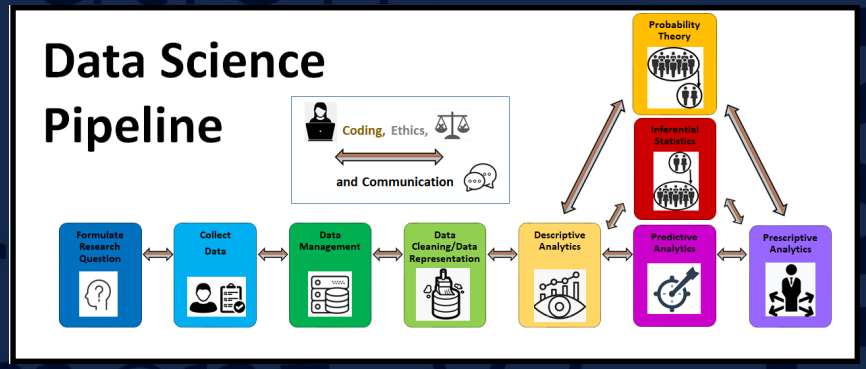
Introduction to the STAT207 Course

August 24, 2021



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Introduction to this Course

About you

About me

What is data science?

Example of the “Full Data Science Pipeline”

Data Science vs. Statistics vs. Computer Science

Why Python?

Why study data science?

Course Website/Syllabus

Learning Content Tips

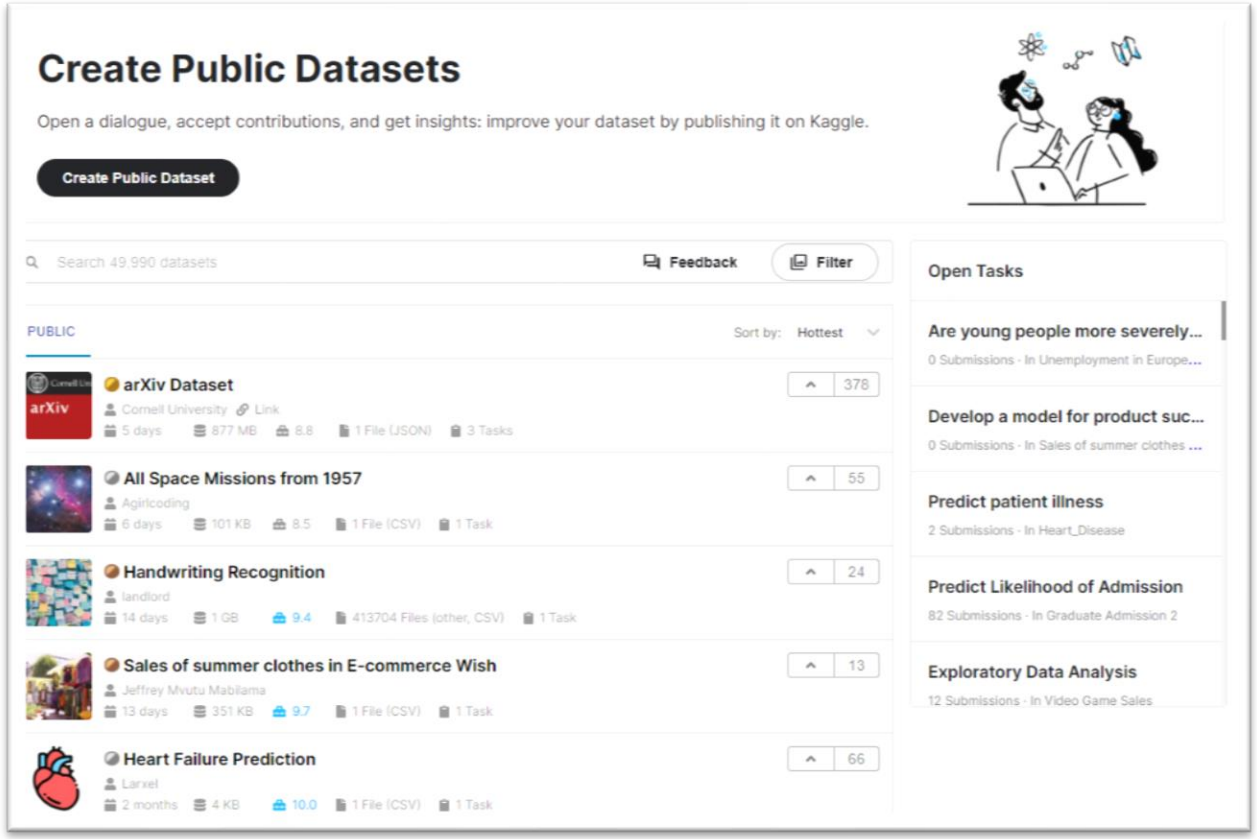
Lecture Tips

General Course Tips

Coding Tips

About You

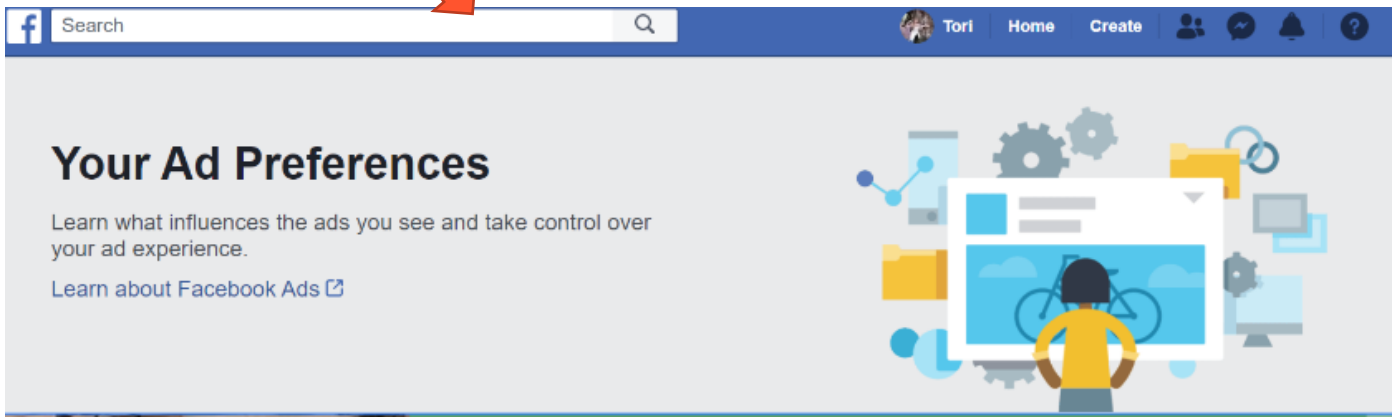
- What types of data sets would you like to **gain insights from, make predictions with,** and/or **use to help make better decisions?**



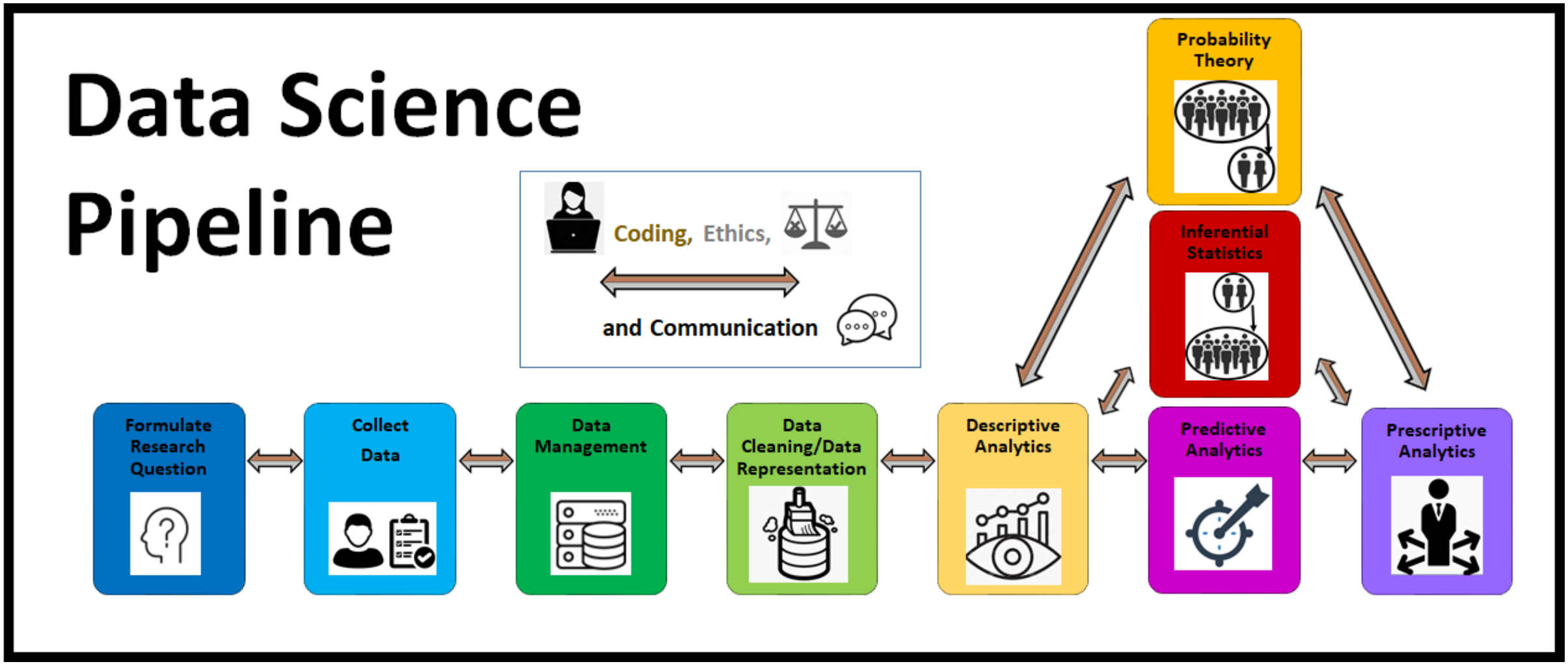
https://www.kaggle.com/datasets

About Me

- Online Advertising
- TV Advertising
- Narcotics Detection
- Gene Expression Analysis
- Get Out the Vote Initiatives

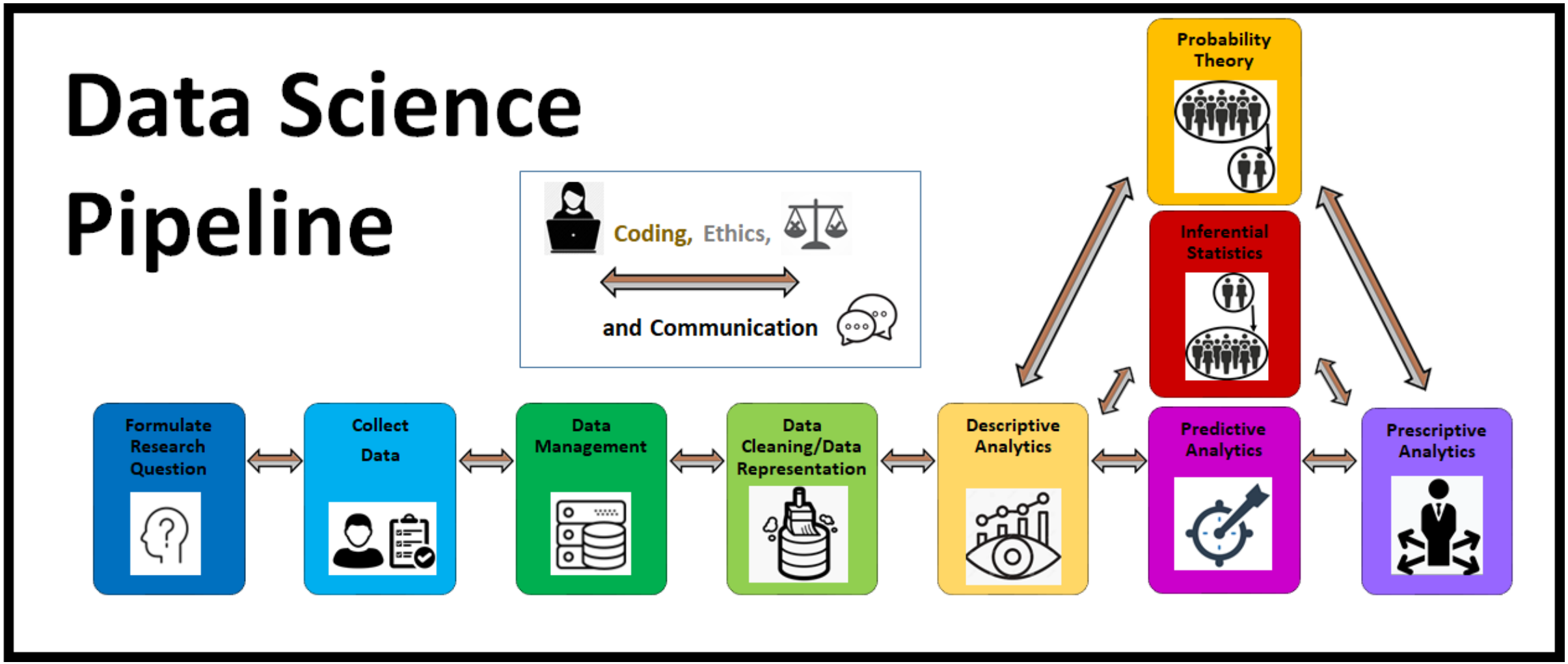


What is data science?



About you
About me
What is data science?
Example of the "Full Data Science Pipeline"
Data Science vs. Stats vs. CS
Why Python?
Why study data science?
Syllabus
Learning Content Tips
Lecture Tips
General Course Tips
Coding Tips

Data Science vs. Statistics



Purpose of STAT207

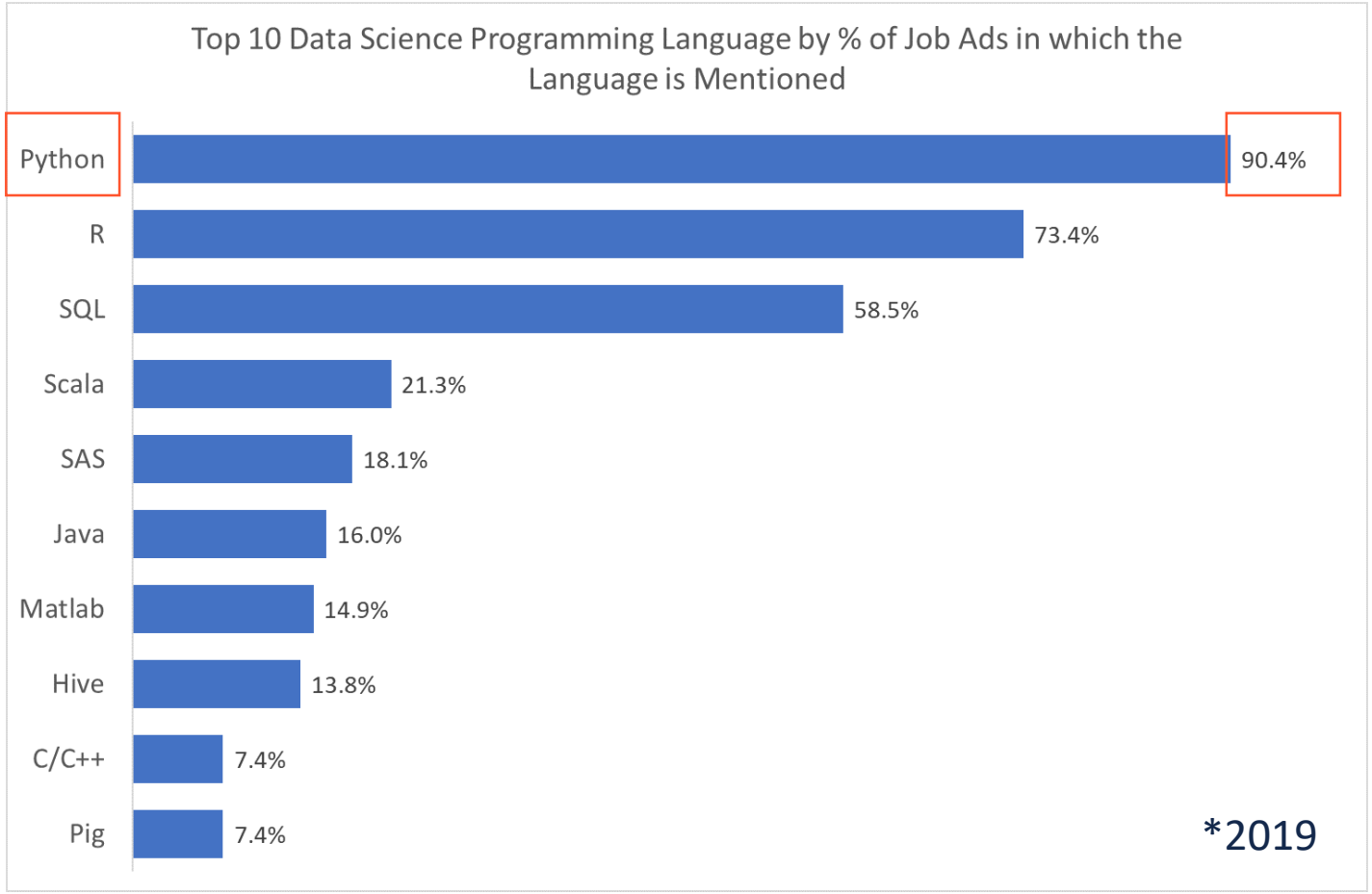
- **Survey** of the data science pipeline

- Using **Python,** complete a **beginning-to-end data science project.**

- When conducting a more advanced data science project, **develop an intuition** for:
  - what **questions to ask**
  - how to efficiently learn **new algorithms**, **models**, **functions** etc
  - What **search terms** to look up
  - **what to research**

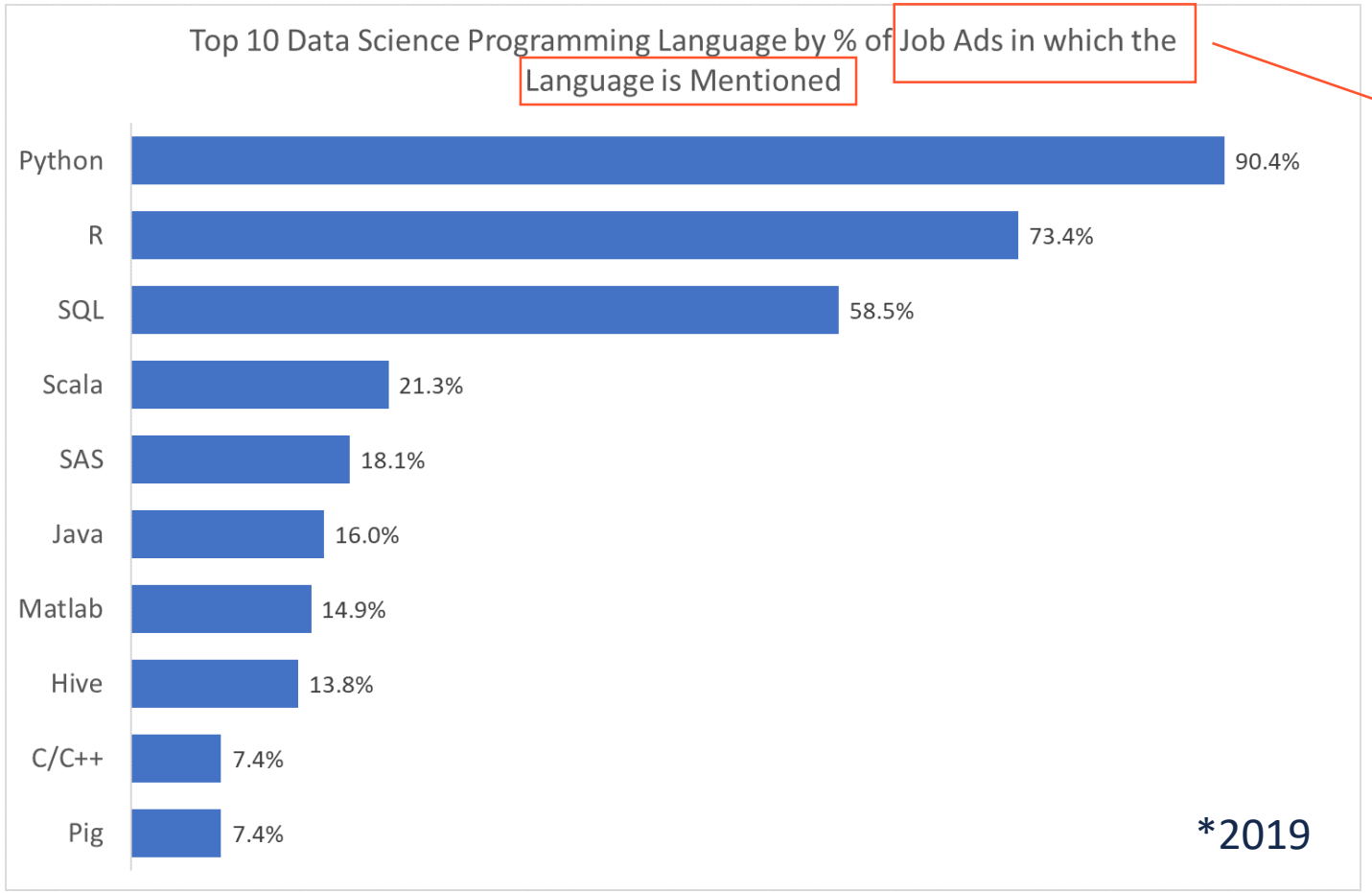
- **Topics covered:**
  - http://courses.las.illinois.edu/fall2021/stat207/course_topics.html

Why use Python for Data Science?



https://towardsdatascience.com/which-programming-language-should-data-scientists-learn-first-aac4d3fd3038

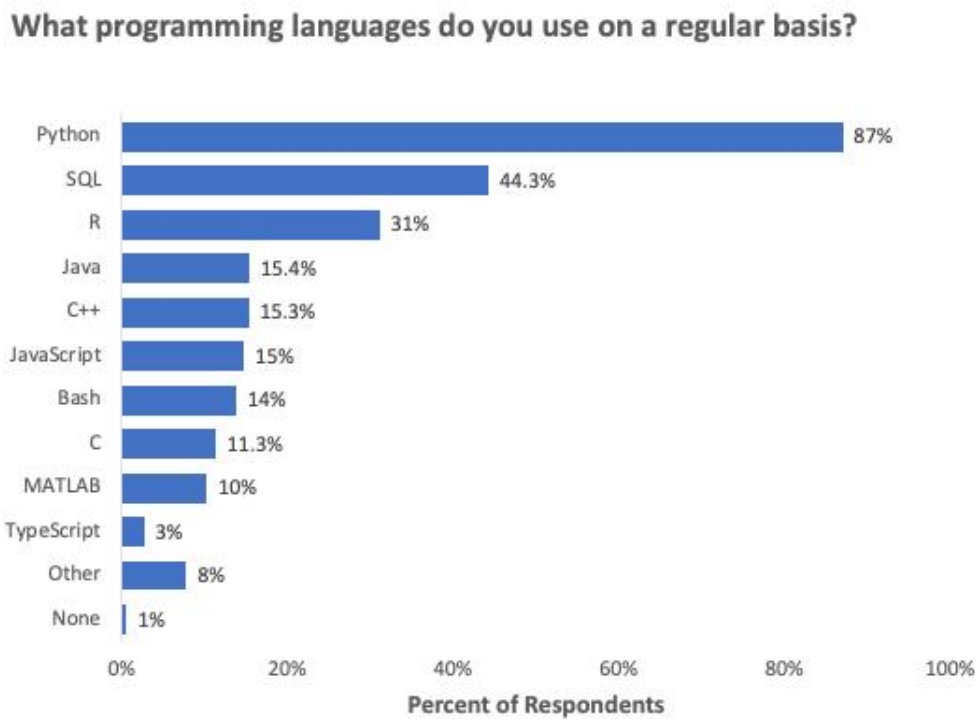
Why use Python for Data Science?



What are some ways we could have collected this data?

https://towardsdatascience.com/which-programming-language-should-data-scientists-learn-first-aac4d3fd3038

Why use Python for Data Science?



Note: Data are from the 2019 Kaggle ML and Data Science Survey. You can learn more about the study here: https://www.kaggle.com/c/kaggle-survey-2019.
A total of 19717 respondents completed the survey; the percentages in the graph are based on a total of 14762 respondents who provided an answer to this question.

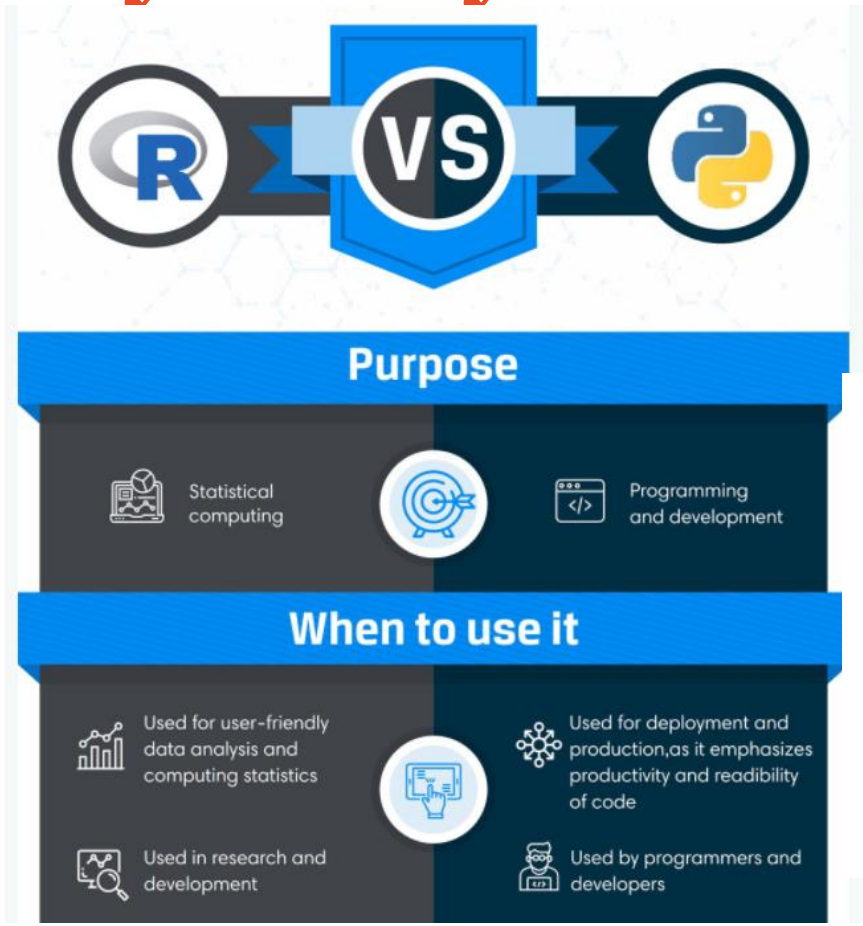
BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

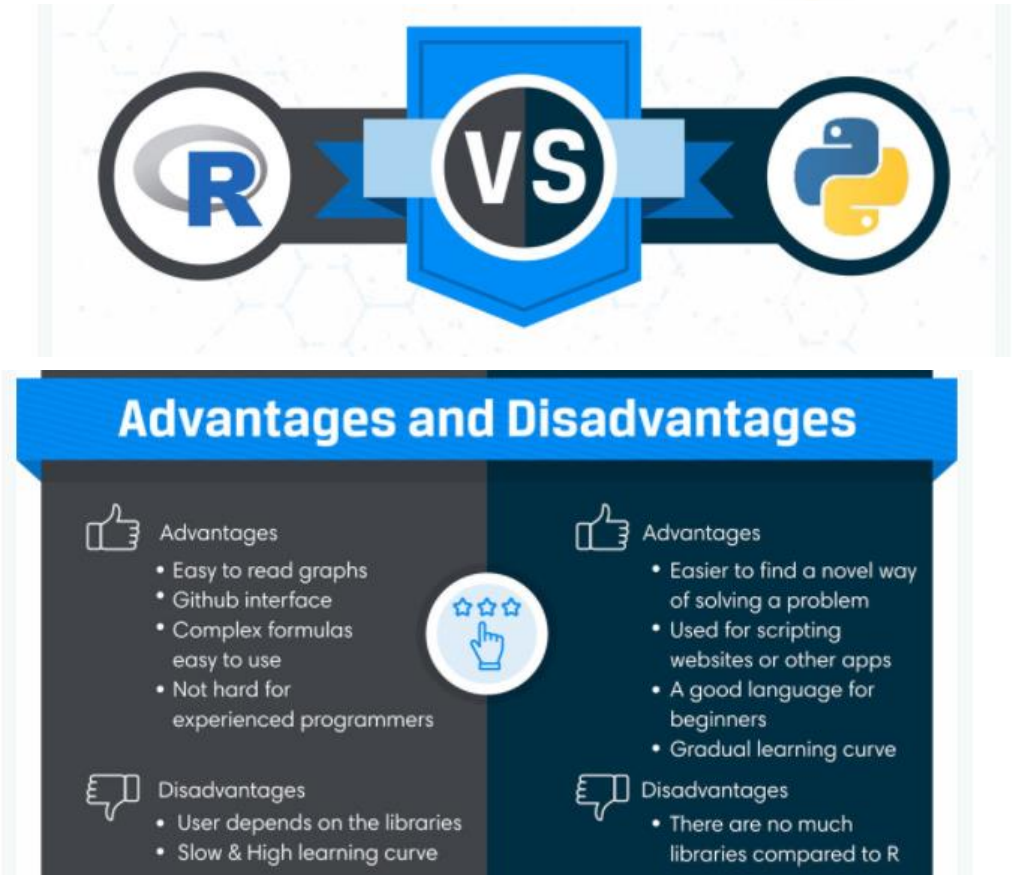
Copyright 2020 Business Over Broadway

What if we wanted make an inference about whether Python is the most used programming language of ALL DATA SCIENTISTS using this sample of data scientists? What might we be interested to know about how the data was collected?

About you
About me
What is data science?
Example of the "Full Data Science Pipeline"
Data Science vs. Stats vs. CS
Why Python?
Why study data science?
Syllabus
Learning Content Tips
Lecture Tips
General Course Tips
Coding Tips

Why use Python for Data Science?





- www.stackoverflow.com has great answers to many of the questions you could ask for Python!
- Working in a big team to automate something? Python is great!

https://www.superdatascience.com/blogs/learn-all-the-pros-and-cons-of-python-vs-r-programming

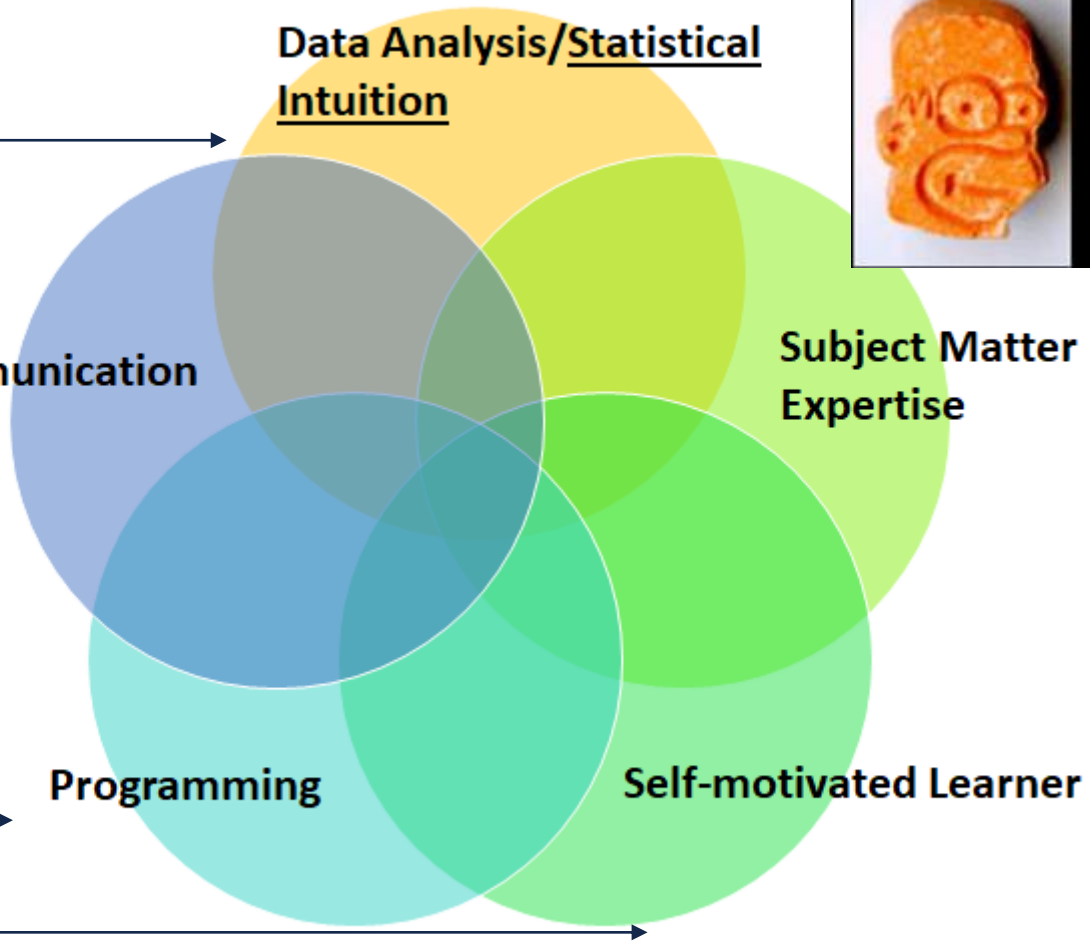
Why study data science?

https://www.superdatascience.com/blogs/learn-all-the-pros-and-cons-of-python-vs-r-programming

Why study data science?

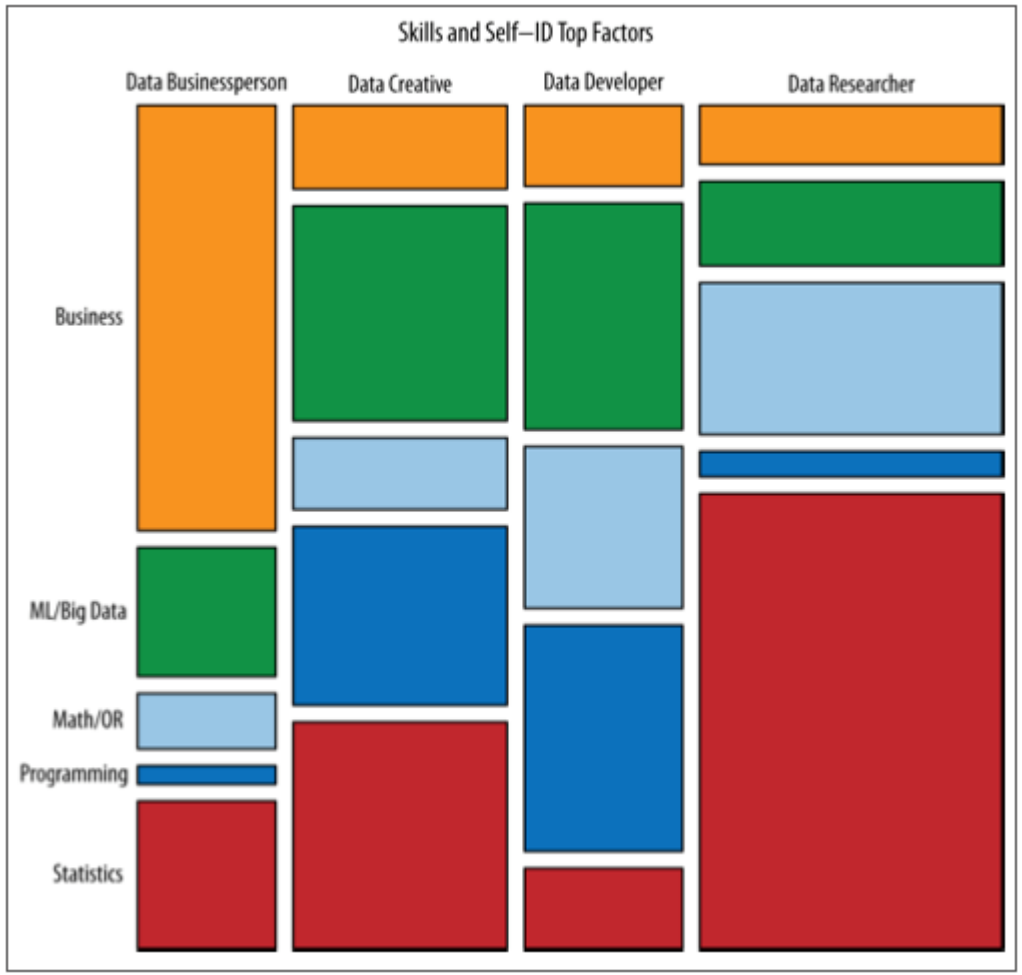
Current State of Data Analysis/Science:

1. Field of data analysis is broad! (Impossible to know *everything.)*
2. New and useful statistical analyses and algorithms are developed everyday!
3. Datasets are larger!



https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

Why study data science?



•**Data Businesspeople** are the product and profit-focused data scientists. They're leaders, managers, and entrepreneurs, but with a technical bent. A common educational path is an engineering degree paired with an MBA.

•**Data Creatives** are eclectic jacks-of-all-trades, able to work with a broad range of data and tools. They may think of themselves as artists or hackers, and excel at visualization and open source technologies.

•**Data Developers** are focused on writing software to do analytic, statistical, and machine learning tasks, often in production environments. They often have computer science degrees, and often work with so-called "big data".

•**Data Researchers** apply their scientific training, and the tools and techniques they learned in academia, to organizational data. They may have PhDs, and their creative applications of mathematical tools yields valuable insights and products.

http://radar.oreilly.com/2013/06/theres-more-than-one-kind-of-data-scientist.html

Course Website and Syllabus

Course Website: http://courses.las.illinois.edu/fall2021/stat207/

- Schedule
- Syllabus
- Course information
- Assignment Information
- Git/Coding/Python Resource Help Pages

Compass Page: https://compass2g.illinois.edu/

- Zoom links for:
  - Assigned Labs (with approval from Yuhan)
  - Open Labs
  - My office hours
- Lecture markups
- Grades, feedback, and solutions on assignments and exams

Course Website and Syllabus

Campuswire: https://campuswire.com/p/G56739ECA

- Registration Code: 1881
- Content and non-personal course related questions.

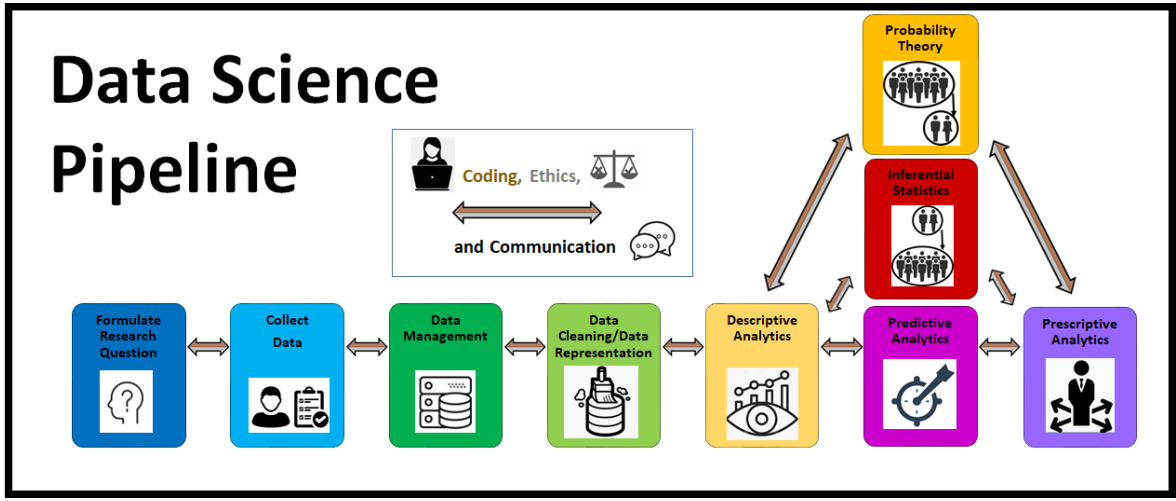
Github: https://github-dev.cs.illinois.edu/stat207-fa21

- **Fetch and merge** your weekly lab assignments, and project
- **Push** your weekly assignments, exams, and project for grading

Learning Content Tips

Real-life Data Science Problems "from scratch" = Asking the right questions

- What are you trying to **achieve** at the end of this analysis?
- What is your **research question**? What would an answer to this research question look like?
- What **kind of data should you collect** to help you answer this question?
  - What language should we careful of using/not using when answering this research question?
- What data procedure, visualization, statistic, model, algorithm should you use to help you answer this question?
  - What is the nature of your data?
    - Which/**how many variables** are involved in answering this question?
    - What **types of variables** are involved? (Categorical vs. Continuous)
    - Does the data fit the **assumptions** of your procedure, statistic, model, or algorithm?



- Are you **communicating** your findings effectively to:
  - Yourself?
  - Your teammates
  - Your boss?
  - Non-technical audiences
  - The general public?
- **Ethics**: Are you thinking about all of the possible people that might be affected by your data decisions and communicating these possible effects?

Lecture Tips - Synchronous

- **Synchronous:** strongly encouraged if you are able to, but not required!

- **Each class download these**
  - Python Notebooks
  - Pdf
  - csv files (sometimes)

- **Note-taking Ideas**
  - Printing the pdf, hand written notes
  - Onenote (or other similar notetaking apps)
  - Shared Google doc with your friends/classmates

- **Following Along with Code**
  - Download .ipynb before class and try to follow along *(not all class notes will be in ipynbs, but all code will be in the pdf)*

- **Engaging during Lecture**
  - Zoom chatroom
  - Private chatroom messages to TA/CA.
- **Breakout Rooms**
  - Ask classmates for help in the breakout rooms.
  - Ask me/CAs for help in the breakout rooms.

General Course Tips

- Check your email regularly!
- Go to open labs
  - **Tuesday 5pm-7pm CST**
- Go to office hours
  - **Monday 11am-1pm CST**
- Get as much done as you can in the assigned labs.
- Campuswire can be helpful!
- Ask questions if you get stuck.
- Code: Keep a list of new Python functions that we learn in the lectures and what they do.