Scalability in Semantic Labeling Approaches

Ahmad Alobaid, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Oscar Corcho, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Axel Ngonga-Ngomo, DICE Group, University of Paderborn,Germany

aalobaid@fi.upm.es
@oeg_upm

9-10-2019
UPM, Montegancedo

What is it?

> the ability of a system to **accommodate** an **increasing number** of elements or objects, to process **growing** volumes of work **gracefully**, and/or to be susceptible to enlargement. (Bondi, 2000)

Types of scalability:

Load scalability => Parallelism

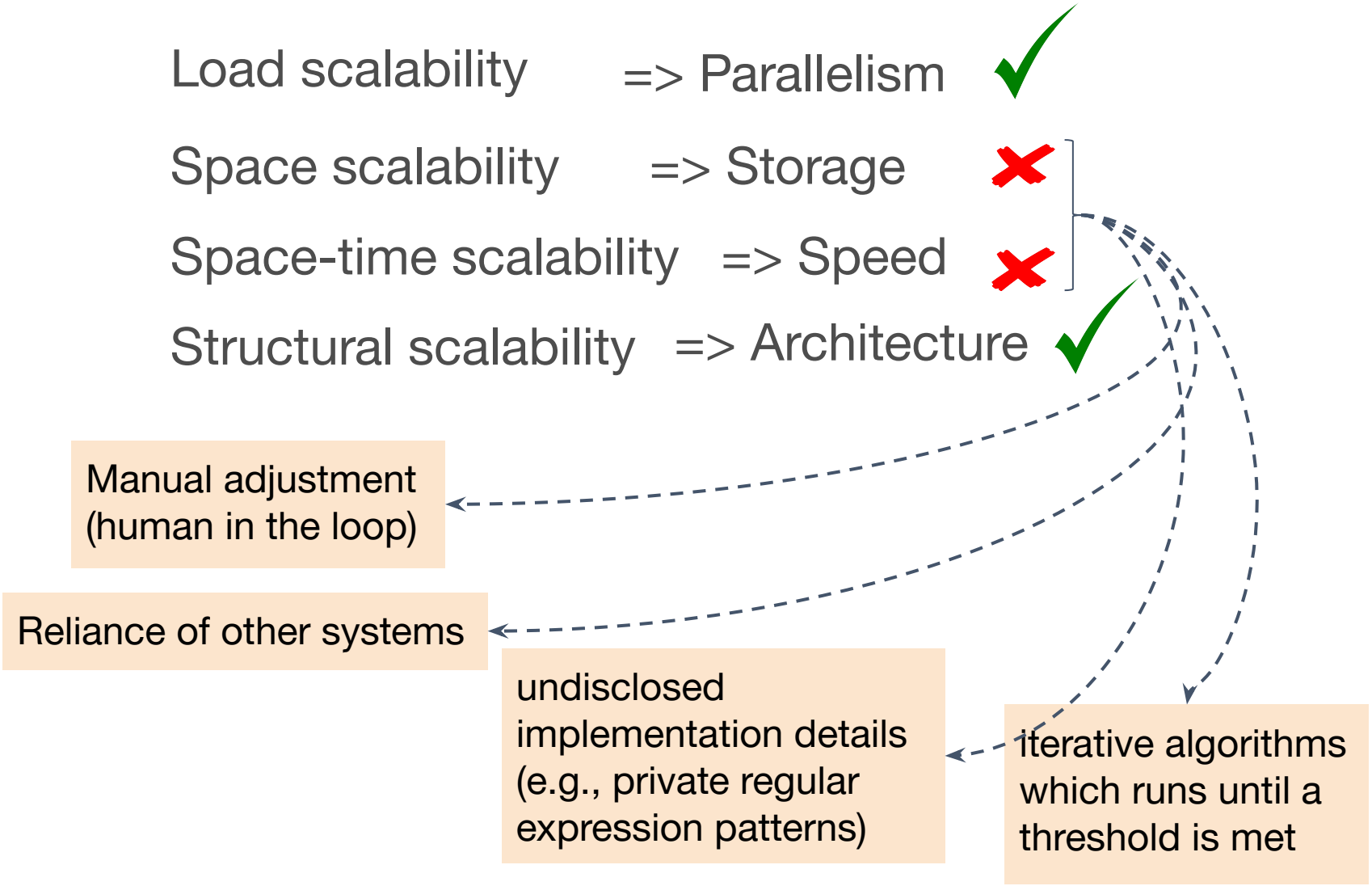
Space scalability => Storage

Space-time scalability => Speed

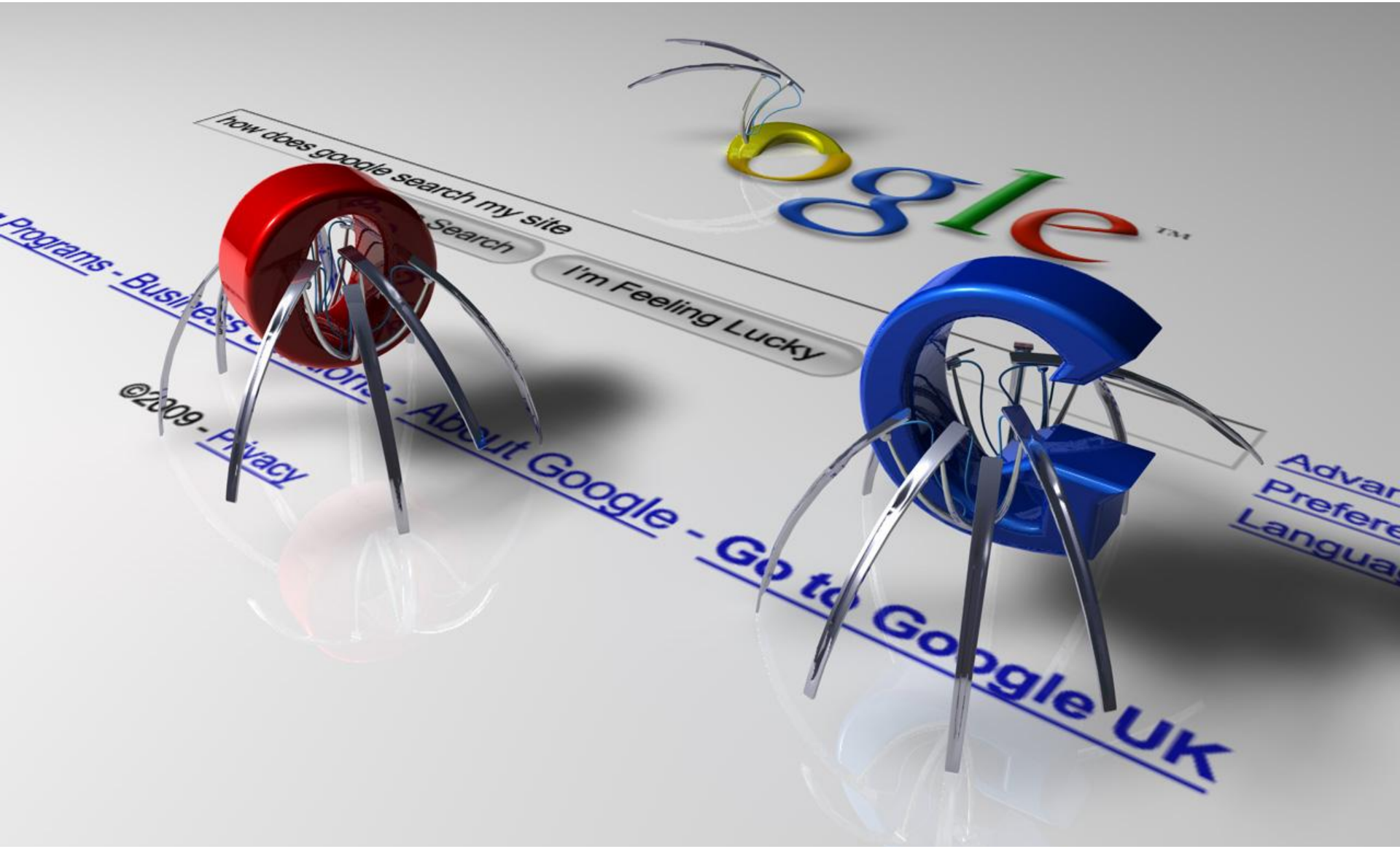
Structural scalability => Architecture

André B Bondi. 2000. Characteristics of scalability and their impact on performance. In Proceedings of the 2nd international workshop on Software and performance. ACM, 195–203.

Scalability Inspection



Web Crawl



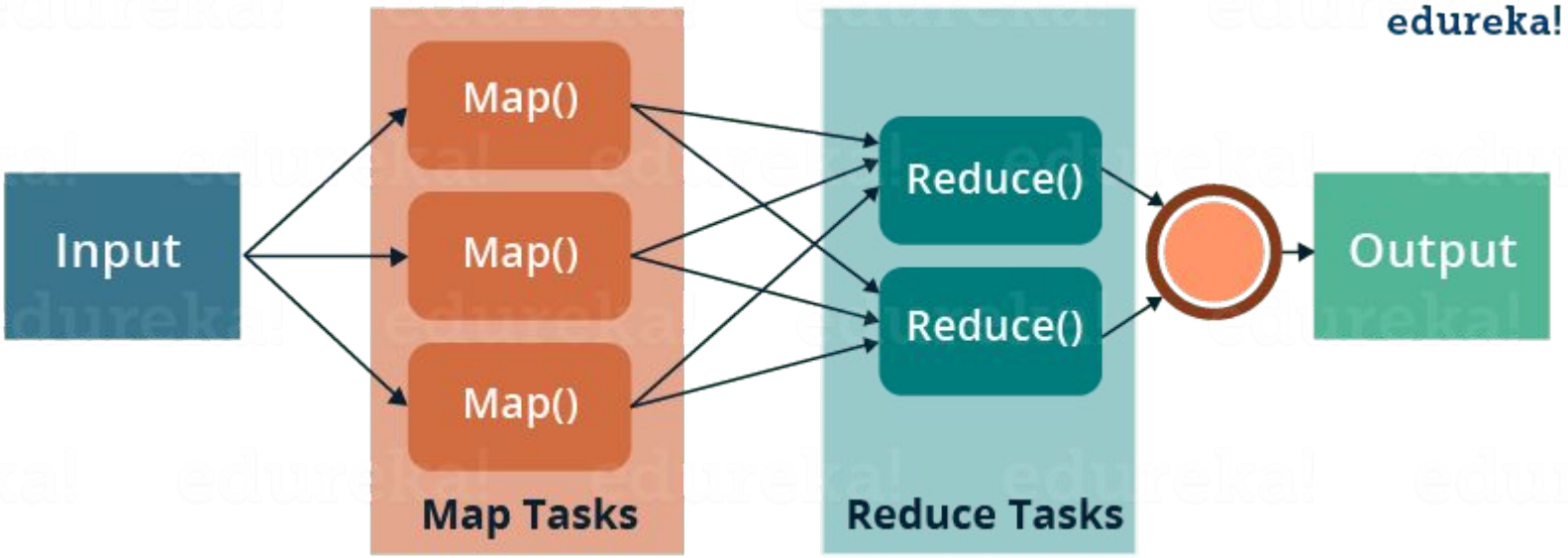
datadeck.com

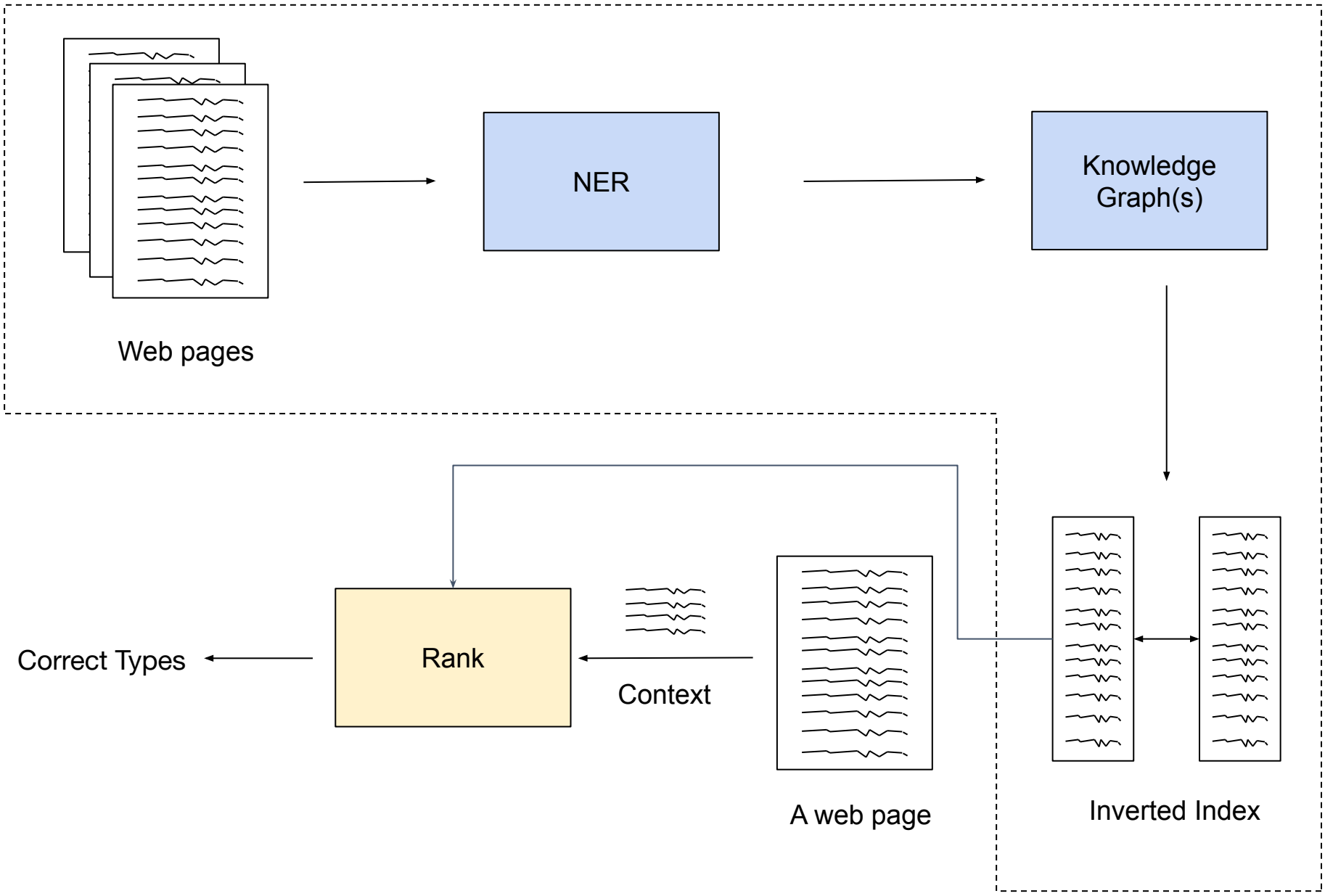

RED-BLACK
INSIDE-OUT UMBRELLA
Gracious Treasures Presents

On board example on Inverted Index

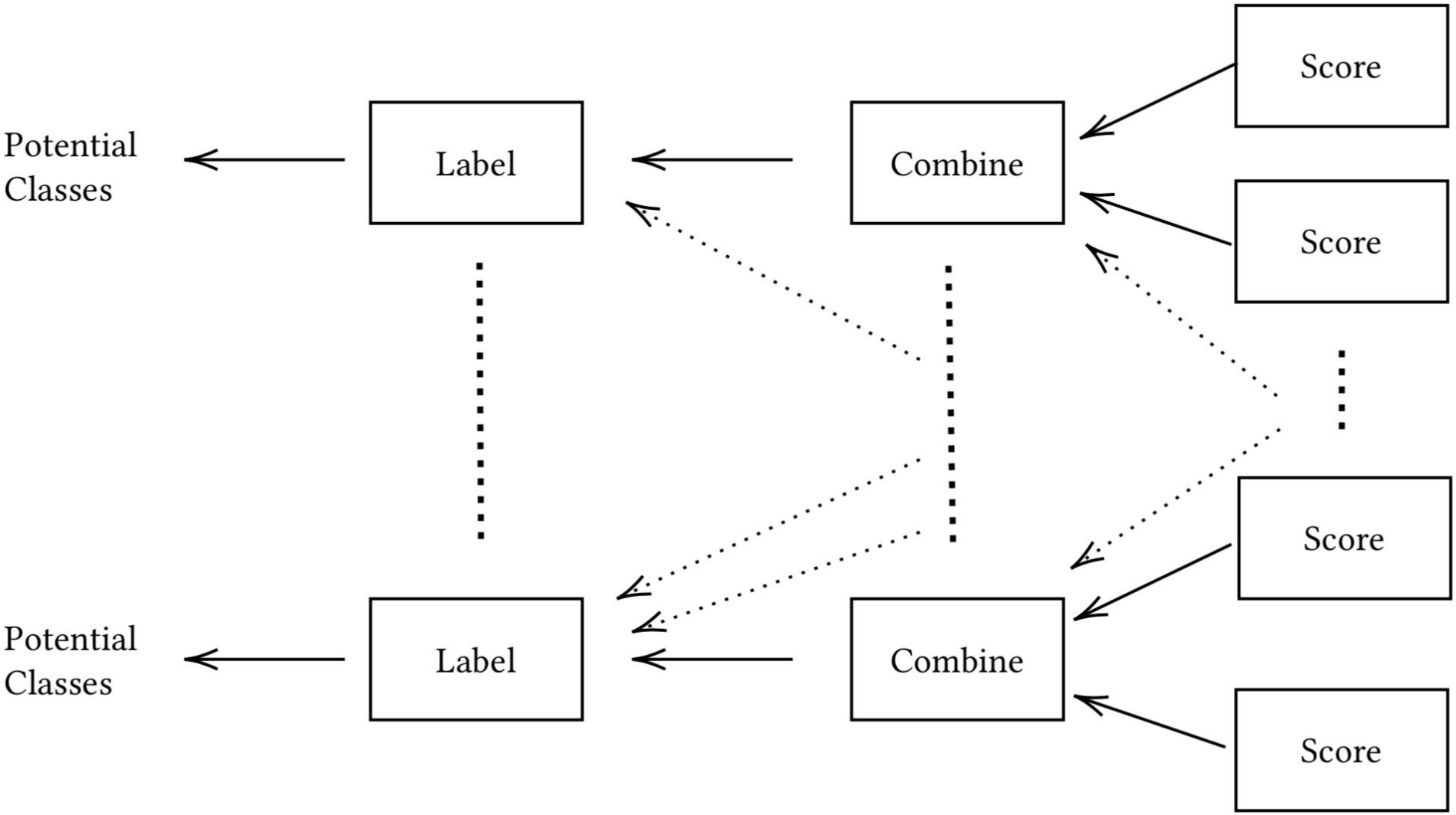
https://github.com/ahmad88me/random-picker

On board example of MapReduce

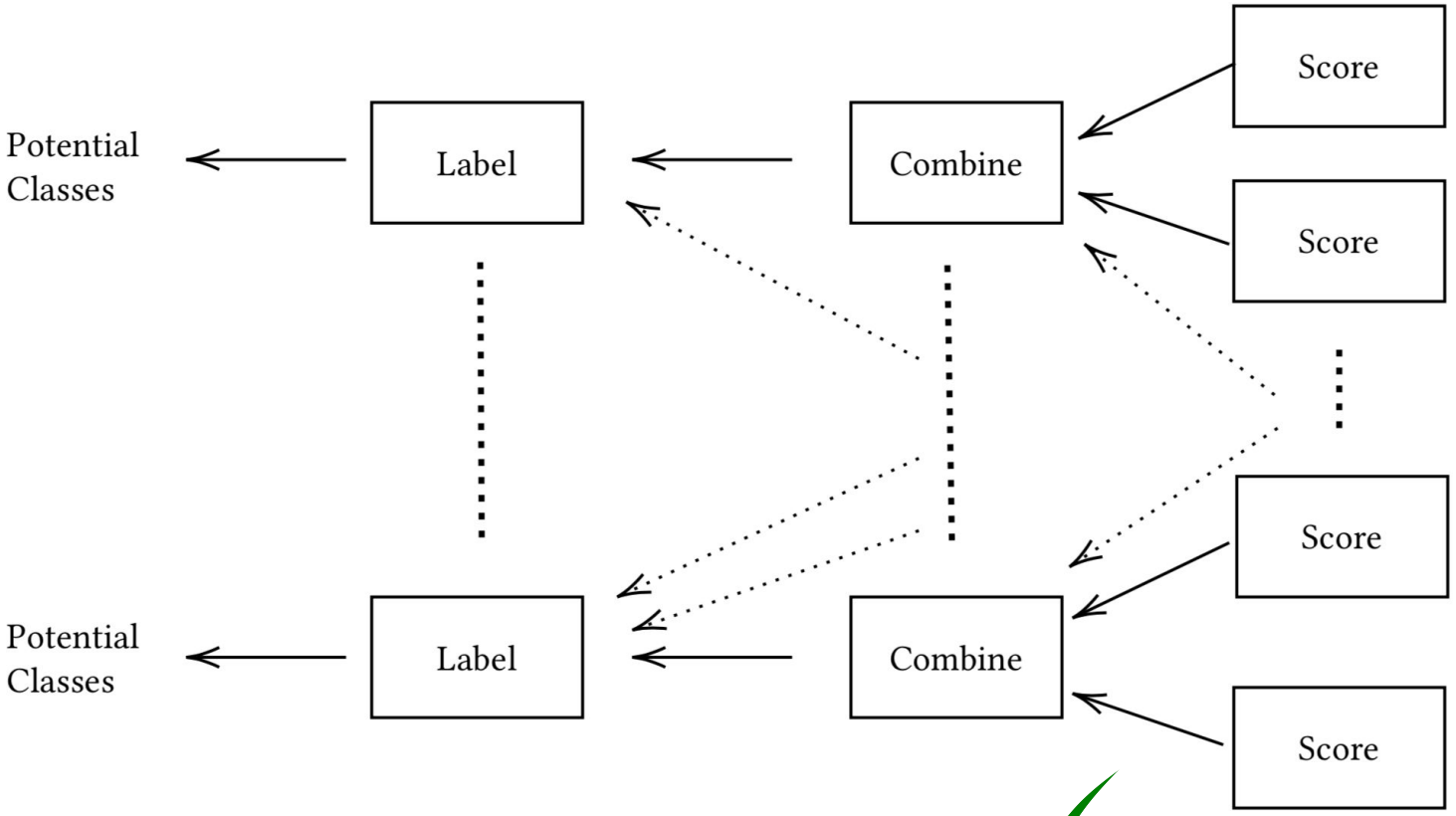



Web pages
NER
Knowledge Graph(s)
Inverted Index
A web page
Context
Rank
Correct Types

Our approach



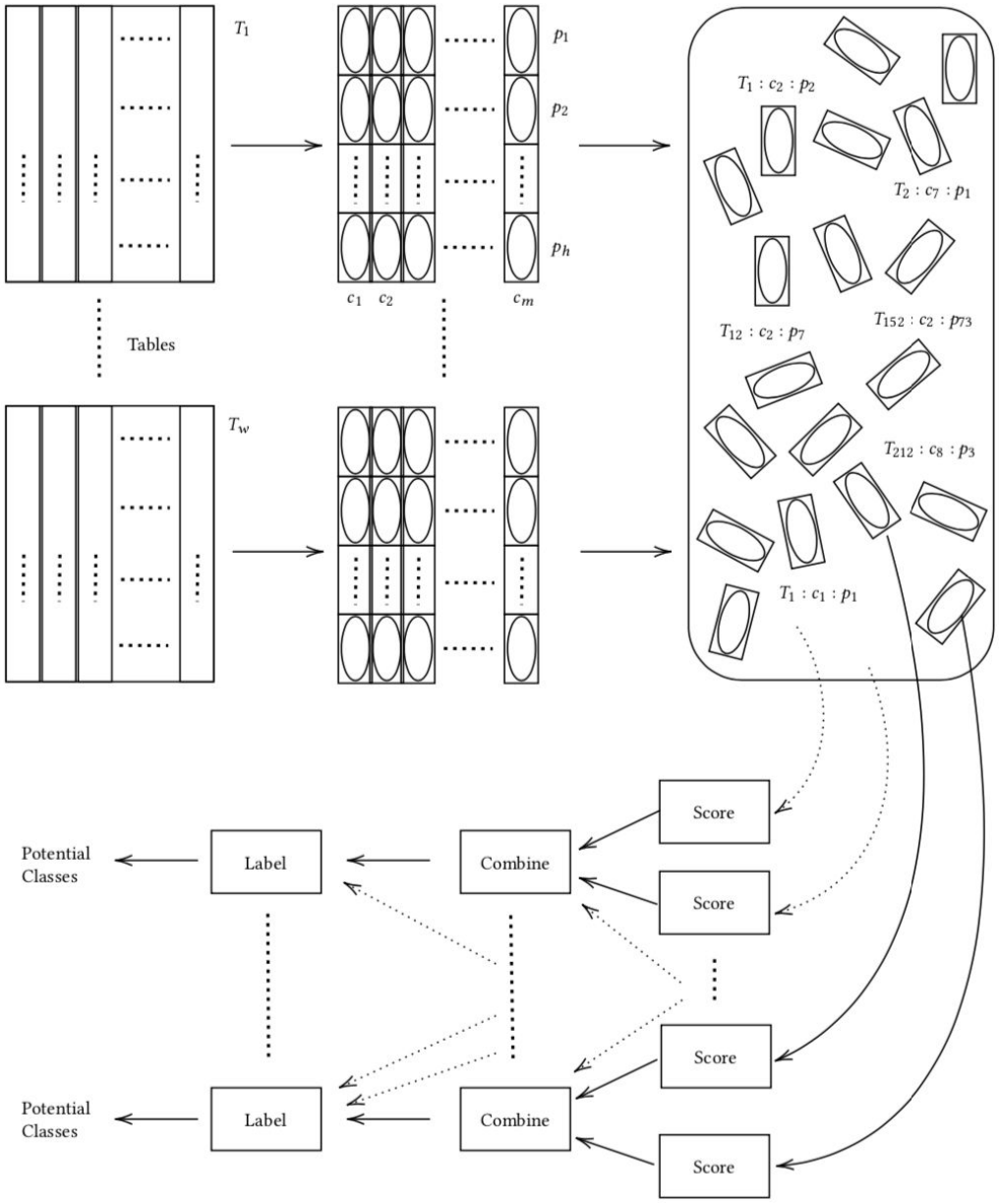
Our approach



Does it run in parallel? Load scalability ✓

Does the architecture handle data growth? Structural scalability ✓

On board example



Compare with others?

Approach	Data (to be annotated)	Published	Accessible	Ease of use
Cafarella et al. [5] (WEBTABLES)	Google Crawl	No	No	?
Limaye et al. [15]	Their Web Crawl	No	No	?
Syed et al. [26]	Tables from Google Square	No ¶	No	?
Venetis et al. [29]	Their Web Crawl	No	No	?
Goel et al. [10]	Their Web Crawl	No	No	?
Zhang et al. [30] (InfoGather+)	Bing Crawl	No	No	?
Nuzzolese et al. [18]	-	-	-	-
Tonon et al. [28] (TRank)	Their Web Crawl	Yes	No †	?
Quercini et al. [20]	Extract from Google Fusion Table	No ¶	No	?
Deng et al. [8]	WWT + WEX	Yes	No †	?
Zhang [32] (TableMiner)	Tables from Wikipedia and IMDB	No	No	?
Ritze et al. [24] (T2K)	Web Data Commons (T2D) *	Yes	Yes	Yes
Ramnandan et al. [22] (SemanticTyper)	Hand picked tables	Yes	Yes	No
Pham et al. [19] (DSL)	Hand picked tables + T2D *	Yes	Yes	No ✠
Neumaier et al. [17]	Extract from DBpedia	Yes	No †	?
Taheriyan et al. [27] (Karma)	US Art Museums	Yes	Yes	No
Ermilov et al. [9]	DBD + T2DT §	Yes	Yes	No
Zhang [31] (TableMiner+)	Tables from Wikipedia + IMDB + MusicBrainz	No	Yes‡	?
Ritze et al. [23]	T2Dv2 *	Yes	Yes	Yes
Alobaid et al. [2]	T2Dv2 *	Yes	Yes	Yes

* These are Gold standards published by Ritze et al. [23, 24].

§ T2DT is a subset taken by Ermilov et al. from the Data Web Commons published by Ritze et al.

✠ We are not referring to the T2D dataset here (which is already described). We are referring to the other datasets that are made available in GitHub as pointed out in the paper.

¶ Google Square and Google Fusion Tables were public, but they are no longer accessible during the time of writing.

† Invalid URLs.

‡ Although the data are not published in the paper, the author actually mention in a footnote of the paper that he can provide the data if asked by an email because the data is large to host.

- Data are not reported

? Unknown

Demo

Other searchengine approaches

What about Google Search?

What about Bing Search?

Questions?

1. Do search engines use sematic web?
2. Are there other techniques for scalable semantic labeling?
3. Why scalability? can we rely on few powerful machines?

Conclusion

Examples of not scalable approaches

Karma

Bing

Input: Knowledge base

Task: Entity type ranking

given entity e in a document d

Te: the types of the entity e

Ce: Context (surrounding text of an entity e) in a document d containing an entity e

Text -> NER -> Inverted Index. Then given a context and an entity, the system will suggest a type

1- Web Crawl

2- NER

3- Inverted Index

4- Given a context and an entity, it suggests a type