

Introduction to the STAT207 Course

August 25, 2020



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Introduction to this Course



About you



Why study data science?



About me



Course Website/Syllabus



What is data science?



Learning Content Tips



Example of the “Full Data Science Pipeline”



Lecture Tips



Data Science vs. Statistics vs. Computer Science



General Course Tips



Coding Tips



Why Python?



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

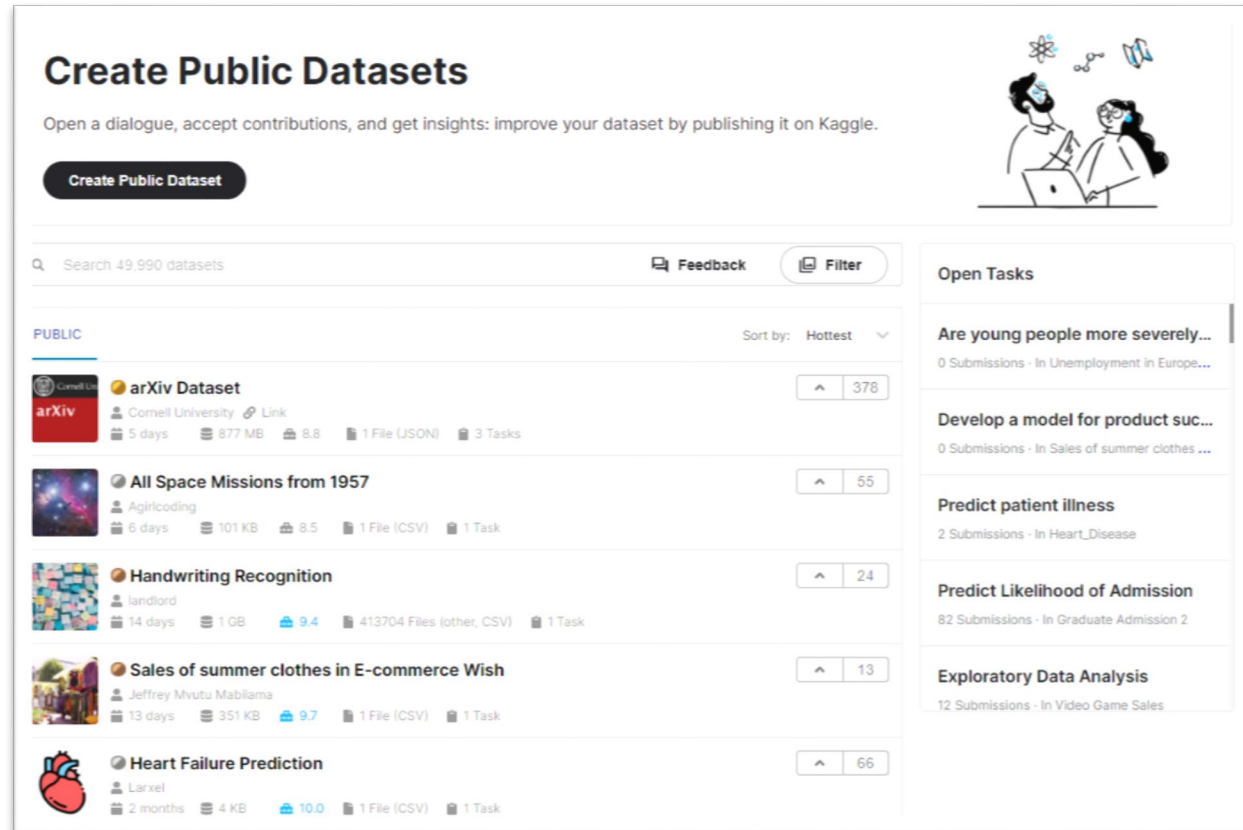
Lecture Tips

General Course Tips

Coding Tips

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?**



The screenshot shows the 'Create Public Datasets' page on Kaggle. At the top, there's a header with the title 'Create Public Datasets' and a sub-header 'Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle.' Below this is a 'Create Public Dataset' button. A search bar indicates 'Search 49,990 datasets'. On the right, there's a 'Feedback' button and a 'Filter' button. The main content area lists several public datasets under the 'PUBLIC' tab, sorted by 'Hottest'. The datasets listed are: 'arXiv Dataset' by Cornell University (877 MB, 8.8 rating, 378 tasks), 'All Space Missions from 1957' by Agirlcoding (101 KB, 8.5 rating, 55 tasks), 'Handwriting Recognition' by landlord (1 GB, 9.4 rating, 24 tasks), 'Sales of summer clothes in E-commerce Wish' by Jeffrey Nvutu Mabilama (351 KB, 9.7 rating, 13 tasks), and 'Heart Failure Prediction' by Larxel (4 KB, 10.0 rating, 66 tasks). On the right side, there's a section titled 'Open Tasks' with five tasks listed: 'Are young people more severely...' (0 submissions), 'Develop a model for product suc...' (0 submissions), 'Predict patient illness' (2 submissions), 'Predict Likelihood of Admission' (82 submissions), and 'Exploratory Data Analysis' (12 submissions).

<https://www.kaggle.com/datasets>



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

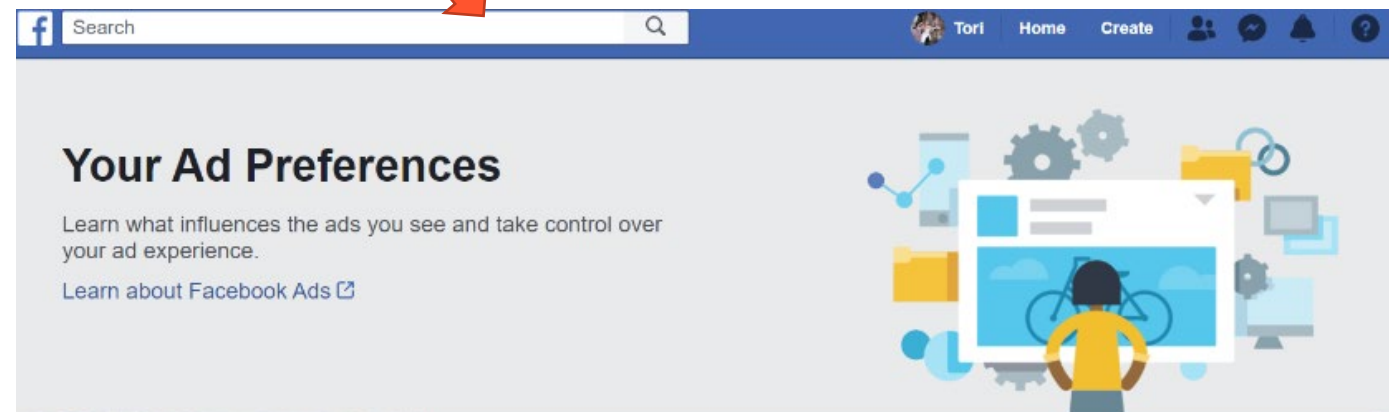
Lecture Tips

General Course Tips

Coding Tips

About Me

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



The Chronicle of
Higher Education



Game of Thrones



Well-being



Data science



Baby boomers



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

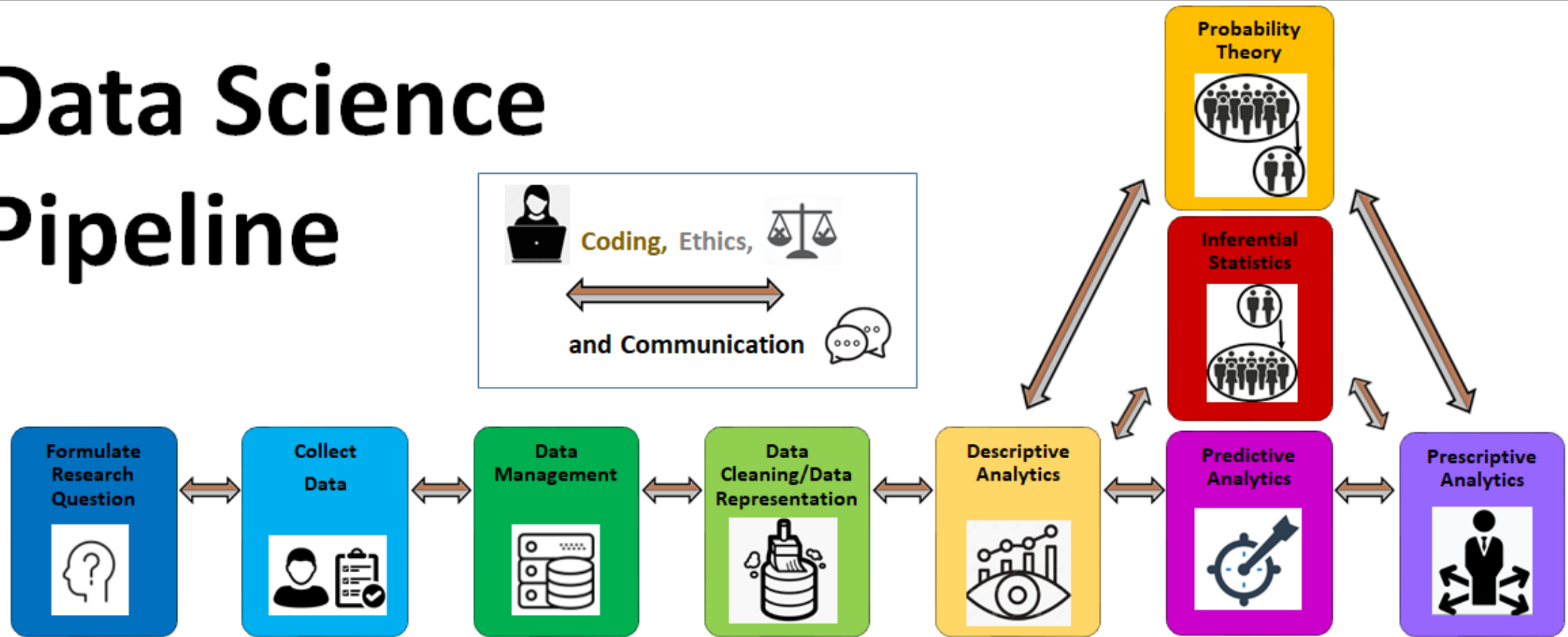
Lecture Tips

General Course Tips

Coding Tips

What is data science?

Data Science Pipeline



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

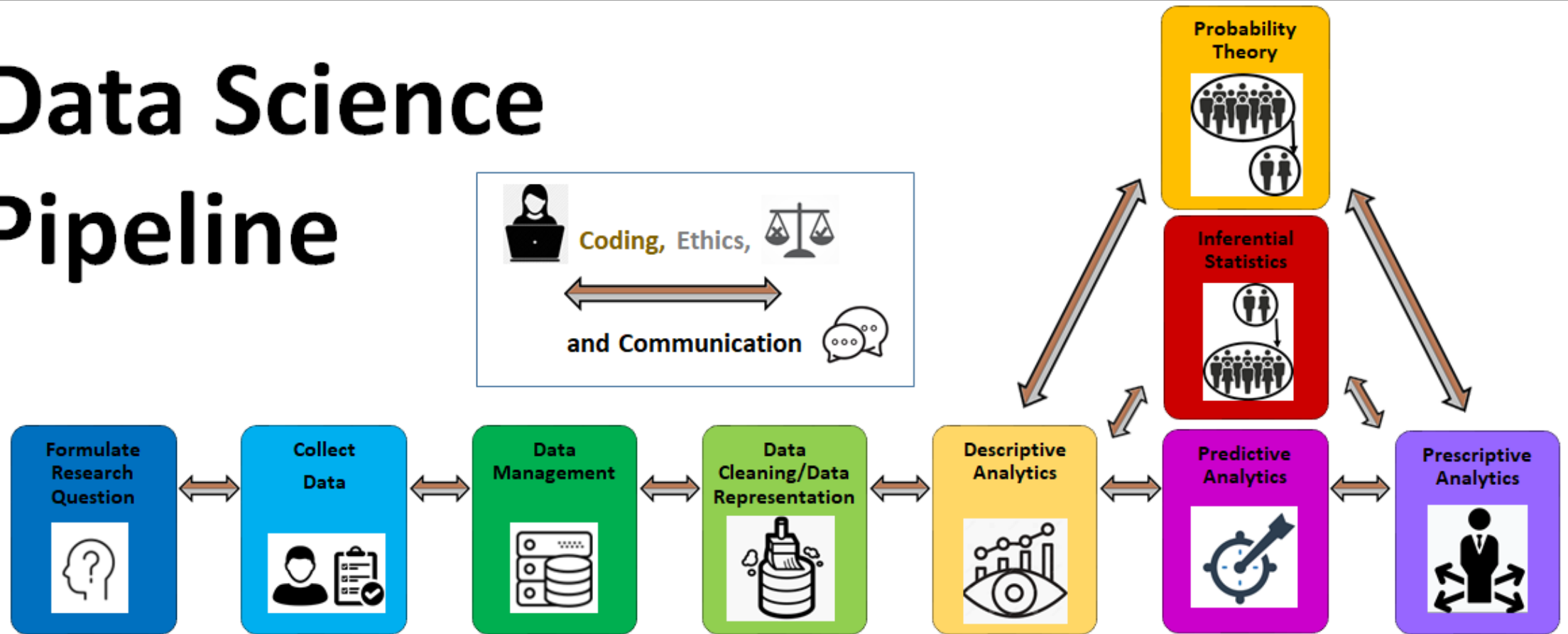
Lecture Tips

General Course Tips

Coding Tips

Example of “Full Data Science Pipeline”

Data Science Pipeline



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

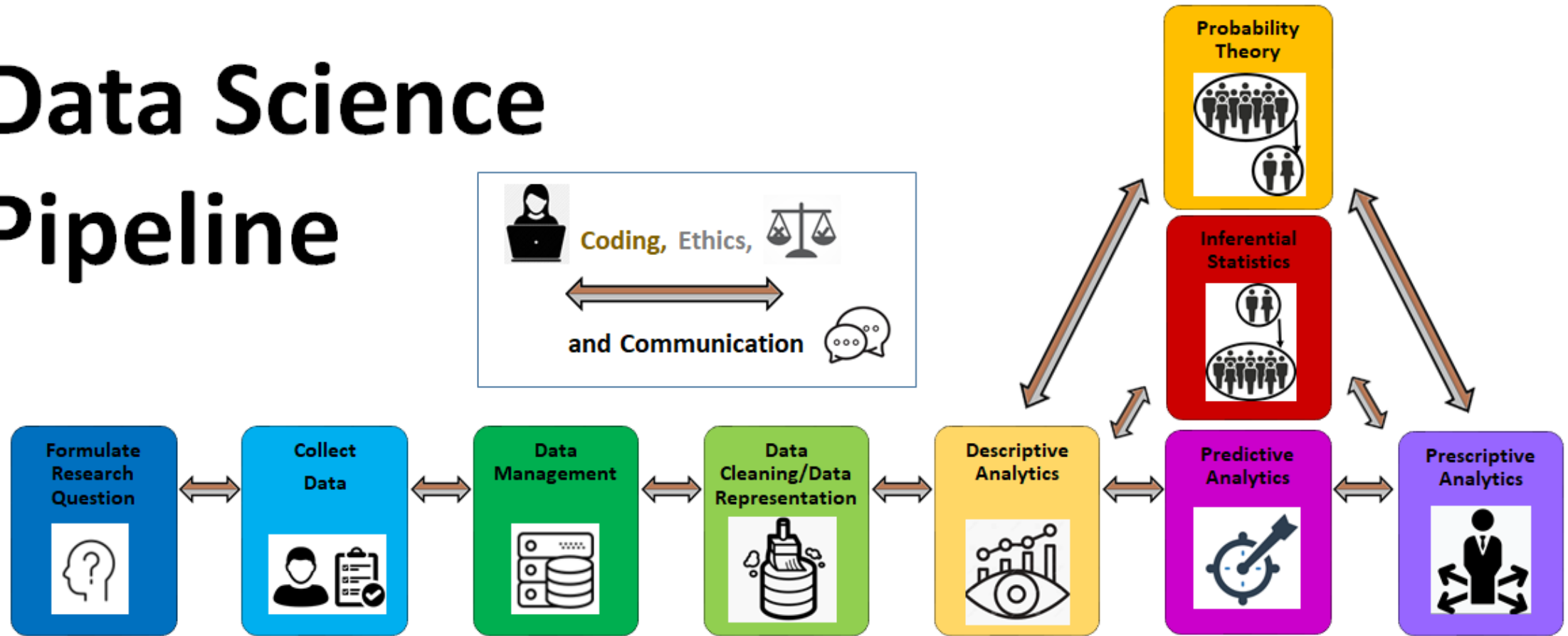
General Course Tips

Coding Tips

Data Science vs. Statistics vs. Computer Science



Data Science Pipeline



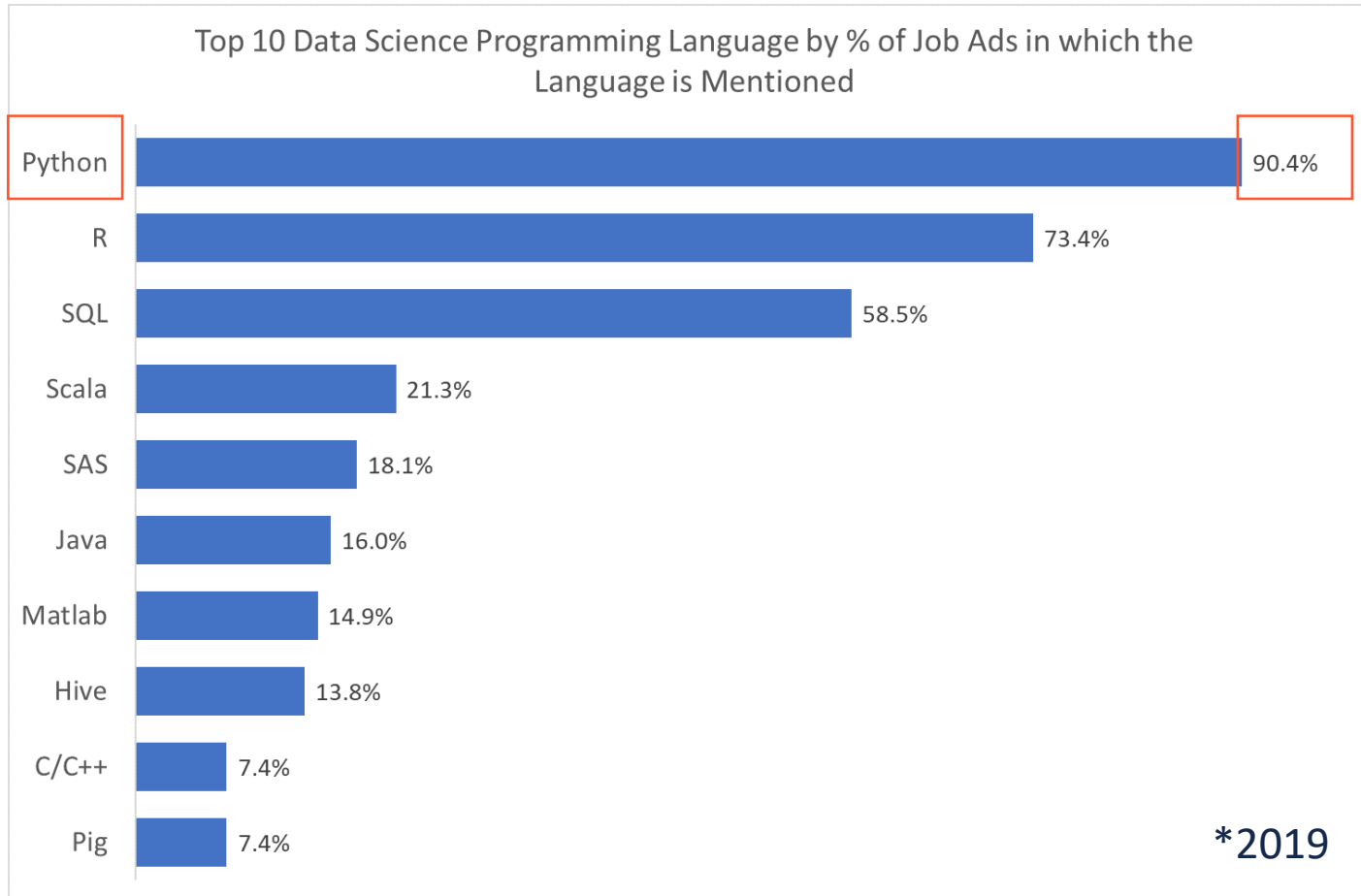
About you
About me
What is data science?
Example of the "Full Data Science

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?



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



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

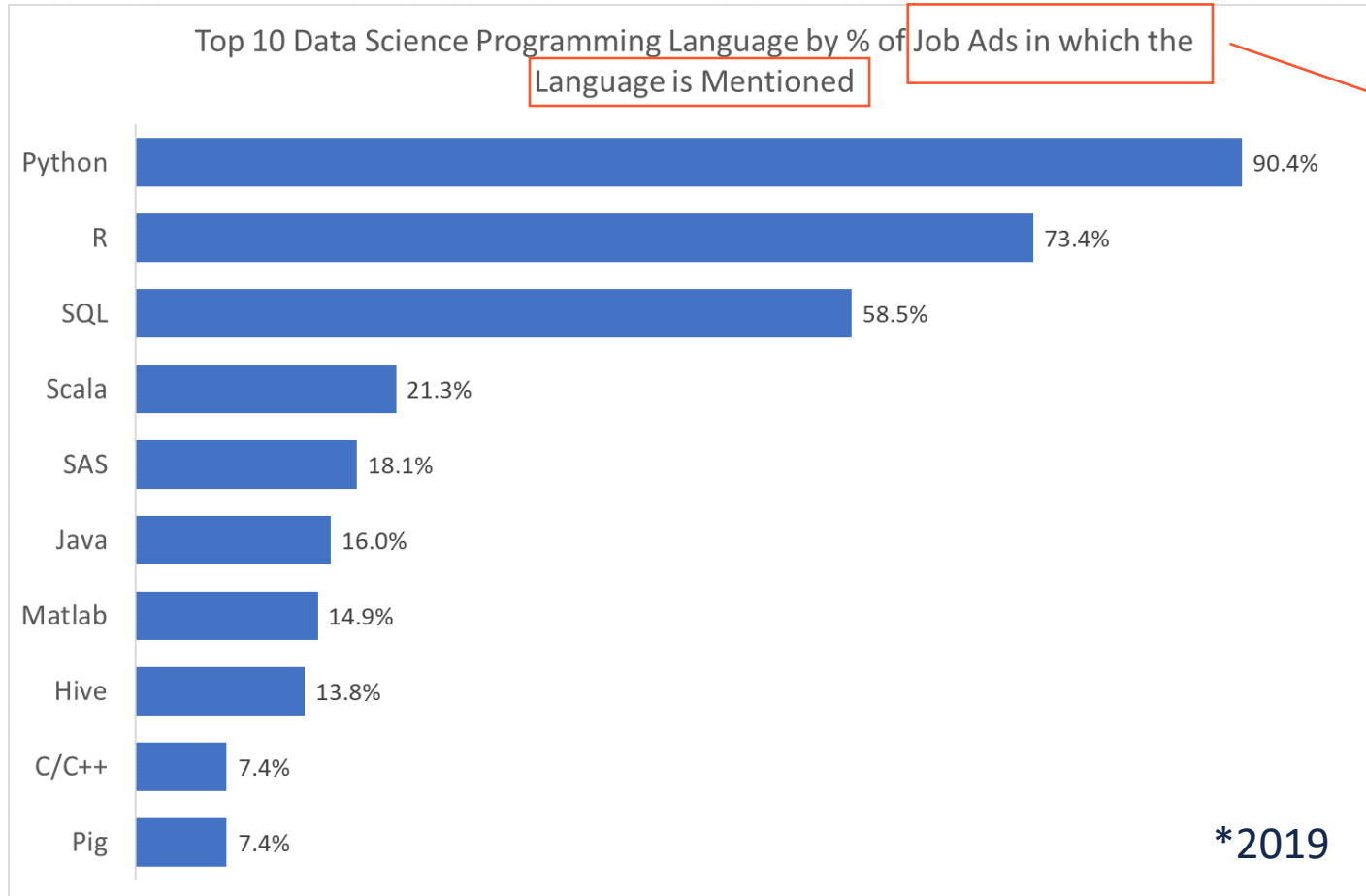
Learning Content Tips

Lecture Tips

General Course Tips

Coding Tips

Why use Python for Data Science?



What are some ways we could have collected this data?

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



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

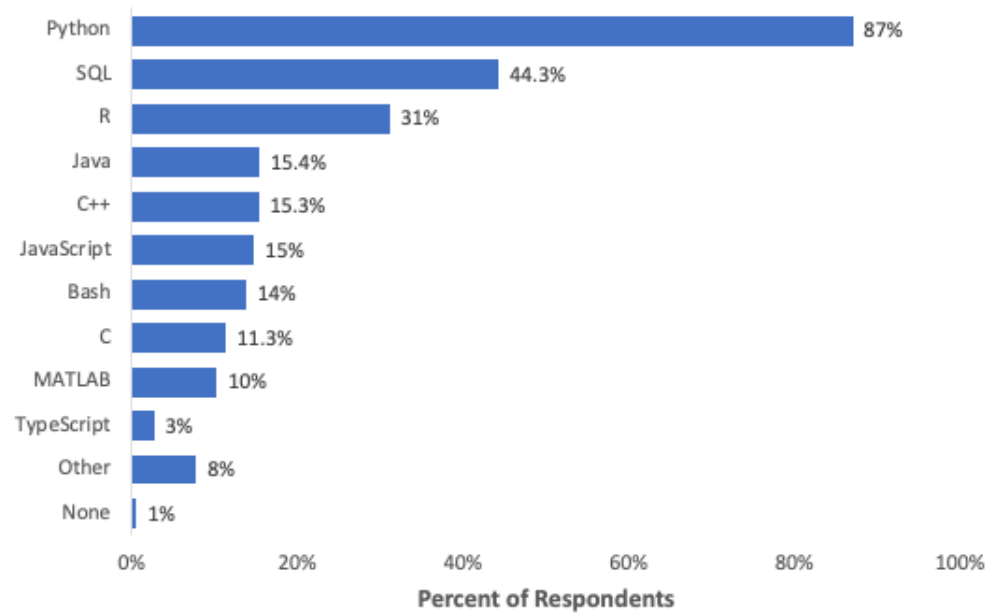
Lecture Tips

General Course Tips

Coding Tips

Why use Python for Data Science?

What programming languages do you use on a regular basis?



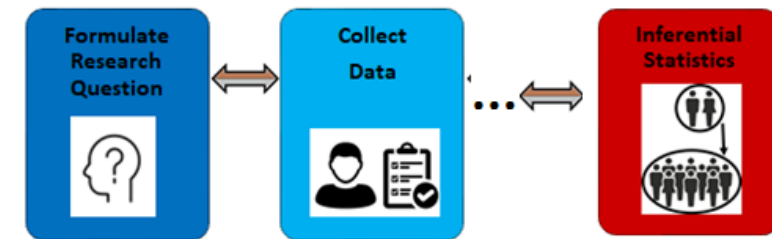
Note: Data are from the 2019 Kaggle ML and Data Science Survey. You can learn more about the study here: <https://www.kaggle.com/c/kaggle-survey-2019>.

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.



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

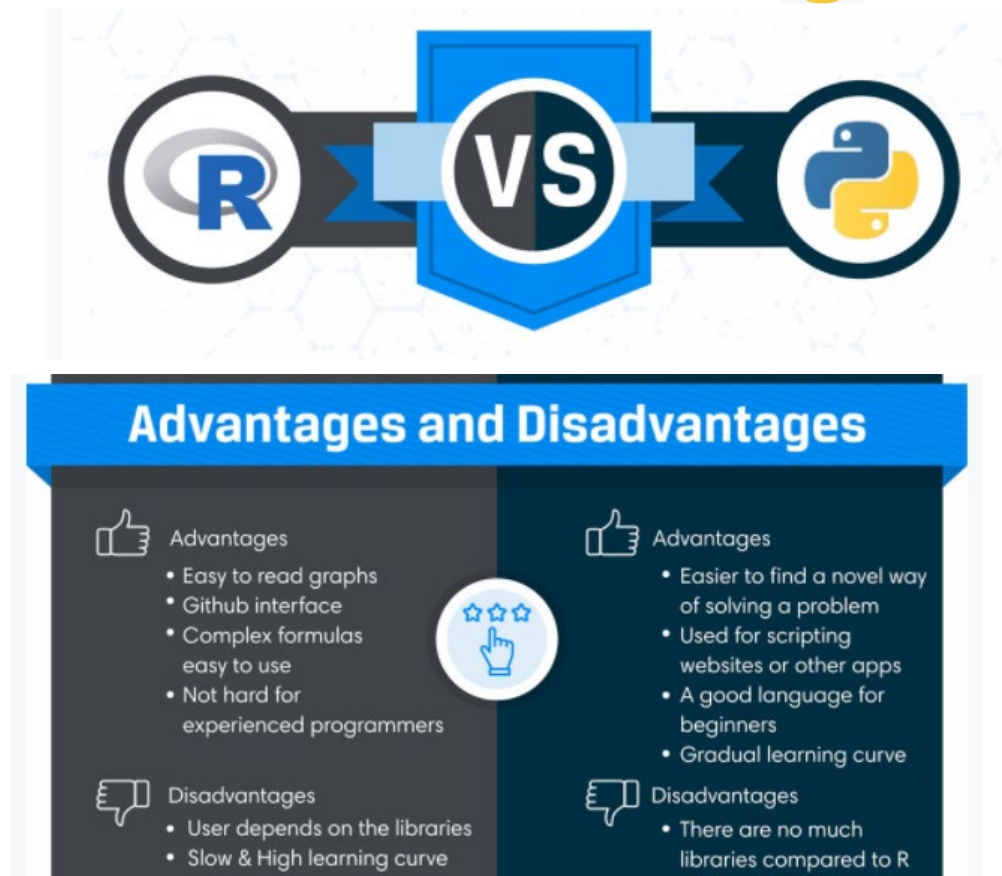
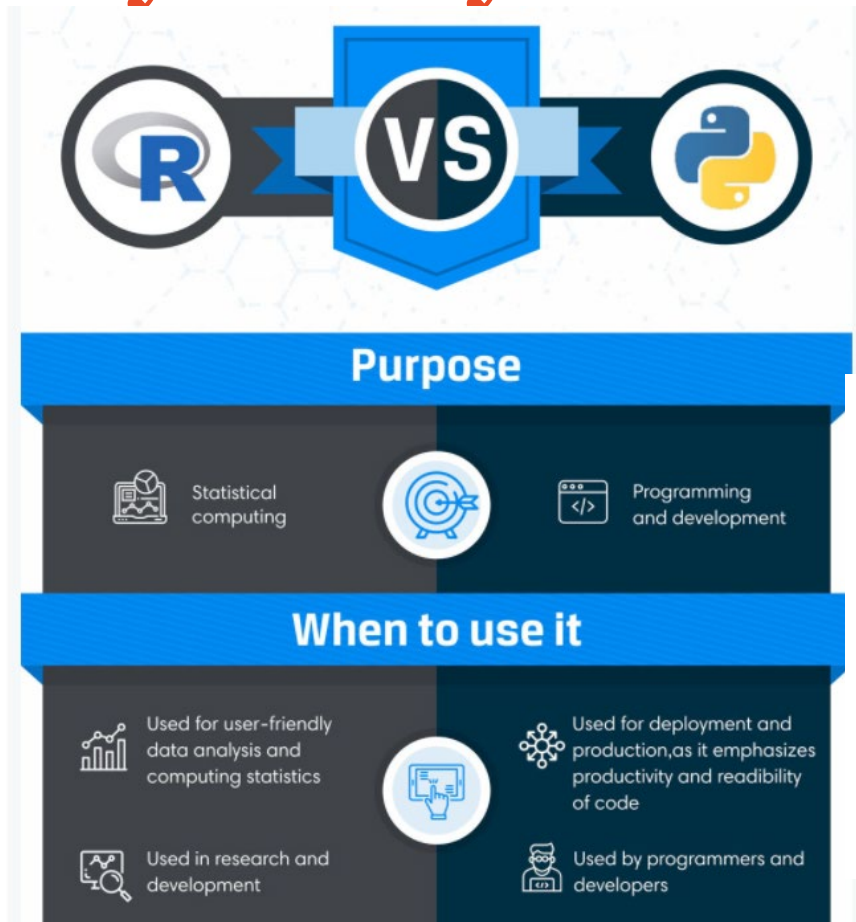
Learning Content Tips

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>



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 study data science?



Data Scientist Roles and Average Salaries (in \$)

Junior/Associate Data Scientist	91,000
Data Scientist	108,000
A.I./Machine Learning Engineer	127,000
Data Science Manager/Architect	140,000
Chief/Senior/Principal Data Scientist	146,000
Director of Data Science	169,000

Source: Dice.com

Dice

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



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

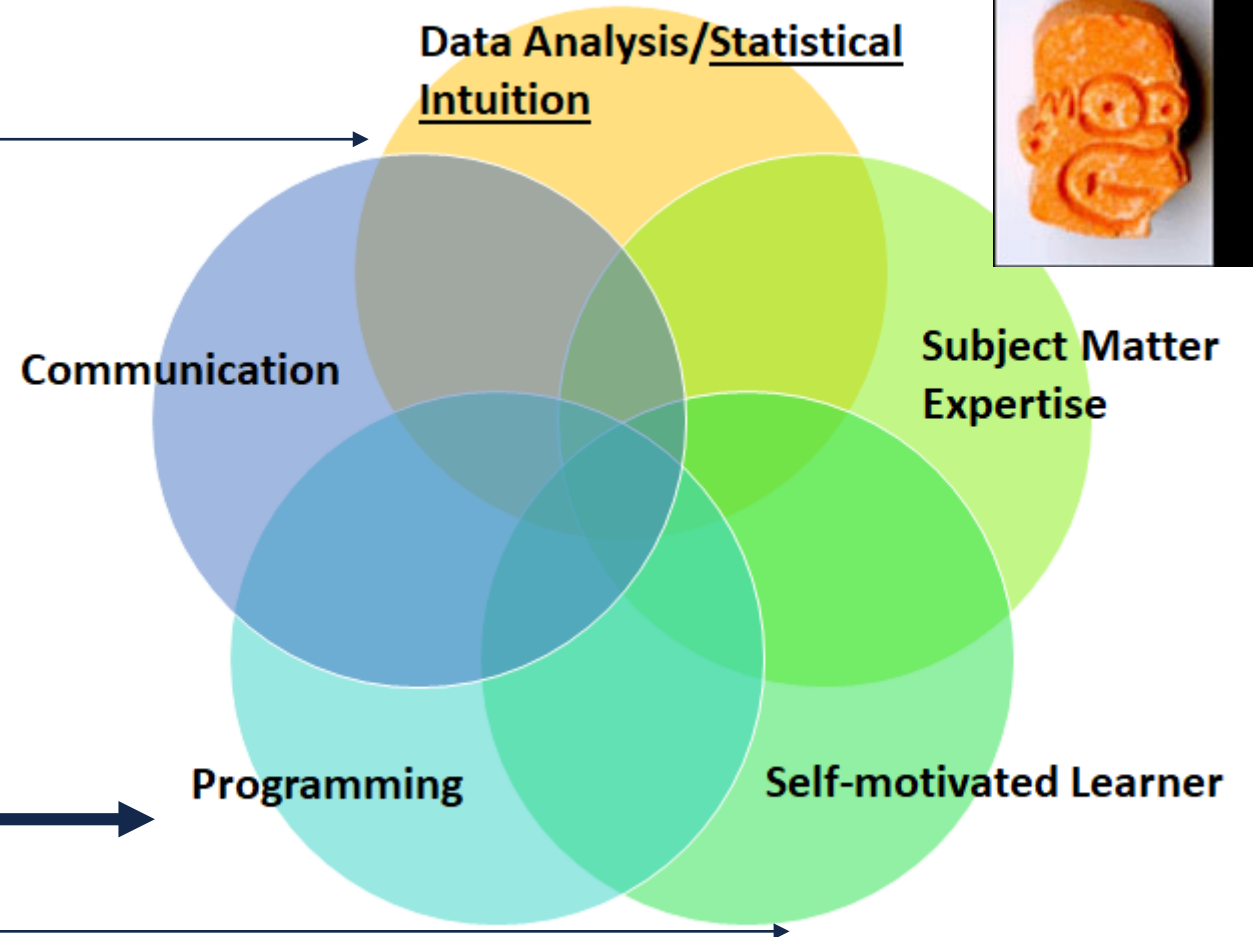
Coding Tips

Why 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>



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

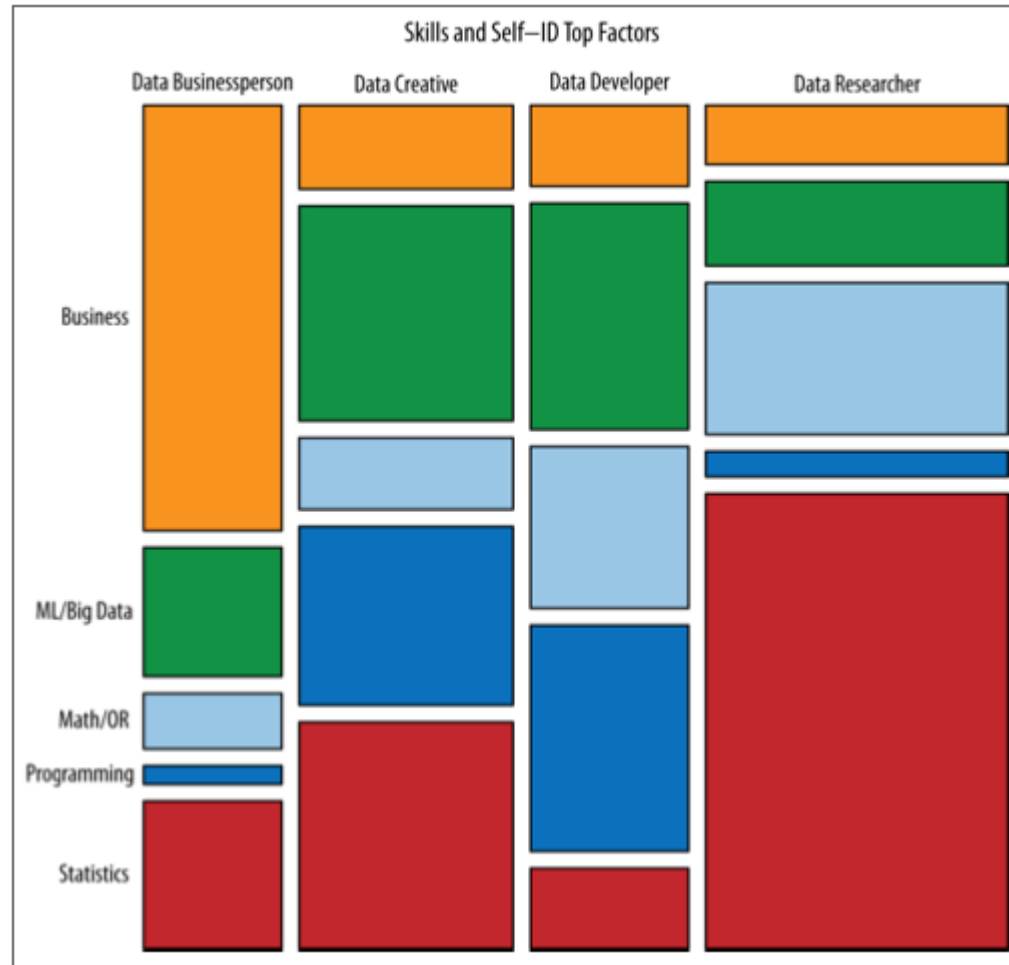
Learning Content Tips

Lecture Tips

General Course Tips

Coding Tips

Why study data science?



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

- **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.



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

Course Website and Syllabus

Course Website: <http://courses.las.illinois.edu/fall2020/stat207/>

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

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

- Zoom links for:
 - Lectures
 - Open labs
 - My office hours
- Videos Posted of the lecture
- Grades



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

Course Website and Syllabus

Piazza: <https://piazza.com/illinois/fall2020/stat207>

- Content and non-personal course related questions.

Github: <https://github-dev.cs.illinois.edu/stat207-fa20>

- **Fetch and merge** your weekly assignment, exams, and (optional) project
- **Push** your weekly assignments, exams, and (optional) project for grading
- Grades on your assignments



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

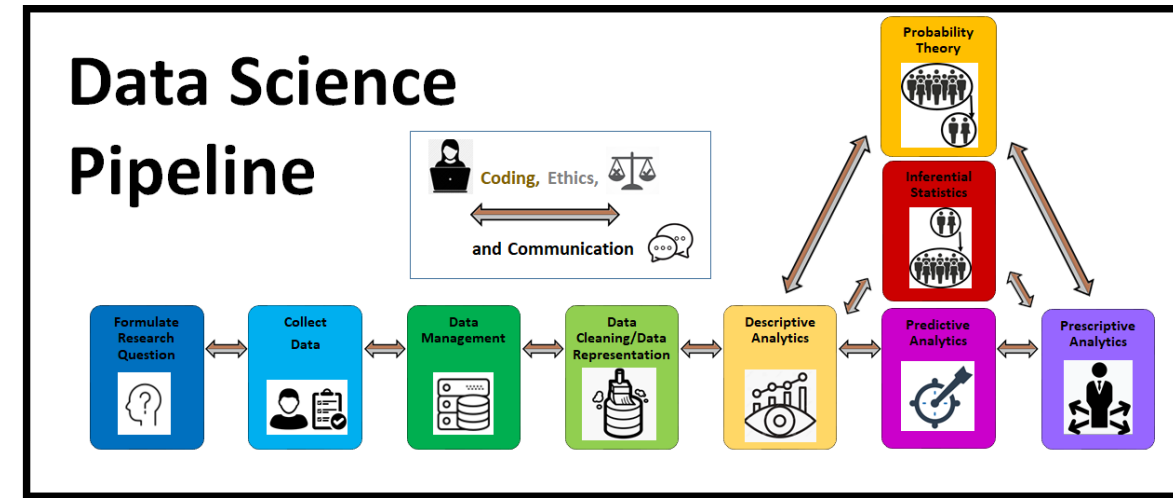
General Course Tips

Coding Tips

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?



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

Lecture Tips - Synchronous

- **Synchronous:** strongly encouraged if you are able to, but not required!
- **Each class download these (posted by 8am CST before class)**
 - Python Notebooks
 - Pdf
- **Note-taking Ideas**
 - Printing the pdf, hand written notes
 - Onenote (or other similar notetaking apps)
- **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.



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

Lecture Tips - Asynchronous

- **Expectation: Watch videos within 24 hours of posting**
 - Try watching with classmates
 - Try watching during office hours/lab to ask question.
 - Ask questions on Piazza.



[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](#)

General Course Tips

- Check your email regularly!
- Go to open labs
 - **Monday-Tuesday 5pm-7pm CST**
 - More coming soon (after survey)
- Go to office hours
 - **Friday 9:30-10:30am CST**
 - More coming soon (after survey)
- Start working on your early.
- Piazza can be helpful!
- Ask questions if you get stuck.

- New Idea: After class, write around 4 sentences describing what you just learned.



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