Huawei Advanced Data Science With Spark Streaming

Jianfeng Qian, Cheng He
Huawei Research Institute

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE

Contents

- streamDM: Stream Mining in Spark Streaming

  (Jianfeng Qian)

- Business scenarios in Huawei with Spark Steaming

  (Cheng He)

Open Source Machine Learning Projects

- Apache Mahout
  - May 2010, v0.1: support Hadoop
  - Apr 2014, Mahout-Samsara, v0.10: support Spark and H2O
  - April 2016,v0.12: R-like DSL, support Flink
- oryx&oryx2
  - Dec 2013, v0.3.0: real-time large-scale machine learning support Hadoop
  - Dec 2015, v2.0: support Spark Streaming
- Apache SAMOA: Scalable Advanced Massive Online Analysis
  - Jul 2015, v0.3.0: support Storm, Samza and Flink

Stream Data Mining?

Stream Data Mining

Data Streams

- Sequence is potentially infinite
- High amount of data: sublinear space
- High speed of arrival: sublinear time per example
- Once an element from a data stream has been processed it is discarded or archived
- Data is evolving

Approximation algorithms

- Small error rate with high probability

streamDM?

streamDM


input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data


Spark

streamDM is incremental

- streamDM is designed specifically to be used inside Spark Streaming.
- All algorithms are incremental

streamDM!

streamDM for users

- Download streamDM

  git clone https://github.com/huawei-noah/streamDM.git

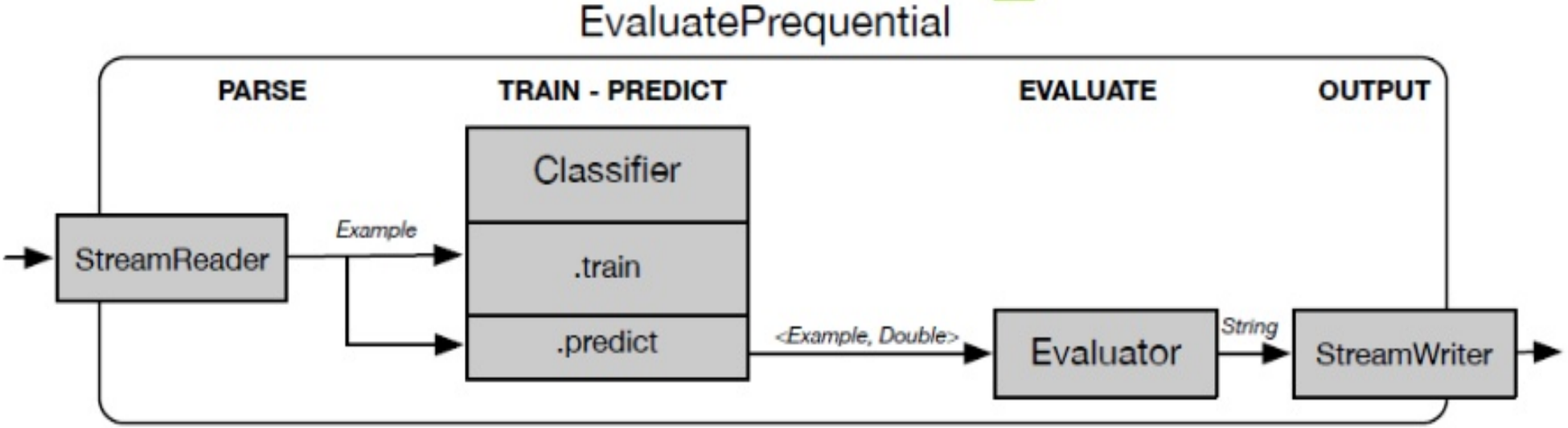
- Build streamDM

  sbt package

- streamDM execute tasks

```
./spark.sh "EvaluatePrequential \
 -l (SGDLearner -l 0.01 –o LogisticLoss -r ZeroRegularizer) \
 –s (FileReader –k 100 –d 60 –f ../data/mydata)" \
1> ../sgd.log 2>../sgd.result
```

streamDM for Programmers



- **StreamReader** read and parse Example and create a stream
- Learner provides the train method from an input stream
- Model data structure and set of methods used for Learner
- **Evaluator** evaluation of predictions
- **StreamWriter** output of streams

streamDM

- Advanced machine learning methods including streaming decision trees, streaming clustering methods as CluStream and StreamKM++.
- Ease of use. Experiments can be performed from the command-line, as in WEKA or MOA.
- High extensibility
- No dependence on third-part libraries, specially on the linear algebra package Breeze.

streamDM

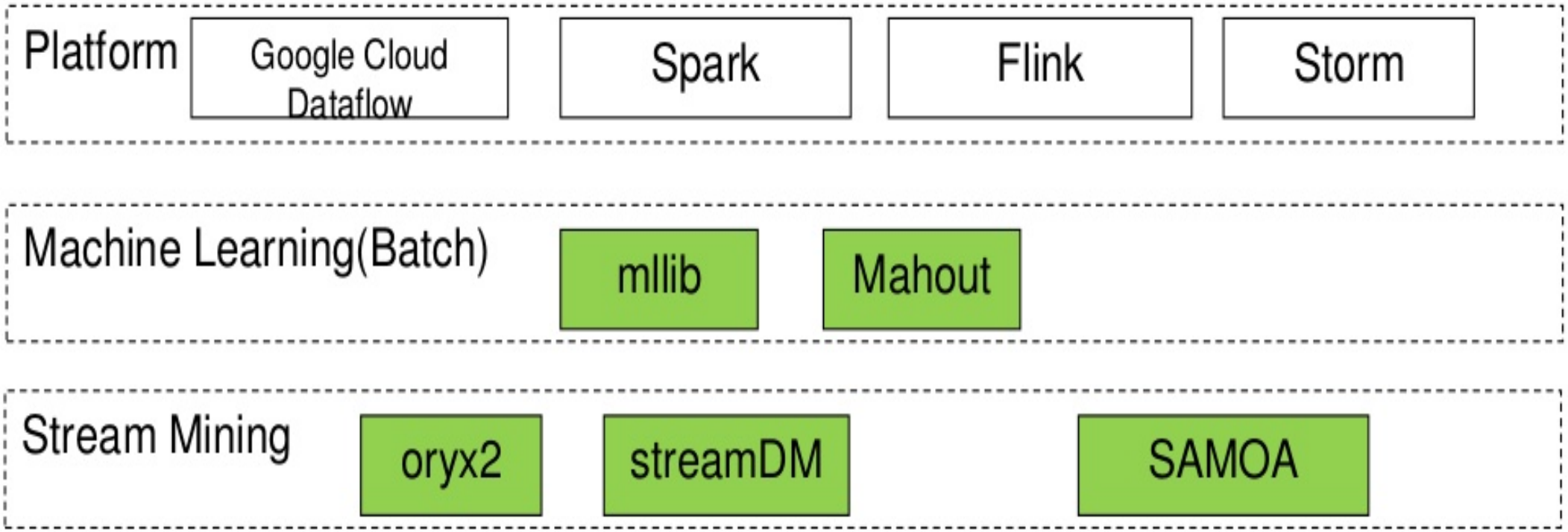
First Release 31/12/15

- SGD Learner and Perceptron
- Naive Bayes
- CluStream
- Hoeffding Decision Trees
- Bagging
- Stream KM++

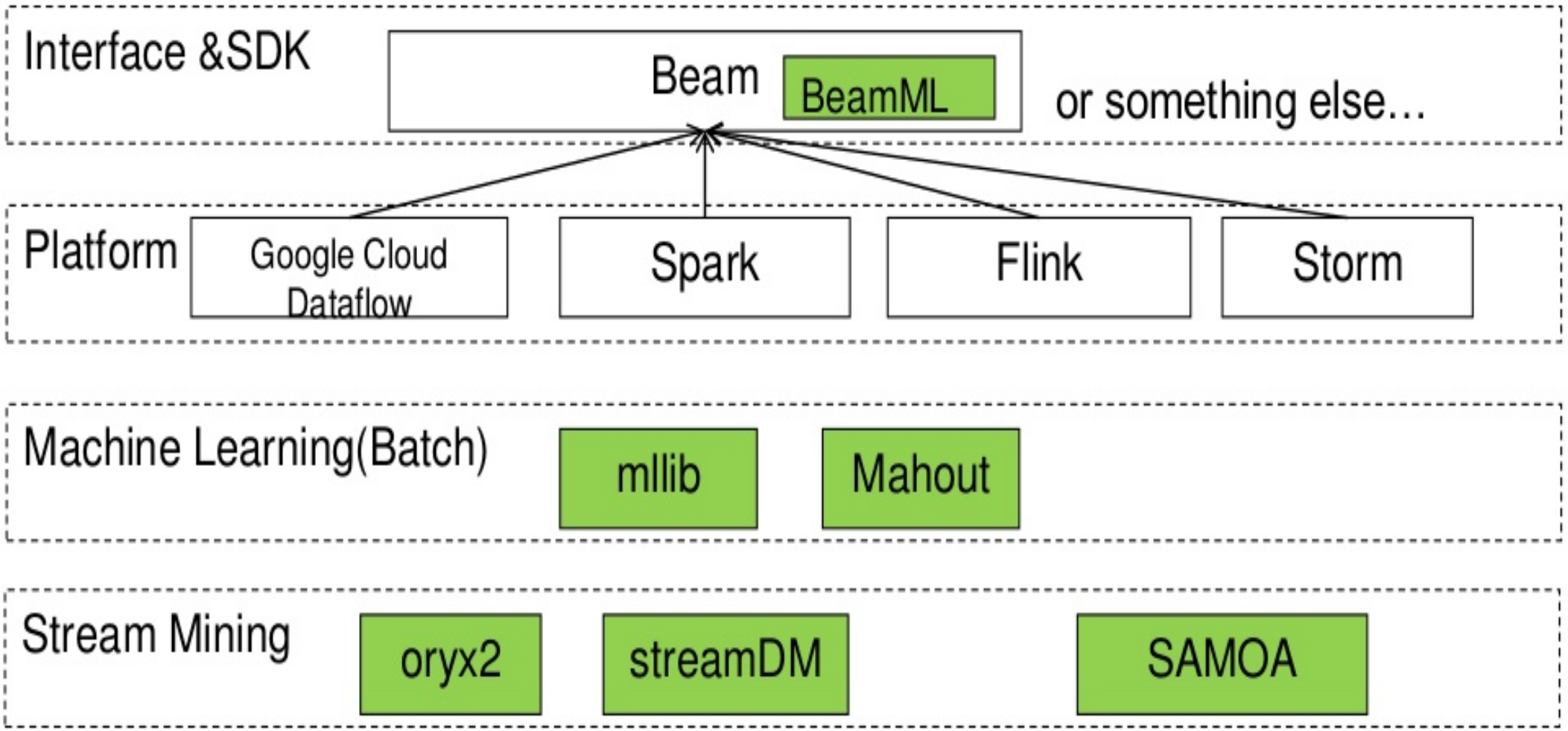
Next Release 31/12/16(support Spark 2.0)

- Random Forests
- Frequent Itemset Miner: IncMine

Something else…



Something else…



Contents

- streamDM: Stream Mining in Spark Streaming

  (Jianfeng Qian)

- Business scenarios in Huawei with Spark Steaming

  (Cheng He)

Special Challenges in Telecom Big Data Analytics

Most unstructured

- Sequence or Time series data;
- Lack of obvious features;

High velocity

- MBB & FBB
- 10G~400Gbps
- Processing capability needed:

1million – 10million pkt or event/s

Extremely imbalanced

- In many applications, positive labeled samples are rare;
- In one specific case: 1/50,000

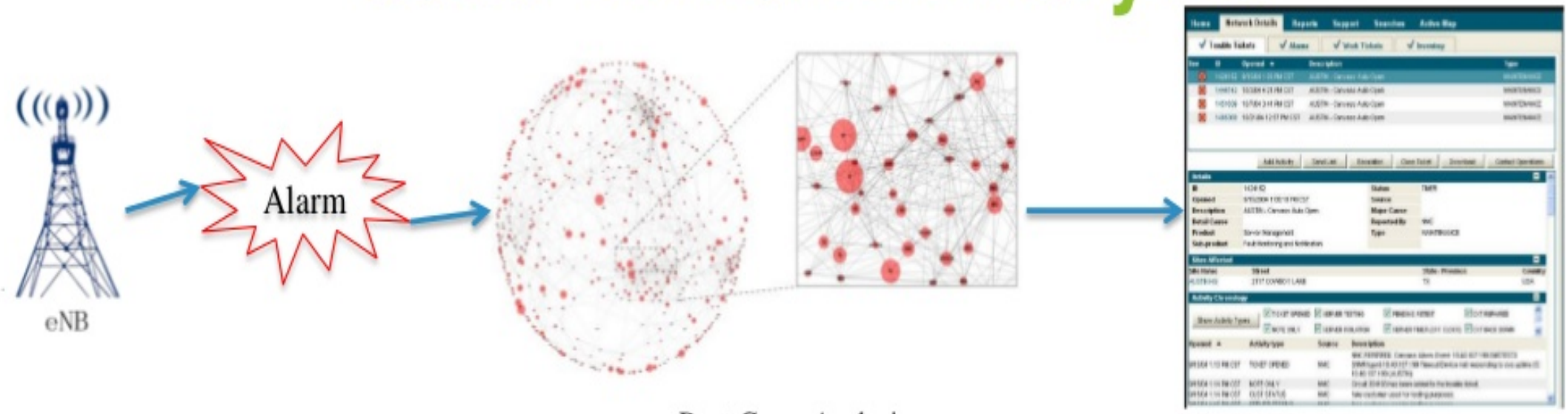
Big data volume & Variety

- Data from multiple domains
- Raw data: 100TB+ / day
- Data need to be analyzed: 5TB+ / day

Concept drift exists

- Because of the dynamic of Network configuration, users' behavior, etc;
- Fast adaptive, incremental modeling is essential;

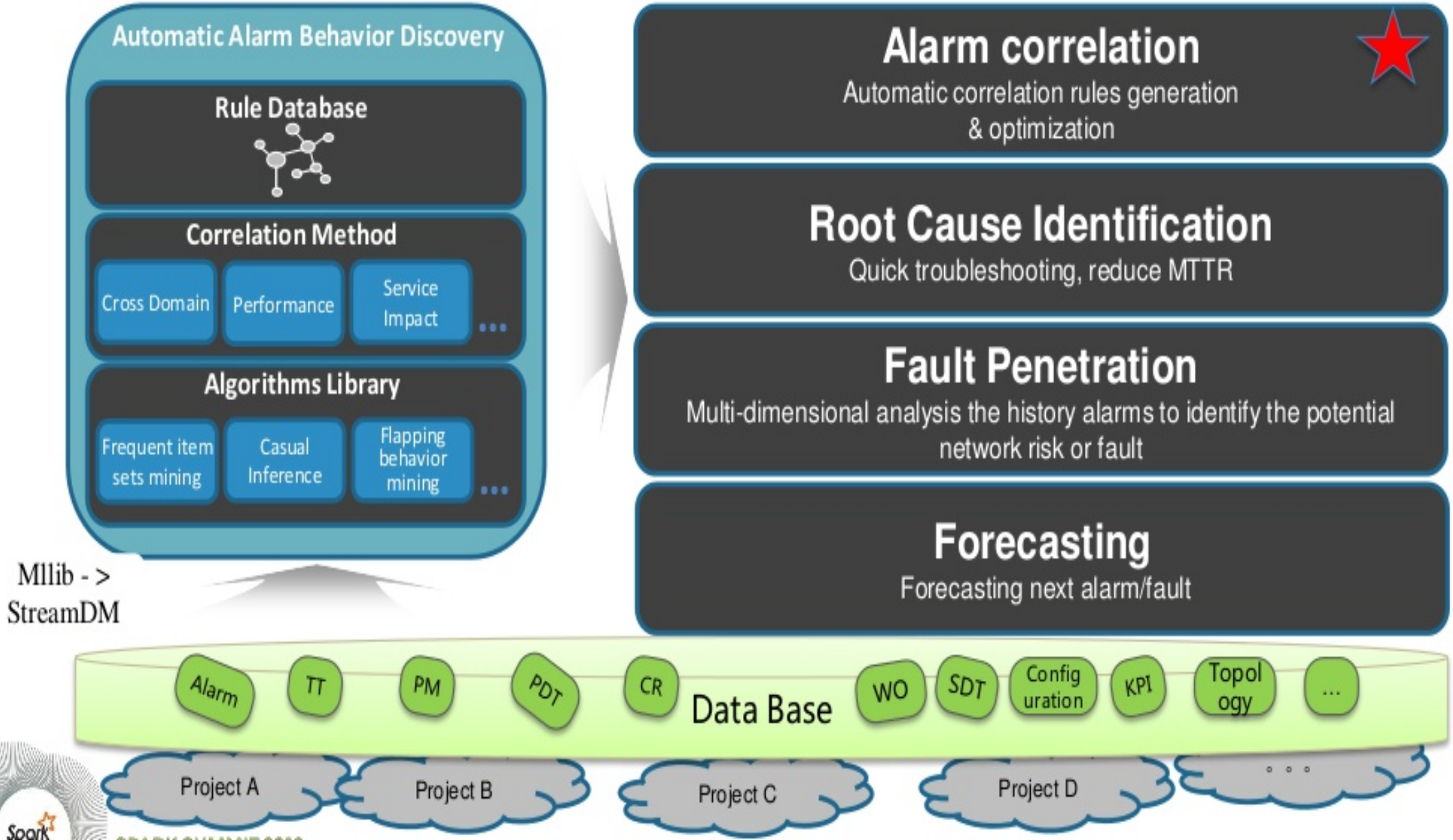
Case 1: Alarm Analysis



`40-100 Millions Alarms / day
`Change rapidly with unnoticed phenomenon

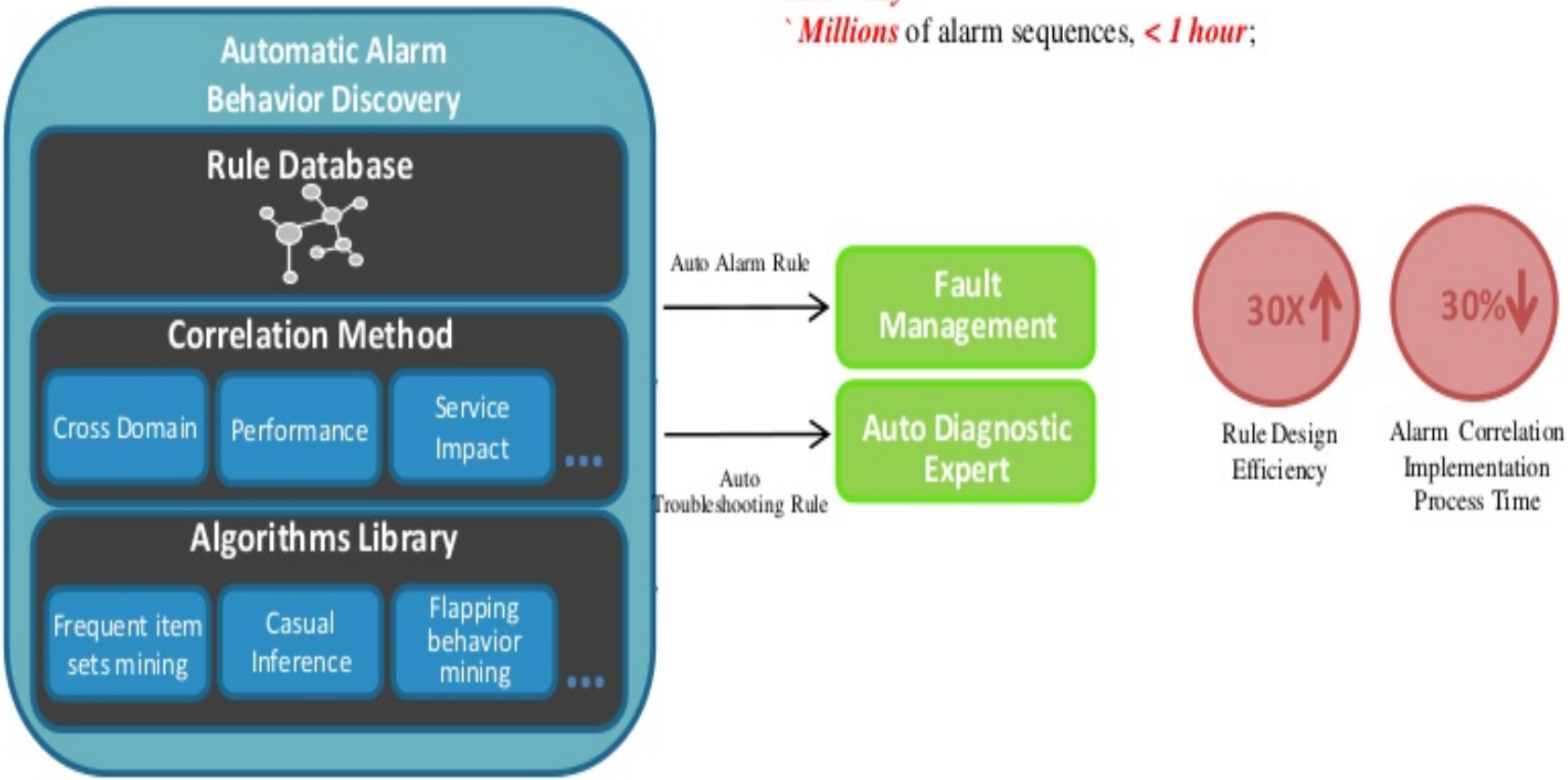
Field Operation / Maintenance

AABD : Automatic Alarm Behavior Discovery

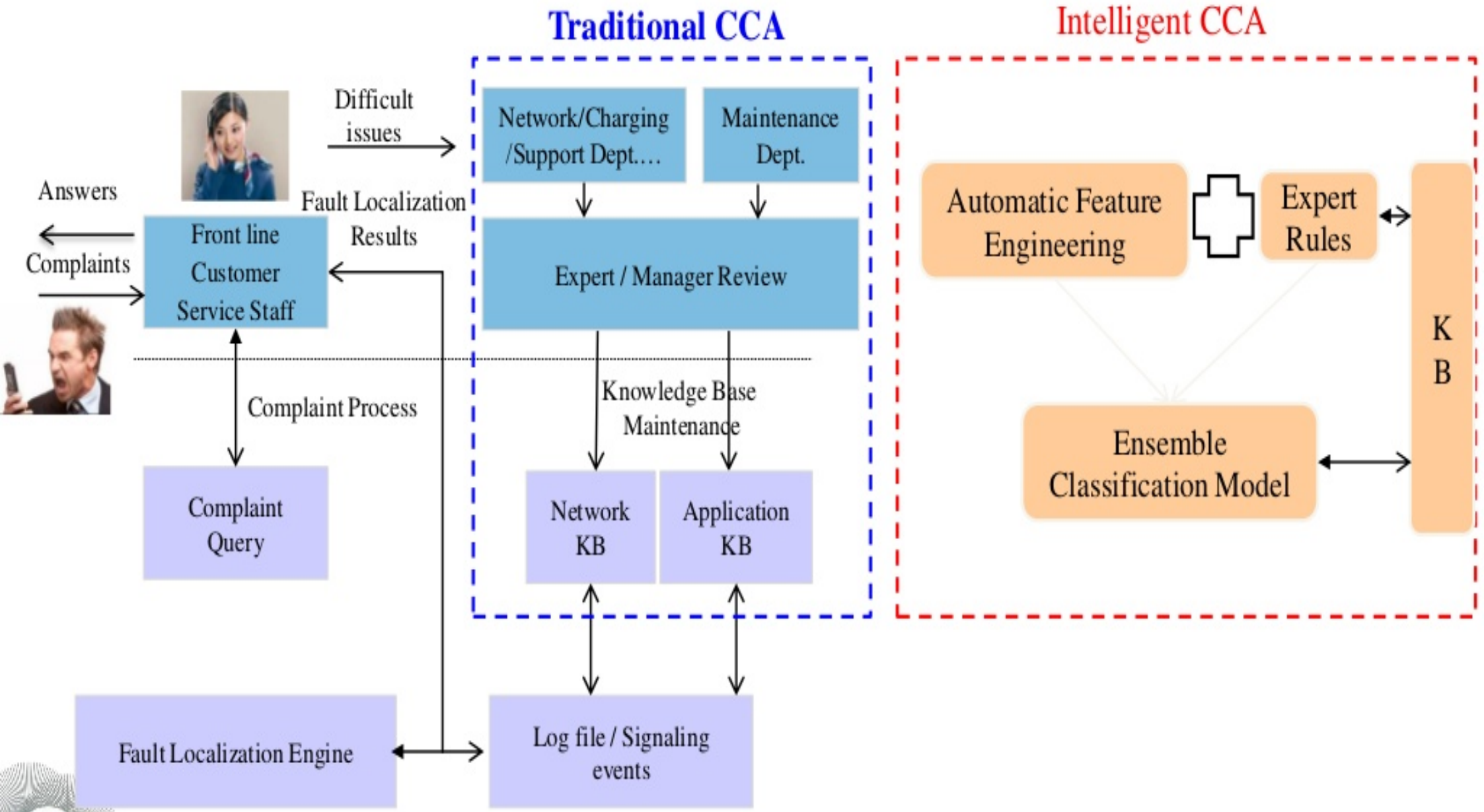


AABD : Results from practical applications

` Deployed in *13* Operator sites all over the world
` Improve efficiency of deployment from *5 man * month* to *3 man * day*
` *Millions* of alarm sequences, *< 1 hour*;

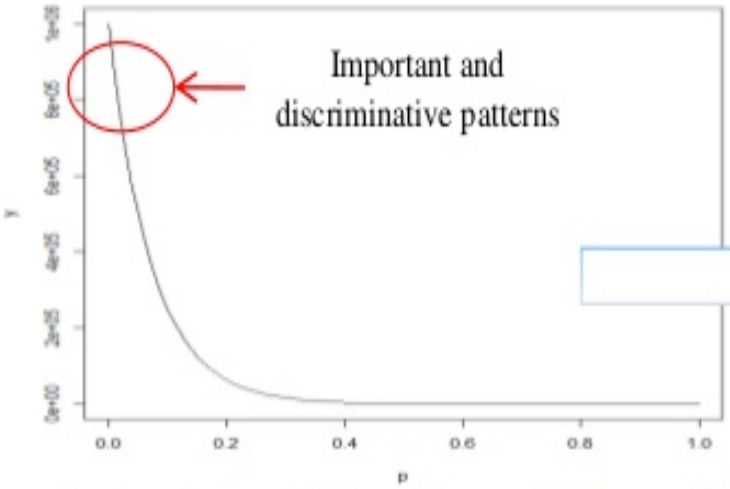


Case 2 : Fault Localization for Customer Care



Challenges on Discriminative sequence pattern mining

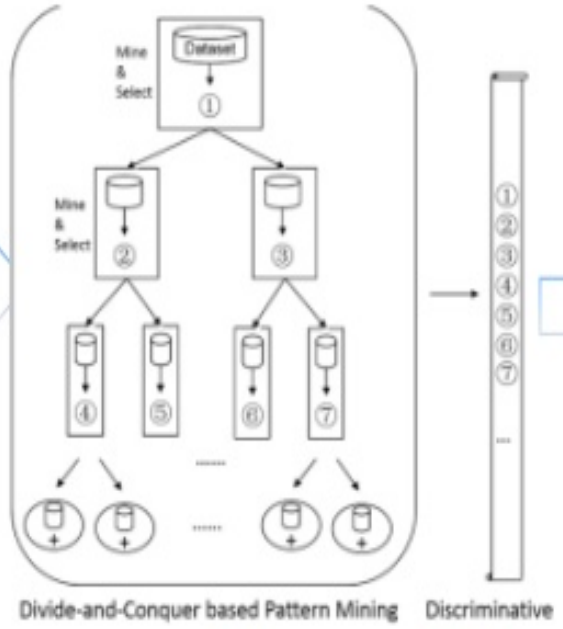
1. It's an NP-Hard problem to mine patterns out from massive sequence data for accurate classification
2. The upper bound of the pattern search space is $O(S^{s(1-p)})$
3. Usually, important patterns occur not so frequently, so we probably will search the total space of $O(S^s)$



1. Open dataset: 192 samples, p = 12%, no. of patterns 8600; p = 4%, No. of patterns 92,000;
2. In CCA, 538 samples, p=5%, No. of patterns 687402;

Original issues

1. Based on Divide-and-conquer, we can reduce the search space, even when we adopt larger p, we can still get discriminative patterns
2. Scale down ratio is $1/s^{sp}$

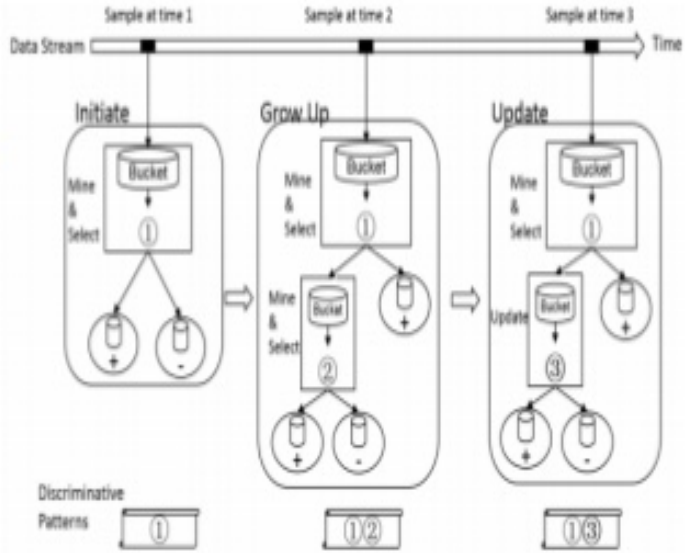


1. In CCA: , 538 samples, p=20%, No. of patterns: 29696;

State of art

$$G(F_{1-nl}) - G(F_{2-nl}) > \sqrt{\frac{R^2 \ln\frac{1}{\delta}}{2n}}$$

1. Online stream pattern mining, dynamically generate trees for pattern mining,, further reduce the search space to $O(2^l m^{m(1-p)})$;
2. Approximation analysis: balance complexity and accuracy;

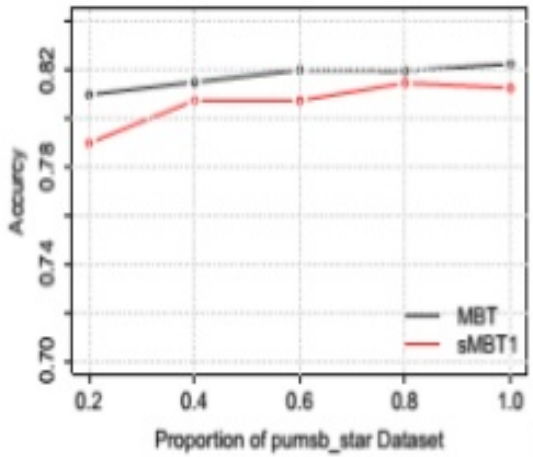


Research goals

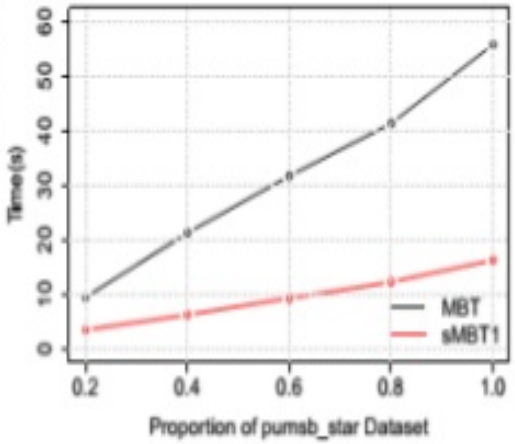
Ref: "Direct mining of discriminative and essential frequent

Experiment results

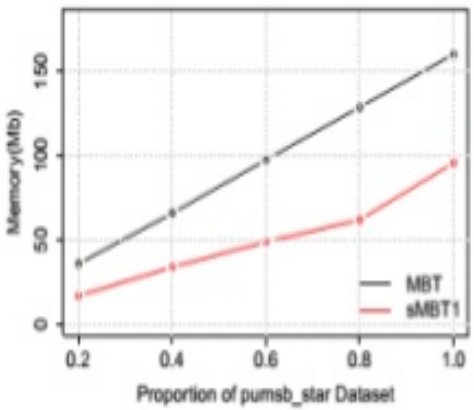
- Mining discriminative and essential patterns with extremely low global support directly.
- Better efficiency for stream sequential data mining while preserving accuracy.
- With the ability to detect concept drift quickly and adapt to new concept fast;



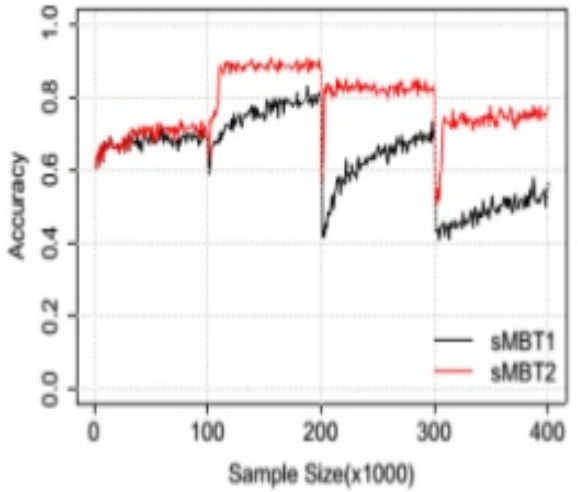
(a) Accuracy



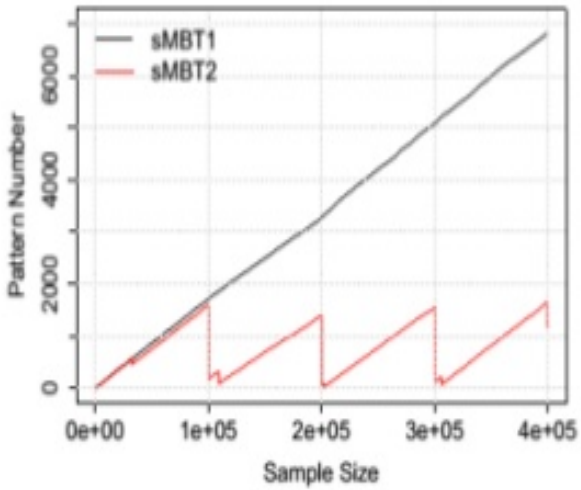
(b) Time



(c) Memory



(a) Accuracy



(b) Pattern Number

Design of Fault Localization Solution

`4-6 TB data / day;

`8M – 10M sequences;

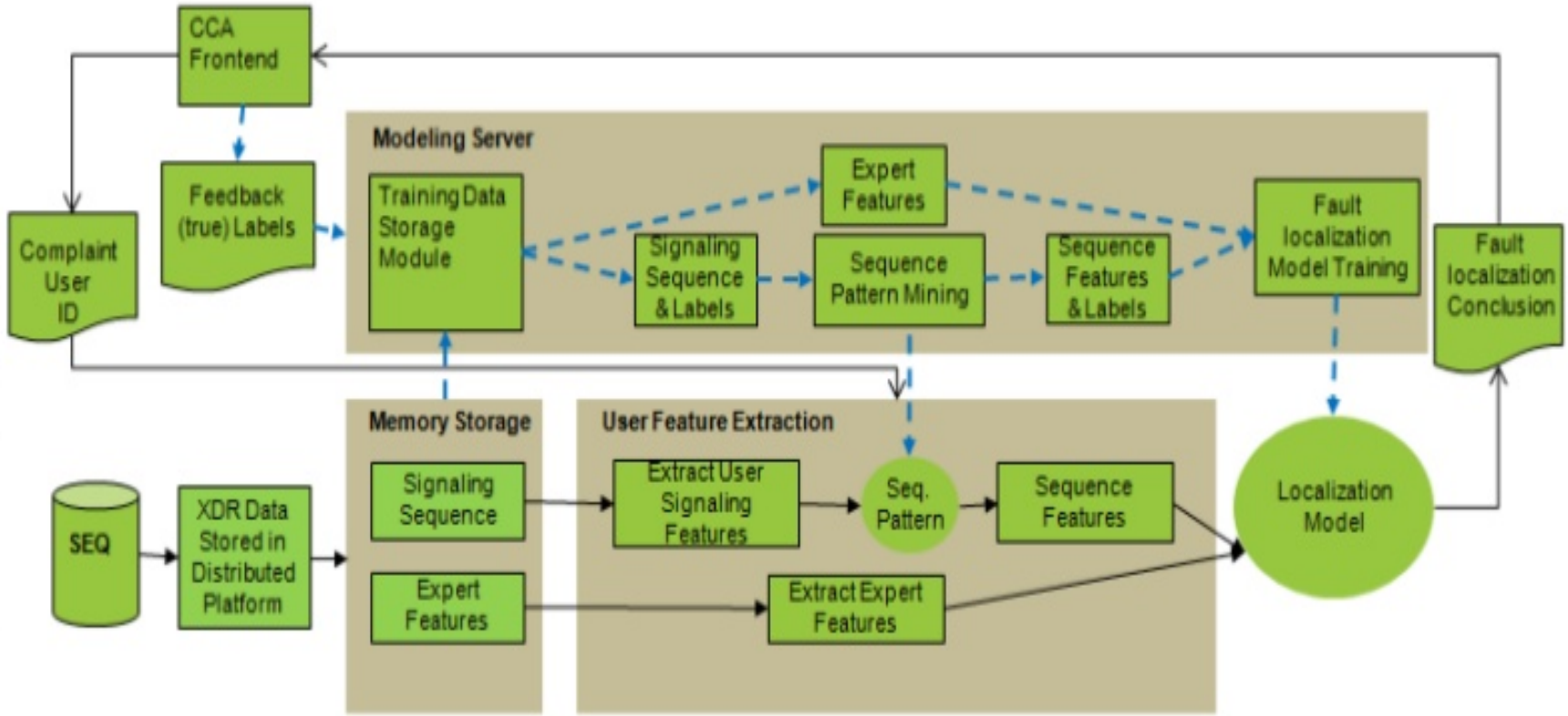
`E-2-E feedback support

< 3.6 sec;

`Model update 5 hours

Spark-streaming

Spark



Data preprocessing	Sequential Pattern Mining	Pattern Matching	Label Feedback
Sequence Encoding	Modeling	Online Classification	Model Update

Lessons from practical applications

- About spark:
  - ✓ Advantages:
    - Easy to use;
    - Perfect community & ecosystem;
  - ✓ Limitations:
    - Delay;
    - Throughput;

Scenarios & Apps:

- App recommendation system;
  - ✓100M+ customers;
  - ✓30M – 100M features;
- Anti-DDoS Solution;
  - ✓4M - 10M flow / sec;

- About big data analytics in Telecom networks:
  - ✓ Efficient sequential pattern mining framework
  - ✓ Deep reinforcement learning;
  - ✓ Robust ML / AI & Domain Knowledge;
  - ✓ Close-loop evaluation;

Huawei Innovation Research Program

- The Huawei Innovation Research Program (HIRP) provides funding opportunities to leading **universities** and **research institutes** conducting innovative research in communication technology, computer science, engineering, and related fields. HIRP seeks to identify and support world-class, full-time faculty members pursuing innovation of mutual interest. Outstanding HIRP winners may be invited to establish further long-term research collaboration with Huawei.
- Call for Proposals for Big Data & Artificial Intelligence
  - HIRPO20160606: Novel Algorithm Design and Use Cases for Data Stream Mining based on streamDM

https://innovationresearch.huawei.com/IPD/hirp/portal/index.html

Join us to build a better connected world

THANK YOU

jianfeng.qian@outlook.com, hecheng@huawei.com