Online classification of performance problems in large distributed systems

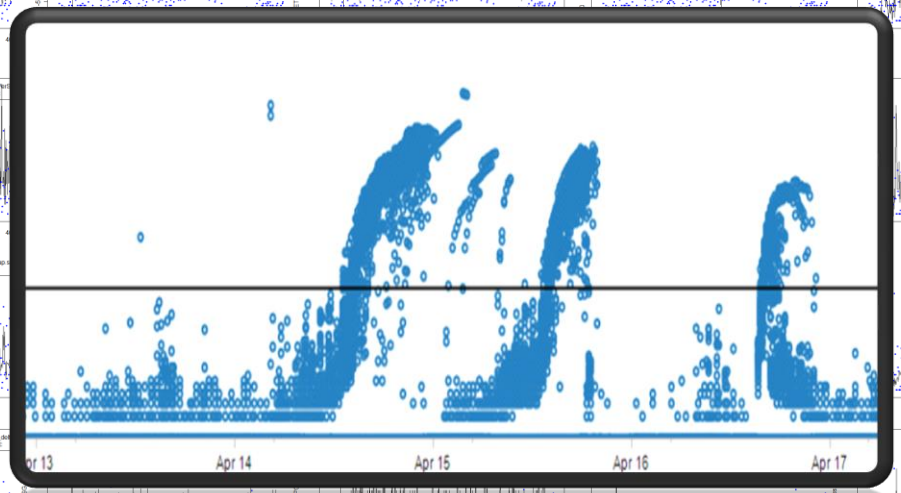
Moises Goldszmidt
Microsoft Research

10/14/09

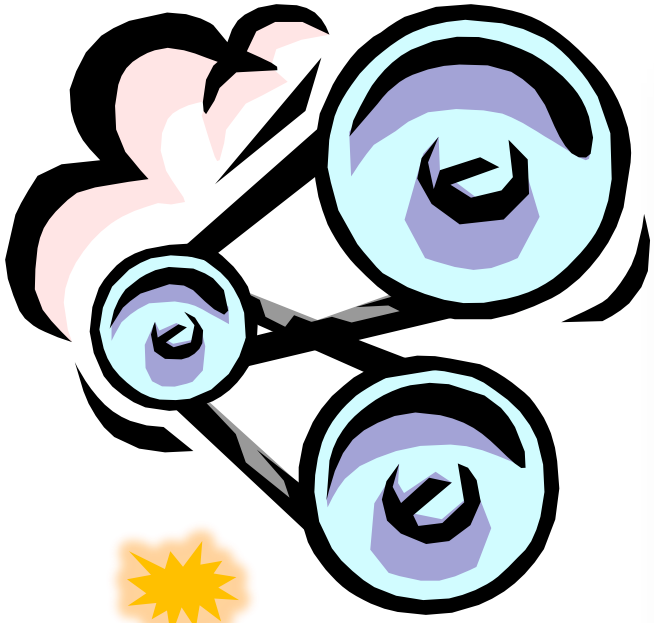
Joint work with:

Peter Bodik, Armando Fox, and Dawn Woodard

1. Scale of distributed systems continues to grow
2. Ability to collect metrics seems to be following the trend
3. Our diagnostic capabilities **are not** following this trend



Diagnosis requires the inspection of 100's of metrics, on 300+ machines



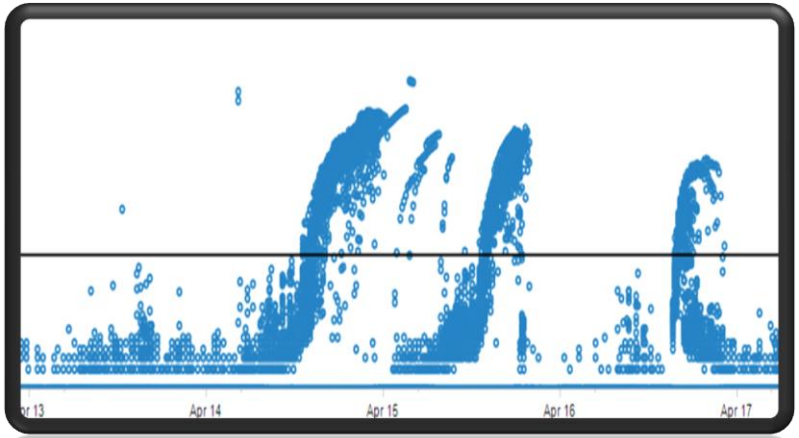
The problem is a backup in the “deliver 3 queue”
The solution is to
Rebalance load
Delay spam post-processing

Datacenter operator

Assumption:
There is a "pattern" of each problem composed by some "function" of the monitored metrics

We use "fingerprints" to refer to the encoding of such patterns

Hypothesis:
We can automatically extract these fingerprints, and use them to uniquely identify each problem



4/14/2008 14:00	4/14/2008 13:30	0	1	0	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 13:45	0	1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:00	1	1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:15	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:30	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:45	1	-1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 15:00	1	0	1	1	1	1	0	1	0	1	0	0	0	0

	5/2008 12:30	0	1	0	0	0	0	0	0	0	1	0	0	1	0
	5/2008 12:45	0	1	0	0	1	1	0	0	0	1	0	0	0	0
	5/2008 13:00	1	1	1	1	1	1	0	1	0	1	0	0	0	0
	5/2008 13:15	1	1	1	1	1	1	0	1	0	1	0	0	0	0
	5/2008 13:30	1	0	1	1	1	1	0	1	0	1	0	0	0	0
	5/2008 13:45	1	1	1	1	1	1	0	1	0	1	0	0	0	0
	5/2008 14:00	1	0	1	1	1	1	0	1	0	1	0	0	0	0

Deliver 3 (queue backup)

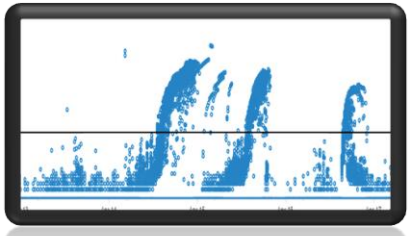
4/17/2008 13:45	4/17/2008 13:15	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 13:30	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 13:45	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:00	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:15	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:30	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:45	1	1	1	1	1	1	0	1	0	1	0	0	0	0

Database Replication

1/ 8 21:30	1/31/2008 21:00	0	0	0	0	0	0	0	0	0	0	0	1	0	
1/3 8 21:30	1/31/2008 21:15	0	0	0	0	0	0	1	1	0	0	0	0	0	
1 1/2 8 21:30	1/31/2008 21:30	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	
1/31/2008 21:30	1/31/2008 21:45	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	
1/31/2008 21:30	1/31/2008 22:00	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	
1/31/2008 21:30	1/31/2008 22:15	-1	-1	0	0	0	0	1	1	0	1	0	0	0	
1/31/2008 21:30	1/31/2008 22:30	0	0	0	0	0	0	1	1	0	1	0	0	0	

Value proposition

Transform diagnosis into automated pattern match to the closest fingerprint



1

2

4/17/2008 13:45	4/17/2008 13:15	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 13:30	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 13:45	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:00	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:15	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:30	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/17/2008 13:45	4/17/2008 14:45	1	1	1	1	1	1	0	1	0	1	0	0	0	0

3

Deliver 3 (queue backup)

4/14/2008 14:00	4/14/2008 13:30	0	1	0	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 13:45	0	1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:00	1	1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:15	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:30	1	0	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 14:45	1	-1	1	1	1	1	0	1	0	1	0	0	0	0
4/14/2008 14:00	4/14/2008 15:00	1	0	1	1	1	1	0	1	0	1	0	0	0	0

Deliver 3 (queue backup)

4/15/2008 12:00	4/15/2008 12:30	0	1	0	0	0	0	0	0	1	0	0	1	0
4/15/2008 12:00	4/15/2008 12:45	0	1	0	0	1	1	0	0	1	0	0	0	0
4/15/2008 13:00	4/15/2008 13:00	1	1	1	1	1	1	0	1	0	1	0	0	0
4/15/2008 13:00	4/15/2008 13:15	1	1	1	1	1	1	0	1	0	1	0	0	0
4/15/2008 13:00	4/15/2008 13:30	1	0	1	1	1	1	0	1	0	1	0	0	0
4/15/2008 13:00	4/15/2008 13:45	1	1	1	1	1	1	0	1	0	1	0	0	0
4/15/2008 13:00	4/15/2008 14:00	1	0	1	1	1	1	0	1	0	1	0	0	0

Database Replication

1/31/2008 21:30	1/31/2008 21:00	0	0	0	0	0	0	0	0	0	0	0	1	0
1/31/2008 21:30	1/31/2008 21:15	0	0	0	0	0	0	1	1	0	0	0	0	0
1/31/2008 21:30	1/31/2008 21:30	-1	-1	0	0	-1	0	1	1	0	1	0	0	0
1/31/2008 21:30	1/31/2008 21:45	-1	-1	0	0	-1	0	1	1	0	1	0	0	0
1/31/2008 21:30	1/31/2008 22:00	-1	-1	0	0	-1	0	1	1	0	1	0	0	0
1/31/2008 21:30	1/31/2008 22:15	-1	-1	0	0	0	0	1	1	0	1	0	0	0
1/31/2008 21:30	1/31/2008 22:30	0	0	0	0	0	0	1	1	0	1	0	0	0

1. Performance problem is detected
2. New fingerprint is generated (real time)
3. Fingerprint is matched against database
4. Solution/repair is generated

4

The problem is a backup in the “deliver 3 queue”
The solution is to
- Rebalance load
- Delay spam post-processing

Remarks…

- Problems will recur often enough
  - Scale and utilization
  - Final resolution may be outside governance
    - Hardware vendor
    - Another data center
    - Priorities on software release
- Mapping from pattern to action ala ROC
  - Start with a linear (in cost) set of actions

    Reboot → reimage → human intervention
  - Recognize patterns and effectiveness of interventions
  - Modified policy
  - Repeat

Fingerprints

time	latency_pf2_cnt_q0.5	latency_pf3_cnt_q0.5	loadavg_1min_q0.5	ms_deliver1_cnt_q0.5	ms_deliver3_cnt_q0.5	ms_smtpd_cnt_q0.5	nAlerts_q0.5	pf1_active_cnt_q0.5	pf1_connect_q0.5	pf1_incoming_cnt_q0.5	pf1_timeout_CONNECT_q0.5	pf2_deferred_cnt_q0.5	type_85_block_cnt_q0.5	type_87_block_cnt_q0.5
1/31/2008 21:00	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1/31/2008 21:15	0	0	0	0	0	0	1	1	0	0	0	0	0	0
1/31/2008 21:30	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	0
1/31/2008 21:45	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	0
1/31/2008 22:00	-1	-1	0	0	-1	0	1	1	0	1	0	0	0	0
1/31/2008 22:15	-1	-1	0	0	0	0	1	1	0	1	0	0	0	0
1/31/2008 22:30	0	0	0	0	0	0	1	1	0	1	0	0	0	0

Fingerprint: a snapshot- summary of the vital metrics of the data center

Summary: the state of a metric is summarized in two ways
1) Across machines → quantiles
2) Across time → hot/cold thresholds

Vital metrics: only the relevant metrics

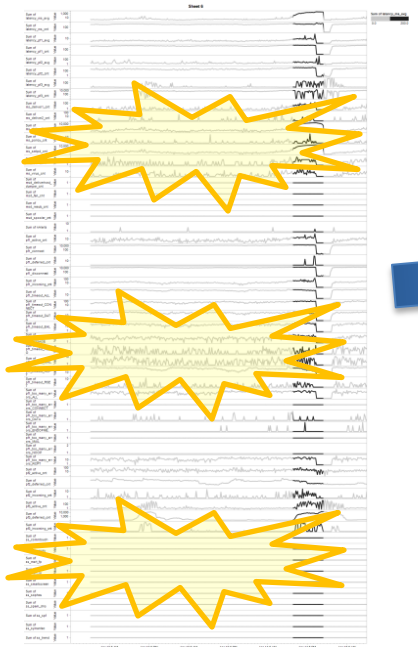
Summary across machines

- Need an effective representation of the "state" of a metric across ALL the machines in the datacenter
- Quantiles are:
  - Robust to outliers
  - Easy to estimate
  - More compact than full P

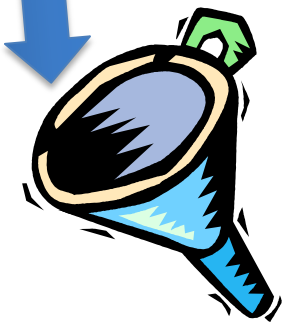
Summary across time

- Discretize the quantiles into 3 states:
  Hot → abnormally high
  Cold → abnormally low
  Normal
- Method: Establish a threshold –
  E.g. 4% of extreme values
- Bypases a lot of normalization issues and also helps with stat modeling (later)

Finding the “vital” metrics

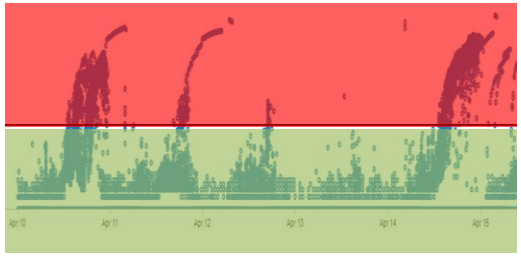


1. Induce “simple” models between metrics and crises
2. Collect the set of most frequent metrics in the last N crises



Logistic regression

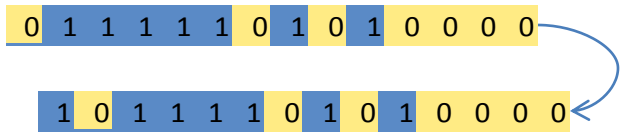
Add L1 Regularization



Pattern matching → crisis ID

Use a statistical model, based on probability for:
a) Crisis modeling
b) Crisis clustering

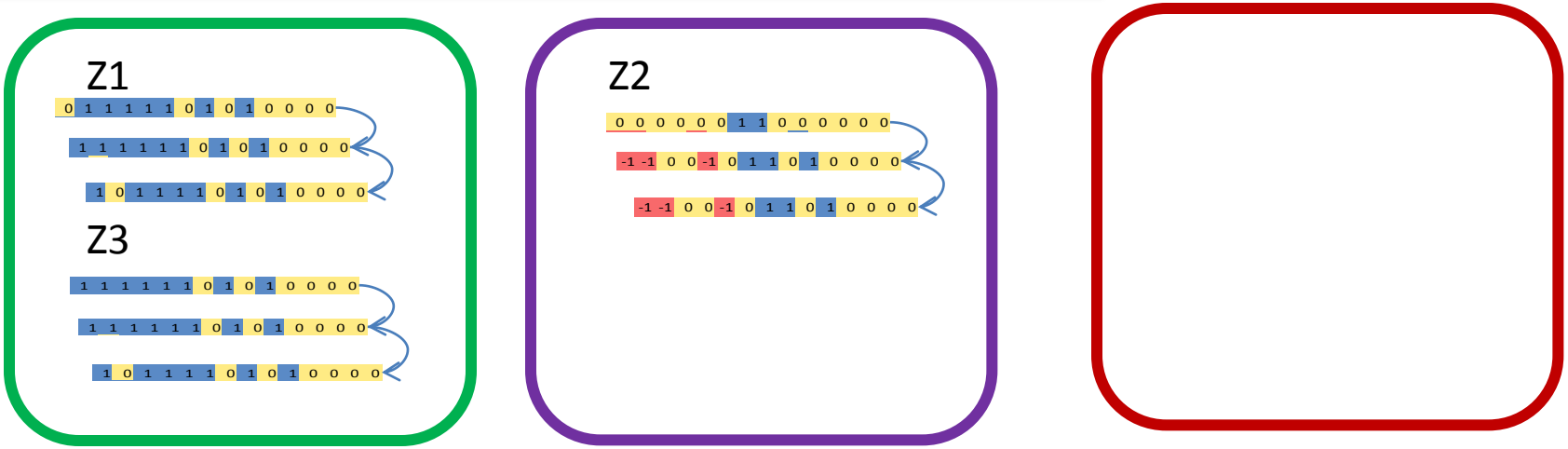
$Pr(Z_{new}=k \mid \{Z\},D)$



.67 .95
.23 .04
.10 .01

The crises is modeled as a time series of fingerprints
Each crisis is associated with a "type" Z_i

To model the crises clustering induce a probability over $\{Z_i\}$



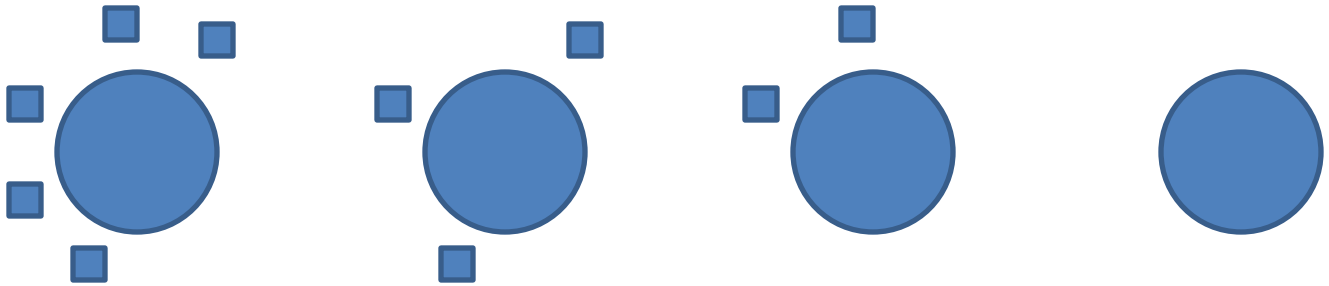
E.g. Pr(Z1 = Z2 = Z3 =) = 0.83

Online clustering

Drichlet process mixture modeling or “the chinese restaurant process”

$$\pi(Z_i = k \mid \{Z_{i'}\}_{i'<i}) \propto \begin{cases} \alpha : if\ (k = new) \\ \sum_{i'<i} 1(Z_{i'} = k) : else \end{cases}$$

Each observation is a new guest who either sits at an occupied table with prob. Proportional to the number of guests a that table, or sits at an empty table



Some math – putting it all together

Prior on clusters Likelihood of data Prior on params

$$\pi(\{Z_i\}_{i=1}^{L} \mid D) \propto \pi(\{Z_i\}_{i=1}^{L}) \int_{\theta} \pi(D \mid \{Z_i\}_{i=1}^{L}, \theta)\pi(\theta \mid \{Z_i\}_{i=1}^{L})$$

- The integral is solved in closed form
- Use MCMC for inference for the restricted space of $\pi(\{Z_i\}_{i=1}^{L})$

(Conceptually) matching is solved!!
On each new crisis compute $\pi(\{Z_i\}_{i=1}^{L+1} \mid D_{new}, D)$

Approximate inference

$$\pi(Z_{new} = Z_i \mid D, D_{new}) = \sum_{\{Z_i\}_{i=1}^L} \pi(Z_{new} = Z_i \mid \{Z_i\}_{i=1}^L, D, D_{new}) \pi(\{Z_i\}_{i=1}^L \mid D, D_{new})$$

$$\pi(Z_{new} = Z_i \mid D, D_{new}) \approx \sum_{\{Z_i\}_{i=1}^L} \pi(Z_{new} = Z_i \mid \{Z_i\}_{i=1}^L, D, D_{new}) \pi(\{Z_i\}_{i=1}^L \mid D)$$

offline

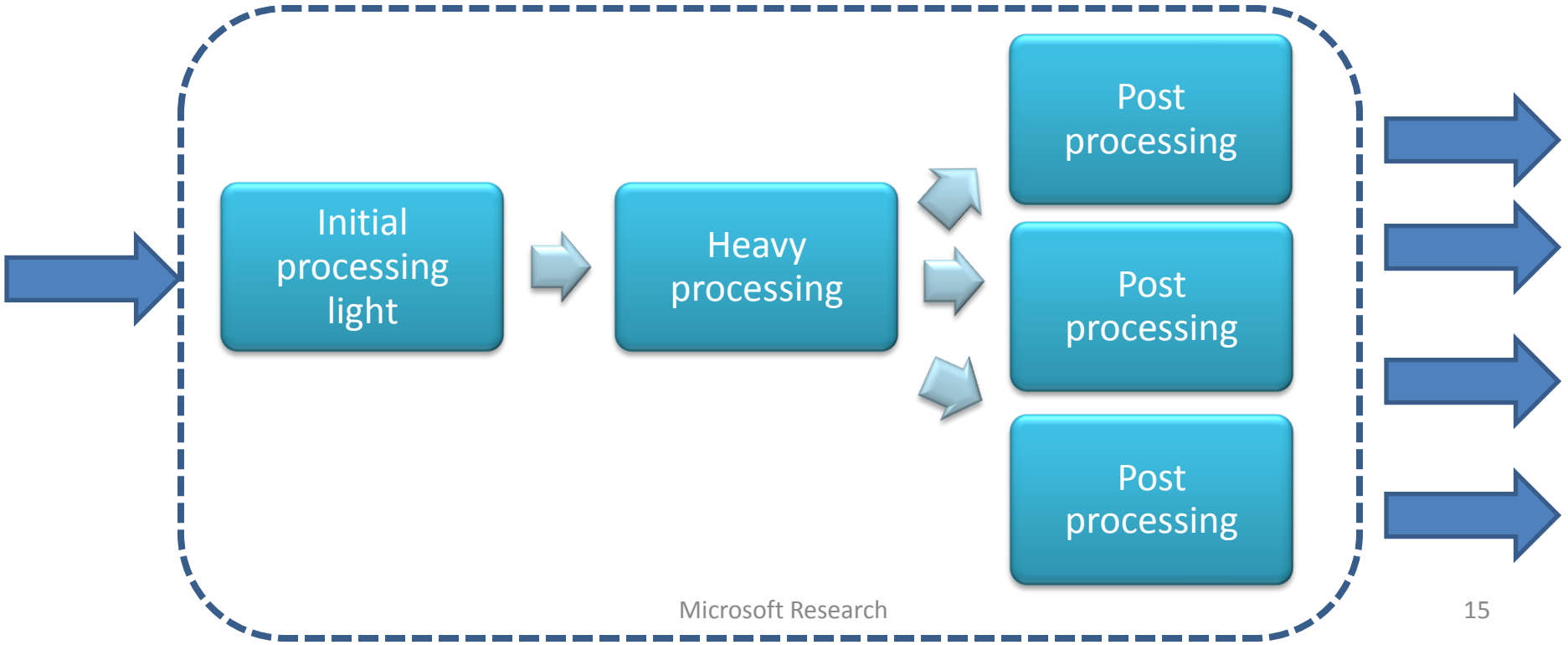
$$\pi(Z_{new} \mid \{Z_i\}_{i=1}^L, D, D_{new}) \propto \pi(Z_{new} \mid \{Z_i\}_{i=1}^L) \pi(D, D_{new} \mid Z_{new}, \{Z_i\}_{i=1}^L)$$

$$\pi(Z_{new} \mid \{Z_i\}_{i=1}^L) \propto \alpha 1(Z_{new} = new) + \sum_{i'=1}^{L} 1(Z_{new} = Z_{i'})$$

Closed form

Evaluation: The system

- Order of low 100's identical servers with balanced workload
- Metrics are mostly workload related + avg cpu + internal alarms → on the order of 100's per server



Evaluation: The data

Period between January-08 and May-08

Crises

# of instances	label
2	overloaded front-end
9	overloaded back-end
1	database configuration error
1	configuration error 1
1	configuration error 2
1	performance issue
1	middle-tier issue
1	request routing error
1	whole DC turned off and on
1	workload spike

Variety of problems:

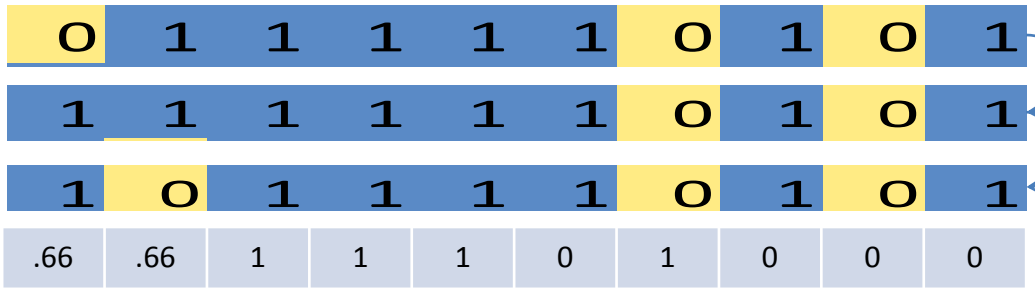
- Straight performance
- Human conf
- Human intervention
- Hardware errors

Crises needed ~ 6 metrics for identification

Experiment: Operational setting

- Start with an empty cluster
- On the event of a crisis
  - epoch by epoch compute a posterior and make decision on type
  - Update parameters
- Results
  - On original ordering of crises:
    83.4% accuracy and avg. of 1.8 periods for ID
  - Given 5 permutations of the ordering
    89% accuracy with avg. of 1.6 periods

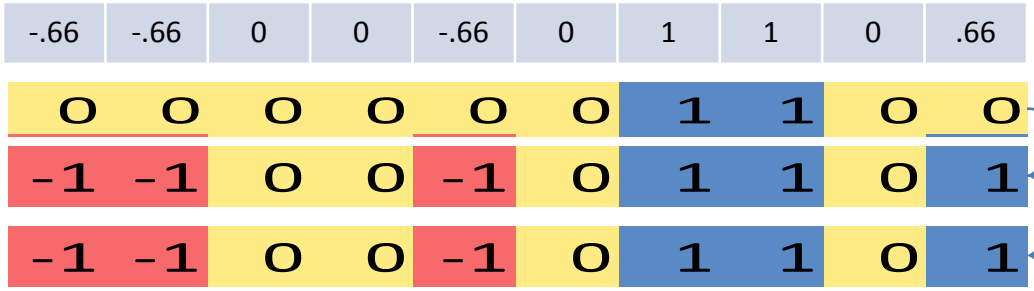
What's the worth of the stats model?



Model a crisis by Averaging each cell

Matching is done by establishing a distance such as L2

Need to take care of threshold T



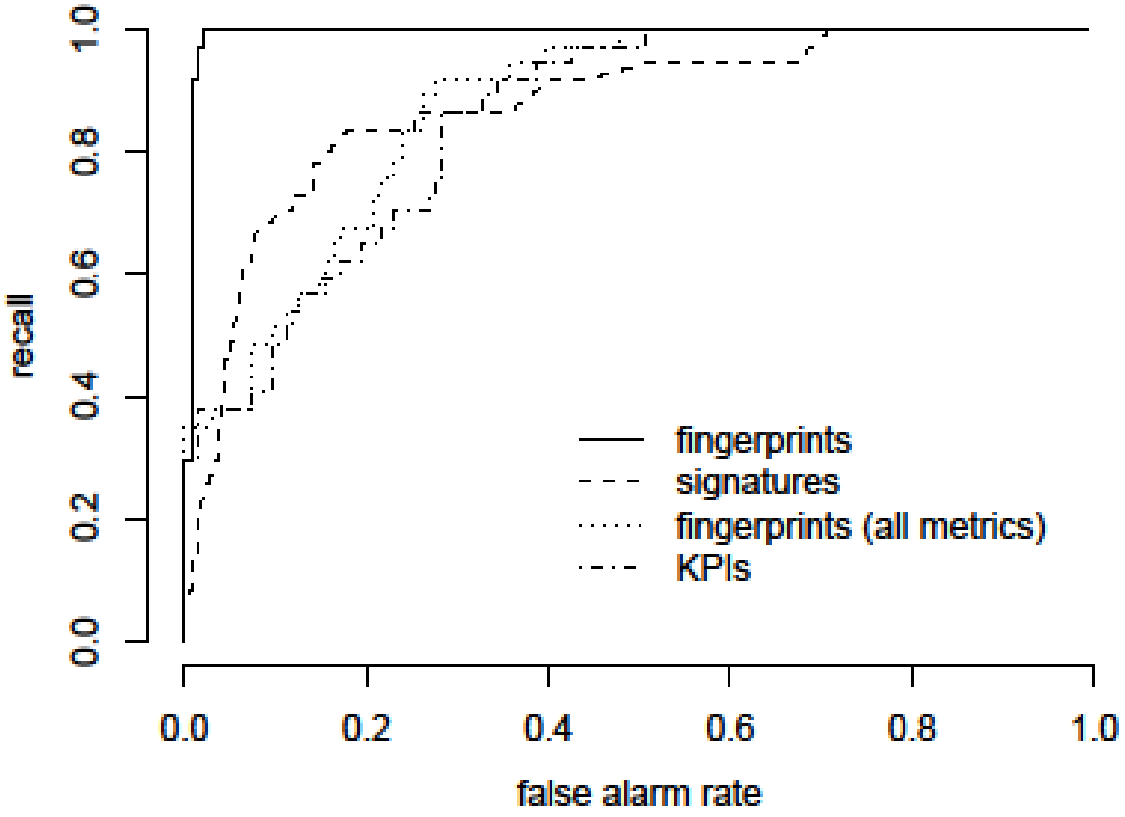
Comparable results to stats model ONLY
when start from 10 labeled crisis!

…And you don't have a probability for decision making!!

What about the other parts of the fingerprint???

1. Do you need really go through the trouble of finding the relevant metrics?
2. Can't you use the "detection" signals?
3. Can't you use the models that find the relevant metrics?

Discriminating between same and diff.



fingerprints 0.994
All metrics 0.873
Just KPIs 0.854
SOSP05 0.876

Conclusions from the experiments

- Fingerprints capture the essence of System state
- It is necessary to select relevant metrics
- It is not enough to look at the KPI
- A full statistical model achieves adaptability from very little initial knowledge

Summary

Described the model and inference machinery for online classification of performance crises and validated its performance on real data

- It provides full posterior for decision making
- Same model can compute posterior on the parameters of the crisis models

More information →
//research.microsoft.com/en-us/people/moises/

Dawn B. Woodard and Moises Goldszmidt, Model-Based Clustering for Online Crisis Identification in Distributed Computing

Peter Bodik, Moises Goldszmidt, Armando Fox, and Hans Andersen, Fingerprinting the datacenter: Automated classification of performance crises

Questions ??