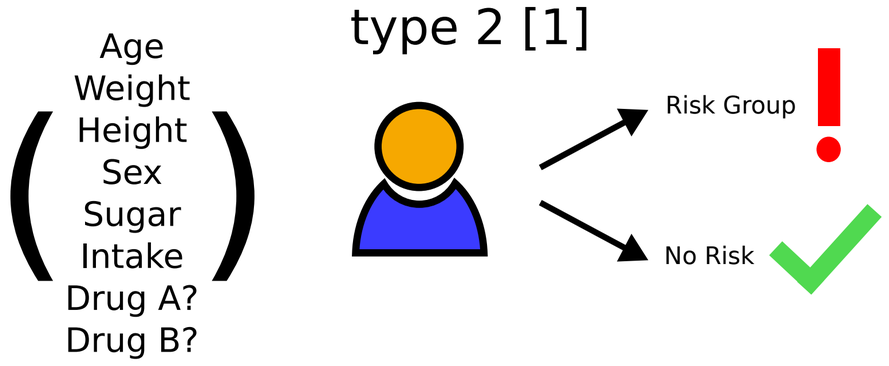
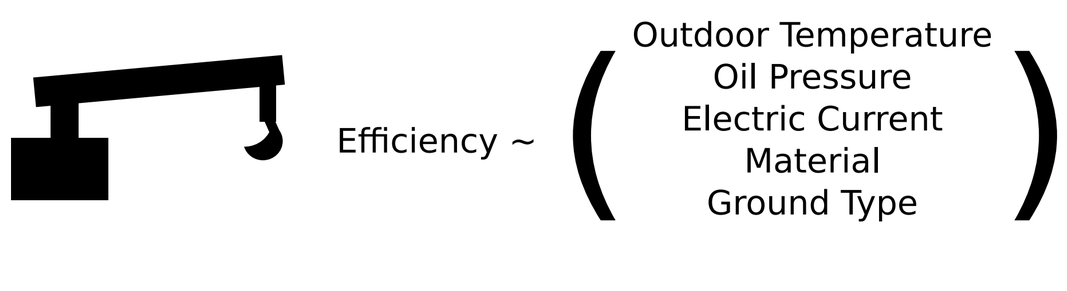
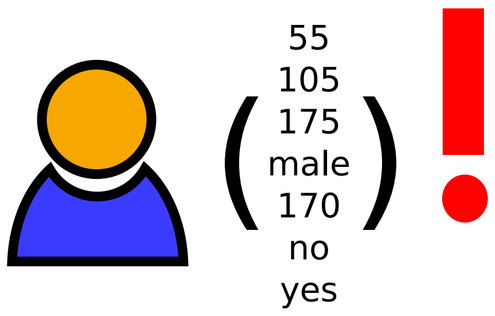
# Discovery of Prediction-Altering Feature Correction (with minimal cost)

**Use Case 1** predict, whether patient is at risk to develop diabetes type 2 [1]

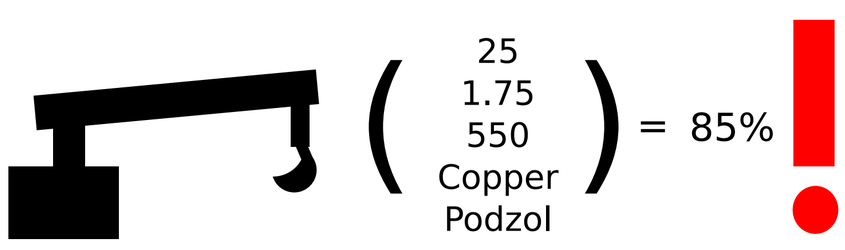



**Use Case 2** predict the efficiency of an excavator; ensure a minimum efficiency of 90% [2]

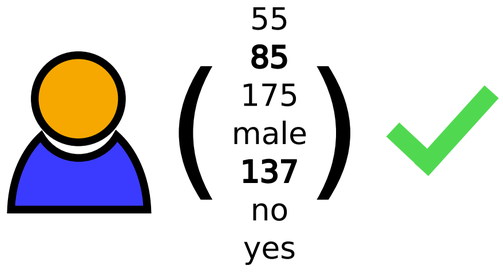



In many machine learning models, we aim to predict a target value. Often, there is a desirable value or range that this value can hold (e.g. high risk in the left example). When we predict an undesirable value, we are interested why it was predicted. For this, explainable models are preferred over, for instance, neural networks. At that, it is as crucial to find out, what we can change to end up with a desired prediction.
We further refer to this problem as feature correction.

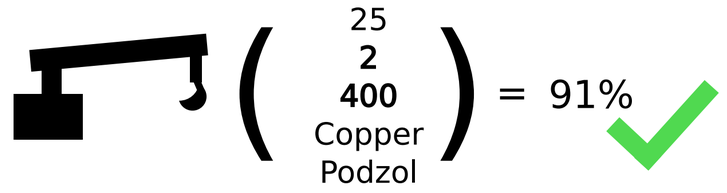



Reduce Weight
Reduce Sugar

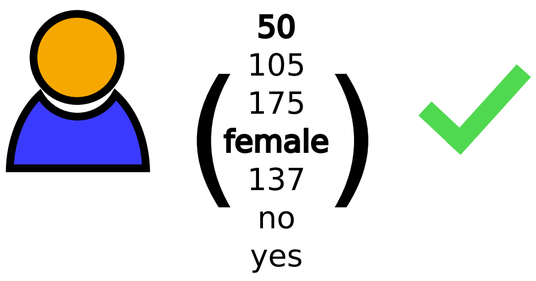



In the examples to the left an right we see an example of feature correction. On the left, we updates weight and sugar intake, on the right we adjusted oil pressure and electric current.
If our model supports gradient update, we can adopt gradient descent to find such a feature correction. We can fix the parameters and propagate the error terms and derivatives back to the inputs.

Increase Oil Pressure
Reduce Electric Current

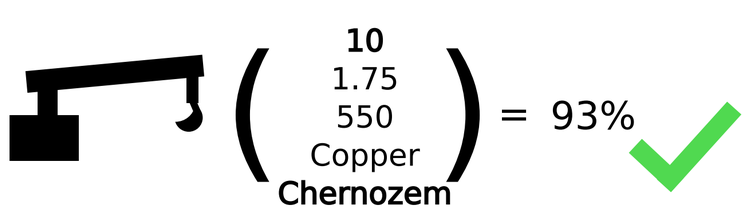



Reduce age
Change gener

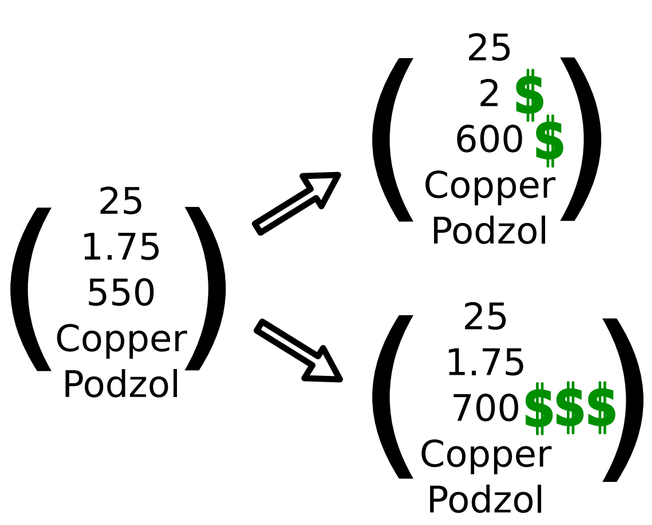



This approach does not solve all problems. For one, it cannot handle discrete values, whose range needs to be relaxed. Second, many models contain features whose value we cannot change (e.g. sex in the left example) or we cannot control (e.g. outdoor temperature) in the right example. Thus, a model incorporate allow additional information, which features may be modified during feature correction.

Change outdoor temperature
Change Soil type



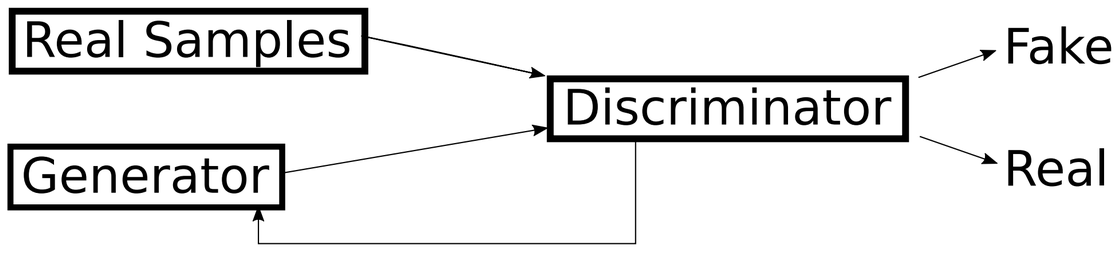

## Extension: Min-cost Feature Correction




An extension of this problem is to associate each feature change with a cost. Clearly, the cost of change heavily varies per feature. This introduces the new problem of "min-cost feature correction". There we want to find the feature correction, which is cheaper than all other feature corrections.

## Related Work

Generative Adversarial Networks[3]




Adverarial Patches[4]

Create a small patch, that can be added to any image. This patch causes a classifier to output a pre-defined class


[1] Vincent Ait-Ammar, Tobias Wieschnowsky: *Case Study: Data Mining & Predictive Maintenance for Energy efficient coal mining*
[2] Christoph Böhm, Alexander Albrecht: *Plugging Data Science into Big Data Processing Workflows*

[3] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[4] Brown, Tom B., et al. "Adversarial patch." arXiv



## Abstract

In machine learning settings, such as disease predictions, there are desirable predictions (healthy) compared to unfavorable ones (high risk, sick). Given an unfavourable outcome, we want to modify the inputs(features) to avert this prediction. To this end, we propose the problem of feature correction. A feature correction between two predictions is a change in features, which result in the model to switch from the first to the second prediction. To account for different financial cost of changing features, we also suggest the min-cost feature correction problem .



Lukas Faber (Master)
lukas.faber@student.hpi.uni-potsdam.de


Discovery of Prediction-Altering Feature Correction (with minimal cost)