
CERN

Disk failures in the EOS setup at CERN: A first systematic look at 1 year of collected data

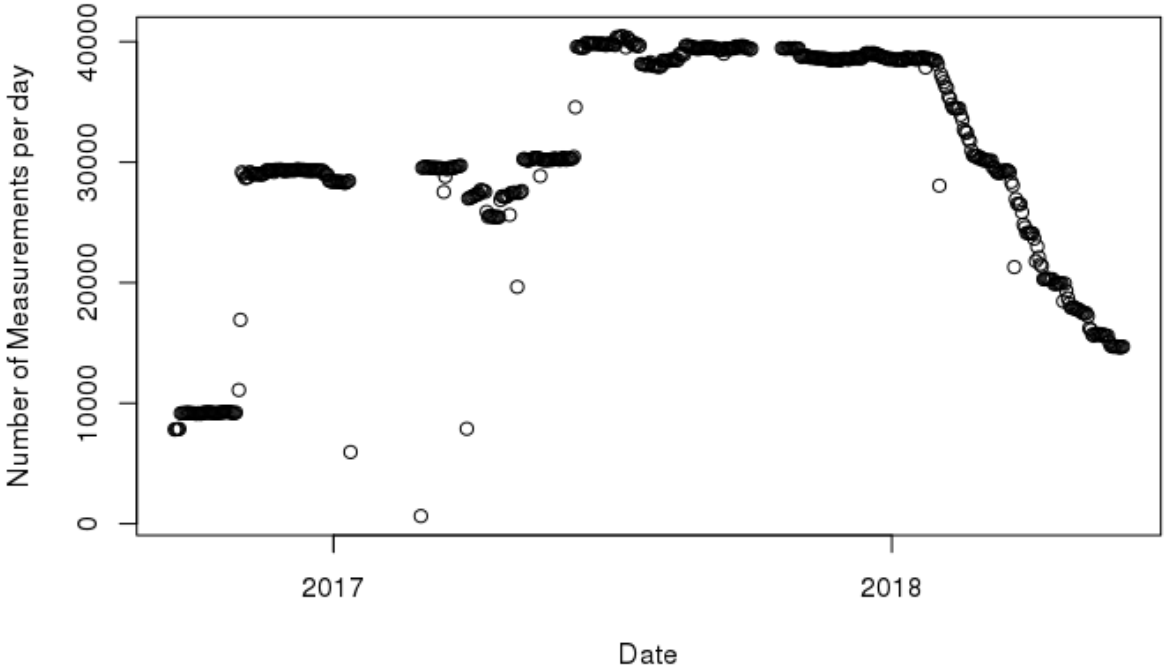
State of the art

- There are some studies in the relation of S.M.A.R.T. metrics and failure in hard drives.
  - Google study: 60 days following the first uncorrectable error on a drive (S.M.A.R.T. 198) the drive was, on average, 39 times more likely to fail. 36% of failed drives did so without recording any S.M.A.R.T. error at all.
  - Backblaze collected data, that was analyzed by IBM. Annualized failure rate of ~1.65%.
- Most of the S.M.A.R.T. metrics are very device dependent.
- Studies reach no clear or widely applicable conclusions.

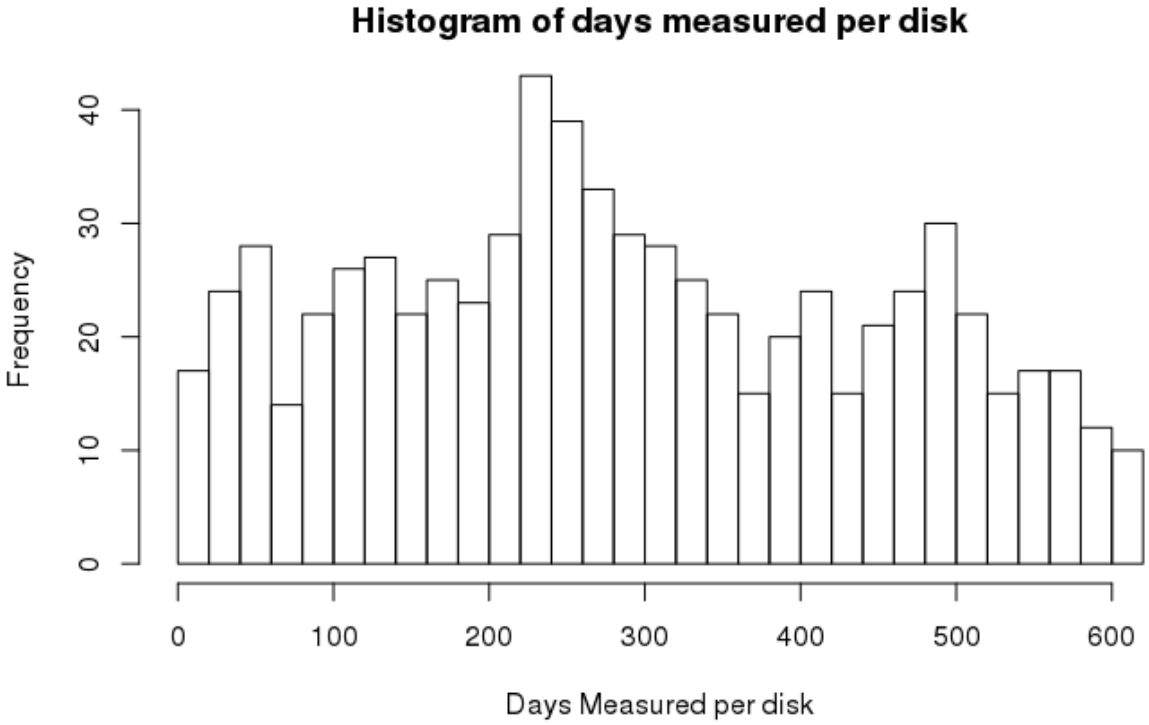
Input data

- Number of days that we have a measurement of: 551
- Number of days since we started recording SMART metrics: 620
- Min / Max number of disks measured in a day: 635 / 40563
- Average disks measured per day: 31770
- Total number of unique disks reporting SMART info: 45874 unique serials
- Only have media information on Vendor on 35.37% of these measurements. This will be solved via an in-house development of a probe, which will provide us organized and easily available data from now on (Soon to be implemented)

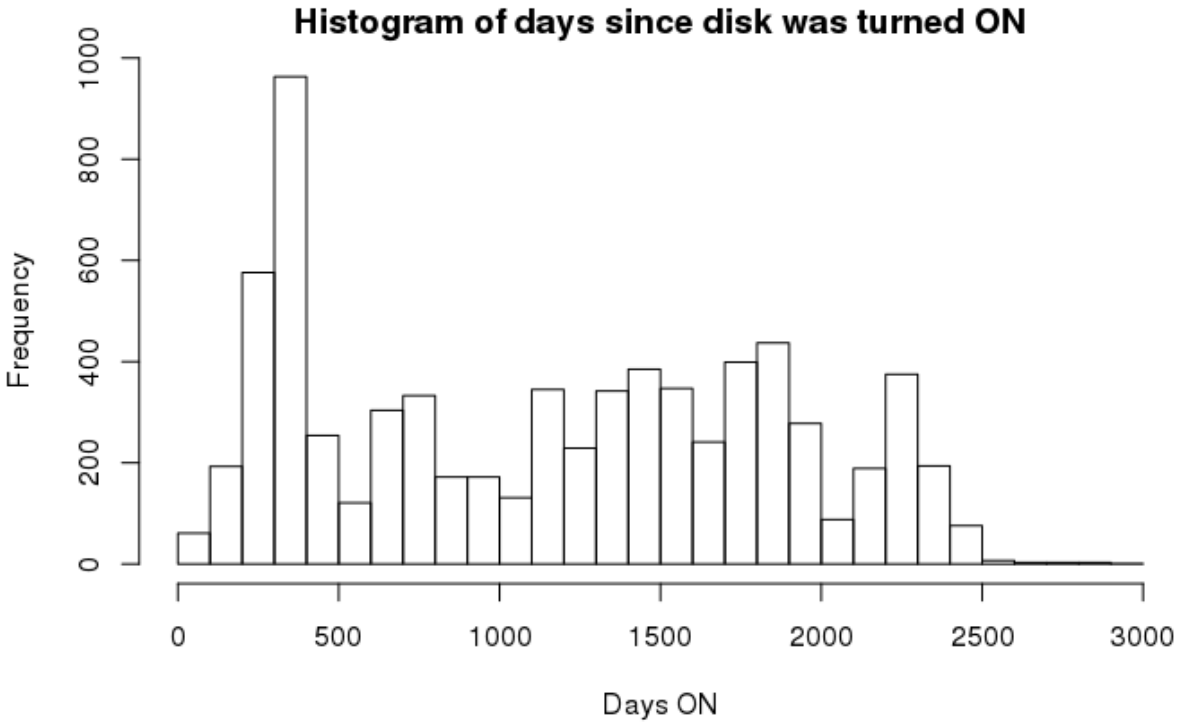
Number of measurements per day



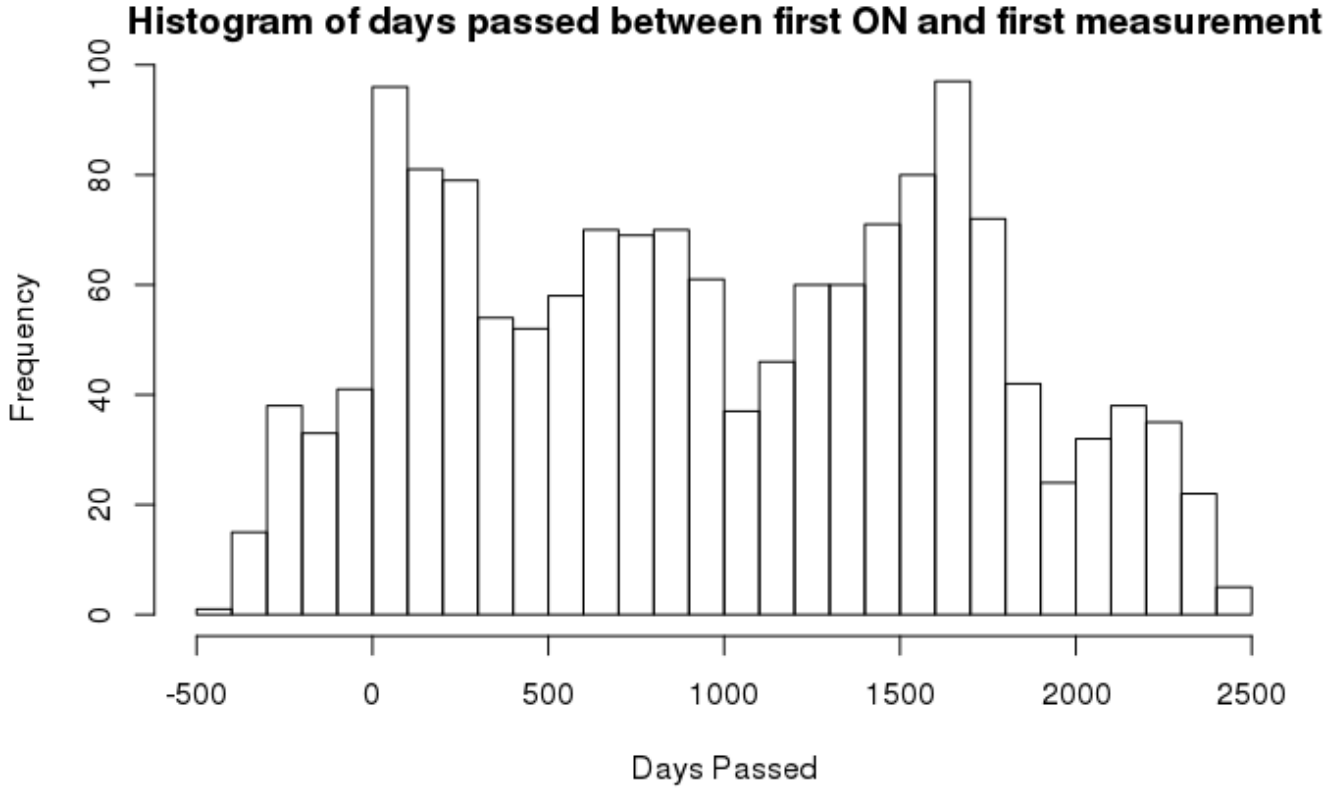
Number of days measured per disk



Number of days ON



Difference On to Measurement



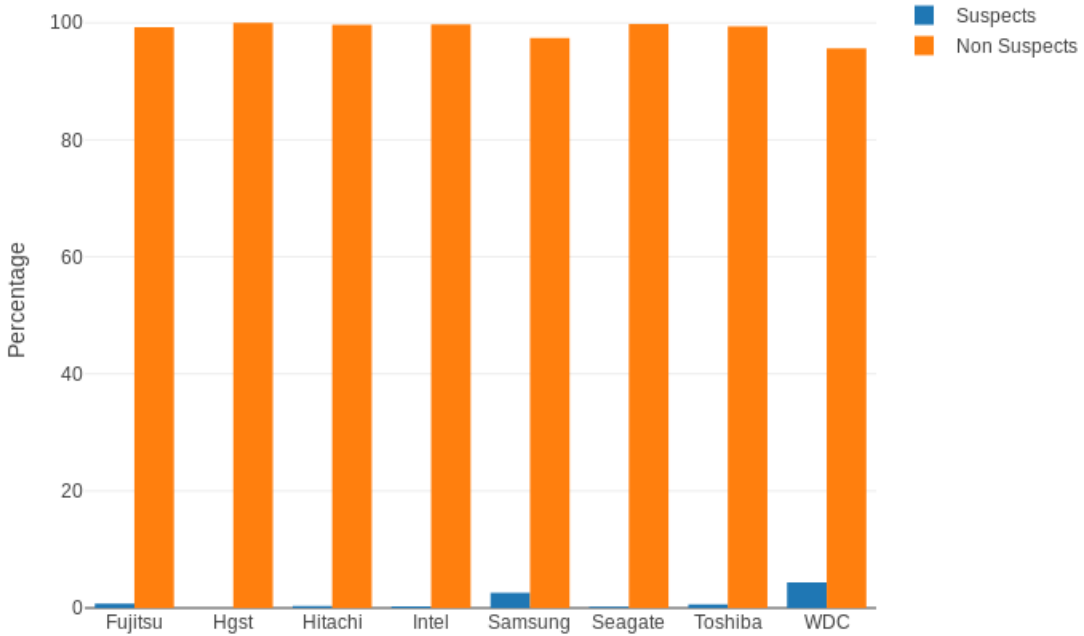
Challenges

- This was not a designed measurement. The data was picked up inside the group by operations. The sensor did not collect detailed metrics, as they were not considered reliable.
- The prediction can only be done by model, as some of the metrics' meaning change heavily between devices (big percentage of metadata is missing). How much data do we need?
- For small sites, or sites with very heterogeneous disk types, it is a very complicated analysis to do.
- Different data sources, and different data structures.
- No information on a disk being an SSD or an HDD.

Suspect and failure classification

- We consider a disk as failed if it disappears from the dataset while the box that contains it is still present, as well as the rest of the disks of the same box. We thus generate a “failed” bit on the last day available of these disks.
- 0.88% of the disks whose information we collected, appear as failed.
- We also generate a “sub” bit, whenever a disk disappears from the dataset.
- 67.97% of the disks have been substituted or have stopped being recorded.
- We use the “failed” classification in order to divide our dataset in Suspects and Non Suspects, and analyze them in order to see if this assumption is correct.

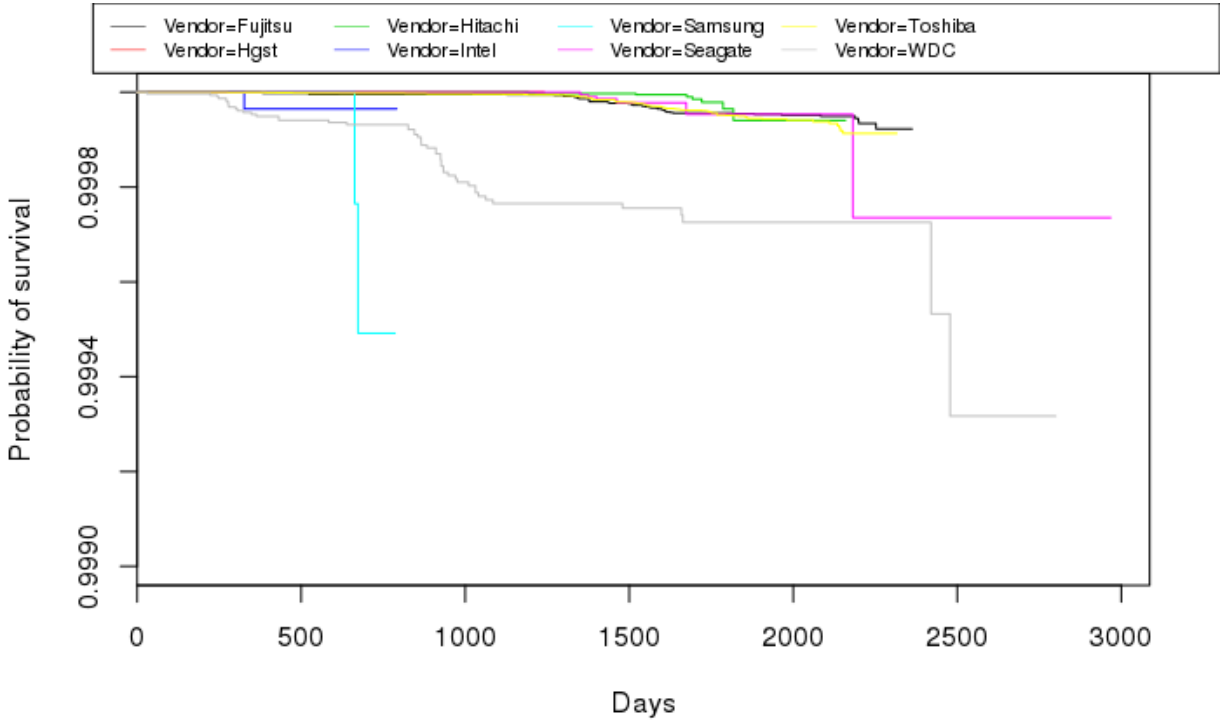
Suspect and failure classification (II)



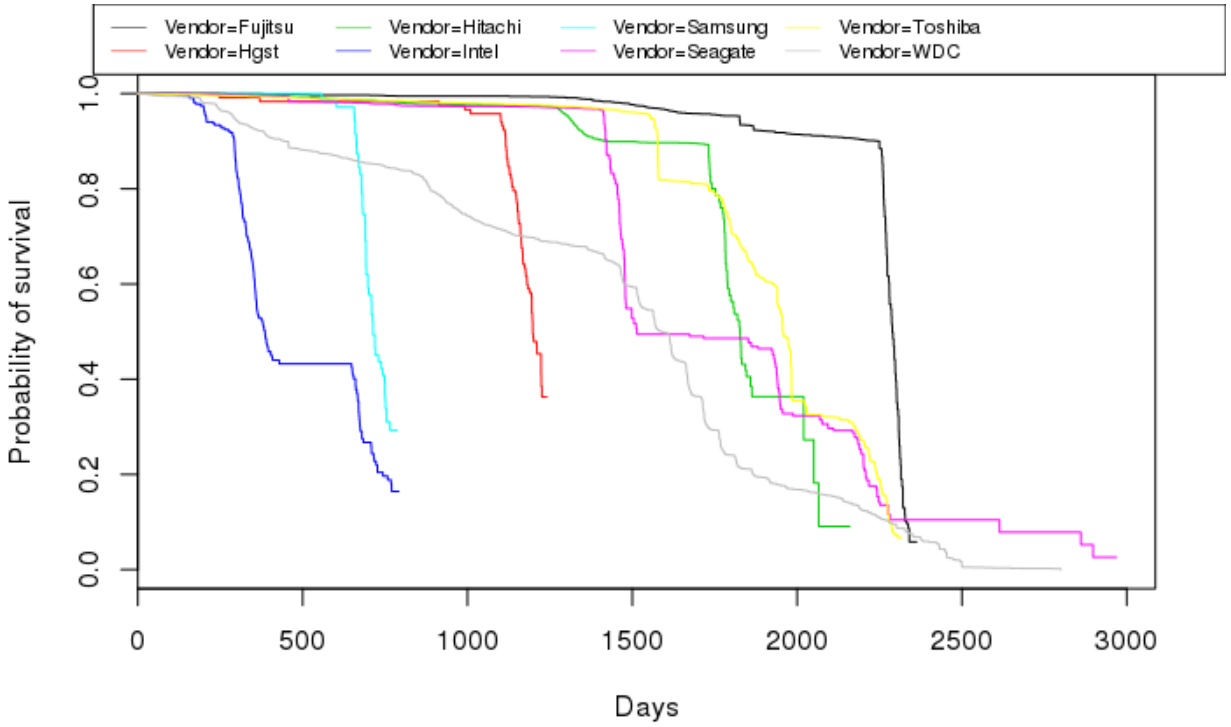
Kaplan-Meier survival curves

- Started with an analyisis based on Kaplan-Meier survival curves, mostly used on clinical studies in hospitals.
- We obtain the information on the survival rate per vendor.
- We generate a survival curve based on failure, and one based on substitution, or dissapearance rate.

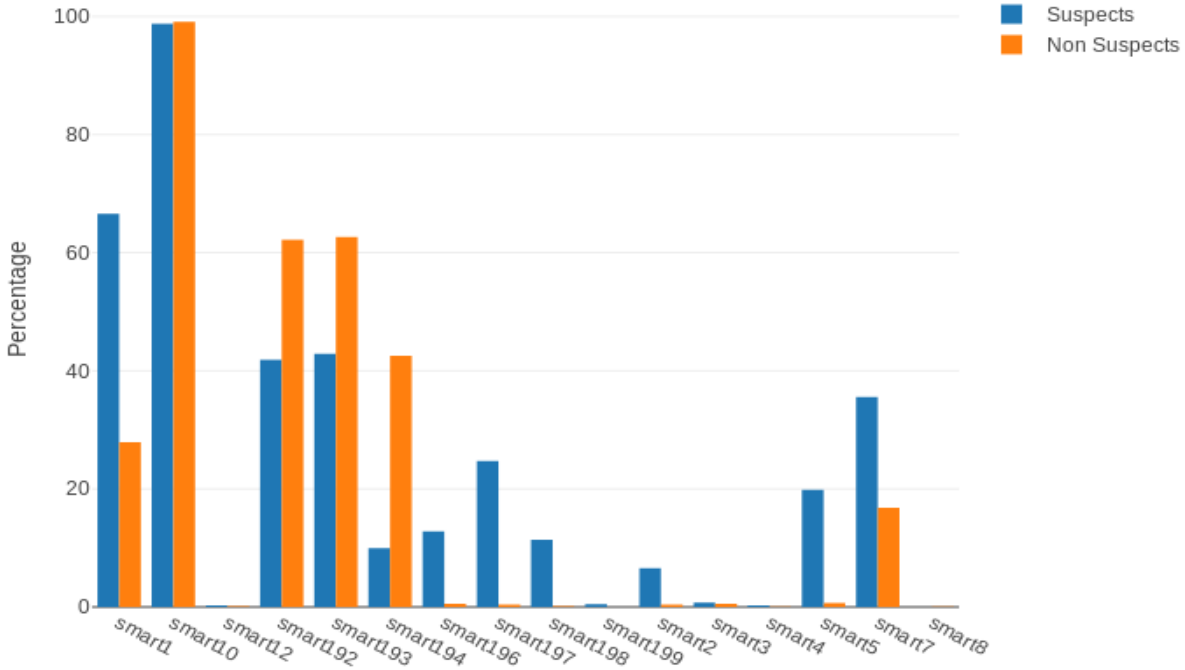
Kaplan-Meier survival curves (II) – Failed disks per vendor



Kaplan-Meier survival curves (III) – Disappeared disks per vendor



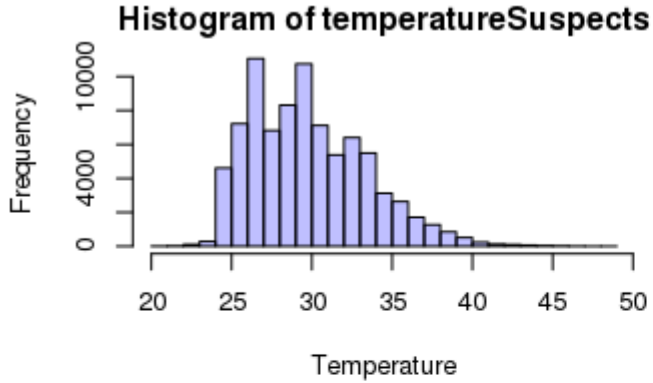
S.M.A.R.T. metrics variation

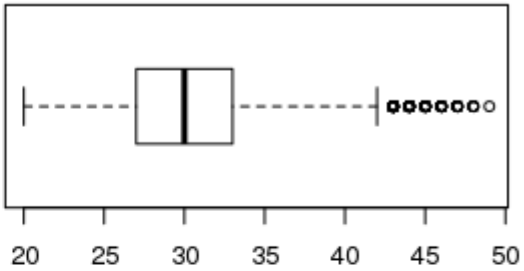


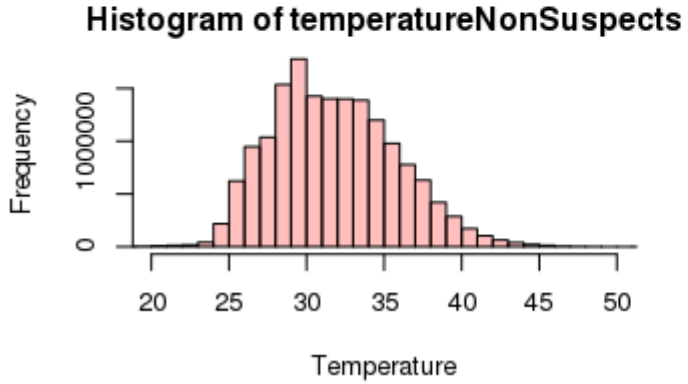
Significant S.M.A.R.T. metrics

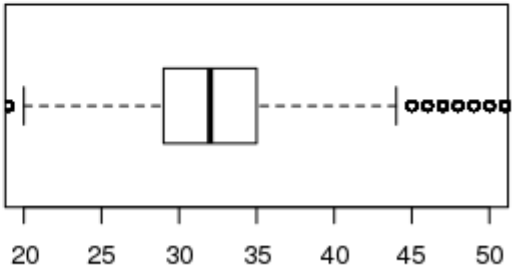
- Based on this, we start analyzing the variation of:
  - Smart2: Throughput Performance
  - Smart5: Reallocated Sectors Count
  - Smart196: Reallocation Event Count
  - Smart197: Current Pending Sector Count
  - Smart198: (Offline) Uncorrectable Sector Count

No correlation to temperature? (Smart194)

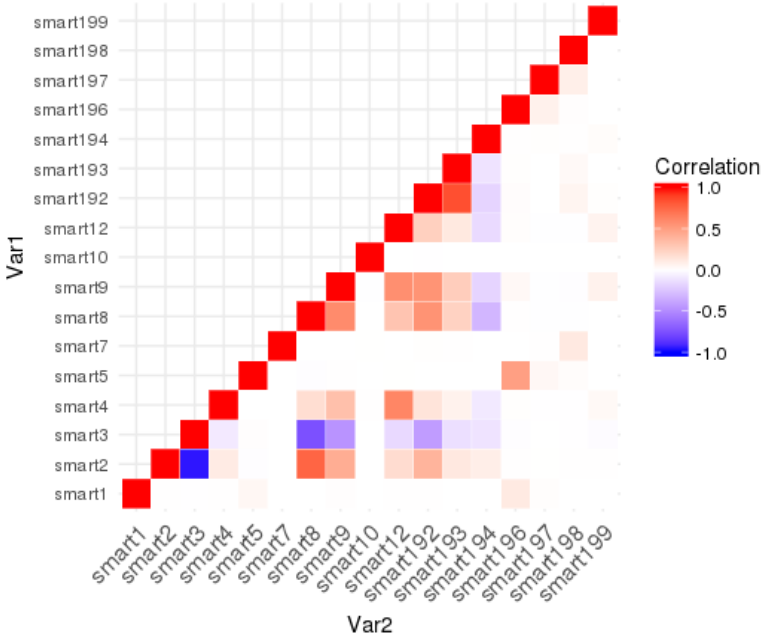




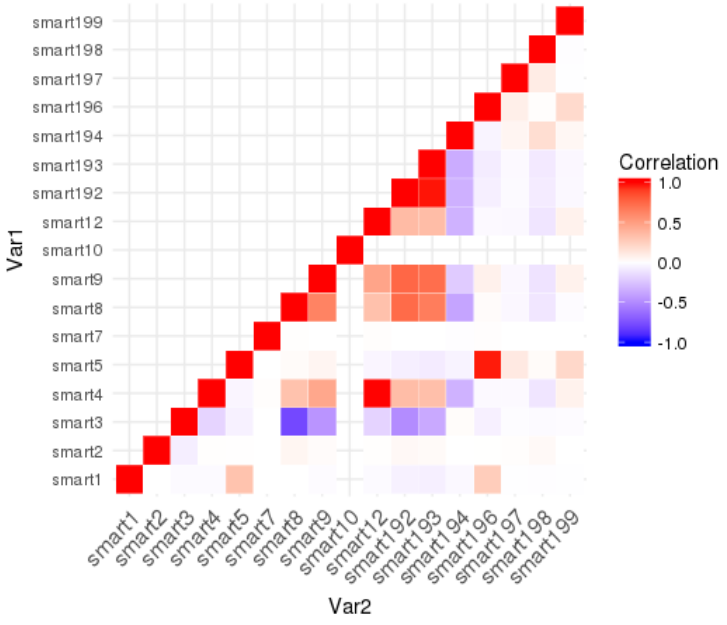




Are any of these values correlated?



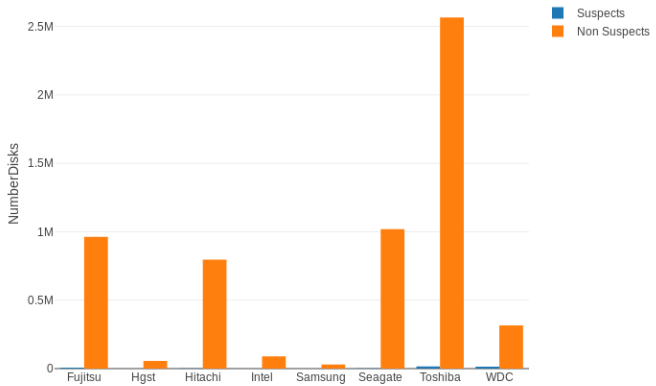
Non Suspects



Suspects

First results

- Global MeanTime Between Failures: 1 failure every 1.6 days
- Global annualized failure rate: 0.89%
- Error on the failure rate: 0.049%
- Global Average Age: 1095 days (~ 3 years)
- Standard deviation of the age: 660.15 days



First results (II)

Vendor	Annualized failure rate	% population w/ metadata	Average age	Standard Deviation of Average Age
Fujitsu	1.84%	16.5%	2214 days	245 days
Hgst	0%	0.96%	1149 days	169.2 days
Hitachi	0.32%	13.58%	1717 days	276.9 days
Intel	0.40%	1.54%	412 days	156.5 days
Samsung	2.39%	0.52%	722 days	50.6 days
Seagate	0.25%	17.37%	1481 days	255.8 days
Toshiba	1.45%	43.9%	1888 days	330 days
WDC	4.52%	5.63%	1424 days	633.2 days

Conclusions and next steps

- We can conclude that there is no apparent relation between temperature and disk failure in the conditions of CERN.
- We can not yet see the impact of failures in the beginning of the lifetime of a disk.
- We obtained five S.M.A.R.T. metrics that seem to be relevant and related to failure.
- Next step: Use these values as predictors in a model, and see if the prediction can be useful to predict failure.
- Next step: See the evolution of disks related to age since the beginning of their life cycle.
- Next step: Analyse correlation between smart196 (Reallocation Event Count) and smart5 (Reallocated Sectors Count), only high on Suspects.