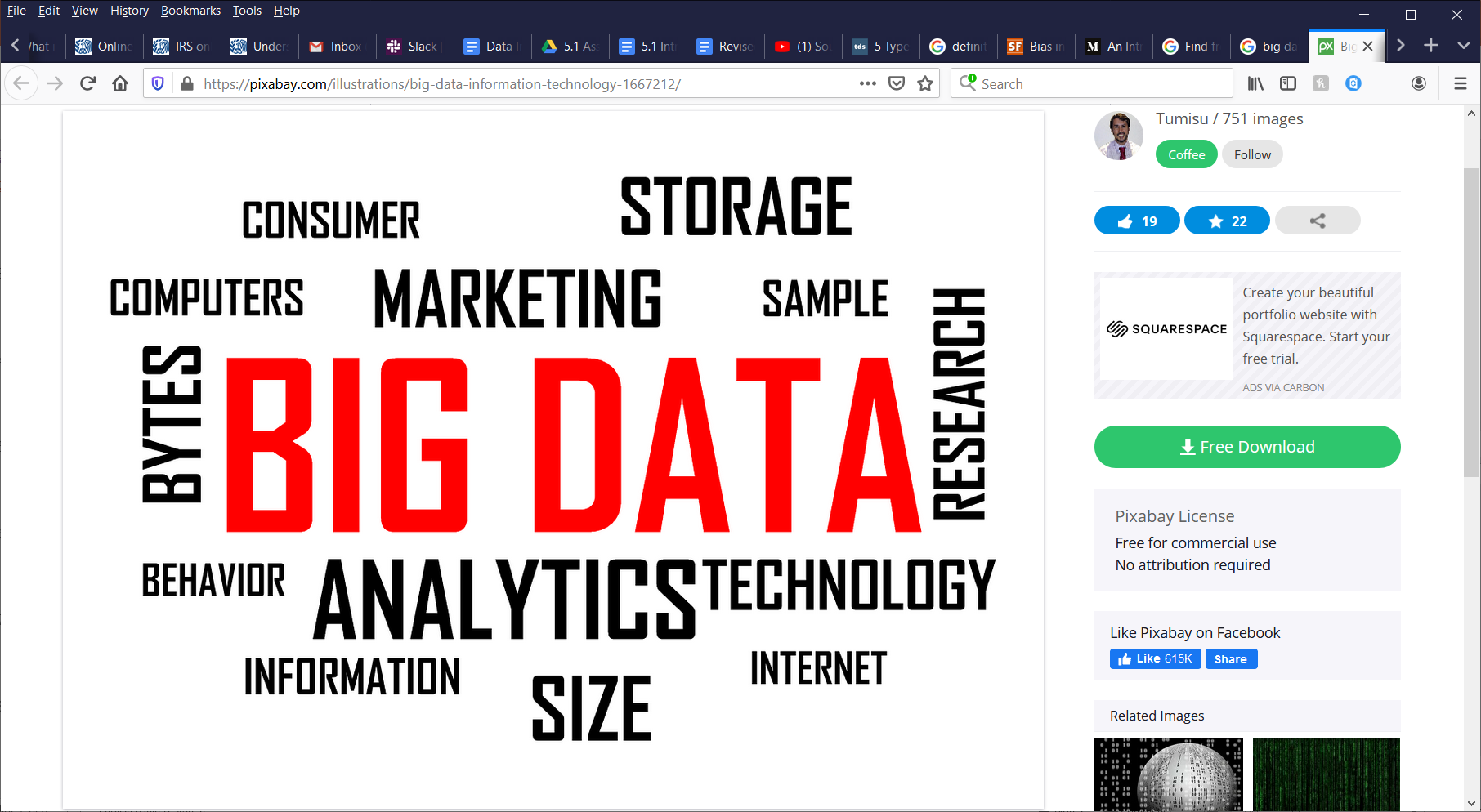
**Introduction to Big Data**

**Overall Word Count (reading and tasks) \_\_\_\_\_\_ (no more than 3000)**

Learning Goals

* **Describe big data and discuss any challenges (e.g. volume of data, structured/unstructured)**
* **Become familiar with how data is collected, stored and processed.**
* **Explain the role of a data analyst in big data**
* **Define scope of a project using big data**
* **Discuss some applications of big data**



Introduction

At this point in the course, you have learned to use some of the tools of a data analyst such as Excel and Tableau. You have developed a basic level of proficiency in SQL and Python. The tools and skills are necessary to creating datasets and visualizations. But where does data come from and how does it get there? In this achievement we will define the concept of Big Data, where it fits in the big picture of data analysis, as well as some of the challenges of discovering knowledge within Big Data.

Imagine big data as a giant well. But instead of collecting and storing water for drinking or growing crops, it collects and stores records of results, actions, and observations for the purpose of increasing our awareness of the world around us. That awareness is what gives scientists the ability to model the effects of climate changes, businesses the advantage of knowing what products they should develop and who is most likely to buy them. Both big data and a well require the use of tools in order to extract the resources they contain. But unlike a well, big data will never run dry and it’s full of data that has varying digress of utility. In fact, data is being added to the well of big data at a rate faster than we can store, process and analyze it. No matter what industry you work in as a data analyst, first it is necessary to have a basic understand how the data you analyze is collected, classified and stored as well as the challenges of applying it.

Big Data Characteristics

The term “Big Data” may appear an abstract concept to many. Up to this point you have been working with data that is structured, clean and organized in a manner making it possible to query and analyze. To someone who has never heard of Big Data, they may ask how big is it? What does it look like? What kind of data is in it and how did it get to be, well, BIG? In an effort to help address these common questions, the data analytics/science community has formulated 5 characteristics of Big Data to help us conceptualize what it is. These characteristics are known as the 5 V’s of Big Data. They are volume, velocity, variety, veracity and value.

Volume

Volume simply refers to the aggregate amount of data in regards to processing power. Whether or not data is considered Big Data or not completely depends on the volume. When you try to attach a file or files that are too big to send in an email, they technically qualify as Big Data. When the volume of a given data source exceeds the ability to process it by traditional means, it qualifies as Big Data.

Velocity

The pace at which data is being generated is what we call velocity. The velocity of data generation is quite fast and getting faster by the second. As technology is better able to track and record every action, result, and occurrence, the velocity of data generation increases. The amount of data that used to be generated over decades will soon be generated every minute!

Variety

Unlike a clean, structured dataset ready to analyze, Big Data exists in many different formats. Messages, transactions, research results, videos, photos, sensor data, model output/input are just some of the forms in which data exists. The various forms of data can be separated into 3 types. They are structured, semi-structured and unstructured. Examples of structured data is data that is stored in databases and organized in tables, rows, columns and ready to be analyzed. Unstructured data refers to data with no consistent format. It consists of emails, texts, videos and photos and makes up the majority of Big Data. Think of unstructured data as a messy heap of data that isn’t organized and easily analyzed; basically, the opposite of structured. Lastly, we have semi-structured data. Think of semi-structured data that has certain common characteristics, but is not organized within a database so that it can be easily queried and analyzed like structured data.

Veracity

Veracity describes the extent to which data is reliable in terms of accuracy and trustworthiness. Never assume that just because you are working with data, it can 100% be relied upon. Always remember that it is humans and machines designed by humans that created big data. Humans can be biased, deceptive, at worst dishonest. Sometimes machines malfunction or are not programmed correctly. These human flaws and machine errors can generate data that when analyzed will lead to false conclusions.

Value

The 5th and final V of Big Data, and arguably the most important, is value. Collecting and storing all of this data is very much pointless unless we can discover value within it. We do that by executing sound analytical techniques that have applications to real world issues. Discovering the value in data leads to everything from cures for diseases, weather predictions and increased awareness of our social and economic environments.

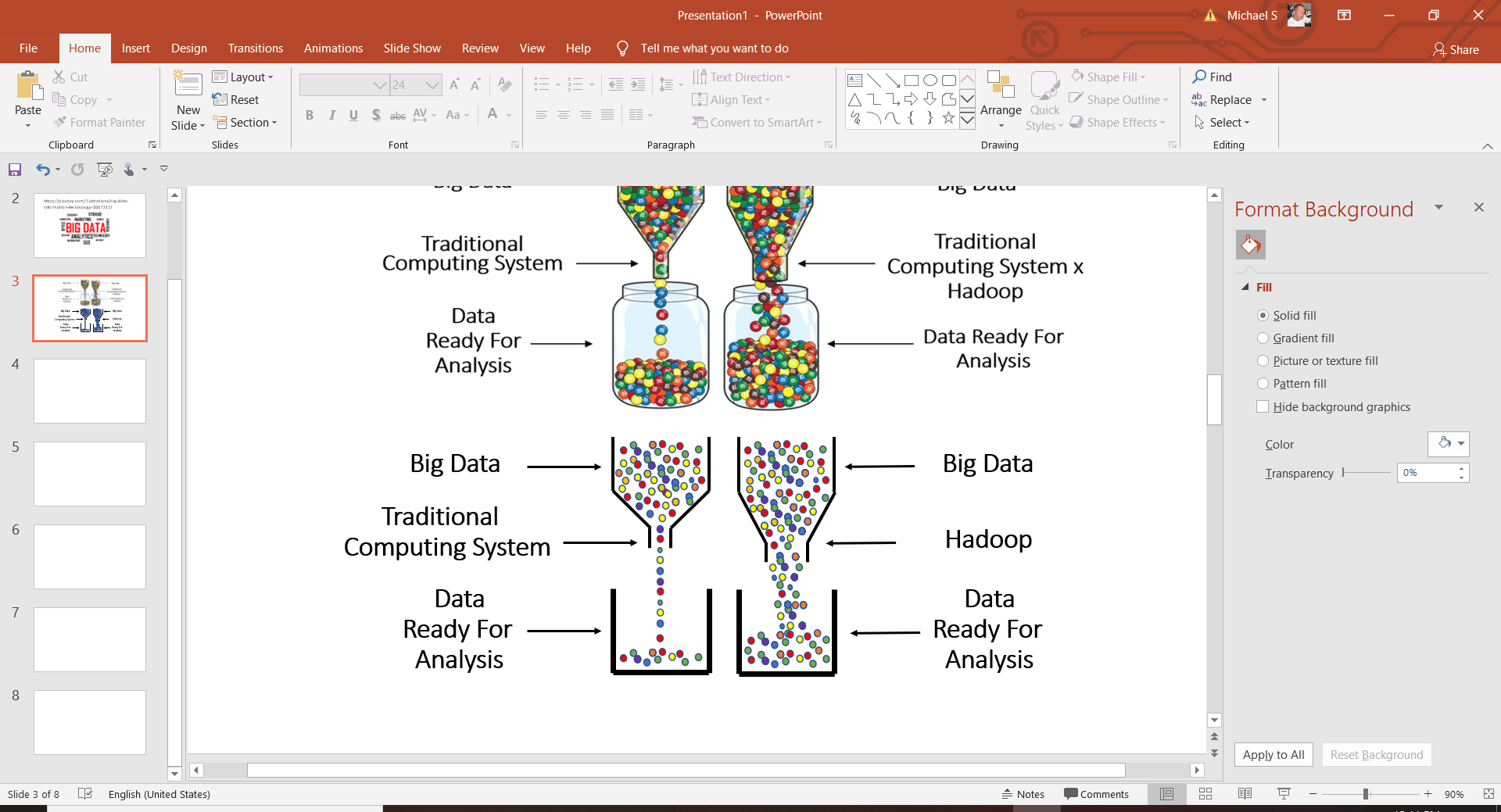
Big Data Lifecycle

Collection

By now we know that data exists in digital media, relational databases, the cloud, and software frameworks such as Hadoop and Apache Spark. But how does it come to be in those places? While some data is collected manually such as when a company simply asks a customer questions, most data collection is automated. Every time you swipe your debit card, the details of the transaction such as what you bought, where you bought it, the amount of the sale, location of the sale, time of day and many others are recorded into one or more of the entities that store data. If you are like most people, you carry your smart phone with you everywhere you go. Chances are, one or more of the apps on your phone is collecting data about you 24 hours a day, 7 days a week. So while you are riding that bus on your way to work, watching videos on youtube.com, liking articles on Facebook News, and browsing online shopping websites, that activity is collected by companies that seek to know what, where, and when to market products to you. A violation of privacy you say? Many of us don’t realize we gave permission for this data to be collected when we downloaded these apps and established online accounts with retailers. We will explore these issues and more later in the Data Ethics lesson. For now, let’s focus on how observations, results, and actions become big data.

Storage

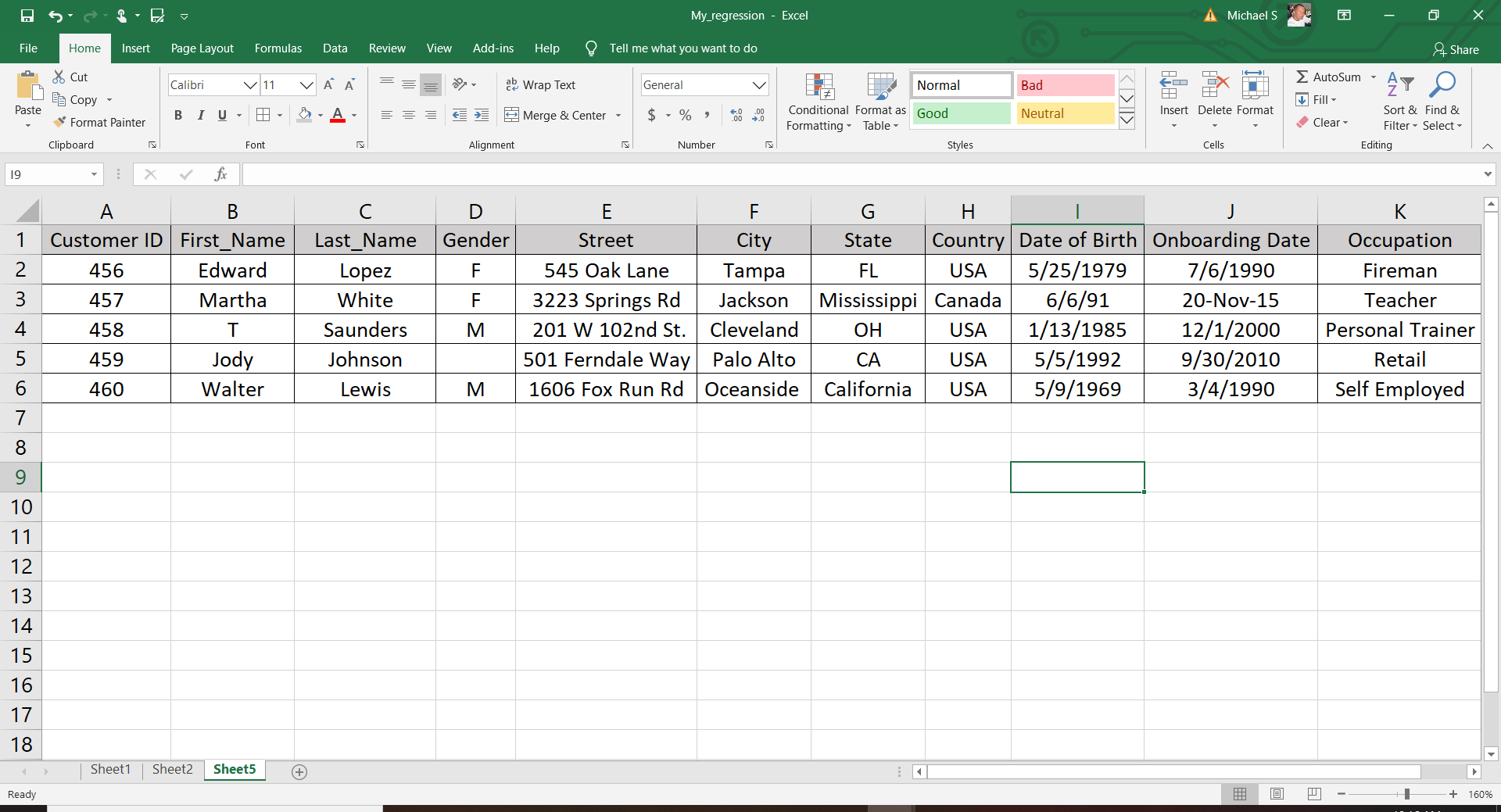
Earlier we spoke of where big data exists and alluded to it’s ever growing size. In 2018, Forbes reported that 2.5 quintillion bytes of data created every day and the pace is accelerating. It would be great if all of big data were organized and centrally located so all you had to do was execute your SQL query and instantaneously get a dataset so you could begin to analyze it. Unfortunately, it’s not that simple. The traditional computing system cannot store and process this immense volume of data. Hence the name “Big Data.” In business terms, this is what is known as a bottleneck.



Hadoop is one of the more popular applications that helps widen that metaphorical bottleneck that exists between Bid Data and analysis. It isn’t necessary that you become an expert in the architecture and inner workings of Hadoop. The key takeaway for you to know as a data analyst is that Hadoop and similar applications are software frameworks used for storing and processing Big Data so that it can be more readily accessible for data analysis.

Analysis

Now that the data has been collected and stored so that you can retrieve it, the lifecycle of data moves into the analysis phase. This is the phase where the data analyst can begin to use various statistical techniques to discover trends, patterns, and gain insights from the data that guide business decision making and policy. But before you perform that cluster analysis or regression, the data must be cleansed to insure quality and integrity. As a junior data analyst, you are probably going to be spending a lot of your time analyzing data for quality and integrity. Data cleansing refers to a process of finding and correcting data that is inaccurate, missing, improperly formatted or incomplete. Observe the example table below containing your company’s client information within a database.



As you can see, for Customer\_ID 459, the gender information is missing. The record for Customer\_ID 457 has Martha White residing in the state of Mississippi, but the country is listed as Canada. Notice the inconsistency in the format of how dates are written, as well as the States where customers reside? These are the types of errors you will find that will throw off the accuracy of the results of your hard work if not cleansed. Suppose you wanted to isolate customers living in California and assess their risk profiles against those residing in other states. Your SQL query will contain “WHERE State = CA” and you will have unknowingly skewed the results of your final analysis because customer records where State is spelled out would have not have been included in your data set. Profit and cost margins are small in the business world. Analyzing dirty data can lead to poor decision making and cost you professional credibility. Quality analysis begins with cleaning data and ends with results in which you can feel confident. The next steps of analysis are to model and visualize your dataset so that you may make predictions and forecasts. These steps will be covered in more detail in later achievements. The primary objective of this achievement is to guide you through the lifecycle of data as it goes from being collected to guiding business decisions.

Implementation

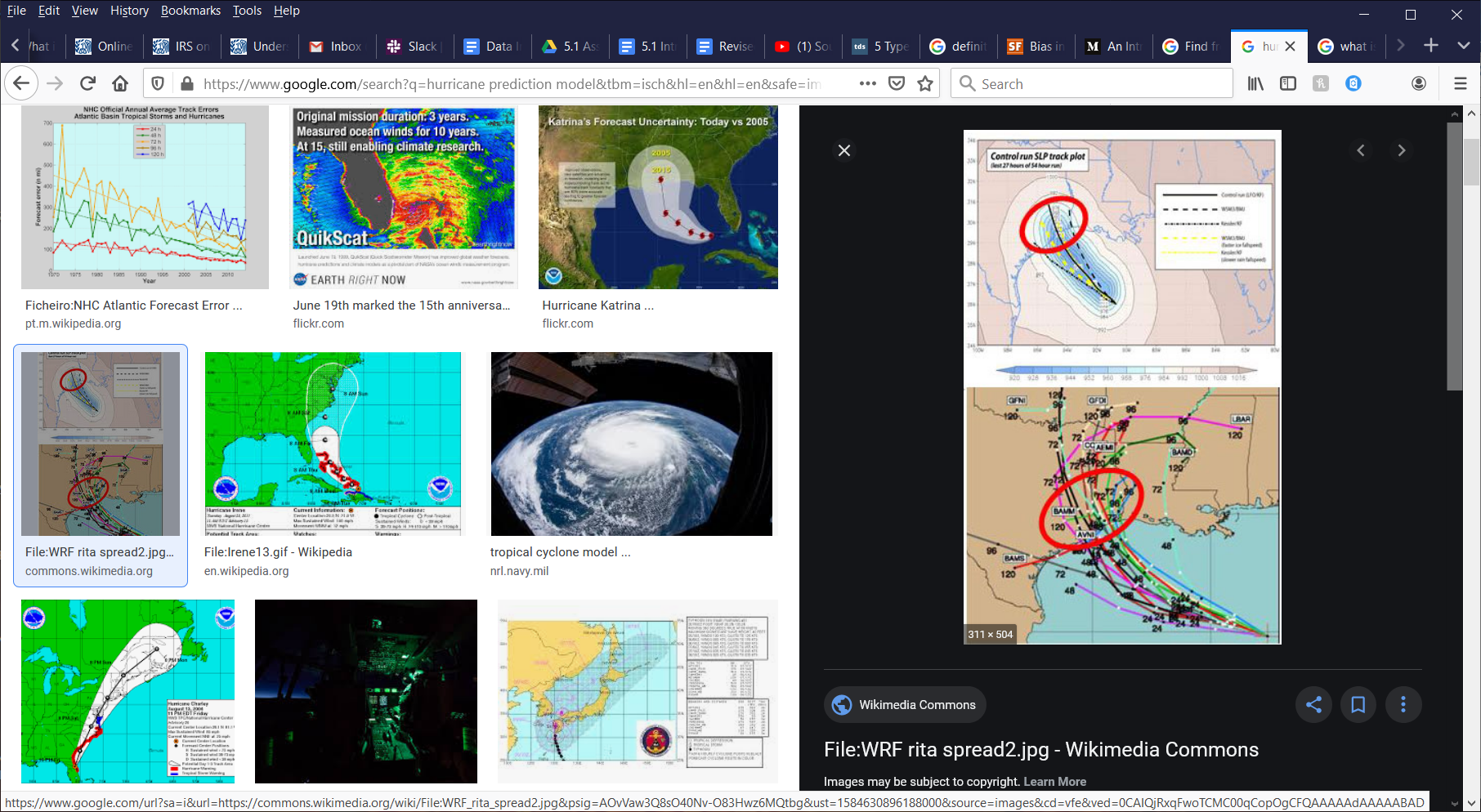
The final stage of the data life cycle occurs when your employer decides to put into operation a prototype model you have built based on your analysis of data. Imagine the global financial services company you work for tells your team to develop a model to be implemented that must predict if Remote Deposit Capture transaction characteristics are likely to be suspected of money laundering. A Remote Deposit Capture transaction is a relatively new transaction type that involves the client taking a digital photo of the front and back of a check and transmitting it to the bank for deposit. The scope of your project would see ou build multiple models that consider different quantitative variables such as the amounts of the transaction activity, frequency of deposits, and how long the funds remain in the account. Your models also consider qualitative information about the client such as if they are an individual or business, geographic location, age of the account, and customer segment. In order to decide which model to implement, they must be run in a test environment. The output of each model must be analyzed by subject matter experts who will determine if the output is A) consistent with the model logic and thresholds, and B) if the transaction activity observed is actually suspicious. In theory, the prototype model that is most efficient at predicting suspicious activity will be implemented into the operational environment. This is process is known as independent model validation and is an imperative part of implementation. After all, anyone can build a model that generates output. But validation tells us if the model works like it is supposed to and works efficiently.

Big Data Applications

The applications of Big Data are everywhere and more of them are being discovered all the time. In many ways, a data analyst is an explorer. They are forever searching for new ways to find and apply the knowledge within Big Data to solve problems, enhance performance, and save time, money and resources. Now let’s talk about some examples of real world Big Data applications.

Weather Prediction

Weather events such as hurricanes can have devastating consequences on people and property. Big Data can help us minimize these consequences by predicting where they will hit and how hard so we can better prepare. The image below is a snapshot of a Hurricane’s location in the Gulf of Mexico and several possible paths it could take as it moves northwest. These models are designed to predict the most likely places the hurricane will hit using data collected from past hurricanes that document how they react to meteorological variables such as air temperature, water temperature, humidity, and barometric pressure. As the hurricane moves, sensors are collecting data and feeding it into these models updating them to reflect the paths it is most likely to take. As we observe more hurricanes and collect more data, we can build models even better at predicting their paths, resulting in saved lives and faster emergency response times.

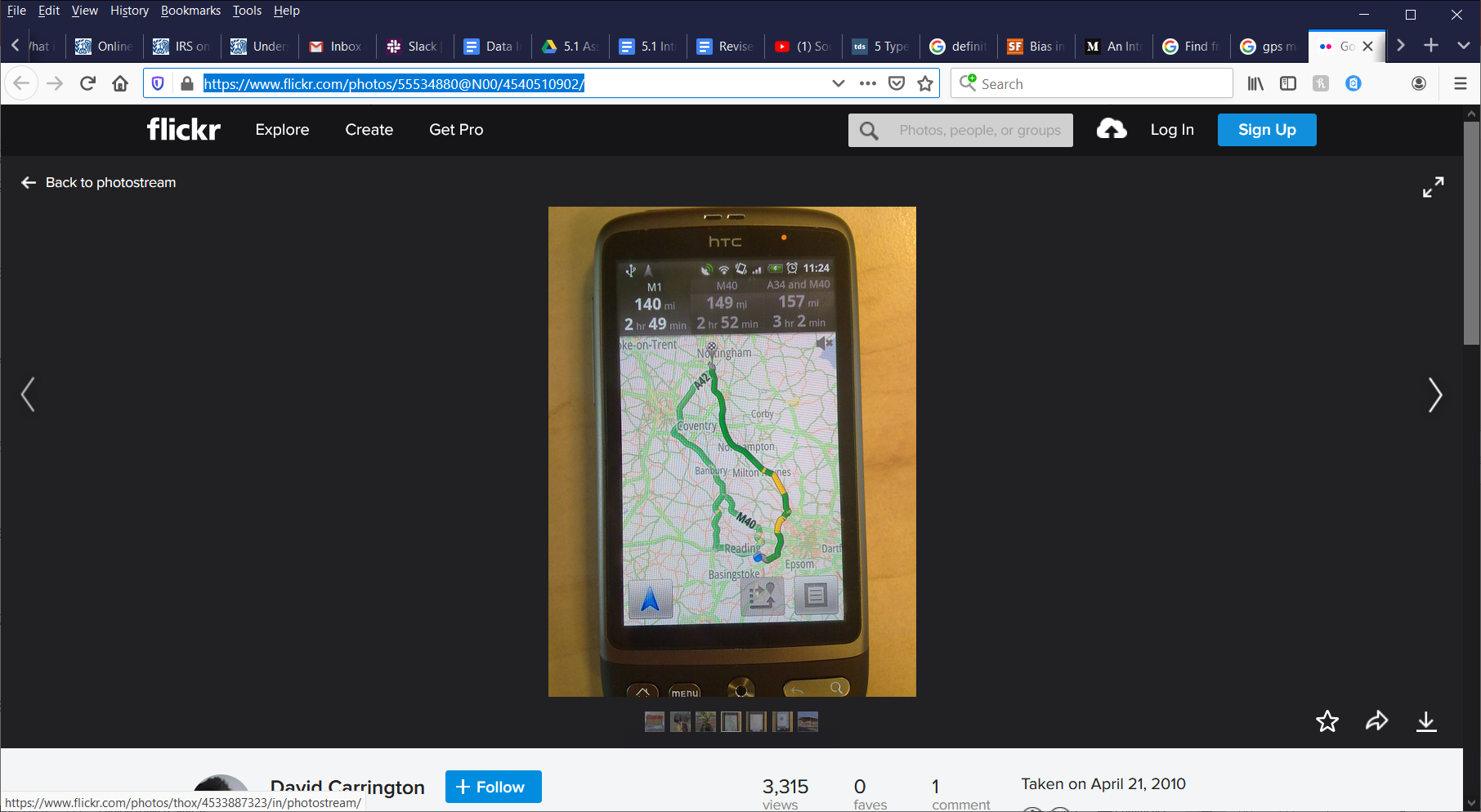


Sports

The existence of Big Data has allowed us to be able to analyze and evaluate athletes and sports that was previously impossible. Various team and player statistical production and attributes can now be stored, queried and analyzed so organizations can better formulate strategies to give themselves a competitive edge. Moneyball is a 2011 film starring Brad Pitt and Jonah Hill and is based on the true story of the 2002 season of Major League Baseball’s Oakland A’s. Tasked with putting together a competitive team on a limited budget, General Manager Billy Beane decided to reject the traditional “inside the box” approach to scouting baseball talent that placed far too much emphasis on a player’s attitude, personality and physique. Instead, he hired data analyst Paulo DePodesta to build a data-driven approach that cut through the human bias factor of talent evaluation and focused exclusively on a player’s statistics and undervalued metrics of production. The A’s began the season with a roster full of nobodies and fans had little hope for the season. But after a rocky start, the A’s had one of their most successful seasons in years which included a then modern era record 20 game winning streak. They revolutionized the approach to scouting baseball talent and now every organization employs an analytics department.

Logistics

The business operation of storing and transporting supplies to consumers is costly and if not done efficiently can significantly cut into a business’s profit margin. Since the invention of Global Positioning Systems (GPS), sensors have collected data on how efficiently our vehicles move from point A to B along our network of roadways. Analysis of this data shows the shortest distance between two points is not necessarily going to be the fastest. Using this data on weather, traffic patterns, time of day and stop lights, our GPS applications now can recommend the fastest routes. This results in less time on the road for drivers, reduced fuel consumption and cost savings.



Summary

By this time, you should have a fundamental understanding of what Big Data is, where it comes from, what it looks like and how a data analyst applies it. Big Data is to a data analyst what steel is to the construction industry. It is the reservoir of raw material that must be collected and processed before structures can be built with it. As you advance in the course and build on your knowledge, it will become more clear why you as a data analyst need to understand the process data goes through from collection and culminating into a nice clean data set ready to be analyzed.

Resources

<https://www.researchgate.net/figure/The-5V-of-Big-Data-Characteristics_fig1_321050765>

<https://paulshepherd.co/innovative-ways-collect-big-data/>

<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#51e2e72560ba>

<https://dsportmag.com/the-tech/education/engine-tech-air-flow-bottlenecks/>

<https://www.tutorialspoint.com/big_data_analytics/big_data_analytics_lifecycle.htm>

<https://www.business2community.com/big-data/what-is-database-cleansing-and-why-is-it-important-02257014>

<https://www.geeksforgeeks.org/5-vs-of-big-data/>

<http://bigdata.black/featured/what-is-big-data/>

<https://lexleader.net/data-power-highlighted-program/>