WORKSHOP ON DATA SCIENCE & MACHINE LEARNING USING R

Tue Vu (PhD)

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CYBER-INFRASTRUCTURE & TECHNOLOGY INTEGRATION (CITI)

CLEMSON COMPUTING & INFORMATION TECHNOLOGY (CCIT)

Fall 2019

CITI cyberinfrastructure technology integration

CLEMSON COMPUTING AND INFORMATION TECHNOLOGY

INTRODUCTION TO

USING

Tue Vu (PhD)

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CYBER-INFRASTRUCTURE & TECHNOLOGY INTEGRATION (CITI)

CLEMSON COMPUTING & INFORMATION TECHNOLOGY (CCIT)

CITI
cyberinfrastructure Technology Integration

CLEMSON
COMPUTING AND
INFORMATION TECHNOLOGY

Wednesday (09/18)	Friday (09/20)
- Data Science - Basic R - Vectors/Matrices - Inputs/Outputs - Control structure (If end, For loop) - Function	- Advanced R - Install packages - R Profiling - Parallel computing - R in HPC - Plotting system - Basic & ggplots

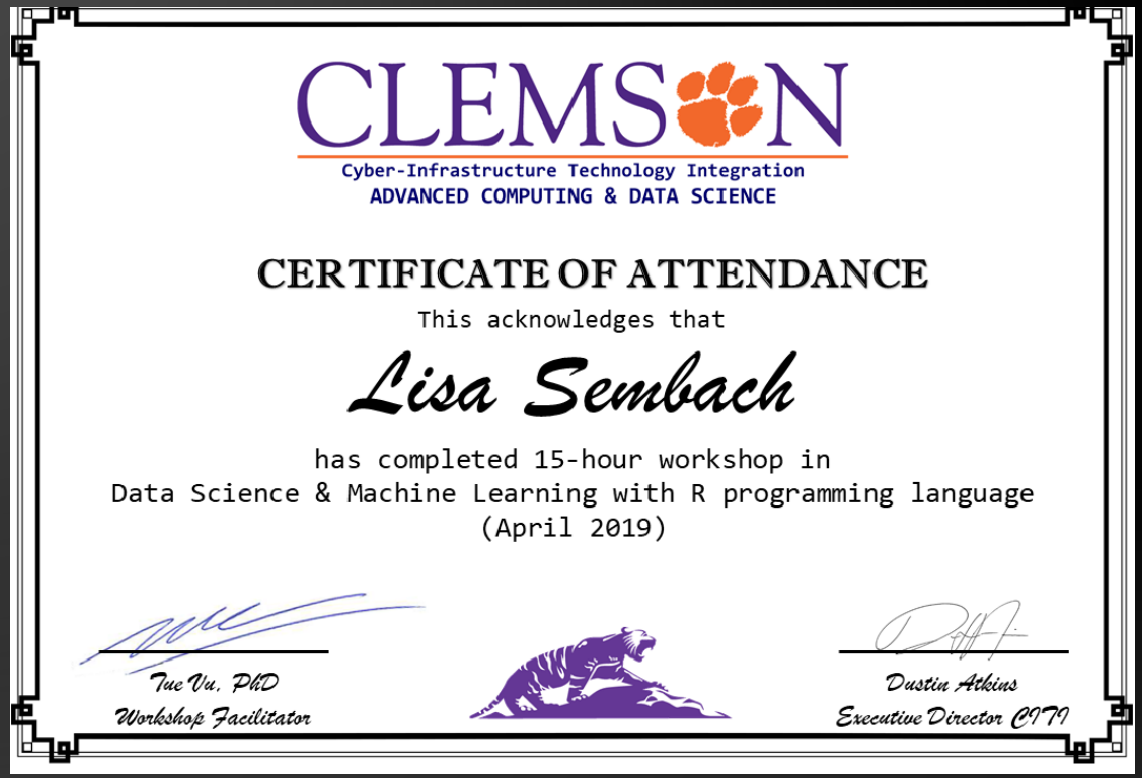
Tuesday (06/18)	Thursday (06/20)
- Data Science - Basic R - Vectors/Matrices - Inputs/Outputs - Control structure (If end, For loop) - Function	- Advanced R - R Profiling - Parallel computing - R in HPC - Plotting system - Basic & ggplots

+

Tuesday (06/25)	Thursday (06/27)
- Introduction to Machine Learning - Why R for Machine Learning? - Types of ML - Caret Package - Supervised Learning (Regression, Decision Tree, Random Forest)	- Supervised Learning (Ensemble prediction, Model based prediction, Regularization & variable selection) - Dimension Reduction - Neural Network - Support Vector Machine - K-Nearest Neighbour - Unsupervised Learning (K-means clustering, Gap-Statistic)

+

ML Projects

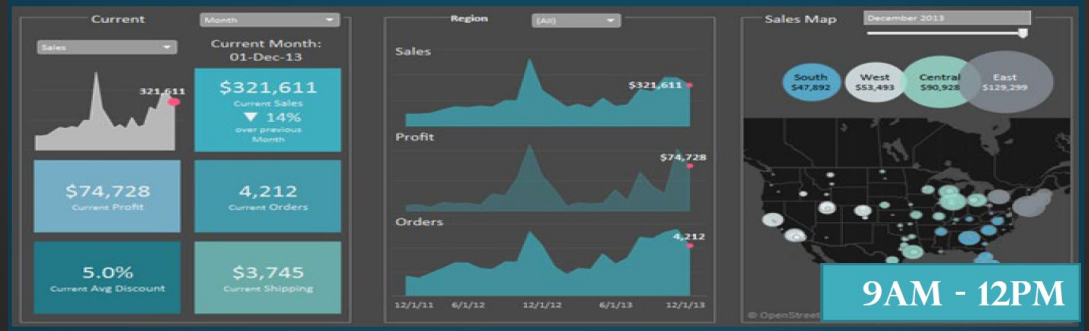


Data Visualization with Tableau

Friday, October 25

Free workshops for students, faculty, and staff on state-of-the-art technologies in data visualization and spatial analytics. No previous experience required!

Morning Session: Tableau Basics



Discover the power of data analytics and visual design with Tableau. Learn how to make basic charts, create interactive dashboards, and publish them online to Tableau public.

Afternoon session: Using R & Python

Learn how to integrate statistical analysis using R. Leverage Tableau and Python to build advanced analytics.

CLEMSON CENTER FOR GEOSPATIAL TECHNOLOGIES

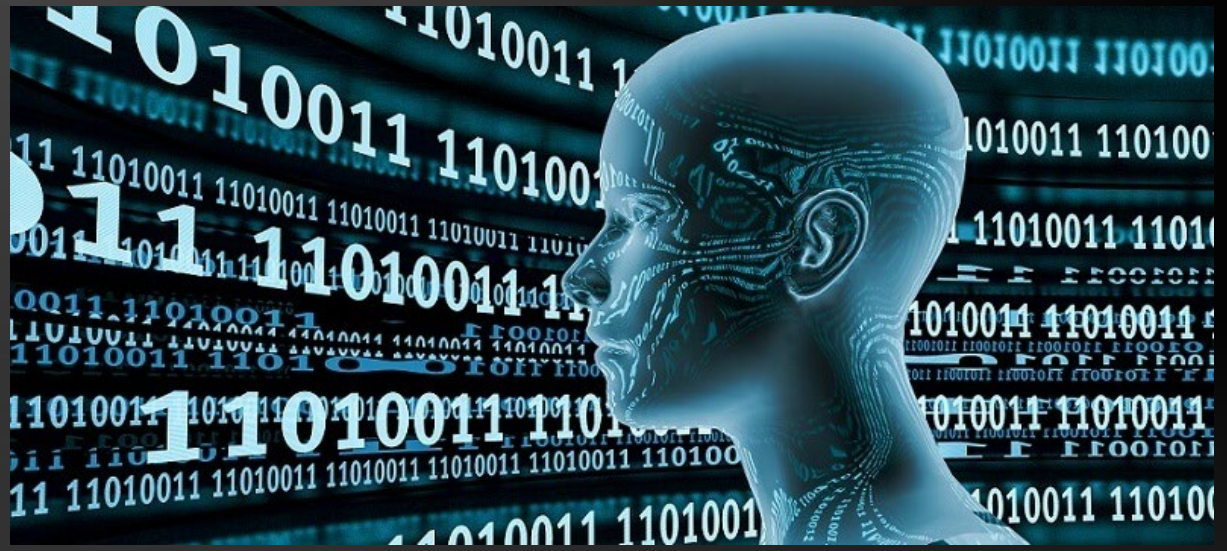
Register: **bit.ly/cutableaubasic**
bit.ly/cutableauadvanced
406A Cooper Library

CITI
cyberinfrastructure Technology Integration

CLEMSON
COMPUTING AND INFORMATION TECHNOLOGY

OUTLINES

1. Introduction to Data Science
2. What is Data Scientist
3. Hands-on R

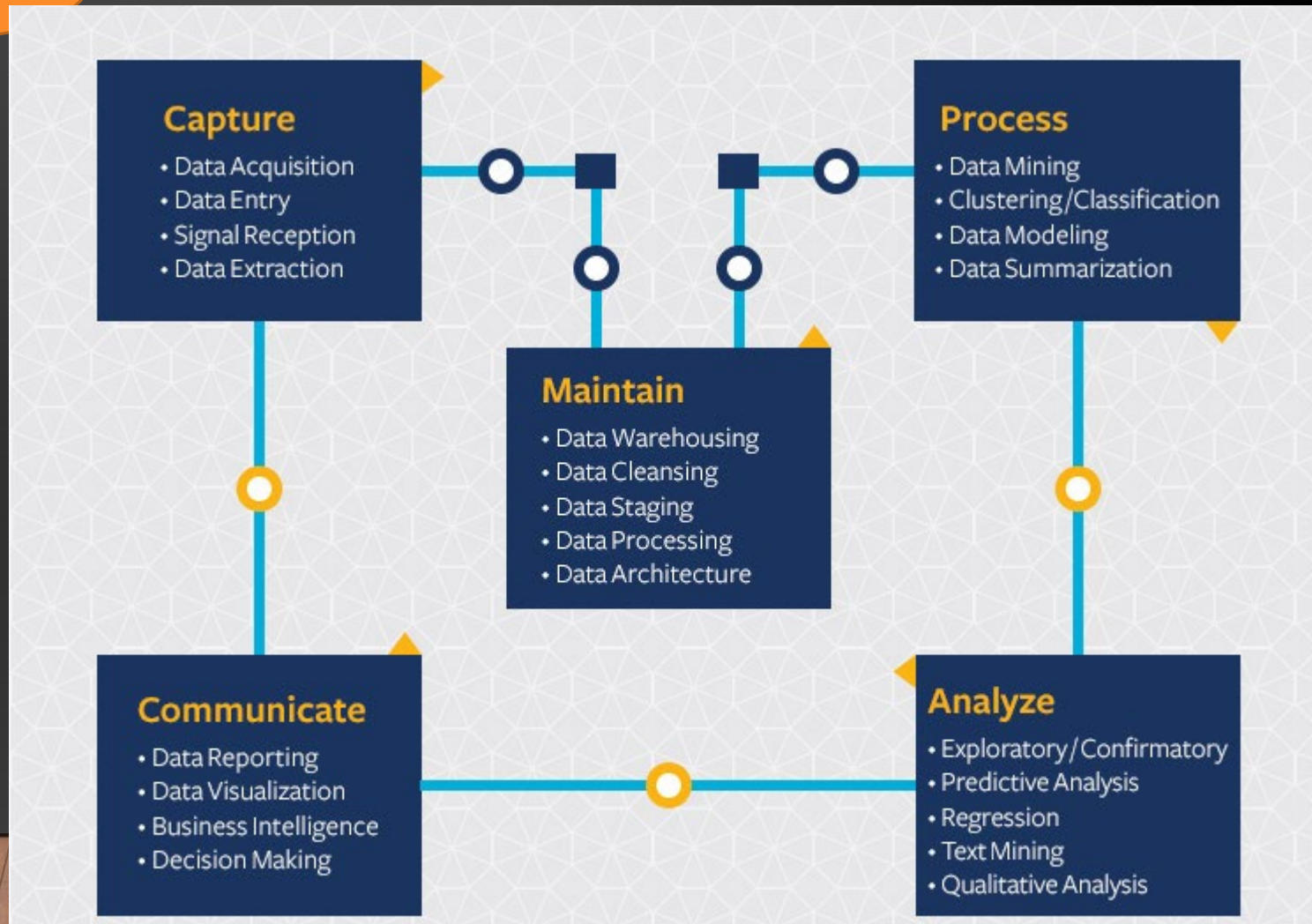


- The keyword for Data Science?

- Data science is only useful when the data are used to answer a question.
- That is the ***Science*** part of the equation.

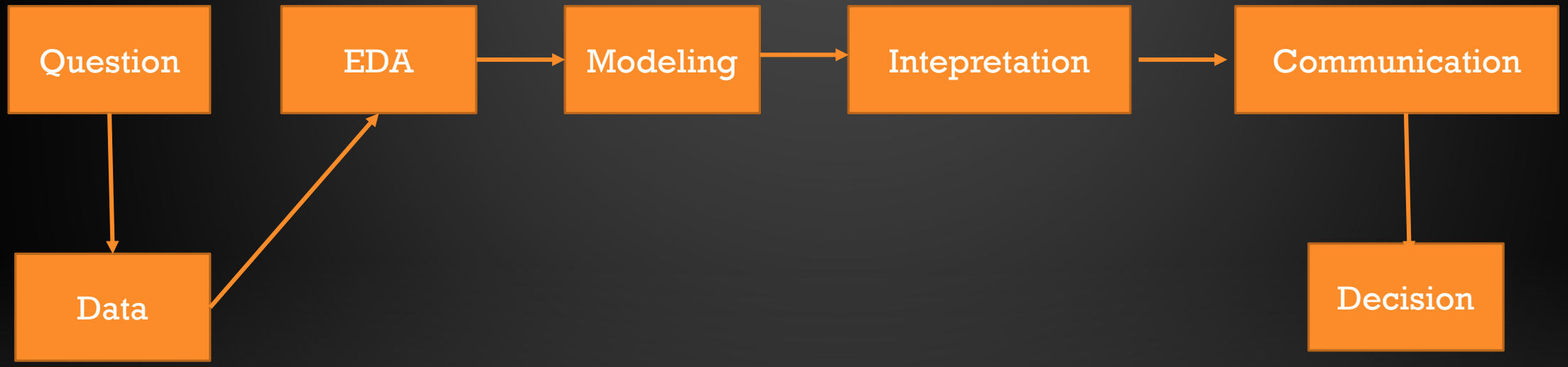
1. Introduction to Data Science

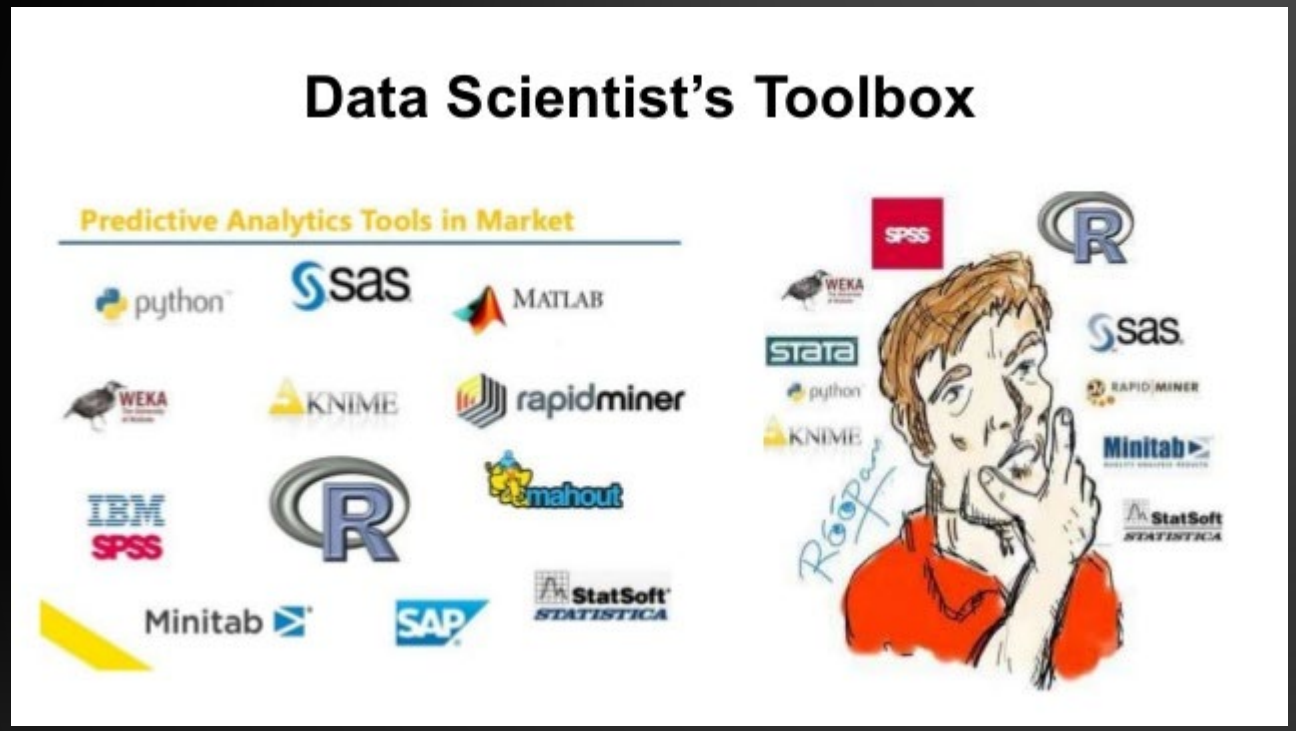
What is Data Science?



datascience.berkeley.edu

- Structure of Data Science Project



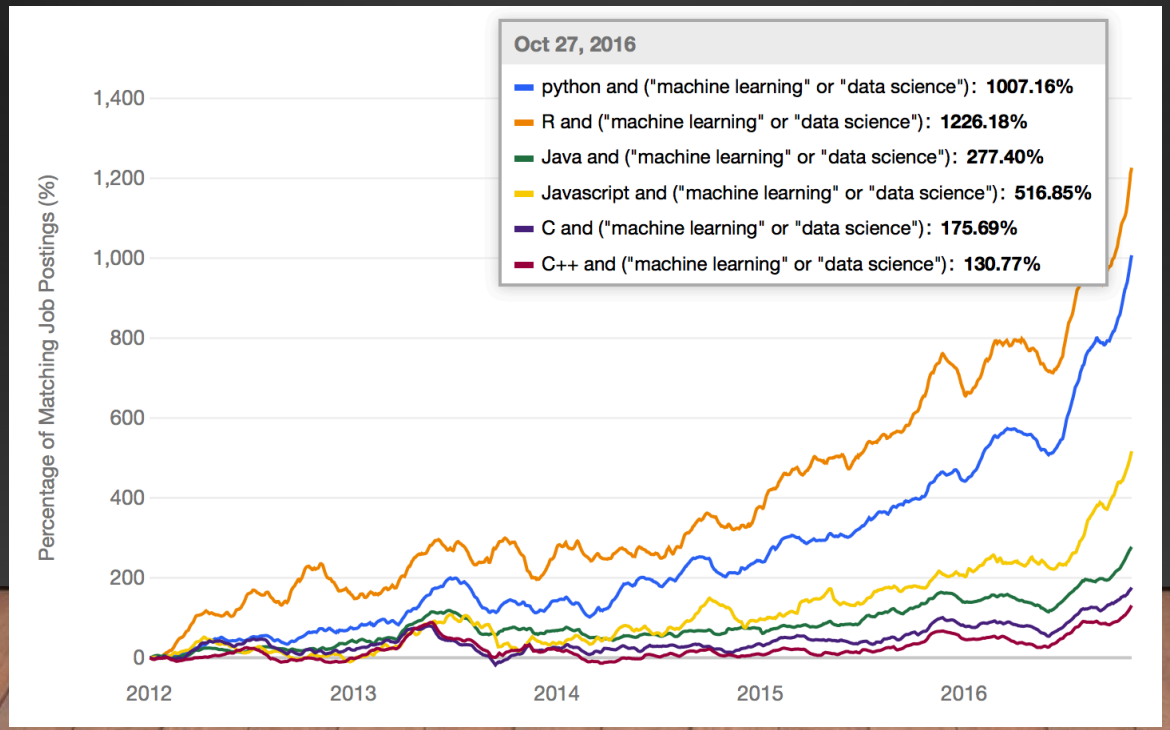


There are two common languages for Data Science

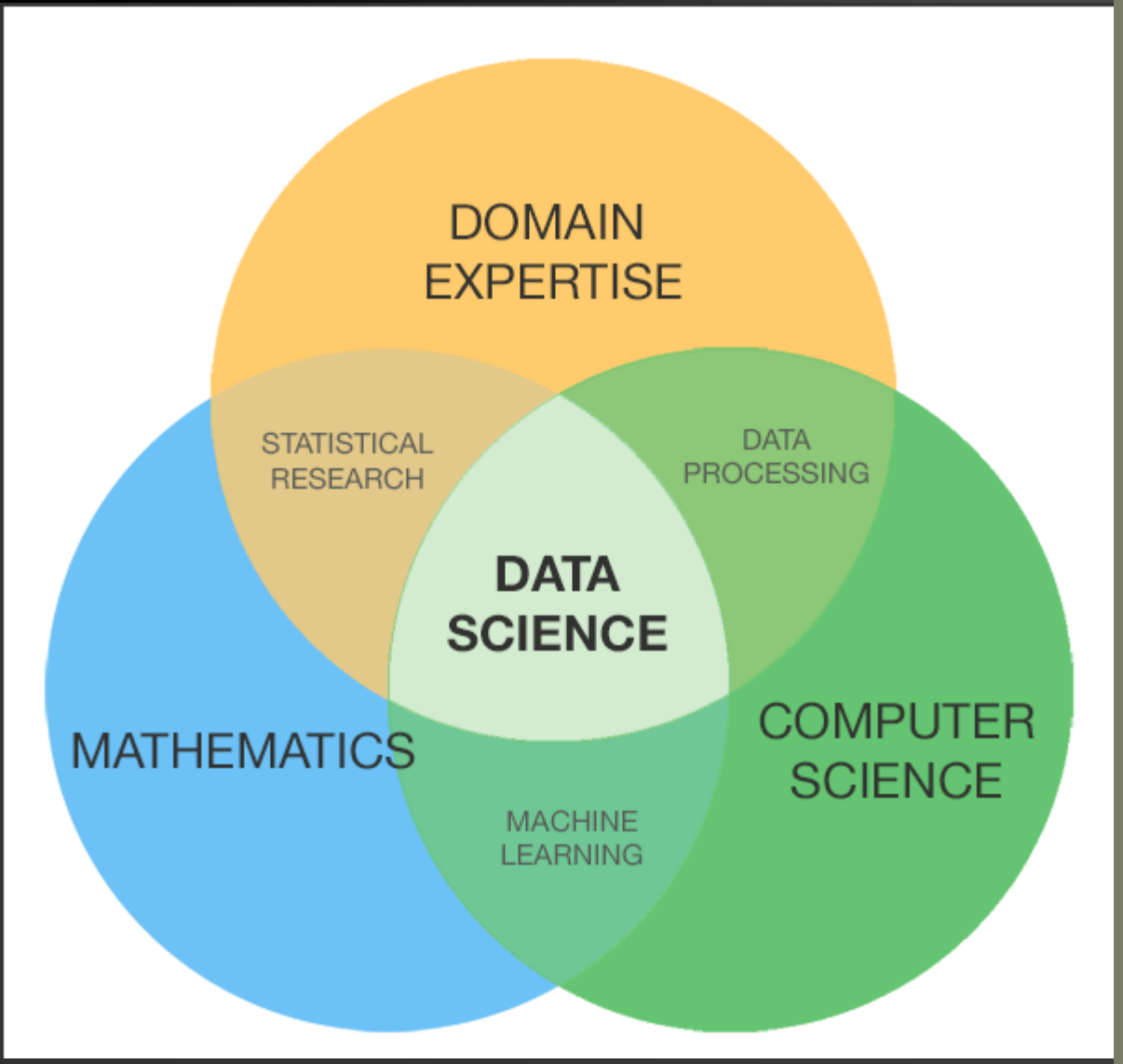
1. R
2. Python

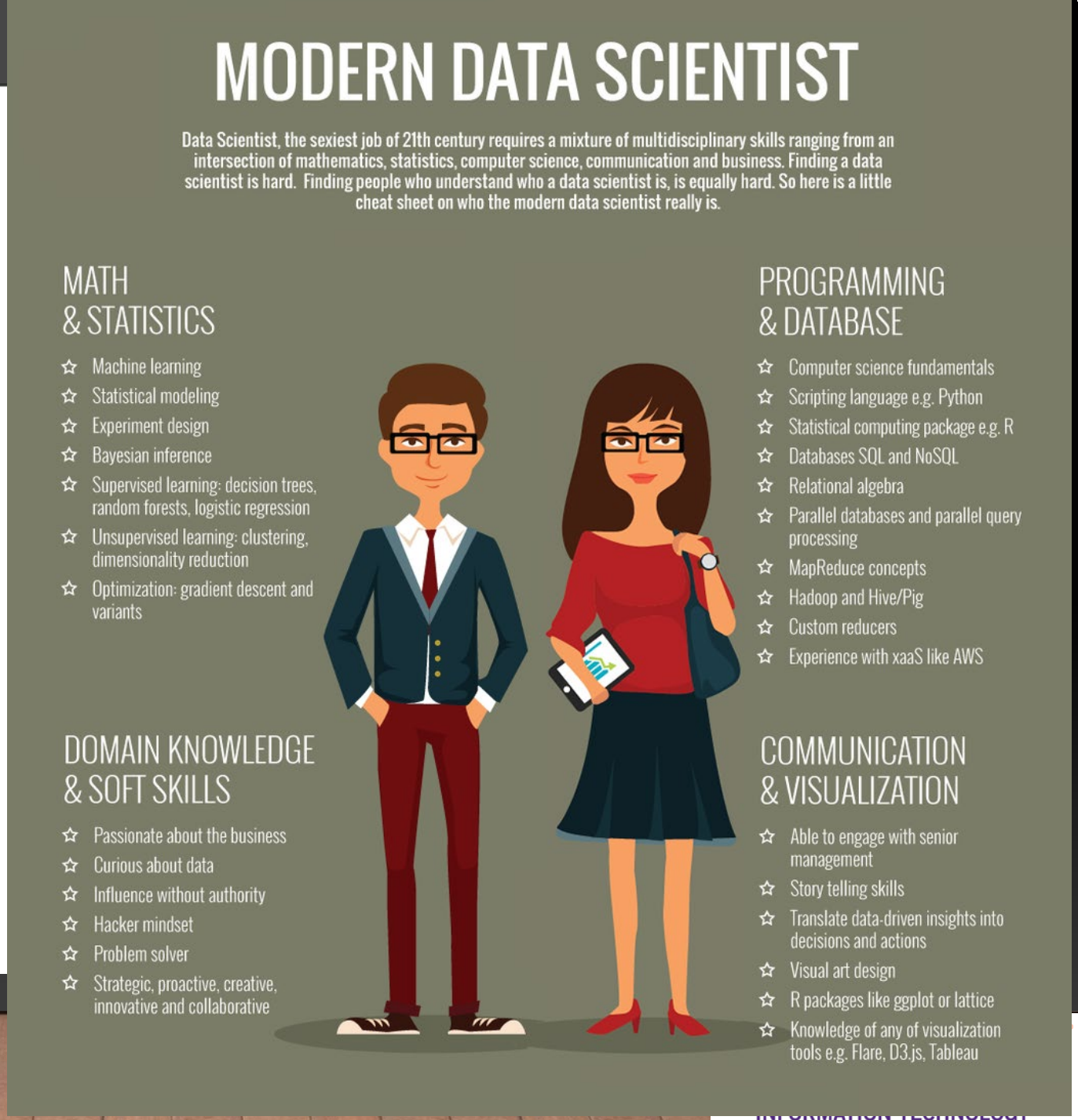
2. Why R?

- **R is used by the best data scientists in the world**. In surveys on Kaggle (the competitive machine learning platform), R is by far the most used machine learning tool.
- **R is powerful because of the breadth of techniques it offers**. The platform has more techniques than any other that you will come across.
- **R is state-of-the-art because it is used by academics/researches**. One of the reasons why R has so many techniques is because academics that develop new algorithms are developing them in R and releasing them as R packages. This means that you can get access to state-of-the-art algorithms in R before other platforms.
- **R is free because it is open source software**. You can download it right now for free and it runs on any workstation platform you are likely to use.



3. Data Scientist





HANDS ON TRAINING WITH R

PLOTTING SYSTEM

Ggplot2: the grammar of graphics

The ggplot works on the philosophy of adding layers to the visualization to visualize your data effectively.

It has 7-layers grammatical elements as shown below:

Ggplot2: the grammar of graphics

```
> library(ggplot2)
> ggplot(data=mtcars)
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt))
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt))+
        geom_point()
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt)) +
        geom_point() +
        facet_grid(gear ~ .)
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt)) +
        geom_point() +
        facet_grid(gear ~ .) +
        stat_smooth()
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt)) +
        geom_point() +
        facet_grid(gear ~ .) +
        stat_smooth() +
        coord_cartesian(xlim = c(13, 30))
```

Ggplot2: the grammar of graphics

```
> ggplot(data=mtcars, aes(x=mpg, y=wt)) +
        geom_point() +
        facet_grid(gear ~ .) +
        stat_smooth() +
        coord_cartesian(xlim = c(13, 30)) +
        theme_dark()
```