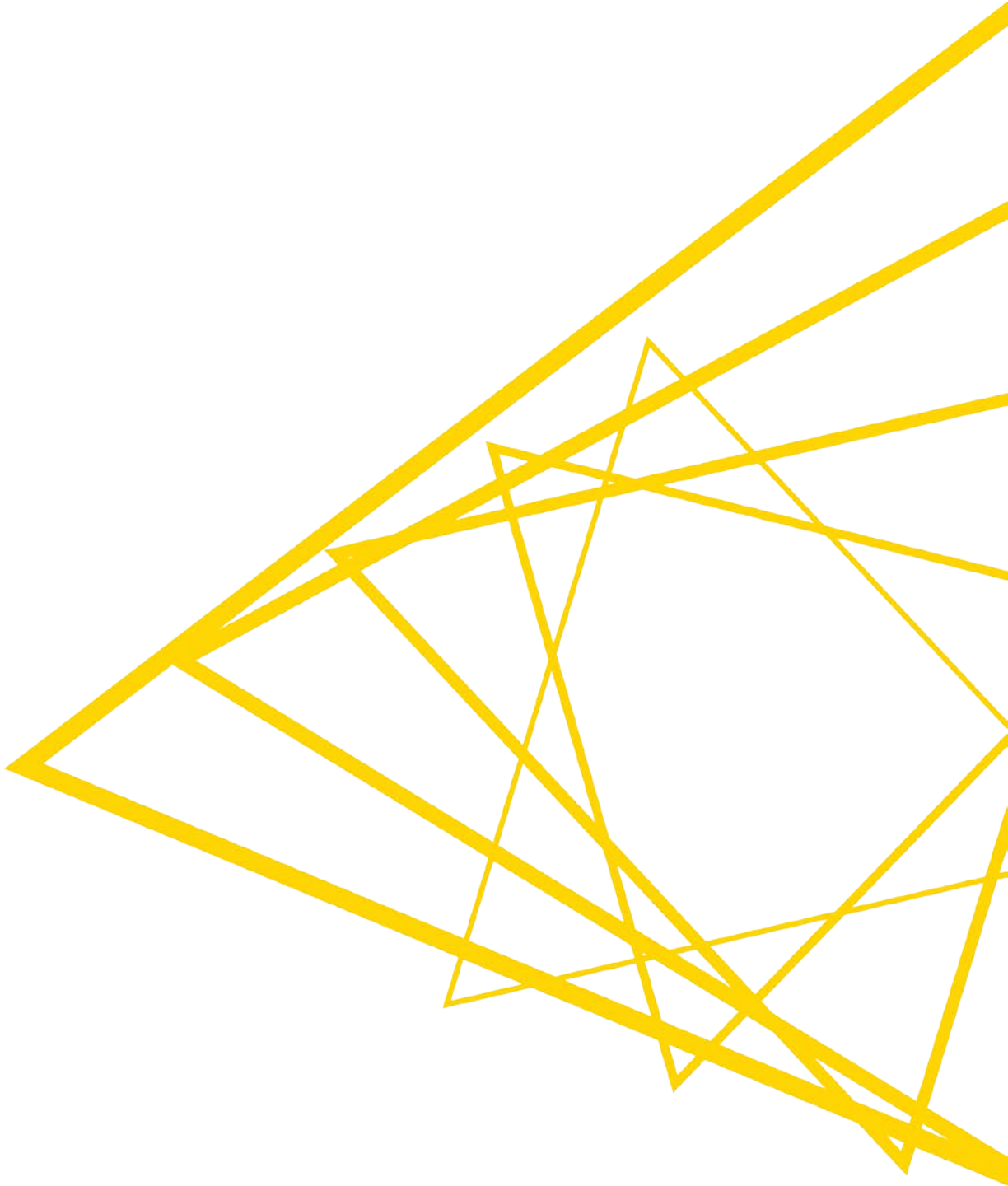
Open for Innovation
KNIME

KNIME & Teacher Bots: From Workflows to Micro Services

Kathrin Melcher
Vincenzo Tursi
Rosaria Silipo
Phil Winters



© 2018 KNIME AG. All Right Reserved.

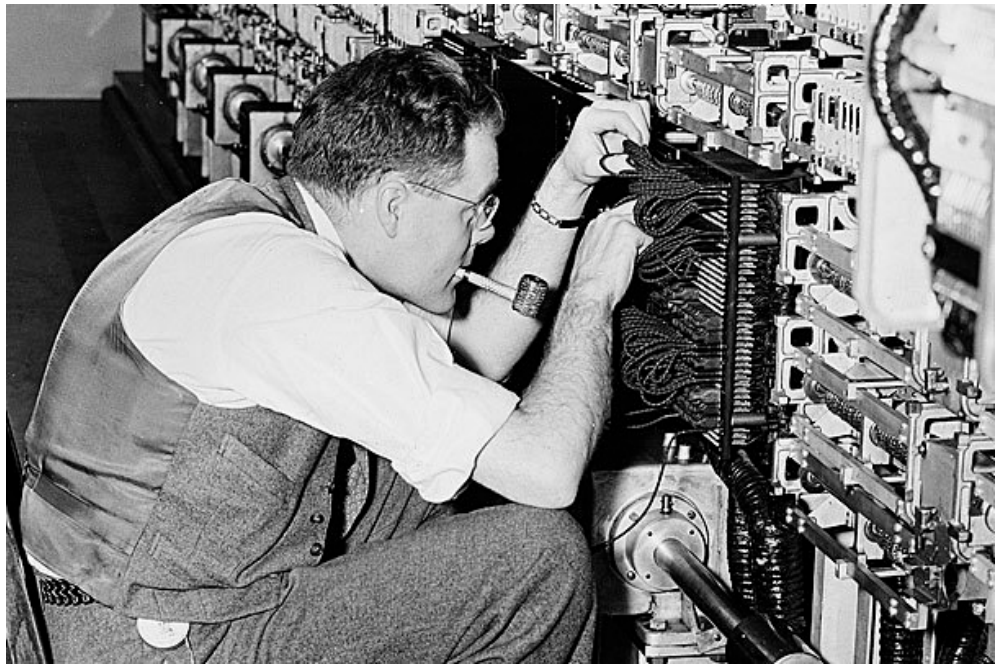
The History of Bots

1950

A.L.I.C.E.

Artificial Linguistic Internet Computer Entity.

Alan Turing



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

Bots

"A bot is software designed to automate the kinds of tasks you would usually do on your own (or another human would do for you)."

Search Bots

Teaching Bots

Communication Bots

Personal Assistant Bots

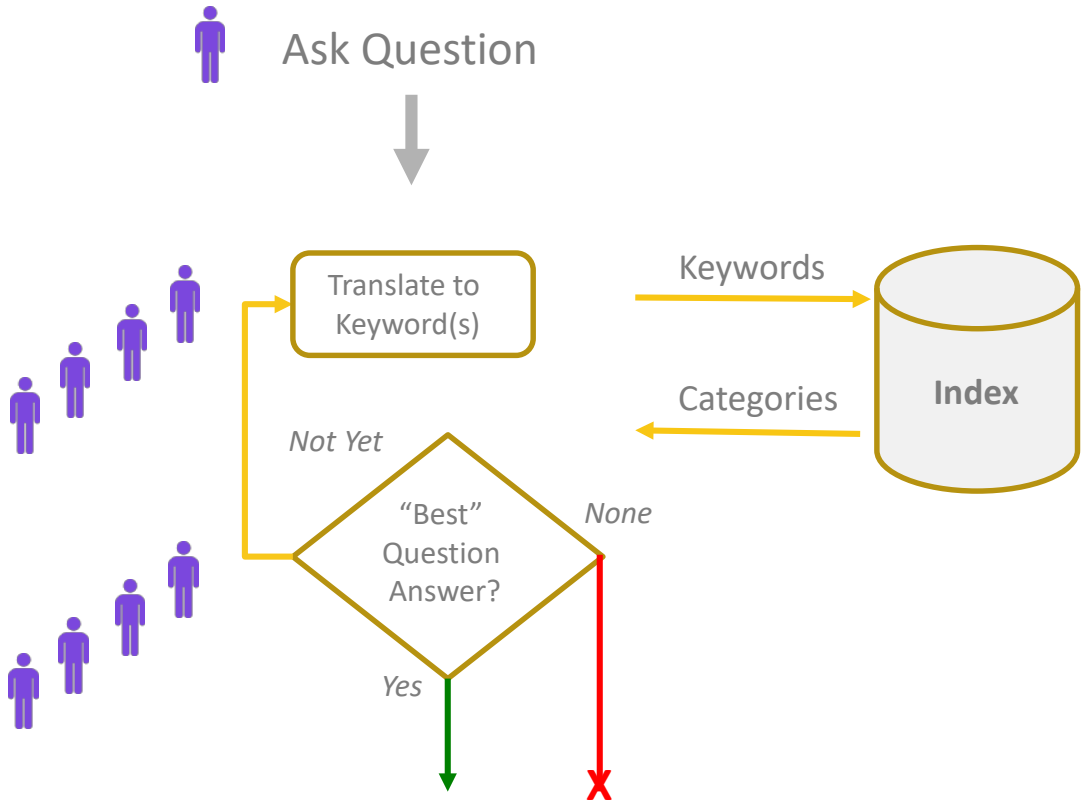
Data & Developer Bots

Team Bots

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

The Human Internet Search Process



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

The Search Challenge: Context

A Jewish Holiday

Paul Shapiro
Professional Search marketer
(and huge KNIME fan…)

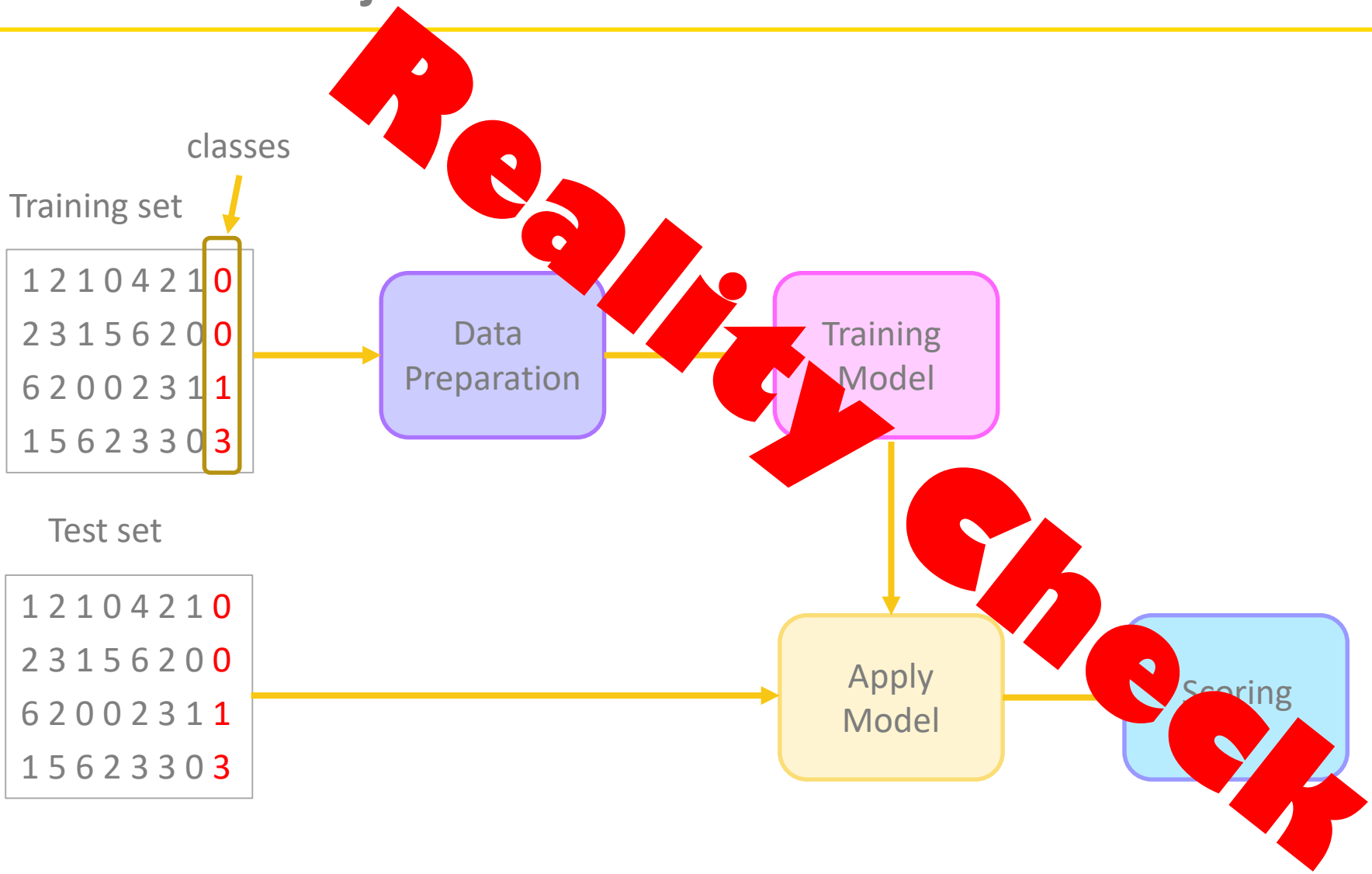
Optimizing for Hanukkah: Sometimes it's still strings, not things

Hannukah
Chanukah
Hanukah
Channukah
Chanuka
Chanukkah
Hanukka
Chanukka
Hannukkah
Ḥanukkah
Channuka
Festival of Lights
Feast of Dedication

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

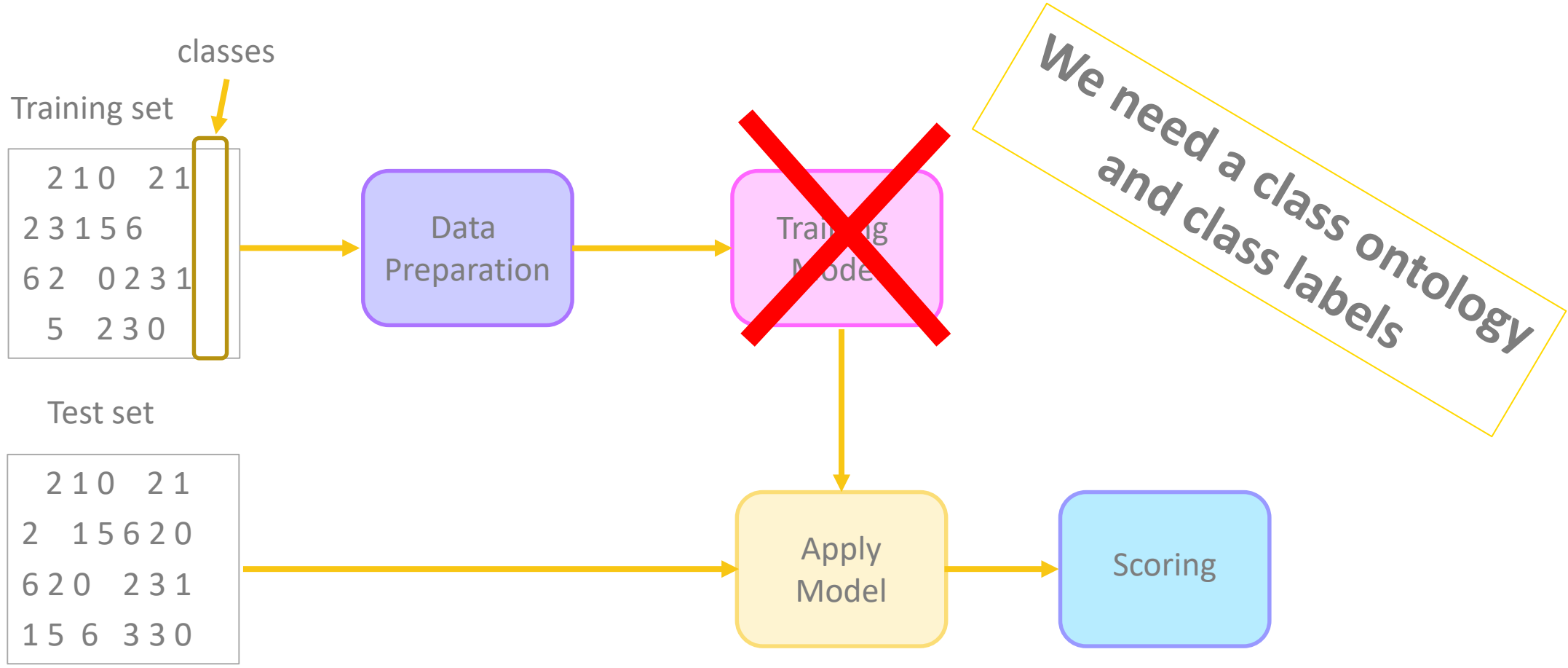
A Data Science Project



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Reality Check

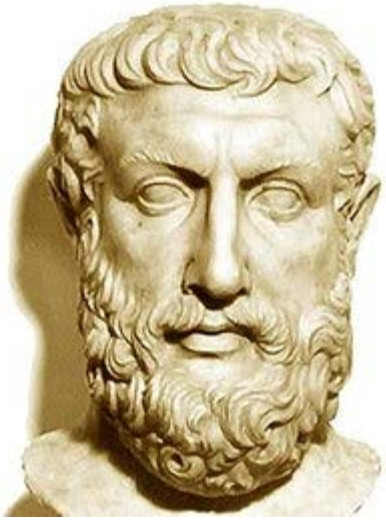


© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Ontology Definition

- **Ontology** is the philosophical study of the nature of being, becoming, existence, or reality, as well as the basic categories of being and their relations
- Introduced by Greek philosophers (Parmenides)
- Parmenides was among the first to propose an ontological characterization of the fundamental nature of reality
- In computer science and information science, an **ontology** is a formal naming and definition of the types, properties, and interrelationships of entities

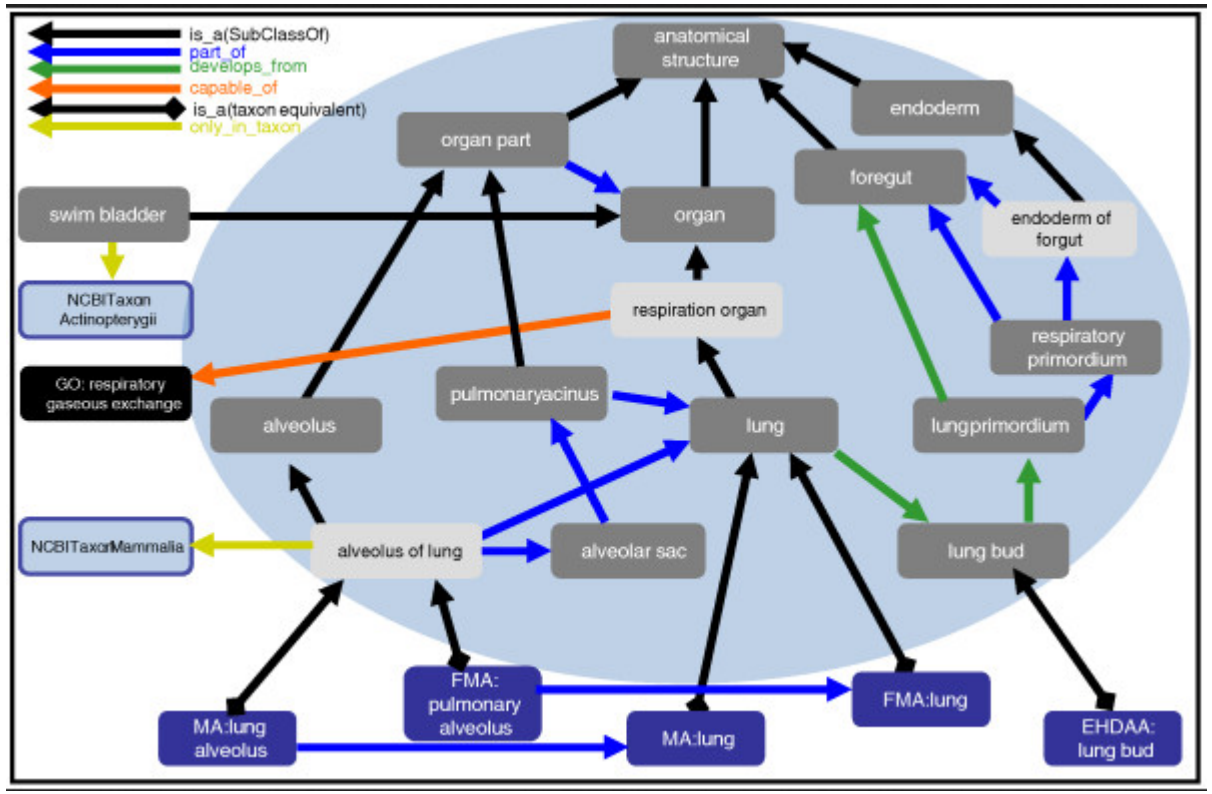


© 2018 KNIME AG. All Rights Reserved.

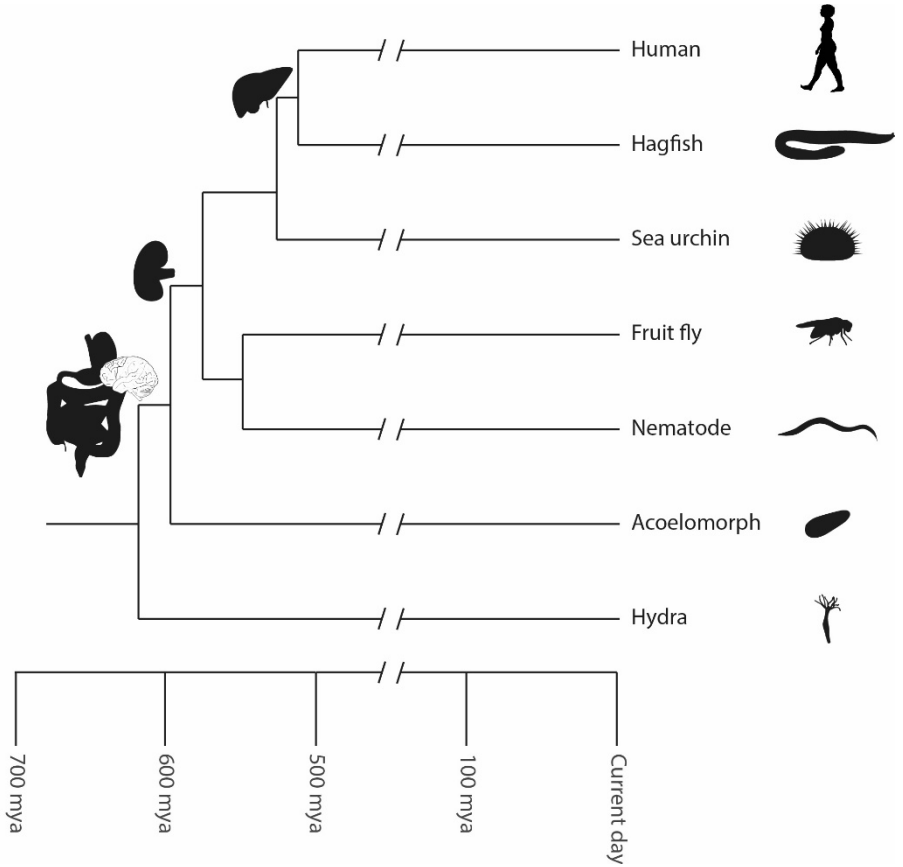
Open for Innovation KNIME

Ontology Example

Uberon-an-integrative-multi-species-anatomy-ontology-gb-2012-13-1-r5-2.jpg



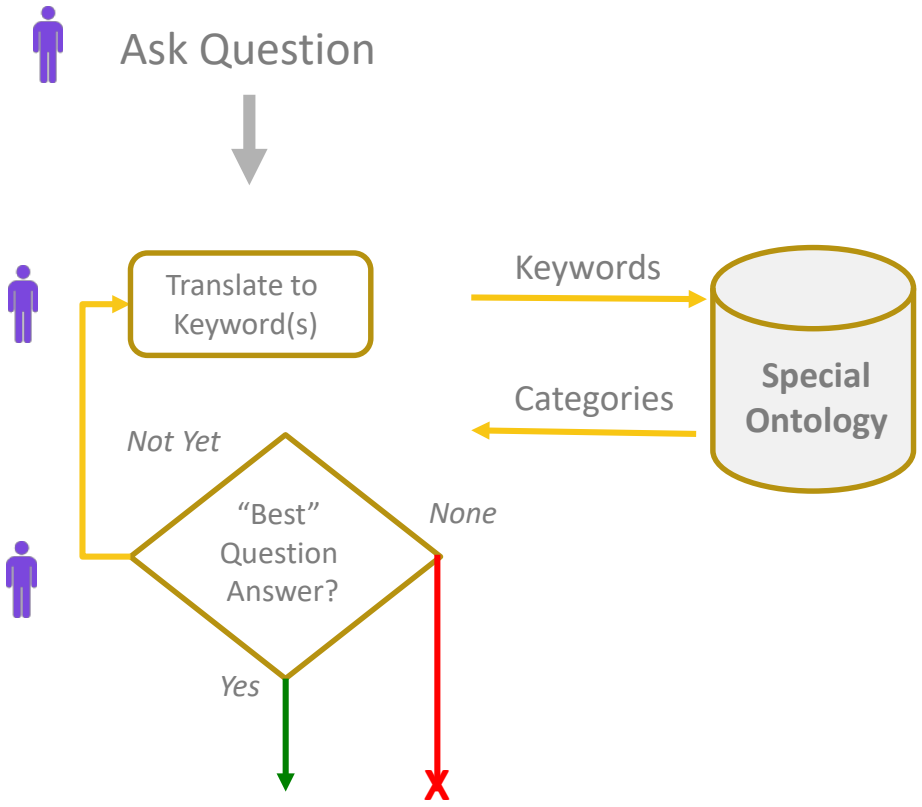
Relationship of major animal lineages with indication of how long ago these animals shared a common ancestor. On the left, important organs are shown, which allows us to determine how long ago these may have evolved.



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

The Specialist Topic Search Process



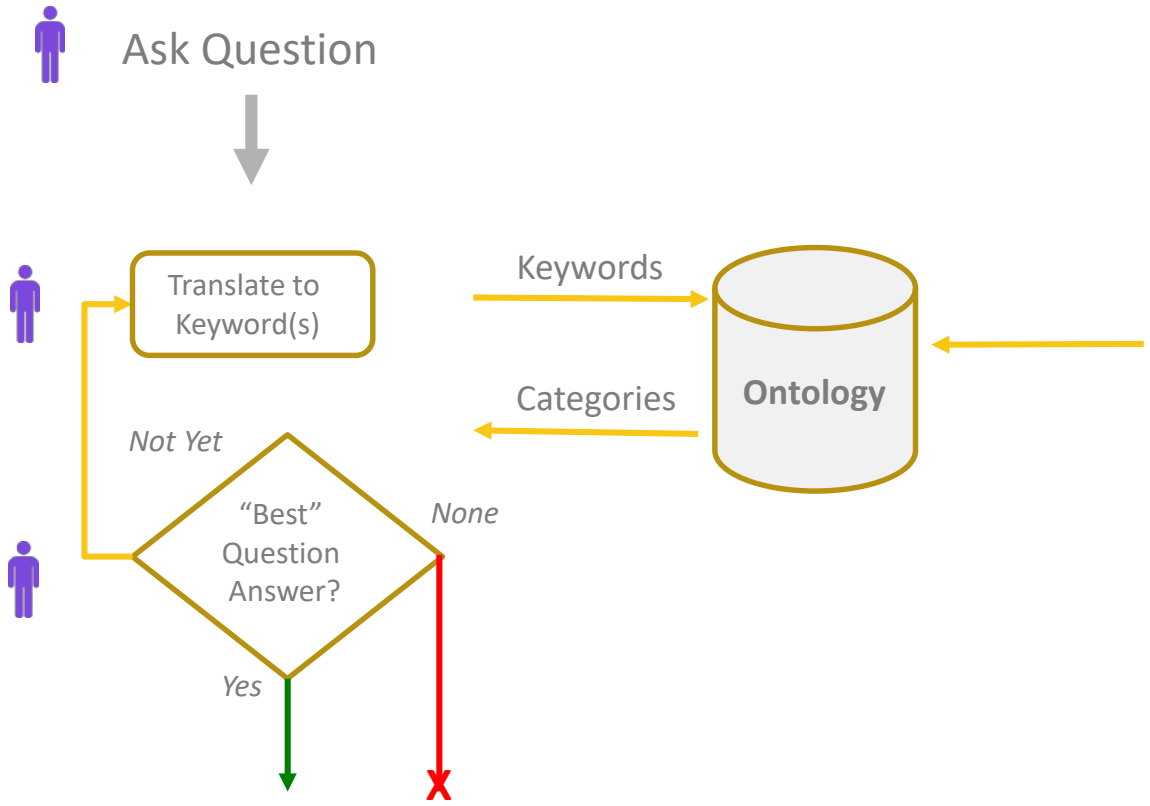
Medicine
- symptoms
- diseases
- treatments

Pharmaceutical
- drugs
- dosages
- allergies

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Creating an Ontology: Simple! Build Context



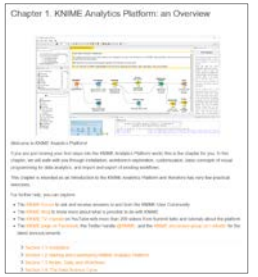
© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

A Real World Ontology Need: I want to learn KNIME

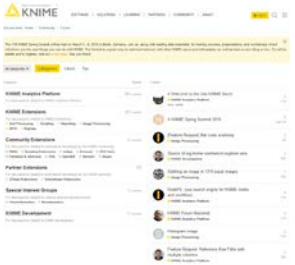
- I have a question…..
  - Terms
  - Concepts
  - Context
  - Background
  - Depth
  - Breadth
  - Language

E-learning



Blog

Forum



Other

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

Our Own Ontology (20 Classes)

From e-Learning Course

- Installation
- Data Access
- ETL
- Mining
- Control
- Deployment
- DataViz

From other Resources

- Use Cases
- Text Processing
- Big Data
- Server
- Image Processing
- Reporting

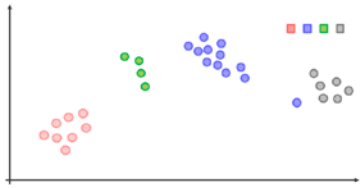
From Experience

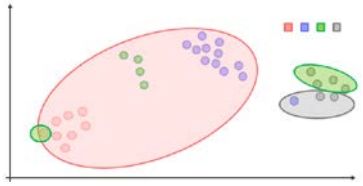
- Development
- Integration
- Optimizing KNIME
- Life Science
- Announcement
- Bug
- Legal

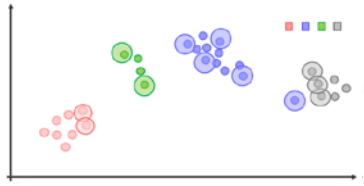
© 2018 KNIME AG. All Rights Reserved.

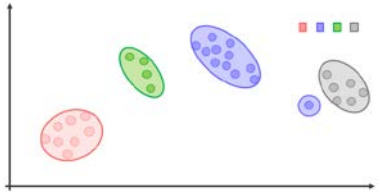
Open for Innovation ® KNIME

Active Learning Cycle









Training Set
[Forum Questions]

1st attempt Class Labels

Model Training

Extract most uncertain predictions

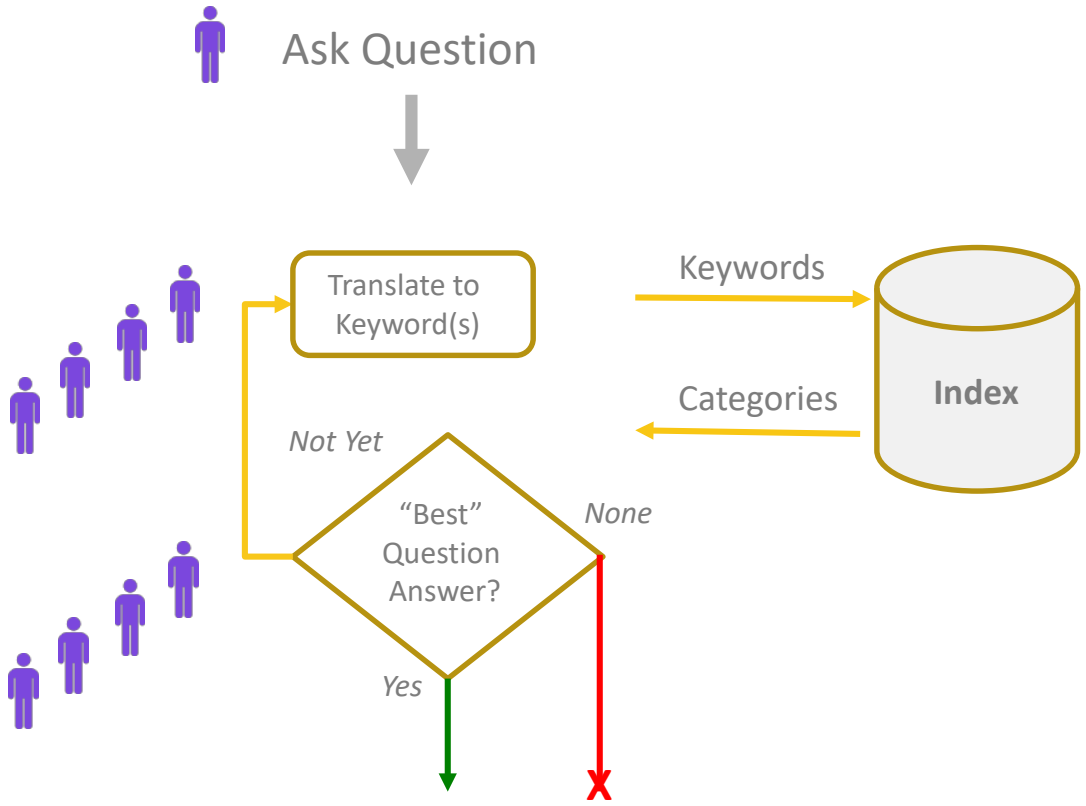
Re-labeling

Class Label Extension

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

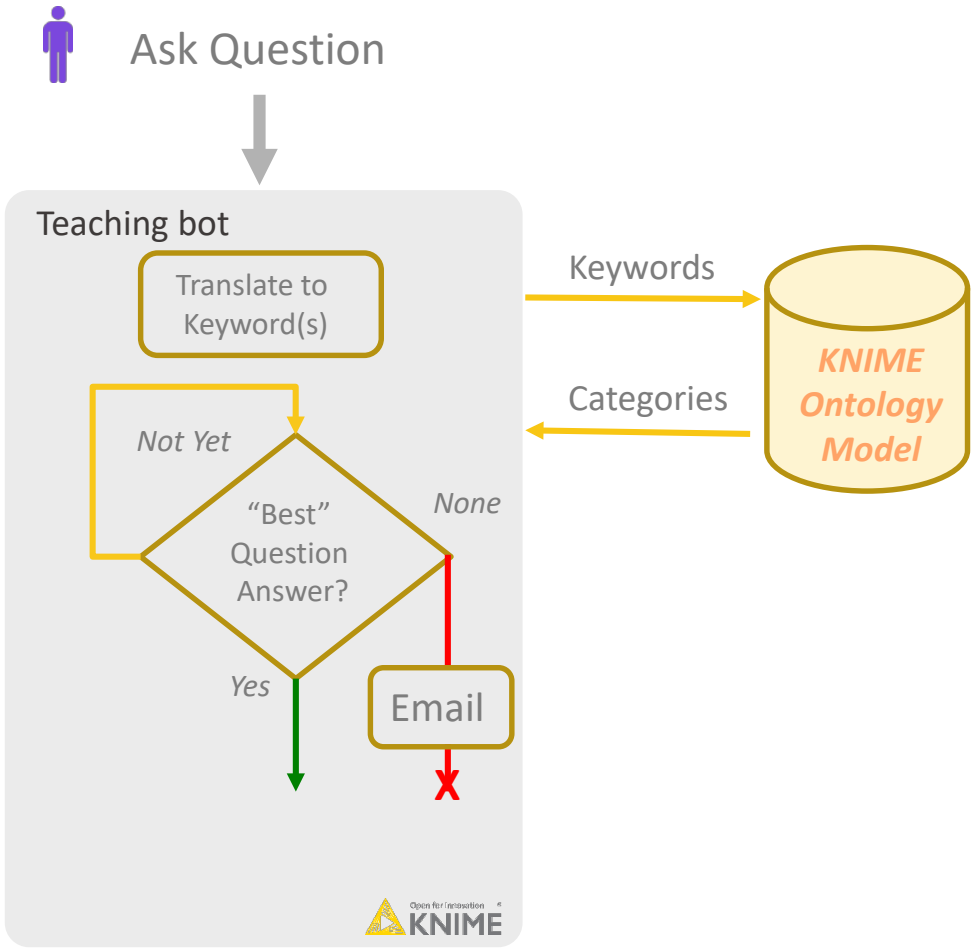
The Human “Learn KNIME” Process



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

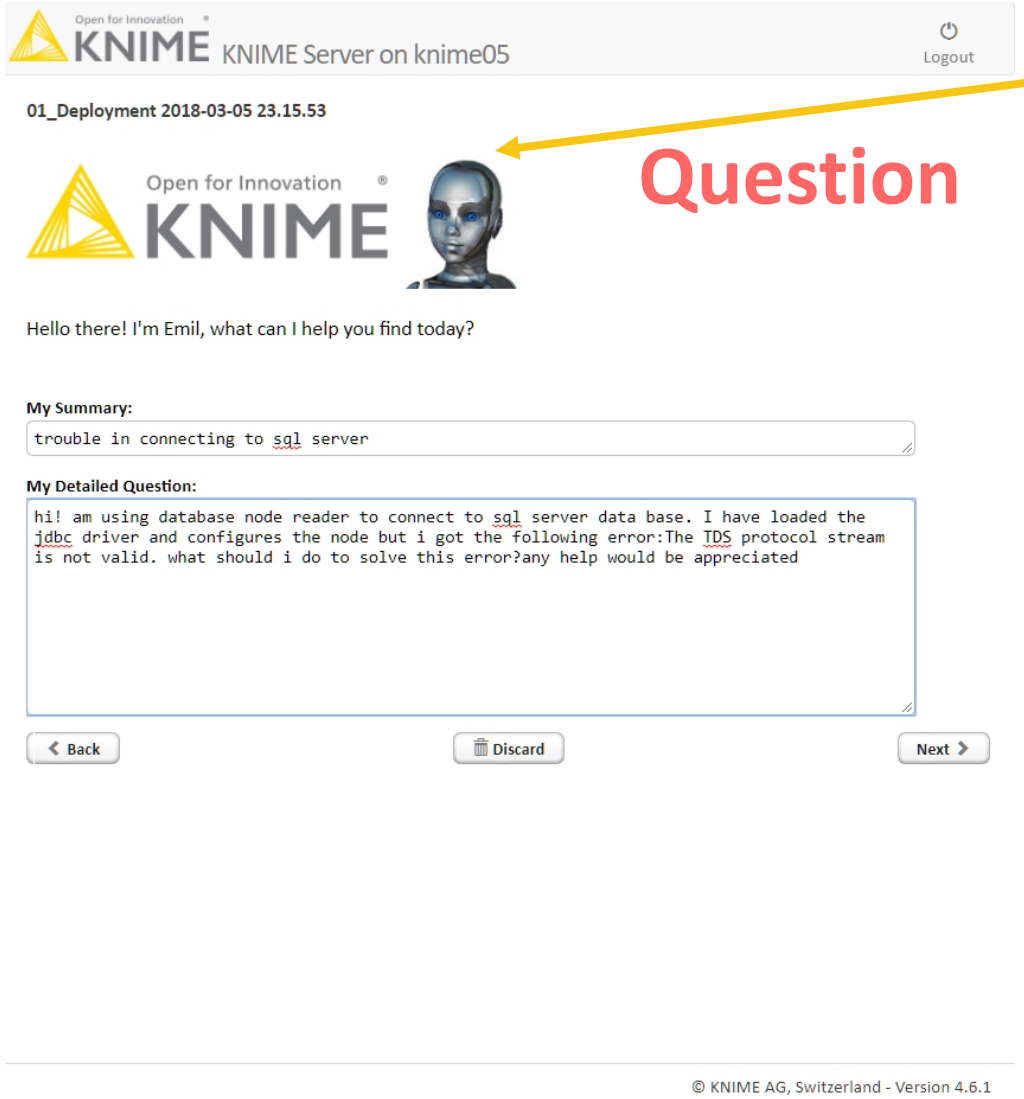
Teacher Bot “Emil”: A Bot to help Learn KNIME



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Teacher Bot “Emil”



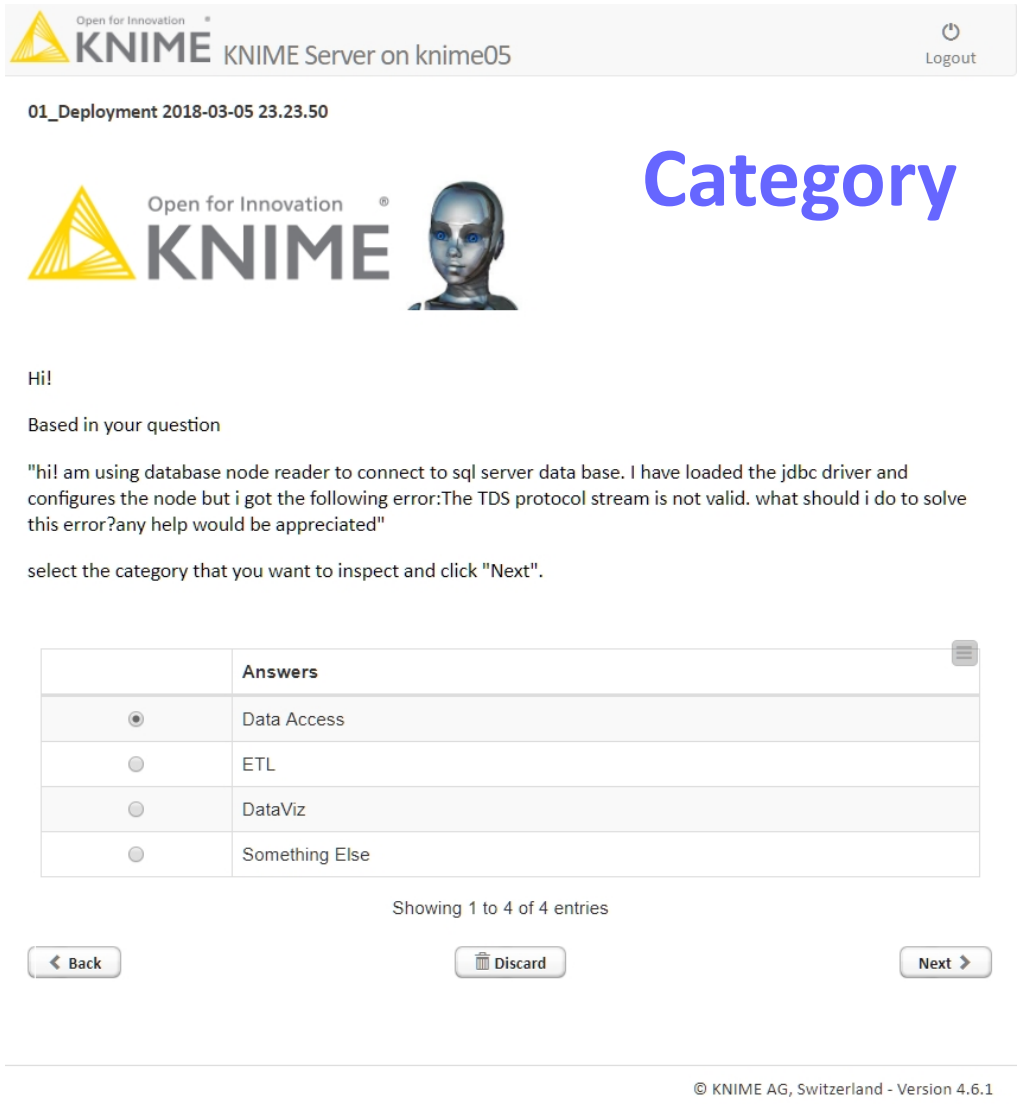
Emil, our Teacher Bot!

Question

© 2018 KNIME AG. All Rights Reserved.

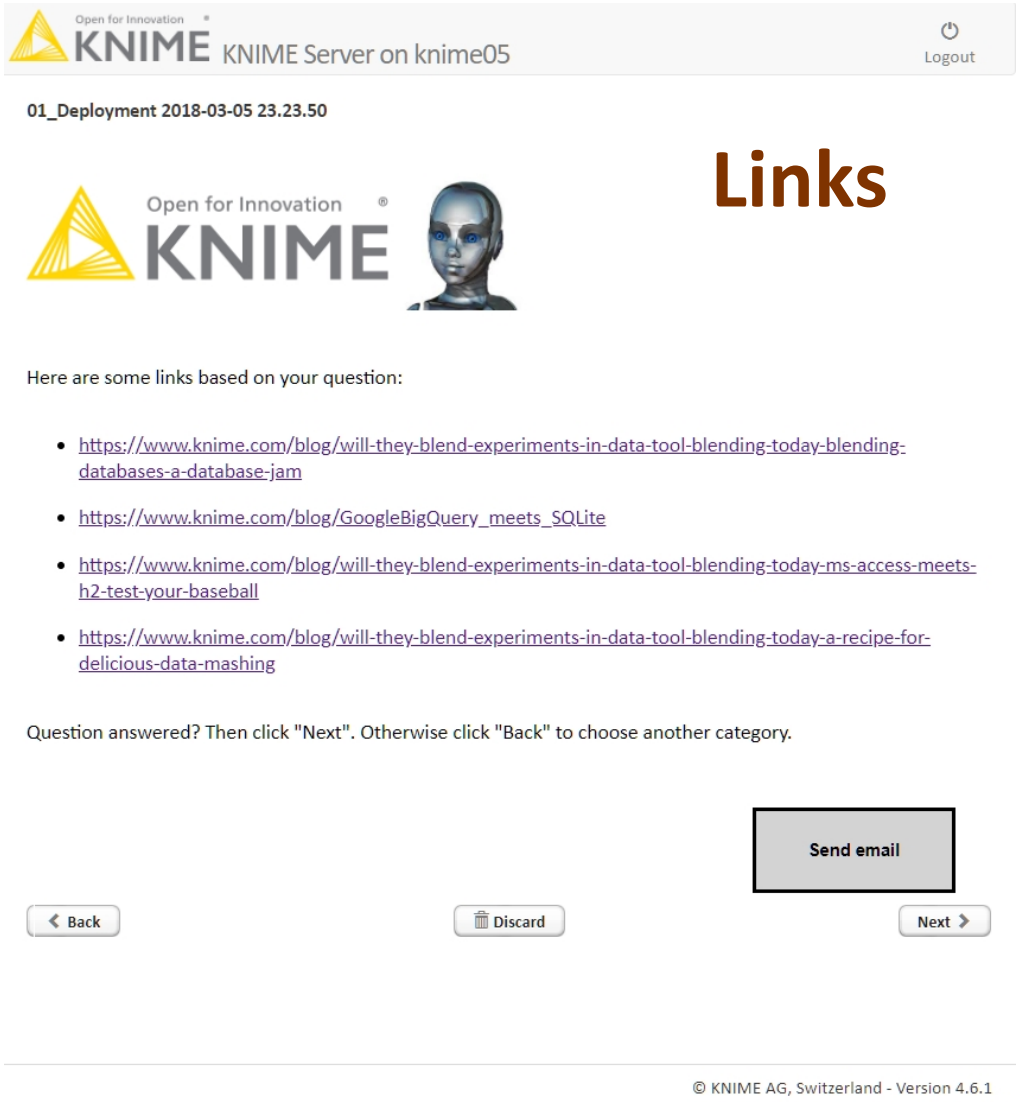
Open for Innovation KNIME

Teacher Bot “Emil”



© 2018 KNIME AG. All Rights Reserved.

Teacher Bot “Emil”



© 2018 KNIME AG. All Rights Reserved.

Teacher Bot “Emil”

Will They Blend? Experiments in Data & Tool Blending. Today: Blending Databases. A Database Jam Session

Links

View Edit Delete Revisions

Mon, 04/10/2017 - 15:19 — rs

In this blog series we'll be experimenting with the most interesting blends of data and tools. Whether it's mixing traditional sources with modern data lakes, open-source devops on the cloud with protected internal legacy tools, SQL with noSQL, web-wisdom-of-the-crowd with in-house handwritten notes, or IoT sensor data with idle chatting, we're curious to find out: will they blend? Want to find out what happens when IBM Watson meets Google News, Hadoop Hive meets Excel, R meets Python, or MS Word meets MongoDB?

Follow us here and send us your ideas for the next data blending challenge you'd like to see at willtheyblend@knime.com.

Today: Blending Databases. A Database Jam Session

The Challenge

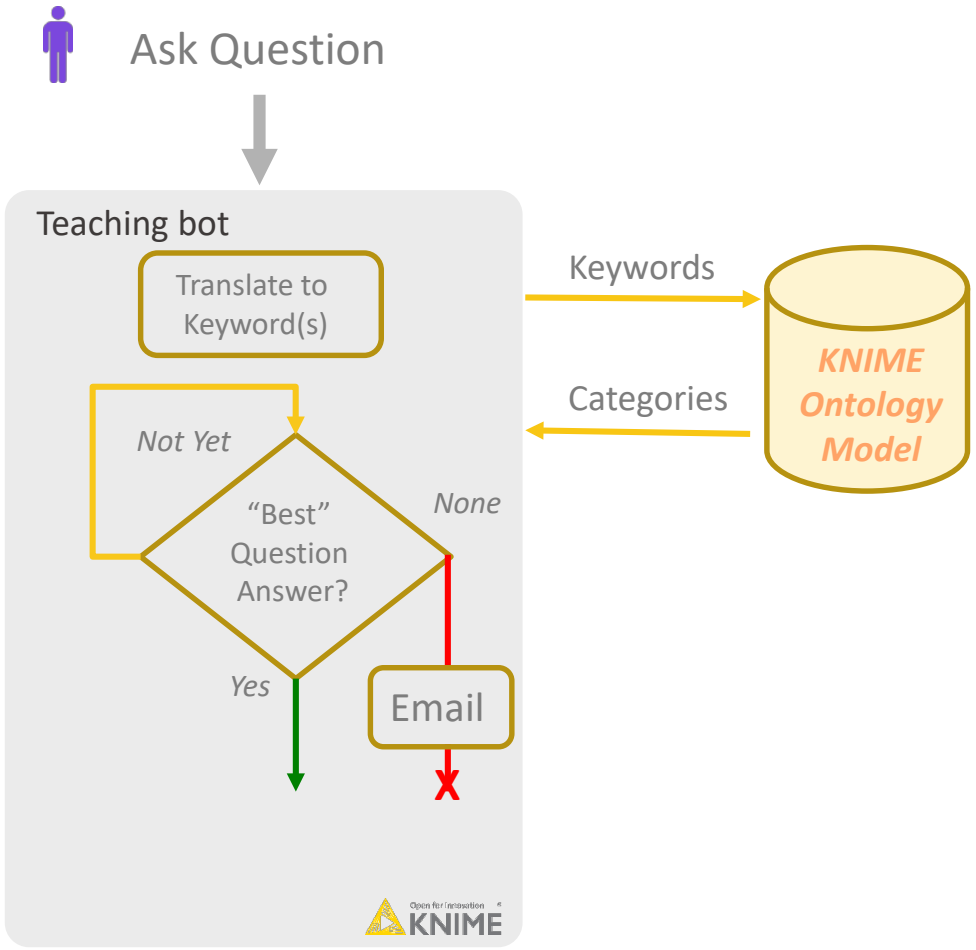
Today we will push the limits by attempting to blend data from not just 2 or 3, but 6 databases!

These 6 SQL and noSQL databases are among the top 10 most used databases, as listed in most database comparative web sites (see DB-Engines Ranking, The 10 most popular DB Engines …, Top 5 best databases). Whatever database you are using in your current data science project, there is a very high probability that it will be in our list today. So, keep reading!

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

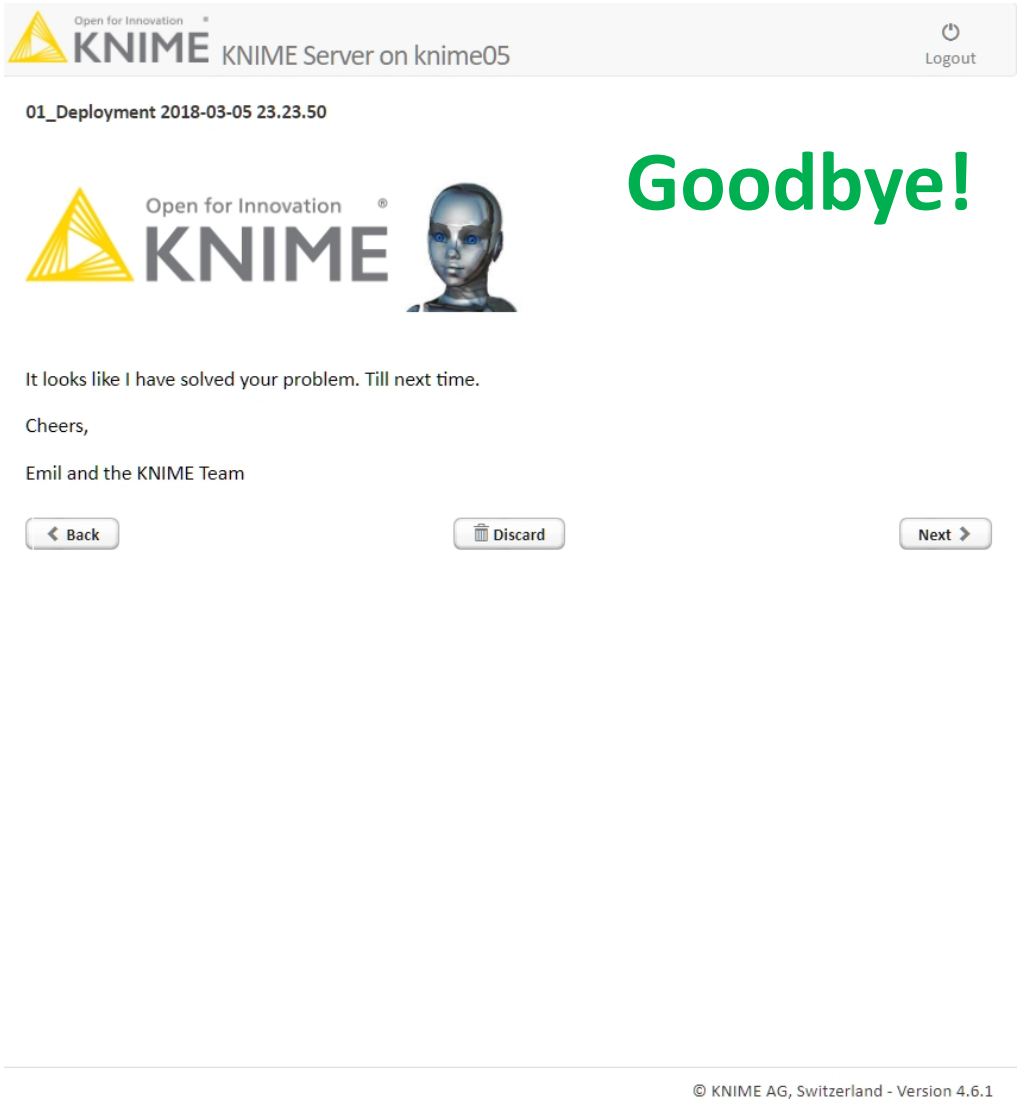
Teacher Bot “Emil”: A Bot to help Learn KNIME



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

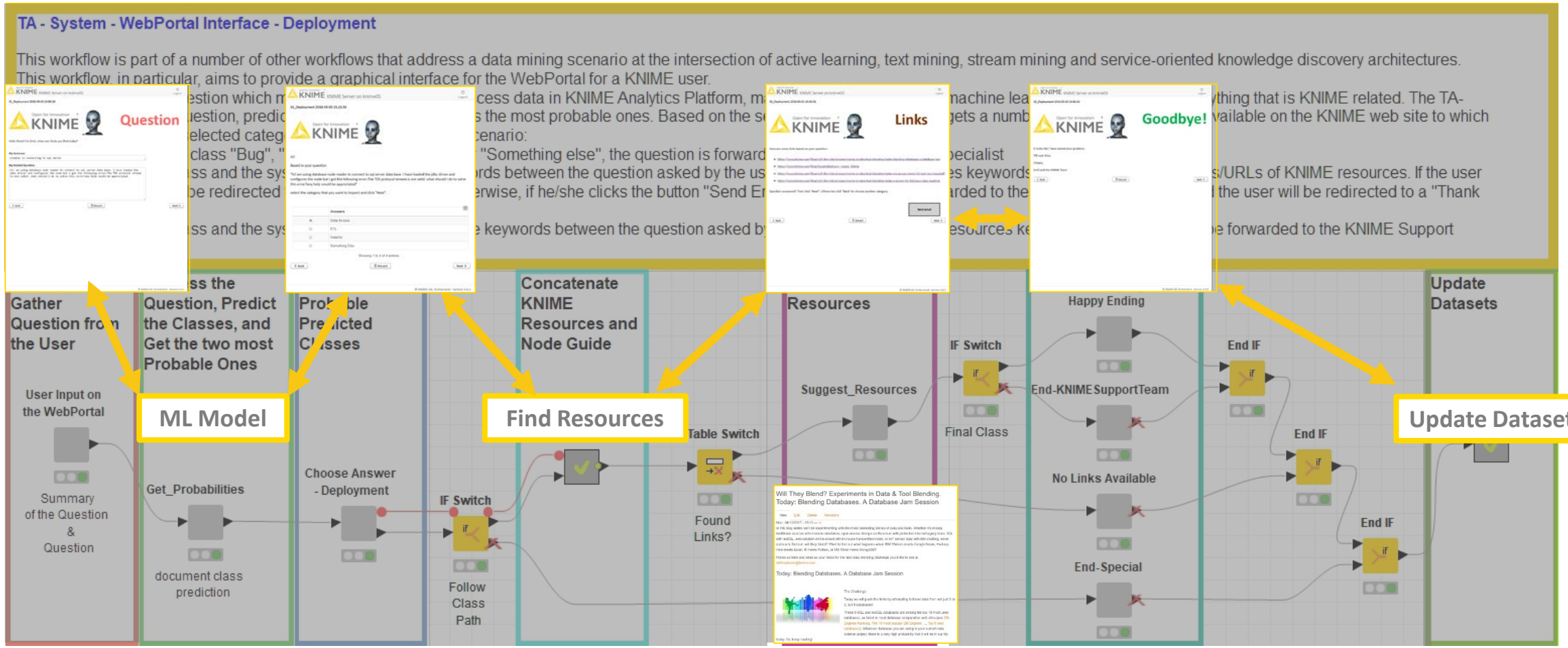
Teacher Bot “Emil”



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

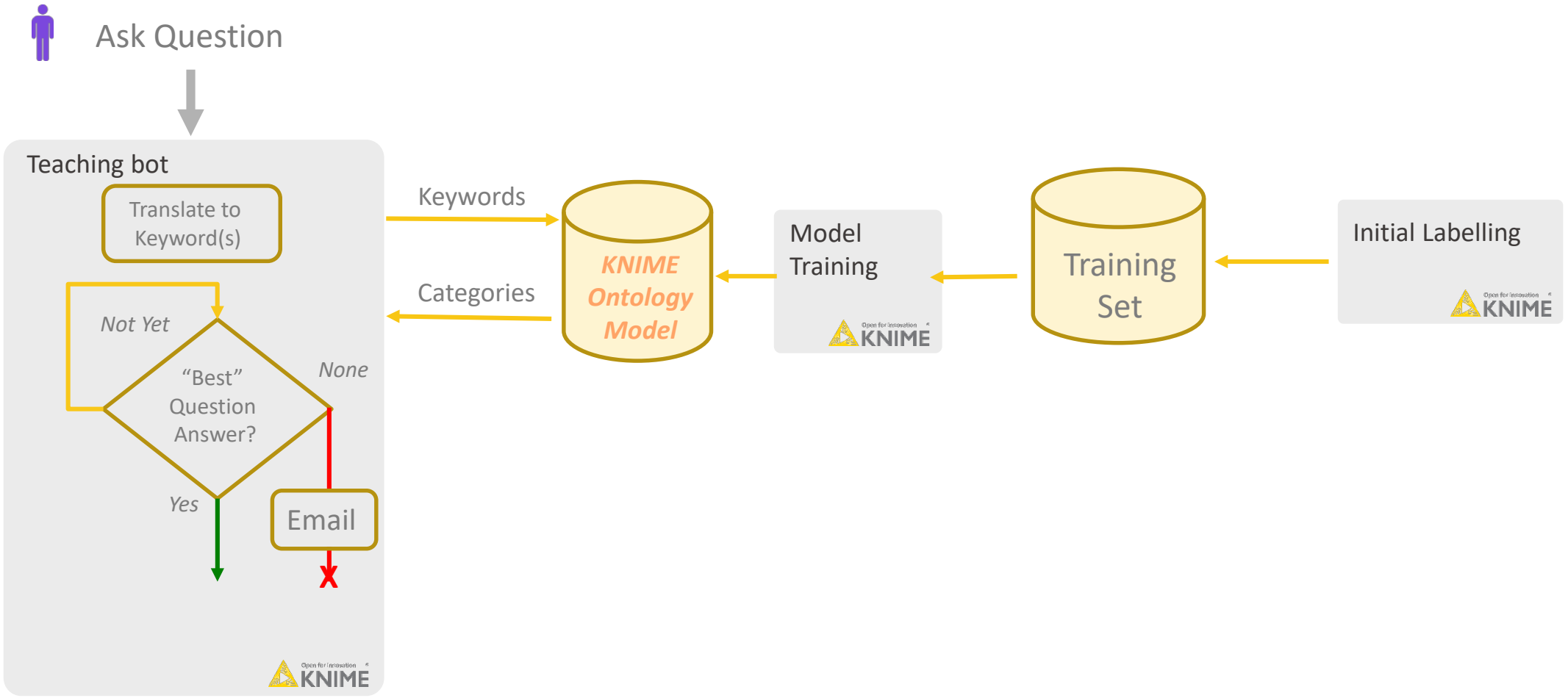
Teacher Bot “Emil”



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Creation of an Initial Ontology

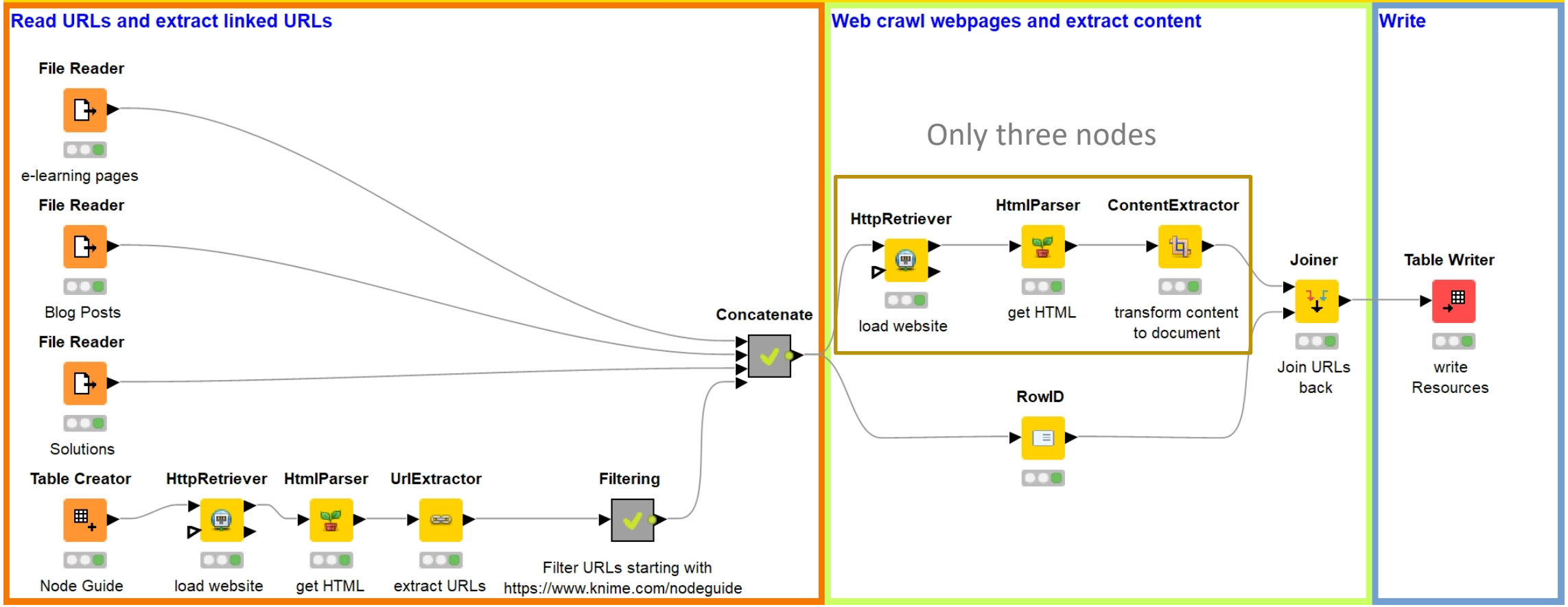


© 2018 KNIME AG. All Rights Reserved.

Web Crawling KNIME Resources

Web Crawling of the KNIME Resources

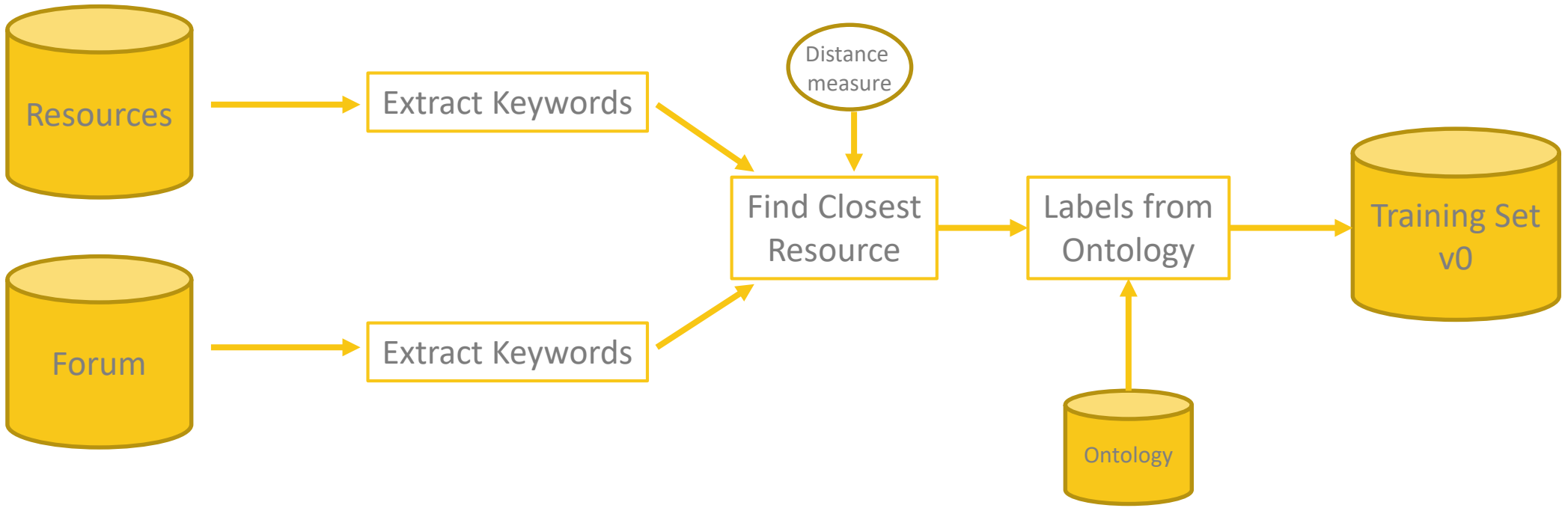
This workflow reads in the URLs of the different resources available on the KNIME website and extracts the content into a document cell.
In case of the node guide only the URL of the first page is given. The metanode extracts all links on the webpage.



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Step 0 – Initial Labeling

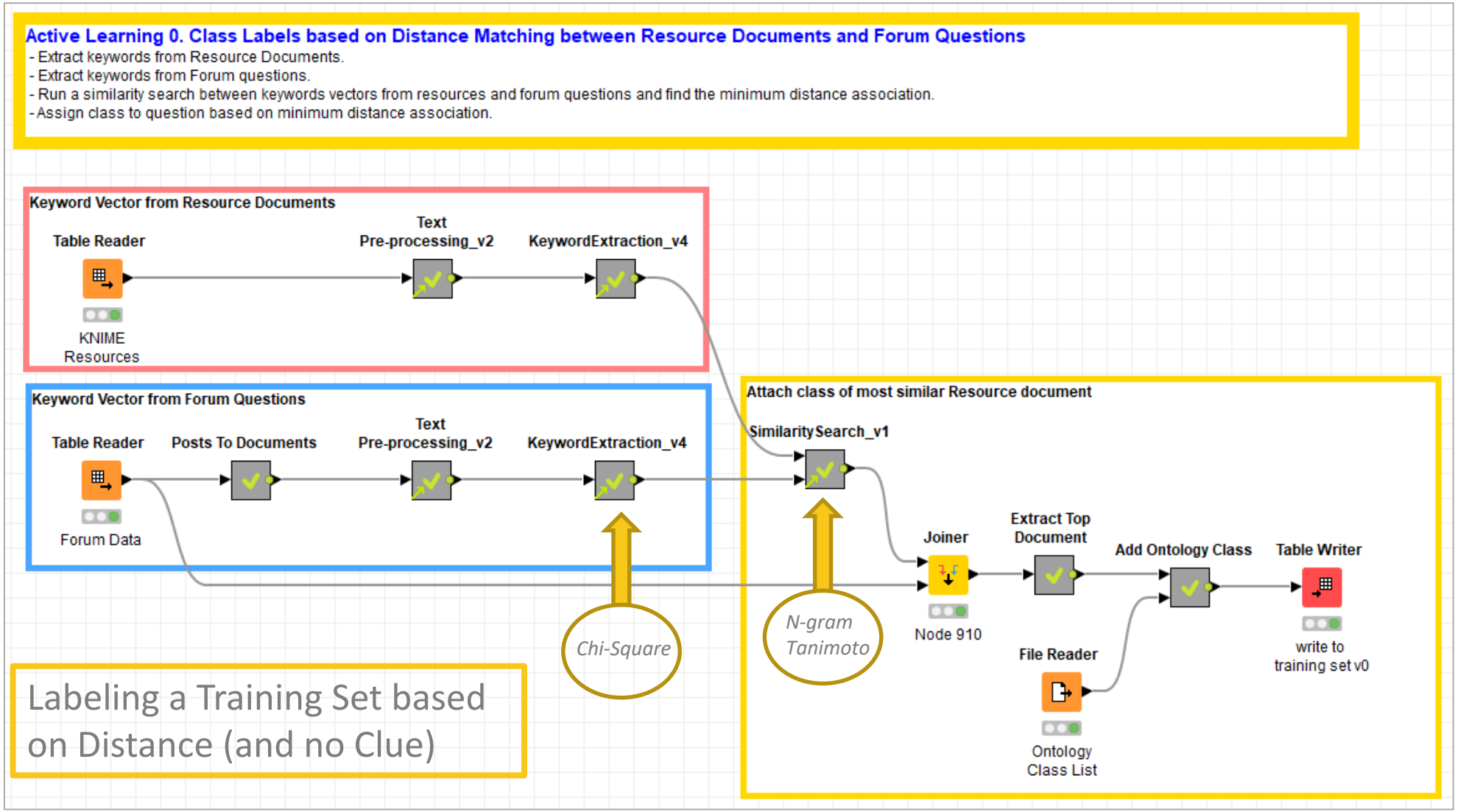


Labeling a Training Set based on Distance (and no Clue)

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Step 0 – Initial Labeling



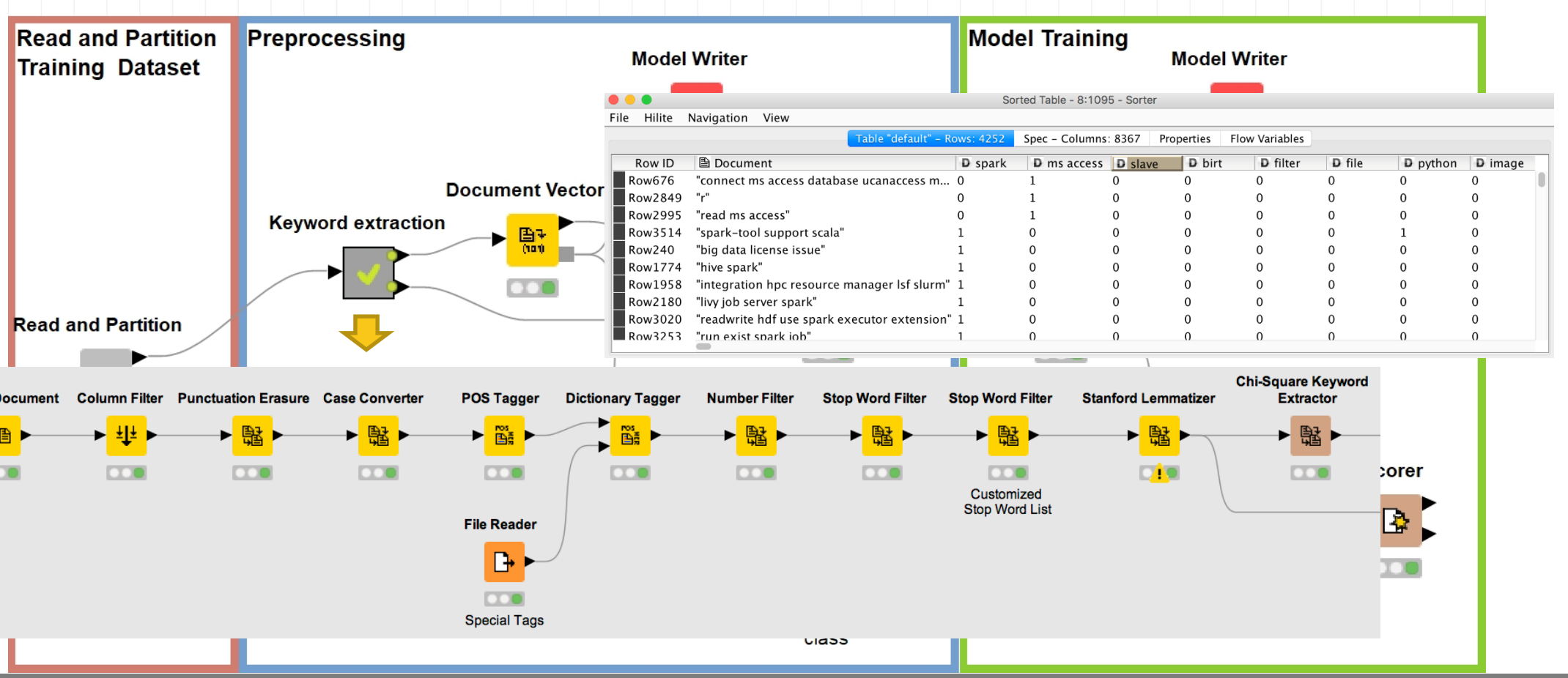
© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Step 1 - Model Training

Train Document Vector and Random Forest Model

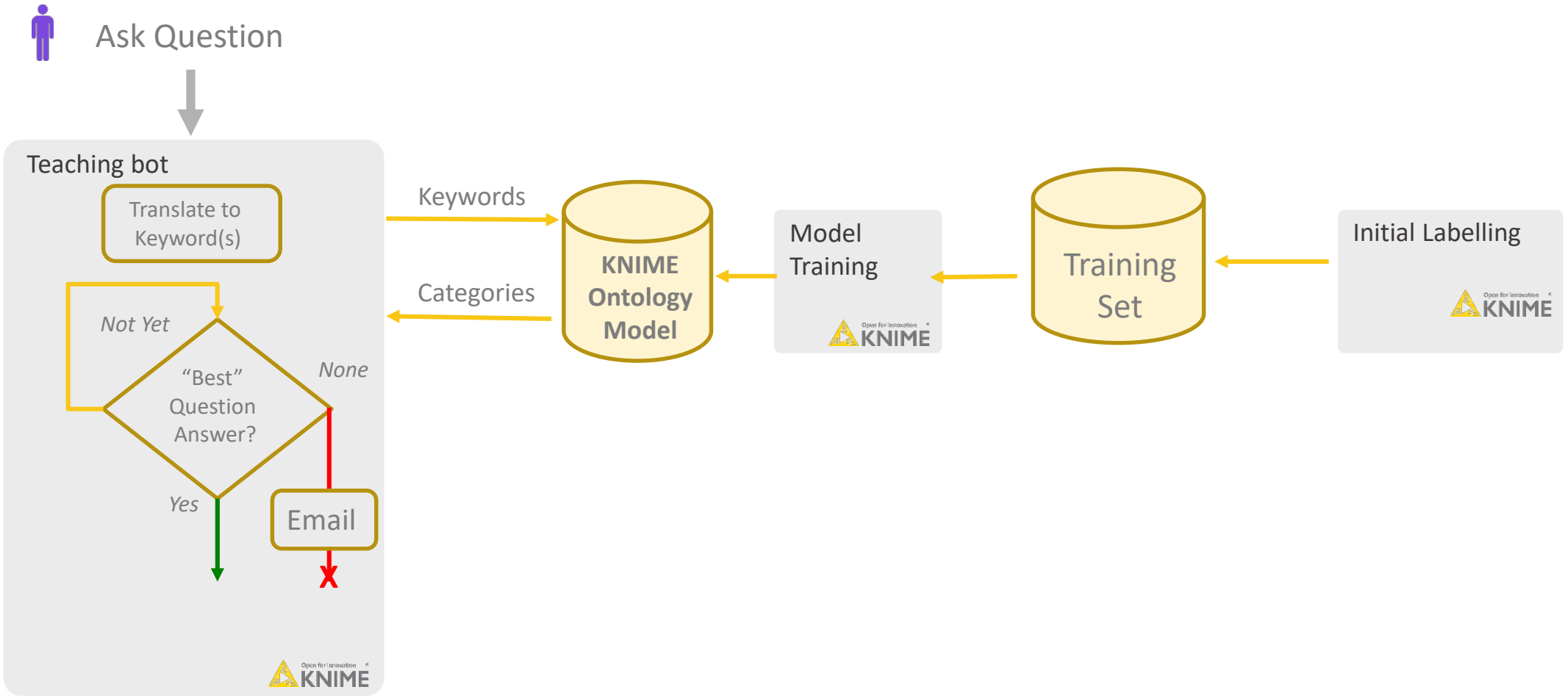
This workflow trains two models. A Document Vector Model depending on the keywords of the trainingset for the PreProcessing and a Random Forest model to make the prediction of the document_class.



© 2018 KNIME AG. All Rights Reserved.

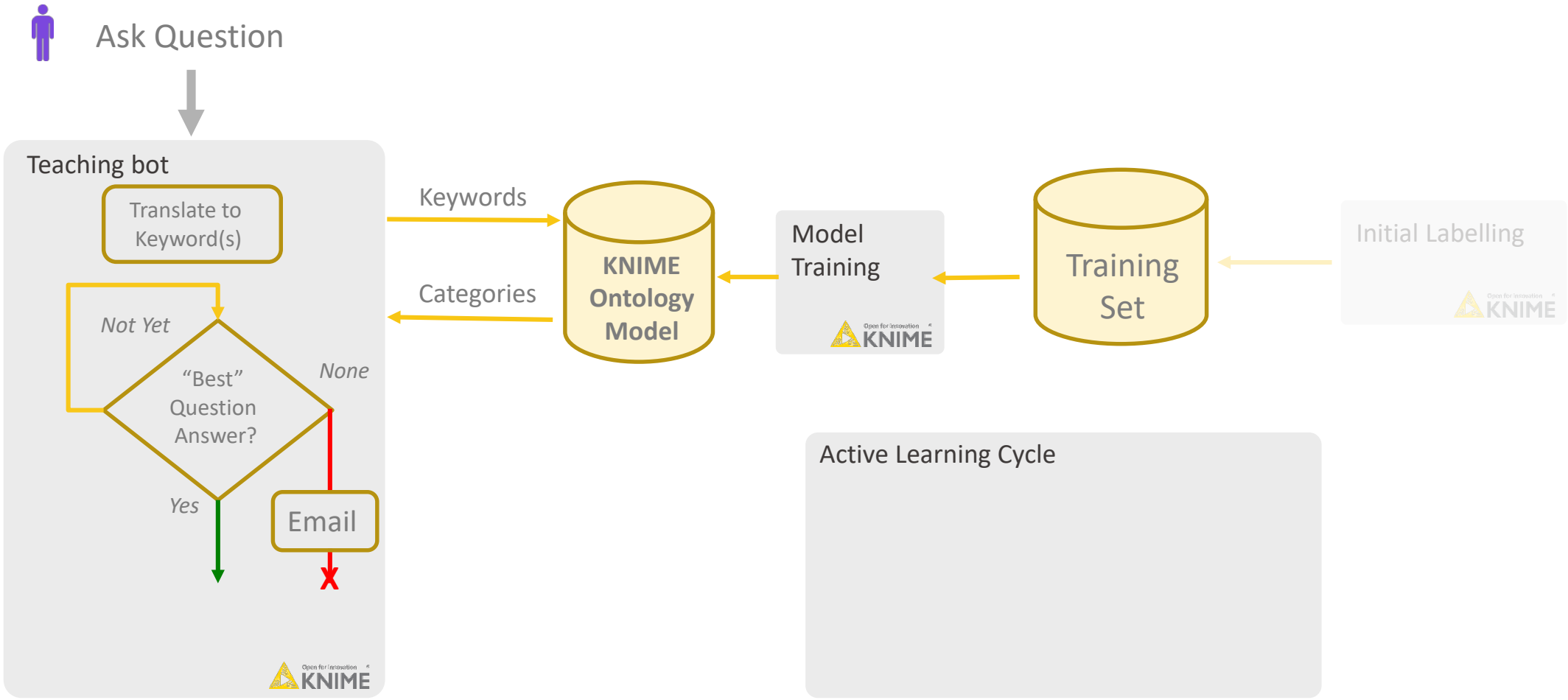
Open for Innovation KNIME

Creation of an Initial Ontology



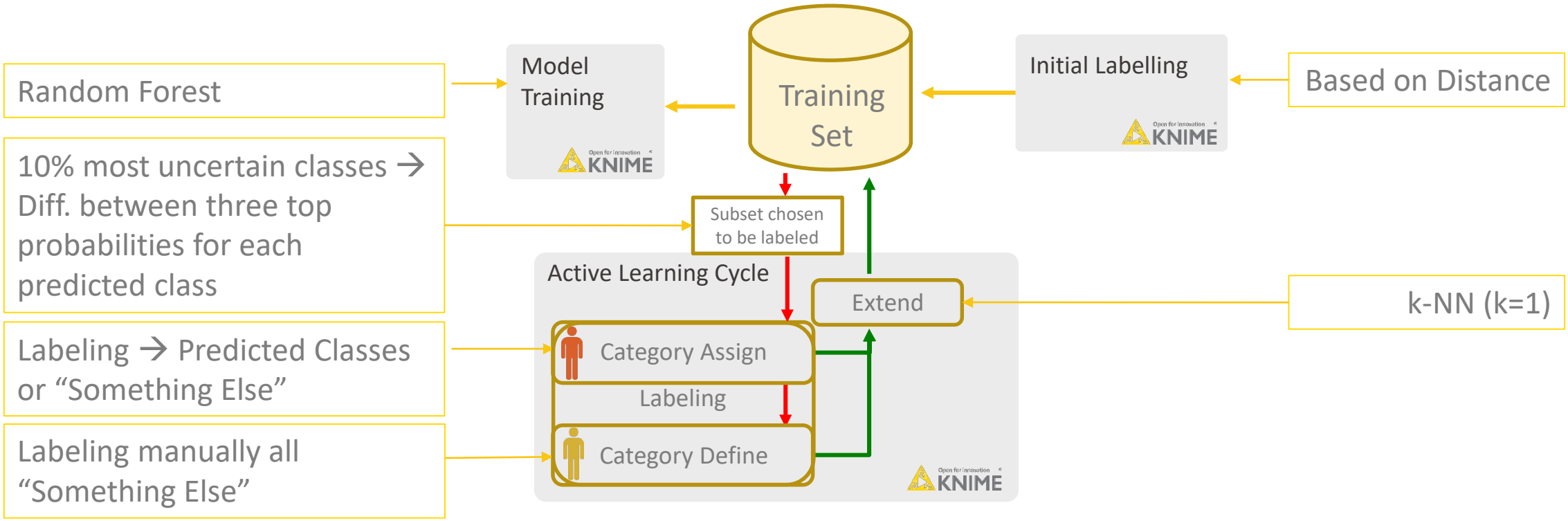
© 2018 KNIME AG. All Rights Reserved.

Adding Active Learning to the Cycle



© 2018 KNIME AG. All Rights Reserved.

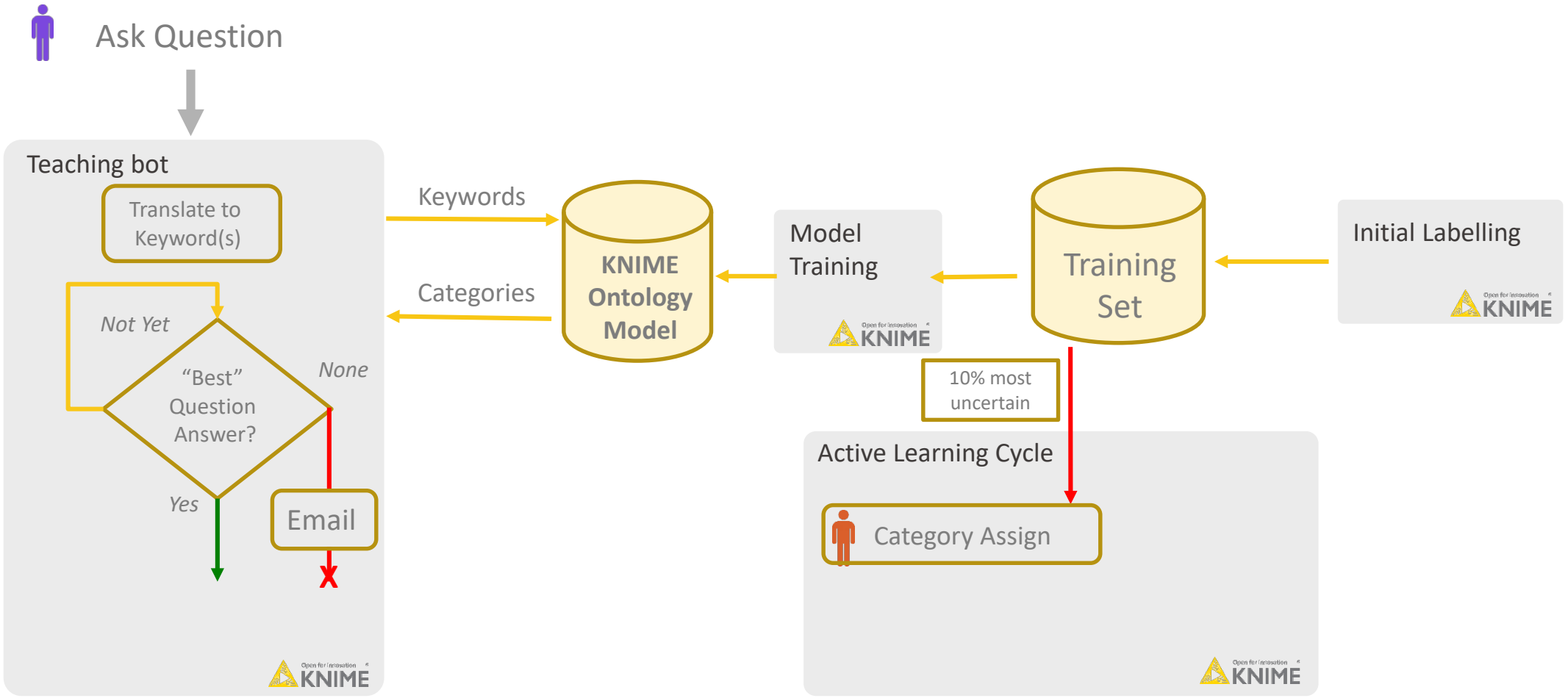
Active Learning



© 2018 KNIME AG. All Rights Reserved.

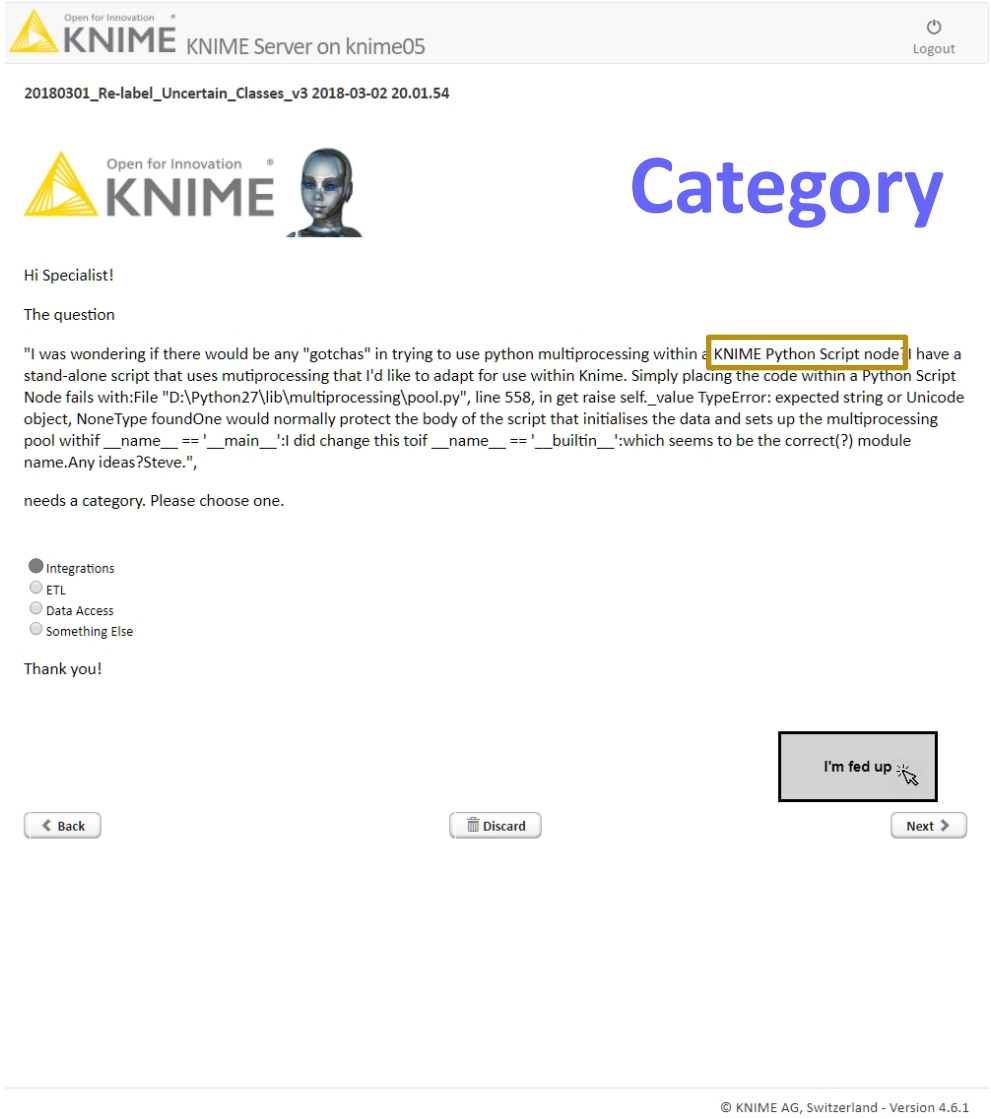
Open for Innovation KNIME

Adding Active Learning to the Cycle



© 2018 KNIME AG. All Rights Reserved.

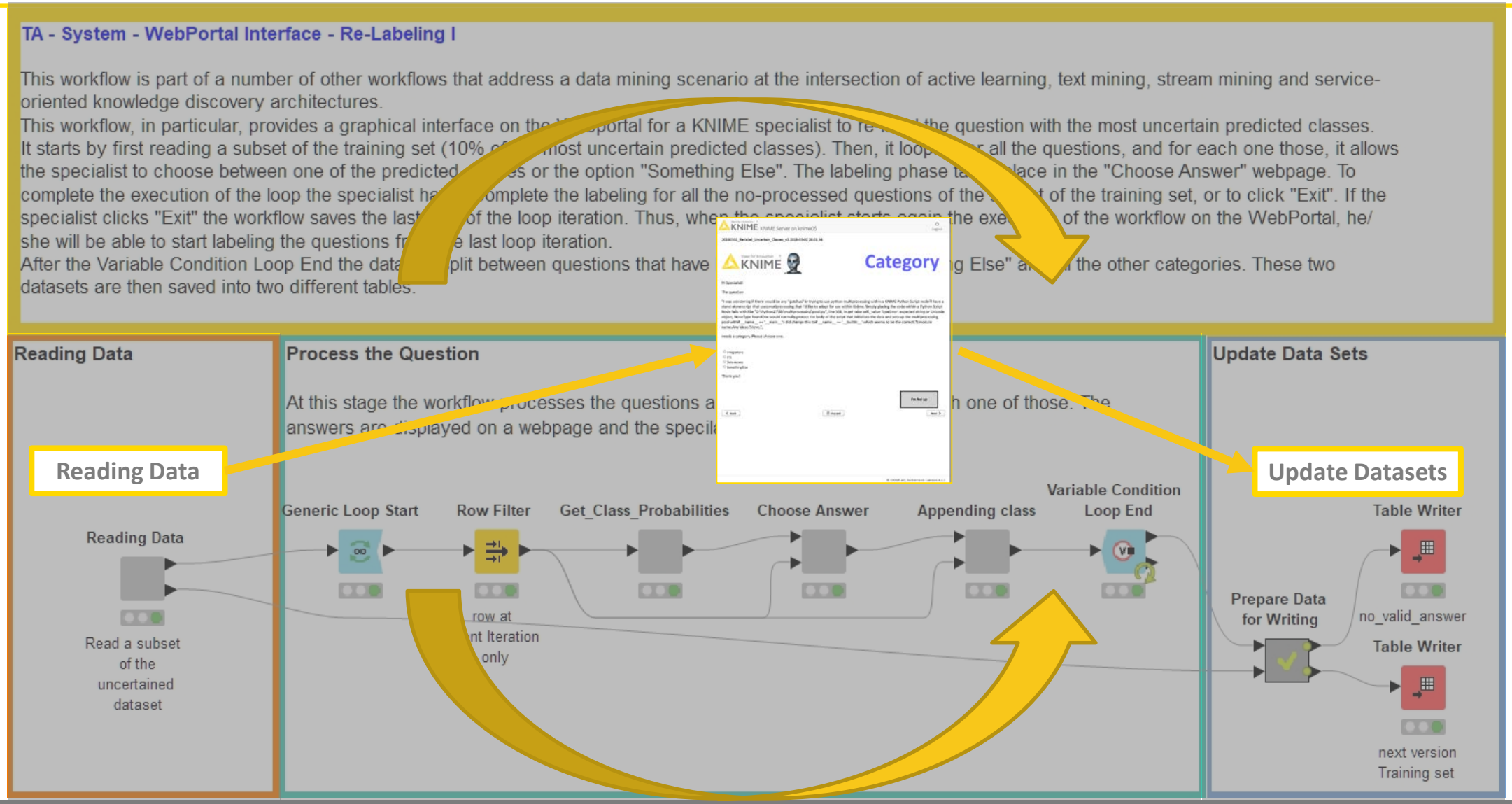
Category Assign



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation
KNIME

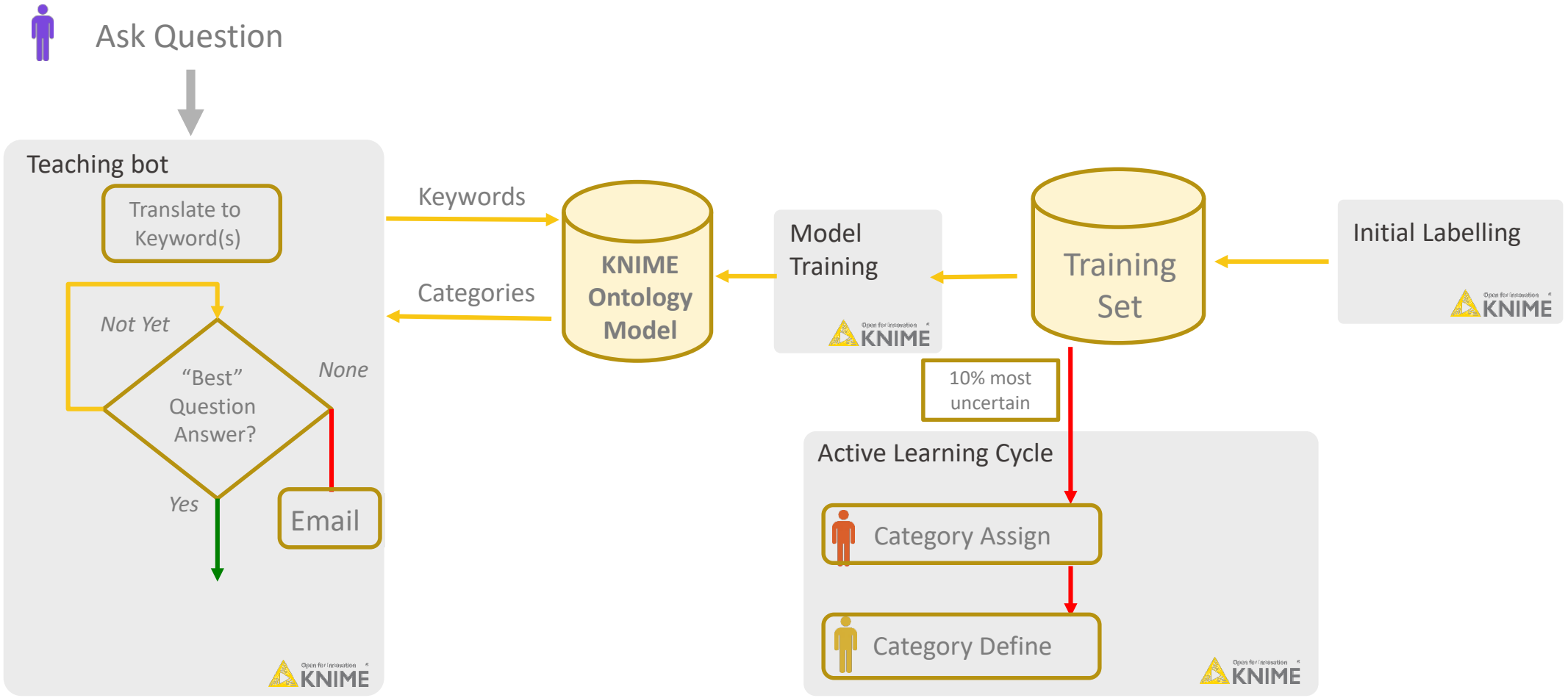
Step 2a - Category Assign



© 2018 KNIME AG. All Rights Reserved.

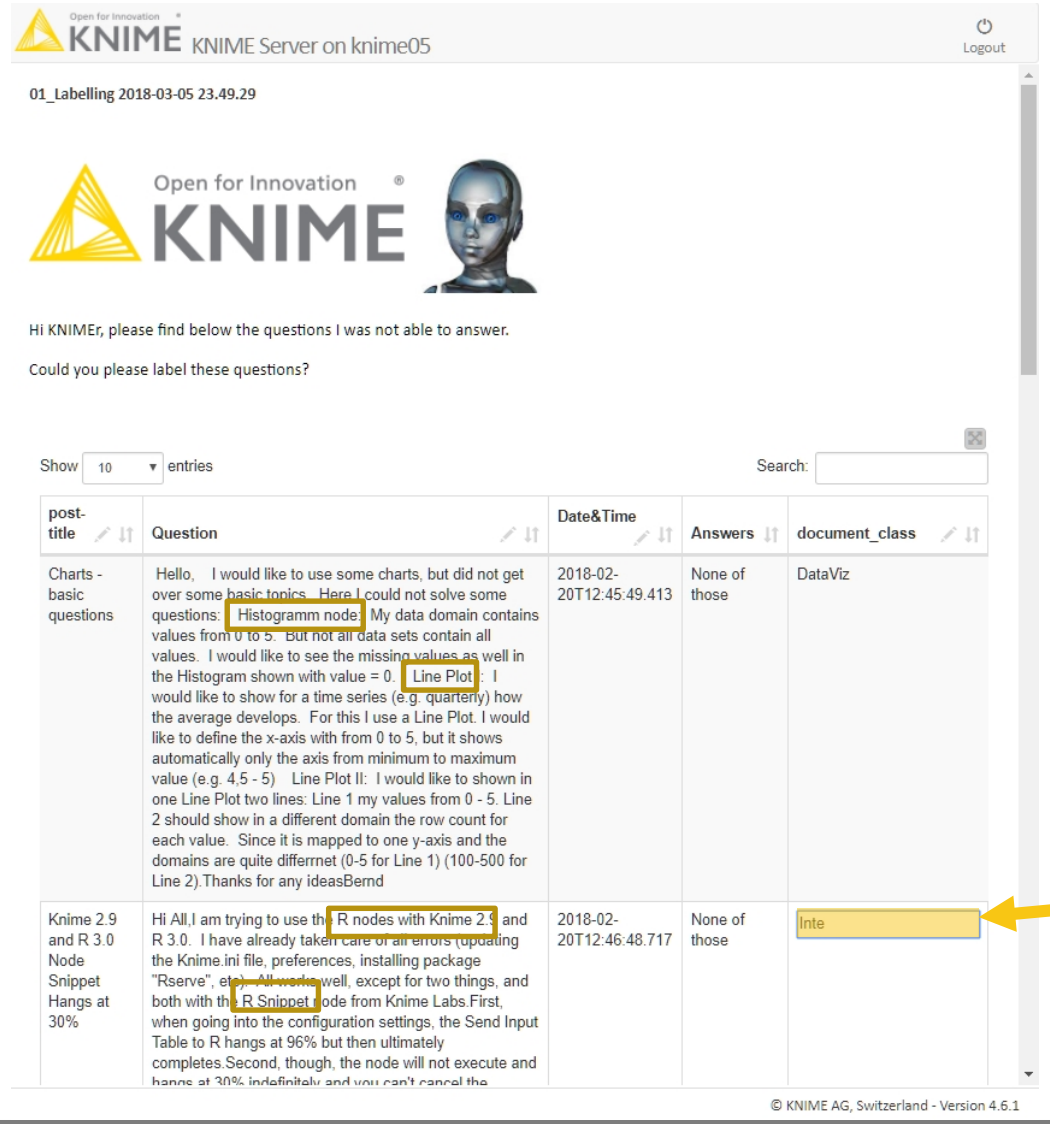
Open for Innovation KNIME

Adding Active Learning to the Cycle



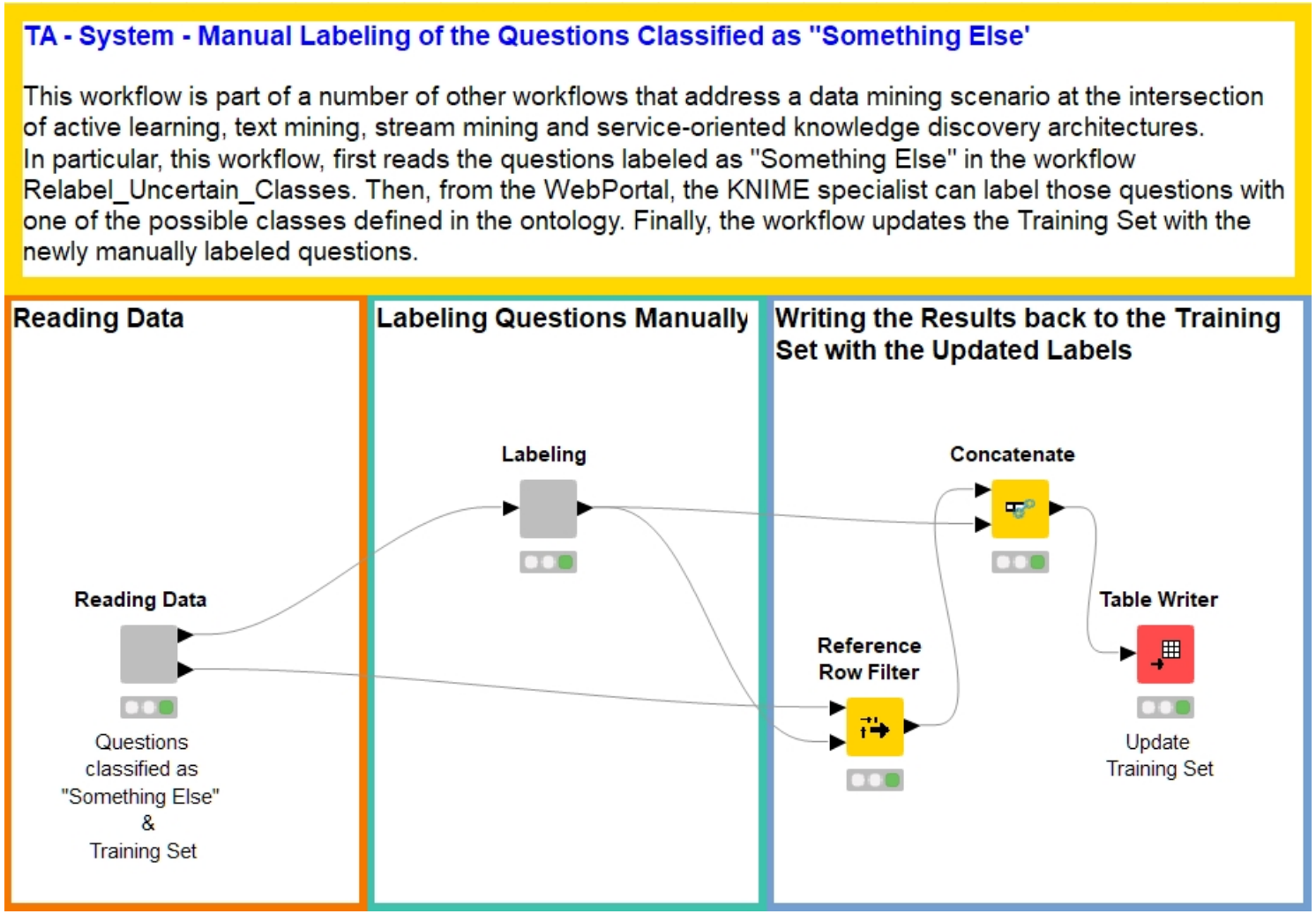
© 2018 KNIME AG. All Rights Reserved.

Step 2b - Category Define



© 2018 KNIME AG. All Rights Reserved.

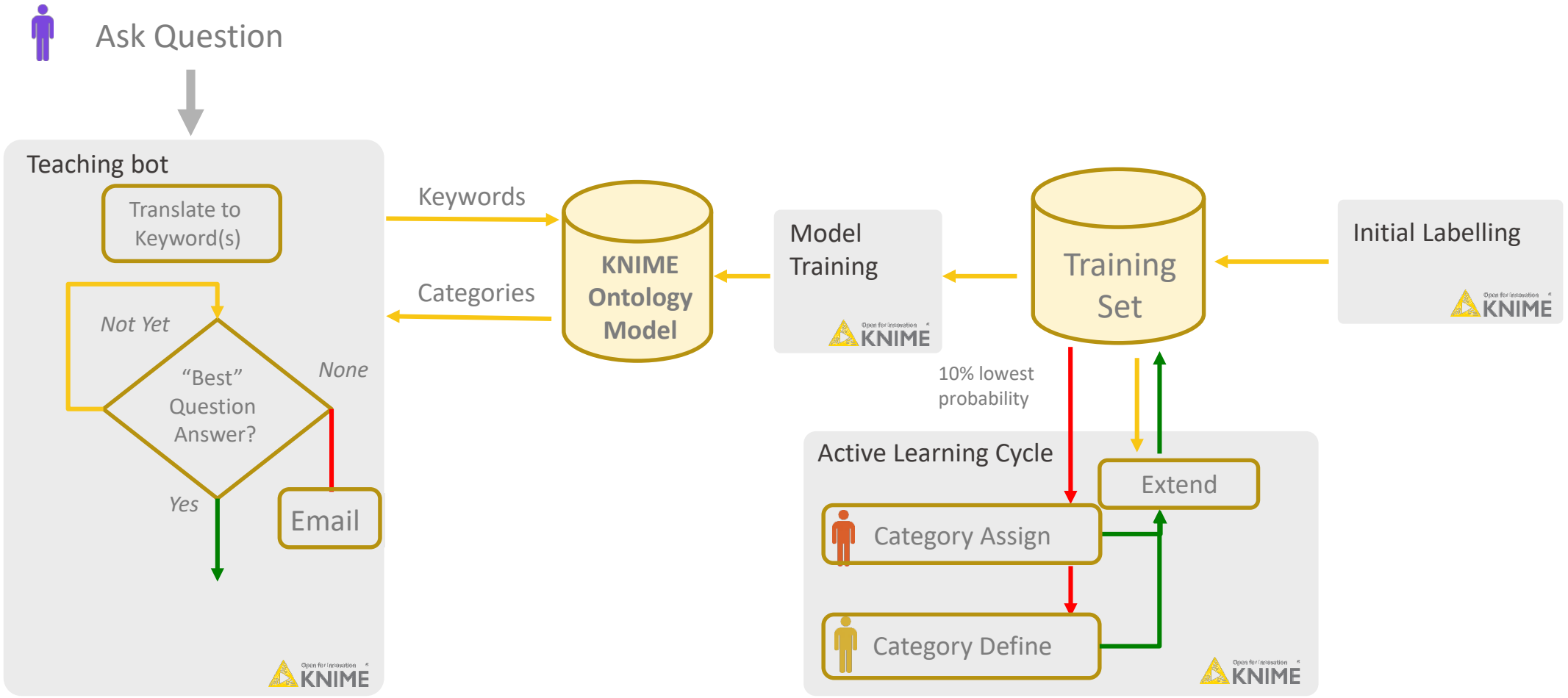
Step 2b - Category Define



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

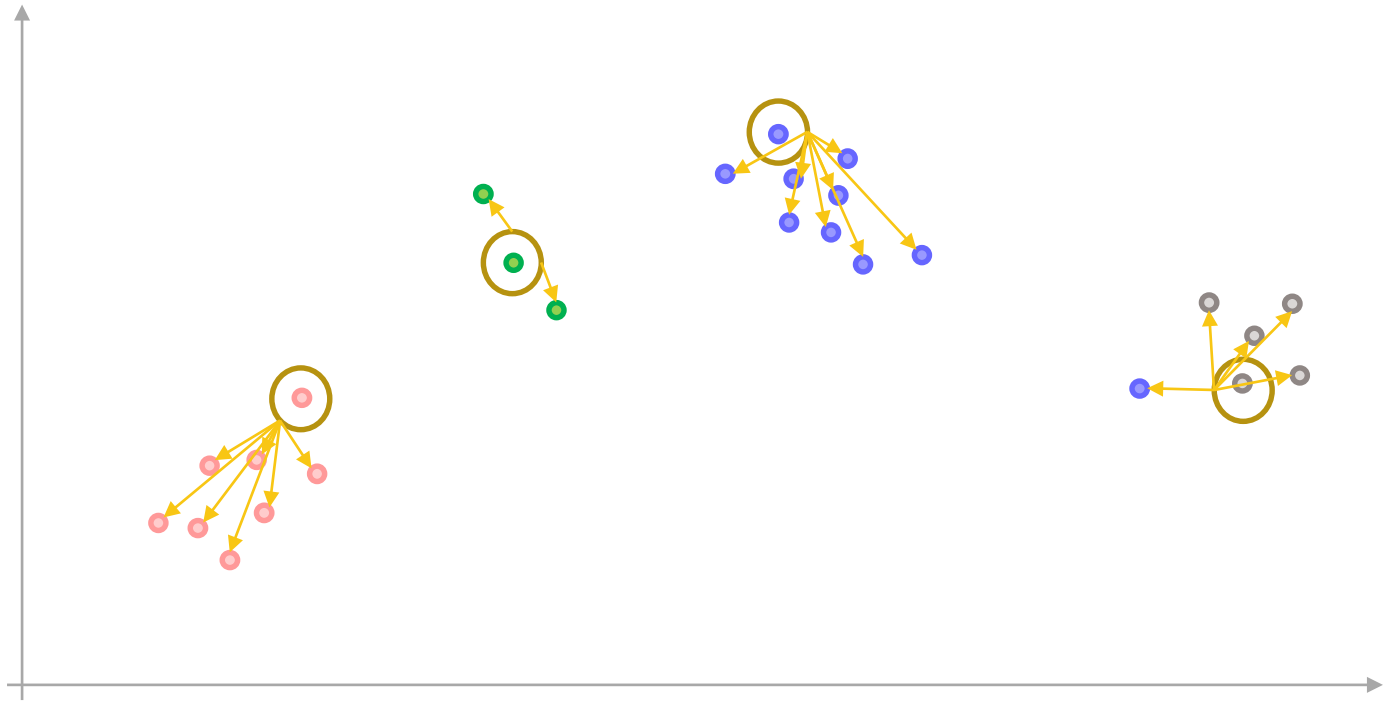
Adding Active Learning to the Cycle



© 2018 KNIME AG. All Rights Reserved.

Step 3 – Extend with k-NN

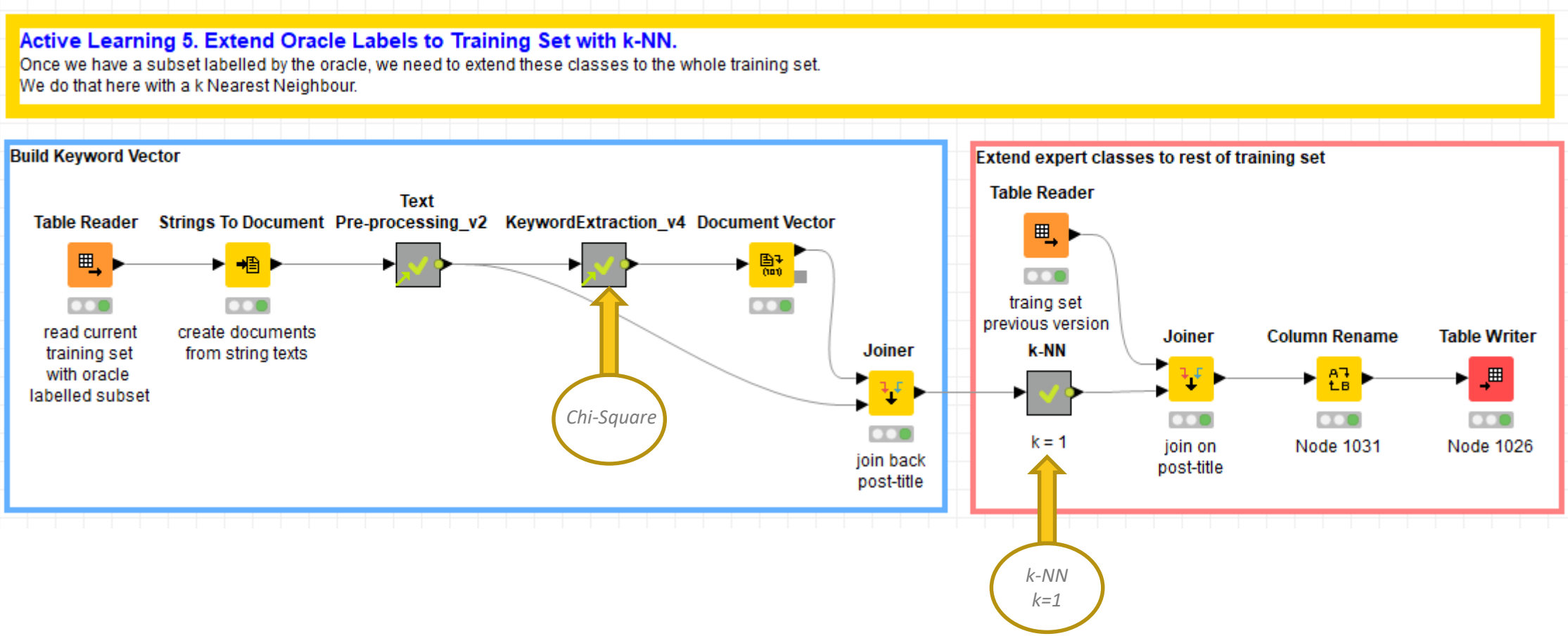
- Expert has labelled uncertain samples
- k-NN (k=1) **extends** the expert classes to their neighbor sample



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

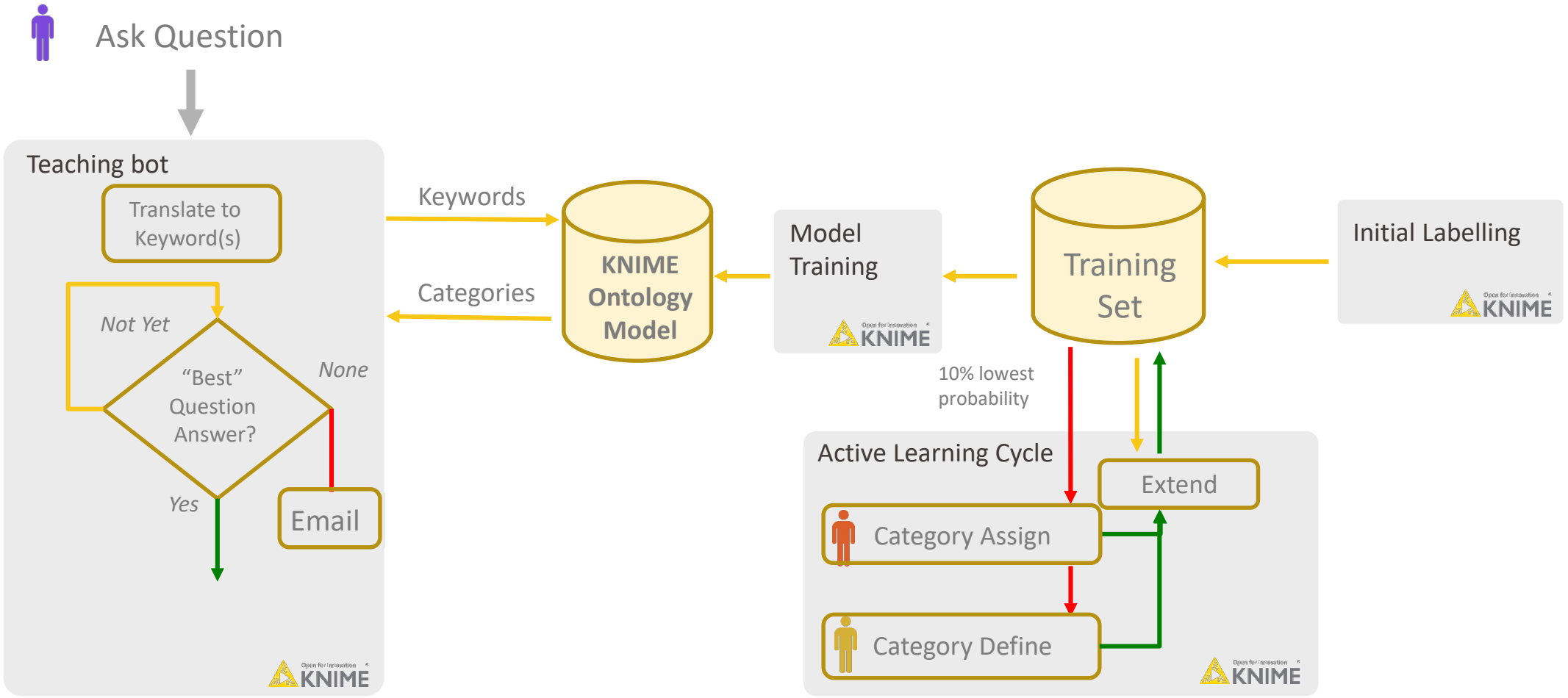
Step 3 - Extend with k-NN



© 2018 KNIME AG. All Rights Reserved.

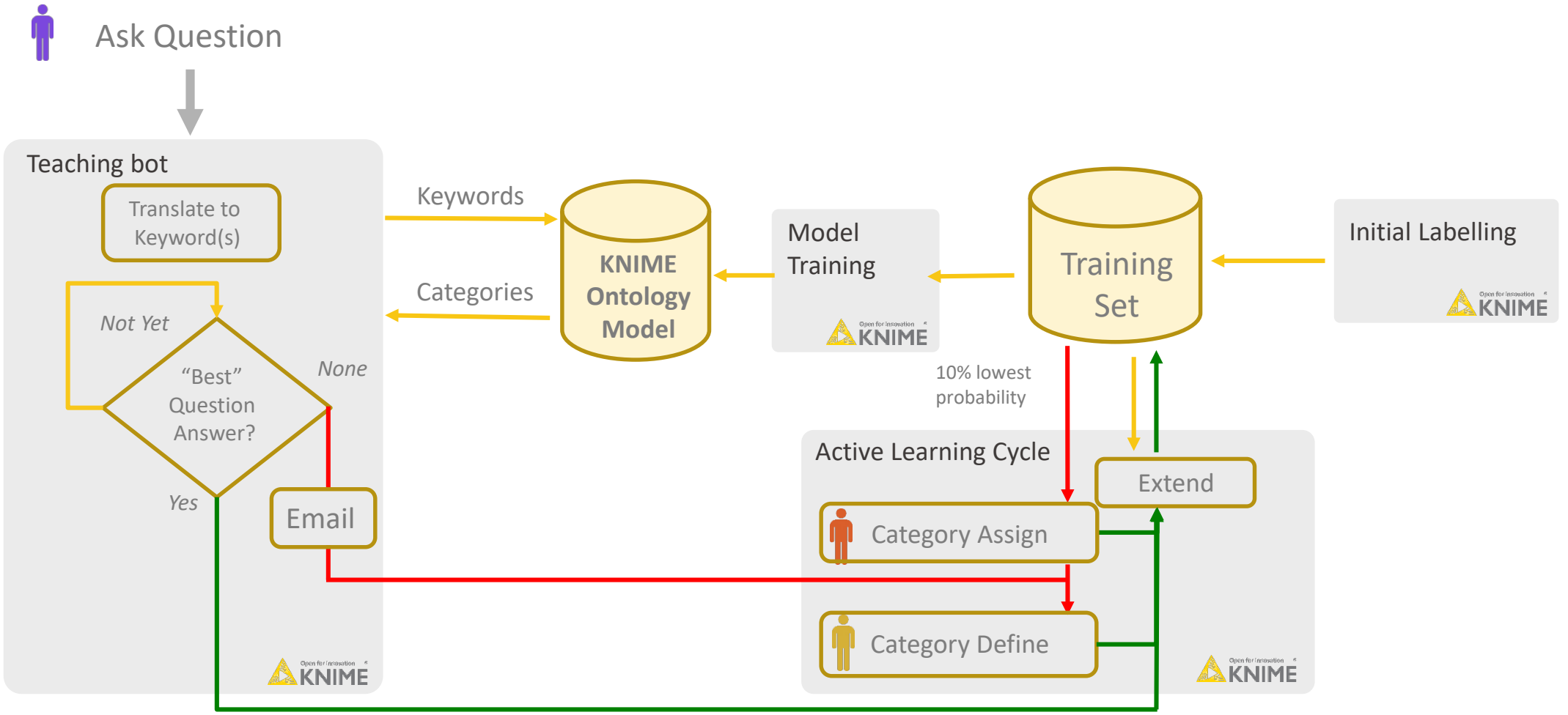
Open for Innovation KNIME

Adding Active Learning to the Cycle



© 2018 KNIME AG. All Rights Reserved.

Combining the Teaching Bot and the Active Learning Cycle

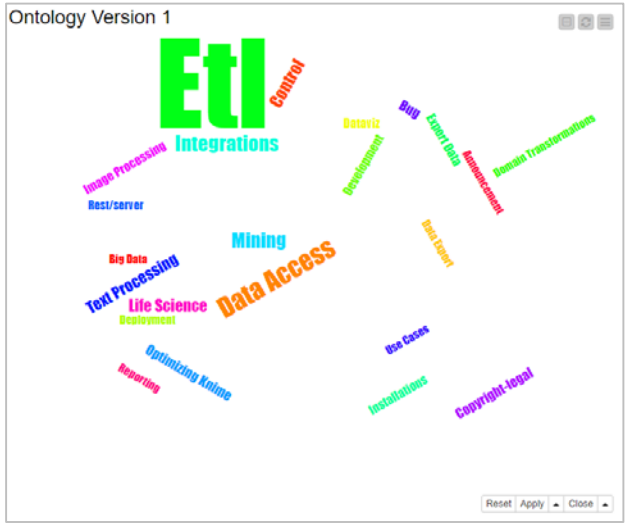


© 2018 KNIME AG. All Rights Reserved.

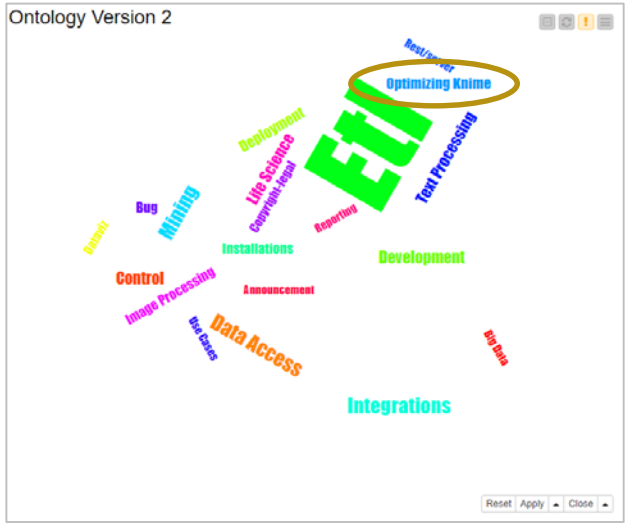
Changes in Training Set

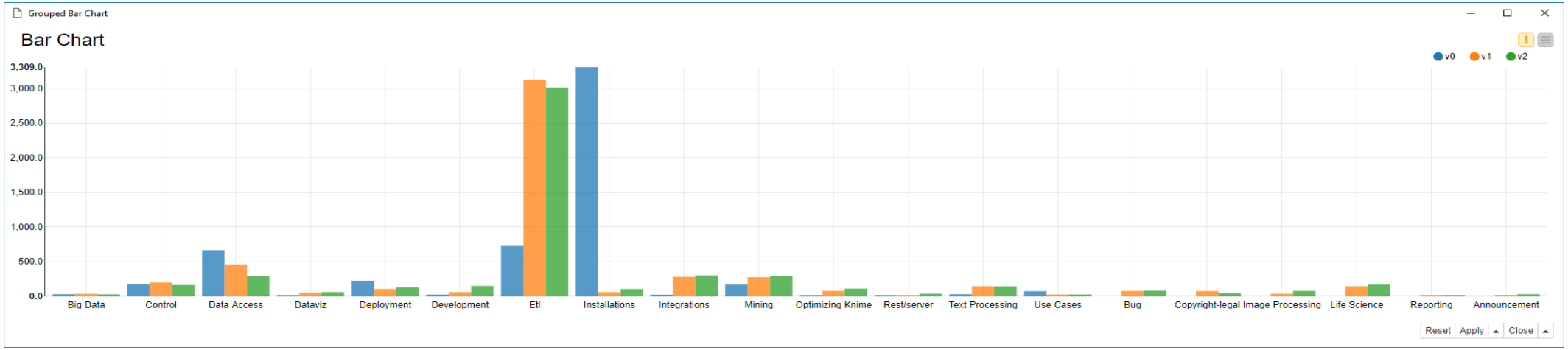
AL Iteration 0

AL Iteration 1



AL Iteration 2



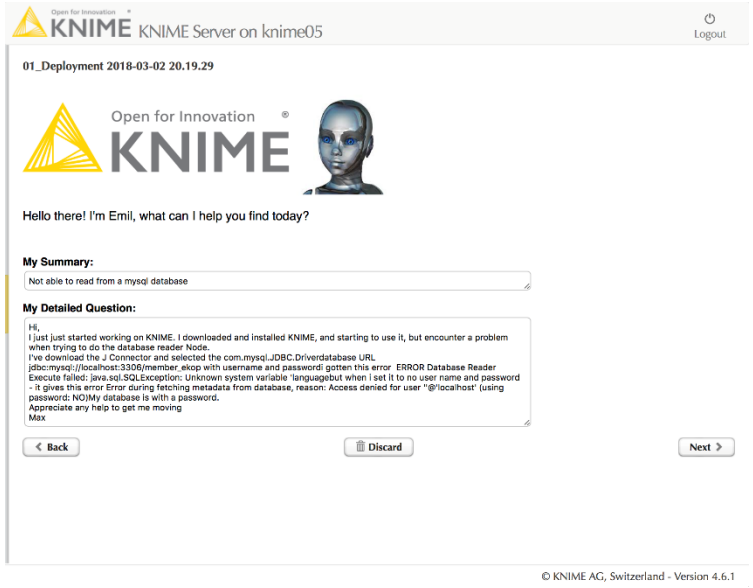


© 2018 KNIME AG. All Rights Reserved.

Open for Innovation
KNIME

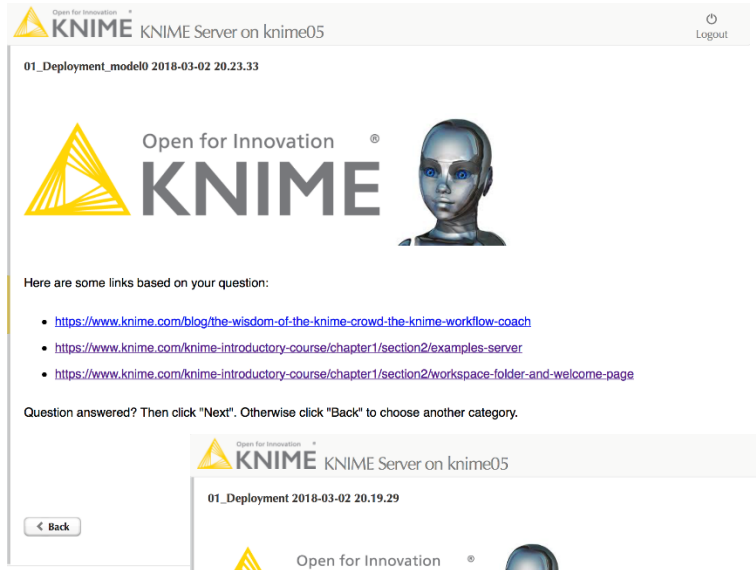
Answer Evolution

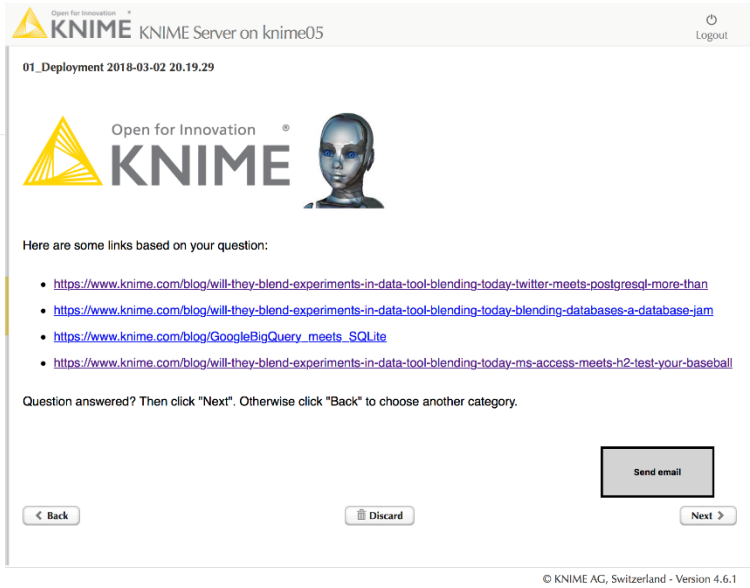
AL #	Input Dataset	Output Model	Ver	Accuracy	Timestamp
0	Traning_Set_v0	Random_Forest_v0	0.0	0.59	19/2/2018
1	Traning_Set_v1	Random_Forest_v1	1.0	**0.56**	23/2/2018
2	Traning_Set_v2	Random_Forest_v2	2.0	0.52	26/2/2018
	...	...	...	...	...



Version 0

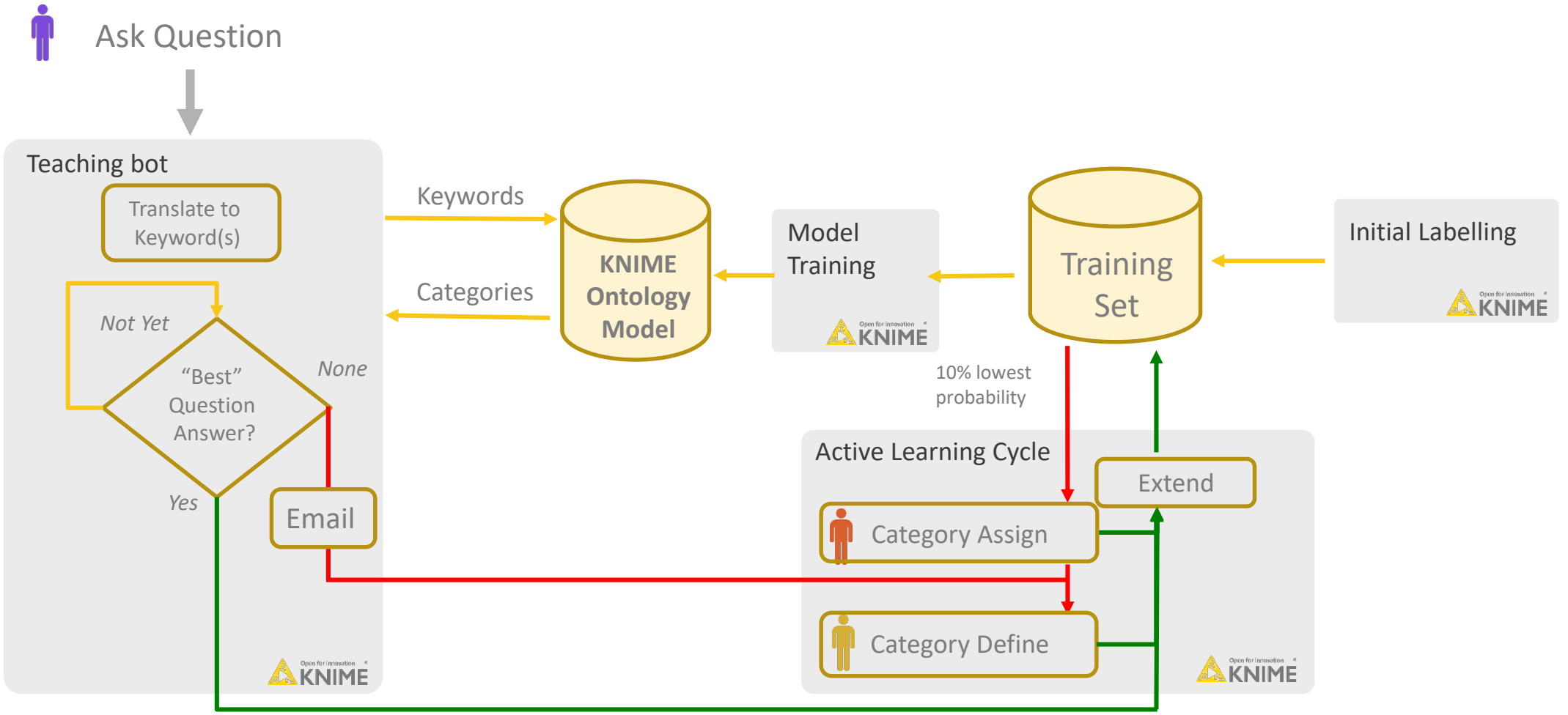
Version 2





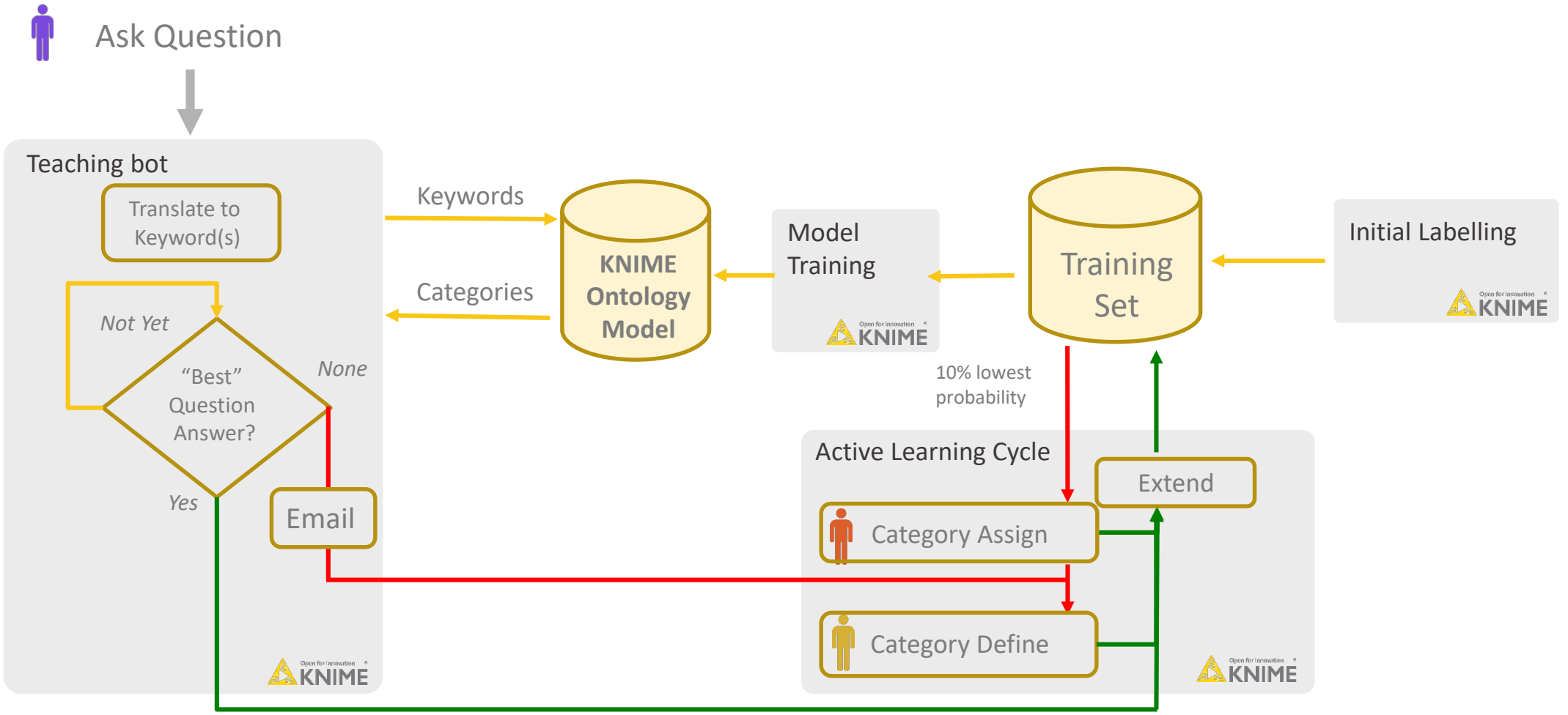
© 2018 KNIME AG. All Rights Reserved.

Combining the Teaching Bot and the Active Learning Cycle



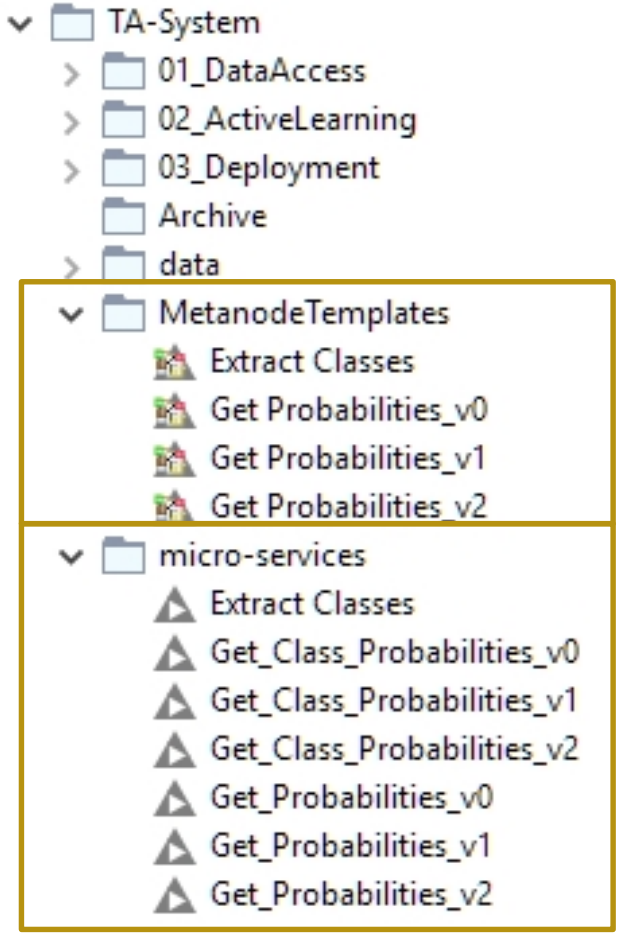
© 2018 KNIME AG. All Rights Reserved.

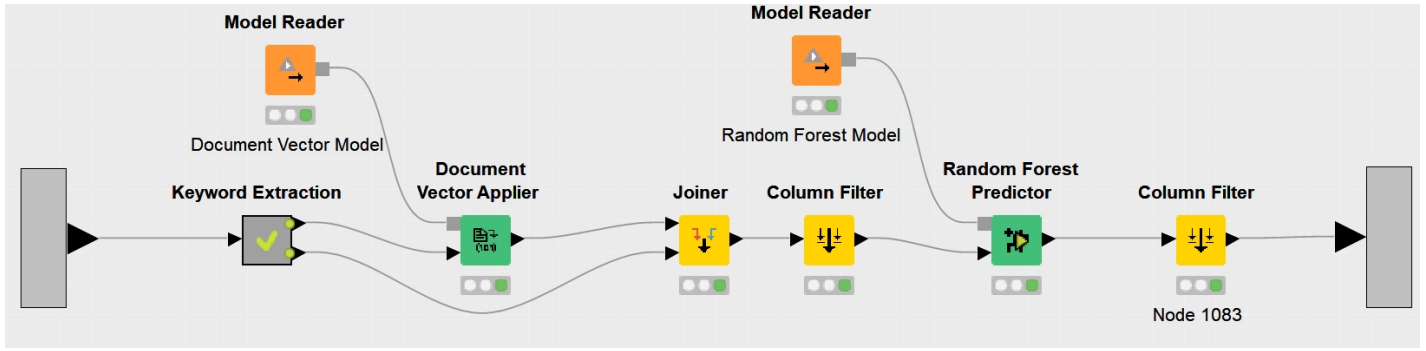
From building one time to reusing Components: MicroServices

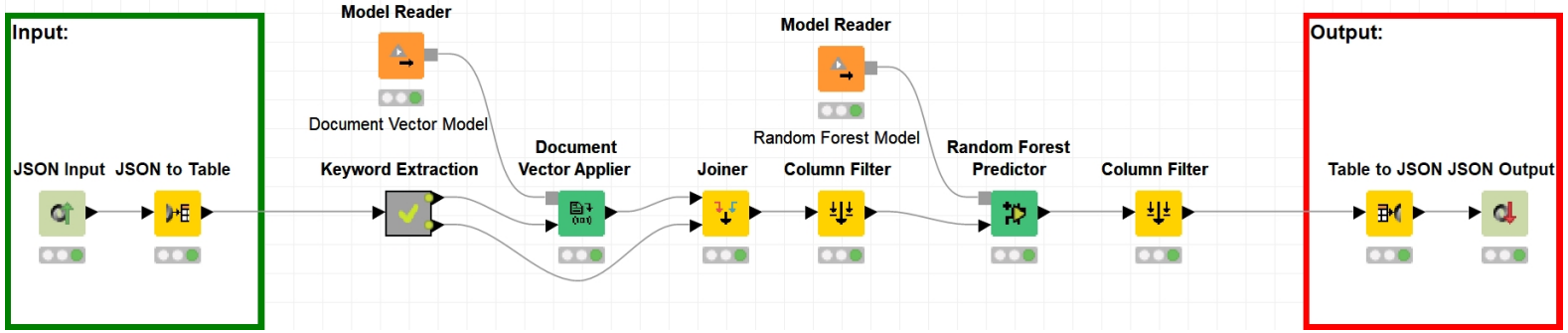


© 2018 KNIME AG. All Rights Reserved.

Microservices





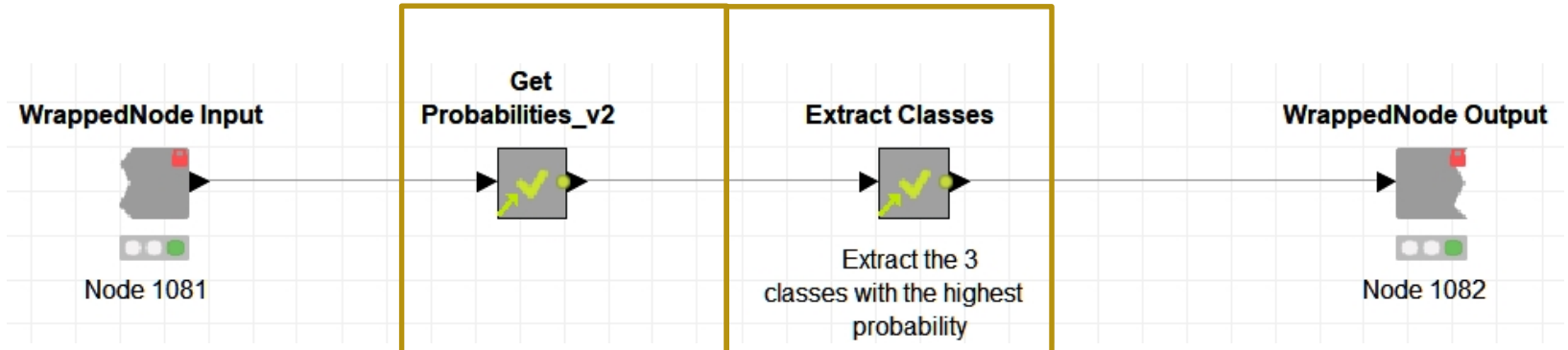


© 2018 KNIME AG. All Rights Reserved.

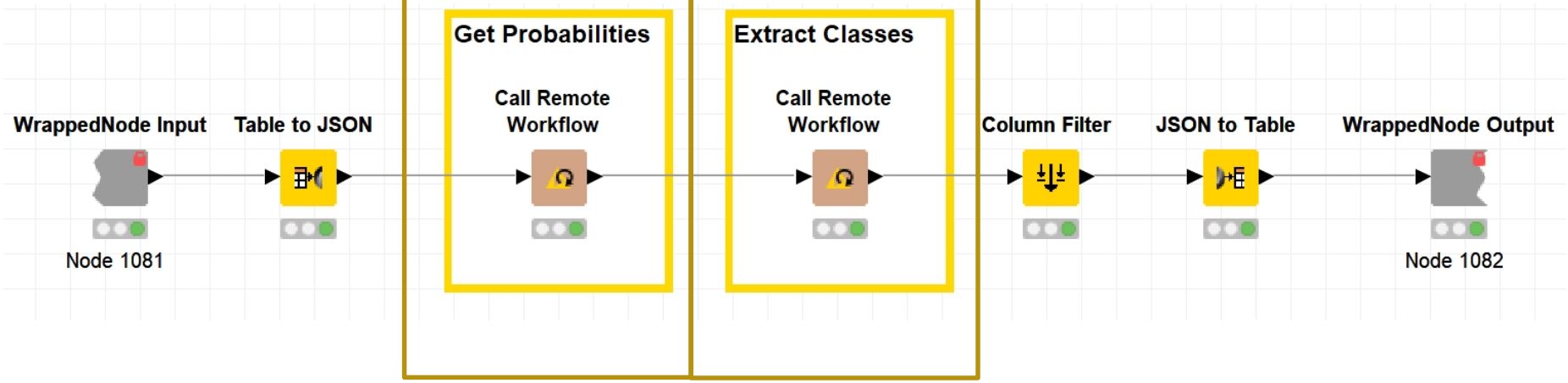
Open for Innovation KNIME

Microservices - Converting reusable Subflows into Microservices

Metanode Templates



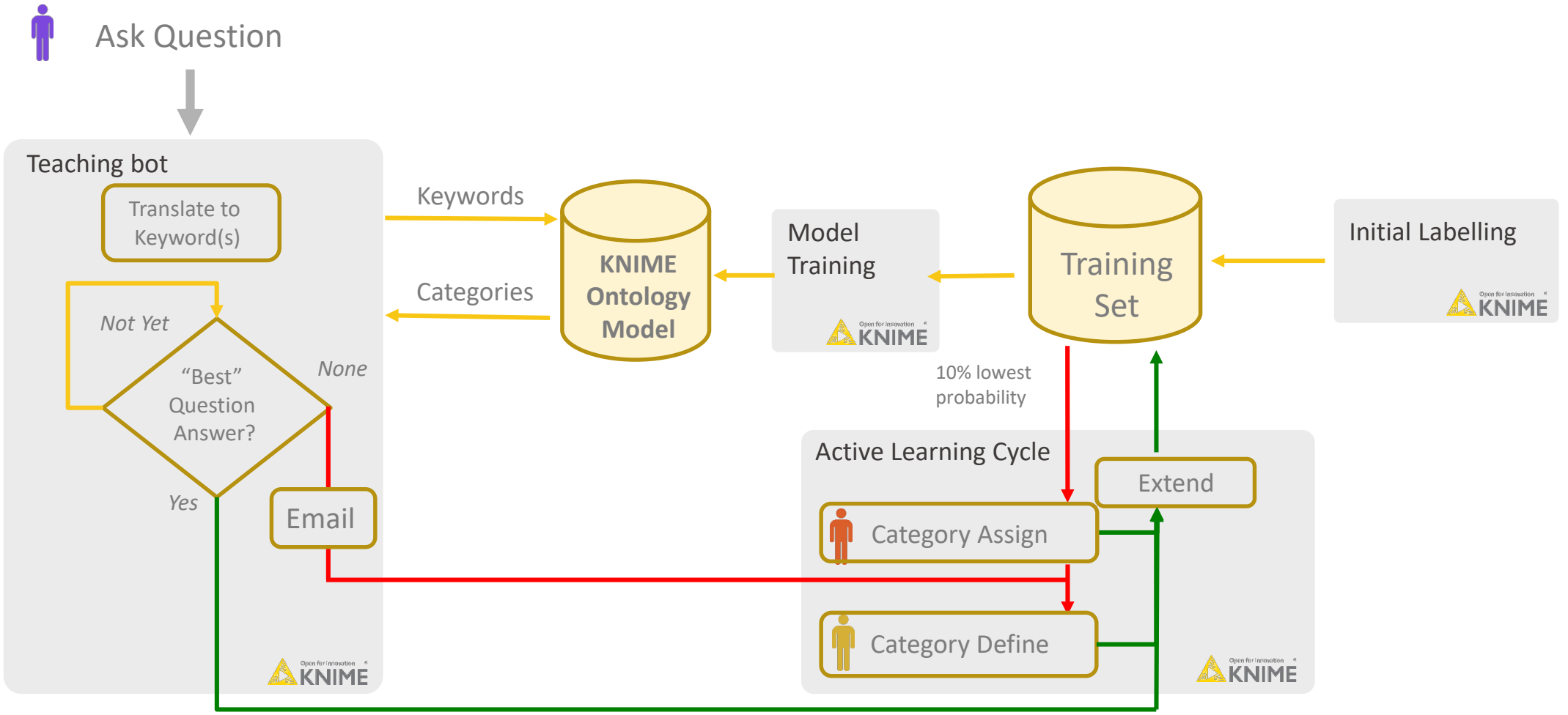
Microservices



© 2018 KNIME AG. All Rights Reserved.

Open for Innovation KNIME

Combining the Teaching Bot and the Active Learning Cycle



© 2018 KNIME AG. All Rights Reserved.

What we have tried to show….

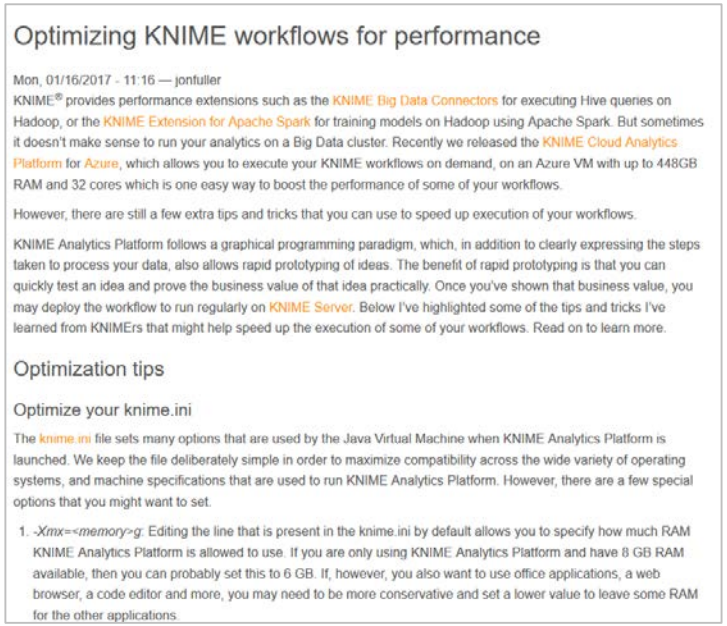
- Creating a basic bot
- Building an Ontology with Active Learning
- Automating the process
- Converting reusable subflows into micro services

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

What did we learn?

- KNIME forum is used as educational tool
- Support is search
- Keyword extraction is a plus with respect to just keyword search
- Re-adjust your class system (and goals) from time to time
- Accuracy is not all
- New educational page on DataViz
- Optimizing KNIME -> Maybe another blog post?



Optimizing KNIME workflows for performance

Mon, 01/16/2017 - 11:16 — jonfuller

KNIME® provides performance extensions such as the KNIME Big Data Connectors for executing Hive queries on Hadoop, or the KNIME Extension for Apache Spark for training models on Hadoop using Apache Spark. But sometimes it doesn't make sense to run your analytics on a Big Data cluster. Recently we released the KNIME Cloud Analytics Platform for Azure, which allows you to execute your KNIME workflows on demand, on an Azure VM with up to 448GB RAM and 32 cores which is one easy way to boost the performance of some of your workflows.

However, there are still a few extra tips and tricks that you can use to speed up execution of your workflows.

KNIME Analytics Platform follows a graphical programming paradigm, which, in addition to clearly expressing the steps taken to process your data, also allows rapid prototyping of ideas. The benefit of rapid prototyping is that you can quickly test an idea and prove the business value of that idea practically. Once you've shown that business value, you may deploy the workflow to run regularly on KNIME Server. Below I've highlighted some of the tips and tricks I've learned from KNIMErs that might help speed up the execution of some of your workflows. Read on to learn more.

Optimization tips

Optimize your knime.ini

The knime.ini file sets many options that are used by the Java Virtual Machine when KNIME Analytics Platform is launched. We keep the file deliberately simple in order to maximize compatibility across the wide variety of operating systems, and machine specifications that are used to run KNIME Analytics Platform. However, there are a few special options that you might want to set.

1. *-Xmx=<memory>g*: Editing the line that is present in the knime.ini by default allows you to specify how much RAM KNIME Analytics Platform is allowed to use. If you are only using KNIME Analytics Platform and have 8 GB RAM available, then you can probably set this to 6 GB. If, however, you also want to use office applications, a web browser, a code editor and more, you may need to be more conservative and set a lower value to leave some RAM for the other applications.

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

How could this be extended?

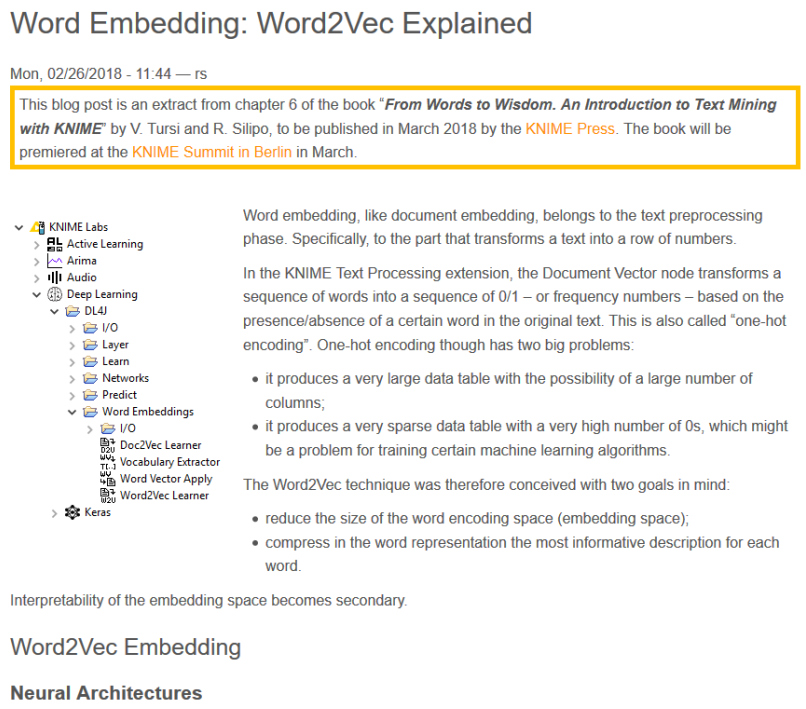
- Improve text processing phase (tagging)
- Use word embedding
  - Problem: Document Vector leads to big and sparse feature spaces
  - Solution: Train a vector representation for each word using the Word2Vec
- Use the Keras integration to replace the Random Forest with a Neural Network which uses LSTM layers.
- Investigate the role of parameters:
  - 10% of uncertain
  - K=1 in k-Nearest Neighbors
  - Forgetting functions?
- Add speech recognition? KNIME YouTube videos as additional resource

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

Where to find more

- Presentation available immediately
- Series of blog posts in the next weeks
- Workflows on EXAMPLE Server
- Collection of blog posts in a whitepaper



Word Embedding: Word2Vec Explained

Mon, 02/26/2018 - 11:44 — rs

This blog post is an extract from chapter 6 of the book "***From Words to Wisdom. An Introduction to Text Mining with KNIME***" by V. Tursi and R. Silipo, to be published in March 2018 by the KNIME Press. The book will be premiered at the KNIME Summit in Berlin in March.

Word embedding, like document embedding, belongs to the text preprocessing phase. Specifically, to the part that transforms a text into a row of numbers.

In the KNIME Text Processing extension, the Document Vector node transforms a sequence of words into a sequence of 0/1 – or frequency numbers – based on the presence/absence of a certain word in the original text. This is also called "one-hot encoding". One-hot encoding though has two big problems:

- it produces a very large data table with the possibility of a large number of columns;
- it produces a very sparse data table with a very high number of 0s, which might be a problem for training certain machine learning algorithms.

The Word2Vec technique was therefore conceived with two goals in mind:

- reduce the size of the word encoding space (embedding space);
- compress in the word representation the most informative description for each word.

Interpretability of the embedding space becomes secondary.

Word2Vec Embedding

Neural Architectures

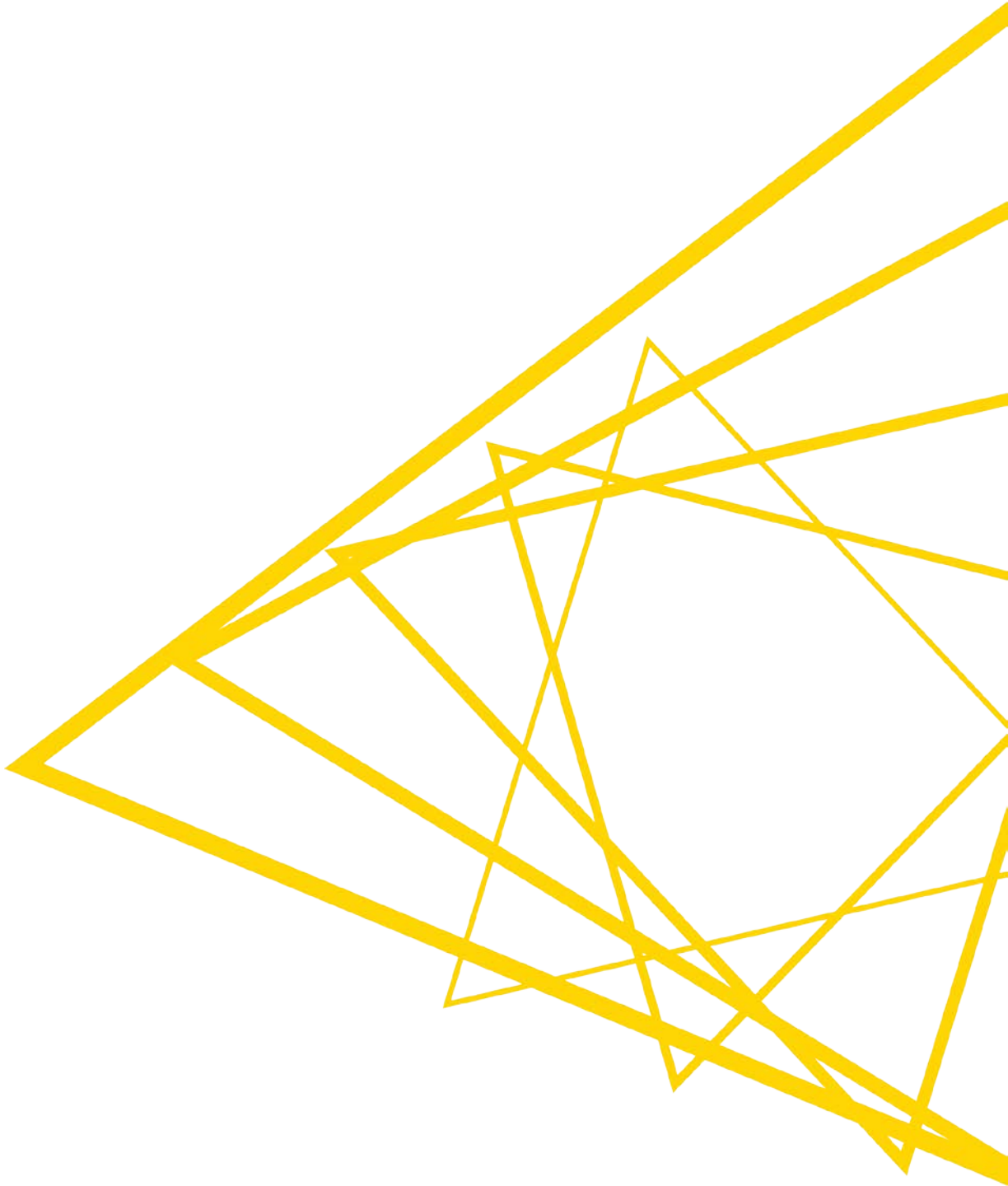
© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ® KNIME

Open for Innovation
KNIME

KNIME & Teacher Bots: From Workflows to Micro Services

Kathrin Melcher
Vincenzo Tursi
Rosaria Silipo
Phil Winters



© 2018 KNIME AG. All Right Reserved.

The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.

© 2018 KNIME AG. All Rights Reserved.

Open for Innovation ®
KNIME