



# Dummy Data Analyzer

ZhiZhuo Zhang

# Motivation

- Expert is expensive
- Expert is not much better Dummy (person without domain knowledge)
- Computer performance should be similar to Dummy
- Table Data is very common

Lib	VenousInfil	Recurrence	OverallSurv	DiseaseFree	Status	EarlyRecurr
ID	C	C	N	N	C	C
11	presence	NO	52.13	52.13	alive	NO
13	presence	NO	52.63	52.63	alive	NO
14	absence	NO	53.63	53.63	alive	NO
17	presence	YES	53.63	10.75	alive	early
19	presence	NO	50.7	50.7	alive	NO
21	presence	NO	52.63	52.63	alive	NO
22	absence	YES	55.17	13.51	alive	late
23	presence	NO	54.67	54.67	alive	NO

# Aim

- Design a system for Dummy User
- Can handle missing data
- Can handle hetero-data types
- Can automatically perform what human Dummy do!
- Input: a set of table files
- Output: knowledge from the input

# Scope

- The system can be treated as “Dummy killer” but not “Expert killer”
- It is ok for the system to report nothing
- It is ok for the system to report trivial thing but true for common sense

# Framework

Human Design  
Criteria

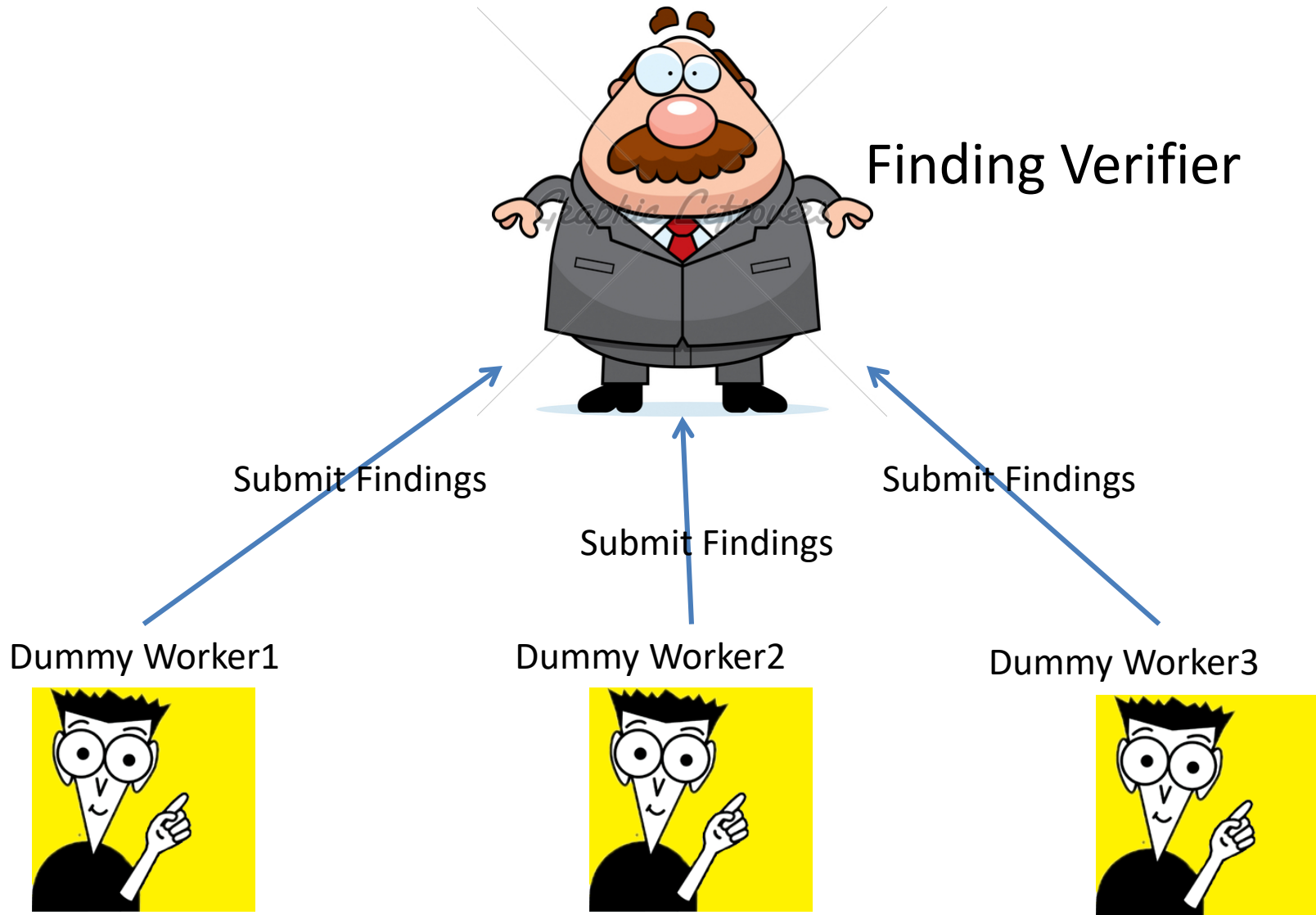


Try all  
combinations



Control  
False Positive

# Framework



# Dummy Worker

- Given the criteria
  - Support, Confidence, P-value,...
- Each Dummy will dig out the knowledge fitting the criteria in his own way



# Finding Verifier



- Ensure Dummies are not mal-function
- Ensure the findings are not false positives
- Techniques:
  - Submit faking data to dummies to estimate false positive rate (FDR)
  - Compare the faking data result to true data result
  - Check if the finding is supported in multiple views



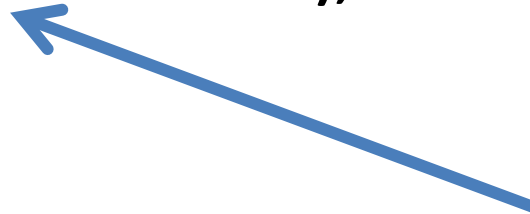
# Implementation

- Java
- Weka

```
public interface AbstractDummy {
```

```
    public List<DummyFinding>  
    DigKnowledge(List<Instances> datasets);
```

```
}
```



Weka Data Object

# DummyFinding Class

```
public class DummyFinding {
```

- **int support;**
  - **double confidence;**
  - **double pvalue;**
  - **boolean isCrossTable;**
  - **HashSet<String> FeatureName;**
  - **String description;**
  - **String DummyName;**
- ```
}
```

# SingleFeatureDummy Class

- Try all the combinations for pairs of column
- Given two columns:
  - Both are numerical
    - Compute the spearman correlation
    - Compute P-value
  - Any one is nominal
    - Use the other one as single feature to do classification (C4.5 decision tree)
    - Compute AUC
- Output DummyFindings satisfying the criteria

# DummyWrapper

- High level Dummy will apply the lower level Dummy to do the findings
- Try different way to preprocess data
- Try to filter results from low level Dummy

```
public abstract class DummyWrapper implements AbstractDummy {  
    AbstractDummy workerDummy;  
    public DummyWrapper(AbstractDummy workerDummy) {  
        super();  
        this.workerDummy = workerDummy;  
    }  
}
```

# MissingValueHanlder Class

- Is DummyWrapper
- Try different ways to fill in the missing data
  - Fill in mean value for missing numerical data
  - Fill in most frequent value for missing nominal data
  - Use EM to estimate the missing data
  - Treat missing value as another nominal value or extreme numerical value

# ShuffleVerifier

- Is DummyWrapper
- Make fake  $n$  datasets by shuffling the value in the original dataset
- Compute FDR by  $\#fake/n*\#true$
- Ignore all the Dummy Findings if  $FDR > 0.05$
- Filter the true data finding no better the fake finding with the same feature set.

# Current Version

Read Table Files



Remove Useless Columns




ShuffleVerifier

MissingValueHandler

SingleFeatureDummy

# Code SVN

 **dummydataanalyzer**  
Analyze Clinical Data or Survey Data for Dummy

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) **Source** [Administer](#)

[Checkout](#) **Browse** [Changes](#)   [Request code review](#)

Source path: svn/

| Directories | Filename                                    |
|-------------|---------------------------------------------|
| ▼ svn       | <a href="#">AbstractDummy.java</a>          |
| ▼ DDA       | <a href="#">Criteria.java</a>               |
| .settings   |                                             |
| src         | <a href="#">DDAmain.java</a>                |
| branches    | <a href="#">DummyFinding.java</a>           |
| tags        | <a href="#">DummyWrapper.java</a>           |
| trunk       | <a href="#">MissingValueHandler.java</a>    |
| wiki        | <a href="#">RemoveUseless_ignoreID.java</a> |
|             | <a href="#">ShuffleVerifier.java</a>        |
|             | <a href="#">SingleFeatureDummy.java</a>     |
|             | <a href="#">Tab2Arff.java</a>               |
|             | <a href="#">common.java</a>                 |

svn checkout ***https://dummydataanalyzer.googlecode.com/svn/DDA***



# Liver Cancer Example

- 88 patients
- Table 1: Clinical Data (Doctor)
  - 18 Features
- Table 2: Gene Expression Data (MicroArray)
  - 228 Features
- Table 3: HBV Integration Data (Sequencing)
  - 40 Features

# Result

SingleFeatureDummy find out:[Gene.MLL4\_log2, HBV.LABEL\_MLL4]

J48 Classification

Support:88 Confidence:0.9615384615384616 pvalue:0.0

SingleFeatureDummy find out:[HBV.HBV\_MLL4, Gene.MLL4\_log2]

J48 Classification

Support:88 Confidence:0.9718706047819972 pvalue:0.0

SingleFeatureDummy find out:[GeneExpression.AFP\_log2,  
Clinical.AFPSerum] Spearman Correlation

Support:88 Confidence:0.8007715344429016 pvalue:0.0

# Future Works

- Implement ClusteringDummy
- Implement more data type support:
  - Fasta, Time-series, Text
- Include more Dummy idea.....

**Dummy Developer is needed!**