

Data Journalism Workflow

Andrea Marchetti

Domanda iniziale

Se vi do un dataset, avete già una storia?

Problematizzare il dato

Dataset $\neq$ Realtà

I dati sono una rappresentazione della realtà non la realtà

Ogni dataset è una scelta

Ogni scelta implica esclusione

Dato pensato da altri

qualcuno ha deciso cosa misurare

qualcuno ha deciso come definirlo

qualcuno ha deciso come registrarlo

qualcuno ha deciso cosa escludere



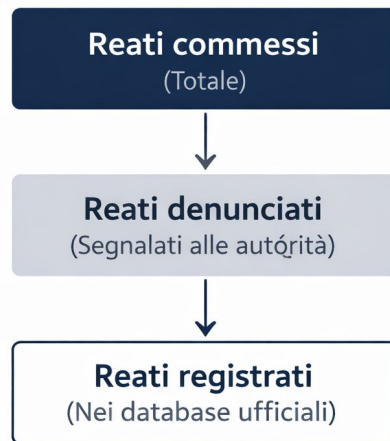
ESEMPIO: Criminalità

Se conto le denunce che arrivano
cosa sto misurando?

Il numero di reati commessi?

Il numero di reati denunciati?

Il dato misura ciò che è registrabile
non necessariamente ciò che
esiste



Cos'è l'epistemologia

L'epistemologia è la disciplina filosofica che studia:

- che cos'è la conoscenza
- come la produciamo
- quali sono i suoi limiti
- quando possiamo dire che qualcosa è “vero”

L'epistemologia si chiede **come sappiamo ciò che sappiamo.**

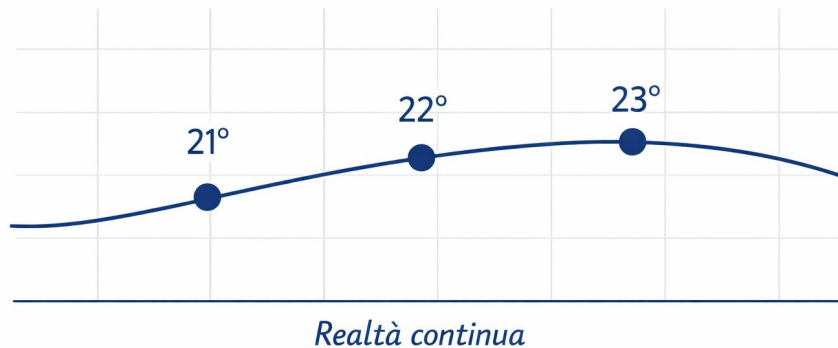
epistēmē = “conoscenza” + lógos = studio

Epistemologia del dato

Consapevolezza critica del limite del dato

Il dato è una discretizzazione della realtà

Pensate alla temperatura: 22° è una approssimazione



I due ingressi nel Data Journalism

Due punti di partenza

Question-driven

Ho una domanda interessante che ha bisogno di dati



Dataset-driven

Ho un dataset interessante (non aggregato) che ha bisogno di domande



Una storia che ha bisogno di dati

Cartomanzia online - Il business dell'incertezza

Perché nei momenti di crisi (economica, sanitaria, affettiva) aumenta il ricorso a cartomanti?

Quanto vale il mercato sommerso della cartomanzia oggi?

Quanto fatturano i consulti telefonici / online?

Chi sono i clienti più frequenti? (Giovani? Professionisti? Persone fragili?)

Ci sono meccanismi di fidelizzazione o dipendenza?



Un interessante dataset che ha bisogno di una storia

Quanto giocano gli italiani

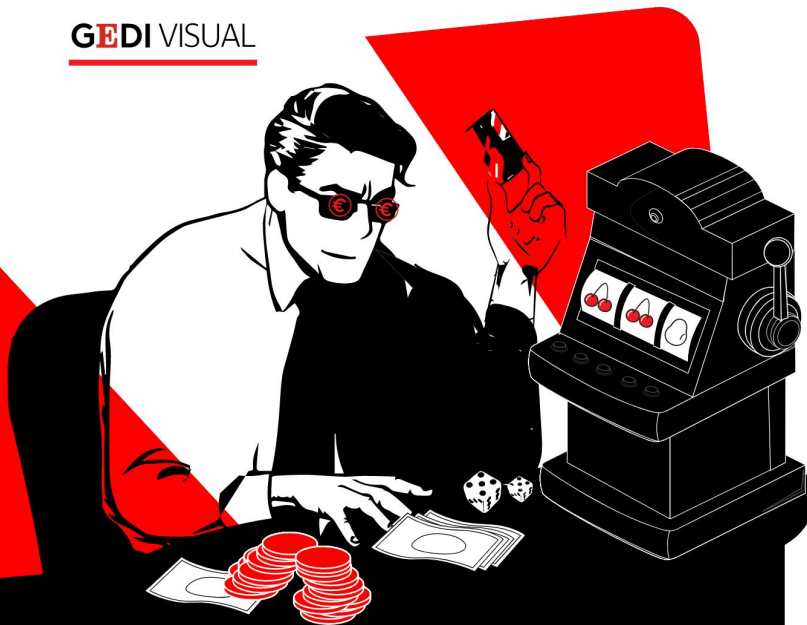
GEDI VISUAL

L'ITALIA DELLE SLOT 2

QUANTO GIOCANO GLI ITALIANI

Dal Superenalotto al Gratta&Vinci,
dal Bingo alle Slot: cerca quanto
pesa la febbre del gioco nel tuo
comune

CERCA IL TUO COMUNE



Question-driven journalism

Vantaggi

- Focalizzazione del problema

- Ipotesi verificabili

- Piano di lavoro più efficiente

- Coerenza narrativa

Rischi

- Confirmation bias

Beffa delle teste di Modigliani

La leggenda: Si credeva che Modigliani avesse gettato alcune sculture nel canale di Livorno nel 1909.



Domanda fatta: Le statue possono essere state prodotte nel 1909?

Confirmation Bias

Domanda corretta: A che epoca risalgono le teste ritrovate?

vere.

La confessione: Si scoprì che una testa era stata realizzata da tre studenti universitari (Pietro Luridiana, Pierfrancesco Ferrucci, Gherarducci) e le altre due dal pittore livornese Angelo Froglia.



Dataset-driven journalism

Vantaggi

I dataset aiutano a quantificare i fenomeni e sostenere affermazioni con numeri verificabili

Si possono scoprire dei trend interessanti o delle anomalie (outlier)

Rischi

Racconti ciò che è misurabile non ciò che conta

Problemi di bias dei dati

Discussione

Quale approccio è più rischioso

Modello pratico: Question-first, data-led

Domanda guida (neutra e testabile)

Quanto è diffuso il fenomeno X? Dove? Da quando? Con quali eccezioni?

Fonti che evidenziano il fenomeno

Dataset che possono rispondere e cosa non possono dire

Ricerca di altri fonti (non-dati) come interviste, osservazioni sul campo

Esplorazione del dataset

Ricerca di pattern/anomalie, ma trattali come ipotesi, non conclusioni

Modello pratico: Question-first, data-led

Domande di controllo (anti bias)

Che spiegazioni alternative ci sono?

Quali casi contrari dovrei vedere se la mia ipotesi fosse falsa?

Verifica incrociata

testimonianze, esperti, sopralluoghi

Storia con trasparenza

Cosa dicono i dati, cosa non dicono, quali assunzioni hai fatto

II Workflow sistemico

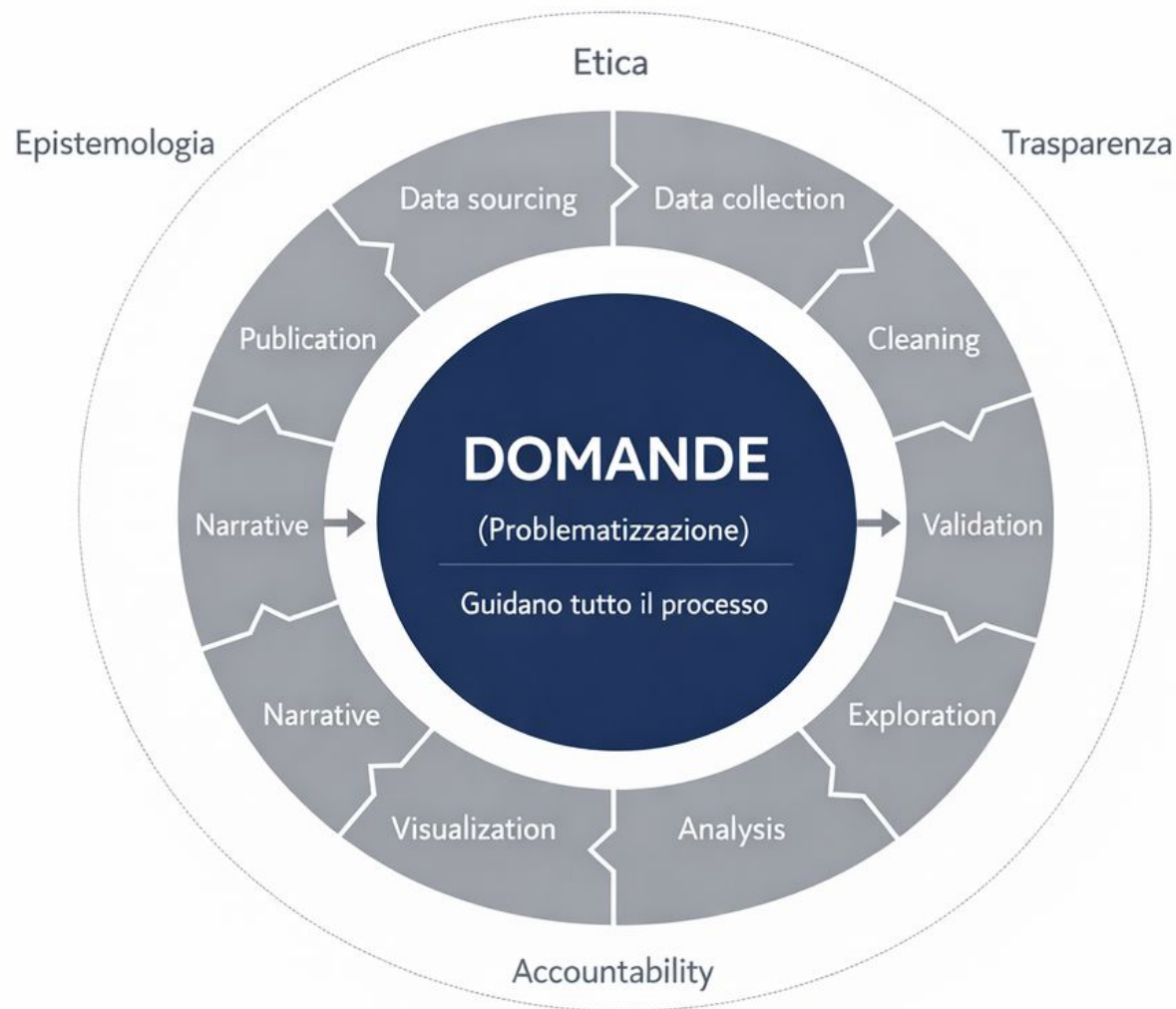
Data Journalism Workflow

Modello iterativo

Livello 1 - domande

Livello 2 - ciclo operativo

Livello 3 - livello critico

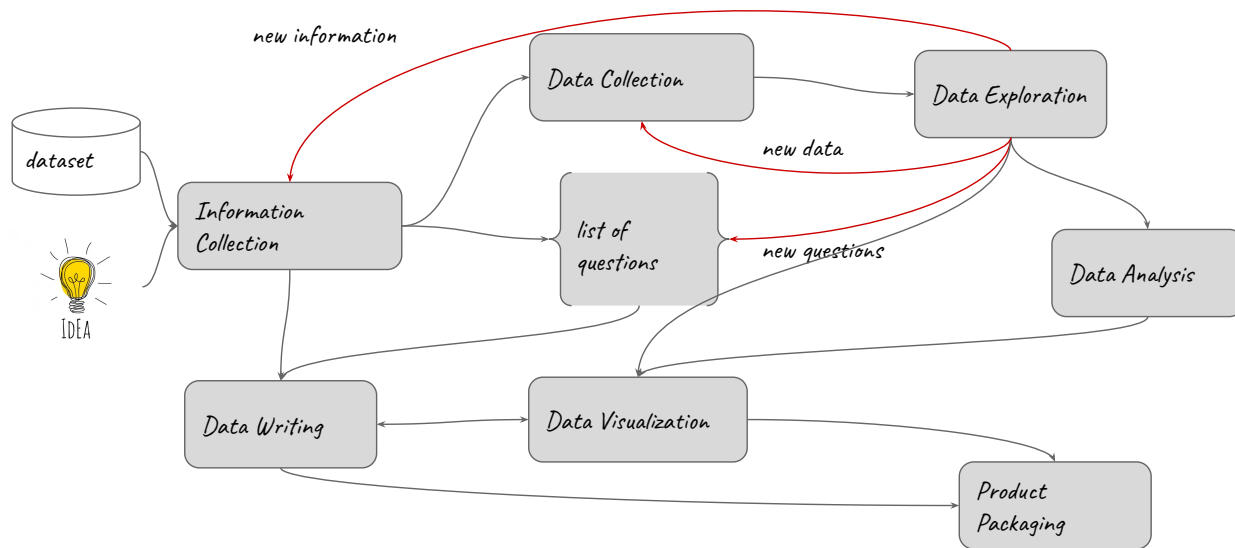


Il Data Journalism NON è lineare

Non è una pipeline in cui da una fase si passa alla successiva

Possiamo ritornare sulle varie fasi

Le domande iniziali possono essere ripensate



Livello 1 - Domande

Al centro ci sono le domande

Le domande determinano quali dati cercare

Le domande cambiano durante il ciclo operativo dei dati in particolare
dopo l'esplorazione e l'analisi

Tipi di domande

Secondo le tre scale: spazio, tempo, categoria

Generali Vs specifiche

	Generale	Specifica
Spazio	Italia	Comune di Pisa
Tempo	2011-2021	25 febbraio 2026
Categoria	Tutta la popolazione	Maschi 15-19

Diventare esperti di dominio: il caso cartomanzia

Concetti chiave	Lessico tecnico	Definizioni operative (misurabili)
Cartomanzia come pratica divinatoria	Tarocchi, Arcani, Stesa, Consulto	Numero di operatori registrati (ATECO), numero di siti/app attive
Mercato della cartomanzia	Linea premium, Revenue share, Call center esoterico	Fatturato dichiarato, bilanci societari, traffico telefonico a tariffa maggiorata
Modello di business	Costo al minuto, Abbonamento, Crediti, Conversione, Retention	Prezzo medio $\times$ durata media consulto $\times$ numero stimato di consulti
Vulnerabilità dei clienti	Cliente fidelizzato, Lead generation, Lifetime value	Frequenza consulti per utente, spesa cumulativa, intervallo medio tra chiamate
Economia sommersa	Intermediazione, società estera, pubblicità ingannevole	Differenza tra fatturato dichiarato e stima mercato, flussi finanziari indiretti
Impatto sociale	Dipendenza, fragilità, tutela consumatore	Correlazione con indicatori di crisi (disoccupazione, inflazione, pandemia)

Domande centrali

Quanto è diffusa la cartomanzia?

Quanto vale il mercato?

Quanto spende un cliente medio?

Esiste dipendenza?

Aumenta nei periodi di crisi?

Chi sono i principali operatori?

Qual è il quadro normativo?

Per ognuna di queste domande devo definire una “definizione operativa” ovvero un modo concreto e misurabile

Es. per la prima domanda posso calcolare il numero di siti web attivi

Rischi epistemologici

Se non diventi esperto di dominio rischi di:

- Trattare tutto come folklore

“Se non conosci il lessico interno di un settore, non stai facendo data journalism.

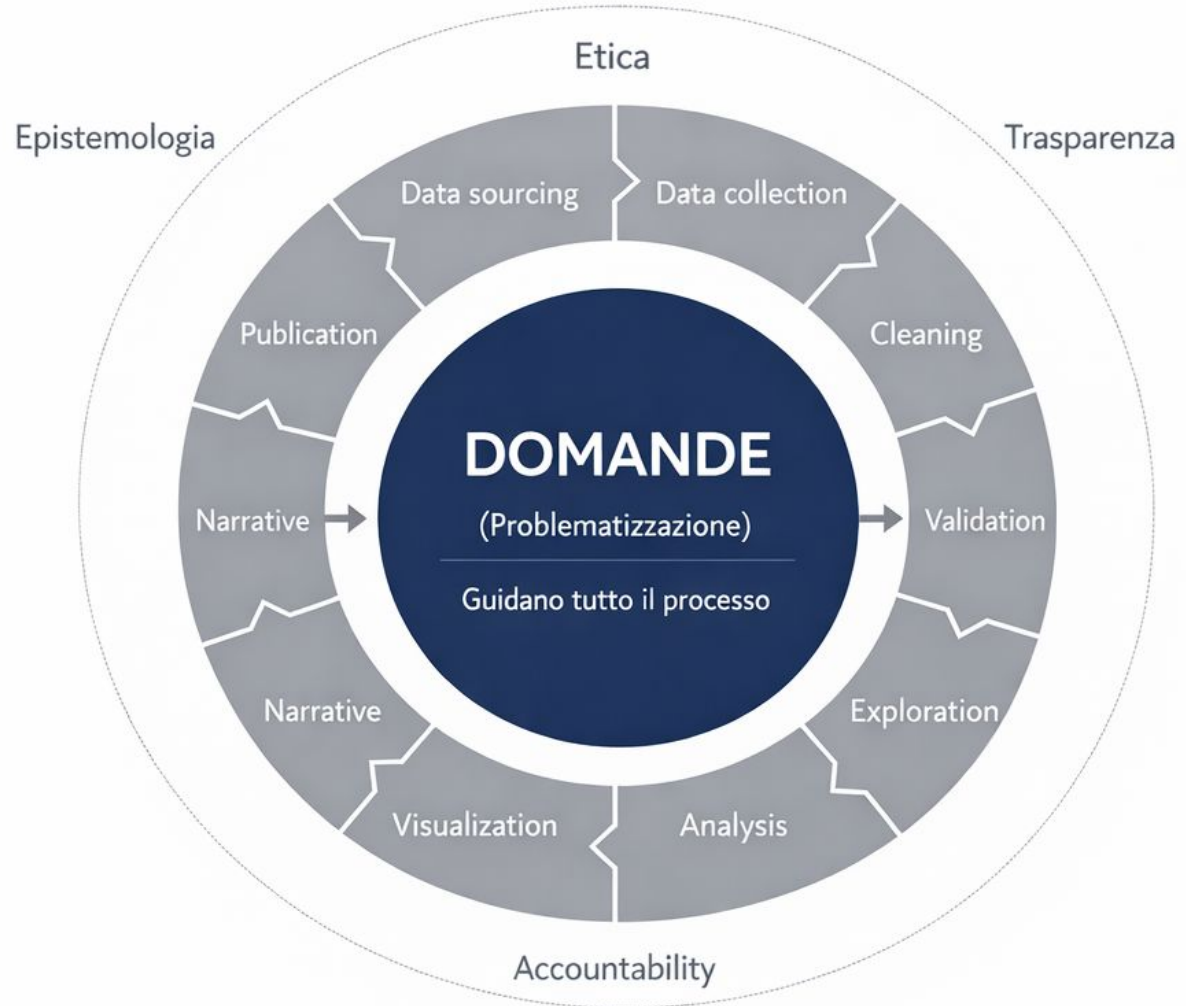
Stai facendo analisi superficiale di numeri.”

- Ignorare il quadro legale

Livello 2 - Fasi del DJ Workflow

Fasi del workflow

1. Data sourcing
2. Data collection
3. Cleaning
4. Validation
5. Exploration
6. Analysis
7. Interpretation
8. Visualization
9. Narrative integration
10. Publication



Data Sourcing

Identificazione e valutazione delle fonti di dati pertinenti rispetto alle domande.

- Ricerca fonti affidabili (fonti istituzionali, open data)
- Valutazione metodologia
- Selezione dati rilevanti

Data Collection

Acquisizione operativa dei dati dalle fonti selezionate.

Download / API / Web scraping

Strutturazione dataset

Documentazione processo

Data Cleaning

Correzione e normalizzazione dei dati grezzi.

Gestione missing values

Rimozione duplicati

Standardizzazione formati

Data Validation

Verifica di affidabilità e qualità dei dati.

Controllo plausibilità

Cross-check fonti

Verifica metodologia raccolta

Data Exploration

Analisi esplorativa orientata alla scoperta di pattern e anomalie.

Statistiche descrittive

Ricerca pattern

Generazione ipotesi e affinare le domande

Data Analysis

Analisi strutturata per testare ipotesi

Modelli statistici

Regressioni

Test di significatività

Produzione evidenza alle domande

Interpretation

Attribuzione di significato ai risultati.

Contestualizzazione storica e sociale

Valutazione implicazioni

Connessione con letteratura

Data Visualization

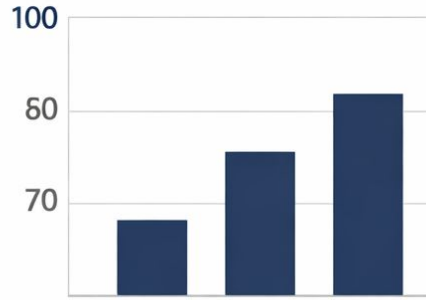
Rappresentazione visiva dei risultati.

Scelta grafico adeguato

Chiarezza visiva

Trasparenza dati

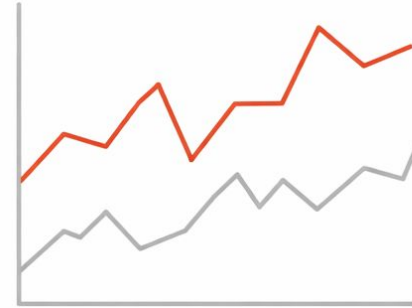
Errori di visualizzazione



Asse manipolato



Troppe categorie



Colore fuorviante

Data Storytelling/Narrative

Costruzione del racconto basato sui dati.

Struttura argomentativa

Integrazione evidenza

Chiarezza espositiva

Publication

Diffusione e documentazione del lavoro.

Pubblicazione online

Condivisione metodologia

Trasparenza e replicabilità

Livello 3: Livello critico

Livello 3 - Il livello critico

Attorno a tutto il sistema:

- Epistemologia
- Etica
- Riproducibilità
- Accountability

Questo livello “abbraccia” tutto.

Perché non è una fase finale.

È una condizione permanente.

Etica: Attenzione ai dati sensibili

Dati personali: qualsiasi informazione che consente di identificare una persona (es. Nome e cognome, data di nascita, codice fiscale, targa, indirizzo IP, dati di localizzazione, ...)

Dati sensibili: sottoinsieme dei dati personali che se divulgati possono creare rischi per la persona come discriminazione, esclusione sociale/lavorativa, furto d'identità, .. (es. origine etnica, opinione politica, convinzione religiosa, dati genetici, dati biometrici, dati sulla salute, orientamento sessuale, ...)

General Data Protection Regulation: Regolamento (UE) 2016/679

Anonimizzazione

Per la pubblicazione di un'intervista devi avere il consenso della persona

Riproducibilità / Replicabilità

- Dichiarare le fonti
- Trasparenza metodologica
- Fornire il codice utilizzato

un risultato è riproducibile se un'altra persona, usando gli stessi dati e lo stesso metodo/codice, ottiene lo stesso risultato

- **Riproducibilità:** stessi dati + stesso codice → stessi output
- **Replicabilità:** dati nuovi + stesso metodo → conclusioni simili

Accountability

Il giornalista è responsabile delle decisioni prese e deve spiegarli e documentarli

La riproducibilità aiuta l'accountability perché rende verificabile ciò che dichiarare

Riproducibilità senza accountability: un'analisi è riproducibile, ma nessuno è formalmente responsabile delle scelte/impatti.

Accountability senza riproducibilità: un ente risponde delle decisioni, ma non può (o non vuole) rendere l'analisi verificabile (dati riservati, processi opachi, ecc.).

Caso studio: Mortalità ISTAT

Il dataset pubblicato da ISTAT

[Decessi e cause di morte: cosa produce l'Istat](#)

Decessi anni 2011-2025 - Dati provvisori

[Dataset con i decessi **giornalieri** \(formato .csv\) in ogni singolo **comune** di residenza per sesso e classi di età quinquennali | gennaio 2011 – settembre 2025](#)

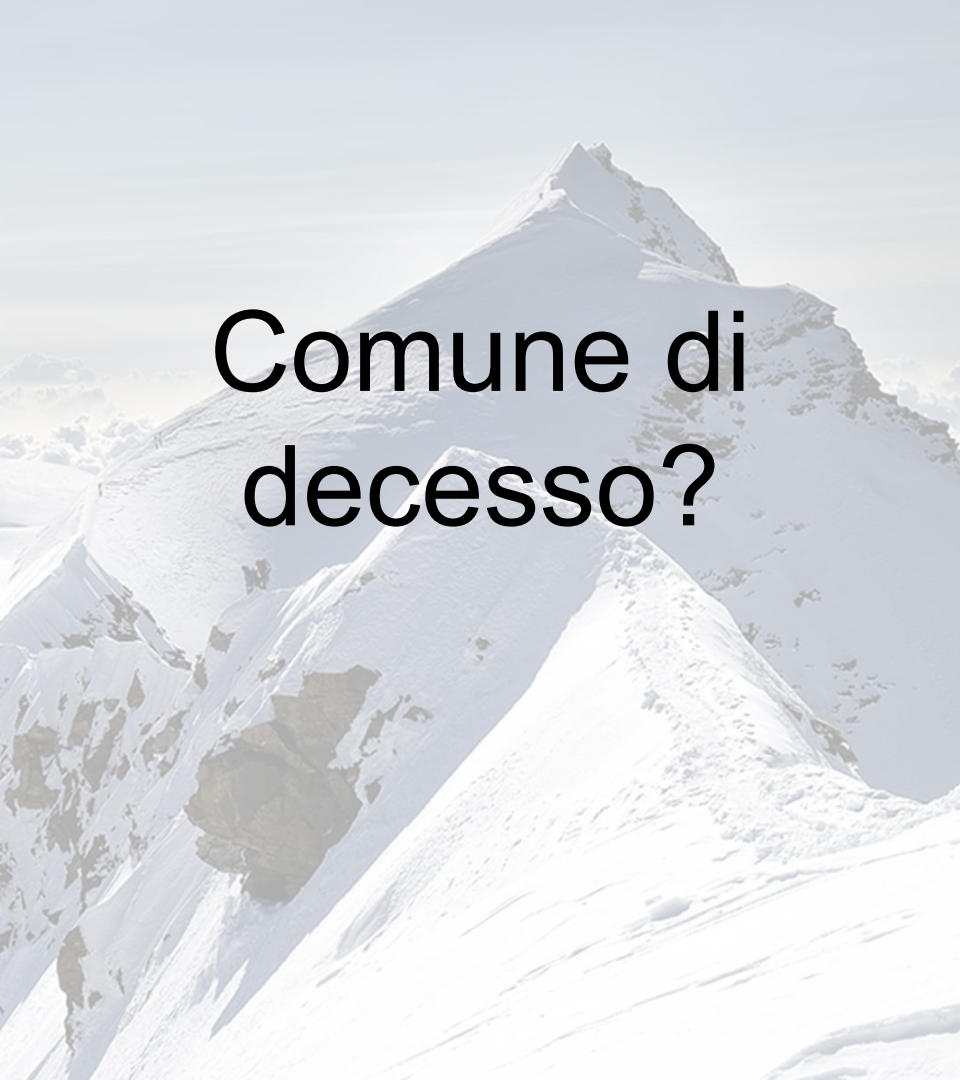
Fonti Istat per i decessi giornalieri

1. **Anagrafe Nazionale della Popolazione Residente** ([ANPR](#))

Anagrafe unica a livello nazionale che raccoglie i dati dei cittadini residenti in Italia

2. Ministero dell'economia e delle finanze per l'acquisizione del flusso dei deceduti tramite l'[Anagrafe Tributaria](#)

Usato durante la fase pandemica



Comune di decesso?

Comune dove è avvenuto l'evento

Comune di residenza

Questo non è così ovvio, ad esempio per la nascita avviene l'opposto, si registra il comune dove avviene l'evento e non il comune di residenza (dei genitori)

Cosa sembra dirci il dataset?

Granularità temporale = giorno

Granularità spaziale = comune

Copertura spaziale = tutti i 7896 comuni di Italia = Italia

Copertura temporale = gennaio 2011- settembre 2025 = circa 14 anni

~7900 Comuni



22 Classi di età

- 0=0
- 1=1-4
- 2=5-9
- 3=10-14
- 4=15-19
- 5=20-24
- 6=25-29
- 7=30-34
- 8=35-39
- 9=40-44
- 10=45-49
- 11=50-54
- 12=55-59
- 13=60-64
- 14=65-69
- 15=70-74
- 16=75-79
- 17=80-84
- 18=85-89
- 19=90-94
- 20=95-99
- 21=100+

Formato CSV

Nel comune di Agliè il 15 aprile del 2013 è morto un maschio di età compresa tra 1 e 4 anni

Formato CSV

01,001,Piemonte,Torino,Agliè,001001,1,6,1018,0,0,0,0,0,0,0,0,0,0,0,0,0,0,**n.d.**,0,0
 ,1,0,0,0,0,0,0,0,0,0,0,0,**n.d.**,0,0,1,0,0,0,0,0,0,0,0,0,0,0,**n.d.**

n.d. Dipende dal fatto che il file è aggiornato al 25 settembre 2025 quindi per tutte le righe che gestiscono un giorno dopo il 25 settembre i campi M25,F25 e T25 non possono essere definiti fino a quando non sono stati raccolti i dati fino al 31 dicembre 2025.

Cosa NON possiamo sapere?

Nome della persona

Età precisa della persona

Ora del decesso

Indirizzo del decesso

La causa di morte

Contesto demografico (questo si può ricavare)

...

Confronto con l'Europa

E il mondo



Fichier-des-personnes-decedees 1970-2025

nomprenom;sexe;datenaiss;lieunaiss;commnaiss;paysnaiss;datedeces;lieudeces;actedeces

MACAPAGAL*RODOLFO/;1;19650920;99220;ANGELES CITY;PHILIPPINES;;20210912;75107;164

MENCHETTI*GIAMPIERO/;1;19471004;99127;BARCHI;ITALIE;;20210913;75107;165

DE CHABOT*ISABELLE MARIE-THERESE FRANCOISE/;2;19390107;75107;PARIS 7E
ARRONDISSEMENT;;20210914;75107;168

ANTONIO*RUI MANUEL/;1;19730317;71014;AUTUN;;20210908;75107;167

DELAGOUTTE*BERNARD/;1;19280314;75058;LE PERREUX-SUR-MARNE;;20210909;75107;163

BELMONDO*JEAN PAUL CHARLES/;1;19330409;75051;NEUILLY-SUR-SEINE;;20210906;75107;160

MOREAU*MICHELINE MADELEINE/;2;19210929;75104;PARIS 4E ARRONDISSEMENT;;20210906;75107;161

GILARDOT*FRANCOISE/;2;19520706;33522;TALENCE;;20210917;75107;170

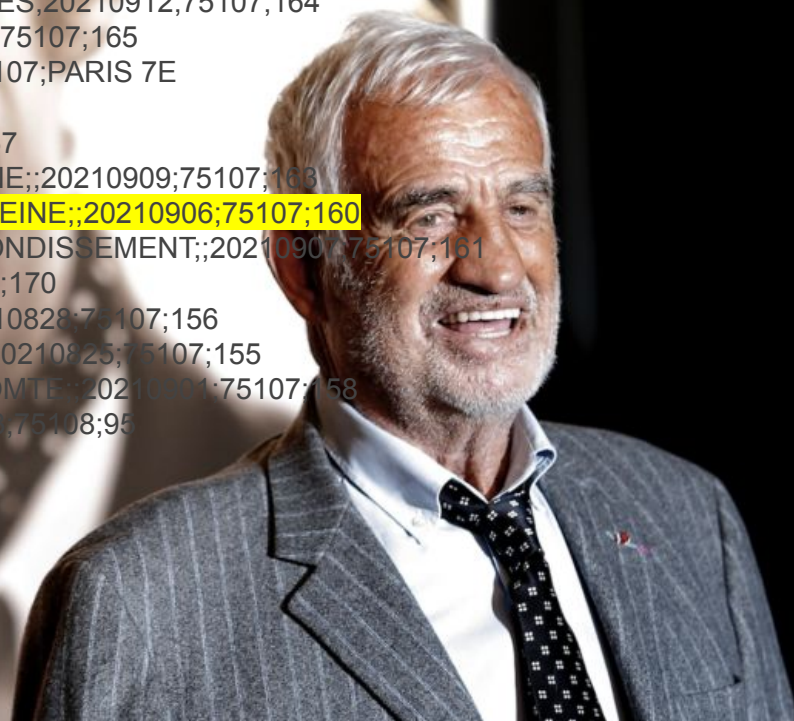
BEAU*PHILIPPE JACQUES RENE/;1;19280912;98506;CHIRONGUI;;20210828;75107;156

ROGGWILLER*DANIEL LOUIS ERNEST/;1;19280626;78582;SANNOIS;;20210825;75107;155

AUGER*MARIE THERESE ANNIE/;2;19200829;85092;FONTENAY-LE-COMTE;;20210901;75107;158

ISMERT*FRANCOIS LOUIS/;1;19460404;57606;SAINT-AVOLD;;20210903;75108;95

Jean Paul Charles Belmondo, nato a
Neuilly-Sur-Seine il 9 aprile 1933 è morto il 6
settembre 2021 a Parigi il 6 settembre 2021



Situazione in Europa

- Germania: 83,1 milioni [DESTATIS](#)
- Francia: 67,4 milioni [INSEE](#) dal 1970 ad oggi
- Regno Unito: 68,2 milioni [ONS](#)
- Italia: 60,5 milioni [ISTAT](#) dal 2011 ad oggi
- Spagna: 47,7 milioni [INE](#)

In sintesi, **solo la Francia rende disponibile pubblicamente un registro dei decessi**. Negli altri Paesi, i certificati di morte sono accessibili solo con restrizioni (parenti, interesse legittimo, o tramite richieste formali), mentre le statistiche aggregate sulla mortalità sono generalmente disponibili presso gli enti nazionali di statistica.

Numeri assoluti Vs Tassi

Bisogna avere il valore della popolazione dei comuni.

Per ogni anno considerato



Problemi metodologici

Bias di raccolta

Missing data

Aggregazioni: spaziali (comune), temporali (24 ore)

Come possiamo testare la validità di questi dati?

Per controllare la copertura del dataset potremmo verificare se sono riportati i dati di decessi certi

Frontale tra moto nel parcheggio, morti due giovani amici. L'ipotesi della sfida: «Stavano facendo acrobazie»

Le vittime avevano 16 e 17 anni, gli inquirenti ipotizzano che alla base dello schianto ci sia un gioco pericoloso

Data: mercoledì 17 settembre 2025

Luogo: Pisa (parcheggio IKEA)

Sesso: maschile

Età: 17 anni (classe età 4=15-19)

Nome: Jacopo Gambini

Data: venerdì 19 settembre 2025

Luogo: Pisa

Sesso: maschile

Età: 16 anni (classe età 4=15-19)

Nome: Leonardo Renzoni

Barbara non ce l'ha fatta: è morta la psichiatra aggredita al Santa Chiara

Nella serata di domenica 23 aprile è stata accertata la morte cerebrale: la dottoressa donerà gli organi. L'omicida è piantonato in carcere

Data: domenica 23 aprile 2023

Luogo: Pisa

Sesso: femminile

Età: 55 anni

Nome: Barbara Capovani

Conclusioni

Cosa abbiamo imparato

Non raccontare ciò che i dati dicono ma interrogare ciò che i dati nascondono

Verificare che le vostre curiosità non siano guidate da pregiudizi (bias)

Diventare esperti di dominio

Lavorare su ipotesi e cercare conferme intervistando esperti