Introduction

Imagine a biology student who needs to analyze and evaluate data measured. in a lab. For her, as a domain expert in her field, the business and the data understanding phases of the data mining process are not a problem. The main **challenge** for her is to preprocess and analyze the data and gain useful knowledge from it.

Our work is aimed at **helping people** to analyze their data in a simple and user friendly way with no previous knowledge of data analysis nor **data**mining.

Problem

How to guide a **non-technical** user through the whole datamining process?

The solution should be:

- user friendly
- easy to use
- fast (no long waitings for results)
- accurate

| ac | curate | bigm® | <pre> openML beta </pre> | WEKA The University of Waikato |
|----|---------------------------------|--------------|--------------------------|--------------------------------------|
| | No installation | \checkmark | \checkmark | |
| | Easy-to-use | ++++ | +++ | + |
| | No technical skills required | \checkmark | √* | |
| | Accurate results | | | \checkmark |
| | Free | | \checkmark | \checkmark |

[†] Pros and Cons of existing solutions applicable to this problem

Web based data-mining assistant Štefan Bocko, Tomáš Horváth Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice

What do these data **say**?



WEKA ()) rapidminer

Our approach

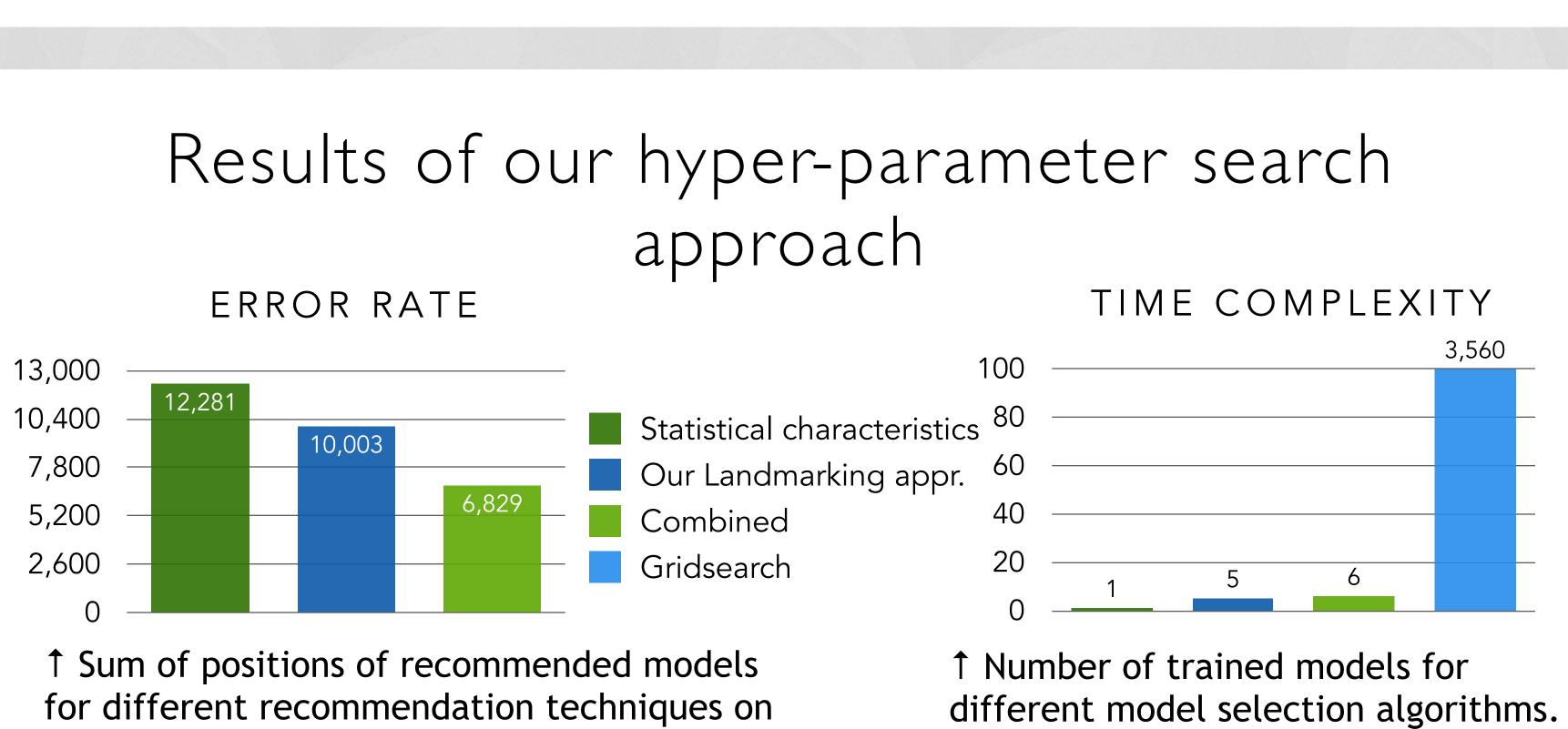
We **designed** and **implemented** a web based application which significantly simplifies data-mining processes.

Discovered challenges:

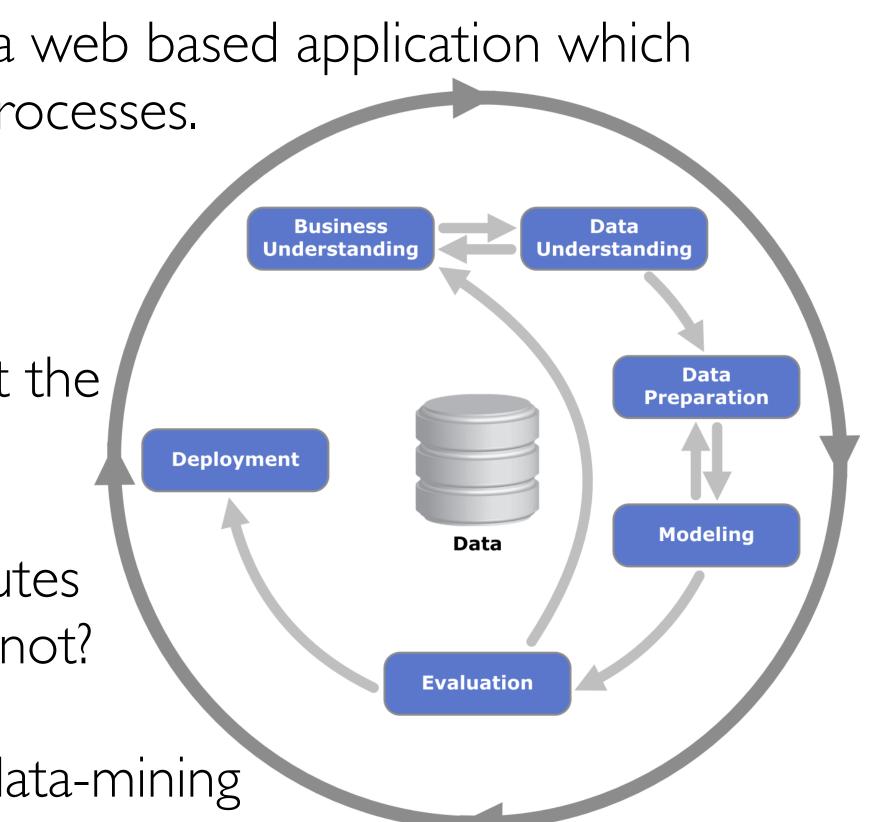
- Data understanding
- How to automatically find out the character of data?
- Data preparation
- How to find out which attributes are important and which are not?
- Modelling
- How to choose the correct data-mining model?
- How to choose the best hyperparameters for that model?
- or days?

Our proposal:

- Hundreds of differently preprocessed data files using computational cluster for fast and reliable results.
- Custom Meta-learning algorithm with Landmarking features speeds up the combined model and hyperparameter selection.



52 datasets (183 484 models total).



• How to get these results within a few seconds instead of hours

• Automatic **conversational system** pre-processes the input data and generates questions for user to determine the further steps.

| | \geq \Box \bigtriangleup \Box \Box | localhost:80 |
|--|--|-----------------------|
| Wins | ton My datasets | |
| Iri | S | |
| E Ho | ome 🔍 Analyze 🛛 🗹 Edit 🗶 Delete | ĺ |
| | Title | Iris |
| | Data File | iris.csv |
| | Number Of Instances | 150 |
| | Description | Iris dataset from UCI |
| # | Title | |
| 1 | sepal length in cm | |
| 2 | sepal width in cm | |
| 1 | and the second second | |
| 3 | petal length in cm | |
| | petal length in cm | |
| 3 | | |
| 3 4 5 Top | petal width in cm class analyzes | |
| 3 4 5 Top | petal width in cm class analyzes Data File | Number Of A |
| 3 4 5 Top # 1 | petal width in cm class analyzes Data File Iris-analysis_1058.arff | 5 |
| 3 4 5 Top # 1 2 | petal width in cm class analyzes Data File Iris-analysis_1058.arff Iris-analysis_1014.arff | 5 |
| 3 4 5 Top # 1 2 3 | petal width in cm class analyzes Data File Iris-analysis_1058.arff Iris-analysis_1014.arff Iris-analysis_1050.arff | 5 5 5 |
| 3 4 5 Top # 1 2 | petal width in cm class analyzes Data File Iris-analysis_1058.arff Iris-analysis_1014.arff | 5 |

prototype.

We will support **regression** and **pattern mining** techniques soon. Planned public release of this software is on 1st. of August 2015. Using the **Meta-learning** we were able to speed up the model recommendation time. For this purpose we combined our Landmarking approach with statistical characteristic approach presented by R. Neruda et al.

- 80-200-1062-9.

| | Marc | e Results | | | | | | |
|----------------------------|-------------------|----------------------------------|--|--------------------------------|------------------------|---------------|--|--|
| | TIOLE | | | | | | | |
| 080/winston/dataset/show/2 | 25 Č 仓 司 | | localhost:8080/winston/svm | nResult/show/6048 | Ċ | | | |
| | About Log ou | Marine Marine | | | | About Log out | | |
| | | | | | | | | |
| | | Model detail | | | | | | |
| | | < Back | | | | | | |
| | | Analysis | Car custom analysis | | | | | |
| | | Model type | Support vector machines | | | | | |
| | | Kernel | RBFKernel | | | | | |
| | | Complexity Constant | 1 | | | | | |
| repository. | | Gamma | 0.1 | | | | | |
| | | Rmse | 0.764 | | | | | |
| | | Mean Absolute Error | 0.583 | | | | | |
| | Туре | Correctly classified instances | 5 | | | | | |
| | numeric | Incorrectly classified instances | 7 | | | | | |
| | numeric | Summary | Correctly Classified Instances5Incorrectly Classified Instances7Kappa statistic-0.3125Mean absolute error0.5833Root mean squared error0.7638 | | 41.6667 % 58.3333 % | | | |
| | numeric | | | | | | | |
| numeric | | | Relative absolute error Root relative squared error | 118.4615 % 153.6993 % 12 | | | | |
| | categorical | | Total Number of Instances | 12 | | | | |
| | | | | | | | | |
| Attributes | Best rmse | 1 Model de | tail overviev | w for | - SVM | algorithm | | |
| | 0.115973054487532 | | | | | | | |
| | 0.115973054487532 | | | | | | | |
| | 0.126970225698421 | | | | | | | |
| | 0.133129641241098 | | | | | | | |
| | 0.146936258710236 | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |



We focused on **classification** problems in our current working

DEMO: <u>http://s.ics.upjs.sk/~sbocko/winston</u>

References

Kazík, O., Pešková, K., Pilát, M., Neruda, R. Combining parameter space search and metalearning for data-dependent computational agent recommendation. I I th International Conference on Machine Learning and Applications (ICMLA 2012): Boca Raton, Florida, USA, 12-15 December 2012. 2 volumes. ISBN 9781467346511 2. Berka, P. Dobývání znalostí z databází. Vyd. I. Praha: Academia, 2003, 366 s. ISBN

3. Vilalta, R., Giruard-CARRIER, C., BRAZDIL, P. SOARES, C. Using Meta-Learning to Support Data Mining. International Journal of Computer Science & Applications, Vol. I, No. I, p. 31–45. 2004.

