



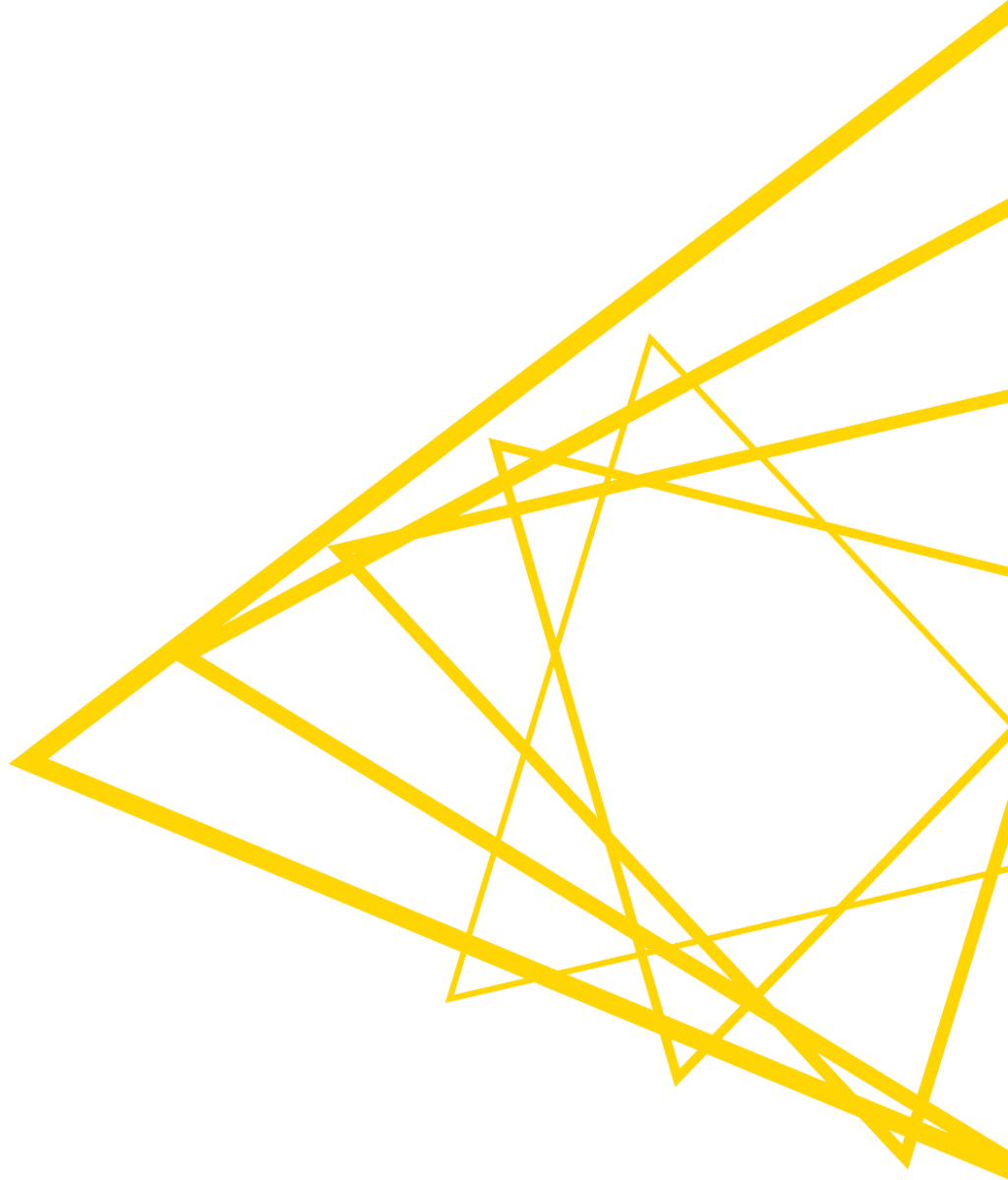
Practicing Data Science

A Collection of Case Studies

Rosaria.Silipo@knime.com



Strata London , May 2 2019

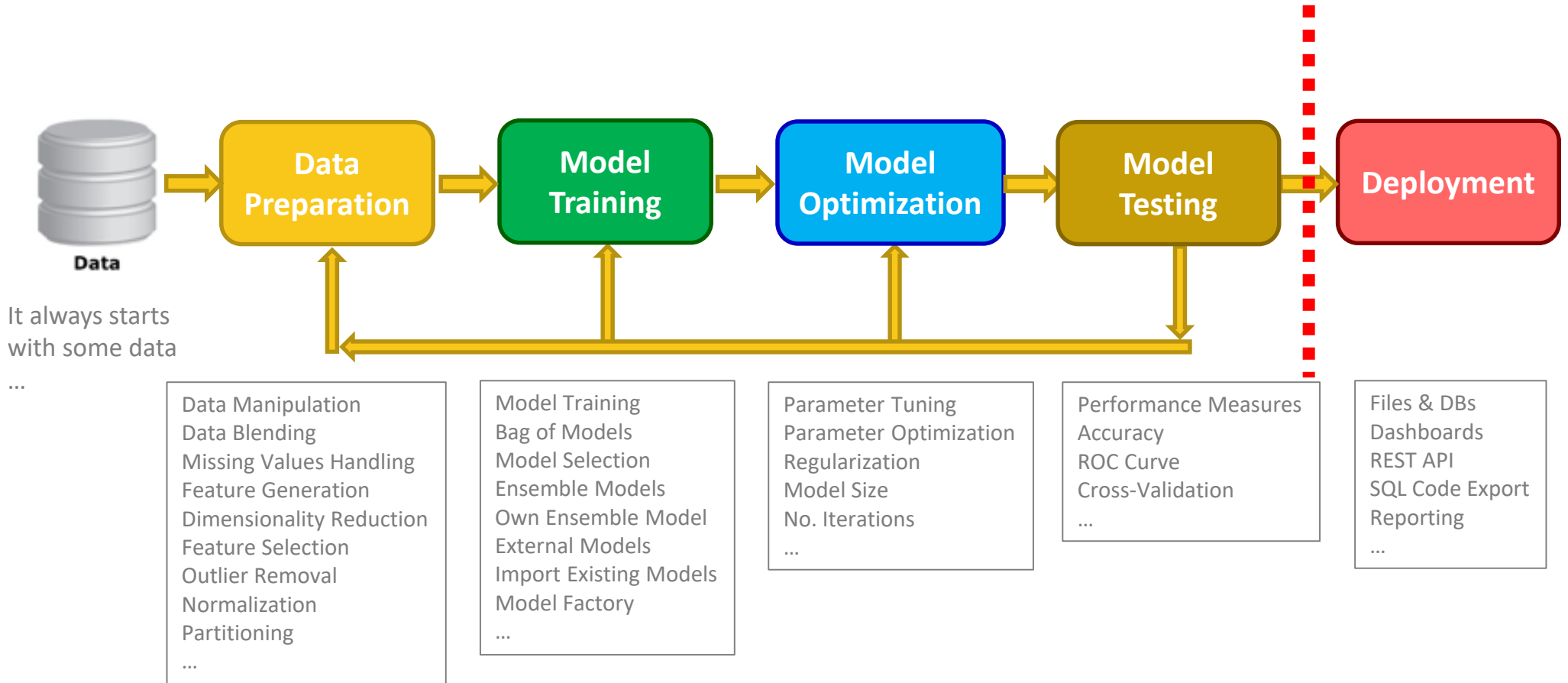


A few Words about me



- I am Rosaria Silipo
 - Principal Data Scientist at KNIME
 - At least 20 years analyzing data
-
- Generally interesting projects become Case Studies
 - 22 case studies collected in a book
 - Almost 23

A Classic Data Science Project



Customer Intelligence: Churn Prediction

Churn Prediction: The Problem



CRM System
Data about your customer

- Demographics
- Behavior
- Revenues

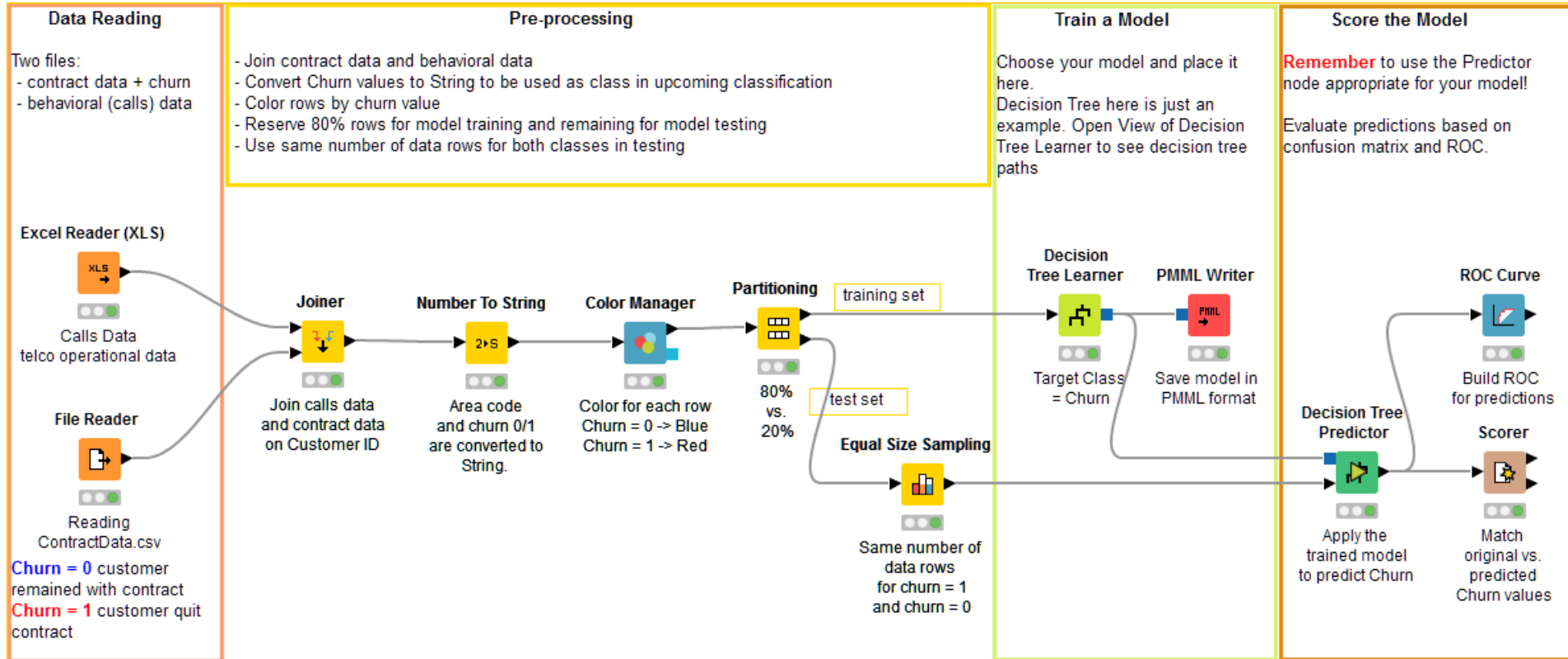


Model



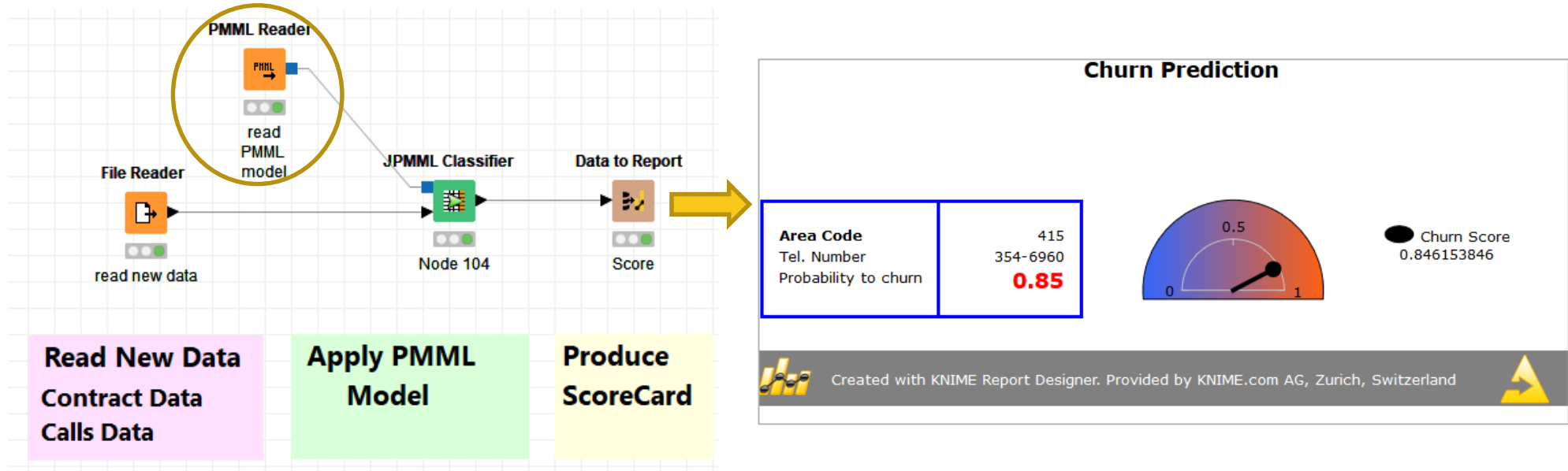
- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- Customer Segmentation
- ...

Churn Prediction: The Training Workflow



Churn Prediction: The Deployment Workflow

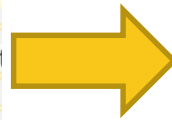
YouTube: "Building a basic Model for Churn Prediction with KNIME" <https://www.youtube.com/watch?v=RHSO10q7e2Y>



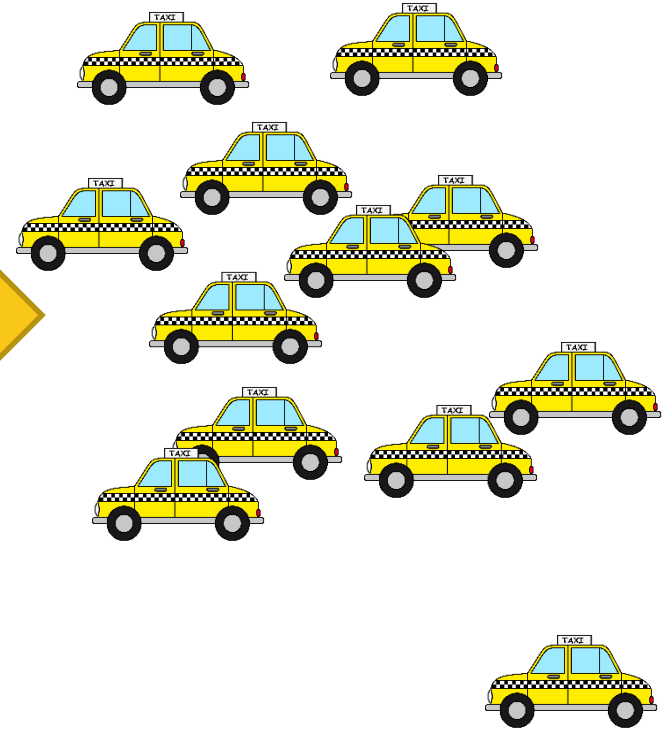
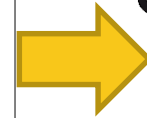
EXAMPLES Server: 50_Applications/18_Churn_Prediction

Demand Prediction (Taxi)

Demand Prediction: The Problem

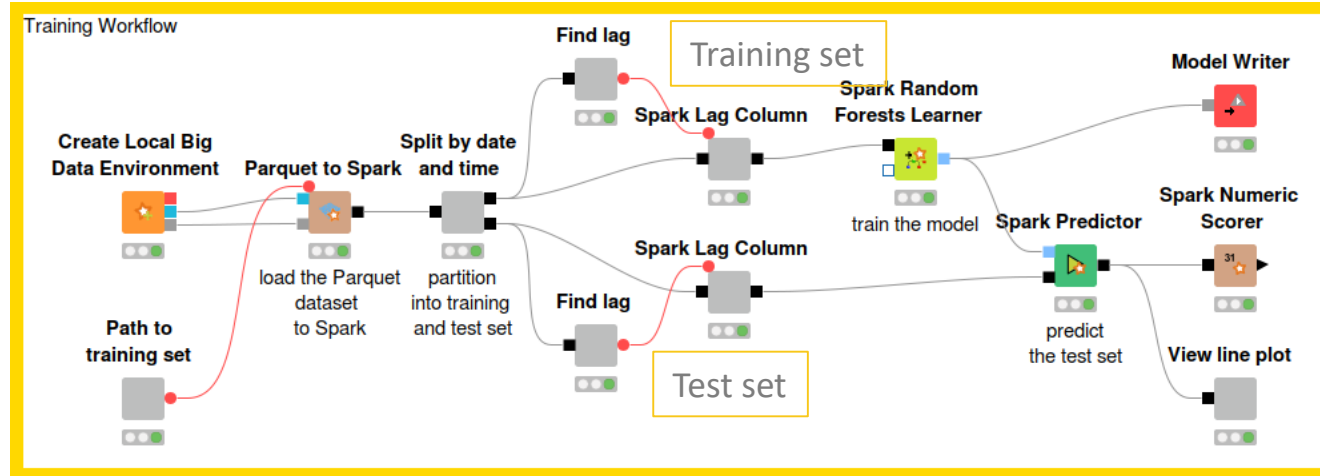


How many taxi
do I need in NYC
on Wednesday
at 12:00?

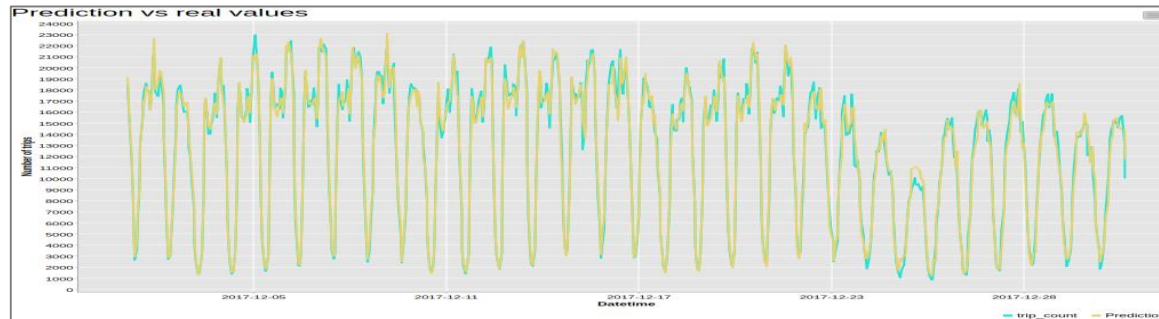


How many customers?
How many kW?
How many diapers?

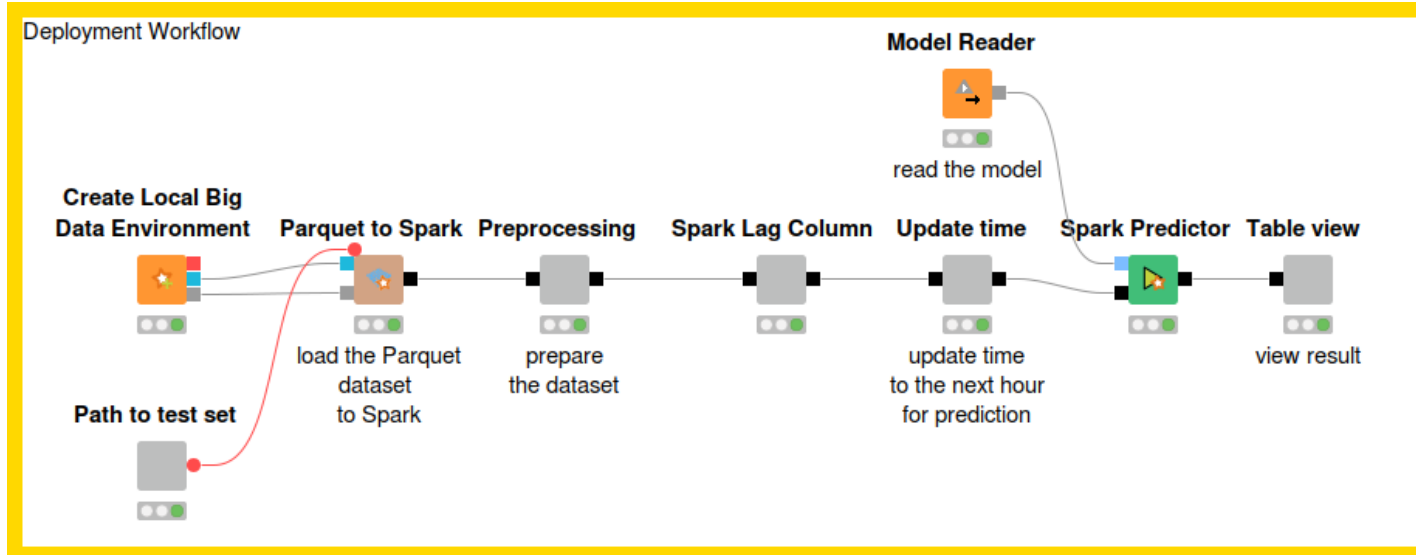
Demand Prediction: The Training Workflow



R2 = 0.81



Demand Prediction: Deployment



On Wednesday
at 12:00 we
need **13k taxis**

Automated Machine Learning



Interaction Points

Business analysts will simply access the *KNIME WebPortal* from any web browser..

The image displays the KNIME WebPortal interface with five numbered interaction points highlighted in yellow boxes. A navigation bar at the top contains icons for each step: 1. Upload Dataset, 2. Select Target, 3. Filter Columns, 4. Select Models, 5. Execution Settings, and a Download Models icon. Arrows indicate the flow between these steps.

1 Upload Dataset

Upload the dataset to be used.

Uploading file "adult.csv" Cancel

2 Select Target

Select the target column whose values should be predicted.

Select:

Row ID	Workclass	Education	Education-Num	Marital St.
Row0	State-gov	Bachelors	13	Never-mar

3 Filter Columns

Set Column Relevance Filter

Use the slider to select a subset of columns based on their relevance. If in doubt, do not change.

4.55

0.00 Overall Column Relevance 100.00

Feature Name	Overall Column Relevance	Correlation with Target (%)	ID/Noise Test (%)	Constant Value Test (%)	Missing Value Test (%)
Age	97.13	33.707	0.7	2.87	0
Occupation	87.12	35.595	0.14	12.88	0

4 Select Models

Choose one or more machine learning models to train for your prediction task.

Simple models

- Naive Bayes
- Decision Tree
- Logistic Regression

Complex models

- Support Vector Machine
- Random Forest
- Generalized Linear Models
- Gradient Boosted Trees
- Deep Learning

5 Execution Settings

Please select the desired distributed environments for the execution of the workflow.

Available options:

- Local execution
- Use Spark cluster if possible
- Use Apache Spark MLlib
- Use other cluster environment

Guide

Set Column Relevance Filter

By default, all columns will be used to train the model that creates the prediction. However, not all columns contribute with the same importance or relevance to the final prediction. In some cases, columns are not informative or contain spurious information. To help you decide, the overall column relevance towards the final prediction is measured.

- Column Relevance** is an overall metric summarizing the metrics below. Use the slider to select the input features based on their Overall Column Relevance.

The additional metrics calculated automatically and used to determine Overall Column Relevance include:

- ID/Noise Test** measures how likely the column is a representation used to identify each row in your table. Row identifiers are uninformative for your model and should be removed.
- Constant Value Test** measures how often the column contains the exact same value. Columns with a constant value should be removed.

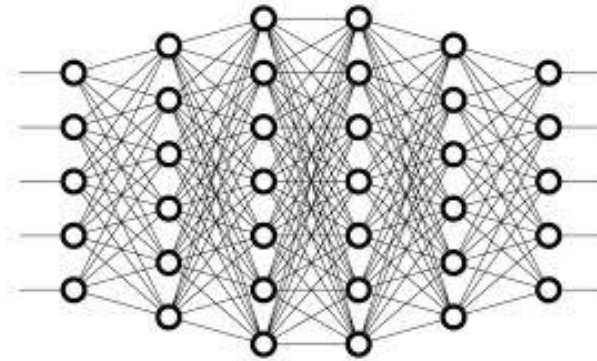
Fraud/Anomaly Detection



Fraud Detection: The Problem

Transactions

- Trx 1
- Trx 2
- Trx 3
- Trx 4
- Trx 5
- Trx 6
- ...



- Good
- Good
- Good

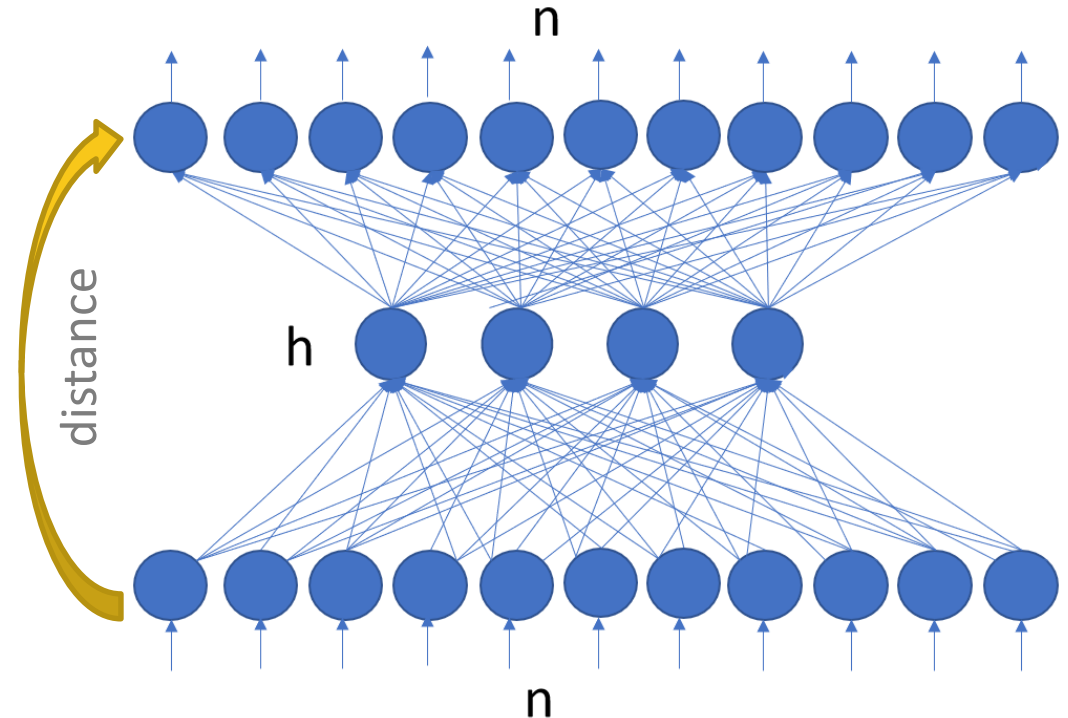
• **Fraud**

- Good
- Good
- ...

Model

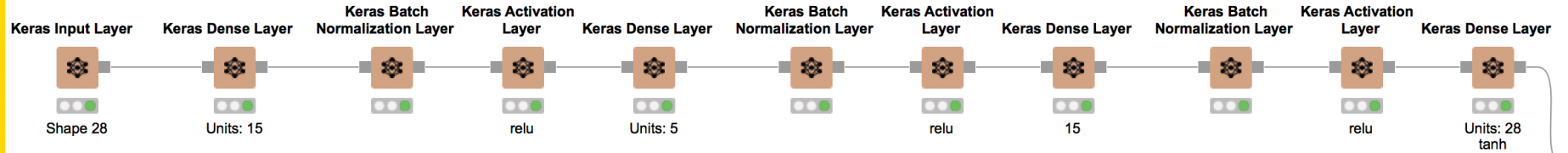
Fraud Detection: without Fraud Examples – Auto-encoder

- Trained with Back-Propagation on just “normal” transactions
- If distance > threshold => possible fraud

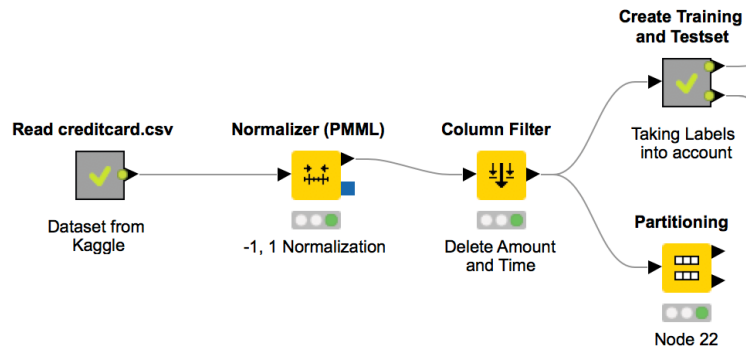


Fraud Detection: without Fraud Examples

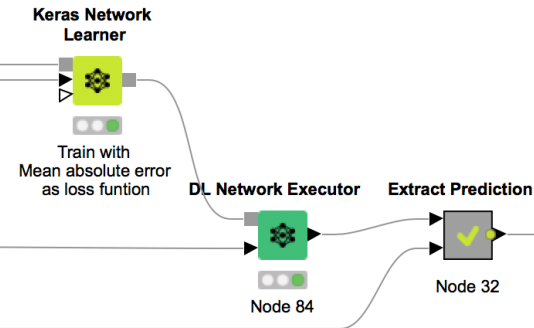
Define Network



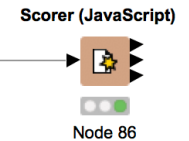
Preprocessing



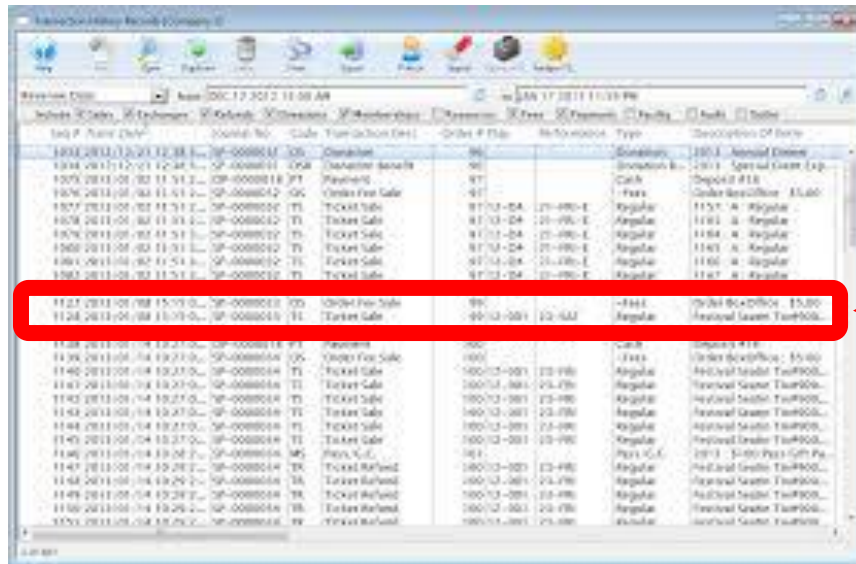
Training and Predicting



Evaluation



Fraud Detection deployed



The screenshot shows a Windows desktop with a data table application. The table has columns for 'Tag #', 'Name (City)', 'Account No.', 'Code', 'Type/Action Desc', 'Order #', 'Date', 'Amount', 'Type', and 'Description Of Item'. A red box highlights two rows: 1127 and 1128. An arrow points from the text 'Suspicious Transaction' to the red box.

Tag #	Name (City)	Account No.	Code	Type/Action Desc	Order #	Date	Amount	Type	Description Of Item
1033	2013-02-21 12:38 L...	SP-0000010	05	Overseas	96			Overseas	100.0 - Annual Dinner
1034	2013-02-21 12:44 L...	SP-0000011	05A	Overseas	96			Overseas	200.0 - Special Guest Exp.
1075	2013-02-22 11:51 L...	SP-0000018	31	Payment	97			Cash	Deposit #18
1076	2013-02-22 11:51 L...	SP-0000012	02	Order Fee Sale	97			Order	Order Receipt - \$5.00
1077	2013-02-22 11:51 L...	SP-0000014	71	Ticket Sale	87	13-04	20-000-E	Regular	1151.0 - Regular
1078	2013-02-22 11:51 L...	SP-0000012	71	Ticket Sale	87	13-04	20-000-E	Regular	1183.0 - Regular
1079	2013-02-22 11:51 L...	SP-0000012	71	Ticket Sale	87	13-04	20-000-E	Regular	1184.0 - Regular
1080	2013-02-22 11:51 L...	SP-0000012	71	Ticket Sale	87	13-04	20-000-E	Regular	1183.0 - Regular
1081	2013-02-22 11:51 L...	SP-0000012	71	Ticket Sale	87	13-04	20-000-E	Regular	1186.0 - Regular
1082	2013-02-22 11:51 L...	SP-0000012	71	Ticket Sale	87	13-04	20-000-E	Regular	1187.0 - Regular
1127	2013-02-28 15:00 L...	SP-0000010	05	Order Fee Sale	96			Order	Order Receipt - \$5.00
1128	2013-02-28 15:00 L...	SP-0000010	11	Ticket Sale	85	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1129	2013-02-28 15:00 L...	SP-0000018	31	Payment	96			Cash	Deposit #18
1130	2013-02-28 15:00 L...	SP-0000014	05	Order Fee Sale	100			Order	Order Receipt - \$5.00
1140	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1141	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1142	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1143	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1144	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1145	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1146	2013-02-28 15:00 L...	SP-0000014	05	Order Fee Sale	100			Order	Order Receipt - \$5.00
1147	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1148	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1149	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1150	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...
1151	2013-02-28 15:00 L...	SP-0000014	71	Ticket Sale	100	13-02	20-000-E	Regular	Annual Dinner Ticket#00...

Suspicious Transaction

Recommendation Engine



Recommendation Engines or Market Basket Analysis



Model

Rule: $X \Rightarrow Y$

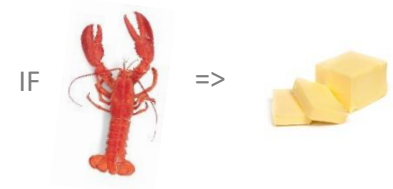
$$Support = \frac{frq(X, Y)}{N}$$

$$Confidence = \frac{frq(X, Y)}{frq(X)}$$

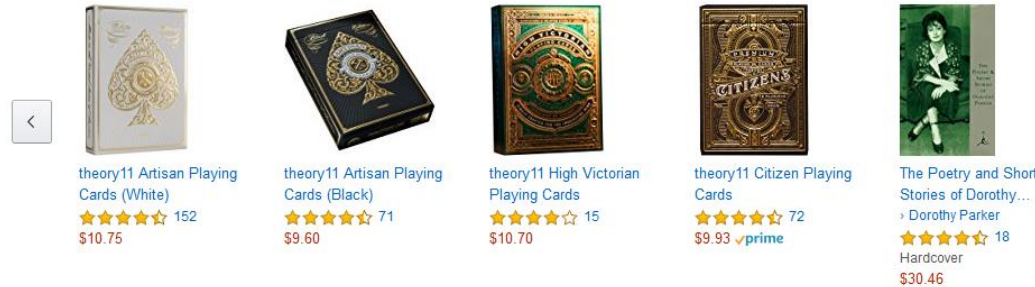
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Recommendation



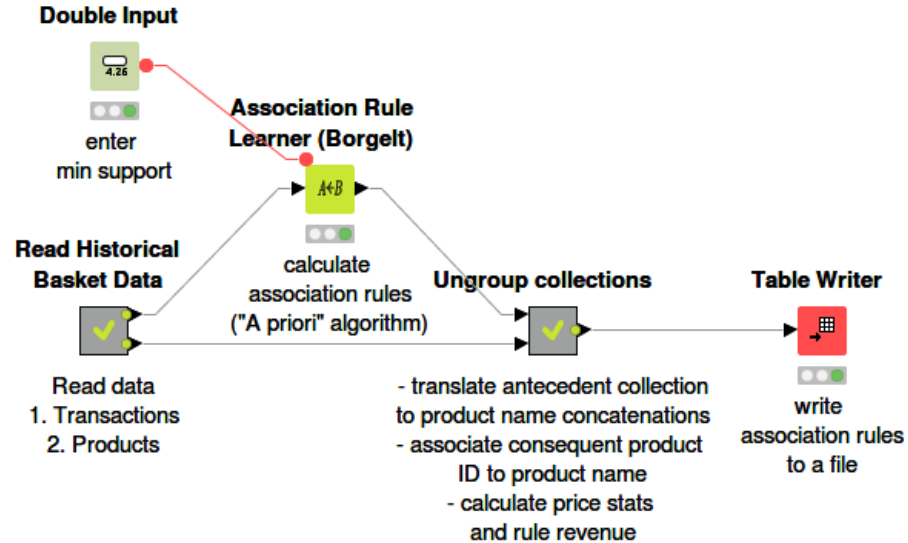
Inspired by your purchases



Market Basket Analysis: with Association Rules

Market Basket Analysis: Build Association Rules

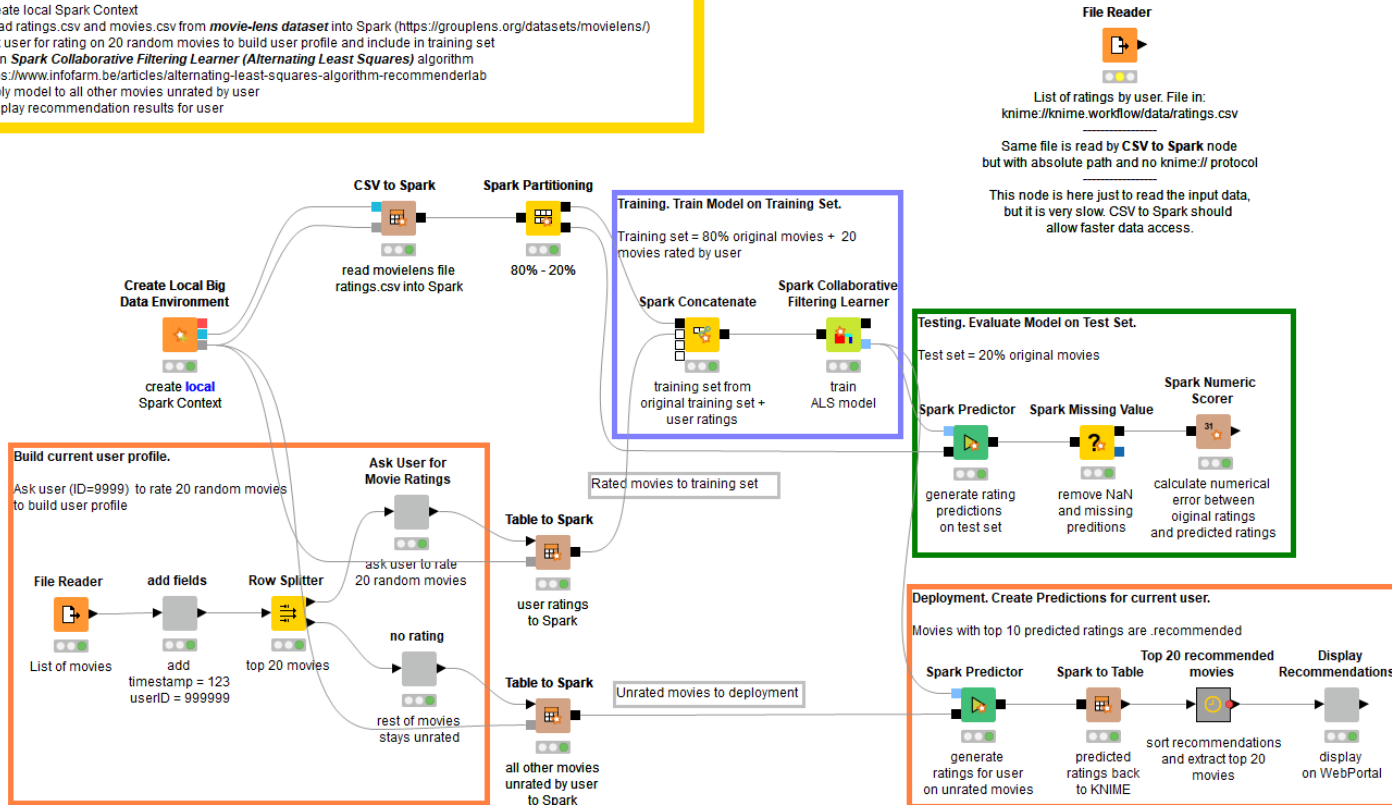
1. Read Transaction/Basket data and Product data
2. Using "A priori" algorithm, build association rule set
 - min. set size = 1
 - min rule confidence = 10%
 - min support is controlled by Double Input Quickform node in %
3. Translate Antecedent collections into product name concatenations
4. Translate Consequent Item ID into Consequent Product Name
5. Calculate price stats and rule revenue
6. Write association rule set to file



Recommendation Engine: with Collaborative Filtering

Movie Recommendation Engine with Spark Collaborative Filtering

1. Create local Spark Context
2. Read ratings.csv and movies.csv from *movie-lens dataset* into Spark (<https://grouplens.org/datasets/movielens/>)
3. Ask user for rating on 20 random movies to build user profile and include in training set
4. Train *Spark Collaborative Filtering Learner (Alternating Least Squares)* algorithm
<https://www.infofarm.be/articles/alternating-least-squares-algorithm-recommenderlab>
5. Apply model to all other movies unrated by user
6. Display recommendation results for user



Recommendation Engine/MBA: Deployment

Basket Analysis Report

Welcome to our Supermarket Chain!

The total price for your current shopping cart is **79.17\$!**

Purchase Advices

1. Try our **lobster** !

Today's price for lobster is just 23.72\$!

... and if you like our **shrimps**, we are sure you will also enjoy the **lobster** !

2. Try our **lobster** !

Today's price for lobster is just 23.72\$!

... and if you like our **cookies**, we are sure you will also enjoy the **lobster** !



Created with KNIME Report Designer. Provided by KNIME.com AG, Zurich, Switzerland



Creative AI

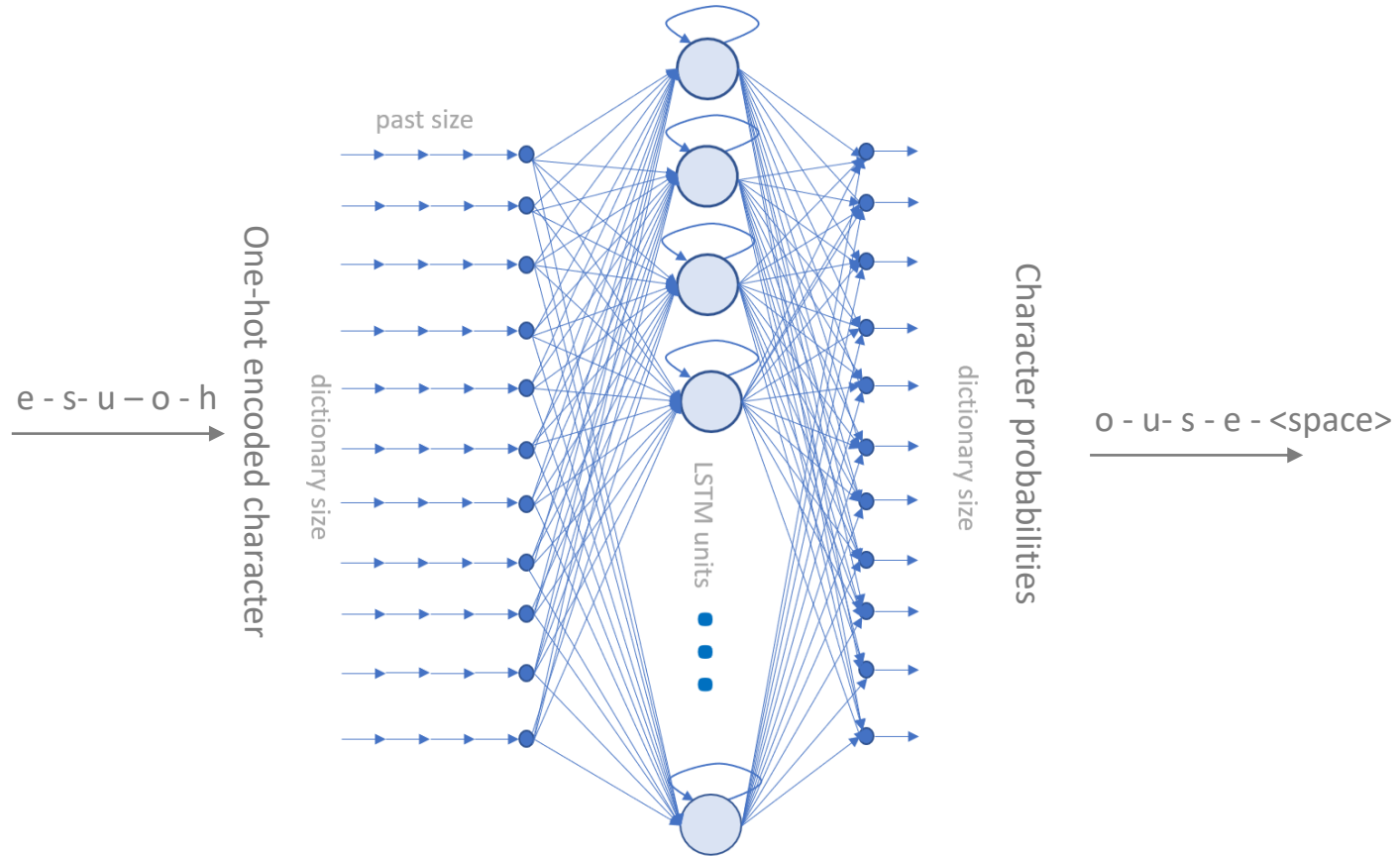


Creative AI: The Problems

- Free Text Generation
 - Simulating a writing style
 - Writing in different languages
 - Providing an answer in a specific style
- Machine Translation
- Generating Candidates for Product Names

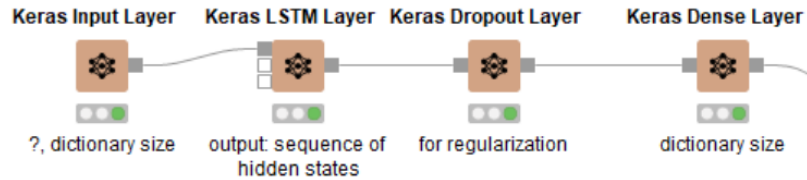


Deep Learning LSTM Network

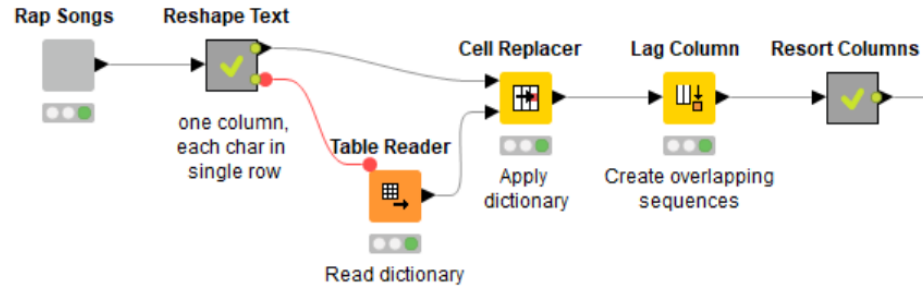


Creative AI: The Training Workflow

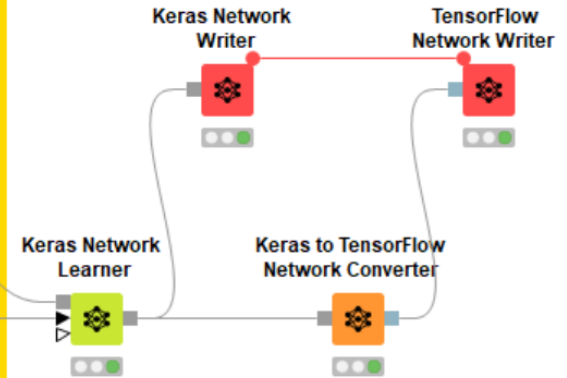
Define Network Structure



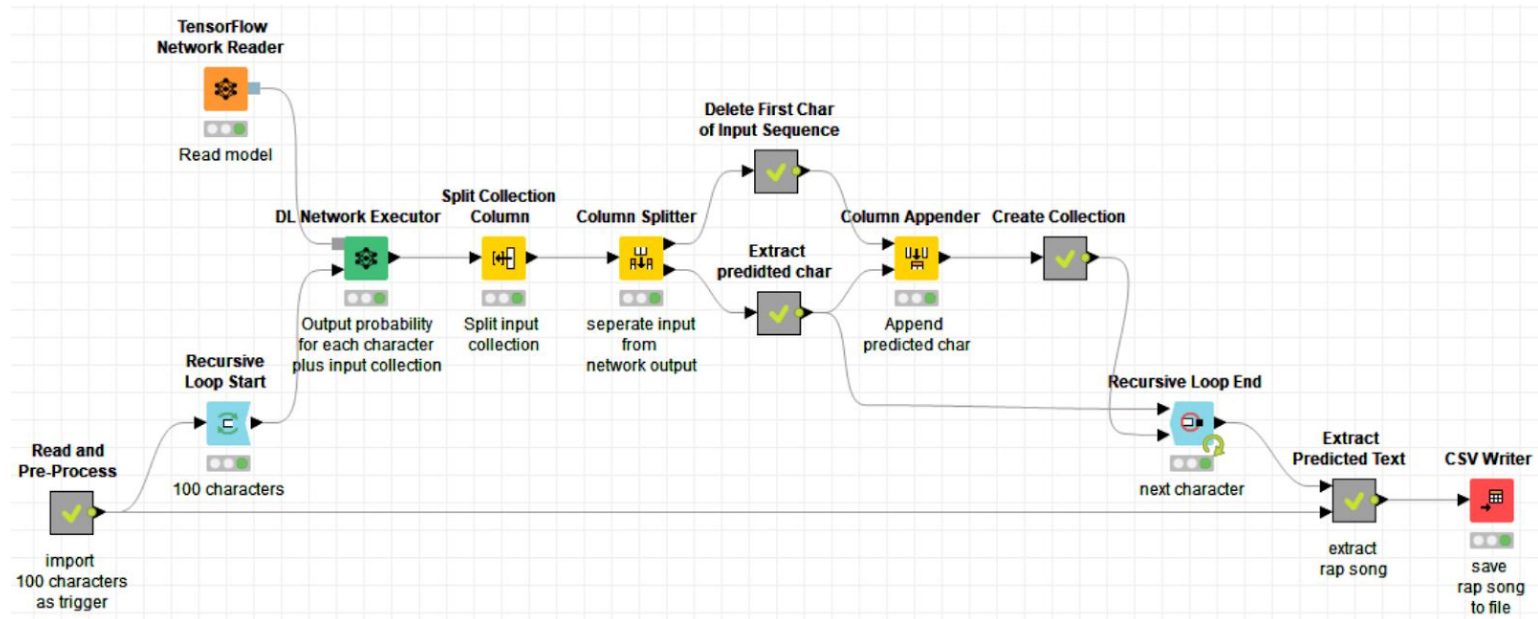
Pre Processing and Encoding



Train and Save Network



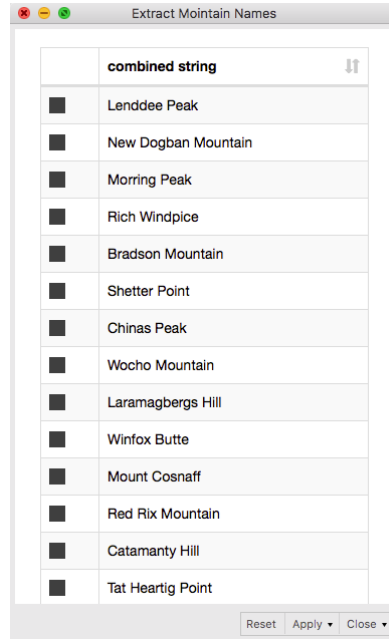
Creative AI: The Deployment Workflow



Creative AI: Deployment and Results

Yo!

*This post is about generating free text with a deep learning network particularly it is about Brick X6, Phey, cabe, make you feel soom the way (I smoke good!) I probably make (What?) More money in six months, Than what's in your papa's safe (I'm serious) Look like I robbed a bank (Okay Okay) I set it off like Queen Latifah 'Cause I'm living single I'm feeling cautious I ain't scream when they served a subpoena (Can't go back to jail) I heard that he a leader (Who pood, what to be f***** up The baugerout Black alro Black X6, Phantom White X6 looks like a panda Goin' out like I'm Montana Hundred killers, hundred hammers Black X6, Phantom White X6, panda Pockets swole, Danny Sellin' bar, candy Man I'm the macho like Randy The choppa go Oscar for Grammy B**** n**** pull up ya panty Hope you killas understand me Hey Panda, Panda Panda, Panda, Panda, Panda I got broads in Atlanta Twistin' dope, lean, and the Fanta Credit cards and the scammers Hittin' off licks in the bando*



	combined string
<input type="checkbox"/>	Lenddee Peak
<input type="checkbox"/>	New Dogban Mountain
<input type="checkbox"/>	Morring Peak
<input type="checkbox"/>	Rich Windpice
<input type="checkbox"/>	Bradson Mountain
<input type="checkbox"/>	Shetter Point
<input type="checkbox"/>	Chinas Peak
<input type="checkbox"/>	Wocho Mountain
<input type="checkbox"/>	Laramagbergs Hill
<input type="checkbox"/>	Winfox Butte
<input type="checkbox"/>	Mount Cosnaff
<input type="checkbox"/>	Red Rix Mountain
<input type="checkbox"/>	Catamanty Hill
<input type="checkbox"/>	Tat Heartig Point

This License refers to version of the GNU General Public License. Copyright also means copyright-bick, Remade me any thing to his sword To his salt and most hidden loose to be so for sings, but not in a libut of his matter than that shall be sure as will be soldye As master compary, do not live in traitor. Bless thy five wits!

-Kent

O pity! Sir, where is the patience now, That this is so far from the sea and some bidings to dismantle So many folds of save and honest.

-Brabantio

I must not think the Turk of Cassio in the strange metting the cribles of a charmer be the reviling of libe to say That I can deceive him to the best advantage, In her prophetic fairs of a little to presently at your powers; whereof I thank you, sir.

-Albany

Gloucester, I will prove upancy of his sport and first accuriors and guard and talking on the white.

-King Lear

Where are the thief? Thou shalt never have the captains at the letter To the Moor and thing we have not the better shall be sure as worth if he be anger—

-Regan

I pray you, have a countend more than think to do a proclaim'd there of my heart, Hot The words save, honest, thief, master, traitor, and deceive seem to fit the context. Notice also that the dialogue sprouting from the start text of the license agreement interestingly involves mainly minor, less tragic characters from the plays.

Caro amico ti scrivo così mi distraigo un po'E siccome sei molto lontano più forte ti scriverò.

Da quella prima folla strana, che aveva preso il suo nome, e di correre alla casa di don Abbondio, con un viso bene di non poterci andar la casa del padre Cristoforo, e gli disse che s'avvicinava all'uscio, e si mise a sparse di corsa, e di stare a sé, verso la strada di servizio, chiesto le parole che gli andavan dall'altra stanza, e con la sua condizione de' cappuccini, e di consigli ricerche di confidenza delle gride, nel suo passaggio, se non pensava con una certa ripugnanza a casa sua, che andavano a scomparire in un campo di buone ragioni che avevan potuto raccogliere i suoi pensieri, e di sopra non senza interrogare, che la sua avventura aveva fatto predicare, e con la forza d'un fatto come fuggitive che aveva preso il suo nome, e di correre alla casa di don Abbondio, con un cappuccino di quella sorte, con un certo sospiro, alzando le sue finestre, e le diede un'occhiata in carrozza. Si vendano a metter nelle mani di chi era stato a sedere sur una strada così fatta con le braccia in

Free Book as a Thank You

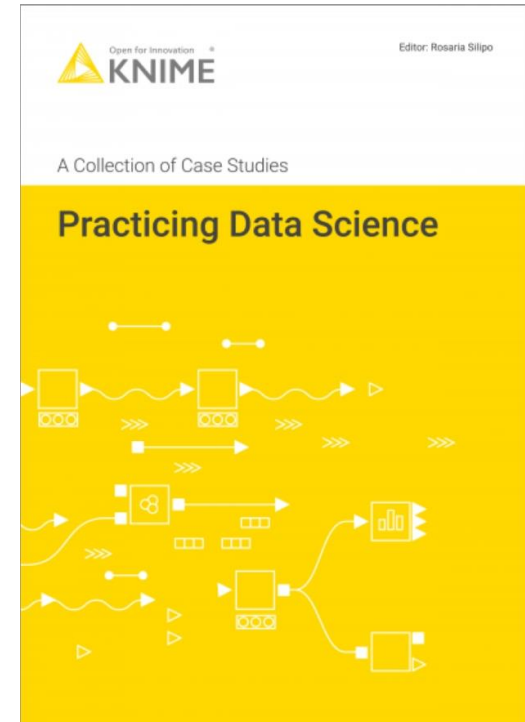
Free Copy of *Practicing Data Science. A Collection of Case Studies* Book from **KNIME Press**

<https://www.knime.com/knimepress>

with this code: **STRATA-LONDON-2019**

Expiration date

Tue, 06/11/2019 - 23:59



Rate today's session

Cyberconflict: A new era of war, sabotage, and fear

David Sanger (The New York Times)
9:55am-10:10am Wednesday, March 27, 2019
Location: Ballroom
Secondary topics: Security and Privacy

See passes & pricing

 Add to Your Schedule
 Add Comment or Question

Rate This Session

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.




Session page on conference website

✓ Attending Notes Remove

Cyberconflict: A new era of war, sabotage, and fear

9:55 AM - 10:10 AM, Wed, Mar 27, 2019


Speakers

 David Sanger
National Security Correspondent
The New York Times

📍 Ballroom

Keynotes

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

 SESSION EVALUATION

O'Reilly Events App

The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.