Predicting Disk Replacement towards Reliable Data Centers

Mirela Botezatu , Ioana Giurgiu , Jasmina Bogojeska , Dorothea Wiesmann ,

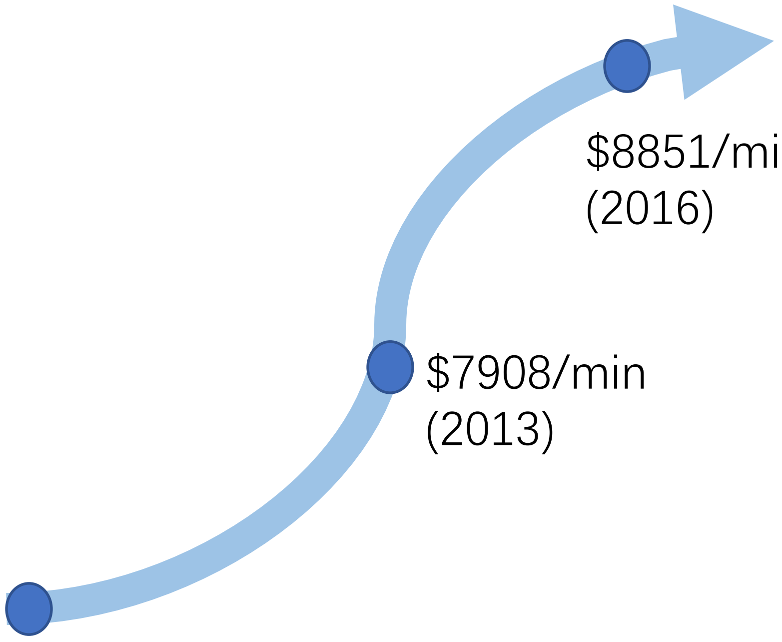
IBM Research

Outline

- Motivation
- Dataset characterization
- Prediction disk replacement
- Experimental results
- Conclusion

Datacenter downtime costs are growing steadily

- IT component failure is a significant contributor to datacenter downtimes.
- Disks are among the most frequently failing components in today’ s IT environments.



63 US data centers

Source: http://www.emerson.com/en-us/News/Pages/

Datacenter downtime costs are growing steadily

Can we mitigate this issue?

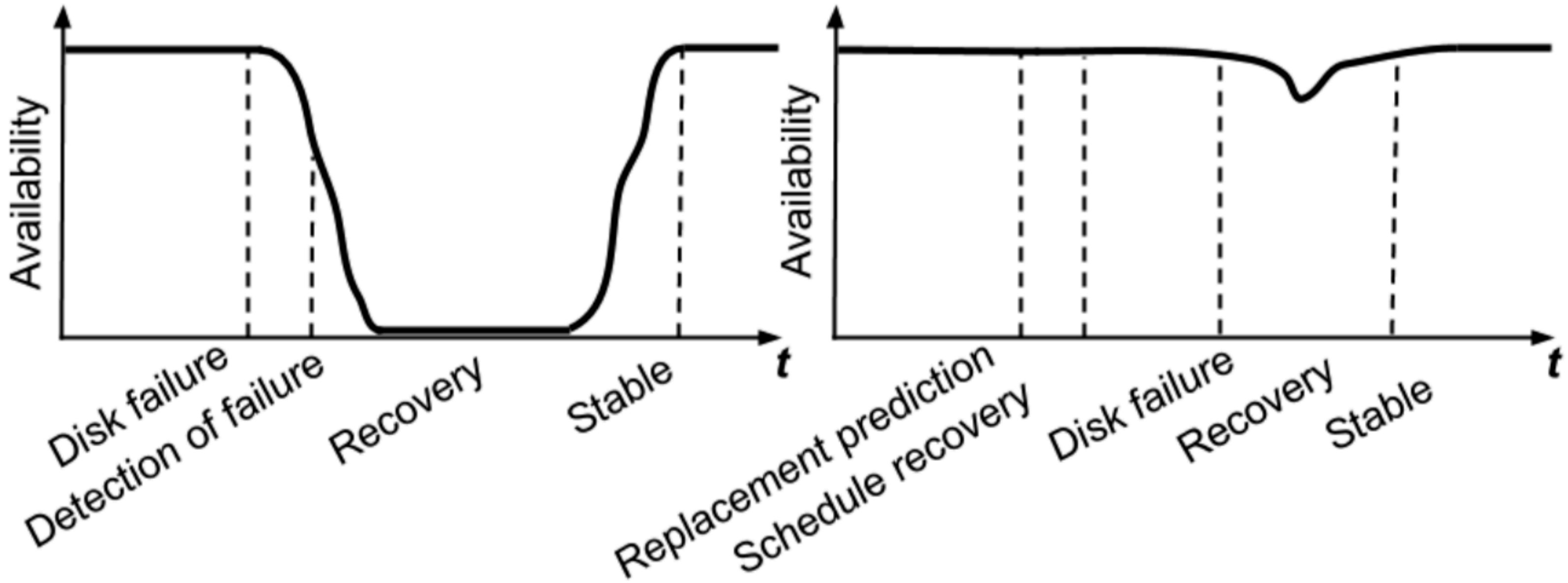


Figure 1: Availability: without proactive replacement (left) vs. with proactive replacement(right)

Objectives

- Given S.M.A.R.T monitoring data for disks (disk sensors' data), provide the subset of S.M.A.R.T attributes that are indicative of an impending disk replacement.

- Use these attributes to build a statistical model that automatically predicts disk replacement with high accuracy.

Legend

ID	193 0xC1	Attribute code in decimal and hexadecimal notations
Ideal	▲ High	Higher raw value is better
	Low ▼	Lower raw value is better
! (Critical)	⚠	Denotes a *Critical* attribute. Specific values may predict drive failure

~80

ID	Attribute name	Ideal	!	Description
01 0x01	**Read Error Rate**	Low ▼		(Vendor specific raw value.) Stores data related to the rate of hardware read errors that occurred when reading data from a disk surface. The raw value has different structure for different vendors and is often not meaningful as a decimal number.
02 0x02	**Throughput Performance**	▲ High		Overall (general) throughput performance of a hard disk drive. If the value of this attribute is decreasing there is a high probability that there is a problem with the disk.
03 0x03	**Spin-Up Time**	Low ▼		Average time of spindle spin up (from zero RPM to fully operational [milliseconds]).
04 0x04	**Start/Stop Count**			A tally of spindle start/stop cycles. The spindle turns on, and hence the count is increased, both when the hard disk is turned on after having before been turned entirely off (disconnected from power source) and when the hard disk returns from having previously been put to sleep mode.[21]
05 0x05	**Reallocated Sectors Count**	Low ▼	⚠ [22][23][24]	Count of reallocated sectors. The raw value represents a count of the bad sectors that have been found and remapped.[25] Thus, the higher the attribute value, the more sectors the drive has had to reallocate. This value is primarily used as a metric of the life expectancy of the drive; a drive which has had any reallocations at all is significantly more likely to fail in the immediate months.[22][26]
06 0x06	**Read Channel Margin**			Margin of a channel while reading data. The function of this attribute is not specified.
07 0x07	**Seek Error Rate**	Varies		(Vendor specific raw value.) Rate of seek errors of the magnetic heads. If there is a partial failure in the mechanical positioning system, then seek errors will arise. Such a failure may be due to numerous factors, such as damage to a servo, or thermal widening of the hard disk. The raw value has different structure for different vendors and is often not meaningful as a decimal number.
08 0x08	**Seek Time Performance**	▲ High		Average performance of seek operations of the magnetic heads. If this attribute is decreasing, it is a sign of problems in the mechanical subsystem.
09 0x09	**Power-On Hours**			Count of hours in power-on state. The raw value of this attribute shows total count of hours (or minutes, or seconds, depending on manufacturer) in power-on state.[27] "By default, the total expected lifetime of a hard disk in perfect condition is defined as 5 years (running every day and night on all days). This is equal to 1825 days in 24/7 mode or 43800 hours."[28] On some pre-2005 drives, this raw value may advance erratically and/or "wrap around" (reset to zero periodically).[29]
10 0x0A	**Spin Retry Count**	Low ▼	⚠ [30]	Count of retry of spin start attempts. This attribute stores a total count of the spin start attempts to reach the fully operational speed (under the condition that the first attempt was unsuccessful). An increase of this attribute value is a sign of problems in the hard disk mechanical subsystem.
11 0x0B	**Recalibration Retries** or **Calibration Retry Count**	▼ Low		This attribute indicates the count that recalibration was requested (under the condition that the first attempt was unsuccessful). An increase of this attribute value is a sign of problems in the hard disk mechanical subsystem.
12 0x0C	**Power Cycle Count**			This attribute indicates the count of full hard disk power on/off cycles.
13 0x0D	**Soft Read Error Rate**	Low ▼		Uncorrected read errors reported to the operating system.
22 0x16	**Current Helium Level**	▲ High		Specific to He8 drives from HGST. This value measures the helium inside of the drive specific to this manufacturer. It is a pre-fail attribute that trips once the drive detects that the internal environment is out of specification.[31]
170 0xAA	**Available Reserved Space**			See attribute E8.[32]
171 0xAB	**SSD Program Fail Count**			(Kingston) The total number of flash program operation failures since the drive was deployed.[33] Identical to attribute 181.
172 0xAC	**SSD Erase Fail Count**			(Kingston) Counts the number of flash erase failures. This attribute returns the total number of Flash erase operation failures since the drive was deployed. This attribute is identical to attribute 182.
173 0xAD	**SSD Wear Leveling Count**			Counts the maximum worst erase count on any block.
174 0xAE	**Unexpected power loss count**			Also known as "Power-off Retract Count" per conventional HDD terminology. Raw value reports the number of unclean shutdowns, cumulative over the life of an SSD, where an "unclean shutdown" is the removal of power without STANDBY IMMEDIATE as the last command (regardless of PLI activity using capacitor power). Normalized value is always 100.[34]
175 0xAF	**Power Loss Protection Failure**			Last test result as microseconds to discharge cap, saturated at its maximum value. Also logs minutes since last test and lifetime number of tests. Raw value contains the following data: • Bytes 0-1: Last test result as microseconds to discharge cap, saturates at max value. Test result expected in range 25 <= result <= 5000000, lower indicates specific error code. • Bytes 2-3: Minutes since last test, saturates at max value. • Bytes 4-5: Lifetime number of tests, not incremented on power cycle, saturates at max value. Normalized value is set to one on test failure or 11 if the capacitor has been tested in an excessive temperature condition, otherwise 100.[34]
176 0xB0	**Erase Fail Count**			S.M.A.R.T. parameter indicates a number of flash erase command failures.[35]
177 0xB1	**Wear Range Delta**			Delta between most-worn and least-worn Flash blocks. It describes how good/bad the wearleveling of the SSD works on a more technical way.
179 0xB3	**Used Reserved Block Count Total**			"Pre-Fail" attribute used at least in Samsung devices.

Data

- Monitoring data (S.M.A.R.T indicators) from a large population of disks (>30000) collected over 17 months.
- Labels indicating whether a disk failed or not.

When is a disk labeled as failed?

- The disk stopped working
- The disk is non-responsive to commands
- The RAID system reports that the drive cannot be written or read, or it shows evidence of failing soon

Goal: Predictive Replacement Component

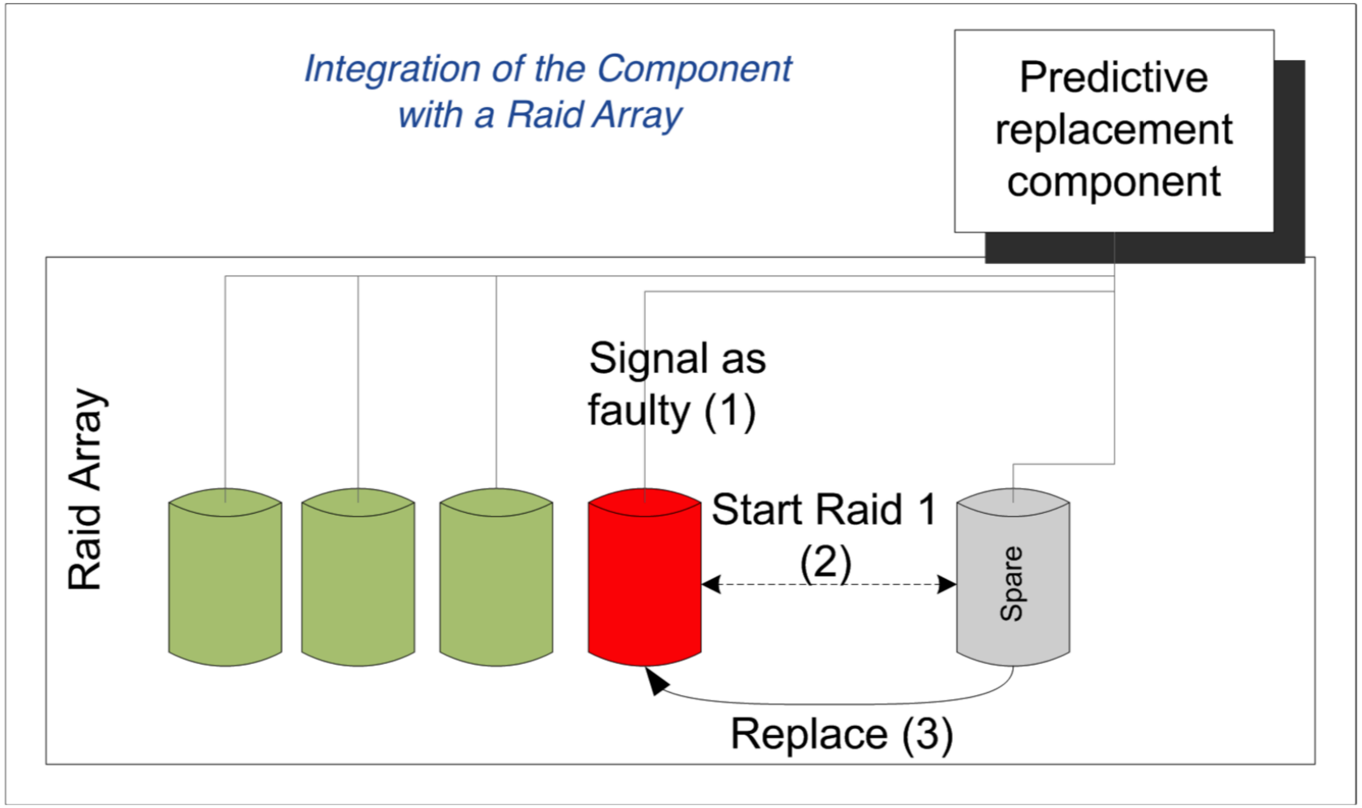


Figure 7: Integration of the predictive replacement component with storage arrays

Prediction pipeline

Changepoint detection → Time series compression → Informed downsampling → Classification

Changepoint detection: Time series changepoint detection. Identify subset of SMART parameters indicative of disk failure.

Time series compression: Exponential smoothing for compact time series representation.

Informed downsampling: Informed down-sampling of the dataset via k-means clustering to address the high class imbalance.

Classification: Build a regularized greedy forests model to distinguish the healthy disks from those likely to fail.

Changepoint detection

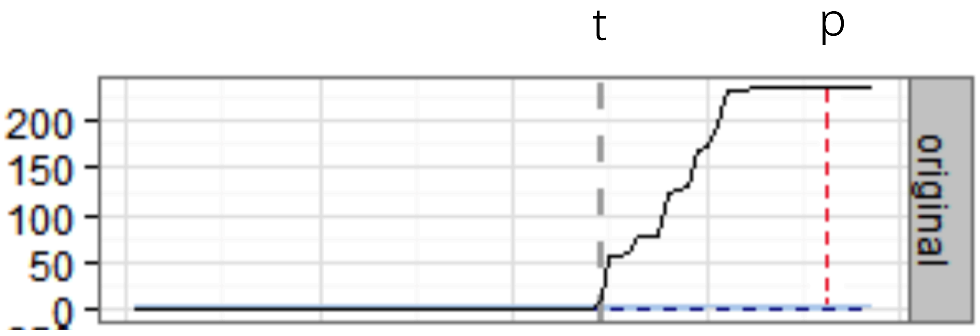
Goal: Reveal the most informative predictors with respect to the disks to the domain experts.

Assumption: When a SMART attribute is informative of disk replacement, we expect a significant shift in its values at some time point before the disk failure.

Approach: Let $S_i = (s_1, s_2, \dots, s_p)$ be the time series for a target SMART attribute.

- If $\exists$ a timestamp $t < p$ when a significant change in the values of the attribute S_i occurs (e.g., the values start increasing), then we consider S_i a potential attribute relevant for the disk replacement

Changepoint detection



Steps towards changepoint detection:

1. Choose a time t that has the largest change:

We take $t = argmax_t ML(\tau)$ where $ML(\tau) = \log\left(p\left(s_{1:t}\middle|\widehat{\theta_1}\right)\right) + \log(p\left(s_{t+1:p}\middle|\widehat{\theta_2}\right))$ provided that $ML(\tau)$ is significantly larger than $\log(p\left(s_{1:p}\middle|\hat{\theta}\right))$

2. We assess whether the change is permanent:

a. We let $\Gamma_t = (s_t, \dots, s_p)$ be the time series observed after point t. We generate $\Psi = \left(\widetilde{s_t}, \dots, \widetilde{s_p}\right)$ that has no changepoint at time t, i.e., we compute the posterior distribution of Ψ given the values in the pre-change period $(s_1, \dots, s_t)$ the values of a control (healthy) time series $x_{1:p}$

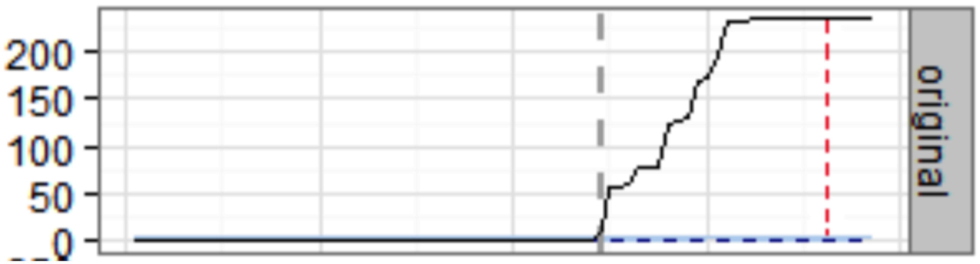
t

Changepoint detection

- b. Finally, a SMART attribute is indicative of a disk replacement if the **probability distributions** of the **actual time series** (measured after the detected change point) and the **synthetic** one generated based on the values of a healthy disk are **significantly different**.

Formally, if Γ and Ψ are drawn from probability distributions P and Q, we check:

$$\begin{cases} H_0 : P = Q \\ H_1 : P \neq Q \end{cases}$$



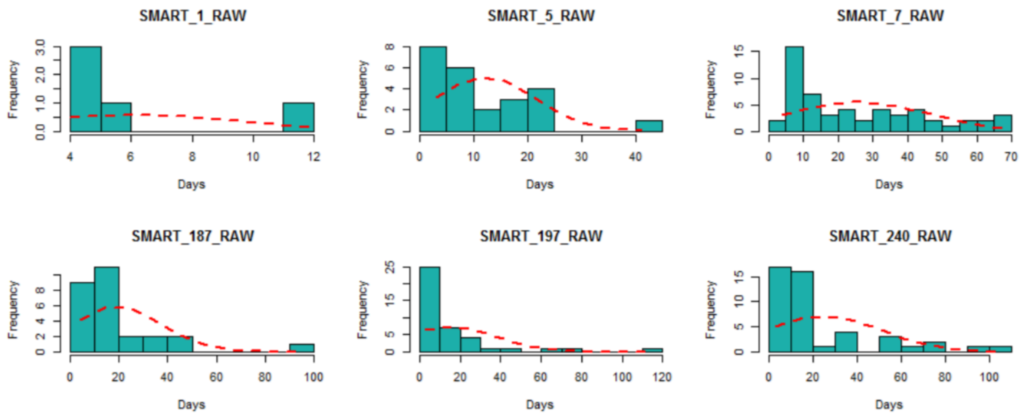


Figure 4: Distribution of the number of days before replacement when the changepoint was observed.

Results-Subset of relevant SMART indicators

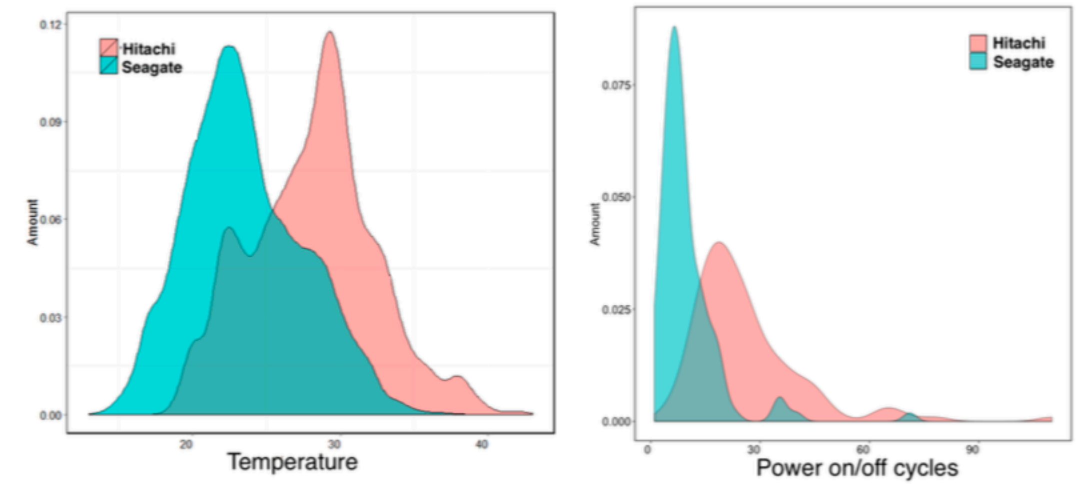


Figure 3: Distribution of the temperature and of the power on off cycles across the replaced disks for Hitachi and Seagate.

	SgtA		HitA	
	Ratio	Inp.	Ratio	Inp.
SMART_1_norm	23%	✓	28%	✓
SMART_1_raw	2%	✓	15%	✓
SMART_3_norm	–	×	13%	✓
SMART_3_raw	–	×	15%	✓
SMART_5_norm	2%	✓	22%	✓
SMART_5_raw	19%	✓	31%	✓
SMART_7_norm	14%	✓	–	×
SMART_7_raw	26%	✓	–	×
SMART_183_norm	0.5%	×	–	×
SMART_183_raw	0.5%	×	–	×
SMART_184_norm	1%	✓	–	×
SMART_184_raw	1%	✓	–	×
SMART_187_norm	21%	✓	–	×
SMART_187_raw	21%	✓	–	×
SMART_188_norm	0%	×	–	×
SMART_188_raw	10%	✓	–	×
SMART_189_norm	1%	✓	–	×
SMART_189_raw	1%	✓	–	×
SMART_190_norm	2%	✓	–	×
SMART_190_raw	2%	✓	–	×
SMART_193_norm	10%	✓	–	×
SMART_193_raw	63%	✓	–	×
SMART_194_norm	2%	✓	31%	✓
SMART_194_raw	2%	✓	2%	✓
SMART_196_norm	–	×	20%	✓
SMART_196_raw	–	×	26%	✓
SMART_197_norm	5%	✓	4%	✓
SMART_197_raw	27%	✓	22%	✓
SMART_198_norm	6%	✓	–	×
SMART_198_raw	27%	✓	–	×
SMART_199_norm	0%	×	–	×
SMART_199_raw	0.5%	×	–	×
SMART_240_norm	0.5%	×	–	×
SMART_240_raw	21%	✓	–	×
SMART_241_norm	0%	–	–	×
SMART_241_raw	15%	✓	–	×
SMART_242_norm	0%	×	–	×
SMART_242_raw	19%	✓	–	×

Table 2: SMART correlation frequencies for SgtA and HitA. A ✓indicates the predictor is included in the classification task.

Compact time series representation

Goal: Provide a compact, highly informative representation of the time series of each indicator.

Observations :

- The single day record is not stable due to the recovery mechanisms embedded in the disk
- For timely predictions, one should not consider as observations for the failed class just the entries from the last day before the disk fails

Approach: We use a window to split the raw data set into segments. We aggregate segments to a single value using exponential smoothing over a specific time window.

Exponential smoothing :$S_t = aY_t + (1-a)S_{t-1}$. For a window length of size k, S_t becomes the weighted average of a k past observations up to Y_{t-k}

Informed downsampling

Observations : Classification algorithms are typically optimized to maximize the accuracy, therefore when trained on imbalanced datasets they exhibit **poor predictive performance**.

Goal: Extract a **subset** of the data for the **dense class** –in our case the **healthy disks**

Approach:

- Cluster the observations from the healthy disk set into k clusters using the K- means clustering algorithm.
    - Choosing k close to the number of samples available for the faulty class samples.
- For each cluster, select the data points closest to the respective cluster centroid as representatives for the healthy disk class.
- We generate a balanced training set: union of the observations for the faulty class and the reduced subset of data points for the healthy class

Disk classification: healthy vs. likely to fail

- Goal: Learn $h: X \to \{0,1\}$ that minimizes the loss $l(h(x); y)$ that quantifies the prediction quality

- Approach: Regularized Greedy Forests (RGF), a variant of Gradient Boosted Decision Trees in which structure search and the optimization step are decoupled:

- RFG introduce an explicit regularization term that takes advantage of individual tree structures. $\hat{h} = argmin_{h \in H}[\ell(h(\mathbf{x}); y) + R(h)]$
- Performs a greedy search on forest structure changing operations by repeatedly evaluating the maximum loss reduction of all the possible structure changes;

Transfer learning

- Observations: Different models of a single disk manufacturer have similar SMART reporting but different distributions of the values reported for the SMART attributes.
- Goal: Transfer the learnings from a specific disk model to a new disk model of the same manufacturer.

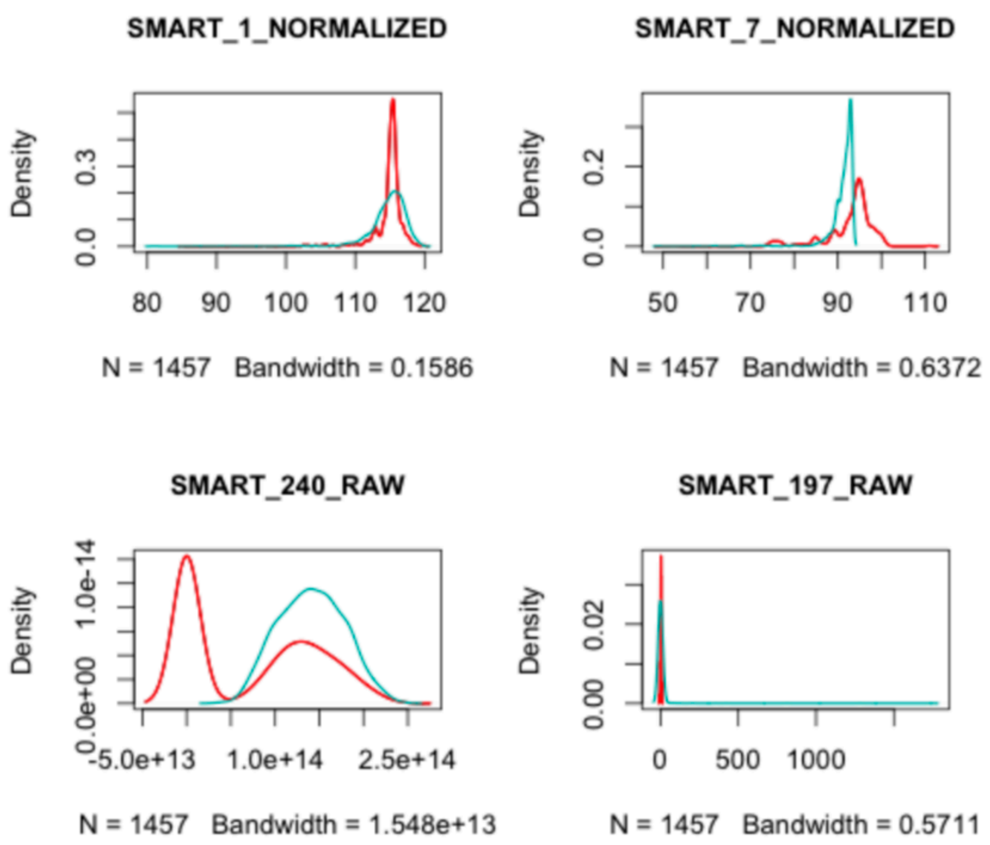


Figure 5: Covariate shift for the two Seagate models

Transfer learning

- Approach: Use the unlabeled data for the target (new) disk model to conduct a sample selection de-biasing

- The idea behind the algorithm is to train a classifier that can rank the observations linked to a source disk model based on their similarity to observations pertaining to the target disk model.

- This enables to sample the observations from the source disk model (which are already labeled) that are more representative for learning the class labels for the target disk model, i.e. that matches the distribution of the source disk model to the target disk model.

Algorithm 3 Transfer learning for different models

Input: $D_{DM_1} = \{x_i, y_i\}_i^n$, the labeled data collected from disk model 1, and $D_{DM_2} = \{x'_i, y'_i\}_i^m$ the unlabeled data from disk model 2.

1. Let $D_{DM_1} = \{x_i, y_i\}_i^n$ be the labeled data collected from disk model 1, and $D_{DM_2} = \{x'_i, y'_i\}_i^m$ be the unlabeled data from disk model 2.
2. Let $D_{aug} = \{x_i, \text{“}DM_1\text{”}\}_i^n \cup \{x'_i, \text{“}DM_2\text{”}\}_i^m$
3. Use D_{aug} to learn a function $f : X \rightarrow [0, 1]$, such that $f(x)$ represents the probability of a disk being of type “DM_1” or “DM_2”.
4. Sample a subset D_{sub} from D_{DM_1} according to f .
5. Use D_{sub} to learn a function $g : X \rightarrow [0, 1]$ (call the procedure in Algorithm 2) such that $g(x)$ represents the probability of a disk of type DM_2 needing replacement.

Output: Predictive model for disk replacement for disk model 2.

Results – Prediction accuracy

		RGF		GBDT		RF		SVM		LR		DT	
		SgtA	**HitA**	**SgtA**	**HitA**	**SgtA**	**HitA**	**SgtA**	**HitA**	**SgtA**	**HitA**	**SgtA**	**HitA**
Replaced	P	0.98	0.84	0.97	0.82	0.93	0.82	0.93	0.72	0.73	0.72	0.89	0.74
	R	0.98	0.79	0.96	0.78	0.94	0.76	0.95	0.65	0.81	0.59	0.87	0.61
	F	**0.98**	**0.81**	**0.96**	**0.80**	**0.94**	**0.79**	**0.94**	**0.68**	**0.77**	**0.65**	**0.88**	**0.67**
	Sd	**0.01**	**0.02**	0.01	0.04	0.05	0.08	0.02	0.05	0.07	0.1	0.04	0.03
Healthy	P	0.99	0.93	0.98	0.92	0.97	0.92	0.97	0.87	0.89	0.85	0.94	0.86
	R	0.98	0.95	0.98	0.94	0.96	0.93	0.96	0.90	0.85	0.90	0.95	0.91
	F	**0.98**	**0.94**	0.98	0.93	0.97	0.92	0.96	0.88	0.87	0.87	0.94	0.88
	Sd	**0.01**	**0.02**	0.02	0.03	0.04	0.05	0.02	0.04	0.08	0.05	0.02	0.02

Table 3: Precision, Recall, F-score, Deviation of different classifiers - median on 100 runs , each of which using randomly-drawn training and test data points

In case of the replaced disks, Seagate has 4x more data points and 2x more non-null SMART indicators than Hitachi, which has a smaller number of drives in the dataset and 60% less predictors.

For the healthy class, Hitachi achieves better performance (as compared to the faulty ones
) because of the lower variability in the values of the SMART parameters recorded for healthy disks.

Results – Comparison with emulated human rules

We train a decision tree on the subset of SMART indicators that is commonly considered when assessing disk health.

		DT on the reduced subset	
		SgtA	**HitA**
Replaced	Precision	0.95	0.66
	Recall	**0.53**	**0.44**
	F-score	0.68	0.51
	Sd	0.06	0.15
Healthy	Precision	0.70	0.84
	Recall	0.98	0.96
	F-score	0.81	0.92
	Sd	0.02	**0.12**

Table 5: Simple decision tree with (insufficient but commonly used) subset of SMART indicators

If one were to do proactive replacement using only this small subset of indicators, the number of disks one could correctly identify drops by almost 50%

Results – Transfer learning

trained on SgtA

trained on HitA

		SgtB		HitB	
		Base	Tr. Learn.	Base	Tr. Learn.
Replaced	P	0.65	**0.90**	0.53	**0.76**
	R	0.52	**0.82**	0.84	**0.78**
	F	0.58	**0.86**	0.65	**0.77**
Healthy	P	0.89	0.96	0.92	0.83
	R	0.93	0.98	0.73	0.82
	F	0.91	0.97	0.81	0.83

Table 4: Precision, recall and F-score to illustrate the importance of transfer learning

Results – High confidence rules from a decision tree model

Line	Model	Rule	Outcome	Confidence
1	Seagate	If $SMART\_197\_raw < 2$ and $SMART\_188\_raw > 0$ and $SMART\_1\_normalized \in [0, 117)$	Healthy	100%
2	Seagate	If $SMART\_197\_raw > 2$	Replace	100%
3	Seagate	If $SMART\_197\_raw < 2$ and $SMART\_188\_raw > 0$ and $SMART\_1\_normalized > 117$	Replace	80%
4	Seagate	If $SMART\_197\_raw < 2$ and $SMART\_188\_raw = 0$ and $SMART\_187\_normalized < 100$ and $SMART\_240\_raw < 14780$ billion	Replace	97%
5	Hitachi	If $SMART\_197\_raw > 1$ and $SMART\_3\_raw > 626$	Replace	100%
6	Hitachi	If $SMART\_197\_raw > 5$ and $SMART\_3\_raw < 626$ and $SMART\_5\_raw > 17$	Replace	92%
7	Hitachi	If $SMART\_197\_raw > 1$ and $SMART\_3\_raw < 626$ and $SMART\_5\_raw < 17$	Replace	100%
8	Hitachi	If $SMART\_197\_raw < 1$ and $SMART\_5\_raw < 7200$ and $SMART\_3\_raw > 629$ and $SMART\_1\_raw \in [0, 109$	Healthy	97%

Table 6: Examples of rules extracted from a decision tree model trained on the Seagate and Hitachi datasets obtained with Algorithm 1.

First, **the primarily important SMART indicators are somewhat different**. The pending sector count (Count of "unstable" sectors , SMART 197 raw) and the read error rate (SMART 1 normalized) seem to be model and even manufacturer agnostic, while the command timeout (The count of aborted operations due to HDD timeout, SMART 188), the average spin up time and the reallocated sectors count are disk model-specific.

Second, we note **a very large difference in the number of read errors (SMART_1_RAW)** that determine a faulty disk state. For Seagate, this threshold is in hundreds of millions, while for Hitachi they are 6 orders of magnitude lower. We attribute this gap to the fact that this indicator is vendor specific, and therefore a comparison across manufacturers is not feasible.

Early vs. late prediction accuracy

- We evaluate how many of the replaced disks our model correctly captures based on snapshots of the SMART indicators taken 1, 3, 10 and 30 days prior to the actual replacement.

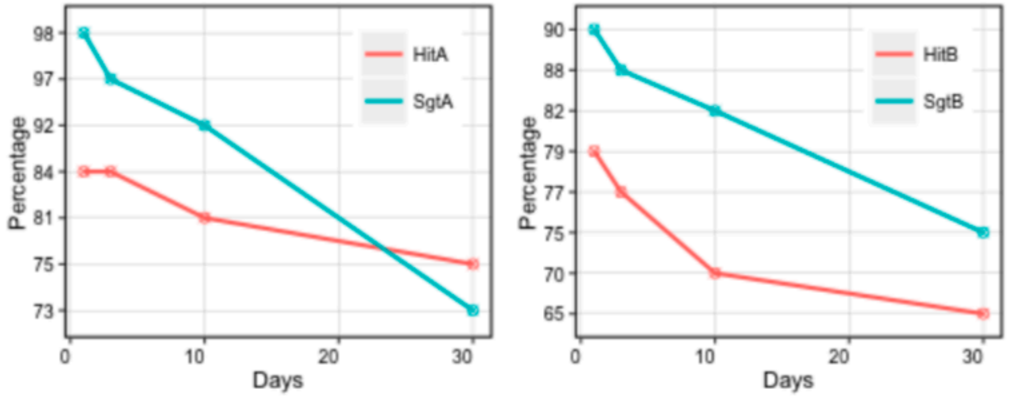


Figure 6: Percentage of disks correctly predicted as replaced on snapshots taken 1,3,10 and 30 days before the actual replacement event.

For both Seagate and Hitachi, an administrator can identify 73 to 75% of the disks to replace a month in advance, which provides her/him with the possibility of planning the replacement in advance, while still using the drives for another 25-30 days.

Conclusion

- The model provides an automatic tool for the disk replacement problem that enables the administrators to identify faulty disks in due time.
- It mitigates the reliability issues of storage service providers by allowing administrators to backup the data and plan the actual replacement in advance.
- Such models are sensitive to the number of SMART attributes they use. This explains the 17% gap in accuracy for the two disk manufacturer.
- Transfer learning can be applied across different models of the same disk manufacturer
- The pipeline can be easily applied to any disk model or manufacturer as long as SMART data is collected.