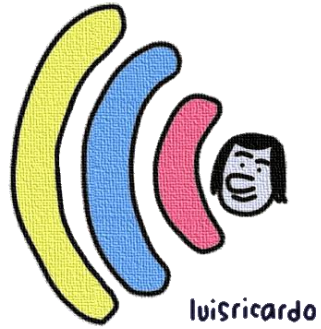
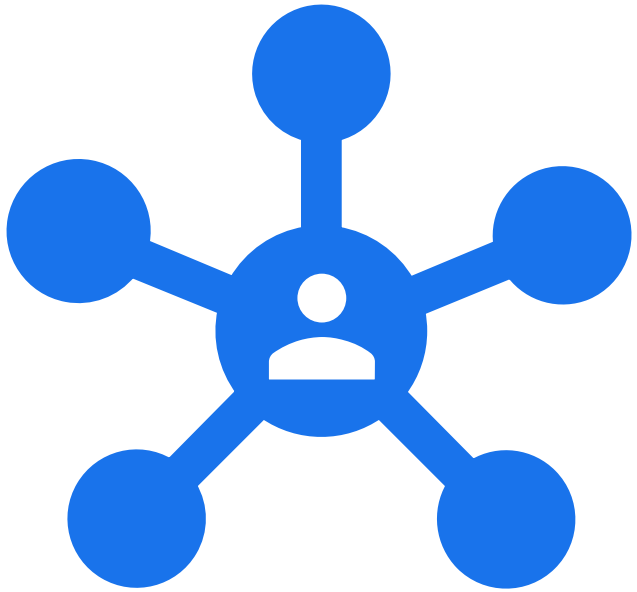
Internet Activity Analysis

STAT405: WENYU DAI, BAHIRAVI RAJKUMAR, JUSTINA JING, STELLA AHN, MASON SMASAL

Defining Internet Activity





- We define internet activity as individual network traffic flows—communications between source and destination IP addresses.
- Each flow contains metadata like duration, protocol, and packet counts. This project analyzes these features to determine which are most important in predicting whether a flow is **benign** or an **intrusion**.

Our Data

- **Source:** CSECICIDS2018 datasets
- **Provider:** Canadian Institute for Cybersecurity
- **Size:** Millions of labeled flows (normal and malicious)
- **Features:** IP addresses, ports, protocols, durations, packet counts, byte rates, etc.
- **Goal:** Identify key metadata predictors for classifying activity

The Variables

Explanatory variable – "Label" (1-Benign, 0-Intrusion/other)

90 predictor variables (eg: Flow Duration, Total Fwd/Bwd Packets, Flow Bytes/s, Timestamp etc)

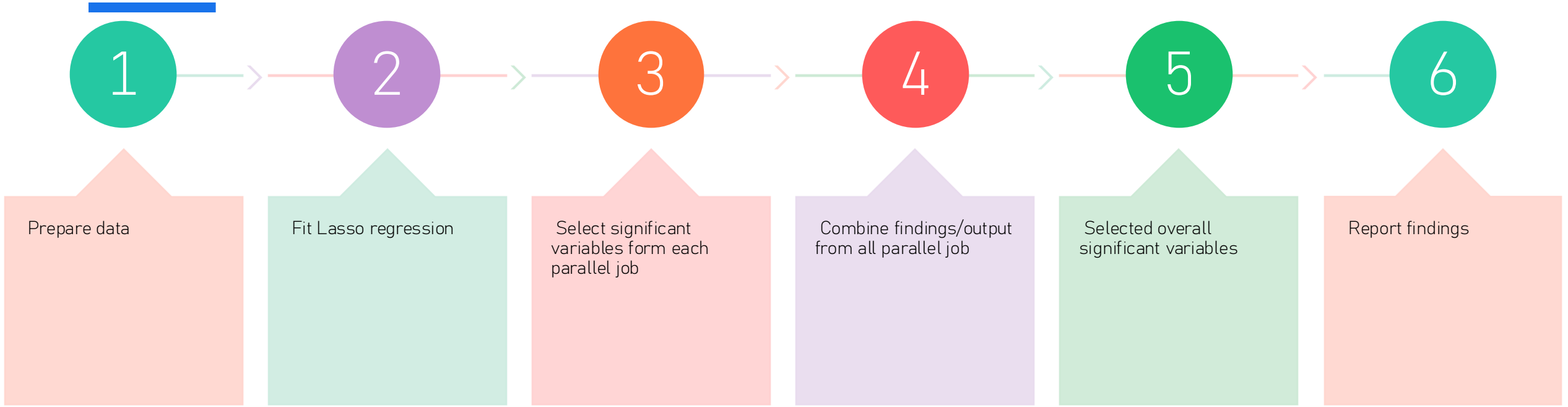
Data cleaning preparing –

- Large data set, hence files were split to smaller chunks to run parallel jobs efficiently.
- Removed variables such as IP address and destination IP, as they are unique identifiers for each activity and do not provide meaningful information for regression analysis.

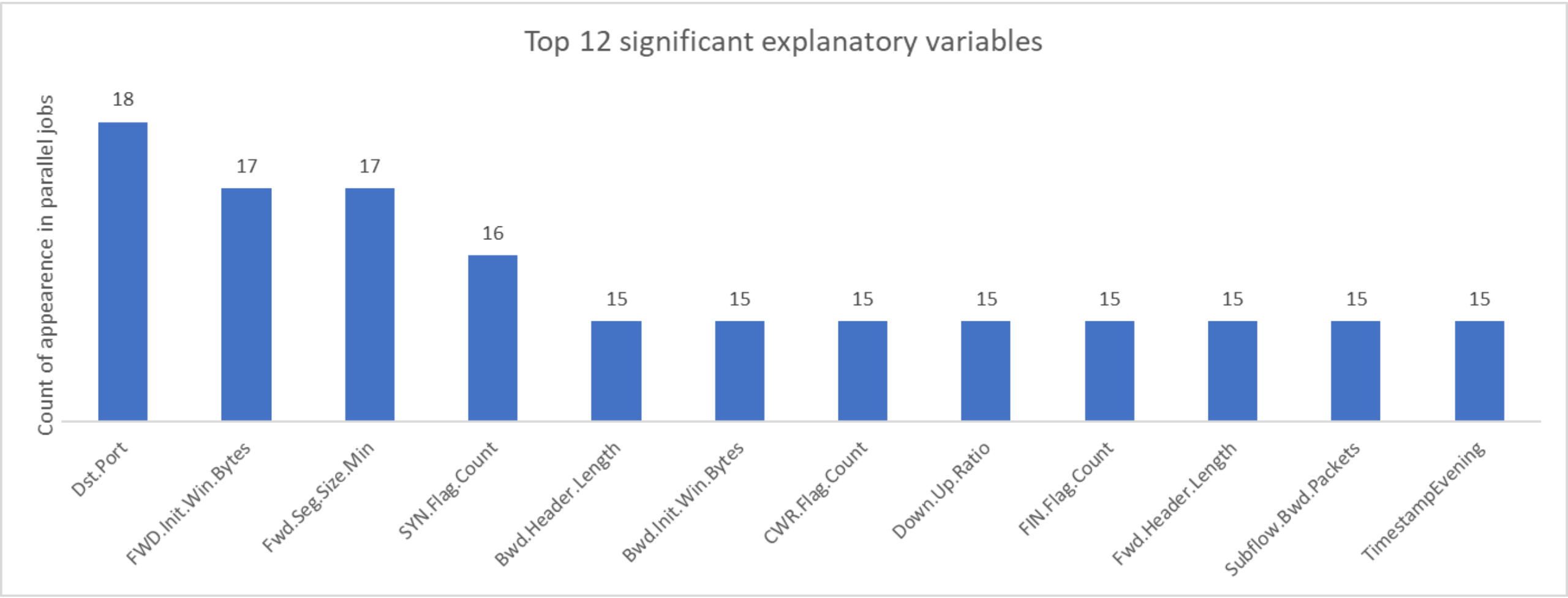
Data preparing –

- Converted variables like Timestamp into categorical format to capture the time of day during which the activity occurred.
- Predictor variable transformed to binary format.

Computational Steps



Visualizations



Conclusion

- This exercise identified key variables associated with distinguishing between **benign** and **intrusive** internet activities.
- We identified stable and important features by analyzing their **appearance counts** after Lasso-based selection.
- Due to the nature of our data — especially after Lasso feature selection — most p-values were either extremely close to **0** or **1**.
- This made the p-values less informative for comparing feature stability or importance. **Therefore**, **instead of relying on noisy or extreme p-values, we focused on counting how consistently each predictor appeared across parallel jobs**.
- This approach selected **10–15** important predictors **out of a total of 90 variables**, which effectively addressed data imbalance and highlighted robust features for future modeling.

Thank you!