Web-Scale Extension of RDF Knowledge Bases from Templated Websites

Ricardo Usbeck[1 2], Lorenz Bühmann[1], Axel-Cyrille Ngonga Ngomo[1], Muhammad Saleem [1], Andreas Both [2], Valter Crescenzi[3], Paolo Merialdo[3], Disheng Qiu[3]

[1]University of Leipzig, Germany

[2]R & D, Unister GmbH, Germany

[3]Università Roma Tre

October 21, 2014

Introduction

- Three domains: `imdb.com`, `goodreads.com` and `espnfc.com`.
- Crawled 10000 pages per domain.
- Projekt page: `rex.aksw.org`

Introduction

- Three domains: `imdb.com`, `goodreads.com` and `espnfc.com`.
- Crawled 10000 pages per domain.
- Projekt page: `rex.aksw.org`

`dbr:The_Lion_King`
`dbo:director`
`dbr:Roger_Allers,`
`dbr:Rob_Minkoff.`

Introduction

- Three domains: `imdb.com`, `goodreads.com` and `espnfc.com`.
- Crawled 10000 pages per domain.
- Projekt page: `rex.aksw.org`

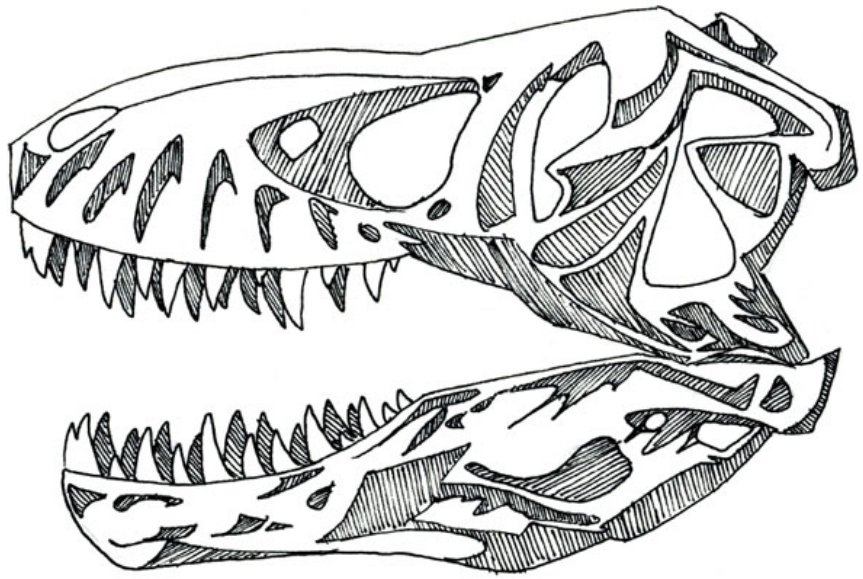
`dbr:The_Lion_King`
`dbo:director`
`dbr:Roger_Allers`,
`dbr:Rob_Minkoff`.

Not in DBpedia but on `imdb.com`

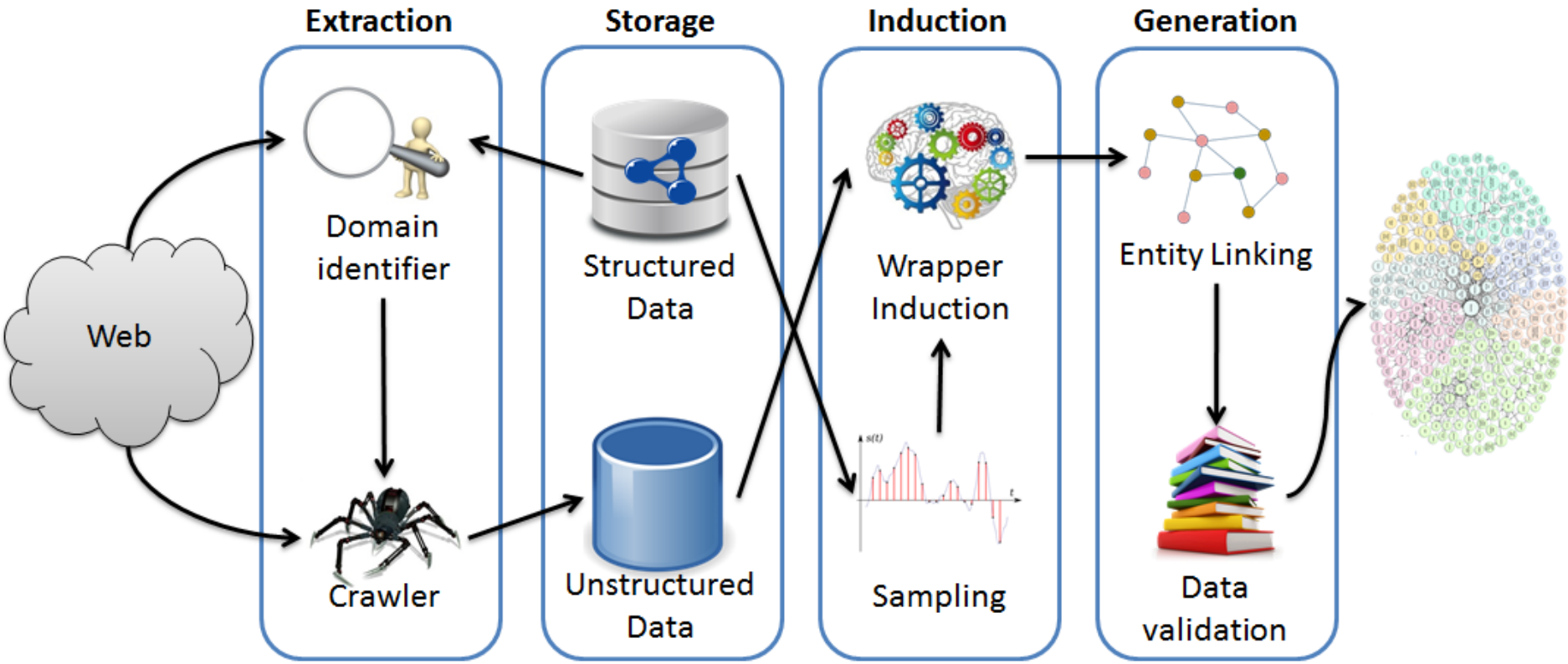
Introduction

REX Features

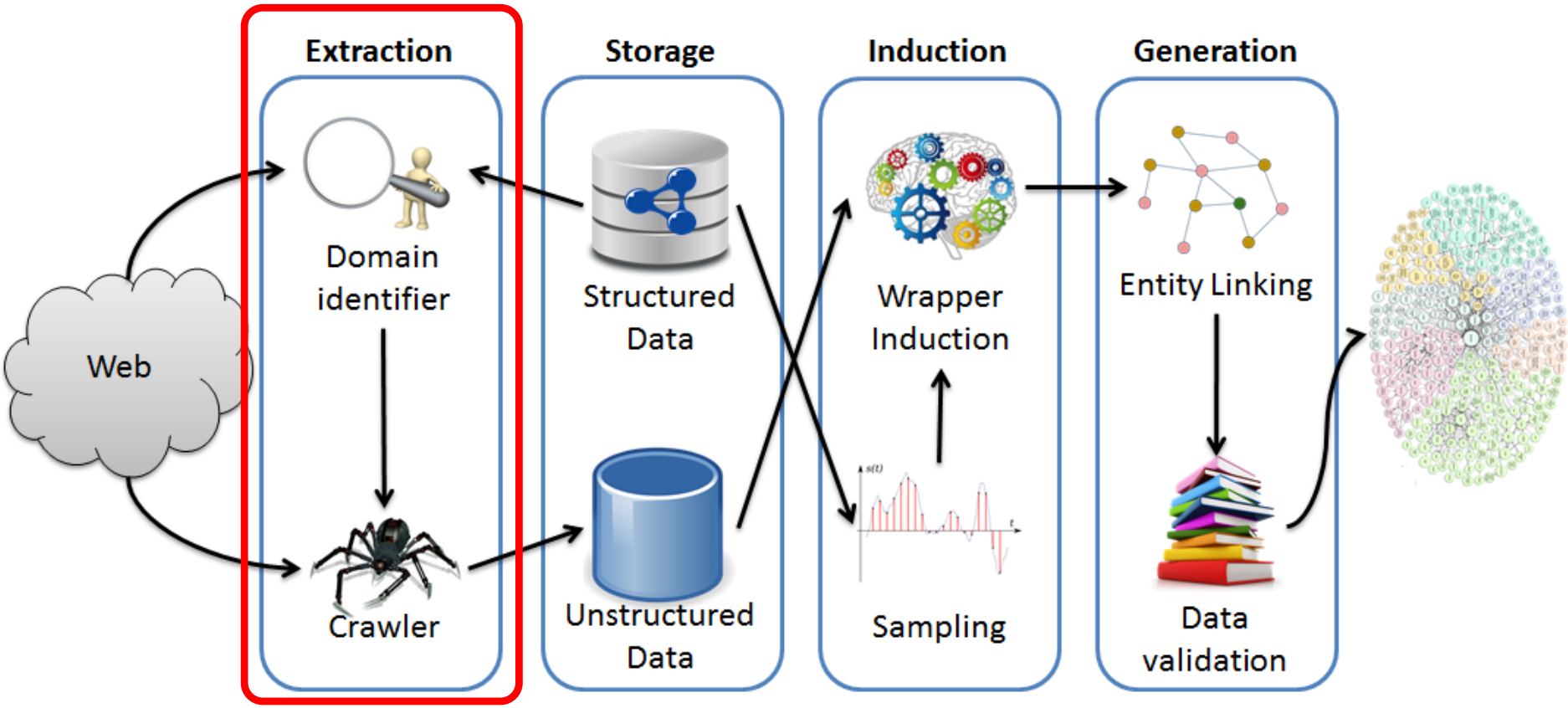
1. *Extensibility*
2. *Use of standards*
3. *Modularity*
4. *Scalability*
5. *Low costs*
6. *Accuracy*
7. *Consistency*



REX Framework

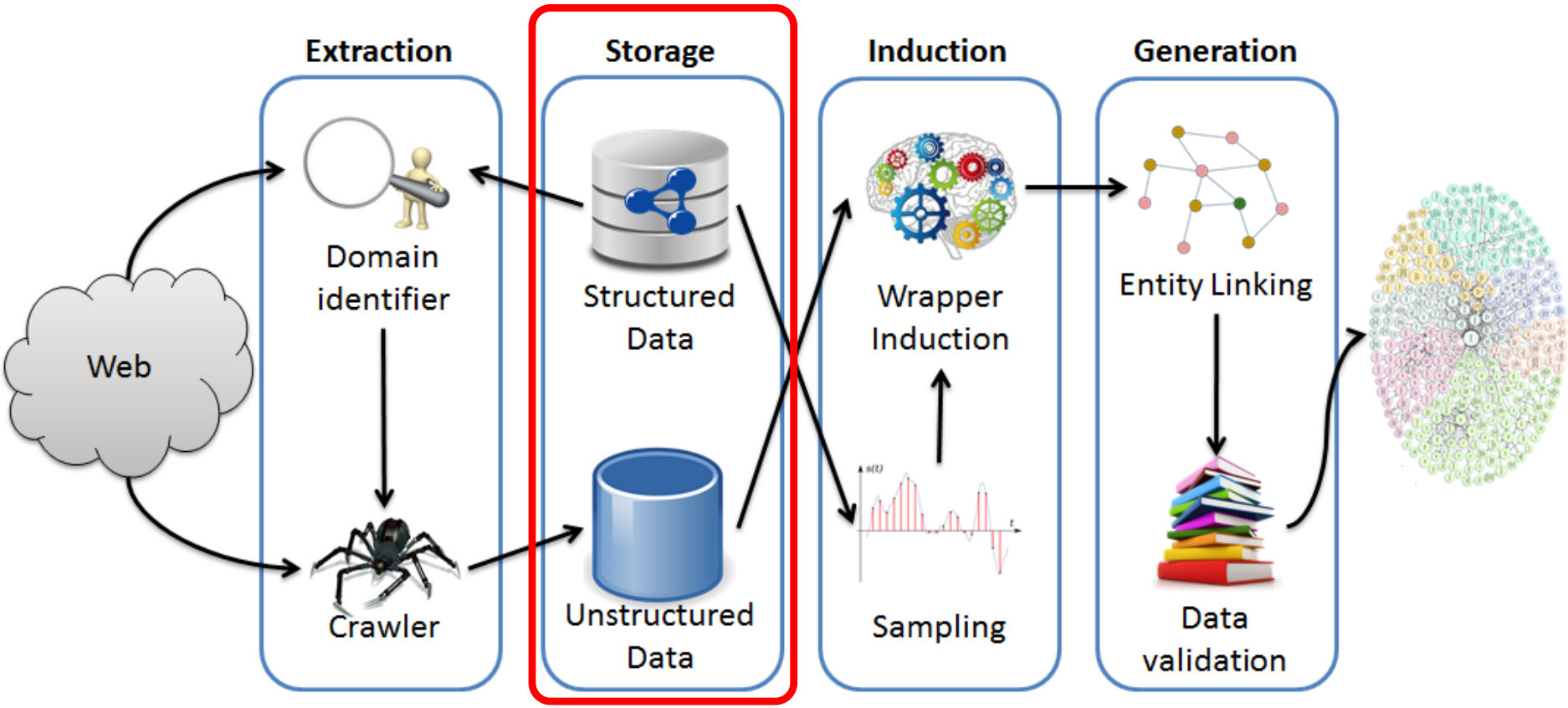


Extraction Layer



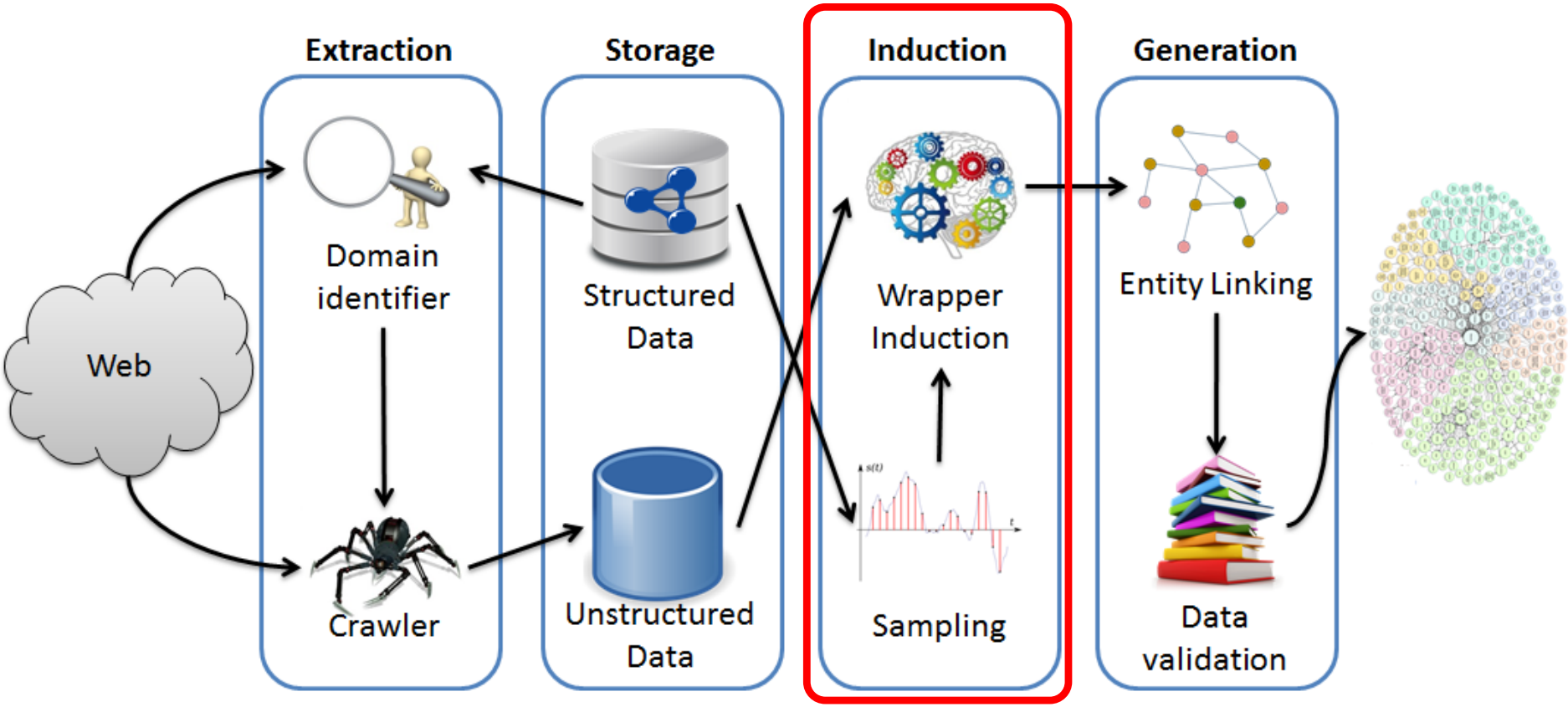
- Set of triples $(s, p, o) \xrightarrow{Google}$ websites with labels of s, p and o
- Top n domains
- Crawl web pages of the detected domain

Storage Layer



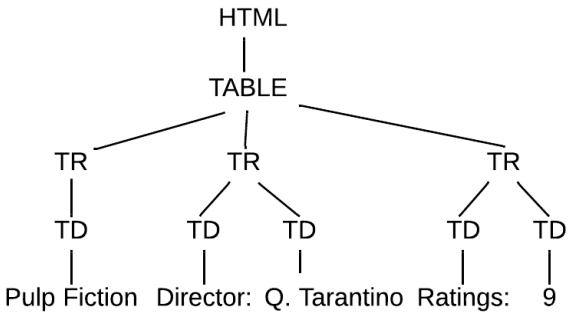
- Structured data via SPARQL
- Unstructured data from Lucene index
- Java APIs allow for exchange of modules

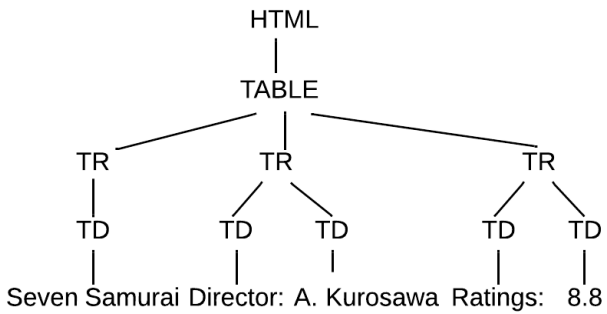
Induction Layer

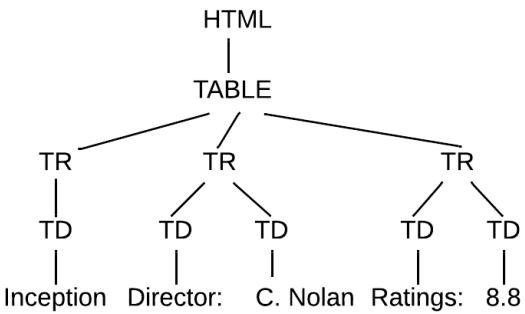


- Sampling over RDF allows to generate examples
- Generate XPaths for sample (s, o) for target p
- Maximize precision and recall via Machine Learning

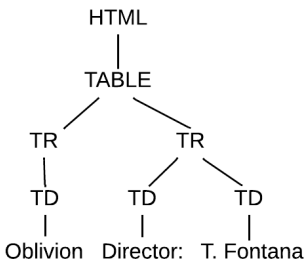
XPath Learning







(a)



(b)

Extraction rules

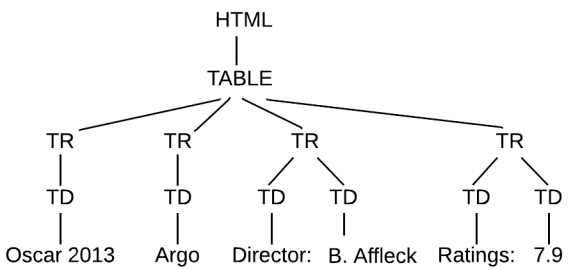
r^1: //*[contains(.,"Ratings:")]/../p-s::tr[2]/td/text()

r^2: //*[contains(.,"Director:")]/../p-s::tr[1]/td/text()

r^3: /html/table/tr[1]/td/text()

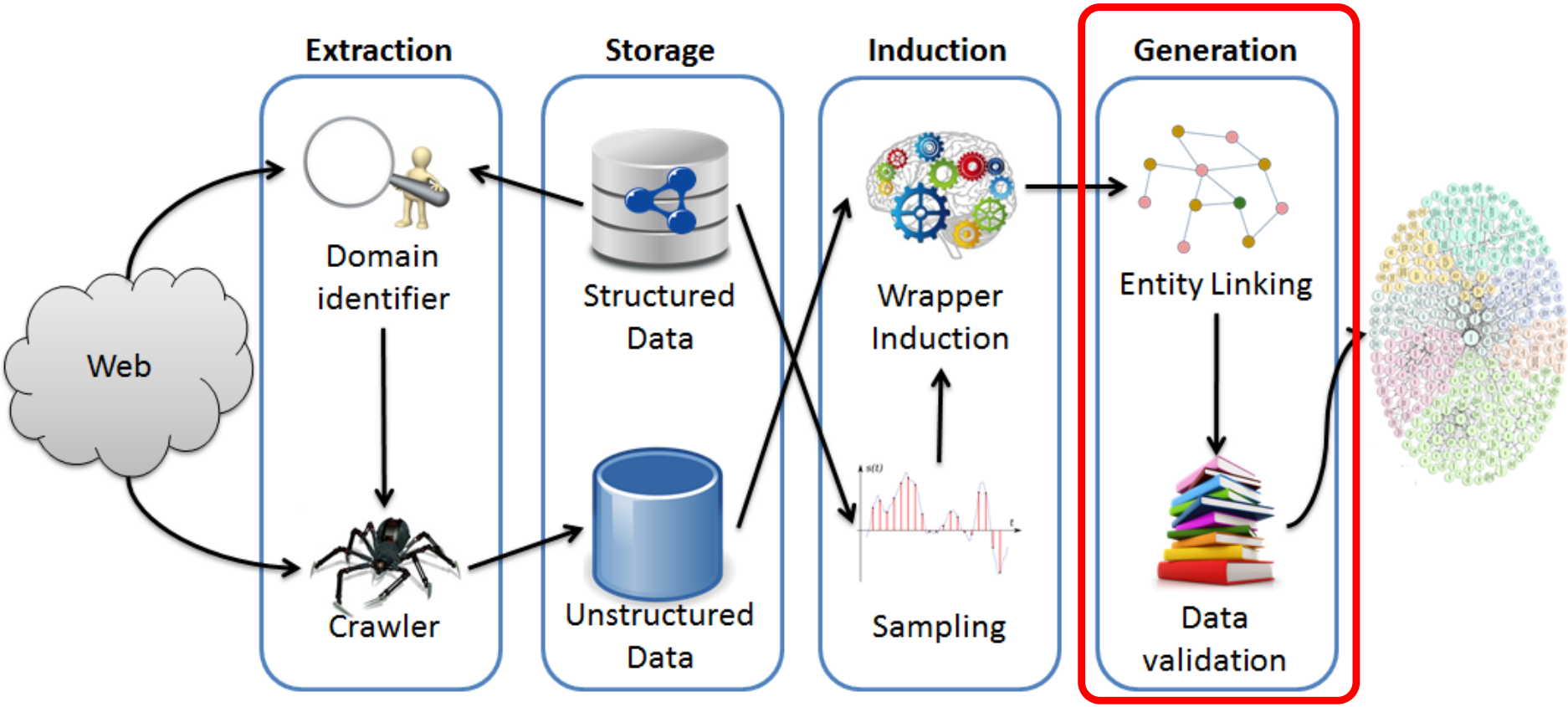
ps = preceding-siblings

(c)



(d)

Generation Layer



- URI disambiguation by using *AGDISTIS* framework
- Consistency check assumes a consistent knowledge base
- Incomplete but sound rule-based approach to ensure maximal consistency

Evaluation - Setup

- Three domains: `imdb.com`, `goodreads.com` and `espnfc.com`.
- Sample 100 pages per domain.

Evaluation - Results I

Property	#Possible triples	#Triples by AlfREX	#Consistent triples	#Correct triples	#New triples
dbo:author^{-1}	54	32	32	22	22
dbo:author	83	83	69	54	54
dbo:team^{-1}	2	1	1	0	0
dbo:team	30	55	42	19	13
dbo:starring^{-1}	40	99	83	35	34
dbo:starring	70	70	44	33	32
dbo:director	61	56	52	41	41

Evaluation - Results I

Property	#Possible triples	#Triples by AlfREX	#Consistent triples	#Correct triples	#New triples
dbo:author^{-1}	54	32	32	22	22
dbo:author	83	83	69	54	54
dbo:team^{-1}	**2**	**1**	**1**	**0**	**0**
dbo:team	30	55	42	19	13
dbo:starring^{-1}	40	99	83	35	34
dbo:starring	70	70	44	33	32
dbo:director	61	56	52	41	41

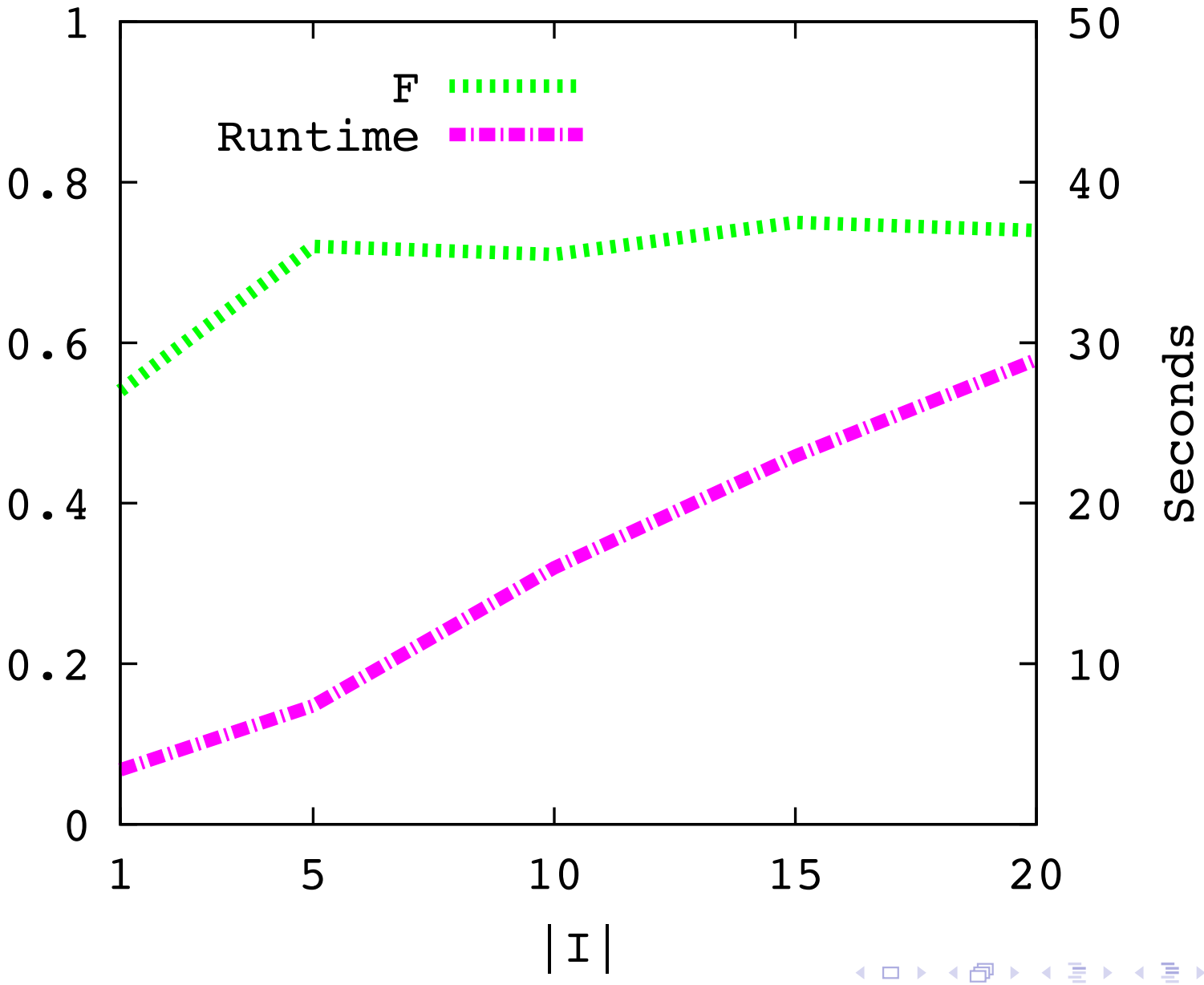
Evaluation - Results I

Property	#Possible triples	#Triples by AlfREX	#Consistent triples	#Correct triples	#New triples
dbo:author^{-1}	54	32	32	22	22
dbo:author	83	83	69	54	54
dbo:team^{-1}	2	1	1	0	0
dbo:team	30	55	42	19	13
dbo:starring^{-1}	**40**	**99**	**83**	**35**	**34**
dbo:starring	70	70	44	33	32
dbo:director	61	56	52	41	41

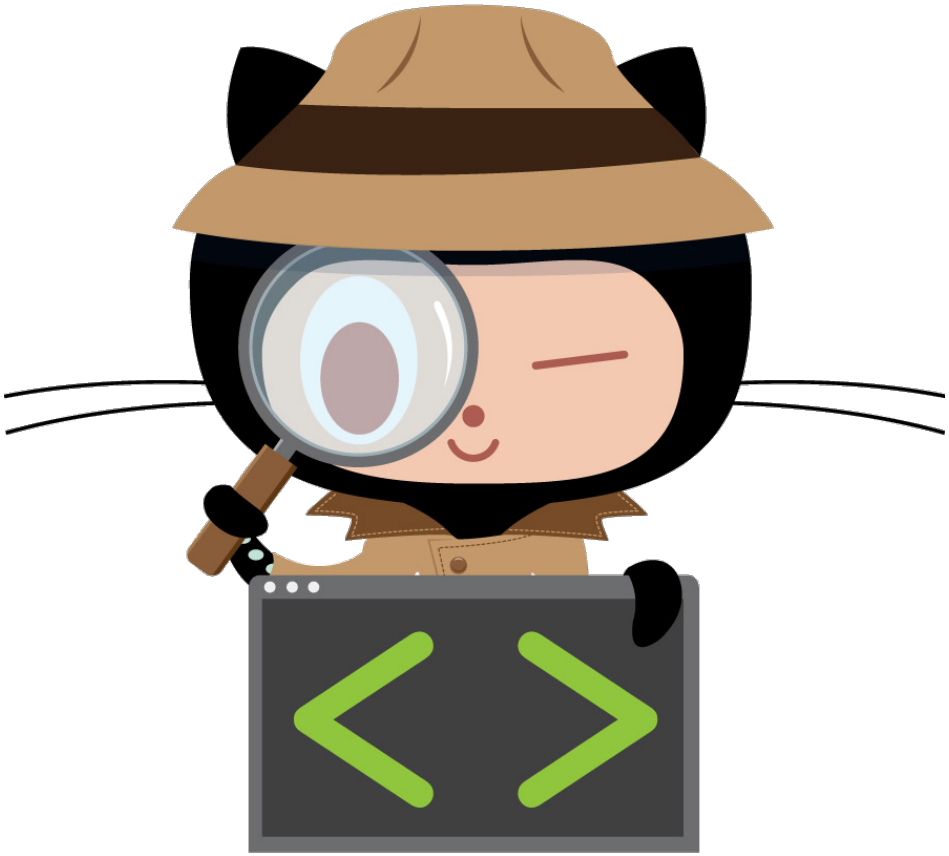
Evaluation - Results I

Property	#Possible triples	#Triples by AlfREX	#Consistent triples	#Correct triples	#New triples
dbo:author^{-1}	54	32	32	22	**22**
dbo:author	83	83	69	54	**54**
dbo:team^{-1}	2	1	1	0	**0**
dbo:team	30	55	42	19	**13**
dbo:starring^{-1}	40	99	83	35	**34**
dbo:starring	70	70	44	33	**32**
dbo:director	61	56	52	41	**41**

Evaluation - Results II



Programmatic Usage of REX



http://vikki.github.io/presentations/jsday.it/images/inspectocat.png

1. Usage as dependency via AKSW Maven Build system
2. Download from Git and install via Maven to be independent of third-party repositories
3. Download from Git and run it from commandline

Conclusion

- REX is available as open source Java project
- LOD Cloud as source for training data to decrease costs
- Improve REX by implementing new modules
- Future work will be about:
  - populate the Web of Data using Web pages

That's all Folks!

Thank you!

Ricardo Usbeck
University of Leipzig
AKSW Research Group
usbeck@informatik.uni-leipzig.de

Code: `http://github.com/AKSW/REX`
Wiki: `http://rex.aksw.org`