csv-detective

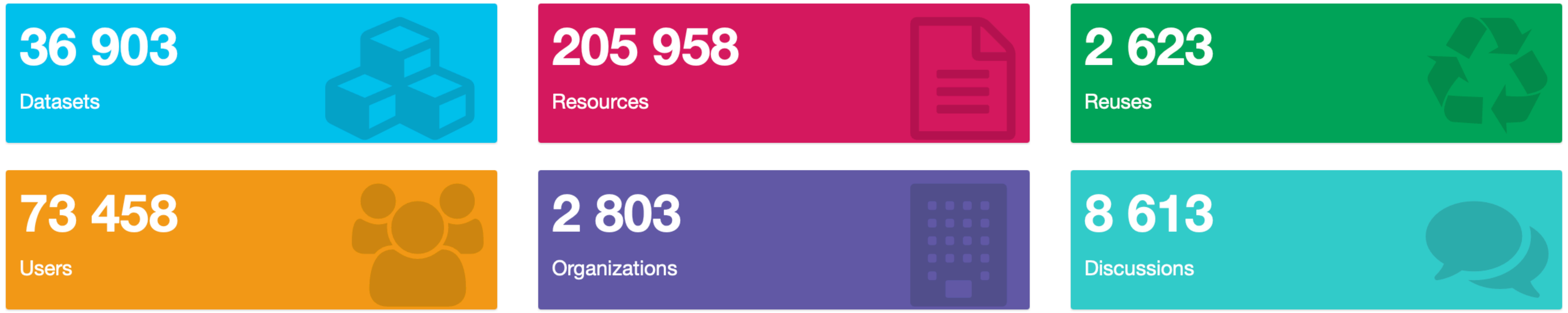
Solving some of the Mysteries of Open Data

Etalab

csv,conf,v6 - Tuesday, May 4 2021

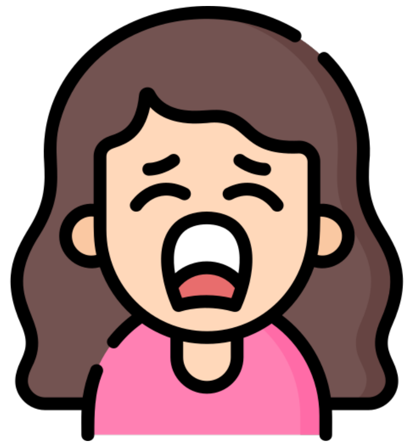
data.gouv.fr / AI Lab

The French Open Data Platform



36 903 Datasets	205 958 Resources	2 623 Reuses
73 458 Users	2 803 Organizations	8 613 Discussions

- Developed and maintained by **Etalab** as a mission under the authority of the Prime Minister
- Anybody can deposit data on the platform.
- Data could be **structured** (very much appreciated) or **not** (we don't want to prevent data to be published)



Data Quality Issues

data.gouv.fr

etalab gouv.fr

How to explore unstructured data

- We can find easily simple data types in datasets with existing tools (ex: pandas profiling)
- But, how can we retrieve complex data types with a business meaning (for further reuses) ?

Information related to a company

Information regarding a locality

Common codes used in administrations

…

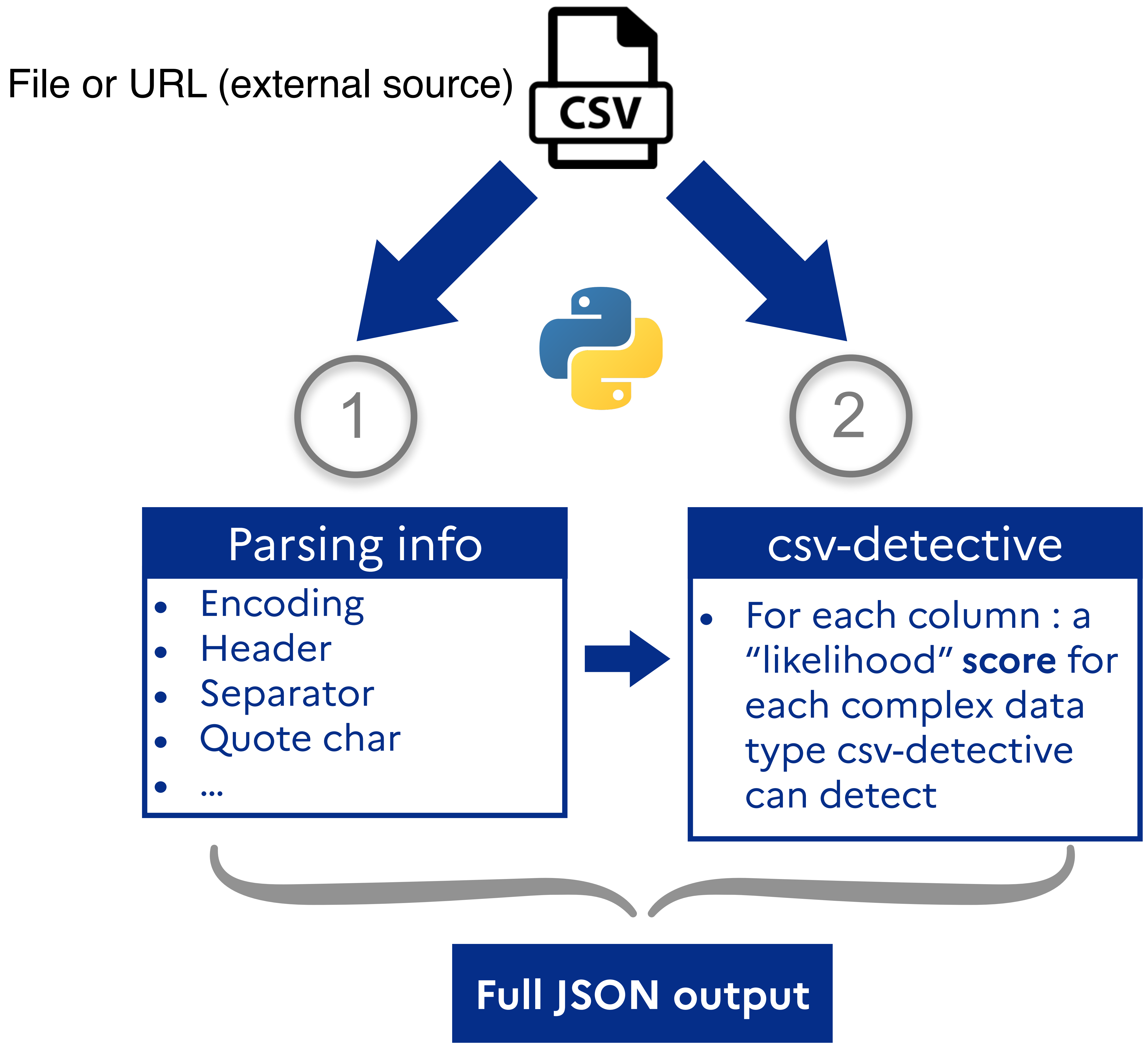
Here comes…

CSV Detective

CSV-DETECTIVE: A PYTHON PACKAGE BY ETALAB

csv-detective

How it works



File or URL (external source)

Parsing info

- Encoding
- Header
- Separator
- Quote char
- …

csv-detective

- For each column : a "likelihood" **score** for each complex data type csv-detective can detect

Full JSON output

List of complex data types that csv-detective is currently able to detect

adresse
booleen
code_commune_insee
code_departement
code_fantoir
code_postal
code_region
code_waldec
commune
csp_insee
date
date_fr
datetime_iso
departement
email
insee_ape700
insee_canton
iso_country_code
jour_de_la_semaine
json_geojson
latitude_l93
latitude_wgs
latitude_wgs_fr_metropole
latlon_wgs
longitude_l93
longitude_wgs
longitude_wgs_fr_metropole
money
other
pays
region
sexe
siren
siret
tel_fr
uai
url
year

csv-detective

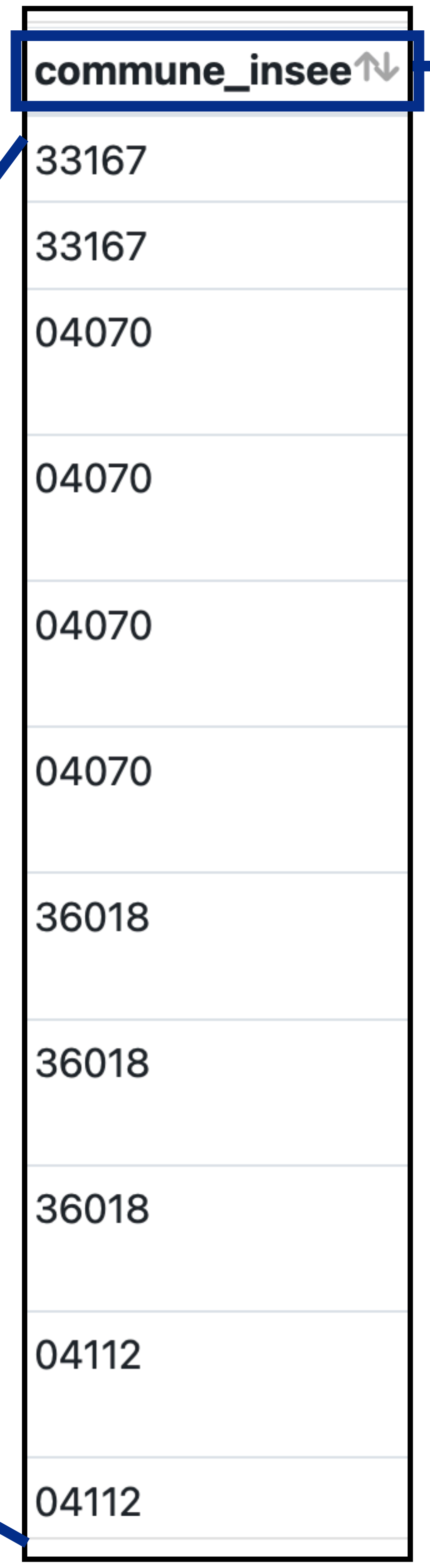
Complex data types scores

In fact, not 1 but 3 different scores are currently returned by csv-detective:

1

field_score

- Analysis of the N first rows of the columns.
- Return the % of values that match the complex data type



commune_insee
33167
33167
04070
04070
04070
04070
36018
36018
36018
04112
04112

Random CSV column

2

label_score

- Analysis based on the header content
- Exact match of header with a known column name —> 1
- Header only contains the known column name —> 0.5

3

ml_score

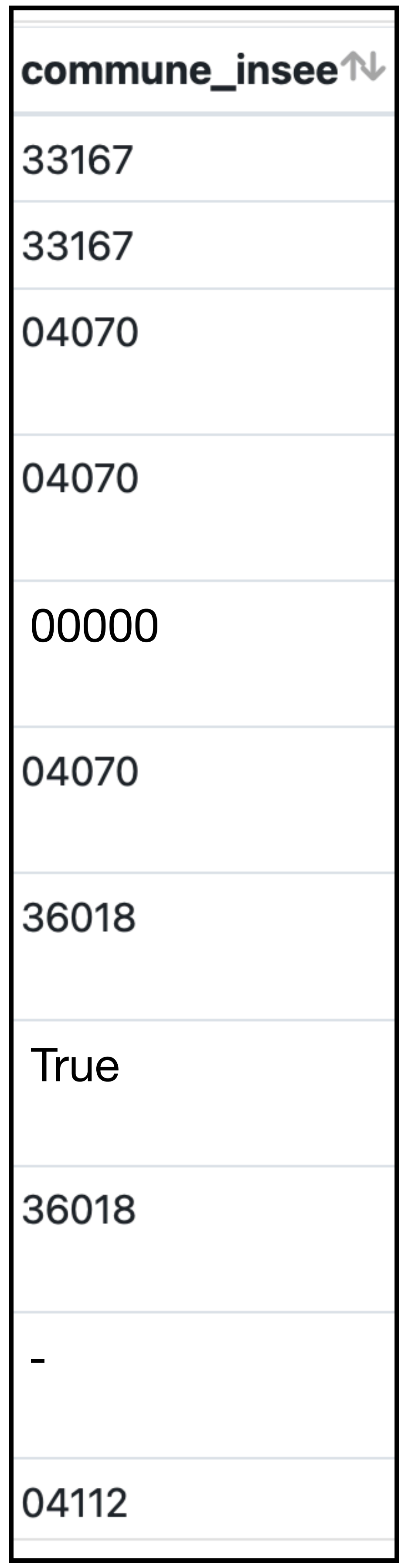
- Likelihood score based on a Machine Learning model trained on annotated data
- Still work in progress

Field Score Calculation Example

Matching test :

This test is based on:

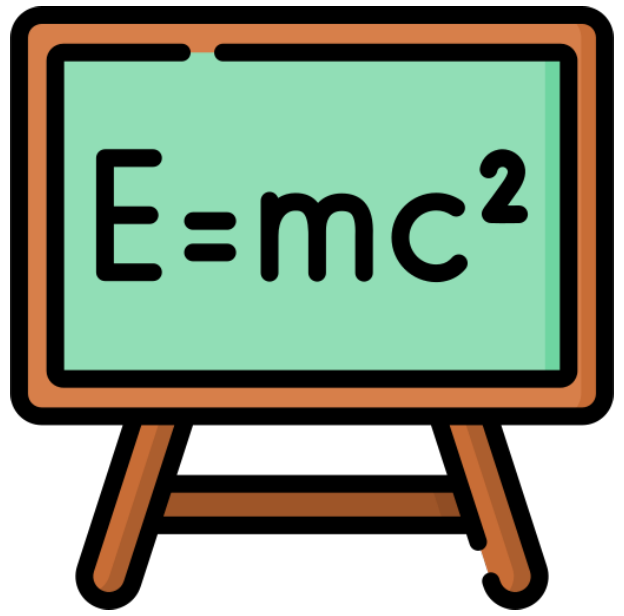
Regex
+
Matching with a comprehensive list of “codes commune Insee” (kind of zip codes)



commune_insee	
33167	✓
33167	✓
04070	✓
04070	✓
00000	✗
04070	✓
36018	✓
True	✗
36018	✓
-	✗
04112	✓

Over the 11 rows, 8 match the “code_commune_insee” format

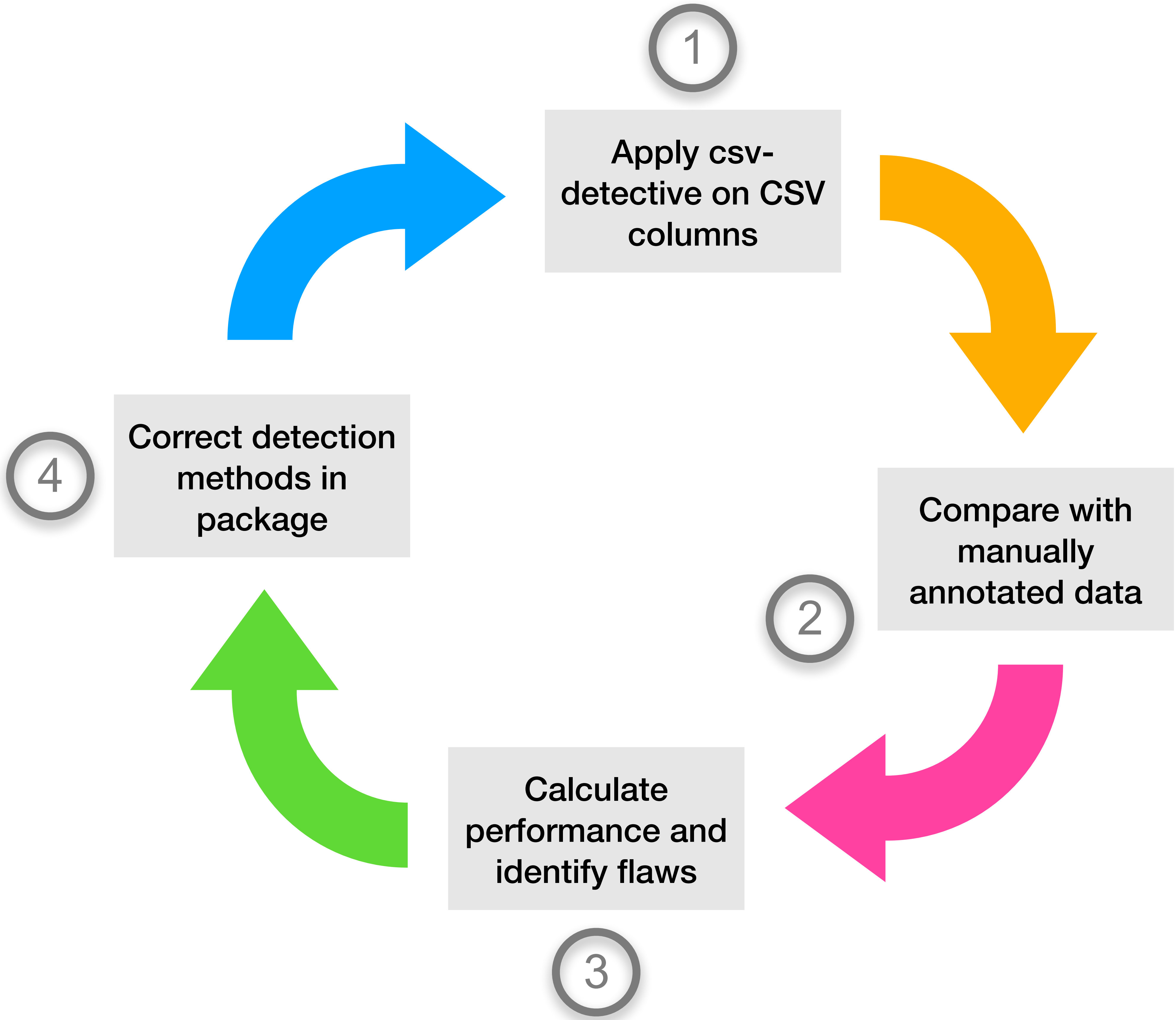
Field Score = 8/11
=
73%



EVALUATION: DOES IT ACTUALLY WORK?

Evaluation

Methodology for evaluation and improvement



1 Apply csv-detective on CSV columns

2 Compare with manually annotated data

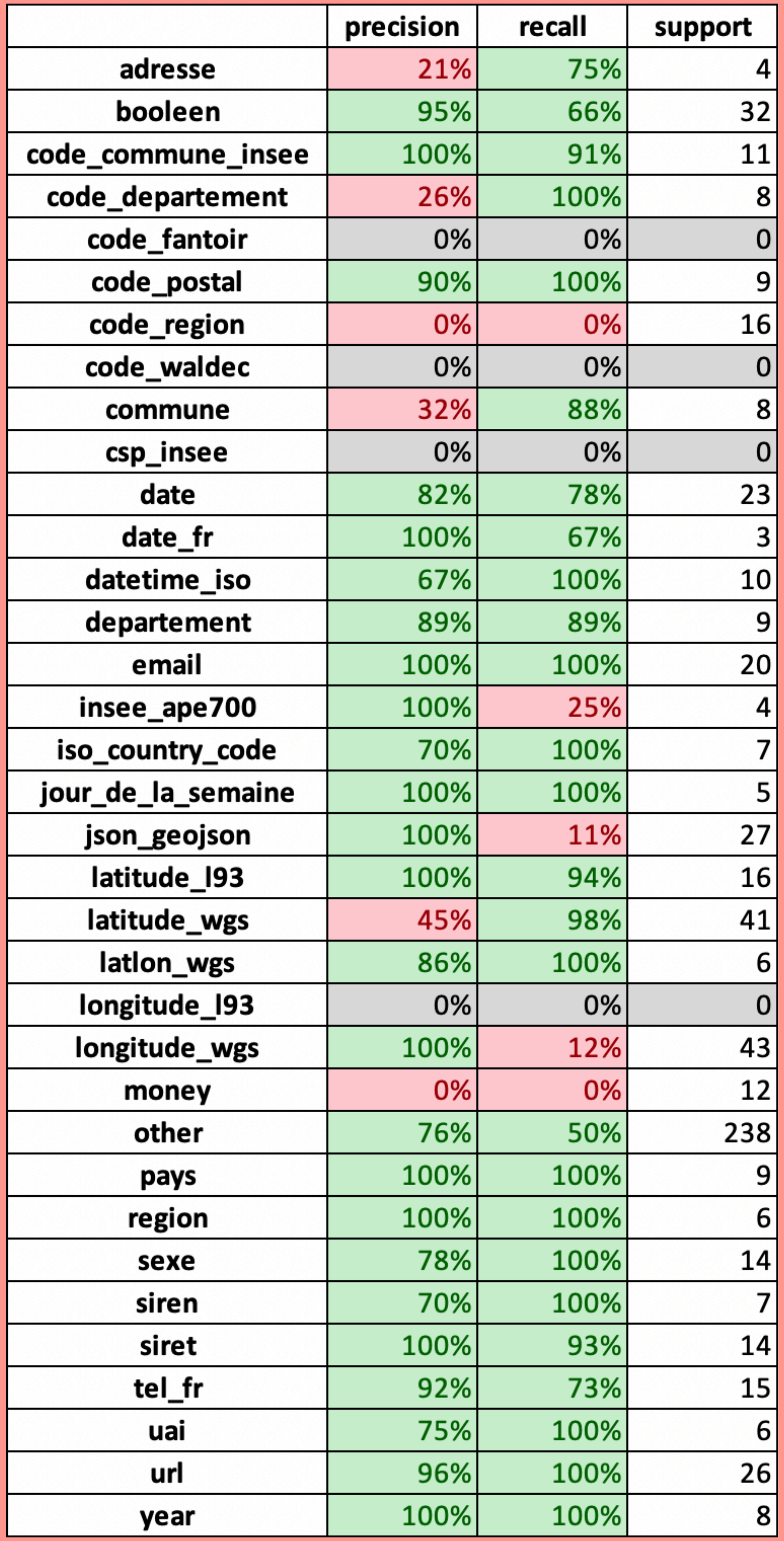
3 Calculate performance and identify flaws

4 Correct detection methods in package

Evaluation

Numerical Evaluation with Precision/Recall

- **Machine Learning method** to evaluate classification models
- **Precision:** “When I predict complex type X, what is the likelihood it is actually true?” (Prediction **quality**)
- **Recall:** “What is the % of actual columns of complex type X that were correctly detected?” (Prediction **comprehensiveness**)
- Well-known **Accuracy** is not relevant here (unbalanced data, different detection method for each complex type…)

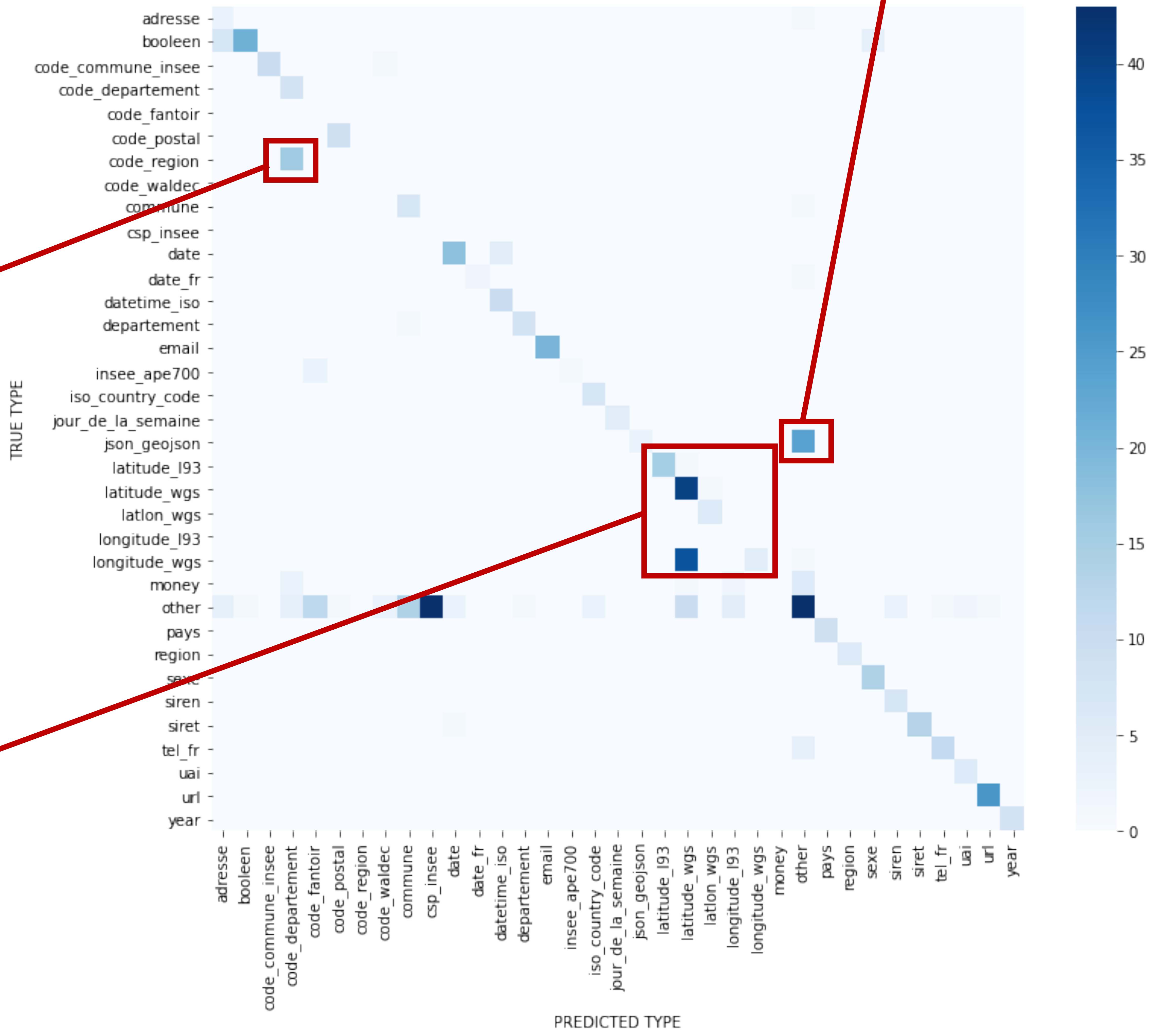


	precision	recall	support
adresse	21%	75%	4
booleen	95%	66%	32
code_commune_insee	100%	91%	11
code_departement	26%	100%	8
code_fantoir	0%	0%	0
code_postal	90%	100%	9
code_region	0%	0%	16
code_waldec	0%	0%	0
commune	32%	88%	8
csp_insee	0%	0%	0
date	82%	78%	23
date_fr	100%	67%	3
datetime_iso	67%	100%	10
departement	89%	89%	9
email	100%	100%	20
insee_ape700	100%	25%	4
iso_country_code	70%	100%	7
jour_de_la_semaine	100%	100%	5
json_geojson	100%	11%	27
latitude_l93	100%	94%	16
latitude_wgs	45%	98%	41
latlon_wgs	86%	100%	6
longitude_l93	0%	0%	0
longitude_wgs	100%	12%	43
money	0%	0%	12
other	76%	50%	238
pays	100%	100%	9
region	100%	100%	6
sexe	78%	100%	14
siren	70%	100%	7
siret	100%	93%	14
tel_fr	92%	73%	15
uai	75%	100%	6
url	96%	100%	26
year	100%	100%	8

Evaluation

Visual Evaluation with Confusion Matrices

- Visualise more precisely how csv-detective tends to **confuse complex types with each other**



json_geojson type is often undetected (classified as “other”)

code_region often confused with **code_departement**

Troubles between **latitudes/ longitudes**

WHAT’S NEXT?

What's next?

csv-detective's future

Methodology Improvement

- **Smarter detection** methods in fields and headers
- **Smarter scores** (combinations, thresholds…)
- Going further on the **ML score** with more annotated data

New Features

- "**Multi**-complex types" detection
- Generalize **file formats other than CSV**

Open Data Ecosystem

- Automatically run csv-detective on every new resource on data.gouv.fr to provide users with **additional information** on datasets
- Use this information for data **quality** and data **previsualisation**
- Allow users to create their **own complex type detection methods**

Summing Up

Key takeaways

- Finding what lives within our CSV files is important !
- An important step for other downstream data cleaning tasks
- Rules + titles may be enough, ML-aided approach has promising potential
- Challenges:
    - Hard to add new rules / Keep up with existing ones
    - Data that lies about their content

ANY QUESTION ?

Source code : https://github.com/etalab/csv-detective

geoffrey.aldebert@data.gouv.fr

anthony.auffret@data.gouv.fr

pavel.soriano@data.gouv.fr