**Abstract**

This project aims to develop a model to accurately predict weather patterns based on historical weather and climate data. A recurrent neural network (RNN) will be leveraged to analyse the sequential and pattern based nature of the weather to forecast real time weather.

[**https://github.com/ljn1999/Traffic-Sign-Detection/blob/master/Report/APS360-Sec01-Team02%20Project%20Proposal.pdf**](https://github.com/ljn1999/Traffic-Sign-Detection/blob/master/Report/APS360-Sec01-Team02%20Project%20Proposal.pdf)

[**https://github.com/tirthajyoti/Deep-learning-with-Python/blob/master/Notebooks/Weather-RNN.ipynb**](https://github.com/tirthajyoti/Deep-learning-with-Python/blob/master/Notebooks/Weather-RNN.ipynb)

* [**https://medium.com/analytics-vidhya/weather-forecasting-with-recurrent-neural-networks-1eaa057d70c3**](https://medium.com/analytics-vidhya/weather-forecasting-with-recurrent-neural-networks-1eaa057d70c3)
* [**https://www.kaggle.com/code/rajatdey/neural-model-for-weather-prediction**](https://www.kaggle.com/code/rajatdey/neural-model-for-weather-prediction)

**\*\* for weather resource:**[**https://www.youtube.com/watch?v=km95-NMT6lU**](https://www.youtube.com/watch?v=km95-NMT6lU)

**Imma watch this:** [**https://www.youtube.com/watch?v=baqxBO4PhI8**](https://www.youtube.com/watch?v=baqxBO4PhI8)

**Completely (ez) different approach (picture of outside → predict cloudy/sunny etc.):** [**https://github.com/berkgulay/weather-prediction-from-image**](https://github.com/berkgulay/weather-prediction-from-image)

**Article + Video + Github:** [**https://medium.com/analytics-vidhya/forecast-weather-using-python-e6f5519dc3c1**](https://medium.com/analytics-vidhya/forecast-weather-using-python-e6f5519dc3c1)

**Google’s Weather Forecast Tutorial:**

[**https://cloud.google.com/blog/topics/sustainability/weather-prediction-with-ai**](https://cloud.google.com/blog/topics/sustainability/weather-prediction-with-ai)

**Introduction (Andrew)**

Motivations - personalise it

* + Weather forecast and prediction is extremely important and something that everyone uses everyday in some form or another
    - Used in agriculture, energy, transportation etc

Goals

* + The goal for the project is to develop a machine-learning model that can accurately predict the future weather using the climate and recent weather of an area

Importance

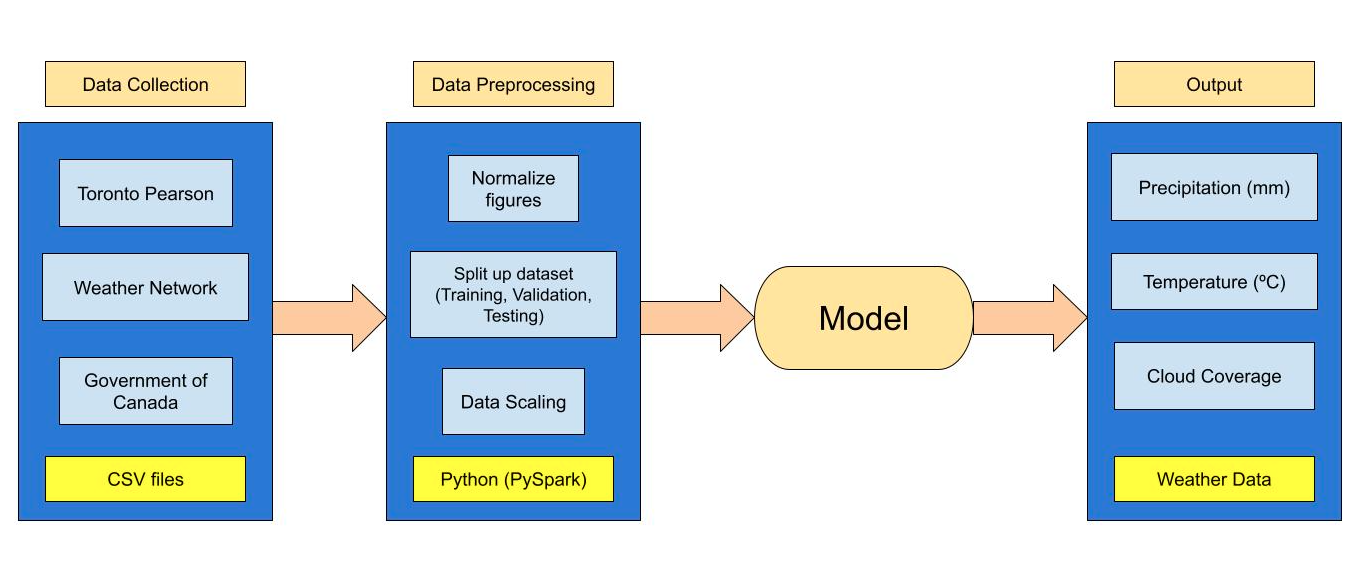
* + Kinda saying the same thing as motivations - just group it
    - Weather is extremely complicated, and is used in so many different fields
    - Using a deep learning model will increase precision and reliability of predictions

Reasonable approach

* + Deep learning is a reasonable approach because the weather is a highly pattern-based phenomenon
    - Models like RNN and CNN are good at detecting patterns
      * Especially RNN for large datasets
    - The Neural Networks are very good at extracting relevant pieces of information from large datasets

**Illustration/Figure (Matthew)**

The Backbone + multi branch/head structure front [this](https://openaccess.thecvf.com/content/CVPR2021W/CVFAD/papers/Parekh_Fine-Grained_Visual_Attribute_Extraction_From_Fashion_Wear_CVPRW_2021_paper.pdf) article, ++ also, it has great visuals, for model schematic

****

**Background & Related Work (Mitchell)**

Since the Advent of computer modelling weather forecasting systems have been in use to provide weather event predictions. Historically, forecasting approaches may be broken into two groups, traditional numerical weather prediction (NWP) systems and artificial intelligence (AI) models. NWP systems are comprised of hard coded physics-based equations modelling physical processes with dynamic multi-variable interactions [[xx](https://www.marktechpost.com/2023/03/01/a-study-on-various-deep-learning-based-weather-forecasting-models/)]. These NWP systems would require extreme amounts of computing power and time to process the results, taking an hour to process the same information that current AI models take one second to process [[xx](https://ai.googleblog.com/2021/11/metnet-2-deep-learning-for-12-hour.html)]. Modern AI solutions include, but are not limited to, Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Units (GRU) and Transformers [[xx](https://arxiv.org/pdf/2204.11115.pdf)]. RNNs are models which use back propagation, a technique that produces outputs as the sigmoid of not only current but also past inputs[[xx](https://www.ibm.com/topics/recurrent-neural-networks)] or [[xx](https://arxiv.org/pdf/2204.11115.pdf)]. LSTM and GRU models are built upon a RNN framework, but implement additional variables that work by learning to “forget” information that is not useful for future predictions, and using hyperbolic tangential activation function instead of the sigmoid[[xx](https://www.ibm.com/topics/recurrent-neural-networks)] or [[xx](https://arxiv.org/pdf/2204.11115.pdf)]. Lastly, recent developments with attentional Transformer models enable the system to select the most important information across prior time steps, not just consolidating all preceding values, or selectively dropping the least useful pieces of information[[xx](https://towardsdatascience.com/attention-for-time-series-classification-and-forecasting-261723e0006d)].

**^^^^** Rough + cited, outlining functioning of best pre-Ai model, as well as 4 common models used for Time Series Forecasting, one being RNN – the one we may do..

**Data Processing (Liza)**

* [https://climate.weather.gc.ca/historical\_data/search\_historic\_data\_e.htm](https://climate.weather.gc.ca/historical_data/search_historic_data_e.html)l
* Note for team: Apache Spark, which is a distributed computing system designed to handle large-scale data processing. PySpark allows you to perform data-cleaning operations efficiently and in parallel across distributed clusters.
* [Pyspark in Google Colab](https://towardsdatascience.com/pyspark-in-google-colab-6821c2faf41c#:~:text=To%20run%20spark%20in%20Colab,Jupyter%20Notebook%20of%20the%20Colab.&text=Our%20Colab%20is%20ready%20to%20run%20PySpark.)

The Government of Canada, Weather Network and Toronto Pearson have amassed hourly weather data for many specific regions in Canada over the course of years. The team will collect the weather data for Toronto over the course of years, directly from their respective historical data portal, in a .CSV file format. When data is stored and the directory is mounted to the Google Colab the team will clean and filter data for relevant metric and non-null fields. To format the data into a usable state, the team will use the PySpark Library (Python API for Apache Spark) to query the data and generate a complete dataset that can be used to train and validate our model. Formatting would not only include only pulling relevant fields, but it would also require ensuring the fields are cast to the correct datatype (i.e. datetime, int, string) and any spacing issues and special characters are dealt with. Developing a query/script to clean our data will allow us to easily scale our dataset. We will be able to source more data from these sources, in their standard template, and clean/format the data quickly and efficiently.

**Architecture (Andrew)**

* Probably just use RNN? - dont actually fully know what it is yet but online seems to think it would be the best
  + RNN best suited for sequential data processing []
  + RNN can take into account the sequential nature of weather to learn weather patterns
* No idea what the hyperparameters would be

**Baseline Model (Andrew)**

* Persistence model
  + A persistence model just takes exactly what the current weather is and predicts the future to be the same
    - This made seem overly simplistic but if this turns out to be more accurate than our model then we know that our model did not learn anything useful
      * Establishes a minimum for our model
    - If it does better than we know that it did pick up some useful weather patterns

**Ethical Considerations (Mitchell)**

**1:**

Four main tenets of ethical AI weather forecasting models regard biased in/output, data handling, model application, and overall transparency. Limitations include data collection resources and natural anomalies.The first ethical concern is surrounding biased input and output, as models are known to understand complex environmental dynamics through the data, it may also be able to learn biases built into the data and its uses. Following this is data handling, from collection to storing and usage, these models require many types of information from various services, all which are extremely sensitive and must be handled with utmost care to prevent any unintended consequences. Beyond improper data usage is inappropriate model usage, as simulations are becoming more accurate and trustworthy, they can be used to reverse engineer a desired outcome. Finally, the last ethical concern is with respect to equitable access to the information, as its necessary for the information processing and output to be made understandable in order to maintain

Or

2:

The four key ethical issues posed by the proposed forecasting model include biases, data handling, model application, and overall transparency in the creation and operation of the model. Biased output can result from biased input, specifically when training on climate data from regions different from those where the model is applied. This will be remedied by training and applying the model strictly within the greater Toronto area. The next issue regards proper handling of data collection, use, and storage. To avoid mishandling all training and inference data used will be collected from publicly available databases provided by the govt of canada. Additionally, model misuse can possibly unlock potentials to use the information produced against groups and regions. Due to the nature of this project, the model is not using state of the art technology, along with only being shared among team 0 and APS360 staff, restricting its overall potential for use and therefore misuse. Lastly, transparency regarding the creation and operation is a necessary element when regarding technology and the environment. This paper, along with all following course documentation will provide reasonable detail to the models processes, and its narrow application and lightweight construction make for a minimal impact on technological resources that could impact the weather. This is in addition to the existing efforts done to reduce the ecological footprint of AI and technology, the details of which extend beyond the scope of this paper.

To be referenced in above^

[Artificial intelligence and climate change: ethical issues](https://www.emerald.com/insight/content/doi/10.1108/JICES-11-2021-0106/full/html)

[The Ethics of AI and Satellite Data in Weather Forecasting](https://www.qlspace.com.au/the-ethics-of-ai-and-satellite-data-in-weather-forecasting/#:~:text=These%20concerns%20include%20the%20potential,to%20weather%20data%2C%20and%20sustainability.)

[Ethical aspects of AI and climate change](https://etairos.fi/en/2021/09/14/ethical-aspects-of-ai-and-climate-change-2/)

[Ai and machine learning doc… (eth § @ end)](https://theconversation.com/ai-and-machine-learning-are-improving-weather-forecasts-but-they-wont-replace-human-experts-182498)

**Project Plan (Liza)**

* General Team Member Roles:
  + Team Lead:
    - Facilitates discussions and encourages team members to share opinions
    - Creates/introduces the goals and objectives for our next meeting
  + PM:
    - Creates agenda before each meeting
    - Proposes meetings when necessary and sets internal deadlines for assignments
    - Organises documents in our shared drive
    - Tracks progress of assignments and reminds members of internal/official deadlines
  + Dev Lead:
    - Responsible for overseeing our final products, conducting quality and efficiency checks for the software deliverables, ensure that there's no confusion among software tasks.
  + Communication Lead:
    - Ensures team’s communication related work (i.e. proposal & presentation) align with deliverable expectations as set out by course

The team has developed a high-level Gantt chart tracking start and completion dates for all course deliverables and for the project progression itself. Each high-level role has been assigned a Task Owner, who is responsible for delegating the work within the task to the appropriate members of the team. All assigned Task Owners will create a detailed subtask tracker, in an Excel format, that will be continuously updated and monitored by the Task Owner. All tasks are to be delegated equally amongst all members, and if any issues with workload arise, the Task Owner is to be notified.

The team has established protocols to encourage progress transparency and ensure continuous communication.

1. All documentation and the Colab Notebook will be located in a shared Google Drive.
2. All task trackers will also be located in this drive.
3. There will be weekly online meetings.
4. All internal communication will be on an iMessage chat & through email.
5. All members must provide replies within 24 hours

**Risk Register (Matthew)**

In this section, we will identify and analyse the major risks associated with the project. It is important to consider any risks and assess the likelihood of each risk so that outlined actions are in place if these risks were to materialise.

One type of risk pertains to an unequal contribution among team members which can arise in a variety of ways including but not limited to:

1. The potential dropout of a team member
2. Conflicting work/school/personal schedules
3. Emergencies
4. Unequal task distributions

To mitigate this risk, we will establish clear communication channels and a supportive environment where redistribution of tasks take place if necessary. We will also maintain comprehensive documentation to minimise the impact of an individual's departure.

Other than the risk of team and personal conflicts, there are also risks associated the the success of the model itself that including:

1. Managing Large Amounts of Data
2. Addressing the Complex Interplay of Variables
3. Vanishing Gradients in RNN’s

Firstly, handling vast volumes of weather data can be challenging. Sorting through the data to identify relevant information is time-consuming and computationally intensive. Failure to address this risk may result in excessive processing times, increased problem complexity, and potential inaccuracies in predictions. Robust data preprocessing techniques, such as filtering algorithms and feature selection, can be done with PySpark and will enhance model performance by ensuring the model will be trained on clean, relevant and accurate data.

Secondly, weather forecasting involves interconnected variables with intricate relationships. Neglecting these complexities may lead to oversimplified models that fail to capture the full range of weather dynamics.

Finally, recurrent neural networks (RNNs) used in weather forecast AI projects are susceptible to the problem of vanishing gradients, where gradients become extremely small during backpropagation, hindering the model's ability to capture long-term dependencies and accurately forecast complex weather phenomena. This limits the RNN's predictive capabilities, especially for events influenced by historical weather patterns. To mitigate this, the model will employ methods like gradient clipping and appropriate weight initialization that can stabilise and improve the training of RNNs, enabling better forecasting performance for weather AI models.

By proactively brainstorming and addressing these risks, the model will obtain enhanced accuracy, improved decision-making, and more efficient weather predictions contributing to the overall success of the project.

Andreq

intro

Weather forecasting is an essential aspect of our daily lives, not only impacting our personal convenience, but affecting entire industries such as agriculture, transportation, and energy systems. The goal of this project is to develop a machine-learning model that can accurately predict the weather based on historical weather and climate patterns of an area.

Accurately forecasting the weather is crucial to both the economy and the safety of people. Accurate weather predictions facilitate better agricultural planning and crop management for farmers. In transportation, weather patterns are critical for optimizing cargo routes and the safety of plane passengers. Additionally, energy systems rely on accurate weather prediction for effective power management and distribution, particularly for the future of renewable energy sources that are heavily reliant on weather conditions.

Deep learning is a reasonable approach to this project as the weather has many pattern-based characteristics and does not follow a rigid structure, making it difficult to compute with traditional methods. Neural Networks excel at learning patterns and are able to learn from the historical weather events of an area and apply that information to make an accurate prediction given real-time data.

arc

The following is the high level architecture of the system we will use:

**Data Collection**

**Data Preprocessing**

**Model**

**Output**

The chosen model for this project is an RNN. An RNN will be used because it excels in areas such as processing sequential data. Weather is inherently sequentially natured and thus the RNN will be the best at interpreting this data and using it to make an accurate prediction. An RNN is good at handling sequential data because it contains memory, time dependency and back propagation, which will all be used to improve the prediction. In contrast, a CNN will not be used as they are best suited for image interpretation and images will not be used for weather prediction.

Baseline Model

The baseline model that will be used is a persistence model. The persistence model is a simple model where you take the current state of the weather and predict the future state to be the exact same as the current state. This model may seem overly simplistic as it does not use any machine learning or complex algorithms, but it establishes a minimum performance for the model. If the model performs better than this baseline then it can be determined that relevant information was learned by the neural network, whereas if the model performs worse, then it can be determined that the model did not learn any relevant information.