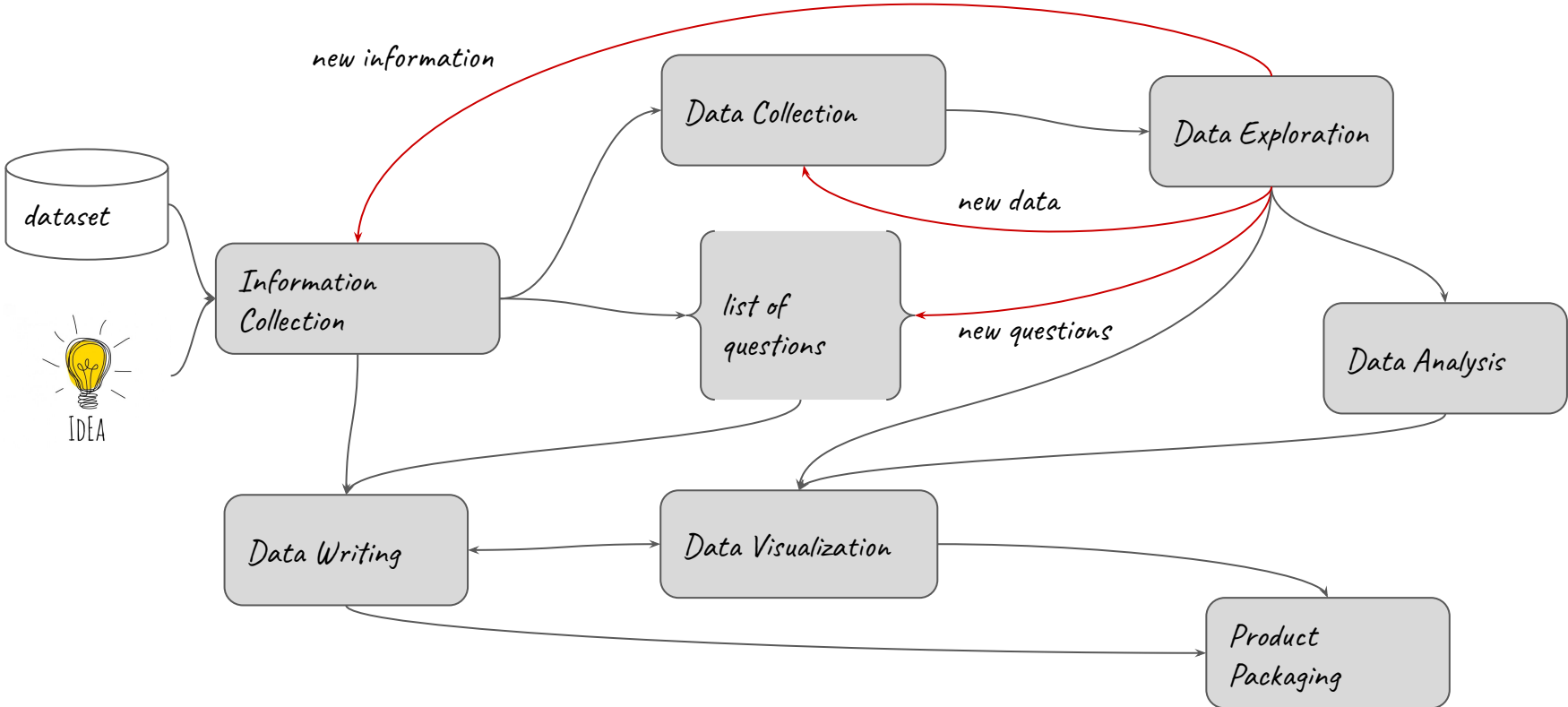


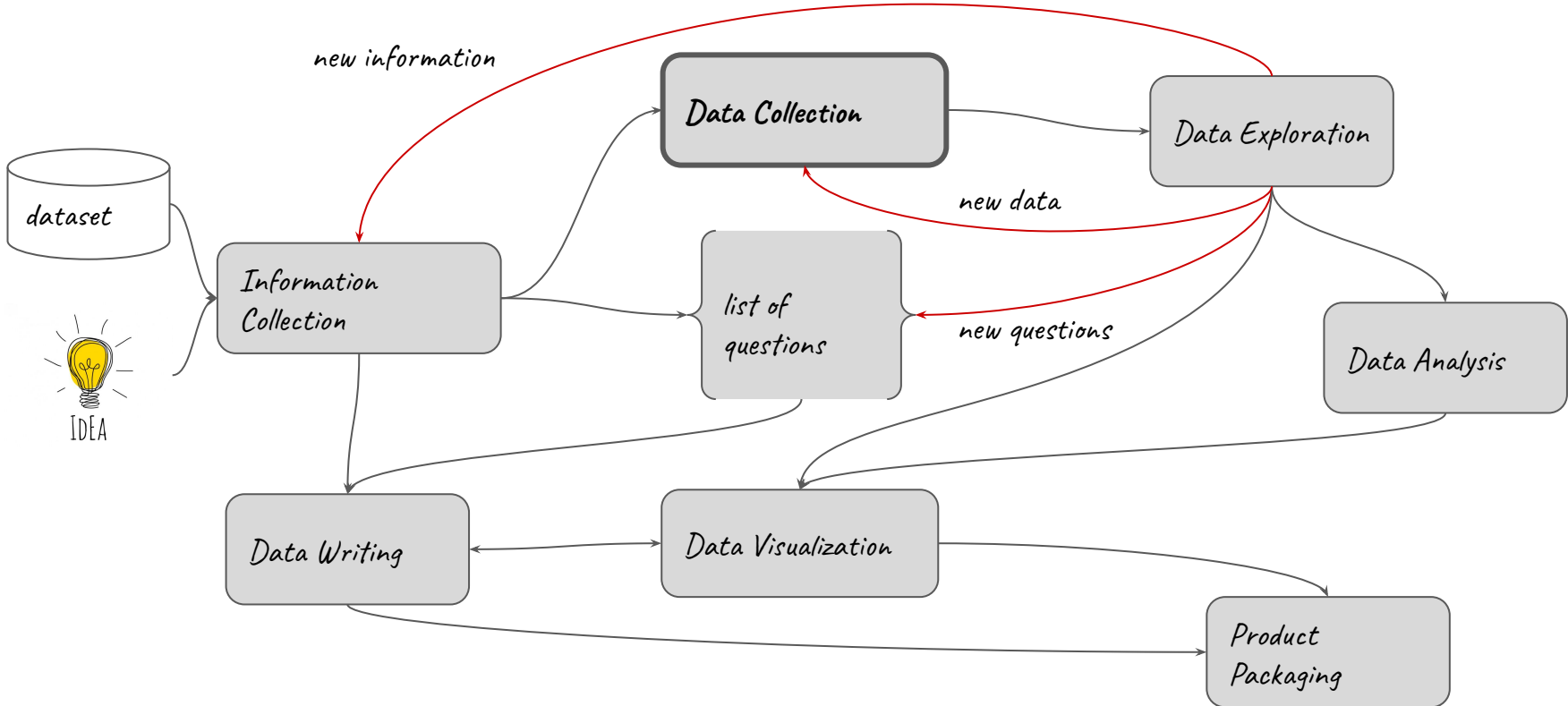
# Estrazione dati da PDF

Angelica Lo Duca  
angelica.loduca@iit.cnr.it

# Data Journalism Workflow



# Data Journalism Workflow



# Licenze dei Dati

# Tipi di Licenza

- Il **Public Domain** afferma che i dati sono liberi di essere utilizzati, distribuiti e modificati senza restrizioni. Questo è il tipo di licenza più permissivo ed è spesso utilizzato per i dati creati dalle agenzie governative o altre entità pubbliche.
- Le licenze **Open Data** richiedono tipicamente che i dati siano disponibili al pubblico gratuitamente e in un formato leggibile dalle macchine.
- La **Free for non-commercial use** ci consente di utilizzare i dati solo per scopi non commerciali.
- La **free academic license** consente l'utilizzo dei dati solo per scopi di ricerca accademica ed educativa.

# Tipi di Licenza

- Una **paid license** richiede il pagamento per l'utilizzo dei dati. I termini e le condizioni della licenza, come l'importo da pagare, la durata della licenza e gli utilizzi consentiti dei dati, sono specificati nell'accordo di licenza.
- Le licenze **Creative Commons** sono disponibili in diversi tipi, tra cui l'Attribuzione (CC BY), l'Attribuzione-Condividi allo stesso modo (CC BY-SA) e l'Attribuzione-Non opere derivate (CC BY-ND).
  - Ogni tipo di licenza ha requisiti diversi, come ad esempio dare credito al creatore originale e permettere ad altri di utilizzare, modificare e distribuire i dati.
  - La più comune è la licenza Creative Commons Zero (CC0). Questa licenza consente di utilizzare i dati per qualsiasi scopo senza attribuzione.

# Tipi di Licenza

La **Public Domain Dedication and License** (PDDL) è simile alla CC0 ma richiede di rinunciare ai nostri diritti morali sui dati. Ciò significa che non possiamo rivendicare la proprietà dei dati o dire di averli creati.

Estrazione da PDF



# Estrazione dati da PDF

- Approccio Manuale
  - copio la tabella
- Approccio Automatico
  - uso un software di conversione da PDF a Word
  - scrivo un programma per l'estrazione della tabella

# Estrazione manuale

- Si può utilizzare con tabelle singole e semplici
- Strumenti necessari
  - editor di testo
  - foglio di calcolo per la verifica
- Esempio
  - [https://www.epicentro.iss.it/coronavirus/bollettino/Infografica\\_22marzo%20ITA.pdf](https://www.epicentro.iss.it/coronavirus/bollettino/Infografica_22marzo%20ITA.pdf)

# Sorveglianza Integrata COVID-19 in Italia

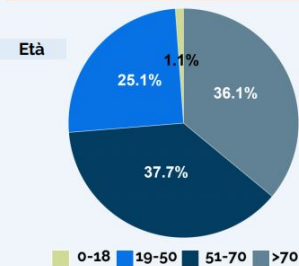
(Ordinanza n. 640 del 27/02/2020)

AGGIORNAMENTO 22 marzo 2020

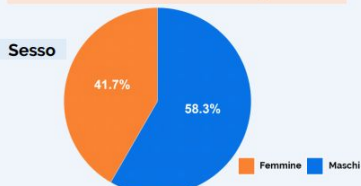
**52.796** casi di COVID-19\* di cui:

**4.824** operatori sanitari <sup>§</sup>

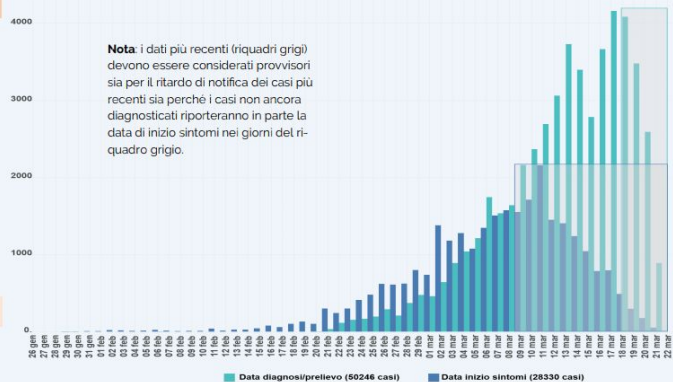
**4.465** deceduti



Età mediana dei casi: **63 anni**



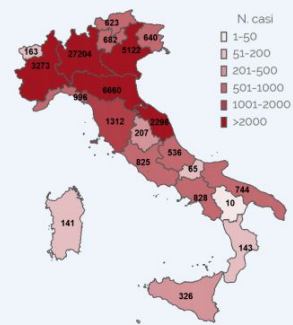
Fascia d'età (anni)	Deceduti [n (%)]	Letalità (%)
0-9	0 (0%)	0%
10-19	0 (0%)	0%
20-29	0 (0%)	0%
30-39	12 (0.3%)	0.3%
40-49	38 (0.9%)	0.6%
50-59	138 (3.1%)	1.3%
60-69	469 (10.5%)	5%
70-79	1585 (35.5%)	15.3%
80-89	1806 (40.4%)	23.3%
>=90	416 (9.3%)	24.1%
Non noto	1 (0%)	0.3%
Totale	4465 (100%)	8.5%



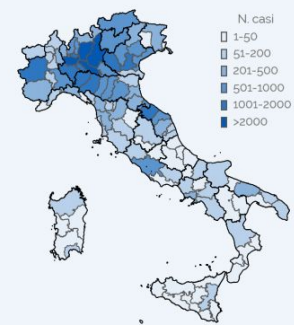
Sono risultati positivi il **99%** dei campioni processati dal Laboratorio nazionale di riferimento presso l'Istituto Superiore di Sanità



## Numero totale di casi di COVID-19 diagnosticati dai laboratori regionali di riferimento



per Regione/PA di diagnosi  
(dato disponibile per 52.796)



per Provincia di domicilio/residenza  
(dato disponibile per 50.235)

\*La definizione internazionale di caso prevede che venga considerata caso confermato una persona con una conferma di laboratorio del virus che causa COVID-19 a prescindere dai segni e sintomi clinici

<https://www.ecdc.europa.eu/en/case-definition-and-european-surveillance-human-infection-novel-coronavirus-2019-ncov>

\*Il flusso ISS raccoglie dati individuali di casi con test positivo per SARS-COV-2 diagnosticati dalle Regioni/PPAA. I dati possono differire dai dati forniti dal Ministero della Salute e dalla Protezione Civile che raccolgono dati aggregati. § Dato non riferito al luogo di esposizione ma alla professione.

# Procedimento

- Con il mouse selezionare tutte le celle della tabella
- Copiare la selezione
- Incollare in un editor di testo
- Sostituire il carattere di separazione (ad esempio lo spazio) con una virgola o un punto e virgola
- Correggere eventuali errori
- Salvare il file come csv
- Aprire il file con un programma per i fogli di calcolo (ad esempio Microsoft Excel)

# Estrazione automatica - convertitori

- Esistono diversi strumenti (anche online)
  - Esempio [Adobe Acrobat online](#)
- Sono rapidi
- Spesso non risolvono il problema, semplicemente convertono il file in un altro formato

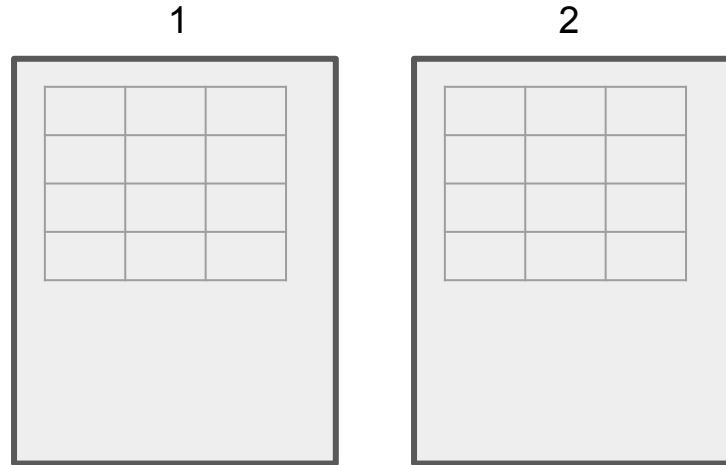
# Estrazione automatica - scrittura programma

- Si usa per tabelle complesse o ripetute
- Strumenti richiesti
  - python
  - libreria tabula-py
    - pip install tabula-py
    - pip3 install tabula-py

# Requisiti

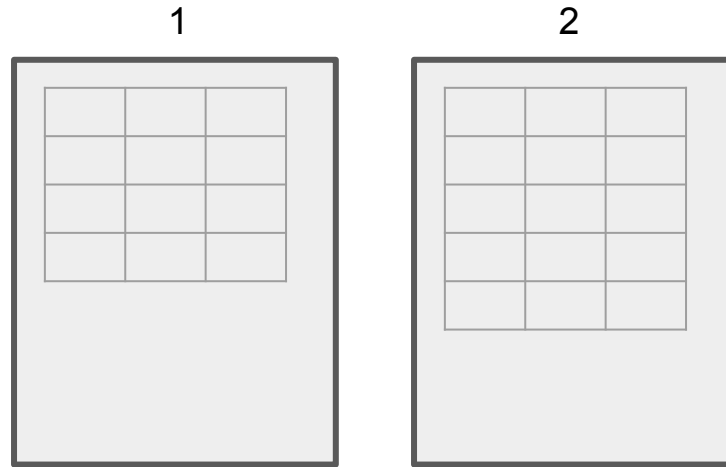
- Le tabelle devono presentare la stessa struttura in tutte le pagine
  - Stesso numero di colonne
  - Più o meno stesso numero di righe
- Le tabelle devono occupare più o meno la stessa posizione nella pagina

# Stessa struttura e posizione su piu' pagine

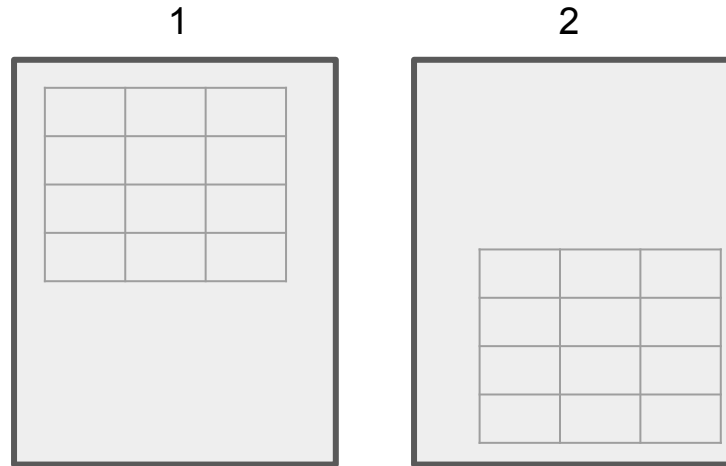




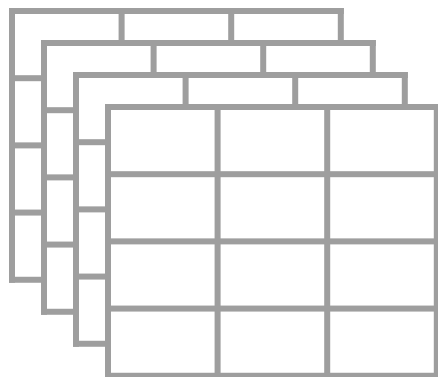
# Struttura simile e stessa posizione su piu' pagine



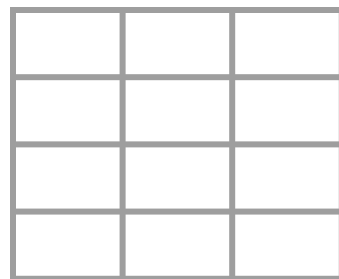
# Stessa struttura e posizione diversa su piu' pagine



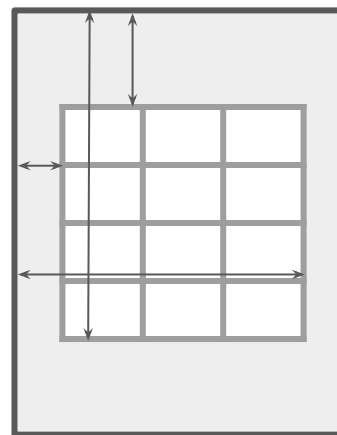
# Estrazione dati da PDF



**Documento PDF**  
con tante tabelle  
ripetute su piu'  
pagine



**Seleziono una  
singola pagina**  
con una singola  
tabella



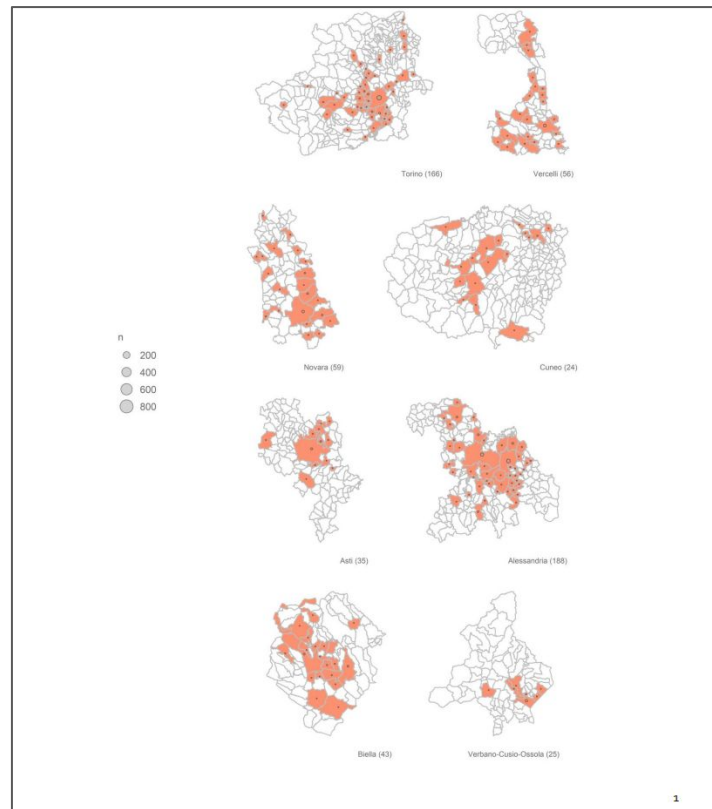
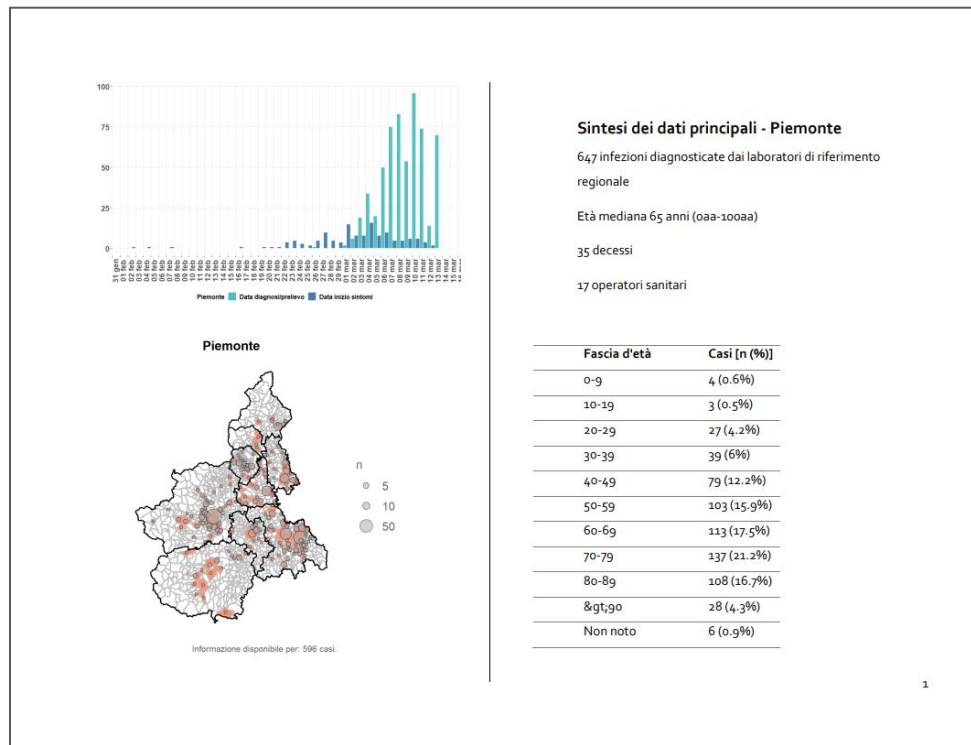
**Calcolo la posizione**  
della tabella nella  
pagina



**Do in input al  
programma il  
PDF e i  
margini**

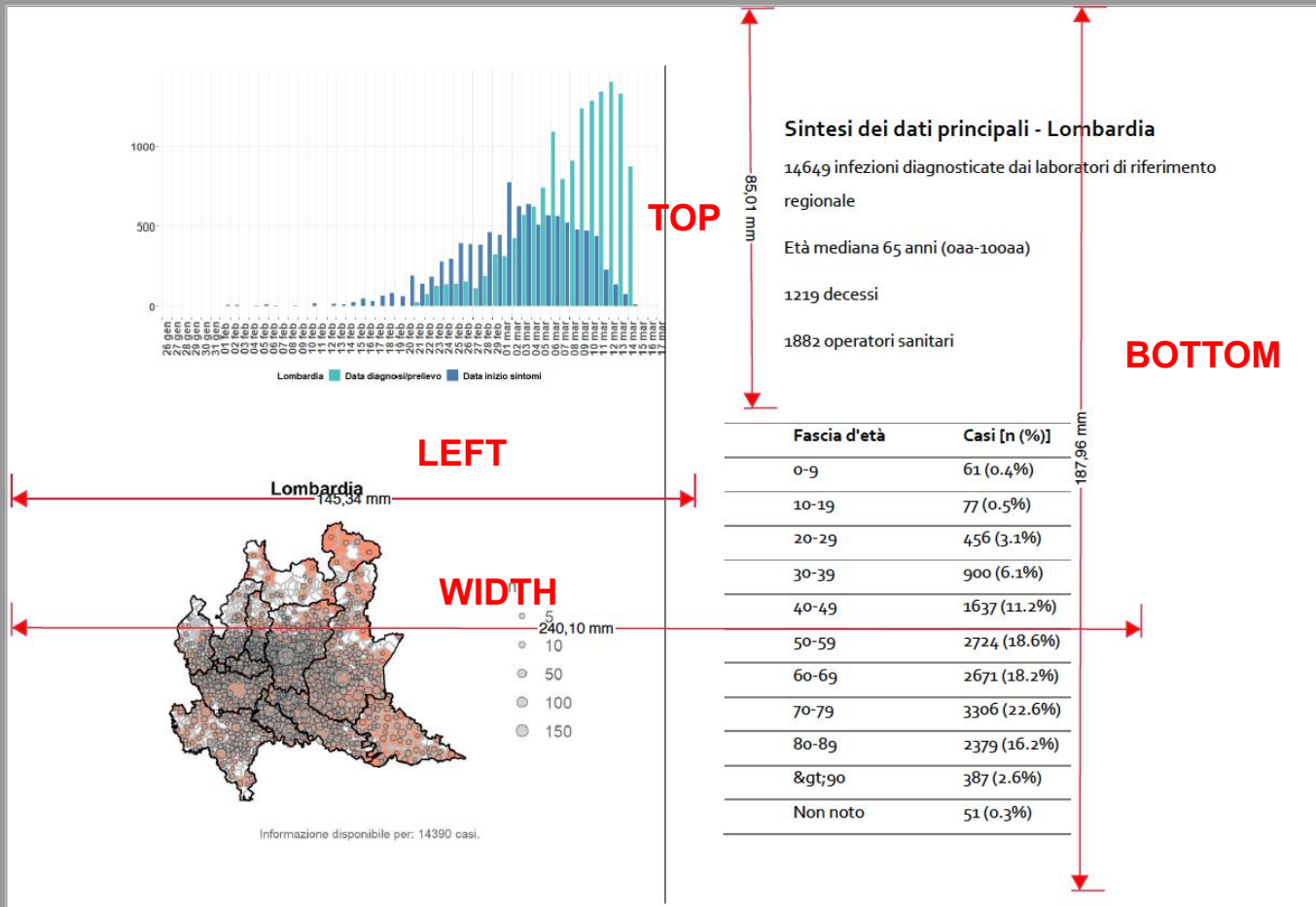
# Esempio - Bollettino Epidemia Covid-19

Due tipi di pagine



# Passi

Selezionare i margini della tabella attraverso un visualizzatore di PDF, come ad esempio Adobe Reader



# Passi (cont.)

Selezionare le pagine da cui estrarre le tabelle

3,5,6,8,9,10,12,14,16,18,22,24,26,28,30,32,34,36,38,40

Utilizzare la libreria tabula-py

<https://tabula-py.readthedocs.io/en/latest/>

# tabula\_py

Richiede l'installazione di Java 8+

Per verificare se Java è installato:

```
java -version
```

Per scaricare Java:

<https://www.java.com/en/download/manual.jsp>

# tabula-py functions

```
import tabula

# Read pdf into list of DataFrame

dfs = tabula.read_pdf("test.pdf", pages='all')

# convert PDF into CSV file

tabula.convert_into("test.pdf", "output.csv", output_format="csv",
pages='all')

# convert all PDFs in a directory

tabula.convert_into_by_batch("input_directory", output_format='csv',
pages='all')
```



# Lattice VS Stream

## Lattice Mode

**lattice=True** forces PDFs to be extracted using lattice-mode extraction. It recognizes each cells based on ruling lines, or borders of each cell.

## Stream Mode

**stream=True** forces PDFs to be extracted using stream-mode extraction. This mode is used when there are no ruling lines to differentiate one cell from the other. Instead, it uses spacings among each cells to recognize each cell.

Estratto da [Parse PDF Files While Retaining Structure with Tabula-py](#)

# Lattice VS Stream

a	b	c
d	e	f
g	h	i

Tabella con bordi → Lattice True  
`tabula.read_pdf(...,lattice=True)`

a	b	c
d	e	f
g	h	i

Tabella senza bordi → Stream = True  
`tabula.read_pdf(...,stream = True)`

# Documentazione tabula-py

<https://github.com/chezou/tabula-py>