

Oppdag det forventede og uventede ved å bruke tekstanalyse i VA

FANS Nettverksmøte – 22. og 23.mai 2024

fans*

Vegard Hansen

Academic Lead @ SAS

Global Academic Program, Education

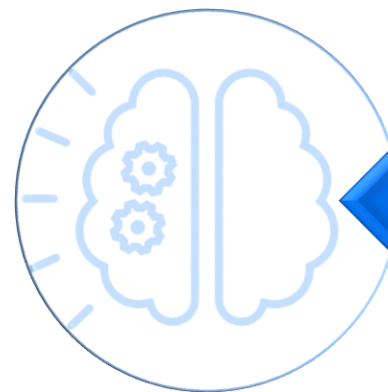
vegard.hansen@sas.com

www.linkedin.com/in/vegard-hansen/



SAS Global Academic Program

Creating a Pipeline of Skilled Future Users



Build SAS Skill Aligned with Workforce Demand



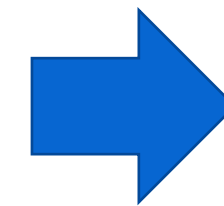
Provide Free Teaching and Learning Resources & Platforms and Recognition



Connect Graduates with SAS Customers for Hiring



SAS (Certified) Specialist



[SAS SKILL BUILDER for Students](#)

Self-study portal for students

Totally free

From Programming to Machine Learning

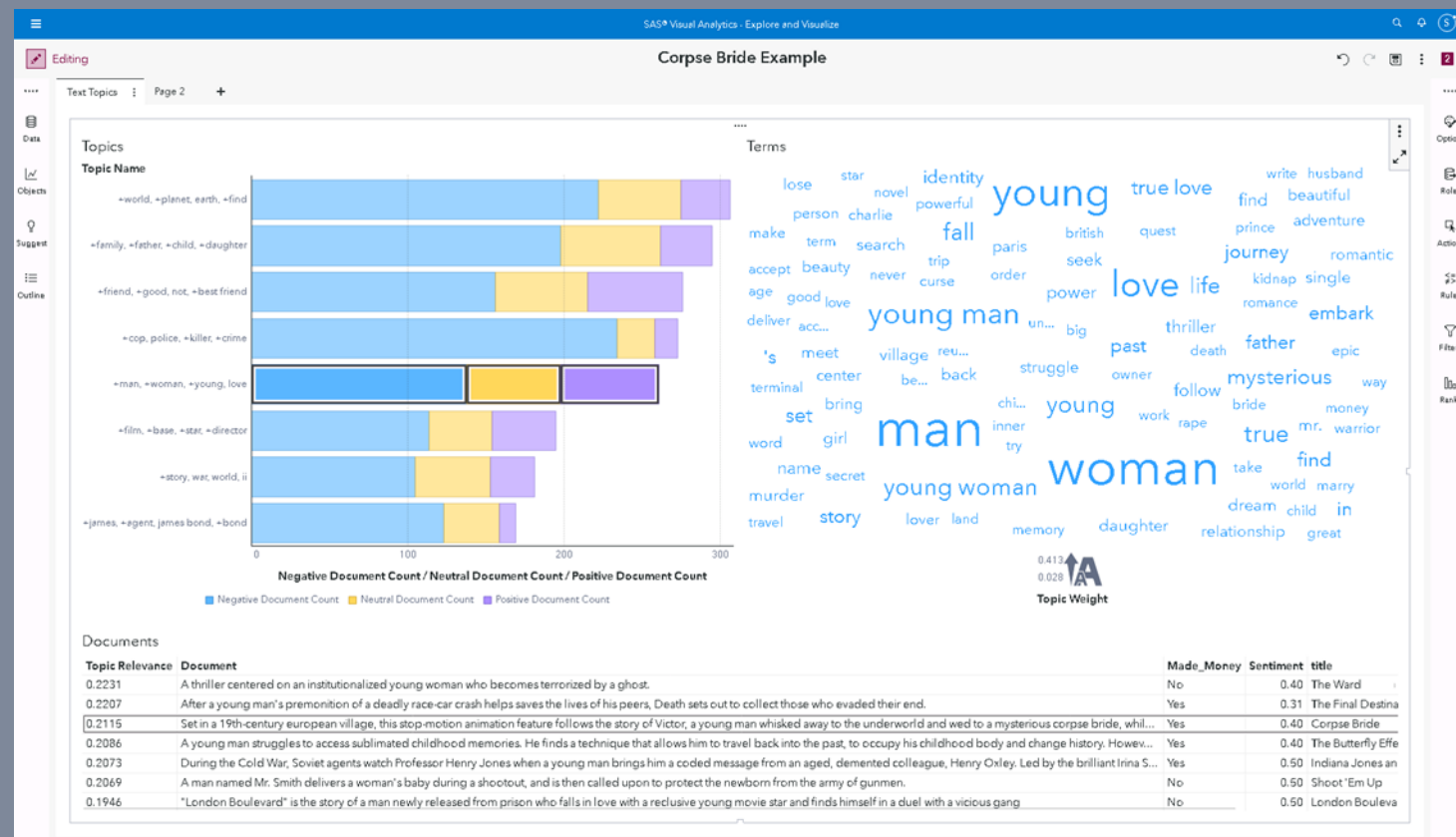
The + of Badges on the CV

[SAS EDUCATORS Portal](#)

Teacher Resource Portal

Totally free

How text/unstructured data can be used to improve the value of your predictive models



Text Topics Object in SAS VA

Text or unstructured data can significantly improve the value of predictive models

1. Sentiment Analysis:

Analyses sentiments from text data like reviews or comments to understand user preferences and trends, which can be used to make more accurate predictions.

2. Topic Modeling:

Identifies topics within the text data to uncover hidden patterns and insights that can be useful for prediction.

3. Named entity recognition (NER)

NER is a text analytics technique used for identifying named entities like people, places, organizations, and events in unstructured text.

4. Term frequency – inverse document frequency

TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term's importance to the document.

5. Event extraction

This is a text analytics technique that is an advancement over the named entity extraction. Event extraction recognizes events mentioned in text content, for example, mergers, acquisitions, political moves, or important meetings.

Meet the Text Topics Object

The screenshot displays the SAS Visual Analytics interface. On the left, the 'Objects' panel shows the 'Text topics' object selected and highlighted with a red circle. A red arrow points from this object to the main visualization area. The main area is titled 'Corpse Bride Example' and shows a 'Text Topics' visualization. This visualization includes a horizontal bar chart with eight bars, each representing a topic. The bars are segmented into three colors: blue (Negative Document Count), yellow (Neutral Document Count), and purple (Positive Document Count). To the right of the chart is a word cloud of terms associated with the topics. Below the chart is a table of documents with columns for Topic Relevance, Document, Made_Money, Sentiment, and title.

Topics

Topic Name	Negative Document Count	Neutral Document Count	Positive Document Count
+world, +planet, earth, +find	~100	~100	~100
+family, +father, +child, +daughter	~100	~100	~100
+friend, +good, not, +best friend	~100	~100	~100
+cop, police, +killer, +crime	~100	~100	~100
+man, +woman, +young, love	~100	~100	~100
+film, +base, +star, +director	~100	~100	~100
+story, war, world, ii	~100	~100	~100
+james, +agent, james bond, +bond	~100	~100	~100

Terms

lose star identity young true love write husband
 person charlie powerful fall paris british quest find beautiful
 make term search trip curse order seek journey romantic
 accept beauty never order power love life kidnap single
 age good love young man un... big past thriller father epic
 deliver acc... 's meet village reu... struggle owner follow mysterious way
 terminal center be... back chi... young work rape bride money
 set bring inner try word girl man inner try work rape true mr. warrior
 name secret young woman woman take find
 murder story lover land memory daughter relationship great

Documents

Topic Relevance	Document	Made_Money	Sentiment	title
0.2231	A thriller centered on an institutionalized young woman who becomes terrorized by a ghost.	No	0.40	The Ward
0.2207	After a young man's premonition of a deadly race-car crash helps saves the lives of his peers, Death sets out to collect those who evaded their end.	Yes	0.31	The Final Destina
0.2115	Set in a 19th-century european village, this stop-motion animation feature follows the story of Victor, a young man whisked away to the underworld and wed to a mysterious corpse bride, whil...	Yes	0.40	Corpse Bride
0.2086	A young man struggles to access sublimated childhood memories. He finds a technique that allows him to travel back into the past, to occupy his childhood body and change history. Howev...	Yes	0.40	The Butterfly Effe
0.2073	During the Cold War, Soviet agents watch Professor Henry Jones when a young man brings him a coded message from an aged, demented colleague, Henry Oxley. Led by the brilliant Irina S...	Yes	0.50	Indiana Jones an
0.2069	A man named Mr. Smith delivers a woman's baby during a shootout, and is then called upon to protect the newborn from the army of gunmen.	No	0.50	Shoot 'Em Up
0.1946	"London Boulevard" is the story of a man newly released from prison who falls in love with a reclusive young movie star and finds himself in a duel with a vicious gang	No	0.50	London Bouleva

Data Roles

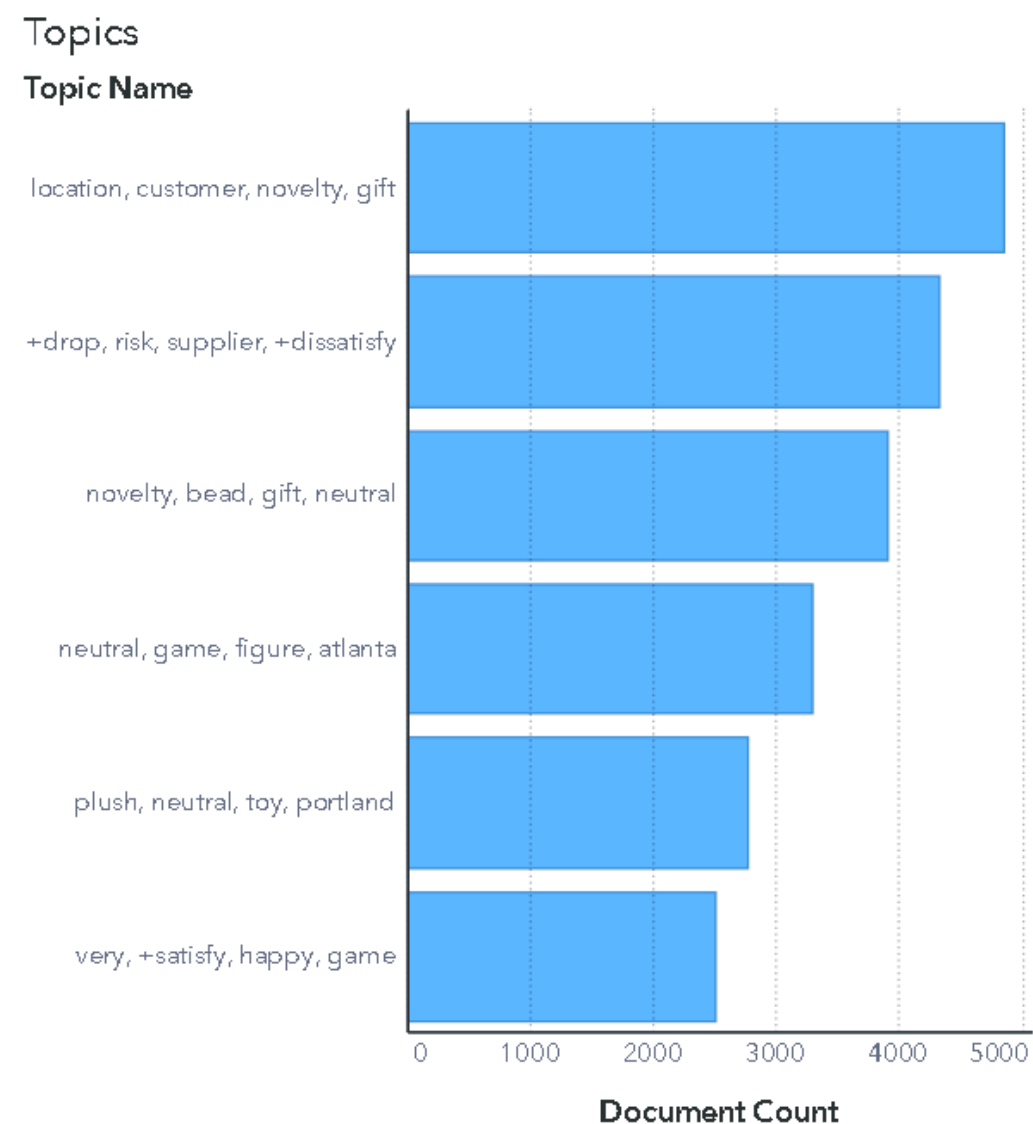
There must be a unique ID in your data!

The screenshot displays the SAS Text Topics interface. At the top, there are navigation tabs for 'Text Topics' and 'Text Topics Changed Options', along with 'Page 3' and a zoom icon. The main area is divided into three panels:

- Topics:** A list of topic names with corresponding blue horizontal bars. The visible names are 'location, cu...', '+drop, risk,...', and 'novelty, be...'. There are also three unlabeled bars at the bottom.
- Terms:** A word cloud of terms associated with the topics. The most prominent terms are 'order', 'supplier', 'toy', 'customer', 'game', and 'place'. Other terms include 'basran', 'jacksonville', 'ok', 'portland', 'neutral', 'orlando', 'happy rock', 'colorado springs', 'charlotte', 'hawaii', 'richmond', 'cleveland', 'las vegas', 'madison', 'nashville', 'wichita', 'please', 'beach', 'rock', 'minneapolis', 'miami', 'el paso', 'washington', 'st. louis', 'atlanta', 'st. louis', 'jefferson', 'bismark', 'beach', 'plush', 'jackson', 'satisfy', 'columbus', 'des', 'concord', 'raleigh', 'promo', 'game', 'dissatisfy', 'dover', 'atlanta ready', 'charlotte', 'reno', 'bellevue', 'locate', 'louisville', 'omaha', 'thrift', and 'hisco'.
- Data Roles:** A panel on the right with a title 'Data Roles'. It contains a dropdown menu set to 'Text topics - OrderNote 1'. Below it are two expandable sections: 'Document collection' containing 'OrderNote' and 'Document details' containing '+ Add'. At the bottom is another dropdown menu set to 'English'. On the far right, there is a vertical toolbar with icons for 'Options', 'Roles', 'Actions', 'Rules', and 'Filters'.

Output

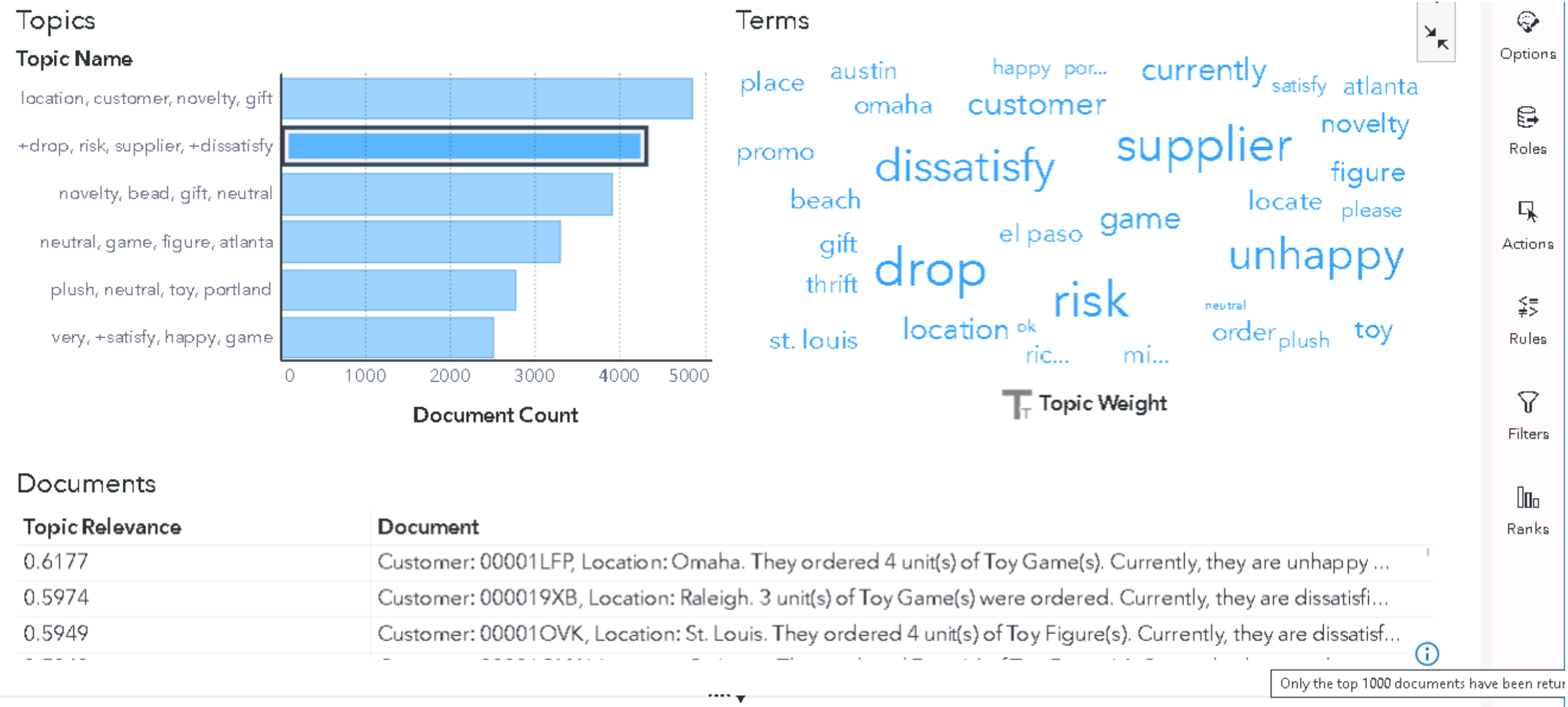
- Suggested topics
- Document count for each topic
- Word Cloud with word frequency



Selected Topic

Click on a topic

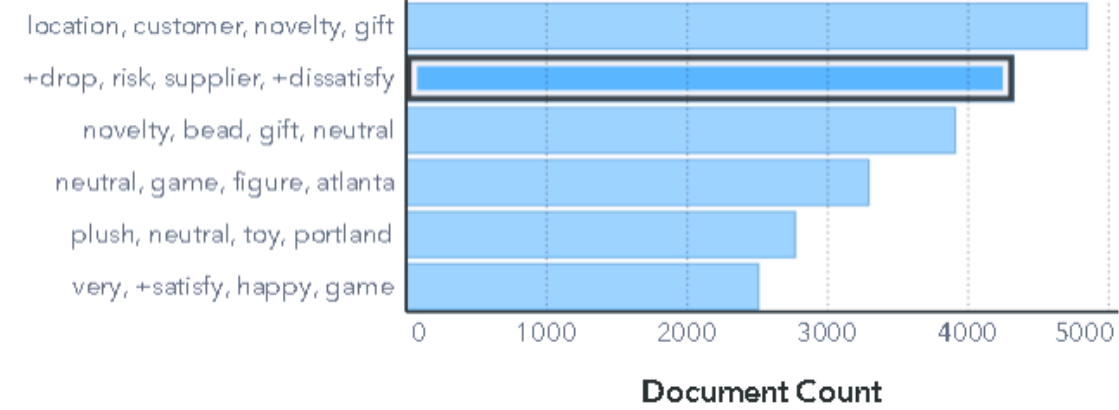
- Word Cloud shows term weights related to the selected topic
- Documents are shown ranked by topic relevance



Details Table

Topics

Topic Name



Terms



Documents

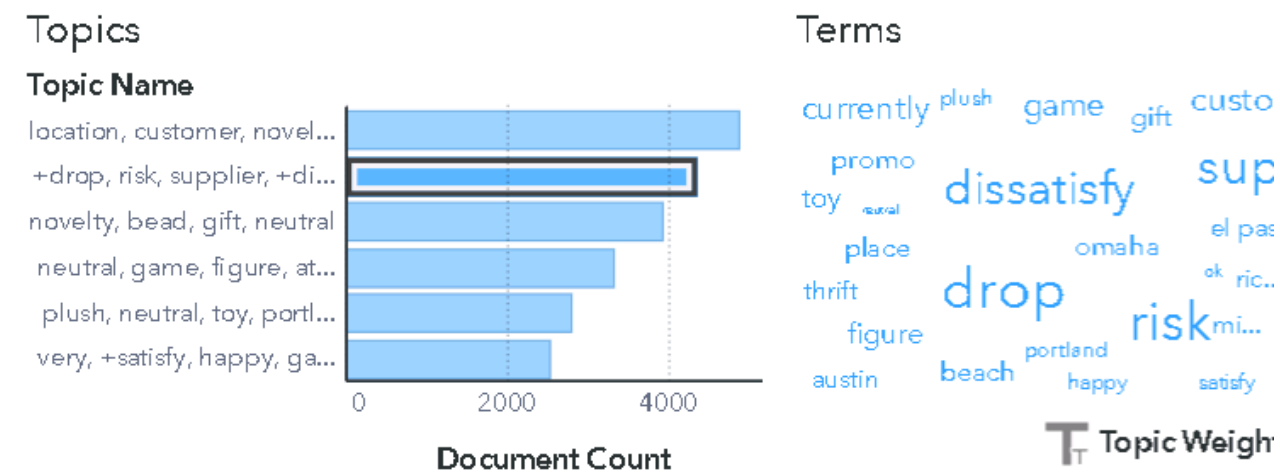
Topic Relevance	Document
0.6177	Customer: 00001LFP, Location: Omaha. They ordered 4 unit(s) of Toy Game(s). Currently, they are unhappy with us as ...
0.5974	Customer: 000019XB, Location: Raleigh. 3 unit(s) of Toy Game(s) were ordered. Currently, they are dissatisfied with u...
0.5949	Customer: 00001OVK, Location: St. Louis. They ordered 4 unit(s) of Toy Figure(s). Currently, they are dissatisfied with ...

Topics Terms Text Topics Summary

Topic Name	Document Count
location, customer, novelty, gift	4859
+drop, risk, supplier, +dissatisfy	4331
novelty, bead, gift, neutral	3913
neutral, game, figure, atlanta	3301
plush, neutral, toy, portland	2777
very, +satisfy, happy, game	2514

Terms Table

is a numerical representation of the Word Cloud



Documents

Topic Relevance	Document
0.6177	Customer: 00001LFP, Location: Omaha. They ordered 4 unit(s) of
0.5974	Customer: 000019XB, Location: Raleigh. 3 unit(s) of Toy Game(s)
0.5949	Customer: 00001OVK, Location: St. Louis. They ordered 4 unit(s)

Term	Topic Weight	Role
drop	0.475	Verb
risk	0.438	Noun
supplier	0.408	Noun
dissatisfy	0.356	Verb
unhappy	0.347	Adjective
customer	0.156	Noun
game	0.139	Proper noun
currently	0.138	Adverb
location	0.106	Noun

Derive Topics

with a right mouse click

The screenshot illustrates the process of deriving topics in SAS. On the left, a 'Topics' table lists topic names and their associated terms. A context menu is open over the first topic, with 'Derive topics...' highlighted. A red arrow points from this menu item to the 'Data' menu in the top right, which is also circled in red. The 'Data' menu shows a list of derived topics, with the first one selected. On the right, a horizontal bar chart shows the 'Event Count' for various topics, with the x-axis ranging from 3000 to 5000.

Topic Name	Terms
location, customer, novelty, gift	omaha, dover, pittsburgh, atl, ok, mac, ansas city, toy
+drop, risk, supplier, +dissatisfy	ok, mac, ansas city, toy
novelty, bead, gift, neutral	toy, ol, batc, orlando, raleigh, orleans
neutral, game, figure, atlanta	toy, ol, batc, orlando, raleigh, orleans

Topics

Topic Name

location, customer, novelty, gift

+drop, risk, supplier, +dissatisfy

novelty, bead, gift, neutral

neutral, game, figure, atlanta

Terms

omaha dover
pittsburgh atl
ok mac
ansas city
toy
ol
batc
orlando
raleigh
orleans

Data

ORDERS43K

Filter

+ New data item

VendorType

Aggregated Measure

Frequency Percent

Derived (Text Topics: December 22, 202...)

Relevance: +drop, risk, supplier, +... - 2

Relevance: location, customer, novelty...

Relevance: neutral, game, figure, atlanta

Relevance: novelty, bead, gift, neutral

Relevance: plush, neutral, toy, portland

Relevance: very, +satisfy, happy, game

Topic: +drop, risk, supplier, +diss - 2

Event Count

3000 4000 5000

in Boise placed an order.

in Birmingham placed an c

is for 00004RPF located in

Options for Text Topics

You can adjust:

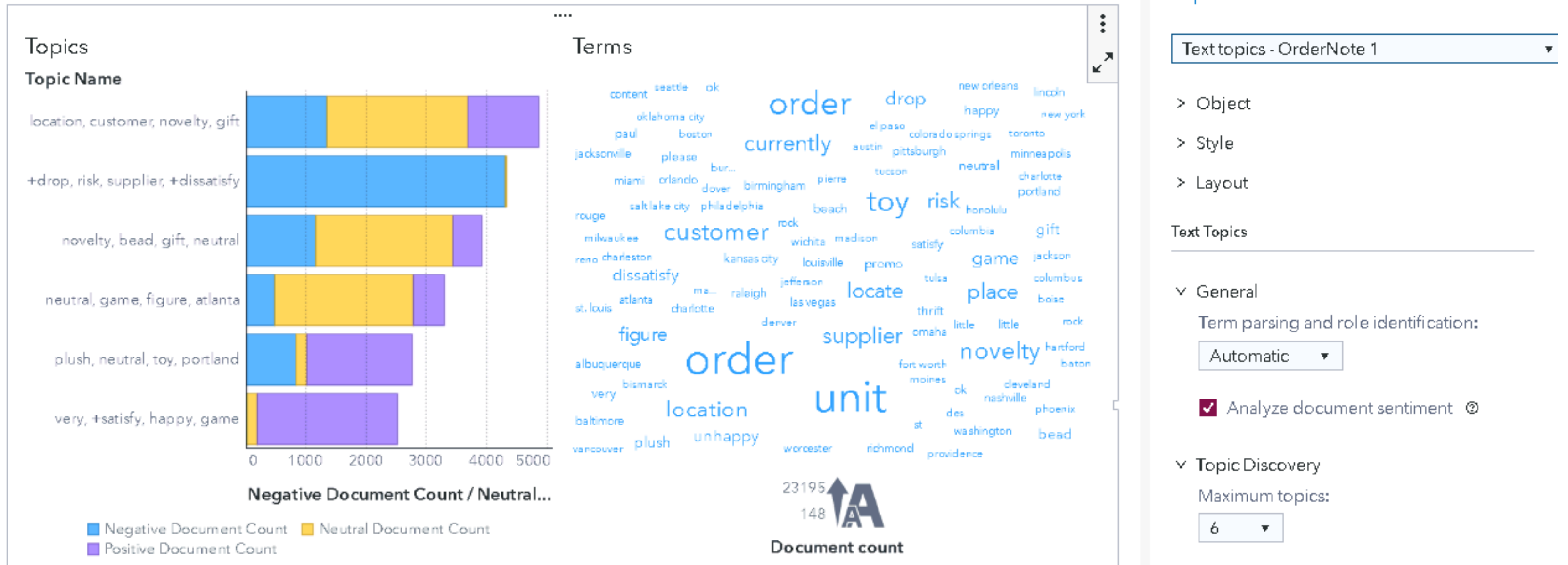
- Parsing (how the terms are built)
- The maximum number of topics
- Emphasize rare terms
- The length of the topic name
- Sentiment Analysis

The screenshot displays the SAS Text Topics interface. On the left, a navigation pane includes 'Data', 'Objects', 'Suggest', and 'Outline'. The main area is titled 'Text Topics' and shows 'Text Topics Changed Options' for 'Page 3'. It features a 'Topics' section with a horizontal bar chart showing document counts for various topic names. The 'Terms' section displays a word cloud of terms. Below the chart and word cloud, there are 'Documents' and 'Document count' sections. On the right, an 'Options' panel allows for configuration of text topics, including 'Text topics - OrderNote 1', 'Object', 'Style', and 'Layout' settings. The 'Text Topics' section includes 'General' options like 'Term parsing and role identification' (set to 'Automatic') and 'Analyze document sentiment' (unchecked). The 'Topic Discovery' section includes 'Maximum topics' (set to 6), 'Cell weight' (set to 'Logarithmic'), 'Term weight' (set to 'Entropy'), and 'Topic label length' (set to 4). The 'Model Display' section is also visible at the bottom.

Topic Name	Document Count
location, customer, ...	~4500
+drop, risk, supplier...	~4000
novelty, bead, gift,...	~3500
neutral, game, figur...	~3000
plush, neutral, toy,...	~2500
very, +satisfy, happ...	~2000

Topic Relevance	Document
	00004RPC in Boise placed an order. 16 unit(s) of Toy Figure(s) were...
	00002L9N in Birmingham placed an order. 8 unit(s) of Toy Figure(s)...
	This order is for 00004RPF located in Boise. They ordered 1 unit(s)...
	000052GE in Pittsburgh placed an order. 5 unit(s) of Toy Figure(s) w...

And if you do choose to analyze sentiments...





- What: A boutique firm specializing in movie consulting
- Specifically: Advise movie production companies on which new movie proposal could be a potential smash hit.
- Analytical task/goal: Based on analysing the characteristics of historical films find what type of movies that maximise Viewer Rating scores.
- Data: dataset of 1500 historical movies with viewer rating (score 1-5) and synopsis (unstructured text data) together with other data

Data - Moviedata

Variables

- Synopsis – overview text
- Title – unique
- MPAA Rating – Rating (R,PG,PG-13 NR,G,NC-17)
- Genre
- Year
- Viewer Rating – Target (0-4)
- Size
- ... (Dummy ++)



What Everyone Should Know About The Movie Rating System.

GENERAL AUDIENCES
G

Nothing that would offend parents for viewing by children.



G GENERAL AUDIENCES

PARENTAL GUIDANCE SUGGESTED
PG

Parents urged to give "parental guidance." May contain some material parents might not like for their young children.



PG PARENTAL GUIDANCE SUGGESTED

PARENTS STRONGLY CAUTIONED
PG-13

Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.



PG-13 PARENTS STRONGLY CAUTIONED

RESTRICTED
R

Contains some adult material. Parents are urged to learn more about the film before taking their young children with them.



R RESTRICTED

NO ONE 17 AND UNDER ADMITTED
NC-17

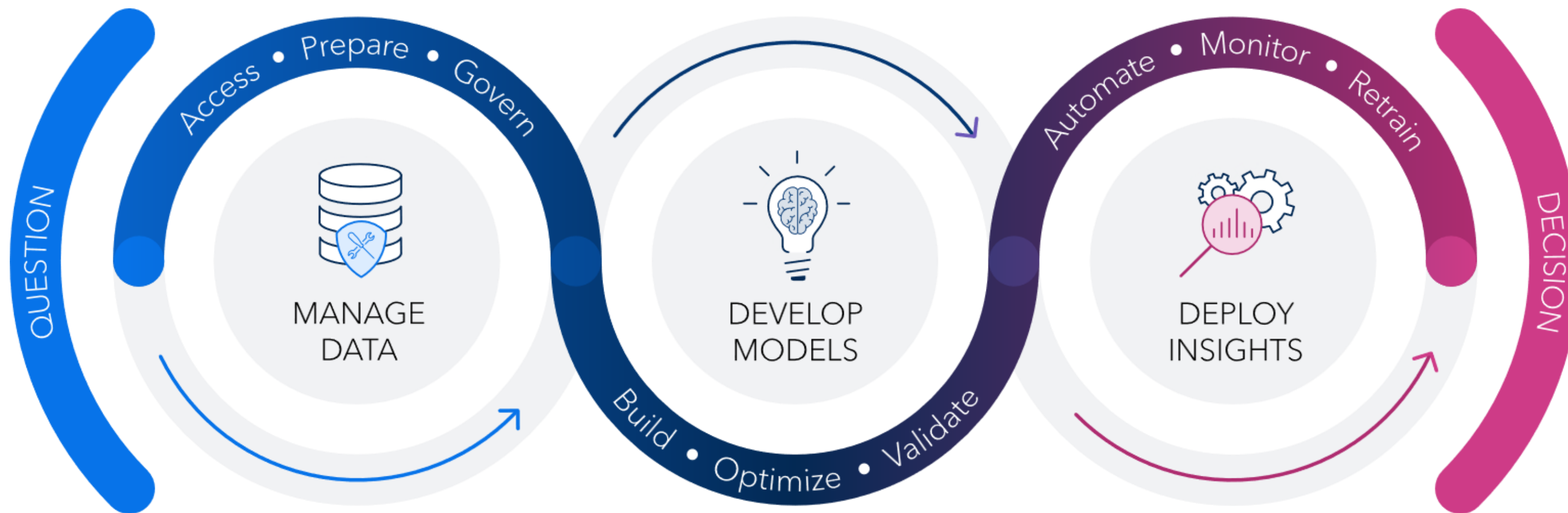
Patently adult. Children are not admitted.



NC-17 NO ONE 17 AND UNDER ADMITTED

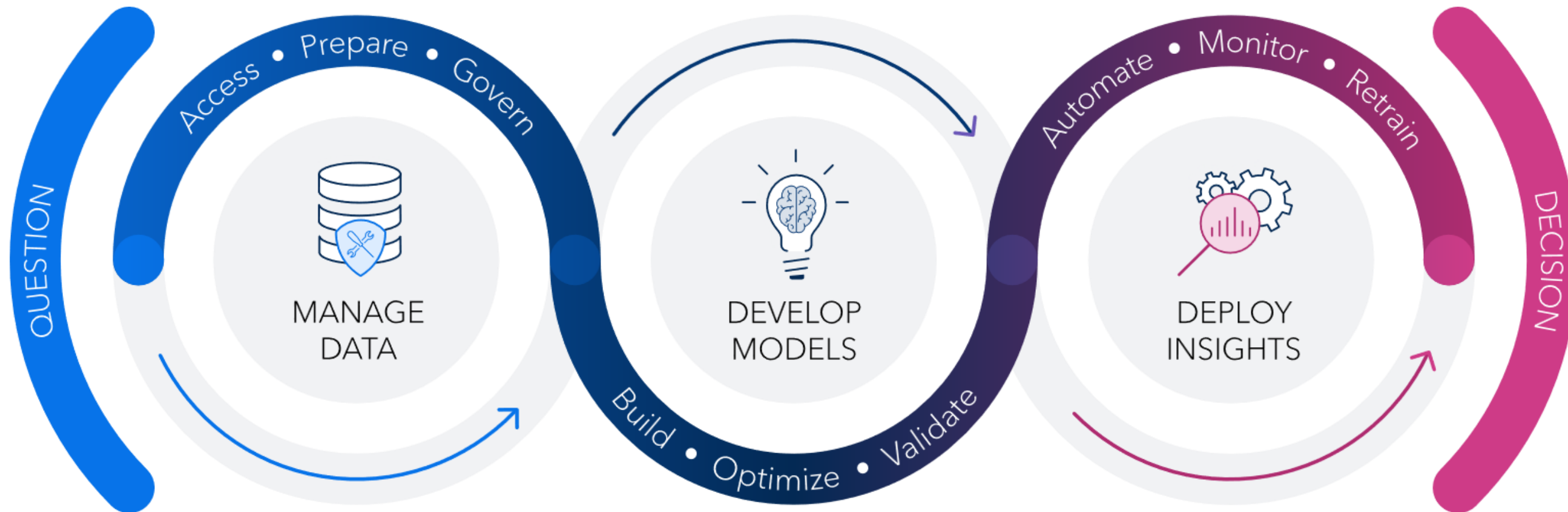
Operationalizing Analytics

The Analytics Lifecycle



Operationalizing Analytics

Which new movie proposal could be a potential smash hit by analysing the characteristics of historical films?



Decide which type of movie that maximizes the Viewer Rating

- Step 1: Data Access: MovieGenres - structured + unstructured data
- Step 1: Understand your data:
 - Data preparation and exploration
 - Define your target – Viewer Rating
 - Initially analytics and explorations
- Step 2: Add text preprocessing:
 - Create topics for modelling

- Step 1: Add a regression model to your (structured) data, measure models performance: ASE
- Step 2: Based on outcome of the text analysis extend your regression model in step 1 with “new” features
- Compare the two models and choose the best one (champion)

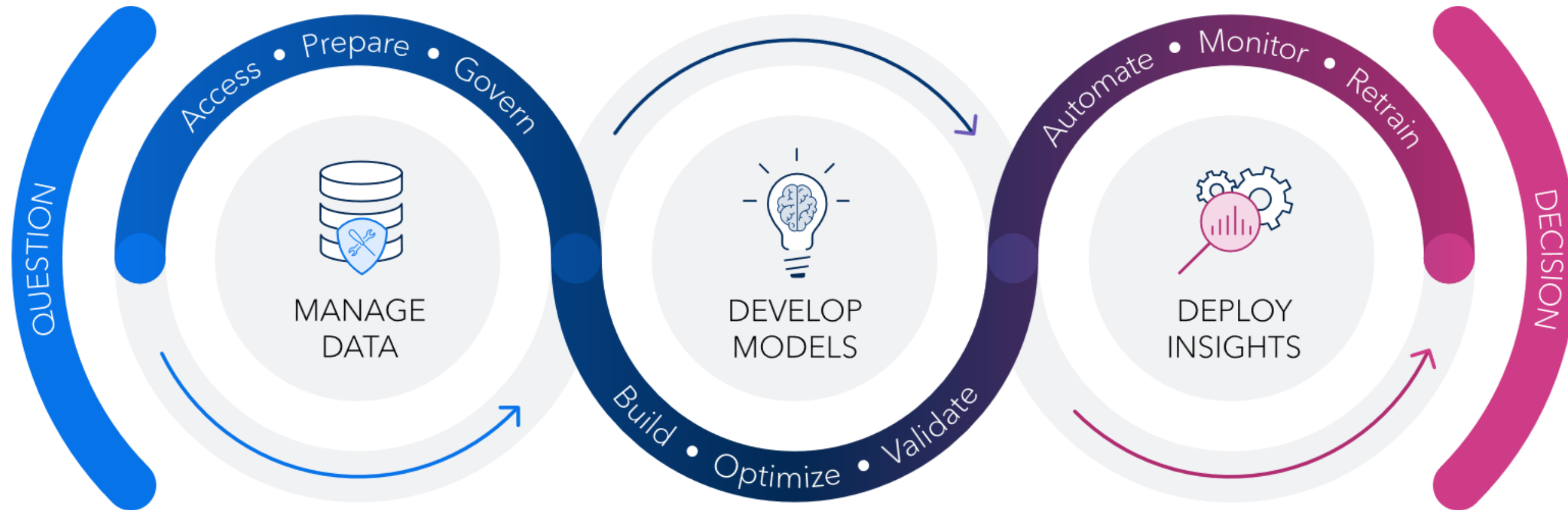
- Finally, you will be ready to;
- Bring your champion model to further analysis and deployment
 - Deploy the results in a Dashboard
 - Validate samples (type of movies) with experts - ASE
 - In long Run:
 - Score the model against new data
 - Monitor periodically (quarterly, yearly, when needed...)

Sentiment Example: Customer review sentiment classifier

How can texted review data improve a predictive model?

Review: *"I absolutely love this product! It's fantastic."*
Sentiment Label: Positive

Decide whether a given review is positive or negative based on the text content.



1. Data Collection – Custom reviews – text + sentiment (target)
2. Understand and know your data, i.e.:
 - Data preparation and exploration
 - Analyse and find (hidden) insight
3. Text preprocessing:
 - Tokenize, start/stop, stemming etc....
 - Feature Extraction

Model Building:

- Train a (ML) model using the feature vectors
 - The model learns to associate certain word patterns with positive or negative sentiments
 - Validate: Assess the model's performance
- Sentiment Prediction:
- Compare models and choose champion

- Deploy the model to the e-commerce platform
- Score new reviews – when customer submits a new review
 - The model predicts whether the review expresses a positive or negative sentiment
- Monitor – performance, retrain, retire, replace...
 - Using metrics like accuracy, precision, recall, or F1-score. Fine-tune hyperparameters if necessary.
 - Preprocess and convert into a feature vector.
 - Feed the vector into the (new) model.

Spørsmål?

#learnsas

#skillbuilder

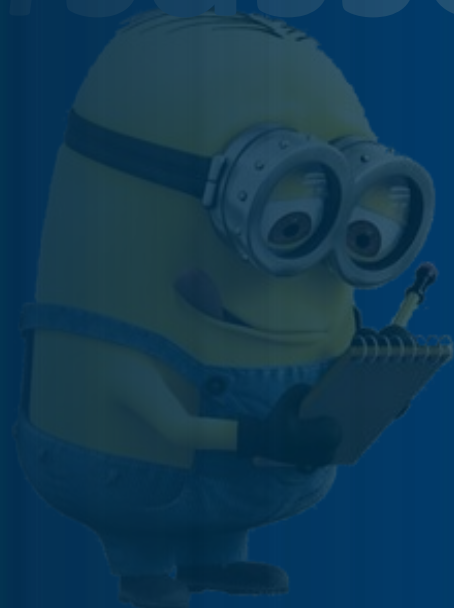
#lifelonglearner

#securethefuture



#sassoftware

#weareallacademics



Vegard Hansen

Academic Lead @ SAS

vegard.hansen@sas.com

www.linkedin.com/in/vegard-hansen/

