

# TODAY'S DATA SCIENTIST

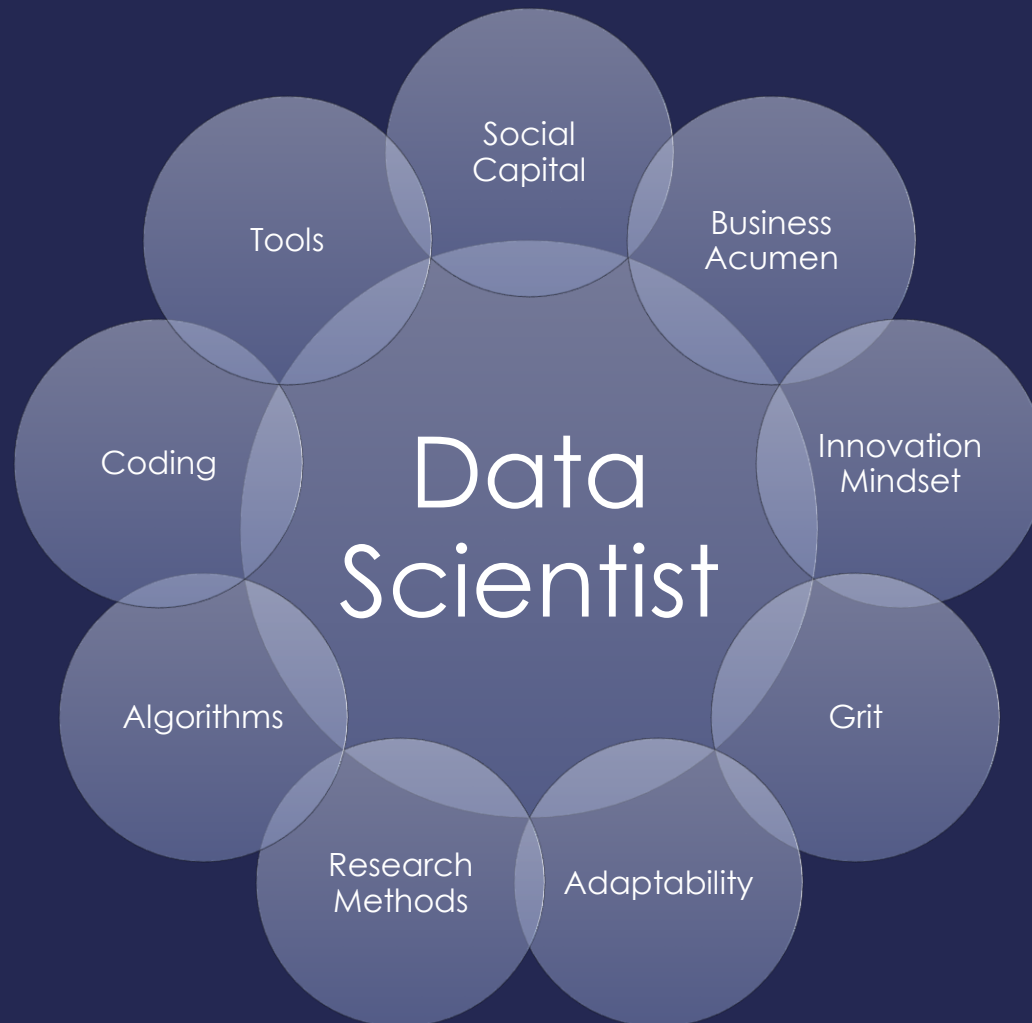
7/27/2023

CHRIS PARRISH

CHRIS@PARRISHHOUSE.NET

GITHUB: [HTTPS://GITHUB.COM/CHRISTOPHER-PARRISH](https://github.com/CHRISTOPHER-PARRISH)

# WHAT'S EXPECTED OF TODAY'S DATA SCIENTIST?



# WHAT'S EXPECTED OF TODAY'S DATA SCIENTIST?

## Experience & Personality

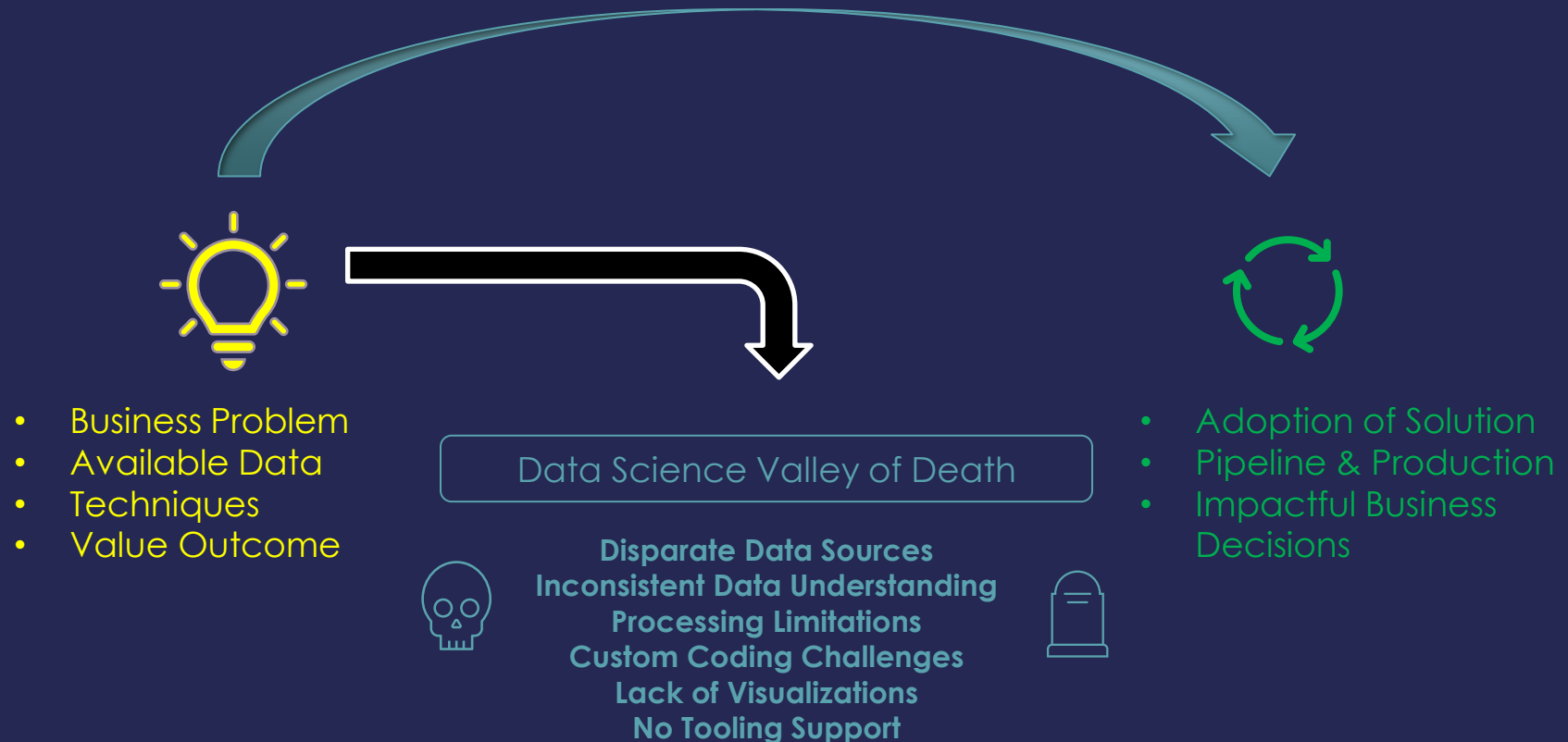
- COMMUNICATION SKILLS THAT CREATE TRUST AND BUILD RELATIONSHIPS
- DOMAIN KNOWLEDGE & PROVEN EXECUTION
- VISIBILITY INTO EXPECTED VALUE AND CREATIVE WAYS TO ACHIEVE GOALS
- FORTITUDE – GRINDING IT OUT TO FIND A SOLUTION
- PIVOT TO BETTER STRATEGIES WHEN NECESSARY AND SAYING “NO” TO KEEP FOCUS
- CONCEPTUALIZING RELATIONSHIPS, DATA GENERATION PROCESS, AND TESTING
- APPROPRIATENESS OF ALGORITHMS TO SOLVE BUSINESS PROBLEMS WITH ANALYTICS

## Skills

- CODE IN EMPLOYER-PROVISIONED LANGUAGES
- ANALYTICS PLATFORM/STACK KNOWLEDGE & DATA ENGINEERING

# WHAT'S EXPECTED OF TODAY'S DATA SCIENTIST?

Experience, Skills, Available Tooling



## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?

- TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
- COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
- FUTURE OF AI/ML AND WHAT TO DO TODAY

# TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS

## 2 EXAMPLES:

1. HOW CAN WE RETAIN CUSTOMERS (AND THEIR ASSETS)?
2. HOW CAN WE INCREASE SALES THROUGH OUR DISTRIBUTION NETWORK?

## EXAMPLE 1

Q: HOW CAN WE RETAIN CUSTOMERS (AND THEIR ASSETS)?

- MOST CUSTOMERS INCUR PENALTY CHARGES FOR EARLY TERMINATION
- HOWEVER, SOME CHOOSE TO LEAVE AND PAY TERMINATION CHARGES
- TERMINATIONS REDUCE ASSETS UNDER MANAGEMENT AND FEES
- COSTLY TO ACQUIRE NEW CUSTOMERS THROUGH DISTRIBUTORS

A: USING PRODUCT CHARACTERISTICS, FORECAST CUSTOMER ATTRITION OVER NEXT 2 YEARS BY PRODUCT AND SCORE BY MOST LIKELY TO ATTRITE

- ALLOWS IDENTIFICATION OF CUSTOMERS THAT MAY POTENTIALLY TERMINATE
- CUSTOMER SERVICE AND MARKETING EFFORTS CAN BE USED FOR RETENTION
- RETENTION PROVIDES FOR CONTINUED ASSETS UNDER MANAGEMENT AND FEES

# EXAMPLE 1



- TARGET
  - Active Customers
- FEATURES
  - Product Type
  - Product Add-ons
  - Product Age
  - Product Benefits

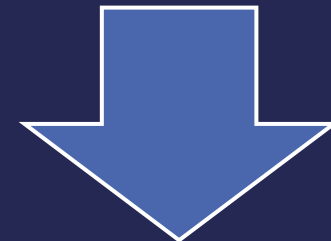


8-year observation window

- FEATURES

2-year performance window

- TARGET

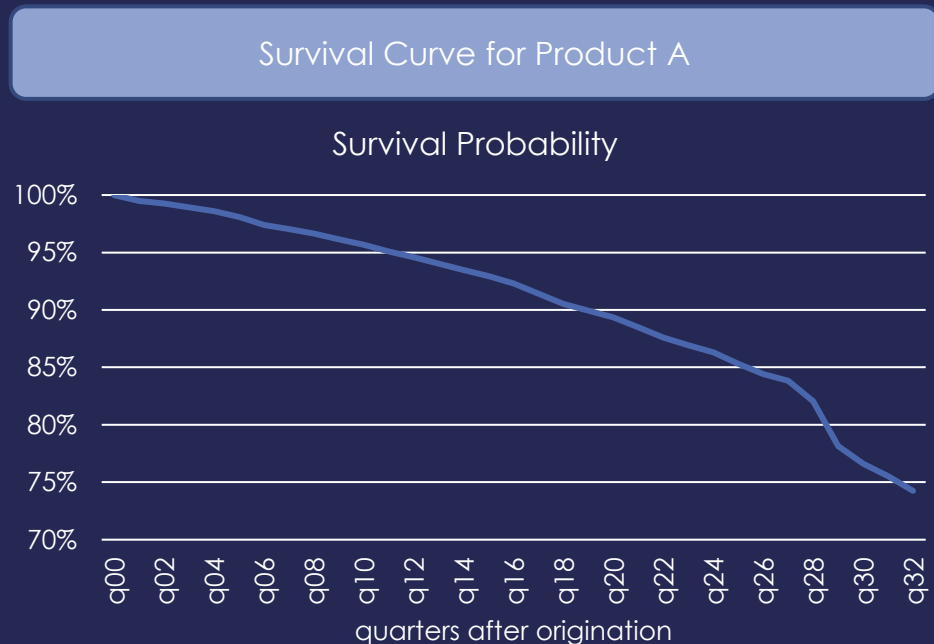




# EXAMPLE 1



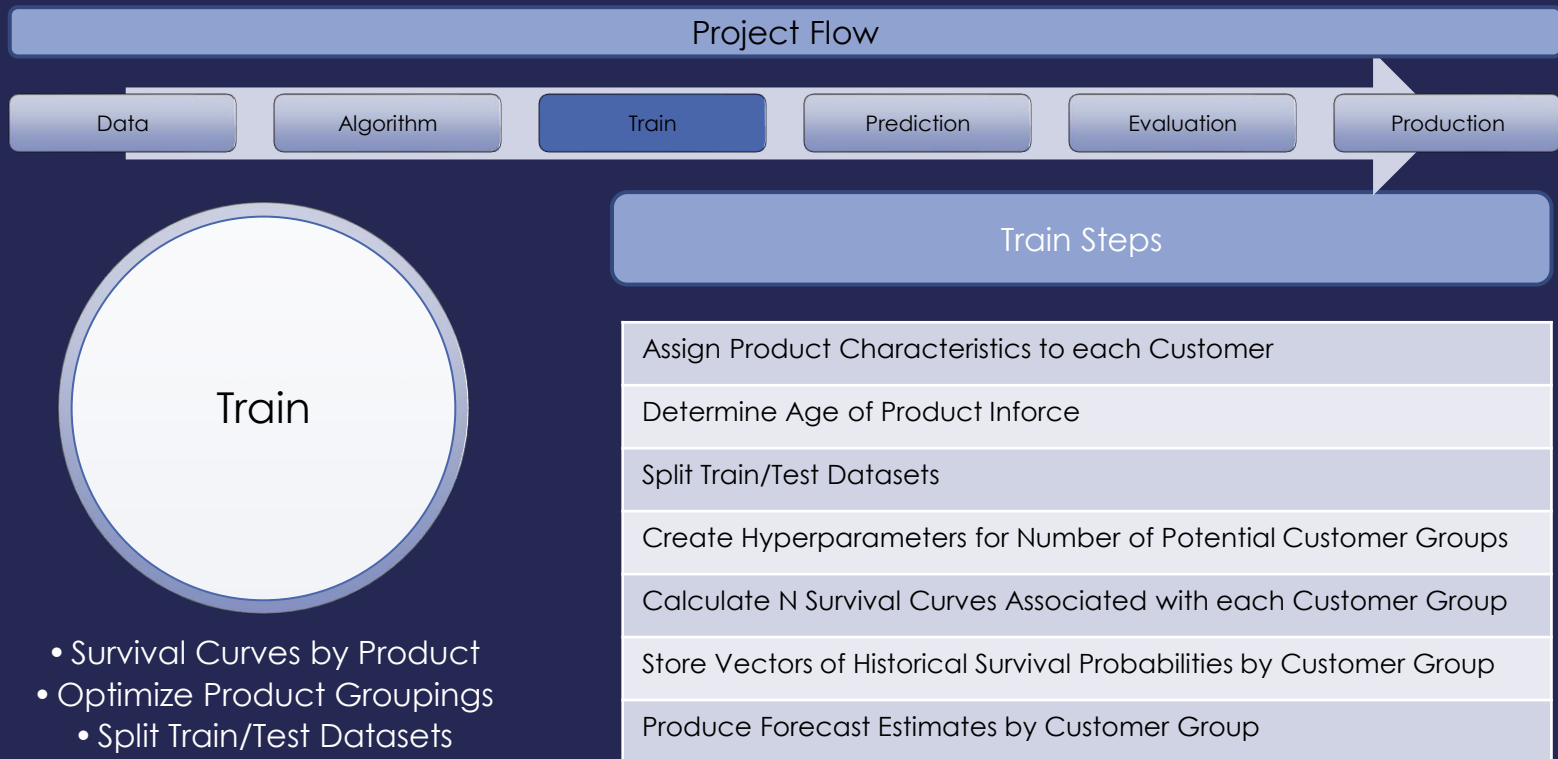
- Supervised ML
- Non-parametric survival



Survival starts off at 100% and decreases over time

For Product A, the bend in the survival curve after 7 years provides the key variation to differentiate survival forecasts by group

# EXAMPLE 1



Since there are N potential customer groups, the algorithm evaluates the hyperparameters and creates multiple sets of survival curves to evaluate for accuracy.

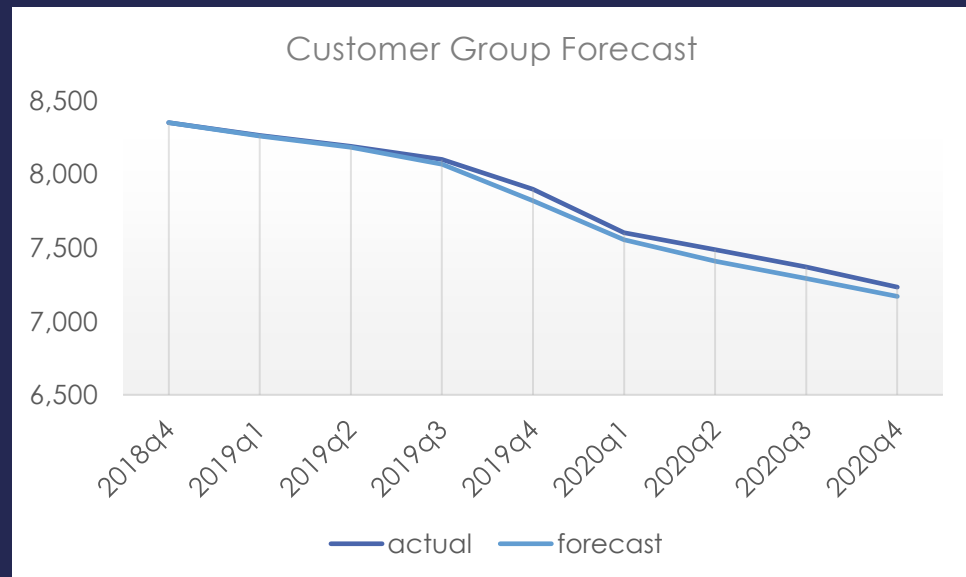
For example, Customer 123 might have Product A with Add-on 1 and Benefit !, and this customer may be in group Product A, Product A Add-on 1, Product A&B, Product A&B with Benefit !, etc. Survival curves are generated for each and evaluated against the test dataset to select the best grouping to maximize accuracy. Forecasts are then generated based on that optimized grouping.

# EXAMPLE 1



- Forecast Estimates

## Example Forecast



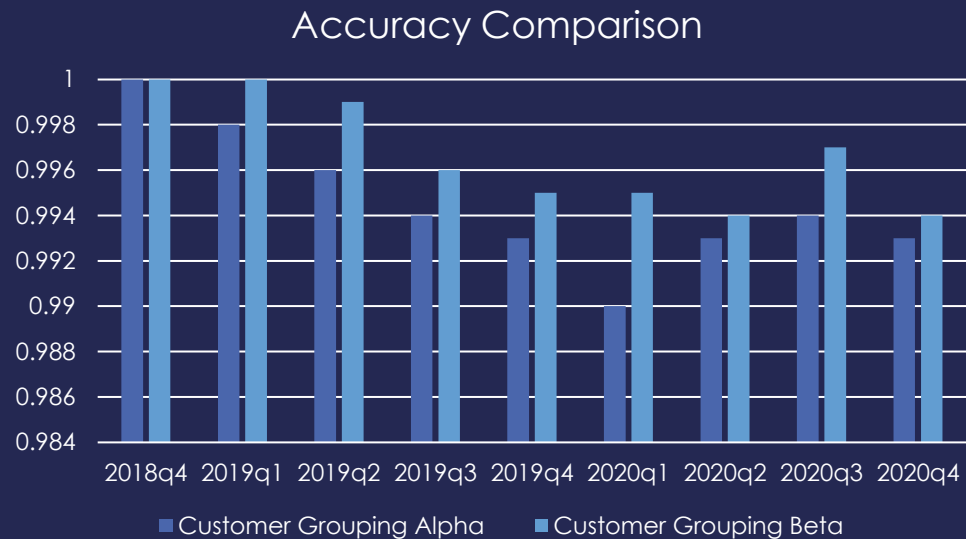
Using a holdout test dataset, a 2-year forecast is generated for each customer group. These forecast values are then compared to actual customer group count to evaluate accuracy.

# EXAMPLE 1



- Accuracy

## Example Actual to Estimated Comparison



Comparing the ratio of actual customer count to expected customer count, Customer Grouping Beta shows better performance (closer to 1) relative to Customer Grouping Alpha. Based on these results, the survival curves of Customer Grouping Beta will be used to generate forecasts.

# EXAMPLE 1



- Adoption
- Measurement
- Visualizations

## Pipeline Stability

Generate Forecasts for Optimized Customer Group

Produce Score List of Customers Most Likely to Attrite

Marketing Can Use Score List to Contact Customers

Rerun Every Quarter

Track Forecasts and Customer Scores from Prior Quarters

Report on Differences

Estimate AUM “Lift” from Assigned Customer Scores

## EXAMPLE 2

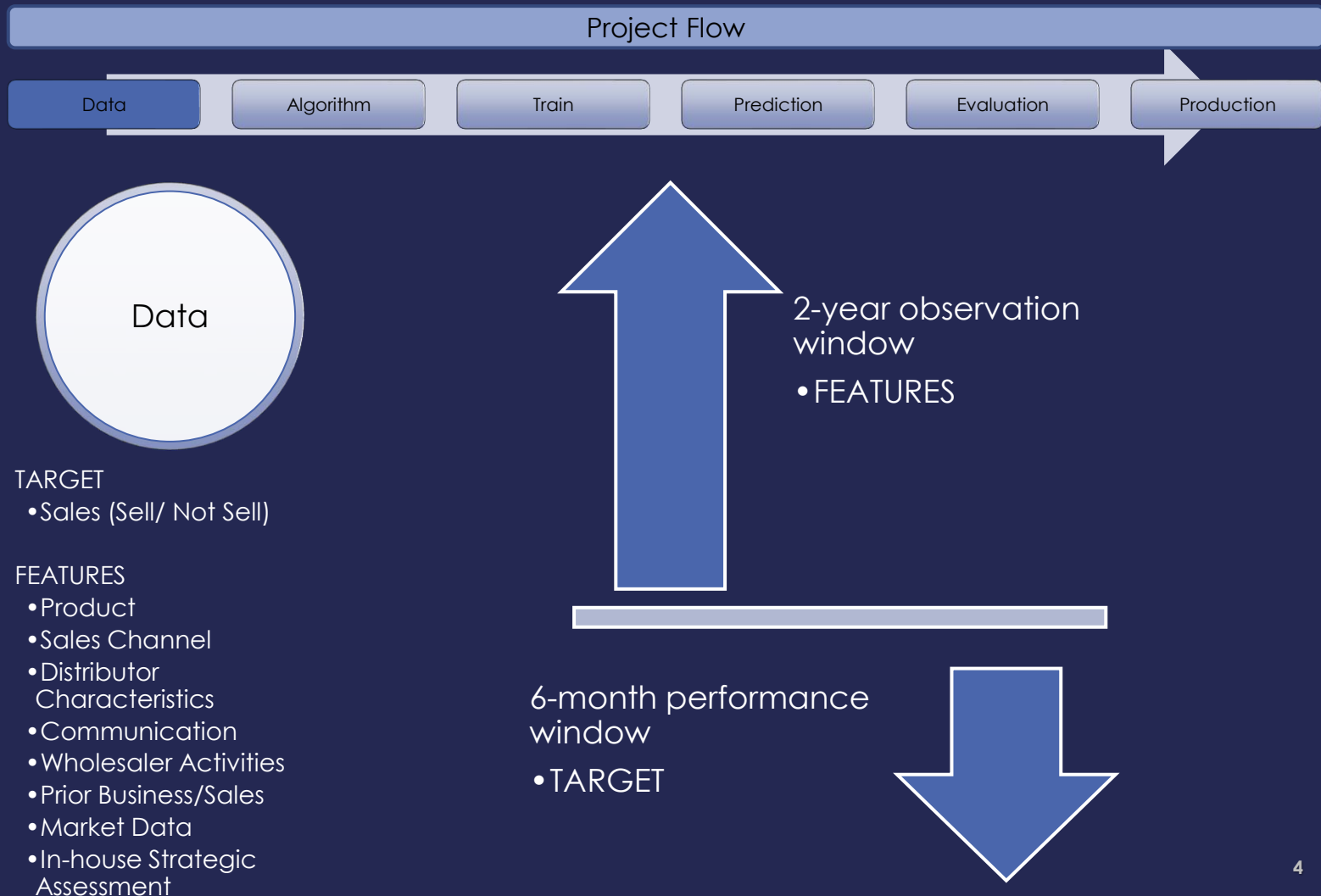
Q: HOW CAN WE INCREASE SALES THROUGH OUR DISTRIBUTORS?

- B2B MODEL THAT BENEFITS FROM IN-PERSON FACE-TO-FACE MEETINGS
- DISTRIBUTORS ARE FRAGMENTED – 200k+
- LIMITED TIME AND SALES STAFF TO MEET NATIONWIDE DISTRIBUTORS

A: RANK DISTRIBUTORS BY THE MOST LIKELY TO SELL PRODUCTS

- PROVIDE EACH WHOLESALER A LIST OF TOP 300 MOST LIKELY TO SELL
- ALLOWS SALES STAFF TO FOCUS ON HIGHEST TARGETS GIVEN TIME CONSTRAINTS
- AUTOMATION OF PROCESS BECOMES BAU FOR MEASURING SALES LIFT

## EXAMPLE 2



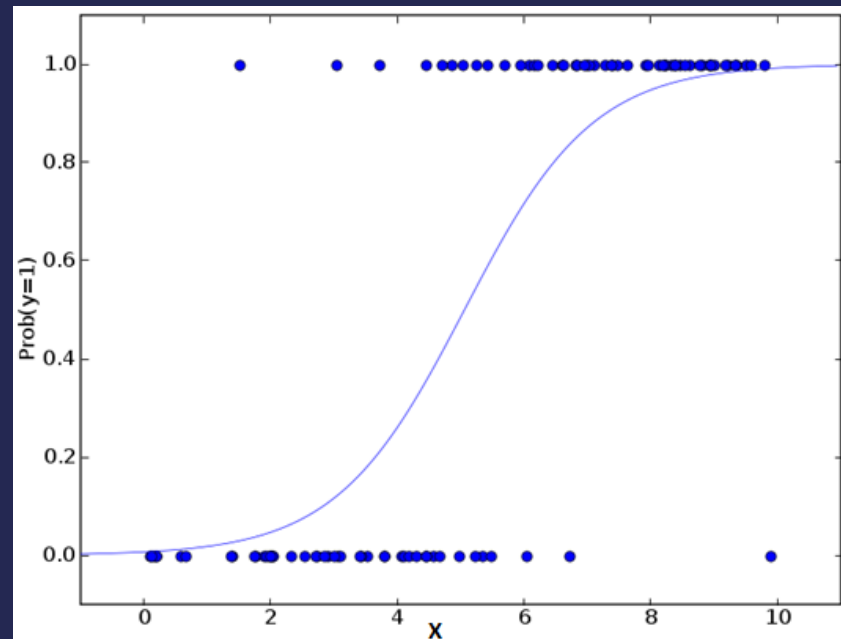
## EXAMPLE 2



Supervised ML

- GLM
- Logistic Regression

Example Logistic Regression Plot





## EXAMPLE 2



- Correlation Analysis
- Fill Missing Values
- Standardize Features
- Feature Selection
- Split Train/Test Datasets

### Example Features Selected

Feature
Average Number of Prior Sales Through Distributor
Number of Face-to-Face Meetings with Distributor in the Past Year
Number of Voicemails Received from Distributor in the Past Year
Number of Distributor Branch Offices
Distributor Total Liquid Assets
Sum of All Prior Sales Through Distributor
Market Opportunity \$ for Distributor in Region
Number of Inbound Calls from Distributor in the Past Year
Difference in Number of Sales Through Distributor Between Year 1 & 2
Distributor Tenure in Business

Features are typically selected using a variety of methods, including correlation analysis, business understanding, past or similar model features, random forest, or stepwise regression

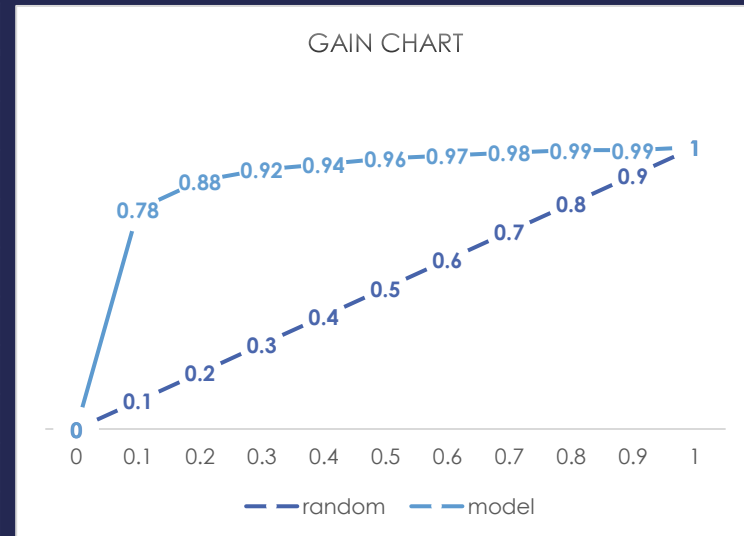
## EXAMPLE 2



Probability to Sell

### Example Probability Ranking

Dec	1s	Cumul %
10	870	78%
9	122	88%
8	35	92%
7	29	94%
6	16	96%
5	13	97%
4	17	98%
3	13	99%
2	3	99%
1	3	100%



- Deciles are sorted in descending order by scored probability
- The higher the cumulative % in the top deciles implies the model is assigned higher probabilities to actual "trues" (target)
- For example, in the (top) 10<sup>th</sup> decile, the model is assigning higher probabilities to 78% of the actual "trues" (870/1,121)
- The same analysis is completed for the test (holdout) dataset to identify potential overfitting

## EXAMPLE 2



Accuracy

Example  
Confusion  
Matrix

Actual	Prediction	
	0	1
0	17,318 (tn)	1,450 (fp)
1	209 (fn)	912 (tp)

Accuracy = (true positive + true negative) / all observations =  
 $(17,318 + 912) / 19,889 = 92\%$

• This shows how well the model predicted propensity to sell (or not to sell) overall

True Positive (Recall) = predicted true positive / all actual positive =  
 $912 / (209 + 912) = 81\%$

• This shows how well the model predicted propensity to sell relative to all distributors that did sell

Precision = predicted true positive / all predicted positive =  
 $912 / (1,450 + 912) = 39\%$

• This shows how well the model predicted propensity to sell relative to all distributors predicted to sell

True Negative (Specificity) = predicted true negative / all actual negative =  
 $17,318 / (17,318 + 1,450) = 92\%$

• This shows how well the model predicted propensity to not sell relative to all distributors that did not sell

## EXAMPLE 2



- Adoption
- Measurement
- Visualizations

### Pipeline Stability

Assign top 300 Distributor "Scores" to each Wholesaler

Deploy into Salesforce.com

Ideally, Wholesaler Uses Score List to Meet with Distributors

Rerun Every Quarter

Compare Scores from Prior Quarters

Report on Differences

Estimate Sales "Lift" from Assigned Distributor Scores

## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?



- TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
- COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
- FUTURE OF AI/ML AND WHAT TO DO TODAY

# DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS



## Languages

- Python/PySpark
- R
- SAS Viya OS APIs
- SQL

## IDEs

- Spyder
- R Studio
- Jupyter
- VS Code
- SAS Studio

## Platforms

- Hadoop/Hive
- Relational DB
- Azure ML (Microsoft Certified Azure Data Scientist)
- GitHub
- SAS Viya

## Visualizations

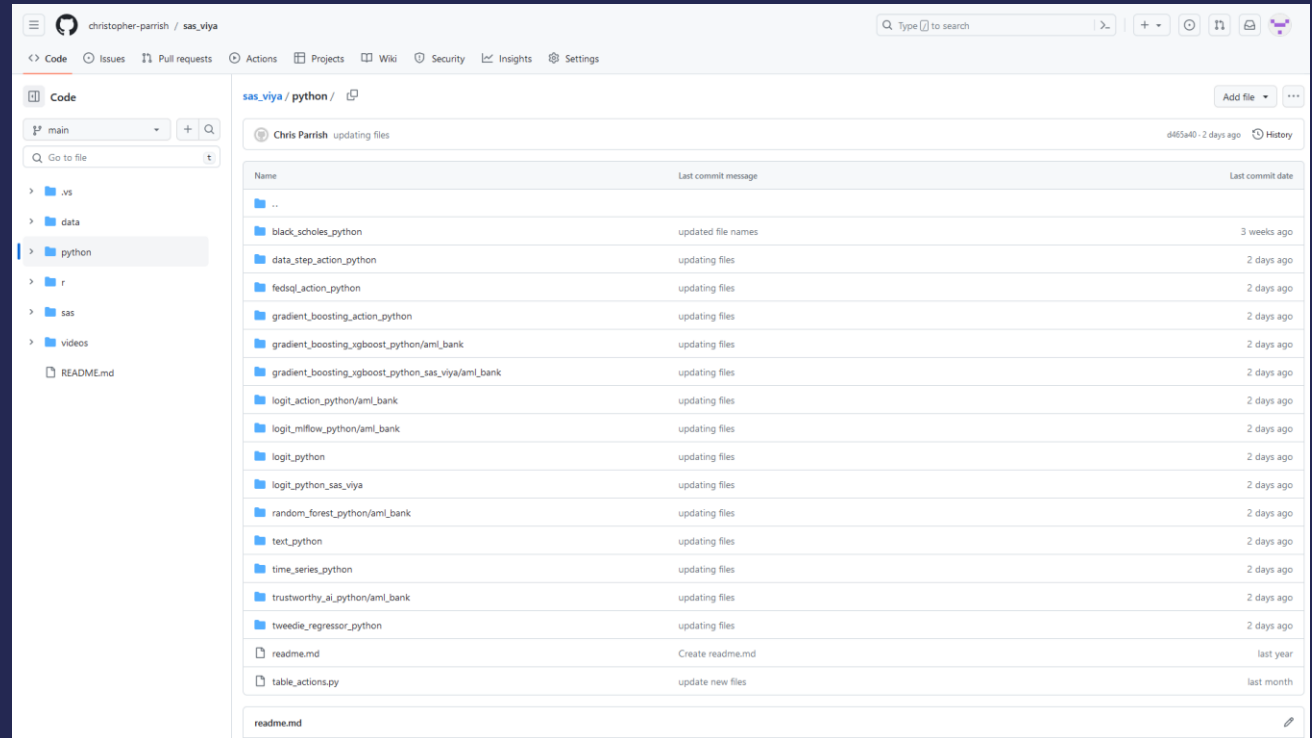
- Power BI
- SAS Visual Analytics

## SAS Viya

- SAS Viya Python/R SWAT & sasctl
- SAS Viya Model Studio
- SAS Viya Model Manager
- SAS Viya Intelligent Decisioning
- SAS Viya Model Risk Management
- SAS Viya Information Catalog
- SAS IML

# DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS

Technologies



Name	Last commit message	Last commit date
..		
black_scholes_python	updated file names	3 weeks ago
data_step_action_python	updating files	2 days ago
fedsql_action_python	updating files	2 days ago
gradient_boosting_action_python	updating files	2 days ago
gradient_boosting_xgboost_python/aml_bank	updating files	2 days ago
gradient_boosting_xgboost_python/sas_viya/aml_bank	updating files	2 days ago
logit_action_python/aml_bank	updating files	2 days ago
logit_mflow_python/aml_bank	updating files	2 days ago
logit_python	updating files	2 days ago
logit_python_sas_viya	updating files	2 days ago
random_forest_python/aml_bank	updating files	2 days ago
text_python	updating files	2 days ago
time_series_python	updating files	2 days ago
trustworthy_ai_python/aml_bank	updating files	2 days ago
tweedie_regressor_python	updating files	2 days ago
readme.md	Create readme.md	last year
table_actions.py	update new files	last month

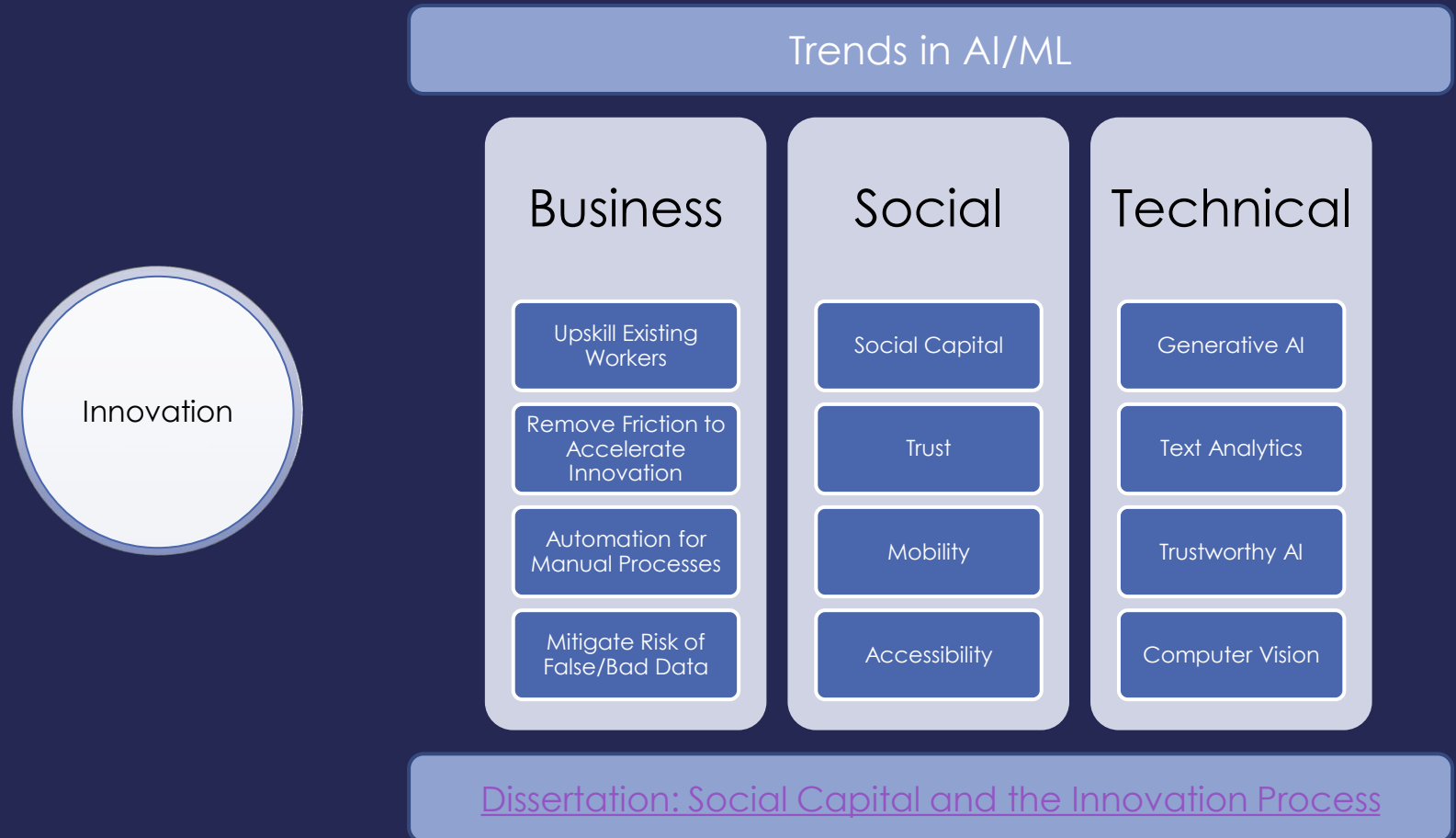
GitHub repositories contain data and code of projects

## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?

- ✓ • TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- ✓ • DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
- COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
- FUTURE OF AI/ML AND WHAT TO DO TODAY



# INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES



# INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES

Research

Firm Social Capital Index



Innovation

Using publicly available data:

- Filings and other public records (e.g., patents)
- Employment websites (Glassdoor, Indeed, LinkedIn)
- News publications

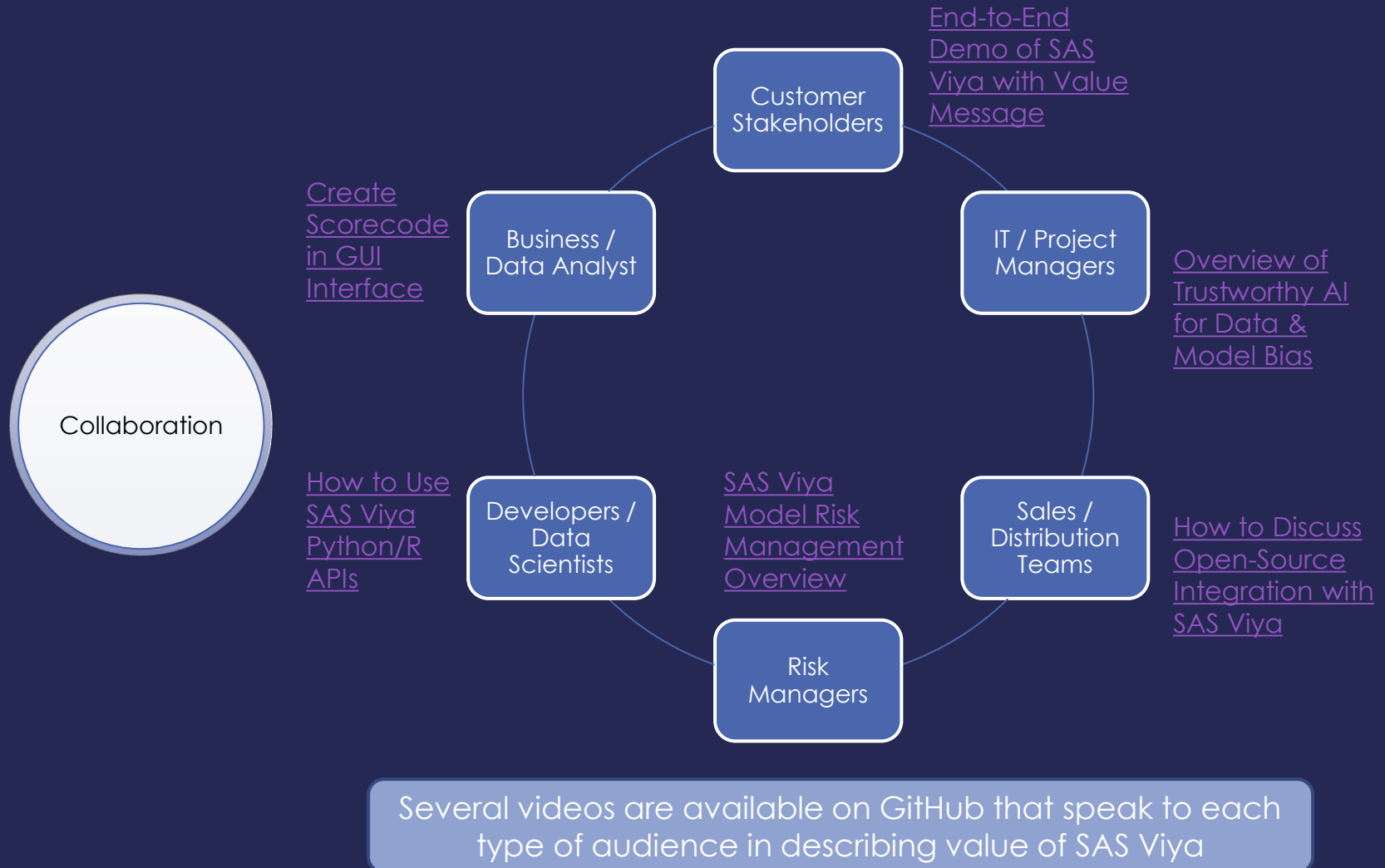
Generate composite using data extraction techniques (ML, NLP):

- Job reductions or labor-related cost-cutting
- Change in the number of employees
- Research/patent activities
- Mentions of “trust”, “collaboration” in filings and news releases
- Indications of types of work environments (WFH, flex-time)
- Sentiment of reviews on employment websites
- Firm on “best” lists, such as best place to work, best CEOs

## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?

- ✓ • TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- ✓ • DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- ✓ • INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
  - COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
  - FUTURE OF AI/ML AND WHAT TO DO TODAY

# COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES



## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?

- ✓ • TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- ✓ • DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- ✓ • INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
- ✓ • COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
- FUTURE OF AI/ML AND WHAT TO DO TODAY

# FUTURE OF AI/ML AND WHAT TO DO TODAY



## FOR SAS:

- Data will ALWAYS be the foundation for AI/ML, so data management tooling is a priority
- Investments in optimized cloud processing is a differentiator
- No/low code tooling combined with supported tech is a differentiator
- A true unified analytics platform that can simply workloads will be preferred
- Ability to interact/connect with multiple vendors across any AI/ML process is mandatory
- AI/ML with a social benefit focus can prove out the “data for good” message

## HOW DOES TODAY'S DATA SCIENTIST DEMONSTRATE THEIR BREADTH OF KNOWLEDGE AND EXPERTISE?

- ✓ • TANGIBLE EXAMPLES OF TRANSLATING IDEAS AND REQUESTS TO AI/ML PROJECTS
- ✓ • DEMONSTRATED UNDERSTANDING AND EXPERIENCE IN VARIOUS TECHNOLOGY STACKS
- ✓ • INNOVATION BY RESEARCHING AND EXPERIMENTING WITH NEW METHODOLOGIES
- ✓ • COLLABORATION AND COMMUNICATION WITH ALL TYPES OF AUDIENCES
- ✓ • FUTURE OF AI/ML AND WHAT TO DO TODAY

Q&A