

WORKSHOP ON DATA SCIENCE & MACHINE LEARNING USING

Tue Vu (PhD)

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CYBER-INFRASTRUCTURE & TECHNOLOGY INTEGRATION (CITI)

CLEMSON COMPUTING & INFORMATION TECHNOLOGY (CCIT)

Fall 2019

INTRODUCTION TO



USING



Tue Vu (PhD)

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CYBER-INFRASTRUCTURE & TECHNOLOGY INTEGRATION (CITI)

CLEMSON COMPUTING & INFORMATION TECHNOLOGY (CCIT)

Wednesday (09/18)	Friday (09/20)
<ul style="list-style-type: none"> - Data Science - Basic R - Vectors/Matrices - Inputs/Outputs - Control structure (If end, For loop) - Function 	<ul style="list-style-type: none"> - Advanced R - Install packages - R Profiling - Parallel computing - R in HPC - Plotting system - Basic & ggplots

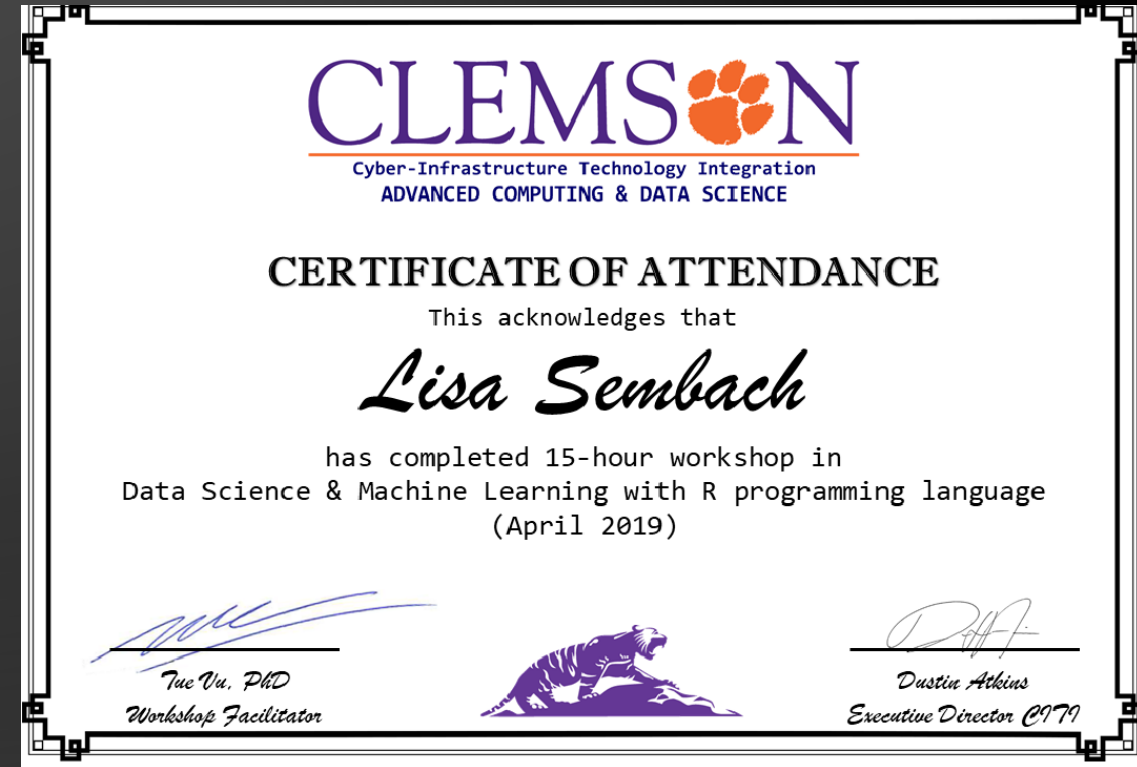
Tuesday (06/18)	Thursday (06/20)
<ul style="list-style-type: none"> - Data Science - Basic R - Vectors/Matrices - Inputs/Outputs - Control structure (If end, For loop) - Function 	<ul style="list-style-type: none"> - Advanced R - R Profiling - Parallel computing - R in HPC - Plotting system - Basic & ggplots



Tuesday (06/25)	Thursday (06/27)
<ul style="list-style-type: none"> - Introduction to Machine Learning - Why R for Machine Learning? - Types of ML - Caret Package - Supervised Learning (Regression, Decision Tree, Random Forest) 	<ul style="list-style-type: none"> - 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

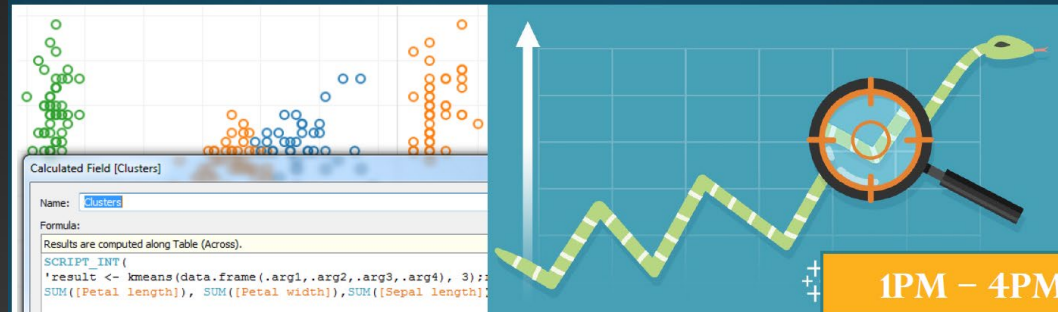
Free workshops for students, faculty, and staff on state-of-the-art technologies in data visualization and spatial analytics. No previous experience required!

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.

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

OUTLINES

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



1. Introduction to Data Science

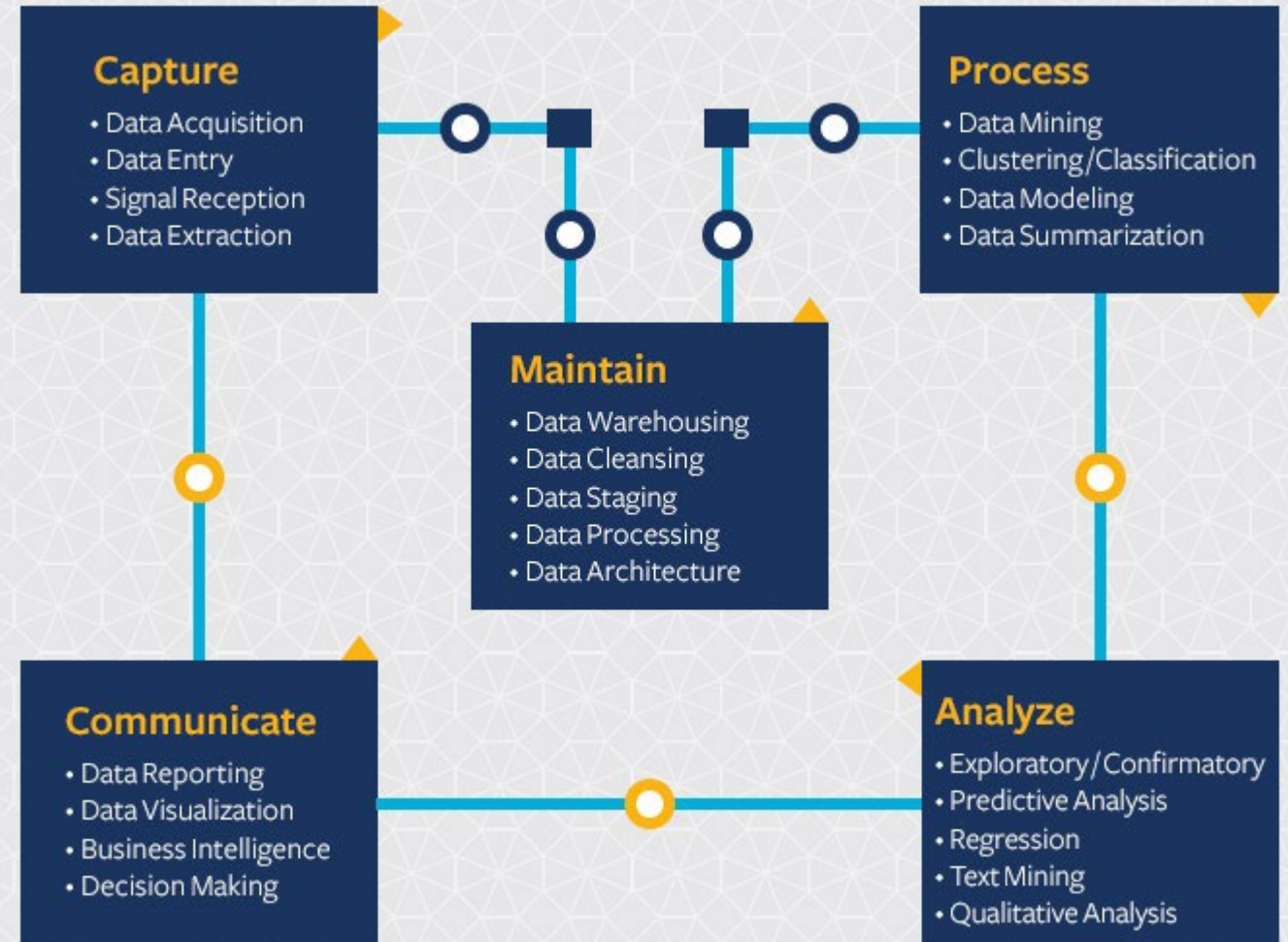
- The keyword for Data Science?



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

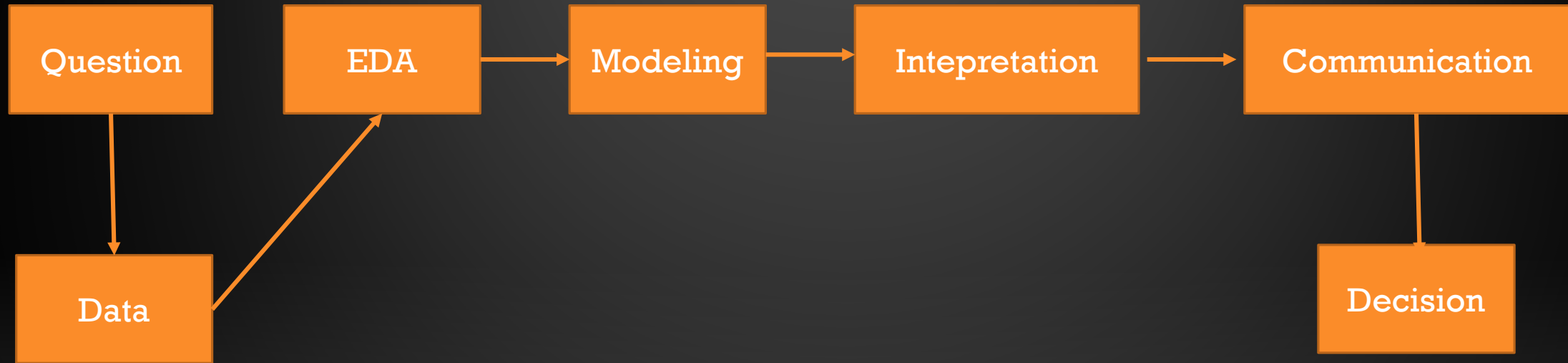
1. Introduction to Data Science

What is Data Science?



1. Introduction to Data Science

- Structure of Data Science Project



1. Introduction to Data Science

Data Scientist's Toolbox

Predictive Analytics Tools in Market

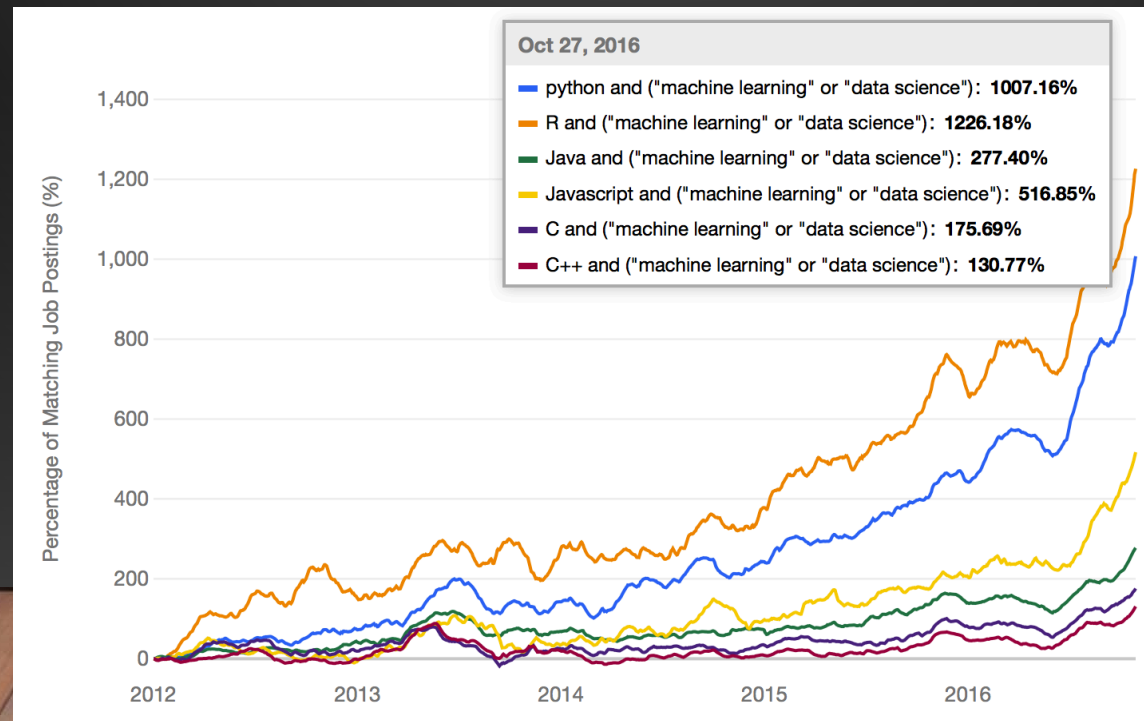


There are two common languages for Data Science

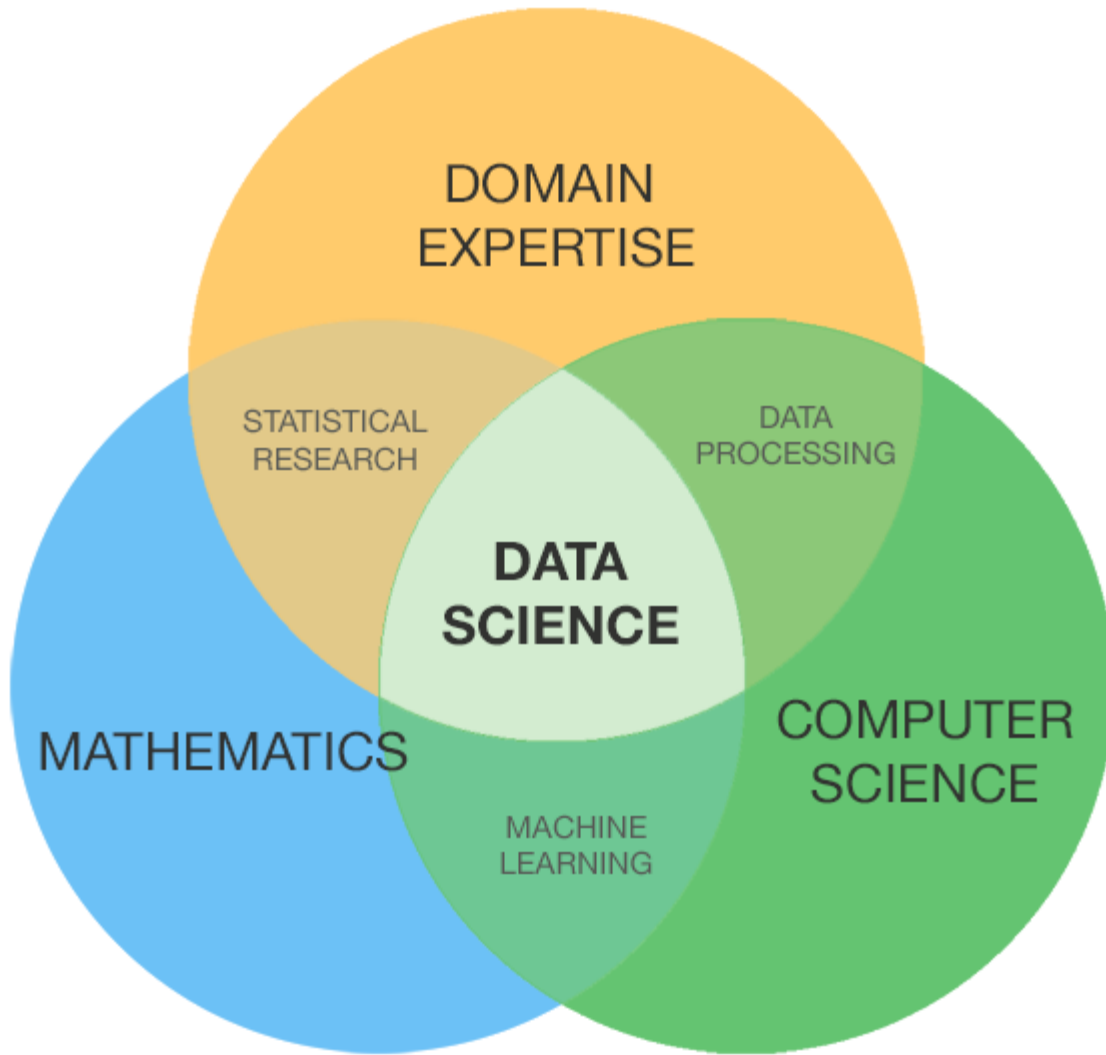
1. R
2. Python

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



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



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 ~ .)
```



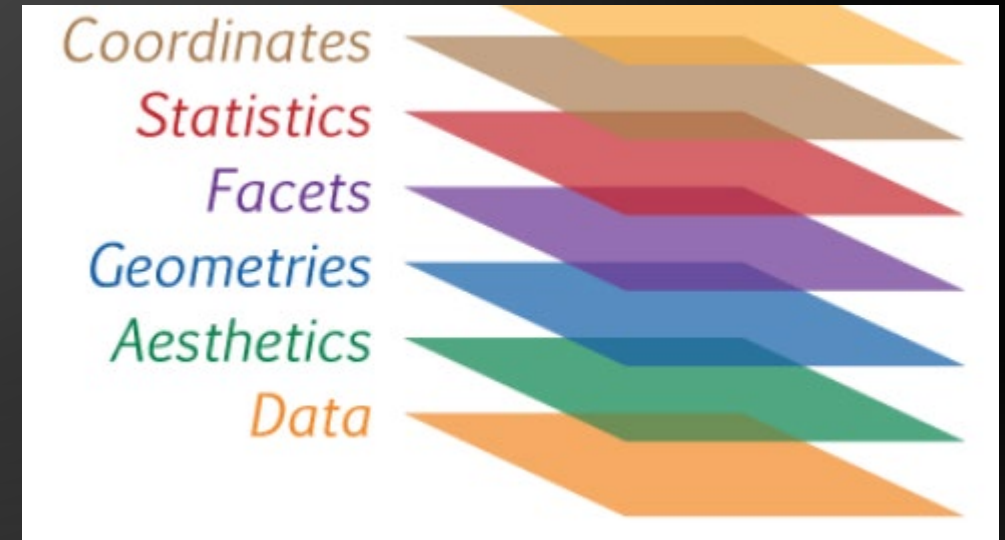
Ggplot2: the grammar of graphics

```
> 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()
```

