Knowledge Engineering and Semantic Web 2015

Identifying Web Tables

Supporting a Neglected Type of Content on the Web

Michael Galkin, Dmitry Mouromtsev, Sören Auer

Outline

- Motivation
- Neglected?
- Pipeline
- Machine Learning
- Evaluation
- Applications

Motivation

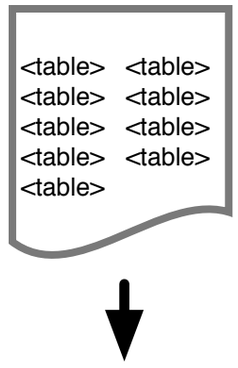
http://commoncrawl.org/ http://stats.lod2.eu/

Motivation

the **DATA**

http://commoncrawl.org/ http://stats.lod2.eu/

Motivation

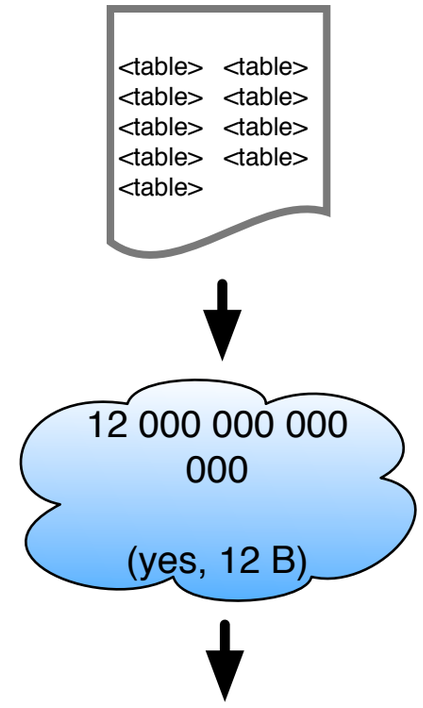


the **DATA**

- 9 tables on an average HTML page

http://commoncrawl.org/ http://stats.lod2.eu/

Motivation



the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted

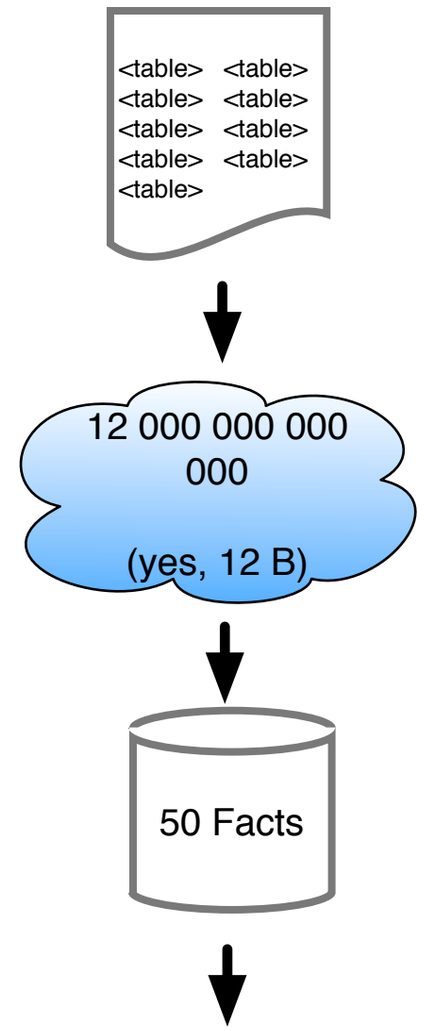
http://commoncrawl.org/ http://stats.lod2.eu/

EIS Fraunhofer IAIS universitätbonn

Motivation

the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts



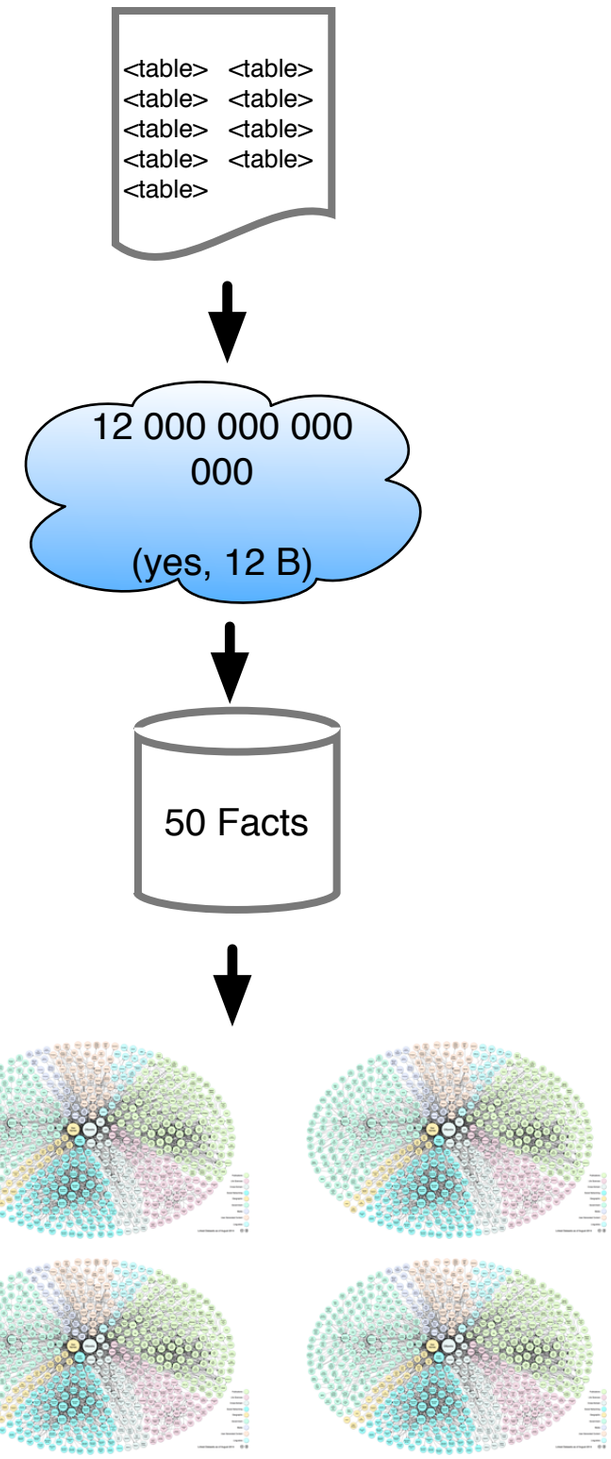
http://commoncrawl.org/ http://stats.lod2.eu/

EIS Fraunhofer IAIS universitätbonn

Motivation

the **DATA**

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts
- 600 B facts



http://commoncrawl.org/ http://stats.lod2.eu/

Motivation

Tables are a natural way how people interact with structured data (for humans)

	Thing1	Thing2	Thing3
Param1	val11	val21	val31
Param2	val12	val22	val32
Param3	val13	val23	val33

Motivation

Tables are a natural way how people interact with structured data (for humans)

	Thing1	Thing2	Thing3
Param1	val11	val21	val31
Param2	val12	val22	val32
Param3	val13	val23	val33

```
<style type="text/css">
.tg  {border-collapse:collapse;border-spacing:0;}
.tg td{font-family:Arial, sans-serif;font-size:14px;padding:10px
5px;border-style:solid;border-width:1px;overflow:hidden;word-break:normal;}
.tg th{font-family:Arial, sans-serif;font-size:14px;font-weight:normal;padding:10px
5px;border-style:solid;border-width:1px;overflow:hidden;word-break:normal;}
.tg .tg-s6z2{text-align:center}
</style><table class="tg" style="undefined;table-layout: fixed; width: 208px">
<colgroup><col style="width: 41px"><col style="width: 55px"><col style="width: 51px">
<col style="width: 61px"></colgroup><tr><th class="tg-031e"></th><th class="tg-s6z2">P1</th>
<th class="tg-s6z2">P2</th><th class="tg-s6z2">P3</th></tr><tr><td class="tg-031e">Obj1</td>
<td class="tg-s6z2">a1</td><td class="tg-s6z2">b2</td><td class="tg-s6z2">c3</td></tr><tr>
<td class="tg-031e">Obj2</td><td class="tg-s6z2">d1</td><td class="tg-s6z2">e2</td>
<td class="tg-s6z2">f3</td></tr><tr><td class="tg-031e">Obj3</td><td class="tg-s6z2">g1</td>
<td class="tg-s6z2">h2</td><td class="tg-s6z2">i3</td></tr><tr><td class="tg-031e">Obj4</td>
<td class="tg-s6z2">j1</td><td class="tg-s6z2">k2</td><td class="tg-s6z2">l3</td></tr><tr>
<td class="tg-031e">Obj5</td><td class="tg-s6z2">m1</td><td class="tg-s6z2">op</td>
<td class="tg-s6z2">xyz</td></tr></table>
```

Neglected?

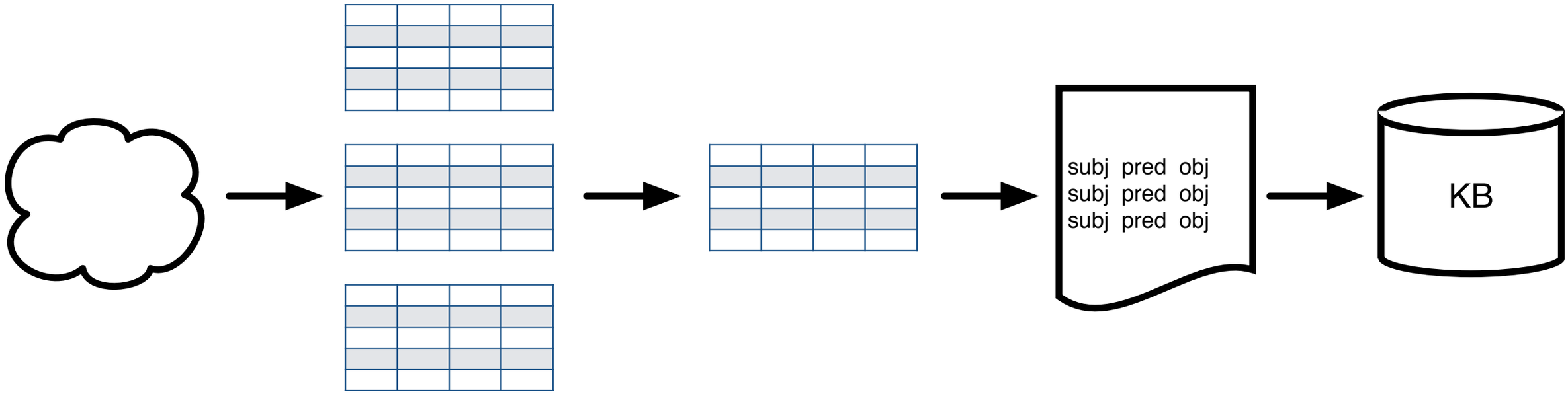
- Processing - search engines index anything except tables
- Annotation - web tables annotation tools are on the edge of extinction
- Retrieval - exists for pictures, audio, video data

Goals

Bring the tables knowledge to the LOD Cloud

- Search & Ingestion
- Explore characteristics of a table
- Extract knowledge & map with the existing knowledge graph

Pipeline



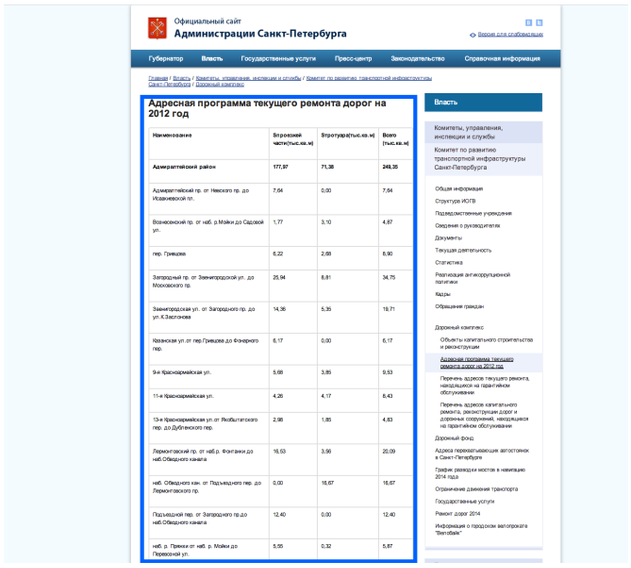
Search for tables | Extract | Genuineness assessment
Orientation assessment | Generate an ontology | Enrich the KB

EIS Fraunhofer IAIS universitätbonn

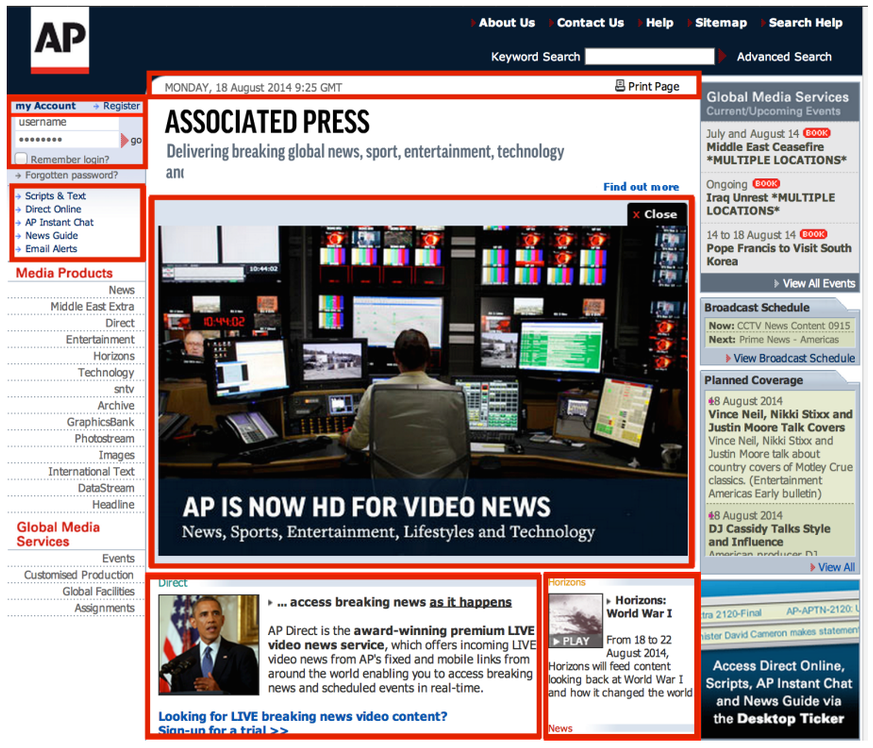
Machine Learning

Did we find a data table or not?

- Genuineness assessment
  a table contains some data, not html formatting



both are represented via <table>



Machine Learning

Which data table did we find?

- Orientation assessment
  it is essential to distinguish concepts from their properties

	Header 1	Header 2	Header 3
Obj1			
Obj2			
Obj3			

horizontal orientation

	Obj1	Obj2	Obj3
Header 1			
Header 2			
Header 3			

vertical orientation

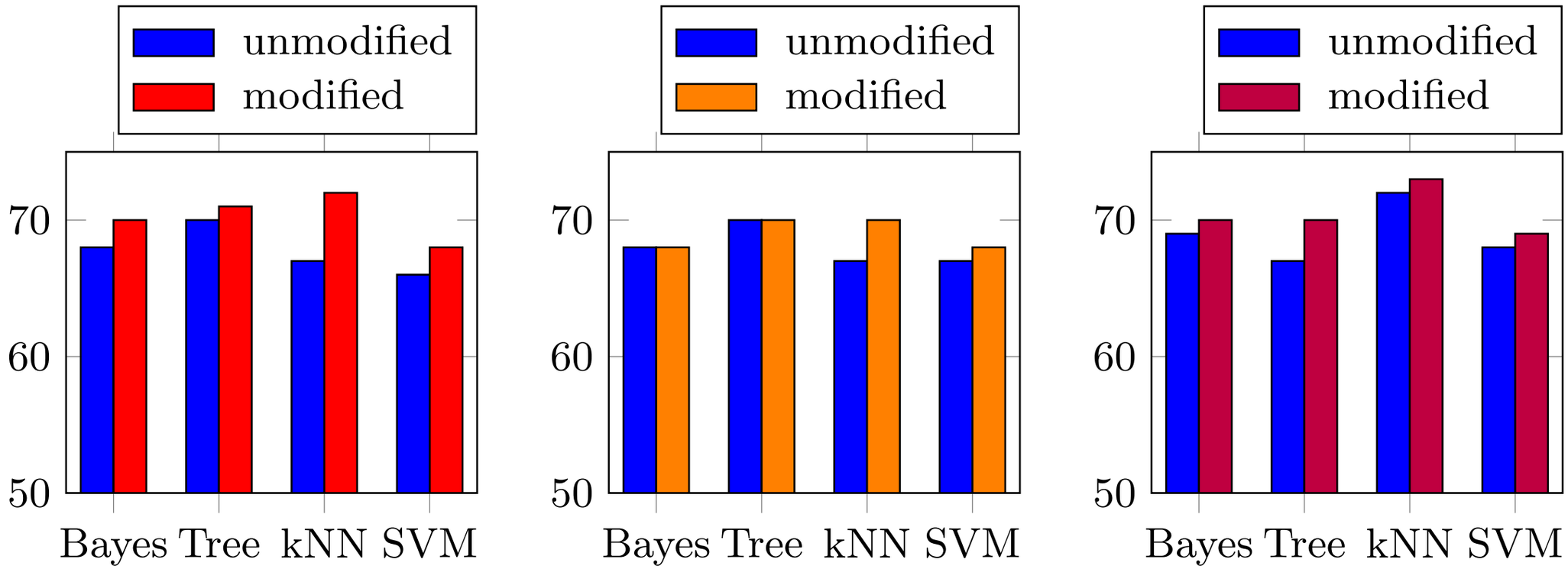
Machine Learning

String Metric → **Heuristics** → **ML**

String Metric	Heuristics	ML
Levenshtein	Horizontal cell similarity	Naive Bayes
Jaro-Winkler		kNN
n-grams	Vertical cell similarity	J48
	Max/avg	SVM

Improvements
1. All numbers are the same
2. Fixed similarity for long sentences

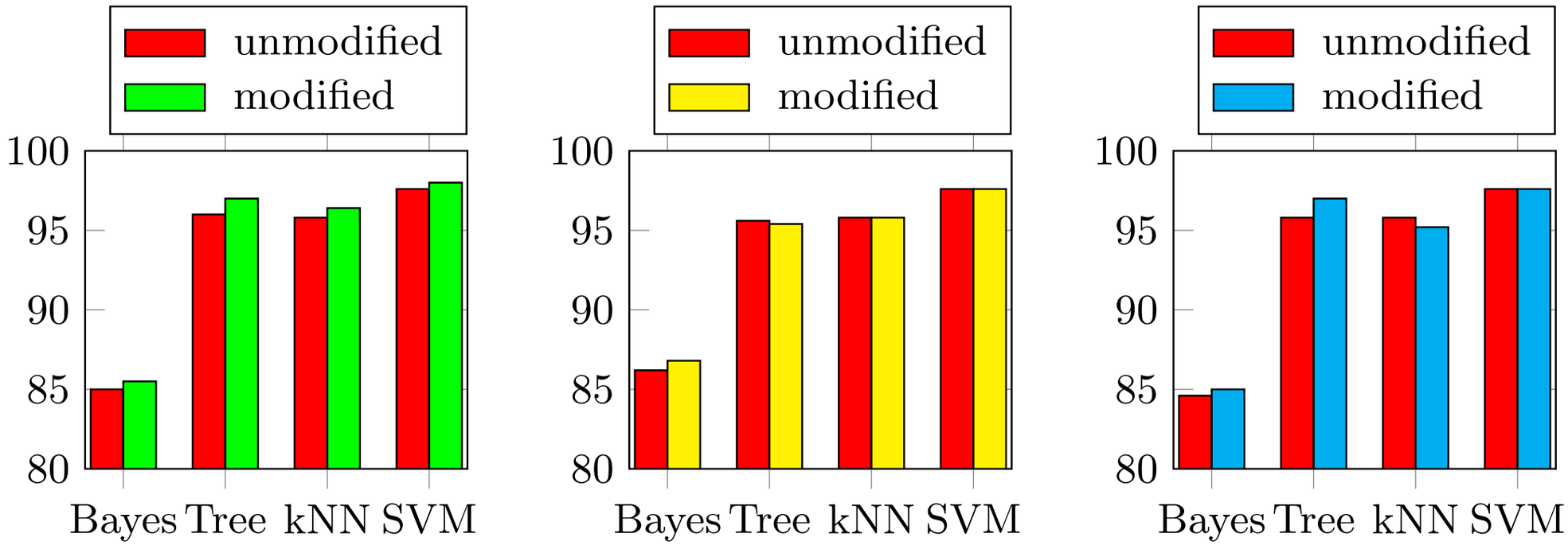
Evaluation



(a) Levenshtein (b) Jaro-Winkler (c) n-grams

Genuineness assessment, F-Measure

Evaluation



(a) Levenshtein (b) Jaro-Winkler (c) n-grams

Orientation assessment, Precision

Applications

- Enrich enterprise or web Knowledge Graphs with billions of facts from the most neglected content type
- Business domain - refine openly published tables with financial KPIs for the data analysis;
  Maintain connection with the Enterprise Knowledge Graph;
  Use semantic technologies for business processes re-engineering.