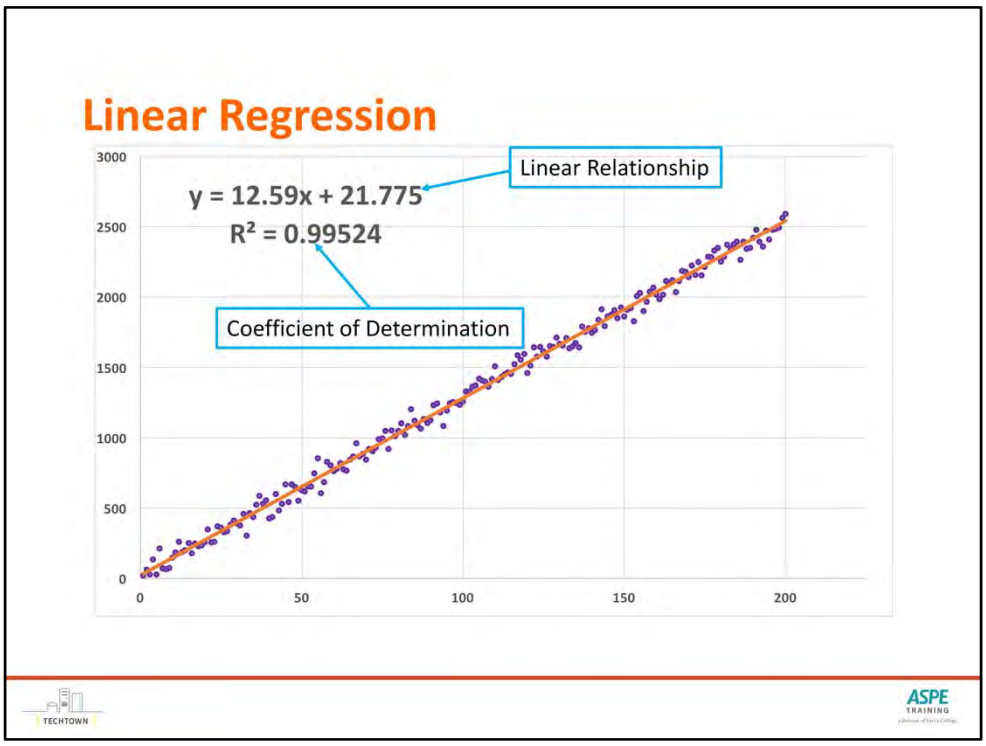
The first thing we will look at is data of two variables. Our best visual aid here is often the scatter plot where we put one variable on the x axis and another on the y axis. We can look at the covariance and correlation of the data set. We use regression to calculate the mathematical fits to the data, then we look at R2 to see which is the best one. Be cautious, the literal best fit may disobey rules of the field you are working in. It is, as always important to remember that the math is not the only thing at work here. What is the problem? What is field of research?



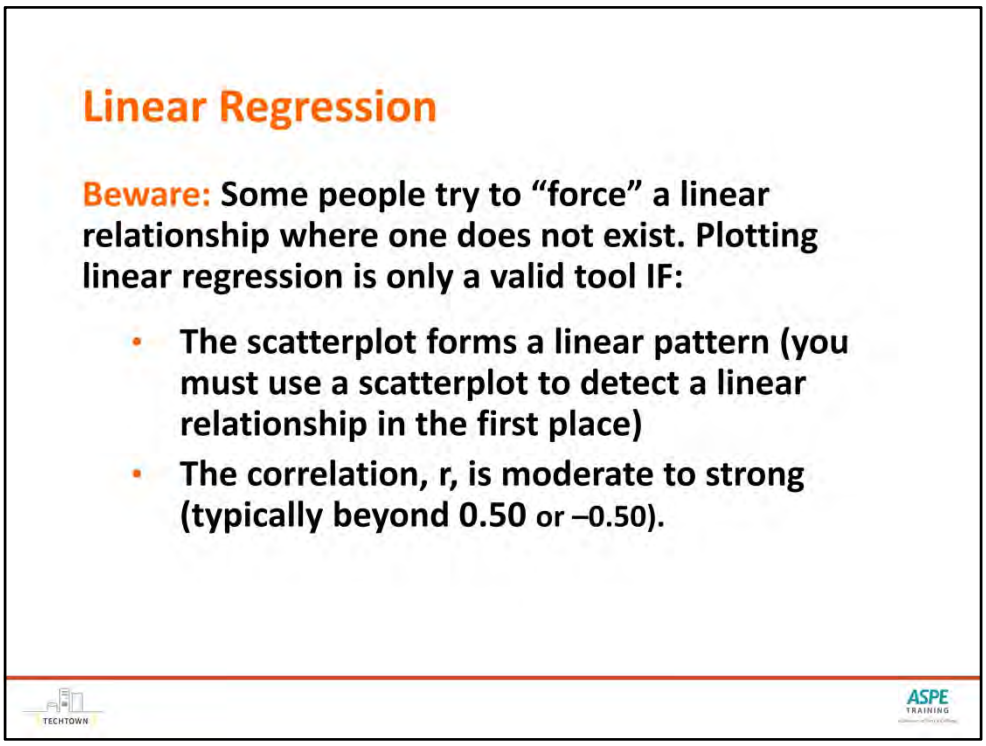
Covariance is, simply put, how the two variables vary together in the set. The formula is shown in the slide. We need it because we really want the correlation, which is the covariance divided by the std dev of both variables. Don't worry, our tools will do this for us.



The “regression line” is a straight-line descriptor. If you have slope and y-intercept of the regression line, you can plug in a value for X and predict the average value for Y.



This fit is calculated by minimizing the sum of the squared differences between each point and the line. This amounts to minimizing the Variance of the data set (see slide on Variance). This is called the **Method of Least Squares**.



These are some warnings about curve fitting. One note about the moderate values of r… in the real world, r = 0.5 is not enough for a good correlation.

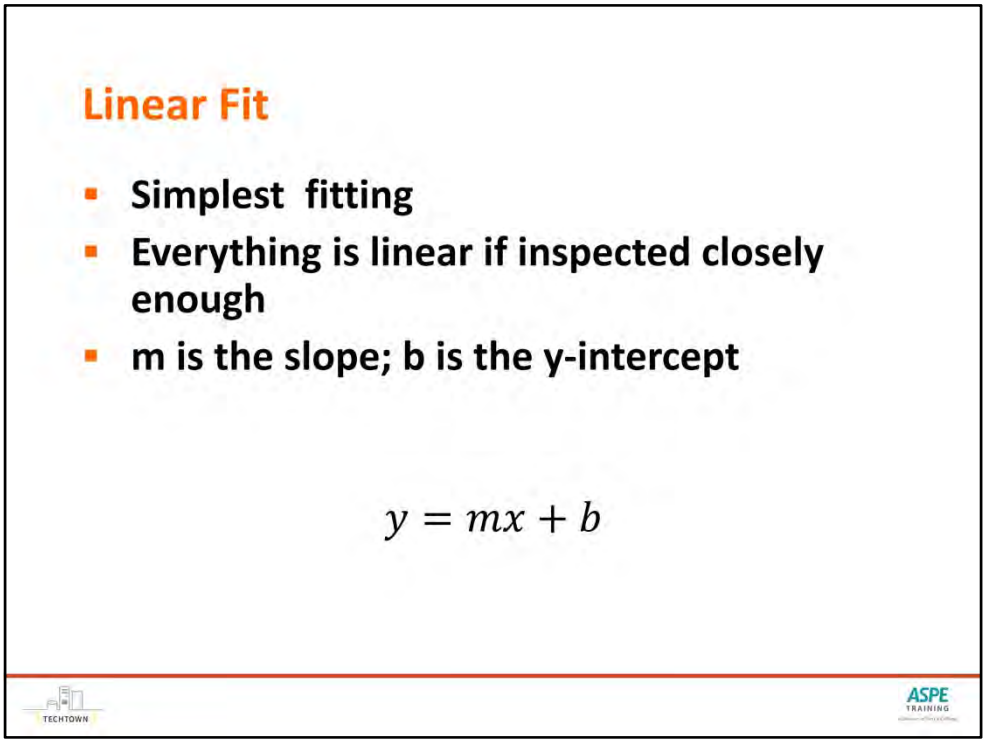
NOTES:

Other Fitting Functions

- Linear – $y = mx + b$
- Polynomial - $y = a + b*x + c*x^2 + \ldots$
- Power Law - $y = a*x^b$
- Exponential – $y = a*b^x$ $(a*e^x)$
- Sinusoidal – $y = a*\sin(b*x + c)$
- Logarithmic, etc.

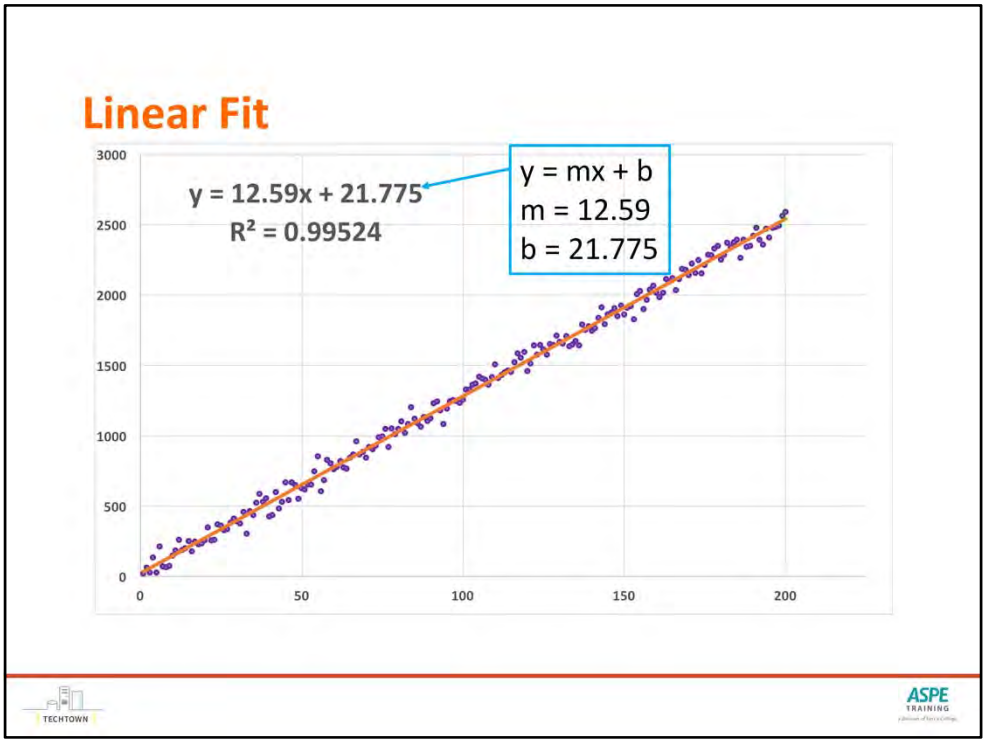
Linear regression is a technique and we can fit a lot of different functions with it. Some of the function types are listed above.

NOTES:



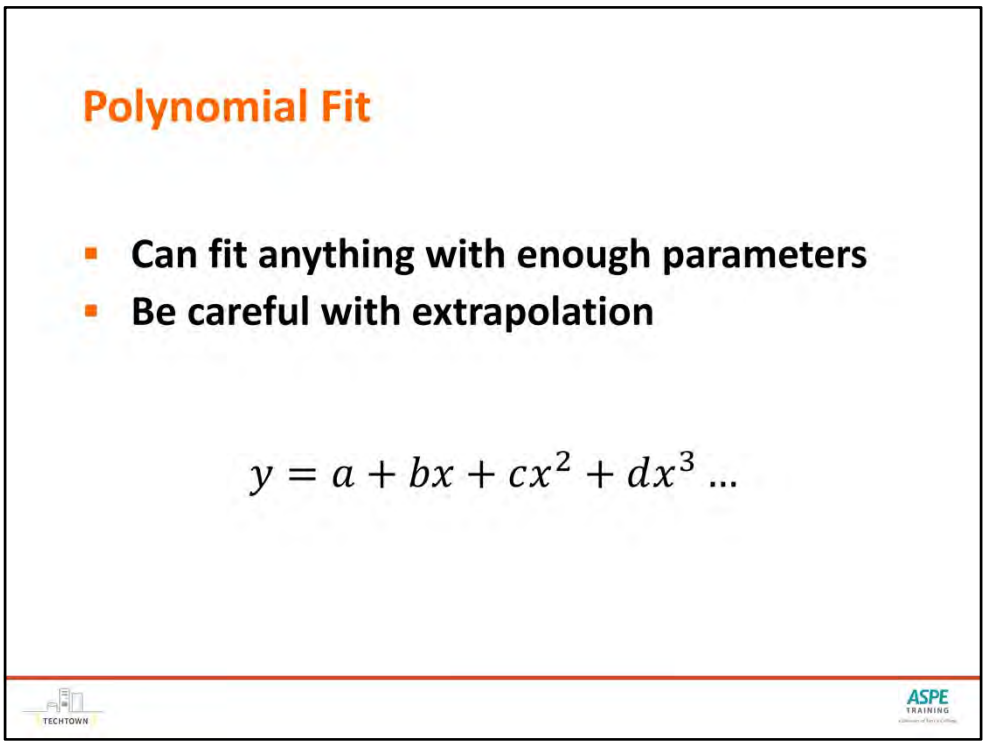
The simplest is the straight line. The equation of the line, when you show it in Excel looks like the equation here, except m and b will be numbers determined by the fit.

NOTES:

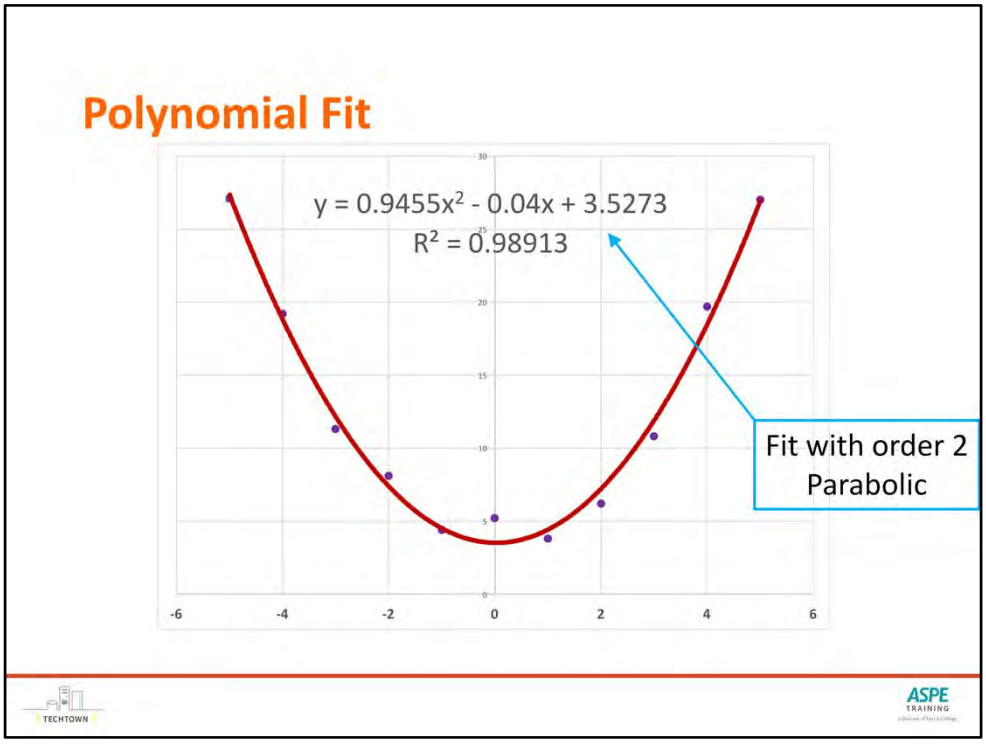


This is what Excel will display.

NOTES:

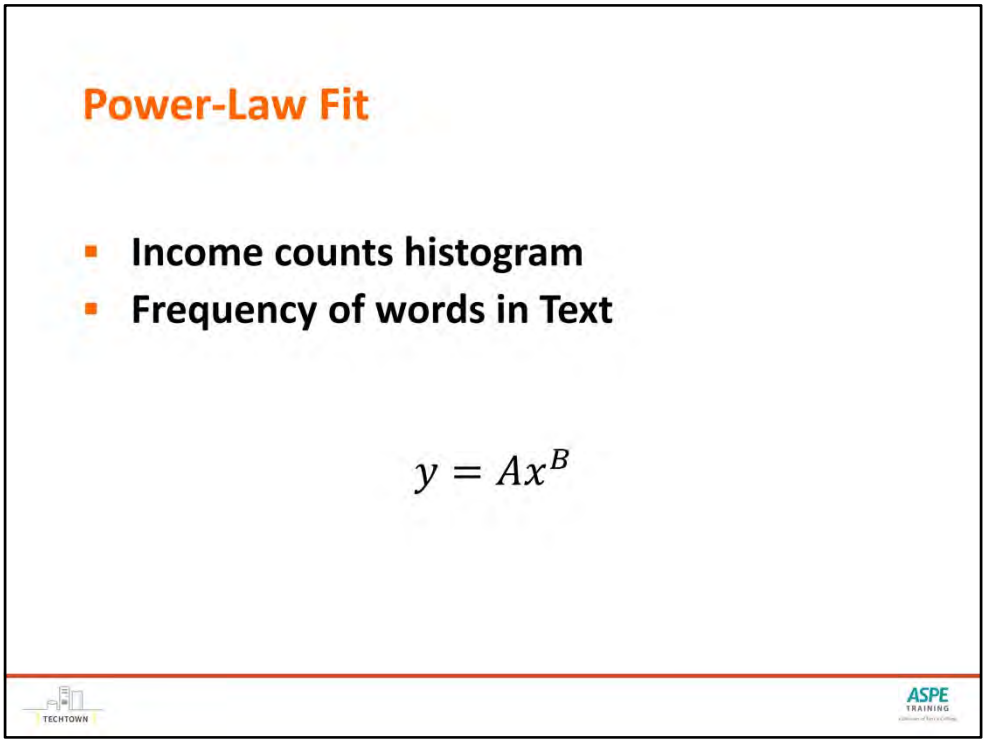


A polynomial is an equation with higher powers of the x variable (the independent one). Be very careful with this… you can almost always fit any data set perfectly by going to enough powers of x, but it does not mean that you are correct. Certain fields of study obey know laws or trends and these determine the type of equation you should use. A bad fit of the data to this equation could mean you have bad data, not require higher order fits. You, as the analyst, should know as much as you can about the field to determine these things.



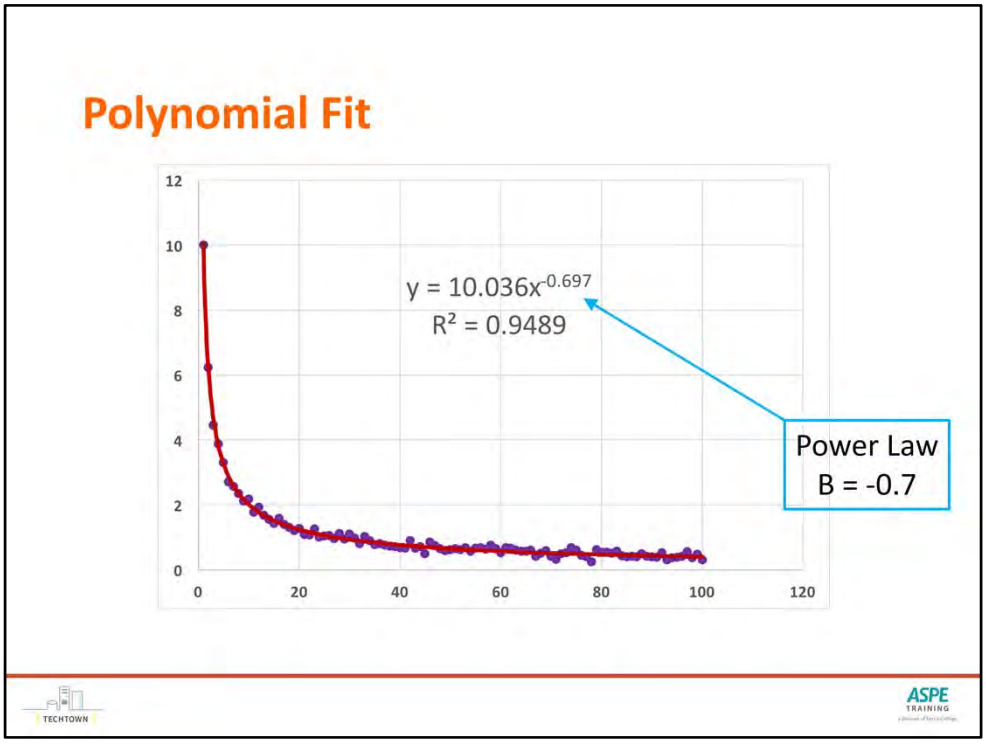
This is an example of a parabola. This is really common in studying new materials in physics and engineering. Notice the equation has a squared x term in it.

NOTES:



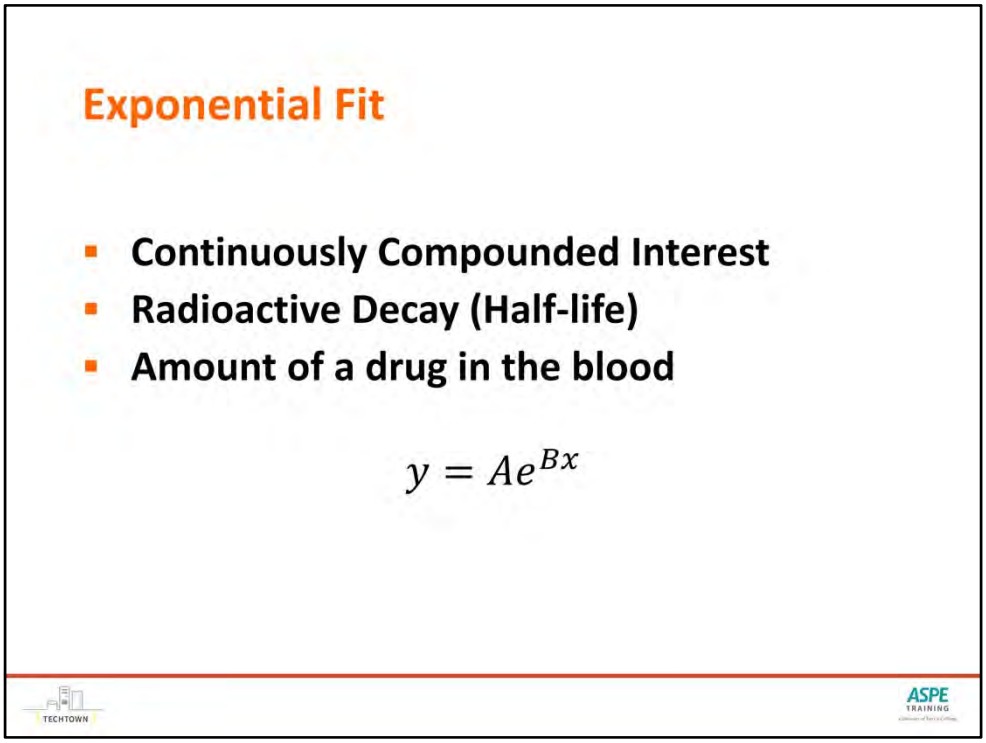
The power law is a fit where B is some number (2,3,4,5,6,etc).

NOTES:



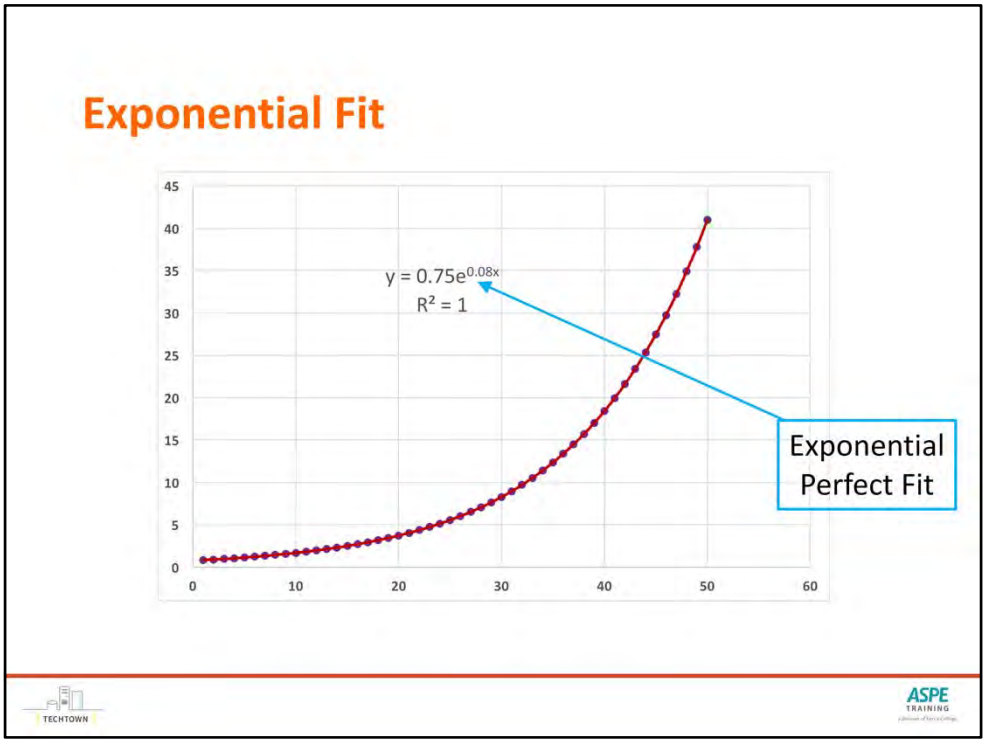
Here is one where B is negative and a fraction.

NOTES:



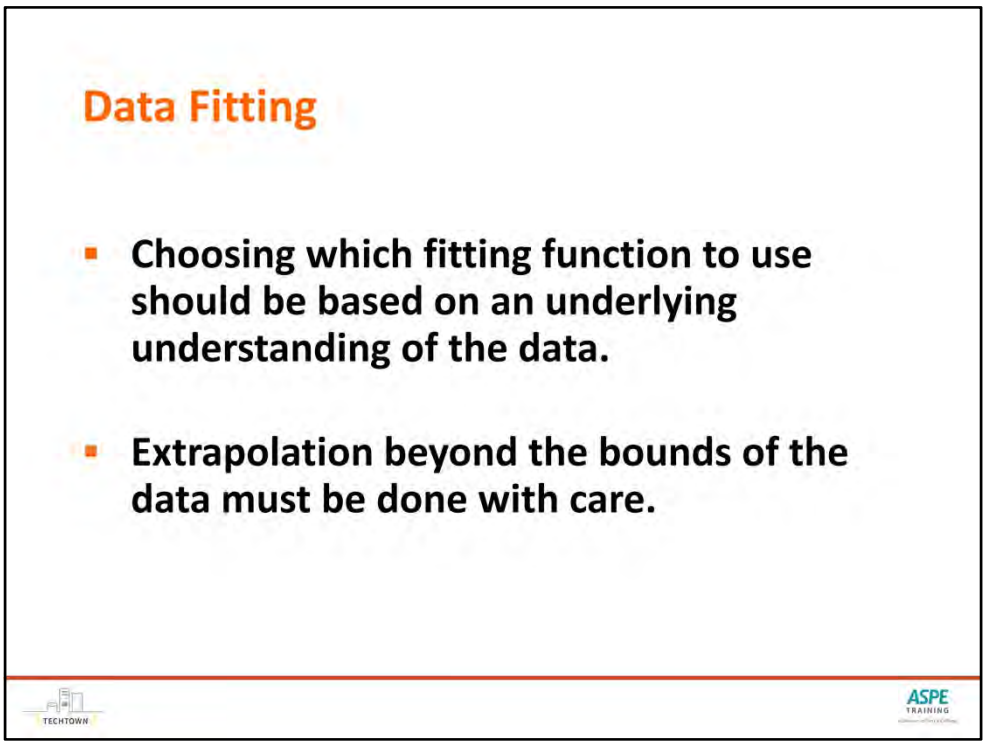
The exponential fit is a common fit to data such as the examples stated.

NOTES:

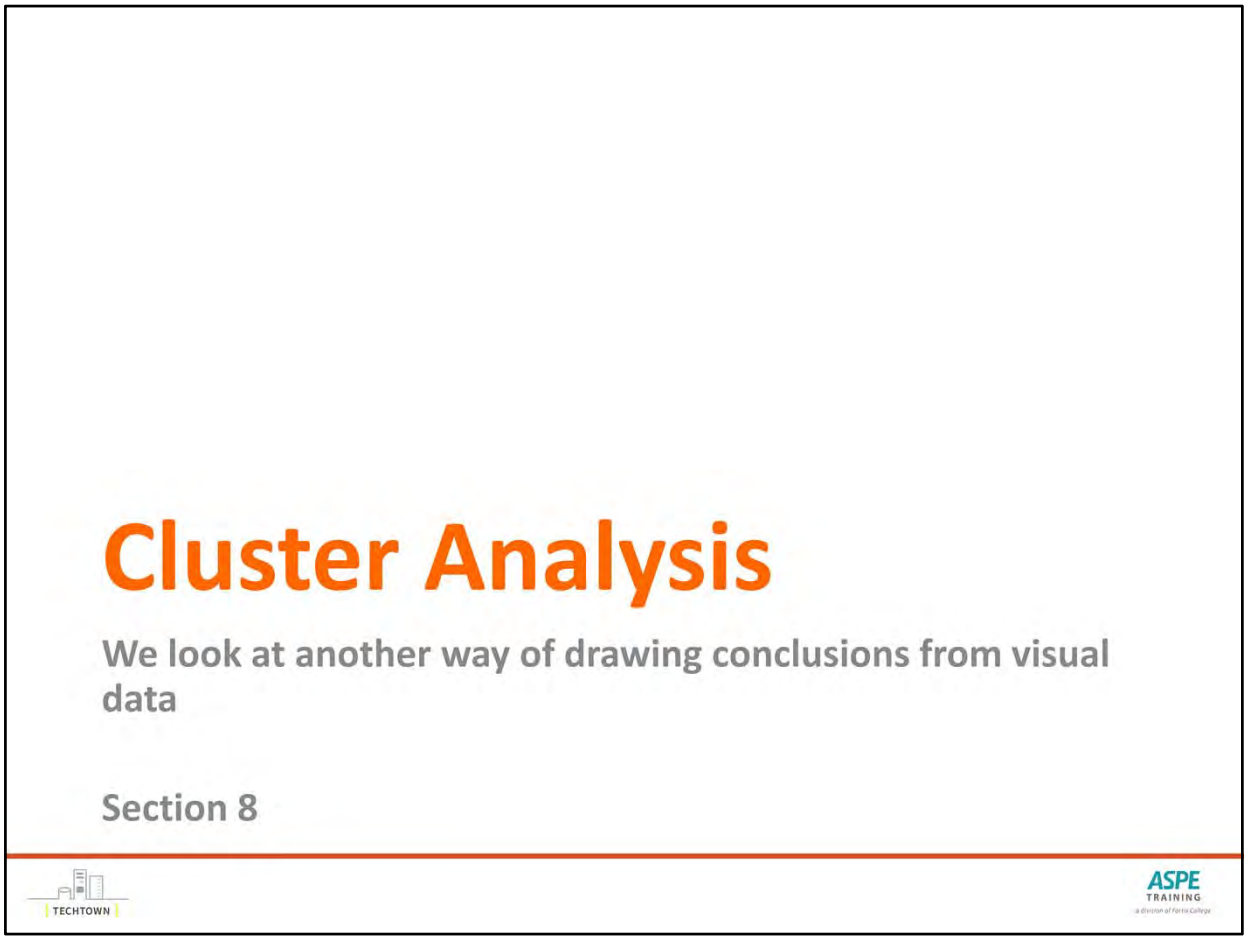


Continuously compounded interested? This is how much money I owe?!

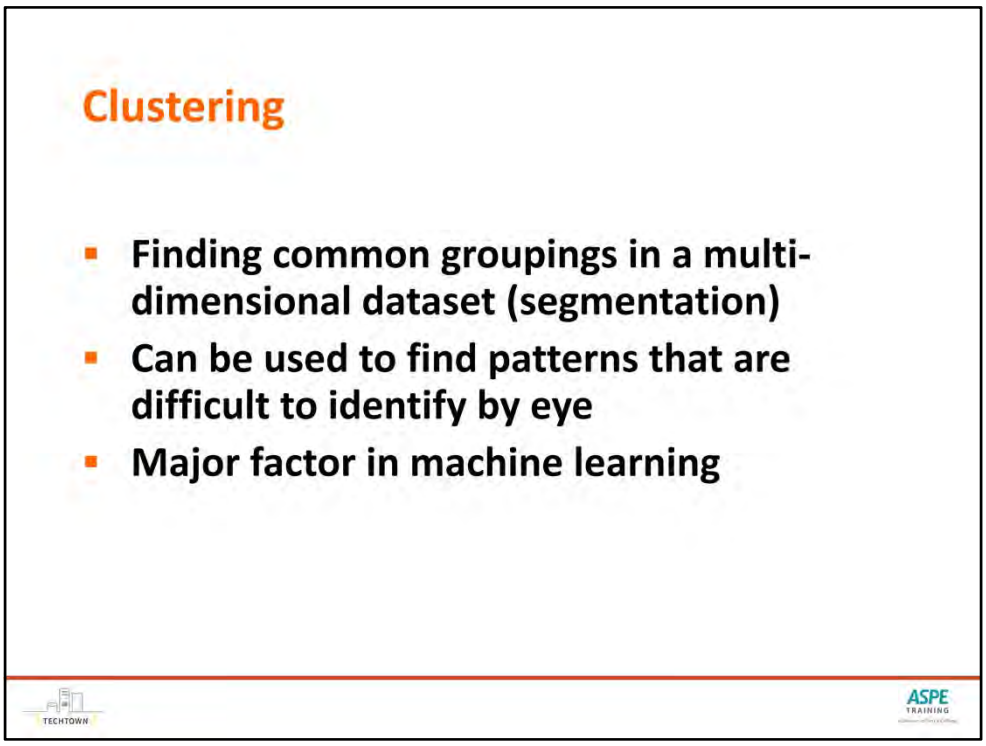
NOTES:



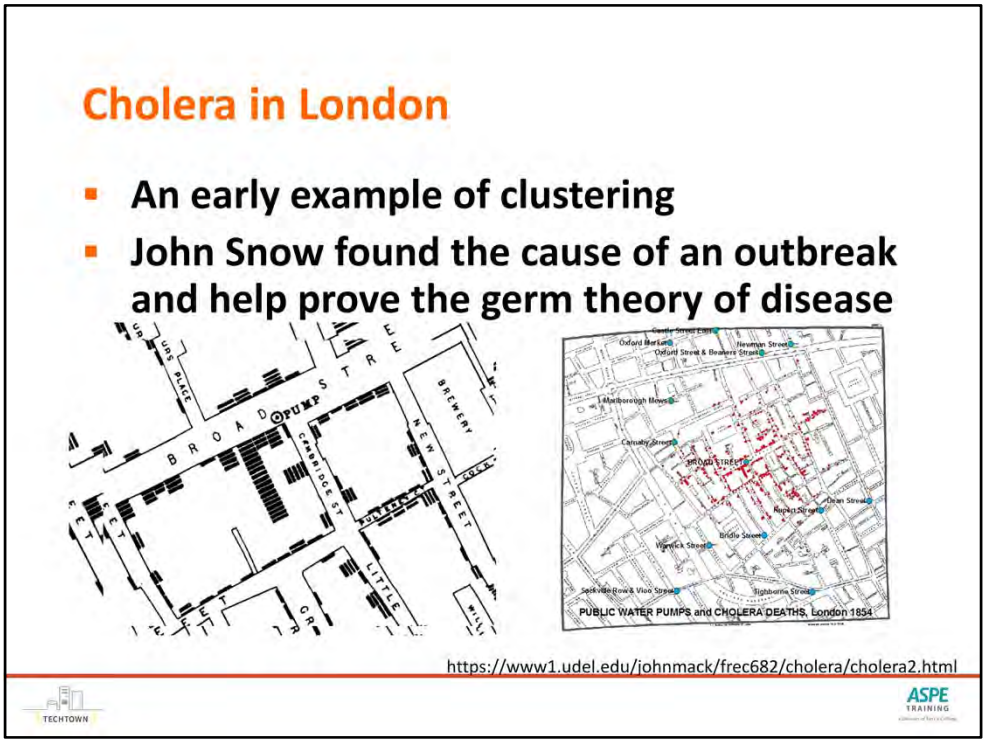
Remember, choosing the fitting function is more than just trying the functions until one gets a high R2 value. It is a powerful tool, but you are fitting things that have already happened. Be careful about predictions you make from it.



Let's look at a different way to visualize data. It still involves scatter plots, but the analysis is different than simply looking at a data fit.



Clustering is where we apply a number of algorithms to find patterns in multidimensional data sets. These patterns are instrumental in machine learning. Remember that we mentioned credit card fraud some time ago in the course, well, that type of analysis is a major clustering and pattern recognition algorithm in which analysts look for patterns and anomalies from the those patterns.



In one of the most famous cases of pattern recognition from clustering analysis, a doctor in 1850's London stopped a deadly disease and at the same time proved the germ theory of disease.

John Snow investigated the cause of a cholera outbreak by documenting where the cases were located. Discovered that people who lived near, or travelled by, the Broad St. water pump were much more likely to get cholera. A nearby cesspool had leaked in to the water supply.
Left – Snows original map of cholera cases in London
Right – Cases shown in red, Water pumps shown in blue.
From: https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html

Voronoi Diagrams

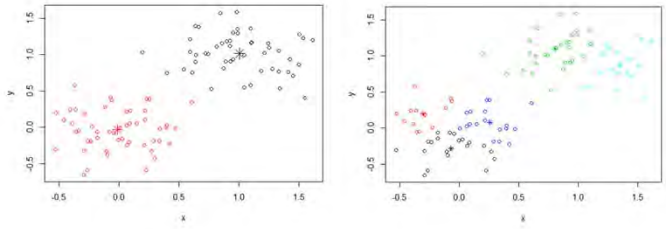
- Show the group of points that are closest to certain centers
- Centers can be chosen by hand, or determined by and algorithm (e.g. K-means)

https://en.wikipedia.org/wiki/Voronoi_diagram

A Voronoi diagram provides a visual map of the distance of any point in space to a specified center point in a region. That center point is usually the mean of a cluster, though it could also be something called a medioid. The regions in the map above show areas in a color in which all points in that color area are closer to the black dot in that same color area than they are to any other black dot. This has been applied to John Snow's map and the computers arrive at the same conclusion he did about the source of the epidemic. Today, these algorithms can tell ambulances the location of the closest hospital in real time, and many other things.

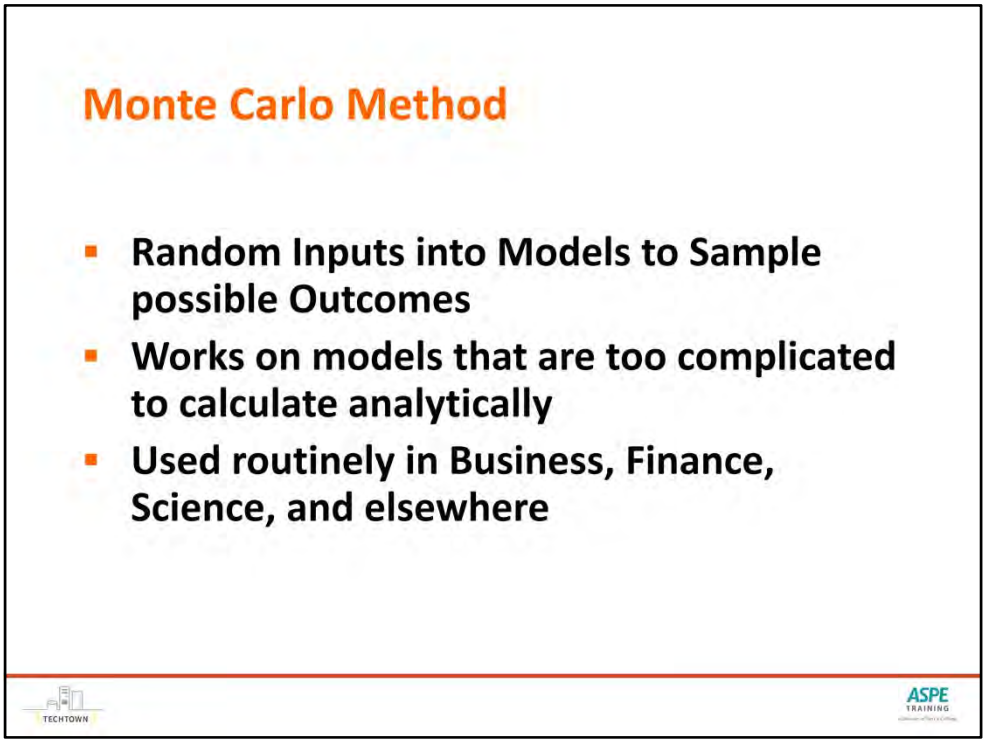
K-Means

- Algorithm that iteratively moves center points to define clusters
- Initial centers are either chosen or placed randomly



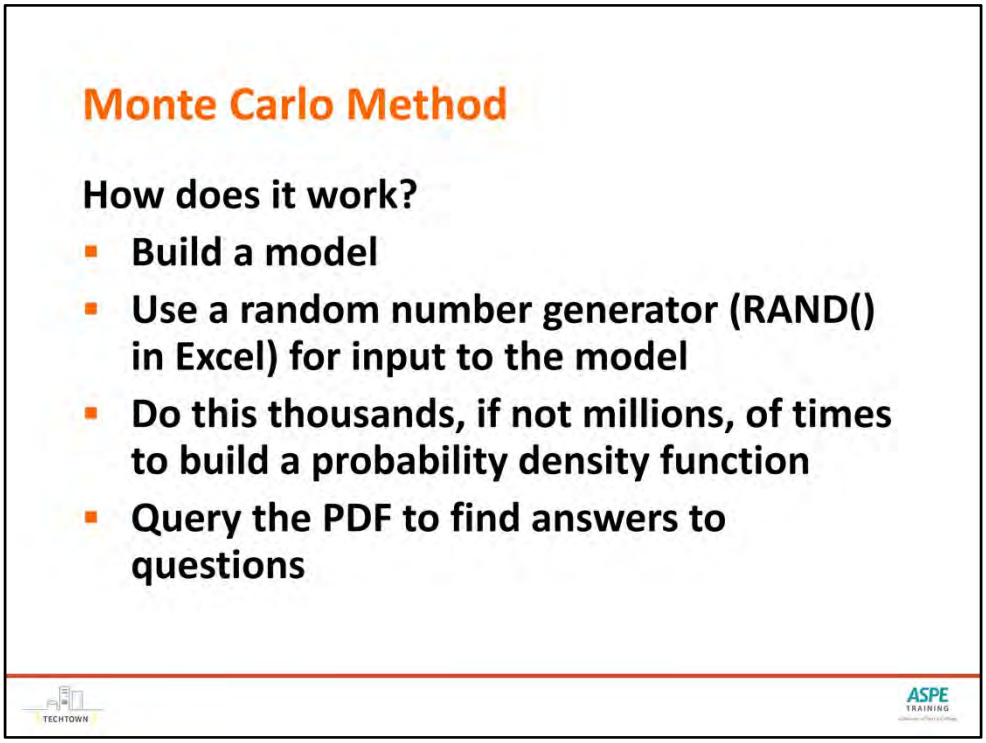
These charts are 2-D examples from R with 2 or 5 centers. The K-Means algorithm works in iterations. It calculates distances from a chosen “center”, then moves the ”center” then calculates distances again, over and over, until the next iteration does not improve the accuracy any. At that point, in K-Means, you have the mean values of clustered data sets.

Up to this point in the class, we have been looking at modelling data for explanatory power, but with limited predictive potential. There are many methods by which we try to add prediction to descriptive data and we have chosen to look at a common one with a simple fundamental premise – Monte Carlo.



The idea behind Monte Carlo is that if you play games of chance long enough, probability defines how often you will win or lose. The only unknown, really, is the model of the game you are playing. In the case of black jack, you are looking to add to 21 given the cards in a deck and the players in the game. Of course, you can never know for sure how one game will end up, but you can calculate the likelihood that you will win any game, and play enough games, you will win that many times.

In Monte Carlo, we develop a model of our system and we use computers to ”play a game” with that model over and over again. We generate random inputs and send them into the model. Then we keep track of what these results are and start adding and subtracting, multiplying and dividing to “ask the questions” of the results. Don’t worry, we will keep out models simple, but in the real world, the power comes from developing complex models and letting probability and powerful computers find our answers.



Here is a brief description of the steps involved. Note that here, PDF is the probability density function from the distribution of results, not the document format.
The next series of slides shows how to do it in Excel.

Monte Carlo Method

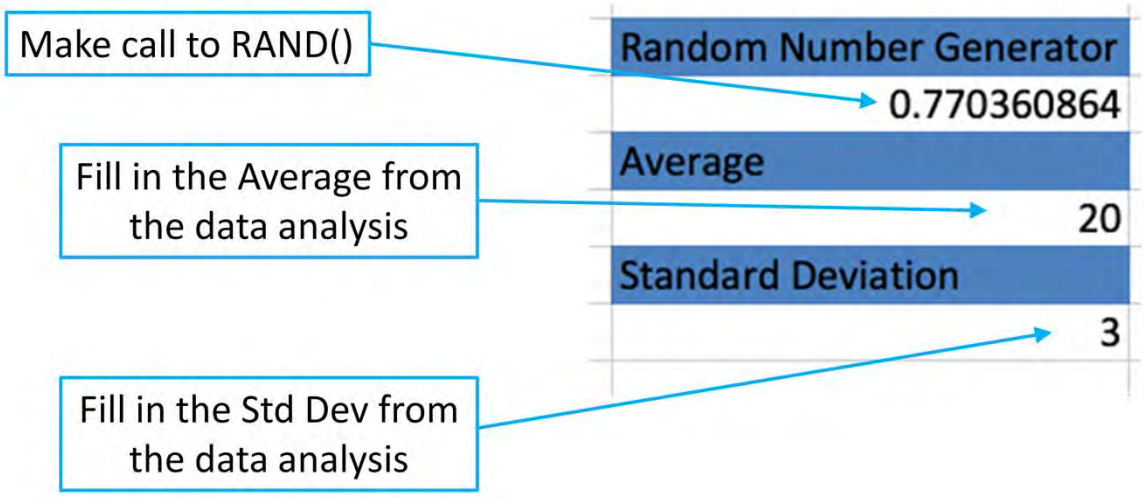
- Data set is 300 previous jobs from DB records
- Analysis shows
  - Average = 20 days
  - Standard Dev. = 3 days
  - Distribution = Normal

Job Number	Days To Deliver
1	18
2	20
3	22
4	17
5	20
6	19
7	20
8	19
9	15
10	23
11	21
12	20

The data set is 300 records, only some of which are pictured here. The first step is to analyze the data with the basics: Average, Std. Dev. and plotting the distribution.

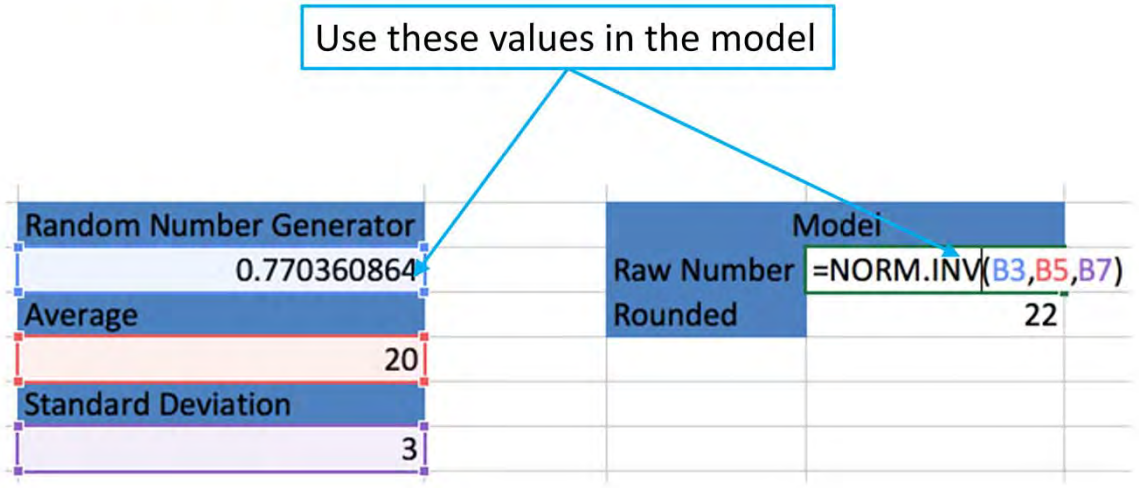
NOTES:

Monte Carlo



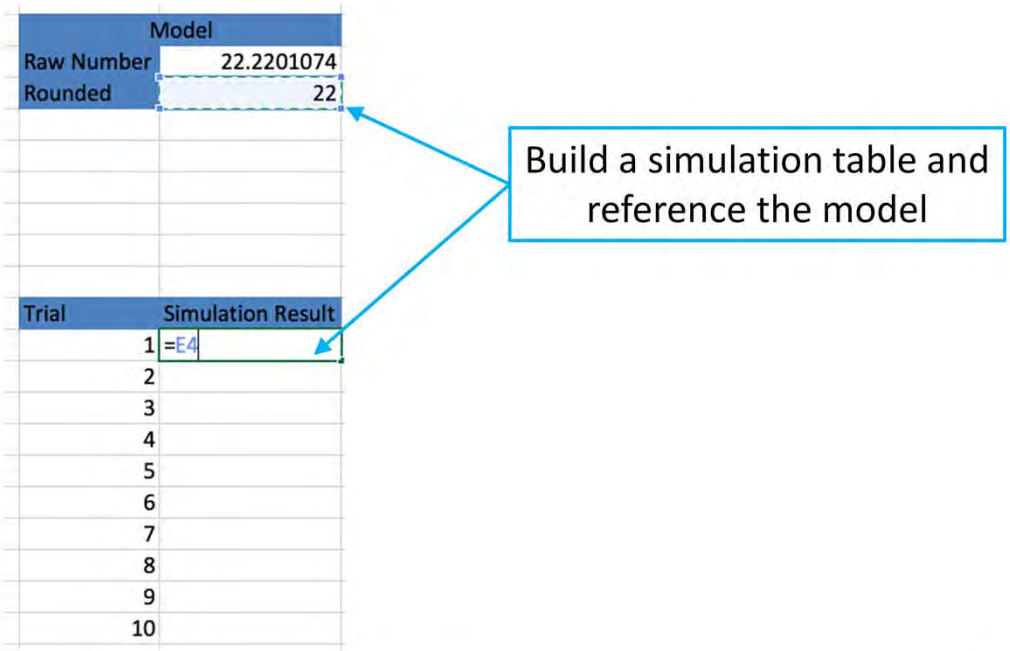
NOTES:

Monte Carlo

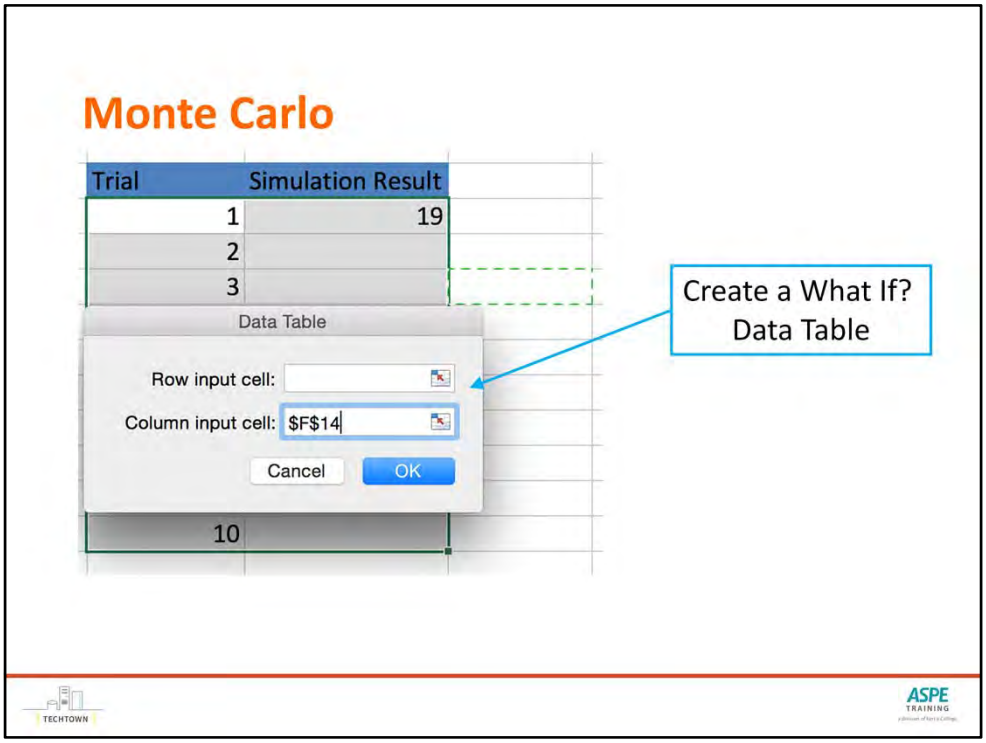


NOTES:

Monte Carlo

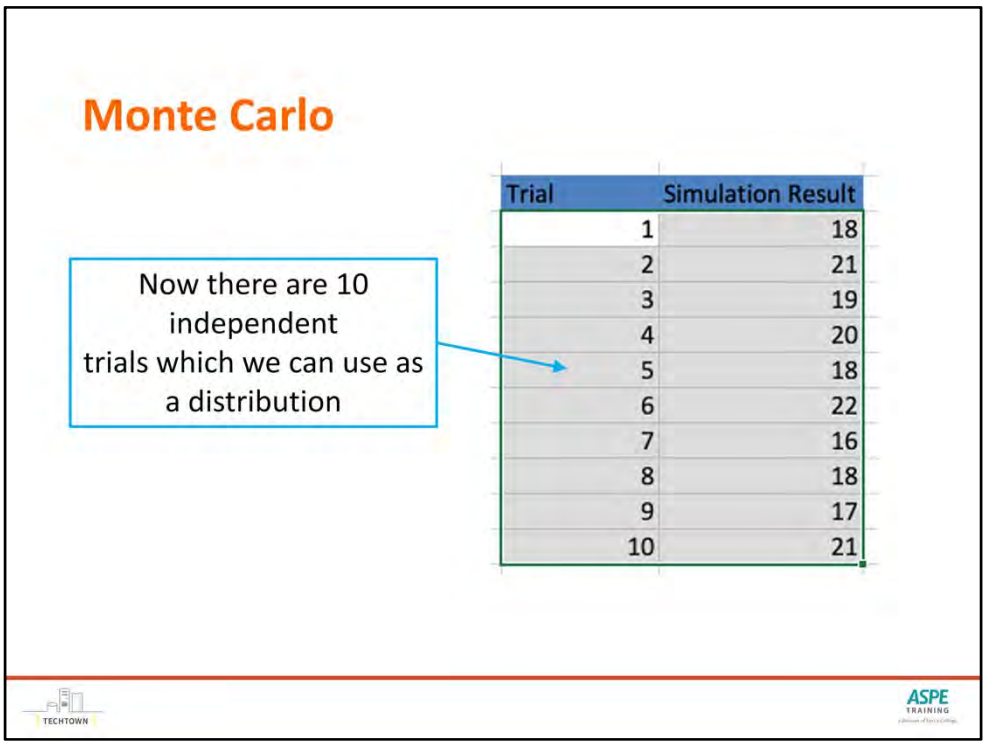


NOTES:



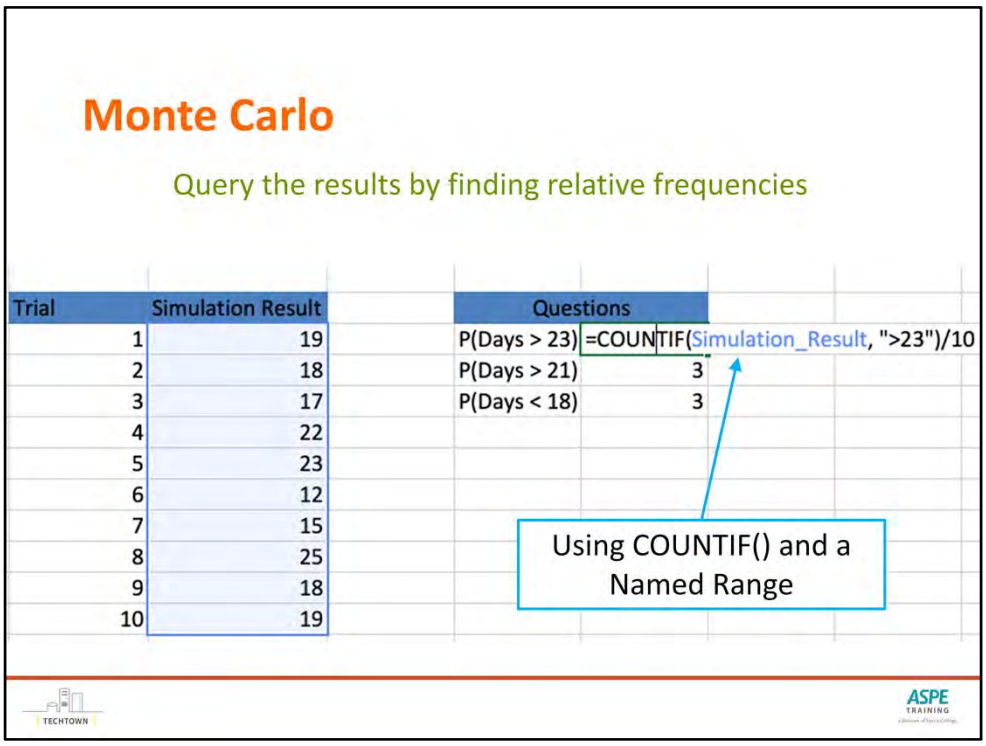
To create a What If? table, you need to click on an ”off, empty cell” for the column input cell

NOTES:



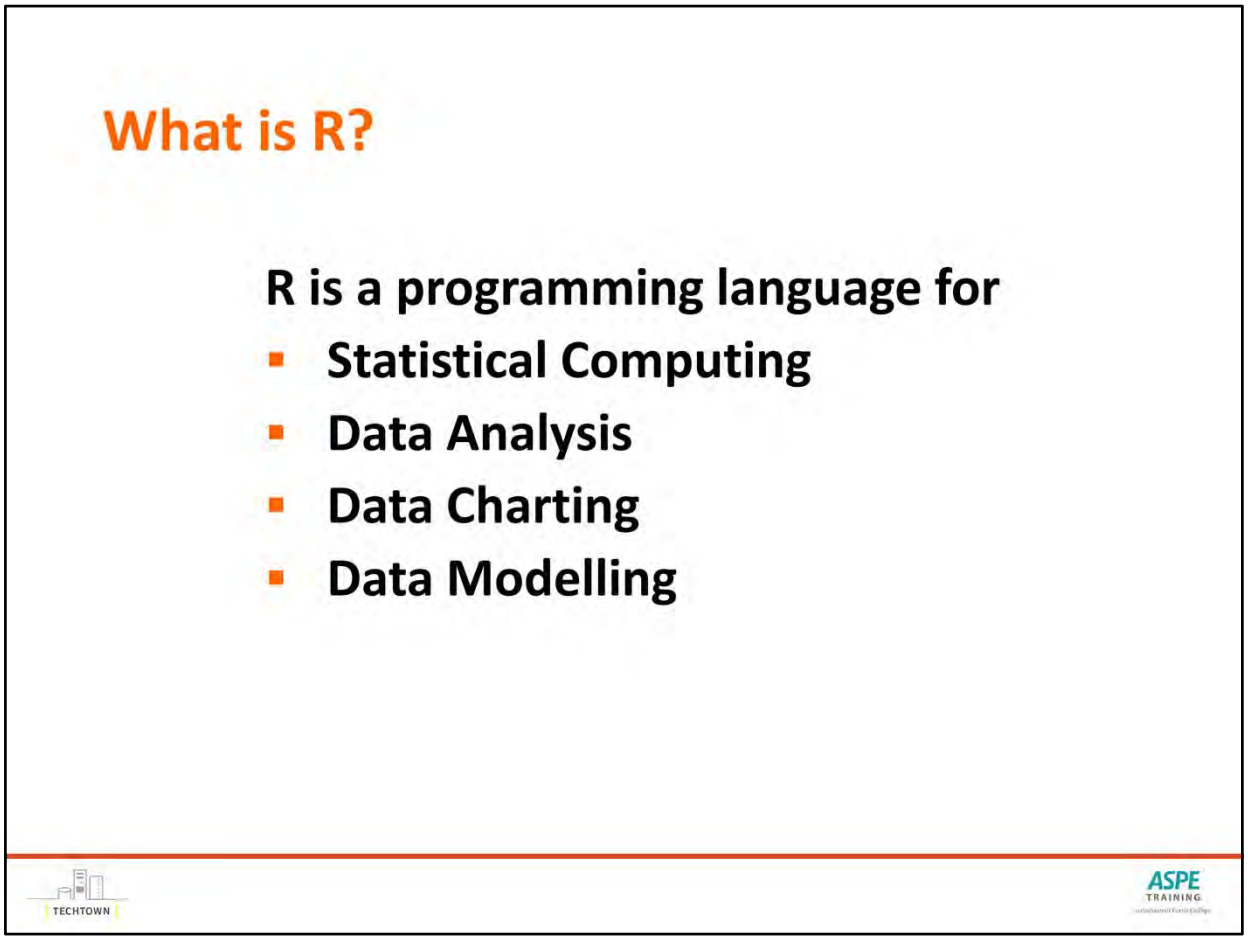
In reality, this only works when we have many more trials (thousands, if not millions).

NOTES:



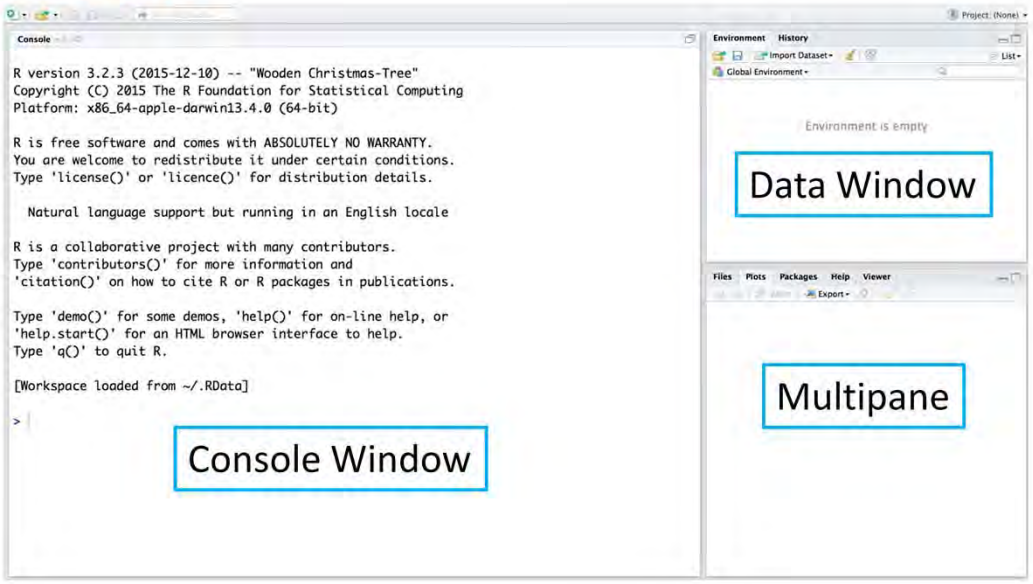
It is important to remember (cannot be stressed enough) that these numbers are not valid with 10 trials. The relative frequencies will converge as the number of trials grows very large.

Programming… this might scare you a little, but that is ok, we will not be doing much actual programming. This is more a section to motivate you to go beyond Excel. A real R programming course will take a substantial amount of time to cover.



R is open source and freely downloadable from https://www.r-project.org/
We recommend R Studio to develop your code in, it is just easier to manage. It can be found at https://www.rstudio.com/

R Studio



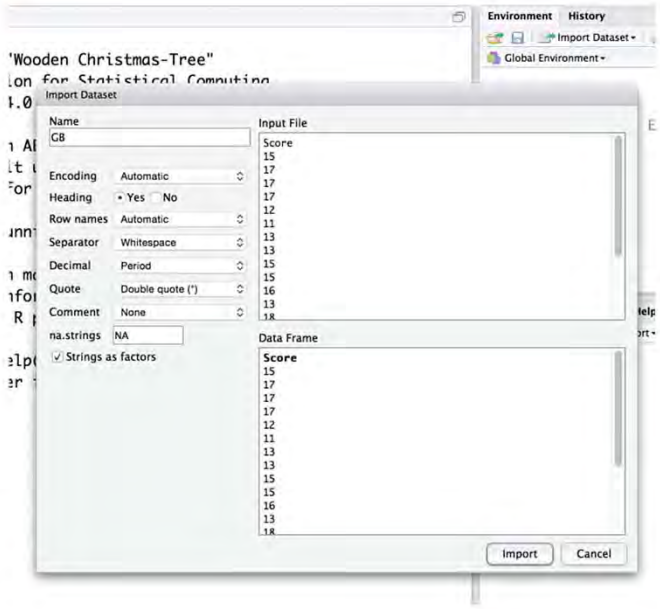
TECHTOWN

ASPE TRAINING

This is a screenshot of R Studio. It is a development environment for code where commands are executed in the console window, data is available in the data window, and the multipane window shows help, charts, and a number of other information displays.

Basic Analysis in R

- Import Grades
- Give data set a name
- Will access with heading as variable



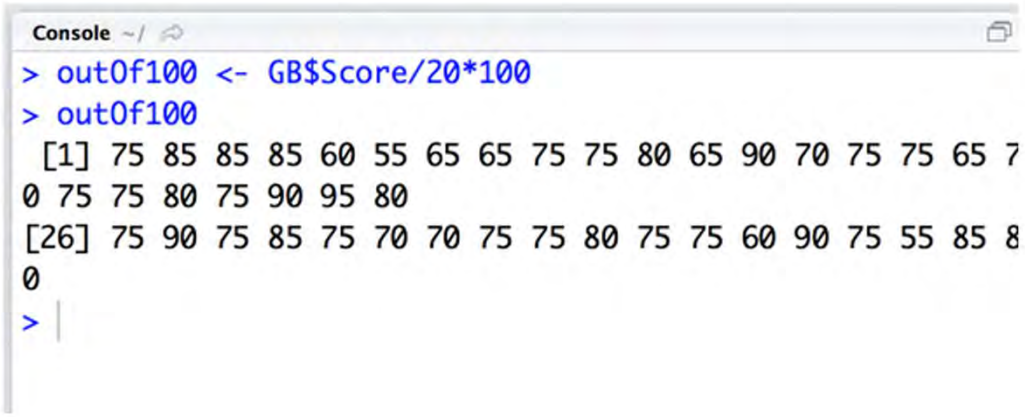
TECHTOWN

ASPE TRAINING

In R Studio, you can import data from a variety of sources – Excel files, CSV files and Tab delimited text files to name a few.
Here we are pulling in grade data from one of the CSV files in the labs.

Basic Analysis in R

Convert scores to percentages
Converts all elements in data set at once



TECHTOWN

ASPE TRAINING

This takes every row of data in the Score directory, divides each value by 20 and multiplies by 100, turning the value into a percentage. The arrow says to store the result of this calculation in a variable called outOf100.

Basic Analysis in R

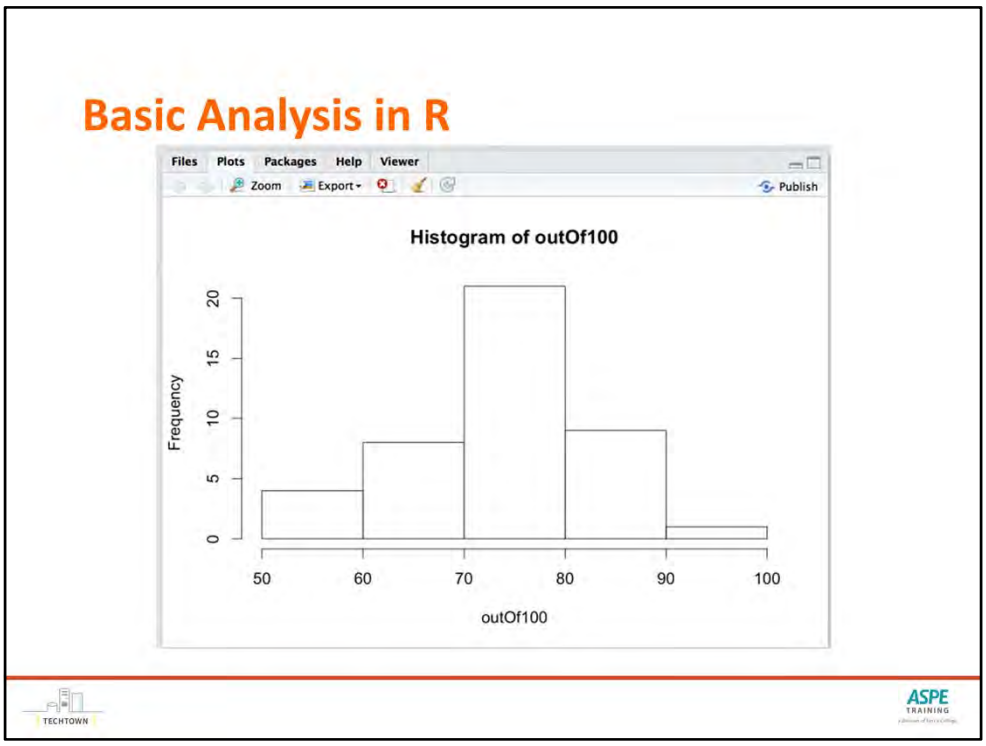
- Calculate central tendency
- Ask for help
- Plot histogram

```
Console ~/
> mean(outOf100)
[1] 75.5814
> median(outOf100)
[1] 75
> sd(outOf100)
[1] 9.335627
> ?hist
> hist(outOf100, breaks=4)
```

TECHTOWN

ASPE TRAINING

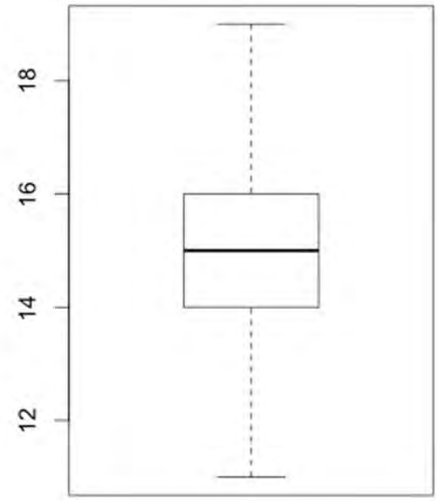
NOTES:



This is the result of the hist() command. It is basic, but R does have a lot of options for setting parameters like labels and ticks. Consult the documentation for more information.

Basic Analysis in R

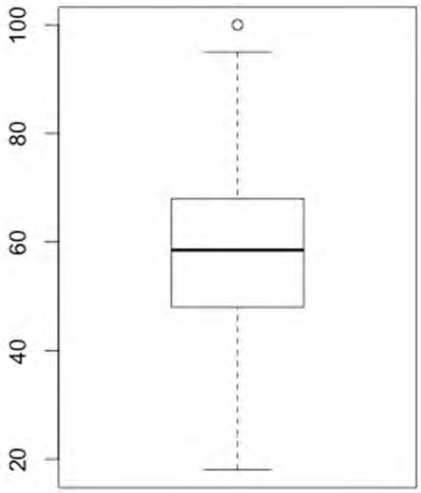
- Create a boxplot with boxplot(GB$Scores)



NOTES:

Basic Analysis in R

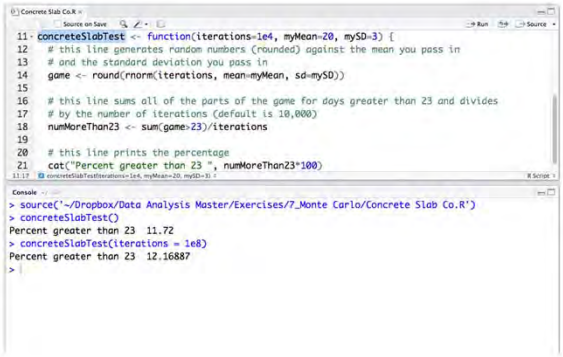
- A different set has an outlier



NOTES:

Monte Carlo in R

- Create a function
- Call the function multiple ways
- Note operations on arrays as variables



TECHTOWN

ASPE TRAINING

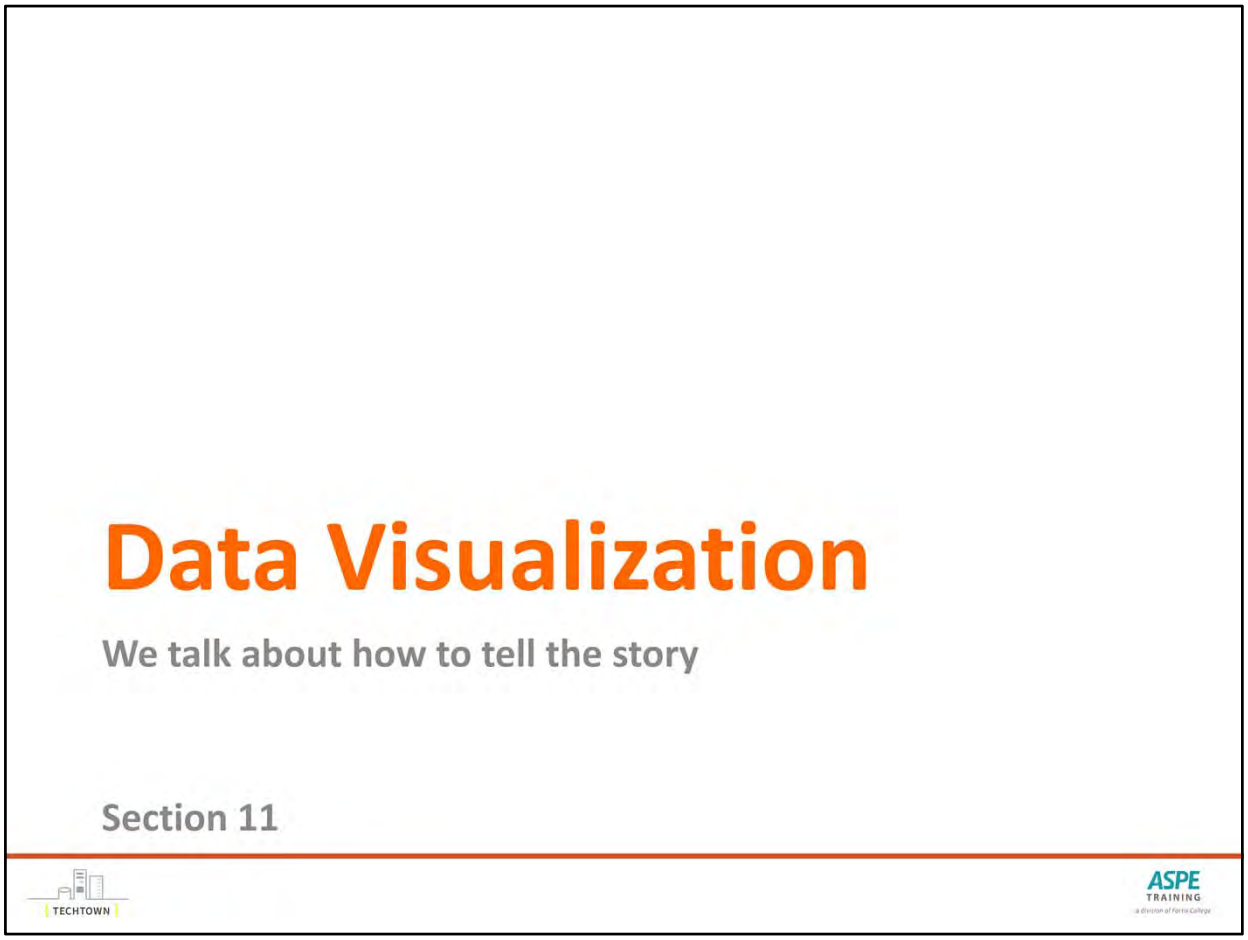
This is a shot of the console window to create a function in R. Functions are ways to create reusable pieces of code.

NOTES:

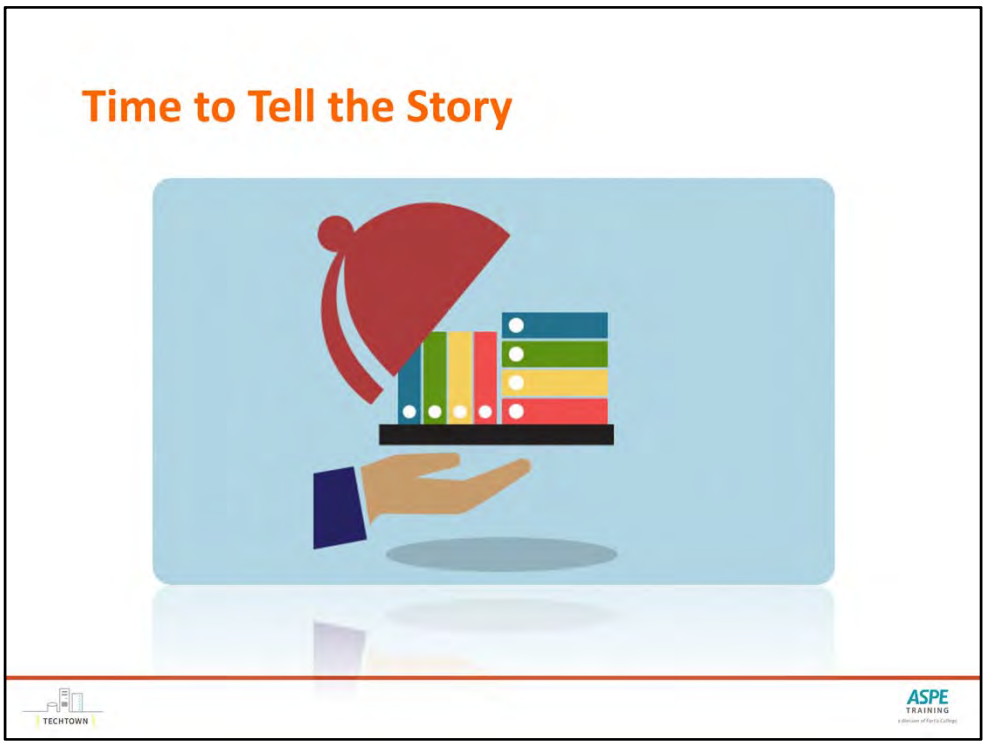
R Summary

- Open-source, well-supported tool for data analysis
- Often more efficient than Excel for repetitive tasks
- Numerous packages are available for complicated analysis techniques and data visualization

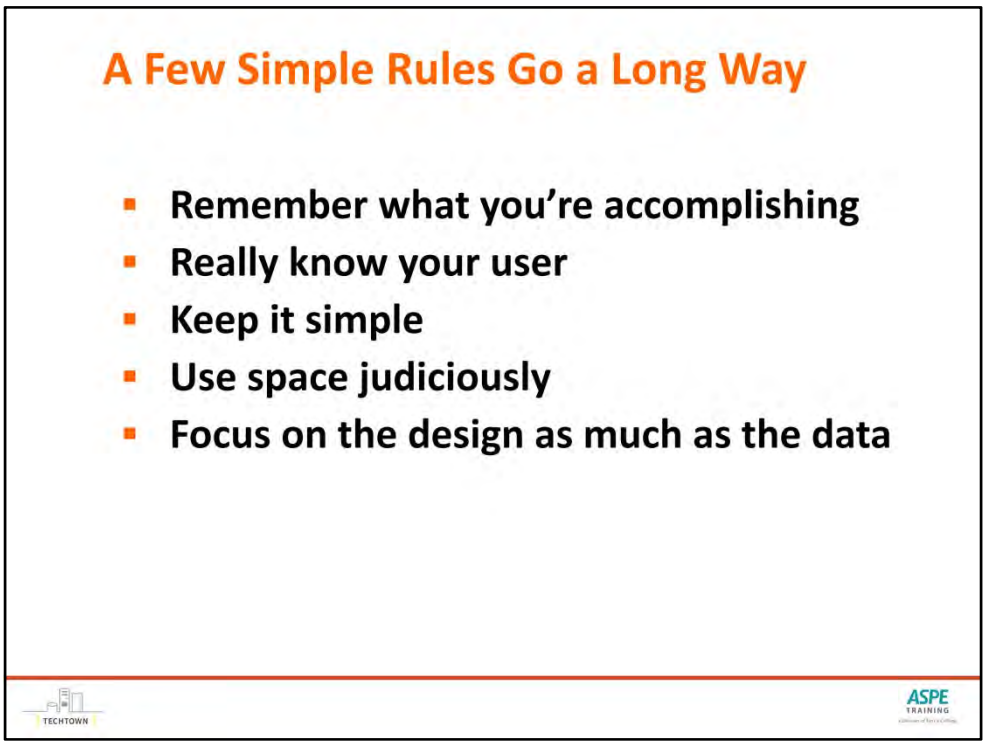
NOTES:



Hopefully, you have seen a number of ways to create data visualization by now, but to end the course lets talk about the final aspect of analysis, telling people about it.



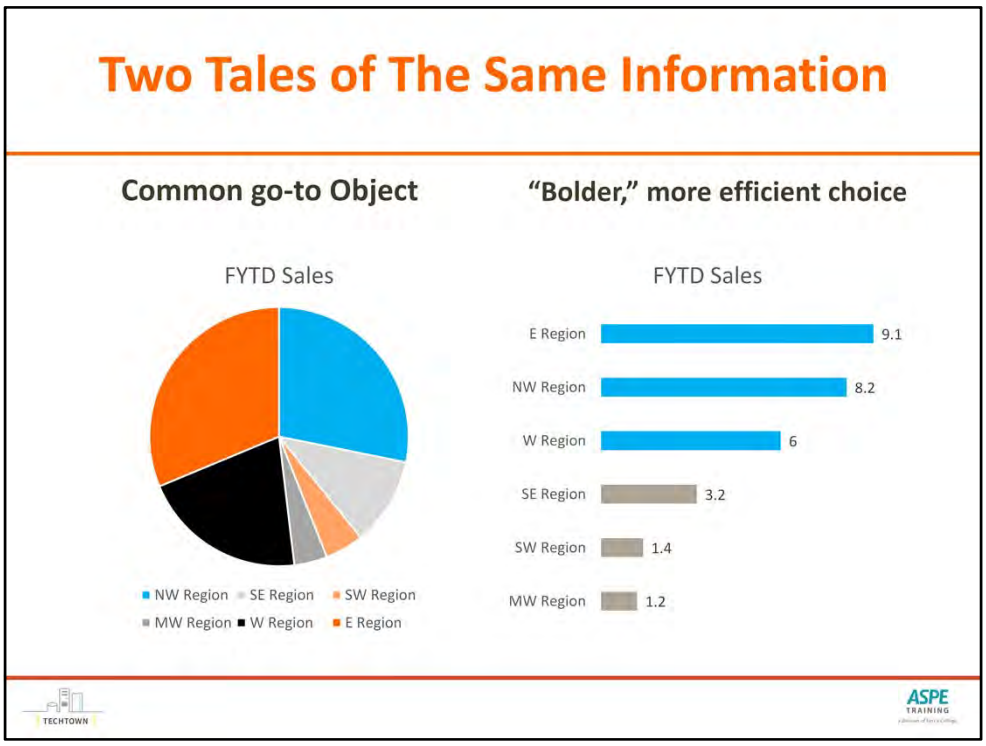
After all the hard work , you have to communicate the results. Results are not just numbers, but also conclusions. You have to tell your story from problem definition to conclusion at multiple levels depending on who is listening.



A story has a beginning, middle, end, a plot, and hopefully, a reader. Telling your story is no different.

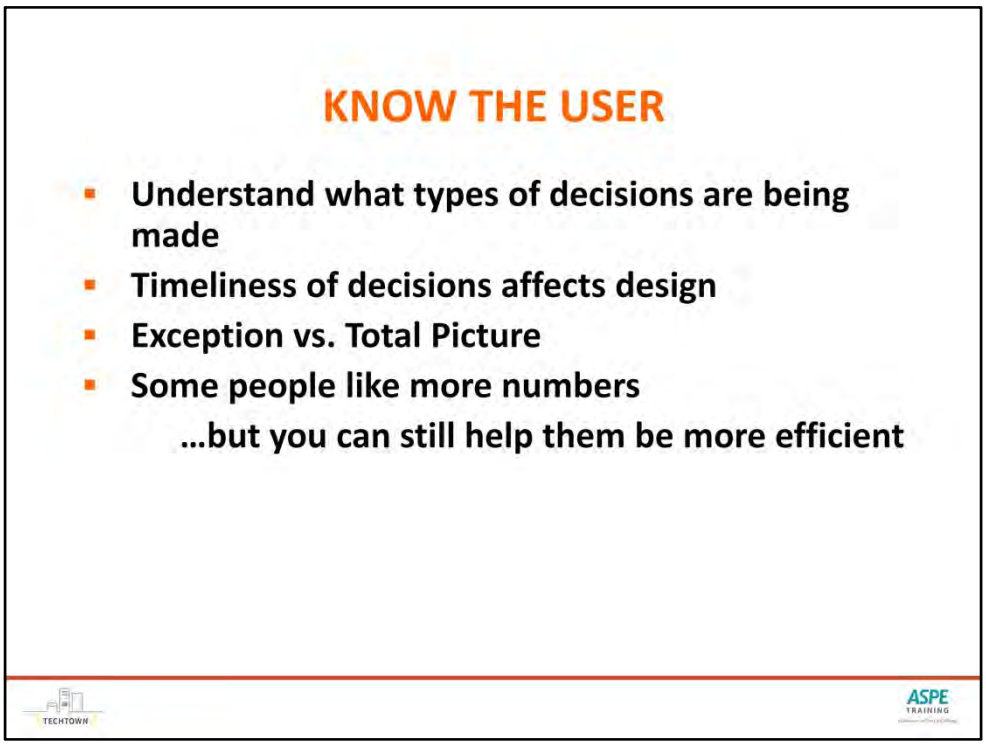
What problem were you attacking?
Why were you taking it on?
What did you do to solve it?
What assumptions, what data was used, rejected?
What was the conclusion?
And as for the reader, what level of depth do they require to ”get it”?

People see patterns very well. We create them where there is nothing a lot of the time (shapes of clouds for instance). Presenting data is leading people to the pattern and telling them the information necessary – not all the information. Charting is great for this. Tables of numbers don't work so well for most people. This is why we call is visualization.

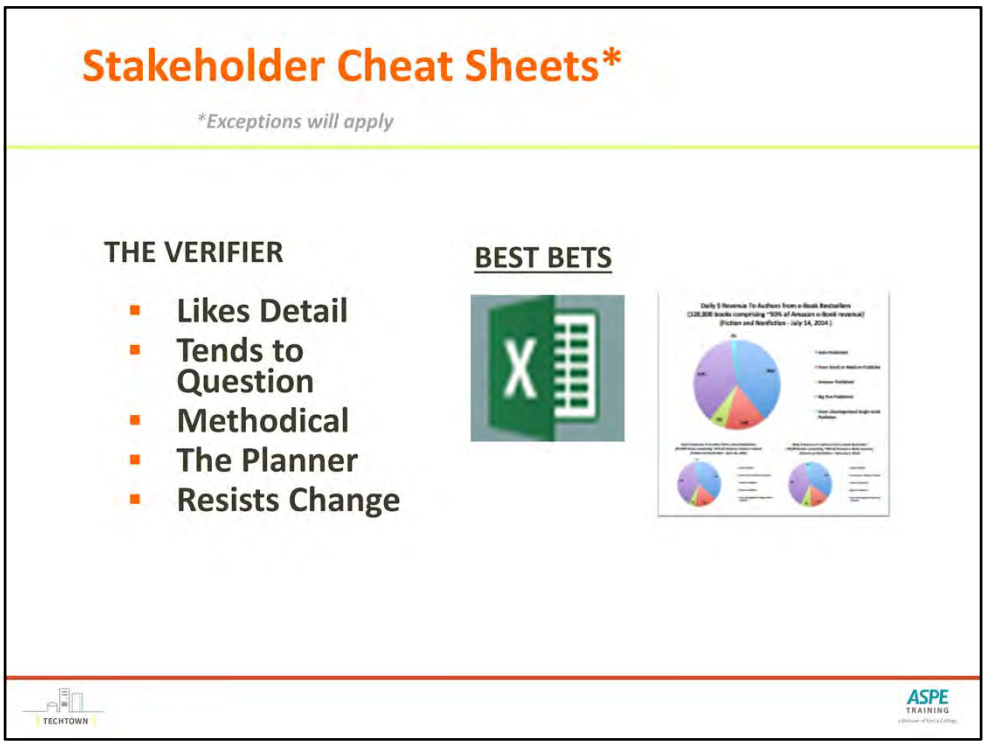


Here are two objects to display information. Which of the two looks "better"? Conveys information more effectively? Which one requires less eye movement and cognitive engagement to extract the information?

Hint: Pie charts are usually frowned upon in any scenario.



Knowing the user is critical. The project leader needs details, the CEO, not so much, they need broad strokes.



These are quick reference slides to stakeholders and what they may like to see. Think about any organization you have been in and see if you agree and can match the personality type with the layout design.

Stakeholder Cheat Sheets*

*Exceptions will apply

THE BIG THINKER

- Shuts Down with Detail
- Tends to trust
- Moves first, Plans In-Route
- Thrives on Change

BEST BET

NOTES:

Stakeholder Cheat Sheets*

*Exceptions will apply

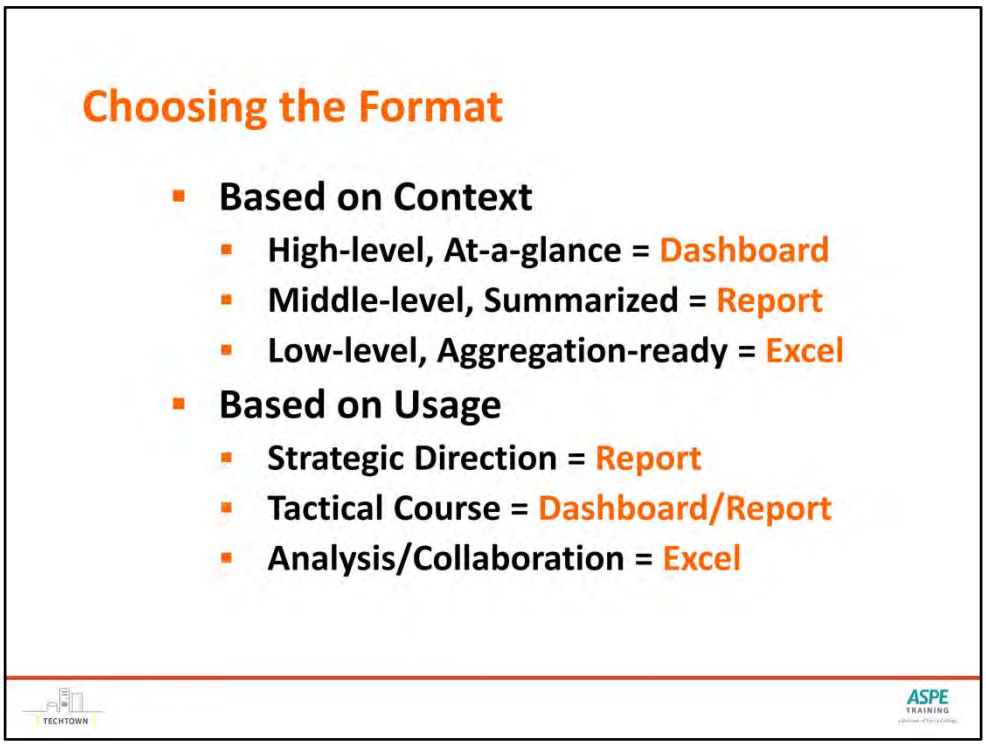
THE INTERPRETER

- Moves quickly to change
- Stops to question logic
- Diverse
- Wants enough detail to feel comfortable

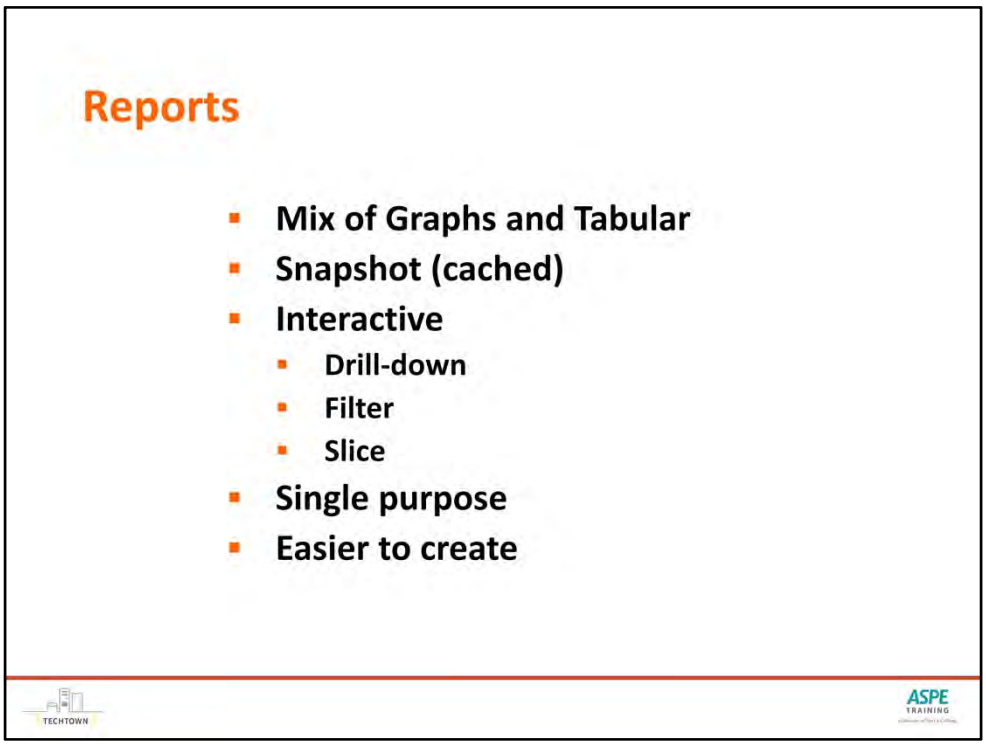
BEST BETS

NOTES:

Avoids these when performing the analysis and preparing the presentation. Don't discard data that are outliers. They are important to indicate the outliers. Don't show graphs just to show graphs, concentrate on displaying the insights. All of the presentation should serve the message.

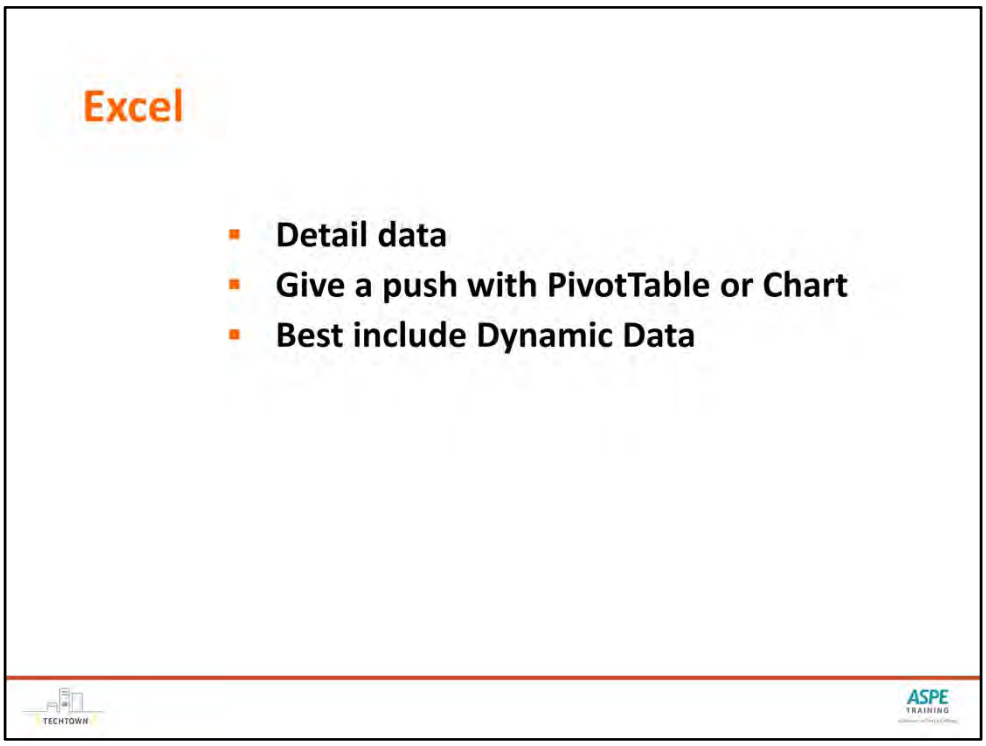


NOTES:



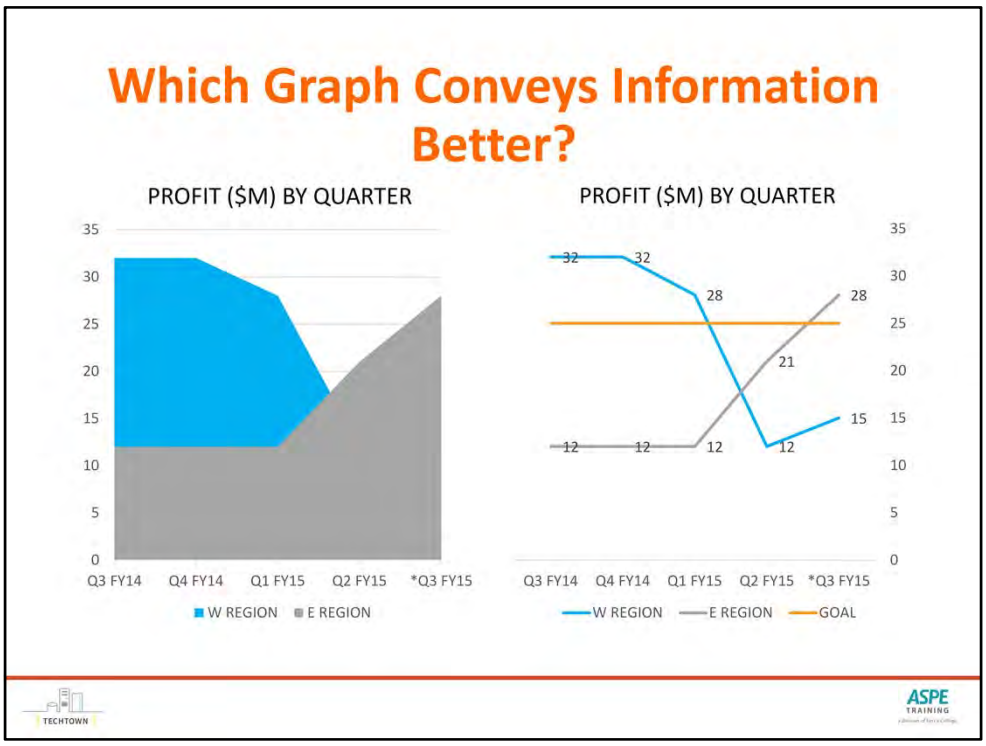
NOTES:

NOTES:

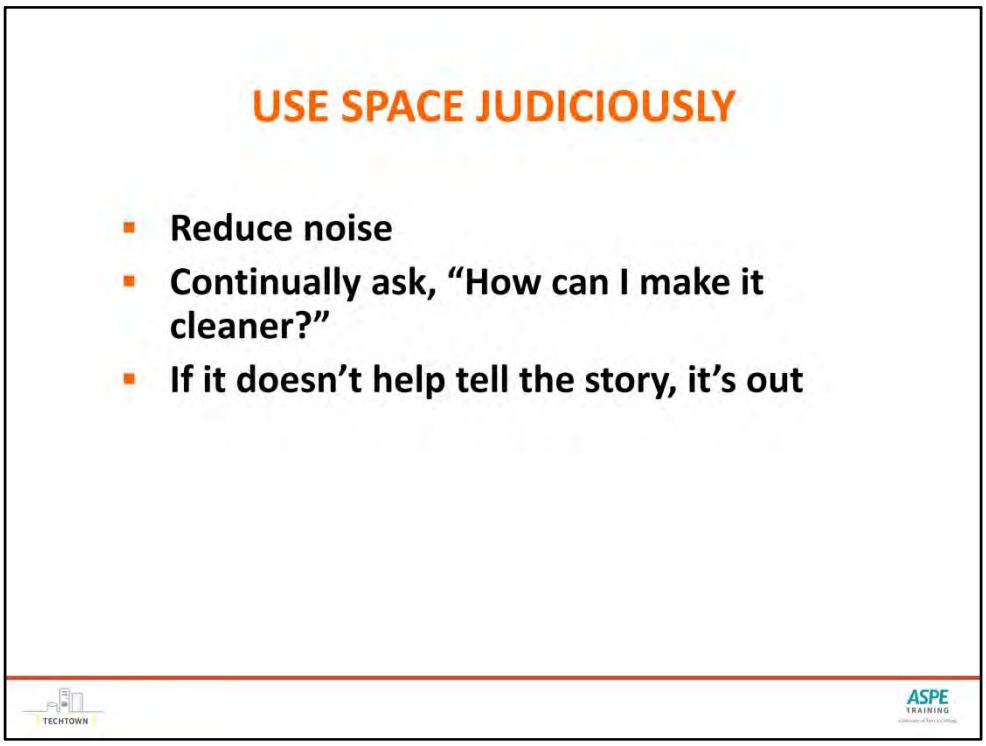


NOTES:

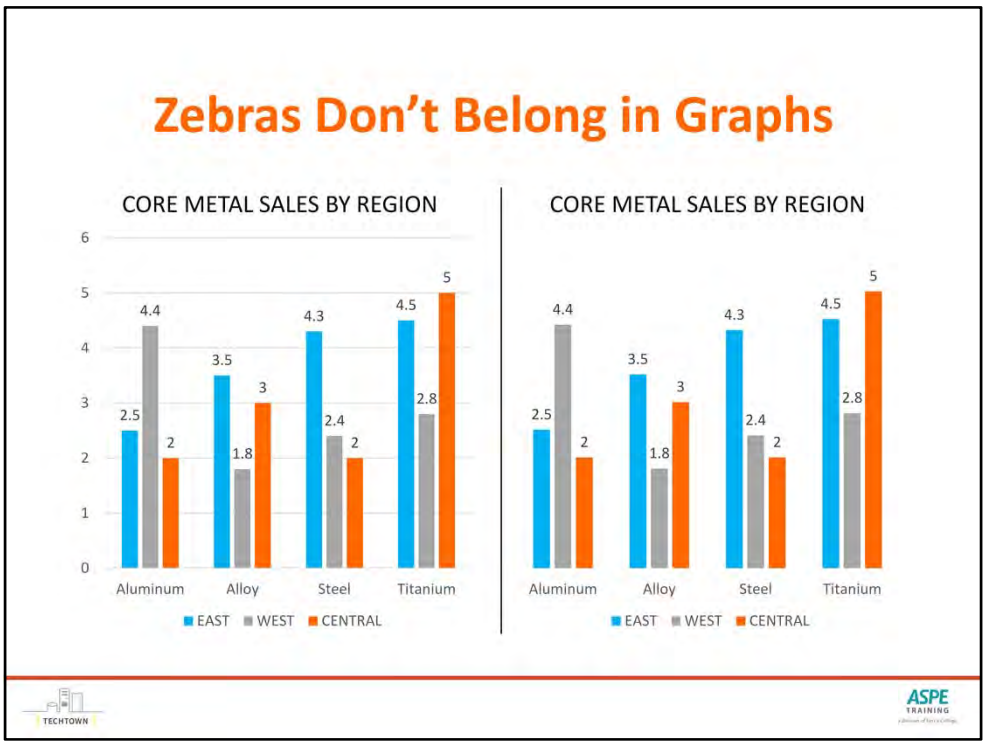
NOTES:



There is no right answer, there is a good answer for the purposes. In this case, the level of detail desired is the key.



NOTES:



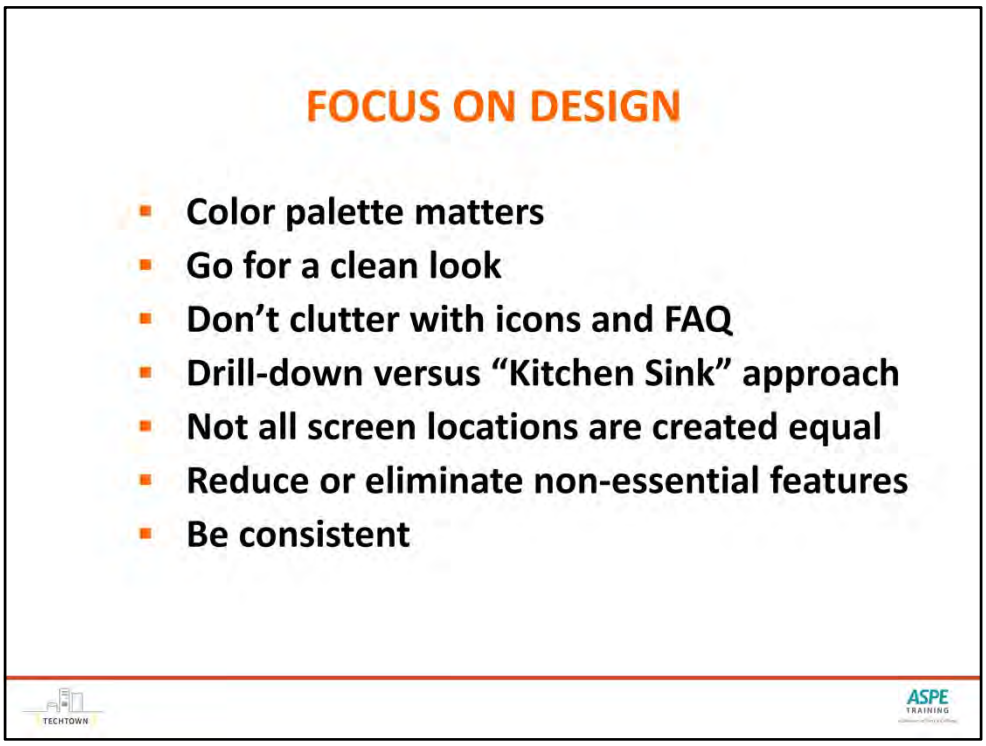
The message here is a little subtle, but horizontal lines to tell you the numbers when the numbers are on the vertical bars. Clean it up and the picture is more pleasing with less distraction.

Grid-based can be more challenging

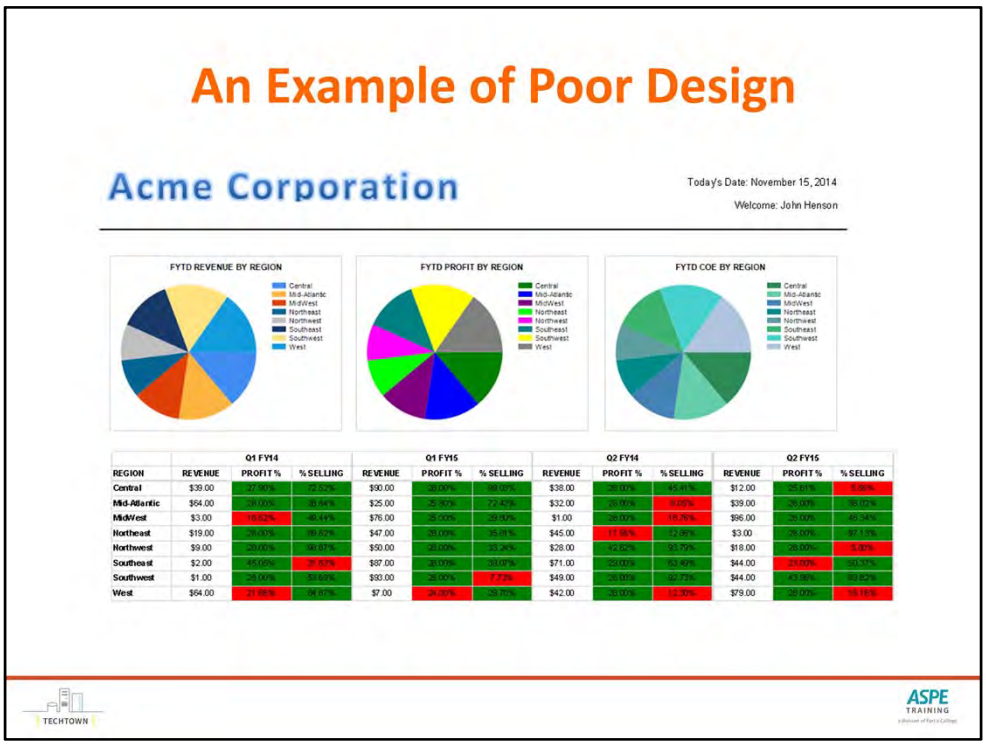
COUNTRY	FYTD Revenue	Profit Margin	Consistency of Execution	Pipeline Growth
Germany	15.0	22.5%	85.0%	25.0%
UK	14.0	18.5%	75.0%	-15.0%
Denmark	8.0	20.0%	80.0%	10.0%
Belgium	5.0	22.0%	75.0%	2.5%
France	5.0	32.0%	92.0%	18.0%
Ireland	5.0	19.0%	64.0%	-16.0%
Norway	2.0	25.0%	98.0%	-25.0%

COUNTRY	Health Indicator	FYTD Revenue	Profit Margin	Consistency of Execution	Pipeline Growth
Germany	▲	15	23%	85%	25%
UK	▼	14	19%	75%	-15%
Denmark	▲	8	20%	80%	10%
Belgium	–	5	22%	75%	3%
France	▲	5	32%	92%	18%
Ireland	▼	5	19%	64%	-16%
Norway	▼	2	25%	98%	-25%

Tables are hard to use to deliver data effectively. However, it is clear, we hope, that the multi colored cells to serve as indicators is a visual onslaught. The bottom table shows how to clean up a table for better, more pleasing display.

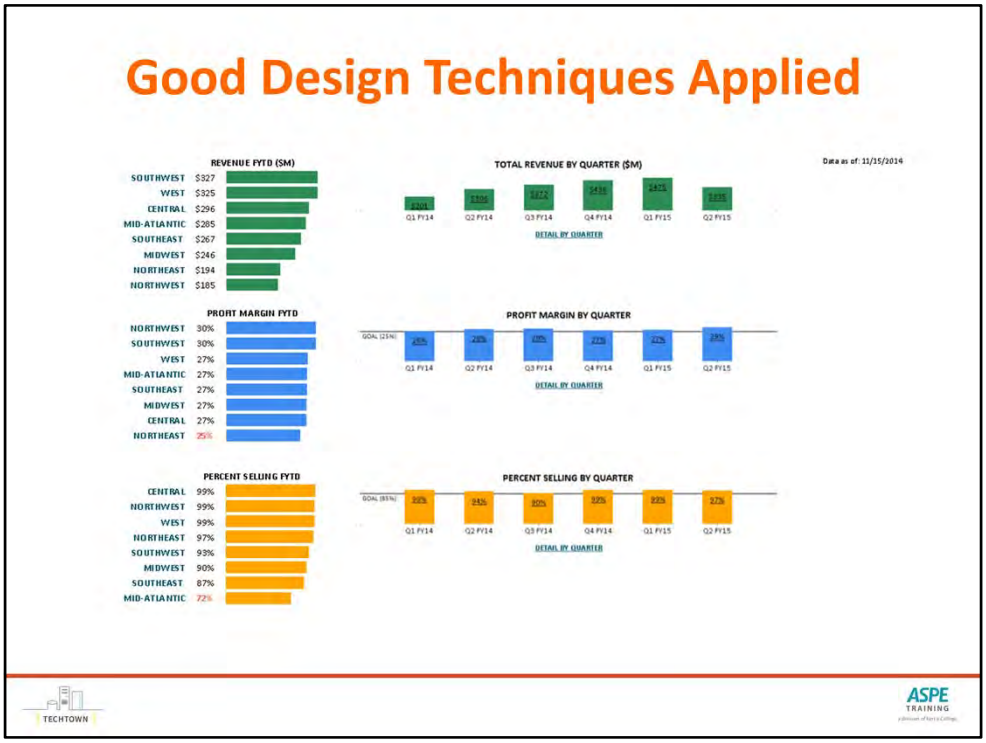


NOTES:



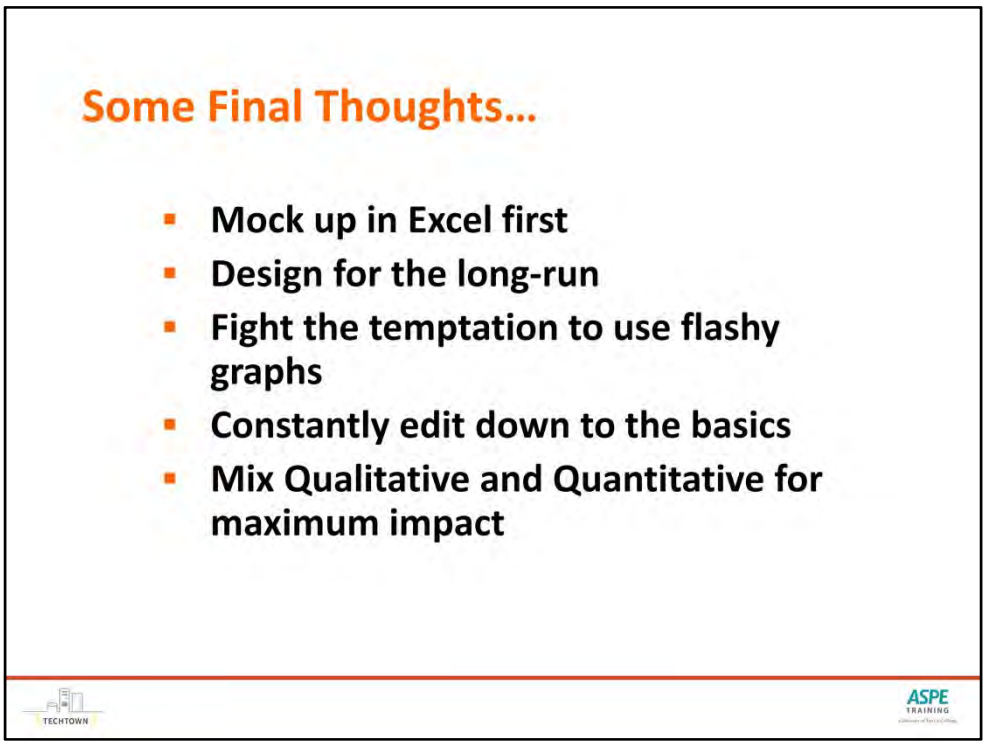
Everything here is meant to offend the eyes, it seems.

NOTES:



This is complicated but clean.

NOTES:



NOTES:

The End of the Show

We take a deep breath and relax before we go and analyze data

NOTES: