

## RELAZIONE SCIENTIFICA SUI RISULTATI DELL'ATTIVITA' DI RICERCA SVOLTA

**Titolo del progetto:** *Development of Big Data Analysis Methodologies for the consent-based data mining*

L'obiettivo del soggiorno di ricerca presso il centro di ricerca Vicomtech, San Sebastian, Spagna, è stato di avviare una collaborazione nell'ambito della cosiddetta *Big-data analysis*, concentrandosi in particolare sullo sviluppo di nuove metodologie di analisi, individuando dei concreti casi di applicazione che possano dar luogo a pubblicazioni su riviste internazionali e, ancora più importante, essere il punto di partenza per la preparazione di nuove proposte di progetti di natura interdisciplinare.

Sono stati individuati i seguenti ambiti di applicazione:

- 1) Analisi di grandi basi di dati di osservazione della Terra
- 2) Analisi di grandi basi di dati di immagini biomedicali, con particolare riferimento a immagini di risonanza magnetica funzionale

Per quanto riguarda la prima di applicazione, sono stati individuati due tipologie di dataset: a) dataset di immagini radar ad apertura sintetica (SAR) polarimetriche; b) immagini radar ad apertura sintetica (SAR) interferometriche. In entrambi i casi le immagini sono state acquisite da differenti sensori satellitari. Le due tipologie di dati sono state scelte perché esistono esempi di grandi database di immagini di questo tipo presso agenzie spaziali (come l'Agenzia Spaziale Europea o l'Agenzia Spaziale Italiana) e alcuni grandi Istituti di Ricerca. Dal dataset radar polarimetrico sono state estratte le mappe di anisotropia, entropia e angolo  $\alpha$  dalle quali poter inferire informazioni sulla copertura del suolo, la tipologia di vegetazione, l'umidità del suolo ed ottenere stime della biomassa disponibile. Dal dataset radar interferometrico sono state estratte le mappe di ampiezza e coerenza interferometrica dalle quali poter inferire tra l'altro, la distribuzione spaziale degli aggregati urbani e della rete infrastrutturale (per esempio, strade, ferrovie, porti, etc.). Individuato questi due esempio di set di dati comuni nell'ambito dell'osservazione della Terra, sono stati sviluppati dei codici in MapReduce per l'elaborazione di grandi moli di dati in genere non facilmente implementabile su una singola macchina. Per questo motivo è stata utilizzata una configurazione di *cloud computing*, usando MapReduce quale ambiente per l'organizzazione del flusso di calcolo sul cluster di macchine. Le funzionalità che sono state implementate hanno riguardato semplici analisi statistiche da utilizzare in ambito di visual analytics. L'impatto che tali semplici strumenti, applicati a grandi mole di dati il più delle volte già pubblici, è enorme e caratterizzato da un elevato grado di interdisciplinarietà. Gli ambiti di applicazione che sono stati investigati hanno riguardato la studio della biodiversità e del cambiamento climatico e quello delle scienze sociali. In entrambi i casi, i dati di osservazione della Terra forniscono misure di variabili ambientali e all'attività umana. I primi risultati di questa ricerca sono stati organizzati in una *table-of-content* di una bozza di articolo da inviare a una rivista internazionale con *referee* dal titolo provvisorio di "*MapReduce N-Dimensional Histograms over large datasets*".

Per quanto riguarda la seconda applicazione è stata effettuata una attività di disseminazione dei risultati da me ottenuti in collaborazione con il Politenico di Bari e l'Unità di Neurologia del Policlinico di Bari su dati fMRI al fine di favorire lo scambio di idee e l'avvio di una attività di ricerca comune. In particolare è stato effettuato seminari presso la Vicomtech al quale hanno partecipato, oltre a ricercatori della Vicomtech, anche ricercatori del Basque Center for Cognition, Brain and Language (<http://www.bcbl.eu/>) e della Fondazione Inbiomed per la Ricerca e la Medicina Rigenerativa (<http://www.inbiomed.org/>). Nell'ambito dei seminari sono stati presentati i risultati dell'analisi di dati di Risonanza Magnetica Funzionale al fine individuare pattern di

attivazione cerebrale in soggetti schizofrenici e confrontarli con i pattern di attivazione normalmente osservati in soggetti di controllo sani. L'enfasi del seminario è stata posta sugli aspetti computazionali (tecniche di Canonical Correlation Analysis, Multicanonical Correlation Analysis e classificazione dei pattern di attivazione di gruppi di soggetti schizofrenici e controlli sani). I risultati presentati nell'ambito del seminario erano stati descritti nell'articolo "Data driven analysis of functional brain networks in fMRI for schizophrenia investigation" accettato dalla rivista International Journal of Imaging Systems and Technology il 27 maggio 2014 e quindi non ancora di dominio pubblico alla data del seminario. A valle del seminario si è discusso delle modalità di estendere tale lavoro nella direzione dello sviluppo di tecniche di analisi di *big-data*. L'applicazione che è stata individuata riguarda l'uso di dati fMRI per studi relativi alla diagnosi dell'Alzheimer, nelle sue prime fasi di sviluppo.

A tale riguardo, si è proceduto con una attività di ricerca preliminare volta alla estrazione di *features* da immagini di Risonanza Magnetica, sia strutturale sia funzionale. È stato sperimentato l'uso della tecnica del Trace transform, un nuovo metodo globale per l'estrazione di *features* sviluppato da ricercatori del centro Vicomtech, per la *semantic context classification*, specialmente in caso di dataset per i quali non è possibile l'applicazione di metodi di estrazione delle *features* del tipo locale non essendoci nessuna conoscenza a priori. I primi risultati ottenuti durante il soggiorno di ricerca verranno ulteriormente analizzati e si studierà l'estensione della metodologia usando operatori differenti da quello di media lungo profili estratti dalle immagini sinora applicato dai ricercatori Vicomtech.

Bari, 3 giugno 2014

Giovanni NICO

A handwritten signature in blue ink, appearing to read "Giovanni Nico", written in a cursive style.