Introduction to Computation for the Humanities and Social Sciences

CS 3
Chris Tanner
DO YOU WRITE "THE DATA IS..." OR "THE DATA ARE..."?
IS IT SINGULAR OR PLURAL?

THAT IS A DEEP PHILOSOPHICAL QUESTION, ACTUALLY.

IT DEPENDS ON WHETHER YOU CONSIDER DATA TO BE FACTS (PLURAL) OR INFORMATION (WHICH IS SINGULAR).

IT'S A FASCINATING GRAMMATICAL CONUNDRUM.

WHAT IF I ONLY HAVE ONE DATA POINT?
THEN YOU HAVE BIGGER PROBLEMS THAN GRAMMAR.
“I like Big Data and I cannot lie”
— Sir Mix-a-Lot
FIRST: SOME REMINDERS

1. The course website is at http://cs.brown.edu/courses/csci0030/ and has all assignments, resources, lectures, etc

2. Submit homework via Canvas (I will soon create the HW week 2 assignment to which you can submit)

3. Homework is due Monday @ 11:59pm

4. TAs (and myself) have office hours today:
   1. me from 1-3pm (CIT 527)
   2. Anna and Caroline from 5-7pm (CIT 203)

5. Further TA Hours are on Sunday and Monday.

6. I will issue override codes today
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
What is Data Collection?

Considerations When Choosing a Dataset

Comprehensive vs Sampled Data

Data Formats

History of Programming Languages

Types of Programming Languages

Python

Pseudocode
What is Data Collection?

- Data collection is the process of collecting information (ideally in a digestible format)
- Either you collect, process, and format the data yourself
- Or you can find a dataset someone else already produced
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Considerations When Choosing a Dataset

- What data is necessary to answer our question?
- How difficult is it to analyze a dataset?
- Comprehensive data vs sampled data?
- What is the allowed usage of data under its license?
- Who collected the data?
- When was the data collected?
- How was the data collected?
- How is the data formatted?
- Does your data collection procedures need to be approved by an IRB?
- Confidentiality Concerns
Considerations When Choosing a Dataset: Difficulty

<table>
<thead>
<tr>
<th>Easy for people</th>
<th>Difficult for people</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>name</strong></td>
<td><strong>age</strong></td>
</tr>
<tr>
<td>1 Michael</td>
<td>46</td>
</tr>
<tr>
<td>2 Jim</td>
<td>31</td>
</tr>
<tr>
<td>3 Pam</td>
<td>29</td>
</tr>
<tr>
<td>4 Meredith</td>
<td>53</td>
</tr>
<tr>
<td>5 Dwight</td>
<td>35</td>
</tr>
</tbody>
</table>

Confusion at Palm Beach County polls
Some Al Gore supporters may have mistakenly voted for Pat Buchanan because of the ballot’s design.

Although the Democrats are listed second in the column on the left, they are the third hole on the ballot.

Punching the second hole casts a vote for the Reform party.

Easy for computers

Difficult for computers

[Image of a cat]

[Image of a maze]

[Image of a maze]

[Image of a video clip]
Considerations When Choosing a Dataset: Difficulty

- We want data that’s easy for the computer to analyze, but often the data we are interested in is easy for humans to analyze.
Considerations When Choosing a Dataset: Difficulty

- We want data that’s easy for the computer to analyze, but often the data we are interested in is easy for humans to analyze.
- Often we can suitably answer our questions from data pulled from these items.

**Speech transcripts**

**Scripts**

**Transcribed Lyrics**
Considerations When Choosing a Dataset: Text

- Text data exists in many different forms
- Analyzing the text can provide insights into the intent and the language of the author
Considerations When Choosing a Dataset: Size

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size Details</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senate votes per year</td>
<td>~15,000 votes, 15KB</td>
<td>Spreadsheets</td>
</tr>
<tr>
<td>The Office (1 episode)</td>
<td>~3,000 words, 13KB</td>
<td></td>
</tr>
<tr>
<td>Alice in Wonderland</td>
<td>~26,000 words, 170KB</td>
<td></td>
</tr>
<tr>
<td>The English Bible</td>
<td>~800,000 words, 4.2MB</td>
<td>Ideally suited for programming language</td>
</tr>
<tr>
<td>Trump’s Tweets</td>
<td>~800,000 words, 4MB</td>
<td></td>
</tr>
<tr>
<td>Encyclopedia Britannica</td>
<td>~40M words, 200MB</td>
<td></td>
</tr>
<tr>
<td>English Wikipedia</td>
<td>~2.9B words, 14.5GB</td>
<td>Analysis difficult</td>
</tr>
</tbody>
</table>
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Comprehensive Data

• We have access to all the data points that exist, which is usually a lot

• Collected and digitized as part of general procedures of an institution

The New York Times

13 million articles

~500 Million tweets per day

CONGRESS.GOV

100,000s votes per year
Sampled Data

- When collecting individual data is relatively expensive
- Only a portion of the population is sampled
- Not just restricted to polling or surveys
Error vs Bias

- Error is the difference between the sampled population and the population as a whole
- Error can be systematic and non-systematic
- Systematic Error is also known as bias
- Non-systematic error is also referred to as the variance
Biases in Sampled Data

• A bias in sampled data occurs when a procedure causes the sample to overrepresent a subpopulation

• They may not necessarily be intentional

• Even if you don’t think overrepresentation of a subpopulation will bias the dataset with regard to your question, it’s still a bias

• Always strive to minimize any biases in your data collection procedures
Biases in Sampled Data

Gallup

- Randomly calls two groups of ~500 people a day by sampling among all possible phone numbers

- For landlines, asks for household member who has the next birthday

- Calls people living in all 50 states

- Tries to assure 70% cellphone, 30% landlines

- Weights data to reflect the demographics of the general population
Biases in Sampled Data

**IMDb**

- Registered users rate films and TV shows on a 1-10 star rating

- Registered users are an overrepresented subpopulation relative to the general population

- Registered users who rate movies in their free time further over represents a specific segment of the general population

- “Men Are Sabotaging The Online Reviews Of TV Shows Aimed At Women” - fivethirtyeight.com

  - 60% who rated Sex in the City were women. Women gave it a 8.1, men gave it 5.8.
Biases in Sampled Data

IMDb

**Men tank the ratings of shows aimed at women**
Average difference between IMDb ratings of TV shows from men and women by share of ratings from women

**Men are more likely to give the crappiest rating**
Share of IMDb ratings of 1 (out of 10) for shows with at least 10,000 ratings by share of ratings from women

*For English language shows with 1,000 or more ratings

*Rounded to nearest 5 percent
Biases in Sampled Data

Yelp

• Registered users rate businesses on a 1-5 star scale

• Registered users tend to represent a certain subset of the population (those who are more social media inclined and opinionated)

• Customers with extreme experiences are more likely to voice their opinions
Biases in Sampled Data

Yelp

3. CVS Pharmacy
   College Hill
   291 Thayer St
   Providence, RI 02906
   (401) 331-1970
   $\$\cdot$ Drugstores

   The location of this CVS right on Thayer St is very convenient for shoppers, diners and students. You might be needing something from the drugstore after having a very large meal,.... read more

4. CVS Pharmacy
   481 Angell St
   Providence, RI 02906
   (401) 521-4340
   Pharmacy

   This is the best pharmacy in Providence, where managers Alex and Nicole extremely caring and helpful for customers.. Alex always takes time and explains in details everything we.... read more
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Data Formats

Bits and Bytes

• Data is stored to the computer in **binary**

• The smallest unit in binary is a **bit** (**binary** + **digit**)

• Every bit is a 1 or 0

• 8 consecutive bits are known as a **byte**:

\[
\begin{array}{c}
\text{byte} \\
10111010
\end{array}
\quad
\begin{array}{c}
\text{byte} \\
01001110
\end{array}
\]
Data Formats

Bits and Bytes

- Data is stored to the computer in **binary**
- The smallest unit in binary is a **bit** (**binary** + **digit**)
- Every bit is a 1 or 0
- 8 consecutive bits are known as a **byte**:
  
  \[
  \begin{array}{c}
  \text{byte} \\
  10111010
  \end{array} \quad \begin{array}{c}
  \text{byte} \\
  01001110
  \end{array}
  \]
Data Formats

**Binary vs Plain Text**

- All data stored to the computer is fundamentally represented in binary.

- Data stored to a file that can be completely interpreted as unformatted text is called **plain text**.

- Binary data is simply data that cannot be interpreted as plain text.

- Generally speaking, if your data file needs to be opened by a specific program, it is probably stored as binary data.

- If you can read your data, as is, on any text editor, it is likely plain text.
Data Formats

Plain Text Examples

- Plain text, what people really mean when they say plain text
- Ends in .txt (generally)
- No formatting, font type, font size, color, etc.
- Text position is provided by whitespace characters (space, tab, return)

ALICE’S ADVENTURES IN WONDERLAND
Lewis Carroll
THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversations?’
Data Formats

Plain Text Examples

- XML (.xml)

- These colors ——> aren’t actually stored in the file, the editor just adds them on your screen to help make it look prettier

```xml
<roll_call_vote>
  <congress>115</congress>
  <session>1</session>
  ...
  <members>
    <member>
      <member_full>Alexander (R-TN)</member_full>
      <last_name>Alexander</last_name>
      <first_name>Lamar</first_name>
      <party>R</party>
      <state>TN</state>
      <vote_cast>Yea</vote_cast>
      ...
    </member>
  </members>
</roll_call_vote>
```
Data Formats

Plain Text Examples

- CSV (.csv)
- Tab-separated (.tsv)

**Delimiter**: The character that separates each value
Data Formats

Plain Text Examples

• **JSON (.json)**

• **JavaScript Object Notation**

• Like XML, data is annotated

• A nesting of key-value pairs

• When whitespace is removed, can be more space efficient than XML

```json
{
    "firstName": "John",
    "lastName": "Smith",
    "isAlive": true,
    "age": 27,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": "10021-3100"
    },
    "phoneNumbers": [
        {
            "type": "home",
            "number": "212 555-1234"
        },
        {
            "type": "office",
            "number": "646 555-4567"
        },
        {
            "type": "mobile",
            "number": "123 456-7890"
        }
    ],
    "children": [],
    "spouse": null
}
```
Plain Text Examples

- YAML (.yaml)

- **YAML Ain’t Markup Language**

- Annotated similarly to JSON

- Programs that can read YAML can read JSON

```plaintext
invoice: 34843

date   : 2001-01-23

bill-to: &id001

given  : Chris

family : Dumars

address:

  lines: |

    458 Walkman Dr.

    Suite #292

  city    : Royal Oak

  state   : MI
```
Plain Text vs XML vs JSON vs YAML

• They can all express the same content

• Plain Text doesn’t have structure, but is universally robust

• XML is the most verbose, harder to parse

• JSON doesn’t have </stuff_here> end tags

• JSON is more succinct than XML (easier to parse)

• YAML is a tad more robust than JSON, and is even more human-readable
Data Formats

File Extensions

• File extensions are the characters that follow the ‘.’ in a filename (hello.txt)

• Sometimes your computer will only allow programs to open files with a specific file extension

• Even if many programs can read a plain text data file, file extensions are used to communicate the type of data inside (.xml, .csv, .yaml, .json)
Graph of Everything on your Computer

Binary Data

Adobe Photoshop

Your Operating System

Collection of Memes

Chrome Browser

PDFs

Adblock Extension

Text Files

Plain Text

.csv .txt .xml .py

.doc .docx .rtf
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
History of Programming Languages

Old School

- **1837** — Invented the digital programmable computer, the “Analytical Engine”.
- Included:
  - arithmetic logic unit
  - control flow (which lines of code to execute when)
  - conditional branching (e.g., “if blah then do blah”)
  - loops (keep doing some code for a certain # of times)
- He never actually built it (funding issues + co-worker drama)
History of Programming Languages

Old School

• **1843** — Wrote the first computer program ever!

• It was to compute Bernoulli Numbers on Babbage’s *Analytical Engine*

• She was precocious:

  > [The Analytical Engine] might act upon other things besides *number*, were objects found whose mutual fundamental relations could be expressed by those of the *abstract science of operations*, and which should be also susceptible of adaptations to the action of the operating notation and mechanism of the engine...Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent.
Diagram for the computation by the Engine of the Numbers of Bernoulli. See Note G. (page 729 et seq.)

<table>
<thead>
<tr>
<th>Number of Operation</th>
<th>Nature of Operation</th>
<th>Variables acted upon</th>
<th>Variables receiving results</th>
<th>Indication of change in the value on any Variable</th>
<th>Statement of Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(V_1 \times V_3)</td>
<td>(V_4, V_5, V_6)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>2</td>
<td>(-V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>3</td>
<td>(+V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>4</td>
<td>(+V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>5</td>
<td>(+V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>6</td>
<td>(-V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>7</td>
<td>(-V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>8</td>
<td>(+V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>9</td>
<td>(+V_1 + V_3)</td>
<td>(V_4)</td>
<td>(V_7)</td>
<td>(V_8 = V_9)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>10</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>11</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>12</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>13</td>
<td>(-V_2 - V_1)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>14</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>15</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>16</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>17</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>18</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>19</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>20</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>21</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>22</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
<tr>
<td>23</td>
<td>(+V_2 + V_4)</td>
<td>(V_6, V_8)</td>
<td>(V_9)</td>
<td>(V_10 = V_11)</td>
<td>2 (n)</td>
</tr>
</tbody>
</table>

Here follows a repetition of Operations thirteen to twenty-three.
History of Programming Languages

Definition

• A programming language is a formal language to describe tasks to a machine or computer

• A set of instructions used to control the behavior of a computer/machine is called a program
History of Programming Languages

Alan Turing’s Universal Turing Machine (1936)

• A mathematical description of a general purpose programmable computer

• Foundational for modern analytical methods in computer science

• Programmer writes instructions, and the execution is done via the tape moving and writing symbols to the tape.
ENIAC (1946)

• One of first electronic general-purpose computers

• Used primarily for ballistic research

• Had 10,000 capacitors (stores data)

• First programmers were all female: Kay McNulty, Betty Jennings, Betty Snyder, Marlyn Meltzer, Fran Bilas, Ruth Lichterman
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Types of Programming Languages

- FORTRAN
- C
- Pascal
- High-Level Language
- Assembly Language
- Machine Language
- Hardware
Types of Programming Languages
Types of Programming Languages

- **Machine Language (aka Machine Code):** The only code a machine can understand. Nobody can program in this language.

  1010101001010110101010101
  01011010101010001010101
  1011011110110101010101010

Illustration:
- Hardware
- Machine Language
- Assembly Language
- High-Level Language
- FORTRAN, C, Pascal
Types of Programming Languages

**Machine Language (aka Machine Code):**
The only code a machine can understand. Nobody can program in this language.

```
1010101010010101010101010
0101101010100100101010101
1011101111101101010101010
```

**Assembly Language:**
Lowest level language to program in; code is translated into Machine Language.

```
mov ecx, 16
mov ebp, ecx
sub esp, 4
push edi
push ecx
add [ebp-4], edi
add eax, [ebp-4]
```
Types of Programming Languages

High-Level Language: (e.g., Python, FORTRAN, C, etc) Allows humans to easily write and read code which the computer will automatically translate to lower-level code and execute.
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Lecture 3

- What is Data Collection?
- Considerations When Choosing a Dataset
- Comprehensive vs Sampled Data
- Data Formats
- History of Programming Languages
- Types of Programming Languages
- Python
- Pseudocode
Python

- Python is a high-level programming language: single lines of Python code generally specify more than one instruction to the computer.

- This means that your Python code is compiled to low-level instructions called byte-code before your code is actually run.

- The byte-code is then interpreted line-by-line into instructions the computer understands (machine code) and executed simultaneously.

- Python is general purpose, and is not particularly specialized to a single task, unlike a language like R, Matlab or Stata.

- Python is open-source, and a large community supports the code.
• You will be writing programs in the Python language

• Often times, these programs will require code written by others known as **libraries**

• In python, a library is a collection of **modules**, and your program will reference the individual modules.
Next lecture, we dive into Python and start to build a programming foundation
Lab Time!