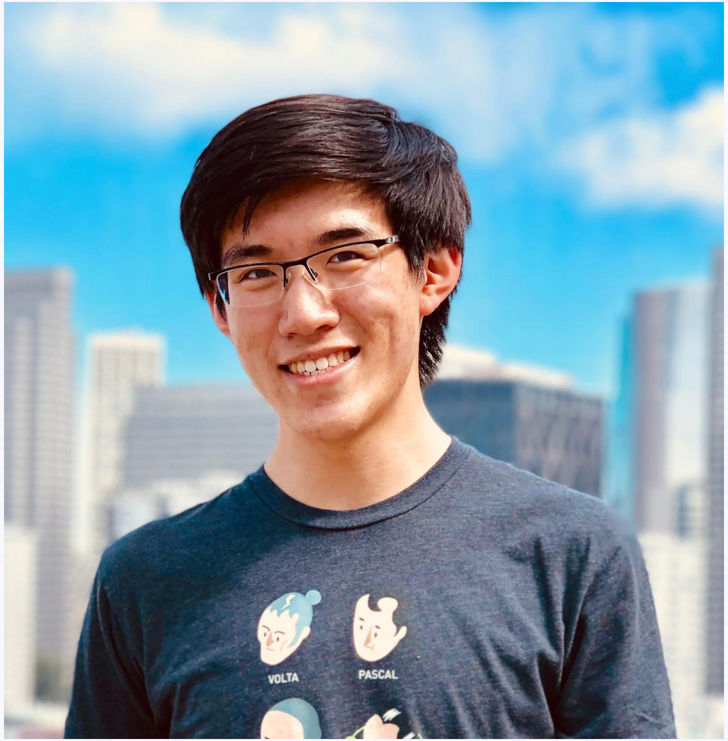
CS196

Data Science

WAIT 45 MINS IN LINE
THEY TELL YOU
TO APPLY ONLINE



Duke Vijit

[Exec] Course Staff CS 196-25
Machine Learning Snake Oil Salesman

@wdv2

Sahil Bhatt

[Exec] Course Staff CS 196-25
Data, Cloud Computing, and Tech Support

@sahilb2

Getting started with Jupyter

- The data scientists' notebook
- Mac / Linux:

```
$ sudo pip install jupyter
$ jupyter notebook
```

- Windows:
  - https://jupyter.readthedocs.io/en/latest/install.html
  - Install Anaconda
  - Follow instructions to install Jupyter

Mini Syllabus

Email Policy: Use Slack. Join the **#datascience** channel

If you must, your subject line must include [CS196]

Office Hours: 1 hour after each Hackerspace (8-9pm Monday)

Grading: Attendance + bi-weekly labs (on Github)

Expectation: Attendance / participation is mandatory!

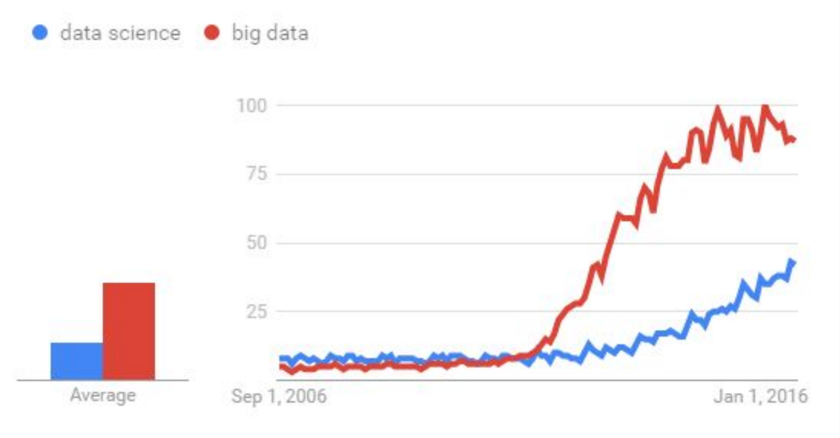
Objectives

- Familiarity with tools of data science, data modeling, and infrastructure
- Several data science projects to get a feel for the breadth of the field
- Given massive loads of data, know what to do.

Course Plan

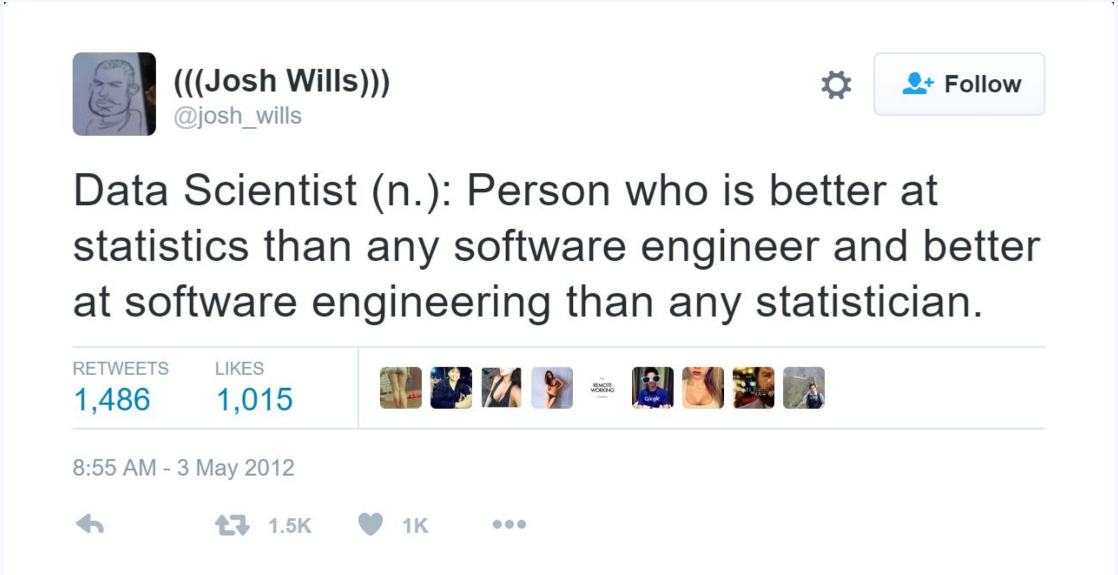
Week 2	Overview and Data Parsing
Week 3	Numpy & Pandas
Week 4	Data Visualization
Week 5	API
Week 6	Data Mining & Web Scraping
Week 7	Data Modeling in SKLearn
Week 8	Unsupervised Learning
Week 9	Supervised Learning
Week 10	SQL Databases
Week 11	NoSQL Databases
Week 12	MapReduce
Week 13	Distributed Systems for Data

Why Data *Science*?



- Data driven thinking is the future
- The scale of data is increasing exponentially
- Data has no intrinsic value: it needs to be interpreted and explained

Data Scientist?



(((Josh Wills)))
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 1,486 LIKES 1,015

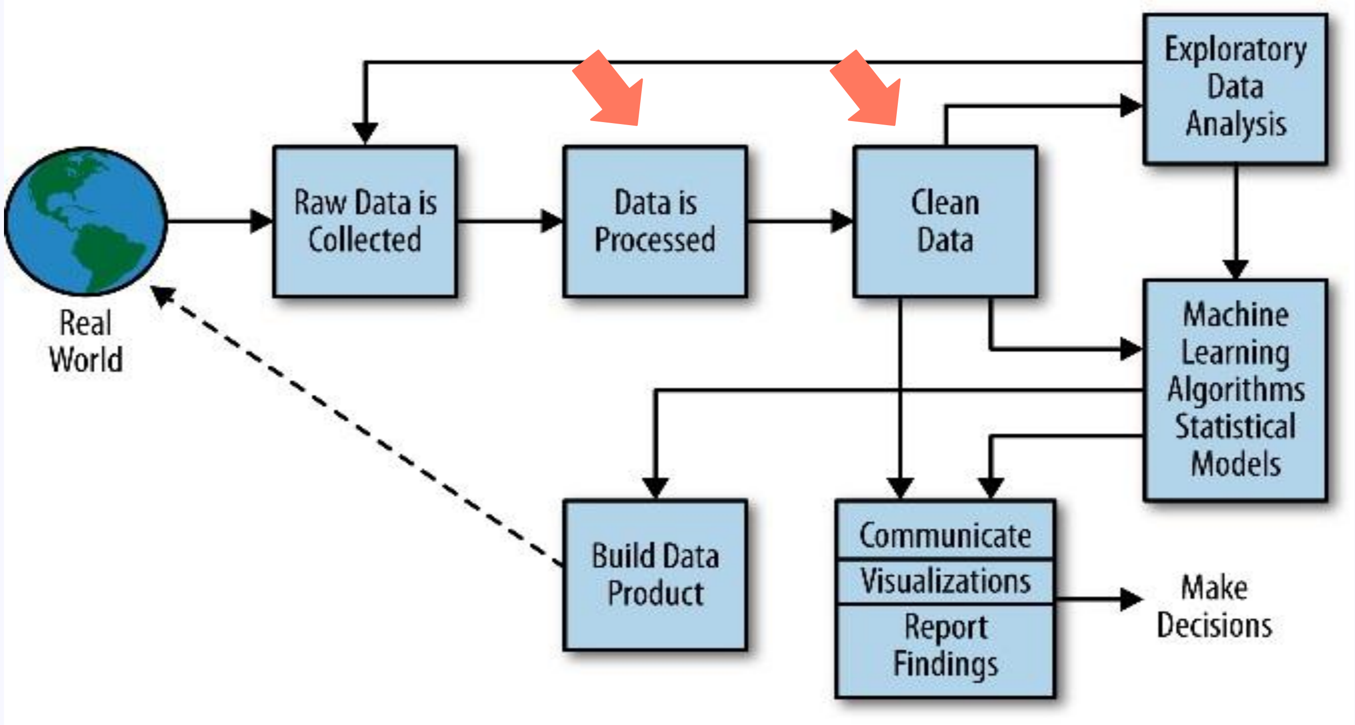
8:55 AM - 3 May 2012

CS196

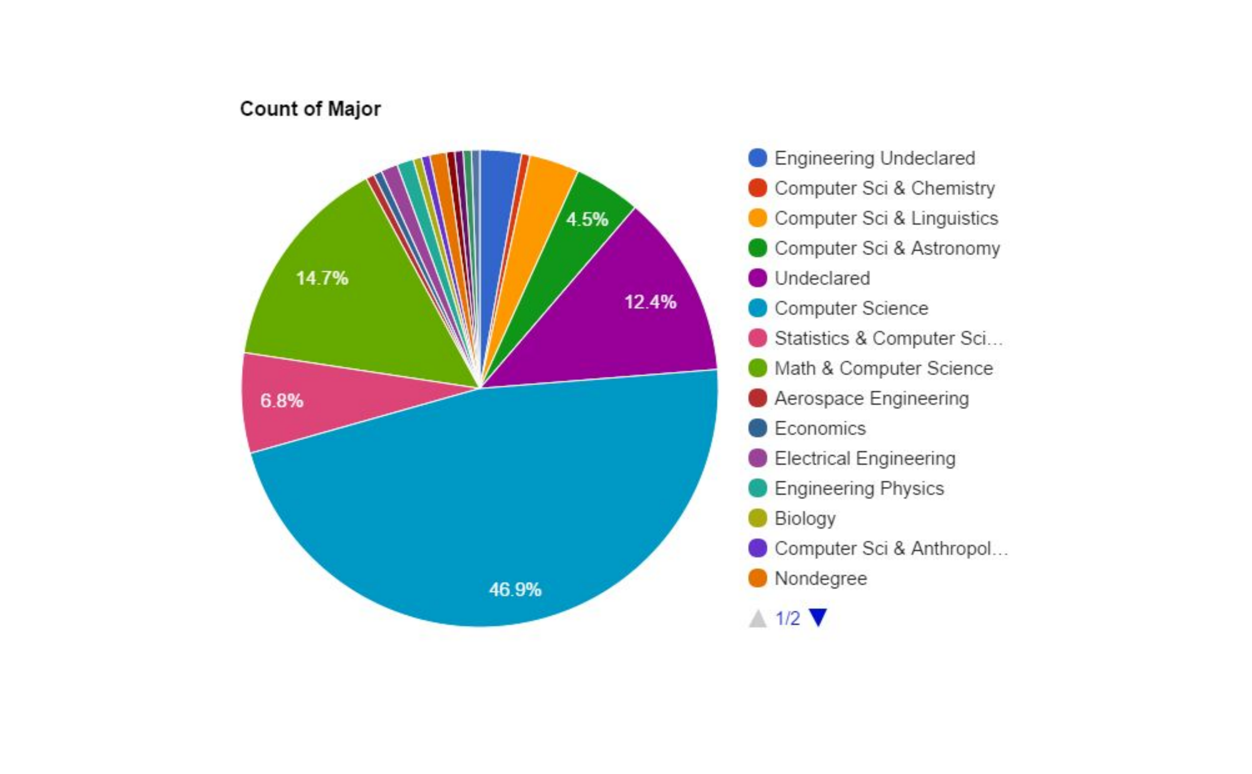
Basic Data Formats

Week 2 of 196 (Week 3 of school) :)

Channel: #datascience



Process of “Data Science”



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

File Types

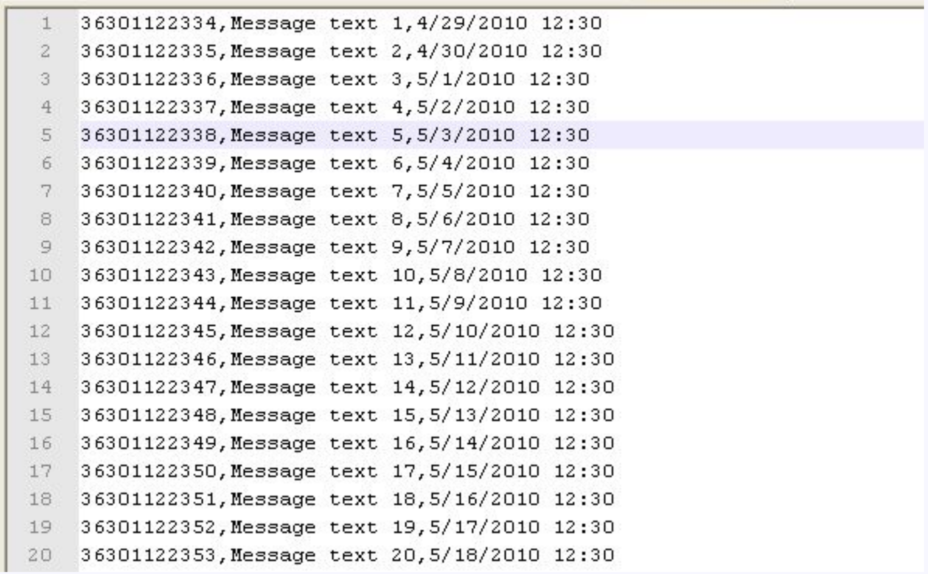
In what form do we store Data?

Structured? Unstructured?

Text? Photo? Video?

CSV - Comma Separated Values

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



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

File Types

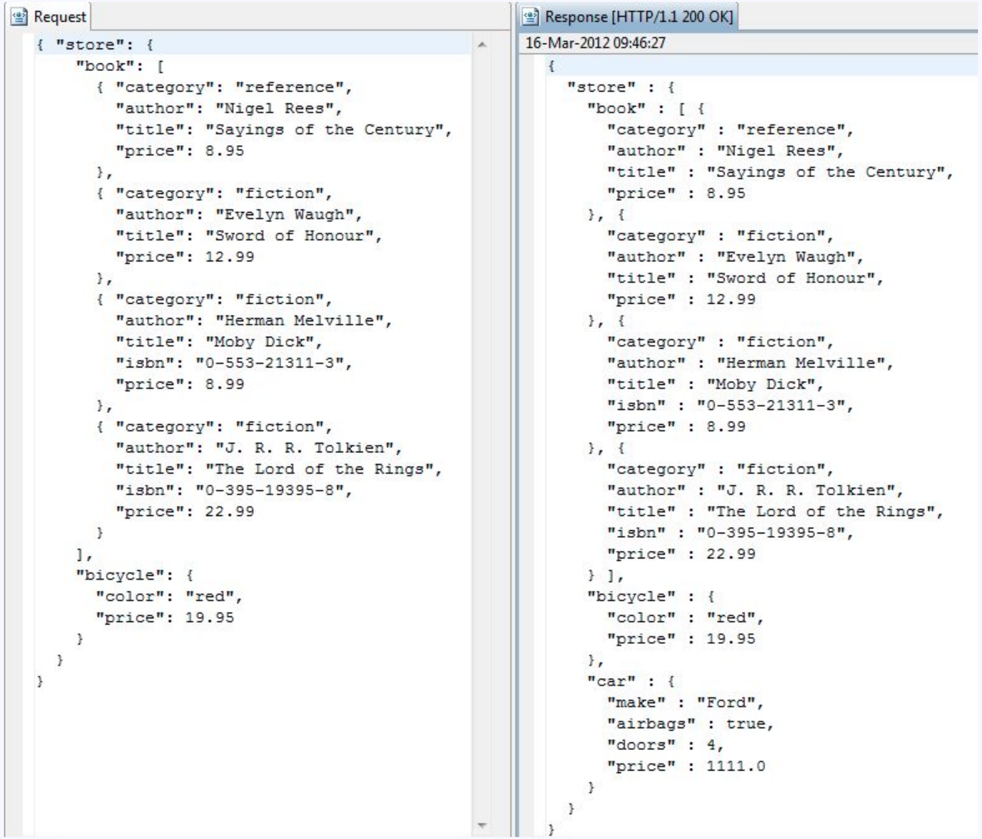
In what form do we store Data?

Structured? Unstructured?

Text? Photo? Video?

JSON - JavaScript Object Notation

- Supported Data Types:
  - Number
  - Boolean
  - String
  - Array
  - Object (Dictionary)
- These are known as *serializable* types.



```
{ "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
    }
  }
}
```

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 ‘Geoffrey Challen’”
- “Exploration”
  - Drill down into fields / entries, do simple visualizations, grasp structure of dataset

Jupyter Time

https://github.com/cs196illinois/data_hackerspace

Questions? AMA