What I have done last 2 days:

1. I took Mushrooms data from Kaggle, because column names are added.
2. Followed the code on [**https://www.kaggle.com/mokosan/mushroom-classification**](https://www.kaggle.com/mokosan/mushroom-classification) and repeated it:
   1. Use the same chunk for splitting the data to train and test parts
   2. The same chunk for choosing the variables x1=odor and x2=spore.print.color for logistic regression classifier
   3. Same schema and functions about computing confusion matrix and plotting ROC curves.
   4. Same code for the 5 classificators
   5. Similar code for presenting the results in table format
   6. Same code for plotting all ROC curves together.

File: **golden\_truth\_classifiers.R**

**Classification results:**

Accuracy Sensitivity Specifity

Logistic Regression 0.9941545 0.9888535 1.000000

Naive Bayes 0.9782881 0.9692429 0.988465

Neural Nets 0.9941545 0.9888535 1.000000

Random Forest 0.9941545 0.9888535 1.000000

Decision Tree 0.9941545 0.9888535 1.000000

1. Next I repeated exactly the same code, but instead the original data **data\_mushrooms\_KAGGLE** I have used **dataConsensus**, obtained from simulated 10 workers noisy data (30% noise for every worker).

File: **consensus\_classifiers.R**

**Classification results:**

Accuracy Sensitivity Specifity

Logistic Regression 0.8981211 0.9076433 0.8876207

Naive Bayes 0.8926931 0.9066774 0.8776042

Neural Nets 0.8981211 0.9076433 0.8876207

Random Forest 0.8981211 0.9076433 0.8876207

Decision Tree 0.8981211 0.9076433 0.8876207

Note:

Till now I have not used workers ID’s at all. It seems that we do not need them when we use majority voting. It may be that we will need these ID’s if we want to use weighted majority voting, where the particular weights are related to the corresponding ID’s. Then during shuffling for splitting train and test data sets we will need to preserve these ID’s as well.

**What I plan to do this evening:**

1. To simulate the mixed population of 1000 workers, 250 of each group (true experts, experts, amateurs, adversaries) and to run **consensus\_classifiers.R** using these worker simulated data.
2. To simulate repeated labeling and repeat the same stuff in **consensus\_classifiers.R**

**What else has to be done:**

1. To repeat the same procedure using two other data sets. We can choose something similar to mushrooms from KAGGLE – no later than the middle of this week (9-15.04; **11-12.04**). .
2. Try to find other working consensus algorithm and apply it to the data. Weighted majority voting should be easiest to realize, but I am afraid we do not have enough time. (deadline also middle of this week, **11-12.04**).
3. Describe the results and prepare slides. Easiest way: to use the slides from previous presentation and modify them. I suppose we can start to do it at the end of this week, **14-15.04.18**.