# Data Science

## Data and Text File Types

Risa Myers
Chris Jermaine
Marmar Orooji

Rice University

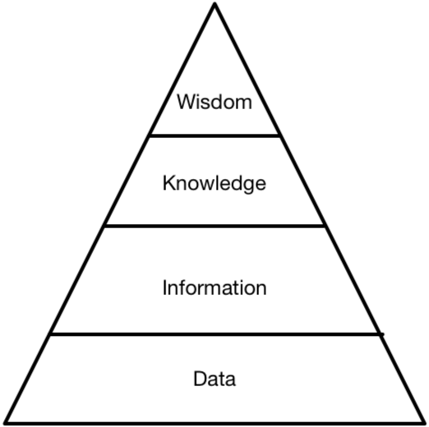
# Introduction to data and common text data file types

- What is data?
- Motivation
- Description of common text/data file types
- Discussion of issues and benefits of each
- List of common challenges
- Introduction to file encoding

# What is data?

- Raw “stuff"

# DIKW hierarchy



- What can be done with the knowledge - e.g. Treatment plan
- What you know - there are disease cells
- Processed data that is human interpretable - an image of cells
- Raw material, often automatically collected - sensor data

# Data example

| | characteristic 1 | characteristic 2 | characteristic 3 | ... |
|---|---|---|---|---|
| object 1 | 1 | red | Jan 1, 2020 | ... |
| object 2 | 3 | blue | Jan 10, 2020 | ... |
| object 3 | 6 | blue | | ... |
| ... | | | | ... |

- Each row represents a single object (patient, dog, weather on a given day)
- Each column represents a single characteristic / attribute (age, breed, temperature)
- Not every object will have a value for every characteristic
- Some characteristic values may be unique

# File formats: Motivation: Why do we care?

- Because data is available in different formats
- Because it’s helpful to understand the characteristics of the different formats

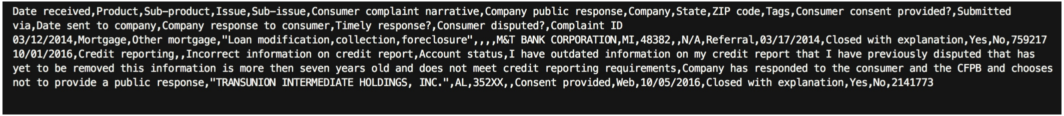
# Common text data file types

1. CSV - comma separated value
2. TXT - ASCII or Unicode text file
3. JSON - Javascript Object Notation
4. XML - eXtensible Markup Language
5. XLS/XLSX - Excel

# CSV – Comma Separated Value

1. Most common
2. One of the oldest formats
3. Frequently used to avoid compatibility problems between architectures
4. Delimited data
5. Sometimes tab or pipe

1. Common issues
   1. Commas that are part of the data
   2. Embedded line breaks
   3. Every field must be included in each row
2. Benefits
   1. Human readable
   2. Native support in many programming languages and tools



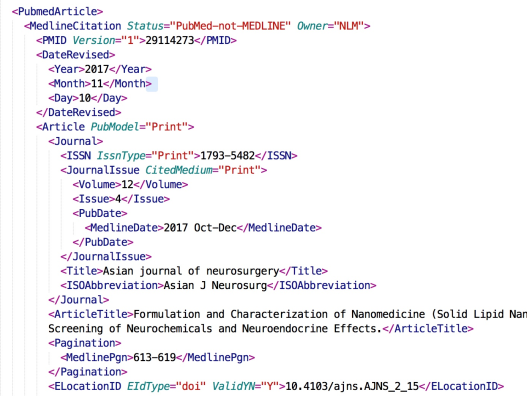
```
Date received,Product,Sub-product,Issue,Sub-issue,Consumer complaint narrative,Company public response,Company,State,ZIP code,Tags,Consumer consent provided?,Submitted
via,Date sent to company,Company response to consumer,Timely response?,Consumer disputed?,Complaint ID
03/12/2014,Mortgage,Other mortgage,"Loan modification,collection,foreclosure",,,,M&T BANK CORPORATION,MI,48382,,N/A,Referral,03/17/2014,Closed with explanation,Yes,No,759217
10/01/2016,Credit reporting,,Incorrect information on credit report,Account status,I have outdated information on my credit report that I have previously disputed that has
yet to be removed this information is more then seven years old and does not meet credit reporting requirements,Company has responded to the consumer and the CFPB and chooses
not to provide a public response,"TRANSUNION INTERMEDIATE HOLDINGS, INC.",AL,352XX,,Consent provided,Web,10/05/2016,Closed with explanation,Yes,No,2141773
```

# TXT – Text

1. Often another name for CSV
2. Could use a different delimiter
3. Could be unformatted
4. Could just be lines of text
5. Common issues
    - May not have a consistent format
6. Benefits
    - Human readable

# XML – Extensible Markup Language

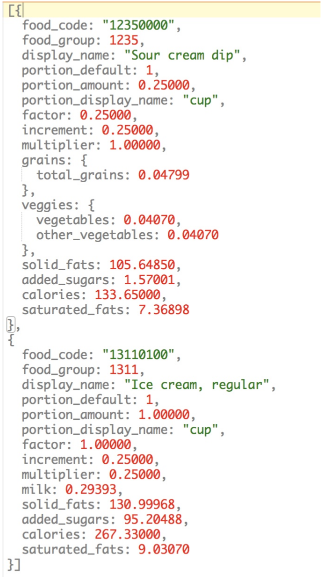
1. Language for self-describing data
2. Uses user defined tags
3. Similar to HTML
4. Designed for use by web browsers
5. Common issues
   1. Can be fragile
   2. Somewhat human readable
6. Benefits
   1. Support for more than just data
   2. Not all elements must be present for each object



```
<PubmedArticle>
  <MedlineCitation Status="PubMed-not-MEDLINE" Owner="NLM">
    <PMID Version="1">29114273</PMID>
    <DateRevised>
      <Year>2017</Year>
      <Month>11</Month>
      <Day>10</Day>
    </DateRevised>
    <Article PubModel="Print">
      <Journal>
        <ISSN IssnType="Print">1793-5482</ISSN>
        <JournalIssue CitedMedium="Print">
          <Volume>12</Volume>
          <Issue>4</Issue>
          <PubDate>
            <MedlineDate>2017 Oct-Dec</MedlineDate>
          </PubDate>
        </JournalIssue>
        <Title>Asian journal of neurosurgery</Title>
        <ISOAbbreviation>Asian J Neurosurg</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Formulation and Characterization of Nanomedicine (Solid Lipid Nan
      Screening of Neurochemicals and Neuroendocrine Effects.</ArticleTitle>
      <Pagination>
        <MedlinePgn>613-619</MedlinePgn>
      </Pagination>
      <ELocationID EIdType="doi" ValidYN="Y">10.4103/ajns.AJNS_2_15</ELocationID>
```

# JSON – JavaScript Object Notation

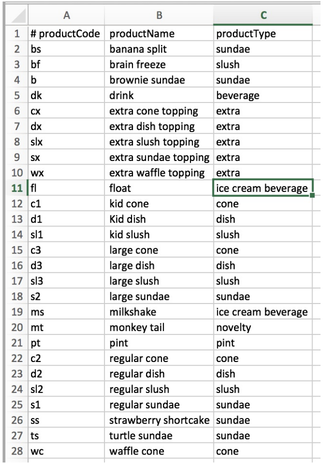
1. Derived from JavaScript
2. Originally designed for server-to-browser communication
3. Open standard
4. Attribute–value pairs
5. Array support
6. Common issues
    1. Can be fragile
7. Benefits
    1. Human readable
    2. Native support in many programming languages and tools
    3. Not all attributes must be present for each object
    4. Lighter-weight than XML



```
[{
  food_code: "12350000",
  food_group: 1235,
  display_name: "Sour cream dip",
  portion_default: 1,
  portion_amount: 0.25000,
  portion_display_name: "cup",
  factor: 0.25000,
  increment: 0.25000,
  multiplier: 1.00000,
  grains: {
    total_grains: 0.04799
  },
  veggies: {
    vegetables: 0.04070,
    other_vegetables: 0.04070
  },
  solid_fats: 105.64850,
  added_sugars: 1.57001,
  calories: 133.65000,
  saturated_fats: 7.36898
},
{
  food_code: "13110100",
  food_group: 1311,
  display_name: "Ice cream, regular",
  portion_default: 1,
  portion_amount: 1.00000,
  portion_display_name: "cup",
  factor: 1.00000,
  increment: 0.25000,
  multiplier: 0.25000,
  milk: 0.29393,
  solid_fats: 130.99968,
  added_sugars: 95.20488,
  calories: 267.33000,
  saturated_fats: 9.03070
}]
```

# XLS/XLSX – Excel

1. Microsoft Excel
2. Extension to XML
3. Compressed text file
4. Common issues
   - Strips leading zeros
   - Makes assumptions about data formats
   - Poor scalability
   - Only machine readable
   - ...
5. Benefits
   - Supports Excel



|  | A | B | C |
|---|---|---|---|
| 1 | # productCode | productName | productType |
| 2 | bs | banana split | sundae |
| 3 | bf | brain freeze | slush |
| 4 | b | brownie sundae | sundae |
| 5 | dk | drink | beverage |
| 6 | cx | extra cone topping | extra |
| 7 | dx | extra dish topping | extra |
| 8 | slx | extra slush topping | extra |
| 9 | sx | extra sundae topping | extra |
| 10 | wx | extra waffle topping | extra |
| 11 | fl | float | ice cream beverage |
| 12 | c1 | kid cone | cone |
| 13 | d1 | Kid dish | dish |
| 14 | sl1 | kid slush | slush |
| 15 | c3 | large cone | cone |
| 16 | d3 | large dish | dish |
| 17 | sl3 | large slush | slush |
| 18 | s2 | large sundae | sundae |
| 19 | ms | milkshake | ice cream beverage |
| 20 | mt | monkey tail | novelty |
| 21 | pt | pint | pint |
| 22 | c2 | regular cone | cone |
| 23 | d2 | regular dish | dish |
| 24 | sl2 | regular slush | slush |
| 25 | s1 | regular sundae | sundae |
| 26 | ss | strawberry shortcake | sundae |
| 27 | ts | turtle sundae | sundae |
| 28 | wc | waffle cone | cone |

# Common issues

1. Unusual encodings
2. Header / no header line
3. Byte order / “endianness"
4. Date format
5. Missing data handling

# Think-Pair-Share

- Which file type would be best for
  1. Sparse data?
  2. Text heavy data?
  3. Dense numeric and categorical data?
  4. Big datasets?

A csv

B TXT

C XML

D JSON

E Excel

# Encoding schemes

- What is an encoding scheme?
  - A mapping of characters to numbers
- Why use an encoding scheme?
  - For efficient transmission and storage
  - So the transmitter and receiver both interpret the text the same way

# ASCII encoding scheme

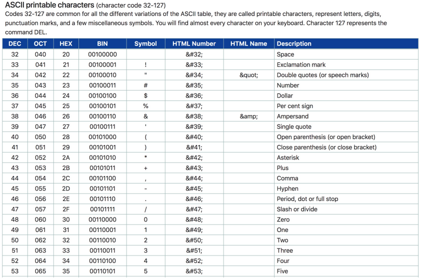
1 Oldest – 1963

2 Based on US characters

3 Encodes 128 characters into 7-bit integers

4 Minimal storage

5 Doesn't include all characters

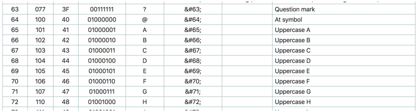


**ASCII printable characters** (character code 32-127)

Codes 32-127 are common for all the different variations of the ASCII table, they are called printable characters, represent letters, digits, punctuation marks, and a few miscellaneous symbols. You will find almost every character on your keyboard. Character 127 represents the command DEL.

| DEC | OCT | HEX | BIN | Symbol | HTML Number | HTML Name | Description |
|---|---|---|---|---|---|---|---|
| 32 | 040 | 20 | 00100000 | | &#32; | | Space |
| 33 | 041 | 21 | 00100001 | ! | &#33; | | Exclamation mark |
| 34 | 042 | 22 | 00100010 | " | &#34; | &quot; | Double quotes (or speech marks) |
| 35 | 043 | 23 | 00100011 | # | &#35; | | Number |
| 36 | 044 | 24 | 00100100 | $ | &#36; | | Dollar |
| 37 | 045 | 25 | 00100101 | % | &#37; | | Per cent sign |
| 38 | 046 | 26 | 00100110 | & | &#38; | &amp; | Ampersand |
| 39 | 047 | 27 | 00100111 | ' | &#39; | | Single quote |
| 40 | 050 | 28 | 00101000 | ( | &#40; | | Open parenthesis (or open bracket) |
| 41 | 051 | 29 | 00101001 | ) | &#41; | | Close parenthesis (or close bracket) |
| 42 | 052 | 2A | 00101010 | * | &#42; | | Asterisk |
| 43 | 053 | 2B | 00101011 | + | &#43; | | Plus |
| 44 | 054 | 2C | 00101100 | , | &#44; | | Comma |
| 45 | 055 | 2D | 00101101 | - | &#45; | | Hyphen |
| 46 | 056 | 2E | 00101110 | . | &#46; | | Period, dot or full stop |
| 47 | 057 | 2F | 00101111 | / | &#47; | | Slash or divide |
| 48 | 060 | 30 | 00110000 | 0 | &#48; | | Zero |
| 49 | 061 | 31 | 00110001 | 1 | &#49; | | One |
| 50 | 062 | 32 | 00110010 | 2 | &#50; | | Two |
| 51 | 063 | 33 | 00110011 | 3 | &#51; | | Three |
| 52 | 064 | 34 | 00110100 | 4 | &#52; | | Four |
| 53 | 065 | 35 | 00110101 | 5 | &#53; | | Five |

⋮



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 63 | 077 | 3F | 00111111 | ? | &#63; | | Question mark |
| 64 | 100 | 40 | 01000000 | @ | &#64; | | At symbol |
| 65 | 101 | 41 | 01000001 | A | &#65; | | Uppercase A |
| 66 | 102 | 42 | 01000010 | B | &#66; | | Uppercase B |
| 67 | 103 | 43 | 01000011 | C | &#67; | | Uppercase C |
| 68 | 104 | 44 | 01000100 | D | &#68; | | Uppercase D |
| 69 | 105 | 45 | 01000101 | E | &#69; | | Uppercase E |
| 70 | 106 | 46 | 01000110 | F | &#70; | | Uppercase F |
| 71 | 107 | 47 | 01000111 | G | &#71; | | Uppercase G |
| 72 | 110 | 48 | 01001000 | H | &#72; | | Uppercase H |

[a]

[a] https://www.ascii-code.com/

# Unicode encoding schemes

1. 1990s
2. All characters from all national alphabets
3. 8, 16 or 32 bits per character
4. First 128 codes match ASCII
5. Uses more space
6. ... but covers all characters

# Wrap up

? What do we know now that we didn’t know before?

1. Discussed common text data formats
    1. CSV/TXT
    2. JSON
    3. XML
    4. Excel
2. (Hopefully) understand the differences
3. Discussed character encodings

? How can we use what we learned today?