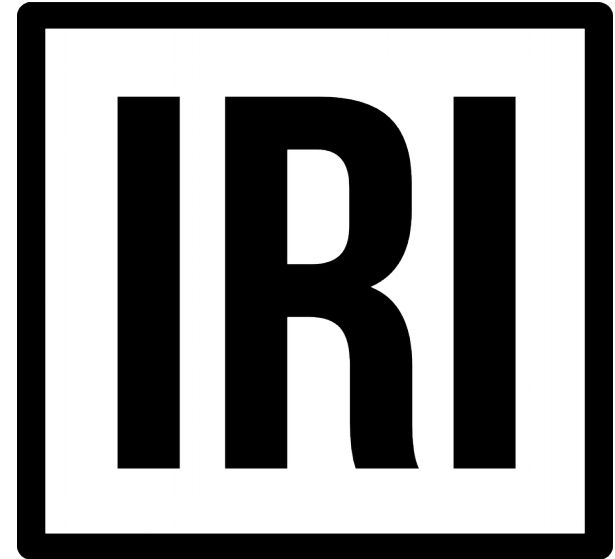


ALICE security research site at University of Frankfurt



9th ALICE Tier-1/Tier-2 Workshop

**Andres Gomez Ramirez, Udo Keschull
IRI - Goethe University Frankfurt
ALICE UF Grid site**

andres.gomez@iri.uni-frankfurt.de



UF Grid Site

- Security research, not production data processing.
- **PhD thesis:** “Deep Learning and Isolation Based Security for Intrusion Detection and Prevention in Grid Computing”.
- **5 Ubuntu 16.04** nodes.
- **Centos 6** containers.
- **CernVM-FS** installed in the hosts and shared as a volume inside Docker containers.



Motivation: Grid Security Challenges

- Users can execute any application: **arbitrary code execution by design.**
- Payloads are frequently executed directly on host Operating System.
- Network sections are **shared.**
- **Hundreds of thousands** of jobs running simultaneously.
- Expensive to have many security experts **monitoring the Grid.**
- Similar to Cloud computing.

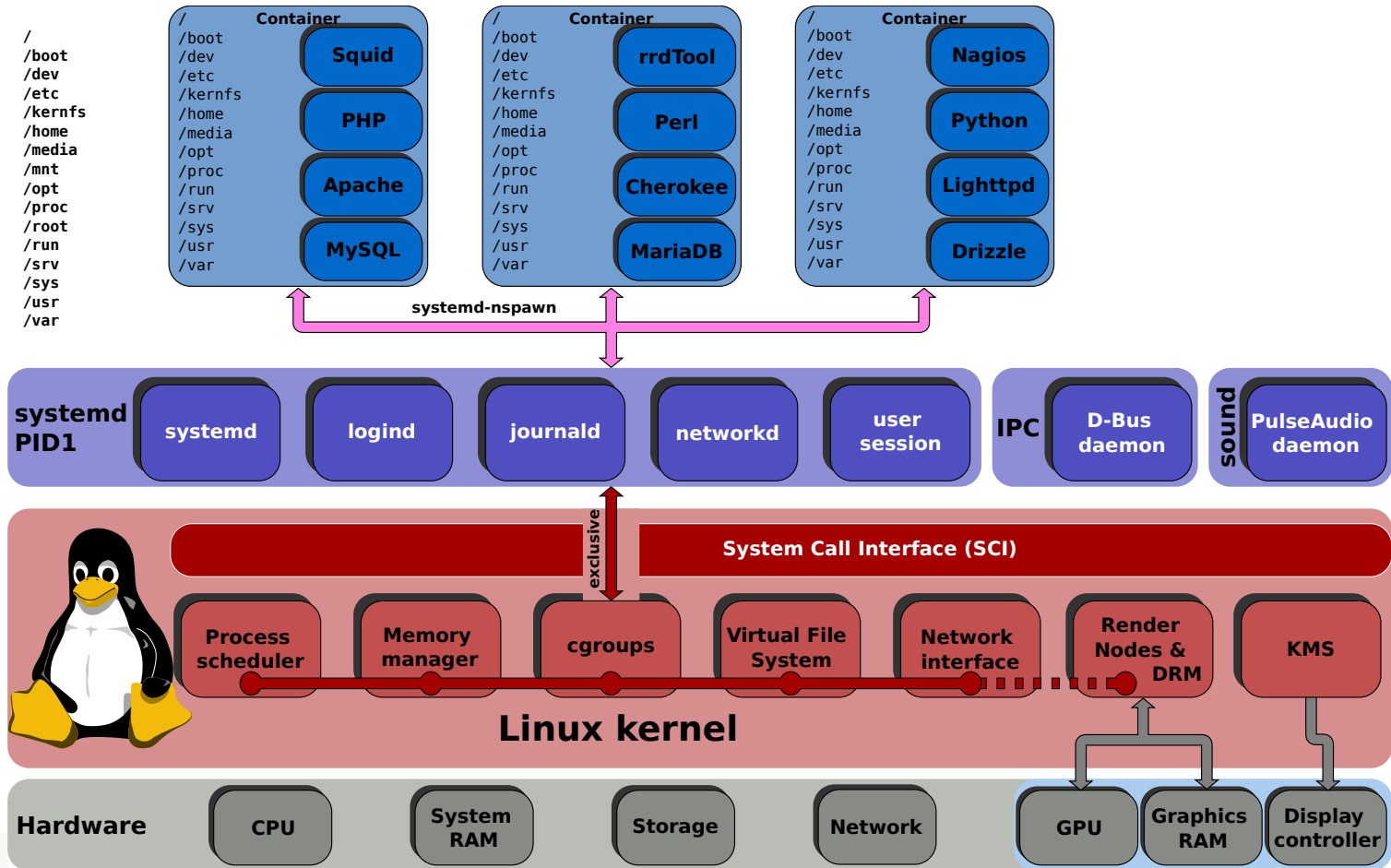


Proposed Solutions

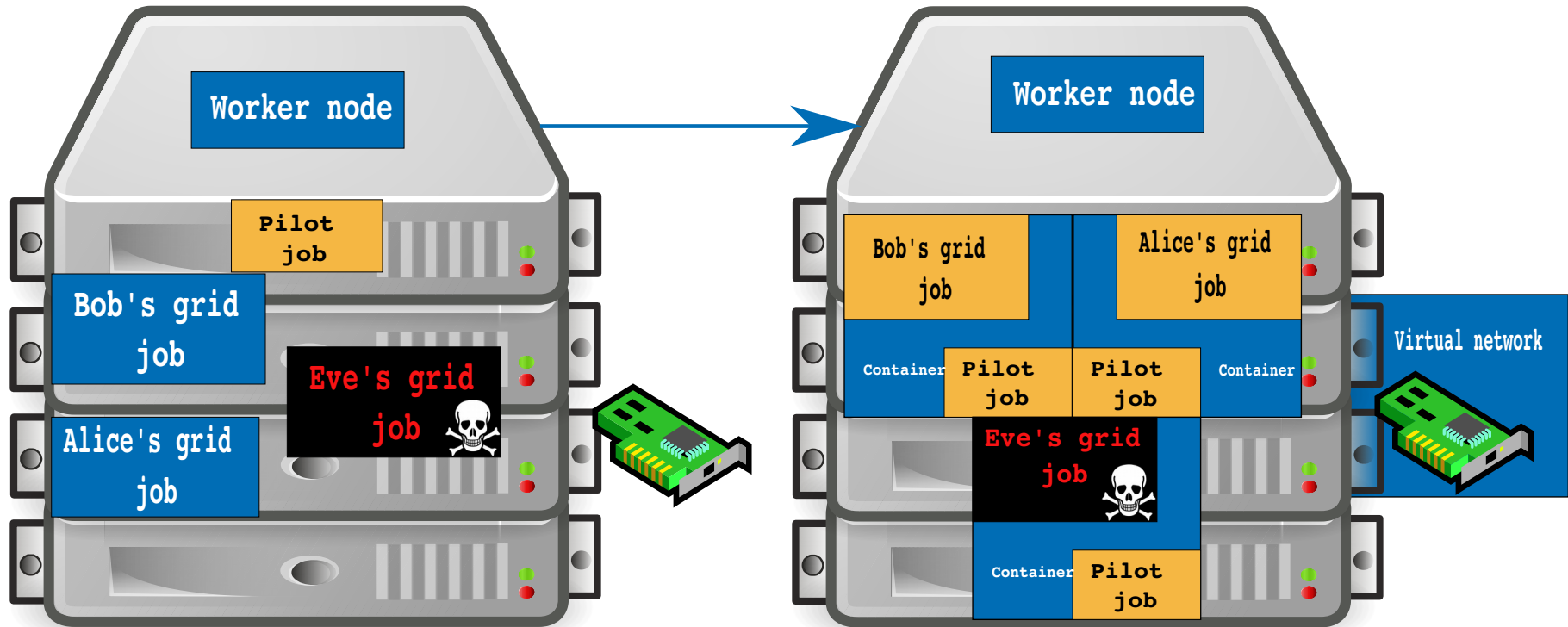
- Linux Containers to execute payloads.
- Network **isolation** with virtual networks.
- Isolation to extract better **payload behavior** data from the host and from the network.
- Automated **Intrusion Detection and Prevention**.
- **Deep Learning** to enable this automation.



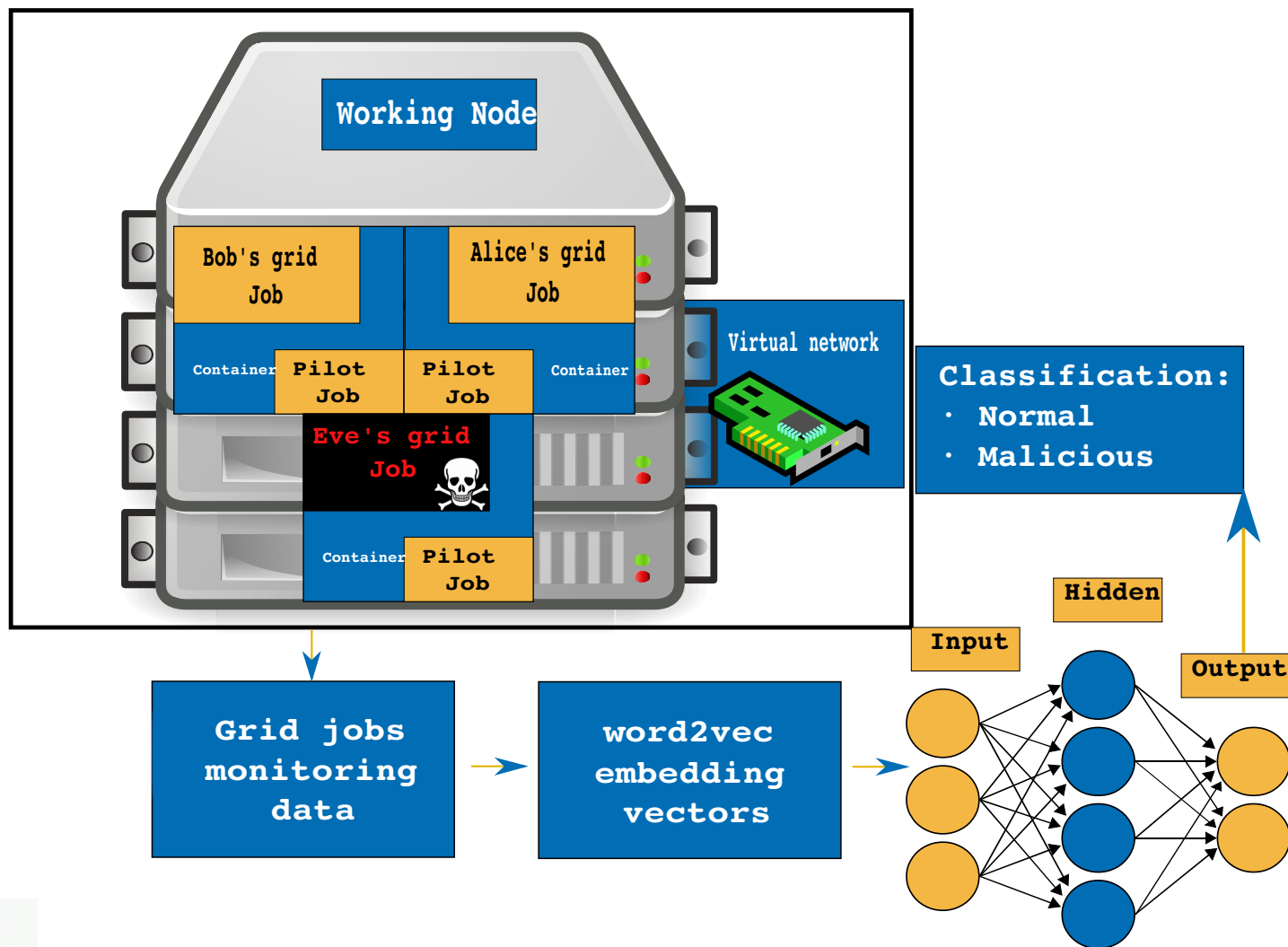
Security by Isolation: Linux Containers



Grid Job Execution and Network Isolation



Behavior Monitoring for the Grid



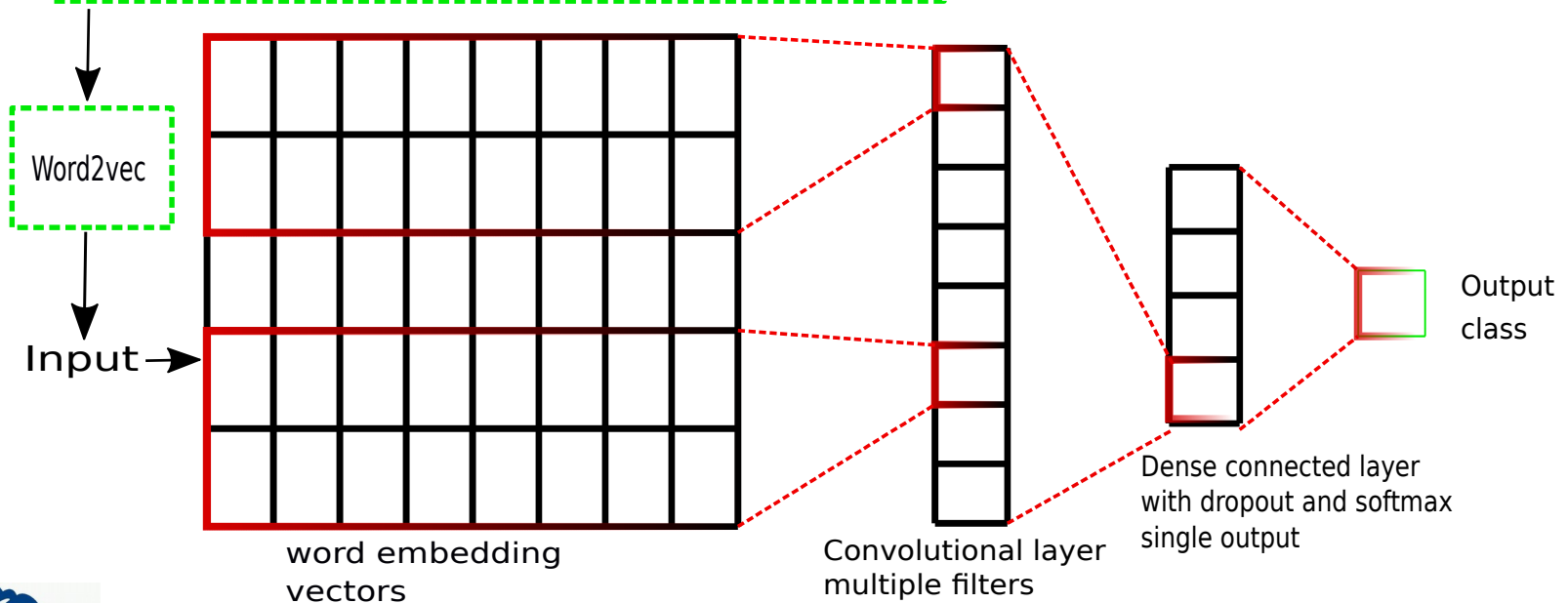
ALICE Grid Security Monitoring

- Linux Containers
- **Deep Learning**
- **Convolutional Neural Networks**
- **Recurrent Neural Networks**
- **Generative** method for improving training
- Grid Jobs – **normal vs malware**

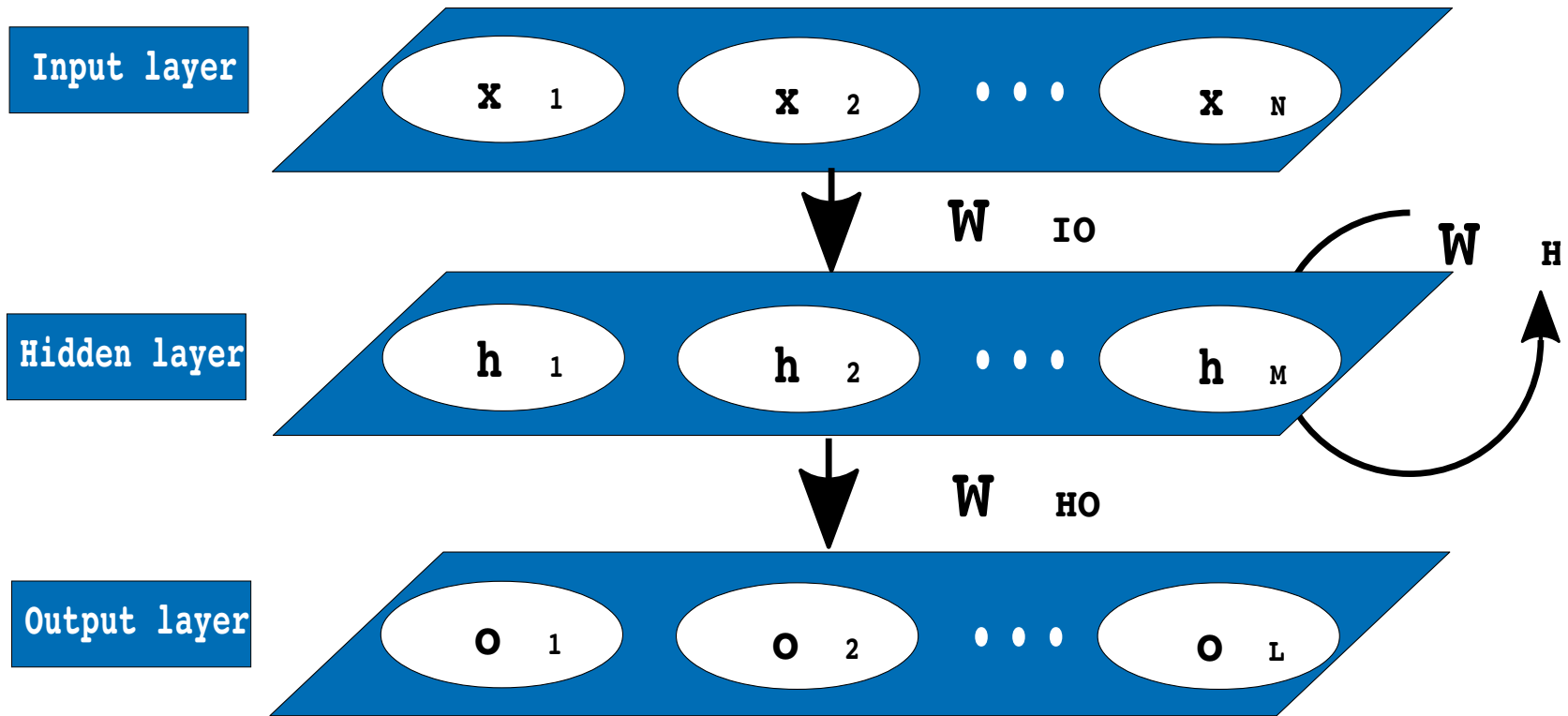


Grid job classification with Convolutional Neural Network

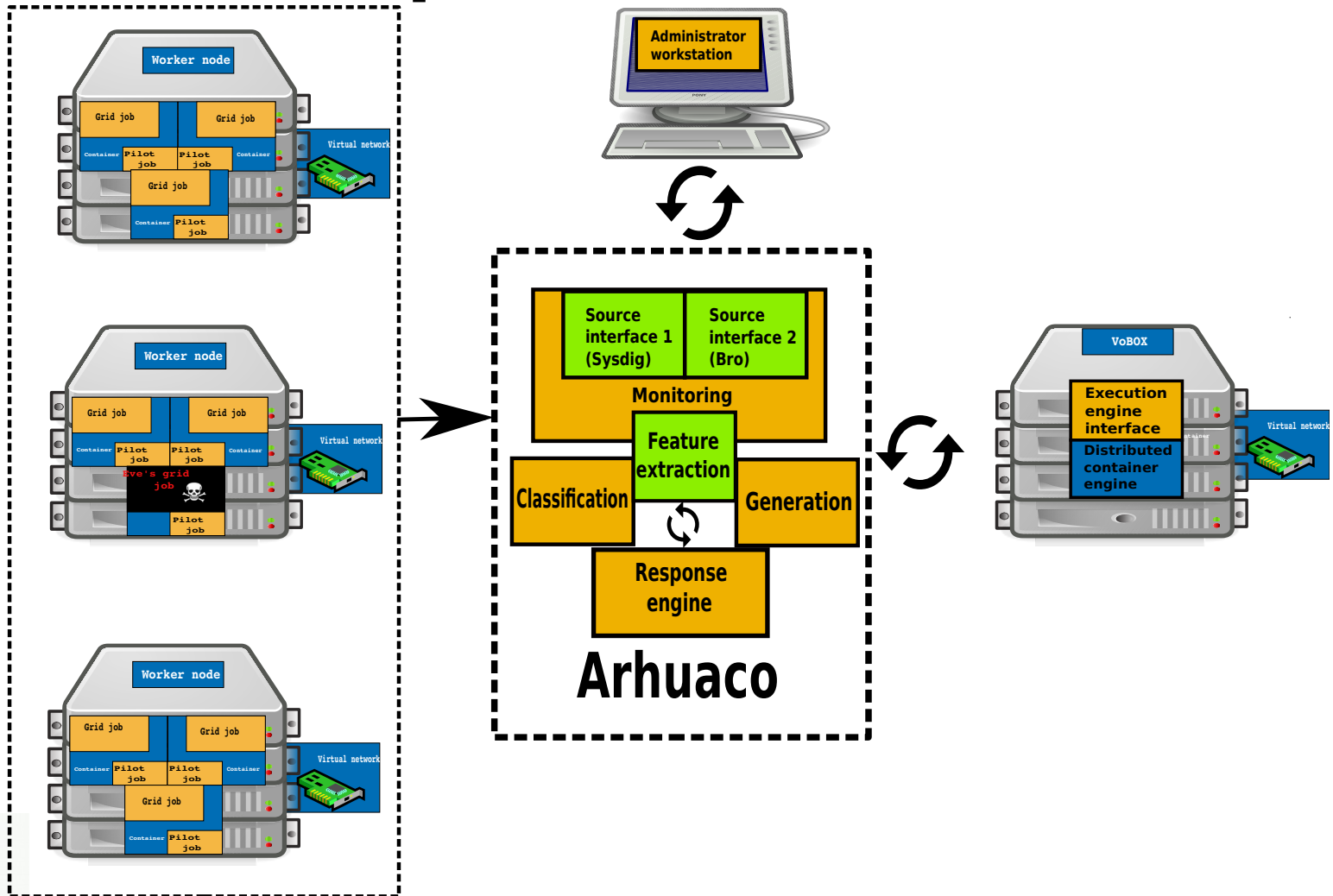
```
Trace log
* file open fd 6 f /etc/passwd name /etc/passwd
  flags 4097 O_RDONLY O_CLOEXEC mode 0
* file lseek fd 6 f /etc/passwd offset 2081 whence 0 SEEK_SET
* file read res 38 data gdm session worker pam/gdm password
* file write res 54 data Traceroute v1.4a5 exploit by sorbo
  sorbox yahoo.com .
* file write res 21 data .telnetd
```



Training data generation with Recurrent Neural Networks



Arhuaco: proof-of-concept implementation



Arhuaco: proof-of-concept Implementation

- Linux Containers: **Docker**, Docker Swarm
- Deep Learning: **Keras**, **Theano**, **TensorFlow**, Python 3.
- Data collection: System Calls - Sysdig, Network connection - The **zeek** Network analysis tool.
- Grid Middleware - **ALICE AliEn**.



Evaluation: Grid Jobs vs Malware

MonALISA Repository for ALICE

Catalogue browser | LEGO Trains ★ | Administration Section | ALICE Reports | Alert XML Feed | Firefox Toolbar | More

/alice/cern.ch/user/a/aliprod/LHC18c11 Welcome agomezra (~) with role agomezra (~)

Permissions	Owner	Timestamp	Size	Filename
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:36	1.424 KB	chunks_1k.txt
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:44	635 B	GeneratorCustom.C
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:55	1.674 KB	JDL
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:32	1.639 KB	JDL_ocdb.jdl
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:44	2.5 KB	JPsiPbPbGenerator.C
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:34	29 KB	QAtrainsim.C
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:44	467 B	rootlogon.C
-rwxr-xr-x	aliprod:aliprod	27 Mar 2018 09:33	5.835 KB	validation.sh

Edit new file 43.15 KB in 8 files

Upload files in this folder (100MB max, multiple selection possible)

Browse... No files selected. Upload...

Create subfolder

Create new folder


virustotal

SHA256: 5ff86d434be5a4011ddcd63b1ddf1ebbb072ad9e27bfc6f40f38dc117cf330d

File name: 341dcb650048862fe07cb53fba4a76ffe9bcd7e_86.tgz

Detection ratio: 21 / 53

Analysis date: 2014-07-22 17:47:44 UTC (3 years, 9 months ago)



Analysis | Additional information | Comments 0 | Votes

Antivirus	Result	Update
Ad-Aware	Application.Linux.BitCoinMiner.A	20140722
AntiVir	LINUX/Procfake	20140722
Avast	ELF:BitCoinMiner-G [Tool]	20140722
BitDefender	Application.Linux.BitCoinMiner.A	20140722
CAT-QuickHeal	Linux.RiskTool.BitCoinMiner.a	20140722
Comodo	Unclassified/Malware	20140722
DrWeb	Linux.CpuMiner.1	20140722
ESET-NOD32	Linux/BitCoinMiner.D	20140722
F-Secure	Application.Linux.BitCoinMiner	20140722



Evaluation: Grid Jobs vs Malware

Dataset	Normal	Malware
System call	12 GB - 127'100,000 lines	8.2 GB - 127'054,763 lines
Network	868 KB - 20,733 lines	108 KB - 2,937 lines

Table 6.1: The complete set of information describing the analyzed grid jobs and malware behavior, as collected log-lines.

Dataset	Training	Validation
System calls traces	10'000,000	100,000
Network traces	20,000	2,000

Table 6.2: Training and validation samples obtained after the feature extraction method.



Results: Impact of Isolation over performance

Setup	ALICE job average runtime (Seconds)	Standard deviation
Native	110.77	10.03
Docker	114.22 (3.12%)	12.58
Arhuaco	117.54 (6.11%)	11.83

Table 6.3: Results of the performance overhead related to the runtime of the ALICE-based jobs.

1600 Grid Jobs in total



Results: Grid job classification – normal vs malware

Testing dataset	CNN ACC	SVM ACC	CNN FPR	SVM FPR
System call	0.9952 (3.24%)	0.9639	0.0068 (−90.10%)	0.0687
Network traces	0.9875 (25.46%)	0.7871	0.0006 (−99.84%)	0.3781

Table 6.8: Comparison of the evaluation metrics between CNN vs. SVM for new testing samples extracted from the system calls and network traces.

- **CNN:** Convolutional Neural Network
- **SVM:** Support Vector Machine
- **FPR:** False Positive rate
- **ACC:** Accuracy



Results: training data generation results

Testing dataset	TPR	SPC	FPR	ACC
Network traces normal	0.9507	0.6219	0.3781	0.7871
Network traces generated	0.9712 (2.16%)	0.6206	0.3839	0.7928 (0.72%)

Table 6.9: Resulting accuracy of the SVM tested with previously unseen data. These results compare the training made with the original network samples vs. the new dataset with generated data.

- **TPR:** Sensitivity or True Positive Rate
- **SPC:** Specificity
- **FPR:** False Positive rate
- **ACC:** Accuracy



Conclusions

- Docker **containers** can be used to isolate and extract behavior information from Grid jobs without big performance impact.
- **Deep learning** is highly effective to identify “malicious” Grid jobs.
- **CNNs with Word2Vec** preprocessing provides improved accuracy than traditional SVM.
- Synthetic **generated data** can improve the training process.



Future work

- Thesis submitted.
- Exploring options: Arhuaco as open source and/or commercial tool.
- UF Grid site probably will not continue operations.
- Future research topics of interest: differential privacy, adversarial machine learning training, privacy preserving intrusion detection.



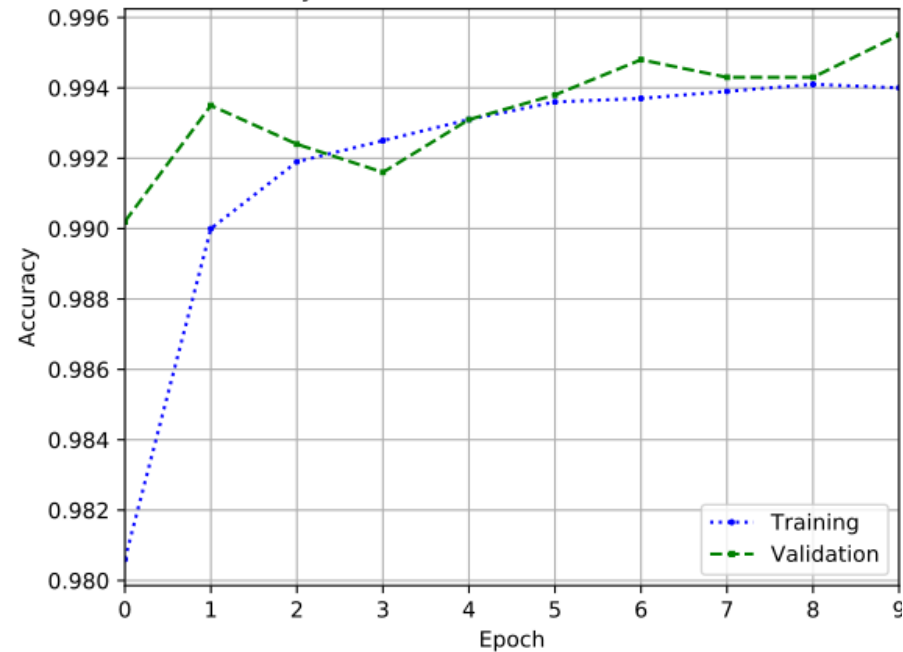


Thank you!

Questions?

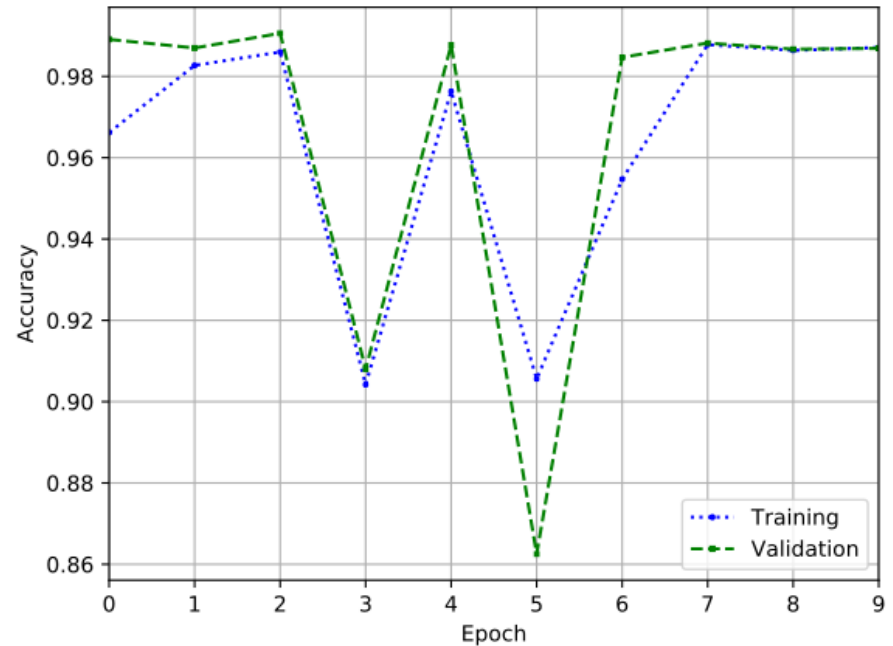
Appendix: CNN results

System call classification with CNN



System Calls

Network trace classification with CNN

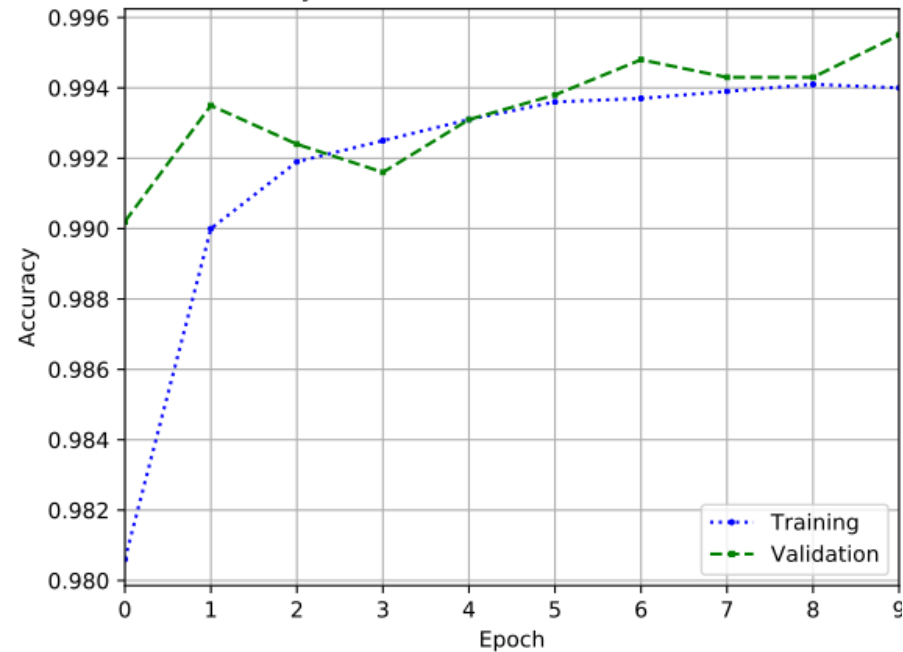


Network connections



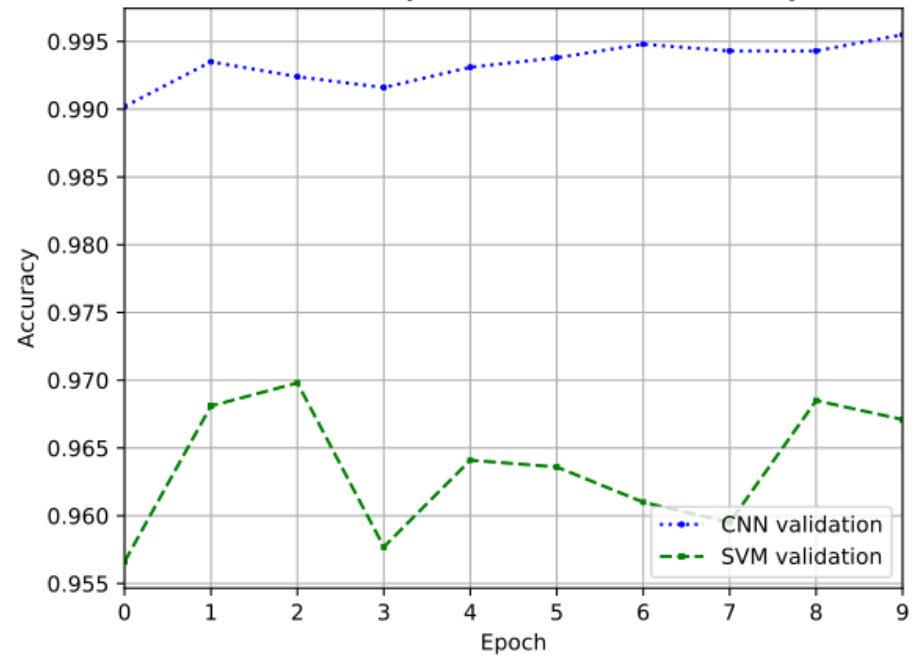
Appendix: System call results

System call classification with CNN



CNN

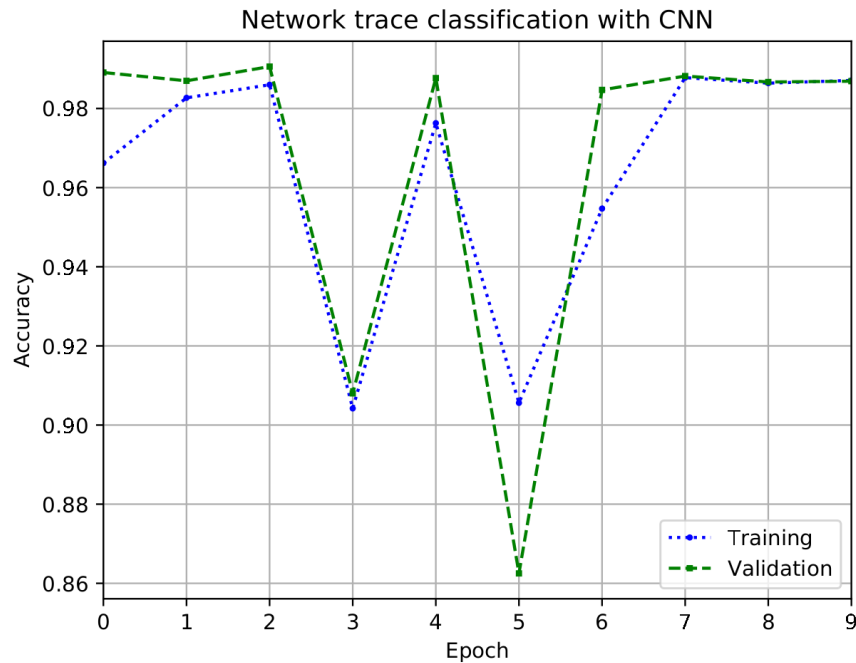
CNN vs SVM system call validation accuracy



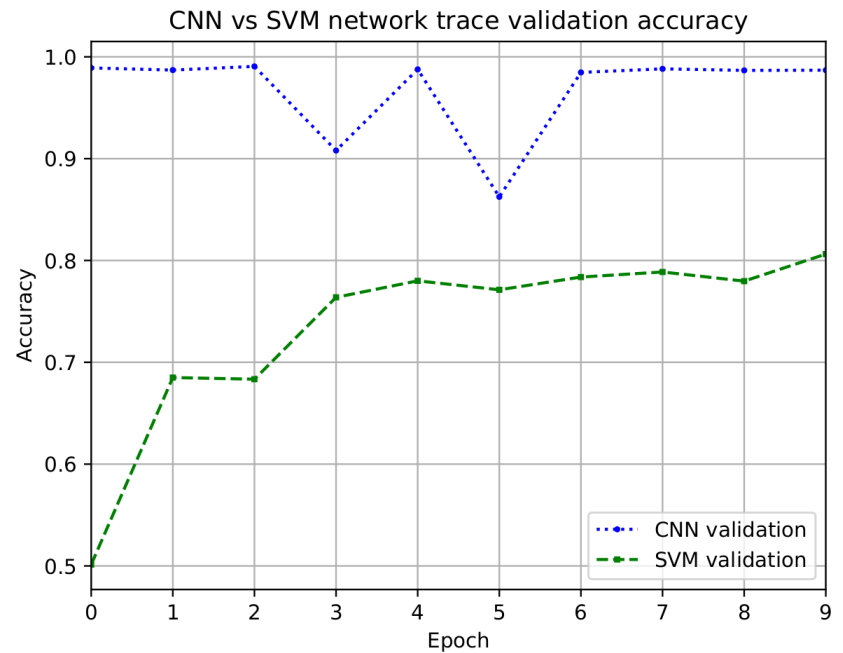
CNN vs SVM



Appendix: Network traces results



CNN



CNN vs SVM



CERN Security Operations Center

