

A Data Science tool for Data Reviewing

Giuseppe Di Monaco, UCB BioSciences GmbH, Monheim, Germany

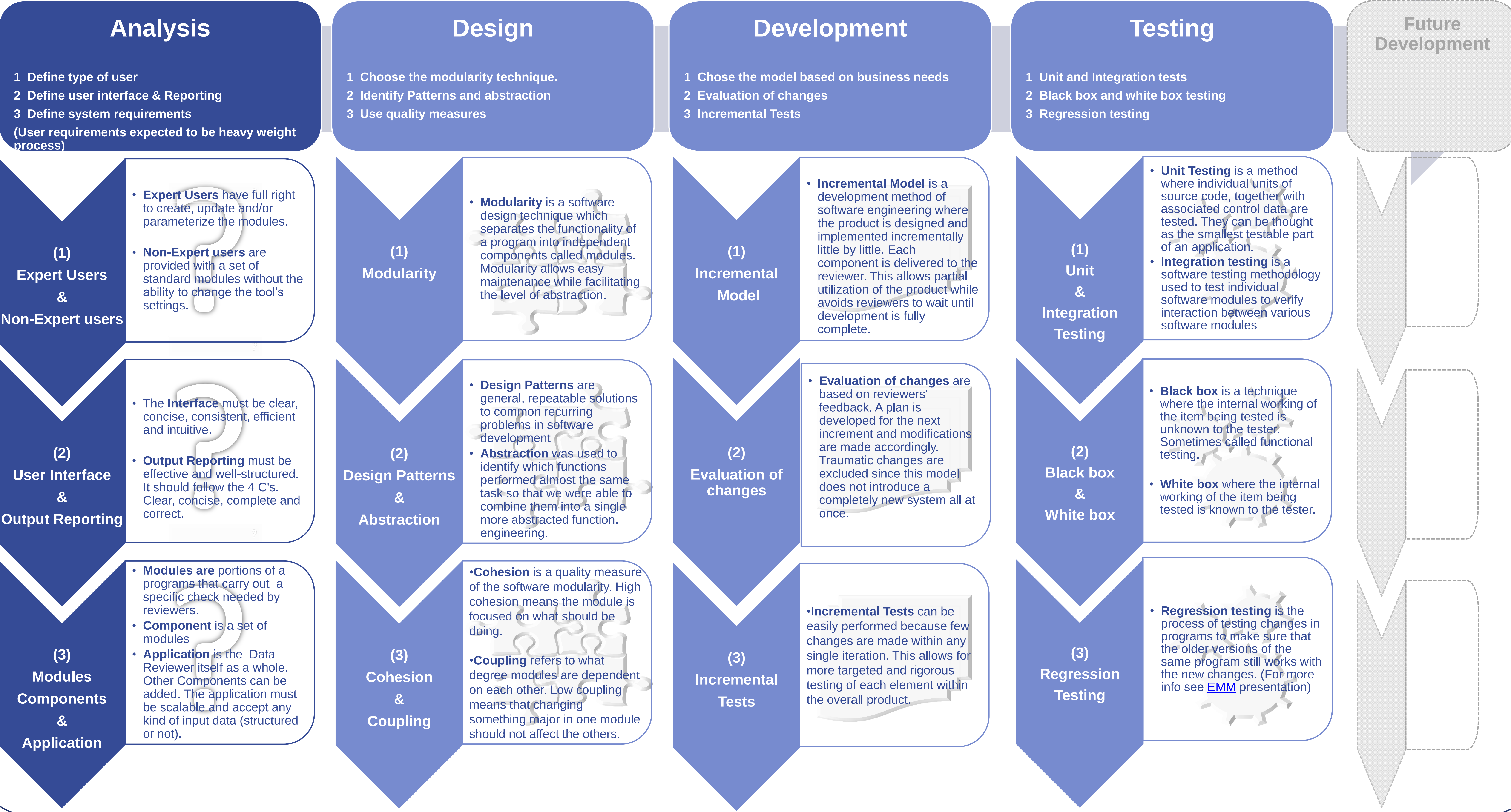
Abstract

Data Science arose from the blending of scientific and engineering methods. This new field requires expertise from different areas such as data cleaning, statistics, and software engineering (or IT). Naturally, a significant challenge is associated with obtaining in-depth knowledge in all of these areas. In order to overcome this, each member of our team is assigned to the tasks that best fit their strengths. We developed tools that cover different parts of the areas mentioned above. The architecture of these tools was designed using rigorous techniques for development, testing and validation with a view to enabling fully scalable and automated maintenance. This paper will focus on a tool called “Data Review” that ensures the data is as clean and consistent as possible. This tool can either work as an independent application or can be part of a greater project covering more aspects of data science.

Introduction (Problem)

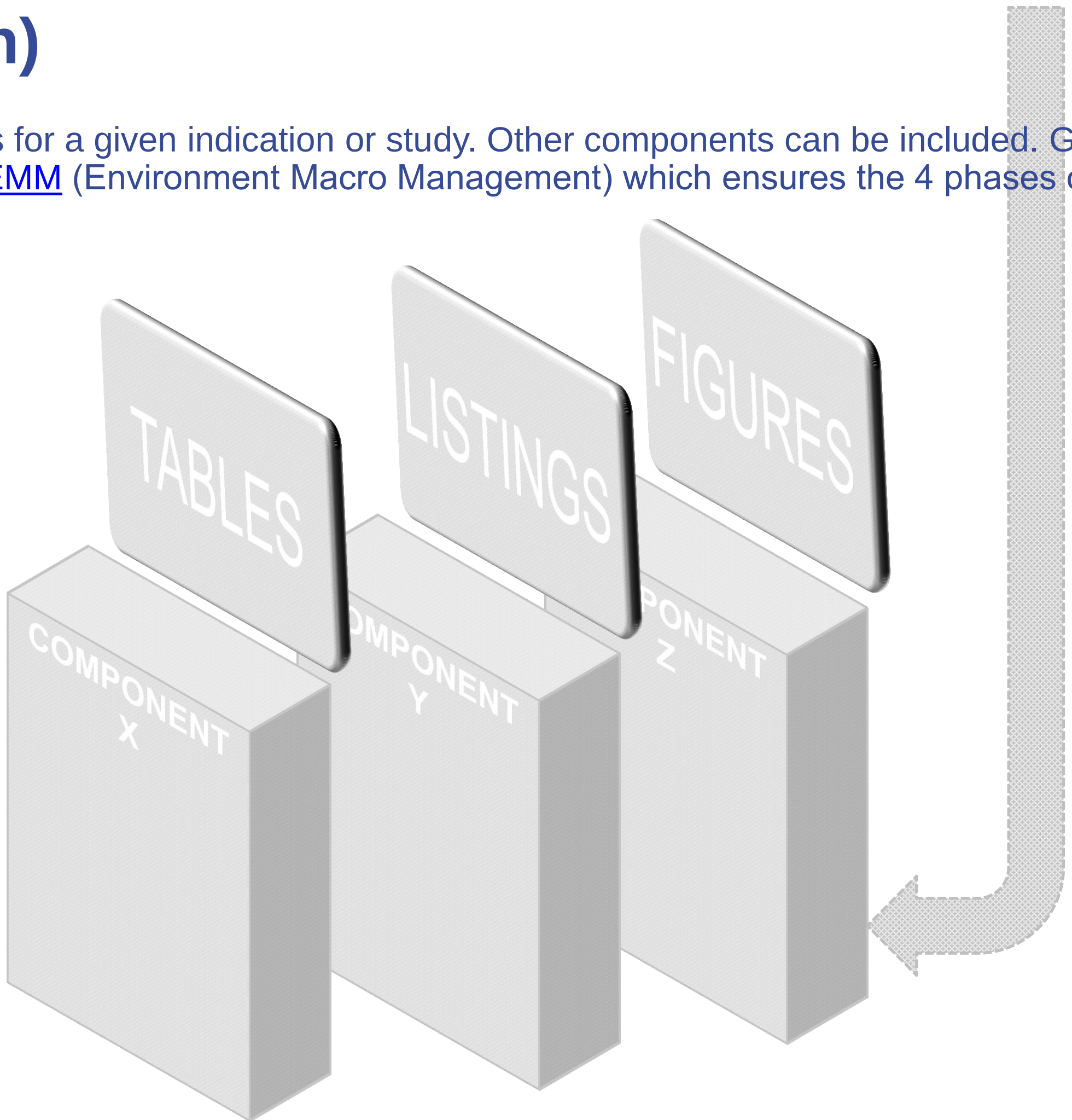
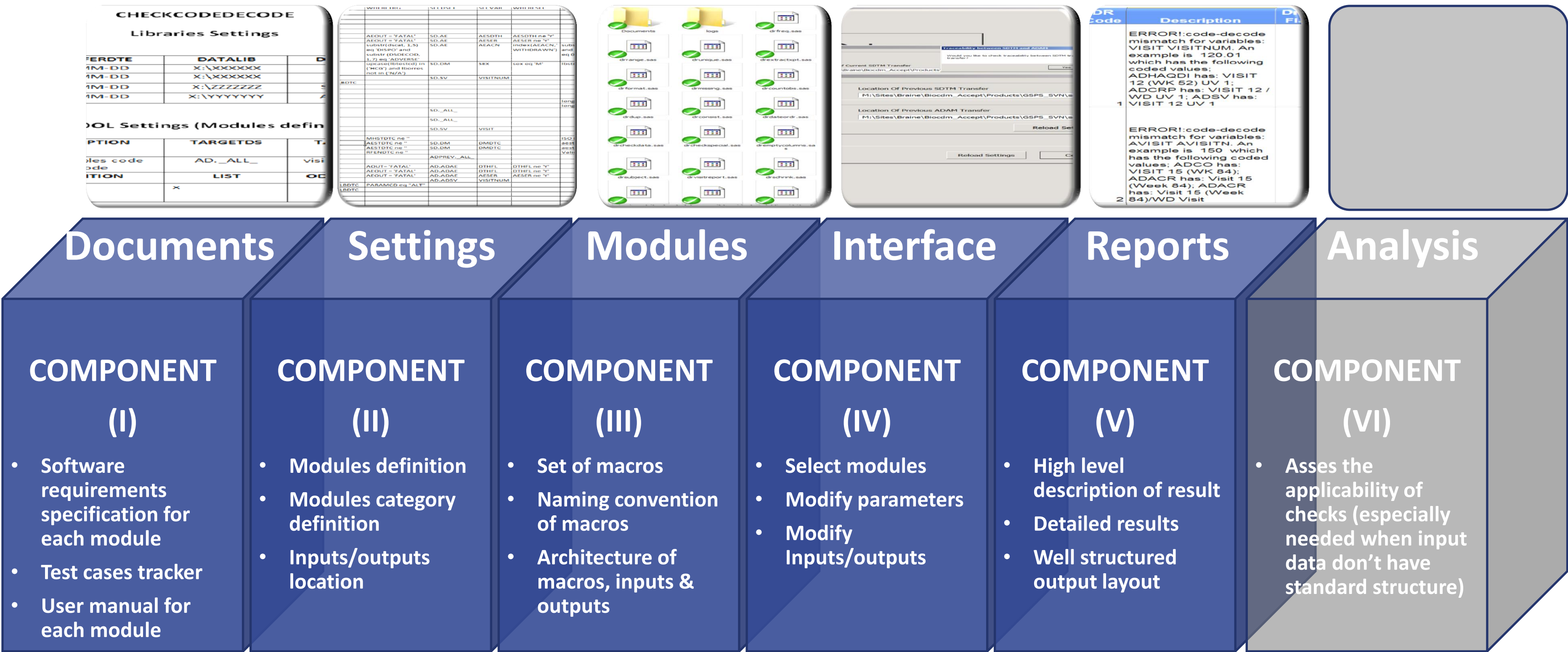
Often reviewers are asked to review data within a short timeframe. Our solution to this challenge is a software application called “Data Reviewer”. Our goal is to improve the quality, efficiency and cost-effectiveness of the data review process. This project was implemented using software engineering methodology. This method allows the application to include ever more components needed during the review process (for example TFL creation or validation.) The application could also become part of a much greater Data Science project.

Formal method for Software Engineering



Result (Data Review Application)

Data Review application is made up of five primary components. The sixth component will perform data analysis to asses the applicability of checks for a given indication or study. Other components can be included. Good candidates would be components to perform TFLs creation and/or validation. The entire application with all its components sits on top of a validation platform [EMM](#) (Environment Macro Management) which ensures the 4 phases of the Tool's lifecycle (Development, Testing, Release, Support) are fully controlled and automated.



Though the overall experience was positive and helped with identifying issues, the user requirements phase turned out to be the most problematic. This reaffirmed that a “Data Scientist” needs to have really broad knowledge.