In 2019, a Machine Learning (ML) model was created that derived the NOC with an accuracy of roughly 82% [1]. This Random Forest (RF) model was trained based on 590 profiles of 2-5 person mixtures, obtained from **TODO:find how many**  donors. The data used for training was not the original electrophoresis results, but consists of 19 features such as the MAC, locus-specific information, and other statistical features of the data.

**Decision Tree [2]**

Derive NOC from

* Decision tree with

Tested various ML approached (RF / MLP / LDA), showing similar performance to the RF19 model. They obtained very high performance (96%) with a RFC 35 model.

Difference with [2] is “Benschop et al. used 1174 unique donors to construct 590 profiles [20], whereas the PROVEDIt dataset only had 26 unique donors within the 766 profiles used”

This means that the classifiers probably overfit to certain donors.

“In conclusion, the decision tree method for NoC assignment has been shown to be over 77% accurate, with increasing performance with improved stutter and artefact filters”

They used a decision tree to classify peaks as stutter or allele.

**Background STR mixture interpretation [3]**

Information about how statistical analysis is done to determine the LR with the Hd and Hp. Showing that the LR is still the de-facto standard method.

“The peak height information is of benefit for analyzing mixed profiles.”

“The effect of incorrect estimation of the number of donors (caused by allele sharing) to the LR value was examined by Benschop (…) and was illustrated to exert a great effect on the LR” [4]

**Background about NFI-used software for LR calculation DNAStatistX**

Shows the importance of correct NOC estimations: under-assigned number of contributors can cause the model the fail calculating the LR because the observed peaks cannot be well explained.

Also includes the NOC model + the generic RF11 model (with a lower accuracy of ~

Also includes the LoCIM method for inferring the major contributor.

**“Short, disjoint rules are easier to interpret than hierarchies like decision lists or trees (Lakkaraju, Bach, and Leskovec 2016)”**

# Evaluation

Explanations should have few features, as humans pick usually just a few reasons. They should be specific to the problem at hand, and every instance should be explained in the same deterministic way [5]. Deterministic, or consistent feature attributions [6].

Exploration using several visual aids [7].

MISC COUNTERFACTUAL METHODS  
Linear optimization [8], learning a local decision tree [9],

Using the gradient of the loss with respect to the input [10], a method that is based on the data, not on the classifier

However, their method is based on highly-dimensional (1000+), behavioural or textual data [11]. Setting a value to zero is not viable in our dataset since this does not correspond to a realistic feature value.

FACE makes use of the training data to form a connected path from one instance to one of the target predictions. This relies heavily on the data having certain areas of higher density, and having connected areas from one prediction to the next. This might not be present in sparser real-life datasets. Their focus lies on if the transition is feasible, not if the end result is. There is not constraint for correlated features.

### Example realism score with NOC estimation

As an example, consider a profile with a Total Allele Count (TAC) of 98 and a Maximum Allele Count (MAC) of 6, that was predicted to have 4 contributors. To generate a counterfactual with a prediction of 3 contributors, the program might propose a profile with a TAC of 30. Though this would make the model predict a NOC of 3, the combination of the original MAC value of 6, with the new TAC value of 30 is impossible. There are a total of 23 loci in the profile; if there are 30 alleles in the entire profile (TAC), that would leave either 1 or 2 alleles per locus (30/23). It would thus be highly unlikely, or even impossible to have a locus with 6 alleles (MAC).

Which model-agnostic counterfactual explanation techniques exist and what problems can they be applied to?

[12] Wachter

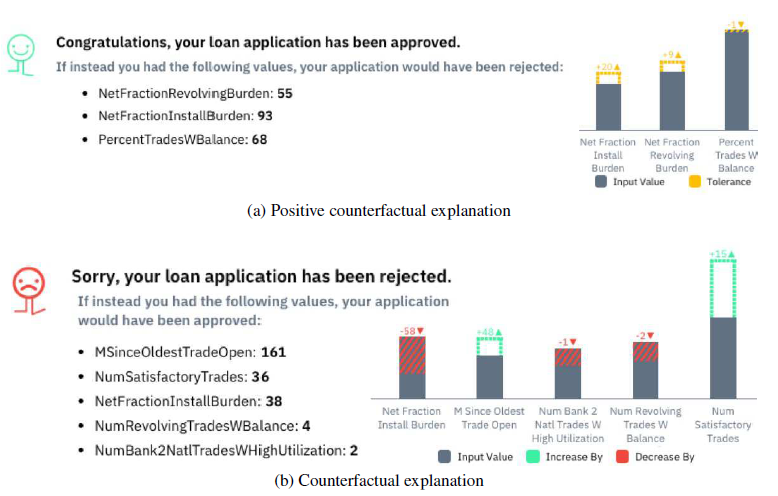
Two reasons that the authors mention which are also relevant to the current case:

1. Understand why a decision was made
2. Understand how a different decision could be made if certain conditions were changed

“Principally, counterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems”

[13] Grath.

They balance the prediction target with the distance measure using a lambda parameter which is optimized using a tolerated mismatch between target and actual prediction.

****

[9] factual and counterfactual explanations: building a simple classifier around a local neighborhood of the point of interest. This local neighborhood is created with a genetic algorithm, after which this feature space is described using a decision tree. The path down the tree that corresponds to the current label thus explains the actual prediction, while any path that leads to a different outcome could be interpreted as a counterfactual. They measure minimal changes to be the number of split conditions that need to be changed in order to change the prediction.

Evaluated based on correct classifications of the instances with the decision tree mimic model as compared to the original black-box. Evaluated on various (binary) classification datasets, mostly obtaining outputs with higher fidelity than LIME, even for the local neighborhood.

“In summary, ANCHOR shows better precision than LORE at the expenses of generality and stability of the produced explanations.”

[10] CADEX

the gradient of the loss with respect to the input, which can be followed over the input space with an optimizer (e.g. Adam) until we find an instance x\*, that typically lies on the decision boundary between y^ and y\*.

Code: <https://github.com/spore1/cadex>

[14] Multi-Objective Counterfactuals (MOC) method.

Only evaluated on the 4 objectives posed, in comparison to other state-of-the-art methods (dice, recourse, whatif, tweaking).

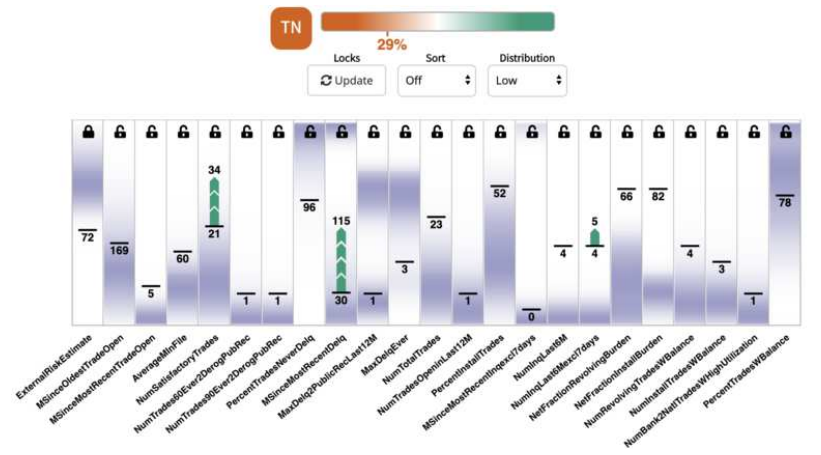
**As for counterfactuals; they can be applied to any model and data type. Some counterfactuals might not be feasible. It is therefore important to use training data and constraints to guide the process.**

**They recommend to determine clear desiderata:**

* **Identify the target for which the explanation is intended.**
* **Understand what they are looking for.**
* **Understand how they will engage with the explanation.**

[15]: Feasible and Actionable Counterfactual Explanations (FACE). Their only focus for evaluation is to visualize the path taken by their algorithm. They do show that the counterfactuals by Wachter et al. (left picture) do not fall close to the data distribution.

[16]: ViCE: visual counterfactual explanations



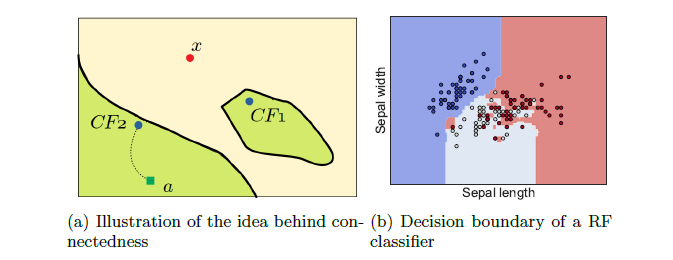
[17] DiCE. Code: <https://github.com/interpretml/DiCE/tree/master/dice_ml>

[18]. Based on the training data. as these counterfactuals are “real experiences” from the problem domain, they have an inherent plausibility”.

[19] MACE: <https://github.com/amirhk/mace>

[20] GeCo: Compares to MACE, CERTIFAI, WIT (what if tool)

A study on justification of counterfactuals [21]. “*Given a classifier f : X ! Y trained on a dataset X, a counterfactual example e <in> X is justified by an instance a <in> X correctly predicted if f(e) = f(a) and if there exists a continuous path h between e and a such that no decision boundary of f is met. Formally, e is justified by a <in> X if: <there is an> h : [0; 1] ! X such that: (i) h is continuous, (ii) h(0) = a, (iii) h(1) = e and (iv) <all>t <in> [0; 1]; f(h(t)) = f(e).”*



They argue that CF2 is better than CF1; in B you can see an application for it (overfit classifier). “CF2 can be connected to a ground-truth instance a without crossing the decision boundary of f and is therefore justified.”

**So instead of choosing instances from the training data per se; they make sure that a counterfactual instance can be “connected” to a ground truth data point without crossing the decision boundary.**

They have written software to asses if counterfactuals are connected to ground-truth data. They sample a circle around an instance, and cluster them using DBSCAN. If they belong to the same cluster as the ground-truth instance, they are justified.

Shows that LORE often creates justified explanations.

<https://github.com/thibaultlaugel/truce>

Desiderata [22]:

* Validity: meaning it has the desired class (originally by wachter et al)
* Proximity
* Actionability
* Sparsity
* Data manifold closeness
* Causality “a counterfactual should maintain any known causal relations between features”
* Diversity

Challenges:

* The evaluation for counterfactual explanations must be done using a user study.
* Counterfactual explanations should be integrated with visualization features.
* Generate robust counterfactual explanations (not overfit to model).
* Unify counterfactuals with traditional XAI

## How to weigh the total closeness scores

* Weigh features according to feature importance
* Weigh distance and features changed together
* Weigh using hyperparameters ([20] and MACE)

1. Benschop, C.C.G., et al., *Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach.* Forensic Science International: Genetics, 2019. **43**: p. 102150.

2. Kruijver, M., et al., *Estimating the number of contributors to a DNA profile using decision trees.* Forensic Science International: Genetics.

3. Tao, R., et al., *Separation/extraction, detection, and interpretation of DNA mixtures in forensic science (review).* International Journal of Legal Medicine, 2018. **132**.

4. Benschop, C.C.G., et al., *The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures.* Forensic Science International: Genetics, 2015. **19**: p. 92-99.

5. Lakkaraju, H., et al., *Interpretable & Explorable Approximations of Black Box Models.* 2017.

6. Lundberg, S. and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*. 2017.

7. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.

8. Russell, C. *Efficient search for diverse coherent explanations*. 2019.

9. Guidotti, R., et al., *Factual and Counterfactual Explanations for Black Box Decision Making.* IEEE Intelligent Systems, 2019. **34**(6): p. 14-23.

10. Moore, J., N. Hammerla, and C. Watkins, *Explaining deep learning models with constrained adversarial examples*. 2019. p. 43-56.

11. Ramon, Y., et al., *A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C.* Advances in Data Analysis and Classification, 2020. **14**(4): p. 801-819.

12. Wachter, S., B. Mittelstadt, and C. Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.* Harvard journal of law & technology, 2018. **31**: p. 841-887.

13. Grath, R.M., et al., *Interpretable Credit Application Predictions With Counterfactual Explanations.* ArXiv, 2018. **abs/1811.05245**.

14. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature – PPSN XVI*. 2020. Cham: Springer International Publishing.

15. Poyiadzi, R., et al., *FACE: Feasible and Actionable Counterfactual Explanations*. 2020. 344-350.

16. Gomez, O., et al. *ViCE*. 2020.

17. Mothilal, R.K., A. Sharma, and C. Tan. *Explaining machine learning classifiers through diverse counterfactual explanations*. 2020.

18. Keane, M. and B. Smyth, *Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)*. 2020.

19. Karimi, A.-H., et al. *Model-agnostic counterfactual explanations for consequential decisions*. in *International Conference on Artificial Intelligence and Statistics*. 2020. PMLR.

20. Schleich, M., et al., *GeCo: Quality Counterfactual Explanations in Real Time*. 2021.

21. Laugel, T., et al., *Unjustified Classification Regions and Counterfactual Explanations in Machine Learning*. 2020. p. 37-54.

22. Verma, S., J.P. Dickerson, and K. Hines, *Counterfactual Explanations for Machine Learning: A Review.* ArXiv, 2020. **abs/2010.10596**.