Pitfalls of Variable Importance Measures in Machine Learning

Berlin School of Economics and Law

M Loecher

Motivation

Prediction versus Understanding

- variables are seldom equally relevant
- Find ranking in "impact"
- Relative importance of regressor variables is an old topic
-

Data

- Sinking of the Titanic

- Kaggle's Two Sigma Connect:

NY Rental Listing Inquiries competition

- Swiss Fertility

Linear Models

Titanic Survival I

Important Predictor Variables

	Survived
Sexmale	-0.489^{***} (0.031)
Pclass	-0.193^{***} (0.023)
SibSp	-0.052^{***} (0.017)
Parch	-0.013 (0.019)
Age	-0.007^{***} (0.001)
Fare	0.0003 (0.0003)
PassengerId	0.0001 (0.0001)
Constant	1.341^{***} (0.082)
Observations	714
R^2	0.401
Adjusted R^2	0.395
Residual Std. Error	0.382 (df = 706)
F Statistic	67.625^{***} (df = 7; 706)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Titanic Survival II

Correlated Features

	Survived
Sexmale	-0.511^{***} (0.032)
Fare	0.002^{***} (0.0003)
Age	-0.002 (0.001)
Constant	0.720^{***} (0.040)
Observations	714
R^2	0.322
Adjusted R^2	0.319
Residual Std. Error	0.406 (df = 710)
F Statistic	112.383^{***} (df = 3; 710)
Note:	*p<0.1; **p<0.05; ***p<0.01

Titanic Survival III

Interactions

	Survived
Sexmale	-0.826^{***} (0.077)
Pclass	-0.244^{***} (0.025)
Pclass:Sexmale	0.138^{***} (0.032)
Constant	1.269^{***} (0.059)
Observations	891
R^2	0.381
Adjusted R^2	0.378
Residual Std. Error	0.384 (df = 887)
F Statistic	181.661^{***} (df = 3; 887)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interest in new rental on RentHop

- bathrooms: number of bathrooms
- bedrooms: number of bathrooms
- latitude
- longitude
- **price**: in USD
- **interest_level**: 'high', 'medium', 'low'
- street_address
- photos: a list of photo links.
- building_id
- created
- description
- display_address
- features: a list of features about this apartment

NY rent data set I

"Location, Location, Location, ..?"

	Dependent variable:
	price
latitude	−6,442.635 (6,600.829)
longitude	−3,559.672 (3,638.745)
bathrooms	1,994.054* (1,060.910)
bedrooms	677.732 (481.168)
Constant	−128.134 (20,096.990)
Observations	10,000
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	44,902.160 (df = 9995)
F Statistic	3.173** (df = 4; 9995)
Note:	*p<0.1; **p<0.05; ***p<0.01

NY rent data set II

	Dependent variable:
	interest_level
bathrooms	-0.178^{***} (0.015)
latitude	-0.083 (0.091)
longitude	-0.051 (0.050)
bedrooms	0.048^{***} (0.007)
price	-0.00000 (0.00000)
Constant	1.146^{***} (0.278)
Observations	10,000
R^2	0.015
Adjusted R^2	0.015
Residual Std. Error	0.620 (df = 9994)
F Statistic	30.774^{***} (df = 5; 9994)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

VI, Linear Models

- *"Variable importance is not very well defined as a concept. There is no theoretically defined variable importance metric..."*
- Change in R^2 when the variable is added to the model last
- Average order-dependent R^2 allocations over all p! orderings (**LMG**)

+ Direction/sign of contribution
+ Uncertainty "for free"
+ Easy to understand !?
- Marginal versus conditional
- Confounding effects
- Slave to linearity
- Interactions must be coded apriori

Machine Learning

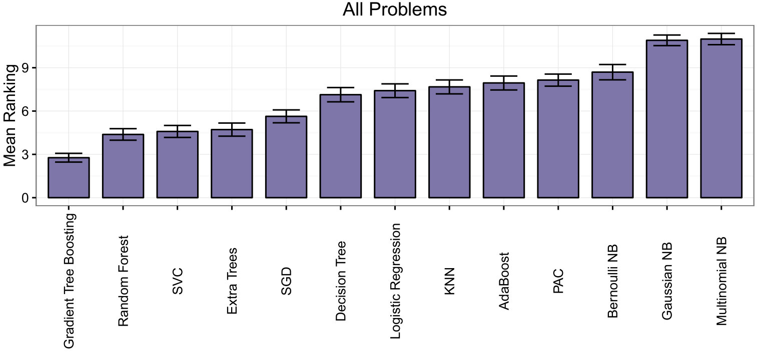
Which machine learning algorithms ?

Data-driven advice for applying machine learning to bioinformatics problems

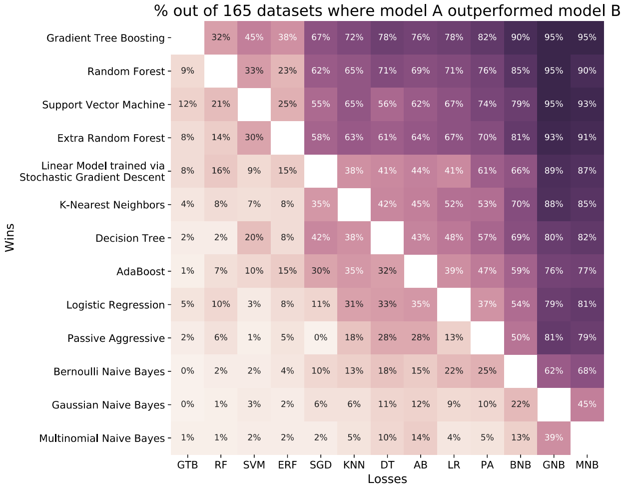
Randal S. Olson†*, William La Cava†*, Zairah Mustahsan, Akshay Varik, and Jason H. Moore†
Institute for Biomedical Informatics, University of Pennsylvania
Philadelphia, PA 19104, USA
†*E-mails: rso@randalolson.com, lacava@upenn.edu and jhmoore@upenn.edu*

As the bioinformatics field grows, it must keep pace not only with new data but with new algorithms. Here we contribute a thorough analysis of 13 state-of-the-art, commonly used machine learning algorithms on a set of 165 publicly available classification problems in order to provide data-driven algorithm recommendations to current researchers. We present a number of statistical and visual comparisons of algorithm performance and quantify the effect of model selection and algorithm tuning for each algorithm and dataset. The analysis culminates in the recommendation of five algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems.

Boosting and Random Forests

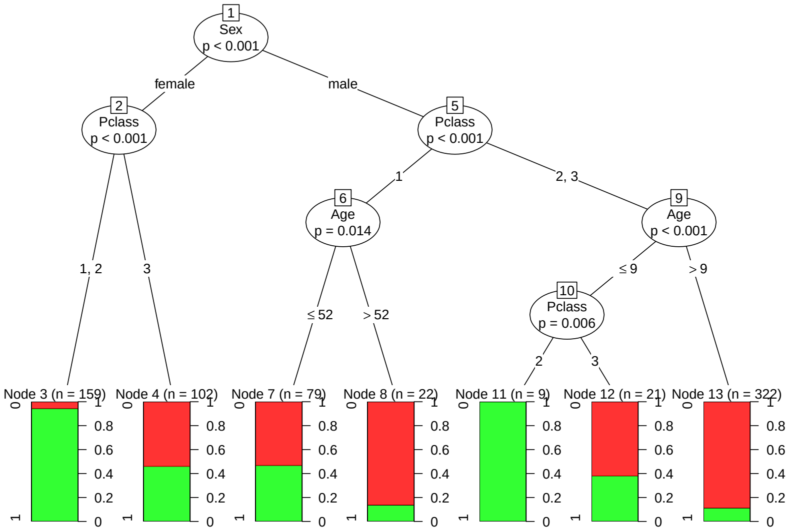


Who wins against whom



Trees

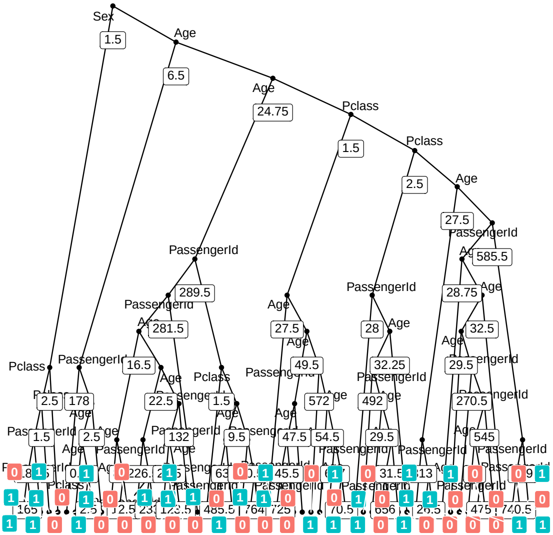
Titanic Tree



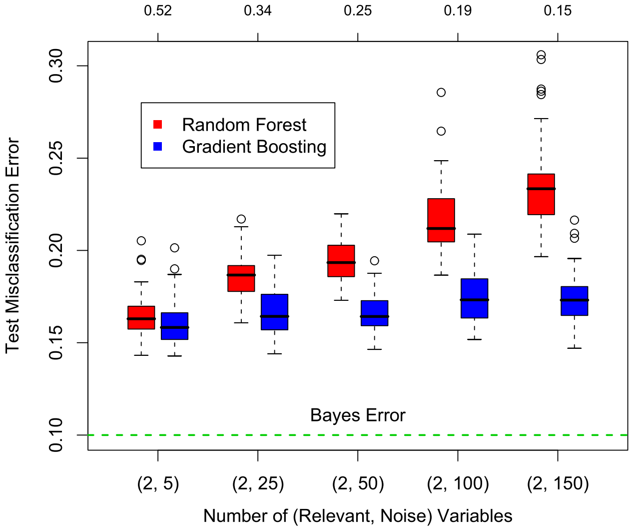
Trees details

- **Greedy**: At each split we minimize *squared error* or *node impurity*
- **All Interactions**: data “thin out” exponentially fast.
- **Piecewise Constant**: no smoothness, inferior for regression.
- **Model complexity**: depth of tree, typically single pruned trees
- **Boosting**: many shallow trees sequentially minimize loss
- **Random Forests**: many deep trees grown in parallel on bootstrapped samples. **Column sampling** leads to additional parameter

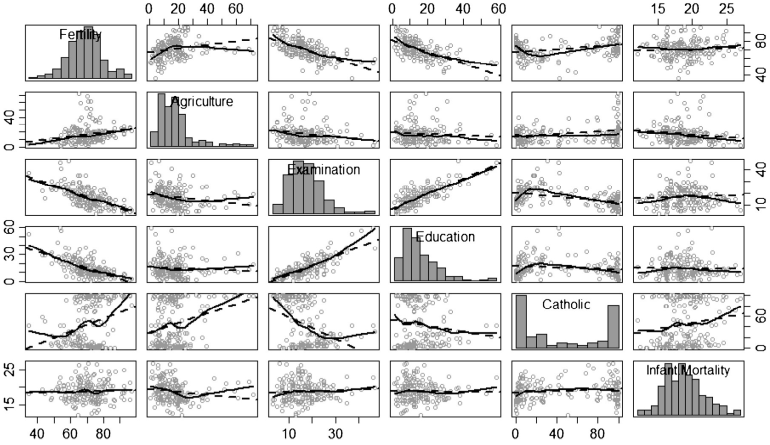
Deep Tree



Column Subsampling



Swiss Fertility



LR versus Forests

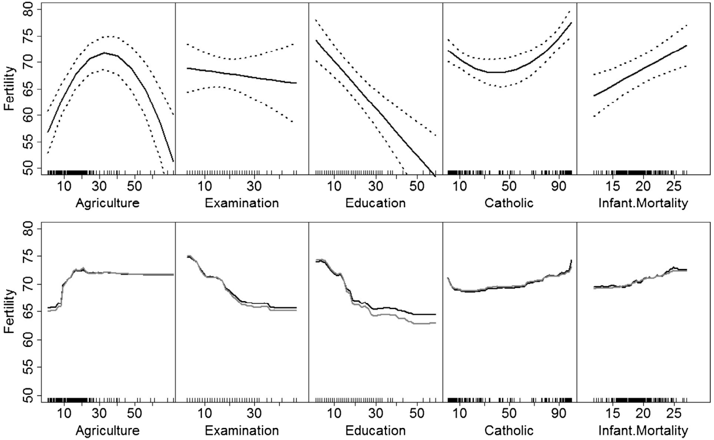


Figure 3. Main effects plot for the linear model (top, with 95% bands) and RF-CART (*mtry* = 1 (black) and *mtry* = 2 (gray)). Rugs at the bottom represent individual data values for the 182 Swiss provinces.

Variable Importance I

- **gini importance**: the mean decrease in impurity of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors)
- For a single decision tree T:

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2(v(t) = l)$$

  as a measure of relevance for each predictor variable X_l.The sum is over the $J - 1$ internal nodes of the tree.
- For ensembles it is simply averaged over the trees

$$GiniImp^2 = \frac{1}{M} \sum_{m=1}^{M} I_l^2(T_m)$$

The textbook story

California Housing data

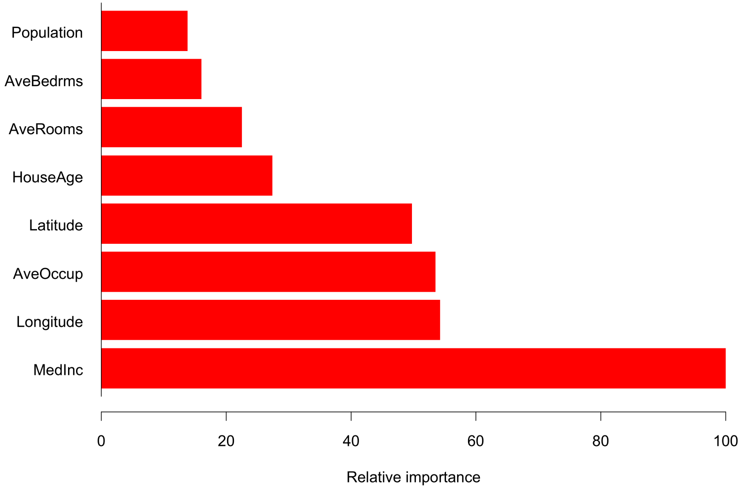


FIGURE 10.14. *Relative importance of the predictors for the California housing data.*

Bonus: Partial Dependence

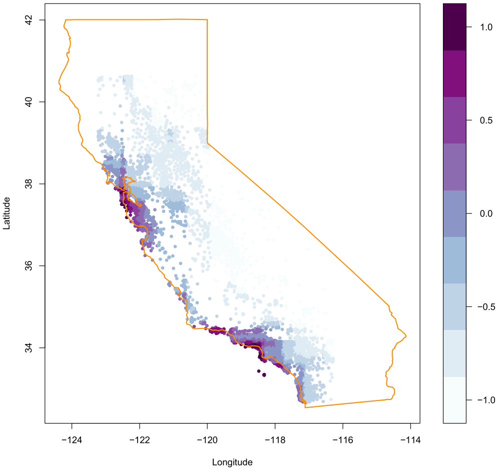


FIGURE 10.17. *Partial dependence of median house value on location in California. One unit is \$100,000, at 1990 prices, and the values plotted are relative to the overall median of \$180,000.*

Overfitting

Predictive vs. interpretational overfitting

- random forests are averages of large numbers of individually grown regression/classification trees.
- both "row and column subsampling": each tree is based on a random subset of the observations, and each split is based on a random subset of *mtry* candidate variables.
- The tuning parameter *mtry* can have profound effects on prediction quality as well as the variable importance measures outlined below.

Bootstrap: OOB

Due to the CART bootstrap row sampling, 36.8% of the observations are (on average) not used for an individual tree; those **"out of bag" (OOB)** samples can serve as a validation set to estimate the test error, e.g.:

$$E\left(Y - \hat{Y}\right)^2 \approx OOB_{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{\hat{y}}_{i,OOB}\right)^2 \tag{1}$$

where $\overline{\hat{y}}_{i,OOB}$ is the average prediction for the ith observation from those trees for which this observation was OOB.

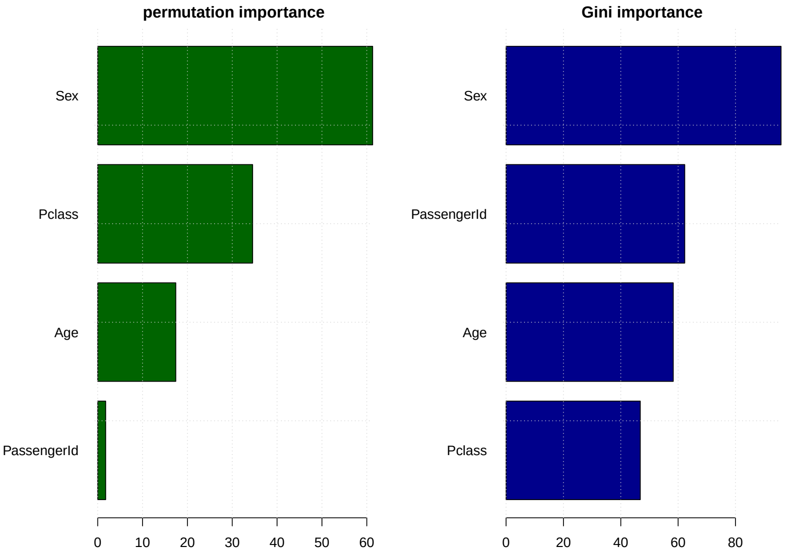
Variable Importance II

The default method to compute variable importance is the *mean decrease in impurity* (or *gini importance*) mechanism: At each split in each tree, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable. Note that this measure is quite like the R^2 in regression on the training set.

The widely used alternative *reduction in MSE when permuting a variable* as a measure of variable importance or short **permutation importance** is defined as follows:

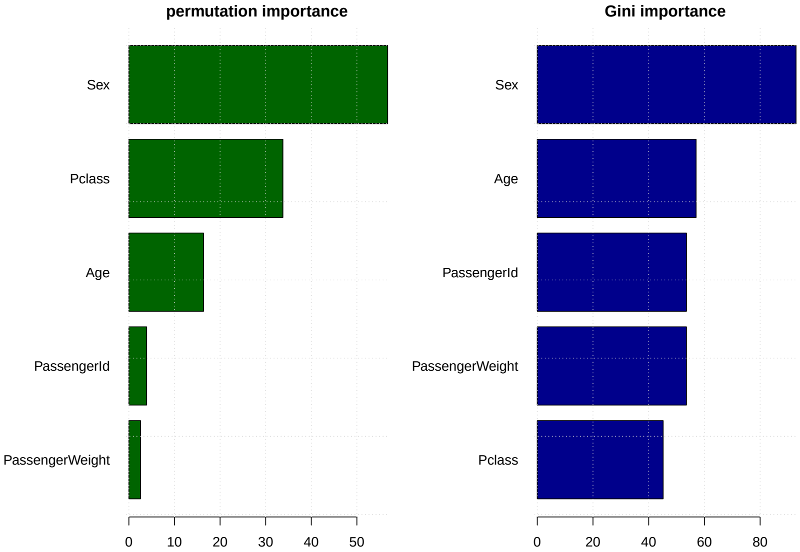
$$\mathrm{VI} = OOB_{MSE,perm} - OOB_{MSE} \tag{2}$$

Gini importance can be highly misleading



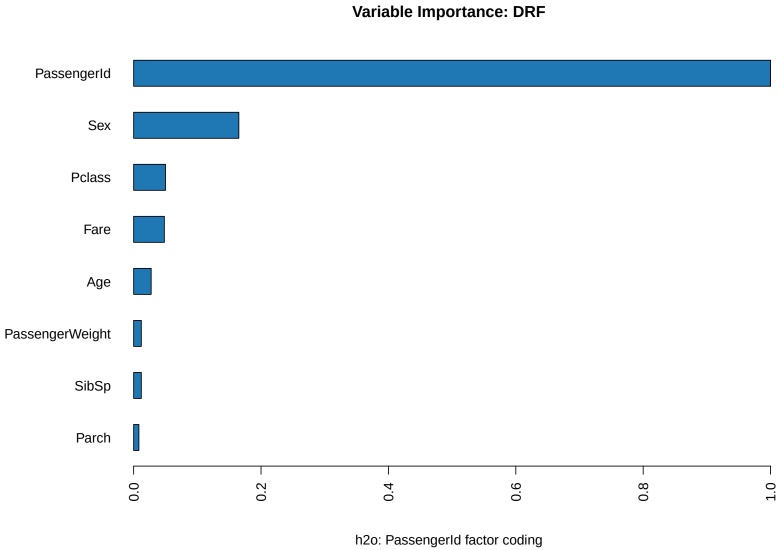
Noise Feature

Let us go one step further and add a Gaussian noise feature, which we call PassengerWeight:



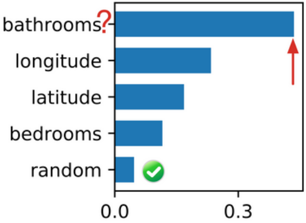
Categorical Features

Coding passenger ID as factor makes matters worse:

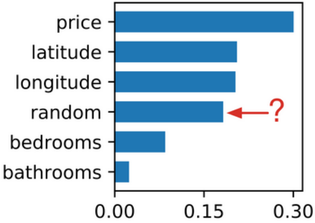


NY Rent, Gini Importance

Random Forest **regressor** predicting apartment rental price from 4 features + a column of random numbers. Random column is last, as we would expect but the importance of the number of bathrooms for predicting price is highly suspicious.

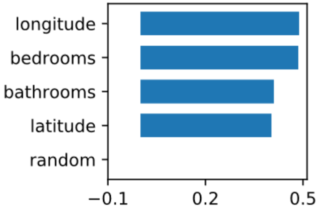


Random Forest **classifier** predicting apartment interest level (low, medium, high) using 5 features + a column of random numbers. Highly suspicious that random column is much more important than the number of bedrooms.

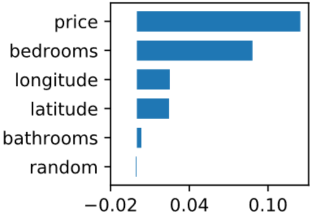


NY Rent, Permutation Importance

permuting each column and computing change in out-of-bag R^2.

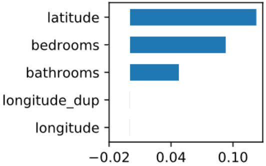


permuting each column and computing change in out-of-bag accuracy

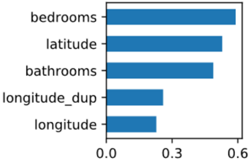


Collinear features

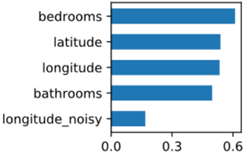
drop column importance



permutation importance

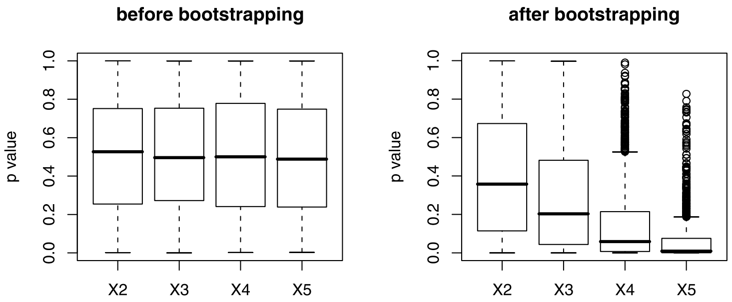


drop column importance with noise



Why Random Forests

- No pruning
- Column Subsampling
- Bootstrap



Distribution of the p values of χ^2 tests of each categorical variable $X_2, \ldots, X_5$ and the binary response for the null case simulation study, where none of the predictor variables is informative.

Outlook

Summary/Recommendations

- RF default importance not reliable
- use permutation importance for all models
- Boosting or extremely randomized trees for VI
- Careful about conditional versus marginal importance
- Social Sciences, Bioinformatics and Economics

```
## [1] "C:/Users/loecherm/Nextcloud/Research/AIZuerich/slid
```