Data Ethics and Introducing Python

Cara Marta Messina
Digital Integration Teaching Initiative
Stefani Anderson
Spring 2020, Virtual Workshop

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to email Cara any questions:
messina.c@husky.neu.edu

Workshop Agenda

Digital and Data Ethics

- Digital data collection
- Digital data research examples
- Activity: Adopt or Not? (answer a series of questions)

Activity: Introduction to Python (OPTIONAL)

- Basics of Python
- Statistical Summaries
- Visualizations

Feel free to ask questions at any point during the presentation!

Learning Objectives

- Understand the basics for how digital data is collected, analyzed, and used in our everyday lives ("big data")
- Explore the ways in which privacy and security are being reshaped and redefined through "big data"
- Understand how big data and data analytics may replicate or resist social and political power
- Learn some basics of Python to analyze data

Feel free to ask questions at any point during the presentation!

What you will learn

- Important definitions:
  - Big data
  - Surveillance capitalism
  - Algorithms/algorithmic bias
- The basics of Python (optional):
  - Data types (strings vs. integers)
  - Basic functions
  - Reading in data and creating dataframes
  - Statistical Summaries
  - Visualizations

Feel free to ask questions at any point during the presentation!

Big Data, Algorithms, and Algorithmic Bias

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Big Data

Big data collects vast amounts of data from vast amounts of users and analyzes that data quickly for particular purposes (advertizing, surveillance, search results, etc).

The goal of big data is to predict individual user behavior based on patterns from the user as well as patterns from “similar” users (based on demographic information, behavioral patterns, etc).

Feel free to ask questions at any point during the presentation!

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

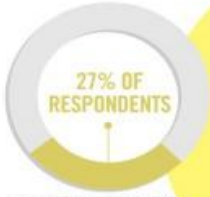
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

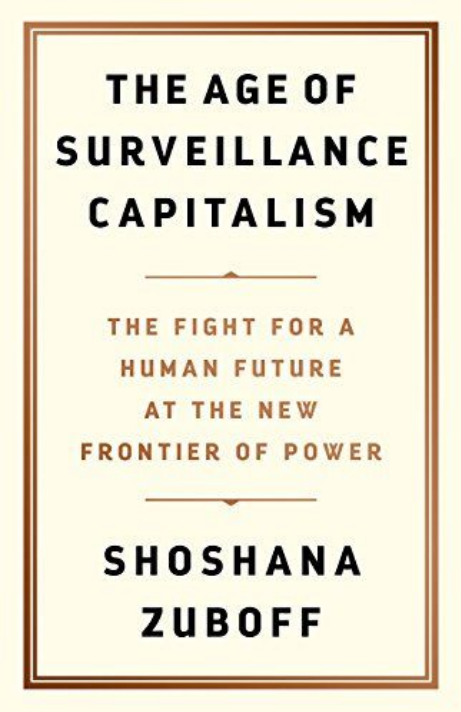
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Surveillance Capitalism

Shoshana Zuboff defines “surveillance capitalism” as the commodification of human behavior. Our “data”--our demographic information, our everyday behaviors online and in person, and who we know--is collected and sold for analysis and advertising purposes. This is one of the main goals of big data.

Our information is one of the most valuable products in America and other countries with similar economic structures.



Feel free to ask questions at any point during the presentation!

Why Talk about Big Data?

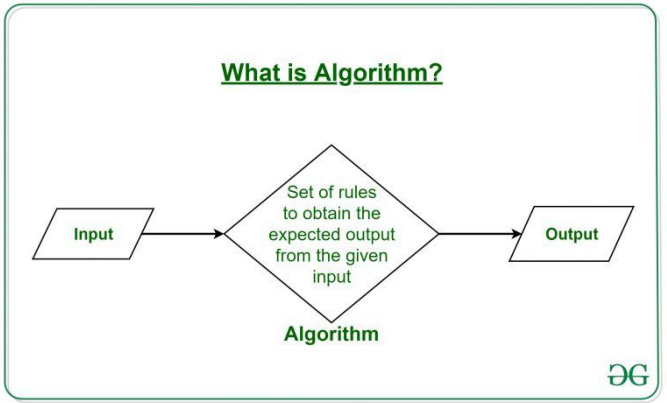
Big data is an extreme way of collecting and analyzing data for particular goals. It is easy to trace the daily impacts of big data, everywhere from your individualized newsfeed on your Twitter to your credit score.

When learning any programming language like Python, you are learning the first steps to creating **algorithms**, which are a series of commands that computers follow to provide some kind of output. Algorithms are the backbone of big data.

Feel free to ask questions at any point during the presentation!

Algorithms

An algorithm is a process of instructions provided, usually for computers to interpret and follow. There is usually an **input**, which is determined by the programmer; then there is a set of rules (the algorithm) that help lead to the **output**, or the results of the program following instructions.



Algorithms can be fairly simple, but they can also be much more complex.

Feel free to ask questions at any point during the presentation!

Algorithmic Bias

Algorithms are *not neutral*. While they do not have minds of their own, people create these algorithms. The processes and data, itself, may reflect particularly biases about society.

For example, Amazon attempted to create an algorithm analyze potential hires' resumes. Their input data was people who had been hired at tech companies and people who were not hired. Because tech companies are known to be a male-dominating field, the input data reflected this. The algorithm interpreted any mention of "women" in the new resumes as negative and rejected these applications.

Feel free to ask questions at any point during the presentation!

Activity: Adopt or Not? (mandatory)

Feel free to ask questions at any point during the presentation!

Activity: Adopt or Not?

For this activity, you will be taking on the role of the algorithm!

You will read a series of applications and answer a few questions based on these applications. These questions are both in future slides, on Blackboard, and on the handout. Please post your answers on Blackboard.

Feel free to ask questions at any point during the presentation!

Adopt or Not?

A dog adoption agency just received an application from a family interested in adopting a dog. Based on four previous applications (two were accepted and two were rejected), decide whether this family's application should be accepted or rejected.

READ THE FIVE APPLICATIONS HERE

Feel free to ask questions at any point during the presentation!

Adopt or Not? Discussion Questions

Please take some time to answer these questions and post your answers on Blackboard:

1. Based on the four previous applications? you think this new application should be accepted or rejected? Why or why not?
2. What questions did you weigh more heavily and why?
3. Who does this question favor and why? (some examples of demographic information includes race, age, gender, socioeconomic class, location, etc.)
4. What problems can you see arising from making a choice about who can adopt a dog or not based on these process?
5. What other questions would you want to include in this application form to mitigate the favorability mentioned in question 3 or the problems mentioned in question 4?

Feel free to ask questions at any point during the presentation!

Adopt or Not: Thinking Like an Algorithm

Algorithms can only interpret the data presented to them. Because you did not have any outside information, you had to make decisions about which questions seemed to matter more in the accepted and rejected applications.

Of course, the data **sample size** is fairly small (only 4), so there is no perfect answer. But even with a ton of data, there is rarely a perfect answer, especially when classifying behavior. If you are trying to determine if someone can be a good dog owner, that can be almost impossible to answer without knowing a person well.

Feel free to ask questions at any point during the presentation!

Activity: Basic Introduction to Python (optional)

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Introduction to Python

For the **second, optional portion** of this virtual workshop, you will engage with a Python notebook specifically created for your class. There will be a series of instructions in the next few slides, as well as some helpful tips if you get stuck.

On the Python notebook, itself, there will be documentation explaining specific codes and steps.

If you ever get too stuck, please email Cara Messina: messina.c@husky.neu.edu. No question is too small!

Feel free to ask questions at any point during the presentation!

What will you be working with?

Azure Notebook: the a cloud-computing software you will use to run the Jupyter notebook. You will not need to download anything.

Jupyter Notebook: A code-composing software. You can compose with regular text or choose to code with Python, R, or Julia (three programming languages)

Python: An object-oriented programming language. Programmers use Python to analyze data, build applications, create visualizations and models, and a bunch of other awesome things.

Feel free to ask questions at any point during the presentation!

Python

Because Python is a computer language, it is impossible to learn overnight (or even in a week-long workshop!). Most programmers Google questions, errors, or specific processes because the language is so large and complex.

This portion of the workshop will teach you the basics of Python, like how to run functions, read in a dataset, analyze the data, and visualize the data.

Feel free to ask questions at any point during the presentation!

Step One: Open Azure Notebooks

First, you will need to **clone** the Azure Notebook that has been created for you. **Follow this link.**

- Make sure you are signed into Azure Notebook through your Northeastern account. It's the same login information as your husky, except it ends in northeastern.edu (lastname.f@northeastern.edu)
- In the folder that the link leads too, click "Clone" in the top right corner.
  - You will be prompted to name the file.
    You can just keep it the name and click "Clone"

- The folder will now be cloned to your Azure Notebook account. You will now be able to edit the "**intro_to_python.ipynb**" file, which is what we will be working with. **Keep readings these slides, first**

Feel free to ask questions at any point during the presentation!

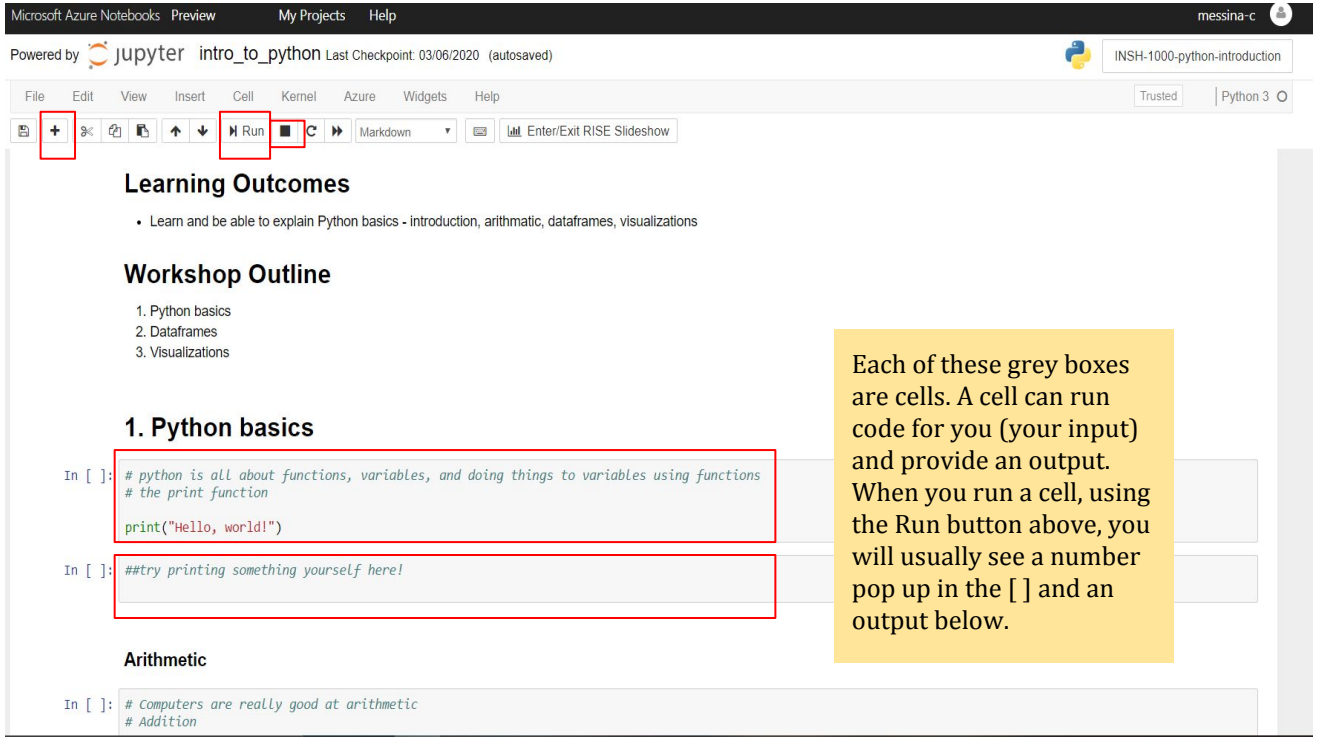
Before You Try Python!

Read the next few slides to have a basic understanding of important vocabulary and potential errors.

Keep these slides open for questions.

“intro_to_python.ipynb”

Your toolbar. The main things you will use are the “add cell” (+), “Run cell,” and “stop running” (square) buttons. These are each highlighted in red.



Each of these grey boxes are cells. A cell can run code for you (your input) and provide an output. When you run a cell, using the Run button above, you will usually see a number pop up in the [ ] and an output below.

Feel free to ask questions at any point during the presentation!

Running a Cell

Before a Cell is Run

- "In" means input
  - A line that begins with # is a COMMENT
  - Anything else is CODE. **print( )** is a function

```
In [ ]: # If you would like to comment in a cell, simply use the hashtag at the beginning of each line!
        # Our first function is the print() function
        # Click this cell and then the run function to print "Hello, world!"

        print("Hello, world!")
```

This cell has not been run yet. We know because the [ ] next to In is empty

Running

When you click "RUN," your cell will run, you may see *, which means the cell is running.

```
In [*]: # If you would like to comment in a cell, simply use the hashtag at the beginning of each line!
        # Our first function is the print() function
        # Click this cell and then the run function to print "Hello, world!"

        print("Hello, world!")
```

Your Cell Ran!

Your cell ran. You know this because:

- there is a number next to In
- there is an *output* below your cell. Note: there may not always be an output

```
In [17]: # If you would like to comment in a cell, simply use the hashtag at the beginning of each line!
         # Our first function is the print() function
         # Click this cell and then the run function to print "Hello, world!"

         print("Hello, world!")

Hello, world!
```

This is our output!!

Potential Errors

Prolonged Run: When you run your cell, it should only take a few seconds to see the number and the output. If you see In[*] for a long time, either click the STOP in the toolbar or, unfortunately, you may have to refresh your notebook.

Error Screens: You will know when there is an error. Ex:

```
In [20]: # How to extract a specific column

         df['reading_scoree'].head(20)

---------------------------------------------------------------------------
KeyError                                  Trac
~/anaconda3_420/lib/python3.5/site-packages/pa
   2133             try:
-> 2134                 return self._engine.ge
   2135             except KeyError:
```

Try checking that you either:

a) Spelled everything correctly
b) Have the proper syntax --
   use parenthesis, brackets, periods, quotation marks, etc when necessary

If everything is working, Google the error (which will be at the bottom of the output). Or email Cara! I'm always happy to help

Feel free to ask questions at any point during the presentation!

Your Turn!

Once you **cloned** the Azure Notebook folder, it's your turn to play!

Again, please let Cara know if you have any questions.

Northeastern University
NULab for Texts, Maps, and Networks