UMBC

DATA 601 – 03 | Spring 2022
Lecture 01: Introduction

Dr. Ergun Simsek

About me…

- Real Office: ITE 325 K
- Virtual Office: https://umbc.webex.com/meet/simsek
- Office Hours: Tuesdays 10:00 am – 12:00 noon
  - Please email me 48 hours in advance to secure a spot in my office hours. If you have questions regarding

- GitHub: https://github.com/simsekergun/DATA601
- Tutors: https://dil.umbc.edu/home/resources/graduate-tutors-spring-2022/

Ground Rules

- Schedule: 7:10 – 8:20, Break, 8:25 - 9:40
- You can take a bathroom break at any time, no need to ask
- I value being punctual (start of class, break, end of class)

- Raise your hand if you have a question
- Don't apologize for asking a question or for not knowing something

- I find it acceptable for you to occasionally not participate

- Tell me if you cannot hear me or if you cannot understand me
- Slides will be provided after lecture

- I value your feedback:
  - Direct: verbal. Indirect: anonymous question/comment sheets on your desk

Ground Rules (Cont…)

- Please email me in advance if you are going to miss a class
- If you have COVID symptoms, please do not come to class, email me (I will share my screen via BB Ultra)
- Discussions on HWs are more than welcome, code sharing, copy+paste+modify is not acceptable
  - First catch: 0 without explanation
  - Second catch: F without explanation

Ground Rules (Cont…)

- Please submit your HWs via Blackboard
- You will need to save your as a jupyter notebook
- The notebook has to be in a working condition
- The grader will do “Restart the kernel and run all cells”, if there is an error, he will not continue grading
- The name of the file should be in the following format

  Lastname_HWXY.ipynb, e.g. Simsek_HW01.ipynb
- Late HWs will not be accepted

Schedule and Grading

- Note that our syllabus and weekly schedule are totally tentative. We might speed up, slow down, remove, add, etc.

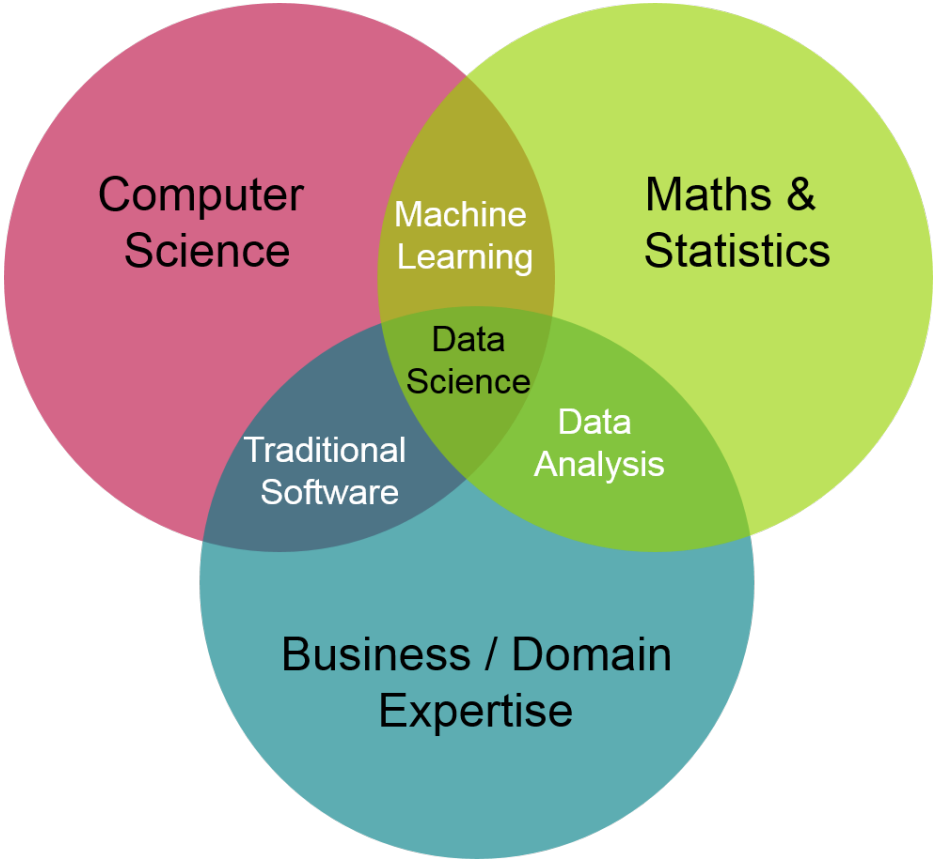
- Tentative Grading
  - Attendance (5 %)
  - 4 Quizzes (9%)
  - 7 Homework (56%)
  - 2 Projects (30%)

94-100	—	A
88 - 93	—	A-
83 - 87	—	B+
77 - 82	—	B
71 - 76	—	B-
66 - 70	—	C+
60 - 65	—	C
50 - 59	—	D
0 - 49	—	F

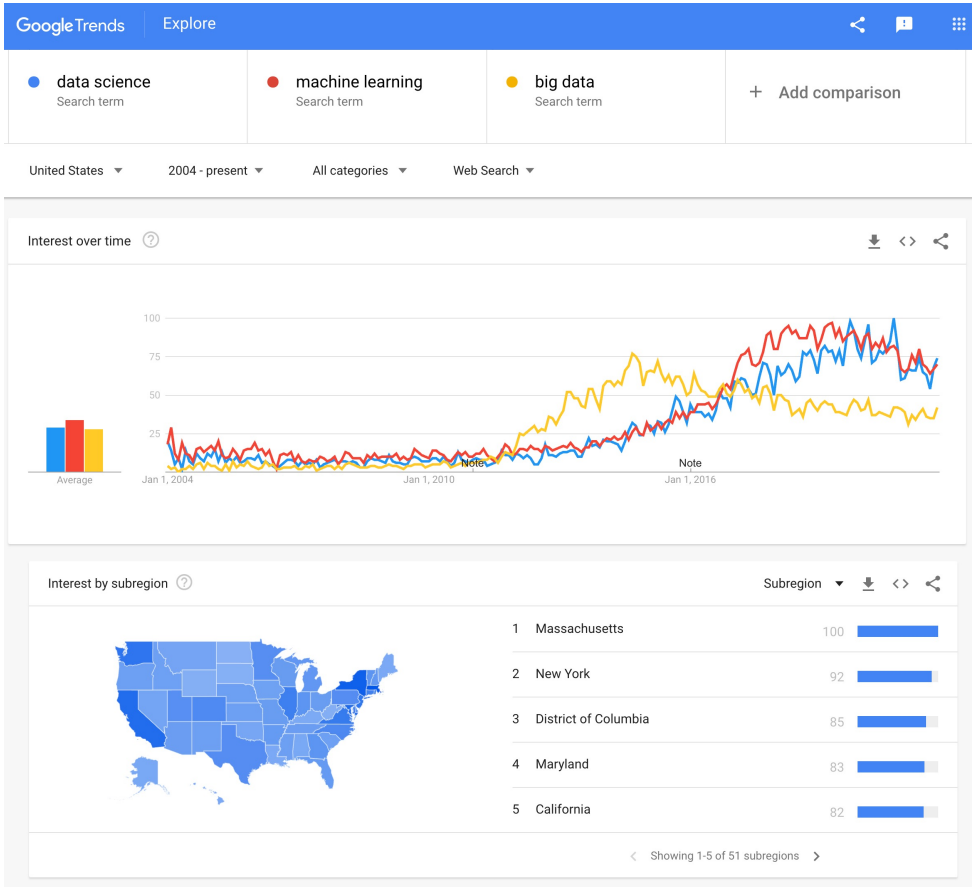
UMBC

WHAT IS DATA SCIENCE?

What is Data Science?



Interest on Data Science



DS is an active field with lots of jargon

There will always be something you haven't heard of before.

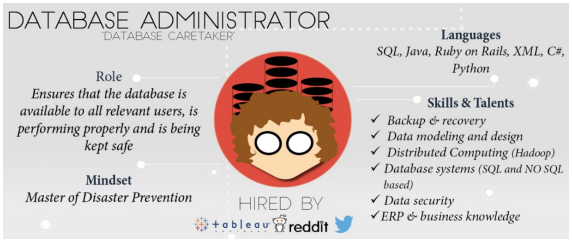
- Know enough to be conversant with peers
- Be curious about new topics
- Research concepts and labels before using them

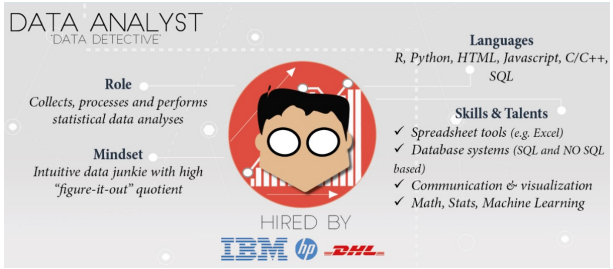
Reference: http://www.datascienceglossary.org/

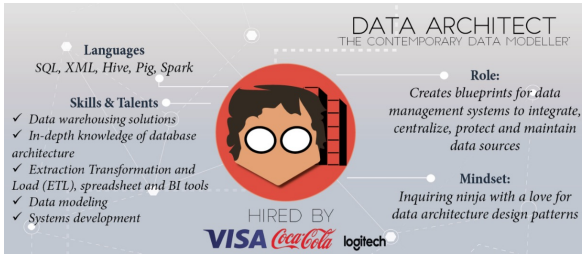
The World of DS is Huge. Don't get lost!

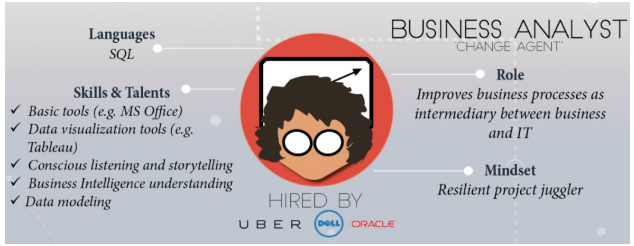
Skills and experience matter

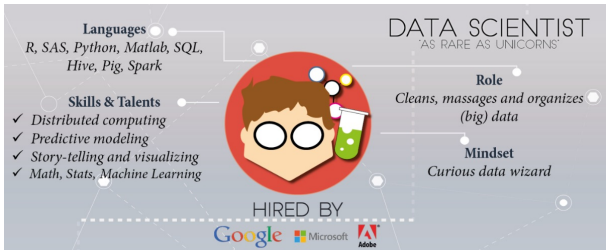
more than title and labels

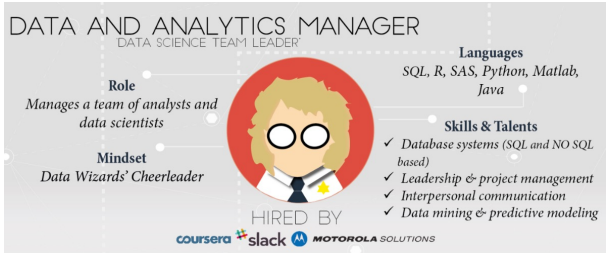












https://www.datacamp.com/community/tutorials/data-science-industry-infographic

Historical progression: data grooming, data mining, data scientist

Why learn data science?

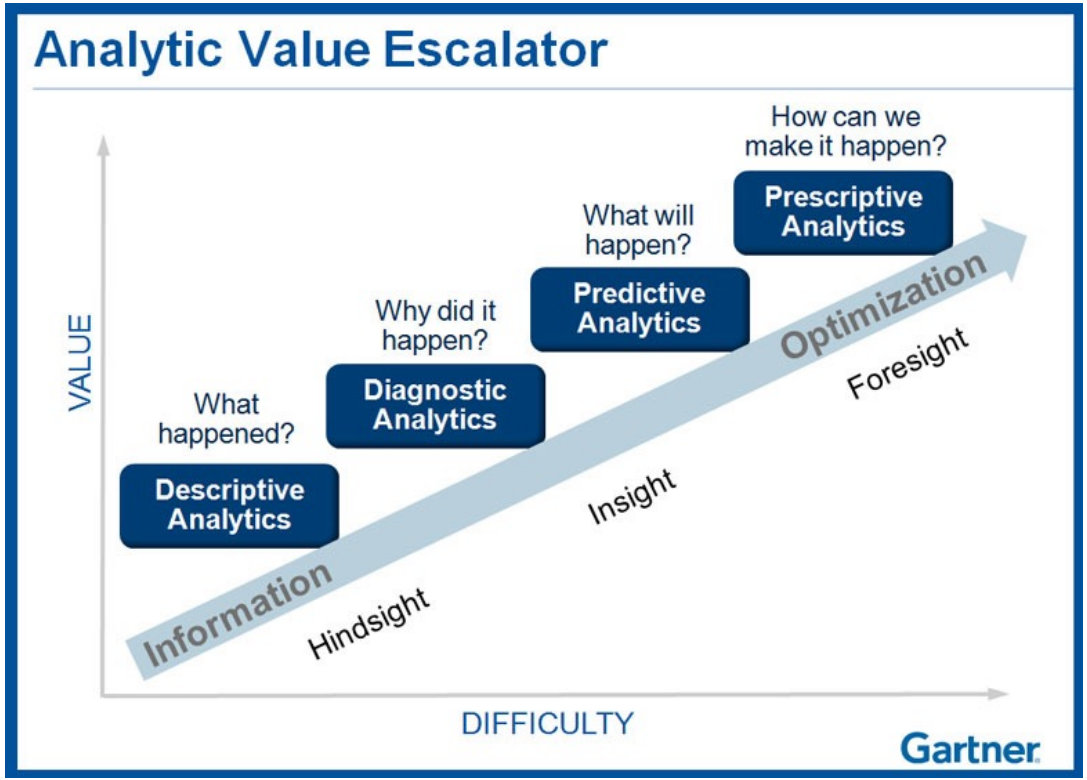
Explore: **identify patterns**

Predict: **make informed guesses**

Infer: **quantify what you know**

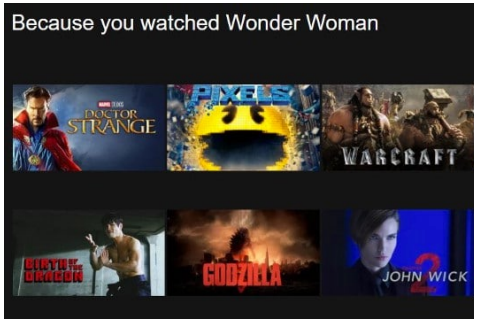
Motives:

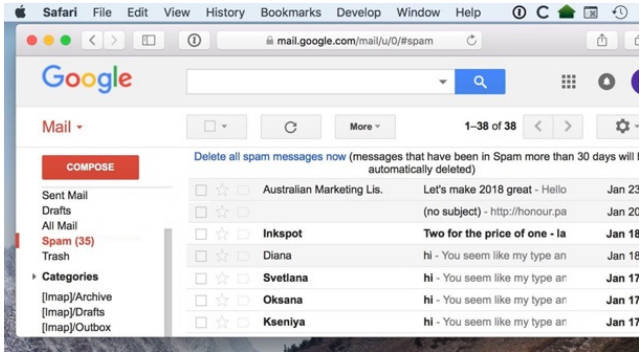
- Make money
  - Employment
  - Promotion
- Help people
- Gain new knowledge

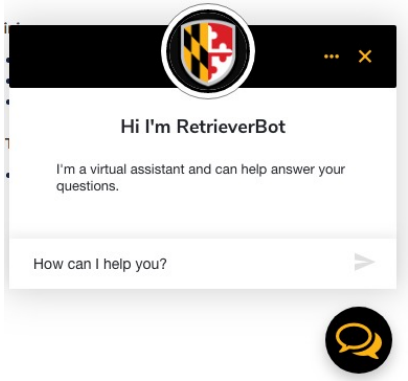


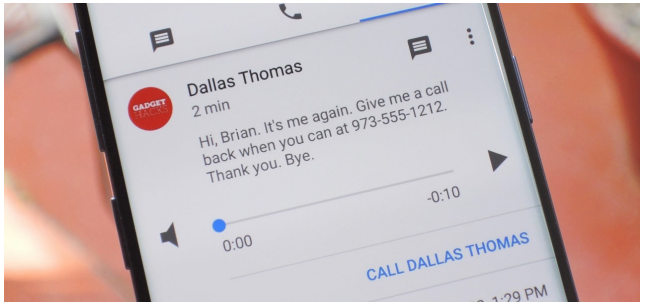
Large scale use cases with lots of data

- Google's search engine
- Bank and Credit Card fraud detection
- Logistics (DHL, UPS) of fleet management
- Healthcare records from patients









Each depends on availability of compute and data

Assumption in this class

- In class we will assume you are a lone data scientist on an island with an internet connection.

- This is not the typical case -- you'll have coworkers, customers, bosses, competitors, collaborators, peers.

Example of how class ≠ real world

- This class will not use competitive grading. (Imagine if it were.)
- As an employee at a company, you may be competing for a bonus or promotion

--> consequence: personal and organizational politics factor into the work environment

Small scale use cases with not much data

As a business employee or bureaucrat or politician

- How do I improve decision making process?
- How do I evaluate the outcome of decisions?
- How do I decrease the risk when faced with an opportunity?
- How do I convince other stakeholders of the best course of action?

While not taking too much time, spending too much money, using the resources I already have access to, and in a way that is convincing?

Logistics

Python with

- Anaconda
  - Jupyter
- Google Colab

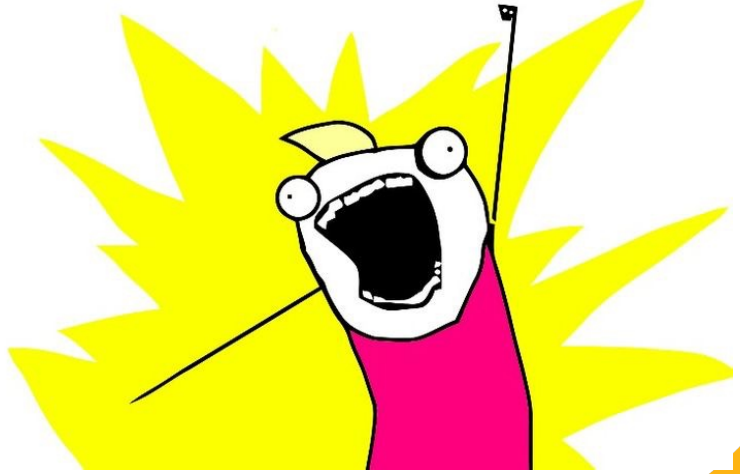
- Google Drive
- GitHub or Blackboard

Why Jupyter + Python for Data 601?

Jupyter is useful for

- Exploration of data (*jargon*: EDA = exploratory data analysis)
- Documenting your activities (to enable reproducibility)
- Figuring out which software is relevant, which algorithms to use, which software libraries are useful
- Visualizing results

And both Jupyter and Python are free!
And both are widely used!



Python and Jupyter do not cover every use case

- For sufficiently large data sets, Jupyter and Python are not the right tool
- For sufficiently complex analytics, Jupyter and Python are not the right tool

Speed and security are typically not your priority during exploration

Knowing when to invest in switching tools is a skill

Evaluate trade-offs of flexibility and security and speed for a given scale

Relevance of infrastructure to data science

Usual explanation when replicating analysis:

1. Get this data
2. (*Documentation*) Apply this transformation to get result

No explanation of

- software used
- software versions
- configurations
- Implementation details

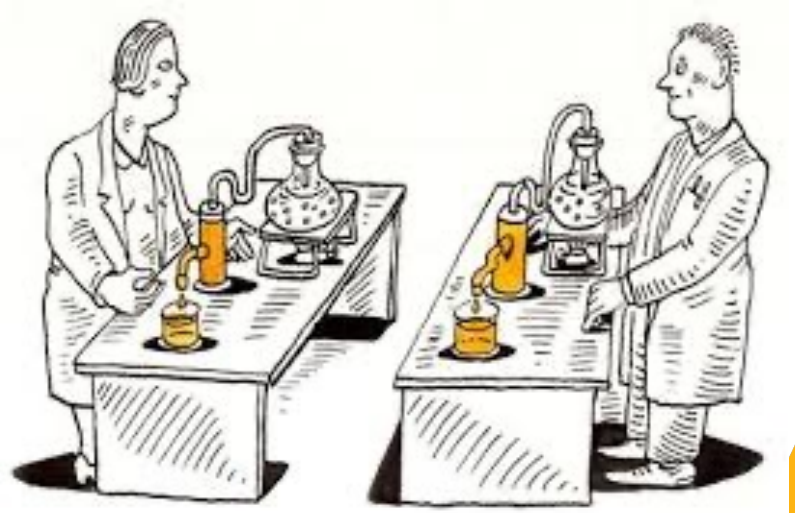
Digital archeology:
Suppose you are to diagnose why someone else's approach doesn't yield same results
Suppose they did their work 20 years ago

Infrastructure in data science to enable

Reproducibility and Portability

In addition to data and analysis, implementation and environment matters

1. Use this Operating System
2. Install this software
3. Configure software this way
4. Add these packages
5. Get this data in this format
6. Run analysis against data
7. Create plots
8. Generate report



Best practices: Version control

- Reproducibility applies to your own attempts (not just other people)

- Regardless of how you develop analytics, you'll be creating or editing software and documents.

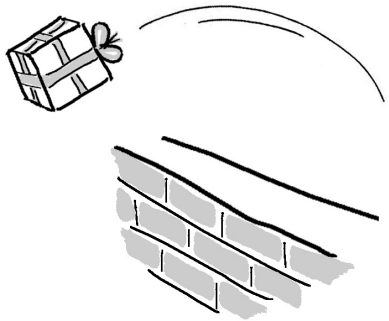
- [*lesson*] Regardless of how you implement best practices, avoid inventing solutions for which someone else already provided a path.

Suggested resource: https://try.github.io/

WHAT ARE WE NOT COVERING?

Not covered: product integration

- There's a complex network of dependencies (i.e. software engineers, managers) of which data science is one component.
- Downstream consumers of your output are likely to be software developers who use containers and support users.
- This class is focused on the data science; not with integration.



See http://dev2ops.org/2010/02/what-is-devops/

Not covered: security

This course is just an introduction course to prepare you for the remaining 8 classes.

Security is not covered at all.

SOFT SKILLS

DS is more than Math and Software

Human interaction in data science

- Discovering stakeholders
- Negotiating with data owners
- Customer engagement

https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

Iterating with customers

- As a data scientist, you'll often be working for someone other than yourself.
- Expect under-specified requirements from customers. Iterate.
- Provide incomplete solutions rather than waiting until the product is perfect.

https://en.wikipedia.org/wiki/Minimum_viable_product

When to persist,
When to change course,
When to seek help

Try attacking the challenge for 30 minutes

Then seek help or do something else for a while

https://en.wikipedia.org/wiki/Pomodoro_Technique

Pro-tip when seeking help

How to ask well-formed questions:

https://stackoverflow.com/help/how-to-ask

[Intentional sidetrack to StackOverflow.]

Ask technical questions:

- *Poor*: "I don't understand Python dictionaries" (--> online tutorials)
- *Better*: "When is it appropriate to use a key-value pair?"

- *Poor*: If I submitted this assignment as is, what score would I get?
- *Better*: I am planning to submit the attached assignment, but currently there's an error in the third cell. I've searched online but don't find any references to the error message. Can you provide guidance?

Emotions in Data Science

- As a data scientist, most of your time will be spent in a desert of uncertainty, frustration, and doubt.

- There will be rare short-lived interspersed spikes of excitement and happiness due to events like getting a new dataset, creating a new analytic, getting a new result, or being thanked by a stakeholder.

This experience is normal and does not go away.

See also the psychology of slot machines

Reading Suggestions

1. 50 years of data science
2. A Very Short History Of Data Science

Action: Read, write, tell

News and blogs

https://www.kdnuggets.com/
https://news.ycombinator.com/
https://hackernoon.com/
https://www.reddit.com/r/datascience/
https://dataelixir.com/newsletters/
https://insidebigdata.com/
https://ai.googleblog.com/

Some Online Resources

- Meetups
  - https://www.meetup.com/topics/data-science/
  - https://www.meetup.com/DataWorks/
  - https://www.meetup.com/Statistical-Seminars-DC/
- Others
  - Salaries: https://www.burtchworks.com/category/salary/
  - A weekly social data project in R: https://github.com/rfordatascience/tidytuesday
- Datasets to work with
  - https://datasetsearch.research.google.com/
  - https://datacatalog.worldbank.org/
  - https://opendata.maryland.gov/