

A visual programming approach for intuitive handling of bioprocess data

Robert Söldner¹, Jonas Austerjost¹, Simon Stumm¹, David J. Pollard¹

¹ Sartorius AG, Corporate Research, August-Spindler-Straße 11, 37097

* Corresponding author: robert.soeldner@sartorius.com

Introduction

During the last decade the expansion of high throughput biology and automated bioprocessing workflows has progressed by the advances in robotics, automation and sensor technology. As a result, the size and number of datasets dramatically grew as did the complexity. An upstream bioprocess engineer or scientist may have to manipulate & wrangle datasets from several different sources, such as online bioreactor profiles, and offline sample analysis, to process one typical experiment. This data wrangling burden to the end users often counteracts any efficiency advantage from using high throughput tools. This generated a requirement for data aware scientists to intervene over the wrangling and processing of complex datasets. This inefficient approach desperately needs a streamlined workflow to eliminate the need of experts for repetitive data processing tasks.

An innovative approach to simplify and democratize basic data processing tasks is the introduction of visual programming.

Visual programming is a paradigm that enables programs or scripts to be created by connecting graphical elements and illustrations instead of writing complex computation as text. This allows an end user, with no prior coding experience, to describe a process using an intuitive drag'n'drop approach. This degree of freedom enables lab scientists to perform rapid data processing based upon functions already implemented as graphical building blocks. In addition, this frees data scientists to focus on more complex challenges, than general data wrangling tasks.

Approach

The initialized visual programming application is based on the free and open-source KNIME Analytics Platform [1]. For the realization of the component, several recurrent data wrangling tasks within an R&D bioprocessing laboratory were identified. For multi-parallel bioreactor system batch data, these data wrangling tasks included the removal of undefined data points (NaN – Not a Number), forward filling of data points, data visualization and more. Based on these identifications, specific building components were implemented that cover the specific tasks. Subsequently the usability and the intuitiveness of the custom components was tested by lab scientists with no prior coding experience.

Results

The implemented components are based on KNIME standard nodes, which were customized to fit to the structure and content of the bioprocessing raw data. These standard nodes were then condensed into single components to remove the complexity of the workflow (see Figure 1).

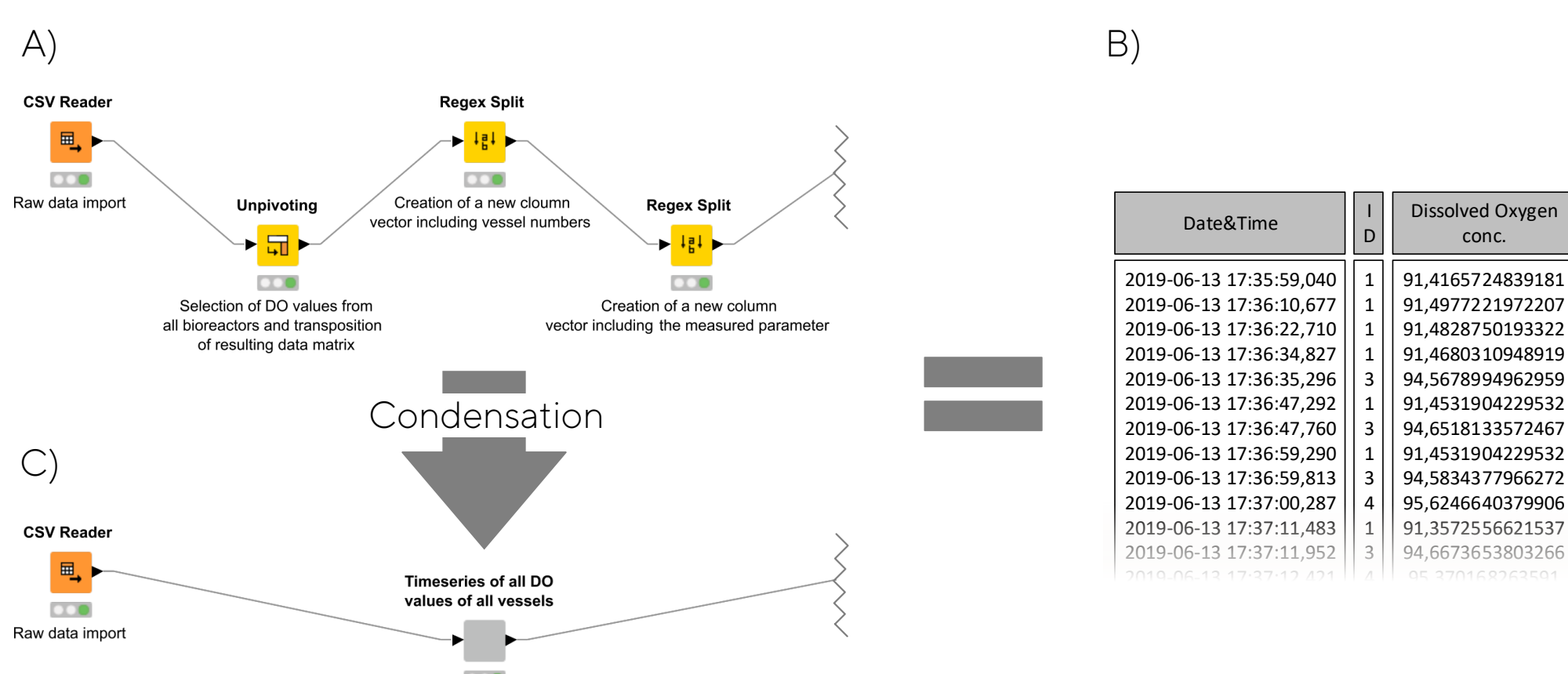


Figure 1: Connected nodes (A), which depict the data wrangling process of multi-parallel bioreactor system batch raw data into a structured table with the desired features (B). A simplification of the data wrangling workflow was introduced by condensing the processing nodes of flow A) into a single component (flow B).

This simplification process allows for the easy combination of workflow steps into “macro” elements, which enhances the clarity of the application. Several of these overarching macro elements were implemented. These included the transfer of specific measured parameters (DO, pH, current volume) of a specific fermentation vessel into designated tables, as well as string to number conversion combined with a forward fill of the imported raw data.

Furthermore, data visualization features were introduced as part of the data processing “assembly line”. Again, the basic components of KNIME were modified to fit the demanded visualization application (see Figure 2).

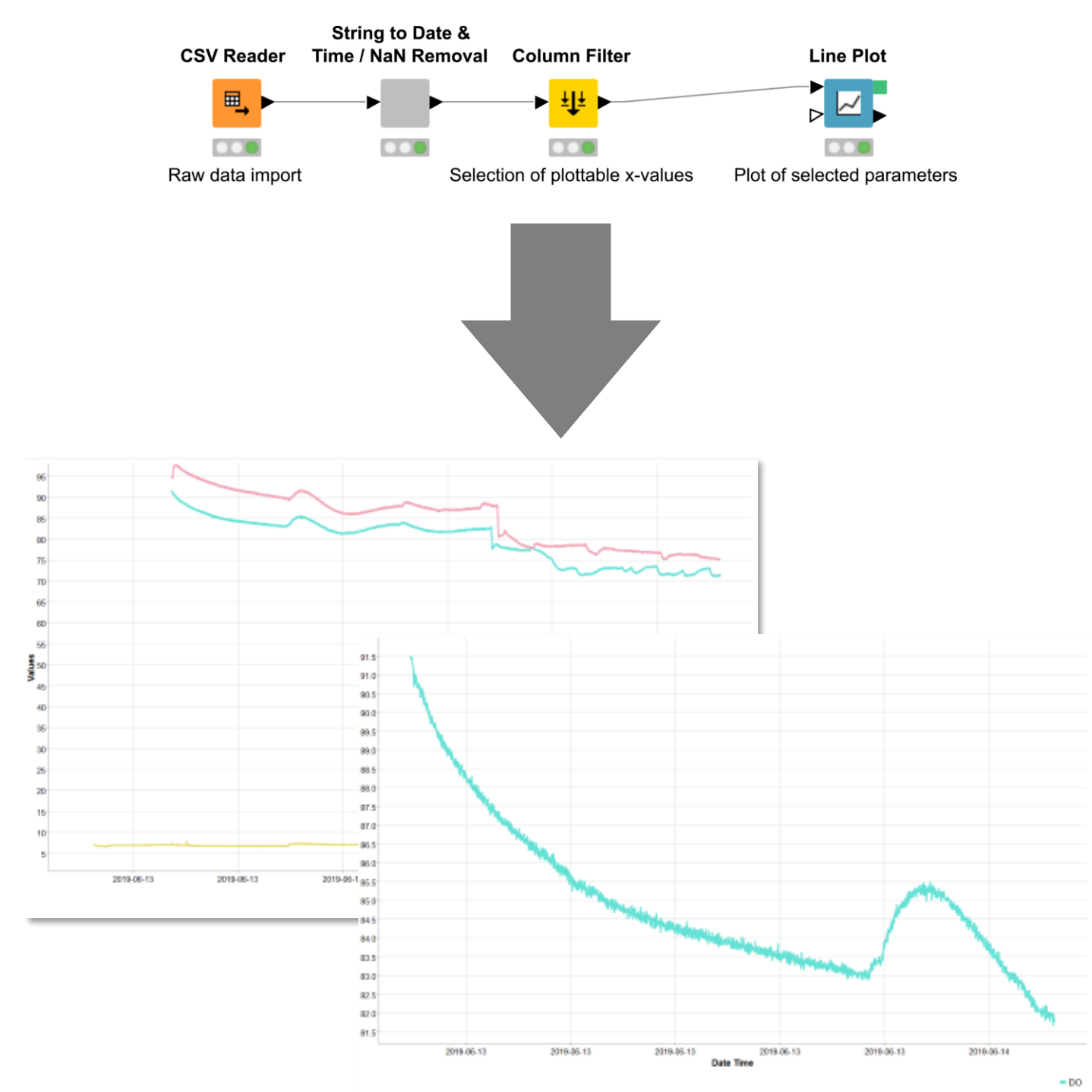


Figure 2: Connected nodes, which depict a data wrangling and visualization process of multi-parallel bioreactor system batch raw data. Plottable features of the preprocessed raw data file were selected and plotted in line plots.

All generated visual programming building blocks were evaluated by laboratory staff. The technique was identified as a fast and simple approach to perform basic data wrangling and visualization tasks.

Conclusion & Outlook

The presented visual programming approach identified as a valuable tool for democratizing and accelerating the data wrangling and processing routine in a bioprocessing laboratory. Work that previously was performed by SMEs can now be executed by all parties in a short amount of time. Although the concept was assessed on bioprocessing data, the initialized approach is applicable to various kinds of structured or unstructured data. A unified, and broadly agreed standard for analytical and process data could make the process of custom building blocks redundant, though. Future implementations are aiming to integrate data analysis components, such as SIMCA-Q [2].

The concept of visual programming is not limited to data processing. The integration of already existing visual hardware, service and software wiring platforms into the laboratory and manufacturing space could facilitate and democratize equipment integration tasks and the setup of process automation pipelines.

References

[1] KNIME Analytics Platform - <https://www.knime.com/knime-analytics-platform>

[2] SIMCA-Q Data Analysis Software - <https://umetrics.com/products/simca-q>