UNIVERSITY OF MARYLAND

INST447 -0101 Fall 2020 Lecture 4

Virtual

Instructor: Bill Farmer

TA: Jonathan Chen

Grader: Jeffrey Chen

September 22, 2020

This Week

Time: Tuesday virtual

- Admin
  - Syllabus Updates
- Readings
- Videos
  - Data basics review
  - Indexing
- Jupyter Examples

Time: Thursday Virtual w/ optional live session

- Live session
  - Jupyter Notebooks subjects from reading
- Videos
  - Concat & Append
  - Aggregations
- Lab & Assignment
- Projects - teams

If you are tired, stand up in the back of class.

Use of phone during class for non-class purposes is rude.

Admin

Admin

- Office Hours (need to schedule a time slot):
  - Monday 8-9 pm
  - Friday 8-10 am
  - Saturday 6-8 pm (changed from am to pm)
  - Sunday 6-7 pm (changed from 4-6 to 6-7)
  - By Appointment * Anytime
- Live class meetings - Thursdays 12:30-1:30
  - Class originally scheduled to start @ 12:30 so I figure this is a good time
  - We can add a couple of these at different times if/when needed
- Piazza vs. Canvas discussions

INST447 General Schedule

- Update 9/15/2020

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					8-10am	
		Noon-ish		12:30-1:30pm		
6-7pm						
	8-9 pm				Lab 11:59pm	6-8pm

Office Hours Live

Office Hours By appointment

Video Ready

Class Live Sessions

Lab Due

Syllabus Updates

- General syllabus schedule still applies
- Updated Canvas to be cleaner
- Contains fairly set schedule….some adjustments may be made

DataFrames / Series

Data Frames

- 2 dimensional arrays with labels
- columns – variables
- rows – each observation
- cell – data value

Data Frames

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

What is a variable in the data set?

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

What is an observation in the data set?

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

Data Frame Labels

- Column labels aka columns
- Row labels aka index

What is the index for Lhotse?

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

Create a DataFrame

- Read from file
  - import pandas as pd
  - df = pd.read_csv(‘path/filename’)
- Create using dictionary
  - dict = {‘col1’:[val1,val2],’col2’:[val3,val4]}
  - df = pd.DataFrame(dict)

Create a DataFrame with fruits and colors

- Fruits: apple, banana, orange
  - dict = {‘fruit’:[‘apple’,’banana’,’orange’], ‘color’:[‘red’,’yellow’,’orange’]}
  - fdf = pd.DataFrame(dict)

	fruit	color
0	apple	red
1	banana	yellow
2	orange	orange

Basic information about a DataFrame

- Prints first 5 rows of the data frame
  - df.head()
- Prints the number of rows and columns
  - df.shape (90, 41) 90 rows, 41 columns

Save a DataFrame

- Write data frame to csv
  - df.to_csv(“path/newfilename.csv”)

Series

- 1 dimensional array with labels
  - e.g. a column of a data frame
- best practice all values are same data type (e.g. int, float, string)
- You can think of a data frame as being made up of a series

	Apples
0	3
1	7
2	5
3	9

Series

+

	Oranges
0	12
1	6
2	1
3	13

Series

=

	Apples	Oranges
0	3	12
1	7	6
2	5	1
3	9	13

DataFrame

Indexing and Slicing

Indexing - rows

- Get first five rows
  - df[:5]
- Get rows 6-10
  - df[6:10]

Slicing - Columns

- Get a column
  - df[“Mountain”]
- Get multiple columns
  - df[[“Mountain”,”Range”]]

How can you get the year of first ascent?

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

How can you get the year of first ascent?

index labels

column names

	Mountain	Height (m)	Range	Coordinates	Parent mountain	First ascent	Ascents bef. 2004	Failed attempts bef. 2004
0	Mount Everest / Sagarmatha / Chomolungma	8848	Mahalangur Himalaya	27°59′17″N 86°55′31″E	NaN	1953	>>145	121.0
1	K2 / Qogir / Godwin Austen	8611	Baltoro Karakoram	35°52′53″N 76°30′48″E	Mount Everest	1954	45	44.0
2	Kangchenjunga	8586	Kangchenjunga Himalaya	27°42′12″N 88°08′51″E	Mount Everest	1955	38	24.0
3	Lhotse	8516	Mahalangur Himalaya	27°57′42″N 86°55′59″E	Mount Everest	1956	26	26.0
4	Makalu	8485	Mahalangur Himalaya	27°53′23″N 87°05′20″E	Mount Everest	1955	45	52.0
5	Cho Oyu	8188	Mahalangur Himalaya	28°05′39″N 86°39′39″E	Mount Everest	1954	79	28.0
6	Dhaulagiri I	8167	Dhaulagiri Himalaya	28°41′48″N 83°29′35″E	K2	1960	51	39.0
7	Manaslu	8163	Manaslu Himalaya	28°33′00″N 84°33′35″E	Cho Oyu	1956	49	45.0
8	Nanga Parbat	8126	Nanga Parbat Himalaya	35°14′14″N 74°35′21″E	Dhaulagiri	1953	52	67.0
9	Annapurna I	8091	Annapurna Himalaya	28°35′44″N 83°49′13″E	Cho Oyu	1950	36	47.0

data

> df[“First ascent”]

Summarizing Data

Variables

- Variable – refers to the property of an object or event that can take on different values
  - e.g. Mountain
  - e.g. Height
  - e.g. Coordinates

How many observations?

- Count observations in Python
  - Data Set (get total number of rows)
    - df.shape (90, 41) 90 rows and 41 columns
  - Variable (get count of non-missing values)
    - df.count()

Types of Variables (Scale of Measurement)

- Categorical
  - Nominal scale - a collection of labels (e.g. Mountain “Lhotse”, “K2”, “Makalu”)
  - Ordinal scale - an ordered rank of labels (e.g. Date of first ascent “1954-09-03”, “1954-10-22”, How satisfied are you with our products. (1-very satisfied, 2-somewhat satisfied…))
- Numeric
  - Interval scale - Equal intervals represent equal differences (e.g. Year of first ascent 1953, 1954, 1955)
  - Ratio scale - Equal intervals represent equal differences and has a true zero (e.g. Heigh 8848, 8611, 8586)

Descriptive Statistics

- **Descriptive Statistics** – Ways to meaningfully show or summarize large amounts of data with only a few values.
- **Measures of Central Tendency** - Typical values for a distribution.
  - mean, median, mode
- **Measures of Variability** - The degree to which individual data points are distributed around the mean.
  - range, standard deviation, frequency distributions

Descriptive Stats for Categorical Variables

- **Central Tendency**
  - mode
- **Variability**
  - frequency distribution
- **In Python**
  - df[“variablename”].value_counts()
  - df[“variablename”].mode()

Descriptive Stats for Numeric Variables

- **Central Tendency**
  - mean, median
- **Variability**
  - range, standard deviation
- **In Python**
  - df[“variablename”].describe()
  - df[“variablename”].mean()
  - df[“variablename”].median()
  - df[“variablename”].sd()

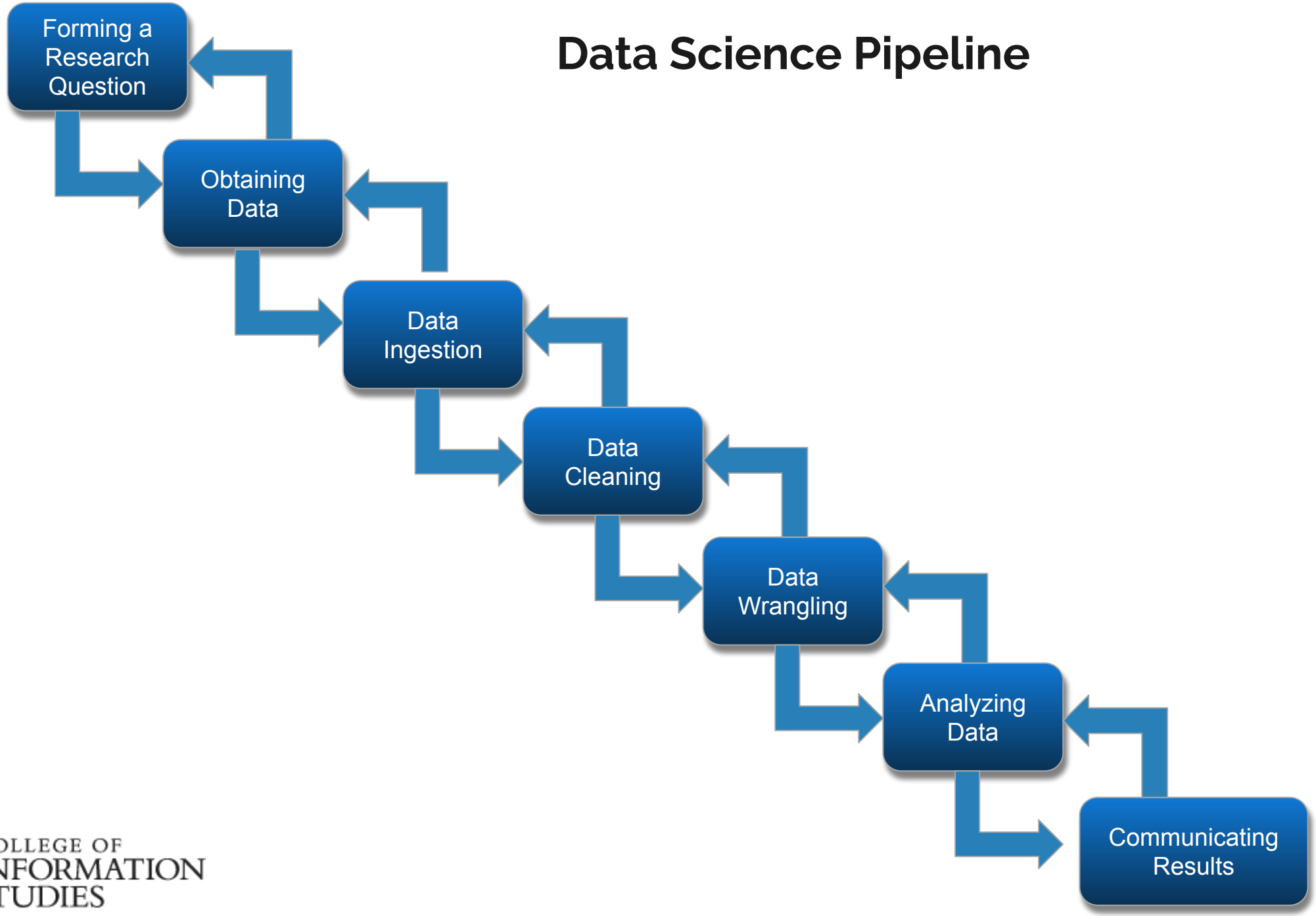
Calculations for Numeric Variables

- **Sum in Python**
  - df[“variablename”].sum()

Lab

Projects

Data Science Pipeline



COLLEGE OF INFORMATION STUDIES

Projects

- Teams of 2, 3, or 4
  - Class is small so I prefer sizes of 2 or 3
  - Project proposals due 2/28, so you have time
- API Keys
  - Twitter (see my submission process)
- Scraping
  - Reddit example (json)
  - Reddit group on data sets https://www.reddit.com/r/datasets/

Data Science Projects

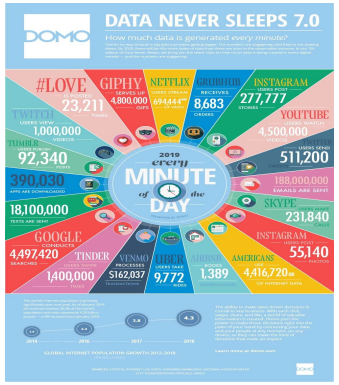
Data science - the ability to take large amounts of data in many different formats and be able to understand it, to process it, to extract value from it, to summarize it, to visualize it, and to communicate it to others.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems that perform tasks which ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users translate into tangible business value. -datarobot.com

FEC.gov

YAHOO! FINANCE

United States® Census Bureau



- Sentiment Analysis
- Customer Segmentation
- Recommending products
- Public Health Issues
- Manufacturing - predicting faults
- Financial Risk Analysis

Illustration from: https://www.socialmediatoday.com/news/what-happens-on-the-internet-every-minute-2019-version-infographic/558793/

COLLEGE OF INFORMATION STUDIES

Projects

salesforce

https://engineering.salesforce.com/ & TLS Fingerprints JA3 and JA3S

Berkeley SETI Research Center

FITS file handling
https://docs.astropy.org/en/stable/io/fits/index.html

astropy:docs

COLLEGE OF INFORMATION STUDIES

Next Week

Next Week

- Project Proposal
- More Python/Pandas

I appreciate your attention

Hope to see you on Thursday!

Reference Material
Install Software

4 Programming Assignments

- Work independently
- Deeper investigation into a data set and research question
- Turn in a well-structured and written report using Jupyter notebooks

Software Tools

- Python & Jupyter Notebook
  - Method 1
    - Python 3 (https://www.python.org/downloads)
      - Pandas Data Analysis Library (pandas)
      - Other modules (e.g. numpy, plotnine)
    - Jupyter Notebooks (aka ipython) (https://jupyter.org/install)
      - blend narrative text
      - code
      - output
      - visualizations
  - Method 2
    - Install Anaconda (includes both) (https://www.anaconda.com/distribution)
- Open Refine
  - http://openrefine.org/download.html
- Data sets
  - https://www.reddit.com/r/datasets/
  - https://opendata.dc.gov/
  - https://datasetsearch.research.google.com/
  - https://www.kaggle.com/datasets