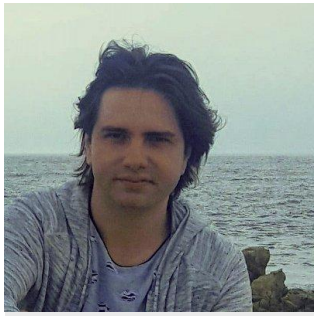
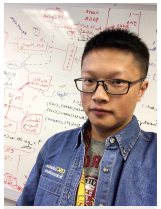
KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis

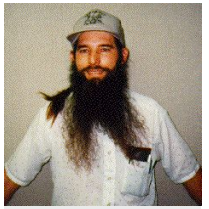

FILIP ILIEVSKI


Daniel Garijo


Hans Chalupsky


Naren Teja Divvala


Yixiang Yao


Craig Rogers


Rongpeng Li


Jun Liu

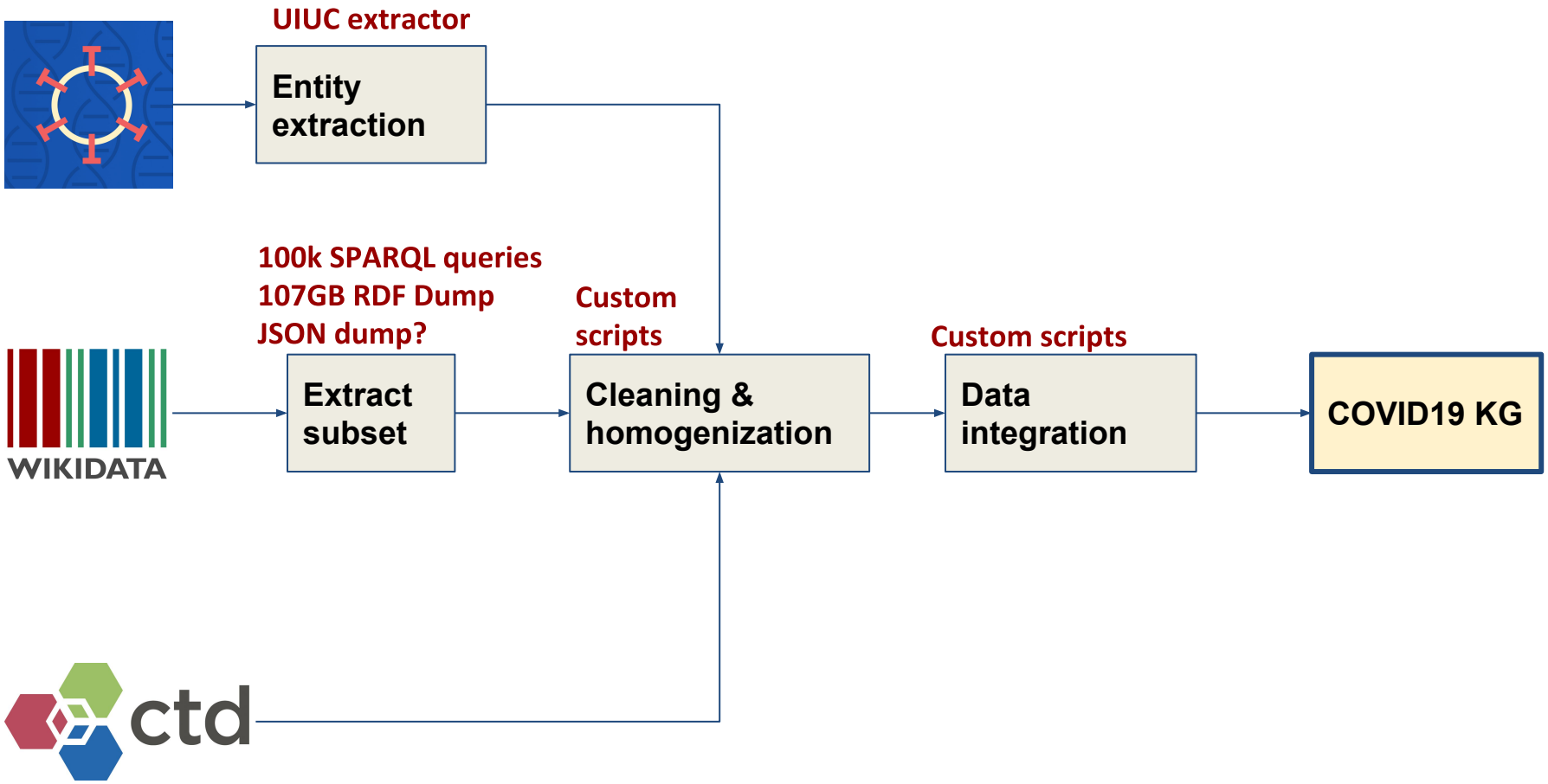

Amandeep Singh


Daniel Schwabe

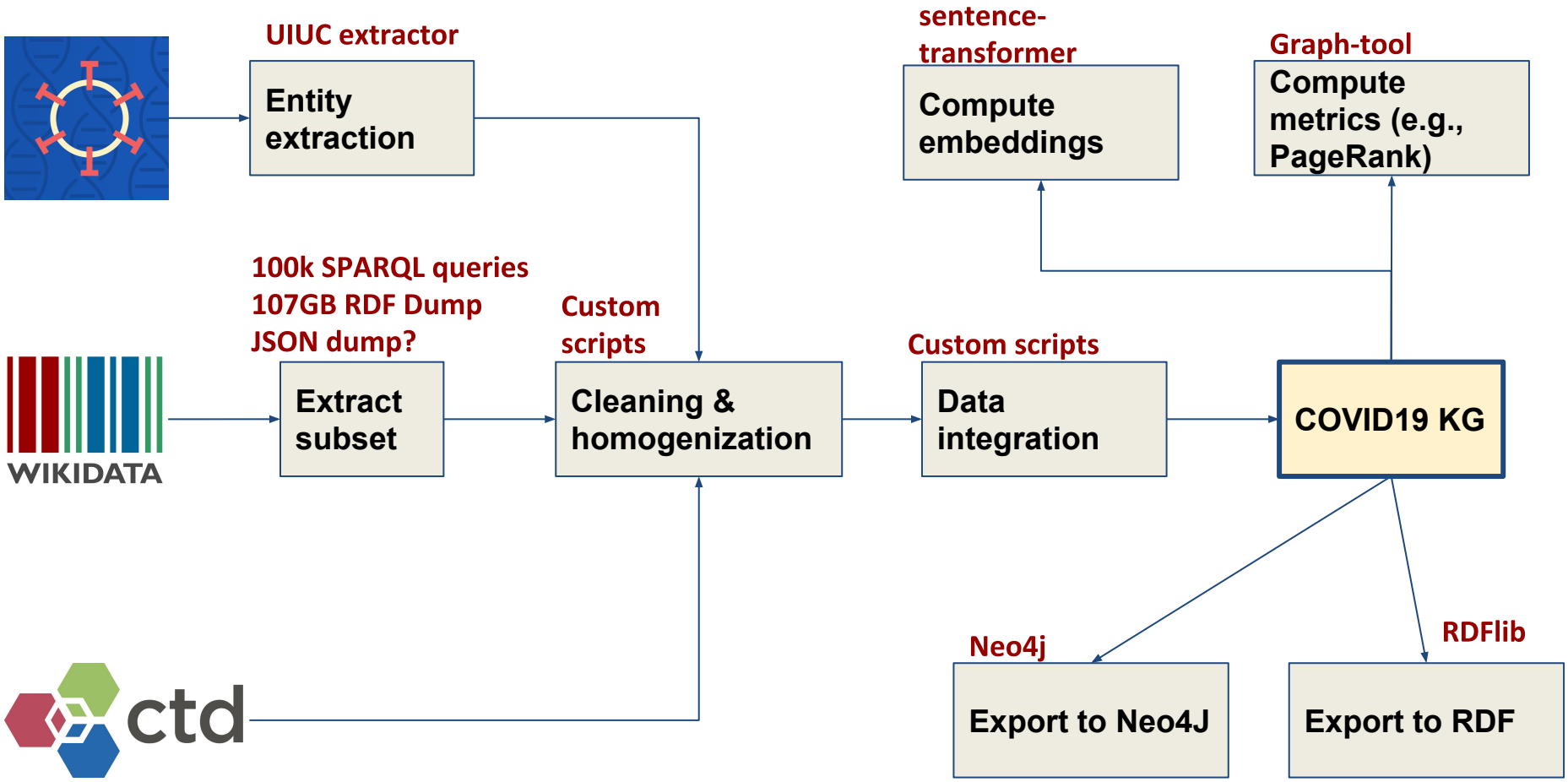

Pedro Szekely

Work supported by the Air Force Research Laboratory under agreement number FA8750-20-2-10002

Integrating KGs at scale can be challenging



Using KGs at scale can be challenging



KG pipelines need many tools

Operation	Tool	Format(s)	Language
graph analytics	graph-tool	GML, GT, TSV/CSV	python
	NetworkX	GML, JSON, TSV/CSV	python
graph database	Neo4J	TSV/CSV	*various*
RDF operations	graphy	RDF	javascript
	Jena	RDF	java
graph embeddings	PyTorch-BigGraph	TSV/CSV	python
entity resolution	RLTK	TSV/CSV	python
entity linking	AGDISTIS	XML	python
	WAT	JSON	python

Requirements for a KG Toolkit

1. **Many capabilities**
   - **R1: simple representation format**
   - **R2: provide the best tool for each job**

Requirements for a KG Toolkit

1. **Many capabilities**
   - **R1: simple representation format**
   - **R2: provide the best tool for each job**

2. **Size**
   - **R3: run Wikidata (billion triples) on an average laptop**

Requirements for a KG Toolkit

1. **Many capabilities**
   - **R1: simple representation format**
   - **R2: provide the best tool for each job**
2. **Size**
   - **R3: run Wikidata (billion triples) on an average laptop**
3. **Ease of use**
   - **R4: common API to all tools**
   - **R5: appeal to all AI practitioners**

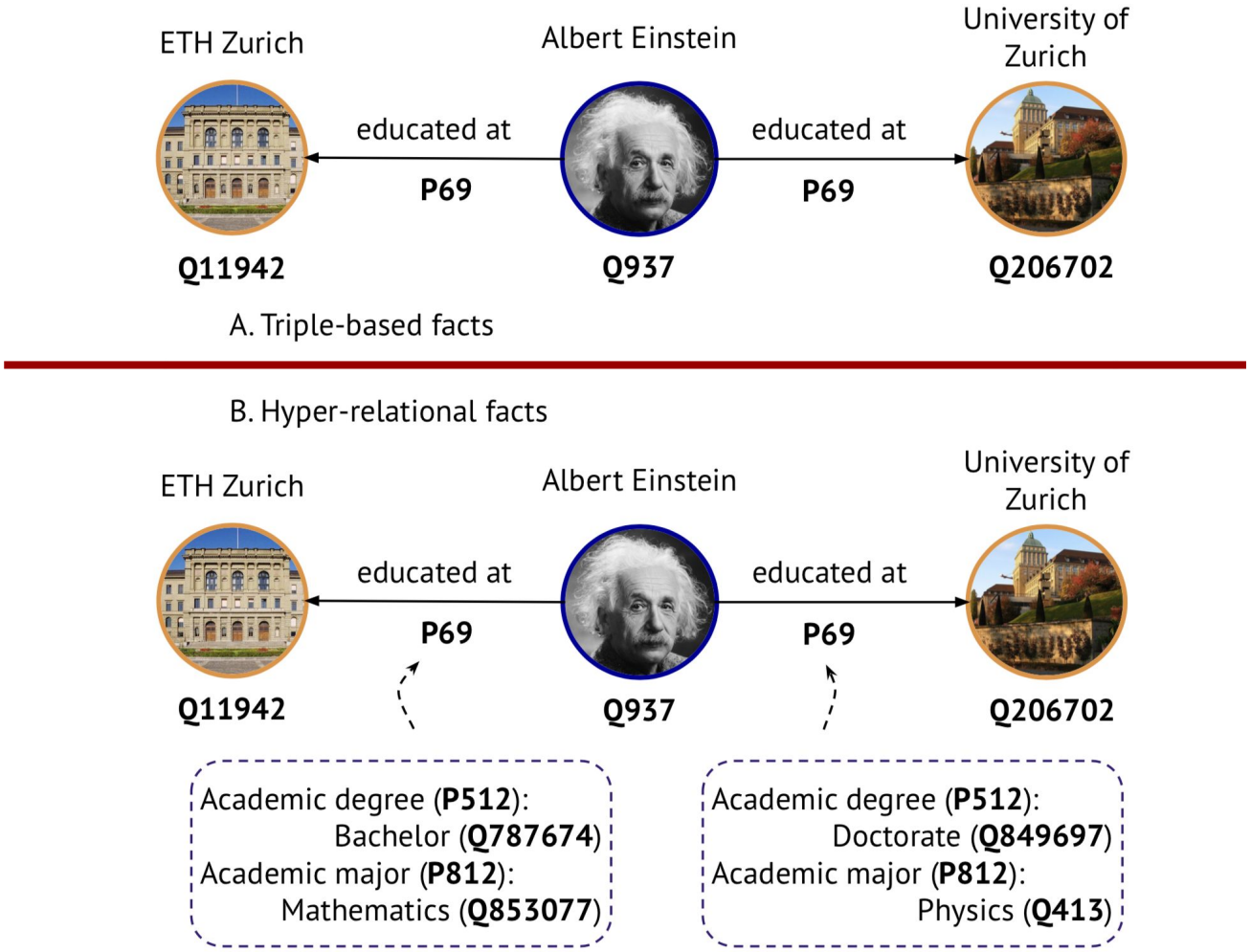
The Knowledge Graph ToolKit

Data format

Operations

Closing remarks

Hyper-relational data format



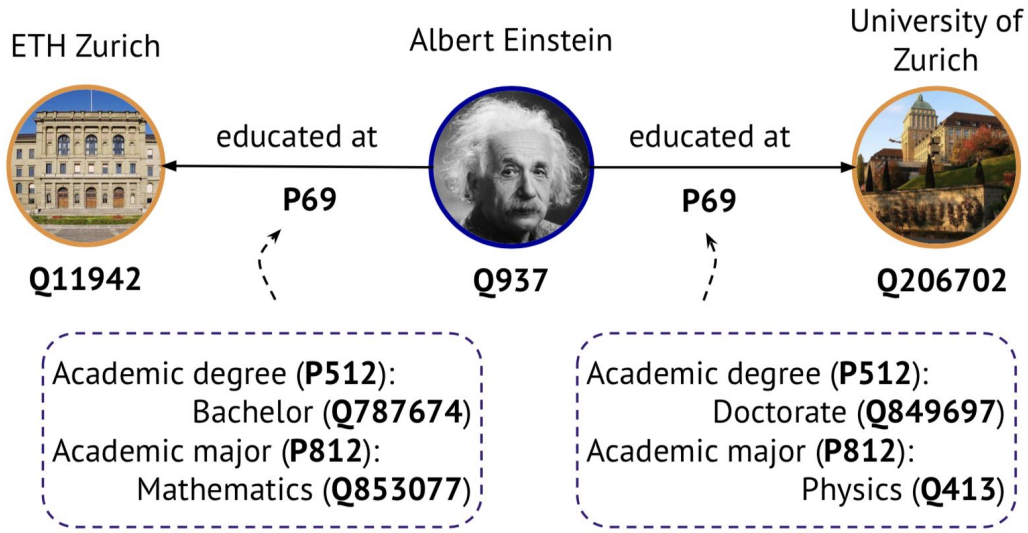
R1: simple representation format

Image by: Michael Galkin et al. (2020)

KGTK hyper-relational data format

node1	property	node2
Q937	*P69*	*Q11942*
Q937	*P69*	*Q206702*

header



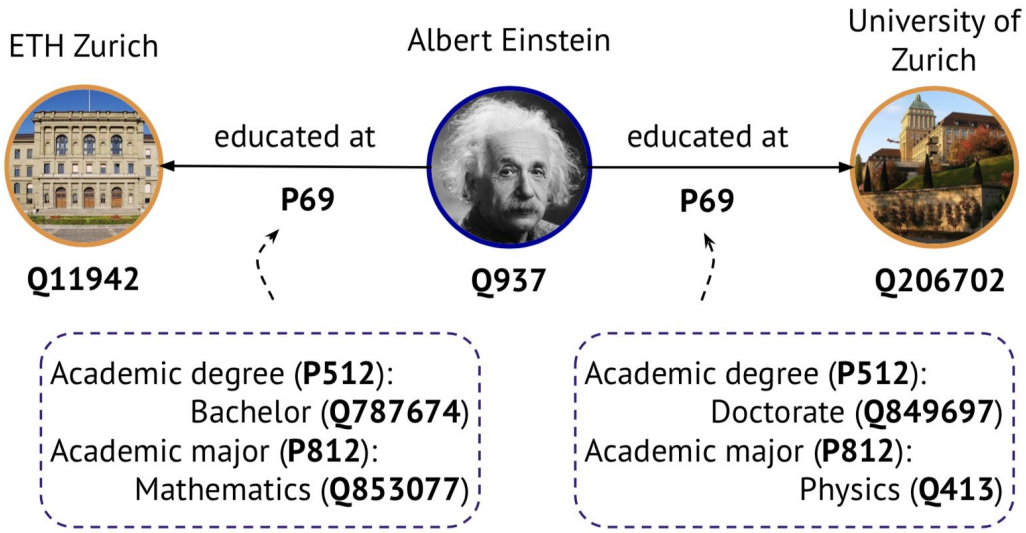
R1: simple representation format

Image by: Michael Galkin et al. (2020)

KGTK hyper-relational data format with qualifiers

node1	property	node2	P512	P812	id
Q937	*P69*	*Q11942*	*Q787674*	*Q853077*	*E1*
Q937	*P69*	*Q206702*	*Q849697*	*Q413*	*E2*

header



R1: simple representation format

Image by: Michael Galkin et al. (2020)

Representing edge metadata

node1	property	node2	P512	P812	id
Q937	*P69*	*Q11942*	*Q787674*	*Q853077*	*E1*

=

node1	property	node2	id
Q937	*P69*	*Q11942*	*E1*
E1	*P512*	*Q787674*	*E4*
E1	*P812*	*Q853077*	*E5*

R1: simple representation format

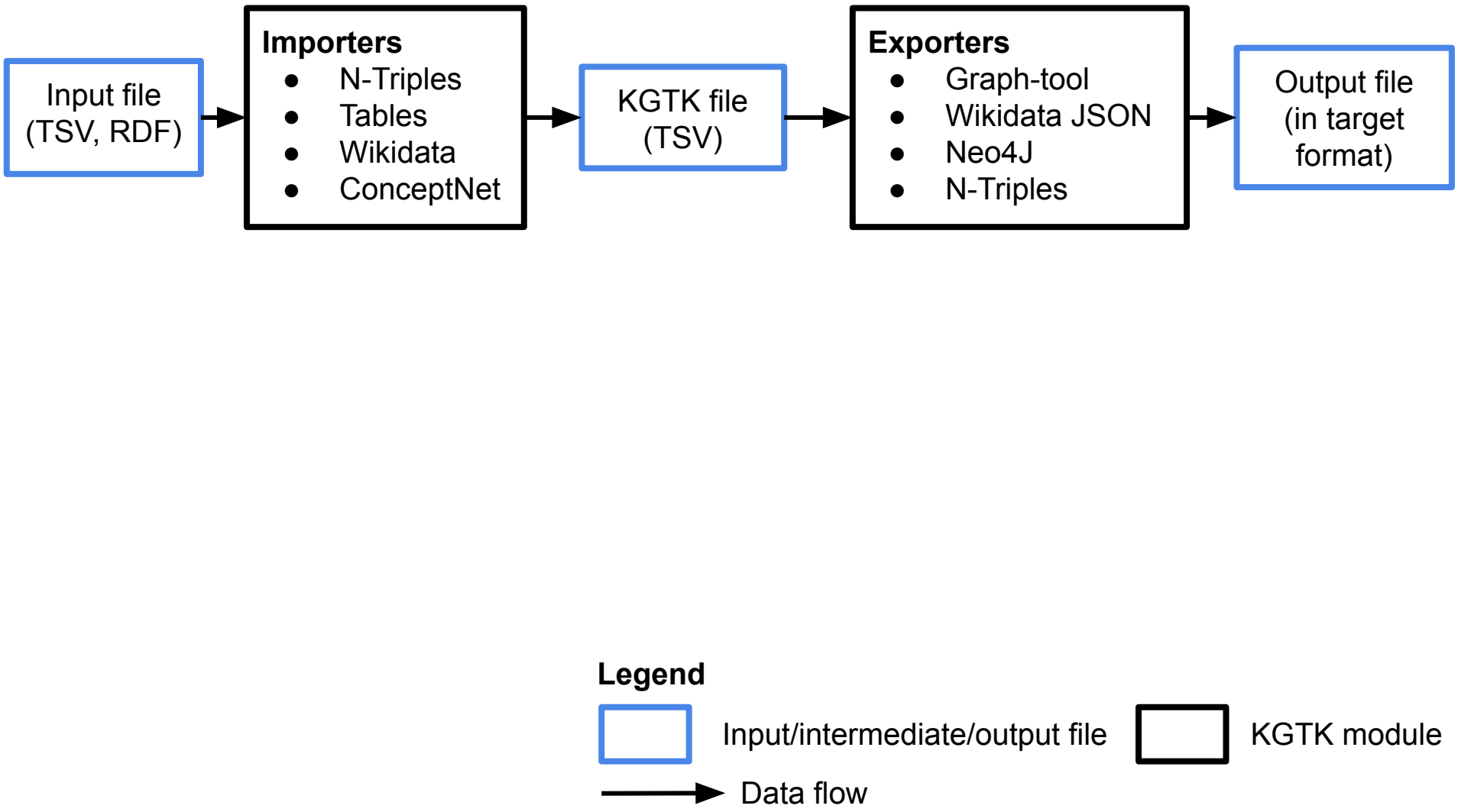
The Knowledge Graph ToolKit

Data format

Operations

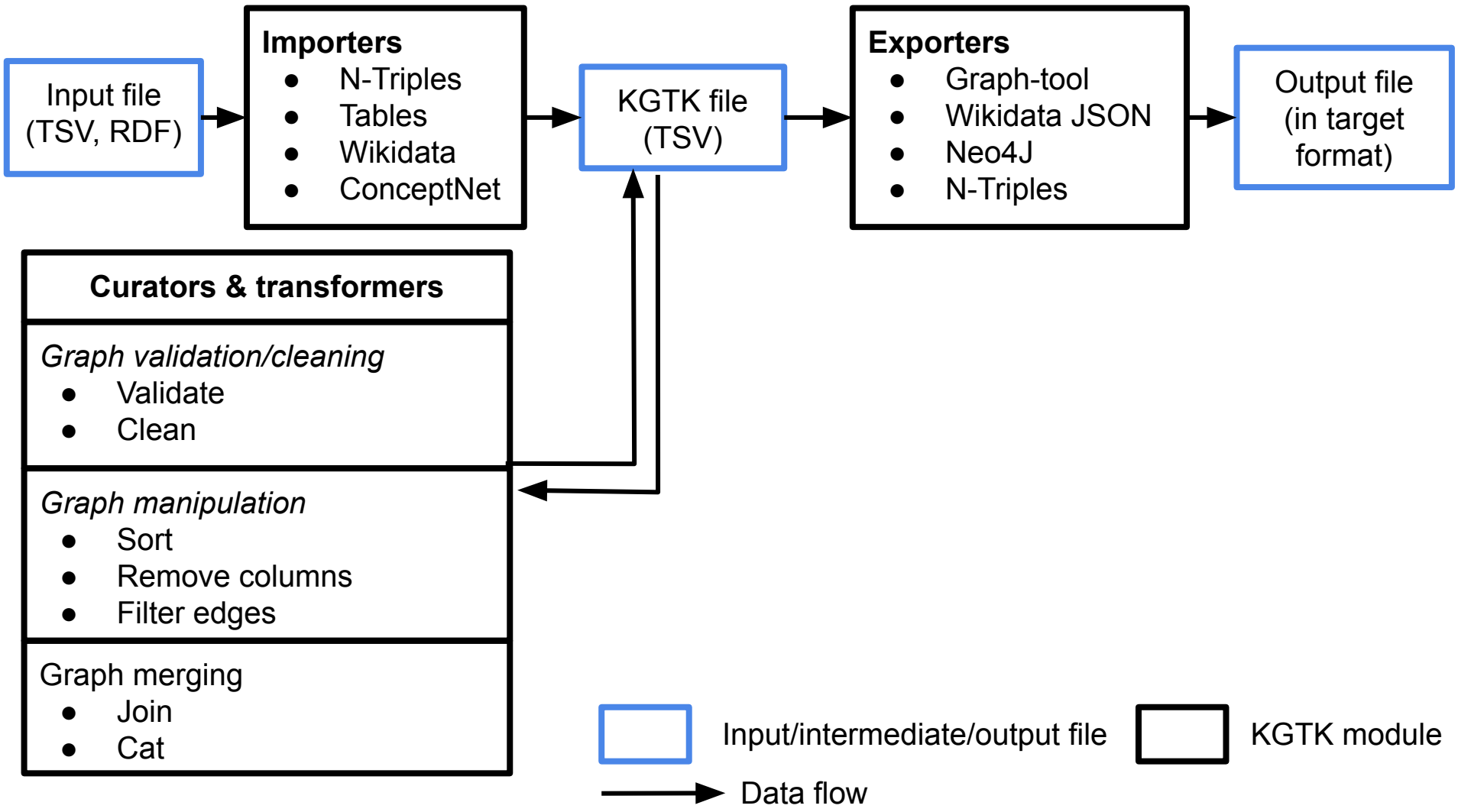
Closing remarks

KGTK supports many formats



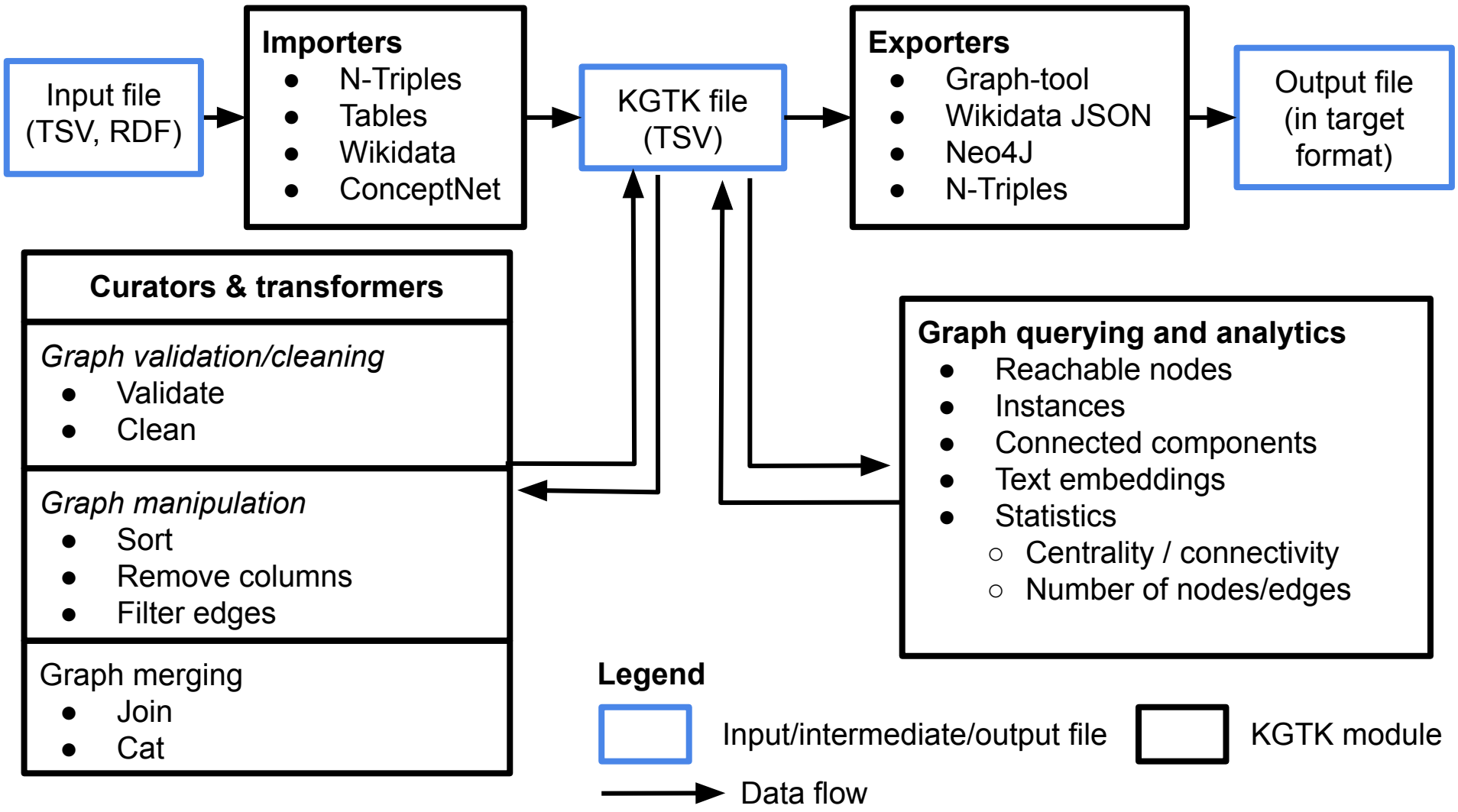
R1: simple representation format

Extensive ETL operations



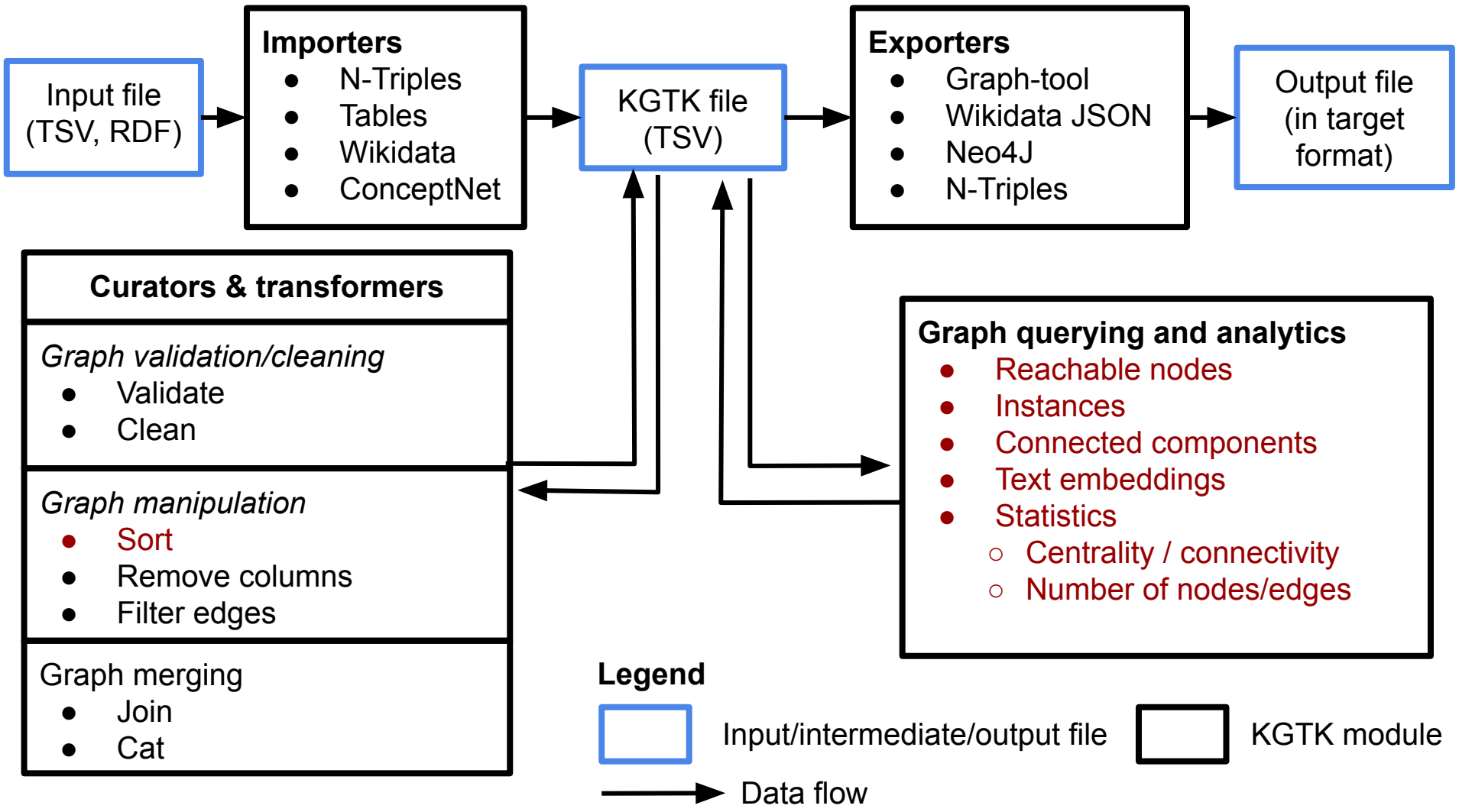
R4: common API to all tools

KGTK advanced analytics



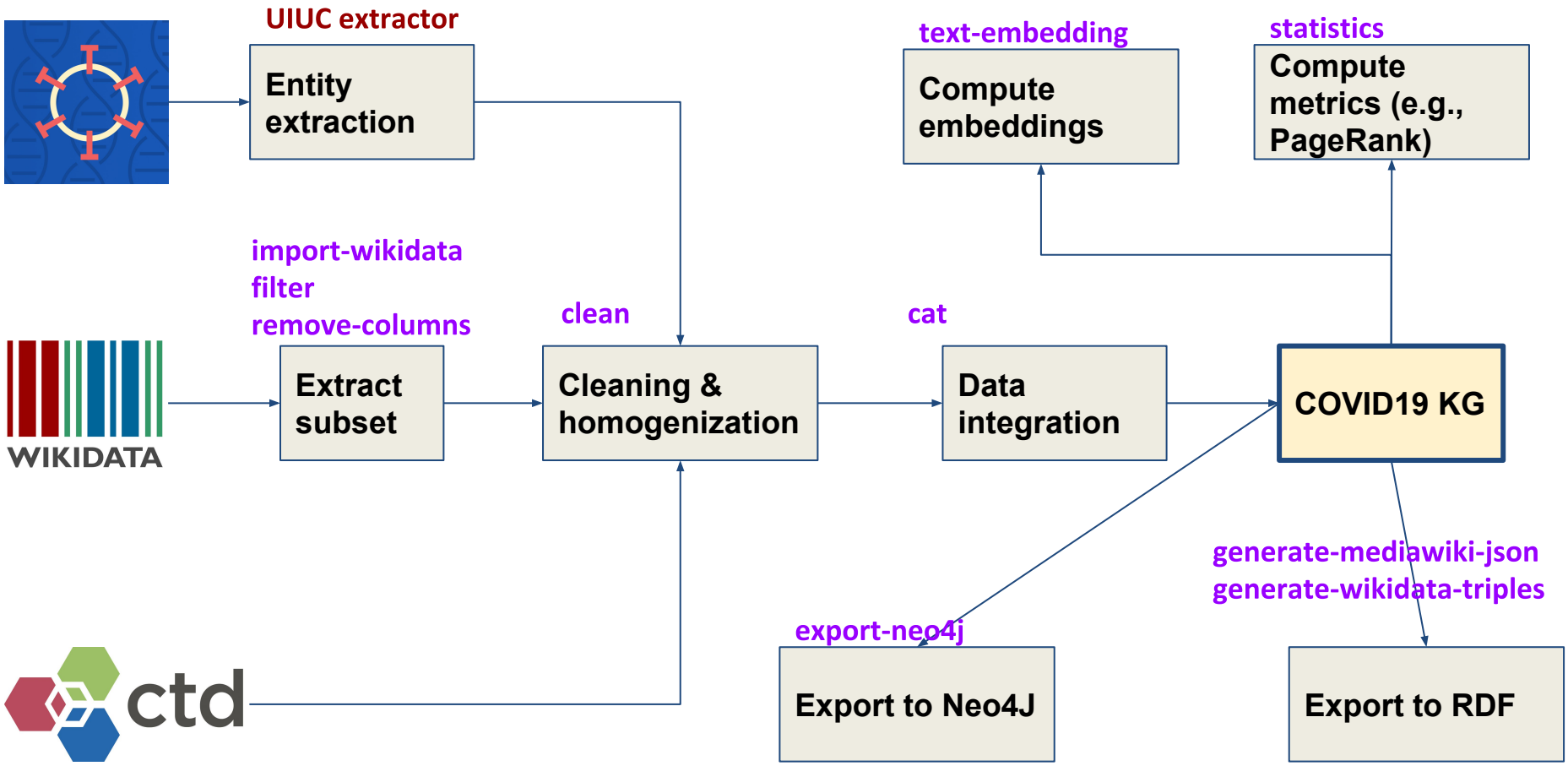
R4: common API to all tools

The best tools for each job



R2: provide the best tool for each job

COVID19 KGTK



KGTK pipelines: ‘member of’ statistics

kgtk import-wikidata ...

1. Import wikidata into KGTK

R3: run Wikidata (billion triples) on an average laptop

KGTK pipelines: ‘member of’ statistics

```
kgtk import-wikidata …/
filter -p ' ; P463 ; '
```

1. Import wikidata into KGTK
2. Select all P463 edges

R3: run Wikidata (billion triples) on an average laptop

KGTK pipelines: ‘member of’ statistics

```
kgtk import-wikidata …/
filter -p ' ; P463 ; ' /
clean
```

1. Import wikidata into KGTK
2. Select all P463 edges
3. Curate the data

R3: run Wikidata (billion triples) on an average laptop

KGTK pipelines: ‘member of’ statistics

```
kgtk import-wikidata … /
filter -p ' ; P463 ; ' /
clean /
remove-columns -c "$ignore_cols" /
```

1. Import wikidata into KGTK
2. Select all P463 edges
3. Curate the data
4. Ignore certain columns

R3: run Wikidata (billion triples) on an average laptop

KGTK pipelines: ‘member of’ statistics

```
kgtk import-wikidata … /
filter -p ' ; P463 ; ' /
clean /
remove-columns -c "$ignore_cols" /
graph-statistics --directed --degrees
--pagerank --degrees -o statistics.tsv
```

1. Import wikidata into KGTK
2. Select all P463 edges
3. Curate the data
4. Ignore certain columns
5. Compute PageRank and degrees

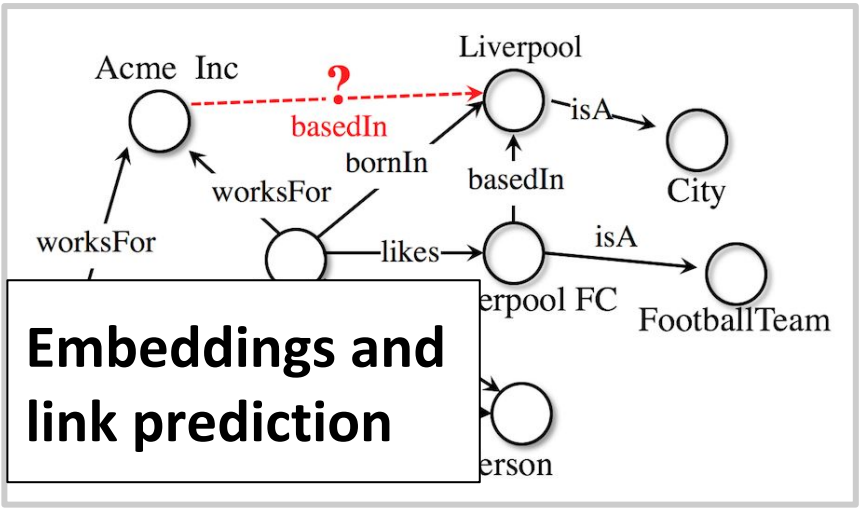
R3: run Wikidata (billion triples) on an average laptop

The Knowledge Graph ToolKit

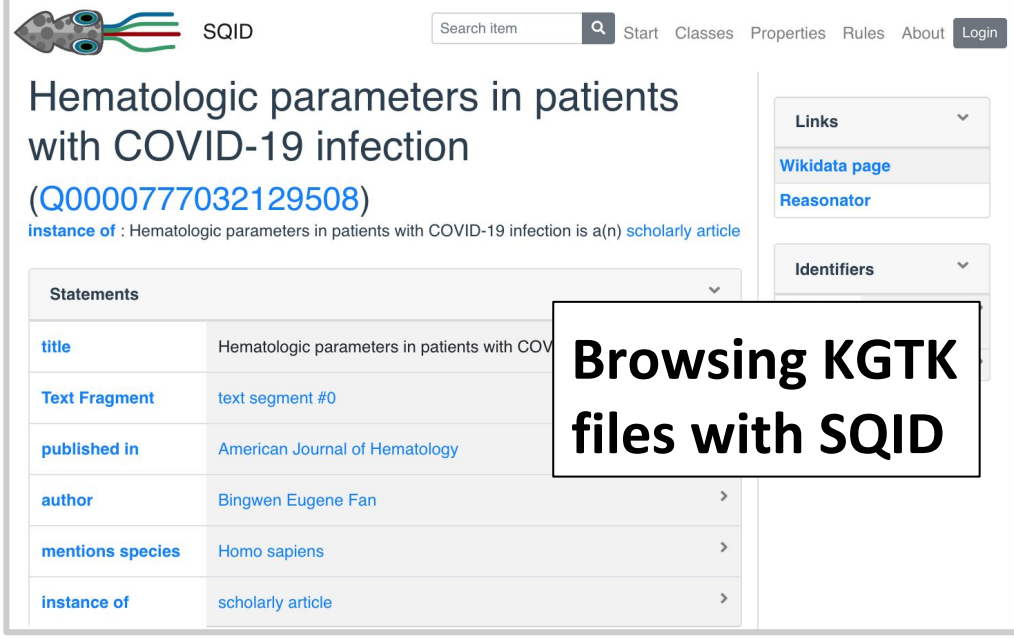
Data format

Operations

Closing remarks

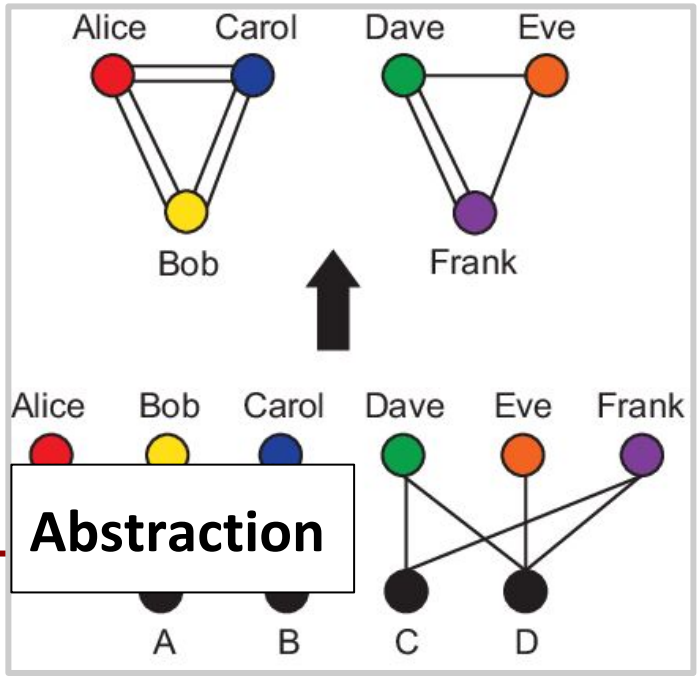


Embeddings and link prediction

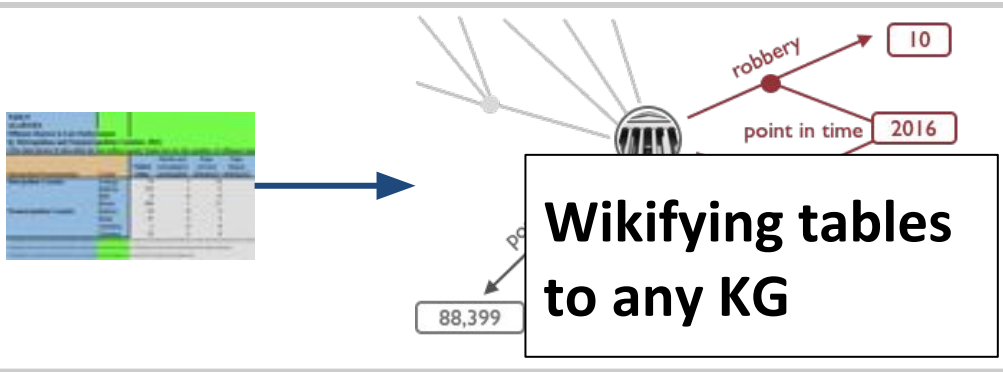


Browsing KGTK files with SQID

Ongoing work

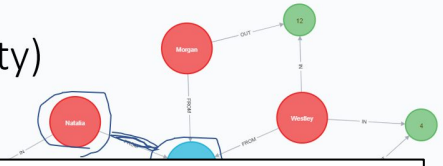


Abstraction



Wikifying tables to any KG

```
MATCH (p:Person)-[:FROM]->(c:City)
WHERE p.name = "Natalia"
RETURN p, c
```

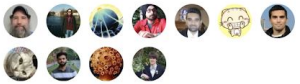


Integration of Cypher-like language

KGTK on GitHub

README_dev.md	add code snippet for add_default_arguments	6 months ago
mkdocs.yml	import framenet first version	2 months ago
requirements-dev.txt	add tox, add mkdocs and precommit to makefile	4 months ago
requirements-full.txt	update req	4 months ago
requirements.txt	Need version 1.13 or later for sh in order to pass FDs properly.	2 months ago
setup.py	proper version number for lite	4 months ago
tox.ini	omit site packages from coverage	last month

Contributors 11



Languages

Python 98.8% Other 1.2%

README.md

KGTK: Knowledge Graph Toolkit

DOI 10.5281/zenodo.3828068 build passing coverage 21%

KGTK is a Python library for easy manipulation with knowledge graphs. It provides a flexible framework that allows chaining of common graph operations, such as: extraction of subgraphs, filtering, computation of graph metrics, validation, cleaning, generating embeddings, and so on. Its principal format is TSV, though we do support a number of other inputs.

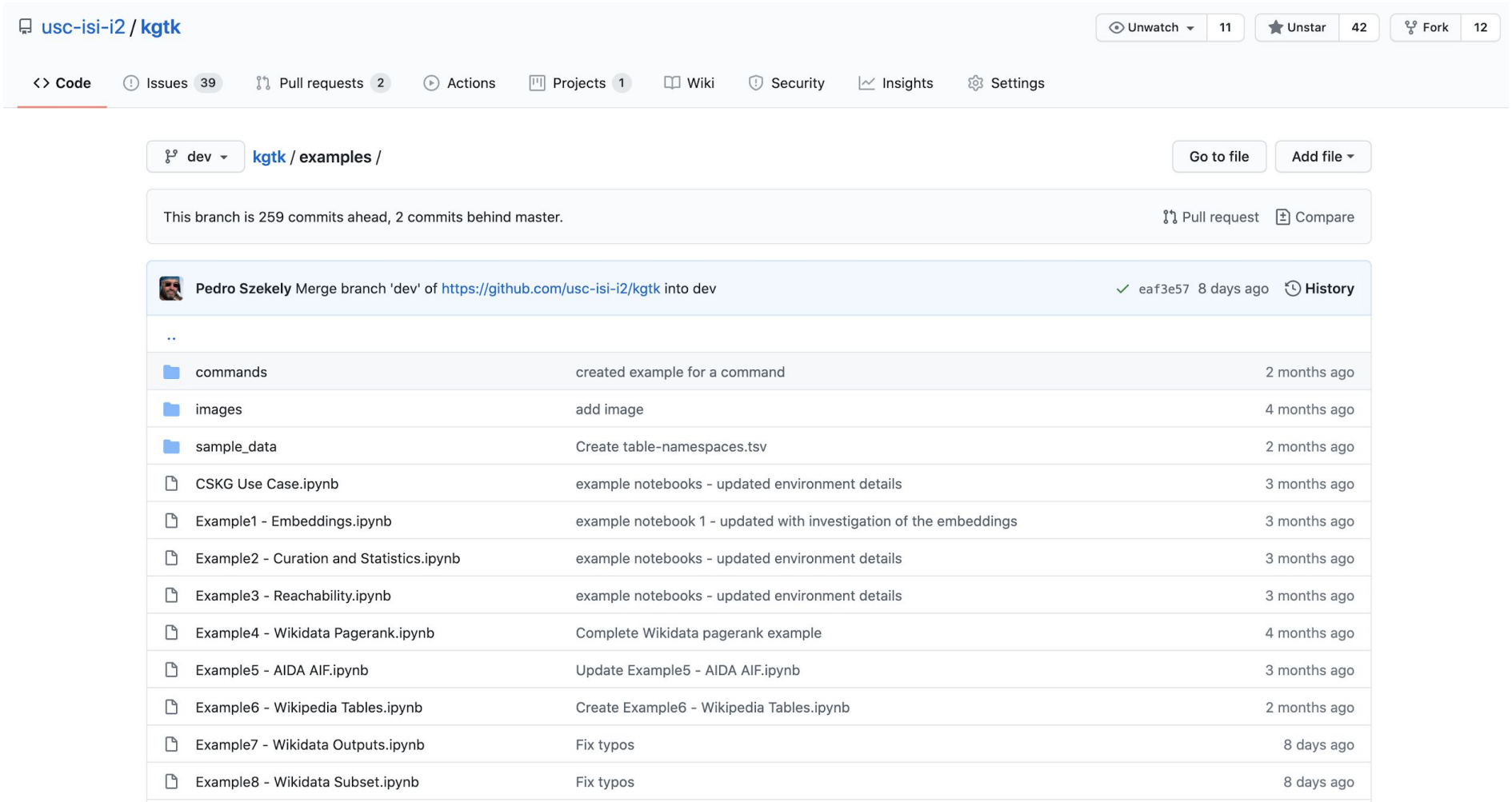
Features

- Computation of reachable nodes
- Filtering based on property values
- Removal of columns
- Sorting
- Computation of embeddings

https://github.com/usc-isi-i2/kgtk/

R5: appeal to AI practitioners

Example notebooks



usc-isi-i2 / kgtk

Unwatch 11 | Unstar 42 | Fork 12

Code | Issues 39 | Pull requests 2 | Actions | Projects 1 | Wiki | Security | Insights | Settings

dev | kgtk / examples / | Go to file | Add file

This branch is 259 commits ahead, 2 commits behind master. | Pull request | Compare

Pedro Szekely Merge branch 'dev' of https://github.com/usc-isi-i2/kgtk into dev — eaf3e57 8 days ago | History

..		
commands	created example for a command	2 months ago
images	add image	4 months ago
sample_data	Create table-namespaces.tsv	2 months ago
CSKG Use Case.ipynb	example notebooks - updated environment details	3 months ago
Example1 - Embeddings.ipynb	example notebook 1 - updated with investigation of the embeddings	3 months ago
Example2 - Curation and Statistics.ipynb	example notebooks - updated environment details	3 months ago
Example3 - Reachability.ipynb	example notebooks - updated environment details	3 months ago
Example4 - Wikidata Pagerank.ipynb	Complete Wikidata pagerank example	4 months ago
Example5 - AIDA AIF.ipynb	Update Example5 - AIDA AIF.ipynb	3 months ago
Example6 - Wikipedia Tables.ipynb	Create Example6 - Wikipedia Tables.ipynb	2 months ago
Example7 - Wikidata Outputs.ipynb	Fix typos	8 days ago
Example8 - Wikidata Subset.ipynb	Fix typos	8 days ago

https://github.com/usc-isi-i2/kgtk/examples

R5: appeal to AI practitioners

Democratizing Knowledge Graphs

Easy to use by all AI practitioners

NLP → spaCy

ML → scikit learn

KG → KGTK