4th International Conference on Computational Systems and Information Technology for Sustainable Solution

# Identification of Criminal Activity Hotspots using Machine Learning to aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates

Ramshankar Yadhunath, Srivenkata Srikanth, Arvind Sudheer, Suja Palaniswamy

Presented by Ramshankar Yadhunath

# AGENDA

## RELEVANCE AND NEED FOR THE RESEARCH

What and Why is this research important?

## PREDICTIVE MODELLING - HOW DOES IT HELP OUR PROBLEM STATEMENT?

A Machine Learning based approach

## THE RESEARCH METHODOLOGY

How did we go about our idea?

## RESEARCH OUTCOME AND RESULTS

How did our models perform? What was the novelty in our work?

## CONCLUSION

Wrapping it Up!

# RELEVANCE AND NEED FOR THE RESEARCH

## What and Why is this research important ?

# AN OVERLOOKED PROBLEM

"There can be no sustainable development without peace and no peace without sustainable development"

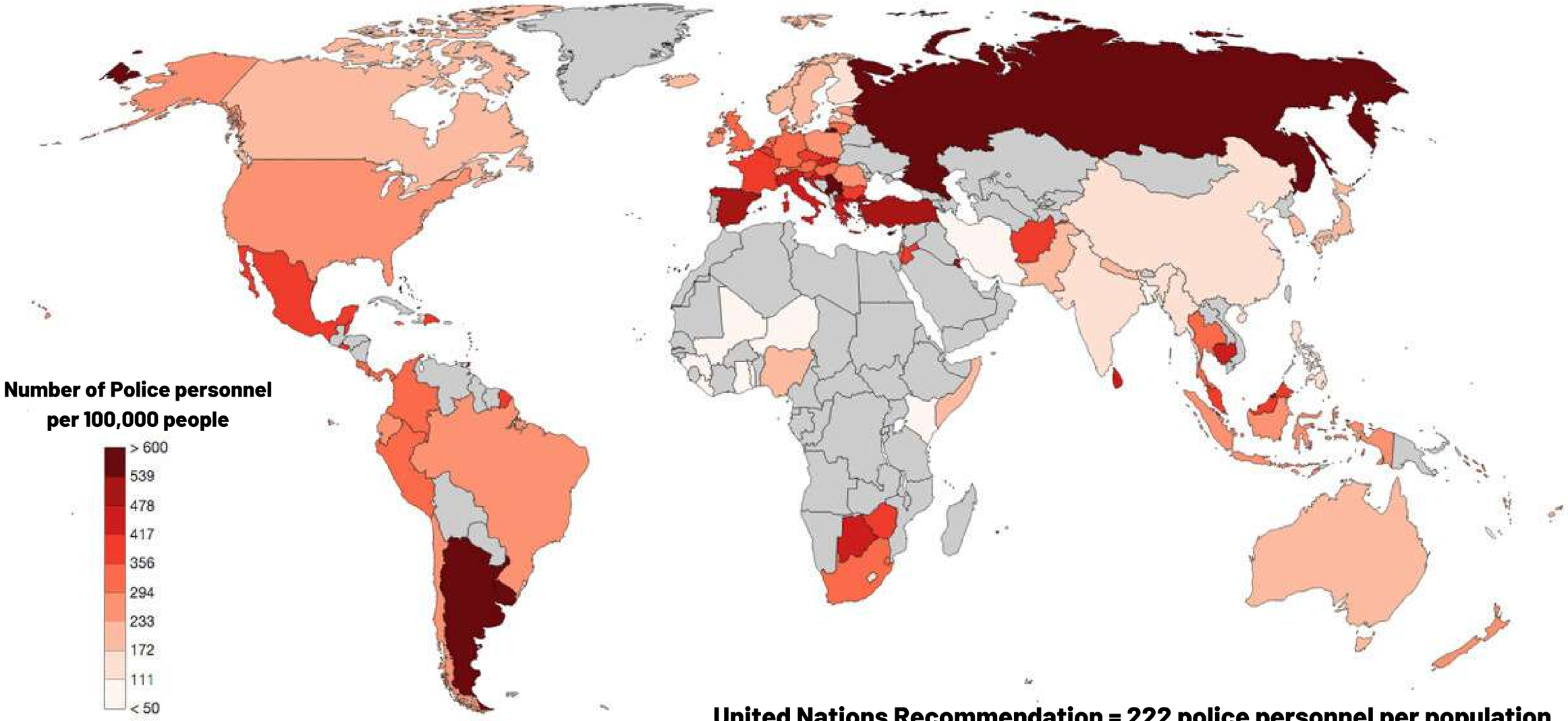
- United Nations 2030 Agenda for Sustainable Development

- Crime is a major deterrent to a peaceful world
- Several organizations are taking steps to reduce crime and its effects

BUT, WHAT IF THERE IS A FACTOR HIDDEN TO THESE ORGANIZATIONS?

"THE PROBLEM OF POOR POLICE-POPULATION RATIOS"

# WORLD (POLICE:POPULATION) RATIOS




**United Nations Recommendation = 222 police personnel per population**

# THE PROBLEM STATEMENT

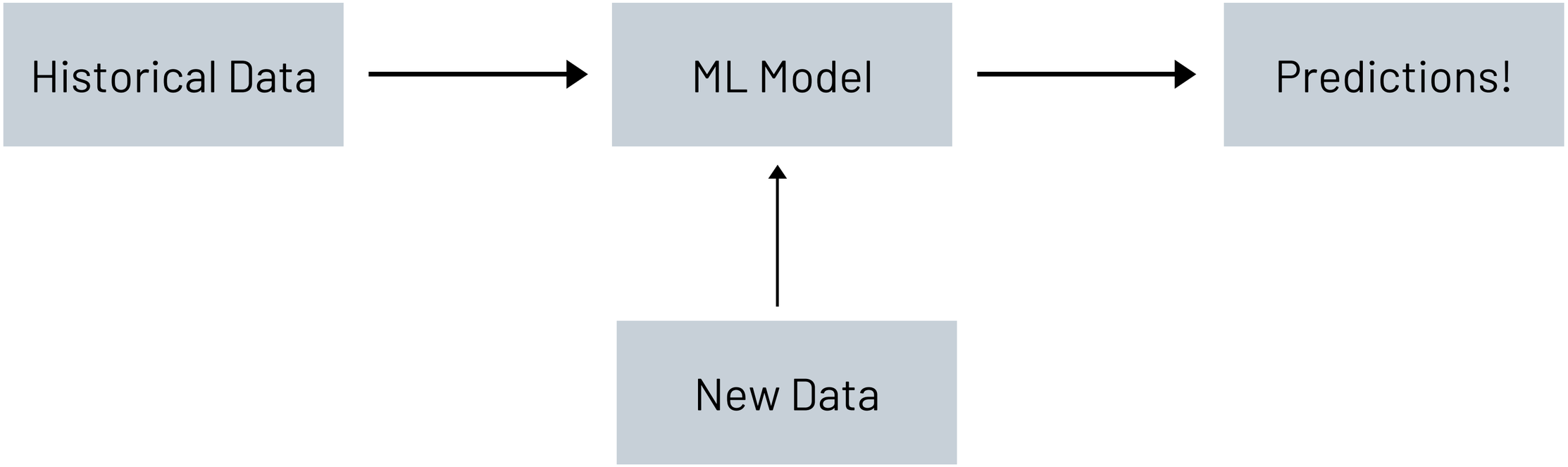
## The Problem Statement

To facilitate effective distribution of police forces in a city among multiple districts based on the extent to which each district is prone to crime at a given hour, in a given day, for a given month.

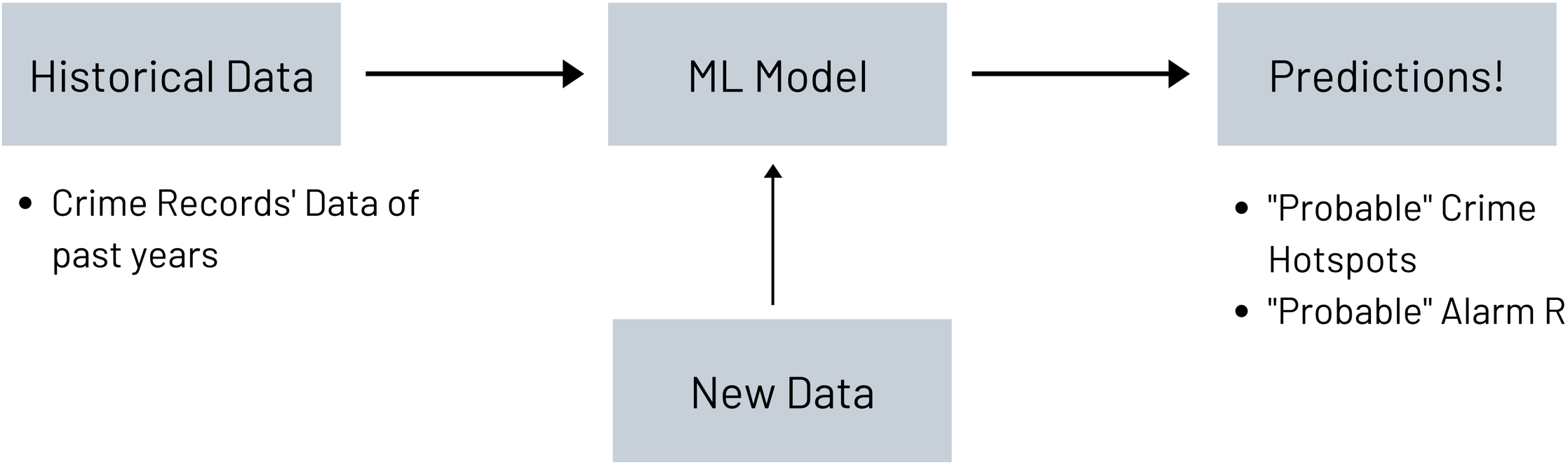
# PREDICTIVE MODELLING

## A Machine Learning Based Approach

# PREDICTIVE MODELLING IN A NUTSHELL



# USING PREDICTIVE MODELLING FOR OUR WORK



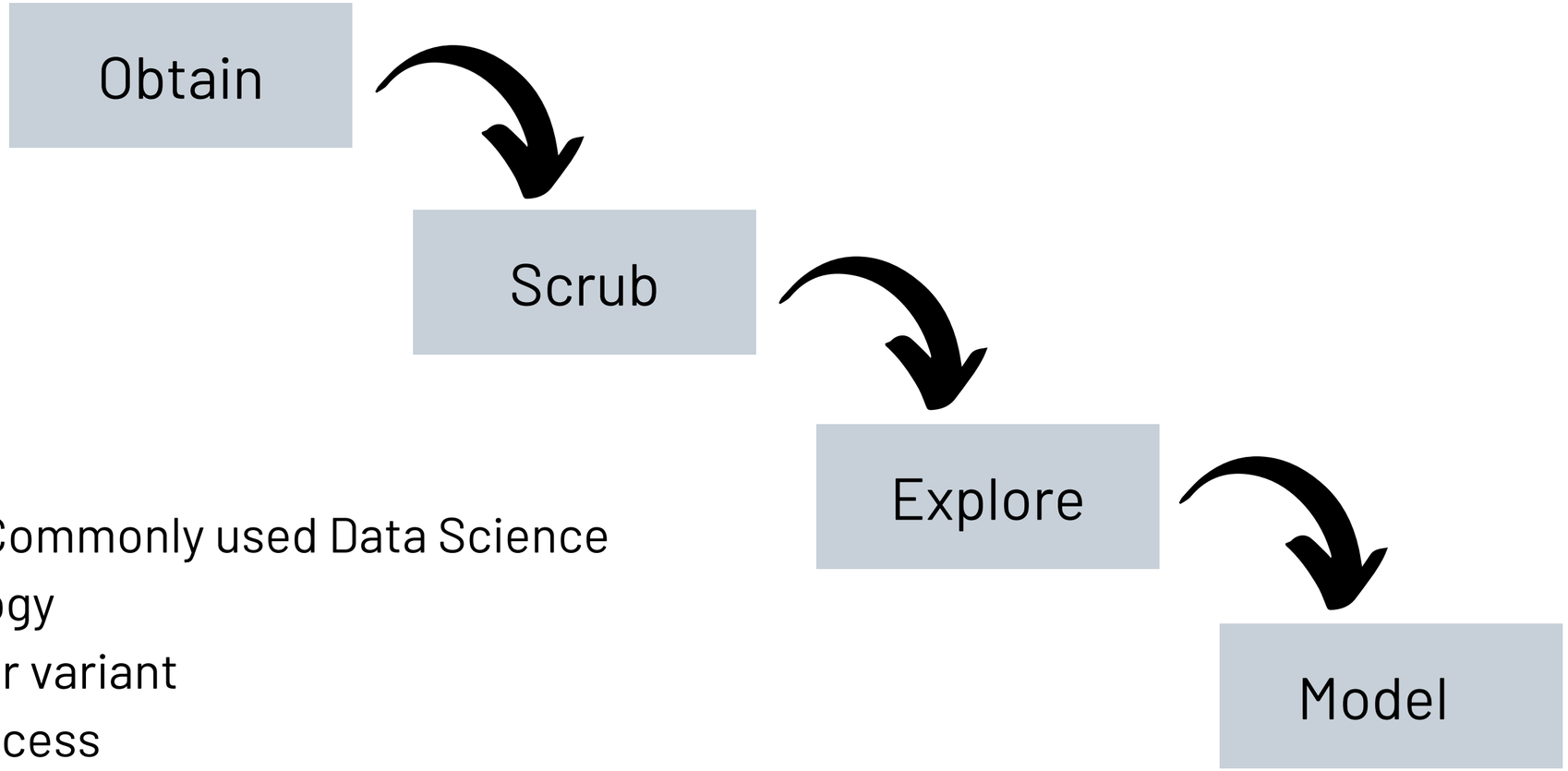

- Crime Records' Data of past years

- "Probable" Crime Hotspots
- "Probable" Alarm Rates

# THE RESEARCH METHODOLOGY

## How did we go about our idea ?

# AN OVERVIEW OF THE METHODOLOGY



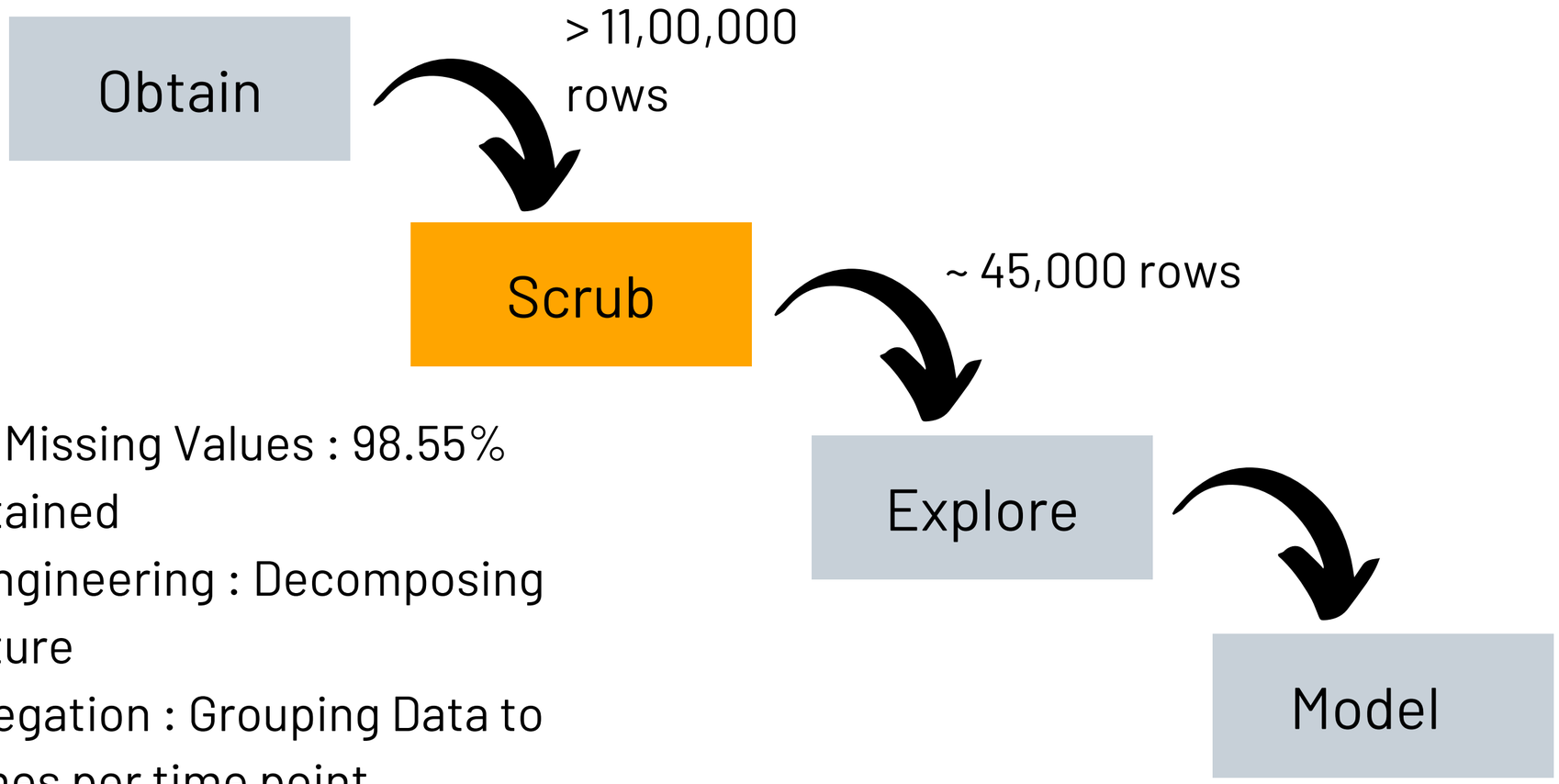

- OSEMN - Commonly used Data Science Methodology
- OSEM - Our variant
- Linear Process

# OBTAIN DATA

Obtain → Scrub → Explore → Model

- City of Chicago Data Portal
- Training Sample : Data from 2015 - 2019 (May)
- Test Sample : Data from 2012 - 2014
- Over 11 lakh entries

# SCRUB (PRE-PROCESS) DATA




- Removing Missing Values : 98.55% entries retained
- Feature Engineering : Decomposing "Date" feature
- Data Aggregation : Grouping Data to count crimes per time point

# EXPLORE DATA

Obtain → Scrub → Explore → Model

- Create a "Target" feature
- Label based on "Normal Distribution" and IQR

TABLE I
FIRST 5 ROWS OF *crimeDat*

| Month | Day | Hour | District | CrimeCount |
|---|---|---|---|---|
| 5 | 2 | 9 | 31 | 1 |
| 1 | 3 | 10 | 31 | 1 |
| 4 | 2 | 10 | 31 | 1 |
| 1 | 2 | 16 | 31 | 1 |
| 1 | 2 | 13 | 31 | 1 |

# EXPLORE DATA

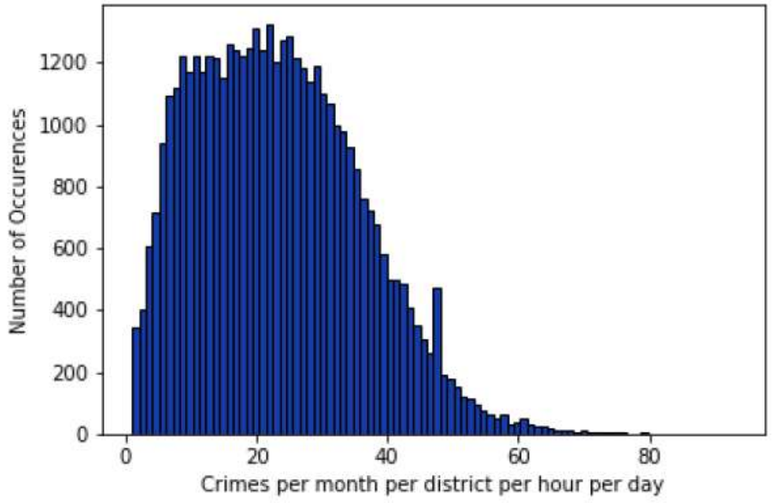
Obtain

Scrub

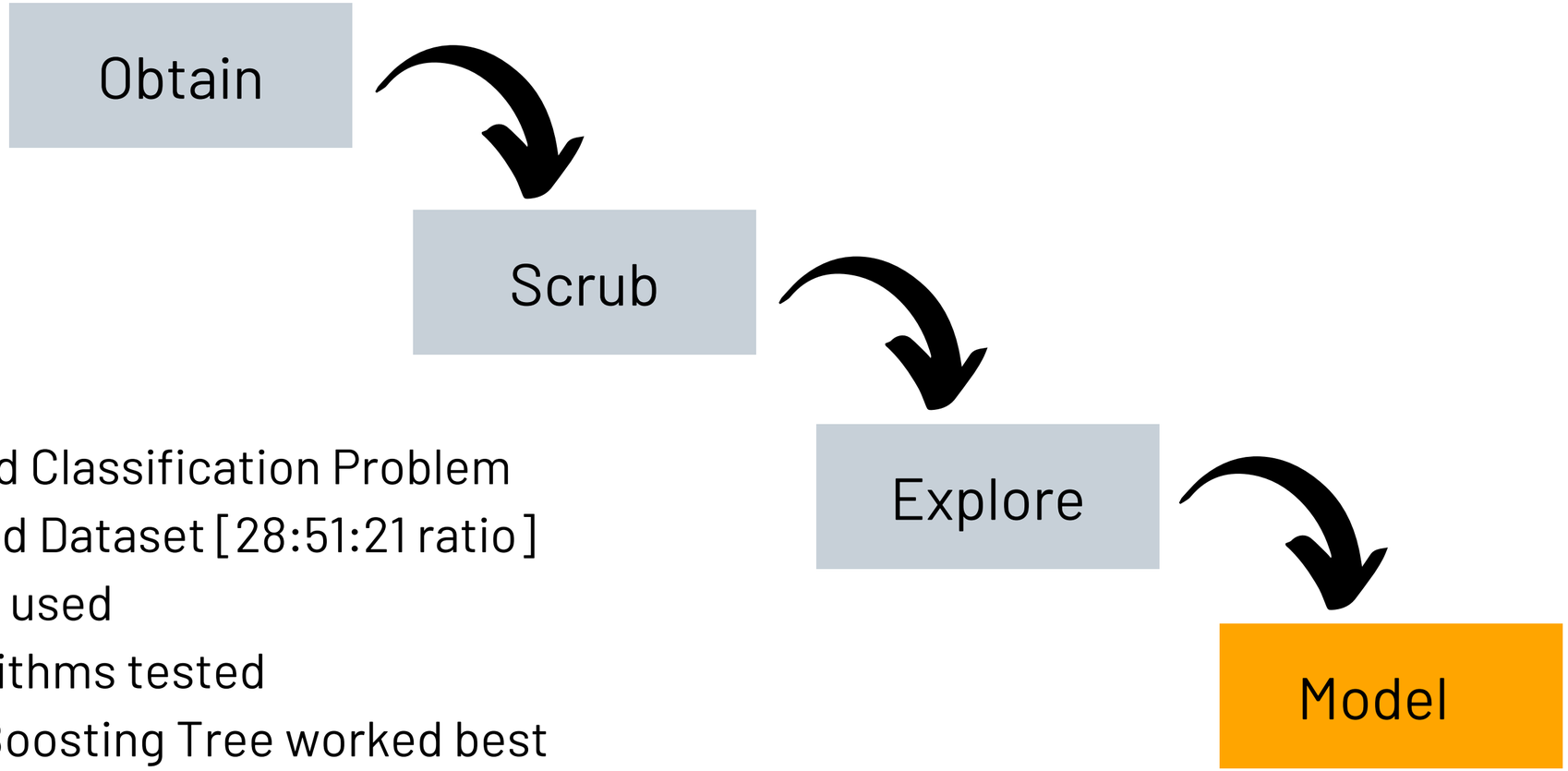
Explore

Model

- Create a "Target" feature
- Label based on "Normal Distribution" and IQR



# MODEL THE DATA




- Supervised Classification Problem
- Imbalanced Dataset [28:51:21 ratio]
- 3 Samples used
- 7 ML algorithms tested
- Gradient Boosting Tree worked best

# RESEARCH OUTCOME AND INNOVATION

## How did our model perform? What was the novelty in our work?

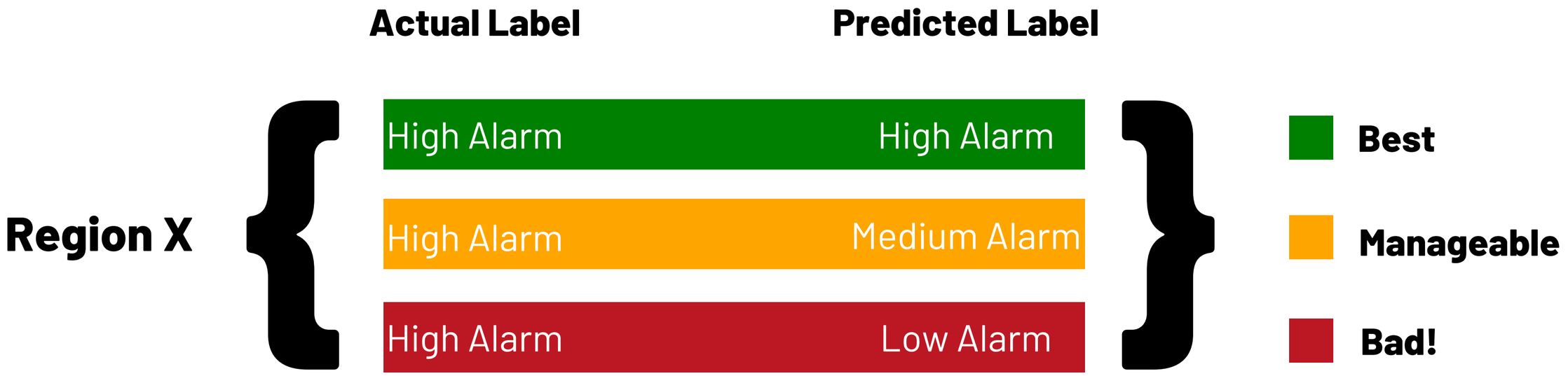
# MODEL EVALUATION METRICS

We have evaluated our models based on the following metrics that are common to most ML problems :

- Accuracy
- Precision
- Recall
- F1 Score
- Unweighted Average Recall

But, we also need a "PROBLEM-SPECIFIC" metric to improve the robustness of our work.

# A NOVEL METRIC FOR OUR PURPOSE




- We must have a model that "minimizes" the ■ scenario
- Problem-specific metric : Percentage of misclassifications of "high alarm" regions as "low alarm" regions
- Let's call this metric "HL-mis" in the further slides

# KEY CONSIDERATIONS WHILE CHOOSING A MODEL

## Key Considerations while choosing a Model :

- High Accuracy
- High F1 score
- Low HL-mis

## Testing Samples :

- Sample 1 : 25% of crimeDat (With class imbalance)
- Sample 2 : 25% of crimeDat (Without class imbalance - Achieved by oversampling)
- Sample 3 : All crime records from 2012-2014

# MODEL COMPARISONS - TRADITIONAL METRICS

TABLE II
COMPARING MODEL PERFORMANCES

| Machine Learning Model | Performance | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sample 1 | | | Sample 2 | | | Sample 3 | | |
| | Accuracy | F1 Score | UAR | Accuracy | F1 Score | UAR | Accuracy | F1 Score | UAR |
| Gradient Boosting Tree | 80.50 | 80 | 78.59 | 81.32 | 81 | 81.47 | 67.48 | 64 | 73.16 |
| Random Forest | 78.35 | 77 | 76.70 | 88.65 | 88 | 88.65 | 66.25 | 62 | 67.39 |
| Decision Tree | 71.73 | 71 | 71.20 | 86.40 | 86 | 86.39 | 65.39 | 61 | 66.39 |
| K Nearest Neighbour | 75.89 | 75 | 73.77 | 80.1 | 79 | 80.10 | 69.68 | 64 | 67.85 |
| Support Vector Machine | 58.32 | 43 | - | 59.25 | 57 | 59.27 | 49.23 | 46 | 56.61 |
| Naive Bayes | 58.51 | 47 | 48.93 | 59.84 | 56 | 59.84 | 47.01 | 44 | 56.98 |
| Logistic Regression | 56.13 | 44 | 46.06 | 58.88 | 57 | 58.89 | 49.63 | 47 | 55.78 |

# MODEL COMPARISONS - OUR METRIC

TABLE III
MISCLASSIFYING "HIGH ALARM" AS "LOW ALARM"

| Machine Learning Model | Percentage of samples that are 2, but wrongly classified as 0 | | |
|---|---|---|---|
| | Sample 1 | Sample 2 | Sample 3 |
| Gradient Boosting Tree | 0.04 | 0.02 | 0 |
| Random Forest | 0 | 0 | 0.11 |
| Decision Tree | 0.04 | 0 | 0.16 |
| K Nearest Neighbour | 0.56 | 0.77 | 0.23 |
| Support Vector Machine | 6.21 | 10.49 | 10.75 |
| Naive Bayes | 6.08 | 11.37 | 11.58 |
| Logistic Regression | 5.91 | 9.53 | 9.69 |

# CONCLUSION

## Wrapping it Up!

# OUR CONTRIBUTIONS

- Our work looks at predictive policing from the angle of "Optimizing low police force" to control crime even in those cities with very high crime rates
- This work can also be incorporated on a state-level or county-level basis and can be the foundation to more complex police force allocation mechanisms
- Our paper is based along the notion of using "Data science as a means of promoting social good"
- The new problem specific metric is an effective way to evaluate the robustness of a model that can predict the alarm rate of a region

# A FEW IMPORTANT REFERENCES


[1] nancy.cao, "Sustainable development goals," Available at https://www.unodc.org/unodc/en/about-unodc/sustainable-development-goals/sdgs-index.html (2019/07/19).

[3] M.-J. Lin, "More police, less crime: Evidence from us state data," *International Review of Law and Economics*, vol. 29, pp. 73–80, 06 2009.

[4] V. P. Sriharsha Devulapalli, "Indias police force among the worlds weakest," Available at https://www.livemint.com/news/india/india-s-police-force-among-the-world-s-weakest-1560925355383.html (2019/07/19).

# A FEW IMPORTANT REFERENCES


[9] T. Almanie, R. Mirza, and E. Lor, "Crime prediction based on crime types and using spatial and temporal criminal hotspots," *arXiv preprint arXiv:1508.02050*, 2015.

[10] S. Das and M. R. Choudhury, "A geo-statistical approach for crime hot spot prediction," *International Journal of Criminology and Sociological Theory*, vol. 9, no. 1, 2016.

[13] "Crimes - 2001 to present — city of chicago — data portal," Available at https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data (2019/05/11).

# THANK YOU