

Federico Gargiulo

Tutor: Pasquale Arpaia – co-Tutor: Nicola Mazzocca

XXIX Cycle - I year presentation

Fault Diagnostics by physical and artificial intelligence  
with heterogeneous measures



UNIVERSITÀ DEGLI STUDI DI NAPOLI  
FEDERICO II

Instrumentation & Measurement  
for Particle Accelerator Lab



# Summary

- Personal background
- Research Activity
- Research Context
- Achievements
- Improvements and future works



# Personal Background

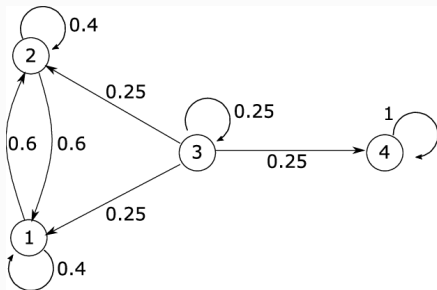
- Education
  - Master Degree in Computer Engineering with a curriculum in Embedded Systems.
    - Master Thesis: Diagnostics with Hidden Markov Model in cryogenic plants for particle accelerators
  - PhD student at ITEE
- Experience
  - Vocational
    - Technical Student in Control Section for Cryogenics at CERN from Feb 2018 until Gen 2019
    - Data Scientist in Analysis Section for Data Center at CERN from May 2019
    - Freelance involved in several project of development in embedded systems
  - Miscellaneous
    - Project Manager in a no-profit organization Horizon Lab  
[www.facebook.com/associationhorizonlab](http://www.facebook.com/associationhorizonlab)
    - Author of several European project accepted and co-financed by European Commission



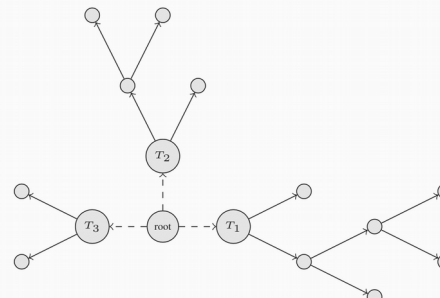
# Research Activity

PhD Project Title: *Fault Detection and Fault Prediction with heterogeneous measures in electromechanical devices*

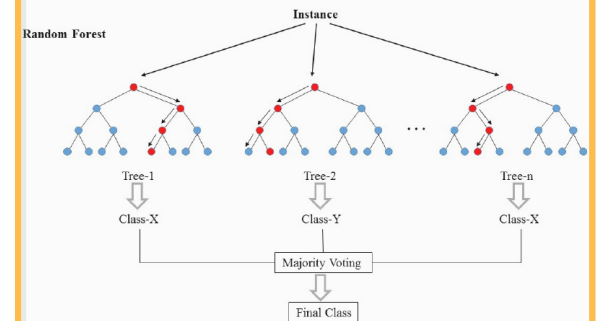
### Hidden Markov Models



### Regularized Greedy Forest

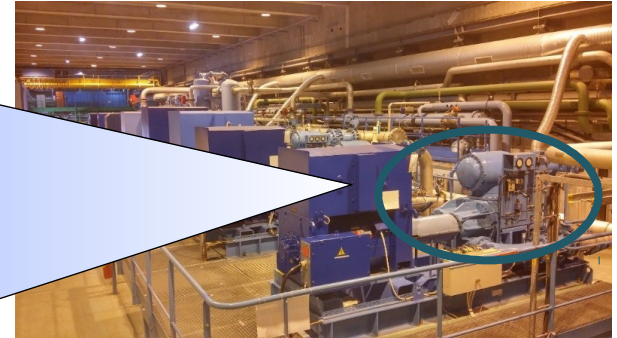
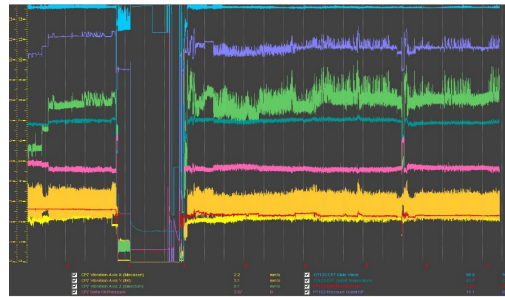


### Random Forest



# Problems

In a complex device the failure of an internal component can cause a complete system failure, sometimes it may be latent and not compromise the usability before a few days of nominal operation.



There are many cases in which it is possible to collect measurements but the complexity of systems does not allow to find an easy and linear relationship between an ordinary measurement and a device failure.

Often machine learning methods are a good way for parametrizing complex models

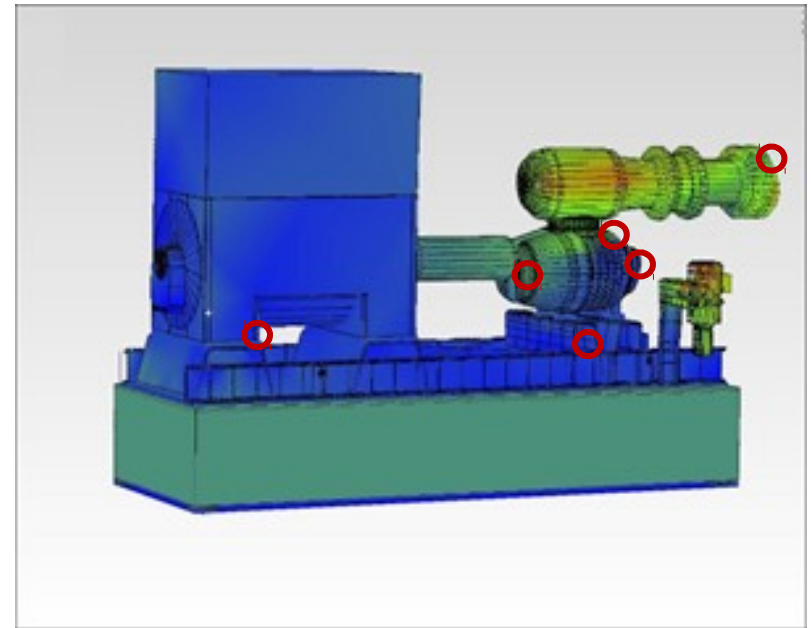






# Problems (1/2)

Fluid machines are composed of hundreds of components and each can present completely different problems, with completely different effects on the final system.

Until now in the literature there was no fault detection methodology valid for the case study (threshold, missing values etc).

We addressed the problem using machine learning techniques based on **Hidden Markov Models**. Results below.



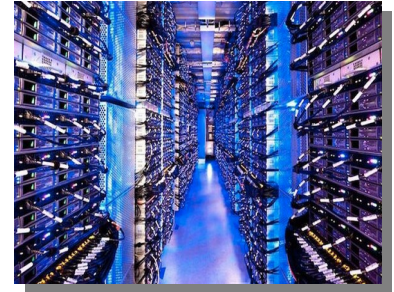
			
<b>IMI 643A00</b> Piezoelectric vibration sensor	<b>PT100</b> Temperature transmitter Platinum Resistance	<b>Rosemount 1151</b> Pressure Transmitter	<b>SINEAX 1538</b> Transducer for AC current



# Problems (2/2)

The world of data centers is highly heterogeneous. Due to the rapid technological progress. So far, fault prediction methods do not provide good results because of several reasons. (early replacing, new models, fault in sockets etc).

Amazon



Currently there are few novelties in the new hard disk models and the technological progress in electromagnetic hard disks is slowed down. So the number of general failures observed is increased as no more hard drives are replaced due to progress but due to failure.

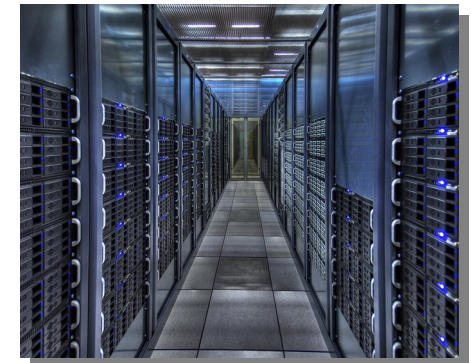
BackBlaze



CERN Data Center in numbers:

- Total number of disk measured per year: ~98k
- Average disk measured per day: ~62k
- Number of days which we have measures of: ~460
- Total replaced HD per year: ~15k
- Number of models used: 120
- Number of Brands: 15

CERN



Federico Gargiulo



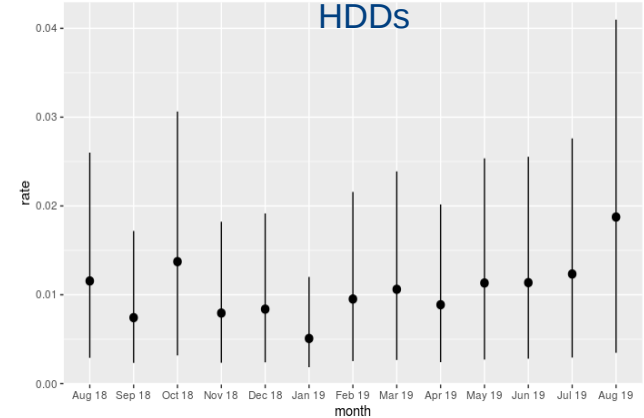
# Problems (2/2)

- Monthly Failure Rate (MFR) & Annual Failure Rate (AFR)
- Projecting yearly MFR to obtain AFR and quantify estimation errors with a confidence level of 95%

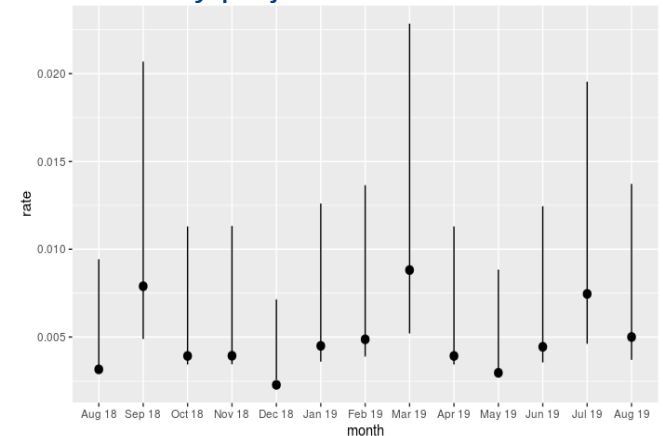
Currently we have an annual failure rate around 1% of running disks.

#	Month	Total HDDs	Broken HDDs	MFR	AFR	Y min Gaussian	Y max Gaussian
1	Aug 18	64230	61	0.09%	1.16%	0.29%	2.60%
2	Sep 18	63900	39	0.06%	0.74%	0.23%	1.72%
3	Oct 18	63800	72	0.11%	1.37%	0.32%	3.06%
4	Nov 18	67431	44	0.07%	0.79%	0.23%	1.82%
5	Dec 18	68231	47	0.07%	0.84%	0.24%	1.92%
6	Jan 19	69469	29	0.04%	0.51%	0.18%	1.20%
7	Feb 19	69001	54	0.08%	0.95%	0.25%	2.16%
8	Mar 19	69919	61	0.09%	1.06%	0.27%	2.39%
9	Apr 19	71265	52	0.07%	0.89%	0.24%	2.02%
10	May 19	71970	67	0.09%	1.13%	0.27%	2.54%
11	Jun 19	67427	63	0.09%	1.14%	0.28%	2.55%
12	Jul 19	67051	68	0.10%	1.23%	0.29%	2.76%
13	Aug 19	72666	112	0.15%	1.88%	0.35%	4.10%

Monthly projection of AFR for HDDs

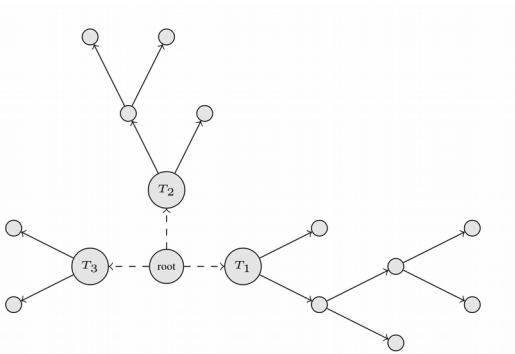


Monthly projection of AFR for SSDs





# Problems (2/2)



*Regularized Greedy Forest (RGF) is a decision forest learning algorithm performed to have a tree-based structure for nonlinear decision. It was introduced in 2014 by R.Jhonson and T.Zhang and has been implemented in libraries for python, R, C++ etc.*

For the 5 most used magnetic models in our data center we can predict with good results which hard disk is close to being replaced and which is not. The methodology has been tested only on the measures of the day before the replacement. Soon we will begin to test even on the last antecedent days.



Context		Training				Validation				Impact			
Model	Splitting Ratio	Leaf	Total number	Num Healthy	Num Broken	Healthy	Score healthy pred.	Broken	Score broken pred.	HD by model	$\frac{HDD \text{ by model}}{\text{Total number of HDD}}$	HD broken by model	$\frac{\text{Broken HDD by model}}{HDD \text{ by model}}$
TOSHIBA_MG04ACA600E	60%-40%	1000	97	50	47	4.9k	83.46%	17	88.23%	12350	15.50%	64	0.52%
HGST_HUS726060ALE610	60%-40%	1000	124	62	62	4.8k	97.29%	32	96.87%	11959	15.01%	94	0.79%
HGST_HMS5C4040BLE640	60%-40%	1000	172	86	86	4k	99.04%	72	93.05%	10395	13.05%	158	1.52%
ST4000NC001-1FS168	60%-40%	1000	438	221	217	3.6k	93.26%	162	96.61%	9565	12.01%	379	3.96%
Hitachi_HUA5C3030ALA640	60%-40%	1000	77	40	37	3.2k	93.01%	25	88.00%	8040	10.09%	62	0.77%

# Achievements

- **Fault Detection on Electromechanical devices for the treatment of fluids**
- **Successfully results using Hidden Markov Models:**  
It has been verified and validated that measures coming from trends slightly off the nominal conditions are well detected.

*The work was successfully completed and published on Measurement Elsevier Journal. Cod. MEAS-D-19-01806R1.*

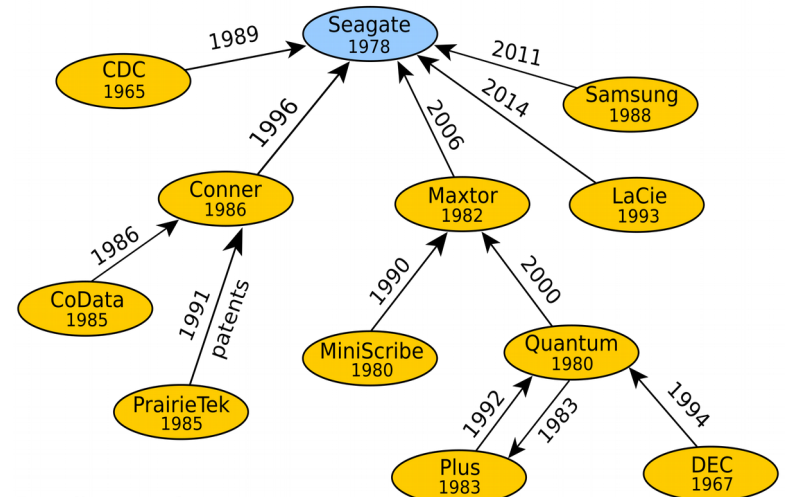
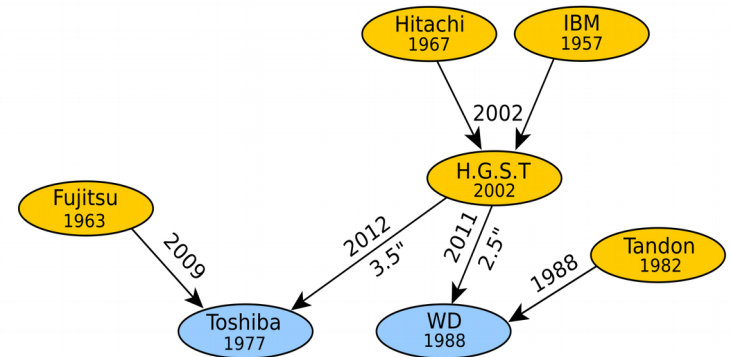
- **Fault Prediction project for storage drives**

So far, good first results through Regularized Greedy Forests. SMART (Self-Monitoring, Analysis and Reporting Technology) measures collected on ~ 90000 in 11 months has lead excellent results: Hard disk close to be broken are well predicted in 80% of general cases. In few models up to 90% of well prediction.



# Improvements and future works

- Generalization of methodologies (several models, typology, models rare used etc)
- Looking for systematic factors and faults in models by vendor, in order to make predictive algorithms highly reusable.
- Analysis of impacts of company fusions, programmed obsolescences, MTTF unrealistic, Enterprises Drives VS. Customers Drives



Founding year for startups  
First shipment for other new entrants

#	Vendor	N
1	HGST	10542473
2	TOSHIBA	5244580
3	HITACHI	4833418
4	INTEL	3912222
5	ST	3419261
6	SEAGATE	1785539
7	WDC	1415733
8	MICRON	1244625

#	Vendor	N
9	SAMSUNG	981664
10	SANDFORCE	161137
11	SANDISK	84912
12	LSI	5395
13	AVAGO	2615
14	SATA	345
15	MK	26



Thanks for your attention

