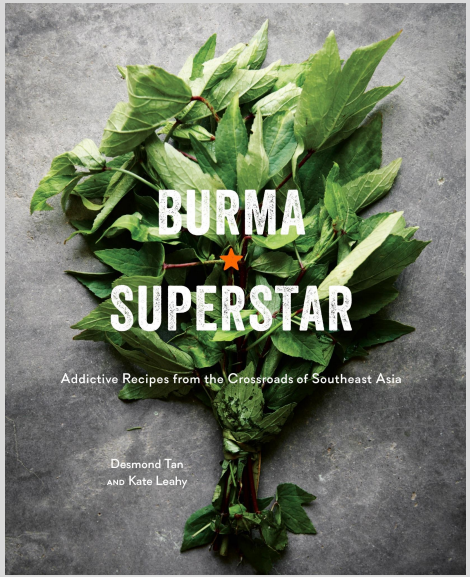
Day 8: Pandas and Data Science

July 19th 2023

Review – AS in import statements



where is recipe **yummy_sauce** ???

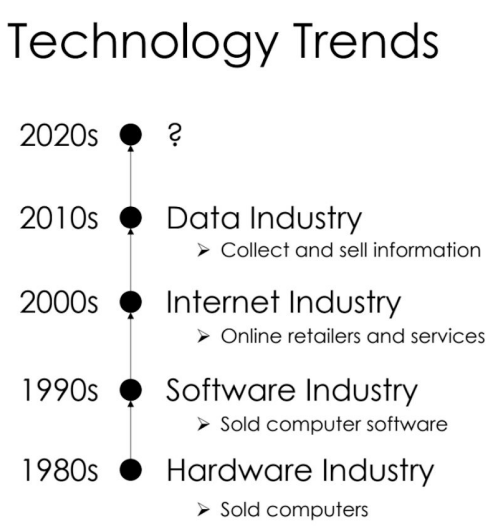
Review – AS in import statements

- using **as** in an import statement (for example: **import numpy as np**) is like giving a nickname to someone else's recipe book so that when you want to use a recipe, you don't have to call it the long book name each time
- importing a package is like bringing it into your kitchen when you are cooking – your code doesn't remember all the recipes every time, you need to tell it what books you are cooking from!
- when you import with . notations (for example: **import matplotlib.pyplot as plt**) you are taking in a specific part of a large set of functions
    - can imagine it is like bringing in a specific book in a book series!
- important to have specifics – what if you name something yummy_sauce and this is also defined in other books!

what is data science?

- at its heart, it is applied math and statistics!
- the amount of data we as a society have amassed is immense
  - what data have you generated today?
- having the computational skills enables you to work in any sector

amazing potential for good: healthcare sector
- single doctors / hospitals cannot sample enough data to see big trends
- researchers don't have access to the types of data hospitals have
- together, create network of data to find trends not visible to single doctors to create life saving policies



Academics

Master's Degree

Certificate in Data Fluency

Data Science Fellows

Certificate in Data Fluency

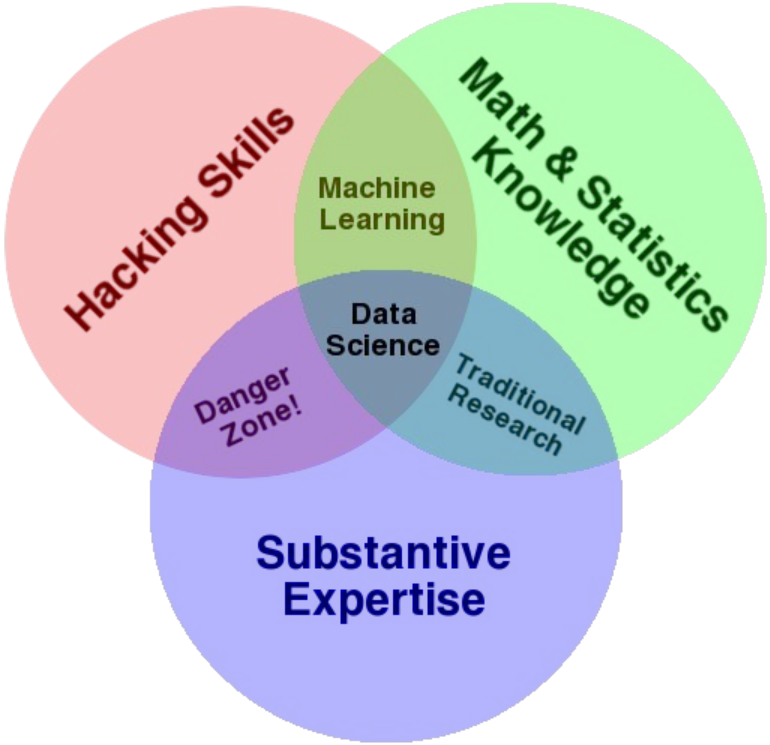
Description

The Certificate in Data Fluency is for undergraduate students who wish to gain fluency and facility with the tools of data analysis and its conceptual framework, but who are not pursuing a concentration in a data-intensive discipline. The program is designed to provide fundamental conceptual knowledge and technical skills to students with a range of intellectual backgrounds and concentrations, while emphasizing a critical liberal learning perspective. For more information about eligibility and requirements:

DATA FLUENCY CERTIFICATE INFO ▸

Purpose

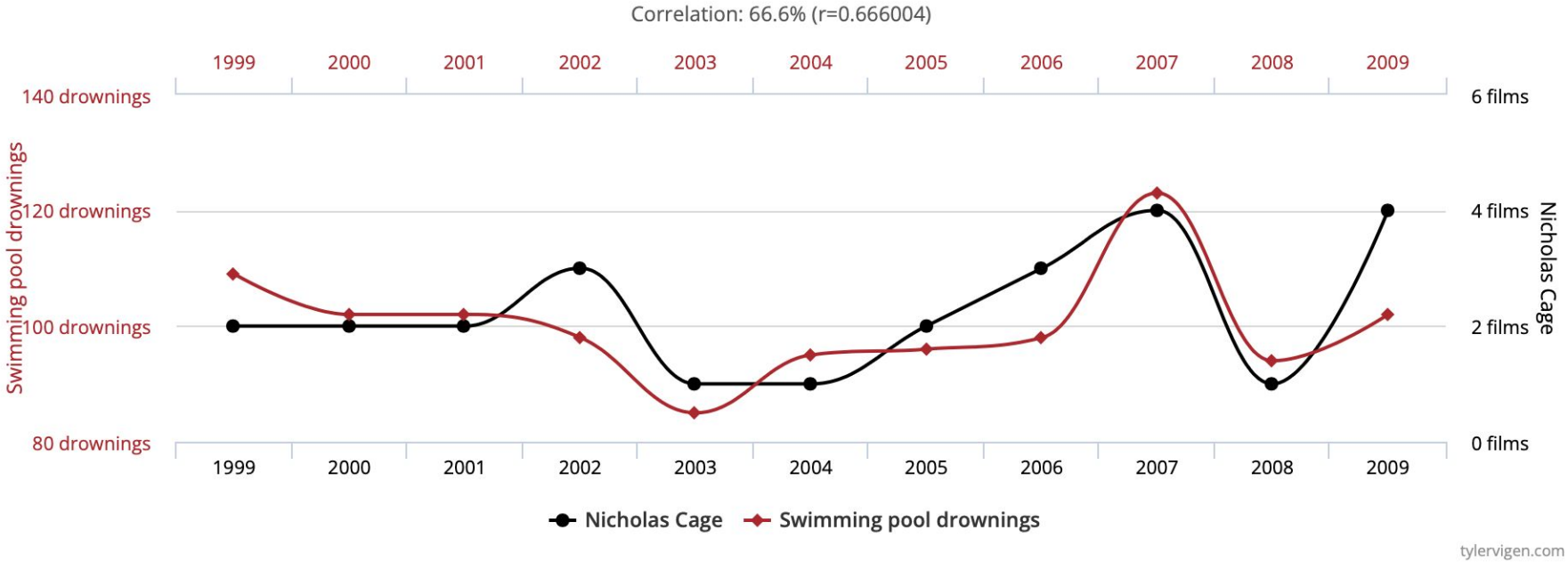
Data fluency implies a familiarity with data science and a basic competency working with data. Many disciplines now require an understanding of how data are collected, stored, analyzed, and visualized. The purpose of this certificate is to prepare


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

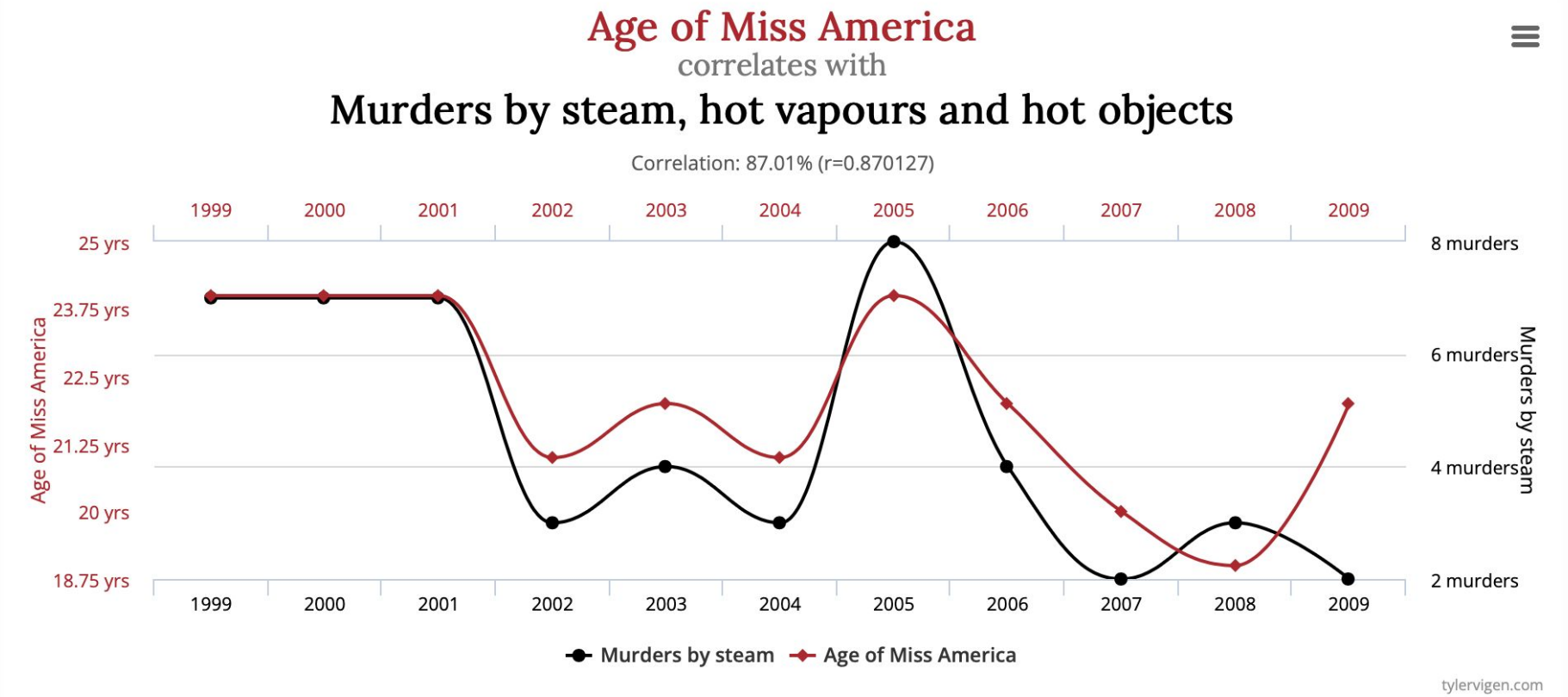
Number of people who drowned by falling into a pool

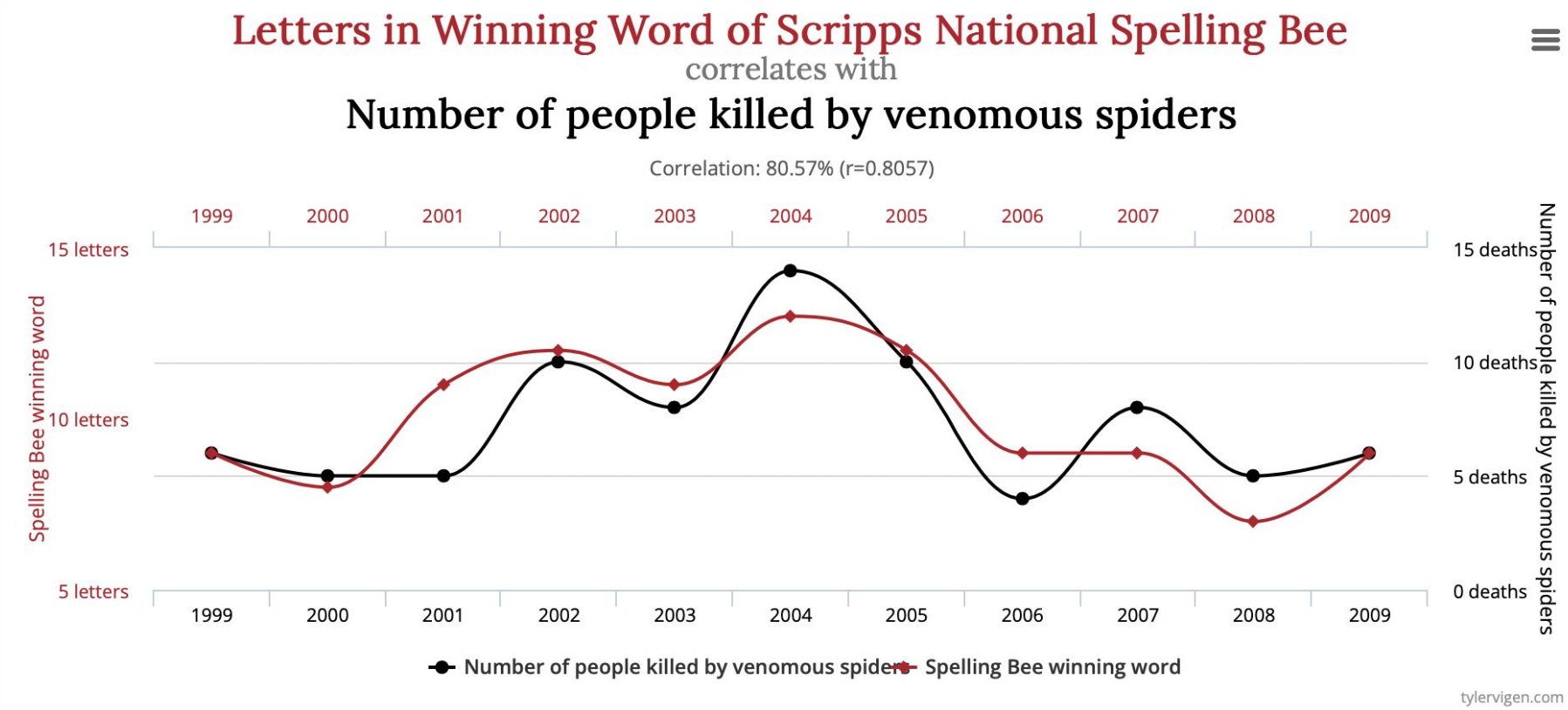
correlates with

Films Nicolas Cage appeared in



Data sources: Centers for Disease Control & Prevention and Internet Movie Database


Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects
Correlation: 87.01% (r=0.870127)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
25 yrs
23.75 yrs
22.5 yrs
21.25 yrs
20 yrs
18.75 yrs
Age of Miss America
8 murders
6 murders
4 murders
2 murders
Murders by steam
Murders by steam
Age of Miss America
tylervigen.com
Data sources: Wikipedia and Centers for Disease Control & Prevention

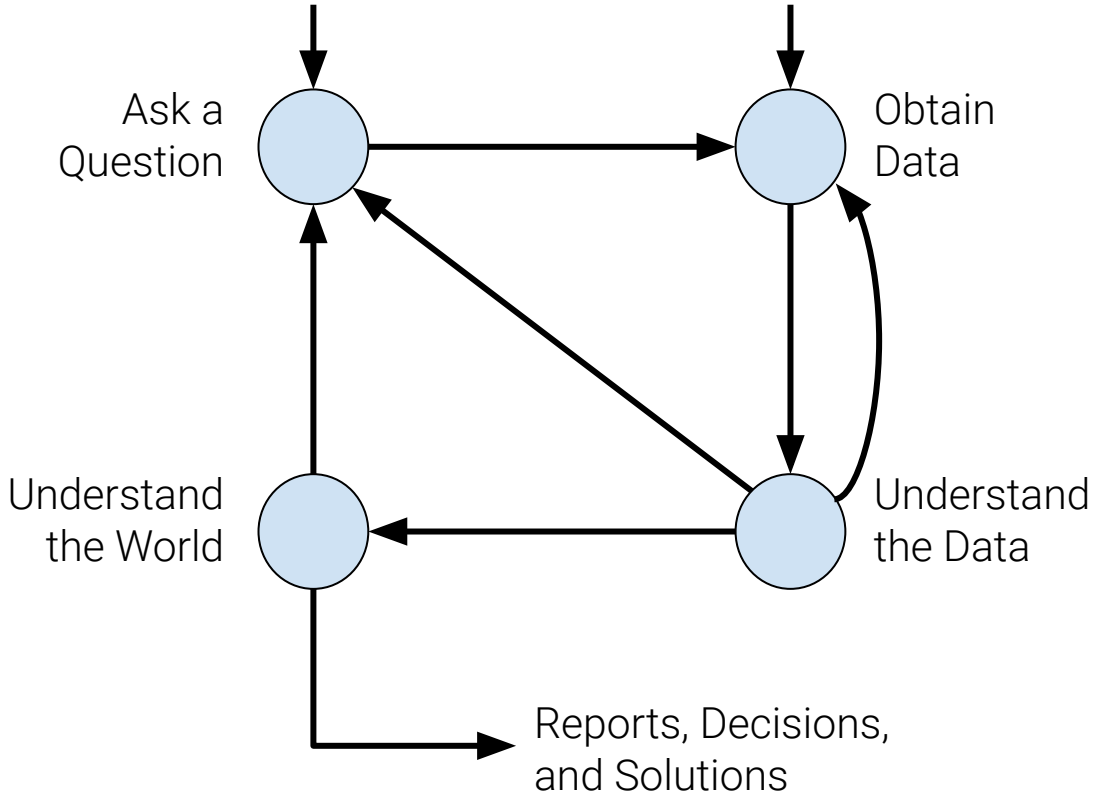


tylervigen.com

Data science lifecycle

The data science lifecycle is a **high-level description** of the data science workflow.

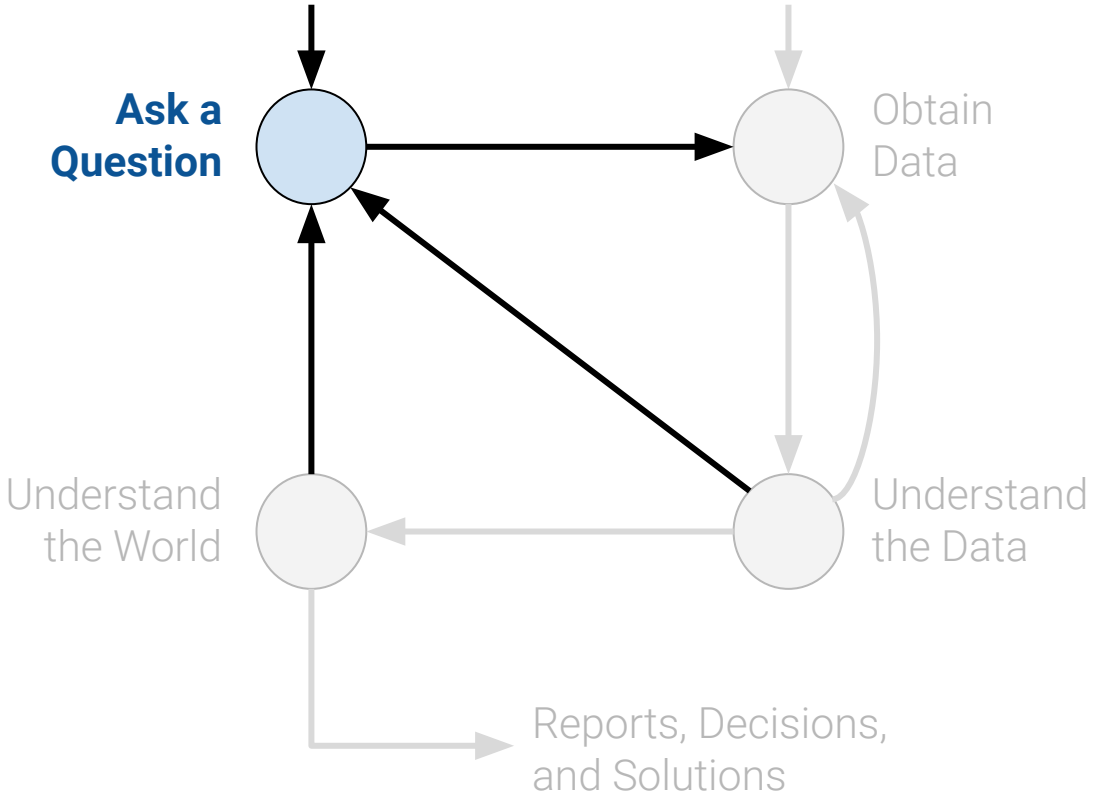
Note the two distinct entry points!



slides taken from Josh Hug @ uc berkeley

1. Question/Problem Formulation

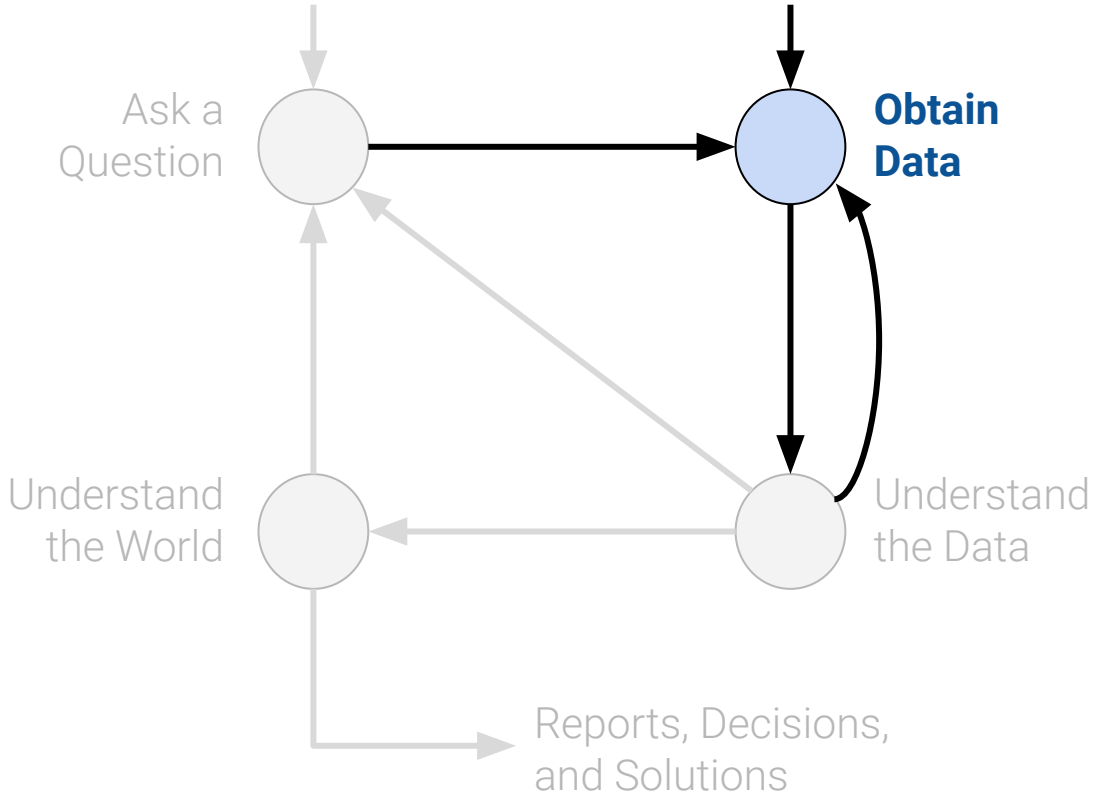
- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?



slides taken from Josh Hug @ uc berkeley

2. Data Acquisition and Cleaning

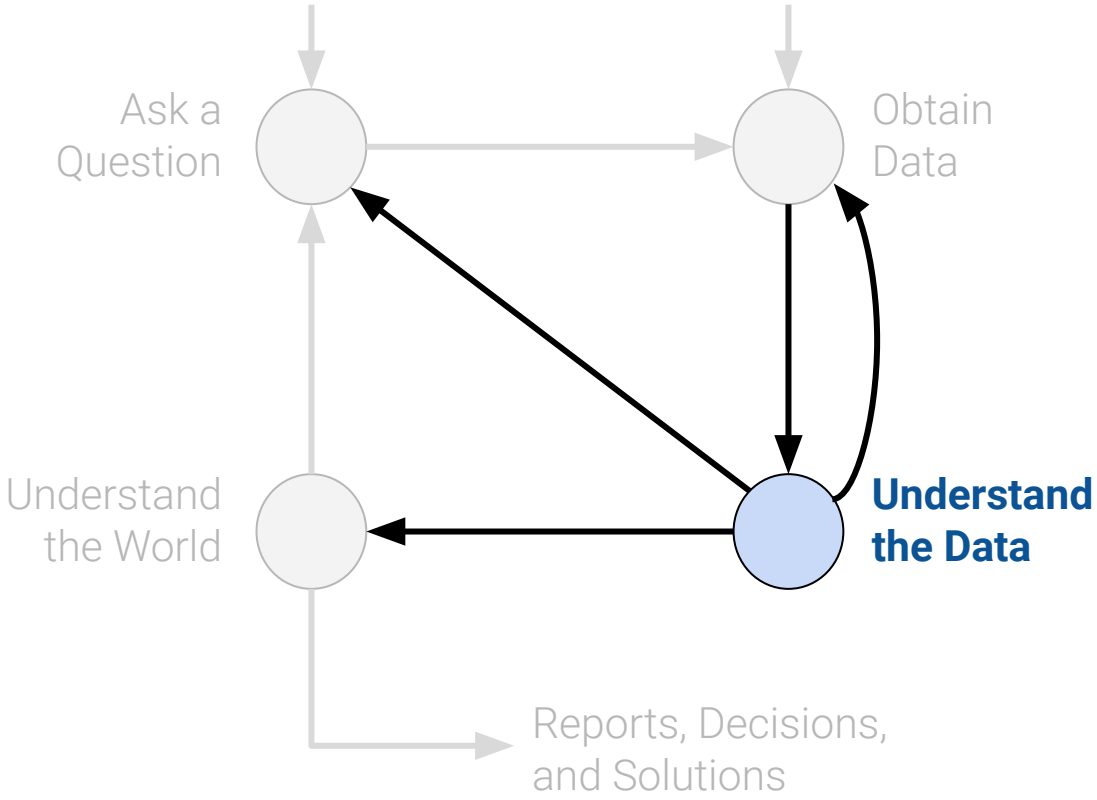
- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



slides taken from Josh Hug @ uc berkeley

3. Exploratory Data Analysis & Visualization

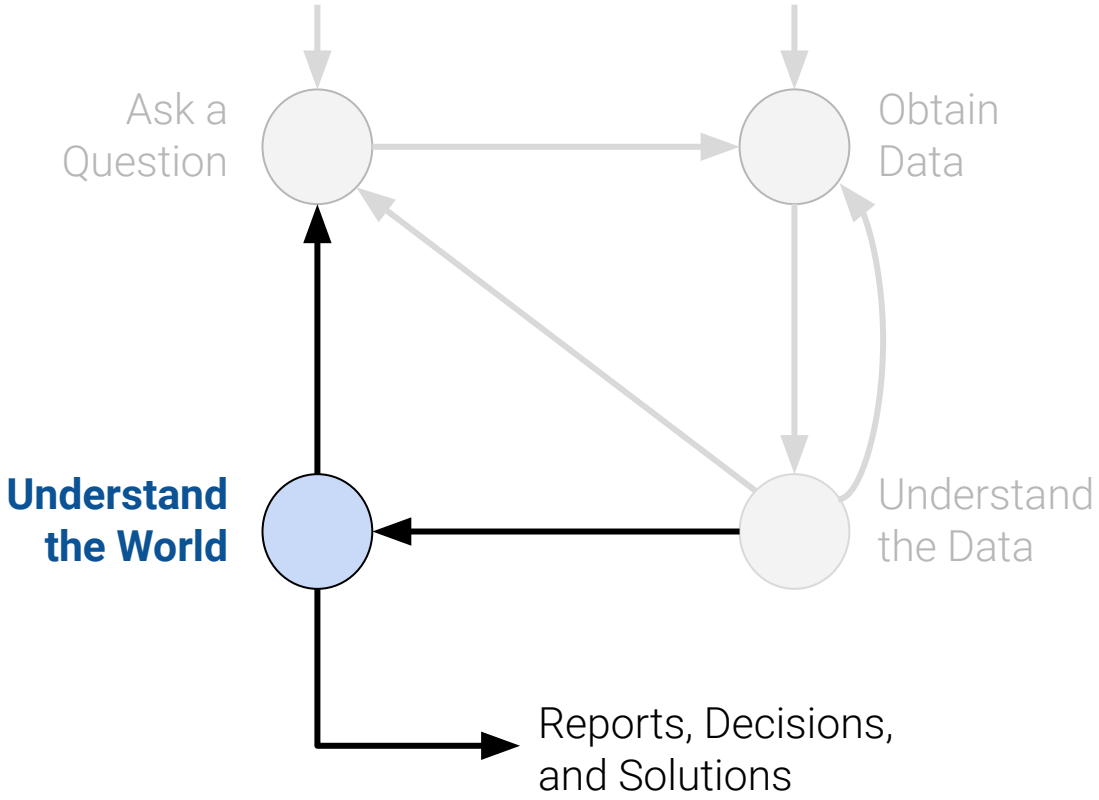
- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



slides taken from Josh Hug @ uc berkeley

4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



slides taken from Josh Hug @ uc berkeley

Today, we will do an overview of pandas

- pandas is a **data science** package to manage data you work with
- I use pandas every day in my research, and took an entire semester long class in college on how to use pandas
- pandas works with any type of data – organizes it into rows and columns so you can manage information!

built on

Example Dataset

animal_data.csv — Open with Visual Studio Code

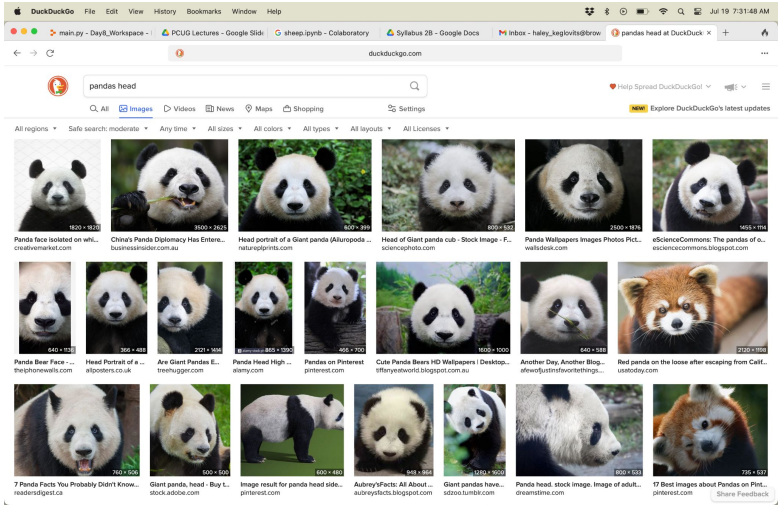
	id	intakedate	istransfer	sheltercode	identichipnumber	animalname	breedname	basecolour	speciesname	animalage	sexname	loca
0	15801	2009-11-28 00:00:00	0	C09115463	0A115D7358	Jadzia	Domestic Short Hair	Tortie	Cat	9 years 2 months.	Female	Ado
1	15932	2009-12-08 00:00:00	0	D09125594	0A11675477	Gonzo	German Shepherd Dog/Mix	Tan	Dog	9 years 1 month.	Male	Ado
2	28859	2012-08-10 00:00:00	0	D12082309	0A13253C7B	Maggie	Shep Mix/Siberian Husky	Various	Dog	6 years 8 months.	Female	Ado
3	30812	2013-01-11 00:00:00	0	C1301091	0A13403D4D	Pretty Girl	Domestic Short Hair	Dilute tortoiseshell	Cat	8 years 11 months.	Female	Stra
5	31469	2013-03-26 00:00:00	0	D1303720	981020007006095	Bonnie	Basenji/Mix	Brown and White	Dog	5 years 11 months.	Female	Stra
7	40705	2015-06-19 00:00:00	0	R15061738		Candy	American Sable/American	Brown	House Rabbit	4 years 6 months.	Female	Lobt
8	40706	2015-06-19 00:00:00	0	R15061739		Charlotte	American Sable/American	Brown	House Rabbit	4 years 6 months.	Female	Lobt
9	46238	2016-10-09 00:00:00	0	D16103207	981020019070093	Tennessee	Heeler/Mix	Brown and Black	Dog	3 years 2 months.	Male	Spe
12	46437	2016-10-26 00:00:00	0	C16103406	981020017650993	Nova	Domestic Long Hair	Black	Cat	12 years 3 months.	Female	Stra
14	46693	2016-11-21 00:00:00	0	D16113659	981020017897009	Baby	Pitbull	White	Dog	4 years 1 month.	Male	Ado
15	47120	2017-01-11 00:00:00	0	C17011137	981020021072899	Vivian	Himalayan	Seal	Cat	5 years 0 months.	Female	Fost
17	47146	2017-01-17 00:00:00	0	D17011163	981020021063793	Dinah	Labrador Retriever/Affenpinscher	Black	Dog	3 years 10 months.	Female	Fost
19	47156	2017-01-17 00:00:00	0	R17011173		Maxwell	American	Black	House Rabbit	3 years 10 months.	Male	Lobt
22	47157	2017-01-17 00:00:00	0	R17011174		Hammer	American	Black	House Rabbit	3 years 10 months.	Male	Lobt
24	47198	2017-01-22 00:00:00	0	C17011213	981020021020775	Mama Mia	Domestic Short Hair	Grey and White	Cat	3 years 3 months.	Female	Ado
26	47239	2017-01-27 00:00:00	0	D17011254	981020021070452	Maria		Brown and Black	Dog	1 year 10 months.	Female	Fost
27	47248	2017-01-27 00:00:00	0	C17011263	981020021068759	Mr. T	DMH/DSH	Black	Cat	9 years 11 months.	Male	Fost
29	47277	2017-01-31 00:00:00	0	C17011292	981020021046006	Coraline	Domestic Short Hair	Brown and Black	Cat	3 years 10 months.	Female	Ado
30	47290	2017-02-01 00:00:00	0	C17021304	981020021057911	Montana	Domestic Short Hair	Torbie	Cat	3 years 4 months.	Female	Ado
32	47319	2017-02-05 00:00:00	0	D17021329		Louise Vada	Pitbull	Black	Dog	3 years 2 months.	Female	Stra
33	47360	2017-02-11 00:00:00	0	C17021370	981020021060759	Calli	DSH/Unknown	Various	Cat	16 years 3 months.	Female	Offic
36	47395	2017-02-13 00:00:00	0	C17021405	981020021035687	Gooseberry	Domestic Short Hair	Tortie	Cat	13 years 9 months.	Female	Ado
37	47402	2017-02-14 00:00:00	0	D17021412	981020021039948	Precious	Shep Mix/Mix	Ruddy	Dog	4 years 9 months.	Female	Ado
40	47414	2017-02-16 00:00:00	0	D17021424	981020021060979	Beemo	Pitbull/Mix	Blue	Dog	6 years 10 months.	Male	Stra
43	47429	2017-02-18 00:00:00	0	C17021438	981020021046343	Ellen	British Shorthair/Unknown	Grey	Cat	11 years 9 months.	Female	Offic
45	47432	2017-02-18 00:00:00	0	R17021441		Fitzgerald	American	Black	House Rabbit	3 years 9 months.	Male	Lobt
47	47502	2017-02-27 00:00:00	0	D17021511	981020015101070	Zoey	Pitbull/Mix	Grey and White	Dog	2 years 8 months.	Female	Stra
49	47510	2017-02-28 00:00:00	0	C17021519	981020021059390	Maureen	Domestic Short Hair	Brown and Black	Cat	2 years 8 months.	Female	Ado
50	47511	2017-02-28 00:00:00	0	C17021520	981020021064555	Janet	DSH/Unknown	Grey	Cat	4 years 2 months.	Female	Ado
51	47520	2017-03-02 00:00:00	0	C17031529	981020021039765	Ruce	DSH/Unknown	Black and White	Cat	3 years 10 months.	Male	Ado
52	47521	2017-03-02 00:00:00	0	C17031530	981020021063627	Ramsay	DSH/Unknown	Orange and White	Cat	3 years 10 months.	Male	Ado
53	47538	2017-03-04 00:00:00	0	D17031547	981020021050526	Gus	Labrador Retriever/Coonhound	Chocolate	Dog	2 years 10 months.	Male	Fost
54	47539	2017-03-04 00:00:00	0	C17031548	981020021038500	Zack Attack	DSH/Unknown	Grey and White	Cat	4 years 9 months.	Male	Ado
55	47540	2017-03-04 00:00:00	0	C17031549	981020021071498	Buggs	DSH/Unknown	Black	Cat	1 year 11 months.	Male	Ado
56	47558	2017-03-06 00:00:00	0	D17031567	981020021074652	Clyde	Golden Retriever/Poodle, Standard	Golden	Dog	6 years 2 months.	Male	Ado

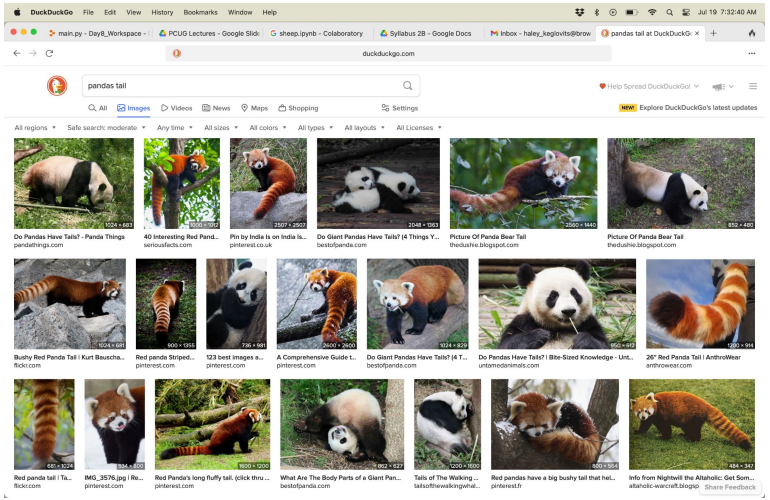
replit note – make your console large today!

Helpful functions we will learn

- view column
- get data from column
- describe
- head and tail
- combine columns
- creating new columns
- math functions:
    - mean, max, min, gt
- unique
- apply
- sample
- loc : pick out only some data!

all of these functions are applied directly to dataframes: use dataframe.function





Explore Dataset

- viewing info:
  - use [bracket] to get data of a specific column from dataset
    - think of the same way as getting the data from a list or vector!
  - need to be specific that we want to see a column information – can't just say 0 need to give the name of the column as a string
  - to turn this into numbers
- what columns are there? use .columns to get list of names
- getting data from column: use .values
- head and tail to see top and bottom of dataframe

Combining Columns

- you can use the same type of combinations from when we learned about string and int/float combinations (using + and - and * etc)
- if you have two columns of strings and combine them with +, you will get a new column which combines each row’s string values in the two columns!

Creating new columns

to make a new column, we will use the same [bracket] notation we do to get out columns that already exist

df[‘newcolumn’] = df[‘oldcolumn’] * 2

don’t forget to use strings with quotes as your column names!

Math functions

- mean
- max
- min
- gt (greater than)

pandas has functions which can be applied the same way we did in numpy! we can apply them to specific **columns** of the dataframe – important to specify because it doesn't make sense to take the maximum of a whole dataset

Unique

unique finds the single values in a specific column of a dataframe!

imagine you have a dataset of multiple animals and you want to know all of the types which are represented. unique is helpful for this!

unique is more useful when you have a column with **repeated** values – if you expect everything to be different, it is not very interesting

this is why it is usually best on string information and not numbers, because there tends to be lots of different numbers whereas there might be fewer strings. but this isn't always the case

Apply

apply takes a **function** and calls it on each row's value in a given column.

df['col'].apply(func)

will apply the function **func** to each value in the column named col!

you can put these results into a new column because it will be the same length of the dataframe

this function is **not** for summary statistics like mean / min / max – this is when you want one value per row of the dataframe

Sample

if you want a **random** subset of your dataframe's rows, you can use **sample** instead of head or tail to pull out a selection of the dataset!

if you don't pass a value inside the function, it will assume you want one row (unlike head and tail which assumed you wanted 5)

Loc

loc is how we will access only specific rows of our dataset! .loc[] expects a logical statement in the brackets which will evaluate each row and decide if it should be included in the new dataframe it returns.

to use this new dataframe, set it to a new variable and use the function .copy() so that you don’t accidentally edit the old data

new_df = df.loc[df.column_name == True].copy()

Drop

drop will remove a column or list of columns from your dataframe for cleaning it up. however, there are two things to note about drop which make it act unexpectedly:

- first, you must pass in the command axis=1 to tell it you want to drop columns. otherwise, it will think you want to remove rows of your dataset!
- second, drop does not work **inplace** meaning it will return a copy of your dataset rather than working directly on the one you have in your variable. to keep these changes (the drops) you can:
  - set your dataframe equal to the line which has the drop (ie, df = df.drop([‘col1’,’col2’],axis=1)
  - pass the command inplace=True after axis=1 (with a comma in between)
  - don’t do both! will not work

Homework Problem 1: Finish Class Exercise

- make a plot which has the age of cats as blue + signs and the age of dogs as red circles
- finish code from class!

Homework Problem 2: Sheep Race

You have a dataset with 5 sheep who you timed at the racetrack, and you want to know if they qualify for a upcoming race. However, they are sheep and a little silly, so they ran around the racetrack a variable number of times. You need to calculate their average time per lap and decide if they qualify for the group race, meaning they run less than or equal to 4 minutes per lap. Runtimes column is already in minutes. Follow the comments in the code to:

- make a column laptimes with the sheep's times
- add another column (with your choice of name) which contains the string 'racer' if they qualify or 'next year' if they don't
- finally, print the names of the sheep who can race with their times all as one string!

There are multiple ways to solve this problem. Think through the logic before doing the code, and go back to the class problems to see what you want to do!

Homework Problem 3: Identify Interesting Plot

tomorrow in class we will be working on our final project, which will be to create an **effective visualization** of one **data relationship** in the set

- an example of a data relationship from today: how does breed of dog correlate with age?
- how does species of animal correlate with age?

Homework Problem 4: Renaming Packages

Your friend really loves pizza, and wants to import the numpy package under the **alias** (name) pizza. Change the code in problem 4 starter so that you import how they want, and so that all of the following code will work.

Optional Homework Challenge: String Apply in Pandas

Problem 5: String Apply in Pandas
Optional challenge: recreate the age in days column in the animal dataset! I am uploading 'animal_data_raw.csv' which contains a column 'animalage' that has the animal's age in words -- how can you convert this to a number?

Hints:

- - use `string.split(' ')` to cut a string every time there is a space
- - make a dictionary which converts something like 'month' to 30 which is number of days for the word
- - take a look at the values in that column -- what patterns do you see? how could you use these patterns? (example: all end in a .! you can get rid of it)