Data Science Bundle (alpha version!)

Data Science Bundle (alpha version!)

Combining clinical and genomic data using i2b2 and tranSMART to perform complex analyses of real-world data

Introduction

This data science bundle supports complex analyses of real-world clinical and genomic data. It includes **i2b2**, which enables query and cohort identification, and **tranSMART**, adds a suite of tools for data exploration, R-based advanced analytics (e.g., correlation analysis, heat maps, PCA, etc.), and genomic modules for Genome Wide Association Studies (GWAS) and high dimensional data analysis such as RNAseq.

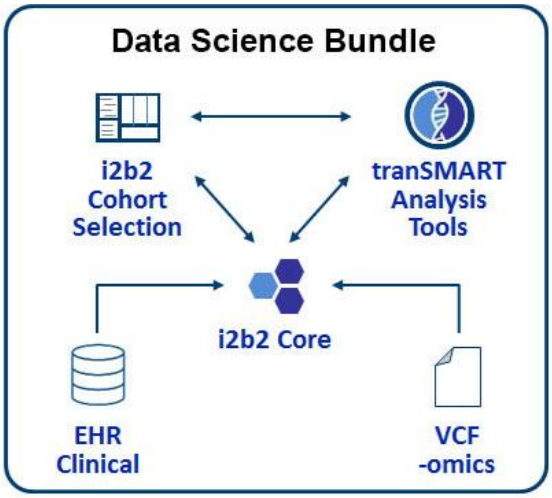

Figure 1. High-level view of the bundle. The i2b2 common data model integrates clinical and genomic data. i2b2 provides tools for query and cohort selection, and tranSMART contains modules for high-dimensional analyses.

Use Cases

Uses cases for this bundle include:

- Translational research
- Genome association studies

Bundle Components

This bundle includes documentation on how to install and configure the following items:

- i2b2 - Local query tool
  - Database
  - Application Layer
  - i2b2 Web Client
  - Sample synthetic data
- tranSMART - Analysis tools
  - Additional database tables
  - Application Layer
  - tranSMART User Interface
- Demo data - Synthetic datasets for testing the software

Demo

A public demo of this bundle is available at the following URL:
http://shrine-node3.i2b2transmartplugins.org/

It consists of both i2b2 and tranSMART running on the same database with Synthea demo data.

Technical Architecture

i2b2 Components

i2b2 consists of independent applications that provide different functionality called "cells" (Figure 1). A collection of cells form an i2b2 "hive". Most i2b2 hives include (1) a Project Management (PM) cell for authentication and authorization; (2) a Clinical Research Chart (CRC) cell, which contains patient data and the query engine; and, (3) an Ontology (ONT) cell, which describes the concepts and codes contained within the CRC cell. Many i2b2 hives also include (4) a Workplace (Work) cell, which enables users to "bookmark" frequently used items in the user interface and share these with collaborators; and (5) an Identity Management (IM), which allows authorized users to retrieve identified patient data. Cells communicate with each other using i2b2 XML messages sent to APIs. When a cell receives a request message, it queries a table in the HiveData database to determine the location of main database for that cell, based on the user's project. An exception is the PM cell, which uses a single database for all projects. The i2b2 Web Client is written entirely in HTML and JavaScript. It communicates with a Web Proxy on a web server, which redirects messages to the appropriate cell.

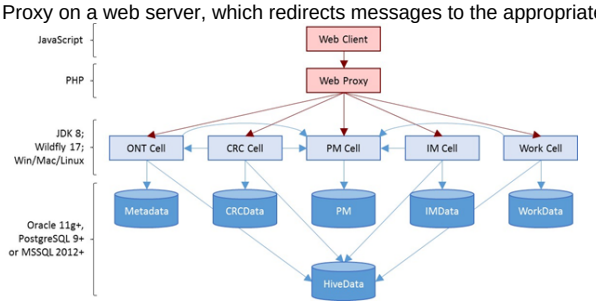

Figure 2. i2b2 components.

tranSMART Components

The TranSMART web user interface is a single tomcat application with an extended set of plugins which may be enabled/disabled in the configuration file.
TranSMART is delivered with a set of supporting applications:

- Transmart-data creates an empty database including all schemas and tables for i2b2, and includes stored procedures for data loading and data management. It also provides targets to install R and saset of required R/BioConductor packages from source.
- Transmart-etl provides Pentaho dataintegration (Kettle) jobs to load clinical and omics datatypes, plus a loader application for genes, pathways, proteins and other metadata.
- RInterface provides an API for scripts to login and extract study data for external analysis
- Transmart-batch and tranSMART-ICE are alternative data loading tools
- Transmart-manual is the online manual to be installed alongside the server using tomcat or a web server
- Scripts provides all-in-one installation scripts for supported operating systems
- GWAVA provides a server for GWAS data visualisation within tranSMART
- Transmart-test is an automated test environment for developers

System Requirements

i2b2

i2b2 requirements can be found here. A summary of the key requirements:

- Database: Oracle (>=11g), PostgreSQL (>=9), MSSQL (>=2012).
- OS: Windows, Mac, or Linux

- Software components: JDK 8, Wildfly 17, web server (no specific requirement)

tranSMART

TranSMART is supported with:

- Oracle Enterprise Edition 12.0.1, PostgreSQL (>=9.4)
- OS: Linux (Ubuntu 20.04, 18.04, 16.04, 14.04, Centos 7, Fedora 33, …)
- Software components: JDK 8, Tomcat (>=7), Groovy, R (>=4.0.0),

The additional database requirements (e.g. Oracle Enterprise) ar eto support partitioning for the large ables used for omics data in the tranSMART-specific schemas.

Installation

i2b2 Install

- Download:
  - https://www.i2b2.org/software/index.html
  - Binary distribution and quick install guide, under "download binary distribution"
  - Or, download the source code from the same page.
- Follow the quick install guide (on https://www.i2b2.org/software/index.html) or the detailed install guide https://community.i2b2.org/wiki/display/getstarted/i2b2+Installation+Guide). There are 3 components:
  - *Data (Chapter 3)*. Install the i2b2 database on MSSQL, Oracle, or Postgres. This provides many metadata tables for querying and authentication, as well as the actual core data tables.
    - It is essential to configure i2b2 with the ACT ontology, to be compatible with the current SHRINE 3.0 bundle. When installing i2b2 data, follow the instructions in the next section on installing the ACT ontology.
  - *Server (Chapter 4)*. A Java program that runs in the Wildfly container which provides an API and data analytic methods on the database. It is divided into components called cells. SHRINE uses some of these cells: CRC to communicate to the database, ONT to provide the query ontology, and PM to manage authentication.
  - *Webclient (Chapter 5).* A web interface to i2b2, which is not required for SHRINE but could be useful for local querying and testing (SHRINE is network-only).

tranSMART Install

Installation instructions are on the transmart wiki. They can be used generally on any Linux operating system.

Install scripts are provided to install on a set of supported operating systems. They are provided for a fresh clean installation of the operating system and can be amended and re-launched in case of problems (e.g. files not in the expected path/format, or changs to the components/requirements for R installation from the public R distributions)