LAB WEEK 3 – CLASSIFICATION & PREDICTION #1

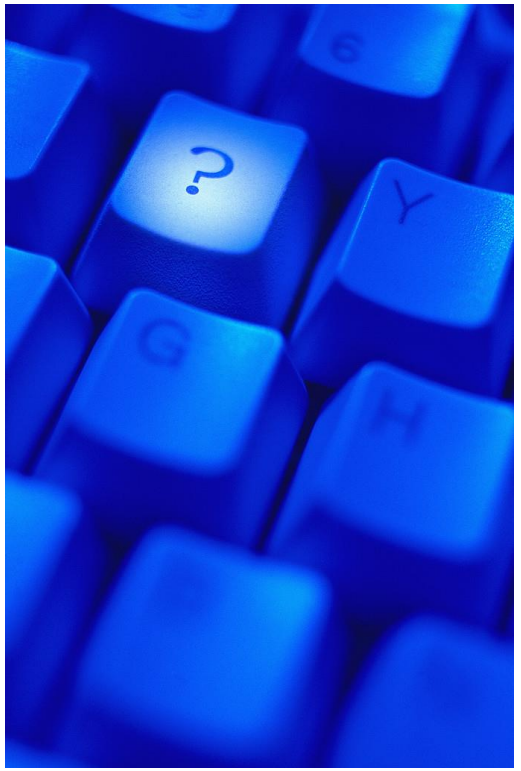
DATA MINING SPRING 2013 | ANDERS HARTZEN (ANDERSHH@ITU.DK) | JENS ANDERSSON GRØN (JANG@ITU.DK)

BUT FIRST

A LITTLE HOUSEKEEPING

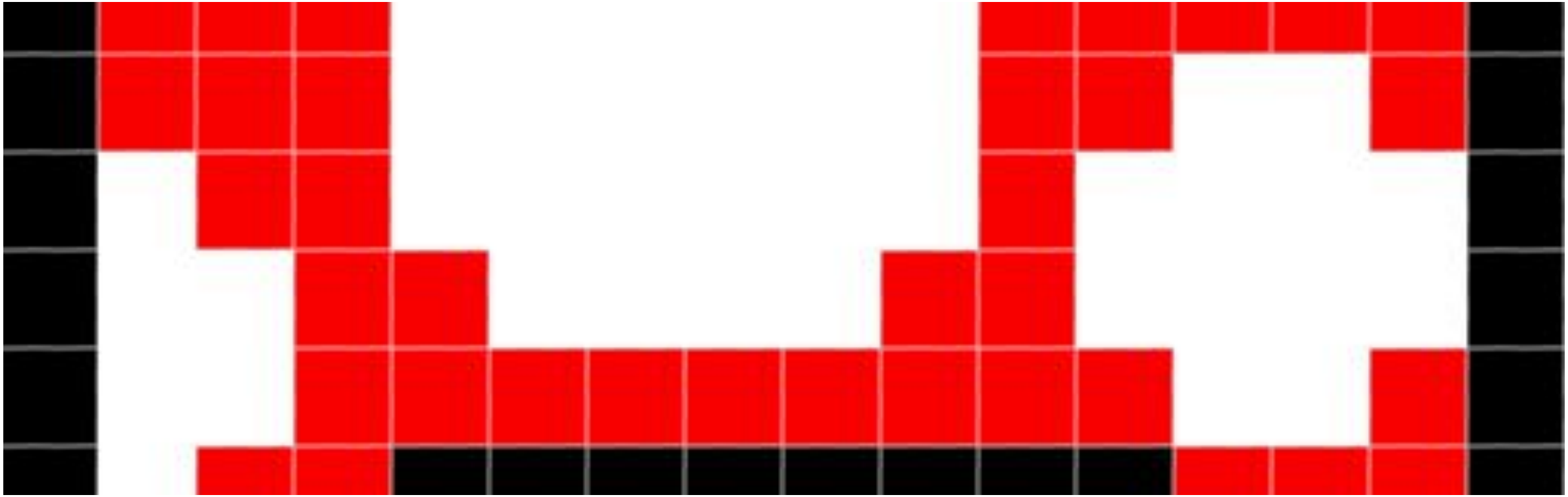
LAB/INDIVIDUAL ASSIGNMENT Q&A FORUM ON LEARNIT

- There is a new forum on the course learnIT page
- Use it to ask questions concerning the labs and the individual assignment
- Instead of emailing your questions individually to the TAs, we recommend you to post them in the forum instead
  - Some of your fellow students may already have asked the same question
  - You will be able to help each other
- TAs will check the forum at least once a day



LAB 1 PROBLEMS AND 2014 QUESTIONNAIRE DATASET

- There were some problems with the code and data given to you at the last lab
  - Some data were skipped by the data loading code (issue with the StringTokenizer class)
  - Some rows in the 2013 questionnaire data lacked some attributes.
    - Many thanks to Niels Abildgaard for detecting these issues and helping correct the data!
  - New version of code and data has been uploaded to the learnIT page
  - Apologies for these issues!
- Data based on your answers to the questionnaire from last week is now available on the learnIT
  - Included in the new code and date .zip file for lab 1

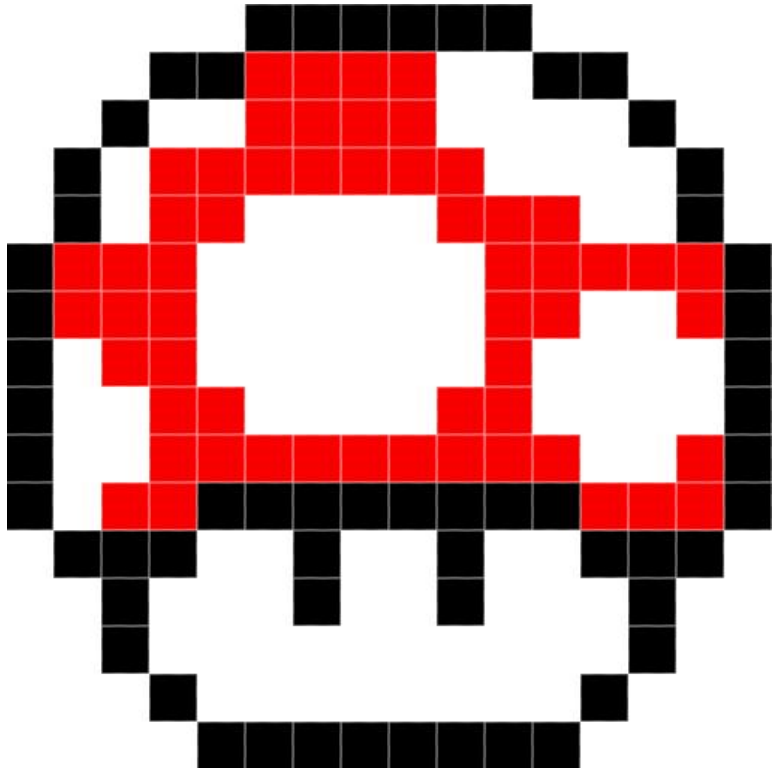


TODAY'S LAB

Mushrooms!

CLASSIFICATION & PREDICTION #1

- Today you will be working with data concerning mushrooms.
- You will try to build a classifier to predict whether not a given mushroom is edible or poisonous.
- You will build two classifiers:
  - One using the ID3 algorithm
  - The other using the kNN algorithm
- Compare their accuracy when classifying the mushroom data.
- Code is provided to help you load data and convert it to Java objects.

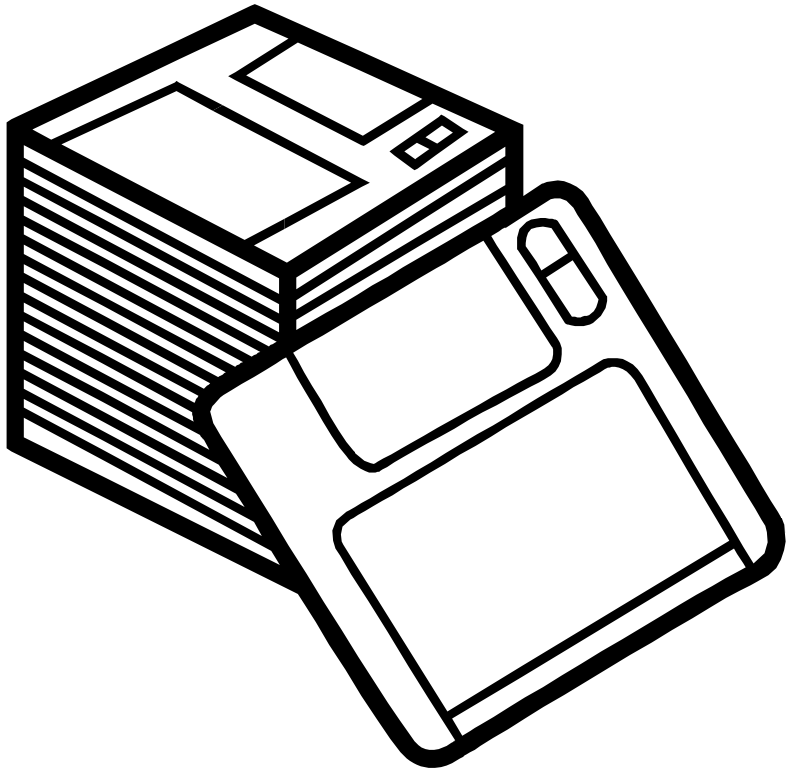


PLAN OF ATTACK

- First take a look at the code provided.
- Then decide which of the two classifiers to begin with
  - kNN is the simplest of the two, and thus easier to implement.
    - Pg. 422-423 in book.
  - ID3 is more complicated and takes a bit longer getting started with
    - Pg. 332-340 in book.
    - Decision tree data structure needed that can be used in the algorithm but also for classification of test tuples.
    - Visualization of decision tree?
    - Where to split data into training and test data?
- After implementing each classifier compare the two's accuracy.

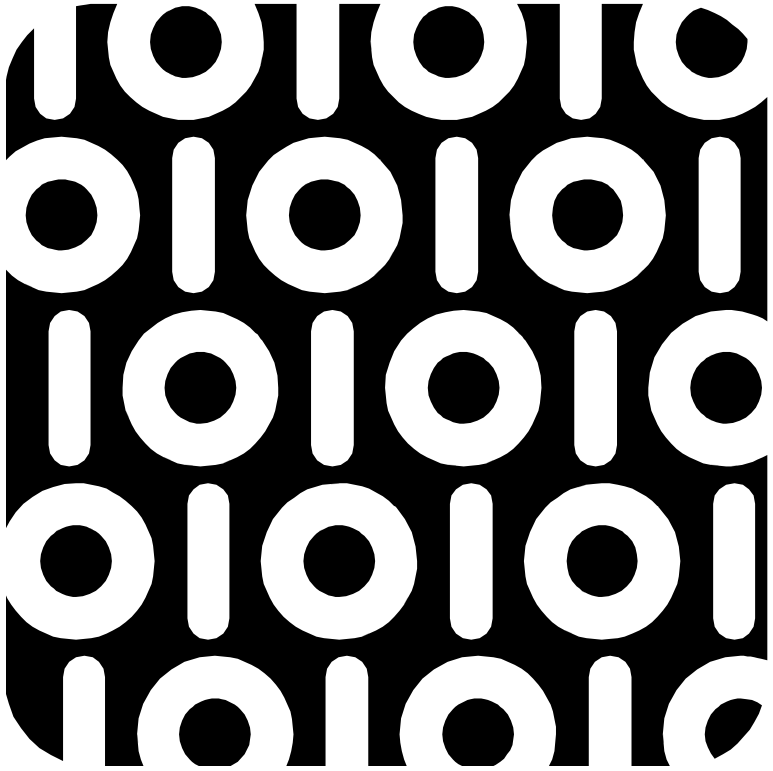
THE DATA

- Has been cleaned beforehand!
- No missing values
- 3000 tuples
- 22 attributes
  - Mix of nominal and binary
- The mushroom data can be found in the agaricus-lepiotadata.txt file in the java-project
- An explanation found in the agaricus-lepiotaexplanation.txt file is also included



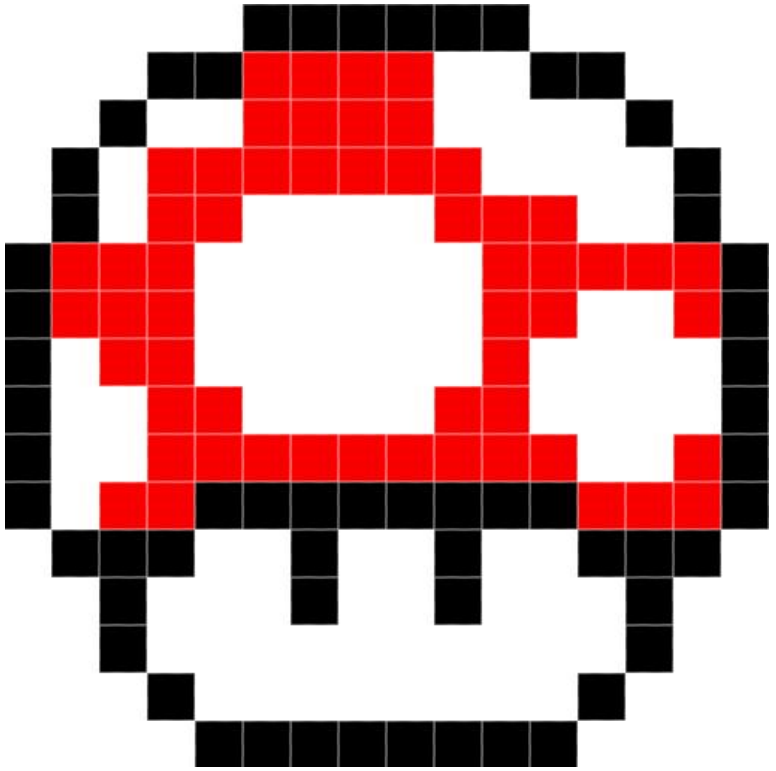
CODE PROVIDED

- Mushroom class used to store data for each mushroom in data.
  - Utilizes a lot of enums!
    - Situated in the ”data” java package.
- Data loading and conversion to Mushroom-objects
  - Done by the CSVFileReader and DataManager class.
- Main-class contains Main-function
  - Currently it calls the LoadData method of the DataLoader which returns an ArrayList of all Mushroom objects loaded in from the data file.



HAVE FUN!

THANK YOU FOR LISTENING!



HELP SLIDES

ITERATING THROUGH ENUM VALUES

EnumSet

```
for(Cap_Shape shap : EnumSet.allOf(Cap_Shape.class))
{
    System.out.println(shap.toString());
}
```

Values()

```
for(Cap_Shape shap : Cap_Shape.values())
{
    System.out.println(shap.toString());
}
```