A Proposal for

Capacity Building for Big Data Science in Manipal University

Submitted by

Dr. M. Devi Prasad
Professor, School of Information Sciences
Manipal University

Table of Contents

Acronyms, Terms and Definitions

Contents

Acronyms, Terms and Definitions

Micro-workshop

Professionals and resourceful people who are busy in their day jobs can spare only a few hours away from their offices. A micro-workshop is specifically tailored to last one or two hours; and on rare and exceptional cases, may be three hours, but never more. The purpose is to either develop a deeper understanding of a subject-matter of interest or solve a specific problem. People come together and engage in intensive brainstorming sessions.

EHR – Electronic Health Records

EMR – Electronic Medical Records

Abstract

The field of *big data science* is all about discovering insights from huge data sets. Almost every human institution has witnessed the benefits of predictive and prescriptive insights drawn chiefly from historical data. Worldwide, big data is applied in different spheres of businesses, governments, science, education, and healthcare. The insights drawn using the methods of data science help in deriving better solutions in a timely manner.

Numerous departments, educational institutions, research units, and hospitals functioning in Manipal University generate, transform and consume large collection of data. It is necessary to note that these data sets are managed by different organizational units with unique privacy and security requirements. This is natural in a large, multi-disciplinary university like ours. In some cases, these data sets may be (directly or indirectly) related, and yet controlled by different entities with distinct objectives. This situation poses a challenge if one desires to draw integrated insights from the seemingly disparate, and yet, semantically related data. This is a classic big data scenario emerging from within Manipal University.

In this proposal, we put forward a three-pronged approach to address the big data requirements of Manipal University. The first effort is to build capacity for dealing with big data. The second task is to demonstrate our competency by applying big data science in the context of community healthcare. The third and the most important goal is to engineer a next generation information architecture that non-intrusively integrates disparate data sets, while taking care of privacy, security, scalability and ease of use requirements.

We propose a detailed project plan spanning four calendar years. The proposed plan identifies many details concerned with building capacity, building competency, and engineering the next generation information architecture. The proposal outlines a model for carrying out interdisciplinary research based on big data methods. To the best of our knowledge this represents first of its kind effort in any university in India.

We believe this important effort will contribute to the IP portfolio of Manipal University. In addition, we propose to host national and international conferences, publish results in peer-reviewed conferences and journals, and invite experts from academia and industries for intellectual exchange and to learn best practices. Apart from meeting the three main objectives, we will create a Web-based outreach program. This is intended to showcase Manipal University's Systems Engineering capabilities in developing multi-disciplinary big data systems.

1. Background - Big Data, Data Science and Visualization

Information technologies are enabling researchers to store, process and share an unprecedented volume of data. Big data technologies are transforming the manner in which data is integrated from a variety of sources. Advances in *data science* have led to computational techniques for analyzing large sets of data, and for automatically constructing statistical models from the underlying data set, all without human intervention.

Big Data systems are characterized by volume, velocity, and variety of data handled in a system. Big data problems pose complex challenges to the state of the art techniques and paradigms of coping with data. Different types of data, their acquisition, integration, storage, modeling, processing, analysis and visualization puts extreme pressure on existing methods and resources. The resources may be computational elements, institutional processes, and even human resources.

Many academic institutions and research centers in Manipal University generate as well as consume significant amount of data. Kasturba medical college hospital alone, for instance, generates exceedingly large amounts of multimodal patient-data everyday (in the form of text, images, and clinical data readings). Given the numbers of patients visiting the hospital each day, and the pace with which EHRs are populated with clinical, imaging, and genome data, we are undoubtedly witnessing a classical Big Data system emerging from within. The center for life sciences too generates and processes terabytes of genome data on a routine basis. The community healthcare centers functioning in and around Udupi have collected data of about fifty-thousand community members in the region (at the time of writing this document). It must be noted that this data is expected to grow in the coming years, and thus, is another source of big data.

There are also many non-healthcare data sources in Manipal. The admission and student information systems in the university, for example, maintain data spanning many decades, pertaining to tens of thousands of students. The IT infrastructure of the entire university is another source of big data. The data center and the computing infrastructure generate hundreds of megabytes of text logs every day. These logs may be inspected for security breaches, intrusion attempts, system failures and many such conditions. These are representative samples of reasonably large data pools in Manipal University.

Over the past few years, along with the data science revolution, *data visualization* has grown immensely in importance as a field in itself. Data visualization is concerned with rendering data for more effective visual inspection, direct manipulation, exploration, communication and analysis. Research and practice in this area has identified a collection

of methods that enhance the descriptive power of visual data. Visualization is preferred when decision makers wish to gain novel insights or recognize deeper patterns in the data (rather than from mathematical models alone) before drawing conclusions or making choices.

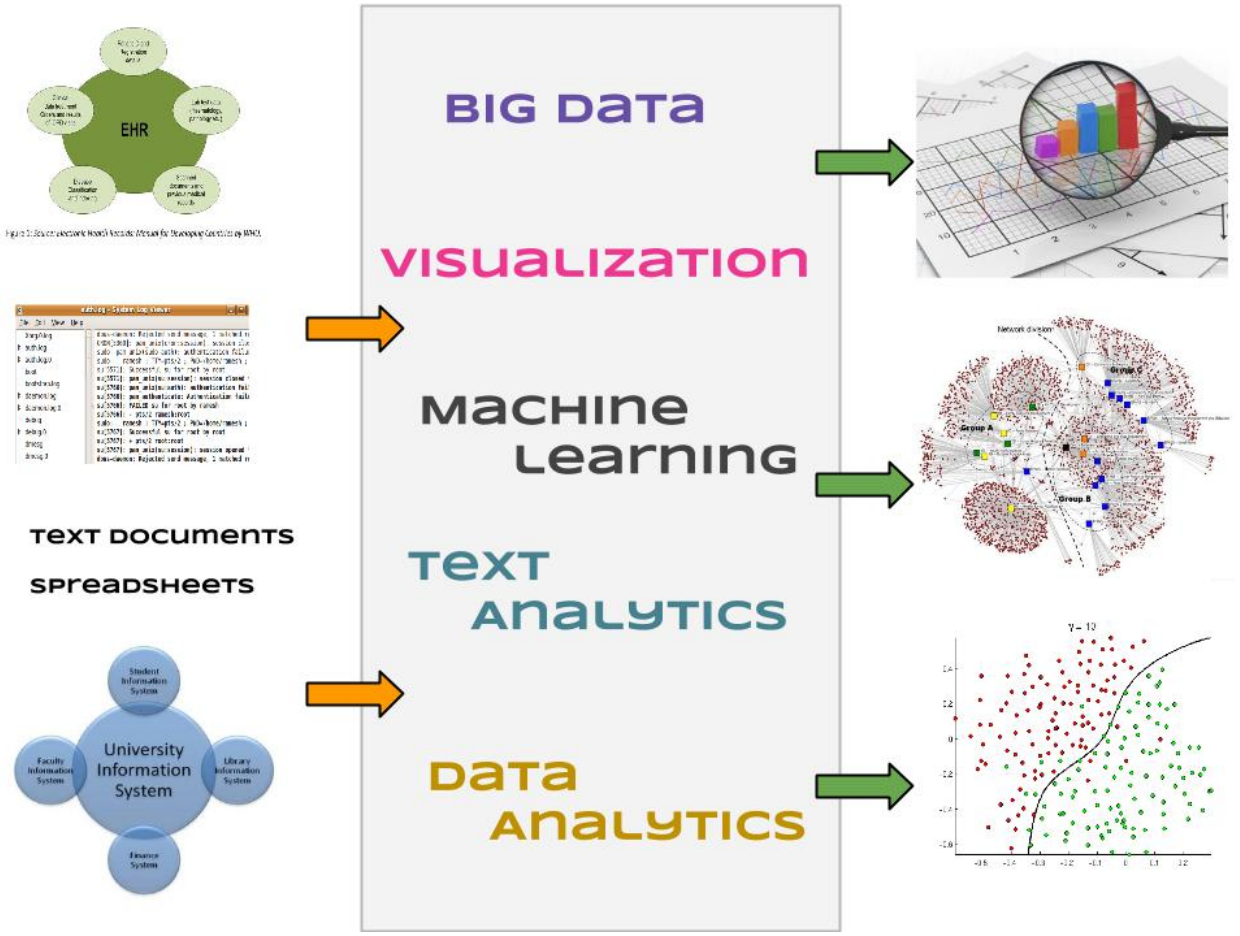

Figure 1 – Big data sources in Manipal University, analytics, and visualization.

There is great value in bracing up to the big data challenge. Data analytics and visualization make it possible to discern unnoticeable and discrete patterns from large data sets. Once an emergent property is made explicit, specialists can direct computational resources to dig deep and draw valuable insights. Insights, in turn, guide human attention in creating value. Insights help focusing on novel pathways for achieving superior solutions.

The picture we are painting is illustrated in Figure 1, above. On the left side of this picture we mention some of the prominent data sources in Manipal University. The right side represents results that may be obtained by data analytics and visualization.

Data science is a complex discipline. Figure 2, below, names some of the fields that constitute data science. A data scientist needs to have a very good grasp of statistics and mathematics. Data analysis is a human activity requiring an aptitude and taste for scientific inquiry and exploration. Although one could learn the underlying principles, it takes practice and experience to develop right intuitions for data analysis.

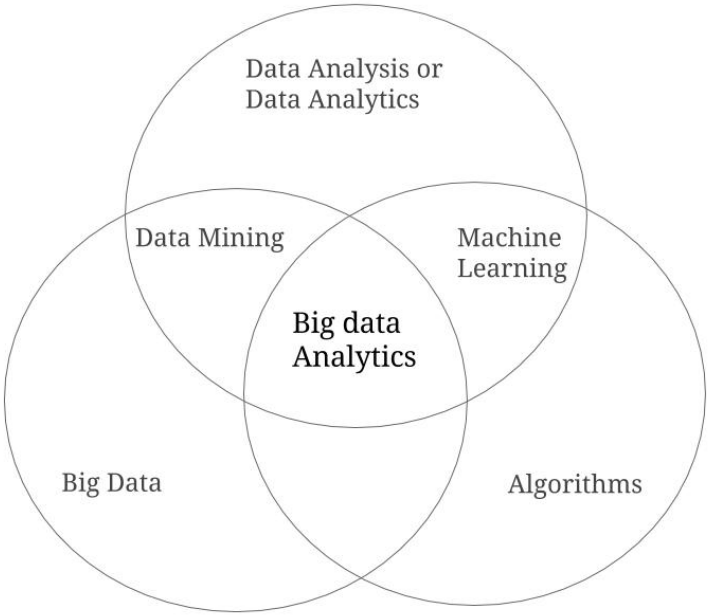

Figure 2 - The fields of data science.

Machine learning, on the other hand, utilizes algorithms to drive data analysis. In machine learning, computers (and not human experts) perform large scale analysis of the data.

Big data technology is also a fairly complex field. Good engineering skills are required to build big data solutions. One needs to prioritize privacy and security - more so in healthcare. It is essential for practitioners to have a firm grasp of the existing workflows, standards and organizational processes. One also needs to be aware of the regulatory aspects, and take into account ethical requirements.

2. The Research Challenge and Opportunities

In this proposal, we detail an exploratory inquiry into the application of data science to fields that are of importance to Manipal institutions. We outline an interdisciplinary research with the primary goal of building capacity for big data science and for attaining competency in engineering systems around big data. We highlight the necessity of building a software platform that enables big data collection, integration, processing, analysis, and sharing of scientific results across the entire research community in Manipal, while being

meticulous about privacy and security at all levels. We recognize that many projects in themselves generate large data sets, and thus, are key sources of big data. Therefore, it is particularly important for us to account for data generation and acquisition in the proposed software platform.

As outlined in the previous section, the skills required to frame data analysis questions in suitable mathematical terms, and to use right techniques demands experience, knowledge, and practice. This could be gained only by working in a team of inspired people with right knowledge, skills and attitudes. It is therefore exceedingly important to constitute a great team of engineers, mathematicians, researchers from Life Sciences, doctors, field workers, and administrators.

Data visualization is also an important tool in the workflow of analysts. This is a rich field with numerous opportunities for innovation. We would like to explore the possibility of designing novel visualization methods for healthcare, especially for exploring genome data of phenotype cohorts. We also intend to develop libraries that can be reused across research projects in Manipal and perhaps released in public domain for general use (in open source projects).

Good engineering skills are an absolute requirement for designing and developing a software platform that integrates multiple data sources of Manipal institutions. Some of the important requirements of the proposed software platform include flexible security architecture, extensible privacy schemes (for instance, static consent versus dynamic consent), virtualization support, distributed processing, access control mechanisms, to name a few. Since this platform will be built over the foundations of open source big data technologies (chiefly from Apache Hadoop ecosystem), it is necessary to build a team of engineers with right competencies.

In the following section, we will state and elaborate the objectives of this project. As a part of this discourse, we identify a few sources of reasonably large-impact data sets in Manipal institutions. Our intent is to develop skills and competency required to harness these data sets in meaningful ways. In this endeavor, we will work with researchers from life sciences and medical Sciences to design solutions that are useful in many ways and are valuable even in the long run.

We recognize that there are many challenges to be tackled. Let us consider, for instance, the existing data sources in Manipal healthcare domain. This includes data silos emerging from the hospital, independently functioning community-centric units, and research centers. The data sources are disparate and heterogeneous. These are managed by different organizational entities with varying security and privacy requirements. We, therefore,

argue that there is an immediate necessity for an information architecture that enables secure and distributed data access. We propose an architecture, and outline the basic design of a software platform that satisfies the requirements. We discuss the engineering challenges in building such a secure and distributed platform.

To the best of our knowledge, no such attempt has been made, so far, in any Indian university. Therefore, we argue that this would be a great exercise in systems engineering. Success in building this system will greatly impact the way big data is leveraged in healthcare research in Manipal, and in our nation, in general.

Privacy and security of healthcare data is a major challenge. Flexibility and security are conflicting requirements. We need to balance privacy and security requirements vis-a-vis the desire to have extensible and flexible policies for accessing data. Designing sound protocols and rigorously proving that these protocols are mathematically correct is another interesting area. We will employ formal verification methods to prove the correctness of our models and the resulting implementation.

Given the interdisciplinary nature of problems addressed in this proposal, we expect greater challenges in building a platform that serves generations of researchers. All kinds of resources - people, teams, knowledge, talent, skills, hardware and software - need to be managed. The following sections describe to some extent our vision in this direction.

3. The Research Objectives

We have three primary objectives for carrying out the research and development described in this proposal:

A. Capacity Building for Big Data Science and Visualization.

We wish to enhance the capacity of existing systems to meet the basic requirements of big data. We mainly intend to scale up the computational capacity and software subsystems to handle the data of past six years. We will leverage the Hadoop ecosystem to equip our system with improved forms of data acquisition and storage (streams, batch data, and real-time data), as well as higher data analysis capacity. This is expected to enhance our competency too, especially in data analytics, predictive modeling, and visualization. In the long run, this will make it possible for Manipal University to develop novel healthcare products and services.

B. Applying Big Data Science in the Healthcare Context.

We will generate proof of concept phenotype-specific genomic data for a specific trait or condition considering unilateral trait transmission from the community and

incorporate into the big data analytics to derive predictive models. This is expected to create innovative community-health solutions by generating statistical models with the application of machine learning and data visualization techniques.

C. Designing a Big Data Information Architecture for Manipal University.

We will develop a secure and scalable next generation information architecture for integrating, managing and provisioning Manipal's big data infrastructure. We will develop tools to integrate and leverage open source big data for trend analysis and customize healthcare products and services. Further, we will harness virtualization technology, and implement powerful security and privacy protocols.

In the following sections, we will elaborate the motivations behind each objective, state the value proposition of each objective, and outline a few key steps required to reach the goal.

4. Capacity Building for Big Data Science and Visualization

Being part of a multidisciplinary university, we enjoy a unique advantage that only a few universities in the world could imagine. In our view, it is important for us to build capacity for data sciences while working with large sets of historical data. This offers an invaluable opportunity to apply data analysis and machine learning techniques on real data. This also serves as an excellent means to build our competency in data sciences.

In order to be able to leverage big data, we need to set up a capable ecosystem. As noted earlier, Manipal has large pools of legacy data. We wish to utilize data from the past six years, at the least, for volume and variety. Therefore, we expect large differences in their schema, storage formats and access strategies. We also expect a large data set to be unstructured or semi-structured. This happens when data resides in text documents or spreadsheets with arbitrary data in their cells.

The first step is to reconcile the differences in electronically available historical data. This has to be achieved in an economical and noninvasive way. This problem has a combined complexity of data integration and classical software engineering. We will develop software adaptors and bridges to modularly integrate data from disparate sources.

There is another challenge. When we electronically import data from papers, we completely lose the implicit structure of the data. This is a problem to be solved using state of the art text processing technologies. Subsequently, text analysis and machine learning techniques could be used to extract structured information for further analysis and visualization.

We will develop an array of visualizations. The goal is to understand visualization for better communication and exploration. We will work closely with the stakeholders (corporate marketing team or doctors or researchers in life sciences) to understand its requirements and help glean insight from visualizations. This effort will undoubtedly enhance our competency in applying data visualization in real problem settings and situations.

The major functional layers and elements of the proposed system are shown below in Figure 3, below. Distributed connectors make it possible to access data stores to be accessed from geographically different locations. The layer just above these connectors represents a collection of abstract (software) adaptors that translate and/or transform from one schema to another. Also, this layer has components that allow version controlling some portion of the base data or data generated during analysis phases.

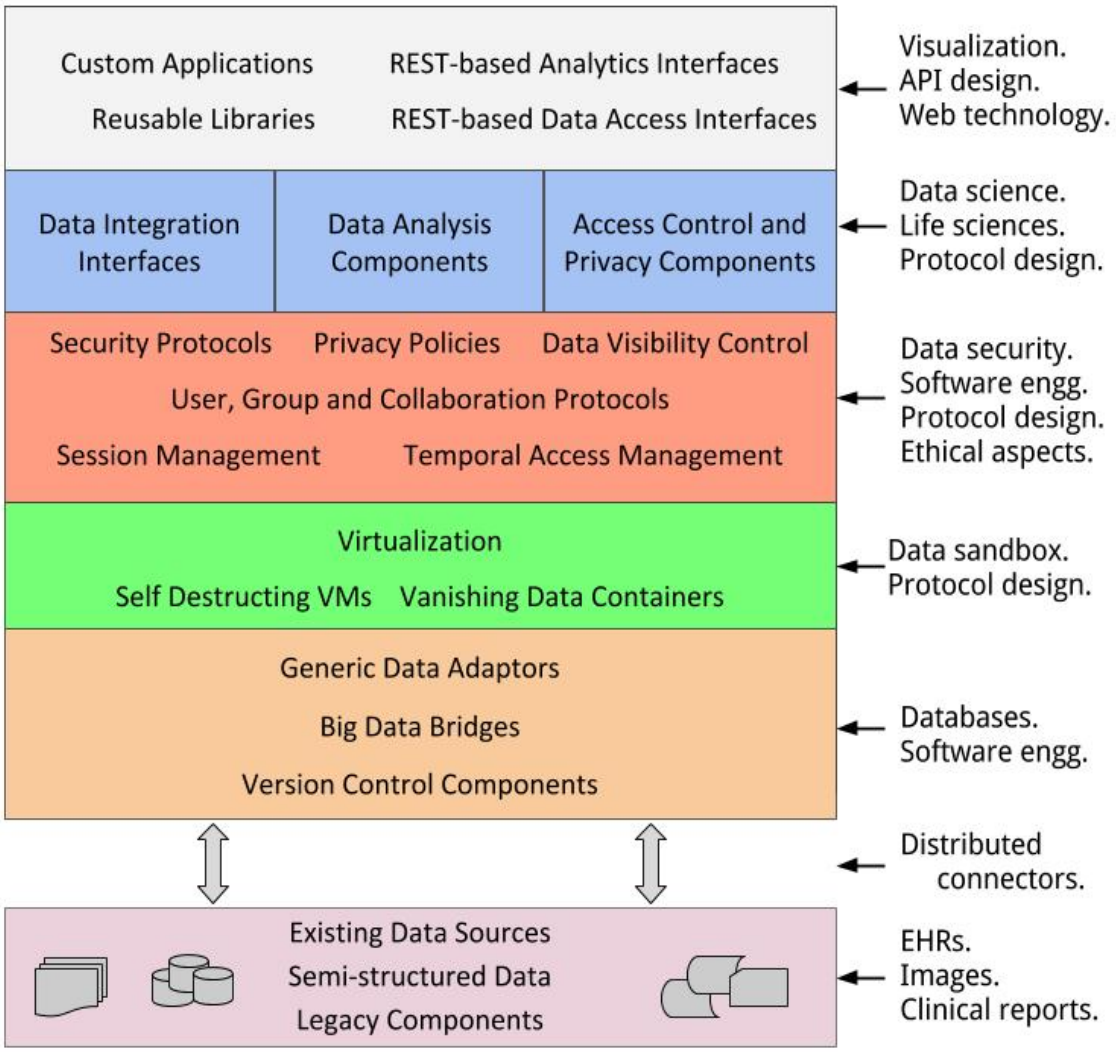

Figure 3 - The software layers of the proposed big data platform.

The privacy and security layer is the lynchpin of this platform. It holds various pieces together in place and is crucial to the success of making data analytics work for healthcare. We wish to use state of the art software engineering principles and best practices to define, validate and enforce privacy and security protocols. We will build formal (mathematical) models, verify their correctness using theorem provers and model checkers, and follow rigorous software development processes to ensure that the implementation is consistent with the automatically verified models. This effort is expected to create invaluable IP for Manipal University.

The next layer represents another major contributor to our IP portfolio. This layer includes various solutions created for data analysis and visualization. Since our approach includes data analysis as well as text analysis, we expect this to be a repertoire of novel solutions packaged as reusable libraries and APIs.

The topmost layer represents the external interfaces to platform services. We envision these to be a collection of interfaces that allow applications to use platform services over the Web. In other words, these are the Application Programming Interfaces (APIs) for the world. In addition, this layer hosts software applications contributed in this research.

5. Applying Big Data Science in the Healthcare Context

We propose to carry out a specific interdisciplinary inquiry working closely with the researchers and practitioners in Community Medicine and Life Sciences. We will be employing data analytics to iteratively analyze many tens of thousands of community health records to identify phenotypes expressing certain characteristics. Our aim is to construct predictive models to identify the high-risk members of the community who may be diagnosed for certain type of diseases, for instance, diabetes and cancer.

We propose to design and execute machine learning techniques multiple times over the data set, gradually expanding the size of the data set. This method, in particular, helps successive iterations to refine models built in the previous steps. The result of this will be fine-tuned models of specific phenotypes.

We will work with the researchers from Community Medicine and Life Sciences to identify phenotype cohorts from the patterns drawn from the application of machine learning. Eventually, we will evaluate the effectiveness of this line of inquiry by genomic analysis of the identified cohorts.

A. Since altered phenotypes are one of the most reliable manifestations of altered gene functions, we will base our initial inquiry around the forty five thousand community health records. Our first goal will be to identify possible phenotypes.

B. We will first apply unsupervised machine learning techniques to identify phenotypes from a subset of the records (may be around twenty thousand records out of the forty five thousand available records). This attempt may yield one or more models that provide some insights by relating one or more health parameters.

C. We will gradually improve these models by letting machine learning algorithms process larger and larger sets of data.

D. In the next step, we will determine phenotype cohorts. Once again, statistical methods and machine learning techniques will be invaluable for this inquiry.

E. If Electronic Health Records (EHRs) for these cohorts is available in KMC, Manipal, we will use that to supplement the findings. We will employ text analytics and machine learning to draw insights from the clinical text contained in the EHRs.

F. Life Sciences Research Center will extract genome data of a few members of the identified cohorts. If concrete genetic causes are identified, this data may be used to provide better advice and patient care for the concerned.

6. Designing a Big Data Information Architecture for Manipal University

Our objective is to reimagine how Manipal's vast data repository can be noninvasively restructured and reorganized in order to offer sophisticated online services for safe access. These services must be engineered from ground up for privacy protection. Supporting stronger privacy and user control involves a serious redesign of key protocols and architectures reinforcing Manipal's ethical standards and practices. This necessitates the development of principled definitions of privacy, matching the needs of the next generation of researchers, medical practitioners, patients and users. In addition, we will need software tools and formal verification methods to evaluate the degree of protection actually afforded.

We will consider two categories of online services:

A. This category consists of systems that process personal information of individual community members and/or patients, but generate information that can be publicly disclosed. A typical example is aggregating the data derived or drawn from the individuals of a community, and extracting more meaningful or semantically richer information and statistics. For example, the likelihood of a specific health condition in a particular section of the community, or the correlation between a health

condition and lifestyle. For such applications, the objective of this project is to devise generic and yet rigorous techniques that minimize the exposure of personal information.

B. The second category of services process personal data, generate additional insights and information, and augment the base data set. For applications of these kinds, we aim to devise architectures that not only minimize the exposure of private information but also enable the specification of flexible access control policies to unambiguously determine under what conditions the resulting sensitive information is shared, with whom, and for what duration of time it shared.

A fundamental objective, common to both categories, is to develop principled and robust definitions of privacy, as well as methods for evaluating the quality of protection offered by many of these mechanisms. We plan to explore and combine variants of obfuscation techniques, cryptography, differential privacy, vanishing or self-destructing data, virtualization and such.

In order to achieve flexibility and extensibility, we need to design methods for role-based visibility and access, specifying how users may interact with data and data services. The goal is to enable administrators to specify new security norms, new privacy protocols, without affecting other parts of the system. Such non-intrusive modifications can be enabled only by a systems that allows declarative specification of requirements. We wish to design expressive domain specific languages (DSLs) for specifying privacy and security requirements.

In order to give strong guarantees about the underlying models, protocols, and methods, we intend to provide rigorous proofs of correctness with the aid of formal methods such as theorem provers, model checkers and verification tools. The need for formal mathematical methods is evident: in the architecture that we propose in the next section, we plan to sandbox many privacy and security aspects within dynamically created virtual machines (VMs) and containers. Activities from within these isolated VMs may access shared data services. Thus, the resulting system will be a concurrent reactive system. In order reason about the behaviors of such systems we need to resort to model checkers and other verification methods.

We intend to build prototypes of these services and evaluate their effectiveness in achieving privacy-friendly goals. We will build libraries, frameworks, define API and develop proof of concept programs to demonstrate the effectiveness of the proposed solutions.