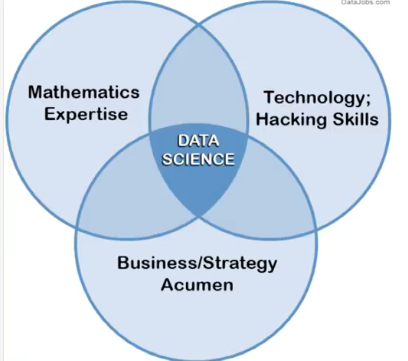
*Getting Value out of Data*

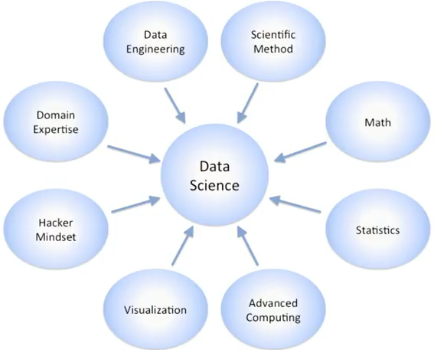
* **Data science** turns data into insights or even actions, but what does that really mean?
* Think of it asbasis for empirical research where data is used to inform hypotheses + provide observations.
* In many cases, this data is used either by businesses or scientists to inform their understanding of a phenomenon.
* B/c there are often large troves of data which we can mine for insights, we often call this **big data**.
* **Insight** refers to the **data product** of data science + is extracted from a diverse amount of data through a combination of exploratory data analysis + modeling.
* The questions we ask are sometimes quite specific + sometimes it takes looking at the data + patterns in it to come up w/ a specific question.
* Another important point to recognize is that data science is *not* a static, one-time analysis.
* It involves a process where the models we generate lead to insights + those insights are then improved by gathering further empirical evidence, or simply, data.
* Ex: Amazon can constantly improve the model of a customer's book preferences using customer demographics, their previous purchases + prior book reviews by the customer.
* The models also likely take into account the similarity of customers to detect common interests.
* They can also use this info to predict which customers are likely to like a new book + take action to market the book to those customers.
* This is where we see insights being turned into action.
* Using data science + analysis of the past (historical) + current (near-real time) info, *data science generates actions* = generation of actionable info for the future = a **prediction**
* When you decide what to wear based on the weather forecast for the day, you're taking action based on insight delivered to you.
* Similarly, business leaders + decision makers take action based on evidence provided by their data science teams.
* B/c companies take action based on their insights, data science teams need to be experts in their practice to ensure those insights are well-reasoned.
* Data science has been around for a very long time + have always used data to gain insight based on observations
* Data science is suddenly on the rise due to 2 things.
* Our **ability to collect data** **in real-time** has ballooned, w/ data coming from a variety of places, including real-time environmental sensors, websites, smart phones, + a variety of other sources.
* In turn, this influx of data has increased demand for **large scale data processing**.
* Data growth combined w/ the advances in storage, networking, + computing at scale has brought us to a new era of data science.
* Many dynamic data-driven applications in this new era build upon data-driven predictions to support decisions, like Amazon book predictions.
* It is nearly impossible to find an industry, scientific discipline, or engineering endeavor today that is not impacted by data science.
* One need only look at the major trends in smart cities, precision medicine, energy management, + smart manufacturing to see how it is shaping our economy today, + all these fields are looking for experts in a combination of advanced data analytics, traditional modeling, + simulations.
* Collected data can include anything from user preferences + purchasing history on websites, to scientific data from remote sensors + instruments + personal health data from variable devices, to social media data related to customer satisfaction, political trends, health epidemics, law enforcements + terrorists activities, as well as medical data from drug trials, treatment options, + patient population.
* Every minute, 204M emails are sent, and 200k photos are uploaded + 1.8M likes are generated on Facebook, and on YouTube, 2.78M videos are viewed + 72 hours of video are uploaded.
* And it is not any different for scientific data.
* HPWREN, the High Performance Wireless Research + Education Network which only connects sensors in San Diego, Riverside, + Imperial Counties, collect 30 TB of data annually.
* We use HPWREN data collected from weather stations throughout San Diego County for wildfire monitoring + modeling
* This consists of a daily amount of 1/2 GB of environmental sensor data + 4 GB of camera data throughout 18 stations.
* This may not sound like a lot, but this is just 1 system for 3 counties.
* NASA's MODIS, or Moderate Resolution Imaging Spectroradiometerl, is a satellite that has imaging instruments on 2 satellites, Aqua + Terra.
* MODIS instruments on these satellites capture images of the entire surface of Earth every 1-2 days, acquiring data in 36 spectral events, which equals 40 science products + produces 600 GB of data per day, or 219 TB of data per year.
* It's not that different in precision medicine.
* 1 of the key promises in precision medicine comes from using individuals' genetic profile to guide decisions regarding prevention, diagnosis, + treatment of disease.
* Genome sequencing is only 1 part of data + it needs to be augmented w/ treatment data, medical histories, + other biomedical data.
* According to a Fast Company article in 2016, the genome sequences of people who will be diagnosed w/ cancer was predicted to equal 4 exabytes (EB), or 10^18 bytes
* Other large volume data sources in scientific research comes from LIGO, Deep Space Network, + Protein Data Bank.
* LIGO, the Laser Interferometer Gravitational-Wave Observatory, is a data source that led to the gravitational wave discovery in 2016.
* The experiment provides large scale physics + observatories to detect cosmic gravitational waves.
* Deep Space Network, which is NASA's network of large antennas + communication sites, located in several countries + are used to support space missions + research asteroids + planets, updates its data stores w/ real time data every 5 seconds.
* Protein Data Bank, which is a repository of info about 3D structures of large biological molecules, which is important for research on human health, disease, drug development.
* Management + analysis of such scientific data sets is a huge challenge for modern scientific research
* 1000 MB = 1GB, 1000 GB = 1 TB, 1000TB = 1PB, 1000 PB = 1 Exabyte/EB, 1000 EB = 1 zettabyte/ZB
* 100 MBs will hold a couple of encyclopedias, a DVD is around 5 GBs, + 1 TB would hold around 300 hours of good quality video.
* A data oriented business currently collects data in the order of TBs, but PBs are becoming more common to our daily lives.
* CERN's Large Hadron Collider generates 15 PBs of data a year.
* According to a report by IDC, sponsored by a big data company called EMC, digital data will grow by a factor of 44 until the year 2020, a growth from .8 ZB in 2009 to 35.2 B.
* A ZB = 1 trillion GBs/10^21 GB
* The effects of it will be huge, considering all the time, cost, + energy that will be used to store + make sense of such an amount of data.
* The next era will be **yottabytes** (YB), 10^24, + **brontobytes**, 10^27, which is really hard to imagine for most of us at this time.
* This is also what we call data at an **astronomical scale**.
* The bottom line is that all of these sources point to an exponential growth in data volume + storage.
* While many of us are excited by the opportunities offered by big data, this rapid growth also comes w/ a number of management + analysis challenges, least of which is information overload.
* Our challenge isn't just to *manage* the data, but to try to see how everything is connected.
* Finding connections between different kinds of data sets has the potential to lead to interesting discoveries.
* Such an endeavor requires proper use of data management, data driven methods, scalable tools for dynamic coordination, scalable execution, + a skilled interdisciplinary workforce.
* By putting time into skills + programming, statistics, machine learning, + big data, you will be ready to take on some of the technical challenges in data science like drug effectiveness analysis, crime pattern detection, + self-driving cars.
* A data science team often comes together to analyze situations or answer questions in business or science which no single person could solve on their own.
* There are lots of moving parts to the solution, but in the end all these parts should come together to provide actionable insight based on data science.
* Being able to use evidence-based insight in your decisions is more important now than ever.

*Why Python*

* You have probably seen diagrams like this one that describes data science.



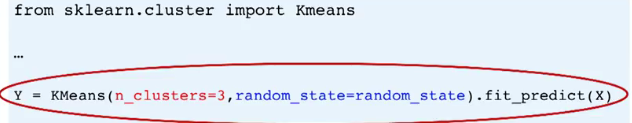
* Data science happens at the intersection of CS, mathematics, + business or scientific expertise.
* If you zoom deeper into this diagram + open up the sets of expertise, we would see a variation of this figure.



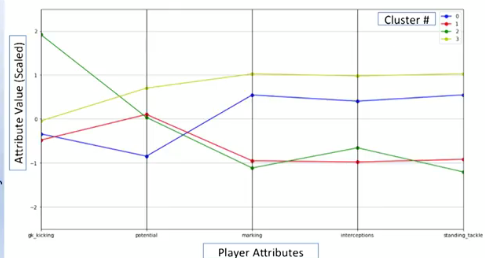
* Even at this level, all of these boxes require deeper knowledge + skills in areas like domain expertise, data engineering, statistics, + computing
* Even deeper analysis of these skills, based on data science job listings, would lead you to skills like ML, statistical modeling, relational algebra, business passion, problem solving, + data visualization.
* That's a lot of skills to have for a single person.
* Given such a wide range of skills across multiple definitions, “data scientists” seems to be impossible
* Some folks have even begun to ask if data scientists are like unicorns, meaning they don't exist.
* There *are* data science experts who has expertise in more than 1 of these skills for sure, but they're relatively rare + still would probably need help from an expert on some of these areas.
* So in reality, data scientists are *teams* of people who act like one.
* This is why we say data science is team sport, referring to the breadth of information + skills it takes to make it happen.
* However, there are still common traits to data scientists.
* For example, they’re passionate about the story + meaning behind data, they understand the problem they are trying to solve + aim to find the right analytical methods to solve this problem, + they all have an interest in engineering solutions to solve problems.
* They also have curiosity about each other's work + have communications skills to interact w/in the team + present their ideas+ results to others.
* These soft skills are very important for success in any data science team.
* But what data tools should you pick?
* According to a recent article on KDnuggets, based on skills + jobs data from Indeed.com, Python is a clear leader in many data science categories.
* Although learning other programming languages (including = Java, C, Scala, R, + Julia) is a good idea.
* Instead of explaining why Python is a good language for data science, let's focus on why data scientists love Python.
* In addition to being an easy-to-learn + readable language, Python is an open-source language w/ a vibrant community.
* Thanks to the efforts of this community, it offers an ever-growing set of data management, analytical processing, + visualization libraries
* Such libraries make Python applicable to every step of the data science process.
* Lastly, but very importantly, **Jupyter Notebooks** make Python-based analysis more producible + repeatable, as well as provides built-in training + communication support to help w/ team communication.
* Some libraries include:
* **NumPy** + **Pandas** to ingest + analyze data efficiently
* visualization libraries, including **Matplotlib**
* ML libraries in **Scikit-Learn** to create models.
* Libraries like **BeautifulSoup** to easily read an XML + HTML-type data,

*Case Study: Soccer Data Analysis*

* This Kaggle European soccer database has more than 25k matches + more than 10k players for European professional soccer seasons from 2008-2016.
* The dataset even has attributes on weekly game updates, team lineup, and detailed match events, but we won’t be using them in this example
* We will use these datasets to demonstrate the basic steps of the process we take for such data science projects for 3 main goals.
* Form meaningful player groups,
* Discover other players that are similar to your favorite athlete,
* Form strong teams by using analytics.
* Since we are looking for questions to solve in data science, we can formulate the question around these goals.
* For example, we would say “how do I form meaningful player groups to find players similar to my favorite player?” and then “how do I use this information to form strong teams?”
* To go even further, we need to ask ourselves why we want to know about strong teams, or what is the benefits of using analytics?
* 1 of the most critical question you can ask yourself is *WHY* am I doing data science in this problem? What insights do I expect?
* This question will lead you to find what you are looking for + you can design strategies in alignment w/ your goals.
* *Our* goal is to take the data + generate insights from it that we can use to take data-driven actions.
* We call these insights **actionable insights**, which are very valuable, but require knowledge of the subject area/business domain.
* For the soccer scenario, the insights we are trying to generate are related to better understanding of player strengths, enhancing performance, + critical attributes of a player's performance.
* We can turn this into a question saying “I want to find the quickest way to improve my favorite player's performance. How do I know what traits impact a player's performance more than the others?”
* The coach can then take these insights + take actions to design programs built upon these insights to improve team strengths in these areas
* There are 5 key steps in the overall process of data science
* Data Acquisition 🡺 Data Prep 🡺 Data Analysis 🡺 Presentation + Reporting of insights 🡺 Turning these insights into data-driven actions.
* In our soccer example, acquire involves downloading the dataset/importing it.
* Acquisition leads into data exploration + visualization + other data prep.
* We then analyze the prepared dataset using statistical analysis, feature selection, model selection, + ML
* Findings from these analysis are typically turned into reports + presented to the **stakeholders**, who need to take actions using such insights
* There are similar steps in any data science project, even though the specifics of each step can be different
* As a 1st step in any data science activity, we need to consider that data can come from many different sources.
* This diversity of data sources will only continue to grow as more innovations are made.
* The broad categories include RDBs + non-RDBs/NoSQL databases, text files in various data formats, + live online streams coming from machines, sensors + online activities.
* In our soccer example, the provider of the dataset gathered the data scattered across many internet sites + did data collection + processing to make the data ready for analysis for us.
* The dataset includes structured data on scores, lineup, info + events, as well as data on betting odds + players + teams' attributes.
* All we had to do in this case was to take that data set + ingest it into Python.
* **Data ingestion** will be one of our focus areas in this course.
* Python has well-defined methods for ingesting data from diverse resources, such as various databases, data access APIs like the Twitter API, text files, + sensors' data streams.
* The next step of our data science process is **exploring** the dataset.
* Python has libraries that can assist in the data prep phase when you want to explore datasets.
* For example, w/ just one line of command, **df.describe().transpose**, we can generate a vital statistical summary of your datasets like mean + SD.
* The data prep also involves data cleaning, as there are many challenges in real world data sets.
* Cleaning can also build on the statistical analysis, like removing outliers, missing values, or in general, weeding out unwanted stuff from your data.
* Although sometimes removing the unwanted entries can be a quick solution, sometimes it can still be a challenge to decide what to remove.
* In those situation, you can **impute** those fields w/ known aggregate values such as mean of the columns, use binary values 0, -1, etc.
* Python offers data cleaning functions to help w/ general data cleaning tasks like finding + removing NULL values.
* In each step of the data science process, **data visualization** is an effective way to capture attention + convey a message in minimal time.
* Python has several open source data visualization libraries that can make this task much easier.
* ***Analysis*** *is the crux of the matter in data science.*
* Once the basic preparatory steps are completed, you get to the algorithms.
* There are three key categories of algorithms: **supervised**, **unsupervised**, + ***semi*-supervised learning**.
* There are vast number of algorithms + techniques as seen in this diagram for dimensionality reduction, clustering, + regression, for example.
* **scikit-learn** provides many tools for ML in Python.
* **Feature Selection =** selecting attributes that have the greatest impact towards the problem you’re solving.
* *It requires some domain knowledge* to narrow down the # of features.
* For example, in the soccer use case, if trying to predict player performance, what are the most critical attributes?
* Agility, reaction time, shot power, + sprint speed? Or hair style, or movie preferences.
* Similarly, if you're grouping players into different sets, what attributes would you choose to assign these groupings or to create new, complex features (i.e. **feature engineering** 🡪 shot power + reaction time)
* *Narrowing the features* has several benefits:
* You get models that are easier to interpret.
* Models train much faster
* Likely to generalize well to newer scenarios.
* You will find almost every top ML algorithm is already implemented in Python.
* Different functionalities are organized into libraries in Python like **scikit-learn**, which contains implementations of fundamental ML algorithms.
* In the soccer example, we will utilize a form of **clustering** algorithm called **K-means**
* **Clustering** = grouping players into similar meaningful sets based on those decided attributes
* Can do w/ just 1 line in Python after importing the library



* After this, we start **interpreting** + **analyzing** the results.
* So how do we analyze the clustering results?
* Do all clusters of them have the same number of players?
* When we look at the attributes we had selected,
* How do these groups differ?
* Plotting these clusters might help w/ interpreting + presenting these results.

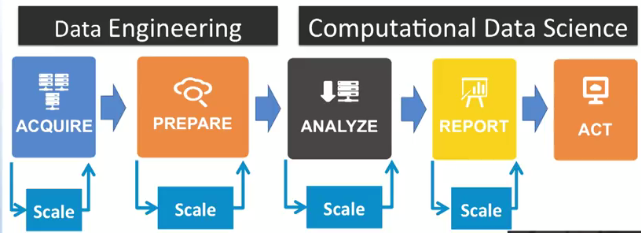


* Once we have done all the work of data cleaning, analyzing + interpretation, it's time to **present** our findings.
* A big part of the presentation/reporting is explaining how we interpreted these results.
* Look at the above graph + think of the four lines as signatures of each of the four clusters our K-means algorithm has found based on the features shown on the X-axis of this plot.
* Do any of the groups have exactly the same signature? 🡪 NO
* Each group is unique in the sense that it differs from the other three in *at least one attribute*.
* To act upon such findings, team coaches can use this info to design customized improvement strategies for each group.
* There are many techniques + best practices for presentation or visualization of these results.
* We need to decide on the graph type + have enough details for the picture to be self-explanatory, like adding label axes, legends, + a readable font size.

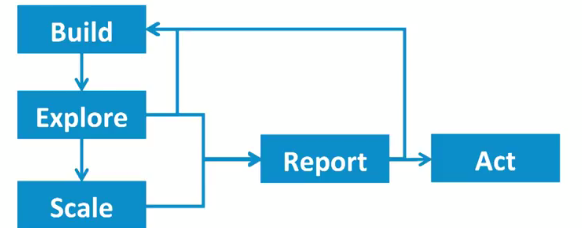
***The DF Process***

*How Does DS happen?*

* Key features/dimensions of **DS,** a multidisciplinary craft combining an interdisciplinary team w/ an application purpose
* DS starts w/ a team of people w/ an overarching broad question + some data to explore
* We start /w data and a question and then build a process around how to come up w/ a data-driven insight
* The process is conceptual in the beginning and defines the core set of steps to solve the question
* From then on, we drill down through different areas of expertise + have some blurred liens between steps
* Can look at the process as 2 district activities containing 5 distinct steps
* **Data engineering 🡺** Data acquisition + prep
* **Data analysis/computational data science 🡺** Analysis, reporting, acting
* DS happens at the boundary of all these steps
* This process above should support experimental work and dynamic **scalability** on big data + cloud platforms



* This 5-step process can be used in alternative ways in real-life big data applications by adding the dependencies of different tools to each other
* The rise of Big Data pushes for alternative scalability approaches at each step
* Another way to look at the process is seeing all the steps as containing reporting needs in different forms



* i.e. drawing all of these activities as an iterative process, including build, explore, + scale for big data as separate steps
* **Scalable data analysis** needs alternative data management techniques, systems, analytical tools + methods, as well as nodes of scalability based on dynamic data + computing load, change in physical infrastructure, and streaming data-specific urgencies arising from special events.
* For simplicity, we will refer to the process as the set of 5 sequential activities that iterate
* However, at the end of the day, the **scalable process** should be programmable through utilization of reusable + reproducible programming interfaces to systems, analytical tools, visualization environments + user reporting environments.

*Asking the right question*

* The 1st step in any process is to **define** what it is you're trying to tackle.
* What is the problem that needs to be addressed or the opportunity that needs to be asserted?
* W/out this, you won't have a clear goal in mind or know when you've solved your problem.
* Ex questions: “How can sales figures in call center logs become viable to evaluate any product?”
* “In a manufacturing process, how can data from multiple sensors on an equipment be used to detect equipment failure?”
* “How can we understand our customers + market better to achieve effective targeted marketing?”
* Next, you need to **assess** the situation w/ respect to the problem or opportunity you have defined.
* This is a step where you need to exercise caution + analyze risks, costs, benefits, contingencies, regulations, resources, + requirements of the situation.
* What are the **requirements** of the problem?
* What are the **assumptions** + **constraints**?
* What **resources** are available to you (in terms of both personnel + capital such as CPU systems, equipment etc.)?
* What are the main **costs** associated w/ this project?
* What are the potential **benefits**?
* What **risks** are there in pursuing a project?
* What are the **contingencies** to potential risks?
* Answers to these questions will help you get a better overview of the situation + better understanding of what the project involves + how you will guide your programming to solve the project w/ all these in mind.
* Then, you need to define your **goals + objectives**.
* Defining success criteria is also very important 🡪 *What do you hope to achieve by the end of this project?*
* Having clear goals + success criteria helps to assess a project throughout its lifecycle.
* Once you know the problem you want to address + understand the constraints + goals, then you can formulate the plan to use to come up w/ the answer/solution to your business problem, or the analytics you are trying to achieve.
* As a summary, defining the questions you're looking to find answers for is a huge factor contributing to the success of any data science project.
* By following the explained set of steps, you can formulate better questions to solve using analytical skills + link them to scientific + business value.

*Steps in DS*

* 1) Acquire 🡪 includes anything that makes us retrieve data (finding, accessing, acquiring, and/or moving data)
* Includes ID of an authenticated access to all related data, transportation of data from different source, + ways to subset + match data to regions/times or interests (**geospatial query)**
* 2) Prep
* 2a) Explore🡪 literally looking at data to understand its nature, what it means, its qualities, its format
* Often takes preliminary analysis of data/samples of data to understand this
* 2b) Pre-process🡪 cleansing, subsetting, filtering
* Creating data that programs can read + understand via modeling raw data into a more-defined data model or by packaging it in a specific data format
* If multiple data sets, this step include integration of the data from different sources/streams
* 3) Analyze 🡪 selection of analytical techniques to use, building a model of the data, + analyzing results
* Can take multipole iterations or require one to go back to steps 1 + 2 to get more data or package data in a different way
* 4) Communicate Results 🡪 evaluation of analytical results, presenting them in a visual manner, + creating reports that include an assessment w/ respect to success criteria
* “interpret”, “summarize”, “visualize”, “post-process”
* 5) Apply Results --< reporting insights from analysis + determining actions from insights based on the purpose initially defined
* ***NOTE:*** This process is *iterative* and findings form 1 step may require previous steps to be repeated w/ this new data/info

*Acquiring Data*

* 1st step in DS process 🡪 determion what data is available 🡪 leave no stone unturned when finding the right data sources
* Want to ID + make use of all suitable data relevant to the problem
* Sometimes, leaving out even a small amount of important data can lead to incorrect conclusions
* May data sources (local + remote) in many varieties (structured vs. unstructured) in various velocities (streaming speed)
* A lot of data is in RDBs (structured data from an organization) we access w/ SQL.
* Data also exists in text and CSV files (Excel)
* Scripting language (high-level programming language that can be general purpose or have specialized purpose) is used to get data from files
* Websites are increasingly popular places to get data
* Via webpages, written using set of standards approved by the **Worldwide Web Consortium (W3C)**
* Variety of formats and services, like XML or JSON, who use markup symbols + tabs to describe web page content
* Sites also host web services to provide programmatic access to their data
* Such as **Representational State Transfer (REST) API’s**, an approach for implementing web services w/ scalability + maintainability in mind
* **Websocket** services allow real-time notifications from websites (growing popularity)
* **NoSQL** are increasingly used to manage variety of data types
* Databases that are NOT in a table format (like RDBs) like mongoDB, HBASE
* NoSQL data stores provide APIs to allow users to access data, either directly or in an application (like a Python script)
* Most NoSQL systems allow data access via a web interface like REST
* Ex:San Diego Supercomputer Center wildfire data analysis + predictions of fire direction + rate of spread
* Acquires data through several different mechanisms
* Stores historical sensor data in an RDB + retrieve it w/ SQL
* Real-time data from listening to websocket services allows one to determine if a weather station is experience fire weather conditions by receiving weather station measurements as they occur
* Data form this is processed \_ compared to patterns from the model
* Tweets retrieved using any hashtags related to fires occurring in the region via Twitter’s REST API to determine tweet sentiment
* Combo of sensor data + tweet sentiments helps to give a sense of urgency of a situation

*Exploring Data*

* After putting the data we need together, it’s important to resist temptations to immediately build models to analyze it
* 1st step after getting data is a preliminary exploratory analysis to understand the specific characteristics of our data
* Looking for correlations, general trends, outliers, etc.
* Won’t be able to use data effectively w/out this step
* **Correlation graphs** can show dependencies between variables, **general trends** show a simple graph of how data progresses over time, and **outliers** show data points unusually far from others
* Outliers help double-check for errors in data due to measurement
* in some cases, they’re not errors + help use find a rare event
* Summary statistics provide numerical values to describe data + the nature of it, + capture various characteristic of a set of values a small set of #’s (mean, median, mode, SD, range)
* Visualizations also provide quick + effective + useful way to look at data in a preliminary analysis
* Heat maps gives ideas about hot spots, histograms + boxplots show distribution shapes + skewness, line graphs show how values change over time + help easily see spikes, scatterplots help see correlations

*Pre-processing Data*