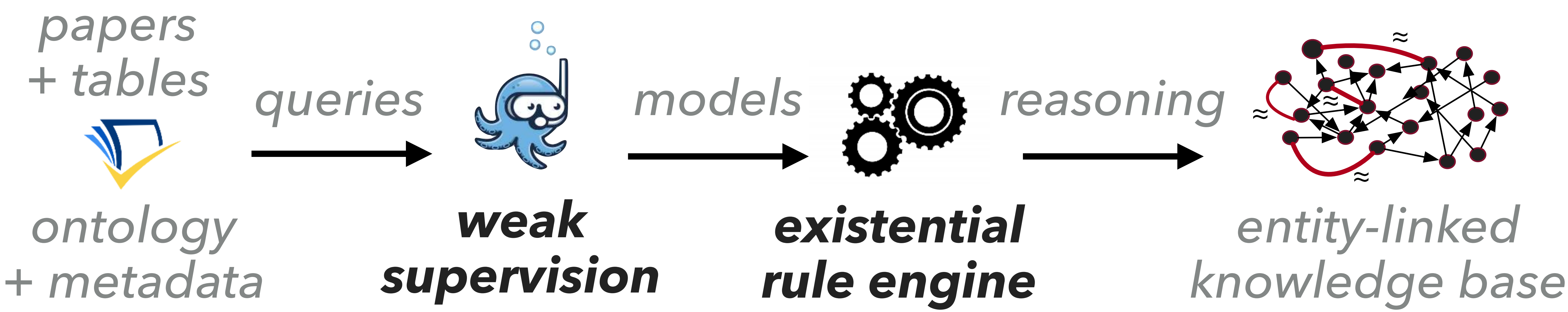
BENNO KRUIT, HONGYU HE, JACOPO URBANI

TAB2KNOW:

BUILDING A KNOWLEDGE BASE FROM TABLES IN SCIENTIFIC PAPERS

MAIN TAKEAWAY: WEAK SUPERVISION + EXISTENTIAL RULES



- **Bootstrap** a KB from tables in scientific papers
  - minimize input from subject matter **experts**
  - maximize **generalization** and **control**
- **Combine** ontology, rules, and machine learning
  - created KB from **73**k tables in context

code: github.com/karmaresearch/tab2know

data: doi.org/10.5281/zenodo.3983013

TABLES IN SCIENTIFIC PAPERS

- Structured information about scientific process
  - similar structure **across documents**
- Could support **reviews** or **search**
- Examples of tables for **human readers**

How do we convert semi-structured data
into high-quality semantic data
with as little effort as possible?

PROBLEMS

- Tables must be **extracted** from PDFs
  - reconstruct from PDF text-boxes!
- Every author uses different **conventions**
  - *e.g.* structure, jargon, layout, formats
- No **Knowledge Base** to link concepts
- Interpretation is **goal-specific**
  - Both **construction** and **querying** must be user-oriented and flexible

TABLE I. RANKING OF SUBMITTED METHODS TO TASK 1.1

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	**82.38**	**93.83**	**87.74**
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

TAB2KNOW

- A system for **constructing** and **querying** a Knowledge Base of information extracted from tables in scientific papers
  1. **Structural** foundation: simple graph of extracted structure
  2. **Semantic** layer: predicted types of tables and columns
  3. **Entity** layer: similar cells resolved to entity clusters
- Based on user-written **rules** and **queries**
  - used as **weak supervision** for machine learning models

1. STRUCTURAL FOUNDATION

TABLE I. RANKING OF SUBMITTED METHODS TO TASK 1.1

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	**82.38**	**93.83**	**87.74**
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

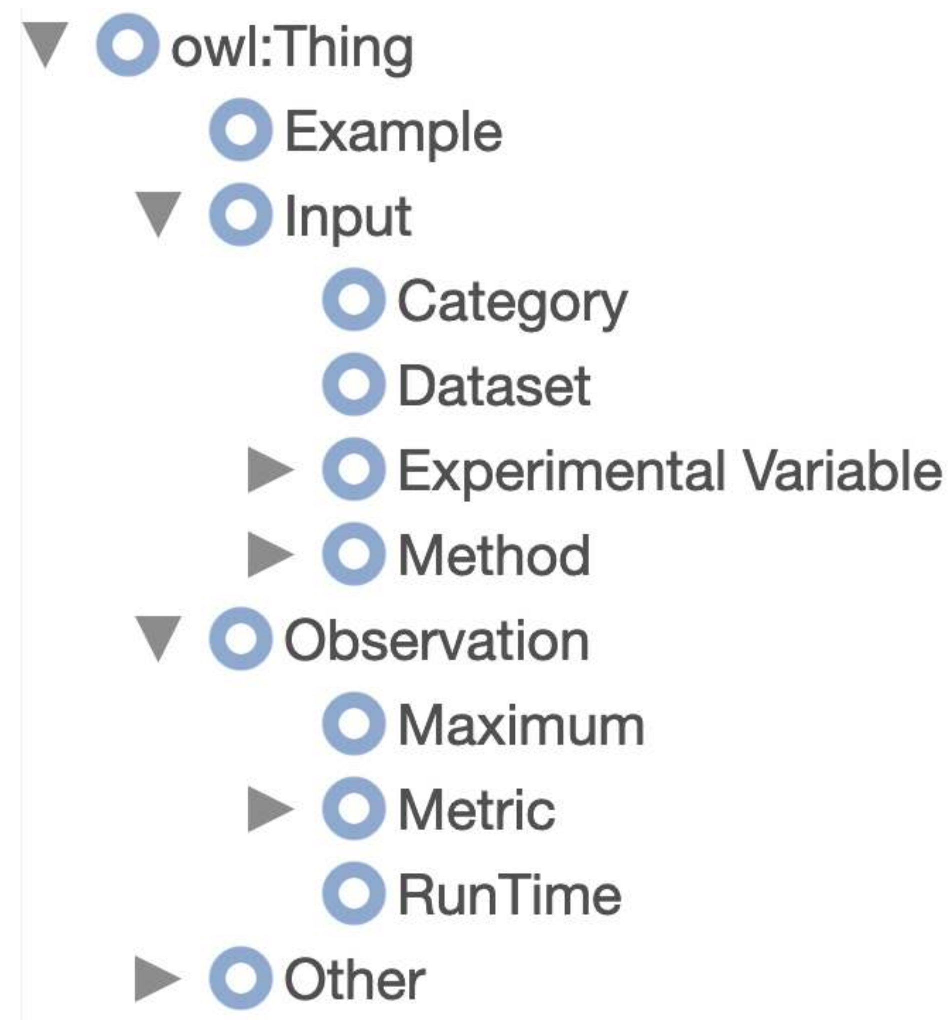
1 **Table Extraction**

```
PREFIX : http://xzy/tab2know
:Table1 :hasRow :Table1-r1
:Table1-r1 rdf:type :Row
:Table1-r1 :rowIndex 1^^⟨xsd:int⟩
:Table1-r1c1 :cellOf :Table1
:Table1-r1c1 :rowIdx 1^^⟨xsd:int⟩
:Table1-r1c1 rdf:value "Method name"
```

- Naïve, **structural** graph + paper metadata in RDF
  - *pdffigures2* for table detection and *Tabula* for extraction
- Already supports **rich queries**
  - frequent cells, co-occurrence, column headers, datatypes, venues, authors, …

2. SEMANTIC LAYER

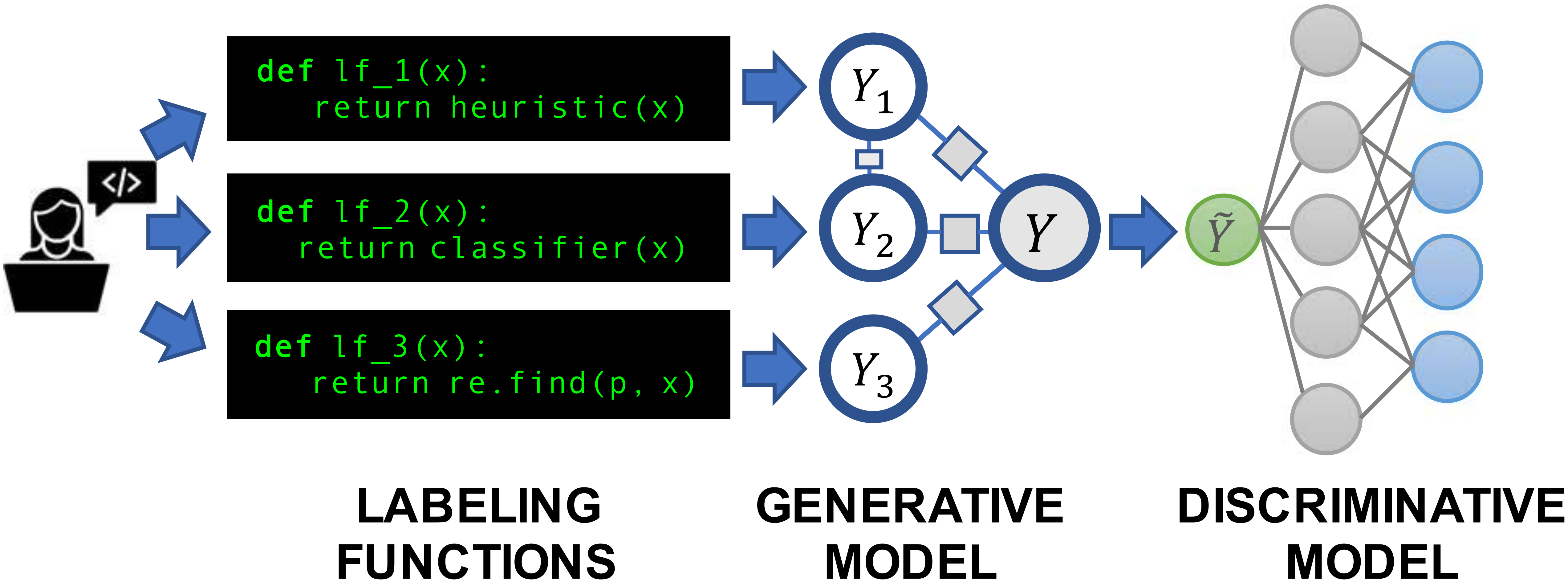
- Table interpretation
  - Much research on web tables + existing KB
  - But we have **expert tables** + **no KB**!
    - Labeling data is **expensive**
- Generalize from expert heuristics using weak supervision



WEAK SUPERVISION

- Hand-labeling data is **expensive** and **inflexible**
  - … so write **labeling functions** instead!
- Aggregate **weak signals** for training ML models
  - exploit labeling function **correlations** and **sparsity**
- **Snorkel** (Ratner et al, VLDB2020)
  - DryBell @ Google (Bach et al., SIGMOD 2019)
  - Overton @ Apple (Christopher Ré et al., 2019)

WEAK SUPERVISION



(Ratner et al)

EXAMPLE: TABLE TYPES

l_1-l_2	#S	#l_1-W	#l_2-W	#l_1-V	#l_2-V
en-de	1.9M	55M	52M	40k	50k
en-fr	2.0M	50M	51M	40k	50k
en-es	1.9M	49M	51M	40k	50k

(a) Input

Models	Rerank size	Beam size	GMV	Latency
miDNN	50	-	2.91%	9%
miRNN	50	5	5.03%	58%
miRNN+att.	50	5	5.82%	401%

(b) Observation

Type	**Example Words**
Offensive	disgusting, filthy, nasty, rude, horrible, terrible, awful, worst, idiotic, stupid, dumb, ugly, etc.
Non-offensive	help, love, respect, believe, congrats, hi, like, great, fun, nice, neat, happy, good, best, etc.

(c) Example

α_c	DP concentration parameter for each $c \in V$
$P_0(e\|c)$	CFG base distribution
$\boldsymbol{x}$	Set of non-terminal nodes in the treebank
$\mathcal{S}$	Set of sampling sites (one for each $x \in \boldsymbol{x}$)
S	A block of sampling sites, where $S \subseteq \mathcal{S}$
$\boldsymbol{b} = \{b_s\}_{s \in \mathcal{S}}$	Binary variables to be sampled ($b_s = 1 \rightarrow$ frontier node)
$\boldsymbol{z}$	Latent state of the segmented treebank
m	Number of sites $s \in S$ s.t. $b_S = 1$
$\boldsymbol{n} = \{n_{c,e}\}$	Sufficient statistics of $\boldsymbol{z}$
$\Delta n^{S:m}$	Change in counts by setting m sites in S

(d) Other

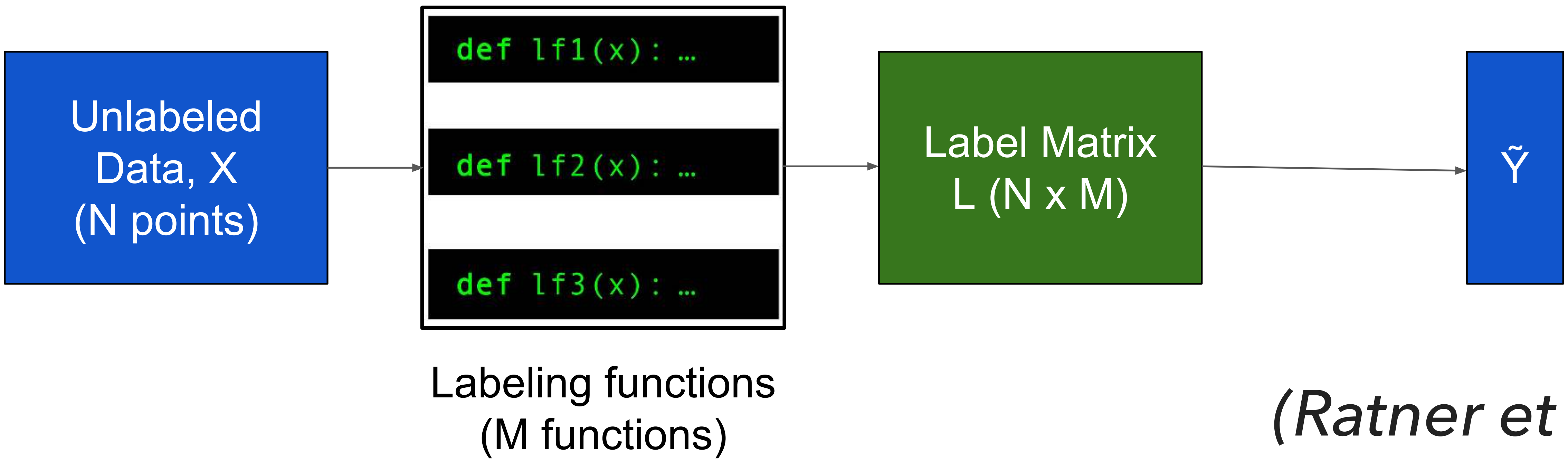
EXAMPLE: TABLE TYPES

Type	Example Words
Offensive	disgusting, filthy, nasty, rude, horrible, terrible, awful, worst, idiotic, stupid, dumb, ugly, etc.
Non-offensive	help, love, respect, believe, congrats, hi, like, great, fun, nice, neat, happy, good, best, etc.

(c) Example

```
select distinct ?table where {
    SERVICE bds:search {
      ?matchedValue bds:search "example" .
    }
    ?table dct:title  ?matchedValue .
}
```

WEAKLY SUPERVISED MODEL TRAINING



(Ratner et al)

- Two options for aggregating labels:
  - Majority Vote
  - Snorkel Model
- Many options for ML model, but ***must not overfit!***

3. ENTITY RESOLUTION

- **Vlog**: Large-scale reasoning on **contexts** of cell values
  - e.g. column header, column type, author, …
  - If similar, merge cell values into entity clusters
- Scales to very large graphs (Best Resource @ ISWC 2019)

$$type(X, \texttt{Column}) \rightarrow \exists Y. colEntity(X, Y)$$
$$type(X, \texttt{Cell}) \rightarrow \exists Y. cellEntity(X, Y)$$

$$\begin{aligned}
ceNoTypLabel(X, L), ceNoTypLabel(Y, L) &\rightarrow X \approx Y \\
eNoTypLabel(X, C, L), eNoTypLabel(Y, C, L) &\rightarrow X \approx Y \\
eTableLabel(X, T, L), eTableLabel(Y, T, L) &\rightarrow X \approx Y \\
eTypLabel(X, S, L), eTypLabel(Y, S, M), STR_EQ(L, M) &\rightarrow X \approx Y \\
eAuthLabel(X, A, L), eAuthLabel(Y, A, M), STR_EQ(L, M) &\rightarrow X \approx Y
\end{aligned}$$

METHOD SUMMARY

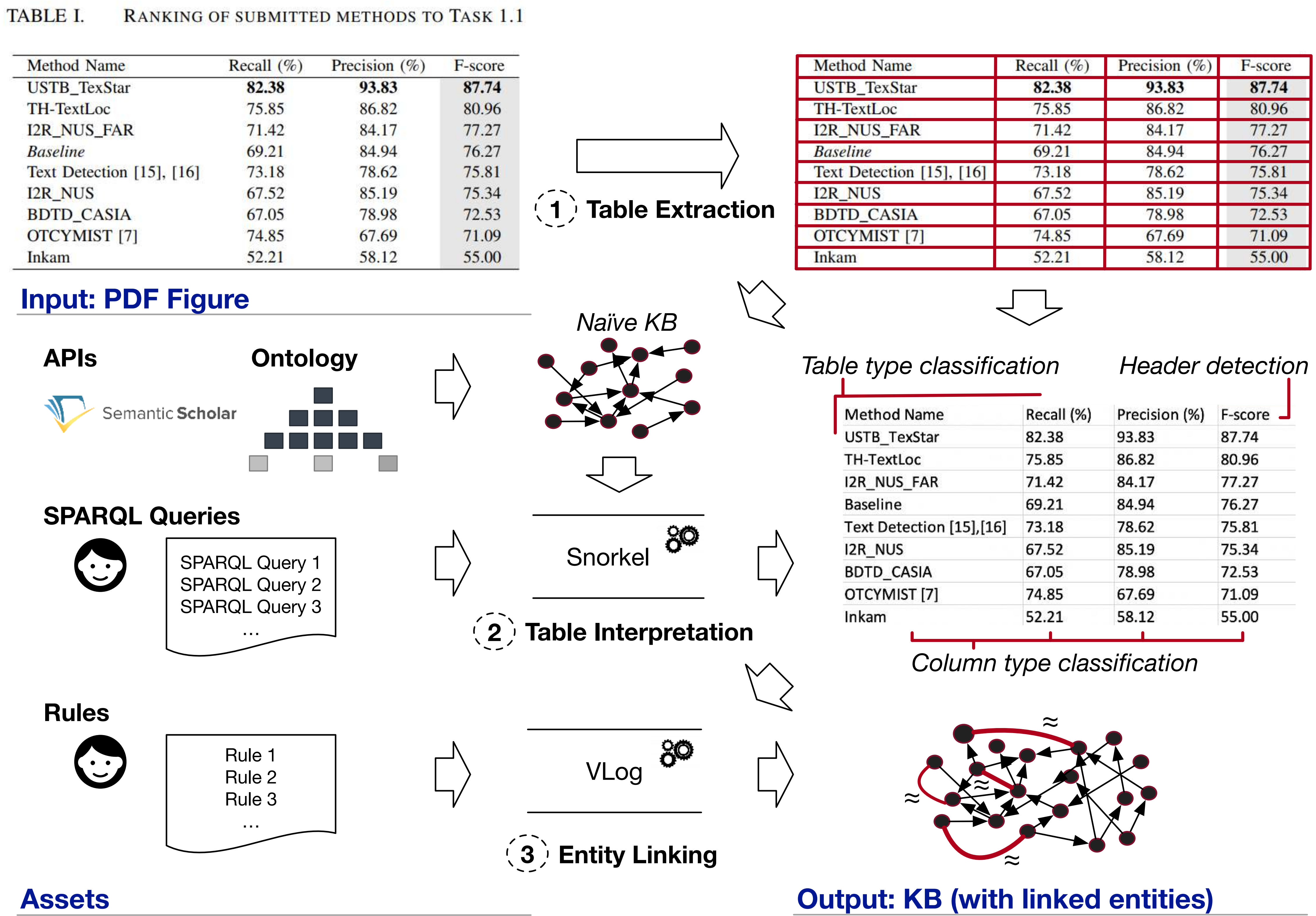


TABLE I. RANKING OF SUBMITTED METHODS TO TASK 1.1

Method Name	Recall (%)	Precision (%)	F-score
USTB_TexStar	**82.38**	**93.83**	**87.74**
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [15], [16]	73.18	78.62	75.81
I2R_NUS	67.52	85.19	75.34
BDTD_CASIA	67.05	78.98	72.53
OTCYMIST [7]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

RESULTS: TABLE TYPES

- Gold standard: **400** sampled tables from 17 AI venues
- Manual annotation: **4** table types, **39** label queries
- Features from table **caption**, **header** cells and **body** cells

Model	Prec.	Recall	F1	AUC
SVM	0.71	0.79	0.74	0.86
LR	0.72	0.79	0.74	0.84
NB	**0.80**	**0.82**	**0.79**	**0.91**

ML model performance on Tab2Know data

RESULTS: COLUMN TYPES

Tab2Know data

- **22** column types
- **55** label queries

Model	Prec.	Recall	F1
NB	0.52	0.48	0.47
SVM	**0.58**	**0.56**	**0.53**
LR	**0.58**	**0.56**	**0.53**

ML model performance on Tab2Know data

Tablepedia data

- **3** column types
- **15** seed concepts

Model	Prec.	Recall	F1
Yu et al.	0.82	0.81	0.81
NB	0.84	0.82	0.81
SVM	0.90	0.89	0.89
LR	**0.92**	**0.91**	**0.91**

ML model performance on Tablepedia data

RESULTS: COLUMN TYPES

Our data is more challenging

Model	Prec.	Recall	F1
NB	0.52	0.48	0.47
SVM	**0.58**	**0.56**	**0.53**
LR	**0.58**	**0.56**	**0.53**

ML model performance on Tab2Know data

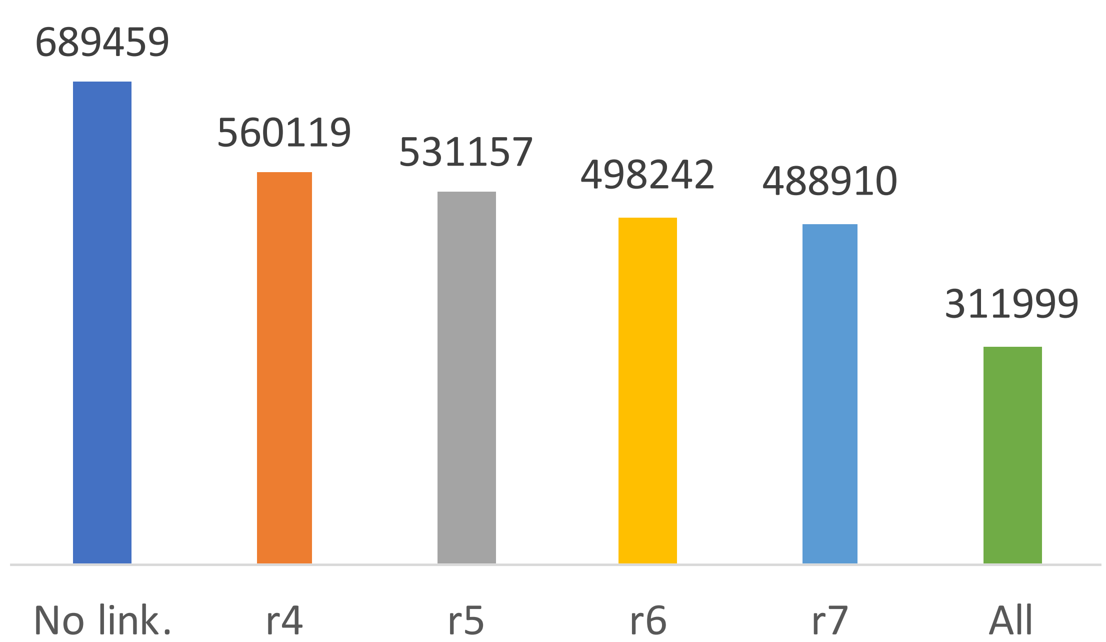
Our model outperforms previous work

Model	Prec.	Recall	F1
Yu et al.	0.82	0.81	0.81
NB	0.84	0.82	0.81
SVM	0.90	0.89	0.89
LR	**0.92**	**0.91**	**0.91**

ML model performance on Tablepedia data

RESULTS: ENTITY RESOLUTION

- **2** entity creation (TGD) + **5** entity merging rules (EGD)
- **65%** entities are sensible, **97%** mergers are good



Number of entities per rule

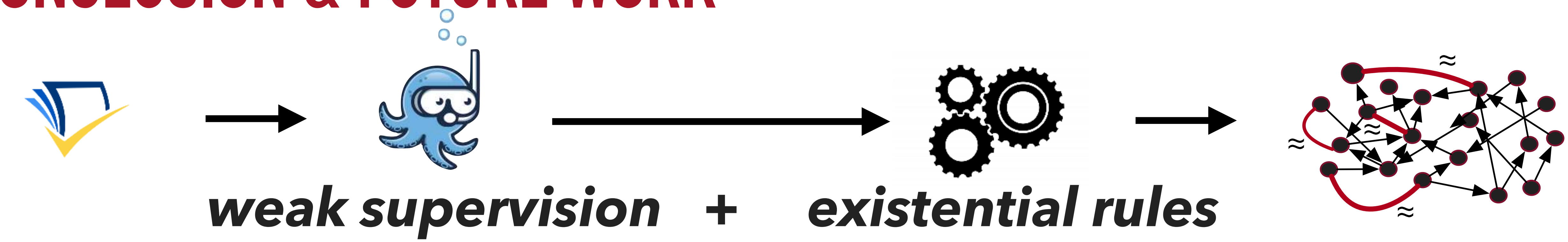
	Label	# Links
Good ✓	mnist	288
	knn	211
	wiki	146
	cifar-10	108
	en-es	65
Bad ✗	after	183
	analysis	66
	subset	49
	0/0/0	9
	f4(x)	6

Examples

RESULTS: KNOWLEDGE GRAPH

- **143k** PDFs from Semantic Scholar for **17 AI venues**
  - **73**k tables extracted
  - **31**M triples in graph
- All data available as public resources
  - data: **doi.org/10.5281/zenodo.3983013**
  - code: **github.com/karmaresearch/tab2know**

CONCLUSION & FUTURE WORK



- **Bootstrapped** a KB from tables in scientific papers
  - **learn** from heuristics + **reason** with rules
- **Outperformed** previous method & **introduced** harder dataset
  - created KB with from **73**k tables
- Future: quantify **effort** needed from experts
  - **goal**: create platform for task-specific KB extraction

code: github.com/karmaresearch/tab2know
data: doi.org/10.5281/zenodo.3983013

BENNO KRUIT kruit@cwi.nl