IBM Analytics

MDM and ML

Manfred Oevers
Manfred.oevers@de.ibm.com
Manager, MDM Development

Martin Oberhofer
martino@de.ibm.com
Executive Architect, IBM Analytics Development

Legal Disclaimer

- The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

- References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

Agenda

1. Why Machine Learning for MDM
2. Use Cases
3. Implementation
4. Demo

Why Machine Learning for MDM

Why Machine Learning for Master Data Management

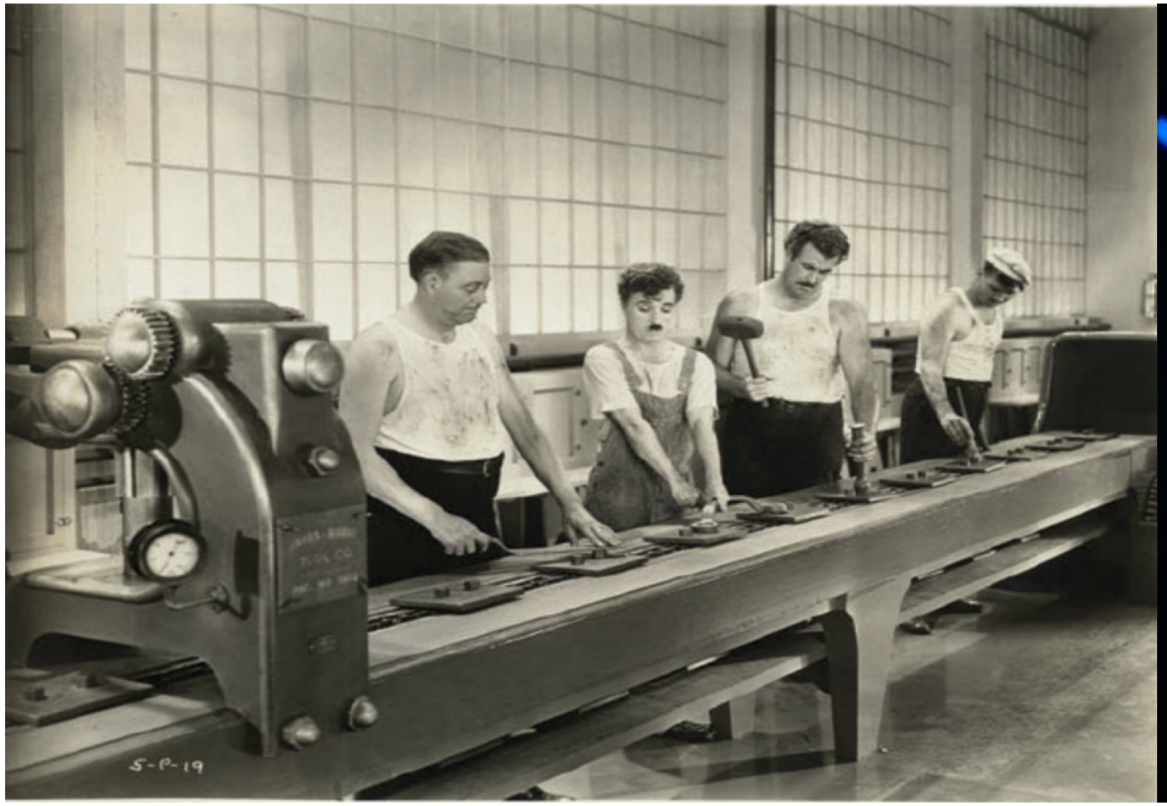
Labor Cost Reduction: Can we automate repetitive clerical tasks by putting ML into MDM?

Deeper Insights: Can we better discover hidden relationships?

Why Machine Learning for Master Data Management – Part 2

Intelligent Matching for Product Master Data: Can we combine techniques from NLP, clustering and ML to build best in breed product matching?

Smart Data Loading: Can we auto-discover, auto-classify and auto-map data on ingest to MDM making data loading seamless?

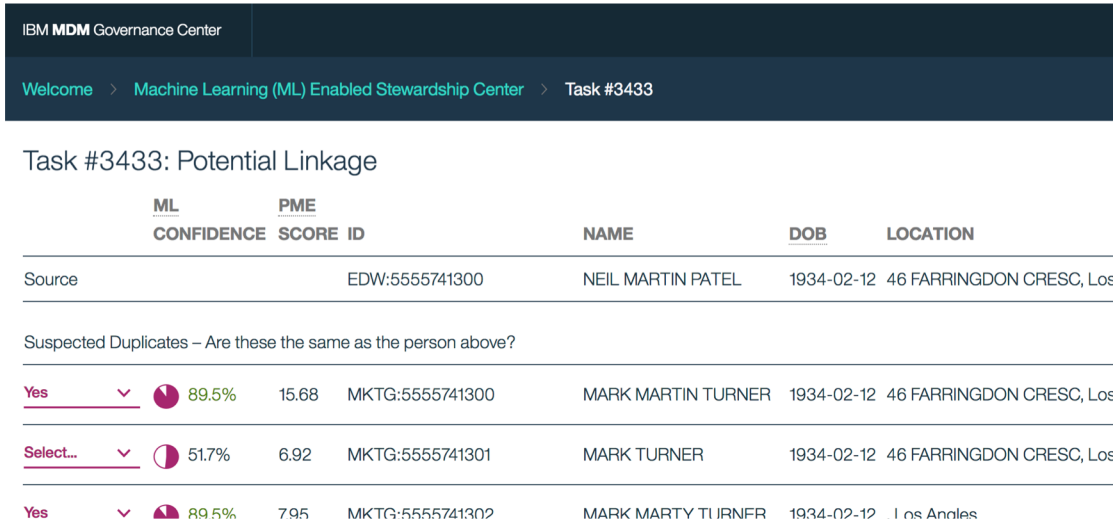


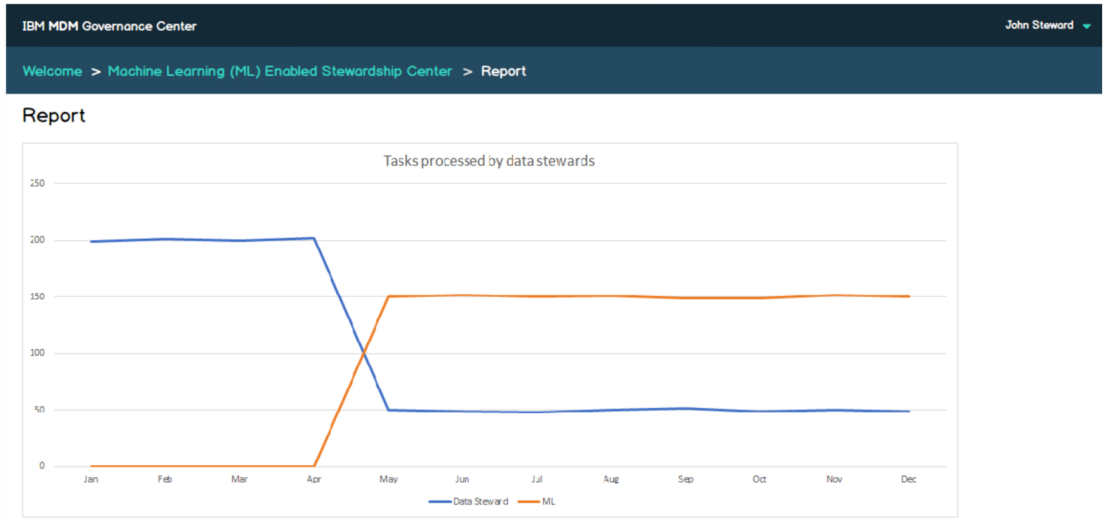
Use Cases

ML for MDM SE / AE

Goal is to improve user efficiency by automating stewardship using Machine Learning.

1. Learn from task resolution history to auto-classify future potential duplicates
   - Prototype
2. Active learning of similar tasks
   - "Give me more like this”

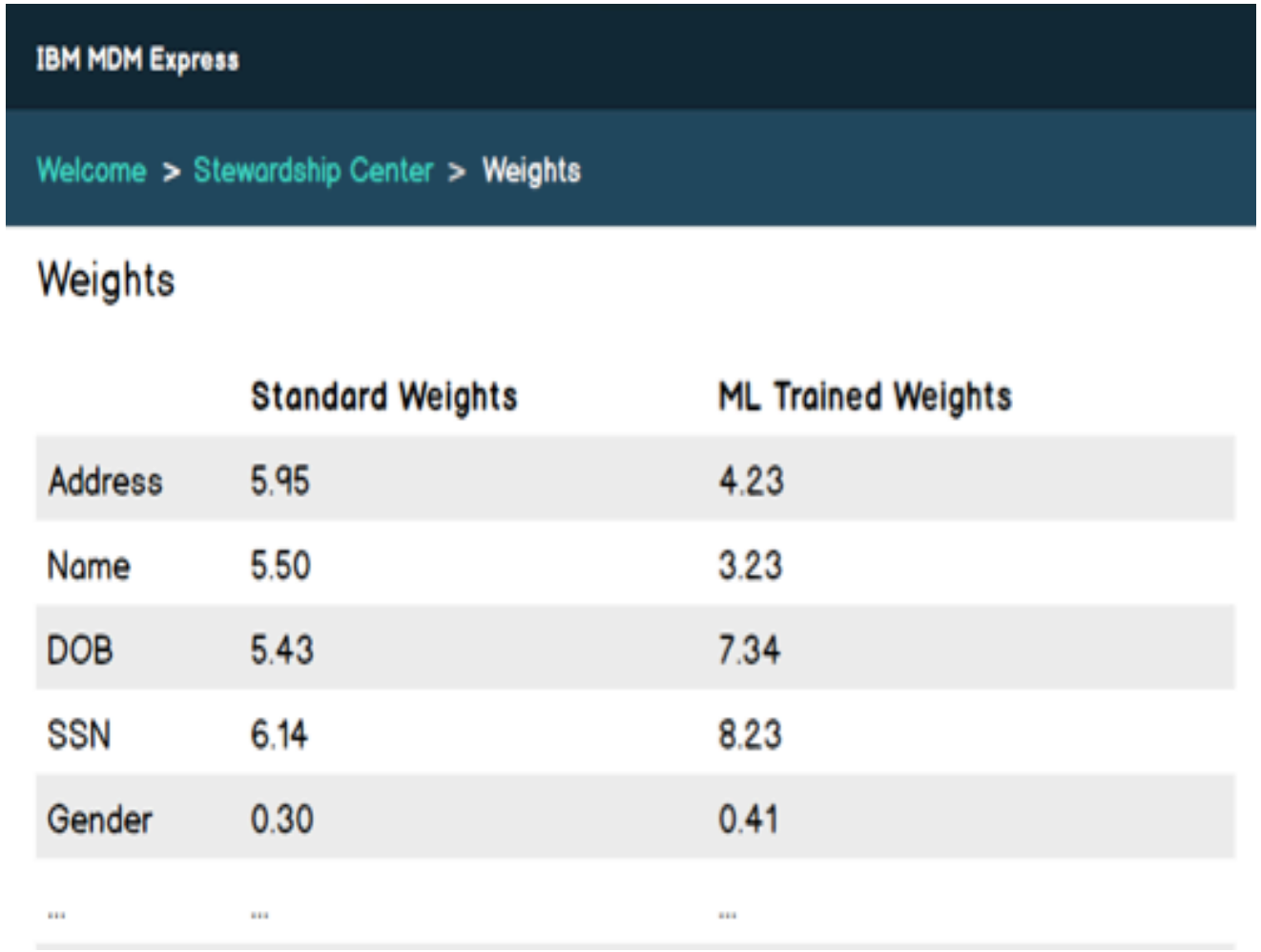




ML for MDM Express

- MDM Express should not require deep skills.
- Matching algorithm configuration out of the box supposed to work globally.
- Might not be perfect for all data sets.
- Should only require lightweight stewardship

➔ Self-tuning of matching weights and thresholds required based on ML



	Standard Weights	ML Trained Weights
Address	5.95	4.23
Name	5.50	3.23
DOB	5.43	7.34
SSN	6.14	8.23
Gender	0.30	0.41
...	...	...

Implementation

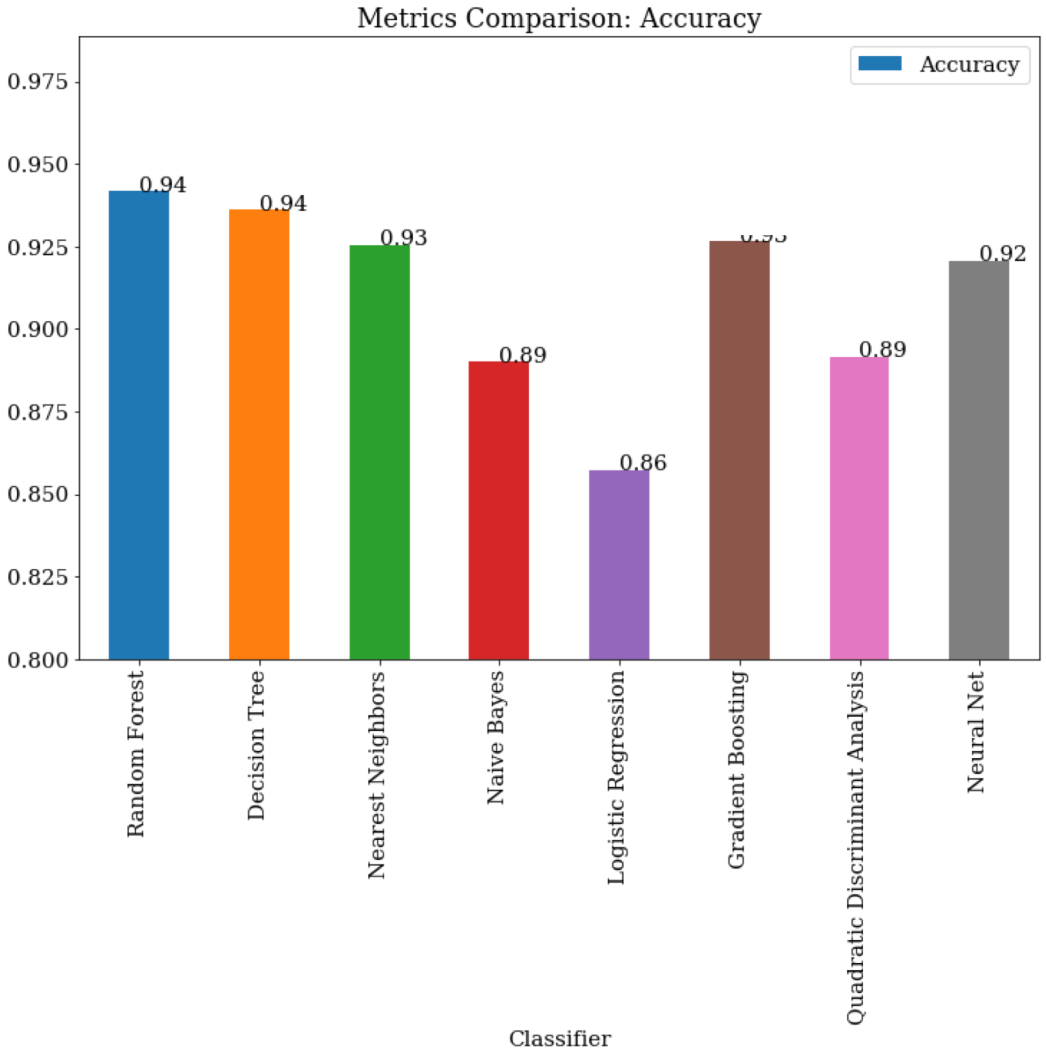
Input for Machine Learning

- Steward decisions retrieved from **mpi_entrule_{type}** table
- Enriched with detailed scores using **mpimcomp** utility
- Resolution history from MDM SE clients with >1 million records
- Training Data Features: **XNM**, AXP, **SSN**, **DOB**, **SEX**, FPF2

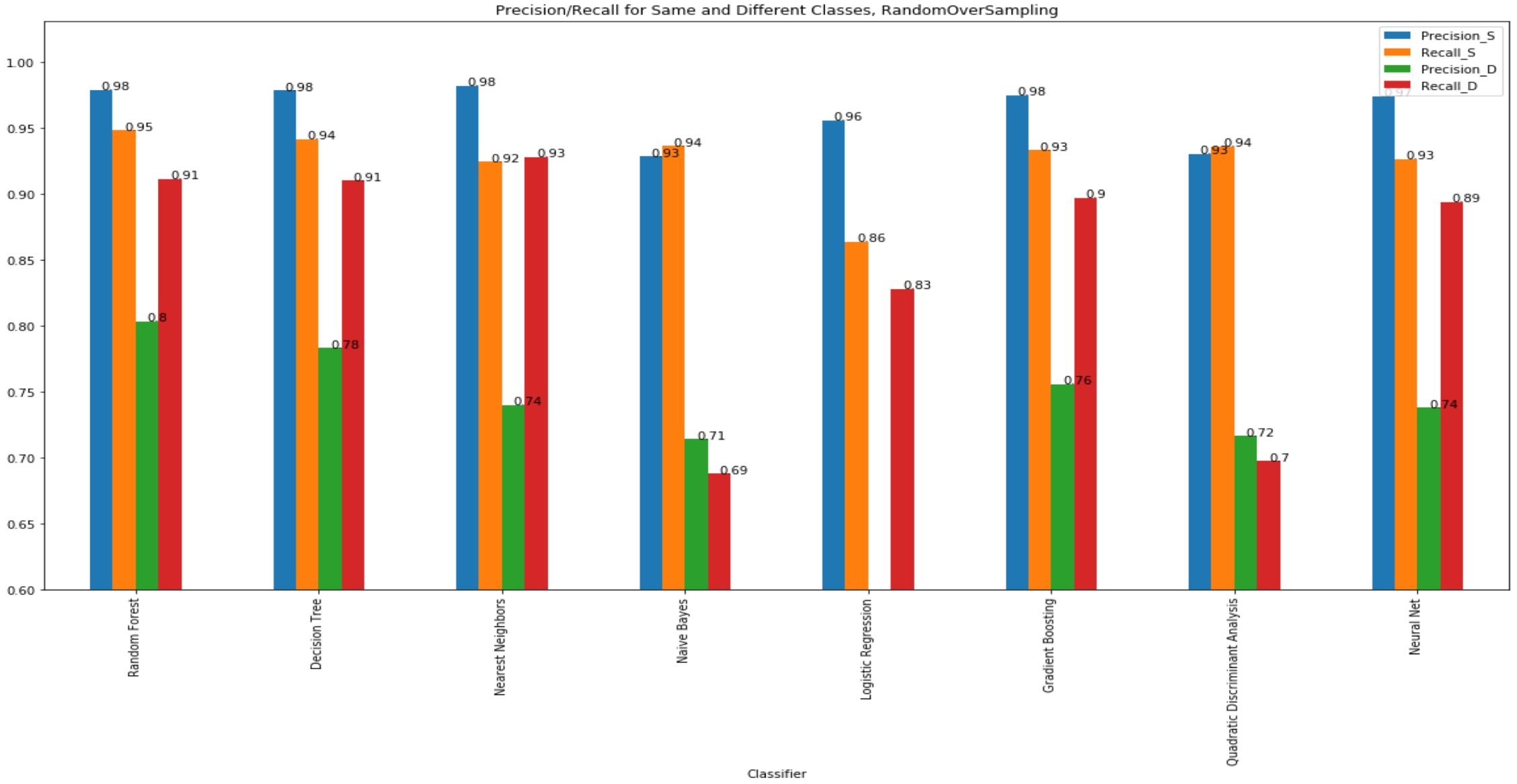
```
MEMRECNO,MEMRECNO2,CAUDTIME,MAUDTIME,RULETYPE,XNM,AXP,SSN,DOB,SEX,FPF2,OVERALL_CMPSCORE
29955364,45928598,2015-01-02 08:07:44,2015-01-02 08:07:44,S,+0.66,+0.13,+0.00,+4.47,+0.26,-3.00,2.5
33087603,45928598,2015-01-02 08:07:44,2015-01-02 08:07:44,S,+0.66,+0.13,+0.00,+4.47,+0.26,-3.00,2.5
32192384,45928598,2015-01-02 08:07:44,2015-01-02 08:07:44,S,+0.66,+3.20,+0.00,+4.47,+0.26,-3.00,5.5
30214332,46274721,2015-01-02 08:10:07,2015-01-02 08:10:07,S,+8.27,+1.33,+0.00,+4.55,+0.26,-
2.00,12.4
46274721,46331036,2015-01-02 08:10:07,2015-01-02
08:10:07,S,+8.27,+4.71,+5.01,+4.55,+0.26,+0.00,22.8
30214332,46331062,2015-01-02 08:10:07,2015-01-02 08:10:07,S,+8.27,+4.71,+0.00,+4.55,+0.26,-
2.00,15.7
46220762,46315567,2015-01-02 09:35:55,2015-01-02 09:35:55,D,+8.07,+4.71,+0.00,+4.45,+0.35,-
6.00,11.5
25754083,46264503,2015-01-02 15:32:23,2015-01-02 15:32:23,D,+2.28,+1.33,+0.00,+4.53,+0.35,-3.00,5.4
25754083,46262360,2015-01-02 15:32:23,2015-01-02 15:32:23,S,+8.27,+1.33,+0.00,+4.53,+0.35,-
2.00,12.4
25754083,36498439,2015-01-02 15:32:23,2015-01-02 15:32:23,S,+8.47,+4.71,+0.00,+4.53,+0.35,-
2.00,16.0
46262360,46264503,2015-01-02 15:32:23,2015-01-02 15:32:23,D,+2.28,+4.71,+0.64,+4.53,+0.35,-6.00,6.5
36532201,46264503,2015-01-02 15:32:23,2015-01-02
15:32:23,S,+8.27,+4.71,+5.01,+4.53,+0.35,+0.00,22.8
36532201,46262360,2015-01-02 15:32:23,2015-01-02 15:32:23,D,+2.28,+4.71,+0.64,+4.53,+0.35,-6.00,6.5
36498439,46264503,2015-01-02 15:32:23,2015-01-02 15:32:23,D,+2.28,+4.71,+0.64,+4.53,+0.35,-6.00,6.5
36498439,46262360,2015-01-02 15:32:23,2015-01-02
15:32:23,S,+8.27,+4.71,+5.01,+4.53,+0.35,+0.00,22.8
```

Exploration of Many Machine Learning Algorithms & Results

- Evaluated multiple classifiers
  - Random forest showed best results
- Skewed Matching Data
  - 80% same, 20% different
  - Evaluated different sampling methods
- Results of tuned model using oversampling
  - Accuracy = 0.94
  - Precision = 0.94
  - Recall = 0.94
- Used 80% of randomly selected data to train model
- Used remaining 20% to verify ML results

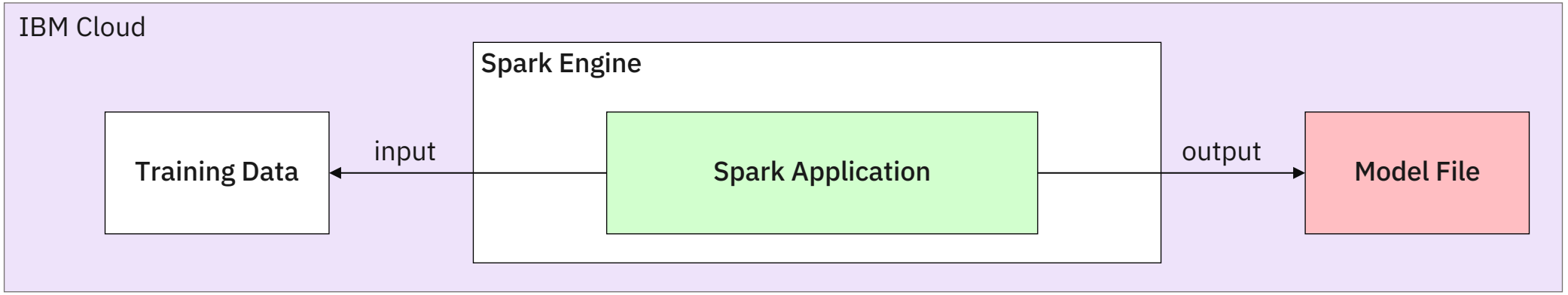


Precision & Recall for Same / Different Match Results across ML Algorithms

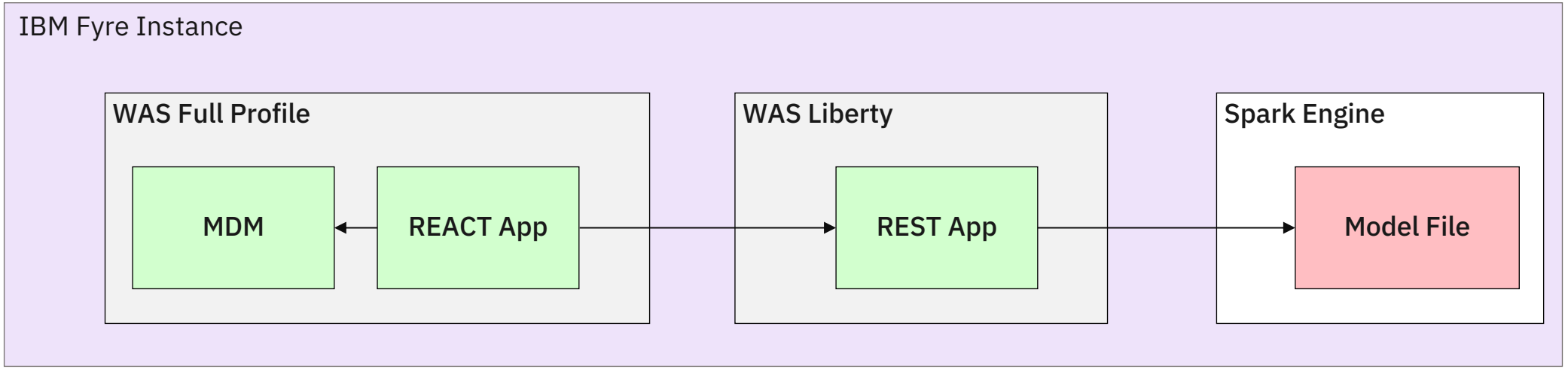


Prototype Using Spark Engine for Machine Learning

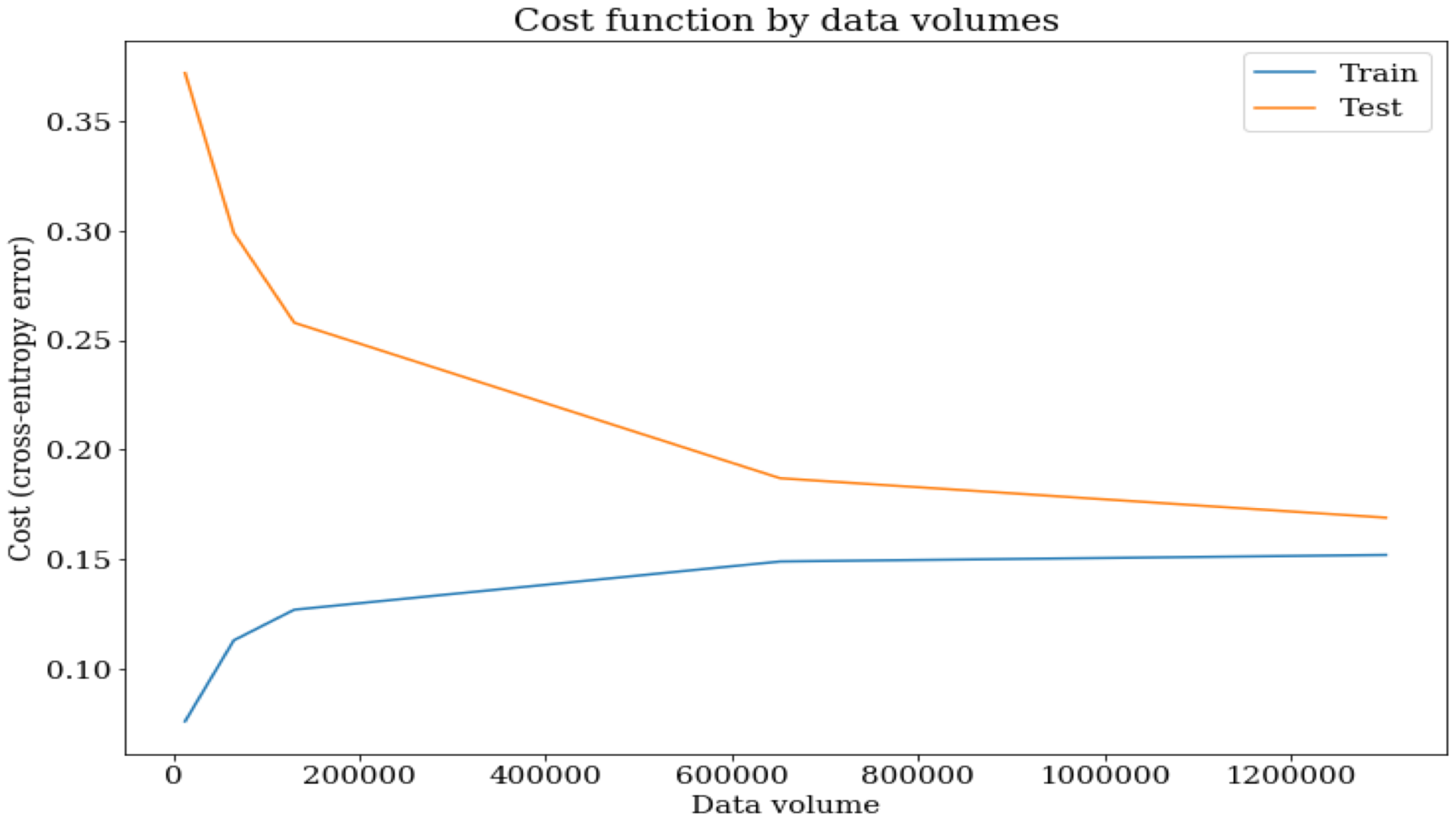
Model Training



Prototype

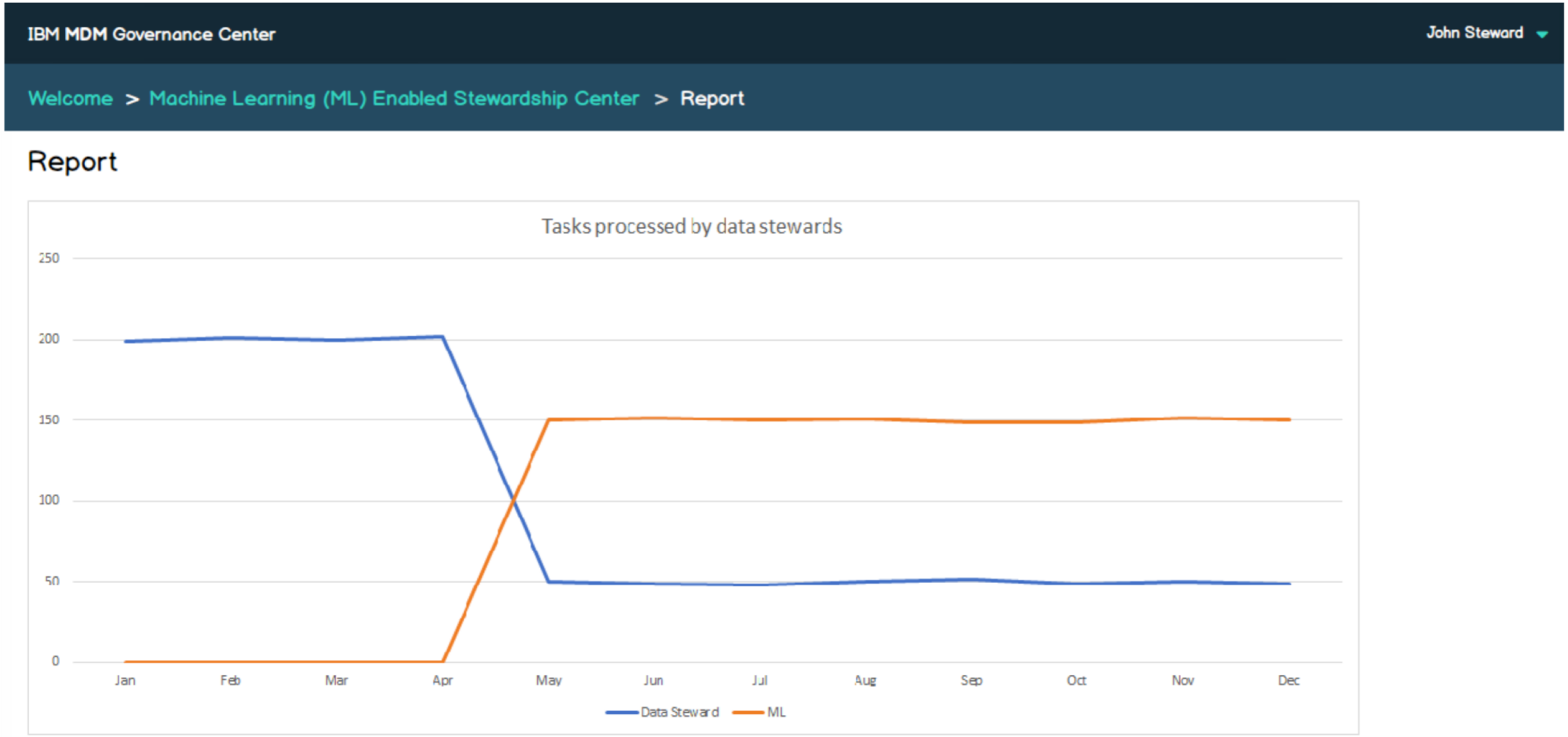


Learning Curve using Cross-Entropy



Run Demo

1) Run with ML for a while
2) Once results are trusted, all clericals above ML confidence threshold get auto-collapsed



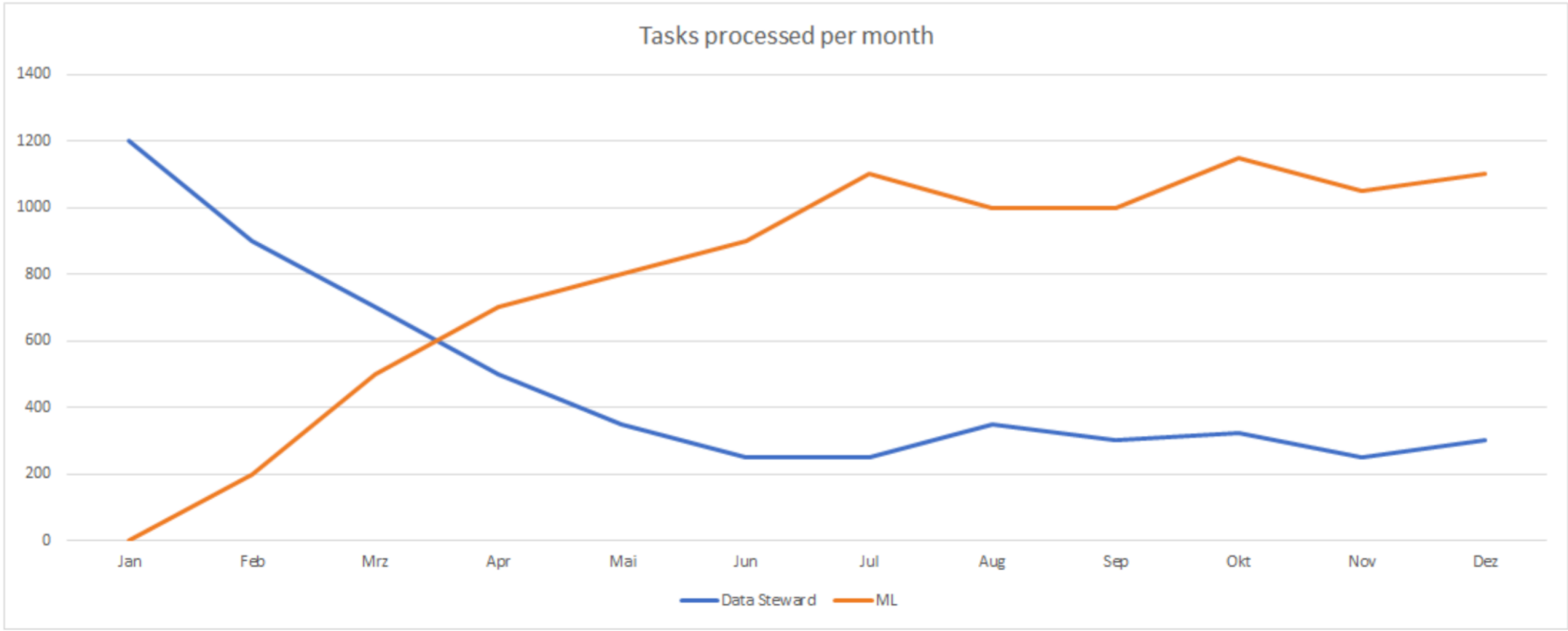
Benefits of using ML for Data Stewardship (Mockup Report)

IBM MDM Governance Center

John Steward

Welcome > Machine Learning (ML) Enabled Stewardship Center > Report

Report



Thank you