

April 2024



HEC PARIS DMI  
Legal Data Analysis

Jorge Portocarrero  
Edouard Tan

# Data Protection

Analysis of the CNIL's sanctions over the recent years

# Agenda

1. Data Protection as key business matter
2. Simplified legal landscape
3. Goals of our data analysis
4. Road map
5. Results
6. What's next

1.

# Data Protection as key business matter

Personal Data has become an exchangeable asset.

It is a driver for the most thriving sectors of our economy.

The collage consists of three news articles:

- U.S. PIRG Education Fund Article:** Headline: "How Mastercard sells its 'gold mine' of transaction data". Subtext: "Mastercard knows where people shop, how much they spend, and on what days - and it sells that information online." Category: SCIENCE / HEALTH.
- The Washington Post Article:** Headline: "Netflix Paid \$100 Million To Access Private Facebook Messages". Subtext: "Paigambar Mohan Raj, April 2, 2024". Category: BUSINESS.
- The New York Times Article:** Headline: "Your tax return is secret by law. TurboTax wants you to give up your privacy.". Subtext: "Shira Ovide, Updated March 1, 2024 at 2:20 p.m. EST | Published March 1, 2024 at 12:30 p.m. EST". Category: HELP DESK.

2.

# Simplified legal landscape

Data Protection Regulations  
enforced by the CNIL

EU Regulation

GDPR

French Law

LIL

French DPA

CNIL

helps and enforces

Entities and individuals  
processing personnal data

# Goals of our data analysis

What do we want to  
achieve?

3.

Goal 1

Identify the number of sanctions issued by the CNIL each year

Goal 2

Identify the Top 5 most frequently sanctioned industries

Goal 3

Identify the spread of the procedures used by the CNIL

Goal 4

Identify the Top 5 industries with the highest fines

Goal 5

Identify the Top 5 entities with the highest fines

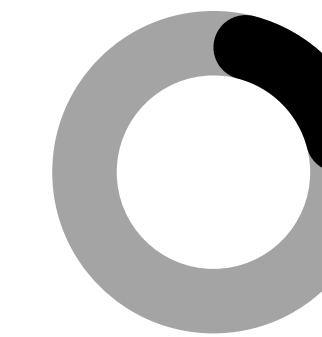
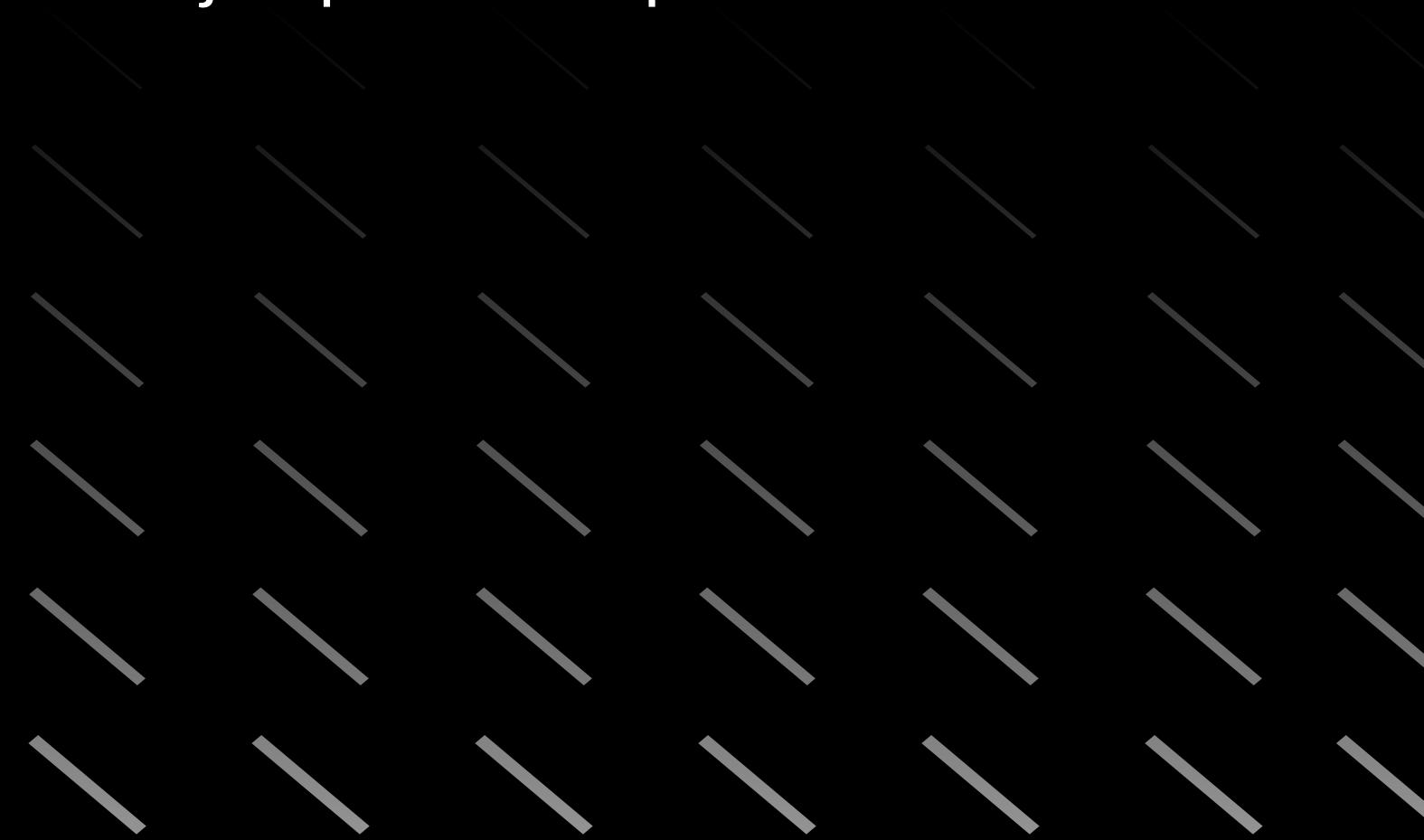
Goal 6

Identify the Top 5 most frequent breaches

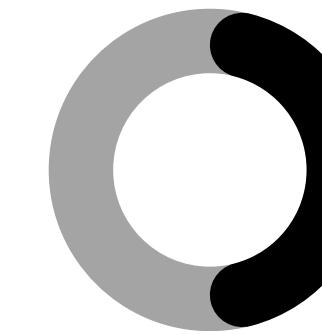
4.

# Road map

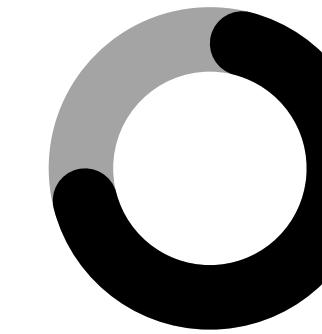
Let's jump into the process



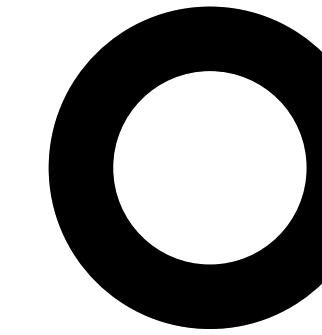
Step 1  
Data collection



Step 2  
Data sanitization



Step 3  
Data optimization



Step 4  
Data analysis

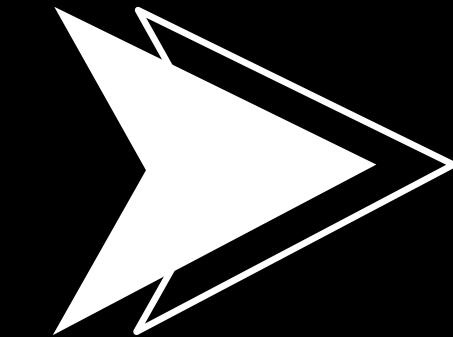
## STEP 1 - DATA COLLECTION

# How to achieve data collection?

legifrance.fr

Website

Web Scraping



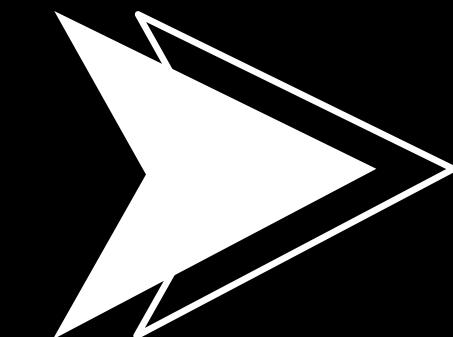
Data retrieved

- Sanction number
- Sanction date
- Sanctioned entity's name
- Type of entity
- GDPR and LIL violated articles
- Fine amount

cnil.fr

Website

Web Scraping



Data retrieved

- Sanction date
- Industry
- Fine amount
- Breaches

## STEP 1 - DATA COLLECTION

CODE VIEW

# Web Scraping

We initialize the process in the IDE



WS\_LEGIFRANCE.py

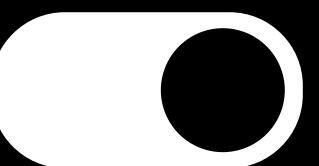
```
1 from selenium import webdriver
2 from selenium.webdriver.chrome.service import Service
3 from selenium.webdriver.chrome.options import Options
4 from bs4 import BeautifulSoup
5 import pandas as pd
6 import re
7 import time
8 import urllib.parse
9
10 options = Options()
11 service = Service(executable_path='/Users/Edouard Tan/LDA/chromedriver')
12 driver = webdriver.Chrome(service=service, options=options)
13 url = "https://www.legifrance.gouv.fr/search/cnil?
    facetteNatureDelib=Sanction&page=1&pageSize=100&searchField=ALL&searchType=ALL&sortValue=DATE_DECISION_DESC&tab_selection=cnil&typePagination=DEFAULT"
14 driver.get(url)
15 time.sleep(5)
16
17 soup = BeautifulSoup(driver.page_source, 'html.parser')
18 articles = soup.find_all('article', class_='result-item')
19
20 data = []
21
22 for article in articles:
23     link = article.find('a')['href']
24     full_link = f"https://www.legifrance.gouv.fr{link}"
25     driver.get(full_link)
```

## STEP 1 - DATA COLLECTION

# Web Scraping

We take a look at the relevant html elements

## WEBSITE VIEW



225 texte(s) trouvé(s) Date de décision antéchronologique ▾ Afficher 100 résultats par page ▾

- > Délibération de la formation restreinte n°SAN-2024-003 du 31 janvier 2024 concernant la société FORIOU
- > Délibération de la formation restreinte n°SAN-2024-002 du 31 janvier 2024 concernant la société DE PARTICULIER A PARTICULIER – EDITIONS NERESSIS
- > Délibération de la formation restreinte n°SAN-2023-025 du 29 décembre 2023 concernant la société TAGADAMEDIA
- > Délibération de la formation restreinte n°SAN-2023-024 du 29 décembre 2023 concernant la société YAHOO EMEA LIMITED
- > Délibération de la formation restreinte n°SAN-2023-023 du 29 décembre 2023 concernant la société NS CARDS FRANCE
- > Délibération de la formation restreinte n°SAN-2023-021 du 27 décembre 2023 concernant la société AMAZON FRANCE LOGISTIQUE
- > Délibération de la formation restreinte n°SAN-2023-018 du 12 décembre 2023 concernant la commune de KOUROU
- > Délibération de la formation restreinte n°SAN-2023-017 du 11 décembre 2023 concernant le ministère de l'intérieur et des Outre-mer et le ministère de l'Europe et des Affaires étrangères

The screenshot shows the browser's developer tools with the 'Elements' tab selected. The page URL is visible in the address bar. The main content area displays a list of legal notices (Délibérations) from the CNIL. The first item in the list is highlighted with a blue selection bar. The corresponding HTML code for this item is shown in the bottom right pane:

```
<div>
  <article class="result-item" id="id-result-item-0">
    <h2 class="title-result-item" data-id="CNILTEXT00049231950">flex</h2>
    <span aria-hidden="true" class="title-icon chevron">...</span>
    <a href="/cnil/id/CNILTEXT00049231950?facetteNatureDelib=Sanction&isAdvancedResult=true&page=1&pageSize=100&query=%7B%28%40ALL%5Bt%22%22%5D%29%7D&sortValue=DATE_DECISION_DESC&tab_selection=cnil&timeInterval=&typePagination=DEFAULT" onclick="eulerianClick(this, 'resultatRecherche')" id="idCNILTEXT00049231950">"Délibération de la formation restreinte n°SAN-2024-003 du 31 janvier 2024 concernant la société FORIOU " = $0</a>
  </article>
<article class="result-item" id="id-result-item-1">...</article>
<article class="result-item" id="id-result-item-2">...</article>
<article class="result-item" id="id-result-item-3">...</article>
<article class="result-item" id="id-result-item-4">...</article>
<article class="result-item" id="id-result-item-5">...</article>
<article class="result-item" id="id-result-item-6">...</article>
<article class="result-item" id="id-result-item-7">...</article>
<article class="result-item" id="id-result-item-8">...</article>
<article class="result-item" id="id-result-item-9">...</article>
<article class="result-item" id="id-result-item-10">...</article>
<article class="result-item" id="id-result-item-11">...</article>
<article class="result-item" id="id-result-item-12">...</article>
<article class="result-item" id="id-result-item-13">...</article>
<article class="result-item" id="id-result-item-14">...</article>
<article class="result-item" id="id-result-item-15">...</article>
<article class="result-item" id="id-result-item-16">...</article>
<article class="result-item" id="id-result-item-17">...</article>
<article class="result-item" id="id-result-item-18">...</article>
<article class="result-item" id="id-result-item-19">...</article>
<article class="result-item" id="id-result-item-20">...</article>
<article class="result-item" id="id-result-item-21">...</article>
<article class="result-item" id="id-result-item-22">...</article>
<article class="result-item" id="id-result-item-23">...</article>
<article class="result-item" id="id-result-item-24">...</article>
```

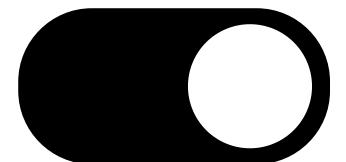
# Web Scraping

We implement them and retrieve data from the sanction title with regex



WS\_LEGIFRANCE.py

```
22 for article in articles:
23     link = article.find('a')['href']
24     full_link = f"https://www.legifrance.gouv.fr{link}"
25     driver.get(full_link)
26     time.sleep(5)
27     detail_soup = BeautifulSoup(driver.page_source, 'html.parser')
28
29     sanction_intro = detail_soup.find('div', {'class': 'intro-under-frame'}).strong.text
30     sanction_titre_regex = r"Délibération de la formation restreinte (?:(n°|no |n°)?\s*(SAN[\s--]*\d+[\s--]*\d+) du (\d+\s?\w+\s?\d+) (?:(concernant|prononçant une|
sanction (?:(pécuniaire )?à l['']encontre de) (?:(la |le |les |l['']))?(société|SOCIETE|sociétés|commune|personne|
physique|GIE|Mme|Monsieur|Association|Fédération)? ?(.+)"
31     match = re.search(sanction_titre_regex, sanction_intro)
32
33     if match:
34         sanction_numero = match.group(1)
35         sanction_date = match.group(2)
36         entite_type = match.group(3) if match.group(3) else "Non spécifié"
37         entite_nom = match.group(4).strip()
38
39         if any(pattern in entite_nom.lower() for pattern in ['x', 'X', 'x et y', '[...']]:
40             query = f"CNIL {sanction_numero}"
41             entite_nom = f"https://www.google.com/search?q={urllib.parse.quote(query)}"
42
43     else:
44         continue
```



# Web Scraping

We locate and identify how fine amounts and infringed articles are written

82. Dès lors, au regard de la responsabilité de la société, de ses capacités financières et des critères pertinents de l'article 83, paragraphe 2, du RGPD évoqués ci-dessus, la formation restreinte estime qu'une amende de trois cent dix mille euros (310 000 €) apparaît justifiée.

83. En troisième lieu, s'agissant de la publicité de la sanction, la formation restreinte considère que celle-ci se justifie au regard de la gravité du manquement en cause, de la position de la société sur le marché, de la portée du traitement et du nombre de personnes concernées.

84. Elle relève également que cette mesure a notamment vocation à informer les personnes concernées par les opérations de prospection de la société. Cette information leur permettra, le cas échéant, de faire valoir leurs droits.

85. Enfin, elle estime que cette mesure est proportionnée dès lors que la décision n'identifiera plus nommément la société à l'expiration d'un délai de deux ans à compter de sa publication.

## PAR CES MOTIFS

**La formation restreinte de la CNIL, après en avoir délibéré, décide de :**

- prononcer une amende administrative à l'encontre de la société FORIOU d'un montant de trois cent dix mille euros (310 000 €) pour manquement à l'article 6 du RGPD ;
- rendre publique, sur le site web de la CNIL et sur le site web de Légifrance, sa délibération, qui ne permettra plus d'identifier nommément la société à l'issue d'une durée de deux ans à compter de sa publication.

Le président

Alexandre LINDEN

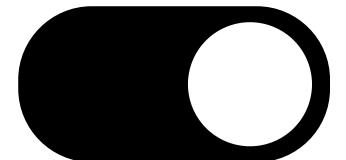
Cette décision est susceptible de faire l'objet d'un recours devant le Conseil d'État dans un délai de deux mois à compter de sa notification.

```


<p></p>
  <strong>PAR CES MOTIFS</strong>
</p>
<p></p>
  <strong>La formation restreinte de la CNIL, après en avoir délibéré, décide de :</strong>
</p>
<p></p>
  <strong>
    • prononcer une amende administrative à l'encontre de la société FORIOU d'un
      montant de trois cent dix mille euros (310 000 €) pour manquement à l'article 6
      du RGPD ; </strong>
</p>
...
  <strong> == $0
    • rendre publique, sur le site web de la CNIL et sur le site web de Légifrance,
      sa délibération, qui ne permettra plus d'identifier nommément la société à
      l'issue d'une durée de deux ans à compter de sa publication.</strong>
</p>
<p></p>
<p></p>
<p></p>
<p>Le président</p>


```

# Web Scraping



We isolate the relevant section and elaborate regex to get fine amounts and infringed articles



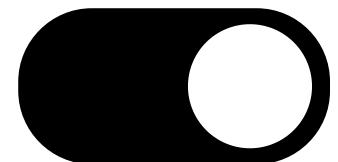
continue

WS\_LEGIFRANCE.py

```
46     sanction_corps = detail_soup.get_text()
47
48     rgpd_ref_regex = r"Vu le règlement \(\UE\)"
49     if not re.search(rgpd_ref_regex, sanction_corps):
50         print("Référence au RGPD non trouvée, arrêt du programme.")
51         break
52
53     start_index = sanction_corps.find("PARCESMOTIFS")
54     end_index = sanction_corps.find("Alexandre LINDEN", start_index)
55     if start_index != -1 and end_index == -1:
56         end_index = len(sanction_corps)
57     sanction_corps_focus = sanction_corps[start_index:end_index] if start_index != -1 else ""
58
59     amendes_regex = r"montant de.*?(\d{1,3}(?:[\s\.,]\?\d{3})*)).*?(?:pour|au regard|au titre)?"
60     amendes_match = re.findall(amendes_regex, sanction_corps_focus, re.IGNORECASE | re.DOTALL)
61     amendes_str = "; ".join(amendes_match) if amendes_match else "Aucune amende pécuniaire prononcée."
62
63     articles_vise_set = set()
64
65     amendes_regex = r"articles? (\d+(:[-.]\d+)?(:-[a-z]\?)?)?(:,? paragraphe\s?\d+)?(:\s?[a-z]\?)?(:\s*(?:,|et)\s*\d+(:[-.]\d+)?(:-[a-z]\?)?)?(:,? paragraphe\s?\d+)?(:\s?[a-z]\?)?)*\b"
66     fondement_regex = r"(du RGPD|du règlement \(\UE\)|du règlement no|de la loi (Informatique et Libertés|informatique et libertés|n° 78-17|du 6 janvier 1978 modifiée)|\"Informatique et Libertés\"|\"informatique et libertés\"|\\"[ ]?Informatique et Libertés[ ]?\")"
67
68     articles_matches = re.findall(amendes_regex, sanction_corps_focus, re.IGNORECASE | re.DOTALL)
69
```

# Web Scraping

We retrieve fine amounts and infringed articles and simplify the legal ground



WS\_LEGIFRANCE.py

```
68     articles_matches = re.findall(amendes_regex, sanction_corps_focus, re.IGNORECASE | re.DOTALL)
69
70     for article_match in articles_matches:
71         subsequent_text = sanction_corps_focus[sanction_corps_focus.find(article_match)+len(article_match):]
72         fondement_match = re.search(fondement_regex, subsequent_text, re.IGNORECASE)
73
74         fondement_reformulation = ""
75         if fondement_match:
76             extracted_text = fondement_match.group()
77             if "RGPD" in extracted_text or "règlement" in extracted_text:
78                 fondement_reformulation = "du RGPD"
79             elif "loi" in extracted_text:
80                 fondement_reformulation = "de la LIL"
81             result = f"Article {article_match} {fondement_reformulation}".strip()
82             articles_vise_set.add(result)
83
84     articles_vise_str = "; ".join(sorted(articles_vise_set)) if articles_vise_set else "Non trouvés"
85
86     data.append([sanction_numero, sanction_date, entite_type, entite_nom, amendes_str, articles_vise_str])
87
88 driver.quit()
89
90 df = pd.DataFrame(data, columns=['Numéro de sanction', 'Date', 'Type d\'entité', 'Nom de l\'entité', 'Montants des amendes', 'Articles violés'])
91 df.to_csv('DS_LEGIFRANCE_RAW.csv', index=False)
92
93 print("La collecte des données est terminée. Le fichier CSV 'DS_LEGIFRANCE_RAW.csv' a été enregistré sur votre ordinateur.")
```

# Web Scraping

We elaborate a data frame with all the data retrieved and export it to a CSV file

```
WS_LEGIFRANCE.py
articles_matches = re.findall(amendes_regex, sanction_corps_focus, re.IGNORECASE, re.DOTALL)
69
70 for article_match in articles_matches:
71     subsequent_text = sanction_corps_focus[sanction_corps_focus.find(article_match) + len(article_match):]
72     fondement_match = re.search(fondement_regex, subsequent_text, re.IGNORECASE)
73
74     fondement_reformulation = ""
75     if fondement_match:
76         extracted_text = fondement_match.group()
77         if "RGPD" in extracted_text or "règlement" in extracted_text:
78             fondement_reformulation = "du RGPD"
79         elif "loi" in extracted_text:
80             fondement_reformulation = "de la LIL"
81         result = f"Article {article_match} {fondement_reformulation}".strip()
82         articles_vise_set.add(result)
83
84     articles_vise_str = "; ".join(sorted(articles_vise_set)) if articles_vise_set else "Non trouvés"
85
86     data.append([sanction_numero, sanction_date, entite_type, entite_nom, amendes_str, articles_vise_str])
87
88 driver.quit()
89
90 df = pd.DataFrame(data, columns=['Numéro de sanction', 'Date', 'Type d\'entité', 'Nom de l\'entité', 'Montants des amendes', 'Articles violés'])
91 df.to_csv('DS_LEGIFRANCE_RAW.csv', index=False)
92
93 print("La collecte des données est terminée. Le fichier CSV 'DS_LEGIFRANCE_RAW.csv' a été enregistré sur votre ordinateur.")
```

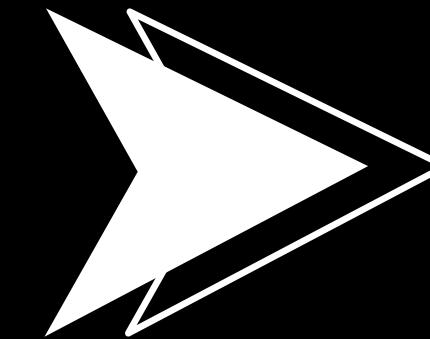
## STEP 2 - DATA SANITIZATION

# How to achieve data sanitization?

### Data from legifrance.fr

- Sanction date (Day Month Year)
- Sanctioned entity's name
- Type of entity ("société")
- GDPR and LIL violated articles

### Data refining



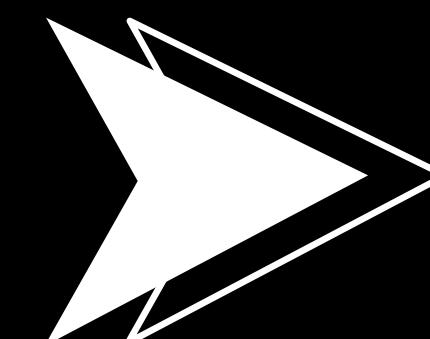
### Data from legifrance.fr

- Sanction date (DD/MM/YYYY)
- Sanctioned entity's name (completed)
- Type of entity ("SOCIETE")
- GDPR and LIL violated articles (completed)

### Data from cnil.fr

- Industry
- Breaches

### Data refining



### Data from cnil.fr

- Industry (simplified with keywords)
- Breaches (simplified with keywords)

## STEP 3 - DATA OPTIMIZATION

# How to achieve data optimization?

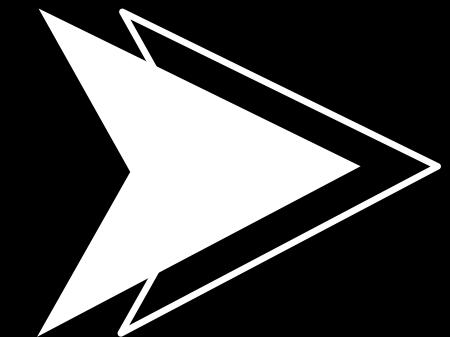
### Data from legifrance.fr

- Sanction number
- Sanction date
- Sanctioned entity's name
- Type of entity
- GDPR and LIL violated articles
- Fine amount

### Data from cnil.fr

- Sanction date
- Industry
- Fine amount
- Breaches

### Data merging



### Merged data

- Sanction number
- Sanction date
- Sanctioned entity's name
- Type of entity
- GDPR and LIL violated articles
- Fine amount
- Sanction date
- Industry
- Fine amount
- Breaches

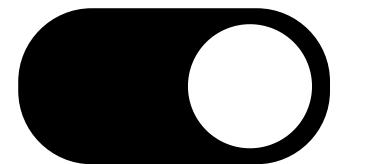
# Data Merging

We combine the two files to get one general file containing relevant data



DF\_MERGING.py

```
1 import pandas as pd
2
3 def fusionner_fichiers_csv(chemin_legifrance, chemin_cnil, chemin_sortie, format_date='%d/%m/%Y'):
4
5     df_legifrance = pd.read_csv(chemin_legifrance)
6     df_cnil = pd.read_csv(chemin_cnil)
7
8     df_legifrance.columns = ['Numero_Sanction', 'Date', 'Type_Entite', 'Nom_Entite', 'Montant_Amende', 'Articles_RGPD', 'Articles_LIL']
9     df_cnil.columns = ['Date', 'Secteur d\'activité', 'Manquements', 'Montant_Amende']
10
11    df_legifrance['Date'] = pd.to_datetime(df_legifrance['Date'], errors='coerce', format=format_date)
12    df_cnil['Date'] = pd.to_datetime(df_cnil['Date'], errors='coerce', format=format_date)
13
14    df_legifrance['Montant_Amende'] = pd.to_numeric(df_legifrance['Montant_Amende'], errors='coerce')
15    df_cnil['Montant_Amende'] = pd.to_numeric(df_cnil['Montant_Amende'], errors='coerce')
16
17    df_merged = pd.merge(df_legifrance, df_cnil[['Date', 'Montant_Amende', 'Secteur d\'activité', 'Manquements']],
18                          on=['Date', 'Montant_Amende'],
19                          how='outer',
20                          suffixes=(None, '_cnil'))
21
22    df_merged.to_csv(chemin_sortie, index=False, date_format=format_date)
23
24 chemin_legifrance = '/Users/Edouard Tan/LDA/DF_LEGIFRANCE_CLEANED.csv'
25 chemin_cnil = '/Users/Edouard Tan/LDA/DF_CNIL_CLEANED.csv'
26 chemin_sortie = '/Users/Edouard Tan/LDA/DF_SANCTIONS_MERGED.csv'
27 format_date = '%d/%m/%Y'
28
29 fusionner_fichiers_csv(chemin_legifrance, chemin_cnil, chemin_sortie, format_date)
```



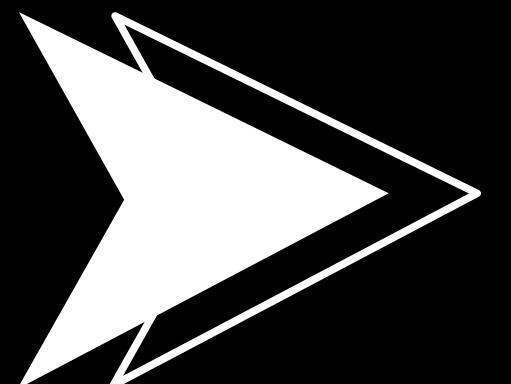
STEP 4 - DATA ANALYSIS

# How to achieve data analysis?

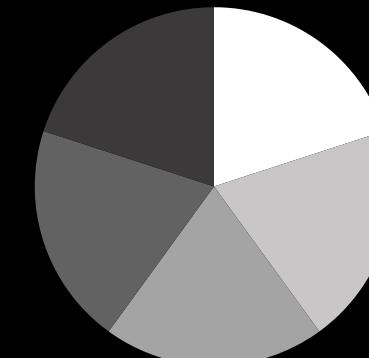
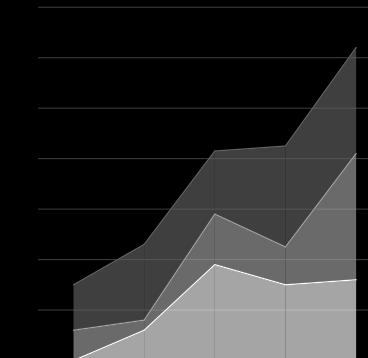
## Merged data

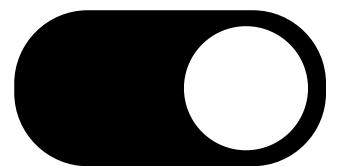
- Sanction number
- Sanction date
- Sanctioned entity's name
- Type of entity
- Fine amount
- Sanction date
- Industry
- Fine amount
- Breaches

## Data analysis



## Results





# Data Analysis & Visualization

We initiate the analysis and start the visualization of data



DF\_ANALYSIS.py

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4
5 sns.set_style("dark", {'axes.facecolor': 'black', 'grid.color': 'none', 'axes.edgecolor': 'white'})
6 plt.rcParams.update({
7     'figure.figsize': (10, 6),
8     'text.color': 'white',
9     'axes.labelcolor': 'white',
10    'axes.titlesize': 18,
11    'axes.labelsize': 14,
12    'xtick.labelsize': 12,
13    'ytick.labelsize': 12,
14    'xtick.color': 'white',
15    'ytick.color': 'white',
16    'axes.spines.top': False,
17    'axes.spines.right': False,
18    'axes.spines.left': True,
19    'axes.spines.bottom': True,
20    'font.family': 'sans-serif',
21    'font.sans-serif': 'Lato',
22 })
23
24 width_px = 1180
25 height_px = 1080
26
27 dpi = 300
```

## STEP 4 - DATA ANALYSIS

CODE VIEW

# Data Analysis & Visualization

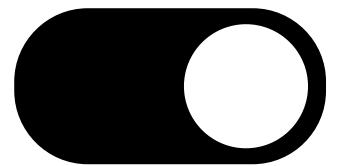
We analyze based on what we want to study

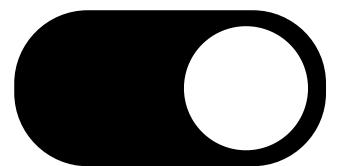


In [1]: In [1]:

DF\_ANALYSIS.py

```
35 df = pd.read_csv("DF_SANCTIONS_MERGED&CLEANED.csv", delimiter=';')
36
37 df['Year'] = pd.to_datetime(df['Date'], format='%d/%m/%Y').dt.year
38
39 # Temporal analysis of the number of sanctions over the years
40 sanctions_per_year = df.groupby('Year').size()
41
42 # Temporal analysis of the types of procedures used
43 procedures_per_year = df.groupby(['Year', 'Type de procédure']).size().unstack(fill_value=0)
44
45 # Analysis of the recurrence of sanctions
46 most_sanctioned_entities = df['Normalized_Entite'].value_counts().head(5)
47 most_sanctioned_sectors = df['Broad_Sector'].value_counts().head(5)
48
49 # Financial analysis
50 fine_stats = df['Montant_Amende'].describe()
51
52 # Group by normalized entity names and sum the fines
53 entities_total_fines = df.groupby('Normalized_Entite')['Montant_Amende'].sum().reset_index()
54 entities_highest_fines = entities_total_fines.sort_values(by='Montant_Amende', ascending=False).head(5)
55
56 # Identification of sectors with the highest total fine amount
57 sectors_highest_fines = df.groupby('Broad_Sector')['Montant_Amende'].sum().nlargest(5)
58
59 # Analysis of the most frequent shortcomings
60 shortcomings = df['Manquements'].str.split('\n').explode().value_counts().head(5)
61
62 # Data Visualization :
```





# Data Analysis & Visualization

We elaborate graphs for each analysed axe

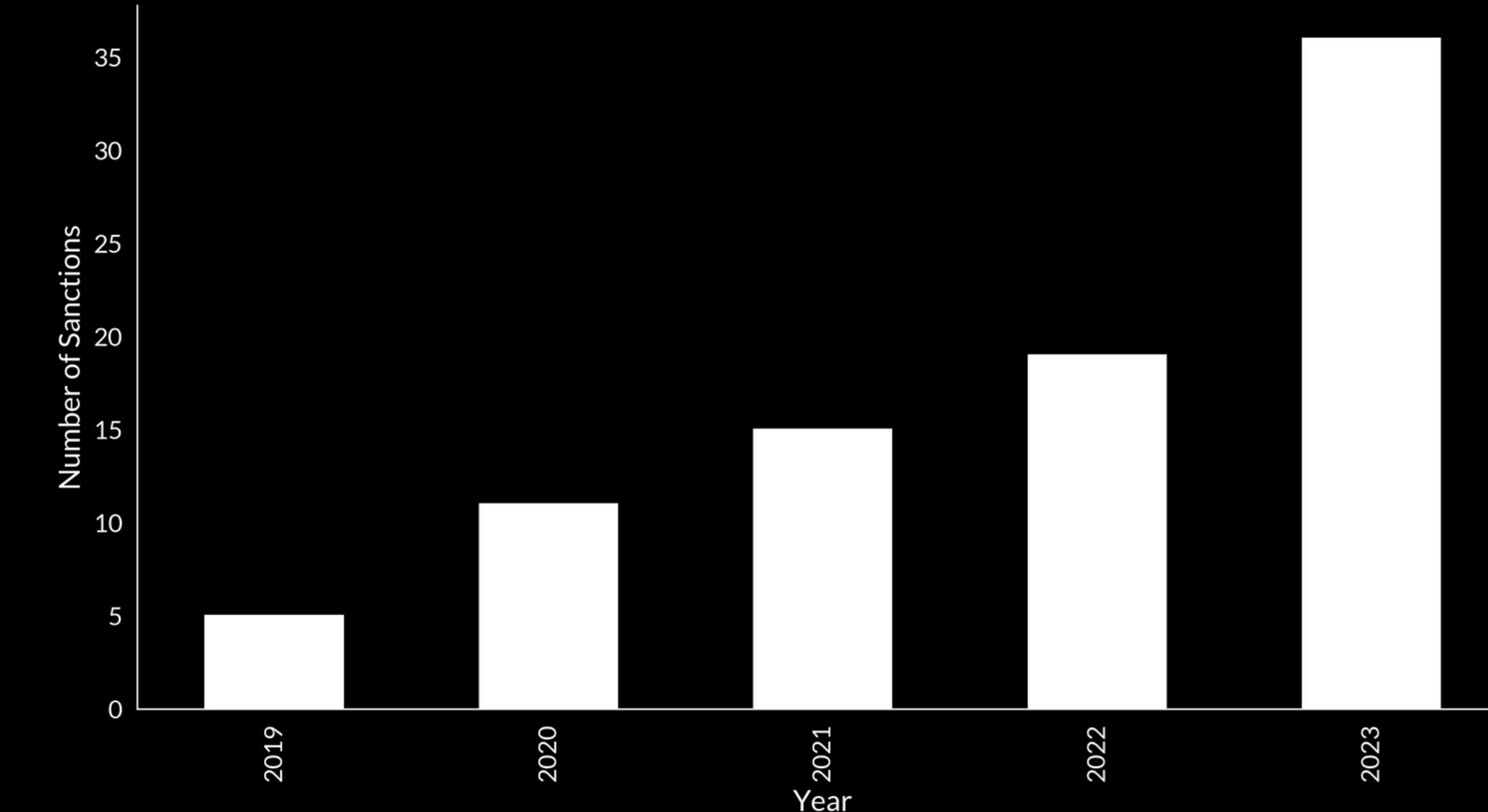
● ● ●

```
...hort countings = df['Mannements'].str.split(';').explode().value_counts() ...  
DF_ANALYSIS.py  
62 # Data Visualization :  
63  
64 # a. Number of sanctions issued by the CNIL each year since 2019  
65 plt.figure(figsize=(10, 6))  
66 sanctions_per_year.plot(kind='bar', color='white')  
67 plt.title('Number of Sanctions Over the Years')  
68 plt.xlabel('Year')  
69 plt.ylabel('Number of Sanctions')  
70 plt.tight_layout()  
71 plt.savefig('a-Number of sanctions issued by the CNIL each year since 2019.png', facecolor='black', dpi=dpi, bbox_inches='tight')  
72  
73 # b. Top 5 most frequently sanctioned Industries  
74 plt.figure(figsize=(10, 6))  
75 most_sanctioned_sectors.plot(kind='barh', color='white')  
76 plt.xlabel('Number of Sanctions')  
77 plt.ylabel('Sector')  
78 plt.tight_layout()  
79 plt.savefig('b-Top 5 most frequently sanctioned Industries.png', facecolor='black', dpi=dpi, bbox_inches='tight')  
80  
81 # c. Types of procedures used  
82 plt.figure(figsize=(10, 6))  
83 procedures_per_year.plot(kind='bar', stacked=True, color=['white', 'grey'])  
84 plt.xlabel('Year')  
85 plt.ylabel('Number of Procedures')  
86 plt.tight_layout()  
87 plt.savefig('c-Types of procedures used.png', facecolor='black', dpi=dpi, bbox_inches='tight')  
88  
89 # d. Top 5 Industries with the highest fines
```

5.

# Results

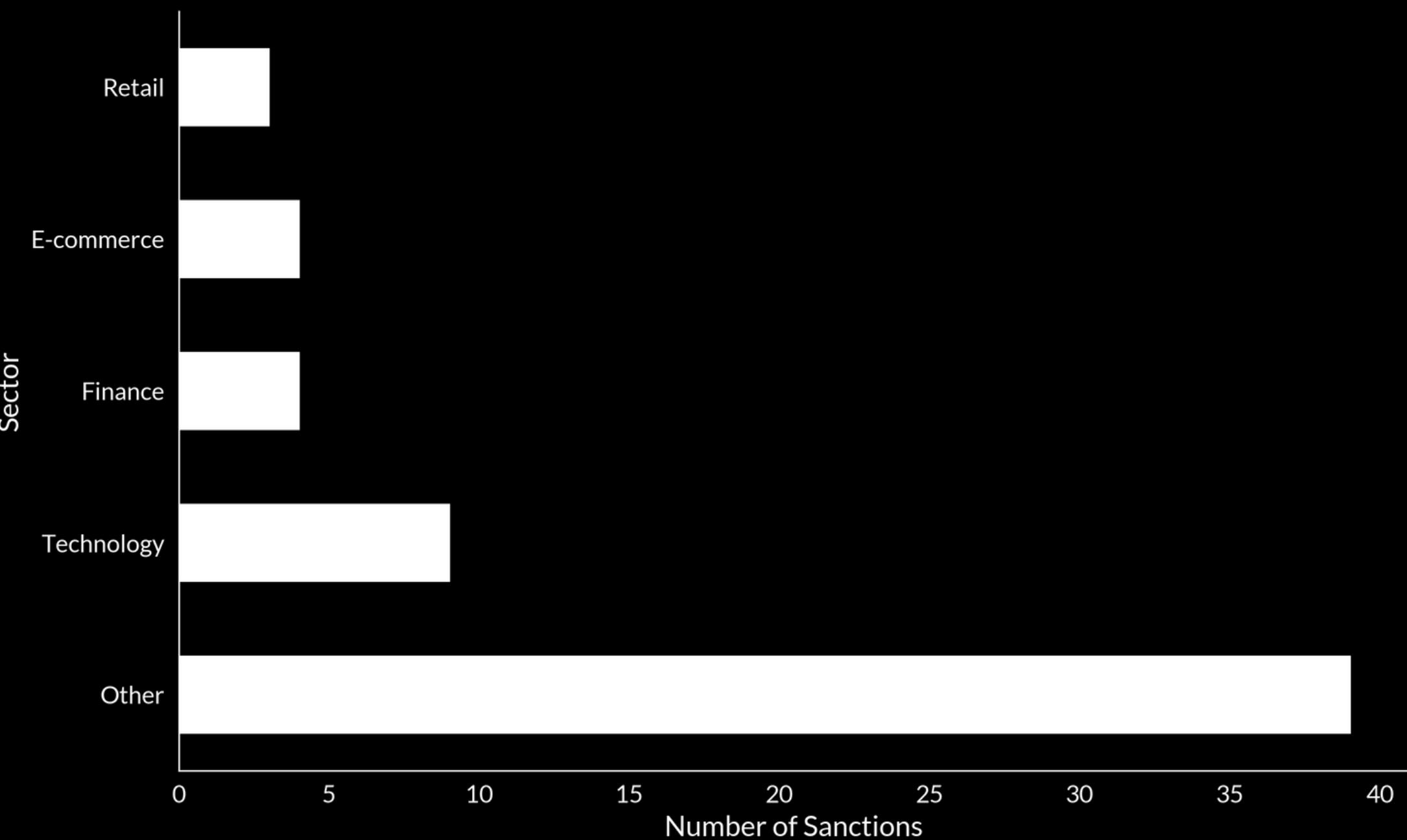
Number of sanctions  
issued by the CNIL each  
year since 2019



5.

# Results

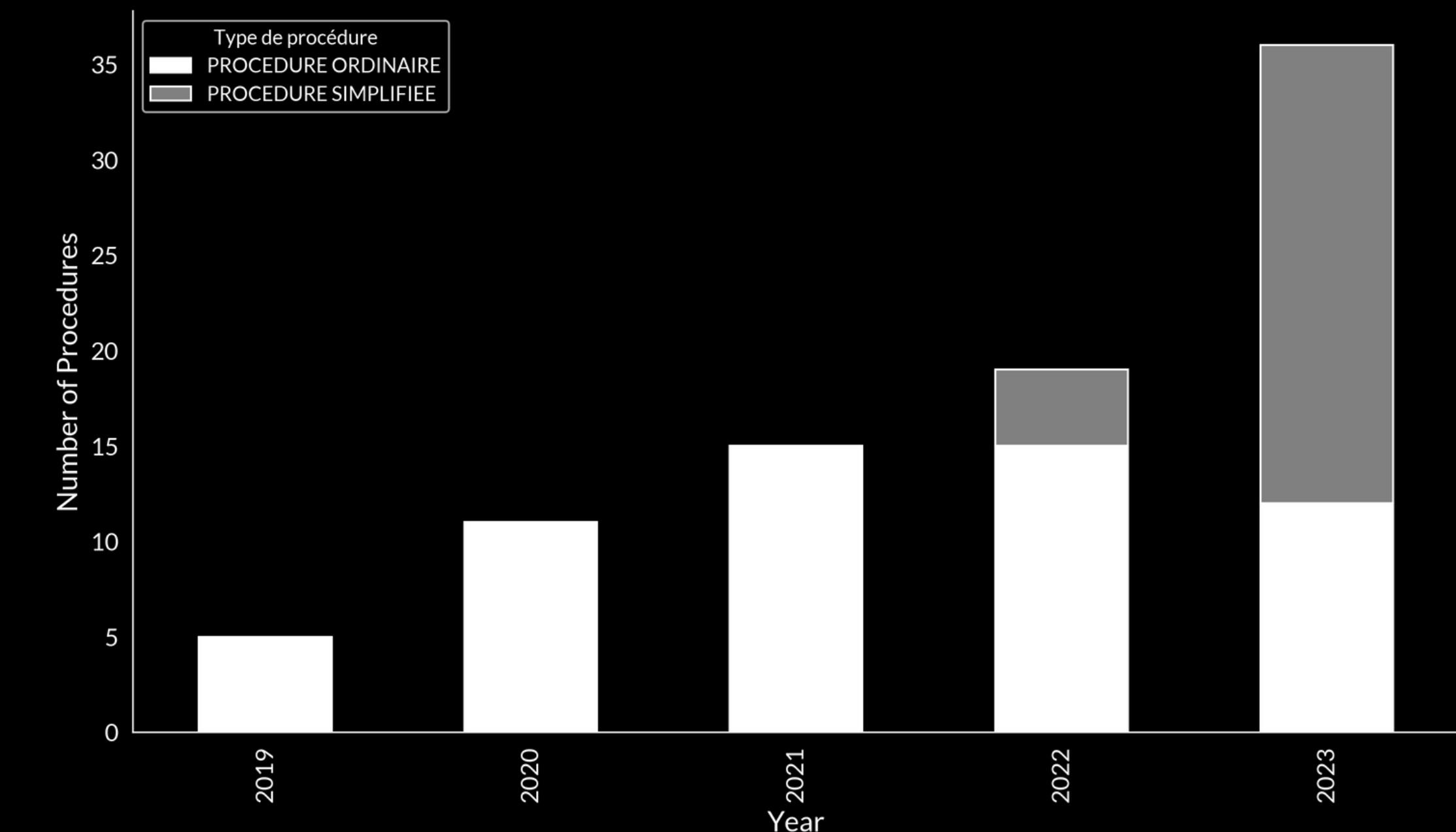
Top 5 most frequently sanctioned industries



# 5.

# Results

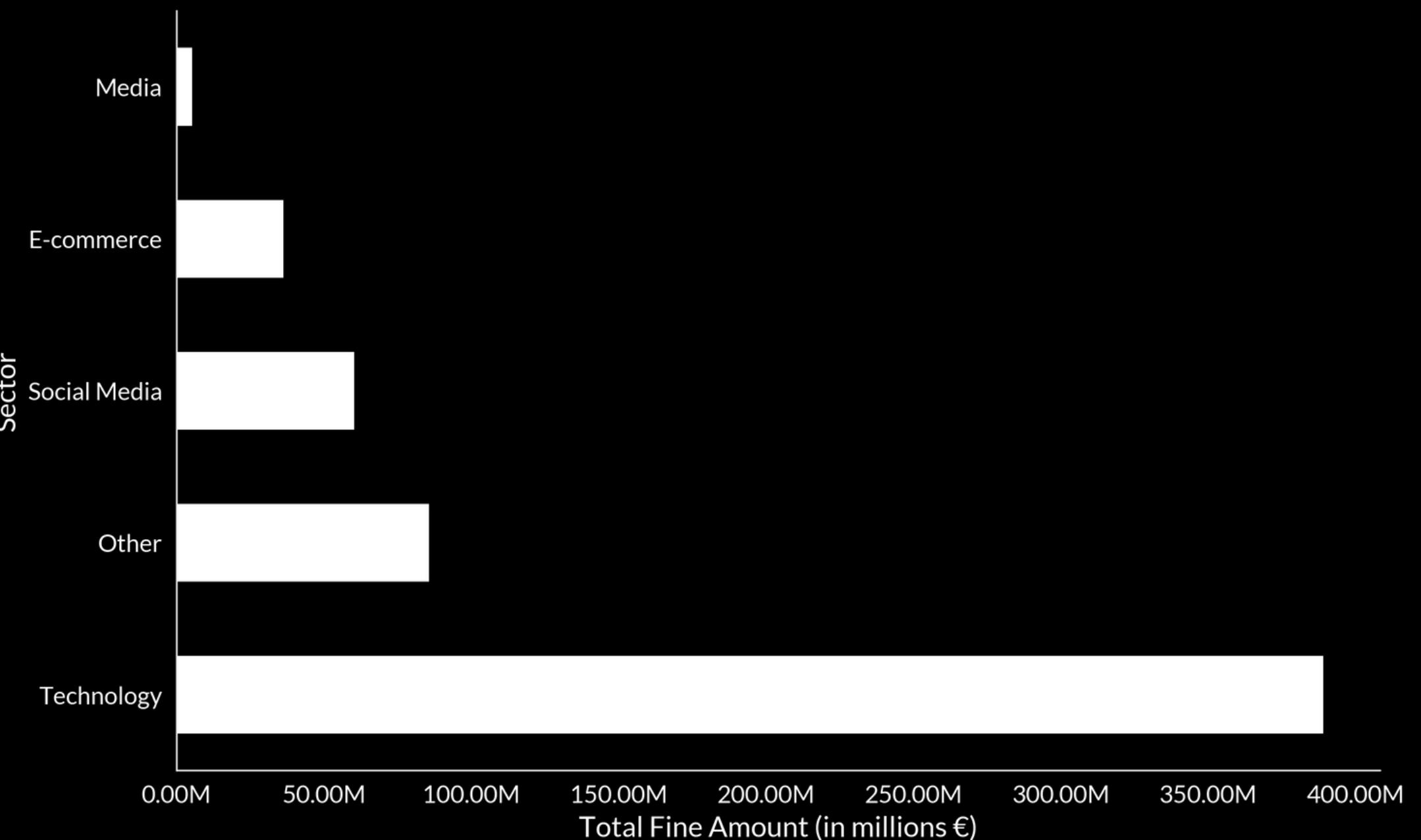
Types of procedures  
used by the CNIL



5.

# Results

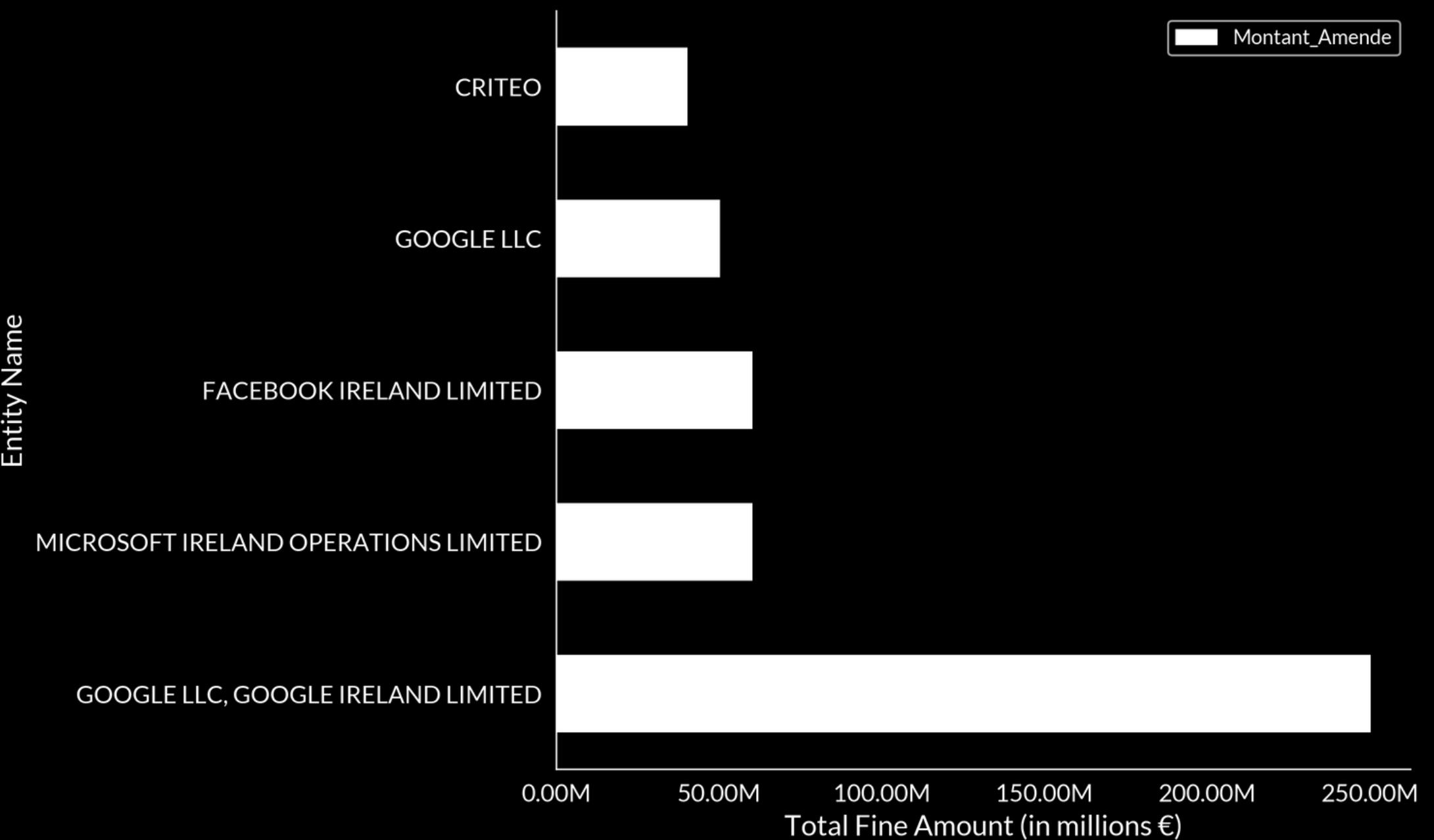
Top 5 Industries with  
the highest fines



5.

# Results

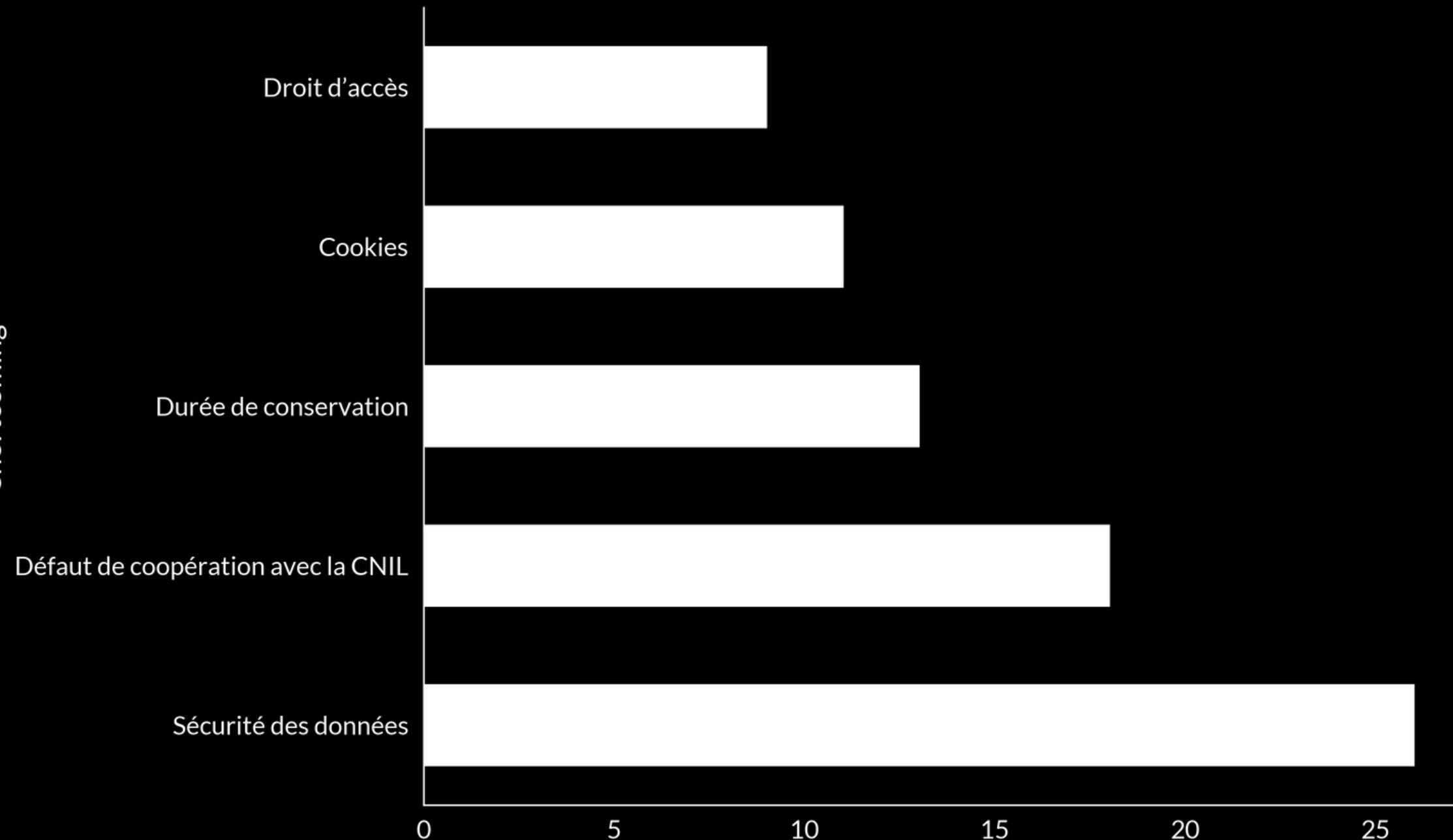
Top 5 entities with the highest fines



# 5.

# Results

Top 5 most frequent breaches



# 6.

## What's next

We want to go further

### Improving the robustness of the code

- Ensuring the code remains clear and easily adaptable
- Implementing Neural Language Processing (NLP)

### Ensuring on real time data retrieval

- Automating the web scraping process
- Providing notifications whenever a sanction is published

### Creating an online website / dashboard

- Creating a comprehensive map of all sanctions
- Facilitating interactive research for legal services

# Let's stay in touch



Jorge  
Portocarrero



Edouard  
Tan