Improving Service Availability of Cloud Systems by Predicting Disk Error

USENIX ATC, July 12, 2018

Yong Xu*, Kaixin Sui, Qingwei Lin, Keceng Jiang,
Wenchi Zhang , Jian-Guang Lou, Dongmei Zhang
Microsoft Research Asia, Beijing, China

Randolph Yao, Yingnong Dang, Murali Chintalapati
Microsoft Azure, Redmond, USA

Hongyu Zhang
The University of Newcastle, Australia

Peng Li
Nankai University, China

Motivation – Towards High Cloud Service Availability

High availability remains one of the top priorities of cloud systems.

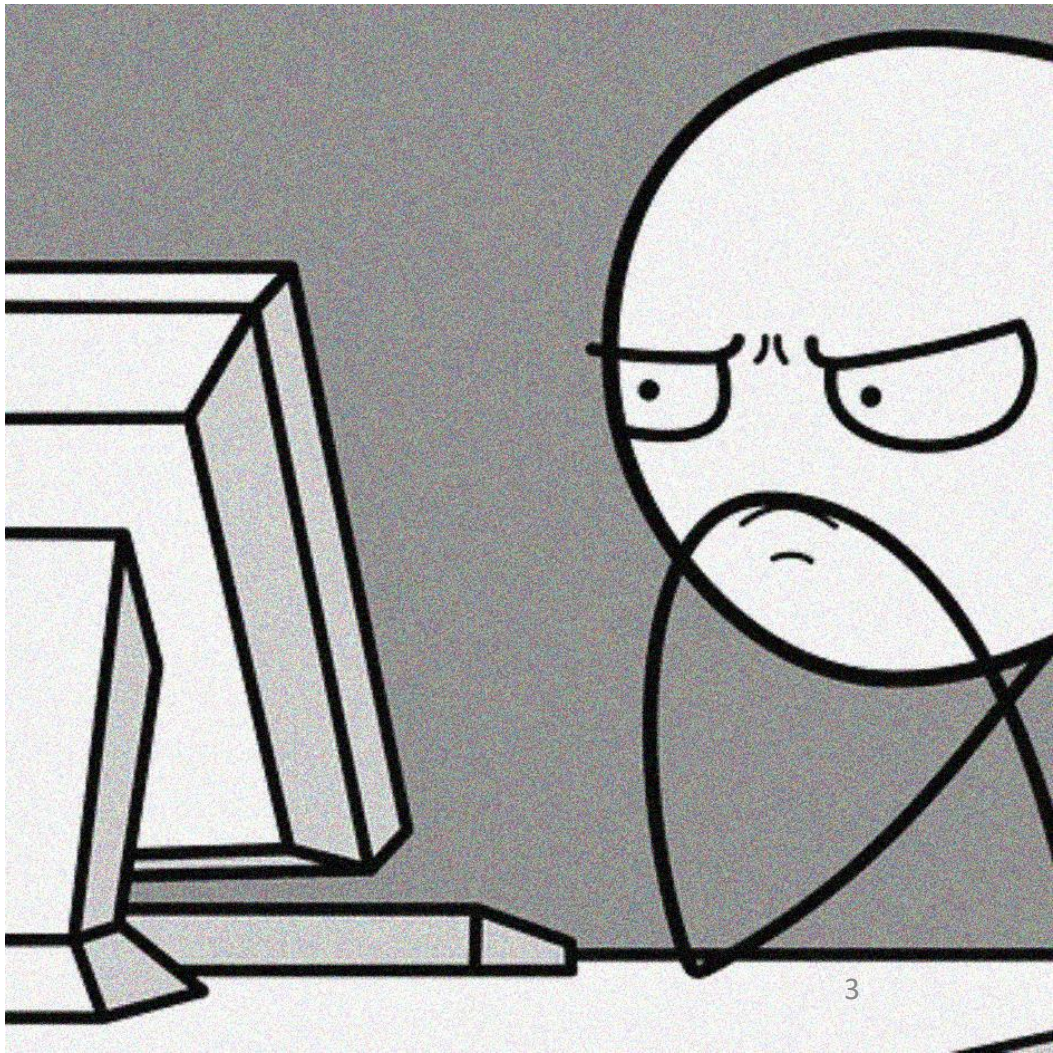
Motivation – Impact of Disk Error on Cloud Service Availability

Unplanned VM downtime is highly painful to customers.

- Hardware issue is one of the top reasons of VM downtime

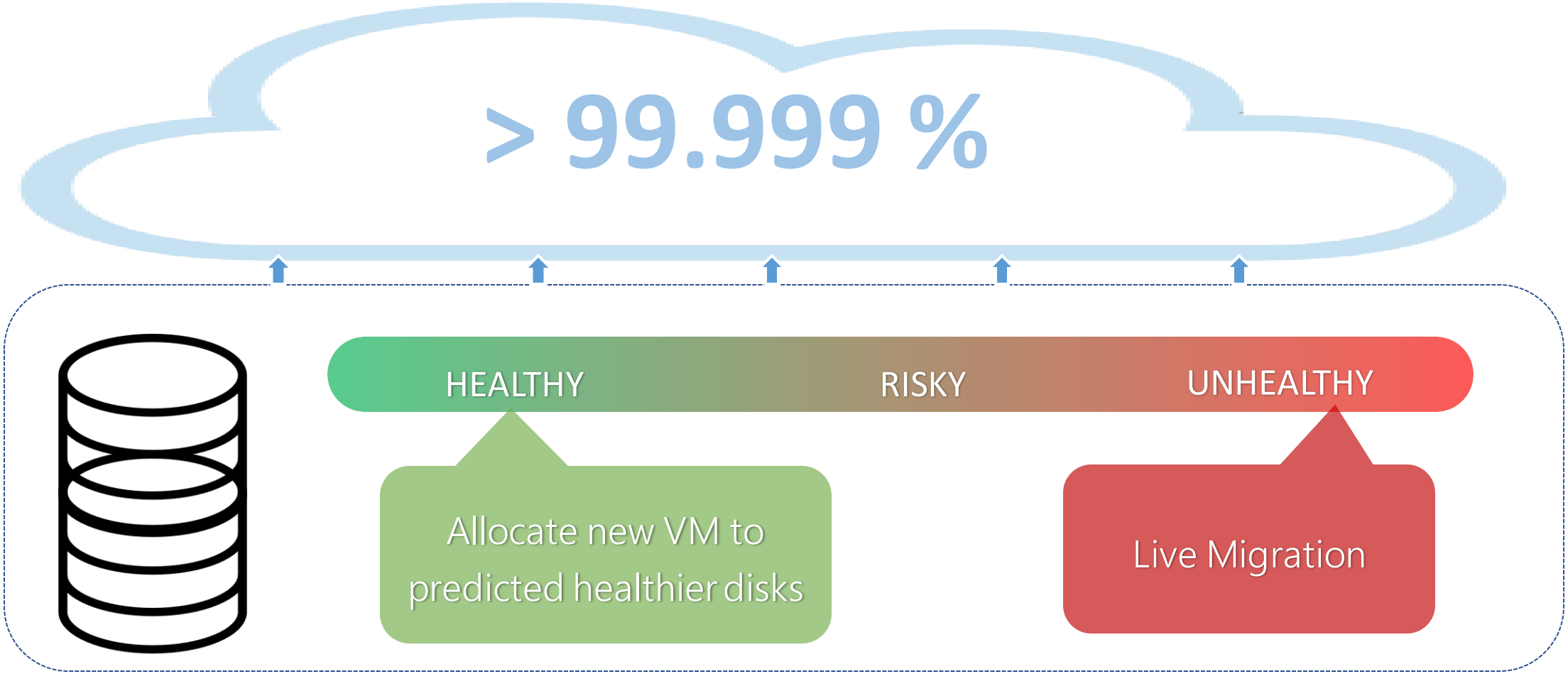
- Disk error contributes most to Hardware issue

- Disk error may result in irreversible data loss disaster



Goal

Improve VM availability by early prediction of disk errors and guide Live Migration (*moving VMs to healthy node without disconnection to the client or application.*



State-of-the-art

Predicting disk errors in industrial settings is difficult.

Disk SMART data

"Self-Monitoring, Analysis and Reporting Technology"

prediction model

Complete failure

Methodology	model	venue
Statistical	Threshold setting	FAST KDD USENIX ATC …
Unsupervised	Clustering Markov chain	
Supervised classification	SVM Neural Network Decision Tree Random Forest	

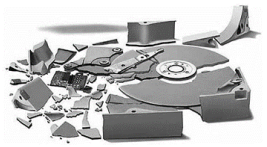
State-of-the-art

Predicting disk errors in industrial settings is difficult.

Disk SMART data

"Self-Monitoring, Analysis and Reporting Technology"

prediction model



Complete failure

Methodology	model	venue
Statistical	Threshold setting	FAST
Unsupervised	Clustering Markov chain	KDD
Supervised classification	SVM Neural Network Decision Tree Random Forest	USENIX ATC ...

No real-production adoption reported in existing work.

Why predicting disk errors in real production is difficult?

The proof of the pudding is in the eating.

- VM downtime occurs far before disk complete failure
- Existing prediction flow(cross-validation guided) goes wrong
- Training with extremely imbalanced health labels of disks is difficult
- …

Insights beyond laboratory work.

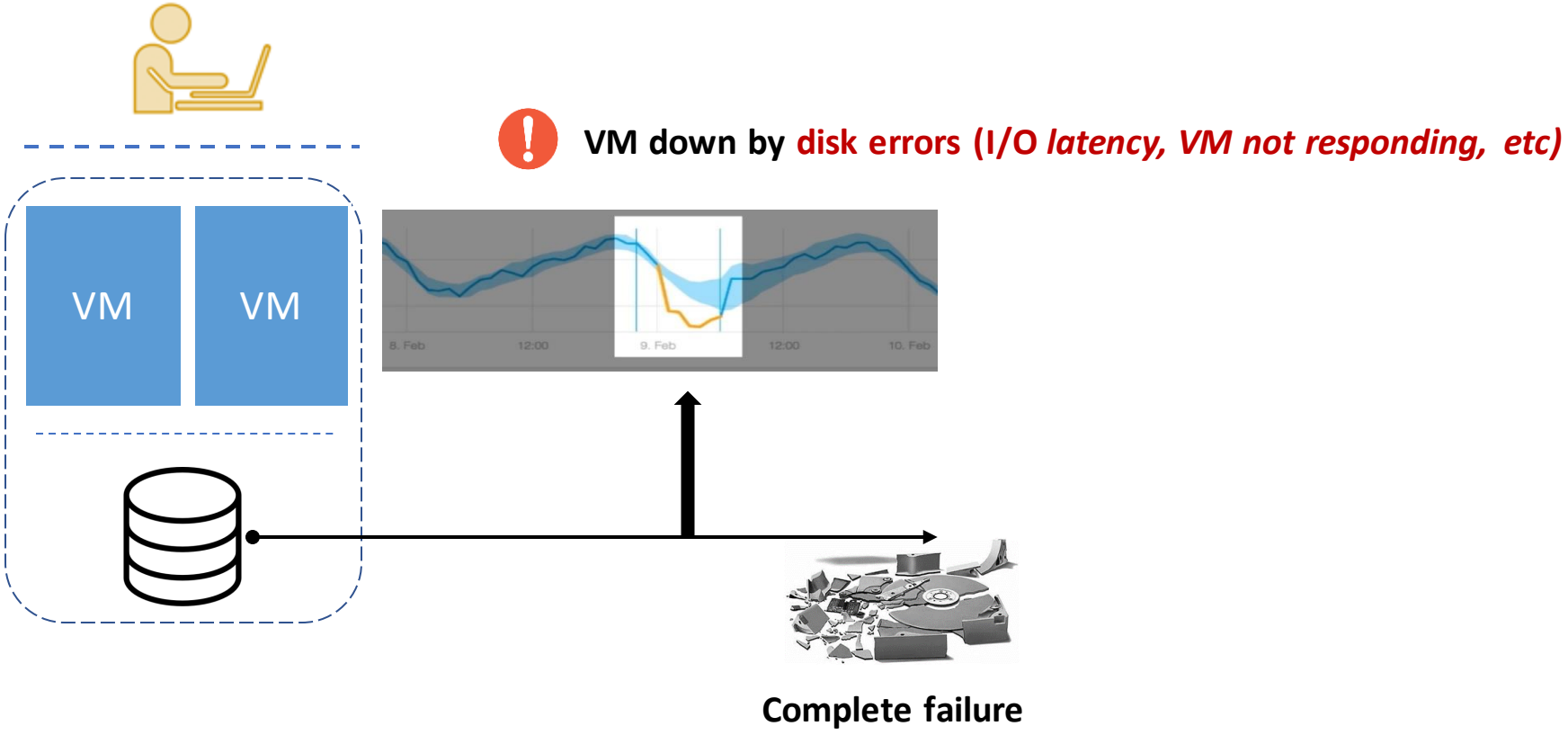
Why predicting disk errors in real production is difficult?

The proof of the pudding is in the eating.

- VM downtime occurs far before disk complete failure
- Existing prediction flow(cross-validation guided) goes wrong
- Training with extremely imbalanced health labels of disks is difficult

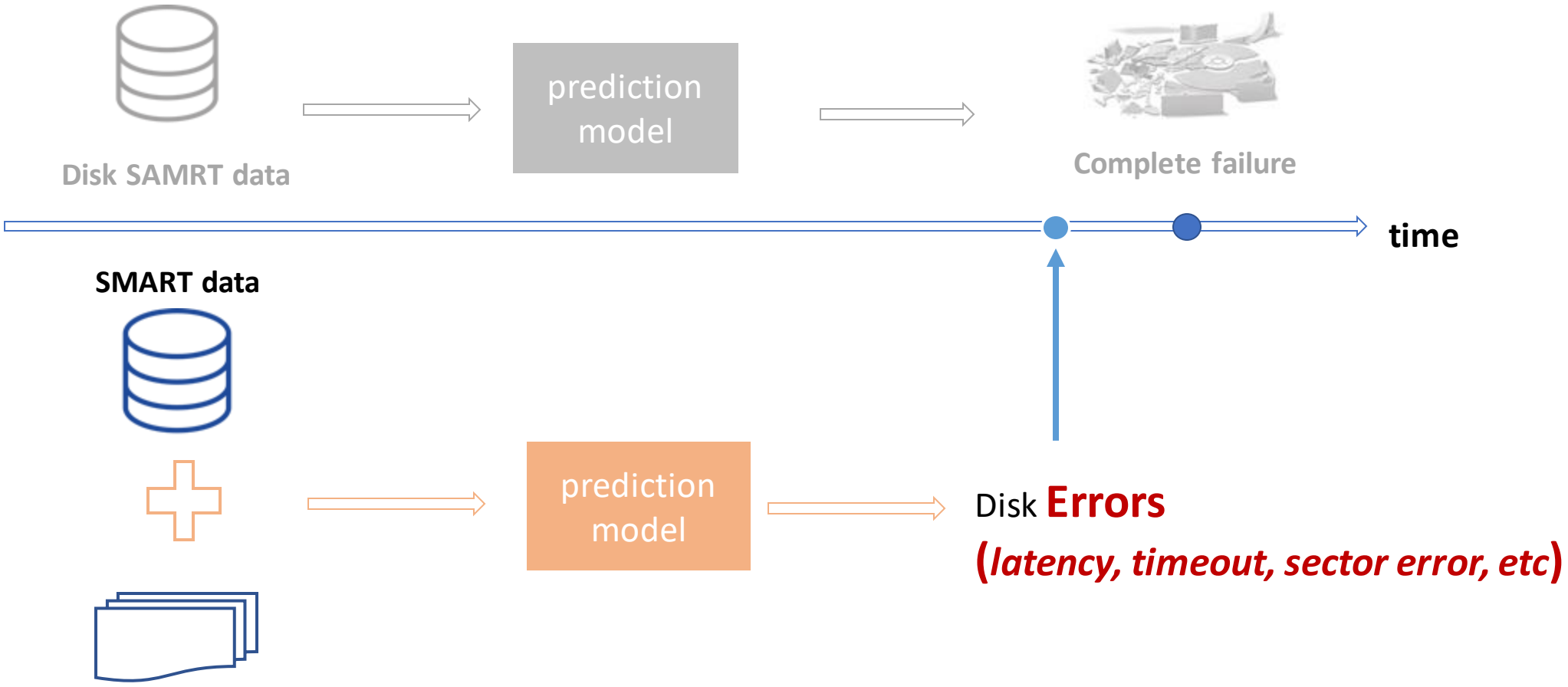
Problem 1 – Predicting complete failure is not helpful to prevent VM downtime

VM downtime occurs far before complete failure of disks.



Solution - Incorporate system-level features

System-level signals manifest earlier symptoms of disk errors.



Why predicting disk errors in real production is difficult?

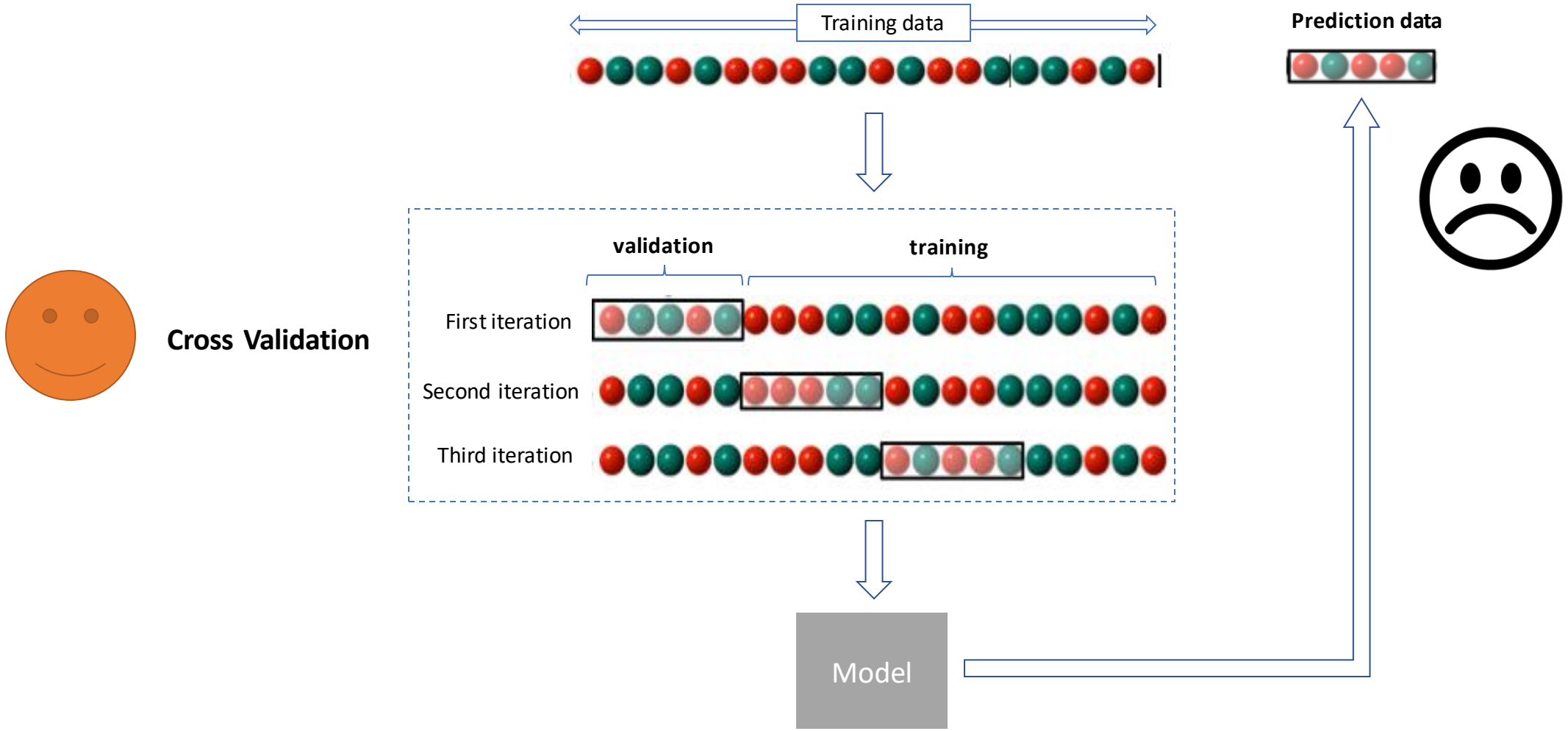
The proof of the pudding is in the eating.

- VM downtime occurs far before disk complete failure
- **Existing prediction flow(cross-validation guided) goes wrong**
- Training with extremely imbalanced health labels of disks is difficult

Problem 2- Cross-Validation Guided prediction goes wrong

State-of-the-art do prediction in cross-validation guided way,

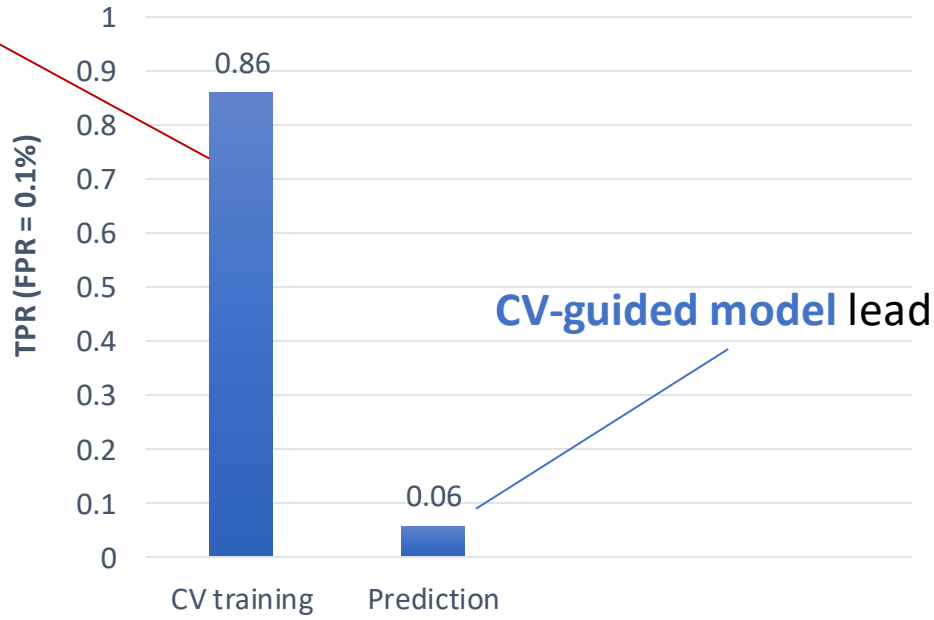
not applicable in real production scenario.



Problem 2- Cross-Validation guided prediction goes wrong

Experiment result shows good result in CV evaluation, but poor result in real online prediction.

Good result of **CV-guided evaluation**.

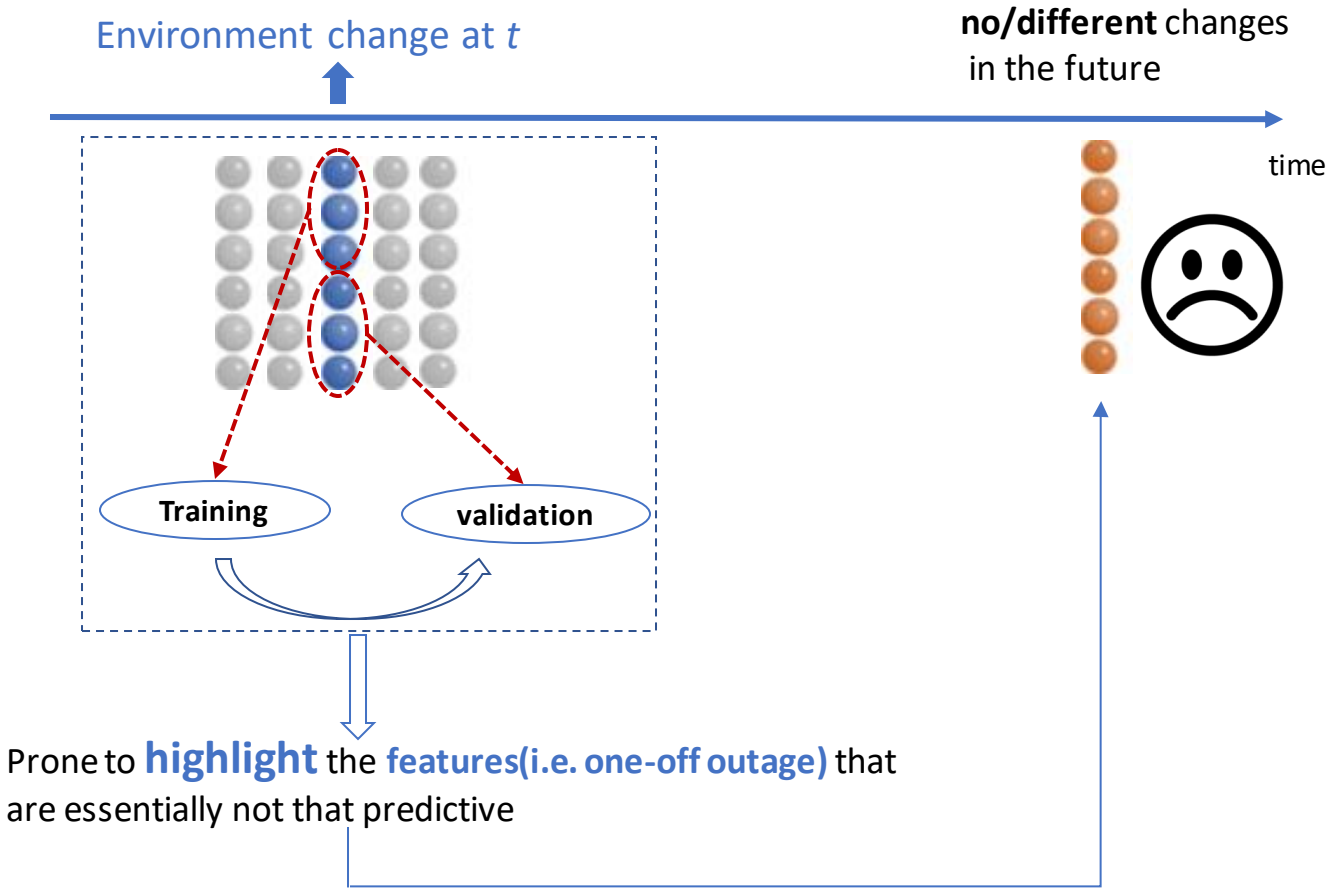


CV-guided model lead to **Low** result in real online prediction

Problem 2- Cross-Validation guided prediction goes wrong

Fundamentally, training phase of Cross-Validation is not applicable for disk error prediction.

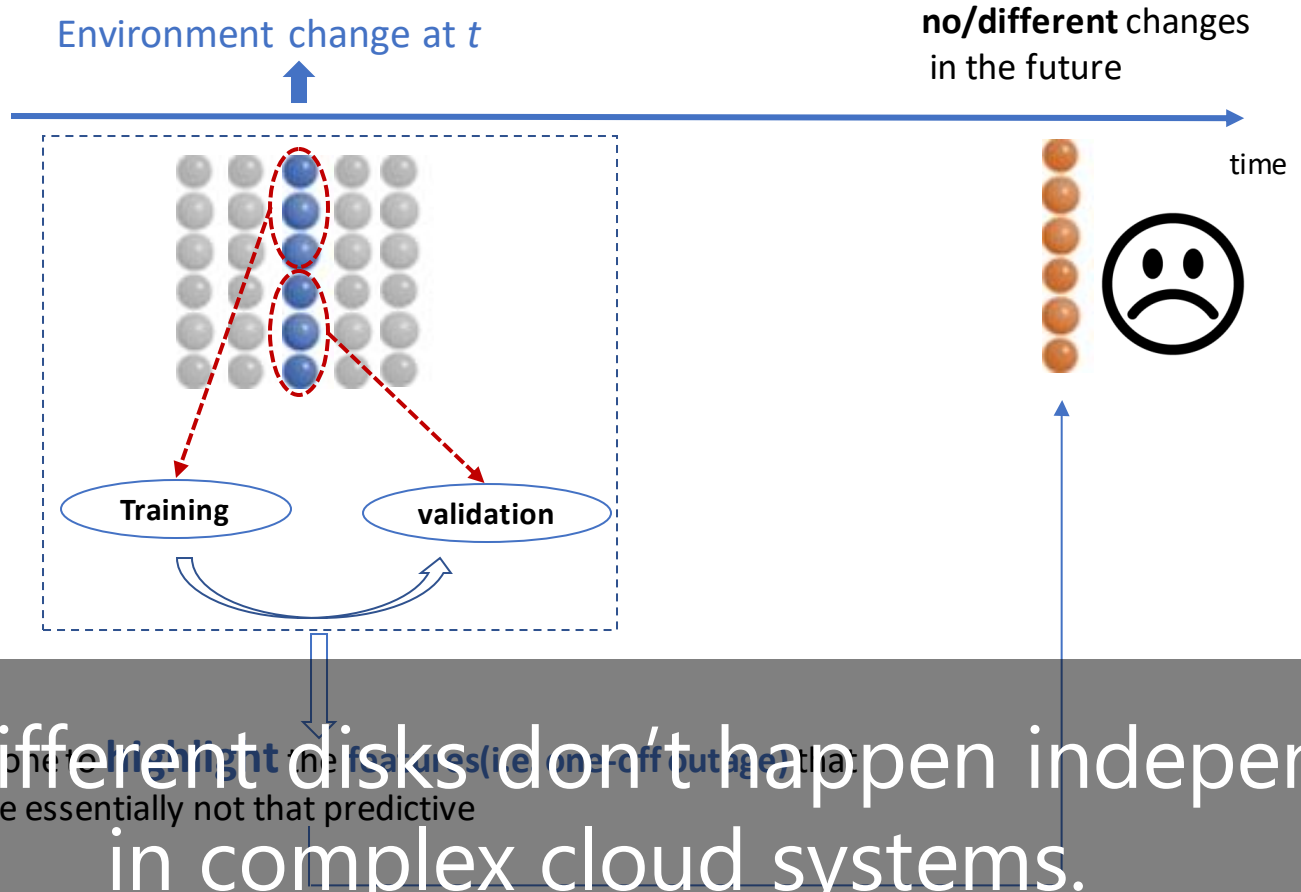
Eg. Rack 3 encounter outage at time t.



Problem 2- Cross-Validation guided prediction goes wrong

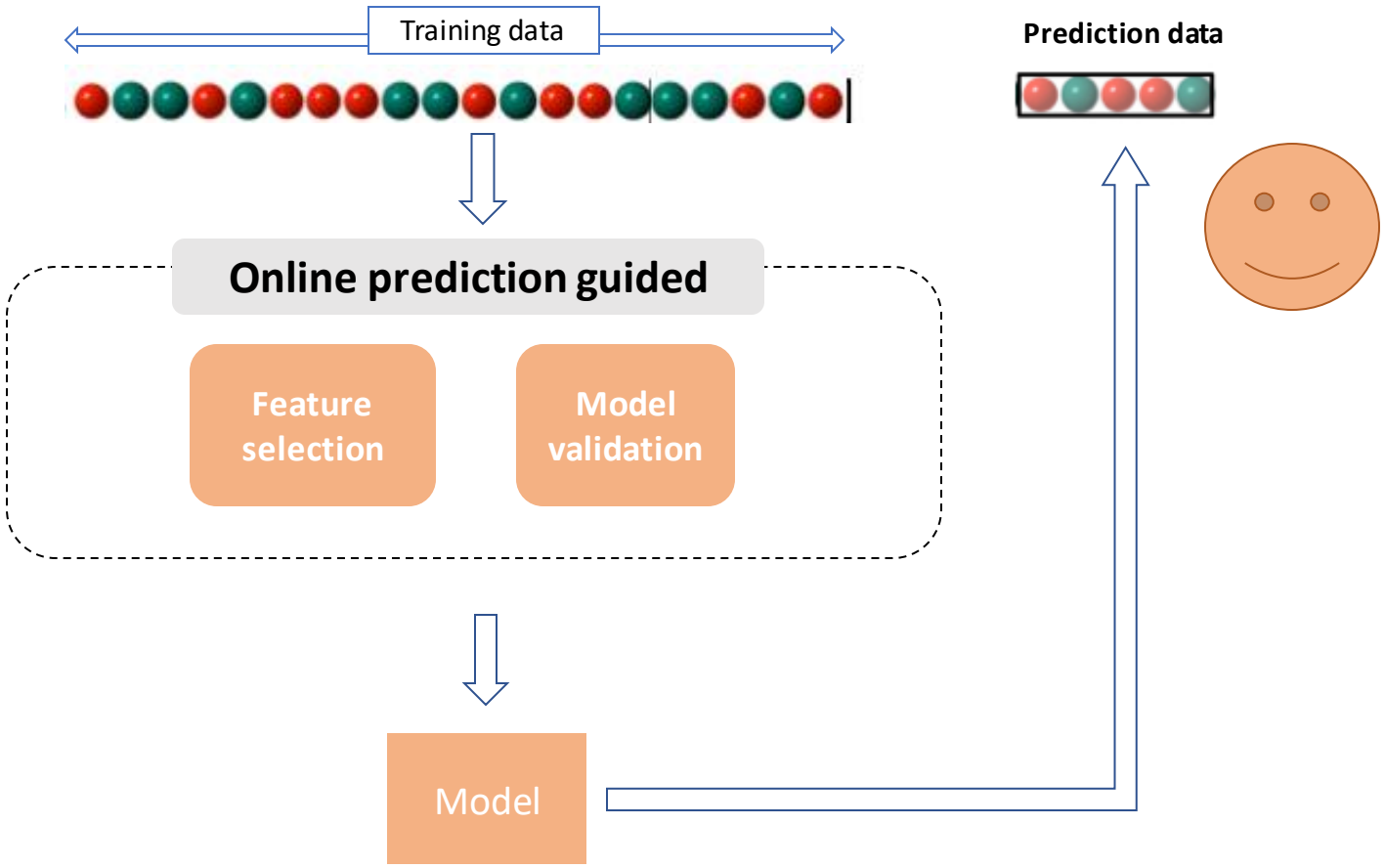
Fundamentally, training phase of Cross-Validation is not applicable for disk error prediction.

Eg. Rack 3 encounter outage at time t.



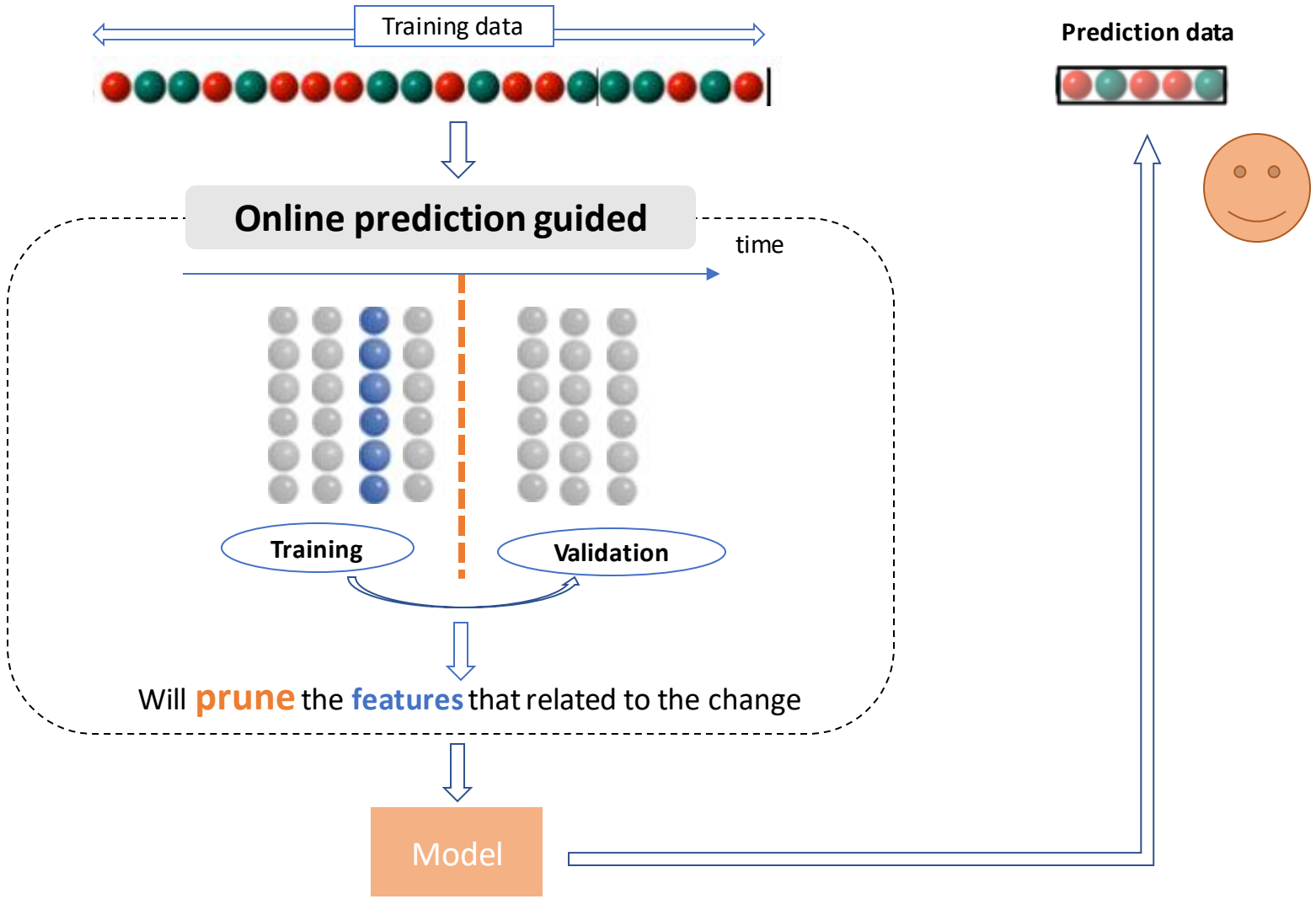
Errors of different disks don't happen independently in complex cloud systems.

Solution – Online prediction guided way



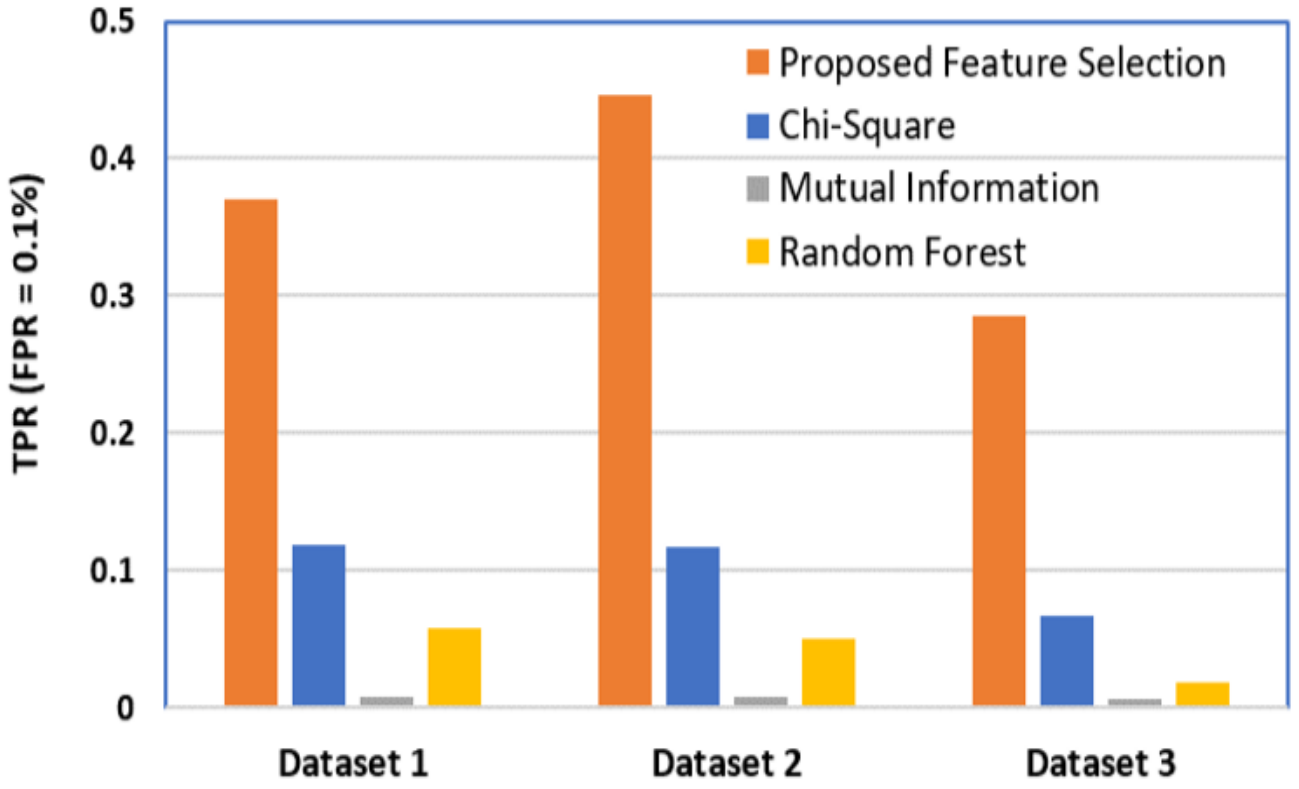
Solution – Online prediction guided way

Strictly separate training and validation set **by time.**



Cross-Validation guided vs. Online prediction guided

Online-prediction guided outperforms.



Why predicting disk errors in real production is difficult?

The proof of the pudding is in the eating.

- VM downtime occur before disk complete failure
- Existing prediction flow(cross-validation guided) go wrong
- **Training with extremely imbalanced health labels of disks is difficult**

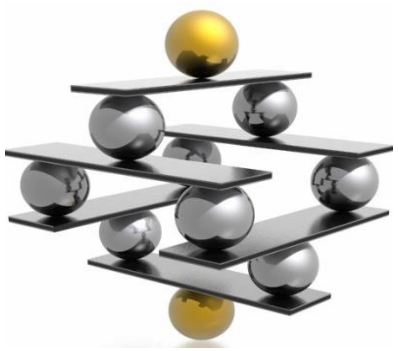
Problem 3 – Extremely imbalanced dataset

Extremely small portion of fault samples leads to low recall using common classification model.

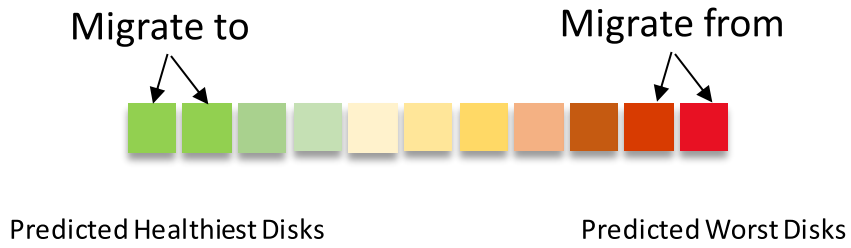
→ prone to predict all to be good → **low recall**

Fault : good
~3 : 10,000

Rethinking the problem



Sensitive to classifier balancing

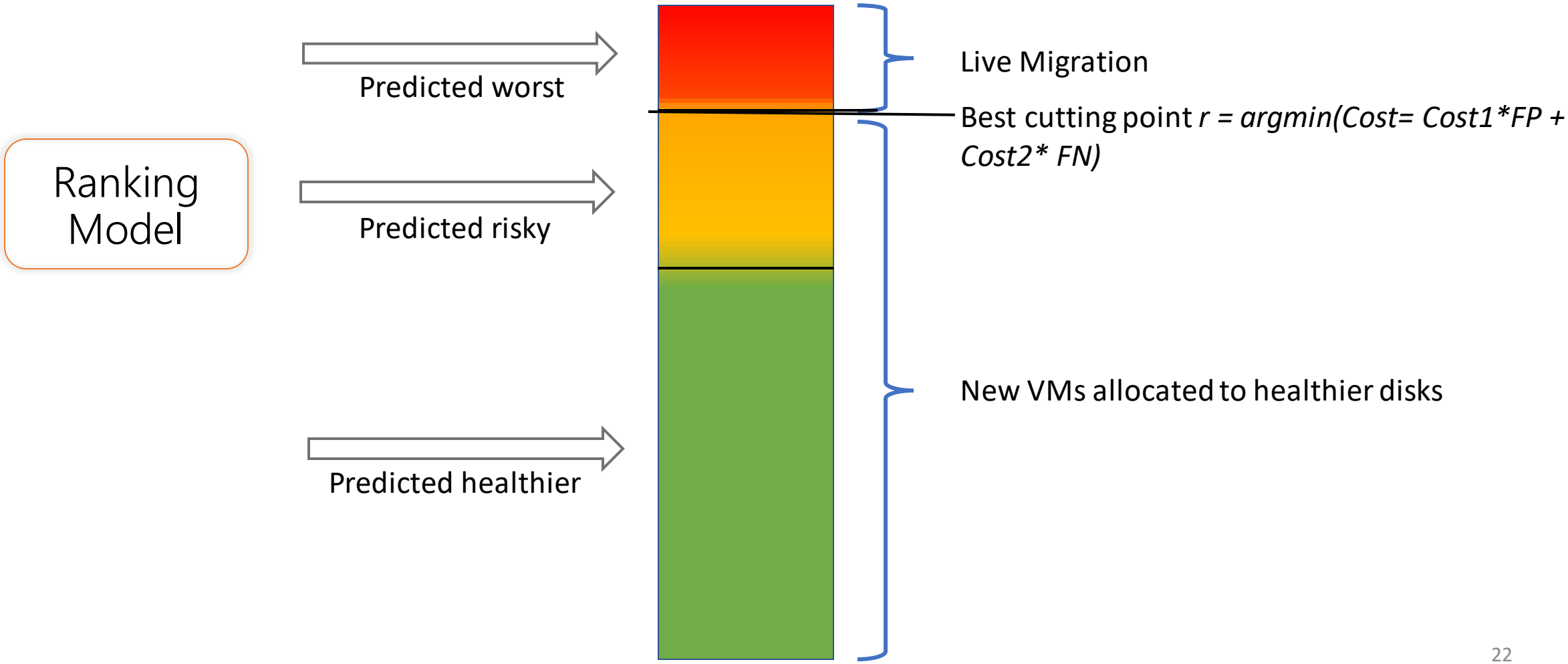


Ordering serves scenario better

Ranking instead of
Classification

Solution - Cost-sensitive ranking model

False predictions, both false positive(FP) and false negative(FN), bring cost to real cloud system.



Evaluation

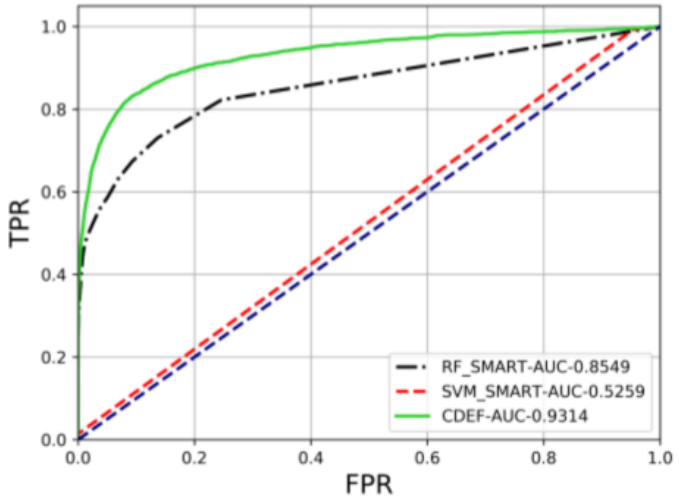
- Dataset
  - Real dataset from Azure
  - Training: October 2017
  - Testing: 3 parts divided from November 2017
  - Healthy disks: faulty disks is ~10,000 : 3

- Setup
  - Data store and process: Microsoft COSMOS
  - Ranking algorithm: FastTree implemented by Microsoft AzureML
  - Windows Server 2012 with Intel CPU E5-4657L v2 @2.40GHz 2.40 with 1.0 TB Memory

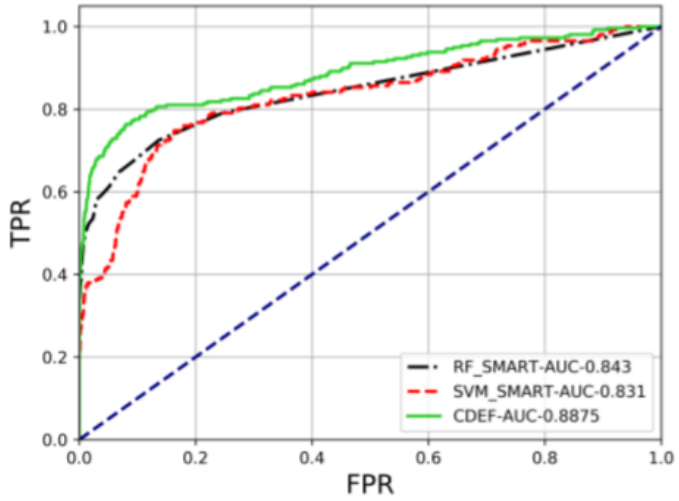
- Evaluation metrics
  - True Positive Rate(TPR) = TP/(TP + FN), under 0.1% False Positive Rate(FPR) = FP/(FP + TN)

Result

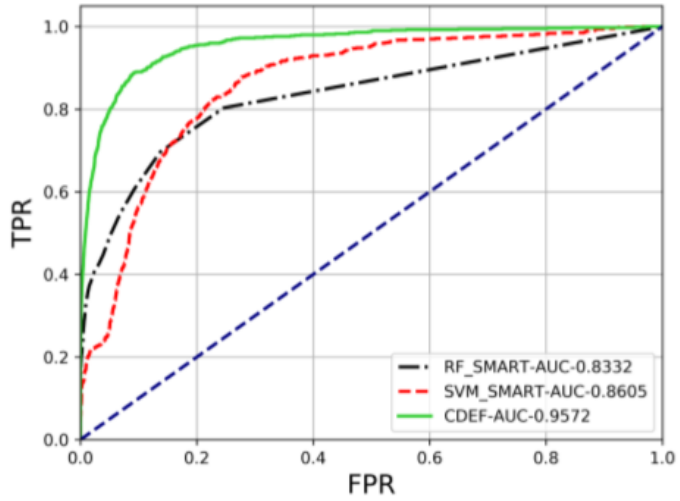
RQ1: How effective is the proposed approach in predicting disk errors?



(a) Dataset 1



(b) Dataset 2



(c) Dataset 3

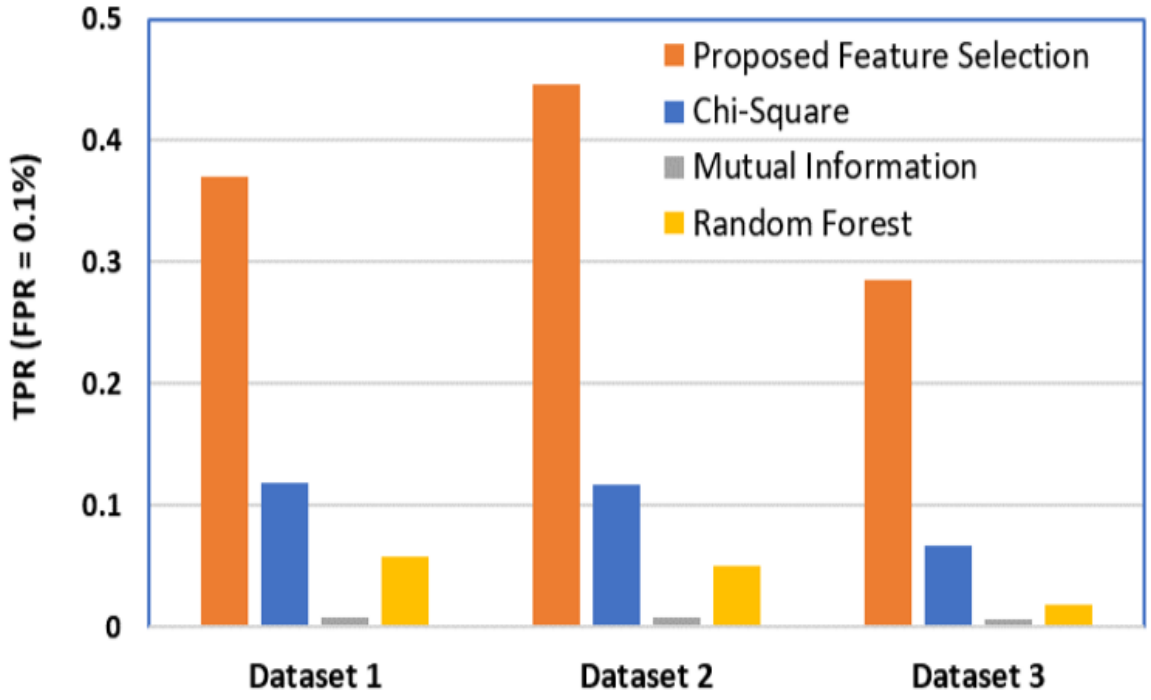
Table 3: Experimental results of CDEF on three datasets

	CDEF		RandomForest		SVM	
	Cost	TPR	Cost	TPR	Cost	TPR
Dataset 1	2508	36.50%	3157	30.51%	2907	15.51%
Dataset 2	234	41.09%	1211	34.11%	258	21.71%
Dataset 3	760	29.67%	1675	18.81%	792	7.20%

42.11% cost(with *Cost1* = 3, *Cost2* = 1) reduction than RandomForest, than **11.5%** SVM.

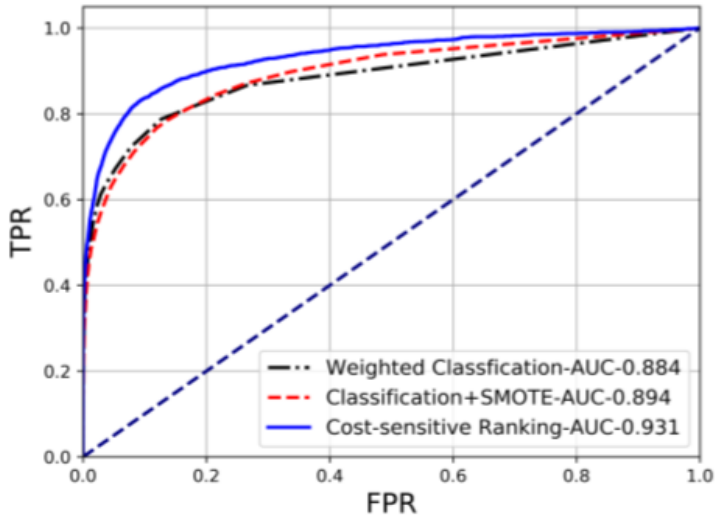
Result

RQ2: How effective is the proposed OnlinePrediction-guided way?

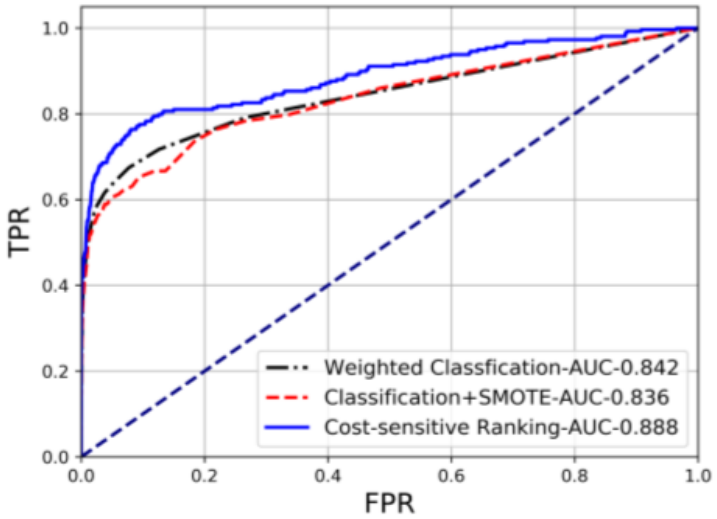


Result

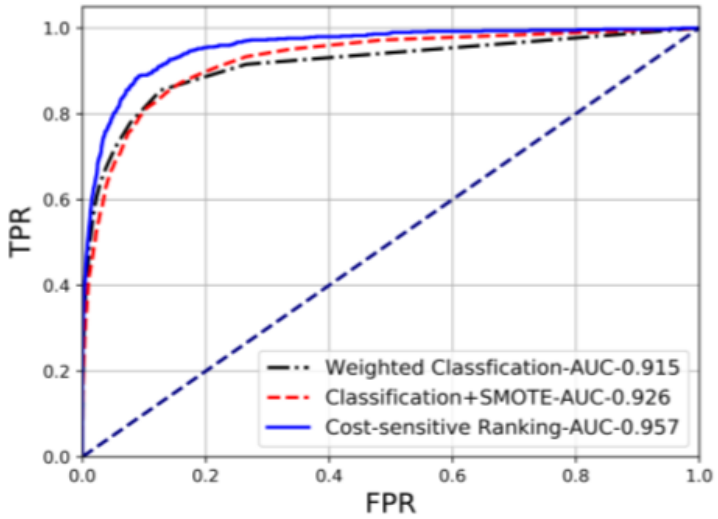
RQ3: How effective is the proposed ranking model?



(a) Dataset 1



(b) Dataset 2



(c) Dataset 3

Table 4: The cost and TPR values (when FPR is 0.1%) achieved by the proposed cost-sensitive ranking model

	Random Guess		Cost-sensitive ranking		Weighted Classification		Classification+SMOTE	
	Cost	TPR	Cost	TPR	Cost	TPR	Cost	TPR
Dataset 1	1447986	0.1%	2508	36.50%	2910	26.52%	9442	24.63%
Dataset 2	1146662	0.1%	234	41.09%	717	27.91%	7812	27.94%
Dataset 3	1446929	0.1%	760	29.67%	1234	17.42%	8239	17.68%

Conclusion

- Point out the CrossValidation-guided prediction does not work for real online prediction in industry settings, and develop an **OnlinePrediction-guided** approach
- Leverage **system-level signals** in additional to SMART data in disk fault prediction
- Propose a **ranking model** to conquer the issue of extremely data imbalance
- **Deployed to large scale industrial cloud system**, Microsoft Azure, and significantly improved Azure service availability