**What is Data Science?**

Data science is the systematic approach to collecting, processing, analysing, and interpreting data to uncover hidden patterns, gain actionable insights, and support informed decision-making. It encompasses a wide range of techniques and tools, including statistics, machine learning, data visualization, and domain knowledge, to extract valuable knowledge and drive advancements in various fields, such as business, healthcare, finance, and research.

Data science can involve statistics, computer science, mathematics, data cleaning and formatting, and data visualization.

**Data science can involve:**

* Statistics, computer science, mathematics
* Data cleaning and formatting
* Data visualization

**Why do we need Data science?**

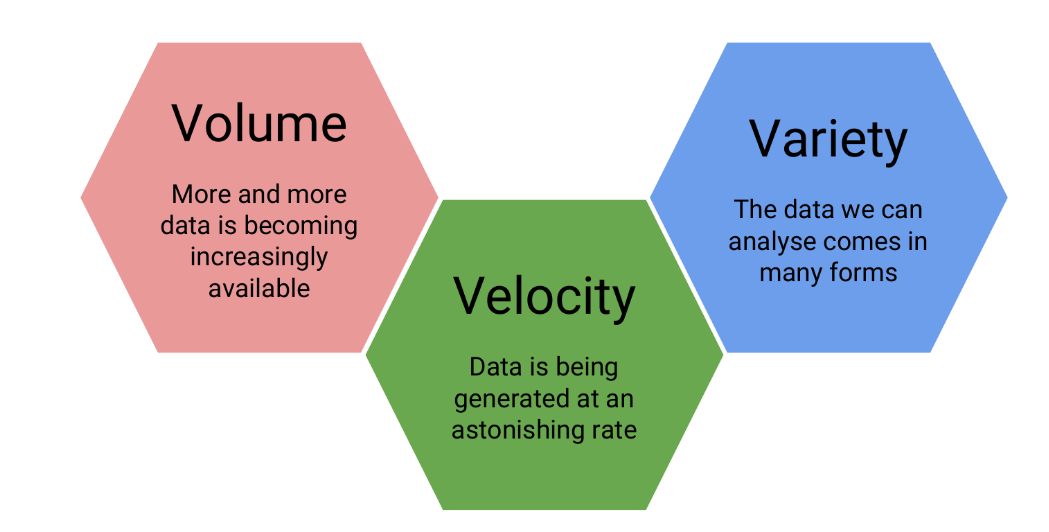
In recent years , the vast amount of data currently available and being generated. And to handle this huge amount of data we need a particular field for this (i.e. Data science) .

**What is Big Data?**

Volume: big data involves large datasets - and these large datasets are becoming more and more routine.

Velocity: Data is being generated and collected faster than ever before.

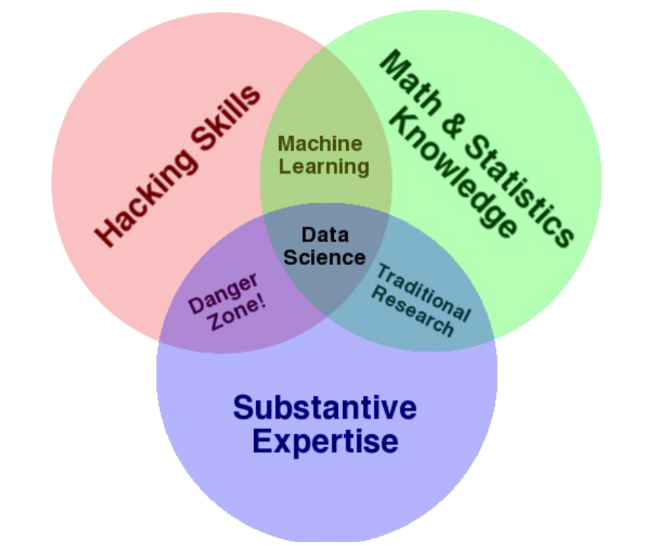
Variety: different types of data available tous like video and audio.



**Who is a Data Scientist?**

A data scientist is broadly defined as someone who combines the skills of software programmer, statistician, and storyteller/artists to extract the nuggets of gold hidden under mountains of data.

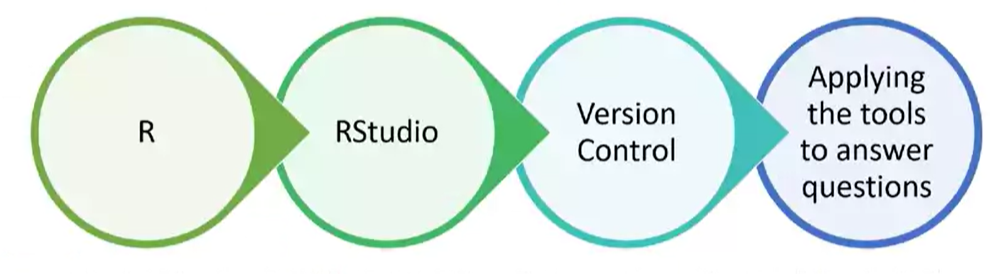
A good data scientist asks questions first and seeks out relevant data second.



\*\*\*Drew Conway’s Venn diagram of data science\*\*\*

Home\_work

* Search for Daryl Morey
* Hilary Mason
* Nate Silver



More common types of messy data

* *Sequencing data*
* *Population census data*
* *Electronic medical records (EMR), other large databases*
* *Geographic information system (GIS) data (mapping)*
* *Image analysis and image extrapolation*
* *Language and translations*
* *Website traffic*
* *Personal/Ad data (eg: Facebook, Netflix predictions, etc)*

Getting Help for Data science related problem prefer these plateforms:

* *StackOverflow*
* *Cross Validated*
* *Coursera*

First steps for solving coding problems

* *Check for typos!*
* *Read the error message and make sure you understand it*
* *Google the error message, exactly*

Forum etiquette

* *Read the forum posting guidelines*
* *Make sure you are asking your question on an appropriate forum!*
* *Describe the goal*
* *Be explicit and detailed in your explanation*
* *Provide the minimum information required to describe (and replicate) the problem*
* *Be courteous! (Please and thank you!)*
* *Follow up on the post OR post the solution*

# Roadmap for Data Science

A roadmap for data science typically includes a series of steps and skills to acquire as you progress in your data science journey. Here's a roadmap that outlines the key stages and topics you should consider:

**1. Fundamental Skills:**

* **Mathematics and Statistics:**
  + Linear algebra
  + Calculus
  + Probability and statistics
* **Programming:**
  + Python or R
  + SQL
* **Domain Knowledge:**
  + Understand the specific domain or industry you want to work in.

**2. Data Collection and Cleaning:**

* **Data Collection:**
  + Collect data from various sources, including databases, APIs, and web scraping.
* **Data Cleaning:**
  + Handle missing data
  + Remove duplicates
  + Handle outliers
  + Data imputation

**3. Data Exploration and Visualization:**

* **Exploratory Data Analysis (EDA):**
  + Understand your data's characteristics and patterns.
* **Data Visualization:**
  + Use libraries like Matplotlib, Seaborn, or ggplot2 for creating plots.
  + Visualize relationships, trends, and distributions in the data.

**4. Feature Engineering:**

* Create relevant features from the data that help improve the performance of machine learning models.

**5. Machine Learning:**

* **Supervised Learning:**
  + Regression
  + Classification
* **Unsupervised Learning:**
  + Clustering
  + Dimensionality Reduction
* **Model Evaluation and Selection:**
  + Cross-validation
  + Metrics (e.g., accuracy, precision, recall, F1-score, RMSE, MAE)
* **Ensemble Methods:**
  + Random Forest, Gradient Boosting
* **Deep Learning:**
  + Neural networks with frameworks like TensorFlow or PyTorch

**6. Model Deployment:**

* Deploy machine learning models into production using cloud services, containers, or platforms like Flask or Django.

**7. Big Data and Distributed Computing:**

* Learn how to work with big data technologies like Hadoop, Spark, and distributed computing frameworks.

**8. Advanced Topics:**

* Dive into more specialized areas of data science, such as:
  + Natural Language Processing (NLP)
  + Computer Vision
  + Time Series Analysis
  + Reinforcement Learning

**9. Data Ethics and Privacy:**

* Understand the ethical implications of data science, including bias, fairness, and privacy.

**10. Portfolio Building:**

* Create a portfolio of data science projects to showcase your skills and experience.

**11. Continuous Learning:**

* Stay up-to-date with the latest developments in data science by reading research papers, blogs, and taking online courses.

**12. Networking and Collaboration:**

* Join data science communities, attend meetups and conferences, and collaborate with other data scientists.

**13. Soft Skills:**

* Develop effective communication and storytelling skills to convey your findings and insights to non-technical stakeholders.

**14. Job Search:**

* Apply for data science positions, internships, or freelance opportunities.

**15. Career Development:**

* Continue to learn, grow, and advance in your data science career.

Remember that this roadmap is just a general guide. Your specific path in data science may vary based on your interests, career goals, and the industry you choose to work in. Adapt and customize your journey to best suit your needs and aspirations.

Top of Form

Link for Documentation of R Programming…

<https://www.r-project.org/>

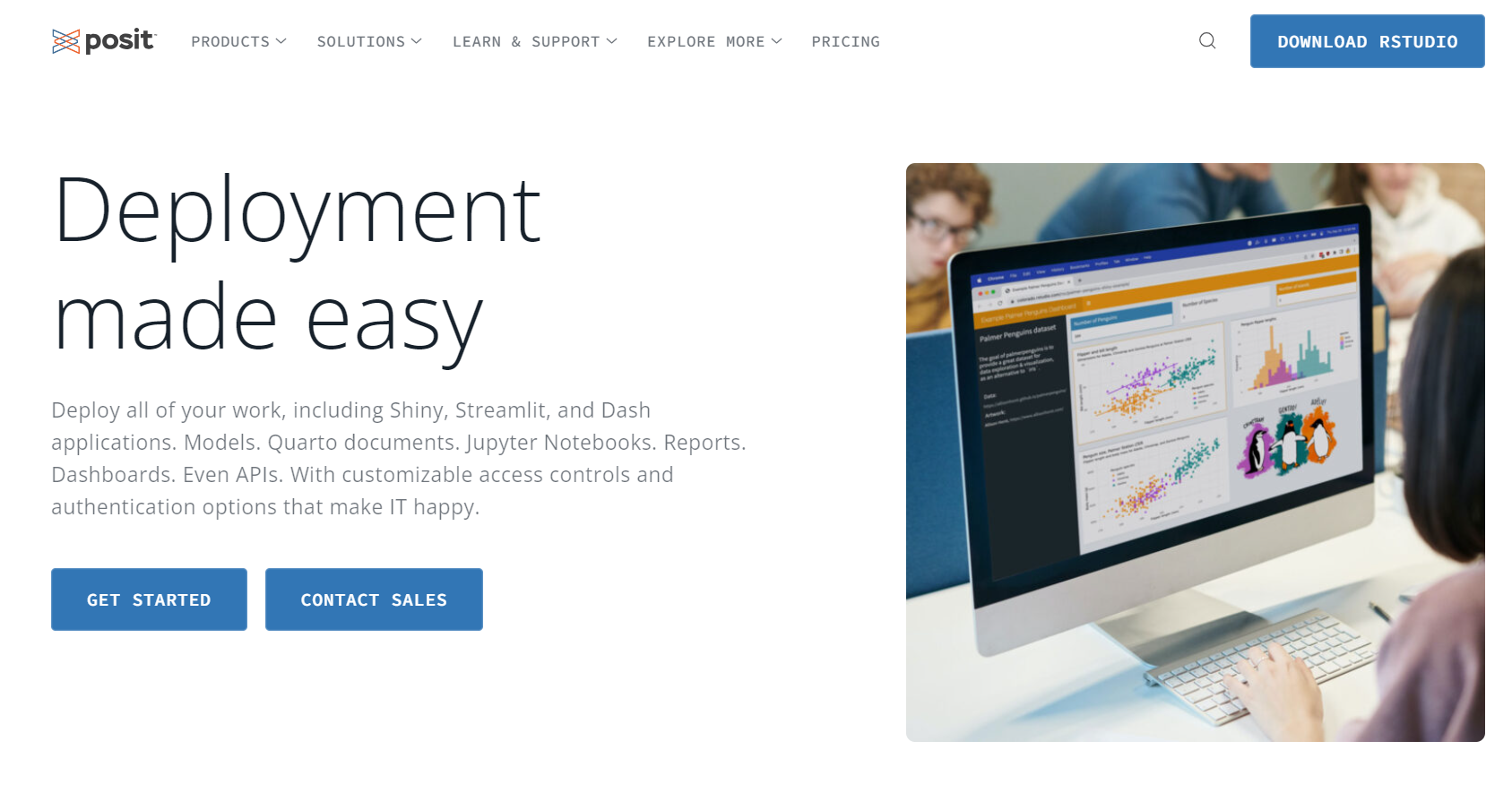
<https://www.programiz.com/r>

<https://www.tutorialspoint.com/r/r_basic_syntax.htm>

What is R?

R is both a programming language and an environment, focused mainly on statistical analysis and graphics. It will be one of the main tools you use in this and following courses.

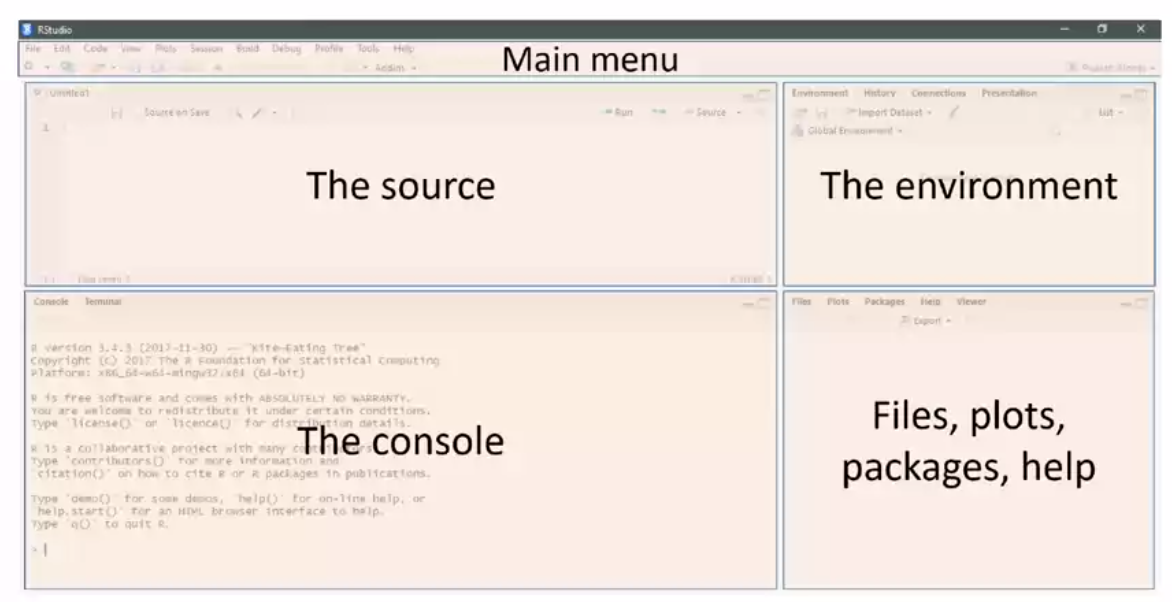
**Posit**

****

What is CRAN?

The “Comprehensive R Archive Network” ( CRAN ) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.

RStudio



The "Iris" dataset in R contains data on iris flowers, and it includes three species of iris:

**Setosa**: Iris setosa

**Versicolor**: Iris versicolor

**Virginica**: Iris virginica

R-PACKAGES

* Available CRAN packages by name…

https://cran.r-project.org/web/packages/available\_packages\_by\_name.html

**Linking GitHub and RStudio**

**Data science podcast:**

<https://soundcloud.com/nssd-podcast>

Sell your book

<https://www.lulu.com/>

Commands:

getwd() //working directory

read.csv(“mydata.csv”)

dir()

read.csv(“mydata.csv”)

ls()

myfunction()

source(“mycode.R”)

second(4)

**Download and install the swirl package for R**

install.packages("swirl")

To install and use packages in R, including the "swirl" package, you typically use the **install.packages()** function to install the package and the **library()** function to load it into your R session. Here are the steps to install and use the "swirl" package:

1. **Install swirl Package:**

Open R or RStudio and run the following command to install the "swirl" package from the CRAN repository:

RCopy code

install.packages("swirl")

This command will download and install the package on your system.

1. **Load the swirl Package:**

After successfully installing the package, you can load it into your R session using the **library()** function:

RCopy code

library(swirl)

This makes the "swirl" functions and lessons available for use.

1. **Start a swirl Session:**

To start a new swirl session and access interactive R tutorials, you can run the following command:

RCopy code

swirl()

This command will guide you through the process of setting up a swirl session, selecting a course, and completing interactive lessons. Follow the on-screen instructions to get started with swirl.

1. **Select a Course:**

After starting a swirl session, you will be prompted to select a course. You can choose from various available courses, including those related to R programming, data analysis, and statistics.

1. **Complete Lessons:**

Once you've selected a course, you can complete the interactive lessons provided by swirl. Follow the prompts, answer questions, and practice R as you progress through the course.

1. **Exit swirl:**

To exit a swirl session, simply type:

RCopy code

bye()

This will save your progress and exit the swirl environment.

That's it! You can now use the "swirl" package in R to learn and practice R programming interactively. Remember that "swirl" offers various courses, so you can choose the one that best fits your learning goals.

**R vs Python**

"R vs. Python" is a common comparison made in the field of data science and statistical computing. Both R and Python are powerful programming languages used extensively for data analysis, machine learning, and scientific research. The choice between them often depends on your specific needs, preferences, and the tasks you want to accomplish. Here are some key points to consider when comparing R and Python:

**R:**

1. **Specialized for Statistics and Data Analysis:** R was specifically designed for statistical analysis and data visualization. It has a wide range of packages and libraries for statistical modeling, data manipulation, and graphing.
2. **Strong Data Visualization:** R has a rich ecosystem of data visualization packages, such as ggplot2, which makes it easier to create high-quality plots and charts.
3. **Community:** R has a strong and active community of statisticians and data scientists, which means there are many resources, packages, and support available.
4. **DataFrames:** R uses data frames as a primary data structure, which is similar to a spreadsheet and is well-suited for data analysis.
5. **Learning Curve:** R may have a steeper learning curve for those with no prior programming experience, especially if you're not familiar with statistical concepts.
6. **Data Cleaning:** R has excellent tools for data cleaning and transformation, which is crucial for data analysis.

**Python:**

1. **General-Purpose Language:** Python is a versatile, general-purpose programming language that can be used for a wide range of tasks, not just data analysis.
2. **Machine Learning and Deep Learning:** Python has become the dominant language for machine learning and deep learning, with libraries like scikit-learn, TensorFlow, and PyTorch.
3. **Community:** Python also has a large and active community, and it is one of the most popular programming languages in the world, which means extensive documentation, resources, and third-party libraries.
4. **Data Structures:** Python uses versatile data structures, including lists, dictionaries, and NumPy arrays, which are suitable for various tasks.
5. **Learning Curve:** Python is often considered more accessible to beginners due to its readable syntax and large community of learners.
6. **Integration:** Python can easily integrate with other technologies and languages, making it suitable for end-to-end data pipeline development.

Ultimately, the choice between R and Python depends on your specific project requirements, your background, and your personal preferences. Some data scientists and analysts choose to use both languages together, taking advantage of their strengths in different areas. It's also worth noting that many data science tools and libraries are available in both R and Python, so you can often achieve similar results in either language.

**History of the S and R programming lectures**

*1.Basic calculations*

*2.Absolute values*

*3.Variable like x<-12*

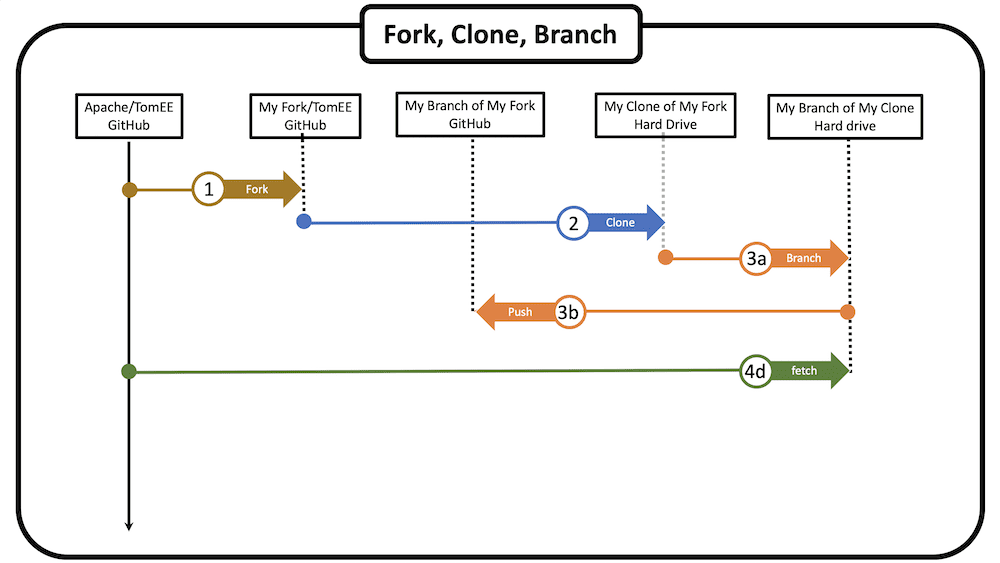
*4.Storing a tuple y<-(-12,6,0, -1)*

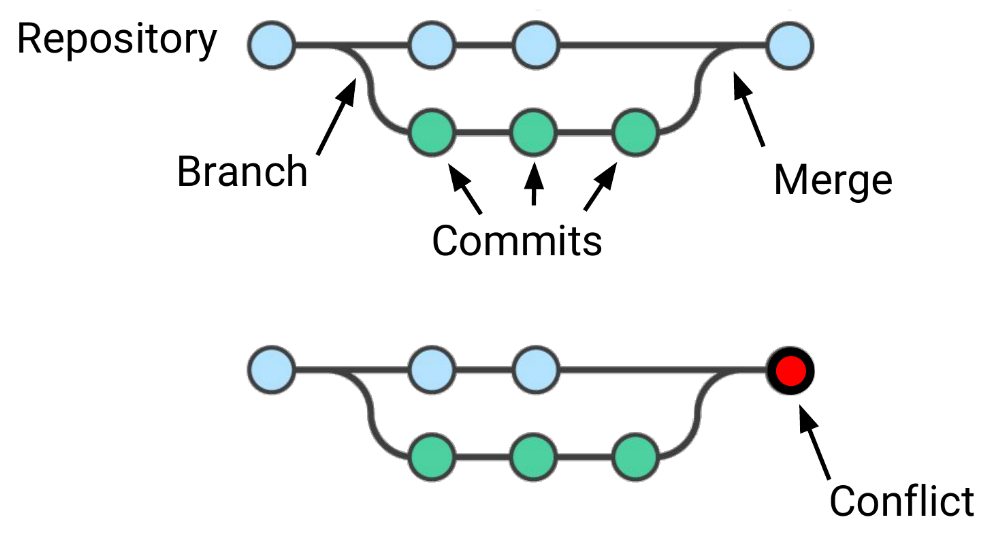
*If 2\*y output----- (-24,12,0,2)*

*abs(y) output-----(12,6,0,1)*

*5.*

Difference between Clone, Branch, and Fork





**Run this in the console**

Install.packages(“rmarkdown”)

**What is R Markdown?**

R Markdown is a way of creating fully reproducible documents, in which both text and code can be combined. In fact, these lessons are written using R Markdown! That’s how we make things:

* bullets
* bold
* italics
* links
* or run inline r code

**Type of Data Analysis…**

* Descriptive (*summarize a set of data*)
* Exploratory (*explore the data and find relationships*)
* Inferential (*use a relatively small sample of data to infer*)
* Predictive (*use current data to make predictions about future data*)
* Casual (*the goal of causal analysis is to see what happens to one variable when we manipulate another variable*)
* Mechanistic (*goal of mechanistic analysis is to understand the exact changes in variables that lead to exact changes in other variables*)