Geographic profiling (GF), in the context of our research, is defined as analyzing an area associated with a series of criminal events to determine/predict one or more of the following outcomes:

1. Residence or base of the offender(s)
2. Identification of the offender(s)
3. Location of future attack(s)
4. Type of future attack(s)

Most of the established methodologies (see CMAP white paper) employ distance based algorithms to statistically measure the target locations (also known as anchor points in GF parlance). Additionally, all of them work under the assumption that analysis data are locally maintained and the dataset is homogeneous. The assumption limits the efficiency and usability of the aforementioned methods used in traditional GF. On one hand, the restructuring process of crime handling tasks in national and regional level fragments the dataset and the individual components fall in the hands of different organizations with different operational scope. On the other hand, the lack of a centralized management of country-wide crime data forces a GF analyst to work with data under his/her jurisdiction only. Therefore, the analyst could miss out on information about potentially related events that happened to have occurred in other parts of the country.

Another critical problem facing the GF research is semantic mismatch of data contents. For instance, consider two organizations that collect crime scene data and store location information. The locations are then geo-coded for visualization applications and the coordinates are also stored along with original data. However, if the organizations use non-identical CRS (coordinate reference system) a GF analyst would be unable to efficiently investigate a crime spanning across both cities. The area of GF borrows ideas from other fields such as psychology and sociology to create more precise location-dependent crime detection mechanisms. This easily leads to an environment where researchers and analysts import terms and interpret them in their own way to fit a particular model or strategy. The issue of semantic mismatch is also referred to as the data heterogeneity problem. To advance GF techniques beyond localized algorithms, we propose methods to transparently integrate the available data sources from the federal and state agencies. Our methods will leverage the vast amount of data stored in these sources to provide a more robust, scalable and effective framework for the GF analysts to carry out their work.

Our integration methods are divided into two categories, namely, inter-domain and intra-domain, each of which is a two step process. Intra-domain level integration pertains to data aggregation of multiple sources, all of which belong to a specific domain. For instance, homicide datasets from different organizations will have much higher degree of similarity than with an identification theft dataset within the same organization since the formers share the same domain. Inter-domain integration refers to complex data fusion techniques whereby data sources from various domains are integrated. Once the first level of integration is performed, analysts can carry out analysis and produce results based on their work. The results from each aggregated component are then integrated in the second phase. The reason for subdividing the integration process is twofold. First, sometimes it is impossible to integrate all the data sources seamlessly because of organizational policies, bureaucratic protocols and other similar factors. Allowing modular integration, therefore, enables us to be flexible and sensitive about the need and environment of a geographic profiling task. Second, because each dataset contains enormous amount of information, integrating them all atomically will make the system highly unscalable.

The integration mechanism relies on the idea of a set of spatio-temporal ontologies. The ontologies define at various levels of granularity and scope, the objects, their attributes and the various states each object can be in related to a particular domain. The ontologies are organized hierarchically to allow for incremental evolution of conceptual details. The geographic contents of a particular criminal event would be instantiated from the ontologies that define contents at a conceptual level. For instance, in the statistical geographic profiling models (e.g., CrimeNet), the cells that fall out of the buffer zone are of particular interest and can be represented as rectangles with certain roles. As time passes and more data becomes available, the status of a cell can change from highly unlikely to very likely for potential anchor point. The temporal aspect of our ontologies captures the time essence in this kind of situation.