**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?**

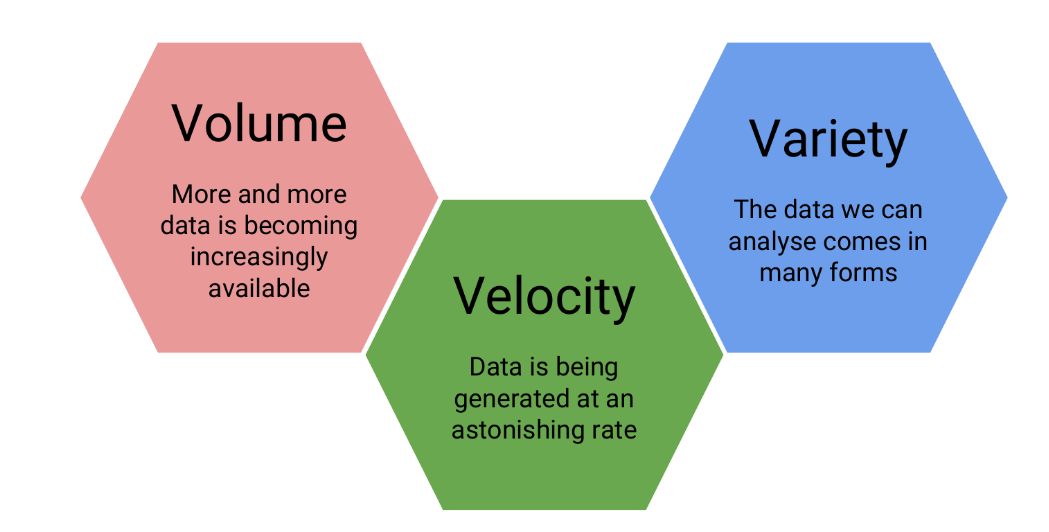
In recent years , the vast amount of data currently available and being generated. And to handle this huge amount of data we need a particular field for this (i.e. Data science) .

**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.

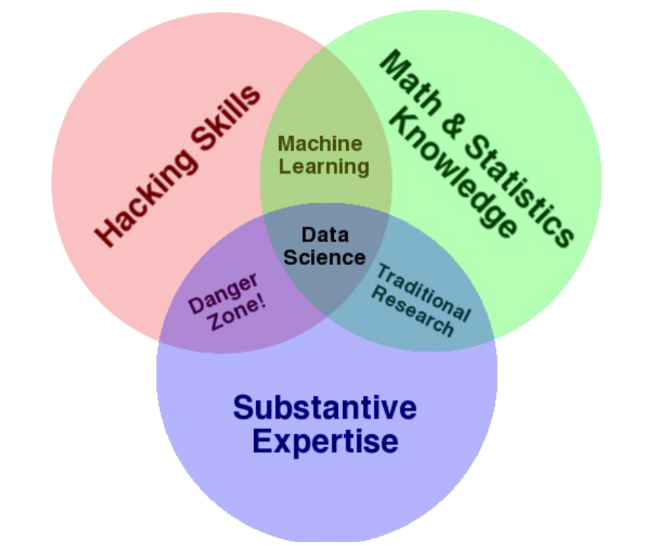
Variety: different types of data available tous like video and audio.



**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.

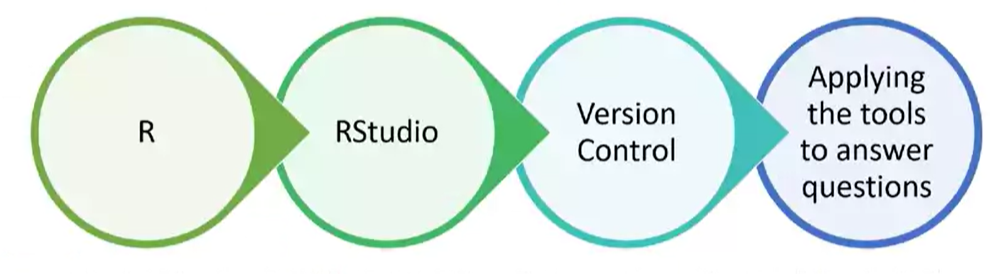
A good data scientist asks questions first and seeks out relevant data second.



\*\*\*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*

Link for Documentation of R Programming…

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

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

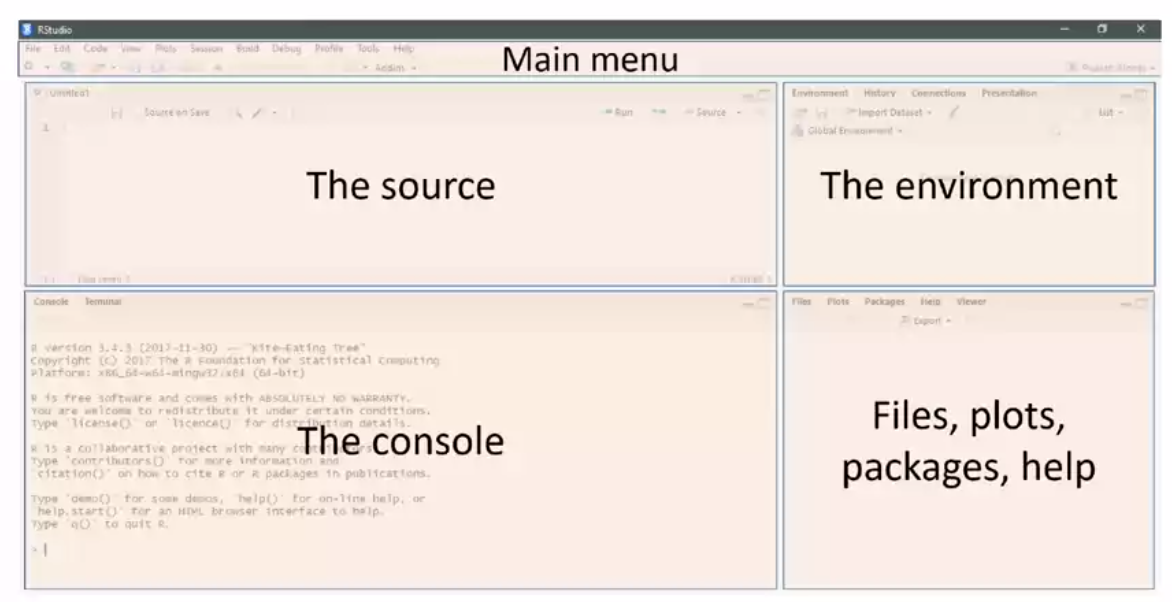
What is R?

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.

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

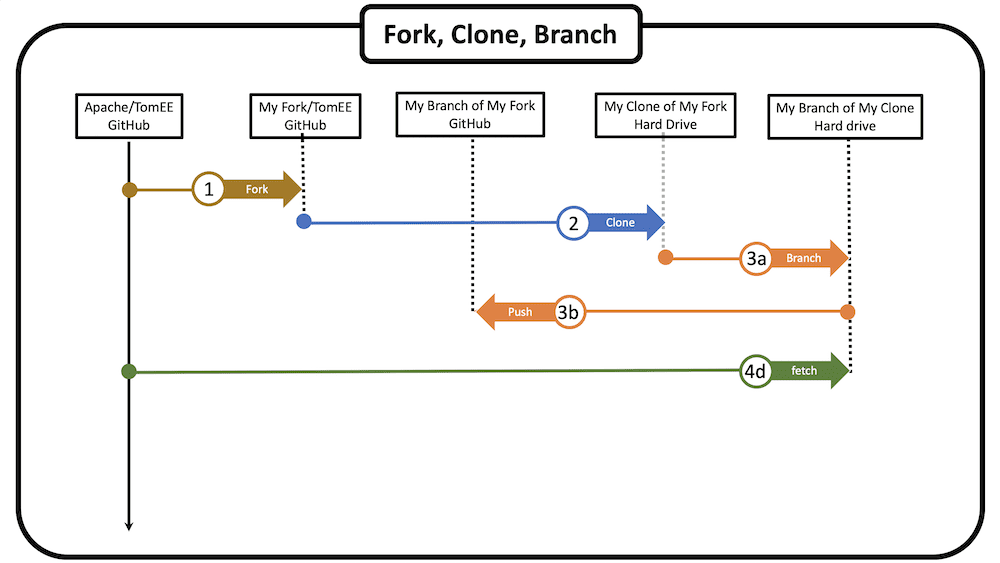
R-PACKAGES

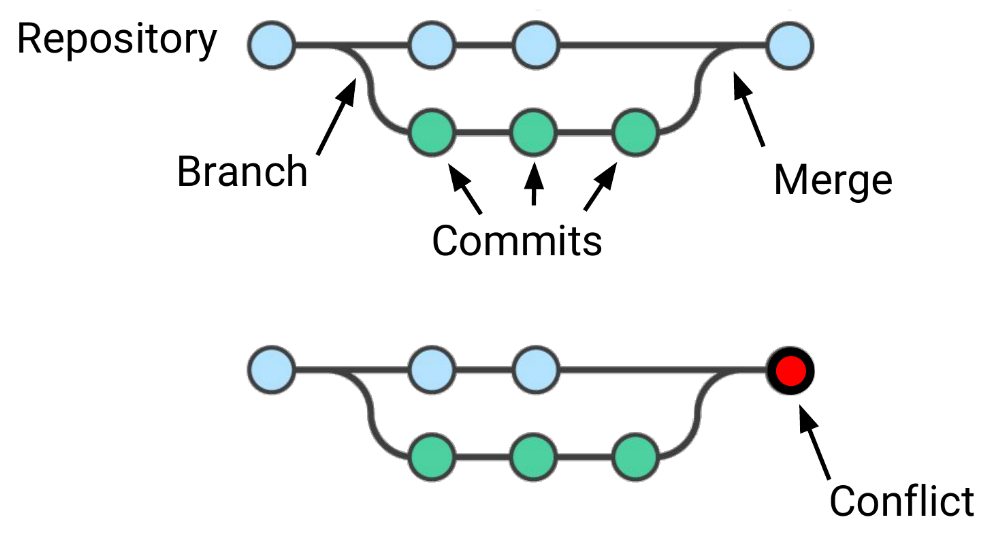
* Available CRAN packages by name…

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

**Linking GitHub and RStudio**

Difference between Clone, Branch, and Fork





**Run this in the console**

Install.packages(“rmarkdown”)