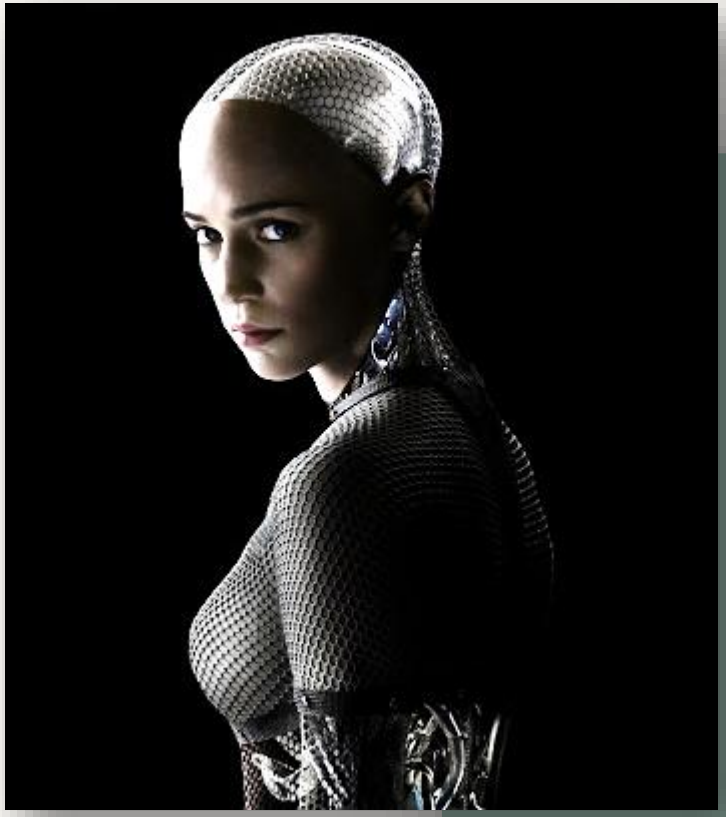
AKADEMIA
GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

AGH

INŻYNIERIA
MECHATRONICZNA

Advanced AI models in engineering

Who will clean this mess?



dr hab. inż. Ziemowit Dworakowski

KATEDRA ROBOTYKI I MECHATRONIKI
WYDZIAŁ INŻYNIERII MECHANICZNEJ I ROBOTYKI

25/10/26

(Example thanks to the courtesy of prof. E. Brzychczy)

Dataset: 156 863 observations x 240 features

Some features were recorded in only some of observations…

It was decided to leave only those with at least 85% coverage

Dataset: 156 863 observations x 91 features

So now, lets just apply standard dropping of rows with values missing…

How many are there?

Dataset: 1341 complete cases (0.9% of all observations)

Should we go lower? Then we'll have more complete cases…

But we are losing information in a biased way
(there is something that connects discarded points…)

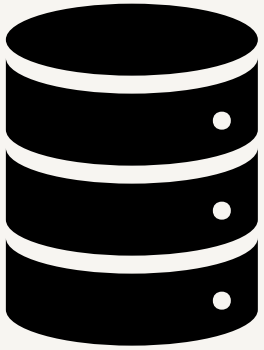
[Example, not refering to any particular real-life problem]



In this (hand-crafted) example, there’s context explanation for every missing data!

Is there a fairy that magically vanishes data points?



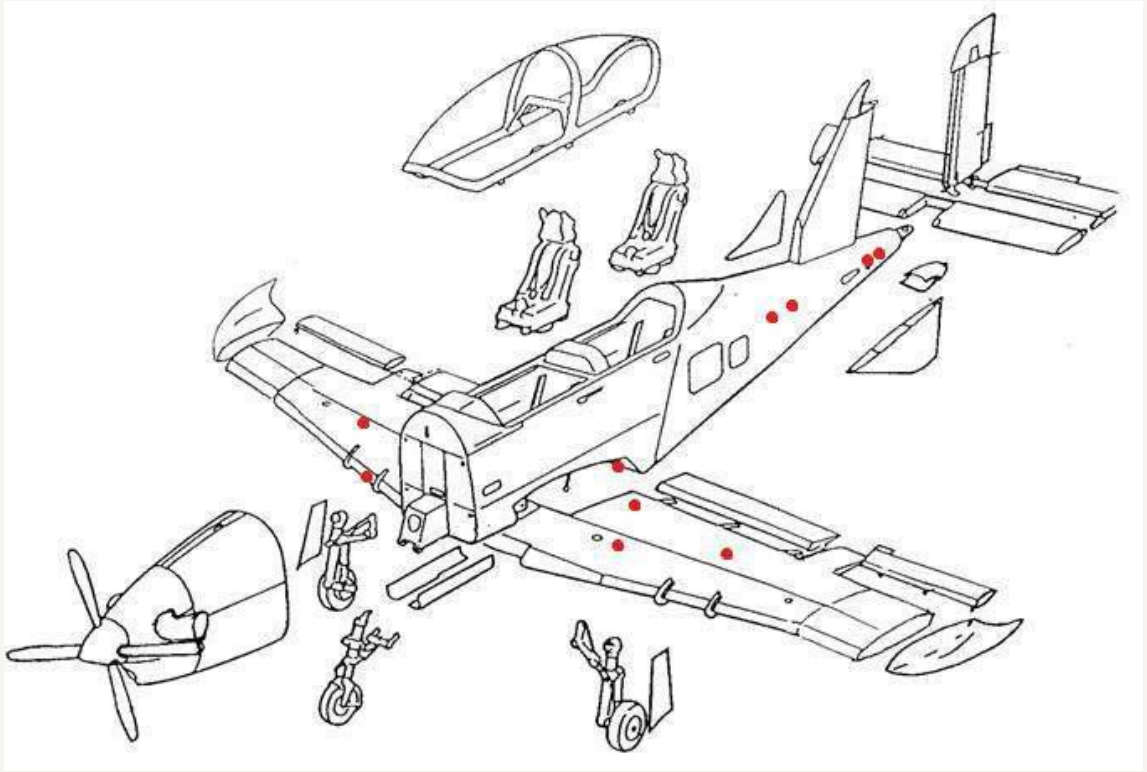


1) Lack of information is also a piece of information...
(Even if only about quality of data acquisition procedures...)

2) Maybe division of dataset into subsets will help?
(If different data are acquired for different operational states...)

Sometimes you wish you had missing data
instead of what you actually have...

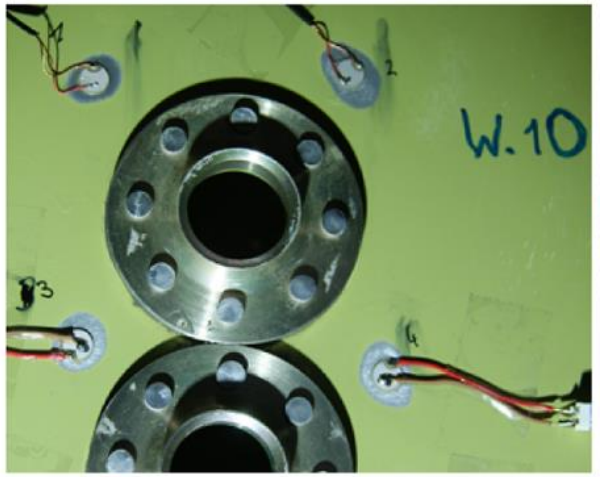
A story about unexpected consequences of tidyness



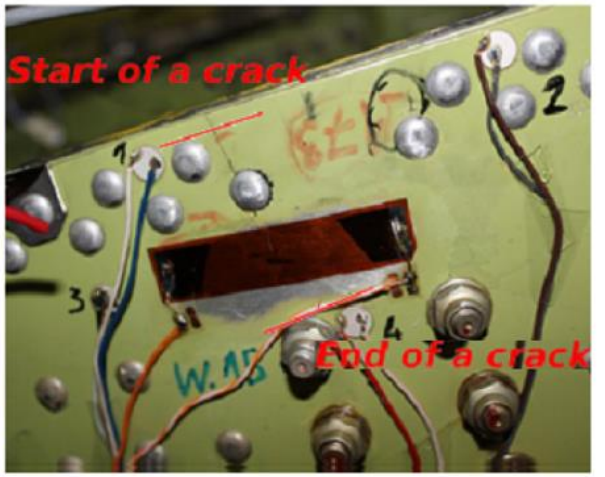
(a)



(b)



(c)



A story about a very tough gearbox

We needed damaged gearbox

So we designed an experiment...

With overloaded gearbox...

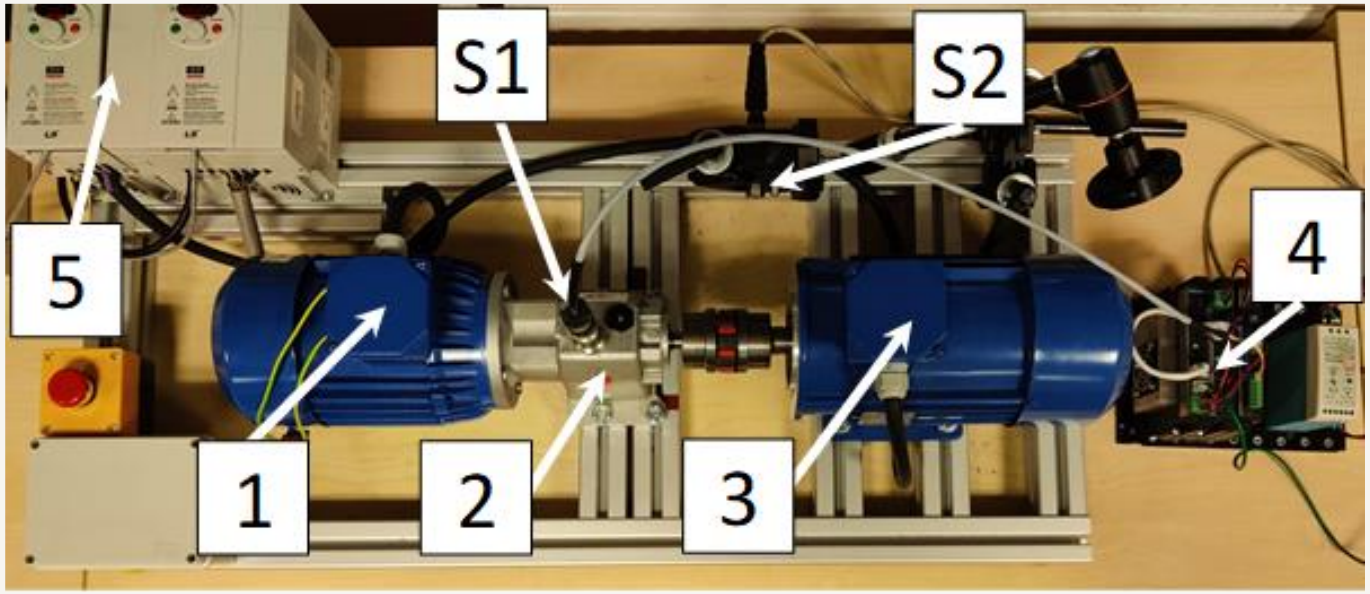
To introduce accelerated wear...

But the gearbox was fine...

So we got rid of lubrication...

But the gearbox was still fine...

For a loooong time...

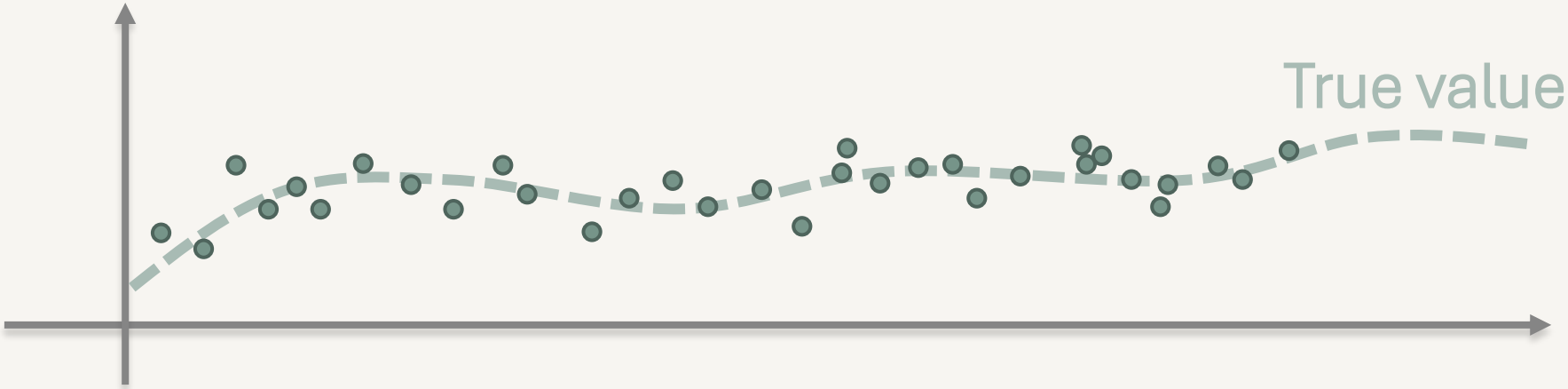


[Image for context, not from the described experiment]

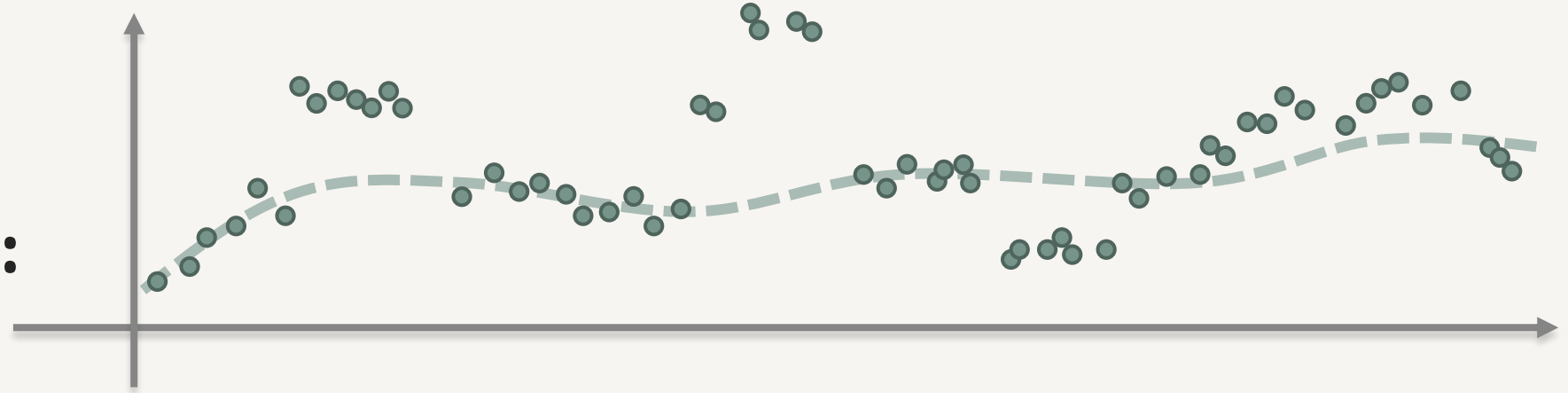
Sometimes you wish you had **relevant** data...

„Measurements affected by external factors”

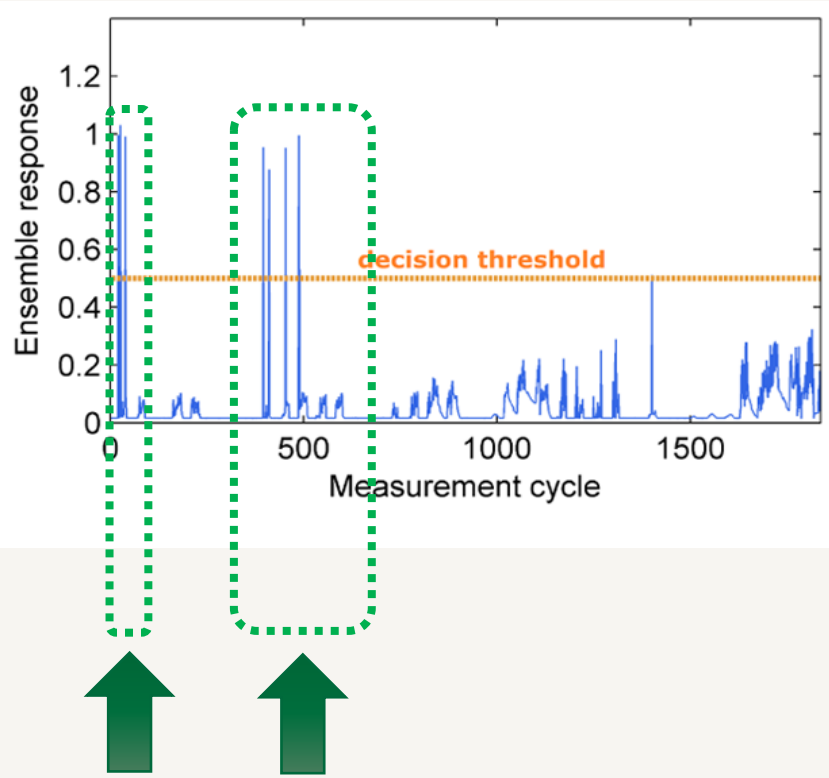
Expectations:

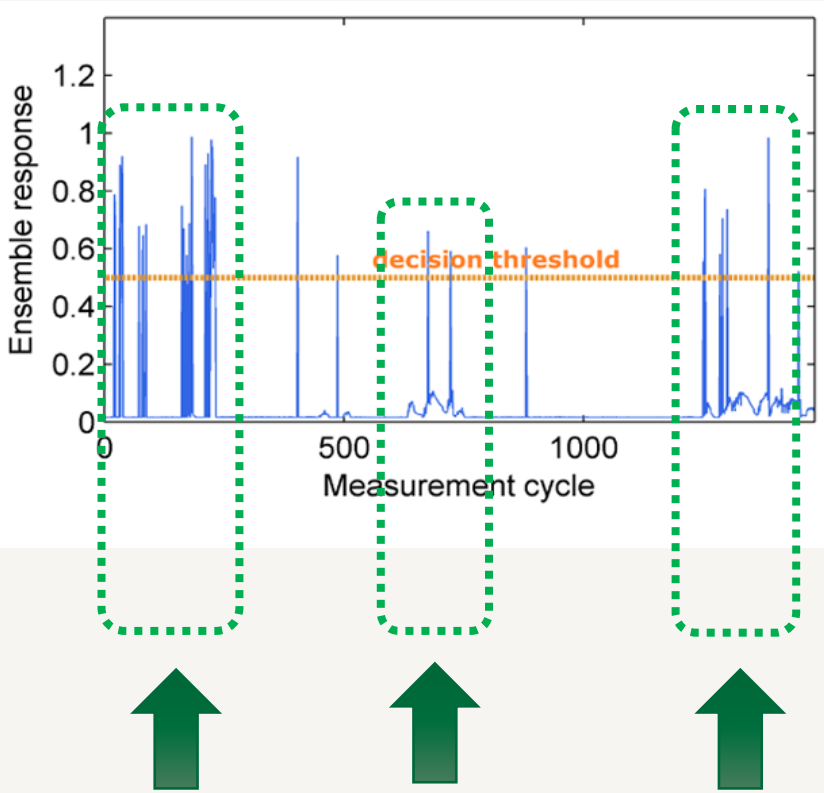


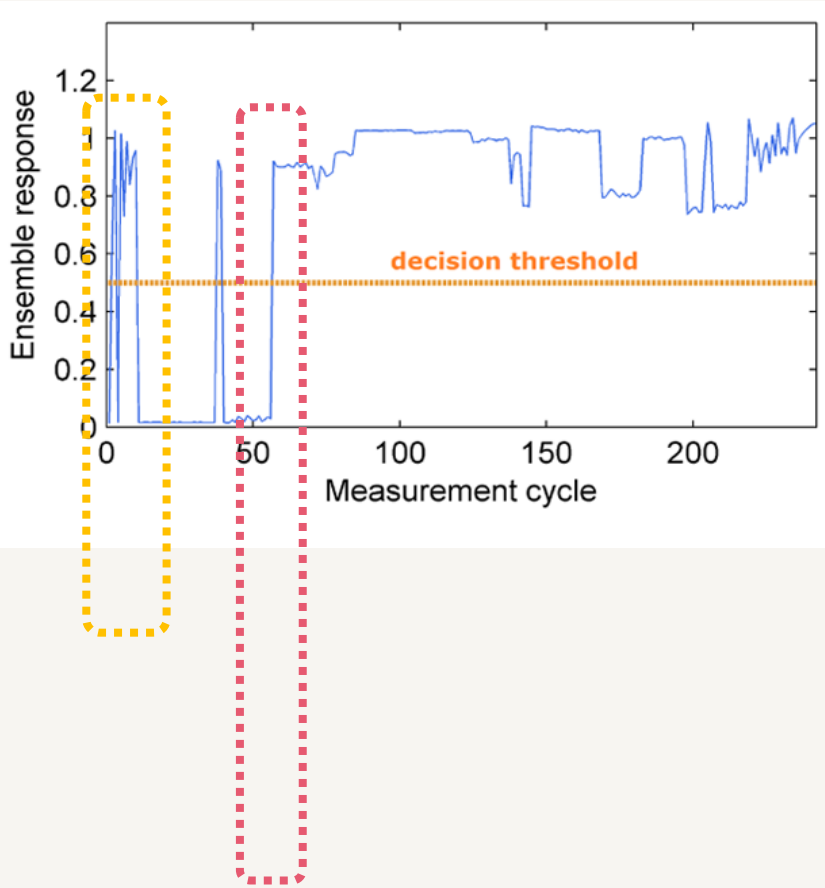
Reality:



Disturbances and noise are often non-stationary!



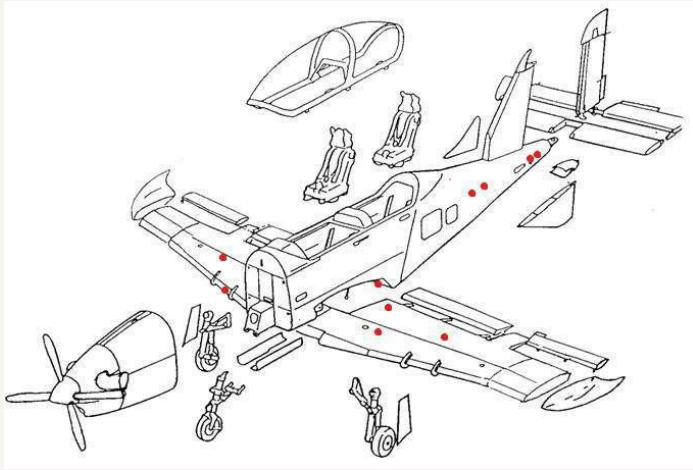


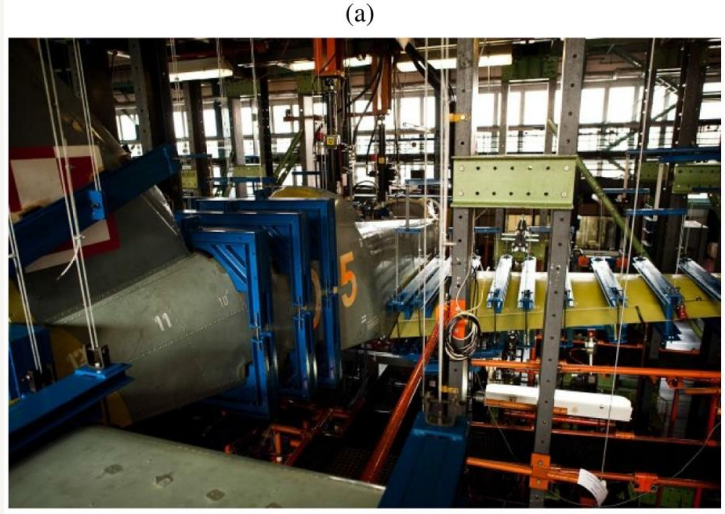


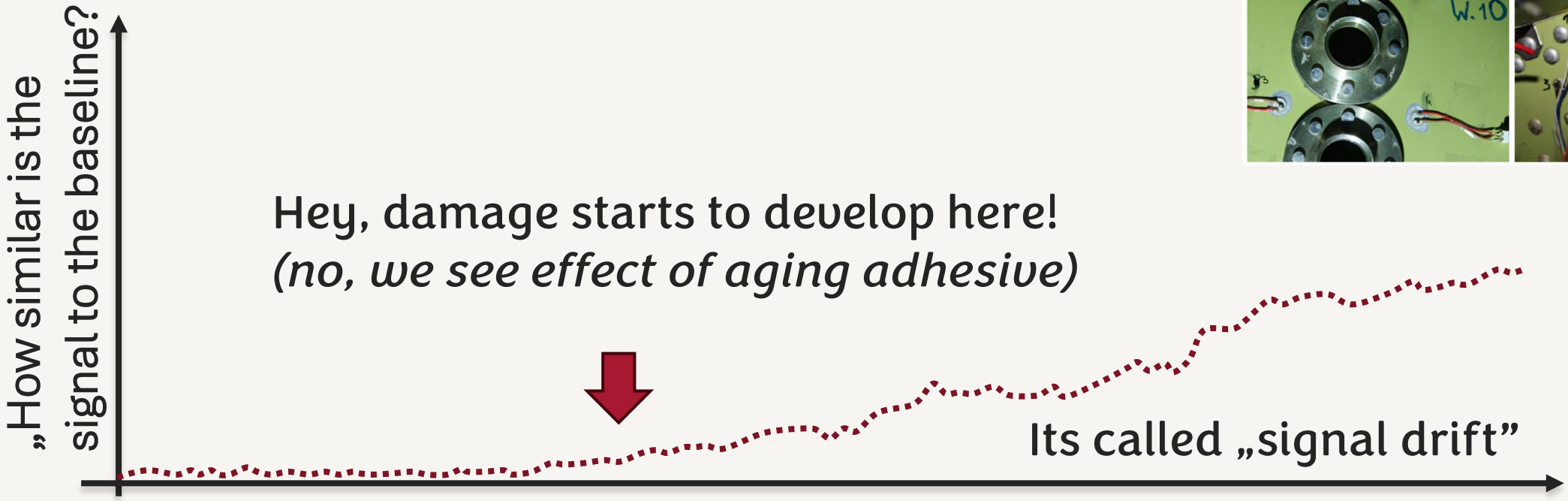
These are just temperature-induced changes in signals, nothing to see here…

That’s damage detected

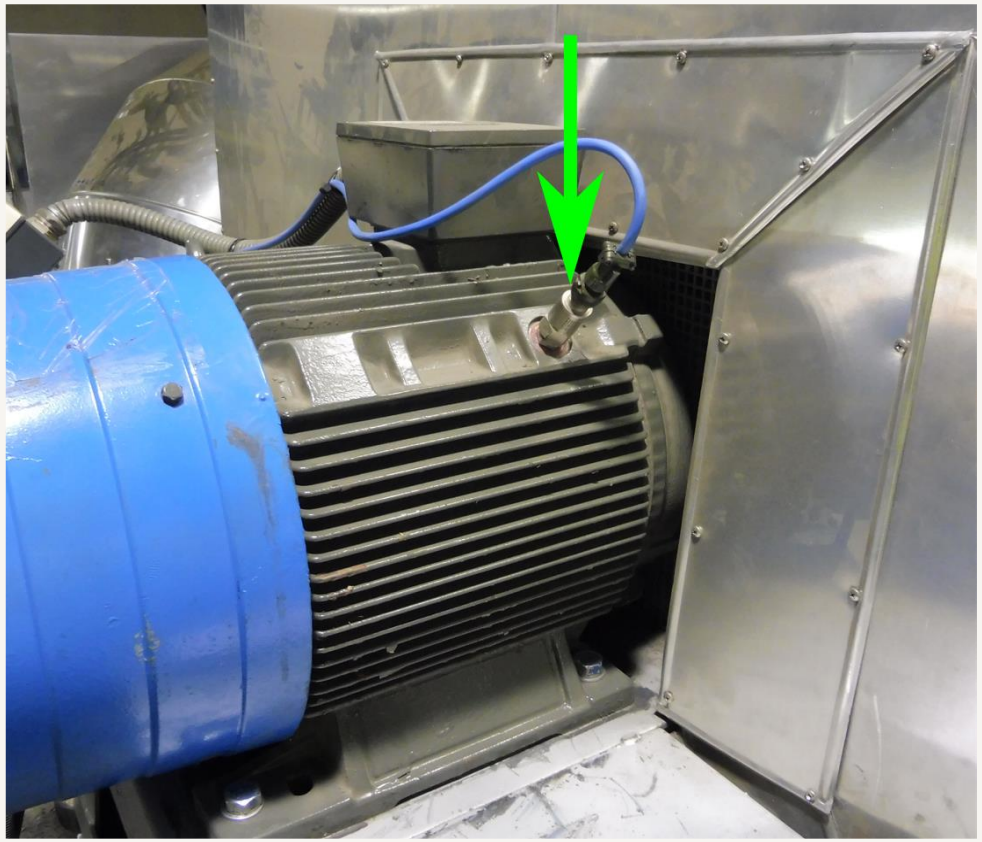
A story about low-grade sensors

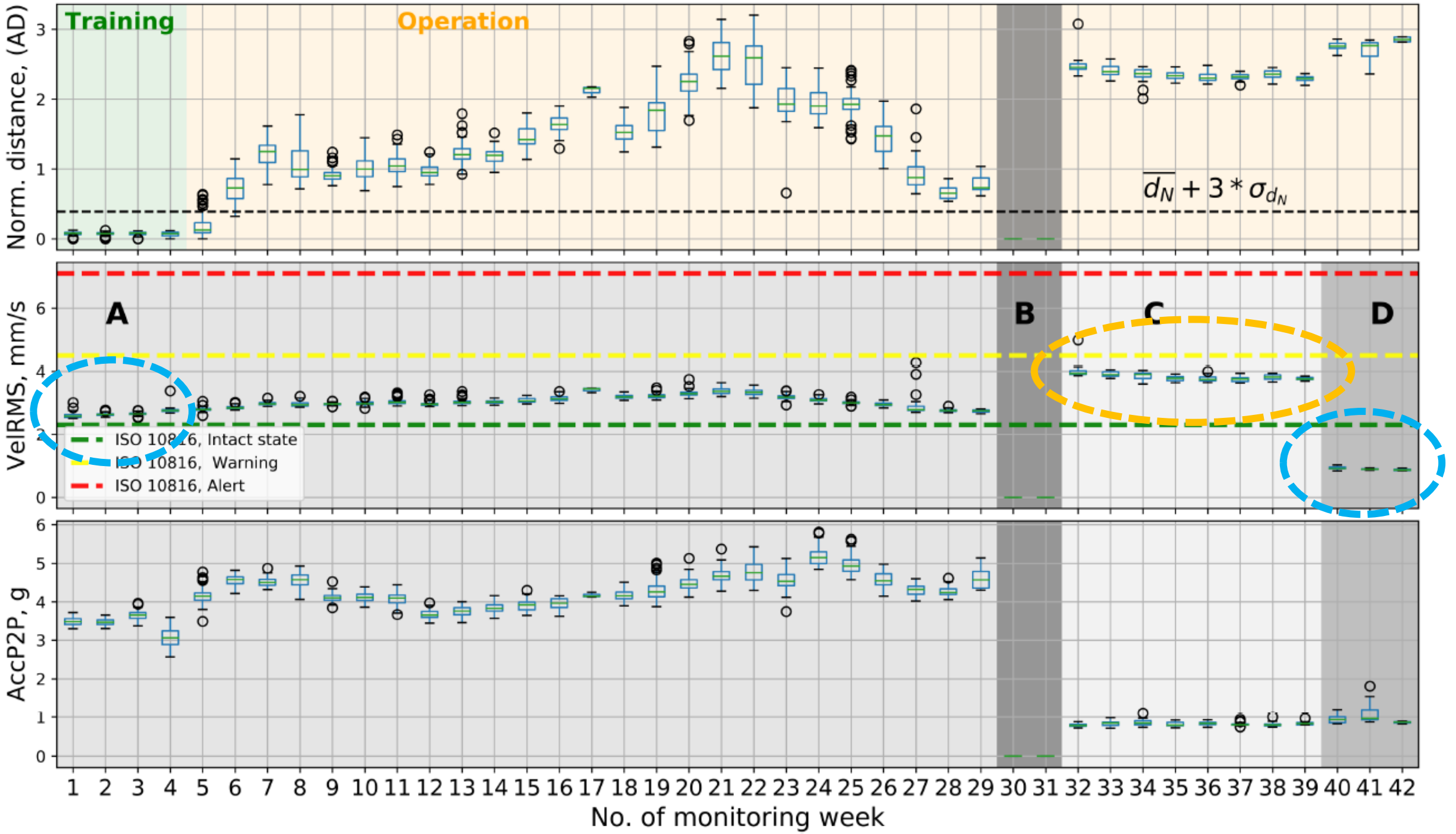




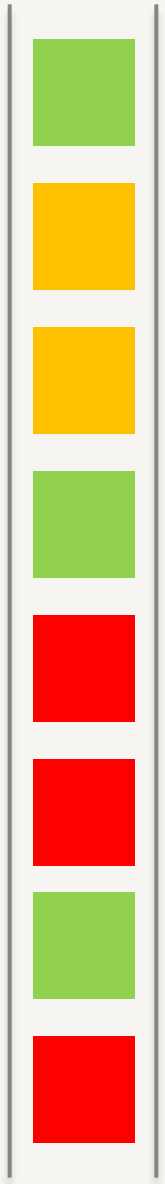


A story about good and even better industrial pumps




Training
Operation
Norm. distance, (AD)
$\overline{d_N} + 3 * \sigma_{d_N}$
A
B
C
D
VelRMS, mm/s
ISO 10816, Intact state
ISO 10816, Warning
ISO 10816, Alert
AccP2P, g
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
No. of monitoring week

Color represents how close are the features to the cluster center



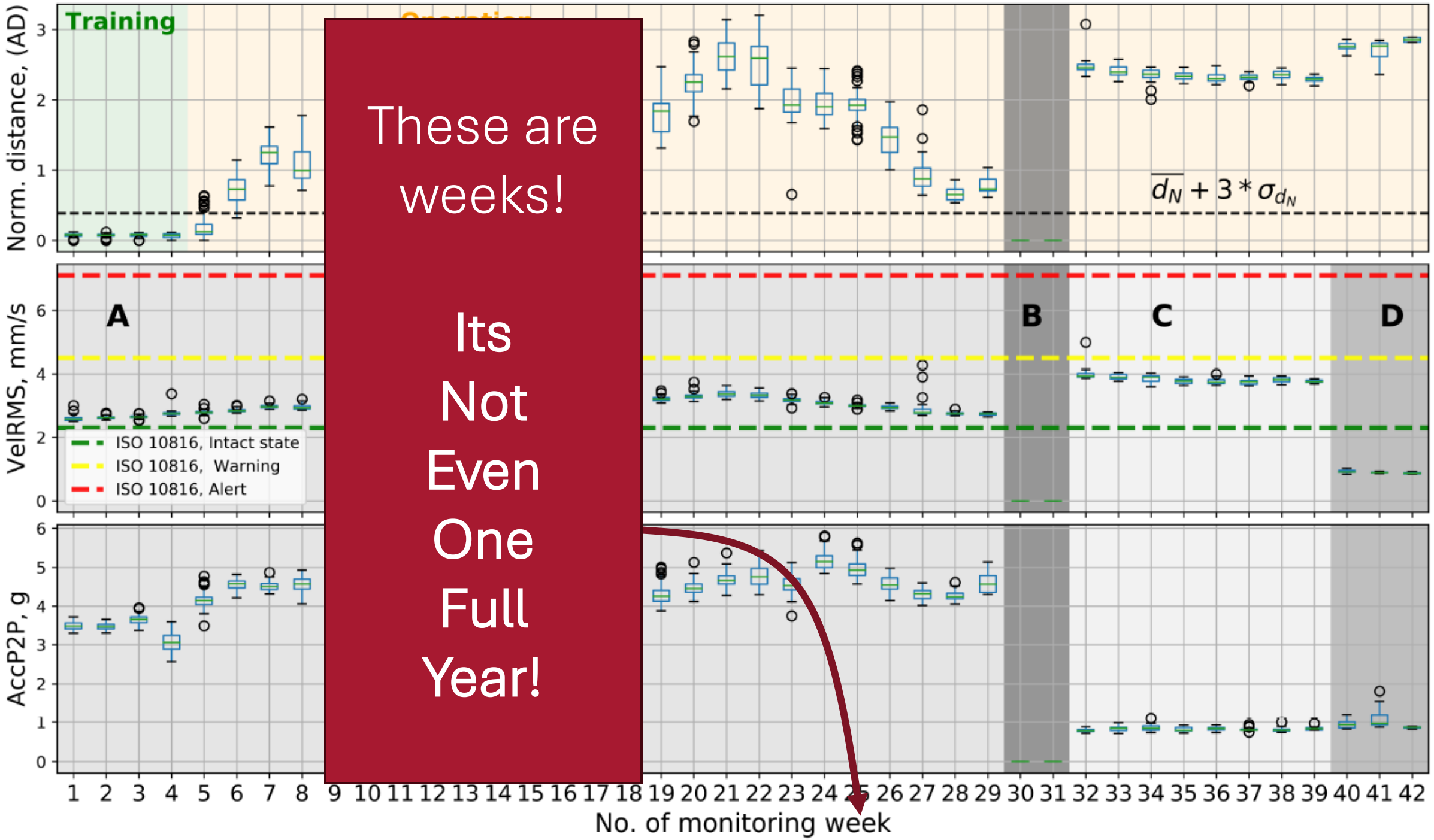
This one is OK, but it stands next to a generator and its vibration is visible in spectrum

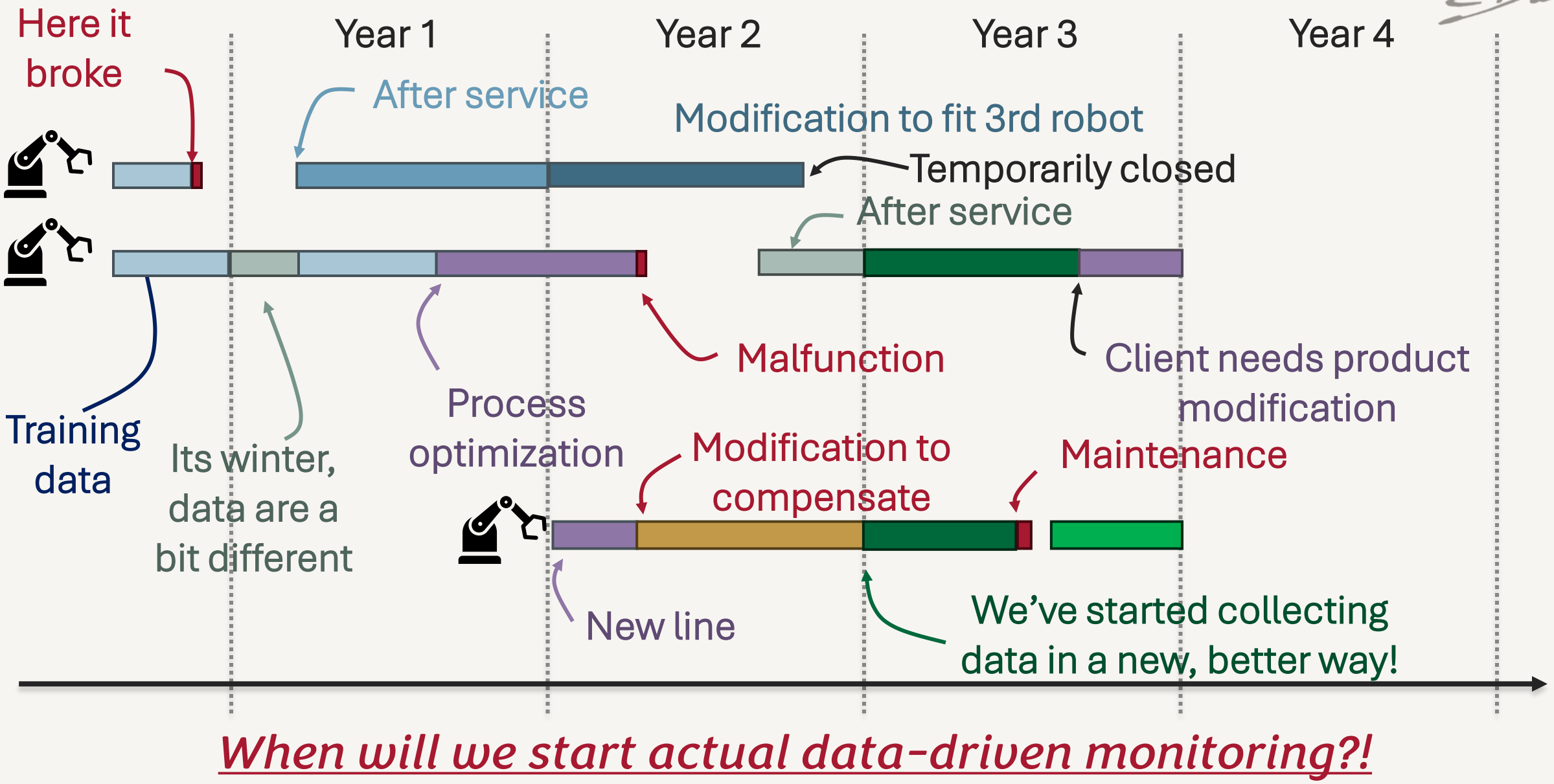
This one starts do develop damage, but will still work for many months, don’t worry

This one is actually damaged

This one is OK, but the sensors for some reason don’t transfer vibration well

This one is brand new – it just looks different in comparison


Training
Norm. distance, (AD)
$\overline{d_N} + 3 * \sigma_{d_N}$
VelRMS, mm/s
A
B
C
D
ISO 10816, Intact state
ISO 10816, Warning
ISO 10816, Alert
AccP2P, g
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
No. of monitoring week
These are weeks!
Its Not Even One Full Year!



When will we start actual data-driven monitoring?!

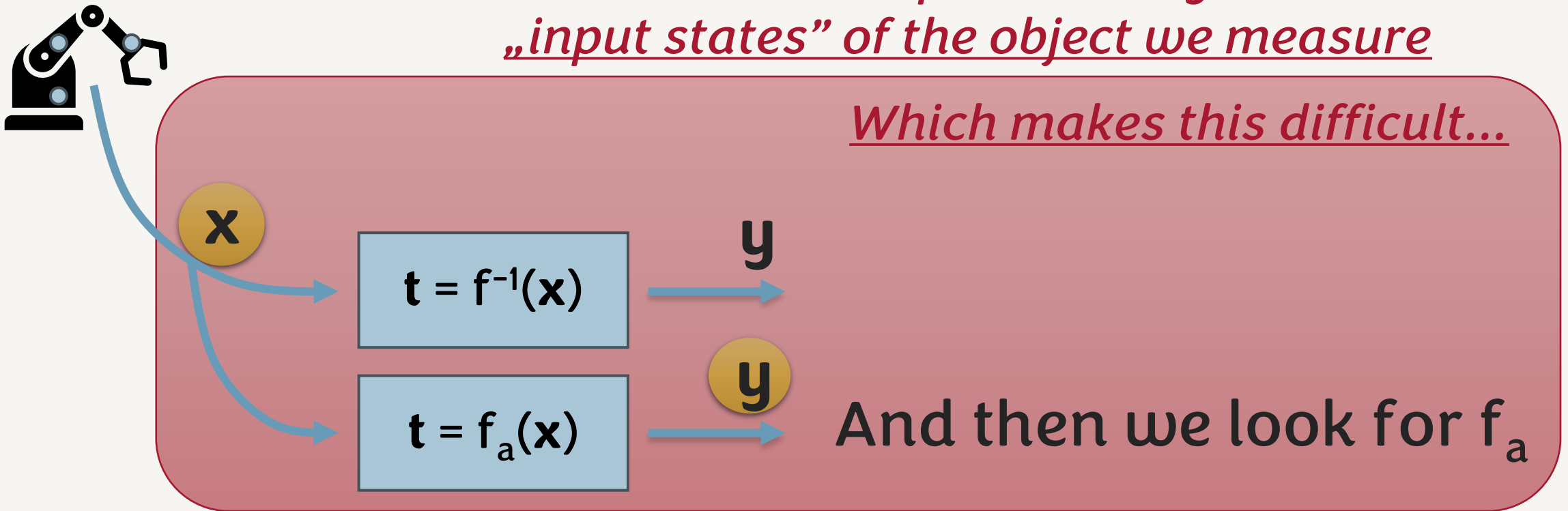
We often assume: $x = f_u(y)$ And hope f_u is bijection

(So features depend on targets only, and in one-to-one way)

In reality: $x = f_u(y,n,e,o,t)$

(So features depend on targets, noise, environment, operation, time and f_u is NOT a bijection)

So: Same features can be produced by different „input states" of the object we measure



To sum up so far…

1. Missing points or outliers can come from changes in operational states
2. People are unpredictable (measurements can and will be messed up by „human factor”)
3. Measurements tend to degrade over time (esp. Important for long experimental programs)
4. Features depend on multiple factors: our targets, but also operational and environmental states, random minute events, non-uniform noise, etc.)
5. We need lots of data (meaning: long acquisition times) but over long time intervals something is bound to change in plant/object operation…
6. Everything is an outlier if you measure hard enough!

And that's still not even half of the potential problems!

We can have unreliable labels/targets (e.g. wrong identification of states)

We can have unreliable or misleading descriptions (e.g. from human operators)

We can have class imbalance and non-representative datasets

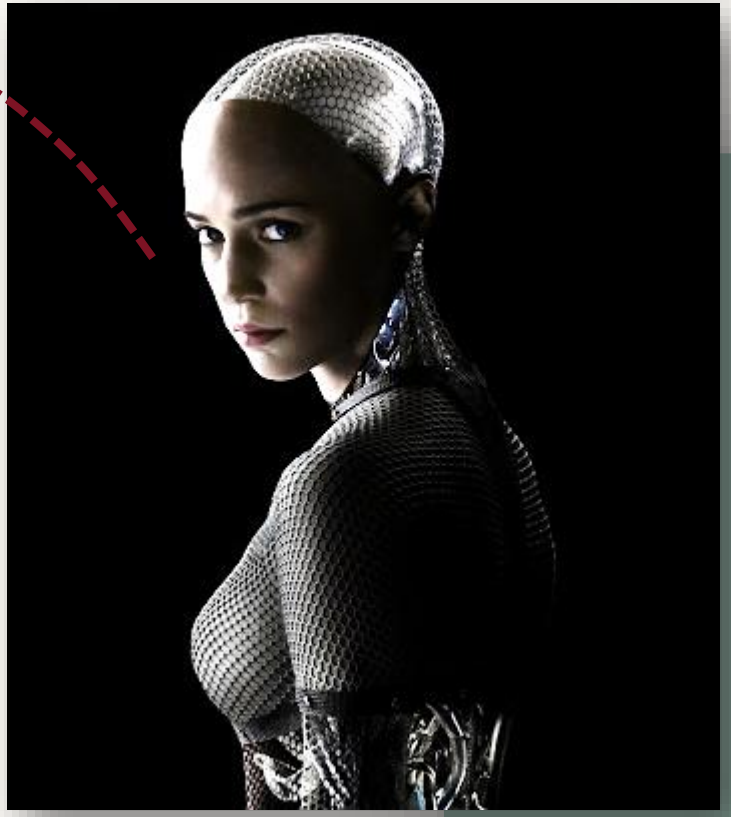
We can have data integration problems (changes in equipment, protocols, sensors, etc.)

Introduction of measurements can alter the process

We can have processing routines that mask target-sensitive features

We can have various forms of data leakage...

That’s for you to decide...



Who will clean this mess?

KATEDRA ROBOTYKI I MECHATRONIKI
WYDZIAŁ INŻYNIERII MECHANICZNEJ I ROBOTYKI

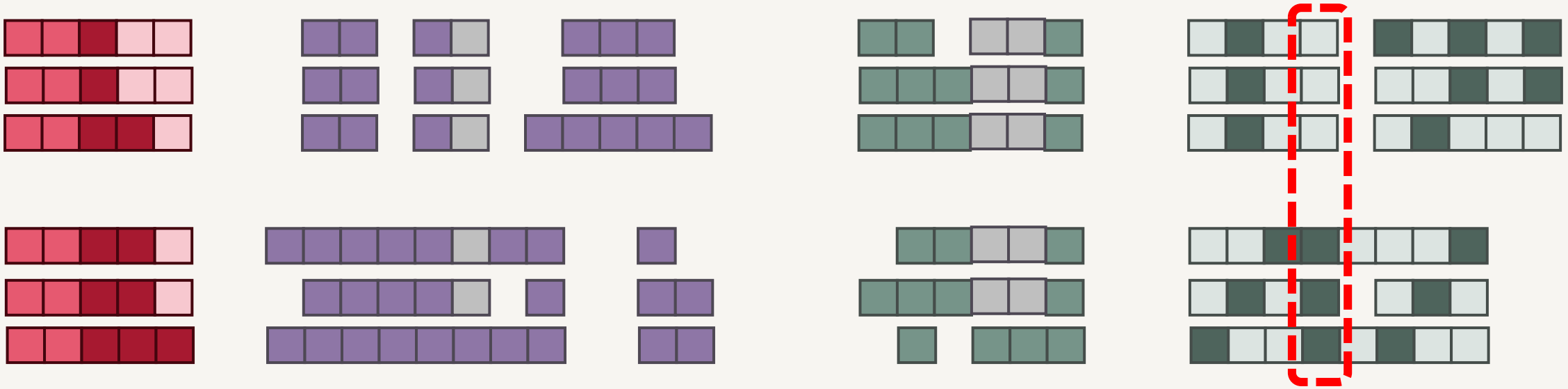
25/10/26

FIX

Concept 1: Manual cleaning of data...

In addition / **as alternative** to „getting rid of outliers" use context knowledge (and variables!) to divide data into subsets





FIX

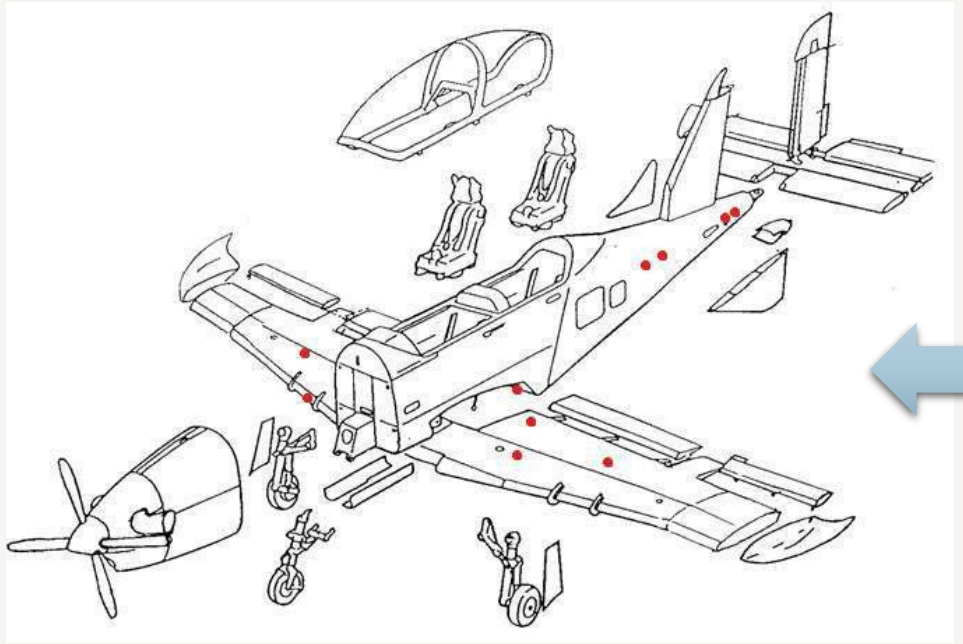
Concept 2: „sanity check”

If context tells you what to expect (e.g. in what ranges the measurements are supposed to be) use that at first layer of decision – to check if measurements are valid

„power generated can’t be negative”

„elevator speedshould not exceed 999 kph”

*„mean of **vibration signal** should be close to 0”*



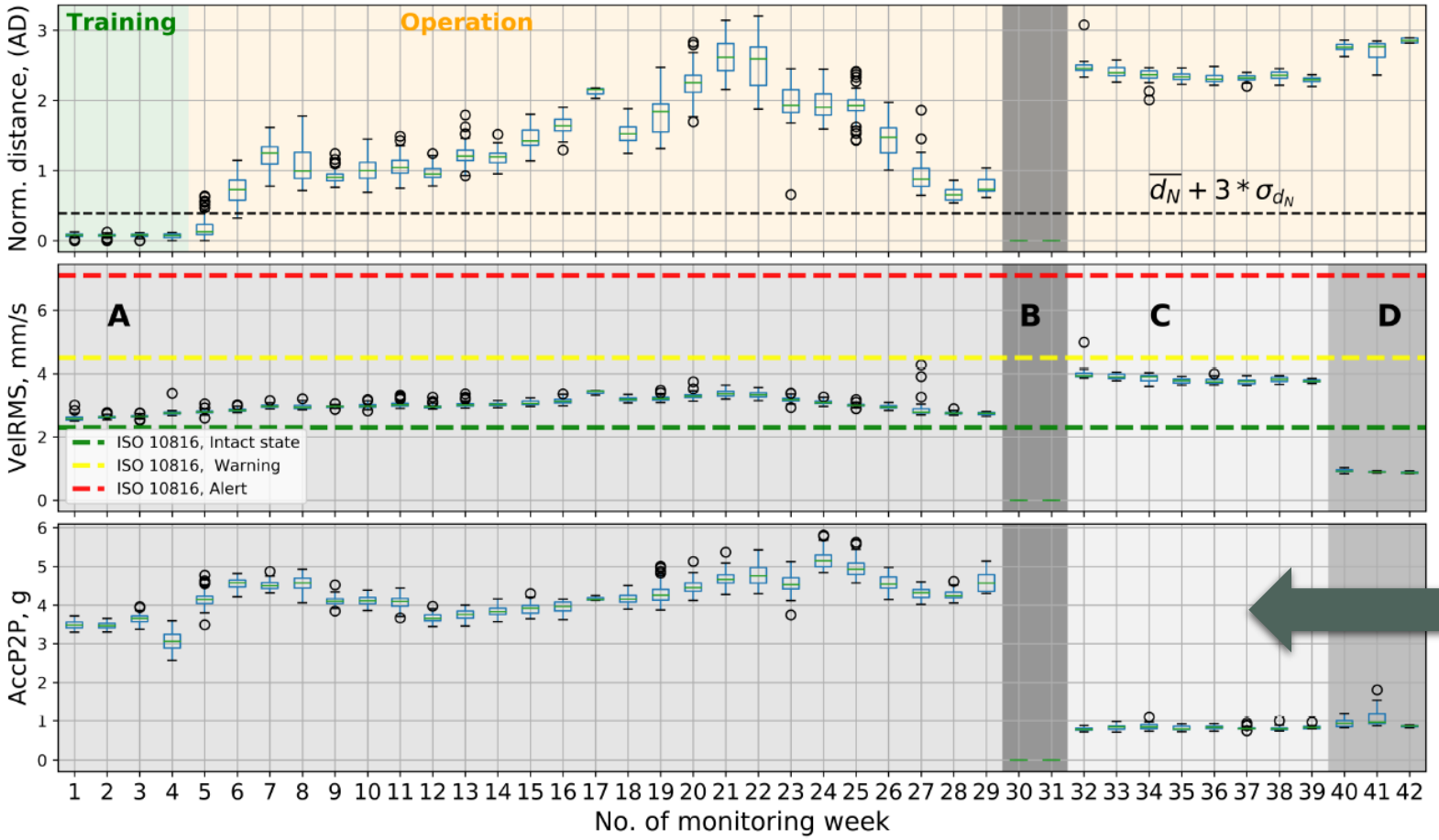
That’s what we did here!

Its unlikely that ALL the measurements suddenly went off!

FIX

Concept 3: „problem detector”

In addition to whatever is your original goal, build state detector to check „how normal is this sample?”

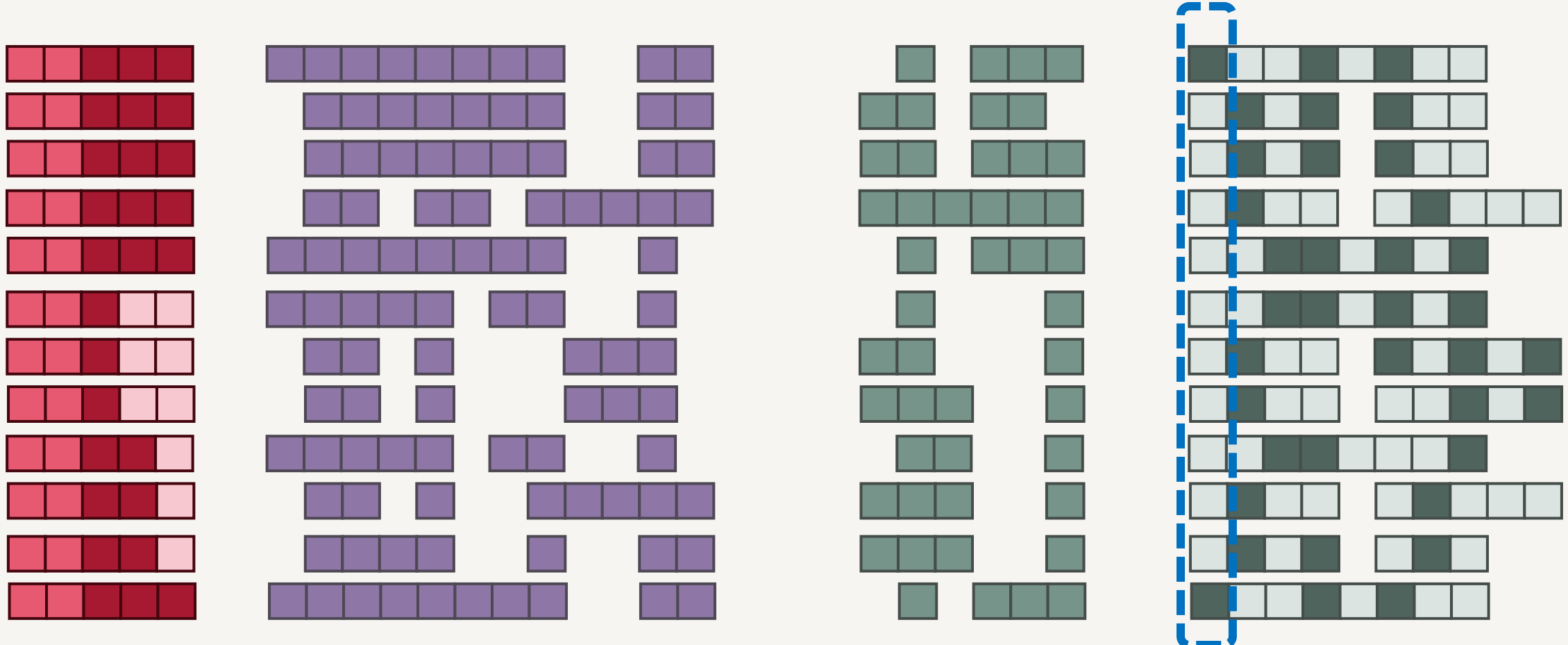


That’s what we did here!

FIX

Concept 4: Introduce decision constraints

Maybe our model can be operational under strict conditions only?
Maybe in such a setup we can actually build a valid decision system?

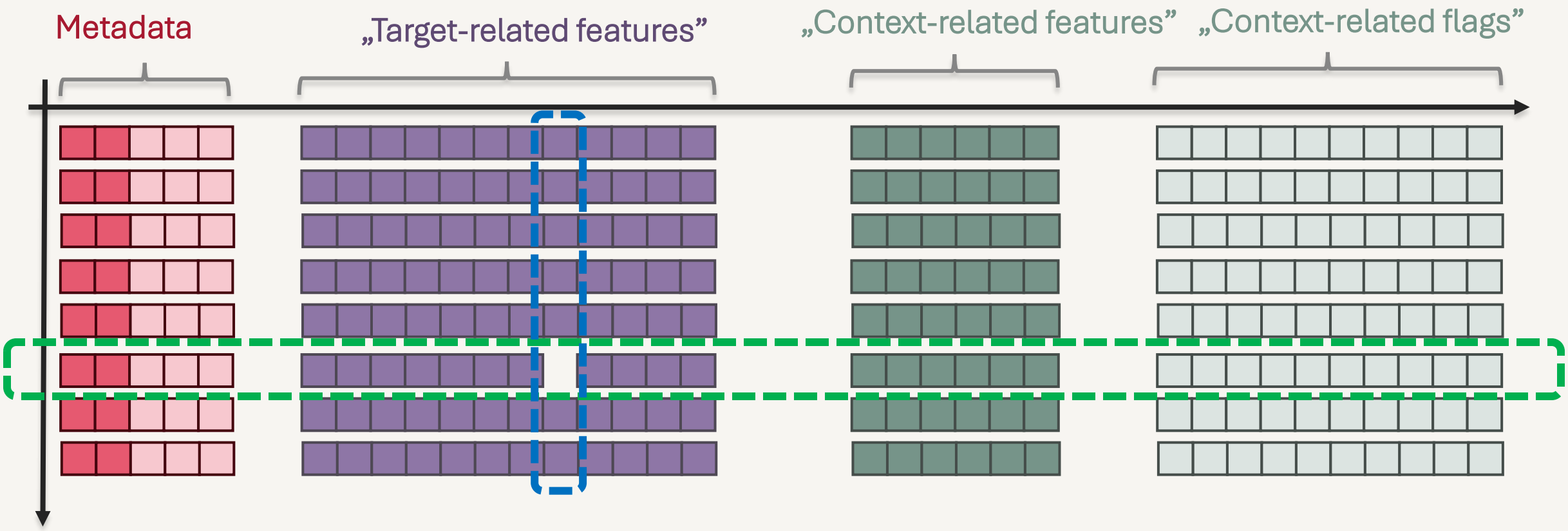


FIX

Concept 5: Fill-in the blanks

Data imputation based on interpolation from previous and next samples or regressed from other features neighbors can be a valid way to deal with missing samples

Why only at 5th position?

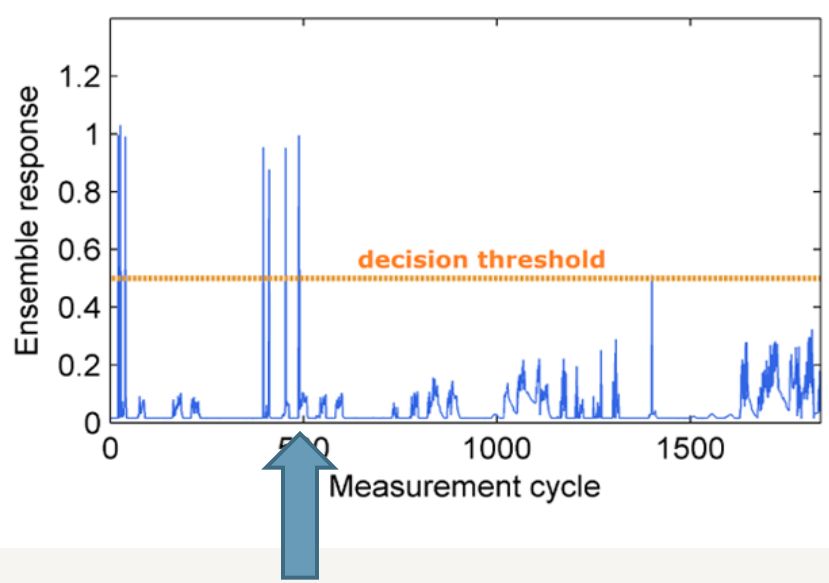


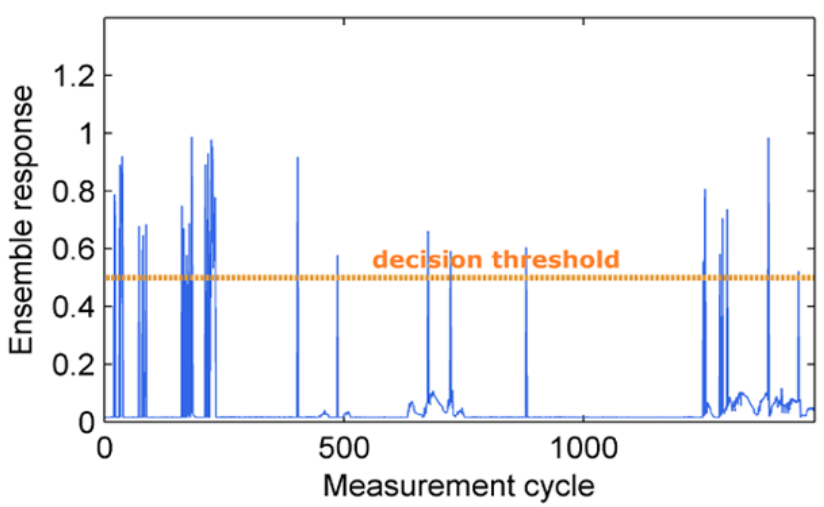
FIX

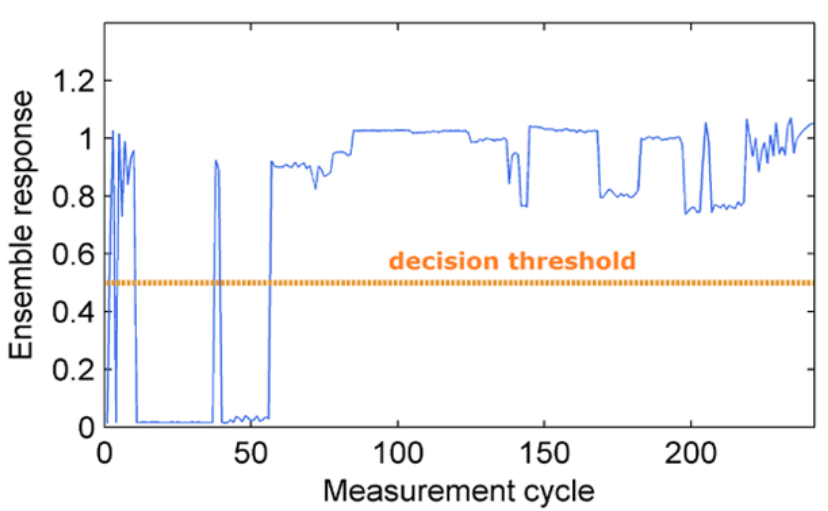
Prevention

Concept 6: Get data from many measurements

If possible, don't hinge your decision on one data sample only – especially if the decision would be different than previous ones... Use **data fusion** to get multiple pieces of evidence (e.g. **bayesian reasoning**)







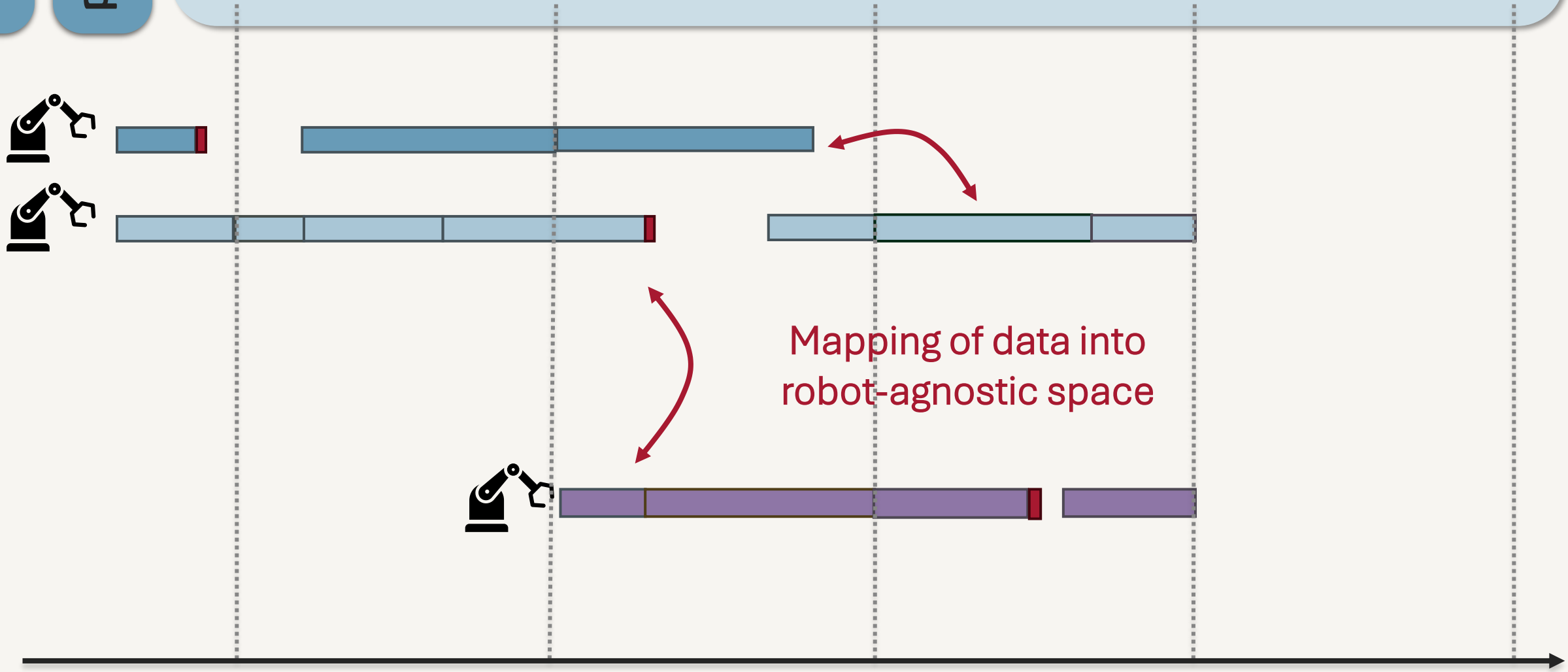
Don't panic, wait for more measurements...

FIX

Prevention

Concept 7: Domain adaptation

Maybe we can learn mapping from one data subset to another and then build large dataset with generalized representation?

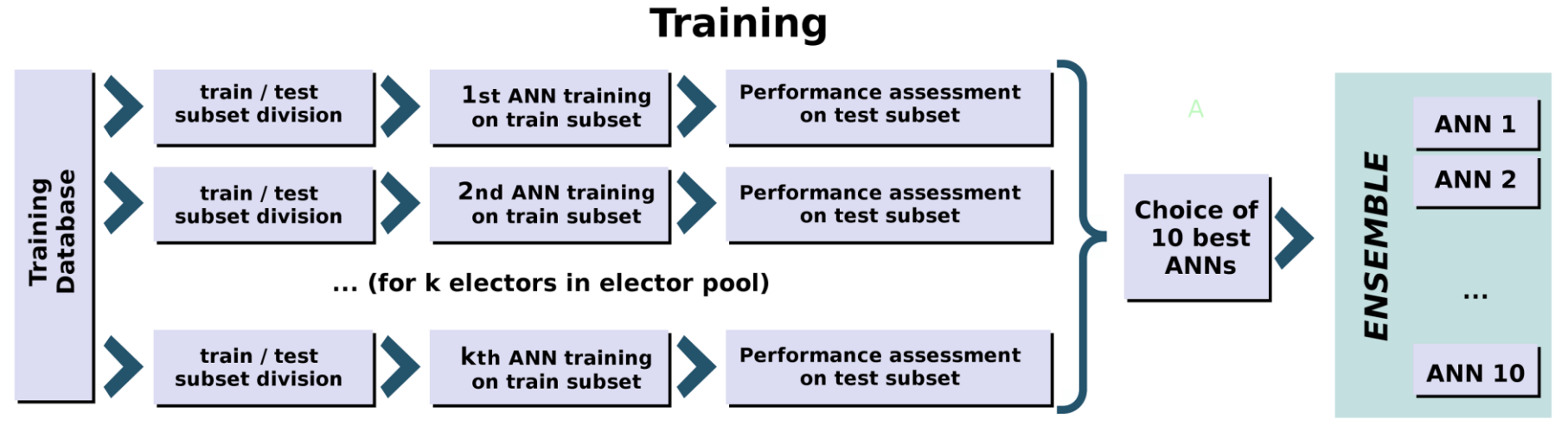


FIX

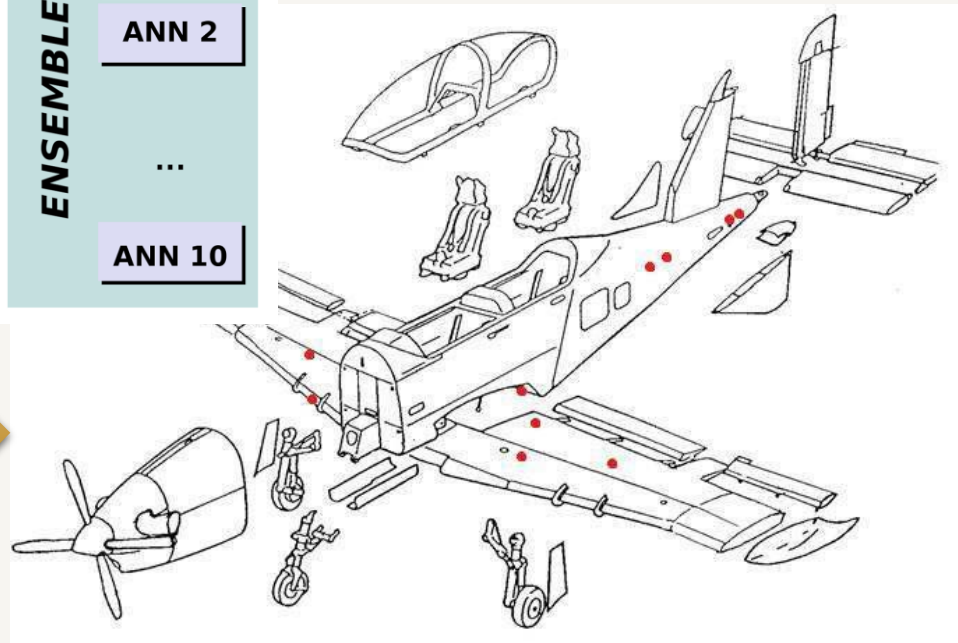
Prevention

Concept 8: Selection of robust models

Ensemble approaches or margin-maximization models (e.g. SVM) can provide reliability if traning and testing distribution do not match exactly



That's what we did here!

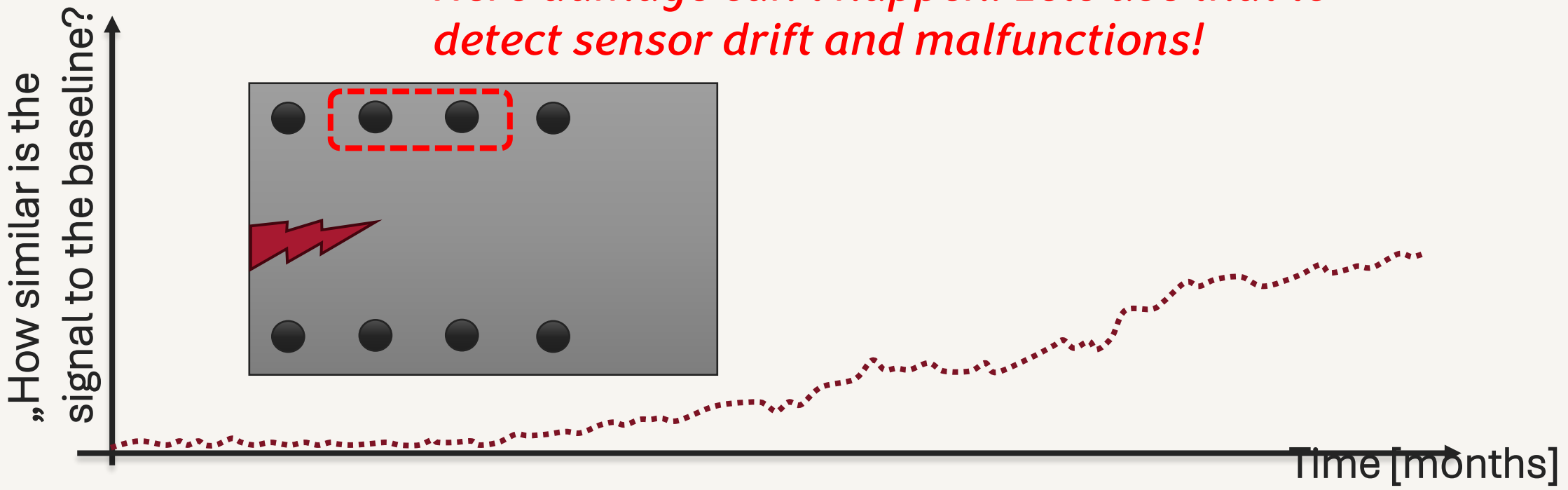


Prevention

Concept 9: self-diagnose as prerequisite

Systems for long-term data acquisition should be periodically testable (preferably in automatic way) – so to check if all the sensors produce valid readings

Here damage can't happen. Lets use that to detect sensor drift and malfunctions!



Prevention

Concept 10: environment & operation monitoring

Even if the goal does not require to know what happens outside
– if possible its better to monitor it as well!
Who knows what may provide additional context for measurements?

