Web Intelligenge Network Conference
From Web to Data

Identifying Official Firm Websites: A Comparison of Machine Learning-Based URL Retrieval Methods and AI-Powered Search Engines

Donato Summa

URL retrieval

All NSIs maintain extensive administrative information on a long list of national enterprises

unfortunately

the corresponding list of official website addresses is largely incomplete (at least in Italy).

URL retrieval

We need the official addresses (URLs) of enterprise websites to extract information from their content

but

manually retrieving official enterprise URLs is a very time-consuming operation

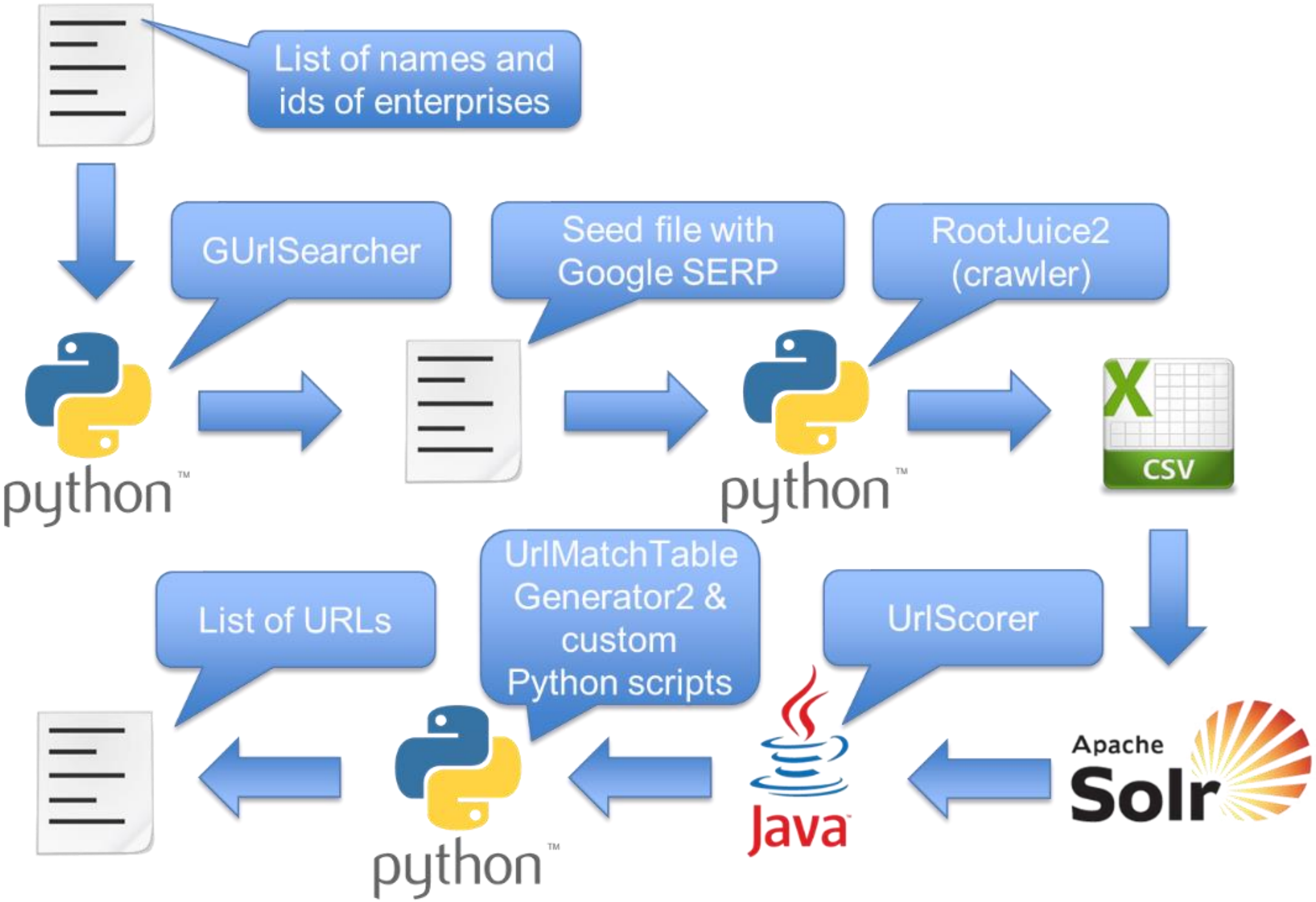
so

the idea is to retrieve them automatically!

URL retrieval

In the previous ESSnet Big Data 1 and Big Data 2 projects, among other things, we developed and improved URL retrieval systems at the national level.

Istat URL retrieval pipeline



List of names and ids of enterprises

GUrlSearcher

Seed file with Google SERP

RootJuice2 (crawler)

UrlMatchTable Generator2 & custom Python scripts

UrlScorer

List of URLs

OBEC annotation exercise

Goal: create an annotated dataset of enterprise-URL pairs

Annotation is used to assess the quality of data processing and retrieval pipelines related to, among other things, enterprise URLs (Does the enterprise have one or more website and what are they?)

For each country, a sample of 500 legal units was drawn from the 2024 ICT sampling population, stratified by:

- NACE section (first-level NACE code)
- enterprise size (10-49, 50-249, 250+ employees)

Additional rules:

- put NACE sections with less than 5% of the sampling population into 1 category
- minimum cell count of 10 (NACE section and enterprise size)

OBEC annotation exercise

Search term: [enterprise name] + [municipality]

The right URL is the dedicated corporate website (social media accounts, yellow pages or enterprise directories are NOT OK)

The website content must include the company's administrative information as recorded in the SBR.

In some cases, website information is contradictory, making it difficult to determine the correct website.

OBEC annotation exercise

To ensure URL correctness, company information was categorized into three tiers:

- **Strong**: VAT ID, tax ID
- **Medium**: trade register ID (with its 3 parts)
- **Weak**: name, address (street + number, zip code, municipality)

<table>
<tr><th>Matches</th><th colspan="2">Contradictions</th><th>Missing</th><th>Website correct?</th></tr>
<tr><td>1+ strong & 1+ weak</td><td colspan="2"></td><td>remaining</td><td>Yes</td></tr>
<tr><td>3 strong</td><td colspan="2"></td><td>remaining</td><td>Yes</td></tr>
<tr><td>1 strong & medium & 1+ weak</td><td colspan="3">remaining</td><td>Yes</td></tr>
<tr><td>2 strong</td><td colspan="3">remaining</td><td>Yes</td></tr>
<tr><td>3 weak</td><td></td><td colspan="2">remaining</td><td>Yes</td></tr>
<tr><td>3 weak</td><td>1+ strong</td><td colspan="2">remaining</td><td>No</td></tr>
<tr><td>1 strong</td><td>remaining</td><td colspan="2"></td><td>No</td></tr>
<tr><td>1+ strong & 2+ weak</td><td>1 strong or medium</td><td colspan="2">remaining</td><td>Unclear, investigation in SBR necessary</td></tr>
</table>

OBEC annotation exercise

Firm ID	Firm administrative info	Official URL (manually found)	Official URL (national procedure)
123	name, address, VAT, NACE, …	**abc.com**	abc.com
234	name, address, VAT, NACE, …		cab.it
345	name, address, VAT, NACE, …	**bcd.it**	aaa.it
456	name, address, VAT, NACE, …		
567	name, address, VAT, NACE, …	**cde.net**	
…	…	**…**	…

Manual URL retrieval (truth) vs:

Istat pipeline
for URL retrieval

you.com

ChatGPT

Gemini

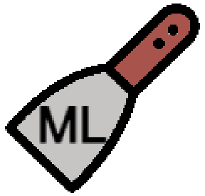
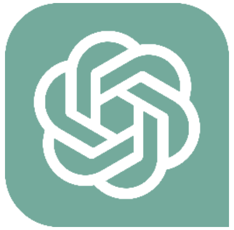
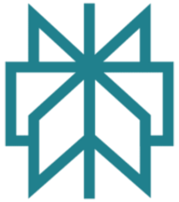
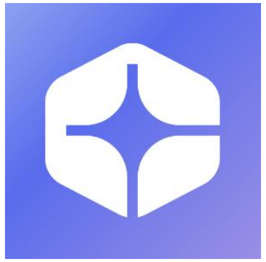
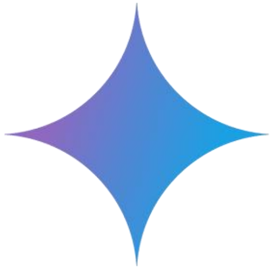
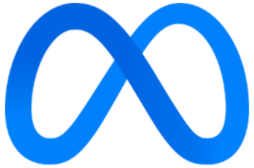
perplexity

Meta
(Llama 3.1 8B)

Manual vs WS&ML vs AI search engines

Firm ID	Firm administrative info	✋	ML				✦	∞
123	name, address, VAT, NACE, …	**abc.com**	abc.com	abc.com	abc.com	abc.com	abc.com	abc.com
234	name, address, VAT, NACE, …		cab.it	cab.it		cab.com		
345	name, address, VAT, NACE, …	**bcd.it**	aaa.it	bcd.it			bcd.it	
456	name, address, VAT, NACE, …				fgv.it		fgv.it	
567	name, address, VAT, NACE, …	**cde.net**	cde.net		cde.net	cde.net		cde.net
678	name, address, VAT, NACE, …	**def.org**	def.org	def.net		def.it		def.org
…	…	…	…	…	…	…	…	…

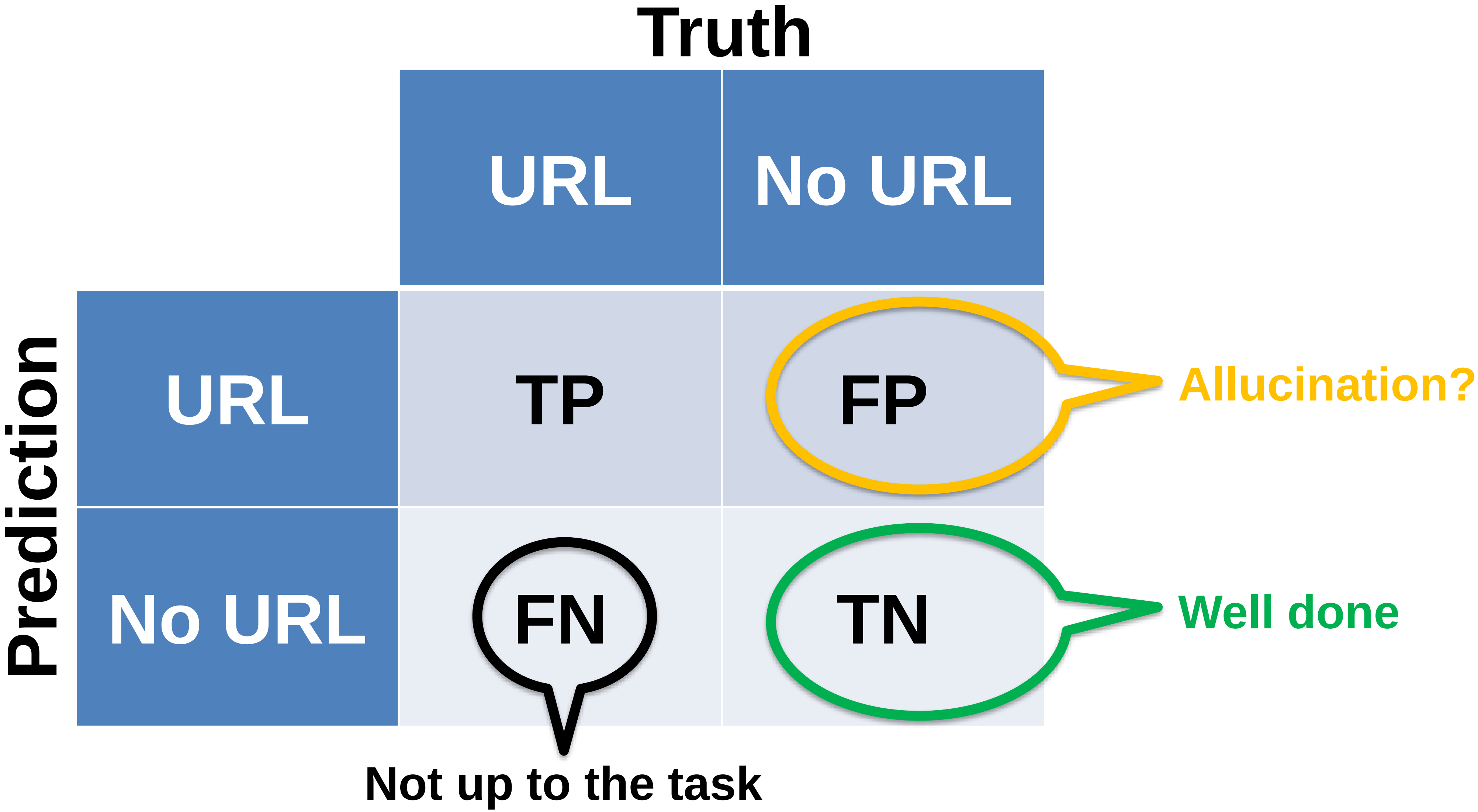
Overall Accuracy (real URL = predicted URL)

 					
0.896	0.778	0.654	0.508	0.692	0.368
1	2			3	

Accuracy by NACE class (real URL = predicted URL)

	ML				✦	∞
C	0.84	0.82	0.62	0.55	0.63	0.33
DEJLMS	0.95	0.79	0.76	0.52	0.68	0.48
F	0.98	0.88	0.78	0.38	0.75	0.4
G	0.91	0.72	0.59	0.51	0.71	0.33
H	0.82	0.56	0.58	0.54	0.68	0.36
I	0.89	0.81	0.64	0.48	0.73	0.36
N	0.91	0.8	0.62	0.56	0.71	0.36

Hypothetical confusion matrix



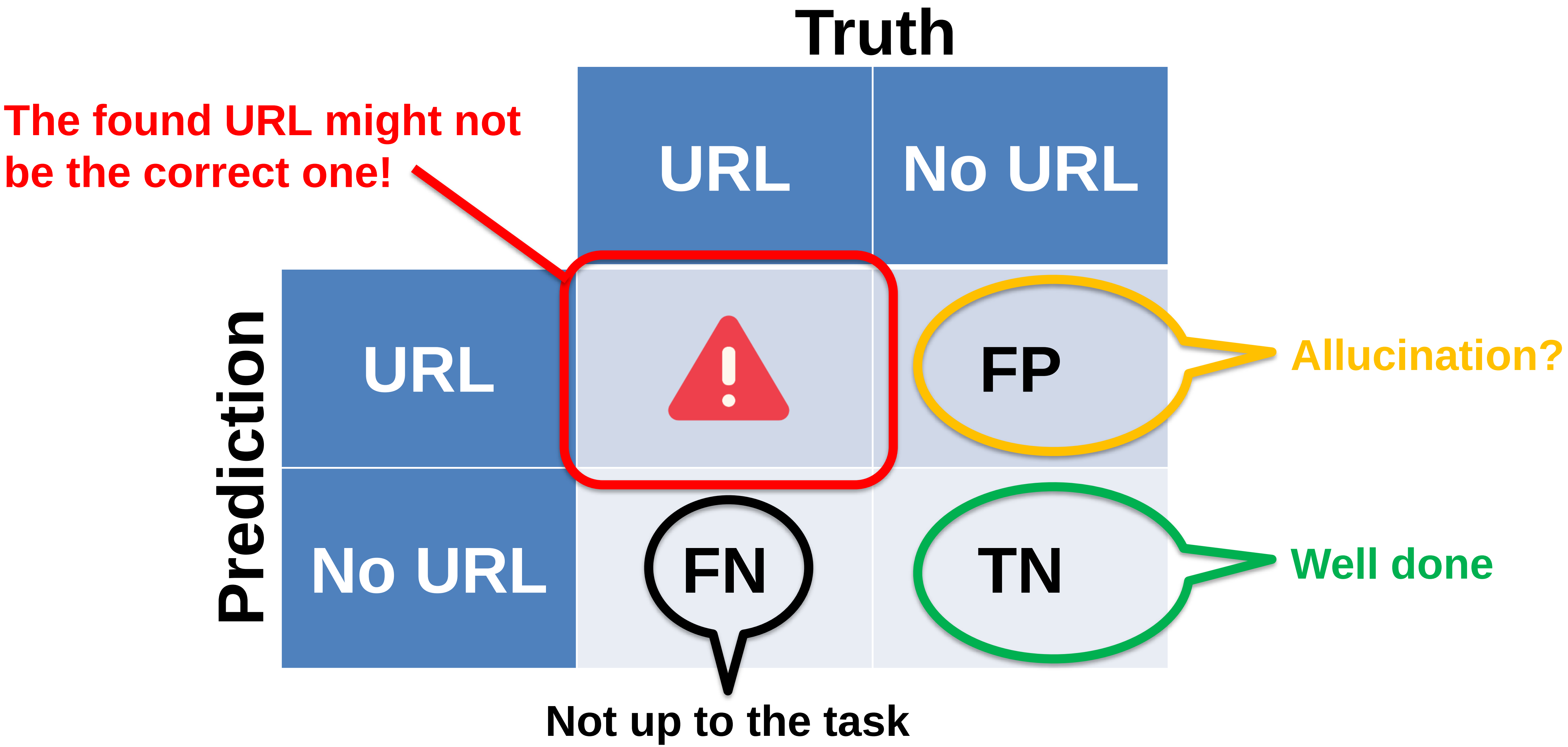
		Truth	
		URL	No URL
Prediction	URL	TP	FP
	No URL	FN	TN

Allucination?

Well done

Not up to the task

Hypothetical confusion matrix



		Truth	
		URL	No URL
Prediction	URL	⚠	FP
	No URL	FN	TN

The found URL might not be the correct one!

Allucination?

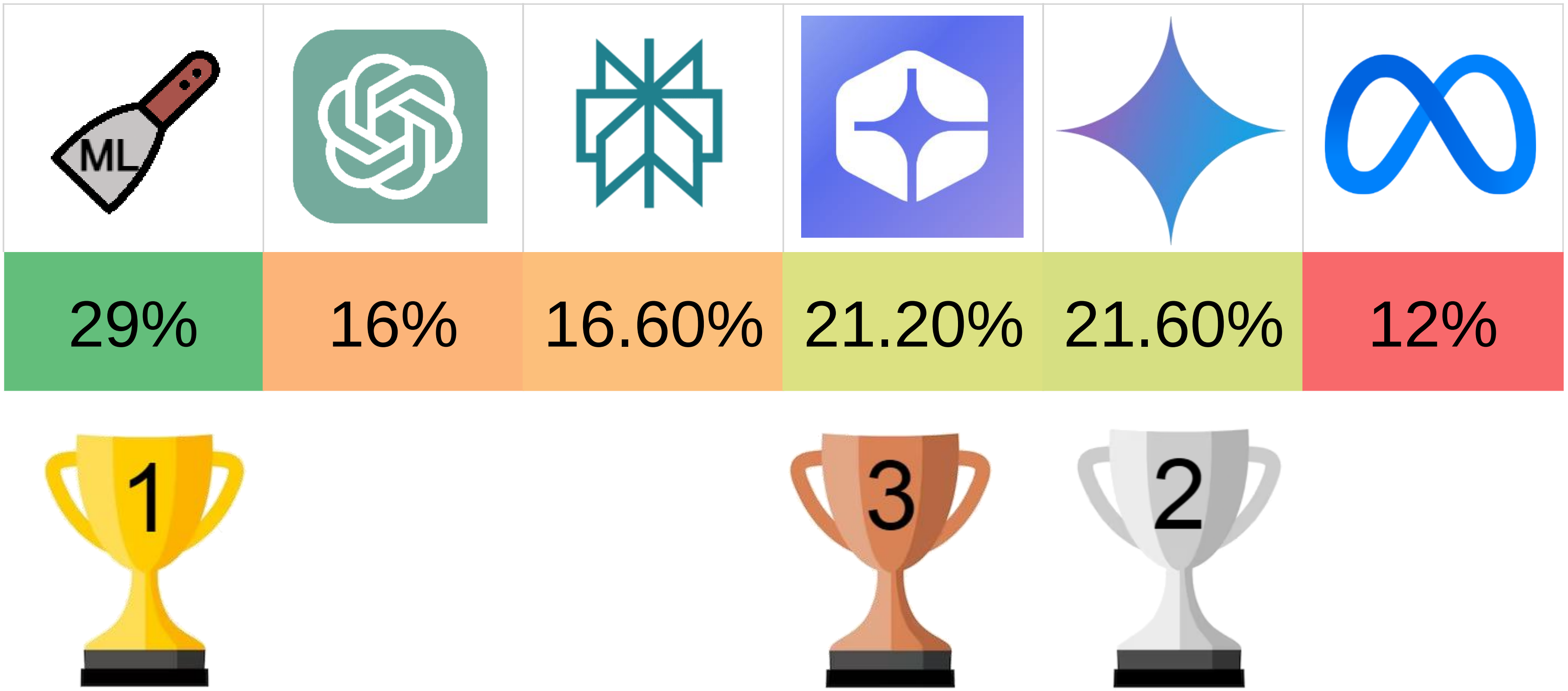
Well done

Not up to the task

True Positives (excluding «false» TPs)

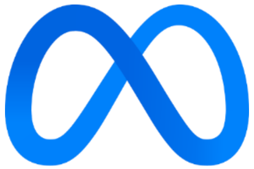
	ML					
real TP ------------- all TP	98%	95.90%	86.20%	56.30%	87.20%	56.40%
	1	2			3	

True Negatives (URL not found and it does not exists)



$\frac{TN}{P + N}$	29%	16%	16.60%	21.20%	21.60%	12%

False Positives (URL predicted, but it should not exist)

						
FP ---------- P + N	3.20%	16.20%	15.60%	11%	10.60%	20.20%
	1			3	2	

NB: if we relax some rules, not all FP are FP (sometimes websites contain just weak administrative info)

False Negatives (URL not found, but it exists)

	ML					
$\frac{FN}{P + N}$	6%	3.40%	11.20%	15.20%	13.20%	23.80%
	2	1	3			

Recap

	ML					
TP	🏆 1	🏆 2	4	6	🏆 3	5
TN	🏆 1	5	4	🏆 3	🏆 2	6
FP	🏆 1	5	4	🏆 3	🏆 2	6
FN	🏆 2	🏆 1	🏆 3	5	4	6

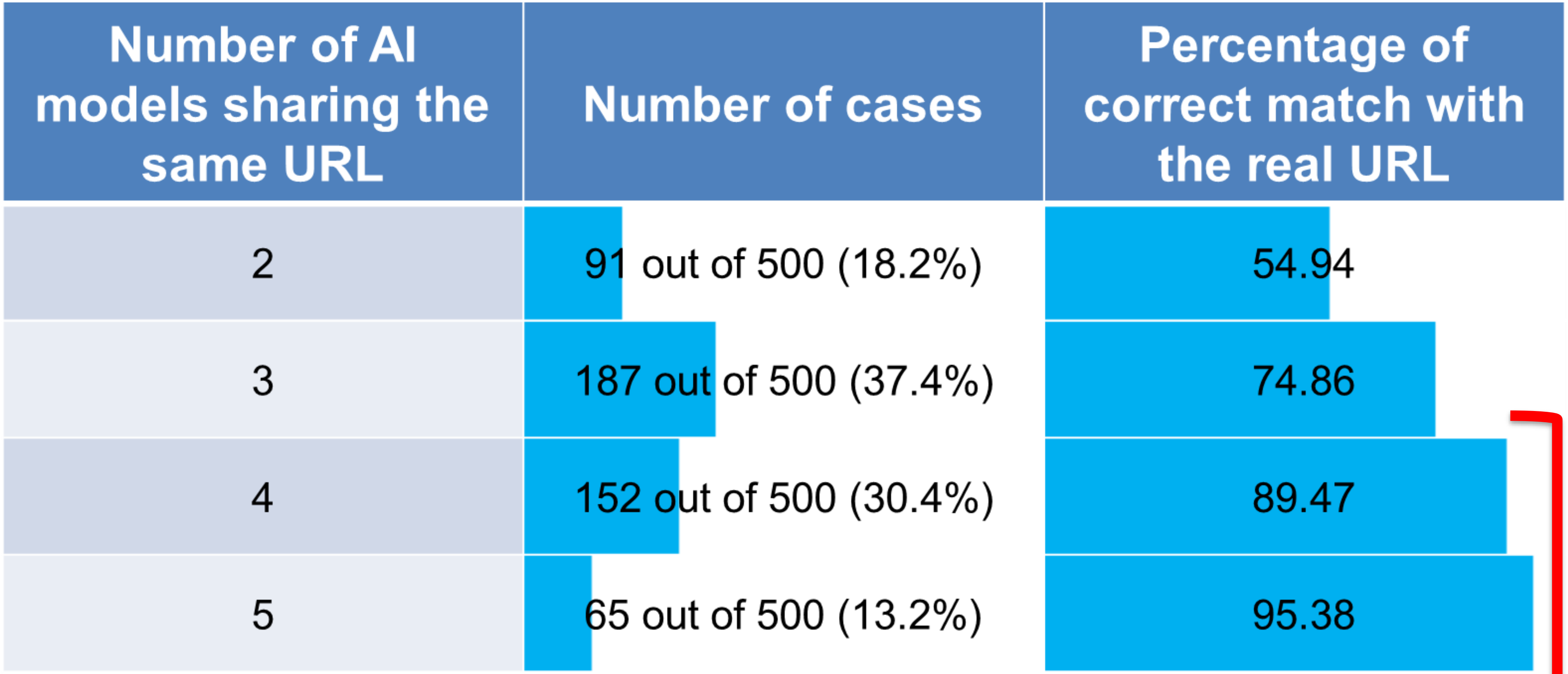
Towards a more efficient URL Retrieval strategy

The Istat pipeline for URL retrieval, based on web scraping and machine learning, currently provides the best results but:

- is time consuming (web scraping)
- requires some manual checks
- needs to be maintained (SW and ML training set)

Can it be matched by an ensemble strategy based on multiple AI search engines?

Accuracy based on AI search engines consensus



Number of AI models sharing the same URL	Number of cases	Percentage of correct match with the real URL
2	91 out of 500 (18.2%)	54.94
3	187 out of 500 (37.4%)	74.86
4	152 out of 500 (30.4%)	89.47
5	65 out of 500 (13.2%)	95.38

43% of cases

Conclusions

- While a task-specific ML model offers superior URL retrieval performance, it requires more effort than general-purpose AI agents, which allow direct queries like

  What is the official website of the enterprise named “Ferrari Spa” and located in “Maranello”?

  and return in a few seconds responses such as

  "www.ferrari.com" or "No official website found“

- Local AI agents with web search capabilities are currently not up for the task. Effective solutions require either commercial offerings or powerful, cloud-deployed AI agents based on open-source LLMs.

- A practical approach might be:

  1) achieve full automation for about 43% of the records (4-5 AI search engine consensus)
  2) rely on the existing system for the remaining records

- The performance gap could decrease in the future, potentially allowing us to use only AI search engines for the URL retrieval task.

Thank you for your attention !