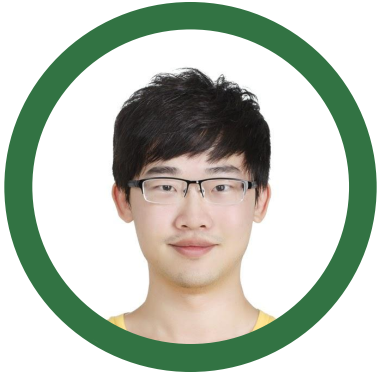
To Fail or not to Fail?

Predicting Hard Drive Health

Guardians of the Memory

Andreas

Dipl.-Ing.
mech. Engineering
with a backround in
energy technology



Chang-Ming

PhD in Physics with
a background in
theoretical modeling

Daniela

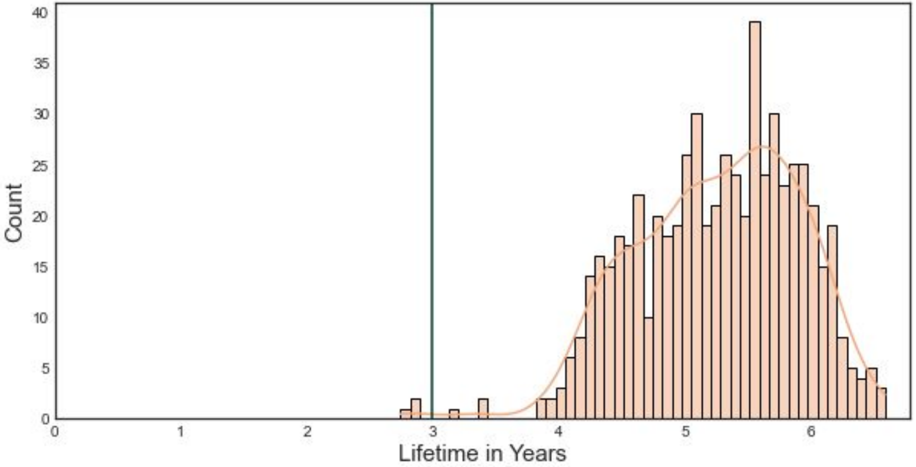
Application
Manager with a
background in
language studies

Felix

PhD in Physics with
strong background
in data analytics

The Stakeholder - Cloudwaver

- Startup offering cloud storage as a service
- Maximize hard drive usage beyond 3 years



The Task

Predict if a hard drive fails in the coming 30 days:

- **Reduce investments** for hard drives by up to **40%**

  (5 years vs 3 years of usage)
- **Enhance sustainability**
- Maintain growth even under global chip shortage

The Hard Drives Dataset

BACKBLAZE

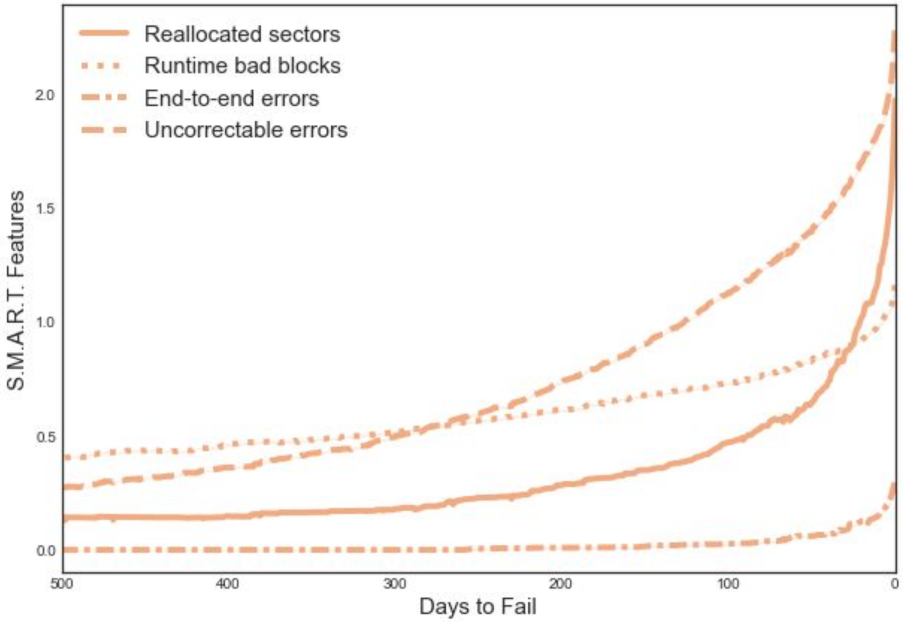
- Around 205k hard drives in 65 models (2021)
- 174 S.M.A.R.T. parameters recorded daily
- **S.M.A.R.T:** Self-Monitoring, Analysis and Reporting Technology
- Focus on the model of interest for Cloudwaver (2019 to 2021)

https://www.backblaze.com/b2/hard-drive-test-data.html

Crucial S.M.A.R.T. Features

Examples of S.M.A.R.T Features:

- **Reallocated sectors**
- **Runtime bad blocks**
- **End-to-end errors**
- **Uncorrectable errors**
- Temperature
- Power on time



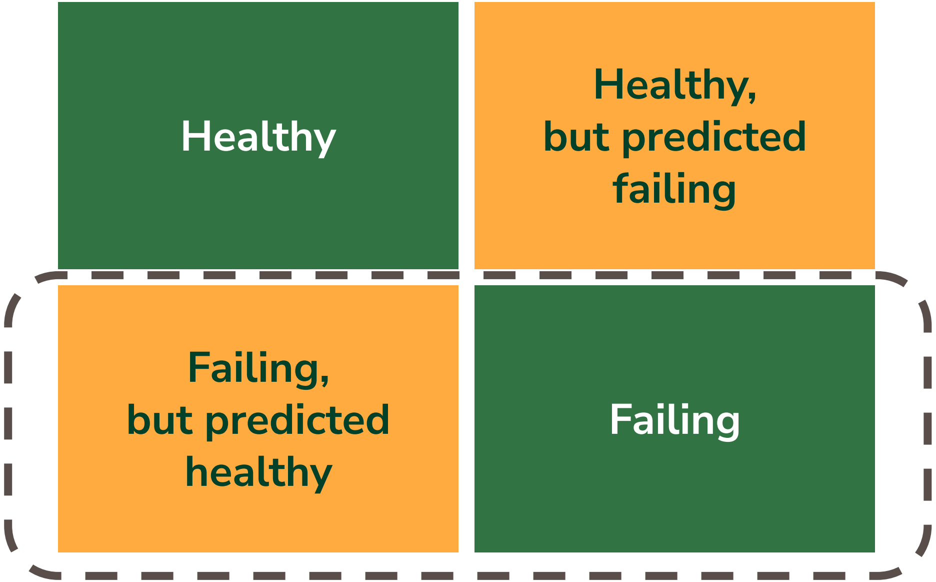
Model Evaluation

F2-score: accounts for recall and precision

Healthy	Healthy, but predicted failing
Failing, but predicted healthy	Failing

Model Evaluation

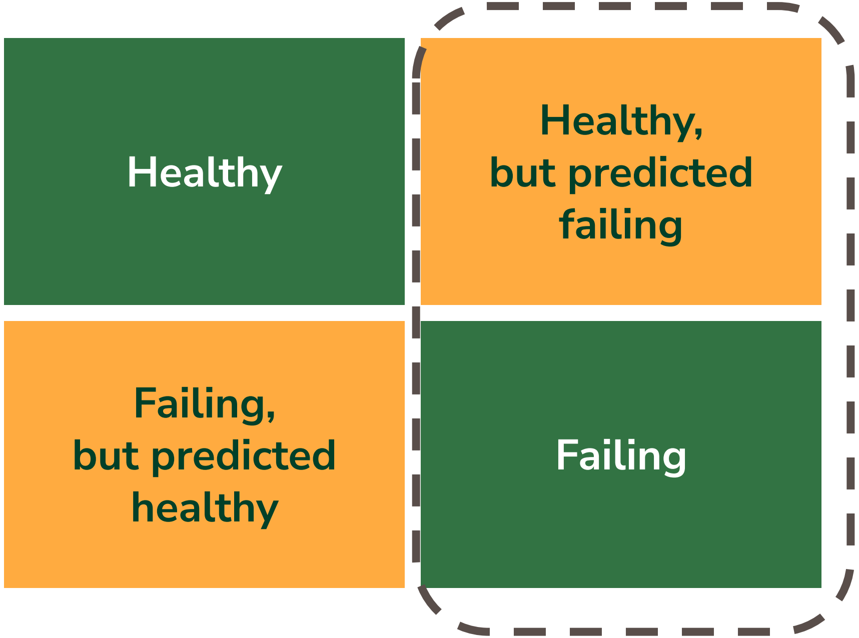
F2-score: accounts for *recall* and precision



Healthy	Healthy, but predicted failing
Failing, but predicted healthy	Failing

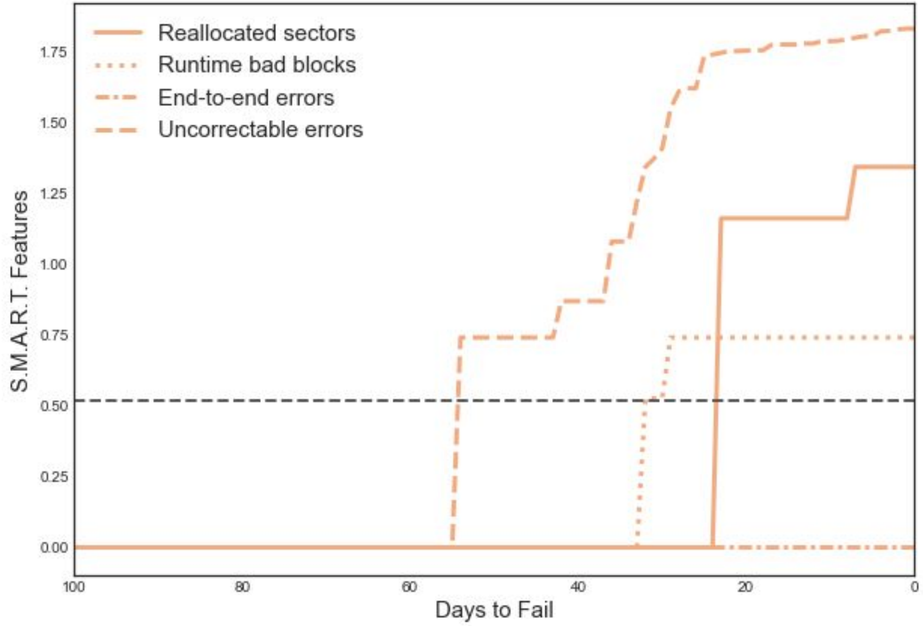
Model Evaluation

F2-score: accounts for recall and *precision*



Healthy	Healthy, but predicted failing
Failing, but predicted healthy	Failing

Baseline Model



29 %

F2-Score

45 %

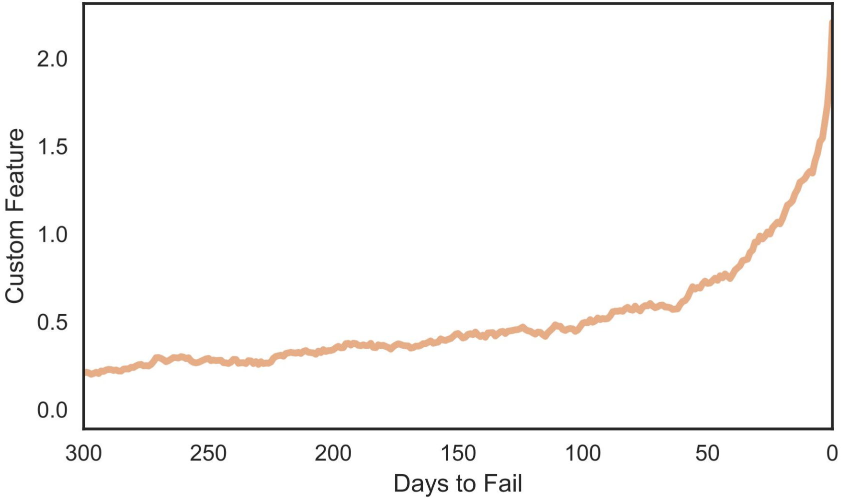
Recall

12 %

Precision

Final Model

- Low-dimensional artificial neural network
- Custom feature captures dynamics of relevant S.M.A.R.T. features



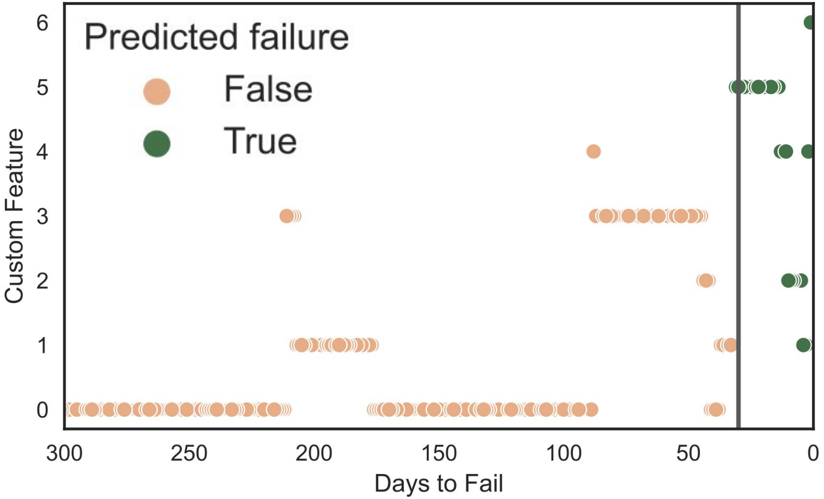
44 %
F2-Score

61 %
Recall

21 %
Precision

Final Model

- Low-dimensional artificial neural network
- Custom feature captures dynamics of relevant S.M.A.R.T. features



44 %

F2-Score

61 %

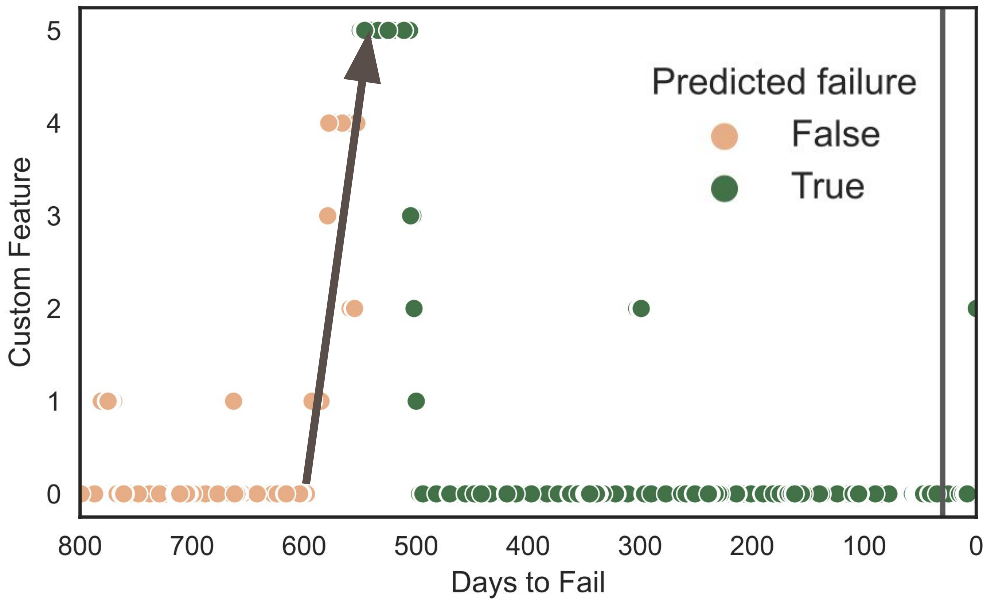
Recall

21 %

Precision

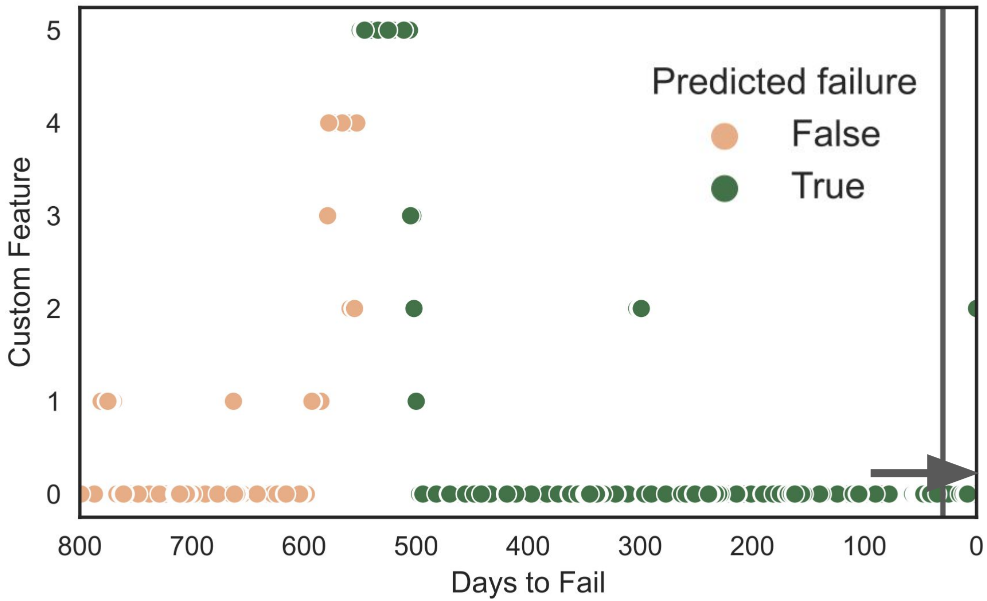
Limitations

- "Fake" failures



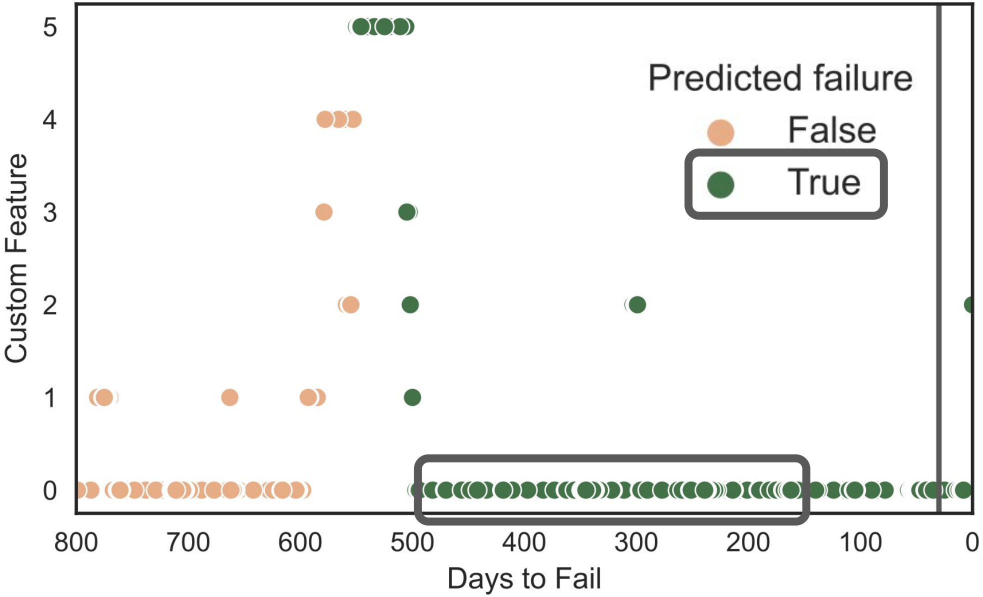
Limitations

- “Fake” failures
- “Silent” failures



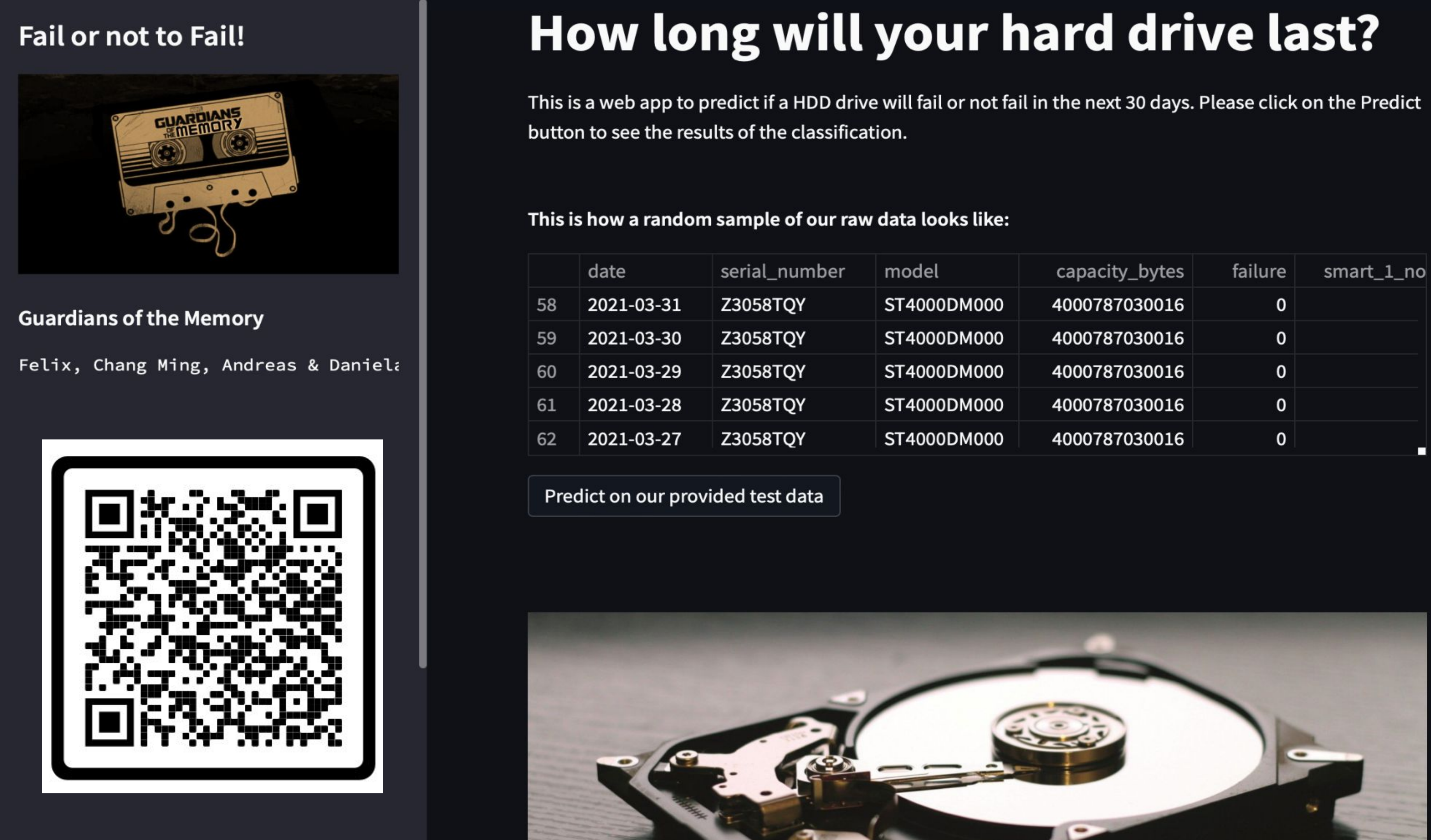
Limitations

- “Fake” failures
- “Silent” failures
- Time-related features



Outlook

- Try to tackle limitations
- Anomaly detection methods
- Include additional hard drive models



Fail or not to Fail!

Guardians of the Memory

`Felix, Chang Ming, Andreas & Daniela`

How long will your hard drive last?

This is a web app to predict if a HDD drive will fail or not fail in the next 30 days. Please click on the Predict button to see the results of the classification.

This is how a random sample of our raw data looks like:

	date	serial_number	model	capacity_bytes	failure	smart_1_no
58	2021-03-31	Z3058TQY	ST4000DM000	4000787030016	0	
59	2021-03-30	Z3058TQY	ST4000DM000	4000787030016	0	
60	2021-03-29	Z3058TQY	ST4000DM000	4000787030016	0	
61	2021-03-28	Z3058TQY	ST4000DM000	4000787030016	0	
62	2021-03-27	Z3058TQY	ST4000DM000	4000787030016	0	

Additional notes Felix - smart999

Calculation:

- 30 day EMA (data as timeseries)
- trigger if 5% increase over EMA
- sum of trigger over all non age-related features

Plots?

- mean over time
- EMA and values for one feature

Additional notes Felix - limitations

Fake failures:

- EMA of smart999, slope/curvature? Idea: if smart999 jumps up but goes down again its a fake failure?
- Need failure prediction e.g. 7 days in a row to say disk will fail

Silent failures:

- Failing on different time scale (minutes, hours)? Not seen in data → make predictions every 10 mins to capture this
- Can we get frequent data short before failure?

Age-related features:

- steady increase over time e.g. power-on-hours, data written,...
- atm improves our model, with better features perhaps not needed any more?
- Way to regularize the importance of those features?

Additional notes Felix - Outlook

Idea to overcome limitations: see last slide

Additional approaches:

- anomaly detection: autoencoders, clustering, dimensionality reduction, isolation forest, oneclassSVM
- survival analysis
- novelty detection
- time-series approach

Additional data for EDA, deployment for different models, more old data (2013?-2019)

MARVEL
GUARDIANS
OF THE
MEMORY

Welcome to our presentation!