# Introduction

1. Over and undersampling – Pros and cons

* Oversampling
* More observations from minority class
* Can amplify noise
* Undersampling
* Lose important info
* Remove noisy observations

1. Combining over and undersampling – combine the pros of 2 methods
2. Which to combine

* Random selection over and undersampling: not very useful
* Generate observations harder to classify + Remove easy to classify observations: More of the observations that are hard to classify – not very useful
* **Remove generation (SMOTE) + removes noisy observations (NCR, Instance hardness, ENN, Tomek Links) – ideal**

# Wrapping up

1. Scalability

* Methods based on KNN do not scale well
* Distance metrics
* Categorical variables