

Predicting the Execution Time of Workflow Activities Based on Their Input Features

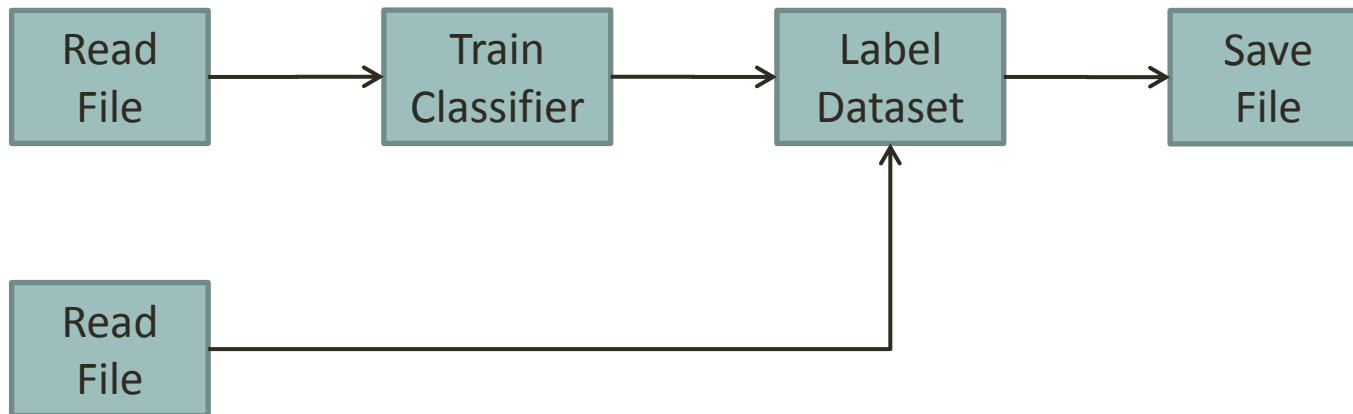
Tudor Miu and Paolo Missier
Newcastle University, UK

In this presentation

- Workflows
 - Execution time prediction
 - Rationale
 - Existing approaches
- Input-based execution time prediction
 - Rationale
 - Method
 - Experiment
 - Results
- Future work

Example of a workflow

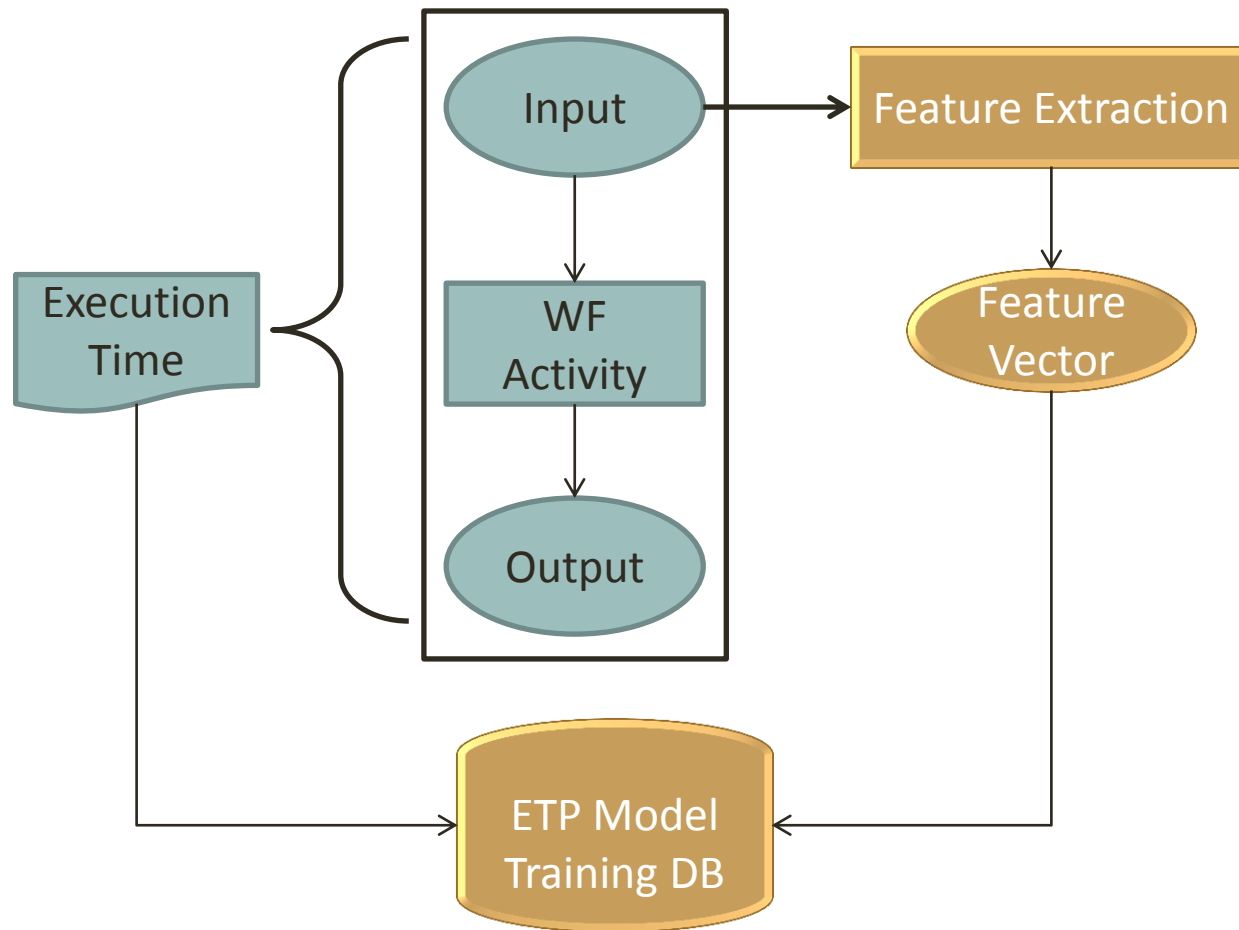
- Machine Learning – generic classification workflow



Execution time prediction (ETP)

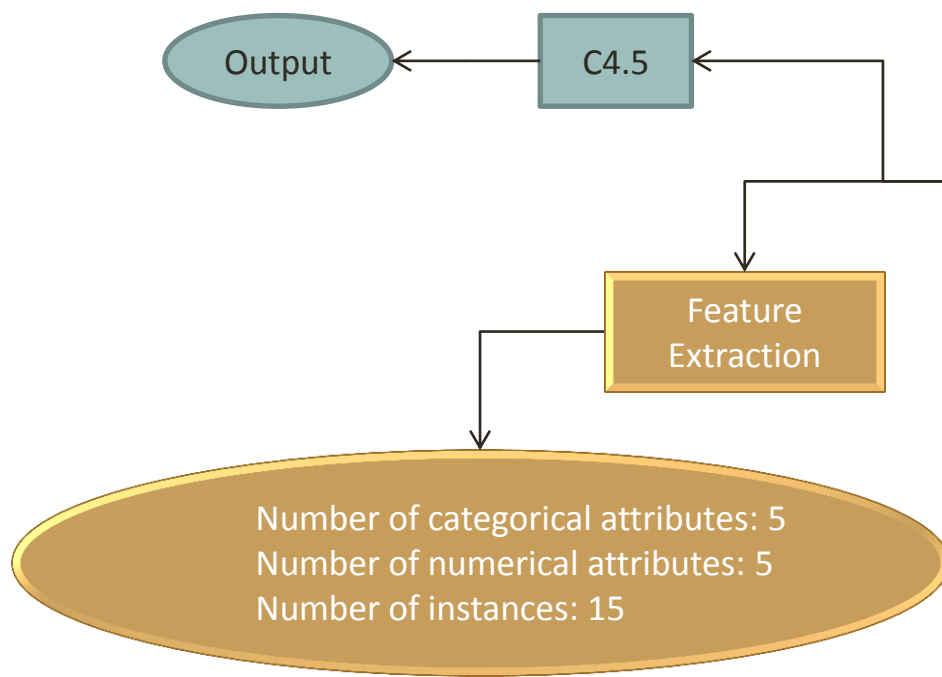
- Why?
 - Scheduling in large computational infrastructures
 - Cost estimation on 'rented' platforms, e.g. cloud
- How it is usually done
 - Time-series analysis
 - Incidental features – group jobs with similar attributes
- We propose to predict execution times for jobs whose execution time is determined:
 - by the input
 - in less incidental ways

Input-based execution time prediction (ETP) – creating an execution history



Feature extraction example: C4.5

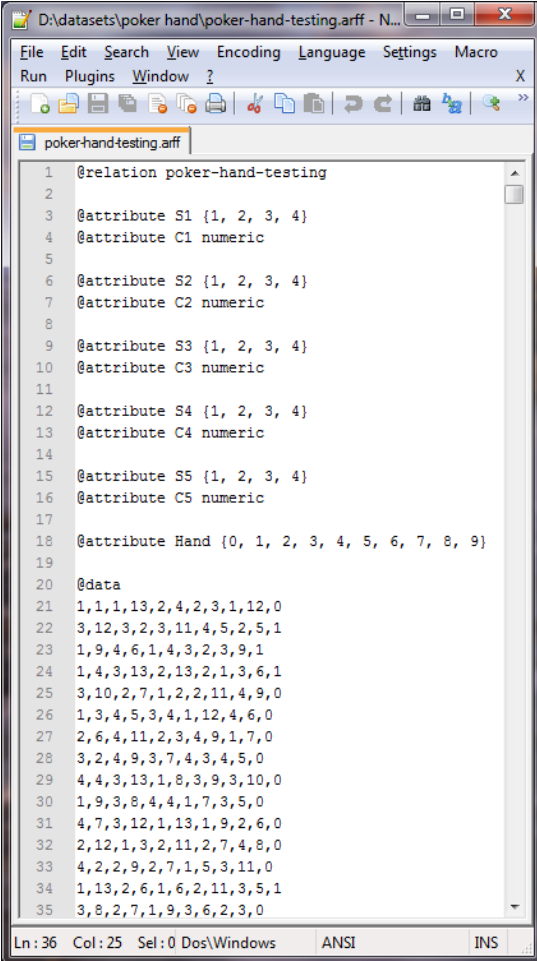
- Measurable quantity
- Summary



```
D:\datasets\poker hand\poker-hand-testing.arff - N...
File Edit Search View Encoding Language Settings Macro
Run Plugins Window ?
poker-hand-testing.arff
1 @relation poker-hand-testing
2
3 @attribute S1 {1, 2, 3, 4}
4 @attribute C1 numeric
5
6 @attribute S2 {1, 2, 3, 4}
7 @attribute C2 numeric
8
9 @attribute S3 {1, 2, 3, 4}
10 @attribute C3 numeric
11
12 @attribute S4 {1, 2, 3, 4}
13 @attribute C4 numeric
14
15 @attribute S5 {1, 2, 3, 4}
16 @attribute C5 numeric
17
18 @attribute Hand {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
19
20 @data
21 1,1,1,1,13,2,4,2,3,1,12,0
22 3,12,3,2,3,11,4,5,2,5,1
23 1,9,4,6,1,4,3,2,3,9,1
24 1,4,3,13,2,13,2,1,3,6,1
25 3,10,2,7,1,2,2,11,4,9,0
26 1,3,4,5,3,4,1,12,4,6,0
27 2,6,4,11,2,3,4,9,1,7,0
28 3,2,4,9,3,7,4,3,4,5,0
29 4,4,3,13,1,8,3,9,3,10,0
30 1,9,3,8,4,4,1,7,3,5,0
31 4,7,3,12,1,13,1,9,2,6,0
32 2,12,1,3,2,11,2,7,4,8,0
33 4,2,2,9,2,7,1,5,3,11,0
34 1,13,2,6,1,6,2,11,3,5,1
35 3,8,2,7,1,9,3,6,2,3,0
Ln: 36 Col: 25 Sel: 0 Dos\Windows ANSI INS
```

Features for C4.5: generic input features

- Either *count* variables
 - Number of categorical attributes
 - Number of numerical attributes
- Or *indicator* variables
 - Attribute included in the input dataset?
 - Input schema description
- Plus *number of instances*



```
D:\datasets\poker hand\poker-hand-testing.arff - N...
File Edit Search View Encoding Language Settings Macro
Run Plugins Window ?
poker-hand-testing.arff
1 @relation poker-hand-testing
2
3 @attribute S1 {1, 2, 3, 4}
4 @attribute C1 numeric
5
6 @attribute S2 {1, 2, 3, 4}
7 @attribute C2 numeric
8
9 @attribute S3 {1, 2, 3, 4}
10 @attribute C3 numeric
11
12 @attribute S4 {1, 2, 3, 4}
13 @attribute C4 numeric
14
15 @attribute S5 {1, 2, 3, 4}
16 @attribute C5 numeric
17
18 @attribute Hand {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
19
20 @data
21 1,1,1,13,2,4,2,3,1,12,0
22 3,12,3,2,3,11,4,5,2,5,1
23 1,9,4,6,1,4,3,2,3,9,1
24 1,4,3,13,2,13,2,1,3,6,1
25 3,10,2,7,1,2,2,11,4,9,0
26 1,3,4,5,3,4,1,12,4,6,0
27 2,6,4,11,2,3,4,9,1,7,0
28 3,2,4,9,3,7,4,3,4,5,0
29 4,4,3,13,1,8,3,9,3,10,0
30 1,9,3,8,4,4,1,7,3,5,0
31 4,7,3,12,1,13,1,9,2,6,0
32 2,12,1,3,2,11,2,7,4,8,0
33 4,2,2,9,2,7,1,5,3,11,0
34 1,13,2,6,1,6,2,11,3,5,1
35 3,8,2,7,1,9,3,6,2,3,0
Ln: 36 Col: 25 Sel: 0 Dos\Windows ANSI INS
```

Features for C4.5: static analysis

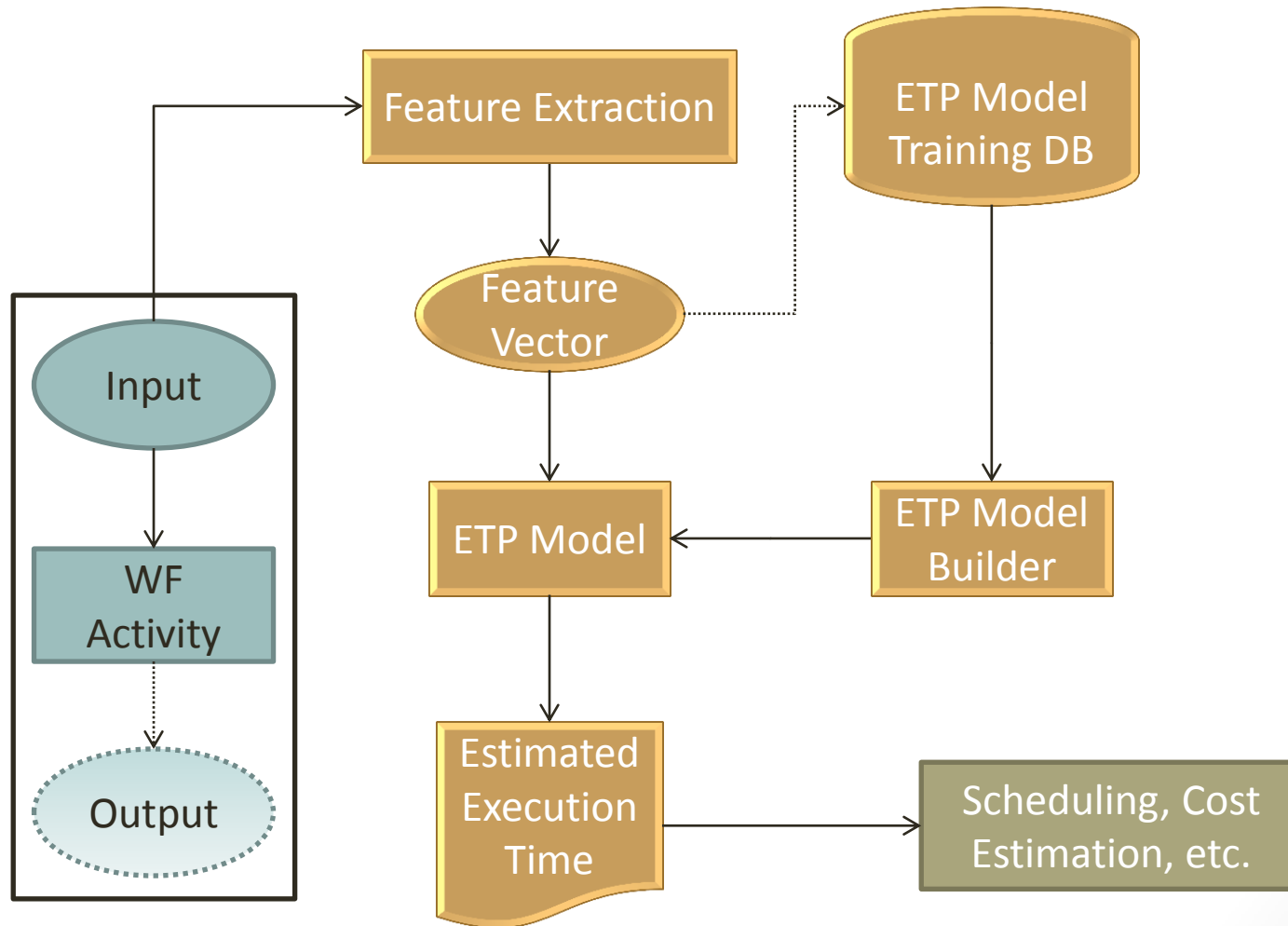
- Analysis of the mechanics of the target algorithm
 - Features
 - N – number of instances
 - p – number of attributes
 - $O(pN \log N) \dots O(pN^2)$
 - Predictions by linear regression

$$\hat{t} = \beta_1 p_{num} N \log N + \beta_2 p_{cat} N \log N + \beta_3 p_{num} N^2 + \beta_4 p_{cat} N^2$$

(plus lower order terms)

- Not always possible/feasible
- Not always best even if available

Input-based ETP – delivering predictions



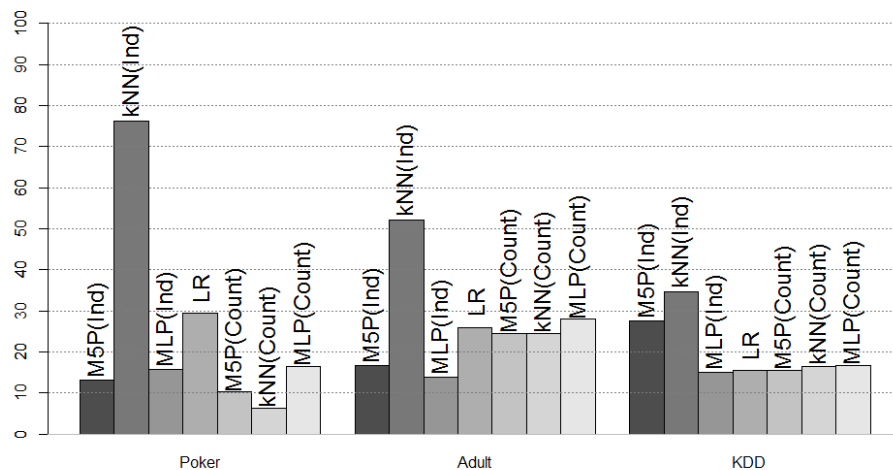
Input-based ETP experiment

- Target algorithm: training phase of C4.5
 - Fixed parameterization
 - Invariable platform (idle computer)
- Three common ML datasets from the UCI ML repository *
- Three common regression models (Weka) used for ETP
 - K Nearest Neighbours (kNN)
 - Multilayer Perceptron (MLP)
 - M5P decision tree (M5P)
- Simulated input variability
 - Random vertical and horizontal sampling
- Measured prediction error
 - For each model/dataset combination
 - As a function of size of execution history

* <http://archive.ics.uci.edu/ml/>

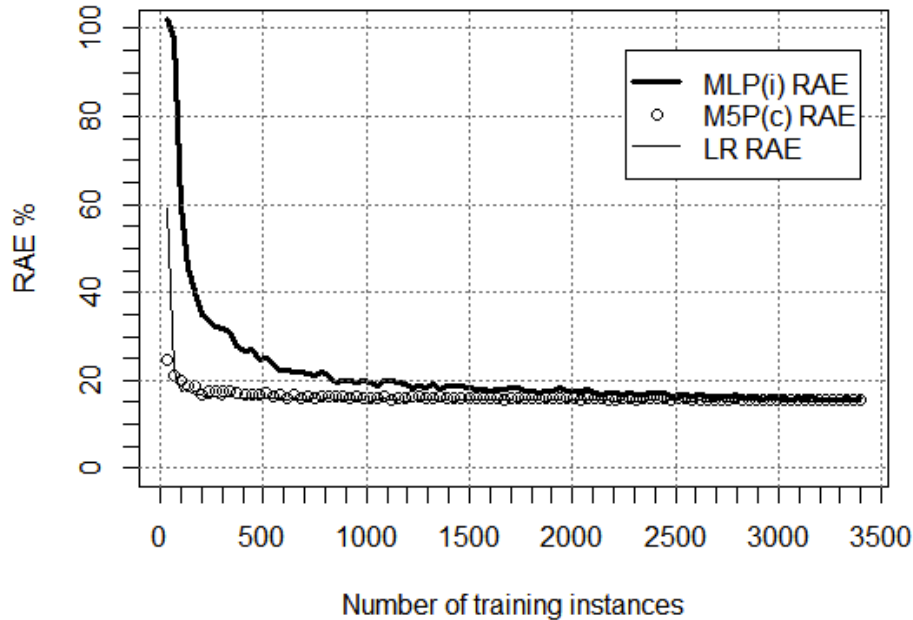
Results

- Relative Absolute Error: $RAE = \frac{\sum_{i=1}^N |a_i - p_i|}{\sum_{i=1}^N |a_i - \bar{a}|}$
- Based on 3400 samples of training sets for C4.5 for each major dataset



* Obtained from 10-fold cross validation

Results



* 10-fold cross validation on 10 samples for each size of the execution time history

- Train model on one major dataset and make predictions on another?
 - Sometimes $RAE > 100\%$

Future work

- Input features complementary to activity parameters
- Input-based ETP may explain some variability
 - Input-based ETP alone is not a panacea
 - Combining input-based ETP with other established methods
- Latent features
 - Detecting change
 - What else?