Data Visualization

Lecture 1: Introduction

Xiao Guo

2023/2/18

A little bit about me

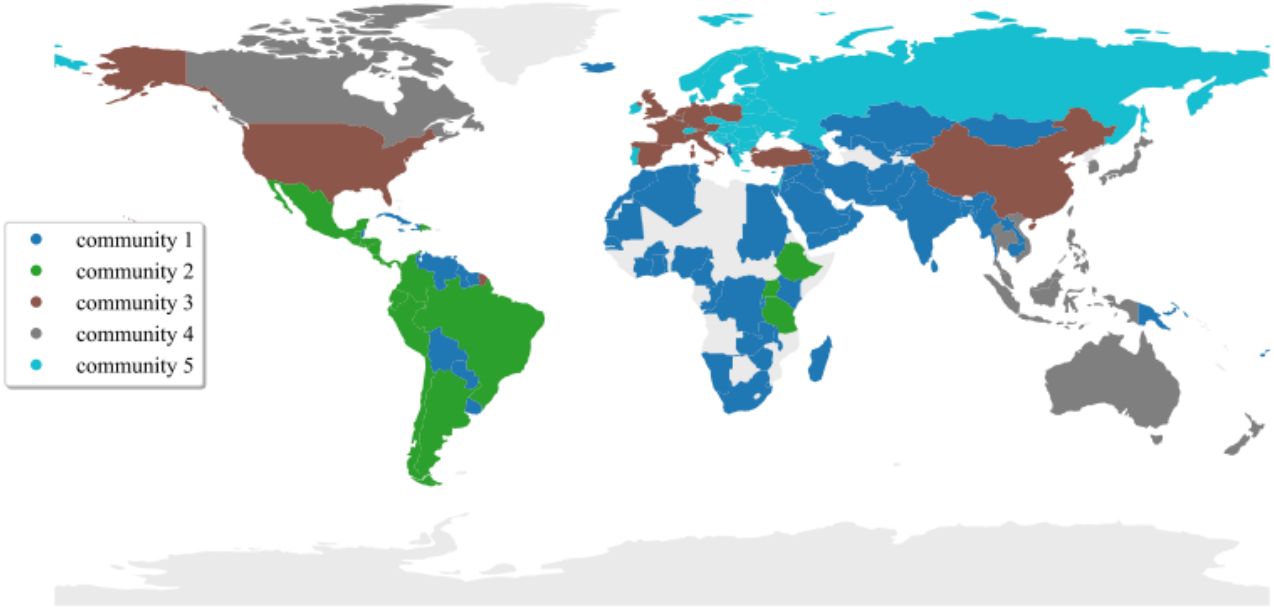
- Northwest University, 2020.04-present
- Northwest University, BS, MS, Ph.D., 2009-2019, supervised by Prof. Hai Zhang.
- Columbia University, Visiting student, 2018.10-2019.10, supervised by Prof. Ming Yuan.
- University of Wisconsin at Madison, Visiting student, 2015.10-2016.01

My current research focuses on high dimensional statistics and statistical machine learning. More specifically,

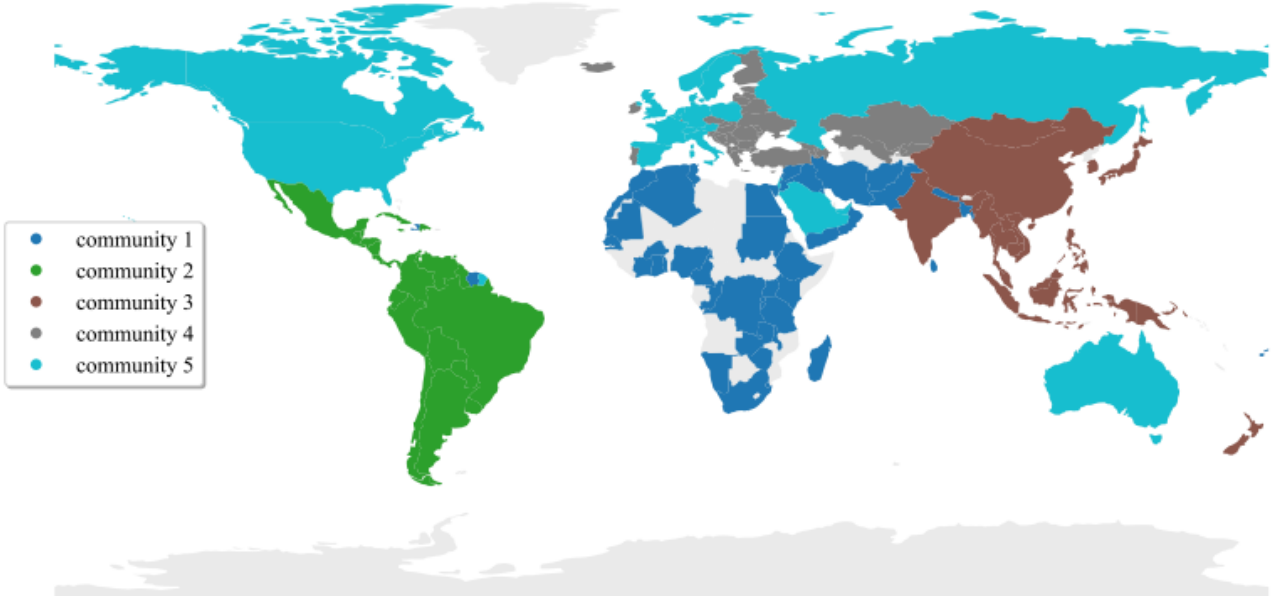
- Network analysis, Clustering, Unsupervised learning methods
- Privacy-preserving statistical data analysis, Large-scale data analysis, Federated learning methods

Welcome to join us!

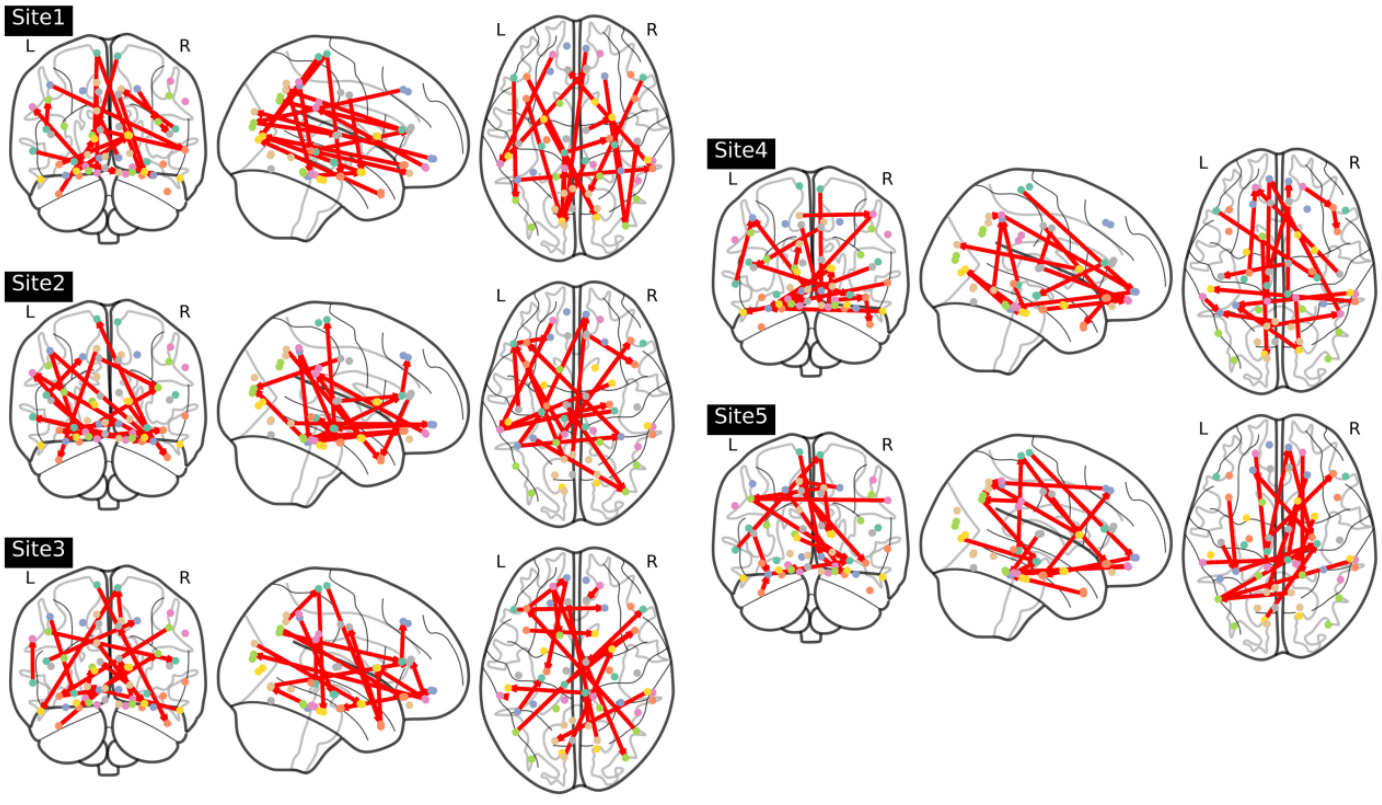
Some examples in our project-World trading patterns



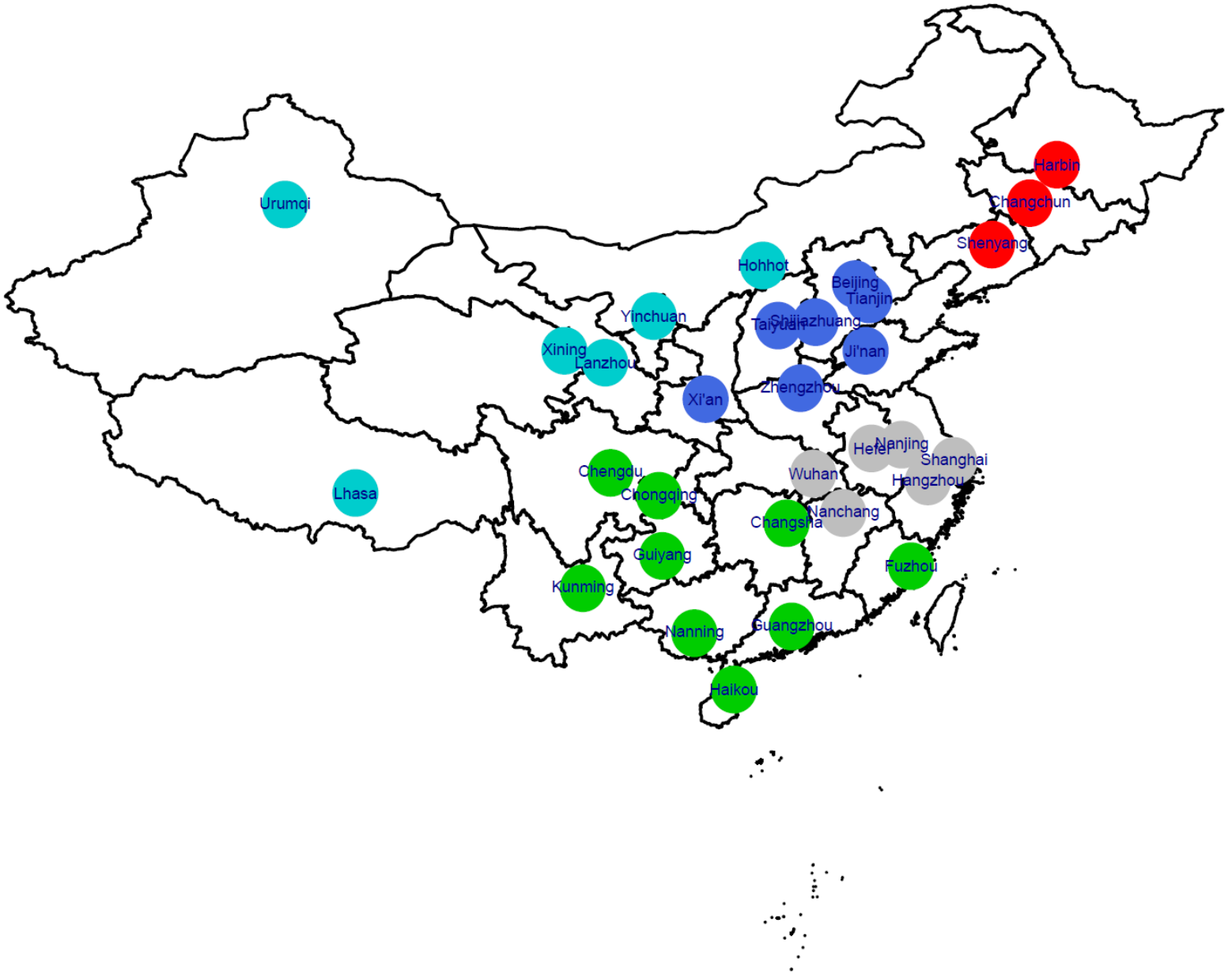
(a) row (export) clustering



Some examples in our project-Multi-site brain functional connectivities



Some examples in our project-Air pollution patterns



Data scientist: the sexy job

October 2012 Issue

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

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."

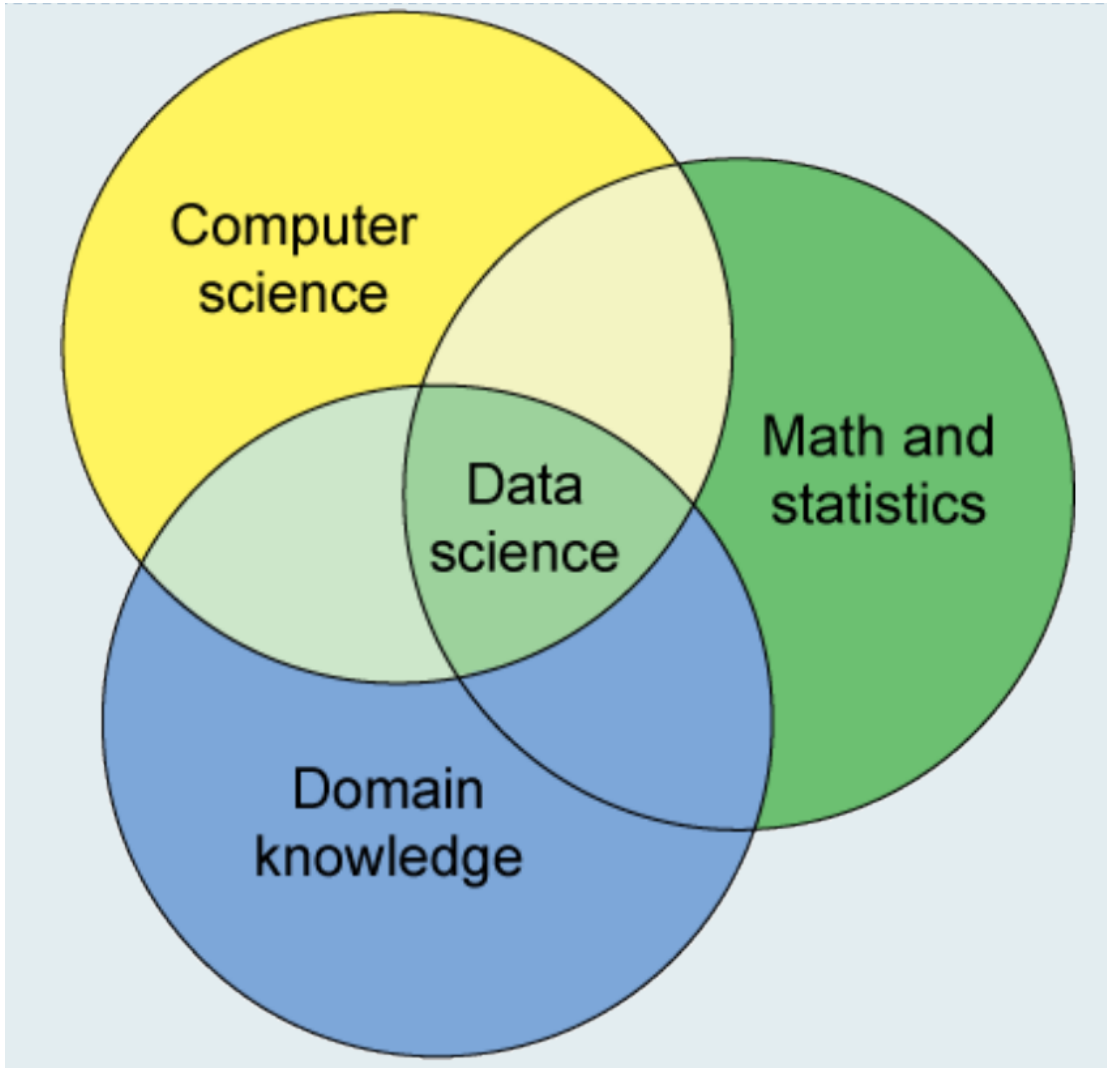
- See also an old article by NYT (2009): For Today's Graduate, Just One Word: Statistics
- And another famous McKinseay 2011 Report: Big data: The next frontier for innovation, competition, and productivity

What is a data scientist?

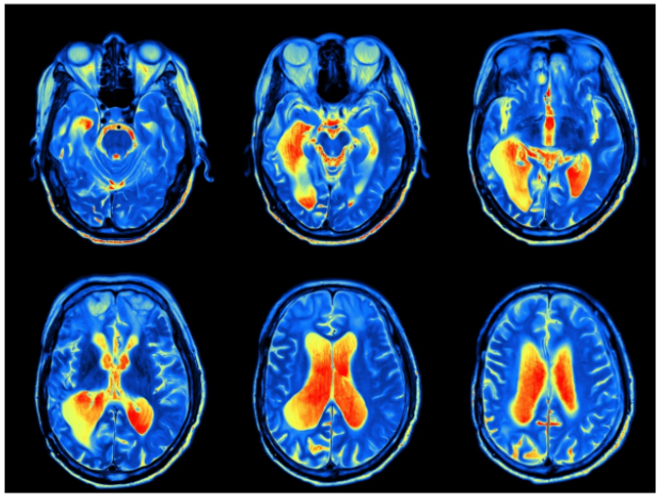
- Nate Silver (FiveThirtyEight, author of The Signal and the Noise): "Data scientist is just a sexed up word for a statistician."

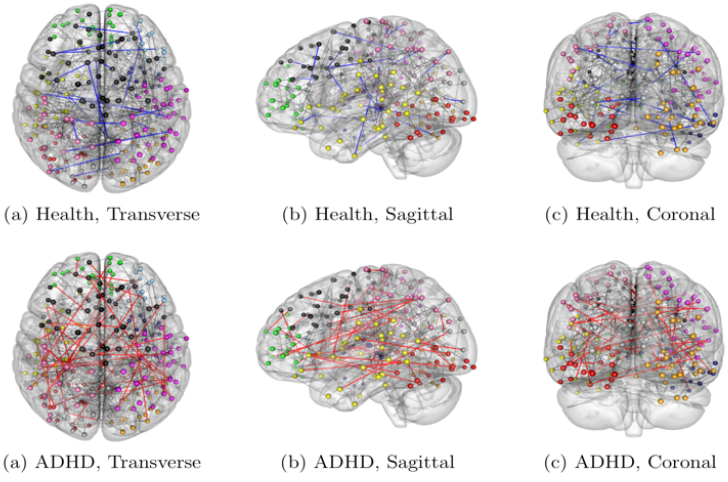
- "A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician." (from Joshua Blumenstock)

Data science is all the rage



An example: structure learning from fMRI data





(a) Health, Transverse (b) Health, Sagittal (c) Health, Coronal

(a) ADHD, Transverse (b) ADHD, Sagittal (c) ADHD, Coronal

- Statistics for modeling
- Computer science for optimizing
- Domain knowledge for explanation

Data science vs Statistics

L. Breiman (2001): Statistical modeling: the two cultures

A probabilist, and statistician, machine learner
(1928 – 2005)

CART, Bagging, Random Forests

"If our goal as field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a diverse set of tools."

A good portrait of data scientist by Bin Yu

- Statistics (S)
- Domain (science) knowledge (D)
- Computing (C)
- Collaboration ("team work") (C)
- Communication (to outsiders) (C)

Data Science = SDC^3

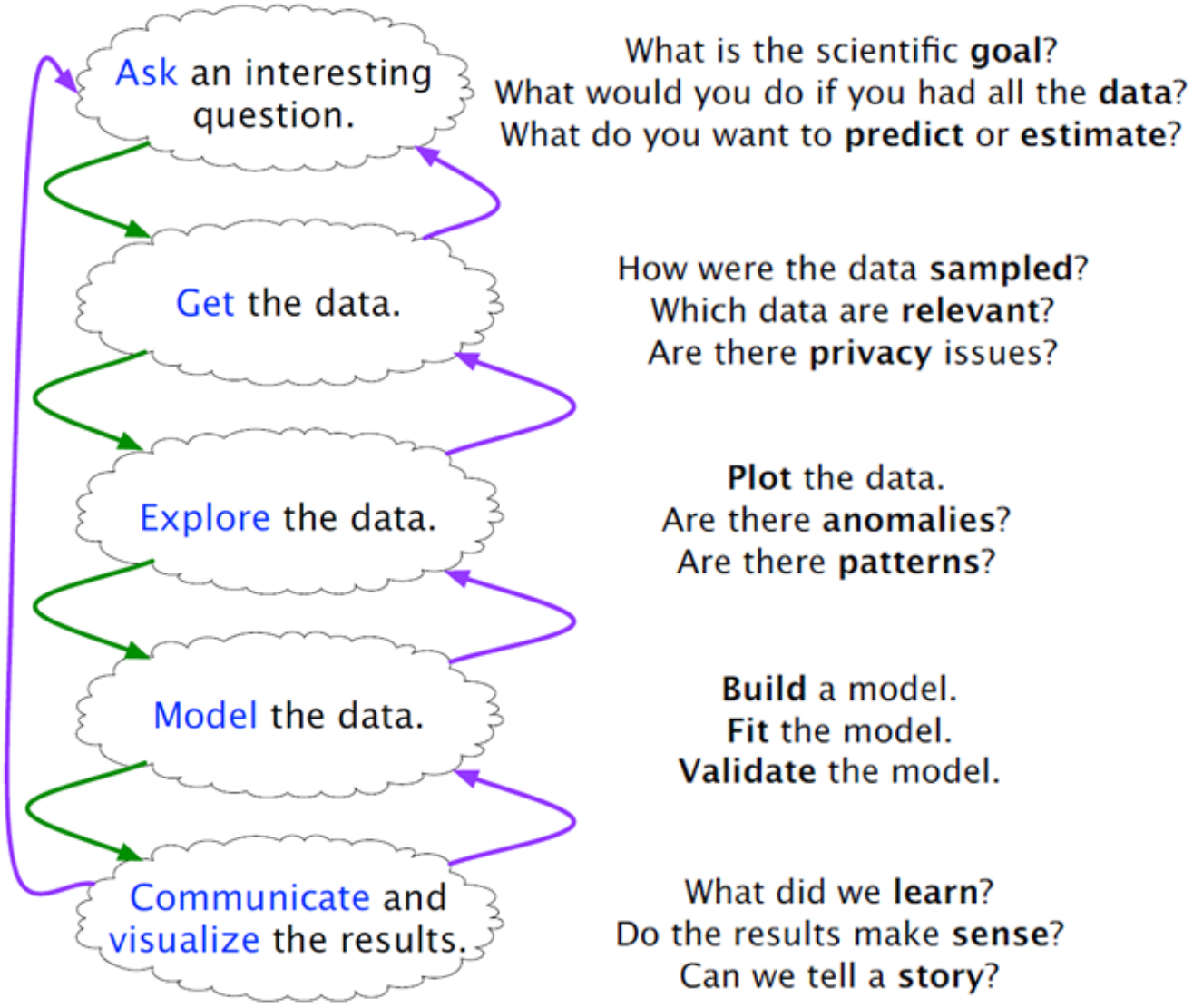
Statisticians do a big part of the job of a data scientist.

No existing discipline does more of the job of a data scientist

To fortify our position in DS, we should focus on

- Critical thinking: enables Statistics + Domain knowledge
- Computing: parallel computation, memory and communication dominate scalability
- Leadership, interpersonal, and communication: abilities enable collaboration + communication with outside

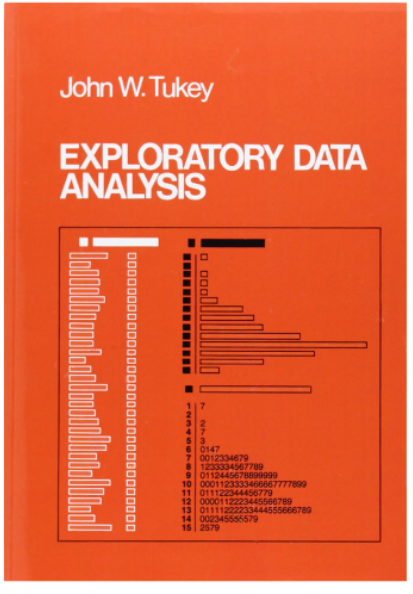
Data science workflow



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

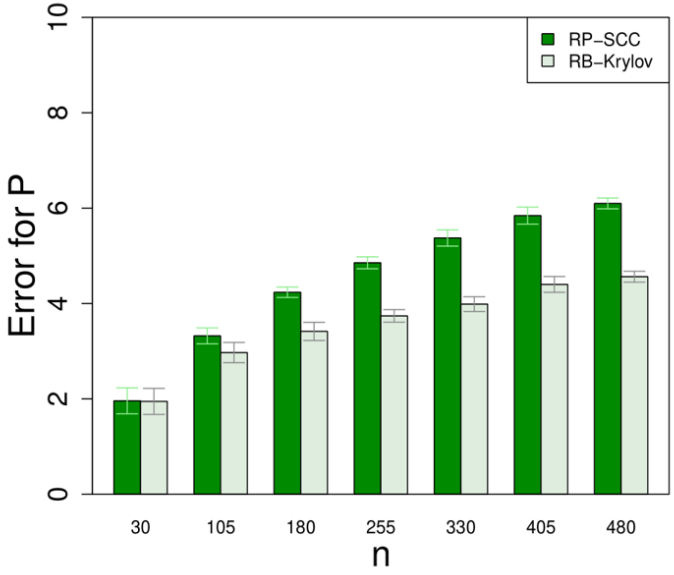
Roles of data visualization

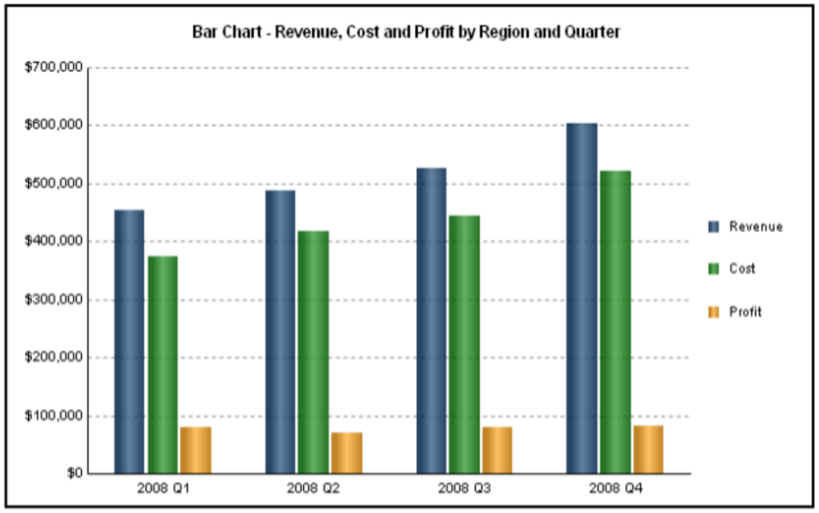
- Role 1: Exploratory data analysis (pre stage);
- Role 2: Visual presentation of results (after stage).
- John W. Tukey (1977; Exploratory Data Analysis): "The greatest value of a picture is when it forces us to notice what we never expected to see."



Principles of data visualization

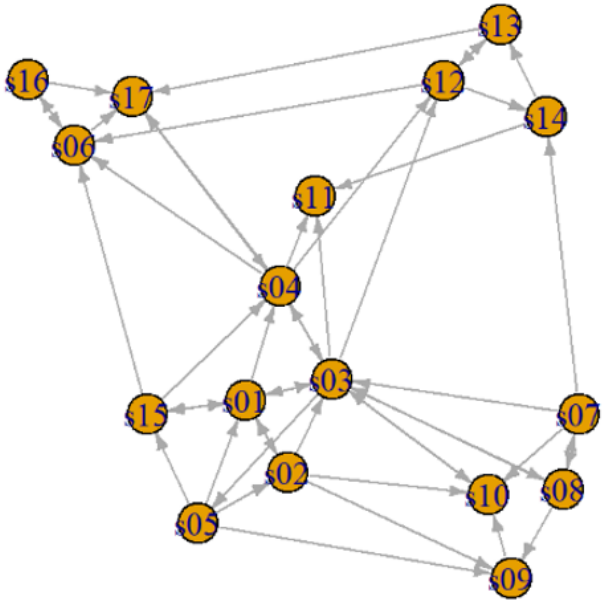
- Determine your audience. What questions will they need answered?

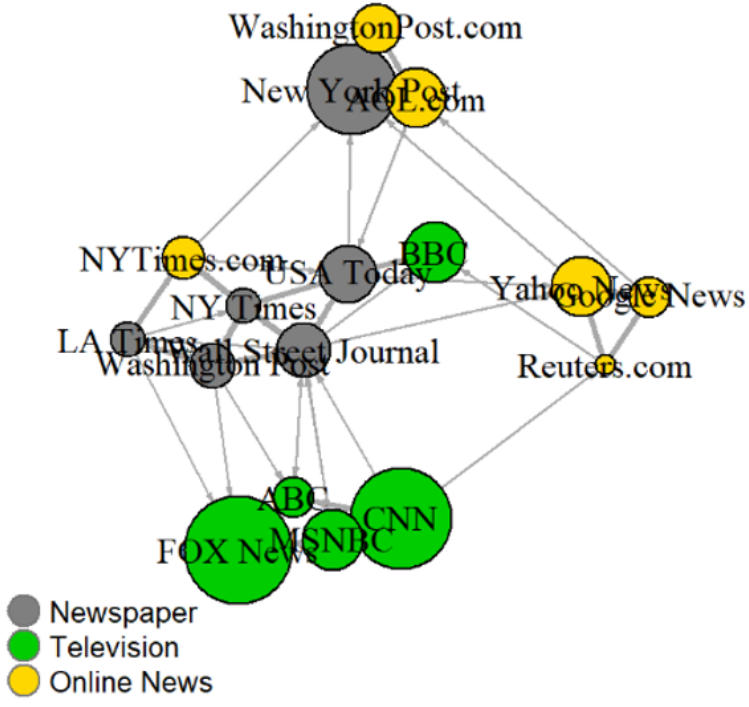




- Choose the right kind of chart (or other visualization) to depict the type of information you have.
- Provide the necessary context for data to be interpreted and acted upon appropriately.
- Keep it simple. Remove any non-essential information.
- Choose colors carefully to draw attention while also considering accessibility issues such as contrast.

- Seek balance in your visual elements, including texture, color, shape, and negative space.



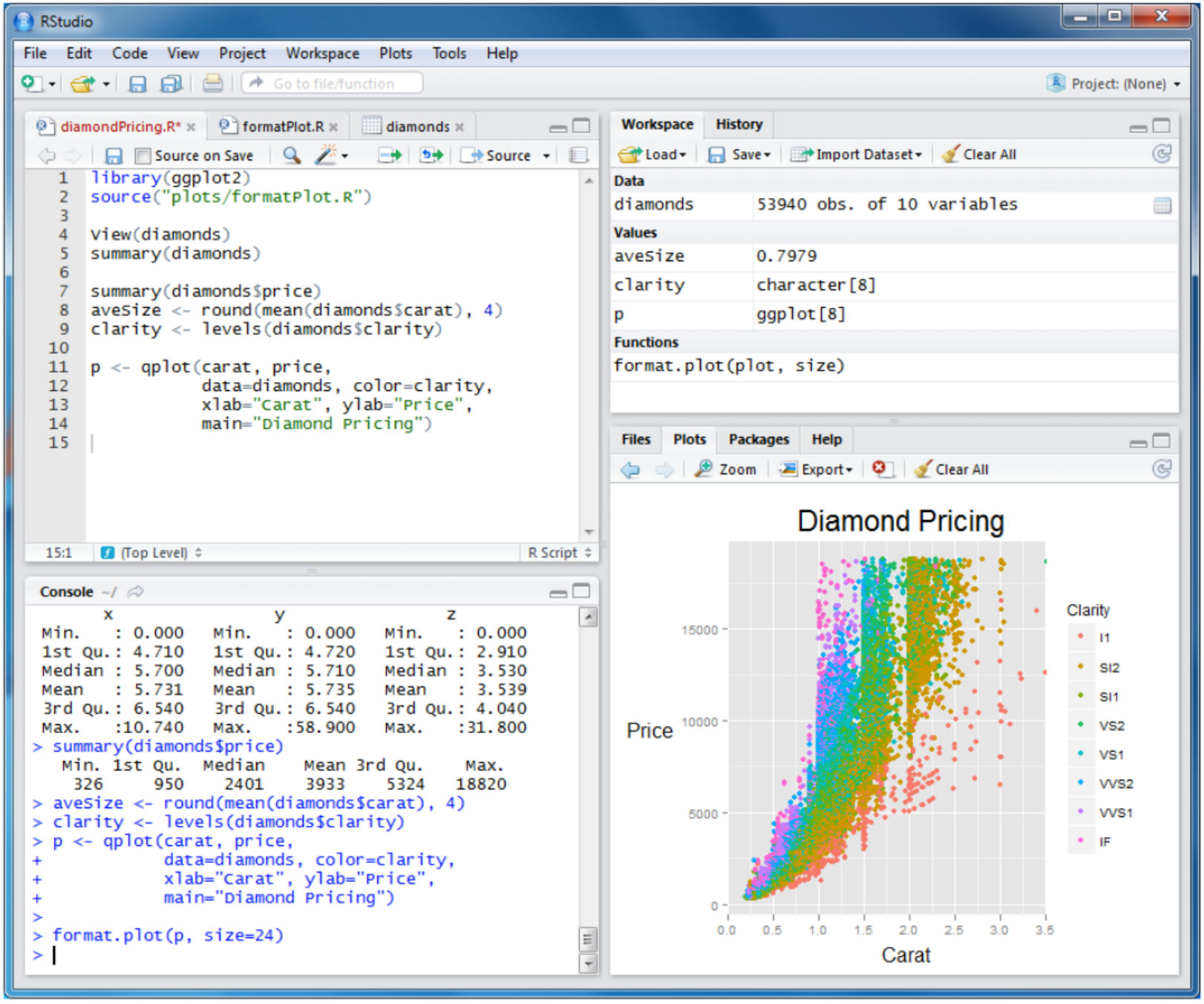


- Represent the data well. What information is missed? What is misinterpreted?

R studio and R markdown

- 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.
- Check out more nice features of RStudio at its official website.

R studio IDE


RStudio
File Edit Code View Project Workspace Plots Tools Help
Project: (None)
diamondPricing.R*
formatPlot.R
diamonds
Source on Save
Source
1 library(ggplot2)
2 source("plots/formatPlot.R")
4 View(diamonds)
5 summary(diamonds)
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15:1 (Top Level) R Script
Console ~/
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$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

R markdown

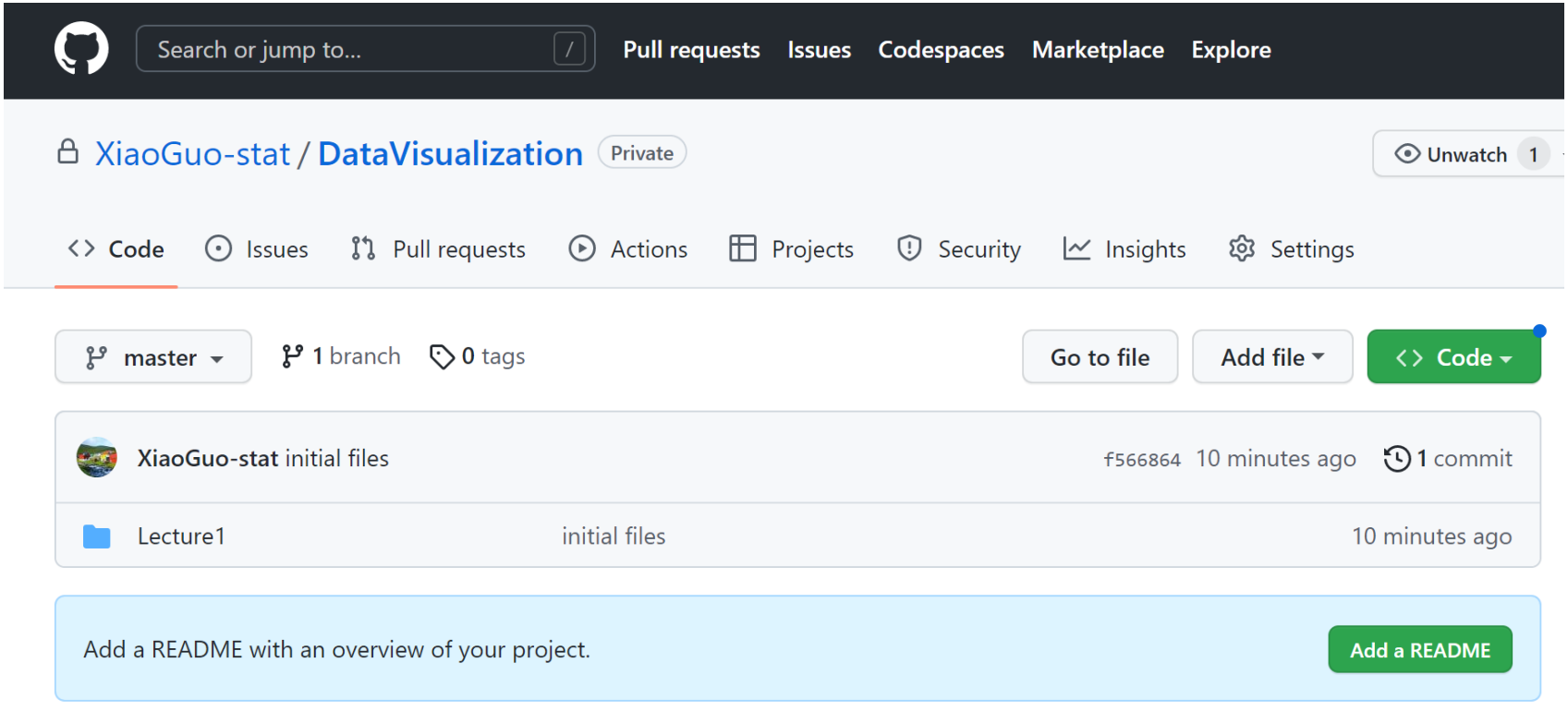
https://rmarkdown.rstudio.com/index.html
https://rmarkdown.rstudio.com/lesson-1.html

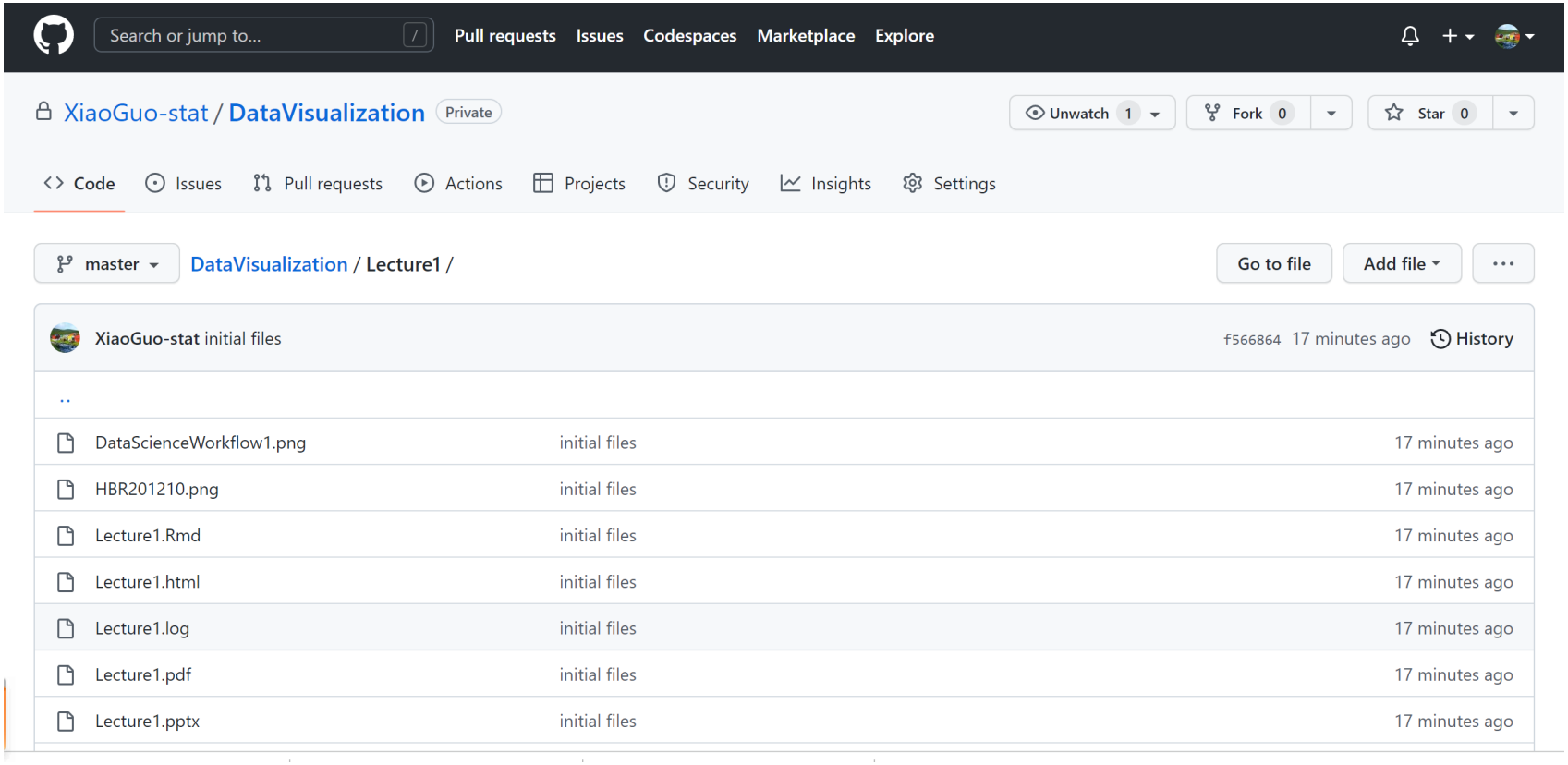
- install.packages("rmarkdown")
- install.packages("knitr")
- install.packages("tinytex")

Course website

- https://github.com/XiaoGuo-stat/DataVisualization

Weekly update.




Search or jump to...
Pull requests
Issues
Codespaces
Marketplace
Explore
XiaoGuo-stat / DataVisualization
Private
Unwatch 1
Fork 0
Star 0
Code
Issues
Pull requests
Actions
Projects
Security
Insights
Settings
master
DataVisualization / Lecture1 /
Go to file
Add file
XiaoGuo-stat initial files
f566864 17 minutes ago
History
..
DataScienceWorkflow1.png initial files 17 minutes ago
HBR201210.png initial files 17 minutes ago
Lecture1.Rmd initial files 17 minutes ago
Lecture1.html initial files 17 minutes ago
Lecture1.log initial files 17 minutes ago
Lecture1.pdf initial files 17 minutes ago
Lecture1.pptx initial files 17 minutes ago

Course materials

- Data Visualization with R by Rob Kabacoff. https://rkabacoff.github.io/datavis/
- HKU Stat3622 Data Visualization. https://ajzhanghk.github.io/Stat3622/
- R for Data Science (2017 O'Reilly) by Grolemund and Wickham. http://r4ds.had.co.nz/
- Learning IPython for Interactive Computing and Data Visualization (2nd) by Rossant, C. (2015). http://ipython-books.github.io/minibook/

References

- HKU Stat3622 Data Visualization. https://ajzhanghk.github.io/Stat3622/
- B. Yu (2014). Let us own data science. IMS Bulletin Institute of Mathematical Statistics (IMS) Presidental Address, ASC-IMS Joint Co nference, Sydney, July, 2014. https://www.stat.berkeley.edu/~binyu/ps/papers2014/IMS-pres-address14-yu.pdf