Log Data Analysis & Software Diagnosis

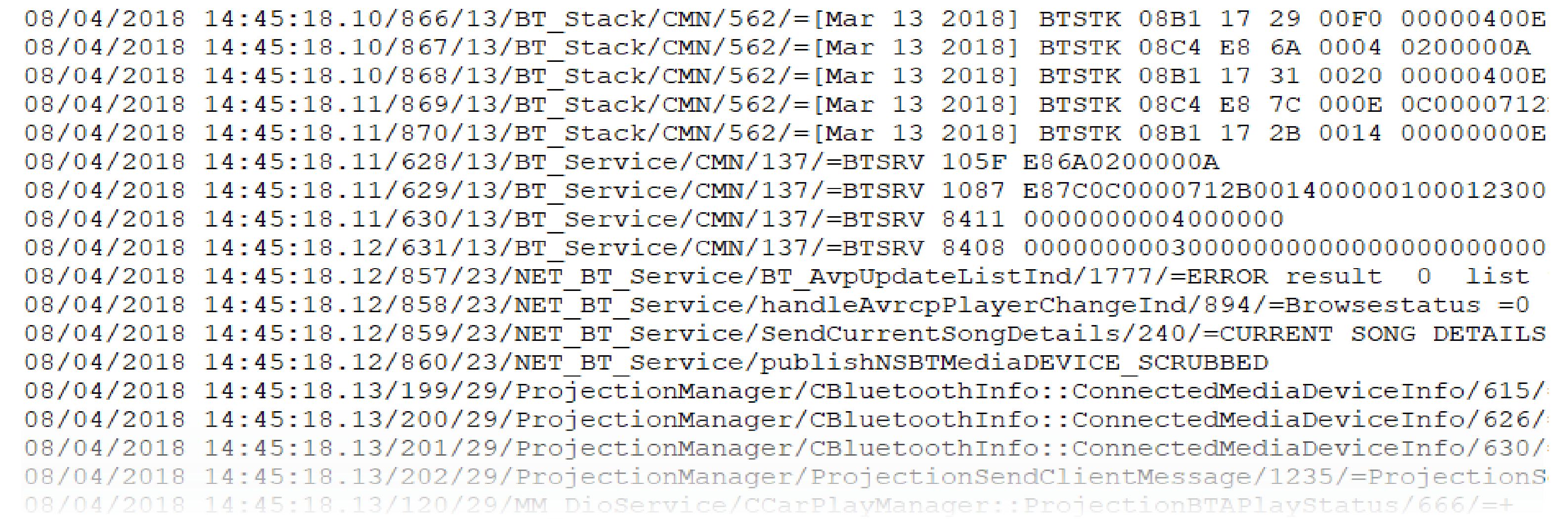
Centre for Mechatronics and Hybrid Technology
McMaster University
Elliott (Yixin) Huangfu

EECOMOBILITY (ORF) & HEVPD&D CREATE

BACKGROUND

Logs are machine data generated constantly by operating system, recording software running events and status. They provide rich information for developers to track issues and reproduce errors.

The goal of this project is to detect software defects automatically by analyzing system log files.

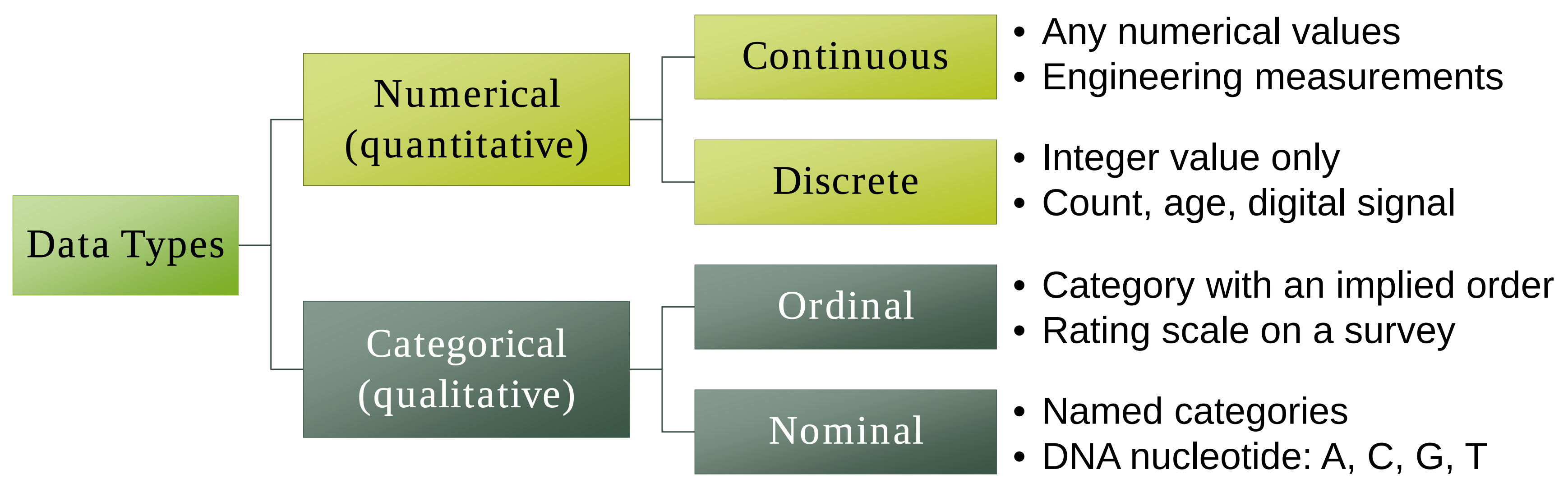


```
08/04/2018 14:45:18.10/866/13/BT_Stack/CMN/562/=[Mar 13 2018] BTSTK 08B1 17 29 00F0 00000400E
08/04/2018 14:45:18.10/867/13/BT_Stack/CMN/562/=[Mar 13 2018] BTSTK 08C4 E8 6A 0004 0200000A
08/04/2018 14:45:18.10/868/13/BT_Stack/CMN/562/=[Mar 13 2018] BTSTK 08B1 17 31 0020 00000400E
08/04/2018 14:45:18.11/869/13/BT_Stack/CMN/562/=[Mar 13 2018] BTSTK 08C4 E8 7C 000E 0C0000712
08/04/2018 14:45:18.11/870/13/BT_Stack/CMN/562/=[Mar 13 2018] BTSTK 08B1 17 2B 0014 00000000E
08/04/2018 14:45:18.11/628/13/BT_Service/CMN/137/=BTSRV 105F E86A0200000A
08/04/2018 14:45:18.11/629/13/BT_Service/CMN/137/=BTSRV 1087 E87C0C0000712B00140000010001230O
08/04/2018 14:45:18.11/630/13/BT_Service/CMN/137/=BTSRV 8411 0000000004000000
08/04/2018 14:45:18.12/631/13/BT_Service/CMN/137/=BTSRV 8408 00000000030000000000000000000000
08/04/2018 14:45:18.12/857/23/NET_BT_Service/BT_AvpUpdateListInd/1777/=ERROR result  0  list
08/04/2018 14:45:18.12/858/23/NET_BT_Service/handleAvrcpPlayerChangeInd/894/=Browsestatus =0
08/04/2018 14:45:18.12/859/23/NET_BT_Service/SendCurrentSongDetails/240/=CURRENT SONG DETAILS
08/04/2018 14:45:18.12/860/23/NET_BT_Service/publishNSBTMediaDEVICE_SCRUBBED
08/04/2018 14:45:18.13/199/29/ProjectionManager/CBluetoothInfo::ConnectedMediaDeviceInfo/615/
08/04/2018 14:45:18.13/200/29/ProjectionManager/CBluetoothInfo::ConnectedMediaDeviceInfo/626/
08/04/2018 14:45:18.13/201/29/ProjectionManager/CBluetoothInfo::ConnectedMediaDeviceInfo/630/
08/04/2018 14:45:18.13/202/29/ProjectionManager/ProjectionSendClientMessage/1235/=ProjectionS
08/04/2018 14:45:18.13/120/29/MM DioService/CCarPlayManager::ProjectionBTAPlayStatus/666/=+
```

ANALYSIS of NOMINAL DATA

Unlike numerical data from most engineering applications, log data are mostly nominal and symbolic. This means mathematical manipulation cannot be applied.

Still, the order information of log messages can be utilized as it contains valuable clues explaining software behaviours.

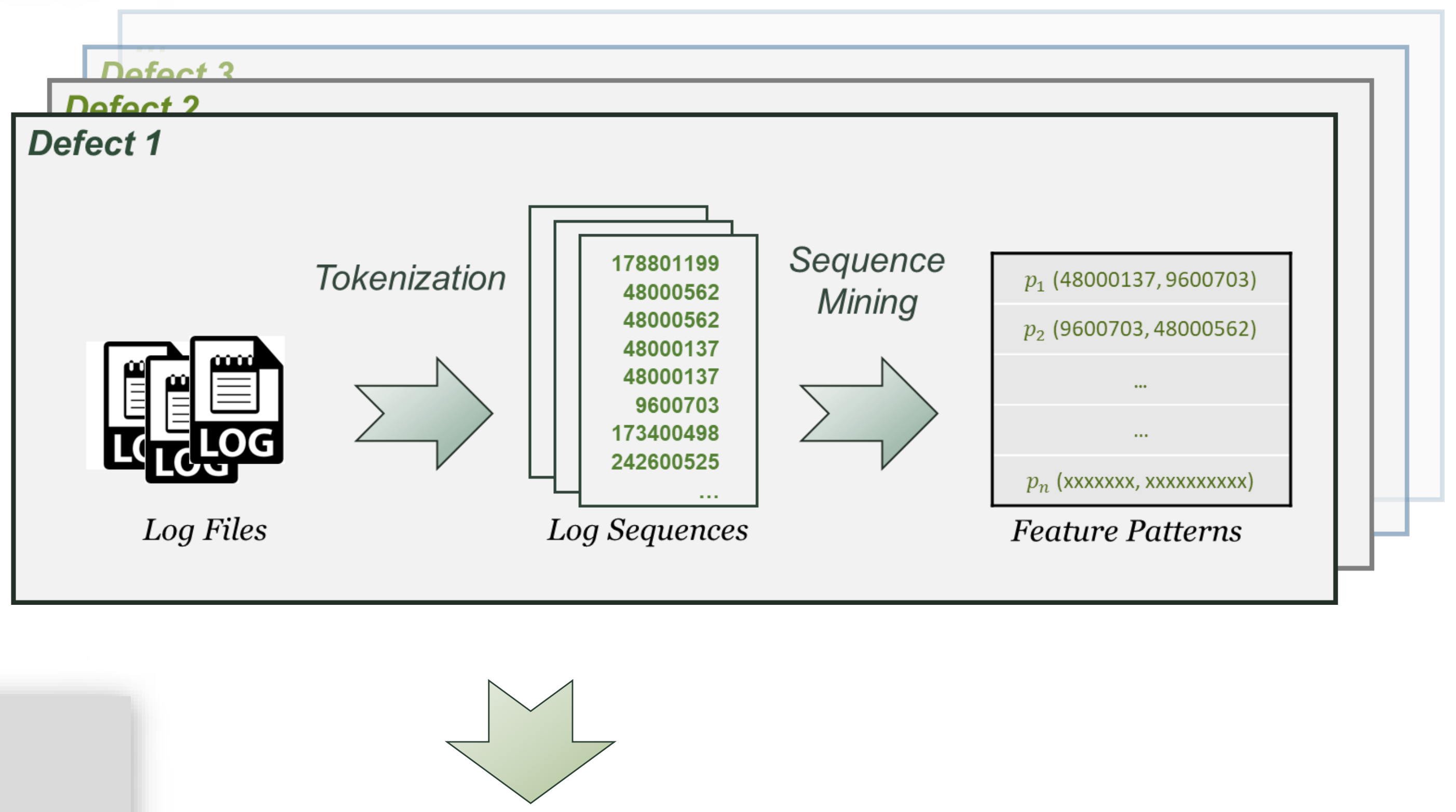


- Any numerical values
- Engineering measurements
- Integer value only
- Count, age, digital signal
- Category with an implied order
- Rating scale on a survey
- Named categories
- DNA nucleotide: A, C, G, T

PATTERN DISCOVERY with DATA MINING

Logs lines are tokenized and converted into sequences of nominal values. The intuition is to extract unique sequences that appear in erroneous log sequences, while not present in regular ones.

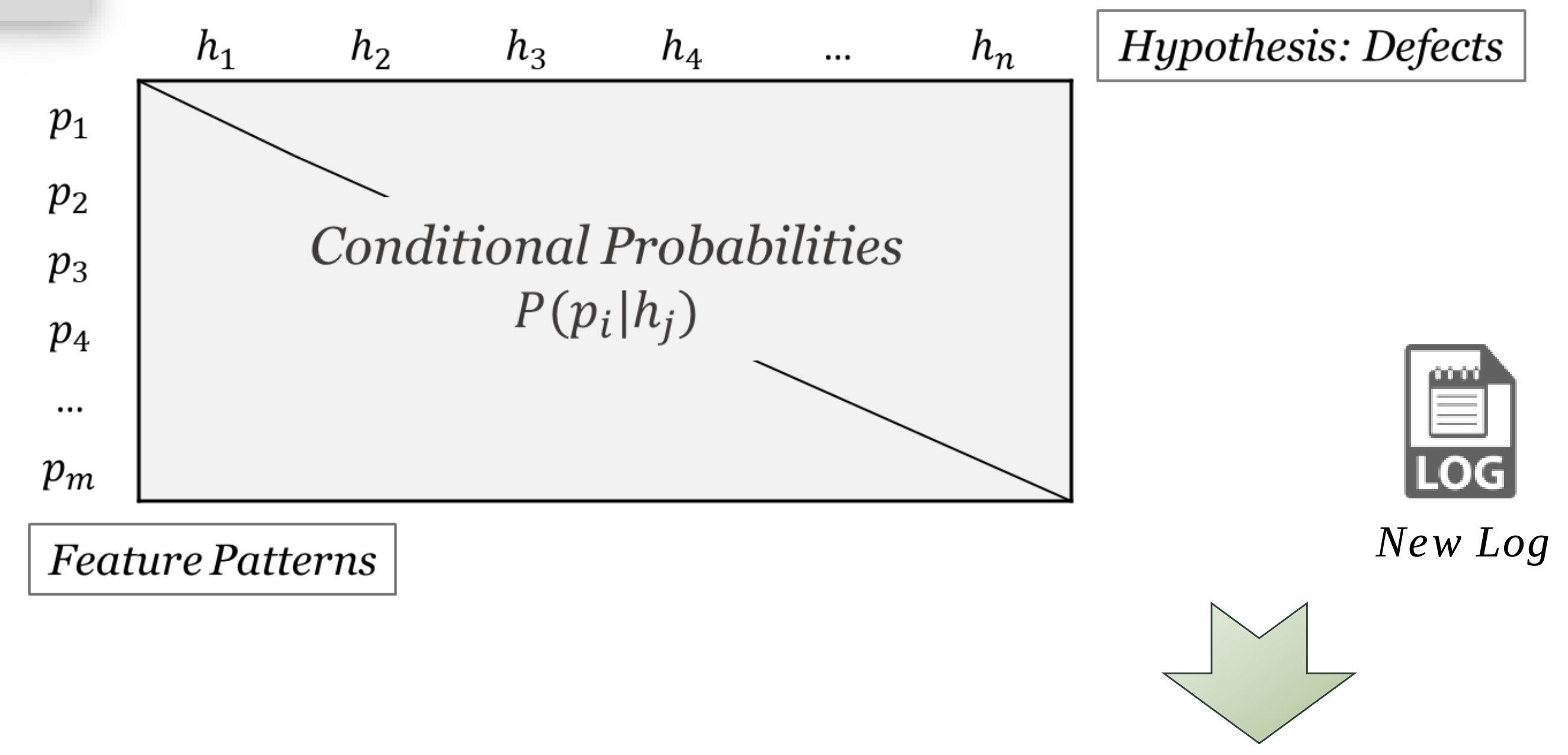
Such approach is called contrast data mining. A prefix tree based search algorithm is implemented, and outputs a set of sequence patterns for every defect.



NAÏVE BAYES CLASSIFIER

Naïve Bayes classifier is a highly practical method in machine learning to predict target values. To apply this approach, every pattern (p_i) is treated as a feature, and a series of hypothesis h_1, h_2, … h_n represents certain defects happening. The conditional probability $P(p_i|h_j)$ is given in a matrix form.

During testing, a new log instance is examined to determine the occurrence of each feature pattern. Then, either $P(p_i|h_j)$ or $P(\neg p_i|h_j)$ will be selected for the calculation. Naïve Bayes Classifier outputs the hypothesis with largest probability product.



$$h_{NB} = \underset{j=1,2,\dots,n}{\operatorname{argmax}} P(h_j) \prod_{0}^{m} P(p_i|h_j)$$

RESULT & NEXT-UPS

Classification Accuracy	
Valid Group (30 cases)	37%
Noise Group (68 cases)	88%
Overall	72%

The method has achieved overall 72% accuracy, and shows good rejection against noise data, which means a low possibility of giving false positives.

However, 37% accuracy in the valid group shows potential for improvement. Future works will focus on enhancing the quality of feature extraction with recurrent neural networks.

We acknowledge the support of the Ontario Research Fund: Research Excellence Program.

We acknowledge the support of the Natural Sciences and Engineering Council of Canada (NSERC), which invests annually over $1 billion in people, discovery and innovation.