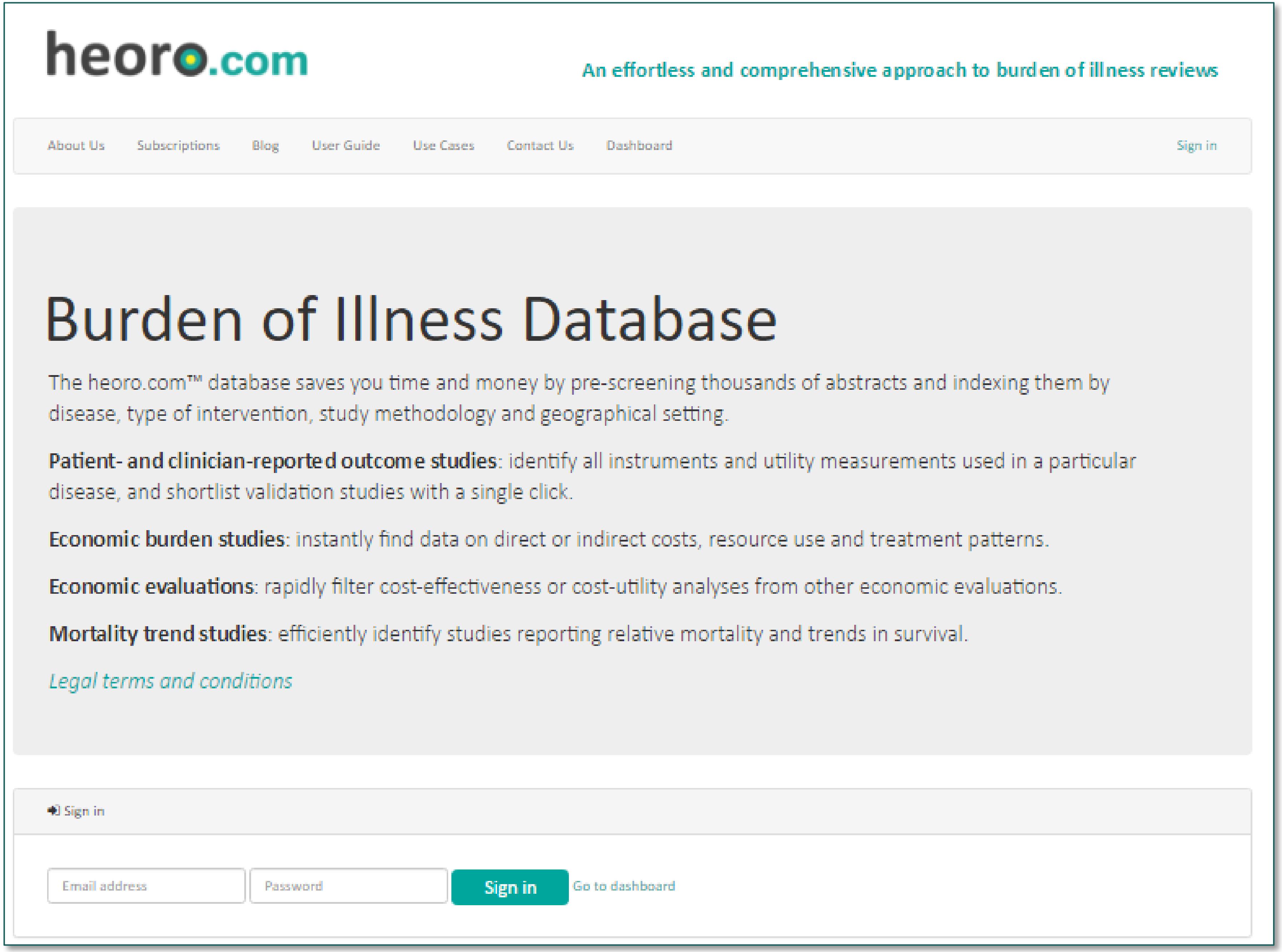
Accuracy of natural language processing-based classifiers for automated identification of studies on humanistic and economic burden of disease

Jean-Baptiste Krohn

Crystallise

heoro.com



An effortless and comprehensive approach to burden of illness reviews

Burden of Illness Database

The heoro.com™ database saves you time and money by pre-screening thousands of abstracts and indexing them by disease, type of intervention, study methodology and geographical setting.

Patient- and clinician-reported outcome studies: identify all instruments and utility measurements used in a particular disease, and shortlist validation studies with a single click.

Economic burden studies: instantly find data on direct or indirect costs, resource use and treatment patterns.

Economic evaluations: rapidly filter cost-effectiveness or cost-utility analyses from other economic evaluations.

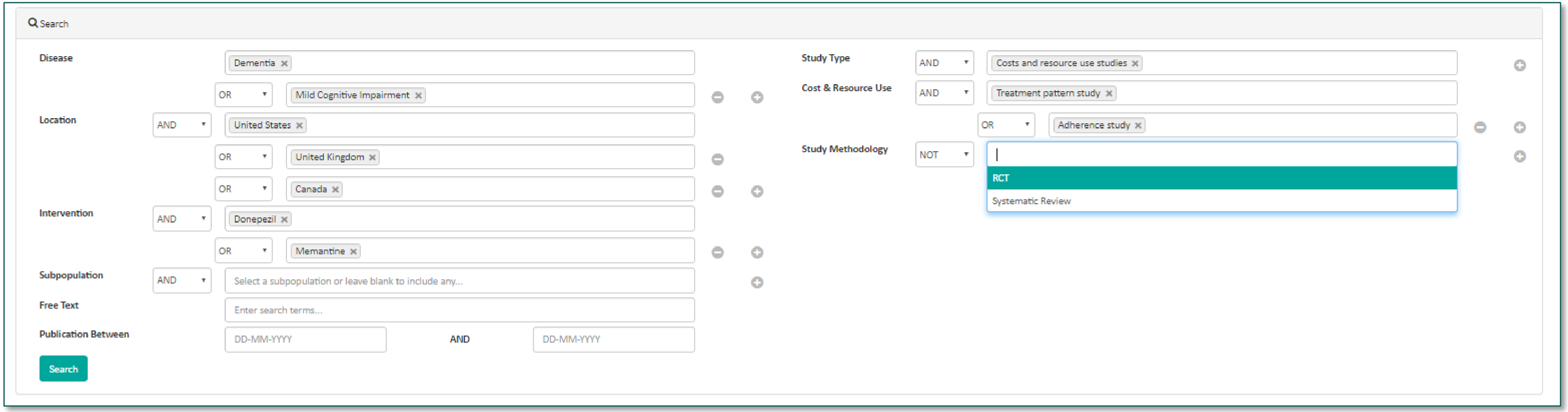
Mortality trend studies: efficiently identify studies reporting relative mortality and trends in survival.

Legal terms and conditions

The problem

- >100,000 abstracts from PubMed search for quality of life, economic burden, economic evaluations and mortality since 2005
- >180,000 abstracts since 1960
- Need a quick, affordable and accurate way to index these

The data



Search

Disease: Dementia ×
OR: Mild Cognitive Impairment ×

Location: AND United States ×
OR: United Kingdom ×
OR: Canada ×

Intervention: AND Donepezil ×
OR: Memantine ×

Subpopulation: AND Select a subpopulation or leave blank to include any...

Free Text: Enter search terms...

Publication Between: DD-MM-YYYY AND DD-MM-YYYY

Study Type: AND Costs and resource use studies ×

Cost & Resource Use: AND Treatment pattern study ×
OR: Adherence study ×

Study Methodology: NOT
RCT
Systematic Review

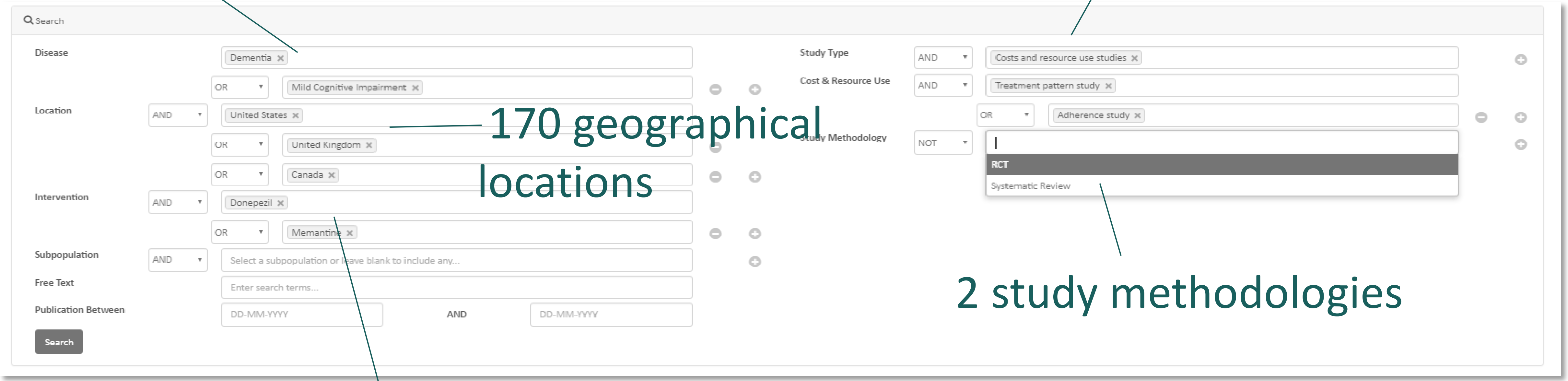
Crystallise

heoro.com

The data

>4,700 diseases

4 study types,
11 sub-types



Search

Disease: Dementia ×
OR Mild Cognitive Impairment ×

Location: AND United States ×
OR United Kingdom ×
OR Canada ×

Intervention: AND Donepezil ×
OR Memantine ×

Subpopulation: AND Select a subpopulation or leave blank to include any...

Free Text: Enter search terms...

Publication Between: DD-MM-YYYY AND DD-MM-YYYY

Study Type: AND Costs and resource use studies ×

Cost & Resource Use: AND Treatment pattern study ×
OR Adherence study ×

Study Methodology: NOT |
RCT
Systematic Review

170 geographical
locations

2 study methodologies

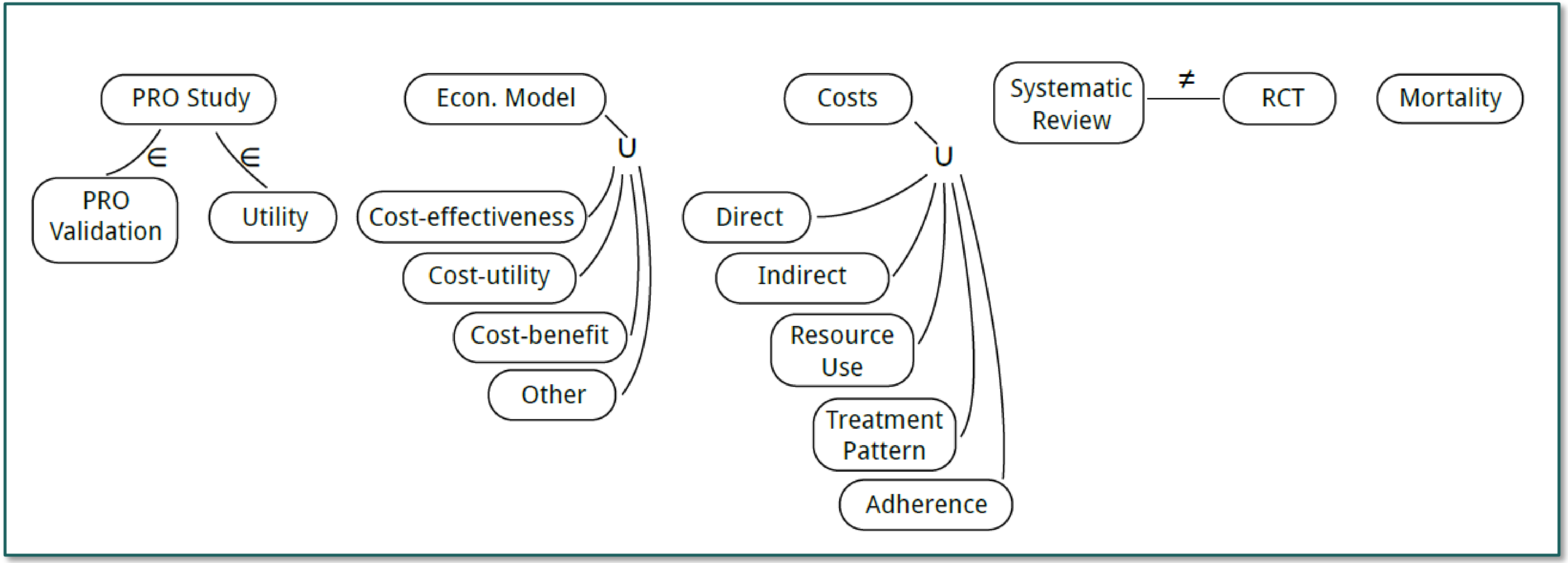
>8,400 interventions

>5,000 PRO instruments

Crystallise

heoro.com

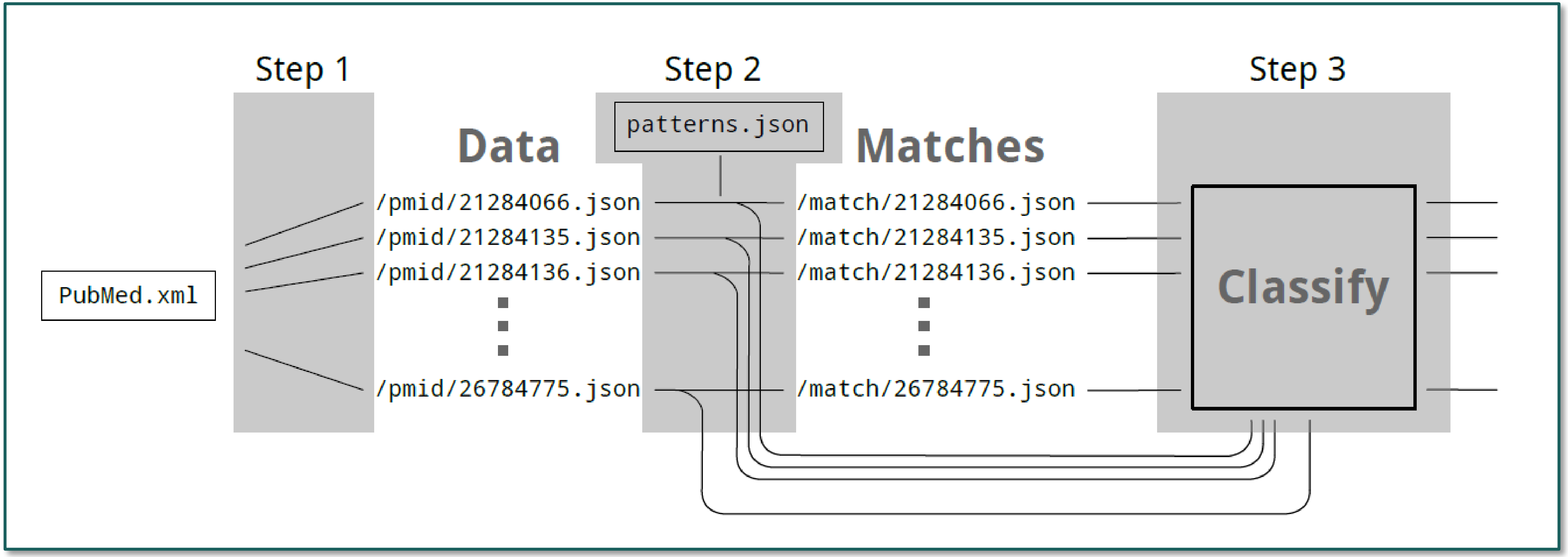
Type constraints



Crystallise

heoro.com

The approach



Crystallise

heoro.com

Step 1: PubMed data extraction

- Text: title and abstract.
- Structure: abstract paragraphs (e.g. OBJECTIVES, METHODS, CONCLUSIONS)
- Metadata: MeSH headings, journal, keywords, authors, etc

Step 2: Pattern matching

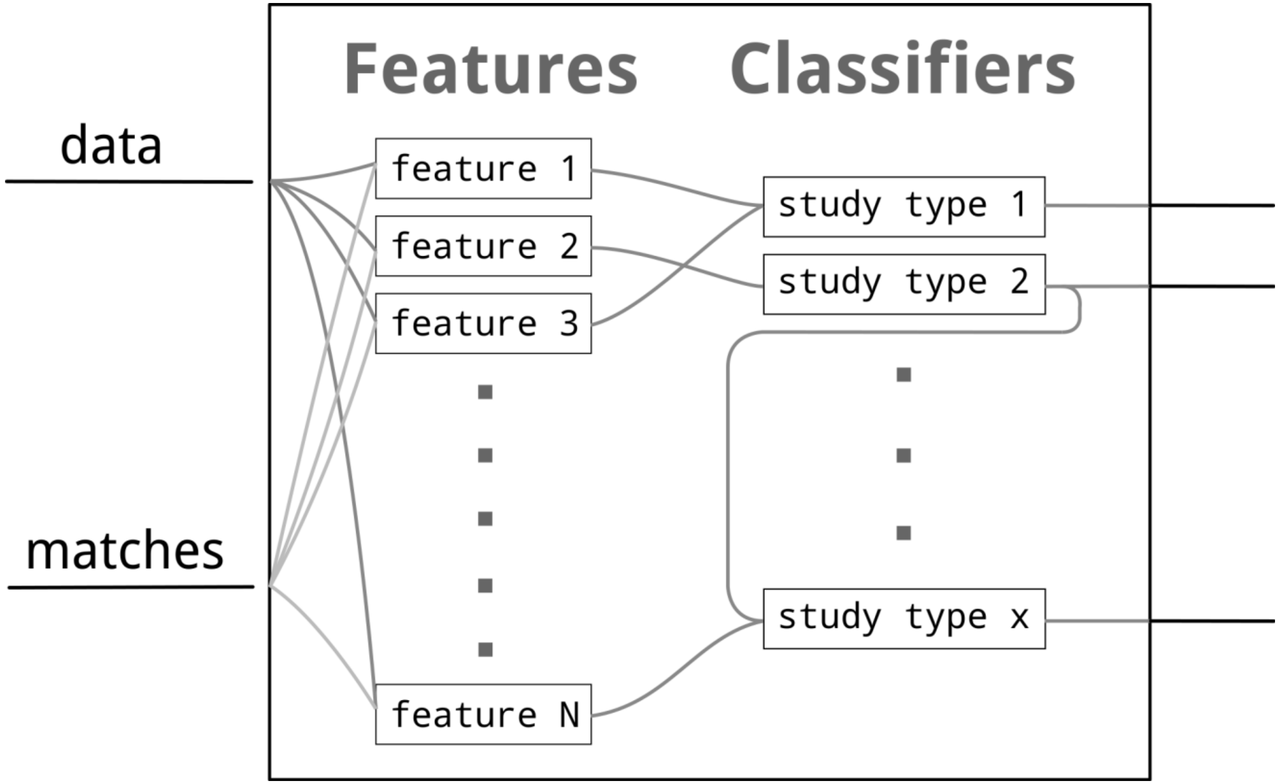
- Example pattern (JSON):

```
“sf6d”: [
  “/\\b[sS]hort(\\s+|-)[Ff]orm((\\s+|)[sS]urvey)?(\\s+|)6[Dd]\\b/”,
  “/\\bSF-6D\\b/i”
]
```

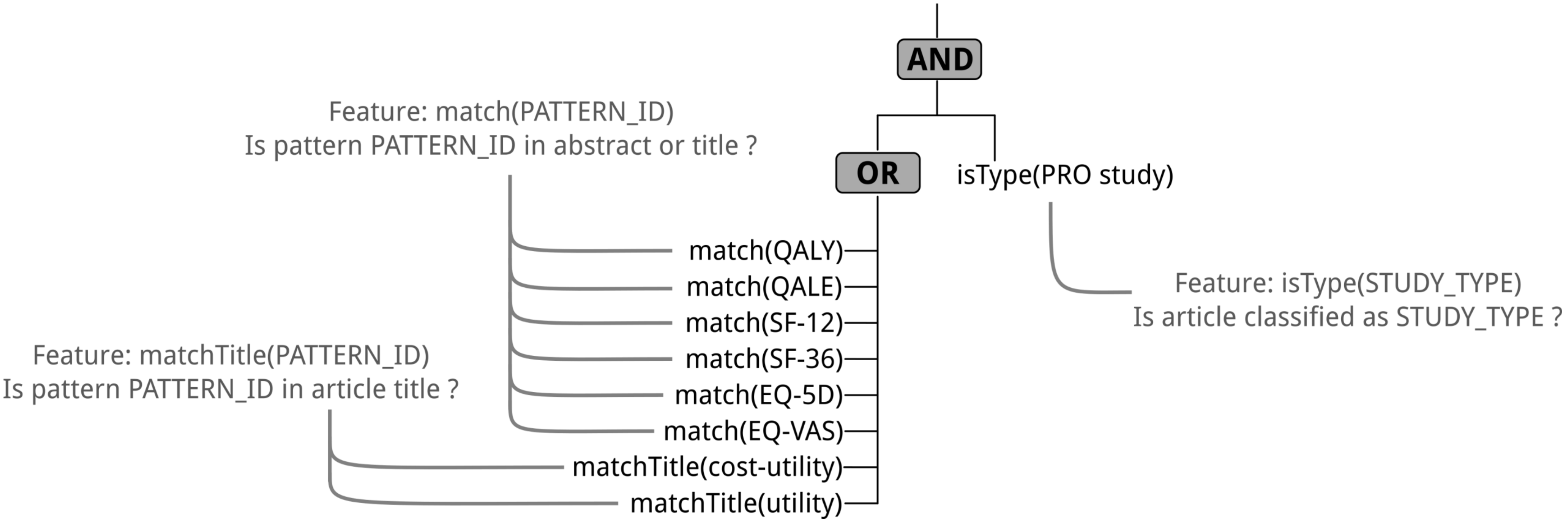
- Example matches (JSON):

```
{“name”:“sf6d”,“path”:[“abstract”,1,“text”],“details”:[292,“Short Form-6D”]}
{“name”:“sf6d”,“path”:[“abstract”,5,“text”],“details”:[ 80,        “SF-6D”]}
```

Step 3: Classification



Step 3: Example classifier



AND

OR

isType(PRO study)

match(QALY)

match(QALE)

match(SF-12)

match(SF-36)

match(EQ-5D)

match(EQ-VAS)

matchTitle(cost-utility)

matchTitle(utility)

Feature: match(PATTERN_ID)
Is pattern PATTERN_ID in abstract or title ?

Feature: matchTitle(PATTERN_ID)
Is pattern PATTERN_ID in article title ?

Feature: isType(STUDY_TYPE)
Is article classified as STUDY_TYPE ?

Learning process

- Human expert assisted by Machine Learning system suggesting improvements based on learning set.

- Human expert intervenes on:
  - Pattern matching: define expressions to detect
  - Feature selection: determine features to use in classifier
  - Classifier structure: define boolean expression combining features

Classifier accuracy: publications from 2005-2016

Study type	Sensitivity	Specificity
PRO studies	96%	96%
Economic models	95%	97%
Costs and resource use	92%	95%
Mortality	82%	97%
RCT	95%	99%
Systematic review	93%	99%
Geographical location	97%	

Classifier accuracy: publications from 2005-2016

Study type	Sensitivity	Specificity
PRO studies	96%	96%
PRO validation	99%	97%
Utilities	100%	98%

Classifier accuracy: publications from 2005-2016

Study type	Sensitivity	Specificity
Economic models	95%	97%
Cost-effectiveness	87%	100%
Cost-utility	99%	98%
Cost-benefit	99%	97%
Other	93%	93%

Crystallise

heoro.com

Classifier accuracy: publications from 2005-2016

Study type	Sensitivity	Specificity
Costs and resource use	92%	95%
Direct costs	79%	98%
Indirect costs	97%	97%
Resource use	90%	97%
Treatment patterns	99%	93%
Adherence	99%	92%

Classifier accuracy: publications from 1960-2004

Study type	Sensitivity	Specificity
PRO studies	90%	90%
PRO validation	98%	91%
Utilities	99%	99%
Cost-effectiveness models	96%	99%
Cost-utility models	100%	99%
Cost-benefit models	100%	100%
Other models	86%	94%

Classifier accuracy: publications from 1960-2004

Study type	Sensitivity	Specificity
Direct costs	80%	93%
Indirect costs	94%	84%
Resource use	76%	91%
Treatment patterns	98%	82%
Adherence	96%	83%
RCTs	97%	98%
Systematic reviews	95%	98%
Mortality	96%	91%

Questions?

- **Contact:**
- Jean-Baptiste Krohn
- mail@jbk.io

- Alison Martin
- alison.martin@crystallise.com
- www.heoro.com