

# COMP6940: BIG DATA AND DATA VISUALISATION

LECTURE #1: COURSE INTRODUCTION AND OVERVIEW

Inzamam Rahaman



# PREFACE

- Course Lecturers:
  - Dr. Kris Manohar (Kris.Manohar@sta.uwi.edu)
  - Mr. Inzamam Rahaman (Inzamam.Rahaman@outlook.com)
- Online Platform: myElearning + YouTube
- Course times: Wednesday 17:00 – 20:00 (GMT-4)
- Course Programming Language: Python (maybe some Scala too)

# AGENDA

1. Course Motivation and Expectations
2. Overview/Summary of content schedule
3. Evaluation Structure
4. Basic introduction to big data concepts
  1. What is Big data?
  2. What is data mining?

# COURSE MOTIVATION

- Over 2 Petabytes of data are generated daily across the globe
  - Financial transactions, tweets, social media posts, video uploads, digital traces, etc...
- Data can lead to powerful and useful insights!
- But first, data must be:
  - Stored
  - Managed
  - Analyzed
  - However...

## DATA IS THE NEW OIL OF THE DIGITAL ECONOMY



Image: verifex/Flickr

Data in the 21st Century is like Oil in the 18th Century: an immensely, untapped valuable asset. Like oil, for those who see Data's fundamental value and learn to extract and use it there will be huge rewards.

# COURSE MOTIVATION

- Extracting usable products and insights from data is difficult engineering challenge
- Large volumes of data “break” some traditional assumptions
- Data storytelling is difficult
- Course objectives in a nutshell:
  - Help equip you with key intuitions and foundational knowledge to tackle these problems



# COURSE EXPECTATIONS



Data science is a rapidly developing field



Best approaches, essential methods, and standards are still in development

Like pinning Jell-O to a wall



We focus on core technologies/approaches to build sound foundation + give a taste of current trends/developments



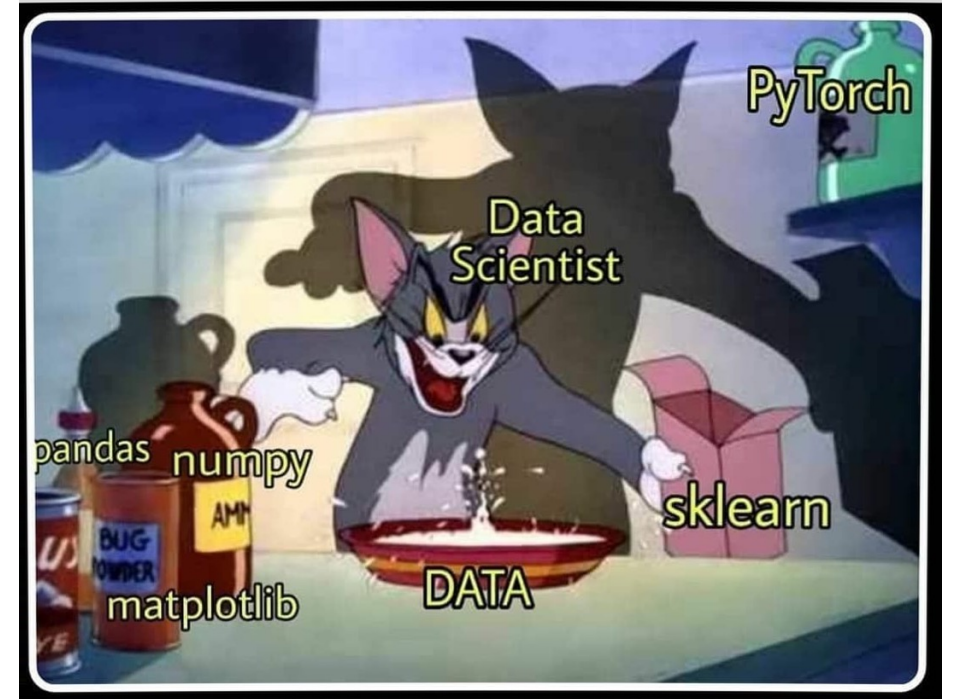
Is a graduate level CS course:

Lot of self-study and research  
Expect foundational programming ability (in Python)  
Expect mathematical maturity of at least undergrad STEM degree



# TECHNOLOGY STACK

- As aforementioned, we expect you to write non-trivial programmes in Python 3.10 or above
  - Suggest installing Anaconda distribution for Python
- In addition, we will also cover tools such as Git
- We recommend using LaTeX for reports through Overleaf
- We suggest using a Unix environment for the course
  - MacOS and Linux distro as already compliant
  - Install WSL on Windows (recommend dual-booting with a Linux distro such as Ubuntu)
- Suggest PyCharm, VS Code, or Data Grip as IDEs for the course
  - Free access to PyCharm Professional and Data Grip through JetBrains student programme
  - <https://www.jetbrains.com/community/education/>
- Remember, this is a graduate level CS course, we are expecting the attendant technological maturity
- Side note: slides use the Goldman Sans font (<https://design.gs.com/foundation/typography/goldman-sans>)





Data, raw and unrefined, A treasure trove, but hard to find. But with visualization as our guide, Insights come, and secrets bide. From patterns hidden deep, To trends on the surface, we'll keep. Big data's value, now clear to see, Thanks to visualization's key. It helps us make sense, Of this complex dense. And with it, we can tell a story, Of data's true worth and its glory.

- ChatGPT

## Course Learning Outcomes

On successful completion of the course, students will be able to:

1. Describe the advantages and limitations of GPUs vs. CPUs related to scaling data analysis from small to large data sets.
2. Explain strategies for big data management in an organization.
3. Extract valuable information by processing large datasets.
4. Use scalable and GPU accelerated frameworks (e.g., Spark, RAPIDS) to analyze large streams of data.
5. Create scalable graphical solutions to specific problems with standard APIs and tools.
6. Build predictive systems that rely on large datasets.
7. Apply data-parallel decomposition to reduce the runtime of machine learning algorithms.
8. Execute basic machine learning algorithms (e.g., clustering algorithms, neural networks) at scale.
9. Collaborate on a big data project using an appropriate version control software (e.g., Git).

| Week | Topic  | Required Reading Learning Resources | Learning Activities | Assignments  |                |
|------|--|-------------------------------------|---------------------|--|----------------|
|      |  |                                     |                     | Name   | Due Date       |
| 1.   | Introduction   | Lecture Notes                       | In-class activities | Ejournal #1  |                |
| 2.   | Data collection and exploratory data analysis  | Lecture Notes                       | In-class activities | <ul style="list-style-type: none"> <li>Asynchronous discussion forum topic #1 provided</li> <li>Assignment #1 provided</li> </ul>              | Due in week 5  |
| 3.   | Data ethics, integration and reducing bias in data sets                                    | Lecture Notes                       | In-class activities |  |                |
| 4.   | Data analytics, concepts and tasks   | Lecture Notes                       | In-class activities | Assignment # 1 code review by peers  | Due in week 4  |
| 5.   | Principles of data visualization   | Lecture Notes                       | In-class activities | Assignment #2 provided   | Due in week 8  |
| 6.   | Data-parallel decomposition strategies (e.g. map reduce).                                  | Lecture Notes                       | In-class activities |  |                |
| 7.   | Introduction to distribute frameworks and common mistakes to avoid with big data           | Lecture Notes                       | In-class activities | <ul style="list-style-type: none"> <li>Assignment # 2 code review by peers</li> <li>Asynchronous discussion forum topic #2 provided</li> </ul> | Due in week 7  |
| 8.   | Collaboration  | Lecture Notes                       | In-class activities | Project provided   | Due in week 13 |
| 9.   | CPU vs GPU Acceleration, Classification  | Lecture Notes                       | In-class activities | Code check-in with TA #1   | Due in week 9  |
| 10   | Clustering, principal component analysis, and least discriminant analysis, neural networks | Lecture Notes                       | In-class activities | Code check-in with TA #2   | Due in week 10 |
| 11   | Text analysis, streaming data  | Lecture Notes                       | In-class activities | Peer reviews of draft presentations  | Due in week 11 |
| 12   | Course Review<br>(No introduction of new subject   | Lecture Notes                       |                     | Project code due   | Due in week 12 |







| # | Assessment Type            | Learning Outcomes<br>(see section above) | Assessment  |          |                   |                     |
|---|----------------------------|--|-------------|----------|-------------------|---------------------|
|   |                            |  | Total Marks | Weight % | Description       | Duration (in weeks) |
| 1 | Assignment #1              | 3, 4 & 5                                 | 40          | 20       | See section above | 3                   |
| 2 | Assignment #2              | 6, 7 & 8                                 | 40          | 20       | See section above | 3                   |
| 3 | E-journals,<br>forum posts | 1 & 2                                    | 5           | 12       | See section above | 1 each              |
| 4 | Project                    | 2, 3, 4, 5, 6, 7,<br>8 & 9               | 100         | 50       | See section above | 5                   |

# BIG DATA

- Rate of data production across all industries and human activities is rapidly increasing every year
- Data generated from multiple sources
- Rapidly accumulated is large and complex
- Constellation of integrated traits emerges
  - The 6 V's of Big Data

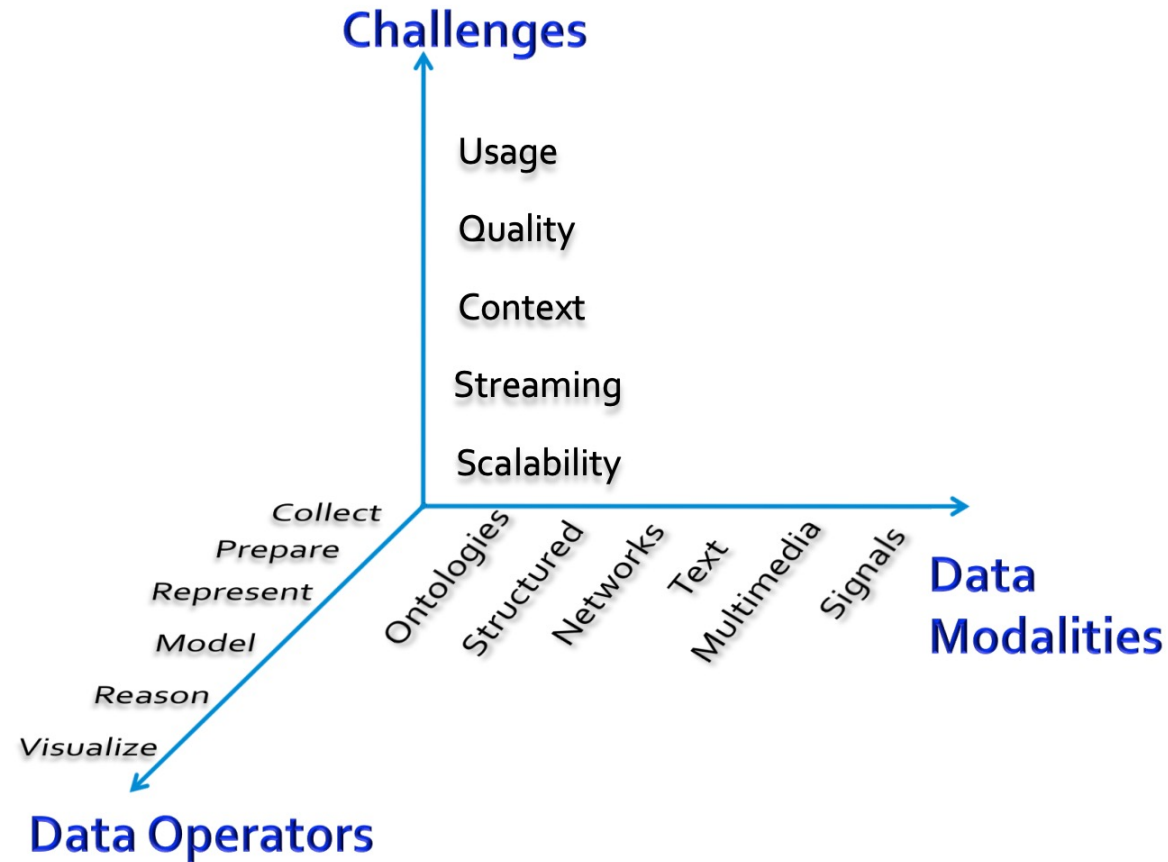
# The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume*, *variety* and *velocity*. Over time, other Vs have been added to descriptions of big data:

| VOLUME   | VARIETY   | VELOCITY   | VERACITY   | VALUE  | VARIABILITY   |
|--|---|--|--|--|---|
| The amount of data from myriad sources.  | The types of data: structured, semi-structured, unstructured.                       | The speed at which big data is generated.  | The degree to which big data can be trusted.   | The business value of the data collected.  | The ways in which the big data can be used and formatted.                             |
|  |  |  |  |  |  |



# CONSIDERATIONS WHEN WORKING WITH BIG DATA



# BIG DATA – DATA CLEANING

- Data from source systems are almost always messy and legion with issues that hamper correct usage
- Form validation didn't exist as we know it today in the 70's
- Inconsistent and erroneous formats, e.g. DDMMYY and MMDDYY in the same data field
- Would discuss more on data cleaning next lecture



# BIG DATA – ETL

- Extract: data is pulled from required source systems and databases
- Transform: data is cleaned and otherwise transformed from source systems
- Load: transformed data is loaded into format or storage amenable to analysis and additional processing for data mining (next slide)
  - Files (e.g. Feather, CSV, etc...)
  - Data Warehouses/Data Lakes (e.g. Snowflake)
- Several tools for ETL (Apache Airflow, SnapLogic)
- In reverse ETL, data is piped from the load area back to the source areas from which the data is extracted



# BIG DATA – DATA MINING

- Goal of data mining is to extract patterns from large quantities of data and build models that allow us to make predictions of the data
- Data Mining can be split into
  - Descriptive: characterize patterns in the data, e.g. clustering for customer segments, stream processing for computing live statistics
  - Predictive: develop models to predict unknown or future values of some variable, e.g. predicting risk of churn, recommender systems to recommend products
- At **best** data mining produces results that are:
  - Valid: patterns and models are representative of current and future state respectively
  - Useful/Actionable: possible to use descriptive or predictive insights
  - Unexpected/Interesting: results are non-trivial
  - Understandable/Explainable: results can be interpreted and understood by human beings
- Will discuss some good practices in detail in later lectures

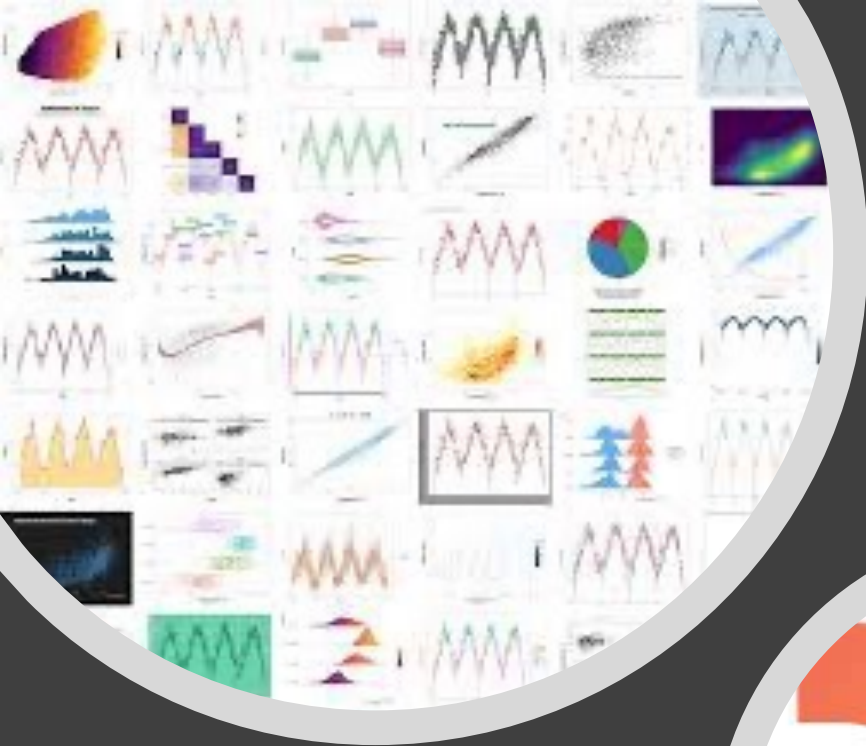
# BIG DATA – DATA MINING

- Suppose we want to increase sales. Number of sales would be KPI
  - We want to increase this KPI
- One way would be more tailored marketing strategies to most valuable segments (i.e. lever)
- Data Mining can help us determine segments from dataset to support and enable business lever
- Data Mining is essential tool in digital transformation
- Good data science, often involves hypotheses – don't just throw random modelling techniques at data and see what sticks

# DATA VISUALIZATION

- Results from data analysis would need to be communicated to stakeholders or fellow researchers
- Data visualization is concerned with pictographically and diagrammatically representing important trends in the data
  - Often with charts that represent the relationships between different quantities in the data
- Good visualizations:
  - Correct: correctly encode the data from the dataset
  - Honest: are created without the intention of misleading the audience
  - Visually appealing: should not use garish and distracting palletes or organization





# DATA VISUALIZATION

- Many different chart types for data viz:
  - Pie, Line, Box plot, etc....
- Visualizations can be standalone or can be embedded in a dashboard to assist in decision making
- Tools for creating visualizations:
  - JavaScript – D3
  - R – ggplot
  - Python – Matplotlib, Seaborn, Plotnone (ggplot in Python)
  - Dashboards – Tableau, PowerBI

# BUILDING BLOCKS OF DATA MINING PRACTICE

Collection

Cleaning

Integration

EDA/CDA

Visualization

Modelling

Presentation/Dissemination

# DATAFRAMES

- Many datasets would be either natively tabular or can be massaged into being tabular
- Most ML/DS libraries expect tabular formats
- Dataframes are common abstract data type for working with data in a tabular format
- Many different implementations:
  - Pandas
  - Polars
  - Dask
- Offer:
  - Data indexing
  - Data selection and filtering
  - Creation and update of columns
  - Etc...

The diagram illustrates a DataFrame structure. At the top, the word "Columns" is written in blue, with blue arrows pointing to the column headers: "Name", "Team", "Number", "Position", and "Age". On the left, the word "Rows" is written in orange, with orange arrows pointing to the row indices 0, 1, 2, 3, 4, 5, and 6. A pink box labeled "Data" in pink text is positioned at the bottom right, with pink lines connecting it to the data cells of the table, specifically highlighting the values "8.0", "NaN", "12.0", "C", and "NaN".

|   | Name            | Team           | Number | Position | Age  |
|---|-----------------|----------------|--------|----------|------|
| 0 | Avery Bradley   | Boston Celtics | 0.0    | PG       | 25.0 |
| 1 | John Holland    | Boston Celtics | 30.0   | SG       | 27.0 |
| 2 | Jonas Jerebko   | Boston Celtics | 8.0    | PF       | 29.0 |
| 3 | Jordan Mickey   | Boston Celtics | NaN    | PF       | 21.0 |
| 4 | Terry Rozier    | Boston Celtics | 12.0   | PG       | 22.0 |
| 5 | Jared Sullinger | Boston Celtics | 7.0    | C        | NaN  |
| 6 | Evan Turner     | Boston Celtics | 11.0   | SG       | 27.0 |

OG

# SELF DEVELOPMENT

- Main scientific conferences/journals: KDD, WebConf, ICDM, IEEE BigData, ASONAM, CHI, RecSys, TKDE, etc...
- TowardsDataScience: medium blog channel. Lots of great ideas – also loads of nonsense. Always read critically
- Reddit – r/datascience and r/machinelearning are pretty good places to learn of new technologies and approaches

# E-JOURNAL #1

- Graphs are becoming increasingly important data structures and abstractions in data mining and big data, thereby leading to the sub-field of graph mining. Research and reflect on this trend and its implications
- Due next week Friday @ 11:59 PM (AoE)

QUESTIONS?



# IMAGE CREDITS

1. SLIDE 5: <https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/>
2. SLIDE 8 & 15: r/DataScienceMememes
3. SLIDE 13: <https://www.quora.com/What-are-the-six-Vs-of-Big-Data>
4. SLIDE 14: <http://www.mmnds.org/mmnds/v2.1/ch01-intro.pdf>
5. SLIDE 16b: <https://siliconangle.com/2017/02/15/snaplogic-expands-cloud-app-kafka-real-time-support-new-release/>
6. Slide 22: <https://www.geeksforgeeks.org/python-pandas-dataframe/>