Audizione CUN sul tema "Data Science" il 14 giugno 2018
Contributo del Gruppo di Informatica (SSD INF/01)

Gli obiettivi culturali di questa classe e i contenuti disciplinari e le competenze trasversali indispensabili che dovrebbero essere fornite da tutti i corsi della classe.

Secondo J.Wing (direttrice del Data Science Institute della Columbia Univ.) la Scienza dei Dati (Data Science) è lo *studio di come estrarre valore dai dati*.
Il "*valore*" è stabilito dall'utente, cioè dall'esperto di dominio. Ad esempio, una grossa azienda tecnologica potrebbe associare valore a un miglioramento nei propri sistemi da un "data product", misurato in termini di click su una pagina, il tempo dedicato ad usare un servizio, il denaro speso per ottenere un risultato. Per chi deve definire delle politiche, sociali o di mercato, il valore potrebbe derivare da un'analisi e visualizzazione di dati che giustificano un cambiamento di politica. Per uno studioso, il valore potrebbe essere associato alla creazione di nuova conoscenza: una scoperta scientifica, una spiegazione di un comportamento sociale, una nuova interpretazione del mondo intorno a noi. La definizione del valore estratto dai dati suggerisce la presenza nella nuova classe LMDS di contenuti disciplinari relativi alle scienze economiche e sociali. Potrebbero essere utili anche competenze di etica e diritto per insegnare agli studenti l'importanza del dato, che va protetto e trattato rigorosamente (e.g., Il nuovo regolamento europeo sulla privacy - GDPR).
Il valore della Data Science si può articolare in modi molto diversi a seconda del dominio. **Le competenze trasversali indispensabili sono comunque quelle informatiche e matematico-statistiche**.
L'azione di "*estrarre*" riguarda l'uso di algoritmi e strumenti il cui scopo finale è di inferire automaticamente pattern e modelli dei dati. I dati seguono un ciclo di vita composto di varie categorie di azioni concatenate: generazione, raccolta, elaborazione, conservazione, gestione, analisi, visualizzazione, interpretazione. Di solito occorrono alcune trasformazioni prima che dai dati si possa estrarre valore. La definizione del ciclo di vita e la messa in opera delle operazioni di trasformazione richiede competenze informatiche. Una competenza informatica tipica di trasformazione dei dati si chiama "data mining", che ingloba l'insieme delle tecniche e delle metodologie che hanno per oggetto l'estrazione di informazioni o modelli utili da grandi quantità di dati. Un'altra competenza tipica si chiama "machine learning" (apprendimento automatico), che utilizza metodi statistici per sviluppare algoritmi che identificano pattern ricorrenti nei dati, e studia algoritmi che apprendono e inducono dai dati modelli predittivi. Le attività correlate richiedono l'uso di grossa potenza di calcolo, quale quella resa disponibile dalle architetture di supercalcolo (competenze informatiche di HPC: High Performance Computing).
Infine, lo "*studio*" riguarda la teoria e la pratica tipici di ogni disciplina scientifica. La Data Science nasce dalla necessità di gestire su base statistica le enormi quantità di dati oggi disponibili per predizioni, analisi e comprensione di fenomeni sociali o

scientifici. È una scienza collegata ai moderni processi di trasformazione digitale ed alle reti che formano i sistemi complessi, in quanto enfatizza il valore e la necessità di approssimazione e semplificazione [Blei and Smyth 2017]. Gli aspetti teorici sono ben ricompresi nei settori delle matematiche applicate, della statistica e dell'algoritmica, mentre quelli pratici hanno bisogno delle competenze dei settori informatici e della statistica computazionale.

Quali potrebbero essere dei naturali sbocchi professionali, o sbocchi verso il proseguimento degli studi, coerenti con gli obiettivi della classe.

La Data Science è molto giovane, ma alcuni profili professionali sono già comparsi e risultano stabilmente rappresentati nel mercato del lavoro; ne elenchiamo alcuni con la denominazione che si è affermata nel mondo anglosassone:
- Data Analyst
- Data Engineer (progettista di infrastrutture big data)
- Data Manager (esperto di Data Governance)
- Machine Learning engineer
- Computational statistician

I "data professionals" vanno comunque dotati di un mix di competenze multidisciplinari che permettano non solo di acquisire dati ed estrarne senso e conoscenza, ma anche di raccontare "storie" attraverso questi dati, a supporto delle decisioni, della creatività e dello sviluppo di servizi innovativi, e di saper gestire le ripercussioni etiche e legali dei Big Data, che spesso contengono informazioni personali e suscitano problematiche relative alla privacy, alla trasparenza, alla consapevolezza, all'uso etico della conoscenza racchiusa nei dati.

In generale gli sbocchi occupazionali sono i seguenti
- nel settore terziario e nelle Pubbliche Amministrazioni, nell'ambito dei servizi innovativi basati sui dati, specie quelli fruibili on-line, in modalità mobile, o basati su social network;
- nel settore industriale, nell'ambito dei processi di trasformazione digitale nei quali le decisioni di livello operativo, tattico/manageriale e strategico/direzionale sono basate su informazioni ottenute in modo tempestivo e sistematico a partire da grandi raccolte di dati costruite internamente ed esternamente agli stessi processi industriali; particolare attenzione va rivolta all'introduzione di automazione massiccia derivante dal programma Industria 4.0
- nel settore primario, con particolare riferimento all'uso delle ICT nello sfruttamento delle risorse naturali (scienze ambientali, meteorologia, scienze agrarie e alimentari, estrazione di minerali, ecc.);

Gli sbocchi verso il proseguimento degli studi:
- nell'ambito scientifico, con particolare riferimento alle scienze dei sistemi complessi come la fisica sperimentale, la chimica e la biologia, ma anche le scienze

informatiche basate sul paradigma della scoperta scientifica guidata dai dati (a titolo di esempio: Artificial Intelligence, Empirical Software Engineering, Bioinformatica, e Brain Informatics)
- nell'ambito economico-sociale, ad esempio in settori quali econometria, marketing quantitativo, digital sociology

In tutti questi ambiti il data scientist è chiamato a progettare e realizzare dei *data products*, risultati valorizzabili basati su analisi descrittive, predittive, o prescrittive per sistemi complessi, utilizzando, quale materia prima, collezioni di dati organizzati e analizzabili prevalentemente attraverso strumenti automatici.

Se sia necessario introdurre altri elementi (per esempio presenza obbligatoria di tirocini o stage, attività laboratoriali, competenze linguistiche, eccetera) indispensabili per il raggiungimento degli obiettivi della classe.

Come per tutte le altre classi di laurea, ai laureati di questa classe si richiede di essere capaci di comunicare efficacemente, in forma scritta e orale, in almeno una lingua dell'Unione Europea, oltre l'italiano, anche con riferimento ai lessici disciplinari.

Ai fini indicati, i curricula dei corsi di laurea magistrale della classe
- prevedono (fra i requisiti curriculari di accesso) almeno la conoscenza della lingua inglese;
- prevedono lezioni ed esercitazioni di laboratorio oltre ad attività progettuali autonome e attività individuali in laboratorio;
- possono prevedere attività esterne, come tirocini formativi, presso aziende e industrie, enti pubblici o istituti di ricerca, laboratori, oltre a soggiorni di studio presso altre università italiane ed europee;
- culminano in un'attività di progettazione o di ricerca o di analisi di caso, che dimostri la padronanza degli argomenti, nonché la capacità di produrre in modo autonomo dei data product.

Osservazioni aggiuntive

In cosa la Data Science differisce dall'Informatica (Computer Science)? A prima vista Data Science è molto più sperimentale. L'Informatica è incentrata sulla teoria della computabilità, e sviluppa in modo sistematico le tecniche di trattamento automatico dell'informazione e di risoluzione esatta/approssimata di problemi di natura computazionale. La Data Science accetta l'incertezza e l'approssimazione come nozioni fondative. Per entrambe usa modelli probabilistici, statistici o subsimbolici capaci però di supportare un ragionamento matematico-formale. L'Informatica è invece fortemente radicata nella Logica simbolica e su modelli componibili per livelli astrazione. Si presuppone che gli algoritmi forniscano risultati precisi o precisamente definibili, e l'incertezza è vincolata nel "non-determinismo".

Molte aree dell'Informatica utilizzano modelli probabilistici, ma spesso questi sono strutturati su elementi discreti o di derivazione logica.

L'ambito di ricerca e di mercato della Data Science è estremamente dinamico, potrebbe essere soggetto ad evoluzione veloce, e quindi se è importante cogliere l'occasione della corrente fase di manutenzione delle classi per introdurre una classe in Data Science, la sua struttura non deve essere troppo rigida, né negli obiettivi formativi né nella specifica degli ambiti e dei settori.

Occorre tenere presenti altre possibili future novità nelle aree multidisciplinari che coinvolgono l'Informatica; segnaliamo come esempio l'importanza della classe LM-91, che fu definita vent'anni fa e venne inizialmente poco utilizzata, ma ha poi permesso l'istituzione di diversi corsi di Data Science.
Nella declaratoria della LM-91 la parola "dato" non compare mai. Tra i requisiti curricolari di accesso, l'unico esplicitato è la conoscenza della lingua inglese. Inoltre, i settori caratterizzanti della LM91 sono molteplici ed eterogenei. Questo dà un'idea abbastanza precisa della natura ibrida della LM91. La nuova LM dovrà invece meglio caratterizzarsi per rigore scientifico e metodologico. Gli SSD INF/01, ING-INF/05, MAT/06, MAT/09 e SECS-S/01 devono essere identificati come "fondazionali" (ovvero caratterizzanti).
Si suggerisce di affiancare la nuova LM alla preesistente LM-91 mantenendo la possibilità di accomodare nuove necessità e tendenze che emergano nel prossimo futuro, e che non trovino collocazione nelle classi già esistenti. LM DS sarà più caratterizzata in senso STEM, mentre la LM91 potrà continuare ad accogliere proposte multidisciplinari come ha fatto in questi ultimi anni.
I requisiti di ingresso della nuova LMDS vanno comunque tenuti laschi, in modo che le nuove lauree magistrali possano accogliere studenti di diverse discipline.

Reference
[Blei and Smyth 2017] D. Blei and P. Symth, "Science and Data Science", Proc. National Academies of Sciences, 114:33(8689-8692), June 2017.