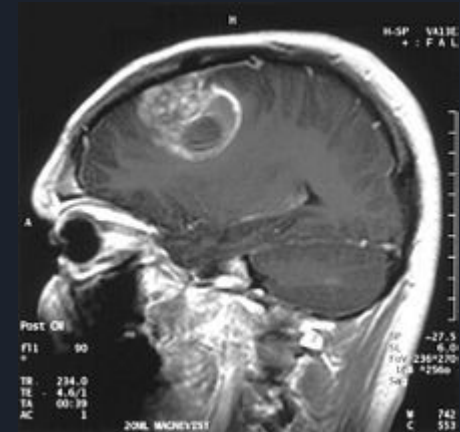
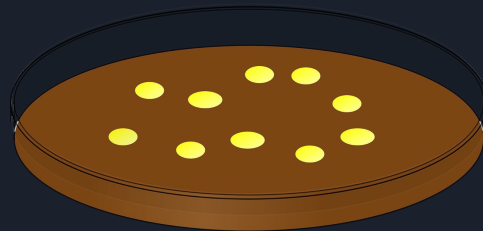


Literature-based knowledge discovery using PubTator and Wikidata

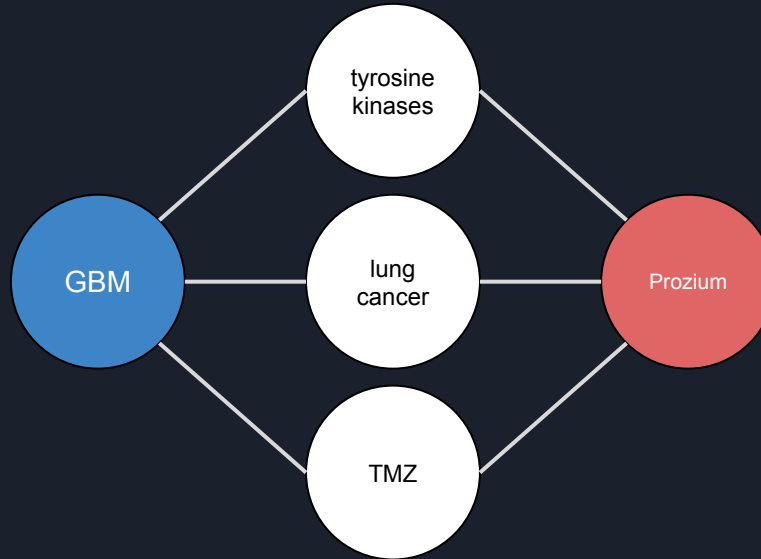
Jake Lever & Steven Jones
BC Cancer Agency & University of British
Columbia, Vancouver, Canada

Our End Goal

What are 10 novel drugs that may treat glioblastoma?



Knowledge discovery - hypothesis generation



Will **GBM** cooccur with **Proziom** in a future publication?

Cooccurrence based

Different levels of linked annotation

Applications

MESH:D000014215

MESH:D005909

Named entity recognition: Proziium treats Glioblastoma

medical condition treated (P2175)

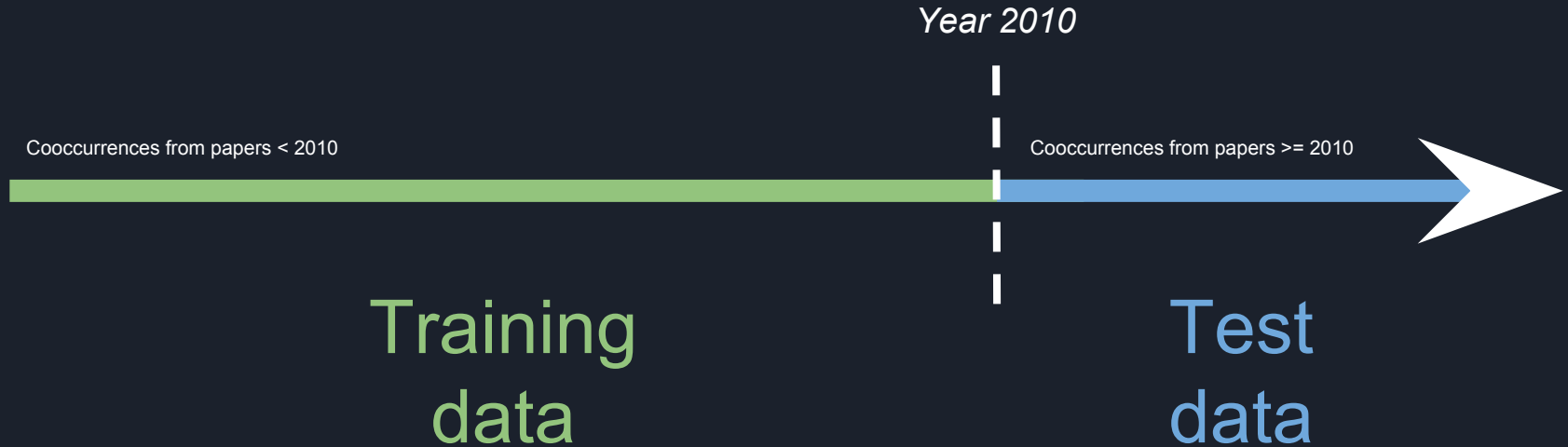
NER + Relation extraction: Proziium treats Glioblastoma

Pubmed Scale
NER + Relation extraction:

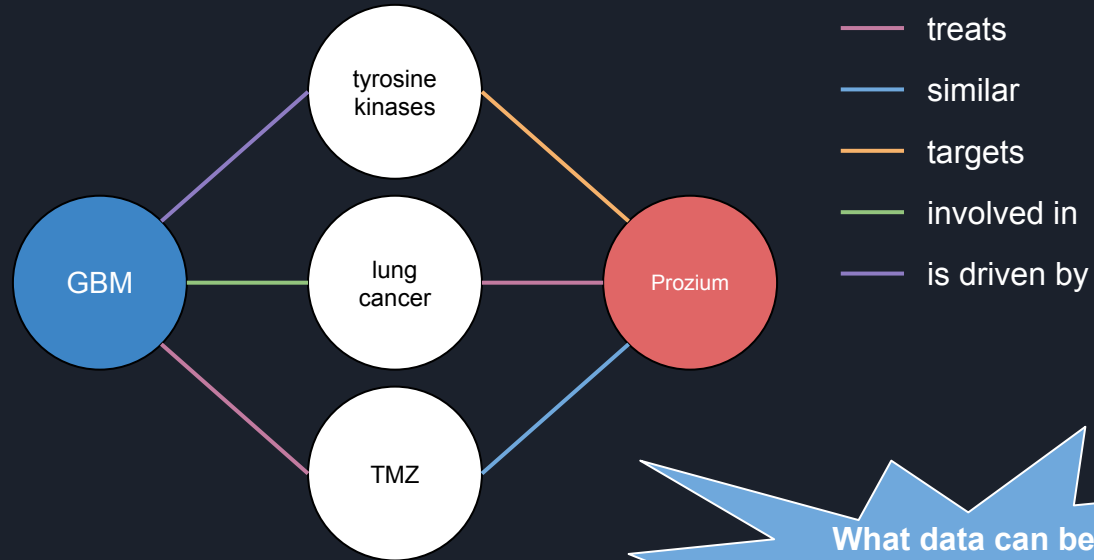
Although the epidermal growth factor receptor (EGFR) inhibitor erlotinib is initially effective in non-small-cell lung cancer (NSCLC) patients with tumors harboring activating mutations of EGFR,

PMID	Relation	ID1	ID2
28008301	treats	MESH:D018967	MESH:D012559
2542067	physically interacts	MESH:D007649	MESH:C564276
28039754	genetic association	MESH:D000212	MESH:D006562
1668116	treats	MESH:D000525	MESH:D001791

Evaluating a hypothesis generation system



Hypothesis generation with tuples



What relation (if any) may occur between GBM and Proziium in a future publication?

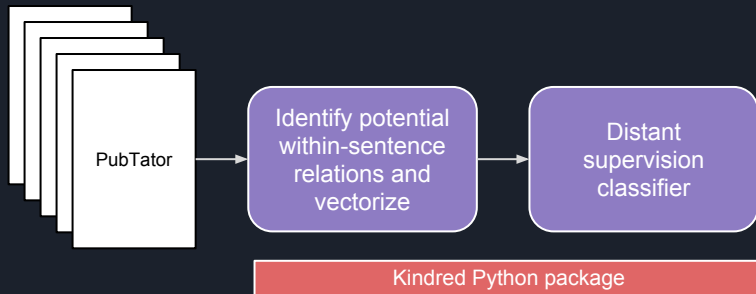
What data can be used to evaluate this system?



The Dataset

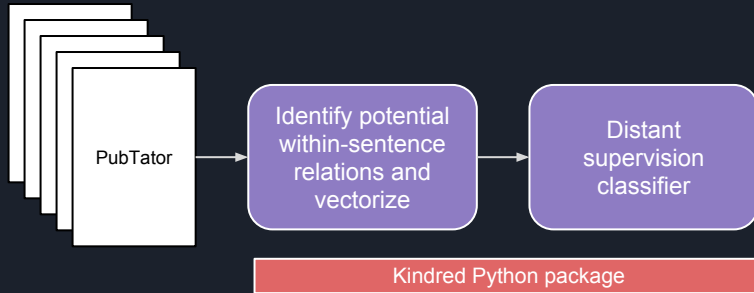


The Dataset



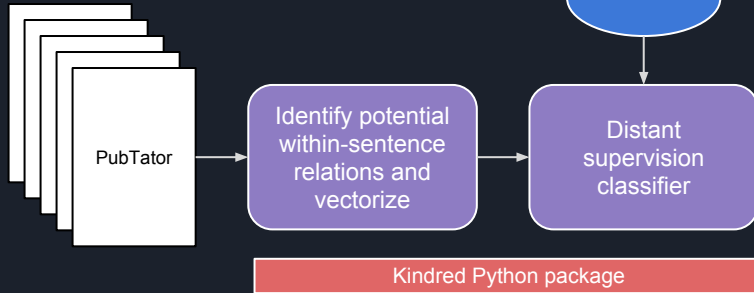
The Dataset

Built to update
easily with
PubRunner



The Dataset

Built to update
easily with
PubRunner



Relation Type	Entity 1	Entity 2	# of relations
Genetic Association	Disease	Gene	2,107
Medical Condition Treated	Chemical	Disease	1,993
Physically Interacts	Chemical	Gene	1,471

The Dataset

Built to update easily with PubRunner



Identify potential within-sentence relations and vectorize

Distant supervision classifier

Kindred Python package

Wikidata Seed Tuples

Relation Type	Entity 1	Entity 2	# of relations
Genetic Association	Disease	Gene	2,107
Medical Condition Treated	Chemical	Disease	1,993
Physically Interacts	Chemical	Gene	1,471

Relation Type	# in Training data	# in Test data
Genetic Association	41,284	20,651
Medical Condition Treated	85,754	25,367
Physically Interacts	70,645	28,702

All with associated PMIDs and publication dates

The Dataset

Built to update easily with PubRunner



Identify potential within-sentence relations and vectorize

Wikidata Seed Tuples

Distant supervision classifier

Kindred Python package

Relation Type	Entity 1	Entity 2	# of relations
Genetic Association	Disease	Gene	2,107
Medical Condition Treated	Chemical	Disease	1,993
Physically Interacts	Chemical	Gene	1,471

Relation Type	# in Training data	# in Test data
Genetic Association	41,284	20,651
Medical Condition Treated	85,754	25,367
Physically Interacts	70,645	28,702

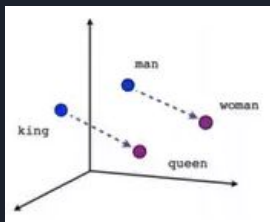
All with associated PMIDs and publication dates

Training data

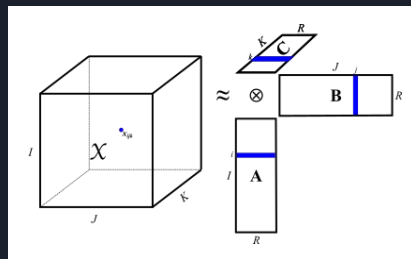
Test data

The Hackathon Proposal

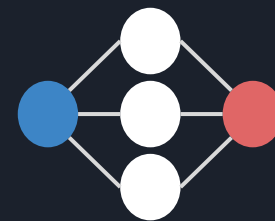
- Try out different hypothesis generation methods on this data set



Embedding-based (e.g. TransE)



Tensor-based (e.g. Rescal)



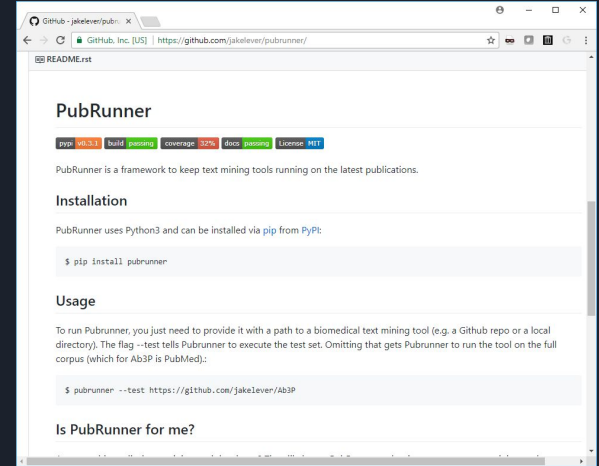
Path-based



Things I'd like to share

- Data
 - Vectorized representations of all within sentences relations in PubTator (using Kindred)
 - Dataset from this project
 - Precision cancer medicine-related annotations
- Software
 - PubRunner - Manage the execution of text mining tools and keep results up-to-date
 - Kindred - Simple relation extraction Python package

<https://github.com/jakelever/>



The screenshot shows the GitHub repository page for 'PubRunner' by Jake Levever. The page includes a header with the repository name, a progress bar showing 98.1% build success, 97% coverage, and MIT license. The main content describes PubRunner as a framework for text mining tools, provides installation instructions using pip, and shows a usage example with a test command.

PubRunner

PubRunner is a framework to keep text mining tools running on the latest publications.

Installation

PubRunner uses Python3 and can be installed via pip from PyPI:

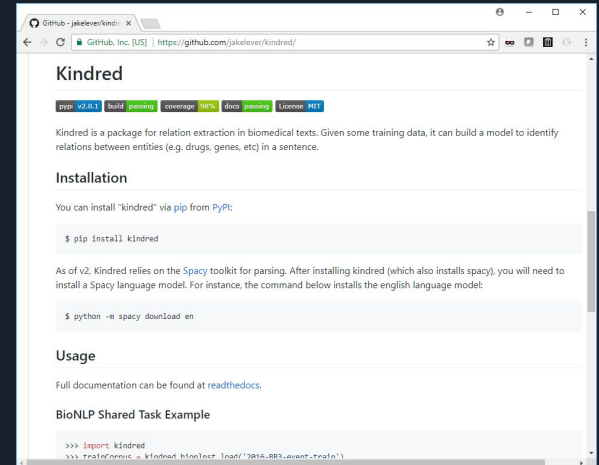
```
$ pip install pubrunner
```

Usage

To run Pubrunner, you just need to provide it with a path to a biomedical text mining tool (e.g. a GitHub repo or a local directory). The flag `--test` tells Pubrunner to execute the test set. Omitting that gets Pubrunner to run the tool on the full corpus (which for Ab3P is PubMed):

```
$ pubrunner --test https://github.com/jakelever/Ab3P
```

Is PubRunner for me?



The screenshot shows the GitHub repository page for 'Kindred' by Jake Levever. The page includes a header with the repository name, a progress bar showing 98% build success, 98% coverage, and MIT license. The main content describes Kindred as a package for relation extraction in biomedical texts, provides installation instructions using pip, and shows a usage example with a test command.

Kindred

Kindred is a package for relation extraction in biomedical texts. Given some training data, it can build a model to identify relations between entities (e.g. drugs, genes, etc) in a sentence.

Installation

You can install "kindred" via pip from PyPI:

```
$ pip install kindred
```

As of v2, Kindred relies on the Spacy toolkit for parsing. After installing kindred (which also installs spacy), you will need to install a Spacy language model. For instance, the command below installs the english language model:

```
$ python -m spacy download en
```

Usage

Full documentation can be found at [readthedocs](#).

BioNLP Shared Task Example

```
>>> import kindred
>>> train_corpus = kindred.MinText.load('2014-RR1-annot_train')
```

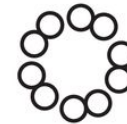
Thank You & Acknowledgements

<https://github.com/jakelever/>

- Steven Jones
- Martin Jones
- Eric Zhao
- Jasleen Grewal
- Jones Lab @ UBC



Bourses d'études
supérieures du Canada
Vanier
Canada Graduate
Scholarships



CANADA'S MICHAEL SMITH
**GENOME
SCIENCES**
CENTRE



compute | **calcul**
canada | canada



CIHR/MSFHR
BIOINFORMATICS
TRAINING PROGRAM FOR HEALTH RESEARCH



BC Cancer Agency
CARE + RESEARCH