Parere sull'opportunità di creare una nuova classe di laurea magistrale in "Data Science"

Audizione dell'Istituto nazionale di statistica

Consiglio Universitario Nazionale

Roma, 14 giugno 2018

Indice

1. Introduzione

In questa audizione l'Istat offre il proprio parere al Consiglio Universitario Nazionale sull'opportunità di creare una nuova classe di laurea magistrale in Data Science.

Il documento introduce il tema della crescita delle informazioni disponibili e della necessità di saperle selezionare e trattare al fine di accrescere le conoscenze e per poterle utilizzare nei processi decisionali. Da queste premesse si illustra il forte interesse delle Autorità statistiche nel modernizzare i propri processi produttivi sfruttando il complesso delle informazioni disponibili, compresi i cosiddetti Big Data. Si giunge quindi a sottolineare, in questo contesto, l'importanza per l'Istituto nazionale di statistica di poter disporre della figura professionale del Data scientist e si propone la riflessione del Sistema statistico europeo (ESS) in merito alle competenze necessarie. Si ricorda, in conclusione, l'impegno e l'interesse già profuso dall'Istat nel promuovere iniziative di formazione in Data Science.

In appendice alcune evidenze relative alla professione del Data scientist rintracciabili nelle indagini dell'Istat sull'inserimento professionale di laureati e dottori di ricerca.

2. La crescita delle informazioni disponibili

I rapidi progressi della tecnologia dell'informazione e l'immenso patrimonio di dati oggi disponibili ha portato molti ricercatori ad asserire che la ricerca scientifica sia di fronte ad un cambio di paradigma. La scienza, inizialmente sperimentale, basata sull'osservazione diretta della natura è diventata teorica per poi evolvere, alcuni decenni fa, a disciplina computazionale e trasformarsi oggi in scienza ad alta intensità di dati. Il Data Science è dunque immaginato da molti come il quarto paradigma della scienza (Jim Gray, vincitore del premio Turing).

L'attuale contesto, caratterizzato da una produzione dati provenienti da variegate fonti, in una quantità e con una velocità fino a qualche anno fa inimmaginabile, offre nuove sfide e opportunità di ricerca, che vedranno la

statistica ufficiale impegnata nell'estrarre 'valore' dai nuovi dati. Dovranno essere messe in atto strategie che permettano l'integrazione di fonti diverse per arricchire l'offerta e al tempo stesso la qualità e la tempestività dell'informazione statistica prodotta. La capacità di estrarre dai dati informazioni che abbiano un significato e siano funzionali richiederà lo sviluppo di sofisticate tecnologie e di competenze interdisciplinari che operino a stretto contatto.

Il nuovo paradigma basato sui dati, che sfrutta le funzionalità tecnologiche per la loro raccolta continua e massiva, per la loro trasmissione istantanea e il loro riutilizzo, espone a nuovi rischi e aumenta in modo esponenziale la vulnerabilità degli individui. Il passaggio a un nuovo modello di produzione statistica dovrà dunque avvenire nel rispetto della Privacy e nella piena consapevolezza del potenziale discriminatorio derivante dalle profilazioni sempre più puntuali ed analitiche ora possibili. Saranno necessarie nuove infrastrutture (metodologiche, tecnologiche, organizzative) nonché un quadro legale ben definito che permettano la gestione delle problematiche connesse al trattamento e alla privacy dei big data e non ne inibiscano l'uso.

3. La modernizzazione dei processi statistici

La disponibilità di dati di diversa natura, sia 'primaria', ovvero raccolti per scopi statistici, sia 'secondaria', come nel caso dei dati amministrativi o dei big data, ottenuti nell'ambito di processi non statistici ma con un elevato potenziale per la produzione di statistiche ufficiali, e il contestuale calo nei tassi di risposta delle indagini condotte nell'ambito del Sistema Statistico Europeo, ha spinto molti Istituti di statistica ad innovare il modello di produzione tradizionale.

Il programma di modernizzazione avviato dall'Istat dalla seconda metà del 2014 ha recepito le esigenze di cambiamento derivanti dalla nuova disponibilità di dati, dalle nuove modalità di misura dei fenomeni, dalla necessità di superare le criticità del sistema tradizionale, onerose sia per il budget dell'istituto sia per l'impegno richiesto ai rispondenti, dalla forte richiesta di integrazione e standardizzazione all'interno del sistema delle statistiche europee.

Arricchire l'offerta e la qualità dell'informazione statistica è uno degli obiettivi principali del programma di modernizzazione dell'Istat e il sistema integrato dei registri ne rappresenta l'elemento fondante. I livelli di integrazione previsti sono di diversa natura: concettuale, logico/fisica e

statistica. Quest'ultima contempla, fra le altre, l'integrazione dei dati 'primari' e 'secondari', comprensivi delle fonti innovative e dei Big Data.

L'Istat ha istituito una commissione tecnica con il compito di definire policy a supporto dell'uso dei Big Data per la Statistica ufficiale e di monitorare ed orientare le scelte dell'Istituto sul tema. In particolare, la commissione tecnica, che prevede l'affiancamento agli esperti di statistica ufficiale di esperti che provengono dall'accademia, dal privato e da altri enti ed istituzioni pubbliche, lavora sui seguenti aspetti: (i) orientamento tecnico scientifico dei progetti che utilizzano fonti Big Data in Istat e (ii) supporto alle partnership e alla cura della data privacy. Tra i principali progetti attivi in Istat sull'uso dei big dati citiamo: (i) uso degli scanner data per la stima dell'inflazione; (ii) uso di Web data per la stima di indicatori sull'uso dell'ICT da parte delle imprese; (iii) uso dei dati di Twitter per il calcolo di indicatori giornalieri di sentiment.

L'Istat partecipa attivamente a progetti europei sul tema Big Data. Ha un ruolo di leadership nel recente progetto "ESSnet Big Data Pilots I", progetto all'interno del Sistema Statistico Europeo (ESS) il cui obiettivo è quello di integrare i big data nella produzione della statistica ufficiale. Il progetto è parte del *Big Data Action Plan and Roadmap 1.0* e rientra nei progetti di attuazione della ESS Vision 2020, ossia la strategia europea per la modernizzazione della produzione delle statistiche europee.

L'obiettivo generale del progetto è di preparare il sistema statistico europeo all'integrazione delle fonti di big data nella statistica ufficiale. L'idea è che l'attività si concentri nell'avvio di progetti pilota che analizzino le potenzialità delle fonti di big data per la produzione o il contributo alla produzione statistica. Scopo di questi pilot è intraprendere azioni concrete nel dominio dei big data e ottenere un'esperienza diretta per il loro utilizzo nella produzione della statistica ufficiale. Esempi di fonti/domini nell'ambito dei quali sono stati realizzati pilot sono: *webscraping job vacancies*, *webscraping enterprise characteristics*; *smart meters*; *AIS Data*; *mobile phone data*; *early estimates*; *multiple domains and methodology*. Istat coordina il *work package* "*web scraping enterprise characteristics*".

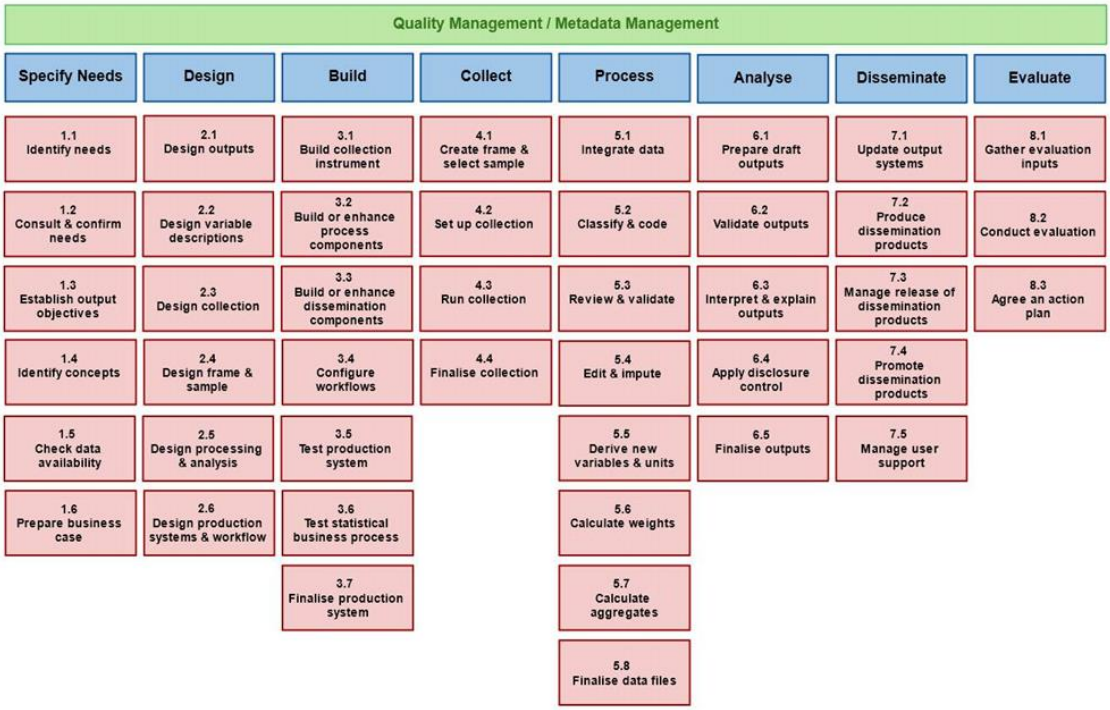
L'apertura della statistica ufficiale verso i Big data richiede l'utilizzo di nuovi strumenti di analisi statistica, dal momento che quelli tradizionali sono messi in crisi dal Volume e dalla Varietà che contraddistingue le nuove fonti di dati. Le tecniche Machine Learning (ML) diventano uno strumento indispensabile per superare i vincoli dei metodi statistici classici, basati su algoritmi adatti a

basi dati di volume contenuto, di elevata qualità e con una struttura a matrice 'unità-variabile'.

L'elevato numero di unità e la ricchezza di variabili disponibili fanno dei registri un ulteriore ambito di sperimentazione delle tecniche ML, per modellare relazioni non lineari e interazioni complesse fra variabili.

Anche all'interno del tradizionale modello di produzione di dati primari (GSBPM) elaborato dall'UNECE, si individuano numerose fasi deputate a beneficiare delle nuove tecniche ML. Si consideri ad esempio il processo 2.4, relativo alla preparazione della lista di campionamento, quando fonti di natura diversa sono combinate tramite procedure di record linkage che utilizzano algoritmi di clustering. Nella fase di selezione del campione (4.1) possono essere utilizzate tecniche di classificazione per stratificare la popolazione e migliorare l'efficienza delle stime. Anche nella fase 4.3 relativa alla fase di raccolta dei dati, l'uso di algoritmi di regressione potrebbero essere utilizzati per prevedere la probabilità di risposta di un'unità sulla base delle informazioni disponibili, permettendo in questo modo una gestione efficiente delle attività di raccolta dati.

Figura 1 - Modello GSBPM (Generic Statistical Business Process Model sviluppato nell'ambito della Commissione economica per l'Europa delle Nazioni Unite



4. La figura professionale del data scientist

Nell'ambito del nuovo assetto produttivo dell'Istat emerge chiaramente l'importanza della nuova figura professionale del data scientist e si auspica pertanto l'istituzione della nuova classe di laurea magistrale in Data Science.

Le risorse attualmente disponibili all'Istat e impiegate in attività legate al Data Science sono decisamente contenute (circa una dozzina). Il confronto con altre realtà suggerisce ampi margini di investimento per l'Istituto per l'acquisizione di figure professionali competenti in materia. Nel futuro sarà pertanto importante per l'Istat poter beneficiare di risorse provenienti da specifici percorsi di istruzione universitaria, dedicati all'approfondimento di tematiche relative al Data Science.

Le indicazioni che seguono sintetizzano le riflessioni avvenute all'interno dell'*ESS Steering group on Big data and Official statstics* di Eurostat. Riportano inoltre le principali evidenze del '*Draft report on the development of a training strategy to bridge the big data skill gap in European official statistics*', elaborato per conto di Eurostat e contenente tra gli altri i risultati di un'indagine condotta presso i focal point europei sui Big Data.

Gli obiettivi culturali della nuova classe di laurea magistrale in Data science dovranno essere interdisciplinari e comprendere gli ambiti delle scienze matematiche, statistiche, informatiche e dell'ingegneria dell'informazione (cfr. Appendice).

In termini generali, lo studente che completerà il percorso di istruzione dovrà essere in grado di svolgere un ruolo attivo nel lavoro di progettazione e implementazione

- dell'architettura dei dati, individuando l'organizzazione più consona alle successive fasi di analisi, visualizzazione e presentazione;
- dell'acquisizione dati e dell'archiviazione dei dati in una forma che li renda altamente riutilizzabili;
- dell'analisi dati attraverso approcci inferenziali, avendo consapevolezza dei limiti e sapendo quantificare l'accuratezza dei risultati.

Le competenze trasversali richieste ad un data scientist sono molteplici e comprendono la capacità di comunicare, di coordinare il lavoro, di problem solving 'creativo', di innovazione, nonché la capacità di comprendere e anticipare questioni etiche come la privacy, in modo da prevenire l'uso

improprio di dati o risultati analitici. La molteplicità di competenze settoriali e trasversali presuppone una modalità di lavoro corale di data scientists ai quali verrà dunque richiesta una spiccata capacità di lavorare in gruppo.

Lo sviluppo delle competenze settoriali dovrà avvenire anche attraverso attività laboratoriali che permetteranno di acquisire capacità nel maneggiare dati di tipologia e provenienza differenti (*web-scraped data*, *mobile phone data*, *sensor data*, *scanner data*), nell'uso dei principali linguaggi di programmazione, dei differenti database, piattaforme di big data, strumenti di visualizzazione e metodi di analisi statistica.

Estremamente utili al raggiungimento degli obiettivi formativi e al successivo inserimento nel mercato del lavoro saranno inoltre le esperienze di stage presso aziende, affrontando specifici casi di studio nell'ambito dei differenti settori di attività (finanza, ICT, media, energia, salute, turismo). Gli stage potranno essere finalizzati all'elaborazione della tesi.

Infine, dal momento che il data scientist dovrà avere una forte interazione con la comunità scientifica internazionale, sarà fondamentale una conoscenza a livello avanzato della lingua inglese

5. L'Istat e la promozione di iniziative di formazione in Data Science

L'Istat partecipa al network EMOS (Master europeo di statistica ufficiale), progetto promosso da Eurostat, con la collaborazione degli istituti nazionali di statistica e le Università, per realizzare attività di alta formazione, finalizzate a garantire che la Statistica Ufficiale entri a far parte dei curricula accademici e a preparare figure professionali con competenze specifiche che aiutino a migliorare il sistema della statistica ufficiale in Europa.

La rete EMOS comprende attualmente quattro università italiane (Firenze - Dipartimento di Statistica, Informatica e Applicazioni; Pisa - Dipartimento di Economia e Management; Roma - Dipartimento di Scienze statistiche; Bergamo – Dipartimento di Scienze Aziendali, Economiche e Metodi quantitativi) che hanno istituito corsi di laurea con curriculum " Official Statistics".

L'impegno dell'Istat all'interno del network EMOS si concretizza nelle attività:

- didattiche: esperti dell'Istituto sono coinvolti in attività seminariali e di docenza su temi della statistica ufficiale
- di tirocinio presso le sedi dell'Istat della durata di 6/8 settimane

- di tesi di laurea, dedicate all'approfondimento di argomenti rilevanti per la Statistica ufficiale

L'interesse dell'Istat nel promuovere iniziative formative è dimostrato inoltre dal patrocinio di master in Data Science promossi da alcuni atenei.

L'Istat ha anche partecipato con una sua squadra alla European Big Data Hackathon. Un evento immersivo in cui i ricercatori dell'Istat hanno sviluppato uno strumento visuale che permette di esplorare e analizzare in maniera interattiva i mercati del lavoro europei, partendo dall'integrazione di Big data (*webscraping* sulle offerte di lavoro e i *repository* di curricula delle agenzie di lavoro) e dati da indagine (*Labour force survey*, Eu-Silc, *Assessment of Adult Competencies*). Il lavoro ha condotto all'elaborazione di indicatori originali, microfondati, su diversi aspetti come il matching delle competenze, la disponibilità alla mobilità e la valutazione delle fonti e delle classificazioni.

Nel futuro l'Istat intende rafforzare e ampliare la collaborazione con le Università e rinnova in tal senso la disponibilità a partecipare attivamente alle fasi di progettazione dei curricula.

Materiale disponibile su richiesta:

- ✓ The use of machine learning in official statistics UNECE Blue Skies Thinking Network https://statswiki.unece.org/display/BST/Blue+Skies+Thinking+Network
- ✓ European Commission 2018 Draft report on the development of a training strategy to bridge the big data skill gap in European official statistics.

Appendice

La professione del Data scientist nelle indagini dell'Istat sul mercato del lavoro

Le indagini sull'inserimento professionale dei laureati nel 2011 (condotta nel 2015) e dei dottori di ricerca del 2012 e 2014 (condotta nel 2018) permettono di quantificare il numero di individui che, nella stringa testuale dedicata al nome della professioni, riportano un riferimento riconducibile al data scientist[1]. Nella popolazione dei dottori che hanno conseguito il titolo nel 2012 o nel 2014 coloro che dichiarano una professione legata al Data Science sono l'1% dei dottori che risultano occupati[2], mentre fra i laureati del 2011 la stessa percentuale è pari allo 0,1%.

Il 75% dei laureati che svolgono la professione di data scientist provengono da corsi di laurea che afferiscono al gruppo Istat-Miur 'Economico-statistico' (43,1%) e 'Scientifico' (31,8%), all'interno del quale vi è una netta prevalenza dell'area delle scienze matematiche e informatiche (29,1%). Dagli stessi gruppi proviene il 79% dei dottori di ricerca, sebbene in questo caso il gruppo modale di provenienza sia quello scientifico (cfr Tabella 1).

Tabella 1 - Laureati del 2011, intervistati nel 2015, e dottori di ricerca del 2012 e 2014, intervistati nel 2018, che hanno dichiarato di svolgere la professione di 'Data Scientist' per gruppo di corso di laurea e dottorato di provenienza
(valori percentuali)

GRUPPO DI CORSO	Laureati	Dottori di ricerca
Economico-statistico	**43,1**	**23,3**
Scientifico	**31,8**	**56,2**
Di cui:		
Scienze matematiche e informatiche	*29,1*	*37,0*
Scienze fisiche	*2,6*	*19,2*
Ingegneria	**19,8**	**20,5**
Politico-sociale	**2,7**	
Chimico-farmaceutico	**1,6**	
Giuridico	**0,9**	
Totale	**100,0**	**100,0**

[1] Nel descrivere la professione gli intervistati hanno utilizzato una delle seguenti parole chiave: 'data scien*', 'big data', 'data min*, 'data anal*', 'computer scien', 'text min*', 'machine ler*'

[2] L'indagine si chiuderà il 15 giugno. Il dato non è corretto per mancata risposta e si riferisce al 72% dei dottori che hanno conseguito il titolo nel 2012 e nel 2014.

Circa l’80% dei data scientist di entrambe i collettivi è impiegato nel settore economico dei servizi, principalmente nei Servizi di informazione e comunicazione. In base alla prevalenza delle attività lavorative svolte, i laureati e dottori di ricerca che svolgono una professione riconducibile al data scientist hanno codificato la loro professione utilizzando i seguenti codici della classificazione Cp2011:

2.1.1.3.2 - Statistici

Le professioni comprese in questa unità conducono ricerche su concetti e teorie fondamentali della scienza attuariale e della statistica, incrementano la conoscenza scientifica in materia, applicano le relative teorie e tecniche per raccogliere, analizzare e sintetizzare informazioni, per definire modelli di interpretazione dei dati, per individuare soluzioni statistiche da adottare nei vari settori della produzione di beni e servizi e della stessa ricerca scientifica.

2.1.1.5.2 - Analisti e progettisti di basi dati

Le professioni comprese in questa unità analizzano, progettano, sviluppano e collaudano i sistemi di gestione di banche dati, garantendone e controllandone le prestazioni ottimali e la sicurezza. Definiscono e predispongono i sistemi di backup e le procedure per preservare la sicurezza e l'integrità dei dati.

2.1.1.4.1 – Analisti e progettisti di software

Le professioni classificate in questa categoria incrementano la conoscenza scientifica nelle scienze dell'informazione e della telematica. Sviluppano, creano, modificano o ottimizzano software applicativi analizzando le esigenze degli utilizzatori; analizzano i problemi di elaborazione dei dati per diverse esigenze di calcolo e disegnano, individuano o ottimizzano appropriati sistemi di calcolo delle informazioni; si occupano dell’ideazione, della realizzazione, dell’integrazione e della verifica dei software impiegati in un sito o in un’applicazione web.