[9]:
```
# Global imports and settings
from preamble import *
%matplotlib inline
plt.rcParams['savefig.dpi'] = 120 # Use 300 for PDF, 100 for slides
#InteractiveShell.ast_node_interactivity = "all"
HTML('''<style>html, body{overflow-y: visible !important} .CodeMirror{min-w
```

```
<IPython.core.display.HTML object>
```

Machine learning pipelines

Preprocessing

- Many of the algorithms that we've seen are greatly affected by *how* you represent the training data
- Scaling, numeric/categorical values, missing values, feature selection/construction
- We typically need chain together different algorithms
  - Many preprocessing steps
  - Possibly many models
- This is called a *pipeline* (or *workflow*)

Outline: * Examples of which preprocessing steps are best combined with learning algorithm * How to apply preprocessing techniques * How to design and optimize pipelines * Some practical advice

Applying data transformations (recap)

- First, we `fit` the preprocessor on the **training data**
  - This computes the necessary transformation parameters
  - For `MinMaxScaler`, these are the min/max values for every feature
- Second, we use the fitted preprocessor to `transform` **training data** and the **test data**
  - No information should `leak` from the test data into the training data
- You can fit and transform the training together with `fit_transform`

[36]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(cancer.data, cancer.ta
                                                    random_state=0)
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Other transformation techniques

We've covered many other transformation techniques before. These can all be used in pipelines. * `OneHotEncoder`: convert categorical to numeric features * `np.digitize`: discretization (binning) of numeric features * `PolynomialFeatures`: construct all interactions with polynomials up to a given degree * `SelectPercentile`: use ANOVA to select most informative features * `SelectFromModel(RandomForestClassifier())`: model-based feature selection * `RFE`: recursive feature elimination * `VarianceThreshold`: removes low-variance (e.g. constant) features

Missing value imputation

- Many sci-kit learn algorithms cannot handle missing value
- `Imputer` replaces specific values
  - `missing_values` (default 'NaN') placeholder for the missing value
  - `strategy`:
    * `mean`, replace using the mean along the axis
    * `median`, replace using the median along the axis
    * `most_frequent`, replace using the most frequent value
- Many more advanced techniques exist, but not yet in scikit-learn
  - e.g. low rank approximations (uses matrix factorization)

```
[14]: from sklearn.preprocessing import Imputer
      X1_train = [[1, 2], [np.nan, 3], [7, 6]];
      imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
      imp.fit(X1_train)
      X1 = [[np.nan, 2], [6, np.nan], [7, 6]]
      print("Missing data: {}".format(X1) )
      print("Imputed data:\n {}".format(imp.transform(X1))  )
```

```
Missing data: [[nan, 2], [6, nan], [7, 6]]
Imputed data:
 [[4.    2.   ]
 [6.    3.667]
 [7.    6.   ]]
```

Dimensionality reduction techniques

- Many techniques
  - `decomposition.PCA` (Principal Component Analysis)
  - `manifold.MDS` (Multi-Dimensional Scaling)
  - `manifold.Isomap`
  - `random_projection.GaussianRandomProjection`
  - `random_projection.johnson_lindenstrauss_min_dim`
  - ...
- See coming lectures
- Reducing the number of features greatly helps distance-based algorithms (kNN, clustering,...)

  - Curse of dimensionality (Bellman's curse): for every new feature, we need exponentially more data

- Very useful in their own right
- Note: not all dimensionality reduction techniques can be used as a transformer.
  - They have a `fit`, but no `transform` method, and can't be applied on test data.

How great is the effect of scaling on (RBF) SVMs?

- First, we train the SVM without scaling

[15]:
```
from sklearn.svm import SVC

svm = SVC()
svm.fit(X_train, y_train)
print("Test set accuracy: {:.2f}".format(svm.score(X_test, y_test)))
```

```
Test set accuracy: 0.63
```

- With scaling, we get a much better model

[16]:
```
# preprocessing using 0-1 scaling
scaler = MinMaxScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)

# learning an SVM on the scaled training data
svm.fit(X_train_scaled, y_train)
# scoring on the scaled test set
print("Scaled test set accuracy: {:.2f}".format(svm.score(X_test_scaled, y
```

```
Scaled test set accuracy: 0.95
```

Hyperparameter Selection with Preprocessing

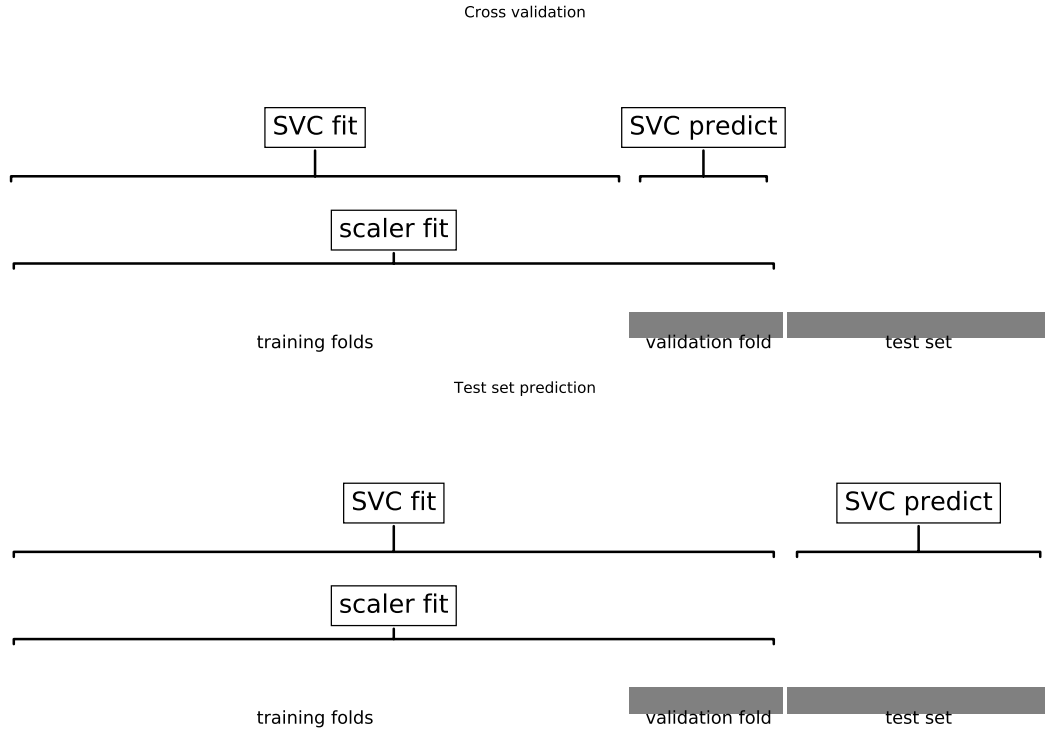
- If we also want to optimize our SVM's hyperparameters, things get more complicated
- Indeed, when we `fit` the preprocessor (`MinMaxScaler`), we used *all* the training data.
- The cross-validation splits in GridSearchCV will have training sets preprocessed with information from the test sets (data leakage)

[17]:
```
from sklearn.model_selection import GridSearchCV
# illustration purposes only, don't use this code
param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100],
              'gamma': [0.001, 0.01, 0.1, 1, 10, 100]}
grid = GridSearchCV(SVC(), param_grid=param_grid, cv=5)
grid.fit(X_train_scaled, y_train)
print("Best cross-validation accuracy: {:.2f}".format(grid.best_score_))
print("Best set score: {:.2f}".format(grid.score(X_test_scaled, y_test)))
print("Best parameters: ", grid.best_params_)
```

```
Best cross-validation accuracy: 0.98
Best set score: 0.97
Best parameters:  {'gamma': 1, 'C': 1}
```

Visualization of what happens in this code * During cross-validation (grid search) we evaluate hyperparameter settings on a validation set that was preprocessed with information in that validation set * This will lead to overly optimistic results during cross-validation * When we want to use the optimized hyperparameters on the held-out test data, the selected hyperparameters may be suboptimal. * To solve this, we need to *glue* the preprocessing and learning algorithms together by building a pipeline

```
[12]: mglearn.plots.plot_improper_processing()
```



Using Pipelines in Grid-searches

- We can use the pipeline as a single estimator in `cross_val_score` or `GridSearchCV`
- To define a grid, refer to the hyperparameters of the steps
  - Step `svm`, parameter `C` becomes `svm__C`

```
[22]: param_grid = {'svm__C': [0.001, 0.01, 0.1, 1, 10, 100],
                   'svm__gamma': [0.001, 0.01, 0.1, 1, 10, 100]}
```

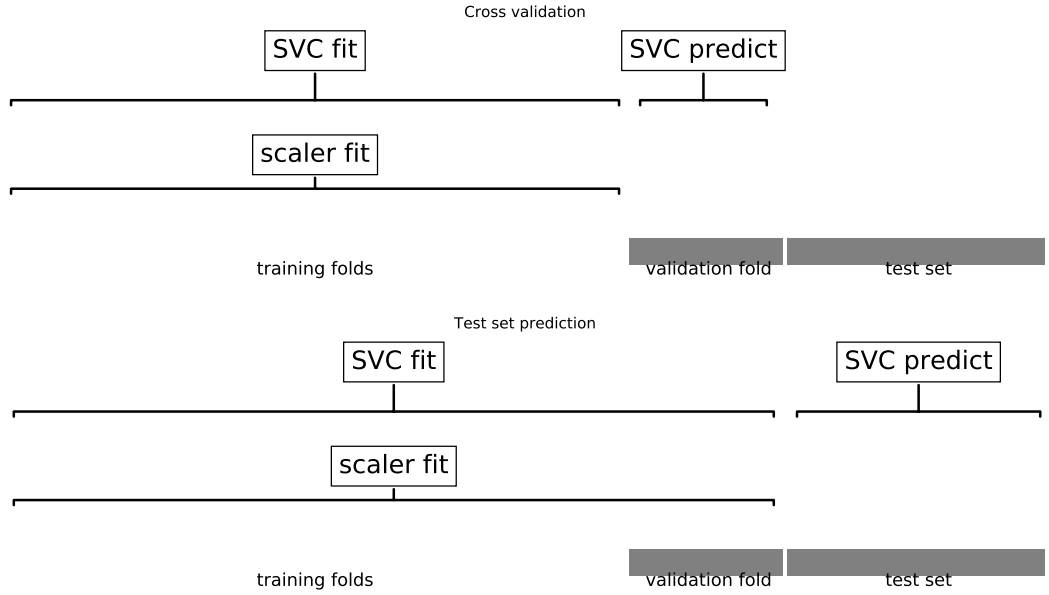
```
[24]: from sklearn import pipeline

      pipe = pipeline.Pipeline([("scaler", MinMaxScaler()), ("svm", SVC(C=100))]
      grid = GridSearchCV(pipe, param_grid=param_grid, cv=5)
```

```
      grid.fit(X_train, y_train)
      print("Best cross-validation accuracy: {:.2f}".format(grid.best_score_))
      print("Test set score: {:.2f}".format(grid.score(X_test, y_test)))
      print("Best parameters: {}".format(grid.best_params_))
```

```
Best cross-validation accuracy: 0.98
Test set score: 0.97
Best parameters: {'svm__C': 1, 'svm__gamma': 1}
```

- Now, the preprocessors are refit with only the training data in each cross-validation split.

```
[25]: mglearn.plots.plot_proper_processing()
```



- When we request the best estimator of the grid search, we'll get the best pipeline

```
[26]: print("Best estimator:\n{}".format(grid.best_estimator_))
```

```
Best estimator:
Pipeline(memory=None,
     steps=[('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))), ('svm', S
  decision_function_shape='ovr', degree=3, gamma=1, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False))])
```

- And we can drill down to individual components and their properties

```
[27]: # Get the SVM
      print("SVM step:\n{}".format(
            grid.best_estimator_.named_steps["svm"]))
```

```
SVM step:
SVC(C=1, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma=1, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
```

```
[28]: # Get the SVM dual coefficients (support vector weights)
      print("SVM support vector coefficients:\n{}".format(
            grid.best_estimator_.named_steps["svm"].dual_coef_))
```

```
SVM support vector coefficients:
[[-1.    -1.    -1.    -1.    -1.    -1.    -1.    -1.    -0.822 -0.609
  -0.93  -1.    -0.665 -1.    -0.793 -1.    -1.    -1.    -1.    -1.
  -1.    -0.63  -1.    -1.    -1.    -0.166 -0.575 -1.    -0.119 -1.
  -0.38  -1.    -0.55  -0.048 -0.606 -1.    -0.071 -1.    -0.126 -0.106
  -0.285 -1.    -1.    -1.    -0.46   1.     0.932  1.     1.     0.901
   1.     0.911  1.     1.     1.     0.653  1.     1.     0.953  0.577
   1.     1.     1.     1.     1.     0.659  1.     1.     0.757  1.
   1.     1.     1.     1.     1.     1.     0.096  0.274  0.329  1.
   1.     1.     0.491  1.     0.41 ]]
```

Information leakage

- See 'Elements of statistical learning' for a great example of data leakage
- Consider a synthetic regression task with 100 samples and 10,000 features with data and labels independently sampled from a Gaussian distribution
  - Hence, there should be no relation between the data `X` and target `y`

```
[29]: rnd = np.random.RandomState(seed=0)
      X = rnd.normal(size=(100, 10000))
      y = rnd.normal(size=(100,))
```

- First, we select the 5% most informative features with `SelectPercentile`, and then evaluate a `Ridge` regressor

```
[30]: from sklearn.feature_selection import SelectPercentile, f_regression

      select = SelectPercentile(score_func=f_regression, percentile=5).fit(X, y)
      X_selected = select.transform(X)
      print("X_selected.shape: {}".format(X_selected.shape))
```

```
X_selected.shape: (100, 500)
```

```
[31]: from sklearn.model_selection import cross_val_score
      from sklearn.linear_model import Ridge
      print("Cross-validation accuracy (cv only on ridge): {:.2f}".format(
            np.mean(cross_val_score(Ridge(), X_selected, y, cv=5))))
```

```
Cross-validation accuracy (cv only on ridge): 0.91
```

- The R^2 performance is 0.91 (a very good model). This can't be right given that our data is random.
- Our feature selection picked out some of the random features which (by chance) correlated with the random target.
- Because we selected the features *outside* of the cross-validation, it could find features that are correlated on the samples in the test folds.
- Hence, information leaked from the test set into the training set (through the selection of features).

- Now, let's do a proper cross-validation using a pipeline:

```
[33]: pipe = pipeline.Pipeline([("select", SelectPercentile(score_func=f_regress
                                                           percentile=5)),
                                ("ridge", Ridge())])
      print("Cross-validation accuracy (pipeline): {:.2f}".format(
            np.mean(cross_val_score(pipe, X, y, cv=5))))
```

```
Cross-validation accuracy (pipeline): -0.25
```

- We get a negative R^2 score (a very poor model), as expected.
- The feature selection now happens *inside* the cross-validation loop, using only the training folds.
- It will still find features that correlate with the labels in the training data, but not with those in the test data.

Grid-searching preprocessing steps and model parameters

- We can use grid search to optimize the hyperparameters of our preprocessing steps and learning algorithms at the same time
- Consider the following pipeline:
  - `StandardScaler`, without hyperparameters
  - `PolynomialFeatures`, with the max. *degree* of polynomials
  - `Ridge` regression, with L2 regularization parameter *alpha*

```
[37]: from sklearn.datasets import load_boston
      boston = load_boston()
      X_train, X_test, y_train, y_test = train_test_split(boston.data, boston.ta
                                                          random_state=0)
      from sklearn.preprocessing import PolynomialFeatures
      pipe = pipeline.make_pipeline(
          StandardScaler(),
          PolynomialFeatures(),
          Ridge())
```

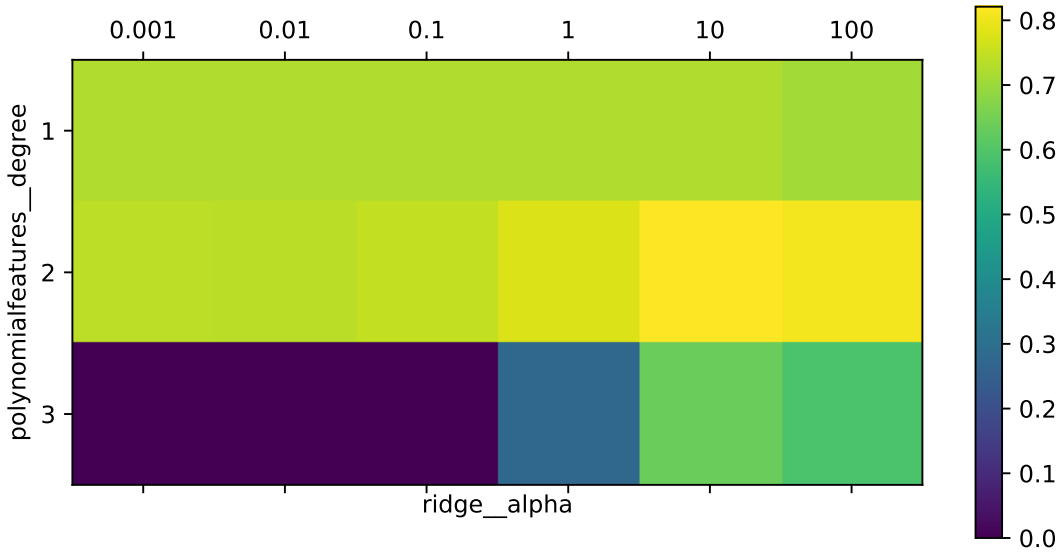
- We don't know the optimal polynomial degree or alpha value, so we use a grid search (or random search) to find the optimal values

```
[38]: param_grid = {'polynomialfeatures__degree': [1, 2, 3],
                   'ridge__alpha': [0.001, 0.01, 0.1, 1, 10, 100]}
      # Note: I had to use n_jobs=1. (n_jobs=-1 stalls on my machine)
      grid = GridSearchCV(pipe, param_grid=param_grid, cv=5, n_jobs=1)
      grid.fit(X_train, y_train)
```

```
GridSearchCV(cv=5, error_score='raise',
       estimator=Pipeline(memory=None,
     steps=[('standardscaler', StandardScaler(copy=True, with_mean=True, with_st
   normalize=False, random_state=None, solver='auto', tol=0.001))]),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'ridge__alpha': [0.001, 0.01, 0.1, 1, 10, 100], 'polynomialfe
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=0)
```

- Visualing the R^2 results as a heatmap:

```
[41]: plt.matshow(grid.cv_results_['mean_test_score'].reshape(3, -1),
                  vmin=0, cmap="viridis")
      plt.xlabel("ridge__alpha")
      plt.ylabel("polynomialfeatures__degree")
      plt.xticks(range(len(param_grid['ridge__alpha'])), param_grid['ridge__alph
      plt.yticks(range(len(param_grid['polynomialfeatures__degree'])),
                 param_grid['polynomialfeatures__degree'])

      plt.colorbar();
```



- Here, degree-2 polynomials help (but degree-3 ones don't), and tuning the alpha parameter helps as well.
- Not using the polynomial features leads to suboptimal results (see the results for degree 1)

```
[39]: print("Best parameters: {}".format(grid.best_params_))
      print("Test-set score: {:.2f}".format(grid.score(X_test, y_test)))
```

```
Best parameters: {'ridge__alpha': 10, 'polynomialfeatures__degree': 2}
Test-set score: 0.77
```

Algorithm selection

- It is also possible to use a grid search to consider various alternative algorithms for a specific step, and tune the pipelines as well
  - StandardScaler or MinMaxScaler?
  - RandomForest or SVM?
- Note that the search space quickly becomes huge
- As an example, let's consider a pipeline that explores:
  - StandardScaler + SVM, varying `gamma` and `C`
  - None + RandomForest, varying `max_features`
- We instantiate a general pipeline, and define everything else in the grid
  - We need to define a list of 2 grids because of dependencies.

```
[40]: pipe = pipeline.Pipeline([('preprocessing', StandardScaler()), ('classifie
```

```
[41]: from sklearn.ensemble import RandomForestClassifier

      param_grid = [
          {'classifier': [SVC()], 'preprocessing': [StandardScaler(), None],
           'classifier__gamma': [0.001, 0.01, 0.1, 1, 10, 100],
           'classifier__C': [0.001, 0.01, 0.1, 1, 10, 100]},
          {'classifier': [RandomForestClassifier(n_estimators=100)],
           'preprocessing': [None], 'classifier__max_features': [1, 2, 3]}]
```

The Scaling+SVM pipeline wins!

```
[42]: X_train, X_test, y_train, y_test = train_test_split(
          cancer.data, cancer.target, random_state=0)

      grid = GridSearchCV(pipe, param_grid, cv=5)
      grid.fit(X_train, y_train)

      print("Best params:\n{}\n".format(grid.best_params_))
      print("Best cross-validation score: {:.2f}".format(grid.best_score_))
      print("Test-set score: {:.2f}".format(grid.score(X_test, y_test)))
```

```
Best params:
{'preprocessing': StandardScaler(copy=True, with_mean=True, with_std=True), 'cla
  decision_function_shape='ovr', degree=3, gamma=0.01, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False), 'classifier__gamma': 0.01}

Best cross-validation score: 0.99
Test-set score: 0.98
```

Algorithm selection: the road ahead

Automating the construction of machine learning pipelines is a hot topic of research, with different possible solutions:

- Model-based optimization: search large algorithm+hyperparameter spaces more effectively (limited to the learning step)
  - E.g. `Auto-SKLearn` and `Hyperopt-SKLearn`
- Multi-armed bandits and Racing: train algorithms first on a small sample, train the best 50% on a larger sample, continue until full dataset
- Genetic programming: very good at exploring large spaces
  - E.g. `sklearn-deap` or `TPOT` (builds pipelines as well)
- Meta-learning: build *meta-models* to predict which algorithms/hyperparameter ranges are most useful on new datasets
  - Requires large amounts of observations on previous datasets, as well as measureable dataset properties (meta-features) -> OpenML
- Combinations of the above
  - Especially meta-learning with all others
  - Contact me if interested :)

Approaching machine learning problems

- Just running your favourite algorithm on every new problem is usually not a great way to start
- Consider the problem at large
  - Do you want exploratory analysis or (black box) modelling?
  - How to define and measure success? Are there costs involved?
  - Do you have the right data? How can you make it better?
- Build prototypes early-on to evaluate the above.
- Analyse your model's mistakes
  - Should you collect more, or additional data?
  - Should the task be reformulated?
  - Often a higher payoff than endless grid searching
- Technical debt
  - Very complex machine learning prototypes are hard/impossible to put into practice
  - There is a creation-maintenance trade-off
  - See 'Machine Learning: The High Interest Credit Card of Technical Debt'

Concept drift

- Data is often a stream, and model building is often part of a feedback cycle
  - Collect new data, cleaning, build models, analyse
  - Continually check that your model is still working
- Concept drift: sudden or gradual changes in the phenomenon that you want to model
  - External: market/weather evolutions, human behavior

  - Feedback: your predictions (and actions) may change future data
    - E.g. movie recommendations, new drugs, treatments, ...

- Different ways to tackle this:
  - Repeated retraining (but often not sure when)
  - Stream mining algorithms (learn on fast streaming data)
  - Change detection techniques (retrain when model starts failing)
  - Meta-learning (switch algorithms depending on data properties)

Real world evaluations

- Accuracy is seldomly the right measure. Usually, costs are involved.
- Don't just evaluate your predictions themselves, also evaluate how the outcome improves *after* you take actions based on them
- Beware of non-representative samples. You often don't have the data you really need.
- Adversarial situations (e.g. spam filtering) can subvert your predictions
- Data leakage: the signal your model found was just an artifact of your data
  - See 'Why Should I Trust You?' by Marco Ribeiro et al.
- A/B testing to evaluate algorithms in the wild
  - More advanced: bandit algorithms

Further reading (Theory)

- The Elements of Statistical Learning (Hastie, Tibshirani, Friedman)
- Deep Learning (Goodfellow, Bengio, Courville) - see later
- Gaussian Processes for Machine Learning (Rasmussen, Williams)
- Machine Learning: An Algorithmic Perspective (Marsland)
- Pattern Recognition and Machine Learning (Bishop)
- Machine Learning: A Probabilistic Perspective (Murphy)
- Foundations of Data Science (Blum, Hopcroft, Kannan)

Other packages

- Deep learning: `TensorFlow`, `Torch`, `Caffe`,...
  - `Keras`, `Lasagne` provide simpler interfaces
- R libraries often provide a richer variation of techniques
  - Powerful statistical analysis and visualization tools
  - mlr (Machine Learning in R)
- `vowpal wabbit` (vw): C++ library for large datasets and streaming data
- MOA (Massive Online Analysis): Java library for streaming data
- MLLib: Scala library on top of Spark, for large distributed systems
- PyMC/Stan: Probabilistic programming, allows to model how likely each data point is correct. Simple, elegant way to build models.

Summary

- Pipelines allow us to encapsulate multiple steps into a single estimator
  - Has `fit`, `transform`, and `predict` methods
- Avoids data leakage, hence crucial for proper evaluation
- Choosing the right combination of feature extraction, preprocessing, and models is somewhat of an art.
- Pipelines + Grid/Random Search help, but search space is huge
  - Smarter techniques are being researched
- Real world applications require careful thought, prototyping, and tireless evaluation.