Data-Driven Network Traffic Analysis: Identifying Attack Signatures

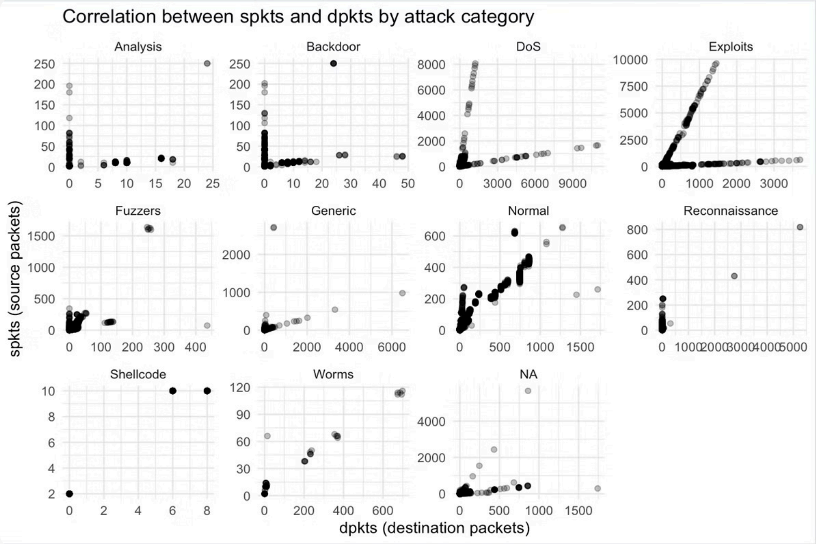
An Analysis of Packet Metrics (spkts & dpkts) for Enhanced Threat Detection

Presenter: Mariia Ivanova
Date: 30/11/2025

The Challenge: Detecting Malicious Network Activity

Network traffic is a blend of normal and malicious activity. Our goal is to **isolate and characterize attack patterns** using fundamental packet-level metrics.

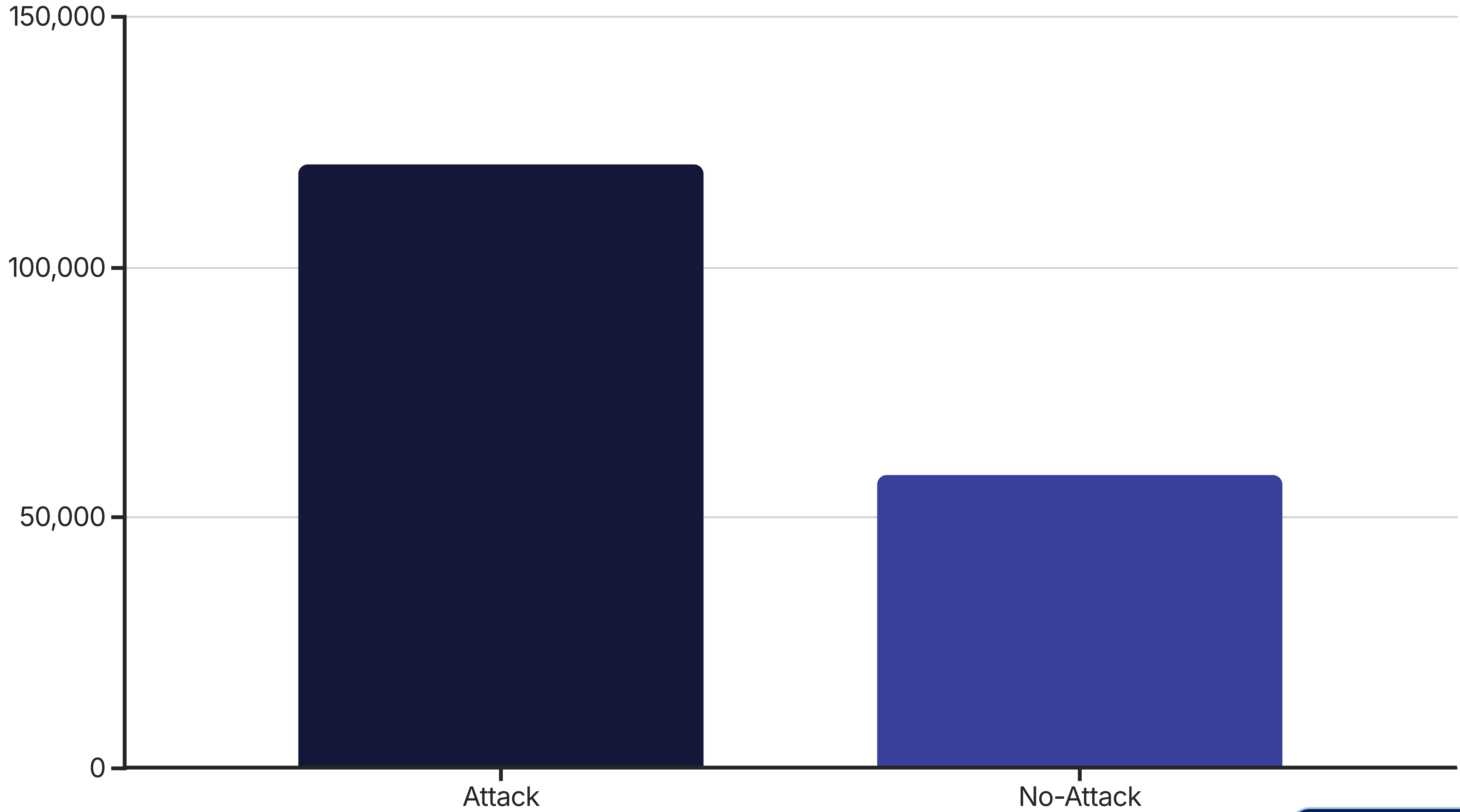
Hypothesis: "There is a significant correlation between source packets (spkts) and destination packets (dpkts) and the likelihood of network traffic being classified as an attack."

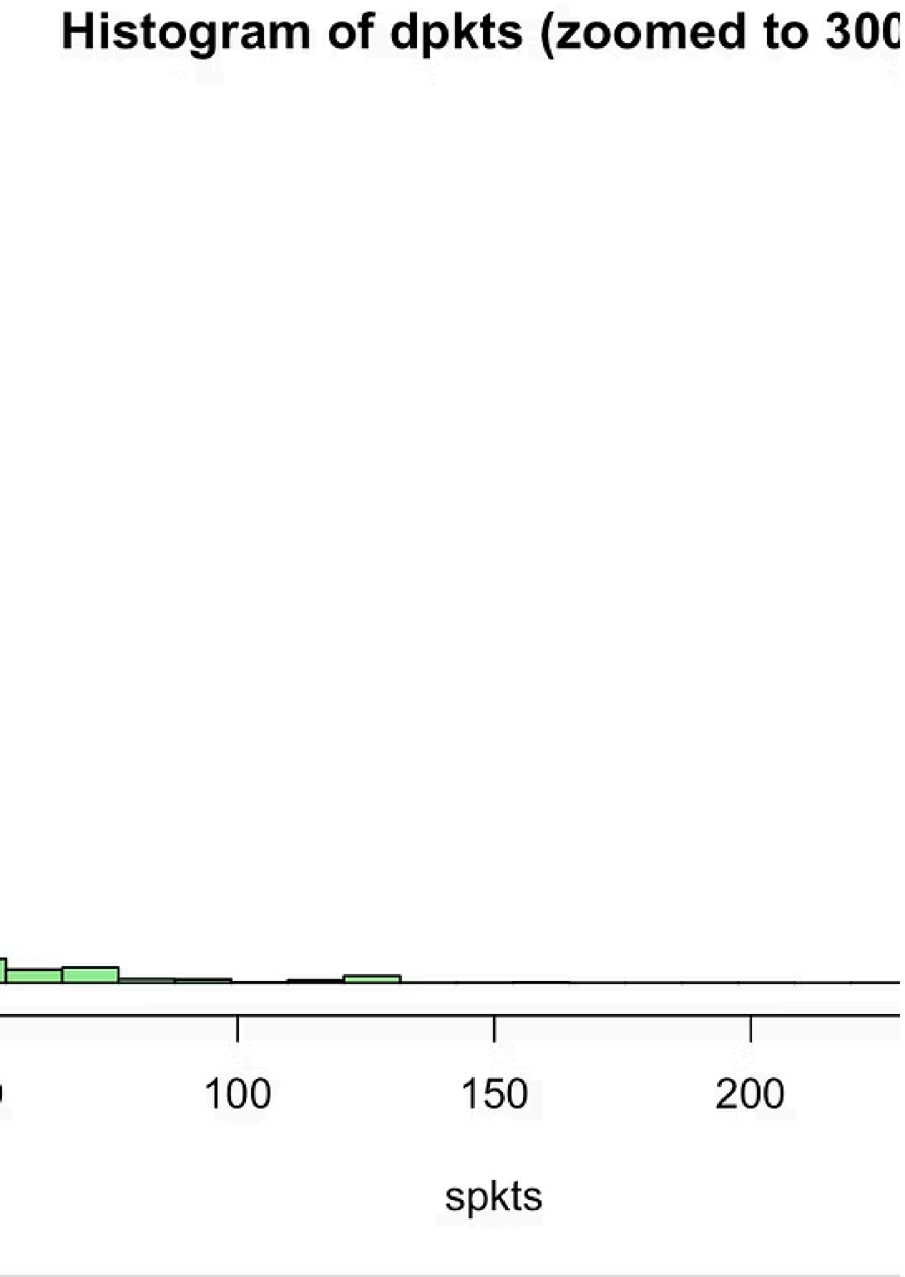


Data Foundation: Cleaning & Validation

The analysis was performed on a real-world network traffic dataset. **Rigorous data cleaning** was essential due to initial inconsistencies in the label and packet count columns (spkts, dpkts).

- **Key Finding 1 (Class Imbalance):** 120,483 Attack instances vs. 58,381 No-Attack instances. This bias must be addressed in modeling.
- **Key Finding 2 (Data Validation):** Packet count columns were converted to numeric, and negative values (logically impossible) were checked and handled.





Distribution Skew: The Majority of Traffic is Small

The distribution of both spkts and dpkts is highly skewed. The **vast majority of values fall between 0 and 50**, with a small number of extreme outliers exceeding 1000.

Implication: This non-normal distribution **precludes the use of parametric tests** (e.g., t-test, ANOVA, Pearson correlation) and necessitates the use of robust or non-parametric methods.

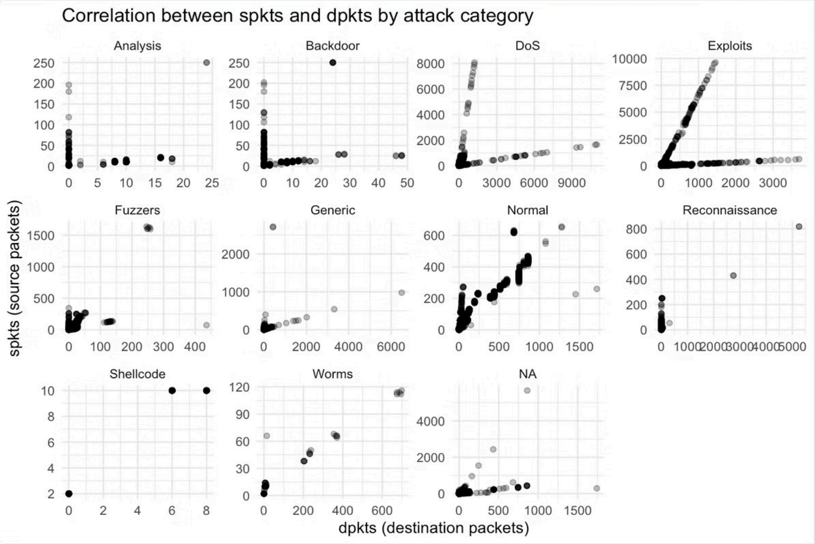
Normal Traffic vs. Attack: A Statistical Divide

Comparing key statistics reveals a clear pattern: **Normal traffic exhibits higher median packet counts** than most attack categories.

- **Key Metrics:** Normal Traffic (No Attack) has a Median dpkts = 10 and Median spkts = 12, resulting in a Ratio (spkts/dpkts) of approximately 1.13 (near symmetrical).
- **Key Observation:** **Attacks often show statistically low medians**, suggesting many attacks are characterized by a few large bursts (high mean) or a high volume of very small packets.

A tibble: 11 × 6

	attack_cat	count	mean_dpkts	median_dpkts	mean_spkts	median_spkts
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	Analysis	1870	2.58	0	6.08	2
2	Backdoor	1630	1.65	0	7.74	2
3	DoS	11604	19.5	0	22.9	2
4	Exploits	31385	22.3	8	32.9	10
5	Fuzzers	17185	6.22	6	14.4	10
6	Generic	37623	0.930	0	2.48	2
7	Normal	52763	38.0	10	30.7	12
8	Reconnaissance	9923	5.37	0	7.06	10
9	Shellcode	1063	3.27	0	5.93	2
10	Worms	123	65.3	6	18.9	10
11	NA	3286	17.3	0	20.0	2



Visualizing Attack Signatures: The Packet Ratio

A scatter plot of spkts vs. dpkts reveals distinct, category-specific patterns that serve as **visual attack signatures**. The **spkts/dpkts ratio** was a key feature engineered from this visual insight and used in subsequent modeling.

- **Signature 1 (Extreme Imbalance):** *Analysis* and *Backdoor* attacks show an extreme ratio (100% to 1000%), where one packet count is very high while the other is low.
- **Signature 2 (Moderate Imbalance):** *DoS* and *Exploits* show a less extreme, but still imbalanced, ratio.
- **Key Finding:** All extreme outliers (high packet counts) belong exclusively to the Attack category.

Model Performance: From Baseline to Robust Detection

Initial Logistic Regression provided a baseline (73% accuracy) but suffered from a high false positive rate (50% of No-Attack instances misclassified). The **Random Forest model** was then deployed to leverage the engineered features and handle the non-linear data distribution.

- **Model Result (Random Forest): Accuracy: 92%**. This model successfully reduced the false positive rate to a manageable level, demonstrating a robust ability to distinguish between normal and malicious traffic.
- **Feature Engineering Success:** The use of **log-transformed features** (log_spkts, log_dpkts) and the **packet ratio** were critical to achieving this performance leap.

```
> confusionMatrix(rf_pred_class, test_data$flag)
Confusion Matrix and Statistics

          Reference
Prediction NoAttack Attack
  NoAttack     7969    424
  Attack       2358  21616

               Accuracy : 0.914
                 95% CI : (0.9109, 0.9171)
    No Information Rate : 0.6809
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7918

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7717
            Specificity : 0.9808
         Pos Pred Value : 0.9495
         Neg Pred Value : 0.9016
             Prevalence : 0.3191
         Detection Rate : 0.2462
   Detection Prevalence : 0.2593
      Balanced Accuracy : 0.8762

       'Positive' Class : NoAttack

>
```

Next Steps: Achieving Production-Ready Security

Recommendation 1 (Bias Mitigation)

Implement advanced strategies to escape bias, such as **cost-sensitive learning** or **ensemble methods** tailored for imbalanced data, to further refine the model's sensitivity to true attacks.

Recommendation 2 (Data Transformation)

Explore alternative data transformations beyond log-scaling, such as **Box-Cox or Yeo-Johnson transformations**, to achieve a more normal distribution for other potential features.

Recommendation 3 (Feature Expansion)

Integrate **other metrics** (e.g., duration, protocol type, service) into the model to capture a broader context of network activity and improve generalization across different attack types.

Conclusion: The Path to Predictive Security

Summary: We successfully validated that spkts and dpkts are **strong indicators of malicious activity**, with attacks exhibiting distinct statistical and visual signatures. The Random Forest model provides a **robust and reliable foundation** for a predictive security system.

Call to Action: The next phase of work will focus on **integrating broader network context and advanced bias mitigation** to deliver a production-ready, low-false-alarm detection system.

Q&A:

Questions?