

# Data Visualization

## Lecture 1: Introduction

Xiao Guo

2023/2/18

## A little bit about me

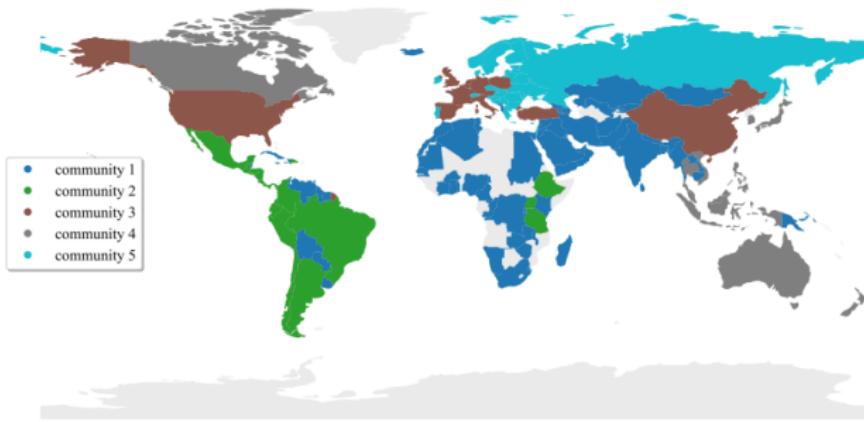
- ▶ Northwest University, 2020.04-present
- ▶ Northwest University, BS, MS, Ph.D., 2009-2019, supervised by Prof. Hai Zhang.
- ▶ Columbia University, Visiting student, 2018.10-2019.10, supervised by Prof. Ming Yuan.
- ▶ University of Wisconsin at Madison, Visiting student, 2015.10-2016.01

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

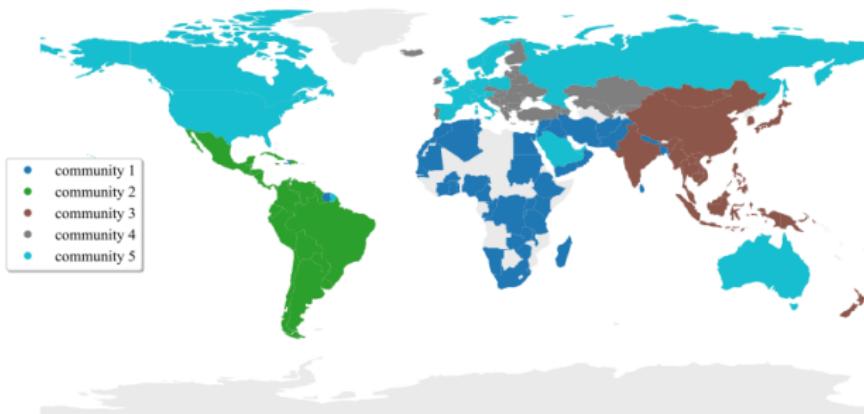
- ▶ Network analysis, Clustering, Unsupervised learning methods
- ▶ Privacy-preserving statistical data analysis, Large-scale data analysis, Federated learning methods

*Welcome to join us!*

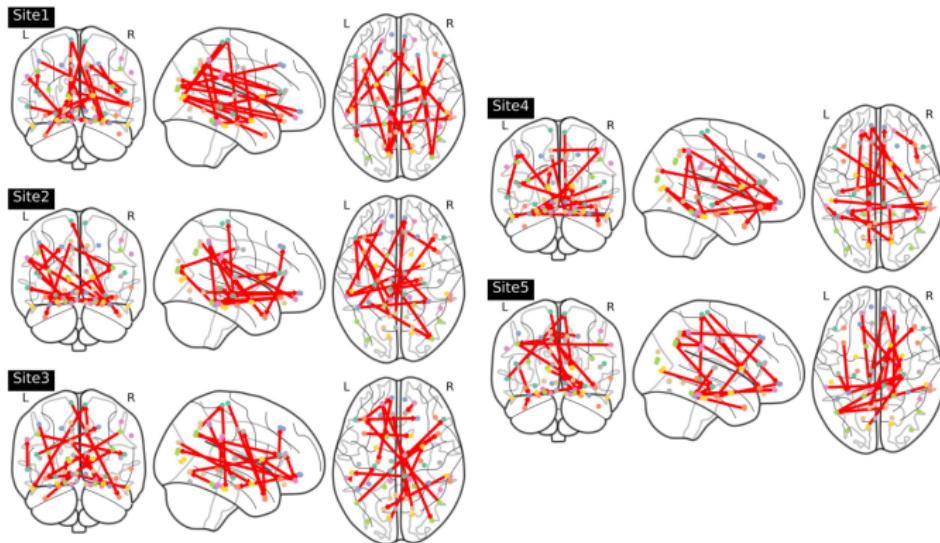
# Some examples in our project-World trading patterns



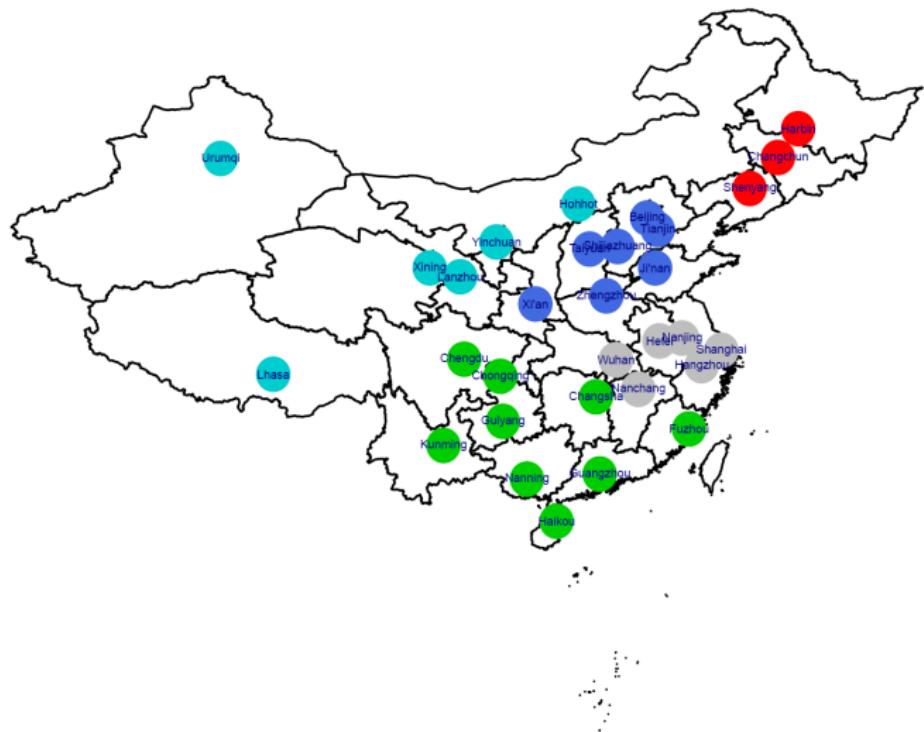
(a) row (export) clustering



# 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

**W**hen 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 McKinsey 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.”

 |  收藏 |  3 |  0

## 内特·希尔沃

 播报

 编辑

 讨论

 上传视频

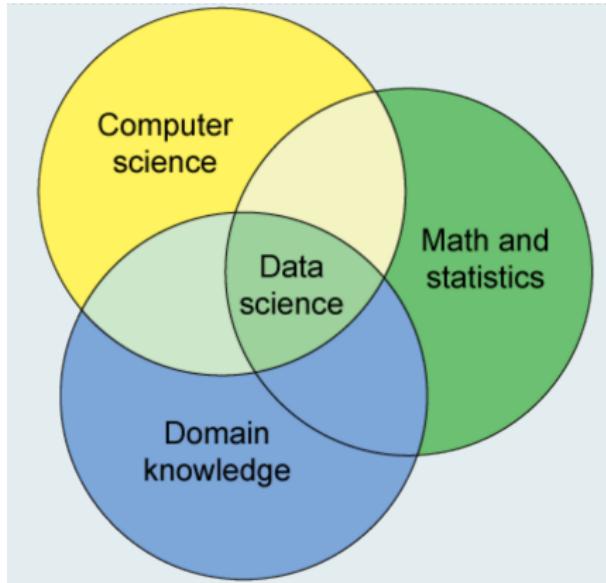
美国专业数模分析人士

 本词条缺少概述图，补充相关内容使词条更完整，还能快速升级。赶紧来[编辑吧！](#)

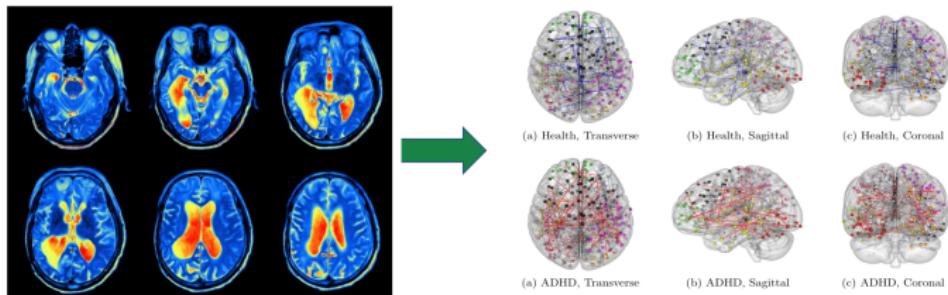
内特·希尔沃，美国专业数模分析人士。借助数学模型，希尔沃成功推断[奥巴马](#)会赢得2012年美国大选，并准确预测了全部50个州的选举结果。

- ▶ “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

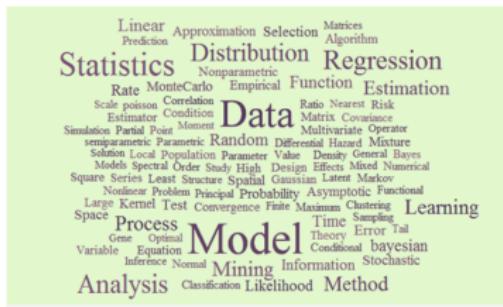


# An example: structure learning from fMRI data



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

# Data science vs Statistics



© morganimation - Fotolia.com

# 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<sup>3</sup>



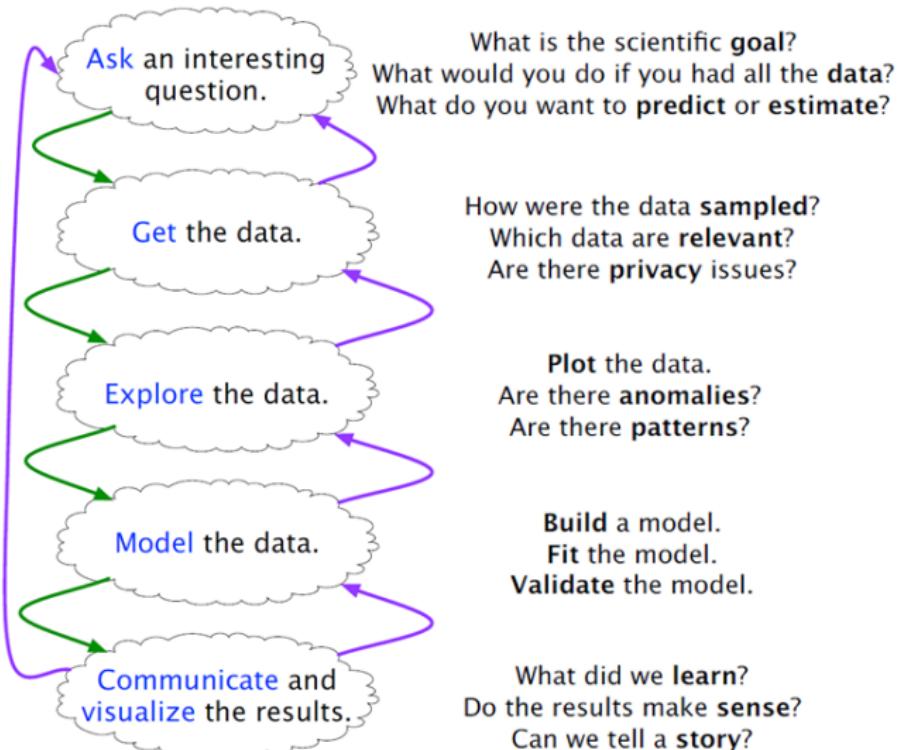
Statisticians do a big part of the job of a data scientist.

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

To fortify our position in DS, we should focus on

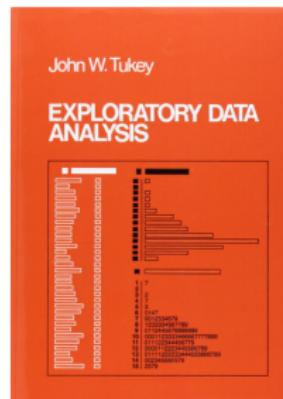
- ▶ Critical thinking: enables Statistics + Domain knowledge
- ▶ Computing: parallel computation, memory and communication dominate scalability
- ▶ Leadership, interpersonal, and communication: abilities enable collaboration + communication with outside

# Data science workflow



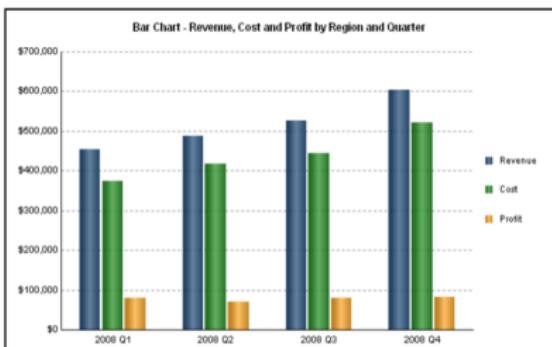
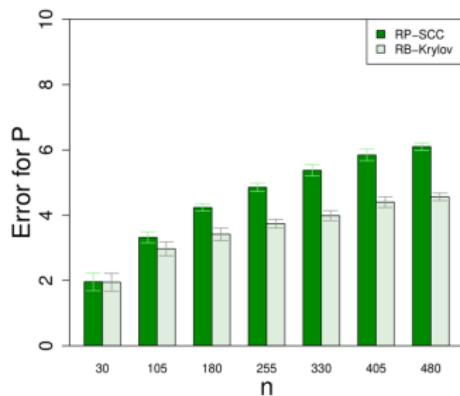
# Roles of data visualization

- ▶ 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.”



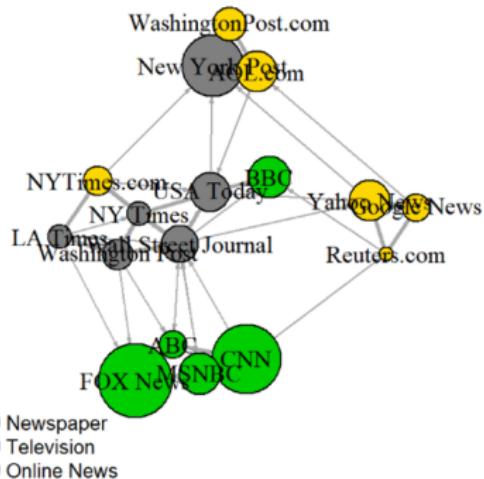
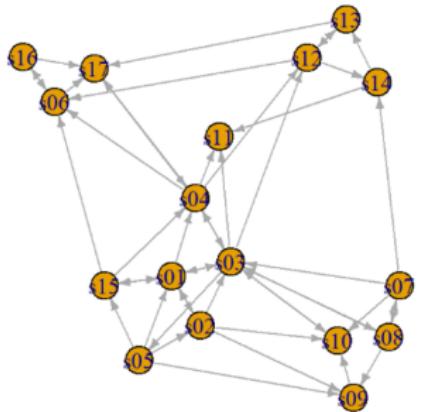
# Principles of data visualization

- 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

File Edit Code View Project Workspace Plots Tools Help

Project: (None)

diamondPricing.R\* x formatPlot.R x diamonds x

Source on Save Source

Go to file/function

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

15:1 (Top Level) R Script

Console

```
library(ggplot2)
source("plots/formatPlot.R")
View(diamonds)
summary(diamonds)
summary(diamonds$price)
aveSize <- round(mean(diamonds$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)
```

Diamond Pricing

Clarity

- I1
- SI2
- SI1
- VS2
- VS1
- VVS2
- VVS1
- IF

Price

Carat

# 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.

The screenshot shows a GitHub repository page for 'XiaoGuo-stat / DataVisualization'. The repository is private. At the top, there is a navigation bar with links for Pull requests, Issues, Codespaces, Marketplace, and Explore. Below the navigation bar, there is a search bar and a user profile icon. To the right of the search bar, there are buttons for Unwatch (with a count of 1) and Watch. The main navigation menu includes Code, Issues, Pull requests, Actions, Projects, Security, Insights, and Settings. The 'Code' tab is currently selected. Below the menu, there are status indicators for the master branch (1 branch, 0 tags), a 'Go to file' button, an 'Add file' button, and a 'Code' dropdown menu. A commit history section shows a single commit from 'XiaoGuo-stat' titled 'initial files' at 10 minutes ago. The commit hash is f566864. Below the commit history, there is a note to 'Add a README with an overview of your project.' and a green 'Add a README' button.

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

XiaoGuo-stat / DataVisualization Private Unwatch 1

Code Issues Pull requests Actions Projects Security Insights Settings

master 1 branch 0 tags Go to file Add file Code

XiaoGuo-stat initial files f566864 10 minutes ago 1 commit

Lecture1 initial files 10 minutes ago

Add a README with an overview of your project. Add a README

 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	fs66864 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) Presidential Address, ASC-IMS Joint Conference, Sydney, July, 2014.  
<https://www.stat.berkeley.edu/~binyu/ps/papers2014/IMS-pres-address14-yu.pdf>