

Turning Privacy Constraints into Syslog Analysis Advantage

Siavash Ghiasvand¹, Florina M. Ciorba², Wolfgang E. Nagel¹

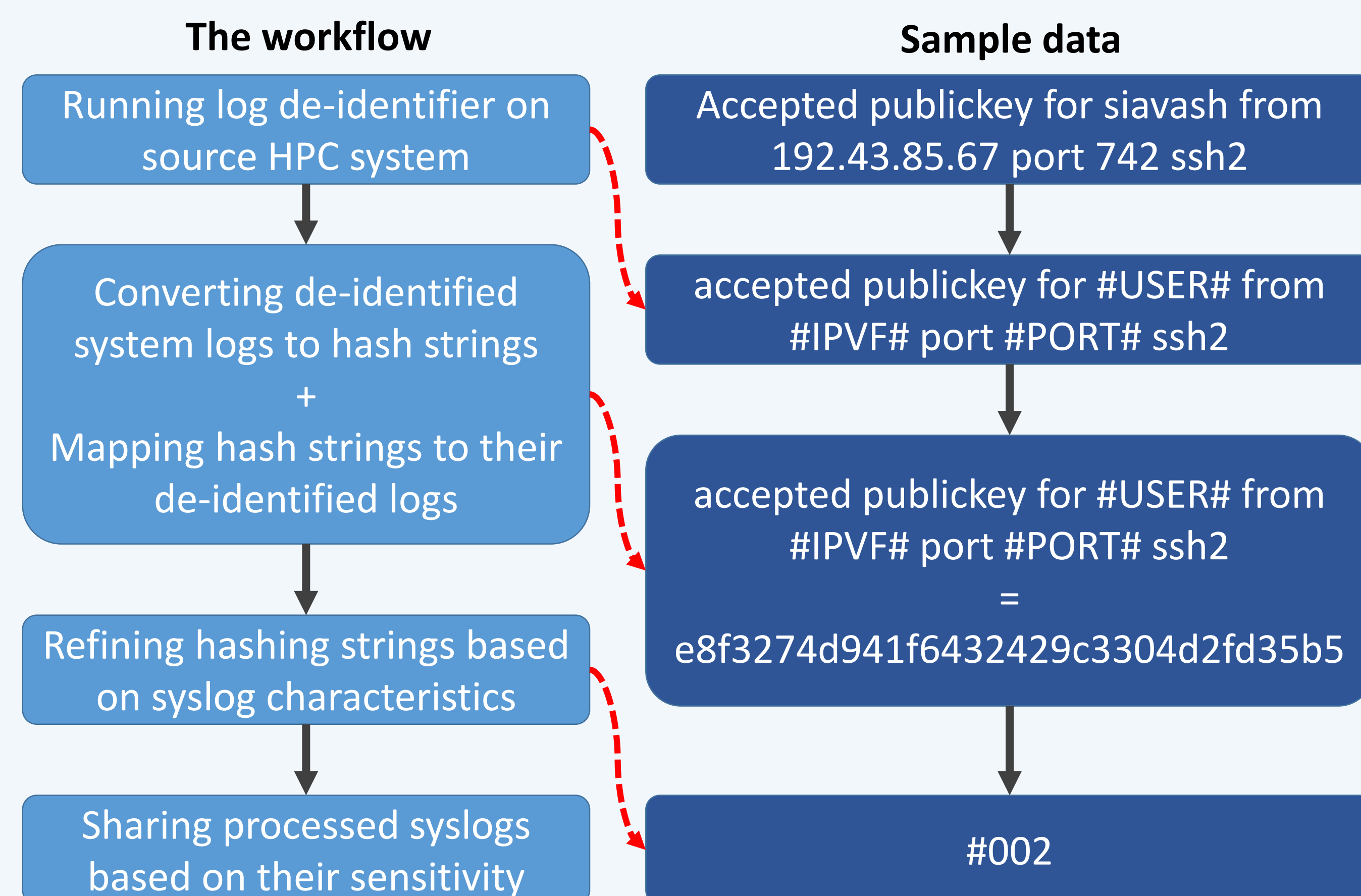
(¹Technische Universität Dresden, ²University of Basel)



1. Motivation

- Size and complexity of high performance computers are growing
- Their failure rate is increasing too
- Efficiency of recovery mechanisms depends on mean time between failures
- De-facto fault tolerance mechanisms lose their effectiveness in near future [1]
- Predicting failures using system logs is an effective alternative solution [4]
- Providing a general solution, requires diverse system logs [2,3]
- System logs contain sensitive data
- Concerns about users privacy prevents open circulation of system logs
- Providing a method to make system log sharing feasible, is necessary

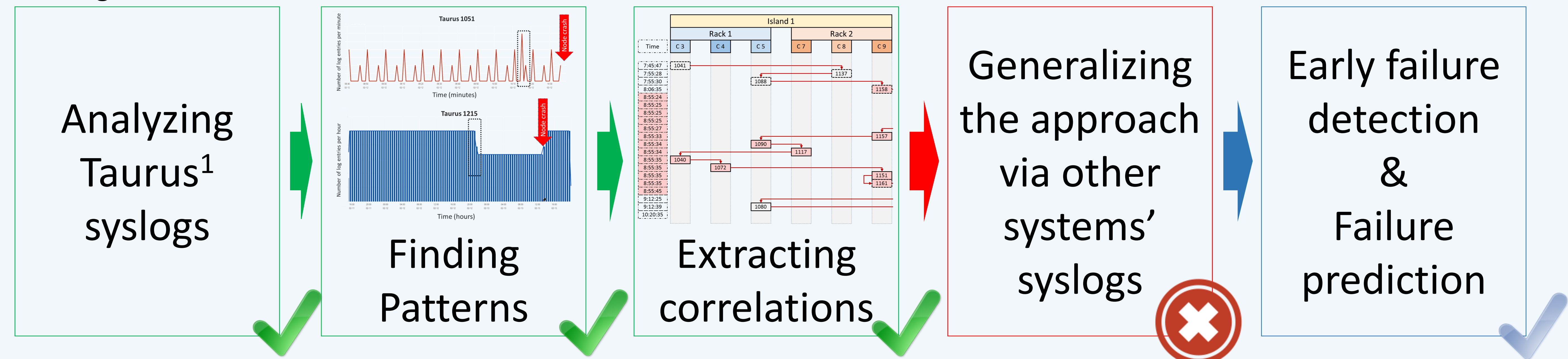
3. Approach



Identifiable and de-identified syslogs

1. (root) CMD (/usr/lib64/sa/sa1 1 1)	identifiable
2. Accepted publickey for siavash from 192.43.85.67 port 742 ssh2	
3. (root) CMD (run-parts /etc/cron.hourly)	
1. (root) cmd (_#PATH# _#HADY#_)	Semi de-identified
2. accepted publickey for #USER# from #IPVF# port #PORT# ssh2	
3. (root) cmd (run-parts _#PATH#_)	
1. (root) cmd (_#PATH# _#HADY#_)	de-identified
2. [REDACTED]	
3. (root) cmd (run-parts _#PATH#_)	
1. 1808e388919ee30bfaf7fdc5aed8823d	de-identified
2. e8f3274d941f6432429c3304d2fd35b5	
3. 2e307f37e1f083ec9e18bd238d4ec38f	

2. Challenge



4. Results

Test data

750,000,000 Syslog entries	500 computing nodes	45.5 GB raw-data size	5,750,000,000 words with 2 - 34 characters
----------------------------	---------------------	-----------------------	--

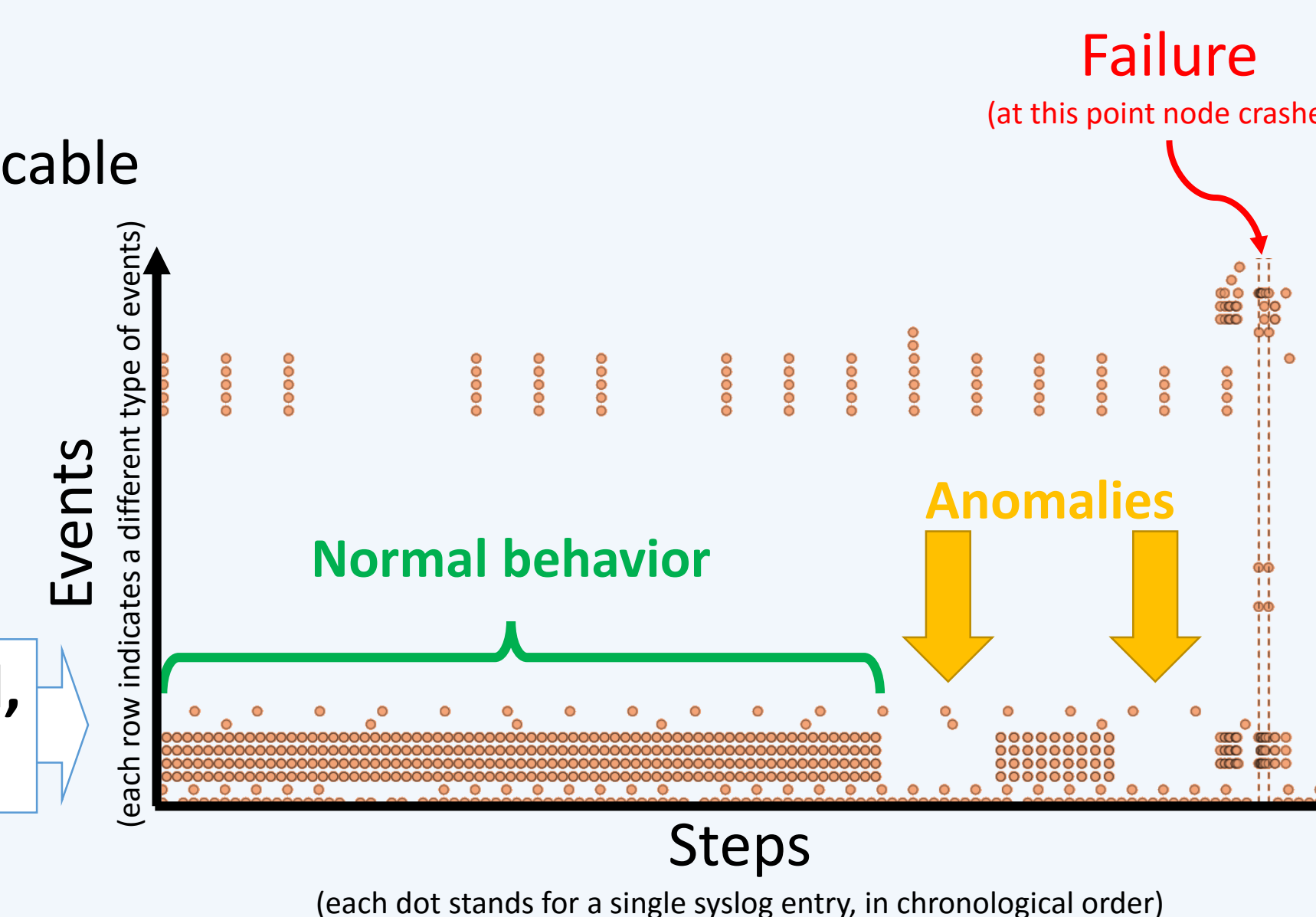
Observations after pre-processing

1000 unique entries	1800 entry/s processing speed Using a single thread	2.5 GB altered-data size	Constant length hash codes
Detected by 38 RegEx queries	Easily scalable	Reducible via better hashing	Dynamic length is also possible

Results

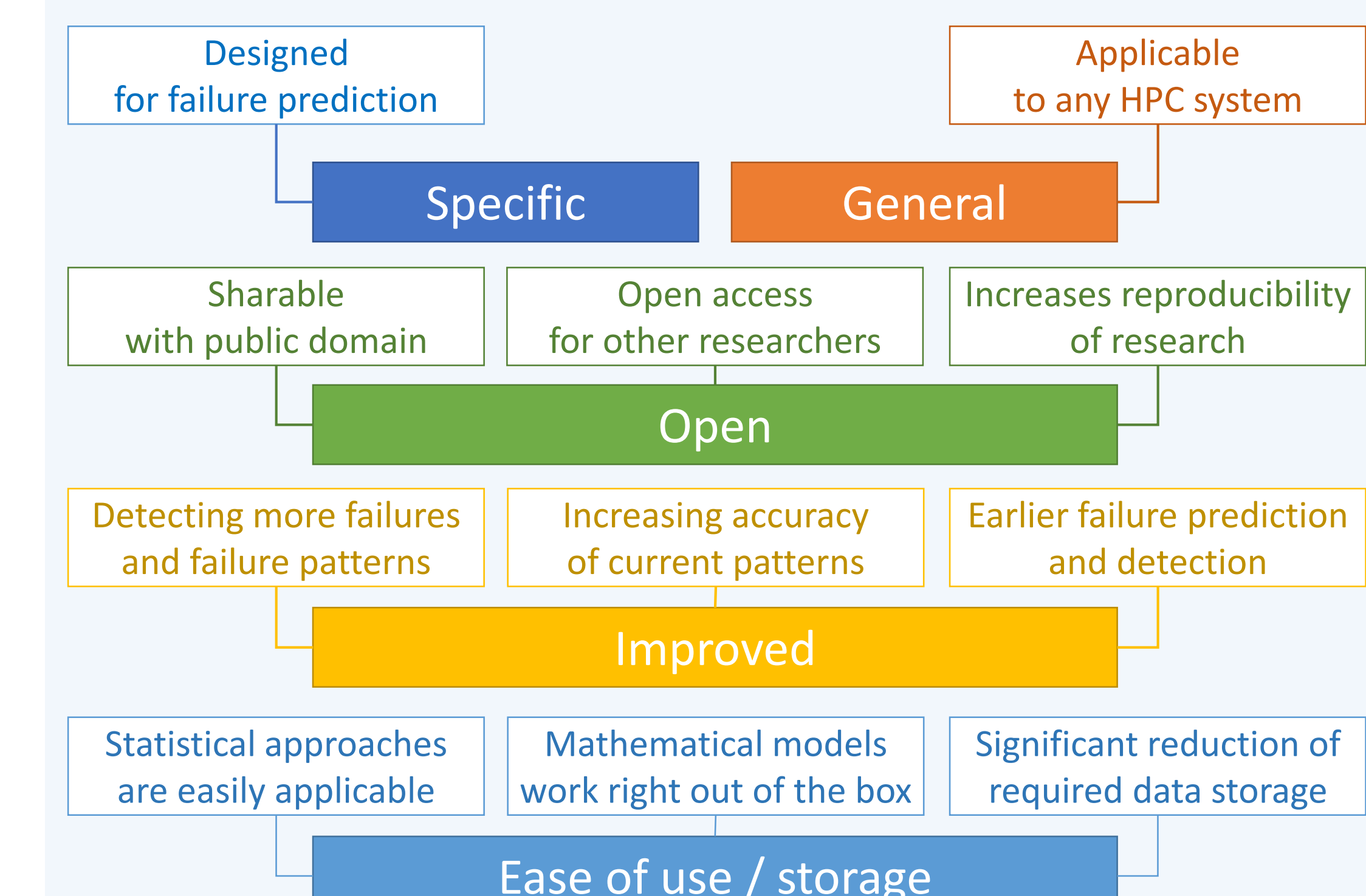
- Pattern-searching domain reduced to a constant number of 1000 entries
 - Performance significantly increased
- Reducing storage size by 94%
 - In-memory mechanisms are now applicable
- Sensitive data eliminated
 - user privacy ensured
 - Processed syslogs are publishable

Despite all changes, data is meaningful and, our pattern detection approach is working.



¹ <https://doc.zih.tu-dresden.de/hpc-wiki/bin/view/Compendium/SystemTaurus>
² <http://www.pub.zih.tu-dresden.de/~ghiasvan/publications/sc16/>

5. Features and benefits



6. Take-home note

Syslogs are invaluable sources of information for predicting failures. Even though, syslogs provide detailed information about system behavior, in our failure prediction approach those information do not yet have an added value. Therefore, simply by replacing all variables with constants, not only we ensure users privacy and prepare a platform for system log sharing, but also we contribute to failure prediction and reproducibility of research.

Acknowledgement

This work is in part supported by the German Research Foundation (DFG) in the Cluster of Excellence "Center for Advancing Electronics Dresden" (cfaed).

References

- F. Cappello, A. Geist, and W. Gropp. Toward Exascale Resilience: 2014 update. Supercomputing Frontiers and Innovations, 2014.
- S. Ghiasvand, F. M. Ciorba, and W. E. Nagel. Toward Resilience in HPC: A Prototype to Analyze & Predict System Behavior. ISC, poster, 2016.
- S. Ghiasvand, F. M. Ciorba, R. Tschuter, and W. E. Nagel. Analysis of Node Failures in High Performance Computers Based on System Logs. SC, poster, 2015.
- S. Ghiasvand, F. M. Ciorba, R. Tschuter, and W. E. Nagel. Lessons Learned from Spatial and Temporal Correlation of Node Failures in High Performance Computers. 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, 2016.