Computational Aspects of KD

- Data Access
  - read.csv
  - write.csv
  - edit
- Visualization
  - scatter plots
- Data Manipulation
  - attribute-oriented approach
  - observation-oriented approach
- Model Building and Evaluation
- Model Deployment

Data Manipulation

Recall that a data frame is a data table representation in R,

```
> mammals.df
  Legs Wings  Fur Feathers Mammal
1    4    no  yes       no   true
2    2   yes   no      yes  false
3    4    no   no       no  false
4    4   yes  yes       no   true
5    3    no   no       no  false
```

Data Manipulation

Attribute-oriented Approach

We can access any attribute in the mammals data frame with the `$` notation.

```
> mammals.df$Legs
[1] 4 2 4 4 3
> mammals.df$Mammal
[1]  true   false  false  true   false
Levels:  false  true
```

R allows us to select groups of attributes with the `subset` function,

```
> subset(mammals.df, select=Fur:Mammal)
   Fur Feathers Mammal
1  yes       no   true
2   no      yes  false
3   no       no  false
4  yes       no   true
5   no       no  false
> subset(mammals.df, select=-Mammal)
  Legs Wings  Fur Feathers
1    4    no  yes       no
2    2   yes   no      yes
3    4    no   no       no
4    4   yes  yes       no
5    3    no   no       no
```

Data Manipulation

Observation-oriented Approach

We can use the *subset* function also for observation-oriented data manipulation.

```
> subset(mammals.df, Legs == 4)
  Legs Wings  Fur Feathers Mammal
1    4    no  yes       no   true
3    4    no   no       no  false
4    4   yes  yes       no   true
```

Another, slightly more complicated example,

```
> mammal.levels <- levels(mammals.df$Mammal)
> mammal.levels
[1] "false" "true"
> true.level <- mammal.levels[2]
> subset(mammals.df, Mammal == true.level)
  Legs Wings Fur Feathers Mammal
1    4    no yes       no   true
4    4   yes yes       no   true
```

Model Building and Evaluation

We use the library 'e1071' (don't ask :) for building support vector machine models. [a]

```
> library(e1071)
```

Now we can construct a support vector machine model of our mammals data with the `svm` function,

```
> model<-svm(Mammal ~ .,data=mammals.df,kernel="linear")
```

At this point we can evaluate our model by checking how it performs on the training set.

```
> mammals.df$Mammal == fitted(model)
[1] TRUE TRUE TRUE TRUE TRUE
```

[a] The library is available through the Package Installer.

Model Deployment

Model deployment means applying your model in an appropriate context. In R we use the `predict` function to compute the value of the dependent attribute for some object. Given that R is a programming language we could program appropriate functionality around the predict function.

```
> independent.df <- subset(mammals.df, select=-Mammal)
> predict(model, independent.df)
     1      2      3      4      5
  true  false  false   true  false
Levels:  false  true
```

How could we test in R whether these predictions are correct with respect to the original data set `mammals.df`?