

Università degli Studi di Milano - Bicocca

Big Data in Public and Social Services

GRAPH4DS

Beatrice Fumagalli - matricola 784549
Stefano Rola - matricola 790383

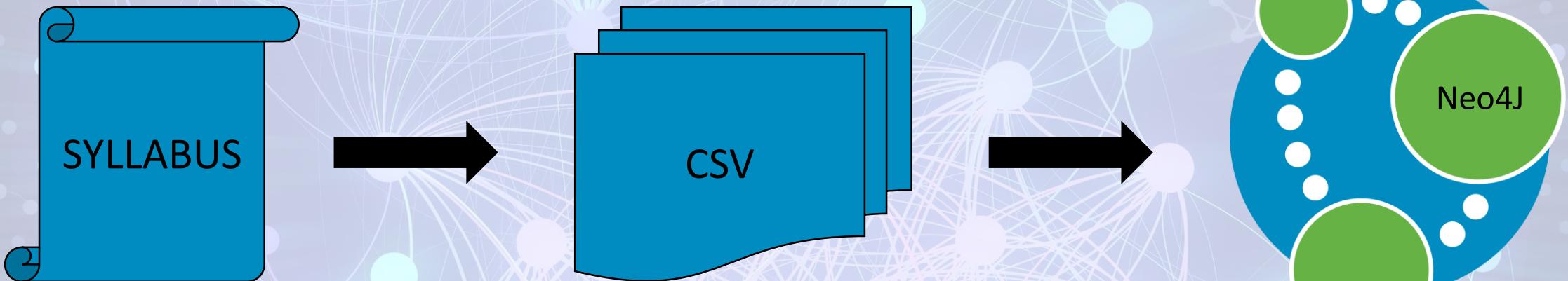
OBIETTIVO

GRAPH4DS: sistema che modella il corso di Laurea magistrale in Data Science dell'Università degli Studi di Milano Bicocca come *Graph Database* per l'esecuzione di *graph-based queries* e *social network analysis*.



Lo scopo è quello di fornire uno strumento utile alla didattica sia per i docenti che per gli studenti.

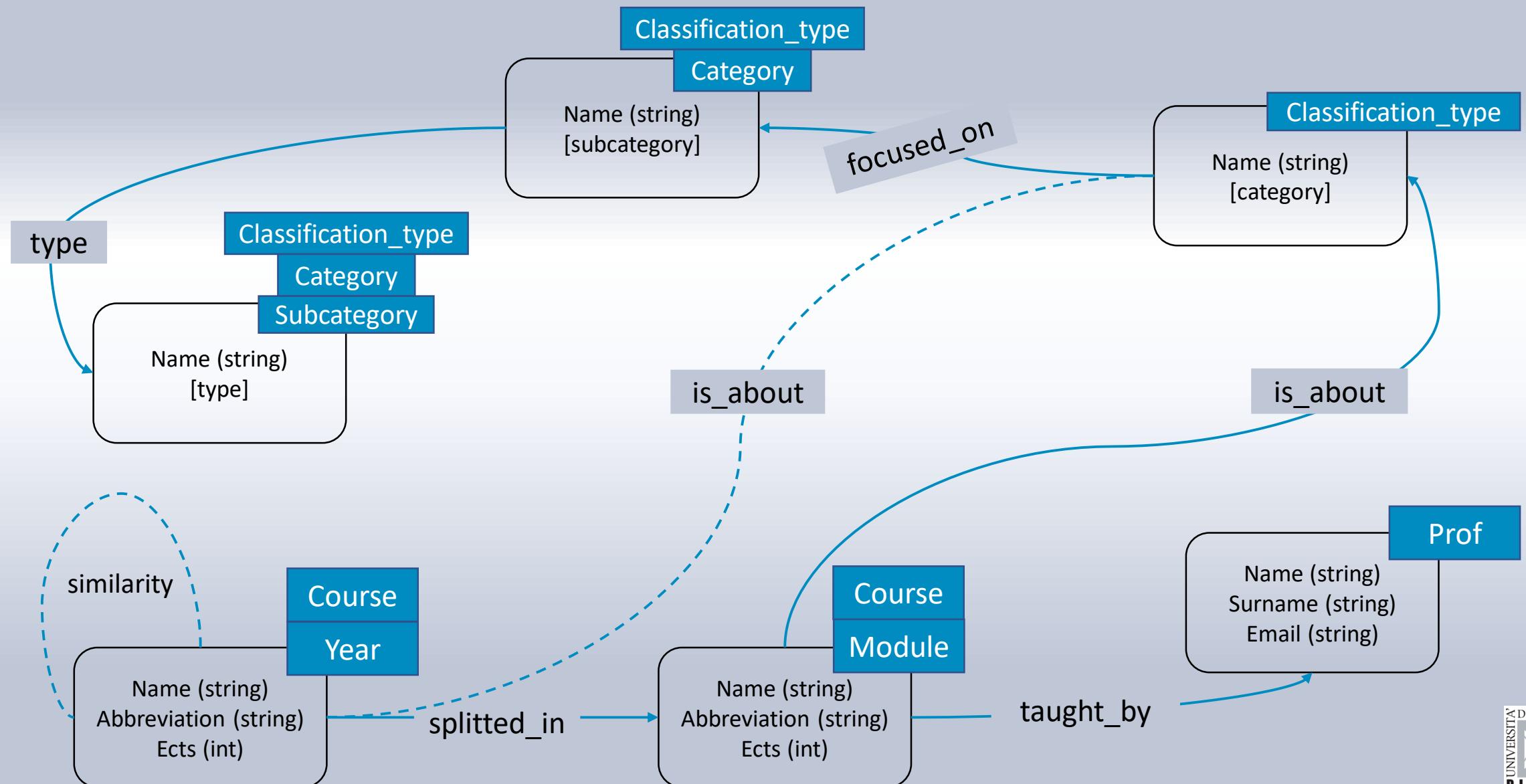
WORKFLOW



SYLLABUS

- Corsi del primo e secondo anno con rispettivi CFU ed eventuale suddivisione in moduli
- 10 tipi di classificazione
- Specifiche di ciascun tipo di classificazione
- Specifiche di ciascun corso/modulo in riferimento alle specifiche dei tipi di classificazione
- Cognome dei docenti di ciascun corso/modulo

DATA MODEL



composto da quattro blocchi elementari:

- **nodi** sono simili alle tuple nei database relazionali comuni. A differenza dei *database* relazionali, i nodi non hanno una struttura fissa ma possono variare di lunghezza contenendo diversi tipi di dati.
- **relazioni** possono esistere tra i nodi. Esse hanno sempre un nodo iniziale e uno finale.
- **proprietà** sono coppie chiave-valore che possono essere incluse in qualsiasi nodo o relazione.
- **etichette** associano un nome comune a un insieme di nodi o relazioni con l'obiettivo di indicizzazione e recupero rapidi. A differenza delle proprietà, le etichette non hanno valori. Un nodo o relazione può avere più di un'etichetta alla volta.

GRAPH4DS: 306 nodi e 1781 relazioni

Course	26
Module	22
Prof	40
Category	121
Subcategory	90
Type	7

GRAPH4DS



SIMILARITY TRA CORSI

Sia S_c l'insieme delle specifiche trattate dal corso c , l'indice di Jaccard tra due corsi c_i e c_j è definito come segue:

$$J_{c_i, c_j} = \frac{|S_{c_i} \cap S_{c_j}|}{|S_{c_i} \cup S_{c_j}|}$$

From	To	Intersection	Jaccard Similarity
"Foundations of probability and statistics"	"Statistical modeling"	63	0.9844
"Statistical modeling"	"High dimensional data analysis"	63	0.8514
"Foundations of probability and statistics"	"High dimensional data analysis"	62	0.8378
"Data science lab"	"Streaming data management"	37	0.8222
"Big data in health care"	"Big data in public health"	23	0.697
"Foundations of computer science"	"Data management"	52	0.65
"Foundations of computer science"	"Machine Learning"	54	0.6207
"Data semantics"	"Data infrastructures"	73	0.6033
"Data management"	"Machine Learning"	51	0.593
"Foundations of computer science"	"Data infrastructures"	58	0.5743

The background of the slide features a complex, abstract network graph. It consists of numerous small, semi-transparent white and light blue dots representing nodes, connected by a dense web of thin, light-colored lines representing edges. This creates a sense of a large, interconnected system or data structure.

Grazie per l'attenzione
