system evolution

Intelligenza Artificiale al servizio del Data Governance

Come determinare la rilevanza del dato?

Business Line IMA

Lina Ferraiolo, Principal Consultant System Evolution – Kirey Group
Claudio Bottari, Machine Learning Engineer System Evolution – Kirey Group

Milano, febbraio 2019

kirey group

Agenda

- **La Data Governance**
  - Il dato è ricchezza
  - L'importanza del dato
  - Il Data Owner
- **L'intelligenza Artificiale**
  - Introduzione al Machine Learning
  - I dati nel contesto bancario
  - Word Embedding Tecnico
  - Applicazioni di ML in ambito di Data Governance

Data Governance

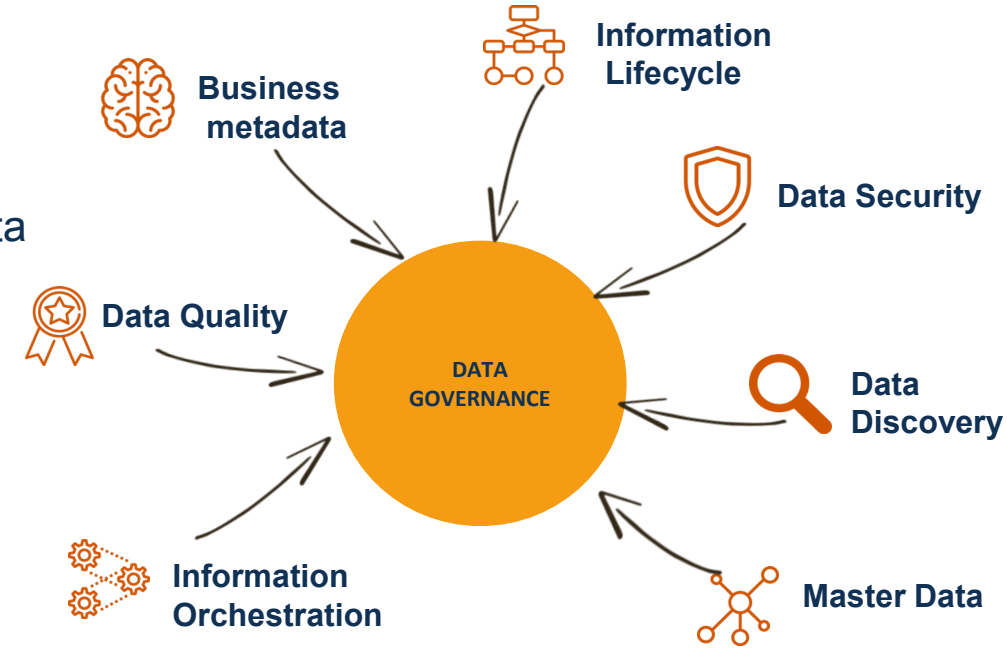
25 febbraio 2019

kirey group

Data Governance: Il dato è ricchezza

Caratteristiche dell'approccio all'attuazione di un programma di Data Governance

- Assessment e valorizzazione situazioni esistenti
- Principio di proporzionalità/Approccio Incrementale
- Sostenibilità



Data Governance: Importanza del dato

La scomposizione in ambiti del patrimonio informativo

La determinazione del perimetro di ciascun ambito

L'attribuzione di un coefficiente numerico, limitato, rappresentativo dell'importanza di un dato in un ambito

Il *Data Owner* è il "proprietario" del dato, colui che ne detiene la responsabilità, avendone fornito le regole funzionali di produzione.

È la figura autorevole che sa associare il «valore» corretto di importanza.

È la figura alla quale si richiede di validare i risultato dell'elaborazione del ML.

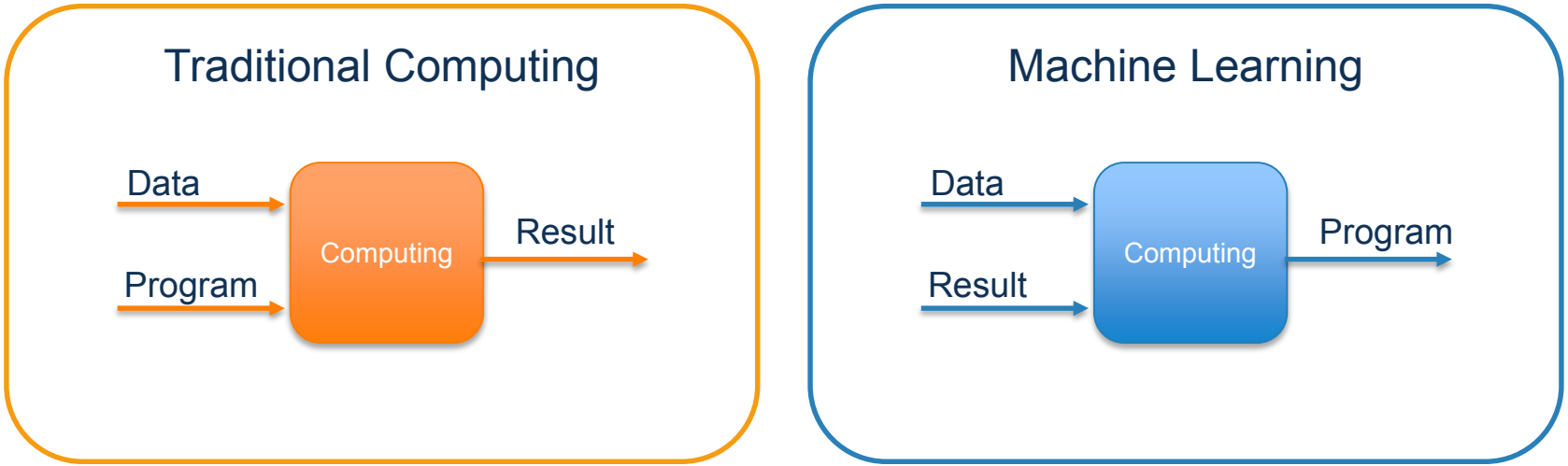
Introduzione al Machine Learning

25 febbraio 2019

kirey group

Traditional Computing vs Machine Learning

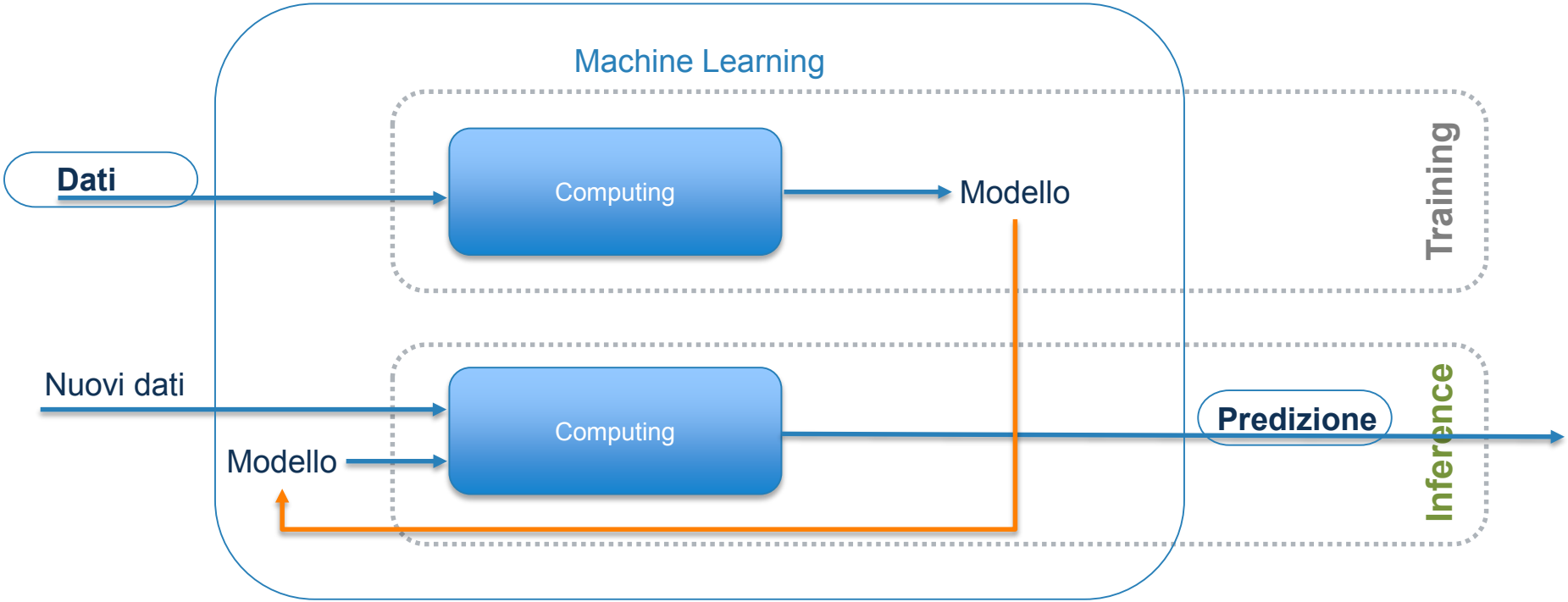
Tipicamente la dicotomia tra questi due paradigmi viene così stigmatizzata:



kirey group

Traditional Computing vs Machine Learning

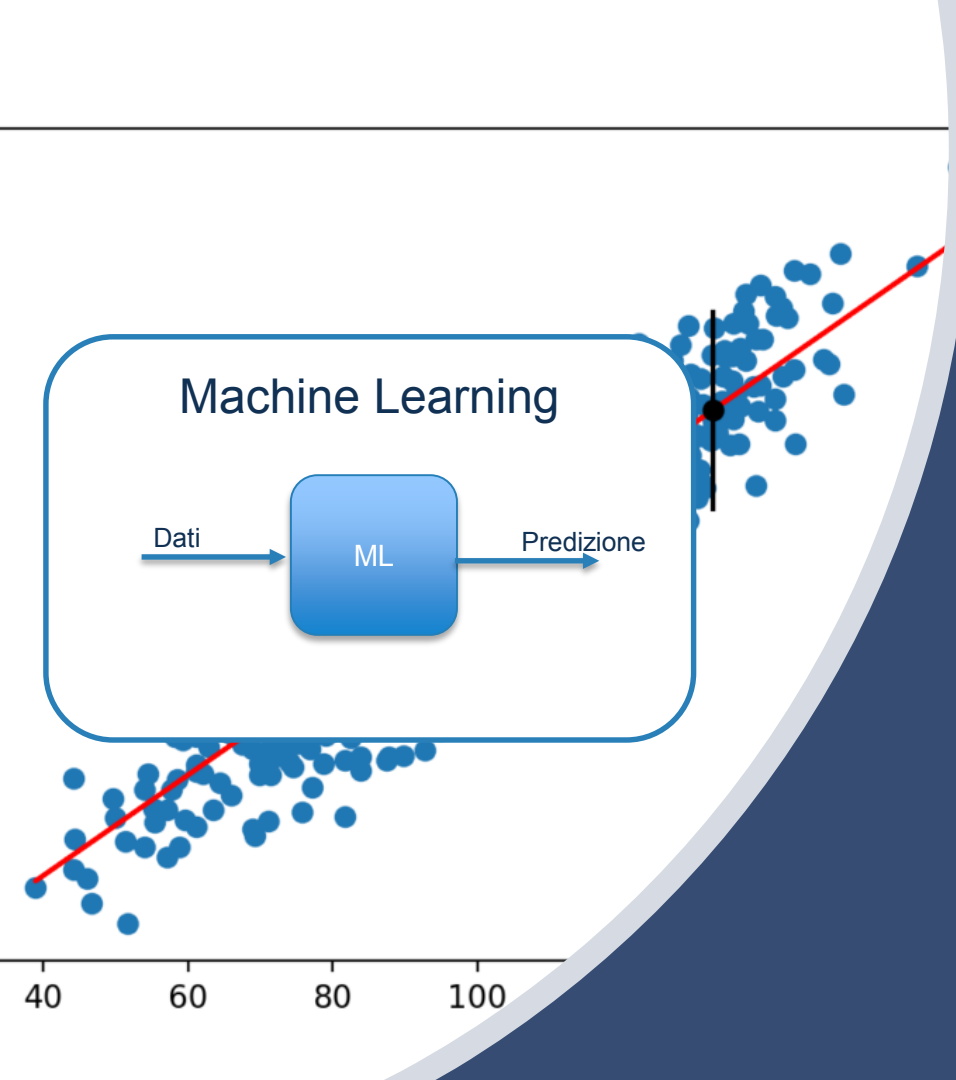
In realtà il funzionamento del processo di machine learning è leggermente più complesso:



kirey group

Quindi il Machine Learning...

- ✓ È il processo tramite il quale è possibile ottenere una **predizione** utilizzando un modello (inference)
- ✓ Il modello di machine learning è creato tramite l'**addestramento** (training)
- ✓ L'input e l'output di questo processo è rappresentato unicamente da **dati**



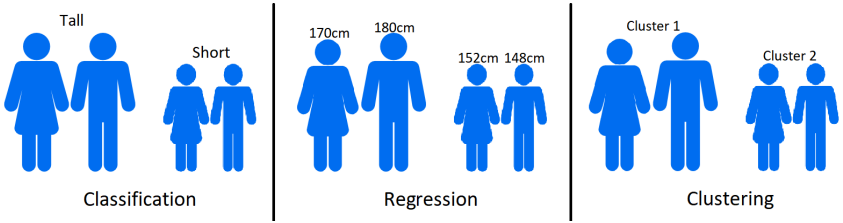
Cos'è una predizione?

Una predizione è la migliore approssimazione per un determinato problema… o più semplicemente è un informazione dedotta dai dati che ho a disposizione

kirey group

Vantaggi del Machine Learning

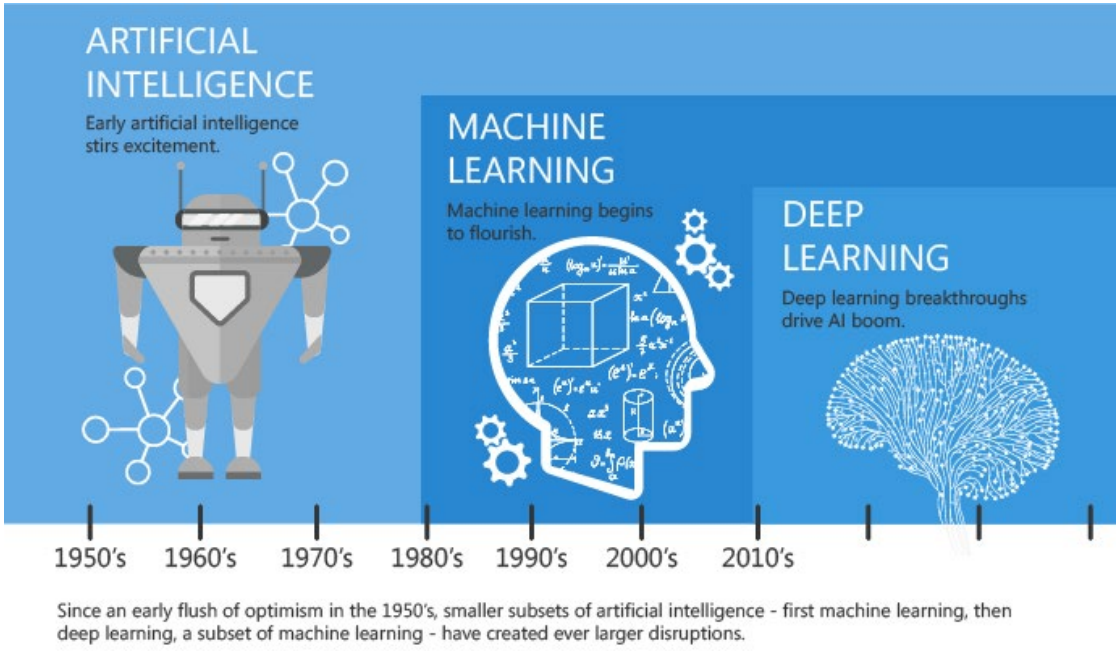
- il machine learning **non sostituisce** la programmazione tradizionale ma è uno strumento che permette di raggiungere risultati altrimenti preclusi
- il machine learning è in grado di **imparare** autonomamente tramite l'osservazione dei dati in fase di **training**
- il machine learning è capace di produrre **predizioni** anche quando non sono note le logiche che determinano la definizione del risultato



Tipi di predizione

- una **qualità** *(classification)*
  predico quale "gruppo“
- una **quantita** (regressione)
  predico un numero
- **raggruppamento** *(clustering)*
  predico i raggruppamenti

kirey group



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

Qual è la differenza tra machine learning e intelligenza artificiale?

- il **machine learning** indica il processo che abbiamo descritto
- l'**Intelligenza Artificiale** indica sia una branca della matematica che si è prefissa di creare modelli ispirati alla neurobiologia e che ha portato alla creazione dei modelli di rete neurali
- PS: il **Deep Learning** è l'ultima evoluzione di questi paradigmi

Intuzioni

- i dati plasmano il modello tramite il training
- la bontà del modello dipenda da molti fattori:
  - ✓ architettura
  - ✓ dati ingresso
- tutto parte dai **dati**...

I dati nel contesto bancario

25 febbraio 2019

kirey group

Cos’è un dato?

Tutto è un dato, ma possiamo identificare diverse caratteristiche che portano a diversi tipi di sfide al momento di doverli processare:

- strutturati/non strutturati
- quantitativi/qualitativi

I dati in contesto bancario

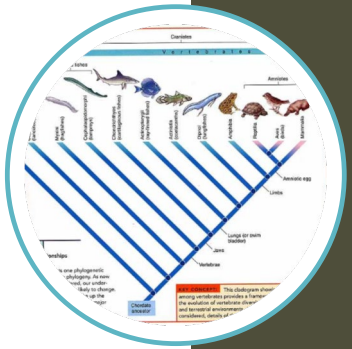
I dati con i quali avremo a che fare in un contesto bancario si possono identificare in queste tipologie:

- Dati **strutturati**: tabelle database
- Dati **non strutturati**: documenti

La sfida dei dati qualitativi

- i computer sono in grado di processare solo numeri (dati **quantitativi**) quindi è necessario tradurre la "classi" (dati **qualitativi**) in numeri

- fortunatamente, ogni dato qualitativo può essere ricondotto a un dato quantitativo tramite definizione di una codifica (**encoding**)

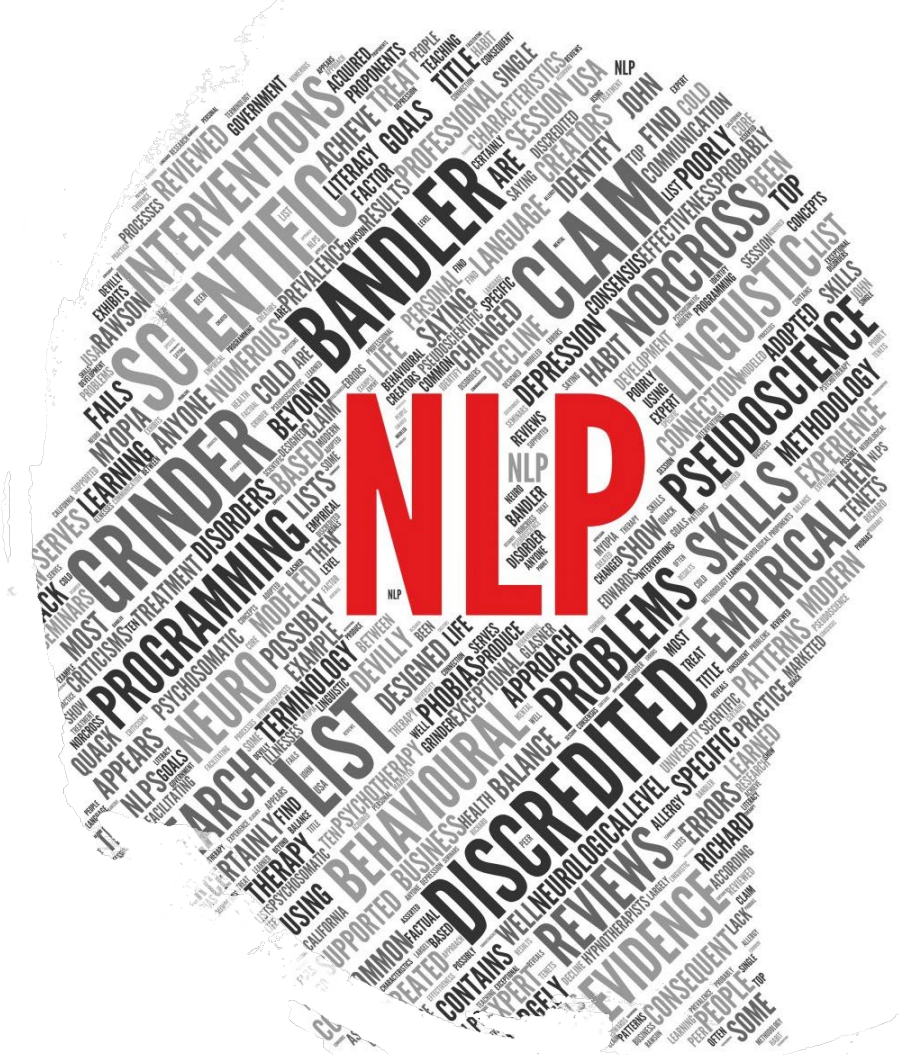


Natural Language Processing

Occorre avvalersi della branca del machine learning che si occupa di processamento del linguaggio naturale (**NLP**) in quanto i dati a nostre disposizione sono le parole.

Le parole sono dati:

- **qualitativi**, quindi sarà necessaria una fase di **ENCODING**
- **non strutturati**, quindi sarà necessario reti adatte ai dati non strutturati come le **CNN**



Word Embedding Settoriale

25 febbraio 2019

kirey group

Perché è necessario

Le reti neurali necessitano di **numeri** ed è quindi necessario che le parole vengano *"tradotte"* per essere elaborate.

Quando il numero di parole da interpretare aumenta occorre avere delle tecniche evolute per tradurle efficacemente.

Nell'esempio a lato, che rappresenta la tecnica più grossolana di encoding, queste parole saranno «equidistanti» le une dalle altre e la loro codifica non aiuterà in alcun modo l'interpretazione di tali testi.

Uomo, donna, ragazzo, ragazza, principe, principessa, regina, re, monarca…

One-hot encoding

	1	2	3	4	5	6	7	8	9
uomo	1	0	0	0	0	0	0	0	0
donna	0	1	0	0	0	0	0	0	0
ragazzo	0	0	1	0	0	0	0	0	0
ragazza	0	0	0	1	0	0	0	0	0
principe	0	0	0	0	1	0	0	0	0
principessa	0	0	0	0	0	1	0	0	0
regina	0	0	0	0	0	0	1	0	0
re	0	0	0	0	0	0	0	1	0
monarca	0	0	0	0	0	0	0	0	1

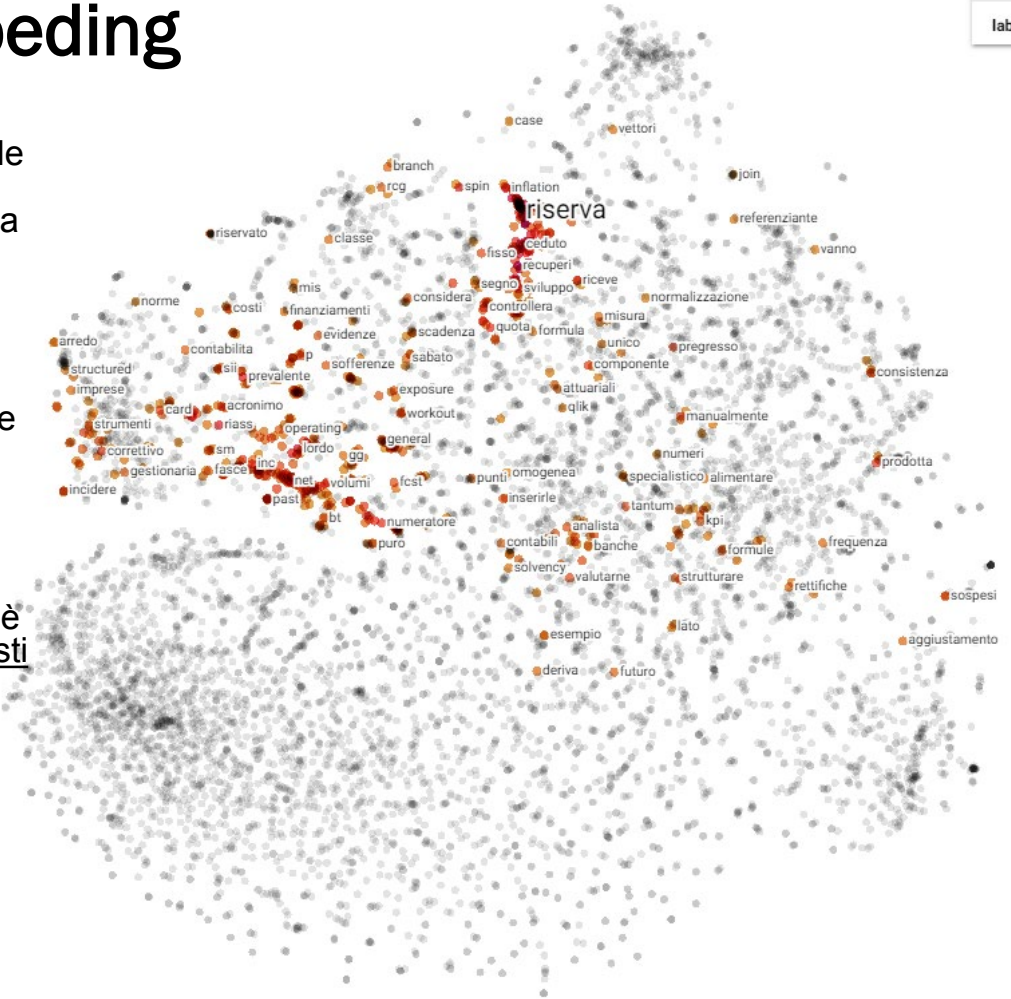
Il moderno Word Embeding

Il Word Embedding è la modalità con la quale le parole vengono posizione in un campo multidimensionali, traducendo ciascua parola in un matrice di numeri.

Dal 2014 in poi, tramite modelli **predittivi** mutuati dal machine learning (word2vec, FastText, BERT ecc), siamo in posizionare le parole in uno spazio multidimensionale in modo coerente col significato della parola stessa.

L'intuizione che ispira questi tipi di algoritmi è che parole simili vengono utilizzate in contesti simili.

Quindi è il **rapporto tra la singola parola e il contesto** nella quale viene utilizzata che ne determina il **significato**.



Embedding tecnico: i vantaggi

Sono presenti tramite l'open source una serie di modelli word Embedding pre-addestrati per tutte le lingue, italiano compreso.

Tali Embeddings sono sempre di carattere generalista, tipicamente addestrati "leggendo" Wikipedia o Reddit. Di conseguenza i significati codificati in tali Embeddings é di carattere generalista e le distanze tra le parole rispecchiano l'approssimazione di tutti i significati attribuiti dall' "uomo comune".

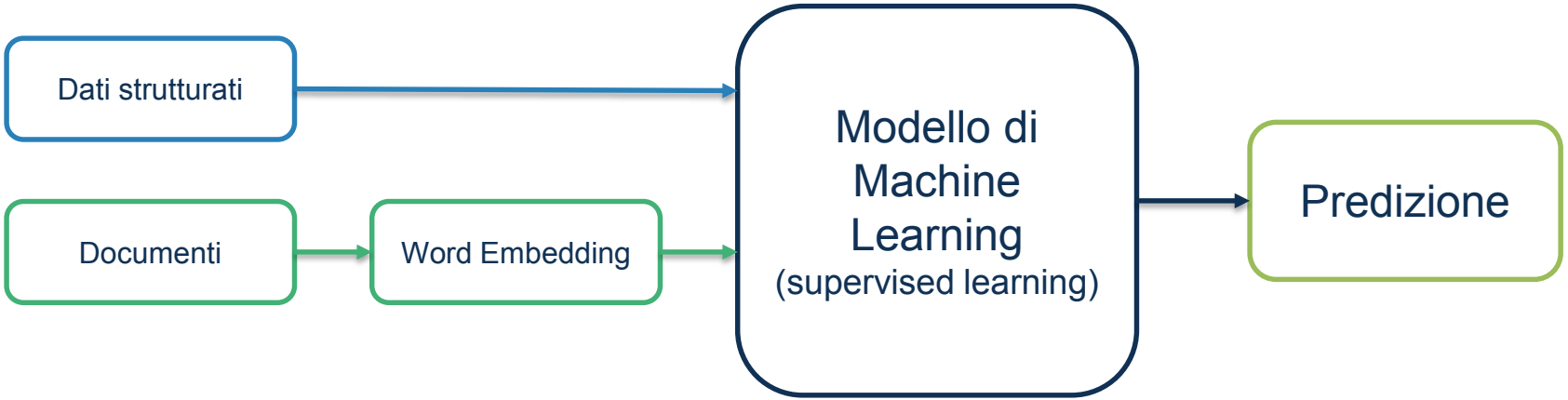
Al contrario un Embedding generato tramite l'utilizzo di il corpus specifico di un particolare settore (assicurativo, bancario, politico ecc.) **rispecchierà il significato specifico e univoco del settore di appartenenza.**

Applicazioni di ML in ambito di Data Governance

25 febbraio 2019

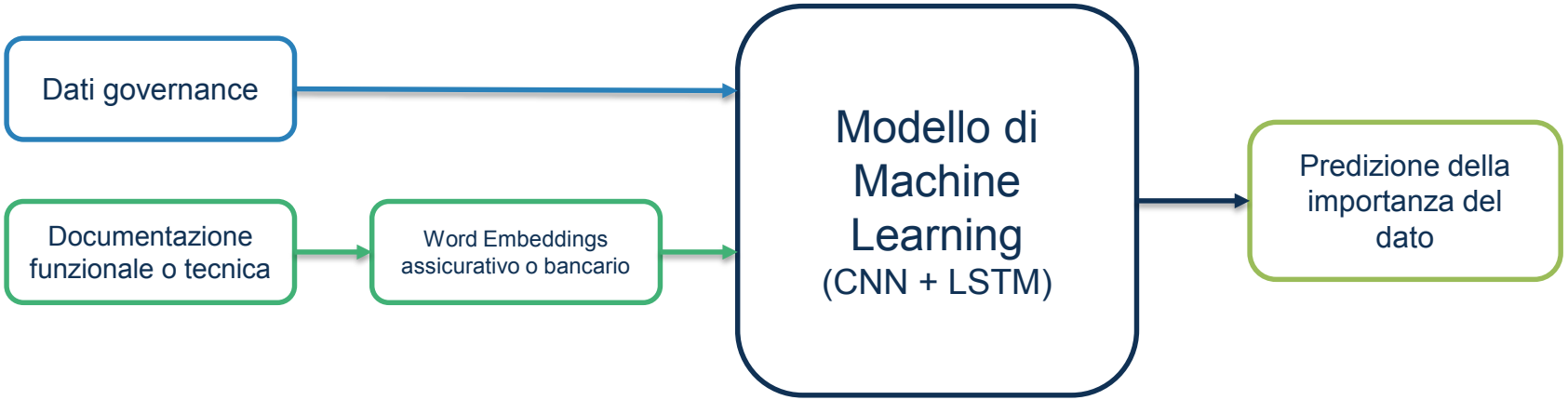
kirey group

Architettura



Questo workflow esemplifica il processo di per il raggiungimento di una predizione tramite tecniche di Deep Learning partendo da dati strutturati e documenti

Esempio di utilizzo



Questo workflow rappresenta un esempio di modello in grado di prevedere l'importanza del dato (Data Governance) tramite l'elaborazione di documentazione (funzionale o tecnica) e una serie di campi funzionali a un certo ambito di utilizzo.
Avendo a disposizione un pregresso di informazioni

system evolution kirey group

Grazie per l'attenzione!

kirey group

info@kireygroup.com

www.kireygroup.com