Presented at the ICEAA 2023 Professional Development & Training Workshop - www.iceaaonline.com/sat2023

CLOUD-BASED MACHINE LEARNING IN THE DOD ENVIRONMENT

Summit2Sea Consulting

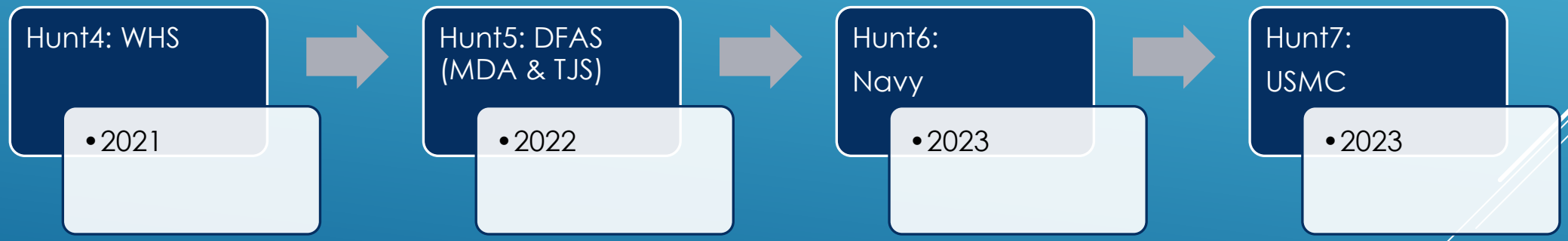
Conner Lawston

REAL DOD CLOUD EXAMPLE: ADVANA

- The Advana (Advanced Analytics) platform is a Cloud system that hosts massive amounts of data for dozens of DoD agencies
- Formed via the consolidation of 1200+ different databases, it holds information including 1 billion+ financial transactions (2018-present)
- 20,000+ Users, with a storage capacity of 500 TB, all on a secure CaC-required environment
- Compatible with data analysis tools (Python/Tableau/etc)- hosts 250+ dashboards
- Built by Booz Allen Hamilton and runs on AWS GovCloud

CASE STUDY: HUMANLESS UNMATCHED TRANSACTIONS (HUNT)

- ML & RPA project to correct 'unmatched' transactions in DoD financial system
- Multiple Stakeholders including OUSD, Missile Defense Agency (MDA), Dept of Navy, Washington Headquarter Service (WHS)



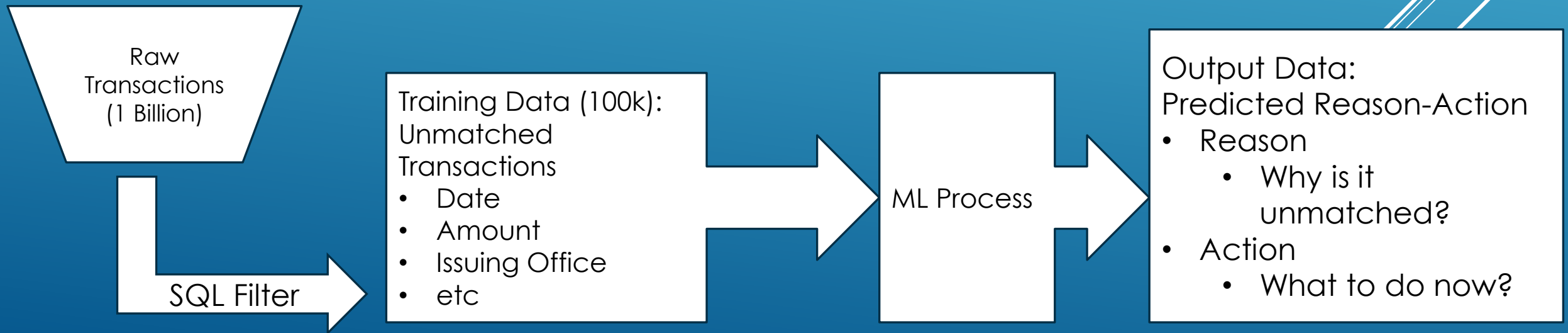
BACKGROUND: UNMATCHED TRANSACTIONS

- DoD Financial Transactions that for some reason did not go though
- May have a typo or missing value in fields
- May have wrong dollar amount
- Reason & Action
  - Reason- Why did it not work?
  - Action- How do we fix this?
  - There are ~20 unique Reason-Action combinations

Reason - Action
Credit to Invoice - Credit to Invoice
Credit to POET - Credit to POET
Incorrect PO Number - Match with correct PO
Direct Invoice - Direct Invoice
Suspense LOA - Suspense LOA
Incorrect or undefined PO Line - Match with correct PO Line
Split Payment to Inv Lines - Split Payment to Inv Lines
Incorrect PO - Use Doc Ref No
Incorrect PO Number - Match with correct PO and LOA from Contract
Incorrect PO docref - Match with correct PO
Incorrect PO docref CLIN REQ ACRN - Match with correct PO
Incorrect PO Number - Match with PSB PO
Insufficient RCV Amt - Add receipt to contract
Insufficient PO Amt - Add funding to contract

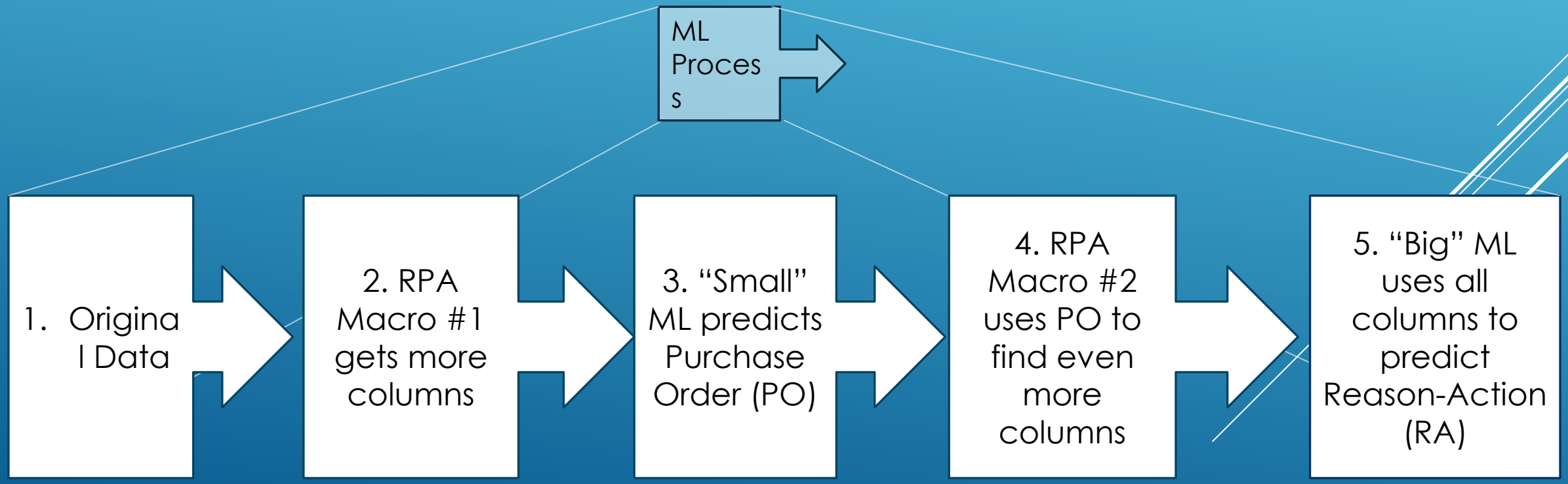
FRAMING THE ML PROBLEM

- SQL to Filter down to only relevant data
- Training Dataset now consists of relevant transactions (100,000 data points)
- Classification Problem with numeric and categorical inputs (90 columns)
- Determine which Reason-Action to give (20 unique classes)



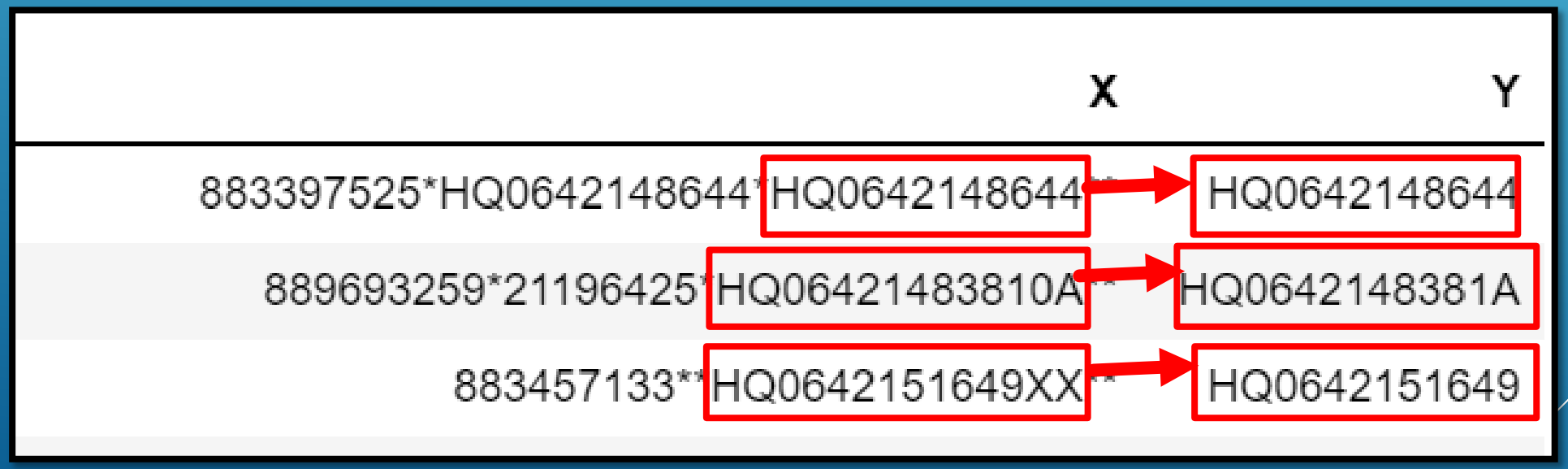
DETAILED VIEW

- Multiple ML Models and RPA Macros used together sequentially
- "Snowball" effect where each step gets more data for final "Big" ML



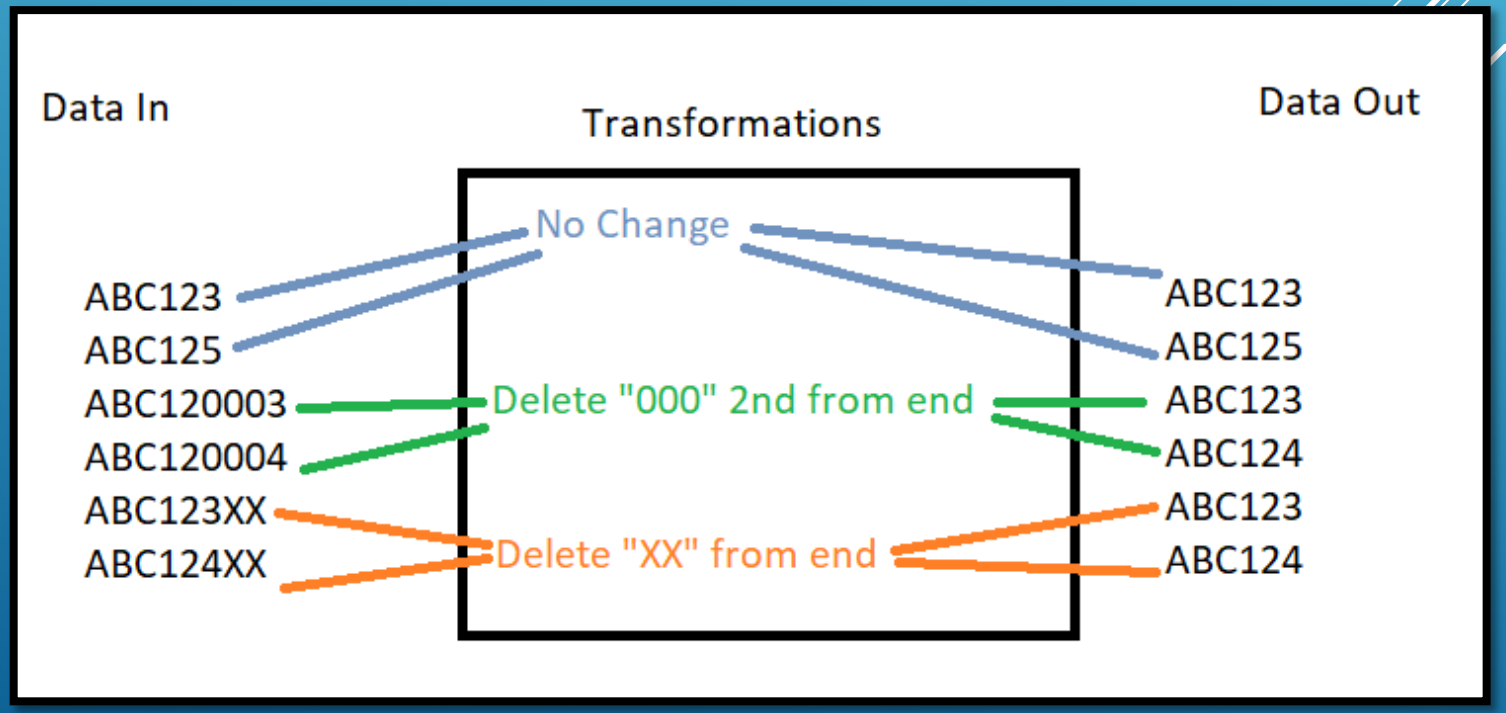
DEEP DIVE: PO ML MODEL

- Objective is to 'Guess' column Y, given column X
- Column Y is usually **similar** to a section of column X
- Each row is unique



PO ML MODEL: PO RULES

- Instead of predicting final value, predict a 'Transformation'
- Once you know each row's appropriate Transformation, apply them to get Y
- Ideally small number of transformations (5?) will fix majority of columns (80%?) [Pareto]
- Group transformations by how different they make start & end columns



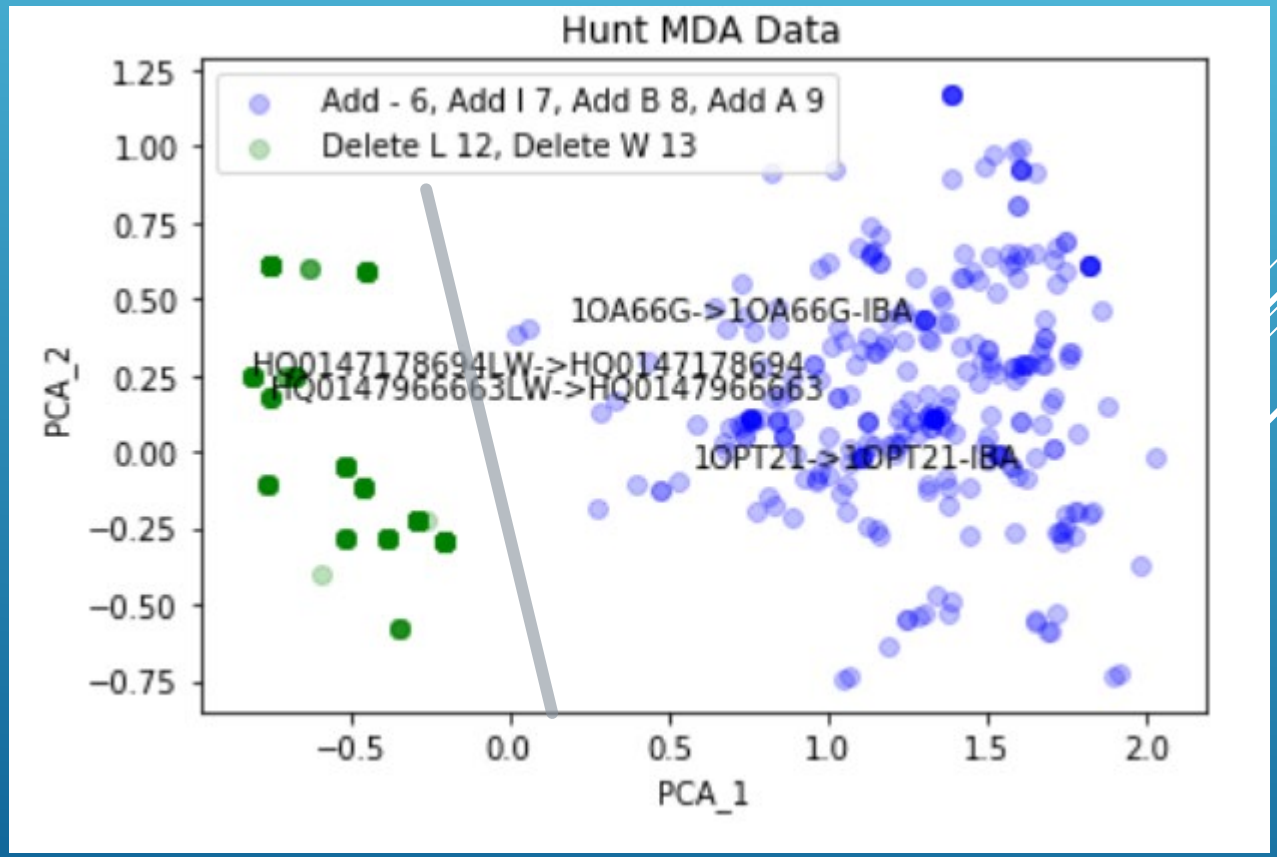
PO ML: PCA

- Prepare data for ML by feature engineering- 'Bag of Letters'
  - How many of each letter does each Input (PO) contain?
  - Apply PCA to letters matrix

Input	Output	A	B	C	D	PCA_1	PCA_2	PO_Rule
ABBA	ABBA-IBA	2	2	0	0	1.871837	0.252884	Add - 4, Add I 5, Add B 6, Add A 7
ACDC	ACDC-IBA	1	0	2	0	-0.176509	-0.882514	Add - 4, Add I 5, Add B 6, Add A 7
CCCD	CV-CCCD	0	0	3	1	-1.360189	-0.385031	Add C 0, Add V 1, Add - 2
DDCC	CV-DDCC	0	0	2	2	-1.350730	0.678367	Add C 0, Add V 1, Add - 2
CCCC	CV-CCCC	0	0	4	0	-1.369648	-1.448429	Add C 0, Add V 1, Add - 2
AAAA	AAAA-IBA	4	0	0	0	1.837511	-0.672732	Add - 4, Add I 5, Add B 6, Add A 7
ABBB	ABBB-IBA	1	3	0	0	1.889000	0.715692	Add - 4, Add I 5, Add B 6, Add A 7
DDDC	CV-DDDC	0	0	1	3	-1.341271	1.741765	Add C 0, Add V 1, Add - 2

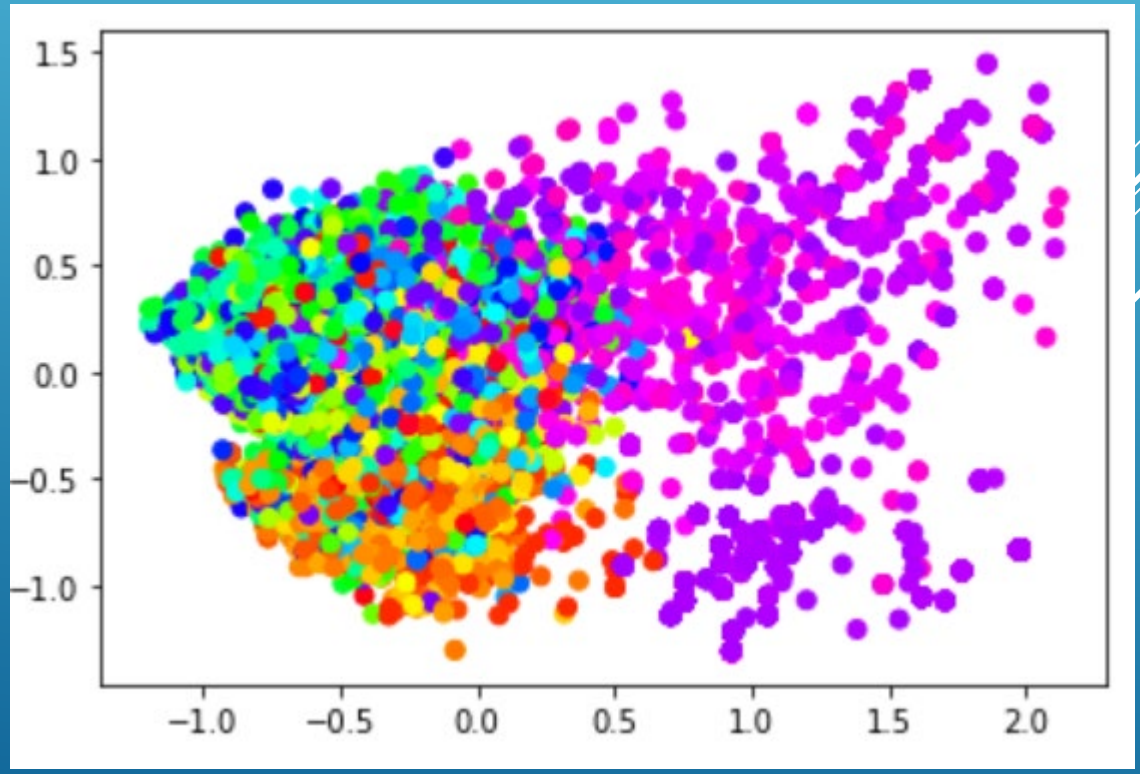
PO ML: VISUALIZE SIMPLE DATA

- Simple Comparison of two rules- easily separable with ML
- ONLY Using PO letters!



PO ML: FULL VISUAL

- More complex example with more rules
- Each color represents a different rule
- Some clusters exist within colors
  - Purple rule seem to be close together
  - Close = similar PCA values
    - = Similar makeup of letters

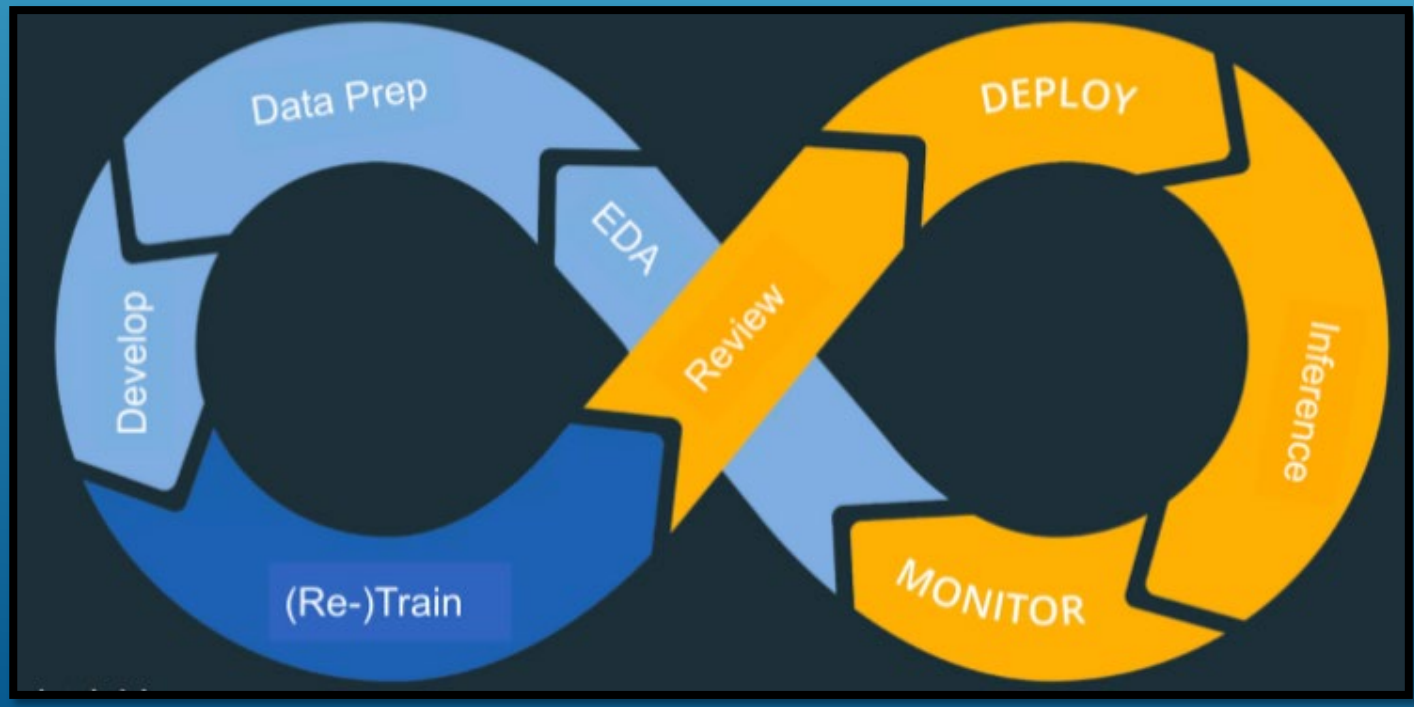


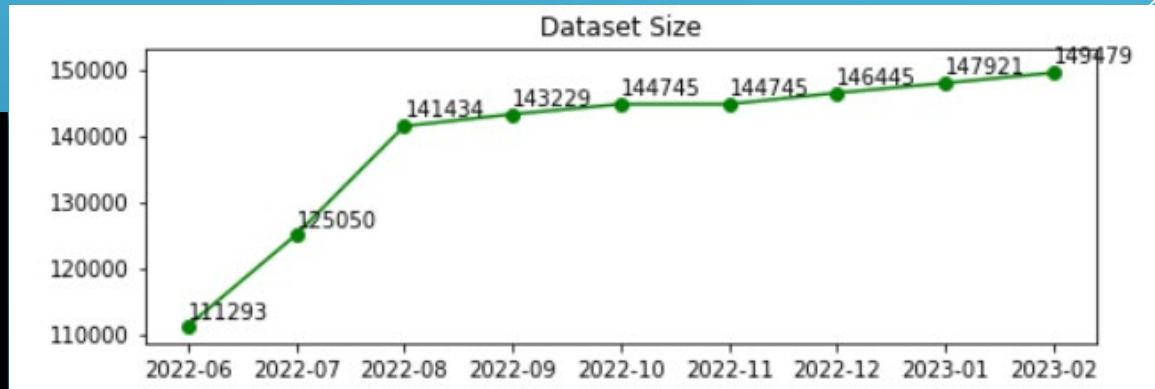
HUNT DEMO

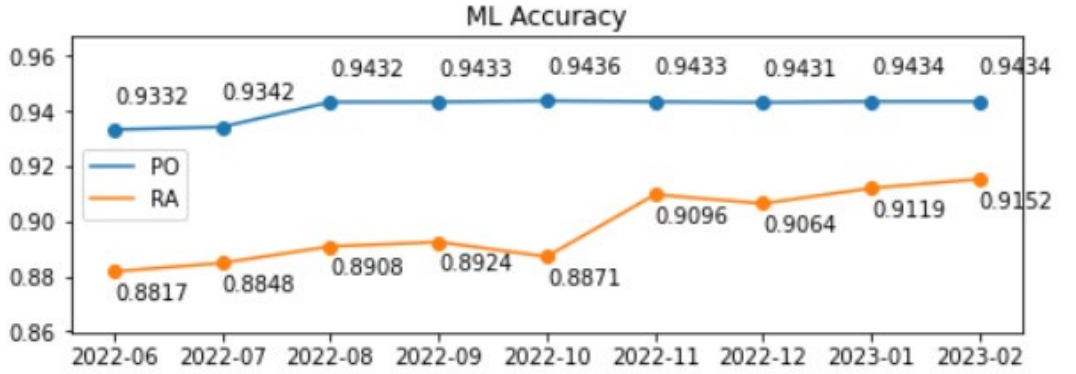
- https://vimeo.com/540679399

MLOPS

- Deploy system into production with minimal intervention needed
- Automatically re-train ML every month when database updates
- Automatically infer on new data every day
- Monitor accuracy metrics and add control logic

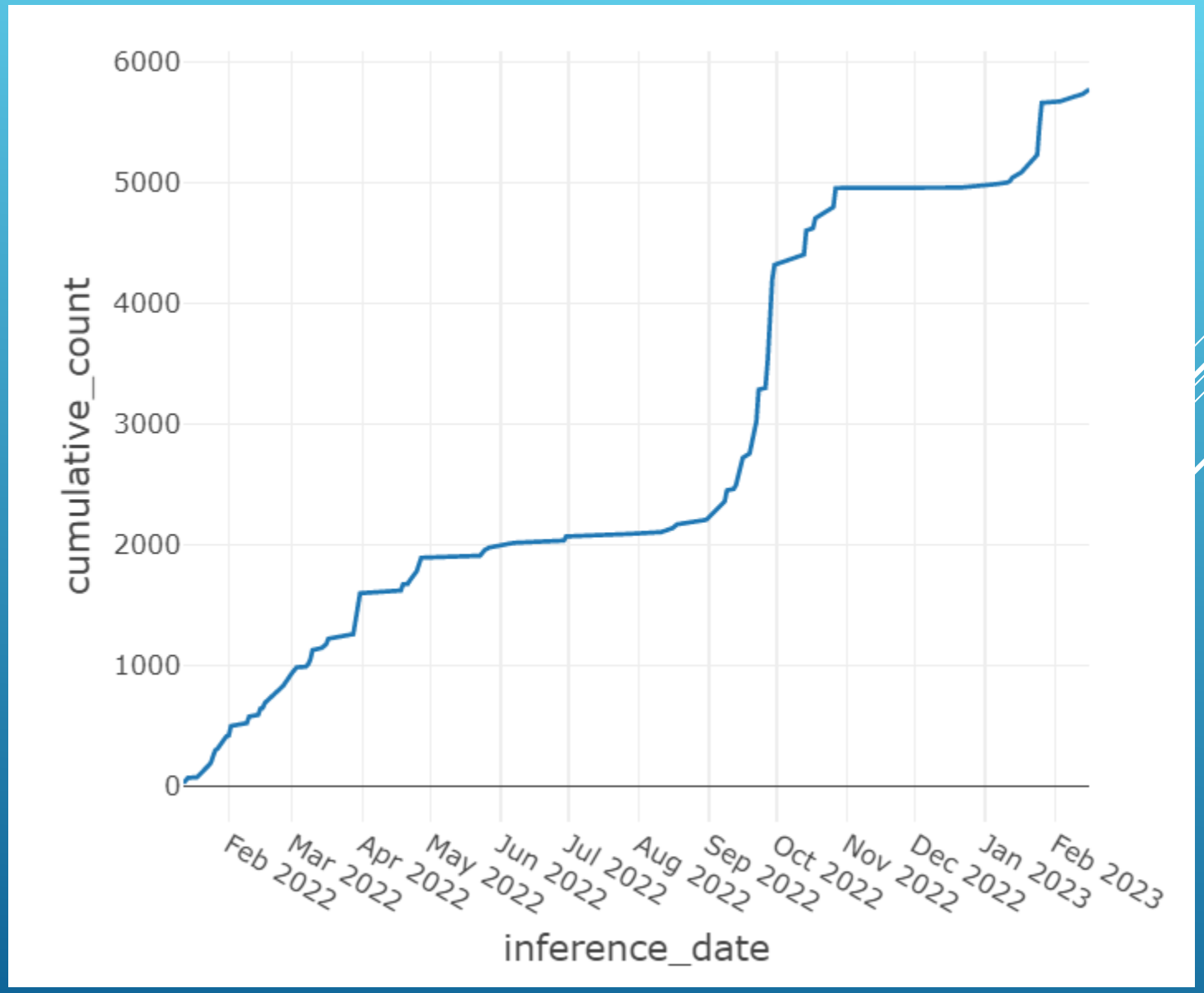






HUNT RESULTS

- Nearly 6,000 ML Predictions made to date!
- Surpassed accuracy goals of 70%
- ML Techniques tested include Logistic Regression, Decision Trees, XGBoost, Neural Networks
- First in DoD to connect Databricks and UiPath
- First in DoD to use 'AutoML'
- WHS asked to run HUNT 24/7 during busy season



HUNT IMPACT

- Our automated system cleared tens of thousands of transactions, saving time and money for the Government

Data Across FY 19/20/21Q1			
Agency	**Total UMT**	**Can Automate**	**Percent**
WHS	50,345	21,950	43.60%
MDA	18,810	8,350	44.39%
DSCA	57,325	8,860	15.46%
DISA	56,975	7,885	13.84%
DHRA	16,075	1,015	6.31%
DoDEA	44,855	29,090	64.85%
TOTAL 6	**244,385**	**77,150**	**31.57%**