

# Databases & SpreadSheets 101

Christan Grant, Ph.D.

Arnold and Lisa Goldberg Rising Star Associate Professor

cgrant@cise.ufl.edu — <https://ceg.me>

 [ufdatastudio.com](https://ufdatastudio.com) 



# Christan Grant, Ph.D.

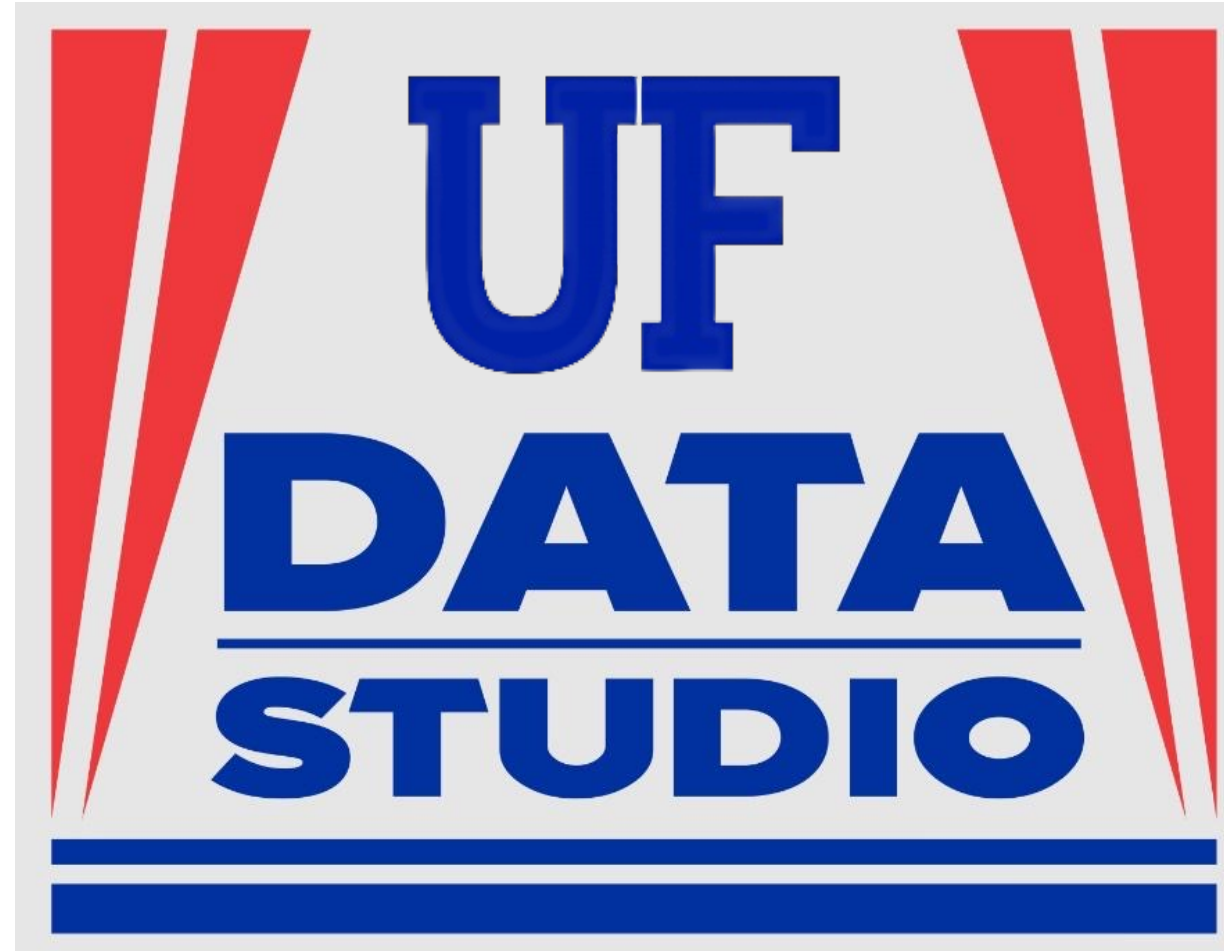
- Completed my doctorate in UF CISE 2015.
- Associate Professor of Computer Science at the University of Oklahoma.





# UF Data Studio

We are passionate about developing new ways humans can interact with their data and understanding the [fundamental research issues all along the data pipeline.](#)



# Outline

- Spreadsheets
- Databases
  - Database Management Systems
- Data Sets
- Data Sheets



☰ [C S-4|5113-001](#) > [Grades](#)

Gradebook ▾ [View](#) ▾ [Actions](#) ▾

Student N Import

🔍 Sea **Export Current Gradebook View**

# Spreadsheets

- Exported spreadsheets go to CSVs to use other software.
- All software reads .csv files.
- Other storage formats include .tsv, .json

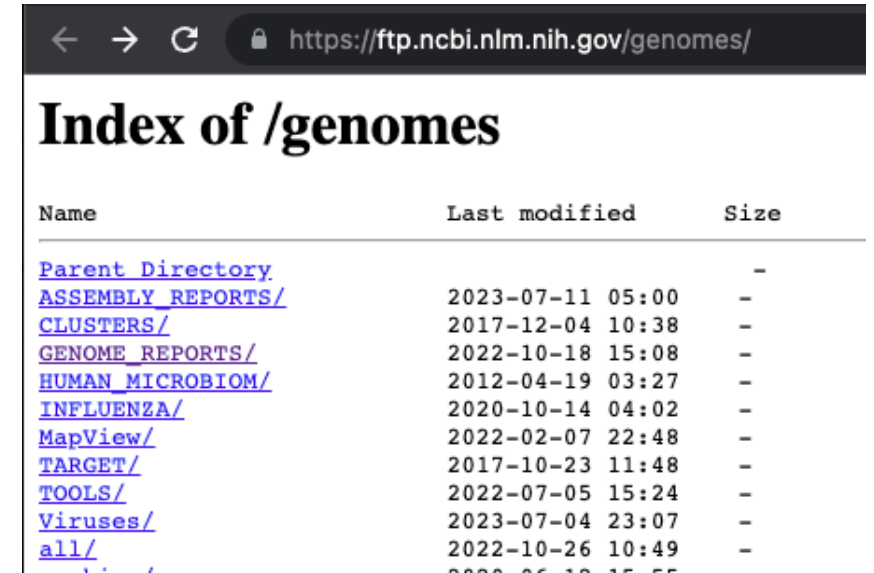
```
{  
  "employee": {  
    "name": "sonoo",  
    "salary": 56000,  
    "married": true  
  }  
}
```

JSON – Java Script  
Object Notation

# Databases

- Collections of data related to a topic.
- Databases are structured, often with many spreadsheets, sometimes referred to as *Tables*.
- It may contain a “code book” or an interpretation of the spreadsheet data.

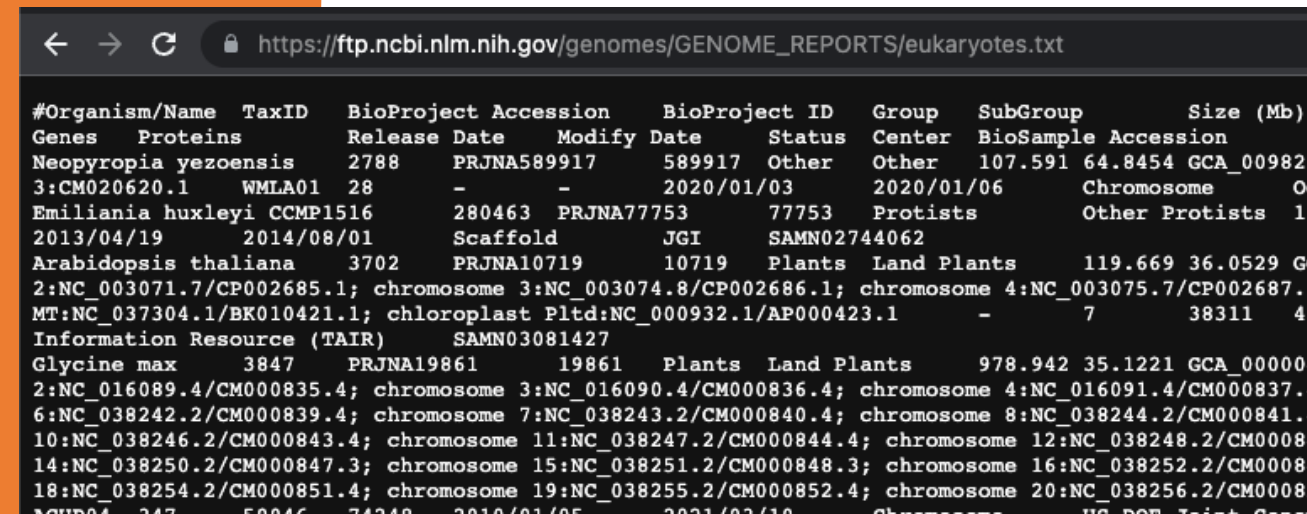
GenBank<sup>®</sup> is the NIH genetic sequence database:



https://ftp.ncbi.nlm.nih.gov/genomes/

### Index of /genomes

| Name                              | Last modified    | Size |
|-----------------------------------|------------------|------|
| <a href="#">Parent Directory</a>  |                  | -    |
| <a href="#">ASSEMBLY_REPORTS/</a> | 2023-07-11 05:00 | -    |
| <a href="#">CLUSTERS/</a>         | 2017-12-04 10:38 | -    |
| <a href="#">GENOME_REPORTS/</a>   | 2022-10-18 15:08 | -    |
| <a href="#">HUMAN_MICROBIOM/</a>  | 2012-04-19 03:27 | -    |
| <a href="#">INFLUENZA/</a>        | 2020-10-14 04:02 | -    |
| <a href="#">MapView/</a>          | 2022-02-07 22:48 | -    |
| <a href="#">TARGET/</a>           | 2017-10-23 11:48 | -    |
| <a href="#">TOOLS/</a>            | 2022-07-05 15:24 | -    |
| <a href="#">Viruses/</a>          | 2023-07-04 23:07 | -    |
| <a href="#">all/</a>              | 2022-10-26 10:49 | -    |



https://ftp.ncbi.nlm.nih.gov/genomes/GENOME\_REPORTS/eukaryotes.txt

| #Organism/Name       | TaxID    | BioProject  | Accession  | BioProject ID | Group       | SubGroup | Size (Mb) |
|----------------------|----------|-------------|------------|---------------|-------------|----------|-----------|
| Neopyropia yezoensis | 2788     | PRJNA589917 | 589917     | Other         | Other       | 107.591  | 64.8454   |
| Emiliana huxleyi     | CCMP1516 | 280463      | PRJNA77753 | 77753         | Protists    |          |           |
| Arabidopsis thaliana | 3702     | PRJNA10719  | 10719      | Plants        | Land Plants | 119.669  | 36.0529   |
| Glycine max          | 3847     | PRJNA19861  | 19861      | Plants        | Land Plants | 978.942  | 35.1221   |

# Data Base Management Systems (DBMS)

- Software to manage the storage and access to databases.
- The data is stored and organized in an efficient binary format (1s and 0s).
- Logically, data are stored as “Tables”.
- Data is accessed through a **Structured Query Language (SQL)**.
- There are many databases types and companies.
- Demo time! <https://sqliteonline.com/>



# Google Sheets (GQL Queries)

Tutorial: <https://spreadsheetpoint.com/google-sheets-query-function/>

H1    -    fx    =QUERY(Dataset, "SELECT \* WHERE B='Manufacturing' ORDER BY E DESC", 1)

|   | G | H              | I             | J        | K  | L            | M           |
|---|---|----------------|---------------|----------|--|--------------|-------------|
| 1 |   | Employee Name  | Department    | DOB      | Address  | Hours Worked | Hourly Rate |
| 2 |   | Miranda Mathew | Manufacturing | 08/01/00 | 9192 10th Avenue<br>Hopewell Junction, NY<br>12533 | 45           | 30          |
| 3 |   | John Leon      | Manufacturing | 05/01/00 | 8412 Pine Rd. Taunton,<br>MA 02780                 | 32           | 20          |
| 4 |   | Cierra Vega    | Manufacturing | 01/08/02 | 941 Bowman Lane<br>Englewood, NJ 07631             | 25           | 15          |

# Outline

- Spreadsheets
- Databases
  - Database Management Systems
- **Data Sets**
- Data Sheets

# Data Sets

<https://www.kaggle.com/datasets/felipeesc/shark-attack-dataset>

- In ML and AI, it isn't enough to just provide a set of data.
- The data are split into {training, validation, testing}

The screenshot shows the Kaggle Datasets interface. At the top, there's a search bar with 'sharks' entered and a 'Filters' button. Below the search bar, there are category filters: 'All datasets', 'Computer Science', 'Education', 'Classification', 'Computer Vision', 'NLP', and 'Data Visualization'. A 'Pre-Trained Model' filter is also visible. The main content area shows a list of 33 datasets. The first four datasets are related to 'Shark Tank India':

| Dataset Name               | Author           | Updated             | Usability | Files        | Size | Rank | Badge  |
|----------------------------|------------------|---------------------|-----------|--------------|------|------|--------|
| Shark Tank India Companies | Devanshu Ramaiya | Updated a year ago  | 10.0      | 1 File (CSV) | 3 kB | 31   | Bronze |
| Shark-Tank-India           | Anshul Mehta     | Updated a year ago  | 10.0      | 1 File (CSV) | 3 kB | 77   | Silver |
| Shark Tank India Dataset   | Shiva Vashishtha | Updated a year ago  | 10.0      | 1 File (CSV) | 4 kB | 86   | Bronze |
| Shark attack dataset       | Felipe_Esc       | Updated 2 years ago | 9.7       | 2 Files      |      | 46   | Bronze |

Kaggle Datasets

The screenshot shows the PyTorch Hub website. The header includes the PyTorch logo and the text 'PYTORCH HUB FOR RESEARCHERS'. Below the header, there's a navigation bar with categories: 'All', 'Audio', 'Generative', 'Nlp', 'Scriptable', and 'Vision' (which is highlighted). A search bar is also present. The main content area features a featured model card for 'YOLOv5' with 40,000 likes and a description: 'Ultralytics YOLOv5 for object detection, instance segmentation and image classification.' To the right of the card is a promotional image for 'ultralytics YOLOv5'.

PyTorch Hub

# Data Sheets

## Datasheets for Datasets

Timnit Gebru<sup>1</sup> Jamie Morgenstern<sup>2</sup> Briana Vecchione<sup>3</sup> Jennifer Wortman Vaughan<sup>1</sup> Hanna Wallach<sup>1</sup>  
Hal Daumé III<sup>1,4</sup> Kate Crawford<sup>1,5</sup>

### Movie Review Polarity

### Thumbs Up? Sentiment Classification using Machine Learning Techniques

#### Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.<sup>1</sup>

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

**Any other comments?**

None.

#### Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them: nodes and edges)? Please provide a description.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up \* non \* -ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example “negative polarity” instance, taken from the file neg/cv452\_tok-18656.txt.

exception that no more than 40 posts by a single author were included (see “Collection Process” below). No tests were run to determine representativeness.

**What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in “Data Preprocessing”).

**Is there a label or target associated with each instance?** If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived



Content



Process



Experience



Fairness in Machine Learning



Privacy



Thank you!

[ufdatastudio.com](http://ufdatastudio.com)

[cgrant@cise.ufl.edu](mailto:cgrant@cise.ufl.edu)



# Bonus

Simple ML for Google Sheets

[sheets.google.com](https://sheets.google.com)