

Edited by
ELISABETH RICHTER

Best of KNIME

The COTM Collection



August 2020 - July 2022



Copyright©2022 von KNIME Presse

Alle Rechte vorbehalten. Diese Veröffentlichung ist urheberrechtlich geschützt und die Erlaubnis muss vom Verleger vor einer verbotenen Reproduktion gewonnen, Lagerung in einer Retrieval System oder Getriebe in beliebiger Form oder auf beliebige Weise elektronisch, mechanisch, Photokopien, Aufzeichnungen oder ebenfalls.

Für Informationen über Berechtigungen und Verkäufe, schreiben Sie an:

KNIME Presse

Talacker 50

800 Zürich

Schweiz

Kimepress@knime.com

www.knime.com

Lernen wir vom Besten

Wenn Sie mehr über die Datenwissenschaft und über KNIME erfahren möchten, sollten Sie lernen aus dem Besten! Hier ist das KNIME Best.

In diesem Booklet haben wir die Top-Geschichten der Top-Beiträge für die KNIME gesammelt Gemeinschaft von August 2020 bis Juli 2022. Die wichtigsten Beiträge wurden ausgewählt und jeden Monat für ihre ausgezeichneten technischen Fähigkeiten und für ihren Beitrag vergeben die Verbesserung der Gemeinschaft in der Datenwissenschaft und in der KNIME.

Während dieser Zeit hatten wir 25 KNIME KNinjas, ein für jeden Monat, mit zwei Auszeichnungen im Januar 2021. In der Gruppe finden Sie Pädagogen, Datenlabor Manager, Datenwissenschaftler, Datenanalysten, Dateningenieure, Akademiker, Biologen, Chemiker, Marketing-Analyse-Experten, Software-Ingenieure, etc. Es wäre nicht überraschend zu finden zwei oder mehr solcher Qualifikationen kombiniert in einem einzigen KNinja.

Wir haben ihre beliebtesten Geschichten gesammelt, die Geschichten entweder weil sie gelöst a langjähriges Problem oder weil sie uns etwas nützliches für unsere Berufsleben. Manchmal war es unmöglich, eine einzelne herausragende Arbeit auszuwählen und wir beschlossen, ein allgemeineres Interview zu veröffentlichen, das alle Bereiche der Tätigkeit des Beiträgers.

In den letzten zwei Jahreszeiten des KNIME-Beitrags des Month-Programms haben wir die wichtigsten Merkmale der KNIME Software von Vijaykrishna Veknatarams Punkt wie man mit Duplikaten in Ihren Daten mit Markus Lauber richtig umgehen kann, dieser Text die Verarbeitung und GUID-Generation kann einfach mit den Komponenten von SJ Porter erfolgen, wie Sie textreiche Visualisierungen von Twitter-Daten um einen bestimmten Hashtag erstellen Angus Veitch. die Macht des niedrigen Codes mit Keith McCormick, über die beiden Gründer und Admins der KNIME Analytics Community-Gruppe auf Facebook, Evan Bristow und Miguel InfMad, dass Armin Ghassemi Ruds Übersetzer und Holen Sie sich Anfrage Plus Komponenten bündeln leistungsfähige Funktionalitäten, wie automatisch neue Charaktere und Setups für eine Dungeons & Dragons-Sitzung mit Philipp Kowalski, was es braucht, um ein Datenwissenschaftler mit Dennis Ganzaroli zu werden, oder was es braucht, um offiziell mit Giuseppe Di Fatta zertifiziert KNIME, wie man neue Drogenkandidaten identifiziert mit KNIME mit Alzbeta Tuerkova, auch viel interessant um die Anwendung von KNIME auf Japanisch mit makkynm, wie man KNIME erkennt betrügerische Kreditkartentransaktion mit Tosin Adekanye, über Excel zu KNIME in Spanisch mit Ignacio Perez, die Erstellung von XML kann einfach und flexibel mit Brian Komponenten von Bates, wie man Python-Code in eine KNIME-Komponente integriert Ashok Harnal, wie man ein Data Science Lab mit Andrea De Mauro, und auch, wie um maschinelles Lernen auf DNA-Analyse mit Malik Yousuf anzuwenden, dass Nick Rivera's YouTube-Kanal ist eine weitere große Quelle, um KNIME zu lernen, dass Sie eine Maschine lehren können Lernmodell mit KNIME mit Paul Wisneskey, wie man Machine Learning in

Marketing Analytics mit Francisco Villarroel Ordenes, was es braucht, um ein gutes zu sein
KNIME Forum-Poster mit BrunoNg, mehrere Möglichkeiten, wie man KNIME für alle Arten von
QSAR-Themen mit Christophe Molina und nicht zuletzt, wie man parsiert und analysiert
PDF-Dokumente mit John Emery.

Lassen Sie uns nicht länger in diese Einführung eintauchen. Lassen Sie uns mehr von der KNIME am besten hören.

Elisabeth, Scott, Corey und Rosaria

Inhaltsverzeichnis

VERWENDUNG DES KNA

<ahref="#page67" style="color:blue; text-decoration:underline;">Verfahren der Klima- und Umweltplanung
<ahref="#page67" style="color:blue; text-decoration:underline;">Verfahren der Klima- und Umweltplanung

[KNIME IN DER DATENWISSENSCHAFT](#page8)

[Home](#) | [Search](#) | [Help](#) | [Log In](#) | [Log Out](#)

VERWENDUNGSEBEREICH

[BILDUNG UND FORSPÅNING](#page98)

[KNIME SUPPORT](#page149)

NODE & 网络设计与实现

Willkommen beim KNIME-Beitrag

Das Programm des Monats

Vielleicht haben Sie Preiskarten oder Abzeichen für KNIME Beitrag des Monats gesehen (COTM) auf sozialen Medien, die wie diese unten aussehen. Wenn Sie noch nie einen gesehen haben oder Sie fragen sich immer noch, was dies feiert, jetzt erhalten Sie eine Antwort.



Die Preiskarte für unseren KNIME-Beitrag des Monats für Dezember 2021 – Andrea De Mauro. Jeden Monat vergeben wir ein Community-Mitglied für ihr herausragendes Engagement und jedes COTM erhält ein Abzeichen.

Was ist ein KNIME COTM?

Die KNIME COTM Auszeichnung wird KNIME-Nutzern vergeben, die sich ausgezeichnet gezeigt haben technische Fähigkeiten und haben zu einer besseren Lernerfahrung als Erzieher beigetragen, schnellere und umfassendere technische Unterstützung, Wissensaustausch über Artikel, Blogs, und YouTube-Videos, zu einem reicheren Repository von Community-Knoten und Komponenten, oder einer stärkeren Präsenz von KNIME in sozialen Medien.

Der COTM-Preis wird in der Anerkennung technischer Fähigkeiten vergeben und Beitrag zur Förderung der KNIME-Gemeinschaft. In der Tat gibt es nur handvoll von solchen Top-Experten auf der ganzen Welt: Dieses Niveau von Know-how und Engagement ist schwer zu finden!

Das Programm hat im August 2020 begonnen und ist seitdem stark! Mit Nominierung von John Emery im Juli 2022 erreichten wir den Punkt, wo unser Programm voll durch zwei ganze Jahreszeiten. Bis dahin haben wir 25 KNIME-Nutzer vergeben, eine pro Monat (mit Ausnahme von Januar 2021, als wir zwei Gemeinschaft vergeben gleichzeitig). Dies ist jedoch nicht das Ende des Programms wie im August 2022 Saison drei hat begonnen...

Wie wird man COTM Awardee?

Zunächst darf jemand – ein Fan, ein Kollege, ein Familienmitglied, sich selbst – nominieren

Sie oder ein anderer KNIME-Nutzer für einen COTM-Preis über die [CO Vorschlag der Kommission](#)

Form. Im Beschreibungsfeld der Form, die Gründe, warum der Nominierte verdient,

Der COTM-Preis muss angegeben werden.

Einmal im Monat werden alle COTM-Nominierungen ausgewertet. Auf der Grundlage ihrer jüngsten Aktivitäten, ihr Beitrag zur KNIME-Gemeinschaft und ihre technischen Fähigkeiten, die besten Nominierten wird für den nächsten Monat mit dem Titel COTM ausgewählt und ausgezeichnet.

Wenn, während des Lesens, der Name eines verabscheuenden KNIME-Benutzers zu beachten, zögern Sie nicht und nominieren sie für das COTM-Preisprogramm in der [CO Antragsformular](#)

Die Liste aller vergangenen COTMs für alle Jahreszeiten finden Sie in der [KNIME COTM Hall of Fame](#)

KNIME im Data Science Lab

In diesem Abschnitt haben wir alle Beiträge zusammengefasst, die sich auf KNIME Analytics konzentrieren Plattform und seine Fähigkeiten. Die hier vorgestellten Artikel zeigen, was KNIME hat die Leistung der KNIME Analytics Platform anbieten und aufdecken. Die Kategorie "KNIME in der Data Science Lab" bietet unsere eager Data Science-Experten:

- **Vijaykrishna Venkataram**
 - Senior Manager, Data Analytics @Relevantz
- **In den Warenkorb**
 - Leiter Daten & Analytics @Vodafone
- **Das ist nicht alles.**
 - Head of Report & Data-Management @Swisscom
- **SJ Porter**
 - Site Reliability Engineer @KNIME
- **Das ist der Hammer!**
 - Professor @FORE Schule des Managements



Vijaykrishna Venkataram wurde nominiert KNIME COTM

für August 2020, was ihn zu unserem ersten COTM macht! Er wurde für die Schaffung einer karte der Merkmale, Erweiterungen und Integrationen von KNIME Server und KNIME Analyseplattform. Die ganze Idee des COTM Programm kam aus seiner Mind Map. Die Mind Map war so detailliert und so allgemein, dass wir dachten, wertvolle Beiträge wie diese sollten belohnt werden. In der

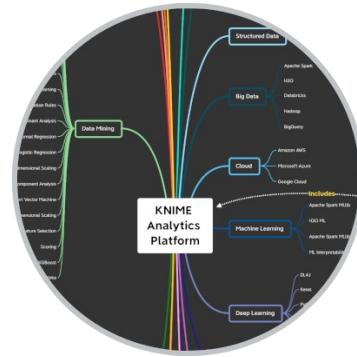
Figur rechts, Sie können einen Teil der besagten Mind-Karte sehen. Wenn es zu klein ist alles zu unterscheiden, können Sie lesen [Linked In den Warenkorb](#) zu diesem Thema.

Vijay hält einen Master of Science in Financial Economics, Finance & Economics von die Madras School of Economics in Chennai, Indien. Er hat mehr als 14 Jahre Erfahrung in der Bank- und Kreditbüro-Branche und ist derzeit Senior Manager, Data Analytics bei Relevantz. Einige seiner Schwerpunkte umfassen Credit Scoring (Anschaffung & Verhaltensanzeigen), Verlust Prognose, Cross-Sell-Analyse und Marktanalyse Analyse. Neben seinen technischen Fähigkeiten, Vijay auch mit Kompetenz in Kommunikation Komplex Modellierungslösungen für technische und nicht-technische Geschäftspublikum.

Besuchen Sie Vijay's Raum auf dem KNIME Hub oder sein Profil

Seite im KNIME Forum (Hub/Forum Griff:

))



Experten diskutieren Werkzeuge und Prozesse für effektive Finanzanalytik

My Data Guest — Ein Interview mit Vijaykrishna
In den Warenkorb

Autor: Rosaria Silipo



Es war mir ein Vergnügen, vor kurzem Interview [Vijaykrishna Venkataraman](#) als Teil der [Mein Gott](#).
Daten Gast Interview-Serie. Er teilte Einblicke in die Arbeit eines Datenwissenschaftlers in der Bank- und Finanzsektor, illustrierte Schlüsselmerkmale der KNIME Software, die seinem Organisation optimiert Prozesse und sprach über die Bedeutung des Modells Governance in stark regulierten Branchen.

Vijaykrishna Venkataraman ist Experte Datenwissenschaftler mit 14 Jahren Erfahrung in Datenwissenschaft vor allem auf den Banken- und Finanzsektor angewandt. Er arbeitete in der Abteilung Risikomodellierung und Analytics der Zentralbank von Oman und ist derzeit zurück nach Indien, wo er als Senior Manager und Data & Analytics Consultant. Vijay war auch der erste [KNIME Beitrag des Monats](#), mit wurde im August 2020 vergeben. Die ganze Idee des Beitrags des Monats Auszeichnung kam von seiner Arbeit. Nach der Ansicht der Mind Map, die Vijay produziert, um zu beschreiben die KNIME-Funktionen, dachten wir, "Such kreative Arbeit sollte belohnt werden. Lassen Sie uns Belohnung es!" und der Rest, wie die Menschen sagen, "es ist Geschichte".

Rosaria: Beginnen wir von Ihrem berühmten [KNIME-Funktion Mind Map](#) . Können Sie es erklären?
Wir?

Vijay: Die Idee der Erstellung der Mind Map kam, während wir bereits verwendet KNIME Analytics Plattform umfassend in unserer Organisation. Jetzt und dann ein Kollege kam zu mir und fragte etwas über die Plattform: „Unterstützung von KNIME das?“ oder „Kann ich das in KNIME tun?“. Ich musste KNIMEs Funktionen jedes Mal erklären. Am ein Punkt Ich dachte: "Warum muss ich immer wieder dasselbe erklären? Warum nicht eine Mind Map erstellen?“. KNIME ist schließlich eine visuelle Programmierplattform, und es sollte möglich sein, seine Merkmale visuell zu illustrieren anstatt die alte Schule zu gehen Art der Vorbereitung von Dokumenten. Die Mind Map deckt nicht das gesamte Spektrum aller

Features KNIME Analytics Platform unterstützt, aber ich denke, dass die wichtigsten sind definitiv enthalten.

Rosaria: Wird auf die neueste Veröffentlichung von KNIME Analytics Platform und KNIME aktualisiert Server? Wo befinden sich die neuen Features auf Ihrer Karte?

Vijay: Ja, diese Mind Map ist die zweite Version und wir haben sie auf die neueste Version aktualisiert (v4.6) der KNIME Software. Wir haben die meisten neuen Features am unteren Rand der die Karte. Die Updates, die wir auf der Mind Map gemacht haben, beinhalten auch die wichtigsten Funktionen KNIME Server.

Rosaria: Welche KNIME-Funktionen nutzen Sie am meisten in Ihrer Arbeit?

Vijay: Unser Geschäft ist sehr reguliert, und als solche arbeiten die meisten Daten mit gehostet On-Premises. Ein Teil unserer Daten wird jedoch über mehrere Plattformen verstreut und Produkte. Wir brauchten also eine Plattform, die alles in eine Platz. Hier entstand die KNIME Analytics Platform. Wir verwenden KNIME's dedizierte DB-Knoten ausgiebig, um die Daten zu extrahieren. Was folgt, sind in der Regel mehrere Datenaggregation und Reinigung , ETL, im Grunde. Dies macht 70-80% von unsere Arbeit. Der nächste Schritt ist die Automatisierung von Prozessen und hier ist KNIME Server zeigt seine Macht. Besonders die Möglichkeit, KNIME Server mit anderen zu verbinden Business Intelligence Tools wie PowerBI sind sehr wertvoll.

Rosaria: Um zu verstehen, warum diese Funktionen, vielleicht können Sie uns mehr über Ihre professionelles Selbst. Ich hätte diese Frage früher stellen sollen. Wie viele Menschen sind in Ihrem Gruppe? Wie viele Berufsprofile?

Vijay: Für die meisten meiner beruflichen Leben habe ich in Daten und Analysen für die Bank- und Finanzsektor. Meine Arbeit konzentriert sich vor allem auf Risikoanalysen wie Risiko Management-, Kredit- oder Betraganalysen. Ich habe zur Einrichtung des Kreditbüros beigetragen für die Zentralbank von Oman, wo meine Aufgabe nicht nur die Gestaltung der Daten Pipeline über datentechnische Lösungen, aber auch datenwissenschaftliche Aufgaben, wie Erstellung von Credit-Scorecards und Geschäftsanalyseverfahren, wie Benchmarking und Markteinblicke Reporting.

Was die professionellen Profile in unserer Gruppe betrifft, haben wir meist Dateningenieure und Daten Analysten.

Rosaria: Stellen Sie sich ein?

Vijay: Ja, ich werde mich bald einstellen. Ich bin gerade dabei, wieder nach Indien zu ziehen und bei einer anderen Organisation, wo ich die Datenanalysepraxis einrichten werde. Einmal Ich nehme diese Rolle auf, es wird sicher sein, dass einige Einstellung kommen die Pipeline in vielleicht 3 bis 6 Monate ab jetzt. Wenn Sie interessiert sind, folgen Sie mir [Linked Im Profil](#) zu bleiben...
Datum.

Rosaria: Vor einiger Zeit schrieb ich den Artikel „ [KNIME Experten wollten!](#)“ wo ich geklagt habe Mangel an KNIME-Experten, um Stellenangebote zu füllen. War es auch für dich wahr? Wie kompliziert war es bisher, Datenwissenschaftler und KNIME-Experten einzustellen?

Vijay: Als ich deinen Artikel las, dachte ich: „Ein Mann hat endlich meinen Verstand gesprochen!“. Wann wir begannen, nach Leuten zu suchen, die unsere Gruppe im Jahr 2019 beitreten, wir waren speziell auf der Suche nach Kandidaten, die KNIME Erfahrung hatten. Leider gab es damals nicht so viele KNIME-Experten verfügbar.

Jetzt werden die Dinge allmählich besser, aber es kann immer noch herausfordernd sein. Daher, wenn Sie sind auf der Suche nach KNIME Experten, lassen Sie mich Ihnen einen Tipp geben: Warte nicht, bis jemand mit KNIME-Expertise erscheint. Suchen Sie lieber nach einem Business oder Data Analyst, der zu Ihrem passt Geschäftsanforderungen und lehren sie KNIME. Ich bin sicher, dass sie innerhalb weniger Wochen wird das Werkzeug meistern. Das haben wir getan, als wir mit KNIME und der Management fragte, wo wir Kandidaten mit KNIME Fähigkeiten einstellen könnten. Ich habe zwei eingestellt Geschäftsanalysten und gab ihnen zwei Wochen, um sich mit KNIME vertraut zu machen. Am Ende in der zweiten Woche konnten sie einige der Probleme knacken, die ich ihnen gab.

Die KNIME Zertifizierung Programm ist auch sehr hilfreich in diesem Sinne, und deshalb sind wir bat unsere Mitarbeiter, L1- und L2-zertifiziert zu bekommen. L1 ist kostenlos und L2 ist nur ein Teil der Kosten, die Sie normalerweise für proprietäre Software ausgeben würden. Als Ergebnis, wir nicht müssen, um Talente auf dem Markt, aber wir trainieren und wachsen unsere eigenen Talente intern.

Rosaria: Was ist die professionelle Kategorie, die am schwersten zu rekrutieren ist? Und warum?

Vijay: Die kompliziertesten Menschen zu mieten sind diejenigen, die ein maschinelles Lernen haben Hintergrund und wollen Vorhersagemodelle sofort aufbauen. Studenten und frisch Talente sind oft unter dem Eindruck, dass Modellbau der wichtigste Teil eines Datenwissenschaft Projekt. Sie wissen nicht, dass in der realen Welt dies nur 10-15% der ganzes Projekt. Die Mehrheit der Arbeit ist, die erforderlichen Daten zu erhalten, verstehen die Business-Fall und die Daten zur Hand, bereiten und reinigen, Funktion Engineering tun, wenn notwendig, definieren Sie Ihr Ziel, etc. Datenmodellierung kommt viel später.

Zum Beispiel sprechen wir in der Bankenbranche viel über die Scorecard-Entwicklung oder Betrugsanalyse. Insbesondere bei Betrugsanalysen ist Ihr Prognoseziel, die Betrugssquote, sehr niedrig. Also, wie kommen Sie mit einem guten Modell für Vorhersage? Hier bist du zuerst müssen die Daten und das Geschäft verstehen, und dann können Sie in Modell springen Entwicklung.

Rosaria: Welche Tools muss ein Datenexperte wissen, in Ihrer Gruppe zu arbeiten?
Python? KNIME? Beide?

Vijay: Zunächst nutzten wir proprietäre Software, um unsere scoring Modelle zu betreiben. Aber die Lizenz wurde sehr teuer, und Talente auf dem Markt für solche proprietären Werkzeuge wurden immer teurer. Die jüngere Generation nutzte mehr Open-Source-Software, die wir wollten

auch annehmen. So begannen wir mit Python, die meisten unserer maschinellen Lernmodelle zu bauen, und SQL für Datenmunging. Derzeit verwenden wir KNIME für alle unsere ETL-Aufgaben und immer noch auf Python für maschinelles Lernen verlassen. Aber wir werden langsam einige von unsere maschinelle Lernarbeit von Python bis KNIME. Besonders mit dem neuesten KNIME Software-Release (v4.6) können wir jetzt Python-basierte Knoten erstellen.

Rosaria: Verwenden Sie neben Python und KNIME weitere Tools in Ihrer Gruppe?

Vijay: Neben KNIME und Python nutzen wir Business Intelligence Software, überwiegend PowerBI, zur Berichterstattung. Darüber hinaus, wie wir auch mit zunehmend große Datensätze, wir experimentieren mit verteilten Systemen. In dem Sinn, Spark sieht vielversprechend aus wegen seiner Geschwindigkeit und weil KNIME hat eine gewidmet [Erweiterung für Apache Spark](#)

Rosaria: Wie hat KNIME Ihnen und Ihrem Team bei Ihrer Arbeit geholfen? Kannst du uns nehmen? durch einen Anwendungsfall oder Beispiel?

Vijay: Okay, lass mich dir ein Beispiel geben. Darlehen im Bankensektor werden in mehrere Währungen und Währungen schwanken jeden Tag. Angenommen ich bereite mich vor eine Portfolio-Review und ich möchte sehen, wie mein Portfolio heute aussieht. Dafür brauche ich den neuesten Währungswert, um die Kredite meines Portfolios in die gegenwärtige Währung umzuwandeln Wert.

Wir haben unser IT-Team zuerst gebeten, eine App zu erstellen, die Wechselkurse von ca. 50-60 aktualisiert Währungen täglich. Es hätte jedoch zu lange gedauert, bis sie bauen.

Eine solche App.

Die nächste beste Lösung war die Verwendung von KNIME. Verwenden [KNIME REST Knoten](#), wir haben eine Arbeitsablauf, der automatisch auf einer täglichen Basis durch die entsprechende Währungstauschanbieter, aktualisiert sie und schreibt sie zurück zu die Datenbank. Diese Lösung wurde in ein paar Tagen umgesetzt , wörtlich. Dank KNIME Server, wir könnten dann den Workflow planen, um jeden Tag um 12 Uhr mit nur ein paar Klicks. Und falls ein Fehler auftritt, benachrichtigt uns KNIME Server sehr bequem pro Email.

Rosaria: Sie haben erwähnt, dass Sie mit zunehmend großen Datensätzen arbeiten. Wie funktioniert das? KNIME griff eine große Anzahl von Platten?

Vijay: Ich werde mit einem anderen Beispiel antworten. Irgendwann entschieden wir uns, von Oracle an Microsoft SQL Server. Das bedeutet, dass wir alle Daten zu einem neuen Server und die Datenstruktur anpassen. Es war eine große Aufgabe, da wir fünf Jahre hatten Daten verlegen. Mit Hilfe von KNIME konnten wir glatt 137,5 migrieren Millionen Datensätze (d.h. etwa 5,8 Milliarden Datenpunkte). Dies geschah mit einem Komplex einen Workflow, der die Daten aus der Datenbank holt und dann in den neuen Struktur. Dieses große Projekt wurde nur von zwei Personen erfolgreich abgeschlossen und ohne Schreiben einer einzigen Zeile des Codes.

Rosaria: Das ist wunderbar! Reden wir über Geld und Zeitersparnis. Wie wichtig ist die Auswirkungen von Datenanalysen in einem Unternehmen? Was sind einige unmittelbare Auswirkungen, die Sie haben Sie in Ihrer Erfahrung gesehen?

Vijay: Ich könnte dies nicht in der Lage sein, in Bezug auf Geld, aber definitiv in Bezug auf Zeit. Mit KNIME können wir viel Zeit sparen. In einem Geschäft wie Banking, Zeit ist alles, was zählt. Sie könnten gewinnen oder verlieren einen Kunden nur wegen der Entscheidungsfindung Prozess, d.h. die Zeit, die der Kunde auf Sie wartet, um ihr Problem zu lösen. A great Wie KNIME unsere Kundenunterstützungsprozesse optimiert hat. Vor dem Schalten KNIME, es dauerte viele Tage, um Kunden zu helfen. Jetzt ist der Prozess viel schneller und wir sind in der Lage, Anfragen zu verfolgen, und helfen Kunden in einem halben Tag, manchmal sogar in 30 min. Die neue KNIME-getriebene Lösung reduzierte auch unsere Churn-Rate, die uns indirekt gerettet hat Geld. Darüber hinaus haben wir auch Ressourcen eingespart. Die Anzahl der Kunden, die wir unterstützen jetzt mit nur zwei Mitarbeitern wäre in der Vergangenheit nicht möglich gewesen. Wir würden brauche mindestens zehn Menschen.

Rosaria: Erzählen Sie uns von Ihrer Erfahrung mit KNIME Server und was dich überzeugt hat Ihr Team, um sich dafür zu entscheiden (wie haben Sie das Management überzeugt, KNIME Server zu wählen über andere Plattformen?)

Vijay: Unser Ansatz, das Management zu überzeugen, war ganz einfach. Wir haben schon sie überzeugt, die KNIME Analytics Platform zu nutzen, als wir ihnen zeigten, dass sie einen Teil ihrer Arbeit automatisieren konnten. Es gab noch eine Hürde: Wie man automatisch Workflows ausführen? Hier wird KNIME Server wirklich wertvoll, wie Sie Jobs planen, Ausführungsschlangen festlegen und Benachrichtigungen senden können scheitern/erfolgen.

Aber das ist nicht alles. Ein weiteres sehr interessantes Feature von KNIME Server ist die [KNIME Webportal](#). Zum Beispiel haben wir einen KNIME Workflow als interaktiver Browser-basierte Daten-App, um die Verkäufe zu überwachen, wo der Endbenutzer (in der Regel das Management) kann Genießen Sie das Dashboard, ohne den Workflow hinter sich zu lassen. Für jemand in Finanzen, Vertrieb oder Buchhaltung, das ist alles, was sie sehen müssen.

Mein Vorschlag an jeden, der einen Weg sucht, das Management zu überzeugen, ist, Wählen Sie einen schwierigen Geschäftsfall und lösen Sie das Problem mit KNIME Software. Zeig sie, zum Beispiel das KNIME WebPortal, so dass sie die Vorteile davon verstehen können, ohne haben jedes KNIME-Wissen an sich.

Rosaria: Wir erreichen das Ende unseres Interviews. Bevor wir uns verabschieden, können wir nicht aber die klassische Frage stellen. Wo sehen Sie die Datenwissenschaft in den nächsten Jahren? Was wird der nächste Hype sein?

Vijay: Ich denke, die Datenwissenschaft selbst ist derzeit am Höhepunkt ihrer Hype-Kurve. Eine wichtige Wir sollten uns mehr auf die Modellführung konzentrieren. In den letzten Jahren haben wir bezeugt die Entwicklung vieler neuer, außergewöhnlicher Modelle, aber die meisten von ihnen sind schwarze Boxen. Der Fokus lag nicht auf Dolmetscherbarkeit und Erklärbarkeit. In

Industrien wie Finanzen oder Versicherungen, die sehr reguliert sind, und wo jeder eingesetzt wird Modell muss durch die Kontrolle des Regulators gehen, Modell Governance ist wahrscheinlich zunehmend prominenter werden. In den nächsten Jahren werden wir uns zurückziehen.

Modellen, die einfacher zu erklären und zu interpretieren sind. Ein parsimonischerer Weg Entwicklungsmodelle müssen zurückgebracht werden. Erstellen von schwarzen Box-Modellen, deren Der zugrunde liegende Entscheidungsprozess ist unsicher für alle Beteiligten wird nicht sein nachhaltig.

Die berühmte Mind Map von Vijay finden Sie auf der [KNIME Community Hub](#). Wenn Sie wollen verbinden mit ihm, bleiben Sie auf dem neuesten Stand über seine zukünftigen Projekte und Jobmöglichkeiten, sein [Linked Im Profil](#) zu Ihrem Netzwerk.

Dieser Artikel wurde erstmals in unserem [Niedriger Code für Advanced Data Science Journal](#) auf Medium. Die Originalversion finden [Hier](#).

Beobachten Sie das ursprüngliche Interview mit Vijaykrishna Venkataraman auf YouTube [Data Guest – Ep 12 mit Vijaykrishna Venkataraman](#)



[In den Warenkorb](#) wurde nominiert KNIME-Beitrag der Monat für Dezember 2021. Er wurde für seine Buch Datenanalyse Ausverkauft (siehe Bild rechts) mit einer vollständigen Einführung in KNIME und mehreren Tutorials zeigen, wie man Maschine lernt KNIME Workflows. Wie Andrea selbst sagt, ist das Buch für alle, die mit Daten arbeiten oder möchten.

Andrea hat mehr als 15 Jahre international die Verwaltung von Data Analytics und Data Science Organisationen. Er ist derzeit Leiter Daten & Analytics bei Vodafone Italien. Andrea hält eine PhD in Management Engineering von der Universität Rom, ein Master of Science in Elektro- und Computertechnik aus dem Universität Illinois in Chicago, ein Masterstudium in IKT Engineering von Polytechnic von Turin, und Diplom in Innovation von Alta Scuola Politecnica in Mailand. In seiner Forschung untersucht er die wesentliche Komponenten von Big Data als Phänomen und die Auswirkungen von AI und Data Analytics auf Unternehmen und Menschen.

[Besuchen Sie Andrea's Raum auf dem KNIME Hub](#) oder [Profil Seite im KNIME Forum](#) (Hub/Forum Griff: Admin)



BI Leader informiert Datentools für Profis

My Data Guest — Ein Interview mit Andrea De Mauro

Autor: Rosaria Silipo



Es war mir ein Vergnügen, vor kurzem Andrea De Mauro als Teil der

[Mein Data Guest](#)

Interview-Serie. Er hat Mythen über die Data Science Superhelden geholt, die IT vs. Frage der Datenwissenschaft und sprach über die Bedeutung, Ihre Hände schmutzig mit Daten und Algorithmen.

In den Warenkorb hat mehr als 15 Jahre internationale Erfahrung in der Verwaltung von Daten analytics und Data Science Teams mit verschiedenen organischen Organisationen. Derzeit ist er der Leiter der Business Intelligence bei Vodafone in Italien. Vorher diente er als Director of Business Analytics bei Procter & Gamble. Er ist Professor für Marketing Analyse und Angewandtes maschinelles Lernen an verschiedenen Universitäten, einschließlich der Internationale Universität Genf (Schweiz) und die Universitäten Bari und Florenz (Italien). Er ist auch Autor von populären Data Science Büchern und der Forschung Papiere in internationalen Zeitschriften.

Rosaria: Wie viele verschiedene Berufsprofile sehen Sie im Bereich Datenwissenschaft?

Andrea: Der traditionelle Mythos eines Datenwissenschaftlers als Superheld, der sich um ganze End-to-End-Prozesse oder die gesamte Landschaft von Komplexitäten rund um die Analytik ist weit von der Realität entfernt. Heute gibt es viele Rollen in der erstaunlichen Welt der Datenanalyse. Ich benutze normalerweise drei Hauptfamilien, um sie zu erklären:

Datenanalysten oder Unternehmensanalysten, die in einem bestimmten Unternehmen fundiertes Fachwissen haben Domain und "Übersetzen" Bedürfnisse zwischen anderen Daten-Praktizierenden und dem Geschäft Teams; Datenwissenschaftler, die sich mehr auf die Algorithmen und die Skalierung der Analysefähigkeiten und Data Engineers, die an der Umsetzung beteiligt sind und Wartung des gesamten Technologiestapels.

Rosaria: Welche professionelle Kategorie ist am schwersten zu rekrutieren?

Andrea: Sie sind alle hart zu rekrutieren für diese Tage! Aber ich denke, der Business Analyst Rolle ist die härteste: Menschen zu finden, die haben, was es braucht, um Geschäft zu erhalten

Komplexitäten, die mit den erforderlichen Algorithmen beantwortet werden, sind hart. Die Rolle ist auch schwierig den letzten Absolventen erklären.

Rosaria: Benötigen die verschiedenen Profile unterschiedliche Datenerziehung in Ihrer Meinung?

Andrea: Ich denke, alle diese Profile erfordern vor allem eine Sache: ein Wachstumsgeist – die Bereitschaft, das Lernen zu halten.

Keiner dieser Profis kann überleben, ohne offen zu sein, kontinuierlich zu lernen: dies gilt insbesondere für Wirtschaftsanalysten und Datenwissenschaftler. Dateningenieure benötigen auch einige vertikale technische Expertise auf einer oder mehreren großen Datenplattformen, wie GCP, AWS oder Azure. Als Data Engineer möchten Sie generell eine Zertifizierung haben. Die gute Nachrichten gibt es viele Möglichkeiten, zertifiziert zu werden.

Ich denke auch, es gibt ein sehr reichhaltiges MOOC-Angebot, um Datenwissenschaft online auf Plattformen zu lernen wie Coursera, Edx und andere Anbieter. Sie bieten Zertifizierungswwege für inspirierende Daten Wissenschaftler oder Analysten. Ich empfehle normalerweise Menschen, die keine Ausbildung haben Hintergrund für die Datenwissenschaft, um diese Lernplattformen zu nutzen.

Rosaria: Welche Tools glauben Sie, dass ein Data Professional lernen sollte?

Andrea: Es ist wirklich wichtig, dass ein anstrebender Datenprofi die richtige Menge an Werkzeuge und Optionen – und zu wissen, wie und wann sie zu nutzen. Sie müssen nicht alle Werkzeuge lernen, aber haben eine gute Mischung von Produkten, die sich als Teil ergänzen einer vielseitigen Toolbox.

Die Art von Toolkit würde ich empfehlen, besitzen würde ein Geschäft Intelligenz-Produkt, Fokussierung auf skalierte Dashboards und Datenvisualisierung Fähigkeiten und eine vielseitige Analyseplattform. KNIME ist ein großes Beispiel für Low-Code Analyseplattformen. Ich würde auch mehr traditionelle codebasierte Analyse-Tools enthalten, die perfekt mit Low-Code-Plattformen wie KNIME integriert werden kann.

Rosaria: Lassen Sie uns jetzt mit dem Lehrer sprechen, verwenden Sie KNIME, um Ihre Datenwissenschaft zu lehren Kurse?

Andrea: Natürlich! Ich habe KNIME schon seit langem an den Universitäten und bei der Arbeit mit Vodafone und P&G. Es ist ein erstaunliches Werkzeug, um Datenwissenschaft für mehrere Gründe. KNIME macht den Prozess der Codierung bequem, so dass Sie sich auf den Kern konzentrieren können analytische Aufgaben.

Diejenigen, die mit der Analytik beginnen möchten, fühlen sich oft durch ihr Fehlen entmutigt kodierende Fähigkeiten. Dafür bietet KNIME eine Lösung. Visual Tools wie KNIME lassen Sie „sehen“ und zu verfolgen, was auf jedem Schritt vor sich geht. Sie können leicht erkennen, wo das Problem ist, wenn Sie an einem bestimmten Punkt festhalten. Dies unterstützt wirklich die Bildungserfahrungen, während Lehrveranstaltungen zur Datenwissenschaft. Kurz gesagt, dies erhöht die Effizienz des Lernens Prozess für die Studenten.

Studenten schätzen es auch. KNIME macht das Lernen zugänglicher und manchmal auch mehr Spaß. Die Erfahrung des schrittweisen Aufbaus eines Workflows ist etwas angenehm für sie. Der Einsatz von KNIME-Knoten macht das Lernmodul und progressiv. Ein Knoten macht Sie „sehen“ was mit Ihren Daten im Fluss läuft sehr leicht. Der Joiner-Knoten kombiniert beispielsweise die beiden Eingangstabellen in eine einzige Ausgabettabelle. Oder die Loop-Knoten verwenden Iteration auf eine Abfolge von Schritten zwischen Start und Ende. Indem Sie es „visuell“ machen, verstehen Sie es besser und reduzieren die Chance Fehler machen.

Rosaria: Also reden wir jetzt mit dem technischen Profi. Sehen Sie sich selbst als Daten einen Analyten, einen Dateningenieur oder einen Datenwissenschaftler?

Andrea: Ich sehe mich mehr als ein Datenanalyst, direkt an der Kreuzung zwischen Geschäftsanforderungen, Geschäftsanforderungen, Daten und Algorithmen. Glücklicherweise hatte ich die Gelegenheit, das vollständige Bild und die Praxis Bits und Stücke aller drei Rollen zu sehen. wenn Ich musste mich entscheiden, ich würde mich mehr als Data Analyst sehen.

Rosaria: Wie wichtig ist die Umsetzung von Datenanalysen in einem Unternehmen, hilft es oder es ist nur eine akademische Übung?

Andrea: Es ist natürlich eine rhetorische Frage. Datenanalyse macht und wird einen großen Unterschied. Es ist ein Spielwechsler für das Geschäft. Es ändert das Betriebsmodell das definiert, wie das Unternehmen arbeitet. Schließlich ändert es die Art und Weise, wie Geschäft getan wird. Es ist nicht nur eine Technik oder eine IT-Komplexität zu verwalten, sondern eine neuartige Möglichkeit, eine Organisation. Emotional bringt es auch große Aufregung. Denken Sie daran, einen Prototypen analytische Lösung, die gut zu Ihrem Business-Fall passt oder tiefe Einblicke finden Sie nie erwartet. Erleben der scheinbar „magischen“ Datenanalyse ist eine potenzielle Moral Booster für alle.

Rosaria: Wie bauen Sie als Manager Ihr Data Science Team? Wo fangen Sie an aus?

Andrea: Fangen wir an, wo du nicht startest! Sie beginnen nicht, indem Sie Dutzende von generische Datenprofis ohne ersten Blick auf die Talente, die Sie bereits in Ihre Familie. Sie können auf jeden Fall Datenwissenschaftler aus den aktuellen Talenten wachsen Sie bereits haben in Ihrem Unternehmen, indem sie diejenigen, die Neugier, Leidenschaft und Bereitschaft haben zu lernen.

Nach dieser Vorgehensweise hat meiner Meinung nach zwei große Vorteile, die ich für richtig halte. erwähnenswert:

- Nur die Person mit Wissen und Verständnis, wie das Unternehmen funktioniert und wie Daten durch sie fließen, kann wirklich verstehen, Möglichkeiten und einige sinnvolle Datenanalysen erstellen.

- Es ist erfrischend für die Profis, egal was ihr Hintergrund ist, zu steigern ihren Karriereweg und ihre Entwicklung durch ernsthafte Datenanalyse. Das die Möglichkeit sollte nicht auf diejenigen beschränkt werden, die eine technische oder eine IT haben Hintergrund.

Rosaria: Über das IT- und Data Science-Team zu sprechen, ist dies eine Frage, die ich in Betracht gezogen habe. für eine lange Zeit: Wo sollte die Bereitstellung von Datenanalyseanwendungen liegen? Menschen argumentieren, ob es mit dem Data Science Team bleiben sollte oder eine IT sein sollte Verantwortung?

Andrea: Es ist schwierig, eine allgemeine Antwort zu haben, da sie wirklich von jedem Fall abhängt und wie das Unternehmen organisiert ist, aber eine Sache ist sicher, dass es eine starke Zusammenarbeit zwischen IT- und Data Science-Teams. Ob diese Teams getrennt oder zusammen hängt davon ab, wo das Unternehmen in Bezug auf die Laufzeit und von viele andere Faktoren, manchmal kraftgetrieben und politisch.

Wenn sie getrennt sind, ist es notwendig, dass jedes Team den Datenlebenszyklus versteht Prozess. Sonst ist es unhaltbar. Kurz gesagt, es ist inhärent eine Zusammenarbeit.

Das Eigentum an den Fähigkeiten und deren Nutzung sollte jedoch in der - weder im Data Science-Team noch im IT-Bereich.

Rosaria: Jetzt spielen wir ein bisschen Mythenbüste auf dem Gebiet der Datenanalyse. Was ist das? Mythen rund um die Datenanalyse, die auf dem Technologemarkt vorherrschen und denken Sie, dass sie Ungerechtfertigt?

Andrea: Datenanalyse wurde mit einem starken Gefühl von Euphorie um sie herum geboren. Das hat führte zur Schaffung vieler Mythen:

Zunächst wurden, wie wir bereits erwähnt haben, einige Leute dazu geführt, zu denken, dass es genug ist, ein gutes Team von starken Datenwissenschaftlern, um mit dem vollen Bedarf zu bewältigen. Das ist ein Mythos. Sie benötigen ein facettenreiches Team von Datenexperten, arbeiten Hand in Hand mit Geschäftspartner. Sie benötigen auch ein Toolkit aus mehreren Werkzeugen, um zu aktivieren das Team, um auf seinem Höhepunkt.

Zweitens behaupten einige, dass der Prozess der Datenwissenschaft vollständig automatisiert werden wird. Wir hören viel über AutoML, die eine wichtige Richtung im maschinellen Lernen und künstlich ist Intelligenz. Es schafft jedoch den Mythos, dass in Zukunft Menschen nicht benötigt werden und Maschinen übernehmen den Prozess, der nicht wahr ist. Die KI, mit der wir zusammenarbeiten Im Moment ist nicht gemein. Es ist eher in der Lage, spezifische Probleme zu lösen und Komplexitäten, die nur mit menschlicher Führung angetrieben werden können. So wird es immer eine Zusammenarbeit zwischen Mensch und Maschine.

Rosaria: Lassen Sie uns ein bisschen über das jüngste Buch, das Sie schrieb [Datenanalyse Ausverkauft](#)
Für wen ist das Buch? Ist das Buch genug, um Datenanalysen zu verstehen und sogar anzuwenden
Es?

»

Andrea: Ich würde sagen, es ist für alle, die mit Daten arbeiten oder arbeiten möchten.

Die Absicht des Buches ist, „Datenanalysen einfach zu machen“. Es gibt praktische Einblicke in wie man bestimmte Aufgaben wie die Erstellung eines Berichts oder die Automatisierung einer Folge von Schritte, die Sie heute in Excel laufen oder eine überzeugende Präsentation erstellen, eine Geschichte erzählen mit Daten. Es ist für Wissensarbeiter und für ihre Manager. Es ist wichtig für die letztgenannte, um bewusst zu sein, was die Datenanalyse alles ist und erste Person über das, was sie fragen von ihrem Team, wenn es um Daten geht. Es ist auch für diese Studenten und Absolventen, die eine Karriere in der Datenanalyse starten möchten.

Jeder, der das Buch liest, kann Datenanalysen in der Praxis durchführen. Die Anleitungen basieren auf Beispielen, die auf viele verschiedene Geschäftsfälle zurückgegriffen werden können. Du muss nicht in einer bestimmten Rolle sein oder einen bestimmten Hintergrund haben, um die Inhalt dieses Buches. Letztlich kann das Buch Ihnen versichern, dass Sie „kann“ lernen Analytik wenig bis wenig und autonom genug, um Daten zu setzen eine Analyse in der Praxis am Arbeitsplatz.

Rosaria: Schließlich möchte ich wirklich wissen, welche Berufsberatung Sie einer Aspiring Data Professional?

Andrea: Warten Sie nicht auf Ihren ersten Job, um die Erfahrung zu bekommen, die Sie brauchen. Sie können starten Ihre Hände schmutzig mit Daten und Algorithmen gut vor Ihren ersten Interviews!

Mein Rat wird immer sein, zu gehen und nach den richtigen Möglichkeiten um Sie zu suchen bewerben Analytiker Erstperson. Suchen Sie nach lokalen Wohlfahrten, die Hilfe mit ihren Daten benötigen. Schauen Sie sich Webseiten wie Kaggle an, um kostenlose Online-Wettbewerbe zu finden und Erfahrung zu sammeln. Dies bietet Ihnen die Möglichkeit, Ihr eigenes erstes Portfolio an Modellen aufzubauen und erfolgreiche Datenanalyseanwendungen.

Dieser Artikel wurde erstmals auf unserer Website veröffentlicht [KNIME Blog](#). Die Originalversion finden [Hier](#).

Beobachten Sie das ursprüngliche Interview mit Andrea De Mauro auf YouTube [Mein Data Guest – Ep. 3 mit In den Warenkorb](#)

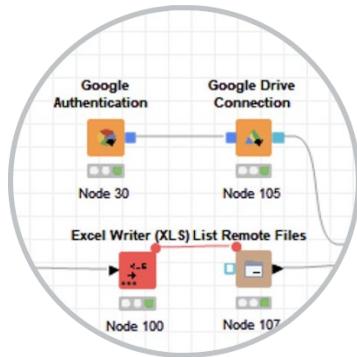


Das ist nicht alles. wurde nominiert KNIME-Beitrag
der Monat für April 2021. Er wurde für seine vielen ausgezeichneten
Artikel, wo er seine Erfahrungen, Rezepte und
Best Practices zur Analyse von Daten mit KNIME Analytics
Plattform - allein und in Kombination mit anderen Werkzeugen.
In den letzten Monaten vor seiner Nominierung, Dennis wirklich
hat gefeuert. Er hat erfolgreich vorhergesagt
alles, wie Nostradamus und der Oktopus Paul, aber

Verwendung von Data Science-Modellen und KNIME Analytics Platform. Zuerst prophezeite er die
Kurve für die COVID Spread weltweit : dann brach er es nieder, um die COVIDIERUNG
Streuland nach Ländern ; und schließlich hat er sogar das Ergebnis dieses Juli vorhergesagt
Europäische Fußballmeisterschaft, UEFA Euro 2020 , mit Hilfe eines neugierigen
Charakter: Yodime. Aber er sagt nicht nur Dinge vorher. Er hat auch geschrieben
Meinungsstück, wie und warum KNIME zum wichtigsten Werkzeug seiner
Tagesarbeit (siehe: Das beste Werkzeug für die Datenmischung ist in meiner Meinung KNIME) In
die Figur rechts, Sie können einen Workflow-Snippet von Dennis sehen' COVID-Spread
weltweiter Artikel.

Dennis ist Data Scientist mit über 20 Jahren
Erfahrung. Er ist derzeit Head of Report und Data-
Management bei Swisscom AG Schweiz. Für die
hält einen Abschluss in Psychologie und Informatik
von der Universität Zürich.

Besuchen Sie Dennis Raum auf dem KNIME Hub oder Profil
Seite im KNIME Forum (Hub/Forum Griff:
Deganza)



Spot auf Euro 2020 Vorwürfe: Zurück zu Classic Techniken

Mein Data Guest – Ein Interview mit Dennis Ganzaroli

Autor: Rosaria Silipo



Es war mein großes Vergnügen, Gastgeber [Das ist nicht alles.](#) in der ersten Episode der [Meine Daten](#)
Gastgewerbe Interview-Serie, die am 22. September 2021 ausgestrahlt wurde. In diesem Gespräch, Dennis teilte einige der Geheimnisse und Best Practices, die dazu führen, ein erfolgreicher Datenwissenschaftler: Zunächst einmal die Leidenschaft für die Daten als die nötige Tool, um Ihren Data Science-Beruf zu unterstützen!

Wir müssen auch einige seiner neuesten Errungenschaften diskutieren, wie die (richtige) Vorhersage der endgültigen Spieler des UEFA Euro 2020 Turniers oder der (richtigen) Schätzung der Die Entwicklung des COVID-19 verbreitete sich auf der ganzen Welt. Nicht zuletzt haben wir sogar einige Empfehlungen zu lesen, um auf dem Laufenden zu halten Entwicklung des Bereichs der Datenwissenschaft.

Dennis arbeitet seit 20 Jahren mit Daten als Data Engineer, Daten Wissenschaftler und manchmal als Datenanalytiker. Wenn Sie mit Dennis reden, erkennen Sie, dass sein Wissen ist viel mehr als "nur" seine professionelle Umgebung. Seine Leidenschaft für Daten nehmen ihn weiter als seine tägliche Routine. In der Tat wendet er klassische und moderne Datenwissenschaft Algorithmen, um den COVID-19-Spread oder den Gewinner der UEFA Euro vorherzusagen 2020 Fußballturnier. Die Schönheit ist in der Tat diese Vielseitigkeit, diese Fähigkeit, Daten anzuwenden Erfahrung zu jedem Aspekt des Lebens. Daten sind schließlich Daten: Alles kann in Zahlen, inspiert und vorhergesagt.

Rosaria: Hallo Dennis, erzählen Sie uns von Ihrem professionellen Selbst und was Sie in Ihrem Job tun.

Dennis: Ich arbeite für ein großes Telco in der Schweiz als Head of Reporting und Data Management. Wir messen die Leistung der Vertriebskanäle und tun alles von der Datenintegration und Datenvermischung bis zur Erstellung von Dashboards, Berichten und So weiter.

Rosaria: Benutzen Sie KNIME in Ihrer täglichen Arbeit?

Dennis: Ja. Wir nutzen die KNIME Analytics Platform vor allem als ETL-Tool und verwenden auch KNIME Server zur Automatisierung unserer Workflows. Wir haben eine Menge Tagesberichte, die am Morgen bereit sein, so sind wir glücklich, eine solche Lösung zu haben. Wir kombinieren auch KNIME mit anderen Werkzeugen - hauptsächlich mit Tableau. Aber ich sage immer den Beteiligten: Tableau ist nur die Karosserie, KNIME ist der eigentliche Motor.

Rosaria: Erzählen Sie uns von der größten Herausforderung, die Sie in Ihrem Berufsleben lösen mussten.

Dennis: Die größte Herausforderung war und wird immer sein, die Geschichte hinter der Daten an Kunden und Interessenvertreter. Data Science dreht sich alles um Geschichten. Sie brauchen stark Kommunikationsfähigkeit und eine gute Visualisierung der Daten. Wie sie sagen: „Ein Bild ist tausend Worte. „

Rosaria: Es besteht eine hohe Nachfrage nach Datenwissenschaftlern auf dem Arbeitsmarkt und doch Datenwissenschaft ist noch nicht Teil des traditionellen Bildungssystems. Menschen müssen oft Lernen Sie Fähigkeiten selbst, melden Sie sich für Online-Kurse und lesen Sie die Literatur. Welche Bücher würden Sie Menschen empfehlen, die neue Fähigkeiten lernen wollen? Sie schreiben auch Ihr eigenes Buch. Bist du nicht?

Dennis: Ich mag dein Buch, Rosaria, über [Codeless Deep Learning](#) . Es ist leicht verstehten und sehr nützlich. Aber ja, ich habe gerade angefangen, mein eigenes Buch mit dem Titel zu schreiben „KNIME Lösungen für reale Weltanwendungen „. Es ist eine Zusammenstellung realer Fälle gelöst mit KNIME zusammen mit anderen Werkzeugen. Ich lese auch regelmäßig Blog-Artikel, um zu halten bis heute. Zum Beispiel die Zeitschriften [Vorwärts Datenwissenschaft](#) , [Niedriger Code für fortgeschrittenen Datenwissenschaft](#) auf Mittel , und alles, was in der Datenwissenschaft angewendet werden kann.

Rosaria: Was ist Ihr Rat für inspirierende Datenwissenschaftler?

Dennis: Wann immer mir jemand diese Frage stellt Ich frage zurück: Was sind Ihre Hobbys? Und wenn die Datenwissenschaft nicht Ihr Hobby ist – Sie müssen Hobbys ändern! Ich glaube, „Lernen“ allein reicht nicht aus. Du musst es leben und es lieben, um erfolgreich zu sein.

Rosaria: Welche Fähigkeiten werden am meisten von Kandidaten unterschätzt, aber ein Plus auf dem Job?

Dennis: Um cool in Stresssituationen zu halten und nie vergessen, dass es ein Job ist und kein Spiel. Obwohl ich stark glaube, dass die Datenwissenschaft Ihr Hobby werden muss [wenn Sie wollen die Arbeit ist kein Hobby. Viel Zeit wirst du Dinge tun, die Sie mögen nicht, aber das sind immer noch sehr wichtig.

Rosaria: Lassen Sie uns darüber sprechen, wie Sie Ihre Expertise mit Daten außerhalb Ihrer Berufsleben. Am 10. Juni, einen Tag vor Beginn der UEFA Euro 2020 Fußball Turnier, Sie konnten richtig vorhersagen, was das letzte Spiel wäre: England vs Italien. Wie hast du das gemacht?

Dennis: Ich fragte Maradona . Nein, ich habe einen ziemlich bekannten Ansatz im Sport benutzt. Wettindustrie. Ich habe ein lineares Regressionsmodell verwendet, um die Bewertungen der Teams zu berechnen.

Rosaria: Aber das Modell sagte, England wäre der Gewinner gewesen.

Dennis: Nicht genau, das Modell berechnet gerade die Leistungsbewertungen der Teams vor das Turnier. Also, England, Italien und Spanien waren in den ersten drei. Übrigens, alle drei Teams machten es zu den Halbfinalen. Obwohl Dänemark eine Überraschung war. Niemand sah, dass kommen, nicht einmal die Fußballexperten. Es ist hier wichtig zu bemerken, dass Fußball ein Spiel ist mit viel Zufall. Scores sind sehr sparsam und ein letztes Spiel kann von einen Strafausstoß. Daher ist es nicht immer möglich, eine präzise Prognose zu erstellen. Alle in all, Ich denke, meine Power-Ratings waren gut, weil die letzten Spiele, die ich als Trainingsdaten, spiegelten die Stärke der Teams gut wider.

Rosaria: Es war also ein ganz einfaches Modell. Kein tiefes Lernen?

Dennis: Genau, kein tiefes Lernen, keine starken GPUs, nur eine einfache lineare Regression Modell.

Der Schlüsselfaktor war hier, Domänenwissen einzubeziehen. Zum Beispiel das Heimfeld Der Vorteil ist ein sehr wichtiger Faktor im Fußball. Selbst ohne Zuschauer – es ist noch da. Ein weiterer Punkt ist, dass Sie genau den richtigen Teil der Daten nehmen müssen - der Teil, der ist am besten das Modell zu trainieren. Zum Beispiel: Nach einem großen Turnier wechseln sich die Trainer oft und die Spieler ändern sich auch. So ist es besser, herauszufiltern, die Spiele passieren richtig bevor solche Änderungen in den Teams vorgenommen werden.

Rosaria: Was ist mit einem anderen Projekt von Ihnen, wo Sie die Verbreitung vorhergesagt COVID19 weltweit und nach Ländern?

Dennis: Meine Motivation hinter diesem Projekt war, die Entwicklung der COVID19 Pandemie. Anfangs nur für China und dann für jedes Land im Welt. Ja, ich wollte die Frage beantworten: Wann wird diese Pandemie vorbei sein?

Rosaria: Welches Modell haben Sie benutzt?

Dennis: Die Entwicklung einer Pandemie ist wie ein Wachstumsprozeß. Am Anfang ist es ein exponentielle Funktion, dann ändert es sich an einer sigmoidalen Kurve. Dies wird am besten durch eine logistische Funktion. Doch dann, wenn mehrere Wellen folgten, dieser einfache Ansatz funktioniert nicht mehr und ein anderer Ansatz wird benötigt. Ich habe herausgefunden, dass Rockefeller Universität hatte bereits eine Methode namens Log-Analyse (auch Wavelets genannt) in die späten 90er Jahre, um die Entwicklung von mehreren überlappenden logistischen Funktionen zu prognostizieren.

Rosaria: War dieses Projekt mit der KNIME Analytics Platform oder mit Jupyter?

Dennis: Ich habe die KNIME Analytics Platform zusammen mit Jupyter verwendet. In der Tat können Sie anrufen Python von KNIME. So wurden die Daten in der KNIME Analytics Platform erstellt, loglet model wurde in Jupyter mit der scipy-package berechnet.

Rosaria: Danke. Dennis, für diesen Einblick in Ihren Job und Ihre anderen Projekte. Wie Können Datenwissenschaftler im Publikum mit Ihnen oder Ihrer Arbeit in Kontakt treten?

Dennis: Ich habe Artikel über Medium geschrieben und ich werde auch einige interessante Sachen auf [Linkedin](#) und [Twitter](#). Ich habe auch eine Facebook-Gruppe geschaffen [Data Science mit Yodime](#) Und auf meine [Youtube Kanal](#) Sie finden einige interessante Videos zu Data Science. und nicht nur. Selbstverständlich können alle meine Workflows von [mein öffentlicher Raum auf Der KNIME Hub](#)

Dieser Artikel wurde erstmals auf unserer Website veröffentlicht [KNIME Blog](#). Die Originalversion finden [Hier.](#)

Sehen Sie das Originalinterview mit Dennis Ganzaroli auf YouTube: “ [Mein Data Guest – Ep 1 mit Das ist nicht alles.](#)”



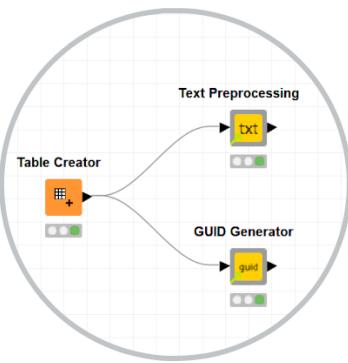
SJ Porter wurde nominiert Beitrag des Monat für Oktober 2020. Er wurde für seine Generator und Textverarbeitung Komponenten, die in der Figur rechts angezeigt. Ab 28.09.2022 beide Komponenten Gesamt ANHANG Downloads. Die erste Komponente war schließlich ein Globally Unique Identifier (GUID) Generator für KNIME. Diese Komponente ist nützlich für Erstellen eines einzigartigen Schlüssels, der nicht auf Row ID basiert. Die

GUID

Textverarbeitungskomponente verwendet extrem schnelle Textverarbeitung Funktionen, um bestimmte Zeichentypen aus einer String-Säule zu entfernen und normalisieren Sie die Daten so weit wie möglich ohne Überverarbeitung. Diese Komponente eliminiert die Notwendigkeit, Text in einen Dokumenttyp umzuwandeln, um ihn vorzubearbeiten.

Zum Zeitpunkt der Auszeichnung war SJ Data Science Team Lead im Verbraucher Berichterstattung. Im Januar 2021 trat er als Data Scientist bei KNIME und macht Er ist ein offizieller KNIMER. Er arbeitet derzeit als Site Reliability Engineer mit Fokus auf den KNIME Edge und KNIME Community Hub Architektur. Neben Gebäude-Workflows und Entwicklungskomponenten, SJ auch geholfen, die Niedriger Code für fortgeschrittenen Datenwissenschaft Zeitschrift auf Medium und aktuell dient als einer der Editoren.

Besuchen Sie SJ's Raum auf dem KNIME Hub oder Profil Seite im KNIME Forum (Hub/Forum Griff:
Sjporter)



KNIME Analytics Platform ist die „Killer-App“ für maschinelles Lernen und Statistik

Ein kostenloses, einfaches und Open-Source-Tool für alle Daten? Ja.
Bitte!

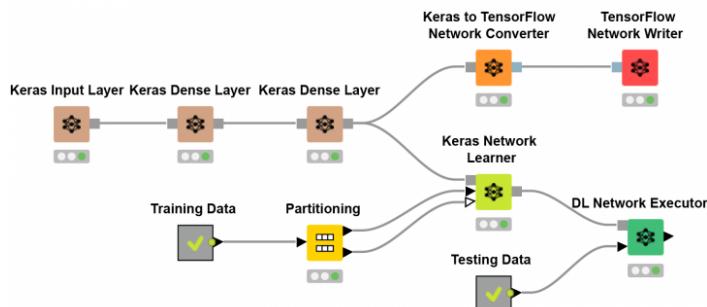
Autor: SJ Porter

Disclaimer:

Ich arbeite jetzt für KNIME als Data Scientist / Data Engineer! Ich schrieb diesen Artikel ungefähr ein Jahr vor I auf KNIME angewendet. Der Artikeltitel und die Inhalte waren (und sind immer noch) meine persönliche Meinung.

– Steven “SJ” Porter

Wenn Sie mit Daten in jeder Kapazität arbeiten, gehen Sie voran und tun Sie sich einen Gefallen: herunterladen [KNIME Analytics Plattform rechts](#) [Hier.](#)

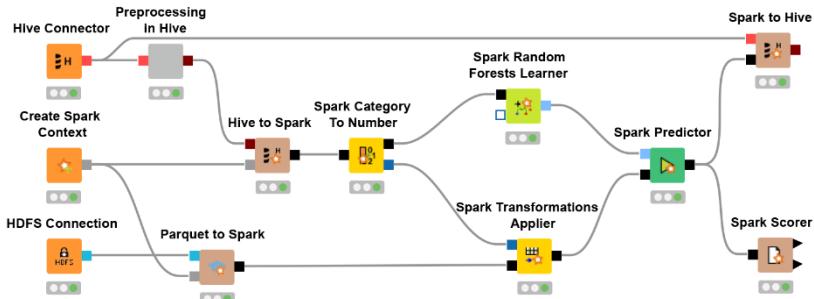


Mehr Datenwissenschaft, weniger Slamming von Maus und Tastatur.

Was ist KNIME Analytics Platform?

KNIME Analytics Platform ist die stärkste und umfassendste kostenlose Plattform Drag-and-Drop-Analysen, maschinelles Lernen, Statistiken und ETL, die ich bisher gefunden habe. Die Tatsache, dass es weder eine paywall noch gesperzte Funktionen bedeutet, dass die Barriere zum Eintritt ist nicht vorhanden.

Steckverbinder mit Datenquellen (beide On-Premises und Cloud) sind für alle verfügbar große Anbieter, so dass es einfach ist, Daten zwischen Umgebungen zu bewegen. SQL Server nach Azure? Kein Problem. Google Sheets zu Amazon Redshift? Klar, warum nicht. Wie geht's? Anwendung eines maschinellen Lernalgorithmus und Filterung/Transformation der Ergebnisse? Du bist abgedeckt.



Big Data? Kleine Daten? Ist nicht zu viel, vorausgesetzt, dass Ihr Computer einen soliden CPU und 16 GB RAM.

Es ist auch erwähnenswert, dass die Gemeinschaft besonders robust ist. Die Entwickler und Produktbesitzer bei KNIME sind eine Explosion zu arbeiten, und die Foren sind überraschend aktiv.

Ich könnte in ein schnelles Start-Tutorial tauchen oder einige der fortgeschritteneren zeigen Fähigkeiten, aber es ist ehrlich sehr intuitiv zu bedienen. Die Dokumentation ist integriert direkt in den Desktop-Client, die UI ist Schmutz einfach, und die UX ist eine tolle Mischung zwischen Komplexität/Kundheit (wenn nötig) und Benutzerfreundlichkeit.

Was kann ich mit der KNIME Analytics Platform machen?

KNIME Analytics Platform ist für folgende Zwecke geeignet:

- ETL-Prozesse (Bewegung von Daten von hier nach dort und Reinigung)
- Lernen von Maschinen
- Deep Learning
- Natürliche Sprachverarbeitung
- API Integration
- Interaktive visuelle Analyse (etwa ein Beta-Feature)

Was ist der Fang?

KNIME Analytics Platform ist 100% kostenlos. Die Dokumentation ist leicht verfügbar bei knime.com und es gibt eine Tonne kostenlose Erweiterungen auf der Plattform. Solange du die bist ein Klick „Los!“ jedes Mal, wenn der Prozess läuft, müssen Sie keine Dime bezahlen. Niemals. Das ist der Vorteil der Zusammenarbeit mit einem Softwarepaket, das Wurzeln in der Wissenschaft hat.

Wenn Sie Workflows planen möchten, KNIME Server ist das Premium-Angebot, das es erlaubt für Prozessplanung unter anderen Features. Die grundlegende tier, KNIME Server Small, ist um [\\$1.67 pro Stunden bei AWS](#). Wenn Sie KNIME Server auf einer EC2-Instanz hosten und

Planen Sie ein [Cron Job](#) die Instanz ein- und ausschalten, es ist ein extrem kostengünstig Option.

Höhere Fliesen von KNIME Server ermöglichen die Nutzung der [REST API](#) und [Webportal](#) . Die Funktionen ermöglichen es Ihnen, die Workflow-Bereitstellung zu automatisieren, Workflows remote auszuführen aus einem anderen Dienst, und erstellen Sie einen interaktiven Hub für Benutzer. Die Automatisierungsfähigkeit Workflows macht KNIME Server Medium eine attraktive Option. Wenn Sie kaufen höchste tier (KNIME Server Large) mit der BYOL-Option, Sie gewinnen die Fähigkeit zu hosten mehrere Instanzen des Servers mit der gleichen Lizenz.

Wie erfahre ich die KNIME Analytics Platform?

Ihre Lernseite ist [Hier](#), aber wenn Sie daten-savvy dann einfach herunterladen die Plattform und versuchen, ein oder zwei. Solange Sie an den Punkt gelangen, an dem einige Daten geladen werden bis in die Anwendung, der Rest ist intuitiv und die Dokumentation ist von innerhalb der Anmeldung. Drag-Knoten aus dem " Excel " auf die Ansicht. Es gibt eine Suchleiste... versuchen, nach " Excel " und gehen von dort.

Dieser Artikel wurde erstmals in der [Vorwärts Zeitschrift für Statistik](#) auf Medium. Finden Sie die Originalversion [Hier](#).



[Das ist der Hammer](#) wurde nominiert Beitrag des Monat für November 2021. Er wurde für seine [Kategorische Eigenschaften Einbettung](#), [Autofeat Generator](#) und [Autofeat Apply](#) Komponenten. Ab 03.10.2022 Komponenten insgesamt 3.171 Downloads. Das sind nur zwei viele Komponenten Ashok entwickelt, die gute Beispiele, wie man Python-Skripte bündelt und teilt ohne Abhängigkeitsprobleme. Alle seine Komponenten sind

Zugang zu seiner KNIME Community Hubprofil mit Gebrauchsanweisung

sie in der Praxis. Besuchen Sie die [Community Component Highlights](#)

Abschnitt auf der

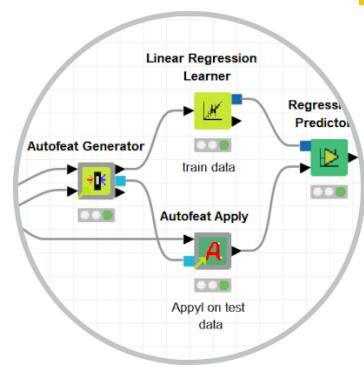
Verified Component Webseite, um mehr zu erfahren.

Ashok hat mehrere Jahre Erfahrung in der Lehre Big Data Technologie und ist derzeit Professor an FORE School of Management in Neu Delhi. Sein Interessengebiete rund um Big Data Systems, einschließlich Machine Learning, Deep Learning, Big Data Speichersysteme und Graph Datenbanken.

Besuchen Sie Ashoks [Raum auf dem KNIME Hub](#) oder [Profil](#)

[Seite im KNIME Forum](#) (Hub/Forum Griff:

Ashokharnal)



Eine Komponentenserie für automatisiertes Feature Ingenieurwesen

Alles gut Die Dinge kommen in Threes – Ashoks Serie
Zubehör und Zubehör

Autoren: Elisabeth Richter & Corey Weisinger

Wie bereits erwähnt, ist Ashok ein sehr aktives Mitglied der KNIME-Gemeinschaft, das gerne Komponenten entwickeln. Er wurde für seine Automatische kategorische Eigenschaften Einbettung Autofeat Generator und Autofeat Apply Komponenten. In diesem Artikel haben wir einen näheren Sehen Sie sich diese drei Komponenten an.

Die Auto-Kategorie Eigenschaften Einbetten Komponente

Die Automatische kategorische Eigenschaften Einbettung Komponenten-Encodes kategorische String Funktionen in numerische Funktionen. Im Allgemeinen, Feature Encoding ist eine Technik, die häufig in der Maschine verwendet wird Lernen. Es bezieht sich auf den Prozess der Umwandlung eines kategorischen (d.h. nicht numerisch) in eine kontinuierliche (d.h. a Variable, die jeden Wert zwischen zwei Punkten annehmen kann. Diese von Ashok entwickelte Komponente automatisiert diese Funktion Kodierungsprozess.

Als Eingaben nimmt die Komponente Trainings- und Testdaten ein. Die Konfigurationsfenster der Komponente wird in der folgenden Figur angezeigt und erlaubt die Angabe der Zielspalte, der zu kodierenden Stringspalte(n), welche Art der kategorischen Kodierung sollte durchgeführt werden, und ob PCA durchzuführen oder nicht. Die zu erstellende Features sind entweder Zählcodierung, sortierte Etikettenzählcodierung, Ziel Kodierung, oder alle drei. So, mit dieser Komponente, mehrere kategorische Variablen kann in einem Schritt angegeben werden.

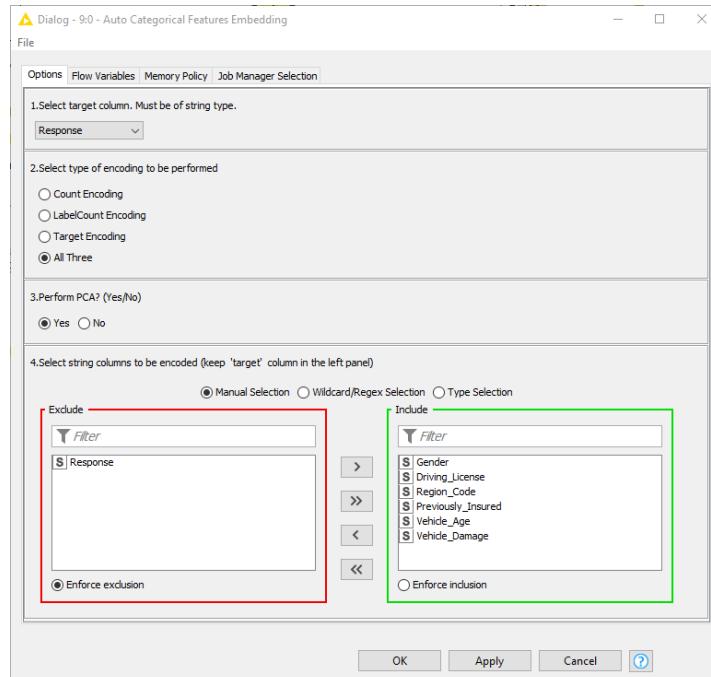
Um Datenleckage bei der Durchführung der Zielcodierung zu vermeiden, wird die Kodierung durchgeführt nur Trainingsdaten verwenden. Die codierten Werte werden dann auf die Testdaten abgebildet. Zu bekommen eine zuverlässige Zielcodierung, der Datensatz sollte ausreichend groß genug sein.

Die Durchführung einer PCA ist optional. Wenn „Ja.“ wird im Konfigurationsfenster ausgewählt, eine Komponente führt einen PCA auf dem Datensatz aus, der entweder aus diesen drei oder allen drei Codierungen zusammen mit anderen numerischen Merkmalen, die bereits in den Daten vorhanden sind. Das PCA-Modell wird auf den Trainingsdaten aufgebaut und dann auf die Testdaten angewendet.



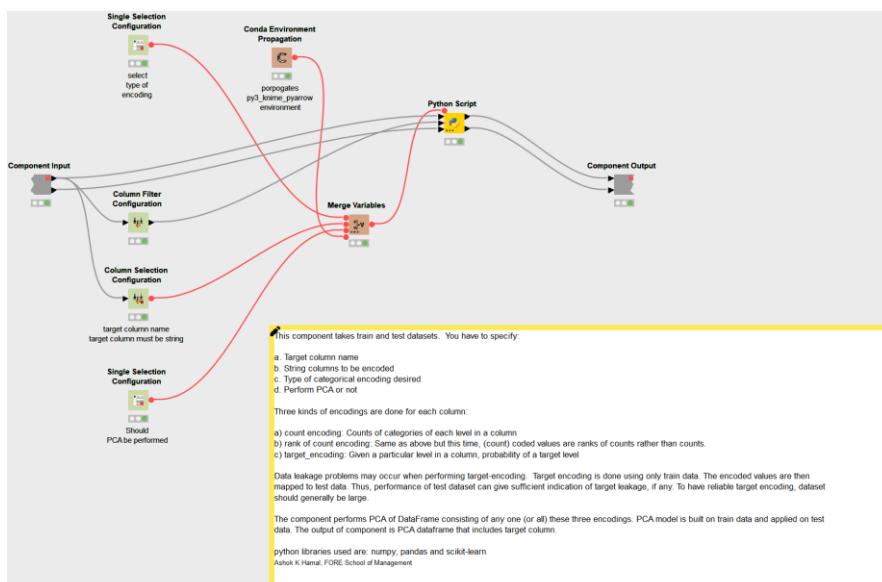
Die Auto-Kategorie
Eigenschaften Einbettung
Bauteil für
kategorischer String
Kodierung.

Eine Komponentenserie für Automatisierte Feature Engineering



Der Konfigurationsdialog des Auto-Kategorischen Feature Embedding
eine Komponente.

Die Ausgabe der Komponente sind zwei Datentabellen – die Trainingsdaten und die Testdaten.
Jede Tabelle enthält entweder die Hauptkomponenten oder die kodierten Spalten entlang
mit den bereits vorhandenen numerischen Spalten und der Zielspalte.



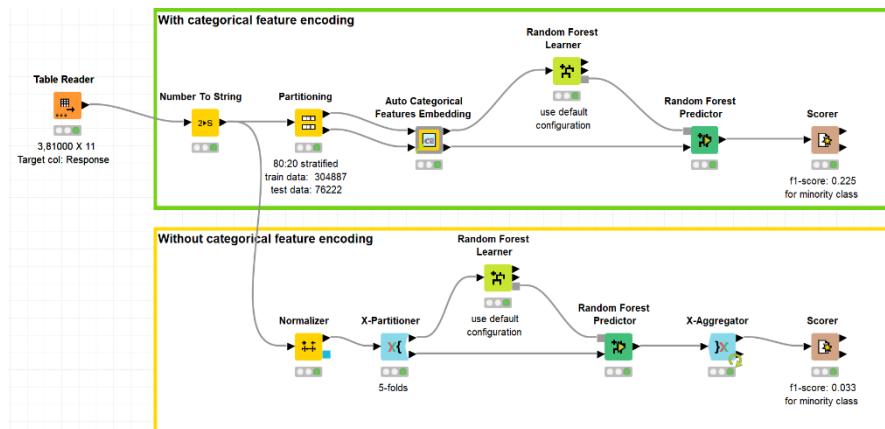
Die Innenseite der Auto Categorical Feature Embedding Komponente.

Im Inneren der Komponente (siehe Bild oben), Ashok fügte ein Python Script Knoten, nutzt die Python Bibliotheken NumPy, Pandas, scikit-learn, und Pyarrow. Bei der Ausführung der Komponente, Ashok stellt sicher, dass KNIME automatisch den erforderlichen Python installiert Pakete über die Conda Umweltausbreitung Knoten.

Das folgende Bild zeigt einen Workflow, der die Verwendung des Auto-Kategorie Eigenschaften Einbettung einer Komponente. Es kann auf Ashoks Hub-Profil gefunden werden ([HealthInsurance Cross-Sell-Kategorie Feature Engineering](#)) Die Daten in diesem Beispiel sind einige persönliche Daten der Kunden eines Versicherungsunternehmens. Das Unternehmen derzeit eine Krankenversicherung für ihre Kunden. Verwendung der vorhandenen Daten, möchte nun vorhersagen, ob die aktuellen Kunden auch an einer neuen Fahrzeugversicherung.

Die Zielspalte ist die "Antwort" Spalte, und der Datensatz ist mit nur ~12% der Kunden bereit, die Fahrzeugversicherung zu kaufen (Responsewert gleich 1)

Nun werden zwei Modelle mit identischer Konfiguration der Lernknoten trainiert: eins mit kategorischer Funktion Kodierung (oben) und eine ohne kategorische Funktion Kodierung (unten)



Ein Beispiel Workflow, der die Verwendung der Auto Categorical Features Embedding demonstriert eine Komponente.

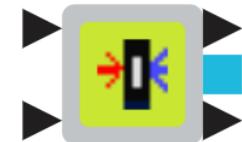
Die jeweilige Scorer Knoten zeigen, dass die F1-Score sowie die Cohen's Kappa sind viel besser bei der Anwendung von Feature-Kodierung. Die Spitze Scorer node meldet eine F1 score 0,2 und ein Cohens Kappa von 0,138, während der Boden Scorer Knoten meldet eine F1 nur 0,034 und ein Cohens Kappa von 0,025.

Anhand der Ergebnisse dieses Beispiels wird der Zweck der Merkmalscodierung unterstrichen: Verbesserung der Klassifizierungsergebnisse.

Der Autofeat Generator und die Autofeat Apply

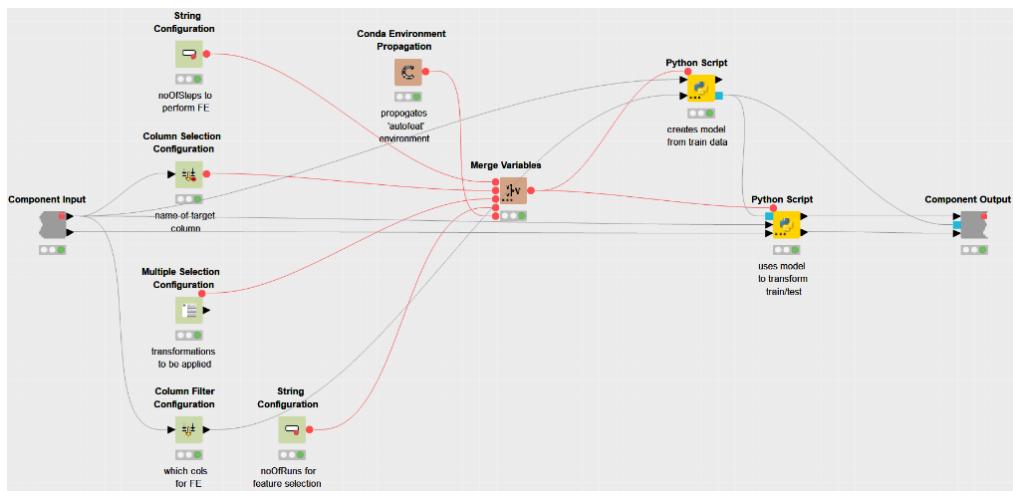
Komponenten

Ashok scheint ein Python-Enthusiast zu sein – zumindest wie oft er verwendet KNIME Python Integration wenn Entwicklung von Komponenten. Die beiden eingesetzten Komponenten den folgenden Abschnitt, Autofeat Generator und Autofeat Anwendung, beide Bündel Python-Skripte, um die zusammen mit NumPy und Pandas. Beide Komponenten enthalten auch die Python Script Knoten und der Conda Umweltausbreitung Knoten, ähnlich wie die Automatische kategorische Eigenschaften Einbettung eine Komponente.



Der Autofeat Generator
eine Komponente.

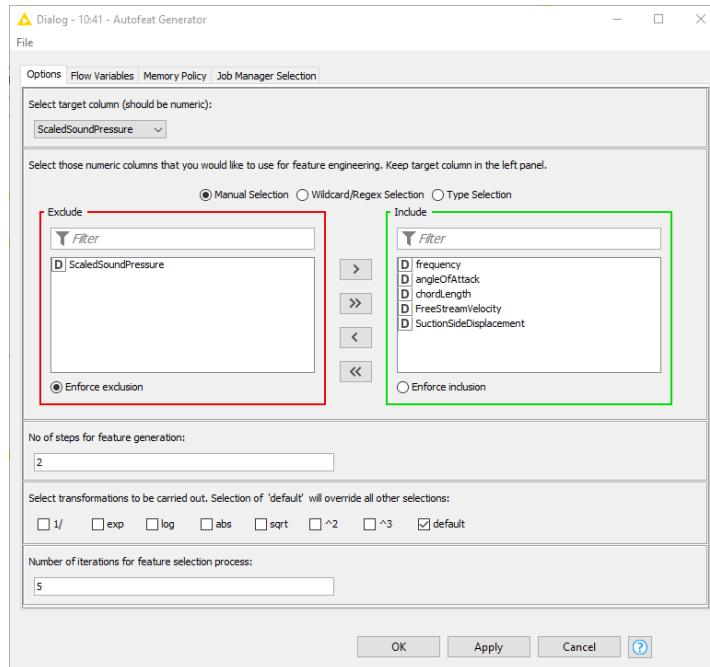
Die Autofeat Generator Komponente erzeugt neue Funktionen, deren Verwendung gerichtet ist zum Aufbau von linearen Modellen. Es nimmt als Eingabe zwei Datentabellen – Training und Test Daten – und baut ein Modell mit den Trainingsdaten. In einem zweiten Schritt das ausgebildete Modell wird dann zur Erzeugung zusätzlicher Spalten auf Basis der bereitgestellten verwendet. Die Komponente gibt die beiden Datentabellen sowie das ausgebildete Modell aus. Das Innere des Eine Komponente ist in der nachfolgenden Abbildung dargestellt.



Die Innenseite der Autofeat Generator-Komponente.

Der Konfigurationsdialog der Autofeat Generator In der die folgenden: Es erlaubt die Spezifikation der Zielspalte (die numerisch sein sollte), welche Spalte(n) im Prozess enthalten sein soll, die Anzahl der Schritte für das Merkmal Erzeugung (die ein wichtiger Parameter ist, je höher die Anzahl der Schritte ist, höher die Anzahl der Merkmale, aber auch eine höhere Wahrscheinlichkeit der Überbelegung), die Anzahl läuft für den Merkmalsauswahlprozess und die auf die Merkmale“ Standard” überschreibt alle anderen Auswahlen).

Eine Komponentenserie für Automatisierte Feature Engineering



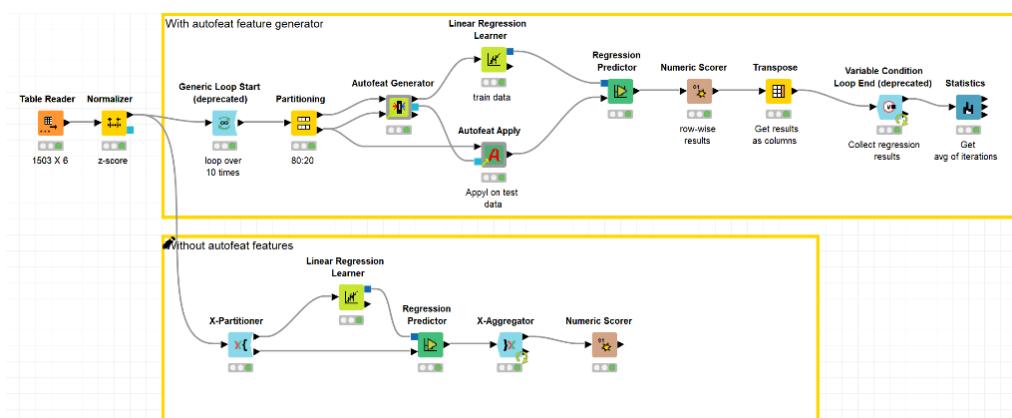
Das Konfigurationsfenster der Autofeat Generator-Komponente.

Die Autofeat Apply eine ähnliche Funktionalität aufweist, jedoch baut es kein Modell. Es nimmt als Eingabe einer Datentabelle und ein “ Autofeat ” Modell (das könnte ein Modell von der Autofeat Generator Komponente) und gibt einen Datensatz mit erzeugte Merkmale, die mit dem Eingabedatensatz konkatiniert sind.

Dieses automatisierte Feature Engineering kommt praktisch, wenn Optimierung der Leistung von maschinellen Lernalgorithmen wie logistische oder lineare Regressionen.



Die Autofeat Apply eine Komponente.



Ein Beispiel-Workflow, der die Verwendung der Autofeat Generator und Autofeat Apply Komponenten demonstriert.

Überprüfen Sie den Beispiel-Workflow oben, um im Detail zu sehen, wie beide Komponenten funktionieren.

Dieser Workflow ist auch auf dem KNIME Community Hub auf Ashoks Hub-Profil verfügbar.

([Airfoil Self-Noise Vorhersage mit Autofeat Generator](#)

)

Der in diesem Beispiel verwendete Datensatz ist der " [Airfoil Selbst Lärmdata Set](#) „die eine NASA ist

Datensatz aus einer Reihe von aerodynamischen und akustischen Tests von zwei und drei

dimensionale Schaufelabschnitte in einem anechoischen Windkanal durchgeführt. In der

Workflow, Ashok baut zwei lineare Regressionsmodelle, eine mit generierten Features

(oben) und eins ohne Autofeat-Funktionen (unten). Die beiden von der

Numerische Scorer nodes show, das mit generierten Features gebaute Modell ist überlegen

das Modell ohne solche Features gebaut.

Datenwissenschaft Anwendungsfälle

In diesem Abschnitt haben wir alle Artikel gesammelt, die zeigen, wie vielseitig KNIME Analytics Platform kann angewendet werden. Einige unserer COTMs schufen spannende, faszinierende, und kreative Anwendungsfälle, die zeigen, dass KNIME ein Werkzeug für (fast) alles ist! KNIME kann sogar verwendet werden, um in die magische Welt der Zwerge und Drachen... Die Kategorie „Data Science Use Cases“ zeigt unsere begeisterten Schöpfer:

- **Armin Ghassemi Rudd**
 - Geschäftsführer und Geschäftsführer Datenwissenschaftler @Intellact
- **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**
 - Ingenieursdirektor @BigBear.ai
- **John Emerging**
 - Hauptberater @phData

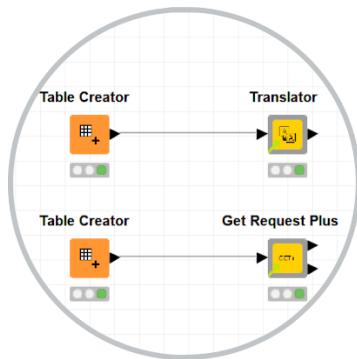


[Armin Ghassemi Rudd](#) wurde nominiert KNIME-Beitrag vom Monat für Februar 2021. Er wurde für seine [Translator](#) und [Anfrage anfordern Plus](#) Komponenten, die in der Figur rechts angezeigt. Ab 28.09.2022 beide Komponenten insgesamt 6567 Downloads. Der Übersetzer Komponente verwendet Google Translate, um jede Eingabe zu übersetzen Text von/zu unterstützten Sprachen. The Get Request Plus Komponente fügt die Option „Retry“ zur GET Request hinzu

Knoten. Neben der Schaffung von fantastischen Komponenten, Armin war eine langfristige und sehr aktiver Beitrag zum KNIME Forum.

Armin ist ein Data Science-Enthusiasten, der eine Leidenschaft für Bildung hat (beide Lernen und Lehren) und erfreut sich der Wahrnehmung von Daten. Er ist derzeit Geschäftsführer und Principal Data Scientist als Intellacct, von dem er auch der Gründer ist. Seine Felder von Interessen sind weit verbreitet und reichen von Verbraucherverhalten bis Astroteilchen Physik. Armin hält einen Master of Science in Informationstechnologie Management mit dem subfield Business Intelligence der Universität Teheran. Nach Abschluss seines Master-Abschlusses, er mit der Universität verbunden bleiben, indem wenige optionale Data Science Kurse.

Besuchen Sie Armin [Raum auf dem KNIME Hub](#) oder [Profil Seite im KNIME Forum](#) (Hub/Forum Griff:
Armierungsruder)



Aufbau eines CV Builders mit BIRT in KNIME — Teil 1

Autor: Armin Ghassemi Rud

In unserem letzten Artikel [Erstellen Sie Ihren Lebenslauf basierend auf LinkedIn Profil mit BIRT in KNIME](#) auf KNIME Blog haben wir einen CV Builder basierend auf LinkedIn Profil. Hier zeigen wir Ihnen, wie Sie den Workflow aufbauen und den Bericht mit BIRT in KNIME.



Erstellung des KNIME-Workflow-Projekts

Lassen Sie uns ein Workflow-Projekt erstellen. Open KNIME Analytics Platform und eine neue Projekt und Name es “ CV_Builder „. Jetzt, bevor Sie etwas anderes in KNIME tun, gehen Sie zur Workspace-Verzeichnis und finden Sie das “ CV_Builder ” Ordner. In diesem Ordner erstellen Sie ein neues Ordner benannt “Daten„ und Bewegung die heruntergeladen wurde LinkedIn Daten Ordner „LinkedInData Ausfuhr ” (erklärt in [unser letzter Artikel zum KNIME Blog](#)) in den Ordner benannt nachDaten” im Workflow-Verzeichnis (CV_Builder) in Ihrem Arbeitsraum.

Wir müssen unser Foto mit dem Namen speichern “ personal_photo.png ” im “ Daten ” Ordner Verzeichnis. Abmessungen sollten 496*516 sein, oder wir müssen das Bildelement konfigurieren in BIRT etwas anders als hier. Dann laden Sie die “ Hintergrund ” Datei und es in diesen Ordner verschieben.

Aufbau des Workflows

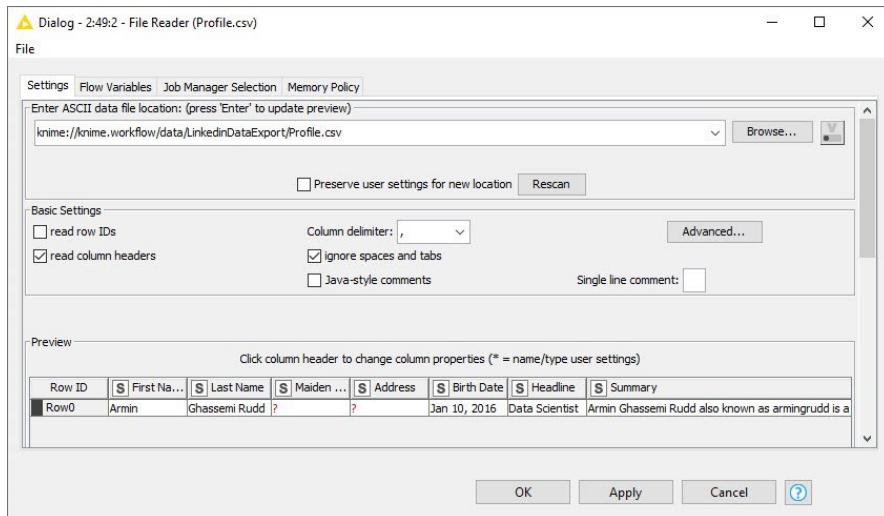
Jetzt müssen wir die CSV-Dateien lesen, die wir heruntergeladen haben und in die verschoben “ Daten „ Ordner. Wir brauchen nur diese 11 Dateien: “ Zertifizierungen.csv „, Bildung.csv „, Kurse.csv „, „E-Mail senden“ Adressen.csv „, „Endpunkte“ empfangen Info „, „Sprachen“ „, „Positionen.csv „, „Profil.csv „, „Projekte.csv „, „Empfehlungen“ Empfangen.csv „, „Skills.csv „,

Da wir die Ausgänge an die Reader Knoten, wir brauchen 11 Datei Reader Knoten. Machen wir es nach einem.

Erstellen Sie eine Metanode und nennen Sie esProfil „. Doppelklicken Sie auf die Metanode, um nach innen zu gehen. Hinzufügen einer Datei-Leser Knoten und gehen Sie zu den Konfigurationen des Knotens. Fügen Sie diesen Pfad als Datei hinzustandort:

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

„knime://knime.workflow/“ bezeichnet das aktuelle Workflow-Verzeichnis. Jetzt Überprüfen Sie die „Lesesäulenkopf“ Option und drücken „Okay.“,

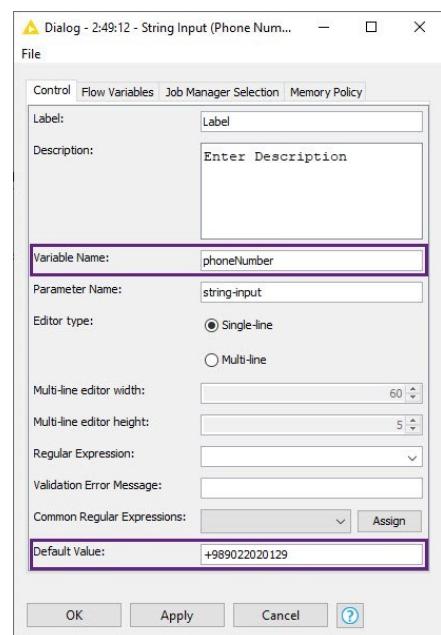


Das Konfigurationsfenster des File Reader Knotens.

Vor der Ausführung des Knotens können wir ein String Input Knoten vor dem Datei-Leser Knotenpunkt um unsere Telefonnummer als Flussvariable hinzuzufügen. Im Konfigurationsfenster des Knotens, die „Variabler Name“ bis „TelefonNumber“ und geben Sie Ihre Telefonnummer vor dem „Standardwert“ Option.

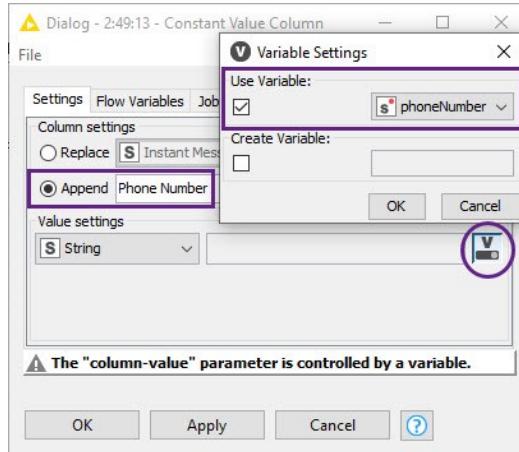
Schließen Sie den Ausgangsport des String Input Knoten zum variablen Eingangsport des Datei Reader knoten und die Knoten ausführen. Wir müssen die Zusammenfassung in unserer Tabelle, so dass die neuen Linien erscheinen erwartet. In der aktuellen Tabelle, Zusammenfassung hat keine neuen Zeilen, auch wenn wir mehrere Absätze in unsere Original Zusammenfassung auf der LinkedIn Website. Die neuen Linien wurden durch Doppel ersetzt Räume. Also verwenden wir ein String Manipulation um dies zu korrigieren. Wir können dies anwenden Ausdruck:

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

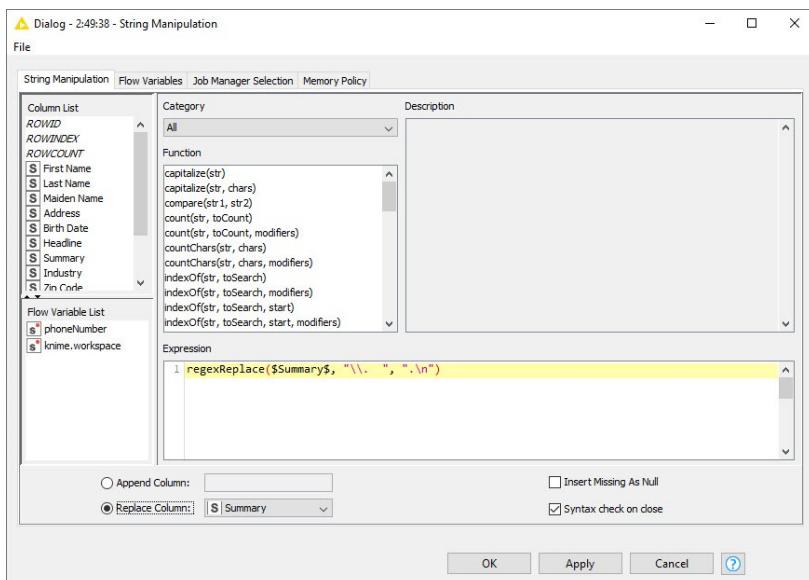


Das Konfigurationsfenster des String Input Knoten.

Jetzt brauchen wir unsere Durchflussvariable `TelefonNumber` „ zur Datentabelle hinzugefügt werden. Die Konstante Wertesäule node übernimmt die Aufgabe für uns. Eine neue Spalte anhängen „ Telefonnummer „ und zuweisen „ `TelefonNumber` „ strömungsvariabel zum „ Werteinstellungen „ Option.



Der Konfigurationsdialog der Konstantwertsäule Knoten.



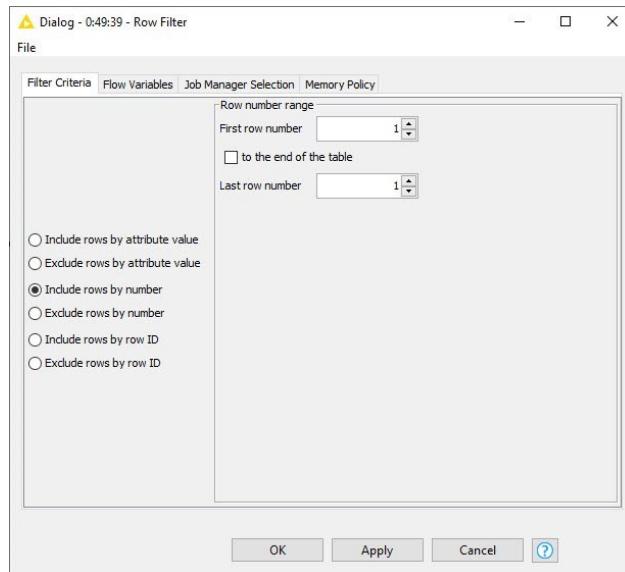
Das Konfigurationsfenster des String Manipulation Knotens.

Als nächstes möchten wir unsere E-Mail-Adresse von `Email Adressen.csv` „ zur Tabelle hinzugefügt auch. Also verwenden wir einen anderen Datei-Leser Knoten und fügen Sie diesen Pfad hinzu:

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

Auch hier müssen wir die „ Spaltenüberschriften lesen „ Option.

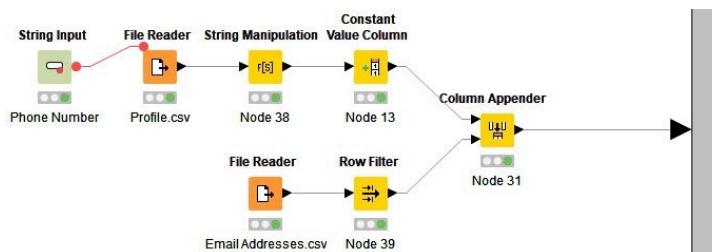
Da wir vielleicht mehrere E-Mail Adressen, a Row Filter Knoten wird verwendet, um den erste in der Liste. Wählen Sie " Zeilen nach Nummer einschließen und geben Sie die Zahl " 1 " für beide „Nummer der ersten Zeile“ und „ Letzte Zeilennummer „,



Der Konfigurationsdialog des Zeilenfilterknotens.

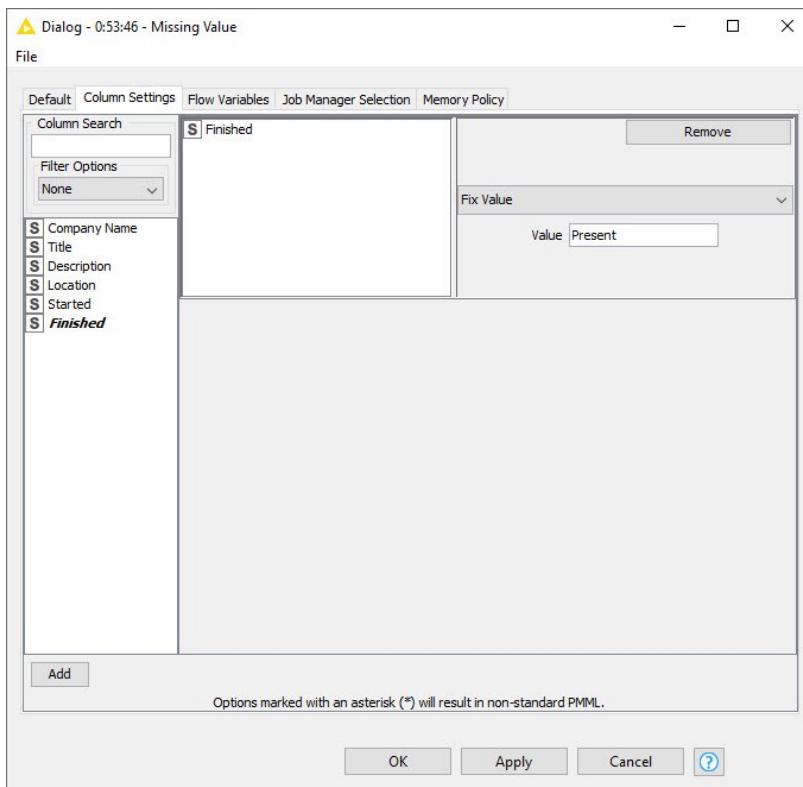
A Antrag auf Einreichungnode wird nun unsere E-Mail-Adresse an die Haupttabelle hinzufügen.

Das haben wir jetzt im Profil Metanode:



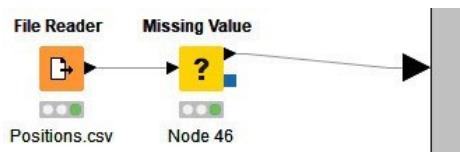
Das Innere des Profils metanode.

Gehen wir nun zur nächsten Metanode: Positionen . A Datei-Leser den Knoten zu lesen „Positionen.csv“ Datei (mit Spaltenüberschriften) und a Fehlender Wert Knoten wird ersetzen fehlende Werte in der „ Fertig „ Spalte mit „ Gegenwart,Wert. Wenn das „ Fertig „Wert Das bedeutet, wir haben noch die Position. Die Linked In der Website zeigt der Begriff „Gegenwart“ aber in den heruntergeladenen Daten erscheint es als fehlender Wert.



Der Konfigurationsdialog des fehlenden Wertknotens.

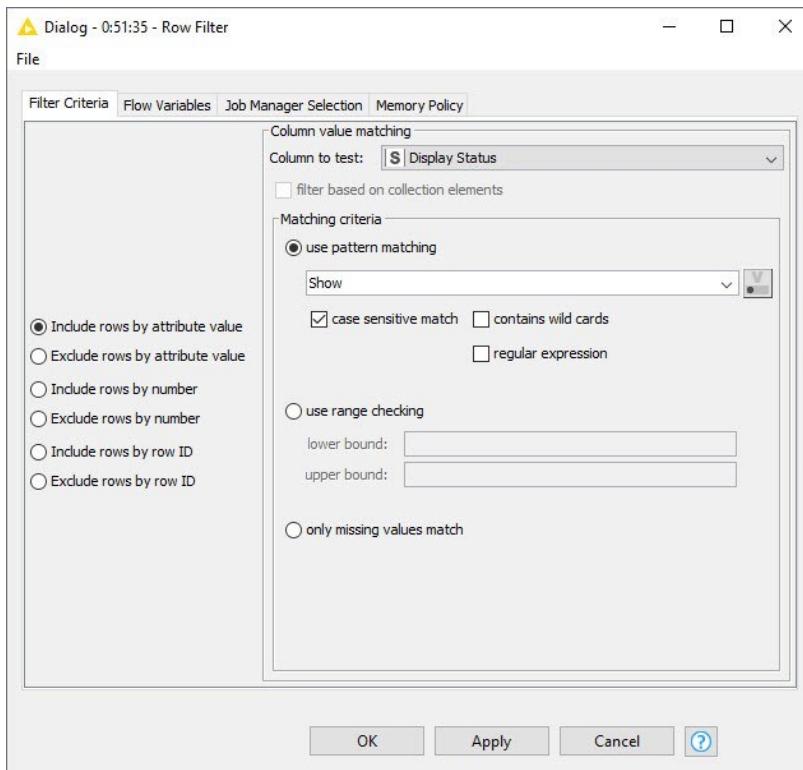
Und das ist es für die Positionen Metanode.



Das Innere der Positions metanode.

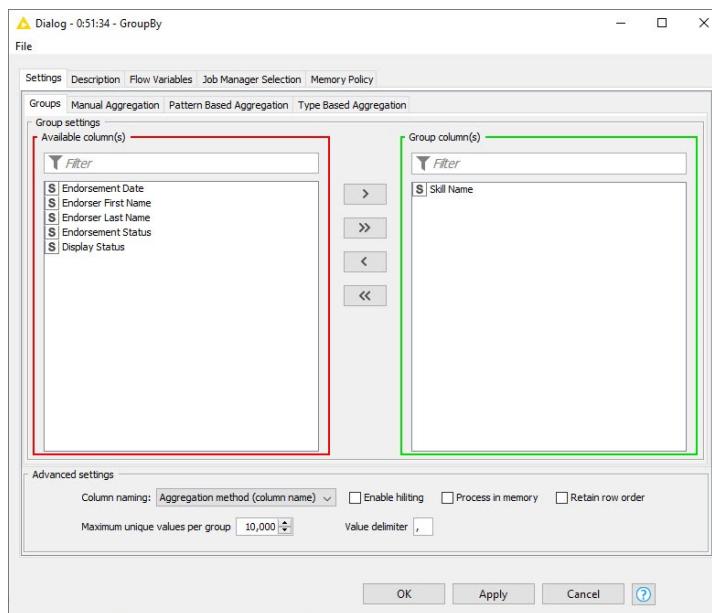
Das gleiche Verfahren kann für die Projekte und Zertifizierungen Metanoden, wir lesen “Projekte.csv” und die “Zertifizierungen.csv” Dateien.

Die nächste Metanode ist diejenige, die “Skill.csv” und die “Endorsement erhalten Info” Dateien. Wir benutzen ein Datei-Leser Knoten für jeden von ihnen. Lassen Sie uns ein Row Filter Knoten nach dem Lesen der Befürwortungen, um die, die wir gewählt haben, angezeigt werden in unserem Linked Im Profil:



Der Konfigurationsdialog des Zeilenfilterknotens.

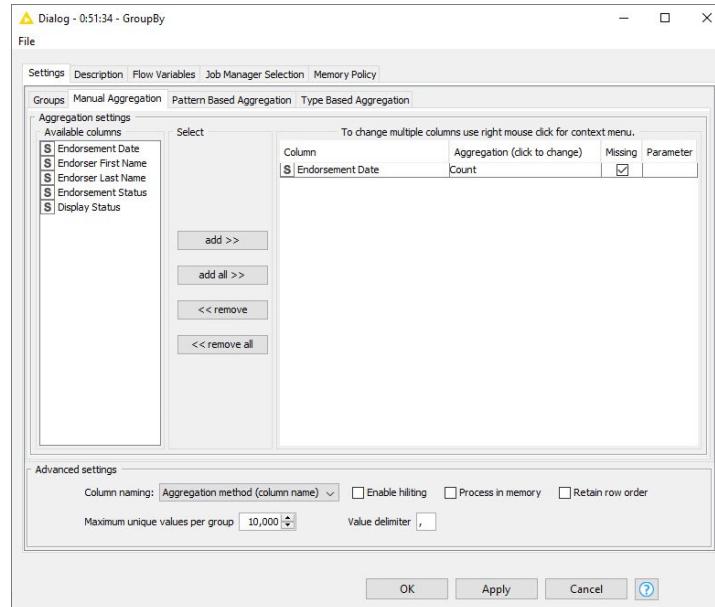
Wie Sie im Bild oben sehen können, haben wir die Zeilen enthalten, in denen der Wert für die „Anzeigestatus“ „Show“ ist



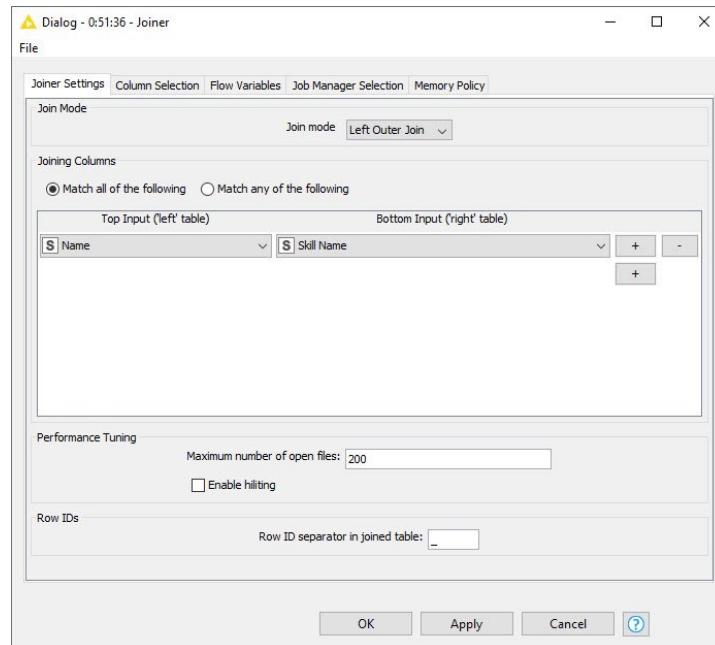
Die Gruppen Tab im Konfigurationsdialog des GroupBy-Knotens.

Datenwissenschaft Use Cases – Armin Ghassemi Rud
Aufbau eines CV Builders mit BIRT in KNIME — Teil 1

Der Nächste, der Gruppe node wird die Bestätigungen für jedes Geschick zählen. Wir wählen die “Name” Spalte als Gruppierungsspalte, und jede der anderen Spalten können ausgewählt werden mit “Anzahl” Aggregationsfunktion während der Option zum Zählen der fehlenden Werte wird überprüft.



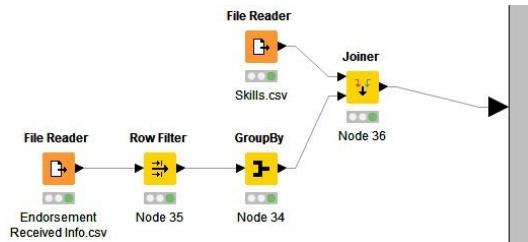
Die Registerkarte Manuelle Aggregation im Konfigurationsdialog des GroupBy-Knotens.



Der Konfigurationsdialog des Joiner-Knotens.

Jetzt, a Mitgliednode wird sich unseren Fähigkeiten und Befürwortungen anschließen. Wir können die „Name“ Spalte aus dem Skills-Datensatz und dem „Skill Name“ aus den Befürwortungen Datensatz, um diese beiden Tabellen zu verbinden.

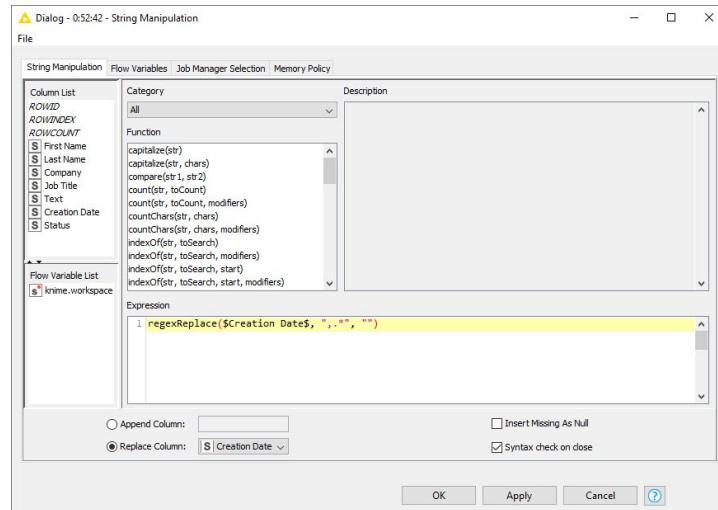
Die Fähigkeiten Metanode ist auch fertig.



Das Innere der Fähigkeiten metanode.

In der Empfehlungen metanode, wir lesen die „Empfehlungen Received.csv“ zuerst. Dann String Manipulation Knoten, um die Zeit von der „Erstellungsdatum“ Spalte sollte verwendet werden, in der wir den folgenden Ausdruck anwenden und die „Erstellungsdatum“ Spalte.

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

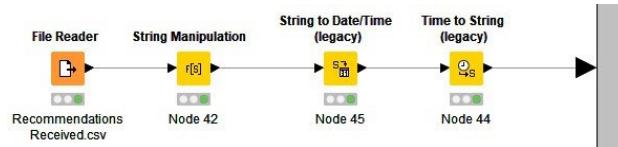


Der Konfigurationsdialog des String Manipulation Knotens.

Nächster, a String to Date/Time (legacy) Node zum Konvertieren von "Erstellungsdatum" Spalte bis Legacy Date&Time, wo wir Monatsnamen wie Jan, Feb, haben... und dann können wir wandeln Sie es zurück in String und halten Sie den Monatnamen und das Jahr mit einem Zeit bis String (Legacy) Knoten.

In den Konfigurationen der String to Date/Time (legacy) Knoten, wir ersetzen die „Schaffung Datum“ Spalte und verwenden Sie dieses Datumsformat MM/dd/y . In der Zeit zum Streichen (Vermächtnis) Knoten,

wir benutzen MMM yyyy als Datumsformat und die Spalte ersetzen. Jetzt, unsere Empfehlungen metanode sieht so aus:



Das Innere der Empfehlungen metanode.

Die letzte Metanode ist die Sprache Metanode, wo wir die „Sprachen“ Datei. Hier müssen wir ein Wörterbuch erstellen, um unsere Fähigkeiten in jeder Sprache in eine Anzahl Skalierung von 1 bis 5. Wir benutzen ein Tabelle Schöpfer Node, um die Leistungsfähigkeit zu erfassen Ebenen und die entsprechenden Zahlen:

Manually created table - 0:55:46 - Table Creator (Language lvl number)		
File Hilitc Navigation View		
Table "default" - Rows: 5 Spec - Columns: 2 Properties Flow Variables		
Row ID	S level	S ▾ 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

Die Ausgabe des Table Creator-Knotens zeigt eine Tabelle mit Leistungsfähigkeit Werte und entsprechende Zahlen.

Wir werden ein Regelmotor (Diktiorär) Knoten, um die Zahlen auf unsere Sprachen basierend auf den Leistungsstufen. Wir müssen also die Regeln schaffen. Das zu tun, wir benutzen ein String Manipulation Knoten nach dem Tabelle Schöpfer Node und verwenden Sie den Ausdruck unten, während die „Niveau“ Spalte.

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

Der Ausgang ist:

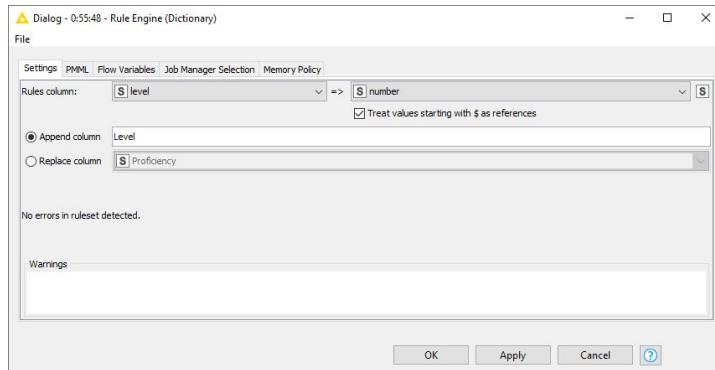
Appended table - 0:55:47 - String Manipulation		
File Hilitc Navigation View		
Table "default" - Rows: 5 Spec - Columns: 2 Properties Flow Variables		
Row ID	S level	S ▾ number
Row0	\$Proficiency\$ = "Native or bilingual proficiency"	5
Row4	\$Proficiency\$ = "Full professional proficiency"	4
Row1	\$Proficiency\$ = "Professional working proficie...	3
Row3	\$Proficiency\$ = "Limited working proficiency"	2
Row2	\$Proficiency\$ = "Elementary proficiency"	1

Der Ausgang des String Manipulation-Knotens nach dem Aufbringen der nach dem Ausdruck Join("\$Proficiency\$ = \"", \$level\$, "\"").

Nun, die Ausgabe der String Manipulation geht zum unteren Eingangsport des Motor (Diktiorär) Knoten und Ausgang aus dem Datei-Leser geht in den oberen Hafen.

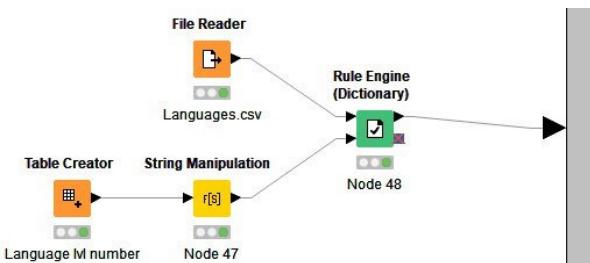
Im Konfigurationsfenster der Regelmotor (Diktiorär) Knoten, die „Regelspalte“, „die Pegelsäule und die Ergebnisse in der „Anzahl“ Spalte. Wir fügen die Ausgabe in eine neue Spalte namens „Ebene“,

Datenwissenschaft Use Cases – Armin Ghassemi Rud
 Aufbau eines CV Builders mit BIRT in KNIME — Teil 1



Der Konfigurationsdialog des Regel Engine (Dictionary) Knotens.

Die Sprachen Metanode sieht nun so aus:



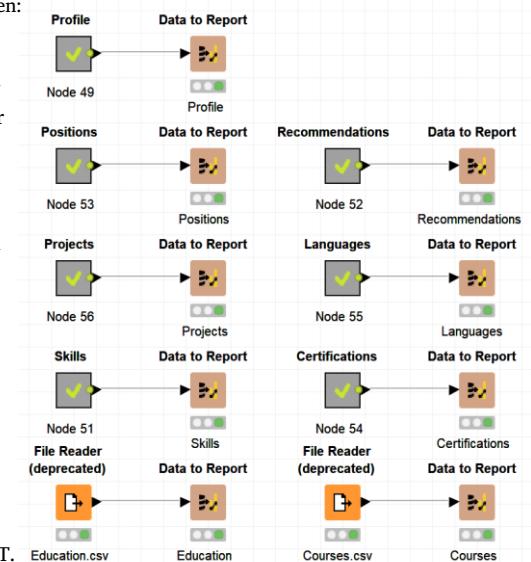
Das Innere der Sprachen metanode.

Es gibt zwei weitere Dateien, die wir hinzufügen müssen:

die „Bildung.csv“ „ und die „Kurse.csv“ Dateien. Wir benutzen zwei Datei Reader Knoten, um diese Dateien zu lesen, und wir sind mit dem Lesen unserer Datensätze beendet.

Jetzt benutzen wir einer Daten zum Bericht Knoten für jeder Metanode und die letzten zwei Reader Knoten.

Vergessen Sie nicht, Node Kommentare für die Daten zum Bericht seit der Namen für die Datentabellen in BIRT werden das gleiche wie diese Kommentare. Jetzt den Workflow ausführen und speichern. Die Workflow ist abgeschlossen, und jetzt können wir Beginnen Sie am Aufbau unseres Lebenslaufs in BIRT.



Der Überblick über den CV Builder Workflow.

Dieser Artikel wurde erstmals in der [Act of Intelligence Accretion Journal](#) auf Medium. Finden die Originalversion [Hier.](#), Die entsprechenden [Lebenslauf](#) finden Sie auf der KNIME Community Hub in [Der öffentliche Raum von Armin](#)

Sie können weiter lesen Teil 2 auf Medium bei [Aufbau eines CV Builders mit BIRT in KNIME – Teil 2](#).



[Angus Veitch](#) wurde nominiert Beitrag des Monat für November 2020. Er wurde für seinen Artikel ausgezeichnet auf seine [Wie geht's? Arbeitsablauf](#) wo er eine KNIME Workflow zur Erstellung textreicher Visualisierungen von Twitter Daten um einen bestimmten Hashtag. Das Bild auf der Rechts zeigt einen Teil der Netzwerkvisualisierung von Nutzern, die über die Melbourne-Verriegelung. Neben dem Schreiben Artikel und die Schaffung von KNIME Workflows, er ist auch ein

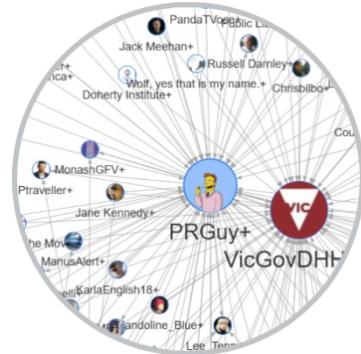
aktives Mitglied der australischen Gemeinschaft und er ist Mitglied der aktuellen [Redaktion unseres Niedriger Code für Advanced Data Science Journal auf Medium.](#)

Angus ist leidenschaftlich an Daten, konzentriert sich auf Textanalyse und Visualisierung, und sogar einen PhD, der die Verwendung von Text erforscht in der Sozialwissenschaften. Er hält zwei Blogs, wo er gerne über lokale Geschichte schreiben, Umpacken in eine verständlichere Form, und über seine Experimente in der Datenanalyse und Visualisierung. Was auch immer er tut, er versucht, es zu kommunizieren in einer klaren und ansprechenden Weise.

Besuchen Sie Angus [Raum auf dem KNIME Hub](#) oder [Profil](#)

[Seite im KNIME Forum](#) (Hub/Forum Griff:

Angust)



Wie geht's? R in KNIME

Ein Workflow zur Erstellung textreicher Visualisierungen von Twitter Daten

Autor: Angus Veitch

Anmerkung des Herausgebers:

Einige der in diesem Artikel verknüpften Knoten im Workflow sind auf einer externen Software verfügbar
Aktualisierung der Website.

Twitter kann heute nicht die am weitesten verbreitete Social Media-Plattform der Welt sein (at [die Zeit des Schreibens, Facebook hat über siebenmal so viele Nutzer](#)), aber es ist sicher die am weitesten untersucht. Dank seiner hoch zugänglichen API (Anwendungsprogrammierung) Schnittstelle), Twitter ist die Go-to-Quelle von Daten für Sozialwissenschaftler, Markt Forscher und andere Analysten, die Trends abbilden und Erkenntnisse aus sozialen Medien. Während einige Social-Media-Plattformen keine einfache Möglichkeit bieten, große herunterzuladen Gehaltsmengen, Twitters API macht es relativ einfach, zehn oder hunderte zu erhalten Tausende von Tweets zu einem bestimmten Thema. Der einzige harte Teil, um Daten zu erhalten von Twitter wählt aus den vielen verfügbaren Skripten, Tools und Services Das wird Ihnen helfen, es zu tun.

Aber was können Sie eigentlich mit hunderttausend Tweets machen?

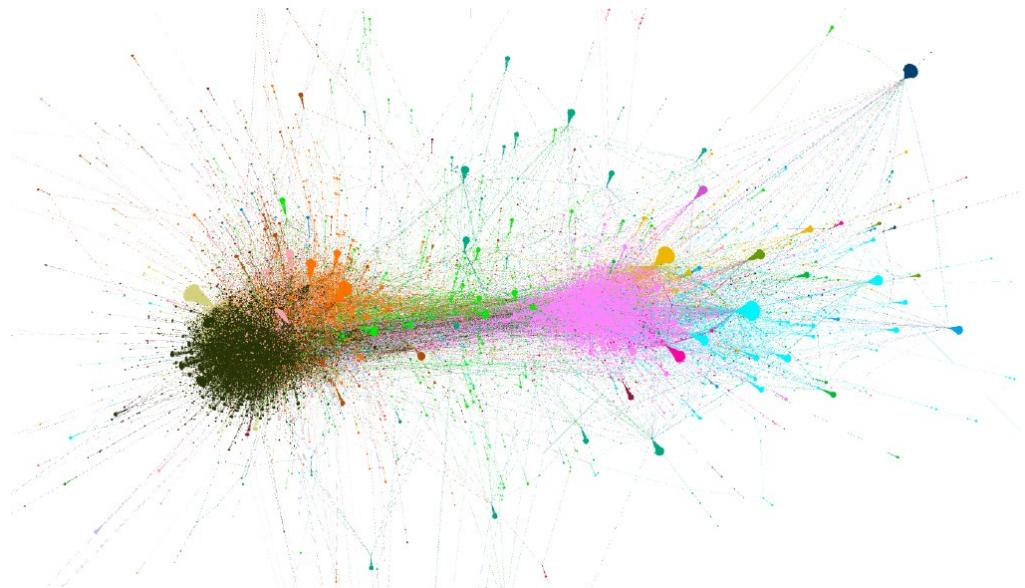
Dieser Artikel stellt eine Antwort auf diese Frage. Die Antwort lautet: TweetKolidR, ein Workflow, den ich für die KNIME Analytics Platform entwickelt habe (Working zusammen mit ausgewählten R-Paketen) zur Erstellung textreicher Visualisierungen von Twitter Datensätze. Nach der ersten Erklärung der Motivation hinter dem TweetKolidR, werde ich illustrieren was es tut und erklärt, wie Sie es verwenden können, um eigene zu erstellen und zu analysieren Twitter Datensätze innerhalb von KNIME.

Die Herausforderung: Sinn für den Twitterverse machen

Twitter-Daten verleihen sich natürlich mindestens zwei analytischen Methoden. Das erste ist Netzwerkanalyse. Beiträge (oder „tweets“) und Nutzer auf Twitter werden über Webs verbunden von Interaktionen wie Retweets (wobei ein Benutzer den Tweet eines anderen Benutzers wiederholt), Erwähnungen (wo ein Benutzer den Benutzernamen eines anderen Benutzers in einem Tweet enthält), und Antworten (wo ein Benutzer auf einen anderen reagiert). Mit kaum Manipulationen überhaupt, die Daten der Twitter-API können in ein Netzwerk von Benutzern oder Tweets umgewandelt werden, die von Retweets und Erwähnungen (Replies sind etwas schwieriger, können aber auch eingebunden werden). Ein beliebter Weg, solche Netzwerke zu visualisieren, ist mit dem Open-Source-Tool Gephi, das

(mit Hilfe von entsprechend abgestimmten Layout- und Community-Detektionsalgorithmen)

ein Netzwerk von 33.000 Nutzern in so etwas verwandeln:



Ein Netzwerk von mehr als 33.000 Nutzern, die im Jahr 2020 über Melbournes Abschaltung schweben, wie in Gephi gezeigt mit dem ForceAtlas2-Algorithmus. Die Farben zeigen eng miteinander verbundene Gemeinschaften.

Es ist sicherlich schön, aber was bedeutet das? Nun, wenn du ins Netzwerk zoomst innerhalb von Gephi würden Sie in der Lage sein, die Namen jedes der 33.000 Benutzer zu sehen, skaliert die Beliebtheit oder den Einfluss jedes Benutzers zeigen. Wenn einige der Benutzernamen sind vertraut, könnten Sie in der Lage sein, Sinn für die Sub-Netzwerke oder Gemeinschaften in die sie gebündelt werden (in dem Bild oben dargestellt).

Dies sind nützliche Informationen. Aber um etwas mehr zu lernen — wie das, was jeder Cluster Benutzer sprechen über, oder welche Attribute sie definieren — Sie müssen durch die ursprünglichen Daten, um die relevanten Tweets und Benutzerbeschreibungen zu finden und zu lesen. Das könnte sei ein mühsamer Prozess! Wenn es nur einen Weg gab, eine sofortige Zusammenfassung jeder Tweets und Benutzerprofile von Clustern ohne die Visualisierung zu verlassen. Wie Sie sehen Der TweetKolid R bietet genau diese Funktionalität.

Die zweite Analysemethode, die häufig auf Twitter-Daten angewendet wird, ist die Zeitreihe Analyse. Da jeder Tweet zeitgenampiert ist, ist es einfach, ein Diagramm zu machen, wie die Lautstärke der Tweets über ein bestimmtes Thema Änderungen im Laufe der Zeit. Aber wie bei den Netzwerk-Visualisierung, müssen Sie zurück zu den Daten, um zu sehen, was wirklich geht auf — zum Beispiel zu sehen, welche spezifischen Themen oder Nutzer Aktivitäten treiben, oder wie sich die Gesamtmischung von Themen im Laufe der Zeit verändert. Der TweetKolid R hilft uns all diese Dinge auf einmal zu sehen.

Kurz gesagt, die Motivation hinter dem TweetKolidR ist, schnelle Einblicke über die Inhalt, Struktur und zeitliche Dynamik der Twitter-Aktivität rund um ein bestimmtes Thema. In es ist in erster Linie darauf ausgelegt, die explorative oder beschreibende Analyse von

Twitter-Daten, könnten aber verwendet werden, um quantitative Analysen zu unterstützen oder zu informieren. Als Zusätzlicher Bonus, der TweetKolid R kann Ihnen helfen, Ihren eigenen Datensatz von Tweets über ein bestimmtes Thema.

Noch besser, der TweetKolid R ermöglicht es Ihnen, all dies zu tun, ohne irgendeinen Code zu verwenden, wie es wird vollständig über eine Punkt-und-Klick-Schnittstelle innerhalb von KNIME Analytics implementiert Plattform.

Installieren des TweetKolid Arbeitsablauf

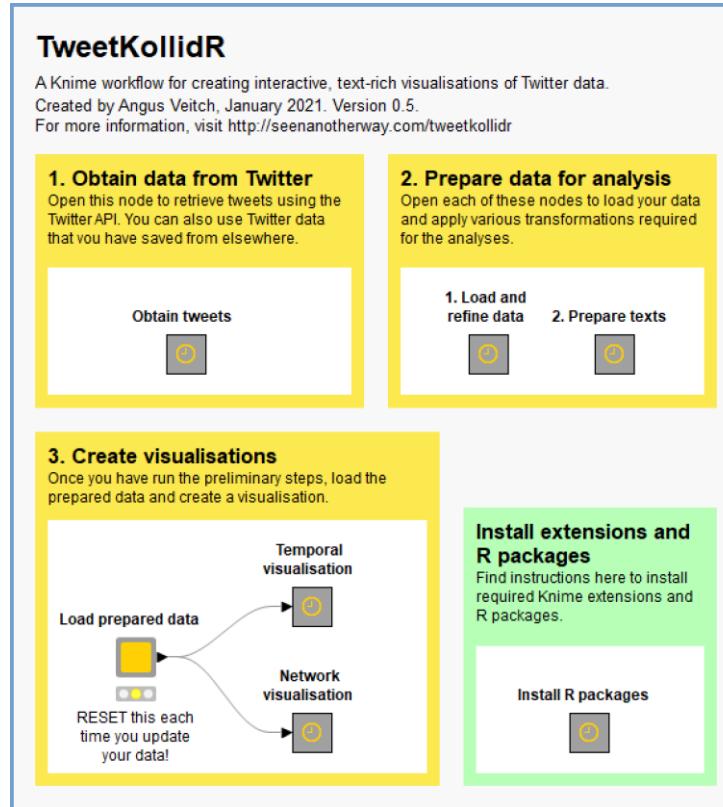
Wie oben erwähnt, wird der TweetKolidR für die KNIME Analytics Platform gebaut, so dass die erste Voraussetzung für die Verwendung ist [herunterladen und installieren KNIME](#) (wenn Sie nicht schon!). Sobald Sie das getan haben, können Sie den Workflow direkt von der [KNIME Hubraum](#). Die Workflow könnte dann verlangen, dass Sie einige Erweiterungen installieren, vor allem die [Interaktive R Statistiken Integration](#), die es KNIME ermöglichen wird, verschiedene Pakete (spezifisch, [igraph](#), [VisNetwork](#) und [Plotly.R](#)) für die Durchführung eines Netzwerks Berechnungen und Erstellung der visuellen Ausgänge. Sobald Sie erfolgreich sind [konfiguriert R Statistik Integration](#) Erweiterung, können Sie diese und andere notwendige installieren R Pakete einfach durch die Ausführung eines vordefinierten Skripts im Workflow. mit eingebautem R-Pakete macht den Workflow etwas schwieriger einzurichten, es ermöglicht es uns, Dinge zu tun, die wir nicht (ja!) allein mit KNIME tun können. Insbesondere gibt es Zugriff auf die wunderbare Welt der Datenvisualisierungstools, die entwickelt wurden für R, von denen igraph, visNetwork und Plotly nur einige Beispiele sind.

Verwendung des Workflows

Der TweetKolid Der R-Workflow soll von KNIME Analytics genutzt werden Plattform wie eine eigenständige App. Wie in der Abbildung unten gezeigt, die „Homepage“ (oder obere Ebene) des Workflows wird in Metanoden unterteilt, die sich auf bestimmte Aufgaben beziehen. In jedem sind weitere Metanoden und Komponenten, mit denen der Benutzer interagiert mit die Aufgabe ausführen. Jede Komponente ist über Point-and-Click-Optionen konfigurierbar, die alle durch die On-Demand-Dokumentation im KNIME ausführlich erläutert werden Beschreibungsbereich.

An keinem Punkt müssen Sie einen beliebigen Code eingeben oder sich mit der inneren Mechanik belästigen des Workflows. Natürlich, wenn Sie unter der Haube sehen wollen, und sogar die Anpassung der Workflow auf Ihre Bedürfnisse, können Sie dies einfach durch Öffnen jeder Komponente und Bohren in die Details. (Aber sei gewarnt, es sieht nicht alle so hübsch aus!)

Die folgenden Abschnitte zeigen die Kernmerkmale des TweetKolidR und Umrisses die Schritte, die an der Verwendung beteiligt sind. Für eine detailliertere Beschreibung siehe die viel längere [Blog Post](#) auf dem dieser Artikel basiert. Am [wohnzimmer.de](#), Sie werden auch finden weitere Analysen mit TweetKolid R-Ausgänge, einschließlich einer [Australien Tag](#) und eine andere über die [GameStop saga](#).



Der Hauptbildschirm des TweetKollidR Workflows. Jede der Boxen enthält weitere Knoten und Anweisungen, um Sie durch den Prozess der Gewinnung zu führen, Erstellung und Visualisierung Ihrer Daten.

Sammeln und vorbereiten Twitter-Daten

Die Suite von KNIME [Twitter Connector Nodes](#) bereits machen es sehr einfach, Daten abzurufen durch die Twitter-API (d.h. vorausgesetzt, Sie haben bereits einen API-Schlüssel, die Sie können durch Antrag auf eine [Entwicklerkonto](#) Der Twitter Search-Knoten zum Beispiel ruft Tweets auf, die einer bestimmten Keywords-Abfrage entsprechen.

Es gibt jedoch einen Fang. Es sei denn, Sie haben für Premium-Zugriff auf die Twitter-API bezahlt, oder haben Zugang zur akademischen Forschungsproduktkategorie, können Sie nur eine begrenzte herunterladen Anzahl der Tweets zu einer Zeit (bis zu 18.000 Tweets in einem 15-minütigen Zeitraum zum Zeitpunkt der Schreiben), und diese Tweets werden nur von der letzten Woche oder so sein. So bauen Sie eine lange oder ein großer Datensatz, müssen Sie mehrere Anfragen im Laufe der Zeit machen und die Ergebnisse. Der TweetKollid R macht diesen Prozess einfach und bietet eine Sequenz von Knoten (in der Abbildung unten dargestellt), die die Ergebnisse wiederholter Recherchen automatisch zusammenfassen. Es erlaubt sogar, mehrere Suchanfragen anzugeben, die dann automatisch um sicherzustellen, dass sie gleich vertreten sind.

Obtain tweets

These nodes enable you to assemble a twitter dataset by sending search queries to the Twitter API. To do this, you will need a Twitter API key, which you can apply for at <https://developer.twitter.com/en/docs/twitter-api/getting-started/guide>. The sequence is designed so that you can run it on a recurring basis to build up a longitudinal dataset while working with the constraints of the API's standard access allowances.

Enter Twitter API credentials

Configure this node to enter your API key and access token. Visit <https://developer.twitter.com> for more information about API access.

Twitter API Connector



Enter search queries

Define one or more queries to send to the Twitter API, with each query appearing on a new line. It's a good idea to try out each query in the Twitter web interface first.

Table Creator



Retrieve tweets

Run these nodes to retrieve tweets matching your queries, merge the results with those saved earlier, and save the outputs. Standard access to the API provides a sample of tweets from the last seven days, up to a limit of about 18,000 per 15-minute interval. Run this process on a recurring basis to build a longer and more complete collection.

Retrieve tweets

Merge with existing data

Table Writer

Reset to do a new search

Save merged data

Review collected data

Here you can see a visual summary of the data you have collected to date, and also compare the output of your search queries. Note that you will first need to have R Integration working and the necessary packages installed.

Review collected data

Dieser Bildschirm führt Sie durch den Prozess der Erstellung eines Datensatzes von Tweets über Ihr Thema von Interesse.

Weil jede Anfrage an den Standard gesendet wird Twitter API gibt nur eine Probe von

Sie können nie sicher sein, dass Ihre Sammlung komplett ist. Allerdings

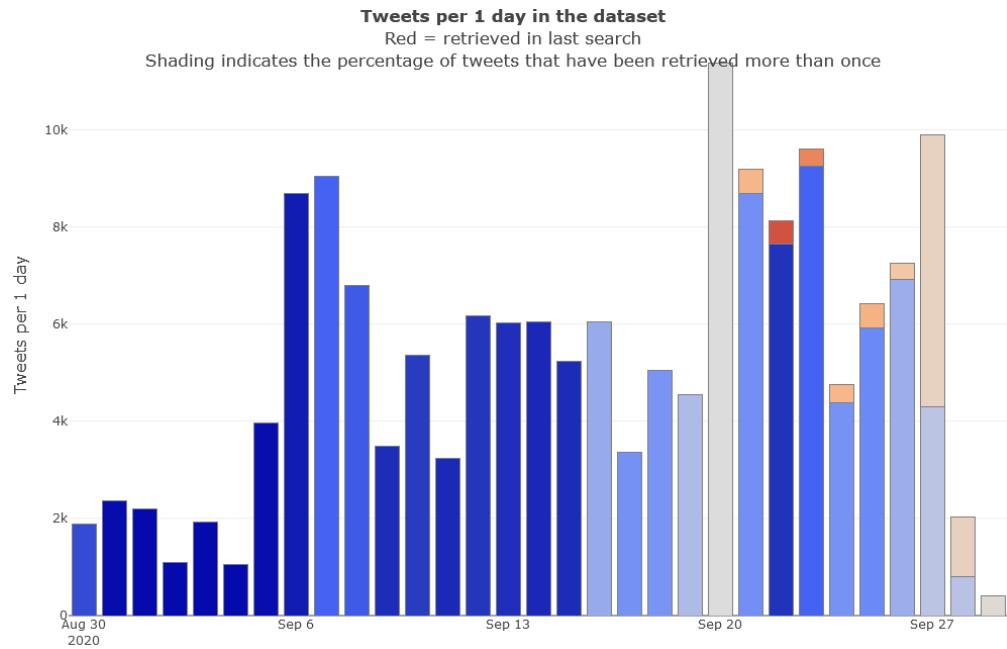
Tracking, wie viele neue Tweets von jeder Anfrage zurückgegeben werden, können Sie zumindest raten wenn Sie so viele abgerufen haben, wie die API bereitstellt. Zu diesem Zweck die

Wie geht's? R liefert die nachfolgende Visualisierung.

In diesem Fall können wir feststellen, dass fast alle Tweets im Datensatz mehrfach abgerufen wurden Zeiten bis zum 16. September, an welchem Punkt gibt es mehrere aufeinanderfolgende Tage, wo die Daten weniger gesättigt sind. Obwohl dieses Bild statisch ist, ist die Ausgabe von TweetKolidR tatsächlich eine interaktive HTML-Version, die vom Plotly-Paket in R erzeugt wird.

Falls Sie sich fragen, die Daten hinter der Abbildung oben, und hinter den Beispielen die in diesem Artikel folgen wird, ist eine Sammlung von Tweets im Zusammenhang mit der 15-wöchigen Lockdown die in Melbourne, Australien zwischen Juni und Oktober 2020 in Reaktion zum Ausbruch von Covid-19. Der Datensatz, den ich über einen Zeitraum von etwa vier zusammengebaut habe Wochen im September 2020, enthält rund 100.000 Tweets, die passende Abfragen wie „Lockdown AND melbourne“.

Wie geht's? R in KNIME



Die Anzahl der Tweets pro Tag in meinem Datensatz am 27. September 2020. Farbe und Schattierung der Balken
Informationen über die wahrscheinliche Vollständigkeit des Datensatzes liefern.

In vielerlei Hinsicht ist der TweetKollidR ein direktes Produkt dieses Lockdowns, wie es war während dieser Zeit, die ich fand die Zeit und die Neugier, um zu sehen, was Twitter könnte zeigen über die lokale Politik der Pandemie. Der TweetKollid R war meine eigene Lockdown Baby!

(Dieser Lockdown hatte ein Happy End, und nicht nur, weil er den TweetKollidR produzierte. Es führte auch zur totalen Eliminierung des Coronavirus aus Melbourne für mehrere Monate.)

Erstellung der Daten zur Analyse

Vor der Visualisierung Ihrer Daten, die TweetKollid R muss es einer Reihe von Vorbereitungen unterwerfen Schritte. Die meisten dieser Schritte umfassen Textvorverarbeitungsvorgänge wie die Entfernung von unerwünschten Zeichen, der Bezeichnungen und n-Grammen, und Filterung und Standardisierung der Begriffe. Diese Schritte werden mit Hilfe von KNIME durchgeführt Native Textverarbeitungsknoten. Der TweetKollid R diese Schritte nicht nur auf die tweets selbst, aber auch auf den in den Benutzerbeschreibungen enthaltenen Text. Außerdem, die TweetKollid R identifiziert Tweets, die doppelt sind, auch wenn sie nicht als Retweets bezeichnet, so dass solche Duplikate von bestimmten ausgeschlossen werden können Analysen.

Der Benutzer führt diese Schritte einfach durch Konfigurieren und Ausführen der relevanten Komponenten, wie sie im Bild unten dargestellt sind. Technische Informationen

Diese Schritte finden Sie in der eingebauten Dokumentation und in derWie geht's? R blogPost :

Load and refine data

These steps will prepare your data for text preprocessing and subsequent analysis. The output of each step will be loaded into the next. Note that if you repeat a step, you will need to **reset the Table Reader** at the start of the next step to ensure that the updated data gets loaded. Double-click the components to access their configuration options, or use ctrl-double-click to browse their contents.

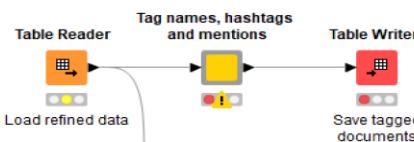
1. Load and refine data

Open the first node to select your data and ensure that it is compatible with the workflow. After cleaning and refining the data, you can optionally filter it to a specific time period.



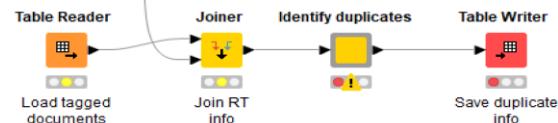
2. Tag names and mentions

This step tags named entities (places, people and organisations) as well as mentions and retweets within the tweet texts. The named entities will appear later in lists of top names, while the mentions and RTs will be used to construct a network of user interactions.



3. Detect duplicates

This step identifies groups of tweets that are at least nearly identical to each other, regardless of whether they are retweets. This information will be used later in term frequency analyses.

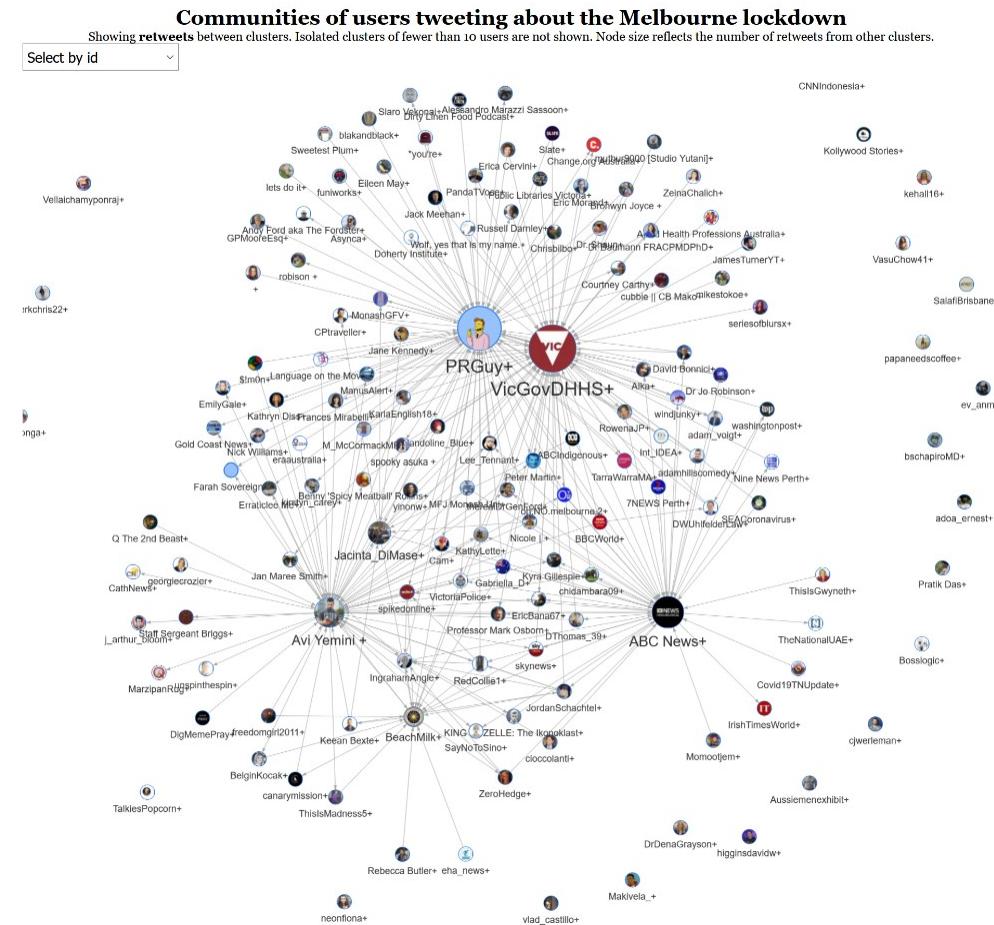


Die Komponenten innerhalb des Datenbereichs Load und verfeinern den Arbeitsablauf.

Communitys von Nutzern entdecken

Erinnern Sie sich an die sehr hübsche, aber nicht sehr informative Netzwerkvisualisierung, die in der Beginn dieses Artikels? Die interaktive Visualisierung der gleichen Daten von TweetKolidR sieht so aus:

Wie geht's? R in KNIME



Das „Zusammenfassungsnetz“ zeigt Verbindungen zwischen Clustern von Nutzern, die über das Melbourne twittern abschließen. Jeder Knoten stellt einen Cluster von überall von mehreren Benutzern zu mehreren tausend dar. Vollständig

Funktionalität, siehe [interaktive Version](#).

Jeder Knoten in diesem Netzwerk ist ein Cluster von Benutzern, die automatisch von einer Community identifiziert werden Erkennungsalgorithmus (spezifisch die **Fast_Grady** Algorithmus aus dem **Igraph R Paket**).

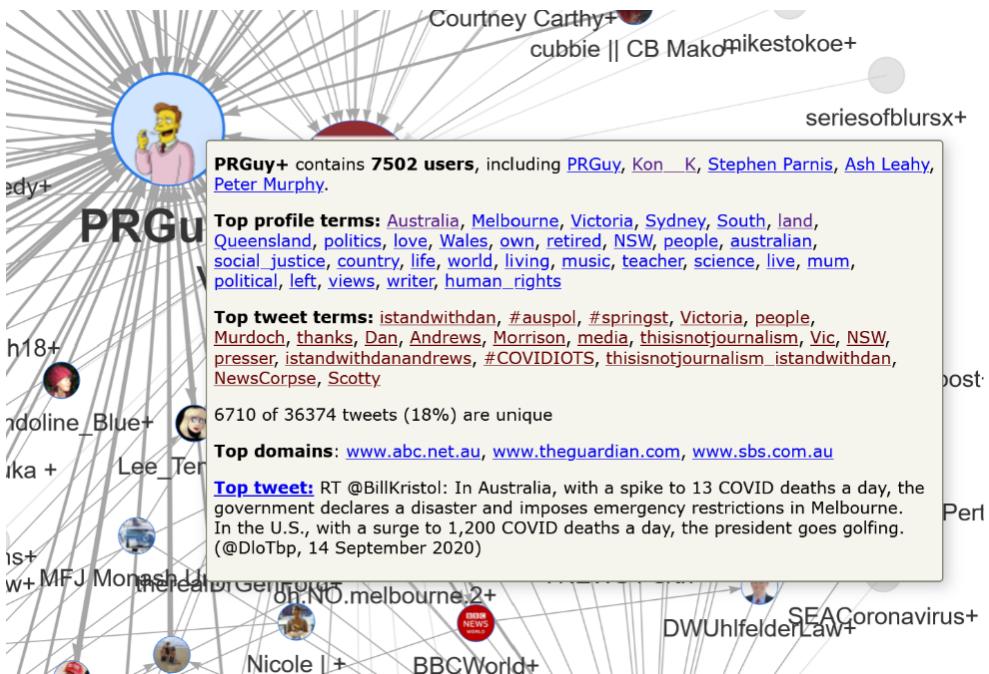
Jeder Cluster wird nach seinem höchst retweeted Benutzer benannt, während die Knotengröße reflektiert die Anzahl der Zeiten, in denen Nutzer aus anderen Clustern Mitglieder der Cluster.

Durch die Darstellung von Clustern anstelle einzelner Benutzer können wir die Hochebene Struktur des Netzes sowie die Namen und Profilbilder der die meisten einflussreichen Benutzer. In vielerlei Hinsicht ist es eine Cartoon-Version des gesamten Netzwerks: es entfernt viel von Detail und Nuance, aber erlaubt uns, die wichtigsten Dinge zu sehen deutlicher.

Unmittelbar zum Beispiel können wir sehen, dass viele der Nutzer über die Abschaltung zu diesem Zeitpunkt könnte in vier große Cluster gruppiert werden, in denen die Konten namens PRGuy, VicGovDHSS, ABC News und Avi Yemini waren die beliebtesten.

Darüber hinaus können wir sehen, dass die ersten beiden dieser Cluster eng mit einander (d.h. sie retweeten einander häufig), während die beiden letzteren looser mit dem Rest verbunden.

Aber wer sonst ist in diesen Clustern, und worüber twittern sie? Um es herauszufinden, Alles, was Sie tun müssen, ist, den Maus-Cursor über einen von ihnen zu schweben (in der interaktiven Version der Visualisierung), an welcher Stelle die folgenden Informationen angezeigt werden:



Die Pop-up-Informationen, die den PRGuy+-Cluster beschreiben.

Dieser Pop-up sagt uns zunächst, dass der Cluster rund um PRGuy 7,502 Nutzer enthält. Auch listet die fünf beliebtesten Benutzer im Cluster, bietet Hyperlinks zu jedem ihrer Twitter Profile.

Der Abschnitt „Top-Profil Begriffe“ enthält die prominentesten und markantesten Begriffe, die erscheinen in den Profilbeschreibungen von Benutzern im Cluster. Die Ortsnamen in dieser Liste Sagen Sie uns, dass die meisten Benutzer wahrscheinlich australisch sein. Inzwischen, Begriffe wie Soziale Gerechtigkeit, links, Menschenrechte, und vielleicht sogar Lehrer und Wissenschaft, schlagen Sie vor, die Benutzer in diesem Cluster sind meist links-Lernen in ihrer Politik.

Die Gegenwart der #IStandWithDan Hashtag an der Spitze der „Top-Tweet-Bedingungen“-Liste gibt Beweise, dass die meisten Benutzer in diesem Cluster unterstützend des viktorianischen premier, Daniel Andrews. Inzwischen, Begriffe wie Murtoch, NewsCorpse und Dies ist nicht Journalismus deuten darauf hin, dass diese Nutzer meist kritisch sind, wie Newscorp News-Outlets haben die Ausschlussdebatte abgedeckt. Dies kann durch einen Klick bestätigt werden jeder dieser Begriffe in der Visualisierung, die einen tatsächlichen Tweet von innerhalb der

Cluster, der den Begriff verwendet. Klicken Sie zum Beispiel auf den Begriff öffnet den folgenden Tweet:

Dies ist nicht Journalismus

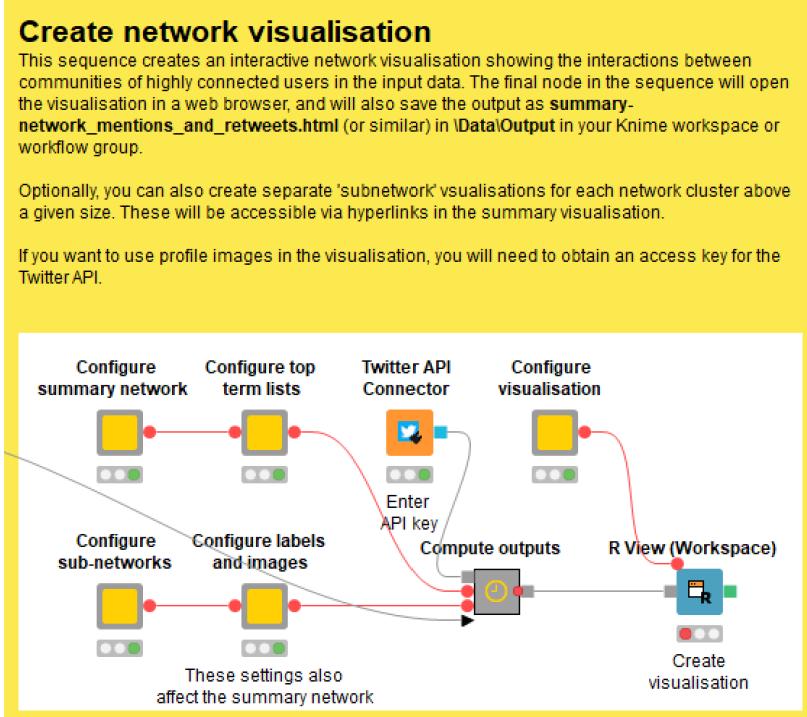
A screenshot of a Twitter post from user @PRGuy17. The tweet reads: "Threat to democracy" - News Corp journalists have been named and shamed for dangerous misreporting on Premier Dan Andrews during #COVID19Vic #ThisIsNotJournalism". It includes a photo of Daniel Andrews. Below the tweet, it says "sbs.com.au" and provides a link to an article. The tweet has 2.1K likes, 1 reply, and 140 replies.

Ein Tweet von @PRGuy17 mit dem Begriff ThisIsNotJournalism.

Die Zusammenfassungsinformationen für diesen Cluster zeigen auch, dass die drei Top-Websites in Tweets aus diesem Cluster sind die ABC, The Guardian und SBS – die alle werden wahrscheinlich von Menschen bevorzugt, die wary von NewsCorp Steckdosen sind.

Dies sind Erkenntnisse, die wir einfach nicht durch einen Blick auf ein Netzwerk in Gephi, selbst wenn wir den Rohdatensatz in der Nähe haben. Durch viel harte Arbeit für uns im Hintergrund (Berechnung von Termfrequenzen usw.) und Erstellung der Ergebnisse in eine interaktive Visualisierung, die TweetKolid R kann uns Stunden Zeit sparen, die würde andernfalls durch Tabellen von Tweets und Benutzerbeschreibungen Rummaging ausgegeben werden.

Viele Aspekte der Visualisierung, wie die Dauer der Termlisten und die Anzahl der Verbindungen, die dargestellt sind, können im Workflow einfach durch Doppel-Klicken Sie auf die unten gezeigten Komponenten.



Mit diesem Teil des Workflows können Sie das Netzwerk konfigurieren und generieren Visualisierung.

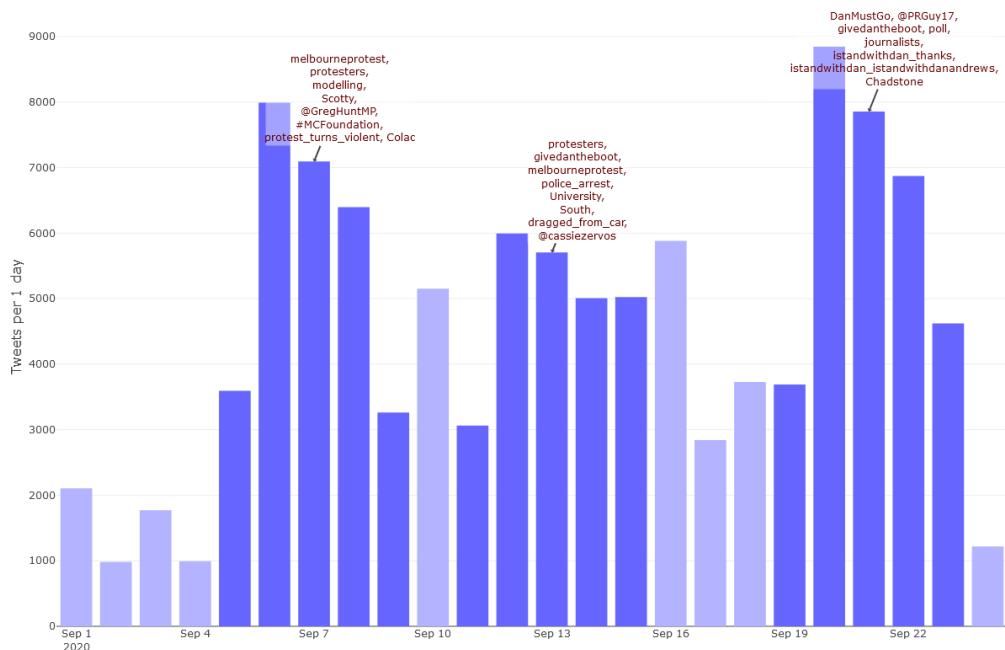
Das Beispiel über Highlights nur einige der Features des TweetKolidR Netzwerks Visualisierung. Um mehr zu erfahren, siehe mein Original[Wie geht's? R Blog Post](#)

Änderungen im Laufe der Zeit entdecken

Wie bei der Netzwerkvisualisierung die zeitliche Visualisierung des TweetKolidR enthält qualitative Informationen und interaktive Elemente, um ein reicheres Bild zu malen Veränderungen im Laufe der Zeit als mit einem herkömmlichen Zeitreihendiagramm erreicht werden können. Die interaktive Ausgänge werden durch das Plotly-Paket für R erzeugt. Das Beispiel unten zeigt die tägliche Anzahl von Tweets über die Melbourne Lockdown vom 1. bis zum [23. September](#). (Um die volle Funktionalität zu sehen, öffnen Sie die [interaktive Version](#))

Neben der Darstellung der Anzahl der Tweets in jedem Zeitschritt zeigt diese Ausgabe eine Liste prominente Begriffe aus jeder „Sprechzeit“ der Tätigkeit. Diese Spitzenzeiten, die schattiert dunkelblau, werden automatisch anhand einiger einfacher Kriterien erkannt, dass die Benutzer kann anpassen. Die Top-Bedingungen sind nicht einfach diejenigen, die am häufigsten in jede Spitzenperiode (zumindest nicht standardmäßig). Vielmehr werden sie nach ihren Frequenz sowie ihre Einzigartigkeit in der Periode. Darüber hinaus schließen sie Begriffe aus, dass in jedem einzelnen Zeitschritt erscheinen und die Aufnahme von mindestens einem Ort und Person oder Organisation.

Wie geht's? R in KNIME



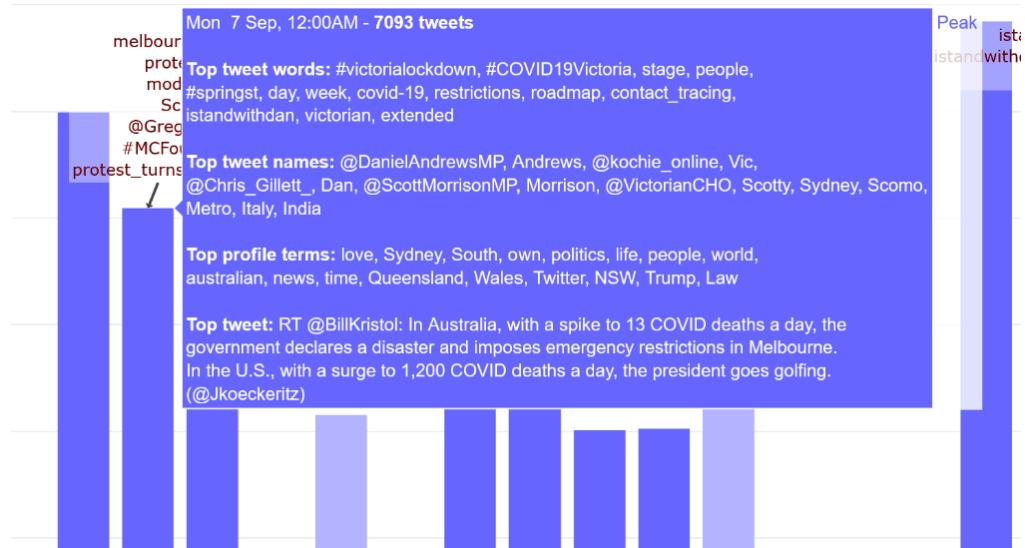
Ein auto-annotiertes Diagramm mit der Anzahl der Tweets, die täglich vom 1. bis 23. September veröffentlicht wurden.
Die Anmerkungen listen prominente Begriffe aus jeder schattierten Periode. Für die volle Funktionalität, siehe [interaktiv](#).

[Version](#) .

Der Zweck der Anmerkungen ist es, eine grobe Vorstellung davon zu vermitteln, was in jedem Periode der Spitzenaktivität. In diesem Fall werden die Begriffe für die erste Peakperiode (der Peak selbst war Sonntag, 6. September) schlagen vor, dass Proteste ein wichtiger Gesprächspunkt waren. sowie das Wort Demonstranten und der Hashtag #melbourneprotest, die Liste enthält die saftige n-Gramm, Protest_turns_violent. Auch diskutiert während und Zeitraum wurde modellings (spezifisch, die von der viktorianischen Regierung verwendet, um ihre Abriegelungsmaßnahmen), der Premierminister (bekannt informell von Australiern als Scotty), und der Bundesgesundheitsminister Greg Hunt. Die Begriffe in der interaktiven Form dieses Visualisierung sind alle Hyperlinks zu aktuellen Tweets, so können Sie auf sie klicken, um zu sehen, wie sie wurden tatsächlich verwendet.

Ein halbes Dutzend Wörter und Namen ist nicht viel, auf dem einen Eindruck von mehrere tausend Tweets über mehrere Tage. Mehr und mehr nuanced Charakterisierung der Aktivität im Laufe der Zeit, die Visualisierung bietet Popup Informationen, die jeden Zeitschritt zusammenfassen. Hier ist ein Beispiel:

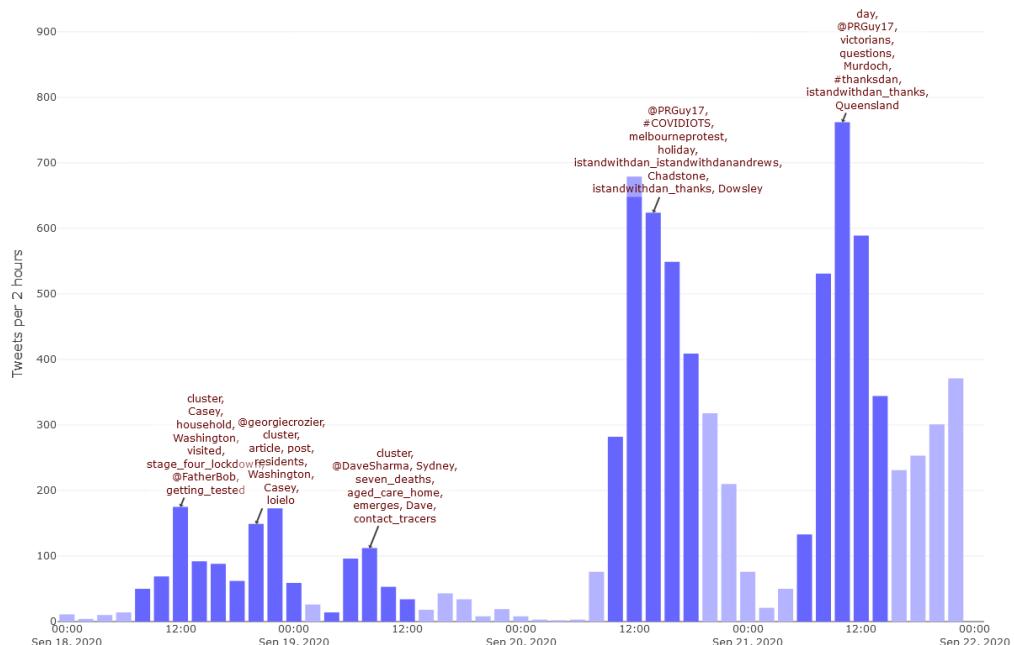
Wie geht's? R in KNIME



Zusammenfassungsinformationen stehen für jeden Zeitschritt in der Visualisierung zur Verfügung.

Diese Popups liefern ähnliche Informationen wie bei der Netzwerkvisualisierung:

separate Listen von prominenten Begriffen und Namen in Tweets verwendet, Top-Bedingungen von Benutzer Profile und die beliebteste Tweet aus der Zeit. Leider sind diese Listen nicht klickbar, weil Plotly (das R-Paket, das die Visualisierung erzeugt) nur ermöglicht es dem Popup-Text anzuzeigen, während der Cursor über der entsprechenden Bar ist.



Aktivität der Nutzer im PRGuy + Netzwerk-Cluster vom 18. bis 22. September.

Sie können die zeitliche Visualisierung konfigurieren, um einen beliebigen Zeitraum innerhalb Ihrer Daten anzuzeigen, mit Zeitschritten jeder Länge. Darüber hinaus können Sie die Daten nur auf die Benutzer filtern aus einem bestimmten Cluster aus der Netzwerkanalyse. Zum Beispiel das Bild oben fasst die Aktivität der Nutzer im PRGuy-Cluster zusammen (vorher diskutiert) vom 18. zum 22. September, mit 2-stündigen Zeitschritten.

Schlussfolgerung

Von den vielen Tools, die zur Erfassung und Analyse von Twitter-Daten zur Verfügung stehen, denke ich der TweetKolidR ist auf verschiedene Weise unverwechselbar. Soweit ich weiß, ist es einzigartig in Bereitstellung von Netzwerk- und Zeitreihen-Visualisierungen, die reich an qualitativen Informationen, die durch rechnerische Textverarbeitung abgeleitet werden. Diese Integration quantitative und qualitative Informationen werden durch zwei Dinge ermöglicht:

- Hintergrundtextverarbeitung durch KNIME und interaktive Visualisierungen generiert durch die visNetwork und Plotly R Pakete.

Zweitens ist der TweetKolidR für seine Nutzung der KNIME Analytics Platform als gegen R oder Python, die häufiger verwendet werden, um Twitter zu sammeln und zu analysieren Daten. Der Hauptvorteil der Verwendung von KNIME zu diesem Zweck ist, dass das resultierende Werkzeug durch eine rein graphische Schnittstelle betrieben werden, die auch von der freundlichste von R-Paketen oder Python-Skripten. Wenn auf diese Weise verwendet, KNIME effektiv löscht die Grenze zwischen der Benutzeroberfläche und dem zugrunde liegenden Code, so dass die neugierige Benutzer, um die inneren Arbeiten jederzeit zu sehen, und sie gegebenenfalls zu ändern.

Dann wieder die inneren Arbeiten des TweetKolid R sind nicht völlig codefrei, da sie enthalten mehrere R-Skripte. Der TweetKolidR ist ein Beispiel für eine Low-Code-Analyse Lösung, wo Code nur verwendet wird, wo es absolut benötigt wird. In diesem Fall ist der Code benötigt, um die interaktiven Visualisierungen zu erzeugen, die die Endausgänge des Workflows sind. Alle vorhergehenden Datenmanipulationen und Transformationen, einschließlich des Textes Verarbeitung, werden von KNIMEs nativer Funktionalität behandelt.

Code oder kein Code, der TweetKolidR ist ein inhärent komplexes Tier, wie Sie sehen, ob Sie blicken unter die oberen Schichten des Workflows. Aus diesem Grund ist es verpflichtet, ein paar Bugs, und es gibt immer Raum für Verbesserungen. Wenn Sie es versuchen, würde ich herzlich willkommen alle Fehlerberichte, Vorschläge oder andere Feedback.

Happy Kolliding!

Dieser Artikel wurde ursprünglich auf Angus' Blog veröffentlicht [in anderer Weg gesehen](#) und wir haben es aufgehoben in unserer [Niedriger Code für Advanced Data Science Journal](#) auf Medium. Der Artikel finden [Hier.](#)

Die entsprechenden [Wie geht's?](#) Der Workflow ist auf dem KNIME Community Hub zu finden [Angus' öffentlicher Raum](#)



[Philipp Kowalski](#) wurde nominiert KNIME-Beitrag der Monat für März 2021. Er wurde für seine [Gewinde](#) in dem er seinen kreativen Anwendungsfall mit KNIME für sein Hobby von Rollenspielen. Mit Hilfe von KNIME Analytics Platform hebt Heldensuche Sitzungen von Dungeons & Dragons. Das Bild rechts zeigt sein standard dice Rolling-Komponente, die mimics das Ergebnis der Würfelrolle. Dies ist wirklich ein einzigartiger Anwendungsfall, aber

[Forum](#)

es war nicht der erste und wird wahrscheinlich nicht das letzte Mal sein, dass Philipp herausgefordert die KNIME Community durch Drücken der KNIME Software auf die Grenzen.

Philipp ist ein No-Code/Low-Code-Enthusiasten und ein erfahrener Trainer. nicht nur [eigene und laufend](#)[BeschaffungZen](#) , [ein Blog und Podcast](#) zentriert auf Beschaffung und Verhandlung Strategien, aber auch einen YouTube-Kanal mit einer breiten Palette von Tutorials zur Nutzung von KNIME und bewährte Verfahren bei der Beschaffung. derzeit arbeitet als Digital Enablement Agent bei Siemens.

Besuchen Sie Philipp [Raum auf dem KNIME Hub](#) oder [Profil](#)

[Seite im KNIME Forum](#) (Hub/Forum Griff:

Kowisoft)



Digitalisierung Evangelist beschäftigt Automatisierung für Unternehmen und Dungeons & Dragons

My Data Guest — Ein Interview mit Philipp Kowalski

Autor: Rosaria Silipo



Es war mir ein Vergnügen, vor kurzem Interview [Philipp Kowalski](#) als Teil der [Mein Data Guest](#)
Interview-Serie. Er teilte Einblicke in seine Rolle als Evangelist der Digitalisierung, erklärt, wie KNIME eine effektive Prozessautomatisierung für Unternehmen vorantreiben kann und gab ein Beispiel der angewandten digitalen Transformation, um Fantasieszenarien für Dungeons & Dragons.

Philipp Kowalski arbeitet als Digital Enablement Agent bei Siemens. Er ist Experte KNIME Benutzer, seit 2018 mit KNIME für viele Datenanalyse-Anwendungen gearbeitet. Er ist auch ein Digitalisierungsexperte und wirklich leidenschaftlich über die Anwendung von Digital Transformation in alle Lebensfelder, vom Job bis zum Hobby, von der Datenanalyse bis kreative Disziplinen. Aber das ist nicht alles. Philipp ist auch ein erfahrener Trainer und besitzt und Läufe [BeschaffungZen](#), ein Blog und Podcast zentriert auf Beschaffung und Verhandlungsstrategien. Er hat einen YouTube-Kanal mit einer breiten Palette von Tutorials auf der Nutzung von KNIME und Best Practices in der Beschaffung. Philipp war auch einer der ersten [KNIME Beitrag des Monats](#), seit März 2021 für seine kreative Nutzung der KNIME Analytics Platform. Tatsächlich baut er KNIME-Workflows auf, um heroische Suche nach Zauberern, Zwergen, Elfen und Menschen für seine Sitzungen von Dungeons & Drachen.

Rosaria: Auf LinkedIn definieren Sie sich als Evangelist der Digitalisierung. Was ist das? Warum wäre die Digitalisierung so nützlich?

Philipp: Eine Digitalisierung evangelist ist jemand, der das Wort über Digitalisierung, insbesondere mit dem Fokus auf KNIME. Digitalisierung ist nützlich, weil es gibt noch so viele redundante und repetitive Aufgaben, die wir loswerden sollten, und bei zur gleichen Zeit haben wir heute so viele Daten zur Verfügung, dass wir in der Lage sind, zu erreichen Das. Die Zeit, die wir durch die Digitalisierung von Prozessen sparen, könnte genutzt werden, um die Daten und implementieren fortschrittliche Analyselösungen. In diesem Sinne ist KNIME wirklich

hilfreich, und ich zeige anderen, wie einfach es ist, es durch meine Arbeit zu verwenden. Mein ultimatives Ziel ist es zu zeigen, dass wir mehr oder weniger alle Spaziergänge des Lebens digitalisieren können.

Rosaria: Ist die digitale Transformation ein großer Teil Ihrer täglichen Arbeit bei Siemens? Verwenden Sie KNIME?

Philipp: Ja, absolut. Ich bin seit fast 20 Jahren ein Beschaffungsprofi.

Seit Februar 2022, in meinem täglichen Job bei Siemens, habe ich die Digitalisierung eingeführt im Bereich der Beschaffung. Meine Aufgabe ist es, meinen Kollegen KNIME beizubringen. Die große Vorteil der Verwendung von KNIME ist, dass der Lernprozess sehr reich, effektiv und schnell ist. Zum Beispiel starten Sie mit der Automatisierung, aber während Sie die Gebäudeautomation sind, Sie Lernen Sie auch andere Aufgaben und Knoten. Mit KNIME zeige ich meinen Kollegen auch, wie man digitalisieren viele redundante Beschaffungsprozesse. Das macht ihr Leben einfacher und besser.

Rosaria: Was sind die KNIME Features, die am meisten für die digitale Transformation helfen?

Philipp: Zu Beginn eines Projekts gibt es in der Regel mehrere ETL-Operationen ausführen. Daher benutze ich viele Datenmanipulationsknoten zur Reinigung, Vorverarbeitung, Zusammenfassung oder Export. Danach verlassen ich mich auf einige Erweiterungen, die sehr nützlich sind in der Beschaffung. Zum Beispiel die Continental-Knoten für XLS-Formatierung, BIRT für Berichterstattung, Lern-Prädiktor-Knoten, um Vorhersagemodelle zu erstellen, oder die Knoten des Textes Ausbau des Bergbaus.

Rosaria: Erzählen Sie uns von einigen typischen Anwendungsfällen in Ihrer Arbeit, wo Sie sich auf KNIME verlassen.

Philipp: Ein Beispiel ist die Meldung. KNIME ist ein fantastisches Werkzeug für die Berichterstattung, beides die BIRT-Erweiterung oder die Erstellung interaktiver Dashboards. Ein weiteres Beispiel ist Muster Anerkennung. Es ist wichtig, dass wir wissen, wann ein Kaufauftrag wahrscheinlich wird Problematisch. Die Identifizierung potenzieller problematischer Ordnungen in einer zeitnahen Weise ist sehr vorteilhaft für das Unternehmen, weil es uns anfordert, diese Aufträge mit mehr zu handhaben Empfindlichkeit. KNIME hilft uns viel, diese Muster zu entdecken und Verluste zu minimieren.

Rosaria: Wie fortgeschritten sollte Ihr Wissen über Datenanalysen in digital arbeiten Transformation?

Philipp: Ich wage zu sagen, dass Sie kein vorheriges Wissen brauchen. Der einzige Code, den ich je verwendet wurde, als ich Daten über das Web-Schrotten sammelte. Für alles andere, KNIME's Low Code-Ansatz ist wirklich genug.

Zusätzlich, wenn Sie in Daten mit Excel proficient sind, haben Sie bereits haben, was es braucht, um reibungslos und produktiv mit KNIME zu arbeiten. Aber KNIME hat wenige Schlüsselvorteile gegenüber herkömmlichen Tabellenkalkulationswerkzeugen. Zuerst die sequentielle, knotenweise Schritt für Schritt Ausführung in KNIME ist ein großer Vorteil, da Sie sofort die Ergebnisse nach jedem Knoten. In traditionellen Tabellenkalkulationstools werden Sie oft sehr geschrieben lange Formeln, und falls Sie eine Klammer oder Semikolon verpassen, ist es schwer, es zu realisieren.

Zweitens ist die KNIME-Gemeinschaft fantastisch. Alle in der freundlich, hilfsbereit und extrem schnell. Es ist nie passiert, dass ich eine Frage auf das KNIME Forum und es wurde nicht gelöst. Ich glaube, das ist ganz besonders.

Anmerkung: [Lesen Sie „Zehn häufige Probleme bei der Verwendung von Excel für Datenoperationen und „Excel zu KNIME“](#) “ [Handbuch zu migrieren schmerzlos von Excel zu KNIME.](#)

Rosaria: Reden wir über Geld und Zeitzersparnis. Wie wichtig ist die Wirkung von Digitalisierung im Geschäft? Was sind einige unmittelbare Auswirkungen, die Sie erlebt haben in Ihrer Erfahrung?

Philipp: KNIME hilft uns, viel Zeit zu sparen, indem wir die Digitalisierung mehrerer Geschäftsprozesse. Zum Beispiel reduzierte der erste Workflow, den ich gebaut habe, meine Berichterstattung, von 3 Stunden im Monat bis 10 Minuten. Für ein anderes Projekt automatisierten Journaling der aktuellen Anforderungen, und dass reduziert den Aufwand pro Mitarbeiter von je 30 Minuten zu fast nichts. Für fünf Mitarbeiter, die täglich 2,5 Stunden sind. Ich glaube, die Beispiele sind ziemlich signifikant.

Rosaria: Es ist jedoch nicht immer der Fall, dass die Analytik reibungslos verläuft. Manchmal Fehler sind gut, denn dann können wir mehr erfahren. Erzählen Sie uns von dem größten Fehler Sie haben gelernt.

Philipp: Ich wollte einmal Lieferantendaten über Web-Schrott extrahieren. Mein Ansatz war, direkt die spezifischen Informationen, die ich von dieser Webseite benötigt, die ich konnte es nicht schaffen. Ich schaute endlich ins KNIME Forum und erhielt große Unterstützung von einigen Forum-Mitglieder. Sie schlugen einen anderen Ansatz vor, der sich als sehr effektiv. Ich sollte zuerst alle Daten von der Website abkratzen und dann nach unten schneiden und sie zu dem, was ich brauchte. Dies ist meine Standardstrategie für viele Daten geworden Probleme.

Rosaria: Lassen Sie uns über Ihre Tätigkeit als YouTuber sprechen. Seit wann bist du ein YouTuber?

Philipp: [Mein Gott YouTube Kanal](#) [existiert seit Juni 2018. Ich habe ein Video hochgeladen](#) jetzt und dann. Seit April 2021 bin ich aktiver und ich begann zu schaffen Inhalte, Kurse und Video-Tutorials mit Schwerpunkt KNIME.

Rosaria: Sie haben zwei Kurse über KNIME auf YouTube. Wie unterscheiden sie sich? Was? Kurs sollten Menschen folgen und warum? Sind sie alle völlig frei?

Philipp: Eigentlich ist es nur ein Kurs. Ich habe das Material ursprünglich auf meiner Website gehostet aber dann entschied ich mich, es auf YouTube zur Verfügung zu stellen, so dass jeder darauf zugreifen konnte. Die Kurs geht von grundlegenden bis fortgeschrittenen Themen und ist speziell auf Zielkomplette KNIME Anfänger und geben ihnen die Möglichkeit zu entdecken, wie groß KNIME ist es. Zuerst stelle ich theoretische Konzepte vor (z.B. warum Automatisierung ist

vorteilhaft, oder was ETL ist), dann beginnen wir mit ein paar Basis-Knoten, bauen wir ein Workflow zusammen, wir erkunden erweiterte Funktionen, etc. Probieren Sie es!

Rosaria: Welche Quellen (z.B. Bücher, Artikel, Blogs, Blueprints, Forum, Social Media, etc.) verwenden Sie, um auf dem neuesten Stand über KNIME und Datenwissenschaft?

Philipp: Alle Arten von Quellen, die von KNIME bereitgestellt werden, sind nützlich und tragen zu einer breiteren Verständnis von Datenwissenschaft und Werkzeug. Die [Datengespräche](#) und [Daten verbinden](#) werden sehr aufschlussreiche Ereignisse, [KNIME YouTube Kanal](#) hat viele hilfreiche Inhalte, und ich immer empfehlen, einen Blick auf die [KNIME Blog](#) und das [KNIME Forum](#).

Rosaria: Wir erreichen das Ende unseres Interviews. Bevor wir uns verabschieden, würde ich wirklich gerne über das Projekt zu sprechen, das Sie den Beitrag des Month-Preises im März verdient 2021: Anwendung von KNIME zur Analyse von Daten für Dungeons & Dragons. Erzählen Sie uns mehr darüber Projekt.

Philipp: Ich bin ein Fantasie-Nerd, und ich habe gespielt [Dungeons & Dragons](#) (ein Rollenspiel) Spiel seit über 30 Jahren. Was mich erstaunt, ist, dass es heute möglich ist, mit Menschen auf der ganzen Welt mit sogenannten virtuellen Tabellen. Andererseits hatte ich auch viele Regelbücher, aber ich wollte nicht die gleichen Dinge immer wieder spielen. Ein Tag, es schlug mich wie Blitz und ich dachte: "Hey, ich kenne ein Werkzeug, das mir helfen kann, zu vermeiden repetitive Aufgaben". Also habe ich mich herausgefordert und versucht, ob ich neue Fantasie aufbauen könnte Welten mit KNIME. Grundsätzlich eine digitale Transformation von Dungeons & Dragons.

Rosaria: Sie haben keine Daten analysiert. Sie nutzten KNIME zur Generierung Szenarien für Ihr Spiel. Hat KNIME dabei geholfen, die Aufgabe schneller oder besser zu machen? Wenn besser, zu Wie weit?

Philipp: Mit KNIME verbesserte sich sowohl die Qualität als auch die Geschwindigkeit des Prozesses. Das ist weil ich die Szenarien nicht mehr aufschreiben musste, aber ich konnte exportieren sie automatisch in ein schönes PDF-Format, das ich dann einfach mit meinem teilen konnte Spieler.

Rosaria: Wie viele Attribute sollen Sie für jeden Dungeons generieren & Drachensitzung? Ich meine, Zeichen, Charaktermerkmale, Ort, etc.

Philipp: Zwischen 100 und 150 verschiedene Attribute aus allen Arten von Tabellen werden erstellt für jede Sitzung. Interessant ist, dass diese Tabellen miteinander verbunden sind. Das bedeutet, dass, wenn Sie eine 4 auf Tisch A rollen, müssen Sie auch auf Tisch B rollen. KNIME ist extrem stark in regelbasierten Anwendungsfällen, die ich mein kleines Siedlungen. Es hat wirklich einige sehr schöne Rollenspiel-Runden gefunkt und es war eine Menge Spaß!

Anmerkung: [Weiterlesen über Philipp's KNIME Forum.](#)

[Dungeons & Dragons Projekt](#)

[in seinem Beitrag auf dem KNIME Forum.](#)

Rosaria: Sir Philipp, eine letzte Frage. Wie können Menschen aus dem Publikum in Kontakt treten mit dir?

Philipp: Sie können mich über [LinkedIn](#) oder [YouTube Kanal](#) . Lass einen. kommentieren Sie in einem meiner Videos, und ich werde Sie erreichen. Alle meine Workflows sind verknüpft auch auf YouTube.

Dieser Artikel wurde erstmals in unserem [Niedriger Code für Advanced Data Science Journal](#) auf Medium. Die Originalversion finden [Hier.](#)

Sehen Sie sich das ursprüngliche Interview mit Philipp Kowalski auf YouTube an [Mein Data Guest – Ep 6 mit Philipp Kowalski](#) „

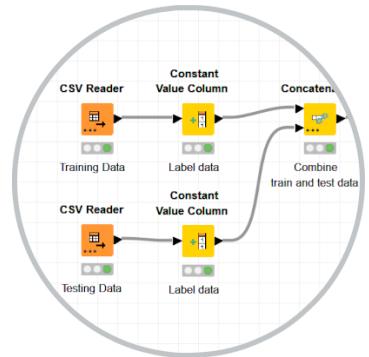


Tosin Adekanye wurde nominiert Beitrag des Monat für August 2021. Sie wurde für ihr wertvolles Beiträge zur [KNIME Community Hub](#), sie unzählige [Social Media Beiträge](#), und zur Moderation der [KNIME, Excel & Reporting Benutzergruppe](#) in der letzten [KNIME Datengespräche – Community Edition](#). Einige von ihr am meisten Zu den beliebten Beiträgen gehören die Arbeit an Daten, die mit der FIFA Arabischen Tasse, Finanztransaktionen, UFO

Sichtungen und Betrugsdetektion. Das Bild rechts zeigt einen Schnipsel von ihr [Nachweis-Workflow](#). Viele von ihnen werden auf ihren Social-Media-Kanälen oder in unserem Low Code for Advanced Data Science Journal on Medium. Sie hat auch lancierte die Serie „30 Days of KNIME“ auf YouTube. Sie hat ihre Mission gemacht ihr Wissen und ihre Erfahrung in der Datenwissenschaft allen interessierten Profis, Entfernen der Codierung Barriere und Fokus auf Konzepte.

Tosin hält einen Master of Business Administration mit dem Fokus auf Business Analytics der Universität Tampa. Ihre akademische Forschung Hintergrund gab sie einen weichen, aber soliden Eintrag in die Datenwissenschaft. Sie ist jetzt versuchen, das Feld für alle zugänglich zu machen indem man Konzepte auf eine Weise erklärt, die jeder kann verstehen. Das ist einer der Fahrer hinter ihr Leidenschaft für die No-Code / Low-Code-Umgebung. Sie ist derzeit Data Scientist im Qatar Financial Centre Regulierungsbehörde.

Besuch von Tosin's [Raum auf dem KNIME Hub](#) oder sie [Profil Seite im KNIME Forum](#) (Hub/Forum Griff: Tosinlitics)



Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und Weitere wichtige Fragen zur Datenwissenschaft

My Data Guest — Ein Interview mit Tosin Adekanye

Autor: Rosaria Silipo

Anmerkung des Herausgebers:

Einige der in diesem Artikel verknüpften Knoten im Workflow sind auf einer externen Software verfügbar

Aktualisierung der Website.



Es war mir ein Vergnügen, vor kurzem Interview [Tosin Adekanye](#) als Teil der [Mein Data Guest](#) Interview-Serie.

Tosin Adekanye begann ihre Reise in die Datenwissenschaft während ihres Studiums in der Psychologie, wo sie stark an der Statistik und akademischen Forschung beteiligt war. Sie entwickelte diese Fähigkeiten während ihrer MBA weiter, wo sie begann zu lernen und Nutzung von ML-Algorithmen und Data Science-Software. Aber es war während der Abschaltung, zuletzt Jahr, dass sie mit der Datenwissenschaft besessen wurde! Und das ist auch, wenn sie Lernen Sie den No Code/Low Code "Powerhouse" – wie sie KNIME Software nennt.

Tosin ist ein sehr aktives Mitglied der KNIME Community und ein Influencer in den Daten Wissenschaftsraum auf sozialen Medien. Sie schreibt Artikel für eine Reihe von Zeitschriften, wie die Zeitschrift "Low code for Advanced Data Science" auf Medium, sehr interessant Lösungen für die unterschiedlichsten Aufgaben. Ich wiederhole „zu den unterschiedlichsten Aufgaben“. Zum Beispiel in einem Artikel stellt sie eine Lösung für die Betrugserkennung bei Kreditkartentransaktionen zur Verfügung; ein weiterer Artikel, den sie über UFO-Sichtungen spricht; dann beschreibt sie in einem anderen Artikel die Möglichkeit, SQL-Codierung mit visueller Programmierung zu umgehen, und so weiter. Sehr vielfältig Projekte.

Rosaria: Hi Tosin, Erzählen Sie uns mehr über diese verschiedenen Projekte. Welches der Projekte Sie haben Artikel über waren für die Arbeit und welche sind Hobbys?

Tosin: Einige der Artikel wurden von Arbeitsprojekten motiviert. Als ich in FIS war, die Globale Kreditkartenverarbeitungsfirma, ein Kollege in der Betrugsabteilung sagte mir

Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und andere wichtige Data Science Fragen

wie sie Modelle verwenden, um Kreditkartenbetrugs vorherzusagen. Das hat mich neugierig gemacht
um so etwas zu bauen Schreiben Sie es , Ein weiterer arbeitsinspirierter Artikel
kam, weil ich täglich Datenbanken verwendet. Das führte zu dem Artikel über SQL :
Die UFO-Artikel Andererseits, das ist ein Hobby von mir. Ich bin besessen von
Science-Fiction und fremde Sichtungen. Aber es ist Hobbys, die mich auch motivieren, ins Blick
verschiedene Themen.

Rosaria: Ich möchte über Ihren Artikel sprechen, um Kreditkartenbetrug vorherzusagen. Können Sie uns sagen?
ein bisschen mehr über Ihr Interesse an der Arbeit mit Finanzdaten?

Tosin: Ja, ich mag Finanzen. Mein Hintergrund ist eigentlich Psychologie und Geschäft,
aber Finanzen sind eine Synergie aller meine Leidenschaften. Betrug ist ein echter Schmerz für beide Kunden
und das Unternehmen. Es kostet so viele Unternehmen in den USA so viel Geld... ich war
neugierig zu sehen, ob ich ein Modell bauen könnte, das diese betrügerischen Transaktionen fangen könnte
und vielleicht sagen Sie uns, welche Variablen mit Betrug zu gehen scheinen. Der verwendete Datensatz war
synthetische Daten, aber es basierte auf realen Trends in diesem Bereich.

Zum Beispiel ist jemand mit einer hohen Wahrscheinlichkeit für Betrug verbunden: Fraudsters
gerne auf potenziell glaubwürdige, ältere Menschen. Online-Betrug findet auch mehr statt
Nacht oder früh morgens. bewusst zu sein, wie dies Unternehmen und
Kunden schützen sich besser.

Rosaria: Haben Sie jemals versucht, Aktienkurse vorherzusagen? Und denken Sie, das ist eigentlich
möglich?

Tosin: Lassen Sie mich etwas zurückspulen! Während meiner MBA in Finanzen haben wir versucht, Aktienkurse vorherzusagen
auf Basis von Dingen wie Finanzquoten aus vergangenen Unternehmensaussagen. Unser Bestes
Modelle wahrscheinlich auf eine Genauigkeit von 10 bis 14 vorhergesagt und nur für eine bestimmte
Varianz. Ich glaube, dass Sie einen Aktienkurs wirklich vorhersagen müssen, um ganzheitlich zu sein.
Sie müssen alles von Nachrichtenartikeln bis zu Social Media Posts bis zu historischen
Aktienpreise an die Finanzen des Unternehmens. Ich arbeite immer noch aktiv daran. Übrigens,
Dieses Projekt war auch der Grund, warum ich mit KNIME begann. Ich brauchte etwas, das
lässt mich Daten aus all diesen Quellen verarbeiten.

In der Vergangenheit habe ich auch eine Analyse zwischen Stimmung auf Twitter und der Bewegung durchgeführt
von Vgl. Werte, und einige Beziehungen dort erkannt. Ja, ich glaube, es ist möglich,
die Aktienpreise vorhersagen. Definitiv nicht zu 100% Genauigkeit, aber genug Daten von Losen gegeben
von verschiedenen Quellen und guten Modellen, ja es ist möglich, mit einer gewissen Genauigkeit vorherzusagen.

Rosaria: In Ihrem Artikel zur Betrugserkennung haben Sie die Leistung von mehreren verglichen.
Modelle (Decision Tree, XGboost, Gradient Boosted Bäume) und erklärt den XGboost
Modell, um das Beste zu sein. Warum denken Sie, dass XGBoost besser als der andere durchgeführt
Algorithmen in diesem speziellen Anwendungsfall?

Tosin: Für die Klassifizierung XGboost tendiert besser aus mehreren Gründen. Es ist ein
Ensemblemodell, so dass es mehrere Bäume baut, die in der Regel eine bessere Leistung haben.

Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und andere wichtige Data Science Fragen

Auch, XGboost verbessert auf schwachen Lernenden - auf Funktionen, die nicht zu gut führen Vorhersagen. Es wird immer besser, wie es vorangeht. Es begrenzt auch die Überbelegung von Einführung einer Kostenfunktion, so dass alle diese verschiedenen Steuerungen in der Regel machen es ausführen besser als andere Modelle.

Rosaria: Sie sagen, es war Ihre Arbeit mit Datenbanken, die Sie dazu führte, den Artikel zu schreiben „in SQL oder nicht in SQL?“. Das ist ein umstrittenes Thema. Einige Leute sagen: "Du musst SQL andernfalls vergessen Sie, wie man in SQL programmiert". Andere sagen: "Wenn ich nicht SQL kann, dann visuell Programmierung erlaubt mir sowieso SQL". Was ist Ihre Meinung?

Tosin: Oft fühle ich, dass in diesem Bereich der Datenwissenschaft Menschen zu werden können an Werkzeugen befestigt. Ich würde sagen, tun, was am besten für Sie funktioniert. Wenn Sie Code sehen möchten und es ist das Beste für Sie dann tun, dass! Ich habe lieber einen gemischten Ansatz. Ich habe Workflows mit vielen SQL-Code, aber Sie werden mich auch mit SQL-Knoten sehen [z.B., die Knoten in den KNIME Datenbank Erweiterung der SQL-Abfragen im Hintergrund] denn oft funktioniert das am besten für mich. Was immer am besten funktioniert und was auch immer am effizientesten für Sie ist, was ich empfehlen würde.

Rosaria: Vor einiger Zeit hat jemand auf Twitter geschrieben, dass die Leute Low-Code-Tools verwenden nur, wenn sie von ihren Bossen erzwungen werden, würden die Menschen sie nie für ihre Hobbys nutzen. Sie scheinen dem Tweet widersprechen. Nutzen Sie die KNIME Analytics Platform für alle Ihre Projekte?

Tosin: Ich bin mir nicht sicher, wie sie diesen Standpunkt erreicht haben. Ich habe KNIME an meinen Arbeitsplatz gebracht! Ich benutze KNIME nicht ausschließlich, aber es wäre schwer für mich, irgendwo hinzugehen und nicht KNIME nutzen können.

Rosaria: Sie nutzen also die KNIME Analytics Platform in Kombination mit anderen Tools?

Tosin: Ja, normalerweise Power BI und KNIME. Manchmal benutze ich PyCharm, wenn ich muss Programm in Python. KNIME hat einen Python-Knoten, so dass Sie auch dort programmieren können, aber ich benutze normalerweise KNIME, PyCharm und Power BI.

Rosaria: Ihre andere große Leidenschaft neben der Datenwissenschaft ist Science Fiction. In Ihrem UFO Artikel, Sie fanden einen Weg, sie zu kombinieren. Erzählen Sie uns mehr über diese Geschichte.

Tosin: Ich bin ein großer Fan von Sci-Fi und Psychologie. In der Psychologie bemerken wir oft, dass mehr wird darüber gesprochen, je mehr Menschen es erleben. Also fragte ich mich wenn es irgendeine Beziehung zwischen Sichtungen von UFOs und Filmen geben könnte über UFOs. Basierend auf Daten von UFO-Sichtungen und Veröffentlichungsdaten von Filmen über Aliens Ich habe die Korrelation visualisiert. Und die Korrelation war sehr wichtig! Aber... Natürlich ist das keine Betrügerei. Ich würde wahrscheinlich mehr graben und bekommen müssen bessere Daten - eine reichere Filmdatenbank plus neuere UFO-Sichtungen und dann sehen, ob ich kann isolieren, was das verursacht.

Rosaria: Wie hat KNIME Ihnen geholfen, die Beziehung zwischen der Anzahl der Filme und viele Sehenswürdigkeiten?

Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und andere wichtige Data Science Fragen

Tosin: Ich hatte mehrere Datensätze, so dass KNIME mir wirklich geholfen, sie zu mischen. Der Film Datensätze sind ziemlich anders, so musste ich etwas Standardisierung tun und alles beitreten zusammen. KNIME war sehr hilfreich für dies sowie für Korrelation. Die Korrelation war super einfach zu laufen [Lineare Korrelation](#) Knoten und dann konnte ich schnell die Beziehung betrachten.

Anmerkung: [Überprüfen Sie den Workflow](#) [UFO Anzeigen Daten Prep](#) [in Tosins Raum auf der KNIME Community Hub.](#)

Rosaria: Planen Sie mehr Artikel wie diese zu schreiben? Ich folge dir und ich kann nicht warten, bis der nächste erscheint.

Tosin: Ich habe mich nicht ganz entschieden über meinen nächsten Artikel, aber es gibt zwei Dinge, die ich arbeite auf. Eins ist Flug Stornierungen - betrachten Sie, welche Faktoren mit Flug Stornierungen gehen und wenn die Flüge höchstwahrscheinlich abgesagt werden.

[Mein anderes Projekt macht Spaß! Da ist ein Twitter API Connector](#) [Knoten in KNIME, macht es](#) super einfach, Tweets zu ziehen. Also werde ich Tweets aus verschiedenen Ländern bekommen, die das Wort Glück nennen und dann sehen, welche Wörter am meisten mit Glück in verschiedenen Teilen der Welt.

Rosaria: Wie können Datenwissenschaftler im Publikum Ihrer Arbeit folgen und auf Ihre Workflows?

Tosin: Ich schreibe Artikel für KNIME's [Niedriger Code für Advanced Data Science](#) [Zeitschrift auf Medium und ich haben auch meine Website](#) [TosinLizenz wo ich meine Arbeit veröffentlichte. Ich kann empfehlen](#) [KNIME Hubraum](#) [wo ich meine Workflows halte. Ich glaube, es ist](#) sehr hilfreich, um zu sehen, was andere Menschen getan haben - nicht nur meine Workflows, sondern alle Workflows im Allgemeinen auf dem KNIME Hub. Es ist toll für Ideengenerierung oder um Ihnen zu helfen Ich werde mich enttäuschen. In der Lage, sich auf Beispiele von anderen zu beziehen hilft viel.

Rosaria: Ihre TomTom-Komponente ist in Ihrem KNIME Hubraum, richtig? Es war sehr beliebte Komponente auf dem KNIME Hub. Was macht das? Kann ich es herunterladen?

Tosin: Ja, ich war motiviert, dies zu tun, weil ich nicht wirklich viele Lösungen um dass Sie schnell den Abstand zwischen zwei Punkten mit Länge und Breite. Ich habe einige Recherchen gemacht und dachte, TomTom war das beste Werkzeug, um mit zu gehen. Also jetzt Sie erhalten nur Ihre API-Schlüssel und setzen es in die Komponente. Sobald Sie Ihre Datendateien haben die Längsinformationen enthalten, können Sie diese durch das Bauteil ausführen und die Fahrzeit und die Distanz, die Verkehrsverzögerungen und alle damit verbundenen Informationen, zu gehen von einem Punkt zum anderen.

Anmerkung: [Download der Komponente](#) [Fahrzeit- und Fernabfrage – Breite](#) [Länge](#) [aus dem KNIME Community Hub.](#)

Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und andere wichtige Data Science Fragen

Rosaria: Planen Sie mehr Komponenten zu implementieren?

Tosin: Ja. Es gibt so viel Sie von der TomTom API bekommen können! Ich will eine noch ein paar. Dies könnte eine Familie von geospatialen Komponenten sein.

Rosaria: Sie sind ein sehr aktives KNIME Community-Mitglied, aber wir gehen zurück in die Zeit: Wie Seit langem hast du KNIME benutzt und wie hast du mit KNIME angefangen?

Tosin: Das mag eine Überraschung sein, aber ich habe erst im Januar angefangen, KNIME zu benutzen. 2021. Ich hatte bereits eine Exposition gegenüber Software wie KNIME - ich verwendete SPSS Modeler ab 2017. Dann benutzte ich Alteryx, aber die Lizenzierung war eine Barriere für mich. Ich brauchte etwas, das effizient war, das mir so viele Dinge für die Datenwissenschaft machen konnte. Da habe ich KNIME gefunden.

Obwohl ich KNIME nicht so lange benutzt habe, kannst du wirklich das Lernen klettern schnell aufgrund der verfügbaren Ressourcen. KNIME hat auch einige der am meisten ansprechbar, leidenschaftlichste Mitarbeiter und das ist wirklich half mir kommen in meiner Lernkurve.

Rosaria: Erzählen Sie uns von der größten Herausforderung, die Sie in Ihrem Berufsleben lösen mussten, als einen Datenwissenschaftler.

Tosin: Umgang mit Textdaten! Ich hatte viele davon weggelaufen Jahre. Aber im Januar wollte ich lernen, wie man Textdaten verarbeitet und analysiert und sich damit wohlfühlen. Beispiel-Workflows haben viel dazu beigetragen, wissen Sie. Ich wollte etwas sentimentale Analyse machen. Ich erkannte, dass mit Textanalyse einmal die Daten richtig gereinigt und verarbeitet, kann es auf Mathematik reduziert werden! Jetzt nur scheint viel einfacher zu sein.

Rosaria: Sagen Sie uns den größten Fehler, von dem Sie gelernt haben.

Tosin: Mein größter Fehler war die Grundlage für einen Artikel, den ich über Klassenungleichgewicht geschrieben habe und warum die Genauigkeit nicht immer das Beste ist - vor allem, wenn Sie unausgeglichen haben Klassen. Ich habe über mein erstes Erkennungsmodell geschrieben, weil es eine Genauigkeit von 99% hatte. Die Linked In der Data Science-Crowd war super hilfreich, weil sie darauf hingewiesen, dass wenn Sie unsymmetrische Klassen haben, ist die Genauigkeit nicht unbedingt die beste Metrik. In der Tat können Sie hohe Genauigkeitswerte haben, aber Ihre Minderheitenklasse kann immer noch wirklich schlecht in Bezug auf die Klassifizierung. Das war etwas, das ich wusste, aber manchmal Sie Wissen Sie etwas in der Theorie, aber Sie wissen es nicht wirklich, bis Sie sehen, dass es in Praxis

Rosaria: Ja, unausgeglichene Klassen können Ihnen falsche Erwartungen darüber geben, wie es funktioniert. Haben Sie einen Rat für alle jungen aspirierenden Datenwissenschaftler, die im Publikum sind?

Tosin: Ich habe drei primäre Ratschläge:

Zu SQL oder nicht zu SQL, UFOs, Sci-fi Movies – und andere wichtige Data Science Fragen

- Eins, viel lesen. Medium ist eine gute Plattform. Sie müssen nicht vollständig verstehen alles, was Sie lesen, aber Sie werden mit dem Thema vertraut gemacht werden und dies wird in die Zukunft.
- Haben Sie auch keine Angst, Ihre Arbeit zu teilen. Die ersten Dinge, die ich geteilt habe, waren nicht immer dass gut, aber wenn ich sie geteilt habe, bekam ich Feedback, die mir geholfen, mehr zu verbessern.
- Fang einfach an! Du musst nicht perfekt sein, aber du wirst wachsen und bauen von dort.

Rosaria: Jedes Buch zu empfehlen für die im Publikum immer eager zu lernen

Etwas Neues?

Tosin: Ja. [Datenanalyse Ausverkauft](#) von Andrea De Mauro. Ich mag es wirklich, weil es lehrt Sie die Theorie für Analytik und für die Datenwissenschaft, die so wichtig ist. Manchmal springen Programme für die Datenwissenschaft direkt in Python, aber ich denke, die Theorie ist wichtiger. Am Ende des Tages ist Python nur ein Werkzeug. Wenn Sie lernen, Theorie, dieses Wissen hilft Ihnen zu wissen, wie Hindernisse in der Praxis zu überwinden und sein besser in diesem Bereich. Dieses Buch lehrt auch, wie man KNIME verwendet.

Rosaria: Bücher sind auf jeden Fall ein wesentliches Werkzeug, um eine solide Basis zu erhalten, aber was sind Ihre üblichen Lesungen, um Sie auf dem Laufenden zu neuen, spannenden Datengeschichten zu halten?

Tosin: Ich las eine Menge Medium Artikel und ich bin sehr aktiv auf LinkedIn, verbunden mit Leute. so sehe ich in der Regel viele Dinge, von denen gesprochen wird. In der Schleife bleiben, Lesen und googling hilft, auf dem neuesten Stand zu halten.

Rosaria: Während ich deiner #30daysofknime-Initiative folgte, entdeckte ich auch, dass du ein sehr talentierter Videomacher und eigentlich... "Überraschen" ... ein sehr talentierter Sänger. Würden Sie möchten dieses Interview mit dem Song, den Sie sang in dem ersten Video veröffentlicht innerhalb die #30daysofknime Initiative?

Finde Tosins Lied – im Video des Originalinterviews in der gelben Box unten verknüpft!

Dieser Artikel wurde erstmals in unserem [KN-Code Blog](#). Die Originalversion finden [Hier.](#).

Sehen Sie sich das ursprüngliche Interview mit Tosin Adekanye auf YouTube an [Mein Data Guest – Ep 2 mit Tosin Adekanye](#).

Darüber hinaus spielte Tosin auch in der fünften Folge der My Data Guest Serie. Sehen Sie die volle Interview hier: “ [My Data Guest – Ep 5 Women in Tech](#)”

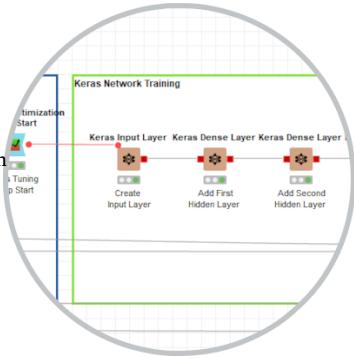


[Paul Wisneskey](#) wurde nominiert Beitrag des Monat für März 2022. Er wurde für seinen zweiteiligen Artikel über American Football, wo er erforscht, ob er kann sein Machine Learning Modell lehren, um einen Pass zu machen: den Ball in Bewegung bringen. Teil 1 behandelt [Parameteroptimierung mit KNIME](#), und Teil 2 [Züge a Neural Network](#). Der Workflow-Schnipsel rechts erfasst den Netzwerktrainingsteil. Alles in allem, Paul ist ein sehr

aktives Mitglied der KNIME Community und hat zahlreiche Blog-Posts geschrieben über KNIME. Viele von ihnen sind in seinem Medium Katalog, die definitiv wert sind einen Blick.

Paul ist ein Softwarearchitekt mit über 25 Jahren Erfahrungsplanung und Implementierung von großen, zuverlässigen Systemen für Big Data, Suche und Analyse, online Informationsdienste, Dokumentdarstellung und verteilte Zusammenarbeit. Er hat umfangreiche Java, Big Data, Web Services und Endeca Suchmaschinenerfahrung. Paul ist derzeit Direktor von Engineering bei BigBear.ai, eine Firma, die AI-leistungsfähige Analytik und Cyber-Engineering-Lösungen die missionskritischen Operationen und die Entscheidung zu unterstützen in komplexen, realen Umgebungen.

Besuchen Sie Paul's [Raum auf dem KNIME Hub](#) oder [Profil in die KNIME Forum](#) (Hub/Forum) Griff:
PWisneskey)



Der Pass, Teil 1: Parameteroptimierung mit KNIME

Von Fußball zu synthetischen Daten für Modellausbildung

Autor: Paul Wisneskey



Foto von [Chris Chow auf Unsplash](#):

Mit dem Super Bowl um die Ecke hier in den Vereinigten Staaten, fand ich vor kurzem gefunden Ich versuche, an einen fußball-themed Blog-Post zu denken. Basierend auf dem Fußballthema, Ich entschied mich, einen Beitrag über einen grundlegenden Aspekt des Spiels zu erstellen: werfen den Ball zu einem Empfänger (a „Reisepass“ in American Football parlance). Es ist eine einfache Operation, mit ein wenig Praxis, die meisten Menschen können intuitiv ohne Kenntnisse von die physikalischen Gleichungen, die die projektile Bewegung der Kugel und die lineare Bewegung bestimmen des Empfängers. Vielleicht wäre es möglich, maschinelles Lernen zu verwenden, um meinen Laptop zu lehren wie man einen Pass zu einem bewegten Ziel abschließt, ohne nur die Programmierung in der Physik Berechnungen?

Betrachten Sie dies in der einfachsten Weise, ich möchte, dass mein Laptop einen Haufen von zufällige Wurfs und lernen, wenn sie erfolgreich sind. Das ist das Wesen der Klasse von maschinellen Lernalgorithmen bekannt als „Beaufsichtigtes Lernen“ „Das Ziel Beaufsichtigtes Lernen ist die Verwendung von Trainingsdaten, um die Funktion so gut zu bestimmen, dass wenn ein neuer Datensatz angegeben wird, der Ausgang kann genau vorhergesagt werden. Im Gegensatz dazu ununterbrochenes Lernen ist, versteckte Muster oder zugrunde liegende Struktur in einer gegebenen

Datensatz, um über die Daten zu erfahren, die nicht gut mit der Aufgabe, die ich habe ausgewählt für meinen Laptop Quarterback im Training.

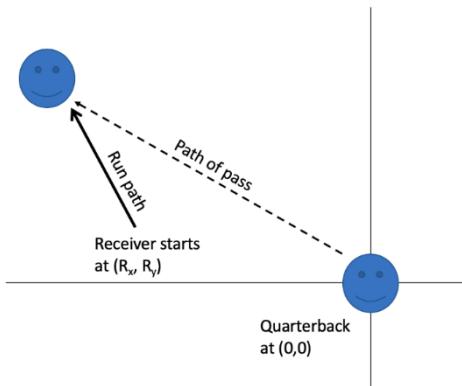
Um den Lernprozess zu beginnen, brauche ich eine Reihe von Trainingsdaten, aber wenn ich wirklich nur verwendet zufällige Passdaten, Ich könnte eine übermäßige Menge von Daten für unvollständig generieren mit sehr wenigen positiven Beispielen von kompletten Pässen. Diese Daten würden wenig Ausbildungswert. Ich muss also mit einem Trainingsdatensatz aus einer entsprechende Anzahl erfolgreicher Pässe. Beachten Sie, dass ich hier hedging, indem ich ein „ANHANG „ wie wir noch nicht wissen, was das sein könnte — die Größe der Ausbildung set ist wahrscheinlich abhängig von der Komplexität des Problems und dem maschinellen Lernen Algorithmus, den wir trainieren wollen.

Wie bekomme ich diese Trainingsdaten? In einer idealen Welt konnte ich mich vervollständigen Daten von einer großen Anzahl von Fußballspielen, aber dies ist weit außerhalb des Geltungsbereichs eines einfachen Blog-Posts. Darüber hinaus eine meiner Modellierungsannahmen, die ich schnucke in den ersten Absatz dieser Buchung war die „ lineare Bewegung eines Empfänger§ In der Realität Der Präsident. - Die Aussprache ist geschlossen. geplante Routen oder ihre Routen basierend auf Echtzeitbedingungen während des Spiels ändern.

Seit Ich habe keine leicht zugängliche Quelle von tatsächlichen Daten, die mit meinem Modellierung Annahmen, Ich werde einige abgeschlossene Passdaten für mit. Meine erste Darm Instinkt war nur zu beginnen, die Mathematik so zu arbeiten, dass Ich könnte eine Reihe von Funktionen für die Berechnung der idealen Pass von einem festen Wurf erzeugen einem Empfänger auf einem beliebigen linearen Weg mit einer konstanten Geschwindigkeit innerhalb eines beliebigen Satzes Zwänge für die maximale Wurfgeschwindigkeit (z.B. kein solches als schneller als Licht).

Aber das fühlte sich ein bisschen wie Betrug, weil ich beide mathematisch liefern würde perfekte Trainingsdaten und ich ignorierte auch das leistungsfähige Werkzeug, das ich in meiner Fingerspitzen: die [KNIME Analytics Plattform](#) „Was, wenn ich KNIME nutzen könnte, um einige zu erzeugen „schmutzig“ abgeschlossene Pass-Trainingsdaten? Habe ich mich nur mit einem Huhn oder dem Eiproblem? Wie kann ich meinen Laptop lehren, um einen Fußball zu werfen, wenn ich es zuerst lehren muss einen Fußball werfen, um Trainingsdaten zu generieren, damit ich es lehren kann, wie man einen Fußball...? Nun, es kann getan werden, und ich werde Sie durch die Wie und warum gehen — wenn Sie würden wie folgt direkt in meinem KNIME Workflow können Sie es von der [KNIME](#) Hubraum „

Aber zuerst lassen Sie uns auf die Aufgabe zurück, die ich meinen Laptop lehren möchte, wie zu erfüllen. Vielleicht Wir definieren es strenger, wir werden einen Ausweg aus dem Konundrum finden. Also, ich habe skizziert das Problem wie folgt heraus:



Eine Skizze, die den Wurf eines Fußballs zeigt.

Beim Zeichnen der Skizze erkannte ich, dass es irgendwelche willkürlichen Entscheidungen gab, die ich treffen konnte, um vereinfacht das Modell. Vor allem würde der Quarterback in einer festen Position bleiben den Ursprung (0,0) des Koordinatensystems. Der Empfänger würde bei einer beliebigen Position R_x , R_y und weg vom Quarterback in einer zufälligen Richtung (aber begrenzt und mit konstanter Geschwindigkeit. Dies ist, was ich bedenke, Empfänger , Laufparameter „

Um sicherzustellen, dass der Pass abgeschlossen werden kann, werde ich auch das Maximum des Empfängers begrenzen Geschwindigkeit deutlich geringer als die maximale Geschwindigkeit der Wurfkugel. Ich bin auch vorausgesetzt, dass der Quarterback den Ball sofort wirft (z.B. zum Zeitpunkt 0 im Modell. Schließlich, um das Modell ein wenig weniger Präzision erfordern, nehme ich an, dass jeder Wurf, der innerhalb einer Einheit des Empfängers gelangt, kann erwischt werden (z.B. gibt es muss kein mathematisch perfekter Schnittpunkt des Pfades von Empfänger und Kugel sein und die genaue Zeit der Landung des Balles.)

Es gibt auch eine andere implizite Parameter Lurking im Modell: Schwerkraft. Da der Weg der Empfänger und die Trajektorie der Kugel sind in Bezug auf ihre Mathematik, entschied ich, die Dinge ein wenig interessanter zu machen, indem sie die Wert der Schwerkraft für jedes laufende Szenario geändert werden. So ist die Schwerkraft ein weiterer Anfang Parameter, der zusammen mit dem "Laufparameter „ als Satz des ursprünglichen Modells Parameter:

Manually created table - 0:1 - Table Creator (Set Variables)					
File	Hilite	Navigation	View	Table "default" – Rows: 1 Spec – Columns: 5 Properties Flow Variables	
Row ID	Gravity	Receiver Start X	Receiver Start Y	Receiver Direction	Receiver Velocity
Row0	2	6.5	4	75	6

Diese Tabelle zeigt die „Laufparameter“.

Basierend auf meiner ersten Skizze wurde klar, dass es nur ein paar „Dinge“ die Quarterback könnte tun, um den Wurf zu beeinflussen, den sie machen. Sie können wählen Sie die

Richtung des Wurfs, des Wurfwinkels und der Wurfgeschwindigkeit. Ich bin
in Anbetracht dieser “ die Parameter ”,

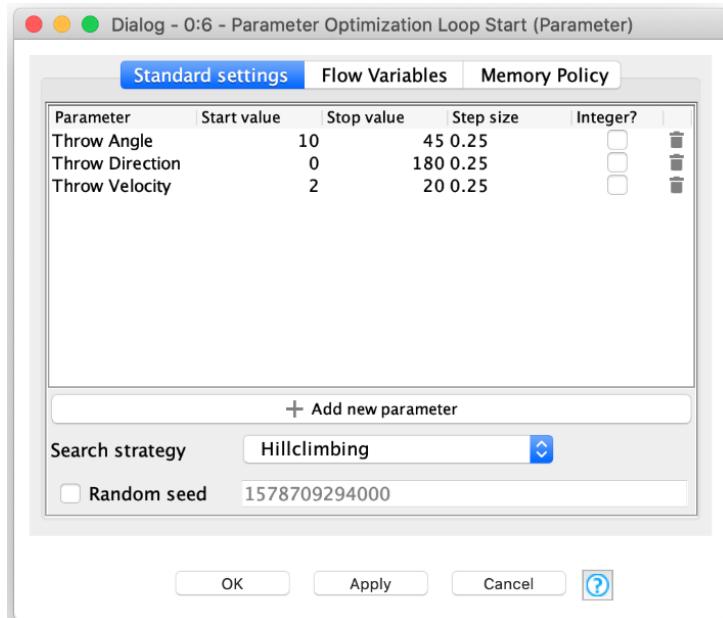
Das Problem wird nun: Für einen beliebigen Satz von Laufparametern, was sind die optimalen
Wurfparameter? Um diese Frage zu beantworten, müssen wir definieren, was wir mit
optimal. Natürlich ist ein unvollständiger Pass nicht optimal, aber da wir mehrere haben
wir können die Parameter variieren, wir können jede Anzahl von abgeschlossenen Pässen machen. Seit
das Ziel des amerikanischen Fußballs ist, den Ball so weit wie möglich nach unten zu bringen,
wir werden in diesem Fall optimal definieren, da der Pass, der am weitesten unter dem Feld gefangen wird
aus dem Quarterback.

Wie ich bereits erwähnt habe, könnten wir nur zufällig Werte für diese Parameter auswählen, bis
wir finden ein Set, das einen Pass abschließt. Dies könnte eine lange Zeit dauern, und es würde nicht
garantieren, dass wir den Satz von Parametern gefunden haben, die den optimalen Pass (basierend
über unsere Definition von optimal). Aber vielleicht können wir einfach genug zufällig abholen.
Sätze von Parametern, bis wir eine große Anzahl von abgeschlossenen Pässen mit den Hoffnungen gefunden
dass einer von ihnen sich dem optimalen Pass nähern wird? Dies ist BFI — Brute Force
und Ignoranz. Wenn wir eine unendliche Anzahl von Pässen machen, muss einer von ihnen sein
Nah genug zum optimalen Pass....

Aber das ist nicht praktisch; vor allem, weil wir eine breite Palette von
Trainingsdaten basierend auf vielen verschiedenen Parameterwerten des Empfängers. Was, wenn wir
können die zufällige Parameterwertauswahl führen? Schließlich, wenn wir einen Wurf in die
echte Welt, wir können sagen, wie Nähe Wir kamen, um unser Ziel zu treffen. Irgendeine Miss wäre
als Fehler betrachtet, und wir können sogar die Größe des Fehlers auf der Grundlage
wie weit weg der Ball vom Empfänger landete. Eine Person könnte und sollte von
den Fehler und den nächsten Versuch entsprechend anpassen.

Nun, KNIME kann das gleiche tun, dank des Paares der Unschuld, aber extrem mächtig
Knoten: die [Parameteroptimierung Loop](#) Knoten. Diese nifty-Knoten ermöglichen es Ihnen,
ein oder mehrere Parameter als Flussgrößen und deren Wertebereich für jede
von ihnen. Die Knoten schleifen dann über eine Reihe von geschachtelten Knoten bis zum optimalen Satz von
Parameterwerte werden gefunden. Klingt vertraut, oder? Hier ist die [Parameteroptimierung](#)
Loop Start für unsere Datenerzeugungsaufgabe konfigurierter Knoten:

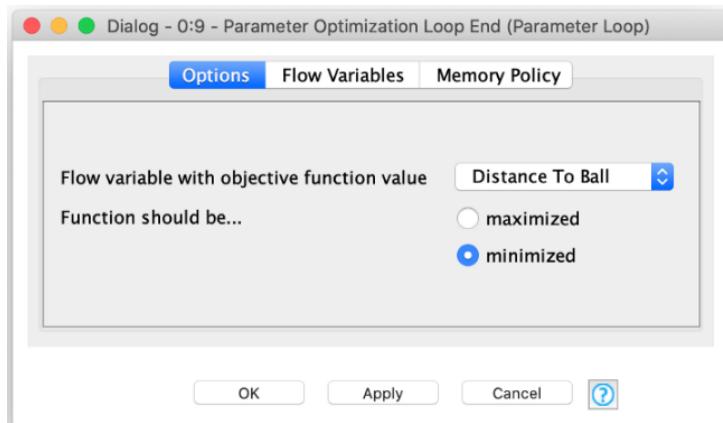
Der Pass, Teil 1: Parameteroptimierung mit KNIME



Der Konfigurationsdialog des Parameteroptimierungs-Loop Start-Knotens.

Sie können die drei Wurfparameter sehen. Ich habe früher zusammen mit den Zwängen diskutiert auf ihren Werten, um innerhalb des Modells I gebaut. Aber Sie können auch bemerken, dass es eine konfigurierbare Suchstrategie – so sagen Sie dem Knoten den besten Weg zu finden die optimalen Parametersätze und wir decken die verschiedenen Strategien in Kürze ab.

Wie bereits erwähnt, ist das Ziel, die Parameterwerte zu optimieren, um die Fehler. In unserem Fall stelle ich den Fehler als bloße Distanz zwischen dem Empfänger fest und der Ball am Ende des Wurfs und es ist wirklich einfach, KNIME zu sagen, dass dies zu optimieren in der [Parameteroptimierung Loop End Node](#):



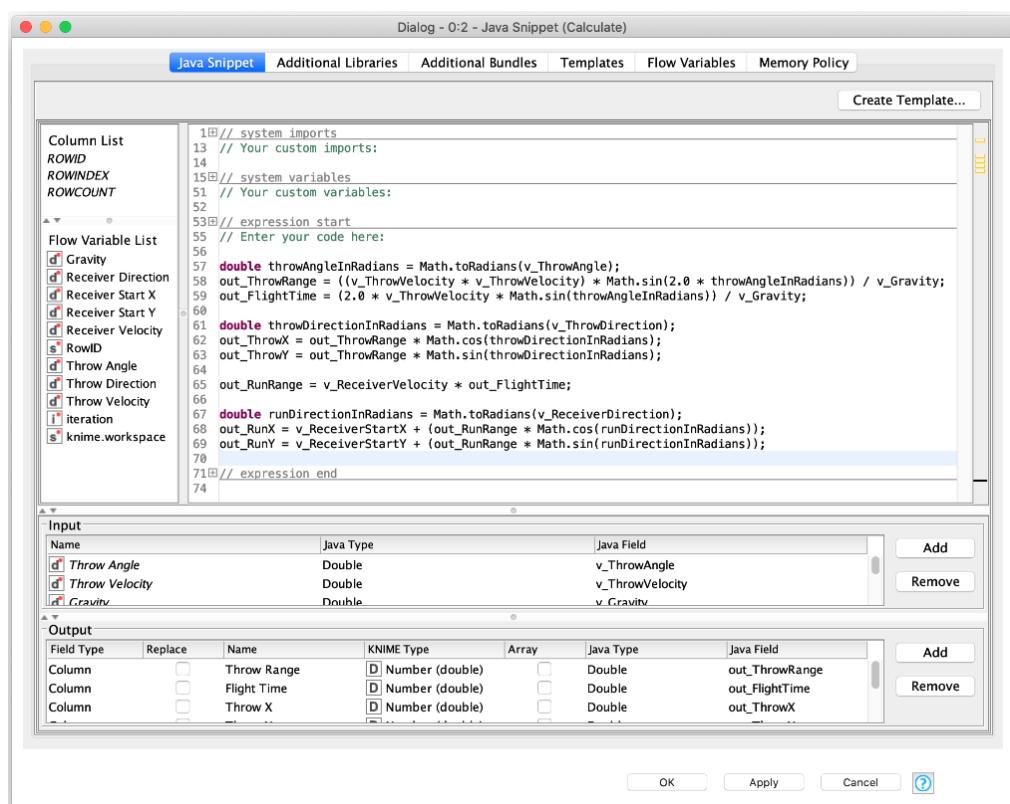
Der Konfigurationsdialog des Parameteroptimierungs-Loop-End-Knotens.

Es gibt zwei Ausgänge vom Schleifenendknoten. Der erste ist der optimale Parametersatz Werte (z.B. die mit dem minimalen Fehlerwert). Die zweite ist das Set aller getesteten

Der Pass, Teil 1: Parameteroptimierung mit KNIME

Parameterwertkombinationen und deren entsprechende Fehlerwerte. Das ist eine wichtige Unterscheidung für unser Szenario zu machen; denken Sie daran, dass wir nicht nur suchen für den genauesten Wurf, aber für den Wurf, der innerhalb einer Einheit des Empfängers ankommt (z.B. Fehler kleiner als 1,0) und ergibt den Empfänger so weit wie möglich unter das Feld. (z.B. die maximale Y-Position für den Empfänger).

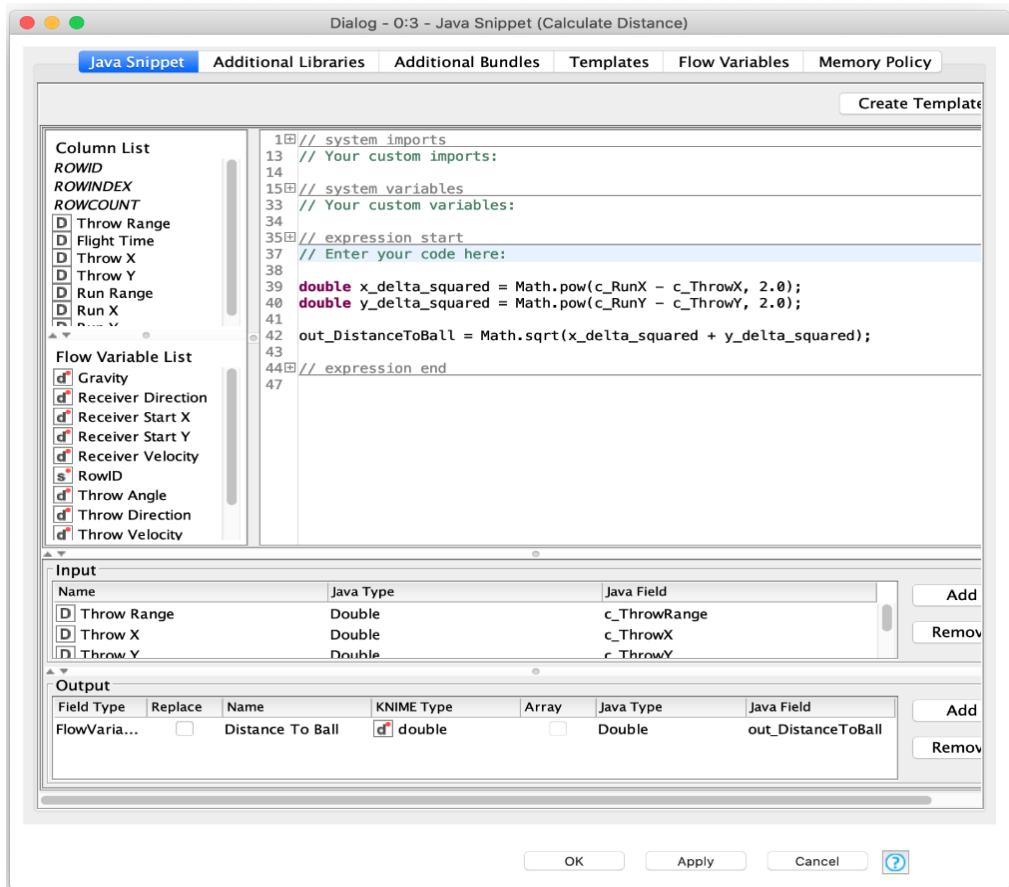
Bevor ich gehe, wie wir diesen optimalen Wurf aus dem Ausgang der Schleife wählen, lassen Sie uns berühren auf was innerhalb der Schleife passiert. Das erste, was die Schleife tut, ist die Berechnung der Merkmale des Werfens basieren auf dem aktuellen Satz von Wurfparametern. Dies schließt ein nicht nur die Koordinaten, wo der Wurf landet, sondern auch seine Zeit im Flug. Die Zeit in Flug wird dann verwendet, um die Position des Empfängers am Ende des Werfens zu berechnen. Das alles wird in einem Java-Snippet-Knoten ausgeführt:



Mit diesem Java-Code-Snippet werden die Eigenschaften des Wurfs berechnet und die Position des Empfängers am Ende der Wurf bestimmt wird.

Jetzt, wo wir wissen, wo der Ball landet und wo der Empfänger ist, wenn der Ball lands, es ist einfach, den Abstand zwischen diesen beiden Punkten zu berechnen, um als unser Fehler zu verwenden Maßnahme:

Der Pass, Teil 1: Parameteroptimierung mit KNIME



Mit diesem Java-Code Schnipsel, der Abstand zwischen dem, wo der Ball landet und wo der Empfänger steht, ist berechnet und als Fehlermaßnahme verwendet.

Das ist alles, was wir brauchen, um jeden Wurf für einen bestimmten Satz von Parametern zu modellieren und das KNIME zu lassen Optimierungsknoten machen ihre Magie. Aber ich habe früher erwähnt, dass die Knoten eine wählbare Suchstrategie zur Suche nach einem optimalen Satz von Parameterwerten.

Die erste Strategie ist „Brute Kraft“ die jede mögliche Kombination aus Parameterwerte mit den in der Schleife konfigurierten Zwängen und Schrittgrößen Startknoten. Wenn wir wollen, dass unser Datengenerator in jeder angemessenen Zeit endet Dies ist kein praktischer Ansatz für unser Modell, da es in vielen Millionen von Iterationen der Schleife für jeden Trainingsdatensatz.

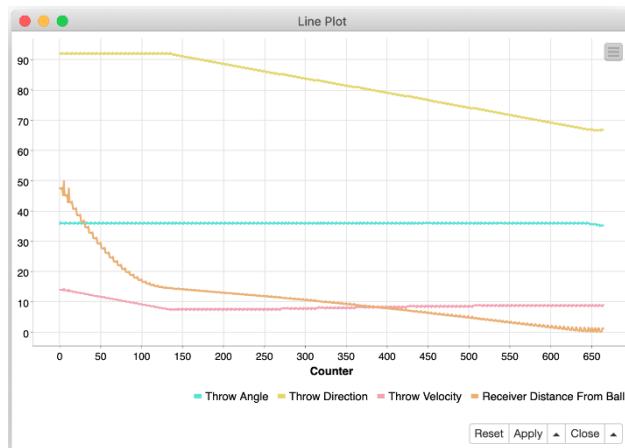
Der nächste unpraktische Ansatz für uns ist der „Die Welt der Welt“ Strategie, wo die Schleife Knoten versuchen zufällige Werte für jeden Parameter in der Hoffnung, dass einer der Sätze wird zu einem minimalen Fehler führen. Sie können dem Knoten sagen, wie viele zufällige Wertesätze zu versuchen und sogar sagen, es vorzeitig zu stoppen, wenn es eine Menge von Werten findet, die gut genug basiert auf einer konfigurierbaren Toleranz. Auch mit unserer großen Menge an potenziellen Werten und Wunsch, finden etwas in der Nähe der optimalen Wurf in Bezug auf Distanz nach unten Feld, wir brauchen eine der Ansatz, der wahrscheinlicher ist, uns die besten möglichen Sätze von Parametern zu geben.

Der Pass, Teil 1: Parameteroptimierung mit KNIME

Hier kommen die übrigen beiden Suchstrategien ins Spiel. Das erste ist „Hillclimbing“ wo die Schleife eine zufällige Startmenge von Parametern und dann wertet den Satz aller Nachbarn aus (z.B. mit den Schrittgrößen, um jeden zu bewegen Parameterwert in beiden Richtungen) zu sehen, welcher Nachbar am meisten in Richtung Minimierung des Fehlers. Dieser Vorgang wiederholt eine konfigurierbare Anzahl von Zeiten oder bis keine Nachbarn führen zu einer Verbesserung der Fehlerquote.

Die letzte Suchstrategie ist „Bayesische Optimierung“ wo mehrere anfängliche Zufall Parameterkombinationen werden berücksichtigt und dann die Bayesische Optimierungsstrategie versucht, die Fehlerwerteigenschaften dieser Kombinationen zu verwenden, um die nächste runde Parameterwertkombinationen zu versuchen. Dies setzt auch für ein konfigurierbares Anzahl der Iterationen.

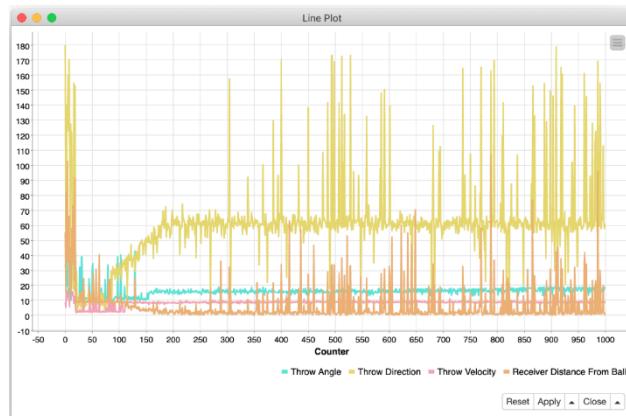
Bei der Entscheidung, zwischen welchen der beiden anspruchsvollereren Suchstrategien zu verwenden, Sie müssen die Art Ihrer Fehlerfunktion kennen: wie glatt und schnell es konvergiert und wenn es lokale Minima anstelle nur eines globalen Minimums haben kann. Glücklicherweise können Sie die Kraft von KNIME nutzen, um die Strategien zu testen und zu visualisieren ihre Operation. Zum Beispiel, wenn wir die „Hillclimbing“ Strategie können Sie die Schleifenstart mit einem relativ großen Fehler von fast 50 Einheiten aus der Kugel und es zunächst Stellt die Geschwindigkeit und beginnt dann bald die Wurfrichtung, bevor es endlich den Winkel des Werfens:



Ein Linendiagramm, das die Parameter der Optimierungsschleife zeigt, wenn mit der Hillclimbing-Strategie.

Wenn wir zum „wechseln Bayesische Optimierung“ Strategie, sehen Sie eine viel mehr chaotisch Suche, da es von mehreren zufälligen Punkten beginnt und eine breitere Variation der Parameterkombinationen:

Der Pass, Teil 1: Parameteroptimierung mit KNIME

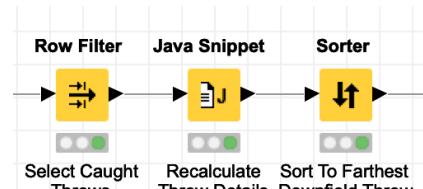


Ein Liniendiagramm, das die Parameter der Optimierungsschleife zeigt, wenn mit der Bayesischen Optimierungsstrategie.

Wie bereits erwähnt, da unsere Fehlerfunktion für den Empfänger nicht optimiert distanzieren Sie das Feld, wir sind nicht nur auf der Suche nach der niedrigsten Fehlerquote, sondern eine akzeptable Fehlerquote (z.B. weniger als 1,0 für einen Fang) mit der größten Y-Distanz vom Empfänger gereist. Auf der Grundlage dieser Anforderungen und des obigen Diagramms der Suchstrategie ist klar, dass die Strategie „Bayesische Optimierung“ viele mehr Potenzial abgeschlossen Wurf Kandidaten auf die kleinen Kosten ein paar hundert zusätzliche das zu tun.

Wir nehmen dann alle versuchten Parametersätze von der Optimierungsschleife versucht und filtert sie bis nur die, die als gefangen (z.B. Fehlerwert von weniger als 1,0). Seit die Optimierungsschleife nur den Parameter ausgibt Werte und ihre damit verbundenen Fehlerergebnisse, wir brauchen die Koordinaten des Fangs neu zu berechnen und dann sortieren wir alle fertigen Werfen basiert auf dem Y-Parameter des Fangs, so dass der erste Rekord enthält unseren optimalen Wurf.

Diese Knoten geben uns eine Tabelle der fertigen Werfen sortiert, um von weitesten nach unten:



Ein Schnipsel der Optimierung
Arbeitsablauf. Mit diesen Knoten die optimale
Werft wird ausgewählt.

Der Pass, Teil 1: Parameteroptimierung mit KNIME

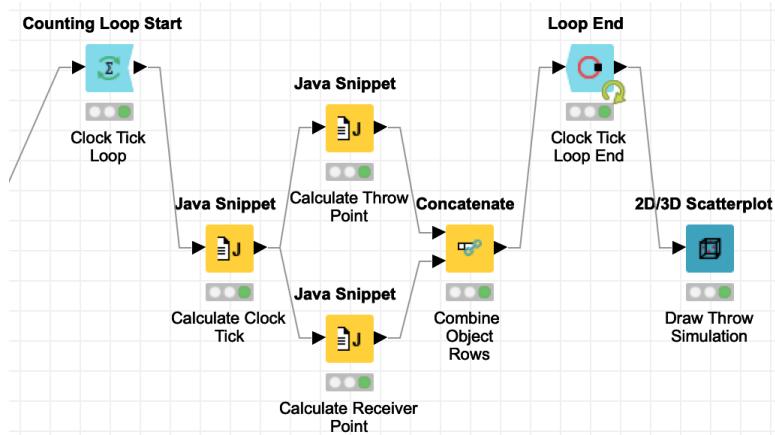
Sorted Table - 0:28 - Sorter (Sort To Farthest)

File	Hilite	Navigation	View	Table "default" – Rows: 216		Spec – Columns: 8	Properties	Flow Variables
Row ID	D Throw Angle	D Throw Direction	D Throw Velocity	D Throw Y	D 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			

Die Ausgabetafel des Sortierknotens. Die Tabelle enthält alle fertiggestellten Wurfs sortiert nach nach unten.

Eine interessante Sache zu beachten ist, dass, weil wir die „Bayesische Optimierung“ Suchstrategie, die besten Ergebnisse wurden bei verschiedenen Iterationen erhalten, wie man sehen kann aus der Zeilen-ID, die während der Parameteroptimierungsschleife erzeugt wurden bevor die Tabelle auf der Throw Y Spalte sortiert wurde.

Schließlich, um die Ergebnisse zu visualisieren und zu überprüfen, ob die ausgewählten Wurfparameter sind tatsächlich genau Ich habe eine zweite Schleife, die die Trajektorie der Kugel simuliert und die Position des Empfängers für 100 „Uhrenbewegungen“ gleichmäßig über den Flug verteilt Zeit des Balles:

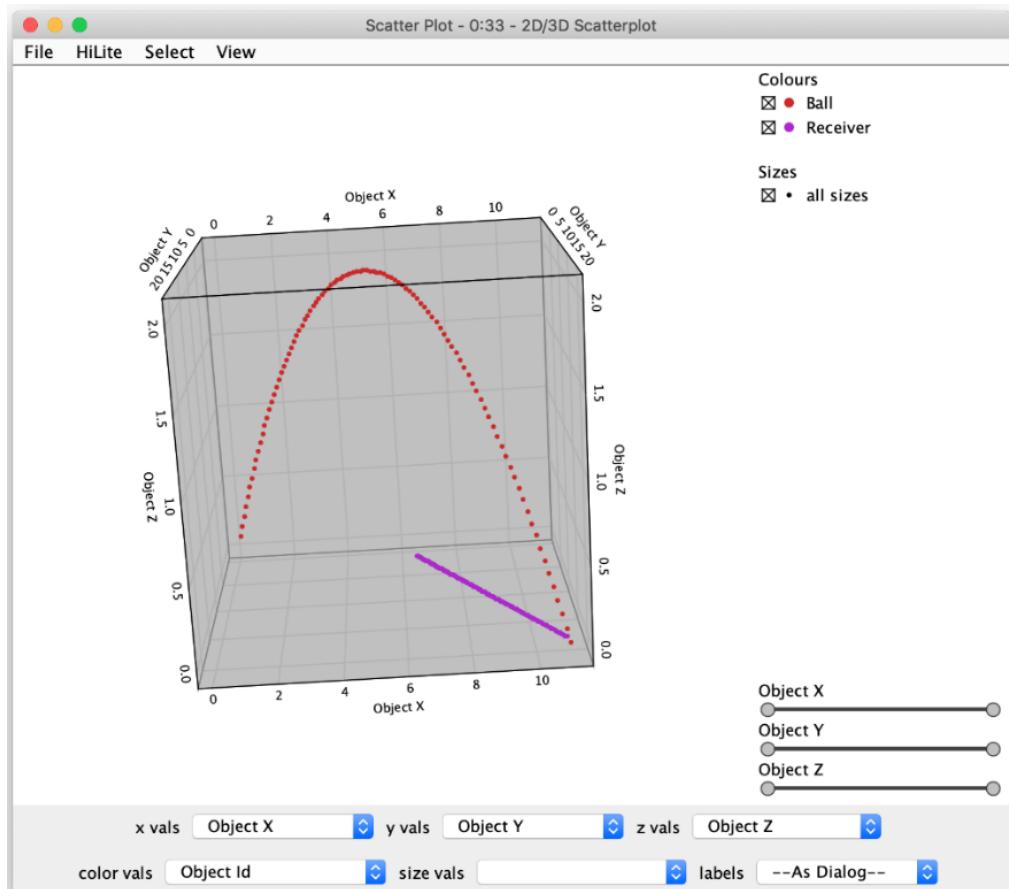


Ein Schnipsel der [Parameteroptimierung Workflow](#). Mit diesen Knoten die optimale Es wird eine Wurfsimulation erzeugt.

Der Pass, Teil 1: Parameteroptimierung mit KNIME

Durch die Verwendung der mächtigen [2D/3D Scatterplot](#) Knoten, der als KNIME frei verfügbar ist

Community-Knoten, können wir dann die Simulationsergebnisse in einem interaktiven Diagramm anzeigen Visualisierung:



Ein 3D-Streudiagramm, das die Simulationsergebnisse in einer interaktiven Graphik-Visualisierung zeigt.

Im obigen Screenshot befindet sich der Quarterback an Position (0, 0) in der linken Rückseite den Würfel. Der Weg des Empfängers wird durch die lila Punkte dargestellt, die auf die Kamera und glühen vom Quarterback weg. Die Trajektorie des Quarterback wird durch die roten Punkte gezeigt und enthält sogar die Trajektorie, um die Beziehung zwischen Pass und Empfänger über die Zeit.

Jetzt, da wir überprüft haben, dass bei einer einzigen Reihe von Startempfänger „Laufparameter“, wir können einen Nah-optimalen Wurf für die Weitergabe an den Empfänger produzieren, wir müssen jetzt bauen eine Umschlingungsschleife zu einer beliebigen Anzahl von Datensätzen mit anfänglichen "run-Parametern" und Berechnung ihrer Wurfparameter mit der Technik, die wir in diesem Blog-Post entwickelt haben. Dies wird unser Trainingsdatensatz sein, um meinen Laptop zu lehren, wie man den Ball geworfen, und mein nächster Blog-Post wird hier abholen und die Lehre von verschiedenen Maschinen abdecken Lernalgorithmen und beinhalten einen Wettbewerb, um zu sehen, welche „lernt das Beste“ für unsere simplistischen Szenario.

Dieser Artikel wurde ursprünglich veröffentlicht [BigBear.AI](#) und wir haben es in unserem [Niedriger Code für Advanced Data Science Amtsblatt](#) auf Medium [Hier](#).

Die entsprechenden [Den Pass machen, Teil 1](#) Arbeitsablauf auf der KNIME Community Hub in Pauls öffentlichem Raum.

Sie können weiter lesen Teil 2 auf Medium bei [Der Pass, Teil 2: Ausbildung einer Neural Netzwerk mit KNIME](#) oder [BigBear.AI](#).

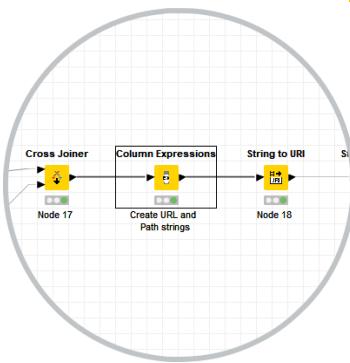


[John Emerging](#) wurde nominiert Beitrag des Monat für Juli 2022. Er wurde für seine unzähligen [Blog Beiträge zum Thema Parsing und Analyse von PDF-Dokumenten](#) . Laufen [baseball hitting streak Simulationen](#) , oder [Rißbildung Wort](#) . Darüber hinaus ist John auch ein angesehener Redner bei Datenwissenschaft Veranstaltungen, und er ist der erste zertifizierte Trainer KN-Code Software (Check out his [Abzeichen](#)) Er ist derzeit Teil der Organisationstafel der [Daten verbinden Nord](#)

[Amerika](#) [Veranstaltungsreihe](#). Das Bild auf der rechten zeigt einen Workflow-Schnippen seiner PDF Parsing Workflow.

John ist ein Principal Consultant ad phData, wo er sich auf die Verwendung von Analytik konzentriert Engineering-Tools wie KNIME helfen Kunden, ihre Daten effektiver zu nutzen. Er Erfahrungen als Operation Research Analyst an der US Army und als Berater in der Finanzierung der Industrie in der Vergangenheit. Er ist besonders Interesse an geospatialer Analyse und außerhalb der Arbeit er gerne Filmfotografie, Wandern und Bergsteigen.

Besuchen Sie John's [Raum auf dem KNIME Hub](#) oder [Profil Seite im KNIME Forum](#) (Hub/Forum Griff: Johnvemery)



Mit KNIME zu Parse und Analyze PDF Dokumente

Tageswetteraufnahmen mit wenigen Klicks extrahieren

Autor: John Emerging

In meinem beruflichen Leben arbeite ich mit Daten; Vorbereitung, Reinigung, Baamodelle und Erstellung von Visualisierungen. Ich verbringe die große Mehrheit meiner Zeit mit Werkzeugen wie Tableau, Power BI und Google Data Studio für Visualisierungsarbeiten, während ich ETL-Workflows in Werkzeuge wie Alteryx, KNIME und Tableau Prep. Außerhalb der Arbeit habe ich ein Ziel am höchsten Punkt aller 50 Staaten stehen. Bis heute bin ich auf 24. Im Oktober Am 31. dieses Jahres flog ich nach New Hampshire, um Mt. Washington zu erreichen. der höchste Punkt des Staates — und mein würde-sein halber Punkt.

WS FORM F-6												STATION MOUNT WASHINGTON OBSERVATORY					
PRELIMINARY LOCAL CLIMATOLOGICAL DATA												MONTH OCTOBER		YEAR 2021			
LATITUDE 44 DEGREES 16 MINUTES NORTH				LONGITUDE 71 DEGREES 18 MINUTES WEST				GROUND ELEVATION (H) 6280 FT				STANDARD TIME EASTERN					
DAY	MAX	MIN	AVG	NORM	DEPART	HEAT	COOL	DEGREE DAYS	TOTAL (EQUIV)	SNOW/ICE ON GROUND-TAM	AVG SPEED	WIND (MPH)	SUNSHINE (MINUTES)	SKY COVER (%)	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	1	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)				PRECIPITATION DATA (IN.)				WEATHER				SYMBOLS USED IN COLUMN 16					
AVERAGE MONTHLY DEPARTURE FROM NORMAL	38.4	TOTAL FOR THE MONTH	12.60	NUMBER OF DAYS:				1 = FOG									
HIGHEST	7.1	DEPARTURE FROM NORMAL	2.61					2 = FOG REDUCING VISIBILITY									
LOWEST	60 on 12th	24 HOUR MAX	4.15 on 30th/31st	CLEAR (SCALE 0-3)	3			TO 1/4 MILE OR LESS									
NUMBER OF DAYS WITH:	14 on 24th	SNOWFALL, ICE PELLETS (IN.)		PARTLY CLOUDY (SCALE 4-7)	3			3 = THUNDER									
		TOTAL FOR THE MONTH	10.7	CLOUDY (SCALE 8-10)	25			4 = ICE PELLETS									
								5 = HAIL									
								6 = GLAZE OR RIME									

Monatliche veröffentlichte Daten des Mt. Washington Observatory.

Mt. Washington ist weltbekannt für sein schreckliches Wetter. Windbögen im Überschuss von 230 Meilen pro Stunde und etwa ein Drittel der Tage des Jahres

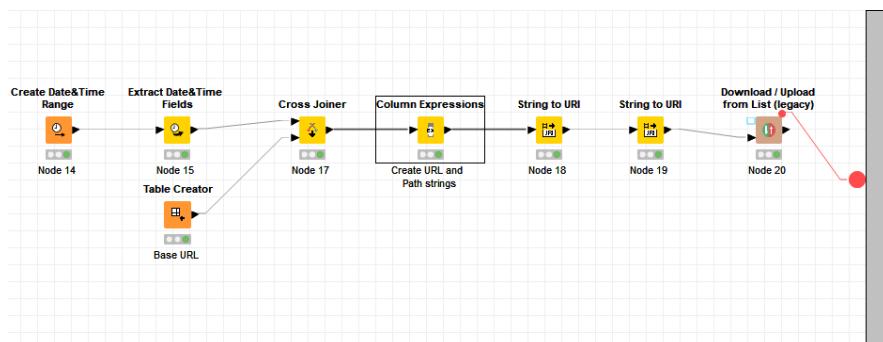
Winde über 100 Meilen pro Stunde erleben. Obwohl viele Besucher fahren oder nehmen Bahn nach oben in den milderer Sommermonaten, das Wetter auf dem Berg kann extrem gefährlich sein — vor allem in den Winter. Natürlich macht das die Idee, es viel attraktiver für mich zu klettern.

Planung einer Wanderung/Klima Endes Oktober bedeutet, Sie müssen das Potenzial für früh navigieren Saison Winterwetter. Als Data Professional war ich daran interessiert zu sehen, was Oktober Das Wetter sah in der Regel nach einer Vorstellung von meiner Wahrscheinlichkeit, den Gipfel zu erreichen. Zum Glück der Mt. Washington Observatorium hält ein Protokoll der täglichen Wetteraufzeichnungen — hohe und niedrige Temperaturen, mittlere und schnellste Winde, Niederschlag usw. — für jede Monat zurück zu 2005.

Erhalten der Daten

Was die Analyse dieser Daten schwierig macht, ist, dass die Beobachtungen als PDF gespeichert. Wie viele Analysten, Datenwissenschaftler und andere Datenexperten können bezeugen, mit Daten in PDFs können eine Herausforderung sein.

Um die Daten zu sammeln, verwendete ich KNIME um jeden Monat PDF auf meinen Computer zu speichern. (Als ein beiseite, KNIME ist ein wunderbar kostenlos und Open-Source-Datenvorbereitung und Data Science-Tool. wenn Sie wollen eine No-oder-Low-Code-Lösung für Ihre Daten Prep Bedürfnisse, es ist eine ausgezeichnete Wahl.)



Dieser Workflow-Snippet erstellt Date&Time-Bereiche, um benutzerdefinierte URLs und Pfadfolgen zu erstellen.

Die URL für die Wetterbeobachtungen von Mt. Washington lautet:

<https://www.mountainwashington.org/uploads/forms/2021/10.pdf>

Das einzige, was sich von einem Monat zum nächsten ändert, ist das Jahr (2021, 2020, 2019...) und den Monat (10, 09, 08...). Verwendung von Werkzeugen wie Datum und Uhrzeitbereich erstellen und Tabelle Schöpfer Ich konnte einfach alle möglichen Kombinationen von Monaten und Jahren zu erzeugen erstellen Sie alle URLs, die bis Januar 2005 zurückgehen. Von dort aus konnte ich die PDFs in einen Ordner auf meinem Computer mit zwei Streichen an URL Knoten. Der erste Knoten erstellt ein URI-String aus der Quelle und der zweite erzeugt URI-String zum Ziel Ziel auf meinem Computer. Wie ich das im November 2021 schreibe, der Download-Prozess dauert etwa eine Minute für über 200 PDFs.

Voreinstellung der Daten

Das Herunterladen der PDFs auf ein Laufwerk auf meinem Computer war relativ einfach. Alles, was ich hatte zu tun wurde Strings für jede URL generieren und dann herunterladen. Daten von ein parsed PDF kann jedoch ein absoluter Albtraum sein. Je nach Struktur das PDF, das Sie parse müssen, diese Aufgabe kann von ganz einfach bis fast unmöglich.

Zum Glück bietet KNIME einen Knoten namens [Tika Parser](#) . Wie KNIME es beschreibt, dieser Knoten „ermöglicht die Parsierung von Dokumenten, die von Tika unterstützt werden.“ Die Tika Parser node ist lächerlich einfach zu konfigurieren. Ich habe einfach das Verzeichnis ausgewählt, das die heruntergeladene PDFs, wählte den Dateityp aus einer Liste und welche Metadaten-Elemente I wollte ausgeben. In diesem Fall habe ich den Dateipfad und den Hauptinhalt ausgewählt. Das Ergebnis Inhaltsausgabe ist eine lange Textfolge aus jedem 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 I	
Row2	C:\Users\John Emery\Documents\DATA\Mt Washington\200503 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row3	C:\Users\John Emery\Documents\DATA\Mt Washington\200504 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row4	C:\Users\John Emery\Documents\DATA\Mt Washington\200505 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row5	C:\Users\John Emery\Documents\DATA\Mt Washington\200506 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row6	C:\Users\John Emery\Documents\DATA\Mt Washington\200507 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row7	C:\Users\John Emery\Documents\DATA\Mt Washington\200508 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row8	C:\Users\John Emery\Documents\DATA\Mt Washington\200509 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row9	C:\Users\John Emery\Documents\DATA\Mt Washington\200510 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row10	C:\Users\John Emery\Documents\DATA\Mt Washington\200511 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row11	C:\Users\John Emery\Documents\DATA\Mt Washington\200512 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
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 I	
Row14	C:\Users\John Emery\Documents\DATA\Mt Washington\200603 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row15	C:\Users\John Emery\Documents\DATA\Mt Washington\200604 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	
Row16	C:\Users\John Emery\Documents\DATA\Mt Washington\200605 Weather.pdf	_ WS FORM F-6 STATION_ MOUNT WASHINGTON OBSERVATORY__PRELIMINARY LOCAL CLIMATOLOGICAL I	

Rohdaten des Tika Parser... nicht die schönste Sache — noch!

Wenn die Daten in ihrer rohen String-Form so sind, können wir beginnen, sie auseinander zu spalten — Felder in verschiedenen Spalten und die Beobachtungen jedes Tages in verschiedenen Zeilen. Um zu beginnen, verwendete ich die Zellteilung Knoten, um die Inhaltsspalte in eine Liste basierend auf der neuen Zeile zu teilen. Abgrenzer (\n). Interessanterweise, in KNIME, wenn Sie einen String in verschiedene Zeilen teilen möchten, Sie gehen zuerst die Daten in eine Liste teilen und dann die Gruppe Knoten, um die Liste in Zeilen. Dieses Verfahren verwandelt unseren 203-reihigen Datensatz in einen 26,566-reihigen Datensatz.

Eine Sache zu beachten ist jedoch, dass die Tika Parser knoten parses die Gesamt PDF. Das enthält die von Ihnen gewünschten Daten und Daten, die Sie möglicherweise nicht wünschen. In meinem Fall war ich nur interessiert an den täglichen Wetterbeobachtungen; Daten zwischen ZAHL und “ 31.” Aufzeichnungen im Bild oben — alles andere musste gehen.

Die dynamische Ausfilterung von Stringdaten ist eine Herausforderung für viele Analytiker. Sie haben nicht der Luxus des Wortes “ X > 100 ” oder jede schöne mathematische Formel. Für diese Übung, ich Ich habe hier gekämpft. Ich wusste, dass ich regelmäßige Ausdrücke verwenden wollte, um die Zeichen zu extrahieren bis zum ersten Raum, aber KNIME hat keine eingebaute Funktion für den regulären Ausdruck Extraktionen. Zum Glück gibt es einen Drittanbieter-Knoten namens [Regex Extractor](#) genau das, was ich brauchte!

Verwendung von Regex Extractor Knoten, ich konnte die erste herausziehen “ Wort ” aus jedem String und dann filtern, wenn es entweder sagte “ ZAHL ” oder war eine Zahl zwischen 1 und 31.

Datenwissenschaft Use Cases – John Emerging
Mit KNIME zu Parse und Analyze PDF Dokumente

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

Wir bekommen dort...

Mit nur noch numerischen Daten spalte ich die Daten schließlich in verschiedene Spalten auf Basis der Raumbegrenzer. Von hier aus standen mir nur Standard-Datenschutzprobleme — Umbenennen von Spalten, die sicherstellen, dass Felder entsprechende Datentypen erhalten wurden, usw.

Die Ergebnisse

Dieser Workflow nahm mich etwa eine Stunde, um von Anfang bis Ende zu bauen. Nur ein paar einfache und einfach zu konfigurierende Knoten, ich konnte URLs erstellen, PDFs herunterladen mein Computer, parse jedes PDF, und dann rekonstruieren Sie die Daten in eine einfach zu bedienende Daten Tisch.

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

Das ist viel besser.

Reine Daten zu haben, ist vielleicht das wichtigste Stück des Puzzles, um Sound durchzuführen Analyse. Fehlerhafte, unvollständige oder schlecht strukturierte Daten verursachen nicht nur Verzögerungen, sondern auch häufig zu einer ungenauen Berichterstattung führen. Es endet eine Zeitverschwendug für alle involviert. Mit einem Tool wie KNIME habe ich bestehende Daten in einer herausfordernden Struktur für Analyse und Formung in die Struktur, die ich brauchte.

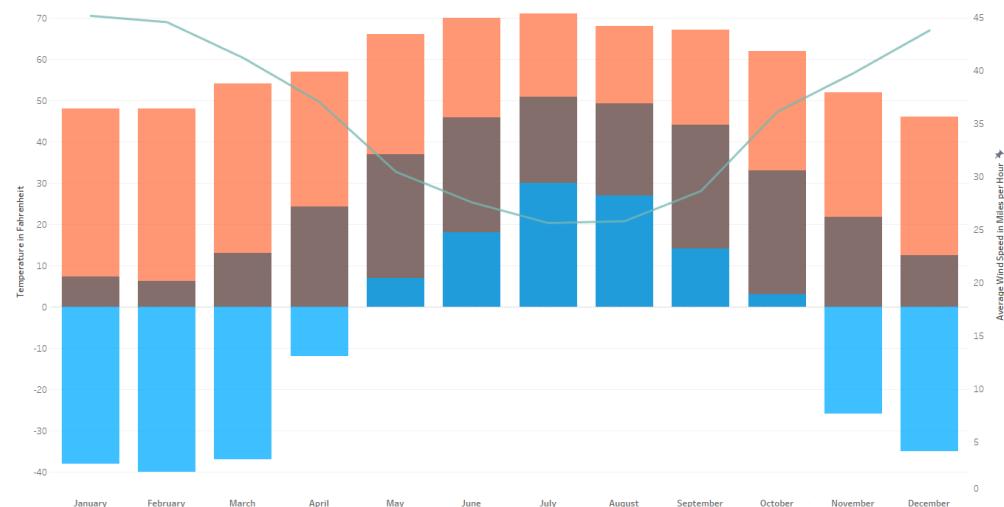
Was ist mit den Ergebnissen von Mt. Washington?

Monthly average wind speed and temperatures

Mount Washington, New Hampshire

bars: minimum / average / maximum

line: average wind speed



Monatliche durchschnittliche Windgeschwindigkeit und Temperaturen, gemessen bei Mt. Washington, New Hampshire.

Der Tag meines Aufstiegs sah fast 4 Zoll Regen (über 5 Zoll insgesamt fiel am Tag des Aufstiegs der Aufstieg und der Tag davor) und durchschnittliche Winde von etwa 50 Meilen pro Stunde und Böen 90. Blick über die historischen Daten, es war der 15. regnerische Tag und um die 80. prozentile für mittlere Windgeschwindigkeit. Kurz gesagt, kein toller Tag zum Klettern!

Dieser Artikel wurde ursprünglich veröffentlicht [LinkedIn Puls](#) und wir haben es in unserem [Niedriger Code für Advanced Datenwissenschaft Amtsblatt](#) auf Medium [Hier.](#).

Bildung und Forschung

In diesem Abschnitt erstellten wir alle Artikel, die einen Lehrzweck enthalten. Das beinhaltet akademische Beiträge, die sich auf Drug Discovery oder Gene Ontology konzentrieren, aber auch andere pädagogische Artikel wie Tutorials, wie bestimmte Konzepte sein können in der KNIME Analytics Platform implementiert. Die Kategorie „Bildung und Forschung“ unsere Lehrer, Wissenschaftler und Wissenschaftler:

- **Keith McCormick**
 - Anerkannt Analytics Leader & Independent Predictive Analytics und Maschine Lernberater
- **Wie ist das?**
 - Professor @Free University of Bozen-Bolzano
- **Alzbeta Tuerkova**
 - Leiter Computer-Aided Drug Design @Celeris Therapeutics
- **Malik Yousef**
 - Leiter des Galilee Centers für digitale Gesundheitsforschung @Zefat Academic Hochschule
- **Nick Rivera**
 - Business Analyst @EMR
- **Francisco Villarroel Ordens**
 - Professor für Marketing @LUISS Guido Carli University
- **Christophe Molina**
 - Freelance Data Analyst, CEO @PIKAÍROS



Keith McCormick wurde nominiert KNIME-Beitrag der Monat für Dezember 2020. Er wurde für seine Kurse auf Linked Im Lernen: Einführung in die Maschine Lernen mit KNIME, und Data Science Foundations - Datenbewertung für prädiktive Modellierung. Im ersten Kurs, Keith zeigt, wie KNIME alle Phasen unterstützt der CRISP-DM-Zyklus in einer Plattform – inklusive Fusion und Aggregation, Modellierung, Datenscoring und R und Python Integrationen. Der zweite Kurs konzentriert sich auf Prinzipien, Leitlinien und Werkzeuge wie KNIME und R für den Datenzugriff, so dass sie für die Maschine geeignet sind Lernen.

Keith ist ein unabhängiger Ausbilder, bei beiden University of California in Irvine (UCI)

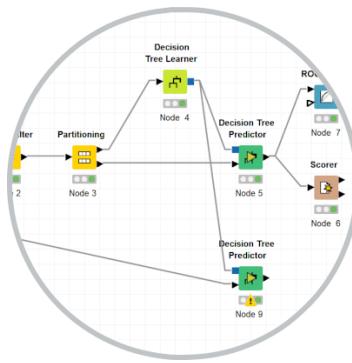
und auf LinkedIn Learning. Keith arbeitet nicht nur als Lehrer, er ist Lehrer im Herzen.

Er Genießen Herstellung sein Studenten für die professionelle Welt, provozieren diese Wirkung der Überraschung wenn er Erkenntnisse aus Daten enthüllt und seine Studenten. Ja, grading, da dies der Moment ist wo er mit der Arbeit der Studenten interagieren kann, nicht nur Kritik, sondern auch zu erklären und zu loben (siehe Keiths Interview im DCE Magazine)

Besuchen Sie Keith's Raum auf dem KNIME Hub oder Profil

Seite im KNIME Forum (Hub/Forum Griff:

Der Präsident)



Wo läuft Data Science Education? Lassen Sie die Experten sprechen

Mein Data Guest – Ein Interview mit Keith McCormick

Autor: Rosaria Silipo



Es war mir ein Vergnügen, vor kurzem Interview live auf LinkedIn die Mein Data Guest Interview-Serie. Er teilte Einblicke in die Welt der Datenwissenschaft

Bildung, erklärte, wie er KNIME in seiner Lehre verwendet, und sprach über die Bedeutung der guten Mentorschft für junge Datenwissenschaftler.

Keith McCormick

als Teil des

Keith McCormick ist ein unabhängiger Datenwissenschaftler, Trainer, Konferenzsprecher und Autor. Seit über 25 Jahren hat er die Data Science Teams geleitet, um hochgradig zu etablieren effektive Analysemethoden in allen Branchen, einschließlich des öffentlichen Sektors, der Medien, Marketing, Gesundheitswesen, Einzelhandel, Finanzen, Produktion und Hochschulbildung.

Rosaria: Sie lehren seit vielen Jahren die Datenwissenschaft? Vielleicht können Sie geben Sie uns eine kleine Zusammenfassung Ihrer Reise als Data Science Lehrer.

Keith: Es ist schon eine Reise von 25 Jahren. In meinen 20er Jahren begann ich als ein unabhängiger Statistik-Software-Trainer, traditionelle Themen in Statistiken wie diskriminierende Analyse und logistische Regression mit SPSS. Es war ursprünglich ein Nebenmann während der Graduiertenschule, um die Rechnungen zu bezahlen, aber schließlich wurde eine Karriere. Die Als nächster wichtiger Meilenstein für mich war die Begegnung mit visuellen Analyse-Tools. In den späten 90s, SPSS kaufte ein viel kleineres Unternehmen, das eine der frühen Vorhersage entwickelt die Analyse arbeitet mit visuellen Datenflüssen zusammen. Es war der Anfang von no-code/low-Code-Tools in der Datenanalyse. Seit sich diese Werkzeuge im Wesentlichen entwickelt haben, erreichen ein sehr hohes Maß an Raffinesse. Ich war schon immer ein großer Fan von Low-Code / No-Code Software, und in der Tat habe ich sie seit 20 Jahren. Vor kurzem, für die Vergangenheit sieben Jahre habe ich bei UC Irvine und für LinkedIn Learning unterrichtet.

Rosaria: Hat sich die Datenwissenschaft in diesen 25 Jahren erheblich verändert?

Keith: Einige Dinge blieben wirklich die gleichen, aber ich sah eine große Veränderung um die Aussprache über no-code/low-code. Zurück in den 90er Jahren, als SAS und SPSS den Raum dominierten,

zwischen den SPSS/SAS Programmierern und Menschen, die verwendete Menüs, um Datenflüsse zu konfigurieren und auszuführen. Es gab die falsche Vorstellung, Datenanalyse war das exklusive Reich der Programmierer, und wenn Sie Menüs verwenden, fühlten Sie Ihre Kompetenzen wurden befragt und Ihre Ergebnisse weniger geschätzt. Das ist verrückt, sogar mehr, wenn Sie denken, dass die Theorie dahinter absolut die gleiche ist unabhängig von ob Sie z.B. eine Regression mit Code oder ohne bauen. Dann, für einige Jahre die Dynamik in Programmiersprachen wie Python und R verschoben, bevor zurück zu No-Code / Low-Code-Tools. Menschen erkennen, dass ein visuelles Werkzeug kann von großer Unterstützung, um Zeit und Prototyp-Lösungen zu sparen, bevor sie zu Produktion. Und KNIME ist fantastisch sowohl im Prototyping als auch in der Produktion!

Rosaria: Wo lehren Sie derzeit?

Keith: Ich lehren seit 2015 für das Erweiterungsprogramm von UC Irvine. Seit Januar dieses Jahres habe ich einen meiner Lieblingskurse dort unterrichtet. Der Kurs konzentriert sich auf die Datenverstehensphase von CRISP-DM, die oft missverstanden ist Meiner Meinung nach. Wenn die Schüler “Datenerkundung” hören, denken sie sofort an Daten Visualisierung. Während die Datenvisualisierung ein wesentlicher Teil der Datenexploration ist, ist es nicht Visualisierung zur Berichterstattung. In der Datenverstehensphase von CRISP-DM wollen wir die Aspekte der Daten, die wir bei der Datenvorverarbeitung beheben müssen, aufdecken oder die uns Einblicke in das geben, was Algorithmen die beste Passform sein könnten. Es sind Daten Visualisierung mit einem Auge zur Modellierung. Das ist oft überraschend für Studenten, die Denken Sie in der Regel an die Datenvisualisierung nur für die Berichterstattung.

Außerdem, manchmal Ich lehren auch einige einleitende maschinelle Lernkurse, wo KNIME mein Werkzeug ist.

Rosaria: Welche Themen decken Sie in Ihren UC Irvine-Kursen und was sind das Lernen Ergebnisse für die Studenten? In welchen dieser Kurse nutzen Sie KNIME (und wie)?

Keith: Im speziellen Fall von UC Irvine starteten sie eine Predictive Analytics Zertifikatsprogramm Ich bin schon seit einigen Jahren dabei, alle Module zu lehren über die Zeit. Das Programm ist um die CRISP-DM Phasen strukturiert: von Problem Definition und Datenaufbereitung, Einführung in Modellierung und Bereitstellung. Also, Ende des Programms sind die Studierenden allen Phasen des Lebenszyklus der Datenwissenschaft ausgesetzt.

In meinen Kursen und in vielen Kursen anderer Ausbilder in unserer Abteilung ist KNIME das Go-to-Tool zum Lehren. Zum Beispiel in einem Kurs über die Einführung in die Modellierung, wo ich jede Woche einen neuen Algorithmus, wie skalierungsabhängige Algorithmen, abdecke, Klassifizierungsalgorithmen, etc., was ich wirklich brauche ist ein Werkzeug, das den Schülern hilft zu sehen, wie die theoretischen Konzepte, die sie in den Büchern lesen, können praktisch umgesetzt werden. Aus pädagogischer Sicht bietet KNIME eine großartige Gelegenheit, wirklich zu praktizieren theoretische Konzepte.

Anmerkung: Im Jahr 2023 werden wir eine drei Kursversion dieses Zertifikats vorstellen wo ich der einzige Ausbilder bin. Es wird mehr Kontinuität geben und Lernende lernen mehr KNIME. Ich bin begeistert davon.

Rosaria: Wie würden Sie Ihren Lehransatz und Ihre Erfahrung beschreiben

Konzepte mit KNIME umsetzen?

Keith: In der ersten Woche bitte ich die Schüler, die einleitenden Kapitel der [Techniken](#) von Michael Berry und Gordon Linoffs Buch und Dean Abbotts

[Daten Bergbau](#)

[Angemeldet](#)

[Predictive Analytics](#), die in der Gemeinschaft sehr beliebt ist. Das sind Lehrbücher

die ihnen ein theoretisches Verständnis der Schlüsselbegriffe im Datenbergbau geben. In

Zusätzlich zu den Lesungen, noch bevor wir mit den häuslichen Hausaufgaben beginnen, bitte ich sie,

KNIME herunterladen und installieren. Als nächstes, um sich mit dem Werkzeug vertraut zu machen, habe ich [sie wählen jeden Workflow verfügbar auf](#) [KNIME Beispiele für Server](#) oder [KNIME Hubraum](#) und

demonstrieren, dass sie es zu arbeiten und verstanden. In der zweiten Woche wähle ich eine

Beispiel -say, binäre logistische Regression- und fragen sie, dass sie

verstanden. Basierend auf dem, was sie in den Lesungen gelernt haben und KNIME praktizieren,

Ich erstelle einen neuen Datensatz und fordere sie heraus, den Workflow zu arbeiten und auszuführen erfolgreich. Ich glaube, dies ist eine effektive Möglichkeit, theoretische Konzepte zu versinken und

geben den Studierenden die Möglichkeit, aus erster Hand zu erfahren, was es bedeutet, Daten zu modellieren.

Darüber hinaus, ergänzend zu meiner Lehre, ich oft die Studenten auf die wunderbare

[kostenlos verfügbare Inhalte](#) [KNIME TV](#) so dass sie ein tieferes Verständnis der

Werkzeug.

Rosaria: Beschreiben Sie in Ihren Kursen Fallbeispiele für Datenwissenschaft? Wenn ja, was

Eins?

Keith: Es gibt definitiv einige praktisches Lernen, wo sie mit Datensätzen experimentieren, insbesondere diejenigen, die Sie auf dem KNIME-Beispielserver finden können, oder die in das Repository von UC Irvine oder Kaggle. Allerdings gibt es eine Grenze zu dem, was Sie mit einem Praxisdatensatz, wie den sehr beliebten Iris-Datensatz. Deshalb ist, was ich versuche zu tun, sich von dem inspirieren lassen, was sie in den Nachrichten oder auf YouTube gesehen haben, und dies in Form von Diskussionen einbeziehen. Idealerweise ist der nächste Schritt, dass sie eine neue Datensätze, die ein tatsächliches Phänomen beschreiben und ihr Wissen anwenden, um real zu bauen Weltbeispiele.

Rosaria: Hat COVID-19 und verschiedene Lockdowns Ihre Lehrweise beeinflusst?

Keith: In den Monaten, in denen Covid-19 tobte, mussten die meisten meiner Konferenzarbeiten

Stoppt. Ich habe in Konferenzen wie der Data Warehouse Institute Conference in

die USA, die gewöhnlich viermal im Jahr stattfanden. Für einige Zeit, Covid-19 auch

Mein Linked In der Lernarbeit. Ich flog nach Santa Barbara, um in einer

schallisoliertes Studio mit meinem Produzenten. Es fühlte sich an, als hätte man ein Audiobuch erzählt, aber das auch musste während des Abriegelns aufhören.

Andererseits war meine Lehre bei UC Irvine nicht von der Pandemie betroffen, wie sie war immer für mich fern.

Rosaria: Wo hat KNIME Ihre Lehre am meisten erleichtert?

Keith: Ich denke, KNIME ist perfekt für Fernunterricht und Aufgaben. Zurück in den Tagen, als ich anfing zu lehren, würde ich um den Raum gehen und Studenten helfen wann immer sie wurden festgeklemmt. Bei der Fernlehre ist das oft nicht mehr der Fall. Ein Faktor ist vielleicht Ich habe gelernt, mehr überlegt zu sein, wie ich Struktur KNIME Aufgaben, die das Lernen erleichtern dürften. Darüber hinaus, Online-Konferenz Werkzeuge haben in den letzten zwei Jahren so viel verbessert, dass, wenn Studenten Hilfe benötigen, sie fordern einfach eine 1-on-1 Sitzung mit mir auf Zoom.

Rosaria: Sie lehren eine Reihe von Kursen auf LinkedIn Learning, und einige von ihnen Funktion KNIME. Was sind die Titel? Was decken diese Kurse ab?

Keith: Das stimmt. Ich habe mehrere Kurse für LinkedIn Learning in der in den letzten Jahren. Genau, 19 Kurse. Einige von ihnen beinhalten KNIME, aber einige nicht alle softwarebezogenen Inhalte einbeziehen - vor allem jene Kurse, die ansprechen Führungskräfte, die in der Regel mehr daran interessiert sind, wie man Wert aus dem maschinellen Lernen auf Unternehmensebene. Unter den Kursen, in denen KNIME das Go-to-Tool ist, bin ich sehr stolz auf zwei immergrüne: [Einführung in maschinelles Lernen mit KNIME](#), [Wissenschaftsstiftungen: Datenbewertung für prädiktive Modellierung](#), [Daten](#)

In Bezug auf die Dauer ist der letzte Kurs eine Ausnahme in der Bildungslandschaft von LinkedIn Lernen. Es ist ein 4-stündiger Kurs, der einen systematischen Ansatz für die Datenverstehensphase für vorausschauende Modellierung, wie Datenformatierung, fehlende Wertanalyse, etc. Ich glaube, KNIME ist ein großartiges Werkzeug, um diese Konzepte zu erfassen.

Rosaria: Gibt es einen neuen LinkedIn Learning Kurs mit KNIME, dass Sie sind besonders glücklich darüber?

Keith: Im Februar 2022 ein neuer Kurs über [maschinelles Lernen und erklärende KI](#) war veröffentlicht. Ich freue mich sehr, dass der Kurs in dieses Jahr geschoben wurde, weil KNIME hat so viele fantastisch hinzugefügt [geprüfte Komponenten für XAI](#) dass ich mich integrieren konnte in meinem Kurs. XAI ist ein so heißes Thema im Moment, aber was Sie in der Regel finden können online ist entweder hochrangige Übersichten oder sehr math-intensive Inhalte. Was fehlt war etwas dazwischen, das von einer Übersicht ausgeht, aber dann eine detaillierte Erläuterung einiger Techniken mit konkreten Anwendungen. Dieser neue Kurs versucht, die Lücke schließen.

Anmerkung: Vor kurzem habe ich meine zwei Decision Trees Kurse für KNIME aktualisiert.

Wenn das abgeschlossen ist, wird es die gesamte KNIME-bezogene Kurse auf fünf bringen.

Rosaria: Wie kompliziert war es, sich auf diese Kurse vorzubereiten?

Keith: Vorbereitungskurse für Linked In Learning unterscheidet sich ziemlich von anderen entfernt Kurse. Es dauert in der Regel etwa 6 Monate von der Konzeption bis zum Endprodukt. Neben der Vorlaufzeit gibt es eine lange Vorlaufzeit, in der Kurse bearbeitet und poliert werden vor der Veröffentlichung. Zum Beispiel, wenn ich einen Kurs entwerfen, der KNIME, I in der Regel einen detaillierten Umriss erstellen und an das Evangelismus-Team weitergeben. Sie nehmen ein schnell aussehen und geben Sie mir Vorschläge, um mir zu helfen, die besten Knoten für die jeweilige umrisse. Als Ergebnis, wenn der Kurs online kommt, ist es so aktuell, wie es kann möglicherweise, da es vom Team überprüft wurde.

Rosaria: Lassen Sie uns zurück zu Ihrer Arbeit als Erzieher, die auch an der Universität oder andere Institutionen. Was glauben Sie, dass die Bildung der Datenwissenschaft heute fehlt?

Keith: Was fehlt, ist ein tiefes Verständnis des Lebenszyklus der Datenwissenschaft, sei es CRISP-DM oder einem anderen Prozessmodell. Heutzutage, in der Datenwissenschaft Bildung sehr wenige Ausbilder lehren Prozessmodelle. In vielen Fällen ist der Fokus auf Codierung so stark, dass Prozess- und Lebenszyklusprobleme funktionieren nie in den Lehrplan. Codierung ist sicherlich wichtig, aber wenn wir nicht integrieren Prozess- und Lebenszyklus Probleme in die Curriculum, wir beenden die Bildung von Studenten, die nicht wissen, wie man Problem zu tun Definition oder haben keine Kompetenz in einer der Phasen vor der Modellierung. Es gibt die falsche Vorstellung, dass alles, was zählt, ist die Modellierung von Algorithmen und Software zu bekommen das Modell ausführen. Es gibt keinen Raum für Datenverstehen, Datenaufbereitung und Business Verständnis.

Rosaria: Im Licht Ihrer langen Erfahrung im Bereich der Datenwissenschaft haben Sie Karriereberatung für junge Datenwissenschaftler?

Keith: Ich würde empfehlen, während ihrer Daten-Wissenschaftsreise ihre Zeit zu widmen und ebenso Anstrengung für Prozess- und Lebenszyklus, Konzepte und Ausführung. Beispielsweise, wenn sie suchen ein Programm, ein Boot-Camp oder irgendwelche Zertifikatkurse, wo der Fokus ist ausschließlich auf der Codierung und die Datenwissenschaft Lebenszyklus und Prozess werden ignoriert, sie sollte einen Weg finden, diese Themen zu behandeln, weil sie sie auf dem Job benötigen.

Eine andere Sache, die sie brauchen, ist, eine echte, praktische Erfahrung zu bekommen, zum Beispiel mit einer Data Science-Lehre, die nicht im Gegensatz zu einem Artaufenthalt ist. Während dieser Zeit, es ist entscheidend, dass sie die richtige Mentorschft von jemandem bekommen, der mehr erfahren und können sie führen und inspirieren.

Rosaria: Wir erreichen das Ende unseres Interviews. Bevor wir uns verabschieden, können wir nicht aber die klassische Frage stellen. Wo sehen Sie die Datenwissenschaft in den nächsten Jahren? Was wird die nächste Innovation sein?

Keith: Es gibt eine Menge Hype um AutoML, und es ist wahrscheinlich, dass es weiter wachsen wird. Ich sehe den Wert von AutoML, vor allem für Prototyping-Lösungen, wo Zeit Geld ist, und wenn es uns mit Modellen präsentiert, die der richtigen Bewertung würdig sind.

Nachdem wir das gesagt haben, müssen wir den Mythos debunken, dass AutoML Menschen ersetzen wird in dem Prozess durch einfaches Drücken der Bereitstellungstaste. In CRISP-DM, nach dem Die Modellierungsphase ist die Evaluationsphase. Bewertung bezieht sich nicht auf das Ranking Modelle nach Genauigkeit - das ist immer noch Teil der Modellierungsphase. Wenn wir zur Bewertung kommen, wir werden erwartet, Wert auf die Organisation zu bringen und uns fragen "Ist das Modell bereit für den Einsatz?" oder gibt es das Potenzial für organisatorischen Widerstand?". Wir muss im Auge behalten, dass wir Modelle für Organisationen bauen, die mit Leute. Das bedeutet, dass wir den menschlichen Aspekt nicht beseitigen können - weder in der Art, wie wir definieren das Problem, noch in der Weise, wie wir Modell-Features entwickeln. Menschen sind dabei zu bleiben.

Rosaria: Wie können sich Leute vom Publikum mit Ihnen in Verbindung setzen?

Keith: Die Leute können mich entweder über [LinkedIn](#) oder auf meiner Website [Keith McCormick](#).

Dieser Artikel wurde erstmals in unserem [Niedriger Code für Advanced Data Science Journal](#) auf Medium. Die Originalversion finden [Hier](#).

Beobachten Sie das ursprüngliche Interview mit Keith McCormick auf YouTube [Mein Data Guest – Ep 4 mit Keith McCormick](#).



Wie ist das? wurde nominiert KNIME-Beitrag der Monat für Mai 2021. Er wurde für seine akademische Tätigkeit und seine Hilfe bei der Gestaltung der [KNIME Zertifizierungsprogramm](#) durch die Gestaltung vieler Prüfung Fragen und Hilfe bei der Strukturierung des Programms. Also, wenn Sie nahmen die Prüfung und fanden die Fragen zu schwer Antwort... Nun, jetzt wissen Sie, was es braucht, um ein KNIME Beitrag des Monats.

Giuseppe ist derzeit Professor an der Fakultät für Informatik der Freien Universität Bozen-Bolzano, Italien, wo er die Kurse unterrichtet und Datenorientierte Entscheidungsfindung Vorher war er Leiter der Abteilung der Informatik am [Universität des Lesens](#), UK. Unter den vielen Werkzeugen für Datenwissenschaft führte er in seinen Kursen auch die KNIME Analytics Platform ein, weil er sieht den Vorteil einer Open Source, Low-Code-Daten-Wissenschafts-Tool in einem Studenten Portfolio für ihre zukünftige Karriere. Nach seinem Abschluss an der Universität von Palermo, Italien, Giuseppe Ventured in der akademischen Welt und sogar in der ersten KNIME-Entwicklungsteam an der Universität Konstanz, Deutschland, bis zur ersten Veröffentlichung von KNIME 1.0 im Jahr 2006.

Besuchen Sie Giuseppe's Raum auf dem KNIME Hub oder
Profil im KNIME Forum (Hub/Forum Griff:
Difatta)



Zertifiziert werden – Holen Sie sich das KNIME Zertifizierungsprogramm

Sind Sie Experte für KNIME Software?

Autoren: Giuseppe Di Fatta & Stefan Helfrich

Anmerkung des Herausgebers:

Dieser Artikel, datiert 2019, beschreibt den ersten Schritt zur Errichtung der KNIME-Zertifizierung Programm. Es ist dank Prof. Giuseppe Di Fatta, dass der erste Pool von Fragen entworfen wurde und getestet für alle zertifizierten KNIME-Nutzer. Seitdem das KNIME Zertifizierungsprogramm hat sich erweitert und verbessert und wird dies weiterhin tun. Dennoch sind wir Prof. Di Fatta, um uns zu helfen, den ersten Schritt zu machen. In der Zwischenzeit, Prof. Giuseppe Di Fatta hat sich geändert seine Zugehörigkeit zur Universität Bozen in Italien.

Es gibt einen offiziellen Weg, diese Frage zu beantworten und sie mit der Welt zu teilen: Du kannst testen Sie Ihre KNIME-Kompetenz mit einem Zertifizierungsprogramm in Zusammenarbeit zwischen Wissenschaft und Industrie.

Berufszertifizierungen sind im Arbeitsprozess besonders nützlich, um zu helfen Schlüsselqualifikationen, die für das von Arbeitgebern angestrebte Arbeitsprofil relevant sind. Sie erleichtern Anpassung der Nachfrage nach Fähigkeiten mit dem Angebot in einem früheren Stadium und auch Förderung der brauchen die richtigen Fähigkeiten. Sie helfen potenziellen Bewerbern, die Anforderungen an den aktuellen Arbeitsmarkt, um ihre Ausbildung und Entwicklung zu planen effektiv.

Mitarbeiter können Zertifizierungen verwenden, um aktuelle Mitarbeiter in kontinuierlich Professional Development (CDP) relevant für kritische Bedürfnisse. Während der Hochschulbildung Grad sind Beweis für eine solide Kenntnis eines Fachgebiets (z.B. BSc Computer Wissenschaft, MSc Data Science), Zertifizierungsprogramme neigen dazu, sich auf sehr spezifische Kenntnisse und Fähigkeiten zu branchenrelevanten Werkzeugen und Prozessen. Zertifizierungsprogramme Unterstützung bei der eindeutigen Identifizierung und Übermittlung des richtigen Kompetenzniveaus. Zertifizierungstests werden verwendet, um Fähigkeiten und Kenntnisse zu diesem Zweck zu bewerten.

Fähigkeiten im Bereich der Datenwissenschaft, des maschinellen Lernens und der Analytik sind in mehr Nachfrage als je zuvor. Die KNIME Analytics Platform ist eine der führenden Plattformen. Die Datenwissenschaft mit KNIME Software Zertifikate aus der KNIME Zertifizierungsprogramm sind Zeugnisse der Effizienz in der Open-Source-Plattform für datengetriebene Innovation: Sie zeigen Ihnen Fähigkeit, Datenanalyseprojekte zu entwickeln, auszuführen und bereitzustellen. Zertifikatsinhaber Verbesserung ihrer beruflichen Glaubwürdigkeit; Arbeitgeber werden leichter identifizieren Kandidaten einen Wettbewerbsvorteil zu gewinnen.

Über die Mitarbeiter

KNIME hat sich mit der Universität Reading zusammengetan, um die KNIME Zertifizierung Programm . Die Motivation war, die Erfahrung und das Know-how aus der Wissenschaft zu ziehen und es anwenden, um ein effektives Zertifizierungsprogramm zu erstellen. Mit Forschungskompetenz in Data Wissenschaft, maschinelles Lernen, Big Data Analytics und High Performance Computing für Wirtschaftliche Wissenschaft, Abteilung Informatik der Universität Reading, geleitet von Dr. Giuseppe Di Fatta, war ein idealer Partner.

Die University of Reading erhielt ihren ersten Abschluss in Informatik genau 50 vor Jahren 1969. Die Abteilung Informatik hat viele Jahre Erfahrung in der Lehre Data Analytics, Data Mining und Machine Learning, und außerdem sie haben die KNIME Analytics Platform in der Lehre Data Analytics und Data Mining angenommen seit über 10 Jahren auf Bachelor-Ebene und vor kurzem auf Postgraduierten Ebene als Gut.

KNIME Zertifizierungsprogramm

Das Zertifizierungsprogramm umfasst drei Ebenen (L1 bis L3) mit zusätzlichem Spezial Thema Prüfungen. Jede Ebene unterstreicht die Kompetenz einer Person mit unterschiedlichen Aspekten und praktische Fähigkeiten für KNIME Software sowie die aktuellsten Data Science-Konzepte und weiß wie.

Passmarken für die Zertifizierungstests sind bei 70% (dies basiert auf der Gradgrenze Typischerweise im britischen Bachelor-Abschluss-Klassifikationssystem für First-Class Auszeichnungen Grad) und Zertifikate werden für 2 Jahre gültig sein. Auf diese Weise Arbeitgeber können versichert werden, dass ein Bewerber mit einem KNIME-zertifizierten Anmeldetag aufsteht bisher mit den neuesten Entwicklungen in KNIME.

Prüfung

- L1
 - Grundkenntnisse in der KNIME Analytics Platform
 - Prüfung: 30-minütige Multiple-Choice-Prüfung (15 Fragen)
- L2
 - Fortgeschrittene Effizienz in der KNIME Analytics Platform
 - Prüfung: 30-minütige Multiple-Choice-Prüfung (15 Fragen)
- L3
 - Kompetenz in der KNIME Software zur Zusammenarbeit und Produktion Datenwissenschaft
 - Prüfung: 90-minütige Multiple-Choice-Prüfung (50 Fragen)

Wie zu studieren?

Um sich auf diese Zertifizierungsprüfungen vorzubereiten, empfehlen wir folgende Methoden der Studie:

- [KNIME Selbst gepflegte Kurse](#)
- [Lehrveranstaltungen für Anfänger und Erweiterete Benutzer](#)
- [Universität des Lesens Kurse für die Statistik](#) auf UG- oder PGT-Ebene (z.B. CS3DM16) und CSMDM16).

Wie kann ich die Prüfung machen?

Erfahren Sie mehr über die verschiedenen Prüfungsstufen und auch Links zur Zertifizierung [Zeitplan in der KNIME Zertifizierungsprogramm](#) :

Dieser Artikel wurde erstmals in unserem [KN-Code Blog](#). Die Originalversion finden [Hier.](#)



Alzbeta Tuerkova wurde nominiert KNIME-Beitrag der Monat für Juni 2021. Sie wurde für sie vergeben Forschungspapier sie veröffentlichte zusammen mit Barbara Zdrrazil über ein effizientes und reproduzierbares, vollständiges KNIME basierte, drogenreaktivierende Anwendung, um neue Medikamente zu identifizieren Kandidaten für seltene Krankheiten und die Covid-19. Die Workflow, Tutorials und Informationen über COVID-19 Daten wurden der Wissenschaft frei zugänglich gemacht

Community für Follow-up-Studien und Lernen (siehe

„Automatisiertes Medikament Repurposing Pipeline in KNIME“

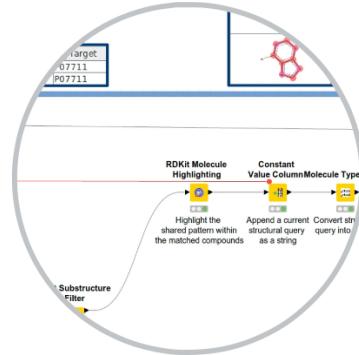
Alzbeta ist derzeit Leiter von Computer-Aided Drug Design bei Celeris Therapeutics. Sie hält einen Doktor in Biologie, mit einer Dissertation über hepatische organische Anion Transport von Polypeptiden des SLCO

Familie, aus dem Universität Wien

Besuchen Sie Alzbeta's Raum auf dem KNIME Hub

(Hub)

Griff: In den Warenkorb)



Ein Jahr der Pandemie: Wie KNIME hilft Neues zu finden

Drogenkonsumenten

Autor: Alzbeta Tuerkova

Die Entwicklung eines neuartigen Medikaments wird allgemein als komplexer Prozess anerkannt, der eine erhebliche Zeit und Kosten genehmigt ist zwischen 12 und 15 Jahren. Um den Prozess der neuen Droge zu beschleunigen In die Pipeline können Entdeckungs-, Medikamentenrückgewinnungsstrategien integriert werden. Drogen Wiederauflösen (auch als Medikamentenrepositionierung bezeichnet) ist eine Neubewertung einer bereits vorhandenen Drogen, um sein Potenzial bei der Behandlung einer neuartigen Krankheit zu testen. Dieser Ansatz wird vorteilhaft, um Zeit, Kosten und Risiken auf die mögliche Seite zu reduzieren Auswirkungen.

Von einem rechnerischen Standpunkt aus können die Drogensätze Text enthalten Bergbau, Netzwerkanalysen, Maschinen (tiefen) Lernmodelle zur Vorhersage von Drogenziel-Krankheitsbeziehung und strukturbasierte Modellierungsmethoden (protein-). Mit kontinuierliches Wachstum von biomedizinischen Datensätzen, rechnerische Arzneimittelrückgewinnungsverfahren haben große Aufmerksamkeit erregt. Insbesondere der Abbau chemischer Strukturen für der Datenvergrößerung (als Andockbibliothek zu verwenden oder eine Maschine zu trainieren) Lernmodell) ist nicht mehr der einzige Zweck. Stattdessen hilft eine groß angelegte Datenanalyse in den chemischen Daten verborgene Disentangle-Muster, die wiederum in der Früh- die Entwicklung der Drogen. Darüber hinaus rechnerische Drogensätze stark profitieren von der Kombination verschiedener Arten von Dateneinheiten, wie Gewebe Expressionsdaten, Gene, Drogenziel-Interaktionen, Krankheitsdatensammlungen und Phenotypen, um einen Hinweis auf die Wirkungsweise eines Arzneimittels zu liefern. Nicht zuletzt, chemische Daten aus verschiedenen Quellen wurden gefunden, um verschiedene Bereiche zu erfassen chemischer Raum. Daher hilft die Kombination von Daten aus mehreren Datenbanken auch Erhöhung der Vielfalt der physikalisch-chemischen Eigenschaften der Endverbindungssätze.

In diesem Beitrag präsentieren wir einen Workflow zur Durchführung von Liganden-basierten *in silico* Drogen Wiederauflösen. Die Anwendbarkeit des entwickelten Workflows wird anhand von von COVID-19. Unsere Strategie basiert auf dem molekularen Ähnlichkeitsprinzip: Strukturell ähnliche Moleküle neigen dazu, ähnliche biologische Aktivitäten zu besitzen. Daher verwenden wir öffentlich hinterlegte strukturelle und bioaktivitätsbezogene Daten, die mit Proteinen verbunden sind Ziele, die an der Krankheit unseres Interesses beteiligt sind. Als nächstes identifizieren wir bereichert molekulare (Sub-)Strukturen, um Unterstruktursuche der Datensätze durchzuführen der verfügbaren Drogen für die Suche nach neuen - potenziell aktiven - Drogenkandidaten. Die Analytik Plattform KNIME dient hier als Werkzeug zur Automatisierung des gesamten Verfahrens und bietet den zusätzlichen Vorteil von Flexibilität, Wiederverwendbarkeit und Transparenz. Zusätzlich, unsere Ein Workflow ist einfach reproduzierbar und kann je nach Einzelprojekt angepasst werden Bedürfnisse.

In einem näheren Detail enthält der Workflow gezieltes Herunterladen von Daten durch Anwendung Programmierschnittstellen (APIs), molekulare Strukturen Standardisierung, Daten Integration, Identifikation struktureller Analoga durch hierarchisches Gerüst-Clustering und maximale gemeinsame Substrukturerzeugung, gefolgt von der retrospektiven Analyse DrugBank und ein Datensatz von antiviralen Arzneimitteln von Chemical Abstracts Service (CAS)

Programmatic Data Access to Life-science Datenbanken und Molekulare Strukturen Standardisierung

Bei der Kombination von Daten aus verschiedenen Repositories ist es vorteilhaft, Datenbanken in eine programmatische Weise. Datenbanken in diesem Workflow genutzt - UniProt, Protein Data Bank (PDB), ChEMBL, Guide-To-Pharmacology (IUPHAR), PubChem und DrugBank - einen gezielten Zugriff der gespeicherten Daten über eine Applikations-Programmierung ermöglichen Schnittstelle (API). Hier wird eine Triade von KNIME-Knoten nacheinander ausgeführt (1) geben eine API-Anforderung (der Knoten "String Manipulation"), (2) Daten von Webdiensten abrufen (der Knoten „GET-Anforderung“) und (3) die entsprechenden Eigenschaften aus empfangenen Dateien (die „XPath“-Knoten). In einem ersten Fall Protein-Ziel-Identifier der Open Targets-Plattform werden auf ihre entsprechenden UniProt-IDs abgebildet. Die abgerufenen UniProt-IDs dienen als Startpunkt zum Abrufen von Protein-Liganden-Strukturdaten von PDB sowie Liganden Bioaktivitätsdaten von ChEMBL, IUPHAR und PubChem.

Voraussetzung für die Verschmelzung von Ligandendaten aus unterschiedlichen Quellen ist die Standardisierung molekulare Strukturen. Um eine einheitliche chemische Datendarstellung zu gewährleisten, eine Kaskade Knoten (involvierende RDKit- und CDK-Knotenerweiterungen) wurden ausgeführt, um (1) entfernen Verbindung Stereochemie (der "String Replacer"-Knoten), (2) Streifensalze durch Weiterleitung einem vordefinierten Satz unterschiedlicher Salz/Salz-Gemische (Knoten "RDKit Salt Stripper"), (3) alle gestreiften Salzkomponenten in der Ausgabetafel (der "Connectivity"-Knoten gefolgt durch den Knoten "Split Collection Column" neutralisiert (4) Ladungen und überprüft nach Möglichkeit atomare Konflikte (der "RDKit Structure Normalizer"), (5) Filterdaten durch Überprüfung spezifische Elemente (der "Element Filter"), (6) erzeugen InChI, InChiKey und Canonical Lächeln Formate mit den entsprechenden RDkit-Knoten.

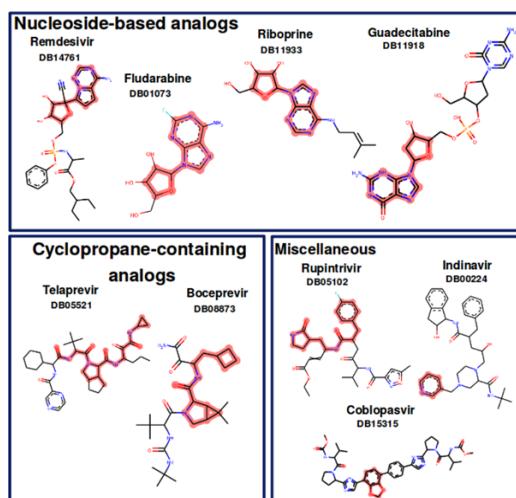
Substruktur sucht potenzielle Drogenkandidaten zu identifizieren geeignet für Arzneimittelrückgewinnung

Molekulare Strukturen werden auf ihre Bemis-Murcko Gerüste reduziert. Erzeugt Gerüste und zugehörige UniProt-IDs (als Liste) werden als Eingabe verwendet (die "Tabelle zu Variabler Knoten). Molekulare Abstände werden für die zurückgehaltenen Gerüste berechnet (MoSS) MCSS Molecule Similarity' Node) und hierarchische Cluster werden entsprechend generiert (der Knoten „Hierarchical Clustering“ [DistMatrix]). Gerüste werden zu einem bestimmten Cluster, indem der Knoten „Hierarchical Cluster Assigner“ verwendet wird. Die Schleife wird dann wiederholt für die restlichen Ziele aus der Eingabetafel. Die Ausgabetafel enthält UniProt-IDs,

zugehörigen Gerüsten und Cluster-IDs. Als nächstes, Schleifen über verschiedene Cluster der zugehörigen Gerüste für ein jeweiliges Ziel erfolgt, um ein Maximum gemeinsam zu schaffen Unterstruktur (der "RDKit MCS"-Knoten) aus allen zugehörigen Gerüsten, die zu einem entsprechende Cluster. Die generierten Unterstrukturen (in SMARTS) werden dann als Abfragen verwendet um Treffer in DrugBank zu finden (Eingabedatensatz beinhaltet Moleküle im SDF-Format, DrugBank IDs und zugehörige Inhalte). Die Strukturen der DrugBank sind standardisiert und dann dem Knoten "RDKit Substructure Filter" unterzogen, um Unterstruktur durchzuführen sucht. Die jeweilige SMARTS-Abfrage wird als Variable an den Eingang weitergeleitet. Entdeckte Unterstrukturen werden durch den Knoten „RDKit Molecule Highlighting“ hervorgehoben. Die Ausgabetabelle enthält identifizierte Treffer (Molekulnamen, assoziierte Ziele, SMARTS Schlüssel, chemische Strukturen), und hervorgehobene Unterstrukturen im SVG-Format.

COVID-19 als Anwendungsfall

Unterstruktursuche halfen, 7836 Verbindungen von DrugBank und 36,521 zu identifizieren Verbindungen aus dem CAS-Datensatz, Aus diesen Hits wurden 135 Verbindungen abgerufen von der DrugBank und dem CAS-Datensatz. Identifizierte MCS können zu fünf kombiniert werden getrennte Cluster: (1) Auf der Grundlage der offenkettigen Strukturschlüssel ermittelte Treffer (59) Treffer), (2) Nucleoside/Nukleotidanaloga (53 Treffer, z.B. Remdesivir, Flaudarabin, Riboprine), (3) Verschiedenes, die ubiquitöse Unterstrukturen enthalten (22 Treffer, z. rupintrivir, indinavir, darunavir), (4) zyklopropanhaltige Treffer (3 Treffer, z.B. Telaprevir), und (5) Adamantanderivate (3 Treffer), siehe unten. Es ist bemerkenswert, Erwähnen Sie, dass einige der identifizierten Medikamente jetzt klinischen Tests unterzogen werden, um ihre Potenzial zur Bekämpfung von COVID-19. Darüber hinaus half der Workflow, Roman zu identifizieren interessante Kandidatenmoleküle, die zukünftige therapeutische Entwicklung informieren können, COVID-19 behandeln.



Beispiele für identifizierte Medikamente mit strukturellen Abfragen hervorgehoben. Von Tuerkova A, Zdravil B. J Cheminf. 2020 Dec;12(1):1-20.

Mit dem Workflow im virtuellen Klassenraum

Der Workflow wurde im Sommersemester 2020 (20.–24. April) innerhalb des Kurses genutzt „Experimentelle Methoden in der Drogenentdeckung und der präklinischen Drogenentwicklung“ das englischsprachige Masterstudiengang Drug Discovery and Development bei Universität Wien (<http://drug-dd.univie.ac.at/>) Durch die Schutzmaßnahmen durch die COVID-19 Pandemie verursacht, wurde dieser Kurs als virtueller Klassenraum strukturiert. Die Studenten besuchten Online-Sessions, in denen die verschiedenen Schritte des Workflows waren erklärt und demonstriert. Tutorials und die verschiedenen Teile des Workflows haben täglich ausgehändigt. Am letzten Tag des 5-Tages-Kurses wählte jeder Schüler ein der Treffer, die von der Unterstruktur-Suche abgerufen und einige Zeit der Literatur gewidmet sucht. Schließlich hat jeder Student einen Bericht vorgelegt, der zusammenfasst, was über die ausgewählte Verbindung und ihre potenzielle Nutzbarkeit für die COVID-19-Behandlung (nach dem, was im April 2020 bekannt war).

Dieser Booklet-Beitrag basierte auf dem veröffentlichten Artikel: Tuerkova A, Zdrazil B. Ein Ligand basierte computergestützte Arzneimittel-Repurcing-Pipeline mit KNIME und Programmatic Data Access: Fallstudien für seltene Krankheiten und COVID-19. Journal of Cheminformatics. 2020 Dec; 12(1):1-20. <http://doi.org/10.1186/s13321-020-00474-z>



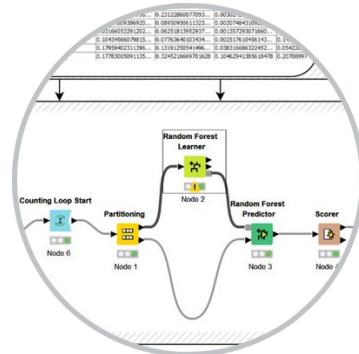
[Malik Yousef](#) wurde nominiert Beitrag des Monat für Januar 2022. Er wurde für seine Forschung ausgezeichnet auf DNA, mRNA und miRNA-Analyse, Gen-Onlogie und Molekularbiologie. Mit KNIME hat er Algorithmen zum maschinellen Lernen [mRNA und ihre Ziele](#), [Genexpression analysieren](#), [und mRNA und miRNA Expressionsprofile über maschinelles Lernen](#). Die Bild rechts zeigt einen Workflow-Snippet von seinem

[Forschungspapier über MiRcorrNet](#) :

Malik ist Data Scientist mit dem Schwerpunkt Bioinformatik mit Anwendungen für verschiedene biomedizinische/biologische Probleme. Er verfügt über umfangreiches Know-how in der Maschine das Lernen und seine Anwendungen,

lange Liste von Publikationen in Top, Peer überprüft wissenschaftliche Zeitschriften, Bücher und US-Patente. Als Professor, er hat hochmoderne Beiträge geleistet in Life Sciences, und er nutzt KNIME aktiv in seiner Forschung. Er ist derzeit Leiter der Galilee Digital Health Research Center (GDH) am Zefat Academic College in Israel.

[Besuchen Sie Malik's Raum auf dem KNIME Hub](#) oder [Profil im KNIME Forum](#) (Hub/Forum Griff: Malik)



Gene Ontologie, Biomarker und Krankheitsprädiktion: Vom Forschungslabor zur Data Science Klasse

My Data Guest — Ein Interview mit Malik Yousef

Autoren: Stefan Helfrich & Aline Bessa



Es war unser Vergnügen, kürzlich live auf LinkedIn Malik Yousef als Teil der

[Mein Data Guest](#) [Interview-Serie](#).

[Malik Yousef](#) ist Leiter des Galilee Digital Health Research Centers am [Zefat Akademiker Hochschule](#) in Israel. Er hat Genom und Genexpression beeinflusst (DNA, mRNA und miRNA) Analyse, Gen-Onlogie und (computational) Molekularbiologie im Allgemeinen, mit seiner Forschung, wie durch eine lange Liste von Publikationen in Top bestätigt, peer-reviewed wissenschaftliche Zeitschriften, Bücher und US-Patente. Er hat maschinelles Lernen entwickelt Algorithmen, um mRNA und ihre Ziele vorherzusagen, Genexpression zu analysieren und zu integrieren mRNA- und miRNA-Expressionsprofile über maschinelles Lernen.

Stefan: Bitte erklären Sie uns in Ihren genaueren Worten, was ist Ihr wissenschaftlicher Bereich Erfahrung.

Malik: Ich habe einen Hintergrund in Mathematik und Informatik und tat meine PhD in Textbergbau, wo ich einige Methoden entwickelt habe, die nur von positiven Beispielen lernen. Das erste Mal, als ich mit Biologie involviert war, war während meines Postdocs am Wistar Institute in die USA. Damals war die Bioinformatik gerade in den frühen Phasen der Gründung. Ich wirklich genossen mit biologischen Daten zu arbeiten und meine Erfahrung in diesem Bereich. Für die Einfachheit halte ich mich nun als Data Scientist, aber die Daten, mit denen ich arbeite aus der Biologie kommen. Derzeit versuche ich Biomarker unter insgesamt 20000 zu finden Genen. Das heißt, ich suche eine kleine, aber bedeutende Gruppe von Genen, die als Biomarker für Krankheitsvorhersage.

Stefan: An welchen Krankheiten arbeitest du?

Malik: Während meines Postdocs arbeitete ich mit Lungenkrebsdaten, aber jetzt arbeite ich hauptsächlich mit öffentlichen Datensätzen aller Krebserkrankungen, wie z.B. Nierenkrebs, Brustkrebs usw. Die Input zu unseren Algorithmen, im Allgemeinen, sind Genexpressionsdaten und die Biomarker können

Gene Ontologie, Biomarker und Krankheitsprädiktion: Vom Forschungslabor zur Data Science Klasse

von jeder Krankheit kommen. Mehr traditionelle Bioinformatik Ansätze Anleihe Algorithmen aus der Informatik und wenden sie auf Biologiedaten an. Aber meine Ansatz ist die Integration von Kenntnissen der Biologie in die maschinellen Lernalgorithmen, und dann Feature-Auswahl durchführen, um die wichtigsten Gene zu identifizieren.

Stefan: Was war Ihr wirkungsreichster wissenschaftlicher Beitrag auf diesem Gebiet?

Malik: Ich würde meine Beiträge in drei Teile teilen. Erstes Lernen von positiv Beispiele. Die meisten maschinellen Lernalgorithmen verlassen sich auf zwei Klassen, zum Beispiel, schwarz vs. rot. Nehmen Sie nun an, dass die Daten nur eine Art Klasse haben, zum Beispiel nur schwarz. In diesem Fall ist die Frage, wie wir Modelle aus nur einer Klasse lernen und bauen können, d.h. vor allem positive Beispiele? Dies ist, was ich während meiner PhD entwickelt, eine Ein-Klasse Textklassifikation.

Zweitens habe ich viel Arbeit im Bereich der Vorhersage von miRNA mit der Sequenz selbst gemacht. Einfach gesagt, eine Sequenz ist nur eine Kombination von Buchstaben, eine Saite so zu sagen. Auch, miRNA ist eine neue Entdeckung, die eine wichtige Rolle in unserem Körper hat: es reguliert und schließt Genen. Und es stellt sich heraus, dass viele Krebserkrankungen von dieser kleinen miRNA kontrolliert werden.

Und drittens möchte ich meinen integrativen Ansatz hervorheben. Dies bedeutet, Biologie zu integrieren Kenntnisse in das maschinelle Lernen über GSM (Gruppierung, Scoring, Modellierung). Das heißt, wir sind auf der Suche, Daten verschiedener Arten und aus verschiedenen Quellen zu integrieren. Zum Beispiel Genexpressionsdaten mit Ontologien und miRNA-Zielen kombinieren. wir haben entwickelt viele Algorithmen in diesem Bereich und sind mit ihm recht erfolgreich.

Stefan: Haben Sie KNIME für Ihre Forschung verwendet?

Malik: Ich liebe KNIME . Ich benutze hauptsächlich KNIME in Kombination mit Python und R. All my Lehren und all meine Kollaborationen beinhalten KNIME zu einem gewissen Grad.

Lassen Sie mich Ihnen eine kurze Geschichte erzählen: Vor ein paar Jahren ging ich auf eine wissenschaftliche Konferenz in Deutschland nur wegen KNIME. Ich entdeckte KNIME, wenn ich mit einem Kollegen spreche über ein Problem, das ich in meinem Java-Code hatte. Nichts Wichtiges, aber um es zu reparieren, hätte es wahrscheinlich eine Woche. Mein Kollege schlug vor, dass ich KNIME nur zur Lösung des Problems verwenden könnte. Ich war beeindruckt und als ich selbst KNIME lernte. Danach Begegnung Ich begann meine Arbeit an maTE, die einer meiner ersten komplexen Algorithmen war in KNIME umgesetzt. Aber auch auf dieser Ebene der Komplexität, die Workflow-Struktur hilft mir, die Arbeit an Nicht-Programmierer zu kommunizieren.

Stefan: Wo glauben Sie, dass KNIME bei Ihren Projekt-Implementierungen am meisten hilft?

Verbindung mit den Datenquellen oder Trainingsmaschinenlernmodellen oder etwas, das Ich kann nicht daran denken?

Malik: Es gibt verschiedene Auswirkungen von KNIME. Zum einen muss ich mit KNIME nicht verbringen Sie viel Zeit auf Debugging nur um einen kleinen Fehler zu finden. Es gibt mir mehr Zeit, Fokus auf meine Forschung.

Gene Ontologie, Biomarker und Krankheitsprädiktion: Vom Forschungslabor zur Data Science Klasse

Zweitens ist die Fähigkeit, Python oder R mit KNIME zu kombinieren, sehr mächtig. Für viele Dinge, die ich tue, nutze ich R und Python-Code. Mit KNIME ist es möglich, bestehende Ansätze, Werkzeuge oder Algorithmen und sie zu kombinieren.

Ein weiterer großer Einfluss von KNIME ist, dass die Zahlen in meiner Veröffentlichungen.

Und schließlich spart ich Zeit mit KNIME, nur weil ich es schneller in KNIME mache. Ich verwendet Matlab vor und zum Beispiel, um den Code in Matlab zu schreiben, nahm es mich mehr als einen Monat, aber um das Gleiche in KNIME zu tun, nahm mich nur eineinhalb Tage.

Aline: Hat KNIME die Zusammenarbeit zwischen den Forschern in Ihrer Gruppe erleichtert? Wie vergleicht sie mit der kodierenden Forschung?

Malik: Die meisten Biologen sind keine rechnerischen Typen. Zeigen sie tatsächlich Python, R, oder Java-Code wäre schwierig. Allerdings mit KNIME und zeigt ihnen einen Workflow in [KNIME Analytics Plattform](#) [erleichtert die Kommunikation mit ihnen.](#)

Auch, in Bezug auf die Kommunikation mit meinen Studenten, ist es einfacher, Projekte aufzuteilen und Arbeit in diesem Entwicklungsumfeld.

Stefan: Wie funktioniert es mit Ihren Schülern? Wie schnell können sie auf Geschwindigkeit und wie schnell können sie mit KNIME gestartet werden, damit sie wirklich produktiv sein können?

Malik: Es gibt zwei Arten von Studenten, die, die ich lehren und diejenigen, die Ich übernehme. Das erste Mal, dass ich KNIME unterrichtete, wurde ich tatsächlich gebeten, eine Daten zu lehren Wissenschaftslabor und durfte KNIME auf Anfrage vorstellen. Ich habe sofort Das Feedback der Studierenden, dass sie KNIME lieben und es ganz einfach und bequem finden. Normalerweise, Ich löse sie aus, indem ich ihnen eine Aufgabe gebe und sie bitten, es in Java zu tun, was in der Regel erfordert sie, um eine Menge Kesselplatten-Code zu schreiben. Dann zeige ich Ihnen, wie viel einfacher es in KNIME. In der Lehre bin ich in die meisten meiner Kurse integriert, sehr oft neben Python.

Einige der Studenten nehmen sogar KNIME an ihre Unternehmen, fragen ihre Chefs, wenn sie kann mit KNIME arbeiten.

Die Studenten, mit denen ich arbeite, wissen in der Regel nichts über KNIME. Aber es ist sehr leicht für sie zu lernen. Sie lernen von selbst nur mit der freien [Lernmaterialien](#) auf der KNIME-Website.

Stefan: [Mitte Juni veröffentlichte KNIME seine](#) [neueste Softwareversion](#) [\(v. 4.6.0\)](#). Einer der die interessantesten neuen Features ist die Möglichkeit, reine-Python-Knoten zu entwickeln, die kann mit anderen wie jedem anderen Knoten geteilt werden. Wie Sie denken, können Forscher profitieren aus dieser neuen Funktion?

Malik: Ich sah diese Veröffentlichung und ich denke, es ist erstaunlich. Ich denke, es ist ein großer Beitrag zu den Menschen, die KNIME verwenden, dass sie jetzt Knoten in Python schreiben können. Dies gibt auch KNIME Community die Möglichkeit, mehr Knoten beizutragen. Obwohl, die Möglichkeit

Komponenten zu bauen und mit anderen zu teilen war auch transformative Arbeit.

Aber wir sollten R hier nicht vergessen. Einige meiner Studenten kennen R aber nicht kennen Python, andere kennen Python aber nicht R. Ich denke, es wäre gut, wenn wir könnten das gleiche mit R.

Stefan: Lassen Sie uns einen Moment über die Studenten reden. Welche Arten von Studenten kommen zu Ihnen Klassen? Welchen Hintergrund haben sie? Sind sie Studenten oder Studentinnen? Sehen Sie eine Welle von älteren Studenten auf der Suche nach einer zweiten Ausbildung?

Malik: Meine Studenten gehören meistens zur Abteilung Informatik oder Informationssysteme. So haben sie im Allgemeinen einen Programmier-Hintergrund, aber dort ist Vielfalt. Es gibt ein paar Studenten, die nicht so stark in der Programmierung sind, aber muss noch etwas kodieren. Besonders für die Studenten ist KNIME sehr gut Lösung. Es erleichtert ihr Leben. Ich denke, die Lehre KNIME könnte auch eine Einführung in Programmierung statt mit Python oder Java zu starten. KNIME ist ein leistungsstarkes Werkzeug Algorithmen und Programmierung ohne eigentliche Programmierung lehren.

Stefan: Welche Themen decken Sie in Ihren Kursen ab? Was werden die Schüler lernen?

Malik: Ich unterrichte ein paar Kurse, in der Regel fortgeschrittenere Kurse. Da, die Studierende lernen alles: Parameter, Flussvariablen, globale Variablen, Daten sammeln in Schleifen, etc. In einigen der Kurse, nach dem College, soll ich lehren Python. Da ich jedoch ein großer Fan von KNIME bin, versuche ich es immer in meine Lehre zu integrieren. Aber ich werde auch einen neuen Kurs im September unterrichten "Coding without Coding" wo ich KNIME verwenden werde, um Studenten zu lehren, wie man ohne tatsächlich schreiben Code kodieren.

Stefan: Wie würden Sie Ihren Lehransatz beschreiben? Wie war die COVID-19 pandemic ändern Sie Ihren Lehrstil?

Malik: Ich denke, wir haben Vorteile von der globalen Pandemie bekommen. Manchmal, Menschen muss gezwungen werden, Dinge zu tun. Die Pandemie zwang uns, über Zoom oder andere zu lehren Plattformen. Jetzt liebe ich die Lehre über Zoom. Ich denke, es erleichtert die Lehre, für die Lehrer und Professoren, aber auch für die Studenten. Außerdem ist es jetzt viel einfacher zu Aufzeichnungsvorträge und ermöglichen ein viel größeres Publikum, meine Kurse „aufzupassen“, für z.B. Teilzeitstudenten.

Stefan: Jetzt die klassische Frage. Wo sehen Sie Forschung in Bioinformatik in in den nächsten Jahren? Was wird die nächste Innovation sein?

Malik: Zurück in den Tagen war ich unter den ersten Menschen in der Bioinformatik. Die Feld begann sich zu etablieren, indem er Menschen aus Informatik, Math, Statistiken usw. Anschließend begannen sie ein Programm in der Bioinformatik zu produzieren. Jetzt haben wir mehr ausgebildete Menschen im Bioinformatik-Programm. Das Lernen ist organisierter, und wir haben mehr Daten zur Verfügung. Als ich begann, hatte ich nur eine Art von Daten — Gen Ausdrucksdaten. Jetzt können wir sehen, dass Algorithmen Daten von zwei kombinieren

oder mehr -Omk Technologien: Wir sind bereits im Bereich der Multi-omics und wir werden müssen mehr Arbeit für eine bessere Integration von N-Datensätzen tun.

Stefan: Gibt es interessante Projekte oder Konferenzen?

Malik: Die meisten Konferenzen dieses Jahres waren abgelegen. Ich war auf der ISMB Konferenz im Juli und nahm in diesem Jahr an den CAMDA Contest Challenges teil. Im Oktober werde ich an der HIBIT 2022 Konferenz in der Türkei. Ich besuchte die Konferenz bereits vor ein paar Jahren, weshalb ich eine Zusammenarbeit mit einem Kollegen habe aus der Türkei. Aber vor allem bin ich an der University of North Carolina als Gastprofessor Mein Stipendium.

Stefan: Wie können sich Leute vom Publikum mit Ihnen in Verbindung setzen?

Malik: Die Leute können mich entweder über LinkedIn , Mein KNIME Hub , oder GitHub :

Dieser Artikel wurde erstmals in unserem Niedriger Code für Advanced Data Science Journal auf Medium. Die Originalversion finden Hier.

Sehen Sie sich das ursprüngliche Interview mit Malik Yousef auf YouTube an Mein Data Guest – Ep 11 mit Malik Yousef .



[Nick Rivera](#) , aka NickyDee auf YouTube wurde nominiert

KNIME Beitrag des Monats für Februar 2022. Er wurde für seine vielen Video-Tutorials, die ein umfangreiche, organisiert, leicht zu verdauen und nützliche Ressource sowohl für Neulinge als auch für erfahrene KNIME-Nutzer ihr Wissen über KNIME. Dazu gehören Explorationen von Sehnsucht Identifikationen, Schwenken und Entflößen, Aggregationen, bedinger Mathematik-, Datums- und Zeitbetrieb, Excel-Funktion

Übersetzungen und viele weitere Datentransformations- und ETL-Operationen. Auf seine

[YouTube Kanal](#) er deckt die meisten Aspekte der Datenmanipulation mit

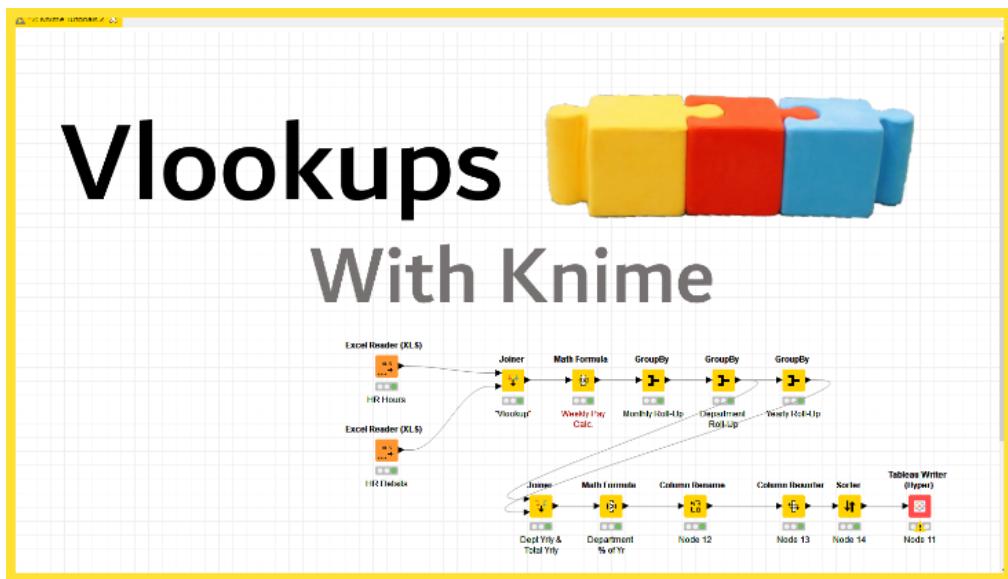
KNIME.

Nick ist ein Data Professional in Business erfahren Modellierung, Reporting und Analyse. Er ist derzeit ein Business Analyst bei der EMR Group. Sein Schwerpunkt ist auf dem Bau von operationellen Modellen, die helfen, die Fähr- und Non-Ferrous-Unternehmen bei EMR. Während Nicks Know-how umfasst ein bisschen von allem, seine bevorzugte Aspekt der Daten ist die Fähigkeit, die es hat, die Entscheidungsfindung beeinflussen.



Wie man ein Excel VLOOKUP in KNIME macht

Autor: Nick Rivera



VLOOKUP in KNIME

Die VLOOKUP ist eine der besten und grundlegendsten Formeln in Excel, und aus gutem Grund auch. Die Fähigkeit, mit (kombin) verwandten Tabellen beizutreten, ist entscheidend, wenn versuchen, Daten aus mehreren Quellen zusammenzubringen, um Daten zu synchronisieren und zu liefern eine succinct Analyse. Keine Sorge mehr, heute werde ich Ihnen zeigen, wie man eine VLOOKUP aus Excel in KNIME.

Geben Sie den Join

Wenn Sie jemals eine SQL-Abfrage geschrieben haben, dann wissen Sie bereits, was ein VLOOKUP ist.
Wenn du es nicht hast, wirst du heute lernen!

Das VLOOKUP in Excel besteht aus 4 Argumenten:

- ANH ANG Der Lookup Wert
- 2. Die Lookup Range
- 3. Der Index der gewünschten Spalte
- Exakte oder annähernde Übereinstimmung

1,347
Version
20.12.20
11, 5, 1,

Um ein VLOOKUP in KNIME durchzuführen, müssen Sie einen Joiner-Knoten verwenden

Die Mitgliednode only 2 "arguments", die Lookup-Spalte(s) und die Spalten, die Sie Ich will vorbeikommen.

Lassen Sie uns zum Beispiel die beiden unten abgebildeten Tabellen verwenden:

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	

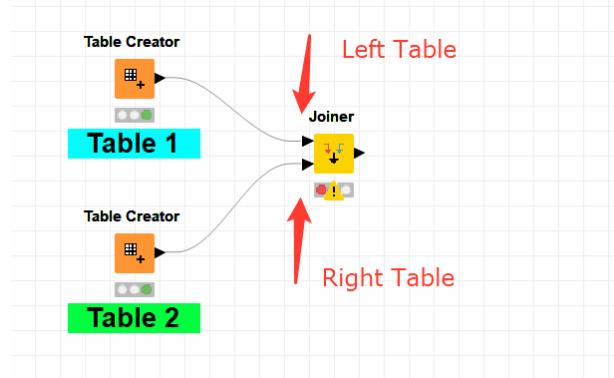
Wir werden VLOOKUP (join) die Bonusrate von der über Tisch zu dieser neuen Tabelle.

Unser Ziel wird VLOOKUP die 2.

Abteilung Bonus Rate

Detail aus Tabelle 1 zur Tabelle

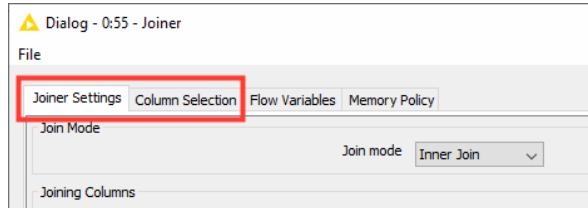
Um ein VLOOKUP durchzuführen, müssen wir die **MitgliedKnotenpunkt**. In Ihrem Node-Repository, **Suche nach “ Mitglied” und dann ziehen und fallen Sie den Knoten in Ihren Arbeitsraum** . Verbindung beide Tabellen, die Sie wollen **VLOOKUP** (aka join) zum **MitgliedKnoten**. Der Tisch, den Sie verbinden mit der oberen Eingabe wird als " links" Beistelltisch, während der Tisch Sie verbinden mit der unteren Eingabe wird als " Recht" Beistelltisch. Das ist wichtig. zu erkennen, weil die **Mitgliednode** gibt Ihnen die Flexibilität, einen linken Join oder einen rechts verbinden, sowie eine innere Verbindung oder eine volle äußere Verbindung. Ich füge mehr Farbe auf den Join-Typen hinzu ein wenig später, aber halten Sie das jetzt auf der Rückseite Ihres Kopfes.



Ein Workflow, der die Verwendung des Joiner-Knotens demonstriert.

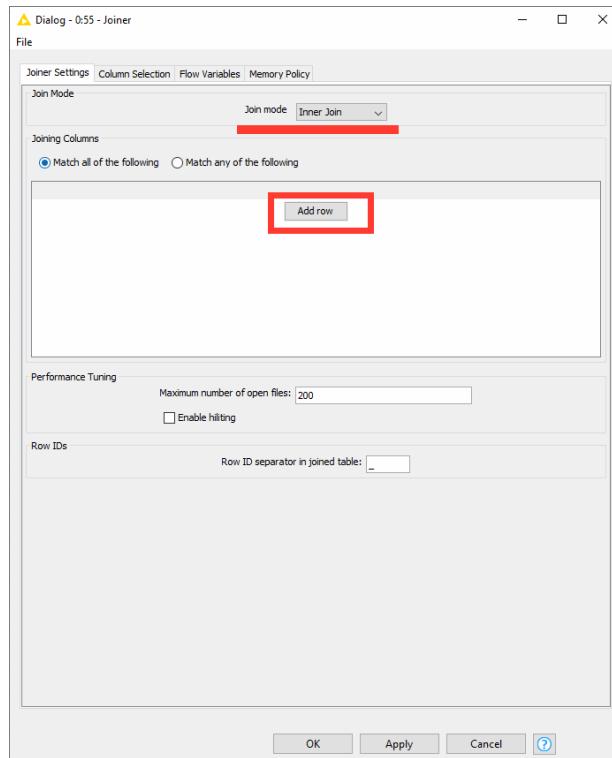
Jetzt, wo Sie die beiden Tische mit den **Mitgliedum** die Konfiguration zu starten. In der Konfiguration sehen Sie: **Einstellungen des Anmelders** und **Auswahl der Spalte** Tabs. Die Registerkarte Joiner Settings ist, wo wir die " **Wert** " Argument, während die Spalte Auswahl Tab ist, wo wir die " **Spaltenindex** " Argument.

MitgliedKnoten, Doppelklick in den **Sehnsucht** und **Sehnsucht**



Die Registerkarten Joiner Settings und Column Selection in den Konfigurationsfenster des Joiner Knotens.

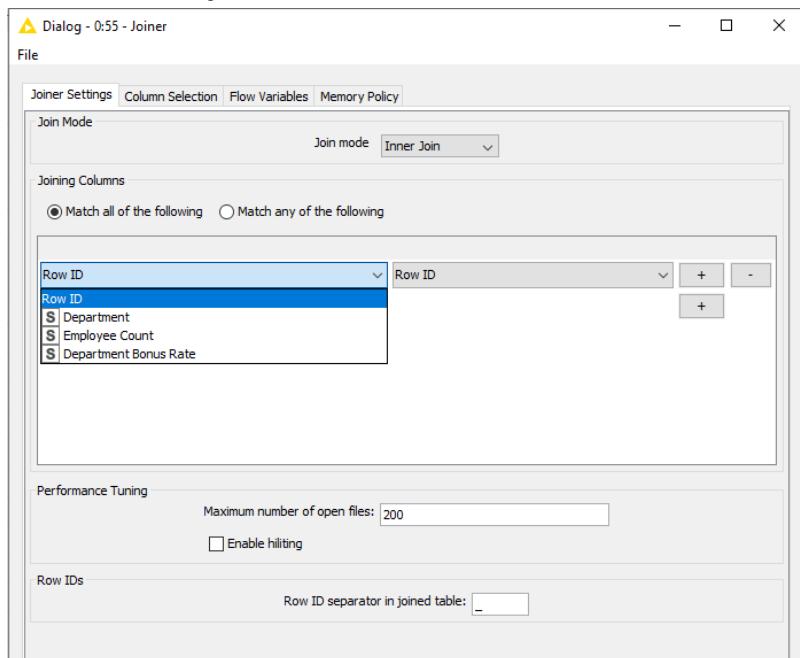
In der „Einstellungen des Anmelders“ Tab müssen wir die Spalte, auf der wir beitreten wollen die beiden Tabellen. Sie können diesen Schritt als die Bereitstellung der Lookup-Wert Argument denken und den Beginn des Lookup-Bereichs. Anders setzen, der Lookup-Wert und der erste Spalte des Lookup-Bereichs sind die Werte, die wir als Fügen bereitstellen müssen Spalten.



Der „Join-Modus“ und „Reihe hinzufügen“ Funktionalität des KNIME Joiner Knoten.

Im Bildschirm oben gedreht, in der roten Box gibt es eine Schaltfläche, die liest "Zeile hinzufügen". Wir werden Klicken Sie dies, um uns eine Zeile zu geben, in der wir die Fügespalten in jeder der beiden Tabellen. Das folgende Bild zeigt das Konfigurationsmenü, nachdem wir auf Row hinzufügen klicken. Wir werden Klicken Sie in die Dropdowns, die markiert sind "Row ID" und dann werden wir beide auf die Fügespalten. Der Dropdown auf der linken Seite entspricht der linken Tabelle aka

Top Input der MitgliedKnoten, während der Dropdown auf der rechten Seite dem entspricht rechte Tabelle aka die untere Eingabe der MitgliedKnoten.



Das Konfigurationsfenster des Joiner Knotens. In der Registerkarte Joiner Settings können Sie die Fügebedingung definieren.

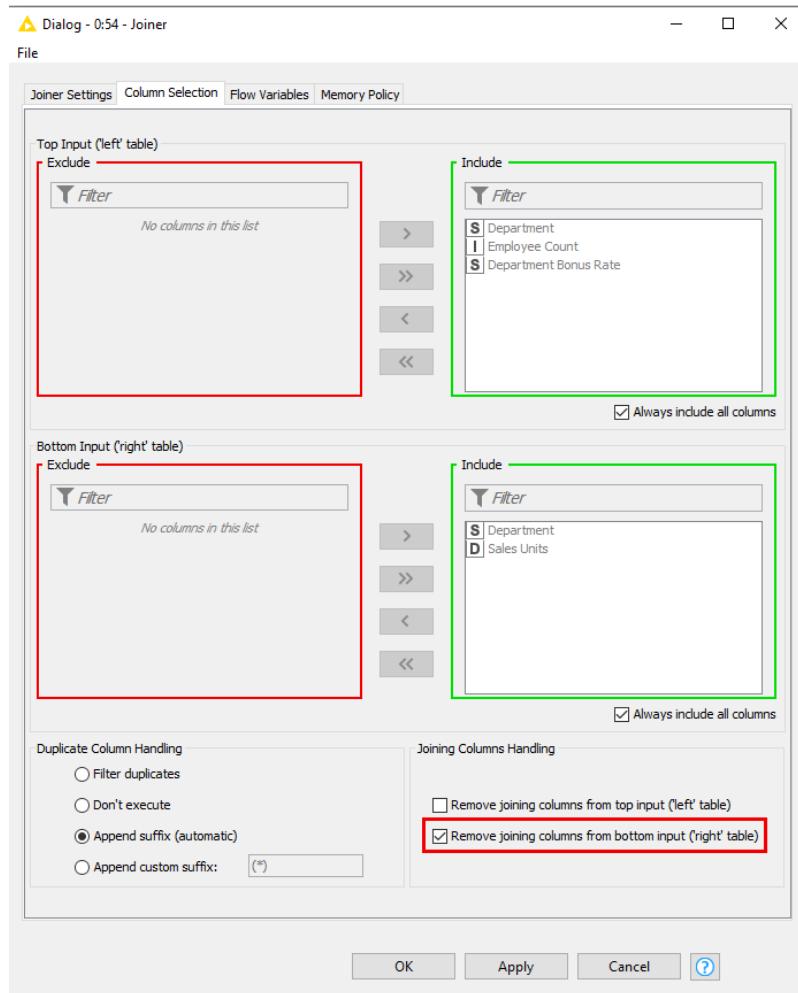
In unserem Beispiel teilen die beiden Tabellen (beide linke und rechte Seite) die Abteilungsspalte, so werden wir unter beiden "Abteilung" auswählen. Jetzt, wo wir die Fügespalten haben delineated, müssen wir die Art von Join oder VLOOKUP wählen, die wir ausführen möchten. Die Optionen unter Teilnahme Modus sind wie folgt:

- ANH
ANG Anmelden: Die letzte Ausgabe hier ist eine neue Tabelle, in der sowohl die linke als auch die rechte Seite Tabellen teilen ähnliche übereinstimmende Zeilen
2. Links beitreten: Hier verbinden Sie Details vom rechten Seitentisch zur LEFT-Seite Tisch - denken Sie an den LEFT-Seitentisch als Ihren Haupttisch
 3. Registrieren Sie sich: Hier verbinden Sie Details vom linken Seitentisch zur rechten Seite Tisch - denken Sie an die RECHT-Seitentabelle als Ihren Haupttisch
- 1.347
100%
20.12.20
13,5 13 Vollständiges Outer Join: Hier treten Sie beide Tische gegeneinander an, unabhängig von deren übereinstimmende Lookup-Spalten oder nicht; daher werden einige Spalten/Reihen null/missing Werte anzeigen

Was hier offensichtlich ist, ist die Flexibilität, die innerhalb des Joiner-Knotens zur Verfügung steht, die nicht wirklich in einem Excel VLOOKUP erhältlich. Wir werden nicht aus dem Weg, um dieses Detail zu überlegen aber, da das Ziel dieses Beitrags ist, ein schnelles DIY der Art sein. Ich werde aufschreiben separater Beitrag, der die Profis des Joins über einen VLOOKUP zeigt, werde ich diesen Beitrag hier einmal verbinden Es ist bereit, also seien Sie auf der Suche.

Unser ursprüngliches Ziel war es, sich an/überholen Sie die Dezernale Bonusrate von Tabelle 1 bis Tabelle 2. Angesichts dieses Ziels werden wir mit einem richtigen Join laufen. Der Gedanke hier ist, dass Tisch auf der rechten Seite ist unsere primäre Analysetabelle, also kommen wir zusammen Details [von der linken Tabelle] zur rechten Tabelle. Lass mich wissen, ob du das nicht befolgst!

Jetzt, da wir die Fügespalten sortiert haben, können wir weitermachen, um die Auswahl Spalten, die wir über die aka Argument 3 eines VLOOKUP bringen wollen. In der Abbildung unten ist die Auswahl der Spalte Tab. Die Konfiguration hier ist ziemlich einfach, wir werden Wählen Sie aus, welche Spalten aus welcher Tabelle wir enthalten möchten & aus unserer Ausgabe Tisch.



Der Joiner-Knoten von KNIME ermöglicht Flexibilität in der Spaltenauswahl von links und rechts Tabellen.

Die Tische, die wir verbinden, sind klein und begrenzt in der Spaltenzahl, also brauchen wir nicht wirklich zu begrenzen, welche Spalten wir von der linken oder der rechten Tabelle überführen wollen. Es kann sein Fälle, in denen Sie Details von zwei Tabellen, die jede mehrere

Spalten verschiedener Details. Sie brauchen vielleicht nicht jedes einzelne Detail von beiden Tabellen, so in diesen Fällen haben Sie die Fähigkeit, zu begrenzen, welche Spalten von jeder Tabelle kommen.

Um Spalten einzuschließen und/oder auszuschließen, die wir im Joiner-Ausgang wünschen, deaktivieren Sie einfach die Box unter jedem Include-Bildschirm, die "Always alle Spalten enthalten".

Unchecking dies ermöglicht es uns, Spalten von der Include-Seite auf die

Ausschließen Sie Seite und umgekehrt. Auch hier ist dies ein weiterer Fall zusätzlicher Flexibilität, die ein Zusammenschluss Angebote, die ein excel VLOOKUP nicht anbietet!

Die " Immer alle Spalten einschließen " Kontrollkästchen ist eine Einstellung, die standardmäßig überprüft wird. Eine weitere Standardeinstellung, die Sie hier beachten müssen, ist, dass die Fügespalten von die untere Tabelle wird zugunsten der Spalten aus der linken Tabelle fallen gelassen. Das ist nicht wirklich ein riesiges Problem in den meisten Fällen, aber es ist etwas zu beachten, wenn Sie mehr mit unterschiedlichen Tabellen. Sie können diese Standardeinstellung einfach überschreiben, indem Sie die Einstellung deaktivieren die Box unten abgebildet. Wenn Sie dieses Kästchen deaktivieren und auch den obenstehenden verlassen unkontrolliert, dann erhalten Sie Duplikate der Fügespalten. Links vom Roten Box in der Abbildung unten sind ein paar Optionen, um die Benennung des Duplikats zu bearbeiten Spalten. Ich benutze gerne das " Hinzufügen von benutzerdefinierten Suffix Option in Fällen, in denen ich suche Berechnung Produktmix oder Marktanteilsrufe.

Nachdem wir die Spalten ausgewählt haben, die wir in unserer Endausgabe wünschen, können wir uns bewerben, ok, und den Knoten ausführen. Die Ergebnisse sehen so aus:

Row ID	Depart...	Employee Count	Department Bonus Rate	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

Diese Tabelle hat die Ausgabe eines Excel VLOOKUP in KNIME!

Tabelle 2 hat nun das Detail der Dekrete Bonusrate aus Tabelle 1. Seit wir nicht alle Spalten ausschließen, wir brachten auch die Spalte Mitarbeiterzähler von links Tisch auf den rechten Tisch – Sie müssen diese Flexibilität lieben!

Dieser Artikel wurde ursprünglich veröffentlicht [Nicks Blog](#). Die Originalversion finden Sie hier [Hier](#).

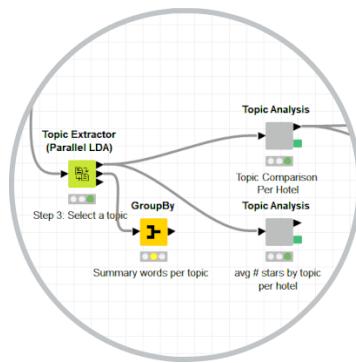
Im Falle von Fragen oder wenn Sie mit Nick in Kontakt kommen möchten, ist er durch [Twitter](#).



Francisco Villarroel Ordens wurde nominiert
Beitrag des Monats für April 2022. Er wurde vergeben
für die Entwicklung gemeinsam mit einem Team bei KNIME a live
Projektarchiv für maschinelles Lernen in Marketing Analytics auf
Die KNIME-Gemeinschaft Hub mit wiederverwendbaren Lösungen für
Kundenbetreuung, Stimmungsanalyse, automatisiertes Bild
Analyse, SEO & CX. Eine umfangreiche Analyse dieses Projekts
in seiner Forschungspapier „Das Bild rechts“
einen Teil der Thema Modellierung des Workflow im Projektarchiv vorgesehen.

Francisco ist derzeit Professor für Marketing und Direktor der MSc. in Marketing-Programm an der LUISS Guido Carli University in Rom, Italien. Er ist ein starker Vertreter von KNIME in seinen Vorträgen, und er verwendet es in seinen Klassen und Forschungsarbeiten umfassend über Marketinganalytik. Seine Interessengebiete sind Analytics, Digital Marketing und Service Forschung, und er ist leidenschaftlich für Verbraucher-Markenkommunikation und die Verwendung von NLP an Online-Konversationen verstehen.

Besuchen Sie Franciscos Raum auf dem KNIME Hub oder
Profil im KNIME Forum (Hub/Forum Griff:
Fvillarroel)



Machine Learning in Marketing Analytics

Marketing Analytics Lösungen auf dem KNIME Community Hub

Autoren: Francisco Villarroel Ordenes & Rosaria Silipo

Viele Unternehmen erweitern derzeit ihre Einführung von Data Science-Techniken auf umfassen maschinelles Lernen. Marketing Analytics ist eine davon. Alles kann reduziert werden zu Zahlen, einschließlich Kundenverhalten und Farbwahrnehmung, und daher alles kann analysiert, modelliert und vorhergesagt werden.

Marketinganalytik umfasst bereits eine breite Palette von Datenerhebungen und Transformationstechniken. Social Media und web-getriebenes Marketing haben eine große schieben in die Digitalisierung des Raumes; Zählen der Anzahl der Besuche, die Anzahl der mag, die Minuten der Betrachtung, die Anzahl der zurückkehrenden Kunden, und so weiter ist häufig Praxis. Wir können jedoch eine Ebene nach oben bewegen und maschinelles Lernen und Statistiken anwenden. Algorithmen zu den verfügbaren Daten, um ein besseres Bild von nicht nur der aktuellen, sondern auch die zukünftige Situation.

The screenshot shows the KNIME Community Hub interface. At the top, there is a navigation bar with the KNIME logo, a search bar containing 'Search workflows, nodes and more...', and various user icons. Below the header, a breadcrumb navigation shows the path: KNIME Community Hub > knime > Spaces > Machine Learning and Marketing. The main content area is titled 'Machine Learning and Marketing' and includes a 'Public space' indicator, a user count of 39, and a 'Copy link' button. A sidebar on the left lists categories such as Consumer Behavior, Consumer Mindset Metrics, Customer Valuation, Data Protection and Privacy, Marketing Mix, Other Analytics, and Segmentation and Personalization, each with a downward arrow icon. At the bottom, a footer states: 'Das öffentliche Workflow-Repository für Marketinganalyselösungen auf dem KNIME Community Hub.'

Marketer können auf maschinellen Lerntechniken Kapitalisieren, um große Datensätze zu analysieren Muster identifizieren oder prognostizierende Analysen durchführen. Beispiele sind die Analyse sozialer Medienbeiträge zu sehen, was Kunden sagen, Bilder zu analysieren, um Einblick in Bilder und Videos, oder vorhersagen Kunden churn – nur drei zu nennen.

In der [maschinelles Lernen und Marketing](#) Raum auf dem KNIME Community Hub, werden Sie finden Sie eine Reihe von Fallstudien Anwendung von maschinellen Lernalgorithmen auf Klassik Marketingprobleme.

In diesem Beitrag werden wir diese Fallstudien beschreiben, eine nach dem anderen, die Besonderheiten zeigen jeder von ihnen und die Einsichten, die sie bringen. Bisher haben wir Lösungen für:

- [Vorhersage des Kunden Churn](#)
- [Messung Bewertungsanalyse in Social Media](#)
- [Bewertung Customer Experience durch Themenmodelle](#)
- [Content Marketing und Image Mining](#)
- [Schlagwortforschung für Suchmaschinenoptimierung](#)

Wir werden dieses Repository weiterhin beibehalten, indem wir die bestehenden Workflows aktualisieren und jedes Mal, wenn eine Lösung aus einem neuen Projekt zur Verfügung steht, neue hinzufügt.

Anmerkung: Dieses Lösungs-Repository wurde von ein gemischtes Team von KNIME-Nutzern und Marketingexperten des KNIME Evangelism Team in Constance (Deutschland), geleitet von [Rosen und Rosen](#), und [Francisco Villaroel Orden](#), Professor für Marketing an der LUISS Guido Carli Universität in Rom (Italien).

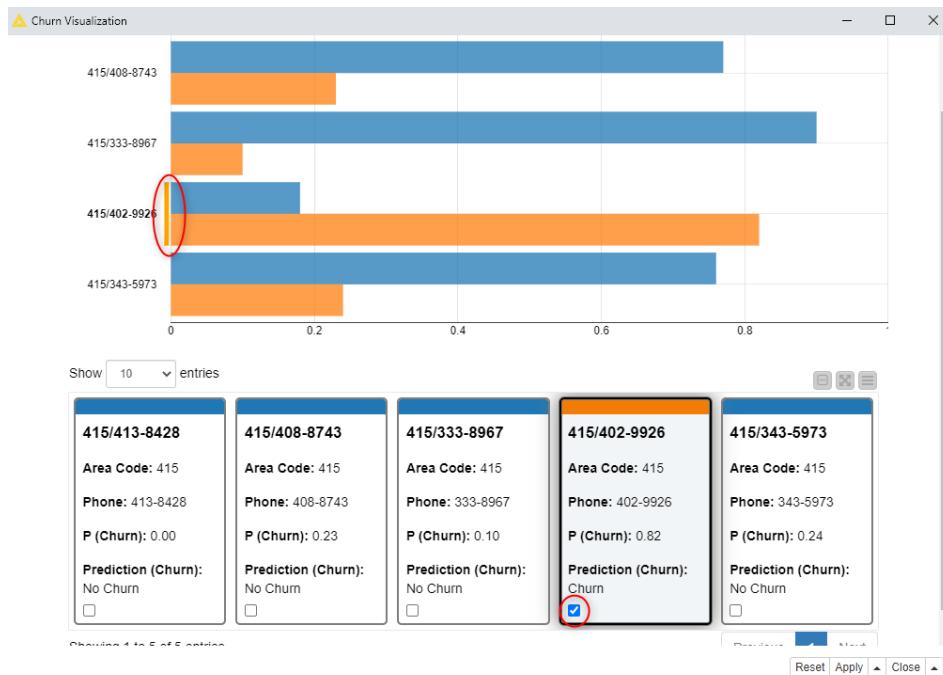
Vorhersage des Kunden Churn

Nutzung bestehender Kundendaten (z.B. Transaktions-, Psychografik-, Attitudinal-) vorausschauende churn-Modelle zielen darauf ab, Kunden zu klassifizieren, die geschürt oder geblieben sind, wie sowie schätzen die Wahrscheinlichkeit neuer Kunden zu churn, alle in einem automatisierten Prozess. Ist die churn Wahrscheinlichkeit sehr hoch und der Kunde ist wertvoll, die Firma Vielleicht wollen Maßnahmen ergreifen, um diesen Mist zu verhindern.

Der Ordner „Consumer Behavior“ > „Churn Prediction“ im Ordner „Consumer Behavior“ [Maschinenlernen und Marketing](#) Raum für die KNIME Community Hub beinhaltet:

- Ein Workflow [Schulung eines ML-Klassifikator](#) (in diesem Fall ein zufälliger Wald) zu unterscheiden Kunden, die geschürt haben und Kunden, die im Training geblieben sind.

- A Bereitstellung von WorkflowAnwendung des zuvor geschulten Modells auf neue Kunden.
Abschätzen ihrer aktuellen Wahrscheinlichkeit, zu churn, und das Ergebnis auf einer einfachen Dashboard (siehe Abbildung unten).



Das Dashboard meldet das Churn-Risiko in Orange für alle neuen Kunden.

Bewertungsanalyse

Sentiment ist eine weitere beliebte Metrik, die im Marketing verwendet wird, um die Reaktionen der Benutzer und Kunden zu einer bestimmten Initiative, Produkt, Veranstaltung, etc. Nach der Popularität von diesem Thema haben wir einige Lösungen für die Umsetzung eines Gefühls gewidmet evaluator für Textdokumente. Solche Lösungen sind im „Consumer Mindset“ enthalten.

Metriken“ > „Sentiment Analysis“ Ordner. Alle Lösungen konzentrieren sich auf drei Stimmungsklassen: positiv, negativ und neutral.

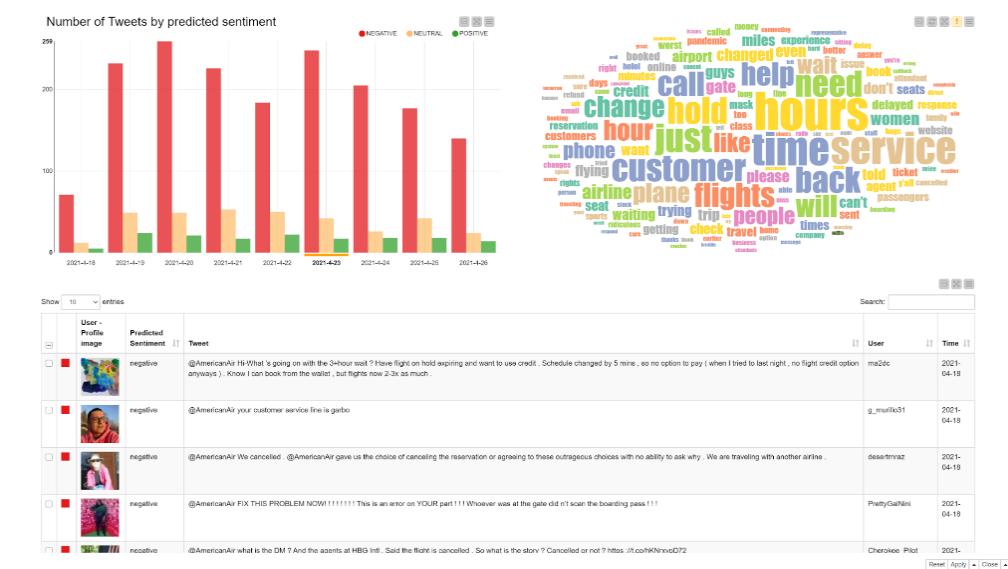
Es gibt zwei Hauptansätze für das Gefühlsproblem:

- **Lexikon basiert.** Hier eine Liste positiver und negativer Wörter (Diktioneare), in Bezug auf die Korpus-Themen, werden zusammengestellt und Grammatik-Regeln gelten auf die Polarität eines bestimmten Textes abschätzen. Erfahren Sie, wie Sie eine Einschätzungsverhersage erstellen Verwendunglexikonbasierte Stimmungsanalyse
- **Machine Learning basiert.** Die Lösungen hier verlassen sich auf keine Regeln, sondern auf die Maschine Lernmodelle. Überarbeitete Modelle werden ausgebildet, um zwischen Negativ zu unterscheiden, positive und neutrale Texte und dann auf neue Texte angewendet, um ihre Polarität zu schätzen.

Machine-learning-basierte Ansätze sind immer beliebter geworden, vor allem wegen ihrer Fähigkeit, alle Grammatikregeln zu umgehen, die eine harte Kodierung benötigen würden. Unter den maschinellen Lernlösungen sind einige Möglichkeiten möglich:

- **Traditionelle maschinelle Lernalgorithmen.** In diesem Fall werden Texte in numerische Vektoren, wobei jede Einheit die Anwesenheit/Absenz oder die Frequenz eines bestimmten Wortes aus dem Korpus Wörterbuch. Danach, traditionell Maschinenlernalgorithmen, wie zufälliger Wald, Support Vector Machine, oder Logistische Regression kann verwendet werden, um die Textpolarität zu klassifizieren. Beachten Sie, dass Eine Vektorisierung wird die Reihenfolge des Wortes im Text nicht erhalten. Weiterlesen [mehr in diesem Tutorial für maschinelles Lernen für die Sentimentanalyse](#) :-
- **Deep Learning basiert.** Deep Learning basierte Lösungen werden immer mehr beliebter für die Stimmungsanalyse, da einige tiefe Lernarchitekturen den Wortkontext (d.h. die Sequenzgeschichte) für eine bessere Stimmung nutzen Schätzung. In diesem Fall werden Texte in Vektoren ein-Hot kodiert, die Sequenz von Solche Vektoren werden dem neuronalen Netz vorgelegt, und das Netzwerk wird ausgebildet, die Textpolarität erkennen. Oft umfasst die Architektur des neuronalen Netzes eine Schicht von Long Short Term Speichereinheiten (LSTM), da LSTM die Aufgabe erfüllt unter Berücksichtigung der Reihenfolge der Erscheinung der Eingabevektoren (die Wörter), [das ist, indem man den Wortkontext berücksichtigt. Erkunden Sie ein Tutorial, um ein tief Lernansatz für die Einschätzungsanalyse](#) :-
- **Sprachmodelle.** Sie werden auch als tiefe kontextualisierte Sprache bezeichnet Modelle, weil sie die kontextabhängige Bedeutung von Wörtern widerspiegeln. hat wurde argumentiert, dass diese Methoden effizienter sind als neurale Netzwerke, da sie eine parallelisierte Kodierung (nicht sequentiell) erlauben Wort- und Unterwort-Token, die auf ihren Kontext hindeuten. Neues Sprachmodell Algorithmen sind ULMFiT, BERT, RoBERTa, XLNet. Im Machine Learning Repository, wir bieten eine unkomplizierte Implementierung von BERT. Wie man benutzt [BERT mit KNIME in diesem Stimmungsanalyse-Tutorial.](#)

Ein Beispiel für all diese Lösungsgruppen finden Sie im „Consumer Mindset Metrics“ > Ordner "Sentiment Analysis" in der [maschinelles Lernen und Marketing](#) Raum.



Visualisierung von Tweets mit geschätzter Stimmung (rot = negativ, grün = positiv, leicht orange = neutral).

Thema Erkennung und Customer Experience

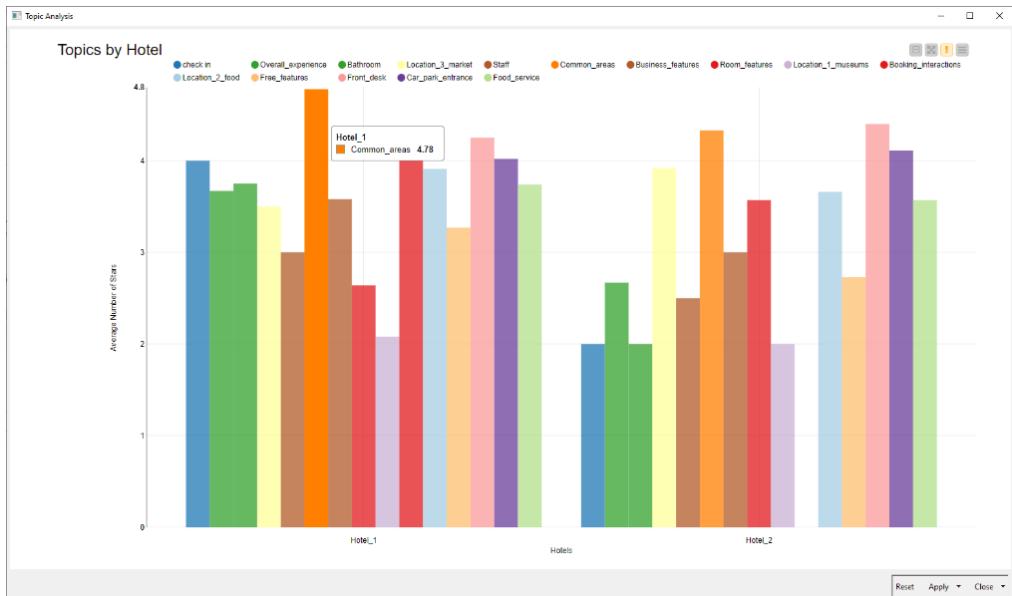
Das Customer Experience Management und die Customer Journey sind eines der am meisten populären Marketingthemen in der Marketingbranche. Viele Informationen über Kundenerfahrung kommt aus den Bewertungen und Feedback und/oder aus dem Star-Ranking-Systemen auf Websites und sozialen Medien.

Die Popularität von Themenmodellen hat zu einer kontinuierlichen Entwicklung von Algorithmen geführt wie Latent Dirichlet Allocation (LDA), Cor related Topic Models (CTM), und Strukturelle Themenmodelle (STM), unter anderem, alle bereits in [forschung. LDA ist in der KNIME Textverarbeitung Verlängerung als KNIME native Knoten](#). Der LDA-Knoten erfasst m-Themen im gesamten Korpus und beschreibt jede von ihnen mit n Keywords, m und n ist ein Teil der Parameter erforderlich um den Algorithmus auszuführen.

Sie finden einen Beispiel-Workflow und zeigen die Nützlichkeit, Themen in Bewertungen, im Ordner „Consumer Mindset Metrics“ > „CX und Topic Models“ im [maschinelles Lernen und Marketing Raum](#).

Der Workflow extrahiert Themen aus Bewertungen mit dem LDA-Algorithmus. Danach schätzt die Bedeutung jedes Themas über die Koeffizienten einer linearen Regression – mit einem KNIME nativen Knoten implementiert – und über die Koeffizienten eines Polynoms regression – in einem R-Skript im KNIME-Workflow implementiert. Es zeigt dann die durchschnittliche Anzahl der Sterne für alle Themen aus den Bewertungen für zwei verschiedene Hotels (siehe Abbildung unten).

Im Bar-Diagramm zum Beispiel, können wir sehen, dass für das Hotel 2 das Thema „Buchung Interaktionen“ wird nie erwähnt. Wir können auch bemerken, dass während Hotel 1 tolle Bewertungen für die „gemeinsame Bereiche“, das Hotel 2 zeichnet sich durch den „Front Schreibtisch“ aus.



Durchschnittliche Anzahl der Sterne nach Bewertungen um eines der 15 erfassten Themen.

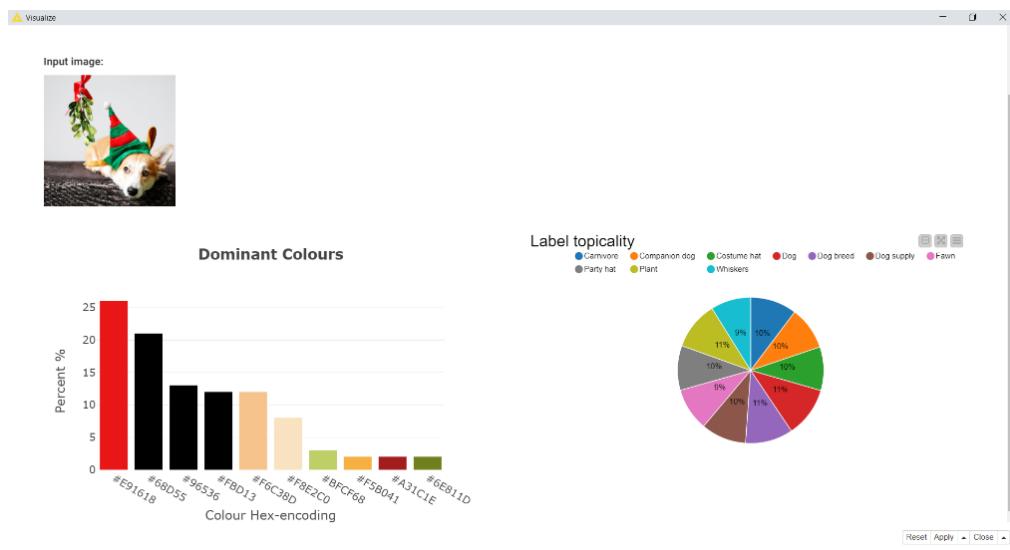
Content Marketing und Image Mining

Die letzten zehn Jahre haben ein exponentielles Wachstum der visuellen Daten einschließlich der Bilder gezeigt und Videos. Dieses Wachstum hat zu einer zunehmenden Entwicklung von Technologien geführt relevante Einblicke aus Bildern zu klassifizieren und zu extrahieren. Dieses Phänomen hatte Auswirkungen auch im Marketing. Da sowohl Verbraucher als auch Unternehmen sich mehr auf Bildmaterial verlassen und Videos zur Kommunikation, Forscher benötigen neue Prozesse und Methoden, um diese zu analysieren Art der Daten.

Das größere Interesse an der Analyse von Visualisierungen und deren Auswirkungen auf die feste Leistung, motiviert uns, einen Workflow zu entwickeln, der bei der Analyse von visuellen Inhalten helfen kann. Der Workflow nutzt die Dienste von Google Cloud Vision (über POST erreichbar) Anforderung), Etiketten (z.B. Menschen) zu erkennen und nuancierte Bildeigenschaften zu extrahieren als Farbkonzentration.

Ein zweiter Workflow nutzt Deep Learning Convolutional Neural Networks zur Klassifikation Bilder von Katzen gegen Hunde. Änderung des Bilddatensatzes und entsprechende Einstellung das Netzwerk, ermöglicht es Ihnen, jede andere Bildklassifikation Task implementieren.

Finden Sie beide Workflows im Ordner „Andere Analytics“ > „Bildanalyse“ der Lernen und Marketing Raum. Die folgende Abbildung zeigt das Ergebnis aus der Analyse eines Bildes über Google Cloud Vision Services.



Analyse des Bildes in der oberen linken Ecke durch Google Vision-Dienste.

Schlagwortforschung für SEO

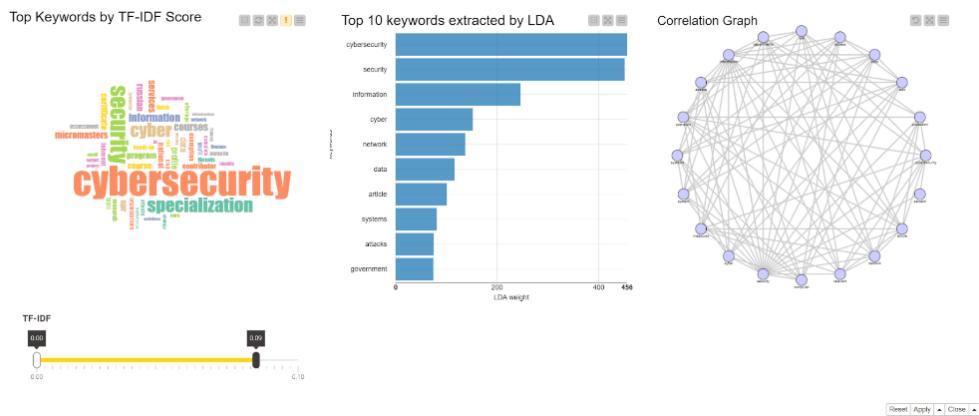
Es ist bekannt, dass Suchmaschinen Webseiten nach dem Vorhandensein von bestimmten Keywords oder Gruppen von Schlüsselwörtern konzeptuell und/oder semantisch verwandt. Außerdem, Schlüsselwörter sollten vom spezialisierten Linguo von Experten sowie von der Gesprächssprache von Neophyten. Beliebte Quellen für solche Keywords sind SERP (Search Engine Result Pages) sowie Social Media.

Im Ordner „Consumer Mindset Metrics“ > „SEO“ der [Maschinenlernen und Marketing Raum](#), Sie finden einen Workflow für semantische Schlüsselwortforschung, implementiert nach dem Artikel „[Semantische Schlüsselwortforschung mit KNIME und Social Media Daten Mining – #BrightonSEO 2015](#)“ geschrieben von Shapiro in 2015.

Der obere Zweig des Workflows verbindet Twitter und extrahiert die neuesten Tweets um einen ausgewählten Hashtag. Der untere Zweig verbindet sich mit der Google Analytics API und extrahiert SERPs um einen bestimmten Suchbegriff. Danach werden URLs isoliert, Webseiten über GET Anfragen an Boilpipe API abgestreift, und Schlüsselwörter werden zusammen mit ihre Frequenzen.

Schlagwörter als: Einzelbegriffe mit höchstem TF-IDF-Score; Mitbewerbsbedingungen mit höchstem co-occurring-Frequenz; Keywords mit höchstem Score aus Themen, die über die Latent Dirichlet Allocation (LDA) Algorithmus.

Als Beispiel haben wir nach Tweets und Google SERPs gesucht „Cybersecurity“. In der folgenden Abbildung werden die Ergebnisse der co-occurring Keywords in der Wortwolke angezeigt. wenn Sie arbeiten im Bereich der Cybersicherheit, dann auch diese Wörter in Ihrem Web Seite sollte Ihre Seite Ranking erhöhen.



Analyse des Bildes rechts durch Google Vision-Dienste.

Erkunden Sie Machine Learning und Marketing Beispiele mit KNIME

Mit diesem Artikel wollten wir die Verfügbarkeit eines öffentlichen Repositorys auf dem [KNIME Community Hub](#), genannt „[maschinelles Lernen und Marketing](#)“, für Marketing Analysten. Ein gemischtes Team von KNIME-Nutzern aus Industrie und Wissenschaft hat geschaffen, einige maschinenlernende Lösungen für einige entwickelt und gepflegt häufig verwendete interessante Anwendungsfälle in der Marketing-Analyse: churn Vorhersage, sentimentanalytik, top-detektion, um kundenerfahrung, bildbergbau, und Keyword-Forschung für Suchmaschinenoptimierung.

Alle Workflows sind kostenlos verfügbar. Sie stellen eine erste Skizze dar, um das Problem zu lösen aber kann natürlich heruntergeladen und nach Ihrem eigenen Geschäft angepasst werden Anforderungen und Daten.

Dieser Artikel wurde erstmals auf unserer Website veröffentlicht [KNIME Blog](#). Die Originalversion finden [Hier](#).

Die in diesem Blog-Post beschriebenen Workflows sind in [KNIMEs maschinelles Lernen und Marketingraum](#) über den KNIME Community Hub.

[KNIME herunterladen Plattform für die Analyse](#) [Hier](#).

Wissenschaftlicher Artikel: F. Villarroel Ordenes & R. Silipo, „Machine-Learning für Marketing über den KNIME Community Hub: Die Entwicklung eines Live-Repository für Marketing Anwendungen, Amtsblatt von Unternehmen Forschung 137(1):393-410, DOI: [10.1016/j.jbusres.2021.08.036](https://doi.org/10.1016/j.jbusres.2021.08.036)



[Christophe Molina](#) wurde nominiert Beitrag des

Monat für Juni 2022. Er wurde für seine Tätigkeit ausgezeichnet innerhalb der KNIME-Gemeinschaft, einschließlich seiner Präsenz auf das KNIME-Forum und Sprecher bei unseren Veranstaltungen. Christophe ist vor allem in der KNIME-Gemeinschaft bekannt seine aktive und resolute Präsenz auf dem Forum, wo er proaktiv unterstützt Benutzer. Darüber hinaus ist er mitautorisiert mehrere wissenschaftliche Beiträge zu verschiedenen QSAR-Themen, wo er

hat KNIME verwendet. Einige davon umfassen: [ADME-Prädiktion mit KNIME: Vorhersage](#)

[Human Oral Bioverfügbarkeit](#)

[ADME Vorhersage mit KNIME: In Silico Aqueous](#)

[Löslichkeit](#) [Automatisiertes Framework für QSAR-Modellierung von hoch ausbalancierten Daten](#)

[ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubilität](#)

[Herausforderung](#) [Isometric Stratified Ensembles: Adaptive Applicability Domain und](#)

[Consensus Klassifizierung der kolloidalen Aggregation](#)

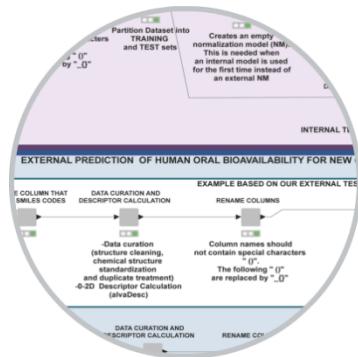
[Das Bild rechts zeigt eine](#)

Schnipsel seines Löslichkeits-Workflows.

Christophe ist ein Data Analyst mit einer tiefen Forschung Erfahrung in pharmazeutisch verwandten Bereichen wie Cheminformatik, Bioinformatik und Genomik und hat mehr als 20 Jahre Berufserfahrung in Data Analytics. Er ist jetzt Freelance Data Analyst und CEO bei PIKÁÍROS, einem privat geführten Data Analytics Unternehmen.

Besuchen Sie Christophe's [Profil im KNIME Forum](#)

(Forumgriff: aworker)



ADME Vorhersage mit KNIME: Eine Retrospektive Beitrag zur zweiten „Solubility Challenge“

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

Anmerkung des Herausgebers:

Dieser Artikel entspricht der Veröffentlichung von Falcón-Cano G, Molina C und Cabrera-Pérez
MÁ: ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubilität
Herausforderung“. ADMET & DMPK. 2021 Jul; 9(3):209-218. <http://dx.doi.org/10.5599/admet.979>

Einleitung

Pharmakokinetische Parameter werden in der Regel durch eine Kombination verschiedener physikochemische Eigenschaften. Unter diesen hat die Löslichkeit eine sehr wichtige Rolle aufgrund seines Einflusses auf den Absorptionsprozess. Die Notwendigkeit, die Löslichkeit auszugleichen, Vermeidung von Überschüssen oder Insuffizienz, ist eine Herausforderung aus der Perspektive der Drogenentdeckung.

In diesem Zusammenhang wurden mehrere Forschungsanstrengungen unternommen, um eine genaue Vorhersage zu ermöglichen. wässrige Löslichkeit durch Quantitative Struktur-Property Relation (QSPR) Ansätze. Zweifellos, die erste und zweite “Solubility Challenges” vorgeschlagen von Llinas et al. waren ein sehr effektiver Indikator für den Fortschritt und den Stand der Technik Löslichkeitsschätzung [1, 2]. Vor kurzem, Llinas et al. die Ergebnisse der zweite „Solubility Challenge“ zur Analyse der Entwicklung der Rechenmethoden verwendet in dieser Vorhersageaufgabe und dem Einfluss der Datenqualität auf die Ergebnisse [3].

In unserer vorherigen Veröffentlichung haben wir eine neue Methode vorgestellt, die auf rekursivem Zufall basiert. Waldansätze zur Vorhersage wässriger Löslichkeitswerte von medikamentösen und medikamentösen Molekülen [4]. Es basierte auf der Entwicklung von zwei neuartigen rekursiven Maschinenlernen Ansätze zur Datenreinigung und zur variablen Auswahl und ein Konsensmodell erzeugt durch die Kombination von Regressions- und Klassifikationsalgorithmen. Dieses Modell im Vergleich zu vielen Modellen eine gute Löslichkeit vorhersage liefern konnte in der Literatur beschrieben. In der Erwägung, dass unser Modell aus einer Datenbank entwickelt wurde von wässrigen Löslichkeitswerten mit begrenzter Information über die Versuchsbedingungen könnte unser Modell die intrinsische Löslichkeit erfolgreich vorhersagen Werte der beiden Drogensätze, die in der zweiten „Solubility Challenge“ verwendet werden?

Die vorliegende Studie beschreibt die Leistung unseres Modells mit den Molekülen der zweite „Solubility Challenge“ und der Vergleich der Ergebnisse mit den erhaltenen mit den besten Performance-Modellen des Wettbewerbs. Es ist notwendig zu klären, daß diese Aufgabe, das Modell wurde nicht geschult, retrainiert oder optimiert basierend auf den Molekülen die Herausforderungstests, d.h. die Modellparameter oder Hyperparameter blieben exakt das gleiche wie die zuvor veröffentlichten Arbeiten [4].

Materialien und Methoden

Herausforderungen

Die zweite "Solubility Challenge" bestand darin, die intrinsische Löslichkeit zu bewerten Schätzung von zwei Drogensätzen. Der erste Satz besteht aus 100 Drogen mit einem Durchschnitt interlaborative Standardabweichung geschätzt von ~0.17 Log-Einheiten. Der zweite Testsatz besteht aus 32 "schwierigen" Medikamenten, gekennzeichnet durch schlechte interlaborative Reproduzierbarkeit: Standard Deviation ~0.62 Log Units. Eine detaillierte Liste dieser Moleküle wurde gezeigt in einem vorherigen Papier [3].

Software

Das Konstanz Information Miner (KNIME) ist ein kostenloses und öffentliches Software-Tool, das zu einer der wichtigsten analytischen Plattformen für Innovation, Data Mining und Maschine Lernen. Die Flexibilität der in KNIME entwickelten Workflows mit unterschiedlichen Werkzeugen ermöglicht Benutzern zu lesen, zu erstellen, zu bearbeiten, zu trainieren und zu testen Maschine Lernmodelle, stark die Automatisierung von Vorhersagen und Anwendung durch jeden Benutzer zu erleichtern [5,6]. In dieser Studie wir nutzten die Open-Source-Software KNIME Analytical Platform Version 4.0.2 [7] und ihre kostenlose komplementäre Erweiterungen für Transformation, Analyse, Modellierung, Daten Visualisierung und Datenvorhersage. Zur Erzeugung von molekularen Deskriptoren aus Strukturen, der „Descriptor“-Knoten aus „alvaDesc“-Erweiterung [8] und der „RDKit Es wurden Deskriptor-Knoten [9] eingesetzt.

Modellierungsdatensatz

Um die Moleküle der zweiten "Solubility Challenge" vorherzusagen, haben wir als Training eingesetzt die in unserem vorhergehenden Papier veröffentlichte kuratierte Menge an wässriger Löslichkeit eingestellt. Dieses Set besteht aus zwei großen wässrigen Löslichkeitsdatenbanken [10, 11]. Für jedes Molekül nehmen den Code SMILES (Simplified Molecular Input Line Entry Specification) als Eingabeformat, Es wurde ein Strukturreinigungs-, Standardisierungs- und Duplikatentfernungsprotokoll entwickelt. Der Code InChi (IUPAC International Chemical Identifier) wurde für Duplikat verwendet Identifizierung und Standardabweichung der experimentellen Messungen berechnet. Eine ausführliche Beschreibung dieses Verfahrens ist in unserem vorherigen Artikel [4]. Obwohl die Hypothese, dass - die Qualität der experimentellen Daten ist die wichtigste Grenzfaktor bei der Vorhersage wässriger Löslichkeit - wurde herausgefördert [12], jede Variabilität in dem experimentellen Protokoll ist immer "Geräusch" für silico Modellierung Zwecke. In diesem Sinn, unser Modell hatte mehrere Herausforderungen wie: 1) der pH-Wert für die Löslichkeit Messung der gesammelten Verbindungen wurde nicht angegeben, 2) die feste Form der Molekül (Polymorphe, Hydrate, Solvate, amorphe) wurde nicht im gemeldete Löslichkeitsmessungen, 3) war es nicht möglich, die Art der Löslichkeit zu überprüfen Messung (kinetisch oder thermodynamisch) und 4) der experimentellen Messung Eine Methode wurde nicht angegeben.

Modellierungsalgorithmus

Aufgrund der Ungewissheit der Datenbank haben wir die Bedeutung einer strengen Protokoll zur Datenauswahl bei der Entwicklung des Originalmodells, um diese Moleküle mit potenzieller Unzuverlässigkeit diskriminieren. Als erster Schritt haben wir eine ANWENDUNGSBEREICH Prüfung Set, bestehend aus Molekülen mit mehr als einem gemeldeten Messung und mit Intersource-Standardabweichung größer als 0 und kleiner als 1 logarithmische Einheit. Wir haben über 1 logarithmische Einheit als Schwelle zur Diskriminierung verwendet unzuverlässige Proben. Dieses RELIABLE Test Set wurde zur Modelloptimierung verwendet.

Aus der QSPR-Perspektive ist es notwendig, eine Reihe von Deskriptoren auszuwählen, die zu das prädiktivste Modell und erleichtert die Modellinterpretation. Zu diesem Zweck haben wir entwickelte einen rekursiven Variablenauswahlalgorithmus basierend auf Regression Zufallswald (RRF) RRF ist eine weit verbreitete Ensemblemethode, die mehrere Entscheidungsbäume zusammenstellt und gibt die Konsensvorhersagen von einzelnen Bäumen aus [13]. Es ist für seine Fähigkeit, „wichtige“ Deskriptoren auszuwählen. Basierend auf dieser Fähigkeit verwenden wir die Anzahl der Auftreten einer Variablen im RRF als Maß für die Bedeutung des Deskriptors, kombiniert mit einer Korrelationsanalyse zwischen Variablen zur Vermeidung von Kollinearität. Jedes Der numerische Deskriptor wurde in der RRF auf zwei Arten injiziert: nicht gedämpft und geschrumpft. Sobald die einzelnen Entscheidung Bäume trainiert und aus dem Ensemble extrahiert wurden, Die Gesamtzahl der Vorkommnisse jeder Variablen wurde berechnet. Nur Variablen mit Eine Anzahl von Ereignissen, die größer als eine Grenzschwelle von 110 sind, wurde beibehalten. von Diese Variablen wurden verworfen, wenn die nicht gedämpfte Variable eine Anzahl von Vorkommnissen, die geringer sind als die Anzahl der Vorkommnisse ihrer homologen geschüttelten variabel. Alle gedämpften Variablen wurden schließlich auch verworfen. Der letzte Satz von Variablen durch anfängliche Berechnung der linearen Korrelation zwischen Variablen rekursiv ausgewählt wurde, und dann nur diejenigen mit der höchsten Anzahl von Ereignissen unter Variablen halten mit einem Korrelationskoeffizienten größer als einer Schwelle von 0,51 zwischen diesen.

In dem Versuch, die Ungewissheit der Daten unabhängig von einem externen Satz zu verringern, Es wurde ein Reinigungsverfahren auf Basis eines RRF-Ansatzes entwickelt. Dieses Verfahren verwendet die Prädiktionsvarianz (PV) der RRF als Metrik zur Unterscheidung unzuverlässiger Proben. Die PV ist eine RRF-Score, die die Variabilität jeder einzelnen Vorhersage mit Respekt vor dem Mittel. Eine hohe PV kann ein Zeichen des anomalen Verhaltens oder der Unsicherheit sein. Dieses Verfahren wurde auf das UNRELIABLE Set, d.h. Moleküle mit wässriger Löslichkeitsstandardabweichung zwischen Quellen gleich 0 oder größer als 1. Um die Parameter dieses Algorithmus, die Minimierung des Wurzelmittels quadratischer Fehler (RMSE) der RELIABLE Test wurde als Zielfunktion verwendet. Erstes, das UNRELIABLE Set wurde zufällig in zwei Sätze von 50 % und 50 % Kardinal geteilt. Eine Regression zufällig Wald wurde auf einem der beiden Sätze trainiert und verwendet, um die wässrige Löslichkeit vorherzusagen und PV des anderen Satzes. Darüber hinaus war die PV der Out-of-Beutel-Proben auch berechnet. Rekursiv wurden Moleküle als innerhalb der PV-Schwelle eingestuft (CLEAN Daten) oder alternativ über die PV-Schwelle (UNCLEAN-Daten), bis keine Moleküle von CLEAN zu UNCLEAN markierten Satz oder umgekehrt geändert.

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

Mit dem CLEAN-Set wurde ein Gradient Boosting Model (GBM) zur Klassifikation ausgebildet. Verwendung von $\log S = -2$ als Cutoff zum Label Moleküle in hochlöslich oder löslich und leicht löslich oder unlöslich. Zwei unabhängige RRF-Modelle wurden basierend auf Diese beiden Teilmengen von markierten Molekülen und ein weiteres RRF-Modell wurden auf alle CLEAN-Daten. Schließlich war die durchschnittliche Prognose unter den drei GBM-Modellen als endgültiger Prädiktionswert angenommen. Die Parameter aller Modelle wurden optimiert basierend auf der RMSE Minimierung des RELIABLE-Testsatzes. Vollständige Details zu unseren Ein entwickelter Algorithmus wird in einem früheren veröffentlichten Papier [4] angegeben.

Zweite „Solubility Challenge“ Vorhersage

Zunächst stellten wir sicher, dass alle Test-Set-Moleküle, die in der Ausgangsquelle gefunden wurden, die als die Das Trainingsset wurde entfernt. Da das Modell zuvor mit der ANWENDUNGSBEREICH Test-Set und durch 5-fache Quervalidierung nutzten wir die gesamte Datenbank (einschließlich des RELIABLE Test Sets) zur Vorhersage der Test-Herausforderungsproben. Zur Analyse die Leistung der Löslichkeitsregressionsmodelle, zwei Arten von Koeffizienten Bestimmung (r^2), Wurzelmittel quadratischer Fehler (RMSE), Mittelwert absoluter Fehler (MAE), Vorspannung und die Prozent der Moleküle mit einem absoluten Fehler kleiner als 0,5 logarithmische Einheiten (% 0,5 log) wurden berechnet.

Ergebnisse und Diskussion

Modellleistung

Die für beide Sätze erhaltenen Statistiken (Test Set 1 = 100 Moleküle und Test Set 2 = 32 In Tabelle 1 sind Moleküle dargestellt. Um die Modellstabilität zu demonstrieren, Ergebnisse werden als Mittelwert und Standardabweichung (Std) gemeldet.

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

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 des endgültigen Konsensusmodells für die Moleküle der zweiten „Solubility Challenge“.

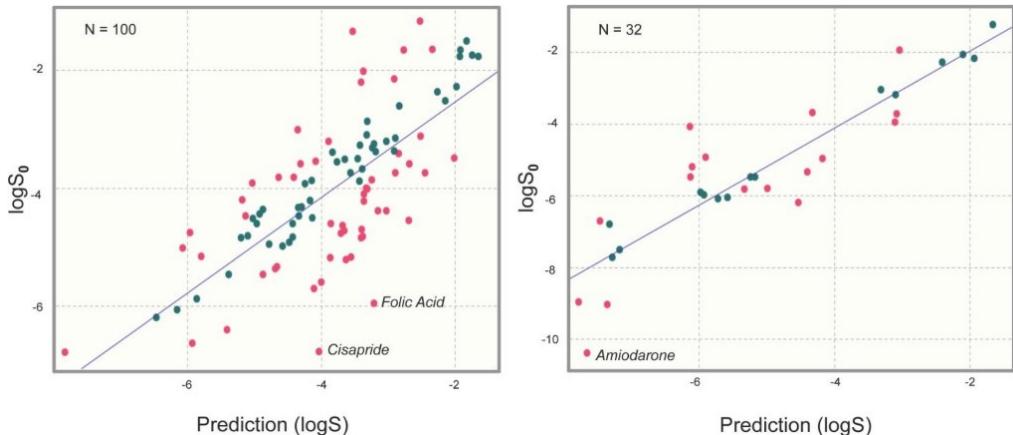


Abbildung 1. Anzahl der Protokolle S (vorhergesagt) vs $\log S_0$ (experimental) für beide Testsätze. Moleküle mit Restwerten höher als 0,5 (Logarithm-Einheiten) sind rot markiert.

Abbildung 2 vergleicht unsere Ergebnisse mit den Top-Rang-Modellen der zweiten „Solubility Herausforderung“. Nach dem mittleren RMSE-Wert, unser Konsens-Modell Rang neun unter den hochrangigen Modellen für die Vorhersage von Test Set 1 und zuerst für die Vorhersage von Test Set 2.

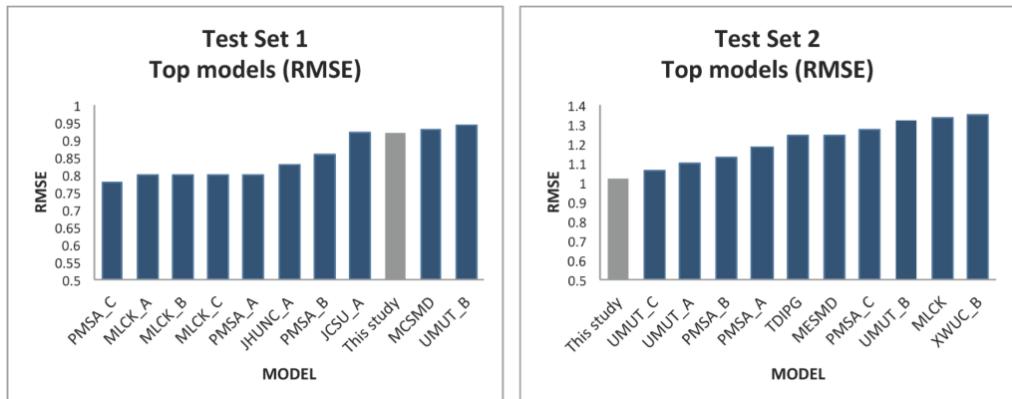


Abbildung 2. Vergleich zwischen den Top-Rang-Modellen der Second Solubility Challenge und unseren Ergebnissen (nach RMSE).

Obwohl es keine signifikanten Unterschiede in Bezug auf die Prognoseleistung gibt, Trainingsset, das wir verwendet haben, enthält wässrige Löslichkeitsmessungen unter nicht-spezifizierte Versuchsbedingungen (pH, Methode und feste Form), ohne Informationen über

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

ihre Art der Löslichkeit (wässrig oder intrinsisch). Es ist bekannt, daß die Anwesenheit von saurem und basischen Gruppen in einem Molekül und der pH-Wert des Mediums beeinflussen den Löslichkeitswert. Intrinsische Löslichkeit entspricht der Löslichkeit der ungeladenen molekularen Spezies, während die wässrige Löslichkeit von dem für Messungen verwendeten pH-Wert abhängt. Daher nicht alle Werte im Trainingssatz sind wahre Eigenlöslichkeitswerte, die die Modellvorhersage des externen Testsatzes mit intrinsischen Löslichkeitsmessungen, führend in einigen Fällen eine höhere Ungewissheit bei Proben, die im Trainingsset enthalten sind.

Wir analysierten die Überlappung unserer Quelle mit den Molekülen aus der zweiten "Solubility Challenge", die auf zwei Überlappungen von 88 und 21 Molekülen, 1. und 2. Test. Nur für den Fall dieser 109 überlappenden Moleküle eine Korrelation. Die Analyse erfolgte zwischen den in der zweiten "Solubility Challenge" und die in unserer Ausgangsquelle gemeldeten wässrigen Löslichkeitswerte gesetzt. Die überlappenden Moleküle wurden aus dem Trainingsset zur Modellierung entfernt. Ziele. Diese Analyse ist in Abbildung 3 dargestellt.

In Anbetracht des Mangels an realen intrinsischen Löslichkeitswerten im Trainingsset, die am meisten problematische Moleküle in der zweiten "Solubility Challenge" sollte die ionisierbare Verbindungen. Die Analyse von Resten ergab, dass Amiodarone (TS2), Cisapride (TS1) und Folsäure (TS1) sind Reaktionsauslöser. Alle enthalten mindestens einen sauren oder basische funktionelle Gruppe und sind praktisch unlösliche Verbindungen. Für diese Moleküle, der wässrige Löslichkeitswert ($\log S_w$) von dem intrinsischen Löslichkeitswert verschieden ist, da nicht genug Solut gelöst wird, um den pH-Wert zu modifizieren, um eine Nah-neutrale Spezies im schlecht gepufferten Medium. Tabelle 2 beschreibt die Werte von $\log S_0$ (zweite "Solubility Challenge"), $\log S_w$ (initial data source), $\log S_w$ (in anderer Quellen) und $\log S_w$ (vorhergesagt).

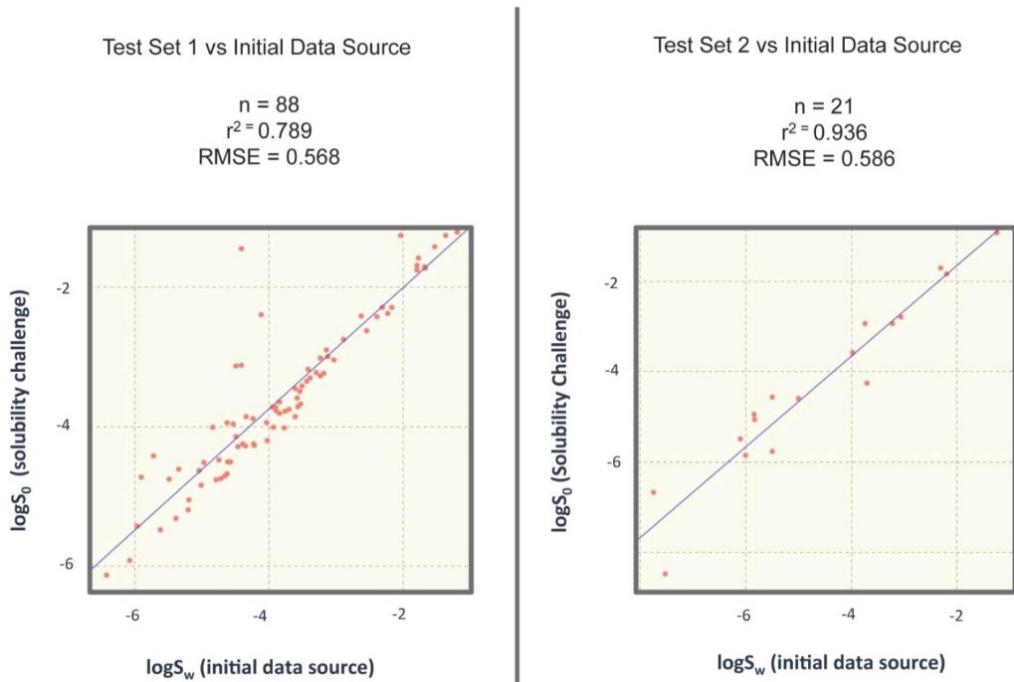


Abbildung 3. Überlappungsprotokoll §mit log S_0 „Analyse der Moleküle der zweiten "Solubilität" Herausforderung“ und das Trainingsset. Für Modellierungszwecke wurden diese überlappenden Moleküle ausgeschaltet das Trainingsset.

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]

^aIntrinsic Aqueous Solubility reported in the second "Solubility Challenge", ^bAqueous Solubility reported for the three outliers in the initial source set, ^cAqueous Solubility reported in other sources

Tabelle 2. Zusammenfassung der Löslichkeitswerte für die Ausreißer.

Um zu beurteilen, ob die Methode in der Lage war, mit der Unsicherheit in den Daten umzugehen, einfaches Experiment wurde durchgeführt. Wie in Abbildung 3 gezeigt, 88 Moleküle aus der ersten Testsatz der Herausforderung überlappt mit unserem ursprünglichen Quellset. Eine Korrelationsanalyse

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

zwischen den beiden von jedem überlappenden Molekül gemeldeten Löslichkeitswerten wurzelförmiges Quadratfehler von 0,568 Log-Einheiten. Wir gehen davon aus, dass der in der Herausforderung bezieht sich auf eine kurierte und zuverlässige Messung, während der in unser anfänglicher Quellsatz könnte von potenzieller Unsicherheit sein. Es gibt einen erheblichen Unterschied zwischen den beiden Wertesätzen für die 88 Moleküle (Konfidenzintervall (CI): 95%; p = 2,9E-5). Als nächstes wurde ein gepaarter t-test für den Vergleich der Leistung entwickelt von zwei Modellen auf Basis zweier unterschiedlicher Trainingseinheiten: (a) die Literaturlöslichkeitsdaten in unserem Ausgangsquellsatz und b) den zuverlässigen intrinsischen Löslichkeitsmessungen in der ersten Reihe der Herausforderung gemeldet. Beide Modelle wurden in der zweiten Herausforderungstest. Es gab keinen signifikanten Unterschied (CI: 95%; p = 0,58) zwischen der Wurzel mittlere quadratische Fehler, die beim zweiten Herausforderungstest mit dem einen oder anderen erreicht werden Trainingseinheiten. Wenn jedoch eine einzige zufällige Waldregression ohne rekursive Auswahl Daten und Variablen ohne Anwendung eines Konsensmodells als Modellierung Der t-Test zeigt einen signifikanten Unterschied (CI: 95%; p = 3.3 E-6). Die Einfluss der Datenqualität auf die Modelleistung hängt vom Modellierungsverfahren ab verwendet. So war die Datenqualität nicht der entscheidende Faktor, wenn eine entsprechende Modellierung Der Ansatz wurde entwickelt, um Datenunsicherheit durch die Wahl der wichtigsten Variablen und die Verwendung eines Konsens-Modells von kombinierten Einzelmodellvorhersagen. Tabelle 3 zeigt eine Überprüfung der Ergebnisse.

Test	Reliable solubility measurements (data challenge)		Literature solubility data (reported in Initial Data Source)	
	n (training) = 88	n (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	0.19 (0.01)	1.93 (0.02)	0.14 (0.06)	1.98 (0.06)
Regression				

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

Tabelle 3. Mittel mit Std-Statistiken auf der Grundlage von zwei Trainingseinheiten bei der Vorhersage des zweiten Tests der zweiten „Solubility Challenge“ mit unserer Methode (Recursive Random Forest (consensus)) gegen eine einzelne RRF: zuverlässig Löslichkeitsmessungen (Datenherausforderung) und Literaturlöslichkeitsdaten.

Automatisiertes System zur wässrigen Löslichkeitsvorhersage

Wir vertrauen, dass es eine Notwendigkeit gibt, öffentlich einen zuverlässigen und vielfältigen Datensatz von Intrinsische Löslichkeitsmessungen für einen strengen Vergleich zwischen Modellierung Algorithmen, aufgrund des relativen Einflusses der Datenqualität auf die Leistung eines Modells. Darüber hinaus sollte die Anwendbarkeit und Reproduzierbarkeit von Löslichkeits-QSPR-Modellen eine Priorität für die Daten F unbrauchbar, A ccessible, I nteroperabel und R (FAIR) [16–18]. In diesem Zusammenhang ist der letzte Zweck des aktuellen Kommentars, öffentlich zugänglich zu machen ein automatisiertes System zur silico wässrigen Löslichkeitsbewertung. Unser Modell ist erfolgreich validiert in einer früheren veröffentlichten Studie und wurde blind getestet mit zweite „Solubility Challenge“, die eine angemessene Leistung zeigt. Das KNIME

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

Der mit dem Papier veröffentlichte Workflow enthält die Ergebnisse unseres Modells auf der zweiten „Solubility Challenge“ und erlaubt die Vorhersage neuer Sets. Der Benutzer kann herunterladen die Arbeitsablauf und folgen die Angaben es enthält von https://pikairos.eu/download/wässrig_solubility_prediction/. Wir entwickelten ein Version basierend auf RDKit und AlvaDesC-Deskriptoren, berechnet mit dem „Descriptor“ Knoten in der "alvaDesc"-Erweiterung enthalten. AlvaDesc 1.0.16 ist mit akademische oder kommerzielle Lizzenzen, die erhalten werden können, indem Sie ein Angebot online verlangen (Registrierung erforderlich) oder direkt per E-Mail (chm@kode-lösungen.net) Nur die SMILES-Codes der Strukturen werden für wässrige Löslichkeit vorhersage, da das Modell keine experimentell ermittelten Wert für die Löslichkeitsberechnung. Das Modell zeichnet sich durch seine Einfachheit aus, da es nur auf Basis von 0-2D-Deskriptoren. Darüber hinaus wird das Modell im Open-Quellanalyseplattform KNIME, eine benutzerfreundliche Software, die für weitere Datenanalyse und Visualisierung.

Schlussfolgerungen

Die mit der Auswertung der zweiten „Solubility Challenge“ erzielten Ergebnisse verstärken die Idee, dass die Datenqualität nicht der Hauptbegrenzerfaktor für eine ausreichende Löslichkeit vorhersage, wenn die implementierte Modellierungsmethodik mit Daten umgehen kann Unsicherheit. In unserem Fall konnte der entwickelte Algorithmus die Datenvariabilität überwinden akzeptable wässrige Löslichkeit vorhersageergebnisse erhalten. Die hier veröffentlichten Ergebnisse eine blinde Vorhersage, da die experimentellen wässrigen Löslichkeitswerte der Herausforderung Zum Zeitpunkt unserer Modellentwicklung und -ausbildung waren Test-Set nicht zugänglich. Obwohl die erzielte Leistung mit denen vergleichbar ist, die in der Überprüfung der die zweite Solubility Challenge, unser Modell basiert nur auf öffentlichen Daten im Vergleich zu einige der besten Modelle der zweiten Solubility Challenge, die auf der riesige wässrige Löslichkeitsdatenbanken von Pharmaunternehmen zur Verfügung. Darüber hinaus ist der Algorithmus unseres Modells global, wie durch den Einsatz von generische Daten ohne die Vorspannung von "Training in der Nähe der Testdaten". Die Automatisierung der vorgeschlagene Methodik und mögliche Anwendung auf größeren Datenbanken, gesammelt unter homogeneren Bedingungen könnte ein Schritt nach vorne sein, um die Löslichkeit zu verbessern Vorhersage während der Drogenentdeckung und Entwicklung. Im Blick auf die Bedeutung des Teilens von Daten und Methoden, um die Reproduzierbarkeit und Anwendbarkeit der Daten zu gewährleisten QSPR-Modelle, wir haben die Daten gemeinsam mit unserem Vorhersagmodell öffentlich zugänglich gemacht auf Basis der KNIME-Analyseplattform als neues kostenloses Tool zur Bewertung wässrige Löslichkeit von Wirkstoffkandidaten.

Referenzen

- [1] A. Llinàs, R. C. Glen, J. M. Goodman. Herausforderung der Solubilität Können Sie vorhersagen 32 Moleküle mit einer Datenbank von 100 zuverlässigen Messungen?. J. Chem. Inf. ANHANG (2008) 1289–1303. <http://doi.org/10.1021/ci000058y>

ADME Vorhersage mit KNIME: Ein retrospektiver Beitrag zur zweiten „Solubility Challenge“

- [2] A. Llinas, A. Avdeef. Solubility Challenge nach zehn Jahren, mit Multilab Shake-Flask Daten, Verwendung von Tight ($SD \leq 0,17 \text{ log}$) und Loose ($SD \geq 0,62 \text{ log}$) Testsets. *J. Chem. Inf. Modell.* 59 (2019) 3036–3040. <https://doi.org/10.1021/acs.jcim.9b00345>
- [3] A. Llinas, I. Oprisiu, A. Avdeef. Ergebnisse der zweiten Herausforderung zu Predict Aqueous Solubilität. *J. Chem. Inf. Modell.* 60, (2020) 4791–4803. <http://doi.org/10.1021/acs.jcim.0c00701>
- [4] G. Falcón-Cano, C. Molina, M. Á Cabrera-Pérez. ADME Vorhersage mit KNIME: In silico Wasserlöslichkeit Konsensus Modell auf Basis überwachter rekursiver zufälliger Wald Ansätze. *ADMET DMPK* 8 (2020) 1–23. <http://doi.org/10.5599/admet.852>
- [5] P.M. Mazanetz, J.R. Marmon, B.T.C. Reisser, I. Morao. Drogenentdeckungs-Anwendungen für KNIME: Eine Open Source Data Mining Platform. *Kurr. Top. Med. Chem.* 12 (2012) 1965–1979. <https://doi.org/10.2174/156802612804910331>
- [6] M.-A. Trapotsi. Entwicklung und Bewertung von ADME-Modellen mit proprietären und opensource data. Universität Hertfordshire, 2017. <https://doi.org/10.18745/th.19719>
- [7] „KNIME Analyse Plattform 4.0.2.“ [Online]. Verfügbar: <https://www.knime.com/download-previous-versions>. [Zugelassen: 17-Mar-2021].
- ANH
ANG A. Mauri, „alvaDesc: Ein Werkzeug zur Berechnung und Analyse molekularer Deskriptoren und Fingerabdrücke“, in Ökotoxikologischen QSARs. Methoden in der Pharmakologie und Toxikologie, K. Roy, Ed. Humana Press Inc., 2020, S. 801–820.
- [9] „RDKit KNIME Integration“. [Online]. Verfügbar: <http://www.knime.com/rdkit>. [Zugelassen: 19-Jun2020].
- [10] M.C. Sorkun, A. Khetan, S. Er. AqSolDB, eine gehärtete Referenzmenge wäßriger Löslichkeit und 2D-Deskriptoren für eine Vielzahl von Verbindungen. *Sci. Daten* 6 (2019) 1–8, Dezember 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. Löslichkeit von neuartigen Verbindungen durch tieferes Lernen. *Vorne, Oncol.* 10. (2017) 1–9. <http://doi.org/10.3389/fonc.2020.00121>
- [12] D.S. Palmer, J.B.O. Mitchell. Ist die experimentelle Datenqualität der limitierende Faktor bei der Vorhersage die wässrige Löslichkeit von medikamentösen Molekülen?. *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: ein Klassifikations- und Regressionstool für die Compoundklassifikation und QSAR-Modellierung. *J. Chem. Inf. Comput. Sci.* 43 (2003) 1947–1958. <http://doi.org/10.1021/ci034160g>
- [14] M. Salahinejad, T.C. Le, D.A. Winkler. Wässrige Solubilitätsvorhersage: Do Crystal Lattice Interaktionen Hilfe?. *Mol. Pharm.* 10. (2013) 2757–2766. <http://doi.org/10.1021/mp4001958>
- [15] S.H. Yalkowsky, Y. He, P. Jain. Handbuch von Aqueous Solubility Data, Second. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, USA: CRC Presse Taylor und Francis Gruppe, 2010.

- [16] M. D. Wilkinson, M. Dumontier, I.J. Aalbersberg et al. Kommentar: The FAIR Guiding Grundsätze für die wissenschaftliche Datenverwaltung und -verwaltung. Daten 3 (2016) 1–9.
<https://doi.org/10.1038/sdata.2016.18>
- [17] J. Wise, A.G. de Barron, A. Splendiani et al. Implementierung und Relevanz der FAIR-Daten Grundsätze der biopharmazeutischen FuE. Drug Discovery Today 24, (2019) 933–938.
<http://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: Methode und Datenaustausch und Reproduzierbarkeit von wissenschaftlichen Ergebnissen. J. Chem. Inf. Modell. 60 (2020) 5868–5869. <https://doi.org/10.1021/acs.jcim.0c01389>

KNIME Unterstützung

Diese Kategorie vereint alle hilfreichen Beiträge unserer COTMs. Das könnte auf dem KNIME-Forum helfen anderen KNIME-Nutzern, indem sie Lösungen für Fragen oder indem Sie Gedanken und Vorschläge für bestimmte Anwendungsfälle teilen. Zum Glück, unsere Heer der Helfer ist nicht nur auf dem KNIME Forum anwesend, sondern auch andere Orte wie Facebook oder Twitter. Wenn Sie jemals einen Rat brauchen, ist es wahrscheinlich, dass einer unserer Unterstützung KNinjas wird vor Ort sein. Die Kategorie „KNIME Support“ umfasst:

- **Evan Bristow**
 - Senior Principal Analyst @Genesys
- **InfMad**
 - Marketing & Digital Analytics Expert @BaseCero Marketing
- **Makkyn**
 - n/a
- **Ignacio Perez**
 - Direktor, Eigentümer @IQuartil
- **Markus Lauber**
 - Senior Data Scientist, Big Data Analytics @Deutsche Telekom
- **Weitere Informationen**
 - Director Data Ops @Triton Digital
- **Brian Bats**
 - Data & Integration Architect @The Walt Disney Company



Evan Bristow wurde nominiert Beitrag des Monat für Januar 2021. Zusammen mit Miguel InfMad, er wurde für den Bau und die Pflege eines [Facebook](#) Gemeinschaft für alle Nutzer – von Newbies bis zu Experten. Die Gruppe ist seit 2019 um und zählt mehr als 1500 Mitglieder. Es ist eine lebhafte, kompetente und sehr hilfreich Gruppe zu Ihren Data Science-Problemen und KNIME Fragen. Seine Unterstützungsarbeit endet nicht

Facebook. Seine Beiträge finden Sie auf dem KNIME Forum und auf dem KNIME Community Hub.

Evan ist eine lange Zeit KNIME Benutzer- und Datenwissenschaft Experte. Er beschreibt sich als begeistert und kreative Person, die die Kunst und Wissenschaft der Forschung. Evan ist derzeit Senior Principal Analyst bei Genesys, wo er KNIME verwendet, um verschiedene analytische und maschinelle Lernansätze Business-Einsichten und Support-Geschäftsbedürfnisse.

Besuchen Sie Evans [Raum auf dem KNIME Hub](#) oder sein Profil in [Das KNIME Forum](#) (Hub/Forum Griff: Evanc)



Expertise im Dienst der Gemeinschaft: Was Unterstützung ist wirklich

Mein Data Guest – Ein Interview mit Evan Bristow

Autor: Rosaria Silipo



Es war mir ein Vergnügen, kürzlich live auf LinkedIn Evan Bristow als Teil der
[Mein Gott.](#) Daten Gast Interview-Serie. Er teilte Einblicke in das, was es bedeutet, ein
[KNIME](#)
Unterstützungsgruppe auf Facebook neue und erfahrene KNIME-Nutzer beim Lernen

Reise, was sind die Dinge, die Sie in einer solchen Rolle erwarten können, und wo Sie zeichnen sollten die Linie in der Unterstützung, die Sie bieten.

Evan Bristow ist ein Senior Principal Data Analyst bei Genesys, wo er verschiedene
 analytische und maschinelle Lernansätze, um geschäftliche Einblicke zu geben und zu unterstützen Geschäftsbedarf. Im Laufe der Jahre hat er sein Know-how und technische Fähigkeiten auf wettbewerbsfähige Marktanalysen durchführen, methodische Unterstützung für Unternehmen bieten Kundenbeziehungen zu verwalten und wiederkehrende und ad-hoc-Operationen zu produzieren und Finanzberichterstattung für Organisationen in der Region AMER. Wann gefragt, was er tut, er möchte sicherstellen, dass die Menschen verstehen, dass er nicht nur das "was" bereitstellt, aber auch das "Warum", das "wie", und das "was wenn". Evan ist auch einer der leidenschaftlichen und Experten-Gehirn hinter der KNIME-Unterstützungsgruppe auf Facebook – zusammen mit Mitbegründer InfMad — und genießt es, seine Expertise in den Dienst der Gemeinschaft zu stellen.

Rosaria: Können Sie erklären, was Sie meinen, wenn Sie sagen, dass Sie nicht nur das "was", aber auch das "Warum", das "wie" und das "was, wenn"?

Evan: Eine Sache, die mir sehr wichtig ist, ist die Analyse auf eine andere Ebene. Nicht nur in der Lage sein, die Zahlen bereitzustellen, sondern ihnen auch Kontext und Bedeutung zu geben.

Sagen wir, Sie arbeiten für ein Unternehmen und sie wollen wissen, wie viel sie sind im nächsten Quartal zu buchen. Das ist eine sehr interessante Frage, die Sie versuchen können modellieren in einer Reihe von verschiedenen Möglichkeiten. Was jedoch entscheidend ist, ist, dass Sie diese verdauen Fragen zu den Bestandteilen: „Sind wir vermissten etwas?“, „Überragen wir?“, „Was sind mögliche Ergebnisse, die wir nehmen müssen“

bei der Erstellung eines Modells berücksichtigt werden?“. Die Beantwortung dieser Fragen definiert das „was wenn“

Wenn Sie Unterschiede in dem sehen, was Sie projizieren vs. was andere Menschen sind Projektierung dann können Sie in Ihr Modell bohren und bestimmen die „Warum“. Auch, wenn Sie sind eine Art von Wurzel verursachen Analyse und Sie sehen Unterschiede in der Make-up von Ihrem Geschäft zwischen einer Periode und einer anderen, Sie können beobachten, was fahren diese Unterschiede.

Sobald diese Ursachen identifiziert werden, sollten Sie in der Lage sein, sie in einen breiteren Kontext zu setzen. Dies ist, wenn Unternehmen Interessenvertreter fragen: „Wie viel werden wir als nächstes buchen Quartal?“, „Was haben wir in diesem Viertel historisch gebucht?“ und „Wie funktioniert das? vergleichen über die Zeit?“. Am Ende, was sie wirklich wissen wollen, ist „Sind wir gut?“.

Rosaria: Was tun Sie genau in Ihrem täglichen Job? Was sind Ihre Aufgaben als Principal? Data Analyst?

Evan: Ich denke gerne an mich als Daten MacGyver. Ich nehme verschiedene Datenquellen oder verschiedene Informationen und ziehen sie durch einen analytischen Prozess, um zu erstellen etwas, das die Frage beantwortet oder das Problem löst. Im Moment bin ich an einigen verschiedenen Modellen arbeiten. Zwei von ihnen projizieren, wie viel Buchungen, die wir aus unserer aktuellen Pipeline erwarten. Der andere berechnet Gelegenheits-Level-Scoring, das ist, wie wahrscheinlich es ist, individuelle Möglichkeiten zu gewinnen. Das gibt unserer Pipeline einen weiteren Blick auf die Risikobewertung. Wann immer die Daten groß werden oder das Spiel wird hart, [KNIME Analytics Plattform](#) ist immer mein Werkzeug.

Ich bin kein Kodierer im Herzen, so ist es für mich viel schneller und zugänglicher zu bauen Arbeitsabläufe in KNIME. Meiner Meinung nach ist der größte Vorteil, wie transparent und Die visuellen Datenflüsse sind leicht zu verstehen. Ich kann immer einen Workflow hochziehen, es einem zeigen Stakeholder, und wir können Fragen zur Fliege ansprechen. Wenn ich das tun sollte in Python und Schreiben Sie Code zur gleichen Zeit, würde ich die Aufmerksamkeit der Unternehmer in der Blink eines Auges.

Rosaria: Wie können Sie alle KNIME-Nutzer auf der Facebook-Gruppe unterstützen? aktiv? Ich meine, du hast einen Job, wie findest du die Zeit?

Evan: Das meiste, was wir normalerweise in der [KNIME Facebook-Gruppe](#) Leute zu bekommen, die sind mit einem anderen Werkzeug oder einer anderen Möglichkeit vertraut, Dinge zu tun, die KNIME akklimatisiert. Wir neigen dazu, mehr von einer Anfängergruppe von Personen zu bekommen, die zum Beispiel mit Excel will aber wissen, wie sie dasselbe mit KNIME tun können. Das sind in der Regel nicht zu komplizierte Themen, wo wir unbedingt jemanden auf die [KNIME Forum](#). Darüber hinaus nutzen wir den Raum, um die Gemeinschaft zur Verfügung zu zeigen Lernmittel. Zum Beispiel die Bücher der [KNIME Presse](#) und insbesondere [Excel zu KNIME](#) Serie, die in verschiedenen Sprachen verfügbar ist. Unser Motto lautet: „Du Unterrichten Sie sie zu Fischen statt nur geben ihnen einen Fisch“.

Allerdings, wenn wir einige seltsame Fragen, die mehr fortgeschrittene Kompetenz erfordern und ich denke, ich beziehe mich normalerweise auf das KNIME Forum.

Rosaria: Was ist der Unterschied zwischen der KNIME Facebook-Gruppe und dem KNIME Forum?

Evan: Das KNIME Facebook-Gruppe ist ein sehr einfacher, Low-Stake-Platz, wo Sie einfach gehen und Fragen stellen. Sehr oft Fragen kommen von KNIME-Nutzern, die die Software auf dem Arbeitsplatz und kann beispielsweise Fragen zur Datenverarbeitung haben — z. Kombinieren, Gruppieren, Filtern usw. – aber nicht auf KNIME-Unterstützung innerhalb ihrer Abteilung oder Organisation. Was diese Benutzer brauchen, ist nur jemand, den sie werfen könnten eine schnelle Frage, und eine sofortige Antwort erhalten. Das ist die Facebook-Gruppe ist für.

Das KNIME-Forum hingegen ist es, dass Benutzer mehr artikulierte, technische oder spezialisierte Fragen und erhalten Unterstützung direkt von KNIMERS oder anderen Experten Benutzer.

Rosaria: Es gibt Fragen, die nicht beantwortet werden sollten - oder sogar gestellt - in einer Unterstützung Gruppe? Neben den offensichtlich unangebrachten Fragen natürlich.

Evan: Wir helfen Ihnen, wenn Sie eine bestimmte Frage haben, wie etwas funktioniert. wenn Sie brauchen Hilfe bei Ihren Hausaufgaben und Sie möchten wissen, was der Pivoting-Knoten Wir helfen Ihnen dabei. Wenn etwas ausgeht oder einen Fehler wirft, gehen Sie vor und fragen Sie auf der KNIME Facebook-Gruppe. Wenn Sie jedoch einen Datensatz haben und fragen wir, wie wir Clustering-Strategien anwenden, werden wir Ihnen nicht helfen, da dies eindeutig eine Hausarbeit ist Zuweisung. Im Allgemeinen erwarten wir, dass die Benutzer ein wenig die Beinarbeit auf ihrer Seite tun bevor sie uns zur Hilfe kommen.

Rosaria: Sie unterstützen auch aktiv KNIME-Nutzer auf dem KNIME Forum?

Evan: Nicht so sehr wie früher. Ich war vor ein paar Jahren ziemlich aktiv, und ich immer noch überprüfen Sie jetzt und dann. Die Wahrheit ist, dass ein Vollzeitjob, Familienpflichten, und die KNIME Facebook Gruppe ist hart genug. Ich kann einfach nicht mithalten alles.

Rosaria: Was sind die häufigsten Fragen, die Sie in der KNIME Facebook Gruppe bekommen?

Evan: Unsere am häufigsten gestellten Fragen folgen dem Thema Werkzeugmigration. Sie in der Regel von Menschen, die wissen, wie man Daten mit Werkzeug X manipuliert und wollen wissen, wie man dasselbe in KNIME tut. Die zweitgrößte Gruppe häufiger Fragen ist in der Regel mit allgemeinen maschinellen Lernfragen verbunden, zum Beispiel: „Warum? Partitionieren Sie Ihre Daten in Zug und Testsatz?“ oder „Wie würden Sie Kreuz tun? Validierung?“. Schließlich erhalten wir oft auch Fragen zu sehr spezifischen Daten wrangling Operationen wie „Wie filtere ich nach bestimmten Kriterien?“. Werkzeug Migration, maschinelles Lernen und Datenknüpfen – das sind die drei Hauptthemen.

Rosaria: Diese Fragen sind etwas zu erwarten. Jetzt bin ich neugierig. Was ist das? die komplizierteste Frage, die jemals in der Gruppe gestellt wurde?

Evan: In der Regel, wenn Fragen zu kompliziert werden, werden wir sie auf die KNIME schieben Forum. Ich denke, die schwierigste Frage musste mit Regex machen. Jemand hatte ein Streichfeld die mit einem Datum und einer eindeutigen Kennung versehen war, und sie wollten parse, dass in das Datum und die eindeutige Kennung. Das Problem war auch, daß Je nach Datenzeile könnte sich eine Anzahl von eindeutigen Kennungen ändern. Mit Text Manipulationsknoten, Spaltzellen und Schwenken war der vorgeschlagene Weg, dies anzupacken Problem — ich glaube. Aber das ist sicherlich nicht der einzige.

KNIME ist wie ein Schweizer Armeemesser der Datenanalyse und es gibt Dutzende von verschiedenen Probleme zu lösen und mit dem gleichen Ergebnis zu kommen. Diese einzigartigen Anwendungsfälle Spaß zu beantworten, aber, und es ist immer interessant zu sehen, wie andere Benutzer antworten die Fragen.

Rosaria: Bevor die Menschen tatsächlich um Unterstützung bitten, müssen sie beginnen, über KNIME zu lernen auf eine Weise oder eine andere. Wie beraten Sie Neulinge, ihre Reise mit KNIME zu beginnen?

Evan: Der beste Weg ist, ein Projekt zu erstellen, in dem Sie die Daten relativ kennen und was das Ziel des Projekts ist. Wenn Sie die Schritte eindeutig definieren können, die Sie muss implementiert werden, um das Ziel zu erreichen, das Sie im Auge haben, Sie können nur verbinden Punkt A bis B. Und wenn Sie stecken, können Sie immer den KNIME Hub für einen Workflow suchen Beispiel, das Ihrem Anwendungsfall nähert, und sich von dem Fluss inspirieren lassen, den jemand auch entworfen. Im Gegensatz zu anderen Analysetools – beispielsweise Excel – ist KNIME visuell, sequentiell und transparent, und das erleichtert wirklich das Lernen.

Rosaria: Wie haben Sie über KNIME gelernt?

Evan: Ich habe KNIME vor vielen Monden entdeckt. Damals arbeitete ich für eine Firma die B2B-Marketing-Forschung und wir nutzten SAS und SPSS Modeler. Das Problem war, dass wir nur eine Kopie von SPSS Modeler hatten, weil es ein hefty hatte Preis-Tag. Dadurch wurde es nur auf dem Computer einer Person installiert, was ziemlich unkomfortabel für die Teamarbeit und für die Skala von Projekten. Also fing ich an, eine Alternative zu suchen Werkzeug, und das ist, wenn ich über KNIME gelernt. Wir haben damit angefangen und erkannt, dass Es war viel besser, Dinge zu tun als andere Werkzeuge. Nicht nur war KNIME besser laufen Analysen, aber es war auch einfacher zu verbinden, zu importieren, zu kombinieren und zu manipulieren verschiedene Datenquellen. Nicht zuletzt, als eine freie Plattform, es machte das Geschäft auch glücklich.

Rosaria: Hat Ihnen jemand in Ihren frühen Tagen bei KNIME geholfen?

Evan: Leider hatte ich keine enge Unterstützungsgruppe, also war das KNIME Forum das Hauptort, wo ich meine Fragen stellen würde. Im Allgemeinen aber, ich mag Dinge zu finden aus eigener Kraft. Ich mag es zu zeigen, zu klicken und herumzuspielen. Ich habe wirklich in die Software zum Beispiel durch das Lesen von Knotenbeschreibungen, und ich habe die

[KNIME](#)

Beispiele für Server viel, um von diesen Workflows inspiriert zu werden. ich habe immer geliebt

Beispiele Server. Es ist sehr nützlich, vor allem für Neulinge. Außerdem habe ich meine Fragen und gefunden ziemlich gute Antworten.

Rosaria: Wenn Sie die drei Top-Features, die Sie mögen über KNIME wählen, was würden sie/Sie sein?

Evan: Die Fähigkeit, verschiedene Datenquellen und Technologien zu integrieren, ist wahrscheinlich eine der besten Dinge in KNIME. Ich kann Daten von im Wesentlichen überall ziehen und speichern im Wesentlichen überall, ohne sich Sorgen darüber, wenn es sich vermischen wird. Das ist etwas, das Sie oft Sie haben Daten auf einem Server gespeichert, Sie haben Daten in einem smartsheet somethere, und jemand sendet Ihnen eine Excel-Datei. Mit KNIME können Sie alle diese verstreuten Informationen zusammenbringen, eine Analyse daraus erstellen und einfach auf Ihren Server setzen, um eine Visualisierung zu erstellen.

Eine weitere große Eigenschaft ist die Fähigkeit, Wissen von einem Workflow zu nutzen einen weiteren Workflow. Ich baue einen Workflow für ein Projekt und ich kann es einfach erweitern oder nehmen Teile dieses Workflows und verwenden es in einem anderen Workflow. Ich fange nie ganz von Kratzern und das spart viel Zeit und Mühe.

Schließlich, KNIME GUI, und die Fähigkeit, Segmente und Prozesse Ihrer Ein Arbeitsablauf in ein Bauteil. Das hilft Ihnen, sich auf das zu konzentrieren, was Sie tun anstatt was du schreibst.

Rosaria: Und die Top-3-Knoten, die Sie nie ohne machen konnten?

Evan: Der erste Pivot und DB Pivot Knoten. Pivozieren in SQL ist schrecklich, einschwenken KNIME ist einfach. Zweitens Statistik Knoten. Das ist ein unschätzbarer Knoten, den Sie Verwenden Sie im Grunde jedes Mal, wenn Sie in frische Daten ziehen. Es macht es einfache zu sehen, was Art von Daten, mit denen Sie es zu tun haben. Und schließlich, Hinterknoten . Jeder hat zu tun eine kleine Filterung an irgendeiner Stelle.

Rosaria: Unterstützung ist nicht die einzige Möglichkeit, wie Sie mit der KNIME Community interagieren. Du warst Mitglied der Redaktion des Medium Journals „Low Code for Advanced Data Wissenschaft“. Wie war die Erfahrung? Was tut ein Mitglied der Redaktion?

Evan: Dies ist wieder eine andere Möglichkeit, der Gemeinschaft zu helfen, nur mit einem anderen Kanal und Medium. Mitglieder des Redaktionsremiums versuchen, den Beitragenden zu helfen, ihre Arbeit dort, formen ihre Datengeschichten, zünden Gespräch auf Low-Code-Tools, und Spot interessante Themen, die es wert sind, mit den Lesern zu teilen, um ihre eigenen Reise zur codelosen Datenanalyse mit KNIME.

Rosaria: Wir erreichen das Ende unseres Interviews. Bevor wir uns verabschieden, gibt es jede Pläne, Ihren Erfolg mit einer Unterstützungsgruppe auf einem anderen Social Media-Kanal zu replizieren?

Evan: Ich habe momentan keine konkreten Pläne, aber ich könnte mir vorstellen, ein LinkedIn zu erstellen oder WhatsApp-Unterstützungsgruppe. Einige der allgemeinen Social Media Chat-Gruppen, wo

Menschen können lässig einfache Fragen fallen und eine Antwort sofort erhalten, ist etwas, an dem ich mich beteiligen möchte. Zu hoch. Für doozy oder sehr technische Fragen bleibt das KNIME Forum der Ort, an dem es geht.

Rosaria: Wie können sich Menschen aus dem Publikum mit Ihnen und Ihrer Arbeit in Verbindung setzen?

Evan: Der beste Weg, mit mir in Kontakt zu treten, ist über die KNIME Facebook Group ([KNIME Analystengemeinschaft](#)) Sie sind immer willkommen, der Gruppe beizutreten, aktiv posten und helfen
Ich und die anderen Admins unterstützen die KNIME-Gemeinschaft.

Sehen Sie das ursprüngliche Interview mit Evan Bristow auf YouTube: [Mein Data Guest – Ep 9 mit Evan Bristow](#) „



InfMad wurde nominiert Beitrag des Monat für Januar 2021. Zusammen mit Evan Bristow, er wurde für den Bau und die Pflege eines [Facebook](#) Gemeinschaft für alle Nutzer – von Newbies bis zu Experten. Die Gruppe ist seit 2019 um und zählt mehr als 1500 Mitglieder. Es ist eine lebhafte, kompetente und sehr hilfreich Gruppe zu Ihren Data Science-Problemen und KNIME Fragen. Seine Unterstützungsarbeit endet nicht

Facebook. Seine Beiträge finden Sie auf dem KNIME Forum und auf dem KNIME Community Hub.

Miguel ist eine lange Zeit KNIME Benutzer- und Datenwissenschaft Experte. Seine berufliche Entwicklung umfasst als Analyst und Projektmanager von Daten Bergbauprojekte für das Marketing. Vor kurzem, er ist ein Experte für digitale Kampagnen von PPC, SMM und SEM. Miguel hält einen Doktortitel in Applied Ökonomie der Complutense Universität Madrid.

Besuchen Sie Miguel's Raum auf dem KNIME Hub (Hubgriff):
miguel_infmad)



Wie man Google Analytics mit KNIME verbindet

Autor: Miguel InfMad; Übersetzt von: Roberto Cadili



In diesem Artikel werden wir erklären, wie man sich mit einem Google Analytics-Konto verbindet

KNIME Analytics Platform. Unser Ziel ist es, Automatisierung der Datenextraktion und erstellen

benutzerdefinierte

Kennzahlen die Leistung unserer Projekte zu überwachen.

Für diejenigen von uns, die keine Experten-Codeer sind, manchmal ist es schwer, reibungslos zu integrieren digitale und analytische Anwendungen. Mit diesem Miniguide hoffen wir Lesern zu helfen

Interesse an digitaler Analytik über die von Google Analytics angezeigten Daten hinausgehen und ihre Analyse mithilfe freier und offener Tools zu verbessern.

Was ist KNIME?

KNIME ist flexibel und skalierbar Open Source, analytische Plattform . Es kann für jede verwendet werden Art der Datenaufgabe, z.B. Zugriff auf und Wrangling von Daten, Bauvorhersagemodele oder Integration eines Rechenrahmens für Big Data.

KNIME ist der perfekte „ Schweizer Armeemesser“ für Digitale Analyse , und auch das Werkzeug, das wir bei BaseCero verwenden (in der Tat das Hauptquartier des Unternehmens ist in Zürich, Schweiz).

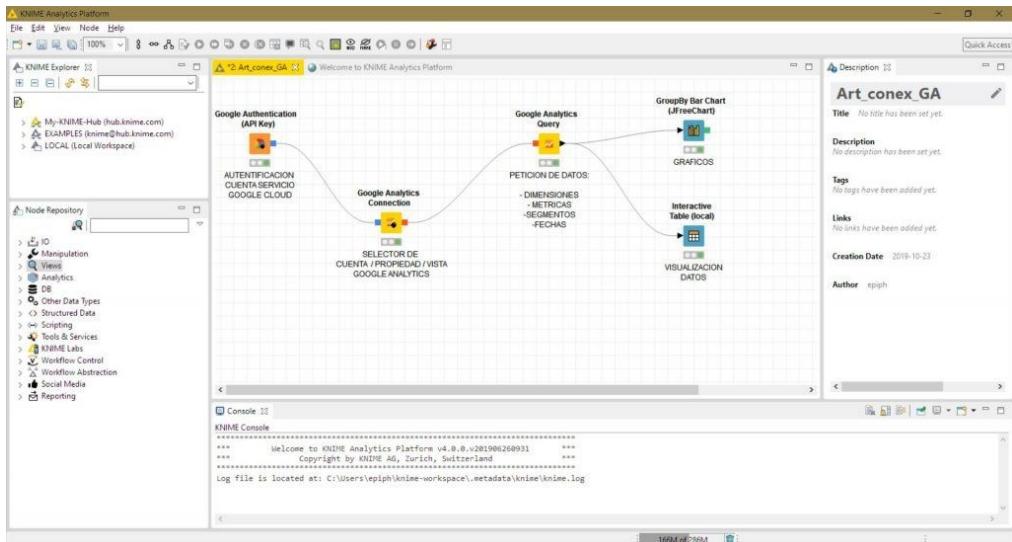
KNIME ermöglicht die Aufnahme und Transformation jedes Datenformats, die Integration mit viele Programmiersprachen, die Entwicklung und Automatisierung von komplexen Prozesse, die Erstellung von prädiktiven Modellen, die Verbindung zu Web-Diensten und Dritt-Party APIs und vieles mehr.

KNIME ist eine visuelle Programmiersoftware. Das bedeutet, du musst nicht wissen, wie man programmieren in einer traditionellen Skriptsprache, um sie nutzen zu können. Die beiden Basisversion und die breite Palette von Erweiterungen und Integrationen, um die KNIME zu erweitern analytics power kann kostenlos heruntergeladen werden und sind voll funktionsfähig. Sie können herunterladen die neueste Version von:

<https://www.knime.com/downloads/download-knime>

Es gibt auch eine Unternehmensversion von KNIME Software zur Bereitstellung von Datenwissenschaft Workflows, Automatisierung der Workflow-Ausführung und Verwaltung der Zusammenarbeit über Anwender und Räume.

Wir ermutigen Sie sehr, es zu versuchen. Zusätzlich zu den oben aufgeführten Fähigkeiten, das Tool mit einer Vielzahl von Workflow-Beispielen für eine breite Palette von Anwendungen angereichert wird, und wenn Sie Hilfe beim KNIME Forum suchen **groß und aktiv** **Gemeinschaft** ist die perfekter Ort, um Ihre Fragen zu stellen.



KNIME Workflow, um sich mit Google Analytics zu verbinden, Anfragen zu senden und Ergebnisse zu erstellen.

Vorteile der Integration von KNIME mit Google Analytics

Während die Daten, die wir abrufen können Integration KNIME und Google Analytics gleich die auf Google Analytics angezeigt werden, mit einem externen Tool wie KNIME bietet mehrere Schlüsselvorteile, die nicht im Originaldienst verfügbar sind. Finden Sie die wichtigsten darunter:

- **Prozessautomatisierung**

Erstellen von Kennzahlen für Ihr Projekt eine nach dem anderen, wie Sie durch die Vielzahl von Daten von Google Analytics können ziemlich zeitaufwendig sein. Entwicklung eines automatisierter Prozess ermöglicht es uns, viel Zeit zu befreien, sich auf das zu konzentrieren, was wirklich Themen: Interpretation und Wahrnehmung der Daten zur Verbesserung des Projekts.

- **Individuelle Extraktion von Funktionen und Metriken**

Manchmal müssen wir benutzerdefinierte Kennzahlen entwickeln, die nicht standardmäßig in Google Analytics. Darauf hinaus ermöglicht die Extraktion von Roh-Webdaten uns, den Analyseprozess anpassen und digitale Ressourcen optimieren.

- **Entwicklung mittels eines Open-Source-Tools**

Unter der GPL-Lizenz entwickelte professionelle Analysewerkzeuge (General Public Lizenz) garantieren transparentere Prozesse und reduzieren Entwicklungskosten.

- **Integration von Datenquellen**

Web-Verkehrsdaten können zusammen mit weiteren Daten, die kommen, reibungslos verarbeitet werden aus anderen Kanälen wie Social Media oder Newsletter. Dies ermöglicht das Gelenk Analyse einer Mehrkanalstrategie zur Bewertung und Optimierung der Entwicklung Projekte und digitale Kampagnen.

Bevor wir tiefer leben

In diesem Artikel konzentrieren wir uns darauf, wie man [Zugriff auf die Web-Verkehrsdaten von Google Analytics KNIME](#). Ein grundlegendes Verständnis der Google Analytics und der KNIME Analytics Platform ist ratsam, den Inhalt jedes Abschnitts richtig zu folgen.

Wie sammelt Google Web-Verkehrsdaten?

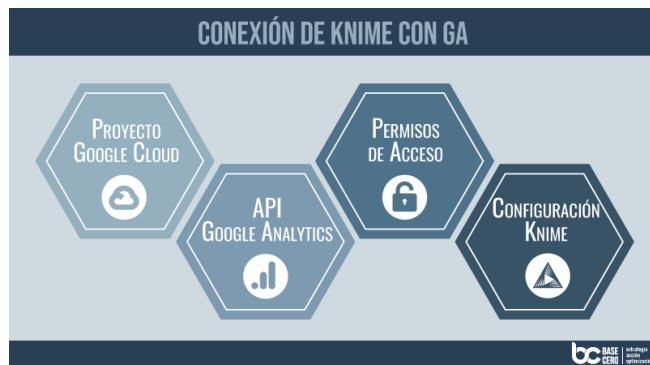
Um es in einfachen Begriffen zu setzen, setzt Google Analytics auf [Verfolgung von Cookies](#) installiert in der Browser, um die Aktivität der Nutzer zu verfolgen und Web-Verkehrsdaten zu sammeln (vorgesehen, dass der Benutzer akzeptiert sie). Gesammelte Daten stehen dann zur Inspektion direkt auf Google zur Verfügung [Analytics, und sie können auch über die dedizierte API](#). Diese zweite Option ist das, was wir in unserer Umsetzung mit KNIME verwenden werden.

Bevor wir diese beiden Technologien verbinden, müssen wir:

- ANH ANG Haben Sie eine Website von Google Analytics mit den darin installierten Tracking-Cookies von Google Analytics.
2. Erstellen Sie ein Google/Gmail-Konto, um Google Analytics und Google nutzen zu können Cloud Platform.
 3. [Installieren KNIME Analytics Plattform](#), die [KNIME Google Connectors](#) [Erweiterung und](#) die [KNIME Twitter Connectors](#) [Erweiterung \(für weitere Informationen über KNIME](#) [Erweiterungen und Integrationen überprüfen die Dokumentation](#))

Im nächsten Abschnitt werden wir zeigen, wie man [Google Analytics mit KNIME verbinden](#) **4 Schritte**.

Google Analytics mit KNIME verbinden

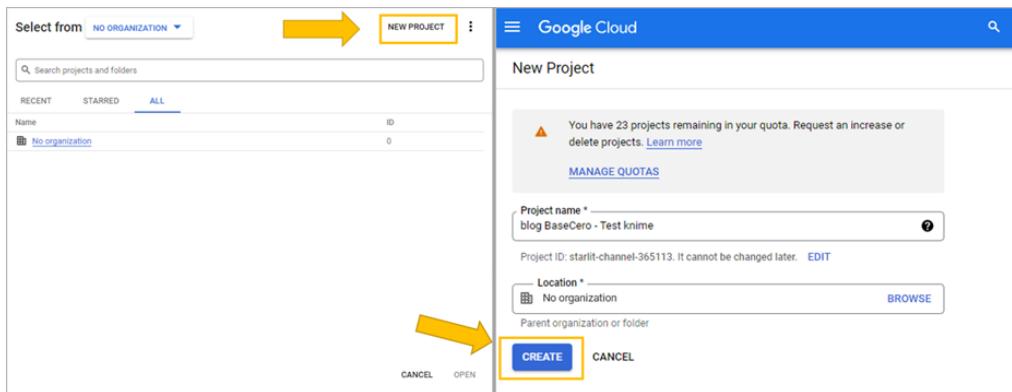


Prozessablauf zur Verbindung mit Google Analytics mit KNIME in 4 Schritten:
1) Erstellen Sie ein Google Cloud-Projekt; 2) Google Analytics API aktivieren; 3)
Zugriffsrechte festlegen; 4) Konfigurieren Sie die KNIME Analytics Platform.

Schritt 1: Erstellen Sie ein Projekt auf Google Cloud Platform

Wir beginnen, eine **Neues Projekt** auf Google Cloud Platform. Dies ist erforderlich
Zeit, die Sie eine Lösung mit einem der von Google angebotenen Analyseprodukte aufbauen möchten.

Wenn Sie keine [Google Cloud](#) Sie können sich über ein Gmail-Konto anmelden.

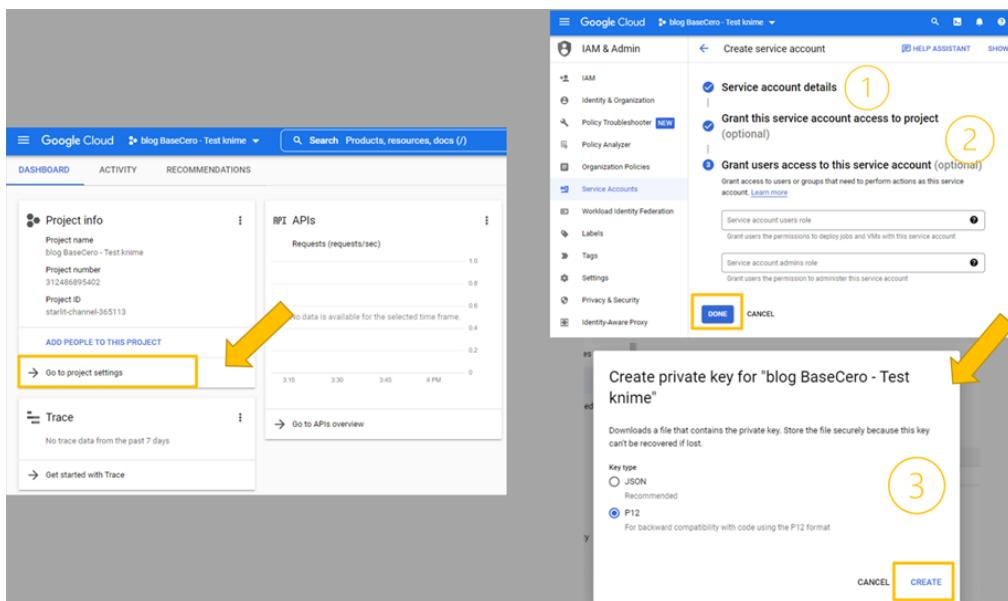


Erstellen eines neuen Projekts auf der Google Cloud Platform.

Im neu erstellten Projekt definieren wir eine **Servicekonto**. Dies ist notwendig, um
die Projektzugriffsberechtigungen auf Google Analytics-Konto(s) verwalten, indem
Sicherheitsdatei eindeutige Identifikationsschlüssel und ein E-Mail-Konto für die Verbindung.

Dazu öffnen wir die Projektkonfigurationen, wählen Sie die Option "Service-Konto" auf
die linke Spalte und eine erstellen.

KNIME Unterstützung – Miguel InfMad
Wie man Google Analytics mit KNIME verbindet

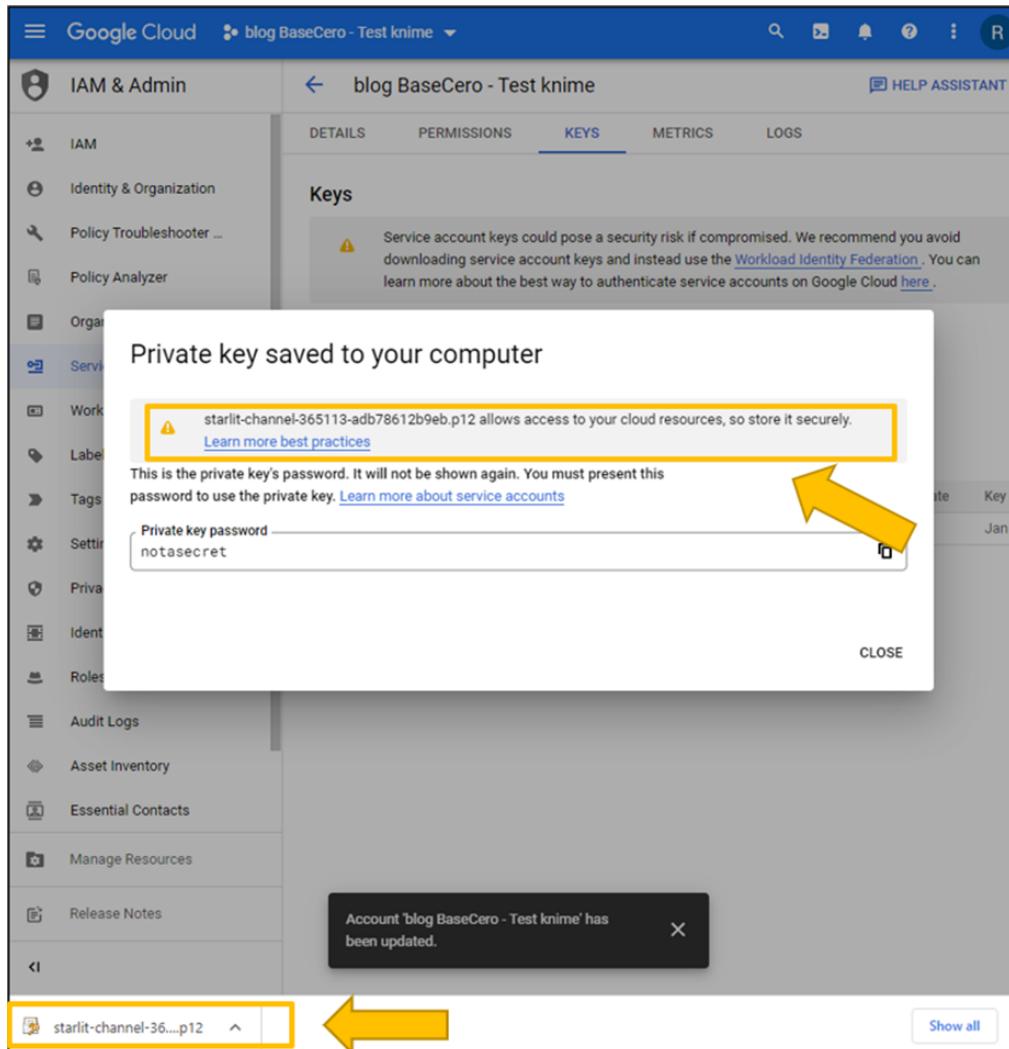


Erstellen eines Service-Accounts innerhalb des neu erstellten Projekts.

Die Erstellung des Service-Accounts beinhaltet 3 Schritte:

- Schritt 1 geht es um die Definition von Service-Account-Details zur einfachen Identifizierung.
- Schritt 2 ist optional und wir können es leer lassen.
- In Schritt 3 erstellen wir private Schlüssel für das Service-Konto und ein downloadbar Sicherheitsdatei, die wir später in KNIME verwenden, um Google Analytics zu authentifizieren.

Wir haben jetzt Identifikations- und Verbindungsinformationen für den neu geschaffenen Service Konto, sowie das zugehörige **Authentifizierungsschlüssel und Sicherheitsdatei** der ein lokales Verzeichnis auf unserer Maschine.



Erstellen eines privaten Schlüssels für das Service-Konto und eine herunterladbare Sicherheitsdatei, die später in KNIME authentifiziert Google Analytics.

Schritt 2: Aktivieren von Google Analytics API

Lassen Sie uns jetzt **das Projekt konfigurieren** dass wir auf Google Cloud erstellt haben. Diese Projekte wie ein **Repository** oder Ordner, wo wir beginnen können, unsere Anwendung auf Google Cloud-Plattform und tun **keine vorbestellte Funktionalität haben**.

Die Anwendung, die wir **Aktivieren** für dieses Projekt **API von Google Analytics**. mit Diese API, wir werden in der Lage, auf und abfragen Web-Verkehrsdaten gesammelt durch Tracking Cookies.

Dazu wählen wir in der oberen Leiste der Projektansicht „**APIs und Services aktivieren**“ „ wir geben “**Google Analytics**” in der Suchleiste und aktivieren:

Das Projekt ist jetzt eingestellt und wir sind bereit zu starten

Verwendung die **Funktionalität und**

Anwendungen von Google Analytics Zugriff auf Webstatistiken sowie die Anmeldeinformationen (d.h., Service-Konto, private Schlüssel, etc.), die wir brauchen, um den Prozess abzuschließen.

The figure consists of three vertically stacked screenshots from the Google Cloud Platform (GCP) web interface.
 1. The top screenshot shows the 'APIs & Services' dashboard. A yellow arrow points from the 'Traffic' section on the left to the '+ ENABLE APIs AND SERVICES' button at the top right.
 2. The middle screenshot shows the 'API Library'. A yellow arrow points from the 'Google Analytics Reporting API' entry to the search bar containing 'google analytics'.
 3. The bottom screenshot shows the 'API Service Details' page for the 'Google Analytics API'. A yellow box highlights the service name 'analytics.googleapis.com'.
 Arrows indicate the flow from enabling the API to searching for it, and finally to viewing its details.

Aktivierung der Google Analytics API.

Schritt 3: Zugriffsberechtigungen einstellen

In diesem Schritt werden wir Zugriffsberechtigungen festlegen, damit unser Projekt die Verbindung zum Verkehr ermöglicht Webdaten.

- Wenn wir die Administratoren des Google Analytics-Kontos sind:

Wir müssen auf das Google Analytics-Konto oder -Eigentum zugreifen und das neue erstelltes Service-Konto (verwenden Sie die mit ihm verbundene E-Mail-Adresse) in der Option „Benutzermanagement“.

This screenshot shows the 'IAM & Admin' section of the GCP console.
 On the left sidebar, 'Service Accounts' is selected.
 In the main area, a 'Create service account' dialog is open.
 Under 'Service account details', there is a checked checkbox for 'Grant this service account access to project (optional)'.
 Below that, another checkbox is checked for 'Grant users access to this service account (optional)'.
 A yellow arrow points from this section to the 'Permissions' panel on the right.
 The 'Permissions' panel shows a table with one row under 'Owner (2)', where the email 'blog-basecer0-test-knime' is listed with a yellow box around it.
 A yellow arrow also points from this row to the 'Role / Principal' dropdown, which is set to 'Owner (2)'.

Wie identifiziert man die E-Mail-Adresse, die mit dem Service-Account verknüpft ist

Die Abbildung oben zeigt, wie man die
Servicekonto.

E-Mail-Adresse zugeordnet

Zu autorisierte Benutzer hinzufügen im Benutzerverwaltungsmenü von Google Analytics, wir Klicken Sie auf „Benutzermanagement“ aus dem Konto oder Eigentum wollen wir gewähren Zugang zu. Als nächstes fügen wir einen neuen Benutzer mit der Service-Account-E-Mail hinzu und setzen die Berechtigungen zu „Lesen und Analysieren“ (weiterlesen über [Benutzer auf Google verwalten Analyse](#))

- **Wenn wir nicht die Administratoren des Google Analytics-Kontos sind:**

Sollte dies der Fall sein, müssen wir den Administrator des Google anfordern Analytics-Konto, um uns Zugang zu gewähren.

Über sein Authentifizierungs- und Identifikationssystem garantiert Google, dass kein Web Informationen können aufgerufen werden, ohne dass die erforderlichen Berechtigungen erteilt werden vom Administrator.

Schritt 4: Konfigurieren von KNIME Analytics Platform und Query Data

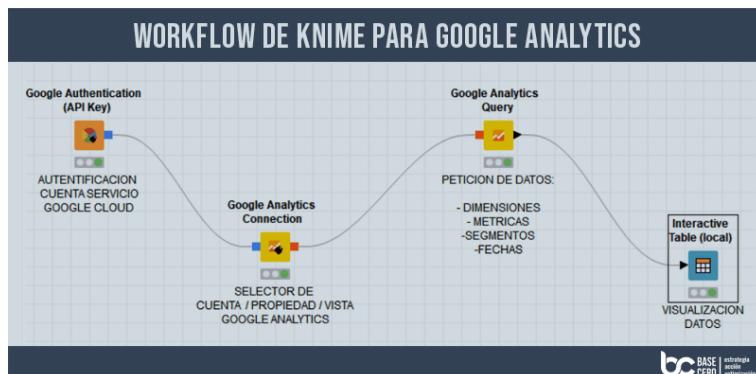
Wir sind nun bereit, mit der KNIME Analytics Platform eine Verbindung zu Google Analytics herzustellen. Zu das tun, müssen wir das Tool konfigurieren, um auf das Google Cloud-Projekt zuzugreifen und zu validieren die Berechtigungen.

Das Abfragen von Google Analytics zum Abrufen von Web-Verkehrsdaten von KNIME ist sehr einfach. Wir nur 3 Knoten benötigen:

ANH ANG Authenticate auf das Service-Konto, das wir verwenden.

2. Zugriff auf das Google Analytics-Konto und die Eigenschaft(-ies), die autorisiert haben Zugang.
3. Abfrage von Google Analytics und Abruf von Dimensionen, Metriken, Segmenten usw.

Die folgende Abbildung zeigt **Beispiel KNIME Workflow** eine Verbindung zu Google Analytics herstellen.



Dieser Beispiel-Workflow verbindet sich mit Google Analytics.

Als nächstes werden wir eine Schritt für Schritt Beschreibung der

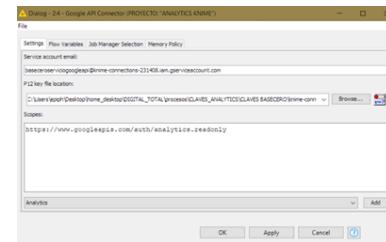
Knotenkonfigurationen in KNIME

ANH ANG [Google Authentication \(API Key\)](#)

Verbindung mit Google API.

Wir geben die E-Mail-Adresse des **Service Rechnung**, das lokale Verzeichnis der Sicherheitsdatei mit **geheime Schlüssel**, und definieren Sie die ^{Anwendungsbereich} **erreich** (Auswahl) „Google Analyse Verbindung (Lesen)“ aus dem Dropdown-Menü. Mit diesen Konfigurationen, wir können eine neue Sitzung in das Google-Projekt und die Verbindung zu Google Analytics API.

Node:



Das Konfigurationsfenster des Google Authentication (API Key) Knoten.

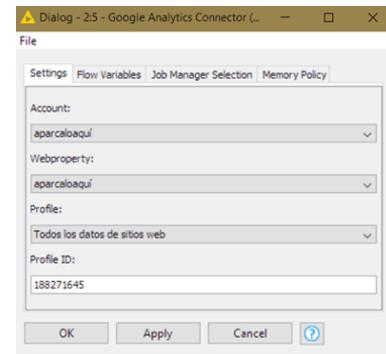
2. [Google Analytics Verbindung](#)

Node: Zugriff

Google Analytics.

Hier müssen wir nur die **Rechnung, Immobilien und Google Analytics** Blick wir würden wie zu fragen. Nur Konten und Eigenschaften welches das Google Cloud-Projekt Zugriff hat werden angezeigt.

Node: Zugriff



Das Konfigurationsfenster des Google Analytics Verbindungsknoten.

3. [Google Analytics Abfrage](#)

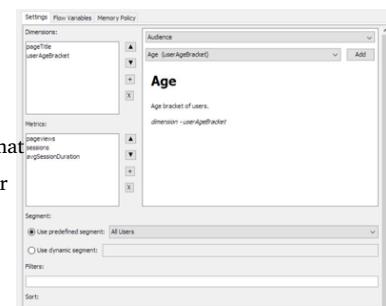
node:

Google Analytics API.

Mit diesem Knoten können wir Google Analytics abfragen API und erhalten Sie die Daten, die wir benötigen. Der Knoten hat eine Schnittstelle zum Erstellen von Abfragen und ermöglicht Benutzer zur Suche und Auswahl verfügbarer Abmessungen und Metriken.

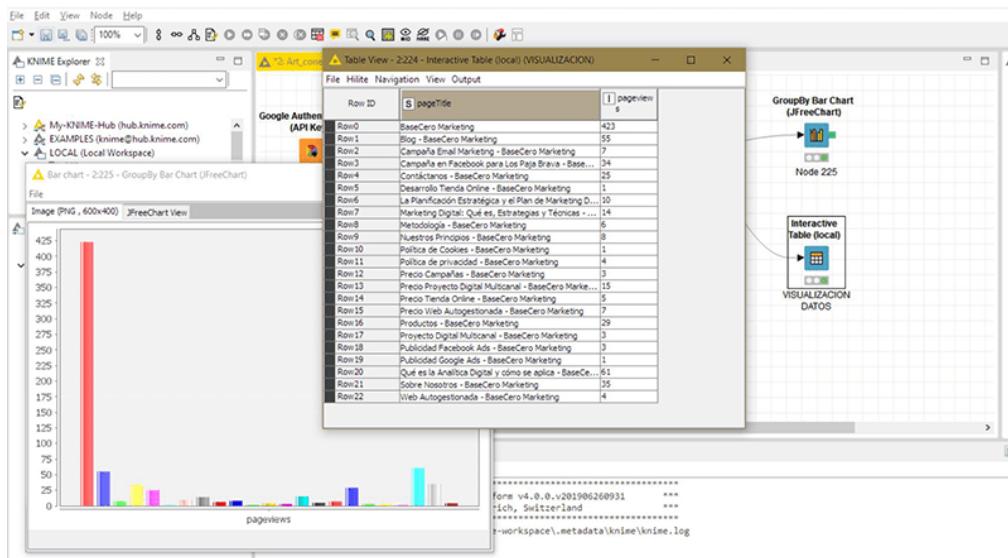
Um den vollen Vorteil dieses Knotens zu nehmen, ist es ratsam, die **Grundkonzepte Google Analytics API** : Abmessungen, Metriken, Segmente, Termine, etc.

Grundkonzepte



Das Konfigurationsfenster des Google Analytics Query Node.

Wir sind jetzt alle bereit und bereit zu beginnen, Erkenntnisse aus Web-Verkehrsdaten zu extrahieren, Vorverarbeitung und Zusammenarbeit mit KNIME, wie jede andere Datenquelle im digitalen Analytik:



Nach erfolgreicher Anbindung an die Google Analytics API können Erkenntnisse aus Web-Verkehrsdaten extrahiert und dann mit KNIME weiterverarbeitet werden.

Stellen Sie klare Ziele für Ihre Web Analytics fest

Mit diesem Artikel wollen wir die [Anwendung digitaler Analytik](#) und frei und Open-Source-Tools wie KNIME Analytics Platform.

Wann immer wir Webdaten analysieren und verstehen wollen, ist es von entscheidender Bedeutung, eine klar [Strategie und Ziele der Analyse](#) . Mit anderen Worten, wir müssen verstehen Was? Aspekte von Webdaten, die wir messen möchten, warum, in welchem Zeitbereich, und für welche Benutzer.

Der Rest ist Nur die Daten zu ordnen und Erkenntnisse zu extrahieren. Das ist, wo KNIME hilft Prozessen zu optimieren, die Implementierungszeit zu reduzieren und leistungsstark zu bauen schlüssellose Analyselösungen.

Dieser Artikel wurde ursprünglich auf Spanisch veröffentlicht [Marketing und Marketing](#) und wurde übersetzt in Englisch von Roberto Cadili. Die Originalversion finden Sie hier [Hier](#).

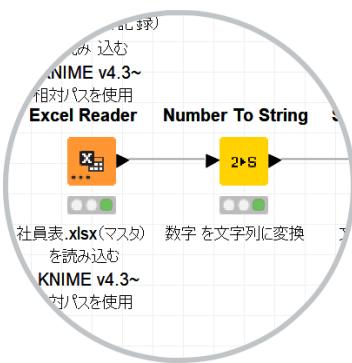


Makkyn wurde nominiert Beitrag des Monat für Juli 2021. Er wurde für seine Aktivitäten ausgezeichnet innerhalb der KNIME Community und für seinen zahlreichen Blog Beiträge und Tutorials auf Japanisch sein es für Anfänger, Intermediate oder fortgeschrittene KNIME-Nutzer. Auf seinem Blog <https://digitalization.hatenablog.jp/> er schreibt über viele Themen wie: String Manipulation, Daten Zugriff, Datenvisualisierung und Zeitreihenanalyse.

Der Inhalt auf seinem Blog ist auf Japanisch geschrieben, und es richtet sich an die japanischen Daten Wissenschaftsgemeinschaft. Er teilt auch seine Wissen aktiv auf Twitter. Achten Sie darauf, zu überprüfen aus ihm [Tweet Geschichte](#):

makkynm ist ein glühender KNIME-Befürworter, und er ist leidenschaftlich, anderen zu helfen, sei es KNIME oder nicht. Er liebt es, anonym zu bleiben, deshalb ist er schickte ihn Twitter Avatar, um ihn hier zu vertreten.

Besuchen Sie makkynm's [Raum auf dem KNIME Hub](#) oder [Profil im KNIME Forum](#) (Hub/Forum Griff:
Makkyn)



Daten aus Datenbanken in KNIME Analytics lesen

Plattform

Verwendung der Microsoft Access Connector Node als Beispiel

Autor: makkynm; Übersetzt von: Elisabeth Richter mit Hilfe von KI



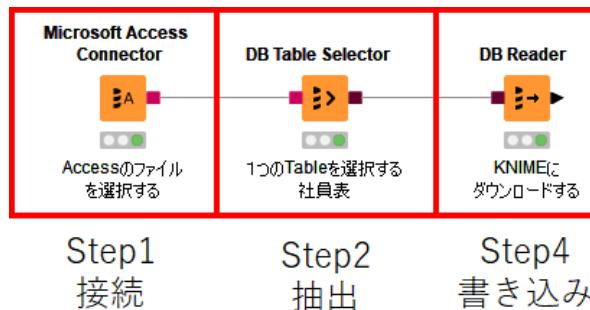
Einleitung

In diesem Artikel werde ich erklären, wie man Daten aus einer Datenbank anschließt und aufruft KNIME Analyseplattform. Ich werde die Datenbankknoten verwenden, die ich Ihnen als ein zum letzten Mal. Wenn Sie diese Datenbankknoten nicht kennen, lesen Sie bitte die [Vorheriger Artikel](#) bevor Sie weiter lesen.

Dieser Artikel konzentriert sich auf den Microsoft Access Connector Knoten, DB Table Selector node und DB Reader node.

Für den Zweck dieses Artikels bin ich mit Microsoft Access verbunden, was ist wahrscheinlich die bekannteste Datenbank mit dem Microsoft Access Connector-Knoten. Die Anbindung an eine Datenbank in der KNIME Analytics Platform erfolgt über eine eigene Verbindungsknoten. In diesem Beispiel ist es jedoch der Microsoft Access Connector-Knoten. wenn Sie Daten aus einer anderen Datenbank verbinden und aufrufen möchten, das einzige, was Sie muss der Microsoft Connector-Knoten durch einen anderen Connector-Knoten ersetzt werden, zum Beispiel einen PostgreSQL Connector-Knoten.

KNIME Unterstützung – makkynm
Daten aus Datenbanken in der KNIME Analytics Platform lesen



Dieser Beispiel-Workflow zeigt, wie man mit und Zugriff verbindet Daten aus einer Datenbank in der KNIME Analytics Platform. Dafür Beispielsweise wird eine Verbindung zu Microsoft hergestellt.

Im Falle des Microsoft Access Connector-Knotens gibt es keine Notwendigkeit zu installieren und konfigurieren Sie weitere Treiber. Für andere Datenbanken müssen Sie jedoch installieren Treiber separat und zusätzliche Einstellungen in den Einstellungen vornehmen.

Ich werde die drei Knoten im folgenden erklären, aber keiner von ihnen benötigt fast alle Einstellungen.

Sie können den Workflow aus meinem KNIME Community Hub-Raum herunterladen:

<http://kni.me/w/Wag-UjltOfjCTEn0>

Dies sind die Schritte, die ich Ihnen zeigen werde:

- ANH
ANG Verbinden Sie mit der Datenbank: Microsoft Access Connector node
2. Wählen Sie die Tabelle aus, aus der Sie die Daten lesen möchten: DB Table Selector node
 3. Abrufen der Daten aus der Datenbank in eine KNIME-Datentabelle: Zu Wunsch Daten: DB Leseknoten

Row ID	社員番号	名前	部署ID	部署	出身	生年月日	勤務地	入社日	ID
Row0	A001	田中 光一	620,000,554	経務	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

Lesedaten, die in einer Datenbank in die KNIME Analytics Platform gespeichert werden.

KNIME Unterstützung – makkynm

Daten aus Datenbanken in der KNIME Analytics Platform lesen

Mit den drei Knoten im oben dargestellten Workflow können Sie Daten lesen, die in einer Datenbank in der KNIME Analytics Platform gespeichert. Diese Daten können dann weiter verwendet werden in KNIME.

In diesem Beispiel werde ich den Mitarbeitertisch und die Teilnahmetabelle lesen, die ich habe in anderen Artikeln vor.

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	

Mitarbeitertabelle

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	

Achtung Tisch.

Der Workflow

Im Bild unten sehen Sie den Workflow. Im Folgenden werde ich die drei erklären Knoten nach dem anderen. In dem vorherigen Artikel, den ich zuvor erwähnte, der Workflow bestand aus vier Schritten (Verbindung, Extraktion, Umwandlung, Schreiben). Dafür jedoch Artikel I weggelassen Schritt 3 (Umwandlung) und wird daher ohne die Verarbeitung der Daten. Daher sind die drei Schritte:

Schritt 1: Mit dem Microsoft Access Connector-Knoten wird die Datenbank erstellt
Verbindung

Schritt 2: Mit dem DB Table Selector-Knoten kann die gewünschte Tabelle in der Datenbank ausgewählt

Schritt 4: Schließlich greift der DB Reader-Knoten auf die im vorherigen Knoten ausgewählten Daten zu und sie in eine KNIME-Datentabelle abrufen

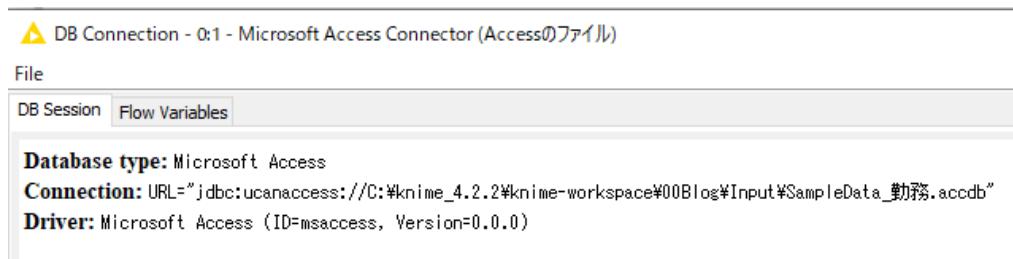


Dieser Beispiel-Workflow zeigt, wie man mit und Zugriff verbindet
Daten aus einer Datenbank in der KNIME Analytics Platform. Dafür
Beispielsweise wird eine Verbindung zu Microsoft hergestellt.

Schritt 1: Wie verwenden Sie den Microsoft Access Connector-Knoten

Dieser Knoten wird verwendet, um mit einer Microsoft Access-Datenbank zu verbinden. Es schafft nur
eine Verbindung zur Datenbank, aber keine Tabelle angibt.

Sie können bestätigen, dass die Datenbank eine Verbindung erfolgreich von rechts aufgebaut hat
Klicken Sie auf den Anschlussknoten und wählen Sie "DB Connection". Hier können Sie die
Datenbanktyp, Datenbank-URL (hier der Dateipfad) und der Treiber in
Verwendung.

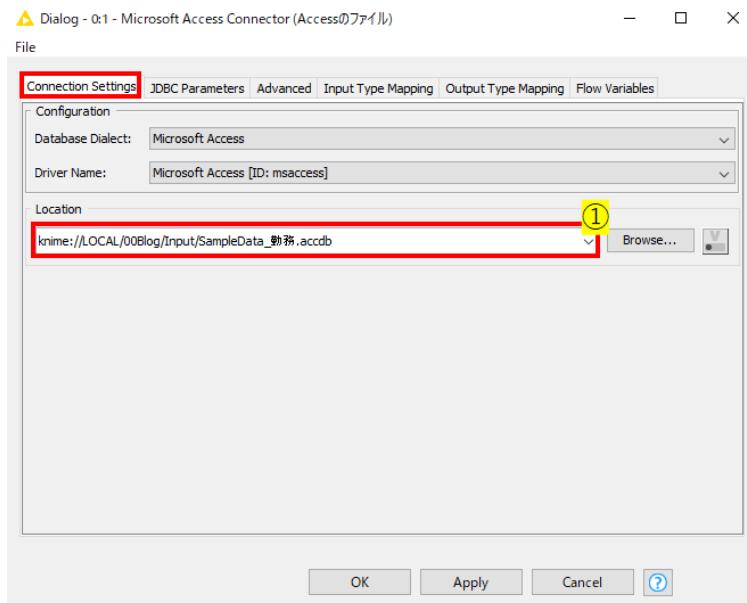


Überprüfen Sie die Datenbankverbindung in der Registerkarte "DB Session" der Ausgabe des Microsoft Access Connector
Knoten.

Um den Knoten zu konfigurieren, müssen Sie nur den Dateipfad des
Datenbank.

In dem oben dargestellten Konfigurationsdialog nutze ich einen relativen Dateipfad. Wenn Sie
gerne mehr über relative Pfade erfahren, siehe diesen Artikel [Datenbankknoten](#)
und KNIME als ETL-Tool

KNIME Unterstützung – makkynm
Daten aus Datenbanken in der KNIME Analytics Platform lesen



Der Konfigurationsdialog des Microsoft Access Connector Knotens.

Schritt 2: Wie verwendet man den DB Table Selector-Knoten

Dieser Knoten nimmt eine DB-Verbindung als Eingabe und ermöglicht es Ihnen, eine Tabelle oder Ansicht von innerhalb der angeschlossenen Datenbank. Sie können eine Tabelle in der Datenbank angeben. In diesem Zum Beispiel wähle ich die Mitarbeitertabelle aus.

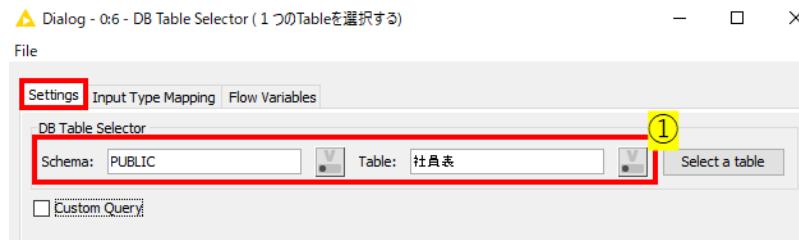
Sie können die aktuell ausgewählte Tabelle durch einen Rechtsklick auf den Knoten und die Auswahl „DB Daten“. Dann klicken Sie auf „Kache Nr. der Zeilen“. An dieser Stelle ist die Tabelle jedoch noch nicht eine KNIME Datentabelle.

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					

Zeigen Sie die Tabelle innerhalb der angeschlossenen Datenbank in der Registerkarte "Tabellenvorschau" vom Ausgang der DB-Tabelle Selector node.

Um diesen Knoten zu konfigurieren, geben Sie einfach die Tabelle an, die Sie in der Konfiguration importieren möchten Dialog des Knotens. „Select a table“ ermöglicht es Ihnen, eine Tabelle aus einer Liste auszuwählen.

KNIME Unterstützung – makkynm
Daten aus Datenbanken in der KNIME Analytics Platform lesen



Der Konfigurationsdialog des DB-Tabellenauswahlknotens.

Schritt 4: Wie man den DB Reader-Knoten verwendet

Dieser Knoten führt die Eingabeabfrage in der Datenbank aus und ruft das Ergebnis in eine KNIME Datentabelle. Dadurch stehen die Daten am Ausgangsport des Knotens zur Verfügung, ähnlich wie in den Anfänger- und Zwischenklassen. Mit anderen Worten:
nun können die Daten in der gleichen Weise behandelt werden wie beispielsweise, wenn eine Excel-Datei mit dem Excel Reader-Knoten gelesen werden.

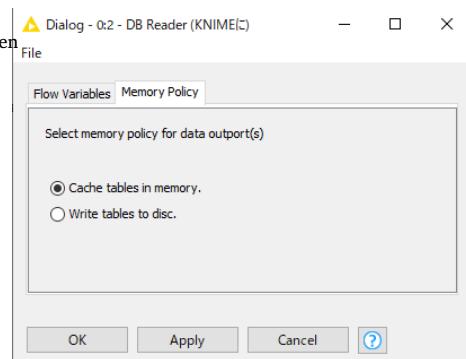
Wenn die Datenmenge in der Datenbank groß ist, empfiehlt es sich, sich zunächst zu verjüngen die Daten mit einem der DB-Knoten (z.B. dem DB-Row-Filter-Knoten) und dann die DB Leseknoten.

Klicken Sie mit der rechten Maustaste auf den Knoten und wählen Sie „KNIME“
Datentabelle“. Es sieht ähnlich wie die Ausgabe des DB-Tabellenauswahl-Knotens vor, jedoch, da die Daten jetzt in einer KNIME-Datentabelle abgerufen werden, die "Cache no. of Zeilen" Knopf wird nicht mehr angezeigt. Zusätzlich werden die Zeilen-IDs (Row ID Spalte) zugeordnet. automatisch.

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						

Die Daten sind nun als KNIME-Datentabelle am Ausgang des DB Reader-Knotens zugänglich.

Um diesen Knoten zu konfigurieren, keine speziellen Einstellungen
werden benötigt!



Der Konfigurationsdialog des DB Reader Knotens.

Ein schnelles Wort

Microsoft Access Connector node - Andere Optionen

In diesem Beispiel wurde bei der Konfiguration des Microsoft Access nur ein Schritt benötigt. Verbindungsknoten. Es gibt jedoch andere Optionen, und ich möchte sie kurz erläutern. Hier.

Verbindungseinstellungen: Konfigurieren der Treibereinstellungen

Wählen Sie einen Treiber aus. Dieses Mal wurde der Treiber automatisch ausgewählt und keine anderen Treiber wurde gezeigt, aber wenn es nicht der Treiber ist, den Sie verwenden möchten, können Sie es hier auswählen. Wenn kein Fahrer ist verfügbar, Sie müssen es in den Einstellungen (KNIME) setzen → Datenbanken).

Referenzlink: [KNIME Leitfaden für die Erweiterung](#)

JDBC Parameter: Erweiterte Treibereinstellungen

In der Registerkarte "JDBC Parameter" im Konfigurationsfenster können Sie JDBC-Treiber einstellen Parameter. Wenn beispielsweise ein Passwort benötigt wird, geben Sie es hier an. Auch bei Verwendung eines Datenbank mit einer großen Datengröße, wird empfohlen, den "Speicher" Punkt auf "false" zu setzen.

Referenzlink: [UCanAccess-A Pure Java JDBC Treiber für den Zugriff](#)

Fortgeschritten: Erweiterte Verbindungseinstellungen

In der Registerkarte "Erweitert" im Konfigurationsfenster, detailliertere Verbindungseinstellungen kann z.B. Timeout-Einstellungen, Wiederherstellen Datenbank Option, etc. angegeben werden.

Eingangstyp Mapping / Ausgangstyp Mapping: Datentyp Mapping

Mit der Registerkarte "Eingabetyp Mapping" im Konfigurationsfenster können Sie Regeln definieren von Datenbanktypen zu KNIME-Typen und der Registerkarte „Output Type Mapping“ auf Karte Regeln von KNIME-Typen zu Datenbanktypen definieren.

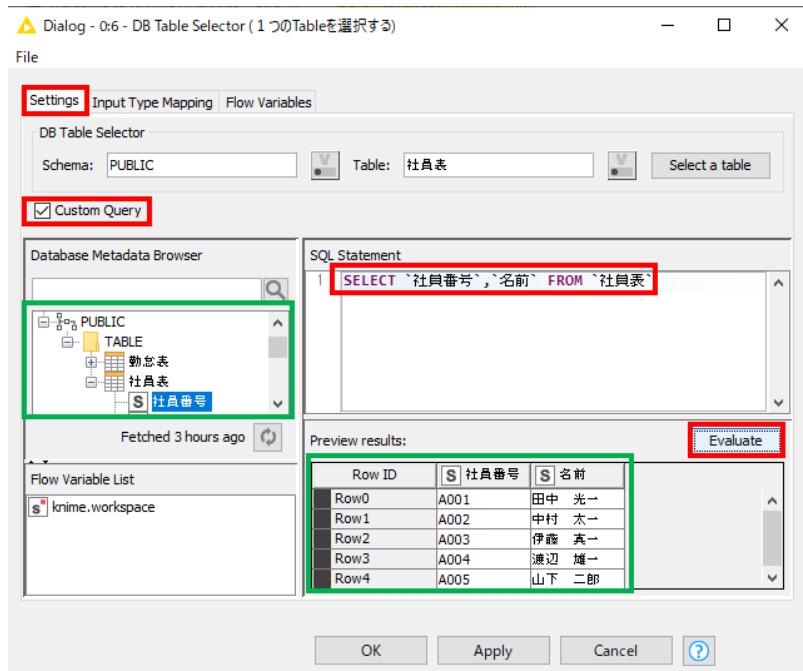
DB Tabelle Selector - Weitere Optionen

Benutzerdefinierte Abfrage: Wählen Sie Tabellenspalten mit SQL-Anweisungen

Überprüfen der „Kundenspezifischen Abfrage“ in der Registerkarte „Einstellungen“ des Konfigurationsfensters ermöglicht oder deaktiviert benutzerdefinierte Abfrage für die gewählte Tabelle. Neben Tischen können Sie auch Spalten angeben.

Mit dem „Database Metadata Browser“ auf der linken Seite können Sie die Tabellenfelder durchsuchen und die SQL-Anweisung angeben.

Wenn Sie „Auswerten“ drücken, können Sie Ihre SQL-Anweisung vorhersehen und dadurch können Sie überprüfen, ob die SQL-Anweisung korrekt ist.



Die Registerkarte "Einstellungen" im Konfigurationsdialog des DB-Tabellenauswahlknotens.

Schlussfolgerung

In diesem Artikel habe ich den Microsoft Access Connector-Knoten als Beispiel verwendet, um erklären, wie man eine Datenbank in der KNIME Analytics Platform verbindet. Wenn Sie wollen verbinden Sie mit einer anderen Art von Datenbank, können Sie einfach den Connectorknoten ändern.

Während Sie die Operationen durchführen können, die ich Ihnen oben in einer Datenbank gezeigt habe, verwenden Sie KNIME kann nicht nur dasselbe tun, sondern hat auch den Vorteil, transparenter zu sein.

Oft sind Datenbanken Blackboxes und die Logik hinter bestimmten Datenbankoperationen ist nicht immer klar. in Erwägung nachstehender Gründe: Analytics Platform ermöglicht die Visualisierung der Logik von bestimmten Operationen.

Für alle, die mit einem solchen „Black-Box-Zugriffswerkzeug“ arbeiten, würde ich sehr empfehlen KNIME Analytics Platform a try! Wir sehen uns nächstes Mal!

Referenzlinks

- NodePit:

[Microsoft Access Connector — NodePit](#)

[DB Tabellenauswahl — NodePit](#)

[DB Reader — NodePit](#)

- KNIME Beispiel Workflows:

[Microsoft Access Connector – KNIME Community Hub](#)

[DB Table Selector – KNIME Community Hub](#)

[DB Reader – KNIME Community Hub](#)

Dieser Artikel wurde ursprünglich auf Japanisch veröffentlicht [makkynms Blog](#) und wurde in Englisch von KNIME. Die Originalversion finden Sie hier [Hier.](#)



[Ignacio Perez](#) wurde nominiert Beitrag des Monat für September 2021. Er wurde für seine Arbeit ausgezeichnet Revolving um die Spanischsprachige KNIME Gemeinschaft und für eine respektierte Referenz. Er etablierte und aktuelle Übersichten

Fragen zur [KNIME Forum](#), regelmäßig veranstaltet Kurse in Spanisch, und er übersetzte auch das e-Book “Von Excel zu KNIME“ auf Spanisch: „[De Excel a KNIME Analytics](#)

[Plattform](#) „Außerdem unterhält er eine [YouTube Kanal](#) wo er KNIME erklärt an die spanischsprachige Gemeinschaft.

Ignacio hält einen Doktortitel in Angewandter Mathematik der Universität Lyon. Er hat über 20 Jahre Erfahrung als Data Analyst in verschiedenen Branchen und ist KNIME seit über 10 Jahren nutzen. Er ist auch der Gründer und Eigentümer [IQuartil](#), ein in 2000 und der Entwicklung gewidmet Beratungsprojekte in der statistischen Analyse, Stichproben, Prognose, Analysen Marketing, Operationen Optimierung, Risikopositionen Verwaltung und Finanzen Planungsanalyse.

[Besuchen Sie Ignacio's Raum auf dem KNIME Hub](#) oder [Profil im KNIME Forum](#) ([Hub/Forum Griff:](#) Iperez)



Der Pionier der KNIME Community en español

Autor: Rosaria Silipo

in Hola data-nautas de la ciencia de datos sin código!

Ignacio ist der Pionier der KNIME Community en español. Am Dezember 2020 begann er

die [KNIME Forum en Español](#) „wo er die meisten Fragen beantwortet“

2021 übersetzte er zusammen mit seinen Kollegen das Buch „

[Von Excel nach KNIME](#)

„

en español; und bis jetzt gibt er regelmäßige Präsentationen, wie er erfolgreich angewendet KNIME in Englisch und en español.

Die folgende Abbildung zeigt den ersten Beitrag im KNIME Forum en Español. Hinweis:

brevity der Ankündigung, etwas, das ein ausgeprägtes Muster aller Ignacio-Posts ist.

Es ist nicht selten, eine seiner auflösenden Antworten mit nur wenigen Worten und einem Workflow zu finden auf dem KNIME Forum. Lassen Sie uns sagen, dass er lieber mit Workflows sprechen lieber als mit Worten.

Nuevo foro en Español

 Community Groups  KNIME en Español



iperez

Dec '20

Aquí estaremos discutiendo y compartiendo sobre KNIME en español!!!. Bienvenidos los aportes

11 ...

created	last reply	1	598	1	11	
 Dec '20	 Jan '21	reply	views	user	likes	

Der erste Beitrag im KNIME Forum en Español.

Wir möchten hier ein Interview von Cynthia (KNIME Interviewer) melden an Arturo (der Kunde), und Ignacio (der technische Enabler) als Beispiel der eine erfolgreiche Unterstützung, die ein Beraterunternehmen den Bedürfnissen des Kunden gerecht werden kann.

[Dieser Artikel basiert auf der Präsentation „](#) [Echtzeit-Informationen zur Produktqualität](#) „

„

von KNIME Herbst Summit 2020, jetzt auf YouTube.

Cynthia: [wir haben](#) [Ignacio Perez](#) [von](#) [In den Warenkorb](#) [und](#) [Arturo Boquin](#) [von](#) [Dinant](#) [und wir sind](#) über Echtzeit-Informationen zur Produktqualität sprechen. Ignacio, sag uns ein bisschen über iQuartil bitte.

Ignacio: iQuartil ist ein KNIME-Trusted Partnerunternehmen mit Sitz in Kolumbien analytics services in Lateinamerika. Wir begleiten unsere Kunden in verschiedenen Branchen beide in der Entwicklung und Herstellung von analytischen Lösungen sowie in der Ausbildung.

Einer unserer Kunden ist Dinant. Dinant ist ein zentralamerikanisches Unternehmen aus Honduras, das arbeitet in der Konsumgüterindustrie verschiedene Arten von Produkten, wie verpackte Lebensmittel, und zu Hause und persönliche Betreuung Produkte. Dinants Marken sind in Mittelamerika, in den Vereinigten Staaten und in der Karibik. Arturo ist verantwortlich für ein Projekt, das wir bei Dinant entwickelt haben. Wir haben hat mit ihnen eine Lösung geschaffen, die Arturo mit uns teilen wird.

Arturo: Ich freue mich sehr, Ihnen die Lösung zu präsentieren, die wir entwickelt haben mit [KNIME Server](#). Zunächst möchte ich Ihnen vor und nach dem Prozess zeigen, dass wir geändert.

Vorher:

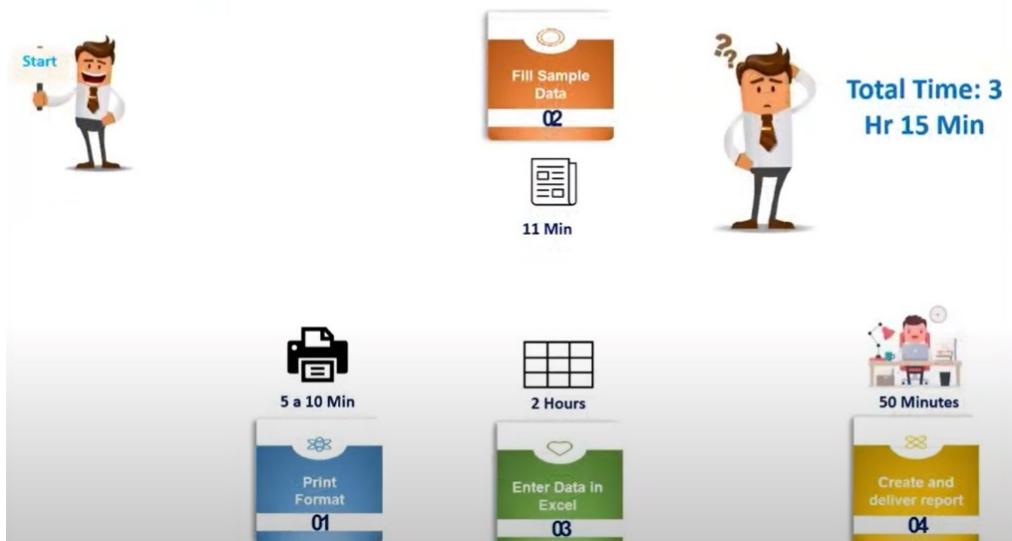
ANH ANG Wir hatten ein Mitglied der Qualitätskontrollabteilung, die in Linie (5-10 min) stand warten, um einen Drucker nur zu verwenden, um viele physische Formate auf Papier zu drucken.

2. Dann hat jemand von der Qualitätskontrolle die Papierdateien mit Daten gefüllt. (11 min)
3. In der Regel warteten sie darauf, viele Papiere zu sammeln, bevor sie weiter zu jemand, der die Daten in Excel digitalisieren würde. (2 Stunden)

1.347
Von
20.12.20
12. X. 21. Viele Tage später würde jemand im selben Team einen Bericht erstellen und liefern um zu sehen, dass etwas schief ging in der Vergangenheit. (50 min)

In der Theorie sollte der Gesamtprozess ca. 3 Stunden und 15 Minuten, aber wir in der Regel auf mehrere Hindernisse, die jeden Teil des Prozesses verzögern.

Before



Prozess vor der Einführung von KNIME Server. Abschlusszeit (theoretisch): ca. 3 Stunden und 15 Minuten.

Nach:

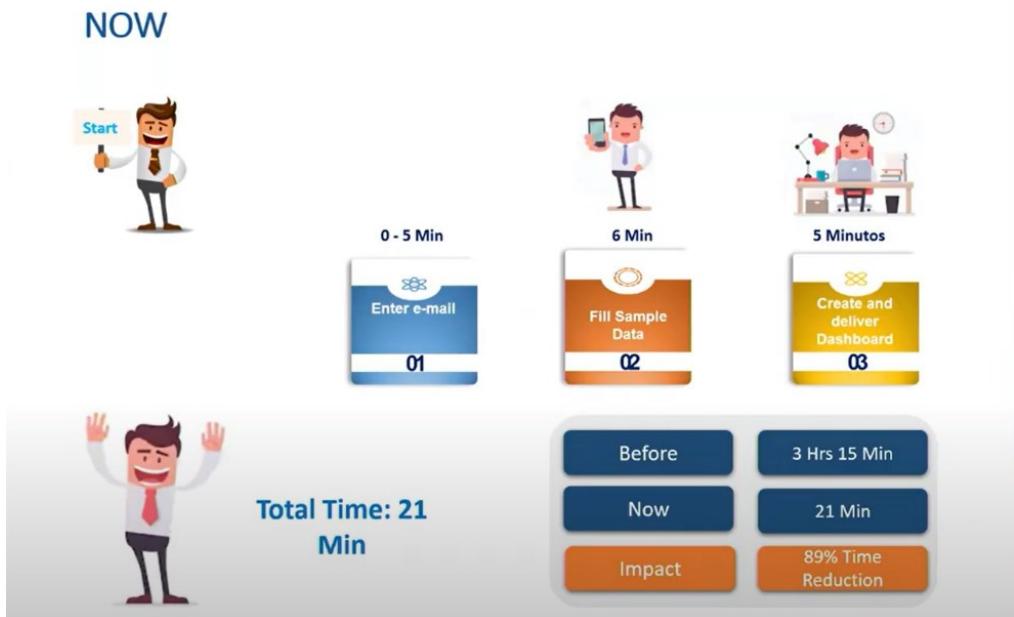
Der Prozess beginnt mit einem Mitglied der Qualitätskontrolle.

ANH ANG Jetzt hat die Person ein Tablet und muss nur die E-Mail eingeben und füllen Sie die Dateien mit Daten.

2. Dann haben wir einen automatisierten Prozess, der das Dashboard für die Berichterstattung aktualisiert alle fünf Minuten.
3. Dank, Menschen in der Fabrik können Ergebnisse fast in Echtzeit sehen und machen Entscheidungen, um negative Änderungen der Produktqualität zu beheben. Wir sind nicht nur die Vergangenheit mehr zu sehen.

Jetzt dauert der Prozess 21 Minuten. Das bedeutet 89% Zeitreduktion und viel Geld in Produktionskosten eingespart. Aus geschäftlicher Sicht ist das die wertvollste Teil des Projekts.

Fragst du dich, wie wir es gemacht haben? Die Antwort lautet: KNIME Server.



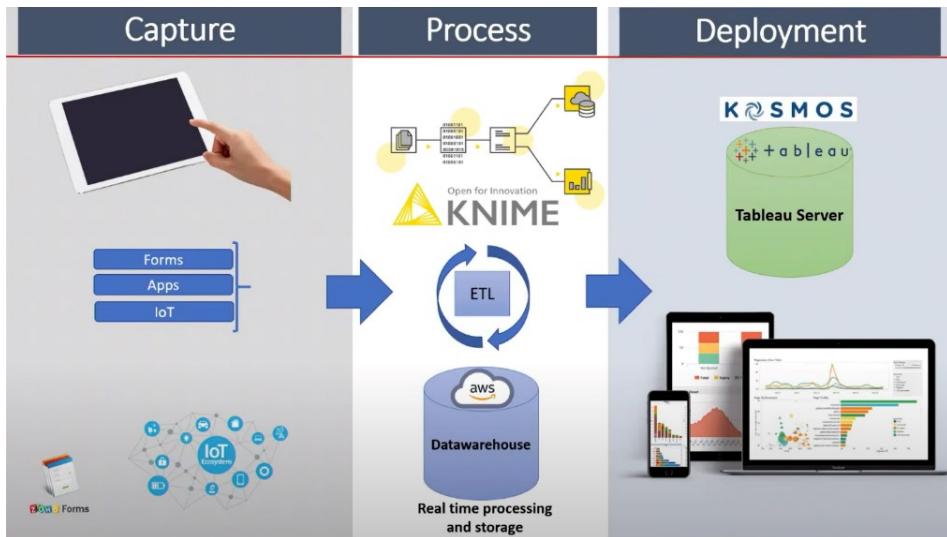
Prozess nach Einführung von KNIME Server. Dauer: 21 Minuten.

Implementierung mit KNIME Server

Wir betrachten den Prozess ganzheitlich.

ANH ANG **Entrückung** Zuerst haben wir die Fangphase. Hier sammeln wir Daten von vielen verschiedenen Quellen, sei es Microsoft Forms, Apps oder IoT-Systeme. Der wichtige Teil ist, dass wir alle diese verschiedenen Quellen mühelos mit KNIME Server verbinden können.

2. **Verfahren**: Sobald die Daten auf dem KNIME Server verfügbar sind, können wir ETL ausführen Operationen, ML-getriebene Anwendungen zur Produktion bewegen und die Ergebnisse in einem Datenlager in Echtzeit.
3. **Bereitstellung**: Schließlich können wir eine TDE-Datei direkt an einen Tableau-Server schreiben, und siehe auf jedem Bildschirm in jedem Teil der Fabrik das Dashboard mit dem betreffenden Informationen, um bessere Entscheidungen zu treffen.



Der 3-stufige Prozess, der von KNIME Server betrieben wird, um Echtzeitinformationen zur Produktqualität zu erhalten.

Cynthia: Es ist erstaunlich zu sehen, wie viel Zeit (und Geld) Sie mit dieser Lösung gespeichert. Während das Endergebnis großartig ist, bin ich sicher, dass es ziemlich einige Arbeit hinter diesem radikaler Prozesswechsel. Was waren die Herausforderungen mit diesem Projekt? hättest du überzeugen das Management des Unternehmens, dass Prozessautomatisierung und Echtzeit Berichterstattung profitiert nicht nur vom Geschäft, sondern auch von den Betreibern ihrer Qualitätskontrolle und Entscheidungsprozess?

Arturo: Das Management zu ändern, ist in der Regel sehr schwierig, und oft ein Hindernis in einigen dieser Projekte. In meiner Erfahrung ist die beste Strategie, eine Prototyplösung in Aktion zu präsentieren, damit die Menschen wirklich sehen können, worum es geht. Für Sie können ein MVP des gesamten Projekts erstellen, in dem Sie Potenzial identifizieren Menschen, die an Bord sein werden und einen Prozess, der über eine Smartphone oder Tablet. Als nächstes können Sie eine einfache Lösung erstellen, in der Sie Daten mischen von Microsoft Forms und SharePoint mit KNIME, und verwenden wrangled Daten, um eine Dashboard mit KNIME oder einem anderen Tool hat KNIME Integrationen für (z.B. Tableau, Power BI, etc.). Auf diese Weise ist es viel einfacher, Ihre Idee zu bekommen, und die Nicht-analytische Menschen in der Organisation sehen den Wert und die Vorteile Ihres Vorschlags.

Cynthia: Hervorragend, danke! Ignacio, möchten Sie etwas hinzufügen?

KNIME Unterstützung – Ignacio Perez
Der Pionier der KNIME Community en español

Ignacio: Ich möchte darauf hinweisen, dass es eine schöne Erfahrung mit Arturo und sein Team. Sie waren bereits schwere und fortgeschrittene Anwender der KNIME Analytics Platform, und in nur ein paar Wochen (literal!) sie konnten diese Lösung entwickeln, die leistungsfähigen Funktionen von KNIME Server sehr schnell zu nutzen und auf hohem Niveau zu präsentieren Manager die Vorteile der Arbeit mit KNIME. Sie haben einen tollen Job gemacht!

Cynthia: Danke, Ignacio und Arturo! Wir freuen uns sehr, dass Sie die Geschichte teilen. Wieder ein weiterer großer Einsatzfall bei der Herstellung und wie KNIME die Werkzeuge für die Betreiber dort, wo sie arbeiten, so dass sie bessere Entscheidungen schnell treffen können.

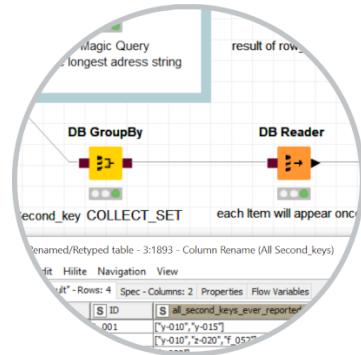


[Markus Lauber](#) wurde nominiert Beitrag des Monat für September 2020. Er wurde für seine [KNIME Forum Thread](#) [Schule der Duplikate – und wie man mit ihnen umgehen](#) wo er verschiedene Möglichkeiten erklärt, wie mit doppelten Werten umgehen und sie beseitigen ohne so wenig potenziell relevante Informationen wie möglich. Das Bild rechts zeigt einen Schnipsel der einem zu diesem Gewinde gehörenden Arbeitsablauf. Markus ist nicht nur ein

regelmäßig auf dem Forum, hält aber auch einen aktiven Raum auf der KNIME Community Hub, wo er viele seiner Workflows teilt, und er ist auch als Sprecher bei unsere Veranstaltungen. Er war also ein sehr aktives und vertrauensvolles Mitglied der KNIME Community schon seit vielen Jahren. Neben KNIME arbeitet er gerne mit R, Python, Spark und H2O.ai integriert sie mit der KNIME Software, insbesondere KNIME Server.

Markus hat über 20 Jahre Erfahrung als Analyst mit dem Fokus auf Data Mining und Big Data. Er ist aktuell ein Senior Data Scientist / Big Data Analytics bei der Deutschen Telekom. Zu seinen Hauptaufgaben gehören: die Leistung und Zuverlässigkeit der Landstriche verbessern und mobile Netzwerke, mit fortschrittlichen Tools auf Big Datenplattformen.

Besuchen Sie Markus [Raum auf dem KNIME Hub](#) oder [Profil in die KNIME Forum](#) (Hub/Forum) Griff:
Mlauber71)



Von H2O.ai AutoML zu Violinlots

Eine beste Sammlung von KNIME Forum Threads von Markus Laube

Autor: Elisabeth Richter

Mit mehr als 1500 Tagen besucht, etwa 11400 Themen betrachtet, und 203 gegebenen Lösungen als von 11.08.2022 ist Markus in der Tat ein sehr aktives Mitglied im KNIME Forum. wir haben folgten Markus seit Jahren und wir haben von vielen seiner letzten Kommentare zum KNIME Forum. Seine Kommentare umfassen alle Phasen der Daten Science Kreation Zyklus: Datenreinigung und Datenaufbereitung, Datenvisualisierung, Maschine Lernen für Vorhersage und Klassifizierung, bis Integration mit anderen Werkzeugen wie R und H2O. In diesem Artikel haben wir seine Reise durch die verschiedenen Schritte der Daten verfolgt Wissenschafts-Erstellungszyklus, indem Sie drei von Markus' populärstes KNIME Forum Threads. Sie können sie auch aus der Kategorie Wissensvermittlung auf unserer KNIME Forum. Im Folgenden zusammenfassen wir diese Best-Of-Markus Gewinde.



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

Schule der Duplikate - und wie man mit Them umgehen

Dies ist ein sehr hilfreicher KNIME Forum Thread, in dem Markus seine besten Praktiken teilt zu einem Thema, das ständig unter Datenwissenschaftlern vorhanden ist: mit Duplikaten zu umgehen. Wann mit Duplikaten zu umgehen, muss man sich über ihre Ziele sicher sein. Obwohl die einfachste Art der Handhabung Duplikate würde einfach loswerden, dies könnte nicht die beste Ansatz als potenziell relevante Informationen könnten verloren gehen.

Jedoch in der Lage, messy Daten in eine aussagekräftige Tabelle mit einer einzigartigen ID zu bringen und ohne zu viel Information zu verlieren, ist eine wertvolle Fähigkeit, die jeder haben sollte.

Dieser Forumsfaden sowie der daran angebrachte Workflow sollen potenziell fördern andere überlegen, was mit ihren Duplikaten zu tun ist und aktiv die Kontrolle übernehmen ihre doppelten Werte.

Die folgende Abbildung zeigt die Daten, mit denen wir uns beschäftigen. Die Datentabelle enthält zwei Identifizierungsspalten, ID und Zweiter Teil, mit beiden Schlüsselspalten, die doppelte Werte enthalten, und weitere Informationen wie die Adresse (Straße und Stadt), den Kaufbetrag (Kauf), sowie das Kaufdatum (Eintrag) und das Datum des letzten Kontakts (Letzter Beitrag). Die Duplikate alle tragen potenziell sinnvolle Informationen und das Ziel ist, die Datentabelle in weniger Zeilen zu reduzieren, ohne irgendwelche (oder so wenig wie möglich) zu verlieren Informationen.

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 Street 10	Boston	21	?
Row7	C_003	z-020	2020-01-02	Little Street 10	Boston	125	?
Row8	C_003	y-010	2020-03-02	Little Street 10	Boston	48	?
Row9	C_003	f-057	2020-04-02	Little Street 10	Boston MA	46	?
Row10	C_003	f-058	2020-04-05	Little Street 10	Boston	96	2020-04-05
Row11	C_003	?	2020-04-05	Little Street 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

Eine Datentabelle mit zwei ID-Spalten, ID und Second_key, die beide Duplikate enthalten, und zusätzliche Informationen wie die Adresse (Straße und Stadt), der Kaufbetrag (Kauf) sowie das Kaufdatum (entry_date) und das Datum des letzten Kontakts (last_contact).

Nun beschreibt Markus in diesem Forum Thread die folgenden Ansätze, wie man umgeht mit Duplikaten:

Option 1: Gruppierung durch Single IDs - Die einfachste Form der Duplikat Entfernung

Verwendung einer Gruppe node ist der einfachste und schnellste Weg, Duplikate von Ihnen zu entfernen. Daten. Gruppe ID und beispielsweise mit mehreren Funktionen aggregiert, Summe (Kauf), Mittel (Kauf), max(entry_date), und max(last_contact).

Obwohl diese Art der doppelten Entfernung ist schnell und einfach, eine Menge potenziell wichtige Informationen werden verworfen.

Option 2: Verwendung von KNIME Duplicate Remover - A Slightly more Sophisticated Ansatz

Mit dem Duplicate Row Filter-Knoten, einem der fortschrittlicheren Filterknoten von KNIME, ermöglicht eine etwas anspruchsvollere doppelte Handhabung. Der Knoten identifiziert Duplikat Zeilen basierend auf einer oder mehreren Spalten und bietet die Möglichkeit, sie entweder zu halten, als „duplicieren“ gekennzeichnet oder zu entfernen.

Im Forum Thread zeigt Markus auch die äquivalente Operation mit SQL.

Diese Methode ist in der Tat eine anspruchsvollere Art der Handhabung Duplikate. Die zweite Schlüsselspalte (Second/key) wird noch ignoriert.

Option 3: Nehmen Sie es auf eine Notch - Betrachten Sie den zweiten Schlüssel

Schließlich zeigt Markus eine Option, bei der auch der zweite Schlüssel berücksichtigt wird.

Er beschreibt verschiedene Wege, wie Zweiter Teil wird gehalten.

Er zeigt auch einen anderen Ansatz mit SQL.

[Lesen Sie den ganzen Faden Schule der Duplikate - und wie man mit ihnen umgeht](#) „auf dem KNIME Forum und den Workflow des gleichen Namens herunterladen vom KNIME Community Hub“

KNIME und R ggplot2 – „die schöne Violine Plot, die es hat alle“

In diesem Forum gibt uns Markus eine Einführung in ggplot2, eine R-Visualisierung

Paket, mit KNIME Analytics Platform. Mit Hilfe der R-Conditional-Violinplot

Bauteil und R Scripting Erweiterung, KNIME Analytics Platform ermöglicht die Erstellung R-basierte Violinfelder mit vielen zusätzlichen Statistiken in einem Diagramm - auch wenn Sie nicht mit R-Code vertraut.

Geige Parzellen sind sehr effektiv, um die Struktur der numerischen Variablen zu zeigen und sie über verschiedene Gruppen vergleichen. Die Breite stellt die Anzahl der Fälle dar, die die Werte auf der y-Achse haben. Es ist weit verbreitet, zum Beispiel als Bevölkerungspyramide.

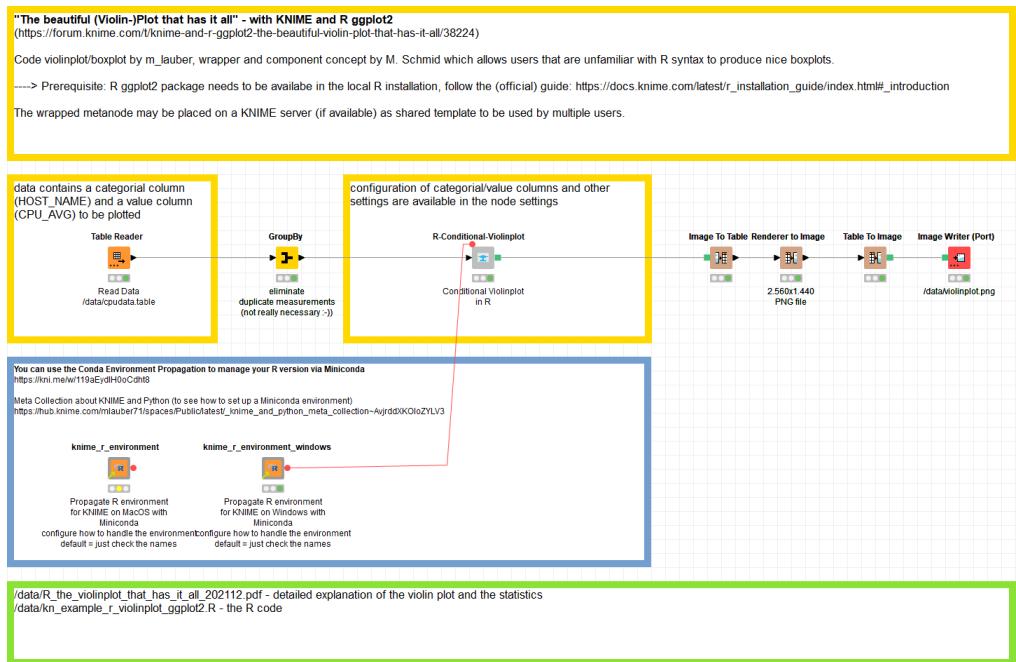
Die Daten, die wir in diesem Anwendungsfall behandeln, zeigen die CPU-Nutzung (in %) in einer Zeit Intervall von einem Monat für zwei verschiedene Server (Server1 und Server2) (siehe Abbildung unten).

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

Die Eingabedaten: CPU-Nutzung (in %) in einem Zeitintervall von einem Monat für zwei verschiedene Server (Server1 und Server2).

Durch die Aufzeichnung dieser Daten in einem Geigenfeld können wir die CPU-Nutzung zwischen die beiden Server durch einen einfachen Vergleich der Formen. Dies erlaubt uns zu beobachten, welche der beiden Server wurde während der Laufzeit des Interesses stärker genutzt

In der Abbildung unterhalb des gesamten Workflows zur Erstellung des Geigendiagramms mit einem R-Skript und Ersparnis als .png eine Datei in den Workflow-Datenbereich angezeigt wird.



Dieser Workflow erstellt ein Violin-Plot und speichert es als .png-Datei.

Die knime_r_environment_windows ein bestimmtes konfigurierbares Bauteil

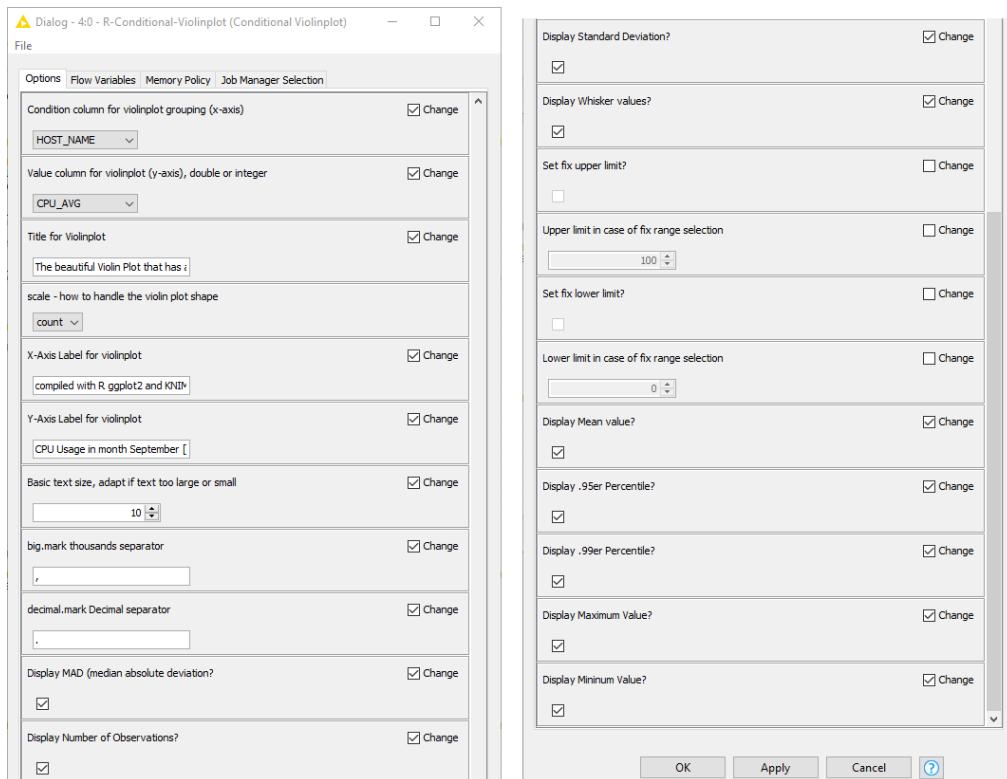
Conda-Umgebung existiert und propagiert die Umgebung an nachgeschaltete Knoten.

Dies geschieht mit Conda Umweltausbreitung Knoten innerhalb der Komponente.

Das Geigenstück wird innerhalb der R-Conditional-Violinplot eine Komponente. Es dauert wie die Datentabelle eingeben und ein .png-Bild des Geigenfeldes ausgeben. Auch die propagierten Die Conda-Umgebung wird den nachgeschalteten Knoten übergeben. Die Komponente weist eine Konfigurationsfenster, in dem Sie die Zustand Spalte und Wert Spalte für das Geigeplot, titel sowie die Achsenbeschriftung und viele andere Einstellungen. Eine grundlegende Konfiguration wäre, wie Sie die Form der Violine zu handhaben wünschen die Anzahl der Artikel. Markus' Standard ist, die Formen proportional zu den Anzahl der Fälle „zählen“. Dabei werden die Flächen proportional zur Anzahl skaliert. von Beobachtungen. Alternativoptionen könnten jedoch „Bereich“ sein, wo alle Geigen haben die gleiche Fläche oder "Breite", wo alle Geigen die gleiche maximale Breite haben.

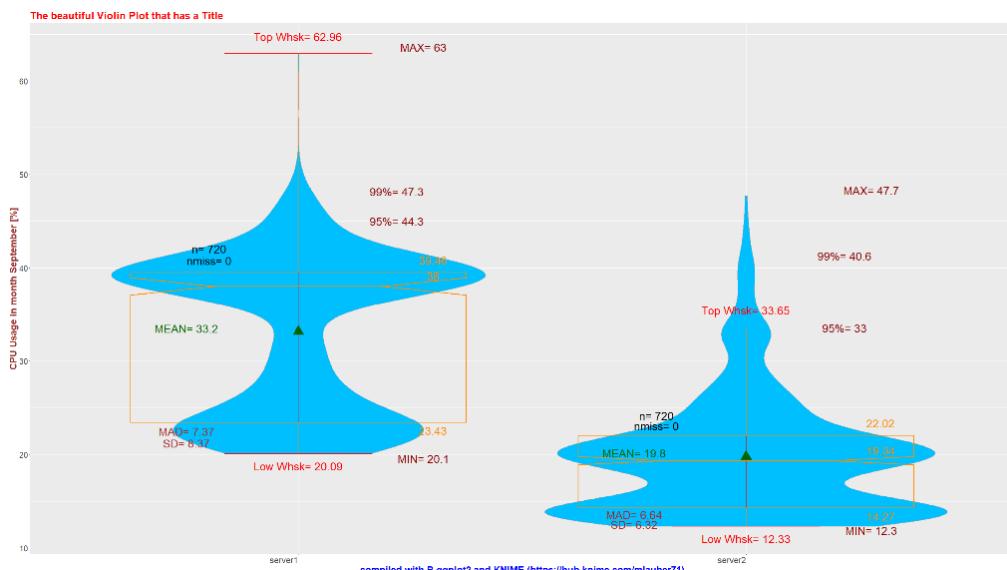
Um tatsächlich das Geigenstück zu schaffen, die R View (Tabelle) Knoten im R-Conditional-Violinplot-Komponente wird verwendet. Im Konfigurationsfenster des Knotens Sie kann Ihr R-Skript hinzufügen. Der Ausgang der R View (Tabelle) Knoten ist ein .png Bild. R befürwortet: Am Ende seines Forums gibt Markus einige Hinweise über den R-Code, zum Beispiel, wie sicherzustellen, dass Etiketten nicht überschneiden alle Informationen.

KNIME Unterstützung – Markus Lauber
Von H2O.ai AutoML zu Violinlots



Das Konfigurationsfenster der R-Conditional-Violinplot-Komponente.

Im folgenden wird das resultierende Geigenstück gezeigt. Von diesem Grundstück ist sichtbar, dass Server1 war stärker während der betrachteten Zeitspanne im Vergleich zu Server2 verwendet. Das ist



Ein Geigendiagramm, das die CPU-Nutzung (in %) in Zeitintervallen über einen Monat (y-Achse) im Vergleich zu zwei Servern zeigt (x-Achse).

weil es mehr Beobachtungen von server1 mit höherer CPU-Nutzung als für Server2. Zusätzliche Informationen können beispielsweise die mittlere CPU-Nutzung für jeder Server, die Mindest- und Maximalwerte oder die Anzahl der fehlenden Werte pro Server.

Lesen Sie den ganzen Faden [KNIME und R ggplot2 – die schöne Violine Plot, die alles hat](#) „ auf dem KNIME Forum and download the workflow of the same name [aus der KNIME-Gemeinschaft](#)

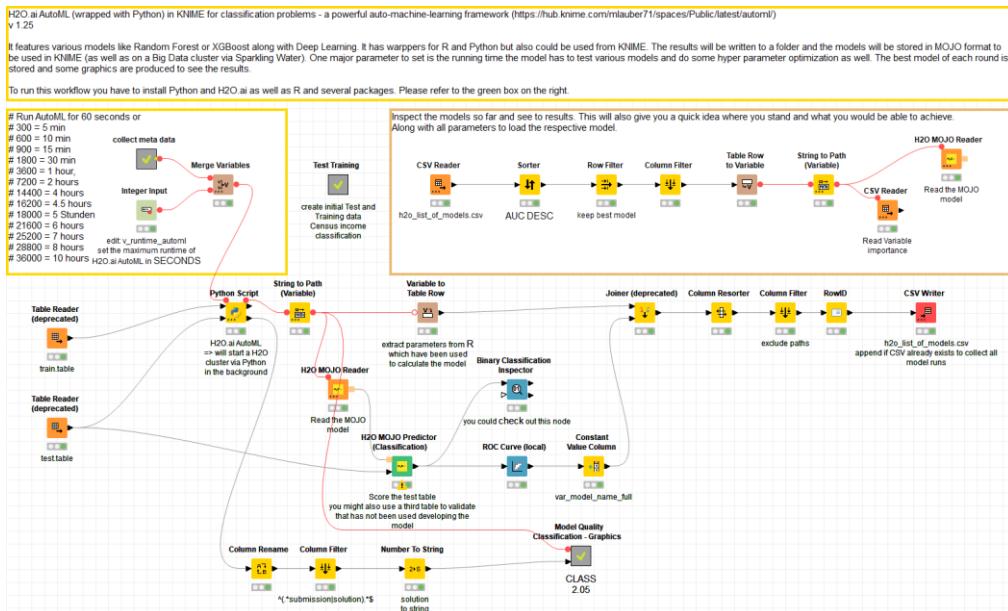
Hubrau

H2O.ai AutoML in KNIME für Klassifizierungsprobleme

Mit 48 Likes ist das Markus' am liebsten auf dem KNIME Forum. In diesem Faden er nimmt das Thema H2O.ai AutoML, ein leistungsfähiges Auto-Maschine-Lerngerüst, und bietet eine Dokumentation zur Verwendung von H2O.ai AutoML in KNIME Analytics Plattform.

AutoML (= Automatisches maschinelles Lernen) ist eine Möglichkeit, Schulungen und Auswertung von maschinellen Lernmodellen. Dies bedeutet, AutoML automatisiert Algorithmus Auswahl, Feature-Generation, Hyperparameter-Tuning, iterative Modellierung und Modell Bewertung. H2O.ai AutoML kann in Schnittstellen wie R und Python verwendet werden, aber es kann auch in der KNIME Analytics Platform mit der H2O-Integration verwendet werden. Es verfügt über verschiedene Modelle wie Random Forest oder XGBoost zusammen mit Deep Learning. Das Ergebnis des AutoML wird in einen Ordner geschrieben und die Modelle werden im MOJO-Format gespeichert, bereit zu sein in verschiedenen KNIME-Workflows wiederverwendet.

In diesem Forum Thread sowie den beigefügten Workflows verwendet Markus Python-Skript, um H2O AutoML und die H2O dedizierten KNIME-Knoten anzuwenden, um ein geschultes Modell in MOJO-Format und um es auf neue Daten anzuwenden. Zusammen mit ihm kommen verschiedene Statistiken und Modellmerkmale, die zur Interpretation des Modells erforderlich sind. Diese werden durch die Python Script-Knoten zu einem .xlsx Datei gespeichert im Workflow-Unterordner unter Modell/validierung/H2O_AutoML_Klassifikation_20210206_1730h.xlsx . Weitere Blätter sind in die Datei mit dem Excel Writer-Knoten in der Model Quality Classification - Grafiken Metanode. Im Folgenden wird Markus' Workflow angezeigt.



Dieser Workflow führt AutoML mit Hilfe von census-income Daten für Schulung und Validierung durch. Der Workflow liest das geschulte Modell im MOJO-Format und wendet es auf neue Daten an, um den Zielwert vorherzusagen. Verschiedene Statistiken und Modellparameter werden in einer .xlsx-Datei gespeichert im Workflow-Unterordner unter model/validate/H2O_AutoML_Classification_20210206_1730h.xlsx.

Die meisten der Magie geschieht innerhalb des Python Script-Knotens. Die zugrunde liegenden Daten der Workflow ist der beliebte [census-income dataset](#) [in Trainingsdaten aufgeteilt](#) (Zug. Tabelle) und Validierungsdaten (Test. Tabelle) Die beiden Datentabellen werden an die Python Script node, zusammen mit einem Wert für die Laufzeit. Die Laufzeit ist tatsächlich eins Hauptparameter eingestellt werden und ist die Zeit, die erforderlich ist, um die verschiedenen Modelle zu testen sowie als Optimierung einiger Hyperparameter. Standardmäßig legt Markus die Laufzeit auf 30 Sekunden können jedoch längere Zeitintervalle gewählt werden.

Innerhalb des Python Script-Knotens wird ein Python-Skript ausgeführt, das die AutoML durchführt. Schließlich wird das beste Performance-Modell als generisches H2O-Modell sowie in MOJO-Format. Mit dem H2O MOJO Reader-Knoten ist das ausgebildete Modell im MOJO-Format lesen. Dies ist dann mit dem H2O MOJO Predictor-Knoten verbunden, um die ausgebildete Modell zu neuen Daten und Vorhersage der Zielwerte.

Der Hauptzweck dieses Forums ist es, zu erklären, wie die Statistiken und Modell Merkmale, die in der begleitenden .xlsx-Datei gespeichert sind, können verwendet werden, um die Modell. Roughly, Markus teilt dieses Verfahren in zwei Teilen: 1) Prüfmodell Grafik und 2) Untersuchung der .xlsx Datei.

Option 1: Inspect Model Graphics

In der Modellqualitätsklassifizierung - Graphics metanode zeigt Markus, wie man mehrere Grafiken, um die Leistung von binären Klassifikationsmodellen zu überprüfen. Alle die Grundstücke werden als .png Dateien zum Workflow-Unterordner-Modell/gültig mit dem

Image Writer (Port) Node. Die erläuterten Metriken sind ROC Curve, TOP Decile Lift, Kolmogorov-Smirnov Goodness-of-Fit Test, und ein Grundstück der normierten Gini Koeffizient, um den besten Abschaltpunkt zu bestimmen.

Option 2: Untersuchung der begleitenden Excel-Datei

In der Tat gibt das Studium der Grafik eine erste Idee, wie das Modell funktioniert, aber für tiefere Einblicke Markus empfiehlt weitere Untersuchung der [.xlsx](#) Datei. Diese Datei besteht aus mehreren Blättern, die verschiedene modell- und schulbezogene Informationen enthalten, einschließlich:

- **Das Leaderboard aus dem Modellset läuft.** Dies gibt Ihnen eine Idee, welche Typen der Modelle berücksichtigt und wie die Werte des AUC verteilt werden zwischen verschiedenen Modelltypen.
- **Die Modellübersicht.** Die Modellübersicht gibt Ihnen die verwendeten Parameter, die ist hilfreich, wenn Sie Ihr Modell weiter tweaken möchten. Darüber hinaus das gesamte Modell Druck wird gespeichert [.txt](#) Datei im gleichen Workflow-Unterordner und enthält weitere Informationen aller Parameter.
- **Die variable Bedeutungsliste.** Diese Liste sammelt die Bedeutung jeder Variablen und sollte sorgfältig untersucht werden. Wenn eine Variable alle Bedeutung erfasst, die Sie könnte ein Leck haben. Es gibt auch an, welche Variablen abgeschnitten werden können.
- **Die Modellübersicht in Bins und Zahlen.** Dieser Tisch gibt uns eine Idee, was ein Cutoff bei einer bestimmten Partitur (d.h. "Submission") würde bedeuten. Eine Abkürzung finden ist sehr abhängig von Ihrem Geschäftsfall. Zum Beispiel wählen Sie einen Cutoff bei 0,8 die Ihnen 92% Präzision und 43% aller gewünschten Ziele geben würde. In Marketing, das wäre ein ausgezeichnetes Ergebnis, aber in der Kredit-Scoring Sie vielleicht nicht möchte mit 8% der Menschen leben, die ihre Kredite nicht zurückzahlen.
- **Blick auf die Kreuzvalidierung.** Im Allgemeinen macht H2O eine Menge Cross-validation durch standardmäßig, um zu vermeiden, überrüsten. Allerdings möchten Sie vielleicht einige Kontrollen machen Sie selbst. Die Grundidee ist, wenn Ihr Modell wirklich einen allgemeinen Trend und hat gute Regeln, die sie an allen (random) Subpopulationen arbeiten sollten und Sie würden erwarten, dass das Modell ziemlich stabil ist. Daher schauen wir uns die Kombination an Standardabweichung. Ein Wert von 0 würde eine perfekte Übereinstimmung zwischen allen darstellen Untergruppen. Also, wenn Sie zwischen mehreren guten Modellen wählen müssen, könnten Sie das Modell mit der geringsten Abweichung betrachten wollen.

Ein letztes Wort für alle Python-Befürworter da draußen: das volle Python-Skript ist verfügbar in die Unterordner von die Arbeitsablauf unter Script „[_automl_h2o_Klassifikation_python.ipynb](#)“.

[Lesen Sie den ganzen Faden H2O.ai AutoML in KNIME für Klassifizierungsprobleme](#) „auf dem KNIME Forum and download the workflow of the same name [vom KNIME Community Hub](#)“

Lernen vom Besten auf dem KNIME Forum

In diesem Artikel haben wir drei der besten Forumsfäden von Markus zusammengefasst. Diese sind unter den Beiträgen, die am liebsten, am meisten angesehen, oder am hilfreichsten sind. Diese Geschichte über die Best-of-Markus Threads ist noch ein weiteres Beispiel dafür, wie große Community-Unterstützung kann sein. Dank Markus' breiter (KNIME-) Weisheit und seiner Aufregung der Lehre Dinge, die er auf jeden Fall aus einem oder anderen Community-Mitglied geholfen hat. Das ist die eine so vielfältige Gemeinschaft. Ich bin sicher, dass Sie ein oder zwei gelernt haben, sogar aus diesem Artikel oder könnte sogar eine neue Perspektive zu bestimmten Themen gewonnen haben - unabhängig davon, ob Sie ein neuer oder erfahrener KNIME-Benutzer sind. Was auch immer deine Motivation ist, zum KNIME Forum beizutragen, wir sind in jedem Fall glücklich über Ihr Engagement.



Weitere
Informationen

wurde nominiert Beitrag des Monats

für Mai 2022. Er wurde für seine unermüdliche Tätigkeit auf die [KNIME Forum](#) und seine unzähligen Beiträge zu den [KNIME Gemeinschaft Hub](#) in seinem öffentlichen Raum. Er ist ein Profi Mitglied des KNIME Forums und der KNIME Community Hub. Er ist bekannt für seine wertvollen Antworten auf die Forum, sowie die vielen wertvollen Workflows und Komponenten, die er in der KNIME Community beitrug.

Bruno greift immer andere KNIMERS mit höchster Höflichkeit und Klarheit ein.

Ab diesem Schreiben hat er 2.2k Posts erstellt, 3.5k Favoriten erhalten und bereitgestellt fast 250 Lösungen.

Bruno ist ein zweisprachiger agiler Teammanager mit mehr als 20 Jahren

Erfahrung in Anwendungsbereich Entwicklung. Er

verfügt über umfangreiche Erfahrung und Wissen über

alle Phasen des Softwareentwicklungszyklus,

Datenbank und Softwarearchitekturdesign. Er ist

gut bei schnellen Mustern und Vorsehen

Probleme und verhindert, dass sie vor

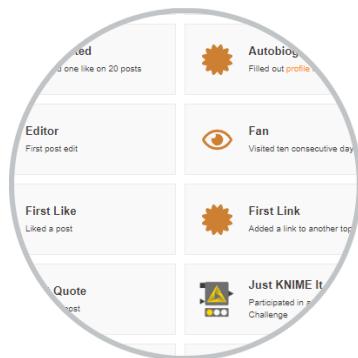
andere erkennen, dass es ein Problem gibt. Bruno ist derzeit

Director Data Ops bei Triton Digital.

Besuchen Sie Bruno's [Raum auf dem KNIME Hub](#) oder [Profil](#)

in die KNIME Forum (Hub/Forum) Griff:

bruno29a)



Unterstützung der Gemeinschaft 24/7 wie ein Champ

Bruno Ngs bestes KNIME Forum Antworten

Autor: Elisabeth Richter

Bruno Ng ist in der Tat einer unserer größten KNIME Forum-Beiträge. Er trat dem Forum bei im November 2020 und ab 24.08.2022 zählt er 535 Tage besucht, ca. 4100 Themen betrachtet, und 258 Lösungen gegeben. Das macht ihn zu einem der besten KNIME Community-Mitglieder, die die meisten Lösungen für Fragen und Probleme bieten. Stimmt. zum Motto „Hilfe mir helfen“, Bruno ist immer sehr eager, anderen KNIME-Nutzern zu helfen auf dem Forum. Im Folgenden möchten wir Brunos fünf gefielen Antworten vorstellen, die vielleicht auch bei Ihren Problemen helfen.



bruno29a

Bruno29a

Regular

Joined Nov 18, '20 Last Post 5 hours Seen 5 hours Views 2258 Trust Level regular

Rundungsnummern

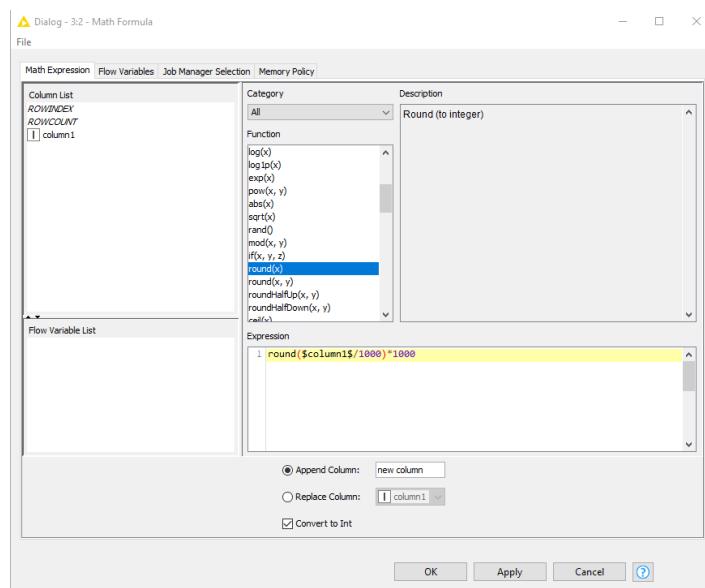
Rundzahlen ist eine Aufgabe, die sicher auf die alltäglichen Dinge für Menschen in alle Arten von Berufen und es ist definitiv nicht Hexerei. Wie man Zahlen in die zehn, hundert oder tausend wurden wahrscheinlich den meisten von uns in den frühen Stadien gelehrt von unserer Ausbildung. Aber wissen Sie, wie man zum Beispiel zu den nächsten tausend mit der KNIME Analytics Platform? Ein guter Ausgangspunkt ist der Math Formel Knoten, Was passiert danach? Die Standard-Rund(x)-Funktion reicht dabei nicht aus, da sie Runden nur Dezimals zu Ganzzahlen. Lassen Sie uns einen Blick auf diese intelligente Lösung vorgeschlagen von Bruno:

Nehmen wir an, Ihr nächster Wert wird durch x definiert (so x ist entweder 1.000 oder 10.000 je nachdem, was du gemeint hast). Machen Sie einfach eine Division von x, um das Ergebnis, und multipliziert mit x. Die Formel wäre also:

`Rund(your_value/x) * x`

Zum Beispiel, wenn ich auf die nächsten 1000:

`Rund ($colum1$/1000) * 1000`



Ergebnisse:

Output data - 3:2 - Math Formula		
File	Edit	Hilite
Table "default" - Rows: 1	Spec	Columns: 2
Row ID	column1	D new column
Row0	181400	181,000

Et voilà - schnelle und einfache Lösung. Ersetzen Sie einfach x mit jedem Wert, den Sie runden möchten bis. Also, wenn Sie auf die nächsten tausend, definieren x=1000, für die nächste 100 x=100 und so weiter.

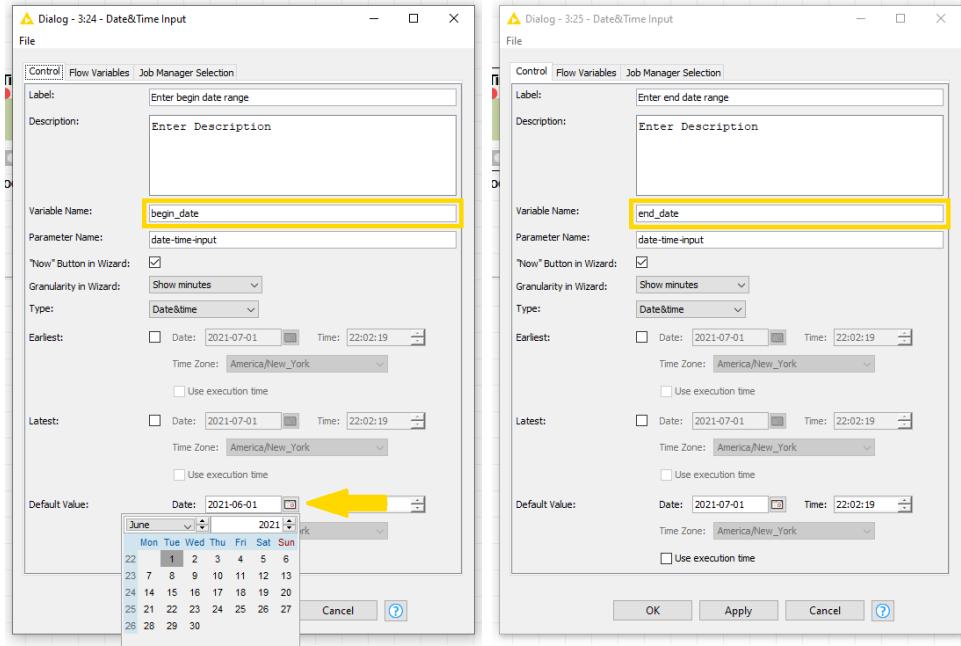
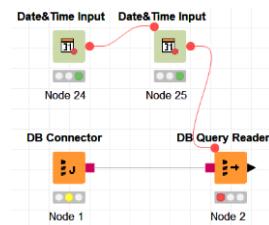
Finden Sie den Originalfaden [Wie runde ich eine ganze Zahl auf die nächsten 1000?](#) auf dem [KNIME Forum](#).

Zeitpunkte als Parameter in SQL Queries einfügen

Umgang mit SQL könnte eine Herausforderung sein, vor allem, wenn ein Newbie, die - Hand im Herzen - jeder von uns war einmal. Zum Glück, KNIME Analytics Platform bietet viele Knoten, die die Verbindung zu JDBC-konformen Datenbanken ermöglichen. Das könnte Dinge machen einfacher, aber dennoch bleibt eine oder andere Sache für einige Benutzer unklar. Auf der KNIME Forum ein Benutzer einmal gefragt, wie mehrere Datumsangaben in die DB Query haben Leselektoren. Zum Glück war Bruno zur Hand und hat diese Frage ausführlich beantwortet.

Wenn Sie wirklich wollen echte Date-Eingabe als Interaktion, dann Sie kann alles, was du ursprünglich gemacht hast, aber du kannst nur 2 Date&Time verwenden Eingangsknoten. Du musst nur sie miteinander verbinden und verschiedene Variablen verwenden Namen (siehe Bild rechts).

Verwenden Sie den richtigen Variablenamen und der Benutzer kann die Datum aus dem Pop-up:



Ebenso wie ich im vorherigen Beitrag erwähnt, können Sie die Ausgabevariablen sehen:

Flow Variable Output - 3:25 - Date&Time Input			
Flow Variables			
Index	Owner ID	Name	Value
0:3:25	s:begin_date	2021-07-01T22:02:19	
0:3:24	s:end_date	2021-06-01T22:02:19	
0	s:knime.workspace	C:\Users\elisabeth.richter\knime-workspace	

Er geht sogar über die Frage hinaus und erklärt, wie die Verwendung einer Komponente würde in dieser Situation profitieren, wie man eine schafft, und wie man sicherstellt, dass der Fluss an den dem Bauteil folgenden Knoten werden Variablen weitergegeben. Wirklich eine Modellantwort.

Finden Sie den Originalfaden [Multiple Date Variable in DB Query Reader?](#) " auf dem KNIME Forum.

Ihr JSON Kleiner Helper, Zellen zu kontaminieren

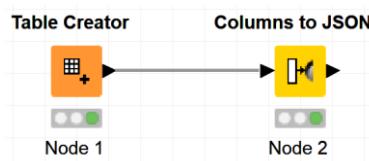
Eine weitere Frage, mit der wir uns wohl alle konfrontiert haben, ist, wie man dynamisch konkatiniert Zellenwerte zusammen aus einer Zahl n der Spalten, wobei n kann sich bei jedem Lauf ändern. Zu illustrieren das erste Ziel des Fragestellers, betrachten wir diese Tabelle. Die ersten vier Spalten, ID, Startdatum, Name und Nachname, sollen in die fünfte Spalte, Anmerkungen. Sowohl die Spaltennamen als auch die Spaltenwerte sollten berücksichtigt.

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

Hier ist, wie Bruno das Problem auf sehr elegante Weise gelöst hat, mit ein wenig Hilfe von die JSON-Knoten.

Sie können Columns zu JSON verwenden, die Ihnen etwas im gleichen Format geben wird dass Sie suchen, und es ist dynamisch, wie in Sie müssen nichts ändern wenn sich Ihre Tabelle ändert (neue Spalten, weniger Spalten, es spielt keine Rolle).

Stecken Sie es einfach an Ihre Daten:



Eingangsdaten:

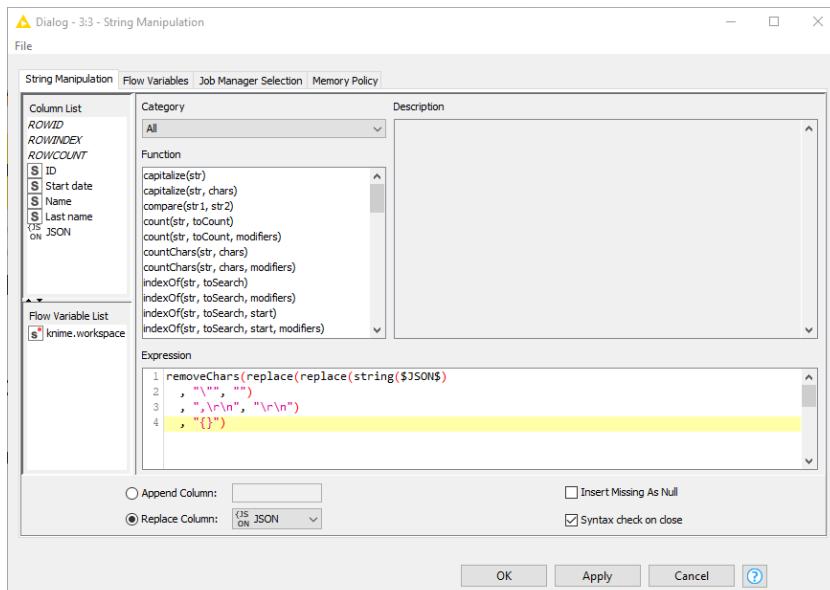
Row ID	ID	Start date	Name	Last name
Row0	111	May-01	Jane	Welch
Row1	112	May-05	Jack	Williams

Ergebnisse:

KNIME Unterstützung – Bruno Ng
Unterstützung der Gemeinschaft 24/7 wie ein Champ

Table with JSON - 3:5 - Columns to JSON (Node 2)

Sie können einige Aufräumarbeiten tun, wenn Sie wirklich wollen, das Format, das erwähnt wird. Du kann die String Manipulation verwenden:



Ergebnisse:

Appended table - 3:3 - String Manipulation

Und hier wieder ging Bruno sogar über und legte zwei weitere Spalten an, um zeigt die Flexibilität dieser Lösung. Es kann an jede Anzahl n angepasst werden Spalten.

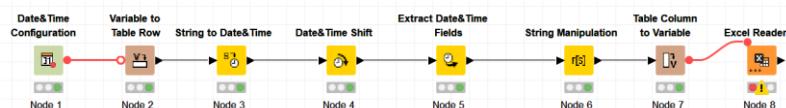
Finden Sie den Originalfaden [Konkatzellen in einen Absatz mit Spaltennamen und formatiert Text](#) auf dem KNIME Forum.

Adressierung der Daten von gestern

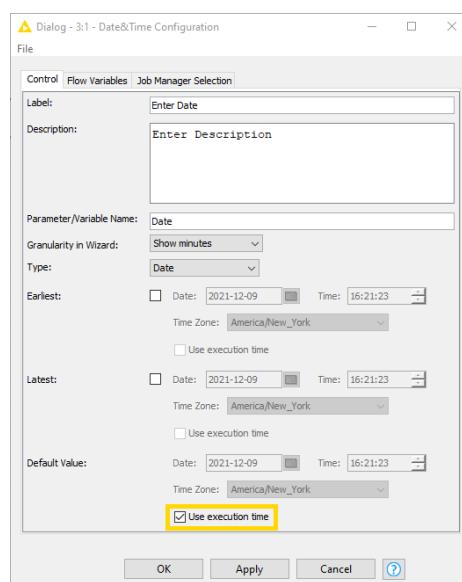
- Ist das nicht ein alter Hut? Nein. Wenigstens nicht für uns alle. Ich bin sicher. Einige von uns hatten das Problem, die Daten von gestern zu behandeln. Oder zumindest einige Daten an einem Tag geliefert, der nur ab heute extrahiert werden kann. Ein KNIME Forum Benutzer wollte wissen, wie man die gestrigen Excel-Bögen dynamisch liest, wo jeder Blatt wird als Datum bezeichnetet (z.B. „08.12.“, „09.12.“ usw.). Hier ist, wie Bruno die Lösung der Problem.

Ich habe etwas zusammengefügt, das Flexibilität bietet, das Datum zu ändern, falls Sie wollen den Workflow für ein anderes Datum ausführen, aber Standard ist das aktuelle Datum. Der Workflow funktioniert auch für jedes Datumsformat.

Mein Workflow sieht so aus:

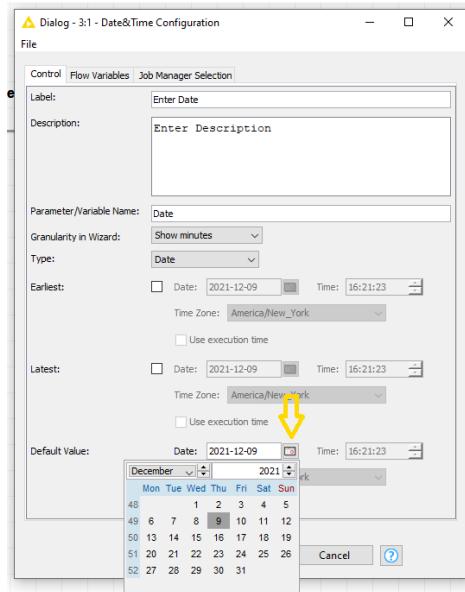


Betrachtet man den ersten Knoten, ist es konfiguriert, das aktuelle Datum zu verwenden:



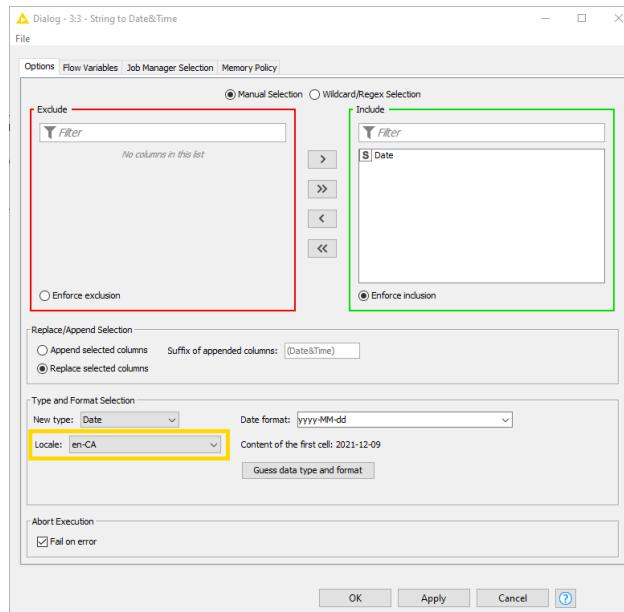
KNIME Unterstützung – Bruno Ng
Unterstützung der Gemeinschaft 24/7 wie ein Champ

Aber sollten Sie es für ein bestimmtes Datum ausführen möchten, nur deaktivieren Sie diese Box, und klicken Sie auf auf dem kleinen Symbol, um einen schönen Kalender Popup zu bekommen, wo Sie wählen können, welches Datum Sie wollen es für:

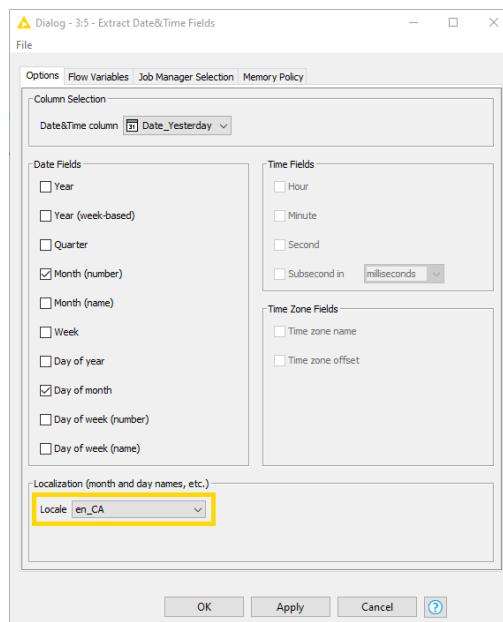


2 Knoten, die Sie basierend auf Ihrem Locale konfigurieren müssen. Die Node 3 und Node 5. Beide sollten das gleiche Lokal verwenden, das sollte Ihr Lokal sein. Ich bin auf Kanada, Also mein ist en-CA. Standardmäßig würde KNIME Ihre Locale automatisch festlegen auf das, was Ihr System/Computer eingestellt ist:

Node 3:



Node 5:

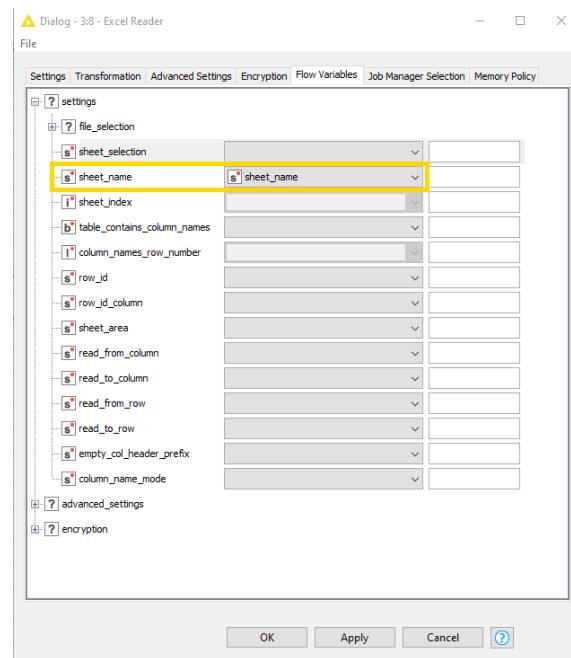


So, für das heutige Datum, wird es eine Variable namens Sheet_name mit Wert generieren „08.12“:

Flow Variables			
Index	Owner ID	Name	Value
0 3:7	\$ sheet_name	\$ sheet_name	08.12
0 3:1	\$ Date	\$ Date	2021-12-09
0	\$ knime.workspace	\$ knime.workspace	C:\Users\elisabeth.richter\knime-workspace

Ich habe es auch so konfiguriert, dass, wenn der Monat ist von 1 bis 9 (Jan bis Sept), Monat wird als 01 bis 09 formatiert, so wenn Sie den Workflow für 4. Feb ausführen, die Sheet_name wird “03.02” sein.

Der Excel Reader ist so konfiguriert, dass der Sheet_name als Sheet_name verwendet wird:



Sie müssen nur auf die Datei zeigen, und das ist es.

Angesichts dieser Antwort gibt es nichts zu sagen. Das war nicht zu hart, oder?

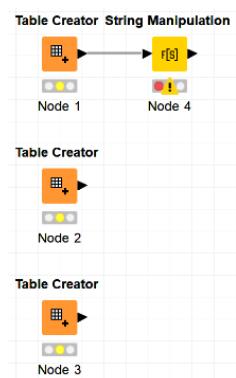
Finden Sie den Originalfaden [Wie dynamisch zu lesen Excel-Blatt](#) „auf dem KNIME Forum.

Eine Node in einer Zeit

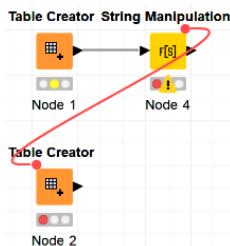
Haben wir nicht alle an der Stelle gewesen, wo wir wollten mehr Kontrolle über unseren Workflow Ausführung und gezielt die Reihenfolge, in der unsere Knoten ausgeführt werden? Nun, wenn Sie haben nicht mindestens einige KNIME-Nutzer auf dem Forum, die speziell gefragt Diese Frage. Auch hier veröffentlichte Bruno eine ausführliche Erklärung, wie man das handhabt Situation.

Grundsätzlich führt KNIME von links nach rechts nacheinander Knoten aus, was bedeutet, wenn 2 Knoten werden miteinander verbunden, es führt zuerst den Knoten von links aus und dann die rechte.

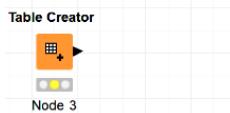
Nun gibt es Fälle, in denen Sie "kann" verbinden 2 Knoten, in dass der Ausgangsporttyp des linken Knotens nicht kompatibel ist mit der Eingangsporttyp des richtigen Knotens oder sogar Fälle, in denen Ihr linker Knoten hat keinen Ausgangsport (weil es keine Operationen, die in Bezug auf diesen Knoten erfolgen sollen, beispielsweise Schreiber (Excel Writer, CSV Writer, etc), oder E-Mail senden, usw.), oder der richtige Knoten hat keinen Eingabeport (Table Creator, B. In diesem Fall würden Sie sie über den Flow verbinden Variabler Port.



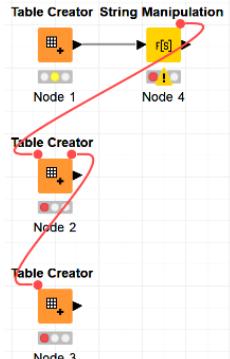
Wenn ich beispielsweise diesen Workflow habe (siehe Bild oben) und ich das Workflow, dann werden die Knoten 1, 2 und 3 alle gleichzeitig beginnen.



Wenn ich den Knoten 4 jedoch so mit dem Knoten 2 verlinke (siehe Bild links), dann wird der Knoten 2 erst nach dem Knoten ausführen 4 ist erledigt. Wenn also dieser Workflow ausgeführt wird, nur Knoten 1 und Knoten 3 wird gleichzeitig beginnen. Node 4 wird ausgeführt erst nach Beendigung des Knotens 1 und der Knoten 2 erst nach Beendigung des Knotens 4.

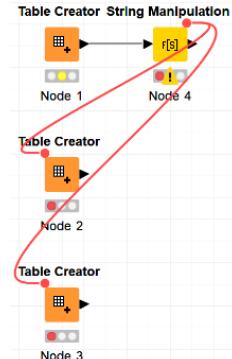


Sie können verschiedene Variationen haben, zum Beispiel:



In der linken Abbildung wird Knoten 2 erst nach Beendigung des Knotens 4 ausführen, und Knoten 3 wird erst nach Knoten ausgeführt 2 abgeschlossen.

Eine weitere Variante ist in der Figur rechts. Hier, Knoten 2 und Knoten 3 gleichzeitig ausgeführt aber erst nach Beendigung des Knotens 4.



Und weiter:

Wenn es um Metanode geht, kann es ein bisschen knifflig sein, je nachdem, was ist Wesen getan und was in der Metanode verknüpft wird. Eine Metanode einfach „zusammenfassen“ Teil Ihres Workflows, so dass die Knoten innerhalb einer Metanode sind unabhängig. Die Metanode ist kein Objekt, so dass einige Knoten bereits starten können wenn sie "linke" Knoten sind bereit. Mit Komponenten ist es etwas anders. Eine Komponente wird erst beginnen, wenn alle Eingangssports bereit sind.

Dies ist wirklich eine wertvolle Antwort, vor allem für KNIME Anfänger, die noch nicht so sind alle Funktionalitäten der KNIME Analytics Platform kennen.

[Das Original finden Forum Thread](#) “ [Stellen Sie sicher, dass ein Knoten erst beginnt, wenn ein anderer erledigt ist auf dem KNIME Forum.](#)

Hilfe vom Besten auf dem KNIME Forum suchen

In diesem Artikel haben wir Brunos herausragendes Engagement auf dem KNIME Forum hervorgehoben und wollte ihn für seine vielen technischen Antworten bestätigen. Er hat sicher half einige neuen und erfahrenen KNIME-Nutzern mit seinen ausführlichen Antworten. Fast jeder Zeit Bruno schlägt eine Lösung vor, er bietet nicht nur viele Screenshots damit, sondern auch stellt einen Workflow fest, um sicherzustellen, dass der Fragesteller wirklich geholfen wird. So bleibt er treu zu seinem Profilslogan „Hilfe mich dir bei, indem du so viele Informationen wie möglich gibst. Je genauer Sie in Ihren Informationen sind, desto genauer wird die Lösung sein.“.



[Brian Bats](#) wurde nominiert Beitrag des Monat für Oktober 2021. Er wurde für seine Tätigkeit ausgezeichnet auf dem KNIME Forum und seinem [Datei öffnen oder Ordner und String Emoji Filter](#) Komponenten. Ab 28.09.2022 beide Komponenten insgesamt 1.926 Downloads. Die erste Komponente ist zum schnellen Öffnen und Überprüfen einer Datei oder Ordner in Ihr KNIME Workspace. Die zweite Komponente ist für Emojis aus E-Mails oder Tweets entfernen. Brian ist ein

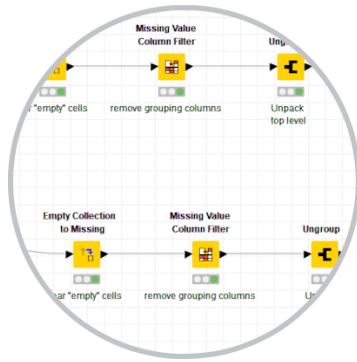
vertrauensvolle Präsenz im KNIME Forum. In diesen ersten sechs Monaten

Forum, er akkumulierte mehr Lieblingsposten als jeder im Vorjahr!

Er bekommt seine Motivation von der Zufriedenheit, mit dem konfrontiert zu werden Herausforderungen im Forum und die Unterstützung anderer mit ihren Workflows.

Brian hat mehr als 35 Jahre breite IT-Erfahrung die es ihm ermöglicht, neue Technologien abzuholen schnell. Er spielt gerne mit Daten und liebt es, Daten in Informationen transformieren. Er ist derzeit ein Daten und Integration Architekt im Walt Disney Unternehmen.

Brians besuchen „[Raum auf dem KNIME Hub](#)“ oder „[Profil im KNIME Forum](#)“ ([Hub/Forum Griff:](#) takbb)



Von anpassbaren XMLs bis zu flexiblen Datum &Time Handhabung

Denken außerhalb der Box mit Brian Bates' Best-of-Forum Gewinde

Autor: Elisabeth Richter

Seit Brian im März 2021 zum KNIME Forum beigetreten ist, sind viele KNIME-Gemeinschaftsmitglieder haben bereits von seinem Engagement profitiert. Ab 17.08.2022 zählt Brian 500 Tage besichtigt, etwa 2800 Themen betrachtet, und 106 gegebene Lösungen. Das macht ihn zu einem sehr aktiven Mitglied im KNIME Forum. Seine Kommentare über verschiedene ETL-bezogene Themen, einschließlich Zugriff auf Daten aus Excel-Dateien, Verbindung zu SQL-Datenbanken, oder Umgang mit verschiedenen Datentypen wie Strings oder Datums- und Zeitdatentypen. Er ist auch viele Feedback und Ideen zur Verbesserung der KNIME Analytics Platform durch Vorschlag von Feature-Anfragen.

Neugierig, mehr über das unermüdliche Engagement dieser KNIME-Unterstützung zu erfahren Champion? Hier ist eine Sammlung der drei Best-Of-Brian Forum Threads, wo er hat Sein Wissen und sein Know-how in den Dienst anderer stellen!



takbb
Brian Bates
📍 London, UK

Joined Mar 5, '21 Last Post 1 day Seen 17 hours Views 2637 Trust Level member

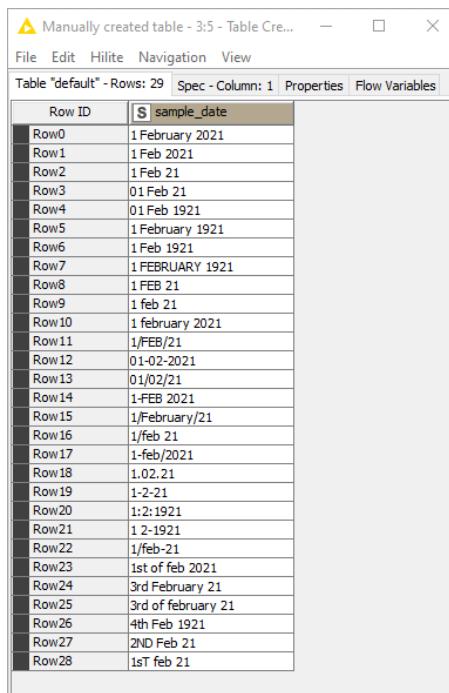
Flexibles Datum Format Handling

Dies ist ein sehr hilfreiches KNIME Forum Thread, in dem Brian ein Thema behandelt, das kann schnell frustrierend: Umgang mit kundenspezifischen Datums- und Zeitformaten. Wenn Sie lesen Datums- und Zeitdaten in der KNIME Analytics Platform werden in der Regel zuerst als String gelesen. Die jeweilige Spalte(n) kann dann in die dedizierten Datums- und Zeitdaten von KNIME umgewandelt werden Typ mit String to Date&Time Knoten. Idealerweise möchten wir einen Datensatz mit konsistente Datums- und Zeitzeichenfolgen, die einfach in Datums- und Zeitformat umgewandelt werden können. Im realen Leben ist dies jedoch oft nicht der Fall, da reale Lebensdaten messy sein können, insbesondere beim Mischen von Daten aus verschiedenen Datenquellen, in denen unterschiedliche Datums- und Zeitformate verwendet werden.

Brian bemerkte, dass es ein paar Fragen zum KNIME Forum über die Umwandlung gab Termine. Manchmal die Formatmaske im Konfigurationsdialog des ~~Datum und Uhrzeit~~ node hat eine harte Zeit, das richtige Datum und Zeitformat zu erkennen, insbesondere wenn Das ist sehr unvereinbar. Es trat ihm auf, dass es schön wäre, eine breite zu handhaben Vielfalt von Datumsformaten mit einer einzigen Maske, oder einige generische Knoten und/oder ein winziges Stück generischen Code.

Ein Problem, das er bemerkte, ist, dass, wenn der Monat ein Name ist, die Maske ist nicht hilfreich zu erkennen inkonsistente Formate, weil es fallempfindlich ist. Zum Beispiel die Monatsmaske MMM wird "Feb", aber es wird nicht in der Lage sein, "FEB" und "feb". Ähnliche Probleme treten auf, z.B., wenn man zwei- und vierstellige Jahre (d.h. yy und yyyy) erreichen will.

Um diese Probleme zu lösen Brian erstellte einen Workflow, der in der Lage ist, eine breite verschiedene Datumsformate auf einmal, d.h. innerhalb desselben Datensatzes. Nur Anforderung: Termine müssen im Format Tag-Monat-Jahr sein. Die Eingabetabelle mit verschiedenen Zufallsdatum Formate werden im folgenden angezeigt.



The screenshot shows a KNIME Table node window titled "Manually created table - 3:5 - Table Cre...". The table has one column named "sample_date". The data consists of 29 rows, each containing a different date string. The dates are mostly variations of "1 February 2021" but include many other formats such as "1 Feb 2021", "01 Feb 21", "01/02/21", "01-02-2021", and "1/02/21".

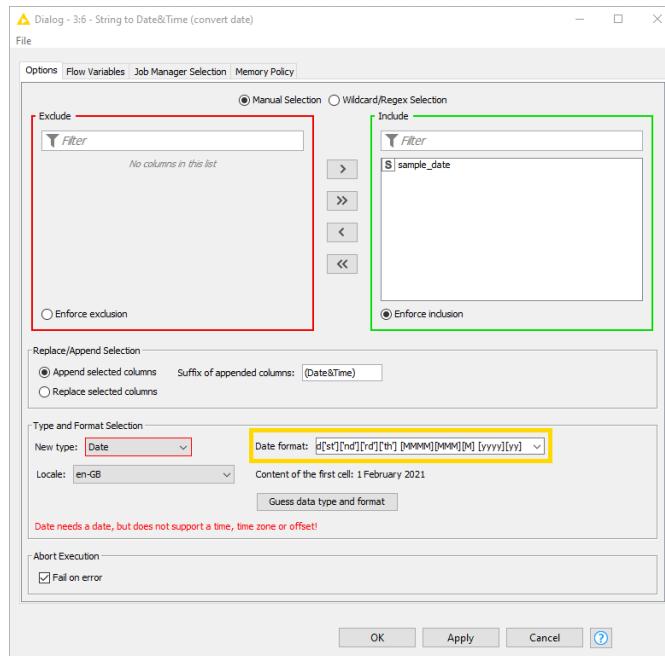
Row ID	sample_date
Row0	1 February 2021
Row1	1Feb 2021
Row2	1.Feb 21
Row3	01 Feb 21
Row4	01 Feb 1921
Row5	1 February 1921
Row6	1 Feb 1921
Row7	1 FEBRUARY 1921
Row8	1 FEB 21
Row9	1 feb 21
Row10	1 february 2021
Row11	1/FEB/21
Row12	01-02-2021
Row13	01/02/21
Row14	1-FEB 2021
Row15	1/February/21
Row16	1/fb 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

Eine Spalte mit verschiedenen Zufallsdaten, alles in der Format Tag-Monat-Jahr.

Nachdem er die Spalte mit zufälligen Daten gelesen hatte, benutzte er den Knoten String Manipulation die Datumsspalte vorbereiten. Dies bedeutet z.B. das Entfernen der Punktum, oder Kapitalisierung des Datums (d.h. „1. Februar 2021“ → „1. Februar 2021“). Dann benutzte er die String to Date&Time Knoten mit einer flexiblen Datumsformatmaske zur Umwandlung der gesamten Spalte in das native Date&Time Format yyyy-MM-dd von KNIME. Siehe Konfiguration Fenster unten unter „Typ und Formularauswahl“ für das flexible Datumsformat.

KNIME Unterstützung – Brian Bates

Von anpassbaren XMLs bis hin zur flexiblen Daten- und Zeitverarbeitung



Das Konfigurationsfenster des String to Date&Time-Knotens mit einem flexible Datum Formatmaske. Mit diesen Einstellungen kann man konvertieren verschiedene zufällige Daten in das native Date&Time-Format von KNIME, gegeben die Strings folgen dem Format Tag-Monat-Jahr.

Die Datumsformatmaske kann manuell nach Ihren Anforderungen angepasst werden. Die Workflow bereitgestellt von Brian enthält ein zweites Beispiel, das verschiedene zufällig Datumsstrings, die im Format Monat-Tag-Jahr sind.

Er könnte seinen Workflow eines Tages in eine nützliche Komponente umwandeln, also...

Lesen Sie den ganzen Faden [Schreiben eines flexiblen Date Format Handling Workflow](#) auf dem KNIME Forum und den dazugehörigen Workflow herunterladen [KNIME Workflow Mit flexibler DATE Formatmaske](#)

XML generieren

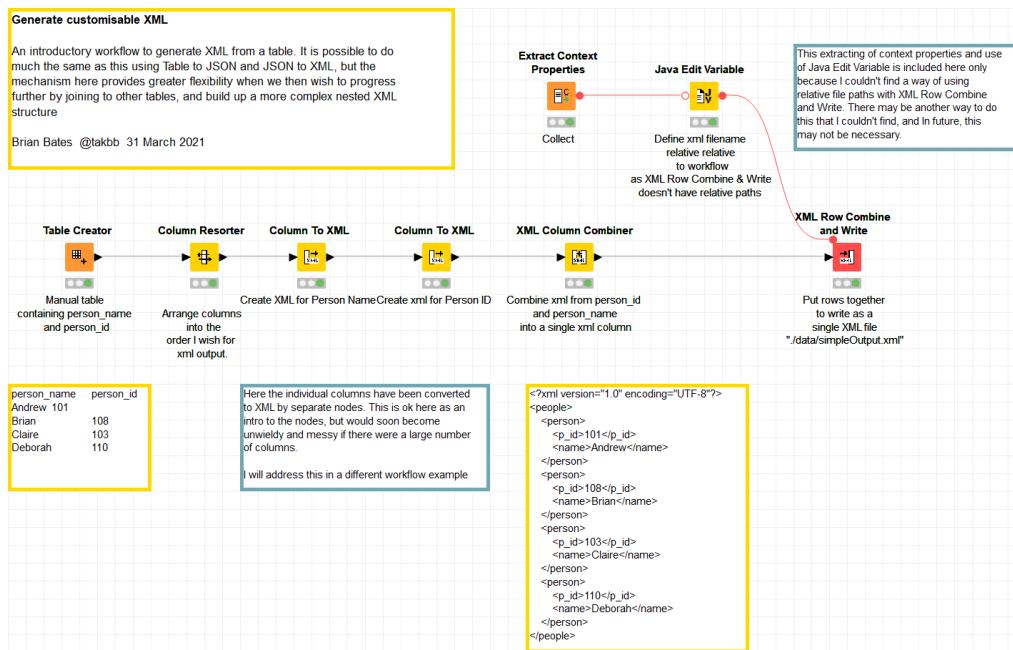
Umgang mit XML (= eXtensible Markup Language) ist etwas, das jeder, der Arbeiten mit Daten werden wahrscheinlich früher oder später behandelt werden müssen. In KNIME Analytics Platform XML-dedicated nodes existieren, die das XML-Handling erleichtern. Dieses Forum Thread adressiert insbesondere eine Aufgabe: [Erstellung XML aus einem oder mehreren Datentabellen](#). Wenn Sie eine XML-Datei in KNIME erstellen möchten, ist eine einfache und schnelle Möglichkeit, die Tabelle JSON Knoten gefolgt von einem JSON zu XML Knoten.

Aber für Brian war dies nicht genug. Er wollte darüber hinausgehen und suchte für eine Möglichkeit, XML nicht nur aus einfachen Tabellen, sondern auch aus komplexeren zu erstellen,

d.h. geschachtelte Struktur. In diesem Forum-Thread dokumentierte Brian seinen Ansatz durch Brechen in drei Stufen. Er baute zunächst einen Workflow, der das Basisgehäuse repliziert, d.h. einen Tabelle wird eine XML-Datei. Er erstellte dann Komponenten, um den Workflow mehr zu machen generisch und flexibel. Schließlich baute er einen dritten Workflow, um zu zeigen, wie man anpassbare, geschachtelten XML-Dateien.

Der erste Workflow: Einfache benutzerdefinierte XML Erzeugung aus Tabelle

Der für diesen Zweck geschaffene erste Workflow ist in der folgenden Abbildung dargestellt. Das Der Workflow deckt den Grundfall ab, eine XML-Datei aus einer Tabelle zu generieren. Im Allgemeinen, die Funktionalität des Workflows ähnelt der Verbindung sequentiell der Tabelle zu JSON und JSON zu XML Knoten. Aber laut Brian, dieses Verfahren eine größere Flexibilität bietet, wenn wir den Anwendungsfall erweitern möchten, z.B. indem wir andere Tabellen und Bau einer komplexeren geschachtelten XML-Struktur.



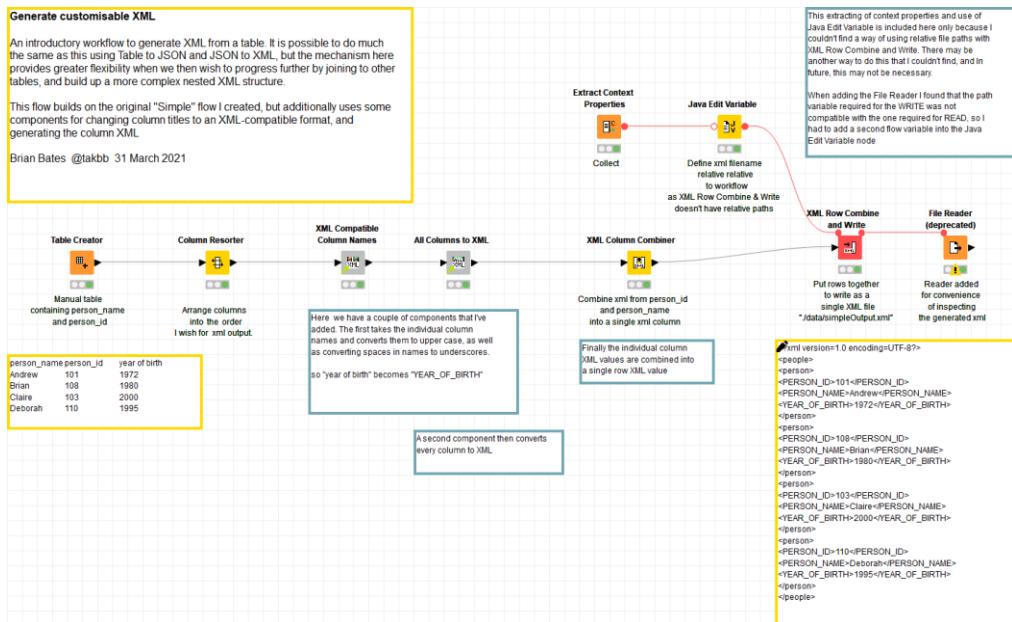
Dieser Workflow repliziert den Fall, dass eine Tabelle eine XML-Datei wird, ohne KNIME JSON-Knoten zu verwenden.

Wie im Screenshot des Workflows gezeigt, wird jede Spalte in der Tabelle in XML separat mit dem Column zu XML-Knoten wiederholt. Obwohl dies gut funktioniert für ein Grundbeispiel in einem komplexeren Anwendungsfall mit einer Vielzahl von Spalten dies würde bald unhandlich und chaotisch werden. Um sich für diesen Fall anzupassen, hat Brian eine zweiter, flexiblerer Arbeitsablauf.

Workflow 2: Einfache anpassbare XML-Generation aus Tabelle mit Komponenten

Um sich für diese Umstände anzupassen, beinhaltet Brians Lösung die Zugabe einer Schleife zu der Workflow, der alle Spalten iterativ in XML umwandelt. Darüber hinaus, Es werden standardisierte Spaltennamen benötigt, um die Schleife nicht zu brechen. Daher, a

eine zweite Schleife hinzugefügt, die alle Spalten umbenannt, bevor sie in die Spalte eingegeben werden zu XML-Knoten. Schließlich, um den Workflow ordentlicher zu machen, werden die Schleifen in einem gekapselt wiederverwendbare Komponente.



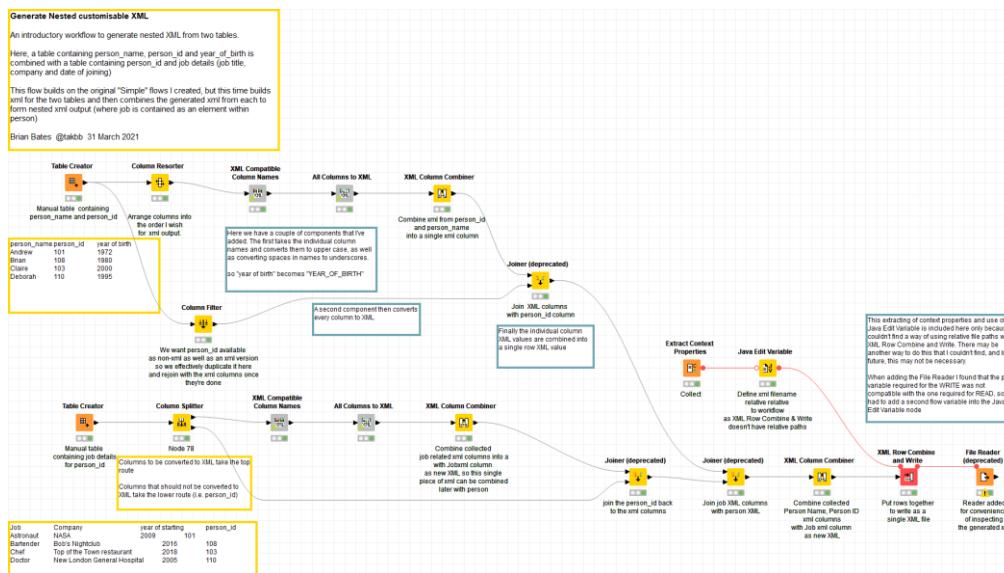
Dieser Workflow ist eine erweiterte Version des ersten Workflows, die Eingabetabellen mit einer größeren Anzahl von Spalten.

Die [XML Kompatible Säulennamen](#) für die Normung der Spaltennamen, indem sie in Großbuchstaben umwandeln und Whitespaces durch unterstrichen. Die [Alle Spalten auf XML](#) für das Durchschleifen verantwortlich alle Spalten in der zur Verfügung gestellten Datentabelle und Konvertierung in XML unter Verwendung der Spalte zu XML-Knoten.

Das letzte Hindernis war, einen Weg zu finden, mit einem komplizierteren Tischstruktur, d.h. geschachtelte Struktur. Daher erstellte Brian einen dritten Workflow.

Workflow 3: Eingebettet Customisable XML Erzeugung aus Tabelle

In diesem dritten Workflow zeigte Brian, wie man eine weitere Datentabelle zum Schluss hinzufügt XML-Datei über den Join-Betrieb. Dies ermöglicht die Erstellung einer XML-Datei mit einem Nest Struktur.



Dieser Workflow ist eine weitere Erweiterung der beiden Workflows oben und ermöglicht die Generierung von XML mit verschachtelte Informationen.

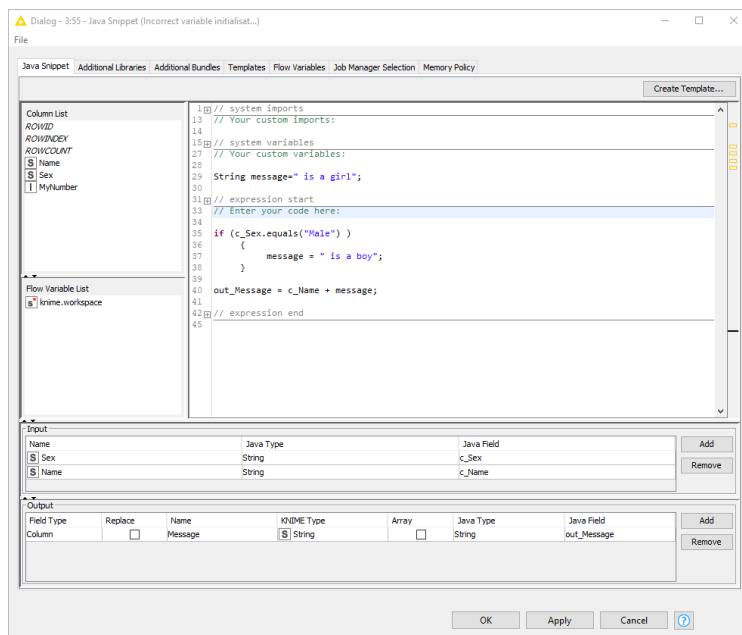
Lesen Sie den ganzen Faden [Einige Workflows, mit denen ich für die Erzeugung von XML gespielt habe](#) auf dem KNIME Forum und Download der drei Workflows des KNIME Community Hub:

- [Einfache benutzerdefinierte XML Erzeugung aus Tabelle](#)
- [Einfach anpassbar XML Erzeugung aus Tabelle mit Komponenten](#)
- [Eingebettet benutzerdefinierte XML Erzeugung aus Tabelle](#)

Benutzerdefinierte globale Variablen in Java Snippet Nodes

In diesem Forum Thread, Brian adressierte ein interessantes Thema zum Java Snippet Knoten. Obwohl dies nicht als Überraschung für alle kommen könnte, stolperte er auf einige interessantes Verhalten des Java Snippet-Knotens und wollte mit jedem teilen, so andere machen nicht den gleichen Fehler. Beim Spielen mit einem Java-Code bemerkte er die in dem Abschnitt mit „Systemvariablen“ bezeichneten Variablen in der Konfiguration Dialog des Java Snippet-Knotens (siehe Abbildung unten) werden nur während der Instantiation des Objekts aus dem Stück Java gebaut. Sie sind im Grunde benutzerdefinierte „globale“ Variablen, die bedeutet, dass sie für jede Zeile nicht wieder initialisiert werden.

KNIME Unterstützung – Brian Bates
Von anpassbaren XMLs bis hin zur flexiblen Daten- und Zeitverarbeitung



Der Konfigurationsdialog des Java Snippet Knotens. Die Variable „Nachricht“ in der Abschnitt „Systemvariablen“ dient als benutzerdefinierte „global“ Variable und ist nicht für jede Zeile neu initialisiert.

Er erklärt dies anhand des folgenden Beispiels:

Manually created table - 3:54 - Table Creator (Sampl...)						
File Edit Hilit Navigation View						
Table "default" - Rows: 5 Spec - Columns: 3 Properties Flow Variables						
Row ID	\$Name	\$Sex	I MyNumber			
Row0	Albert	Male	10			
Row1	Brenda	Female	20			
Row2	Charlie	Male	30			
Row3	Deborah	Female	40			
Row4	Edward	Male	50			

Die Eingabedatentabelle mit den Attributen „Name“, „Sex“ und „MyNumber“.

Sein Ziel war es, für jede Datenezeile die Nachricht zu drucken “[Name] ist ein Junge” wenn Sex=Male, und „[Name] ist ein Mädchen“ anders. Jedoch der Java Snippet-Knoten, der wie gezeigt konfiguriert ist oben erzeugt ein falsches Ergebnis, da es die Nachricht druckt “[Name] ist ein Junge” für jeden Zeile, unabhängig vom Wert für „Sex“.

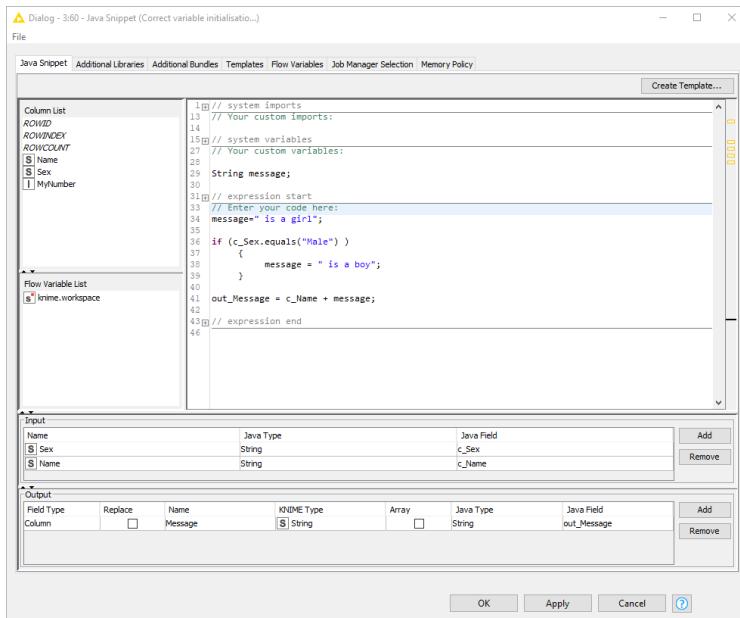
Nach einigen graben, Brian gelernt, der Grund dafür ist, dass die Saite “Message” ist definiert im Abschnitt „Systemvariablen“. Die Variablen in diesem Abschnitt sind jedoch Globale Variablen und damit werden sie für jede Zeile nicht neu initialisiert. Das heißt, die Wert für die Nachricht wird global definiert, um “ist ein Mädchen” aber sobald der Wert ändert an “ist ein Junge” in der ersten Reihe, es hält diesen Wert für die restlichen Zeilen.

KNIME Unterstützung – Brian Bates
Von anpassbaren XMLs bis hin zur flexiblen Daten- und Zeitverarbeitung

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

Das Ergebnis des in der Figur konfigurierten Java Snippet-Knotens oben. Die Spalte „Message“ fügt für jede Zeile falsch hinzu Satz “[Name] ist ein Junge.

Ändern Sie jedoch die Konfiguration des Java Snippet-Knotens und definieren Sie den Wert für „Botschaft“ im Abschnitt „Expressionsstart“ (siehe Abbildung unten) korrektes Ergebnis. Daher ist „Botschaft“ nicht mehr global definiert und daher wiederinitialisiert nach der Bearbeitung jeder Zeile.



Der Konfigurationsdialog des Java Snippet-Knotens beim korrekten initialisiert. Die Variable "Message" wird im Abschnitt "Expressionsstart" definiert und dadurch die Variable nach der Bearbeitung jeder Zeile wieder initialisiert.

Aber das ist nicht alles. Brian enthält auch Beispiele, für die eine Definition von global Variablen können z.B. eine laufende Summe berechnen, eine Lag-Spalte erstellen, Hinzufügen einer progressiven Namensverkettung, zur Sequenzerzeugung oder für Gegengenerator.

[Lesen Sie den ganzen Faden “ Java Snippets haben lange Erinnerungen!... ” und finden Sie die zugehörigen Workflow auf dem KNIME Forum.](#)

Inspiration aus dem Besten auf dem KNIME Forum

Nachdem wir das Ende dieses Artikels erreicht haben, können Sie feststellen, dass Sie etwas gelernt haben. oder zwei. Oder Sie haben vielleicht sogar eine ganz neue Perspektive zu einigen Themen gewonnen. Du hast noch nie darüber nachgedacht. Was auch immer der Fall sein mag, dank Brians breiter Wissen und sein unermüdliches Engagement im KNIME Forum hat er definitiv geholfen neue und erfahrene KNIME-Nutzer inspiriert.

Dieser Artikel zeigt, wie große Community-Unterstützung sein kann - und wie vielseitig er kann sein. Der Grund, im KNIME Forum aktiv zu sein, kann von unterschiedlicher Motivation sein. Sei es, Hilfe mit einer bestimmten Fehlermeldung suchen, Fehler melden oder neue Funktionen vorschlagen, unsere Entwickler, oder einfach Ihre Gedanken und Wissen mit der Gemeinschaft zu teilen. Was auch immer Ihre Motivation ist, wir sind in jedem Fall glücklich über Ihre Beiträge.

Node & Topic Index

A

ADME-Prädiktion	133
PDF herunterladen Dokumente	88
Apache Spark	5.
API	23, 47, 68, 153
Automatisierte Feature Encoding	27
Automatisiertes Feature Engineering	27
Automatisierte Feature Generation ..	30
Automatisierung	27, 174
AutoML	185

B.

Banken	5.
BERT	ANHANG
Biomarker	ANHANG
BIRT-Erweiterung	35, 62
Business Automation	62, 174
Business Intelligence	12

C

Zellverflüssigung.....	193
Zellteilung	88.
Churn-Prädiktion	125
Klassifikationsproblem	185
Cloud-Verbindung	153
Kolumn-Antragsteller	35
Kolumnen an JSON	193
Gemeinschaftsunterstützung	146, 180, 190, 202
Konstante Wertekolumn	35
Inhalt Marketing	129
Datum und Uhrzeitbereich	88

D

Datenanalyse	12
Datenextraktion	153
Data Science Education	95
Daten zum Bericht	35.

Datenwerkzeuge	12
Datenbankverbindung	164
Datum und Uhrzeitkonfiguration	195
Datum und Uhrzeitbehandlung	202
DB Querleser	19
DB Reader	ANHANG
DB Tabellenauswahl	ANHANG
Deep Learning	ANHANG
Krankheitsprädiktion.....	111
Drogenentdeckung	ANHANG
Drogenrückgewinnung	ANHANG
Dungeons & Dragons	62
Duplikate Handhabung.....	180
Duplicate Row Filter	180
Dynamischer Datenzugriff	195

E

Bildung	95, 102
Excel Reader	195
Excel zu KNIME	117.
Auszug Datum und Uhrzeit	195

F

Facebook Group	146
Feature Einbettung	27
Funktion Kodierung	27
Feature Engineering	27
Feature Generation	30
Dateilesler	35.
Finanzen	5, 68
Finanzanalytik	
Strömungsvariablen	193, 195, 198
Fraud Detection	ANHANG

G

Gene Ontologie	ANHANG
Gephi	47.
Globale Variablen	207

- Google Analytics 153
- Google Analytics API 153
- Google Analytics Verbindung 153
- Google Analytics Abfrage 153
- Google Authentication (API Key) 153
- GruppeBy..... 35, 180

H

- H2O.ai 185

I

- Image Mining ANHANG

J

- Java Snippet 75, 207
- Teilnehmer 35, 117
- JSON 193

K

- Schlüsselwort Forschung ANHANG
- KNIME Zertifizierungsprogramm 102
- KNIME-Funktionen 5.
- KNIME Hub ANHANG
- KNIME WebPortal 5, 23

L

- Sprachmodelle 126
- Lebenswissenschaften..... ANHANG
- Lebenswissenschaften 106, 133
- Lineare Regression 183

M

- Machine Learning 5, 18, 23, 68, 75, 111, 124, 185
- Marketing Analytics ANHANG
- Mathematische Formel ANHANG
- Microsoft Access Connector 164
- Fehlender Wert..... 35
- Modellleistung 68
- Modellvorschrift 183

- N**
- Netzwerkanalyse..... 47.

P

- Parametereingabe 19
- Parameteroptimierung 75
- Parameteroptimierung Loop End..... 75
- Parameteroptimierung Loop Start..... 75
- E-Mail-Adresse PDF Dokumente 88
- Prädiert Fußballpass 75
- Beschaffung ANHANG
- Produktqualität 174
- Python 5, 18, 68, 111

R

- R 111
- R Scripting Extension 182
- R Statistik Integration 47
- R Ansicht (Tabelle) 182
- Regex Extractor 88
- REST..... 5.
- REST API 23.
- Risikoanalyse 5.
- Rundungsnummern ANHANG
- Row Filter 35.
- Regelmotor (Diktiorär)..... 35

S

- Suchmaschinenoptimierung 130
- Sentiment Analysis ANHANG
- Folgenabnahme 198
- Solubilitätsherausforderung 133
- Sortierer 75
- SQL Query 19
- Statistik 23.
- String Input 35.
- String Manipulation 35, 193
- String to Date&Time 195, 202
- String to Date/Time (legacy) .. 35
- String an URI 88.

T

- Tischler 35, 88

Lehren	95, 102	vgl	
Tika Parser	88.	.	
Zeit zum Streichen (Recht).....	35	Geige Plot	182
TomTom API	ANHANG	VLOOKUP	117.
Thema Modellierung	128		
Twitter API	47, 68		
		W	
		Web Analytics	153
U			
UFO Sichtungen	ANHANG	X	
Ungruppe	88.	XML Generation	204

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