**Incontro dei candidati alla presidenza SIS con il gruppo Statistica e Data Science,**
*Piattaforma Zoom, giovedì 11 giugno 2020, ore 15:00-16.15.*

Risposte ai quesiti posti al candidato Giuseppe Arbia

Domanda 1. Quale ritenete sia e come pensate possa evolvere il rapporto fra Statistica e Data Science?

La domanda è articolata e richiede due risposte. La prima relativa al rapporto attuale fra Statistica e Data Science, la seconda relativa a tale rapporto, ma visto in prospettiva.
Sulla prima domanda, in realtà, da come è formulata, si dà già per scontato che Statistica e Data Science siano due campi distinti, mentre la disciplina Data Science, da principio (diciamo quando se ne iniziò a parlare una decina di anni fa) si presentava come una sintesi delle *skills* dello statistico unite a quelle dell'informatico di fronte alle sfide dei Big Data.
Detto ciò è pur vero che, almeno in Italia i due campi sono intesi ormai come distinti ancorché in dialogo l'un l'altro.
Parlando più in generale, va osservato come il rapporto tra Statistica e Data Science sia molto variegato nel mondo. In paesi diversi si osservano forme di convivenza diverse. In particolare, nei paesi caratterizzati da una cultura statistica più avanzata e dove la figura dello statistico è maggiormente riconosciuta e consolidata (parlo, ad esempio, del mondo anglo-sassone che conosco meglio), il rapporto è più facile e meno di sudditanza della nostra disciplina. In Italia (ma non solo) dobbiamo purtroppo ancora, invece, combattere per evidenziare la specificità del nostro contributo. Inoltre, osserviamo come stia, purtroppo, prevalendo l'idea (pericolosa) che basti accumulare dati (senza attenzione ai criteri di raccolta e alla qualità del dato) e disporre di algoritmi (spesso slegati dalla realtà che si vuole analizzare) per giungere ad una migliore conoscenza senza bisogno della statistica. Un clamoroso esempio negativo in tal senso è rappresentato dalla catena di email, uscita nei mesi scorsi di pandemia, dove venivamo tutti invitati a partecipare ad una raccolta di dati relativi a manifestazioni del Covid, chiedendo di rispondere all'appello numerosi con la motovazione che « più siamo e più il campione diventa rappresentativo ». Basterebbe qui ricordare il caso di scuola di George Gallup che nel 1936 previde la vittoria di Franklin Delano Roosevelt alle elezioni presidenziali degli USA utilizzando un campione di 50,000 unità e battendo le previsioni effettuate dal Literary digest che ne aveva contattate 2,000,000. (Salvo poi ad incorrere in un errore banale 12 anni dopo nell'occasione delle elezioni presidenziali che vedevano Truman opposto a Dewey).
Ecco, ad esempio, rimarcare la differenza tra campione di convenienza e campione probabilistico, affermare con precisione che un campione deve essere un campione ben fatto e non necessariamente un campione grande per essere utile, questo è un compito che spetta solo a noi non ad altri! Il *data scientist* poco accorto analizza i dati senza ragionare sulla loro provenienza, senza interrogarsi sulla loro qualità. Ma è giusto così. Non è il suo mestiere. E' il nostro!

Passando al secondo aspetto della domanda, piuttosto che dire come evolverà il rapporto tra Statistica e Data Science preferisco dire come dovrebbe evolvere. Infatti, l'evoluzione futura di tale rapporto è ancora largamente nelle nostre mani e dipenderà da ciò che faremo.
Il processo di integrazione tra le due discipline, infatti, è ancora aperto. Quello che dobbiamo fare è governare il processo di transizione verso nuove figure senza farsene travolgere. Nella situazione attuale, infatti, corriamo il rischio essere sopraffatti in due modi. Per irrilevanza o per assorbimento. Da un lato possiamo rinchiuderci nella nostra riserva indiana a costudire

acceso il focolare delle nostre antiche tradizioni (altezzosamente chiusi nel nostro complesso di superiorità), ovvero possiamo farci travolgere e snaturare dal nuovo (depressi nel nostro complesso di inferiorità). La risposta dovrà essere, invece, quella di aprirci a questa grande sfida dando il nostro fondamentale contributo e senso critico senza perdere la nostra identità così che nella nuova figura che nascerà ci sia forte l'impronta del nostro DNA.
La Società di Statistica che immagino io è una Società che, con la sua autorevolezza scientifica e con umiltà, governi questo grande processo di transizione, non se ne faccia travolgere, che non si rinchiude in casa facendo finta che fuori non stia accadendo nulla e che si possa continuare a fare ricerca e ad insegnare la statistica come se nulla fosse. Che apra la porta e ne accolga i rischi, ma anche, con positività, le sfide. In questo processo, naturalmente, il gruppo SDS avrà un ruolo centrale.
Come questo possa realizzarsi in pratica lo dettaglierò meglio nella prossima risposta

Domanda 2. Quale strategia pensate possa essere utile per creare, a livello nazionale e internazionale, una solida connessione fra le comunità scientifiche che si occupano, da diverse angolazioni, di Data Science anche al fine di coordinare gruppi di ricerca?

Qui la risposta è semplice. La connessione tra comunità scientifiche, infatti, conosce un unico collante: **la qualità della ricerca scientifica**! Il resto sono solo scorciatoie che non portano da nessuna parte. Occorre (noi per primi) fare lo sforzo di pubblicare lavori di qualità su riviste di data science e ospitare noi articoli di data science sulle nostre riviste mostrandoci reciprocamente i vantaggi dei due approcci e favorendone le contaminazioni. Solo così verremo accolti senza pregiudizi nei consessi e nei momenti istituzionali preposti alla formazione di programmi didattici e di ricerca. In tal senso per essere operativi occorrerà selezionare alcune riviste internazionali e nazionali per non disperdere i nostri contributi in mille rivoli. In tale processo il gruppo SDS dovrà svolgere, evidentemente, un ruolo fondamentale. Tale elenco di riviste andrà possibilmente concordato anche con altre società statistiche internazionali che sentono lo stesso problema. Penso alla Royal Statistical Society, all'ISI a FenStat e ad altre società europee ed internazionali.
Questa strategia, ovviamente, apre un tema molto più ampio che è quello del riconoscimento delle riviste fuori raggruppamento nei processi di valutazione quali ASN e VQR.
Mi impegno, faccio una grande fatica per farmi pubblicare un lavoro su riviste di data science e poi il lavoro non mi viene riconosciuto!
Ed è questo un problema che travalica il rapporto col data science e riguarda molti altri campi di statistica applicata. Il principio secondo me dovrebbe essere che il 5% dei top journals di altre *subject categories* (data science, ma anche medicina, scienza della terra, ecc) dovrebbe essere presente nelle nostre fascie A. Se è vero che siamo *pervasivi*, (come spesso ci viene detto), allora vediamone di coglierne la positività.
Da questo punto di vista mi impegno in prima persona a presentare domanda come esperto del Gruppo di lavoro per la classificazione delle riviste ai fini dell'ASN (in base alla call uscita l'8 giugno) e invito tutti a farlo, pronto a fare un passo indietro se qualcun altro di noi risultasse selezionato. Ben sapendo che questo impegno unito a quello della presidenza cancellerebbe definitivamente ogni mia altra velleità di fare altro in quegli anni, ma anche con la coscienza che questo mi consentirebbe di lavorare operativamente su un tema importantissimo per la SIS e centrale al mio programma di presidenza.

Domanda 3. In che modo la SIS, i suoi gruppi, e in special modo il gruppo SDS, può porsi come interlocutore privilegiato in tutte quelle nuove iniziative per la formazione dei data scientists (lauree, lauree magistrali, masters) soprattutto quando non sono direttamente organizzate dagli statistici stessi?

Questa è la domanda più difficile delle tre, perchè qui la partita non si gioca solo sul campo della qualità scientifica come nel caso della ricerca.
Qui il lavoro di diffusione di una cultura statistica, di cui parlavo prima, porterà sì sperabilmente a dei risultati, ma solo con un'onda lunga per la quale occorreranno anni (a seguito di quanto ho detto nella risposta 2) quando potrebbe essere ormai troppo tardi per i suoi riflessi sulla definizione dei programmi dei corsi di studio.
Quando nel 2012 ho trascorso un semestre di insegnamento come visiting presso la New York University, partivano simultanemente due programmi : uno di Data Science, presso la facoltà di Ingegneria ed uno di Business Analitics presso la Leonard N. Stern Business School.
Anche da noi le cose stanno andando in maniera simile ed io suggerisco di perseguire una strategia mista nei confronti dei diversi corsi di studio a seconda che siano organizzati da noi statistici o da altri.

- Se siamo noi ad organizzarli, occorrerà essere molto vicini alla domanda di mercato individuando quelle figure che raramente sono di *data scientist* puro, ma che possono riguardare diversi campi quali l'economia, il business, ma anche la biostatistica e la salute. Qui dobbiamo essere propositivi mettendo dentro tutta la nostra fantasia e realismo. Ad esempio, presso la mia università, ho proposto un corso di "Metodi statistici e di data science per il decision-making in sanità". Allo stesso modo possiamo proporre e gestire corsi di Business analytics presso le business school dove avremmo ampi spazi o in altri campi ancora. Con questi corsi andremmo ad intercettare fette di mercato ancora non occupate da altri e che forniscono al mercato figure professionali che sono molto richieste.
- D'altro canto nel secondo caso nei corsi non gestiti da noi, e tipicamente gestiti da ingegneri informatici, occorrerà tentare di essere comunque presenti con corsi di metodologia, ma anche con corsi relativi a statistiche applicate (magari come corsi opzionali) i quali possono suggerire utili sbocchi occupazionali ai partecipanti a laurea e master marcatamente di tipo ingegneristico-informatico che poi possono incontrare scogli all'ingresso del mercato del lavoro, ad esempio presso le aziende, se si presentano con una formazione troppo orientata in forma esclusiva ai sistemi di dati,.

Il gruppo SDS sostenuto dalla SIS tutta ha l'autorità per svolgere un'attività di monitoraggio, controllo e indirizzo dei programmi dei CDS a livello nazionale individuando le situazioni dove è possibile/necessario intervenire.

Giuseppe Arbia

Roma, 11 giugno 2020