DS2000
9/17 - tues.

Admin

- Hw2 due Fri 9pm
  - ↳ styleguide
  - ↳ check gradescope!!

Agenda

1. Data Sources
2. File processing
3. Python

1. Data Sources

Any DS program:

- gather data
- computations → arith. operators
- communication → print()
  plt.plot()

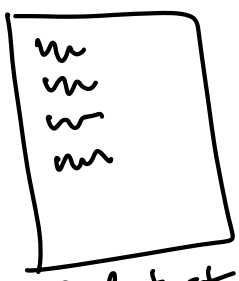
ex) plt.plot(—, —, —)
x, y, "o" ●
"x" x
"s" ■

Gather data
↳ how can we get data?:
input() → from the user
↳ other sources of data

- a file (text file, excel file)
- scraping a website
- online function (API)
- generate random #s

responsibly using a file:

- source: data.boston.gov
  [accuracy: ok]
- type: salary of city employees
- consent/knowledge: ok
- responsible use:
  only what we want to learn
  don't amplify personal data



one number per line
(one person's salary)
2023

sal.txt (text file)

2. File Processing in Python

- gather data: from user, get the filename
  then, open + read file
  - read one line at a time
  - top to bottom
  - save each number in a variable

- computations: 5 Boston salaries (2023)
  ↳ what might we want to calculate?
  - range (min vs. max)
  - mean (avg)
  - median
  - std deviation
  - under/over threshold
- communication
  ↳ plot the data (row # (x) vs. salary (y))

3. Python

↳ save the data file
(same PyCharm project as code)

In Python:
↳ open the file

```
with open(filename, "r") as infile:
```

- Python needs!
- string (from user)
- reading
- file as a variable

↳ read each line one at a time

```
szl = float(infile.readline())
```

(similar to input())

↓

string, unless we say otherwise