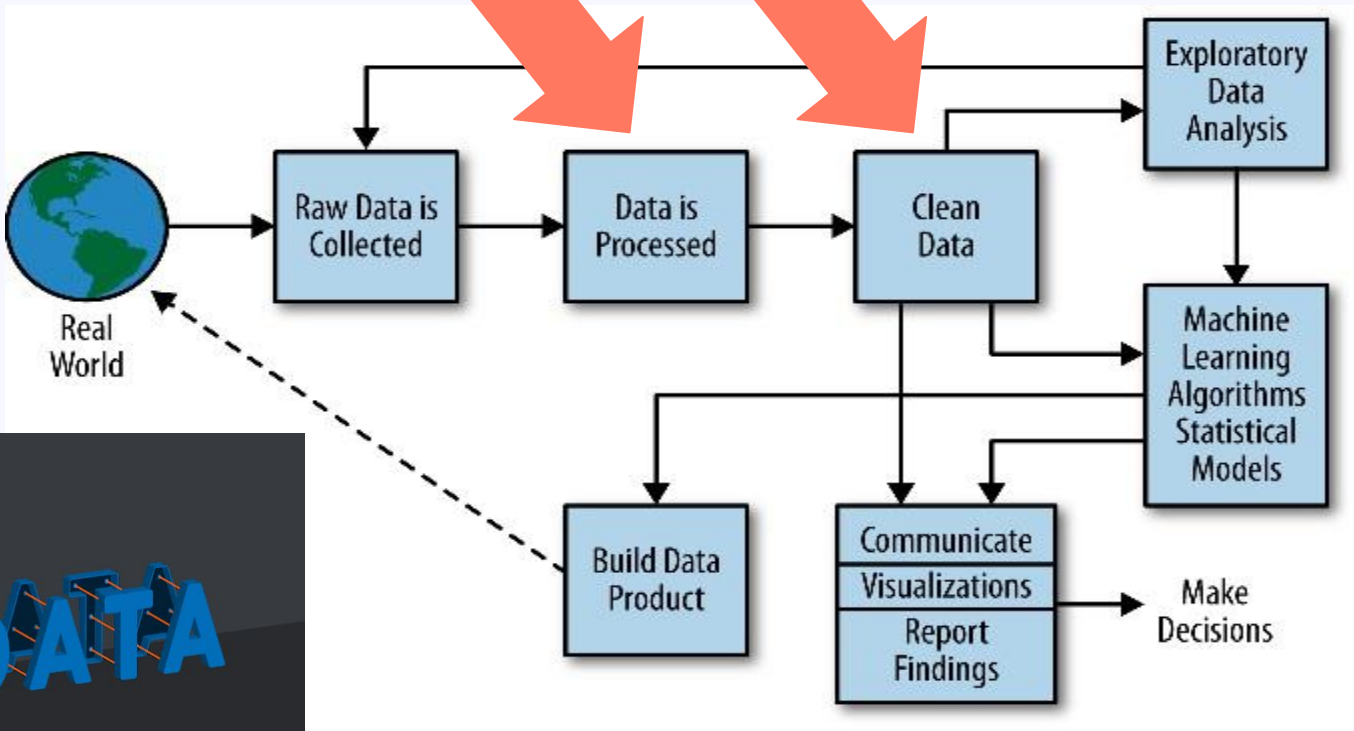
CS196

Data Science

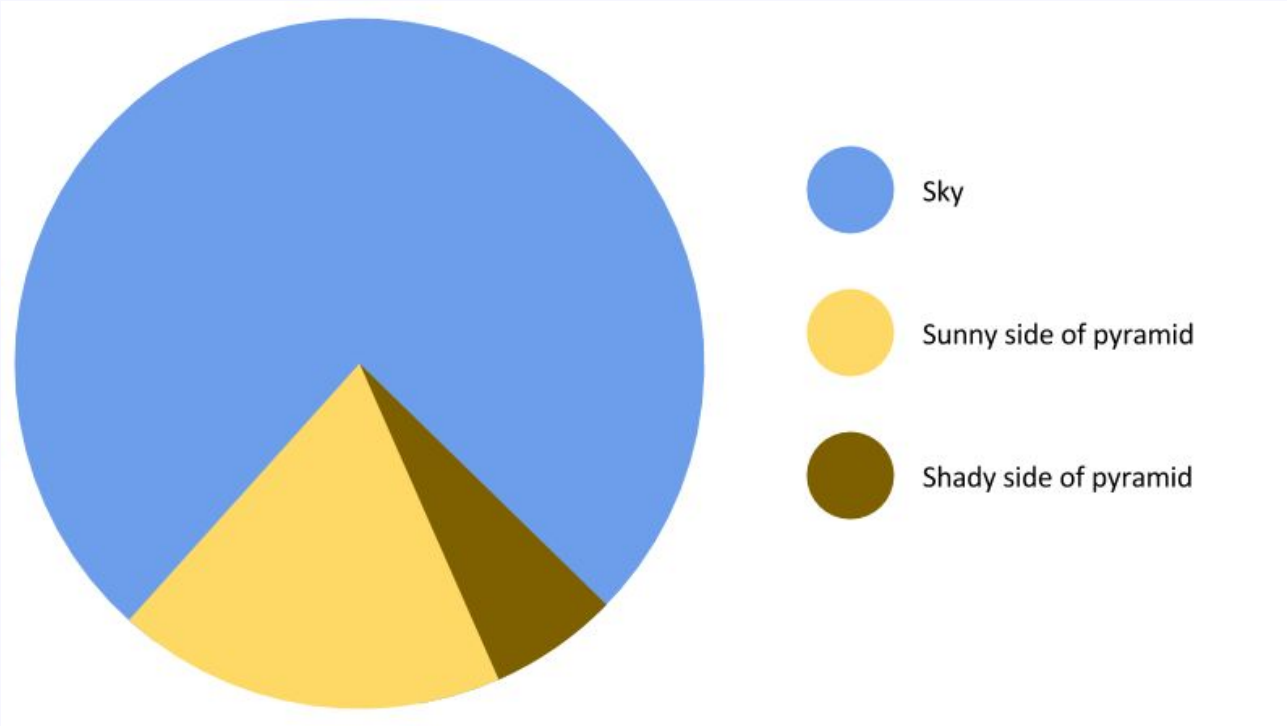
CS196

Basic Data Structures

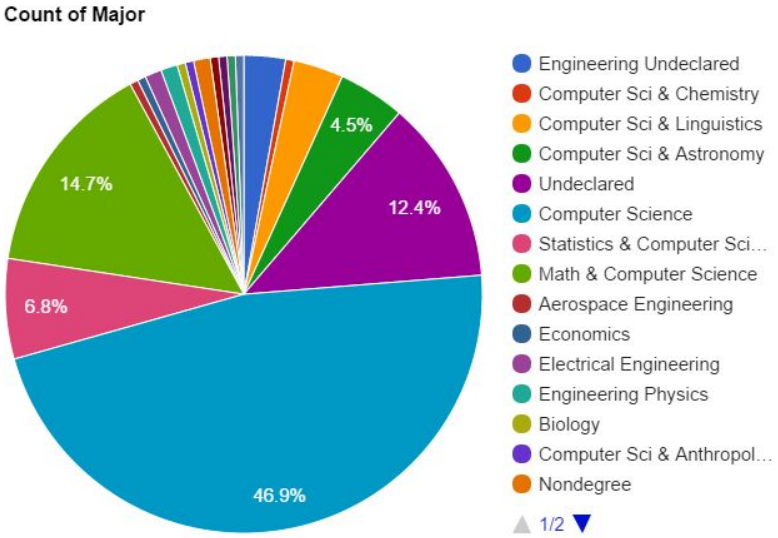
Week 3



Process of “Data Science”



Once Data set is processed and cleaned, we can easily communicate through Data/Visualization.



Once Data set is processed and cleaned, we can easily communicate through Data/Visualization.

File Types

In what form do we store Data?

Structured? Unstructured?

Text? Photo? Video?

CSV - Comma Separated Values

- Fields separated by a delimiter (a comma)
- Rows separated by newlines

```
1  36301122334,Message text 1,4/29/2010 12:30
2  36301122335,Message text 2,4/30/2010 12:30
3  36301122336,Message text 3,5/1/2010 12:30
4  36301122337,Message text 4,5/2/2010 12:30
5  36301122338,Message text 5,5/3/2010 12:30
6  36301122339,Message text 6,5/4/2010 12:30
7  36301122340,Message text 7,5/5/2010 12:30
8  36301122341,Message text 8,5/6/2010 12:30
9  36301122342,Message text 9,5/7/2010 12:30
10 36301122343,Message text 10,5/8/2010 12:30
11 36301122344,Message text 11,5/9/2010 12:30
12 36301122345,Message text 12,5/10/2010 12:30
13 36301122346,Message text 13,5/11/2010 12:30
14 36301122347,Message text 14,5/12/2010 12:30
15 36301122348,Message text 15,5/13/2010 12:30
16 36301122349,Message text 16,5/14/2010 12:30
17 36301122350,Message text 17,5/15/2010 12:30
18 36301122351,Message text 18,5/16/2010 12:30
19 36301122352,Message text 19,5/17/2010 12:30
20 36301122353,Message text 20,5/18/2010 12:30
```

File Types

In what form do we store Data?

Structured? Unstructured?

Text? Photo? Video?

CS196

JSON - JavaScript Object Notation

- Supported Data Types:
  - Number
  - Boolean
  - String
  - Array
  - Object (Dictionary)
- Good for representing hierarchical data
- Used ubiquitously in the web

File Types

In what form do we store Data?

Structured? Unstructured?

Text? Photo? Video?

JSON - JavaScript Object Notation

Request

```
{ "store": {
    "book": [
      { "category": "reference",
        "author": "Nigel Rees",
        "title": "Sayings of the Century",
        "price": 8.95
      },
      { "category": "fiction",
        "author": "Evelyn Waugh",
        "title": "Sword of Honour",
        "price": 12.99
      },
      { "category": "fiction",
        "author": "Herman Melville",
        "title": "Moby Dick",
        "isbn": "0-553-21311-3",
        "price": 8.99
      },
      { "category": "fiction",
        "author": "J. R. R. Tolkien",
        "title": "The Lord of the Rings",
        "isbn": "0-395-19395-8",
        "price": 22.99
      }
    ],
    "bicycle": {
      "color": "red",
      "price": 19.95
    }
  }
}
```

Response [HTTP/1.1 200 OK]

16-Mar-2012 09:46:27

```
{
  "store" : {
    "book" : [ {
      "category" : "reference",
      "author" : "Nigel Rees",
      "title" : "Sayings of the Century",
      "price" : 8.95
    }, {
      "category" : "fiction",
      "author" : "Evelyn Waugh",
      "title" : "Sword of Honour",
      "price" : 12.99
    }, {
      "category" : "fiction",
      "author" : "Herman Melville",
      "title" : "Moby Dick",
      "isbn" : "0-553-21311-3",
      "price" : 8.99
    }, {
      "category" : "fiction",
      "author" : "J. R. R. Tolkien",
      "title" : "The Lord of the Rings",
      "isbn" : "0-395-19395-8",
      "price" : 22.99
    } ],
    "bicycle" : {
      "color" : "red",
      "price" : 19.95
    },
    "car" : {
      "make" : "Ford",
      "airbags" : true,
      "doors" : 4,
      "price" : 1111.0
    }
  }
}
```

Data Parsing + Manipulation

- Converting to native data structures
  - Python has csv, json parsing modules
- Selecting fields
  - I.e. “from phonebook select last_names”
- Selecting entries
  - I.e. “from phonebook select ‘Tyler Kim’”
- “Exploration”
  - Drill down into fields / entries, do simple visualizations, grasp structure of dataset

Jupyter Time

http://bit.ly/aesthetichackerspace

Week 3 Challenge

Should you choose to accept...
(you should)

CS196

To Do:

1. Git checkout the Data_Hackerspace repo

2. Mass parse all the json files at once!
   (Hint: `import glob, os`**)**

3. Print only the “lead_paragraph” section from all 27 json files. Each files has up to 12 of them.

4. Extra Credit if you can do it in less than 5 lines
   Extra Extra Credit if you can it in less than 3

Attendance: http://bit.ly/dataweek3