A Case Study of Deploying R2R Strategy with Deep Learning Model in High-Mix Semiconductor Manufacturing

Yulei Sun, Ph.D.

OCTOBER 6, 2021

APCSM Conference

innovation™ onto

Outline

- Introduction to Onto Innovation
- Overview of AI/ML and applications in semiconductor manufacturing
- R2R case study for CVD with RNN deep learning model
  – Challenges
  – Results
- Conclusion

Onto Innovation Snapshot

High tech capital equipment company specializing in optical process solutions for semiconductor and related markets

COMPANY
NYSE (ONTO)

CASH ON BALANCE SHEET (3/27/2021)
$393M Cash
No Debt

CASH FLOW FROM OPRERATIONS
30% of Revenue[3]

TTM REVENUE[1]
$613 Million

GROSS MARGIN[5]
54%

YOY GROSS MARGIN IMPROVEMENT[4]
+18%

COST REDUCTION[2]
$24 Million

OPERATING MARGIN[5]
25%

YOY OPERATING MARGIN IMPROVEMENT[4]
+56%

Note 1: Trailing 12 months actual results
Note 2: Implemented cost synergies from the Nanometric / Rudolph merger
Note 3: Q1-2021 Cash from Operations: $51 Million, 30% of Revenue
Note 4: Q1-2021 YoY change compared with Q1-2020
Note 5: Q1-2021 Non-GAAP Gross Margin & Operating Margin.

Onto Enterprise Business Unit (Software)

Manufacturing Process Control and Analysis

Software Solutions

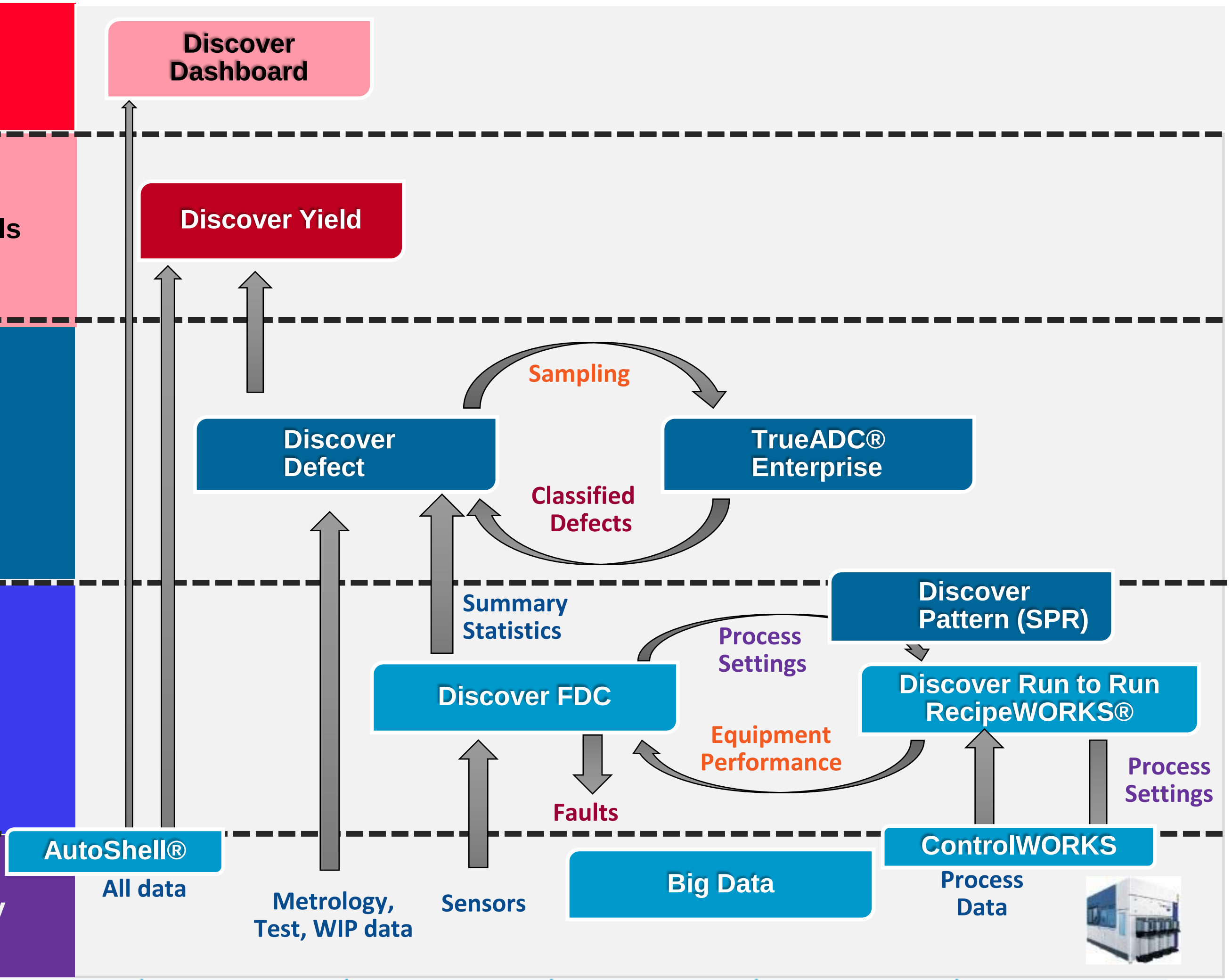
VISUALIZATION
- Interactive drill down
- Configurable dashboards

MODELING
- Use traditional statistics and MVA techniques to enhance process models
- Automate yield learning

MONITORING
- Focus on yield limiting defects to decrease yield production times
- Use process data to improve analysis and tighten limits

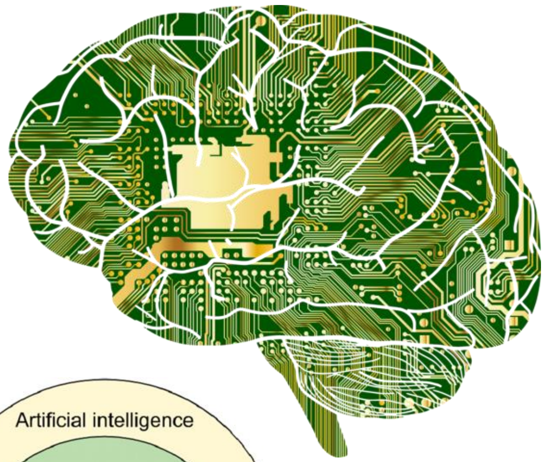
CONTROL
- Use R2R and FDC to control and monitor process
- Prevent model adjustments on runs with faults

COLLECT
- Gather wafer & tool data at every step

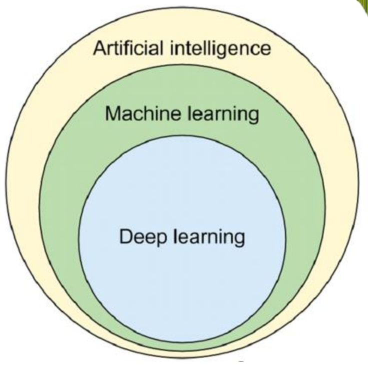


	Defect Analysis	Yield Analysis	Auto Defect Classification	Adv. Process Control	Fault Detection	Factory Automation	Equipment Control
Onto	✓	✓	✓	✓	✓	✓	✓

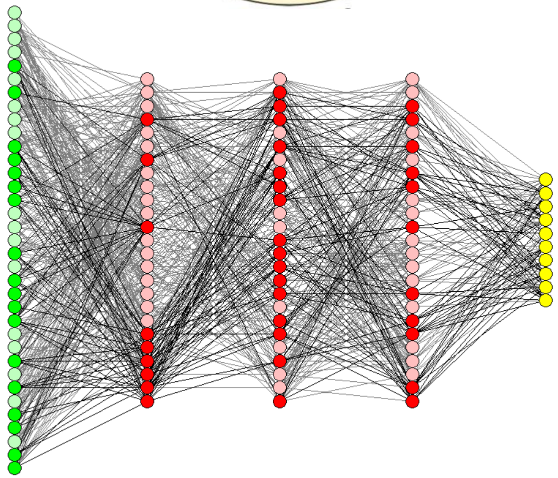
Artificial Intelligence (AI) and Machine Learning (ML)



- In the past ten years, AI and ML technologies have found their way into many different areas and disruptively changed our life and way of solving problems.
  – Refine results in online search and shopping, customize advertising, tailor news feeds, and even drive cars
  – Primarily driven by tremendous expansion in availability of data and computing power
- Terminologies
  – **Artificial Intelligence (AI)**: used famously by Alan Turning about possibility of machines being able to imitate human beings doing intelligent things, e.g., playing chess.
  – **Machine learning (ML)**: subset of AI that allows automation of learning based on evaluation of past results against specified criteria
  – **Deep learning (DL):** subset of ML that refers to a multi-layered learning hierarchy in which the output of each layer is the input for next layer, e.g. neural networks

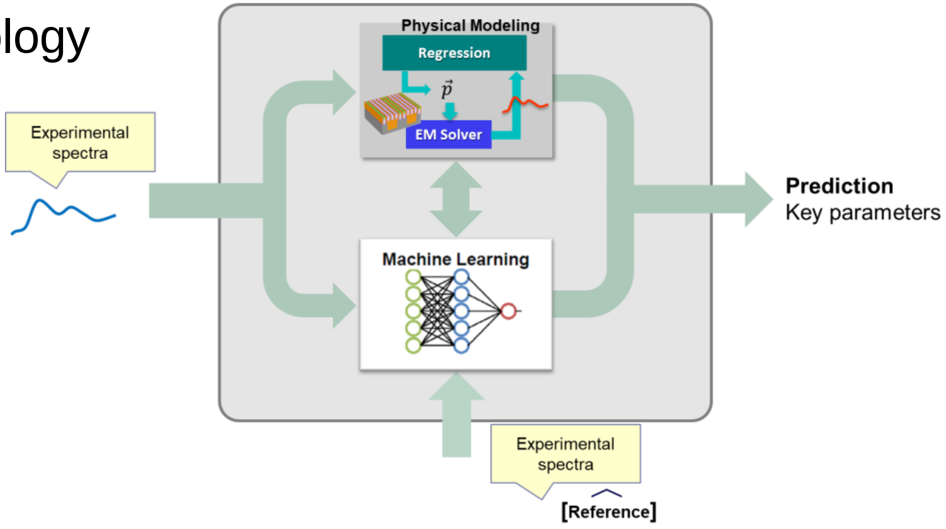


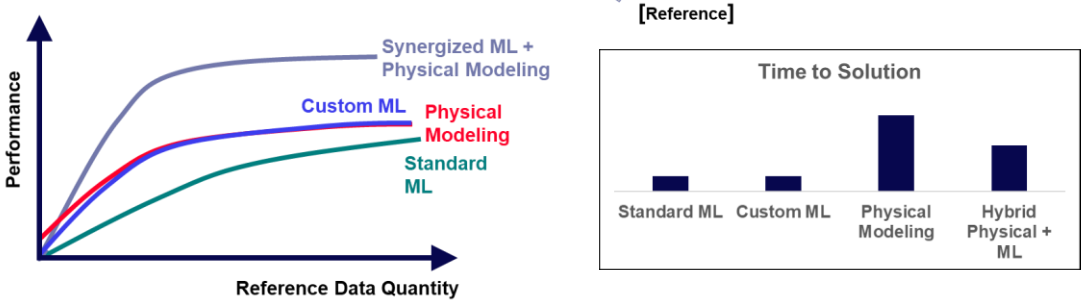
Wikipedia



AI/ML in Semiconductor Manufacturing

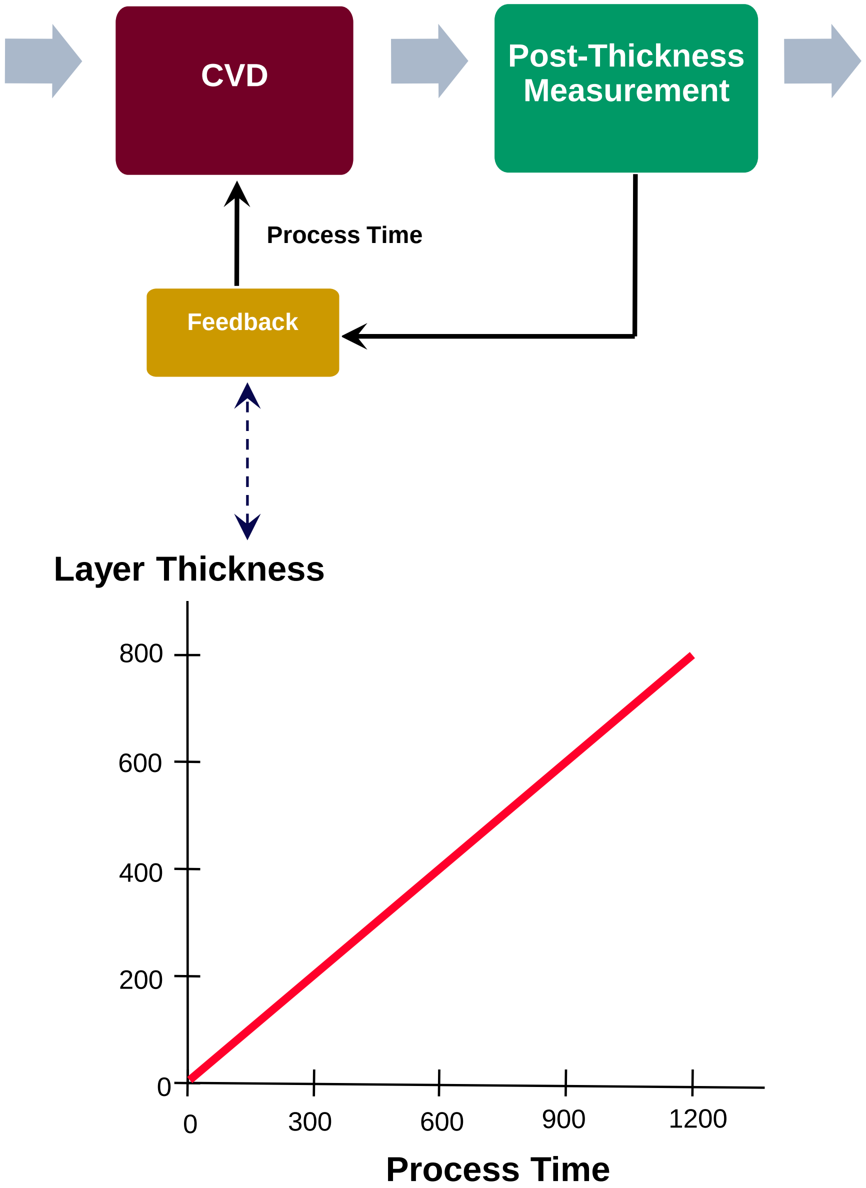
- Ability of AI/ML to learn from data autonomously and quickly find patterns and correlations has found its applications in ONTO metrology and inspection systems
  - AI based spatial pattern recognition (SPR) system for inline wafer monitoring *(Bachiraju, 2021 APCSM)*
  - Automatic defect classification (ADC) system with ML models
  - Machine learning based optical critical dimension (OCD) metrology *(Wong et al, 2021 SPIE)*
- Pure AI/ML approach greatly challenged by new and uncharacterized situations. Combination of traditional approach and AI/ML techniques gives better performance.
- Great work applying AI/ML technologies in run-to-run (R2R) control strategy design and implementation
  - Tutorial on AI for semiconductor manufacturing *(Redman et al, 2018 APC Conference)*
  - Evaluated different neural networks in R2R control modeling and compared their performance with traditional exponentially weighted moving average (EWMA) models *(Chen et al, 2019 APC Conference & 2020 APCSM)*





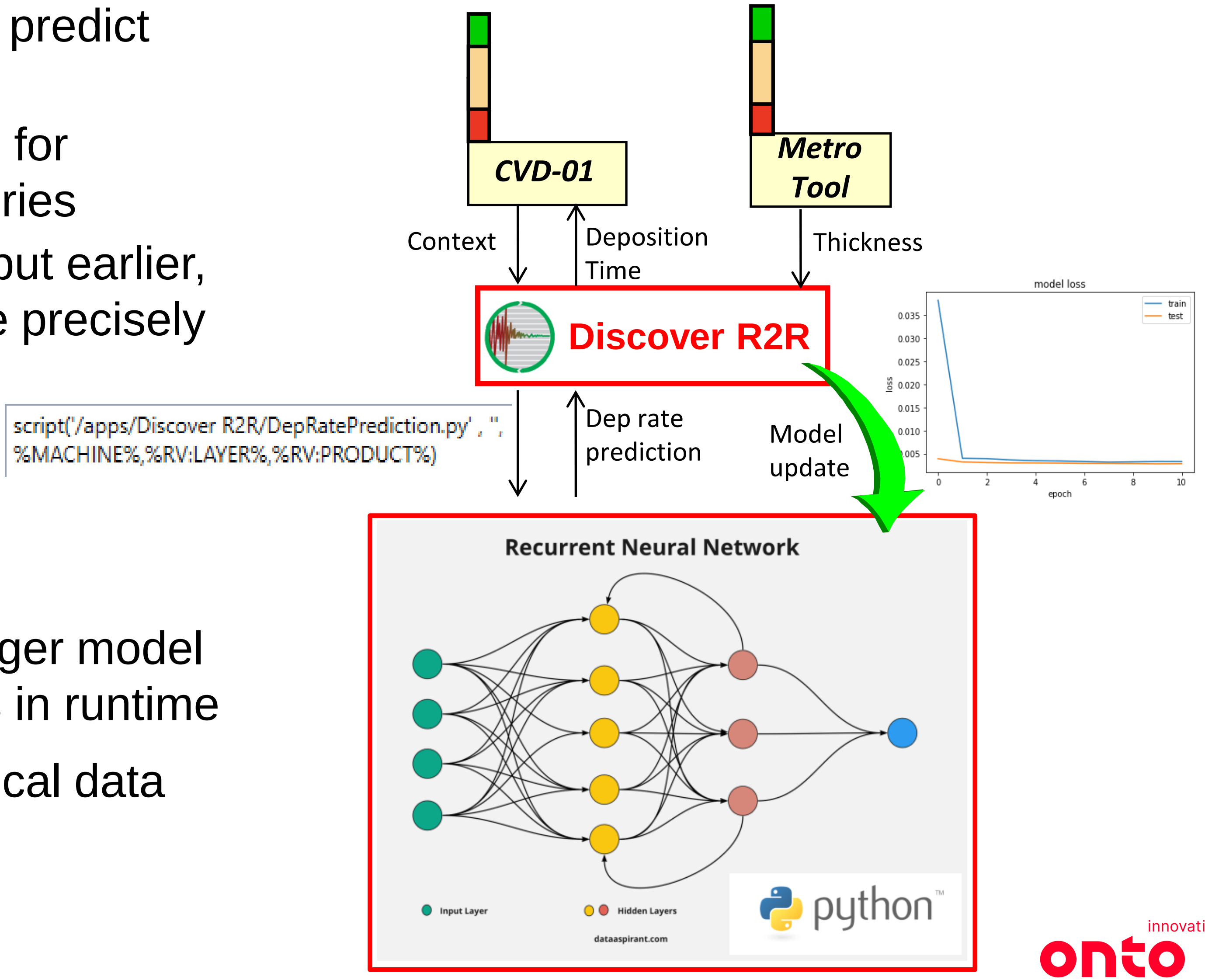
Synergized physical modeling & ML hybrid approach for OCD, 2021 SPIE

CVD Process Flow and Traditional R2R Design



R2R Control with RNN Deep Learning Model

- Long Short-Term Memory (LSTM) recurrent neural network (RNN) model built in Python to predict deposition rate
  - Internal memory of RNN make it ideal for sequential data analysis, e.g., time series
    - RNN capable of remembering an input earlier, crucial for predicting outcomes more precisely

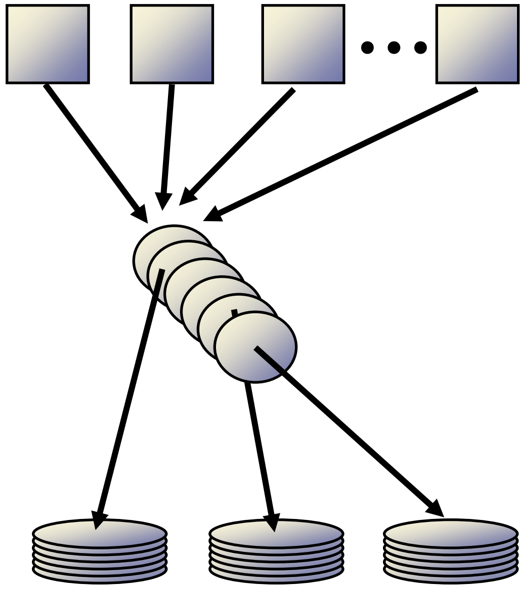


- Discover R2R calls Python script to trigger model prediction and calculates setting values in runtime
- RNN model re-trained with latest historical data

Challenge 1: Collecting Enough Good Data

- Building a solid and effective ML model requires guidance from domain expertise
  - Correlation and meaning inherent in the dataset to be extracted
  - Enough features (or inputs to the model), preferably independent of each other, need to be incorporated in the model.
    - e.g., manufacturing context, equipment hardware parameters, consumable usages, and upstream parametric data, etc., that can impact the output values of the model (or labels).
    - Similar to partition/control thread definition in traditional R2R

- Large dataset does not assure successful ML model
  - We need good data containing enough variabilities for model to extract.
  - Sometimes, inline production data does not contain enough variation within the settings in a short timeframe, e.g., running fixed setting values for extended period of time.



features (Machine, Layer, Product) — label (ObservedDepRate)

	Machine	Layer	Product	ObservedDepRate
0	CVD-01	1	3059	0.685098
1	CVD-01	1	3059	0.690010
2	CVD-01	1	3059	0.683857
3	CVD-01	1	3061	0.695134
4	CVD-01	1	3061	0.696837
...	...	...	...	...
2995	CVD-01	4	3059	0.598119
2996	CVD-02	3	3061	0.648666
2997	CVD-01	3	3059	0.687119
2998	CVD-01	3	3059	0.686212
2999	CVD-02	3	3131	0.655170

Challenge 2: Data Preparation and Pre-processing

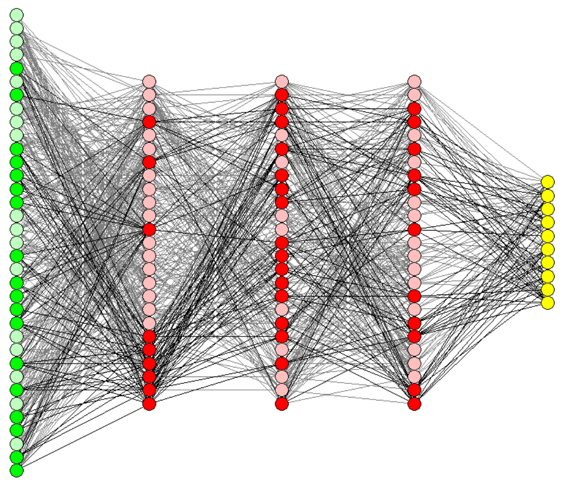
- Categorical manufacturing context data, e.g., strings, must be converted to numerical data, e.g., one-hot encoding
  - Most ML algorithms typically only handle numerical values, e.g., matrices
  - ML models can only predict outputs from known inputs, therefore any new category, e.g., new product, requires the entire model to be re-built
    - Similar issue with non-threaded control
    - Large dataset may not be available in high-mix manufacturing environment, especially for low-running products.
  - Feature scaling is another critical step during pre-processing of data before creating a ML model
    - Better performance obtained with scaled data so that features are evaluated by their proportions, not sizes.

	Machine	Layer	Product	ObservedDepRate
0	CVD-01	1	3059	0.685098
1	CVD-01	1	3059	0.690010
2	CVD-01	1	3059	0.683857
3	CVD-01	1	3061	0.695134
4	CVD-01	1	3061	0.696837
...	...	...	...	...
2995	CVD-01	4	3059	0.598119
2996	CVD-02	3	3061	0.648666
2997	CVD-01	3	3059	0.687119
2998	CVD-01	3	3059	0.686212
2999	CVD-02	3	3131	0.655170

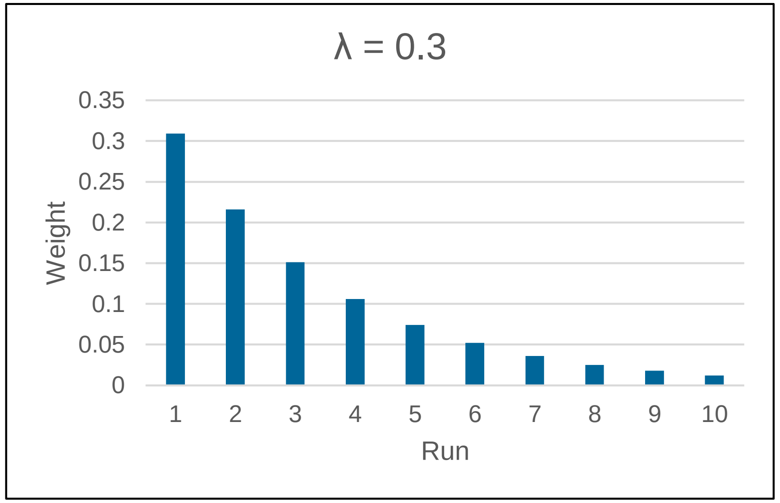
	0	1	2	3	4	5	6	7	8	9	...	11	12	13	14	15	16	17	18	19	ObservedDepRate
0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.685098
1	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.690010
2	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.683857
3	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.695134
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.696837
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2995	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.598119
2996	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.648666
2997	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.687119
2998	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.686212
2999	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.655170

Challenge 3: Hyperparameter Tuning

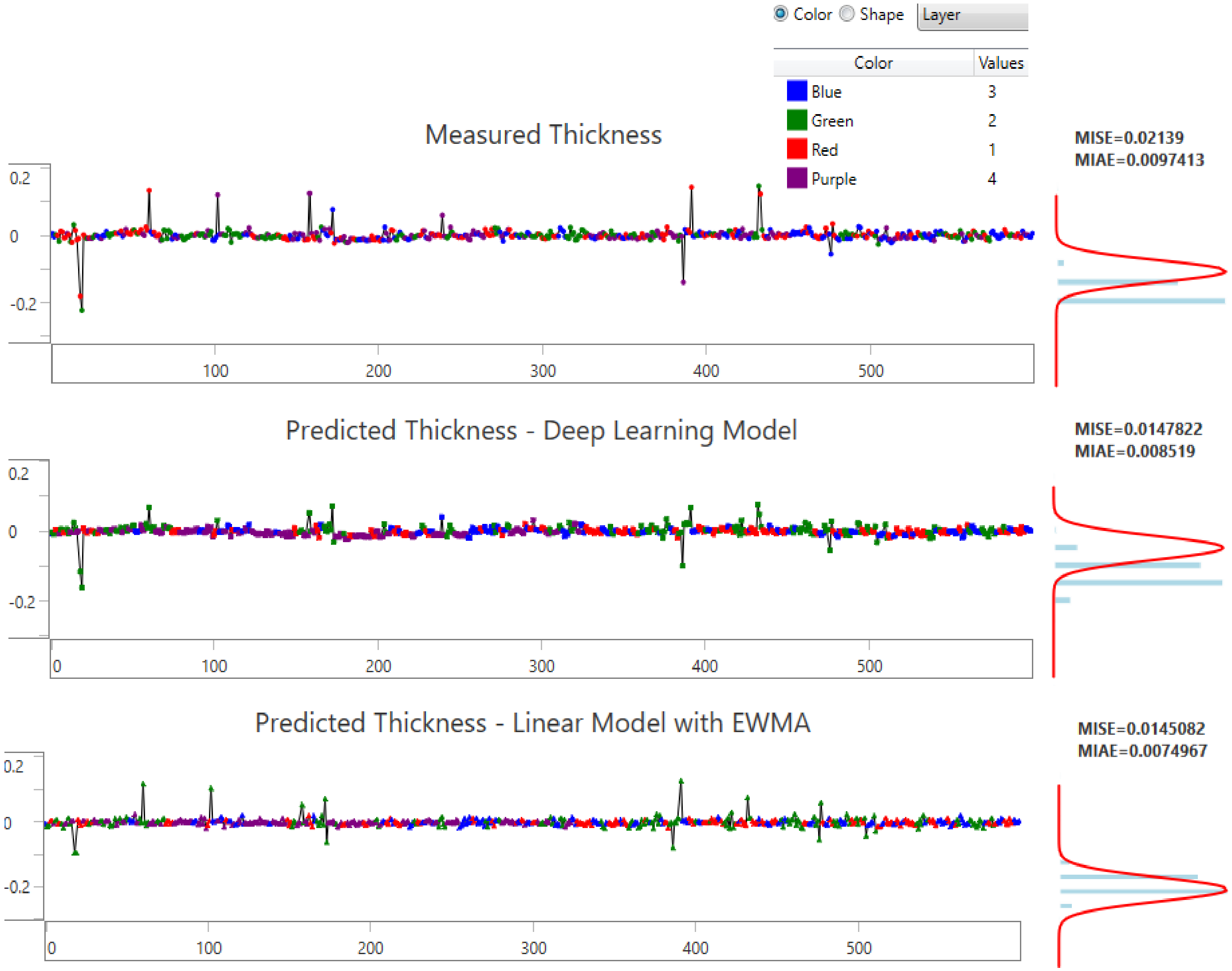
- Variables used to control learning process and can significantly affect model performance
  - e.g., number of hidden layers in neural network, number of neurons at each layer, batch size, epoch, dropout, etc.
  - Their values cannot be estimated from data and must be set manually by the practitioners, e.g., data scientists, using heuristics.
  - One cannot know the best value for a hyperparameter on a given problem, and may use
    - Rules of thumb
    - Copy values used on other problems
    - Search for the best value by trial and error



VS



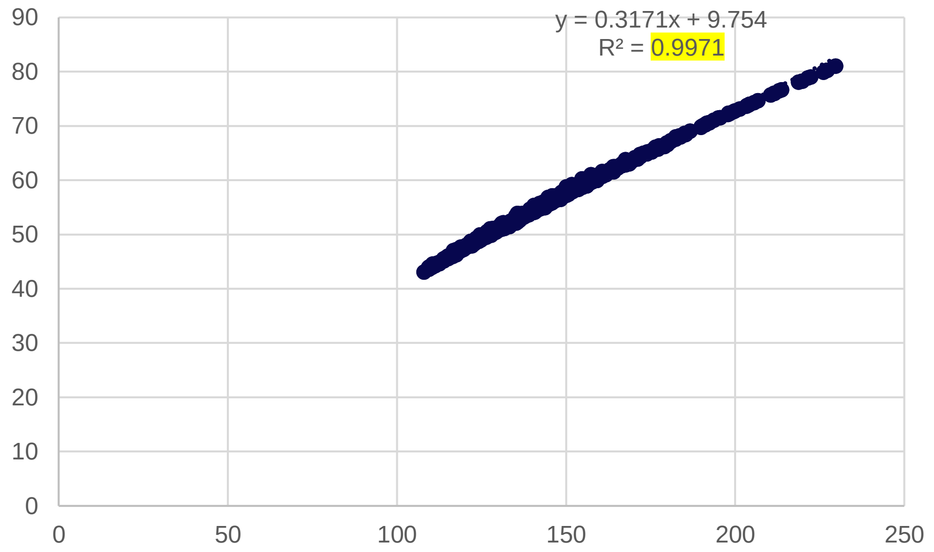
Results



- RNN model performs similarly to traditional linear model and EWMA algorithm but not better
  - e.g., 31% RMSE reduction vs 32%
- May not be beneficial to deploy such complex model in high-mix production due to large overhead

Summary

- Areas ideal for AI application
  - Repetitive, mental labor-intensive
  - Good example data available
  - Analytical methods not applicable or too hard
  - Interpolation
- Tasks NOT suitable
  - Accurate measurement
  - Single-shot situation
  - Generalization
- Unlikely to replace traditional R2R solutions, AI/ML solutions can provide complementary capability
  - Complex, time sensitive situations where accurate physical or statistical model not ready yet, e.g., MOCVD
  - Edge cases that traditional R2R finds challenging, e.g., first run after model reset due to PM, extremely low running products
- Right tool and approach can significantly speed up search for an accurate statistical model *(Sun, 2020 APCSM Conference)*
  - A simple model of R^2 (goodness-of-fit) close to 1 makes any complex model unnecessary



Thank You

谢谢 | 謝謝

ありがとう

Obrigado

Danke

감사합니다

Merci

info@ontoinnovation.com
www.ontoinnovation.com