Report dalla Summer school CODATA-RDA “Research Data Science”

I dati della ricerca nel 21° Secolo

Antonella Zane

Università degli Studi di Padova

Dall'1 al 12 agosto ho partecipato a Trieste presso ***The Abdus Salam International Centre for Theoretical Physics*** (**ICTP**)[1] ad una scuola estiva sui metodi, strumenti e competenze nella scienza dei dati indispensabili per la ricerca del 21° Secolo.
Il corso, gratuito e idealmente orientato ai giovani ricercatori di tutto il mondo, è stato organizzato in collaborazione con **CODATA** (The Committee on Data for Science and Technology)[2], **RDA** (Research Data Alliance)[3] e **TWAS** (The World Academy of Science)[4]. Ai partecipanti provenienti dai Paesi in via di sviluppo è stato concesso un contributo per le spese di viaggio, vitto e alloggio grazie anche ad alcuni sponsor quali **Godan** (Global Open Data and Agriculture and Nutrition)[5] e **GEO** (Group on Earth Observation)[6]. Partner del progetto sono state le organizzazioni no-profit **Data Carpentry**[7] e **Software Carpentry**[8] che hanno l'obiettivo di fornire ai ricercatori competenze e abilità nell'elaborazione dei dati di ricerca (research computing).

Gli 80 partecipanti, selezionati sulla base del curriculum e di un questionario di ingresso tra 315 candidati, sono stati coinvolti per 11 intense giornate in lezioni teoriche, laboratori informatici *hands-on*, lavoro di gruppo e seminari serali facoltativi di *Author Carpentry*[9] tenuti dalla *resident librarian* del corso Gail Clement (CalTech University), per un totale di 93 ore.

I temi del corso sono stati affrontati da 15 docenti[10] provenienti da Paesi diversi con il supporto di una decina di tutor grazie ai quali è stato possibile gestire con successo i laboratori per un numero così consistente di participanti. Tutte le attività si sono svolte in inglese, in un clima cordiale e coinvolgente, di scambio continuo tra le persone.

L'obiettivo comune degli Enti organizzatori è quello di **promuovere la qualità, le politiche per la disponibilità e le competenze necessarie per il migliore uso dei dati della ricerca** che oggi, grazie alle nuove tecnologie, possono essere raccolti e riprodotti in modo più efficiente del passato:

> *Contemporary research – particularly when addressing the most significant, transdisciplinary research challenges – cannot be done effectively without a range of skills relating to data. This includes the principles and practice of Open Science and research data management and curation, the use of a range of data platforms and infrastructures, large scale analysis, statistics, visualisation and modelling techniques, software development and annotation and*

1 https://www.ictp.it/.
2 http://www.codata.org/.
3 https://rd-alliance.org/.
4 https://rd-alliance.org/.
5 http://www.godan.info/.
6 http://www.earthobservations.org/index.php.
7 http://www.datacarpentry.org/.
8 http://software-carpentry.org/.
9 http://libguides.caltech.edu/authorcarpentry.
10 http://indico.ictp.it/event/7658/speakers.

more. We define 'Research Data Science' as the ensemble of these skills.[11]

Le motivazioni

Uno degli aspetti su cui i relatori hanno richiamato più volte la nostra attenzione è stato quello della sfida posta dai *Big Data* cioè le grosse moli di dati, anche non strutturati ed eterogenei, che vengono sempre più di frequente prodotti come risultato di sistemi diversi, e della loro interrelazione, sempre più comuni nella vita di tutti i giorni (Internet of Things, GPS, social network, etc.)

Nell'ambito della ricerca, nuove competenze e nuovi profili professionali sono necessari per la gestione dei dati quali ad esempio il *data scientist* e il *data engineer.* Il primo crea modelli e continua ad aggiornarli mentre il secondo crea sistemi.

Ho partecipato a questo corso per un duplice motivo: da un lato, come ex ricercatrice, per vedere come vengono gestiti oggi i dati della ricerca e per conoscere gli strumenti attualmente disponibili per questa attività; dall'altro, come bibliotecaria, per capire le attuali necessità dei ricercatori e quindi dei servizi che il Sistema Bibliotecario di Ateneo potrebbe fornire a loro supporto.

Scopo di questo resoconto è fare conoscere questa interessante iniziativa, che verosimilmente avrà delle repliche nei prossimi anni e, possibilmente, suscitare interesse in chi potesse ravvedervi un'occasione di formazione avanzata, in uno stimolante contesto internazionale.

Il valore dei dati

Dati e metadati sono la base del patrimonio conoscitivo generato dall'attività di ricerca e possono essere essi stessi lo spunto per nuove ricerche. La loro conservazione e manutenzione nel tempo sono indispensabili per garantire il riuso e la riproducibilità dei risultati.

Good research needs good data. Un'intera giornata del corso è stata dedicata, con lezioni frontali e tanto lavoro di gruppo, al *Research Data Management (RDM),* l'attività di gestione e valutazione dei dati che si effettua durante tutto il periodo in cui questi rivestono un interesse scientifico, con l'obiettivo di ottenere il massimo dai dati della ricerca, anche ai fini della condivisione e del riuso. L'RDM comprende il *Data Management Plan*, documento che contiene informazioni dettagliate sui dati prodotti nell'ambito di progetti, sempre più richiesto dalle istituzioni che finanziano progetti di ricerca.

Il nuovo paradigma richiede che il ricercatore, tradizionalmente immerso nell'attività di ricerca che gli è consona, pianifichi con anticipo anche l'uso, la produzione e la gestione dei dati, non solo in funzione dei suoi obiettivi specifici ma anche considerando i dati *per sé.* I dati eventualmente prodotti da lui/lei vanno a confluire potenzialmente nel patrimonio condiviso dalla comunità scientifica, patrimonio che è alla base della d*ata-driven research.*

Nel nuovo paradigma, quindi, anche il bibliotecario ha il suo ruolo. Con le parole di Anelda van der Walt[12], del team di organizzatori della Summer School:

> "*Library Carpentry aims to teach librarians skills they will need to support their researchers in the 21st century – just as researchers are learning about new tools and methodologies, librarians also need new skills to work with bigger data, harness the power of the internet, etc."*

11 http://indico.ictp.it/event/7658.
12 https://www.linkedin.com/in/aneldavanderwalt.

Working the black seam

Per la gestione efficiente ed efficace di queste moli di dati è disponibile un vasto ecosistema di programmi e sistemi informatici che consentono maggiore efficienza (statistiche, elaborazione dei dati anche con calcolo parallelo e loro visualizzazione) e risparmio di tempo rispetto a quello che potrebbe essere l'approccio col tradizionale foglio di calcolo.
Di seguito presento una breve panoramica dei programmi, sistemi e concetti che abbiamo utilizzato, per ciascuno dei quali segnalo alcuni aspetti dell'esperienza svolta.

Unix shell

La Shell (*command line*) è il programma interattivo, a riga di comando, che consente di avviare altri programmi; al contempo incorpora nativamente un insieme proprio di istruzioni con le quali si possono creare nuovi programmi (*script*) anche molto complessi.

Per il corso abbiamo lavorato in ambiente GNU/Linux (*Linux Mint*) con la *Bash*, una delle shell più usate e come editor di testo abbiamo utilizzato *Nano*.
Ma perché usare la shell, e fare più fatica? Perché con un solo comando posso fare molte cose!
La riga di comando, infatti, consente di fare interagire tra loro diversi programmi, a loro volta molto potenti (*cat*, *curl*, *cut*, *find*, *grep*, *head*, *sort*, *uniq*, ...), concatenando il flusso di dati in entrata o in uscita da ciascuno (concetto di *pipe*), ad esempio per elencare i file e ordinarli per dimensione, verificarne la consistenza, verificare, ordinare e confrontare i contenuti di due file, concatenare più comandi in una sola riga e ottenere i risultati in pochi istanti.

Git

E' un sistema per il controllo delle versioni, in pratica un programma che aiuta a tenere sotto controllo lo sviluppo del proprio software e dei propri documenti. Anche Git è un programma a riga di comando (ma esistono interfacce grafiche).

Durante il corso abbiamo inizializzato un *repository*, familiarizzato con i comandi principali e simulato diverse situazioni che si possono presentare quando si utilizza Git da soli o in team con altre persone.

R/ RStudio (ggplot2, tmap, Shiny)

R è sia un linguaggio che un programma che fornisce un ampio ventaglio di tecniche statistiche e di visualizzazione grazie alla facilità di installazione di *packages* quali ad esempio *ggplot2* e *tmap* che consentono un'elevata personalizzazione dei grafici. R non produce solo grafici statici ma anche visualizzazioni interattive grazie alla sua applicazione web *Shiny*. R è una alternativa libera all'analogo programma proprietario *S*.

Per l'attività di laboratorio abbiamo scaricato il dataset pubblico *gapminder* da un mirror della rete CRAN[13] ed effettuato, con questi dati, alcune operazioni esemplificative di visualizzazione, anche avanzate, di tabelle e grafici.

13 Rete di server FTP e WEB che immagazzinano copie identiche di versioni di codice.

Visual design

Una giornata è stata dedicata al tema del *visual data analysis*, sia dal punto di vista teorico che sotto il profilo degli strumenti software indicati per l'elaborazione grafica di grandi moli di dati (es. *Tableau*, *VizQL*, *Protovis*, etc.).

Durante le esercitazioni pratiche abbiamo fatto esperienza con la metodologia *5-design-sheet* e disegnato manualmente diverse soluzioni grafiche per la rappresentazione dei dati secondo i modelli *What, Why, How.*

SQL e SQLite

SQL è il linguaggio per interagire con le basi di dati relazionali. *SQLite* è un DMBS leggero e semplice, di facile installazione, disponibile sui più diffusi sistemi operativi, molto utilizzato come back-end per altri programmi. Non è adatto per la manipolazione di grosse moli di dati ma va benissimo per applicazioni di piccole o medie dimensioni e per studio personale.

Durante il corso abbiamo operato su una base di dati creata con SQLite, sia tramite l'apposito plugin per Firefox, *SQLite Manager*, sia con comandi SQL diretti, per accedere e manipolare i dati.

Machine Learning e Recommender System

Machine Learning (ML) è un settore dell'intelligenza artificiale che si occupa di rendere le macchine “intelligenti” cioè capaci di apprendere in modo automatico grazie ad algoritmi che imparano sulla base dei dati che vengono forniti.
La differenza tra la programmazione tradizionale e ML consiste nel fatto che nel primo caso il programmatore definisce tutte le condizioni logiche e le reazioni predefinite del sistema al momento della progettazione del software, mentre nel secondo caso i computer “imparano da soli”, in parte, che cosa fare. Le tecniche utilizzate da ML sono le raccomandazioni (es. i suggerimenti forniti da *Netflix* e *Amazon* sulla base delle preferenze indicate dai clienti), il *clustering* (es. *Google news*) e la classificazione (es. i filtri per lo *spam*).

Durante il laboratorio abbiamo lavorato con i dataset di *MovieLens*[14], sito web che raccomanda ai propri utenti quali film guardare sulla base delle loro preferenze (*rating* e *tag*).

Internet of things

Si calcola che entro il 2020 ci saranno più di 25 miliardi di dispositivi connessi attraverso piattaforme dedicate (es. smartphone, smart TV) che porteranno sorprendenti sviluppi ad esempio nell'ambito dei domini della domotica e delle città intelligenti.

Data Science applications and use cases

Che cosa posso fare con i Big data? Vista la loro quantità posso aggregarli e fare analisi statistiche (es. *data warehouse*), *indexing*, *querying and searching*, *knowledge discovery* (*data mining*, *stats modeling*), *data-driven* (*predictive*, *deep learning*) *research*. La scienza dei dati si applica a tutti i dati, è ad esempio *data science* raccogliere informazioni sugli eventi criminali avvenuti in un quartiere in un determinato periodo di tempo al fine di calcolare le forze di polizia necessarie.

Negli USA si calcola che ci sarà presto bisogno di competenze in questo ambito e si stanno già cercando – per il 2018 – circa 190.000 analisti predittivi e 1,5 milioni di manager/analisti capaci di

14 https://movielens.org/.

prendere decisioni sulla base dell'analisi di queste grandi moli di dati.

Le reti neurali

Nella penultima giornata abbiamo affrontato il tema dell'analisi dei dati con il supporto delle *reti neurali,* con accenni ad alcuni degli approcci e delle tecniche più diffuse.

Nel laboratorio abbiamo utilizzato R con i package *kohonen* e *neuralnet.*

Landscape of Research Computing

Cosa posso fare se ho una conferenza tra una settimana e devo ancora elaborare una montagna di dati? Posso aumentare la capacità di calcolo aggiungendo *nodi* (un nodo è un elaboratore, che spesso è a sua volta una macchina virtuale), eventualmente presi in prestito da fornitori cloud.

Durante l'esercitazione pratica ci sono state fornite le credenziali per accedere alle risorse di *high throughput computing* dell'Open Science Grid[15] su cui abbiamo fatto esperienza con i seguenti applicativi: *Condor* per la sottomissione dei *job* di calcolo e *Dagman* per la gestione delle dipendenze. Infine abbiamo creato una Virtual Machine sull'infrastruttura *OpenStack*[16] del centro *Jetstream*[17].

Approfondimenti

In relativamente poco tempo, grazie all'entusiasmo, oltre che alla competenza ed esperienza di docenti e organizzatori, siamo riusciti, tra l'altro, ad acquisire gli elementi più significativi di una dozzina di strumenti/sistemi software (famiglie di) che possono essere impiegati in sinergia per la gestione, elaborazione e visualizzazione dei dati della ricerca.

Non mi è ovviamente possibile dare qui un resoconto esaustivo di quanto visto e trattato negli undici intensissimi giorni di corso, né è lo scopo di questo documento, resto quindi a disposizione per chi desiderasse ulteriori informazioni.

Il programma del corso e tutti i materiali didattici presentati sono disponibili all'indirizzo http://indico.ictp.it/event/7658/.

E' possibile organizzare un workshop di Data & Software Carpentry presso la propria istituzione:

- informazioni sulle persone coinvolte e la progettazione del workshop: http://software-carpentry.org/workshops/operations/;
- pubblicizzare il workshop presso i colleghi: http://software-carpentry.org/workshops/pitch/;
- modulo per compilare la richiesta: http://software-carpentry.org/workshops/request/.

15 https://www.opensciencegrid.org/.
16 https://www.openstack.org/.
17 http://jetstream-cloud.org/.

Letture consigliate dai docenti

- Bouiton, G., Babini, D., Hodson, S., Li, J., Marwala, T., Musoke, M. G. N., Uhlir, P. F. & Wyatt, S., *Open data in a big data world*: http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_short_en.pdf (short version) http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_long_en.pdf (extended version)

- Gray, D., Brown, S. & Macanufo, J. (2010), *Gamestorming: A Playbook for Innovators, Rule-breakers, and Changemakers* , O'Reilly Media, 290 p.

- Munzer, T. (2014), *Visualization Analysis and Design*, CRC Press, 428 p.

Strumenti / Tecnologie / Progetti segnalati

- colorbrewer2.org - Uno strumento online progettato per aiutare a scegliere buone combinazioni di colori per le mappe e altri elementi grafici.

- hadoop.apache.org – Framework per l'elaborazione distribuita di grosse moli di dati, come per esempio i dati non strutturati che vengono raccolti dalle reti sociali e da Internet of Things.

- www.kaggle.com (start up tecnologica che mette a disposizione dei data scientist un ambiente di lavoro e un set di servizi per la gestione dei dati).

- Lupi, G. (2015), *Sketching with Data Opens the Mind's Eye* http://news.nationalgeographic.com/2015/07/2015704-datapoints-sketching-data/.

- Lupi, G.& Posavec, S., *"Dear Data" Project:* http://www.dear-data.com/about/.

Ringraziamenti

Ringrazio le compagne di corso Elena Bertossi e Tanja Wissig per la lettura di questo testo e l'incoraggiamento.