Generating Code from Natural Language

Bernhard Stadler (TU Dresden), Klaudia Thellmann (TU Dresden)

AM
Understanding
Language

MOTIVATION

Support data scientists in creating analysis workflows:

o Reduce learning overhead and simplify handling of libraries

o Explain analysis results in an understandable way

TASK
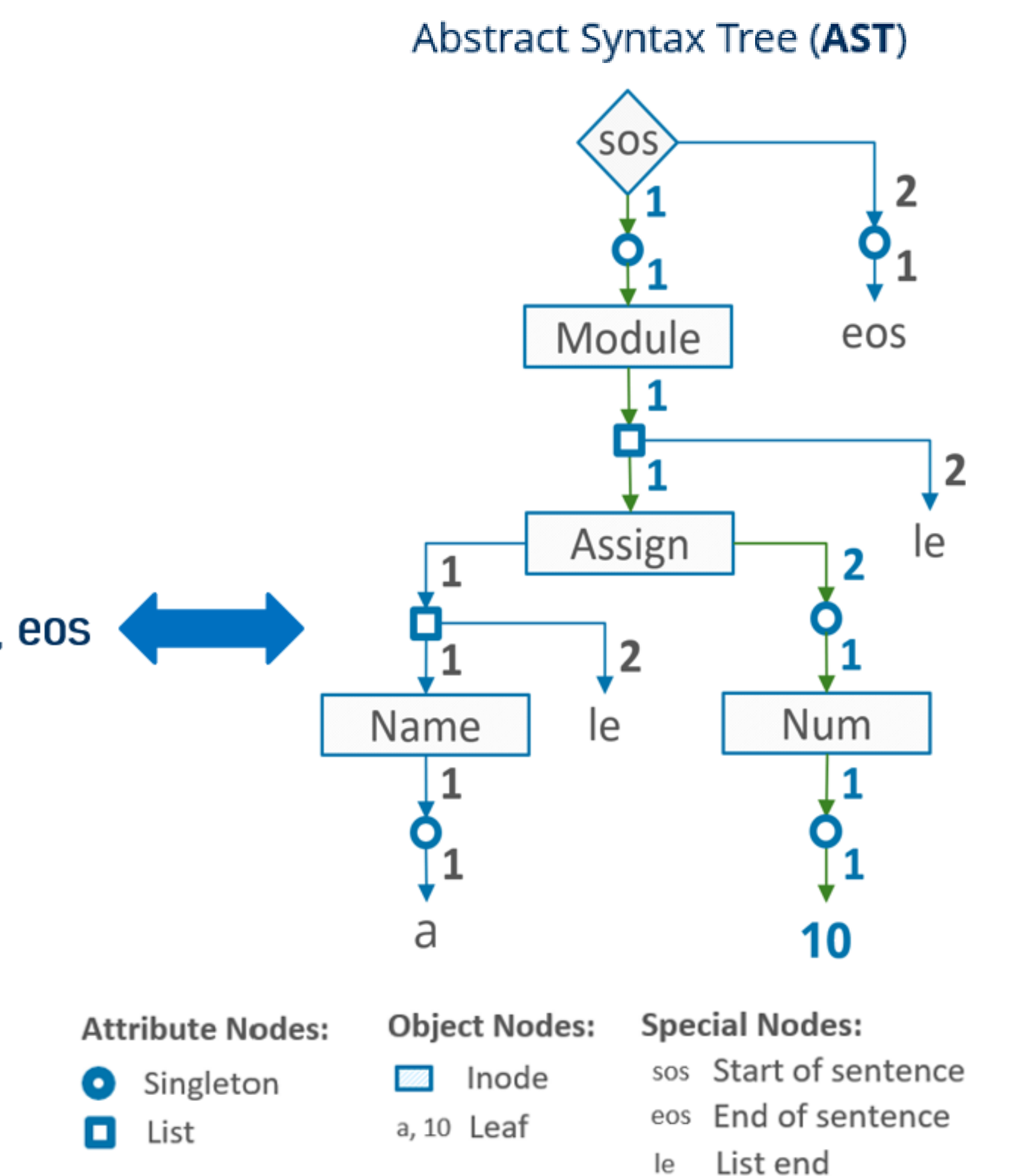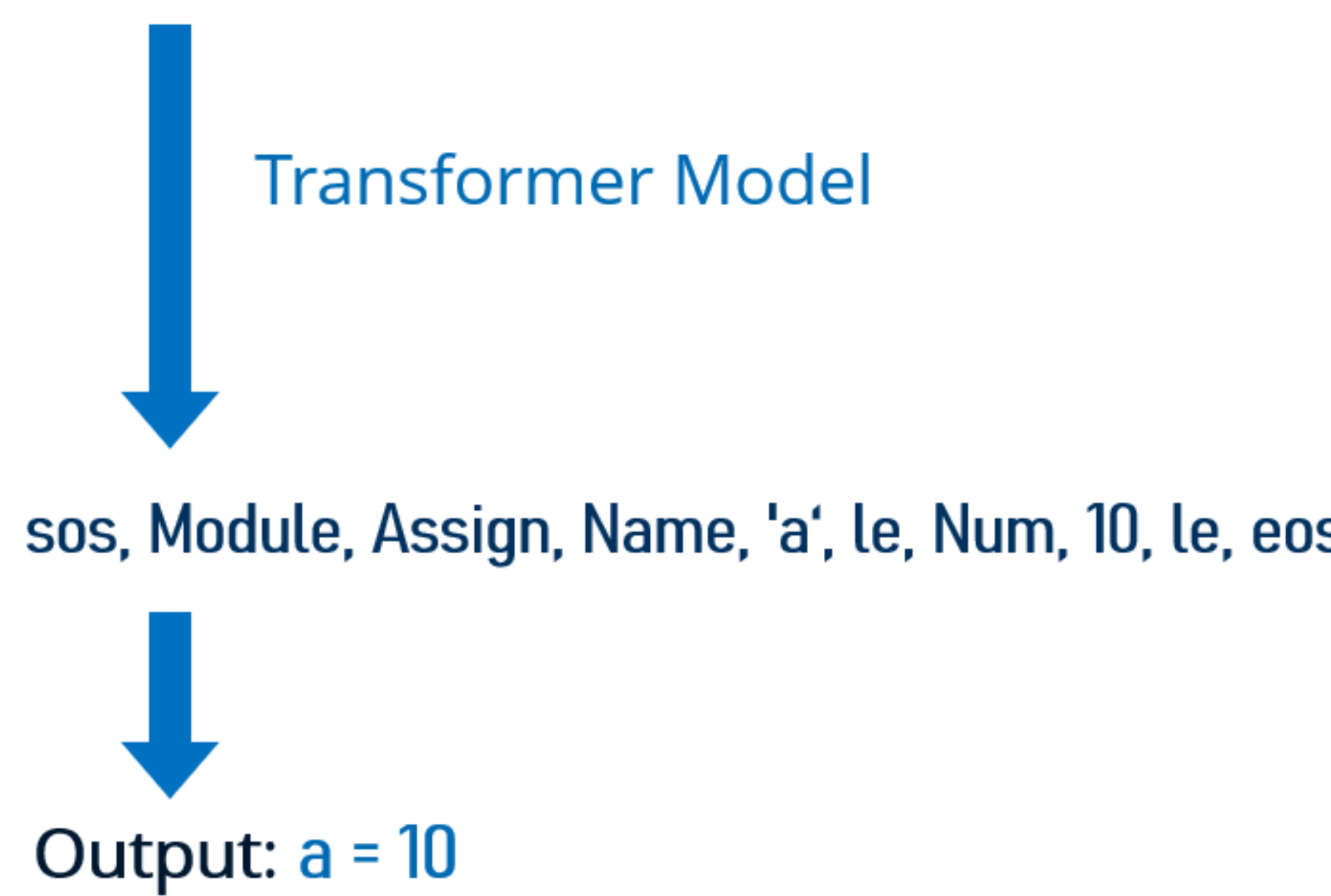
Develop dialogue system for creating analysis workflows:

o Specify analysis steps on a high abstraction level

o Automatically map natural language to Python code

APPROACH

o Transformer for mapping natural language to a tree-based code representation

o Tree-based encoding scheme for incorporating hierarchical structure

o Grammar model to ensure the well-formedness of the generated code

Input: „Assign value 10 to variable a"

Transformer Model

sos, Module, Assign, Name, 'a', le, Num, 10, le, eos

Output: a = 10

Abstract Syntax Tree (**AST**)

Attribute Nodes:	Object Nodes:	Special Nodes:	
Singleton	Inode	sos	Start of sentence
List	a, 10 Leaf	eos	End of sentence
		le	List end

IMPLEMENTATION

RASA-based conversational assistant:

o Integrated into a JupyterLab environment as an extension

o Generated code snippets inserted in a Jupyter Notebook

EVALUATION

o Natural-language-to-code benchmarks: CoNaLa, Hearthstone

o Tree encoding 3.0% improvement over sequential encoding

o CoNaLa: 18,1% BLEU (SotA: 32,3%)

o Hearthstone: 70,7% BLEU (SotA: 81,8%)
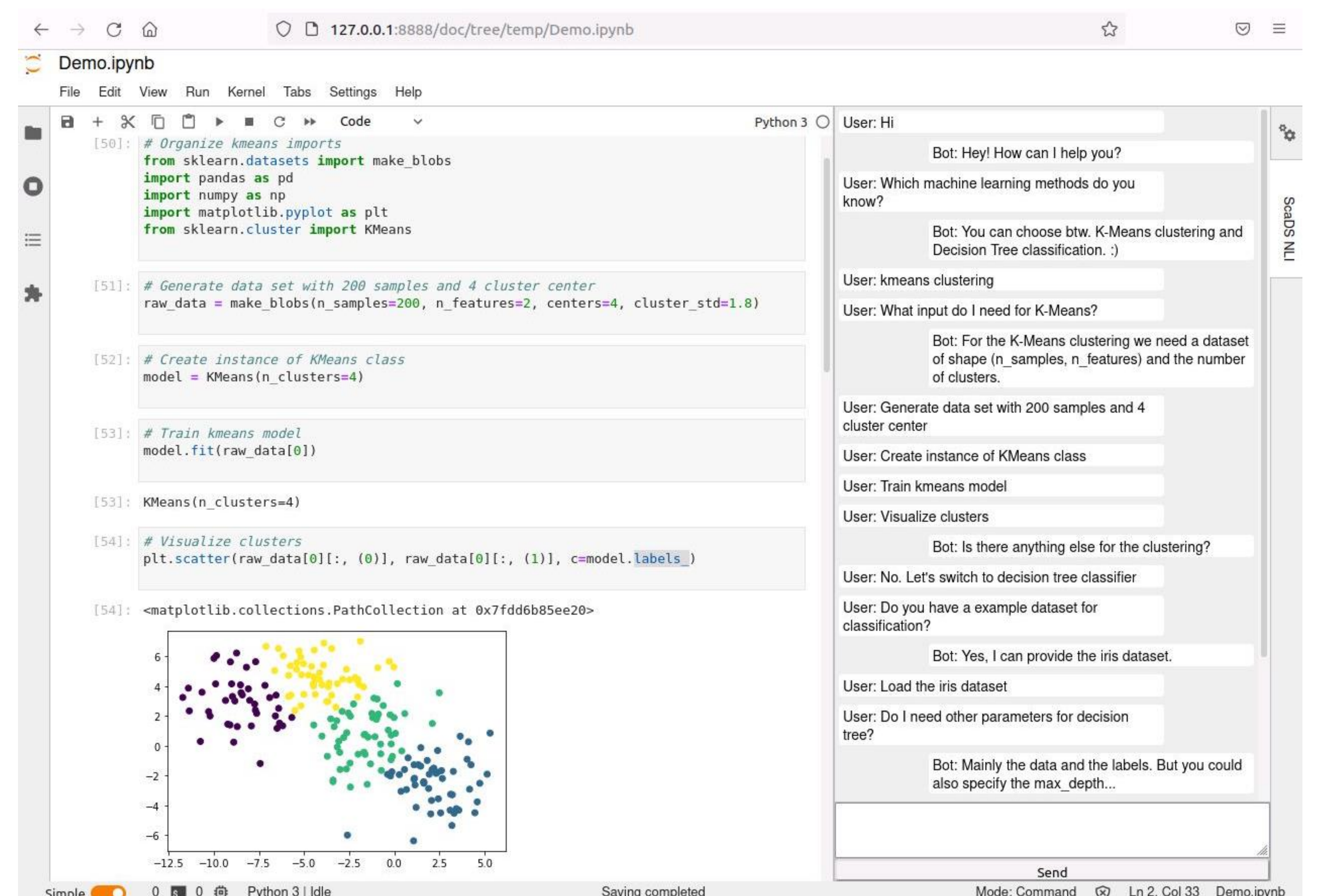
AUTHORS

TU Dresden
Bernhard Stadler, M. Sc. Inf.
bernhard.stadler@tu-dresden.de

TU Dresden
Dipl. Inf. Klaudia Thellmann
klaudia-doris.thellmann@tu-dresden.de

PARTNER

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

CBG Max Planck Institute of Molecular Cell Biology and Genetics

Leibniz-Institut für ökologische Raumentwicklung

UFZ HELMHOLTZ Centre for Environmental Research

HZDR

HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF