A TEXT MINING APPROACH FOR EXTRACTING EVENT LOGS FROM UNSTRUCTURED DATA

Sutaraj Dutta
Mark McGann

Supervisor:
Dr Sean McGarraghy, UCD
Colin Melody, Deloitte
Michael Bridges, Deloitte

BUSINESS PROBLEM

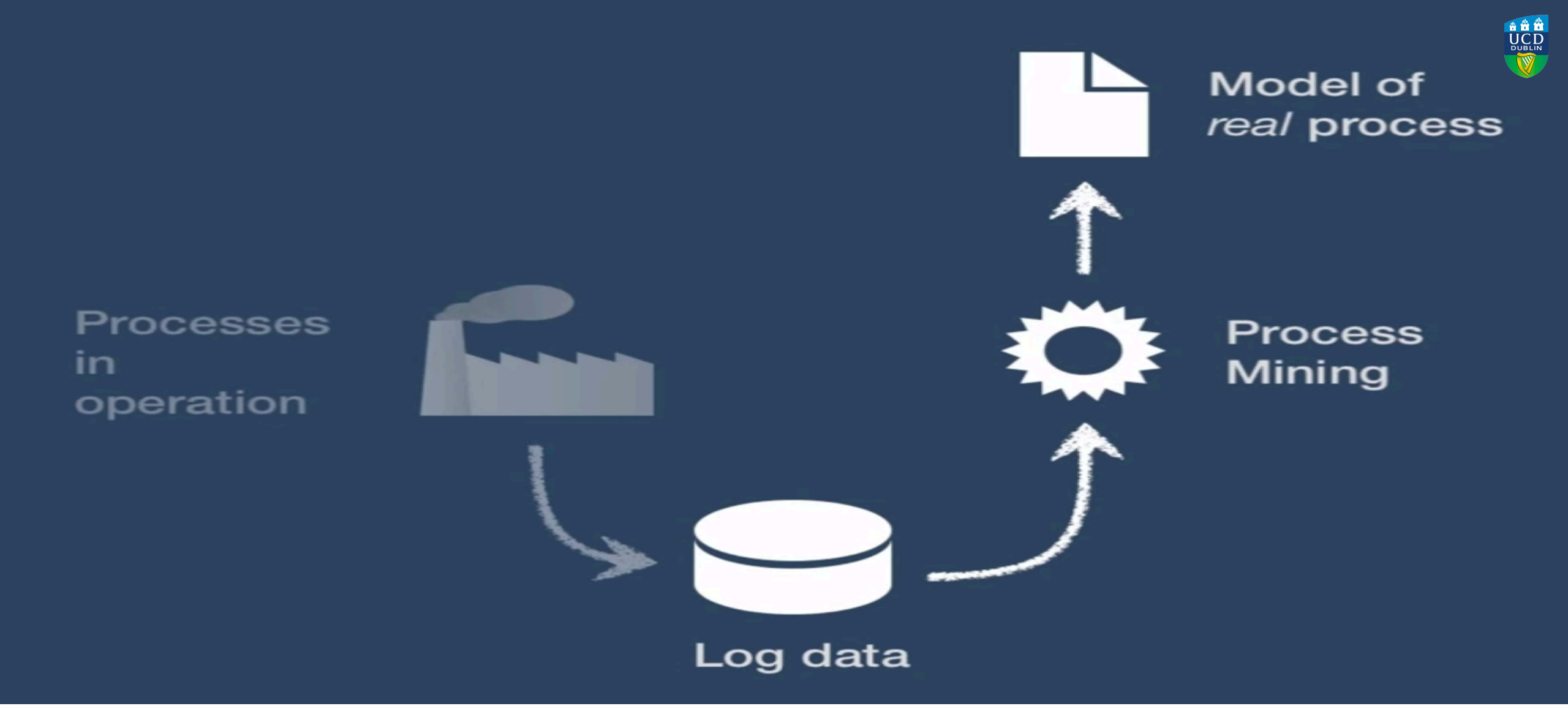
Process Mining

Harvest insight from records of business events to gain deeper understanding of underlying business process

Fundamental Assumption

Process data is recorded in an event log format

An atypical IT infrastructural requirement



PROCESS MINING

VISUALISING THE APPROACH

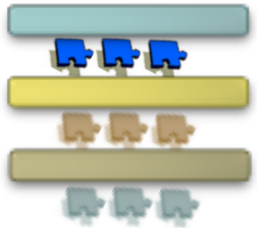
HYBRID ALGORITHM

Document Libraries

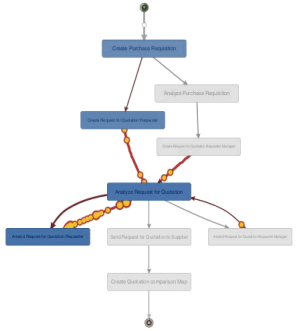
Unstructured Data
SELECT, TRANSFORM, MINE

Extracted Information

Structured Event Log



Process Mining



Visualisation of Process Map

METHODOLOGY OVERVIEW

Minimum requirements of Event Log

1. Activity Label
2. Case ID
3. Timestamp

Techniques used

Classification

Information Extraction

ASSUMPTIONS

Assumptions of Feasibility

- All transactions of process are recorded, and in an unstructured manner

- Only email and letter documents are considered

- All documents converted to .txt format

ASSUMPTIONS

Data Assumptions

- Every document contains at least ONE Case ID and Timestamp
- Each document records a SINGLE activity of process
- A document's content contains term(s) that are unique to each individual activity

ARTIFICIAL DATA GENERATION

Decompose business documents into principal components

- Document Type
- Document Contents
- Activity-specific words and phrases
- Log information

ARTIFICIAL DATA GENERATION

Data Requirement

Acquired a real-life event log of a Dutch Financial Institution's Loan Application Process

- 35 Activities
- 13,087 cases
- 262,200 events

ARTIFICIAL DATA GENERATION

Simulating Randomness

Simulate real world scenarios by varying domain characteristics:

- Number of Observations
- Process Complexity

ARTIFICIAL DATA GENERATION

Simulating Randomness

Vary Document Parameters:

- Document Type
- Document Length
- Level of Noise
- Description of Timestamp and Case ID
- Language distribution – certainty of occurrence of keywords

ARTIFICIAL DATA GENERATION

Sampling Activity-Specific Key Words

X the word we select is a discrete random variable with a probability of occurrence p

Generate u ~ U[0,1]

Essentially map the p for each x into subsets of [0,1]

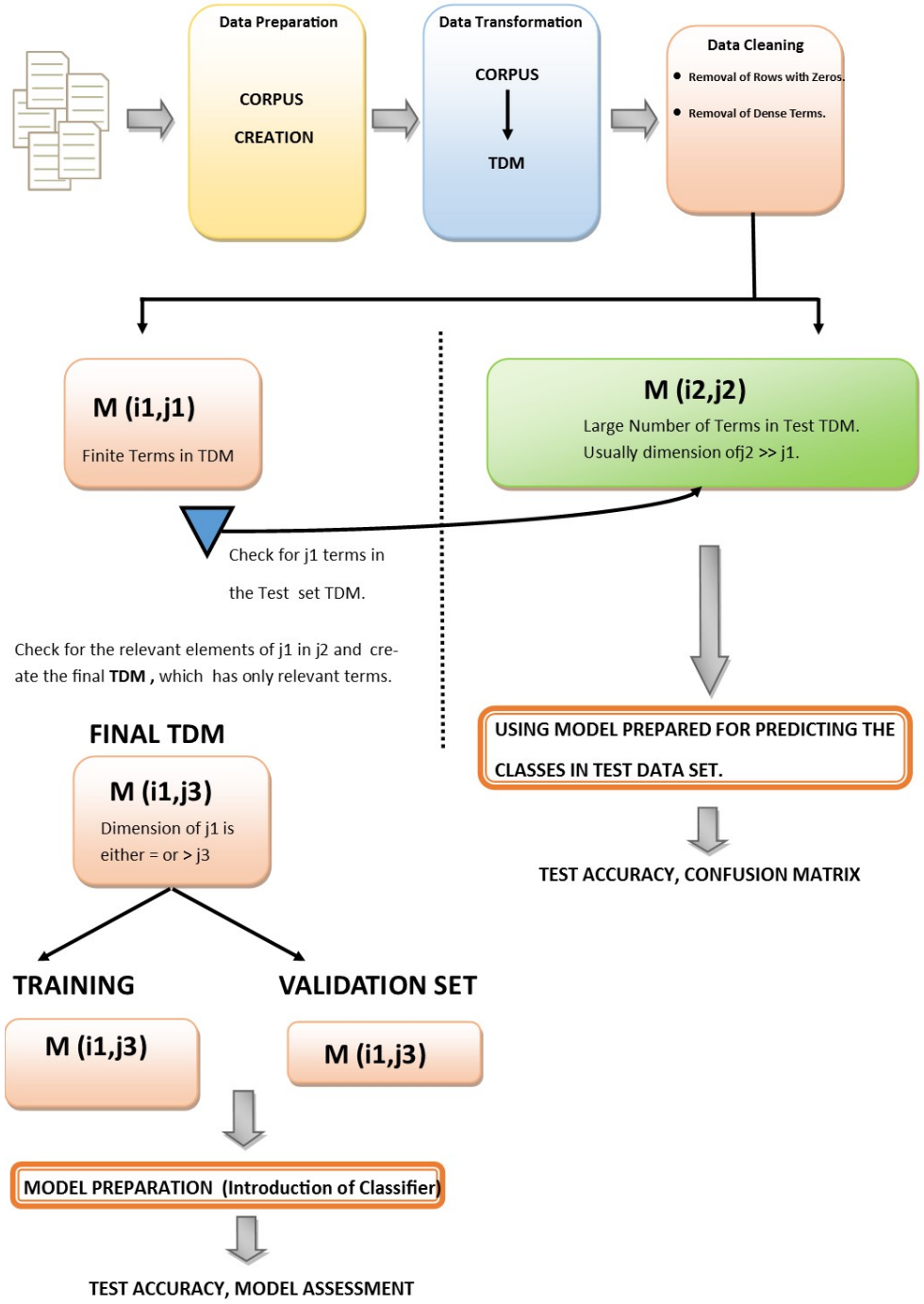
Probability that uniform random value u falls into any range is the length of that range

$u >= p_x$ and $u < p_x + p_{x+1}$

DATA PREPARATION AND CLASSIFICATION

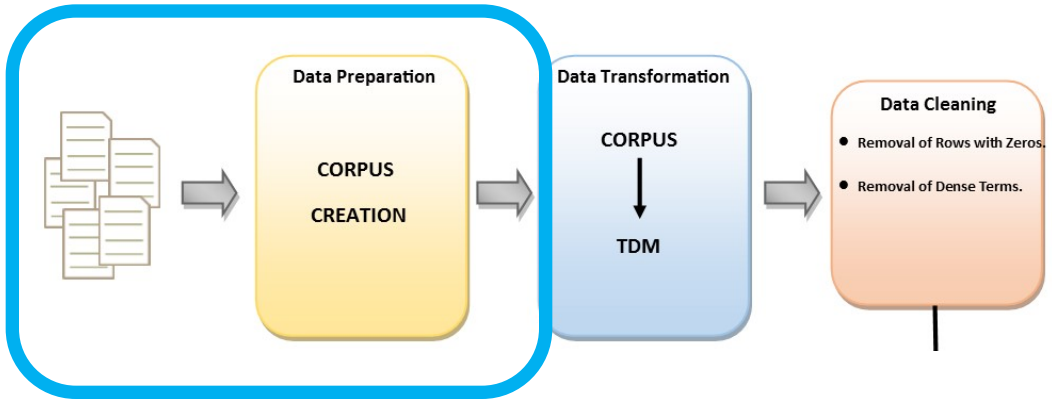
PRE PROCESSING

- Data Preparation
- Data Transformation
- Data Cleaning
- Data Sampling



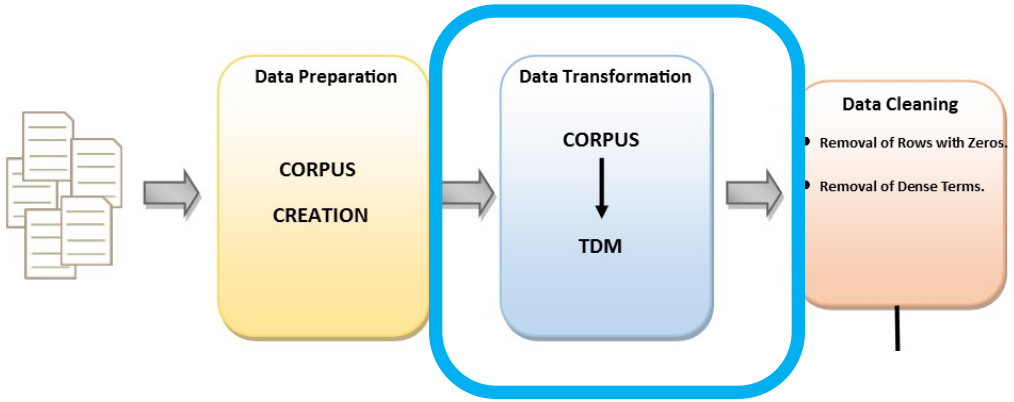
PRE PROCESSING

- Data Preparation
  - Corpus Creation
  - Large and structured sets of text
  - Text Manipulation Techniques



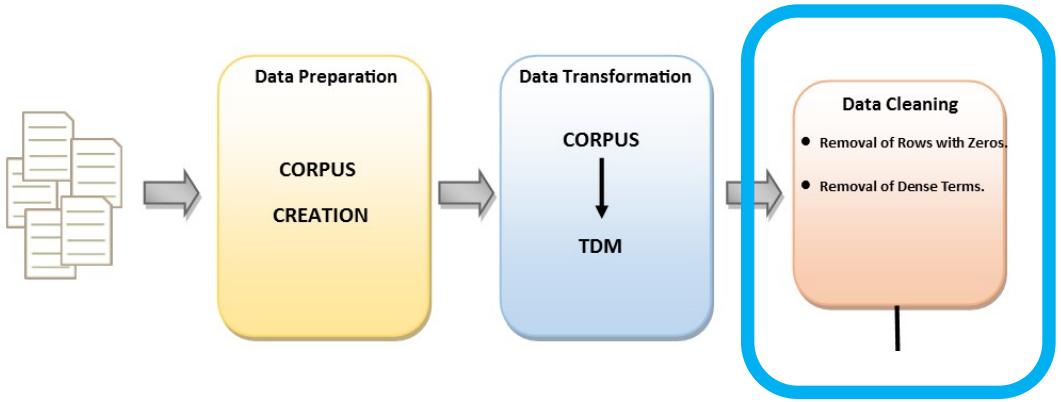
PRE PROCESSING

- Data Transformation
  - Term Frequency (TF)
  - Sparsity
  - Matrix Structure
    - Rows – Documents
    - Columns – Terms
    - Cells - TF



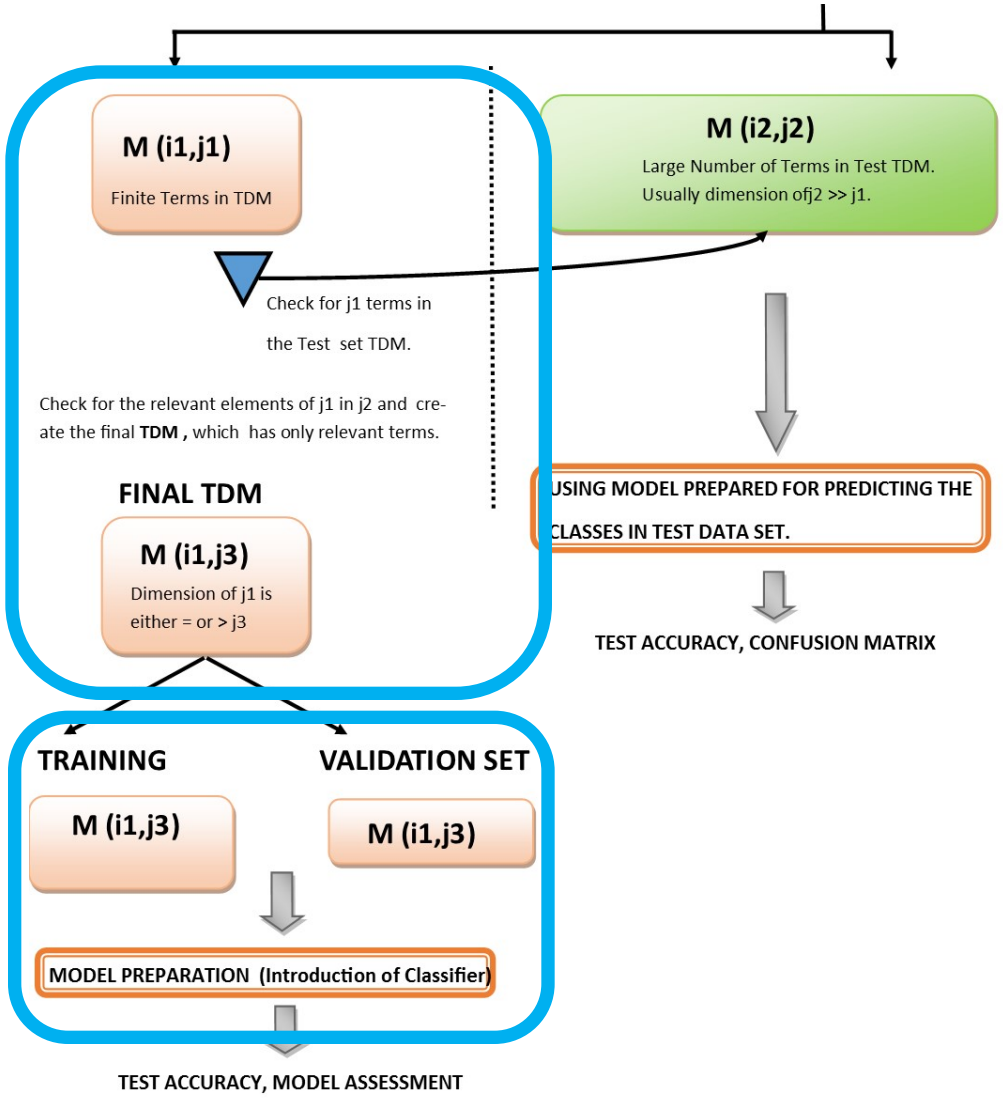
PRE PROCESSING

- Data Cleaning
  - Sparsity (while Transformation)
  - Dense Terms
  - Cumulative zero frequency



PRE PROCESSING

- Data Sampling
  - Standardisation
  - Final TDM
  - Sampling Training Set
    - Training
    - Validation



PROCESSING

Objective is devise generic framework

- No assumptions about domain
- No assumptions about underlying data distributions

PROCESSING

Challenges with High Dimensional Sparse Matrices

- Equidistant data points
- Unreliable parameter estimation
- Poor generalization ability

We therefore considered an ensemble of the benchmark algorithms for document classification.

PROCESSING

Algorithms Considered

- **kNN**
- **Naïve Bayes**
- **Support Vector Machines**

INFORMATION EXTRACTION

INFORMATION EXTRACTION

Two approaches considered

- Artificially Intelligent: Named Entity Recognition & Part Of Speech tagging

- Rule-Based: Regular Expression

INFORMATION EXTRACTION

Problem Space Reduction

- Common factor for Case ID and Timestamp?

- Only consider lines in a document containing numeric values

INFORMATION EXTRACTION

Case ID

- Generic method based on occurrence of ID-related terms
- Weighted Score
- Regular expression for extracting Case ID can be refined with the inclusion of company-specific knowledge of Case ID pattern

INFORMATION EXTRACTION

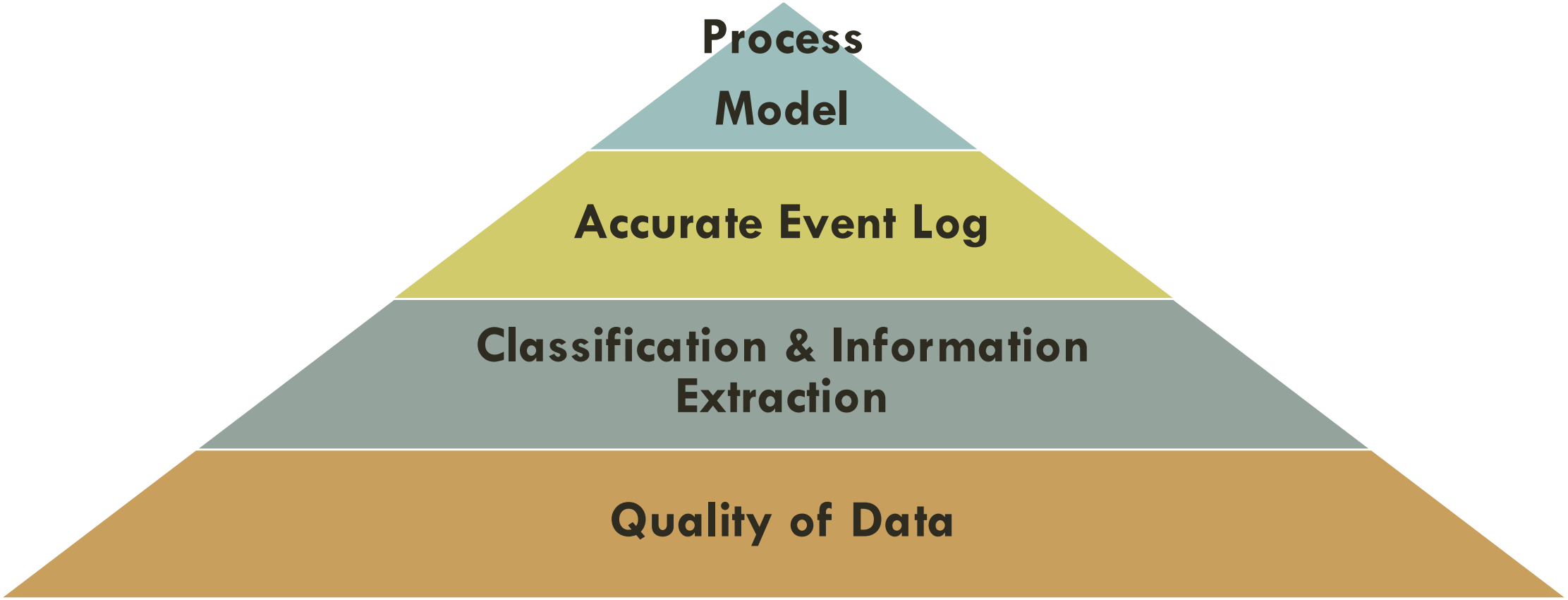
Timestamp

- Domain knowledge of document types – date appears at the top of a document

Regular Expressions to identify dates described by both

- Standard formats
- Natural language

UNDERLYING CRITERIA FOR SUCCESS



RESULTS

RESULTS

Designed Test Cases

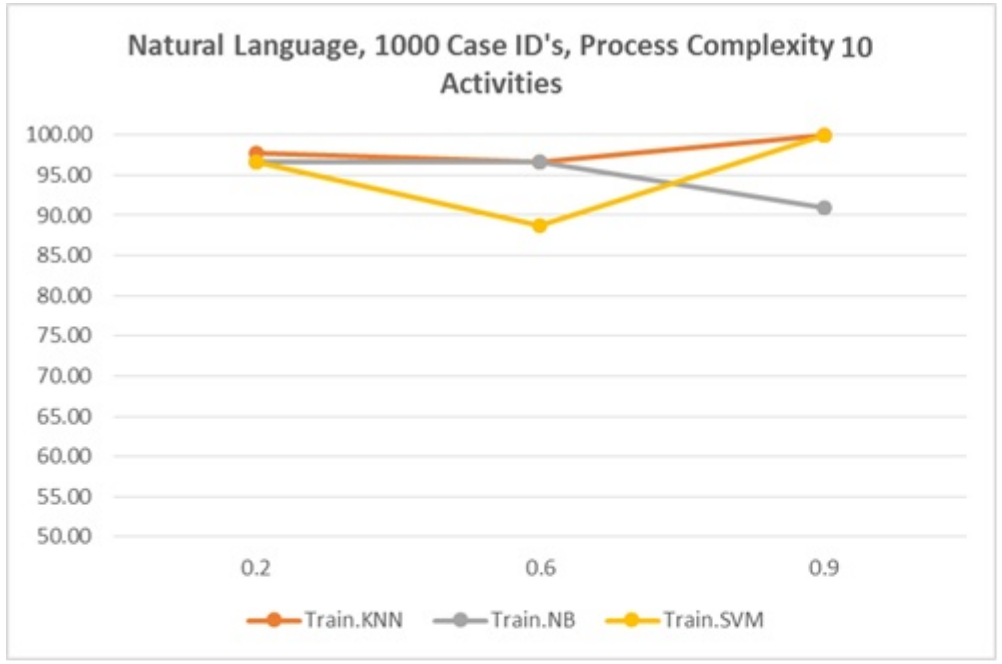
Evaluation of Accuracy with regards to

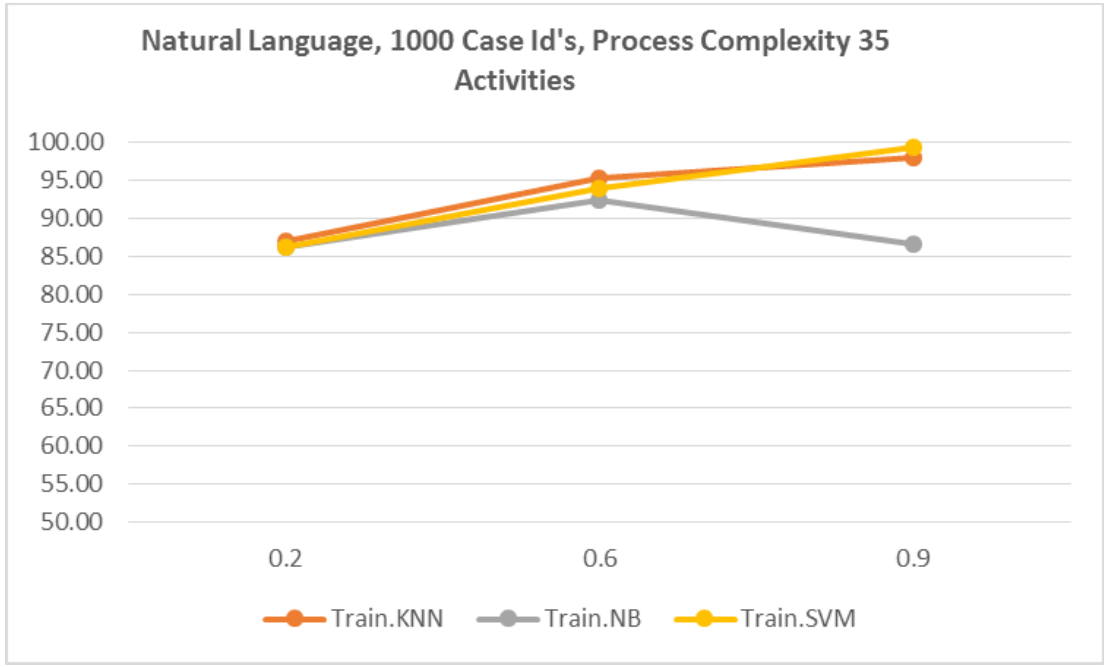
- Sparsity
- Process Complexity
- Language Complexity

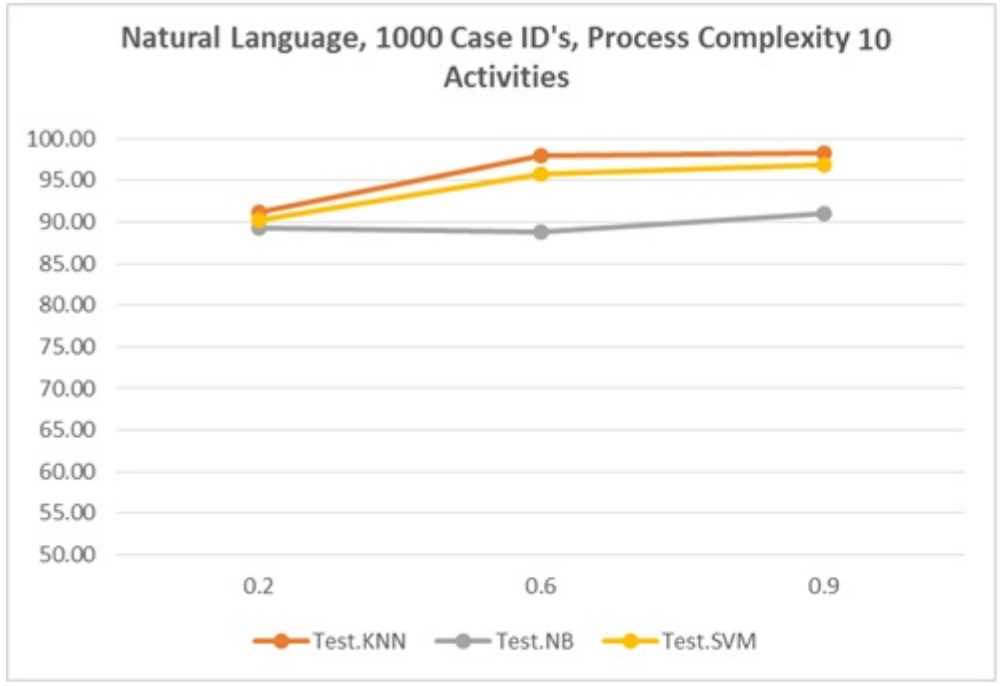
- F1 Score Evaluation

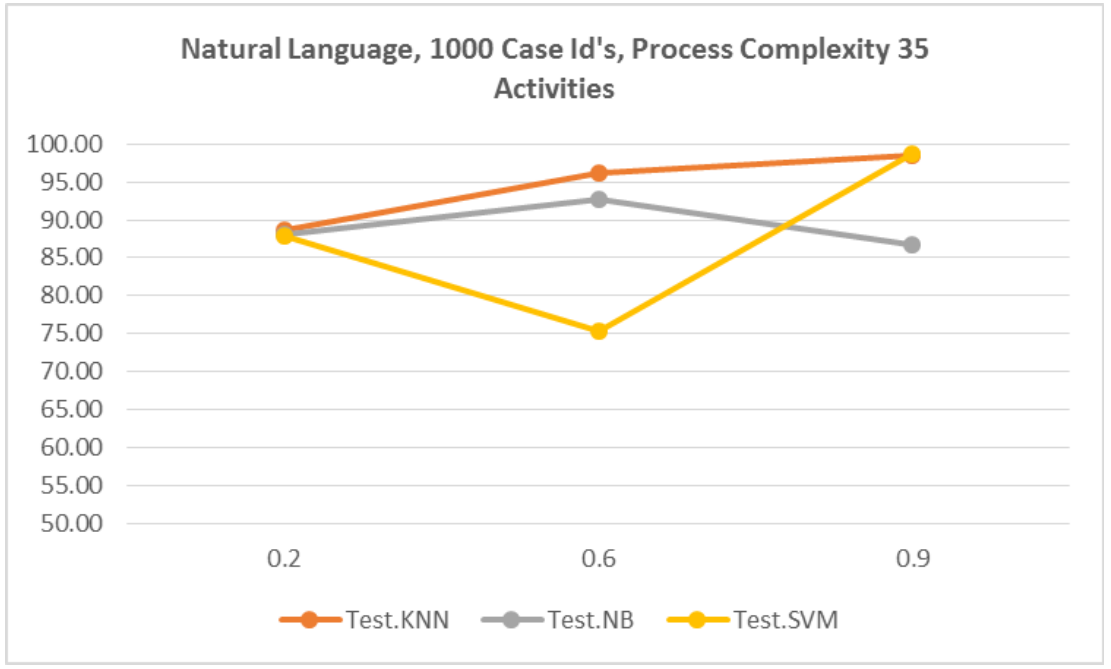
RESULTS / PERFORMANCE

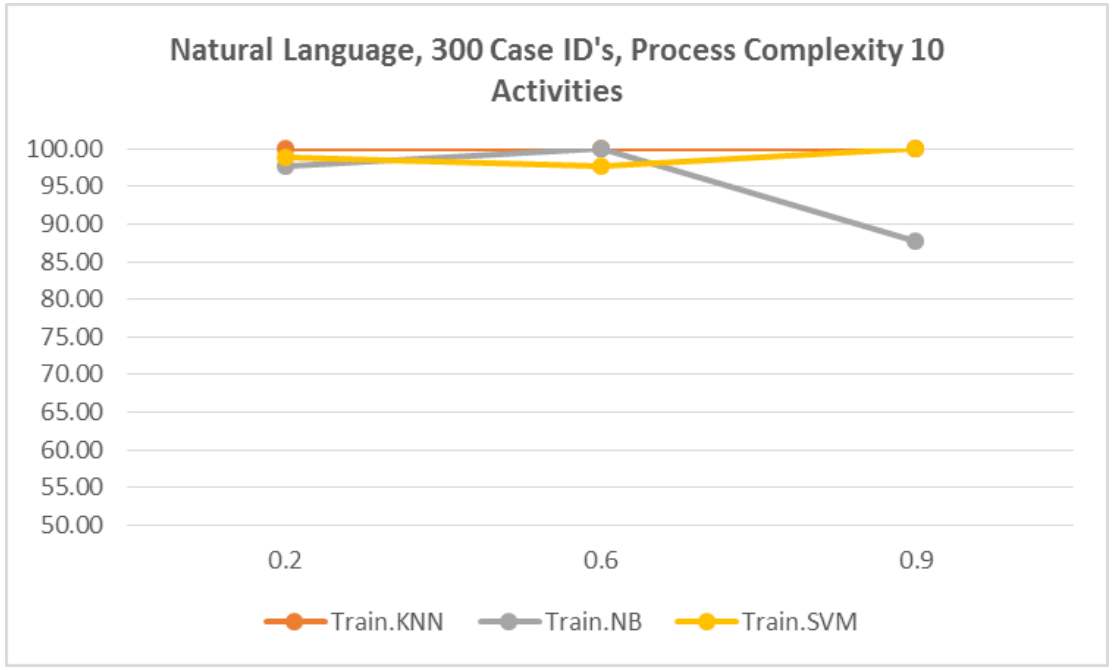
- Sparsity and Complexity
  - Significant correlation of Sparsity and Complexity with Accuracy
  - Low Sparsity v/s High Sparsity

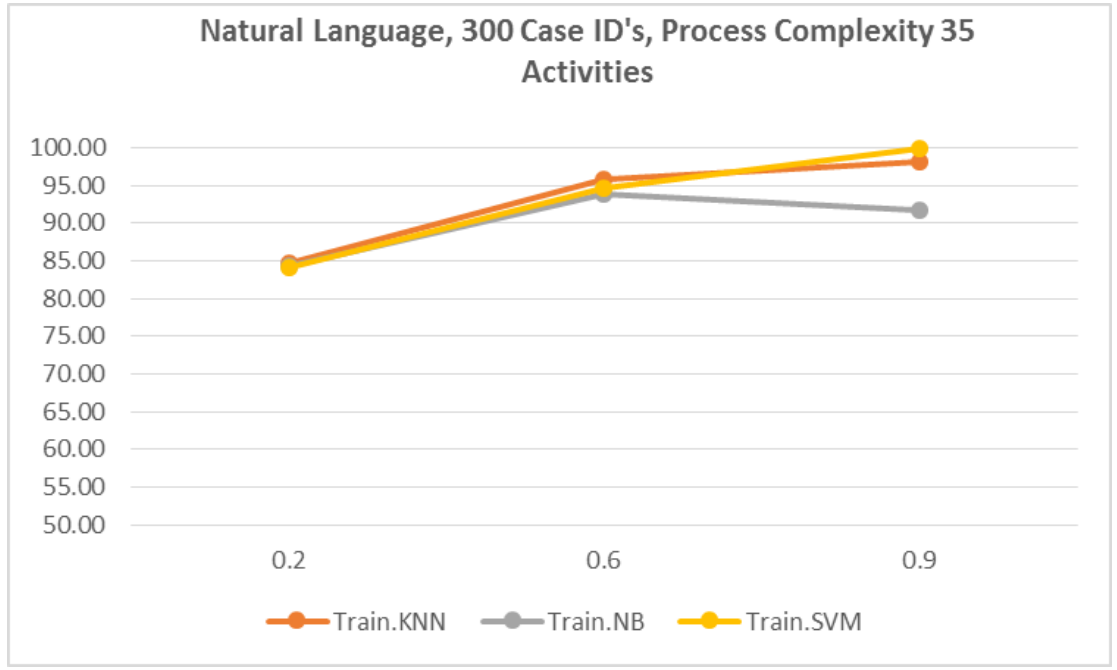

Natural Language, 1000 Case ID's, Process Complexity 10 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Train.KNN
Train.NB
Train.SVM

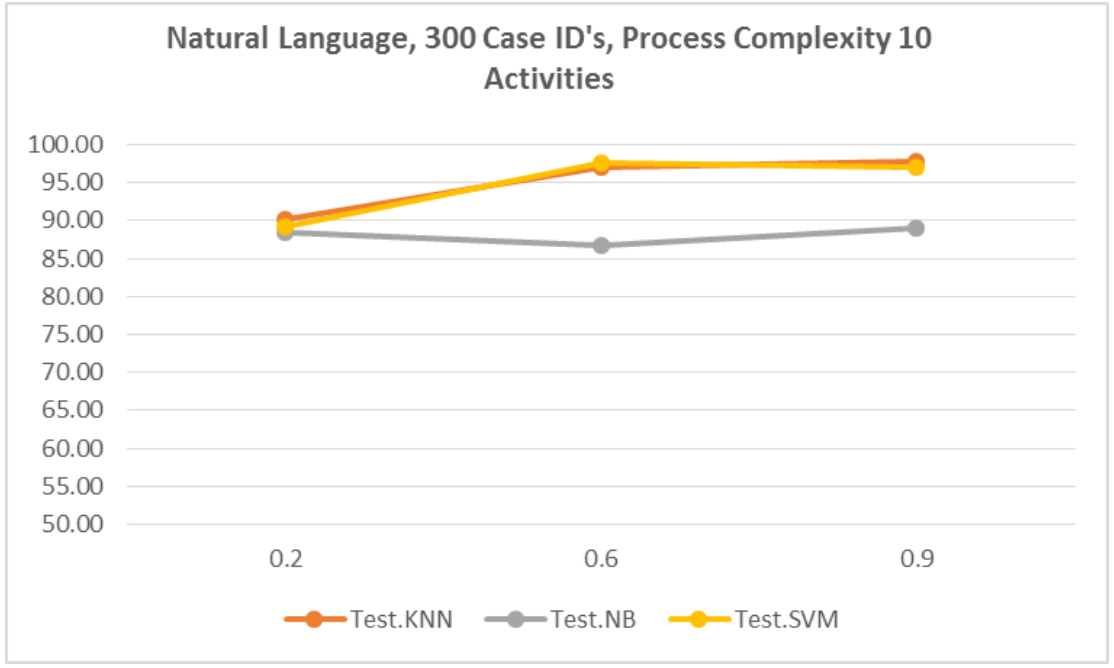

Natural Language, 1000 Case Id's, Process Complexity 35 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Train.KNN
Train.NB
Train.SVM

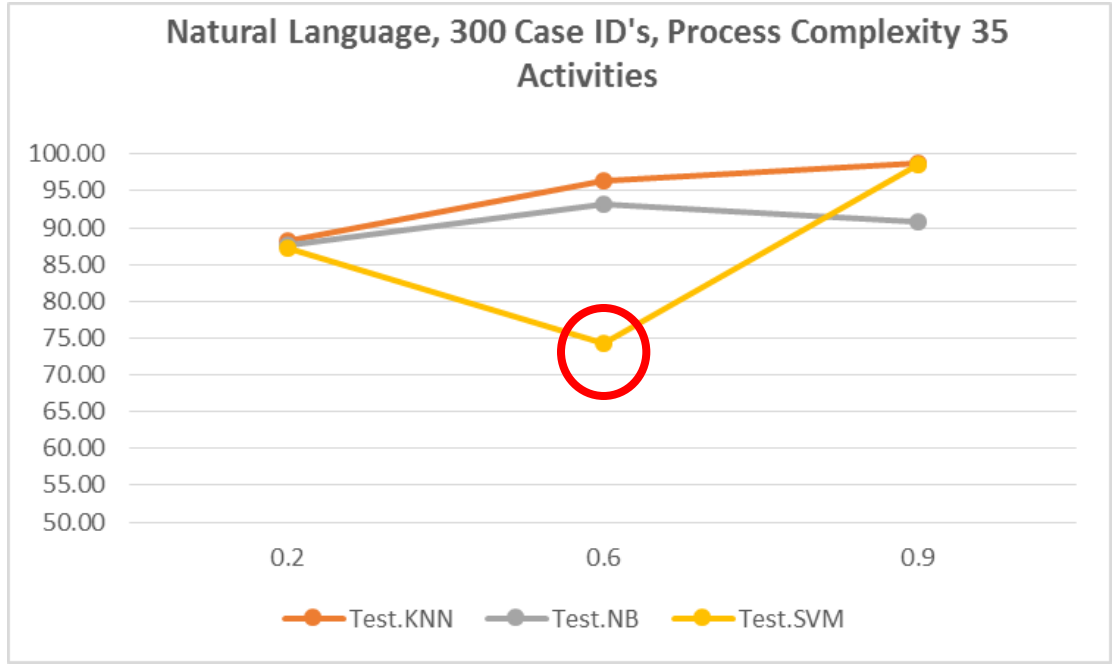

Natural Language, 1000 Case ID's, Process Complexity 10 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Test.KNN
Test.NB
Test.SVM


Natural Language, 1000 Case Id's, Process Complexity 35 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Test.KNN
Test.NB
Test.SVM

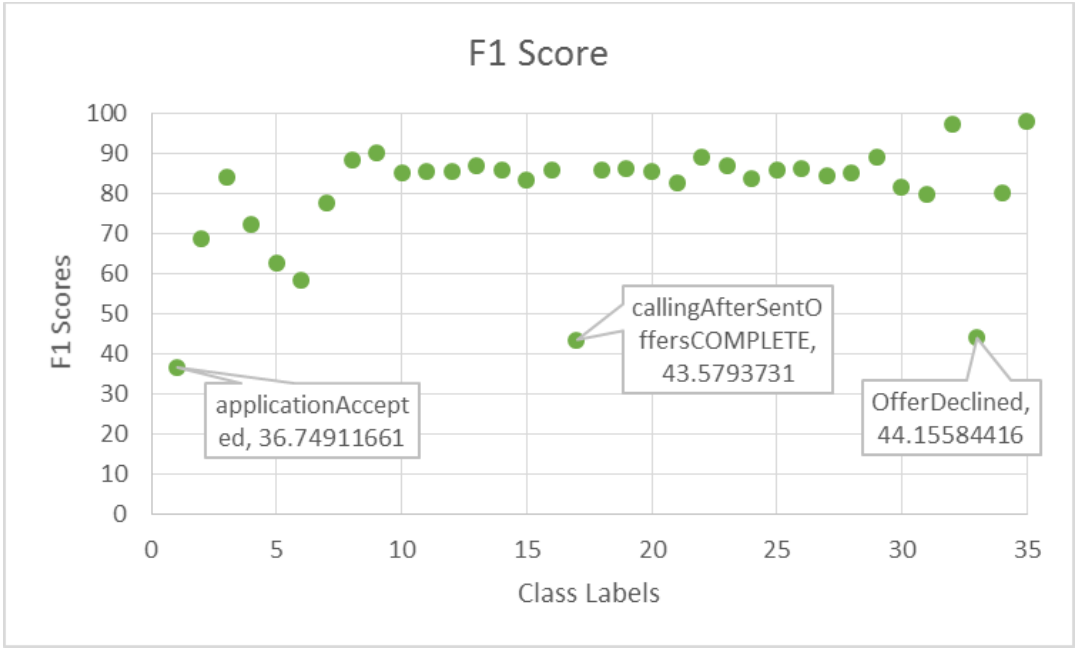

Natural Language, 300 Case ID's, Process Complexity 10 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Train.KNN
Train.NB
Train.SVM


Natural Language, 300 Case ID's, Process Complexity 35 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Train.KNN
Train.NB
Train.SVM


Natural Language, 300 Case ID's, Process Complexity 10 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Test.KNN
Test.NB
Test.SVM


Natural Language, 300 Case ID's, Process Complexity 35 Activities
100.00
95.00
90.00
85.00
80.00
75.00
70.00
65.00
60.00
55.00
50.00
0.2
0.6
0.9
Test.KNN
Test.NB
Test.SVM

SVM Classification F1 Score (300 Case Id's, Sparsity 0.6, Complexity 35 Classes)



- Extend overall accuracy to class wise recall, precision, f1 scores.

- Identify poorly classified classes and trace back to the features of the class.

RESULTS / PERFORMANCE

- Process Map Evaluation
  - Order and Trace evaluation
  - Effect of errors in the event log

RECALL

Process: comprised of activities

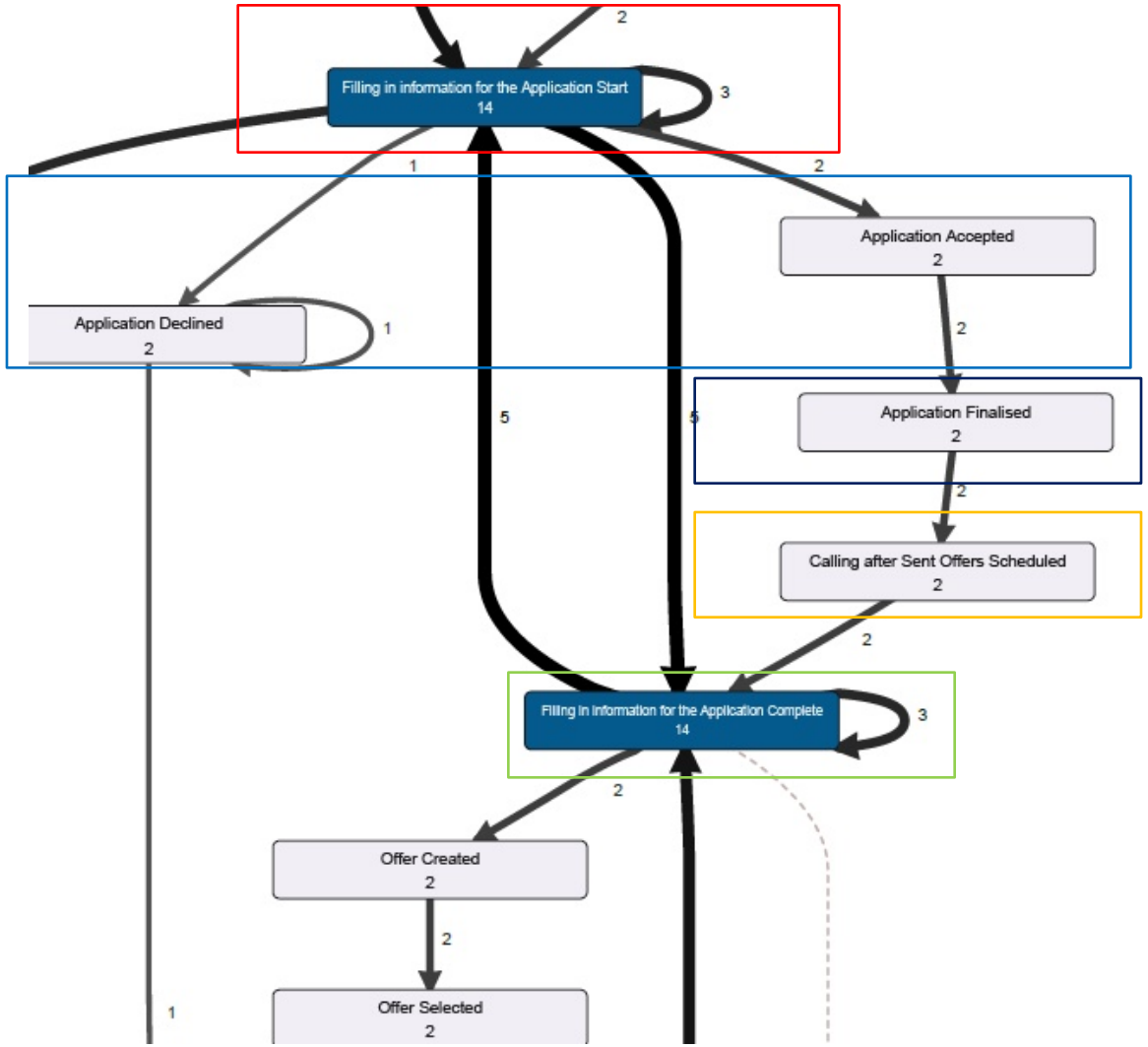
Event: occurrence of an activity

Trace: a sequence of events

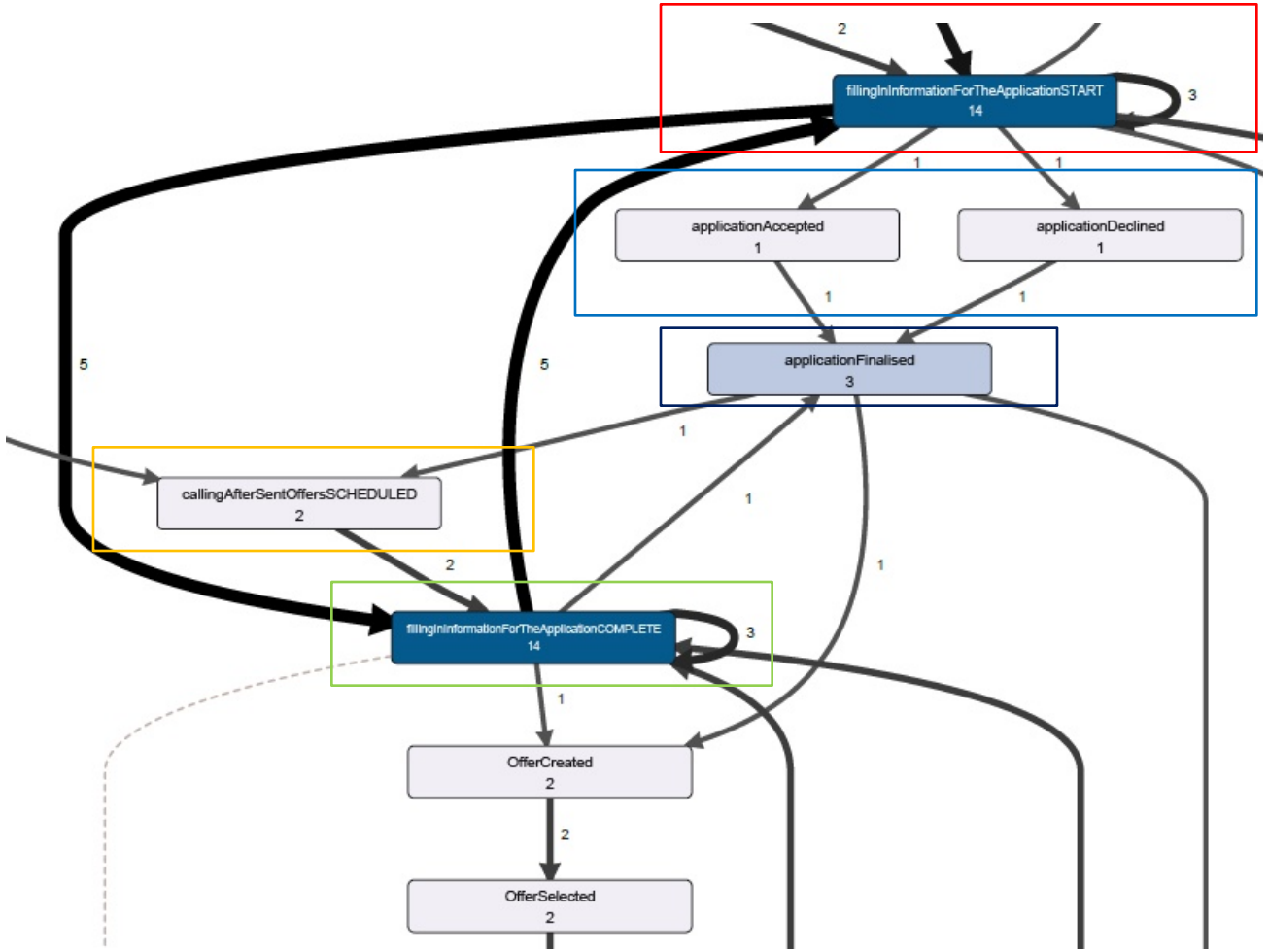
DISCO plots a 100% truthful process map visualization – deterministic

- Consequence of inaccurate event log:
  - Falsely identifies and models traces of events that do not exist
  - Spaghetti model

Benchmark Process Map

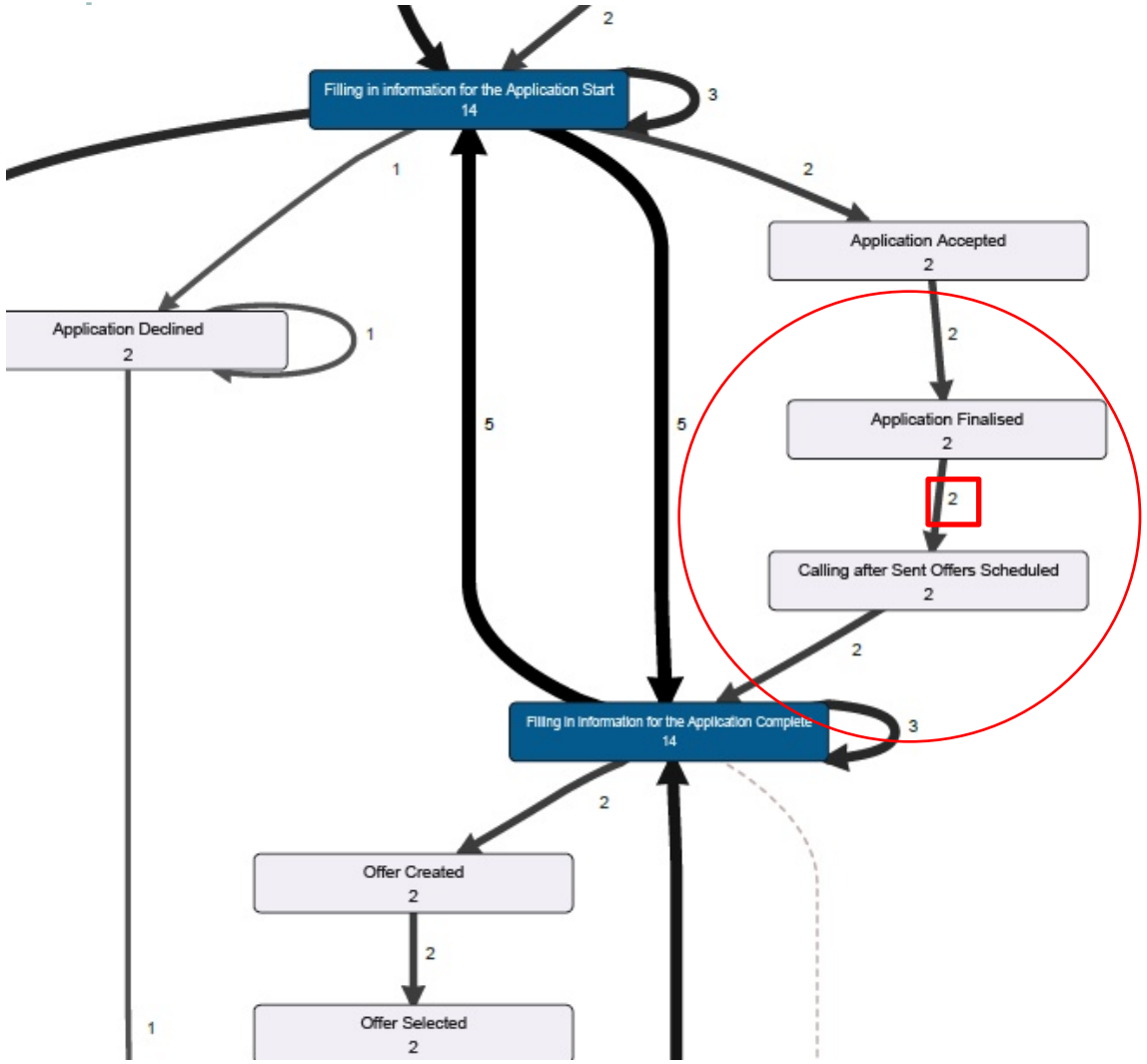


Predicted Process Map

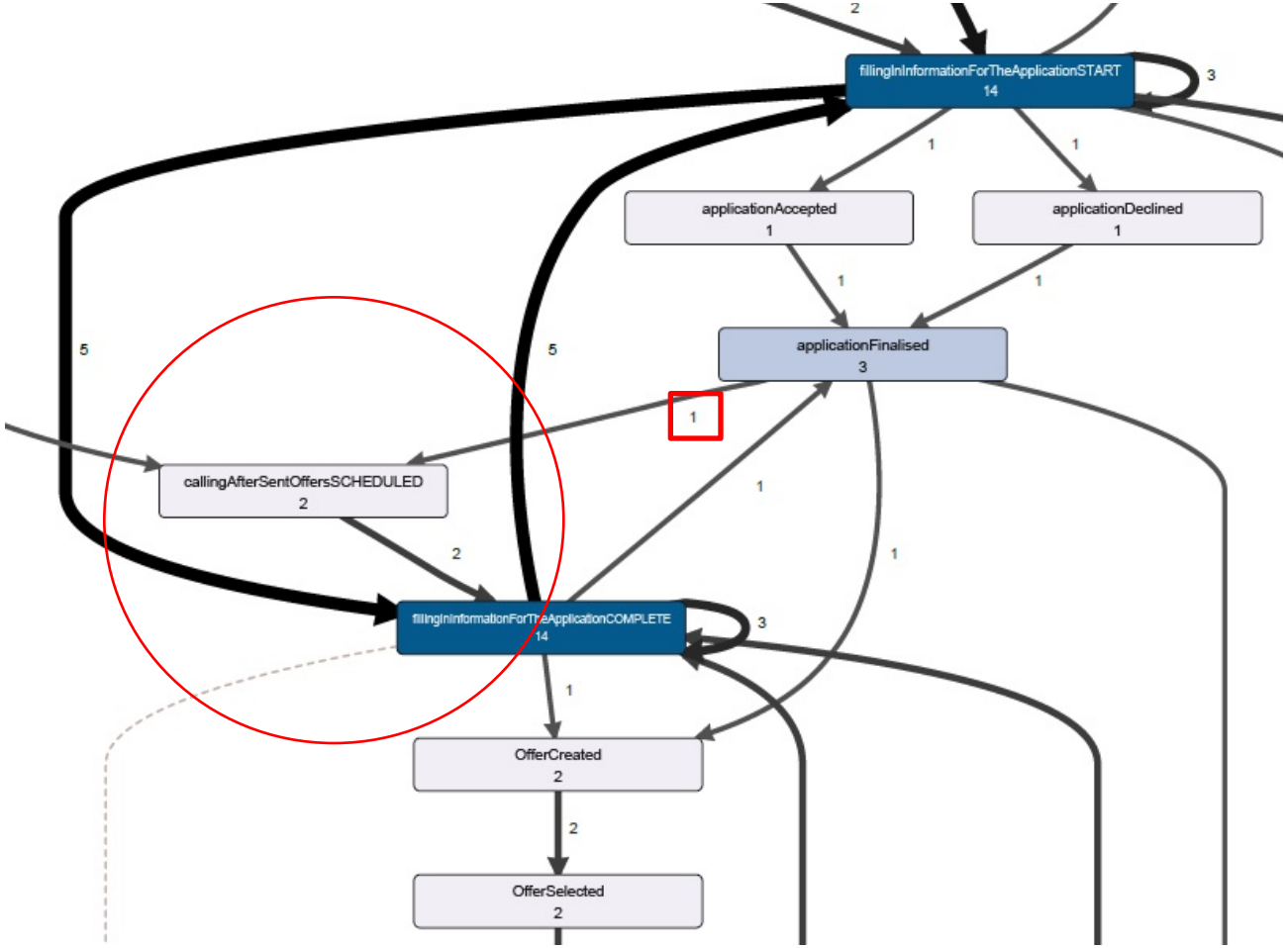


Process Map developed using DISCO™ demo license

Benchmark Process Map



Predicted Process Map



Process Map developed using DISCO™ demo license

ACADEMIC AND BUSINESS CONTRIBUTION

ACADEMIC CONTRIBUTION

Proof of Concept

- Extended an analytical technique via Natural Language Processing
- Identified impact of NLP performance on process mining
- Collated comparative results of algorithmic performance under varying realistic domain specifications

ACADEMIC CONTRIBUTION

Investigation of Factors of Success
in the context of unstructured process relevant documents

- Process Complexity (# of classes)
- Distribution of Language
- Sparsity of Term Document Matrix (TDM)

BUSINESS CONTRIBUTION

- Developed modular framework which aids in event log extraction from unstructured data formats
- Ability to mine business processes for enterprises that do not have PAIS or formally record an event log
- Enables model discovery and facilitates model enhancement

CONCLUSION

LEARNING OUTCOMES

Greater understanding of the significance of Data Quality and necessity of meticulous Data Preprocessing

Gravity of making and identifying research assumptions

- Establishes the merit of findings
- Helps maintain a practical scope
- Necessitates the development of solutions in a modular fashion, which provides a platform for further research

CONCLUSIONS

Difficulties we encountered indicated the richness of this topic for further research and development of methods.

- Proved the concept's feasibility
- Demonstrated importance of key relationships
- Established a platform from which to build on

THANK YOU | Q/A?