Use Case Collections

Edited by
ELISABETH **RICHTER**

Best of KNIME

The COTM Collection

August 2020 - July 2022

Open for Innovation
KNIME

KNIME Press
Talacker 50
8001 Zurich
Switzerland

knimepress@knime.com

www.knime.com

Let’s Learn from the Best

If you want to learn more about data science and about KNIME, then you should learn it from the best! Here is the KNIME best.

In this booklet we have collected the top stories of the top contributors to the KNIME community from August 2020 until July 2022. The top contributors have been selected and awarded every month for their excellent technical skills and for their contribution to the upskilling of the community in data science and KNIME.

During this time, we have had 25 KNIME KNinjas, one for each month, with two awardees in January 2021. In the group you can find educators, data lab managers, data scientists, data analysts, data engineers, academics, biologists, chemists, marketing analytics experts, software engineers, etc. It would not be surprising to find two or more such qualifications combined in one single KNinja.

We have collected their most popular stories, the stories either because they solved a long-standing problem or because they taught us something useful for our professional life. Sometimes, it was impossible to select one single outstanding work and we decided to publish a more general interview covering all fields of the contributor’s activity.

In these past two seasons of the KNIME Contributor of the Month program we have learned the key features of the KNIME Software from Vijaykrishna Veknataram’s point of view, how to properly deal with duplicates in your data with Markus Lauber, that text processing and GUID generation can be easily done using SJ Porter’s components, how to create text-rich visualizations of Twitter data around a given hashtag with Angus Veitch. the power of low code with Keith McCormick, about the two founders and admins of the KNIME Analytics Community group on Facebook, Evan Bristow and Miguel InfMad, that Armin Ghassemi Rudd's Translator and Get Request Plus components bundle powerful functionalities, how to automatically generate new characters and setups for a Dungeons & Dragons session with Philipp Kowalski, what it takes to become a data scientist with Dennis Ganzaroli, or what it takes to become officially KNIME certified with Giuseppe Di Fatta, how to identify new drug candidates using KNIME with Alzbeta Tuerkova, also much interesting revolving around the application of KNIME in Japanese with makkynm, how to use KNIME to detect fraudulent credit card transaction with Tosin Adekanye, about Excel to KNIME in Spanish with Ignacio Perez, that creating XML can be easy and flexible with Brian Bates’ components, how to integrate Python code within a KNIME component with Ashok Harnal, how to set up a data science lab with Andrea De Mauro, and also, how to apply machine learning to DNA analysis with Malik Yousuf, that Nick Rivera’s YouTube channel is another great source to learn KNIME, that you can teach a Machine Learning model using KNIME with Paul Wisneskey, how to use Machine Learning in

Marketing Analytics with Francisco Villarroel Ordenes, what it takes to be a good KNIME Forum poster with BrunoNg, multiple ways how to use KNIME for all kinds of QSAR topics with Christophe Molina, and last but not least, how to parse and analyze PDF documents with John Emery.

Let’s not indulge longer into this introduction. Let’s hear more from the KNIME best.

Elisabeth, Scott, Corey, and Rosaria

Table of Contents

Welcome to the KNIME Contributor of the Month Program

You might have seen award cards or badges for KNIME Contributor of the Month (COTM) on social media that look like this one below. If you have never seen one or if you are still wondering what this celebrates, now you will get an answer.

The award card for our KNIME Contributor of the Month for December 2021 – Andrea De Mauro. Each month we award one community member for their outstanding commitment, and each COTM receives a badge.

What is a KNIME COTM?

The KNIME COTM award is assigned to KNIME users who have shown excellent technical skills and have contributed to a better learning experience as educators, to faster and more exhaustive technical support, to knowledge sharing via articles, blogs, and YouTube videos, to a richer repository of community nodes and components, or to a stronger KNIME presence on social media.

In a nutshell, the COTM award is assigned in recognition of technical skills and contribution to the advancement of the KNIME community. Indeed, there are only a handful of such top experts around the world: this level of expertise and dedication is hard to find!

The program has started in August 2020 and is going strong ever since! With the nomination of John Emery in July 2022 we reached the point where our program fully passed through two entire seasons. Until then, we have awarded 25 KNIME users, one per month (with the exception of January 2021 when we awarded two community members at the same time). However, this is not the end of the program as in August 2022 season three has started…

How to become a COTM Awardee?

First of all, somebody – a fan, a colleague, a family member, yourself – must nominate you or any other KNIME user for a COTM award via the COTM Candidate Proposal form. In the description field of the form, the reasons why the nominee deserves to receive the COTM award must be stated.

Once a month all COTM nominations are evaluated. Based on their recent activities, their contribution to the KNIME community, and their technical skills, the best nominee is selected and awarded the title of COTM for the next month.

If, while reading, the name of a deserving KNIME user springs to mind, do not hesitate and nominate them for the COTM award program in the COTM Proposal form.

The list of all past COTMs for all seasons can be found in the KNIME COTM Hall of Fame.

KNIME in the Data Science Lab

In this section we have assembled all contributions that focus on KNIME Analytics Platform and its capabilities. The articles presented here reveal what KNIME has to offer and uncover the power of KNIME Analytics Platform. The category "KNIME in the Data Science Lab" features our eager Data Science experts:

- **Vijaykrishna Venkataram**

  – Senior Manager, Data Analytics *@Relevantz*

- **Andrea De Mauro**

  – Head of Data & Analytics *@Vodafone*

- **Dennis Ganzaroli**

  – Head of Report & Data-Management *@Swisscom*

- **SJ Porter**

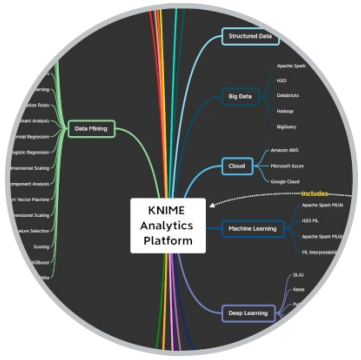
  – Site Reliability Engineer *@KNIME*

- **Ashok Harnal**

  – Professor *@FORE School of Management*

Vijaykrishna Venkataram was nominated KNIME COTM for August 2020 which makes him our very first COTM! He was awarded for creating a mind map of the features, extensions, and integrations of KNIME Server and KNIME Analytics Platform. Indeed, the whole idea of the COTM program came from his mind map. The mind map was both so detailed and so general that we thought that valuable contributions like this should be rewarded. In the figure on the right, you can see a part of said mind map. If it is too small to distinguish anything, you can read his LinkedIn post on this same topic.

Vijay holds a Master of Science in Financial Economics, Finance & Economics from the Madras School of Economics in Chennai, India. He has more than 14 years of experience in the Banking and Credit Bureau industry and is currently Senior Manager, Data Analytics at Relevantz. A few of his focus areas include Credit Scoring (Acquisition & Behavioral Scorecards), Loss Forecasting, Cross-sell Analysis, and Market-sizing Analysis. Besides his technical skills, Vijay also brings expertise in communicating complex modeling solutions to technical as well as non-technical business audiences.



Visit Vijay's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `vijayv2k`).

Expert Discusses Tools & Processes for Effective Financial Analytics

My Data Guest — An Interview with Vijaykrishna Venkataraman

Author: Rosaria Silipo

It was my pleasure to recently interview Vijaykrishna Venkataraman as part of the *My Data Guest* interview series. He shared insights into the work of a data scientist in the banking and financial sector, illustrated key features of KNIME Software that help his organization optimize processes, and spoke about the importance of model governance in highly regulated industries.

Vijaykrishna Venkataraman is an expert data scientist with 14 years of experience in data science applied mainly to the banking and financial sector. He used to work in the Risk Modeling and Analytics department of the Central Bank of Oman, and is currently moving back to India, where he will be working as a Senior Manager and Data & Analytics Consultant. Vijay was also the first KNIME Contributor of the Month, having been awarded in August 2020. Indeed, the whole idea of the Contributor of the Month award came from his work. After seeing the mind map that Vijay produced to describe the KNIME features, we thought "Such creative work should be rewarded. Let's reward it!" and the rest, as people say, "it's history".

Rosaria: *Let's start from your famous KNIME feature mind map. Can you explain it to us?*

Vijay: The idea of creating the mind map came about while we were already using KNIME Analytics Platform extensively in our organization. Every now and then a colleague came to me and asked something about the platform: "Does KNIME support this?" or "Can I do that in KNIME?". I had to explain KNIME's features every time. At some point I thought "Why do I have to explain the same things over and over again? Why not create a mind map?". KNIME is a visual programming platform after all, and it should be possible to illustrate its features visually rather than going the old-school way of preparing documents. The mind map does not cover the entire spectrum of all

features KNIME Analytics Platform supports, but I think that the most important ones are definitely included.

Rosaria: *Is it updated to the latest release of KNIME Analytics Platform and KNIME Server? Where are the new features located in your map?*

Vijay: Yes, this mind map is the second version and we updated it to the latest release (v4.6) of KNIME Software. We have added most of the new features at the bottom of the map. The updates we have made to the mind map also include key features of KNIME Server.

Rosaria: *What KNIME features do you use the most in your work?*

Vijay: Our business is highly regulated, and as such most of the data we work with is hosted on-premises. However, part of our data is scattered across multiple platforms and products. Hence, we were in need of a platform that brought everything into one place. This is where KNIME Analytics Platform came into the picture. We use KNIME's dedicated DB nodes extensively to extract the data. What follows are usually several data aggregation and cleaning operations – ETL, basically. This makes up 70-80% of our work. The next step is the automation of processes and this is where KNIME Server demonstrates its power. Especially the possibility to connect KNIME Server with other Business Intelligence tools, such as PowerBI, is very valuable.

Rosaria: *To understand why those features, maybe you can tell us more about your professional self. I should have asked this question earlier. How many people are in your group? How many professional profiles?*

Vijay: For most of my professional life, I have been working in data & analytics for the banking and financial sector. My work focuses mostly on risk analytics, such as risk management, credit scores, or fraud analytics. I contributed to set up the credit bureau for the Central Bank of Oman, where my job involved not only the design of the data pipeline via data engineering solutions, but also data science tasks, such as the creation of credit scorecards, and business analytics processes, such as benchmarking and market insights reporting.

As for the professional profiles in our group, we mostly have data engineers and data analysts.

Rosaria: *Are you hiring?*

Vijay: Yes, I'll be hiring soon. I'm currently in the process of moving back to India and joining another organization, where I'll be setting up the data analytics practice. Once I take up that role, there will be for sure some hiring coming down the pipeline in maybe 3 to 6 months from now. If you're interested, follow my LinkedIn profile to stay up-to-date.

Rosaria: *Some time ago I wrote the article "KNIME Experts wanted!" where I lamented the lack of KNIME experts to fill job vacancies. Has it been true for you too? How complicated has it been so far to hire data scientists and KNIME experts?*

Vijay: When I read your article I thought "Somebody has finally spoken my mind!". When we started looking for people to join our group in 2019, we were specifically looking for candidates who had KNIME experience. Unfortunately, back then, there were not so many KNIME experts available.

Now, things are gradually getting better but it can still be challenging. Therefore, if you are looking to hire KNIME experts, let me give you a tip: Don't wait until someone with KNIME expertise appears. Rather, look for any business or data analyst who fits your business requirements and teach them KNIME. I am sure that within a few weeks they will master the tool. This is what we did when we started working with KNIME and the management asked where we could recruit candidates with KNIME skills. I hired two business analysts and gave them two weeks to get familiar with KNIME. At the end of the second week, they were able to crack some of the problems I gave them.

The KNIME Certification program is also very helpful in that sense, and that's why we asked our employees to get L1- and L2-certified. L1 is free of charge, and L2 is just a fraction of the costs you would normally spend on proprietary software. As a result, we don't have to scout talents on the market but we train and grow our own talents internally.

Rosaria: *What is the professional category which is the hardest to recruit? And why?*

Vijay: The most complicated people to hire are the ones who have a machine learning background and want to build predictive models straight away. Students and fresh talents are often under the impression that model building is the most crucial part of a data science project. They don't realize that in the real world this is just 10-15% of the entire project. The majority of the work is to get the required data, understand the business case and the data at hand, prepare and clean it, do feature engineering if necessary, define your target, etc. Data modeling comes much later.

For example, in the banking industry, we talk a lot about scorecard development or fraud analytics. Especially in fraud analytics your prediction target, the fraud rate, is very low. So how do you come up with a good model for prediction? This is where you first need to understand the data and the business, and then you can jump into model development.

Rosaria: *What tools does a data professional need to know to work in your group? Python? KNIME? Both?*

Vijay: Initially, we used proprietary software to run our scoring models. But the license got very costly, and talents in the market for such proprietary tools were getting dearer. The younger generation was using more open-source software, which we wanted to

adopt too. So we started using Python to build most of our machine learning models, and SQL for data munging. Currently, we are using KNIME for all our ETL tasks, and still rely on Python for machine learning. However, we are slowly migrating some of our machine learning work from Python to KNIME. Especially with the latest KNIME Software release (v4.6) we can now create Python-based nodes.

Rosaria: *Besides Python and KNIME, do you use other tools in your group?*

Vijay: Besides KNIME and Python, we use Business Intelligence software, predominantly PowerBI, for reporting. Moreover, as we are also working with increasingly large datasets, we are experimenting with distributed systems. In that sense, Spark looks promising because of its speed and because KNIME has a dedicated extension for Apache Spark.

Rosaria: *How has KNIME helped you and your team in your work? Can you take us through a use case or example?*

Vijay: Okay, let me give you an example. Loans in the banking sector are issued in multiple currencies, and currencies keep fluctuating every day. Suppose I'm preparing a portfolio review and I want to see how my portfolio looks today. To do that, I need the latest currency value to convert the loans of my portfolio into the present currency value.

We first asked our IT team to build an app that updated exchange rates of about 50-60 currencies on a daily basis. However, it would have taken too long for them to build such an app.

The next best solution was to use KNIME. Using KNIME REST nodes, we created a workflow that automatically crawls on a daily basis exchange rates provided by the corresponding currency exchange providers, updates them, and writes them back to the database. This solution was implemented in a couple of days – literally. Thanks to KNIME Server, we could then schedule the workflow to run every day at 12 o'clock with just a few clicks. And should an error arise, KNIME Server notifies us very conveniently per email.

Rosaria: *You mentioned that you are working with increasingly large datasets. How does KNIME handle a large number of records?*

Vijay: I'll answer with another example. At some point, we decided to switch from Oracle to Microsoft SQL Server. This means that we had to migrate all the data to a new server, and adjust the data structure. It was a huge task as we had five years of data laying around. With the help of KNIME, we were able to smoothly migrate 137.5 million records (i.e., about 5.8 billion data points). This was done using a complex workflow which fetched the data from the database and then brought it in the new structure. This big project was successfully completed by two people only, and without writing one single line of code.

Rosaria: *That's wonderful! Let's talk about money and time saving. How significant is the impact of data analytics in a business? What are some immediate impacts that you have witnessed in your experience?*

Vijay: I might not be able to quantify this in terms of money but definitely in terms of time. With KNIME, we are able to save a lot of time. In a business like banking, time is all that matters. You could win or lose a client just because of the decision-making process, i.e., the time the customer is waiting for you to solve their problem. A great example is how KNIME optimized our customer support processes. Before switching to KNIME, it took us many days to help clients. Now, the process is way faster and we are able to track requests, and help clients in half a day, sometimes even in 30 min. The new KNME-driven solution also reduced our churn rate, which indirectly saved us money. In addition, we also saved resources. The number of customers we support now with only two employees would not have been possible in the past. We would have needed at least ten people.

Rosaria: *Tell us about your experience with KNIME Server and what convinced you and your team to opt for it (how did you convince the management to choose KNIME Server over other platforms?)*

Vijay: Our approach to convince the management was quite simple. We already convinced them to use KNIME Analytics Platform when we showed them that they were able to automate part of their work. Now, there was one hurdle left: How to automatically execute workflows? This is where KNIME Server becomes really valuable as you can schedule jobs, set execution queues, and send notifications upon fail/success.

But that's not all. Another very interesting feature of KNIME Server is the KNIME WebPortal. For example, we deployed a KNIME workflow as an interactive, browser-based data app to monitor sales where the end user (usually the management) can enjoy the dashboard without having to walk through the workflow behind it. For someone in finance, sales, or accounting, this is all they need to see.

My suggestion to anybody who is looking for a way to convince the management is to pick a tricky business case, and solve the problem using KNIME Software. Show them, for example, the KNIME WebPortal so they can understand the benefits of it without having any KNIME knowledge per se.

Rosaria: *We are reaching the end of our interview. Before we say goodbye, we cannot but ask the classic question. Where do you see data science going in the next few years? What will the next hype be?*

Vijay: I think data science itself is currently at the peak of its hype curve. One important aspect we should focus more on is model governance. In the last few years, we have witnessed the development of many new, extraordinary models but most of them are black boxes. The focus was not set on interpretability and explainability. However, in

industries like finance or insurance that are highly regulated, and where every deployed model has to go through the scrutiny of the regulator, model governance is likely to become increasingly more prominent. In the next few years, I believe we will move back to models that are easier to explain and interpret. A more parsimonious way of developing models has to be brought back. Creating black box models whose underlying decision making process is obscure for all stakeholders is not going to be sustainable anymore.

Vijay's famous mind map can be found on the KNIME Community Hub. If you want to connect with him, stay up-to-date about his future projects and job opportunities, add his LinkedIn profile to your network.

This article was first published in our Low Code for Advanced Data Science Journal on Medium. Find the original version here.

Watch the original interview with Vijaykrishna Venkataraman on YouTube: "My Data Guest – Ep 12 with Vijaykrishna Venkataraman".

Andrea De Mauro was nominated KNIME Contributor of the Month for December 2021. He was awarded for his book *Data Analytics Made Easy* (see image on the right), featuring a full introduction to KNIME and multiple tutorials showing how to build machine learning KNIME workflows. As Andrea says himself, the book is for everyone who works – or would like to work – with data.

Andrea has more than 15 years of international experience managing Data Analytics and Data Science organizations. He is currently Head of Data & Analytics at Vodafone Italy. Andrea holds a PhD in Management Engineering from the University of Rome, a Master of Science in Electrical and Computer Engineering from the University of Illinois at Chicago, a master's degree in ICT Engineering from Polytechnic of Turin, and a diploma in Innovation from Alta Scuola Politecnica at Milan. In his research, he investigates the essential components of Big Data as a phenomenon and the impact of AI and Data Analytics on companies and people.

Visit Andrea's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `adm`).

BI Leader Discusses Data Tools for Professionals

My Data Guest — An Interview with Andrea De Mauro

Author: Rosaria Silipo

It was my pleasure to recently interview Andrea De Mauro as part of the *My Data Guest* interview series. He busted myths about data science superheroes, settled the IT vs. data science question, and spoke about the importance of getting your hands dirty with data and algorithms.

Andrea De Mauro has more than 15 years of international experience in managing data analytics and data science teams with various organic organizations. Currently, he is the Head of Business Intelligence at Vodafone in Italy. Prior to that, he served as Director of Business Analytics at Procter & Gamble. He is a professor of Marketing Analytics and Applied Machine Learning at various universities including the International University of Geneva (Switzerland) and the Universities of Bari and Florence (Italy). He is also the author of popular data science books and of research papers in international journals.

Rosaria: *How many different professional profiles do you see in the data science space?*

Andrea: The traditional myth of a data scientist as a superhero, who takes care of entire end-to-end processes or the full landscape of complexities around analytics is far from the reality. Today, there are plenty of roles available in the amazing world of data analytics. I normally use three main role families to explain them:

Data analysts or business analysts, who have deep business expertise in a specific domain and "translate" needs between other data practitioners and the business teams; Data scientists, who focus more on the algorithms and on the scaling of the analytics capabilities; and Data engineers, who are involved with the implementation and maintenance of the full technology stack.

Rosaria: *Which professional category is the hardest to recruit?*

Andrea: They are all tough to recruit for these days! But I think the Business Analyst role is the hardest one: finding people who have what it takes to get business

complexities answered with the required algorithms is hard. The role is also difficult to explain to recent graduates.

Rosaria: *Do the different profiles need different data education in your opinion?*

Andrea: I think all of these profiles require primarily one thing: a growth mindset – the willingness to keep learning.

None of these professionals can survive without being open to learn continuously: this is particularly true for business analysts and data scientists. Data engineers require also some vertical technical expertise on one or more big data platforms, like GCP, AWS, or Azure. In general, as a data engineer, you want to have a certification also. The good news is there are plenty of opportunities to get certified.

I also think there is a very rich MOOC offering to learn data science online on platforms like Coursera, Edx, and other providers. They offer certification paths for aspiring data scientists or analysts. I normally recommend people who do not have an educational background dedicated to data science to make use of these learning platforms.

Rosaria: *What tools do you think a data professional should learn?*

Andrea: It's really important for an aspiring data professional to have the right set of tools and options – and to know how and when to leverage them. You don't need to learn all the tools, but have a good mix of products that complete each other as part of a versatile toolbox.

The type of toolkit I would recommend possessing would include a business intelligence product, focusing on enabling scaled dashboards and data visualization capabilities, and a versatile analytics platform. KNIME is a great example of low-code analytics platforms. I would also include more traditional code-based analytics tools, which can be integrated perfectly with a low-code platforms like KNIME.

Rosaria: *Now let's talk to the teacher, do you use KNIME to teach your data science courses?*

Andrea: Of course! I have used KNIME for a long time now, both at the universities and at work with Vodafone and P&G. It's an amazing tool to teach data science for multiple reasons. KNIME makes the process of coding convenient so you can focus on the core analytical tasks.

Those who would like to start using analytics feel often discouraged by their lack of coding skills. KNIME offers a solution to this. Visual tools like KNIME let you "see" and track what's going on at each step. You can easily identify where the problem is if you are stuck at a certain point. This really supports the educational experiences while teaching data science courses. In short, this increases the efficiency of the learning process for the students.

Students appreciate it as well. KNIME makes the learning more accessible and sometimes also more fun. The experience of building a workflow step by step is somewhat enjoyable for them. The usage of KNIME nodes makes the learning modular and progressive. A node makes you "see" what is going on with your data in the flow very easily. The Joiner node, for example, combines the two input tables into one single output table. Or the Loop nodes apply iteration to a sequence of steps between Start and End. By making it "visual", you understand it better and reduce the chance of making mistakes.

Rosaria: *So let's talk now to the technical professional. Do you see yourself as a data analyst, a data engineer, or a data scientist?*

Andrea: I see myself more as a data analyst, right at the intersection between the business requirements, the business needs, data and algorithms. Fortunately, I've had the opportunity to see the full picture and practice bits and pieces of all three roles. If I had to choose, I would see myself more as a Data Analyst.

Rosaria: *How significant is the impact of implementing data analytics in a business, does it help or it is just an academic exercise?*

Andrea: It's a rhetorical question, of course. Data analytics is making and will make a huge difference. It's a game changer for the business. It changes the operating model that defines how the company works. Eventually, it changes the way business is done. It's not just a technicality or an IT complexity to manage, but a novel way of running an organization. Emotionally, it also brings great excitement. Think about prototyping an analytics solution that fits well to your business case or finding deep insights you never expected. Experiencing the seemingly "magic" of data analytics is a potential morality booster for everyone.

Rosaria: *As a manager how do you build your data science team? Where do you start from?*

Andrea: Let's start where you don't start from! You don't start by hiring dozens of generic data professionals without first looking at the talents that you already have in your family. You can definitely grow data scientists from the current talents you already have in your company by upskilling those who have curiosity, passion, and willingness to learn.

Following this course of action, in my opinion, has two major benefits, which I feel are worth mentioning:

- Only the person with knowledge and understanding about how the company operates and how data flows through it can really understand the real opportunities and build some meaningful data analytics.

- It's refreshing for the professionals, no matter what their background is, to boost their career path and development by getting serious on data analytics. This opportunity should not be restricted to those who have a technical or an IT background.

Rosaria: *Talking about the IT and data science team, this is a question I've considered for a long time: Where should the deployment of data analytics applications reside? People argue whether it should stay with the data science team or be an IT responsibility?*

Andrea: It is difficult to have a general answer as it really depends on each case and how the company is organized, but one thing for sure is that there is a strong collaboration need between IT and Data science teams. Whether these teams are separated or together depends on where the company is in terms of maturity and on many other factors, sometimes power-driven and political.

If they are separated, it is necessary that each team understands the data lifecycle process. Otherwise it's unsustainable. So, in short, it's inherently a collaboration.

The ownership of the capabilities and their utilization, however, should reside in the business - neither in the data science team nor in the IT one.

Rosaria: *Now let's play a bit of myth busters in the field of data analytics. What are the myths around data analytics prevalent in the tech market and do you think they are unjustified?*

Andrea: Data analytics was born with a strong sense of euphoria around it. This has led to the creation of many myths:

First, as we mentioned earlier, some people were led to think that it is enough to have a good team of strong data scientists to cope with the full set of needs. This is a myth. You need a multifaceted team of data professionals, working hand in hand with engaged business partners. You also need a toolkit made of multiple tools to enable that team to perform at its peak.

Second, some claim that the process of data science will be fully automated. We hear a lot about AutoML, which is an important direction in machine learning and artificial intelligence. However, it creates the myth that in the future humans will not be needed and machines will take over the process, which isn't true. The AI we are working with right now is not meant to be generic. It is rather able to solve specific issues and complexities that can only be driven with human guidance. So, it will always be a collaboration between humans and machines.

Rosaria: *Let's talk a bit about the recent book you wrote "Data Analytics Made Easy". Who is this book for? Is the book enough to understand data analytics and to even apply it?*

Andrea: I would say it's for everyone who works – or would like to work – with data. The intent of the book is "to make data analytics easy." It gives practical insights on how to perform specific tasks such as creating a report or automating a sequence of steps that today you run in Excel, or building a compelling presentation, telling a story with data. It's for knowledge workers and for their managers. It's important for the latter to be aware of what data analytics is all about and learn first-person about what they are asking from their team when it comes to data. It's also for those students and graduates who would like to start a career in data analytics.

Everyone who reads the book will be able to put data analytics in practice. The tutorials are based on examples that can be reapplied to many different business cases. You don't have to be in a specific role or have a certain background to understand the content of this book. Ultimately, I think the book can give you reassurance that you "can" learn analytics little by little and become autonomous enough to put data analytics in practice at work.

Rosaria: *Finally, I would really like to know what career advice you would give to an aspiring data professional?*

Andrea: Do not wait for your first job to get the experience you need. You can start getting your hands dirty with data and algorithms well before your first interviews!

My advice will always be to go out and look for the right opportunities around you to apply analytics first-person. Look for local charities in need of help with their data. Check out websites like Kaggle to find free online competitions and gain experience. This will give you opportunities to build your own first portfolio of models and successful data analytics applications.

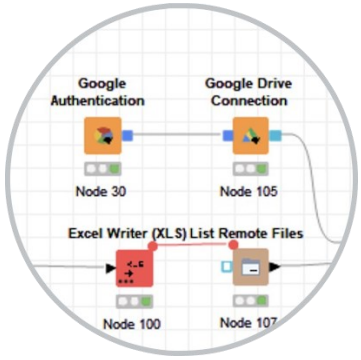
This article was first published on our KNIME Blog. Find the original version here.

Watch the original interview with Andrea De Mauro on YouTube: "My Data Guest – Ep 3 with Andrea De Mauro".

Dennis Ganzaroli was nominated KNIME Contributor of the Month for April 2021. He was awarded for his many articles, where he shares his experiences, recipes, and best practices for analyzing data with KNIME Analytics Platform - both alone and in combination with other tools. In the past months before his nomination, Dennis really has been on fire. He has been successfully predicting everything, like Nostradamus and the octopus Paul, but using data science models and KNIME Analytics Platform. First, he predicted the curve for the COVID spread worldwide; then he broke it down to predict the COVID spread country by country; and lastly, he even predicted the outcome of this July's European soccer championship, UEFA Euro 2020, with the help of a curious character: Yodime. But he is not only predicting things. He has also written an opinion piece on how and why KNIME has become the most important tool in his day-to-day work (see: The Best Tool for Data Blending Is in My Opinion KNIME). In the figure on the right, you can see a workflow snippet of Dennis' COVID spread worldwide article.

Dennis is a Data Scientist with over 20 years of experience. He is currently Head of Report and Data-Management at Swisscom AG Switzerland. Dennis holds a degree in Psychology and Computer Science from the University of Zurich.

Visit Dennis' space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `deganza`).



Spot-on Euro 2020 Predictions: Back to Classic Techniques

My Data Guest—An Interview with Dennis Ganzaroli

Author: Rosaria Silipo

It was my great pleasure to host Dennis Ganzaroli in the first episode of the *My Data Guest* interview series, which was aired on September 22, 2021. In this conversation, Dennis shared some of the secrets and best practices that lead to becoming a successful data scientist: First among all, the passion for data as the most necessary tool to support your data science profession!

We also got to discuss some of his latest achievements, like the (correct) prediction of the final players in the UEFA Euro 2020 tournament or the (correct) estimate of the evolution of the COVID-19 spread all over the world. Last, but not least, we even got some recommendations on what to read to keep up to date with the constantly evolving field of data science.

Dennis has been working with data for easily 20 years as a data engineer, data scientist, and sometimes as a data analyst. When you talk to Dennis you realize that his knowledge is much more than "just" his professional environment. His passion for data takes him further than his daily routine. Indeed, he applies classic and modern data science algorithms to predict the COVID-19 spread or the winner of the UEFA Euro 2020 soccer tournament. The beauty is indeed this versatility, this ability to apply data expertise to every aspect of life. Data is data after all: Everything can be converted to numbers, inspected, and predicted.

Rosaria: *Hi Dennis, tell us about your professional self and what you do in your job.*

Dennis: I work for a big Telco in Switzerland as Head of Reporting and Data Management. We measure the performance of the sales channels and do everything from data integration and data blending through to creating dashboards, reports and so on.

Rosaria: *Do you use KNIME in your everyday work?*

Dennis: Yes. We use KNIME Analytics Platform mainly as an ETL tool and we also use KNIME Server to automate our workflows. We have a lot of daily reports that have to be ready in the morning, so we are happy to have such a solution. We also combine KNIME with other tools - mainly with Tableau. But I always say to the stakeholders: Tableau is just the car body, KNIME is the real engine.

Rosaria: *Tell us about the biggest challenge you've had to solve in your professional life.*

Dennis: The biggest challenge was, and will always be, explaining the story behind the data to clients and stakeholders. Data science is all about storytelling. You need strong communication skills and a good visualization of the data. As they say: "A picture is worth a thousand words."

Rosaria: *There is a high demand for data scientists in the employment market and yet data science is still not part of the traditional education system. People often have to learn skills themselves, sign up for online courses, and read the literature. Which books would you recommend to people who are wanting to learn new skills? You're also writing your own book. Aren't you?*

Dennis: I very much like your book, Rosaria, about Codeless Deep Learning. It's easy to understand and very useful. But yes, I just started writing my own book with the title "*KNIME Solutions for Real World Applications*". It is a compilation of real-world cases solved with KNIME together with other tools. I also read blog articles regularly to keep up to date. For example the journals *Towards Data Science*, *Low Code for Advanced Data Science* on *Medium*, and everything that can be applied in data science.

Rosaria: *What is your advice for aspiring data scientists?*

Dennis: Whenever somebody asks me this question I ask back: What are your hobbies? And if data science is not your hobby – you have to change hobbies! I think that "learning" alone is not enough. You must live it and love it to succeed.

Rosaria: *What skills are most underestimated by candidates but a plus on the job?*

Dennis: To keep cool in stress situations and never forget that it's a job and not a game. So although I believe strongly that data science must become your hobby [if you want to be successful], the job is not a hobby. A lot of the time you will be doing things that you don't like but that are still very important.

Rosaria: *Let's talk about how you've used your expertise with data outside of your professional life. On June 10, one day before the start of the UEFA Euro 2020 soccer tournament, you managed to correctly predict what the final game would be: England vs Italy. How did you do that?*

Dennis: I asked Maradona 🙂. No, I used a fairly well-known approach in the sports betting industry. I used a linear regression model to calculate the ratings of the teams.

Rosaria: *But the model predicted that England would have been the winner.*

Dennis: Not exactly, the model just calculated the power ratings of the teams before the tournament. So, England, Italy, and Spain were in the top three. By the way, all three teams made it to the semi-finals. Though Denmark was a surprise. Nobody saw that coming, not even the soccer experts. It's important here to notice that soccer is a game with a lot of randomness. Scores are very sparse and a final match can be decided by a penalty shootout. Therefore, it's not always possible to make a precise forecast. All in all, I think my power ratings were good, because the past games that I used as training data, reflected the strength of the teams well.

Rosaria: *So, it was quite a simple model. No deep learning?*

Dennis: Exactly, no deep learning, no strong GPUs, just a simple linear regression model.

The key factor here was to include domain knowledge. For example, the home field advantage is a very important factor in soccer. Even without spectators – it's still there. Another point is that you have to take just the right portion of data - the portion that's best to train the model. For example: after a big tournament, often the coaches change and the players change too. So it's better to filter out those games happening right before such changes are made in the teams.

Rosaria: *What about another project of yours where you predicted the spread of COVID19 worldwide and country by country?*

Dennis: My motivation behind this project was to forecast the evolution of the COVID19 pandemic. In the beginning just for China and then for every country in the world. Yes, I wanted to answer the question: When will this pandemic be over?

Rosaria: *Which model did you use?*

Dennis: The evolution of a pandemic is like a growth process. At the beginning it's an exponential function, then it changes to a sigmoidal curve. This is best described by a logistic function. However, then, when several waves followed, this simple approach does not work anymore and another approach is needed. I found out that Rockefeller University had already used a method called *Loglet Analysis* (also called wavelets) in the late 90s to forecast the evolution of multiple overlapping logistic functions.

Rosaria: *Was this project with KNIME Analytics Platform or with Jupyter?*

Dennis: I used KNIME Analytics Platform together with Jupyter. Indeed, you can call Python from KNIME. So the data was prepared in KNIME Analytics Platform but the loglet model was calculated in Jupyter with the scipy-package.

Rosaria: *Thank you, Dennis, for this insight into your job and your other projects. How can data scientists in the audience get in touch with you or your work?*

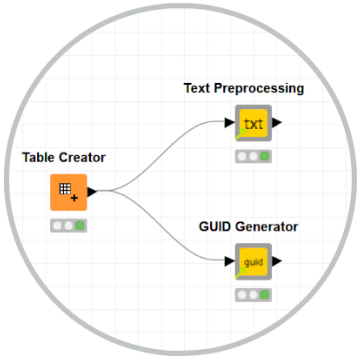
Dennis: I've written articles on Medium and I'm also posting some interesting stuff on Linkedin and Twitter. I have also created a Facebook group Data Science with Yodime. And on my Youtube channel you will find some interesting videos about Data Science, and not only. Of course all my workflows can be downloaded from my public space on the KNIME Hub.

This article was first published on our KNIME Blog. Find the original version here.

Watch the original interview with Dennis Ganzaroli on YouTube: "My Data Guest – Ep 1 with Dennis Ganzaroli".

SJ Porter was nominated KNIME Contributor of the Month for October 2020. He was awarded for his GUID Generator and Text Processing components, which are displayed in the figure on the right. As of 28.09.2022 both components total 15.564 downloads. The first component was, finally, a Globally Unique Identifier (GUID) generator for KNIME. This component is useful for creating a unique key that is not based on Row ID. The Text Preprocessing component uses extremely fast regex-based text processing functions to remove specific types of characters from a String column and normalize the data as much as possible without over-processing. This component eliminates the need to convert text to a Document type in order to preprocess it.

At the time of the award, SJ was Data Science Team Lead in the consumer reporting industry. In January 2021, he joined KNIME as a Data Scientist, making him an official KNIMEr. He currently works as a Site Reliability Engineer focusing on the KNIME Edge and KNIME Community Hub architecture. Besides building workflows and developing components, SJ also helped creating the Low Code for Advanced Data Science journal on Medium and currently serves as one of the editors.



Visit SJ's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `sjporter`).

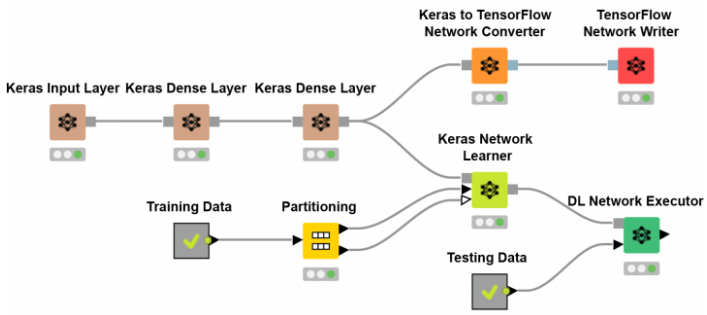
KNIME Analytics Platform is the “killer app” for machine learning and statistics

A free, easy, and open-source tool for all things data? Yes, please!

Author: SJ Porter

> ***Disclaimer:***
>
> *I now work for KNIME as a Data Scientist / Data Engineer! I wrote this article roughly one year before I applied to KNIME. The article title and contents were (and still are) my personal opinion.*
>
> *– Steven “SJ” Porter*

If you work with data in any capacity, go ahead and do yourself a favor: download KNIME Analytics Platform right here.

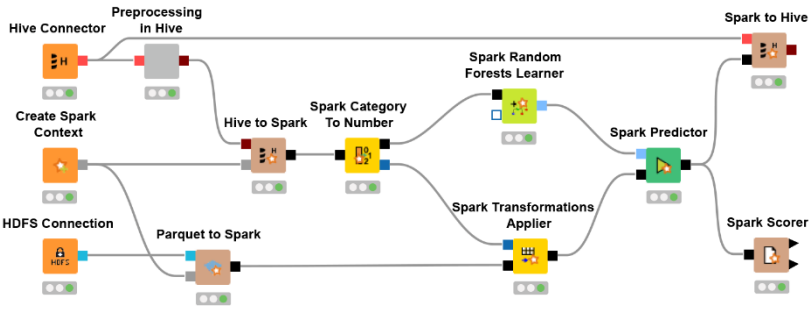


More data science, less slamming of the mouse and keyboard.

What is KNIME Analytics Platform?

KNIME Analytics Platform is the strongest and most comprehensive free platform for drag-and-drop analytics, machine learning, statistics, and ETL that I’ve found to date. The fact that there’s neither a paywall nor locked features means the barrier to entry is nonexistent.

Connectors to data sources (both on-premises and on the cloud) are available for all major providers, making it easy to move data between environments. SQL Server to Azure? No problem. Google Sheets to Amazon Redshift? Sure, why not. How about applying a machine learning algorithm and filtering/transforming the results? You’re covered.



Big data? Small data? Doesn't matter too much provided that your computer has a solid CPU and 16+ GB RAM.

It's also worth mentioning that the community is particularly robust. The developers and product owners at KNIME are a blast to work with, and the forums are surprisingly active.

I could dive into a quick start tutorial or show off some of the more advanced capabilities, but it's honestly very intuitive to use. The documentation is integrated directly into the desktop client, the UI is dirt simple, and the UX is a great blend between complexity/customizability (when needed) and user-friendliness.

What can I do with KNIME Analytics Platform?

KNIME Analytics Platform is well-suited for the following:

- ETL processes (moving data around from here to there and cleaning it up)
- Machine learning
- Deep learning
- Natural language processing
- API integration
- Interactive visual analytics (somewhat of a beta feature)

What's the catch?

KNIME Analytics Platform is 100% free. The documentation is readily available at knime.com and there are a ton of free extensions to the platform. So long as you're the one clicking "*Go*" every time the process runs, you won't have to pay a dime. Ever. That's the benefit of working with a software package that has roots in academia.

If you want to schedule workflows, KNIME Server is the premium offering that allows for process scheduling among other features. The basic tier, KNIME Server Small, is around $1.67 per hours on AWS. If you host KNIME Server on an EC2 instance and

schedule a cron job to turn the instance on and off, it's an *extremely* cost-effective option.

Higher tiers of KNIME Server allow for use of the REST API and WebPortal. Those features allow you to automate workflow deployment, execute workflows remotely from another service, and create an interactive hub for users. The ability to automate workflows makes KNIME Server Medium an attractive option. If you purchase the highest tier (KNIME Server Large) with the BYOL option, you gain the ability to host multiple instances of the server using the same license.

How do I learn KNIME Analytics Platform?

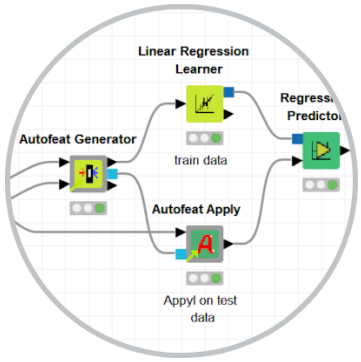
Their learning page is here, but if you're data-savvy then simply download the platform and try out a thing or two. So long as you get to the point in which some data is loaded up into the application, the rest is intuitive and the documentation is accessible from within the application. Drag nodes from the "*Node Explorer*" onto the view. There is a search bar… try searching for "*Excel*" and go from there.

This article was first published in the Towards Data Science Journal on Medium. Find the original version here.

Ashok Harnal was nominated KNIME Contributor of the Month for November 2021. He was awarded for his Auto Categorical Features Embedding, Autofeat Generator and Autofeat Apply components. As of 03.10.2022 the components total 3.171 downloads. These are just two of many components Ashok developed which provide good examples on how to bundle and share Python scripts without dependency issues. All his components are accessible on his KNIME Community Hub profile with instructions on how to use them in practice. Visit the Community Component Highlights section on the Verified Component web page to learn more.

Ashok has several years of experience in teaching Big Data technology and is currently a Professor at FORE School of Management in New Delhi. His areas of interest revolve around Big Data Systems, including Machine Learning, Deep Learning, Big Data storage systems, and Graph Databases.

Visit Ashok's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `ashokharnal`).



A Component Series for Automated Feature Engineering

All Good Things comes in Threes – Ashok's Series of Feature Engineering Components

Authors: Elisabeth Richter & Corey Weisinger

As mentioned above, Ashok is a very active KNIME Community member who likes to develop components. He was awarded for his *Auto Categorical Features Embedding*, *Autofeat Generator* and *Autofeat Apply* components. In this article, let's have a closer look at these three components.

The Auto Categorical Features Embedding Component

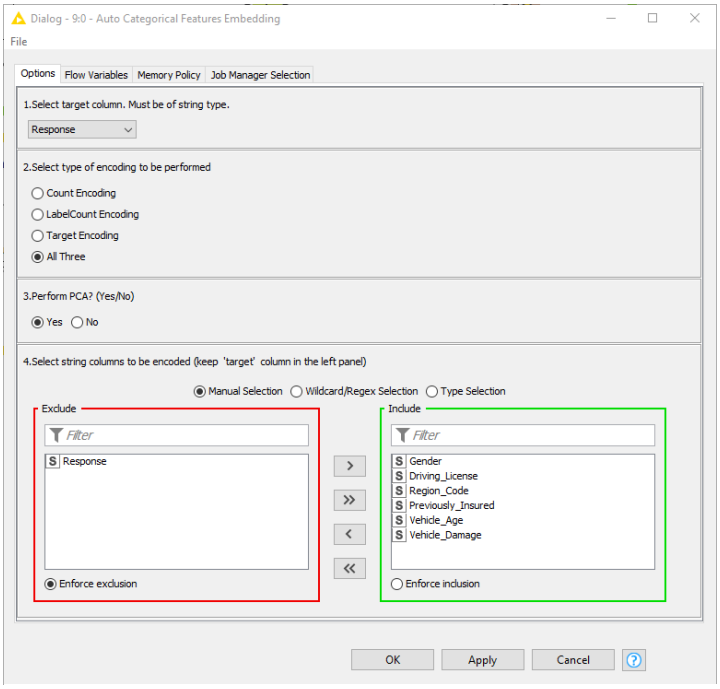
The *Auto Categorical Features Embedding* component encodes categorical string features into numeric features. In general, Feature Encoding is a technique often used in machine learning. It refers to the process of transforming a categorical (i.e., non-numerical) variable into a continuous one (i.e., a variable that can take on any value between any two points). This component developed by Ashok automates this feature encoding process.

The Auto Categorical Features Embedding component for categorical string encoding.

As inputs, the component takes training and test data. The configuration window of the component is displayed in the following figure and allows the specification of the target column, the string column(s) to be encoded, which type of categorical encoding should be performed, and whether to perform PCA or not. The features to be created are either count encoding, ranked label count encoding, target encoding, or all three of them. So, with this component, multiple categorical variables can be specified in one go.

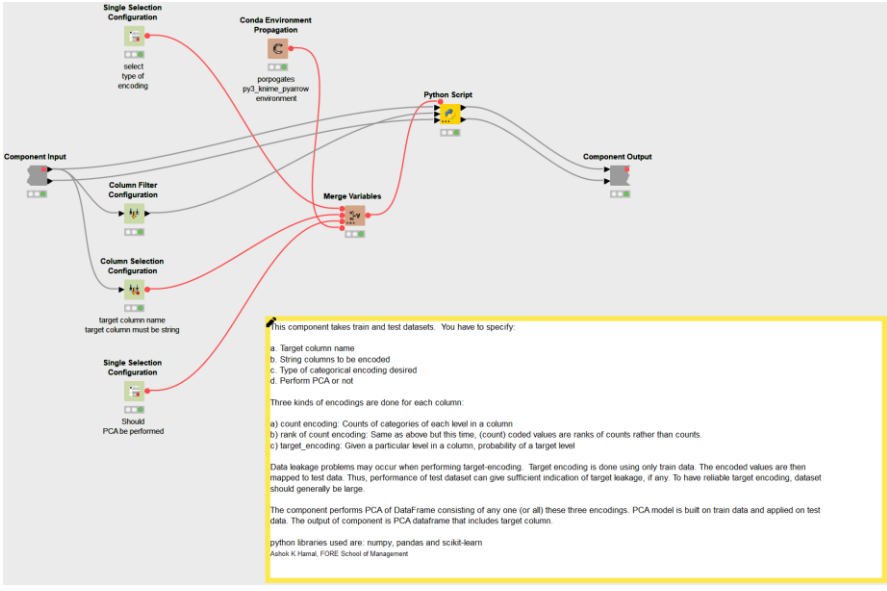
To avoid data leakage while performing target encoding, the encoding is performed using only training data. The encoded values are then mapped to the test data. To get a reliable target encoding, the dataset should be sufficiently large enough.

Performing a PCA is optional. If "*yes*" is selected in the configuration window, the component performs a PCA on the dataset consisting of either of these three or all three encodings along with other numeric features that are already present in the data. The PCA model is built on the training data and then applied to the test data.



The configuration dialog of the Auto Categorical Feature Embedding component.

The output of the component are two data tables – the training data and the test data. Each table includes either the principal components or the encoded columns along with the already present numeric columns and the target column.



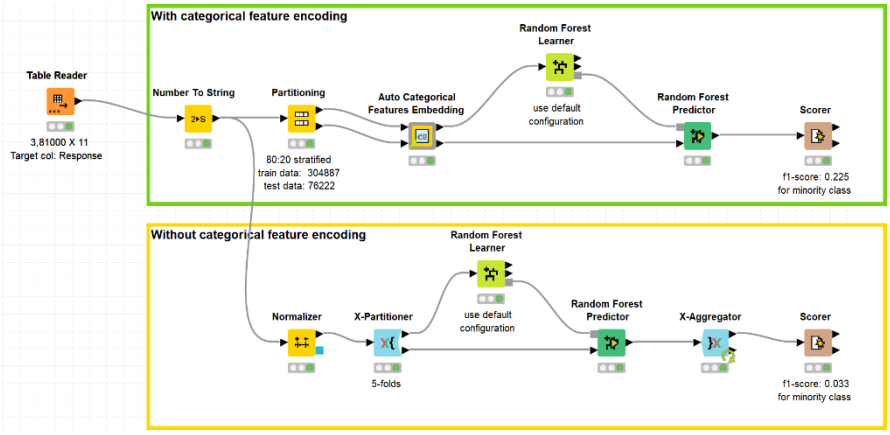
The inside of the Auto Categorical Feature Embedding component.

Inside the component (see the image above), Ashok added a *Python Script* node that uses the Python libraries *numpy*, *pandas*, *scikit-learn*, and *pyarrow*. When executing the component, Ashok makes sure that KNIME automatically installs the required Python packages via the *Conda Environment Propagation* node.

The image below shows a workflow that demonstrates the use of the *Auto Categorical Features Embedding* component. It can be found on Ashok's Hub profile (*HealthInsurance Cross-Sell--Categorical Feature Engineering*). The data in this example is some personal data of the customers of an insurance company. The company currently provides a health insurance to their customers. Using the existing data, they would now like to predict whether the current customers would also be interested in a new vehicle insurance.

The target column is the "*Response*" column, and the dataset is highly imbalanced with only ~12% of the customers being willing to buy the vehicle insurance (response value equal to 1).

Now two models are trained using identical configuration of the learner nodes: one with categorical feature encoding (top) and one without categorical feature encoding (bottom).



An example workflow that demonstrates the use of the Auto Categorical Features Embedding component.

The respective *Scorer* nodes show that the F1 score as well as the Cohen's Kappa are much better when applying feature encoding. The top *Scorer* node reports an F1 score of 0.2 and a Cohen's Kappa of 0.138, whereas the bottom *Scorer* node reports an F1 score of only 0.034 and a Cohen's Kappa of 0.025 respectively.

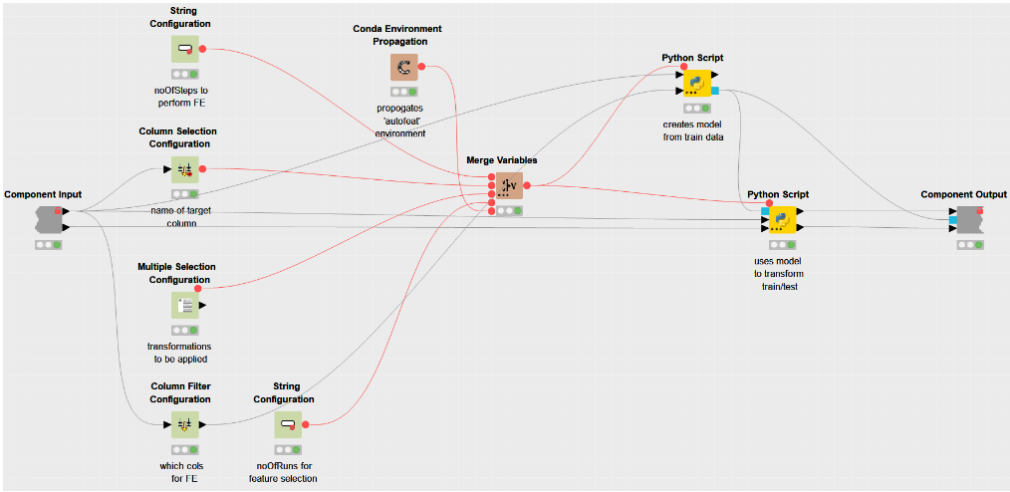
Based on the results of this example the purpose of feature encoding is underlined: to improve classification results.

The Autofeat Generator and the Autofeat Apply Components

The Autofeat Generator component.

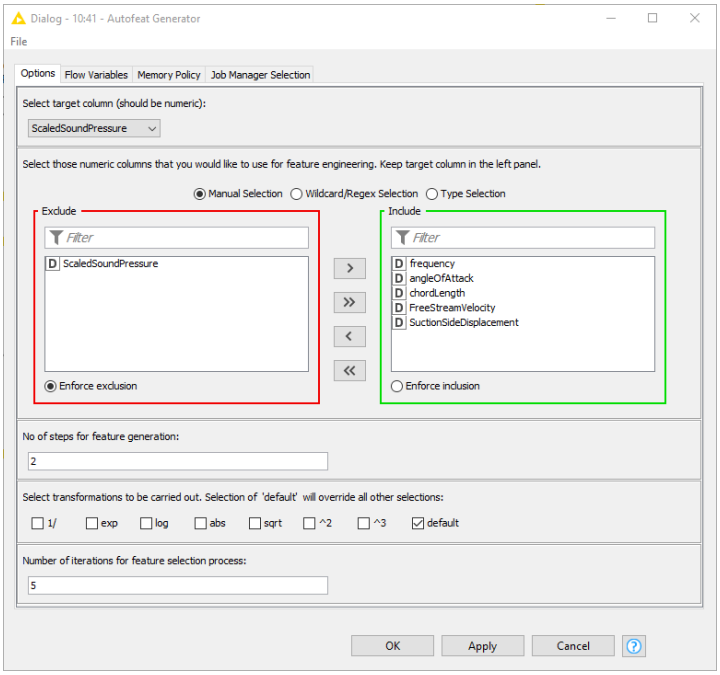
Ashok seems to be a Python enthusiast – at least judging on how frequently he uses the *KNIME Python Integration* when developing components. The two components introduced in the following section, the *Autofeat Generator* and the *Autofeat Apply*, both bundle Python scripts to run the *autofeat* library along with *numpy* and *pandas*. Both components also contain the *Python Script* node and the *Conda Environment Propagation* node, similar as the *Auto Categorical Features Embedding* component.

The *Autofeat Generator* component generates new features whose use are directed towards building linear models. It takes as input two data tables – training and test data – and builds a model using the training data. In a second step, the trained model is then applied to generate additional columns based on those provided. The component outputs the two data tables as well as the trained model. The inside of the component is shown in the figure below.



The inside of the Autofeat Generator component.

The configuration dialog of the *Autofeat Generator* component is displayed in the following. It allows the specification of the target column (which should be numeric), which column(s) to be included in the process, the number of steps for the feature generation (which is an important parameter as the higher the number of steps, the higher the number of features, but also a higher chance of overfitting), the number of runs for the feature selection process, and the transformation(s) to be applied to the features ("*default*" overwrites all other selections).

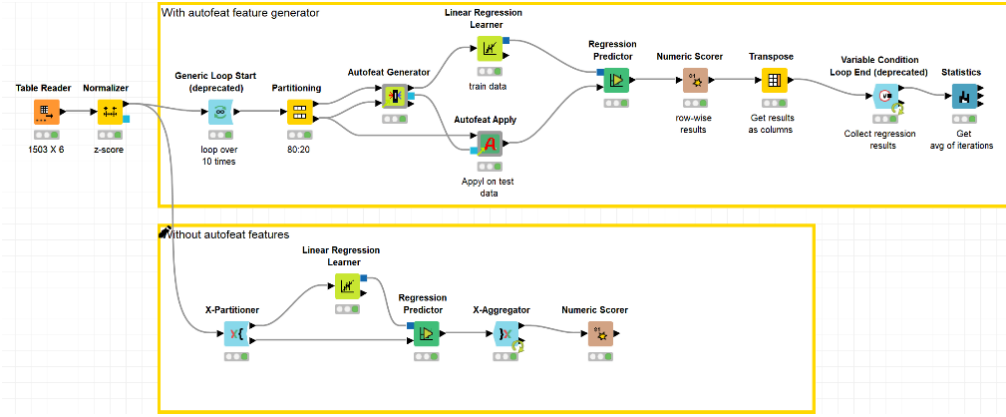


The configuration window of the Autofeat Generator component.

The *Autofeat Apply* component has a similar functionality, however, it does not build a model. It takes as input a data table and an "*autofeat*" model (this could be a model created by the *Autofeat Generator* component) and the outputs a dataset with generated features concatenated to the input dataset.

The Autofeat Apply component.

This automated feature engineering comes in handy when optimizing the performance of machine learning algorithms such as logistic or linear regressions.



An example workflow that demonstrates the use of the Autofeat Generator and Autofeat Apply components.

Check out the example workflow above to view in detail how both components work. This workflow is also available on the KNIME Community Hub on Ashok's Hub profile (*Airfoil Self-Noise prediction using Autofeat Generator*).

The dataset used in this example is the "*Airfoil Self-Noise Data Set*" which is a NASA dataset obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. In the workflow, Ashok builds two linear regression models, one using generated features (top) and one without autofeat features (bottom). The two statistics outputted by the *Numeric Scorer* nodes show, the model built using generated features is superior to the model built without such features.

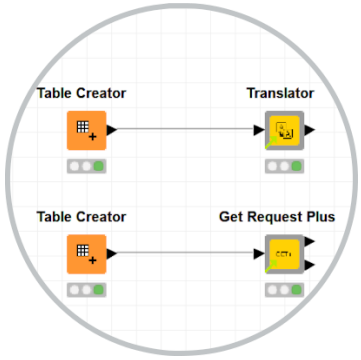
Data Science Use Cases

In this section we have collected all the articles that show how multifaceted KNIME Analytics Platform can be applied. Some of our COTMs created thrilling, fascinating, and creative use cases that demonstrate that KNIME is a tool for (almost) everything! KNIME can even be used to dive into the magical world of dwarves and dragons... The category “Data Science Use Cases” features our enthusiastic creators:

- **Armin Ghassemi Rudd**
  – Managing Director & Principal Data Scientist *@Intellacct*
- **Angus Veitch**
  – Data Analytics Consultant *@Forest Grove*
- **Philipp Kowalski**
  – Digital Enablement Agent *@Siemens Industry Software GmbH*
- **Tosin Adekanye**
  – Data Scientist *@Qatar Financial Centre Regulatory Authority (QFCRA)*
- **Paul Wisneskey**
  – Director of Engineering *@BigBear.ai*
- **John Emery**
  – Principal Consultant *@phData*

Armin Ghassemi Rudd was nominated KNIME Contributor of the Month for February 2021. He was awarded for his Translator and Get Request Plus components, which are displayed in the figure on the right. As of 28.09.2022 both components total 6567 downloads. The Translator component uses Google Translate to translate any input text from/to supported languages. The Get Request Plus component adds the "Retry" option to the GET Request node. Besides creating awesome components, Armin has been a long term and very active contributor to the KNME Forum.

Armin is a Data Science enthusiast who has a passion for education (both learning and teaching) and enjoys making sense of data. He is currently Managing Director and Principal Data Scientist as Intellacct of which he is also the founder. His fields of interests are widely spread and range from consumer behavior to astroparticle physics. Armin holds a Master of Science in Information Technology Management with the subfield Business Intelligence from the University of Teheran. After completing his master's degree, he stayed connected to the university by instructing a few optional Data Science Courses.



Visit Armin's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `armingrudd`).

Building a CV Builder with BIRT in KNIME — Part 1

Author: Armin Ghassemi Rudd

In our last article *Build your CV based on LinkedIn profile with BIRT in KNIME* on KNIME Blog we introduced a CV Builder based on LinkedIn profile. Here we will show you how to build the workflow and the report with BIRT in KNIME.

Creating the KNIME workflow project

Let's create a workflow project. Open KNIME Analytics Platform and create a new project and name it "*CV_Builder*". Now before doing anything else in KNIME, go to the workspace directory and find the "*CV_Builder*" folder. Inside this folder, create a new folder named "*data*" and move the downloaded LinkedIn data folder "*LinkedInDataExport*" (explained in our last article on the KNIME Blog) to the folder named "*data*" under the workflow directory (*CV_Builder*) in your workspace.

We need to save our photo with the name "*personal_photo.png*" in the "*data*" folder directory. Dimensions should be 496*516, or we will need to configure the image item in BIRT a bit different from what we do here. Then download the "*background.png*" file and move it to this folder.

Building the workflow

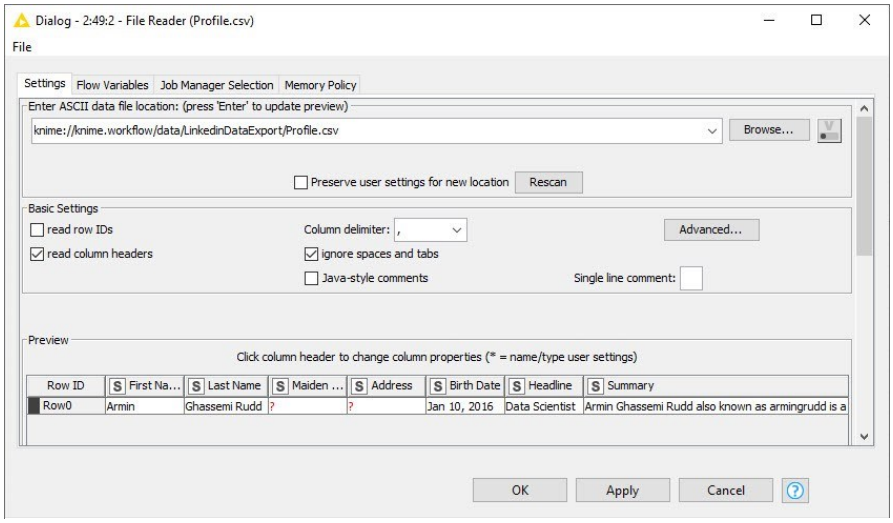
Now, we need to read the CSV files we have downloaded and moved to the "*data*" folder. We only need these 11 files: "*Certifications.csv*", "*Education.csv*", "*Courses.csv*", "*Email Adresses.csv*", "*Endorsement Received Info.csv*", "*Languages.csv*", "*Positions.csv*", "*Profile.csv*", "*Projects.csv*", "*Recommendations Received.csv*", "*Skills.csv*".

Since we are going to pass the outputs to the *Data to Report* node, we need 11 *File Reader* nodes. Let's do it one by one.

Create a metanode and name it "*Profile*". Double click the metanode to go inside. Add a *File Reader* node and go to the configurations of the node. Add this path as the file location:

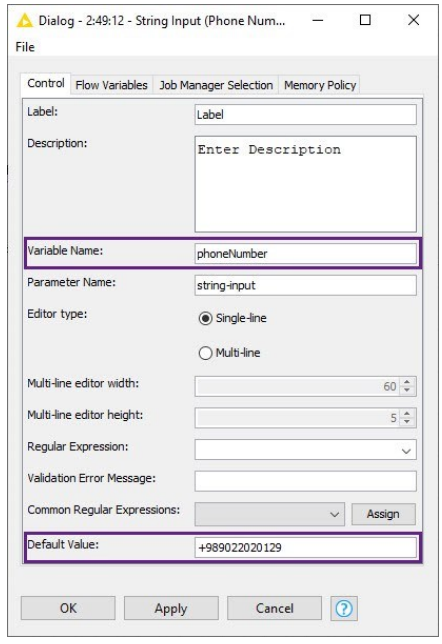
```
knime://knime.workflow/data/LinkedinDataExport/Profile.csv
```

"*knime://knime.workflow/*" refers to the current workflow directory. Now Check the "*read column headers*" option and press "*OK*".



The configuration window of the File Reader node.

Before executing the node, we can add a *String Input* node before the *File Reader* node to add our phone number as a flow variable. In the configuration window of the node, set the "*Variable Name*" to "*phoneNumber*" and enter your phone number in front of the "*default value*" option.

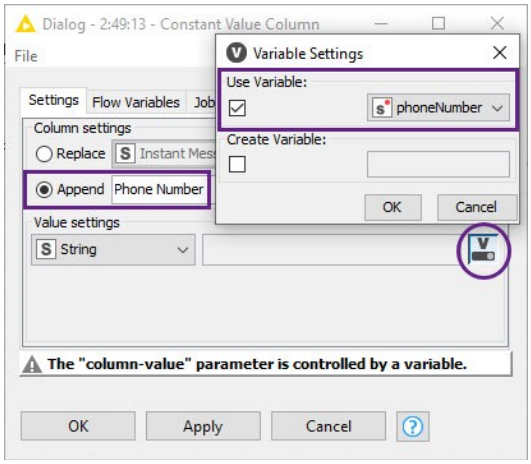


The configuration window of the String Input node.

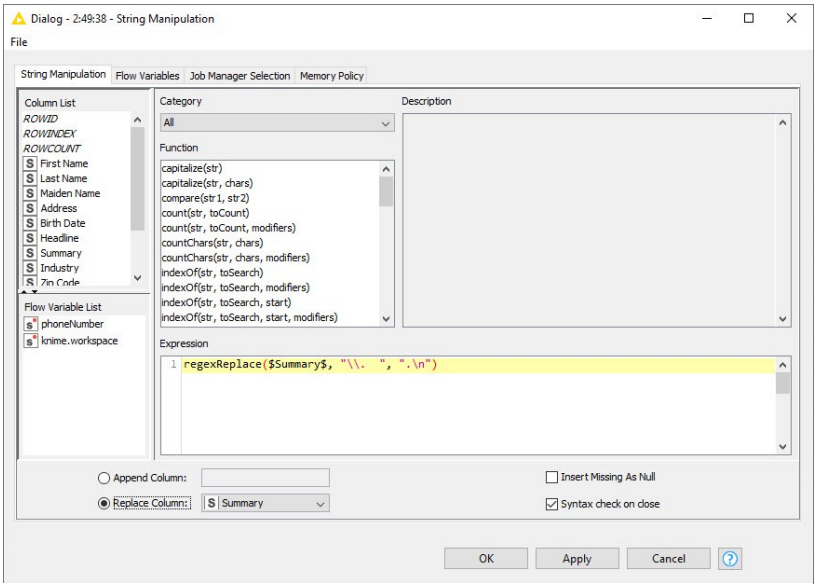
Connect the output port of the *String Input* node to the variable input port of the *File Reader* node and execute the nodes. We need to manipulate the summary in our table so that the new lines will appear as expected. In the current table, the summary has no new lines even if we have several paragraphs in our original summary on the LinkedIn website. Those new lines have been replaced with double spaces. So, we use a *String Manipulation* node to correct this. We can apply this expression:

```
regexReplace($Summary$, "\\. ", ".\n")
```

Now, we need our flow variable "*phoneNumber*" to be added to the data table. The *Constant Value Column* node does the task for us. Append a new column named "*Phone Number*" and assign the "*phoneNumber*" flow variable to the "*Value settings*" option.



The configuration dialog of the Constant Value Column node.



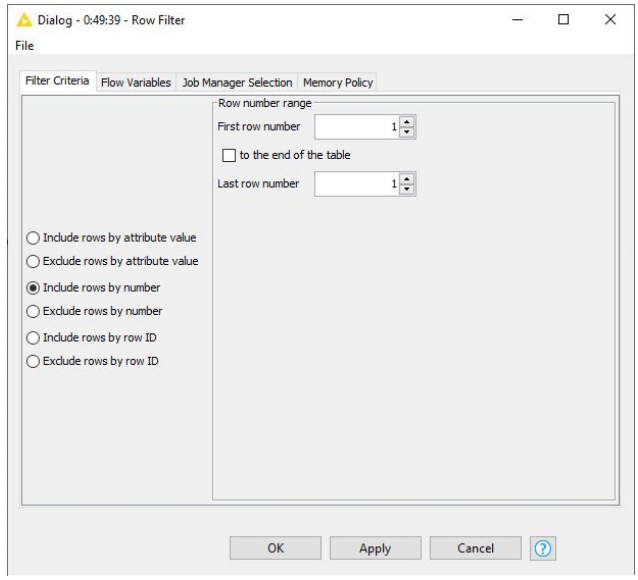
The configuration window of the String Manipulation node.

Next, we want our Email address from "*Email Addresses.csv*" to be added to the table as well. So, we use another *File Reader* node and add this path:

```
knime://knime.workflow/data/LinkedinDataExport/Email%20Addresses.csv
```

Again, we need to check the "*Read column headers*" option.

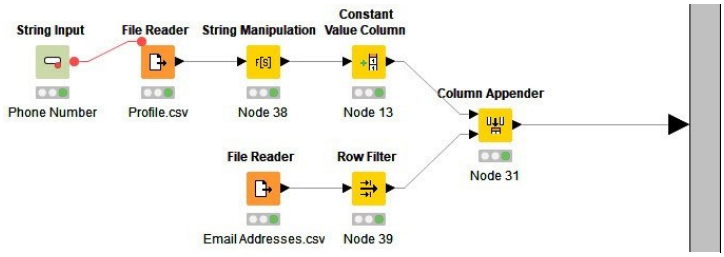
Since we may have several Email addresses, a *Row Filter* node will be used to keep the first one in the list. Select "*Include rows by number*" and input the number "*1*" for both "*First row number*" and "*Last row number*".



The configuration dialog of the Row Filter node.

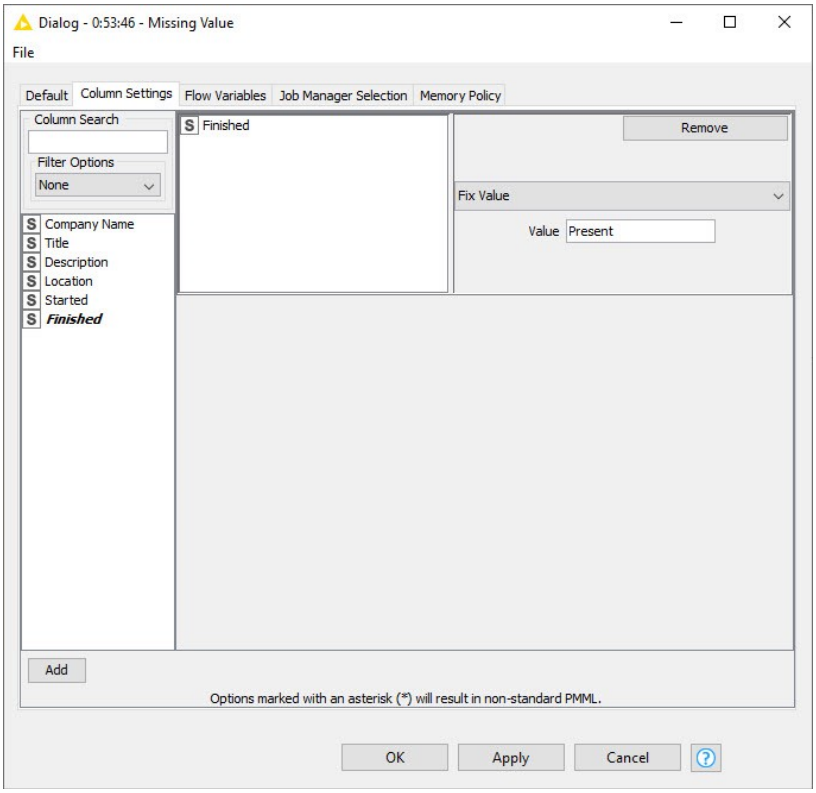
A *Column Appender* node will now add our Email address to the main table.

This is what we have now in the *Profile* metanode:



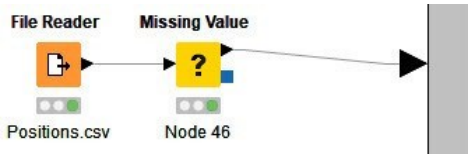
The inside of the Profile metanode.

Now, let's go to the next metanode: *Positions*. A *File Reader* node to read the "*Positions.csv*" file (with column headers) and a *Missing Value* node will replace the missing values in the "*Finished*" column with "*Present*" value. When the "*Finished*" value is missing, it means we still have the position. The LinkedIn website shows the term "*Present*" but in the downloaded data, it appears as a missing value.



The configuration dialog of the Missing Value node.

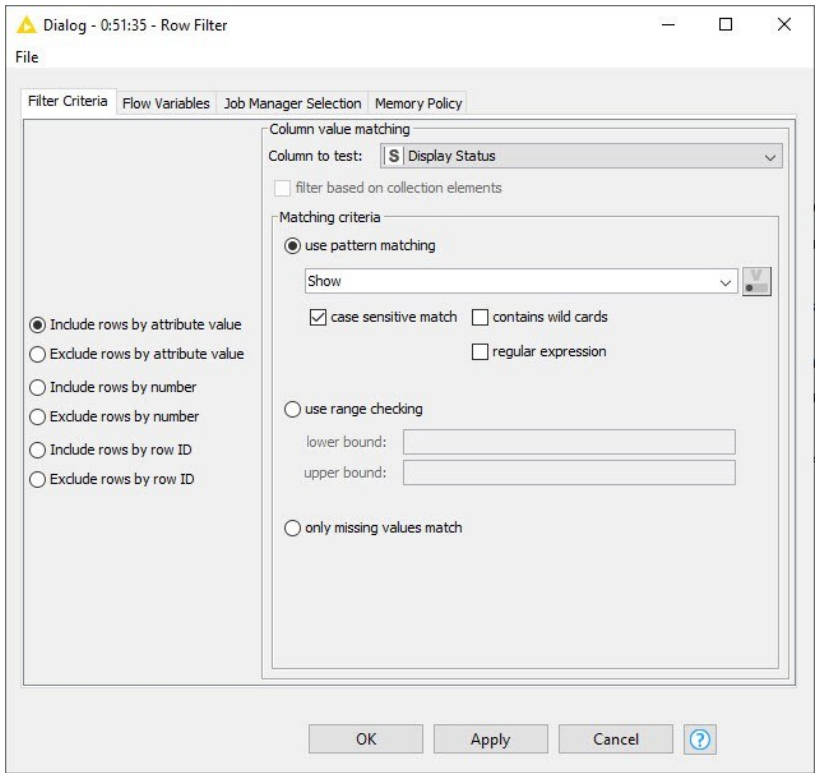
And that's it for the *Positions* metanode.



The inside of the Positions metanode.

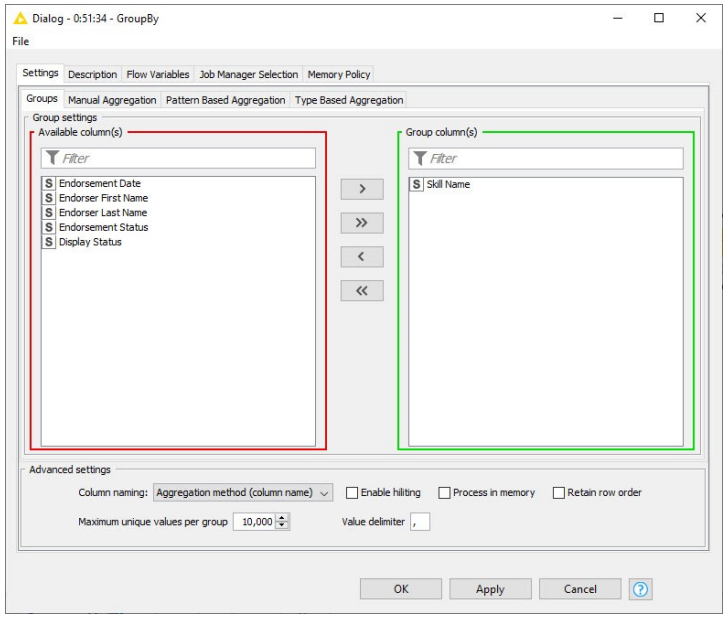
The same method can be used for the *Projects* and the *Certifications* metanodes where we read the "*Projects.csv*" and the "*Certifications.csv*" files.

The next metanode is the one to read the "*Skill.csv*" and the "*Endorsement Received Info.csv*" files. We use a *File Reader* node for each one of them. Let's use a *Row Filter* node after reading the endorsements to include the ones we have chosen to be shown in our LinkedIn profile:



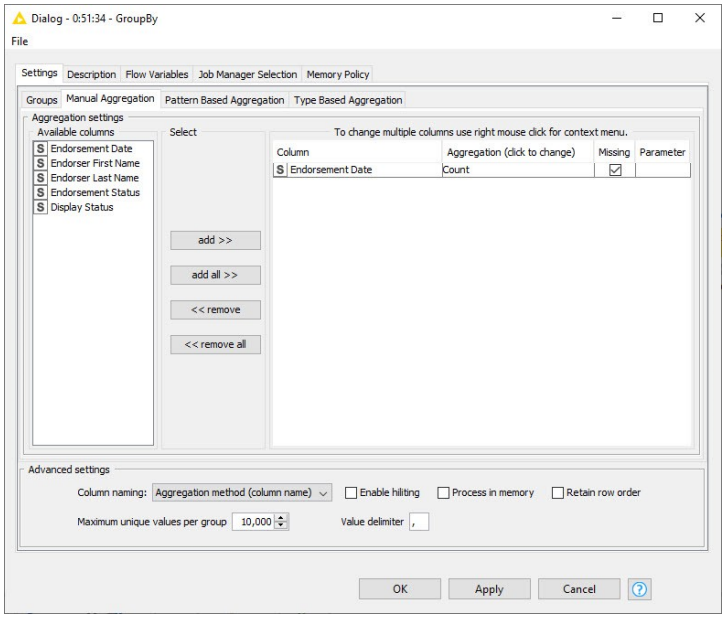
The configuration dialog of the Row Filter node.

As you can see in the image above, we have included the rows where the value for the "*Display Status*" is "*Show*".

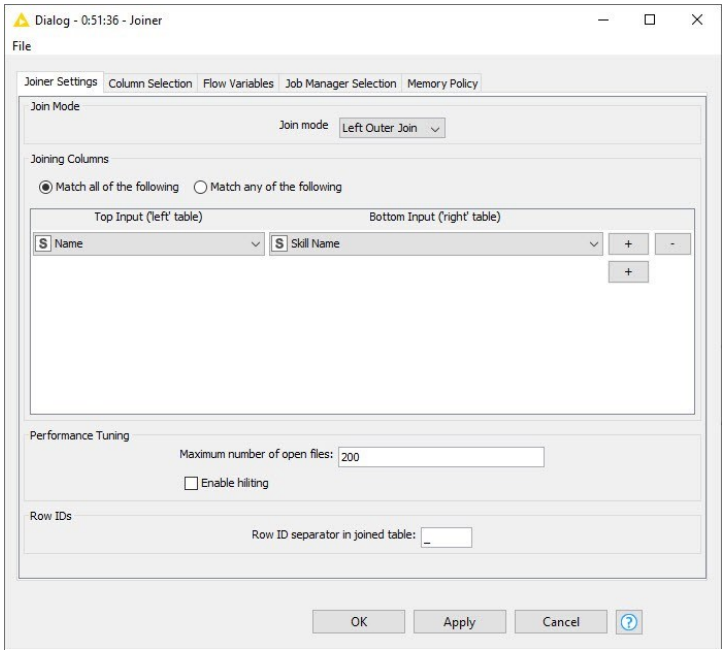


The Groups tab in the configuration dialog of the GroupBy node.

Next, the *GroupBy* node will count the endorsements for each skill. We select the "*Skill Name*" column as the grouping column, and any of the other columns can be selected with the "*Count*" aggregation function while the option for counting the missing values is checked.



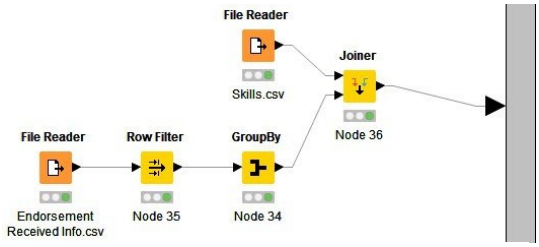
The Manual Aggregation tab in the configuration dialog of the GroupBy node.



The configuration dialog of the Joiner node.

Now, a *Joiner* node will join our skills and endorsements to each other. We can use the "*Name*" column from the Skills dataset and the "*Skill Name*" from the endorsements dataset to join these two tables.

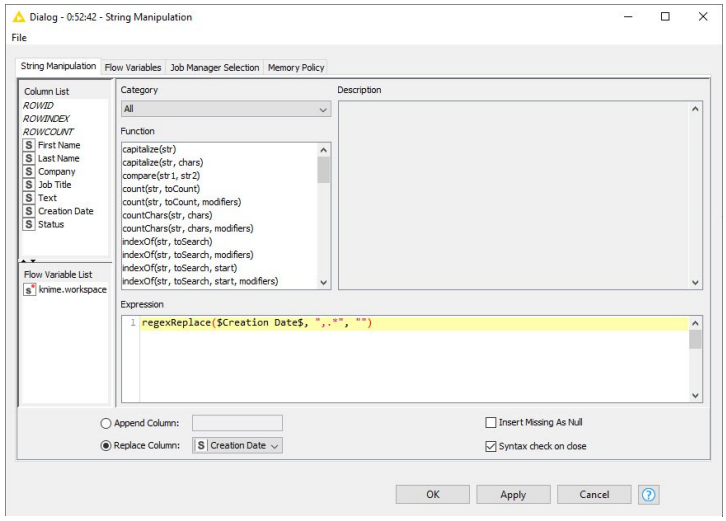
The *Skills* metanode is finished as well.



The inside of the skills metanode.

In the *Recommendations* metanode, we read the "*Recommendations Received.csv*" first. Then a *String Manipulation* node to remove the time from the "*Creation Date*" column should be used in which we apply the expression below and replace the "*Creation Date*" column.

```
regexReplace($Creation Date$, ",.*", "")
```

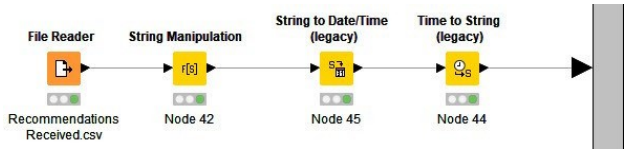


The configuration dialog of the String Manipulation node.

Next, a *String to Date/Time (legacy)* node to convert the "*Creation Date*" column to legacy Date&Time where we have month names like Jan, Feb, … and then we can convert it back to string and keep the month name and the year by using a *Time to String (legacy)* node.

In the configurations of the *String to Date/Time (legacy)* node, we replace the "*Creation Date*" column and use this date format: *MM/dd/yy*. In the *Time to String (legacy)* node,

we use *MMM yyyy* as the date format and replace the column. Now, our *Recommendations* metanode looks like this:



The inside of the Recommendations metanode.

The last metanode is the *Language* metanode, where we read the "*Languages.csv*" file. Here we need to create a dictionary to convert our proficiency in each language to a number scaling from 1 to 5. So we use a *Table Creator* node to enter the proficiency levels and the corresponding numbers:

Manually created table - 0:55:46 - Table Creator (Language lvl number)

File Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	level	number
Row0	Native or bilingual proficiency	5
Row4	Full professional proficiency	4
Row1	Professional working proficiency	3
Row3	Limited working proficiency	2
Row2	Elementary proficiency	1

The output of the Table Creator node shows a table containing proficiency levels and corresponding numbers.

We are going to use a *Rule Engine (Dictionary)* node to apply the numbers to our languages based on the proficiency levels. So we need to create the rules. To do that, we use a *String Manipulation* node after the *Table Creator* node and use the expression below while replacing the "*level*" column.

```
join("$Proficiency$ = \"", $level$, "\"")
```

The output is:

Appended table - 0:55:47 - String Manipulation

File Hilite Navigation View

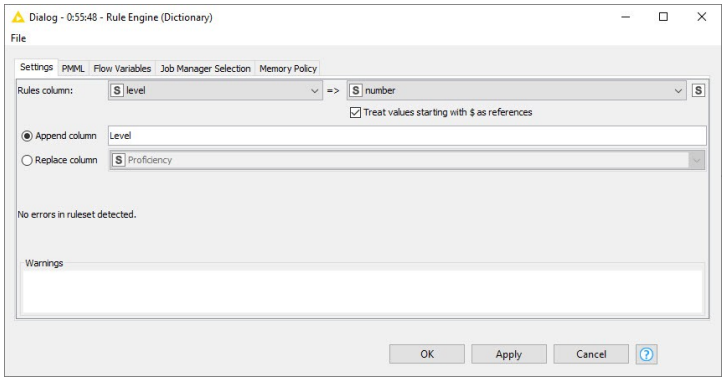
Table "default" - Rows: 5 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	level	number
Row0	$Proficiency$ = "Native or bilingual proficiency"	5
Row4	$Proficiency$ = "Full professional proficiency"	4
Row1	$Proficiency$ = "Professional working proficie...	3
Row3	$Proficiency$ = "Limited working proficiency"	2
Row2	$Proficiency$ = "Elementary proficiency"	1

The output of the String Manipulation node after applying the following expression join("$Proficiency$ = \"", $level$, "\"").

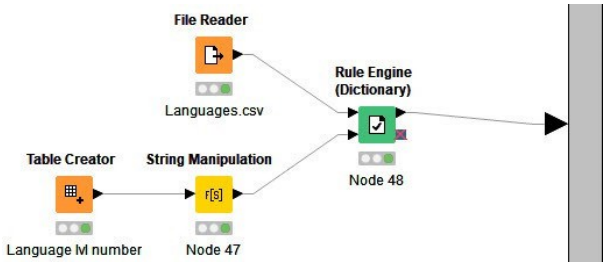
Now, the output of the *String Manipulation* goes to the bottom input port of the *Rule Engine (Dictionary)* node, and the output from the *File Reader* goes to the top port.

In the configuration window of the *Rule Engine (Dictionary)* node, the "*Rules column*" is the level column, and the results are in the "*number*" column. We append the output in a new column named "*Level*".



The configuration dialog of the Rule Engine (Dictionary) node.

The *Languages* metanode now looks like this:

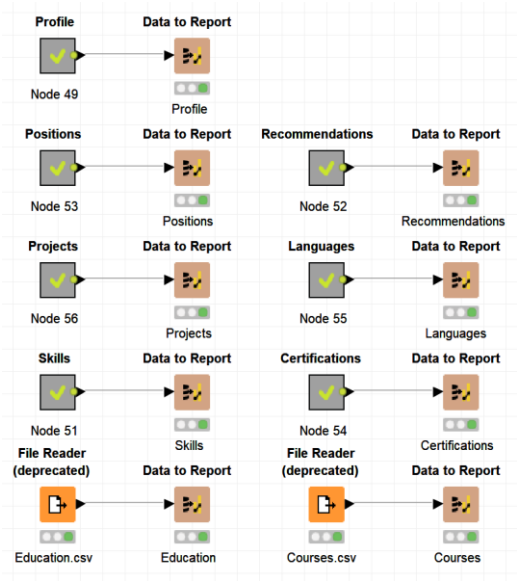


The inside of the Languages metanode.

There are two more files we need to add: the "*Education.csv*" and the "*Courses.csv*" files. We use two *File Reader* nodes to read these files, and we are finished with reading our datasets.

Now we use one *Data to Report* node for each metanode and the last two *File Reader* nodes.

Do not forget to use node comments for the *Data to Report* nodes since the names for the data tables in BIRT will be the same as these comments. Now execute the workflow and save it. The workflow is completed, and now we can start working on building our CV in BIRT.



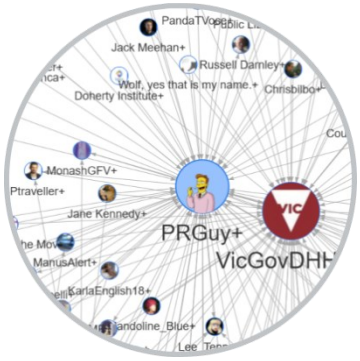
The overview of the CV Builder workflow.

This article was first published in the Act of Intelligence Accretion Journal on Medium. Find the original version here. The corresponding CV_Builder workflow can be found on the KNIME Community Hub in Armin's public space.

You can continue reading part 2 on Medium at Building a CV Builder with BIRT in KNIME – Part 2.

Angus Veitch was nominated KNIME Contributor of the Month for November 2020. He was awarded for his article on his TweetKollidR workflow where he describes a KNIME workflow for creating text-rich visualizations of Twitter data around a given hashtag. The image on the right shows part of the network visualization of users who tweeted about the Melbourne lockdown. Besides writing articles and creating KNIME workflows, he is also an active member of the Australian community and he is a member of the current editorial board of our Low Code for Advanced Data Science Journal on Medium.

Angus is passionate about data, focusing on text analysis and visualization, and even obtained a PhD researching the use of text analytics in social science. He is maintaining two blogs where he likes to write about local history, repacking it into a more intelligible form, and about his experiments in data analysis and visualization. Whatever he is doing, he strives to communicate it in a clear and engaging way.



Visit Angus' space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `angusveitch`).

TweetKollidR in KNIME

A workflow for creating text-rich visualisations of Twitter data

Author: Angus Veitch

Editor's Note:

Some of the nodes in the workflow linked in this article are available on an external software update site.

Twitter may not be the most widely used social media platform in the world today (at the time of writing, Facebook has about seven times as many users), but it is surely the most widely studied. Thanks to its highly accessible API (application programming interface), Twitter has become the go-to source of data for social scientists, market researchers and other analysts who want to map trends and mine insights from social media. Whereas some social media platforms provide no easy way to download large amounts of content, Twitter's API makes it relatively easy to obtain tens or hundreds of thousands of tweets about a given topic. The only hard part about obtaining data from Twitter is choosing from among the many available scripts, tools and services that will help you to do it.

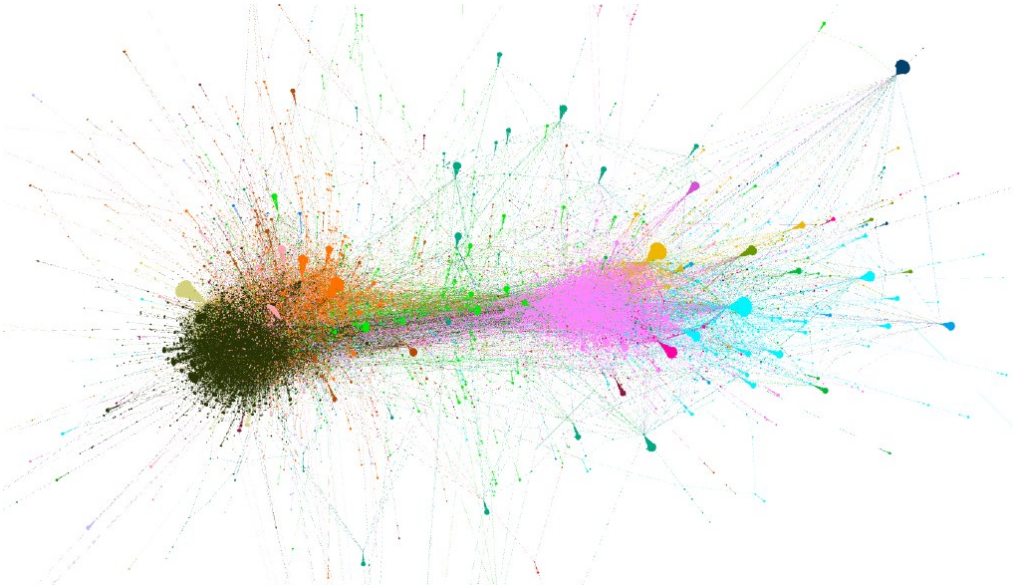
But what can you actually do with a hundred thousand tweets?

This article presents one answer to that question. The answer goes by the name of TweetKollidR, a workflow that I developed for KNIME Analytics Platform (working together with selected R packages) to create text-rich visualisations of Twitter datasets. After first explaining the motivation behind the TweetKollidR, I will illustrate what it does and explain how you can use it to create and analyse your own Twitter datasets from within KNIME.

The challenge: making sense of the Twitterverse

Twitter data lends itself naturally to at least two analytical methods. The first is network analysis. Posts (or 'tweets') and users on Twitter are connected through webs of interactions such as retweets (where one user repeats another user's tweet), mentions (where one user includes another user's username in a tweet), and replies (where one user tweets in response to another). With hardly any manipulation at all, the data from Twitter's API can be turned into a network of users or tweets linked by retweets and mentions (replies are a little more difficult, but can also be incorporated). A popular way to visualise such networks is with the open-source tool Gephi, which

(with the help of appropriately tuned layout and community detection algorithms) can turn a network of 33,000 users into something like this:



A network of more than 33,000 users tweeting about Melbourne's lockdown in 2020, as rendered in Gephi using the ForceAtlas2 algorithm. The colours indicate tightly connected communities.

It's certainly pretty, but what does it mean? Well, if you were to zoom into the network within Gephi, you would be able to see the names of each of the 33,000 users, scaled to show each user's popularity or degree of influence. If some of the usernames are familiar, you might be able to makes sense of the sub-networks or communities into which they are clustered (indicated by colours in the image above).

This is useful information. But to learn anything more — such as what each cluster of users is talking about, or what attributes define them — you will need to search through the original data to find and read the relevant tweets and user descriptions. This could be a laborious process! If only there was a way to see an instant summary of each cluster's tweets and user profiles without leaving the visualisation. As you will see shortly, the TweetKollidR provides exactly this functionality.

The second analytical method that is commonly applied to Twitter data is time series analysis. Since each tweet is timestamped, it is easy to make a plot showing how the volume of tweets posted about a given topic changes over time. But, as with the network visualisation, you will have to go back to the data to see what is really going on — to see, for example, which specific topics or users are driving spikes in activity, or how the overall mixture of topics is changing over time. The TweetKollidR helps us to see all of these things at once.

In short, the motivation behind the TweetKollidR is to provide rapid insights about the content, structure and temporal dynamics of Twitter activity around a given topic. In doing so, it is designed primarily to facilitate the exploratory or descriptive analysis of

Twitter data, but could be used to support or inform more quantitative analyses. As an added bonus, the TweetKollidR can help you to build your own dataset of tweets about a given topic.

Even better, the TweetKollidR allows you to do all of this without using any code, as it is implemented entirely through a point-and-click interface within KNIME Analytics Platform.

Installing the TweetKollidR workflow

As mentioned above, the TweetKollidR is built for KNIME Analytics Platform, so the first requirement for using it is to download and install KNIME (if you haven't already!). Once you've done that, you can grab the workflow straight off the KNIME Hub. The workflow might then require you to install a few extensions, most notably the Interactive R Statistics Integration, which will allow KNIME to call upon various R packages (specifically, igraph, visNetwork and Plotly.R) to perform some network calculations and create the visual outputs. Once you have successfully configured the R Statistics Integration extension, you will be able to install these and other necessary R packages simply by running a pre-defined script in the workflow.

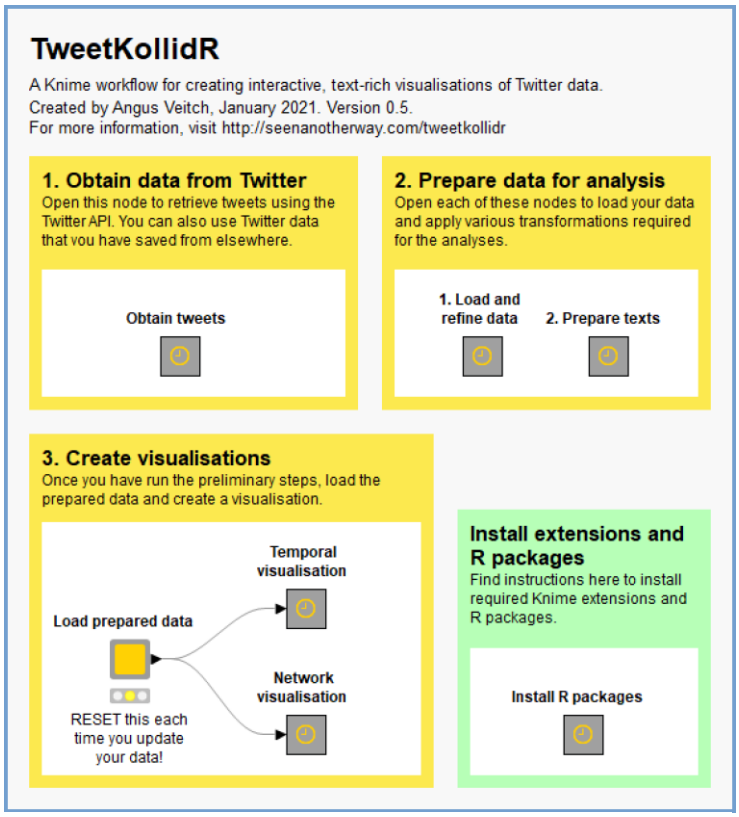
While incorporating R packages makes the workflow a little more difficult to set up, it allows us to do things that we cannot (yet!) do with KNIME alone. In particular, it gives us access to the wonderful world of data visualisation tools that have been developed for R, of which igraph, visNetwork and Plotly are just a few examples.

Using the workflow

The TweetKollidR workflow is designed to be used from within KNIME Analytics Platform much like a standalone app. As shown in the figure below, the 'home page' (or top level) of the workflow is divided into metanodes that relate to specific tasks. Within each one are further metanodes and components that the user interacts with to the perform the task. Each component is configurable through point-and-click options, which are all explained in detail through the on-demand documentation in KNIME's description pane.

At no point will you have to enter any code, or bother yourself with the inner mechanics of the workflow. Of course, if you want to see under the hood, and even customise the workflow to your needs, you can do this easily by opening up each component and drilling down into the details. (But be warned, it doesn't all look as pretty as this!)

The following sections demonstrate the core features of the TweetKollidR and outline the steps involved in using them. For a more detailed description, see the much longer blog post on which this article is based. At seenanotherway.com, you will also find other analyses done with TweetKollidR outputs, including one about Australia Day and another about the GameStop saga.

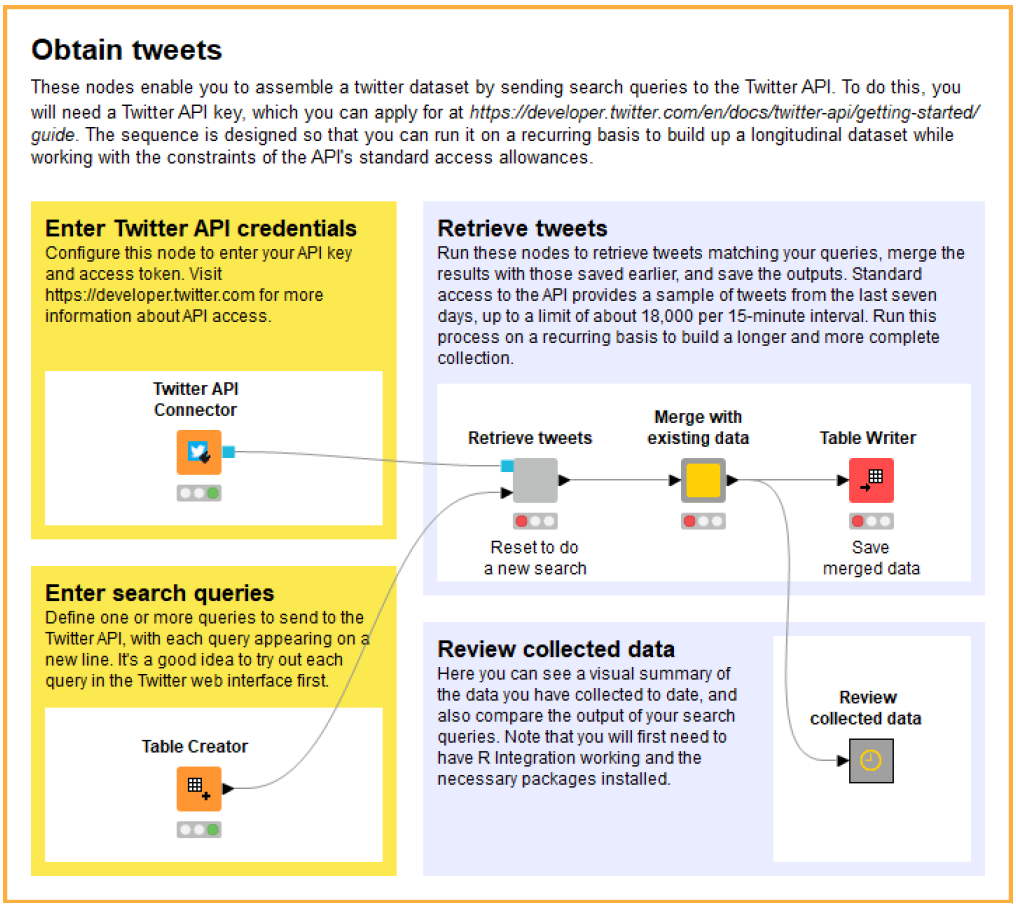


The main screen of the TweetKollidR workflow. Each of the boxes contains further nodes and instructions to guide you through the process of obtaining, preparing and visualising your data.

Gathering and preparing Twitter data

KNIME's suite of Twitter Connector nodes already make it very easy to retrieve data through Twitter's API (that is, assuming you already have an API key, which you can obtain by applying for a developer account). The Twitter Search node, for example, retrieves tweets that match a specified keyword query.

However, there's a catch. Unless you have paid for premium access to Twitter's API, or have access to the Academic Research product tier, you can only download a limited number of tweets at a time (up to 18,000 tweets in a 15-minute period at the time of writing), and these tweets will only be from the last week or so. So to build up a long or a large dataset, you will need to make multiple requests over time and collate the results. The TweetKollidR makes this process easy, providing a sequence of nodes (shown in the figure below) that automatically collate the results of repeated searches. It even allows you to specify multiple search queries, which are then automatically shuffled to ensure that they are equally represented.

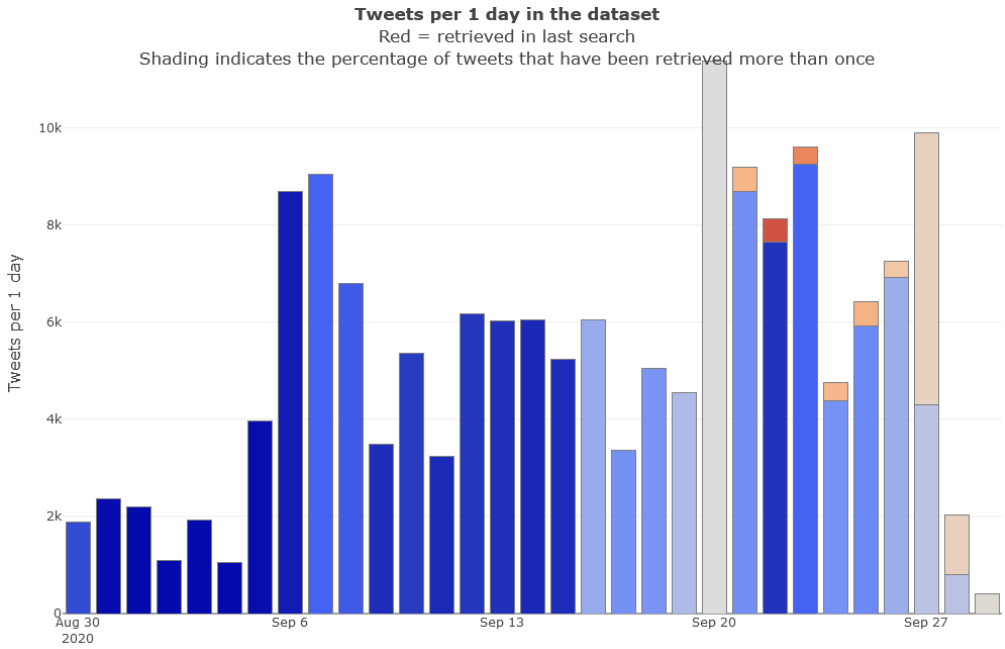


This screen guides you through the process of building a dataset of tweets about your topic of interest.

Because each request sent to the standard Twitter API returns only a sample of matching tweets, you can never be sure that your collection is complete. However, by tracking how many new tweets are returned from each request, you can at least guess when you have retrieved as many as the API is going to provide. To this end, the TweetKollidR provides the following visualization shown below.

In this case, we can see that nearly all tweets in the dataset were retrieved multiple times up until September 16th, at which point there are several successive days where the data is less saturated. Although this image is static, the TweetKollidR's output is actually an interactive HTML version generated by the Plotly package in R.

In case you're wondering, the data behind the figure above, and behind the examples that will follow in this article, is a collection of tweets relating to the 15-week lockdown that was enforced in Melbourne, Australia between June and October 2020 in response to the Covid-19 outbreak. The dataset, which I assembled over a period of about four weeks in September 2020, contains around 100,000 tweets matching queries such as 'lockdown AND melbourne'.



The number of tweets per day in my dataset on 27 September 2020. The colour and shading of the bars provide information about the likely completeness of the dataset.

In many respects, the TweetKollidR is a direct product of that lockdown, as it was during this period that I found the time and the curiosity to see what Twitter could reveal about the local politics of the pandemic. The TweetKollidR was my very own lockdown baby!

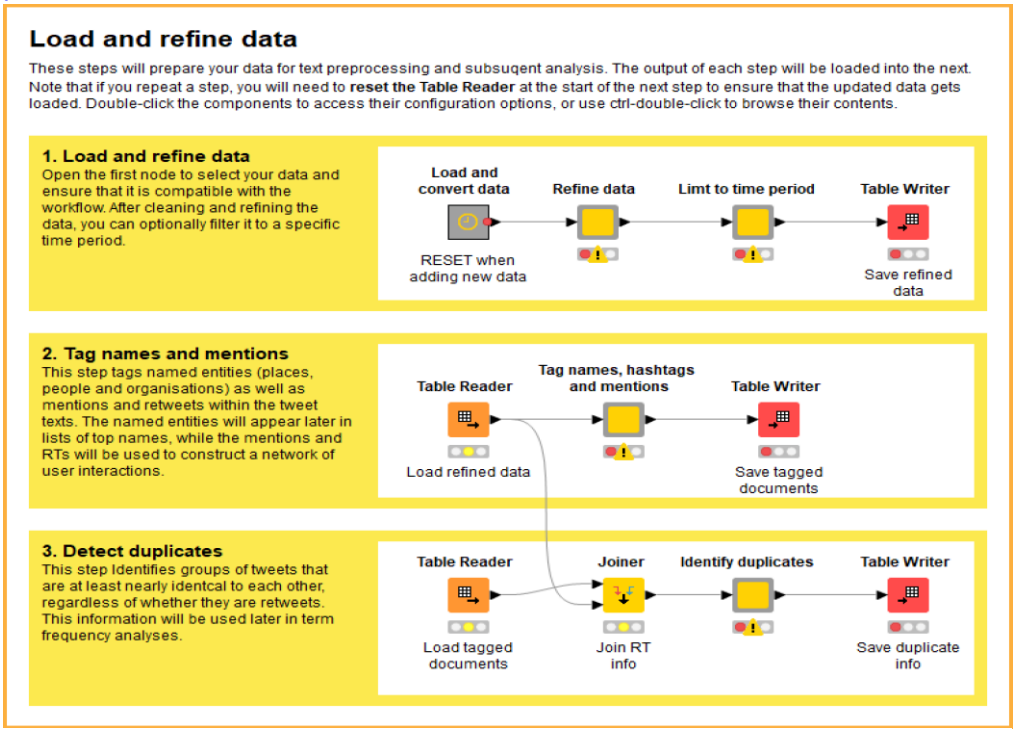
(This lockdown had a happy ending, and not just because it produced the TweetKollidR. It also resulted in the total elimination of the coronavirus from Melbourne for several months.)

Preparing the data for analysis

Before visualising your data, the TweetKollidR must subject it to a series of preparatory steps. Most of these steps include text preprocessing operations such as the removal of unwanted characters, the tagging of names and n-grams, and filtering and standardisation of terms. These steps are all performed with the help of KNIME's native text processing nodes. The TweetKollidR applies these steps not only to the tweets themselves, but also to the text contained in the user descriptions. In addition, the TweetKollidR identifies tweets that are duplicates of one another even if they are not marked as retweets, allowing such duplicates to be excluded from certain analyses.

The user performs these steps simply by configuring and executing the relevant components, such as those shown in the image below. Technical information about

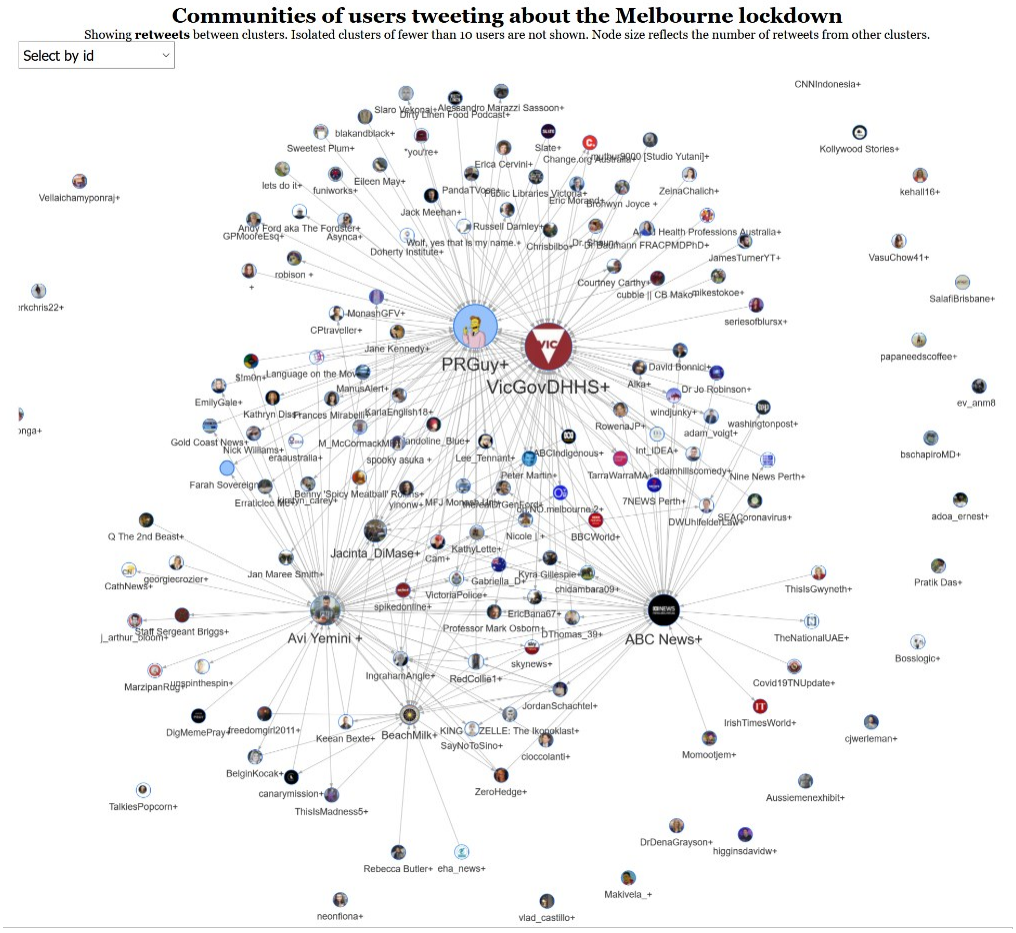
these steps can be found in the inbuilt documentation and in the [TweetKollidR blog post](#).



The components within the Load and refine data section of the workflow.

Exploring communities of users

Remember the very pretty but not very informative network visualisation shown at the start of this article? The TweetKollidR's interactive visualisation of the very same data looks something like this:



The 'summary network' showing connections between clusters of users tweeting about the Melbourne lockdown. Each node represents a cluster of anywhere from several users to several thousand. For full functionality, see the interactive version.

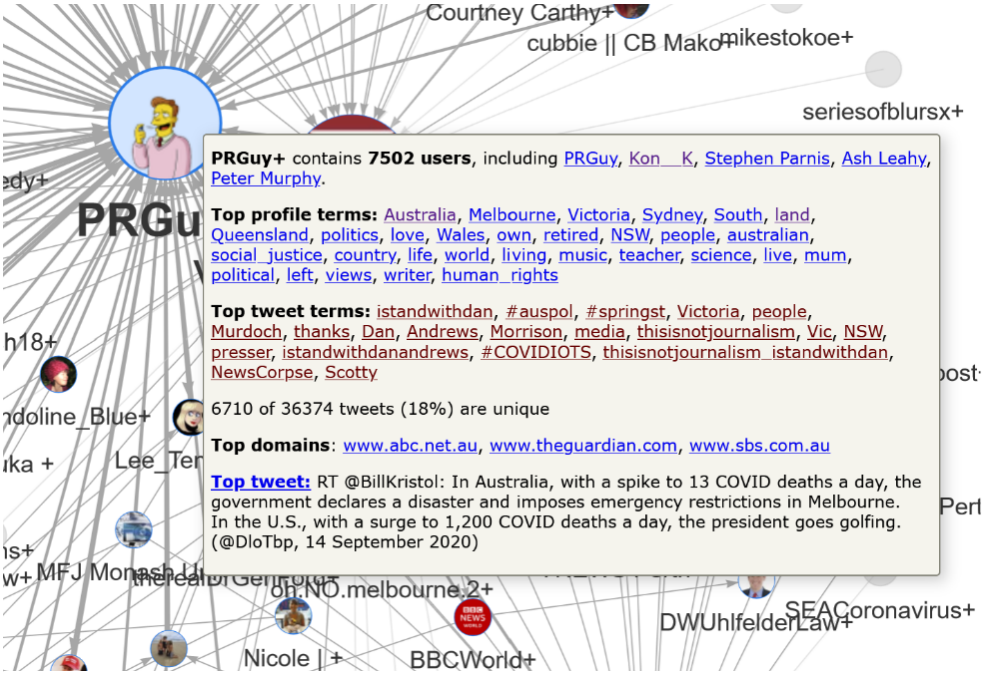
Each node in this network is a cluster of users, identified automatically by a community detection algorithm (specifically, the *fast_greedy* algorithm from the *igraph* R package). Each cluster is named after its most highly retweeted user, while the node size reflects the number of times that users from other clusters have retweeted members of the cluster.

By showing clusters instead of individual users, this visualisation allows us to see the high-level structure of the network as well as the names and profile pictures of the most influential users. In many ways it is a cartoon version of the full network: it removes much of the detail and nuance, but allows us to see the most important things more clearly.

Immediately, for example, we can see that many of the users tweeting about the lockdown at this time could be grouped into four large clusters in which the accounts named PRGuy, VicGovDHSS, ABC News and Avi Yemini were the most popular.

Moreover, we can see that the first two of these clusters are closely connected with one another (meaning they frequently retweet one another), while the latter two are more loosely connected to the rest.

But who else is inside these clusters, and what are they tweeting about? To find out, all you need to do is hover the mouse cursor over one of them (in the interactive version of the visualisation), at which point the following information appears:



The pop-up information describing the PRGuy+ cluster.

This pop-up tells us, firstly, that the cluster around PRGuy contains 7,502 users. It also lists the five most popular users in the cluster, providing hyperlinks to each of their Twitter profiles.

The 'top profile terms' section lists the most prominent and distinctive terms that appear in the profile descriptions of users in the cluster. The place names in this list tell us that most of the users are likely to be Australian. Meanwhile, terms such as *social justice*, *left*, *human rights*, and perhaps even *teacher* and *science*, suggest that the users in this cluster are mostly left-leaning in their politics.

The presence of the *#IStandWithDan* hashtag at the top of the 'top tweet terms' list provides evidence that most users in this cluster are supportive of the Victorian premier, Daniel Andrews. Meanwhile, terms like *Murdoch*, *NewsCorpse* and *ThisIsNotJournalism* suggest that these users are mostly critical of how Newscorp news outlets have covered the lockdown debate. This can be confirmed by clicking on any of these terms in the visualisation, which opens up an actual tweet from within the

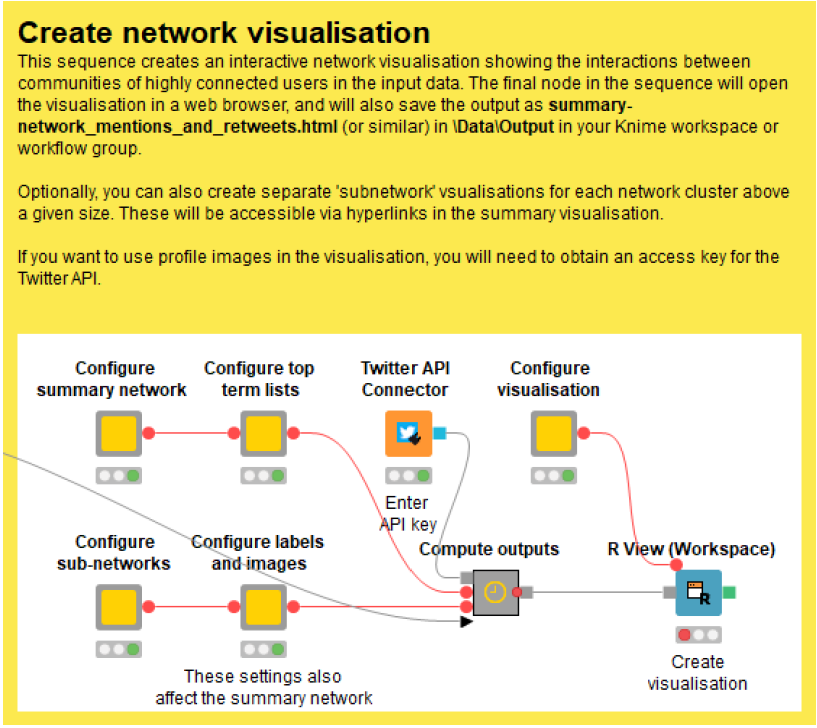
cluster that uses the term. For example, clicking on the term *ThisIsNotJournalism* opens the following tweet:

A tweet by @PRGuy17 containing the term ThisIsNotJournalism.

The summary information for this cluster also reveals that the top three websites linked in tweets from this cluster are the ABC, The Guardian, and SBS — all of which are likely to be preferred by people who are wary of NewsCorp outlets.

These are insights that we simply could not have gained by looking at a network in Gephi, even if we had the raw dataset close at hand. By doing a lot of hard work for us in the background (calculating term frequencies, and so on), and compiling the results into an interactive visualisation, the TweetKollidR can save us hours of time that would otherwise be spent rummaging through tables of tweets and user descriptions.

Many aspects of the visualisation, such as the length of term lists and the number of connections that are shown, can be configured in the workflow simply by double-clicking on the components shown below.



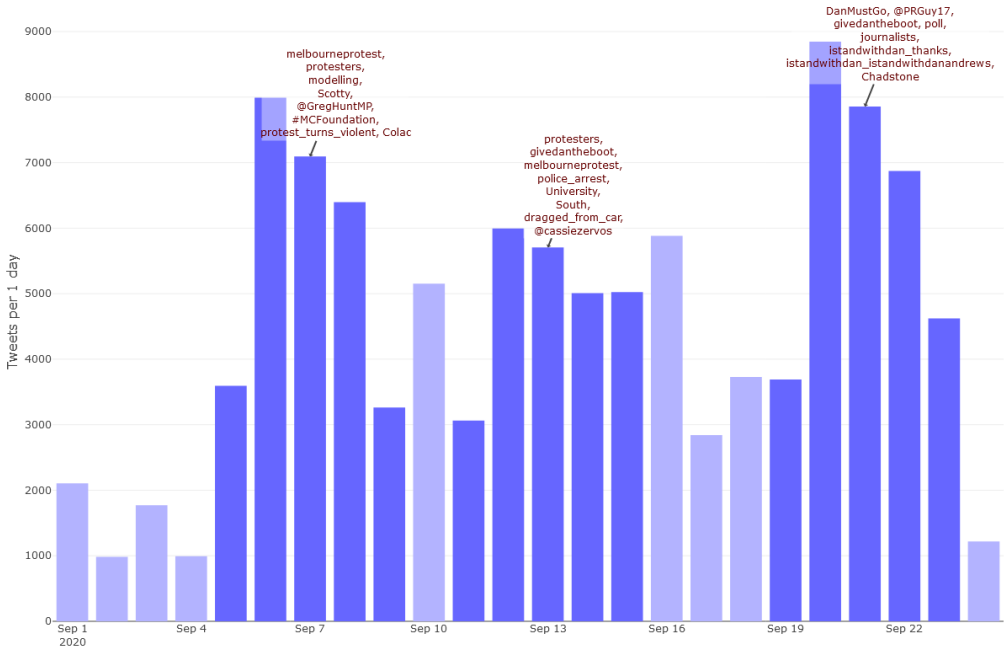
This part of the workflow enables you to configure and generate the network visualisation.

The example about highlights just some of the features of the TweetKollidR's network visualisation. To learn more, see my original TweetKollidR blog post.

Exploring changes over time

As with the network visualisation, the TweetKollidR's temporal visualisation incorporates qualitative information and interactive elements to paint a richer picture of changes over time than can be achieved with a conventional time series graph. The interactive outputs are produced by the Plotly package for R. The example below shows the daily number of tweets about the Melbourne lockdown from the 1st to the 23rd of September. (To see the full functionality, open the interactive version.)

As well as showing the number of tweets in each timestep, this output shows a list of prominent terms from each 'peak period' of activity. These peak periods, which are shaded dark blue, are automatically detected based on some simple criteria that the user can customise. The top terms are not simply those that occur most frequently in each peak period (at least not by default). Rather, they are selected according to their frequency as well as their uniqueness to the period. In addition, they exclude terms that appear in every single timestep, and force the inclusion of at least one location and person or organisation.



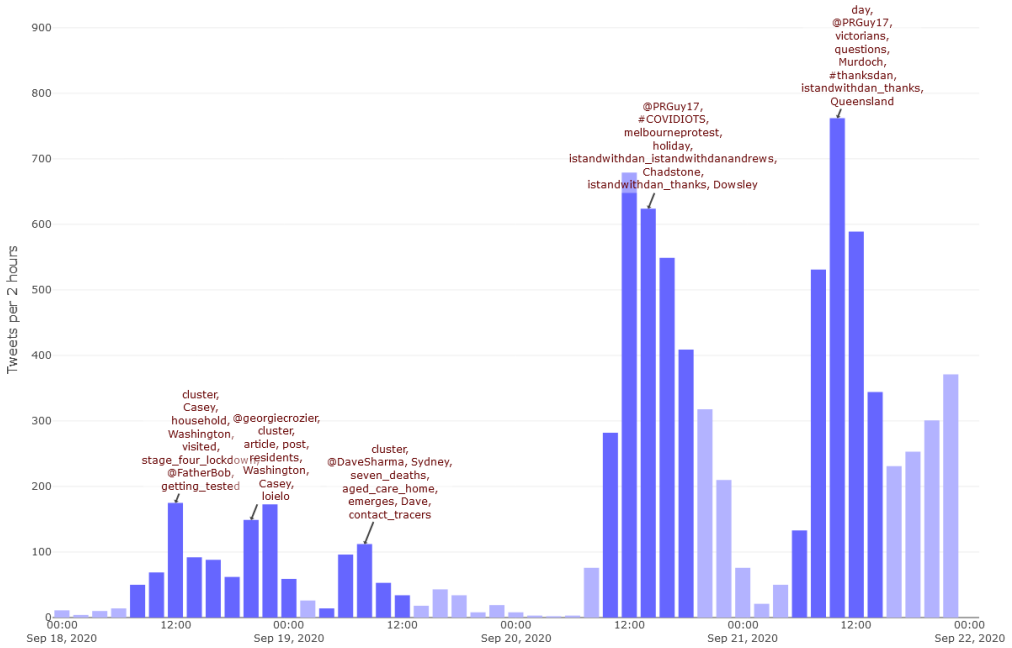
An auto-annotated chart showing the number of tweets posted each day from the 1st to 23rd of September. The annotations list prominent terms from each shaded period. For the full functionality, see the interactive version.

The purpose of the annotations is to provide a rough idea of what is going on in each period of peak activity. In this case, the terms for the first peak period (the peak itself was Sunday, 6 September) suggest that protests were a major talking point. As well as the word protesters and the hashtag #melbourneprotest, the list includes the juicy n-gram, protest_turns_violent. Also discussed during this period was modelling (specifically, that used by the Victorian Government to justify their lockdown measures), the Prime Minister (known informally by Australians as Scotty), and the federal health minister, Greg Hunt. The terms in the interactive form of this visualisation are all hyperlinked to actual tweets, so you can click on them to see how they were actually used.

Half a dozen words and names is not much upon which to form an impression of several thousand tweets posted across several days. To allow a fuller and more nuanced characterisation of activity over time, the visualisation provides popup information summarising each timestep. Here is an example:

Summary information is available for every timestep in the visualisation.

These popups provide similar information to those in the network visualisation: separate lists of prominent terms and names used in tweets, top terms from user profiles, and the most popular tweet from the timestep. Unfortunately, these lists are not clickable, because Plotly (the R package that produces the visualisation) only allows the popup text to display while your cursor is over the relevant bar.



Activity by users in the PRGuy+ network cluster from the 18th to 22nd of September.

You can configure the temporal visualisation to show any time period within your data, using timesteps of any length. Furthermore, you can filter the data to just the users from a given cluster from the network analysis. For example, the image above summarises the activity of users in the PRGuy cluster (discussed earlier) from the 18th to the 22nd of September, using 2-hour timesteps.

Conclusion

Of the many tools that are available for gathering and analysing Twitter data, I think the TweetKollidR is distinctive in several ways. As far as I am aware, it is unique in providing network and time series visualisations that are rich in qualitative information derived through computational text processing. This integration of quantitative and qualitative information is made possible by two things: the background text processing performed by KNIME, and the interactive visualisations generated by the visNetwork and Plotly R packages.

Secondly, the TweetKollidR is notable for its use of KNIME Analytics Platform as opposed to R or Python, which are more commonly used to gather and analyse Twitter data. The main benefit of using KNIME for this purpose is that the resulting tool can be operated through a purely graphical interface, which cannot be said of even the friendliest of R packages or Python scripts. When used in this way, KNIME effectively erodes the boundary between the user interface and the underlying code, allowing the curious user to see the inner workings at any time, and to modify them if necessary.

Then again, the inner workings of the TweetKollidR are not entirely code-free, as they include several R scripts. The TweetKollidR is an example of a low-code analytics solution, where code is used only where it is absolutely needed. In this case, code is needed to generate the interactive visualisations that are the workflow's final outputs. All of the preceding data manipulations and transformations, including the text processing, are handled by KNIME's native functionality.

Code or no code, the TweetKollidR is an inherently complex beast, as you will see if you peek beneath the top layers of the workflow. For this reason, it is bound to contain a few bugs, and there is always room for improvement. If you do give it a try, I would very much welcome any bug reports, suggestions or other feedback.

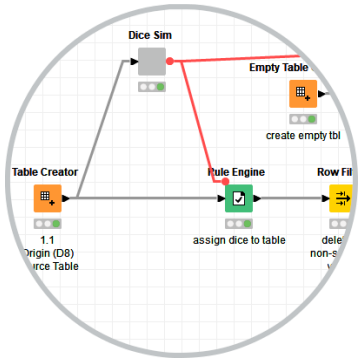
Happy Kolliding!

This article was originally published on Angus' blog Seen Another Way and we republished it in our Low Code for Advanced Data Science Journal on Medium. Find the article here.

The corresponding TweetKollidR workflow can be found on the KNIME Community Hub in Angus' public space.

Philipp Kowalski was nominated KNIME Contributor of the Month for March 2021. He was awarded for his Forum thread in which he shared his creative use case using KNIME for his hobby of role-playing games. With the help of KNIME Analytics Platform he crates heroic quests for sessions of Dungeons & Dragons. The image on the right shows his standard dice rolling component which mimics the result of dice roll. This is truly a unique use case, but it wasn't the first and will probably not be the last time that Philipp challenged the KNIME community by pushing the KNIME software to the limits.

Philipp is a no-code/low-code enthusiast and an experienced trainer. He not only owns and runs ProcurementZen, a blog and podcast centered around procurement and negotiation strategies, but also maintains a YouTube channel with a wide range of tutorials on the usage of KNIME and best practices in procurement. Philipp currently works as Digital Enablement Agent at Siemens.



Visit Philipp's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `kowisoft`).

Digitalization Evangelist Discusses Automation for Businesses and Dungeons & Dragons

My Data Guest — An Interview with Philipp Kowalski

Author: Rosaria Silipo

It was my pleasure to recently interview Philipp Kowalski as part of the *My Data Guest* interview series. He shared insights into his role as a digitalization evangelist, explained how KNIME can drive effective process automation for businesses, and gave an example of applied digital transformation to generate fantasy scenarios for Dungeons & Dragons.

Philipp Kowalski works as Digital Enablement Agent at Siemens. He is an expert KNIME user, having worked with KNIME since 2018 for many data analytics use cases. He is also a digitalization expert and truly passionate about the application of digital transformation to all fields of life, from the job to the hobbies, from data analytics to creative disciplines. But that's not all. Philipp is also an experienced trainer, and owns and runs ProcurementZen, a blog and podcast centered on procurement and negotiation strategies. He has a YouTube channel with a wide range of tutorials on the usage of KNIME, and best practices in procurement. Philipp was also one of the first KNIME Contributors of the Month, having been awarded in March 2021 for his creative usage of KNIME Analytics Platform. Indeed, he builds KNIME workflows to create heroic quests for wizards, dwarves, elves, and humans for his sessions of Dungeons & Dragons.

Rosaria: *On LinkedIn, you define yourself as a digitalization evangelist. What is that? Why would digitalization be so useful?*

Philipp: A digitalization evangelist is someone who spreads the word about digitalization, especially with the focus on KNIME. Digitalization is useful because there are still so many redundant and repetitive tasks we should get rid of, and at the same time nowadays we have so much data available that we are able to accomplish that. The time we save by digitalizing processes could be used to make sense of the data and implement advanced analytics solutions. In that sense, KNIME is really

helpful, and I'm showing others how easy it is to use it through my work. My ultimate goal is to demonstrate that we can digitalize more or less all walks of life.

Rosaria: *Is digital transformation a big part of your daily job at Siemens? Do you use KNIME?*

Philipp: Yes, absolutely. I have been a procurement professional for almost 20 years. Since February 2022, in my daily job at Siemens, I've started introducing digitalization in the field of procurement. My task now is to teach KNIME to my colleagues. The big advantage of using KNIME is that the learning process is very rich, effective, and quick. For example, you start with automation but while you are building automation, you learn other tasks and nodes as well. Using KNIME, I also show my colleagues how to digitalize a lot of redundant procurement processes. This makes their life easier and better.

Rosaria: *What are the KNIME features that help the most for digital transformation?*

Philipp: In the beginning of a project, there are usually several ETL operations to perform. Hence, I use a lot of data manipulation nodes for cleaning, preprocessing, summarization, or export. After that, I rely on quite a few extensions that are very useful in procurement. For example, the Continental nodes for XLS formatting, BIRT for reporting, learner-predictor nodes to build predictive models, or the nodes of the text mining extension.

Rosaria: *Tell us about some typical use cases in your work where you rely on KNIME.*

Philipp: One example is reporting. KNIME is a fantastic tool for reporting, using both the BIRT extension or creating interactive dashboards. Another example is pattern recognition. It is important for us to know when a purchase order is likely to become problematic. Identifying potential problematic orders in a timely manner is very beneficial for the company because it prompts us to handle these orders with greater sensitivity. KNIME helps us a lot to discover these patterns and minimize losses.

Rosaria: *How advanced should your knowledge of data analytics be to work in digital transformation?*

Philipp: I dare to say that you don't need any previous knowledge. The only code I ever used was when I collected data via web scraping. For everything else, KNIME's low code approach is really enough.

Additionally, if you are proficient in data wrangling operations with Excel, you already have what it takes to work smoothly and productively with KNIME. But KNIME has a few key advantages over traditional spreadsheet tools. First, the sequential, node-wise step-by-step execution in KNIME is a huge benefit as you can immediately see the results after each node. In traditional spreadsheet tools, you often end up writing very long formulas, and in case you miss a bracket or semicolon it's hard to realize it.

Second, the KNIME community is fantastic. Everyone in the KNIME Forum is so friendly, helpful, and extremely fast. It has never happened that I posted a question on the KNIME Forum and it was not solved. I think this is really special.

> ***Note.*** Read "*Ten Common Issues When Using Excel for Data Operations*", and the "*Excel to KNIME*" handbook to migrate painlessly from Excel to KNIME.

Rosaria: *Let's talk about money and time saving. How significant is the impact of digitalization in a business? What are some immediate impacts that you have witnessed in your experience?*

Philipp: KNIME helps us save a lot of time by powering the digitalization of several business processes. For example, the first workflow I built reduced my reporting effort from 3 hours a month to 10 minutes. For a different project, we automated journaling of current demands, and that reduced the effort per employee from 30 minutes each day to almost nothing. For five employees that's 2.5 hours every day. I think these examples are pretty significant.

Rosaria: *It's not always the case that analytics goes smoothly, though. Sometimes mistakes are good because then we can learn more. Tell us about the biggest mistake you have learned from.*

Philipp: I once wanted to extract supplier data via web scraping. My approach was to directly extract the specific information that I needed from that web page, which I couldn't manage. I finally looked into the KNIME Forum and got great support from some Forum members. They suggested a different approach, which turned out to be very effective. I was to first scrape all the data from the website, and then trim it down and wrangle it to what I needed. This has become my standard strategy for many data problems.

Rosaria: *Let's talk about your activity as a YouTuber. Since when have you been a YouTuber?*

Philipp: My YouTube channel has existed since June 2018. I used to upload a video every now and then. Since April 2021, I have been more active and I started creating content, courses and video tutorials with a focus on KNIME.

Rosaria: *You have two courses about KNIME on YouTube. How do they differ? Which course should people follow and why? Are they all completely free?*

Philipp: Actually, it's only one course. I originally hosted the material on my website but then I decided to make it available on YouTube so everyone could access it. The course goes from basic to advanced topics, and is especially designed to target complete KNIME beginners and give them the opportunity to discover how great KNIME is. First, I introduce theoretical concepts (e.g., why automation is

advantageous, or what ETL is), then we start using a few basic nodes, we build a workflow together, we explore advanced features, etc. Give it a try!

Rosaria: *What sources (e.g., books, articles, blogs, blueprints, forum, social media, etc.) do you use to keep up to date about KNIME and data science?*

Philipp: All kinds of sources provided by KNIME are useful and contribute to a broader understanding of data science and the tool. The Data Talks and the Data Connects are very insightful events, the KNIME YouTube channel has a lot of helpful content, and I always recommend having a look at the KNIME Blog and the KNIME Forum.

Rosaria: *We are reaching the end of our interview. Before we say goodbye, I would really like to talk about the project that earned you the Contributor of the Month award in March 2021: applying KNIME to analyze data for Dungeons & Dragons. Tell us more about that project.*

Philipp: I'm a fantasy nerd, and I have been playing Dungeons & Dragons (a role-playing game) for over 30 years. What amazes me is that nowadays it is possible to play with people all around the world using so-called virtual tables. On the other hand, I also had a lot of rule books but I didn't want to play the same things over and over again. One day, it struck me like lightning and I thought "Hey, I know a tool that can help me avoid repetitive tasks". So I challenged myself and tried whether I could build new fantasy worlds using KNIME. Basically a digital transformation of Dungeons & Dragons.

Rosaria: *You were not actually analyzing data. You were using KNIME to generate scenarios for your game. Did KNIME help to make the task faster or better? If better, to what extent?*

Philipp: Using KNIME improved both the quality and the speed of the process. This is because I didn't have to write down the scenarios anymore but I was able to export them automatically into a nice PDF format, which I could then easily share with my players.

Rosaria: *How many attributes are you supposed to generate for each Dungeons & Dragons session? I mean, characters, character features, location, etc.*

Philipp: Between 100 and 150 different attributes out of all kinds of tables are created for each session. The interesting part is that these tables are interconnected. This means that if you roll a 4 on table A, you also have to roll on table B. I learned that KNIME is extremely strong in rule-based use cases, which I used to create my little settlements. It really sparked some very nice role-playing rounds and it was a lot of fun!

Note. Read more about Philipp's Dungeons & Dragons project in his post on the KNIME Forum.

Rosaria: *Sir Philipp, one last question. How can people from the audience get in touch with you?*

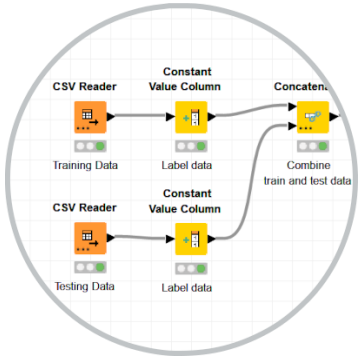
Philipp: You can reach me via LinkedIn or on my YouTube channel. Just leave a comment in one of my videos, and I will reach out to you. All my workflows are linked on YouTube as well.

This article was first published in our Low Code for Advanced Data Science Journal on Medium. Find the original version here.

Watch the original interview with Philipp Kowalski on YouTube: "My Data Guest – Ep 6 with Philipp Kowalski".

Tosin Adekanye was nominated KNIME Contributor of the Month for August 2021. She was awarded for her valuable contributions on the KNIME Community Hub, her countless social media posts, and for moderating the KNIME, Excel & Reporting user group at the last KNIME Data Talks – Community Edition. Some of her most popular contributions include working on data dealing with the FIFA Arab cup, Financial Transactions, UFO sightings, and Fraud Detection. The image on the right shows a snippet of her Fraud Detection workflow. Many of those are published on her social media channels or in our Low Code for Advanced Data Science journal on Medium. She has also launched the "30 Days of KNIME" series on YouTube. She has made it her mission to divulge her knowledge and experience in data science to all interested professionals, removing the coding barrier and focusing on concepts.

Tosin holds a Master of Business Administration with the focus on Business Analytics from the University of Tampa. Her academic research background gave her a soft but solid entry into data science. She is now trying to make the field accessible to everyone by explaining concepts in a way everyone can understand. This is one of the drivers behind her passion for the no-code/low-code environment. She is currently a Data Scientist at Qatar Financial Centre Regulatory Authority.



Visit Tosin's space on the KNIME Hub or her profile page in the KNIME Forum (Hub/Forum handle: `tosinlitics`).

To SQL or Not to SQL, UFOs, Sci-fi Movies – And Other Important Data Science Questions

My Data Guest — An Interview with Tosin Adekanye

Author: Rosaria Silipo

Editor's Note:

Some of the nodes in the workflow linked in this article are available on an external software update site.

It was my pleasure to recently interview Tosin Adekanye as part of the *My Data Guest* interview series.

Tosin Adekanye started her journey towards data science during her undergrad studies in psychology, where she was heavily involved in statistics and academic research. She developed these skills further during her MBA, where she started learning and utilizing ML algorithms and data science software. But it was during the lockdown, last year, that she became obsessed with data science! And that is also when she got to learn the No Code/Low Code "powerhouse" – as she calls KNIME software.

Tosin is a very active member of the KNIME Community and an influencer in the data science space on social media. She writes articles for a number of journals, like the "Low code for Advanced Data Science" journal on Medium, describing very interesting solutions to the most diverse tasks. I repeat "to the most diverse tasks". For example, in an article she provides a solution for fraud detection on credit card transactions; in another article she talks about UFO sightings; then in another article she describes the possibility to bypass SQL coding with visual programming, and so on. Quite a diverse array of projects.

Rosaria: *Hi Tosin, Tell us more about these different projects. Which of the projects you've written articles about were for work and which ones are hobbies?*

Tosin: Some of the articles were motivated by work projects. When I was at FIS, the global credit card processing company, a colleague in the fraud department told me

about how they use models to predict credit card frauds. That's what got me curious to build something like that myself and write about it. Another work-inspired article came about because I use databases heavily every day. That led to the article on SQL. The UFO article on the other hand, well, that's a hobby of mine. I'm obsessed with science fiction and alien sightings. But it's hobbies that also motivate me to look into different topics.

Rosaria: *I'd like to talk about your article for predicting credit card fraud. Can you tell us a bit more about your interest in working with financial data?*

Tosin: Yes, I really like finance. My background is actually psychology and business, but finance is a synergy of all my passions. Fraud is a real pain for both the customer and the company. It's cost so many companies in the U.S. so much money... I was curious to see if I could build a model that could catch these fraudulent transactions and maybe tell us which variables seem to go with fraud. The dataset I used was synthetic data, but it was based on real-life trends in this field.

Someone's age, for example, is associated with a high likelihood for fraud: Fraudsters like to pick on potentially gullible, older people. Online fraud also takes place more at night or early in the mornings. Being aware of insight like this can help companies and customers better protect themselves.

Rosaria: *Have you ever tried to predict stock prices? And do you think this is actually possible?*

Tosin: Let me rewind a bit! During my MBA in finance we tried to predict stock prices based off of things such as financial ratios from past company statements. Our best models probably predicted to an accuracy of 10 to 14 and only accounted for a certain amount of variance. I believe that to really predict a stock price you need to be holistic. You need to look at everything from news articles to social media posts to historical stock prices to the company's financials. I am still actively working on this. By the way, this project was also the reason why I started using KNIME. I needed something that lets me process data from all these kinds of sources.

In the past, I also ran some analysis between sentiment on Twitter and the movement of S&P values, and detected some relationships there. So yes, I think it's possible to predict stock prices. Definitely not to 100% accuracy, but given enough data from lots of different sources and good models, yes it's possible to predict with some accuracy.

Rosaria: *In your article on fraud detection, you compared the performance of several models (Decision Tree, XGboost, Gradient Boosted trees) and declared the XGboost model to be the best. Why do you think XGBoost performed better than the other algorithms in this particular use case?*

Tosin: For classification XGboost tends to perform better for several reasons. It's an ensemble model so it's building multiple trees, which usually have better performance.

Also, XGboost improves on weak learners - on features that are not leading to good predictions. It keeps getting better as it progresses. It also limits overfitting by introducing a cost function, so all these different controls usually make it perform better than other models.

Rosaria: *You say that it was your work with databases that led you to write the article "to SQL or not to SQL?". This is a controversial topic. Some people say, "You must SQL otherwise you forget how to program in SQL". Others say, "If I cannot SQL, then visual programming allows me to SQL anyway". What is your opinion?*

Tosin: Oftentimes I feel that in this field of data science people can become too attached to tools. I would say do what works best for you. If you like to see code and it's the best thing for you then do that! I prefer to have a blended approach. I have workflows with lots of SQL code, but you'll also see me using SQL nodes [e.g., the nodes in the KNIME Database extension implementing SQL queries in the background] because oftentimes that's what works best for me. Essentially whatever works best and whatever is most efficient for you is what I would recommend.

Rosaria: *Some time ago, somebody posted on Twitter that people use low code tools only if forced by their bosses, that people would never use it for their hobbies. You seem to contradict that tweet. Do you use KNIME Analytics Platform for all your projects?*

Tosin: I'm not sure how they got that point of view. I brought KNIME to my workplace! I don't use KNIME exclusively but it would be hard for me to go somewhere and not be able to use KNIME.

Rosaria: *So, you use KNIME Analytics Platform in combination with other tools?*

Tosin: Yes, usually Power BI and KNIME. Sometimes I use PyCharm if I need to program in Python. KNIME does have a Python node, so you can program in there too, but I usually use KNIME, PyCharm, and Power BI.

Rosaria: *Your other big passion in addition to data science is science fiction. In your UFO article, you found a way to combine them. Tell us more about that story.*

Tosin: I'm a huge fan of sci-fi and psychology. In psychology we often notice that the more something is talked about the more people seem to experience it. So I wondered if there could be some sort of relationship between sightings of UFOs and movies about UFOs. Based on data of UFO sightings and release dates of movies about aliens I did a visualization of the correlation. And the correlation was highly significant! But of course that's not causation. I would probably need to do some more digging and get better data - a richer movie database plus more recent UFO sightings and then see if I can isolate what's causing what.

Rosaria: *How did KNIME help you decipher the relationship between the number of movies and number of sightings?*

Tosin: I had multiple datasets so KNIME really helped me to blend them. The movie datasets are pretty different, so I had to do some standardization and join everything together. KNIME was very helpful for that as well as for correlation. The correlation was super easy to run just using the Linear Correlation node and then I was able to quickly look at the relationship.

> ***Note.*** Check out the workflow UFO Sightings Data Prep in Tosin's space on the KNIME Community Hub.

Rosaria: *Are you planning to write more articles like those? I am following you and I cannot wait till the next one appears.*

Tosin: I haven't quite decided about my next article but there are two things I'm working on. One is flight cancellations - looking at what factors go with flight cancellations and when flights are most likely to be canceled.

My other project is fun! There's a Twitter API Connector node in KNIME that makes it super easy to pull tweets. So I'm going to get tweets from different countries that mention the word happiness and then see which words are most associated with happiness in different parts of the world.

Rosaria: *How can data scientists in the audience follow your work and access your workflows?*

Tosin: I write articles for KNIME's Low Code for Advanced Data Science journal on Medium and I also have my website TosinLitics where I publish my work. I can recommend checking out my KNIME Hub space where I keep my workflows. I think it's very helpful to see what other people have done - not just my workflows but all workflows in general on the KNIME Hub. It's great for idea generation or to help you get unstuck. Being able to refer to examples by others helps a lot.

Rosaria: *Your TomTom component is in your KNIME Hub space, right? It was a very popular component on the KNIME Hub. What does it do? Can I download it?*

Tosin: Yes, I was motivated to do this because I didn't really see many solutions around that let you quickly overview the distance between two points using longitude and latitude. I did some research and figured TomTom was the best tool to go with. So now you just get your API key and put it in the component. Once you have your data files containing the longitudinal information, you can run those through the component and get the drive time and the distance, traffic delays, and all related information, to go from one point to the other.

> ***Note.*** Download the component Drivetime and Distance Query – Latitude Longitude from the KNIME Community Hub.

Rosaria: *Are you planning to implement more components?*

Tosin: Yes. There's so much you can get from the TomTom API! I'm planning to do a couple more. This could be a family of geospatial components.

Rosaria: *You are a very active KNIME community member but let's go back in time: How long have you been using KNIME, and how did you get started with KNIME?*

Tosin: This might come as a surprise but I actually only started using KNIME in January 2021. I'd already had some exposure to software like KNIME - I used SPSS Modeler from 2017. Then I used Alteryx but the licensing was a barrier for me. I needed something that was efficient, that could let me do so many things for data science. That's when I found KNIME.

Even though I haven't been using KNIME that long, you can really climb the learning curve quickly because of the resources that are available. KNIME also has some of the most approachable, most passionate employees and that's really helped me come along in my learning curve.

Rosaria: *Tell us about the biggest challenge you had to solve in your professional life as a data scientist.*

Tosin: Dealing with textual data! I had been running away from this for many many years. But in January I wanted to learn how to process and analyze textual data and get comfortable with it. Example workflows have helped a lot with this, you know. I wanted to do some sentiment analysis. I realized that with textual analysis, once the data's cleaned properly and processed, it can be reduced to mathematics! Now it just seems a lot easier.

Rosaria: *Tell us the biggest mistake you've learned from.*

Tosin: My biggest mistake was the basis for an article I wrote about class imbalance and why accuracy is not always the best - especially when you have imbalanced classes. I posted about my first detection model because it had an accuracy of 99%. The LinkedIn data science crowd was super helpful because they pointed out that when you have imbalanced classes, accuracy is not necessarily the best metric. Indeed, you can have high accuracy values, but your minority class can still perform really badly in terms of classification. That was something I knew but sometimes you know something in theory but you don't really realize it until you see it happening in practice

Rosaria: *Yes, imbalanced classes can give you false expectations of how it's performing. Do you have any advice for all young aspiring data scientists who are in the audience?*

Tosin: I have three primary pieces of advice:

- One, read a lot. Medium is a good platform. You don't have to fully understand everything you read but you'll be familiarized with the topic and this will help in the future.
- Also, don't be afraid to share your work. The first things I shared weren't always that good but having shared them I got feedback which helped me improve more.
- Just get started! You don't have to be perfect but you're going to grow and build up from there.

Rosaria: *Any book to recommend for the ones in the audience always eager to learn something new?*

Tosin: Yes, *Data Analytics Made Easy* by Andrea De Mauro. I really like it because it teaches you the theory for analytics and for data science, which is so important. Sometimes, programs for data science just jump right into Python, but I think the theory is more important. At the end of the day Python is just a tool. When you learn about theory, this knowledge helps you know how to overcome obstacles in practice and be better in this field. This book also teaches how to use KNIME.

Rosaria: *Books are definitely an essential tool to get a solid basis but what are your usual readings to keep you up-to-date on new, exciting data stories?*

Tosin: I read a lot of Medium articles and I'm very active on LinkedIn, connected with people. so I usually see a lot of things that are being talked about. Staying in the loop, reading, and googling helps to keep up to date.

Rosaria: *While following your #30daysofknime initiative, I also discovered that you are a very talented video-maker and actually ... "surprise" ... a very talented singer. Would you like to conclude this interview with the song you sang in the first video posted within the #30daysofknime initiative?*

Find Tosin's song – in the video of the original interview linked in the yellow box below!

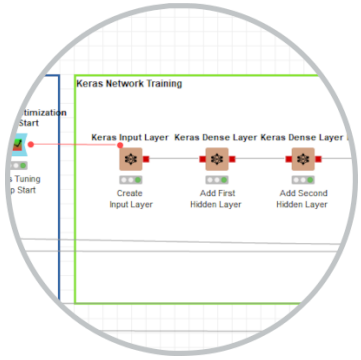
This article was first published in our KNME Blog. Find the original version here.

Watch the original interview with Tosin Adekanye on YouTube: "My Data Guest – Ep 2 with Tosin Adekanye".

Additionally, Tosin also starred in the fifth episode of the My Data Guest series. Watch the full interview here: "My Data Guest – Ep 5 Women in Tech".

Paul Wisneskey was nominated KNIME Contributor of the Month for March 2022. He was awarded for his two-part article on American football where he explores whether he can teach his Machine Learning model to make a pass: throw the ball to a receiver in motion. Part 1 deals with Parameter Optimization with KNIME, and Part 2 Trains a Neural Network. The workflow snippet on the right captures the network training part. All in all, Paul is a very active member of the KNIME community and has written numerous blog posts about KNIME. Many of them are in his Medium catalog which are definitely worth a look.

Paul is a software architect with over 25 years of experience designing and implementing large scale, reliable systems for big data, search and analytics, online information services, document imaging, and distributed collaboration. He has extensive Java, Big Data, Web Services, and Endeca search engine experience. Paul is currently Director of Engineering at BigBear.ai, a firm that delivers AI-powered analytics and cyber engineering solutions to support mission-critical operations and decision-making in complex, real-world environments.



Visit Paul's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `pwisneskey`).

Making the Pass, Part 1: Parameter Optimization with KNIME

From football to synthetic data for model training

Author: Paul Wisneskey

Photo by Chris Chow on Unsplash.

With the Super Bowl just around the corner here in the United States, I recently found myself trying to think of a football-themed blog posting. Based on the football theme, I decided to create a post about a fundamental aspect of the game: throwing the ball to a receiver (a "*pass*" in American football parlance). It is a simple operation that, with a little bit of practice, most people can perform intuitively without any knowledge of the physics equations that govern the projectile motion of the ball and the linear motion of the receiver. Maybe it would be possible to use machine learning to teach my laptop how to complete a pass to a moving target without just programming in the physics calculations?

Looking at this in the most simplistic manner, I want my laptop to make a bunch of random throws and learn from when they are successful. This is the essence of the class of machine learning algorithms known as "*supervised learning*". The goal of supervised learning is to use training data to determine the function so well that when a new data set is given, the output can be accurately predicted. In contrast, unsupervised learning is to model hidden patterns or underlying structure in a given

data set in order to learn about the data, which does not match well with the task I've selected for my laptop quarterback in training.

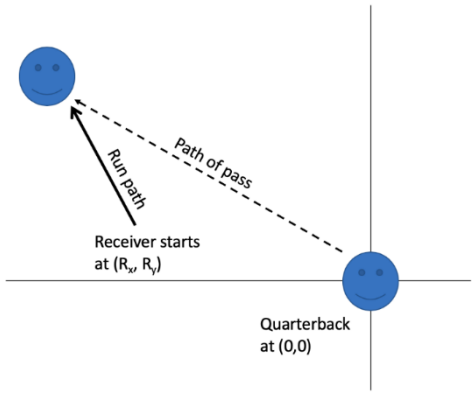
To start the learning process, I need a set of training data but if I genuinely just used random pass data, I could generate an excessive amount of data for incomplete passes with very few positive examples of complete passes. This data would have little training value. So instead, I need to come up with a training data set consisting of an appropriate number of successful passes. Note that I'm hedging here by saying an "*appropriate number*" as we don't yet know what that might be — the size of the training set is likely dependent on the complexity of the problem and the machine learning algorithm that we are trying to train.

So how am I going to get this training data? In an ideal world, I could capture completed pass data from a large number of football games, but this is way outside of the scope of a simple blog posting. Furthermore, one of my modeling assumptions that I snuck into the first paragraph of this posting was the "*linear motion of a receiver*". In the real world, there are no such constraints on the motion of a receiver — they can run pre-planned routes or change their routes based on real-time conditions during the play.

Since I don't have any easily accessible source of actual data that fits with my modeling assumptions, I'm going to have to synthesize some completed pass data for training with. My first gut instinct was just to start working out the mathematics so that I could produce a set of functions for calculating the ideal pass from a fixed thrower to a receiver on an arbitrary linear route at a constant velocity within a set of arbitrary constraints for the maximal throw velocity (e.g., no such thing as a faster than light throw).

But this felt a bit like cheating because I would both be providing mathematically perfect training data and I was also ignoring the powerful tool I had sitting at my fingertips: the KNIME Analytics Platform. What if I could use KNIME to generate some "*dirtier*" completed pass training data? Did I just reason myself into a chicken or the egg problem? How can I teach my laptop to throw a football if I first must teach it to throw a football to generate training data so that I can teach it how to throw a football…? Well, it can be done, and I will walk you through the how and why — if you'd like to follow along directly in my KNIME workflow you can download it from the KNIME Hub.

But first let us back up to the task I want to teach my laptop how to perform. Maybe if we more rigorously define it, we will find a way out of the conundrum. So, I sketched out the problem as follows:

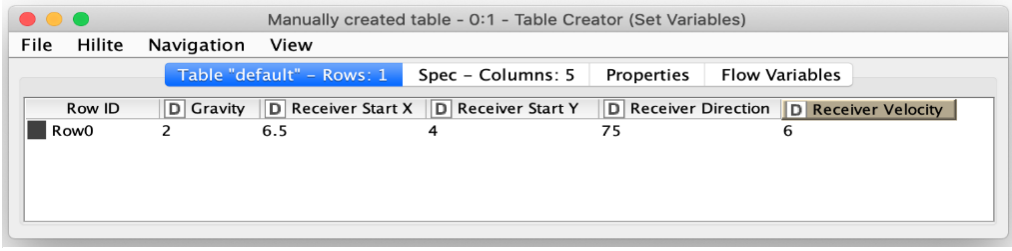


A sketch displaying the throw of a football.

While drawing the sketch, I realized there were some arbitrary choices I could make to simplify the model. Most importantly, the quarterback would stay in a fixed position at the origin (0,0) of the coordinate system. The receiver would start at an arbitrary position *Rx*, *Ry* and run away from the quarterback in a random direction (but limited to down field) and at a constant velocity. These are what I am considering the receiver's "*run parameters*".

To make sure the pass can be completed, I will also limit the receiver's maximum velocity to be well less than the maximum velocity of the thrown ball. I am also assuming that the quarterback throws the ball immediately (e.g., at time 0 in the model). Finally, to make the model require a little less precision, I am assuming that any throw that gets to within one unit of the receiver can be caught (e.g., there does not have to be a mathematically perfect intersection of the path of the receiver and ball and the exact time of the landing of the ball.)

There is also one other implicit parameter lurking in the model: gravity. Since the path of the receiver and the trajectory of the ball are not particularly complex in terms of their mathematics, I decided to make things a little more interesting by allowing the value of gravity to be changed for each run scenario. Thus, gravity is another initial parameter that is supplied along with the "*run parameters*" as the set of initial model parameters:



This table shows the "run parameters".

Based on my initial sketch, it became clear that there were only a few "things" the quarterback could do to affect the throw that they are making. They can choose the

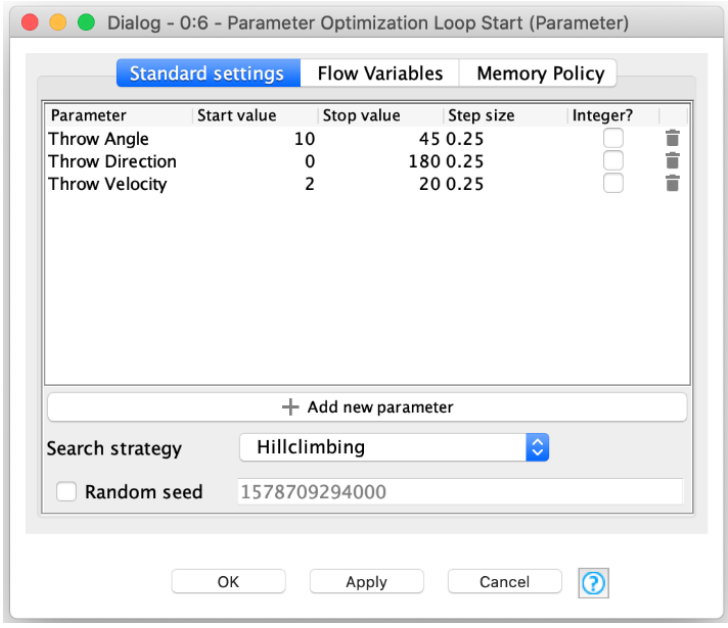
direction of the throw, the angle of the throw, and the speed of the throw. I'm considering these the "*throw parameters*".

The problem now becomes: for an arbitrary set of run parameters, what are the optimal throw parameters? To answer that question, we have to define what we mean by optimal. Obviously, an incomplete pass is not optimal but given that we have multiple throw parameters we can vary, we can make any number of completed passes. Since the goal of American football is to advance the ball as far down the field as possible, we will define optimal in this case as the pass that is caught the farthest down field from the quarterback.

As I mentioned earlier, we could just randomly pick values for these parameters until we find a set that completes a pass. This could take a long time and it would not guarantee that we found the set of parameters that results in the optimal pass (based on our definition of optimal). But maybe we could just keep picking enough random sets of parameters until we found a large number of completed passes with the hopes that one of them will prove to be close to the optimal pass? This is BFI — Brute Force and Ignorance. If we make an infinite number of passes, one of them is bound to be close enough to the optimal pass….

But this is not practical; particularly since we might need to generate a broad set of training data based on many different receiver's run parameter values. What if we could guide the random parameter value selection? After all, if we make a throw in the real world, we can tell how *close* we came to hitting our target. Any miss would be considered an error, and we can even measure the magnitude of the error based on how far away the ball landed from the receiver. A person could and should learn from that error and adjust their next try accordingly.

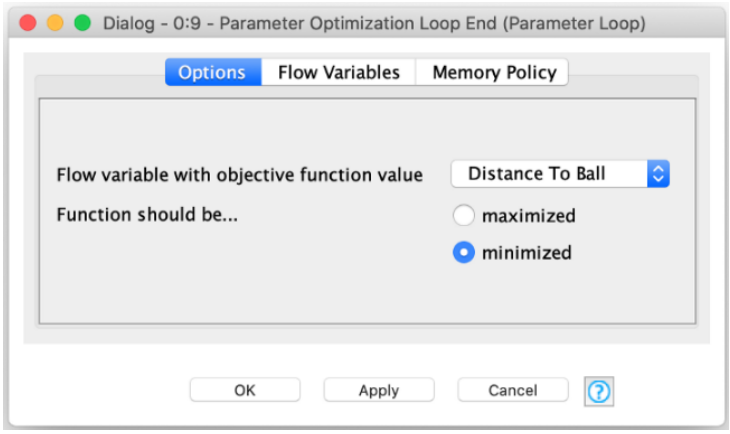
Well, KNIME can do the same thanks to the pair of innocuous but extremely powerful nodes: the *Parameter Optimization Loop* nodes. These nifty nodes allow you to specify one or more parameters as flow variables and designate their range of values for each of them. The nodes will then loop over a set of nested nodes until the optimal set of parameter values is found. Sounds familiar, right? Here is the *Parameter Optimization Loop Start* node as configured for our data generation task:



The configuration dialog of the Parameter Optimization Loop Start node.

You can see the three throw parameters I discussed earlier along with the constraints on their values to fit within the model I constructed. But you may also notice that there is a configurable search strategy — this is how you tell the node the best way to find the optimal parameter sets and we cover the various strategies shortly.

As I mentioned earlier, the goal is to optimize the parameter values to minimize the error. In our case, I'm defining the error as merely the distance between the receiver and the ball at the end of the throw and it's really simple to tell KNIME to optimize that in the *Parameter Optimization Loop End* node:

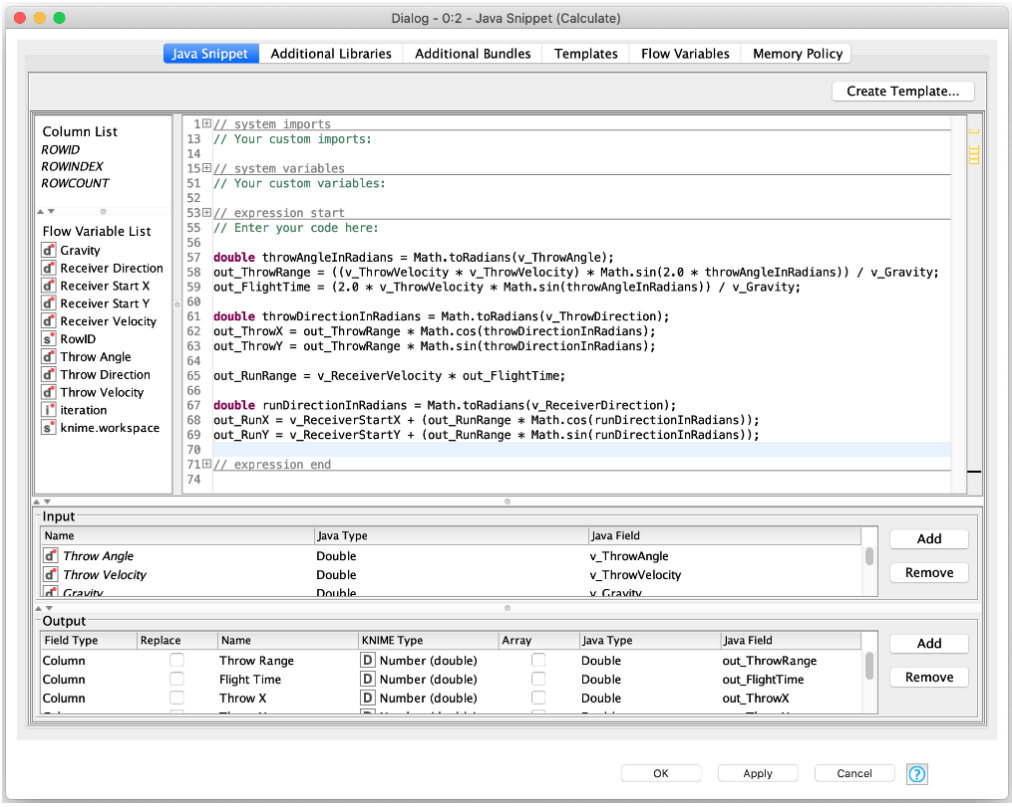


The configuration dialog of the Parameter Optimization Loop End node.

There are two outputs from the loop end node. The first is the optimal set of parameter values (e.g., the one with the minimal error value). The second is the set of all tested

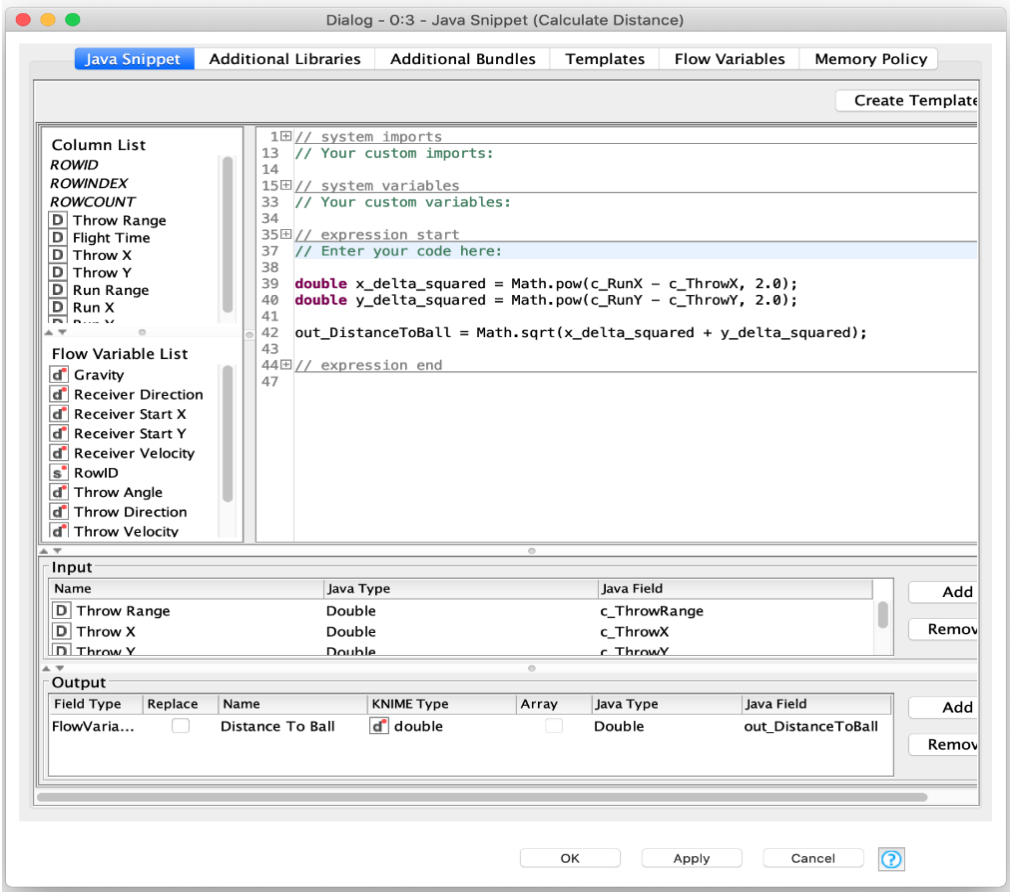
parameter value combinations and their corresponding error values. This is an important distinction to make for our scenario; remember that we are not just looking for the most accurate throw but for the throw that arrives within one unit of the receiver (e.g., error less than 1.0) and results in the receiver as far down the field as possible (e.g., the maximal Y position for the receiver).

Before I go into how we select this optimal throw from the loop's output, let us touch on what is happening inside of the loop. The first thing the loop does is calculate the throw's characteristics based on the current set of throw parameters. This includes not just the coordinates of where the throw lands but also its time in flight. The time in flight is then used to calculate the receiver's position at the end of the throw. All of this is done in a Java Snippet node:



With this Java code snippet, the throw's characteristics are calculated and the receiver's position at the end of the throw is determined.

Now that we know where the ball lands and also where the receiver is when the ball lands, it is easy to calculate the distance between those two points to use as our error measure:



With this Java code snippet, the distance between where the ball lands and where the receiver stands is calculated and used as our error measure.

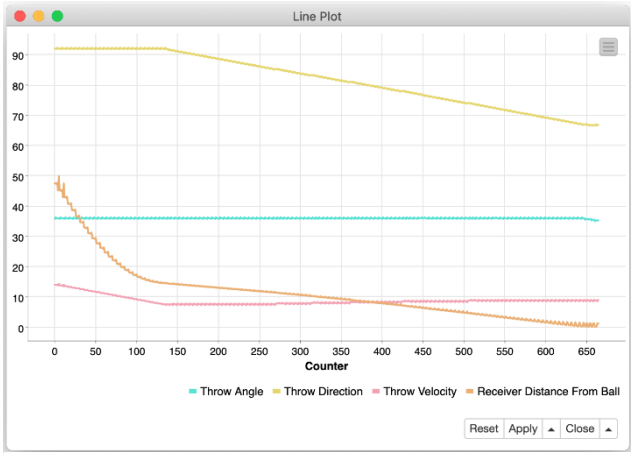
This is all we need to model each throw for a given set of parameters and let the KNIME optimization nodes do their magic. But I did earlier mention that the nodes support a selectable search strategy for how to search for an optimal set of parameter values. The first strategy is "*Brute Force*" which will try every possible combination of parameter values using the constraints and step sizes that are configured in the loop start node. If we want our data generator to finish in any reasonable amount of time this is not a practical approach for our model since it would result in many millions of iterations of the loop for each training data record.

The next impractical approach for us is the "*Random Search*" strategy where the loop nodes will try random values for each parameter in the hopes that one of the sets will result in a minimal error. You can tell the node how many random sets of values to try and even tell it to stop prematurely if it finds a set of values that is good enough based on a configurable tolerance. Again, with our large set of potential values and desire to find something close to the optimal throw in terms of distance down field, we need an approach that is more likely to give us the best potential sets of parameters.

This is where the remaining two search strategies come into play. The first is "*Hillclimbing*" where the loop picks a random starting set of parameters and then evaluates the set of all the neighbors (e.g., using the step sizes to move each parameter value in both directions) to see which neighbor moves the most towards minimizing the error. This process is repeated a configurable number of times or until no neighbors result in an improvement of the error score.

The final search strategy is "*Bayesian Optimization*" where multiple initial random parameter combinations are considered and then the Bayesian Optimization strategy attempts to use the error value characteristics of these combinations to select the next round of parameter value combinations to try. This also continues for a configurable number of iterations.

When deciding between which of the two more sophisticated search strategies to use, you need to know the nature of your error function: how smoothly and quickly it converges and if it may have local minima instead of just one global minimum. Fortunately, you can leverage the power of KNIME to test the strategies and visualize their operation. For example, when we used the "*Hillclimbing*" strategy you can see the loop start with a relatively large error of almost 50 units from the ball and it initially adjusts the velocity and then soon starts adjusting the throw direction before finally tweaking the angle of the throw:



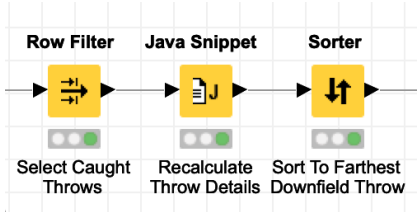
A line plot showing the parameters of the optimization loop when using the Hillclimbing strategy.

If we switch to the "*Bayesian Optimization*" strategy, you see a much more chaotic search since it is starting from multiple random points and considering a wider variation of parameter combinations:

A line plot showing the parameters of the optimization loop when using the Bayesian Optimization strategy.

As I mentioned previously, since our error function does not optimize for the receiver's distance down the field, we are not just looking for the lowest error score but an acceptable error score (e.g., less than 1.0 for a catch) with the largest Y distance traveled by the receiver. Based on these requirements and the above graph of the search strategy, it is clear that the "Bayesian Optimization" strategy produces many more potential completed throw candidates at the small cost a few hundred extra iterations to do so.

We then take all of the attempted parameter sets tried by the optimization loop and filter them down to just the ones that are considered to be caught (e.g., error value of less than 1.0). Since the optimization loop only outputs the parameter values and their associated error scores, we need to recalculate the coordinates of the catch and then we sort all of the completed throws based on the Y parameter of the catch so that the first record contains our optimal throw.



A snippet of the parameter optimization workflow. With these nodes, the optimal throw is selected.

These nodes give us a table of the completed throws sorted in order from farthest downfield to least downfield:

Sorted Table - 0:28 - Sorter (Sort To Farthest)

File Hilite Navigation View

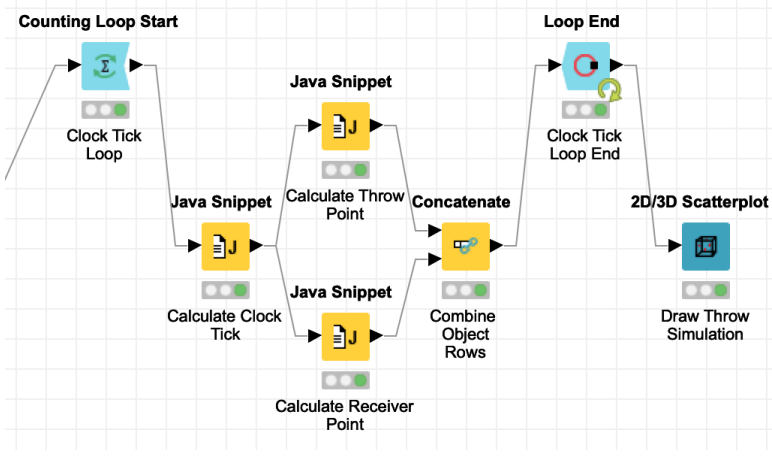
Table "default" - Rows: 216 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Throw Angle	Throw Direction	Throw Velocity	Throw Y	Receiver Distance From Ball
Row885	18.677	62.948	8.938	21.583	0.998
Row983	18.542	63.175	8.896	21.294	0.907
Row907	18.663	61.924	8.921	21.289	0.856
Row714	18.088	61.644	8.985	20.968	0.936
Row581	18.062	61.909	8.977	20.958	0.903
Row987	18.19	63.399	8.879	20.907	0.907
Row804	17.905	61.406	9.012	20.864	0.987
Row827	18.292	61.488	8.907	20.772	0.723
Row635	17.862	62.938	8.919	20.679	0.845
Row950	18.587	63.794	8.733	20.672	0.844
Row653	17.631	62.628	8.98	20.672	0.909
Row964	18.253	61.617	8.88	20.637	0.614
Row755	17.851	61.773	8.947	20.581	0.743
Row823	18.732	62.952	8.707	20.534	0.486
Row945	18.36	62.076	8.806	20.484	0.411
Row657	17.56	61.732	8.99	20.476	0.813
Row757	17.501	63.453	8.924	20.434	0.993

The output table of the Sorter node. The table contains all completed throws sorted in order from farthest downfield to least downfield.

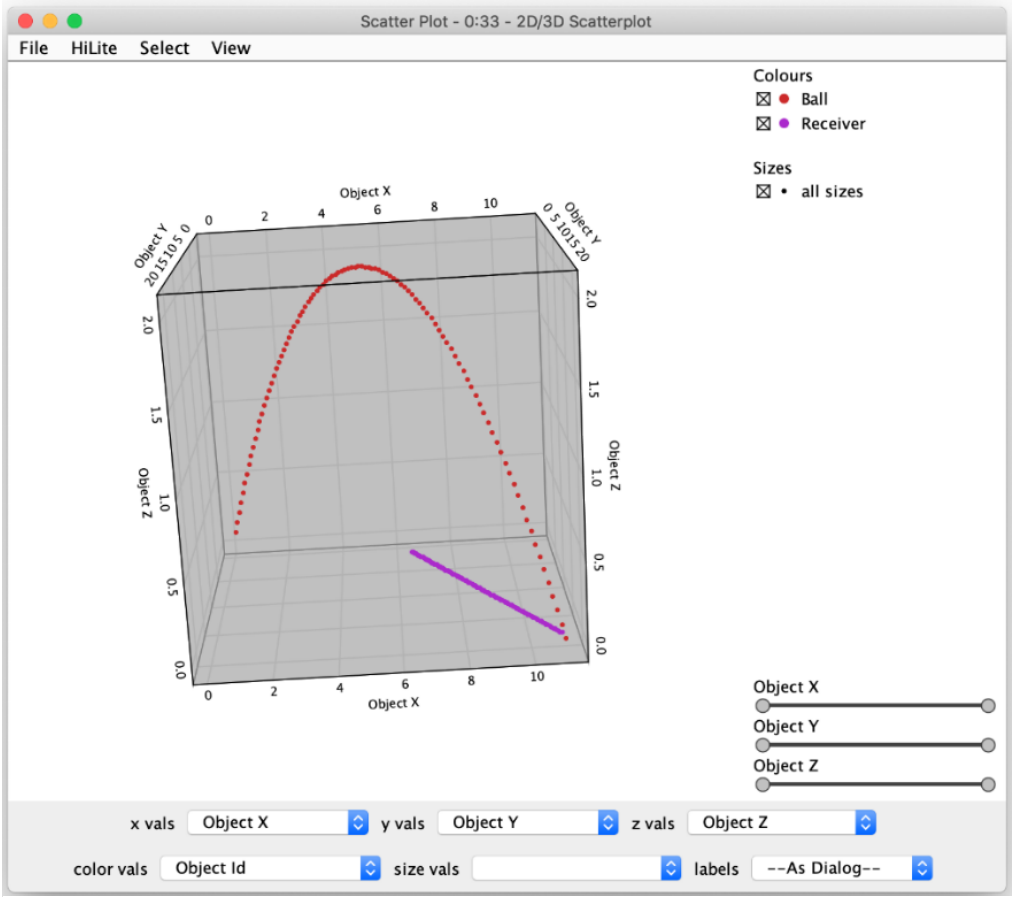
One interesting thing to note is that because we used the "*Bayesian Optimization*" search strategy, the best results were obtained at different iterations as can be seen from their row ID's which were generated during the parameter optimization loop before the table was sorted on the Throw Y column.

Finally, to help visualize the results and verify that the selected throw parameters are indeed accurate I've included a second loop that simulates the trajectory of the ball and the position of the receiver for 100 "*clock ticks*" evenly distributed over the flight time of the ball:



A snippet of the [parameter optimization workflow]. With these nodes, the optimal throw simulation is generated.

By using the powerful 2D/3D Scatterplot node, which is freely available as a KNIME community node, we can then display the simulation results in an interactive graph visualization:



A 3D scatter plot that shows the simulation results in an interactive graph visualization.

In the screenshot above, the quarterback is located at position (0, 0) in the back left of the cube. The receiver's path is represented by the purple points coming towards the camera and angling away from the quarterback. The quarterback's throw's trajectory is shown by the red dots and even includes the trajectory to better illustrate the relationship between the pass and the receiver over time.

Now that we have verified that given a single set of starting receiver "*run parameters*" we can produce a near-optimal throw for passing to the receiver, we now need to build an enclosing loop to an arbitrary number of records with initial "run parameters" and calculate their throw parameters using the technique we evolved in this blog posting. This will be our training data set for trying to teach my laptop how to throw the ball, and my next blog posting will pick up here and cover the teaching of various machine learning algorithms and include a contest to see which one "*learns the best*" for our simplistic scenario.

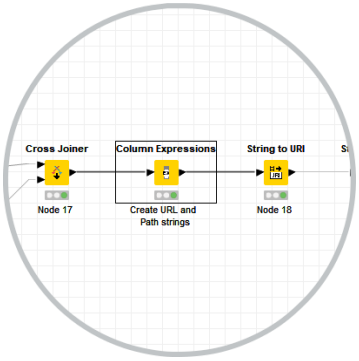
This article was originally published on BigBear.AI and we republished it in our Low Code for Advanced Data Science Journal on Medium here.

The corresponding Making the Pass, Part 1 workflow can be found on the KNIME Community Hub in Paul's public space.

You can continue reading part 2 on Medium at Making the Pass, Part 2: Training a Neural Network with KNIME or on BigBear.AI.

John Emery was nominated KNIME Contributor of the Month for July 2022. He was awarded for his countless blog posts about parsing and analyzing PDF documents, running baseball hitting streak simulations, or cracking Wordle. Furthermore, John is also a respected speaker at data science events, and he is the first certified trainer of KNME Software (check out his badge). He is currently part of the organizing board of the Data Connect North America series of events. The image on the right shows a workflow snippet of his PDF parsing workflow.

John is a Principal Consultant ad phData where he focuses on using analytics engineering tools such as KNIME to help clients use their data more effectively. He gained experiences working as Operation Research Analyst at the US Army and as consultant in the financing industry in the past. He is particularly interested in geospatial analysis and outside of work he likes to do film photography, hiking, and mountaineering.



Visit John's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `johnvemery`).

Using KNIME to Parse and Analyze PDF Documents

Extracting daily weather recordings with just a few clicks

Author: John Emery

In my professional life I work with data; preparing, cleaning, building models, and creating visualizations. I spend the vast majority of my time using tools like Tableau, Power BI, and Google Data Studio for visualization work, while I build ETL workflows in tools such as Alteryx, KNIME, and Tableau Prep. Outside of work, I have a goal of standing at the highest point of all 50 states. To date, I have been to 24. On October 31st this year, I flew up to New Hampshire in an attempt to summit Mt. Washington, that state's highest point — and my would-be halfway point.

WS FORM F-6

PRELIMINARY LOCAL CLIMATOLOGICAL DATA

STATION: MOUNT WASHINGTON OBSERVATORY

MONTH: OCTOBER — YEAR: 2021

LATITUDE: 44 DEGREES 16 MINUTES NORTH — LONGITUDE: 71 DEGREES 18 MINUTES WEST — GROUND ELEVATION (H): 6280 FT — STANDARD TIME: EASTERN

	TEMPERATURE (°F)					DEGREE DAYS		PRECIPITATION (IN.)			WIND (MPH)			SUNSHINE		SKY	
DAY	MAX	MIN	AVG	NORM	DEPART	HEAT	COOL	TOTAL (EQUIV)	SNOW & ICE	SNOW/ICE ON GROUND-7AM	AVG SPEED	FASTEST MILE SPEED	FASTEST MILE DIR	(MINUTES) TOTAL	% POSS	COVER (TENTHS)	WEATHER OCCUR.
1	32	27	30	37	-7	35	0	0.04	T	1	49.2	77	310 (NW)	0	0	10	1246
2	42	31	37	37	0	28	0	0.85	T	1	37.1	55	300 (NW)	0	0	10	1246
3	45	41	43	37	6	22	0	0.98	0.0	0	23.6	44	280 (W)	0	0	10	12
4	48	41	45	36	9	20	0	0.02	0.0	0	6.8	14	210 (SW)	235	33	9	12
5	50	43	47	36	11	18	0	0.01	0.0	0	5.0	14	310 (NW)	60	9	10	12
6	52	41	47	35	12	18	0	0.00	0.0	0	6.6	15	330 (NW)	697	99	3	12
7	53	41	47	35	12	18	0	0.00	0.0	0	14.3	38	290 (W)	699	100	3	12
8	52	38	45	34	11	20	0	0.00	0.0	0	8.8	27	340 (N)	680	98	5	12
9	51	38	45	34	11	20	0	0.00	0.0	0	9.4	33	230 (SW)	654	95	8	
10	48	39	44	34	10	21	0	0.00	0.0	0	17.4	34	290 (W)	285	41	10	12
11	50	43	47	33	14	18	0	0.00	0.0	0	9.1	34	280 (W)	370	54	10	12
12	60	47	54	33	21	11	0	0.00	0.0	0	7.8	18	320 (NW)	683	100	8	
13	57	45	51	32	19	14	0	0.00	0.0	0	12.1	28	190 (S)	676	99	5	12
14	48	44	46	32	14	19	0	T	0.0	0	13.0	30	300 (NW)	61	9	10	12
15	50	43	47	32	15	18	0	0.01	0.0	0	19.3	34	250 (W)	90	13	10	12
16	52	38	45	31	14	20	0	2.01	0.0	0	32.3	69	180 (S)	0	0	10	12
17	39	28	34	31	3	31	0	0.26	0.0	0	39.4	61	290 (W)	0	0	10	126
18	28	23	26	30	-4	39	0	**0.72**	**4.8**	1	41.0	92	290 (W)	0	0	10	126
19	26	20	23	30	-7	42	0	0.45	2.3	4	67.0	92	290 (W)	0	0	10	126
20	38	25	32	30	2	33	0	0.00	0.0	5	54.5	88	300(NW)	420	64	8	126
21	42	36	39	29	10	26	0	**0.12**	0.0	4	32.0	57	260 (W)	105	16	10	12
22	41	30	36	29	7	29	0	0.33	0.0	T	34.2	64	220 (SW)	0	0	10	126
23	34	18	26	29	-3	39	0	0.00	0.0	T	20.6	47	290 (W)	535	82	8	126
24	24	14	19	28	-9	46	0	0.00	0.0	T	46.4	81	290 (W)	380	59	8	126
25	40	24	32	28	4	33	0	0.38	3.6	2	22.1	46	170 (S)	0	0	10	1246
26	41	33	37	28	9	28	0	**1.02**	0.0	T	48.9	83	100(E)	0	0	10	12
27	39	30	35	27	8	30	0	0.39	0.0	0	55.8	85	060 (NE)	85	13	10	12
28	37	33	35	27	8	30	0	0.00	0.0	0	23.3	40	090 (E)	635	100	5	
29	39	30	35	26	9	30	0	0.00	0.0	0	24.8	39	120 (SE)	633	100	0	
30	39	29	34	26	8	31	0	1.74	T	0	42.9	92	110 (E)	0	0	10	1246
31	41	32	37	26	11	28	0	3.27	T	0	49.4	89	110 (E)	0	0	10	124
SUM	1338	1045	----	----	----	815	0	12.60	10.7	----	874.1	----	----	7983	----	260	----
AVG	43.2	33.7	----	----	----	----	----	----	----	----	28.2	FASTEST	DIR	POSS	38%	8.4	----
											MISC. ->	**92**	**110 (E)**	20805			----

TEMPERATURE DATA (°F)

AVERAGE MONTHLY 38.4
DEPARTURE FROM NORMAL 7.1
HIGHEST 60 on 12th
LOWEST 14 on 24th

NUMBER OF DAYS WITH:

PRECIPITATION DATA (IN.)

TOTAL FOR THE MONTH 12.60
DEPARTURE FROM NORMAL 2.61
24 HOUR MAX 4.15 on 30th/31st

SNOWFALL, ICE PELLETS (IN.)

TOTAL FOR THE MONTH 10.7

WEATHER

NUMBER OF DAYS:

CLEAR (SCALE 0-3) 3
PARTLY CLOUDY (SCALE 4-7) 3
CLOUDY (SCALE 8-10) 25

WITH .01" OR MORE PRECIP 17

SYMBOLS USED IN COLUMN 16

1 = FOG
2 = FOG REDUCING VISIBILITY TO 1/4 MILE OR LESS
3 = THUNDER
4 = ICE PELLETS
5 = HAIL
6 = GLAZE OR RIME

Monthly published data by the Mt. Washington Observatory.

Mt. Washington is world-renowned for its horrendous weather. Wind gusts in excess of 230 miles per hour have been recorded, and about one-third of the days of the year

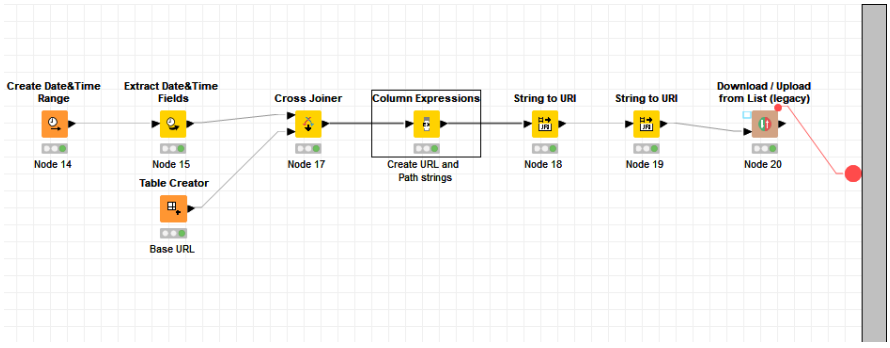
experience winds above 100 miles per hour. Although many visitors drive or take the railway to the top during the milder summer months, the weather on the mountain can be extremely dangerous — especially heading into winter. Of course, this makes the idea of climbing it much more appealing to me.

Planning a hike/climb in late October means you have to navigate the potential for early season winter weather. As a data professional, I was interested in seeing what October weather typically looked like to get an idea of my likelihood of reaching the summit. Thankfully, the Mt. Washington Observatory keeps a log of daily weather recordings — high and low temperatures, average and fastest winds, precipitation, etc. — for each month going back to 2005.

Getting the Data

What makes analyzing this data tricky is that the observations are saved as a PDF. As many analysts, data scientists, and other data professionals can attest, dealing with data in PDFs can be a challenge.

To collect the data, I used KNIME to save each month's PDF to my computer. (As an aside, KNIME is a wonderful **free** and open-source data prep and data science tool. If you want a no-or-low code solution to your data prep needs, it is an excellent choice.)



This workflow snippet creates Date&Time ranges to build custom URLs and path strings.

The URL for the Mt. Washington weather observations is:
`https://www.mountwashington.org/uploads/forms/2021/10.pdf`

The only thing that changes from one month to the next is the year (2021, 2020, 2019…) and the month (10, 09, 08…). Using tools such as *Create Date&Time Range* and *Table Creator* I was able to easily generate all possible combinations of months and years to build out all URLs going back to January 2005. From there, I was able to download the PDFs to a folder on my computer using two *String to URI* nodes. The first node creates a URI path string from the source and the second creates URI path string to the target destination on my computer. As I write this in November 2021, the download process takes about a minute for over 200 PDFs.

Prepping the Data

Downloading the PDFs onto a drive on my computer was relatively easy. All that I had to do was generate strings for each URL and then download them. Prepping data from a parsed PDF can be an absolute nightmare, however. Depending on the structure of the PDF that you need to parse, this task can range from quite simple to nearly impossible.

Thankfully, KNIME offers a node called *Tika Parser*. As KNIME describes it, this node "allows parsing of any kind of documents that are supported by Tika." The *Tika Parser* node is ridiculously easy to configure. I simply selected the directory that housed the downloaded PDFs, selected the file type from a list, and then which metadata items I wanted to output. In this case I selected the file path and main content. The resulting content output is a long text string from each PDF.

Port Output | Port 0 | Load data | Rows: 203, Columns: 2

ID	Filepath	Content
Row0	C:\Users\John Emery\Documents\Data\Mt Washington\200501 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL DA
Row1	C:\Users\John Emery\Documents\Data\Mt Washington\200502 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row2	C:\Users\John Emery\Documents\Data\Mt Washington\200503 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row3	C:\Users\John Emery\Documents\Data\Mt Washington\200504 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row4	C:\Users\John Emery\Documents\Data\Mt Washington\200505 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row5	C:\Users\John Emery\Documents\Data\Mt Washington\200506 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row6	C:\Users\John Emery\Documents\Data\Mt Washington\200507 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row7	C:\Users\John Emery\Documents\Data\Mt Washington\200508 Weather.pdf	_WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL DATA M
Row8	C:\Users\John Emery\Documents\Data\Mt Washington\200509 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row9	C:\Users\John Emery\Documents\Data\Mt Washington\200510 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row10	C:\Users\John Emery\Documents\Data\Mt Washington\200511 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row11	C:\Users\John Emery\Documents\Data\Mt Washington\200512 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row12	C:\Users\John Emery\Documents\Data\Mt Washington\200601 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL DA
Row13	C:\Users\John Emery\Documents\Data\Mt Washington\200602 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row14	C:\Users\John Emery\Documents\Data\Mt Washington\200603 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row15	C:\Users\John Emery\Documents\Data\Mt Washington\200604 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL D
Row16	C:\Users\John Emery\Documents\Data\Mt Washington\200605 Weather.pdf	WS FORM F-6 STATION MOUNT WASHINGTON OBSERVATORY PRELIMINARY LOCAL CLIMATOLOGICAL D

Raw data from the Tika Parser… not the prettiest thing — yet!

Once the data is in its raw string form like this, we can begin splitting it apart — putting fields in distinct columns and each day's observations on distinct rows. To start, I used the *Cell Splitter* node to split the Content column into a list based on the newline delimiter (\n). Interestingly, in KNIME when you want to split a string into distinct rows, you go first split the data into a list and then use the *Ungroup* node to split the list into rows. This procedure turns our 203-row data set into a 26,566-row data set.

One thing to note, however, is that the *Tika Parser* node parses the **entire** PDF. This includes the data that you want and data that you may not want. In my case, I was only interested in the daily weather observations; data between the "*DAY*" and "*31*" records in the image above — everything else had to go.

Filtering out string data dynamically is a challenge for many analysts. You don't have the luxury of saying "*X > 100*" or any nice mathematical formula. For this exercise, I struggled here. I knew I wanted to use regular expressions to extract the characters up until the first space, but KNIME doesn't have a built-in function for regular expression extractions. Thankfully, what does exist is a third-party node called *Regex Extractor* — precisely what I needed!

Using the *Regex Extractor* node, I was able to pull out the first "*word*" from each string and then filter when it either said "*DAY*" or was a number between 1 and 31.

Content_SplitResultList
1 29 1 15 6 9 50 0 0.51 1.9 13 70.1 113 W 0 0 10 1234569
2 30 5 18 6 12 47 0 0.33 0.3 13 39.5 80 W 355 65 9 1246
3 30 16 23 6 17 42 0 0.05 0.0 13 58.3 80 W 0 0 10 126
4 20 5 13 6 7 52 0 0.03 0.1 13 43.0 69 W 515 94 4 126
5 7 -3 2 6 -4 63 0 0.00 0.0 13 50.0 70 W 120 22 9 126
6 22 7 15 6 9 50 0 0.50 3.2 13 29.0 64 S 0 0 10 1269
7 20 1 11 6 5 54 0 0.20 1.3 14 52.9 87 W 150 27 8 1269
8 25 12 19 5 14 46 0 0.23 3.1 14 24.5 42 NW 0 0 10 126
9 22 8 15 5 10 50 0 T 0.2 15 21.2 47 SW 525 94 5 1269
10 21 2 12 5 7 53 0 0.38 2.0 15 55.9 108 NW 0 0 10 1269
11 16 -11 3 5 -2 62 0 T 0.2 13 50.2 104 W 425 76 5 1269
12 32 14 23 5 18 42 0 0.17 1.0 13 23.5 62 W 0 0 10 1246
13 41 32 37 5 32 28 0 0.00 0.0 12 55.8 103 SW 0 0 10 126
14 43 -6 19 5 14 46 0 1.64 3.3 5 64.1 103 W 0 0 10 1269
15 -4 -10 -7 5 -12 72 0 T 0.2 6 44.7 79 W 268 47 7 126
16 4 -7 -2 5 -7 67 0 0.04 1.0 5 19.5 45 W 0 0 10 126

We're getting there…

With only numeric data left, I finally split the data into distinct columns based on the space delimiter. From here, only standard data prep issues stood in my way — renaming columns, ensuring fields were given appropriate data types, etc.

The Results

This workflow took me about one hour to build from beginning to end. Using just a few simple and easy-to-configure nodes, I was able to construct URLs, download PDFs to my computer, parse each PDF, and then reconstruct the data into an easy-to-use data table.

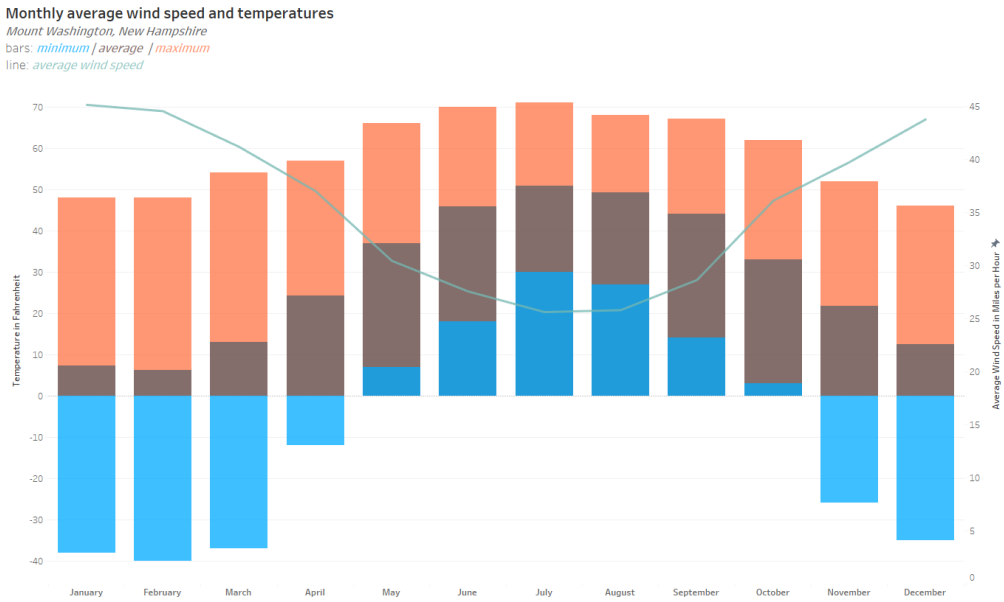
Port Output | Port 0 | Load data | Rows: 6179, Columns: 20

ID	date	Day	Max Temp	Min Temp	Avg Temp	Normal Temp	Departure	Heating Degree Days	Cooling Degree Days	Total Precipitation	Snow and Ice	Snow and Ice on Ground	Avg Wind Speed
Row0	200501	1	29	1	15	6	9	50	0	0.51	1.9	13	70.1
Row1	200501	2	30	5	18	6	12	47	0	0.33	0.3	13	39.5
Row2	200501	3	30	16	23	6	17	42	0	0.05	0.0	13	58.3
Row3	200501	4	20	5	13	6	7	52	0	0.03	0.1	13	43.0
Row4	200501	5	7	-3	2	6	-4	63	0	0.0	0.0	13	50.0
Row5	200501	6	22	7	15	6	9	50	0	0.5	3.2	13	29.0
Row6	200501	7	20	1	11	6	5	54	0	0.2	1.3	14	52.9
Row7	200501	8	25	12	19	5	14	46	0	0.23	3.1	14	24.5
Row8	200501	9	22	8	15	5	10	50	0	0.001	0.2	15	21.2
Row9	200501	10	21	2	12	5	7	53	0	0.38	2.0	15	55.9
Row10	200501	11	16	-11	3	5	-2	62	0	0.001	0.2	13	50.2
Row11	200501	12	32	14	23	5	18	42	0	0.17	1.0	13	23.5
Row12	200501	13	41	32	37	5	32	28	0	0.0	0.0	12	55.8
Row13	200501	14	43	-6	19	5	14	46	0	1.64	3.3	5	64.1
Row14	200501	15	-4	-10	-7	5	-12	72	0	0.001	0.2	6	44.7
Row15	200501	16	4	-7	-2	5	-7	67	0	0.04	1.0	5	19.5

This is much better.

Having clean data is perhaps the most important piece of the puzzle to perform sound analysis. Messy, incomplete, or poorly structured data not only cause delays but also frequently result in inaccurate reporting. It ends up being a waste of time for all involved. Using a tool like KNIME, I took existing data in a challenging structure for analysis and molded it into the structure I needed.

….. so what about the results from Mt. Washington?



Monthly average wind speed and temperatures, measured at Mt. Washington, New Hampshire.

The day of my climb saw nearly 4 inches of rain (over 5 inches total fell on the day of the climb and the day before) and average winds of about 50 miles per hour and gusts nearing 90. Looking over the historical data, it was the 15th rainiest day and around the 80th percentile for average wind speed. In short, not a great day for climbing!

This article was originally published on LinkedIn Pulse and we republished it in our Low Code for Advanced Data Science Journal on Medium here.

Education and Research

In this section, we compiled all the articles that contain a teaching purpose. This includes academic contributions focusing on Drug Discovery or Gene Ontology, but also other educational articles like tutorials on how certain concepts can be implemented in KNIME Analytics Platform. The category “Education and Research” features our teachers, scientists, and academics:

- **Keith McCormick**

  – Recognized Analytics Leader & Independent Predictive Analytics and Machine Learning Consultant

- **Giuseppe Di Fatta**

  – Professor *@Free University of Bozen-Bolzano*

- **Alzbeta Tuerkova**

  – Head of Computer-Aided Drug Design *@Celeris Therapeutics*

- **Malik Yousef**

  – Head of Galilee Center for Digital Health Research *@Zefat Academic College*

- **Nick Rivera**

  – Business Analyst *@EMR*

- **Francisco Villarroel Ordenes**

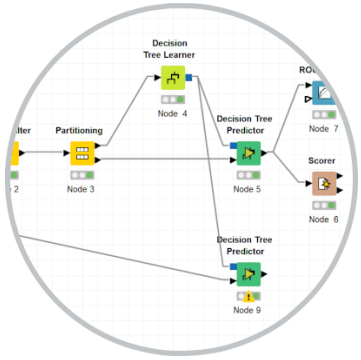
  – Assistant Professor of Marketing *@LUISS Guido Carli University*

- **Christophe Molina**

  – Freelance Data Analyst, CEO *@PIKAÏROS*

Keith McCormick was nominated KNIME Contributor of the Month for December 2020. He was awarded for his courses on LinkedIn Learning: Introduction to Machine Learning with KNIME, and Data Science Foundations - Data Assessment for Predictive Modeling. In the first course, Keith shows how KNIME supports all phases of the CRISP-DM cycle in one platform – including merging and aggregation, modeling, data scoring, and the R and Python integrations. The second course focuses on principles, guidelines, and tools such as KNIME and R for data access, making them suitable for machine learning.

Keith is an independent instructor, at both the University of California at Irvine (UCI) and on LinkedIn Learning. Keith not only works as a teacher, he is a teacher at heart. He enjoys preparing his students for the professional world, provoking that effect of surprise when revealing insights from data, and grading his students. Yes, grading, since this is the moment where he can interact with the work of the students, not only to criticize, but also to explain and praise (see Keith's interview in DCE Magazine).



Visit Keith's space on the KNIME Hub or his profile page in the KNIME Forum (Hub/Forum handle: `keith_mccormick`).

Where is Data Science Education going? Let the Expert speak

My Data Guest – An Interview with Keith McCormick

Author: Rosaria Silipo

It was my pleasure to recently interview live on LinkedIn Keith McCormick as part of the *My Data Guest* interview series. He shared insights into the world of data science education, explained how he uses KNIME in his teaching, and spoke about the importance of good mentorship for young data scientists.

Keith McCormick is an independent data scientist, trainer, conference speaker, and author. For over 25 years, he has guided data science teams to establish highly effective analytical practices across industries, including the public sector, media, marketing, healthcare, retail, finance, manufacturing, and higher education.

Rosaria: *You have been teaching data science for how many years now? Maybe you can give us a little summary of your journey as a data science teacher.*

Keith: Well, it has been a journey of 25 years already. In my 20s, I started out as an independent statistics software trainer, teaching traditional topics in statistics such as discriminant analysis and logistic regression with SPSS. It was originally going to be a side-hustle during graduate school to pay the bills but eventually became a career. The next critical milestone for me was the encounter with visual analytics tools. In the late 90s, SPSS bought a much smaller company that developed one of the early predictive analytics workbenches using visual data flows. It was the beginning of no-code/low-code tools in data analytics. Ever since these tools have evolved substantially, reaching a very high level of sophistication. I've always been a huge fan of low-code/no-code software, and indeed I've been using them for 20 years. More recently, for the past seven years, I've been teaching at UC Irvine, and for LinkedIn Learning.

Rosaria: *Has data science changed considerably in these 25 years?*

Keith: Some things really stayed the same but I have seen a major change around the no-code/low-code debate. Back in the 90s when SAS and SPSS dominated the space,

there was almost a "civil war" between the SPSS/SAS programmers and people who used menus to configure and execute data flows. There was the misconception that data analytics was the exclusive realm of coders, and if you were using menus you felt your competences were questioned and your results less valued. This is crazy, even more so if you consider that the theory behind it is absolutely the same regardless of whether you're building, for example, a regression with code or without. Then, for some years the momentum shifted to programming languages, such as Python and R, before shifting back to no-code/low-code tools. People are realizing that a visual tool can be of great support to save time and prototype solutions before moving them to production. And KNIME is fantastic both in prototyping and production!

Rosaria: *Where do you teach currently?*

Keith: I have been teaching for UC Irvine's extension program since 2015. Since January of this year, I have been teaching one of my favorite courses there. The course focuses on the data understanding phase of CRISP-DM, which is often misunderstood in my opinion. When students hear "data exploration", they immediately think of data visualization. While data visualization is a crucial part of data exploration, it's not visualization for reporting. In the data understanding phase of CRISP-DM, we want to uncover aspects of the data that we either have to fix during data preprocessing, or that will give us insights into what algorithms might be the best fit. It's data visualization with an eye to modeling. Realizing that is often surprising to students who usually think of data visualization just for reporting.

Besides that, sometimes I also teach some introductory machine learning courses, where KNIME is my go-to tool.

Rosaria: *What topics do you cover in your UC Irvine courses and what are the learning outcomes for the students? In which of those courses do you use KNIME (and how)?*

Keith: In the case of UC Irvine specifically, they started a Predictive Analytics Certificate Program I have been a part of for quite a few years, teaching all modules over time. The program is structured around the CRISP-DM phases: from problem definition and data preparation, to introduction to modeling and deployment. So at the end of the program, students are exposed to all phases of the data science life cycle.

In my courses and in many courses of other instructors in our department, KNIME is the go-to tool for teaching. For example, in a course about introduction to modeling, where I cover every week a new algorithm, such as scale-dependent algorithms, classification algorithms, etc., what I really need is a tool that helps students see how the theoretical concepts that they read in the books can be implemented practically. From an educational perspective, KNIME offers a great opportunity to really practise theoretical concepts.

Note. In 2023, we are going to introduce a three course version of this certificate where I'm the only instructor. There will be more continuity, and learners will learn more KNIME. I'm excited about that.

Rosaria: *How would you describe your teaching approach and your experience of implementing concepts with KNIME?*

Keith: The first week I ask students to read the introductory chapters of Data Mining Techniques by Michael Berry and Gordon Linoff's book, and Dean Abbott's Applied Predictive Analytics, which is really popular in the community. Those are textbooks that give them a theoretical understanding of the key concepts in data mining. In addition to the readings, even before we start with hands-on homework, I ask them to download KNIME and install it. Next, to familiarize themselves with the tool, I have them pick any workflow available on the KNIME Examples Server or KNIME Hub and demonstrate that they got it to work and understood it. In the second week, I pick an example -say, binary logistic regression- and ask them to demonstrate that they understood it. Based upon what they have learnt in the readings and practicing KNIME, I provide a new dataset and challenge them to make the workflow work and execute successfully. I believe this is an effective way to let theoretical concepts sink in and give students a chance to experience first-hand what it means to model data.

Moreover, complementarily to my teaching, I often point students to the wonderful content available for free on KNIME TV so they can get a deeper understanding of the tool.

Rosaria: Do you describe any data science use cases in your courses? If yes, which ones?

Keith: There is definitely some practical learning where they experiment with datasets, especially those that you can find on the KNIME Examples Server, or those available in the repository of UC Irvine, or Kaggle. However, there is a limit to what you can do with a practice dataset, such as the very popular Iris dataset. Therefore, what I try to do is to take inspiration from what they may have seen on the news or on YouTube, and incorporate that in the form of discussions. Ideally, the next step is that they choose a new dataset describing an actual phenomenon and apply their knowledge to build real world examples.

Rosaria: Has COVID-19 and various lockdowns impacted your way of teaching?

Keith: In the months when Covid-19 was raging, most of my conference work had to stop. I used to teach in conferences like the Data Warehouse Institute Conference in the US, which usually took place four times a year. For quite some time, Covid-19 also changed my LinkedIn Learning work. I used to fly to Santa Barbara to record in a soundproof studio with my producer. It felt like narrating an audiobook, but that also had to stop during lockdown.

On the other hand, my teaching at UC Irvine was not affected by the pandemic as it has always been remote at least for me.

Rosaria: *Where has KNIME facilitated your teaching the most?*

Keith: I think KNIME is perfect for remote teaching and assignments. Back in the days, when I started teaching, I would walk around the room and help students whenever they got stuck. With remote teaching, that is often no longer the case. One factor is that perhaps I've learnt to be more thoughtful about how I structure KNIME assignments, which is likely to facilitate learning. In addition to that, online conference tools have improved so much in the last two years that whenever students need help, they simply request a 1-on-1 session with me on Zoom.

Rosaria: *You teach a number of courses on LinkedIn Learning, and some of them feature KNIME. What are the titles? What do these courses cover?*

Keith: That's right. I have been recording several courses for LinkedIn Learning in the last few years. To be precise, 19 courses. Some of them involve KNIME but some don't involve any software-related content at all - especially those courses that are targeting executives who are usually more interested in how to get value out of machine learning at the enterprise level. Among the courses where KNIME is the go-to tool, I am very proud of two ever-green ones: Introduction to Machine Learning with KNIME, and Data Science Foundations: Data Assessment for Predictive Modeling.

In terms of duration, the last course is quite an exception in the educational landscape of LinkedIn Learning. It's a 4-hour course that introduces a systematic approach to the data understanding phase for predictive modeling, such as data formatting, missing value analysis, etc. I think that KNIME is a great tool to grasp those concepts.

Rosaria: *Is there any new LinkedIn Learning course involving KNIME that you're especially happy about?*

Keith: In February 2022, a new course about machine learning and explainable AI was released. I am very happy that the course got pushed into this year because KNIME has added so many fantastic verified components for XAI that I was able to integrate in my course. XAI is such a hot topic at the moment but what you can usually find online is either high-level overviews or very math-intensive content. What was missing was something in between that starts from an overview but then provides a detailed explanation of a few techniques with concrete applications. This new course tries to close the gap.

> ***Note.*** More recently, I have updated my two Decision Trees courses for KNIME. When that is complete, it will bring the total of KNIME-related courses to five.

Rosaria: *How complicated was it to prepare for those courses?*

Keith: Preparing courses for LinkedIn Learning is fairly different from other remote courses. It usually takes about 6 months from conceptualization to the final product. Besides the prep time, there is a long lead time where courses are edited and polished before publication. For example, if I am designing a course that involves KNIME, I usually create a detailed outline and pass it on to the Evangelism team. They take a quick look and give me suggestions to help me pick the best nodes for the particular outline. As a result, when the course comes out online, it's as up-to-date as it can possibly be since it was reviewed by the team.

Rosaria: *Let's get back to your work as an educator who also teaches at university or other institutions. What do you think data science education is missing today?*

Keith: What is missing is a deep understanding of the data science life cycle, be it CRISP-DM or any other process model. Nowadays, in data science education very few instructors teach process models. In many cases, the focus on coding is so strong that process and life cycle issues never work their way into the curriculum. Coding is certainly important but if we don't integrate process and life cycle issues in the curriculum, we end up educating students who don't know how to do problem definition or don't have a competence at any of the phases prior to modeling. There is the misconception that all that matters is modeling algorithms and getting software to run the model. There is no room for data understanding, data preparation and business understanding.

Rosaria: *In the light of your long experience in the data science space, do you have any career advice for young data scientists?*

Keith: I would recommend that during their data science journey they devote their time and effort equally to process and life cycle, concepts, and execution. For example, if they are seeking out a program, a boot camp or any certificate courses where the focus is exclusively on coding and the data science life cycle and process are ignored, they should find a way to address those topics because they will need them on the job.

Another thing that they need is to get some real, practical experience, for example with a data science apprenticeship not unlike a medical student's residency. During this time, it's crucial that they get the right mentorship from somebody who is more experienced and can guide and inspire them.

Rosaria: *We are reaching the end of our interview. Before we say goodbye, we cannot but ask the classic question. Where do you see data science going in the next few years? What will the next innovation be?*

Keith: There is a lot of hype around AutoML, and it's likely that it will continue to grow. I see the value of AutoML, especially for prototyping solutions where time is money, and if it's presenting us with models that are worthy of proper evaluation.

Having said that, we need to debunk the myth that AutoML is going to replace humans in the process by simply pressing the deployment button. In CRISP-DM, after the Modeling phase comes the Evaluation phase. Evaluation does not refer to ranking models by accuracy - that is still part of the modeling phase. When we get to evaluation, we are expected to bring value to the organization and ask ourselves "Is the model ready for deployment?" or "Is there the potential for organizational resistance?". We have to keep in mind that we are building models for organizations that are filled with people. This means that we can't eliminate the human aspect - neither in how we define the problem, nor in the ways we engineer model features. Humans are in the process to stay.

Rosaria: *How can people from the audience get in touch with you?*

Keith: People can reach me either via LinkedIn, or on my website Keith McCormick.

This article was first published in our Low Code for Advanced Data Science Journal on Medium. Find the original version here.

Watch the original interview with Keith McCormick on YouTube: "My Data Guest – Ep 4 with Keith McCormick".

Giuseppe Di Fatta was nominated KNIME Contributor of the Month for May 2021. He was awarded for his academic activity and his help in shaping the KNIME certification program by designing many examination questions and helping in structuring the program. So, if you took the exam and found the questions too hard to answer… well, now you know what it takes to become a KNIME Contributor of the Month.

Giuseppe is currently a Professor at the Faculty of Computer Science of the Free University of Bozen-Bolzano, Italy, where he teaches the courses *Machine Learning* and *Data-driven Decision Making*. Before that, he served as Head of the Department of Computer Science at the University of Reading, UK. Among the many tools for data science, he also introduced KNIME Analytics Platform in his courses because he sees the benefit of an open source, low code data science tool in a student's portfolio for their future career. After his graduation at the University of Palermo, Italy, Giuseppe ventured into the academic world and even joined the initial KNIME development team at the University of Konstanz, Germany, until the first release of KNIME 1.0 in 2006.

Visit Giuseppe's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `difatta`).

Get Certified – Get the KNIME Certification Program

Are you an expert in KNIME Software?

Authors: Giuseppe Di Fatta & Stefan Helfrich

Editor's Note:

This article, dated 2019, describes the first step taken to establish the KNIME certification program. It is thanks to Prof. Giuseppe Di Fatta that the first pool of questions was designed and tested for all the KNIME users to be certified. Since then, the KNIME certification program has expanded and improved and will keep doing so. However, we are still grateful to Prof. Di Fatta for helping us take the first step. In the meantime, Prof. Giuseppe Di Fatta has changed his affiliation to the University of Bolzano in Italy.

There is an official way to answer this question and share it with the world: You can test your KNIME proficiency with a certification program developed in collaboration between academia and industry.

Professional certifications are particularly useful in the employment process to help identify key skills relevant to the job profile sought by employers. They facilitate matching the demand for skills with the offer at an earlier stage and also promote the need for the right skills. They help prospective applicants to understand the requirements in the current job market to plan their training and development more effectively.

Employers can use certifications to engage current employees in Continuous Professional Development (CDP) relevant to critical needs. While Higher Education degrees are evidence of a solid knowledge of a subject area (e.g., BSc Computer Science, MSc Data Science), certification programs tend to focus on very specific expertise and skills on industry-relevant tools and processes. Certification programs help to ensure the right competence level is clearly identified and communicated. Certification tests are used to assess skills and knowledge for this purpose.

Skills in the field of data science, machine learning, and analytics are in more demand than ever. KNIME Analytics Platform is one of the leading platforms. The *Data Science with KNIME Software* certificates from the *KNIME Certification Program* are testimony of proficiency in the open source platform for data driven innovation: they show your ability to develop, execute and deploy data analytics projects. Certificate-holders will boost their professional credibility; employers will more easily identify the right candidates to gain a competitive advantage.

About the collaborators

KNIME teamed up with the University of Reading to develop the *KNIME Certification Program*. The motivation was to draw on the experience and know-how from academia and apply it to build an effective certification program. With research expertise in Data Science, Machine Learning, Big Data Analytics, and High Performance Computing for Computational Science, the Department of Computer Science of the University of Reading, headed by Dr. Giuseppe Di Fatta, was an ideal partner.

The University of Reading awarded their first degree in Computer Science exactly 50 years ago in 1969. The Department of Computer Science has many years of experience in teaching Data Analytics, Data Mining, and Machine Learning, and moreover they have adopted KNIME Analytics Platform in teaching Data Analytics and Data Mining for over 10 years at undergraduate level and more recently at postgraduate level as well.

KNIME Certification Program

The certification program comprises three levels (L1 to L3) with additional special topic exams. Each level highlights a person's expertise with different aspects and practical skills on KNIME Software as well as most current data science concepts and know how.

Pass marks for the certification tests are at 70% (this is based on the grade boundary typically set in the UK undergraduate degree classification system for first-class honors degrees) and awarded certificates will be valid for 2 years. In this way employers can be reassured that an applicant with a KNIME certified credential is up to date with the latest developments in KNIME.

Examinations

- L1
  - Basic Proficiency in KNIME Analytics Platform
  - Examination: 30-minute multiple-choice exam (15 questions)
- L2
  - Advanced Proficiency in KNIME Analytics Platform
  - Examination: 30-minute multiple-choice exam (15 questions)
- L3
  - Proficiency in KNIME Software for Collaboration and Productionizing of Data Science
  - Examination: 90-minute multiple-choice exam (50 questions)

How to study?

To prepare yourself for these certification exams, we recommend the following methods of study:

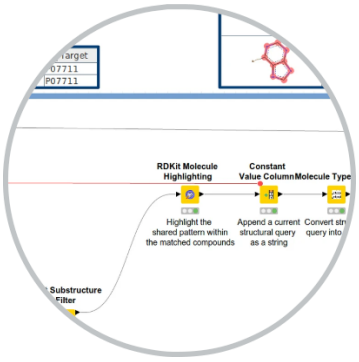
- KNIME Self-Paced Courses
- Instructor-led Courses for Beginners and Advanced Users
- University of Reading Data Science courses at UG or PGT level (e.g., CS3DM16 and CSMDM16).

How can I take the exam?

Find more information about the different exam levels and also links to the certification schedule in the KNIME Certification Program.

This article was first published in our KNME Blog. Find the original version here.

Alzbeta Tuerkova was nominated KNIME Contributor of the Month for June 2021. She was awarded for her research paper she published together with Barbara Zdrazil about an efficient and reproducible, fully KNIME based, drug-repurposing application to identify new drug candidates for rare diseases and the Covid-19. The workflow, tutorials, and information gained on COVID-19 data have been made freely available to the scientific community for follow-up studies and learning (see "*Automated Drug Repurposing Pipeline in KNIME*").



Alzbeta is currently Head of Computer-Aided Drug Design at Celeris Therapeutics. She holds a PhD in biology, with a thesis on hepatic organic anion transporting polypeptides belonging to the SLCO family, from the University of Vienna.

Visit Alzbeta's space on the KNIME Hub (Hub handle: `atuerkova`).

A Year of Pandemic: How KNIME Helps Find New Drug Candidates

Author: Alzbeta Tuerkova

Developing a novel drug is commonly recognized as a complex process, requiring a considerable amount of time and costs**.** The current timeline for a new medicine to get approved is ranging between 12 to 15 years. To accelerate the process of novel drug discovery, drug repurposing strategies can be incorporated into the pipeline. Drug repurposing (also known as drug repositioning) is a re-evaluation of an already existing drug to test its potential in the treatment of a novel disease. This approach becomes advantageous to reduce time, expenses, and risks attributed to the possible side effects.

From a computational standpoint, drug repurposing approaches can involve text mining, network analyses, machine (deep) learning models to predict drug-target-disease relationship, and structure- (protein-) based modeling methods. With the continuous growth of biomedical data sets, computational drug repurposing methods have attracted considerable attention. Specifically, mining of chemical structures for the sake of data augmentation (to be used as a docking library or to train a machine learning model) is no longer its sole purpose. Instead, large-scale data analysis helps disentangle patterns hidden in the chemical data, which in turn can assist in the early-stage drug development. Furthermore, computational drug repurposing approaches strongly benefit from combining different kinds of data entities, such as tissue expression data, genes, drug-target interactions, disease data collections, and phenotypes, to deliver an indication about a drugs' mode-of-action. Last but not least, chemical data originating from different sources were found to cover diverse areas of chemical space. Therefore, combining data from multiple databases also helps increase diversity of the physico-chemical properties of the final compound sets.

In this contribution, we present a workflow for performing ligand-based *in silico* drug repurposing. The applicability of the developed workflow is exemplified on the basis of COVID-19. Our strategy is based on the molecular similarity principle: Structurally similar molecules tend to possess similar biological activities. Therefore, we utilize publicly deposited structural- and bioactivity-ligand data associated with protein targets that are involved in the disease of our interest. Next, we identify enriched molecular (sub)structures in order to perform substructure searches of the datasets of available drugs for finding new - potentially active - drug candidates. The analytics platform KNIME serves here as a tool to automate the entire procedure, and provides the additional advantage of flexibility, re-usability, and transparency. In addition, our workflow is easily reproducible, and can be adapted according to individual project needs.

In a closer detail, the workflow includes targeted download of data through Application Programming Interfaces (APIs), molecular structures standardization, data integration, identification of structural analogs by hierarchical scaffold clustering and maximum common substructure generation, followed by the retrospective analysis of DrugBank and a data set of antiviral drugs provided by Chemical Abstracts Service (CAS).

Programmatic Data Access to Life-science Databases and Molecular Structures Standardization

When combining data from different repositories, it is beneficial to query databases in a programmatic manner. Databases utilized in this workflow - UniProt, Protein Data Bank (PDB), ChEMBL, Guide-To-Pharmacology (IUPHAR), PubChem, and DrugBank - permit targeted access of the stored data through an Application Programming Interface (API). Here, a triad of KNIME nodes is consecutively executed to (1) specify an API request (the “String Manipulation” node), (2) retrieve data from web services (the “GET request” node), and (3) extract relevant properties from received files (the “XPath” node). In a first instance, protein target identifiers of the Open Targets platform are mapped to their corresponding UniProt IDs. The retrieved UniProt IDs serve as a starting point to retrieve protein–ligand structural data from PDB, as well as ligand bioactivity data from ChEMBL, IUPHAR, and PubChem.

A prerequisite for merging ligand data from diverse sources is to standardize molecular structures. To ensure unified chemical data representation, a cascade of nodes (involving RDKiT and CDK node extensions) was executed to (1) remove compound stereochemistry (the “String Replacer” node), (2) strip salts by forwarding a predefined set of different salts/salts mixtures (the “RDKit Salt Stripper” node), (3) list all stripped salt components in the output table (the “Connectivity” node followed by the “Split Collection Column” node), (4) neutralize charges and check for possible atomic clashes (the “RDKit Structure Normalizer” node), (5) filter data by checking specific elements (the “Element Filter” node), (6) generate InChI, InChiKey, and Canonical smiles formats using the corresponding RDkit nodes.

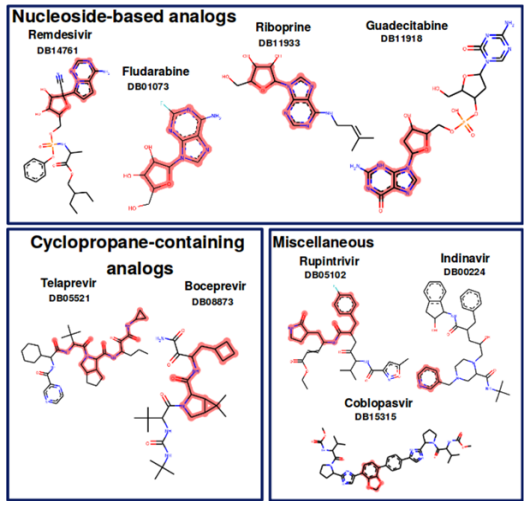
Substructure searches to identify potential drug candidates suited for drug repurposing

Molecular structures are reduced to their Bemis-Murcko scaffolds. Generated scaffolds and associated UniProt IDs (as a list) are used as an input (the “Table to Variable” node). Molecular distances are computed for the retained scaffolds (‘MoSS MCSS Molecule Similarity’ node) and hierarchical clusters are generated accordingly (the “Hierarchical Clustering [DistMatrix]” node). Scaffolds are grouped into a specific cluster by applying the “Hierarchical Cluster Assigner” node. The loop is then repeated for the remaining targets from the input table. The output table contains UniProt IDs,

associated scaffolds, and cluster IDs. Next, looping over distinct clusters of associated scaffolds for a respective target is done in order to create a maximum common substructure (the "RDKit MCS" node) from all associated scaffolds belonging to a respective cluster. The generated substructures (in SMARTS) are then used as queries to find hits in DrugBank (input data set includes molecules in SDF format, DrugBank IDs, and associated content). The structures from DrugBank are standardized, and then subjected to the "RDKit Substructure Filter" node to perform substructure searches. The respective SMARTS query is forwarded to the input as a variable. Detected substructures are highlighted by the "RDKit Molecule Highlighting" node. The output table contains identified hits (molecule names, associated targets, SMARTS keys, chemical structures), and highlighted substructures in SVG format.

COVID-19 as a use case

Substructure searches helped identify 7836 compounds from DrugBank and 36,521 compounds from the CAS data set, Out of those hits, 135 compounds were retrieved from both DrugBank and the CAS data set. Identified MCSs can be combined into five separate clusters: (1) Hits identified on basis of the open-chain structural keys (59 hits), (2) Nucleoside/nucleotide analogs (53 hits, e.g., remdesivir, fludarabine, riboprine), (3) miscellaneous, which contain ubiquitous substructures (22 hits, e.g., rupintrivir, indinavir, darunavir), (4) cyclopropane-containing hits (3 hits, e.g., telaprevir), and (5) adamantane derivatives (3 hits), see the figure below. It is noteworthy to mention that some of the identified drugs are now undergoing clinical trials to test their potential to combat COVID-19. In addition, the workflow helped identify novel interesting candidate molecules, which can inform future therapeutics development to treat COVID-19.



Examples of identified drugs with structural queries highlighted. Taken from Tuerkova A, Zdrazil B. J Cheminf. 2020 Dec;12(1):1-20.

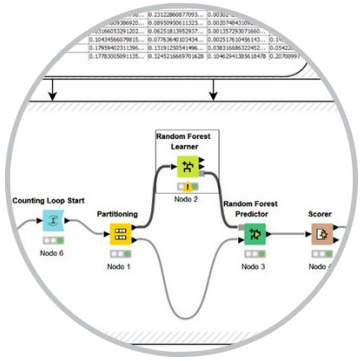
Using the workflow in the virtual classroom

The workflow was used in the summer semester 2020 (April 20–24) within the course "Experimental Methods in Drug Discovery and Preclinical Drug Development", a part of the English-language Master's Degree Program Drug Discovery and Development at the University of Vienna (https://drug-dd.univie.ac.at/). Due to the protective measures caused by the COVID-19 pandemic, this course was structured as a virtual classroom. The students attended online sessions, in which the various steps of the workflow were explained and demonstrated. Tutorials and the different parts of the workflow have been handed out daily. On the last day of the 5-days course, each student selected one of the hits retrieved by the substructure search and dedicated some time to literature searches. Finally, every student submitted a report summarizing what is known about the selected compound and its potential usefulness for COVID-19 treatment (according to what was known in April 2020).

This booklet contribution was based on the published article: Tuerkova A, Zdrazil B. A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19. Journal of Cheminformatics. 2020 Dec; 12(1):1-20. https://doi.org/10.1186/s13321-020-00474-z

Malik Yousef was nominated KNIME Contributor of the Month for January 2022. He was awarded for his research on DNA, mRNA, and miRNA analysis, gene ontology, and molecular biology. Using KNIME, he has developed machine learning algorithms to predict mRNA and their targets, analyze gene expression, and integrate mRNA and miRNA expression profiles via machine learning. The image on the right shows a workflow snippet from his research paper on miRcorrNet.

Malik is a Data Scientist with the focus on bioinformatics with applications to various biomedical/biological problems. He has significant expertise in machine learning and its applications and has notched up a long list of publications in top, peer reviewed scientific journals, books, and US patents. As a professor, he has made cutting-edge contributions in life sciences, and he uses KNIME actively in his research. He is currently Head of the Galilee Digital Health Research Center (GDH) at Zefat Academic College in Israel.



Visit Malik's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `malik`).

Gene Ontology, Biomarkers and Disease Prediction: From the Research Lab to the Data Science Class

My Data Guest — An Interview with Malik Yousef

Authors: Stefan Helfrich & Aline Bessa

It was our pleasure to recently interview live on LinkedIn Malik Yousef as part of the My Data Guest interview series.

Malik Yousef is the Head of Galilee Digital Health Research Center at Zefat Academic College in Israel. He has impacted genome and gene expression (DNA, mRNA, and miRNA) analysis, gene ontology, and (computational) molecular biology, in general, with his research, as confirmed by a long list of publications in top, peer-reviewed scientific journals, books, and US patents. He has developed machine learning algorithms to predict mRNA and their targets, analyze gene expression, and integrate mRNA and miRNA expression profiles via machine learning.

Stefan: *Please explain to us in your more precise words, what is your scientific area of expertise.*

Malik: I have a background in mathematics and computer science and did my PhD in text mining, where I developed some methods that learn from positive examples only. The first time I was involved with biology was during my postdoc at Wistar Institute in the USA. At that time, bioinformatics was just in the early stages of establishment. I really enjoyed working with biological data and contributing my experience to this field. For simplicity, I now consider myself a Data Scientist, but the data that I work with come from biology. Currently, I'm trying to find biomarkers among a total of 20000 genes. That is, I'm looking for a small but significant group of genes that can serve as biomarkers for disease prediction.

Stefan: *What kind of diseases are you working on?*

Malik: During my postdoc, I worked with lung cancer data, but now I mainly work with public datasets from all cancers that exist, e.g., kidney cancer, breast cancer, etc. The input to our algorithms, in general, is gene expression data and the biomarkers can

come from any disease. More traditional bioinformatics approaches borrow algorithms from computer science and apply them to biology data. However, my approach is to integrate knowledge of biology into the machine learning algorithms, and then perform feature selection to identify the most important genes.

Stefan: *What has been your most impactful scientific contribution in this field?*

Malik: I would divide my contributions into three parts. First, learning from positive examples. Most machine learning algorithms rely on two classes, for example, black vs. red. Now, assume that the data has only one kind of class, for example, only black. In this case, the question is how can we learn and build models from just one class, i.e., mainly positive examples? This is what I developed during my PhD, a one-class text classification.

Second, I did a lot of work in the area of predicting miRNA using the sequence itself. Simply said, a sequence is just a combination of letters, a string so to say. Also, miRNA is a new discovery that has an important role in our body: it regulates and shuts down genes. And it turns out that a lot of cancer diseases are controlled by this small miRNA.

And third, I want to highlight my integrative approach. This means, integrating biology knowledge into machine learning via GSM (grouping, scoring, modeling). That is, we are looking to integrate data of different types and from various sources. For instance, combining gene expression data with ontologies and miRNA targets. We have developed a lot of algorithms in this area and are quite successful with it.

Stefan: *Have you used KNIME for your research?*

Malik: I love KNIME. I mainly use KNIME in combination with Python and R. All my teaching and all my collaborations involve KNIME to some extent.

Let me tell you a brief story: A few years ago, I went to a scientific conference in Germany just because of KNIME. I discovered KNIME when talking to a colleague about an issue I had in my Java code. Nothing major but to fix it, it would have probably taken me a week. My colleague suggested I could just use KNIME to solve the problem. I was impressed and that was when I started learning KNIME by myself. After this encounter I started my work on maTE which was one of my first complex algorithms implemented in KNIME. But even at this level of complexity, the workflow structure helps me communicate the work to non-programmers.

Stefan: *Where do you think KNIME helps the most in your project implementations? Connecting to the data sources or training machine learning models, or something that I can't think of?*

Malik: There are different impacts of KNIME. For one thing, with KNIME I don't have to spend a lot of time on debugging only to find a small bug. It gives me more time to focus on my research.

Second, the ability to combine Python or R with KNIME is very powerful. For a lot of things I'm doing, I use R and Python code. With KNIME, it's possible to use existing approaches, tools, or algorithms and to combine them.

Another great impact of KNIME is that the figures are directly usable in my publications.

And lastly, I save time using KNIME just because I'm faster doing it in KNIME. I used Matlab before and, for example, to write the code in Matlab it took me more than one month but to do the same thing in KNIME took me only one and a half days.

Aline: *Has KNIME facilitated the collaboration among the researchers in your group? How does it compare with coding-based research?*

Malik: Most biologists are no computational guys. Showing them actual Python, R, or Java code would be difficult. However, using KNIME and showing them a workflow in KNIME Analytics Platform made it easier to communicate with them.

Also, in terms of communication with my students, it's easier to split up projects and work in this development environment.

Stefan: *How does it work with your students? How fast can they get up to speed and how fast can they get started with KNIME so that they can really be productive?*

Malik: There are two kinds of students, the ones that I'm teaching and the ones that I'm supervising. The first time I taught KNIME, I actually got asked to teach a data science lab and was allowed to introduce KNIME upon request. I immediately got the students' feedback that they love KNIME and find it quite easy and convenient. Usually, I trigger them by giving them a task and ask them to do it in Java, which usually requires them to write a lot of boilerplate code. Then, I show them how much easier it is in KNIME. In terms of teaching, I integrate KNIME in most of my courses, very often alongside Python.

Some of the students even take KNIME to their companies, asking their bosses if they can work with KNIME.

The students I'm working with usually don't know anything about KNIME. But it's very easy for them to learn. They study by themselves only using the free learning materials on the KNIME website.

Stefan: *In mid-June, KNIME released its latest software version (v. 4.6.0). One of the most interesting new features is the possibility of developing pure-Python nodes that can be shared with others like any other node. How do you think researchers can benefit from this new feature?*

Malik: I saw this release and I think it's amazing. I think it's a huge contribution to the people using KNIME that they can write nodes in Python now. This also gives the KNIME community the opportunity to contribute more nodes. Although, the possibility

to build components and share them with others was also transformative work. However, we shouldn't forget about R here. Some of my students know R but don't know Python, others know Python but don't know R. I think it would be good if we could do the same with R.

Stefan: *Let's talk a moment about the students. What kinds of students come to your classes? What background do they have? Are they college students or grad students? Do you see a wave of older students in search of a second education?*

Malik: My students mostly belong to the department of Computer Science or Information Systems. So, in general, they have a programming background but there is variety. There are a couple of students that are not so strong in programming but still need to do some coding. Especially for those students KNIME is a very good solution. It eases their lives. I think teaching KNIME could also be an introduction to programming rather than starting with Python or Java. KNIME is a powerful tool to teach algorithms and programming without actual programming.

Stefan: *What topics do you cover in your courses? What will the students learn?*

Malik: I'm teaching a couple of courses, usually more advanced courses. There, the students learn everything: parameters, flow variables, global variables, collecting data in loops, etc. In some of the courses, according to the college, I'm supposed to teach Python. However, as I'm a big fan of KNIME I always try to integrate it in my teaching. But I'll also be teaching a new course in September called "Coding without Coding" where I'll be using KNIME to teach students how to code without actually writing code.

Stefan: *How would you describe your teaching approach? How did the COVID-19 pandemic change your teaching style?*

Malik: I think we did get some benefits from the global pandemic. Sometimes, humans need to be forced to do things. The pandemic forced us to teach via Zoom or other platforms. Now I love teaching via Zoom. I think it makes teaching easier, for the teachers and professors but also for the students. In addition, it is now much easier to record lectures and enable a much bigger audience to "attend" my courses, for instance, part-time students.

Stefan: *Now the classic question. Where do you see research in bioinformatics going in the next few years? What will the next innovation be?*

Malik: Back in the days, I was among the first people working in bioinformatics. The field started to establish itself by attracting people from Computer Science, Math, Statistics, etc. They then started to produce a program in bioinformatics. Now we have more educated people in the bioinformatics program. The learning is more organized, and we have more data available. When I started, I had only one kind of data — gene expression data. Now we can see the need for algorithms to combine data from two

or even more-omics technologies: We are already in the area of multi-omics and we'll need to do more work for a better integration of N datasets.

Stefan: *Any interesting upcoming projects or conferences?*

Malik: Most of the conferences this year were remotely. I was at the ISMB conference in July and participated in the CAMDA Contest Challenges this year. In October, I will be attending the HIBIT 2022 conference in Turkey. I attended that conference already a couple of years ago which is the reason why I have a collaboration with a colleague from Turkey. But mostly, I'll be at the University of North Carolina as a visiting professor doing my scholarship.

Stefan: *How can people from the audience get in touch with you?*

Malik: People can reach me either via LinkedIn, my KNIME Hub, or GitHub.

This article was first published in our Low Code for Advanced Data Science Journal on Medium. Find the original version here.

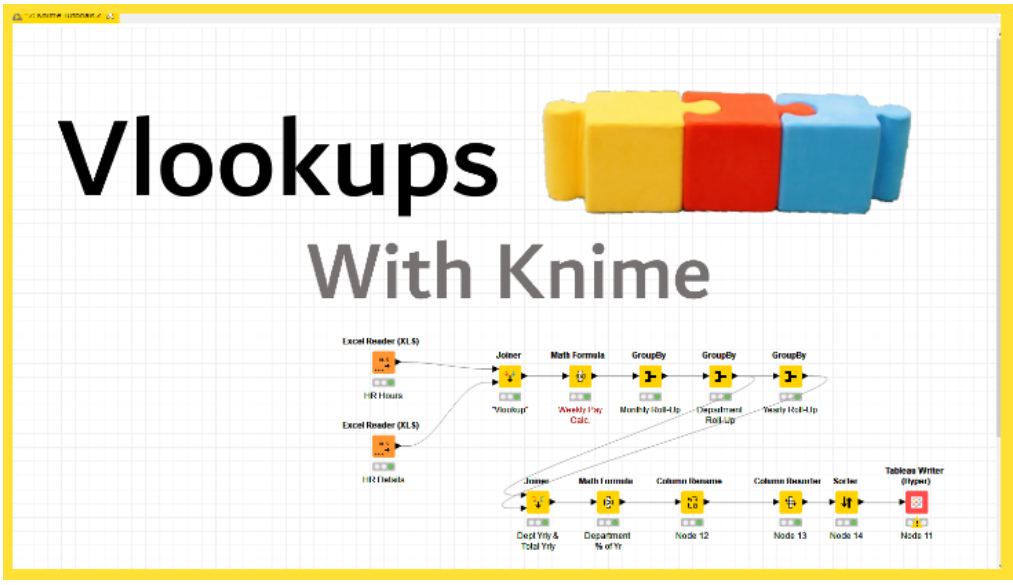
Watch the original interview with Malik Yousef on YouTube: "My Data Guest – Ep 11 with Malik Yousef".

Nick Rivera, aka NickyDee on YouTube, was nominated KNIME Contributor of the Month for February 2022. He was awarded for his many video tutorials representing an extensive, organized, easy to digest, and useful resource for both newbies and expert KNIME users to progress in their knowledge of KNIME. This includes explorations of lookup IDs, pivoting and unpivoting, aggregations, conditional math, date and time operations, excel function translations, and many more data transformation and ETL operations. On his YouTube channel he covers most aspects of data manipulation operations with KNIME.

Nick is a data professional experienced in Business Modeling, Reporting, and Analysis. He is currently a Business Analyst at EMR Group. His primary focus is on building operational models that help drive the Ferrous and Non-Ferrous businesses at EMR. While Nick's expertise includes a bit of everything, his favorite aspect of data is the ability it has to influence informed decision making.

How to do an Excel VLOOKUP in KNIME

Author: Nick Rivera



VLOOKUP in KNIME

The *VLOOKUP* is one of the best and most fundamental formulas available in Excel, and for good reason too. The ability to join (combine) related tables is crucial when trying to bring together data from multiple sources, in order to synchronize and deliver a succinct analysis. Worry no more, today I will show you how to perform a *VLOOKUP* from Excel in KNIME.

Enter the Join

If you've ever written a SQL query before, then you already know what a VLOOKUP is. If you haven't, then today you will learn!

The VLOOKUP in Excel is made up of 4 arguments:

1. The Lookup Value
2. The Lookup Range
3. The index of the column you want bring over
4. Exact or approximate match

To perform a VLOOKUP in KNIME you will have to use a Joiner node

The *Joiner* node only needs 2 "arguments", the lookup column(s) and the columns you want to bring over.

For our example, let's use the two tables pictured below:

Table 1		
Department	**Employee Count**	**Department Bonus Rate**
Sales	35	10%
Finance	6	2%
HR	5	2%
Accounting	7	1%
Operations	150	5%

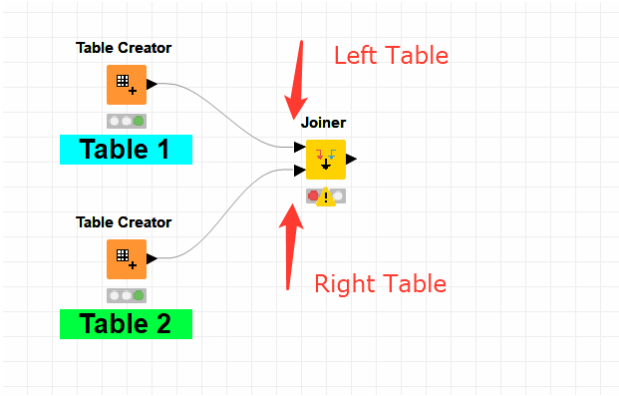
VLOOKUP in KNIME

Table 2		
Department	**Sales Units**	**Bonus Rate**
Operations	80	
Accounting	10	
HR	20	
Sales	130	
Finance	15	

We'll VLOOKUP (join) the bonus rate from the above table to this new table.

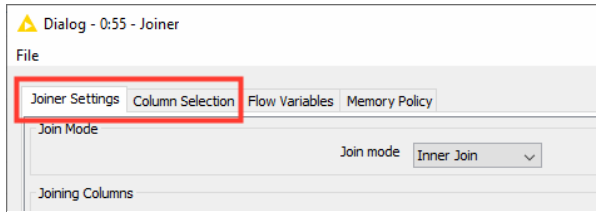
Our goal will be to VLOOKUP the *Department Bonus Rate* detail from Table 1 to Table 2.

To perform a VLOOKUP, we'll need to select the *Joiner* node**. In your node repository, search for "*Joiner*" and then drag and drop the node into your workspace**. Connect both tables that you want to *VLOOKUP* (aka join) to the *Joiner* node. The table that you connect to the top input will be considered the "*left*" side table, while the table you connect to the bottom input will be considered the "*right*" side table. This is important to recognize because the *Joiner* node will give you the flexibility to run a left join or a right join as well as an inner join or a full outer join. I'll add more color on the join types a little later, but just keep that in the back of your head for now.



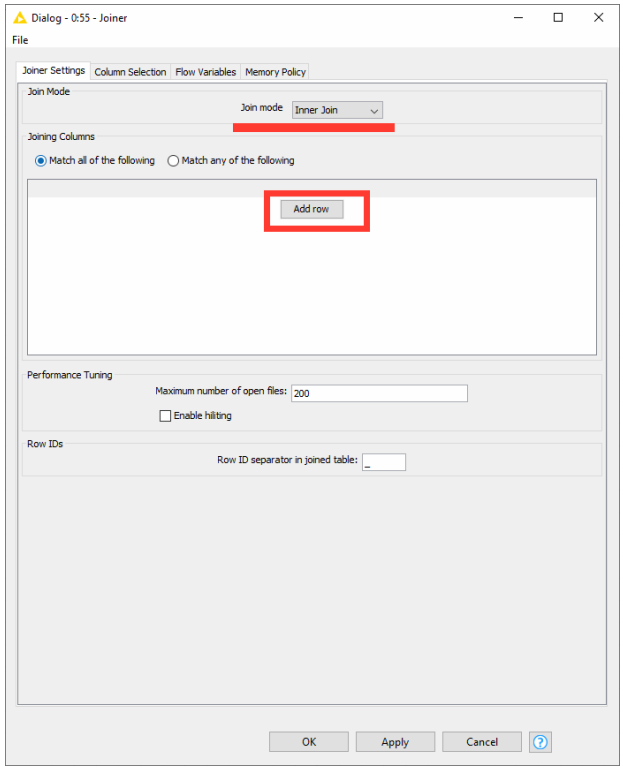
A workflow that demonstrates the use of the Joiner node.

Now that you've got the two tables connected to the *Joiner* node, double click into the *Joiner* node to start the configuration. In the configuration you will see the following tabs (see figure below). The most important tabs here are the "*Joiner Settings*" and "*Column Selection*" tabs. The Joiner Settings tab is where we will set up the "*lookup value*" argument while the column selection tab is where we'll set up the "*lookup column index*" argument.



The Joiner Settings and Column Selection tabs in the configuration window of the Joiner node.

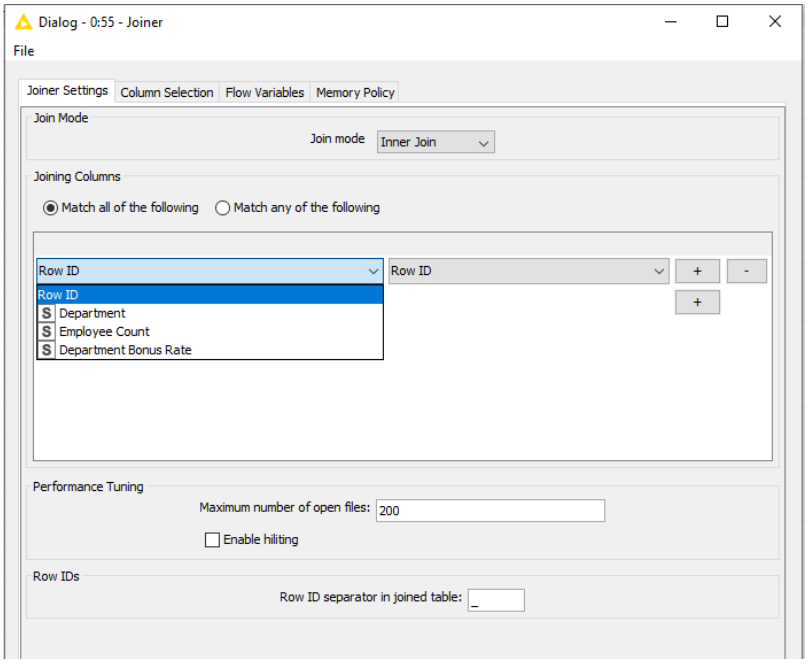
In the "*Joiner Settings*" tab we'll need to outline the column on which we want to join the two tables. You can think of this step as the providing the Lookup value argument and the beginning of the lookup range. Put differently, the lookup value and the first column of the lookup range are the values that we need to provide as the joining columns.



The "Join mode" and "Add row" functionality of the KNIME Joiner node.

In the screen shot above, in the red box there is a button that reads "*Add row*". We'll click this to give us a row where we can outline the joining columns in each of the two tables. The below image shows the configuration menu after we click Add Row. We'll click into the dropdowns that are labeled "*Row ID*" and then we'll set these both to the joining columns. The dropdown on the Left side corresponds to the left table aka the

Top Input of the *Joiner* node while the dropdown on the right side corresponds to the Right table aka the Bottom Input of the *Joiner* node.



The configuration window of the Joiner node. In the Joiner Settings tab, you can define the join condition.

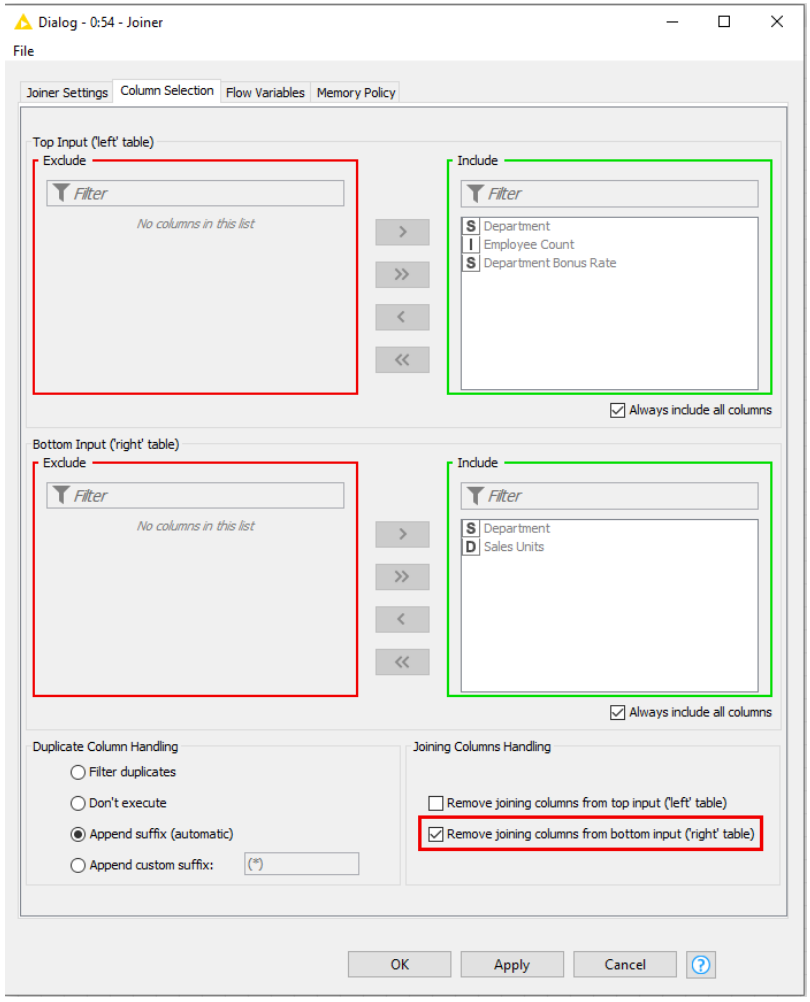
In our example, the two tables (both left and right side) share the Department column, so we will select "Department" under both. Now that we've got the joining columns delineated, we need to select the type of join or VLOOKUP we want to run. The options under Join Mode are as follows:

1. Inner Join: The final output here is a new table where both the left & right side tables share similar matching rows
2. Left Join: Here you're joining details from the right side table to the LEFT side table - think of the LEFT side table as your primary table
3. Right Join: Here you're joining details from the left side table to the RIGHT side table - think of the RIGHT side table as your primary table
4. Full Outer Join: Here you're joining both tables against each other, regardless of their being matching lookup columns or not; therefore, some columns/rows will show null/missing values

What's glaringly obvious here is the flexibility available within the joiner node that isn't really available in an excel VLOOKUP. We won't get off track fully outlining that detail here though, since the goal of this post is to be a quick DIY of sorts. I will write up a separate post showing the pros of the join over a VLOOKUP, I'll link that post here once it's ready, so be on the lookout.

Our original goal was to join/lookup the Departmental Bonus rate from table 1 over to table 2. Given that goal, we're going to run with a right join. The thought here is that the table on the table on the right side is our primary table of analysis, so we're joining details [from the Left Table] to the Right Table. Let me know if you don't follow that!

Now that we've got the joining columns sorted, we can move on to selecting the columns we want to bring over aka argument 3 of a VLOOKUP. In the screenshot below is the *Column Selection* tab. The configuration here is pretty straightforward, we'll select which columns from which table we want to include & exclude from our output table.



KNIME's Joiner node allows flexibility in column selection from both left and right tables.

The tables we're joining are small and limited in column count, so we don't really need to limit which columns we want bring over from the left or the right table. There may be cases though, where you're joining detail from two tables which each hold multiple

columns of different detail. You might not need every single detail from both tables, so in those cases you have the ability to limit which columns come over from each table. To include and/or exclude columns that we want in the joiner output, simply uncheck the box under each Include screen that reads "Always include all columns". Unchecking this will allow us to free move columns from the Include side to the Exclude side and vice-versa. Again, this is another case of added flexibility that a join offers which an excel VLOOKUP doesn't offer!

The "*Always include all columns*" checked box is a setting that's checked in by default. Another default setting you've got to be aware of here is that the joining columns from the lower table are dropped in favor of the columns from the left table. This isn't really a huge issue in most cases, but it's something to keep in mind when you get more involved with varying tables. You can simply override this default setting by unchecking the box pictured below. If you uncheck this box and also leave the one above it unchecked, then you will get duplicates of the joining columns. To the left of the red box in the screenshot below are a few options on handling the naming of the duplicate columns. I like to use the "*Append custom suffix*" option in cases where I'm looking to calculate Product Mix or Market Share values.

After we've selected the columns we want in our final output, we can hit apply, ok, and execute the node. The results look like this:

Joined table - 0:54 - Joiner

File Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	S Depart...	I Employee Count	S Department Bonus Rate	D Sales Units
Row0_Row0	Sales	35	10%	35
Row1_Row1	Finance	6	2%	6
Row2_Row2	HR	5	2%	5
Row3_Row3	Accounting	7	1%	7
Row4_Row4	Operations	150	5%	150

This table has the output of performing an excel VLOOKUP in KNIME!

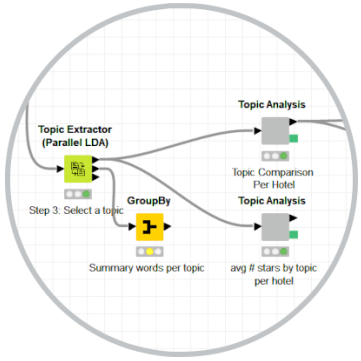
Table 2 now has the Departmental Bonus Rate detail from Table 1. Since we didn't exclude any columns, we also brought over the Employee Count column from the left table over to the right table – you've got to love that flexibility!

This article was originally published on Nick's blog. You can find the original version here.

In case of questions or if you'd like to get in contact with Nick, he's reachable through Twitter.

Francisco Villarroel Ordenes was nominated KNIME Contributor of the Month for April 2022. He was awarded for developing together with a team at KNIME a live repository for Machine Learning in Marketing Analytics on the KNIME Community Hub with reusable solutions for customer churn, sentiment analysis, automated image analysis, SEO & CX. An extensive analysis of this project is provided in his research paper. The image on the right shows a part of the Topic Modelling workflow provided in the repository.

Francisco is currently an assistant professor of marketing and the director of the MSc. in Marketing program at LUISS Guido Carli University in Rome, Italy. He is a strong proponent of KNIME in his lectures, and he uses it extensively in his classes and research work regarding marketing analytics. His areas of interest are Analytics, Digital Marketing, and Service Research, and he is passionate about consumer-brand communications and the use of NLP to understand online conversations.



Visit Francisco's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `fvillarroel`).

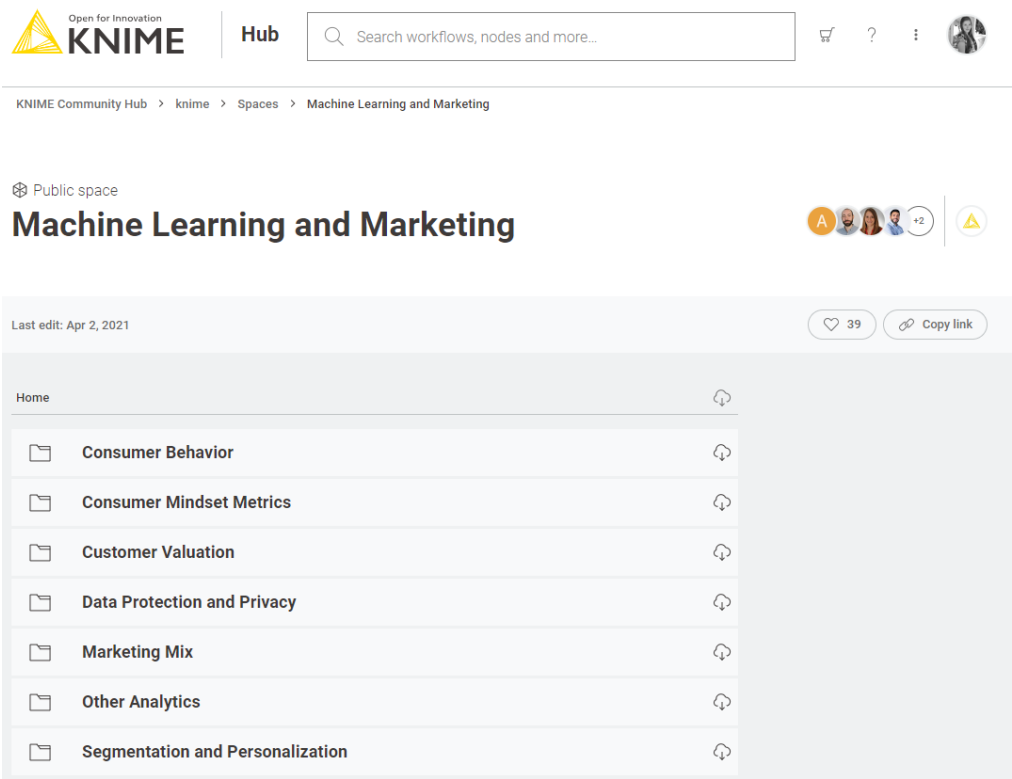
Machine Learning in Marketing Analytics

Marketing Analytics Solutions on the KNIME Community Hub

Authors: Francisco Villarroel Ordenes & Rosaria Silipo

Many businesses are currently expanding their adoption of data science techniques to include machine learning. Marketing analytics is one of them. Anything can be reduced to numbers, including customer behavior and color perception, and therefore anything can be analyzed, modeled, and predicted.

Marketing analytics already involves a wide range of data collection and transformation techniques. Social media and web driven marketing have given a big push in the digitalization of the space; counting the number of visits, the number of likes, the minutes of viewing, the number of returning customers, and so on is common practice. However, we can move one level up and apply machine learning and statistics algorithms to the available data to get a better picture of not just the current but also the future situation.



The public workflow repository for marketing analytics solutions on the KNIME Community Hub.

Marketers can capitalize on machine learning techniques to analyze large datasets to identify patterns or perform predictive analytics. Examples include analyzing social media posts to see what customers are saying, analyzing images to extract insight into pictorials and videos, or predicting customer churn – to name just three.

In the Machine Learning and Marketing space on the KNIME Community Hub, you will find a number of case studies applying machine learning algorithms to classic marketing problems.

In this post, we will describe these case studies, one by one, showing the particularity of each one of them and the insights they bring. So far, we have solutions for:

- Prediction of Customer Churn
- Measuring Sentiment Analysis in Social Media
- Evaluation of Customer Experience through Topic Models
- Content Marketing and Image Mining
- Keyword Research for Search Engine Optimization

We will continue to maintain this repository by updating the existing workflows and adding new ones every time a solution from a new project becomes available.

Note. This solution repository has been designed, implemented, and maintained by a mixed team of KNIME users and marketing experts from the KNIME Evangelism Team in Constance (Germany), headed by Rosaria Silipo, and Francisco Villaroel Ordenes, Professor of Marketing at LUISS Guido Carli university in Rome (Italy).

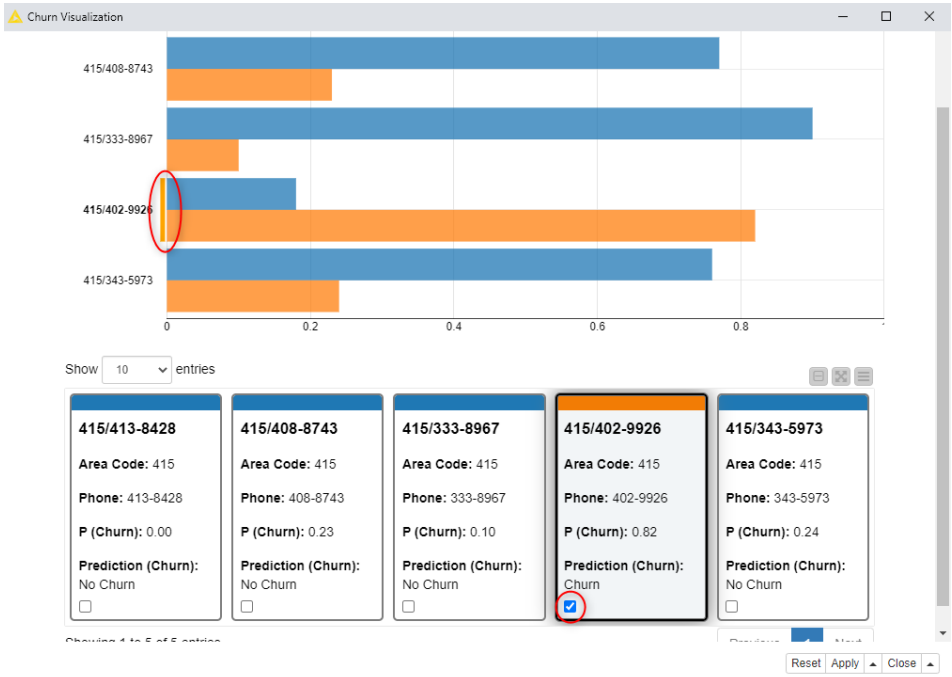
Prediction of Customer Churn

Using existing customer data (e.g., transactional, psychographic, attitudinal), predictive churn models aim to classify customers who have churned or remained, as well as estimate the probability of new customers to churn, all in an automated process. If the churn probability is very high and the customer is valuable, the firm might want to undertake actions to prevent this churn.

The “Consumer Behavior” > “Churn Prediction” folder in the Machine Learning and Marketing space on the KNIME Community Hub includes:

- A workflow training a ML classifier (a random forest in this case) to distinguish customers who have churned and customers who have stayed in the training set.

- A deployment workflow applying the previously trained model to new customers, estimating their current probability to churn, and displaying the result on a simple dashboard (see the figure below).



The dashboard reporting the churn risk in orange for all new customers.

Sentiment Analysis

Sentiment is another popular metric used in marketing to evaluate the reactions of users and customers to a given initiative, product, event, etc. Following the popularity of this topic, we have dedicated a few solutions to the implementation of a sentiment evaluator for text documents. Such solutions are contained in the “Consumer Mindset Metrics” > “Sentiment Analysis” folder. All solutions focus on three sentiment classes: positive, negative, and neutral.

There are two main approaches to the problem of sentiment:

- **Lexicon based.** Here, a list of positive and a list of negative words (dictionaries), related to the corpus topics, are compiled and grammar rules are applied to estimate the polarity of a given text. Learn how to build a sentiment predictor using lexicon-based sentiment analysis.
- **Machine Learning based.** The solutions here rely on no rules, but on machine learning models. Supervised models are trained to distinguish between negative, positive, and neutral texts and then applied to new texts to estimate their polarity.

Machine-learning-based approaches have become more and more popular, mainly because of their capability to bypass all grammar rules that would need hard-coding. Among the machine learning based solutions, a few options are possible:

- **Traditional machine learning algorithms.** In this case, texts are transformed into numerical vectors, where each unit represents the presence/absence or the frequency of a given word from the corpus dictionary. After that, traditional machine learning algorithms, such as random forest, Support Vector Machine, or Logistic Regression can be applied to classify the text polarity. Notice that in the vectorization process the order of the word in the text is not preserved. Read more in this tutorial for machine learning for sentiment analysis.
- **Deep Learning based.** Deep learning based solutions are becoming more and more popular for sentiment analysis, since some deep learning architectures can exploit the word context (i.e. the sequence history) for better sentiment estimation. In this case, texts are one-hot encoded into vectors, the sequence of such vectors is presented to the neural network, and the network is trained to recognize the text polarity. Often, the architecture of the neural network includes a layer of Long Short Term Memory units (LSTM), since LSTM performs the task by taking into account the order of appearance of the input vectors (the words), that is by taking into account the word context. Explore a tutorial to set up a deep learning approach to sentiment analysis.
- **Language models.** They are also referred to as deep contextualized language models because they reflect the context-dependent meaning of words. It has been argued that these methods are more efficient than recurrent neural networks because they allow parallelized encoding (rather than sequential) of word and sub-word tokens contingent on their context. Recent language model algorithms are ULMFiT, BERT, RoBERTa, XLNet. In the Machine Learning repository, we provide a straightforward implementation of BERT. See how to use BERT with KNIME in this sentiment analysis tutorial.

Find an example of all these solution groups in the “Consumer Mindset Metrics” > “Sentiment Analysis” folder in the Machine Learning and Marketing space.

Visualization of tweets with estimated sentiment (red = negative, green = positive, light orange = neutral).

Topic Detection and Customer Experience

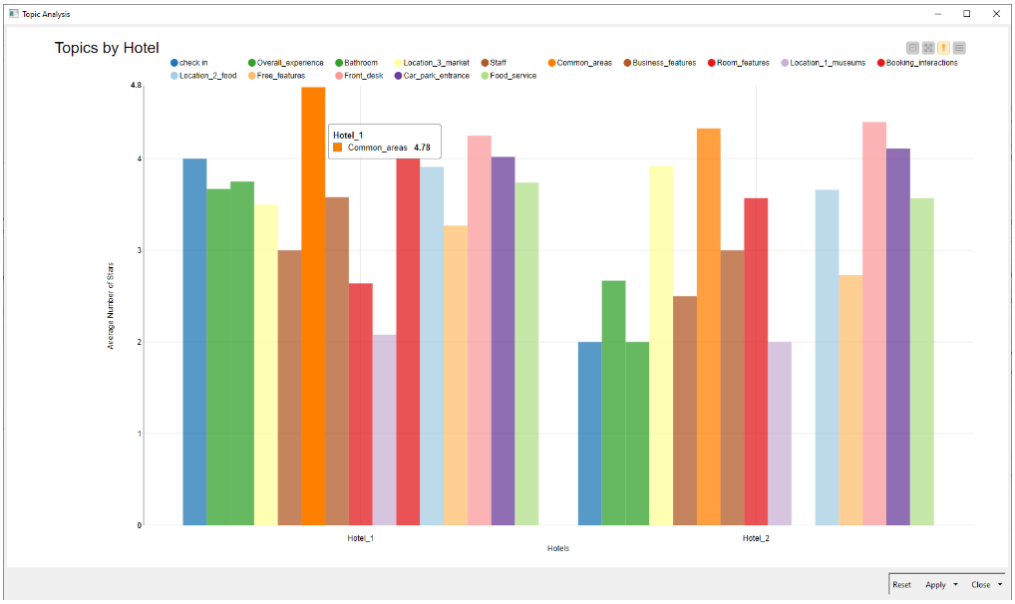
Customer experience management and the customer journey is one of the most popular marketing topics in the marketing industry. A lot of information about customer experience comes from the reviews and feedback and/or from the star-ranking systems on websites and social media.

The popularity of topic models has resulted in a continuous development of algorithms such as Latent Dirichlet Allocation (LDA), Correlated Topic Models (CTM), and Structural Topic Models (STM), among others, all of them already implemented in business research. LDA is available in the KNIME Text Processing extension as a KNIME native node. The LDA node detects m topics in the whole corpus and describes each one of them using n keywords, m and n being some of the parameters required to run the algorithm.

You'll find an example workflow, showing the usefulness of discovering topics in reviews, in the folder "Consumer Mindset Metrics" > "CX and Topic Models" in the Machine Learning and Marketing space.

The workflow extracts topics from reviews using the LDA algorithm. After that it estimates the importance of each topic via the coefficients of a linear regression – implemented with a KNIME native node – and via the coefficients of a polynomial regression – implemented in an R script within the KNIME workflow. It then displays the average number of stars for all topics extracted from the reviews for two different hotels (see the figure below).

In the bar chart for instance, we can see that for hotel 2 the topic “booking Interactions” is never mentioned. We can also notice that while hotel 1 gets great reviews for the “common areas”, hotel 2 excels for the “Front desk”.



Average number of stars by reviews around one of the 15 detected topics.

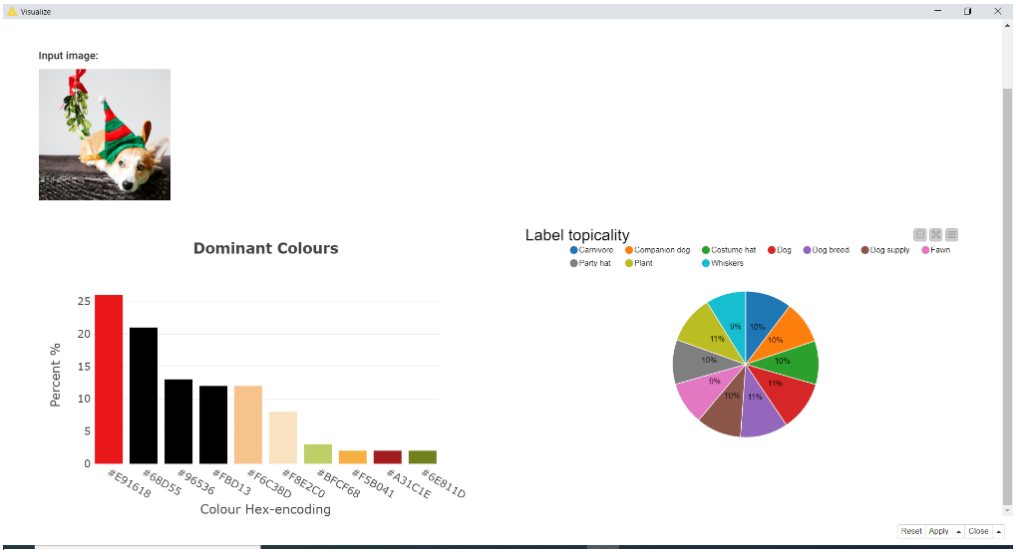
Content Marketing and Image Mining

The last ten years have shown an exponential growth of visual data including images and videos. This growth has resulted in an increasing development of technologies to classify and extract relevant insight from images. This phenomenon has had an impact on marketing as well. As both consumers and firms are relying more on pictorials and videos to communicate, researchers need new processes and methods to analyze this type of data.

The greater interest in the analysis of visuals and its implications for firm performance, motivated us to develop a workflow that can help with the analysis of visual content. The workflow takes advantage of Google Cloud Vision services (accessed via POST Request), to detect labels (e.g., humans) and extract nuanced image properties such as color concentration.

A second workflow uses deep learning Convolutional Neural Networks to classify images of cats vs. dogs. Changing the image dataset and correspondingly adjusting the network, allows you to implement any other image classification task.

Find both workflows in the folder “Other Analytics” > “Image Analysis” of the Machine Learning and Marketing space. The figure below shows the result obtained from the analysis of an image via Google Cloud Vision services.



Analysis of the image in the top left corner through Google Vision services.

Keyword Research for SEO

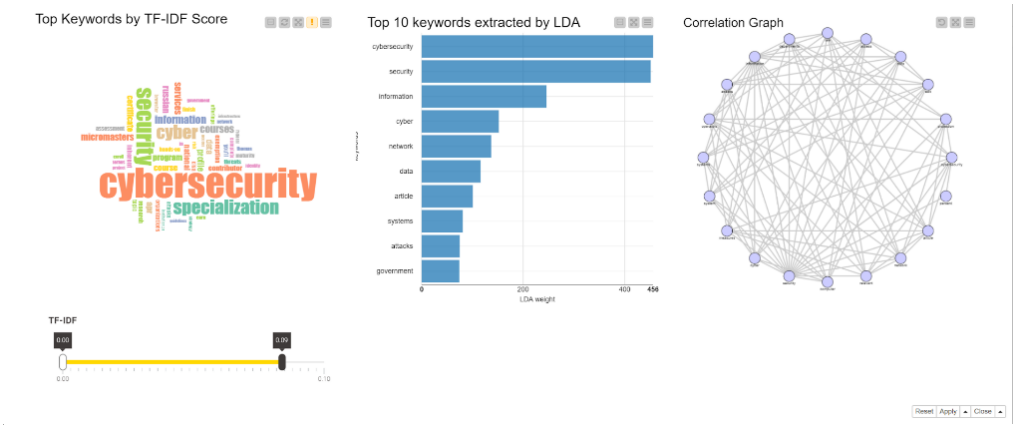
It is known that search engines rank web pages according to the presence of specific keywords or groups of keywords conceptually and/or semantically related. In addition, keywords should be taken from the specialized lingo by experts as well as from the conversational language by neophytes. Popular sources for such keywords are SERP (Search Engine Result Pages) as well as social media.

In the "Consumer Mindset Metrics" > "SEO" folder of the Machine Learning and Marketing space, you'll find a workflow for semantic keyword research, implemented following the article "Semantic Keyword Research with KNIME and Social Media Data Mining – #BrightonSEO 2015" written by Shapiro in 2015.

The upper branch of the workflow connects to Twitter and extracts the latest tweets around a selected hashtag. The lower branch connects to Google Analytics API and extracts SERPs around a given search term. After that, URLs are isolated, web pages scraped via GET Requests to Boilerpipe API, and keywords are extracted together with their frequencies.

Keywords as: single terms with highest TF-IDF score; co-occurring terms with highest co-occurring frequency; keywords with highest score from topics detected via the Latent Dirichlet Allocation (LDA) algorithm.

As an example, we searched for tweets and Google SERPs around *"cybersecurity"*. Resulting co-occurring keywords are shown in the word cloud in the figure below. If you are working in the field of cybersecurity, then including these words in your web page should increase your page ranking.



Analysis of the image on the right through Google Vision services.

Explore Machine Learning and Marketing Examples with KNIME

With this article we wanted to announce the availability of a public repository on the KNIME Community Hub, named "*Machine Learning and Marketing*", for marketing analysts. A mixed team of KNIME users from industry and academia has created, developed, and maintained, some machine learning based solutions for a few commonly used interesting use cases in marketing analytics: churn prediction, sentiment analysis, topic detection to evaluate customer experience, image mining, and keyword research for Search Engine Optimization.

All workflows are available for free. They represent a first sketch to solve the problem but can of course be downloaded and customized according to your own business requirements and data specs.

This article was first published on our KNIME Blog. Find the original version here.

The workflows described in this blog post are available in KNIME's Machine Learning and Marketing space on the KNIME Community Hub.

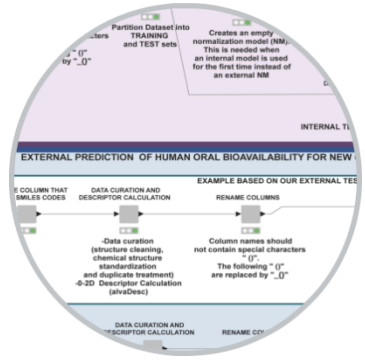
Download KNIME Analytics Platform here.

Referenced scientific article: F. Villarroel Ordenes & R. Silipo, "Machine learning for marketing on the KNIME Community Hub: The development of a live repository for marketing applications", Journal of Business Research 137(1):393-410, DOI: 10.1016/j.jbusres.2021.08.036

Christophe Molina was nominated Contributor of the Month for June 2022. He was awarded for his activity within the KNIME community, including his presence on the KNIME Forum and being a speaker at our events. Christophe is mostly known in the KNIME community for his active and resolutive presence on the Forum, where he proactively assists users. Furthermore, he co-authored several scientific papers on various QSAR topics where he has used KNIME. Some of them include: *ADME Prediction with KNIME: Prediction of Human Oral Bioavailability*, *ADME prediction with KNIME: In Silico Aqueous Solubility*, *Automated Framework for QSAR Modeling of Highly Imbalanced Data*, *ADME prediction with KNIME: A retrospective contribution to the second "Solubility Challenge"*, *Isometric Stratified Ensembles: Adaptive Applicability Domain and Consensus Classification of Colloidal Aggregation*. The image on the right shows a snippet of his solubility workflow.

Christophe is a Data Analyst with a deep research experience in pharmaceutical related fields such as Cheminformatics, Bioinformatics, and Genomics and has more than 20 years of working experience in Data Analytics. He is now Freelance Data Analyst and CEO at PIKAÏROS, a privately held Data Analytics company.

Visit Christophe's profile in the KNIME Forum (Forum handle: `aworker`).



ADME prediction with KNIME: A retrospective contribution to the second “Solubility Challenge”

Authors: Gabriela Falcón-Cano, Christophe Molina & Miguel Ángel Cabrera-Pérez

Editor's Note:
This article corresponds to the publication by Falcón-Cano G, Molina C, and Cabrera-Pérez MÁ: ADME prediction with KNIME: A retrospective contribution to the second “Solubility Challenge”. ADMET & DMPK. 2021 Jul; 9(3):209-218. http://dx.doi.org/10.5599/admet.979

Introduction

Pharmacokinetic parameters are usually influenced by a combination of different physicochemical properties. Among these, solubility has occupied a very important role due to its influence on the absorption process. The need to balance solubility, avoiding excess or insufficiency, is a challenge from the perspective of drug discovery.

In this regard, several research efforts have been made to provide accurate prediction of aqueous solubility through Quantitative Structure-Property Relationship (QSPR) approaches. Undoubtedly, the first and second “Solubility Challenges” proposed by Llinas et al. have been a very effective indicator of the progress and state-of-art of solubility estimation [1, 2]. Recently, Llinas et al. have reviewed the results of the second “Solubility Challenge” to analyse the evolution of the computational methods used in this prediction task and the influence of data quality on the results [3].

In our previous publication, we presented a new method based on recursive random forest approaches to predict aqueous solubility values of drug and drug-like molecules [4]. It was based on the development of two novel recursive machine-learning approaches used for data cleaning and variable selection, and a consensus model generated by the combination of regression and classification algorithms. This model was able to provide good solubility prediction compared to many of the models described in the literature. Considering that our model was developed from a database of aqueous solubility values with limited information on the experimental conditions of the solubility assay, could our model successfully predict the intrinsic solubility values of the two sets of drugs used in the second “Solubility Challenge”?

The present study describes the performance of our model with the molecules of the second “Solubility Challenge” and the comparison of the results with those obtained with the best performing models of the competition. It is necessary to clarify that, for this task, the model was not trained, retrained or optimized based on the molecules of the challenge tests, i.e., the model parameters or hyper-parameters remained exactly the same as those set in previously published work [4].

Materials and methods

Challenge sets

The second "Solubility Challenge" consisted of evaluating the intrinsic solubility estimation of two sets of drugs. The first set is composed of 100 drugs with an average inter-laboratory standard deviation estimated of ~0.17 log units. The second test set consists of 32 "difficult" drugs, characterized by poor inter-laboratory reproducibility: Standard Deviation ~0.62 log units. A detailed list of these molecules have been shown in a previous paper [3].

Software

The Konstanz Information Miner (KNIME) is a free and public software tool that has become one of the main analytical platforms for innovation, data mining and machine learning. The flexibility of workflows developed in KNIME to include different tools allows users to read, create, edit, train and test machine learning models, greatly facilitating the automation of predictions and application by any user [5,6]. In this study, we used the open-source software KNIME Analytical Platform version 4.0.2 [7] and its free complementary extensions for transformation, analysis, modelling, data visualization and data prediction. For the generation of molecular descriptors from structures, the "Descriptor" node from "alvaDesc" extension [8] and the "RDKit Descriptor" node [9] were employed.

Modelling dataset

To predict the molecules of the second "Solubility Challenge", we used as the training set the curated set of aqueous solubility published in our previous paper. This set consists of two large aqueous solubility databases [10, 11]. For each molecule, taking the SMILES (Simplified Molecular Input Line Entry Specification) code as input format, a structure cleaning, standardization, and duplicate removal protocol was developed. The InChi (IUPAC International Chemical Identifier) code was used for duplicate identification and the standard deviation among experimental measurements was computed. A detailed description of this procedure has been shown in our previous article [4]. Although the hypothesis that - *the quality of the experimental data is the main limiting factor in predicting aqueous solubility* - has been challenged [12], any variability in the experimental protocol is always "noise" for in silico modelling purposes. In this sense, our model had several challenges such as: 1) the pH value for the solubility measurement of the collected compounds was not stated, 2) the solid form of the molecule (polymorphs, hydrates, solvates, amorphous) was not characterized in the reported solubility measurements, 3) it was not possible to verify the type of solubility measurement (kinetic or thermodynamic) and 4) the experimental measurement method was not specified.

Modelling algorithm

Due to the uncertainty of the database, we considered the importance of a rigorous protocol for data selection in the development of the original model, in order to discriminate those molecules with potential unreliability. As a first step, we selected a RELIABLE Test Set, consisting of molecules with more than one reported measurement and with inter-source standard deviation greater than 0 and less than 1 logarithmic unit. We used beyond 1 logarithmic unit as a threshold to discriminate unreliable samples. This RELIABLE Test Set was used for model optimization.

From the QSPR perspective, it is necessary to select a set of descriptors that leads to the most predictive model and facilitates model interpretation. To this end, we developed a recursive variable selection algorithm based on regression random forest (RRF). RRF is a widely used ensemble method that assembles multiple decision trees and outputs the consensus predictions from individual trees [13]. It is recognized for its ability to select "important" descriptors. Based on this ability, we use the number of occurrences of a variable in the RRF as a measure of the descriptor's importance, combined with a correlation analysis between variables to avoid collinearity. Each numerical descriptor was injected in the RRF in two ways: nonshuffled and shuffled. Once the individual decision trees were trained and extracted from the ensemble, the total number of occurrences of each variable was calculated. Only variables with a number of occurrences greater than a marginal threshold of 110 were retained. Among those, variables were discarded if the non-shuffled variable had a number of occurrences lower than the number of occurrences of its homologous shuffled variable. All shuffled variables were eventually discarded too. The final set of variables was selected recursively by initially computing the linear correlation between variables, and then keeping only those with the highest number of occurrences among variables with a correlation coefficient greater than a threshold of 0.51 between them.

In an attempt to reduce the uncertainty of the data, independent of any external set, a cleaning procedure based on an RRF approach was developed. This procedure uses the Prediction Variance (PV) of the RRF as a metric to discriminate unreliable samples. The PV is an RRF score that highlights the variability of each individual prediction with respect to the mean. A high PV can be a sign of anomalous behaviour or uncertainty. This procedure was applied to the UNRELIABLE Set, i.e. molecules with aqueous solubility standard deviation between sources equal to 0 or greater than 1. To set the parameters of this algorithm, the minimization of the root mean squared error (RMSE) of the RELIABLE Test was used as the objective function. First, the UNRELIABLE Set was randomly divided into two sets of 50 % and 50 % cardinal. A regression random forest was trained on one of the two sets and used to predict the aqueous solubility and PV of the other set. In addition, the PV of the out-of-bag samples was also calculated. Recursively, molecules were classified as within the PV threshold (CLEAN data) or alternatively as beyond the PV threshold (UNCLEAN data), until no molecules changed from CLEAN to UNCLEAN labelled set or vice versa.

Using the CLEAN set, a Gradient Boosting Model (GBM) was trained for classification using logS = -2 as the cut-off to label molecules into highly soluble or soluble and slightly soluble or insoluble. Two independent RRF models were developed based on these two subsets of labelled molecules and one more RRF model was trained on all CLEAN data. Finally, the average prediction among the three GBM models was assumed as the final prediction value. The parameters of all models were optimized based on the RMSE minimization of the RELIABLE test set. Full details on our developed algorithm are given in previous published paper [4].

Second "Solubility Challenge" prediction

First, we ensured that all test set molecules found in the initial source set used as the training set were removed. Since the model was previously validated using the RELIABLE Test Set and by 5-fold crossvalidation, we used the entire database (including the RELIABLE Test Set) to predict the test challenge samples. To analyse the performance of the solubility regression models, two types of coefficient of determination (r2), root mean squared error (RMSE), mean absolute error (MAE), bias and the percent of molecules with an absolute error less than 0.5 logarithmic units (% 0.5 log) were calculated.

Results and Discussion

Model performance

The statistics obtained for both sets (Test Set 1 = 100 molecules and Test Set 2 = 32 molecules) are shown in Table 1 and Figure 1. To demonstrate model robustness, the results are reported as mean and standard deviation (Std).

Test	r^2 (validation)		r^2 (Pearson)		RMSE (validation)		MAE (validation)		Bias		% 0.5 log	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Test Set 1 (N = 100)	**0.458**	0.01	0.58	0.01	**0.925**	0.03	0.74	0.03	-0.234	0.01	40	1
Test Set 2 (N = 32)	**0.777**	0.02	0.78	0.01	**1.019**	0.1	0.77	0.1	-0.278	0.02	40	6

Performance of the final consensus model for the molecules of the second "Solubility Challenge".

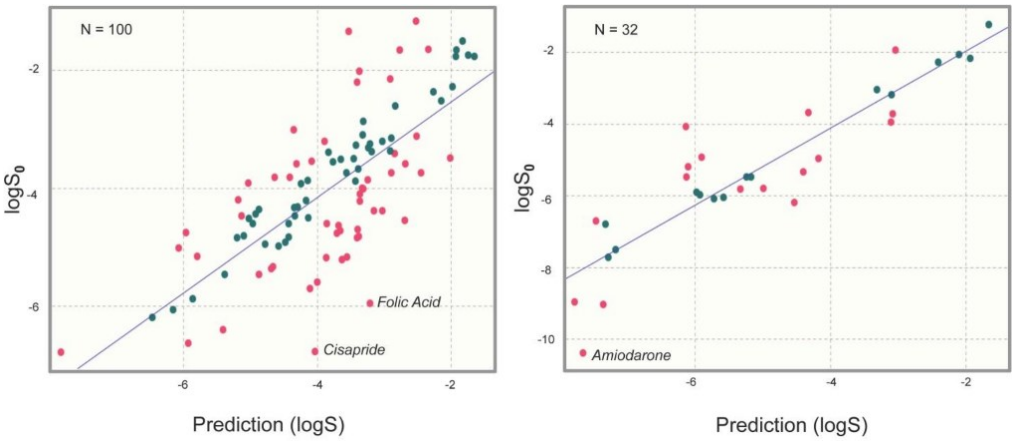


Figure 1. Plot of log S (predicted) vs log S0 (experimental) for both test sets. Molecules with residual values higher than 0.5 (logarithm units) are highlighted in red.

Figure 2 compares our results with the top-rank models of the second "Solubility Challenge". According to the mean RMSE value, our consensus model ranks ninth among the top-ranked models for the prediction of Test Set 1 and first for the prediction of Test Set 2.

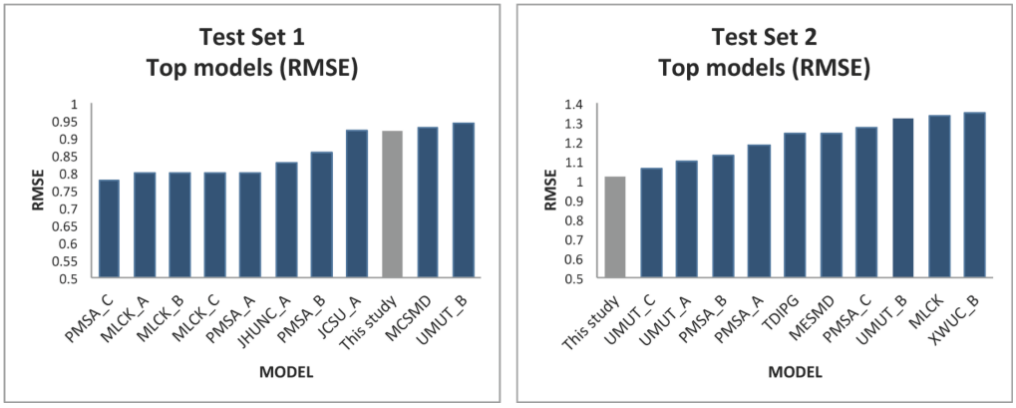


Figure 2. Comparison between the top-rank models of the Second Solubility Challenge and our results (according to RMSE).

Although there are no significant differences in terms of prediction performance, the training set we have used contains aqueous solubility measurements under non-specified experimental conditions (pH, method and solid form), without information on

their type of solubility (aqueous or intrinsic). It is known that the presence of acidic and basic groups in a molecule and the pH of the medium affect the solubility value. Intrinsic solubility corresponds to the solubility of the uncharged molecular species, whereas aqueous solubility depends on the pH used for measurements. Therefore, not all the values in the training set are true intrinsic solubility values, which influences the model prediction of the external test set with intrinsic solubility measurements, leading in some cases to higher uncertainty for samples contained in the training set.

We analysed the overlap of our source set with the molecules from the second "Solubility Challenge", resulting on two overlaps of 88 and 21 molecules, 1st and 2nd test respectively. Only for the case of these 109 overlapping molecules, a correlation analysis was performed between the intrinsic solubility values reported in the second "Solubility Challenge" and the aqueous solubility values reported in our initial source set. The overlapping molecules were eliminated from the training set for modelling purposes. This analysis is shown in Figure 3.

Considering the lack of real intrinsic solubility values in the training set, the most problematic molecules in the second "Solubility Challenge" should be the ionizable compounds. The analysis of residuals showed that Amiodarone (TS2), Cisapride (TS1) and Folic Acid (TS1) are response outliers. All of them contain at least one acidic or basic functional group and are practically insoluble compounds. For these molecules, the aqueous solubility value (log Sw) is different from the intrinsic solubility value, since not enough solute is dissolved to modify the pH in order to maintain a near-neutral species in the poorly buffered medium. Table 2 describes the values of log S0 (second "Solubility Challenge"), log Sw (initial data source), log Sw (reported in other sources) and log Sw (predicted).

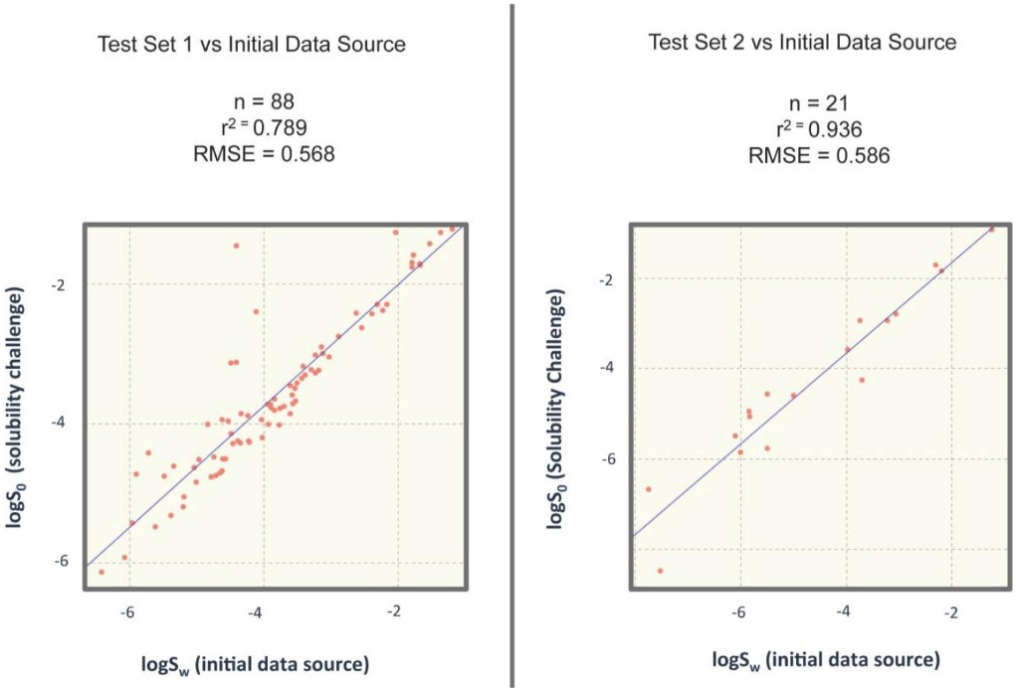
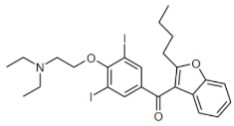
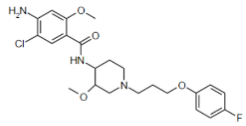
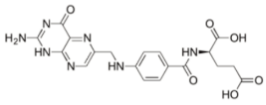


Figure 3. Overlapping log S_0 against log S_w analysis between the molecules of the second "Solubility Challenge" and the training set. For modelling purposes, these overlapping molecules were eliminated from the training set.

Structure	Name	log S_0 [a]	log S_w [b] (initial source set)	log S_w (predicted)	log S_w [c] (other sources)
	Amiodarone	-10.4	-9.35	-7.54	-7.17 [14]
	Cisapride	-6.78	-5.23	-4.27	-4.7 [15]
	Folic Acid	-5.96	-5.44	-3.12	> -2.87 [15]

[a]Intrinsic Aqueous Solubility reported in the second "Solubility Challenge", [b]Aqueous Solubility reported for the three outliers in the initial source set, [c]Aqueous Solubility reported in other sources

Table 2. Summary of solubility values for the outliers.

To assess whether the method was able to deal with the uncertainty in the data, a simple experiment was performed. As shown in Figure 3, 88 molecules from the first test set of the challenge overlapped with our initial source set. A correlation analysis

between the two solubility values reported by each overlapping molecule showed a root mean squared error of 0.568 log units. We assume that the value reported in the challenge refers to a curated and reliable measurement, whereas the value reported in our initial source set could be of potential uncertainty. There is a significant difference between the two sets of values for the 88 molecules (Confidence interval (CI): 95%; p = 2.9E-5). Next, a paired-sample t-test was developed for comparing the performance of two models based on two different training sets: (a) the literature solubility data reported in our initial source set and (b) the reliable intrinsic solubility measurements reported in the first set of the challenge. Both models were evaluated on the second challenge test. There was no significant difference (CI: 95%; p = 0.58) between the root mean squared errors achieved on the second challenge test using one or the other training sets. However, if a single random forest regression without recursive selection of data and variables and without applying a consensus model is used as the modelling algorithm, the t-test highlights a significant difference (CI: 95%; p = 3.3 E-6). The influence of data quality on model performance depends on the modelling procedure used. Thus, data quality was not the determinant factor when an appropriate modelling approach was designed to address data uncertainty by selecting the most important variables and using a consensus model of combined single model predictions. Table 3 shows a review of the results.

Test	Reliable solubility measurements (data challenge) n (training) = 88		Literature solubility data (reported in Initial Data Source) n (training) = 88	
	r^2 (validation)*	RMSE (validation)*	r^2 (validation)*	RMSE (validation)*
Recursive Random Forest (consensus)	0.30 (0.05)	1.79 (0.06)	0.29 (0.05)	1.80 (0.05)
Single Random Forest Regression	0.19 (0.01)	1.93 (0.02)	0.14 (0.06)	1.98 (0.06)

*The results are reported as Mean (Std). The Std was computed by repeating 10-times the modelling procedure.

Table 3. Mean with Std statistics based on two training sets when predicting the second test of the second "Solubility Challenge" using our method (Recursive Random Forest (consensus)) versus a single RRF: reliable solubility measurements (data challenge) and literature solubility data.

Automated system for aqueous solubility prediction

We trust there is a need to make publicly available a reliable and diverse data set of intrinsic solubility measurements for a rigorous comparison between modelling algorithms, due to the relative influence of data quality on the performance of a model. Furthermore, applicability and reproducibility of solubility QSPR models should be a priority for data to be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR) [16–18]. In this regard, the final purpose of the current commentary is to make publicly available an automated system for in silico aqueous solubility assessment. Our model has been successfully validated in a previous published study and has been blind tested with the second "Solubility Challenge", showing an adequate performance. The KNIME

workflow published with the paper contains the results of our model on the second "Solubility Challenge" and allows the prediction of new sets. The user can download the workflow and follow the instructions it contains from https://pikairos.eu/download/__aqueous_solubility_prediction/. We developed a version based on RDKit and AlvaDesC descriptors, calculated using the "Descriptor" node contained in the "alvaDesc" extension. AlvaDesc 1.0.16 is available with academic or commercial licenses, which can be obtained by requesting a quote online (registration required) or by contacting them directly by email (chm@kode-solutions.net). Only the SMILES codes of the structures are needed for aqueous solubility prediction, as the model does not require any experimentally determined value for solubility calculation. The model is characterized by its simplicity since it is only based on 0-2D descriptors. In addition, the model is implemented in the open-source analytics platform KNIME, which is a user-friendly software suitable for further data analysis and visualization.

Conclusions

The results obtained with the evaluation of the second "Solubility Challenge" reinforce the idea that data quality is not the major limiting factor for obtaining adequate solubility predictions if the implemented modelling methodology can cope with data uncertainty. In our case, the developed algorithm was able to overcome data variability to obtain acceptable aqueous solubility prediction results. The results published here are a blind prediction, since the experimental aqueous solubility values of the challenge test set were not accessible at the time of our model development and training. Although the achieved performance is comparable to those reported in the review of the second Solubility Challenge, our model is only based on public data compared to some of the best models of the second Solubility Challenge, which were based on the huge aqueous solubility databases available from pharmaceutical companies. Furthermore, the algorithm of our model is global, as demonstrated by the use of generic data without the bias of "training close to the test data". The automation of the proposed methodology and its possible application on larger databases, collected under more homogeneous conditions, could be a step forward to improve solubility prediction during drug discovery and development stages. In attention to the importance of sharing data and methods to ensure reproducibility and applicability of QSPR models, we made the data publicly available along with our predictive model based on the KNIME Analytical Platform as a new free tool for the assessment of aqueous solubility of drug candidates.

References

[1] A. Llinàs, R. C. Glen, J. M. Goodman. Solubility Challenge : Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?. J. Chem. Inf. Model. 48 (2008) 1289–1303. https://doi.org/10.1021/ci800058v.

[2] A. Llinas, A. Avdeef. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets. J. Chem. Inf. Model. 59 (2019) 3036– 3040. https://doi.org/10.1021/acs.jcim.9b00345.

[3] A. Llinas, I. Oprisiu, A. Avdeef. Findings of the Second Challenge to Predict Aqueous Solubility. J. Chem. Inf. Model. 60, (2020) 4791–4803. https://doi.org/10.1021/acs.jcim.0c00701.

[4] G. Falcón-Cano, C. Molina, M. Á. Cabrera-Pérez. ADME prediction with KNIME: In silico aqueous solubility consensus model based on supervised recursive random forest approaches. ADMET DMPK 8 (2020) 1–23. https://doi.org/10.5599/admet.852.

[5] P.M. Mazanetz, J.R. Marmon, B.T.C. Reisser, I. Morao. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. Curr. Top. Med. Chem. 12 (2012) 1965–1979. https:/doi.org/10.2174/156802612804910331.

[6] M.-A. Trapotsi. Development and evaluation of ADME models using proprietary and opensource data. University of Hertfordshire, 2017. https://doi.org/10.18745/th.19719.

[7] "KNIME Analytics Platform 4.0.2." [Online]. Available: https://www.knime.com/download-previous-versions. [Accessed: 17-Mar-2021].

[8] A. Mauri, "alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints," in Ecotoxicological QSARs. Methods in Pharmacology and Toxicology, K. Roy, Ed. Humana Press Inc., 2020, pp. 801–820.

[9] "RDKit KNIME Integration." [Online]. Available: https://www.knime.com/rdkit. [Accessed: 19-Jun2020].

[10] M.C. Sorkun, A. Khetan, S. Er. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Sci. Data 6 (2019) 1–8, Dec. 2019. https://doi.org/10.1038/s41597-019-0151-1.

[11] Q. Cui, S. Lu, B. Ni, X. Zeng, Y. Tan, Y.D. Chen, H. Zhao. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. Front. Oncol. 10 (2017) 1–9. https://doi.org/10.3389/fonc.2020.00121.

[12] D.S. Palmer, J.B.O. Mitchell. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules?. Mol. Pharm. 11 (2014) 2962–2972. https://doi.org/10.1021/- mp500103r.

[13] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958. https://doi.org/10.1021/ci034160g.

[14] M. Salahinejad, T.C. Le, D.A. Winkler. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help?. Mol. Pharm. 10 (2013) 2757–2766. https://doi.org/10.1021/mp4001958.

[15] S.H. Yalkowsky, Y. He, P. Jain. Handbook of Aqueous Solubility Data, Second. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, USA: CRC Press Taylor & Francis Group, 2010.

[16] M. D. Wilkinson, M. Dumontier, I.J. Aalbersberg et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3 (2016) 1–9. https://doi.org/10.1038/sdata.2016.18.

[17] J. Wise, A.G. de Barron, A. Splendiani et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24, (2019) 933–938. https://doi.org/10.1016/j.drudis.2019.01.008.

[18] K.M. Merz, R. Amaro, Z. Cournia, M. Rarey, T. Soares, A. Tropsha, H.A. Wahab, R. Wang. Editorial: Method and Data Sharing and Reproducibility of Scientific Results. J. Chem. Inf. Model. 60 (2020) 5868–5869. https://doi.org/10.1021/acs.jcim.0c01389.

KNIME Support

This category combines all the helpful contributions made by our COTMs. This could be on the KNIME Forum helping out other KNIME users by proposing solutions to questions, or by sharing thoughts and suggestions for specific use cases. Luckily, our army of helpers is not only present on the KNIME Forum but other places like Facebook or Twitter. If you ever end up needing advice, it's likely that one of our support KNinjas will be on the spot. The category "KNIME Support" features:

- **Evan Bristow**

  – Senior Principal Analyst *@Genesys*

- **Miguel InfMad**

  – Marketing & Digital Analytics Expert *@BaseCero Marketing*

- **makkynm**

  – n/a

- **Ignacio Perez**

  – Director, Owner *@IQuartil*

- **Markus Lauber**

  – Senior Data Scientist, Big Data Analytics *@Deutsche Telekom*

- **Bruno Ng**

  – Director Data Ops *@Triton Digital*

- **Brian Bates**

  – Data & Integration Architect *@The Walt Disney Company*

Evan Bristow was nominated KNIME Contributor of the Month for January 2021. Together with Miguel InfMad, he was awarded for building and nurturing a Facebook community for all users – from newbies to experts. The group has been around since 2019 and counts more than 1500 members. It is a lively, competent, and very helpful group regarding your data science problems and KNIME questions. His work of support does not end on Facebook. You can find his contributions on the KNIME Forum and on the KNIME Community Hub.

Evan is a long time KNIME user and data science expert. He describes himself as an enthusiastic and creative person who enjoys the art and science of research. Evan is currently Senior Principal Analyst at Genesys where he uses KNIME to utilize various analytical and machine learning approaches to give business insights and support business needs.

Visit Evan's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `evanb`).

Expertise at the Service of the Community: What Support is really about

My Data Guest – An Interview with Evan Bristow

Author: Rosaria Silipo

It was my pleasure to recently interview live on LinkedIn Evan Bristow as part of the *My Data Guest* interview series. He shared insights into what it means to run a KNIME support group on Facebook to help new and expert KNIME users in their learning journey, what are the things you can expect in such a role, and where you should draw the line in the support you provide.

Evan Bristow is a Senior Principal Data Analyst at Genesys, where he utilizes various analytical and machine learning approaches to give business insights and to support business needs. Over the years, he has applied his expertise and technical skills to conduct competitive market analyses, provide methodological support for business-to-business customer relationships, and produce recurring and ad-hoc operations and financial reporting for organizations in the AMER region. When asked about what he does, he likes to make sure people understand that he is not just providing the "what", but also the "why", the "how", and the "what if". Evan is also one of the passionate and expert brains behind the KNIME support group on Facebook —together with co-founder Miguel InfMad— and enjoys putting his expertise at the service of the community.

Rosaria: *Can you explain what you mean when you are saying that you are not just providing the "what", but also the "why", the "how", and the "what if"?*

Evan: One thing that is very important to me is taking analytics to another level. Not only being able to provide the numbers but also giving them context and meaning.

Let's say you are working for a company and they want to know how much they are going to book next quarter. That's a very interesting question which you can attempt to model in a number of different ways. What's crucial though is that you digest those types of questions into the constituent components: "Are we going to miss something?", "Are we over projecting?", "What are possible outcomes we need to take

into account when setting up a model?". Answering those questions defines the "what if".

If you're seeing differences in what you are projecting vs. what other people are projecting then you can drill into your model and determine the "why". Also, if you are doing some kind of root cause analysis and you are seeing differences in the makeup of your business between one period and another, you can observe what's driving those differences.

Once those causes are identified, you should be able to put them in a broader context. This is when business stakeholders may ask: "How much are we going to book next quarter?", "What have we booked in that quarter historically?" and "How does that compare over time?". In the end, what they really want to know is "Are we doing well?".

Rosaria: *What do you do exactly in your daily job? What are your tasks as a Principal Data Analyst?*

Evan: I like to think of myself as a data MacGyver. I take different data sources or different pieces of information and pull them through an analytical process to create something that answers the question or solves the problem. At the moment, I'm working on a couple of different models. Two of them are projecting how much bookings we're expecting out of our current pipeline. The other one computes opportunity level scoring, that is how likely it is to win individual opportunities. This gives another view on risk assessment to our pipeline. Whenever the data grows big or the game gets tough, KNIME Analytics Platform is always my go-to tool.

I'm not a coder at heart so for me it's a lot faster and more accessible to build workflows in KNIME. In my opinion, the greatest advantage is how transparent and easy to understand visual data flows are. I can always pull up a workflow, show it to a stakeholder, and we can address questions on the fly. If I were to do that in Python and write out code at the same time, I would lose business stakeholders' attention in the blink of an eye.

Rosaria: *How do you manage to support all KNIME users on the Facebook group so actively? I mean, you have a job, how do you find the time?*

Evan: Most of what we usually do in the KNIME Facebook group is getting people who are familiar with another tool or another way of doing things acclimatized to KNIME. We tend to get more of a novice group of individuals who are familiar, for example, with Excel but want to know how they can do the same thing using KNIME. Those are usually not too complicated topics where we necessarily want to refer somebody to the KNIME Forum. Additionally, we use the space to point the community to available learning resources. For example, the books of the KNIME Press and in particular the Excel to KNIME series, which is available in different languages. Our motto is "You teach them to fish rather than just giving them a fish".

However, when we get some oddball questions that require more advanced expertise and thinking, I usually refer people to the KNIME Forum.

Rosaria: *What is the difference between the KNIME Facebook group and the KNIME Forum?*

Evan: The KNIME Facebook group is a very easy, low-stake place where you just go and ask questions. Very often questions come from KNIME users who utilize the software on the job and may have, for example, data wrangling questions —e.g., combining, grouping, filtering, etc.— but cannot count on KNIME support within their department or organization. What these users need is just somebody they could throw a quick question to, and get an immediate response. This is what the Facebook group is for.

The KNIME Forum, on the other hand, is for users to ask more articulated, technical or specialized questions, and receive support directly from KNIMErs or other expert users.

Rosaria: *Are there questions that should not be answered —or even asked— in a support group? Besides the obviously inappropriate questions, of course.*

Evan: We will help you if you have a specific question about how something works. If you need help with your homework and you would like to know what the Pivoting node does, we will help you with that. If something is bugging out or throwing an error, go ahead and ask on the KNME Facebook group. However, if you have a data set and ask us how to apply clustering strategies, we won't help you as this is clearly a homework assignment. In general, we expect users to do a little bit of the legwork on their side before coming to us for help.

Rosaria: *Are you also actively supporting KNIME users on the KNIME Forum?*

Evan: Not as much as I used to be. I was pretty active a couple of years ago, and I still check in every now and then. The truth is that balancing a full-time job, family duties, and the KNIME Facebook group is hard enough. I simply cannot keep up with everything.

Rosaria: *What are the most frequent questions you get in the KNIME Facebook group?*

Evan: Our most frequently asked questions follow the theme of tool migration. They usually come from people who know how to manipulate data using tool X and want to know how to do the same in KNIME. The second largest group of frequent questions is usually related to general purpose machine learning questions, for example: "Why do you partition your data into train and test set?" or "How would you do cross validation?". Finally, quite often, we also receive questions about very specific data wrangling operations, such as "How do I filter based on certain criteria?". Tool migration, machine learning, and data wrangling —those are the three main themes.

Rosaria: *Those questions are somewhat to be expected. Now I am curious. What is the most complicated question that was ever asked in the group?*

Evan: Usually if questions get too complicated, we'll shove them off to the KNIME Forum. I think the hardest question had to do with Regex. Somebody had a string field that was oddly populated with a date and some unique identifier, and they wanted to parse that into the date and the unique identifier. The problem was also that the number of unique identifiers could change depending on the data row. Using text manipulation nodes, splitting cells and pivoting was the suggested way to tackle this problem —I believe. But that is certainly not the only one.

KNIME is like a Swiss army knife of data analytics and there are dozens of different ways to tackle problems, and come up with the same result. Those unique use cases are fun to answer though, and it's always interesting to see how other users answer the questions.

Rosaria: *Before people actually ask for support, they need to start learning about KNIME in a way or another. How do you advise newbies to start their journey with KNIME?*

Evan: The best way is to come up with a project where you know the data relatively well and what the goal of the project is. If you can clearly define the steps that you need to implement to reach the goal that you have in mind, you can just connect point A to point B. And if you are stuck you can always search the KNIME Hub for a workflow example that approximates your use case, and get inspired by the flow that somebody else designed. Unlike other analytics tools —say Excel, for example— KNIME is visual, sequential and transparent, and that really facilitates learning.

Rosaria: *How did you learn about KNIME?*

Evan: I discovered KNIME many moons ago. Back then, I was working for a company that was doing B2B marketing research and we were using SAS and SPSS Modeler. The problem was that we only had one copy of SPSS Modeler because it had a hefty price tag. As a result, it was installed only on one person's computer, which was quite inconvenient for team work and to scale projects. So I started looking for an alternative tool, and this is when I learnt about KNIME. We started working with it and realized that it was a lot better at doing things than other tools. Not only was KNIME better to run analyses, but it was also easier to connect to, import, combine and manipulate different data sources. Last but not least, being a free platform, it made the business happy too.

Rosaria: *Did anybody help you in your early days with KNIME?*

Evan: Unfortunately, I didn't have a close support group, so the KNIME Forum was the main place where I would post my questions. In general, though, I like to figure things out on my own. I like pointing, clicking, and playing around. I was really digging into the software by reading node descriptions, for example, and I was using the KNIME

Examples Server a lot to get inspired by those workflows. I have always loved the Examples Server. It's very useful, especially for newbies. Besides that, I googled my questions and found pretty good answers.

Rosaria: *If you were to choose the top three features that you like about KNIME, what would they be?*

Evan: The ability to integrate different data sources and technologies is probably one of the best things in KNIME. I can pull data from essentially anywhere and store it essentially anywhere without worrying about if it will blend. That's something you often have to deal with in business: you've got data stored on a server, you have data in a smartsheet somethere, and someone sends you an Excel file. With KNIME you can bring all those scattered pieces of information together, build an analysis out of it, and easily put it back on your server to create a visualization.

Another great feature is the ability to leverage knowledge from one workflow to another workflow. I build a workflow for one project and I can easily extend it or take pieces of this workflow and use it in a different workflow. I never start completely from scratch and this saves a lot of time and effort.

Lastly, KNIME GUI, and the ability to abstract segments and processes of your workflow into a component. That helps you focus on what you are doing instead of what you are writing.

Rosaria: *And the top 3 nodes you could never do without?*

Evan: First, the Pivoting and DB Pivot nodes. Pivoting in SQL is terrible, pivoting in KNIME is easy. Second, the Statistics node. That's an invaluable node that you can basically use every time when you are pulling in fresh data. It makes it easy to see what kind of data you are dealing with. And lastly, the Filter nodes. Everybody has got to do a little filtering at some point.

Rosaria: *Support is not the only way you interact with the KNIME community. You were a member of the editorial board of the Medium journal "Low Code for Advanced Data Science". How was the experience? What does a member of the editorial board do?*

Evan: This is again another way of helping the community, just using a different channel and medium. Members of the editorial board try to help contributors get their work out there, shape their data stories, ignite conversation on low code tools, and spot interesting how-to topics that are worth sharing with the readers to kick off their own journey towards codeless data analytics with KNIME.

Rosaria: *We are reaching the end of our interview. Before we say goodbye, are there any plans to replicate your success with a support group on another social media channel?*

Evan: I don't have specific plans at the moment, but I could imagine creating a LinkedIn or WhatsApp support group. Some of the general social media chat groups, where

people can casually drop simple how-to questions and get an answer straight away, are something I would like to be involved in. Noting too high-stake. For doozy or very technical questions the KNIME Forum remains the place to go to.

Rosaria: *How can people from the audience get in touch with you and your work?*

Evan: The best way to get in touch with me is via the KNIME Facebook group (KNIME Analyst Community). You are always welcome to join the group, actively post, and help me and the other admins provide support to the KNIME community.

Watch the original interview with Evan Bristow on YouTube: "My Data Guest – Ep 9 with Evan Bristow".

Miguel InfMad was nominated KNIME Contributor of the Month for January 2021. Together with Evan Bristow, he was awarded for building and nurturing a Facebook community for all users – from newbies to experts. The group has been around since 2019 and counts more than 1500 members. It is a lively, competent, and very helpful group regarding your data science problems and KNIME questions. His work of support does not end on Facebook. You can find his contributions on the KNIME Forum and on the KNIME Community Hub.

Miguel is a long time KNIME user and data science expert. His professional development includes working as analyst and project manager of data mining projects applied to marketing. More recently, he has become an expert in digital campaigns of PPC, SMM, and SEM. Miguel holds a PhD in Applied Economics from the Complutense University of Madrid.

Visit Miguel's space on the KNIME Hub (Hub handle: `miguel_infmad`).

How to connect to Google Analytics with KNIME

Author: Miguel InfMad; Translated by: Roberto Cadili

In this article, we will explain how to connect to a Google Analytics account using KNIME Analytics Platform. Our goal is to **automate data extraction** and create **custom KPIs** to monitor the performance of our projects.

For those of us who are not expert coders, sometimes it's hard to smoothly integrate digital and analytical applications. With this miniguide, we hope to help readers interested in digital analytics **go beyond the data displayed by Google Analytics** and enhance their analysis using free and open-source tools.

What's KNIME

KNIME is a flexible and scalable **open source, analytic platform**. It can be used for any kind of data task, e.g., accessing and wrangling data, building predictive models, or integrating a computational framework for Big Data.

To put it straight, KNIME is the perfect "**Swiss Army knife**" for **digital analytics**, and also the tool that we use at BaseCero (as a matter of fact, the company's headquarter is in Zurich, Switzerland).

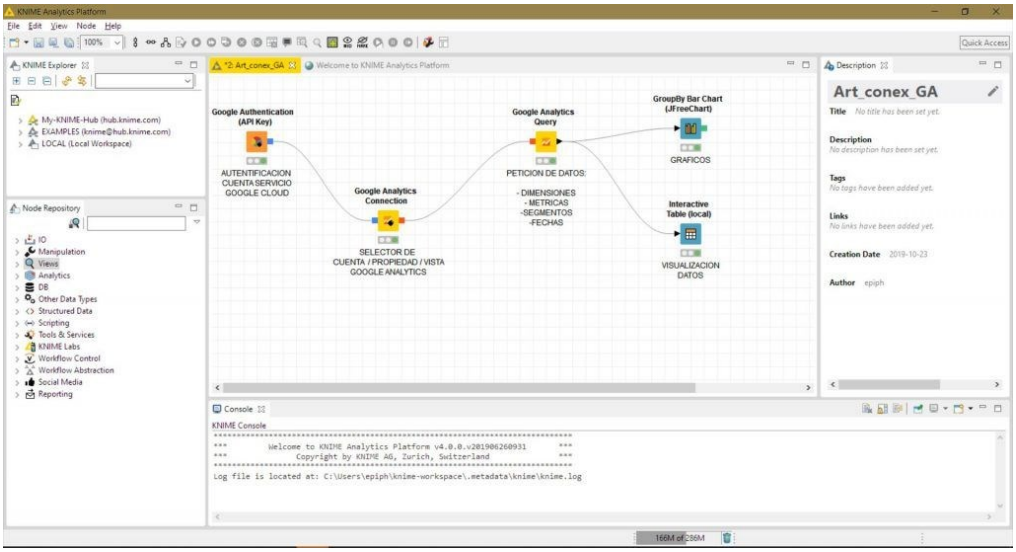
KNIME allows the ingestion and transformation of any data format, the integration with many programming languages, the development and automation of complex processes, the creation of predictive models, the connection to web services and third-party APIs, and much more.

KNIME is a visual programming-based software. This means that **you don't need to know how to program** in a traditional scripting language to be able to use it. Both the base version and the wide range of extensions and integrations to expand KNIME's analytics power can be downloaded for free and are fully functional. You can download the latest version from:

https://www.knime.com/downloads/download-knime

There is also an enterprise version of KNIME Software for deploying data science workflows, automating workflow execution, and managing collaboration across users and spaces.

We highly encourage you to give it a try. On top of the capabilities listed above, the tool is enriched with a large number of workflow examples for a wide range of applications, and if you seek help the KNIME Forum and its **large and active community** is the perfect place to post your questions.



KNIME workflow to connect to Google Analytics, send requests, and plot results.

Advantages of Integrating KNIME With Google Analytics

While the data we can retrieve integrating KNIME and Google Analytics is the same as that displayed on Google Analytics, using an external tool like KNIME offers several **key advantages that are not available in the original service**. Find the most important ones below:

- **Process Automation**

  Building KPIs for your project one by one as you browse through the multitude of data offered by Google Analytics can be fairly time-consuming. Developing an automated process allows us to free up a lot of time to focus on what really matters: interpreting and making sense of the data to improve the project.

- **Custom Extraction of Features and Metrics**

  Sometimes we need to develop custom KPIs that are not available by default in Google Analytics. Additionally, the extraction of raw web data allows us to customize the analytics process and optimize digital resources.

- **Development Using an Open-Source Tool**

  Professional analytics tools developed under the GPL license (General Public License) guarantee more transparent processes and reduce development costs.

- **Data Source Integration**

  Web traffic data can be processed smoothly together with additional data coming from other channels, such as social media or newsletters. This allows for the joint analysis of a multi-channel strategy to evaluate and optimize the development of projects and digital campaigns.

Before We Dive Deeper

In this article, we will focus on how to **access Google Analytics's web traffic data using KNIME**. A basic understanding of Google Analytics and KNIME Analytics Platform is advisable to properly follow the content of each section.

How does Google collect web traffic data?

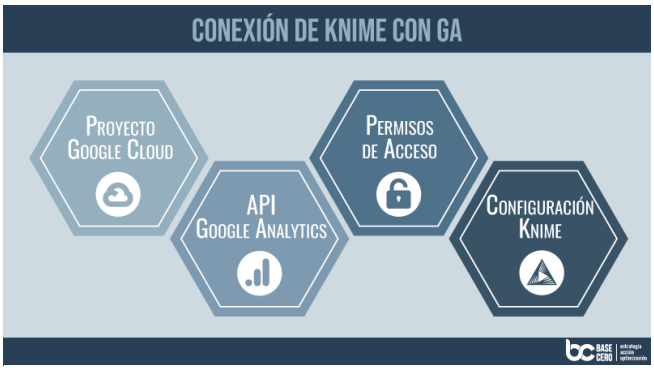
To put it in simple terms, Google Analytics relies on **tracking cookies** installed in the browser to track users' activity and collect web traffic data (provided that the user accepts them). Collected data are then available for inspection directly on Google Analytics, and they can also be accessed using the dedicated API. This second option is what we are going to use in our implementation with KNIME.

Before we start blending these two technologies, we need to:

1. Have a property website with Google Analytics's tracking cookies installed in it.
2. Create a Google/Gmail account to be able to use Google Analytics and Google Cloud Platform.
3. Install KNIME Analytics Platform, the KNIME Google Connectors extension and the KNIME Twitter Connectors extension (for more information on KNIME extensions and integrations check the documentation).

In the next section, we'll illustrate how to **connect to Google Analytics with KNIME in 4 steps**.

Connect To Google Analytics With KNIME

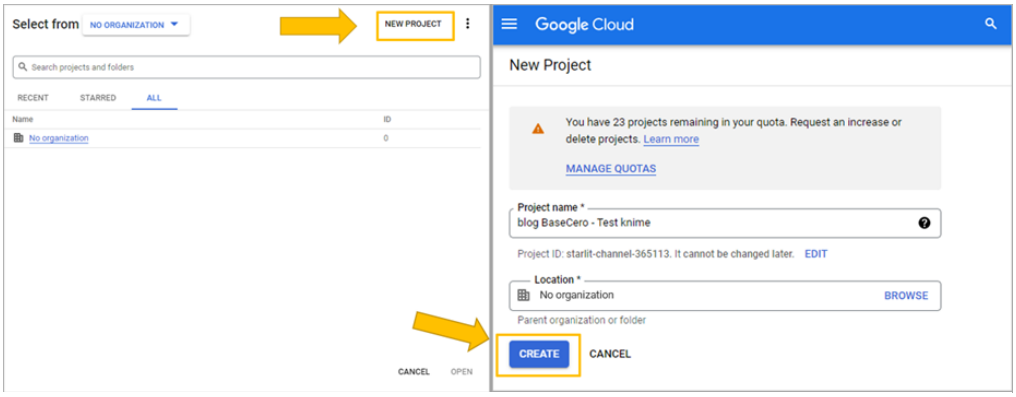


Process flow to connect to Google Analytics with KNIME in 4 steps: 1) Create a Google Cloud project; 2) Activate Google Analytics API; 3) Set access permissions; 4) Configure KNIME Analytics Platform.

Step 1: Create A Project on Google Cloud Platform

We start off setting up a **new project** on Google Cloud Platform. This is required every time you want to build a solution using any of the analytics products offered by Google.

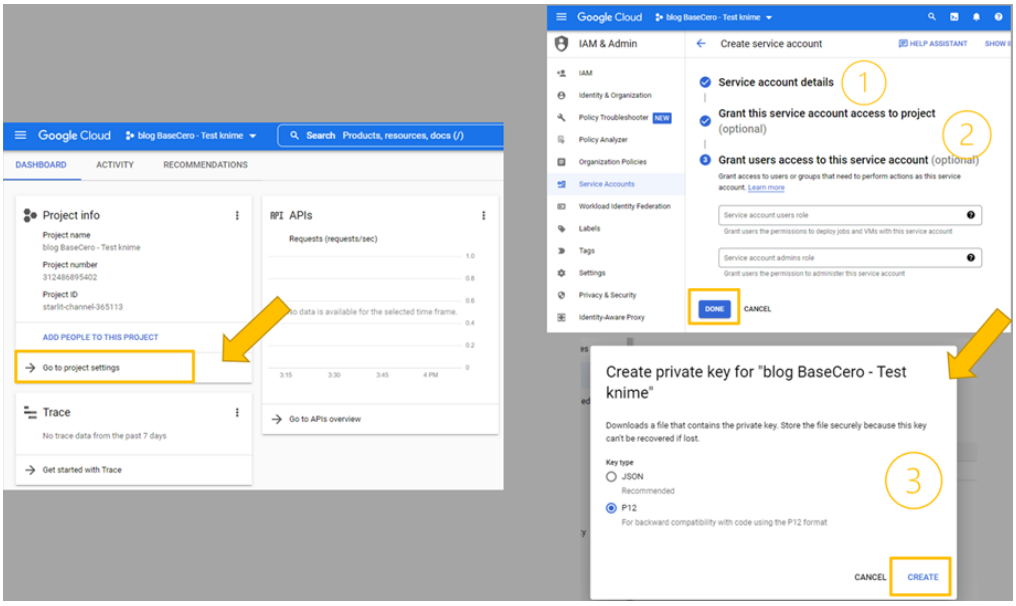
If you don't have a Google Cloud account, you can sign in using a Gmail account.



Creating a new project on Google Cloud Platform.

Within the newly created project, we define a **service account**. This is needed to manage the project access permissions to Google Analytics account(s) by providing a **security file**, unique identification keys, and an email account for connection.

To do that, we open the project configurations, select the option "service account" on the left column, and create one.

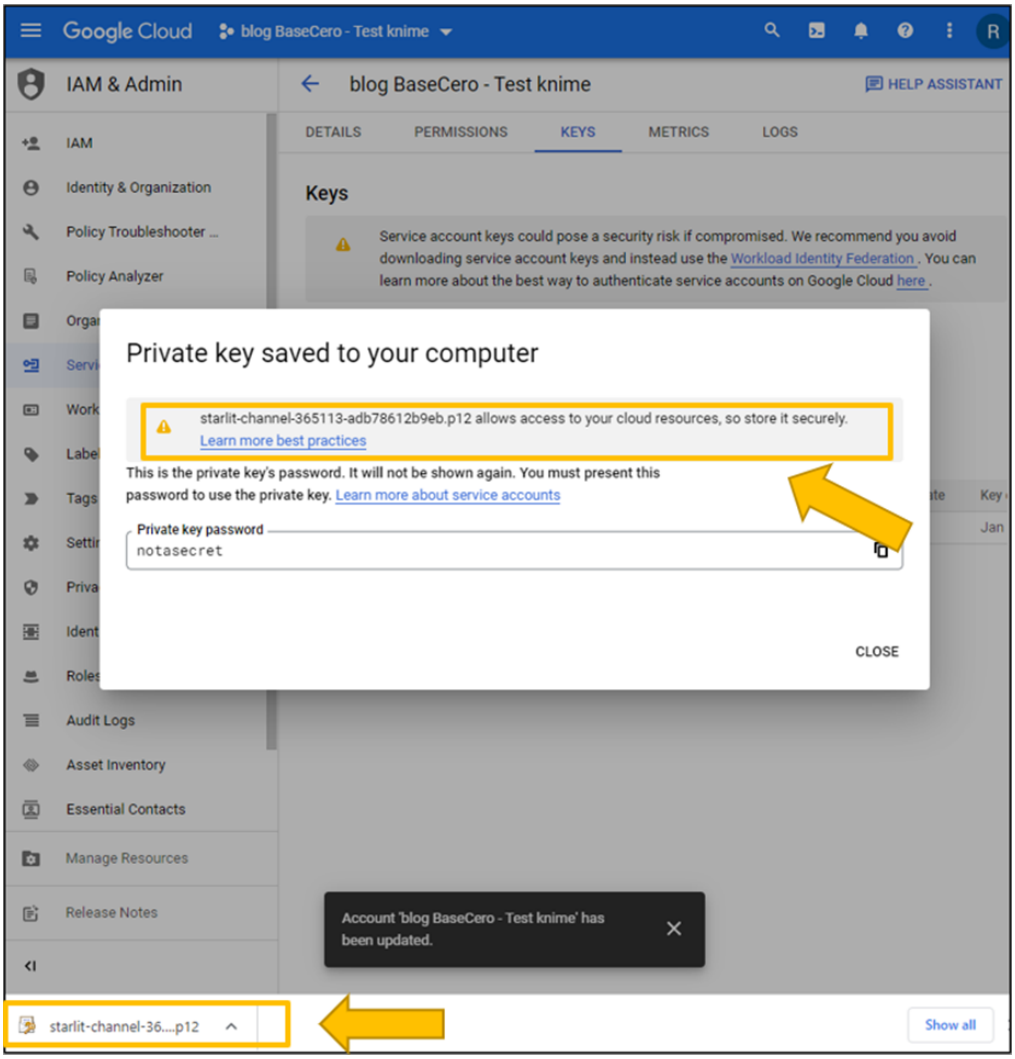


Creating a service account within the newly created project.

The creation of the service account entails 3 steps:

- Step 1 is about defining service account details for ease of identification.
- Step 2 is optional and we can leave it blank.
- In step 3, we create private keys for the service account, and a downloadable security file that we will later use in KNIME to authenticate to Google Analytics.

We now have identification and connection information for the newly created service account, as well as the associated **authentication keys and *p12* security file** stored in a local directory on our machine.



Creating a private key for the service account and a downloadable security file that will later be used in KNIME to authenticate to Google Analytics.

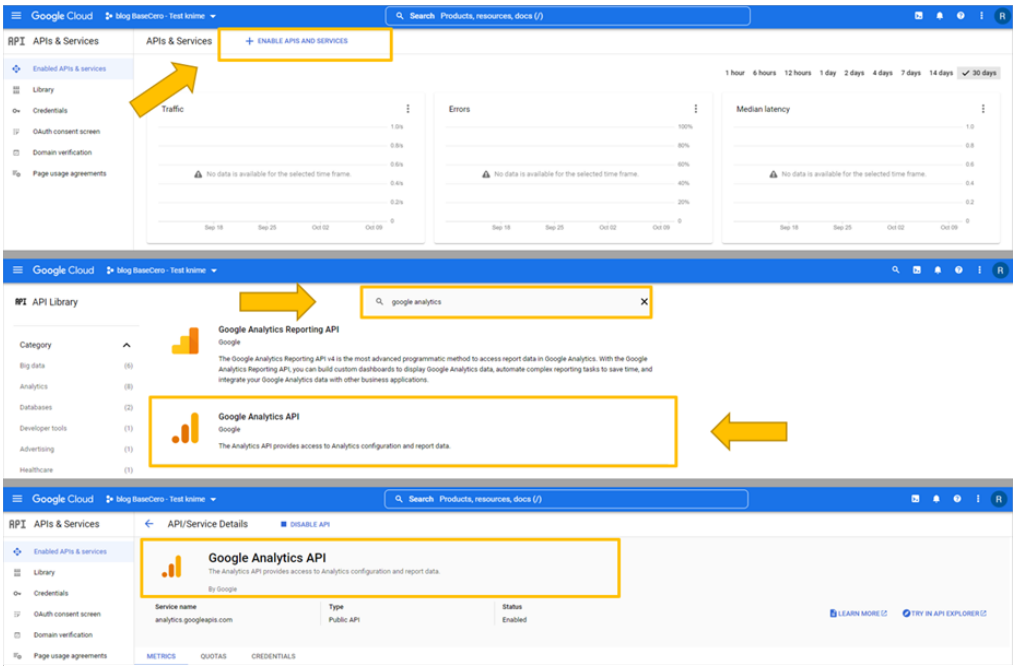
Step 2: Activate Google Analytics API

Let's now **configure the project** that we created on Google Cloud. These projects behave like a **repository** or folder where we can start developing our application on Google Cloud platform, and do **not have any preassigned functionality**.

The application that we **activate** for this project is the **API of Google Analytics**. With this API, we will be able to access and query web traffic data collected by tracking cookies.

To do that, in the upper bar of the project view, we select "**Enable APIs and Services**", we type "**google analytics**" in the search bar, and activate it:

The project is now set, and we are ready to start **using** the **functionality and applications of Google Analytics** to access web statistics, as well as the credentials (i.e., service account, private keys, etc.) that we need to finalize the process.



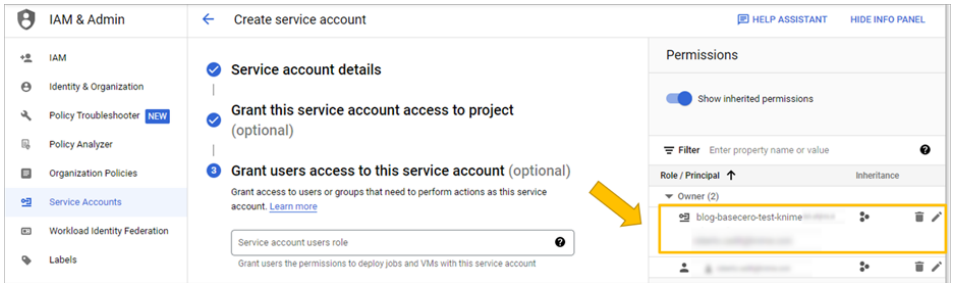
Activating the Google Analytics API.

Step 3: Set Access Permissions

In this step, we will set access permissions to allow our project to connect to traffic web data.

- **If We Are the Administrators of The Google Analytics Account:**

  We need to access the Google Analytics account or property and add the newly created service account (use the email address associated with it) in the option “User Management”.



How to identify the email address associated with the service account

The figure above shows how to identify the **email address associated with the service account**.

To **add authorized users**, in the user management menu of Google Analytics, we click on "*User Management*" from the account or property we want to grant access to. Next, we add a new user using the service account email, and set the permissions to "Read and Analyze" (read more about managing users on Google Analytics).

- **If We Are not the Administrators of The Google Analytics Account:**

  Should this be the case, we need to request the administrator of the Google Analytics account to grant us access.

  Via its authentication and identification system, Google guarantees that no web information can be accessed without having the necessary permissions granted by the administrator.

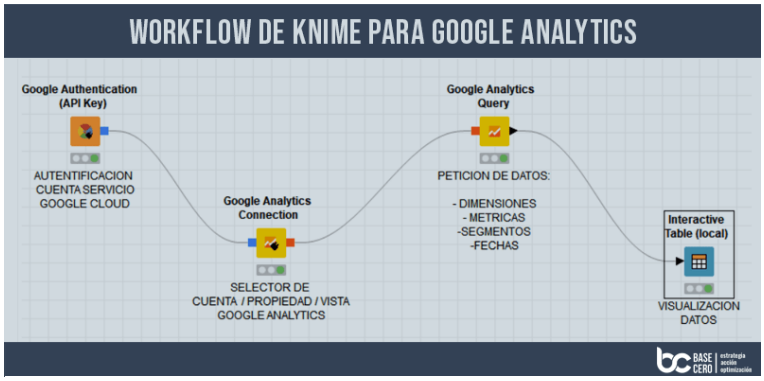
Step 4: Configure KNIME Analytics Platform and Query Data

We are now ready to connect to Google Analytics using KNIME Analytics Platform. To do that, we need to configure the tool to access the Google Cloud project and validate the permissions.

Querying Google Analytics to retrieve web traffic data from KNIME is very easy. We only need 3 nodes to:

1. Authenticate to the service account that we will use.
2. Access the Google Analytics account(s) and property(-ies) that have authorized access.
3. Query Google Analytics and retrieve dimensions, metrics, segments, etc.

The figure below shows an **example KNIME workflow** to connect to Google Analytics.

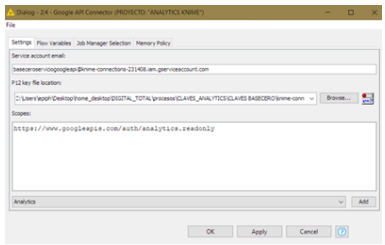


This example workflow connects to Google Analytics.

Next, we will provide a step-by-step description of the **node configurations in KNIME**.

1. **Google Authentication (API Key) node: Connecting to Google API.**

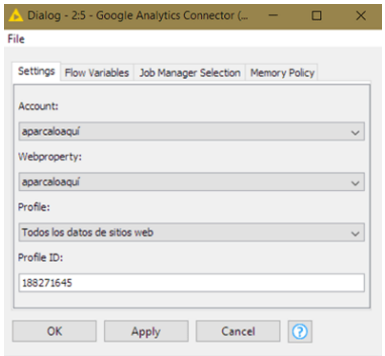
   We input the email address of the **service account**, the local directory of the security file with the **secret keys**, and define the **scope** (selecting "Google Analytics Connection (Read)" from the drop down menu). With these configurations, we can start a new session in the Google project and connect to Google Analytics API.



The configuration window of the Google Authentication (API Key) node.

2. **Google Analytics Connection node: Accessing Google Analytics.**

   Here we only need to select the **account**, **property and Google Analytics view** we would like to query. Only accounts and properties to which the Google Cloud project has access will be displayed.

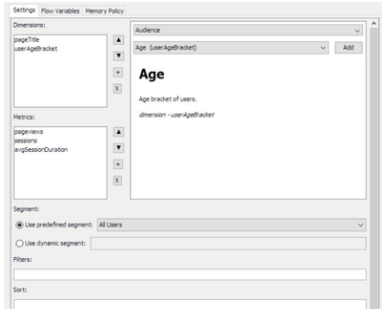


The configuration window of the Google Analytics Connection node.

3. **Google Analytics Query node: Querying Google Analytics API.**

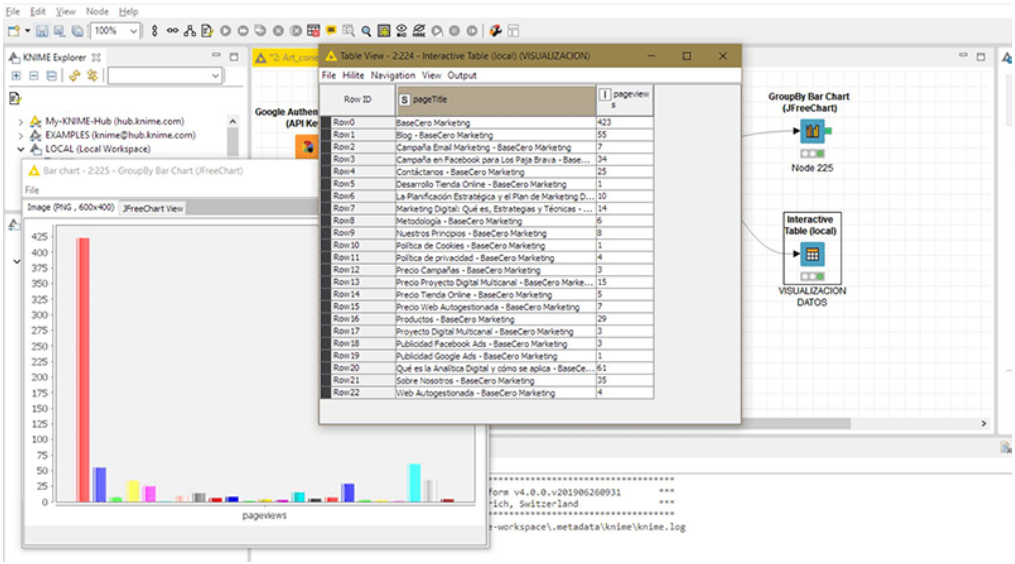
   With this node we can query Google Analytics API and obtain the data we need. The node has an **interface to build queries** and allows users to search and select available dimensions and metrics.

   To take full advantage of this node, it is advisable to know the **basic concepts of Google Analytics API**: dimensions, metrics, segments, dates, etc.



The configuration window of the Google Analytics Query node.

We are now **all set and ready** to start extracting insights from web traffic data, preprocessing, and working with it using KNIME, like any other data source in digital analytics:



After successfully connecting to Google Analytics API, insights from web traffic data can be extracted and then be further processed using KNIME.

Set clear Goals for your Web Analytics

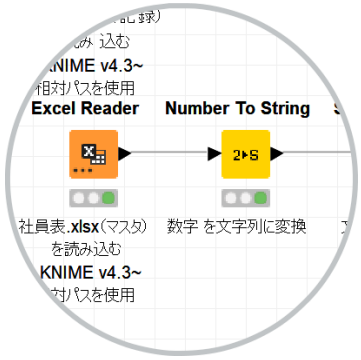
With this article we aimed at promoting the application of digital analytics, and free and open-source tools like KNIME Analytics Platform.

Whenever we want to analyze and make sense of web data, it is crucial to define a clear strategy and goals of the analysis. In other words, we need to understand **what aspects of web data we want to measure, why, in what time range, and for which users**.

The rest is *just* a matter of wrangling the data and extracting insights. That's where KNIME helps streamline processes, cut implementation time, and build powerful codeless analytics solutions.

This article was originally published in Spanish on BaseCero Marketing and was translated into English by Roberto Cadili. You can find the original version here.

makkynm was nominated KNIME Contributor of the Month for July 2021. He was awarded for his activities within the KNIME community and for his numerous blog posts and tutorials in Japanese be it for beginners, intermediate, or advanced KNIME users. On his blog https://degitalization.hatenablog.jp/ he writes about many topics including: String Manipulation, Data Accessing, Data Visualization, and Time Series Analysis. The content on his blog is written in Japanese, and it is aimed at the Japanese data science community. He is also sharing his knowledge actively on Twitter. Make sure to check out his Tweet history.



makkynm is an ardent KNIME advocate, and he is passionate about helping others, be it KNIME-related or not. He likes to stay anonymous, that's why he sent his Twitter avatar to represent him here.

Visit makkynm's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `makkynm`).

Reading Data from Databases in KNIME Analytics Platform

Using the Microsoft Access Connector Node as an Example

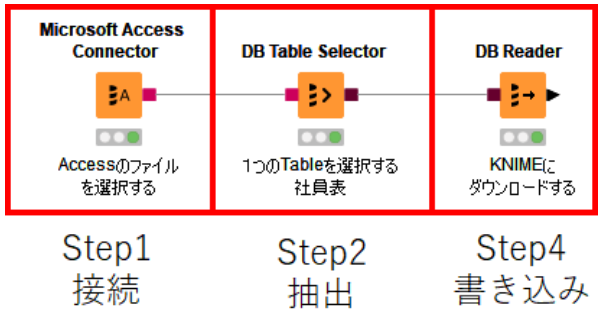
Author: makkynm; Translated by: Elisabeth Richter with the help of AI

Introduction

In this article, I will explain how to connect to and access data from a database in KNIME Analytics Platform. I will be using the database nodes that I showed you as an example the last time. If you do not know these database nodes, please read the previous article before you continue reading.

This article's focus lies on the Microsoft Access Connector node, DB Table Selector node, and DB Reader node.

For the purpose of this article, I will be connecting to Microsoft Access, which is probably the most familiar database, using the Microsoft Access Connector node. Connecting to a database in KNIME Analytics Platform is done via a dedicated connector node. In this example, it is the Microsoft Access Connector node, however, if you want to connect to and access data from a different database, the only thing you need to do is to replace the Microsoft Connector node with a different connector node, for example, a PostgreSQL Connector node.



This example workflow shows how to connect to and access data from a database in KNIME Analytics Platform. For this example, a connection to Microsoft is established.

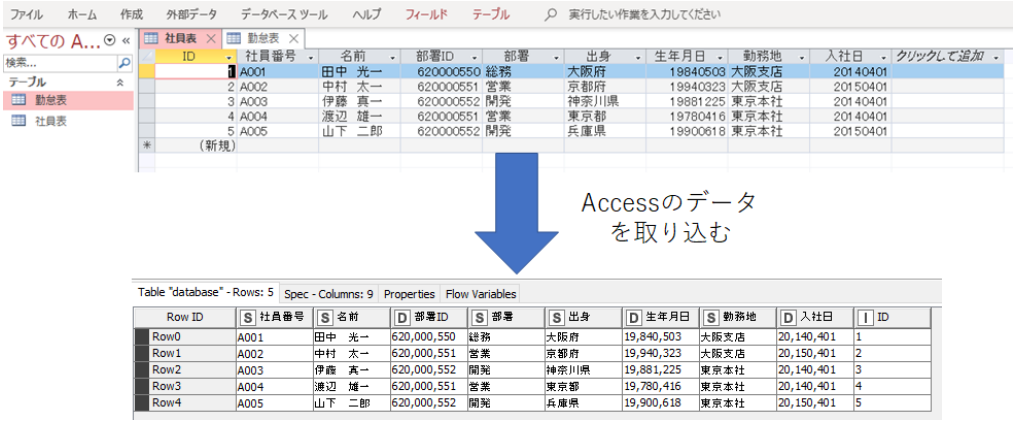
In the case of the Microsoft Access Connector node, there is no need to install and configure any additional drivers. However, for other databases, you will need to install drivers separately and make additional settings in the Preferences.

I will explain the three nodes in the following, however, none of them requires almost any settings.

You can download the workflow from my KNIME Community Hub space: https://kni.me/w/Wag-UjltQfjCTEn0

These are the steps I'm going to show you:

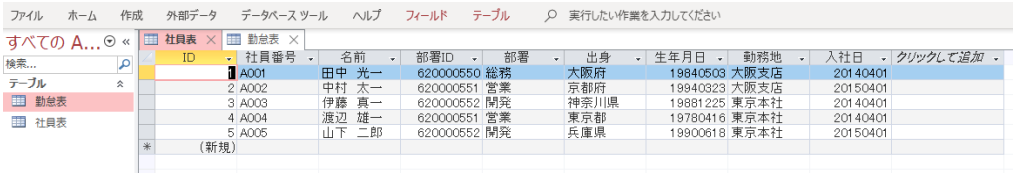
1. Connect to the database: Microsoft Access Connector node
2. Select the table from which you want to read the data: DB Table Selector node
3. Retrieve the data from the database into a KNIME data table: Access desired data: DB Reader node



Reading data that is stored in a database into KNIME Analytics Platform.

With the three nodes in the workflow shown above, you are able to read data that is stored in a database in KNIME Analytics Platform. This data can then be further used in KNIME.

In this example, I'm going to read the employee table and attendance table that I have been using in other articles before.



ID	社員番号	名前	部署ID	部署	出身	生年月日	勤務地	入社日	クリックして追加
1	A001	田中　光一	620000550	総務	大阪府	19840503	大阪支店	20140401	
2	A002	中村　太一	620000551	営業	京都府	19940323	大阪支店	20150401	
3	A003	伊藤　真一	620000552	開発	神奈川県	19881225	東京本社	20140401	
4	A004	渡辺　雄一	620000551	営業	東京都	19780416	東京本社	20140401	
5	A005	山下　二郎	620000552	開発	兵庫県	19900618	東京本社	20150401	
*	(新規)								

Employee table

ID	出勤日	社員番号	残業時間 h	クリックして追加
1	2020/03/31	A001	2	
2	2020/03/31	A002	3	
3	2020/03/31	A003	5	
4	2020/04/01	A001	1	
5	2020/04/01	A002	1	
6	2020/04/01	A003	4	
7	2020/04/05	A001	2	
8	2020/04/05	A002	3	
9	2020/04/05	A003	7	
10	2020/04/07	A001	3	
11	2020/04/07	A002	2	
12	2020/04/07	A003	4	
*	(新規)			

Attendance table.

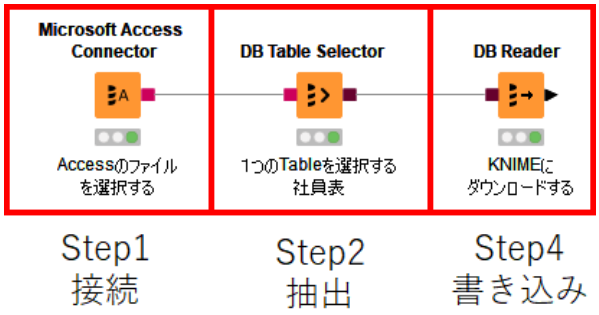
The Workflow

In the image below you can see the workflow. In the following, I will explain the three nodes one after the other. In the previous article I mentioned before, the workflow consisted of four steps (connection, extraction, conversion, writing). However, for this article I omitted Step 3 (conversion) and hence will continue without processing the data. Therefore, the three step are:

Step 1: Using the Microsoft Access Connector node establishes the database connection

Step 2: With the DB Table Selector node the desired table in the database can be selected

Step 4: Lastly, the DB Reader node accesses the data selected in the previous node and retrieves it into a KNIME data table

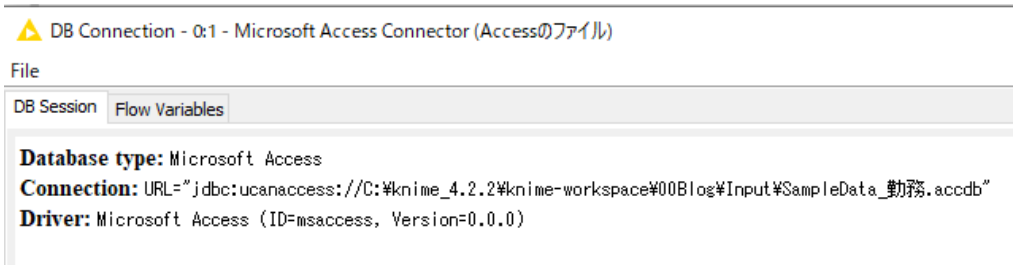


This example workflow shows how to connect to and access data from a database in KNIME Analytics Platform. For this example, a connection to Microsoft is established.

Step 1: How to use the Microsoft Access Connector node

This node is used to connect with a Microsoft Access database. It only creates the connection to the database but does not specify a table.

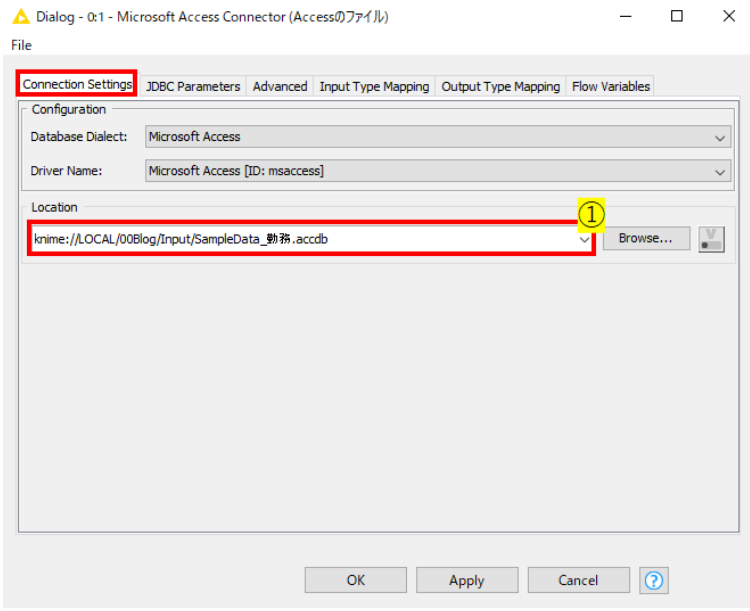
You can confirm that the database established a connection successfully by right clicking the connector node and selecting “DB Connection”. Here you can check the database type, the database URL (in this case this is the file path), and the driver in use.



Check the database connection in the “DB Session” tab of the output of the Microsoft Access Connector node.

To configure the node, the only thing you need to do is to select the file path of the database.

In the configuration dialog shown above, I am using a relative file path. If you would like to learn more about relative paths, please see this article about database nodes and KNIME as an ETL tool.



The configuration dialog of the Microsoft Access Connector node.

Step 2: How to use the DB Table Selector node

This node takes a DB Connection as input and allows you to select a table or view from within the connected database. You can specify one table in the database. In this example, I select the employee table.

You can view the currently selected table by right clicking the node and selecting “DB Data”. Then click “Cache no. of rows”. However, at this point, the table is not yet a KNIME data table.

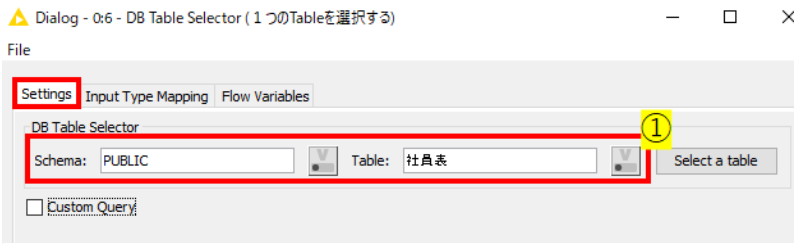
Table Preview | DB Spec - Columns: 9 | DB Query | DB Session | Flow Variables

Cache no. of rows: 100

Row ID	S 社員番号	S 名前	D 部署ID	S 部署	S 出身	D 生年月日	S 勤務地	D 入社日	I ID
Row0	A001	田中　光一	620,000,550	総務	大阪府	19,840,503	大阪支店	20,140,401	1
Row1	A002	中村　太一	620,000,551	営業	京都府	19,940,323	大阪支店	20,150,401	2
Row2	A003	伊藤　真一	620,000,552	開発	神奈川県	19,881,225	東京本社	20,140,401	3
Row3	A004	渡辺　雄一	620,000,551	営業	東京都	19,780,416	東京本社	20,140,401	4
Row4	A005	山下　二郎	620,000,552	開発	兵庫県	19,900,618	東京本社	20,150,401	5

View the table from within the connected database in the “Table Preview” tab of the output of the DB Table Selector node.

To configure this node, simply specify the table you want to import in the configuration dialog of the node. “Select a table” allows you to select a table from a list.



The configuration dialog of the DB Table Selector node.

Step 4: How to use the DB Reader node

This node executes the input query in the database and retrieves the result into a KNIME data table. As a result, the data is available at the output port of the node, similar to the one dealt with in the beginner and intermediate classes. In other words, now the data can be handled in the same way as, for example, when an Excel file is being read using the Excel Reader node.

If the amount of data in the database is large, it is recommended to first narrow down the data with one of the DB nodes (e.g., the DB Row Filter node) and then use the DB Reader node.

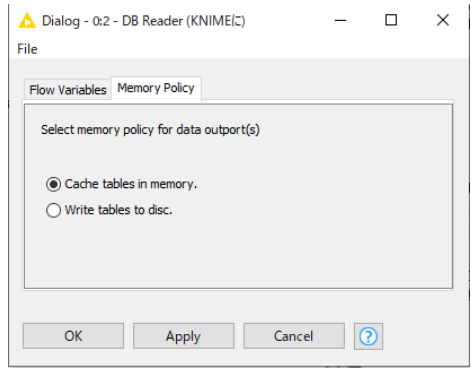
To access the output of the DB Reader node, right-click the node and select “KNIME data table”. It looks similar to the output of the DB Table Selector node shown before, however, as the data is now retrieved in a KNIME data table, the “Cache no. of rows” button is not shown anymore. In addition, the row ids (Row ID column) are assigned automatically.

Table "database" - Rows: 5 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	S 社員番号	S 名前	D 部署ID	S 部署	S 出身	D 生年月日	S 勤務地	D 入社日	I ID
Row0	A001	田中　光一	620,000,550	総務	大阪府	19,840,503	大阪支店	20,140,401	1
Row1	A002	中村　太一	620,000,551	営業	京都府	19,940,323	大阪支店	20,150,401	2
Row2	A003	伊藤　真一	620,000,552	開発	神奈川県	19,881,225	東京本社	20,140,401	3
Row3	A004	渡辺　雄一	620,000,551	営業	東京都	19,780,416	東京本社	20,140,401	4
Row4	A005	山下　二郎	620,000,552	開発	兵庫県	19,900,618	東京本社	20,150,401	5

The data is now accessible as a KNIME data table at the output port of the DB Reader node.

To configure this node, no special settings are required!



The configuration dialog of the DB Reader node.

A quick word

Microsoft Access Connector node - Other Options

In this example, only one step was required when configuring the Microsoft Access Connector node. However, other options exist and I would like to briefly explain them here.

Connection Settings: Configure the Driver Settings

Select a driver. This time the driver was automatically selected and no other drivers were shown, but if it is not the driver you want to use, you can select it here. If no driver is available, you need to set it in the Preferences (KNIME → Databases).

Reference link: KNIME Database Extension Guide

JDBC Parameters: Advanced Driver Settings

In the “JDBC Parameters” tab in the configuration window you can set JDBC driver parameters. For example, if a password is required, specify it here. Also, when using a database with a large data size, it is recommended to set the "memory" item to "false".

Reference link: UCanAccess-A Pure Java JDBC Driver for Access

Advanced: Advanced Connection Settings

In the “Advanced” tab in the configuration window, more detailed connection settings can be specified, for example, timeout settings, restore database option, etc.

Input Type Mapping / Output Type Mapping: Data Type Mapping

The “Input Type Mapping” tab in the configuration window allows you to define rules to map from database types to KNIME types and the “Output Type Mapping” tab to define rules from KNIME types to database types respectively.

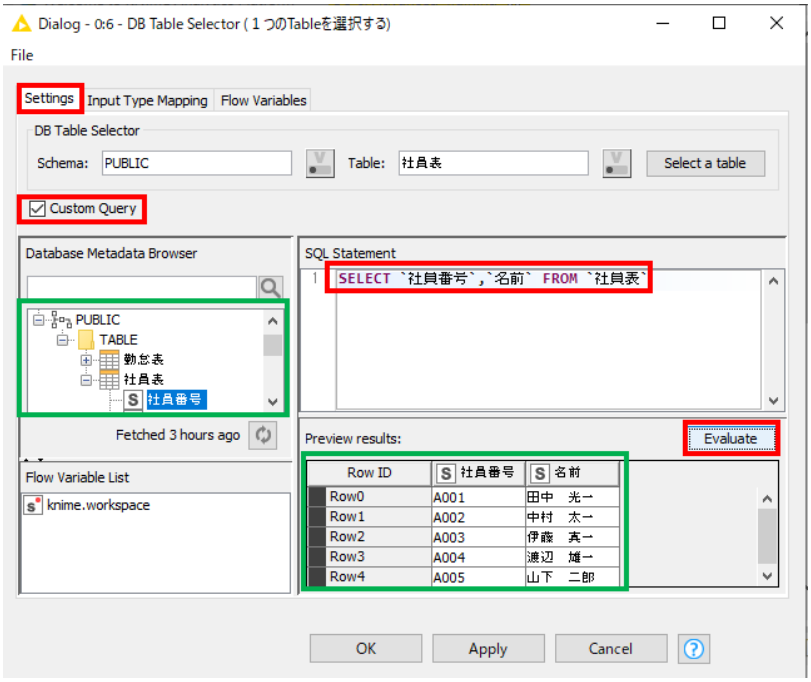
DB Table Selector - Other Options

Custom Query: Select Table Columns with SQL Statements

Checking the “Custom Query” box in the “Settings” tab of the configuration window enables or disables custom query for the selected table. In addition to tables, you can also specify columns.

With the “Database Metadata Browser” on the left, you can browse the table fields and specify the SQL statement.

When pressing “Evaluate”, you can preview your SQL statement and by that you can check whether the SQL statement is correct.



The “Settings” tab in the configuration dialog of the DB Table Selector node.

Conclusion

In this article, I have used the Microsoft Access Connector node as an example to explain how to connect to a database in KNIME Analytics Platform. If you want to connect to a different type of database, you can simply change the connector node.

While you can perform the operations I showed you above in a database directly, using KNIME can not only do the same but also has the advantage of being more transparent. Often, databases are black-boxes and the logic behind certain database operations is not always clear. Whereas KNIME Analytics Platform allows you to visualize the logic of certain operations.

For everyone that works with such a “black-box access tool” I would highly recommend giving KNIME Analytics Platform a try! See you next time!

Reference links

- NodePit:

  Microsoft Access Connector — NodePit

  DB Table Selector — NodePit

DB Reader — NodePit

- KNIME Example Workflows:

  Microsoft Access Connector – KNIME Community Hub

  DB Table Selector – KNIME Community Hub

  DB Reader – KNIME Community Hub

This article was originally published in Japanese on makkynm's blog and was translated into English by KNIME. You can find the original version here.

Ignacio Perez was nominated KNIME Contributor of the Month for September 2021. He was awarded for his work revolving around the Spanish-speaking KNIME community and for being a respected reference. He established and currently overviews the Spanish questions on the KNIME Forum, regularly hosts courses in Spanish, and he also translated the e-book "From Excel to KNIME" into Spanish: "de Excel a KNIME Analytics Platform". Furthermore, he maintains a YouTube channel where he explains KNIME to the Spanish-speaking community.

Ignacio holds a PhD in Applied Mathematics from the University of Lyon. He has over 20 years of experience as a Data Analyst working in different industries and is using KNIME for over 10 years now. Also, he is the founder and owner of IQuartil, a firm established in 2000 and dedicated to the development of consulting projects in statistical analysis, sampling, forecasting, analytical marketing, operations optimization, risk management and financial planning analysis.

Visit Ignacio's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `iperez`).

The pioneer of the KNIME Community en español

Author: Rosaria Silipo

¡Hola data-nautas de la ciencia de datos sin código!

Ignacio is the pioneer of the KNIME community en español. On Dec 2020, he started the KNIME Forum en Español, where he answers most of the questions en español; in 2021, together with his colleagues, he translated the booklet "From Excel to KNIME" en español; and up to now he gives regular presentations about how he successfully applied KNIME in English and en español.

The figure below shows the first post in the KNIME Forum en Español. Notice the brevity of the announcement, something that is a distinct pattern of all Ignacio's posts. It is not rare to find one of his resolving answers with just a few words and a workflow attached on the KNIME Forum. Let's say that he prefers to speak with workflows rather than with words.

The first post in the KNIME Forum en Español.

We would like to report here an interview conducted by Cynthia (the KNIME interviewer) to Arturo (the customer), and Ignacio (the technical enabler) as an example of the successful support that a consultant company can provide to meet the client's needs. This article is based on the presentation "*Real-Time Information on Product Quality*" from KNIME Fall Summit 2020, now available on YouTube.

Cynthia: *We have Ignacio Perez from iQuartil and Arturo Boquin from Dinant and we're going to talk about real-time information on product quality. Ignacio, tell us a little bit about iQuartil please.*

Ignacio: iQuartil is a KNIME-Trusted partner company based in Colombia that provides analytics services in Latin America. We accompany our clients in different sectors both in developing and productionizing analytical solutions as well as in training.

One of our clients is Dinant. Dinant is a Central American company from Honduras that operates in the consumer goods industry manufacturing and commercializing different types of products, such as packaged foods, and home and personal care products. Dinant's brands are well known in Central America, in the United States and in the Caribbeans. Arturo is responsible for a project we developed at Dinant. We've worked with them to put in place a solution that Arturo is going to share with us.

Arturo: I'm very glad to present to you the solution that we have developed with KNIME Server. First, I would like to show you the before and after of the process that we changed.

Before:

1. We had a member of the quality control department who stood in line (5-10 min) waiting to use a printer just to print many physical formats on paper.
2. Then, someone from quality control filled the paper files with data. (11 min)
3. Generally, they waited to accumulate many papers before passing them on to someone who would digitize the data in Excel. (2 hours)
4. Many days later someone in the same team would create and deliver a report just to see that something went wrong **in the past**. (50 min)

In theory, the total process should take approx. 3 hours and 15 minutes, but we usually encounter several obstacles that delay each part of the process.

Process before introducing KNIME Server. Completion time (in theory): approx. 3 hours and 15 minutes.

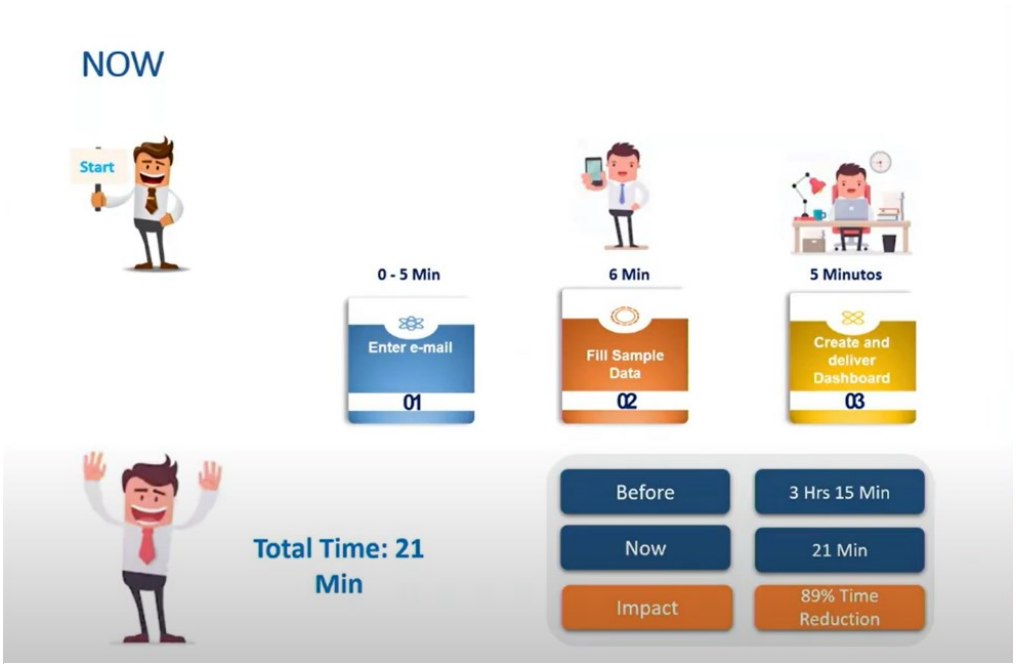
After:

The process begins with a member of quality control.

1. Now the person has a tablet and just has to enter the email and fill the files with data.
2. Then, we have an automated process that updates the dashboard for reporting every five minutes.
3. Thanks to that, people in the factory can see results almost in real time and make decisions in order to fix any negative variation on product quality. We are not seeing only the past anymore.

Now the process takes 21 minutes. Which means 89% time reduction and a lot of money saved in production cost. From a business perspective, that's the most valuable part of the project.

Are you wondering how we did it? The answer is: KNIME Server.



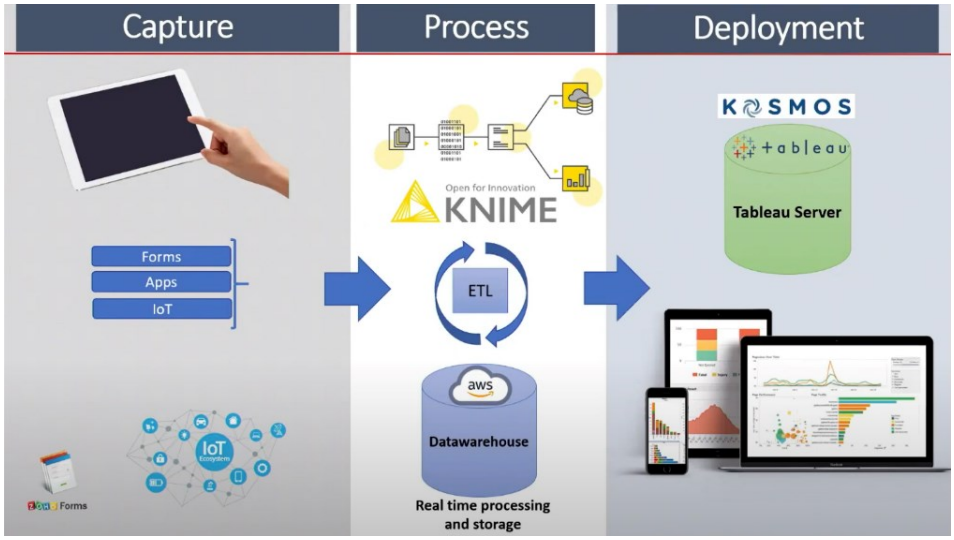
Process after introducing KNIME Server. Completion time: 21 minutes.

Implementation with KNIME Server

We look holistically at the process.

1. ***Capture*.** First, we have the capture stage. Here, we gather data from many different sources, be it Microsoft Forms, apps, or IoT systems. The important part is that we can connect all these different sources effortlessly to KNIME Server.

2. ***Process.*** Once the data is available on KNIME Server, we can perform ETL operations, move ML-driven applications to production, and store the results in a data warehouse in real time.
3. ***Deployment***. Finally, we can write a TDE file directly to a Tableau server, and see on any screen in any part of the factory the dashboard with the relevant information to make better decisions.



The 3-step process powered by KNIME Server to obtain real-time information on product quality.

Cynthia: *It's amazing to see how much time (and money) you saved with this solution. While the final outcome is great, I'm sure that there was quite some work behind this radical process change. What were the challenges with this project? Did you have to convince the management of the company that process automation and real-time reporting benefits not only the business but also the operators in their quality control and decision-making process?*

Arturo: Convincing management to change processes is usually very difficult, and often an obstacle in some of these projects. In my experience, the best strategy is to present a prototype solution in action so people can really see what it is about. For example, you can create an MVP of the whole project where you identify potential people that will be on board and one process that could be modernized via a smartphone or tablet. Next, you can create a simple solution where you blend data from Microsoft Forms and SharePoint with KNIME, and use wrangled data to build a dashboard with KNIME or some other tool KNIME has integrations for (e.g., Tableau, Power BI, etc.). In this way, it's much easier to get your idea across, and make the non-analytical people in the organization see the value and benefits of your proposal.

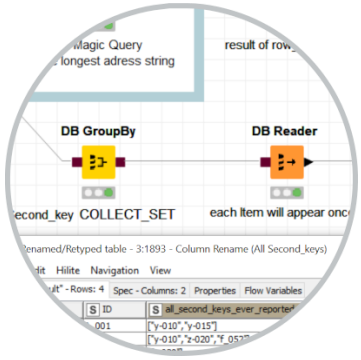
Cynthia: *Excellent, thank you! Ignacio, would you like to add something?*

Ignacio: I would like to point out that it was a nice experience working with Arturo and his team. They were already heavy and advanced users of KNIME Analytics Platform, and in just a couple of weeks (literally!) they were able to develop this solution, capitalize on the powerful features of KNIME Server very fast, and present to high-level managers the benefits of working with KNIME. They did a great job!

Cynthia: *Thanks, Ignacio and Arturo! We really appreciate you sharing the story. Again another great use case on manufacturing and how KNIME can provide the tools to the operators right there where they work so they can make better decisions quickly.*

Markus Lauber was nominated KNIME Contributor of the Month for September 2020. He was awarded for his KNIME Forum thread *School of duplicates – and how to deal with them* where he explains different ways of how to deal with duplicate values and how to eliminate them without losing as little potentially relevant information as possible. The image on the right shows a snippet of the workflow belonging to this thread. Markus is not only a regular on the Forum but also maintains an active space on the KNIME Community Hub where he shares many of his workflows, and he is also active as a speaker at our events. So all in all, he has been a highly active and trusted member of the KNIME community for many years already. Besides KNIME he likes to work with R, Python, Spark and H20.ai and integrates them with the KNIME software, particularly KNIME Server.

Markus has over 20 years of experience as Analyst with the focus on Data Mining and Big Data. He is currently a Senior Data Scientist / Big Data Analytics at Deutsche Telekom. His main tasks include to improve the performance and reliability of landlines and mobile networks, using advanced tools on Big Data platforms.



Visit Markus' space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `mlauber71`).

From H2O.ai AutoML to Violin Plots

A Best-of Collection of KNIME Forum Threads by Markus Lauber

Author: Elisabeth Richter

With over 1500 days visited, roughly 11400 topics viewed, and 203 given solutions as of 11.08.2022, Markus is indeed a very active member on the KNIME Forum. We have been following Markus for years by now and we have benefitted from many of his final answer comments on the KNIME Forum. His comments span all phases of the data science creation cycle: data cleaning and data preparation, data visualization, machine learning for prediction and classification, till integration with other tools like R and H2O. In this article we have followed his journey through the various steps in the data science creation cycle, by picking out three of Markus' most popular KNIME Forum threads. You can also find them linked from the Knowledge Sharing category on our KNIME Forum. In the following, we summarize these Best-Of-Markus threads.

mlauber71

mlauber71

Regular

Featured Topic H2O.ai AutoML in KNIME for classification problems

twitter.com/mlauber71

Joined Feb 27, '18 Last Post 2 days Seen 3 hours Views 18936 Trust Level regular

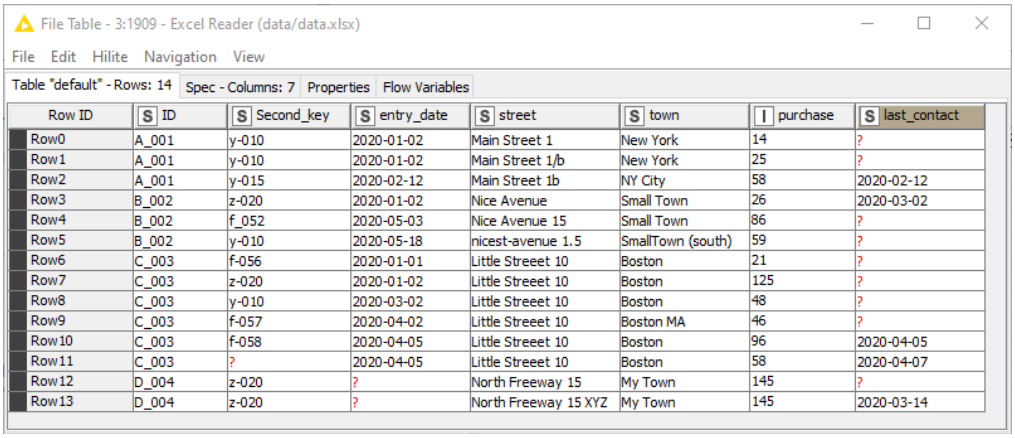
School of Duplicates - And How to Deal with Them

This is a very helpful KNIME Forum thread in which Markus shares his best practices on a topic constantly present among data scientists: dealing with duplicates. When dealing with duplicates, one must be certain about their goals. Although the easiest way of handling duplicates would be just getting rid of them, this might not be the best approach as potentially relevant information might get lost.

However, being able to bring messy data into a meaningful table with a unique ID and without losing too much information is a valuable skill everyone should have.

This forum thread as well as the workflow attached to it, aims to potentially encourage others to think about what to do with their duplicates and to actively take control over their duplicate values.

The following figure shows the data we are dealing with. The data table contains two ID columns, *ID* and *Second_key*, with both key columns containing duplicate values, and additional information like the address (*street* and *town*), the purchase amount (*purchase*), as well as the purchase date (*entry_date*) and the date of last contact (*last_contact*). The duplicates all carry potentially meaningful information and the goal is to reduce the data table into fewer lines without losing any (or as little as possible) information.


File Table - 3:1909 - Excel Reader (data/data.xlsx)

File Edit Hilite Navigation View

Table "default" - Rows: 14 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	ID	Second_key	entry_date	street	town	purchase	last_contact
Row0	A_001	y-010	2020-01-02	Main Street 1	New York	14	?
Row1	A_001	y-010	2020-01-02	Main Street 1/b	New York	25	?
Row2	A_001	y-015	2020-02-12	Main Street 1b	NY City	58	2020-02-12
Row3	B_002	z-020	2020-01-02	Nice Avenue	Small Town	26	2020-03-02
Row4	B_002	f_052	2020-05-03	Nice Avenue 15	Small Town	86	?
Row5	B_002	y-010	2020-05-18	nicest-avenue 1.5	SmallTown (south)	59	?
Row6	C_003	f-056	2020-01-01	Little Streeet 10	Boston	21	?
Row7	C_003	z-020	2020-01-02	Little Streeet 10	Boston	125	?
Row8	C_003	y-010	2020-03-02	Little Streeet 10	Boston	48	?
Row9	C_003	f-057	2020-04-02	Little Streeet 10	Boston MA	46	?
Row10	C_003	f-058	2020-04-05	Little Streeet 10	Boston	96	2020-04-05
Row11	C_003	?	2020-04-05	Little Streeet 10	Boston	58	2020-04-07
Row12	D_004	z-020	?	North Freeway 15	My Town	145	?
Row13	D_004	z-020	?	North Freeway 15 XYZ	My Town	145	2020-03-14

A data table with two ID columns, ID and Second_key, both containing duplicates, and additional information like the address (street and town), the purchase amount (purchase), as well as the purchase date (entry_date) and the date of last contact (last_contact).

Now, in this forum thread Markus describes the following approaches on how to deal with duplicates:

Option 1: Grouping by Single IDs - The Simplest Form of Duplicate Removal

Using a *GroupBy* node is the simplest and quickest way to remove duplicates from your data. Group by *ID* and aggregate with several functions, for example, *sum(purchase)*, *mean(purchase)*, *max(entry_date)*, and *max(last_contact)*.

Although this type of duplicate removal is fast and easy, a lot of potentially important information is discarded.

Option 2: Using KNIME's Duplicate Remover - A Slightly more Sophisticated Approach

Using the Duplicate Row Filter node, one of KNIME's more advanced filter nodes, allows a slightly more sophisticated duplicate handling. The node identifies duplicate rows based on one or more columns and offers the possibility to either keep them, marked as "duplicate", or to remove them.

In the forum thread, Markus also shows the equivalent operation using SQL.

This method is indeed a more sophisticated way of handling duplicates. However, the second key column (*Scond_key*) is still ignored.

Option 3: Taking it up a Notch - Considering the Second Key

Lastly, Markus shows an option where the second key is also taken into consideration. He describes different ways how *Second_key* is kept.

He also demonstrates another approach using SQL.

Read the whole thread "School of duplicates - and how to deal with them" on the KNIME Forum and download the workflow of the same name from the KNIME Community Hub.

KNIME and R ggplot2 – „the beautiful Violin Plot that has it all"

In this forum thread Markus gives us an introduction to ggplot2, an R visualization package, with KNIME Analytics Platform. With the help of the *R-Conditional-Violinplot* component and the *R Scripting extension*, KNIME Analytics Platform allows to create R-based violin plots with lots of additional statistics in one chart - even if you are not familiar with R code.

Violin plots are very effective to show the structure of numeric variables and to compare them across different groups. The width represents the number of cases that have the values on the y-axis. It is widely used, for example, as a population pyramid.

The data we are dealing with in this use case shows the CPU usage (in %) in a time interval of one month for two different servers (server1 and server2) (see figure below).

Read table - 4:3 - Table Reader (Read Data)

File Edit Hilite Navigation View

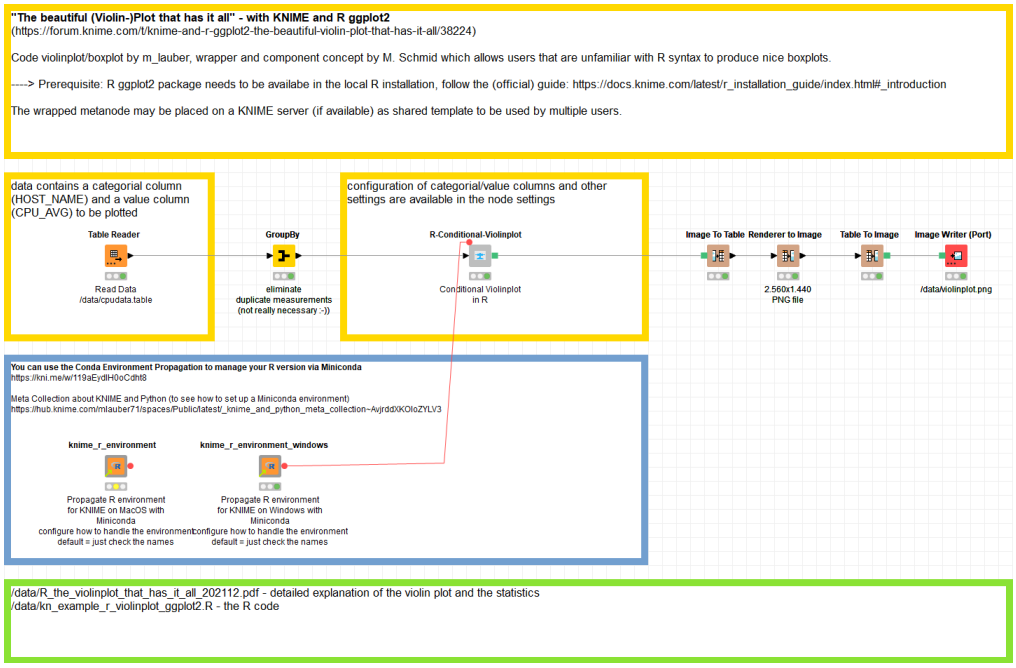
Table "default" - Rows: 2880 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	STARTINTERVALL	HOST_NAME	CPU_AVG
Row0	01.09.2016 00:00	server1	42.417
Row1	01.09.2016 00:00	server1	42.417
Row2	01.09.2016 01:00	server1	38.375
Row3	01.09.2016 01:00	server1	38.375
Row4	01.09.2016 02:00	server1	40.021
Row5	01.09.2016 02:00	server1	40.021
Row6	01.09.2016 03:00	server1	39.521
Row7	01.09.2016 03:00	server1	39.521
Row8	01.09.2016 04:00	server1	39.646
Row9	01.09.2016 04:00	server1	39.646
Row10	01.09.2016 05:00	server1	48.271
Row11	01.09.2016 05:00	server1	48.271
Row12	01.09.2016 06:00	server1	39.333
Row13	01.09.2016 06:00	server1	39.333
Row14	01.09.2016 07:00	server1	38.458

The input data: CPU usage (in %) in a time interval of one month for two different servers (server1 and server2).

By plotting this data in a violin plot, we are able to compare the CPU usage between the two servers by simply comparing the shapes. This allows us to observe which one of the two servers was more heavily used during the period of interest.

In the figure below the entire workflow for creating the violin plot using an R script and saving it as a *.png* file to the workflow data area is shown.

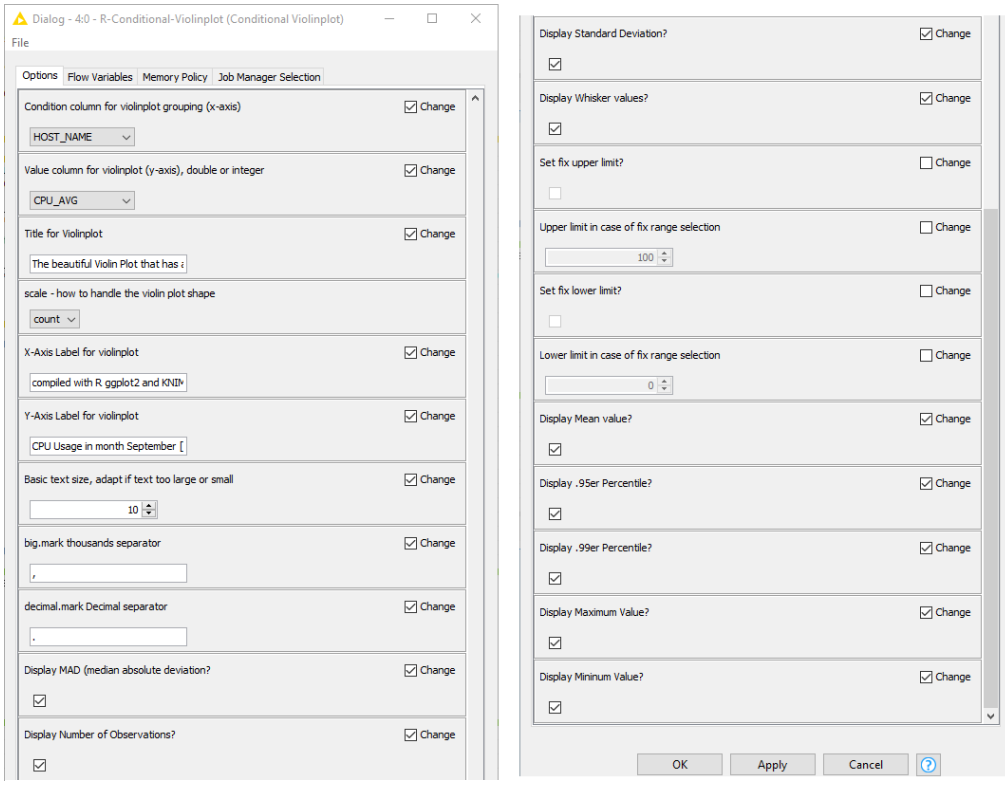


This workflow creates a violin plot and saves it as a .png file.

The *knime_r_envionment_windows* component ensures that a specific configurable Conda environment exists and propagates the environment to downstream nodes. This is done with the *Conda Environment Propagation* node inside the component.

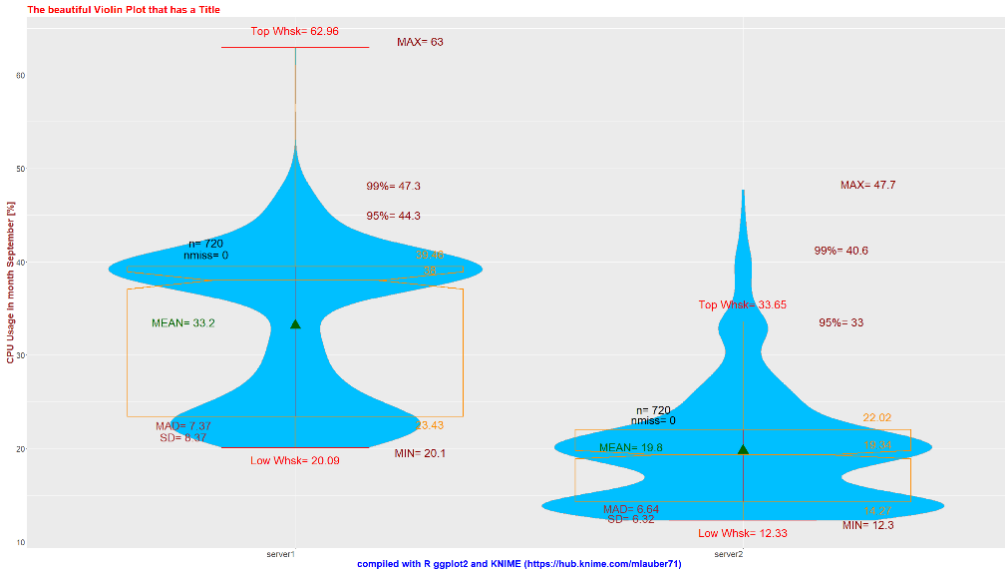
The violin plot is created within the *R-Conditional-Violoinplot* component. It takes as input the data table and outputs a .png image of the violin plot. Also, the propagated Conda environment is passed to the downstream nodes. The component has a configuration window in which you can define the *condition column* and the *value column* for the violinplot, the *plot title* as well as the *axis labeling*, and many other settings. One basic configuration would be how you want the violin's shape to handle the number of items. Markus' standard is to have the shapes proportional to the number of cases "count". Using this, the areas are scaled proportionally to the number of observations. However, alternative options could be "area", where all violins have the same area, or "width", where all violins have the same maximum width.

In order to actually create the violin plot, the *R View (Table)* node inside the R-Conditional-Violinplot component is used. In the configuration window of the node you can add your R script. The output of the *R View (Table)* node is a *.png* image. R advocates watch out: at the end of his forum thread, Markus gives a few hints regarding the R code, for example, how to make sure that labels are not overlapping any information.



The configuration window of the R-Conditional-Violinplot component.

The resulting violin plot is shown below. From this plot it is visible that server1 was more heavily used during the considered time period compared to server2. This is



A violin plot showing the CPU usage (in %) in time intervals over one month (y-axis) compared to two servers (x-axis).

because there are more observations of server1 with higher CPU usage than for server2. Additional information can be retrieved, for example, the mean CPU usage for each server, the minimum and the maximum values, or the number of missing values per server.

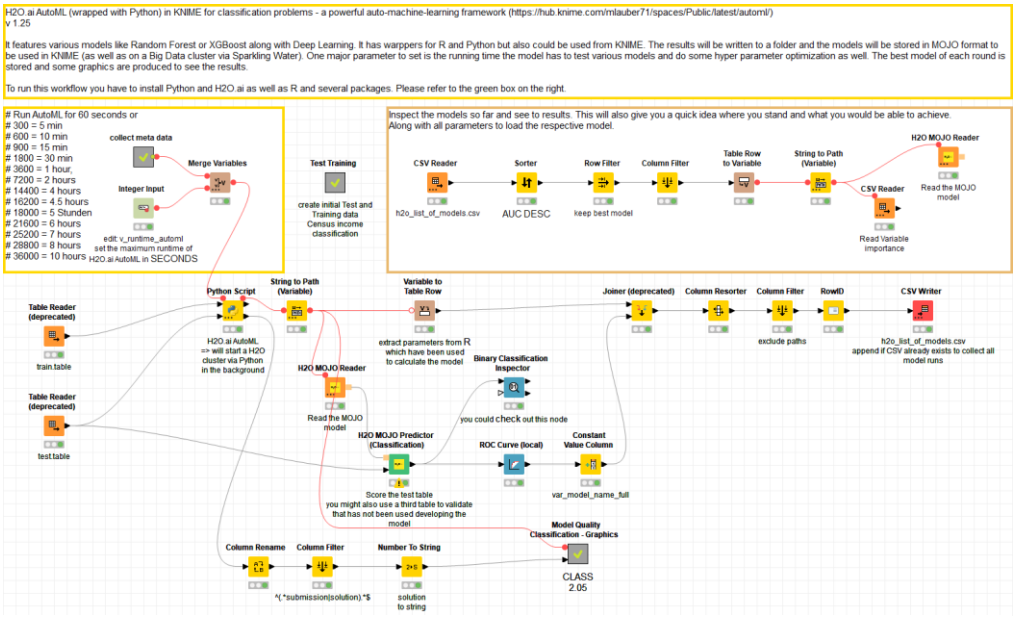
Read the whole thread "KNIME and R ggplot2 – the beautiful Violin Plot that has it all" on the KNIME Forum and download the workflow of the same name from the KNIME Community Hub.

H2O.ai AutoML in KNIME for classification problems

With 48 likes, this is Markus' most liked thread on the KNIME Forum. In this thread he takes up the topic of H2O.ai AutoML, a powerful auto-machine-learning framework, and provides a documentation on how to use H2O.ai AutoML in KNIME Analytics Platform.

AutoML (= Automatic Machine Learning) is a way of automating training and evaluating machine learning models. This means, AutoML automates algorithm selection, feature generation, hyperparameter tuning, iterative modeling, and model assessment. H2O.ai AutoML can be used in interfaces like R and Python, but it can also be used in KNIME Analytics Platform with the H2O integration. It features various models like Random Forest or XGBoost along with Deep Learning. The result of the AutoML is written to a folder and the models are stored in MOJO format, ready to be reused in different KNIME workflows.

In this forum thread as well as the attached workflows, Markus uses Python script to apply H2O AutoML and the H2O dedicated KNIME nodes to read a trained model in MOJO format and to apply it to new data. Along with it come various statistics and model characteristics required for interpreting the model. These are saved by the Python Script node to an *.xlsx* file stored in the workflow subfolder under *model/validate/H2O_AutoML_Classification _20210206_1730h.xlsx*. Further sheets are added to the file with the Excel Writer node in the *Model Quality Classification - Graphics* metanode. In the following, Markus' workflow is displayed.



This workflow performs AutoML using some census-income data for training and validation. The workflow reads the trained model in MOJO format and applies it to new data to predict the target value. Various statistics and model parameters are saved in an .xlsx file saved in the workflow subfolder under model/validate/H2O_AutoML_Classification_20210206_1730h.xlsx.

Most of the magic happens within the Python Script node. The underlying data of the workflow is the popular *census-income dataset*, partitioned into training data (*train.table*) and validation data (*test.table*). The two data tables are inputted to the Python Script node, together with a value for the run time. The run time is in fact one major parameter to be set and is the time required to test the different models as well as optimizing some hyperparameters. By default, Markus set the run time to 30 seconds, however, longer time intervals can be chosen.

Inside the Python Script node, a Python script is executed that performs the AutoML. Eventually, the best performing model is saved as a generic H2O model as well as in MOJO format. With the H2O MOJO Reader node, the trained model in MOJO format is read. This is then connected with the H2O MOJO Predictor node to apply the trained model to new data and predict the target values.

Now, the main purpose of this forum thread is to explain how the statistics and model characteristics stored in the accompanying .xlsx file can be used to interpret the model. Roughly, Markus divides this procedure in two parts: 1) inspecting model graphics, and 2) investigating the *.xlsx* file.

Option 1: Inspect Model Graphics

In the Model Quality Classification - Graphics metanode, Markus demonstrates how to plot several graphics to inspect the performance of binary classification models. All the plots are saved as *.png* files to the workflow subfolder model/validate with the

Image Writer (Port) node. The metrics explained are ROC Curve, TOP Decile Lift, Kolmogorov-Smirnov Goodness-of-Fit Test, and a plot of the Normalized Gini Coefficient to determine the best cutoff point.

Option 2: Investigating the Accompanying Excel File

Indeed, studying the graphics gives a first idea of how the model performs but for deeper insights Markus recommends further investigating the *.xlsx* file. This file consists of multiple sheets containing various model- and training-related information, including:

- **The leaderboard from the set of models run.** This gives you an idea what types of models were considered and how the values of the AUC are distributed between different model types.
- **The model summary.** The model summary gives you the parameters used which is helpful if you want to further tweak your model. In addition, the whole model print is saved to a *.txt* file in the same workflow subfolder and includes further information of all parameters.
- **The variable importance list.** This list collects the importance of each variable and should be studied carefully. If one variable captures all the importance you might have a leak. It also indicates which variables might be cut off.
- **The model overview in bins and numbers.** This table gives us an idea what a cutoff at a certain score (i.e., "submission") would mean. Finding a cutoff point is highly dependent on your business case. For example, you choose a cutoff at 0.8 which would give you 92% precision and 43% of all your desired targets. In marketing that would be an excellent result but in credit scoring you might not want to live with 8% of people not paying back their loans.
- **Looking at cross-validation.** In general, H2O does a lot of cross-validation by default in order to avoid overfitting. However, you might want to do some checks of your own. The basic idea is, if your model is really catching a general trend and has good rules they should work on all (random) sub-populations and you would expect the model to be quite stable. Hence, we're looking at the combined standard deviation. A value of 0 would represent a perfect match between all subgroups. So, if you have to choose between several good models you might want to consider the model with the least deviation.

One last word to all the Python advocates out there: the full Python script is available in the subfolder of the workflow under *script/ kn_automl_h2o_classification_python.ipynb*.

Read the whole thread "H2O.ai AutoML in KNIME for classification problems" on the KNIME Forum and download the workflow of the same name from the KNIME Community Hub.

Learning from the Best on the KNIME Forum

In this article, we've summarized three of Markus' best Forum threads. These are among the posts that are most liked, most viewed, or most helpful. This story about the Best-of-Markus threads is yet another example of how great community support can be. Thanks to Markus' broad (KNIME-) wisdom and his excitement of teaching things he has definitely helped out one or the other community member. That's the perks of such a diverse community. I'm sure that you've learned a thing or two even from this article or might have even gained a new perspective on certain topics - irrespective of whether you're a new or expert KNIME user. Whatever your motivation is to contribute to the KNIME Forum, we're in any case happy about your engagement.

Bruno Ng was nominated KNIME Contributor of the Month for May 2022. He was awarded for his tireless activity on the KNIME Forum and his countless contributions to the KNIME Community Hub in his public space. He is a prolific member of the KNME Forum and the KNIME Community Hub. He is well-known for his valuable answers on the Forum, as well as the many valuable workflows and components he has contributed to the KNIME Community Hub. Bruno always engages other KNIMErs with the utmost courtesy and clarity. As of this writing, he has created 2.2k posts, received 3.5k favorites, and provided almost 250 solutions.

Bruno is an accomplished bilingual agile team manager with more than 20 years of experience in application development. He possesses extensive experience and knowledge of all stages of the software development cycle, database, and software architecture design. He is good at quickly spotting patterns and foreseeing issues and preventing them from occurring before others realize there is an issue. Bruno is currently Director Data Ops at Triton Digital.

Visit Bruno's space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `bruno29a`).

Supporting the Community 24/7 like a Champ

Bruno Ng's Best KNIME Forum Answers

Author: Elisabeth Richter

Bruno Ng is indeed one of our largest KNIME Forum contributors. He joined the forum in November 2020 and as of 24.08.2022, he counts 535 days visited, roughly 4100 topics viewed, and 258 solutions given. This makes him one of the top KNIME community members that provide the most solutions to questions and problems. True to the motto "help me help you", Bruno is always very eager to help other KNIME users on the Forum. In the following, we'd like to present Bruno's five most liked replies which might be of help with your problems as well.

bruno29a

Bruno29a

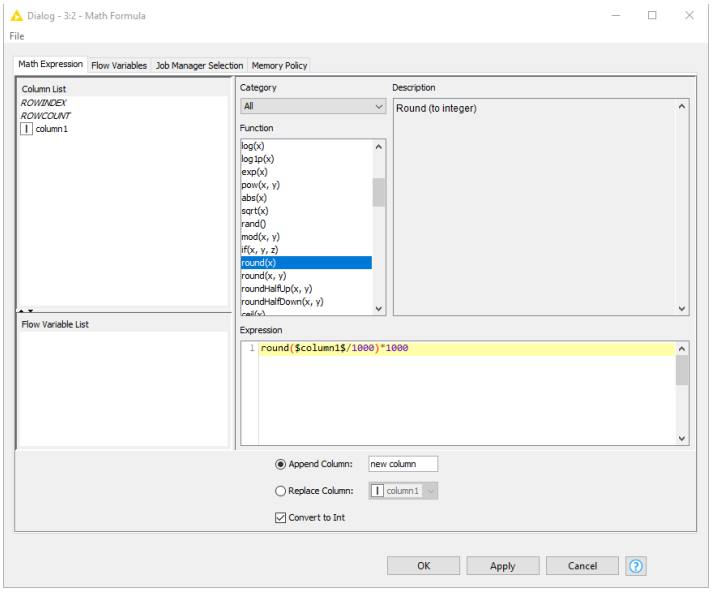
Regular

Joined Nov 18, '20 Last Post 5 hours Seen 5 hours Views 2258 Trust Level regular

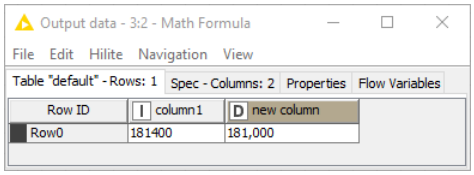
Rounding Up Numbers

Rounding numbers is a task that for sure counts to the everyday things for people in all kinds of professions and it's definitely not witchcraft. How to round numbers to the nearest ten, hundred, or thousand was probably taught to most of us in the early stages of our education. But do you know how to round, for example, to the nearest thousand using KNIME Analytics Platform? A good starting point is the Math Formula node but what happens after? The standard round(x) function is not sufficient in this case as it rounds only decimals to integers. Let's have a look at this smart solution proposed by Bruno:

> *Let's assume your nearest value is defined by x (so x is either 1,000 or 10,000 depending on which one you meant). Just do a division by x, round the result, and multiply by x. So the formula would be:*
>
> `round(your_value/x) * x`
>
> *For example, if I round to the nearest 1,000:*
>
> `round($column1$/1000) * 1000`



Results:



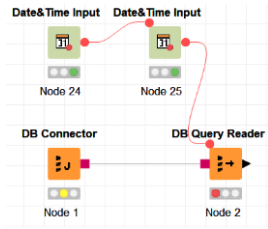
Et voilà - quick and easy solution. Just replace x with whatever value you want to round to. So, if you want to round to the nearest thousand, define x=1000, for the nearest hundred x=100, and so on.

Find the original thread "How do I round a whole number to the nearest 1,000?" on the KNIME Forum.

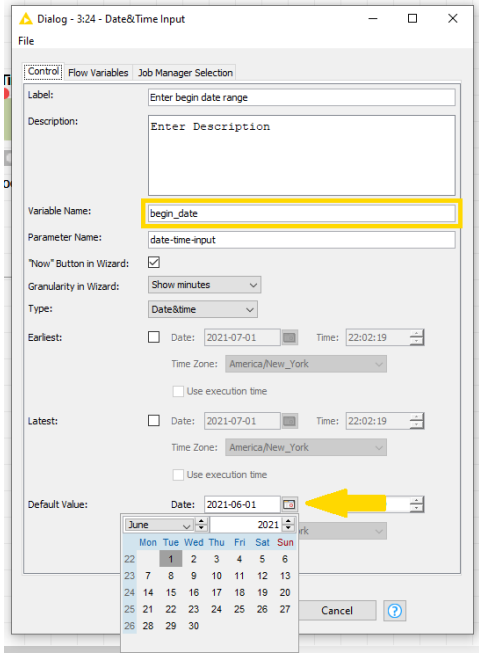
Inserting Dates as Parameters in SQL Queries

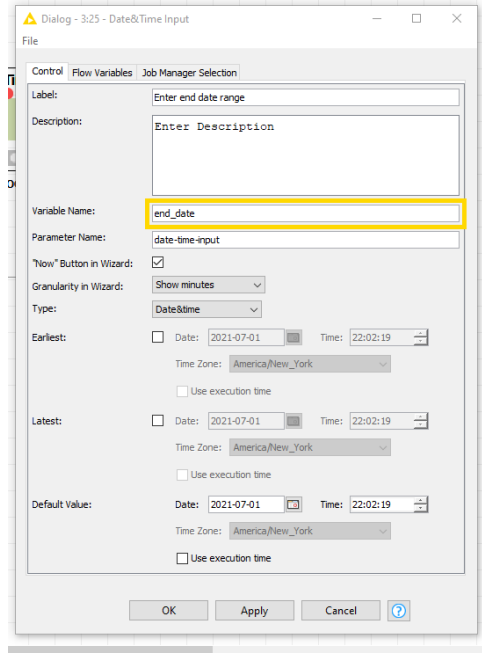
Dealing with SQL might be challenging, especially when being a newbie which - hand on heart - each of us once was. Luckily, KNIME Analytics Platform offers a bunch of nodes that allow connecting to JDBC-compliant databases. This might make things easier but nonetheless one or the other thing remains unclear to some users. On the KNIME Forum a user once asked how to have multiple date inputs into the DB Query Reader node. Luckily, Bruno was on hand and answered this question in great detail.

If you really want real date input as interaction, then you can use whatever you were doing originally, but you can just use 2 Date&Time Input nodes. You just need to connect them to each other and use different variable names (see image on the right).

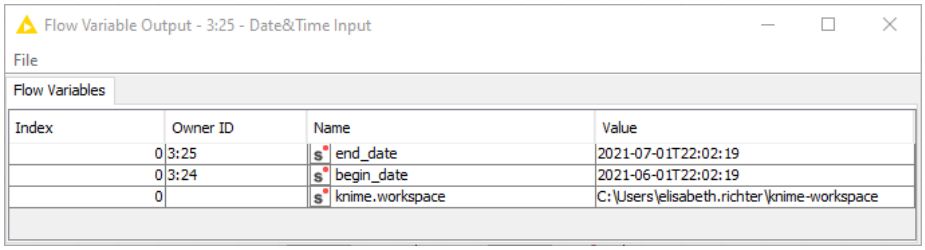


Use the proper variable name, and user can choose the date from the pop-up:





Similarly as I mentioned in the previous post, you can see the output variables:



Index	Owner ID	Name	Value
0	3:25	end_date	2021-07-01T22:02:19
0	3:24	begin_date	2021-06-01T22:02:19
0		knime.workspace	C:\Users\elisabeth.richter\knime-workspace

He even goes beyond the question and explains how the usage of a Component would be benefitting in this situation, how to create one, and how to make sure that the flow variables are passed on to the node following the component. Truly a model answer.

Find the original thread "Multiple Date Variable into DB Query Reader?" on the KNIME Forum.

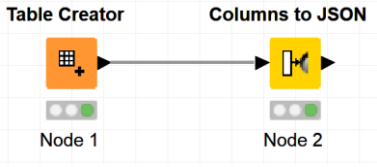
Your JSON Little Helper to Concatenate Cells

Another question that we probably all had to face is how to dynamically concatenate cell values together from a number *n* of columns, where *n* can change at every run. To illustrate the initial goal of the questioner, let's consider this table. The first four columns, ID, Start date, Name, and Last name, are supposed to be concatenated into the fifth column, Notes. Both the column names and the column values should be considered.

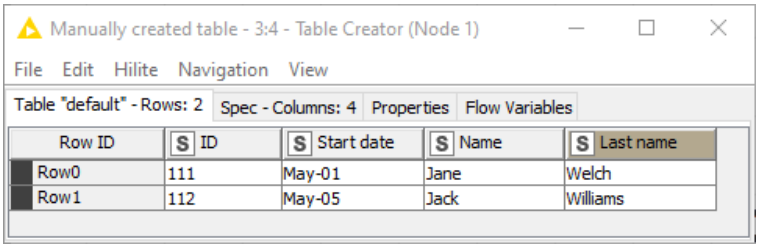
ID	Start date	Name	Last name	Notes
111	May-01	Jane	Welch	ID: 111 Start date: May-01 Name: Jane Last name: Welch
112	May-05	Jack	Williams	ID: 112 Start date: May-05 Name: Jack Last name: Williams

Here is how Bruno has solved the problem in a very elegant way, with a little help from the JSON nodes.

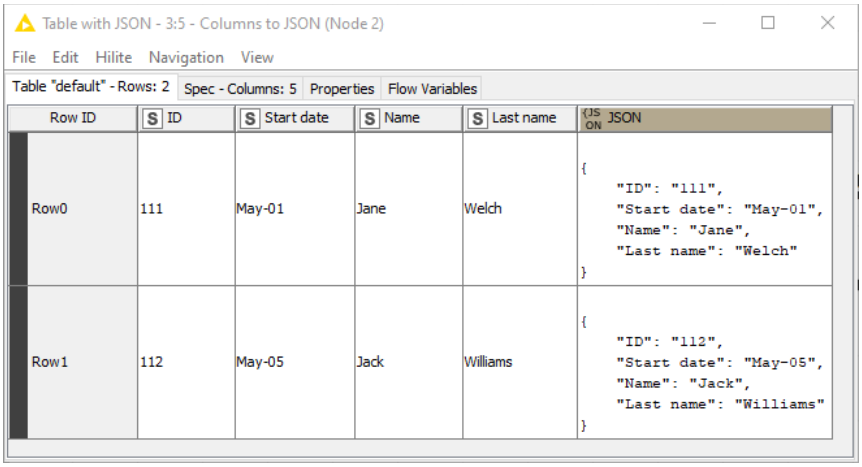
> *You can use Columns to JSON which will give you something in the same format that you are looking for, and it's dynamic as in you don't have to change anything if your table changes (new columns, less columns, it does not matter).*
>
> *Just plug it to your data:*



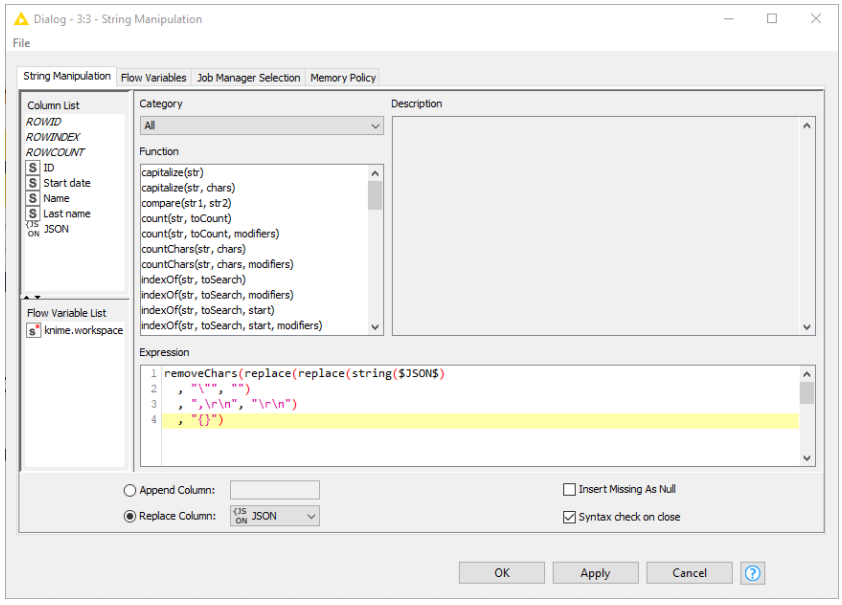
> *Input data:*



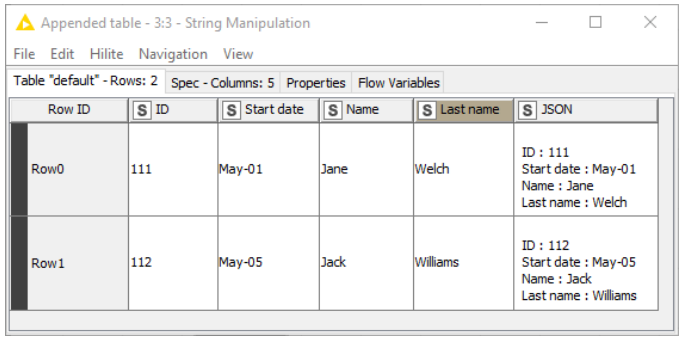
> *Results:*



You can do some cleanups if you really want the format that is mentioned. You can use the String Manipulation:



Results:



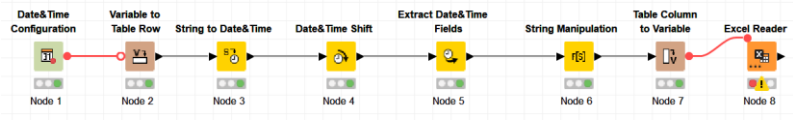
And here again, Bruno even went beyond and attached two more columns just to demonstrate the flexibility of this solution. It can be adapted to any number n of columns.

Find the original thread "Concatenate cells into a paragraph with column names and formatted text" on the KNIME Forum.

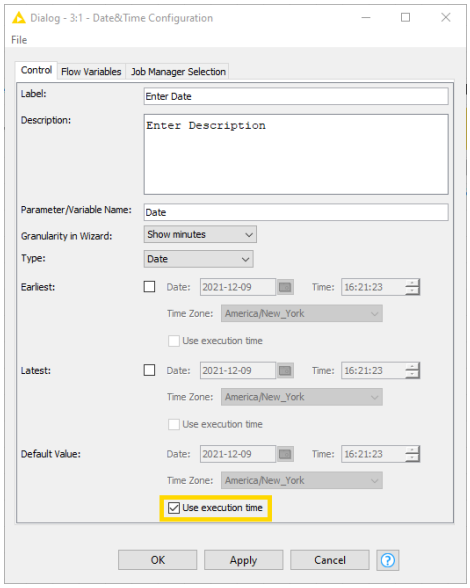
Addressing Yesterday's Data

Yesterday's data - isn't that an old hat? Well... no. At least not for all of us. I'm sure some of us have had the problem to address yesterday's data. Or at least some data delivered on a day that can only be extracted starting from today. One KNIME Forum user wanted to know how to dynamically read yesterday's Excel sheets, where each sheet is named as a date (e.g., "08.12.", 09.12.", etc.). Here is how Bruno solved the problem.

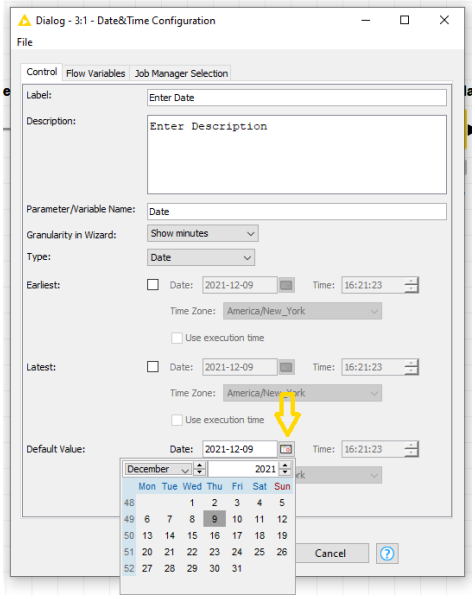
> *I've put something together that offers flexibility of changing the date in case you want to run the workflow for some other date, but default is the current date. The workflow also will work for any date format.*
>
> *My workflow looks like this:*



> *Looking at the first node, it's configured to use the current date:*

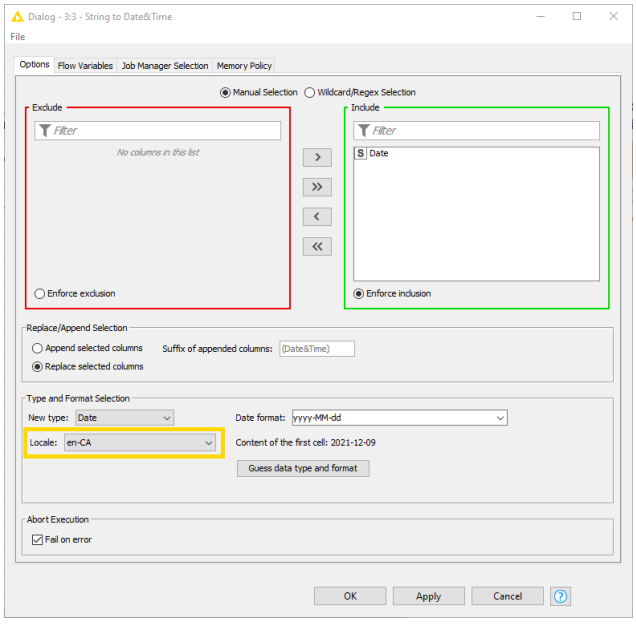


But should you want to run it for a specific date, just uncheck that box, and click on that little icon to get a nice calendar popup where you can choose which date you want to run it for:

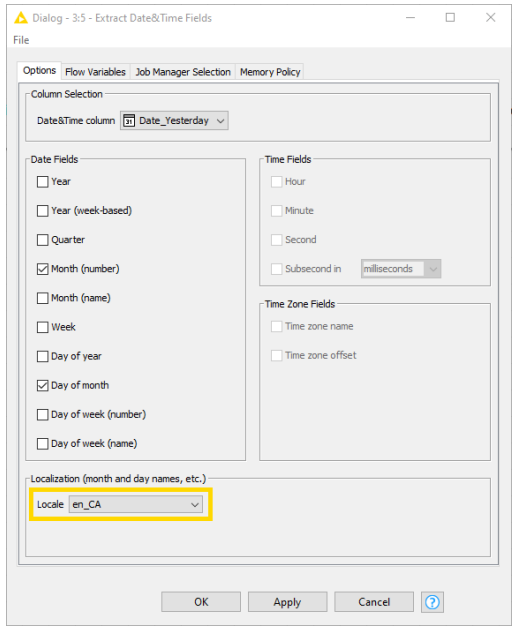


2 nodes that you need to configure based on your locale. The Node 3 and Node 5. Both should use the same locale, which should be your locale. I'm on Canada, so mine is en-CA. By default, KNIME would set your locale automatically based on what your system/computer is set to:

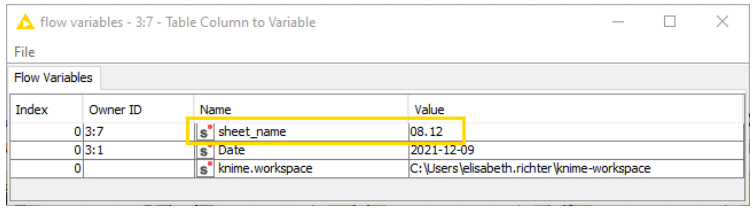
Node 3:



Node 5:

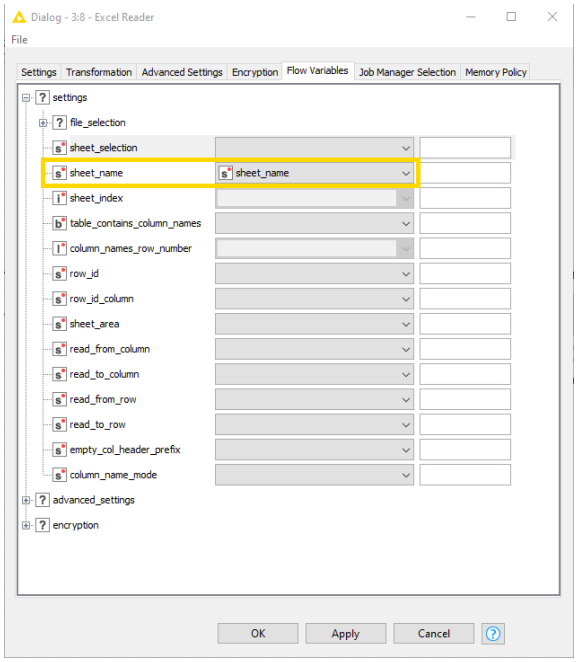


So, for today's date, it will generate a variable called sheet_name with value "08.12":



I have also configured it so that if the month is from 1 to 9 (Jan to Sept), the month will be formatted as 01 to 09, so if you run the workflow for Feb 4th, the sheet_name will be "03.02".

The Excel Reader is configured to use the sheet_name as sheet_name:



You just need to point to the file, and that's it.

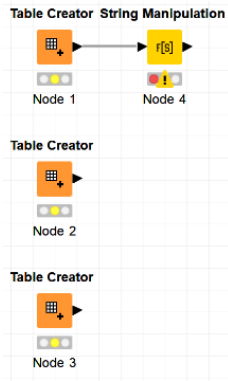
Given that answer there is nothing left to say. That wasn't too hard, wasn't it?

Find the original thread "How to dynamically read excel sheet" on the KNIME Forum.

One Node at a Time

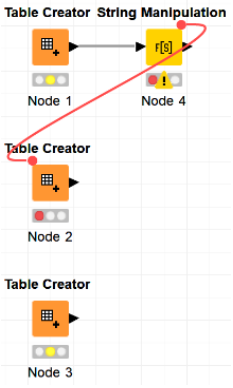
Haven't we all been at the point where we wanted more control over our workflow execution and specifically control the order in which our nodes are executed? Well, if you haven't there are at least some KNIME users on the Forum who asked specifically this question. Also here, Bruno posted an extensive explanation on how to handle this situation.

Basically, KNIME executes nodes sequentially from left to right, meaning if 2 nodes are connected together, it will execute the node from the left first and then the one to its right.



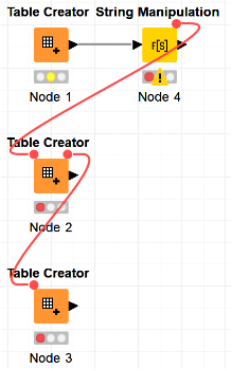
Now, there are cases where you "can't" connect 2 nodes, in that the output port type of the left node is not compatible with the input port type of the right node, or even cases where your left node does not have an output port (because there are no operations to be done in relation to that node, for example, Writers (Excel Writer, CSV Writer, etc), or Send Email, etc), or the right node does not have any input port (Table Creator, etc). In that case, you would connect them via the Flow Variable port.

For example, if I have this workflow (see image above) and I execute the workflow, then node 1, 2 and 3 will all start at the same time.



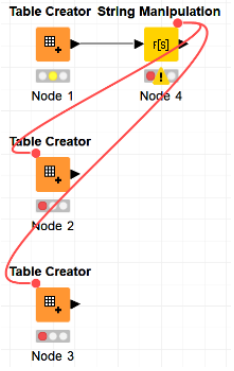
However, if I link the node 4 to the node 2 like this (see image on the left), then node 2 will execute only after node 4 is done. So when this workflow is executed, only node 1 and node 3 will start at the same time. Node 4 will execute only after node 1 is completed, and node 2 will execute only after node 4 is completed.

You can have different variations, for example:



In the figure on the left, node 2 will execute only after node 4 is completed, and node 3 will execute only after node 2 is completed.

Another variation is shown on in the figure on the right. Here, node 2 and node 3 are executed at the same time but only after node 4 is completed.



And further:

When it comes to metanode, it can be a bit tricky, depending on what's being done and what's being linked inside the metanode. A metanode simply "summarize" part of your workflow, so the nodes within a metanode are independent. The metanode is not one object, so some nodes can start already if they "left" nodes are ready. With components, it's a bit different. A component will not start until all input ports are ready.

This is really a valuable answer, especially for KNIME beginners who are not yet so familiar with all the functionalities of the KNIME Analytics Platform.

Find the original Forum thread "Ensure one node does not start until another is done?" on the KNIME Forum.

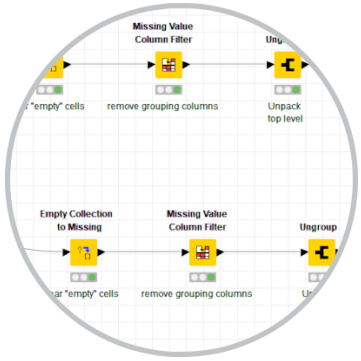
Seeking Help from the Best on the KNIME Forum

In this article, we highlighted Bruno's outstanding engagement on the KNIME Forum and wanted to acknowledge him for his many technical answers. He has for sure helped some new as well as expert KNIME users with his detailed replies. Almost every time Bruno proposes a solution, he not only provides many screenshots with it but also attaches a workflow to make sure that the questioner really is helped. So he remains true to his profile slogan "Help me help you by giving as much information as possible. The more accurate you are in your information, the more accurate the solution will be.".

Brian Bates was nominated KNIME Contributor of the Month for October 2021. He was awarded for his activity on the KNIME Forum and his Open File or Folder and String Emoji Filter components. As of 28.09.2022 both components total 1.926 downloads. The first component is for quickly opening and checking on a file or folder in your KNIME workspace. The second component is for removing emojis from emails or tweets. Brian is a ubiquitous trusted presence in the KNIME Forum. In this first six months on the Forum, he accumulated more favorited posts than anyone over the preceding year! He gets his motivation from the satisfaction of being confronted with the challenges on the Forum and from helping others with their workflows.

Brian has more than 35 years of broad IT experience which allows him to pick up new technologies quickly. He enjoys playing with data and likes to transform data into information. He is currently a Data & Integration Architect at the Walt Disney Company.

Visit Brians' space on the KNIME Hub or his profile in the KNIME Forum (Hub/Forum handle: `takbb`).



From Customizable XMLs to Flexible Date&Time Handling

Thinking Outside the Box with Brian Bates' Best-of Forum Threads

Author: Elisabeth Richter

Since Brian joined the KNIME Forum in March 2021, many KNIME community members have already benefited from his engagement. As of 17.08.2022, Brian counts 500 days visited, roughly 2800 topics viewed, and 106 given solutions. Indeed, this makes him a very active member on the KNIME Forum. His comments span over various ETL-related topics, including accessing data from Excel files, connecting to SQL databases, or dealing with various data types like strings or date & time data types. He's also providing lots of feedback and ideas for improving KNIME Analytics Platform by proposing feature requests.

Curious to know more about the tireless commitment of this KNIME support champion? Here is a collection of the three Best-Of-Brian forum threads where he has brilliantly put his knowledge and expertise at the service of others!

takbb

Brian Bates

London, UK

Joined Mar 5, '21 Last Post 1 day Seen 17 hours Views 2637 Trust Level member

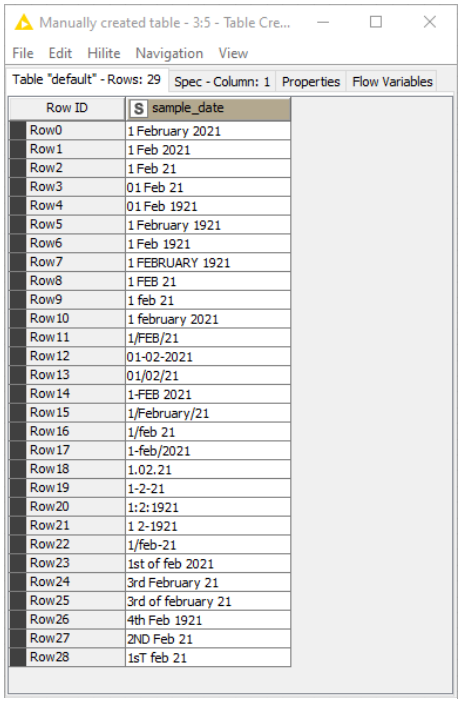
Flexible Date Format Handling

This is a very helpful KNIME Forum thread in which Brian deals with a topic that can quickly become frustrating: handling custom date and time formats. When reading date & time data in KNIME Analytics Platform it is usually read as string first. The respective column(s) can then be converted to KNIME's dedicated Date&Time data type using the *String to Date&Time* node. Ideally, we want to have a data set with consistent date & time strings that can easily be converted to date & time format. However, in real life this is often not the case as real life data can be messy, especially when blending data from different data sources where different date and time formats are used.

Brian noticed that there were a few questions on the KNIME Forum about converting dates. Sometimes, the format mask in the configuration dialog of the *String to Date&Time* node has a hard time detecting the right date & time format, especially if this is highly inconsistent. It occurred to him that it would be nice to handle a wide variety of date formats with a single mask, or some generic nodes and/or a tiny bit of generic code.

One problem he noticed is that if the month is a name, the mask is unhelpful to detect inconsistent formats because it is case-sensitive. For example, the month mask MMM will match "Feb", but it won't be able to match "FEB" and "feb". Similar problems occur, for example, when wanting to match two- and four-digit years (i.e., yy and yyyy).

To address these problems Brian created a workflow that is able to handle a wide variety of date formats all at once, i.e., within the same data set. Only requirement: the dates must be in the format day-month-year. The input table with various random date formats is displayed in the following.


Manually created table - 3:5 - Table Cre...

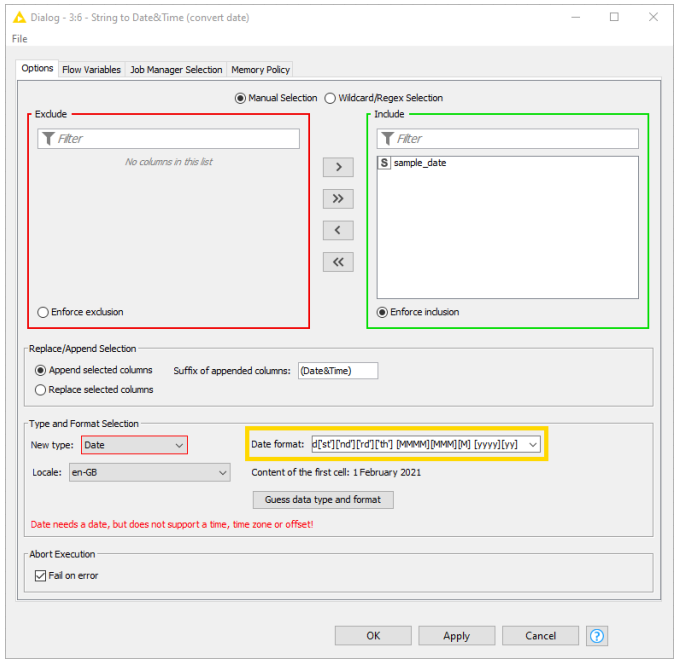
File Edit Hilite Navigation View

Table "default" - Rows: 29 | Spec - Column: 1 | Properties | Flow Variables

Row ID	S sample_date
Row0	1 February 2021
Row1	1 Feb 2021
Row2	1 Feb 21
Row3	01 Feb 21
Row4	01 Feb 1921
Row5	1 February 1921
Row6	1 Feb 1921
Row7	1 FEBRUARY 1921
Row8	1 FEB 21
Row9	1 feb 21
Row10	1 february 2021
Row11	1/FEB/21
Row12	01-02-2021
Row13	01/02/21
Row14	1-FEB 2021
Row15	1/February/21
Row16	1/feb 21
Row17	1-feb/2021
Row18	1.02.21
Row19	1-2-21
Row20	1:2:1921
Row21	1 2-1921
Row22	1/feb-21
Row23	1st of feb 2021
Row24	3rd February 21
Row25	3rd of february 21
Row26	4th Feb 1921
Row27	2ND Feb 21
Row28	1sT feb 21

A column with various random dates, all in the format day-month-year.

After reading the column with random dates, he used the String Manipulation node to prepare the date column. This means, for example, removing punctuation, or capitalizing the date (i.e., "1 february 2021" → "1 February 2021"). Then, he used the *String to Date&Time* node with a flexible date format mask to convert the entire column into the uniform KNIME's native Date&Time format yyyy-MM-dd. See the configuration window below under "Type and Form Selection" for the flexible date format.



The configuration window of the String to Date&Time node using a flexible date format mask. With these settings it is possible to convert various random dates into the KNIME's native Date&Time format, given the strings follow the format day-month-year.

The date format mask can be manually adapted according to your requirements. The workflow provided by Brian contains a second example that converts various random date strings that are in the format month-day-year.

He might convert his workflow into a useful component one day, so stay tuned…

Read the whole thread "Writing a flexible Date Format handling workflow" on the KNIME Forum and download the associated workflow "KNIME_Workflow With flexible DATE format mask" from the KNIME Community Hub.

Generating XML

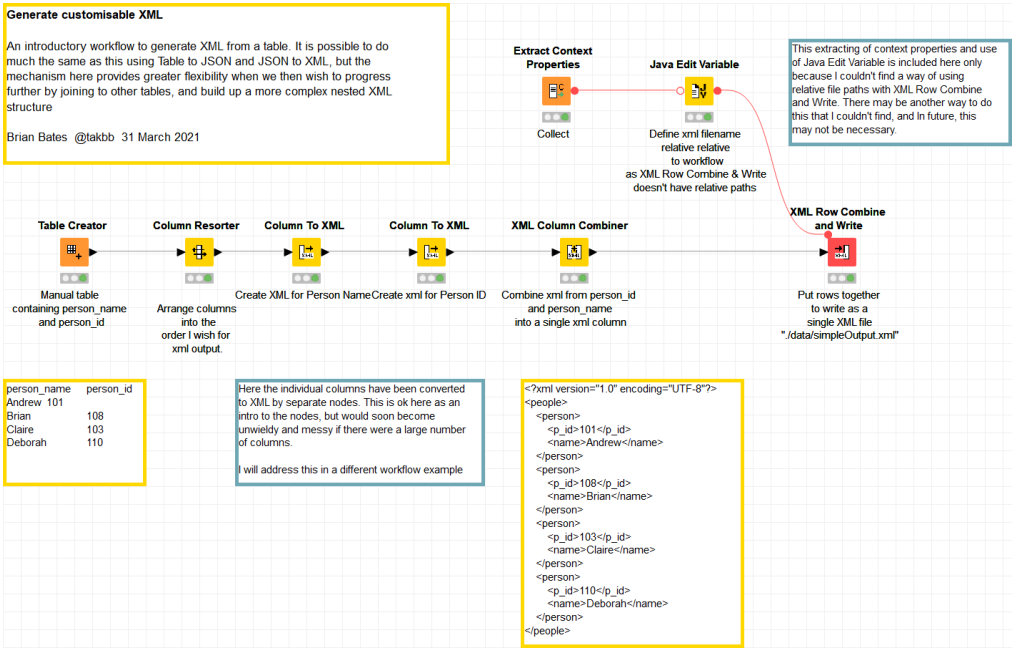
Dealing with XML (= eXtensible Markup Language) is something that everyone who works with data will likely have to deal with sooner or later. In KNIME Analytics Platform XML-dedicated nodes exist that ease XML handling. This forum thread addresses one task in particular: **creating** XML from one or more data tables. If you want to create an XML file in KNIME, one easy and quick way is to use the *Table to JSON* node followed by a *JSON to XML* node.

However, for Brian this was not enough. He wanted to go beyond this and was looking for a way to create XML not only from simple tables but also from more complex ones,

i.e., nested structure. In this forum thread, Brian documented his approach by breaking it down into three steps. He first built a workflow replicating the base case, i.e., one table becomes one XML file. He then created components to make the workflow more generic and flexible. Lastly, he built a third workflow to demonstrate how to create customizable, nested XML files.

The First Workflow: Simple Customisable XML Generation from table

The first workflow he created for this purpose is shown in the following figure. This workflow covers the basic case of generating an XML file from one table. In general, the functionality of the workflow resembles that of connecting sequentially the *Table to JSON* and the *JSON to XML* nodes. However, according to Brian, this procedure provides greater flexibility when we wish to extend the use case, for example by joining other tables and building a more complex nested XML structure.



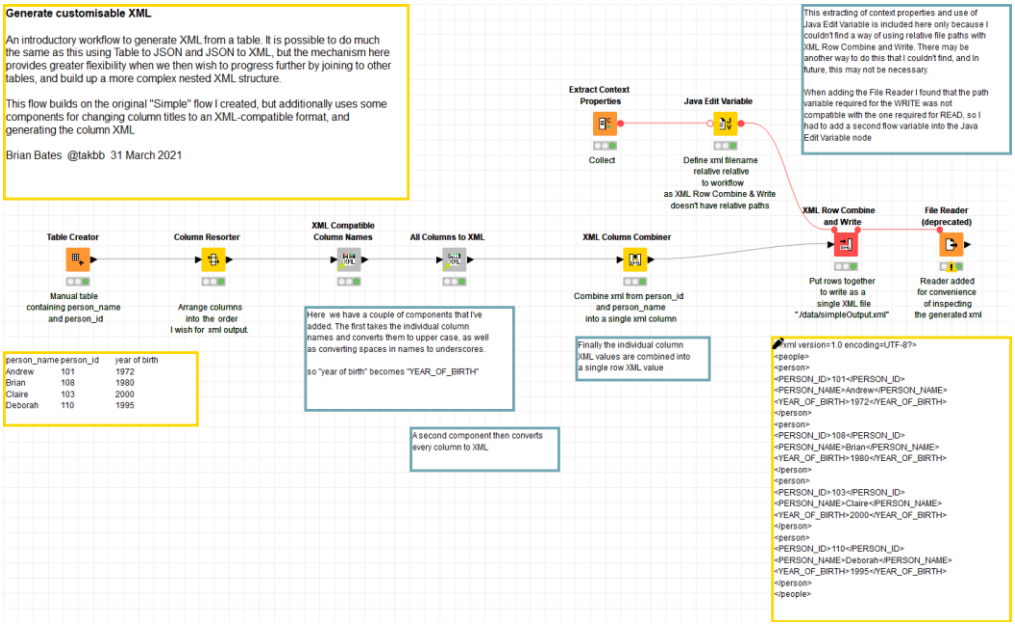
This workflow replicates the case where one table becomes an XML file without using KNIME JSON nodes.

As shown in the screenshot of the workflow, each column in the table is converted to XML separately using the Column to XML node repetitively. Although this works fine for a basic example, in a more complex use case with a large number of columns this would soon become unwieldy and messy. To adjust for this case, Brian created a second, more flexible workflow.

Workflow 2: Simple Customizable XML Generation from Table with components

In order to adjust for these circumstances, Brian's solution includes adding a loop to the workflow that iteratively converts all columns to XML. In addition to that, standardized column names are required in order for the loop not to break. Hence, a

second loop is added that renames all columns before inputting them to the Column to XML node. Lastly, to make the workflow neater, the loops are encapsulated in one reusable component each.



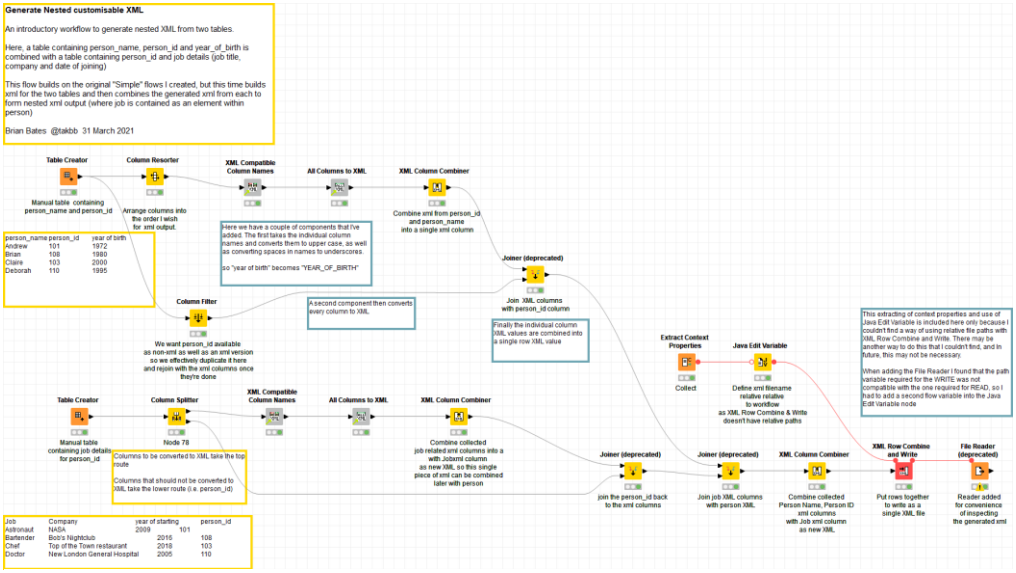
This workflow is an extended version of the first workflow that allows for input tables with a larger number of columns.

The XML Compatible Column Names component is responsible for standardizing the column names by converting them to uppercase letters and replacing whitespaces by underscores. The All Columns to XML component is responsible for looping through all columns in the provided data table and converting them to XML, using the Column to XML node.

Now, the last obstacle to tackle was to find a way to deal with a more complicated table structure, i.e., nested structure. Therefore, Brian created a third workflow.

Workflow 3: Nested Customisable XML Generation from table

In this third workflow, Brian demonstrated how to add another data table to the final XML file via the join operation. This allows for the creation of an XML file with a nested structure.



This workflow is another extension of the two workflows above and enables generating XML containing nested information.

Read the whole thread "Some workflows that I played with for generating XML" on the KNIME Forum and download the three workflows from the KNIME Community Hub:

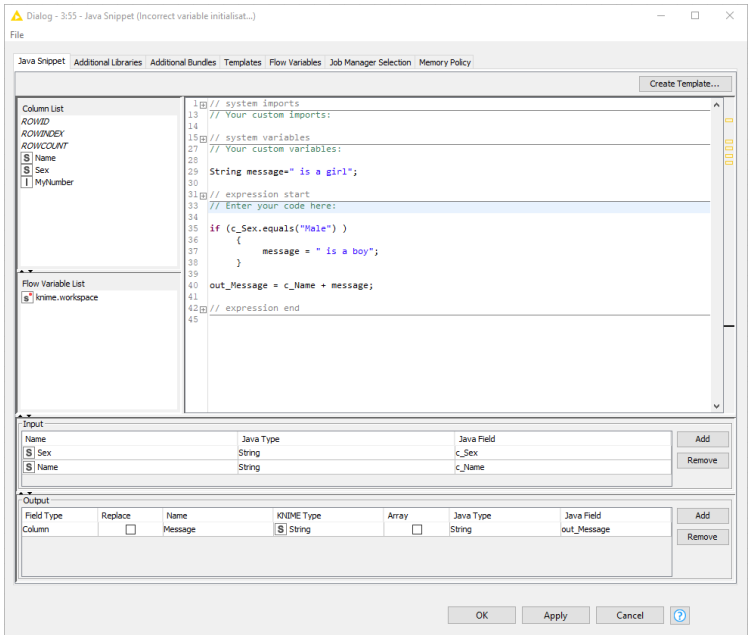
"Simple Customisable XML Generation from table"

"Simple Customisable XML Generation from table with components"

"Nested Customisable XML Generation from table"

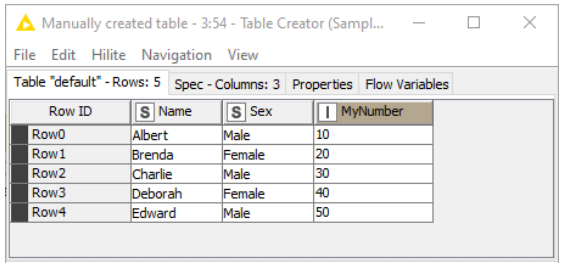
Custom Global Variables in Java Snippet Nodes

In this Forum thread, Brian addressed an interesting topic regarding the Java Snippet node. Although this might not come as a surprise to everyone, he stumbled on some interesting behavior of the Java Snippet node and wanted to share with everyone, so others don't make the same mistake. While playing with some Java code he noticed that variables defined in the section marked "system variables" in the configuration dialog of the Java Snippet node (see the figure below) are only processed during the instantiation of the object built from the piece of Java. They are basically custom "global" variables which means they are not re-initialized for each row.



The configuration dialog of the Java Snippet node. The variable "message" in the section "system variables" serves as a custom "global" variable and is not re-initialized for each row.

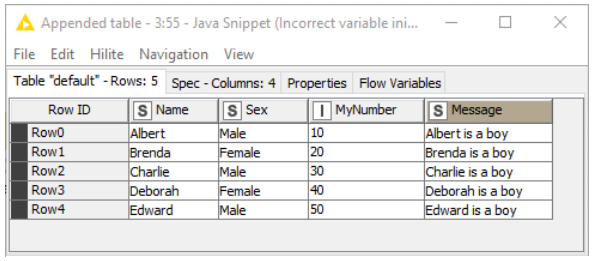
He explains this by means of the following example:



The input data table with the attributes "Name", "Sex", and "MyNumber".

His aim was to print for each data row the message "[name] is a boy" if Sex=Male, and "[name] is a girl" otherwise. However, the Java Snippet node configured as shown above produces an incorrect result as it prints the message "[name] is a boy" for each row, regardless of the value for "Sex".

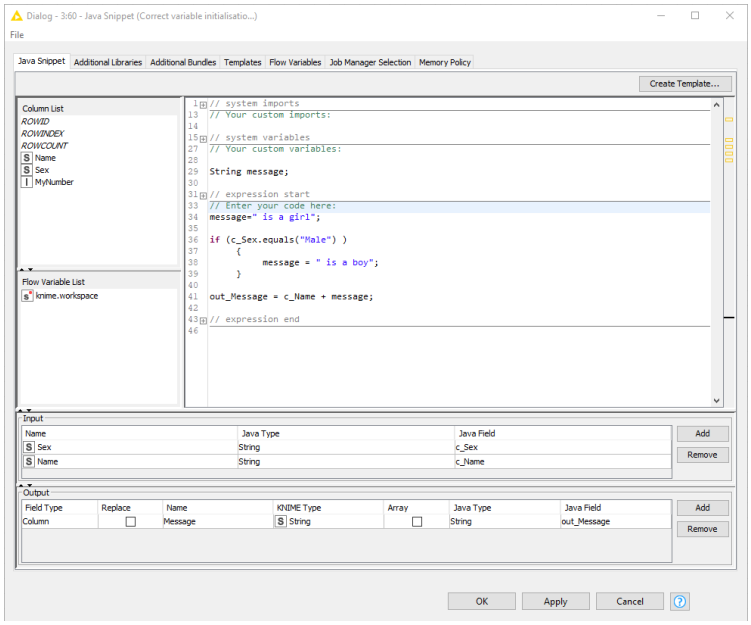
After some digging, Brian learned the reason for that is that the string "Message" is defined in the section "system variables". However, the variables in this section are global variables and hence they are not re-initializes for each row. This means, the value for the message is defined globally to "is a girl" but once the value changes to " is a boy" in the first row, it keeps this value for the remaining rows.



Row ID	Name	Sex	MyNumber	Message
Row0	Albert	Male	10	Albert is a boy
Row1	Brenda	Female	20	Brenda is a boy
Row2	Charlie	Male	30	Charlie is a boy
Row3	Deborah	Female	40	Deborah is a boy
Row4	Edward	Male	50	Edward is a boy

The result of the Java Snippet node as configured in the figure above. The column "Message" wrongly adds for each row the sentence "[name] is a boy".

However, changing the configuration of the Java Snippet node and defining the value for "message" in the section "expression start" (see the figure below) produces the correct result. Hence, "message" is not globally defined anymore and therefore re-initializes after processing each row.



The configuration dialog of the Java Snippet node when being correctly initialized. The variable "message" is defined in the section "expression start" and by that re-initializes the variable after processing each row.

But that's not all. Brian also included example use cases for which a definition of global variables can be useful, for example, calculating a running total, creating a lag column, adding a progressive concatenation of names, for sequence generation, or for a counter generator.

Read the whole thread "Java snippets have long memories!..." and find the associated workflow on the KNIME Forum.

Getting Inspiration from the Best on the KNIME Forum

Now that we've reached the end of this article, you may find that you've learned a thing or two after all. Or you might even have gained a whole new perspective on some topics you never thought about before. Whatever the case may be, thanks to Brian's broad knowledge and his tireless engagement in the KNIME Forum, he has definitely helped out and inspired new and expert KNIME users.

This article shows how great community support can be - and also how versatile it can be. The reason to be active in the KNIME Forum can be of different motivation. Be it to seek help with a particular error message, to report bugs or propose new features to our developers, or to simply share your thoughts and knowledge with the community. Whatever your motivation is, we're in any case happy about your contributions.

Node & Topic Index

H

I

J

K

L

M

N

P

R

S

T

U

V

W

X

Best of KNIME
The COTM Collection

We collected the top contributions of our KNIME COTMs from August 2020 to July 2022. This booklet contains 25 stories that teach you more about data science and KNIME. Let's learn from the best.

Elisabeth Richter holds a master's degree in Social and Economic Data Science. During her studies, she developed a keen interest in Machine Learning, Deep Learning, and various NLP-related techniques. Her research focused on understanding media bias and examining user behavior in social media. She is part of the Evangelism team at KNIME and works as a Data Science Publisher with a particular focus on the books published under KNIME Press.

COTM : August 2020 - July 2022

Vijaykrishna Venkataraman
Markus Lauber
SJ Porter
Angus Veitch
Keith McCormick
Evan Bristow
Miguel InfMad
Armin Ghassemi Rudd
Philipp Kowalski
Dennis Ganzaroli
Giuseppe Di Fatta
Alzbeta Tuerkova
makkynm
Tosin Adekanye
Ignacio Pérez
Brian Bates
Ashok K Harnal
Andrea De Mauro
Malik Yousef
Nick Rivera
Paul Wisneskey
Francisco Villarroel Ordenes
Bruno Ng
Christophe Molina
John Emery