CS 133 - Introduction to Computational and Data Science

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University
Spring 2017

Introduction to Python II

- Quiz 2
  - Average score drops from 17.21 to 15.6
  - Max is 19.5
  - Go through the quiz

- In-class exercise is due today
  - Questions about this exercise?
  - Go through the exercise together

Introduction to Python II

- Reading (Data Science from Scratch):
  - Read Chapter 3: Visualizing Data
  - Read Chapter 4: Linear Algebra

Visualize data

- In the previous class, you have learned processing files, generating random numbers.
- Today we are going to learn some new fancy features, drawing graphics

Why visualizing data?

1. To explore data
2. To communicate data

Both are equally important!!!!

matplotlib

“matplotlib is a python 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms”

http://matplotlib.org/

You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc, with just a few lines of code

How to install matplotlib

Method 1: Use the official website (http://matplotlib.org) and follow the instructions………

Method 2: Install anaconda!!!

https://www.continuum.io/downloads

Get python 2.7

Follow prompts

Enjoy

For Linux: sudo apt-get install python-matplotlib

For Mac: curl -O https://bootstrap.pypa.io/get-pip.py

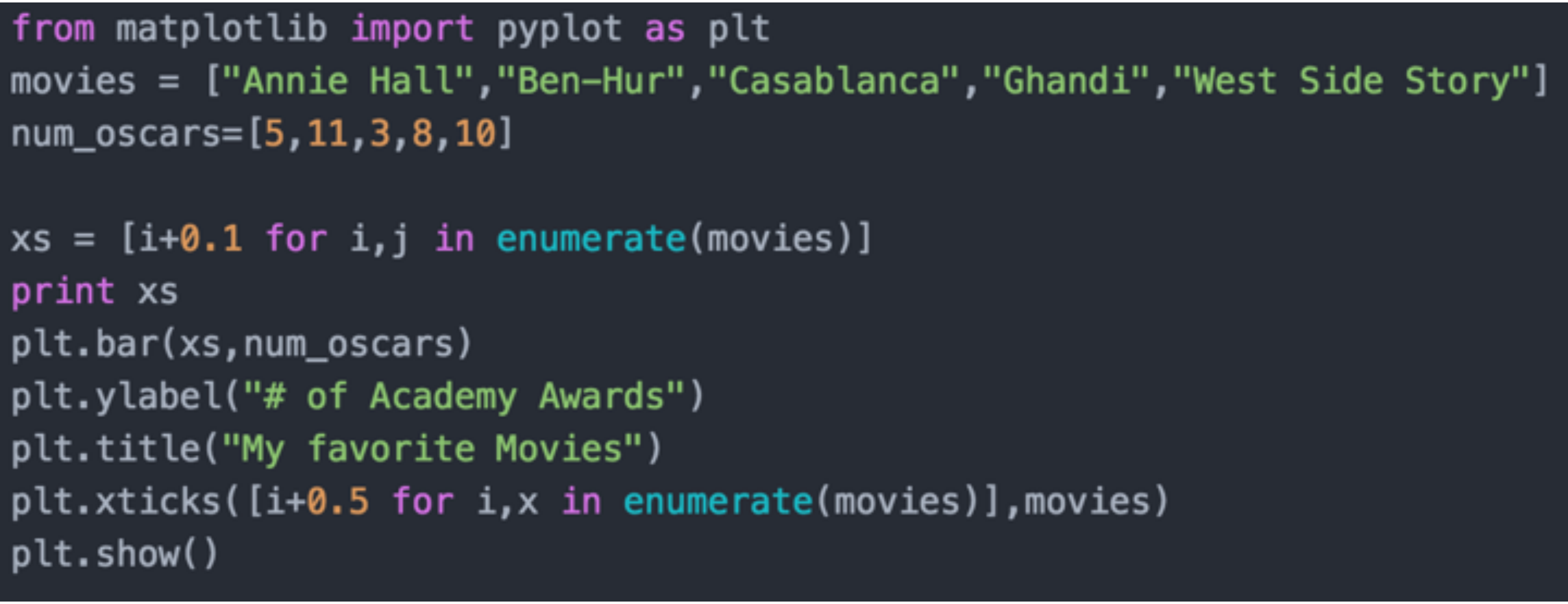
python get-pip.py

pip install matplotlib

Examples using matplotlib

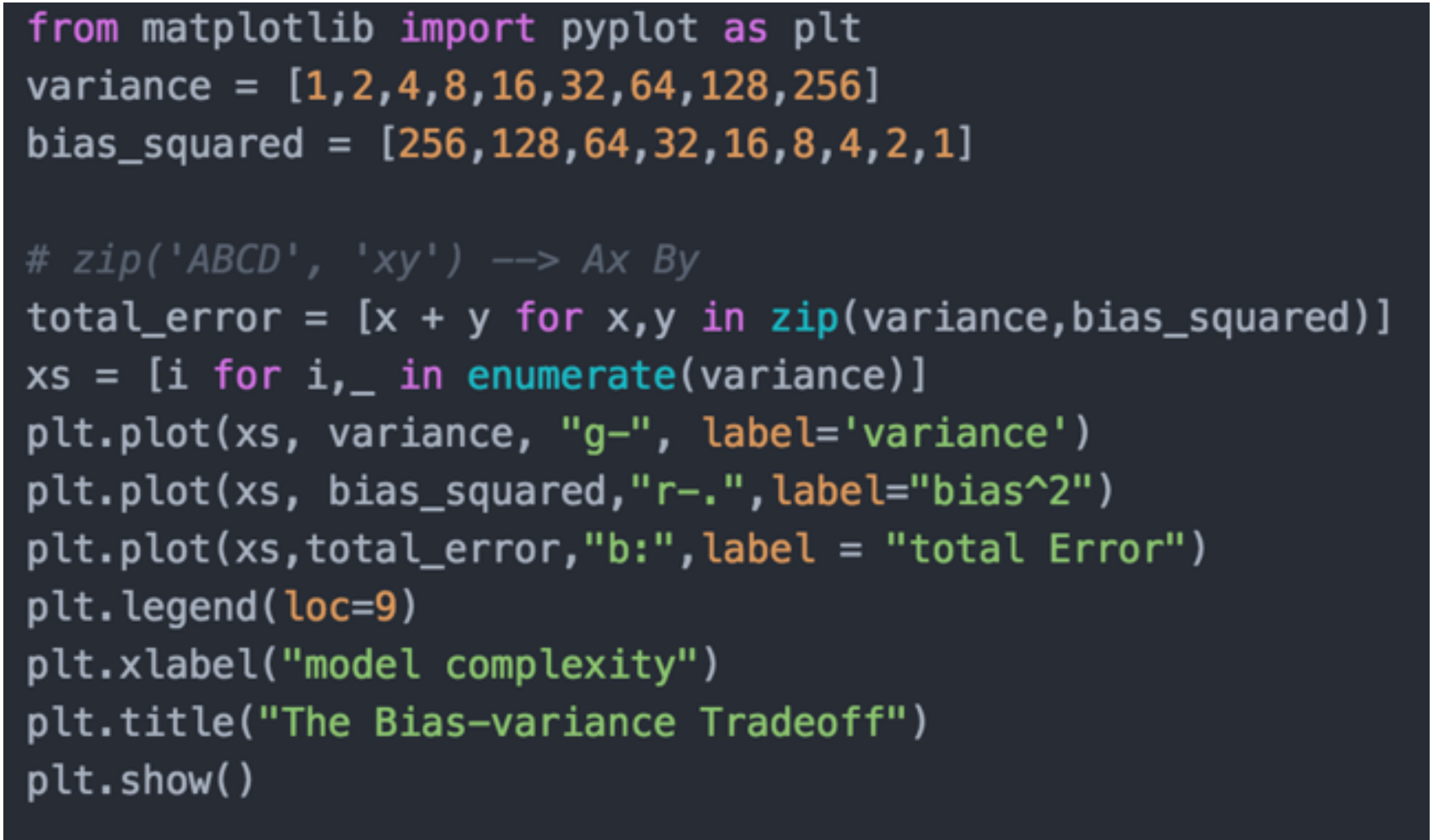
Make sure to take notes of the different things we will talk about

Bar plot



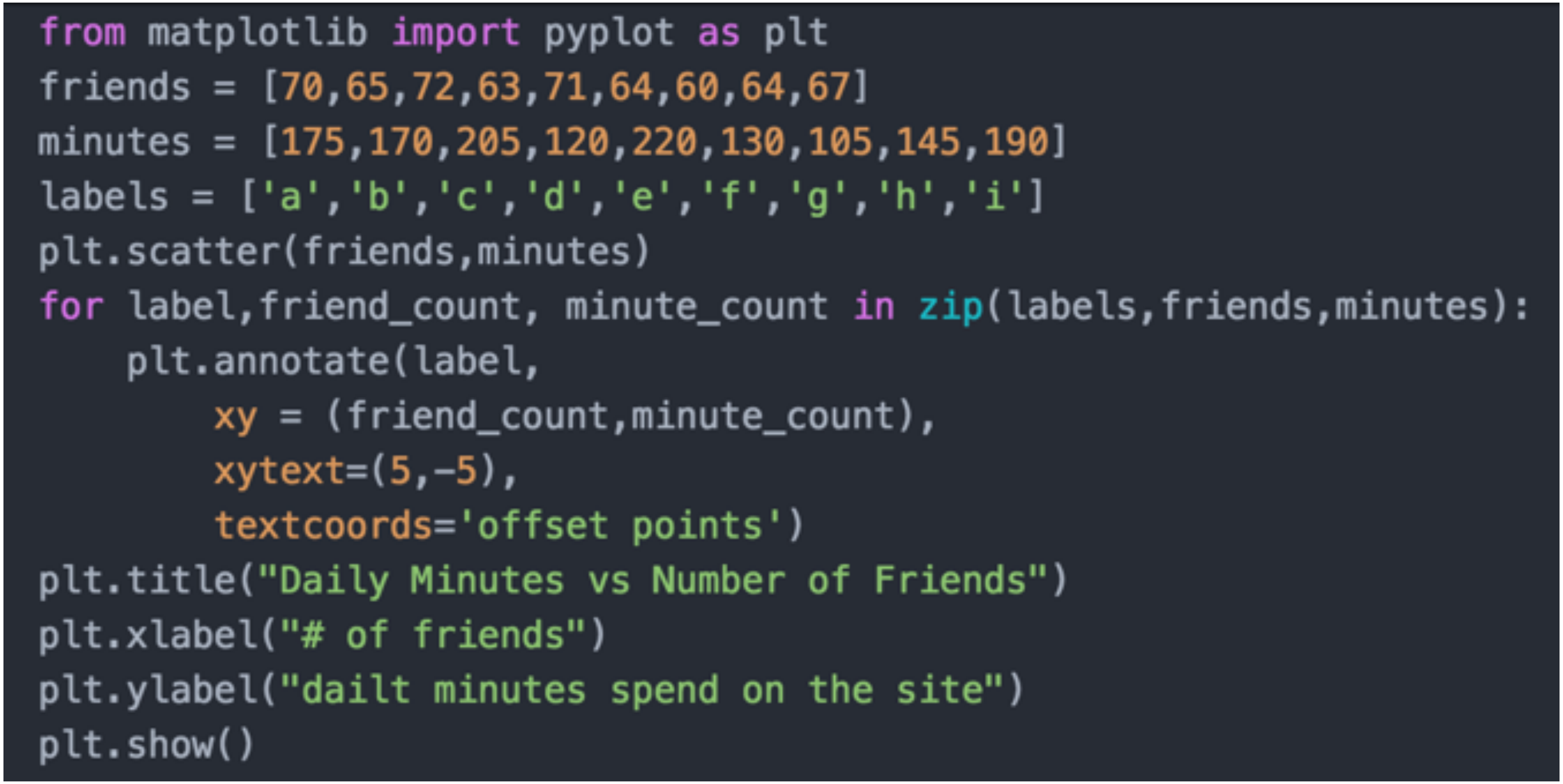
```
from matplotlib import pyplot as plt
movies = ["Annie Hall","Ben-Hur","Casablanca","Ghandi","West Side Story"]
num_oscars=[5,11,3,8,10]

xs = [i+0.1 for i,j in enumerate(movies)]
print xs
plt.bar(xs,num_oscars)
plt.ylabel("# of Academy Awards")
plt.title("My favorite Movies")
plt.xticks([i+0.5 for i,x in enumerate(movies)],movies)
plt.show()
```

lines plot



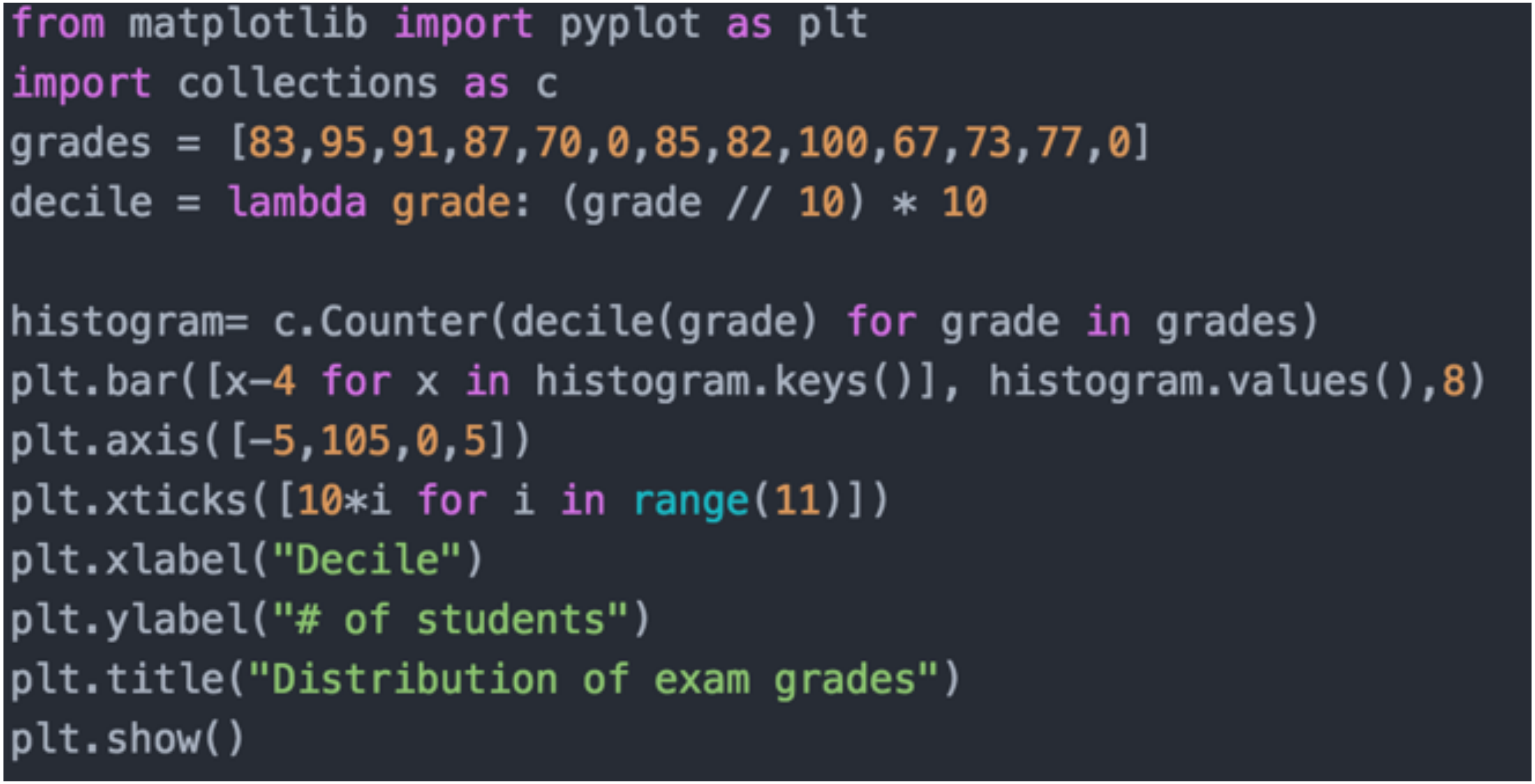
```python
from matplotlib import pyplot as plt
variance = [1,2,4,8,16,32,64,128,256]
bias_squared = [256,128,64,32,16,8,4,2,1]

# zip('ABCD', 'xy') --> Ax By
total_error = [x + y for x,y in zip(variance,bias_squared)]
xs = [i for i,_ in enumerate(variance)]
plt.plot(xs, variance, "g-", label='variance')
plt.plot(xs, bias_squared,"r-.",label="bias^2")
plt.plot(xs,total_error,"b:",label = "total Error")
plt.legend(loc=9)
plt.xlabel("model complexity")
plt.title("The Bias-variance Tradeoff")
plt.show()
```

Scatter plot



```python
from matplotlib import pyplot as plt
friends = [70,65,72,63,71,64,60,64,67]
minutes = [175,170,205,120,220,130,105,145,190]
labels = ['a','b','c','d','e','f','g','h','i']
plt.scatter(friends,minutes)
for label,friend_count, minute_count in zip(labels,friends,minutes):
    plt.annotate(label,
        xy = (friend_count,minute_count),
        xytext=(5,-5),
        textcoords='offset points')
plt.title("Daily Minutes vs Number of Friends")
plt.xlabel("# of friends")
plt.ylabel("dailt minutes spend on the site")
plt.show()
```

Decile plot



```
from matplotlib import pyplot as plt
import collections as c
grades = [83,95,91,87,70,0,85,82,100,67,73,77,0]
decile = lambda grade: (grade // 10) * 10

histogram= c.Counter(decile(grade) for grade in grades)
plt.bar([x-4 for x in histogram.keys()], histogram.values(),8)
plt.axis([-5,105,0,5])
plt.xticks([10*i for i in range(11)])
plt.xlabel("Decile")
plt.ylabel("# of students")
plt.title("Distribution of exam grades")
plt.show()
```

Exercise

Read the file data.txt and store its contents in a list

1. First element should go in list l1
2. Second element should go in list l2
3. Create a line plot that includes both lines.
4. Create a bar chart for each list
5. Create a bar chart with the decile
6. Create a scatter plot