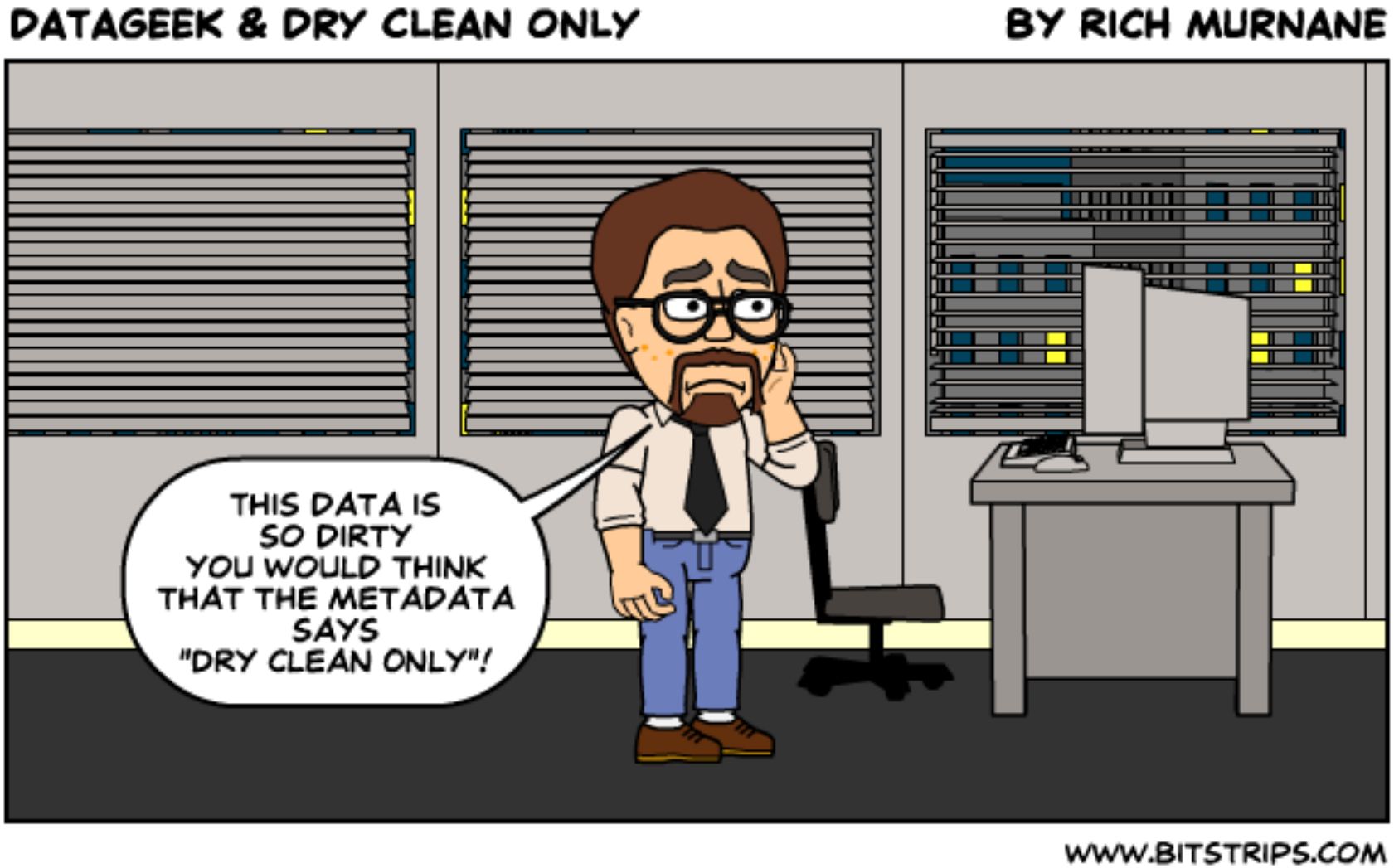
Analisi dei Dati

Fonti dati e data pre-processing

Domenico De Stefano

a.a. 2022/2023

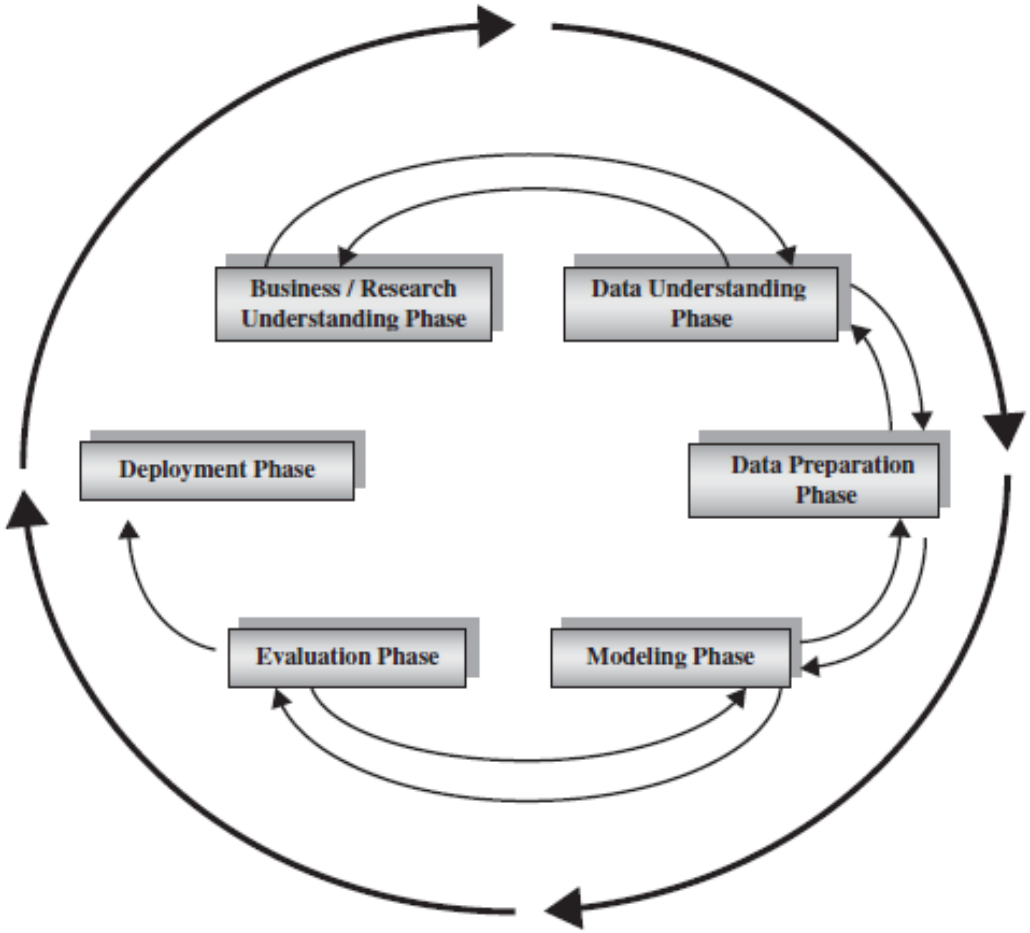

DATAGEEK & DRY CLEAN ONLY
BY RICH MURNANE
THIS DATA IS SO DIRTY YOU WOULD THINK THAT THE METADATA SAYS "DRY CLEAN ONLY"!
WWW.BITSTRIPS.COM

Data preparation o data pre-processing (preparazione dei dati)

- L'analisi dei dati è il processo di pulizia, trasformazione, esplorazione e modellazione di dati con il fine di evidenziare informazioni che suggeriscano conclusioni e supportino le decisioni
- Per data pre-processing si intendono le fasi (indicata in CRISP-DM come data preparation) che consentono di ottenere un insieme di dati pronto per le analisi successive (modeling)
- nel contesto delle indagini statistiche (e dello schema del disegno di indagine) questa fase include anche alcuni aggiustamenti post-rilevazione

Analisi dei dati come processo (iterativo)

Riproponiamo il Cross-Industry Standard Process for Data Mining (CRISP-DM)

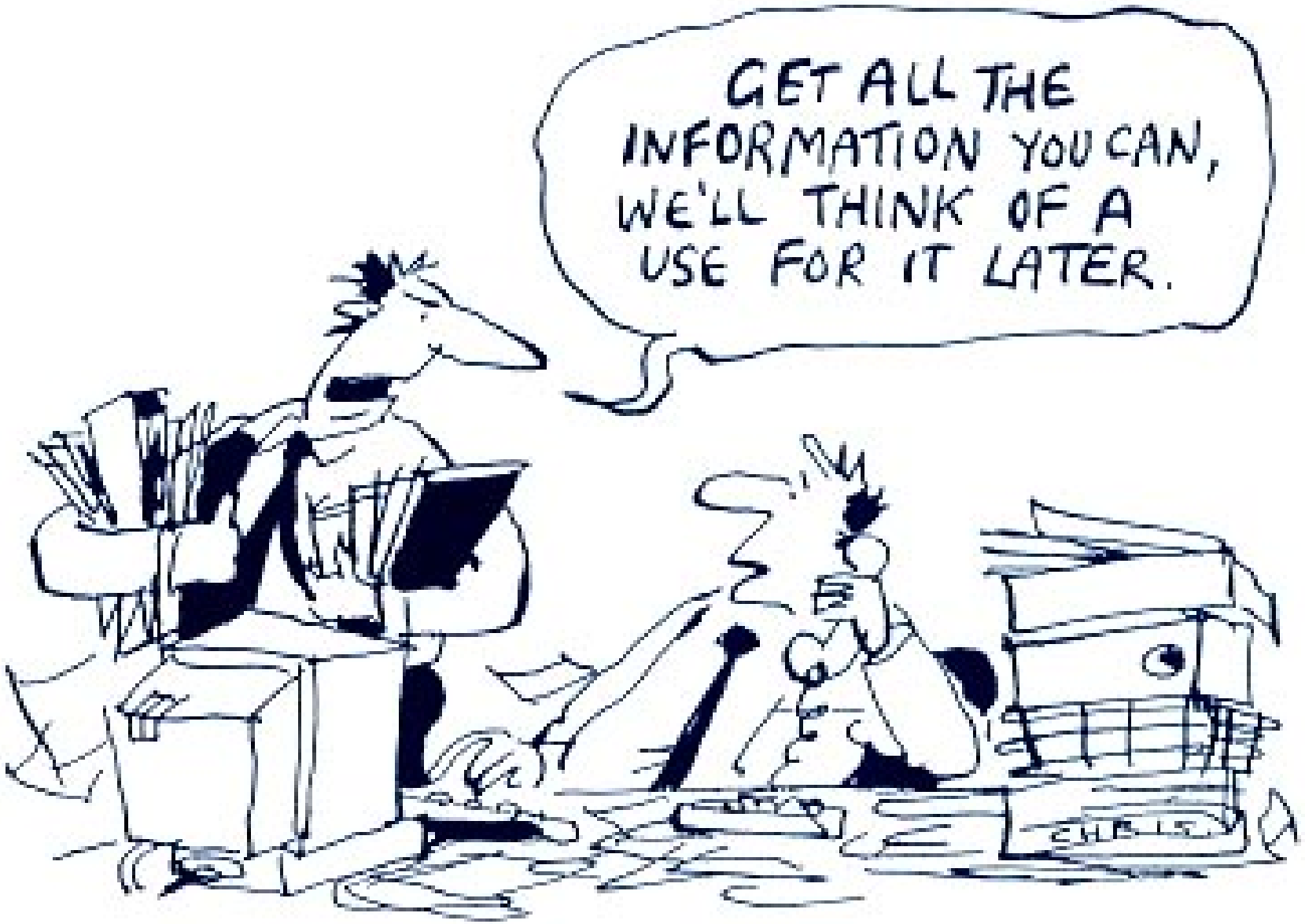


Analisi dei dati come processo (iterativo) I

1. Data Preparation Phase
   - lavoro di preparazione del dataset che contenga una versione pulita e sistematizzata dei cosiddetti dati grezzi raccolti (tipicamente struttura dati casi$\times$variabili)
   - Selezionare casi e variabili che si vogliono analizzare e che siano appropriati per raggiungere gli obiettivi conoscitivi del progetto
   - Se necessario, operare opportune trasformazioni di variabili
   - Organizzare ulteriormente i dati per usare determinate tecniche o modelli statistici

Gli obiettivi e le operazioni della fase di pre-processing dipendono dalla fonte da cui si reperiscono i dati

Indice


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.

Dati, informazione, conoscenza I

- I dati statistici possono provenire da
  - rilevazione appositamente eseguita (progettazione indagine)
  - da indagini effettuate su uno specifico argomento o dall'elaborazione, finalizzata a scopo statistico, di informazioni raccolte all'interno di procedimenti di tipo amministrativo.
  - da nuove fonti di dati (es: web, transazioni elettroniche, ecc.) che danno luogo ai cossidetti "big data"
- Fruibilità dei dati
  - I dati possono essere resi disponibili a terzi (utenti) o nella forma rilevata (banche dati, microdati) o, più frequentemente, in forma elaborata (indicatori, tabelle, grafici, ...).
  - I dati possono essere però ormai anche essere disponibili in forma non strutturata in nessuno dei modi di cui sopra (ad esempio: messaggio scambiato su un social)

Qualità dei dati I

Concetto cardine di qualunque analisi statistica è la qualità delle informazioni statistiche prodotte e dell'intero processo di raccolta dati

- La valutazione della qualità di un'indagine statistica e dei dati raccolti è un'attività ormai fondamentale
- Le indagini statistiche in campo sociale servono a fornire informazioni per analizzare l'evoluzione dei processi sociali in atto, per soddisfare e anticipare i bisogni informativi su cui impostare gli interventi volti a ridurre le emergenze sociali e promuovere il benessere del Paese
- per questa ragione il dato statistico è considerato alla stregua di qualunque bene/servizio
  - qualità definita dalle norme **Iso 8402-1986** come il possesso della totalità delle caratteristiche che portano al soddisfacimento delle esigenze, esplicite o implicite, dell'utente
- La qualità dei dati statistici (e dell'informazione) è un concetto multidimensionale che investe le diverse caratteristiche desiderabili di tale tipo di prodotto

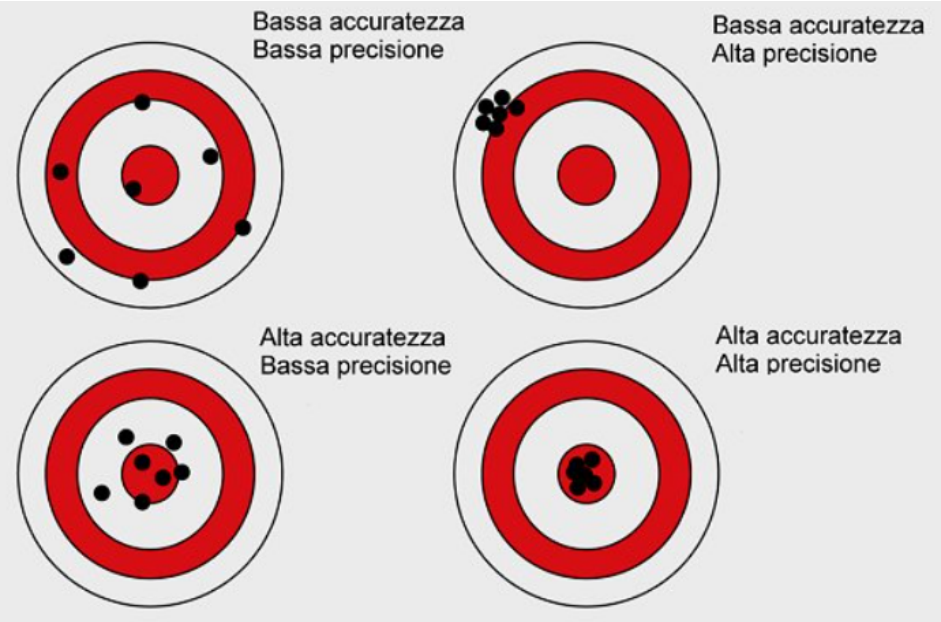
Qualità dei dati II

Dimensioni della qualità del dato

L'Eurostat (ufficio statistico dell'UE) ha definito le caratteristiche cui deve soddisfare l'informazione statistica, identificando le seguenti **dimensioni**: rilevanza, accuratezza, tempestività, accessibilità, confrontabilità, coerenza e completezza

Qualità dei dati III

- Rilevanza. Capacità del dato di soddisfare le esigenze conoscitive degli utenti. Tale caratteristica è strettamente collegata con gli obiettivi dell'indagine
- Accuratezza. Grado di corrispondenza tra la stima ottenuta dall'indagine e il vero (ma ignoto) valore della caratteristica di interesse. Il dato in esame è rispondente al dato dell'intera popolazione? Il campione è rappresentativo?



Qualità dei dati IV

- Tempestività. Brevità dell'intervallo di tempo compreso tra il momento della diffusione dell'informazione prodotta e l'epoca di riferimento della stessa
- Accessibilità e trasparenza. Semplicità per l'utente di reperire, acquisire e comprendere l'informazione disponibile in relazione alle proprie finalità.
- Confrontabilità. Possibilità di paragonare nel tempo e nello spazio le statistiche riguardanti il fenomeno di interesse
- Coerenza. non contraddittorietà tra dati Si può parlare di coerenza 'interna' di dati in una dataset, come di una coerenza 'esterna' tra dati di dataset diversi. La coerenza può risultare, ad esempio, nella comparazione di dati di serie storiche o di fenomeni correlati (anche proventienti da indagini o dataset diversi).
- Completezza. Capacità di integrazione dei singoli dati al fine di fornire un quadro informativo soddisfacente del dominio di interesse. Dipende dal fenomeno studiato: ad es. un indirizzo senza un codice postale, l'occupazione di un soggetto senza l'anno di assunzione; un'analisi sui CFU acquisiti dagli studenti senza l'anno di iscrizione

Processo di produzione conoscenza statistica I

Conoscenza statistica: conoscenza quantitativa di un fenomeno collettivo

Relativamente ad un fenomeno di interesse (es: performance degli studenti di scuola secondaria, caratteristiche individuali e contesto sociale):

1. Informazioni da dati già disponibili
   - raccolte per fini specifici da indagini campionarie mediante questionario $\Rightarrow$ Es: PISA (Programme for International Student Assessment)
   - raccolte per altri fini (fonti amministrative) $\Rightarrow$ Es: Registri di classe; archivi iscrizioni scolastiche
2. Informazioni raccolte da nuove fonti dati $\Rightarrow$ Es: piattaforme e-learning, gruppi di studio su whatsapp, post su blog scolastici ecc.

La natura dei dati. Le indagini statistiche

Indagine statistica (censuaria o campionaria)

- Vantaggi
  - Indagini pianificate ad hoc (specificità rispetto all'obiettivo conoscitivo)
  - Specifica popolazione obiettivo (target population)
  - Definizioni, concetti e variabili definite ex-ante
  - Quesiti mirati e possibilità di testare il questionario
  - Stime basate sul paradigma inferenziale tradizionale (nel caso di indagini campionarie)
  - Garantite più dimensioni della qualità dei dati
- Svantaggi
  - Costi elevati (dipende da ampiezza del campione, territorio da indagare, lunghezza del questionario)
  - Elevato carico statistico sui rispondenti (ad es: su argomenti sensibili rischio di ottenere risposte non attendibili)
  - le fonti di errore citate nella progettazione di indagine (lato rappresentazione/ lato misurazione)

Fonti dati da indagine

In Italia:

- Sistema statistico nazionale (SISTAN)
  - rete di soggetti pubblici e privati che dal 1989 fornisce l'informazione statistica ufficiale. Oltre all'Istat, 57 istituzioni pubbliche e private realizzano oltre 800 lavori statistici tra indagini, elaborazioni e sistemi informativi statistici.

A livello internazionale (alcune banche dati):

- Eurostat (The Statistical Office of the European Communities)
- FAO (Food and Agriculture Organization of the United Nations)
- UNESCO (Education Information Service)
- WHO (The World Health Organization)
- OECD (Organization for Economic Co-operation and Development)

La natura dei dati. Gli archivi amministrativi

Archivi Amministrativi (Anagrafi, Banche dati ministeriali o istituzionali, ecc.)

- Vantaggi
  - Riduzione dei costi e del carico statistico sui rispondenti
  - Possibilità di aumentare il dettaglio (sottopopolazione e livelli territoriali)
  - Coerenza del contesto in cui vengono prodotti i dati
- Svantaggi
  - La popolazione amministrativa potrebbe non coincidere con la target population
  - Accesso ai dati può essere problematico
  - La qualità dei dati non è sempre ottimale (es. errori nell'inputazione dati, differenze di classificazione a seconda della normativa, data management diverso, ecc.)

È necessario tradurre il dato amministrativo in informazione statistica di qualità (a volte può non essere possibile)

La natura dei dati. Le nuove fonti

Big Data (per semplicità: dati generati dall'uso degli strumenti digitali)

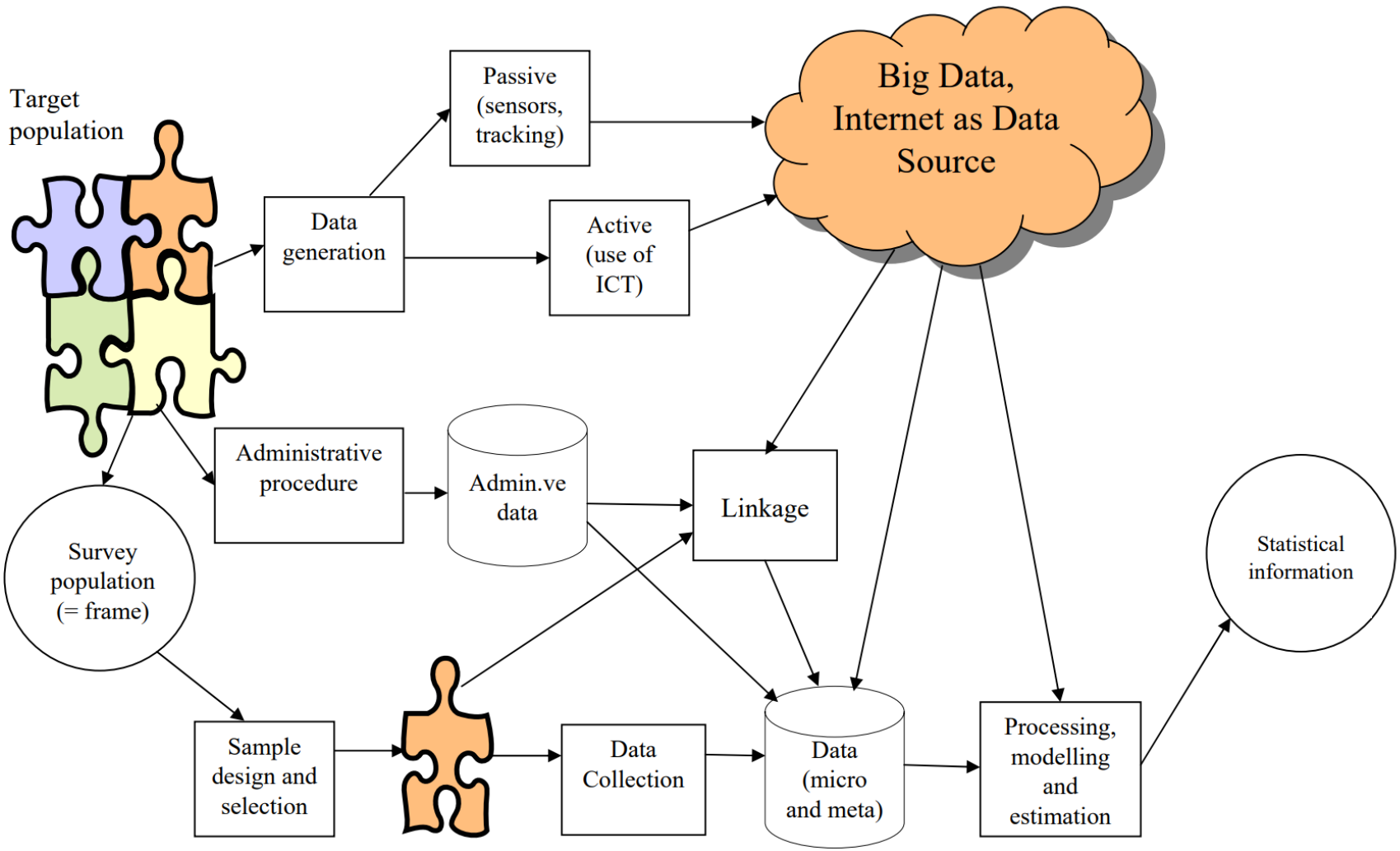
- Vantaggi
  - Registrano eventi, spesso registrano "comportamenti" (spontanei)
  - Ampliano le opportunità di analisi e le informazioni disponibili
  - Dati tempestivi, generati ad un costo estremamente contenuto
- Svantaggi
  - Assenza della target population
  - Definizioni e variabili di solito non coincidono con quelle della statistica ufficiale e possono non essere pertinenti rispetto all'obiettivo conoscitivo
  - Accesso ai dati può essere problematico
  - Problemi tecnologici legati al trattamento di ingenti quantità di dati
  - Difficoltà nell'estrarre l'informazione rilevante (presenza del cosiddetto 'rumore' nei dati)

È necessario un grande impegno per estrarre valore dai big data. Metodi finora utilizzati non sono ancora sufficienti e non sono strutturati come quelli relativi alle indagini classiche

Integrazione tra fonti diverse

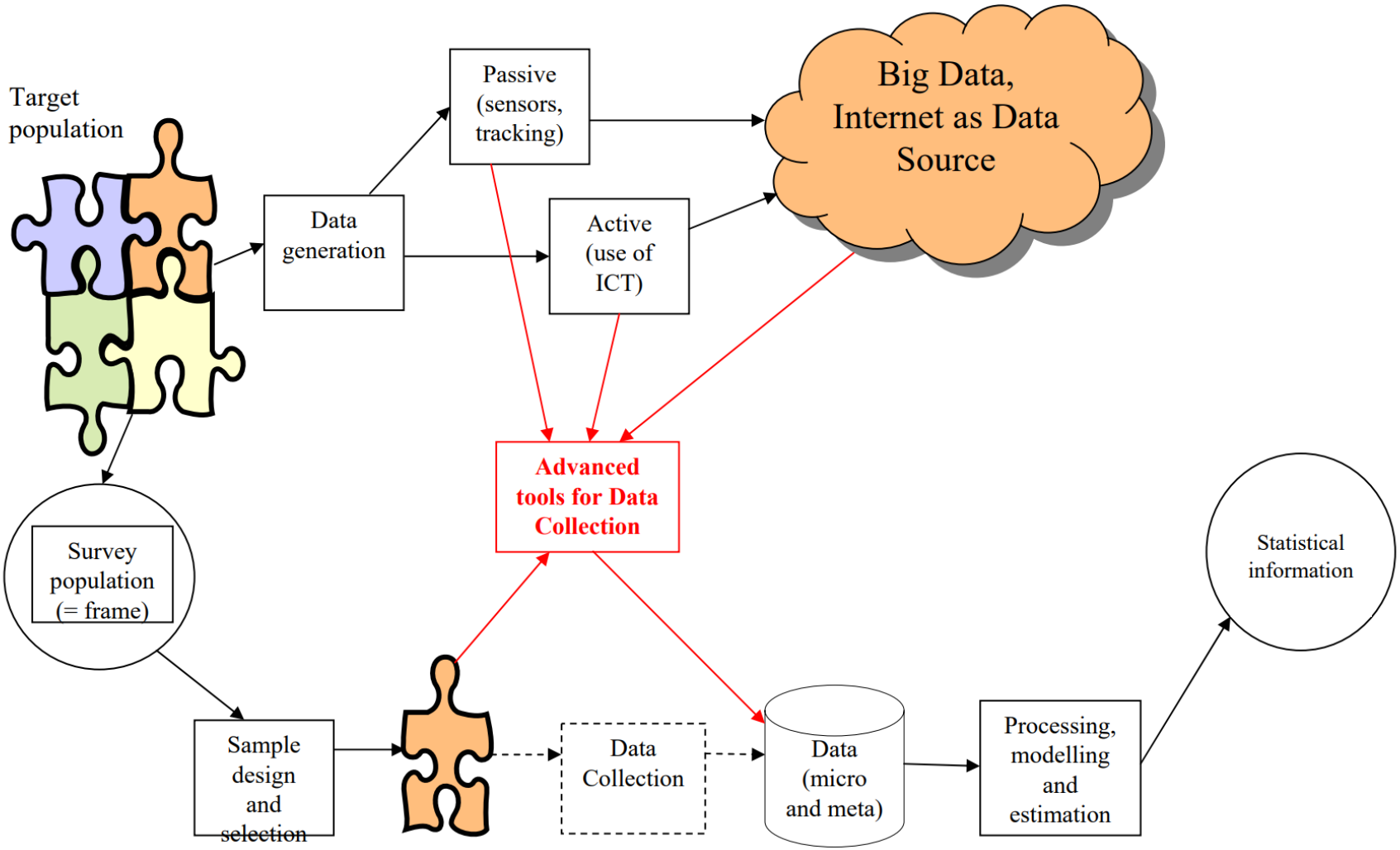
Il futuro della produzione dati ormai corre verso l'itegrazione tra fonti diverse

Il frame teorico



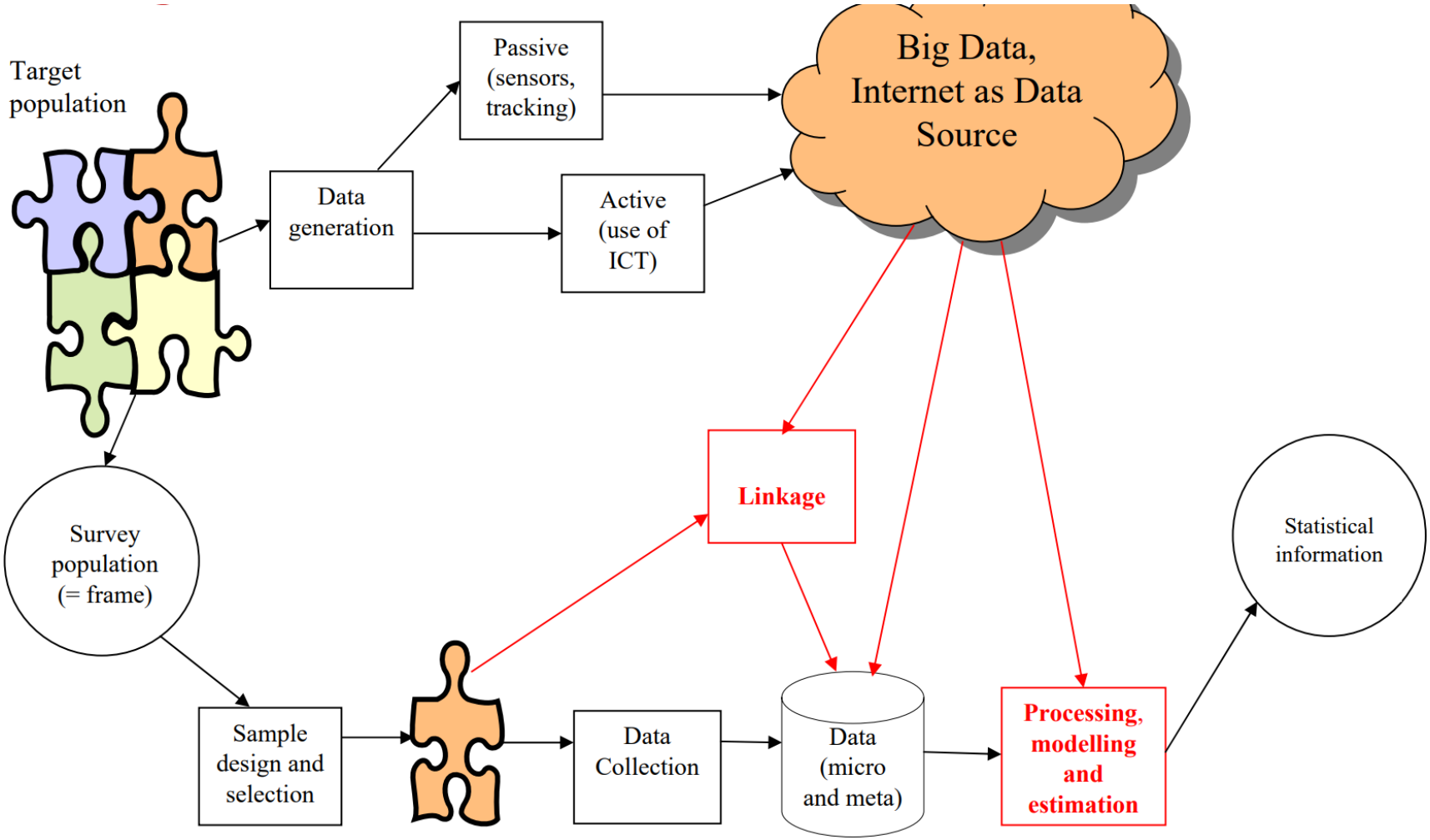
Il frame teorico

Scenario 1: tecniche alternative per la raccolta dati



Il frame teorico

Scenario 2: Uso integrato di dati di indagine e Big Data



Scenario 2: un esempio (ISTAT)

Utilizzando le footprint generate dai **tracking device** (cellulari, GPS) è possibile individuare i bacini di movimento che possono essere utilizzati, ad esempio, per ridefinire i Sistemi Locali del Lavoro.
L'idea è di utilizzare congiuntamente:

- dati di indagine campionarie (indagine Forze Lavoro ISTAT);
- dati di censimento;
- dati amministrativi (Inps);
- Big Data originati da:
  - tracking devices e sensori;
  - interrogazioni su Internet;

Big Data: un focus

Con il termine big data si indicano si indica un insieme enorme di dati (dovreste averlo capito) che richiede la definizione di nuovi strumenti e metodologie per estrapolare, gestire e processare informazioni entro un tempo ragionevole

I dati possono essere di vario tipo

- Strutturati: del tipo unità statistiche $\times$ variabili (matrice dati)
- Non strutturati: ad esempio testi, immagini, video, documenti, email, tweet ecc.

Big Data: un focus

Big Data: un focus

Le fonti da cui si generano i big data

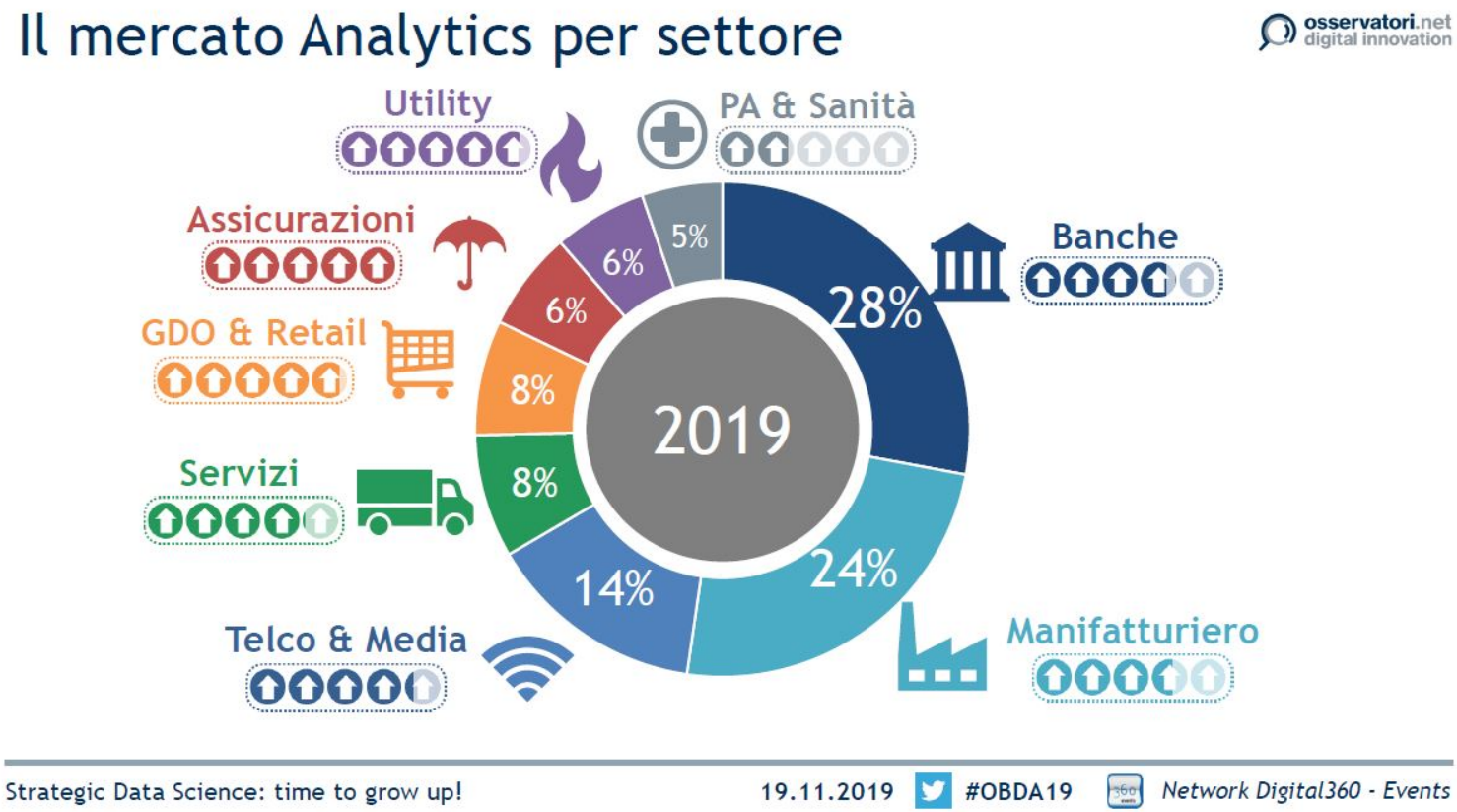
- Human-generated data
  ad es.: dati da Social Media, Blog, SMS, e-mail, User generated contents e maps, ecc.
- Process-mediated data
  quali Sistemi transazionali, commerciali e bancari tradizionali, e-commerce, carte di credito, dati prodotti da Enti Pubblici e/o privati
- Machine-generated data
  tipicamente ciò che va sotto il nome di Internet of Things , come sensori fissi (home-automation, sensori ambientali/meteorologici, sistemi per il controllo del traffico, ecc.) e mobili (dispositivi mobili, sensori su automezzi, immagini satellitari).

Big Data: un focus

I big data hanno delle caratteristiche riassumibili nelle cosiddette 5 V (originariamente l'analista Doug Laney aveva definito il modello di crescita come tridimensionale con le 3 V originarie: Volume, Varietà, Velocità)

Big Data: un focus

Chi usa (o vorrebbe) i Big Data? (fonte: Osservatorio sui Big Data - Politecnico di Milano, 2019)



Indice

Motivazione

I dati raccolti spesso sono "sporchi":

- incompleti: manca il valore di alcuni attributi, o mancano del tutto alcuni attributi interessanti.
- inaccurati: contengono valori errati o che si discostano sensibilmente da valori attesi.
  - Ad esempio, nel campo età di un impiegato si trova il valore di 120 anni
- inconsistenti
  - ad es:, se un soggetto di 15 anni indica di avere 5 figli.

Queste inesattezze non influivano sullo scopo iniziale per cui i dati sono stati raccolti, per cui vengono scoperte solo ora.

GIGO: garbage in – garbage out

se i dati in input non sono di buona qualità, neanche le analisi basate su di essi lo possono essere

Aspetti di pre-processing dei dati I

Principali aspetti e tecniche nella fase di pre-processing dei dati:

- data cleaning (pulizia dei dati): riempire i campi con i valori mancanti, i dati rumorosi, rimuovere i valori non realistici.
- data integration (integrazione dei dati): integrare dati provenienti da database diversi (o più ambiziosamente da fonti diverse) risolvendo le inconsistenze.
- data transformation (trasformazione dei dati): preparare i dati per l'uso di alcune applicazioni di analisi dei dati
- data reduction (riduzione dei dati): ridurre la mole dei dati in input, ma senza compromettere la validità delle analisi (campionamento, riduzione delle variabili, ecc.)

Indice

Data Cleaning I

Le attività eseguite durante la fase di data cleaning sono:

- risolvere gli attributi che hanno valori mancanti
- correggere le inconsistenze o le inesattezze
- identificare gli outliers (dati molto diversi da quelli che caratterizzano l'intera distribuzione di una variabile)

Dati mancanti I

Possibili approcci quando si hanno dati con valori mancanti:

- ignorare le istanze con valori mancanti o eliminare tutte le unità statistiche che presentano uno o piu dati mancanti
  - non molto efficace se la percentuale di dati mancanti è alta.
  - si usa spesso quando il dato che manca è la classe in un problema di classificazione
- inputare a mano i dati mancanti sulla base delle carattersitiche del fenomeno o ricorrendo al parere di esperti
  - non fattibile se i dati sono di grande dimensione
  - potrebbe produrre distorsioni non controllabili nei risultati finali
- Sostituire i dati mancanti con un valore costante come "NA" (in R).
  - occorre usare algoritmi e metodi che gestiscono la presenza di dati mancanti categorizzati in tale modo
  - se si usano altri valori (ad esempio 0 o "missing" a seconda del tipo di dato) occorre che si specifichi all'algoritmo che tali valori sono da considerarsi mancanti

Dati mancanti II

Altri possibili approcci:

- Usare la media della variabile (se numerica) o la moda (se la variabile è qualitativa) per i dati mancanti
  - la media potrebbe essere un valore non rappresentativo di una distribuzione molto eterogenea
  - possibile uso della mediana
- predirre il valore dell'attributo mancante sulla base degli altri attributi noti
  - la predizione può avvenire usando regressione lineare, alberi di classificazione, ecc.
  - si usano algoritmi di analisi dei dati per preparare i dati in input ad altri metodi di analisi dei dati

Esempi di dati mancanti

Sostituzione con valori costanti

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	0	71	Europe
3	17.000	302	140	US
4	15.000	400	150	Missing
5	37.700	89	62	Japan

Esempi di dati mancanti

Sostituzione con medie e mode

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	200.65	71	Europe
3	17.000	302	140	US
4	15.000	400	150	US
5	37.700	89	62	Japan

Dati inaccurati I

Cause specifiche di inesattezze nei dati

- errori tipografici in variabili qualitative o errori che inficiano la natura di variabili che dovrebbero essere quantitative: ad es. per l'altezza si riporta 180cm
- inconsistenze nelle modalità: ad es. Friuli Venezia Giulia e FVG oppure Friuli (il software non riconosce che sono la stessa modalità)
- errori tipografici o di misura in attributi numerici: ad es. riportare 1,80 per l'altezza espressa in cm
- errori deliberati: durante un sondaggio, l'intervistato può fornire un CAP falso oppure alcuni errori causati da sistemi di input automatizzati: se il sistema insiste per un codice ZIP (come il CAP ma negli USA) e l'utente non lo possiede?

Dati inaccurati II

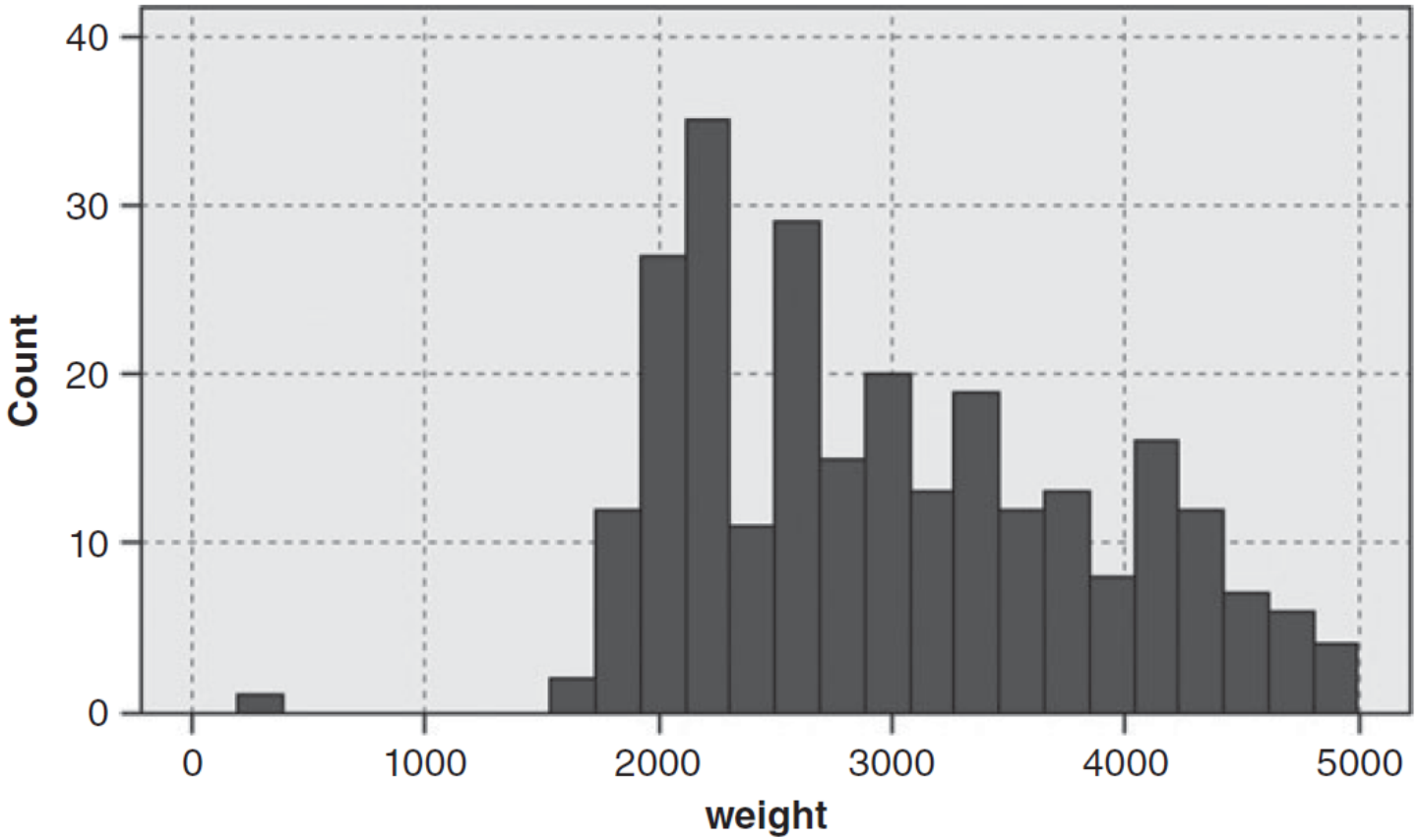
Occorre imparare a conoscere i propri dati!

- capire il significato di tutti i campi e variabili
- individuare gli errori che sono stati commessi. Semplici visualizzazioni grafiche consentono di identificare rapidamente dei problemi (ad es. semplicemente dei boxplot o anche tabelle di frequenza):
  - la distribuzione è consistente con ciò che ci si aspetta?
  - c'è qualche dato ovviamente sbagliato?

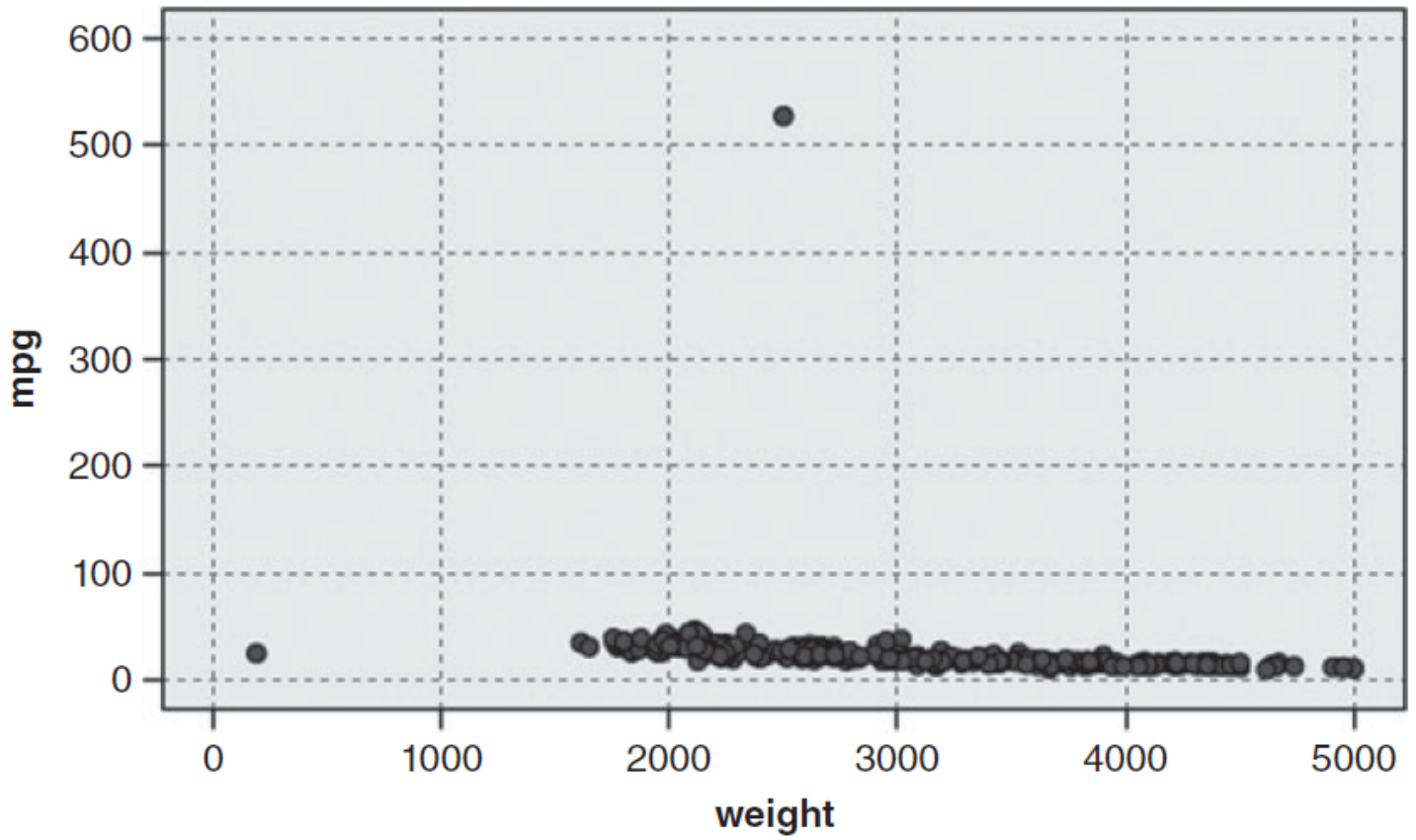
Esempi di dati inaccurati

Brand	Frequency
USA	1
France	1
US	156
Europe	46
Japan	51

Esempi di dati inaccurati



Esempi di dati inaccurati



Indice

Trasformazione dati I

I dati sono consolidati e trasformati in forme più appropriate per le analisi.

- aggregazione: costruire dati aggregati prima dell'analisi (ad esempio classi di valori)
- costruzione degli attributi: costruire nuovi attributi a partire da quelli presenti (ad esempio media voti a partire dai voti ai singoli esami)
- cambiare scala di misura: trasformare i dati in gradi C in gradi F
- normalizzazione (o in generale rescaling): modificare la scala dei dati in modo che cadano in intervalli stabiliti (ad esempio da -1 ad 1)

Normalizzazione dei dati I

Alcune variabili possono avere range di variazione molto diversi. Ad esempio i voti all'esame variano tra 18 e 30 mentre i cfu da 0 a 180 (in una triennale).
Queste differenze di variazione causano problemi in alcuni metodi di analisi dei dati perciò spesso occorre normalizzare i valori (in particolare metodi di classificazione o metodi basati sulle distanze come il clustering).

- Un metodo di normalizzazione molto usato è il metodo min-max

$$X^*_{\text{mm}} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Per altre applicazioni basta invece solo centrare la variabile:

$$X_c = X - \overline{X}$$

Entrambi i metodi però sono molto influenzati dagli outliers

Normalizzazione dei dati: standardizzazione I

Trasforma una variabile X in una variabile Z con distribuzione "standard", ossia di media 0 e varianza pari a 1 eliminando qualsiasi effetto dell'unità di misura.

$$Z = \frac{X - \overline{X}}{\sigma_X}$$

Ora l'unità di misura è in unità di deviazioni standard dalla media (che è 0)

- utile quando non si conosce minimo e massimo
- i valori normalizzati non hanno un minimo e un massimo fissato (ma tipicamente si muovono in un range che va tra -3 e 3)
- non influenzato dagli outlier (o almeno non altrettanto dei metodi precedenti)
- i valori standardizzati vengono anche detti Z-score: di solito gli outliers sono oltre il range -3, 3

Indice

Data reduction I

Necessità di effettuare una riduzione dei dati per ottenere una rappresentazione ridotta dei dati per rendere più intelligibili i dati e produrre gli stessi (o comunque simili) risultati oppure per risparmiare spazio di storage degli stessi.
Varie strategie

- selezione di attributi rilevanti
- aggregazione
- riduzione della dimensionalità
- riduzione della numerosità
- discretizzazione o generazione delle gerarchie di concetto

Selezione di attributi rilevanti I

Eliminare variabili non utili dall'intero dataset (non sempre necessario, dipende dall'estensione del dataset)

- ad esempio, eliminare gli attributi irrilevanti, come può essere ad esempio una costante (variabile priva di variazione) o identificativi (delle unità statistiche, del questionario, dell'intervistatore, ecc.)
- eliminare variabili non utili perché non determinanti nello studio di un certo fenomeno di interesse (ricorrere al parere di un esperto del campo)

o per un'analisi specifica

- Serve una "misura" della bontà di un insieme di variabili, in modo che si possa scegliere l'insieme migliore

Selezione di attributi rilevanti II

- step-wise forward selection
  - si parte da un insieme vuoto di variabili (o dall'insieme minimo di variabili per un certo metodo, ad es. una sola variabile indipendente e una dipendente per un modello)
  - ad ogni passo aggiungo la variabile che massimizza la qualità dell'insieme risultante secondo un certo criterio di bontà (vedremo di fitting di un modello in seguito)
  - insieme iniziale:
    $\{X_1, X_2, X_3, X_4, X_5, X_6\} \rightarrow \{\} \rightarrow \{X_1\} \rightarrow \{X_1, X_3\} \rightarrow \{X_1, X_3, X_6\}$
- step-wise backward selection
  - si parte da tutte le variabili
  - ad ogni passo si toglie la variabile che massimizza la qualità dell'insieme risultante secondo un certo criterio di bontà (vedremo di fitting di un modello in seguito)
  - insieme iniziale: $\{X_1, X_2, X_3, X_4, X_5, X_6\} \rightarrow \{X_1, X_2, X_3, X_4, X_5, X_6\} \rightarrow \{X_1, X_3, X_4, X_5, X_6\} \rightarrow \{X_1, X_3, X_5, X_6\} \rightarrow \{X_1, X_3, X_6\}$

Selezione di attributi rilevanti III

Necessità di stabilire dei criteri di arresto. Ci si ferma quando:

- si è raggiunto un numero fissato di variabili

oppure

- si è raggiunta una qualità minima desiderata per l'insieme
- l'aggiunta o la rimozione di una variabile danno un incremento minimo della misura di bontà da massimizzare (se fosse da minimizzare si farebbe un ragionamento analogo)