Exploratory Data Analysis with R

Introduction

Xuemao Zhang
East Stroudsburg University

August 29, 2022

What's covered in this lecture?

- Math 318 Course Outline
  - Course Objectives
  - Tentative Contents
  - Assessments
  - References
- Why programming?
- R, RStudio and R Markdown
- SQL
- Power BI
- Introduction to Data Science

Math 318 Course Outline

Course Admin Information

- Instructor: Dr. Xuemao Zhang
  - Office: SciTech Rm 128
  - Email: xzhang2@esu.edu
- Lecture Hours:
  - MWF: 12:00–12:50Pm (Science & Technology Room 145)
- Department Secretary: Christine Getz
  - Office: SciTech Rm 118
  - Email: cgetz@esu.edu
  - Telephone: 570-422-3447

Course Objectives

- This course (as part of data science) will focus on data query, data cleaning/wrangling and visualizations using SQL (Structured Query Language), R and Power BI.
- **Programming:**
  - SQL with PostgreSQL and pgAdmin
  - R and Rstudio
- **You will learn:**
  - Data query using SQL
  - Managing data: data import/export and data cleaning/wrangling using R
  - Communicate effectively using statistical graphics using R and Power BI
    - ★ Bar charts, histograms, box-plots, Error bars, scatter plots, clustering and dimension reduction
    - ★ BI with dashboard https://www.youtube.com/watch?v=yKTSLffVGbk
  - Types of data visualizations:
    - ★ Static visualizations
    - ★ Animated graphics
    - ★ Interactive visualizations
    - ★ Visualization of spatial data
- **Prerequisites:** Elementary Statistics (Math 110 or similar courses)

Tentative Contentes

- Introduction to data science and EDA
- Introduction to base R
- Data wrangling using package `dplyr` and `tidyr`
- Sampling techniques and statistical inferences with R (optional)
- Data visualizations using R
  - Data visualization using `ggplot2`
  - Animated graphics using `magick` and `gganimate`
  - Interactive visualizations using `plotly`
  - Visualization of spatial data using `maps`, `usmap` and `tmap`
  - Building dashboard and web apps
- Basic SQL queries
- Business intelligence using Power BI, SQL and R

Assessments

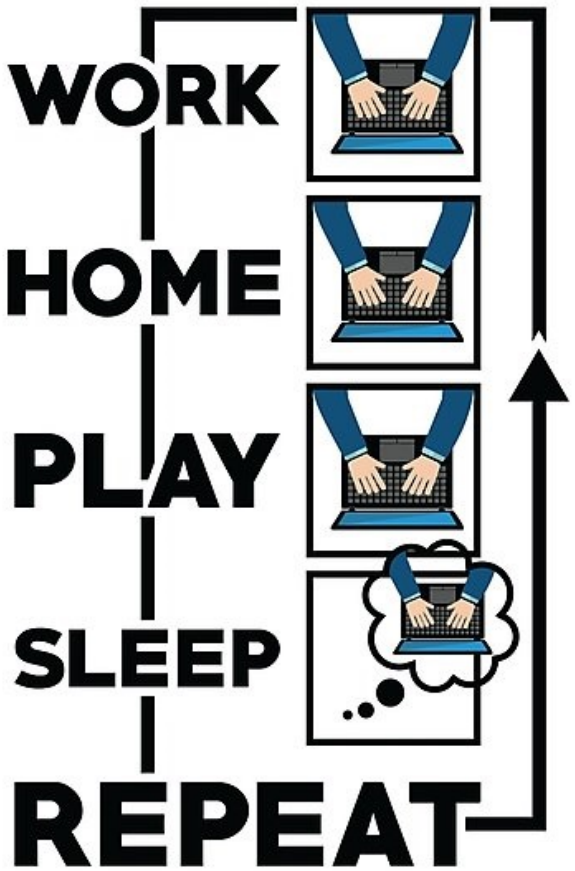
- It is a skill-development course.
  - 20% in-class quizzes (4 sets)
  - 30% Homework Assignments (6 sets)
  - 30% Projects (2 sets)
  - 20% Final project, consisting of
    - Oral presentation: 5%
    - Written report: 15%

R References

- Venables, W. N., Smith, D. M. and the R Core Team (2020). *An Introduction to R*. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Grolemund, G. and Wickham, H.(2017 O'Reilly). *R for Data Science*. http://r4ds.had.co.nz/
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd). Springer. http://had.co.nz/ggplot2/
- Xie, Y. (2018). *R Markdown: The Definitive Guide*. https://bookdown.org/yihui/rmarkdown/
- Xie, Y., Dervieux, C. and Riederer, E. (2022). *R Markdown Cookbook*. https://bookdown.org/yihui/rmarkdown-cookbook/
- RStudio Cheat Sheets. https://www.rstudio.com/resources/cheatsheets/

Why programming?

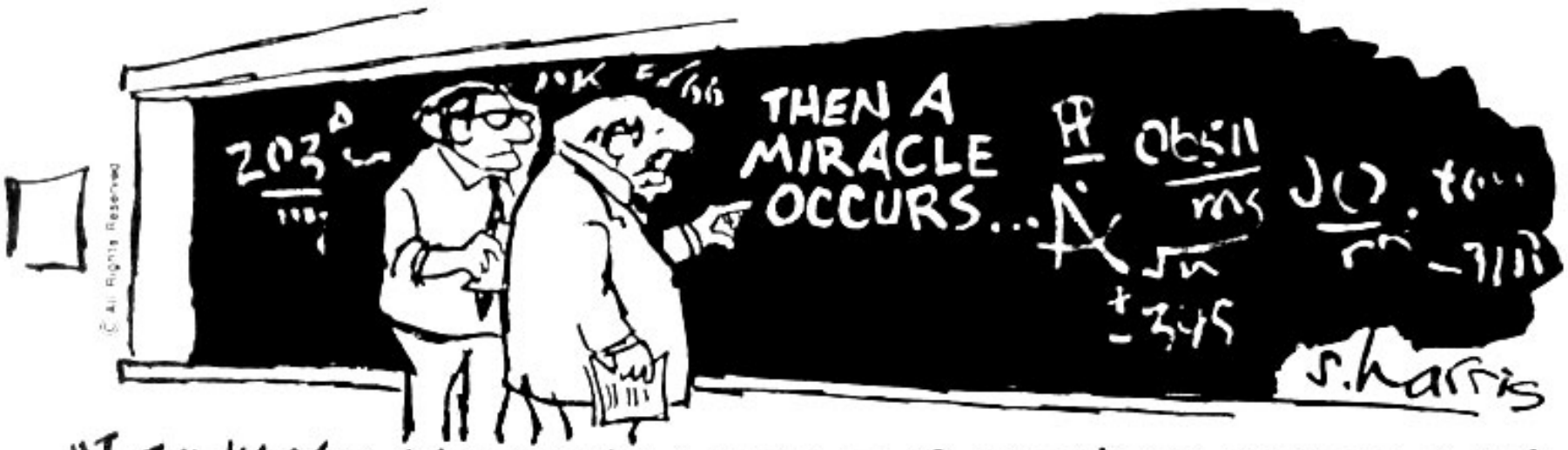
- To be able to easily repeat your own work.



Source: https://www.redbubble.com

Why programming?

- The workflow of using a script makes your research reproducible.



Source: Malanris.ru

Why programming?

- Python, R and SAS are the main data analytics tools which require programming.
- SQL (Structured Query Language) is a programming language to manage relational databases
- Programming isn't scary. If you've written formulas in Excel, you've already done "programming".

R

- It's a software environment for statistical computing and graphics, free and open source.
- It is availalbe for three plat forms: Linux, (Mac) OS X, and Windows.

R: https://www.r-project.org/

RStudio(an IDE, integrated development environment, for R): https://www.rstudio.com/

R

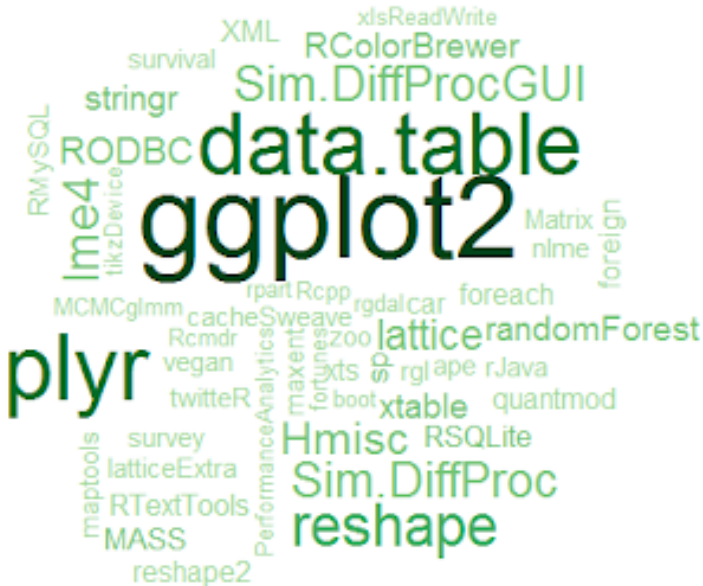
- It's designed to analyze data. The spreadsheet-like data structure "data frame" makes it easy to apply calculations.

```
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),
   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15",
        "2014-05-11", "2015-03-27")), stringsAsFactors = FALSE)
print(emp.data)
```

```
##   emp_id emp_name salary start_date
## 1      1     Rick 623.30 2012-01-01
## 2      2      Dan 515.20 2013-09-23
## 3      3 Michelle 611.00 2014-11-15
## 4      4     Ryan 729.00 2014-05-11
## 5      5     Gary 843.25 2015-03-27
```

R

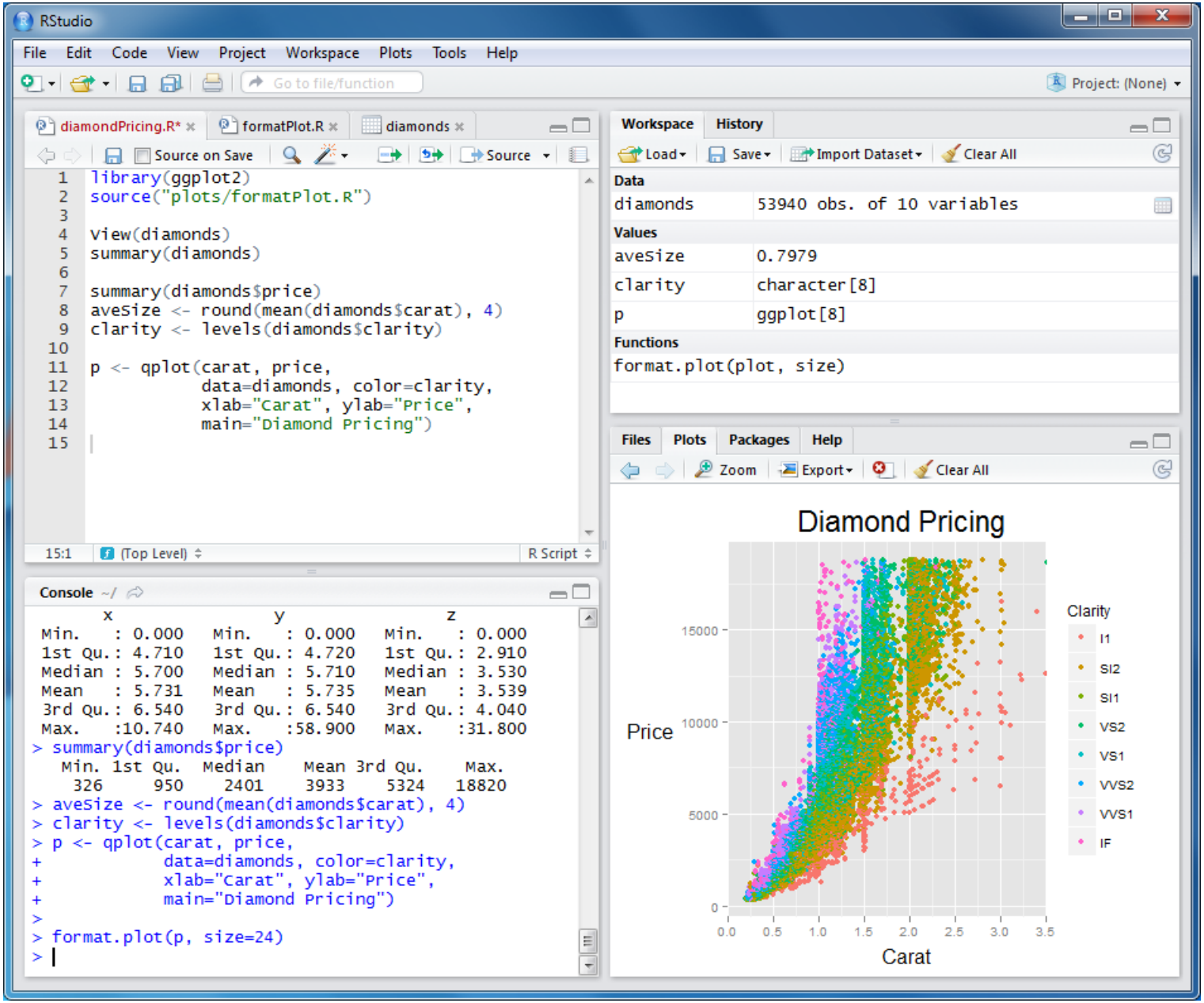
- There is a large and growing community of R users. For example, DataCarpentry(https://www.datacarpentry.org/), Software Carpentry (https://software-carpentry.org/), R-bloggers (https://www.r-bloggers.com/) and ROpenSci(https://ropensci.org/community/).
- The functionality of R is extended by more than 18,000 user-contributed packages.



RStudio IDE

- RStudio is a popular IDE (Integrated Development Environment) for R programming
- It is a powerful editor for R coding and debugging.
- It is a powerful generator for HTML, PDF, dynamic documents and slide shows.
- RStudio can be run on both Desktop and Cloud (https://rstudio.cloud/).
  - You may use Rstudio Cloud for the first several weeks if you have difficulties installing R and Rstudio.
- Check out more nice features of RStudio at its official website(https://www.rstudio.com/products/rstudio/features/)

RStudio IDE


RStudio
File Edit Code View Project Workspace Plots Tools Help
diamondPricing.R*
formatPlot.R
diamonds
Source on Save
Source
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15
15:1 (Top Level) R Script
Console
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 950 2401 3933 5324 18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")
>
> format.plot(p, size=24)
>
Workspace History
Load Save Import Dataset Clear All
Data
diamonds 53940 obs. of 10 variables
Values
aveSize 0.7979
clarity character[8]
p ggplot[8]
Functions
format.plot(plot, size)
Files Plots Packages Help
Zoom Export Clear All
Diamond Pricing
Price
Carat
Clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
0
5000
10000
15000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

R Markdown (Demonstrations)

```
knitr::kable(head(iris), format = 'html')
```

- Dynamic documentation: report, table, graphics ...
- R packages by Yihui Xie: knitr, bookdown, xaringan, etc

```
plot(iris, col=iris$Species)
```

- Data-generated graphics that are reproducible

R Markdown

- Click here to view a fantastic micro-video tutorial
- Browse here for a gallery of creative Rmarkdown works

- We will see more about Rmarkdown in next lecture.

SQL

- SQL (Structured Query Language) is a standardized programming language that is used to manage relational databases and perform various operations on the data in them.
  - The term SQL is pronounced ess-kew-ell or sequel.
- What Can SQL do?
  - SQL can execute queries against a database
  - SQL can retrieve data from a database
  - SQL can insert records in a database
  - SQL can update records in a database
  - SQL can delete records from a database
  - SQL can create new databases
  - SQL can create new tables in a database
  - SQL can create stored procedures in a database
  - SQL can create views in a database
  - SQL can set permissions on tables, procedures, and views

SQL

- PostgreSQL

PostgreSQL 

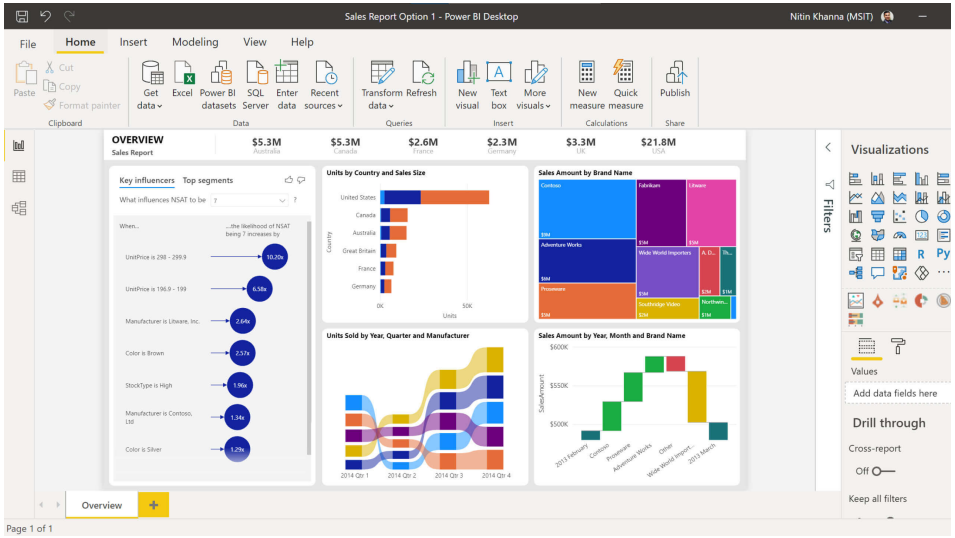
- It is a **free and open-source** relational database management system (RDBMS).
- pgAdmin: The pgAdmin package is a **free and open-source** graphical user interface (GUI) administration tool for PostgreSQL.
- Please install PostgreSQL and pgAdmin on your laptop following the youtube video PostgreSQL (Postgres) - Installation & Overview.

SQL

- Although SQL is an ANSI/ISO standard, there are different versions of the SQL language.
  - However, to be compliant with the ANSI standard, they all support at least the major commands (such as SELECT, UPDATE, DELETE, INSERT, WHERE) in a similar manner.
  - Most of the SQL database programs also have their own proprietary extensions in addition to the SQL standard!

Power BI

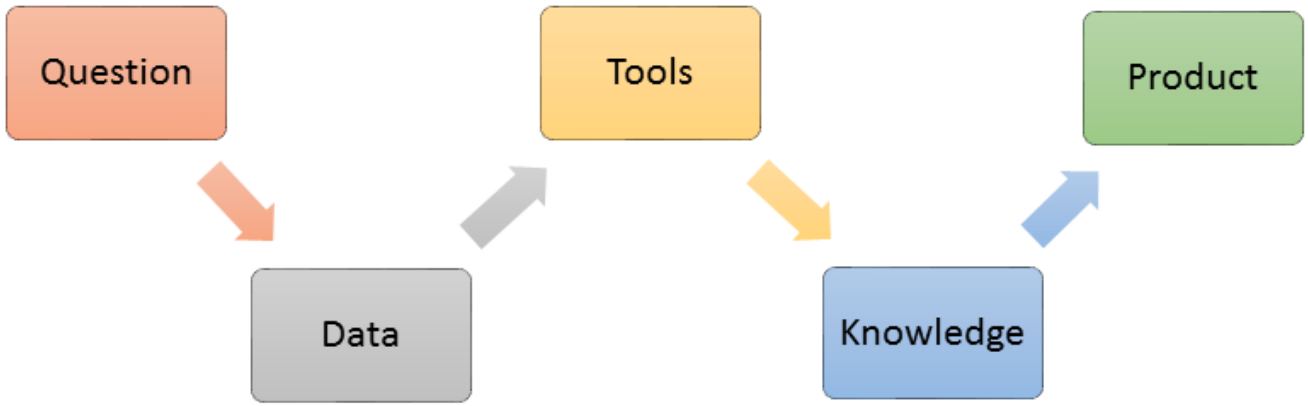
- Power BI



- Power BI is an **interactive** data visualization software product developed by Microsoft with primary focus on business intelligence.
  - Data may be input by reading directly from a **database**, webpage, or structured files such as spreadsheets, CSV, XML, and JSON.
  - The functionality can be extended by R and Python.
- Power BI Desktop is free.
- Please install Power BI to your laptop following the video How to install Power BI on Windows 10 64-bit.

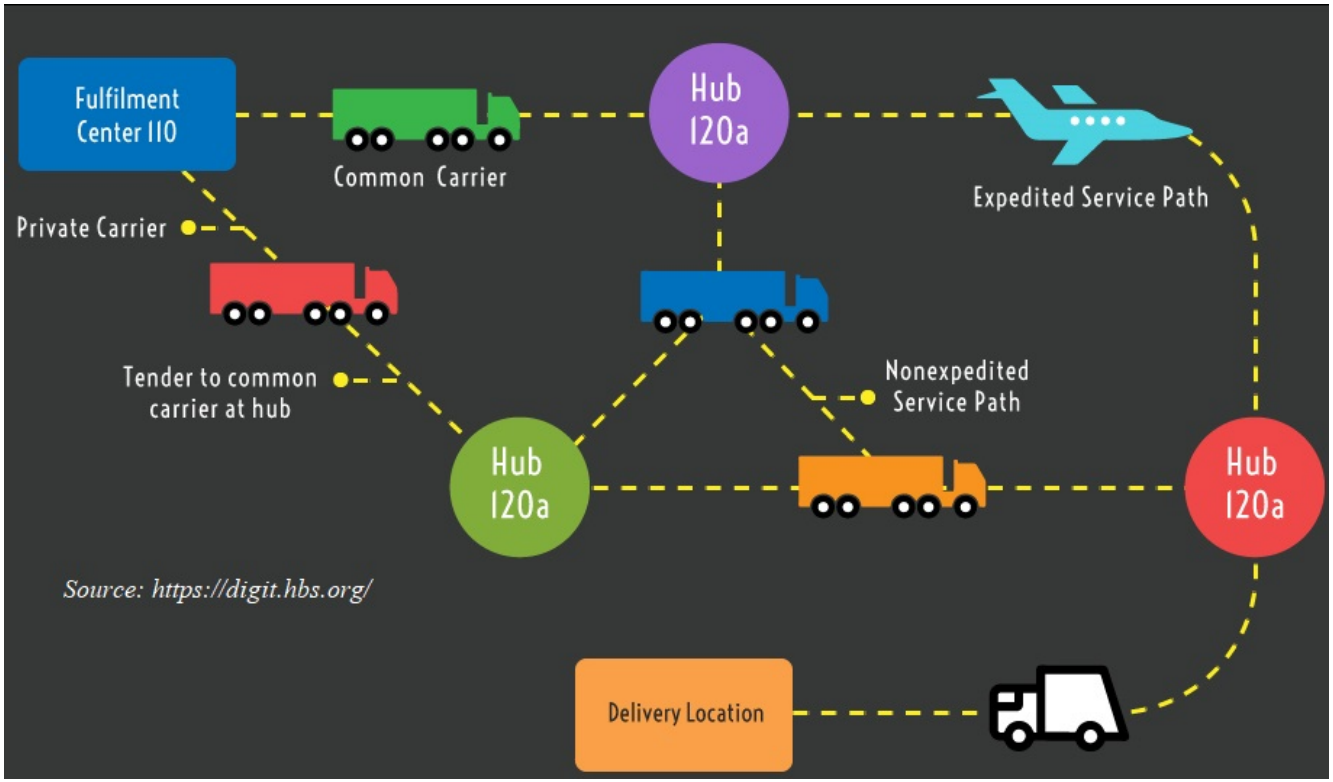
Introduction to data science

- Data science is the study of large sets of data, using computers to look for patterns and trends.



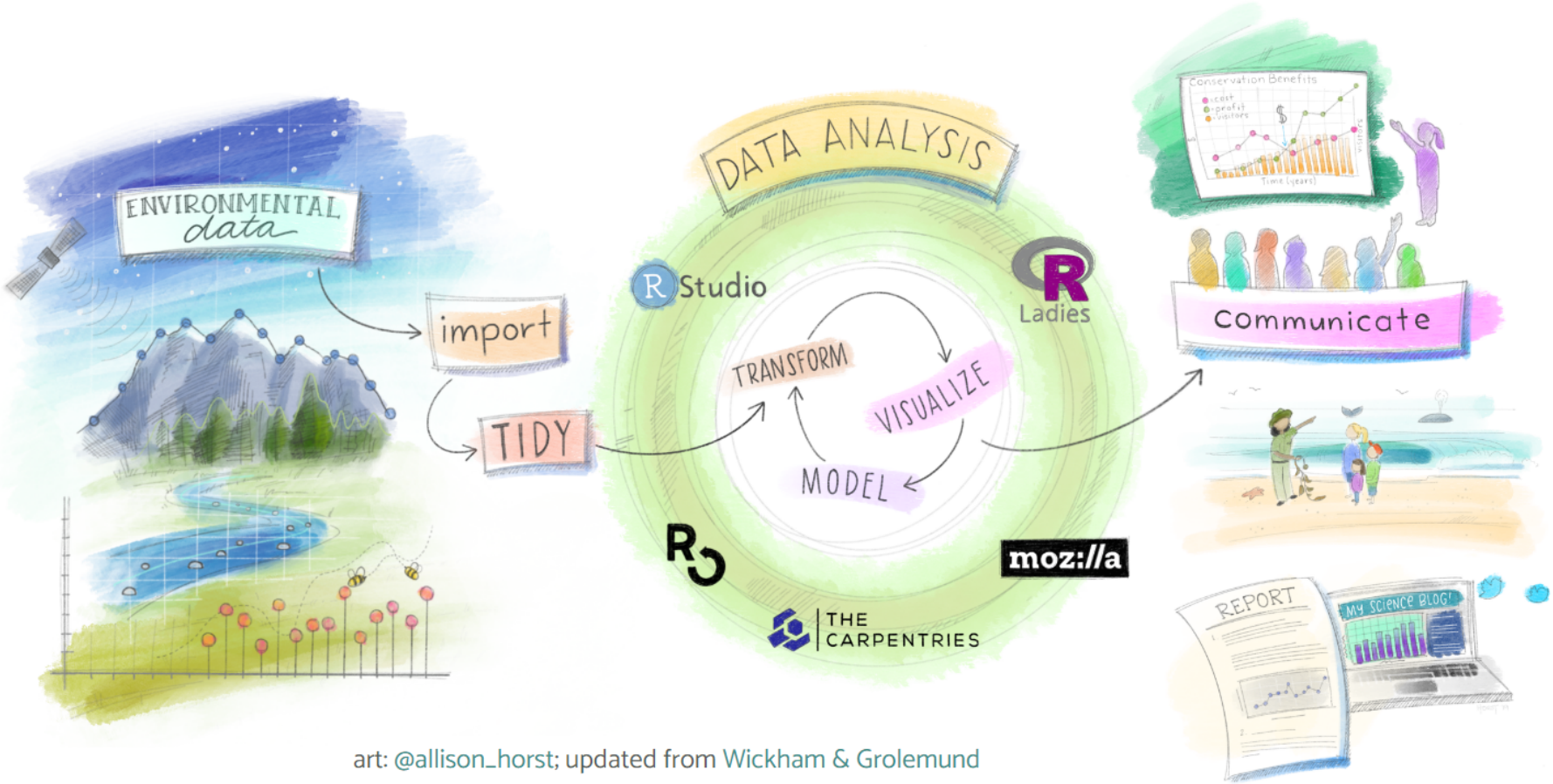
Introduction to data science

- For example, Amazon uses Big Data gathered from customers to predict consumer shopping behavior: when customers will buy items, which items they will buy, when and where the items will ship? Amazon ships your items before you order it using an "anticipatory shipping" system!



Introduction to data science

- Data science is the discipline of turning raw data into understanding



art: @allison_horst; updated from Wickham & Grolemund

Data Scientist The Sexy Job

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

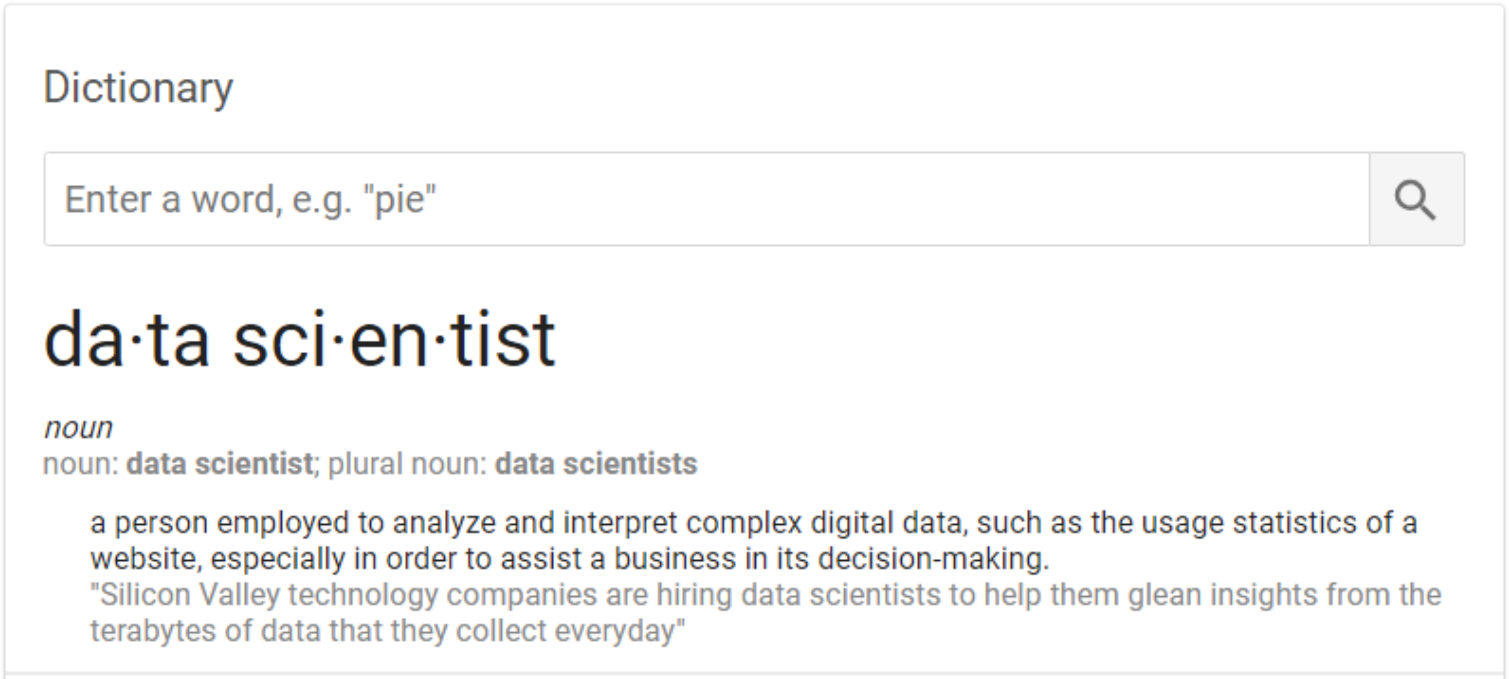
- See also an old article by NYT (2009): For Today's Graduate, Just One Word: Statistics
- And another famous McKinseay 2011 Report: Big data: The next frontier for innovation, competition, and productivity
- Is Data Scientist Still the Sexiest Job of the 21st Century?

Job Market in Data Science or Data Analytics

Search Data Analytics or Data Scientist on https://www.indeed.com/

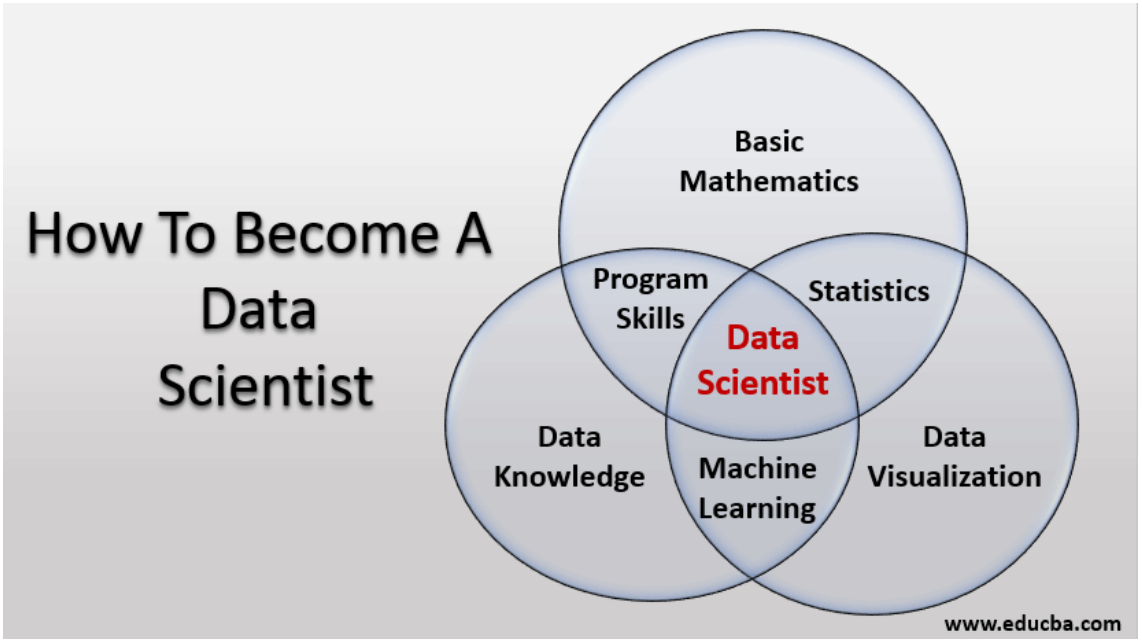
What is a data scientist?

- "A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician." (from Joshua Blumenstock, 2013).



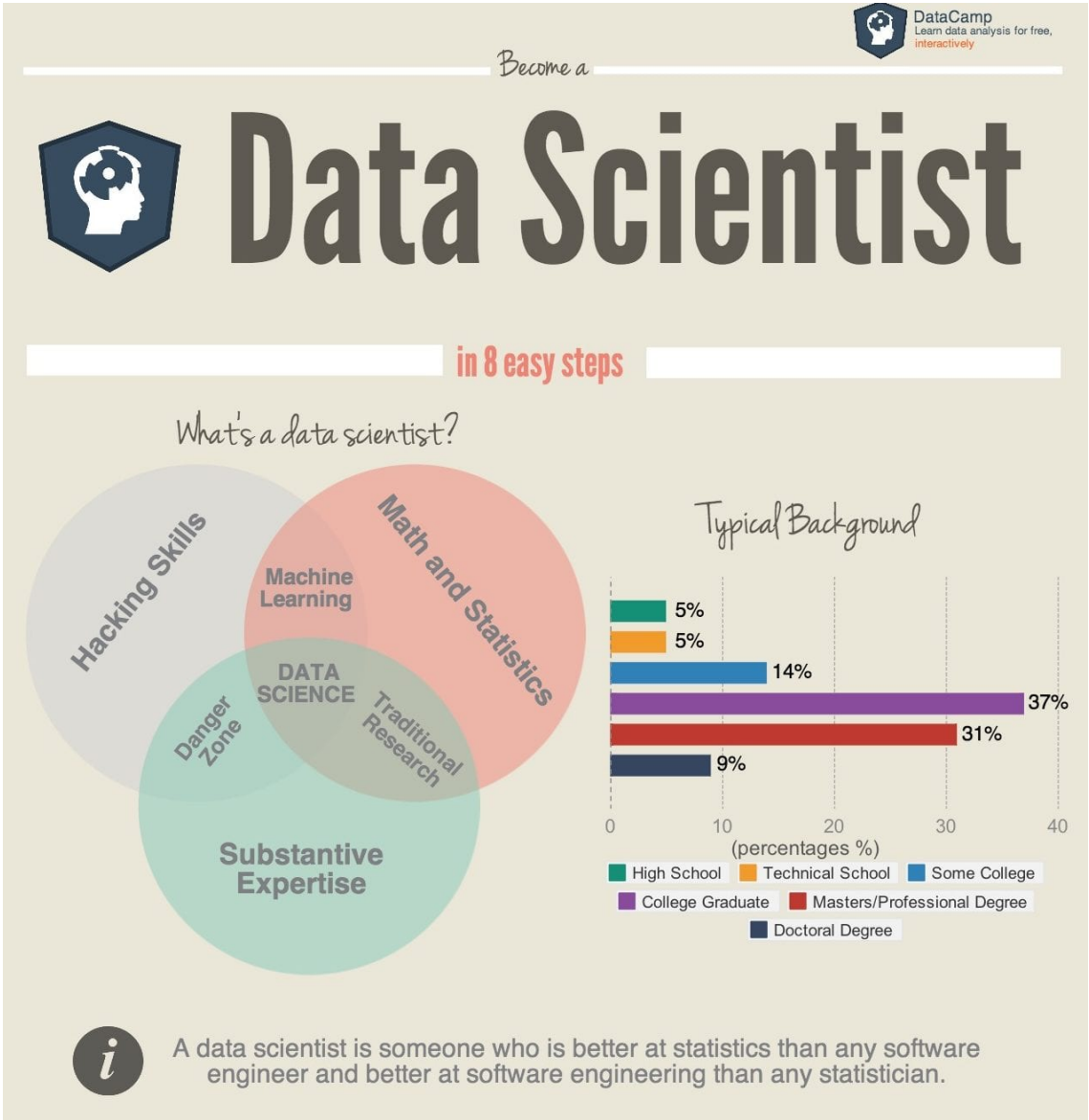
How to become a data scientist

- An article on coursera: How to Become a Data Scientist?
- How to Become a Data Scientist



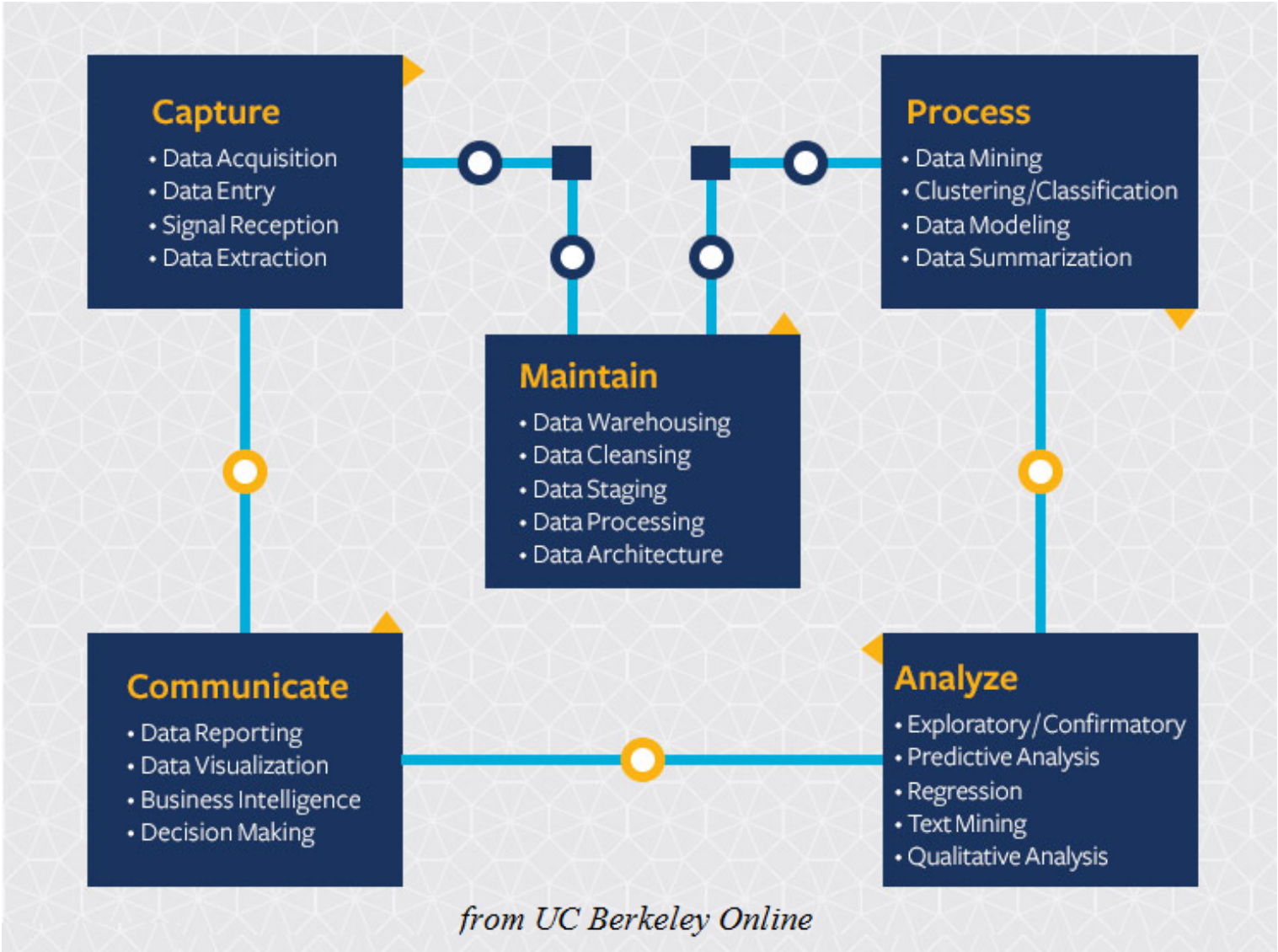
How to become a data scientist

- Become A Data Scientist in 8 Steps: Infographic



Tools in data science

- The Data Science Life Cycle



Programming languages

The 10 Best Data Science Programming Languages to Learn in 2021

5 Best Data Analysis Programming Languages in 2022

- Programming languages for data analysis:
  - R
  - Python
  - SAS
  - Julie
- General purpose programming languages
  - Python
  - Java
  - C/C++
  - Scala
  - Julie
- Databases query: SQL(structured query language)
- Web app development: JavaScript

Data visualization tools

8 Best Data Visualization Tools that Every Data Scientist Should Know

- Microsoft Power BI
- Tableau
- Plotly (free library in R and Python) and Dash (commercial product by https://plotly.com/)
- SAS Visual Analytics
- ⋮
- Excel with VBA

Tools for big data

Top 7 Big Data Analytics Tools | Its Technology And Techniques

- Apache Airflow: "Airflow is a platform to programmatically author, schedule and monitor workflows." — Airflow documentation
- Apache Hadoop
- Apache Spark
- MongoDB
- RapidMiner
- Microsoft Azure
- Zoho Analytics

License