Finding unusual messages with ADE
The hunt for context in system logs

James Caffrey

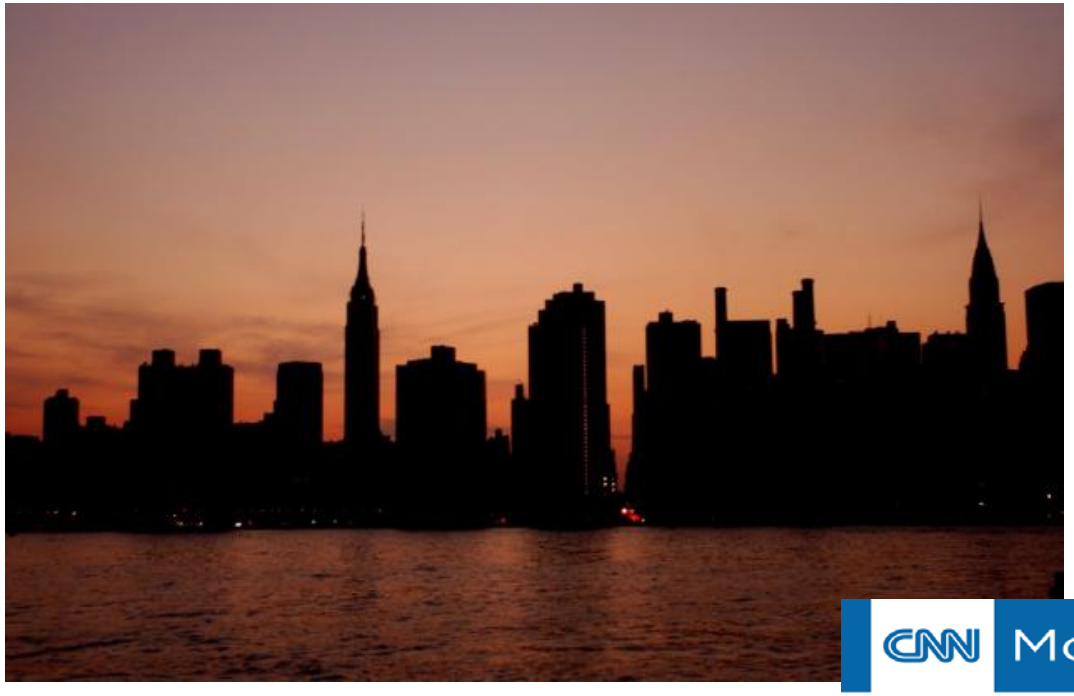
Maintainer for Open Mainframe project ADE

Anomaly Detection Engine for Linux Logs (ADE)

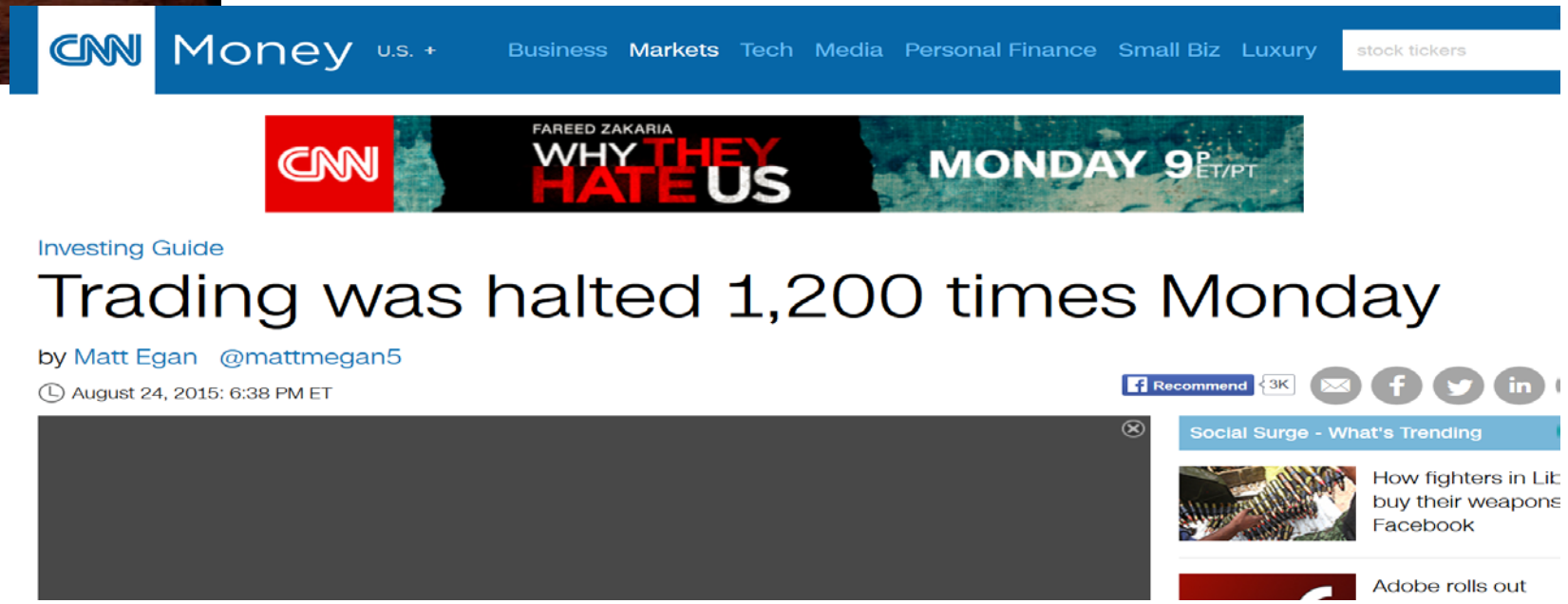
- ADE
  - Anomaly Detection Engine for Linux Logs
  - Sponsored by Linux Foundation Open Mainframe project
- Members of Linux Foundation Open Mainframe project
  - Universities
  - Corporations
  - Vendors

- Agenda
  - An example of why context important is?
  - Context from ADE
    - How context helps to answer critical questions when diagnosing a problem?
    - Example “Is a message occurring when expected?”
  - Creating context
    - Using data science framework to explain the process
    - Example “Do the logs contain sufficient information for ADE to work?”
  - Improving problem diagnosis – providing more insight

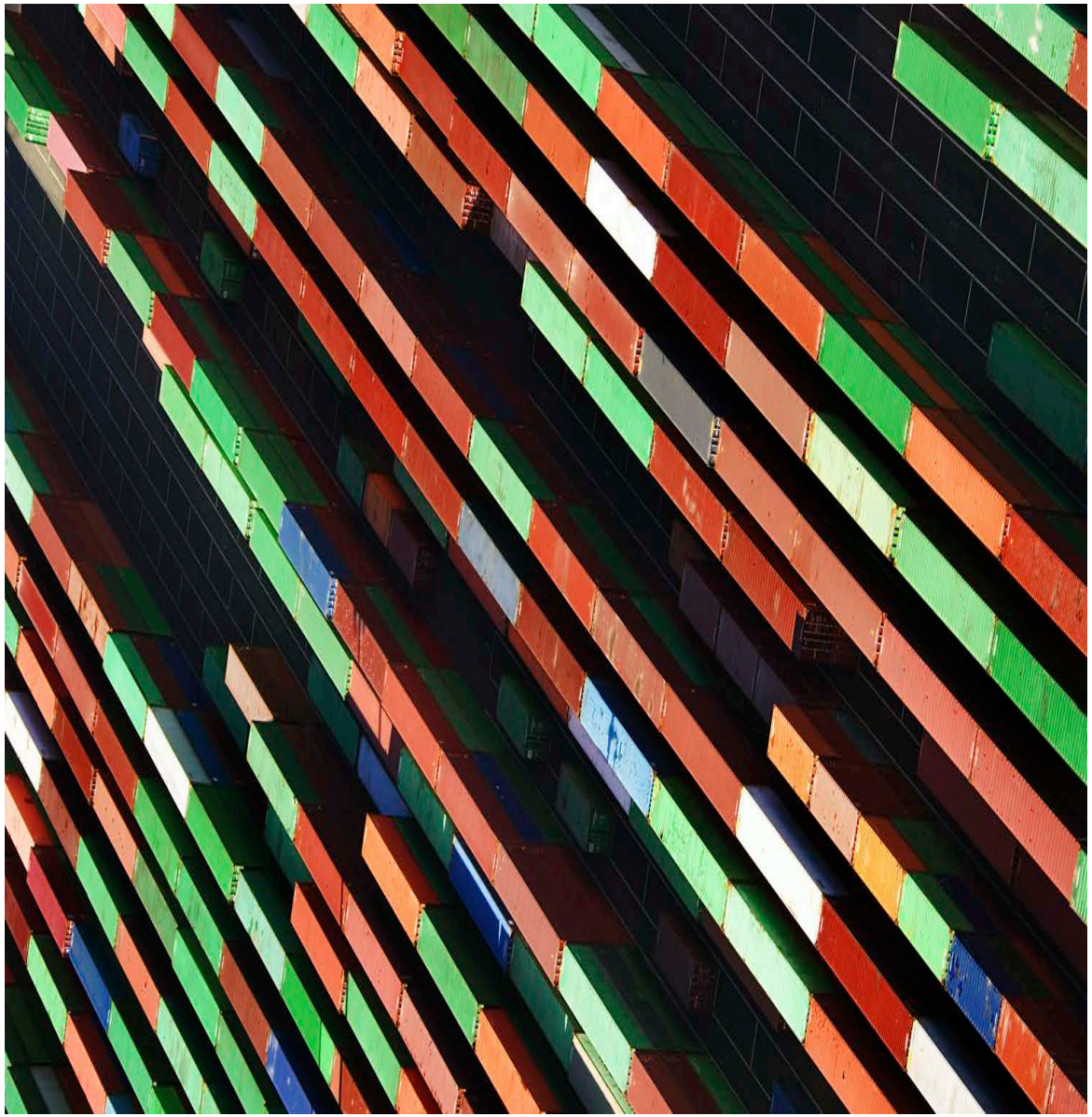
Failure of complex systems



When complex systems fail, people are needed to recover the system and prevent a re-occurrence.



Finding the problem / Finding the cause requires context



- Day 14:15:46 your_system sshd[17348]: pam_unix2(sshd:session): session finished for user support: service=sshd, tty=ssh, rhost=8439fc486bfe4e8c612f575df236838bdf996e9d.boulder.ibm.com
- Day 14:15:46 your_system sshd[17348]: pam_gsa: setcred called but not implemented
- Day 14:20:01 your_system /usr/sbin/cron[17703]: pam_unix2(crond:session): session started for user root: service=crond, tty=cron
- Day 14:20:01 your_system /usr/sbin/cron[17704]: (root) CMD (/usr/local/support/workfstest)
- Day 14:20:01 your_system /usr/sbin/cron[17702]: pam_unix2(crond:session): session started for user root: service=crond, tty=cron
- Day 14:20:01 your_system /usr/sbin/cron[17702]: pam_unix2(crond:session): session finished for user root: service=crond, tty=cron
- Day 14:20:02 your_system /usr/sbin/cron[17703]: pam_unix2(crond:session): session finished for user root: service=crond, tty=cron
- Day 14:20:43 your_system sshd[17937]: Accepted publickey for support from 861719a3aa0caa2cf4c5a6fffe21f5088c53eae8 port 34990 ssh2
- Day 14:20:43 your_system sshd[17937]: pam_gsa: setcred called but not implemented.
- Day 14:20:43 your_system sshd[17937]: pam_unix2(sshd:session): session started for user support: service=sshd, tty=ssh, rhost=8439fc486bfe4e8c612f575df236838bdf996e9d.boulder.ibm.com
- Day 14:20:43 your_system sshd[17939]: pam_gsa: setcred called but not implemented.
- Day 14:20:47 your_system sshd[17939]: Received disconnect from 861719a3aa0caa2cf4c5a6fffe21f5088c53eae8: 11: disconnected by user
- Day 14:20:47 your_system sshd[17937]: pam_unix2(sshd:session): session finished for user support: service=sshd, tty=ssh, rhost=8439fc486bfe4e8c612f575df236838bdf996e9d.boulder.ibm.com
- Day 14:20:47 your_system sshd[17937]: pam_gsa: setcred called but not implemented.
- Day 14:24:01 your_system /usr/sbin/cron[18113]: pam_unix2(crond:session): session started for user root: service=crond, tty=cron
- Day 14:24:01 your_system /usr/sbin/cron[18113]: pam_unix2(crond:session): session finished for user root: service=crond, tty=cron
- Day 14:24:51 your_system kernel: dasd_erp(3990): 0.0.0591: Equipment check or processing error
- Day 14:24:51 your_system kernel: klogd 1.4.1, ---------- state change ----------
- Day 14:25:01 your_system /usr/sbin/cron[18252]: pam_unix2(crond:session): session started for user root: service=crond, tty=cron
- Day 14:25:01 your_system /usr/sbin/cron[18253]: (root) CMD (/var/adm/perfmgr/bin/verify.srm)

Context provided by ADE to help diagnosing a problem

- Can logs help figure out the problem?
- When did it start?
- Where did it start?

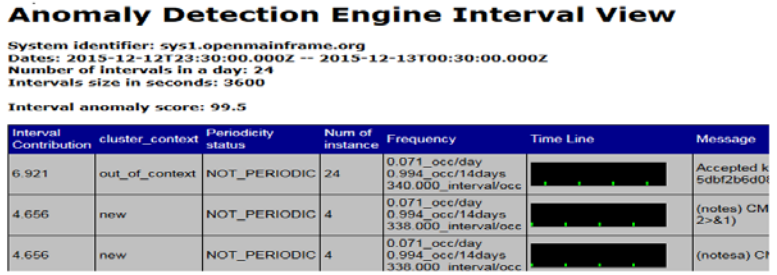
Systems or Groups of systems

- Are there a lot of messages?
- Are there any new messages, any rare messages?
- Are the messages within the interval unusual?

Time slices

- Did it appear with the other messages that it normally appears with it?
- Did it appear when it normally appears?
- Where there more occurrences of messages than expected?
- Is the message unusual?
- Are there similar messages?

Messages



Anomaly Detection Engine Interval View

System identifier: sys1.openmainframe.org
Dates: 2015-12-12T23:30:00.000Z -- 2015-12-13T00:30:00.000Z
Number of intervals in a day: 24
Intervals size in seconds: 3600

Interval anomaly score: 99.5

Interval Contribution	cluster_context	Periodicity status	Num of instance	Frequency	Time Line	Message
6.921	out_of_context	NOT_PERIODIC	24	0.071_occ/day 0.994_occ/14days 340.000_interval/occ		Accepted k 5dbf2b6d0
4.656	new	NOT_PERIODIC	4	0.071_occ/day 0.994_occ/14days 338.000_interval/occ		(notes) CM 2>&1)
4.656	new	NOT_PERIODIC	4	0.071_occ/day 0.994_occ/14days 338.000_interval/occ		(notesa) C

Context example – when a message occurs, is it expected

Is this message or a similar message occurring when expected?

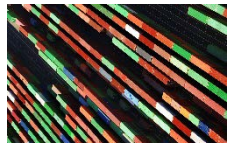
Context	Explanation
New	ADE has not previously detected this message
IN_SYNC	ADE expects this message to be issued in a periodic pattern, and the message was issued as expected
NOT_IN_SYNC	ADE expects this message to be issued in a periodic pattern, but the message was not issued as expected
NOT_PERIODIC	ADE does not expect this message to be issued in a periodic pattern

/usr/sbin/cron(root): /var/adm/perfmgr /bin/verify.srm_20		in context (15)	2	1.000	23.929_occ/day 1.003_interval/occ	NOT_IN_SYNC	0.003
/usr/sbin /cron(root):root_77		in context (1)	1	95.588	1.059_occ/day 22.667_interval/occ	NOT_PERIODIC	NaN
postfix/cleanup_15		in context (15)	26	1.000	23.929_occ/day 1.003_interval/occ	IN_SYNC	3.871
postfix/local_17		in context (4)	13	1.000	23.929_occ/day 1.003_interval/occ	IN_SYNC	7.246
postfix/pickup_13		in context (4)	13	1.000	23.929_occ/day 1.003_interval/occ	IN_SYNC	7.246

Creating context using a data science approach

What is data science

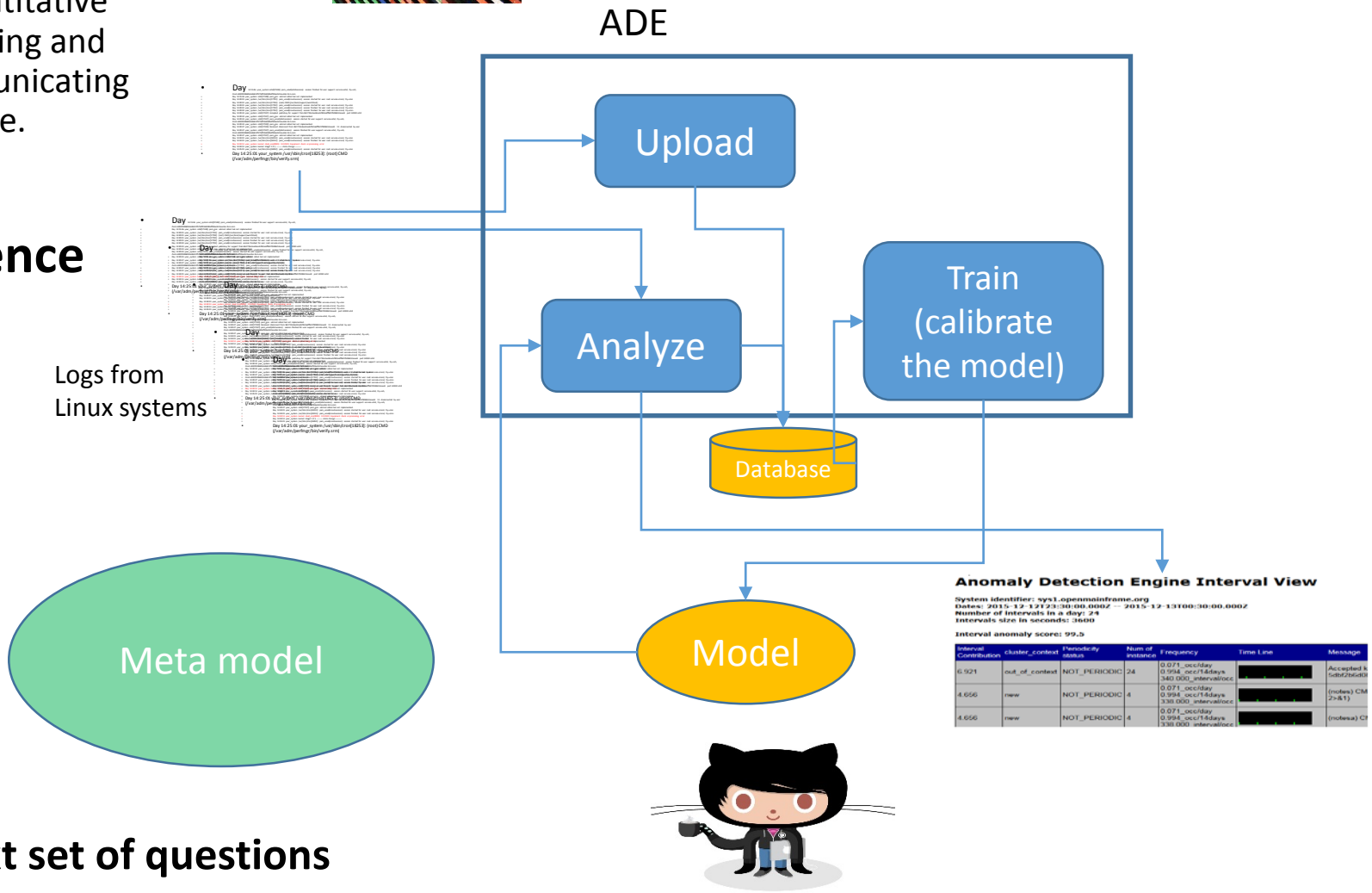
Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.



Steps involved in a data science solution

- Define the question of interest
- Get the data
- Clean the data
- Explore the data
- Fit statistical models
- Communicate the results
- Make your analysis reproducible

Wash , rinse, repeat Answer the next set of questions



Example of data science in ADE

Question: Will the logs work or are more logs needed?

Hypothesis: A rule can be written to determine if the model created from the log(s) are able to detect if the anomaly scores are unusual

Anomaly score for 99.5 different from anomaly score for 100

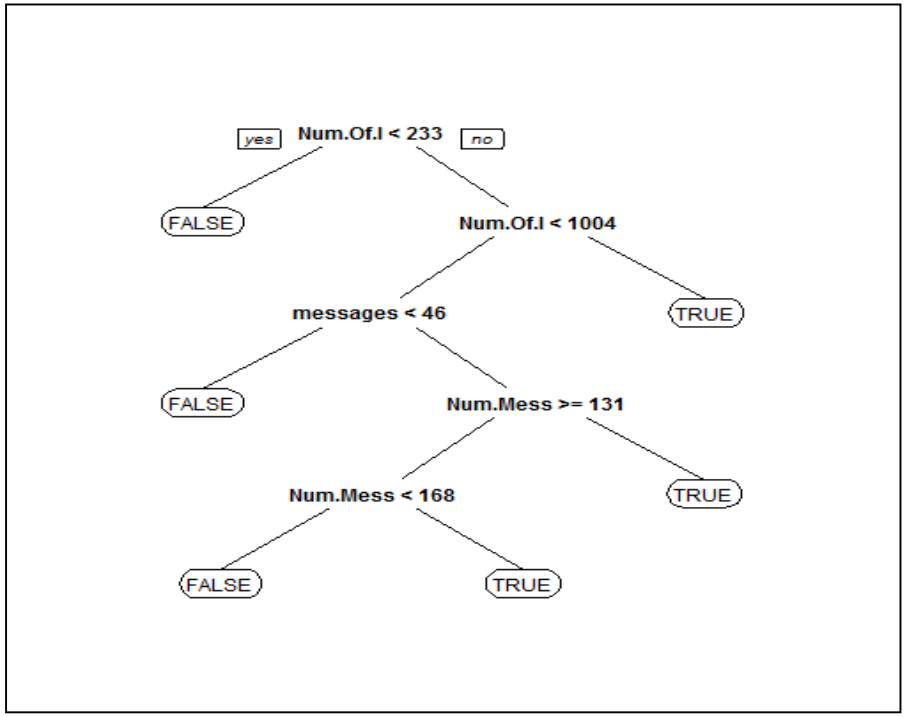
Use a data science approach to create classification model

- Build large number of models
- Applied statistical technique (linear discriminate analysis) using that data
- Examine the results

Implement ADE command to verify sufficient data in training period to build model based on values determined by analysis

More than 1000 intervals

More than 180 message ids and more than 20 intervals



Providing context for problem diagnosis is a journey

Improvements driven by Community Contributions to ADE

- Today - Infrastructure to detect anomalies
  - Parsing of logs
  - Splitting logs into time slices
  - Wrapper messages
  - Statistical methods
    - Counting
    - Pattern recognition
  - Output management
    - Storing data in database
    - Writing results to file system
- Data for researchers
  - Single data set from IBM

OPEN MAINFRAME PROJECT

- ADE is sponsored by the open mainframe project of the Linux Foundation
- Community open to
  - Corporations
  - Universities

- Tomorrow -
  - Better understanding of how to diagnose a problem
  - Better cleaning of the data
  - Better statistical methods - analytics
  - Better Visualization
- Is unusual behavior important
- Progress depends on data for researchers
  - 80% of the work in any data science project is acquiring the data

- **Less impact from problems**

What I learned from nose works about creating context Finding problems in complex systems

- Trusted your dog – trust the analytics
  - Both will surprise you with their accuracy and ability to find things
- Training is important
  - The better the training of either the dog or ADE the more accurate it becomes
- Take advantage of all of the clues / context provided.
  - A dog will often work to isolate the scent. Use the context provide by tools like ADE to isolate the problem.

Questions?

References

Open mainframe project – ADE

Description of Data Science