DB Management Systems
Getting Data

Joel Klein – jdk514@gwmail.gwu.edu

Overview

- Where does data come from?
- What formats does this data take?
- Dataset for this course

Data Sources

Where Do We Get Data?

- Data can come from several sources, but is typically sourced from the following:
  - Colleagues
  - Clients/Customers
  - API's
  - Sensors
  - Online Collections

- Real World Examples??

Data Formats

Structured Data

- This is the type of data typically used when first learning about data science, and what is found in SQL-like databases
  - Think Pandas Dataframes

- Here structured means:
  - The information conforms to a set data-model

- This is great for learning the ropes of analyzing data, but it is not the format in which we usually receive data

Layout Example:

Col 1	Col 2	Col 3	Col 4
Type Int	**Type Str**	**…**	**…**
Int 1	Str 1	…	…
Int 2	Str 2	…	…
Int 3	Str 3	…	…

Data Example:

Name	Date	Genre	MPAA
Interstellar	Oct 2014	Sci-Fi	PG-13
…	…	…	…

Typical formats: CSV, Excel, SQL DBs

Semi-Structured

- A large amount of online data is transmitted in this form
- This type of data has enforced rules (e.g. data-types, hierarchy, etc.), but is not as restrictive (e.g. recursive objects)
- JSON and XML are two common formats for semi-structured data
- The benefits to this format is that it can convey more nuanced information and modified on the fly, but at the cost of certain guarantees (what features exist, defaults, etc.)

Sample JSON:

```
{
   "Name": "Interstellar",
   “Release Date": Oct 2017,
   “Genres": [
          “Science Fiction”,
          “Drama”,
   ]
}
```

JSON

- JSON stands for **J**ava**S**cript **O**bject **N**otation
    - It started out primarily as a structure for data used in AJAX (Asynchronous Javascript and XML) calls
    - Currently the predominant method for sharing data online

- JSON is comprised of JSON arrays and objects
    - These are effectively a 1:1 with python lists and dictionaries
    - These elements can be infinitely nested

JSON Data Types
String: “Example”
Integer: 10
Float: 1.34
Object: {more_data}
Array: [data1, data2, etc.]
Boolean: true or false
Null: null

Unstructured Data

- Any data without a semblance of structure, typically data designed for the consumption of people and not machines

- Typical formats include:
  - Text
  - Audio
  - Video

- This format of data usually requires intensive ETL/ML algorithms to transform the data into a machine usable state

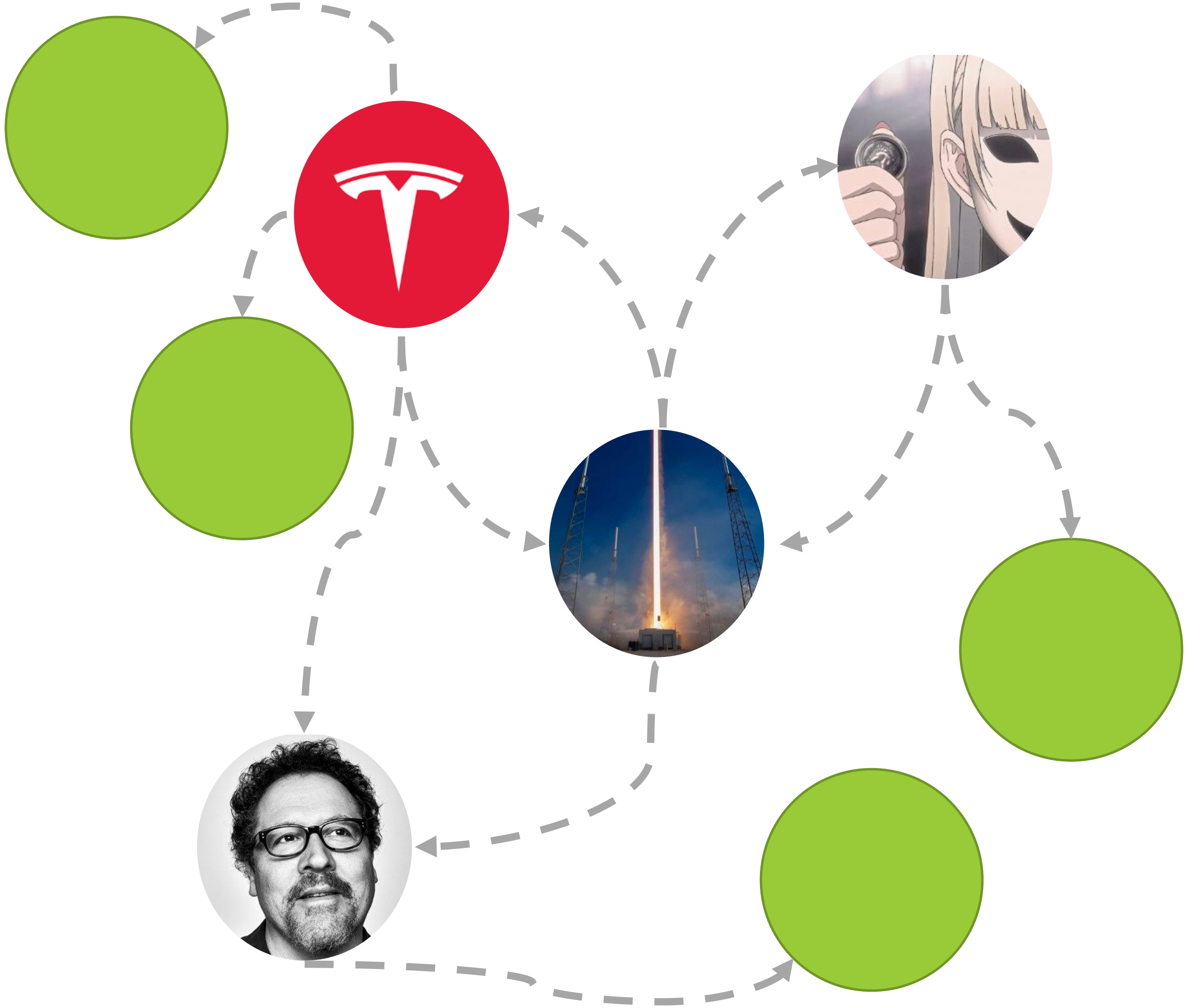
Course Data

Data for the Course

- For this course we will be using data that has been pulled down from twitter using their API
    - ▫ Given that it has been pulled from the twitter API, what format do you think it takes?

- This data focuses on Elon Musk and expands outward through his network of twitter friends, favorites, statuses, etc.

Elon’s Friend Network



Data Acquisition

```
function get_twitter_data(user):
        •   get_friends
        •   get_favorites
        •   get_lists
        •   get_statuses
        •   get_retweets
        •   return data

trumps_friends = get_friends(trump)
for friend in trumps_friends:
        •   data = get_twitter_data(friend)
        •   secondary_friends.add(data[‘friends’])

for friend in secondary_friends:
        •   data = get_twitter_data(friend)
```

End Slide

EMSE 6992 – DBMS for Data Analytics