UMBC

DATA 601 – Lecture 01
Introduction

UMBC Data 601 Fall 2022
Instructor: Felix Gonzalez

About me…

- Office Hours: Prior to class, Mondays 6:00-6:30pm
- Virtual Office by Request: https://umbc.webex.com/meet/fgonzale
- Resources: https://dil.umbc.edu/

About You

- Name
- Background
- Why Data Science?

Ground rules

- Schedule: 6:30pm – 7:25, Break, 7:30 – 8:25, Break, 8:30 – 9:10pm
- Also, it is acceptable to get up at any time and take a bathroom break
- I value being punctual (start of class, breaks, 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

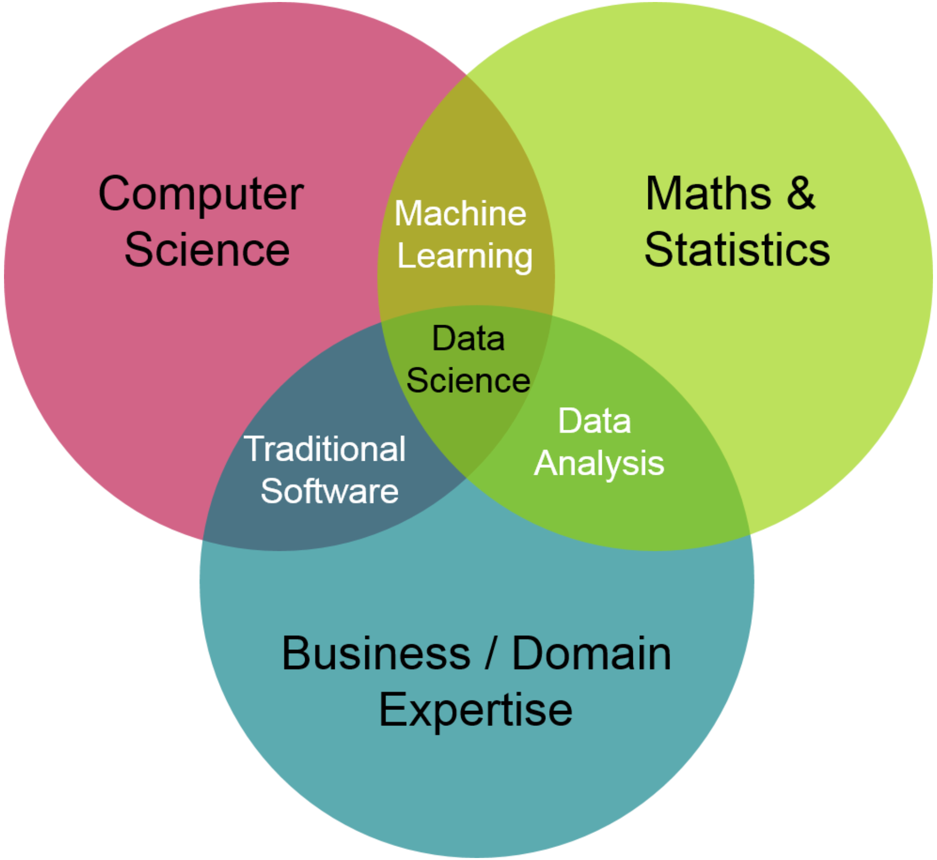
Schedule and Grading

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

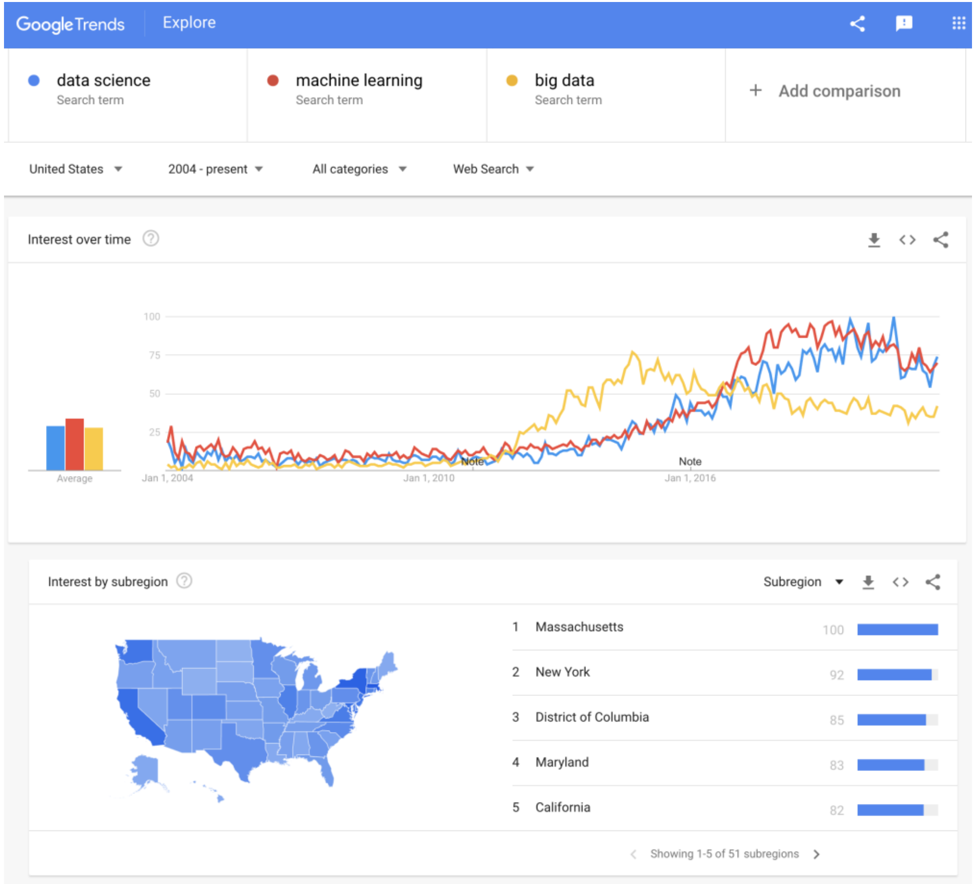
- Tentative Grading
  - Attendance (5%)
  - 4 Quizzes (9%)
  - 7 Homework (56%)
  - 2 Projects (Each project is 30%)

WHAT IS DATA SCIENCE?

What is Data Science?



Interest on Data Science



● data science — Search term

● machine learning — Search term

● big data — Search term

United States ▾ 2004 - present ▾ All categories ▾ Web Search ▾

1	Massachusetts	100
2	New York	92
3	District of Columbia	85
4	Maryland	83
5	California	82

‹ Showing 1-5 of 51 subregions ›

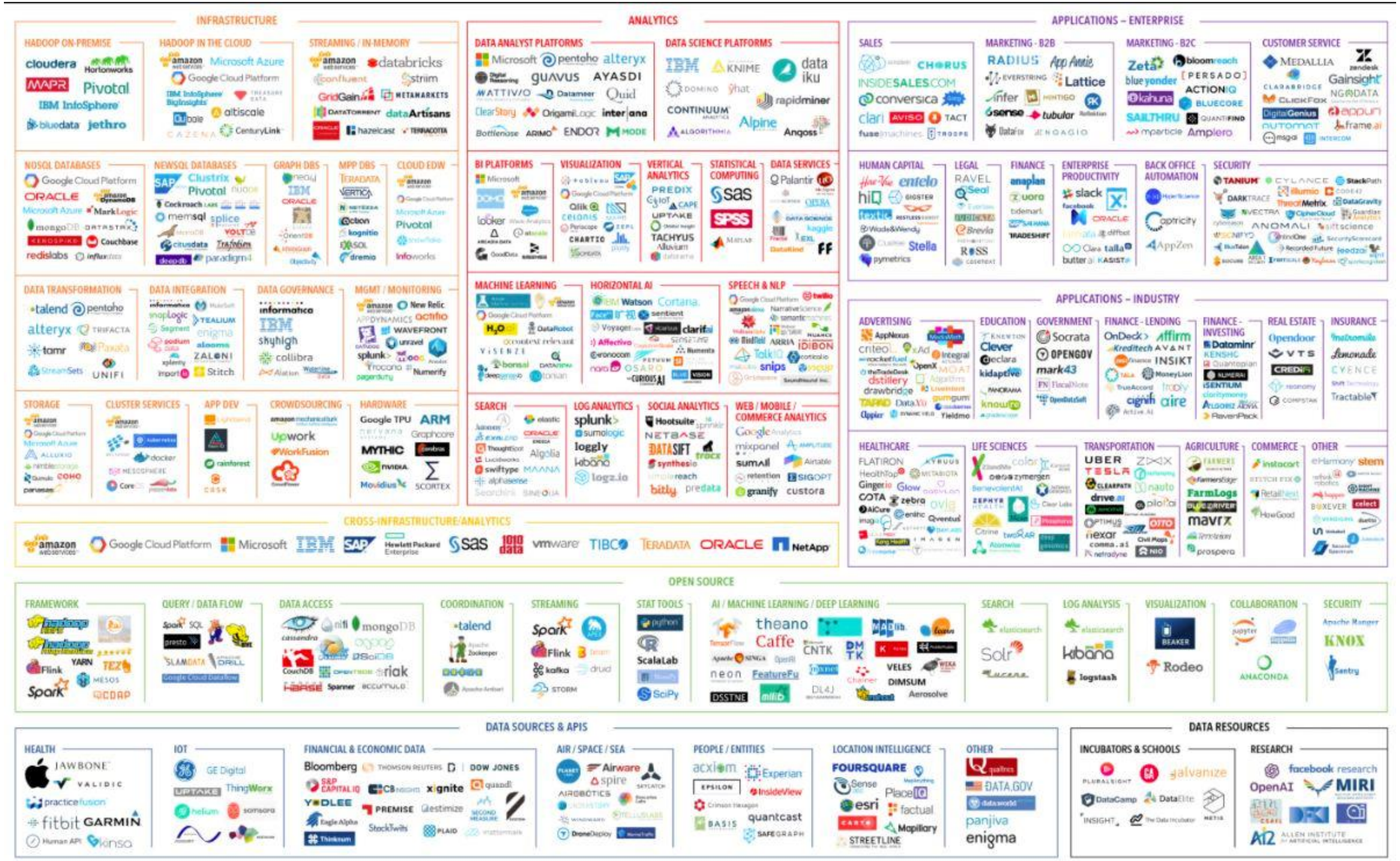
DS is an active field with lots of jargon

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

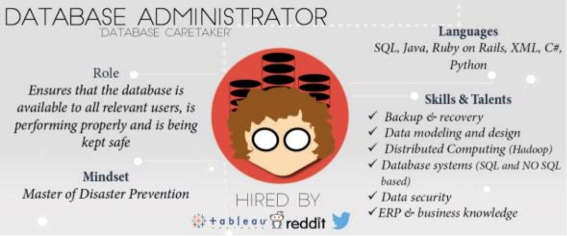
- Know enough to be conversant with peers
- Be curious about new topics
- Research concepts and labels before using them

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

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



Skills and experience matter more than title and labels



DATABASE ADMINISTRATOR
"DATABASE CARETAKER"

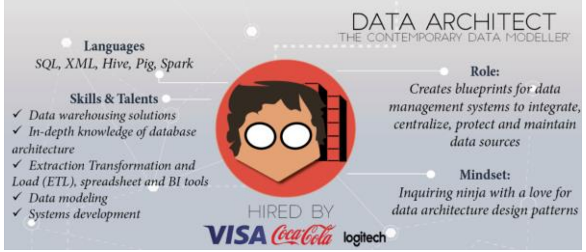
Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset
Master of Disaster Prevention

Languages
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents
- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

HIRED BY
tableau reddit



DATA ARCHITECT
"THE CONTEMPORARY DATA MODELLER"

Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents
- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns

HIRED BY
VISA Coca-Cola logitech

DATA ENGINEER
"SOFTWARE ENGINEERS BY TRADE"

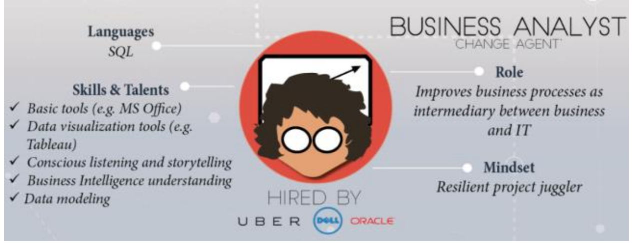
Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset
All-purpose everyman

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents
- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

HIRED BY
Spotify a



BUSINESS ANALYST
"CHANGE AGENT"

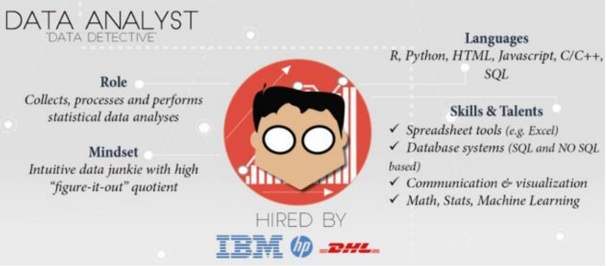
Languages
SQL

Skills & Talents
- ✓ Basic tools (e.g. MS Office)
- ✓ Data visualization tools (e.g. Tableau)
- ✓ Conscious listening and storytelling
- ✓ Business Intelligence understanding
- ✓ Data modeling

Role
Improves business processes as intermediary between business and IT

Mindset
Resilient project juggler

HIRED BY
UBER DELL ORACLE



DATA ANALYST
"DATA DETECTIVE"

Role
Collects, processes and performs statistical data analyses

Mindset
Intuitive data junkie with high "figure-it-out" quotient

Languages
R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents
- ✓ Spreadsheet tools (e.g. Excel)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Communication & visualization
- ✓ Math, Stats, Machine Learning

HIRED BY
IBM hp DHL

DATA SCIENTIST
"AS RARE AS UNICORNS"

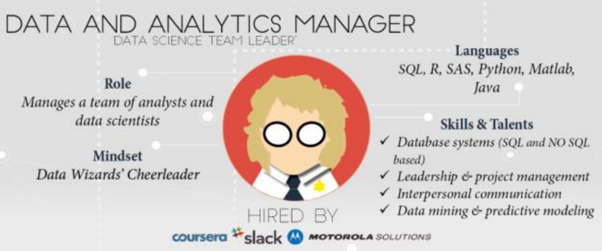
Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents
- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

HIRED BY
Google Microsoft Adobe



DATA AND ANALYTICS MANAGER
"DATA SCIENCE TEAM LEADER"

Role
Manages a team of analysts and data scientists

Mindset
Data Wizards' Cheerleader

Languages
SQL, R, SAS, Python, Matlab, Java

Skills & Talents
- ✓ Database systems (SQL and NO SQL based)
- ✓ Leadership & project management
- ✓ Interpersonal communication
- ✓ Data mining & predictive modeling

HIRED BY
coursera slack MOTOROLA SOLUTIONS

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

Historical progression: data grooming, data mining, data scientist

Why learn data science?

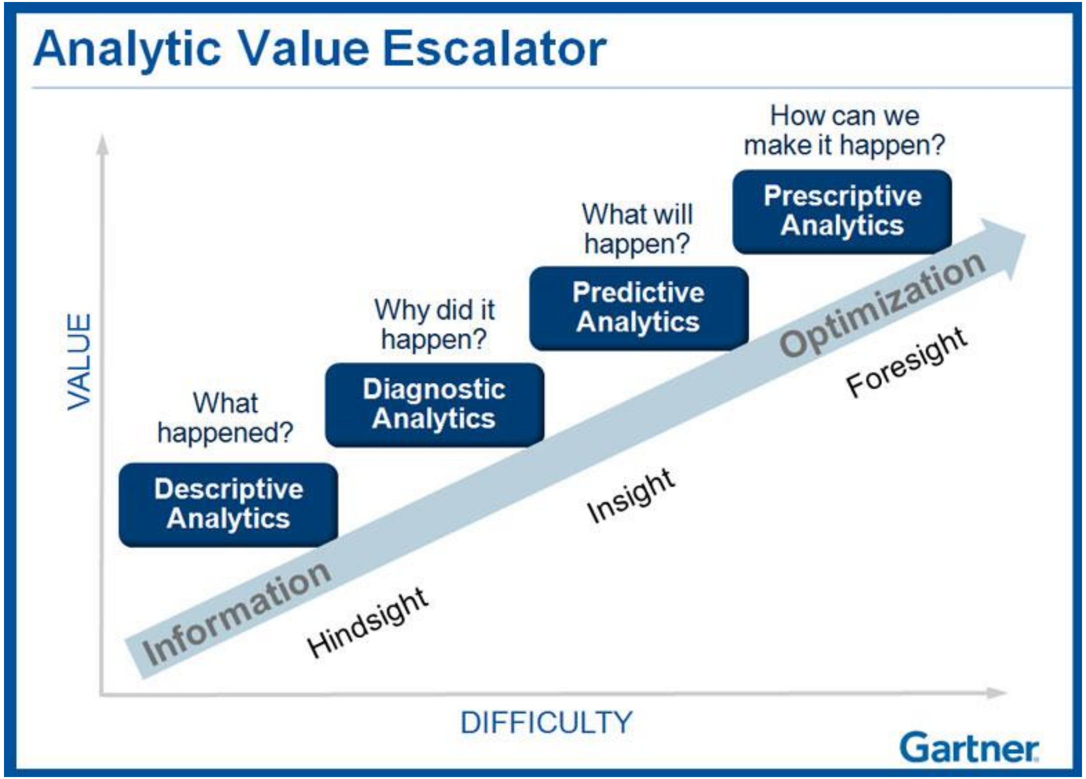
Explore: **identify patterns**

Predict: **make informed guesses**

Infer: **quantify what you know**

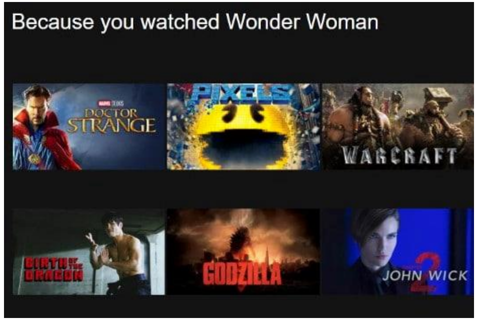
Motives:

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

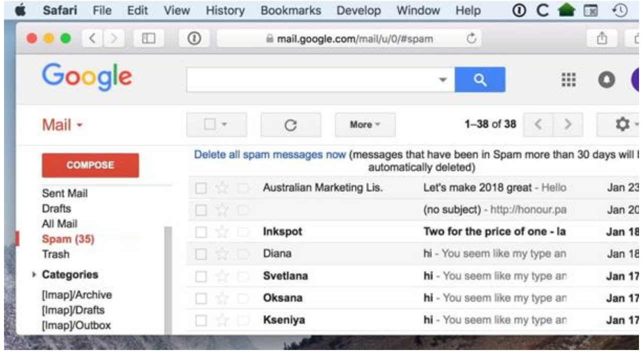


Large scale use cases with lots of data

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

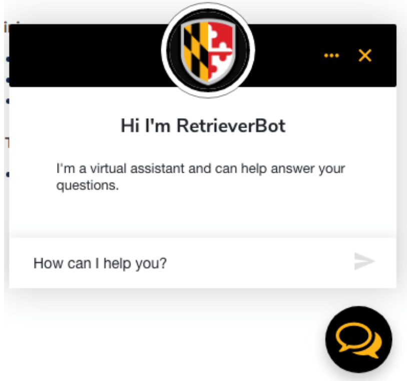


Because you watched Wonder Woman



Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

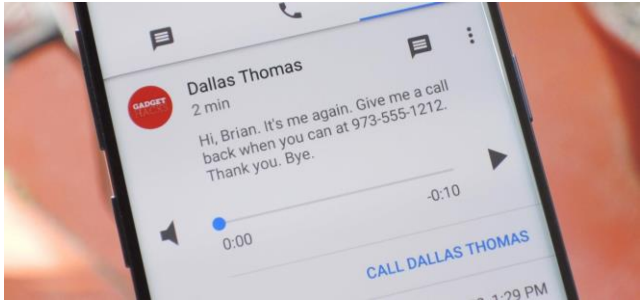
Australian Marketing Lis.	Let's make 2018 great - Hello	Jan 23
	(no subject) - http://honour.pa	Jan 20
Inkspot	Two for the price of one - la	Jan 18
Diana	hi - You seem like my type an	Jan 18
Svetlana	hi - You seem like my type an	Jan 17
Oksana	hi - You seem like my type an	Jan 17
Kseniya	hi - You seem like my type an	Jan 17



Hi I'm RetrieverBot

I'm a virtual assistant and can help answer your questions.

How can I help you?



Dallas Thomas

2 min

Hi, Brian. It's me again. Give me a call back when you can at 973-555-1212. Thank you. Bye.

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

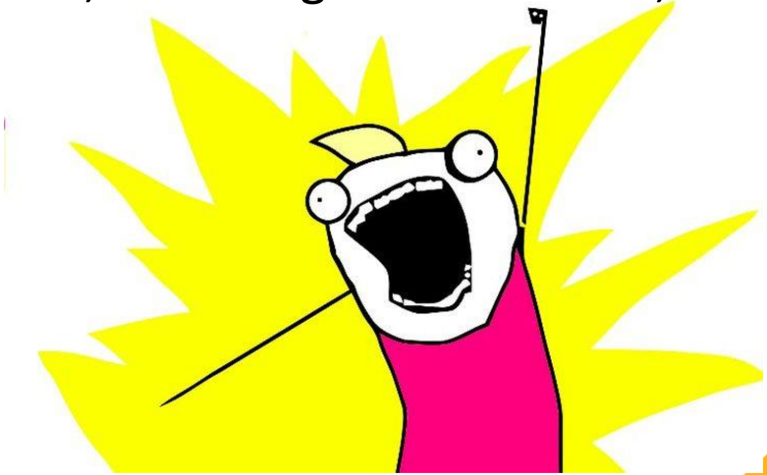
Python with Jupyter Notebooks

- Anaconda Data Science Platform (https://www.anaconda.com/)
- Google Colab (https://colab.research.google.com/)
- Google Drive (https://drive.google.com)
- GitHub (https://github.com/fgonzaleumbc/Data601_fall2022)
- Blackboard (https://my.umbc.edu/)

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 may not be the right tool
- For sufficiently complex analytics, Jupyter may not be 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

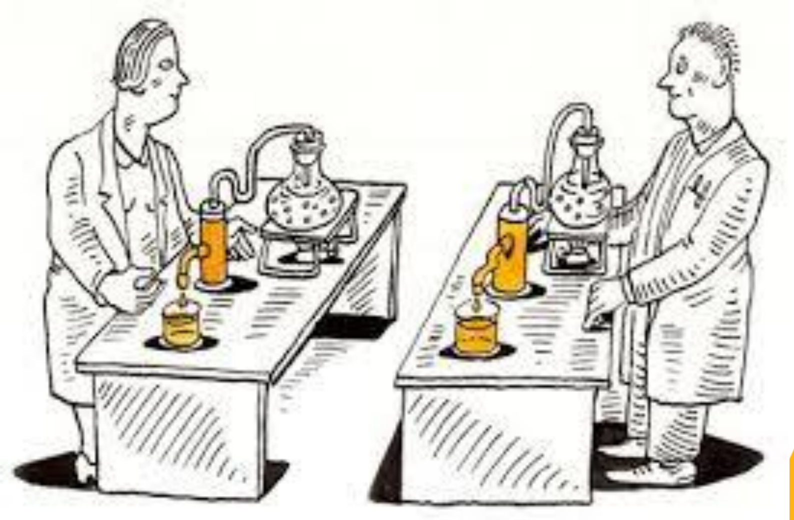
Digital archeology:
Suppose you are to diagnose why someone else's approach doesn't yield same results
Suppose they did their work 20 years ago

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

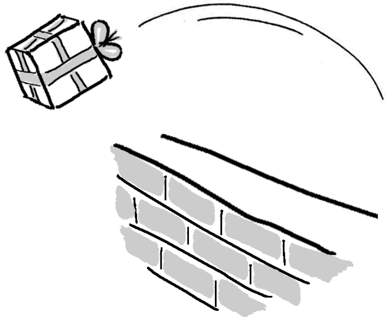
- Reproducibility applies to your own attempts (not just other people)
- Regardless of how you develop analytics, you'll be creating or editing software and documents.
- [*lesson*] Regardless of how you implement best practices, avoid inventing solutions for which someone else already provided a path.

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 classes.

Security is not covered at all.

UMBC

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/

Data Sets Online Resources

Title	Description	Link	Comment
Github	Used by many developers to share code and collaborate.	https://github.com/	Class code is hosted in this website.
U.S. Government’s Open Data	Data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.	https://data.gov/	Real world datasets that in some cases may be extremely large.
Kaggle	Host data science competitions, datasets, Jupyter notebooks, etc.	https://www.kaggle.com/	Good source for datasets and Jupyter Notebooks.
Google Research Data	Google Dataset Search Engine	https://datasetsearch.research.google.com/	Search engine for datasets. May send you to data in other websites in this table.
World Bank Data	Data collections by the World Bank	https://datacatalog.worldbank.org/	Also has collections from various sources.
Maryland Government Data	MD Open Data Website.	https://opendata.maryland.gov/	Real world datasets from Maryland’s State Government

Visualization Examples and Online Resources

Title	Description	Link	Comments
MatPlotLib Example Gallery	Python Plotting Library	https://matplotlib.org/stable/gallery/index.html	Library included in Anaconda.
Seaborn Example Gallery	Python Plotting Library	https://seaborn.pydata.org/examples/index.html	Library included in Anaconda.
Plotly Example Gallery	Python Plotting Library	https://plotly.com/python/	Library included in Anaconda.
D3.JS	Data Driven Documents Visualization Library	https://d3-graph-gallery.com/	Used to deploy dynamic website plots.
Five Thirty Eight	Interactive Dashboard Examples	https://projects.fivethirtyeight.com/	Excellent Website that provides great examples on the capabilities of dashboards and data analytic visualizations.