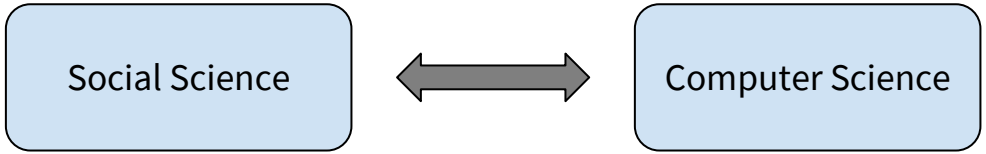
Social Media Mining in KNIME: Democratising Access to Libraries for Text Analysis (DELTA)

Riza Batista-Navarro, Thomas Flavel and Conor Gaughan

DIGISURVOR Workshop
14th January 2026

Interdisciplinary Research



Examples of computational tasks:

- Pre-processing and cleaning of data
- Extraction of variables through analysis of social media/user-generated content: text but also other modalities
- Harnessing cutting-edge natural language processing (NLP) models, i.e., LLMs

Technical Barriers

1. **Tool silos**

   Codebases are in individual repositories

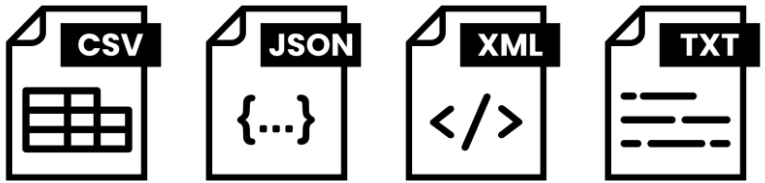
   No one-stop catalogue of analytics tools

Technical Barriers

2. **Lack of interoperability between tools**

   Written in different programming languages

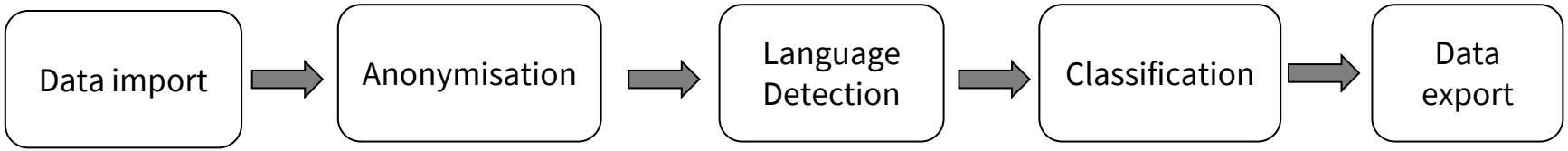
   Follow different input and output formats



File format icons from Freepik (LAB Design Studio)

Technical Barriers

Example: **a simple hate speech detection pipeline (workflow)**



Multiple analytics tools are required

Each tool specialises on a task

Technical Barriers

3. **Required programming skills**

   Calling APIs

   Writing bespoke code

4. **Disadvantages of cloud-based platforms**

   Cost of analysing huge datasets on the cloud

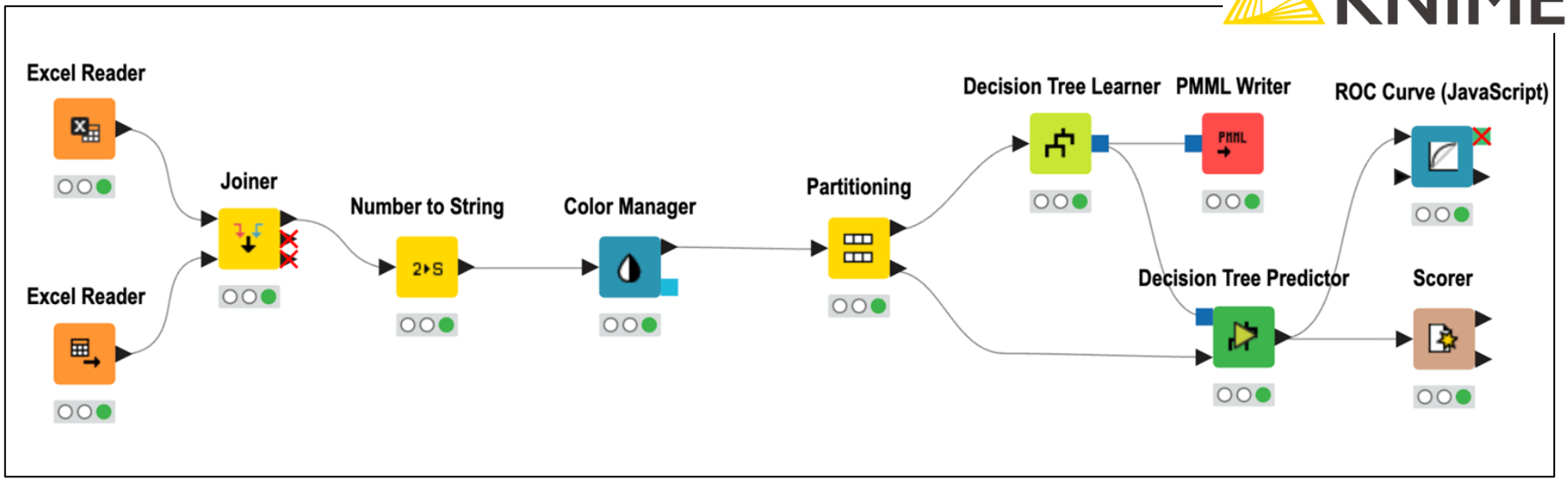
   Data privacy risks of transferring potentially sensitive data to external servers

Existing Solutions for Social Media Analytics

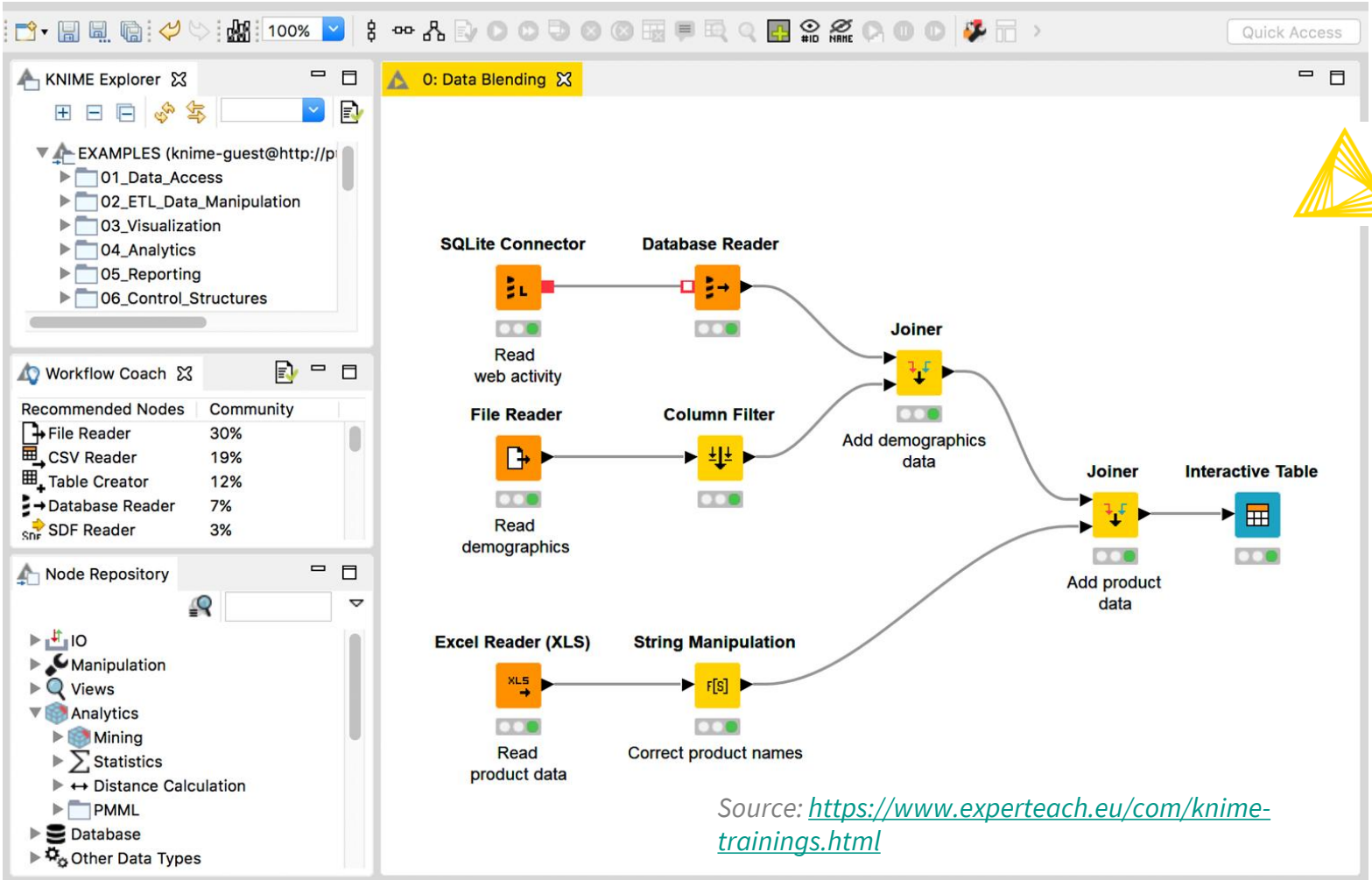
	Description	Link	Current Gap(s)
HuggingFace	Repository of machine learning-based models	https://huggingface.co/	Requires writing of Python code to build workflows
spaCy	Framework for developing NLP workflows	https://spacy.io/	Requires writing of Python code to build workflows
GATE	Workflow development platform offering cloud-based analytics	https://cloud.gate.ac.uk	Incurs costs; requires transferring data to the cloud
Communalytic	Cloud-based social media workflows	https://communalytic.org	Incurs costs; requires transferring data to the cloud; workflows are fixed
KNIME	Generic data analytics platform	http://knime.com	Limited support for social media processing/analytics

Our Proposed Solution: Building upon KNIME

KNIME: Open-source workflow development platform for building and running data processing pipelines based on visual programming



Source: https://www.knime.com/why-visual-workflows



Source: https://www.experteach.eu/com/knime-trainings.html

Open for Innovation
KNIME

- ✓ Uses a **drag-and-drop, no-code** interface
- ✓ Serves as a **repository/central collection** of tools, i.e., "nodes"
- ✓ Promotes **interoperability**
- ✓ **Free**
- ✓ **Locally installed** software: runs locally

- ✗ Limited nodes for text analysis (majority of current nodes are for general domain and biomedical data)
- ✗ Lacks support for analysing social media content
- ✗ Requires high-spec local machines/servers

DELTA:
Democratising Access to Libraries for Text Analysis

Goal:

To develop new KNIME nodes by wrapping publicly available NLP tools for social media analytics supporting a variety of tasks

To facilitate the no-code creation of end-to-end social media analysis workflows

Data gathering: API access, scraping, data donation

Pre-processing: de-identification/anonymisation, normalisation, language detection

Analysis: topic identification, sentiment analysis, detection of emotions, stance, irony, hate speech, ideological leaning, etc (variables) -- **for various languages**

Evaluation: Metrics

Wish List of Nodes?

Data Gathering

Tool/Library Name	Platform	Method	Programming language	Link to code	Licence	Languages?
snscrape	Facebook, Instagram, Mastodon, Reddit, Telegram, Twitter, VKontakte, Sina Weibo	Scraping	Python	https://github.com/JustAnotherArchivist/snscrape	GPL-3.0	N/A
tweepy	X	API access	Python	https://github.com/tweepy/tweepy	MIT	N/A
praw	Reddit	API access	Python	https://github.com/praw-dev/praw	BSD-2-Clause	N/A
Google API Client	YouTube	API access	Python	https://github.com/google	Apache-2.0	N/A
Telethon	Telegram	API access	Python	https://github.com/LonamiWebs/Telethon	MIT	N/A
Unofficial TikTok API	TikTok	Scraping	Python	https://github.com/davidteather/TikTok-Api	MIT	N/A
Port	Any	Data donation	Python	https://github.com/d3i-infra/data-donation-task	AGPL-3.0	N/A
Feldspar	Any	Data donation	Python	https://github.com/eyra/feldspar	AGPL-3.0	N/A

Wish List of Nodes?

Data Pre-processing

Tool/Library Name	Input	Output	Programming language	Link to code	Licence	Languages?
TextWash	Personally identifying text	Deidentified text	Python	https://github.com/ben-aaron188/textwash	GPL-3.0	English, Dutch
PyTesseract	Image containing text	Text	Python	https://github.com/h/pytesseract	GPL-3.0	> 100 languages
Preprocessor	Tweets containing hashtags, urls, emojis, etc	Tweets where hashtags etc are normalised/cleansed	Python	https://github.com/s/preprocessor	GPL-3.0	Agnostic
LinguaPy	Unstructured text	Detected language(s)	Python	https://github.com/pemistahl/lingua-py	Apache 2.0	75 languages
NLTK	Unstructured text	Various NLP tasks	Python	https://github.com/nltk/nltk	Apache-2.0	Many; varies by function

Wish List of Nodes?

Analysis/Variable Generation

Tool/Library Name	Variable measures	Variable draws on	Programming language	Link to code		Languages?
Twitter SES	Socioeconomic status	Twitter accounts followed (brands)	R	https://github.com/yuanmohe/Twitter_SES		Agnostic but US-focused
TwitterGenderPredictor	Gender	Microblog text	Python	https://github.com/jtwool/TwitterGenderPredictor	MPL-2.0	English
M3-Inference	Gender, age, organization-or-person	Twitter profiles	Python	https://github.com/euagendas/m3inference	AGPL-3.0	32 languages
Gensim	Topics	Unstructured text	Python	https://github.com/piskvorky/gensim	LGPL-2.1 license	Agnostic
BERTopic	Topics	Unstructured text	Python	https://github.com/MaartenGr/BERTopic	MIT	Agnostic
deepIdeology	Ideological leaning	Microblog text	R	https://github.com/alex-gottlieb/deepIdeology		English
PoliBERTweet	Stance	Microblog text	Python	https://github.com/GU-DataLab/PoliBERTweet	MIT	English
VADER	Sentiment	Unstructured text	Python	https://github.com/cjhutto/vaderSentiment	MIT	English

Wish List of Nodes?

Evaluation

Tool/Library Name	Metric measures	Metric draws on	Programming language	Link to code		Languages?
MoverScore	Semantic similarity of sentence pairs	Pair of sentences	Python	https://github.com/AIPHES/emnlp19-moverscore	MIT	Multilingual
SemScore	Semantic similarity of sentence pairs	Pair of sentences	Python	https://github.com/geronimi73/semscore	MIT	English
TextDescriptives	Text quality, readability, dependency_distance, pos_proportions, coherence, descriptive_stats	Original and/or generated text	Python	https://github.com/HLasse/TextDescriptives	Apache-2.0	Multilingual
Evaluate	Various metrics used to evaluate a range of NLP models	Model outputs and gold standard labels	Python	https://github.com/huggingface/evaluate	Apache-2.0	Agnostic

Community

KNIME has a wide user-base (hundreds of thousands)

Nodes and workflows are being shared via the **KNIME Community Hub**

Active user forums

Additionally:

- Outreach to other social science audiences
- Domain-specific examples
- Tailored documentation
- Support for languages beyond English

Call for Contributions to the Wish List

Third-party tools you've come across/used?

Your own in-house tools/code that you want to share with the wider community?

Let us know!

Thank you!

Any questions or suggestions?

Get in touch:

riza.batista@manchester.ac.uk

thomas.flavel@manchester.ac.uk

conor.gaughan@manchester.ac.uk