DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning

(to appear in CCS'17)

Min Du, Feifei Li, Guineng Zheng, Vivek Srikumar
University of Utah

Background

```
15/07/31 12:20:17 INFO SparkContext: Running Spark version 1.3.0
15/07/31 12:20:18 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
15/07/31 12:20:18 INFO SecurityManager: Changing view acls to: zhouliang
15/07/31 12:20:18 INFO SecurityManager: Changing modify acls to: zhouliang
15/07/31 12:20:18 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users
with view permissions: Set(zhouliang); users with modify permissions: Set(zhouliang)
15/07/31 12:20:18 INFO Slf4jLogger: Slf4jLogger started
15/07/31 12:20:18 INFO Remoting: Starting remoting
15/07/31 12:20:18 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://
sparkDriver@head:60626]
15/07/31 12:20:18 INFO Utils: Successfully started service 'sparkDriver' on port 60626.
15/07/31 12:20:18 INFO SparkEnv: Registering MapOutputTracker
15/07/31 12:20:18 INFO SparkEnv: Registering BlockManagerMaster
15/07/31 12:20:18 INFO DiskBlockManager: Created local directory at /tmp/spark-3799bc3c-5275-499c-8b89-
fa93e6b0131e/blockmgr-f7e603b7-c8c3-4faf-be6c-2af1620dc1e3
15/07/31 12:20:18 INFO MemoryStore: MemoryStore started with capacity 10.4 GB
15/07/31 12:20:19 INFO HttpFileServer: HTTP File server directory is /tmp/spark-c01a992b-
d9d3-4751-8f2e-05c2a64cb329/httpd-b9f5fc86-0f7c-434c-aed4-20f27b9b3731
15/07/31 12:20:19 INFO HttpServer: Starting HTTP Server
15/07/31 12:20:19 INFO Server: jetty-8.y.z-SNAPSHOT
15/07/31 12:20:19 INFO AbstractConnector: Started SocketConnector@0.0.0.0:43664
15/07/31 12:20:19 INFO Utils: Successfully started service 'HTTP file server' on port 43664.
15/07/31 12:20:19 INFO SparkEnv: Registering OutputCommitCoordinator
15/07/31 12:20:19 INFO Server: jetty-8.y.z-SNAPSHOT
15/07/31 12:20:19 INFO AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
15/07/31 12:20:19 INFO Utils: Successfully started service 'SparkUI' on port 4040.
15/07/31 12:20:19 INFO SparkUI: Started SparkUI at http://head:4040
15/07/31 12:20:19 INFO SparkContext: Added JAR file:/home/zhouliang/experiments/knn-join/./target/
scala-2.10/knn-join_2.10-1.0.jar at http://192.168.1.2:43664/jars/knn-join_2.10-1.0.jar with timestamp
1438316419295
15/07/31 12:20:19 INFO AppClient$ClientActor: Connecting to master akka.tcp://sparkMaster@head:7077/user/
Master...
15/07/31 12:20:19 INFO SparkDeploySchedulerBackend: Connected to Spark cluster with app ID
```

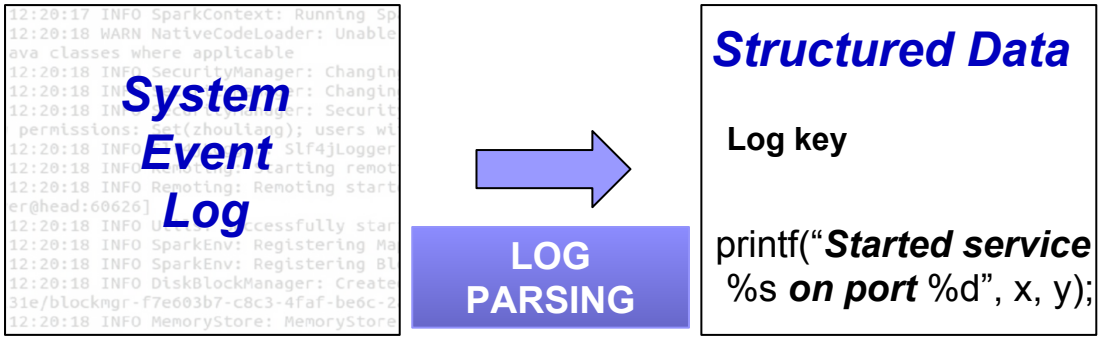
Background


System Event Log
Exist practically on every computer system!
Automatic Analysis?

Background

Started service A on port 80
Executor updated: app-1 is now LOADING
……

Background



Started service A on port 80
Executor updated: app-1 is now LOADING
……

Started service * on port *
Executor updated: * is now LOADING
……

Background

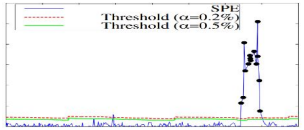
LOG PARSING

Structured Data

Log key

printf("***Started service*** %s ***on port*** %d", x, y);

Anomaly Detection



LOG ANALYSIS

Started service A on port 80
Executor updated: app-1 is now LOADING
……

Started service * on port *
Executor updated: * is now LOADING
……

DeepLog

log message (log key underlined)	log key	parameter value vector
t_1 Deletion of **file1** complete	k_1	$[t_1 - t_0, \text{file1Id}]$
t_2 Took **0.61** seconds to deallocate network ...	k_2	$[t_2 - t_1, 0.61]$
t_3 VM Stopped (Lifecycle Event)	k_3	$[t_3 - t_2]$
...	...	...

DeepLog

log message (log key underlined)	log key	parameter value vector
t_1 Deletion of **file1** complete	k_1	$[t_1 - t_0, \text{file1Id}]$
t_2 Took **0.61** seconds to deallocate network ...	k_2	$[t_2 - t_1, 0.61]$
t_3 VM Stopped (Lifecycle Event)	k_3	$[t_3 - t_2]$
...	...	...

log message **log key** **parameters**

Deletion of file1 complete. ⇨

Deletion of file2 complete. ⇨

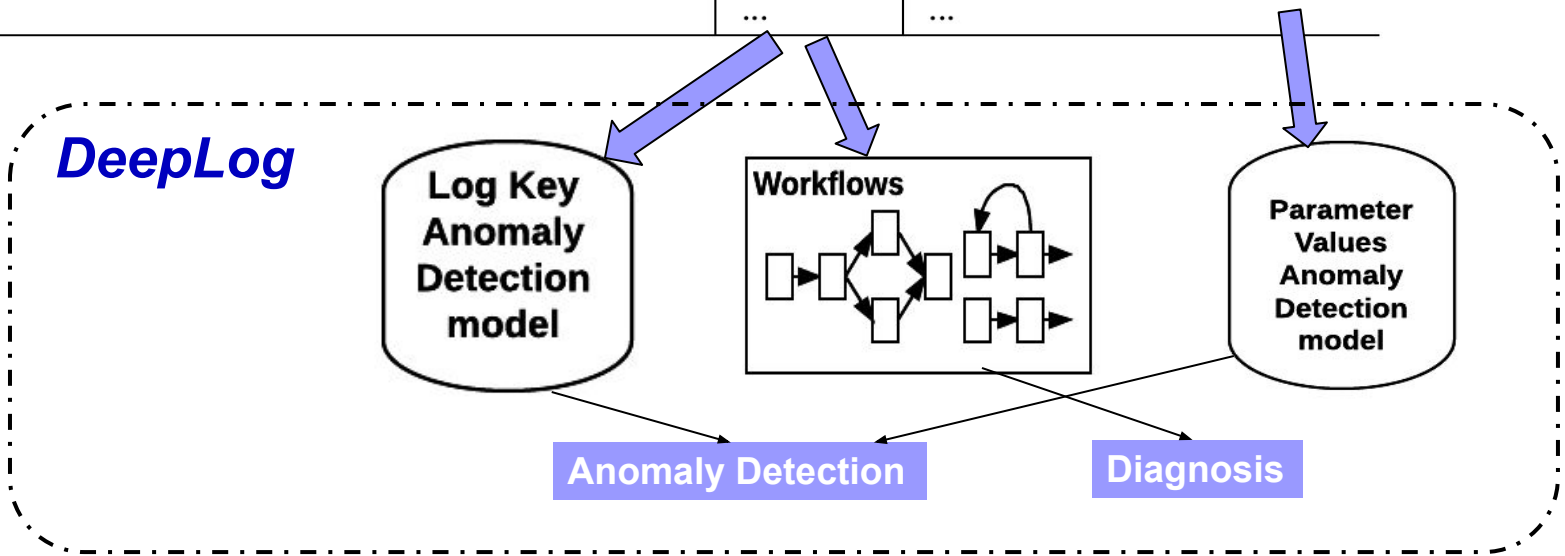
SPELL
A streaming log parser published in ICDM'16

⇨ *Deletion of file1 complete.* [ ]

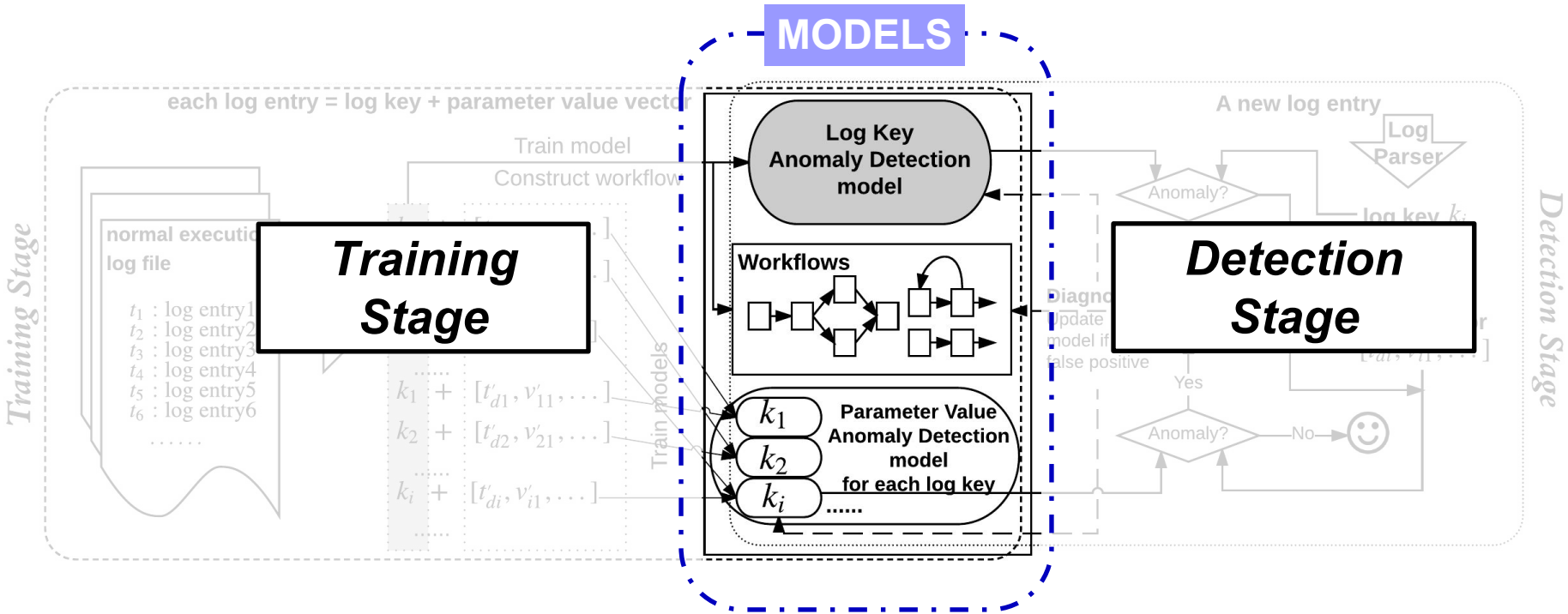
⇨ *Deletion of * complete.* [*file2*]

DeepLog

log message (log key underlined)	log key	parameter value vector
t_1 Deletion of **file1** complete	k_1	$[t_1 - t_0, \text{file1Id}]$
t_2 Took **0.61** seconds to deallocate network ...	k_2	$[t_2 - t_1, 0.61]$
t_3 VM Stopped (Lifecycle Event)	k_3	$[t_3 - t_2]$
...	...	...



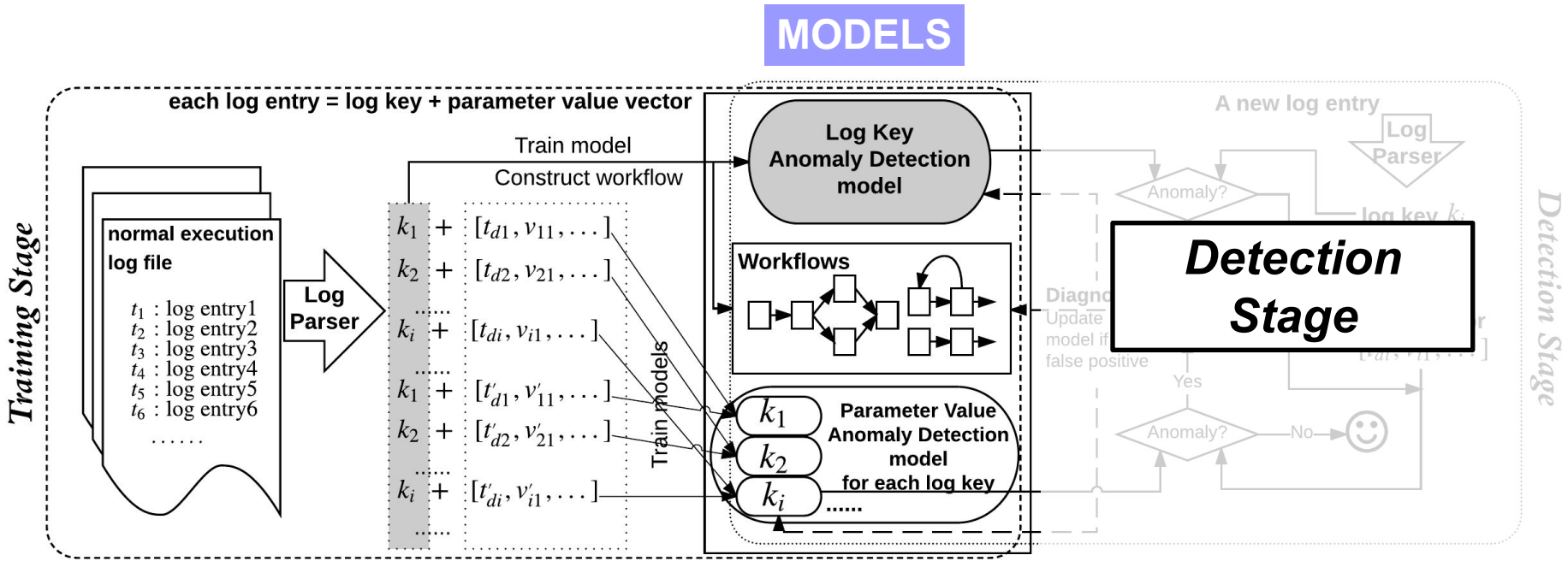
DeepLog Architecture



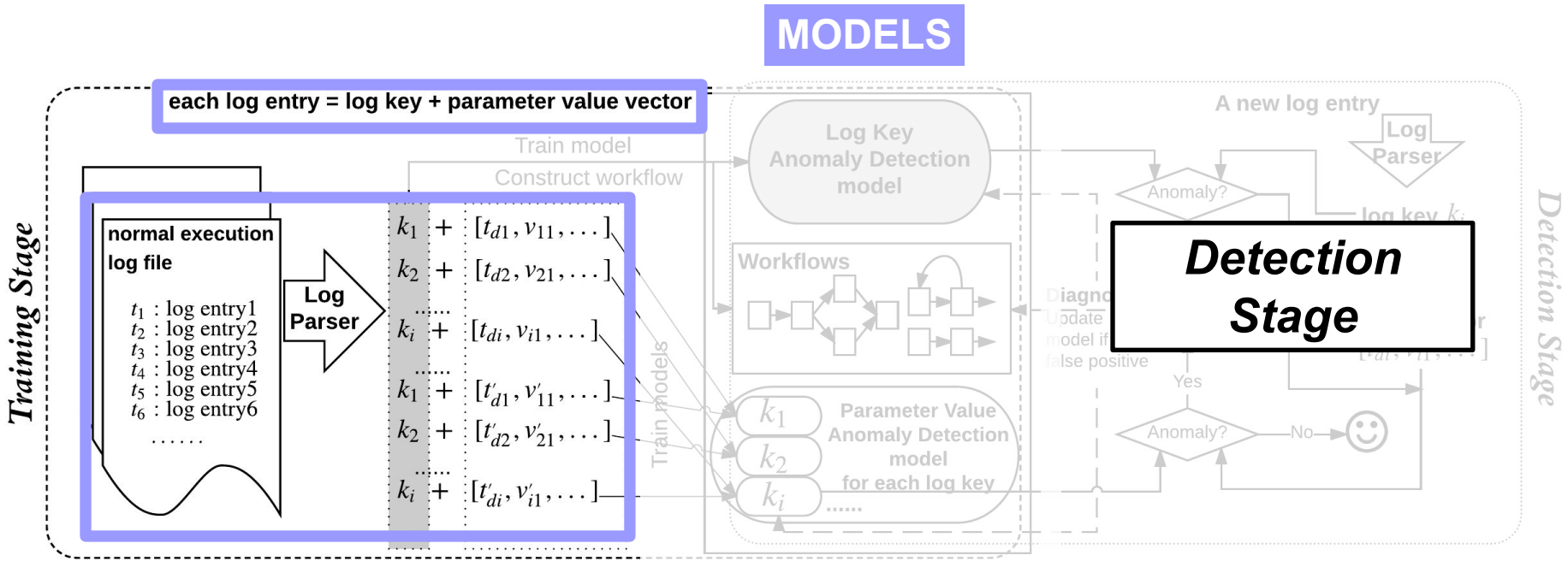
Training Stage

Detection Stage

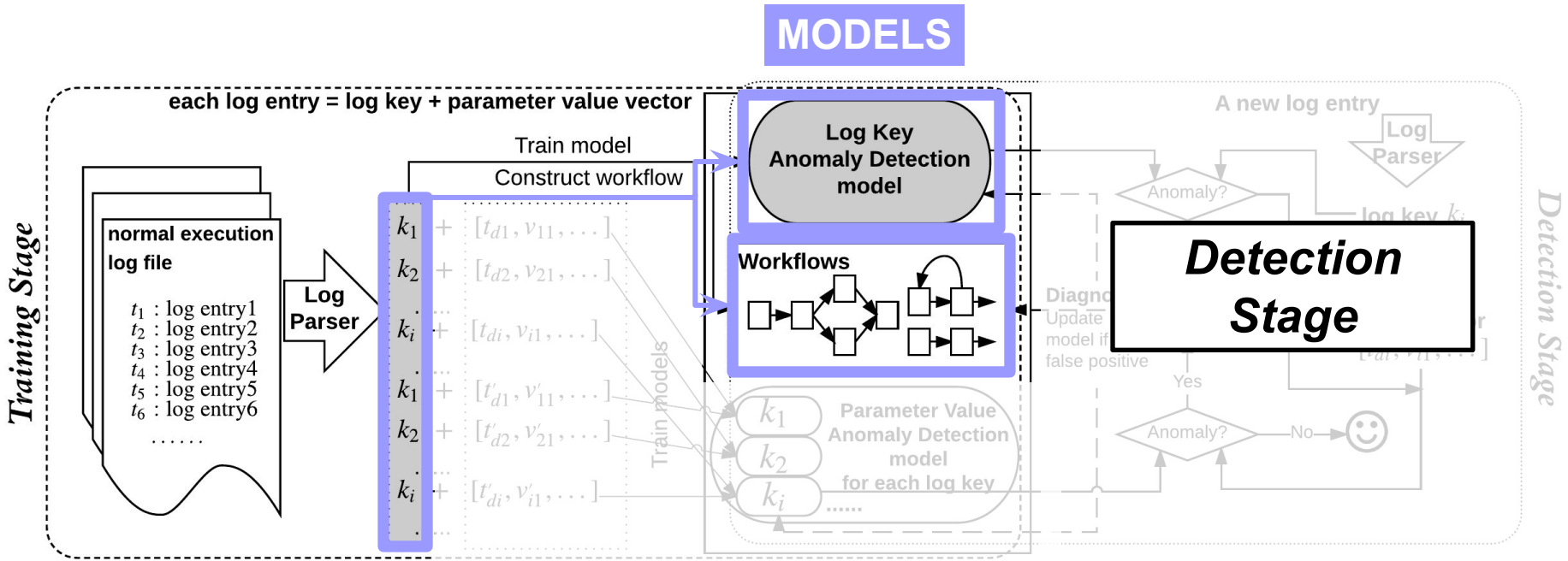
DeepLog Architecture



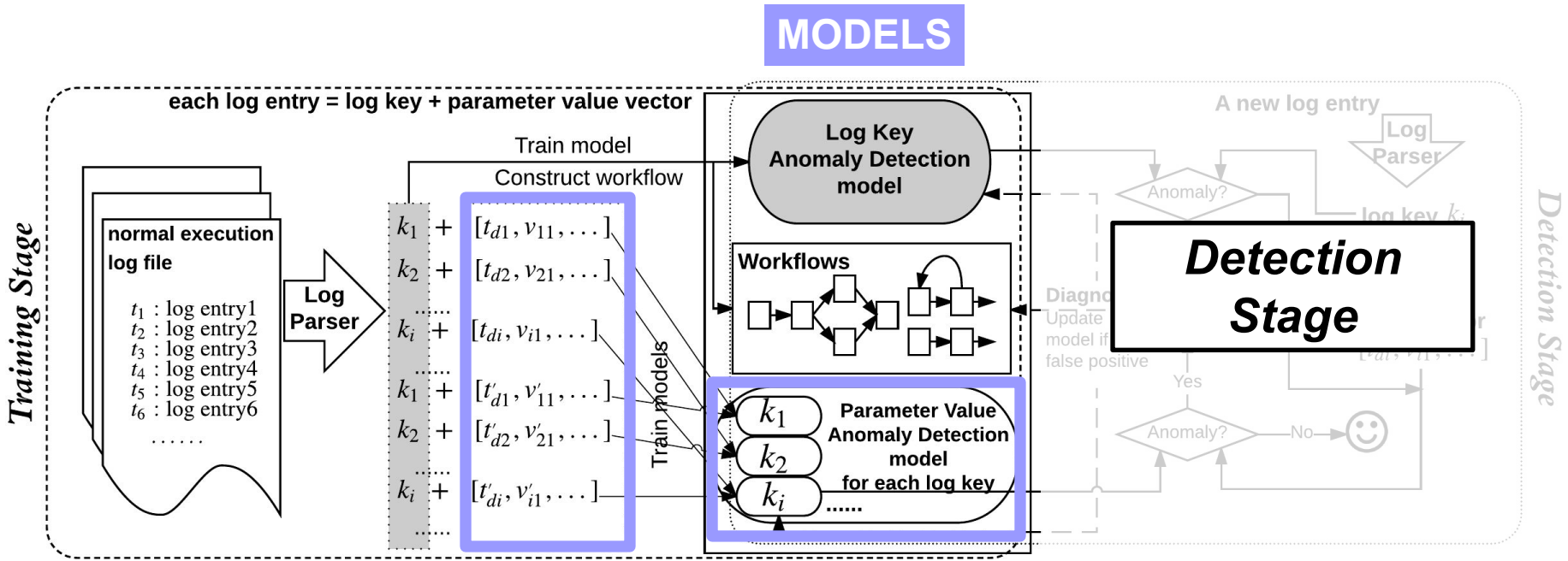
DeepLog Architecture



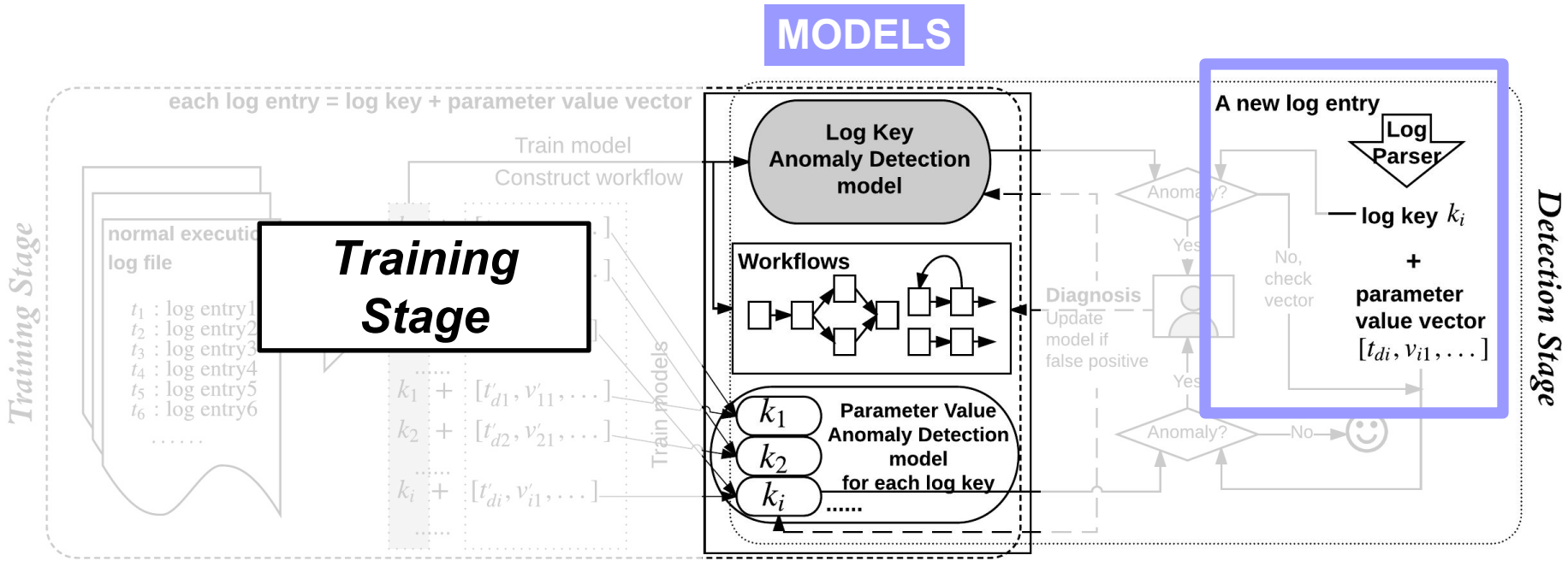
DeepLog Architecture



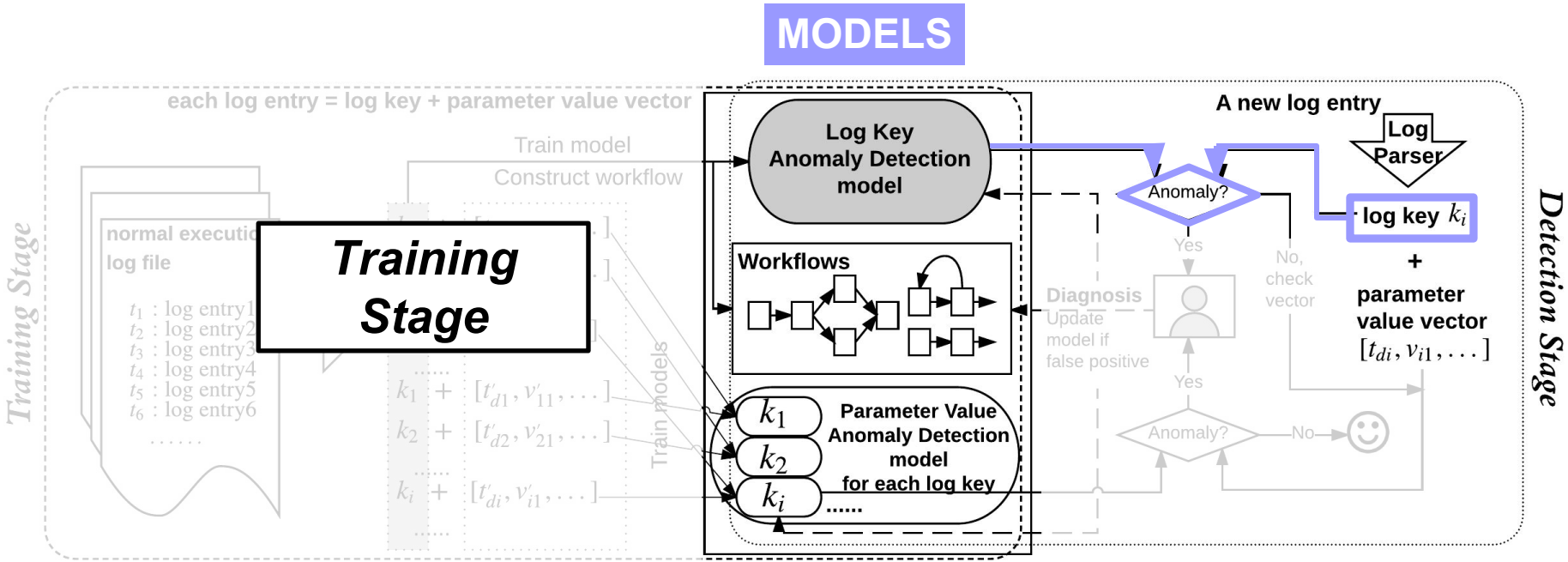
DeepLog Architecture



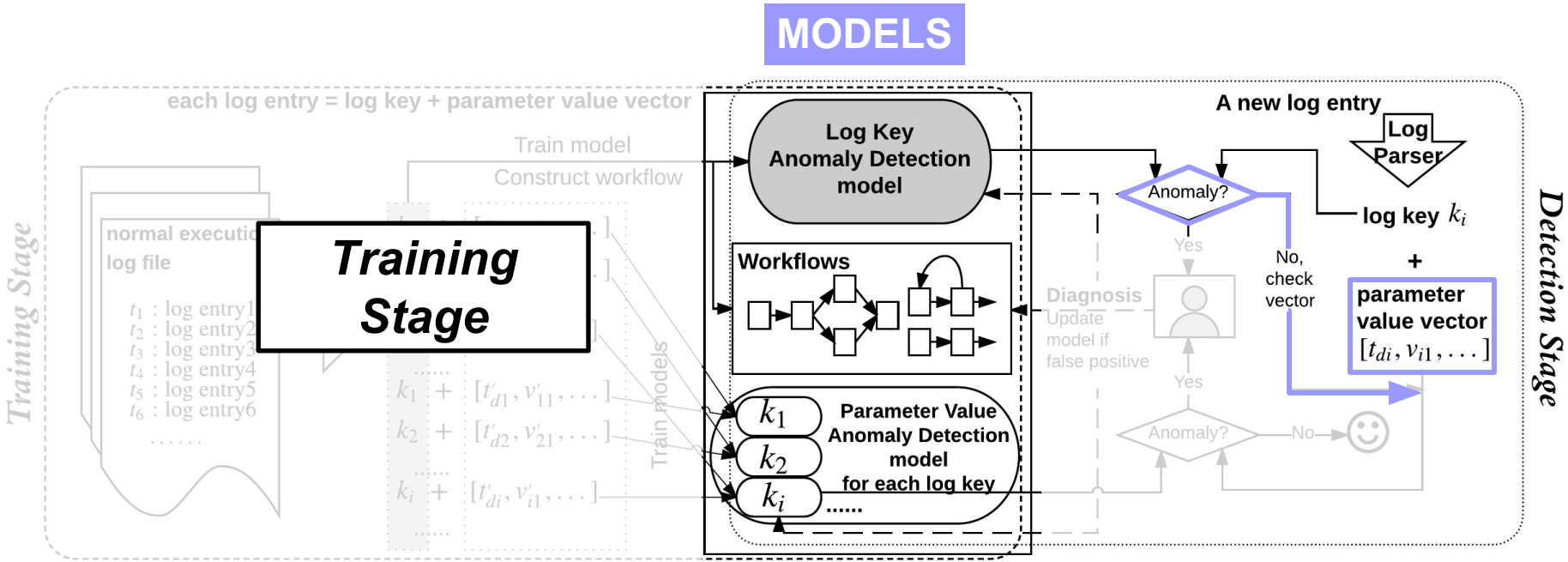
DeepLog Architecture



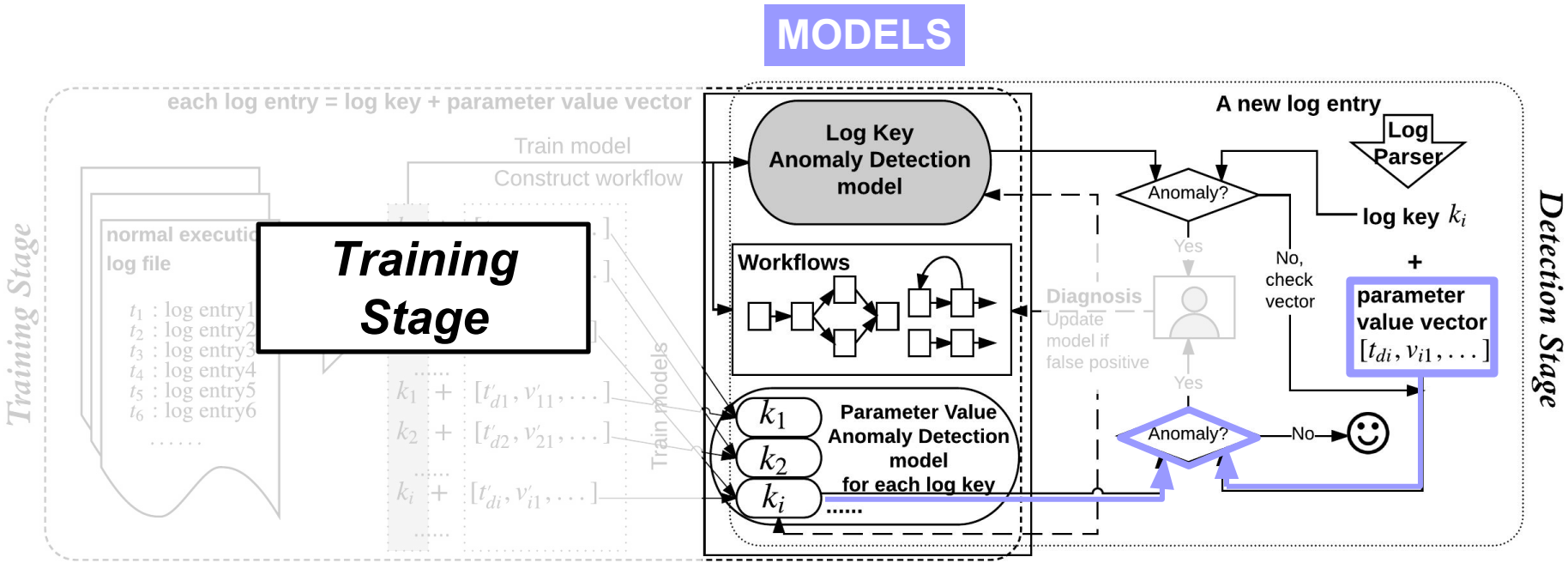
DeepLog Architecture



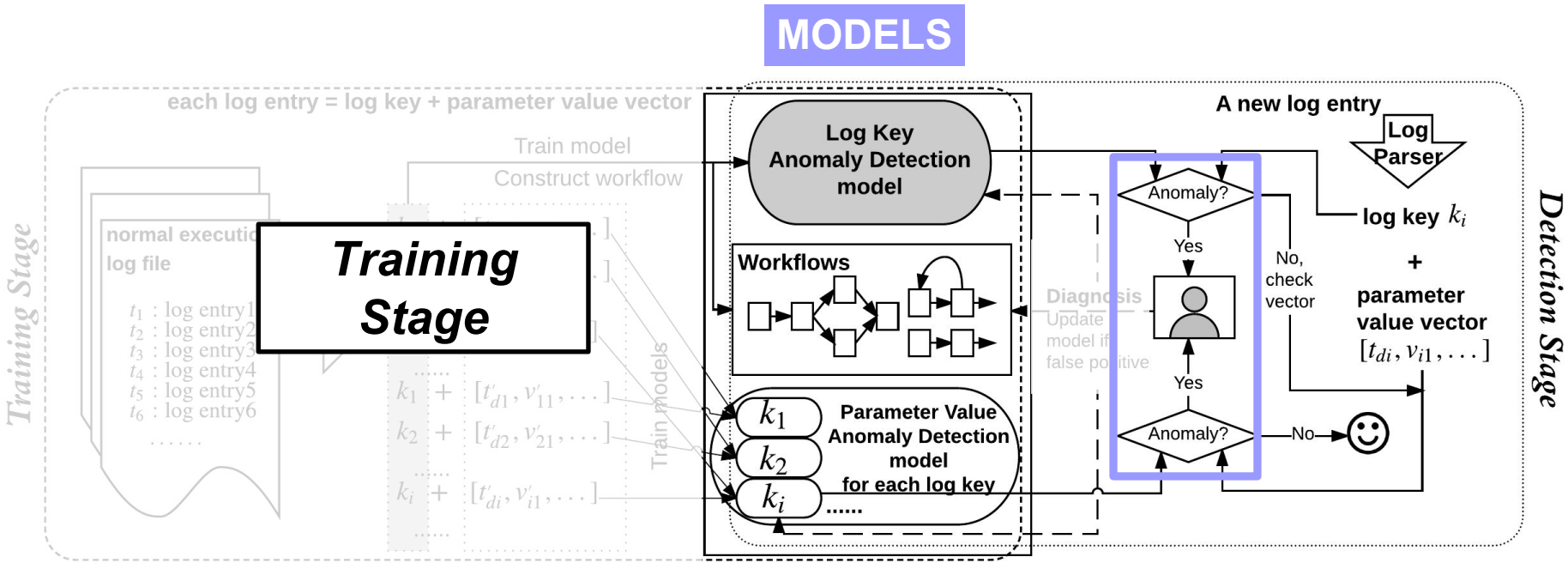
DeepLog Architecture



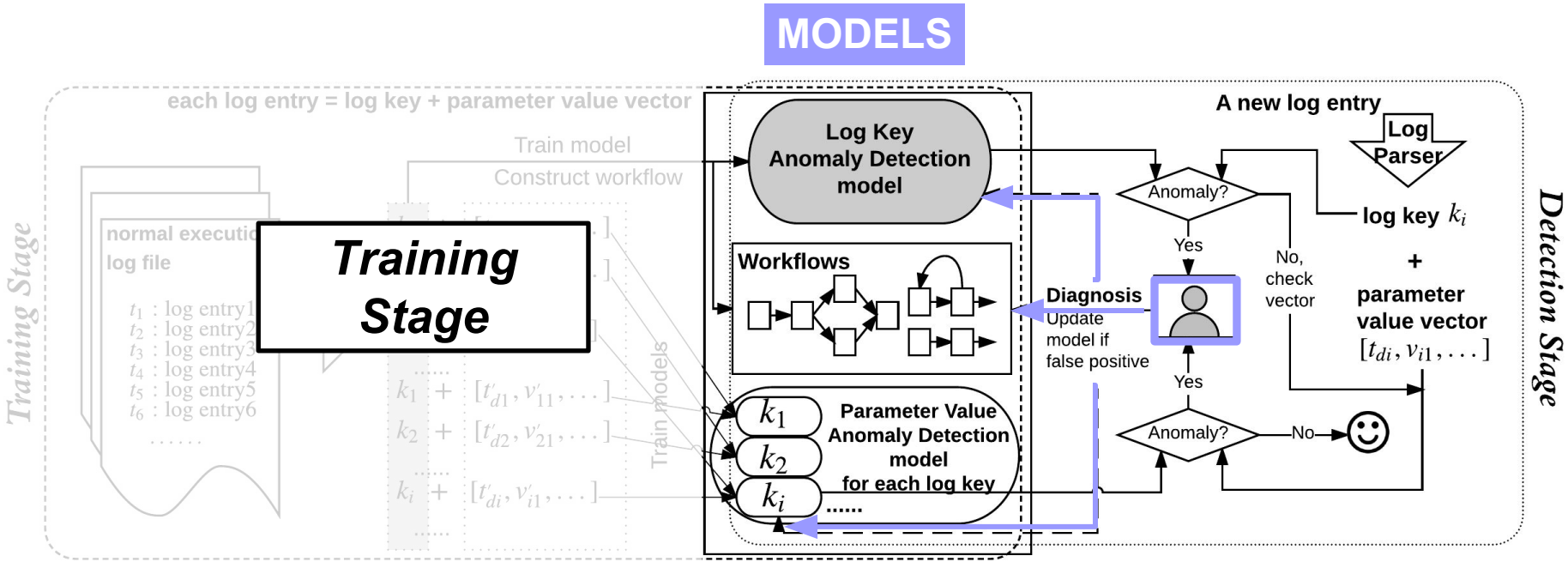
DeepLog Architecture



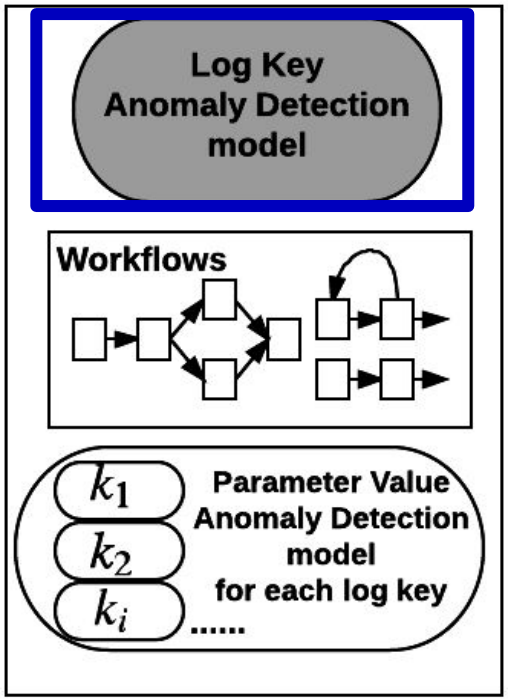
DeepLog Architecture



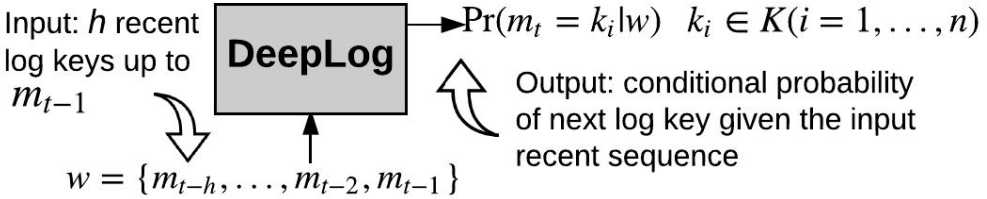
DeepLog Architecture



Log Key Anomaly Detection model

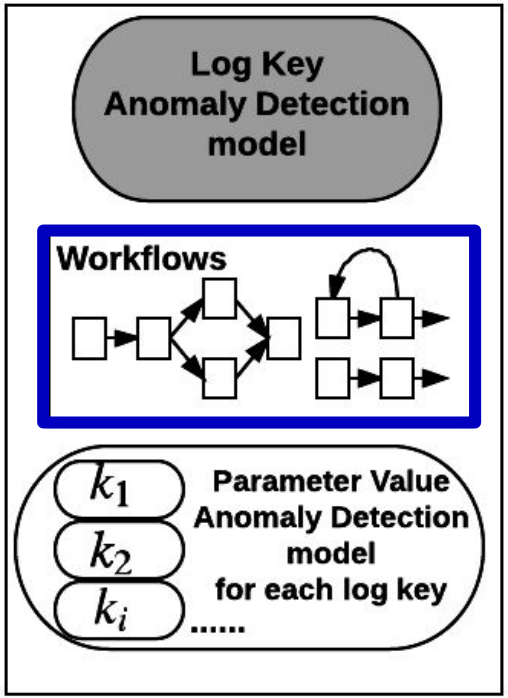


Use long short-term memory (LSTM) architecture



In detection stage, DeepLog checks if the actual next log key is among its top *g* probable predictions.

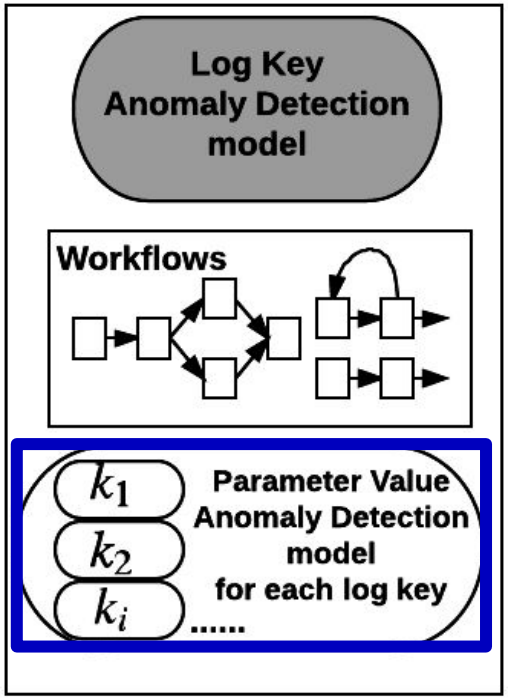
Workflow Construction



Method 1: Using LSTM prediction probabilities

Method 2: Using co-occurrence matrix

Parameter Value Anomaly Detection model



Example:

Log messages of a particular log key:

t_2: *Took* 0.61 *seconds to deallocate network* ...

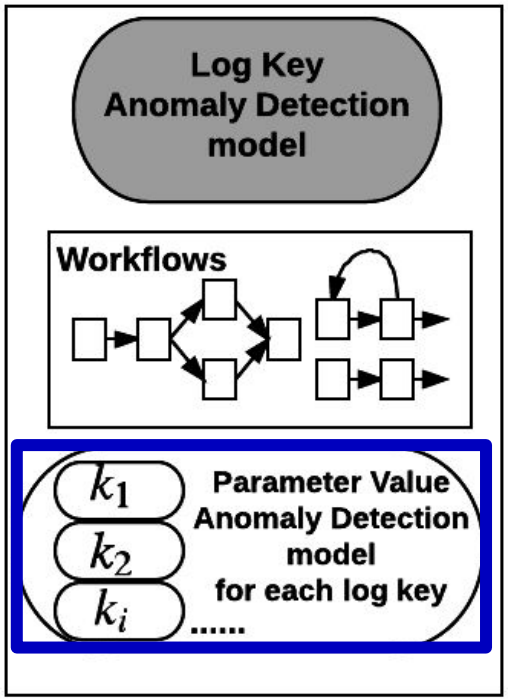
t'_2: *Took* 1.1 *seconds to deallocate network* ...

....

Parameter value vectors overtime:

[t_2- t_1, 0.61], [t'_2- t'_1, 1.1], ...

Parameter Value Anomaly Detection model



Example:

Log messages of a particular log key:

t_2: *Took 0.61 seconds to deallocate network ...*

t'_2: *Took 1.1 seconds to deallocate network ...*

....

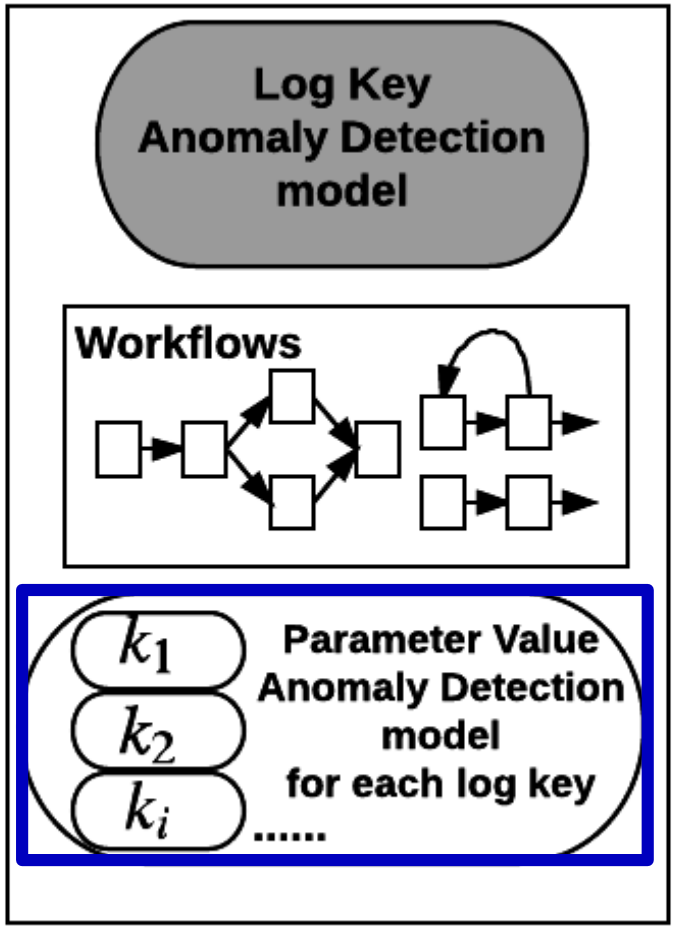
Parameter value vectors overtime:

$[t_2 - t_1, 0.61]$, $[t'_2 - t'_1, 1.1]$, ...

Multi-variate time series data anomaly detection problem!

--- Leverage LSTM to check reconstruction error.

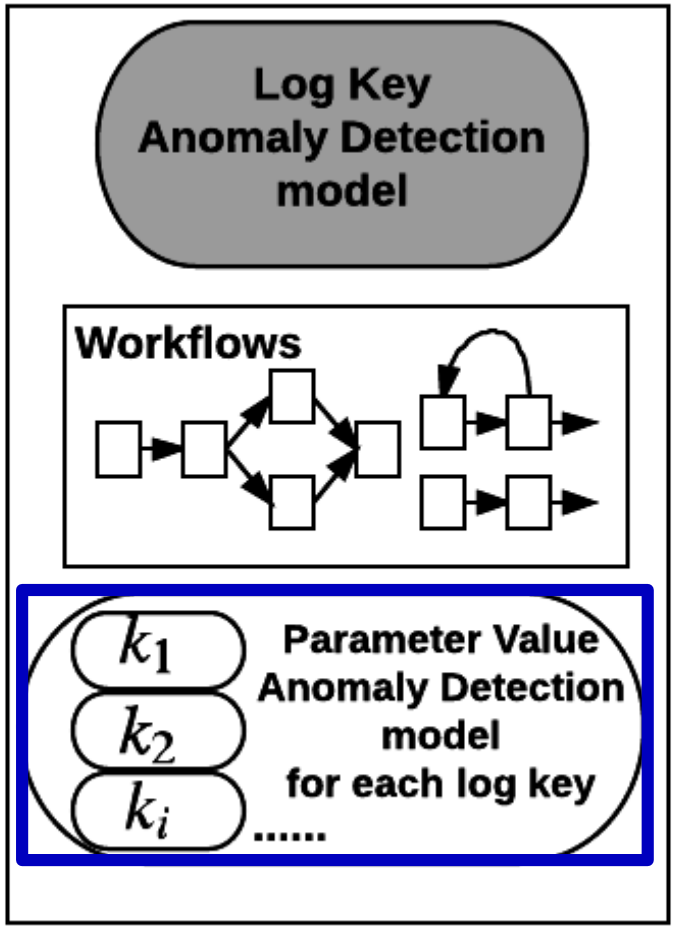
Parameter Value Anomaly Detection model



Multi-variate time series data anomaly detection problem

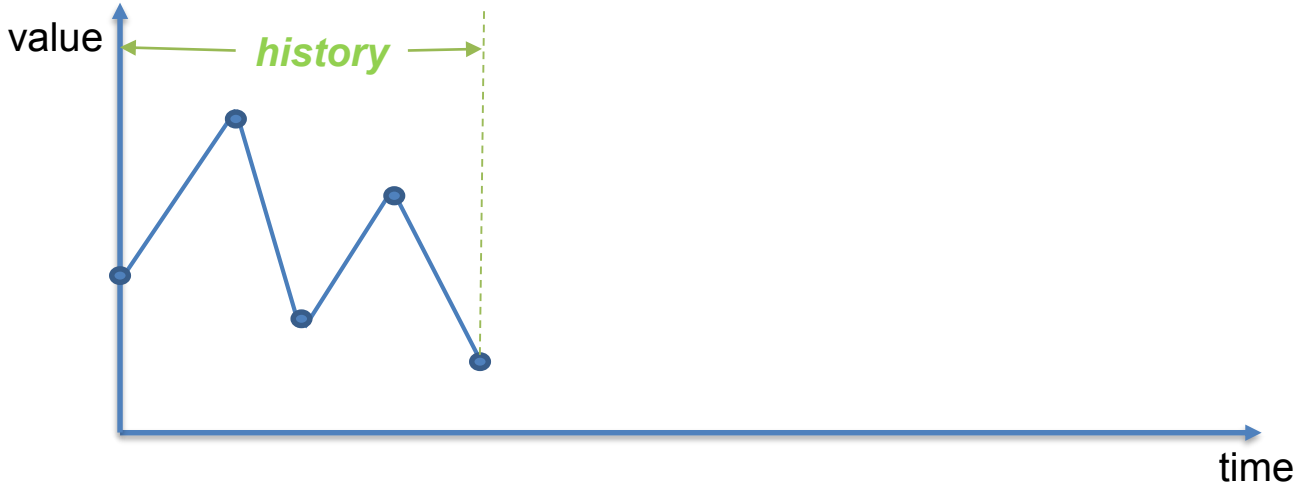
- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

Parameter Value Anomaly Detection model

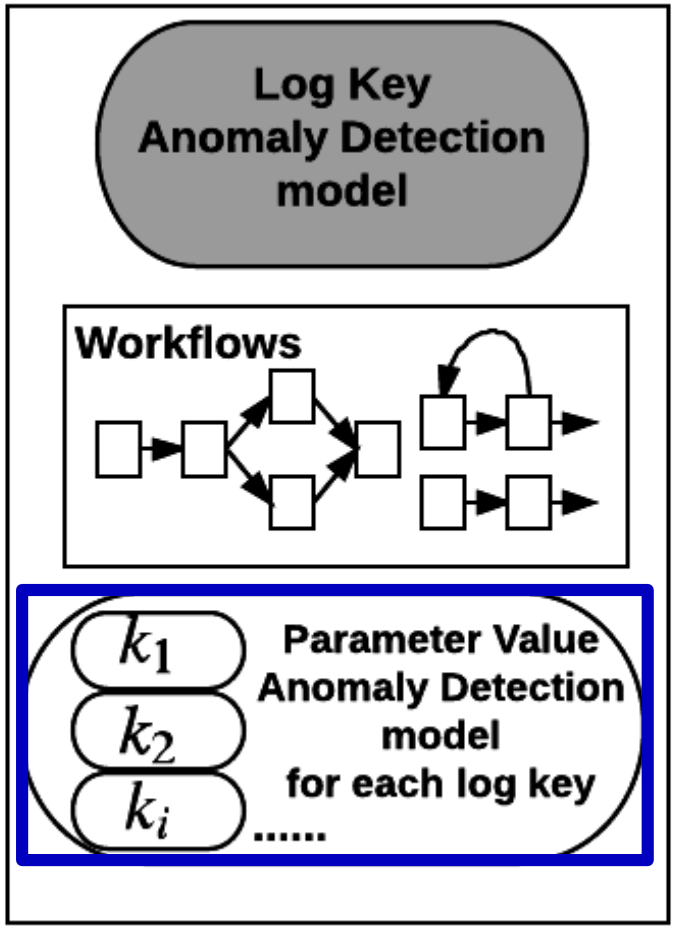


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

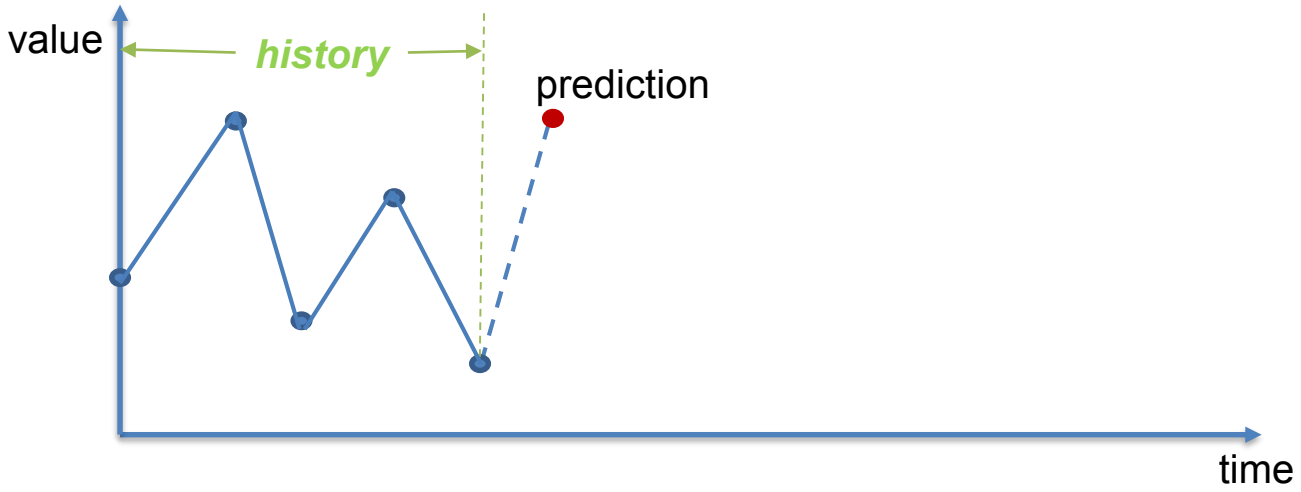


Parameter Value Anomaly Detection model

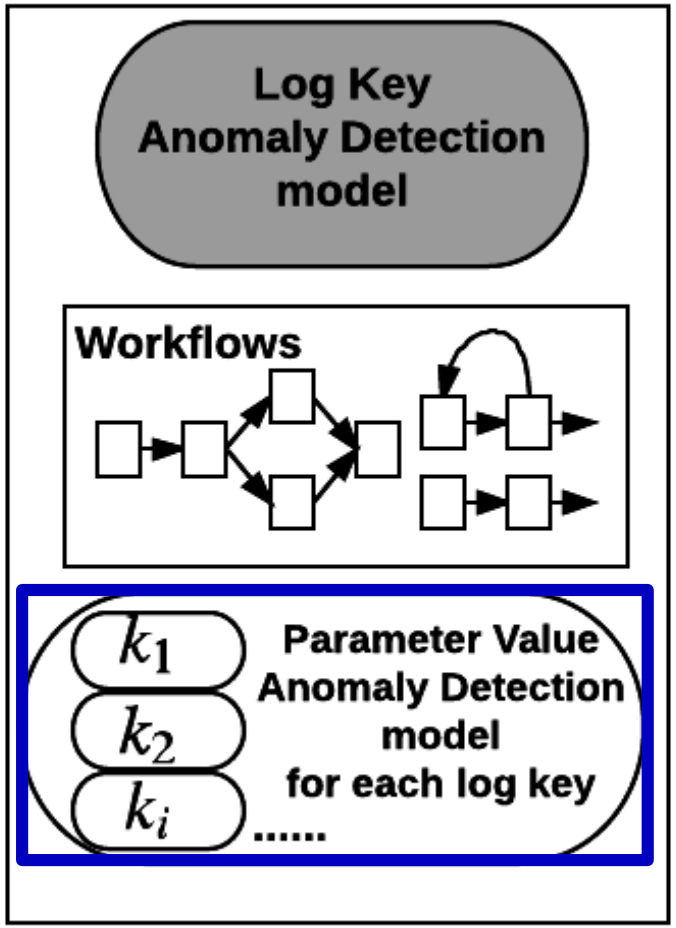


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

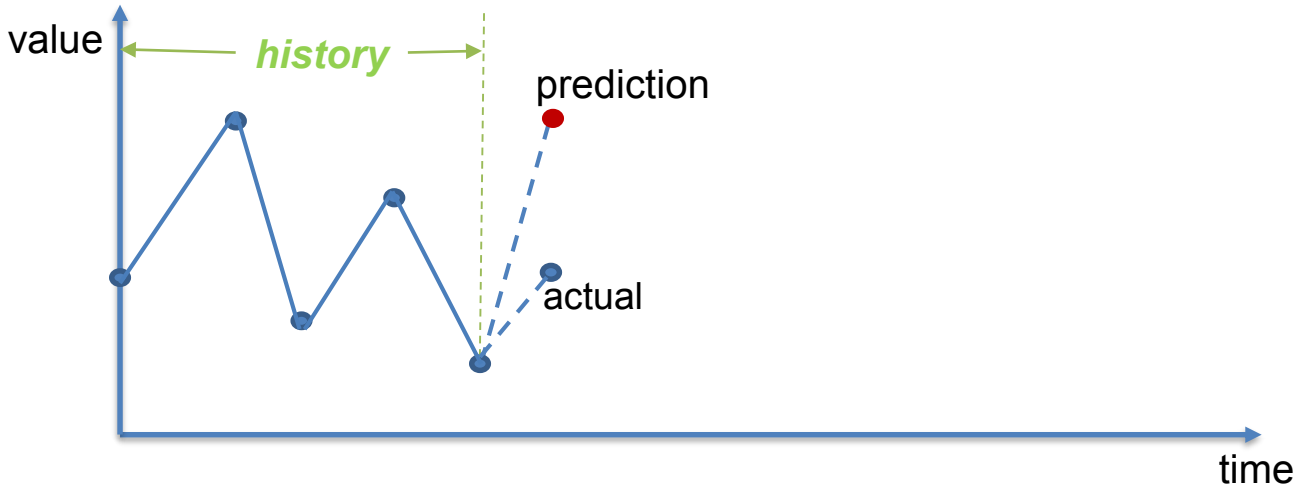


Parameter Value Anomaly Detection model

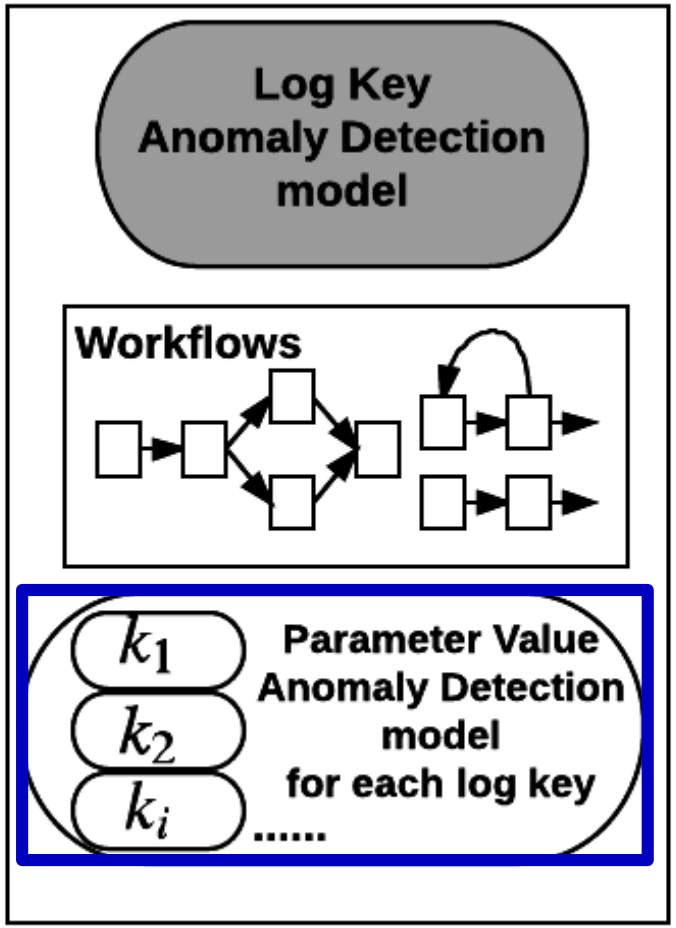


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

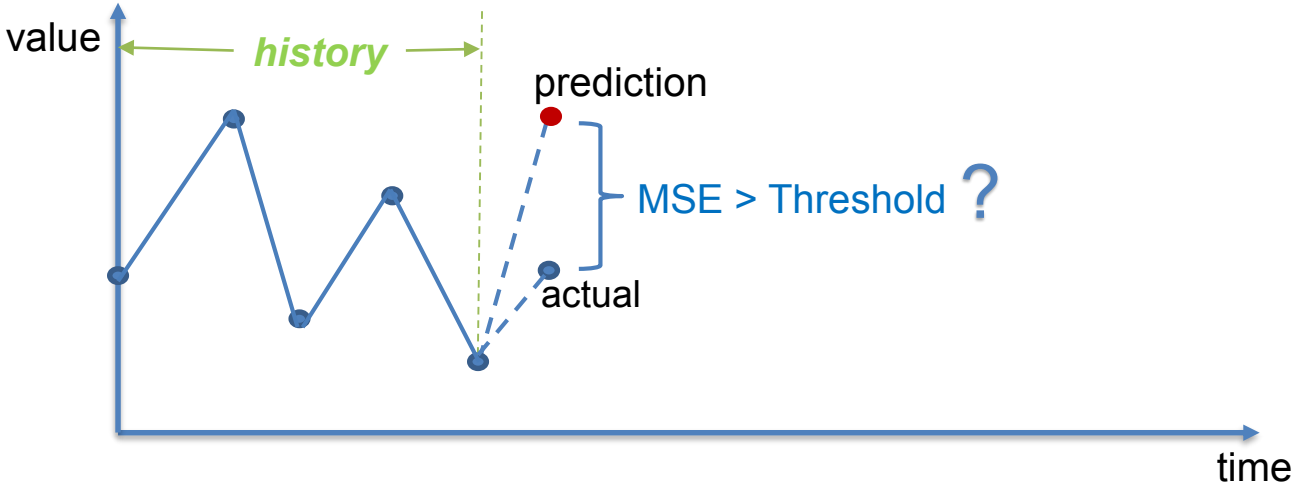


Parameter Value Anomaly Detection model

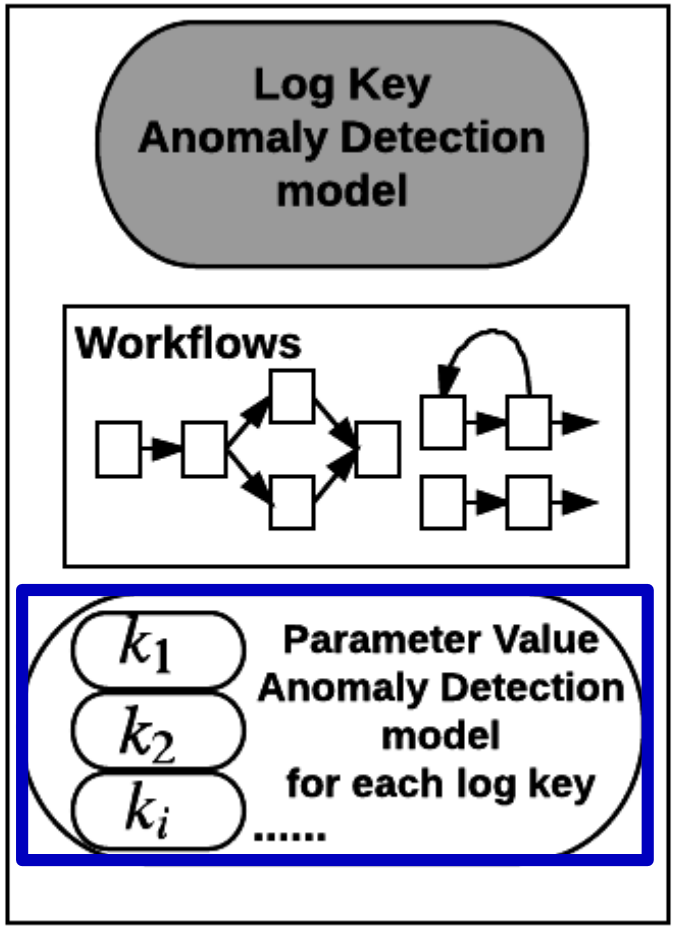


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

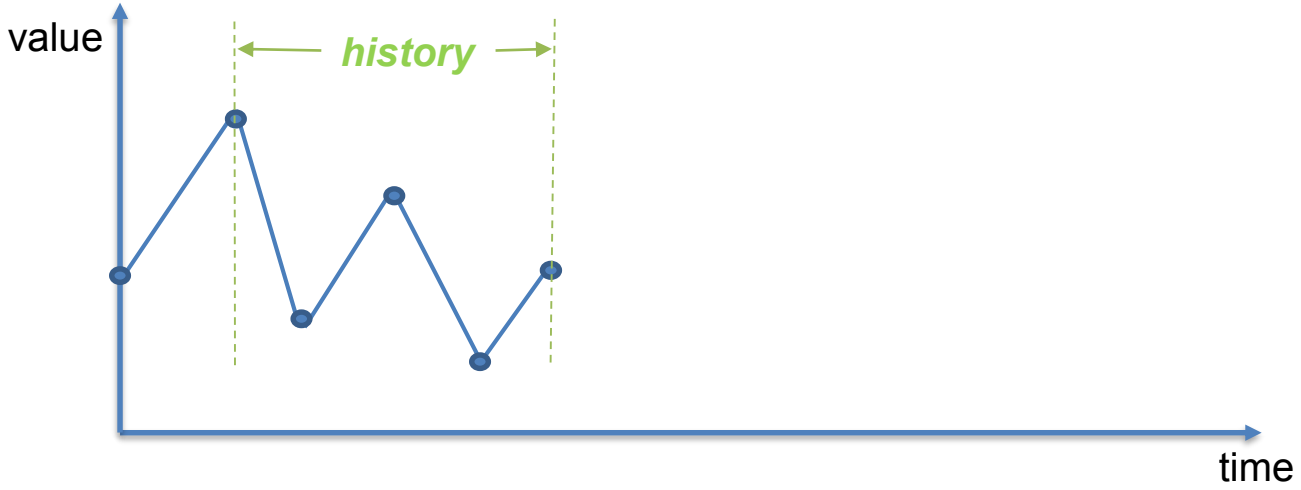


Parameter Value Anomaly Detection model

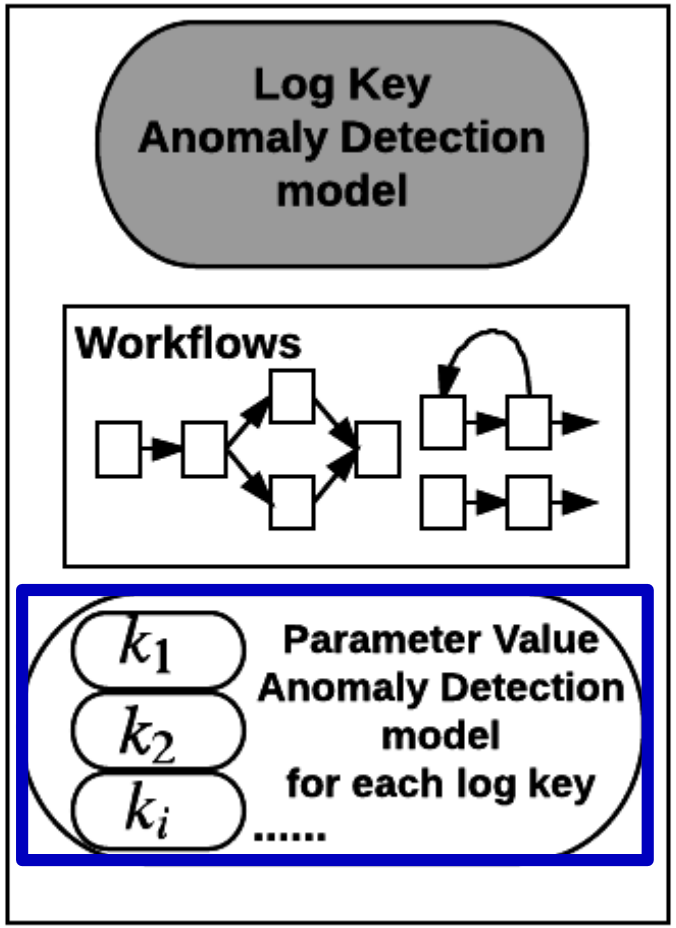


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

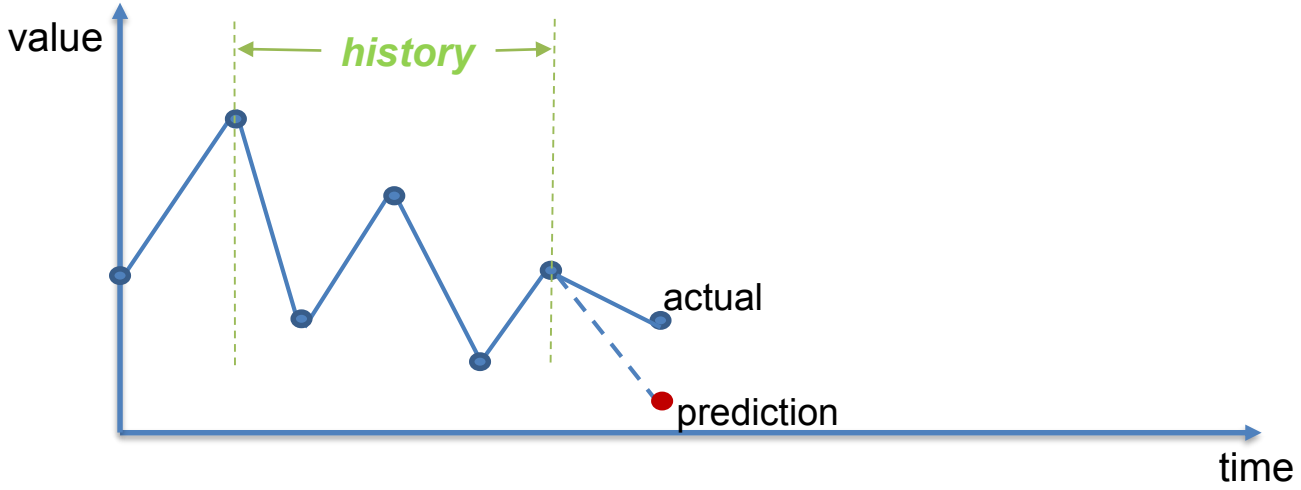


Parameter Value Anomaly Detection model

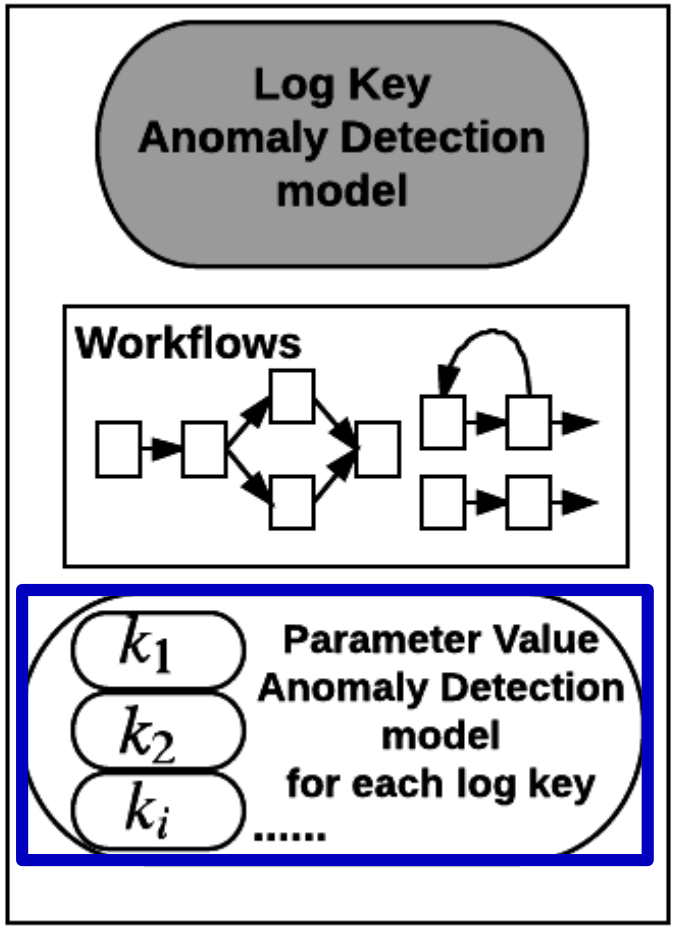


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

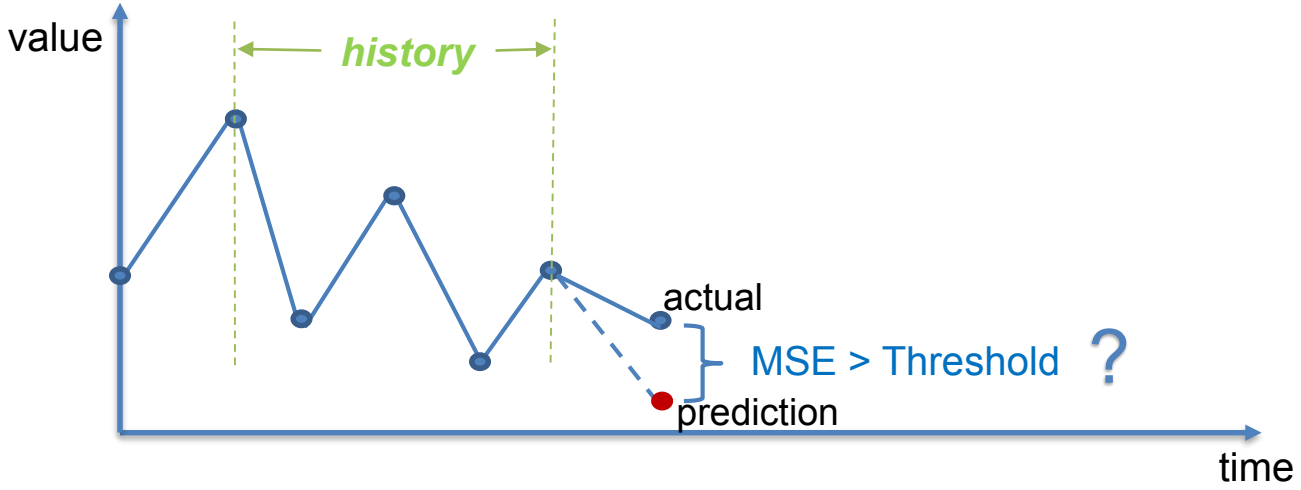


Parameter Value Anomaly Detection model

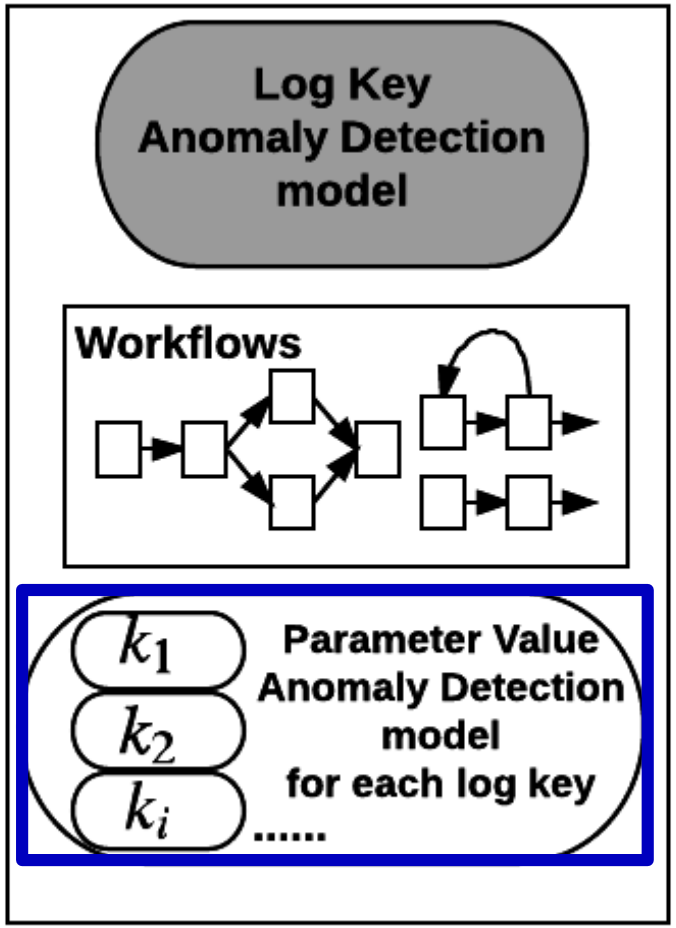


Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

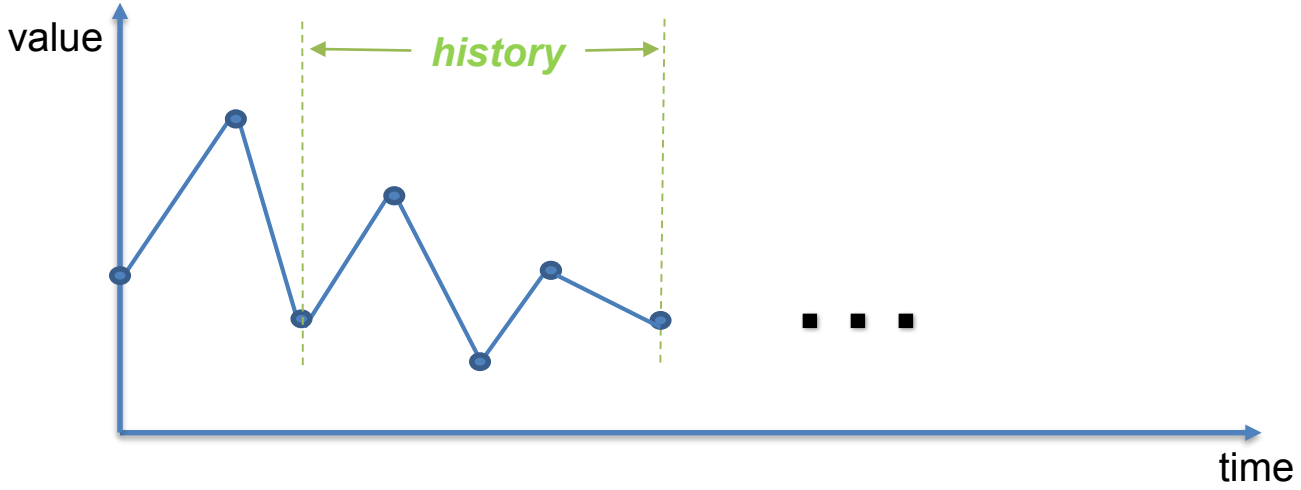


Parameter Value Anomaly Detection model



Multi-variate time series data anomaly detection problem

- ✓ Leverage LSTM-based approach;
- ✓ A parameter value vector is given as input at each time step;
- ✓ An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.



LSTM model online update

Q: How to handle false positive?

LSTM model online update

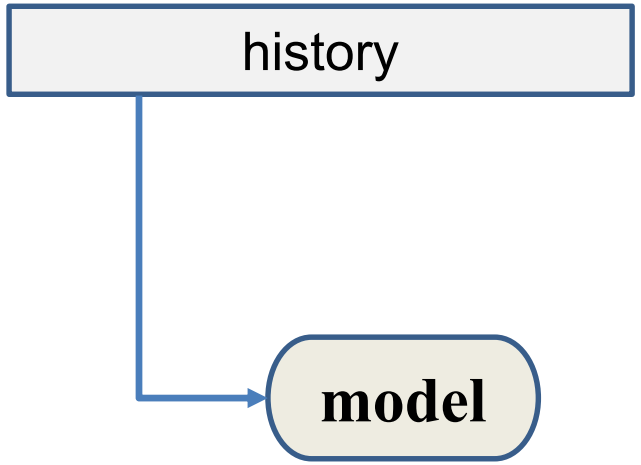
Q: How to handle false positive?

Log sequence: history

LSTM model online update

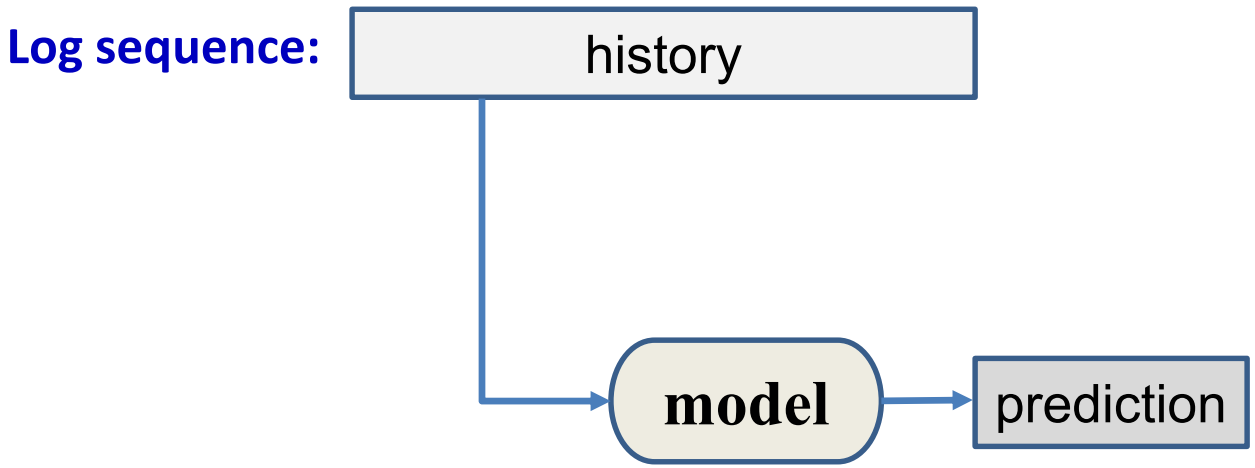
Q: How to handle false positive?

Log sequence:



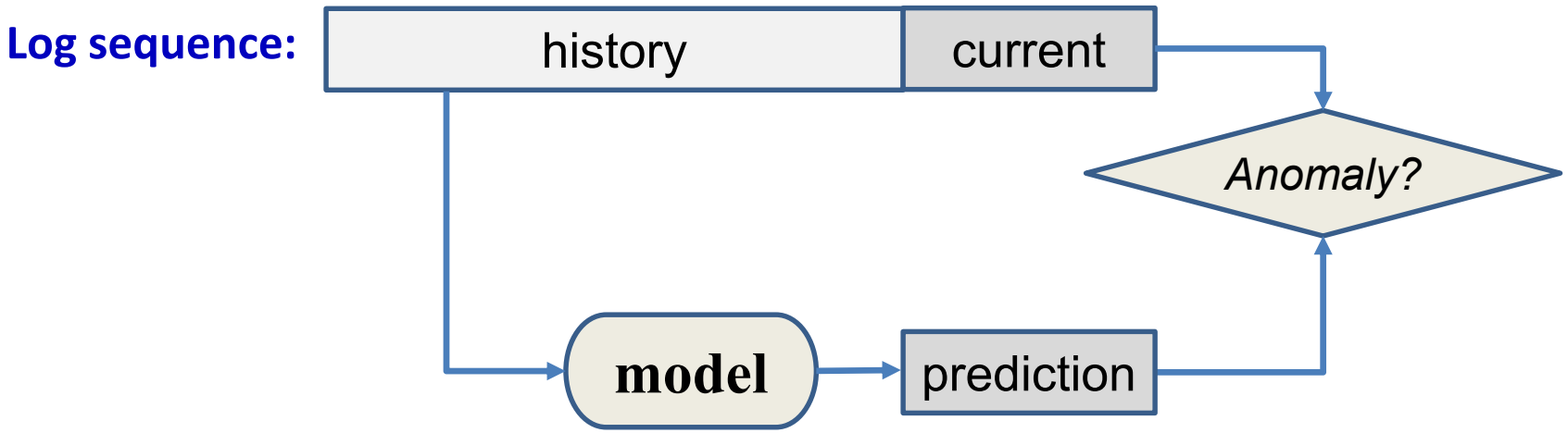
LSTM model online update

Q: How to handle false positive?



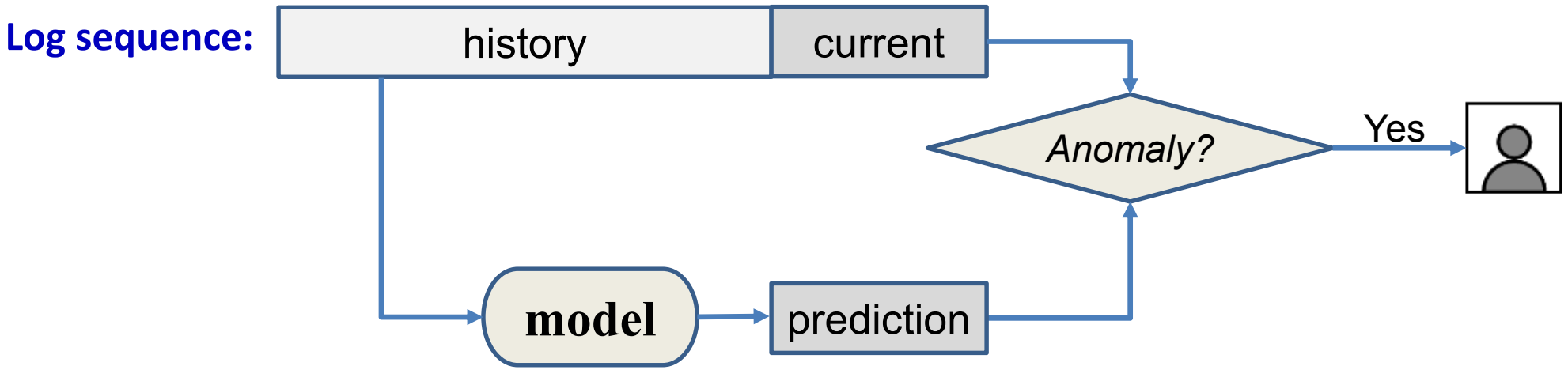
LSTM model online update

Q: How to handle false positive?



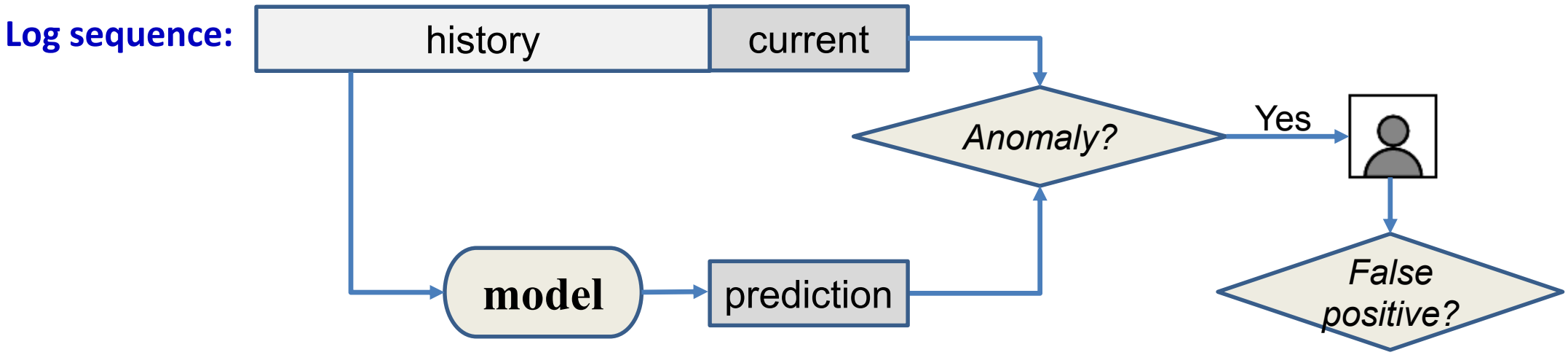
LSTM model online update

Q: How to handle false positive?



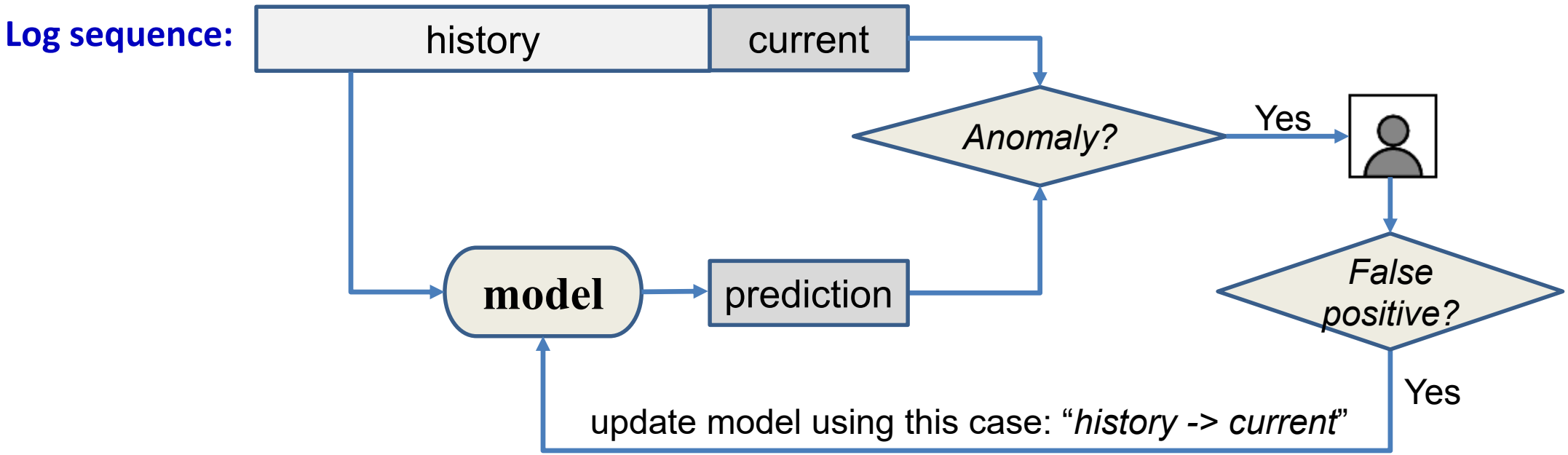
LSTM model online update

Q: How to handle false positive?

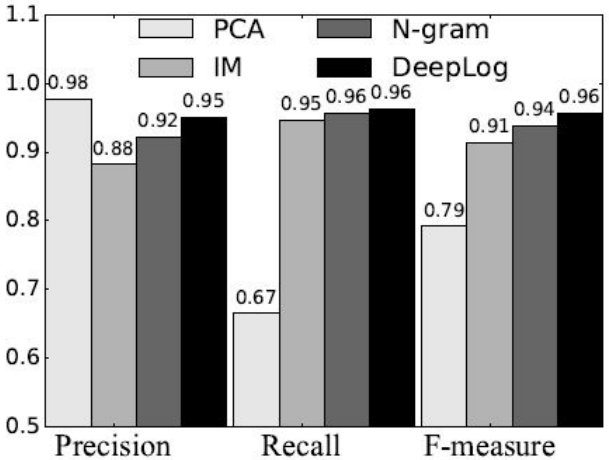


LSTM model online update

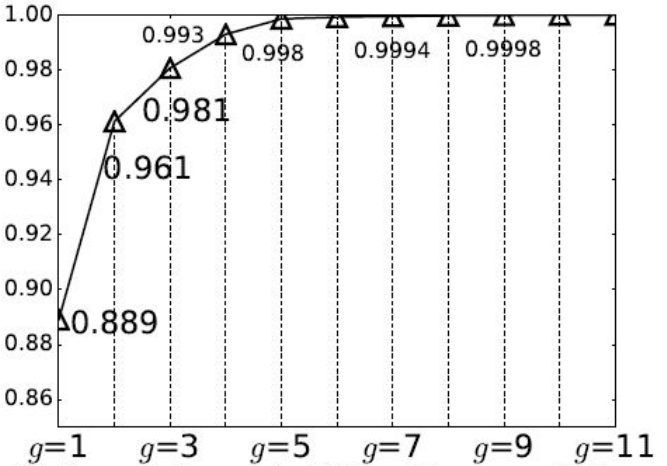
Q: How to handle false positive?



Evaluation – log key anomaly detection



(a) Accuracy on HDFS.



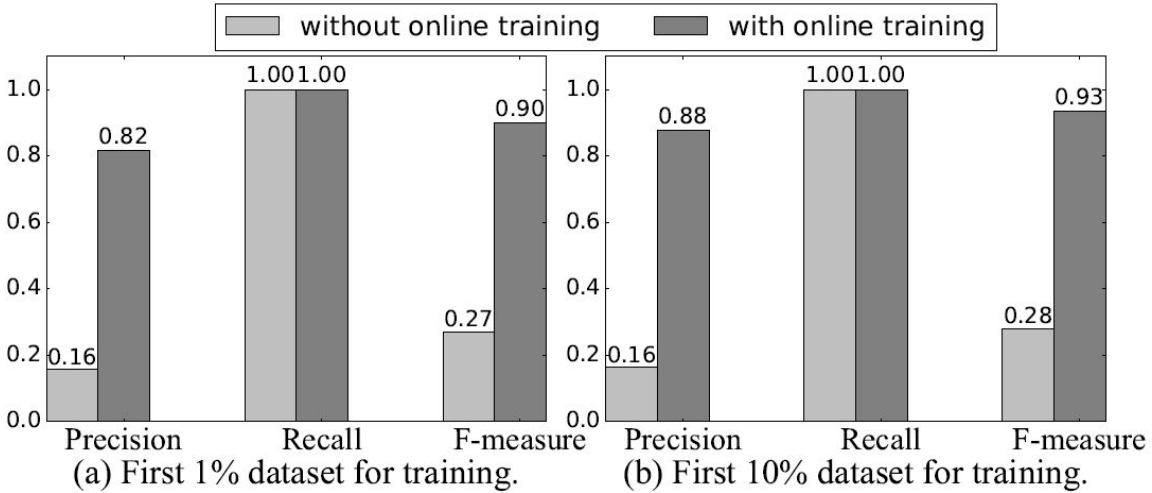
(b) Cumulative probability of top g predictions.

Evaluation results on HDFS log data.

(over a million log entries with labeled anomalies)

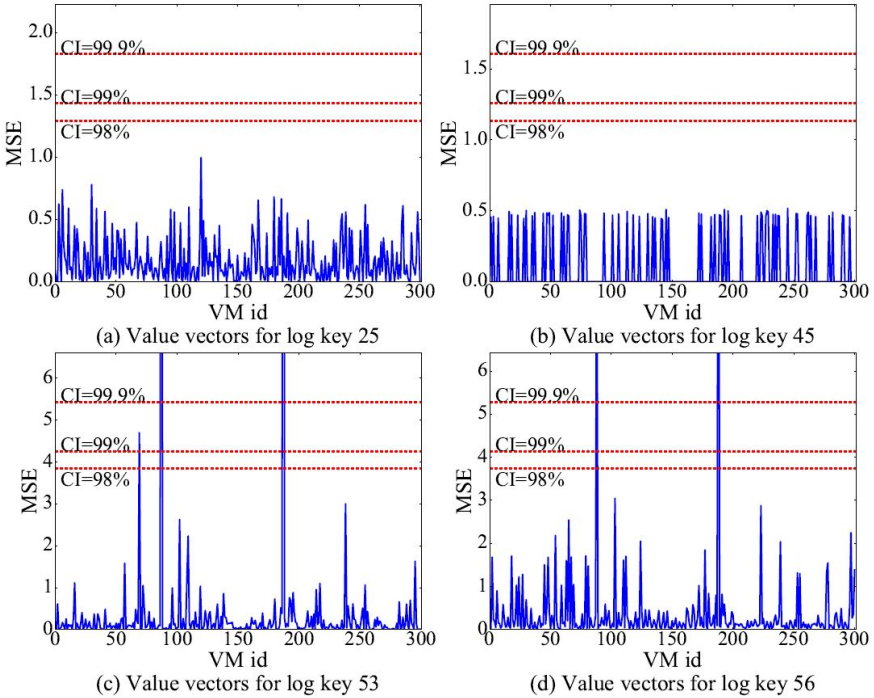
PCA (SOSP'09), IM (UsenixATC'10), N-gram (baseline language model)

Evaluation – LSTM model online update



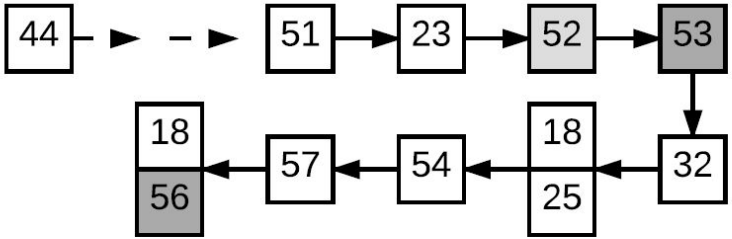
Evaluation results on Blue Gene/L log, with and without online model update.

Evaluation – parameter value anomaly detection



Evaluation results on OpenStack cloud log with different confidence intervals (CIs)

Evaluation – workflow construction



44: instance: * Attempting claim: memory * disk * vcpus * CPU
51: instance: * Claim successful
23: instance: * GET * HTTP/1.1" status: * len: * time: *
*52: instance: * Creating image*
53: instance: * VM Started (Lifecycle Event)
32: instance: * VM Paused (Lifecycle Event)
18: instance: * VM Resumed (Lifecycle Event)
.....
56: instance: * Took * seconds to build instance

Diagnosis using constructed workflow.

Injected anomaly: during VM creation, network speed from controller to compute node is throttled.

Summary

DeepLog

- A realtime system log anomaly detection framework.
- LSTM is used to model system execution paths and log parameter values.
- Workflow models are built to help anomaly diagnosis.
- It supports online model update.

Min Du
mind@cs.utah.edu