caret walkthrough

Hunyong Cho
Department of Biostatistics

Install the following packages (It takes 5 minutes).

```
install.packages(c("caret", "RANN", "e1071", "randomForest", "neuralnet", "gbm"))
```

Why need a training and a test set?

If we use data to fit a model,
and use the same data to measure the error (e.g. MSE),

the measured error will be optimistic (overfitting problem).

⬇

Pretend to have only a part of the data to fit a model (training set),
and use the hold-out sample to measure the performance (test set)

Tuning parameters of Machine Learning Tools

penalty constant	LASSO, ridge, Elastic Net
min node size, ...	Trees (CART)
node size, # trees, # vars, ...	Random Forest
penalty coef, kernel type	SVM
...	

Tuning parameters should be given to fit a model.

What values should be given? We have to choose

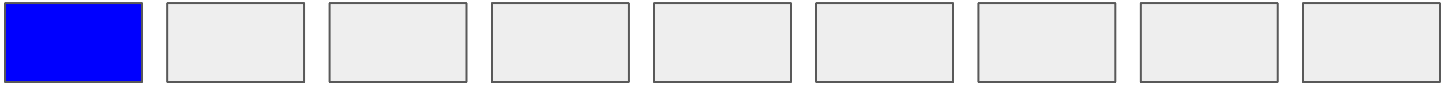
Need for Training/Validation/Test set splitting

Training	Fit models with different tuning parameters
Validation	Find the optimal set of parameters (min error, or best performance)
Test	Measure the realistic performance

K-fold Cross-Validation

Why use training set and validation set just once?

Partition the data into K chunks and ...



- Hold out the first set as the validation set and fit models using the rest.
- Repeat with the second set as the validation set, and so on.
- Average the performance across K measures.

* Test set should still be held out.