

Data Collection - Open data

Andrea Marchetti

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

Esercizio (5')

Definire 3-5 domande generali/specifiche legate al dataset dei decessi

1. ???
2. ???
3. ???
4. ???
5. ???

Esercizio (10')

Impadronirsi dei seguenti concetti:

- Demografia
- Popolazione residente
- Tasso di mortalità,
- Aspettativa di vita alla nascita,
- Tavola di mortalità
- Eccesso di mortalità
- Stock vs Flusso

Demografia

La demografia studia le popolazioni umane: quante persone ci sono, come sono distribuite e come cambiano nel tempo. In pratica si occupa di:

- **Dimensione** della popolazione (quanti residenti, stock)
- **Struttura** (età, sesso, composizione per famiglie, cittadinanza, ecc.)
- **Dinamica** (i “flussi” che la fanno cambiare):
 - nascite (fertilità/natalità)
 - decessi (mortalità)
 - migrazioni (immigrazione/emigrazione, trasferimenti di residenza)
- **Distribuzione territoriale** (comuni, province, regioni; urbanizzazione)
- **Indicatori sintetici**: tassi di natalità/mortalità, fecondità, speranza di vita, crescita naturale e totale, ecc.

Demografia: Etimologia

Demografia viene dal greco:

- **dêmos** (δῆμος) = “popolo”, “popolazione”
- **-grafia** da gráphein (γράφειν) = “scrivere”, “descrivere/registrare”

Quindi letteralmente significa “**descrizione della popolazione**” (cioè misurare e descrivere come è fatta e come cambia una popolazione).

Demografia - Indicatori sintetici

Tasso di mortalità = decessi / popolazione $\times$ (1.000 o 100.000)

Limite: più anziani $\Rightarrow$ più decessi

Aspettativa/Speranza di vita alla nascita

Tavola di mortalità / sopravvivenza = (<https://demo.istat.it/app/?i=TVM>)
aspettativa di vita per ogni età

Eccesso di mortalità: differenza tra decessi osservati e “attesi” (media storica o modello). Serve per eventi tipo ondate di calore/epidemie.

Stock Vs Flusso

Stock è una quantità misurata in un istante (una “fotografia”).

Esempi:

- Popolazione residente al 1° gennaio
- Numero di detenuti in un giorno specifico
- Posti letto disponibili oggi

Si esprime come: “quanti ci sono a una data”.

Flusso è una quantità misurata su un intervallo di tempo (un “film”).

Esempi:

- Decessi/Nascite in un anno
- Ingresso/uscita da un carcere in un mese
- Migrazioni durante un anno

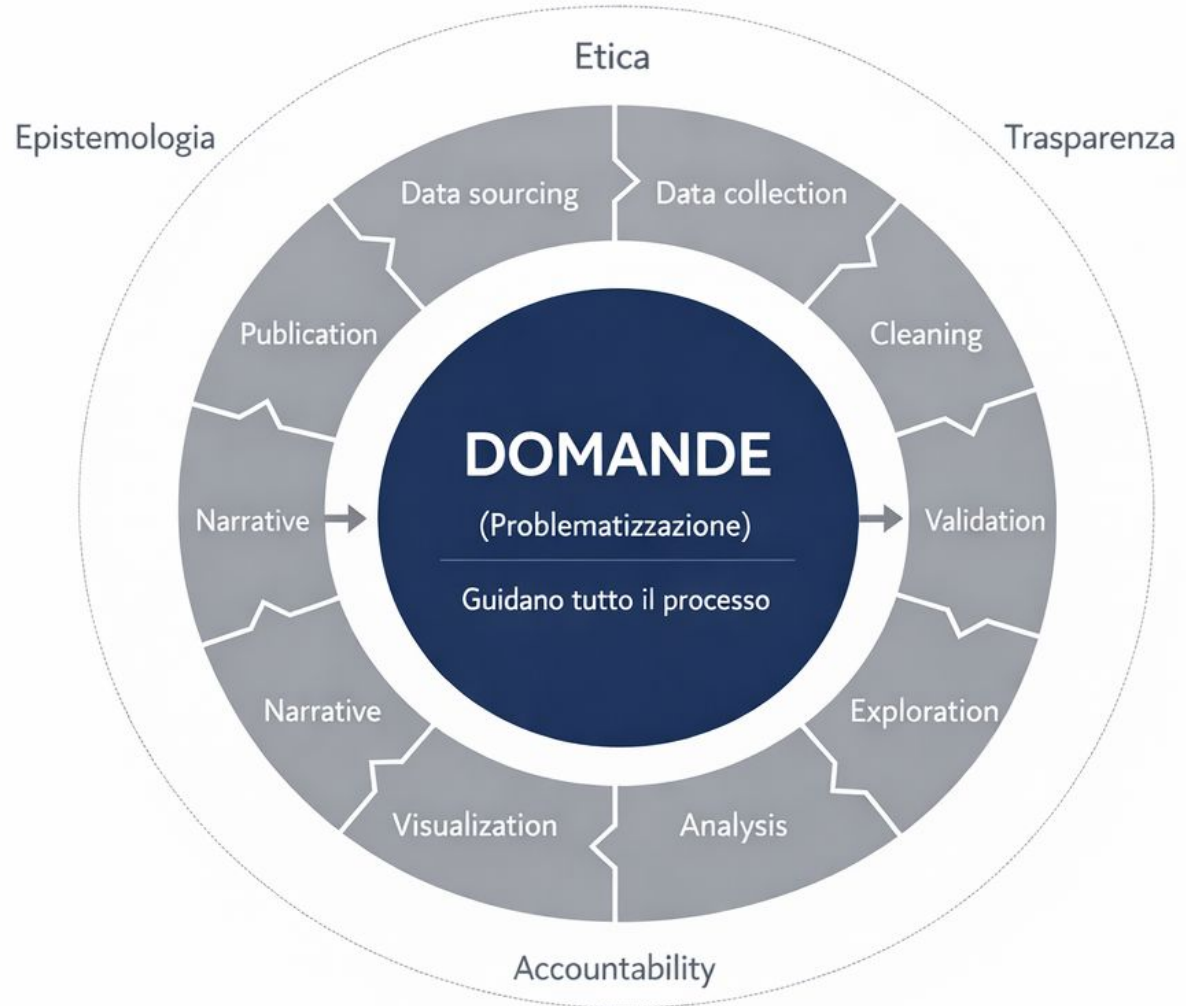
Si esprime come: “quanti eventi accadono tra due date”.

Esercizio (5')

Modificare le domande generali/specifiche legate al dataset dei decessi sulla base della conoscenza del dominio

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: principali metodi

1. **Download open data:** [data.gov](#), [istat](#), [zenodo](#), ...
2. Web Crawling through API: API ufficiali di servizi
3. **Web Scraping:** librerie per estrarre dati da pagine web
4. Estrazione da documenti: PDF, OCR/HTR di scansioni
5. Dati generati: survey, sensori
6. Richieste formali di dati: FOIA/accesso civico

1 Download open data

Cosa include

- Portali open data: data.gov, data.gov.it,
- Dataset statistici: istat, eurostat, onu, ocse
- Dataset terzi e repository:
Zenodo/Dataverse/Figshare, GitHub,
Kaggle/data.world

Perché conviene

- Dati strutturati a volte già pronti per l'uso,
- Autorevolezza e metodologia di acquisizione,
- Comparabilità nel tempo

Rischi

- Granularità limitata,
- aggiornamenti limitati

Output tipico

CSV, XSLX, GeoJSON, bulk

2 Web Crawling attraverso web api

Cosa include

- Molti servizi tra cui Spotify, Meta/Facebook, Google, X mettono a disposizione delle web api o dump massivi (bulk)

Perché conviene

- Dati aggiornati basati su comportamenti sociali,
- query mirate

Rischi

- Rate limit,
- Autenticazione,
- Campi che cambiano,
- Costi

Output tipico

Json, XML compressi

3 Web Scraping

Possibilità di usare librerie python per fare web scraping

Conviene quando i dati sono solo su pagine web: es. Tripadvisor, Booking, Amazon

- Pagine statiche → requests + (bs4 o lxml) / pandas.read_html
- Molte pagine → Scrapy
- Dinamico (JS) → **Playwright** (o Selenium)

Rischi:

- Fragilità,
- Codice complesso,
- Blocchi

4 Estrazione da documenti

Conviene

- Dati presenti solo su report PDF o su documenti acquisiti come immagini

Rischi

- Errori di estrazione,
- data cleaning/validazione

5 Dati creati da te (survey & osservazione)

Conviene

Quando i dati mancano

Rischi

- Bias campione (campioni limitati e non rappresentativi)
- Privacy/consenso
- Trasparenza metodologica

6 Richieste formali di dati

FOIA significa Freedom of Information Act: è il nome (nato negli USA) delle leggi che danno ai cittadini il diritto di chiedere documenti e dati alla pubblica amministrazione.

Conviene

- Quando esistono dei dati interessanti in possesso di un ente pubblico (es gioco d'azzardo)

Rischi

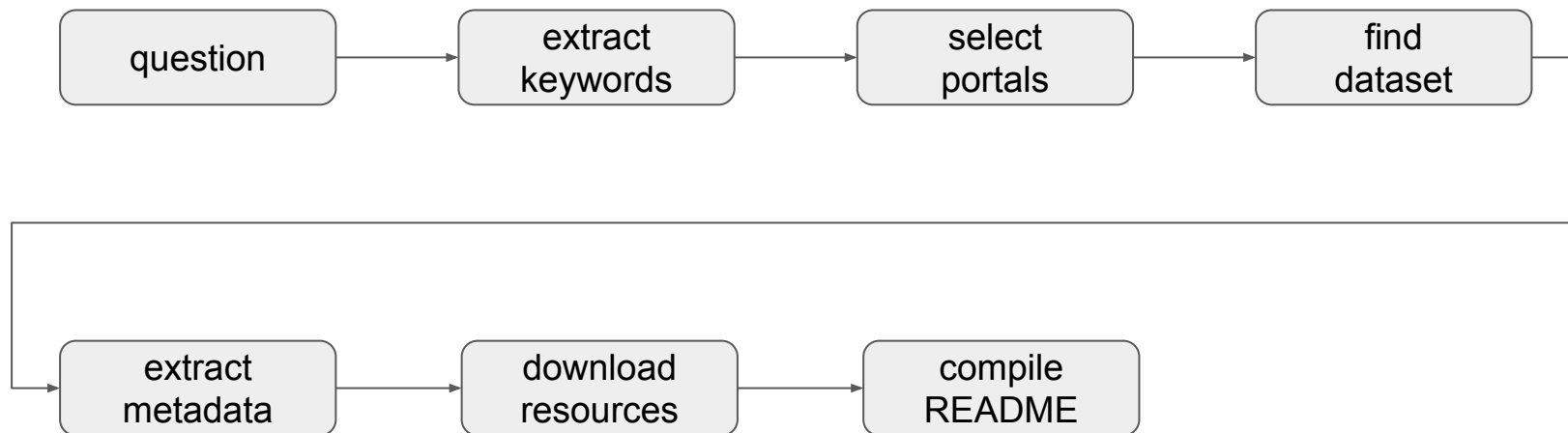
- Procedura lenta e laboriosa

Confronto

Metodo	Caratteristiche	Complessità
Open data download	(più pulito)	bassa
API	(più stabile, limitato)	media
Scraping	(più fragile)	alta
PDF/OCR/HTR	(più sporco)	media-alta
dati generati sondaggi	(più controllo, bias)	bassa
FOIA/partnership	(più lento ma potente)	media

Open data

Percorso operativo



Dalla domanda alle keyword

Domanda: es. andamento della mortalità in Italia

Sinonimi + termini amministrativi

(es. “mortalità”, “decessi”, “anagrafe”, “popolazione residente”, “comune”, “serie storica”)

Select Portals

Cataloghi nazionali → scoperta rapida “chi pubblica cosa” → *dati.gov.it*, *data.gov*, *data.europa.eu*

Istituti statistici ufficiali → dati standardizzati + metodologia → [*ISTAT*](#), [*Eurostat*](#), *ONS (UK)*

Portali di enti locali (Regioni/Comuni) → dettaglio territoriale e servizi → portali open data regionali/comunali (spesso indicizzati dai cataloghi)

Agenzie/Ministeri tematici → dati operativi su un settore specifico → ambiente, trasporti, sanità, lavoro (portali/dashboards dedicati)

Organizzazioni internazionali → confronti tra paesi + serie lunghe → [*World Bank*](#), *ONU*, [*OECD*](#), *WHO*

Geodati / portali cartografici (GIS) → confini, reti, suolo, POI per mappe e join → geoportali nazionali/regionali, *OpenStreetMap*, *ISTAT (confini comunali)*

Repository scientifici & community dataset → nicchia, dataset citabili o riusabili → *Zenodo/Dataverse/Figshare* (research), *GitHub/Kaggle* (community)

Find dataset

Google: operatori di ricerca

decessi	trova documenti che contengono la parola decessi
“decessi giornalieri”	trova documenti che contengono la frase “decessi giornalieri”
filetype: csv	trova file con una specifica estensione
site: dominio	trova documenti solo in un determinato dominio
intitle: “decessi comunali”	trova pagine con il titolo indicato
inurl: opendata	trova url con una precisa parola

Combinazione di operatori: [site:istat.it "decessi giornalieri"](#)

[Google Search Operators: The Complete List \(44 Advanced Operators\)](#)

Google Dataset Search

Dataset Search

Search for Datasets



Try [coronavirus covid-19](#) or [water quality site:canada.ca](#).



Si possono usare gli operatori search di google

Modelli LLM con Chat per fare Web Search

Perplexity

<https://www.datastudios.org/post/does-perplexity-always-show-sources-citation-quality-and-transparency>

ChatGPT (OpenAI)

<https://help.openai.com/en/articles/9237897-chatgpt-search>

Gemini (Google) <https://ai.google.dev/gemini-api/docs/google-search>

Claude (Anthropic) <https://claude.com/blog/web-search>

...

Estrazione dei metadata

1. **Fonte:** chi pubblica? istituzione/affidabilità
2. **Scopo:** perché è stato raccolto
(amministrativo/statistico/ricerca)
3. **Unità di analisi:** cosa rappresenta una riga (comune-anno? individuo? evento?)
4. **Copertura temporale e spaziale:**
territorio, periodo, popolazione
inclusa/esclusa
5. **Granularità:** livello
(comune/provincia/regione), frequenza
(annuale/mensile)
6. **Encoding:** utf-8, iso-8859-1 (Latin-1)
7. **Definizioni:** significato colonne, unità di misura
8. **Metodologia:** come è stato prodotto
(stima, censimento, registri, revisioni)
9. **Qualità:** missing, duplicati, outlier, coerenza interna
10. **Aggiornamento/versioni:** ultimo update, revisioni, serie ricostruite
11. **Licenza/riuso:** CC BY? vincoli di attribuzione
12. **Tracciabilità:** URL, data download, file originale conservato

Tabella Encoding

Encoding	Caratteri Speciali	Diffusione	Ideale per
UTF-8	Eccellente (Globale)	Massima	Qualsiasi uso, scambio dati
Windows-1252 / ISO-8859-1 (Latin-1)	Limitato (Europa)	Alta (Windows)	Sistemi legacy, dati solo europei
UTF-8+BOM	Eccellente	Media (Excel)	Importazione in Excel

Chiavi e granularità: come unire due dataset

1. **Allinea la granularità**

stesso livello territoriale? (comune vs provincia)
stesso tempo? (anno vs mese; data riferimento pop)

2. **Scegli la chiave corretta**

preferisci **codici** (codice ISTAT comune) > nomi
chiave tipica: **codice_comune + anno**

3. **Normalizza le chiavi**

zeri iniziali, tipi (stringa vs numero), encoding
attenzione a maiuscole, spazi, accenti nei nomi

4. **Controlla unicità**

la chiave deve essere unica in entrambi i dataset (one-to-one)
se non lo è: stai unendo a granularità diversa
→ prima aggrega

5. **Esegui il merge e verifica**

quante righe “perdono” il match?
controlla campioni manuali (5 comuni, 2 anni)

Caveat tipici

Confini e fusioni comunali

comuni che cambiano codice/nome o si fondono
→ serie spezzate

Codici vs denominazioni

join per nome = errori (omonimie, spelling) →
usare codice ISTAT

Date di riferimento diverse

decessi = flusso annuale; popolazione = stock
(1° gen / 31 dic)

dichiarare la scelta (es. pop al 1° gennaio come
proxy)

Standardizzazione per età (se non la fai,
dichiaralo)

“confronto descrittivo, non causale”

Piccoli numeri

comuni piccoli → tassi instabili (outlier
“rumorosi”)

preferire aggregazioni / medie mobili / soglie
minime

Interpretazione

evitare causalità: i dati descrivono, non
spiegano da soli

Template README (riproducibilità+accountability)

Titolo dataset: “Decessi ISTAT + Popolazione comunale – esercizio”

Fonti (link)

decessi: URL + data download

popolazione: URL + data download

Descrizione

cosa contiene il dataset finale (1–2 righe)

Unità di analisi

es. “comune–anno”

Variabili principali

decessi, popolazione, tasso (specificare

moltiplicatore: per 1.000/100.000)

Trasformazioni effettuate

pulizia colonne, filtri, aggregazioni, join (chiave usata)

Controlli qualità

missing/duplicati/range + risultati sintetici

Caveat (limiti)

3 punti (confini/codici, tasso grezzo, piccoli numeri, date riferimento)

File prodotti

raw/..., clean/..., script/notebook, grafico

Licenza/riuso

licenza dichiarata dalle fonti + attributo richiesto

Esercizio (10')

Trovare la serie della popolazione residente in Italia per comune 2011-2024

Non basta “popolazione per comune”: serve anche **anno** e **codice comune** (meglio del nome).

Controllare sempre **data di riferimento** della popolazione (1° gennaio/31 dicembre) e allinearla ai decessi.

Annotare nel README: **link fonte + data download + definizioni + caveat.**

Conclusioni

Cosa abbiamo imparato

Fare domande

Come diventare esperti di dominio aiuta a costruire la lista di domande

Come cercare open dataset (data sourcing) e pipeline

Template README da compilare per ogni dataset