Università di Genova | MaLGa

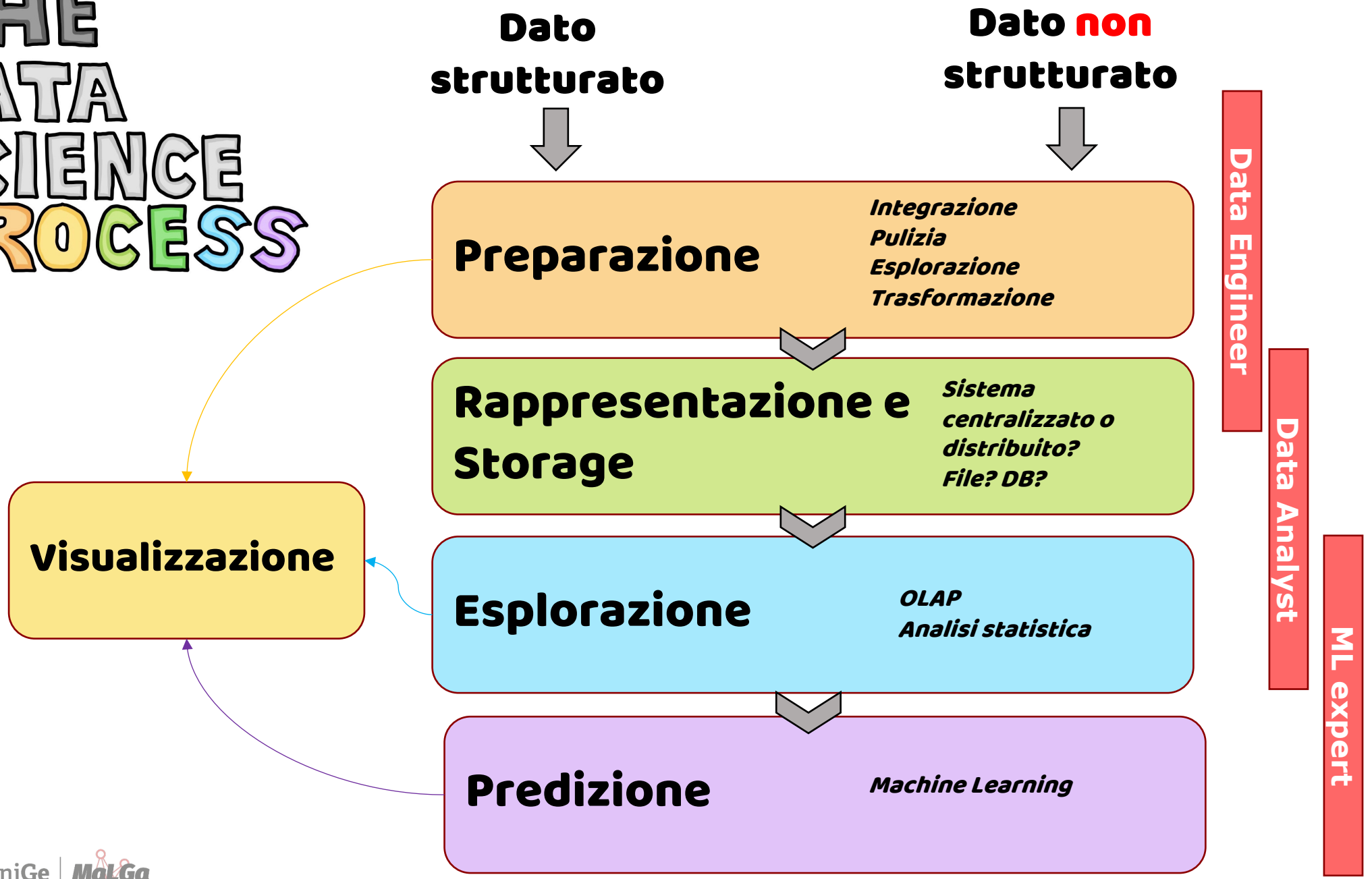
Preparazione e rappresentazione del dato

Nicoletta Noceti

UniGe | MaLGa

Dove eravamo arrivati?

THE DATA SCIENCE PROCESS



I passi della data science

- Cominciamo a conoscere e capire i passi della data science
- Il primo passo riguarda la preparazione, seguito dalla rappresentazione e storage
- Oggi parliamo principalmente del primo, e facciamo qualche cenno al secondo

UniGe | MaLGa

Preparazione del dato

Preparazione del dato può significare...

- Pulizia
- Integrazione

Ma anche...

- Esplorazione
- Trasformazione

Pulizia del dato (aka Data Cleaning)

I dati reali possono essere afflitti da tante diverse tipologie di errore/rumore, facciamo qualche esempio

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Typo

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Valori non corretti

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Diverse rappresentazioni per la stessa informazione

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Valori nulli

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Violazione della chiave

Pulizia del dato (aka Data Cleaning)

Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Dati inconsistenti

Pulizia del dato (aka Data Cleaning)

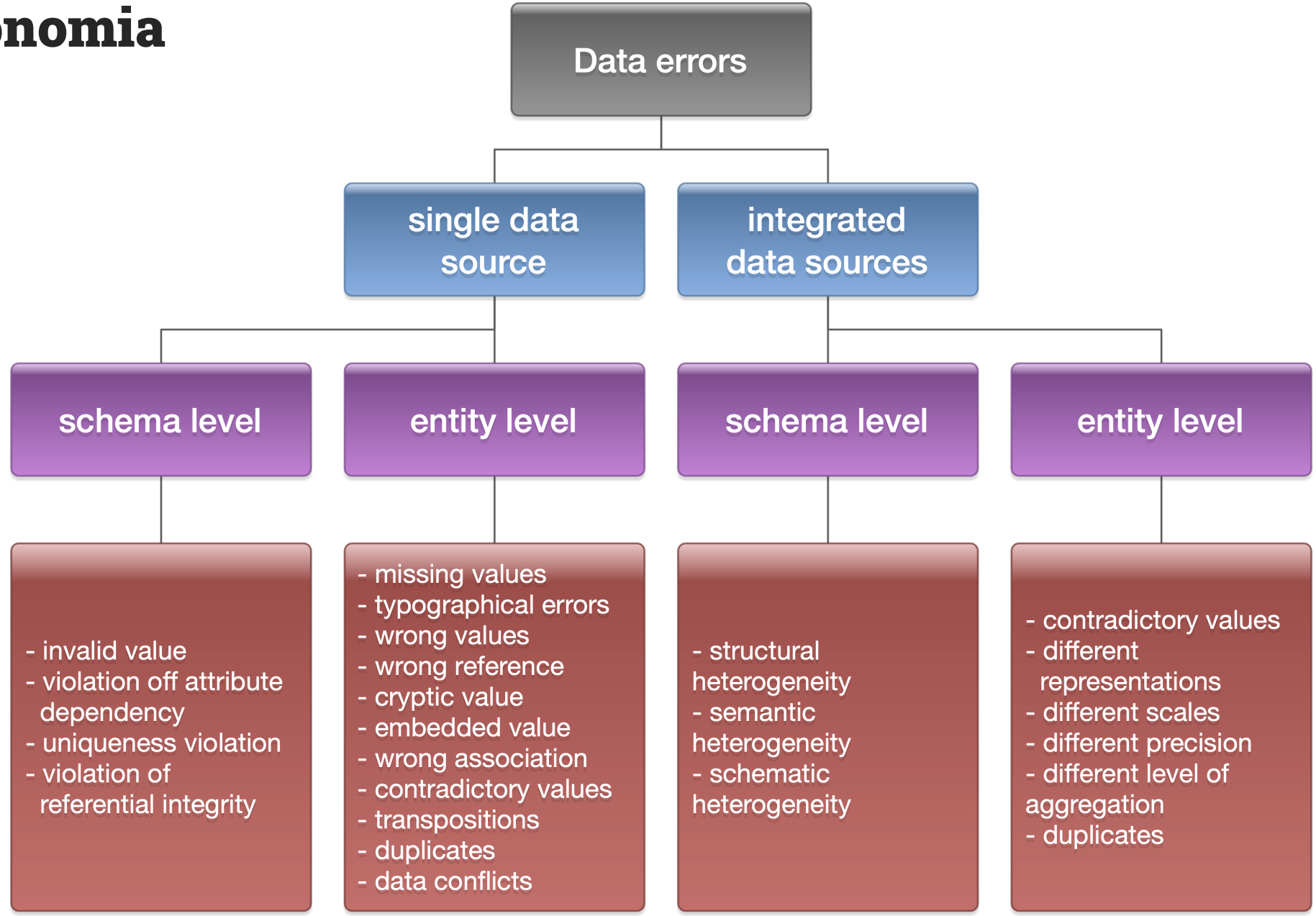
Esempi

NI#	name	DoB	Age	Gender	phone	zip
1234	Smith, John	18/2/1980	30	M	null	70567
1234	Jane Doe	32/4/1970	47	F	768-4511	5555
1235	John Smith	18/2/1980	37	M	567-3211	70567

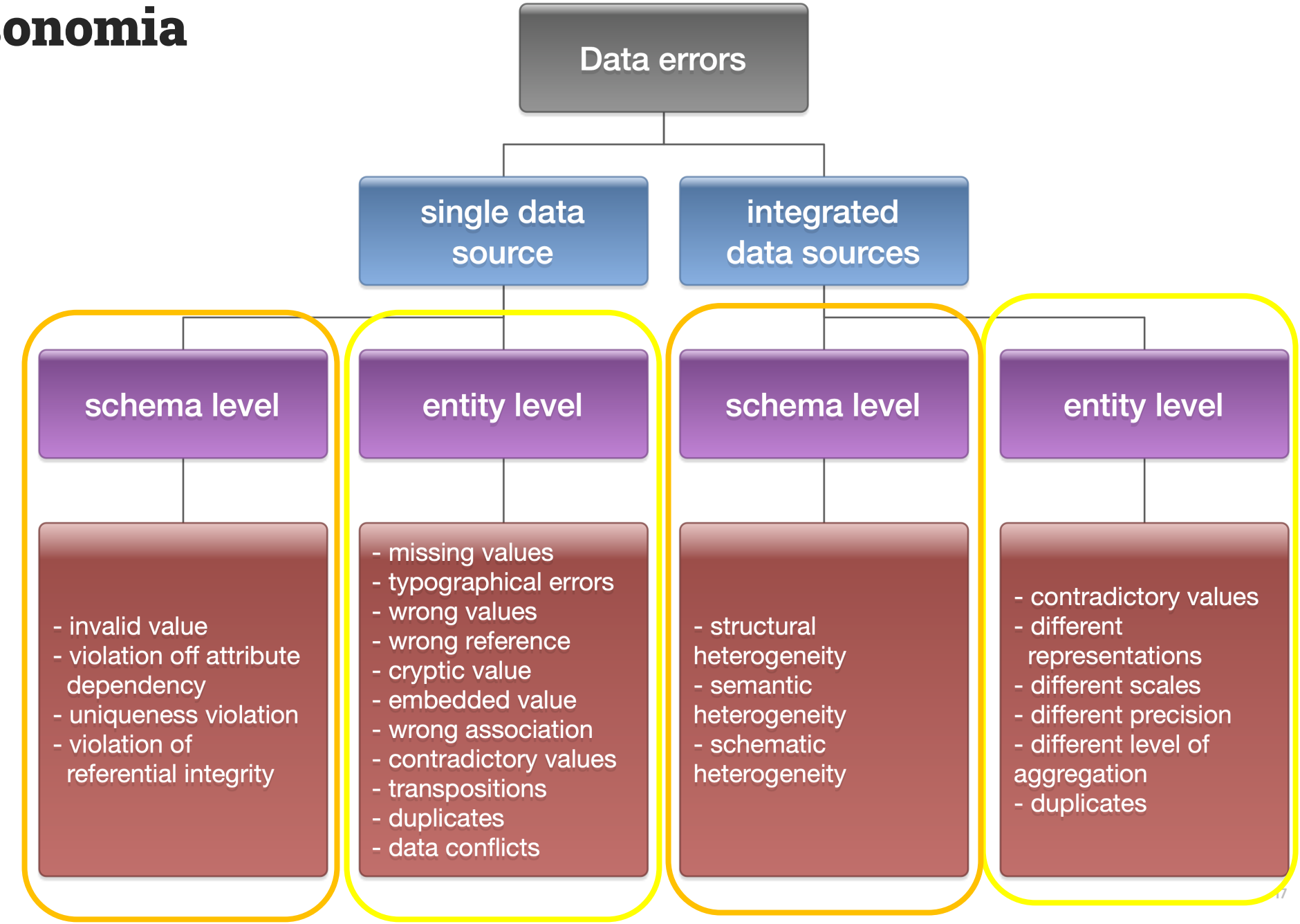
zip	city
70567	Stuttgart
70567	Stutgart
70569	Germany

Duplicati

Una tassonomia



Una tassonomia



Pulizia del dato (aka Data Cleaning)

- E’ fondamentale preoccuparsi della pulizia del dato prima di procedere con le successive fasi di analisi!!
- Ci sono alcune caratteristiche che è bene valutare

Pulizia del dato (aka Data Cleaning)

Consistenza

Un dato potrebbe contenere errori o conflitti che emergono come violazioni di regole semantiche

Esempio: age = 82 e age = 20 per la stessa persona

Pulizia del dato (aka Data Cleaning)

Precisione

Quanto vicino è un valore che rappresenta una entità reale al vero valore che ci si aspetta per tale entità

Esempio: età di studenti di scuola superiore espresso come <= 40 oppure 15

Pulizia del dato (aka Data Cleaning)

Completezza

La possibilità di rispondere a determinare domande sui dati a partire dalle informazioni a disposizione

Esempio: la presenza di valori null nei dati rappresenta un problema

Pulizia del dato (aka Data Cleaning)

Timeliness

Il fatto che i dati usati per rispondere ad una domanda su un certo periodo temporale siano relativi allo stesso periodo

Esempio: non posso fare analisi sull'anno 2023 se ho soprattutto dati che provengono dagli anni precedenti

Riassumiamo

Cosa significa che un dato è “sporco”

- E’ **incompleto**: mancano valori (“”; null; …)
- E’ **rumoroso**: contiene errori di grammatica, typo, di traduzione, o valori non compatibili con il tipo (esempio salario=-10)
- E’ inconsistente: contiene discrepanze (esempio age=42, birthday=“03/07/1997”)

Riassumiamo

Perchè un dato è “sporco”

- Gli errori e le inconsistenze possono essere accidentalmente introdotti in tante fasi della manipolazione dei dati, dalla raccolta, allo storage, alla trasmissione, trasformazione, integrazione, ...
- Possono essere dovuti a dati incompleti alla fonte, errori umani, incosistenze tra fonti, ...
- Ragioniamo con un esempio su come possiamo procedere, e cominciamo a ragionare in termini di obiettivi analitici

Esempi

ID_Film	Titolo	Genere	Regista	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	Tim Burton	2022	1	86
10009	Lucifer	Fantasy	Null	2015	6	85
10004	Stranger Things	Fantascienza, avventura	Shawn Levy	2016	4	86
10005	Cobra Kai	Azione, sportivo	Null	2018	5	82
10003	1899	Mistery	Null	2022	1	Null
10011	La casa di carta	Drama	Null	2017	3	83
10013	Dark	Mistery	Baran do Obar	2017	3	84
10010	Black Mirror	Dark comedy	Null	2011	5	83
10014	The umbrella academy	Fantasy	Null	2019	3	86
10001	The crown	Drama	Null	2016	5	82
Null	Resident evil	Horror	Null	Null	Null	Null
10007	Narcos	Drama	Null	2015	3	80
10012	The watcher	Thriller	Joe Charbanic	2022	1	74
10002	Orange is the new black	Drama	Null	2013	7	77

Esempi

ID_Film	Titolo	Genere	Regista	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	Tim Burton	2022	1	86
10009	Lucifer	Fantasy	Null	2015	6	85
10004	Stranger Things	Fantascienza, avventura	Shawn Levy	2016	4	86
10005	Cobra Kai	Azione, sportivo	Null	2018	5	82
10003	1899	Mistery	Null	2022	1	Null
10011	La casa di carta	Drama	Null	2017	3	83
10013	Dark	Mistery	Baran do Obar	2017	3	84
10010	Black Mirror	Dark comedy	Null	2011	5	83
10014	The umbrella academy	Fantasy	Null	2019	3	86
10001	The crown	Drama	Null	2016	5	82
Null	Resident evil	Horror	Null	Null	Null	Null
10007	Narcos	Drama	Null	2015	3	80
10012	The watcher	Thriller	Joe Charbanic	2022	1	74
10002	Orange is the new black	Drama	Null	2013	7	77

Esempi

ID_Film	Titolo	Genere	Regista	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	Tim Burton	2022	1	86
10009	Lucifer	Fantasy	Null	2015	6	85
10004	Stranger Things	Fantascienza, avventura	Shawn Levy	2016	4	86
10005	Cobra Kai	Azione, sportivo	Null	2018	5	82
10003	1899	Mistery	Null	2022	1	Null
10011	La casa di carta	Drama	Null	2017	3	83
10013	Dark	Mistery	Baran do Obar	2017	3	84
10010	Black Mirror	Dark comedy	Null	2011	5	83
10014	The umbrella academy	Fantasy	Null	2019	3	86
10001	The crown	Drama	Null	2016	5	82
~~Null~~	~~Resident evil~~	~~Horror~~	~~Null~~	~~Null~~	~~Null~~	~~Null~~
10007	Narcos	Drama	Null	2015	3	80
10012	The watcher	Thriller	Joe Charbanic	2022	1	74
10002	Orange is the new black	Drama	Null	2013	7	77

Esempi

ID_Film	Titolo	Genere	~~Regista~~	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	~~Tim Burton~~	2022	1	86
10009	Lucifer	Fantasy	~~Null~~	2015	6	85
10004	Stranger Things	Fantascienza, avventura	~~Shawn Levy~~	2016	4	86
10005	Cobra Kai	Azione, sportivo	~~Null~~	2018	5	82
10003	1899	Mistery	~~Null~~	2022	1	Null
10011	La casa di carta	Drama	~~Null~~	2017	3	83
10013	Dark	Mistery	~~Baran-do-Obar~~	2017	3	84
10010	Black Mirror	Dark comedy	~~Null~~	2011	5	83
10014	The umbrella academy	Fantasy	~~Null~~	2019	3	86
10001	The crown	Drama	~~Null~~	2016	5	82
~~Null~~	~~Resident evil~~	~~Horror~~	~~Null~~	~~Null~~	~~Null~~	~~Null~~
10007	Narcos	Drama	~~Null~~	2015	3	80
10012	The watcher	Thriller	~~Joe Charbanic~~	2022	1	74
10002	Orange is the new black	Drama	~~Null~~	2013	7	77

Esempi

ID_Film	Titolo	Genere	~~Regista~~	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	~~Tim Burton~~	2022	1	86
10009	Lucifer	Fantasy	~~Null~~	2015	6	85
10004	Stranger Things	Fantascienza, avventura	~~Shawn Levy~~	2016	4	86
10005	Cobra Kai	Azione, sportivo	~~Null~~	2018	5	82
10003	1899	Mistery	~~Null~~	2022	1	82
10011	La casa di carta	Drama	~~Null~~	2017	3	83
10013	Dark	Mistery	~~Baran-do-Obar~~	2017	3	84
10010	Black Mirror	Dark comedy	~~Null~~	2011	5	83
10014	The umbrella academy	Fantasy	~~Null~~	2019	3	86
10001	The crown	Drama	~~Null~~	2016	5	82
~~Null~~	~~Resident evil~~	~~Horror~~	~~Null~~	~~Null~~	~~Null~~	~~Null~~
10007	Narcos	Drama	~~Null~~	2015	3	80
10012	The watcher	Thriller	~~Joe Charbanic~~	2022	1	74
10002	Orange is the new black	Drama	~~Null~~	2013	7	77

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
002	10007	2021-09-10	84
001	10003	2022-12-27	70
002	10005	2020-10-21	80
002	10002	2020-03-03	77
001	10001	2022-09-01	56
002	10009	2022-12-01	63
002	10011	2022-12-28	91

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
002	10007	2021-09-10	84
001	10003	2022-12-27	70
002	10005	2020-10-21	80
002	10002	2020-03-03	77
001	10001	2022-09-01	56
002	10009	2022-12-01	63
002	10011	2022-12-28	91

- Quali sono le domande a cui possiamo rispondere?
- Può essere utile partire “riordinando” un po’ i dati

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

- Quali sono le domande a cui possiamo rispondere?
- Può essere utile partire “riordinando” un po’ i dati

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

- **Quali sono le domande a cui possiamo rispondere?**
- **Può essere utile partire "riordinando" un po' i dati**
- **Possiamo fare un paio di osservazioni:**
  - L'utente 001 è "più recente" dell'utente 001
  - Il voto medio di 001 è 75.5, quello di 002 è 79

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

- **Quali sono le domande a cui possiamo rispondere?**
- **Può essere utile partire "riordinando" un po' i dati**
- **Possiamo fare un paio di osservazioni:**
  - L'utente 001 è "più recente" dell'utente 001
  - Il voto medio di 001 è 75.5, quello di 002 è 79
- **Questo è solo un esempio con pochi dati, ma se ne avessimo molti potremmo...**

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

- **Quali sono le domande a cui possiamo rispondere?**
- **Può essere utile partire "riordinando" un po' i dati**
- **Possiamo fare un paio di osservazioni:**
  - L'utente 001 è "più recente" dell'utente 002
  - Il voto medio di 001 è 75.5, quello di 002 è 79
- **Questo è solo un esempio con pochi dati, ma se ne avessimo molti potremmo…**
- **Le domande possono aumentare se mettiamo insieme informazioni che arrivano da più tabelle**

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

ID_Film	Titolo	Genere	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	2022	1	86
10009	Lucifer	Fantasy	2015	6	85
10004	Stranger Things	Fantascienza, avventura	2016	4	86
10005	Cobra Kai	Azione, sportivo	2018	5	82
10003	1899	Mistery	2022	1	82
10011	La casa di carta	Drama	2017	3	83
10013	Dark	Mistery	2017	3	84
10010	Black Mirror	Dark comedy	2011	5	83
10014	The umbrella academy	Fantasy	2019	3	86
10001	The crown	Drama	2016	5	82
10007	Narcos	Drama	2015	3	80
10012	The watcher	Thriller	2022	1	74
10002	Orange is the new black	Drama	2013	7	77

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

ID_Film	Titolo	Genere	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	2022	1	86
10009	Lucifer	Fantasy	2015	6	85
10004	Stranger Things	Fantascienza, avventura	2016	4	86
10005	Cobra Kai	Azione, sportivo	2018	5	82
10003	1899	Mistery	2022	1	82
10011	La casa di carta	Drama	2017	3	83
10013	Dark	Mistery	2017	3	84
10010	Black Mirror	Dark comedy	2011	5	83
10014	The umbrella academy	Fantasy	2019	3	86
10001	The crown	Drama	2016	5	82
10007	Narcos	Drama	2015	3	80
10012	The watcher	Thriller	2022	1	74
10002	Orange is the new black	Drama	2013	7	77

Esempi

ID_Utente	ID_Film	Data	Score
001	10006	2022-10-10	86
001	10009	2022-12-20	90
001	10003	2022-12-27	70
001	10001	2022-09-01	56
002	10007	2021-09-10	84
002	10005	2020-10-21	80
002	10002	2020-03-03	77
002	10009	2022-12-01	63
002	10011	2022-12-28	91

ID_Film	Titolo	Genere	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	2022	1	86
10009	Lucifer	Fantasy	2015	6	85
10004	Stranger Things	Fantascienza, avventura	2016	4	86
10005	Cobra Kai	Azione, sportivo	2018	5	82
10003	1899	Mistery	2022	1	82
10011	La casa di carta	Drama	2017	3	83
10013	Dark	Mistery	2017	3	84
10010	Black Mirror	Dark comedy	2011	5	83
10014	The umbrella academy	Fantasy	2019	3	86
10001	The crown	Drama	2016	5	82
10007	Narcos	Drama	2015	3	80
10012	The watcher	Thriller	2022	1	74
10002	Orange is the new black	Drama	2013	7	77

Esempi

ID_Utente	ID_Film	Data	Score	Titolo	Genere	Anno	Stagioni	TMDBScore
001	10006	2022-10-10	86	Mercoledi	Dark fantasy	2022	1	86
001	10009	2022-12-20	90	Lucifer	Fantasy	2015	6	85
001	10003	2022-12-27	70	1899	Mistery	2022	1	82
001	10001	2022-09-01	56	The crown	Drama	2016	5	82
002	10007	2021-09-10	84	Narcos	Drama	2015	3	80
002	10005	2020-10-21	80	Cobra Kai	Azione, sportivo	2018	5	82
002	10002	2020-03-03	77	Orange is the new black	Drama	2013	7	77
002	10009	2022-12-01	63	Lucifer	Fantasy	2015	6	85
002	10011	2022-12-28	91	La casa di carta	Drama	2017	3	83

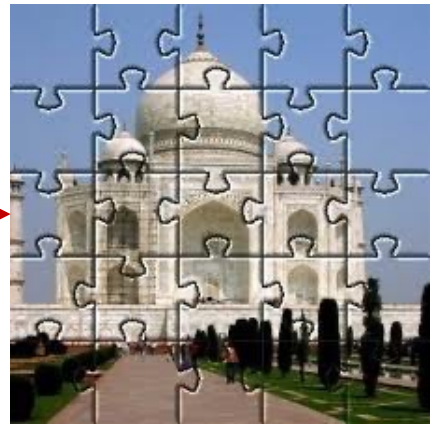
Esempi

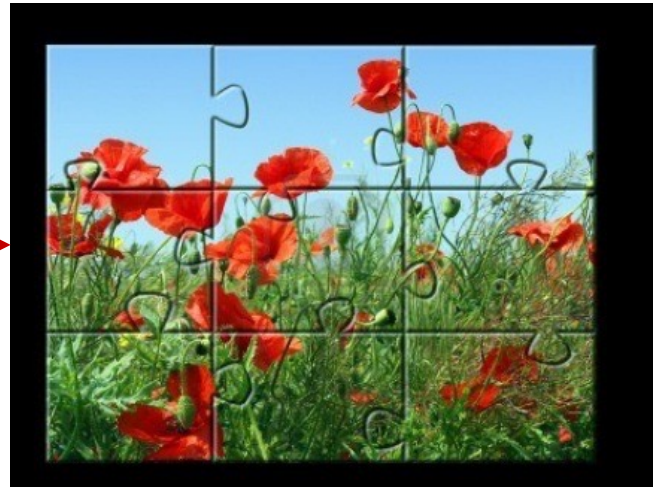
A quali domande possiamo rispondere adesso?

ID_Utente	ID_Film	Data	Score	Titolo	Genere	Anno	Stagioni	TMDBScore
001	10006	2022-10-10	86	Mercoledi	Dark fantasy	2022	1	86
001	10009	2022-12-20	90	Lucifer	Fantasy	2015	6	85
001	10003	2022-12-27	70	1899	Mistery	2022	1	82
001	10001	2022-09-01	56	The crown	Drama	2016	5	82
002	10007	2021-09-10	84	Narcos	Drama	2015	3	80
002	10005	2020-10-21	80	Cobra Kai	Azione, sportivo	2018	5	82
002	10002	2020-03-03	77	Orange is the new black	Drama	2013	7	77
002	10009	2022-12-01	63	Lucifer	Fantasy	2015	6	85
002	10011	2022-12-28	91	La casa di carta	Drama	2017	3	83

Integrazione

Cosa significa fare integrazione di dati? E’ come risolvere tanti puzzle





Un altro esempio

- Immaginiamo di avere a disposizione una seconda tabella con dei dati riferiti ad un’altra piattaforma

Esempi

ID_Film	Titolo	Genere	Anno	TMDBScore
aa016	Cinderella	animation	1950	7,035
aa039	The jungle book	animation	1967	7,281
aa073	Robin Hood	animation	1973	7,3
aa021	Pinocchio	animation	1940	7,1
aa027	The three Caballeros	animation	1935	6,34
aa095	Pluto and the gopher	animation	1950	6,7
aa024	Rescue Dog	animation	1947	6,3
aa001	The Muppet Show	comedy	1976	8,047
aa011	Mary Poppins	fantasy	1964	7,569

Un altro esempio

- Immaginiamo di avere a disposizione una seconda tabella con dei dati riferiti ad un’altra piattaforma

- Come possiamo arricchire la tabella precedente con questi nuovi dati?

- Occorre capire quali siano le informazioni in comune e se serva fare trasformazioni

Esempi

ID_Film	Titolo	Genere	Anno	TMDBScore
aa016	Cinderella	animation	1950	7,035
aa039	The jungle book	animation	1967	7,281
aa073	Robin Hood	animation	1973	7,3
		……		

ID_Film	Titolo	Genere	Anno	Stagioni	TMDBScore
10006	Mercoledi	Dark fantasy	2022	1	86
10009	Lucifer	Fantasy	2015	6	85
10004	Stranger Things	Fantascienza, avventura	2016	4	86
10005	Cobra Kai	Azione, sportivo	2018	5	82
		……			

Esempi di trasformazioni

- Conversioni (es. unità di misura)
- Traduzioni
- …

Ma anche creare nuove colonne:

- ”Separare” le informazioni (es. Data ‘2023-03-07’ → Anno 2023, Mese 03, AnnoMese 2023,03)
- Raggruppare le informazioni (es. Stato –> Continente)

Inoltre alcune colonne potrebbero essere inutili → Il concetto di utilità/inutilità dipende dall’obiettivo analitico

Una nota sull'efficienza

- In presenza di molte tabelle
  - Fare join è costoso
  - Trasmettere informazioni è costoso

- Se la "dimensione del problema" lo consente, conviene costruire una tabella unica → denormalizzazione
- Dopo il laboratorio 1, noi ragioneremo sempre su singola tabella (possiamo immaginare di avere una tabella per obiettivo analitico)

Cosa farete in laboratorio

- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia

Cosa farete in laboratorio

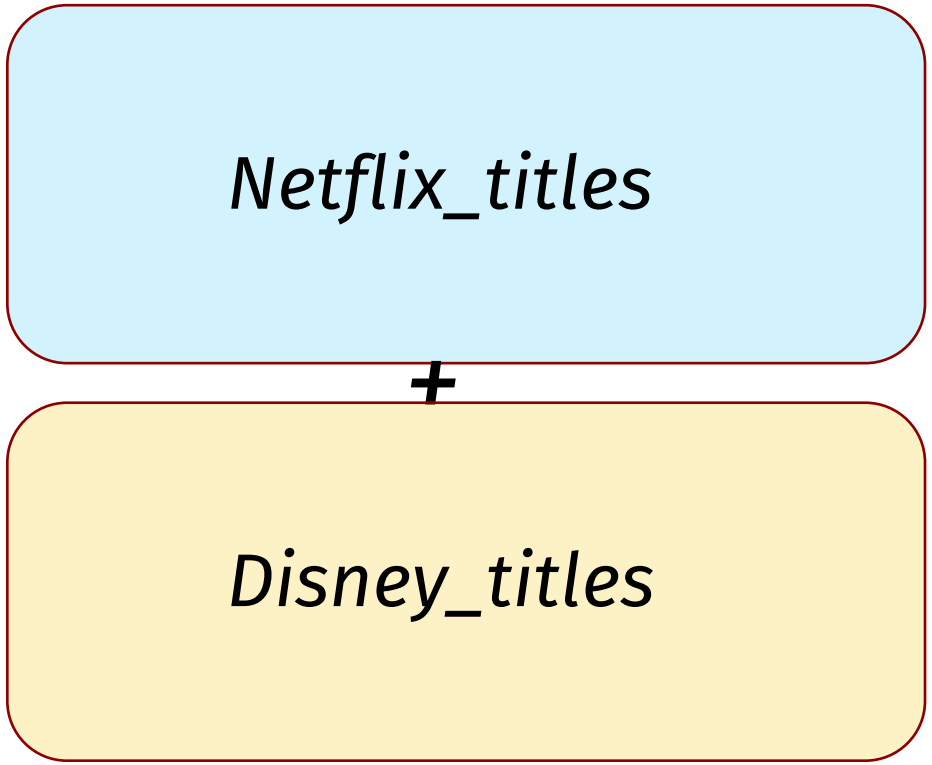
- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia

Netflix_titles

Disney_titles

Cosa farete in laboratorio

- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia



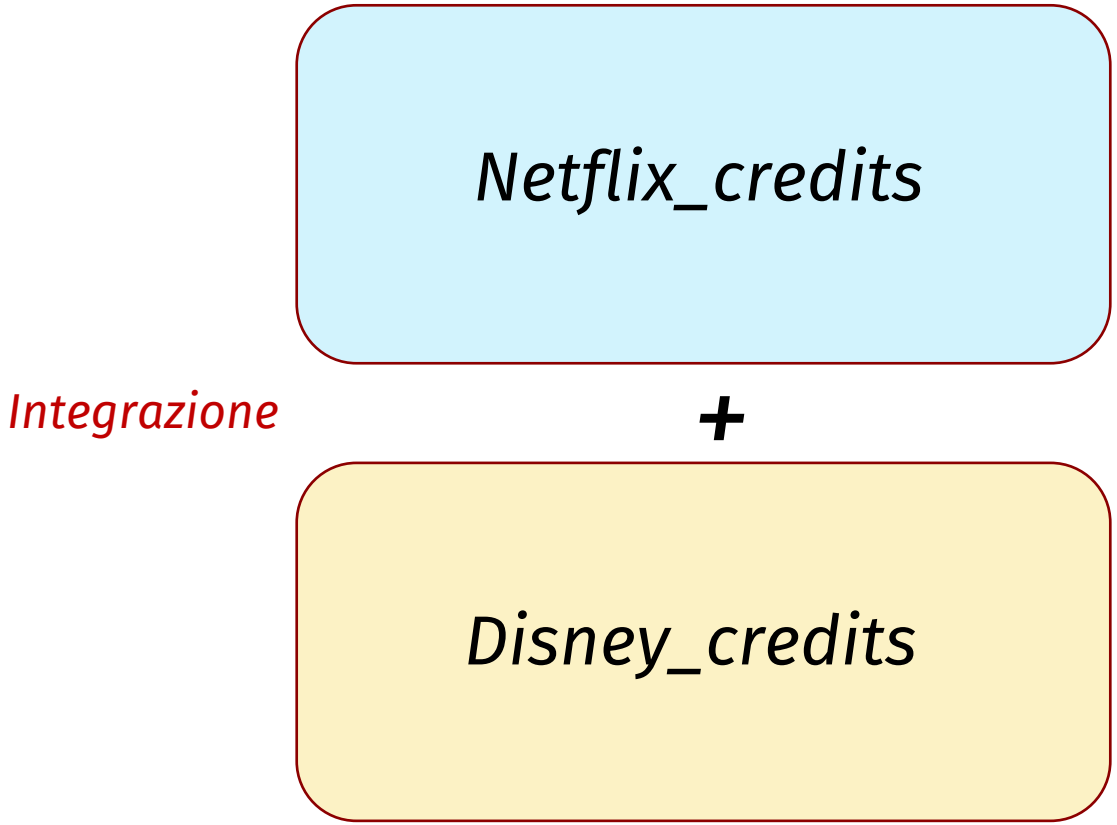
Integrazione

Cosa farete in laboratorio

- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia

Cosa farete in laboratorio

- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia



Cosa farete in laboratorio

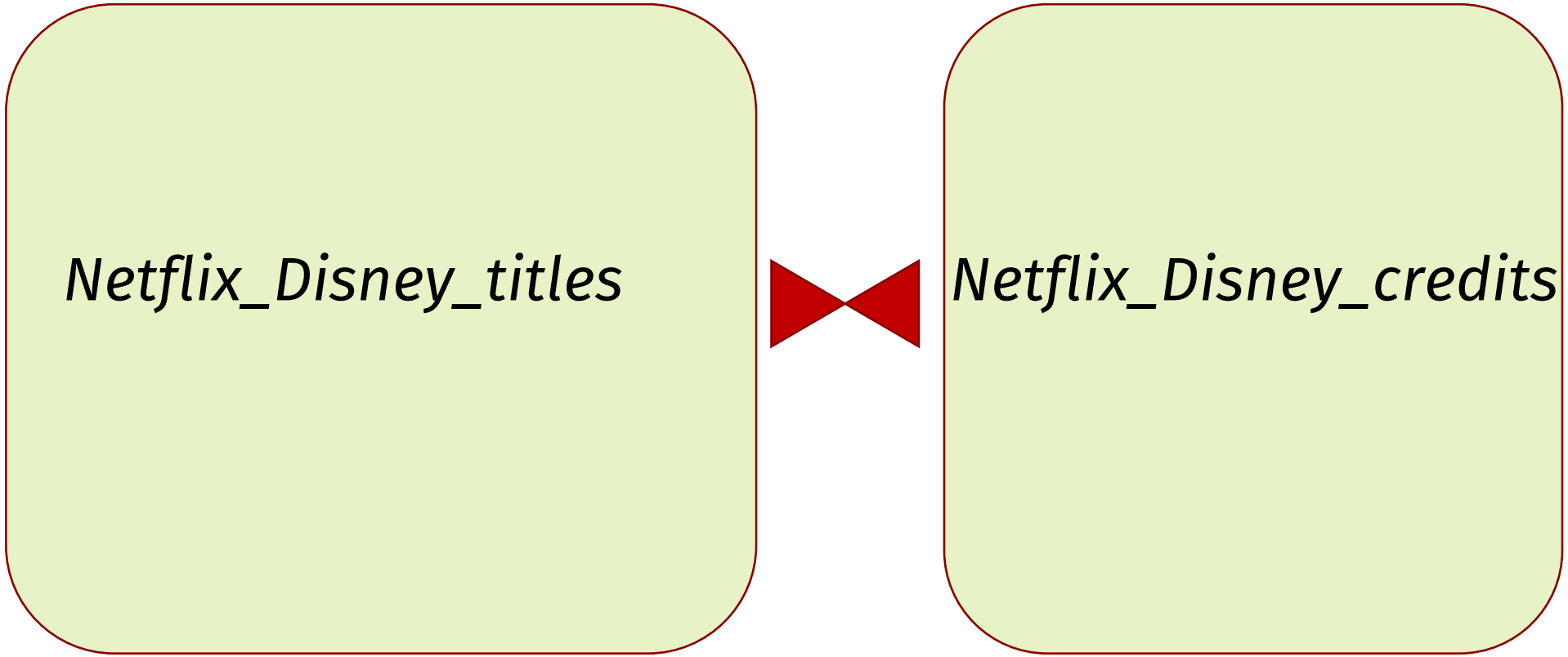
- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia

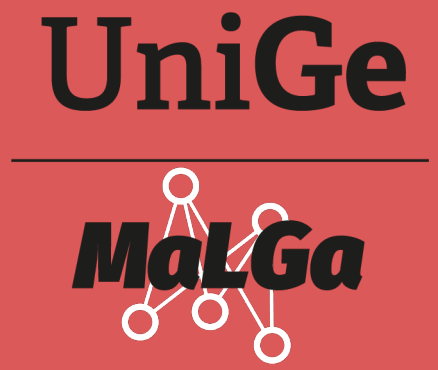
Netflix_Disney_titles

Netflix_Disney_credits

Cosa farete in laboratorio

- Avrete a disposizione diverse tabelle su cui applicare metodi di integrazione e pulizia




UniGe
MaLGa