CERN

Disk Failures in the EOS Setup at CERN

A first systematic look at one year of collected data.

Alfonso Juan Portabales Gonzalez, Dirk Duellmann
CERN, Storage Group

Motivation

- All LHC experiments are preparing for run periods with significant increase in data volume and rate
  - storage and **media cost an important planning input**
    - CERN deploys almost 100k disk devices
  - data access failures and service recovery after media failures require **human effort from users and sites**
- Can we predict and prevent data access problems?
  - **identify less reliable hw types** or deployment modes
  - **proactively relocate data** to reduce human effort
- Can we collect and **share failure information for HEP workloads**?
  - among sites, users and storage hw and sw developers

SMART Disk Metrics -

Self-**M**onitoring **A**nalysis and **R**eporting **T**echnology

- SMART metrics tend to be vendor/model dependent:
  - Initial studies did not reach clear or widely applicable conclusions
  - **Reasonably sized data set is required** to use more sophisticated **statistical or ML methods**
- Recent studies of SMART based failure models for hard drives
  - **Backblaze**: collects and publishes drive data since 2013(!)
    - MSST 2017: Annualised Failure Rate around ~3.33%
    - 77% of failed drives show smart attributes, IBM ML model
  - **Google**: 60 days after the first uncorrectable error on a drive (Smart[198]) a drive is 39 times more likely to fail
    - but 36% of failed drives showed no smart error at all

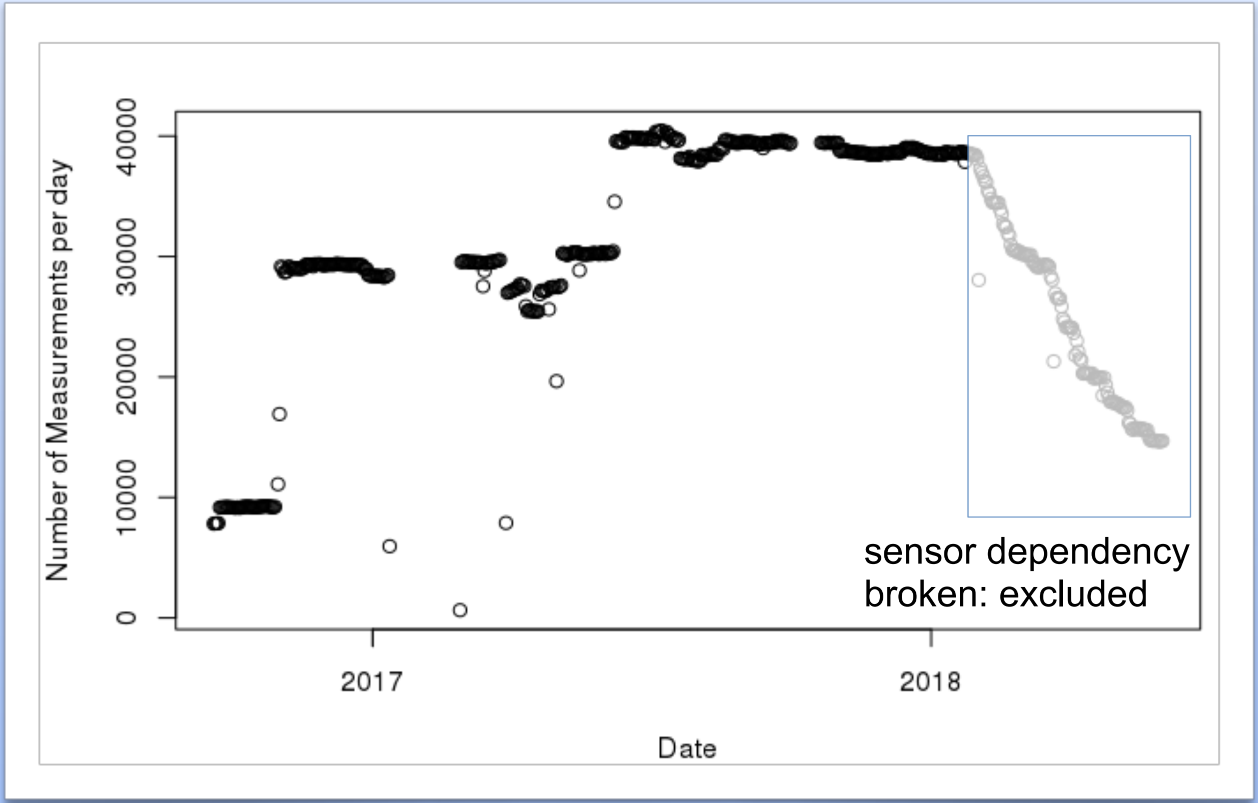
Challenges: Data Availability & Quality

- This study was not a designed measurement!
  - (previous) Fabric disk sensor: collected only smart summary (1-bit)
  - EOS operations: smart metrics with ~daily collection
  - Disk model information: scraped periodically “by-hand”
  - EOS scrubbing: analysis of checksum failures has started, but is not yet included here
- Different data sources, and different data structures
  - For smaller sites this may more complicated (due to smaller statistics) or more easy due to fewer data sources involved
- **Data that is not actively analysed is usually wrong or not existing!**
  - **Metrics for daily operation $\neq$ metrics for analytics**

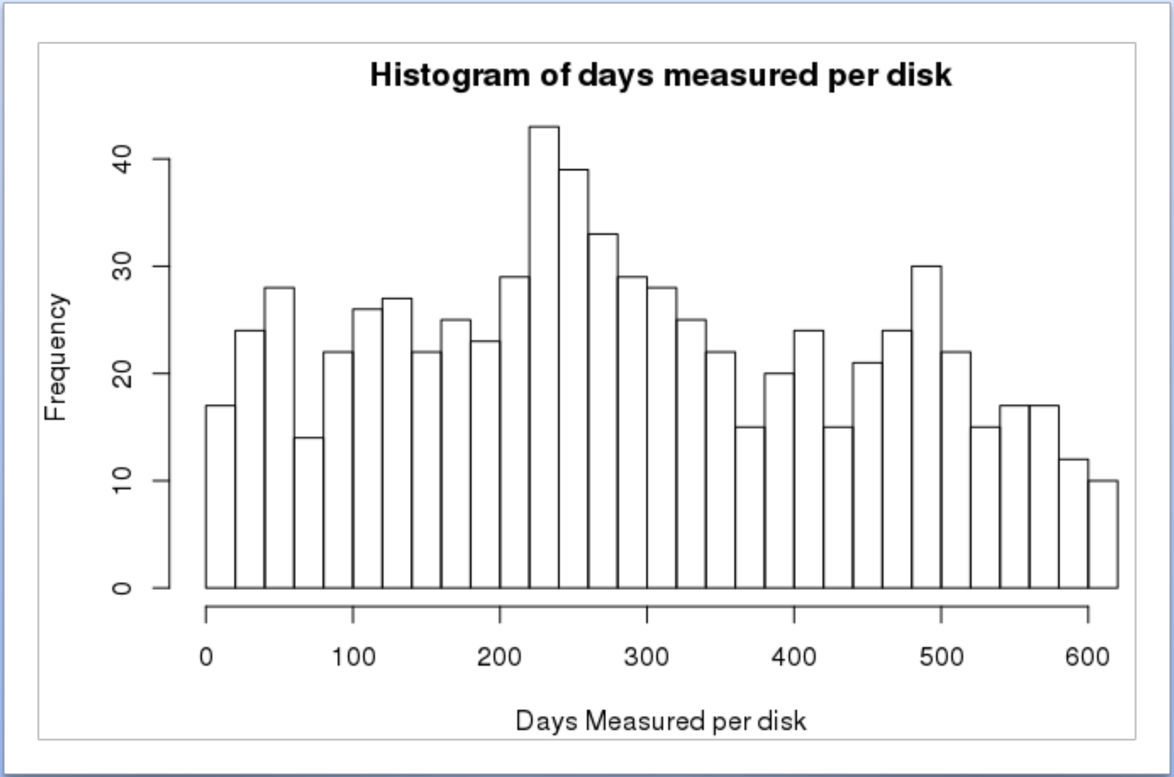
Input Dataset: Some Overall Statistics

- Days with smart measurements: **551**
  - oldest measurement included: 620 days ago
- Number of EOS disks measured per day:
  - between 635 and 40563
  - average per day: 31770
- Total number of unique disks: **45874**
- Complete vendor device information for **35%** of all measurements.
- **Deployment of a new fabric disk probe is imminent**
  - **will provide more complete drive meta data and smart info for all production drives at CERN**

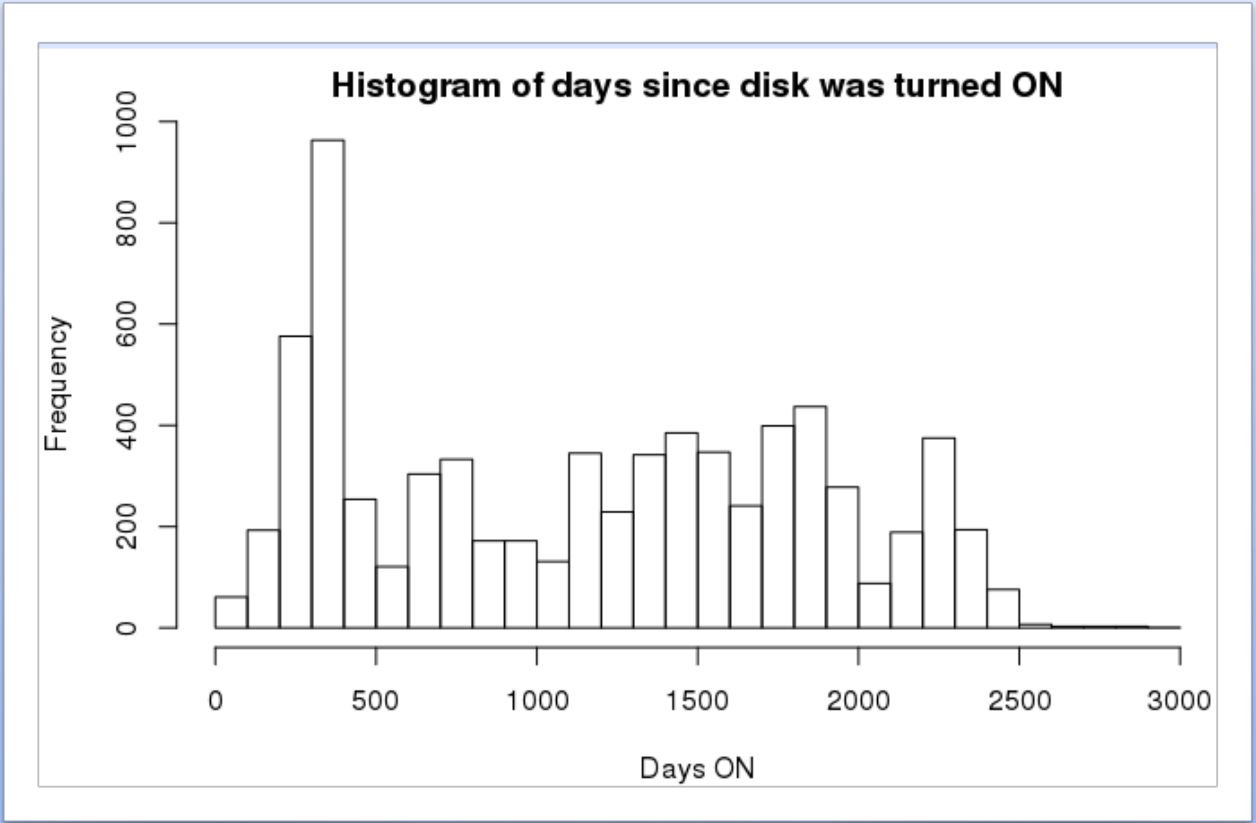
Metric Collection: Measurements per Day



Number of Days Measured per Disk



Per Disk: Days in Operation

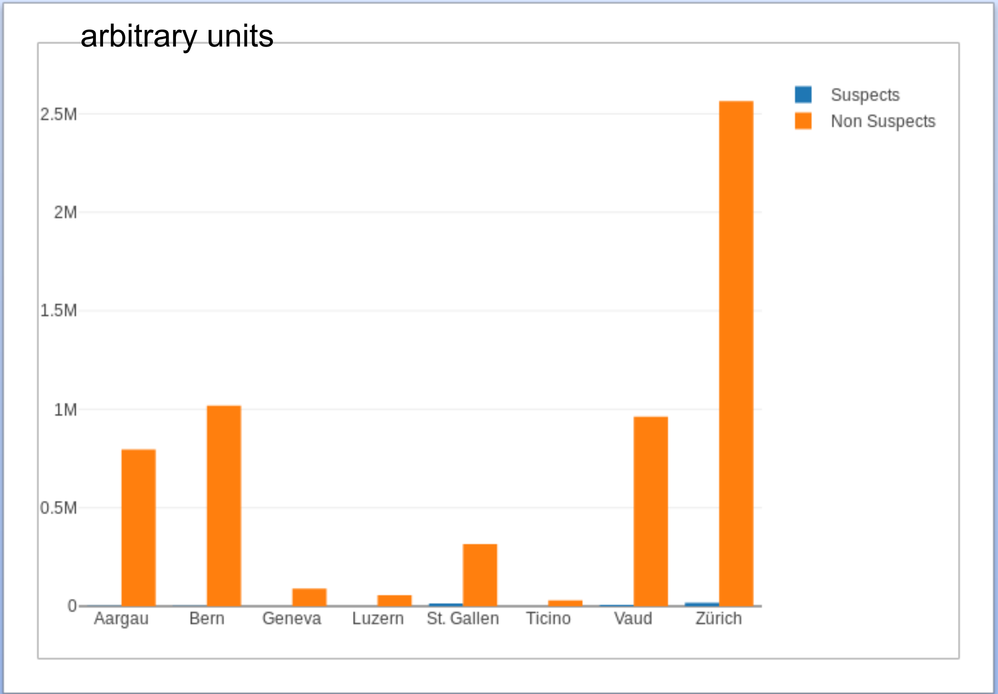


How to Define Drive Failure?

- A disk is considered as “suspect” of failure
  - when it disappeared from the daily smart data collection
  - while the other disks in the machine continue to report
- This basic label divides our population into two groups
  - **Suspects** and **Non Suspects**
- We checked this rule against other possible causes
  - eg disk exchange within the centre, correlated outages etc.
  - we can trace disks through the centre via their unique serial
- Note: ~68% of all disks have been substituted or stopped being recorded
  - Also the replacement data is useful to review the **hardware flow through the data center**

Overall Results

- Annualised Failure Rate: 0.89% +/- 0.05% (stat.)
- Average disk age: 1095 days
- EOS at CERN: MTBF
  - 1 failure every 1.6 days
- Relative vendor contributions - names replaced by CH cantons



Results by “Vendor”

Vendor	Failure Rate [%/yr]	MD complete [%]	Average Age [days]	SD Age [days]
Vaud	1.84	17	2214	245
Luzern	0.00	1	1149	169
Aargau	0.32	14	1717	277
Geneva	0.40	2	412	157
Ticino	2.39	1	722	51
Bern	0.25	17	1481	256
Zürich	1.45	44	1888	330
St.Gallen	4.52	6	1424	633

D

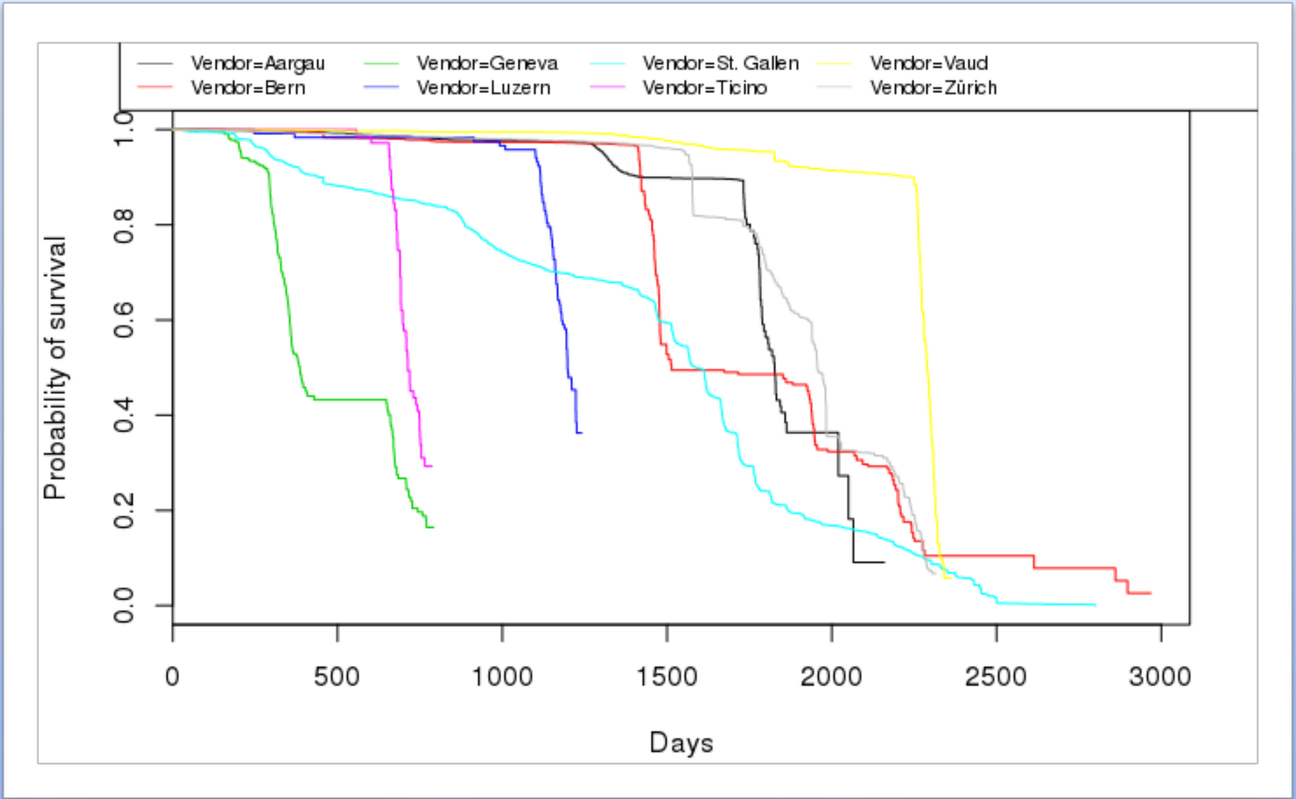
Kaplan-Meier Survival Curves

- Analysis is based on Kaplan-Meier survival curves
  - used eg on clinical trials in medicine
  - easy to calculate eg via R package **`survival`**, or python **`lifelines`**

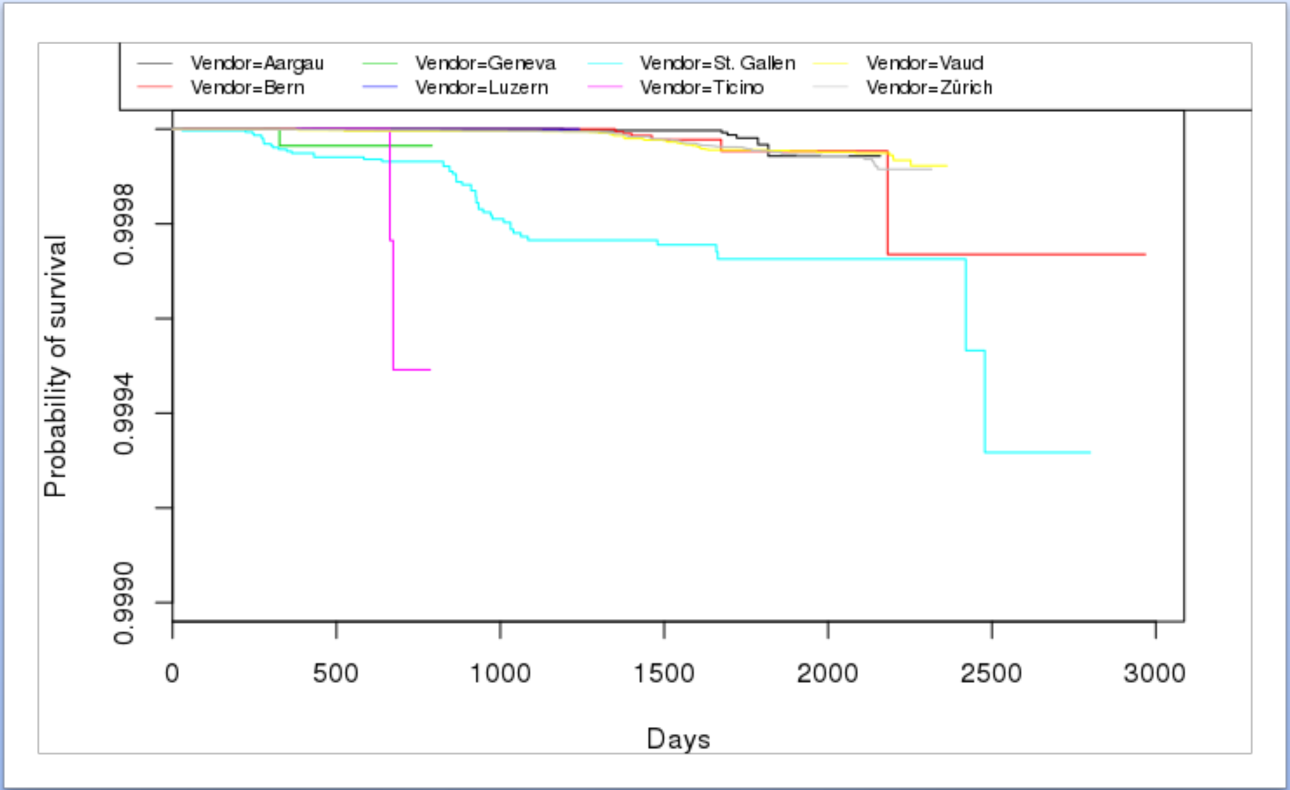
$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

- Initially producing survival rate per vendor:
  - increased statistics will allow model based analysis
- We consider two survival curves:
  - one based on single drive failure
  - one based on drive set substitution

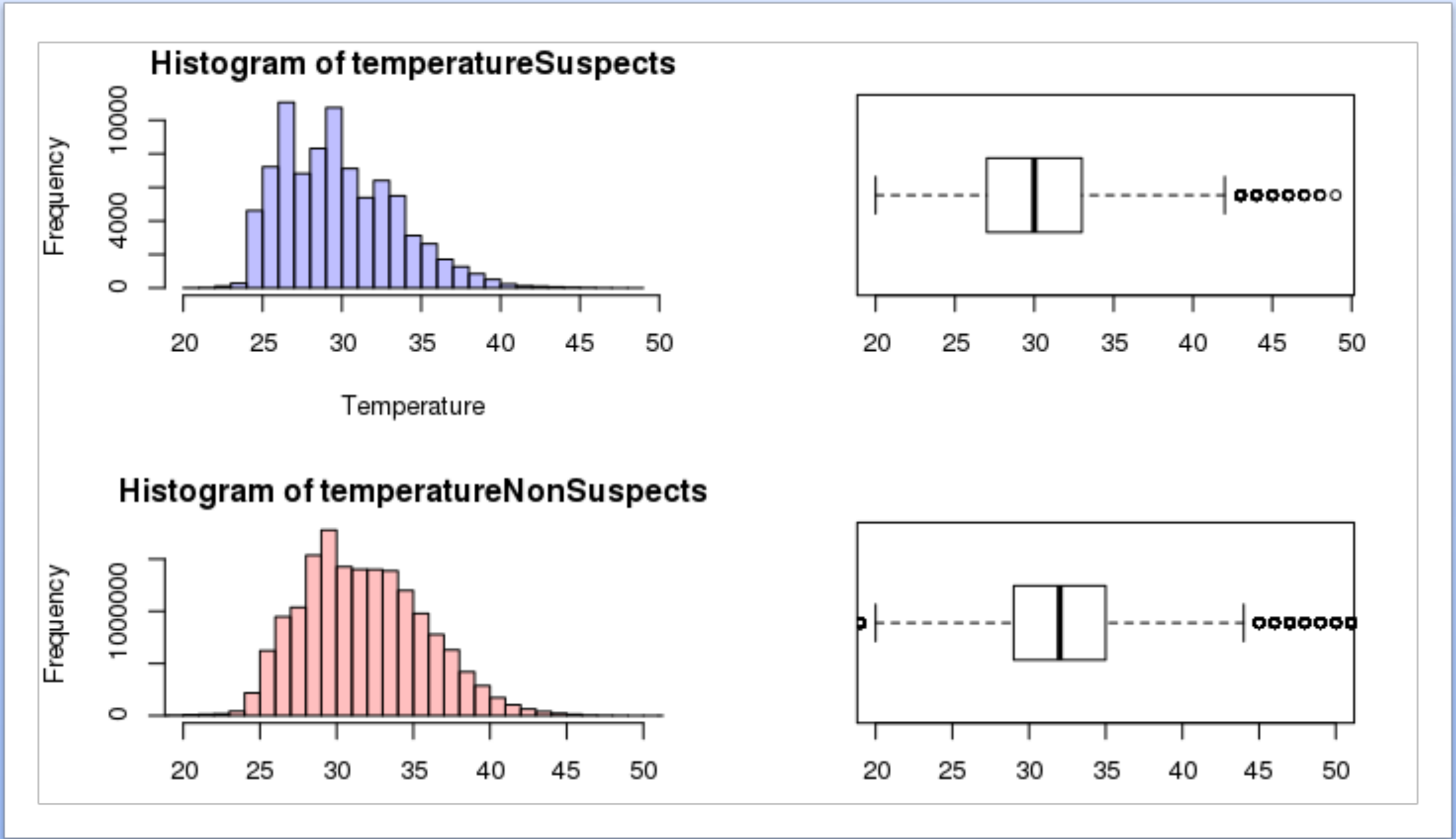
Survival Curves: Disk Replacements by Vendor



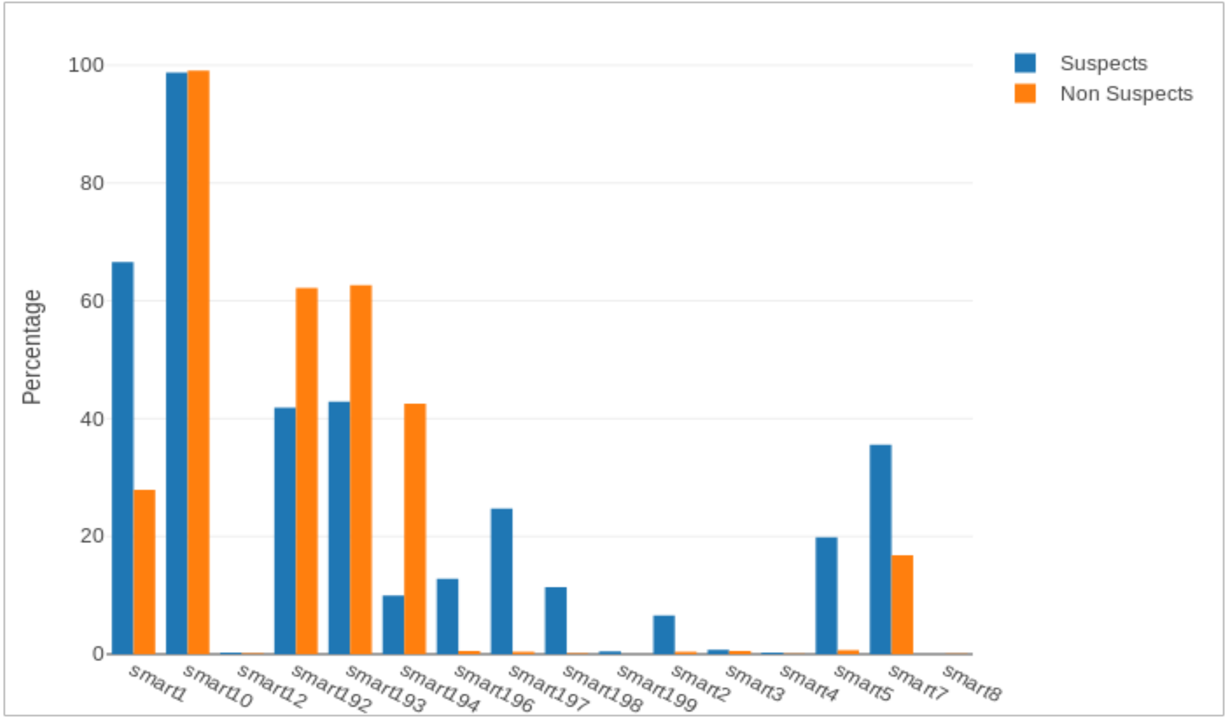
Survival Curves: Failed Disks per Vendor



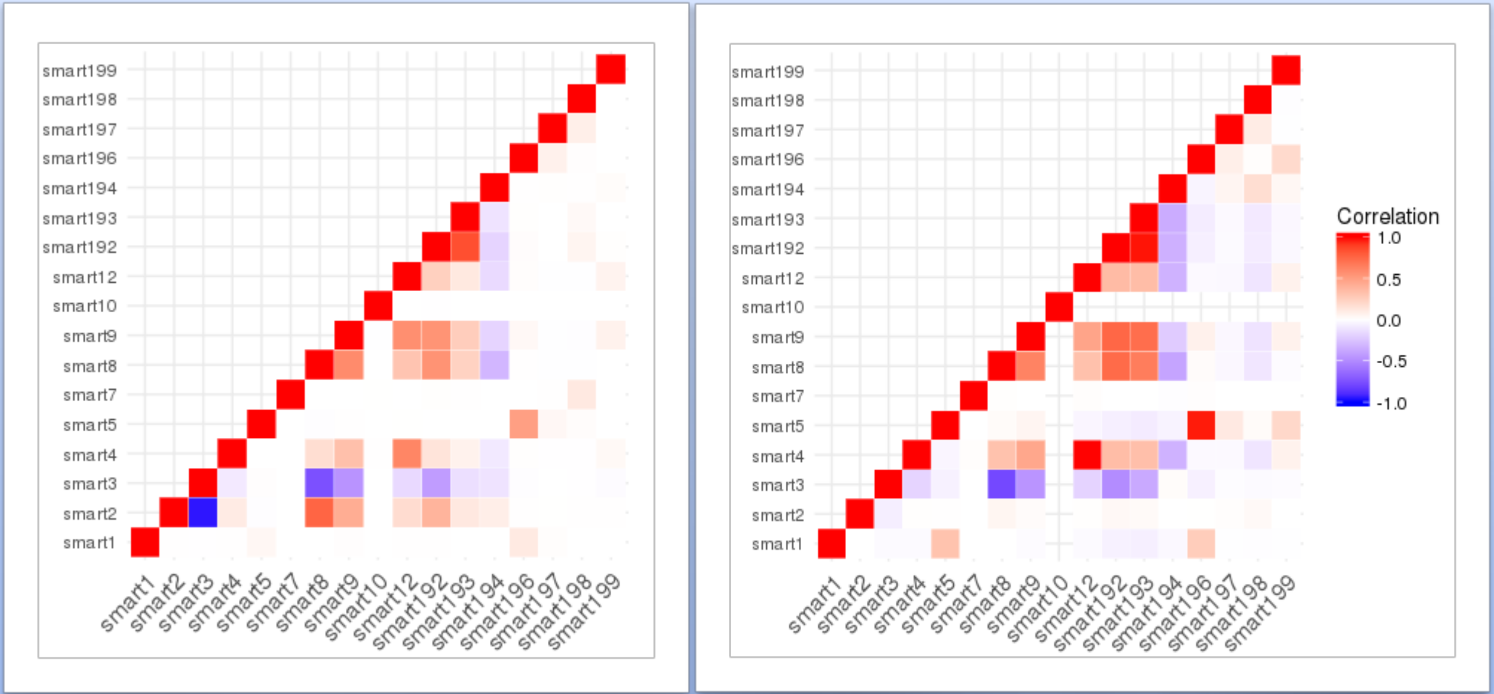
Are Failures correlated with Temperature?



SMART Metric Variation



Are the Metrics Correlated?



Non Suspects

Suspects

Conclusions and Next Steps

- With current statistics and under CERN conditions and workload, we
  - measured overall annualised failure rate (AFR) as 0.89%+/-0.05%
  - no visible correlation between disk temperature
  - no increased failure rate for young disks (burn-in period sufficient)
  - identified relevant SMART metrics as input for a failure prediction

- With more than **tripled statistics expected from new fabric disk probe**
  - train a RNN for failure prediction model
  - review failures by model and by age with full CERN population
  - quantify impact of media faults wrt. other sources of unavailability

Thanks for your attention! Questions?