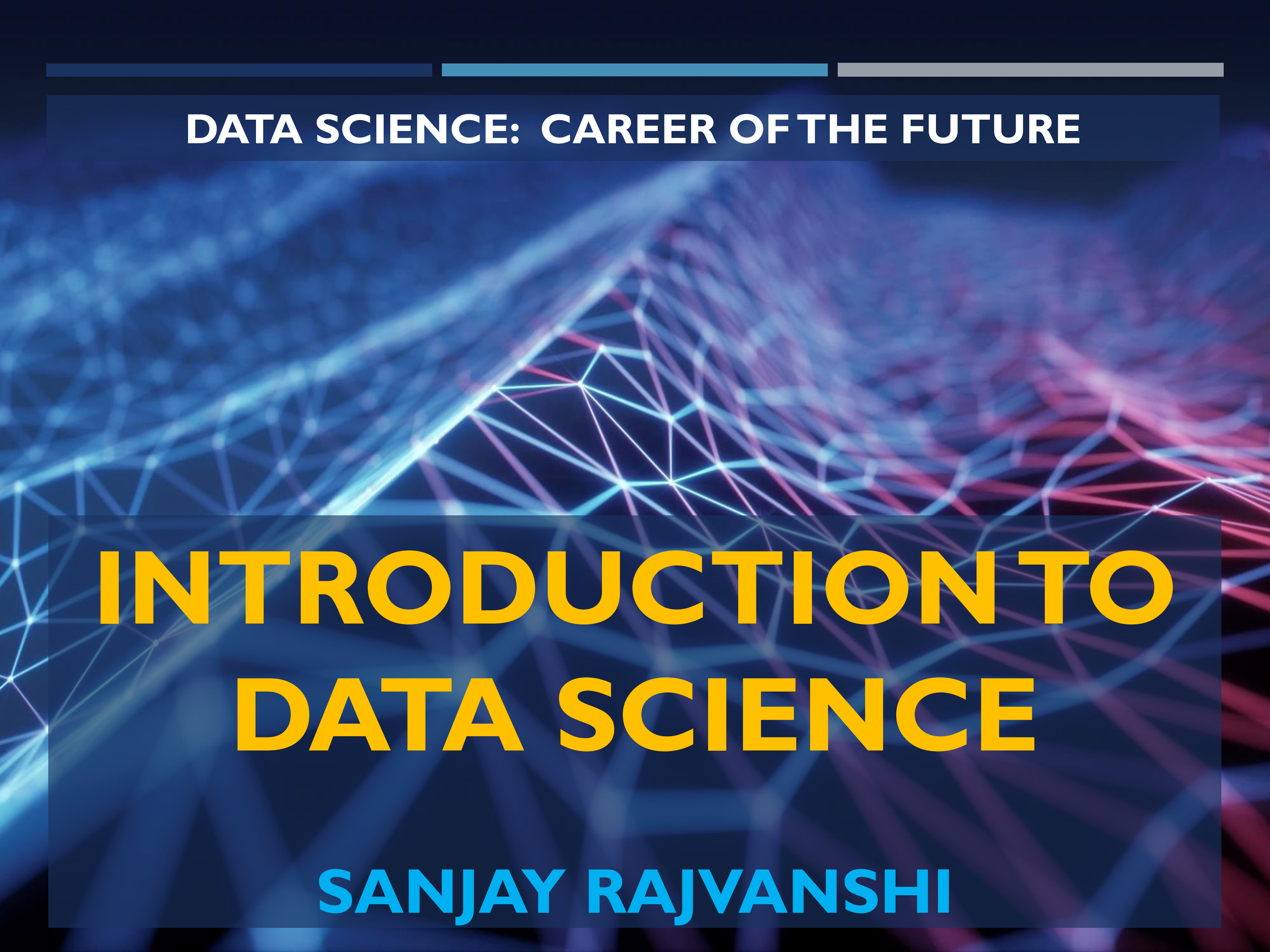


DATA SCIENCE: CAREER OF THE FUTURE

INTRODUCTION TO DATA SCIENCE

SANJAY RAJVANSHI

INTRODUCTIONS

- Your first name
- Your background – student (high school or college), professional
- Your motivation to take this class

LOGISTICS

- Interactive sessions
- Raise any questions and comments anytime – important to catch in the context
- No planned or scheduled breaks due to limited time – use your best judgement
- My contact email:
  - sr4public@gmail.com
- Course material will always be made available within 24-48 hours of class at:
  - https://github.com/sr4public/Intro-to-Data-Science-09252019
- Send an email to the library or me if you need additional information, instructions or want to provide inputs on what you would like to see covered in the class.
- Exercise extreme caution when on internet!

MY MOTIVATION

- Sr. Technology Leader with experience in technology and engineering for complex IT systems
- 20+ years of work experience at IBM, Hughes Network Systems
- Reasons that are driving me to teach this class:
  - Help prepare all generations for Data Science
    - Data Science is an emerging technology that is already impacting how we work
    - Offers great opportunities in coming years
    - Provide enough introductory information to enable your consideration as a career option (and college choices)
  - Give back to my IT profession
  - Give back to our community

SCHEDULE

Session	Date	Time	Topic
1	Sep 25	7:00 pm – 8:00 pm	Introduction to data science and associated tools.
2	Oct 2	7:00 pm – 8:00 pm	Introduction to Python. Learn how to use Python for data analysis. Python is simple, yet powerful language that is often used in data science.
3	Oct 9	7:00 pm – 8:00 pm	Data wrangling with Python. Learn how to gather data and make it useful for analysis.
4	Oct 16	7:00 pm – 8:00 pm	Data visualization and analysis with Python. Learn how to create useful visualizations to aid in the analysis of the data.
5	Oct 23	7:00 pm – 8:00 pm	Brief introduction to artificial intelligence and machine learning. Get a peek into how to make data based predictions.

Note: All classes are on Wednesdays.

SESSION 1: INTRODUCTION TO DATA SCIENCE AND ASSOCIATED TOOLS

DATA SCIENCE: WHAT

- Data science is a multi-disciplinary field that uses various types of analytical techniques to discover new insights and trends from the data. It uses tools to gain new understanding that may not be easily possible to gain by just a human review of the data. [1]
- *“Data science is the discipline of making data useful.”* [2]
- Prominent examples are Netflix and Amazon. Both use complex data science techniques to present list of options to a customer that may be more appealing to him/her (like which movie or shows to watch next in case of Netflix and which other products might be of interest to buy next in case of Amazon).
- University of Berkley also has a very good article on data science. [3]
- Note: you are not required to sign-up for an account on any of the sites to read these articles.

DATA SCIENCE: WHAT

- Example: Which colleges should I apply to?
  - Goal: Narrow down choices to your favorite schools with very high chances of success.
  - Process: Collect and analyze college admission related data to answer the question.

Available Information about Recently Admitted Students in Universities	Size of Data	Quality of Trends or Insights	Need for Tools, Automation
Your friends and relatives	--	--	--
All students from Montgomery county in last year (~11K, 12th grade, 2017, [4])			
All students from Maryland state in last year (~60K, 12th grade, 2017, [4])			
All students for all states in last year			
All students for all states multiple years			

DATA SCIENCE: WHAT

- Identified the problem, data, need for tools, automation.
- Iteratively analyze, visualize.

Available Information about Recently Admitted Students in Universities	Size of Data	Quality of Trends or Insights	Need for Tools, Automation
Your friends and relatives	--	--	--
All students from Montgomery county in last year (~11K, 12th grade, 2017, [4])	Higher	Better	May be
All students from Maryland state in last year (~60K, 12th grade, 2017, [4])	Higher	Better	High
All students for all states in last year	Much higher	Much better	Much higher
All students for all states multiple years	Much, much higher	Much, much better	Much, much higher

DATA SCIENCE: WHAT

- Data Analytics [5]
  - *"Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software."*
- Big Data [6]
  - *"Big data is high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."*
- Artificial Intelligence [7]
  - *"a branch of computer science dealing with the simulation of intelligent behavior in computers"*
  - *"the capability of a machine to imitate intelligent human behavior"*
- Machine Learning [8]
  - *"The intention of ML is to enable machines to learn by themselves using the provided data and make accurate predictions."*
  - *"ML is a subset of artificial intelligence; in fact, it's simply a technique for realizing AI."*

DATA SCIENCE: WHAT

- Various Data Science related roles [3]

Role	Description	Skills Needed
Data Scientist	Reviews what problems need to be solved and where to find the related data. Businesses use them to "*source, manage, and analyze large amounts of unstructured data*".	*"Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning"*
Data Analyst	Given a problem, they "*organize and analyze data to find results that align with the high-level business strategy*"	*"Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization"*
Data Engineer	Develop, deploy, manage, and optimize aspects of data and infrastructure.	*"Programming languages (Java, Scala), NoSQL databases (MongoDB, Cassandra DB), frameworks (Apache Hadoop)"*

DATA SCIENCE: WHY

- Data Science is an emerging technology that is quickly becoming mission-critical to more and more businesses. Recruiting data scientists or data science professionals is a top challenge these businesses are facing.
- A 2018 KDnuggets blog published on the internet states that "*Job growth in the next decade is expected to outstrip growth during the previous decade, creating 11.5M jobs in the Data Science/Analytics area by 2026, according to the U.S. Bureau of Labor Statistics.*" [9]
- *"There's a widening gulf between the needs of organizations and the abilities of job candidates to fulfill those needs."* [10]
- *Data science — the study of computer-generated "big data" — is the hottest career in the U.S., according to Glassdoor. And now it's the hottest math class at a growing number of California high schools. About 30 high schools in California have started offering data science classes for juniors and seniors, in some cases as an alternative to Algebra 2.* [11]
- Georgetown University offered a summer program in data science in 2019. [12]

DATA SCIENCE: HOW

- System implementation in general follows the pattern below:
  - Problem Definition (or Requirements)
  - Analysis
  - Design
  - Implementation
  - Maintain and support
- Implementation is supported by
  - Lifecycle process
  - Tools
  - Standards, guidelines, etc.

DATA SCIENCE: HOW

- Data Science solution lifecycle (iterative):
  - Problem identification
  - Identify data
  - Clean, transform data
  - Analyze, visualize
  - Identify algorithm(s)
  - Implement
  - Maintain and support

- Tools
  - Python and associated libraries
  - Jupyter notebooks
  - Kaggle.com
  - Github
  - Anaconda
- Standards, guidelines, etc.
  - Defer for now
- Always pay attention to and follow the T&Cs.

SETUP ENVIRONMENT

- Environment:
  - Python and associated libraries
  - Jupyter Notebook
  - Optional:
    - Kaggle.com
    - Github
- Refer to the environment setup document ("Intro to Data Science-Env Setup-09252019.pdf") provided earlier for installation instructions.

REFERENCES

- No text book is required for this course.
- Below are some helpful resources if interested.

Website	Description
https://www.python.org/	Official website for Python
https://pandas.pydata.org/	Open source library providing data structure and data analysis tools

Book Name	Author	Publisher	Comments
Data Science from Scratch: First Principles with Python	Joel Grus	O'Reilly	Purely based on topics listed in the table of contents and reviews on Amazon
The Signal and the Noise: The Art and Science of Prediction	Nate Silver	Penguin	Interesting book

FIRST EXERCISES

- Two Exercises
  - Hello World!
  - Retrieve data from a file
- Open Jupyter Notebook. Go to the corresponding browser window.
- Create a new folder “DSIntro”
  - Click on “New” and select “Folder”.
  - You will see a new folder “Untitled Folder”. Click on the checkbox next to it. And click on “Rename”. Enter “DSIntro” in the box and click “Rename”.
  - Double click on “DSIntro” folder to open it.
- Create a new folder “Exercise” in that new folder ‘DSIntro” folder using similar steps as listed above. Double click on “Exercise” folder to open it.

EXERCISE-1: HELLO WORLD!

- Create a Python file by clicking on “New” and selecting “Python 3”. Change the name of the notebook to “S1-Ex1”.
- Enter the following line of code in the first cell (create new cells by clicking on “+”)

```
print (“Hello World!”)
```

- Click on “Run”.
- You will see the “Hello World!” output below the cell.
- If you see “Hello World!” then Congratulations! You have verified your installation successfully.

EXERCISE-2: RETRIEVE DATA FROM A FILE

- Create a text file by clicking on “New” and selecting “Text File”.
- Enter the following in that file:

```
Number,Name,Term
1,George Washington,1789-1797
2,John Adams,1797-1801
3,Thomas Jefferson,1801-1809
4,James Madison,1809-1817
5,James Monroe,1817-1825
6,John Quincy Adams,1825-1829
7,Andrew Jackson,1829-1837
8,Martin Van Buren,1837-1841
9,William Henry Harrison,1841-1841
10,John Tyler,1841-1845
```

EXERCISE-2: RETRIEVE DATA FROM A FILE

- Change the name of the file to “S1-Ex2-US-Presidents.csv” by clicking on “untitled.txt” at the top and entering the new name in the box. Click “OK”.
- Save the file by clicking on “Save” under the “File” menu.
- Now switch to the other tab (name of the tab should be “DSIntro”).
- Now create a Python file by clicking on “New” and selecting “Python 3”. Change the name of the notebook to “S1-Ex2”.
- Enter each line of following code in one cell each (click on "+" to create a new cell)

```
import pandas as pd

listOfPresDf = pd.read_csv ('S1-Ex2-US-Presidents.csv')

listOfPresDf.head ()

listOfPresDf.head (10)
```

EXERCISE-2: RETRIEVE DATA FROM A FILE

- Click on “Run” four times to execute code in each of the four cells.
- You will see the content of the file under the last two cells – first one shows a default number of data rows and the second one shows 10 data rows from the file.
- Now enter all lines of code in one cell and observe the behavior when you click on "Run". What changes would you make to display outputs like the first run?

EXAMPLE – A BRIDGE TOO FAR?

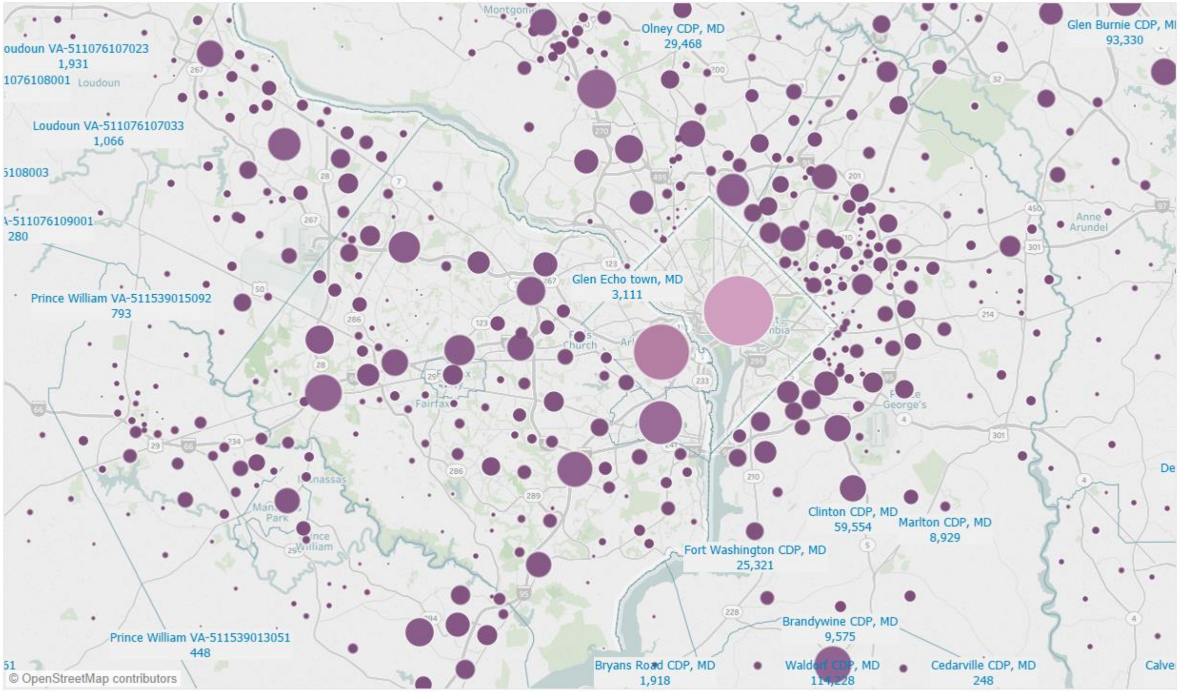
- Problem: Where can we build an additional bridge over the Potomac River to ease the traffic bottleneck from the American Legion bridge?
- Data:

Column Name	Type	Description
census_block_group	String	Census block group (CBG) visited
visitor_home_cbgs	JSON {String:Int}	Origin home CBGs for each CBG visited
visitor_work_cbgs	JSON {String:Int}	Work-location CBGs for each CBG visited

- Identified data from a company called SafeGraph.
- Cleaned, transformed data; Analyzed, visualized; Identified algorithm; Implemented
- Objective for discussing this example is to give you a feel for analysis, visualization for a potential real-life problem.
- Work in progress, pictures shown are not final and could change.

EXAMPLE – A BRIDGE TOO FAR?

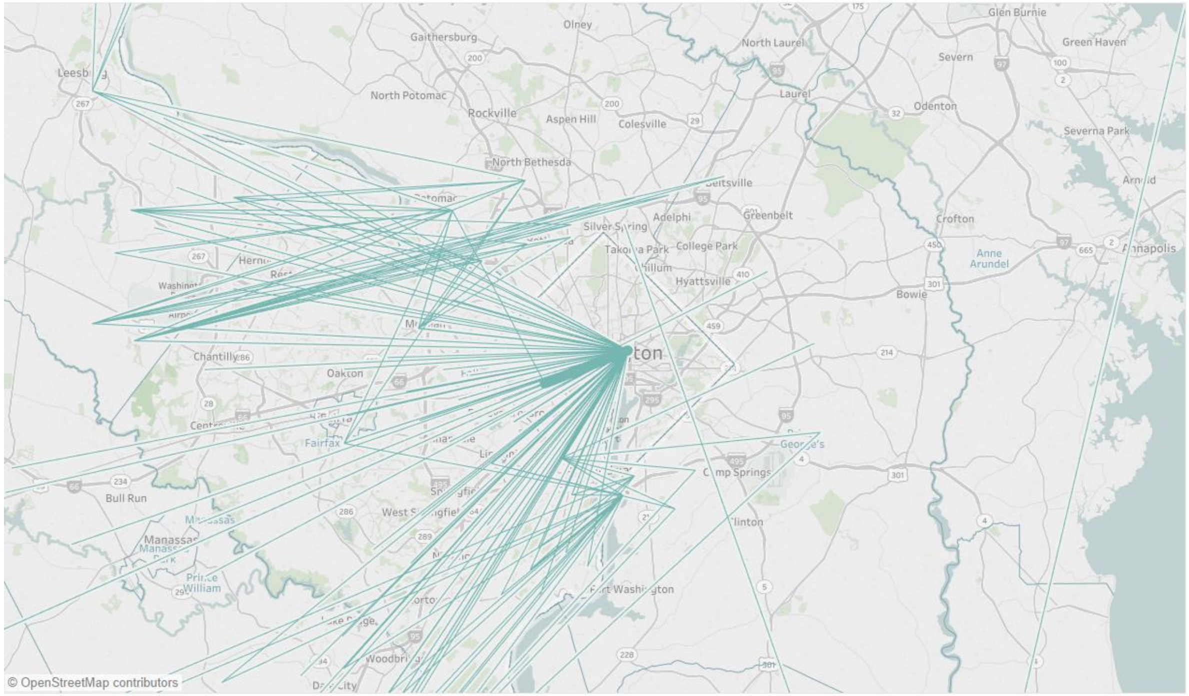
- Picture shows various places in DMV with size representing the number of visitors.
- Tableau software package was used to create the pictures.



© OpenStreetMap contributors

EXAMPLE – A BRIDGE TOO FAR?

- Applied clustering technique.
- Picture shows traffic pattern in one of the clusters in DC-MD-VA metro area.



© OpenStreetMap contributors

SESSION 1 – RECAP

- Data Science: What, Why, How
- Associated tools and environment
- Examples
- Exercises

SESSION 1 – HOMEWORK

- Do more Python programming practice from Python website
  - Refer to the website: https://docs.python.org/3/tutorial/
  - Read sections 1, 2.
  - Try out examples from sections 3, 4.

SESSION 2 – AGENDA

- Python is simple, yet powerful language that is often used in data science.
- Introduction to Python – continue learning about Python. Focus will be on aspects more relevant to Data Science.
- Learn how to use Python for data analysis. We will start to learn how to make the data suitable for the problem, clean/convert/transform it – sometimes referred to as data wrangling or data munging.

REFERENCES

Note: you are not required to sign-up for an account on any of the sites to read these articles.

1. *Data Science* – https://en.wikipedia.org/wiki/Data_science
2. *What on earth is data science?* – https://www.kdnuggets.com/2018/09/what-is-data-science.html
3. *University of Berkley data science article* – https://datascience.berkeley.edu/about/what-is-data-science/
4. *Enrollment in Maryland Public Schools* – http://www.marylandpublicschools.org/about/Documents/DCAA/SSP/20172018Student/2018EnrollbyRace.pdf
5. *data analytics (DA)* – https://searchdatamanagement.techtarget.com/definition/data-analytics

REFERENCES

6. *Data Science vs. Big Data vs. Data Analytics* – https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
7. *Merriam-Webster* – https://www.merriam-webster.com/dictionary/artificial%20intelligence
8. *Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences* – https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb
9. *How many data scientists are there and is there a shortage?* – https://www.kdnuggets.com/2018/09/how-many-data-scientists-are-there.html
10. *What's Driving the Demand for Data Scientists?* – https://knowledge.wharton.upenn.edu/article/whats-driving-demand-data-scientist/

REFERENCES

11. *'Big data' classes a big hit in California high schools* – https://edsource.org/2018/big-data-classes-a-big-hit-in-california-high-schools/593838

12. *Georgetown University summer program in data science* – https://summer.georgetown.edu/programs/SHS02/introduction-to-data-science-academy