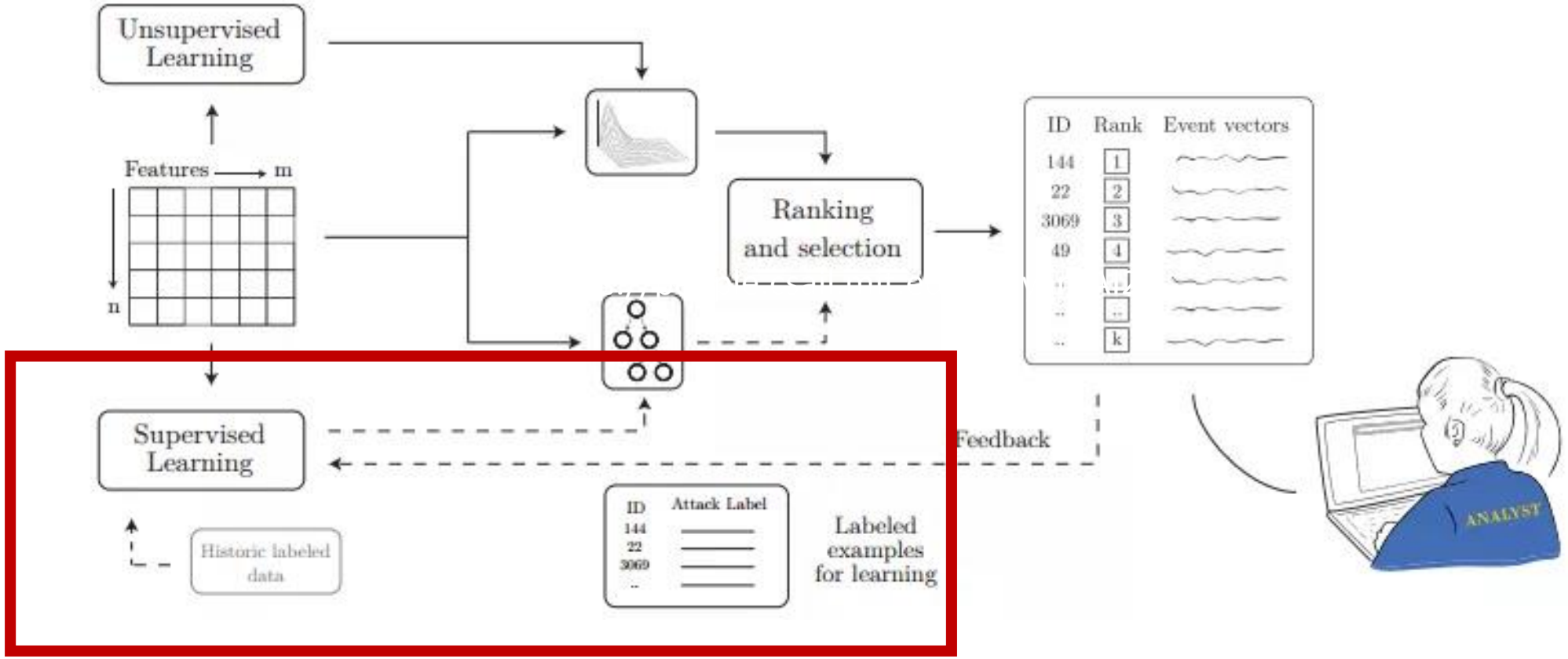
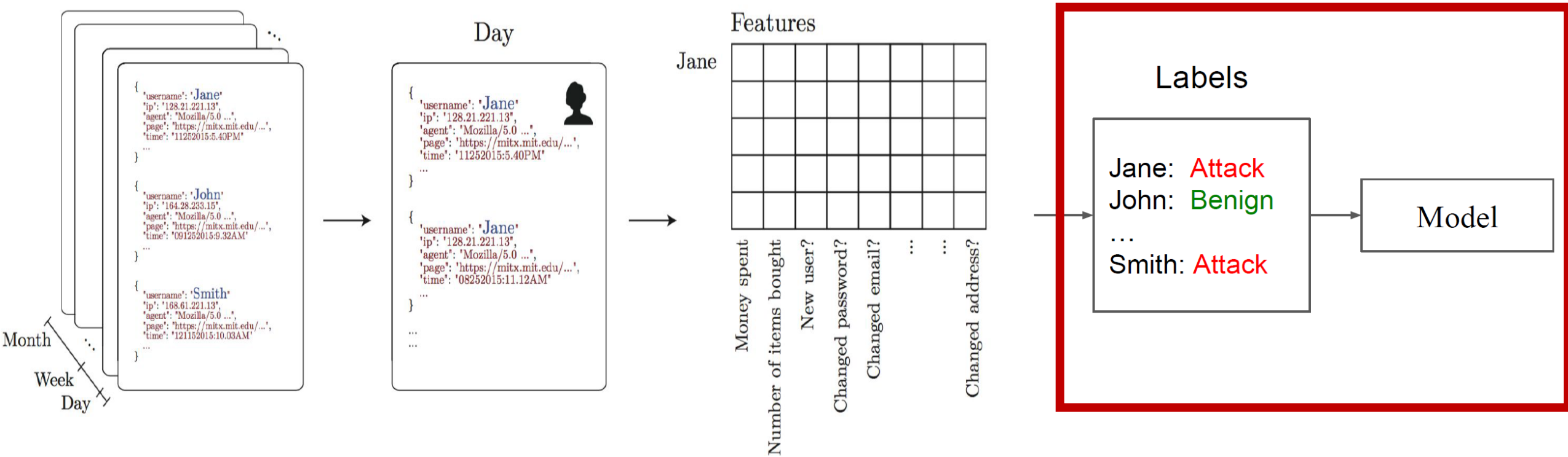
Machine Learning Capstone Project 2019 Task 3

Ai^2: A big data machine to defend

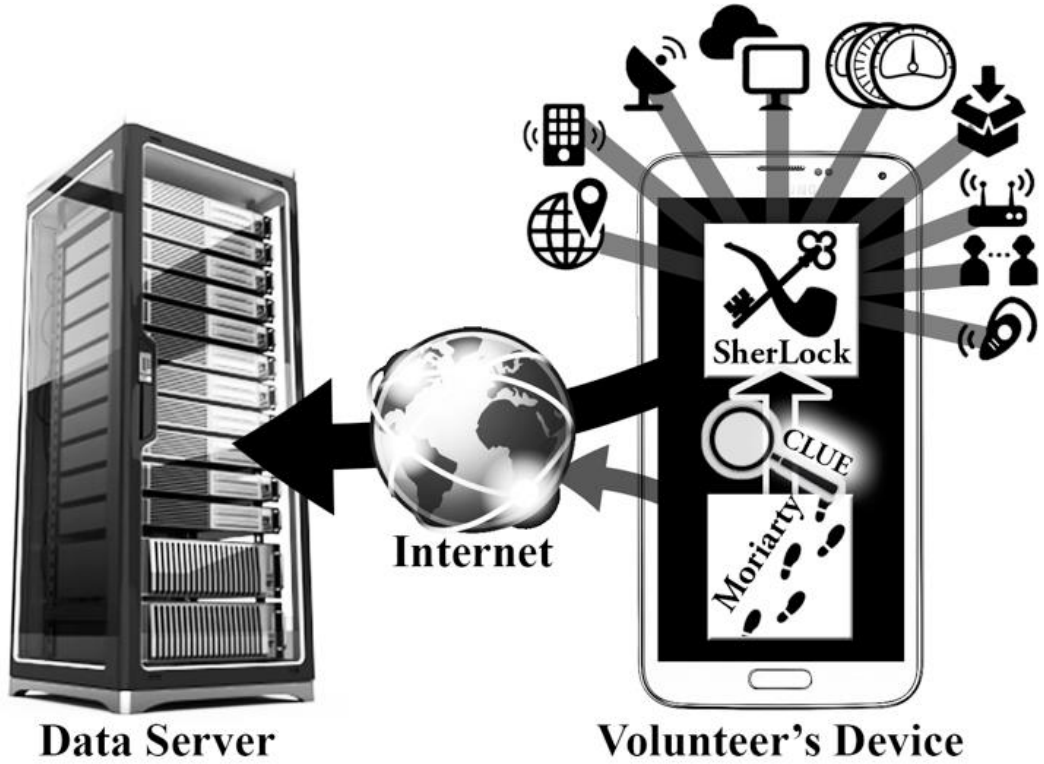


https://people.csail.mit.edu/kalyan/AI2/


Month
Week
Day
Day
Features
Jane
Money spent
Number of items bought
New user?
Changed password?
Changed email?
...
...
Changed address?
Labels
Jane: Attack
John: Benign
...
Smith: Attack
Model

Introduction

- The available dataset is from an experiment which two smartphone agents:
  - SherLock collects a wide variety of software andsensor data at a high sample rate.
  - Moriarty perpetrates various attacks on the user and logs its activities, thus providing labels for the SherLock dataset.



Milestone 1

- In Milestone 1 we carry out a complete Exploratory Data Analysis process of a subset of this dataset.
- This system learns a **descriptive model** of those features extracted from the data via unsupervised learning, using unsupervised learning methods.
- Then we known a lot about the data collected by Sherlock.
- Then, it’s the moment to meet Moriarty

Target

- Moriarty is a benign application paired with a malicious behavior.
- Both the benign application and its malicious behavior are changed every few weeks.
- Moriarty logs both the benign and malicious activities it performs.
- Moriarty leaves clues on the device for SherLock to collect.
- The clues serve as explicit labels for the time series dataset collected by SherLock.

Ai^2: A big data machine to defend

Big data processing system.

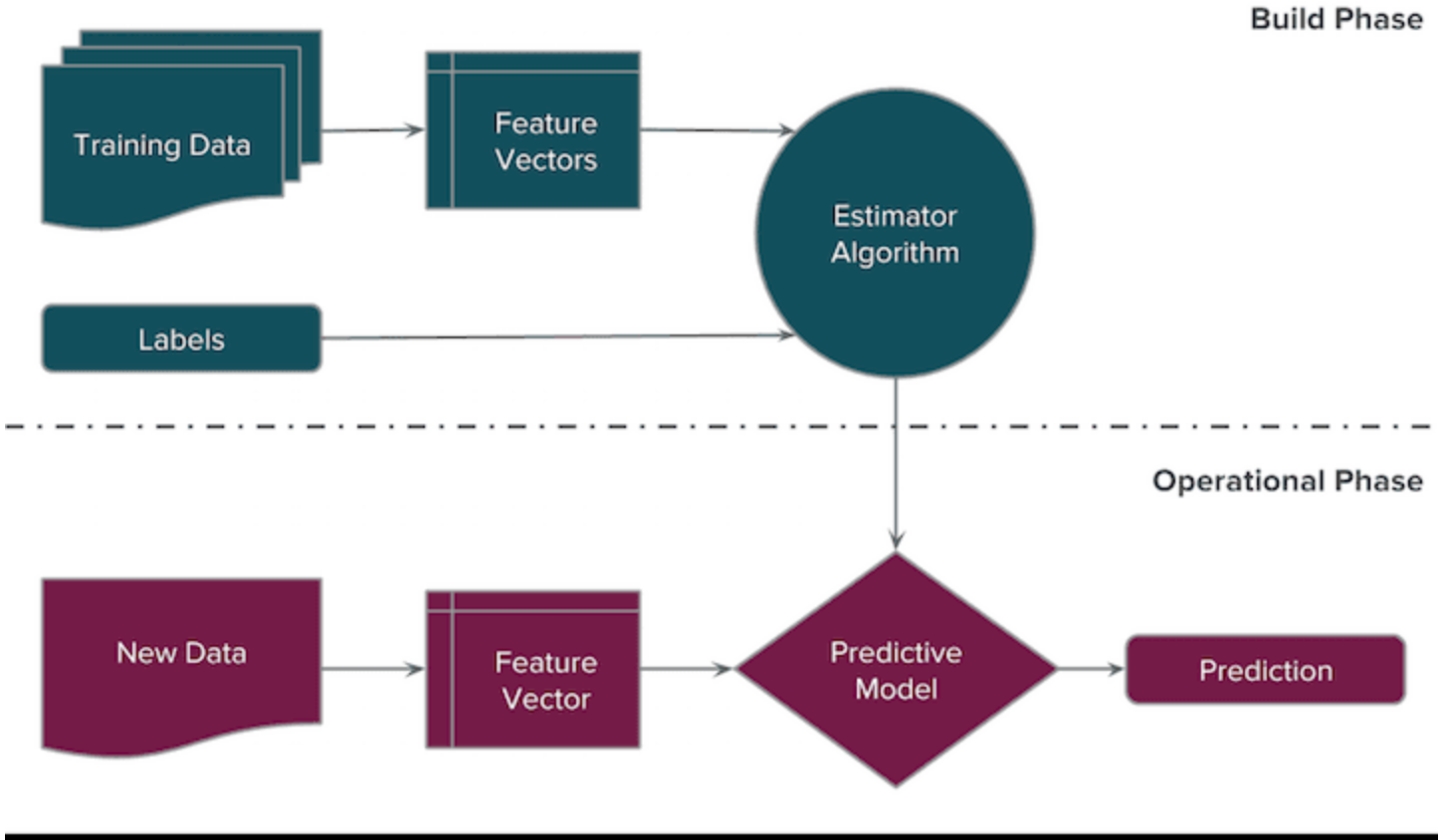
Outlier detection system.

Feedback mechanism and continuous learning.

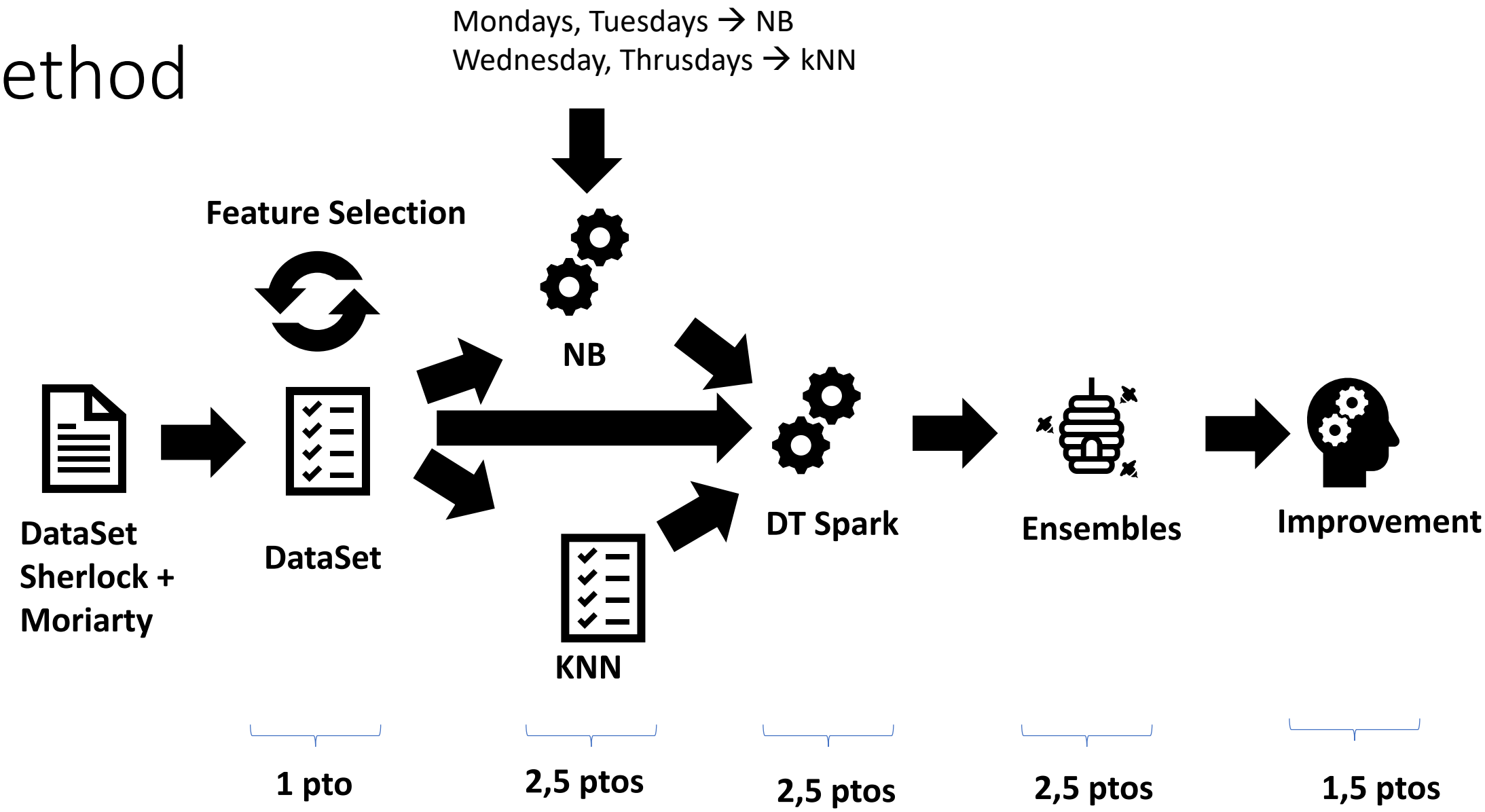
Supervised learning module

Target

- We have data (from Sherlock) and we have labels (from Moriarty) then: **We can build a model to predict the Moriarty attach**



Method


Mondays, Tuesdays → NB
Wednesday, Thrusdays → kNN
Feature Selection
NB
DataSet
Sherlock +
Moriarty
DataSet
KNN
DT Spark
Ensembles
Improvement
1 pto
2,5 ptos
2,5 ptos
2,5 ptos
1,5 ptos

Deliverables

- Public Presentation (10 min, Dec 18)
- Github (with label/branch) (Dec 22)
- Report (ACM format) or Labbook

Feature Selection

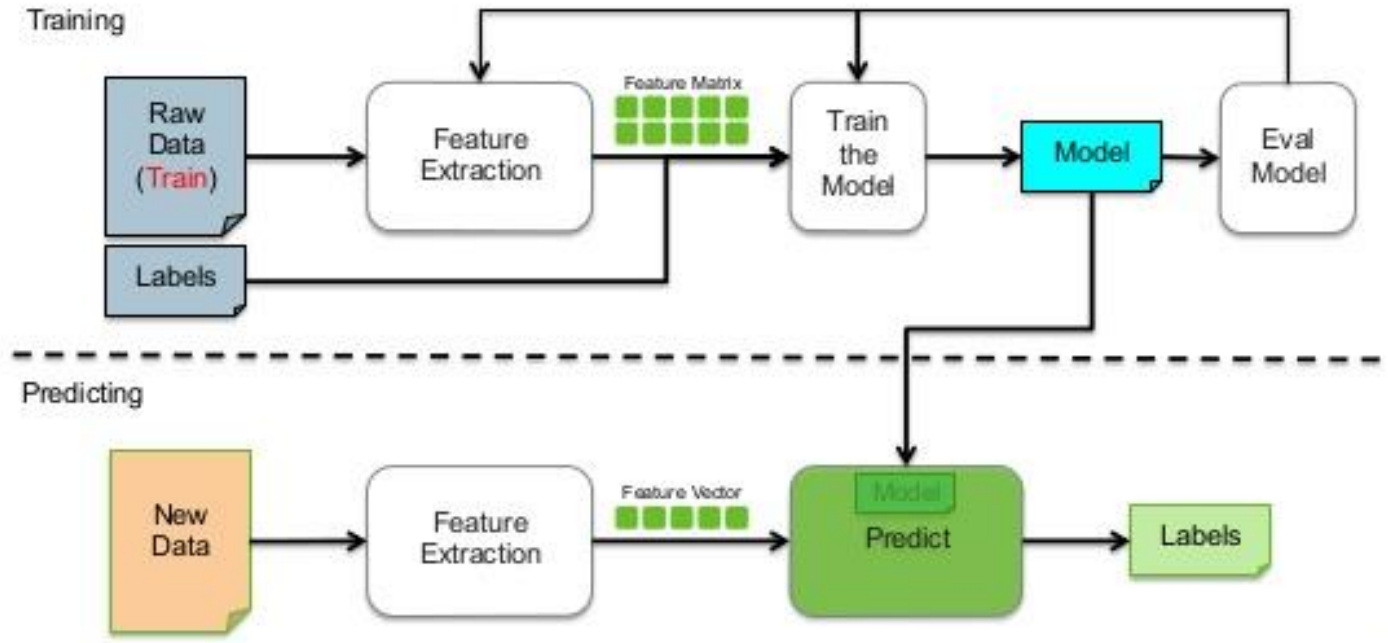
Probe	*Sample Interval*	*Sensors*	*Num. Fields*	*Description*
T0	1 day	Telephony Info	15	Information on the current telephony configuration.
		Hardware Info	6	The device's hardware configuration.
		System Info	5	Kernel, SDK, baseband, and general information.
T1	1 minute	Location	15	{longitude, latitude, altitude, (anonymized via clustering)}, speed, and accuracy.
		Cell Tower	5	Cell tower ID, type, and reception info.
		Device Status	14	Brightness, volume levels, orientation, and modes.
		WiFi Scan	4	**For each visible AP:** identifiers, encryption, frequency, and signal strength.
		Bluetooth Scan	9	**For each visible device:** identifiers, device class (type), parameters, and signal strength.
T2	15 seconds	Accelerometer	51	Statistics on 800 samples captured over a duration of 4 seconds at 200Hz. For each respective axis: mean, median, variance, covariance between axis, middle sample, FFT components and their statistics. A subset of these features is extracted from the orientation, rotation, and barometer sensors.
		Linear Accelerometer	51	
		Gyroscope	51	
		Orientation	9	
		Rotation Vector	12	
		Magnetic Field	51	
		Barometer	16	
T3	10 seconds	Audio	21	Statistics over 5 seconds.
		Light	3	Luminosity
T4	5 seconds	Global App Stats	98	Information on the CPUs, memory, network traffic, IO interrupts, and connected WiFi AP.
		Local App Stats	70	**For each running application:** statistics on CPU, memory and network traffic. Linux level process information from the system /proc folder.
		Battery	14	Configuration and statistics on power consumption and temperature.

Choose a selection of sensors from the available information from Sherlock.

In the first approach it is not necessary to use a complex feature selection criterion

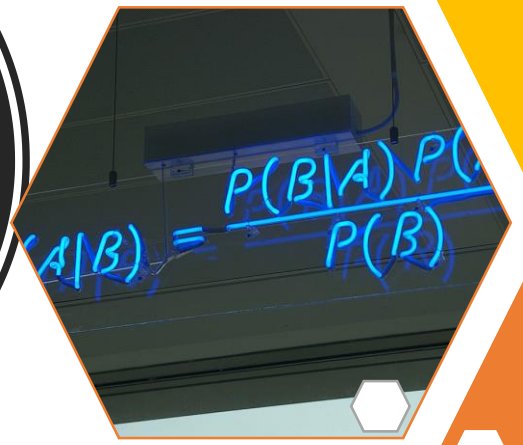
Modelling

Supervised Learning Workflow

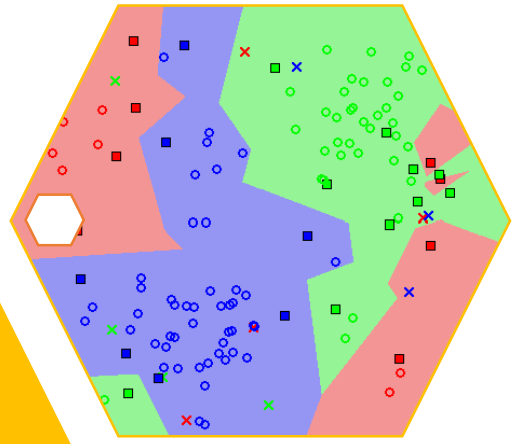


Hortonworks

Modelling



KNN



NaiveBayes

Tree Models & Ensembles