

Intelligenza artificiale simbolica e gestione dei dati

Maurizio Lenzerini

Dipartimento di Ingegneria Informatica, Automatica e Gestionale Antonio Ruberti
Sapienza Università di Roma, Italy

HUMAN TOUCH 2.0: OLTRE L'IA NELL'UNIVERSITA' DEL DOMANI

Siena, 7 – 9 novembre 2024

Dati e intelligenza artificiale (AI): un legame imprescindibile

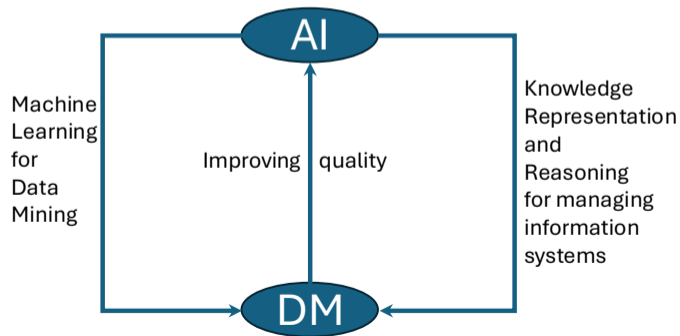
- AI per Data Analytics (Data mining)
- Data Management for AI (Data-centric AI)
 - monitorare e assicurare la qualità dei dati
 - modellare opportunamente dati di “training” e di “test”
 - ideare ed applicare “stress tests” basati sulla semantica dei dati
- AI per Data management

... legame caratterizzato da diverse sottodiscipline dell'AI

- AI (Machine Learning) per Data Analytics (Data mining)
- Data Management per AI (Machine Learning): Data-centric AI
 - monitorare e assicurare la qualità dei dati
 - modellare opportunamente dati di “training” e di “test”
 - ideare ed applicare “stress tests” basati sulla semantica dei dati
- AI (Semantic Domain Modeling e Knowledge Representation) per Data management

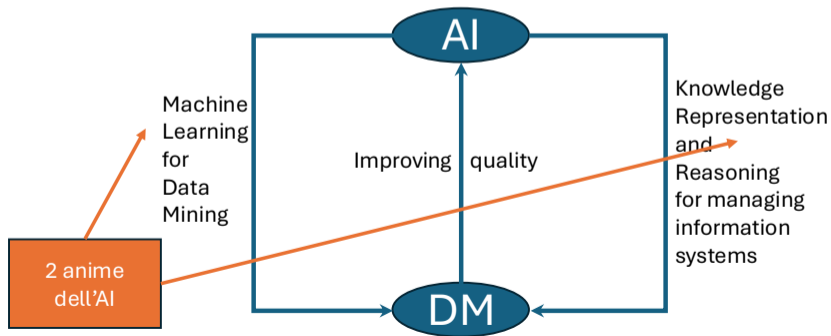
... legame caratterizzato da diverse sottodiscipline dell'AI

- AI (Machine Learning) per Data Analytics (Data mining)
- Data Management per AI (Machine Learning): Data-centric AI
 - monitorare e assicurare la qualità dei dati
 - modellare opportunamente dati di “training” e di “test”
 - ideare ed applicare “stress tests” basati sulla semantica dei dati
- AI (Semantic Domain Modeling e Knowledge Representation) per Data management



... legame caratterizzato da diverse sottodiscipline dell'AI

- AI (Machine Learning) per Data Analytics (Data mining)
- Data Management per AI (Machine Learning): Data-centric AI
 - monitorare e assicurare la qualità dei dati
 - modellare opportunamente dati di "training" e di "test"
 - ideare ed applicare "stress tests" basati sulla semantica dei dati
- AI (Semantic Domain Modeling e Knowledge Representation) per Data management



Due anime dell'AI

- **data-driven (connectionist AI)** – decisioni che mostrano caratteri di intelligenza vengono assunte basandosi su una struttura di nodi interconnessi e la consapevolezza del mondo emerge dalla percezione e dall'esperienza, codificata nei dati raccolti ed estratta mediante il machine learning
- **representation-driven (symbolic AI)** – il mondo è concettualizzato in termini di una struttura simbolica espressa mediante un determinato formalismo, con associata **semantica esplicita**, che il sistema usa per esplicitare la **conoscenza** che ha del mondo stesso e per **ragionare** su di esso mediante la logica (operazioni su simboli)

Due anime dell'AI

- **data-driven (connectionist AI)** – decisioni che mostrano caratteri di intelligenza vengono assunte basandosi su una struttura di nodi interconnessi e la consapevolezza del mondo emerge dalla percezione e dall'esperienza, codificata nei dati raccolti ed estratta mediante il machine learning
- **representation-driven (symbolic AI)** – il mondo è concettualizzato in termini di una struttura simbolica espressa mediante un determinato formalismo, con associata **semantica esplicita**, che il sistema usa per esplicitare la **conoscenza** che ha del mondo stesso e per **ragionare** su di esso mediante la logica (operazioni su simboli)

Knowledge Representation and Reasoning is the area of Artificial Intelligence (AI) concerned with how knowledge about a domain can be explicitly represented symbolically and manipulated in an automated way by reasoning procedures. More informally, it is the part of AI that is concerned with (slow) thinking (based on semantics), and how (slow) thinking contributes to intelligent behavior. [Brachman & Levesque, 2004]

Da dove viene l'AI

- [Aristotele \(384 a.c.\)](#): ontologia e logica
- [Leibniz \(1646\)](#): calculus ratiocinator
- [Boole \(1815\)](#): algebra per la logica
- [Hilbert \(1862\)](#): “Mathematics must be formulated on a solid and complete logical foundation”
- [Russell \(1872\)](#): “all Mathematics is Symbolic Logic”
- [Godel \(1906\)](#): Sistema deduttivo “corretto e completo” per la logica del primo ordine nella sua tesi di dottorato
- [Frege \(1848\)](#) riceve una lettera da Russell (1903)
- Teorema di incompletezza di Godel (1930)
- Teorema di [Church \(1903\)](#) - [Turing \(1912\)](#) Non esiste alcun procedimento meccanico che in tempo finito sia in grado di stabilire se una formula sia un teorema oppure no (1936)
- Turing machine (1936) + Turing test (Computing machine and Intelligence, 1948)
- [Pitts](#) and [McCulloch](#) studied networks of idealized artificial neurons and showed how they might perform simple logical functions (1943)
- Conferenza di Dartmouth (1956), con [McCarthy](#), [Minsky](#), [Rochester](#), [Shannon](#)

Da dove viene la Knowledge Representation – The early days ...



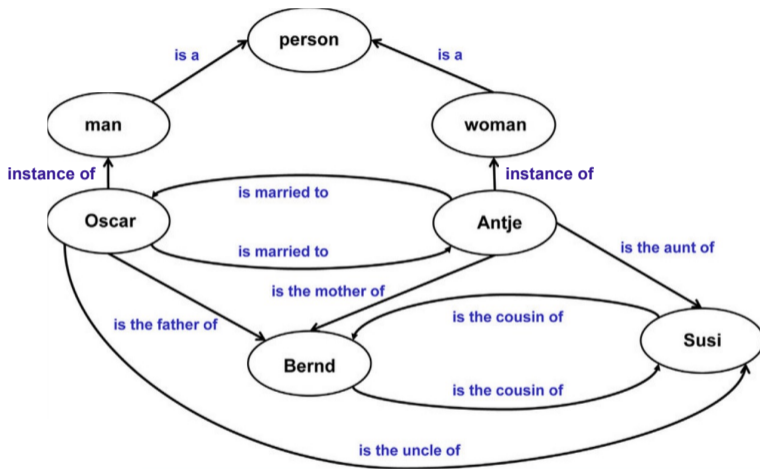
Da dove viene la Knowledge Representation (e la sua relazione con la Data Management)

- '50s: "Semantic Nets" (R. H. Richens 1956), an "interlingua" for machine translation
- '60s: Network data model (IDS) and Hierarchical data model (IMS) no modeling or design theory
- '60s: Semantic Networks (R. Quillian, 1968)
- '70s: Relational data model (E. Codd, 1970) normalization theory
- '70s: Frame-based systems (M. Minsky, 1974)
- '70s: Database modeling (Roussolopoulos & J. Mylopoulos, 1975)
- '70s: "What's in a link" (W.A. Woods, 1975)
- '70s: Entity-Relationship model (P. Chen, 1976)
- '70s: "Conceptual Graphs" (J. Sowa, 1976)
- '70s: "Aggregation and Generalization" (Smith & Smith, 1977)
- '70s: "KL-ONE" (R. Brachman et al, 1977) automated reasoning
- '70s: "Data and reality" (W. Kent, 1978)

Da dove viene la Knowledge Representation (e la sua relazione con la Data Management)

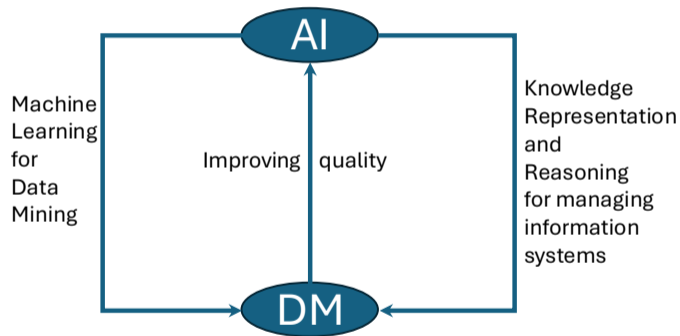
- '80s: Design methods based on normalization (C. Date, 1981)
- '80s: “Towards a Logical Reconstruction of Relational Database Theory” (R. Reiter 1982)
- '80s: “Conceptual modeling” (Broadie et al 1984)
- '80s: “Expressiveness and tractability in knowledge representation and reasoning” (H. Levesque & Brachman 1987)
- '90s: Unified Modeling Language (Fowler & Scott 1997)
- '90s: Description Logics
- '90s: Semistructured data models and NoSQL
- '00s: OWL (Ontology Web Language)
- '00s: Graph databases
- '00s: Tools for reasoning on DL KBs
- '10s: Google announces its knowledge graph (2012)
- '10s -'20: Ontology-based Data Management

Semantic networks and knowledge graphs

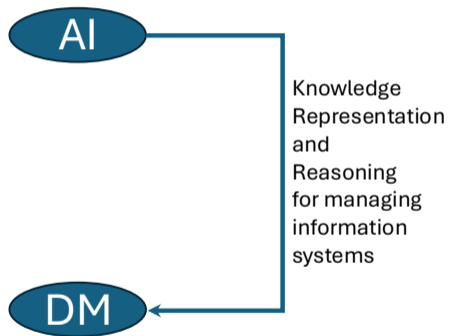


Knowledge Graph (2022)

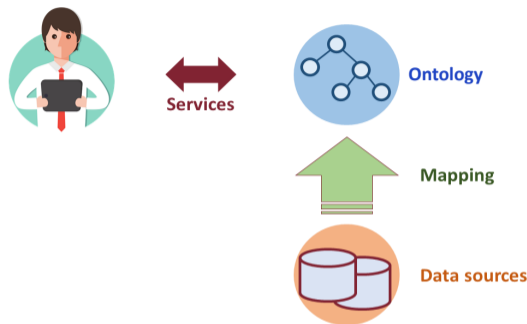
Il resto dell'intervento: Knowledge Representation for managing Information Systems



Il resto dell'intervento: Knowledge Representation for managing Information Systems



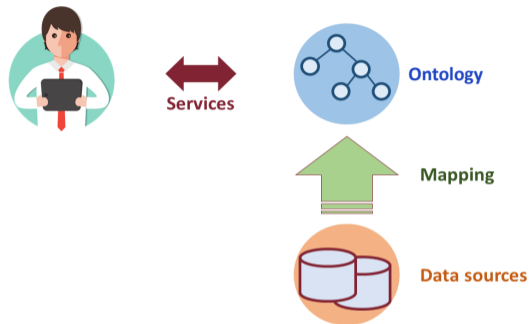
Ontology-based data management: la semantica nella gestione dei dati



Based on three main components [Poggi & al. 2008, InfSys]:

- **Data sources**, set of heterogeneous data repositories managed by the information system
- **Ontology**, declarative specification of the knowledge about the domain of interest (and even the whole information system) enabling formal reasoning over the system
- **Mappings**, used to semantically link data at the sources to the ontology

Ontology-based data management: la semantica nella gestione dei dati



- **Ontology**, declarative specification of the knowledge about the domain of interest (and even the whole information system) enabling formal reasoning over the system

Abbiamo davvero bisogno di una rappresentazione del dominio?

Fragment of a relational table in a Bank Information system:

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Abbiamo davvero bisogno di una rappresentazione del dominio?

Negative value denotes a holding

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Abbiamo davvero bisogno di una rappresentazione del dominio?

S means that the customer is the leader of the group it belongs to

S means that the customer is the head of the group it belongs to

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Abbiamo davvero bisogno di una rappresentazione del dominio?

*N means that the
FATTURATO field is not valid*

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Rimaniamo al livello delle sorgenti?

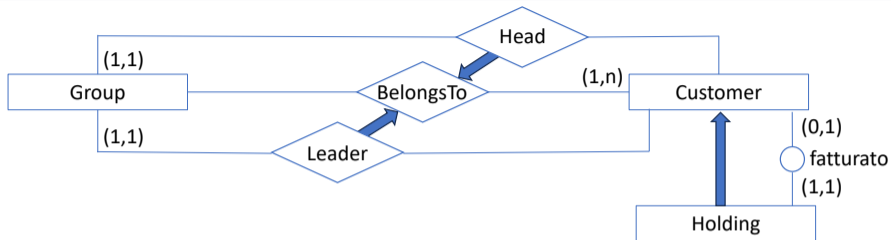
Davvero vogliamo usare (accedere, interrogare, verificare la qualità, interoperare) i dati a questo livello?



CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Sorgente

... oppure costruiamo il modello del dominio (ontologia)?

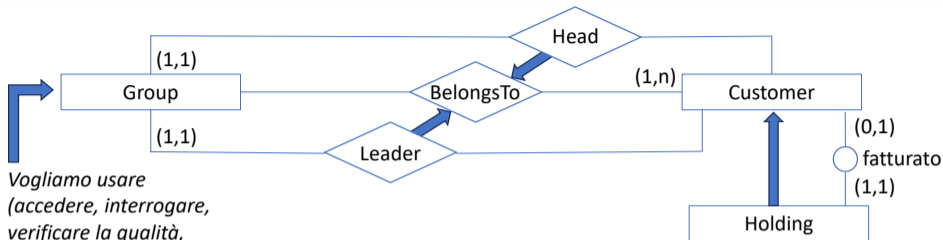


Ontologia

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Sorgente

Una volta che abbiamo l'ontologia ...



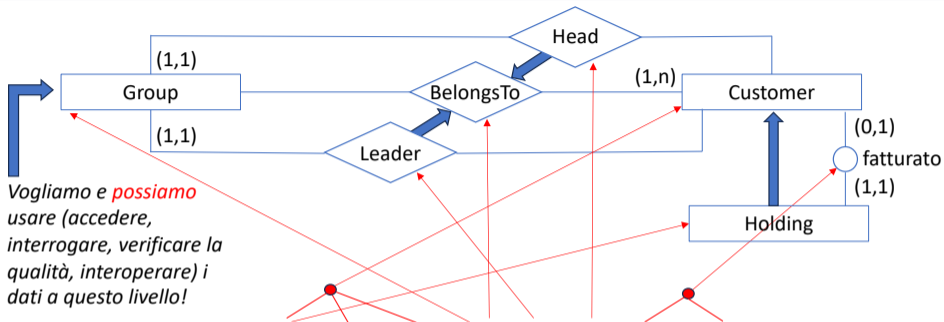
Ontologia

Vogliamo usare
(accedere, interrogare,
verificare la qualità,
interoperare) i dati a
questo livello!

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N	
130976	7-mag-2001	9-lug-2003	75680					

Sorgente

Ontology-based data management reso possibile dal mapping



Vogliamo e **possiamo** usare (accedere, interrogare, verificare la qualità, interoperare) i dati a questo livello!

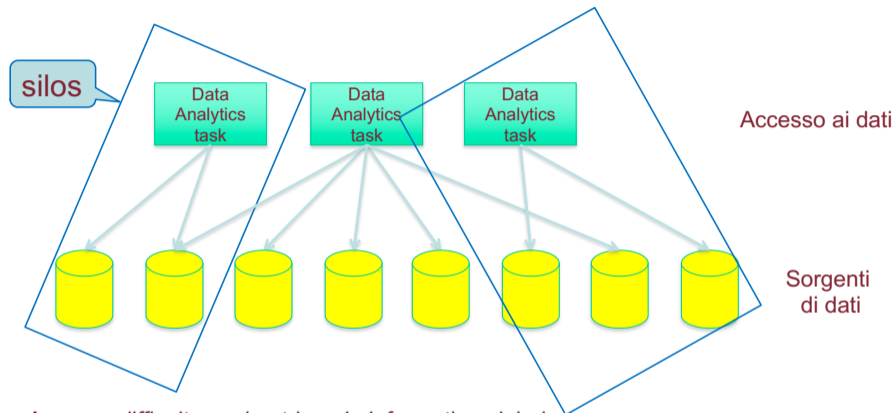
Ontologia

Mapping

Sorgente

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S
-472900	10-mag-2001	1-gen-9999	62010	S	N	0 00	N
130976	7-mag-2001	9-lug-2003	75680				

Da una a più sorgenti: superare l'organizzazione a silos

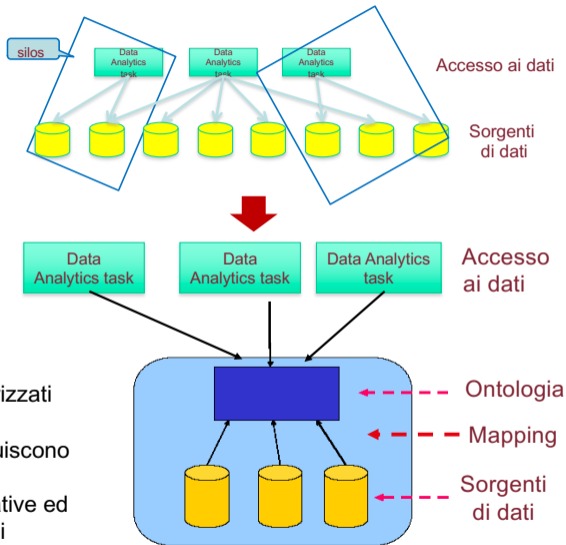


- Accesso difficoltoso al patrimonio informativo globale
- Ridondanze ed incoerenze nei dati si moltiplicano a livello di task
- Dipendenza dei risultati dal modo in cui i task considerano i dati
- Difficoltà nell'interoperabilità (difficile confronto e armonizzazione dei risultati di diversi task)

Da una a più sorgenti: superare l'organizzazione a silos

OBDM
paradigma
knowledge-driven

- L'ontologia è una **base di conoscenza** che rappresenta concettualmente il dominio di interesse, e non i dati memorizzati nelle sorgenti
- I servizi di cui gli utenti usufruiscono sono espressi sull'ontologia
- Il mapping tra risorse informative ed ontologia riconcilia i due livelli



Ontology-based data management: la semantica nella gestione dei dati

Il fulcro è la **conoscenza condivisa del patrimonio informativo** dell'organizzazione, espressa in modo dichiarativo nell'ontologia (**modello del dominio**), vero punto di discontinuità del sistema.

- Verso l'alto

- “Data and Metadata Querying” mediato da una visione concettuale
- Data analytics svolto su una visione unitaria/condivisa/riconciliata/coerente
- Cooperazione ed interoperabilità basate sulla semantica
- “Explanation and fairness”
- “Semantic data publishing” ⇒ grafo della conoscenza

- Verso il basso

- Gestione di sorgenti autonome, anche esterne, ed eterogenee (anche “data lake”) e loro relazioni
- Semantic data integration
- Capacità di introspezione: “provenance”, profilazione, verifiche di qualità
- Supporto alla re-ingegnerizzazione

NOTA: indipendente dalla realizzazione (data warehousing, ETL, ELT, data federation, virtual data integration, ...)

Benefici

- **Consapevolezza e controllo del patrimonio informativo dell'organizzazione**
- Modelli di conoscenza su domini stanno proliferando in diversi campi (upper ontologies, biologia, finanza, pubblica amministrazione, health, assicurazioni, trasporto, sport, specifici ambiti scientifici, geografia ...)
- L'integrazione di specifiche ontologie con tali modelli è possibile

Alcune frasi prese dalla presentazione del convegno:

- Nel nuovo panorama si delinea un ecosistema in cui la tecnologia deve agire come mezzo per migliorare le condizioni umane [e la condizione delle istituzioni pubbliche], **anzichè imporre un suo dominio autonomo**.
- In tale prospettiva la persona è chiamata a programmare la tecnologia, secondo i propri valori fondanti e le proprie necessità, **senza doverla subire**.
- La centralità della persona rappresenta un principio basilare del mondo universitario anche nell'epoca della transizione digitale. **L'essere umano mantiene la propria centralità nel pensiero consapevole e costruttivo**, offrendo un contributo insostituibile all'evoluzione sociale ed economica.

Criticità

- complessità della modellazione ⇒ metodologie,
- complessità nella gestione ⇒ tools (OBDA Systems – obdasystems.com, Ontopic – ontopic.ai, Stardog – stardog.com, ...)
- costi ⇒ costi/benefici
- rigidità ⇒ flessibilità, evoluzione e manutenzione dell'ontologia
- bisogno di adattabilità a diverse architetture ⇒ flessibilità, evoluzione e manutenzione delle sorgenti e dei mapping

Alcune criticità sia dei LLMs sia dei KGs possono essere affrontate dalla **integrazione neuro-simbolica in AI**

- **LLMs per supportare ontologie**
 - LLMs possono aiutare ad analizzare grandi quantità di documenti ed **estrarre conoscenza intensionale** da esprimere poi come ontologia
 - LLMs possono aiutare ad analizzare grandi quantità di documenti ed **estrarre conoscenza estensionale** per popolare ontologie
 - Large Language Models possono contribuire alla **modellazione di ontologie e nella loro evoluzione**
 - Large Language Models possono supportare la **formulazione di queries in linguaggio naturale**

Alcune criticità sia dei LLMs sia dei KGs possono essere affrontate dalla **integrazione neuro-simbolica in AI**

- **Ontologie per supportare LLMs**

La caratteristica dominante è che il modello del dominio è sotto il controllo del “progettista” del sistema di AI → **modularità, estendibilità, affidabilità, spiegabilità, manutenibilità**

- **Allucinazioni**: le ontologie forniscono una sorta di “ground truth”
- **Limitate capacità di ragionamento**: le ontologie sono equipaggiate con metodi per il ragionamento automatico
- **Carente conoscenza su domini specifici**: le ontologie supportano basi di conoscenza su specifici domini
- **Obsolescenza della conoscenza**: le ontologie necessitano di manutenzione, ma non di fasi di addestramento costose
- **“Bias”, privatezza e risposte offensive**: le ontologie possono controllare meglio la deriva di bias o accessi malevoli

Alcune criticità sono insiti nel paradigma “representation-driven”

Esempio: da un bando di concorso pubblico...

- (Art. 4) Il concorso è articolato in
 - una prova scritta
 - una prova orale
 - un colloquio in lingua inglese
- (Art. 7) Il punteggio totale massimo è di 120 punti. La prova scritta è valutata fino ad un massimo di 60 punti
- (Art. 7) La prova orale consiste in un colloquio sulle materie tecniche e in un colloquio in lingua inglese
- (Art. 7) La prova orale è valutata fino ad un massimo di 60 punti ed il colloquio in lingua inglese fino ad un massimo di 5 punti.

ChatGTP: la prova orale comprende anche il colloquio in inglese

Alcune criticità sono insiti nel paradigma “representation-driven”

Esempio: da un articolo in prima pagina di uno dei maggiori quotidiani nazionali...

- C'è qualcuno che sostiene che il rigore nei conti pubblici è in contrasto con lo sviluppo dell'economia, ma anni di storia italiana con spesa facile e stagnazione economica dimostrano esattamente il contrario.
- **ChatGTP**: È possibile confutare l'affermazione che il rigore nei conti pubblici è in contrasto con lo sviluppo dell'economia citando il fatto che ci sono stati anni della storia italiana in cui i conti pubblici non sono stati gestiti con rigore e, tuttavia, l'economia italiana ha registrato una stagnazione.