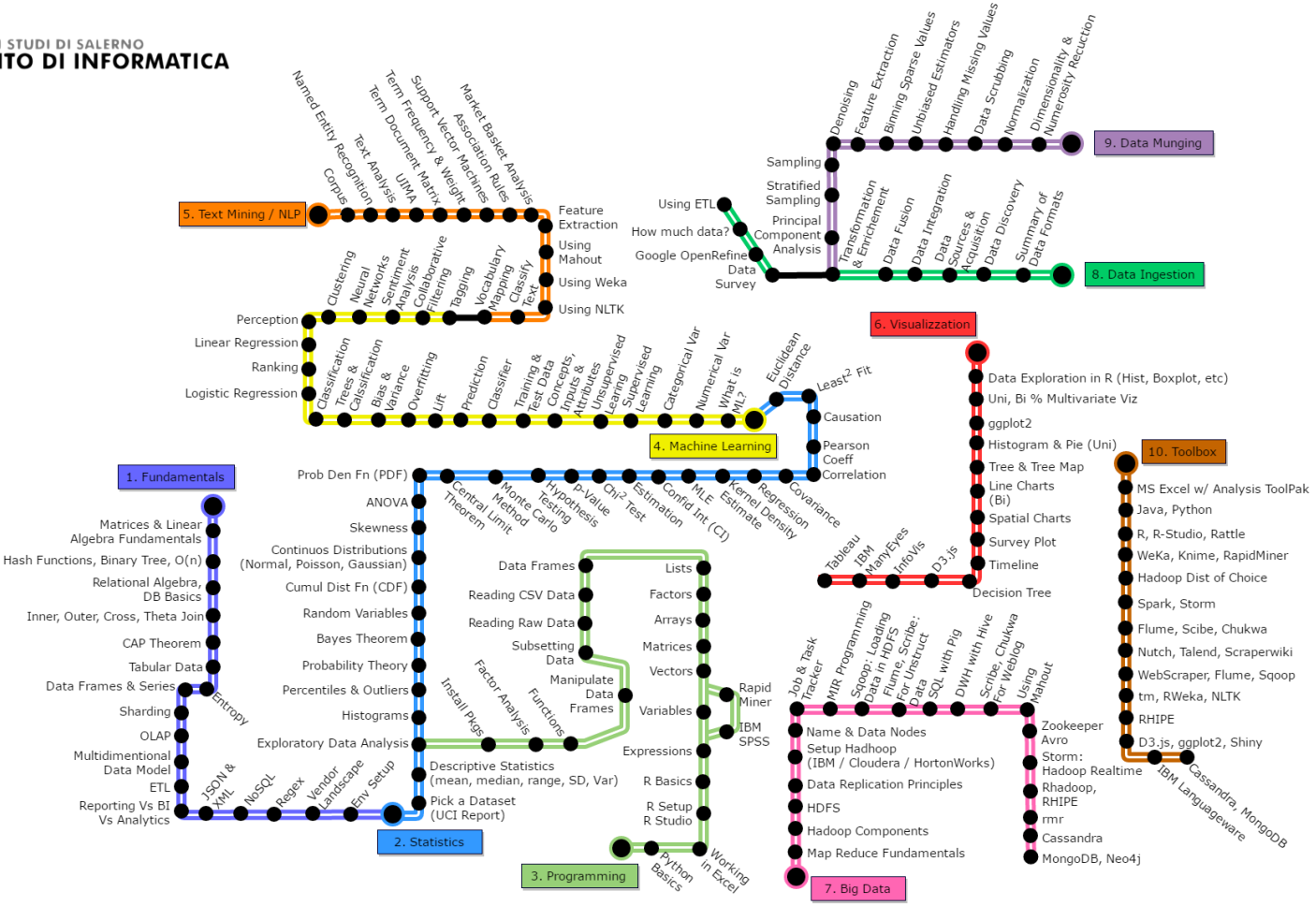
UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA



Fondamenti di Data Science & Machine Learning

Data Integration

Prof. Giuseppe Polese, aa 2024-25

Outline

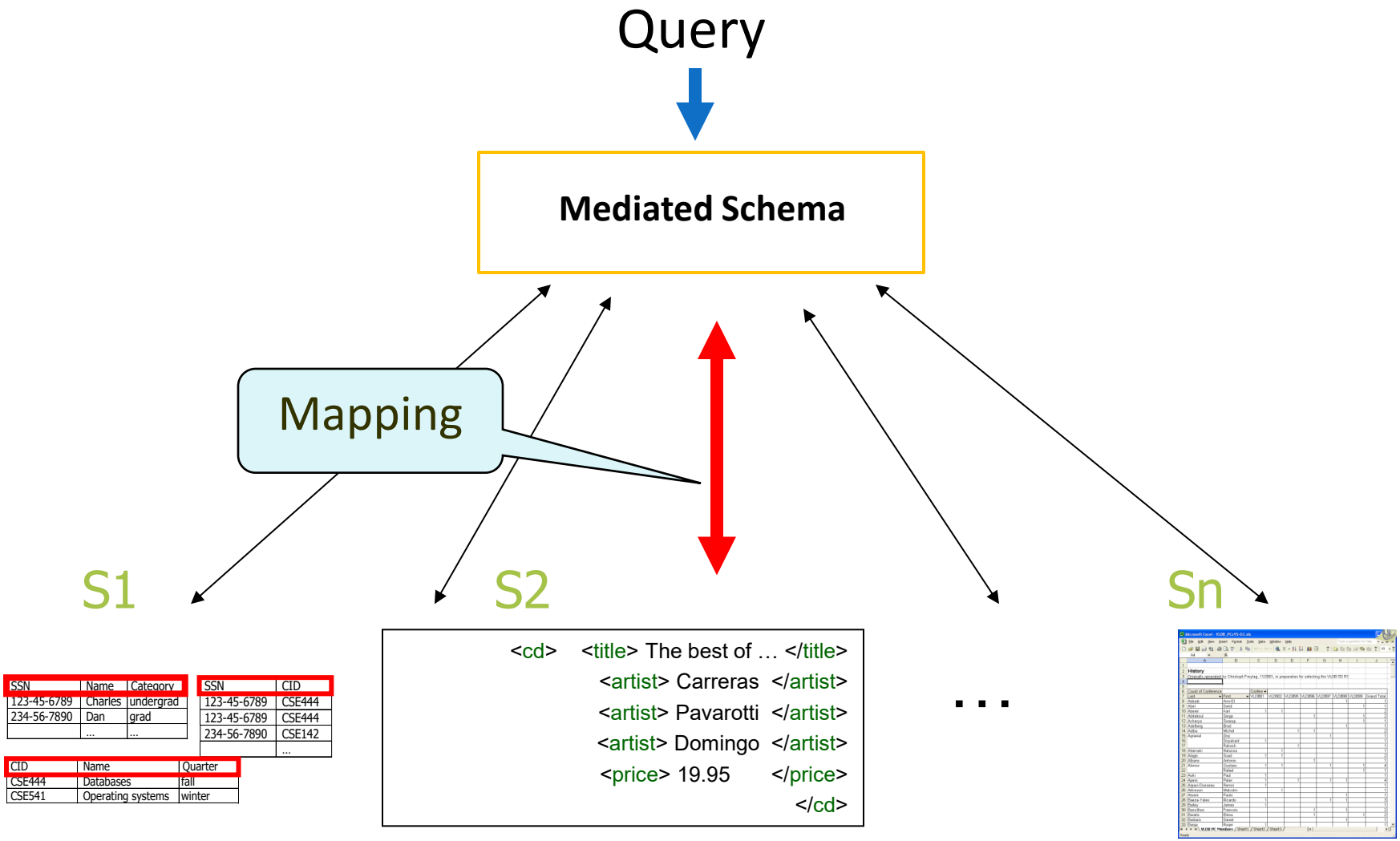
- Data Integration
- Esempi di applicazione
- Obiettivi e problematiche
- Architetture di integrazione
  - Database materializzati
  - Database virtuali (non materializzati)
- Integrazione degli schemi

Data Integration

- I database vengono utilizzati per gestire una grossa quantità di dati, assumendo che siano stati definiti degli schemi
- In realtà
  - spesso vengono creati insiemi di dati in maniera indipendente
  - Successivamente può nascere la necessità di combinarli

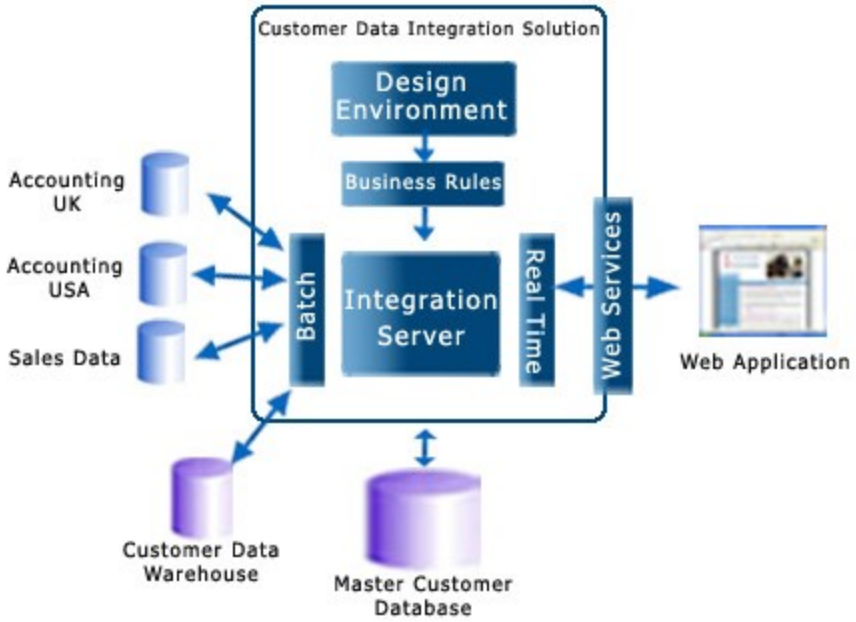
La data integration rappresenta il processo di integrazione di dati provenienti da sorgenti informative multiple, distribuite, autonome ed eterogenee, con lo scopo di fornire agli utenti l'accesso uniforme agli stessi

Un'astrazione ad alto livello



Esempi di applicazione: Business

- Motivazioni
  - Proliferare incontrollato dei sistemi informativi aziendali
  - Disgregazione delle informazioni tra i vari reparti
  - Erogazione di servizi inter-aziendali
- Tipologie
  - Enterprise Resource Planning
  - CRM
  - Portali

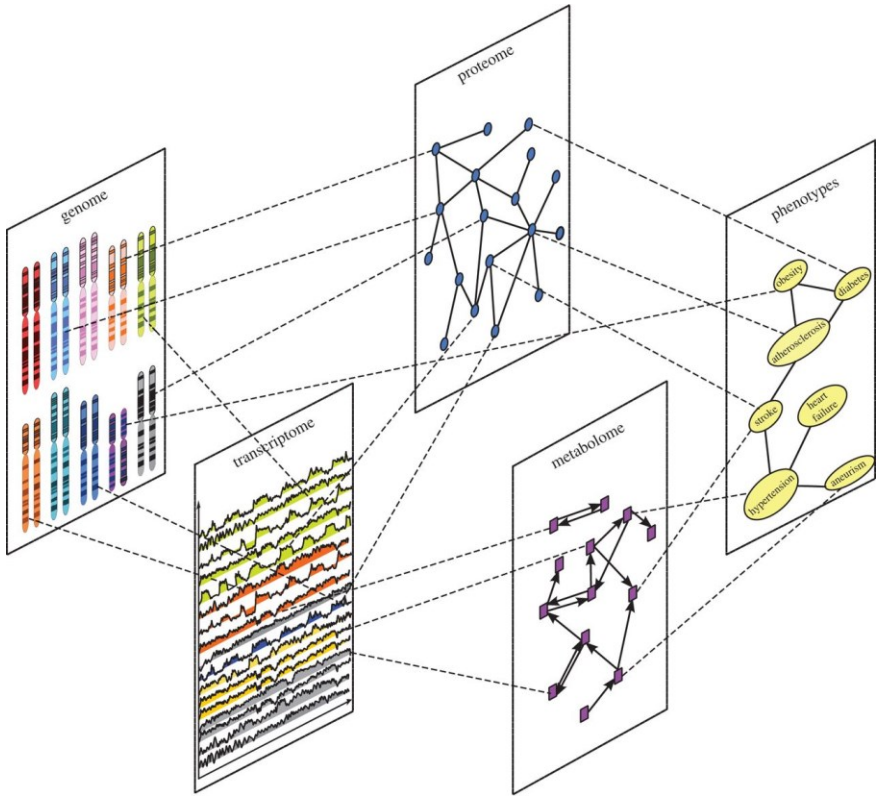


Esempi di applicazione: Scienza

- Motivazioni
  - È possibile avere a disposizione centinaia di sorgenti di dati di specifici contesti scientifici
- Tipologie
  - Genetica

    Integrazione di dati genetici
  - Astrofisica

    Monitoring di eventi in cielo



Esempi di applicazione: Web

- Motivazioni
  - Sono disponibili milioni di moduli HTML di alta qualità
- Tipologie
  - Integrazione di dati provenienti da siti Web differenti
  - Shopping comparativo
  - B2B
  - Mercati elettronici

Obiettivi

La data integration rappresenta il processo di integrazione di dati provenienti da sorgenti informative multiple, distribuite, autonome ed eterogenee, con lo scopo di fornire agli utenti l'accesso uniforme agli stessi

- Sorgenti Informative: non solo database
- Multiple: molte sorgenti, ma anche per due risulta complesso
- Distribuite: su LAN, WAN, o Internet
- Autonome: non si conoscono i DBA delle singole sorgenti
- Eterogenee: i modelli dei dati possono essere differenti
- Accesso: effettuare interrogazioni, e in qualche caso modifiche
- Uniforme: stessa interfaccia di interrogazione

Problematiche da gestire

- Nel processo di integrazione diventa necessario gestire problematiche su diversi livelli di astrazione
- Livello di sistema
  - Gestire diverse piattaforme
  - Utilizzare SQL su sistemi multipli non è molto semplice
  - Query processing distribuito
- Livello logico
  - Eterogeneità degli schemi (e dei dati)
- Livello “sociale”
  - Localizzare e catturare solo i dati rilevanti nelle aziende
  - Convincere le persone a condividere (sicurezza, privacy, …)

Problematiche da gestire

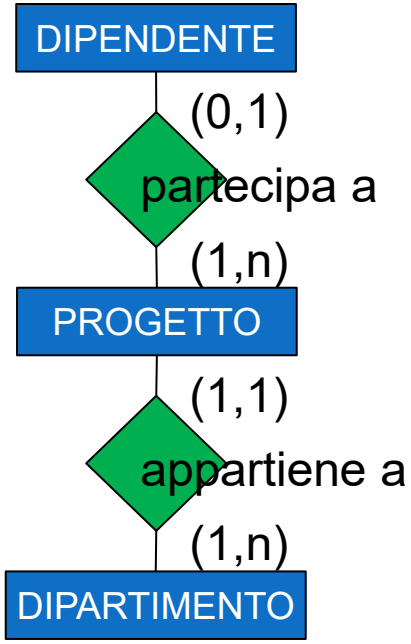
- Nel processo di integrazione diventa necessario gestire problematiche su diversi livelli di astrazione
- Livello di sistema
  - Gestire diverse piattaforme
  - Utilizzare SQL su sistemi multipli non è molto semplice
  - Query processing distribuito
- Livello logico
  - Eterogeneità degli schemi (e dei dati)
- Livello "sociale"
  - Localizzare e catturare solo i dati rilevanti nelle aziende
  - Convincere le persone a condividere (sicurezza, privacy, …)

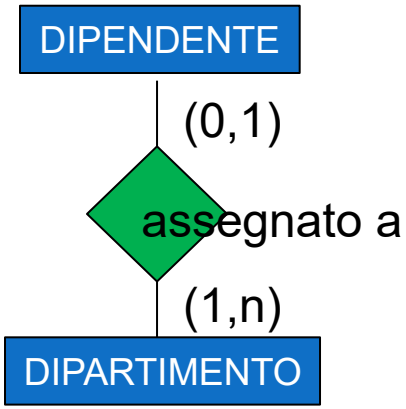
Eterogeneità degli schemi

- L'eterogeneità degli schemi può portare ad avere problemi in fase di integrazione causati da:
  - Diversità di prospettiva
  - Equivalenza dei costrutti del modello
  - Incompatibilità delle specifiche
  - Concetti comuni
  - Concetti correlati

Diversità di prospettiva

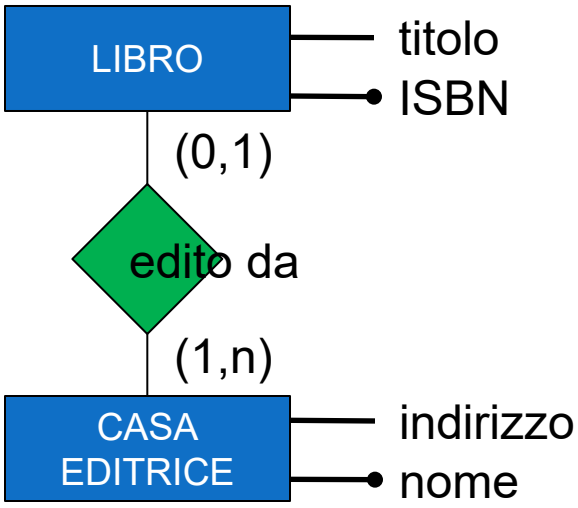
- Il punto di vista rispetto al quale diversi gruppi di utenti vedono uno stesso oggetto del dominio applicativo può differenziarsi notevolmente in base agli aspetti rilevanti per la funzione a cui essi sono preposti

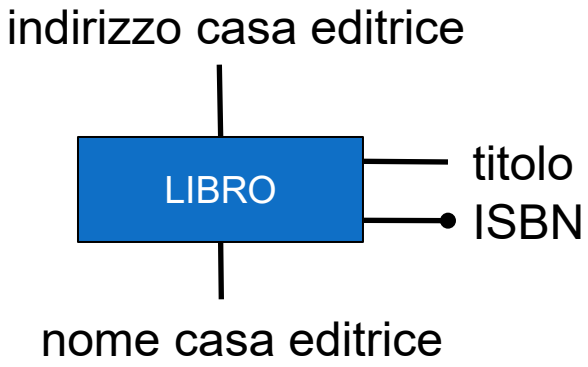




Equivalenza dei costrutti del modello

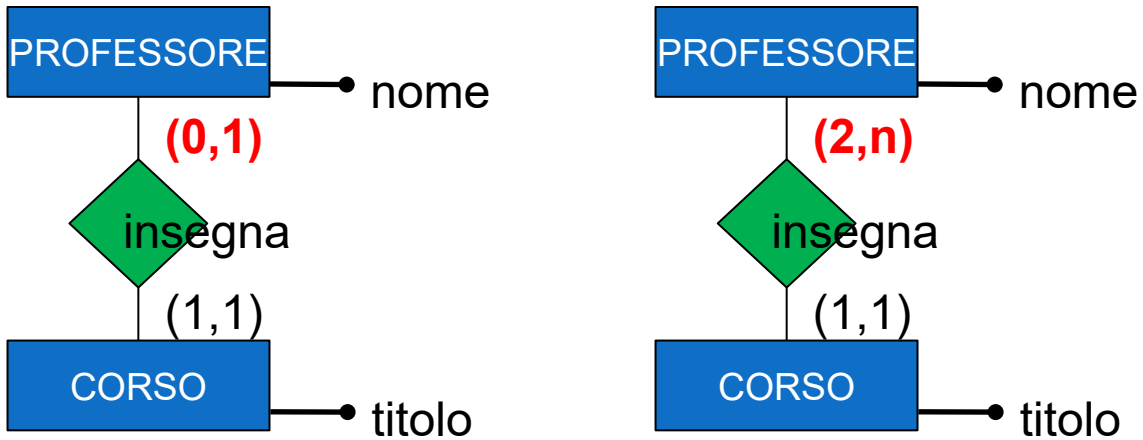
- Tipicamente, i formalismi di modellazione permettono di rappresentare uno stesso concetto utilizzando combinazioni diverse dei costrutti a disposizione





Incompatibilità delle specifiche

- L'incompatibilità delle specifiche indica che schemi diversi che modellano una stessa porzione del dominio applicativo racchiudono concetti in contrasto tra loro
  - Tale diversità deriva da errate scelte progettuali che possono coinvolgere ad esempio la scelta dei nomi, dei tipi di dati e dei vincoli di integrità
  - **Esempio:** in un caso un professore non può tenere più di un corso, nell'altro deve tenerne almeno 2



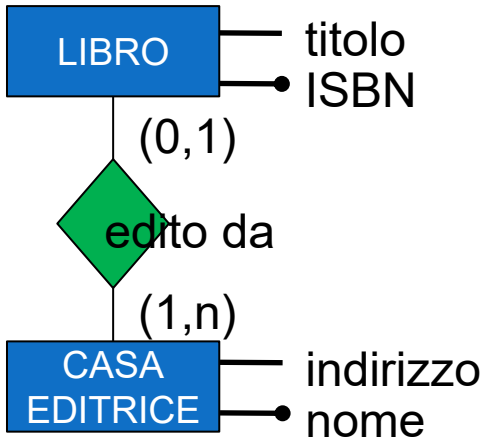
Concetti comuni

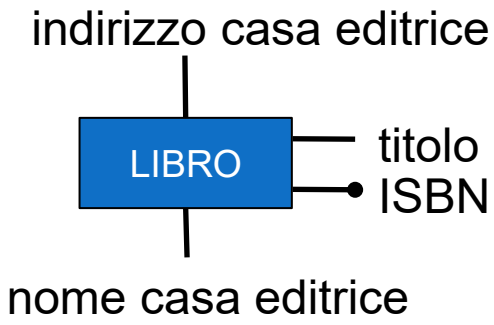
- Quattro sono le possibili relazioni esistenti tra due distinte rappresentazioni R_1 e R_2 di uno stesso concetto:
  - Identità
  - Equivalenza
  - Comparabilità
  - Incompatibilità

Identità ed Equivalenza

- **Identità**: si verifica quando vengono utilizzati gli stessi costrutti, il concetto è modellato dallo stesso punto di vista, quindi R_1 e R_2 coincidono
- **Equivalenza**: si verifica quando R_1 e R_2 non sono le stesse poiché sono stati utilizzati costrutti diversi ma equivalenti
  - Tra le varie definizioni di equivalenza:
    - Due schemi R_1e R_2 sono equivalenti se le loro istanze possono essere messe in corrispondenza 1-a-1

Esempio di Equivalenza





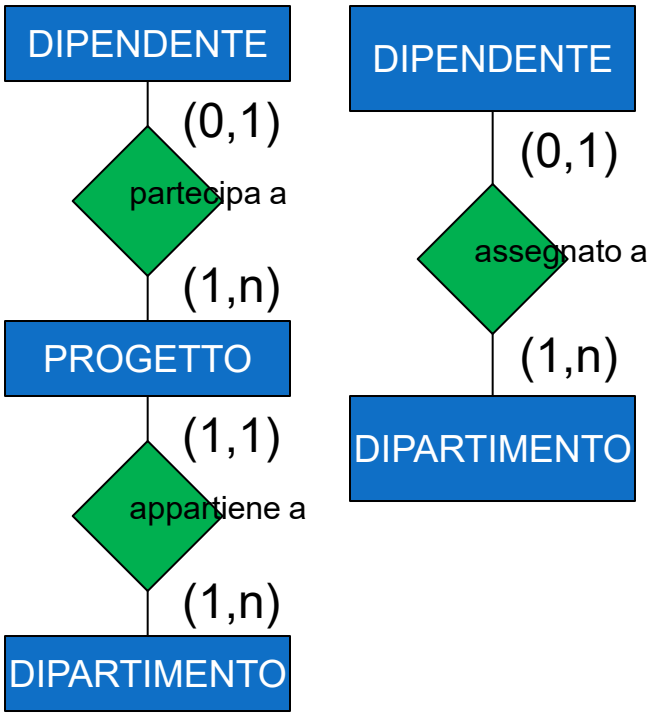
LIBRO		
ISBN	**titolo**	**casa editrice**
123445	Il DFM	McGraw-Hill
435454	Mi sembra logico	Apogeo
…	…	…

CASA EDITRICE	
nome	**indirizzo**
McGraw-Hill	Via Ripamonti, 89
Apogeo	Via Verdi, 45
…	…

LIBRO			
ISBN	**titolo**	**nome c.e.**	**Indirizzo c.e.**
123445	Il DFM	McGraw-Hill	Via Ripamonti, 89
435454	Mi sembra logico	Apogeo	Via Verdi, 45
…	…	…	

Comparabilità e Incompatibilità

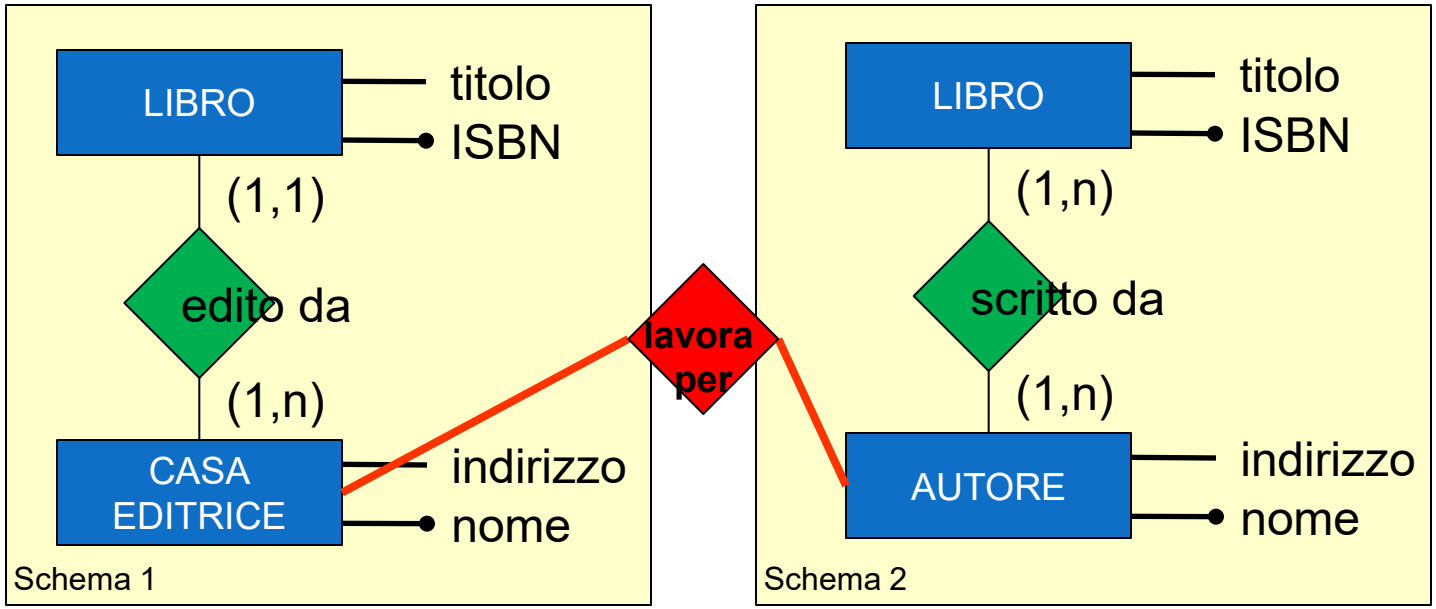
- **Comparabilità**: questa situazione si verifica quando R_1 e R_2 non sono né identiche né equivalenti, ma i costrutti utilizzati e i punti di vista dei progettisti non sono in contrasto tra loro



- **Incompatibilità**: questa situazione si verifica quando R_1 e R_2 sono in contrasto a causa dell'incoerenza nelle specifiche:
  - in altre parole quando la realtà modellata da R_1 nega la realtà modellata da R_2

Concetti correlati

- A seguito dell'integrazione, molti concetti diversi, ma correlati, verranno a trovarsi nello stesso schema, dando vita a nuove relazioni che non erano percepibili in precedenza
  - Tali relazioni sono dette proprietà inter-schema



Eterogeneità dei dati

- L'eterogeneità può sussistere anche dal punto di vista della semantica dei dati
  - Diverse valute: Euro, Dollari, …
  - Diversi sistemi di misurazione:
    - Chilogrammi vs Libbre
    - Centigradi vs Fahrenheit
  - Diverse granularità: Grammi, Chilogrammi, …

Architetture di integrazione

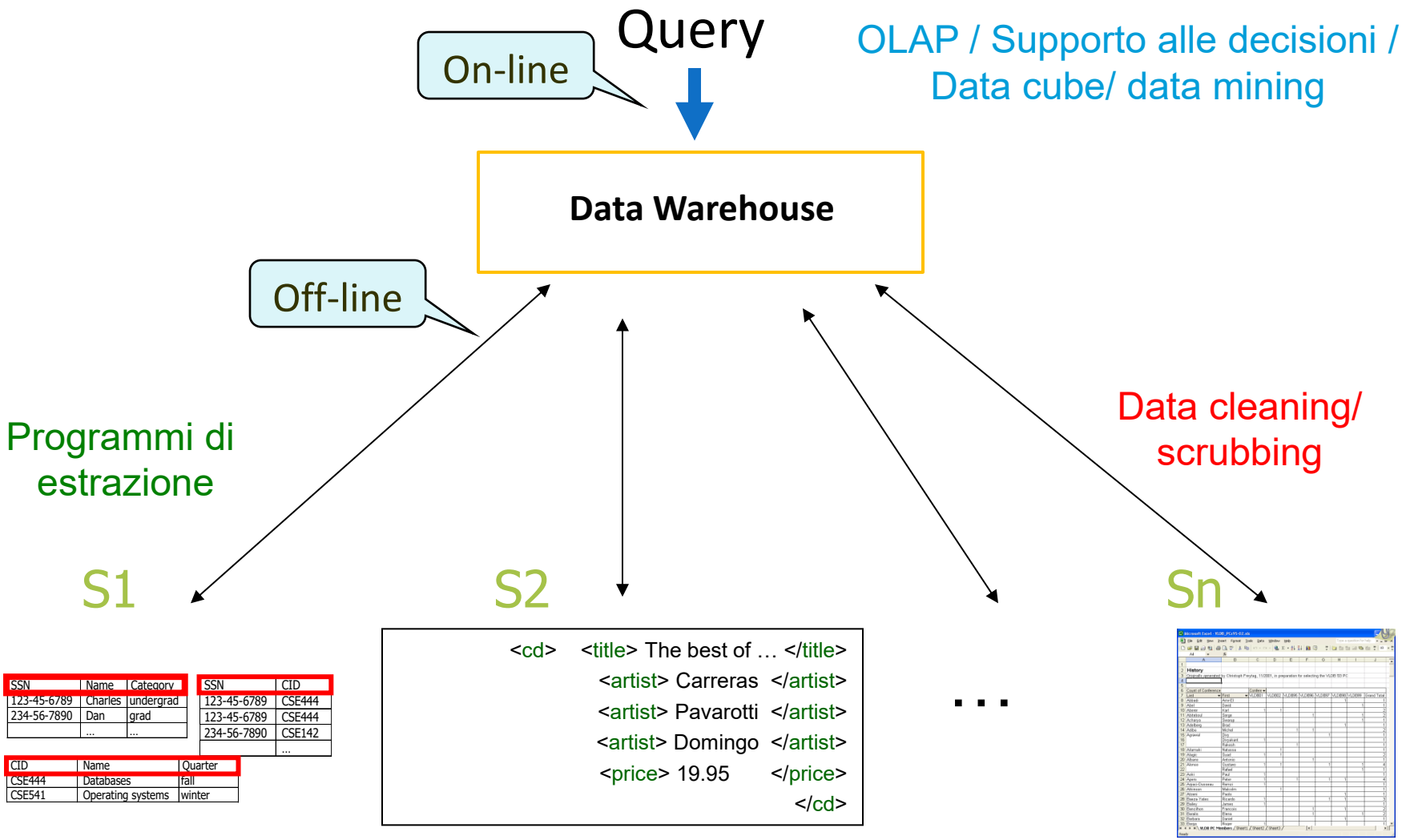
- Dal punto di vista architetturale esistono due principali approcci di integrazione

1. Usare un database materializzato -> Data Warehouse
   - Extract-Transform-Load Systems
   - I dati vengono spostati in un nuovo database integrato
2. Usare un database virtuale (non materializzato) -> Vista Virtuale
   - Enterprise Information Integration Systems, Data Integration Systems, Data Exchange Systems (source-to-target)
   - I dati restano nelle sorgenti

Data Warehouse

- I dati integrati vengono caricati periodicamente in un data warehouse (off-line)
  - Separa i DBMS operativi dai DBMS a supporto alle decisioni
    - In generale, i data warehouse non vengono utilizzati solo per l'integrazione dei dati
  - Le prestazioni sono buone
  - I dati potrebbero non essere aggiornati

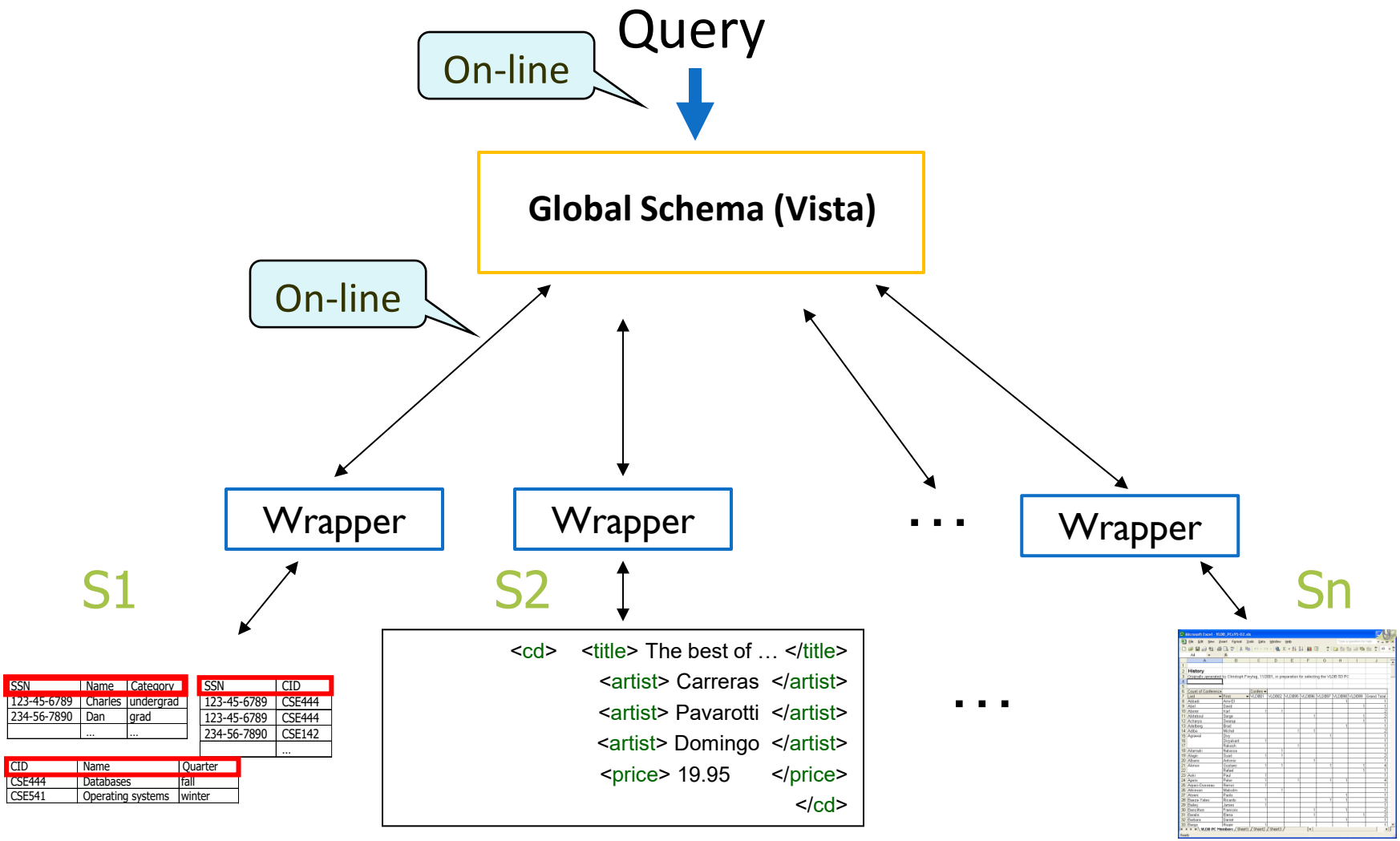
Data Warehouse



Vista Virtuale

- I dati vengono lasciati nelle sorgenti e le query inviate allo schema globale verranno riformulate nei formati delle sorgenti locali (on-line)
  - Devono essere utilizzati i wrapper
    - Modulo che si occupa di inviare le query alle sorgenti e trasformare le risposte in tuple (o in un altro modello di dati interno)
  - I dati saranno sempre aggiornati
  - Aspetti critici:
    - Correlazione tra le sorgenti informative e lo schema di mediazione
    - Query Reformulation

Vista Virtuale



SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	…	…

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

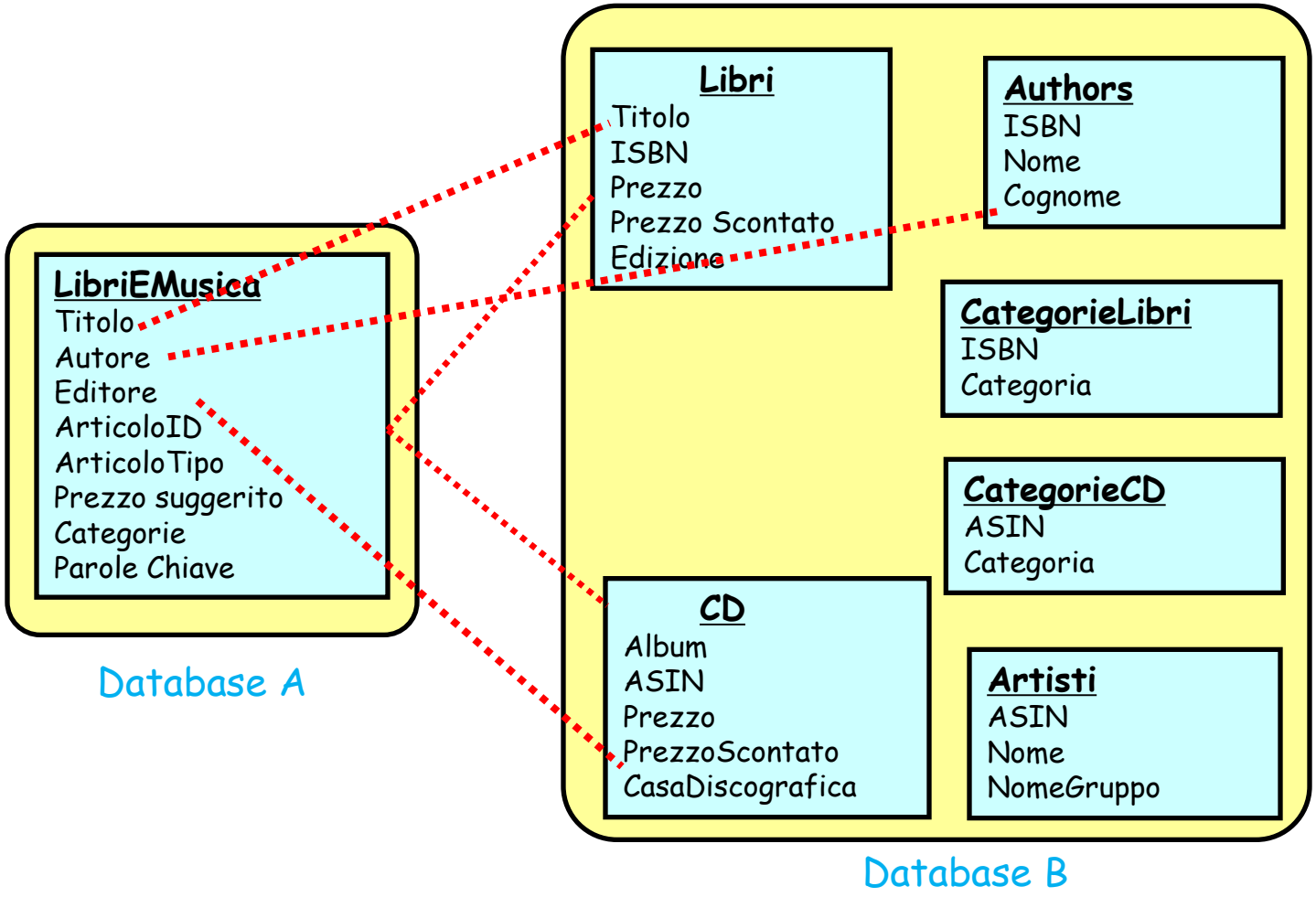
Integrazione degli schemi

- Da un punto di vista logico l'integrazione di sorgenti di dati eterogenee consiste nel
  - Identificazione di concetti correlati
  - Analisi dei conflitti e risoluzione
  - Integrazione degli schemi a livello concettuale

Identificazione dei concetti correlati

- Individuazione di corrispondenze tra i concetti degli schemi locali
  - Se le diverse sorgenti dati modellano porzioni distinte del mondo reale, il problema dell'integrazione non esiste
- Questa fase del processo di integrazione viene anche detta Schema Matching

Schema Matching

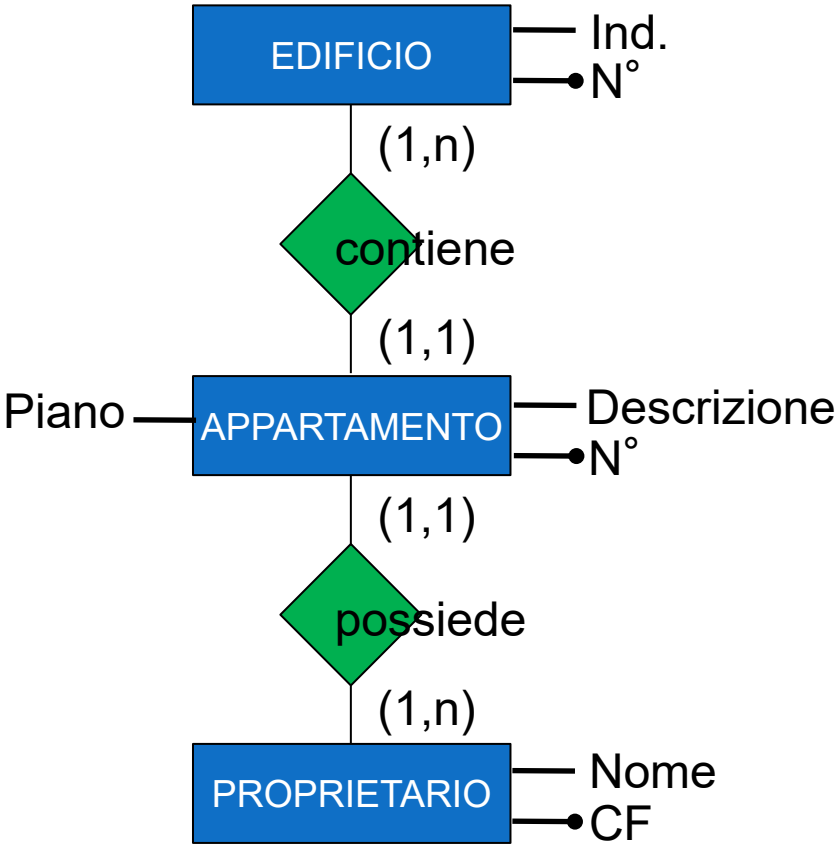


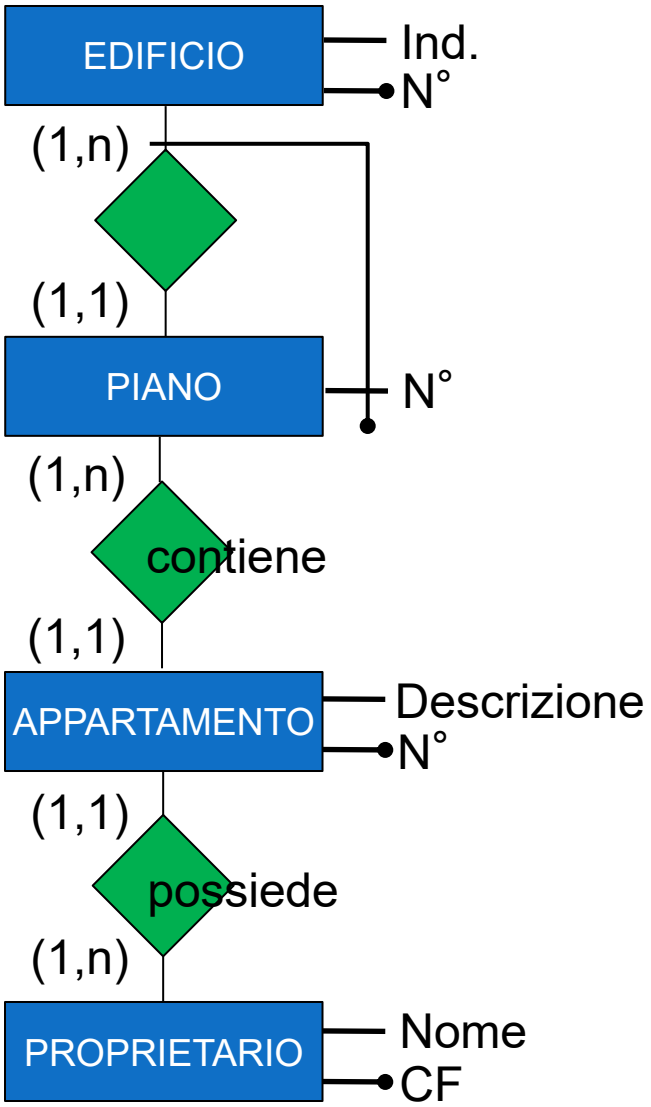
Eventualmente… LibriEMusica(x:Titolo,…) = Libri(x:Titolo,…) ∪ CD(x:Album,…)

Analisi dei conflitti

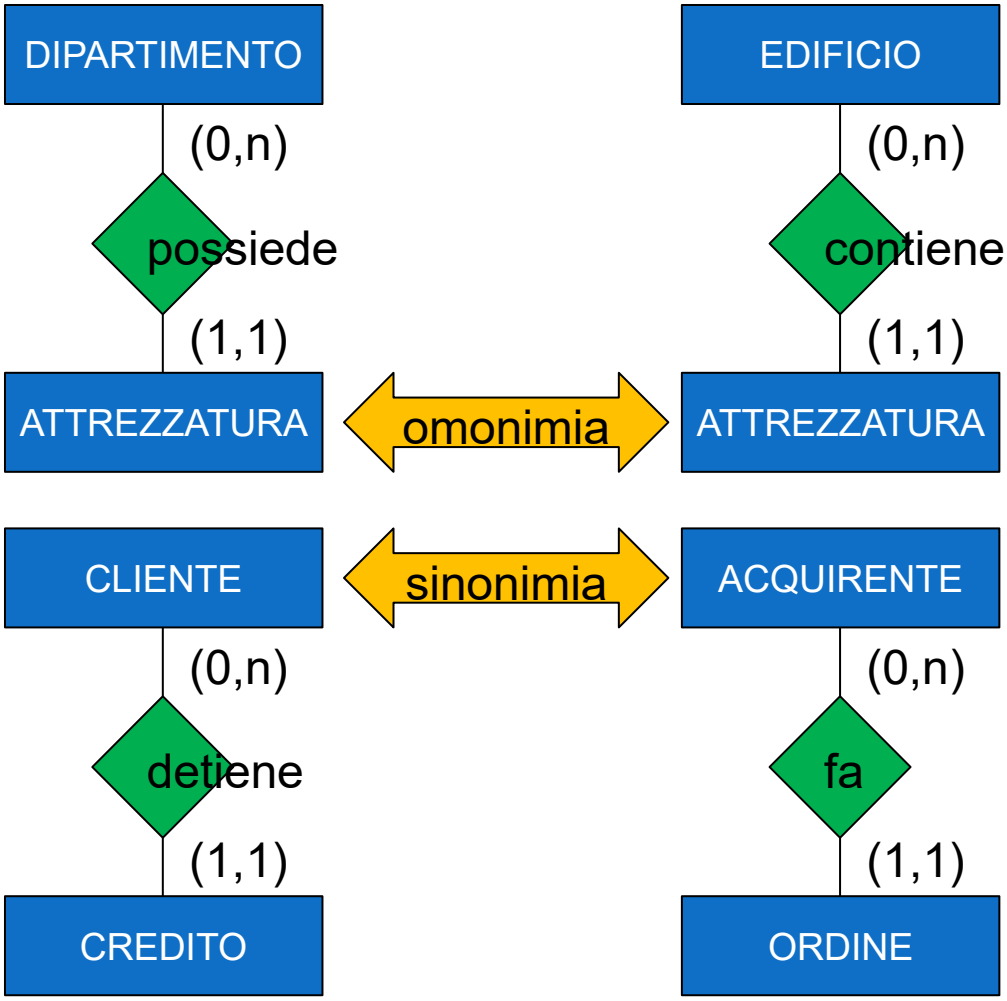
- I tipi di conflitti che possono essere evidenziati ricadono nelle seguenti categorie
  - **Conflitti di eterogeneità**: indicano discrepanze dovute all'utilizzo di formalismi con diverso potere espressivo negli schemi sorgenti
  - **Conflitti semantici**: si verificano quando due schemi sorgenti modellano la stessa porzione di mondo reale a un diverso livello di astrazione e dettaglio
  - **Conflitti sui nomi:** si verificano a causa delle differenze nelle terminologie utilizzate nei diversi schemi sorgenti
  - **Conflitti strutturali:** sono causati da scelte diverse nella modellazione di uno stesso concetto

Conflitti Semantici





Conflitti sui nomi



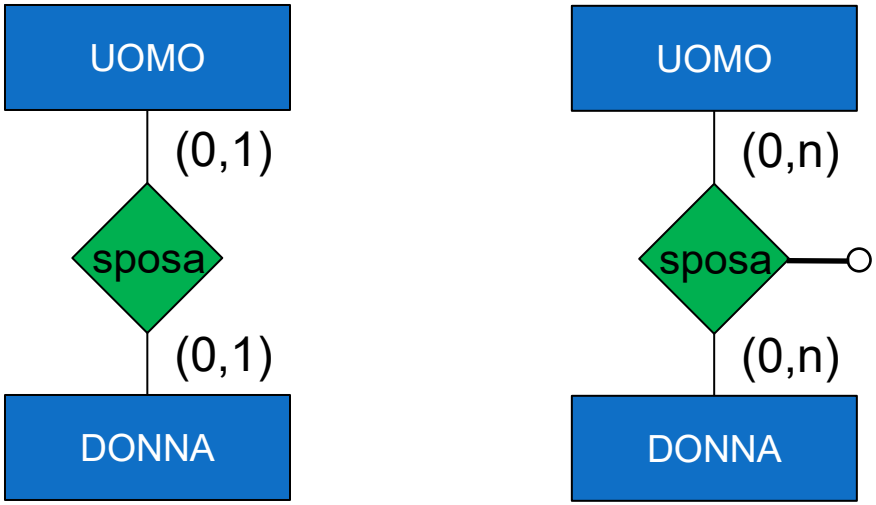
Conflitti strutturali

- In particolare, i conflitti strutturali possono essere:
  - **Conflitti di tipo**: si verificano quando uno stesso concetto è modellato utilizzando due costrutti diversi
  - **Conflitti di dipendenza**: si verificano quando due o più concetti sono correlati con dipendenze diverse in schemi diversi
  - **Conflitti di chiave**: si verificano quando per uno stesso concetto vengono utilizzati identificatori diversi in schemi diversi
  - **Conflitti di comportamento**: si verificano quando diverse politiche di cancellazione/modifica dei dati vengono adottate per uno stesso concetto in schemi diversi

Conflitti di tipo

- A livello di attributo (NUMERICO, ALFANUMERICO, ...)
  - Esempio: l'attributo"sesso"
    - Maschio/Femmina
    - M/F
    - 0/1
- A livello di entità
  - Diverse astrazioni dello stesso concetto del mondo reale producono insiemi di proprietà differenti (attributi)

Conflitti di dipendenza



- Esempio: Due diversi schemi per modellare il matrimonio
  - Lo schema a destra permette la storicizzazione delle informazioni

Integrazione degli schemi a livello concettuale

- Da un punto di vista logico l'integrazione di sorgenti di dati eterogenee consiste nel
  - Risoluzione dei conflitti
  - Produzione di un nuovo schema concettuale che esprime (per quanto possibile) la stessa semantica degli schemi che abbiamo voluto integrare
  - Produzione delle relazioni tra concetti degli schemi sorgenti e quello integrato (mapping)

Mapping

- Le query al sistema integrato vengono poste in termini di schema globale (o mediato)
  - specificano quali dati del database virtuale che ci interessano
  - Il problema è capire quali dati reali (nelle sorgenti di dati) corrispondono a quelli virtuali
- È necessario definire un mapping tra lo schema logico globale (mediato) ed ogni sorgente di dati (definizione di vista logica)
- Due approcci di base
  - GAV (Global As View)
  - LAV (Local As View)

Global As View

- Ad ogni concetto dello schema globale deve essere associata una vista il cui significato è definito in base ai concetti che risiedono sugli schemi sorgenti
  - Riduce l'estensibilità dello schema riconciliato poiché l'aggiunta di una nuova sorgente richiederà la modifica di tutti i concetti dello schema globale che la utilizzano
  - Facilita la modalità di definizione delle interrogazioni poiché per capire quali concetti degli schemi sorgenti sono coinvolti sarà sufficiente sostituire a ogni concetto dello schema globale la definizione della vista che lo definisce rispetto ai concetti sugli schemi locali (unfolding)

GAV: Un esempio

```
// DB1 Magazzino
ORDINI2011(chiaveO, chiaveC, data ordine, impiegato)
CLIENTE(chiaveC, nome, indirizzo, città, regione, stato)

// DB2 Amministrazione
CLIENTE(chiaveC, piva, nome, tel, fatturato)
FATTURE(chiaveF, data, chiaveC, importo, iva)
STORICOORDINI2010(chiaveO, chiaveC, data ordine, impiegato)
.............
CREATE VIEW CLIENTE AS
SELECT CL1.chiaveC, CL1.nome, CL1.indirizzo, CL1.città, CL1.regione,
   CL1.stato, CL2.tel, CL2.fatturato
FROM DB1.CLIENTE AS CL1, DB2.CLIENTE AS CL2
WHERE CL1.chiaveC = CL2.chiaveC;

CREATE VIEW ORDINI AS
SELECT * FROM DB1.ORDINI2011
UNION
SELECT * FROM DB2.STORICOORDINI2010;
```

Local As View

- Lo schema globale è espresso indipendentemente dalle sorgenti, i cui concetti saranno invece definiti come viste sullo schema globale
  - Richiede trasformazioni complesse (query rewriting) *per* capire quali elementi degli schemi sorgente devono essere presi in considerazione per ricreare il concetto espresso nello schema globale
  - Favorisce l'estensibilità dello schema riconciliato e la sua manutenzione
    - l'aggiunta di una nuova sorgente al sistema richiederebbe solo la definizione della vista sullo schema globale che non verrebbe quindi necessariamente modificato

LAV: Un esempio

```
// DB Globale
ORDINI(chiaveO, chiaveC, data ordine, impiegato)
CLIENTE(chiaveC, piva, nome, indirizzo, città, regione, stato, tel,
   fatturato)
.................
// DB1 Magazzino
CRETE VIEW CLIENTE AS
SELECT chiaveC, nome, indirizzo, città, regione, stato
FROM DB.CLIENTE;

CREATE VIEW ORDINI2011 AS
SELECT * FROM DB.ORDINI
WHERE data > '31/12/2010' and data < ''1/1/2012'';
```

- Nota: la definizione delle sorgenti locali è semplice ma come faccio a esprimere le interrogazioni dallo schema globale a quello locale?