Developing Executable Phenotype Algorithms using the KNIME Analytics Platform

Instructors:

William Thompson, PhD
Department of Medicine
Feinberg School of Medicine
Northwestern University, Chicago, IL

Huan Mo, MD, MS
Department of Biomedical Informatics
Vanderbilt University, Nashville, TN

Jennifer Pacheco
Center for Genetic Medicine
Feinberg School of Medicine
Northwestern University, Chicago, IL

Robert Carroll, PhD
Department of Biomedical Informatics
Vanderbilt University, Nashville, TN

Summary:
KNIME Analytics (www.knime.org) is an open source platform that integrates data access, data transformation, statistical analysis, data-mining tools, and snippets of different programming languages in a visual workbench. It is the sixth most popular data science tool in the 2015 KDNuggets poll. The Electronic Medical Records and Genomics (eMERGE) network has implemented increasing numbers of phenotype algorithms, such as colon polyps and type 2 diabetes mellitus (T2DM)[1], as KNIME workflows to enhance their inter-institutional portability[3,4]. The PhEMA group has also demonstrated an executable implementation of the NQF Quality Data Model (QDM) on KNIME[2]. In addition to the applications that we will demonstrate in this tutorial, KNIME provides toolkits for Next Generation Sequencing (NGS) analyses, data mining, text processing, and social media research. It also provides support for integrating code written in SQL, Java, R, Python, and other programming languages.

Part one of the presentation will cover fundamental KNIME concepts and operations, the development of extract, transform, and load (ETL) workflows, algorithm implementation, and effective communication in KNIME. After introducing basic concepts, we will show how ETL workflows can facilitate intra- and inter-institutional collaboration and that these workflows are executable artifacts which enable reproducible research. We will then implement a phenotype algorithm, developed and validated by eMERGE, to demonstrate the value of KNIME in a more complex workflow. We will conclude this section by emphasizing collaboration as one of the core benefits of KNIME, dedicating time to discuss effective communication using KNIME. In the second part of the presentation, we will discuss how to leverage the extensibility of KNIME for more sophisticated applications, including

phenome-wide association studies (PheWAS)[5,6], natural language processing (NLP), XML processing, and incorporating RESTful web services.

We will limit the lecture time to around 30% in order to allow the audience to have ample hands-on experience editing and executing workflows for most of the demonstration. Instructors will be available to answer questions during this hand-on session.

Outline of Topics:

This tutorial will include introduction of KNIME concepts, data I/O and transformation, and real world applications.

Part I: Programming and Collaborating with KNIME.

- KNIME introduction: KNIME concepts, data structures, data I/O, and basic transformations.

- Case study (eMERGE T2DM algorithm): We will demonstrate the transportability of phenotype algorithms, and teach audience how to adapt imported workflows to read and execute on local data from their institutions. We will also lead the audience to add new logic to the algorithm.

- KNIME best practices in collaborations: defining and using KNIME file paths, modularization, and documentation of input fields.

Part II: KNIME Extensions

- KNIME, R, and PheWAS: We will introduce the integration of R into KNIME workflows, and we will guide the audience in using the PheWAS R package in the KNIME environment.

- KNIME, Java, and NLP: We will demonstrate integrating Java NLP packages into KNIME workflows using Java snippet nodes, feeding in unstructured text and extracting structured data output. We will also demonstrate KNIME native NLP nodes.

- KNIME, XML and RESTful web services: We will use KNIME modules to make connections to the NLM (National Library of Medicine) Value Set Authority Center (VSAC), i2b2, and RxNav API to demonstrate using the XML and RESTful nodes in KNIME.

Tutorial Length: 3 hours.

Learning Objectives:
After this session, the participants should be able to:
- Create KNIME workflows, perform data I/O, and data transformation;
- Adapt downloaded KNIME workflows for existing phenotype algorithms and adopt them with local data, and create sharable workflows for collaboration;
- Have a picture of the variety of KNIME extensions and possible applications.

Intended Audience:
The target audience consists of data researchers, statisticians, educators, and clinicians interested in data handling and analysis. This tutorial is particularly designed for the audience with the following needs and interests:

- For those who desire to use a more graphically oriented tool, rather than a tool that requires intensive coding;
- Need to work in teams;
- Interested in doing reproducible research that can be shared with other researchers

Content Level:
Basic 40%, Intermediate 40%, Advanced 20%.

Prerequisites:
- Experience with tabular data, spreadsheets, and a basic understanding of data models of relational databases will be helpful, but not required.
- Java knowledge is not required, but it is helpful for those who want to use the more advanced features of KNIME such as developing custom nodes.
- Please download and install KNIME with all free extensions before the tutorial.
- Please download and install R statistics on your laptop, with rJava and R PheWAS package. (https://phewas.mc.vanderbilt.edu/)

Instructor experience:

William K Thompson, PhD, Research Assistant Professor: Dr. Will Thompson began his career at Motorola Labs, during which time he implemented a variety natural language processing (NLP) software components that have been commercially released in Motorola handsets and other products, as well as resulting in several U.S. patents. At Northwestern University's Feinberg School of Medicine, he began developing NLP algorithms for the eMERGE and Pharmacogenomics Research Network (PGRN) projects. In collaboration with Dr. Joshua Denny (Vanderbilt University) and Dr. Jyoti Pathak (Weill Cornell), he was awarded a grant from the NIH to create a national infrastructure for standardized and portable EHR phenotyping algorithms (aka. PhEMA Project). Dr. Thompson has been leveraging the KNIME data analytics platform for these projects, and has developed and shared KNIME-based phenotyping algorithms across multiple sites in the eMERGE network.

Huan Mo, MD, MS, Research Fellow: With a strong background in both medicine and biomedical informatics, Huan Mo has been devoted to research in EHR-driving phenotyping and pharmacogenomics. He is an active advocate of an interoperative and collaborative academic framework in clinical data research. He led the publication of desiderata for computable representations of EHR-driven phenotype algorithms, with 30+ coauthors across the nation. Following the lead of Dr. Thompson, Ms. Pacheco, and many other investigators in the eMERGE network, Huan Mo has participated in developing and adopting KNIME implementations of many phenotypes in PheKB. Last year, Huan led the publication of "A prototype for executable and portable electronic clinical quality measure using the KNIME analytics platform", which was recognized with the 2015 Distinguished CRI Paper Award.

Jennifer Pacheco, Informatics Lead for NUgene: Jennifer leads the technical operation for the NUgene Project, a long-term research study to examine the role genes play in the development and treatment of diseases and phenotypes. Jennifer has a background in computer science and bioinformatics, and is currently working on a master's degree in biomedical informatics. Prior to joining NUgene in 2005, she was a data mining technology consultant, and a scientific systems analyst at a pharmaceutical company. Since the beginning of the eMERGE project in 2007, she has devoted most of her time to designing, developing, and executing algorithms to extract phenotype data from the EHR. In particular, she collaborates with informatics professionals within and outside of NU to develop best practices for sharing standardized transportable phenotype algorithms that are both computable

and human readable, and to that end, has been using KNIME to create and share executable versions of phenotype algorithms across the eMERGE network and beyond.

Robert J Carroll, PhD, Research Assistant Professor: Dr. Robert Carroll is a phenotyping expert with experience in genetic analysis. He has participated in cross institutional research to evaluate the performance of statistics-based phenotyping methods and in large scale genetic analyses through the eMERGE network and other consortia. Dr. Carroll is the primary author and maintainer of the R PheWAS package, which integrates a complete PheWAS workflow (data transformation, statistical analysis, and plotting) in the R environment. He has applied the KNIME platform to implement a machine learning workflow for phenotype identification.

Funding: This tutorial is funding by PhEMA (R01 GM105688), PheWAS (R01 LM010685), eMERGE (U01 HG006379, U01 HG006378, U01 HG006388), and iPGx (R01 GM103859).

References:

1. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. (2012) Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 19(2):212-8. doi: 10.1136/amiajnl-2011-000439.
2. Mo H, Pacheco JA, Rasmussen LV, Speltz P, Pathak J, Denny JC, et al. (2015) A Prototype for Executable and Portable Electronic Clinical Quality Measures Using the KNIME Analytics Platform. AMIA Jt Summits Transl Sci Proc. 2015:127–31.
3. Tromp G, Borthwick KM, Smelser DT, Bock JA, Elmore JR, et al. (2015) Ephenotyping for Abdominal Aortic Aneurysm in the Electronic Medical Records and Genomics (emerge) Network: Algorithm Development and Konstanz Information Miner Workflow. Biomedical Data Mining 4: 113. doi: 10.4172/2090-4924.1000113.
4. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, et al. (2015) Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc 22:1220-30. doi: 10.1093/jamia/ocv112.
5. Carroll RJ, Bastarache L, Denny JC. (2014) R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 30(16):2375-6. doi: 10.1093/bioinformatics/btu197.
6. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology. 31(12):1102-1110. doi:10.1038/nbt.2749.