Scuola Superiore di Catania

Corso di Laboratorio

"Big data: nuovi paradigmi per l'analisi dei dati"

a.a.2015-2016

Responsabile: prof. Giovanni Gallo

Motivazioni
La digitalizzazione delle comunicazioni e dei servizi, la disponibilità a costi sostenibili di estese reti di sensori per il monitoraggio di quasi ogni aspetto dei fenomeni fisici e sociali ha reso disponibili per lo studio e l'analisi scientifica nuove tipologie di dati che aprono interessanti prospettive ma che necessitano allo stesso tempo di nuovi strumenti di analisi. Negli ultimi anni si è anche assistito ad una rapida crescita della quantità di dati biologici, biomedici ed epidemiologici prodotti mediante le tecnologie "omics". Pertanto, la biologia moderna presenta ora nuove sfide in termini di gestione ed analisi dei dati. Le caratteristiche principali di queste nuove sorgenti di informazioni sono:
a) il formato digitale;
b) l'elevatissimo numero di record raccolti;
b) il superamento del concetto di "campionamento statistico" a favore della raccolta ed analisi sistematica e completa di "tutti" i dati;
c) ricerca euristica di correlazioni tra fenomeni per la costruzione di modelli predittivi.

Obiettivi
Il corso si pone come obiettivi:
a) fornire conoscenze di base nell'analisi "Big data";
b) familiarizzare con gli strumenti di calcolo tipici del settore;
c) far acquisire la capacità di utilizzare l'analisi "Big data" in modo critico e competente;
d) fornire degli spunti per una riflessione metodologica, epistemologica e delle conseguenze politiche e sociali di questo nuovo paradigma per l'analisi dei dati;
e) conoscere direttamente alcune delle esperienze pilota e delle applicazioni dei "Big data";
f) svolgere attività pratiche di laboratorio/progetto di analisi di "Big data".

Destinatari
Il corso è orientato agli allievi della Classe Scientifica impegnati nella formazione Magistrale (secondo livello). Esso potrà anche rivolgersi ad allievi di primo livello ben motivati e a allievi della Classe Umanistica, specialmente se interessati agli studi sociali.
La presenza di uditori tra i laureandi e dottorandi della Università di Catania che ne facciano richiesta sarà regolata dalle disposizioni della Scuola Superiore di Catania.

Pre-requisiti
Le nozioni propedeutiche per una proficua frequenza del corso includono:
a) Statistica elementare e esperienza, anche elementare, nella analisi dei dati;
b) Esperienza elemantare con dati in formato tabellare e conoscenza di base di linguaggi di tipo SQL;
c) Capacità di base nella programmazione con strumenti di alto livello (MATLAB, R, C#);

Articolazione del corso e valutazione
Il corso prevede trenta ore di didattica frontale e laboratorio affidato a docenti ed esperti provenienti in massima parte dalla Università di Catania ed articolati in moduli dettagliati di seguito.
A completamento della parte di formazione di base seguiranno interventi mirati di esperti internazionali invitati.
La valutazione di profitto sarà fatta valutando la qualità dei risultati progettuali ottenuti dagli allievi.

Struttura modulare del corso

1. Introduzione ai Big Data.

Il tema Big Data è molto vasto e le sue applicazioni spaziano dalla ricerca alla sicurezza o al business. In questo modulo diamo le definizioni e le coordinate principali del mondo dei Big Data; presentiamo alcuni esempi di applicazioni reali di Big Data; e diamo infine una visione generale di quali strategie seguire per attuare un progetto di Big Data con successo.
(4 ore) (docente: Luca Naso)

2. Algoritmi per i big data
Algoritmi per matrici di grandi dimensioni. Strutture dati per Big Data. Memoria Esterna e cache-obliviousness. Riduzione delle Dimensionalità. Tecniche Algoritmiche per l'analisi dei Big Data.
(6 ore) (docente: Giuseppe Nicosia)

3. Workshop di introduzione al software R
Il workshop si prefigge di familiarizzare gli studenti all'utilizzo della suite R per l'elaborazione statistica. In particolare verranno presentati, mediante esercitazioni pratiche, i "modi d'uso" più comuni con tale software nell'ambito della esplorazione ed elaborazione statistica dei dati.
(4 ore) (docente: Giovanni Gallo)

4. Workshop sugli strumenti tipici dei Big data
Una recente ricerca di O'Reilly (Oct 2014) ha ribadito che esiste una grande moltitudine di strumenti per chi voglia lavorare con i Big Data. Un pool di circa 800 data scientist ha nominato circa 300 strumenti diversi. La ricerca ha inoltre mostrato come i più pagati specialisti del settore utilizzino costantemente più di 20 strumenti.
Nella prima parte del modulo cerchiamo di razionalizzare questo mare di strumenti fornendo una tassonomia di base degli strumenti di Big Data. La restante (e più corposa) parte del modulo è dedicata all'utilizzo di un piccolo sotto-insieme di strumenti di Big Data (principalmente HDInsight, Hadoop, Hive, Excel) al fine di completare un piccolo progetto di Big Data. Il progetto consisterà nell'analisi degli accessi fatti ad un sito web. Qualora fossero disponibili, il progetto potrebbe studiare i dati del sito della SSC.
(10 ore) (docente: Luca Naso)

5. Casi di studio dei big data biologici per la biomedicina traslazionale e la sanità pubblica
Oggi i dati biologici si presentano in molte forme relative a vari livelli dei sistemi biologici (genoma, trascrittoma, epigenoma, proteoma, metaboloma) oltre ai dati epidemiologici, clinici e biomedici. Gli strumenti e le tecniche per l'analisi dei big data biologici permettono di tradurre la massiccia quantità di informazioni in una migliore comprensione dei meccanismi biomedici di base che possono essere ulteriormente applicati alla biomedicina traslazionale e alla sanità pubblica.
(2 ore) (docente: Antonella Agodi)

6. Casi di studio dei big data nel contesto delle tecnologie forensi
Rassegna di casi di studio in ambito legale e forense nei quali la metodologia dei Big Data offre potenzialità di nuove e più efficaci applicazioni. In particolare si esaminerà l'uso di dati

visuali provenienti da dispositivi di ripresa audio-video sempre più onnipresenti nella nostra società tecnologica. Verranno in particolare discussi casi di nell'ambito del riconoscimento di volti e della video sorveglianza evidenziando le prospettive e i limiti dei Big Data in tale contesto .
(2 ore) (docente: Sebastiano Battiato).

7. Casi di studio dei big data per lo studio dei sistemi complessi
(2 ore) (docente: Vincenzo Nicosia)

8. Contributi di esperti internazionali invitati: fino a tre conferenze di 2 ore, con esperti nel settore che illustreranno l'utilizzo di Big Data per la modellizzazione di dinamiche complesse di tipo sociale ed economico.