

# ESU 2019 Data Visualization Summer Institute

## Introduction

Xuemao Zhang  
East Stroudsburg University

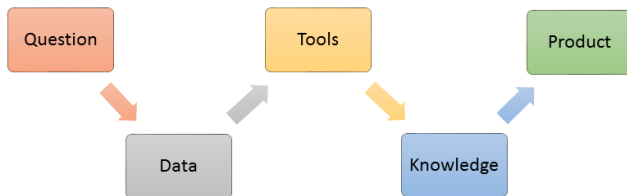
June 24, 2019

# Outline

- Why data science?
- Why programming?
- Why R?
- Why this data camp?

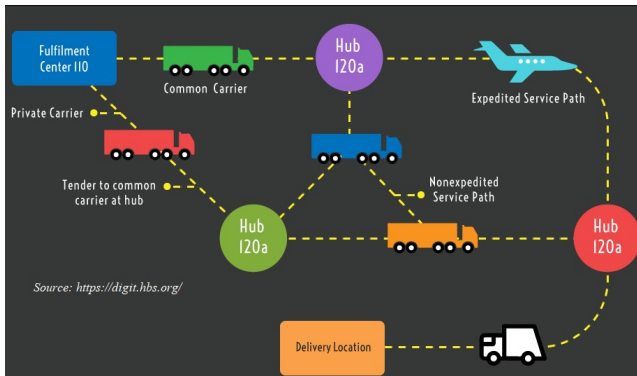
# Why data science?

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

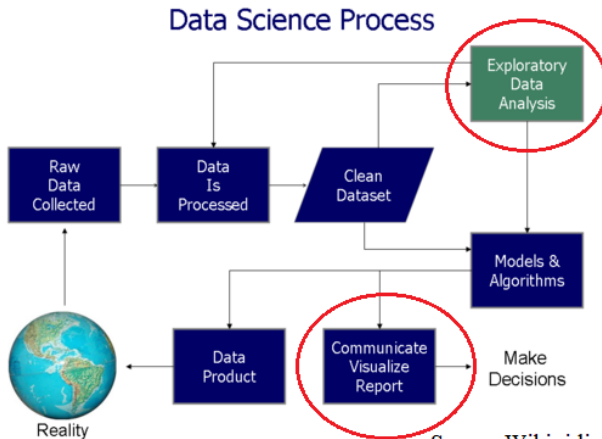


# Why data science?

- 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!



# Data Science Workflow



Source:Wikipedia

# 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 McKinsey 2011 Report: Big data: The next frontier for innovation, competition, and productivity

# Job Market of Data Scientist



**What**  
Job title, keywords, or company

data scientist

Q

**Where**  
City, state, or zip code

PA

data scientist jobs in Pennsylvania

Sort by:  
relevance - [date](#)

Salary Estimate

\$55,000	(795)
\$70,000	(658)
\$90,000	(489)
\$105,000	(342)
\$120,000	(180)

Job Type

Full-time	(928)
Contract	(48)
Part-time	(46)
Internship	(10)
Temporary	(6)

**New! Join Indeed Prime** - Get offers from great tech companies

**Data Scientist - Conshohocken, PA**  
RS Energy Group  
Conshohocken, PA  
Sponsored [save job](#)

**Software Engineer (Data Science Team) - Conshohocken, PA**  
RS Energy Group  
Conshohocken, PA  
Sponsored [save job](#)


**Manager, Scientist, Swiftwater, PA**  
Sanofi ★★★★★ 3,170 reviews  
Swiftwater, PA 18370  
Sponsored [save job](#)

**Data Scientist Others**  
citius tech ★★★★★ 68 reviews

# 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).

Dictionary

Enter a word, e.g. "pie" 

da·ta sci·en·tist

*noun*  
noun: **data scientist**; plural noun: **data scientists**

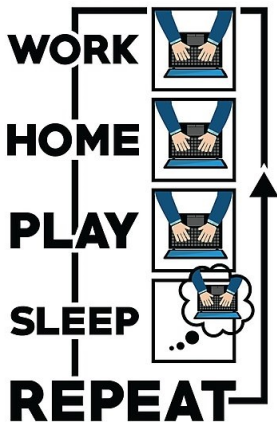
a person employed to analyze and interpret complex digital data, such as the usage statistics of a website, especially in order to assist a business in its decision-making.

"Silicon Valley technology companies are hiring data scientists to help them glean insights from the terabytes of data that they collect everyday"



# Why programming?

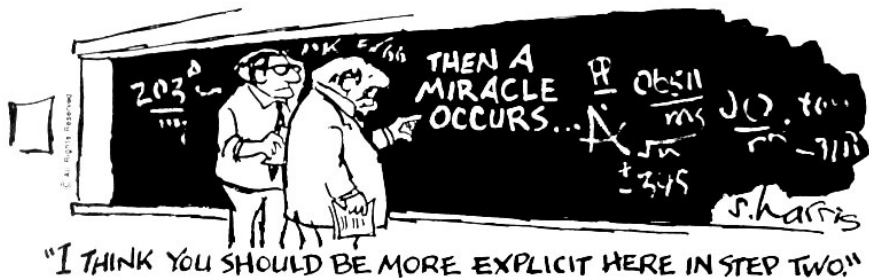
- To be able to easily repeat your own work.



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.
- Programming isn't scary. If you've written formulas in Excel, you've already done "programming".



# Why R?

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

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

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



# Why R?

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

```
##      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
```



## Why R?

- R offers various advanced visualization tools. For example, maps and animation.

# Why this data camp?

## You will learn:

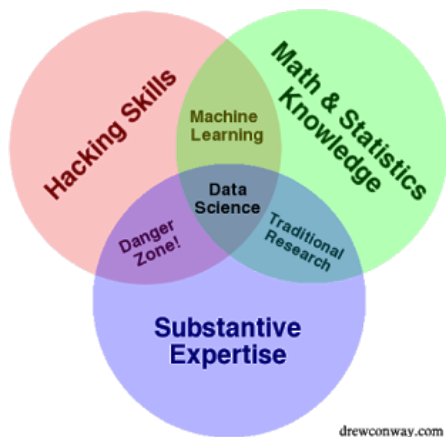
- Choose the best chart that fits the data
- Communicate effectively using graphics
- Create compelling visualization via R programming tools





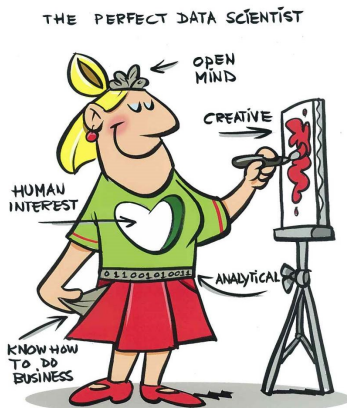
# Why this data camp?

- Data science is the application of analytical skills, scientific method, and computational skill to solve problems across professions: e.g. Biology, business, chemistry, physics, economics, mathematics, and computer science.



# Why this data camp?

- It motivates you to choose a STEM(Science, Technology, Engineering and Mathematics) major if you want to conduct data analysis even if you do not pursue data science as a career.
- You may find yourself a good fit to be a data scientist.



©Marion van de Wiel 2014

Source: <https://decisionata.com>

# Contentes of the Data Camp

- Introduction
- R and RStudio
- Data manipulation
- Data visualization
- R-Markdown

## Data Visualization

- R: base plot
- R: advanced visulization using ggplot2
- R: animation
- R: Map visualization

# Questions?



- Ask us if you need help with R and RStudio installations.
- Ask us if you need help with R package installations.