32nd Annual INCOSE international symposium
hybrid event
Detroit, MI, USA
June 25 - 30, 2022

Automatic Text Classification of PDF Documents using NLP Techniques

www.incose.org/symp2022

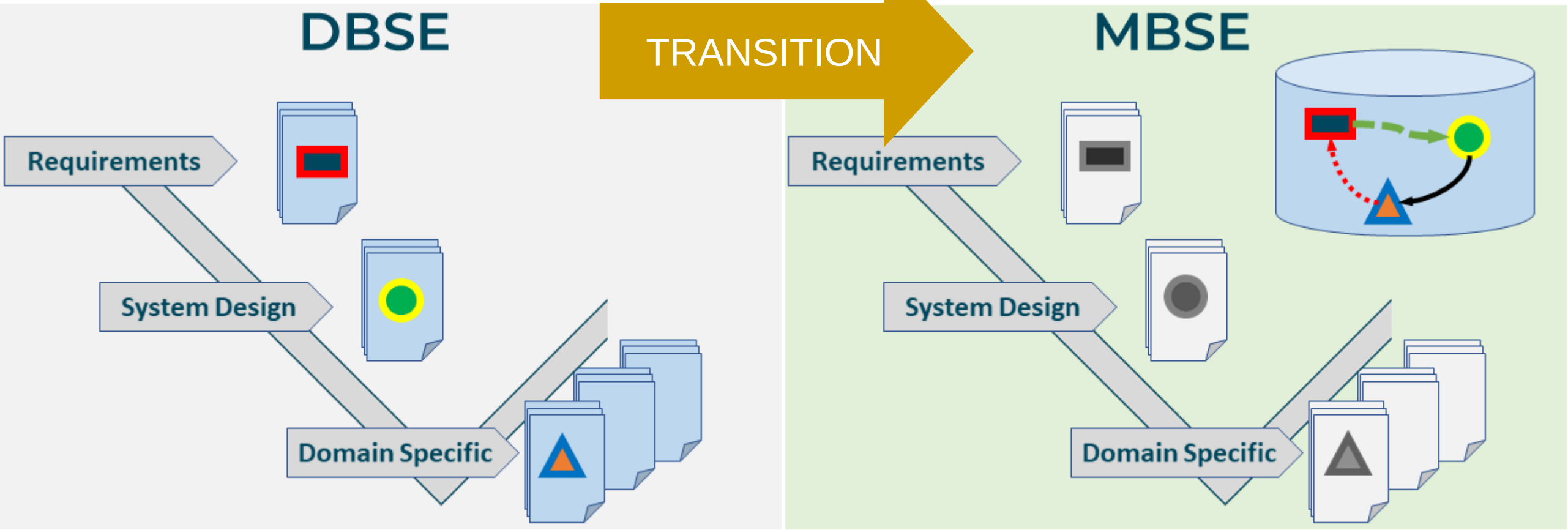
Authors

Nabil Abdoun
SysDICE GmbH
Franz-Volhard-Str. 5,
68167 Mannheim, Germany
nabil.abdoun@sysdice.com

Mohammad Chami
SysDICE GmbH
Franz-Volhard-Str. 5,
68167 Mannheim, Germany
mohammad.chami@sysdice.com

SysDICE
KNOWLEDGE FOR IMPACT

Problem Formulation – MBSE Transition

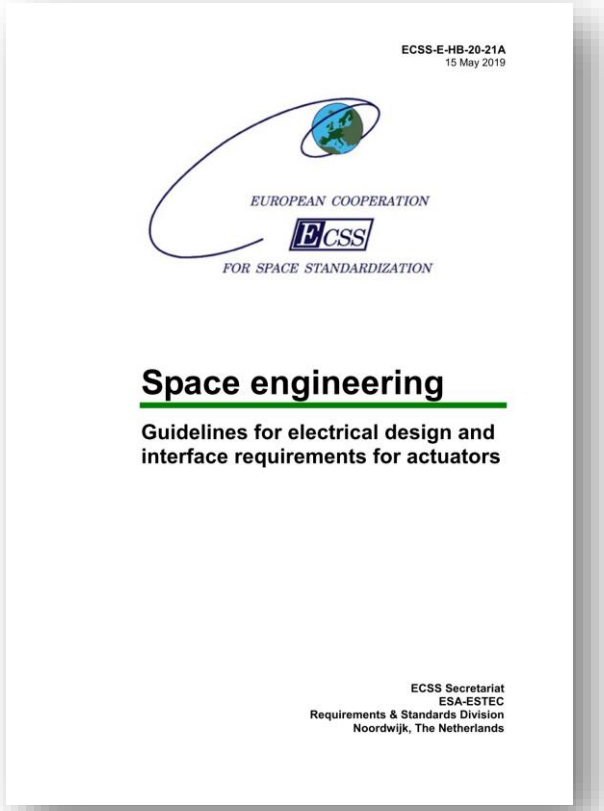


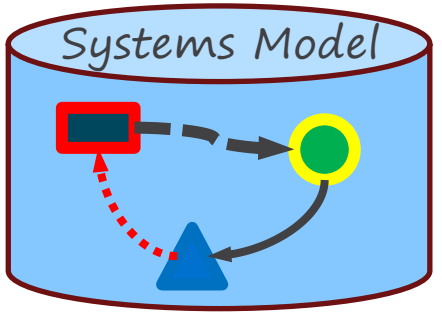
Documents As Primary Artefacts

Models As Primary Artefacts

The Approach: Text-to-Model

With an example from the space sector (ECSS)*



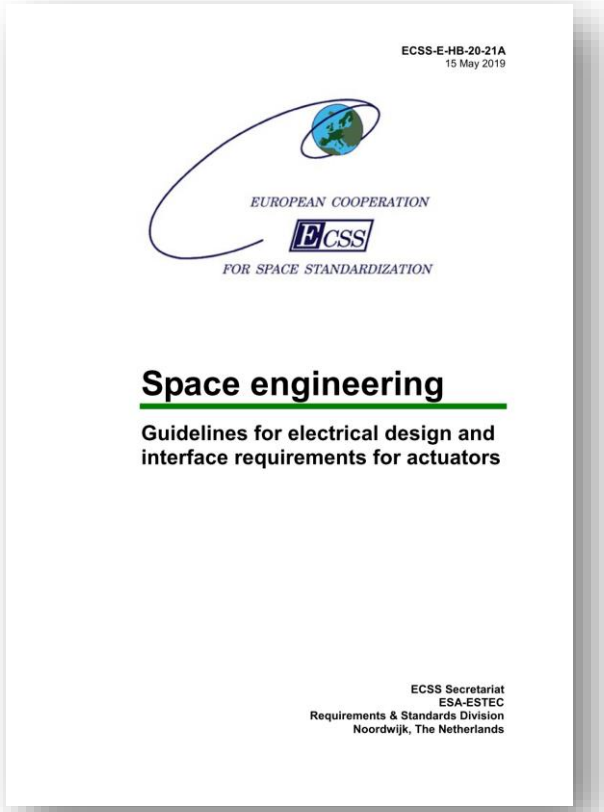


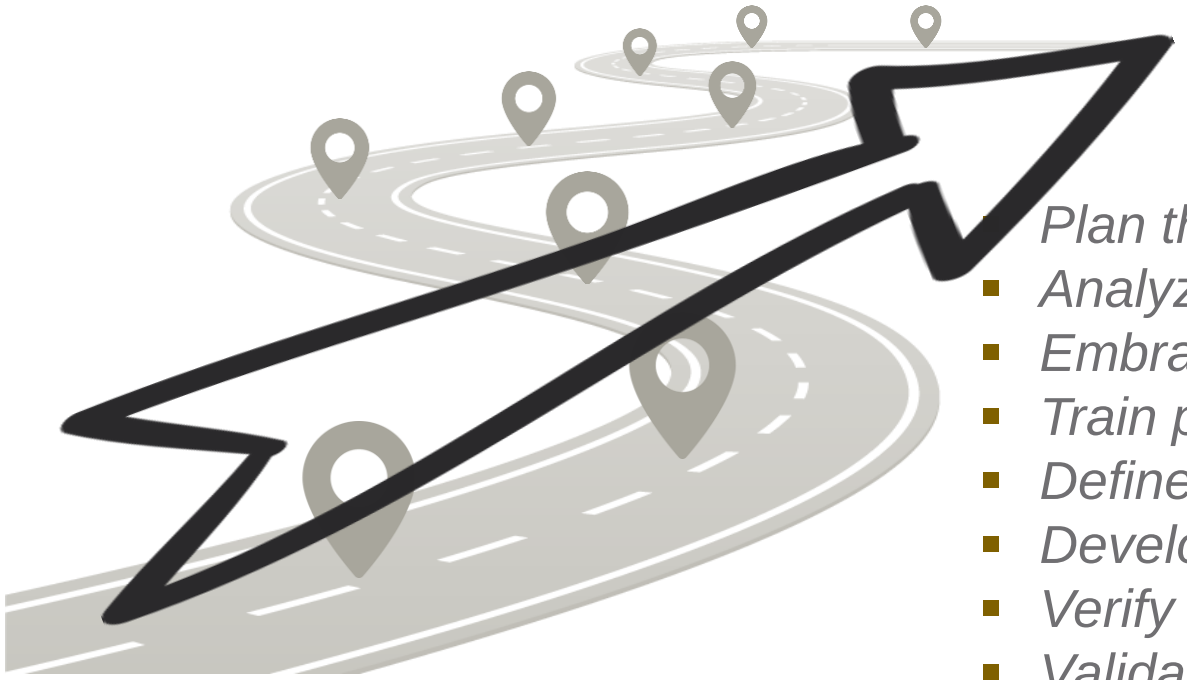
- *Plan the MBSE adoption*
- *Analyze SE information*
- *Embrace process & methods*
- *Train personnel*
- *Define purpose & scope*
- *Develop system models*
- *Verify system models*
- *Validate system models*
- *Manage system models*
- *…*

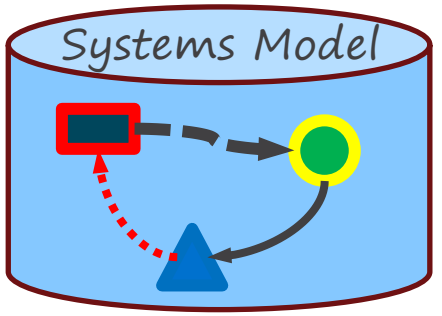
**ECSS: https://ecss.nl/standards/*

The Approach: Text-to-Model

With an example from the space sector (ECSS)*





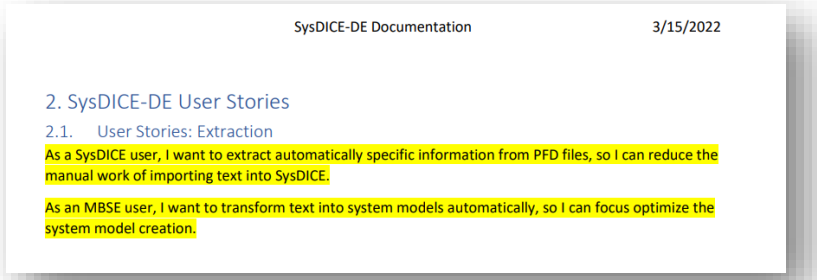


- *Plan the MBSE adoption*
- *Analyze SE information*
- *Embrace process & methods*
- *Train personnel*
- *Define purpose & scope*
- *Develop system models*
- *Verify system models*
- *Validate system models*
- *Manage system models*
- *…*

**ECSS: https://ecss.nl/standards/*

Background and Problem Scope

From "Text-to-Model" To "PDF-to-Model"

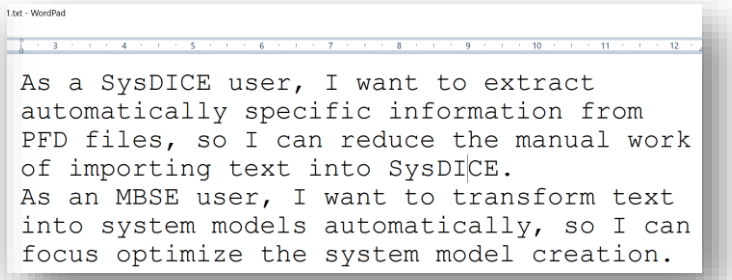

SysDICE-DE Documentation 3/15/2022

2. SysDICE-DE User Stories

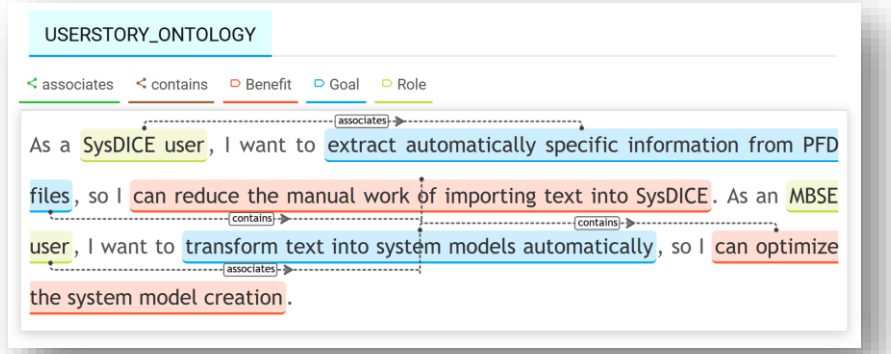
2.1. User Stories: Extraction

As a SysDICE user, I want to extract automatically specific information from PFD files, so I can reduce the manual work of importing text into SysDICE.

As an MBSE user, I want to transform text into system models automatically, so I can focus optimize the system model creation.


1.txt - WordPad

As a SysDICE user, I want to extract automatically specific information from PFD files, so I can reduce the manual work of importing text into SysDICE.
As an MBSE user, I want to transform text into system models automatically, so I can focus optimize the system model creation.



Textual Information

System Model Information

Extracting specific SE information (e.g., requirements, user stories) from PDF documents was performed manually. ✗

Automatic identification of entities and relationships from text using AI and transforming them into system models (e.g., SysML model). ✓

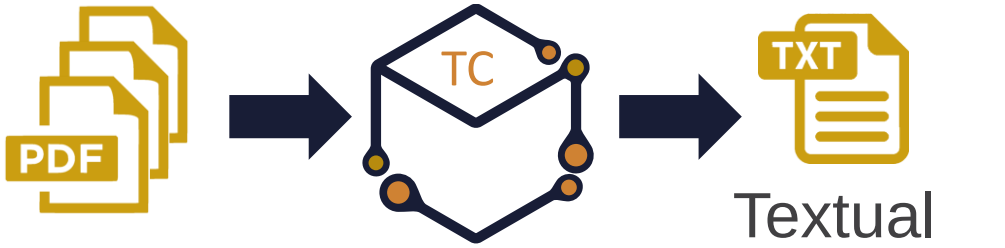
SysDICE Text Classification (TC)

Proof of Concept

Automatically **extract systems engineering textual information** from documents using AI

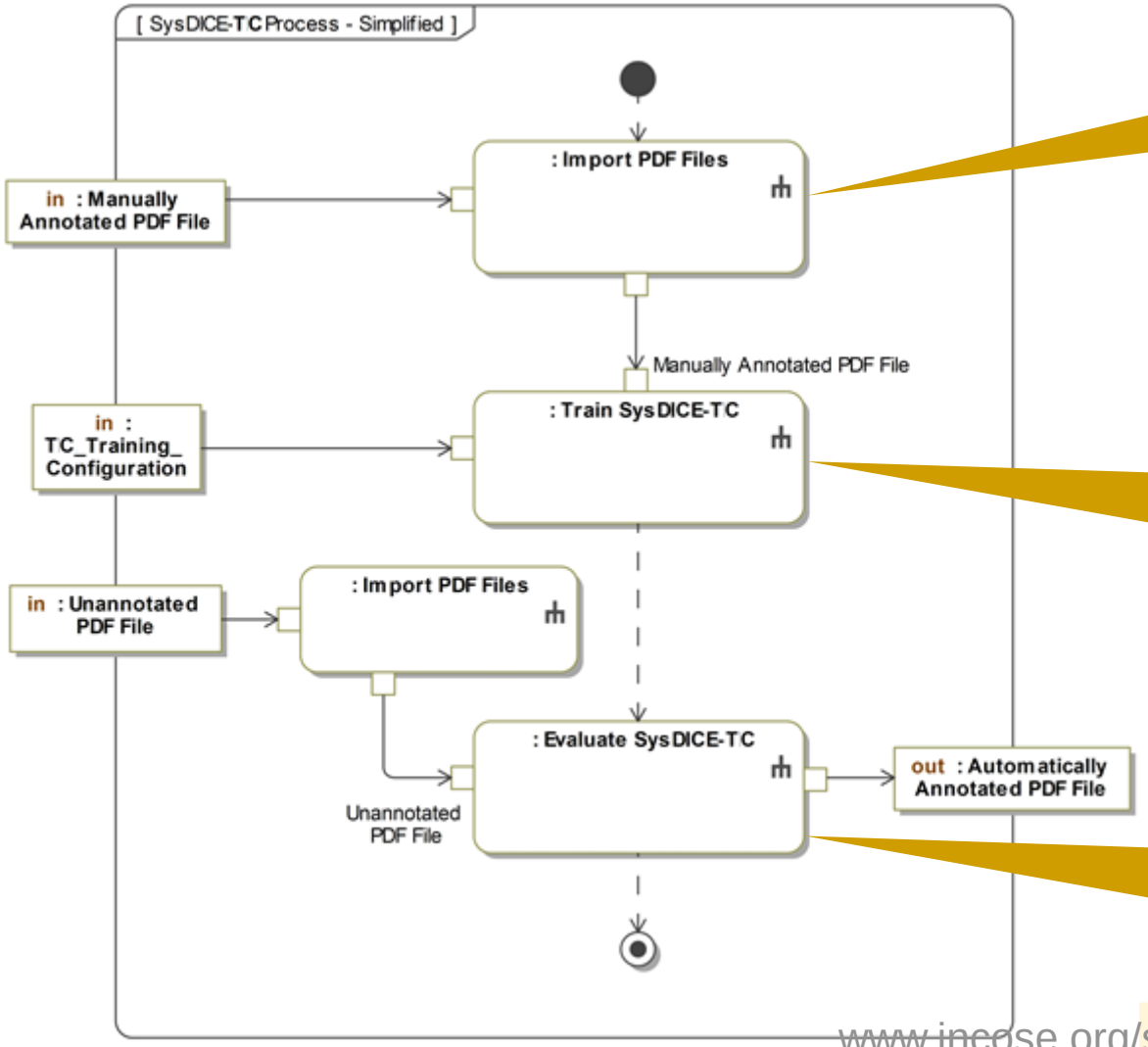
- Actual PoC scope: PDF documents

1. A **flexible** solution: Works for a wide spectrum of systems engineering industries without the need to specifically tailor its AI algorithms
2. A **Simple end-user annotation** of the SE information using any PDF application in order to train the AI algorithms
3. After the successful evaluation of the trained AI algorithms by the end-user, the **SE information are automatically extracted** and provided for the second module of SysDICE, i.e., Text-to-Model



SysDICE-TC Process Activities

The Simplified Version



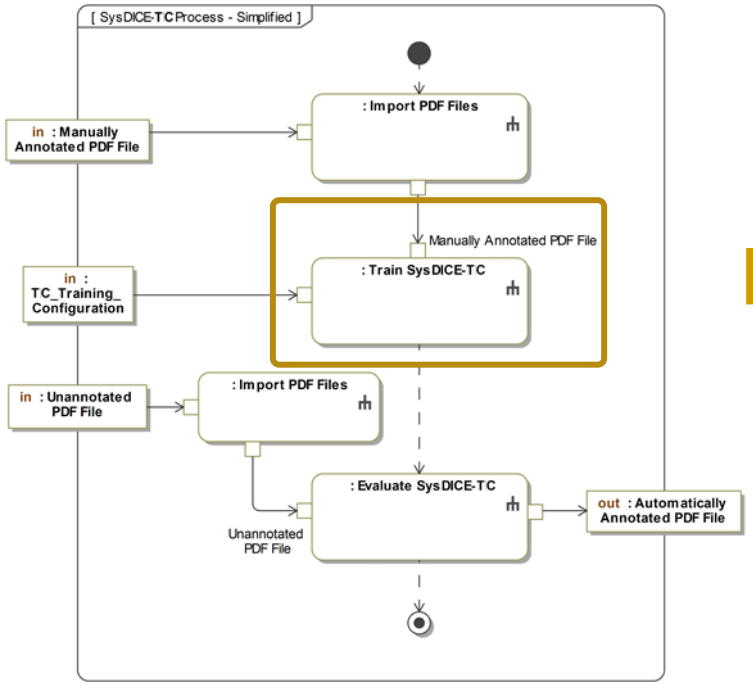
Import the manually annotated PDF files into SysDICE-TC.

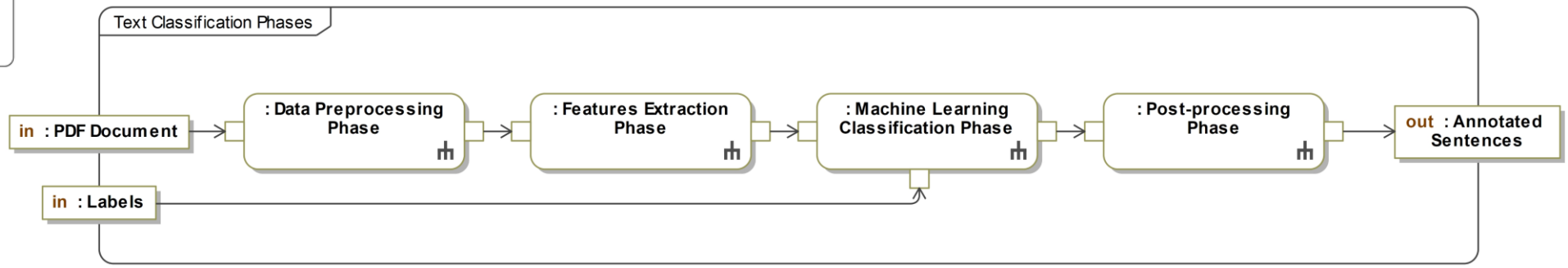
Train SysDICE-TC machine learning algorithms based on the user input from the manually annotated PDF files and the selected training configurations.

Evaluate the trained SysDICE-TC on new unannotated PDF files and generate the automatically annotated PDF file.

SysDICE-TC Process Activities

The Simplified Version – Train SysDICE-TC



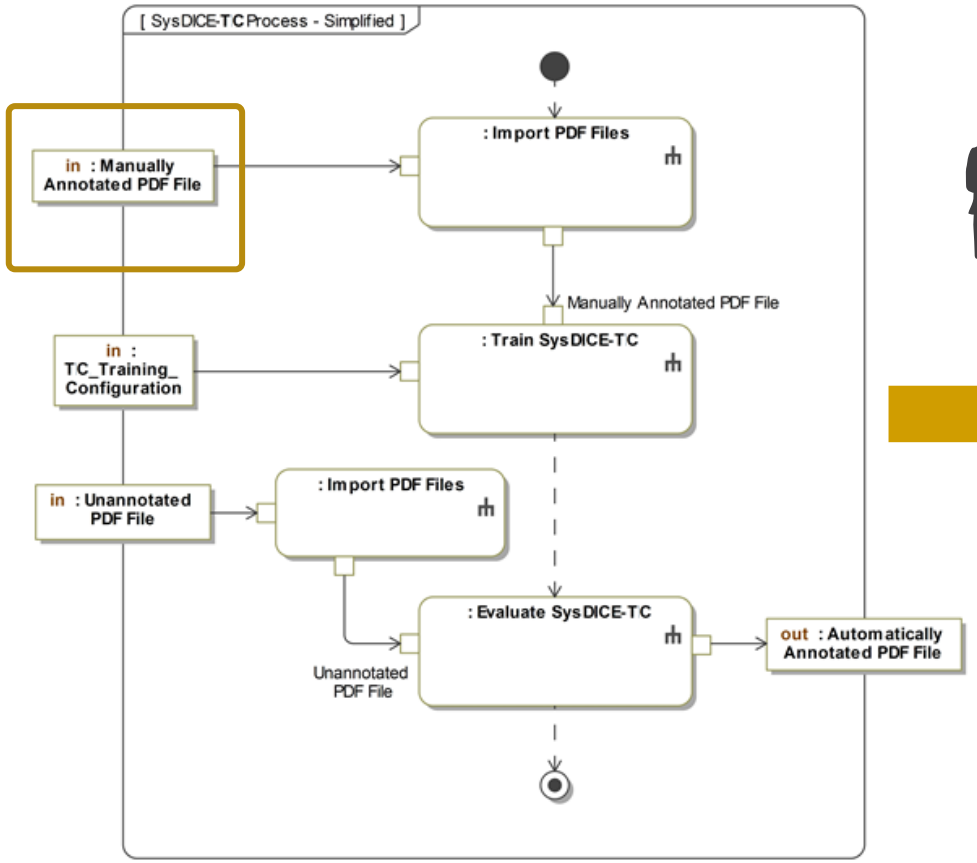


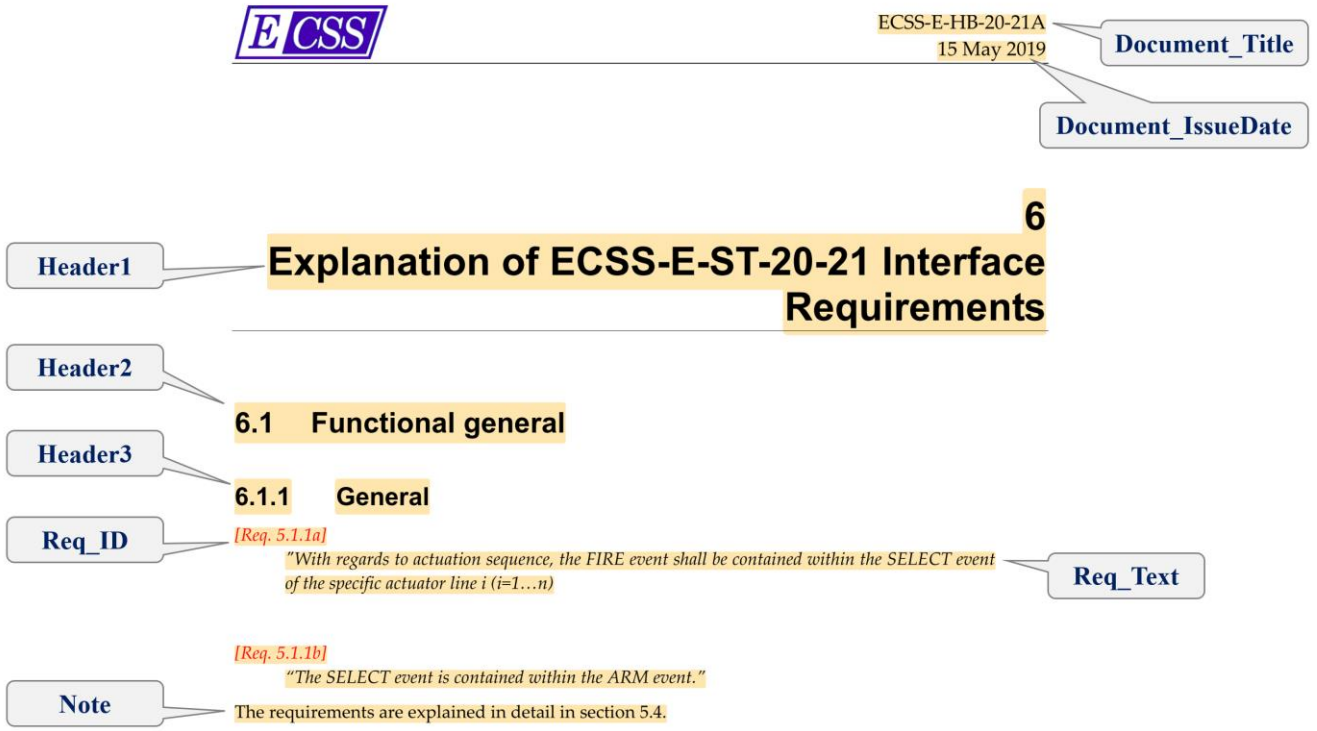
SysDICE Text Classification method phases

SysDICE-TC Process Activities

The Simplified Version – Annotate PDF Files Manually

- Use Case Application: **European Cooperation for Space Standardization** ECSS* from the aerospace sector.

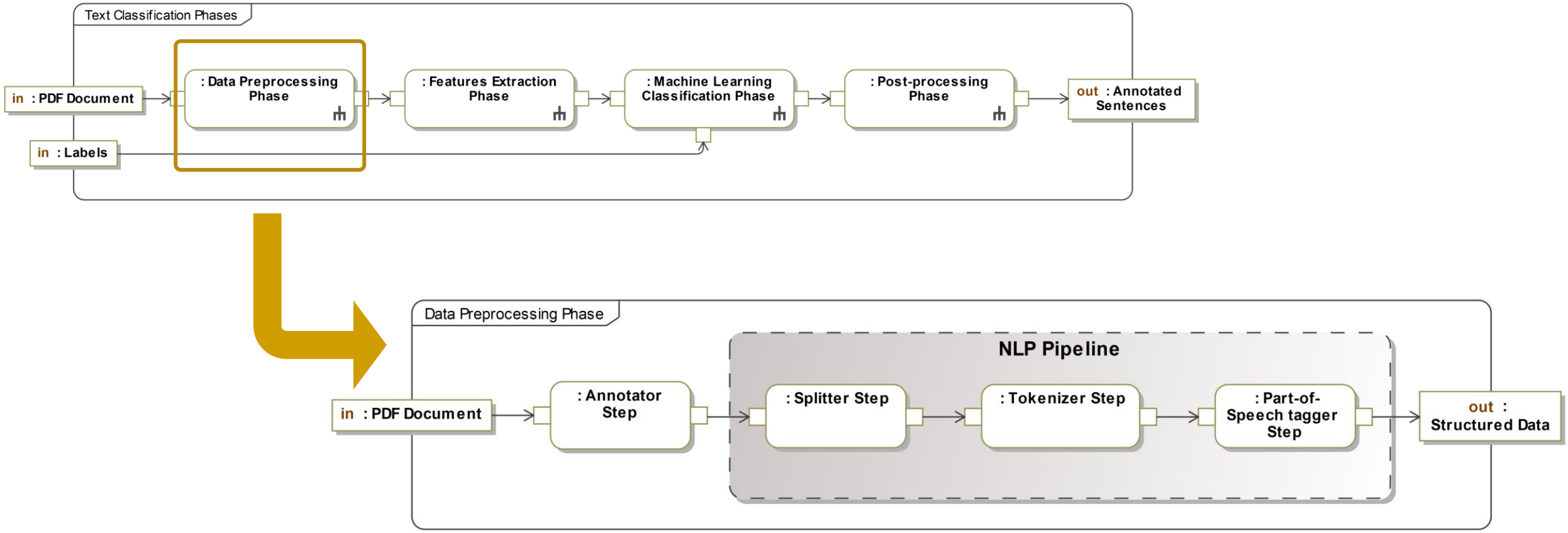




*ECSS: https://ecss.nl/standards/

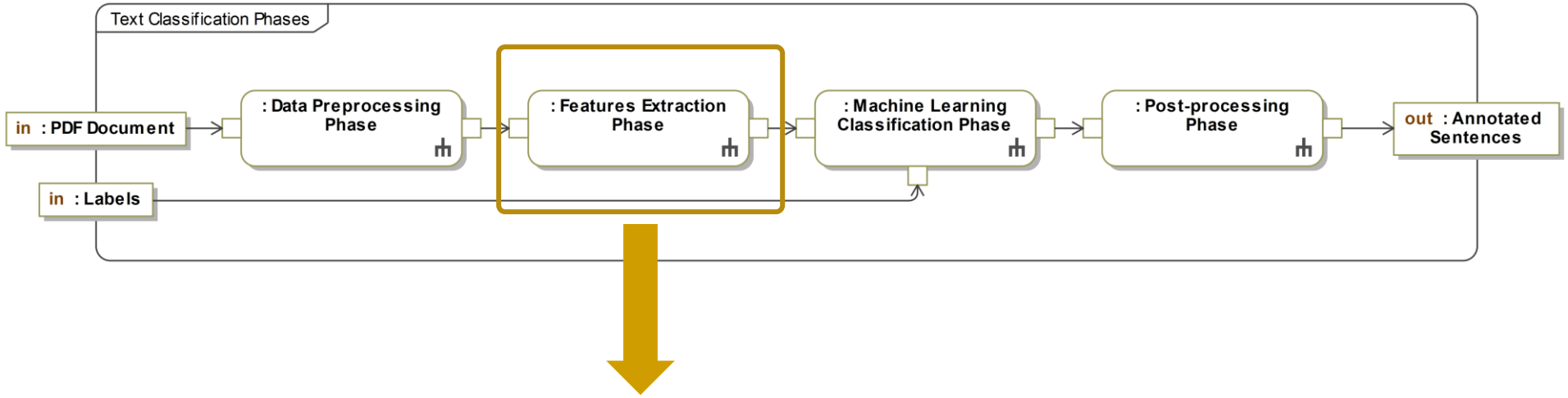
SysDICE-TC Process Activities

The Simplified Version – Data Preprocessing Phase



SysDICE-TC Process Activities

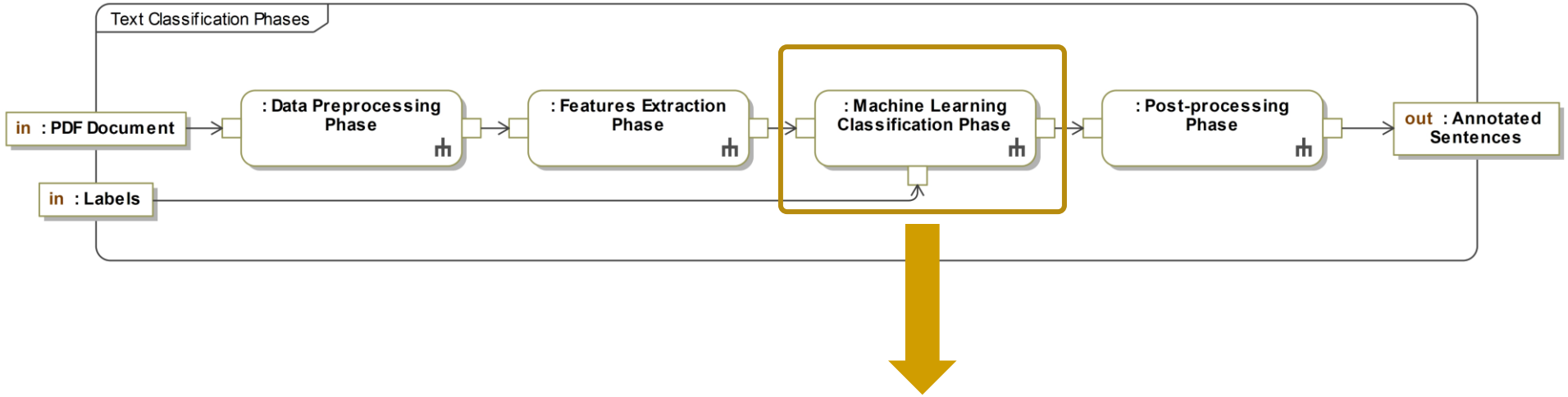
The Simplified Version – Features Extraction Phase



Feature Category Name	Description
Tokenizer Features	Features based on the tokenizer, such as the number of tokens, the name/type of the first token, etc.
Layout and Structural Features	Features based on the layout of the document, e.g., x-y, etc.
Grammatical and POS Features	Features based on the grammar and the POS, e.g., contains_past_verbs, etc.
Most frequent-modal features	Features based on the most common modal verb in the requirements texts.

SysDICE-TC Process Activities

The Simplified Version – Evaluate SysDICE-TC



ML classifier	Accuracy (%)	Precision (%)	Recall (%)
Decision Tree	98.2	99.2	99.0
Random Forest	98.0	99.4	98.8
Naïve Bayes	74.0	76.3	82.7
Support Vector Machine	84.2	88.0	89.0
Feedforward Neural Network	87.9	92.0	91.0

SysDICE-TC Process Activities

The Simplified Version – Evaluate SysDICE-TC

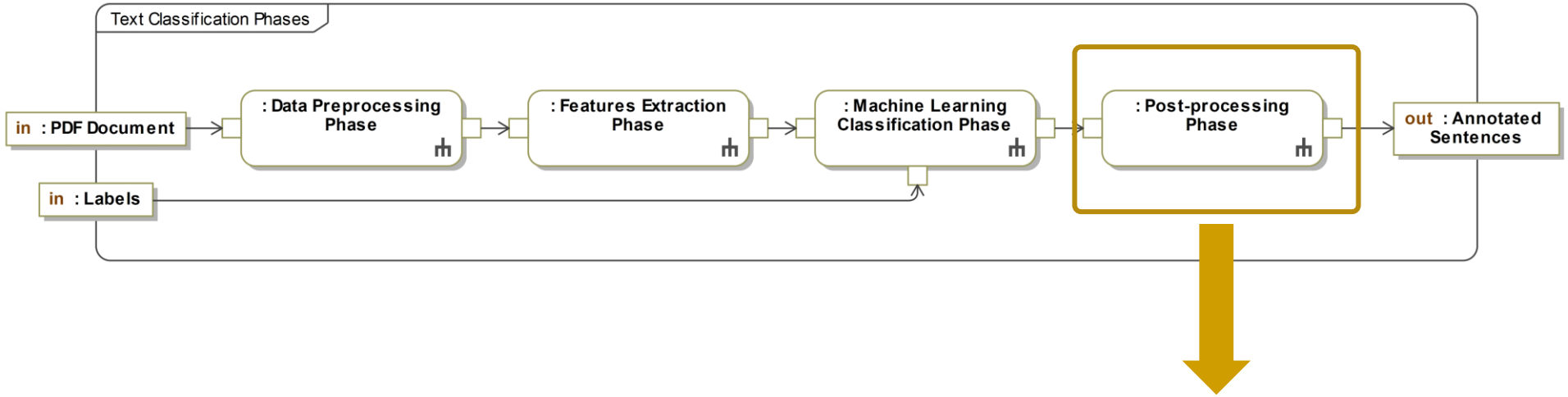
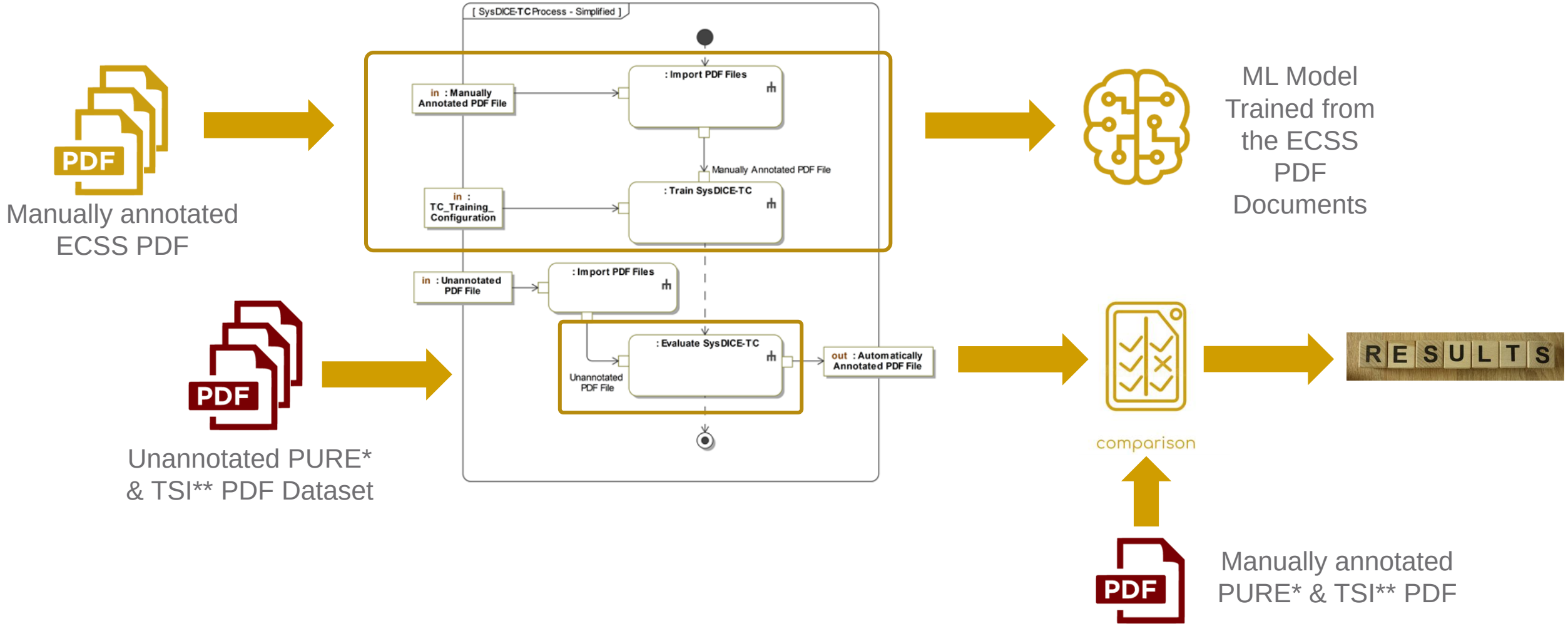


Table of contents

SysDICE-TC

Test Case: Machine Learning Generalization



*Ref: PURE: a Dataset of Public Requirements Documents: https://zenodo.org/record/1414117#.YkGbuDWxURJ

**Ref: TSI: Technical specification for Interoperability

Takeaways

First step

A first step but it works.

AI4MBSE competences

New activities would require competences in AI and MBSE.

Machine learning model

An NER model can be trained and achieved without any previously provided library.

Next action

Looking forward to enhance and work further on the text-to-model tool part.

AI4MBSE

Lack of a coherent foundation for enabling the application of AI for MBSE.

32nd Annual INCOSE international symposium

hybrid event

Detroit, MI, USA
June 25 - 30, 2022

www.incose.org/symp2022