# Lab – Classification & Prediction #1

*Data Mining, Spring 2016*

IT UNIVERSITY OF COPENHAGEN

*Follow Up from last week*

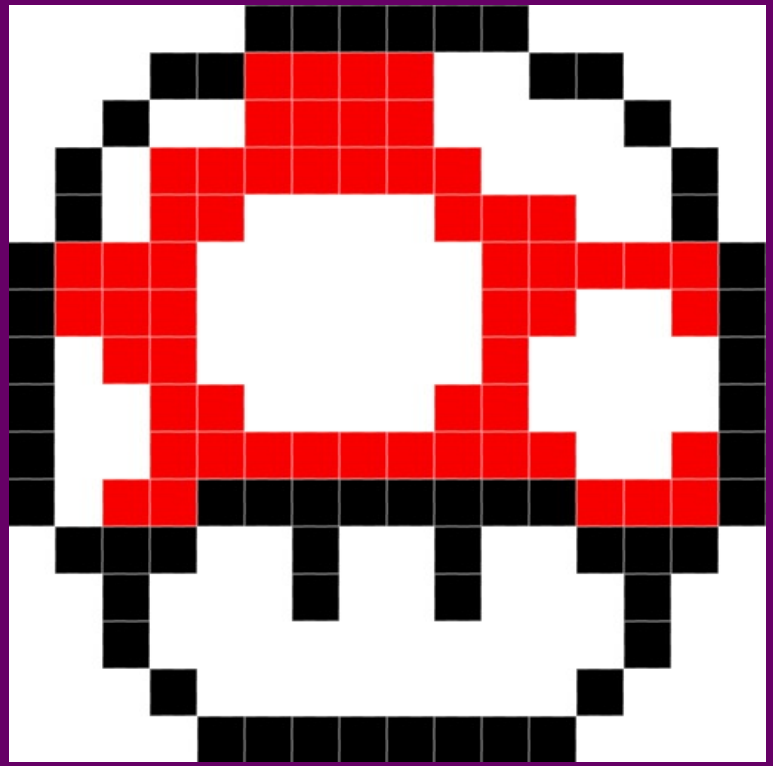
# Follow up from last week

- How did it go in the pre-processing lab last week?
- Important: You are not required to pre-process all the attributes from the questionnaire
  - Select about 4-5 attributes to focus on for the mandatory assignment and pre-process/clean them
- Instead of working on today's lab you can continue working on last weeks lab
  - TAs are more than happy to answer questions concerning last weeks lab!
- Remember to use Q&A Forum on learnIT if you have any questions about anything really (Java programming, course, mandatory assignment, group project, lectures etc.)

*Today's Lab: Classification and Prediction #1*

# 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.

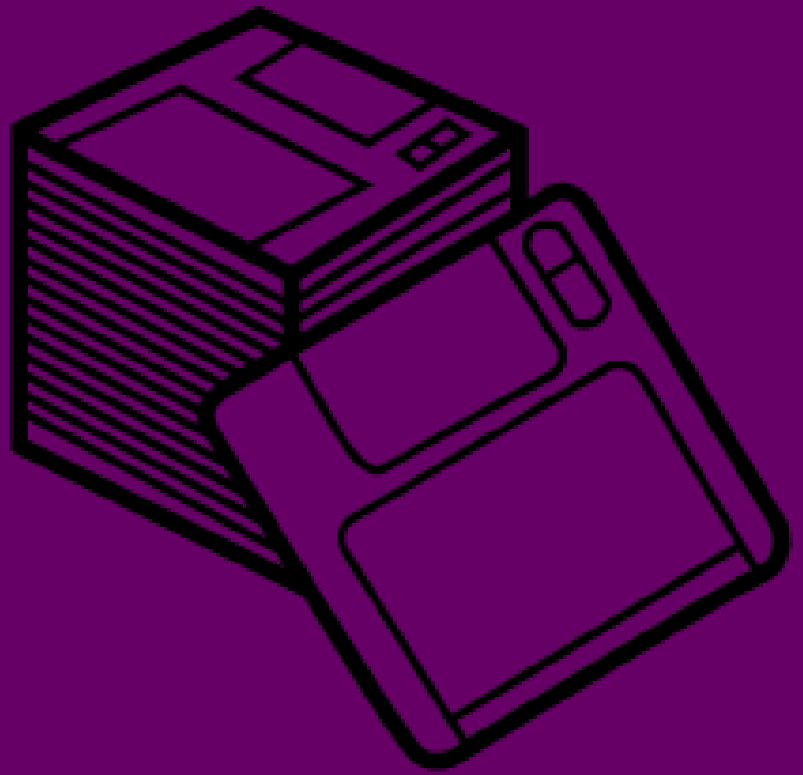


# Overview of today's Lab

- 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.
- Other helper methods have been made, which you may or may not find helpful when working with ID3.

*Thanks for listening!*

*Help Slides*

# Iterating through Enum Values

- Enum Set

```
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());
}
```