Dalla Governance della IA a quella dei Big Data

Donato Malerba
Università degli Studi di Bari Aldo Moro
CINI - Big Data Lab

donato.malerba@uniba.it

Nexa Center for Internet & Society
Politecnico di Torino

Conferenza Nexa su Internet & Società 2017
Etica e governance dell'Intelligenza Artificiale
Torino, 18 Dicembre 2017

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Chatbot del servizio clienti

- Nei portali delle amministrazioni i dati ci sono, anche in formato aperto, ma il cittadino non li usa direttamente per le proprie esigenze.
- Le chatbot possono aiutare il cittadino a ritrovare le informazioni di suo interesse.
- Vantaggi:

  - meno personale dedicato a rispondere a domande spesso banali e frequenti

  - più personale dedicato ad altre funzioni,

  - servizio al cittadino più efficiente.

Chatbot del servizio clienti

Potrebbe accadere, come sta accadendo in moderni help desk nel settore IT, di scoprire che molte domande sono relative a semplici esigenze informative. Sicché una chatbot può aiutare a rispondere risparmiando il tempo degli operatori reali per chiamate più complesse.

Di più, la chatbox può aiutare ad acquisire informazioni da chi chiama, nell'attesa che qualcuno prenda la chiamata.

Il binomio AI e Big Data

- **Ovunque ci siano dati in grande quantità, c'è spazio per l'IA.**

  L'IA può aiutare a gestire i grandi volumi di dati, ad analizzarli e a trovare delle regolarità (*pattern*) che non sarebbero altrimenti scopribili dagli esseri umani.

- **Le tecnologie di IA non possono prescindere dai dati.**

  Molti successi recenti dell'IA si basano su una enorme quantità di dati.

Il binomio AI e Big Data

Non ci può essere governance dell'IA senza governance dei dati.

La governance dei dati

- **Che vuol dire?**

  Significa saper rispondere tempestivamente e con certezza a domande su:

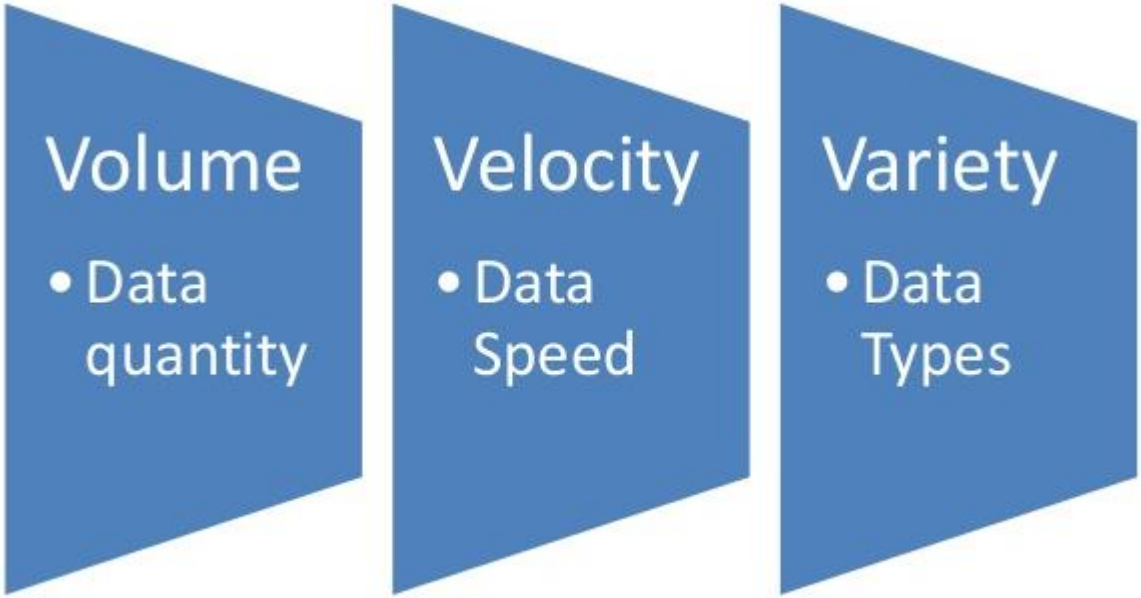
  - provenienza del dato
  - stato di aggiornamento del dato
  - responsabilità del dato
  - contesto normativo e suo impatto sull'uso del dato
  - uso effettuato del dato

Governance dei dati

- Nella governance dei dati si definisce un framework per impostare le **politiche di uso dei da**ti e si implementano controlli per assicurare che l'informazione resti **accurata**, **consistente** e **accessibile**.
- Senza un efficace processo di governance dei dati, si possono causare molti problemi, con costi inattesi.

Caratteristiche dei Big Data

- Le principali sono volume, velocità e varietà



Cosa comporta la governance dei dati

Partire dai processi e dalle persone

- La governance dei dati, siano essi “BIG” o meno, deve **incoraggiare le buone pratiche** e limitare quelle rischiose.
- Occorre individuare il **responsabile dei dati**, stabilire accordi su **come e per cosa usare i dati**, determinare **l’impatto del cambiamento** nell’uso dei dati sull’organizzazione.

Cosa comporta la governance dei dati

Usare la tecnologia per fare tutto questo 1/2

- I processi di data governance sono fortemente variabili e interessano molti soggetti nell'organizzazione. Per di più devono rispettare delle tempistiche, che sono particolarmente critiche nel caso dei big data. Questo richiede dei **sistemi dedicati e un alto livello di automazione**.
- **È la stessa governance dei dati a dover essere misurata e gestita**, come accade in tutti i processi, in modo da migliorarne la qualità e l'utilità. L'automazione aiuta.

Cosa comporta la governance dei dati

Usare la tecnologia per fare tutto questo 2/2

- **Criticità**: i sistemi di data governance si devono adattare a coloro che conoscono il significato e l'uso del dato, ma non hanno familiarità con aspetti tecnici di gestione del dato stesso.

Cosa comporta la governance dei dati

Anticipare gli aspetti della gestione del dato che sono critici per gli scopi dell'organizzazione

- È più importante sapere "**cosa**" governare di "**come**" implementare una giusta governance dei dati.
- Se è vero che tutti i dati potrebbero essere soggetti a qualche forma di governance, generalmente è solo un **sottoinsieme** a contenere l'informazione cruciale su cui andrebbero concentrati gli sforzi.

Non strafare

- Non occorre mai perdere di vista il fine ultimo
- Molti big data vengono dai social media, ma solo una piccola parte è di interesse. È un errore provare a controllare tutti I dati che giungono, perché si sprecherebbero tempo e risorse per dati poco rilevanti.
- Occorre conoscere quali siano le esigenze dell'organizzazione per capire quali dati conviene trattare.
- La valutazione della qualità dei dati in ingresso è una delle priorità nella governance dei dati.
- Si possono usare controlli di qualità proattivi, per i big data

Cosa comporta la governance dei dati

Fare leva su processi e best practice esistenti

- La governance dei dati **non deve essere un peso**. Occorre fare leva su “workflows” o “business rule” esistenti per avviare una strategia che miri a estrarre valore dai big data.
- Molte organizzazioni hanno già riconosciuto l’importanza di queste attività e messe a sistema, anche se **non necessariamente generano valore**. Ciò perché la transizione verso un’organizzazione guidata dai dati può richiedere nuove forme di analisi e di integrazione dei dati.

Requisiti chiave della governance dei dati

Identificare i dati e mantenerne un catalogo

- È fondamentale **catalogare i dati** e renderne il catalogo **disponibile agli utenti**. Non si tratta solo di mantener traccia dei metadati ma di automatizzare il processo. La categorizzazione manuale non è una scelta possibile, in molti casi.
- **L'accessibilità agli utenti ne favorisce l'utilizzo**.
- Occorre tener traccia di tutti gli *asset* legati ai dati, al loro utilizzo e alla loro elaborazione (jobs di MapReduce, visualizzazioni , modelli dei dati).

Requisiti chiave della governance dei dati

- Una sfida nella governance dei Big Data è rappresentata proprio dalla **categorizzazione**, dalla **modellazione** e dal **mapping** dei dati man mano che vengono catturati e memorizzati, particolarmente a causa della natura non strutturata dell'informazione.
- Per ottenere informazione utile dai big data, occorre fare molte cose, come l'analisi semantica del dato, che è poi resa in modelli concettuali e ontologie.

Requisiti chiave della governance dei dati

- Un problema fondamentale è che **i big data sono più orientati all'esplorazione e alla scoperta** che alle convenzionali analisi di business intelligence.

Un ciclo vizioso:

I big data non possono essere ben "governati" o "amministrati" fino a quando non vengono modellizzati, ma non possono neppure essere modellizzati se non quando vengono esplorati dall'analista dei dati.

Requisiti chiave della governance dei dati

Esplorare i dati per identificarne le opportunità

- Dopo aver catalogato i dati, occorre saperli **ritrovare** nel catalogo, tenendo conto dei diversi ruoli:
  - **Professionista IT**: vede i dati in un contesto applicativo;
  - **Security team**: vede i dati in un contesto di politiche di privacy;
  - **Auditor**: ha necessità di capire da dove proviene il dato e le sue relazioni al contesto.

  Si possono usare tecniche di machine learning e di AI per aiutare a trovare i dati appropriati.

Evitare i silos

- È importante fare il mapping dei nuovi dati a quelli di riferimento usati dall'organizzazione per categorizzare l'informazione.
- Se non fatto bene, l'informazione raccolta potrebbe essere fuorviante, imprecisa o incompleta.
- Va cercato un approccio unificato alla gestione e governo dei big data.
- **Occorre evitare che all'interno dell'organizzazione si costruiscano dei silos dei dati**.

Requisiti chiave della governance dei dati

Mantenere la validità dei dati

- Assicurarsi prima che **l'accuratezza del dato sia adeguata all'uso.**
- Quindi determinare l'**impatto dei cambiamenti nel dato**.
- Implementare politiche di valutazione di quali dati violano i requisiti minimi di accuratezza.

Requisiti chiave della governance dei dati

Google Flu Trends: previsione in base all'oggetto delle ricerche condotte con *Google search* ➔ ugualmente accurate ma in tempo reale

Stati Uniti: dati ILI (Influenza-Like Illness) forniti pubblicamente dagli U.S. Centers for Disease Control.

Detecting influenza epidemics using search engine query data. ***Nature 457, 1012-1014 (19 February 2009)***

Requisiti chiave della governance dei dati

Assicurare affidabilità e disponibilità dei dati

- Le strategie di governance dei dati devono includere quelle per l'**elevata disponibilità** e per il **disaster recovery**, pena il deprezzamento del valore dei dati.
- Un effetto benefico della **diffusione dei dati a livello globale** è quello di facilitare il disaster recovery. Tuttavia c'è la tendenza a trattenere nel paese nativo di un individuo i suoi dati personali.

Requisiti chiave della governance dei dati

Proteggere i dati sensibili

1. Identificare i dati più sensibili
2. Stabilire ciò che può essere fatto con quei dati e da chi.
3. Sviluppare metodi sistematici per acquisite questa informazione e disseminarla sia ai consumatori che ai tecnologi.

È importante il controllo della fase di ingestion dei dati per identificare il livello di protezione adeguato.

Assegnare policy predefinite o definirne di nuove per la privacy e sicurezza.

Governance dell'IA?

- La governance dei dati aiuta le organizzazioni a prendersi cura dei dati che posseggono, a trarne valore, e a rendere visibili agli utenti aspetti importanti dei dati che posseggono.
- La criticità non è solo dovuta a possibili errori o incompletezze del dato, ma ai nuovi usi del dato, che spesso richiedono nuovi attributi e nuovi metadati per supportarli.
- La governance dei big data è un tema nuovo e ancora da esplorare.

Saremo pronti a implementare una adeguata governance dell'IA solo quando saremo in grado di implementare una adeguata governance dei big data

Il progetto

Toreador

- TOREADOR (www.toreador-project.eu) è uno dei principali progetti di ricerca finanziati dalla Commissione Europea nell'ambito della call Big Data Research del programma H2020.
- Coordinato dal Laboratorio Nazionale Big Data del CINI
- Principal Investigator: *Prof. Ernesto Damiani*.
- Partecipano SAS, Atos, Engineering che fanno parte della Big Data Value Association.

cini Big Data national lab

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Il progetto

Toreador

- **Obiettivo primario:** aiutare le organizzazioni e le PMI europee che desiderano eseguire computazioni di analitiche Big Data, diminuendo l'investimento iniziale e il livello di competenza tecnologica necessari.
- **Come raggiungerlo?**

Il progetto

Toreador

- Sviluppando *modelli* ontologici indipendenti dalla tecnologia per rappresentare
- (i) le entità (dati, algoritmi, metriche) coinvolte nelle computazioni di analitiche su Big Data e
- (ii) le attività che tali computazioni richiedono: l'acquisizione, memorizzazione e preparazione dei dati, le analitiche da calcolare, la distribuzione e parallelizzazione dei calcoli, la presentazione e interpretazione dei risultati.

Il progetto

Toreador

- Metodologia: accompagnare il committente di un'analitica su Big Data attraverso tre fasi:
- *Individuazione* degli obiettivi (funzionali e di prestazioni) della computazione da eseguire (*modello dichiarativo*)
- *Derivazione* assistita di una descrizione della computazione indipendente dalla tecnologia (*modello procedurale*).
- *Traduzione* semi-automatica del modello procedurale in una procedura operativa (*modello di messa in opera*) eseguibile nell'ambiente di calcolo desiderato (sistema ICT del committente o in **cloud**)

Il progetto

Toreador

- Un elemento importante della prima fase della metodologia TOREADOR è la gestione dei conflitti tra scelte diverse nei modelli delle varie fasi (ad esempio, tra la rappresentazione del dato e l'accuratezza dei risultati dell'analitica).

Il progetto

Toreador

- La metodologia TOREADOR comprende la modellazione dello *status giuridico* del dato (dato personale/non personale, consenso, scopo).
- TOREADOR ha affidato a un partner del consorzio, lo studio legale internazionale Bird&Bird, uno studio delle **casistiche** che permettono di modellare lo stato giuridico del dato alla luce della recente **normativa europea**.

Il progetto

Toreador

- Duplice scopo della modellazione dello stato giuridico del dato:

1. aiutare il committente di una analitica Big Data a specificare obiettivi di anonimizzazione adeguati allo status giuridico dei dati su cui opera, nel contesto dell'analitica che vuole eseguire;
2. includere automaticamente nei modelli procedurali che descrivono la computazione dell'analitica, ed in particolare nel modello della fase di preparazione del dato, le necessarie procedure di anonimizzazione (ad esempio, hashing non reversibile).

Il progetto

Toreador

- La metodologia TOREADOR individua poi, in sede di risoluzione dei conflitti tra gli obiettivi delle varie fasi, quali sono le analitiche compatibili con le procedure di anonimizzazione da adottare e gli obiettivi di accuratezza raggiungibili.

Il progetto

Toreador

- Oltre ad essere traducibile nella messa in opera di una computazione eseguibile, il modello TOREADOR che specifica lo status giuridico assegnato ai vari dati e le corrispondenti procedure di anonimizzazione da adottare è utilizzabile ai fini contrattuali (ad esempio, per specificare le condizioni di fornitura dell'analitica) e per ispezione/validazione da parte di terzi.
- Questo consente di definire un Privacy Seal per modelli, che permetterebbe di certificare il modello e poi di individuare la tecnologia che lo realizza.

Il progetto

Toreador

- In questo modo si intende anche affrontare il problema della privacy che spesso impedisce l'uso di sistemi cloud pubblici in applicazioni di big data analytics.

Cercare gli indizi nei Big Data

- La precisone di dati che provengono da sorgenti esterne è difficile da validare
- Il significato e il contesto dei dati testuali non è sempre evidente
- I big data pongono sfide ai manager IT, ai programmatori, agli architetturisti del dato, a coloro che modellizzano I dati e ai professionisti della governance dei dati
- Servono anche elevate prestazioni per consegnare i dati richiesti ove richiesto