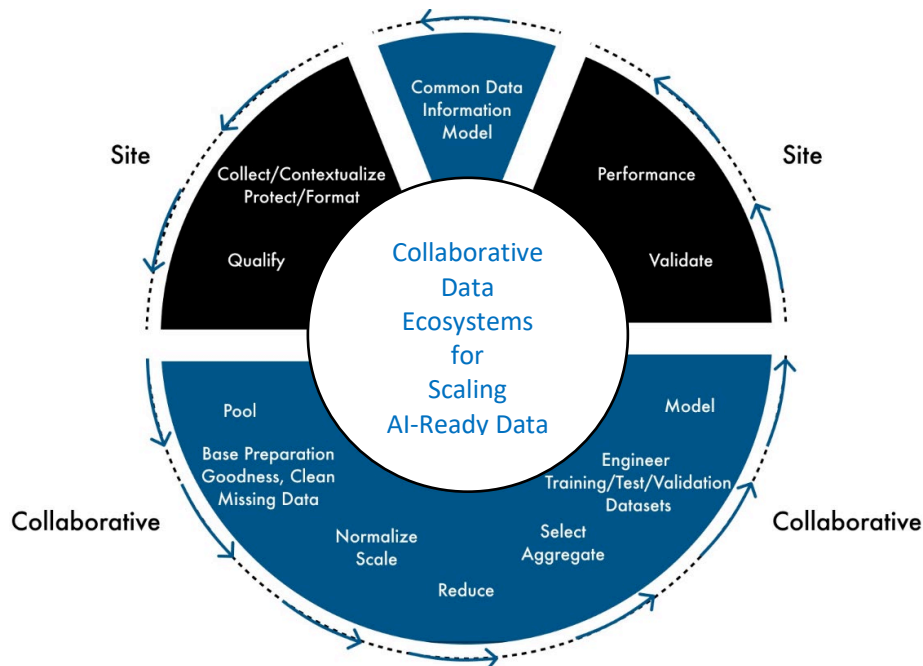




NIST Advanced Manufacturing Series NIST AMS 100-72

Artificial Intelligence with Open and Scaled Data Sharing in Semiconductor Manufacturing – Workshop Report

Sthitie Bom
Jim Davis
Said Jahanmir
Bruce Kramer
Don Ufford
Gregory W. Vogl



This publication is available free of charge from:
<https://doi.org/10.6028/NIST.AMS.100-72>

**NIST Advanced Manufacturing Series
NIST AMS 100-72**

Artificial Intelligence with Open and Scaled Data Sharing in Semiconductor Manufacturing – Workshop Report

Sthitie Bom
Seagate Technology

Jim Davis
UCLA

Said Jahanmir
*Office of Advanced Manufacturing
NIST*

Bruce Kramer
*Office of Advanced Manufacturing
NIST*

Don Ufford*
*Office of Advanced Manufacturing
NIST*

Gregory W. Vogl
*Engineering Laboratory
NIST*

**Former NIST employee; all work for this publication was done while at NIST.*

This publication is available free of charge from:
<https://doi.org/10.6028/NIST.AMS.100-72>

November 2025



U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST Disclaimer

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

NIST Technical Series Policies

[Copyright, Use, and Licensing Statements](#)

[NIST Technical Series Publication Identifier Syntax](#)

Publication History

Approved by the NIST Editorial Review Board on 2025-11-14

How to Cite this NIST Technical Series Publication

Bom S, Davis J, Jahanmir S, Kramer B, Ufford D, Vogl GW (2025) Artificial Intelligence with Open and Scaled Data Sharing in Semiconductor Manufacturing – Workshop Report. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Advanced Manufacturing Series (AMS) NIST AMS 100-72.

<https://doi.org/10.6028/NIST.AMS.100-72>

Author ORCID iDs

Jim Davis: 0009-0001-6340-6310

Gregory W. Vogl: 0000-0003-4238-851X

Contact Information

said.jahanmir@nist.gov

bruce.kramer@nist.gov

gregory.vogl@nist.gov

Abstract

The processing of manufacturing data is still not yet a priority. It is often not done well or consistently for use with artificial intelligence (AI). It is rarely scaled across applications or operations and remains largely closed for cost-effective application and training development. Like mined minerals, raw operational data in manufacturing has little value until it is qualified, refined, concentrated, and readied in sufficient amounts for use. The Workshop, “Artificial Intelligence with Open and Scaled Data Sharing in the Semiconductor Industry,” sponsored by the National Science Foundation (NSF award 2334590) and supported by the National Institute of Standards and Technology (NIST), was organized to evaluate the promise of Scaled Data Sharing in manufacturing ecosystems while also addressing how to overcome barriers.

This report articulates and benchmarks significant industry-defined economic opportunities, many untapped, with scaled and more open, but managed and cost-effective data sharing for producing AI-Ready data, and AI, Machine Learning (ML) and Digital Twin (DT) applications. The report also describes how to proceed with factory and company data ecosystems. It further lays out important business principles for Data-First strategies, as well as how and when to share data for individual company benefit. Workforce training will eventually be driven by the business value of data sharing. Starting out, however, on-the-job workforce training programs on data processing are needed to drive the ability to build out all facets of data sharing value.

A coalition of manufacturers in the semiconductor industry addressed how to pool and process data for current and new operational opportunities and the greater orchestration and scaling of AI/ML/DT solutions. A Seagate Technology/UCLA (University of California, Los Angeles) team benchmarked multiple economic value points when processing and using data from multiple machines as a Collaborative Data Ecosystem (CDE) for an AI/ML metrology application. The result is a scalable, cross-factory/company approach that greatly expands the value of Scaled Data Sharing. The economies of consistently processed data, data aggregation, and application when acting together are maximized. Model performance is improved.

The Workshop demonstrated that manufacturers, even competitors, can agree on a meaningful process for Scaled Data Sharing to achieve significant factory and company operational benefits. Factories and companies can act together to prepare data consistently and build collaborative AI/ML models, e.g., for wafer operations, while acting separately on their own products and applications. Importantly, data, processed consistently from operations at different sites, leads to more robust modeling, open cross-operational opportunities and insights, and suggests new data service possibilities, as well as a grand market-driven vision for implementing AI/ML systems that draw value from cross-company data. All of this was accomplished cost effectively by working with industrial participants as a CDE using existing workforces and known technologies.

Keywords

Advanced Manufacturing; AI; Artificial Intelligence; Collaborative Data Ecosystem; Data Sharing; Machine Learning; Manufacturing; ML; Semiconductor Manufacturing; Smart Manufacturing

Table of Contents

1. Executive Summary.....	1
2. Summary of Scaled Data Sharing Findings for an Industry Ecosystem	5
2.1. Benchmarking an Industry CDE	6
2.2. Benchmarked Findings.....	9
2.3. Critical Industry Mindset Factors.....	15
2.4. A Line of Sight to a Grand Vision at the Scale of the Internet.....	17
3. Background on Data Sharing	21
3.1. Economic and Energy Productivity Potential Drawn from Data	22
4. Semiconductor Industry Opportunity.....	25
5. Framework for Forming a CDE.....	28
6. Workshop Conclusions	30
Appendix A. NSF/NIST Workshop 2023-2025 Participants.....	34
Appendix B. Integrating the Semiconductor Use Case and Workshop Roundtables	39
B.1. Designed to Build an Industry Ecosystem and Benchmark the Outcomes	39
B.2. A Unique Workshop on Business and Technical Execution with Data	39
B.3. The Workshop Roundtables.....	42
B.3.1. Semiconductor Industry Use Case Roundtables 1 and 2 on Oct. 10 and 24, 2023: The Business Case for Increasing Productivity of Metrology	42
B.3.2. Data Sharing Infrastructure Roundtable 3 on Nov. 14, 2023: Minimum Viable Infrastructure.....	43
B.3.3. AI Ready Data Roundtable 4 on Feb. 28, 2024: Individual Factory and CDE Responsibilities as Coalition Governed Workflows	44
B.3.4. Data Application Roundtable 5 on Dec. 9, 2024	44
Appendix C. Use Case Specifics.....	46
Appendix D. Data Processing Use Case Baseline	50
Appendix E. Analysis Details on Benchmark Findings.....	53
E.1. Sustaining Data Processing Consistency in Operational Use	53
E.2. Benchmark Data Processing Consistency Economics	58
E.3. Benchmark Outcomes with Aggregating Data for Model Robustness	59
E.4. Benchmarking Site Contextualization, Qualification, Categorization, and Formatting	61
E.5. Benchmarking Operating Model and Governance.....	63

List of Tables

Table 1. Key Benchmarked Findings 10
Table 2. Eight Key Execution Principles for Industry Data Sharing 15
Table 3. Rank Ordered Priority Application Areas..... 29

List of Figures

Figure 1-1. Scaling Data Sharing and Data Value 2
Figure 2-1. Roundtable and Benchmark Progression as Played Out in Workshop 7
Figure 2-2. Consistent and Collaborative Data Refinement and Model Building (see Figure E-1) 9
Figure 4-1. Quantified Potential of SM and AI/ML/DTs Developed by and for the CDE 26
Figure B-1. Semiconductor Manufacturing Process Flow..... 40
Figure B-2. Factory Line Operation Used for Benchmark Studies 41
Figure C-1. Schematic of a Typical Etch Tool 46
Figure C-2. Etch Tool Information Model: Ion Beam Subsystems and Processing Context 47
Figure C-3. The Elements of an Etch Machine Dataset – Single Row 37 Features and Variables 48
Figure D-1. Workflow of AI-Ready Steps Used for the Etch Machine Benchmark 50
Figure E-1. Analysis of the Benchmark Workflow for Data Processing Consistency 55
Figure E-2. Receiver Operating Characteristic Curve for Two Tools..... 61
Figure E-3. Projected Data Sharing Performance Requirements with Different Kinds of Applications 63

1. Executive Summary

What became abundantly clear during the Workshop was that qualified (operationally acceptable) data, processed consistently, was key to enhancing the value of data, a value that increased multi-fold with Scaled Data Sharing. Algorithms, methods, and standards were important, but data consistency was key to scaling applications, drawing full value from artificial intelligence (AI) / machine learning (ML) / digital twin (DT) systems, trusting data from many sources for other applications, and categorizing data for reuse and richer operational analytics. From data capture to data contextualization, data consistency to data sharing, and data aggregation to better models, every success in this study was driven by qualified and consistently processed data.

Contrary to many opinions, Smart Manufacturing solutions grounded in AI/ML/DT applications can be implemented in a cost-effective and governed manner for near-term benefits to a manufacturing operation. A Collaborative Data Ecosystem (CDE) that focuses on Scaled Data Sharing with data processing consistency makes it possible to: scale consistency and boost the value of data; build models collaboratively; build sharable data inventories; support data update cycles for sustained use; update and improve data processing and model building methods and technologies more effectively; and conduct single-source training with much greater effectiveness and a much lower individual company cost. Compared to top-down industry business and/or ontology frameworks, CDEs are market-driven, bottom-up business entities that form around increasing data value and common application objectives, driving higher value for each individual factory site when execution is collaborative.

Reports and trade articles, including the 2022 NIST report,¹ “Towards Resilient Manufacturing Ecosystems through Artificial Intelligence (AI),” have discussed the advantages of a Smart Manufacturing (SM) industry powered by scaling AI and data sharing from factory floors to supply chains. These reports describe the potential for transformational improvements in industry-wide productivity, precision, demand dynamics, supply chain resilience, and workforce effectiveness. They also describe significant technical, business, market, and mindset barriers that inhibit wide-scale, accelerated adoption. This report provides an industry perspective and demonstration on how to proceed.

Figure 1-1 from the 2022 NIST report illustrates the importance of Scaled Data Sharing for scaling both digital (black) and physical (blue) impacts. The impacts of scaling physical interoperability are shown in the vertical (blue) direction, where digital actions with AI applications create value not only at the level of physical machines and operations on the factory floor but also across factory/company enterprises and cross-company supply chains, i.e., scaled operational interoperability. The scaled industry data and digital application interoperability that supports these scaled operational impacts are shown horizontally (black), i.e., Scaled Data Sharing. At the intersection are shared data, tools, and applications, interconnectedness with trust, and network effects that drive data and digital scaling and integration for operational benefits.

¹ <https://www.nist.gov/publications/towards-resilient-manufacturing-ecosystems-through-artificial-intelligence-symposium>

Digital and Physical Network Effects at Scale



Smart Manufacturing (SM) is (1) scaled data sharing and data interoperability (horizontal), that is (2) synchronized, orchestrated, and scaled operationally with advanced sensor, control, platform, and software application systems (vertical & horizontal), for (3) preventive, predictive, proactive, zero incident, control, physical operational interoperability, and enterprise management, from (4) factory floors to supply chains (vertical) to achieve:

- increased productivity (energy, material, operations, and workforce),
- improved quality, optimized capacity utilization, and demand response,
- improved demand-dynamic performance (better, faster, and cheaper),
- decreased cost of regulatory compliance.

Artificial Intelligence (AI) in real-time manufacturing operations refers to software systems that learn and improve with synchronized data from the operations to enable, improve, and automate asset management, quality improvement, and operational interoperability at scale with real-time human and machine action:

- **Machine Learning (ML)** refers to algorithms that use prior data to identify the current state and predict future states, with the goal of improving productivity, precision, and performance.
- **Digital Twins (DT)** are virtual information constructs that mimic physical systems, which are dynamically updated with data, have predictive capabilities, and inform decisions that realize value (definition from 2024 National Academies Digital Twins Consensus Study).

Artificial Intelligence (AI) for shared data, tools, and applications refers to: (1) discoverability of scaled, secure, and managed datasets, (2) network search, discovery, and accessibility tools, and (3) selection and usability of qualified, categorized datasets, tools, and models.

Figure 1-1. Scaling Data Sharing and Data Value
[\[https://doi.org/10.6028/NIST.AMS.100-47\]](https://doi.org/10.6028/NIST.AMS.100-47)

This Workshop, Artificial Intelligence with Open and Scaled Data Sharing in the Semiconductor Industry, sponsored by the National Science Foundation (NSF award 2334590) and supported by the National Institute of Standards and Technology (NIST),² was organized to evaluate the promise of Scaled Data Sharing in data and operational ecosystems while also addressing significant barriers. The Workshop was designed as an industry demonstration about how a coalition of manufacturers in the semiconductor industry could agree to pool data to seek new operational opportunities with AI/ML/DT systems and return benefits to individual manufacturers. The Workshop demonstrated and benchmarked a scalable approach to producing qualified data that maximizes the value of data sharing by optimizing data processing, aggregation, usage, and model development when factories and companies act together.

The Workshop was conducted as a series of five Roundtables between July 2023 and July 2025. The central element was an agreed-upon AI/ML virtual metrology use case for which the data and modeling execution requirements could be considered in detail and ecosystem results could be benchmarked. The semiconductor coalition committed to testing ecosystem data sharing as a business priority. This emphasis highlighted the importance of a Data-First strategy. There was a cross-company recognition that, without data as a business priority, companies would continue to manage data as a second-class asset, undermining the value of AI/ML/DT systems.

An active decision was made to constrain the demonstration to existing technologies (no research and development) and to benchmark near-term benefits only. These decisions made it possible to focus only on practical business and technical data management and execution. Information and operation technology (IT/OT) providers, equipment builder perspectives, standards groups, consultants, and the CESMII³ Manufacturing USA Institute participated in various roundtables to provide technology perspectives, alternatives, and insights. The manufacturers conducted numerous pre- and post-roundtable meetings to make implementation decisions unencumbered with legacy services, data ownership models, and considerations of new technologies. As decisions were made, a Seagate Technology/University of California, Los Angeles (UCLA) project team benchmarked economic value points associated with the execution of the data processing and engineering steps needed to refine raw data from multiple similar machine sources into qualified (operationally acceptable), AI-Ready data for model building. The focus on multiple similar machines made it possible to study data aggregation for building more robust datasets. Benchmarking was done by comparing collaborative model performances for each machine with cross-machine vs. individual machine datasets.

² This material is based upon work sponsored by the National Science Foundation (NSF) under Grant 2334590 and further supported by the National Institute of Standards and Technology (NIST). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF or NIST. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States. Approved for public release; distribution is unlimited.

³ CESMII (Collaborative Ecosystems for Smart Manufacturing Innovation Institute) sponsored by the DOE as the 9th national Manufacturing USA Institute in 2017, see <https://www.cesmii.org>

The Workshop demonstrated that manufacturers, even competitors, can agree on a meaningful process for scaled data sharing to achieve needed consistency for significant factory and company operational benefits. Barriers needed to be addressed, and business and technical guardrails needed to be implemented. Factories and companies needed to act together to prepare data and build collaborative AI/ML models, e.g., wafer processing/flatness metrology, while still acting separately on their own products and applications. On the benefits side, pooling data from similar operations in different factories can lead to more robust modeling, open cross-operational opportunities and insights, and suggest new data service possibilities. AI/ML systems developed collaboratively can draw greater value from processing and using cross-company data. All of this was accomplished more cost effectively by working as a CDE compared to each company on its own. Also, this was done with technologies available today.

2. Summary of Scaled Data Sharing Findings for an Industry Ecosystem

Smart Manufacturing (SM) defines the orchestration of advanced digital technologies used to construct scaled software systems. Data is used to repeatedly learn, enable automation/autonomy, and interface with physical and human control and management systems. At scale, better products are made and substantially less energy and materials are used. Manufacturing operations are far more productive, demand-responsive, efficient, and resilient. The workforce is more effective and more engaged. Industry-wide adoption and scaled application of these software systems are expected to give birth to broad new industrial, business, and workforce capabilities that enhance U.S. leadership in manufacturing. The goals for using the data are not just about building single Artificial Intelligence/Machine Learning/Digital Twin (**AI/ML/DT**) applications. They are about scaling and sustaining industry-qualified data across multiple applications, aggregating data for more robust modeling, and opening insights from cross-operation analyses. All manufacturers – small, medium, and large – have valuable data and are a part of larger manufacturing systems.

This NSF/NIST Workshop brought industry, academic, and government experts together to define, demonstrate, and benchmark a cross-factory, cross-company business and technical strategy for scaling data availability and modeling capacity needed for factory, enterprise, and industry-wide operational impacts. The Workshop recognized that information and operation technology providers and equipment builders have been monetizing data services for some time to improve product and data service offerings. Manufacturers, however, have not developed the value and ownership of their own data for orchestrating and scaling AI-based SM in the context of their respective manufacturing operations.

Inside the box, a Collaborative Data Ecosystem is defined. SM that integrates and scales AI/ML/DT applications into solutions that are orchestrated and comprehensive must draw from large volumes of distributed operational data to produce **AI-Ready data**. AI-Ready data is consistently and persistently contextualized, qualified, prepared, and engineered for multiple applications at multiple operational scales. **Consistency** in data

Industry Smart Manufacturing Collaborative Data Ecosystem for Scaling Data and AI

- **Network of Manufacturing Factories** organized as collaborative, consistency focused, and scalable data ecosystems to create mutual value with data, solve common problems, and address shared goals.
- **Data Sharing** for using and reusing data from the multiple site locations from which it was collected.
- **Data-First Strategy** that emphasizes the importance of identifying mechanisms for gathering and sharing qualified data for use in real-time SM applications.
- **Data Interoperability** with shared multi-use data models about physical devices to enable persistent and consistent contextualization, formatting, categorization, and qualification of data at the source to make it reusable and sharable.
- **AI-Ready Data** prepared from interoperable data that is qualified, contextualized, and selected to extract operational value from AI applications.
- **Operational Interoperability** from factory floor aggregation, cross-factory/company operational exchange, and resilient supply chain use of AI-Ready Data in orchestrated AI applications integrated into physical and human systems for **Smart** (Preventive, Predictive, Proactive, Zero Incident) **Manufacturing**.

processing is the key objective. A **Data-First strategy** recognizes the need to first process raw operational data into AI-Ready data for any application to work. Consistently processed data is needed to make a highly valuable manufacturing resource usable across multiple applications. Today, a great deal of data is collected, but the critical processing is often not done, being constrained by legacy business and technical practices that embed data in single use applications or segment and isolate data in compartments or “silos.” These practices neglect the opportunity to use and scale data across factories, companies, and supply chains in multiple applications and scaled solutions. Compartmentalization and data embedding reduce both the usability and value of data and create real barriers to the availability of qualified and consistently processed data for the adoption of SM and scaled application of robust AI/ML/DT systems.

Furthermore, today’s legacy of data compartmentalization, vendor business models that rely on product lock-in and vendor data ownership, intellectual property (IP) protections that are one-size-fits-all, security risks associated with data sharing across a siloed infrastructure, and Do-It-Yourself strategies that fragment consistency and block data and model sharing, all combine to limit the value of data, prevent scaling, increase security risks, and lead to costly “reinvent it here” solutions.

This Workshop identified, examined, and evaluated the details of a manufacturing business strategy referred to as a **Collaborative Data Ecosystem (CDE)** to address scaled qualification and processing of data into AI-Ready data for multi-machine AI/ML/DT models and cross-operation systems. Key elements of the strategy were demonstrated and/or benchmarked with an industrial use case. The CDE was organized as an interconnected coalition of factories across multiple manufacturers that embraced scaled site-level **Data Interoperability**, cross-site **Data Sharing**, and consistently/collaboratively processed **AI-Ready data** with ecosystem-oriented **Data-First** business strategies. These digital features are referenced in the horizontal capabilities in **Figure 1-1** that power the growth and robustness of SM and the AI/ML/DT systems that are needed for scaling and orchestrating their operational orchestration. As viewed in this Workshop report, a CDE is comprised of sites within and across factories that can solve common problems, implement common solutions, and share the results for the benefit of all participants, while also enabling standardized approaches to security and the guardrails needed to protect proprietary content. By addressing data interoperability and consistent processing of AI-Ready data, CDEs facilitate data aggregation, exchange, and sharing for factory, company, and supply chain **Operational Interoperability** while also ensuring IT security and IP protections. Data and digital systems are applied not only within individual factories but are also positioned for factory and cross-company enterprise applications (i.e., the full range of the vertical elements, bottom to top, in **Figure 1-1**).

2.1. Benchmarking an Industry CDE

A key recommendation in the 2022 NIST report was to organize a CDE demonstration to benchmark how to start, collaborate, and achieve timely benefits with data sharing, while

applying only existing technologies. The recommendation emphasized a cross-company investment strategy for the coalition to start, build experience, build capability, and grow benefits that lead to an improvement cycle. It also recognized that multiple manufacturers, including competitors, need to agree on technology, business, workforce, and governance foundations to support a practical implementation of data sharing and interconnectedness with trust as noted in **Figure 1-1**. Given that current market drivers, industry value systems, and IP and security risk concerns are significant barriers to data sharing, there also needs to be a change in business mindset that makes it possible for manufacturers to consider how to place a value on their own data. The Workshop followed directly from this prior recommendation and focused on data sharing for producing pooled and scalable datasets as the foundational aspect to demonstrate. One ML application of common interest was collaboratively developed with the pooled data so that model performance could be benchmarked. **The premise that was born out is that if the data are qualified, consistent, scalable and trusted, cross-operational advantages follow.**

The Workshop brought together 32 factory engineers and data scientists from 12 semiconductor manufacturing companies. Data scientists from academic institutions across the country, industry experts on Information Technology (IT) and Operational Technology (OT) infrastructure, experts on price analysis and equipment building, and government leaders in advanced manufacturing comprised an additional 27 participants who challenged, proposed, and reviewed paths forward. As shown in **Figure 2-1**, five Roundtables addressed different questions with different groups of experts to collect recommendations on how to structure technical and business pathways, initiate plans, and execute them as a CDE. The Workshop was uniquely designed to also benchmark key intra- and intercompany requirements, governance, and technical recommendations by simultaneously building a collaborative AI/ML/DT application for semiconductor industry use as the Workshop evolved. This benchmarking, in parallel with the Workshop, played a major role in organizing each Roundtable.

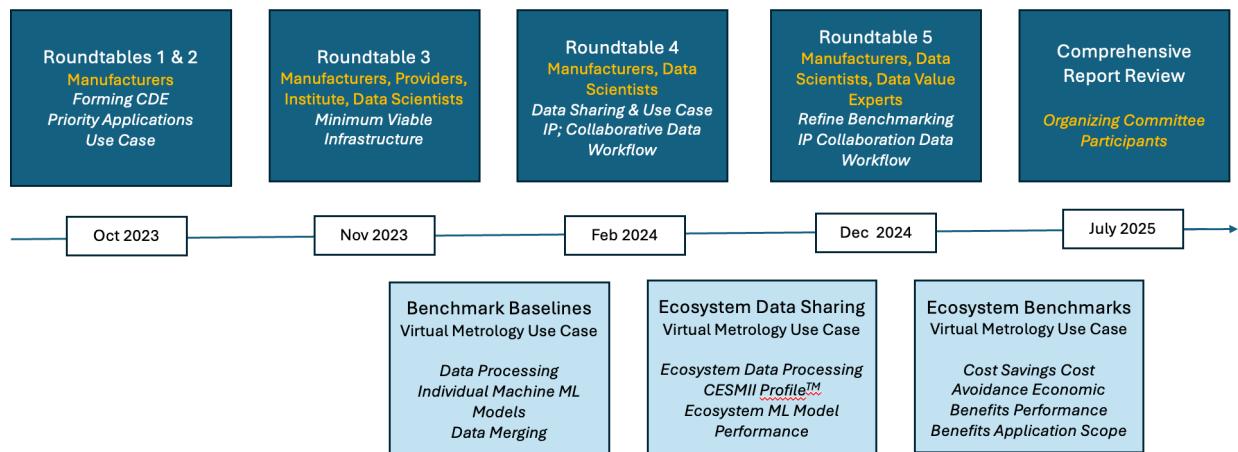


Figure 2-1. Roundtable and Benchmark Progression as Played Out in Workshop

With reference to **Figure 1-1**, the Workshop focused on a production improvement/quality assurance use case in the “AI for the Factory Floor” category. It was selected because it was of common interest among the semiconductor companies. Workshop Roundtables were developed to address targeted questions that challenged conventional compartmentalization (siloing), dug deeply into how data needed to be processed and qualified for building an AI/ML/DT model to predict wafer flatness from machine tool data during production, and focused on collaboratively processing AI-Ready data from multiple, yet similar machines. Again, this emphasis on scaling data processing, sharing data, and model development as a CDE is framed by the three horizontal blocks in **Figure 1-1**. Consistency with data contextualization, qualification, formatting, the processing of AI-Ready data, and the engineering of training, validation, and test sets was the central focus throughout the Workshop.

Figure 2-2 brings out the foundational finding that data preparation and refinement consistency is best achieved as a workflow of consistent and repeated data processing steps that include (counterclockwise): (1) eliminating contextualization and formatting inconsistencies with a common data information model built as a *collaborative step*, (2) ensuring consistent qualification (operational acceptability) and formatting (including categorization of key distinguishing operational features) as *onsite steps*, (3) ensuring consistency by maximizing pooled data processing as a workflow of *collaborative steps*, and (4) *site validation and deployment* with shared but individually applied solutions and methods. The steps shown in **Figure 2-2** were applied to the Workshop use case to ‘refine’ raw operational data from multiple machines at multiple sites into the AI-Ready data used to build the collaborative AI model used for each machine. The findings showed that maximizing the extensive collaboration (shown in blue) while ensuring minimal inconsistencies from site steps (in black) were key to data consistency. The entrée into the cycle is a Common Data Information Model. **Figure 2-2** is derived from **Figure E-1** and discussed in detail in **Appendix E**.

Figure 2-2 additionally brings out that consistency also means consistently selected and applied methods for each step. CDEs within and across factories can solve common problems, implement common solutions, and share the results for the benefit of all participants, while also enabling coalition approaches to security and the guardrails needed to protect proprietary content. The common data information model provided the key workflow step for deciding the IP sensitivity of each data type being used. It also provided a useful workflow mechanism for masking IP-sensitive product recipe and substrate information. These configuration specifications together with machine type also became important for setting up machine-operation-searchable dataset categories. The consistency of the entire workflow process involving multiple sites forced discussions on cross-site consistency with security approaches.

To speed up and simplify multi-site data storage logistics, the use case involved five etch machine tools that were functionally similar and used to perform similar operations for different products in different factories for one company. A common data information model was worked out for

all five machine tools across all five sites using the CESMII SM Profile^{TM4} to encode the model into a digitally standard form. Data information modeling was additionally demonstrated on Chemical Mechanical Planarization (CMP) machines at three company sites to show data information modeling extended across companies and different machines. Taken together, the five-site use case and the three-company common data information model building demonstrated that cross-company data sharing was readily doable. Wafer production datasets from multiple etch machines were qualified, categorized, prepared, and engineered into AI-Ready data. The use of the aggregated AI-Ready datasets was benchmarked by building a collaborative AI/ML/DT wafer metrology management model to demonstrate how a factory CDE focused on data processing can work.

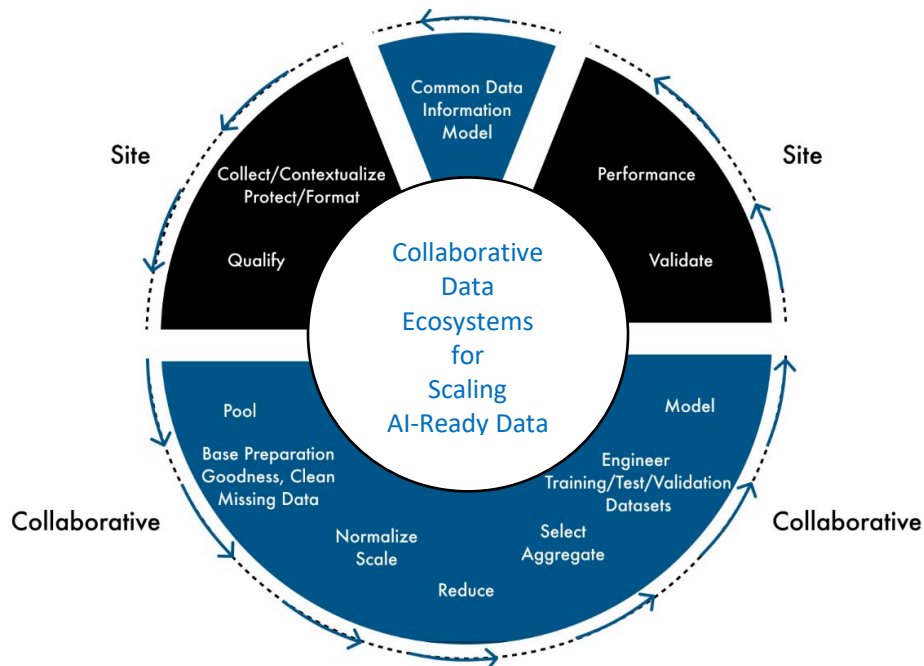


Figure 2-2. Consistent and Collaborative Data Refinement and Model Building (see Figure E-1)

2.2. Benchmarked Findings

Numerous points of benefit to the individual factories/companies acting collaboratively were documented during the Workshop. Ways to overcome barriers to data sharing became evident when executing the workflow steps in **Figure 2-2**. **Table 1** below summarizes the seven key benchmarked findings with descriptions of each listed below. More detailed descriptions can be found in **Appendix E**.

⁴ see <https://www.cesmii.org/technology/sm-profiles/> for further information on the CESMII SM Profiles and the associated Interoperability Platform

Table 1. Key Benchmarked Findings

Benchmarked Ecosystem Benefits		Etch Machine Tool & Wafer Flatness
1	Qualified dataset aggregation for more robust model building	Consistently processed batch run datasets from four machines at different sites aggregated to form a 100,000 batch run super dataset
2	Documented untapped benefits with qualified cross-factory data	Optimizing across operations; staging data exchange; collaborative AI models; analyzing data over time; cross-operation material/product insights; better algorithms; better data methods; data processing automation
3	Data processing staffing and costs	4 FTE headcount avoidance and 3x cost avoidance while processing consistent, validated, and more usable data
4	Model performance with aggregated data	30 % to 50 % better ML model performance compared to using siloed data
5	Data information modeling (CESMII SM Profile™ structure)	Minimum viable infrastructure; common data information model in operational context; achievable across companies; data type, kind, source, and sensitivity at data item-by-data item granularity
6	Additional data information modeling benefits	Cross-site staff engaged; vehicle for data qualification, addressing IP sensitivity, and dataset searchability by process features; machine level of componentization confirmed
7	Governance and data sharing	Trust; methods; site workforce engagement; single source training; data value and ownership; mindset; executive principles

1. **Benchmarked Data Processing Consistency and Qualification:** Data processing consistency proved to be key to data sharing across common machine tool functions, underscoring the importance of a Data-First strategy. Consistent pooling of data from different machines depended heavily on minimizing inconsistencies during site collection, contextualization, qualification, protecting, and formatting. With reference to **Figure 2-2**, data sharing and collaborative processing are enabled by consistent site **collection** (contextualization, qualification, protection, and formatting). Processing then continues with **preparing** (pooling, base preparation, normalizing, and scaling) and **engineering** (reducing, aggregating, and selecting) data to produce AI-Ready-data training, test, and validation datasets for ML **model building** and implementation.

Ensuring consistent qualification of data at the source site made it possible to pool the cross-machine data and scale the processing of all datasets together. By design, the datasets, when processed together, produced a much larger combined dataset consistently processed into available AI-Ready data. In the benchmarking exercise, data from the five tools comprised 40,000, 36,000, 23,000, 2,700, and 800 batch run datasets, respectively (see **Figure C-3**). These datasets were qualified at each site and then shared for collaborative processing. The datasets from the five machines were combined and processed together, producing a combined dataset of 103,000 batch run datasets. Each batch run dataset was processed from about 40,000 time-series sensor measurements over each batch run.

2. Documented Operational Benefits with Using and Processing Data Consistently: Cross-factory and cross-company AI-Ready data consistency, qualification, reuse, validation, and categorization are difficult and expensive to achieve after siloed data has been processed independently. Even though the same categorical steps may be used, variations and inconsistencies in data processing are introduced and compounded at each of the twelve steps in **Figure 2-2**. Usefulness of the data from a given machine is effectively limited to that machine because it is so difficult to reconcile after processing. It is well documented that data collected and processed only for a single application (i.e., siloed) is not reusable or extendable. Also, manufacturers have reported that the cost of data processing for a single application can be as high as 70 % of the total cost. Given that data collection and processing have a significant cost, a strong focus on data processing consistency makes it possible to scale the investment and open economic opportunities beyond a single application in ways that accrue with use in other applications that are just not accessible if data are not processed consistently. The Workshop documented the multiple points of economic value by using and processing consistently developed, ready-to-go, and qualified AI-Ready datasets:

- Optimizing with site, factory, and company cross-machine, cross-operation, cross-factory comparative analyses.
- Staging the data to be exchanged between machines for operational interoperability.
- Building collaborative AI models for multiple AI applications with scaled data reusability.
- Staging the potential for aggregating datasets and analyzing models over time to increase predictive capacity.
- Identifying untapped opportunities for cross-operational insights about materials, recipes, and products, and for synergistically improving ML models and digital twins.
- Identifying untapped opportunities for manufacturers to seek new and better algorithms, models, and solutions with more open data and models sharing.
- Evolving and assimilating new data processing technologies as co-developed methods with greater cross-operation validation.
- Automating data processing workflows together to improve and scale consistency.

3. Benchmarked Data Processing Cost: Staffing costs for processing data consistently across five machines in five different locations were benchmarked. Compared to processing data at each site independently, pooled processing with site-dependent steps could be done (per application) in 1.5 full time equivalent (FTE) data scientist months for all sites vs. 5.5 FTE months if each site acted independently. This is an approximately 3x cost avoidance in staff and a 4 FTE avoidance in increased headcount to process data for five machines in support of one kind of application. When concentrating data processing as a shared activity, a full “CDE” FTE becomes easily justified in supporting more machines, more applications, sustaining applications with updated data and models, updating methods, and developing automated procedures. The cost avoidance takes the form of increasing qualified data availability

without increased staffing. The value of better qualified and processed data availability is readily realized in model performance that meets or exceeds true and false positive rate criteria. It is also realized with the ease and readiness to do cross-operational analyses. The staff avoidance potential is a conservative estimate because it did not account for the savings of an additional data scientist who would have been required to address the more difficult task of reconciling data consistency when interfacing data from multiple sites after the data had been independently processed. These savings on staff effort and headcount for data processing provided significant justification for data processing as a CDE.

However, this shared staffing investment is still a means to an end. The operational advantages documented in Benchmark 2 are the major economic drivers. Acting as a CDE not only makes it possible to have access to qualified and scaled data, but it also makes it possible to significantly lower the cost to achieve the consistency and qualification required for AI-Ready data.

4. **Benchmarked Data Aggregation:** An important economic opportunity that occurs only with scaled, multi-machine datasets is the ability to construct ML models that can use and apply data across a range of machines. This capability enables the application of a foundational model derived from previous production to new applications that are being initiated. For the use case, the associative properties of ML modeling capture the data experiences of all machines together (i.e., increased variation and coverage of the data) to improve the model performance in predicting wafer flatness as a pass or fail. This Scaled Data Sharing involves aggregating consistently processed data from multiple machines into cross-machine training datasets to build a collaboratively developed ML model for all machines together. For the use case, training a common model with cross-machine data consistently improved model performance for all machines. Given baseline performances with individual machine models trained only with data from respective machines, data aggregation consistently lowered false positive rates by 30 % to 50 % while maintaining an acceptable true positive rate.
5. **Benchmarked Data Information Modeling (CESMII SM Profile™ structure)** (see footnote 4): With reference to Benchmark 1, data processing consistency, whether as an individual site or acting as a CDE, depends on consistent and persistent contextualization and formatting for reuse and scaled sharing. Data information models are static descriptions of data with respect to machine function, data type, name, and units as well as how the machine is used in the context of the manufacturing operation. They also describe how data are to be collected and formatted and they facilitate decisions about the different data types at the granularity of each data item collected for modeling purposes. For example, each of the streamed data items measured from machine sensors needed to be understood for further processing, but these data items did not need to be treated as IP sensitive. Data items about substrate materials and product recipes were in the form of static configuration data for a run and were highly IP sensitive. CESMII's tools for structuring and managing data models as SM Profiles™ were validated and established as minimum viable infrastructure for enabling data

information modeling, data decisions, and the data interoperability that are needed for data processing. For this benchmark study, a procedure for building a common, cross-factory/company data information model using CESMII's SM Profile™ structure was established with factory site operators and data engineers. Factory-level data information modeling by site personnel was important in resolving and reaching acceptance on how to describe and structure machine functions and agree on sensor names, data types, units, and collection requirements.

- 6. Benchmarked Factory Staff Acceptance and Use of Data Information Models:** An SM Profile™ is a software structure for capturing a Data Information Model that needs to be filled-in with a description of data from machines/process operations in the context of the specific manufacturing operation in which they are used. With reference to Benchmark 2, once the description is constructed for a device or operation, much of that description can be reused, modified, and/or updated now that it is structured as a data information model. CESMII has a tool called the SM Profile Designer™ for entering data-operational information to build an SM Profile, i.e., data information model that is automatically encoded for software system use with standards-based data formatting and hardware and software interfaces.

For this benchmark study, a procedure for building a common, cross-factory/company data information model was established with factory site operators and data engineers. Two common machine tools were “profiled” as cross-factory and cross-company demonstrations of collaborative data information model building, i.e., Etch and Chemical Mechanical Planarization (CMP) tools. Factory-level SM Profile building by site personnel was important in resolving and reaching acceptance on how to describe and structure machine functions. Site personnel also needed to agree on sensor names, data types, units, and collection requirements. This process of data modeling served to engage the site staff not only in building the data model, but also in developing a common understanding of the data and positioning everyone at all sites to better qualify data in operational context together. The act of building an SM Profile informed how the data needed to be qualified with consistency across multiple sites and suggested ways to automate site requirements. It led to a better shared understanding of the how the machine is being used at multiple sites. It also led to decisions on which data items are IP sensitive and those that are not. For those which were IP sensitive, it proved to be a mechanism for deciding and acting collaboratively on IP sensitivities.

By observation, factory-level benchmarking confirmed that data information models developed for machine tool/unit operations provided a natural “sweet spot” for operations staff in all factories and companies to reach agreement on a common data information model that could be implemented in the CESMII SM Profile™ template. A machine/unit operation proved to be the more natural and useful level of physical componentization for a data information modeling because machine-level components, associated functions, and installed sensors tended to be tightly coupled by machine. Considering subcomponents of machines and building multiple data models was too granular. Considering the machine itself

as the base component proved to be a level at which machine functions and components were sufficiently more coupled relative to interrelations with other machines. It was therefore more natural to build data information models for multi-machine lines from individual machine-level models.

Settling on this level of componentization was needed for a common data information model and it helped address what data was CDE sharable, what data needed to be anonymized as confidential, and how to mask it – key aspects for an IP Protection plan. Site experts keyed on using the naming convention requirement for an SM Profile to mask the configuration specifications needed in the ML model but that were considered IP sensitive. The organizing committee additionally recognized the value of these configuration attributes and how they could be used as dataset categorizations. Machine type, brand and model, and company-confidential designations of materials and product processing were configuration specifications that could be searchable, given the consistent format of the data information model. These attributes made it possible to search and select qualified datasets by operational features.

The CDE cross-factory/cross-company process of building a data information model involved multiple iterations that each participating company undertook independently before bringing the models together, one at a time, to resolve a common model. The process was a facilitated process involving site data scientists and machine tool/process engineers converging on functional descriptions, naming common functions, naming sensors, naming categorizations, and agreeing on data types and units. The exercise demonstrated that common profiles are achievable, and benchmarking showed that the process becomes faster and easier as more sites were added.

7. **Benchmarked CDE Governance:** Executing data processing as a CDE required a commitment to a governance structure that ensured trust in qualification and consistency processes, security, IP protections, and validation. Factory site data engineers and scientists needed to engage in the solutions and work together to reach agreement and train together to qualify data and to build and sustain a common data model. The governance structure is also needed to address and sustain consistency by co-evolving/developing/buying data processing methods that are used in the ecosystem workflow. The orchestration of methods into a workflow that could be automated was called an AI-Ready data processing playbook. Governance was needed to manage the value of the pooled data with the participants providing and using it, and to manage the cost and benefit advantages of the data consistency across the ecosystem. This consistency with data processing that drives standardized approaches also supports better security and IP management. Single-source training on data processing goes hand-in-hand with workforce training on data security and IP protections, the methods for which can be managed better as a CDE. While operational use of the data intersects with data collection and processing, it became useful to govern a data ecosystem separately from the operational use of the data. It was then possible to scale data processing

and the availability of the data while making separate decisions on cross-factory and company operational interoperability and supply chain applications.

Importantly, governance was underpinned with a “mindset” that involved a willingness to challenge conventional thinking, a conviction that sharing information can have benefits for all parties, and a belief that data has potential value as a business asset. Adhering to the eight execution principles in **Table 2** was critical for sustaining the ecosystem effort.

Table 2. Eight Key Execution Principles for Industry Data Sharing

Stay within existing technology
Engage workforce on data and the value of the solution
Commit to contribute data
Build alignment with business motivations
Build from a use case that allowed granular review of data
Converge quickly on minimum viable infrastructure
Plan to share data more broadly than just the coalition
Address the \$ value of the data

2.3. Critical Industry Mindset Factors

To organize the Workshop, Seagate Technology (Seagate) with support from UCLA’s Office of Advanced Research Computing and UCLA/CESMII formed a joint project coalition with twelve Small- and Medium-Sized Manufacturers (SMMs) in semiconductor manufacturing to organize a test SMM Semiconductor CDE, with the participants noted in **Appendix A**. The Workshop was organized to determine if a CDE can realize the value and advantages projected for open data sharing. Worldwide semiconductor revenues reached \$630 billion in revenues in 2024 and are on track to exceed \$700 billion in 2025.⁵ The manufacturing process is based largely on similar machine operations from a small group of equipment suppliers. Participating manufacturers in the CDE represented about 20 % of the industry with more than \$130 billion in combined annual revenues. A potential collective gain of \$28 billion was calculated from projections of yield improvement, labor efficiency, and capital avoidance. The projected economic potential was reviewed by the coalition. The business opportunities combined with significant commonalities increased the motivations for this coalition to form the test CDE. There was an expectation of successfully demonstrating potential for data sharing, aggregation, and scaling.

Led by Seagate Technology, all semiconductor manufacturer participants committed up front to a common goal of an interconnected, governed, and scalable SM ecosystem that leverages Scaled Data Sharing and the cooperative development of an AI/ML/DT software application model. Over the duration of the Workshop, commitments were re-checked to make sure each participating company saw a line of sight to immediate economic improvements in their respective wafer

⁵ See <https://www.semiconductors.org/policies/tax/market-data/?type=post>

production operations without increased staffing and facilities. Throughout the Workshop, the Organizing Committee sought to unite stakeholders across industry, academia, and government to tackle challenges such as fragmented data systems, limited AI adoption, and protection of confidential information.

Sustaining this Workshop and use case required a “mindset” that involved a willingness to challenge conventional thinking, a conviction that sharing information can have benefits for all parties, and a belief that data has potential value as a business asset. Actively establishing this mindset proved to be critical and invaluable. It allowed the collaborative to achieve consensus on numerous considerations that formed key execution pathways for data sharing and the adoption of SM as a CDE, something that has never been put into practice in the semiconductor industry. Active mindset commitments took the form of a collective agreement on the eight execution principles listed in **Table 2**. These principles provided the essential mindset, technology, and business foundations that had to be defined before the coalition could actively engage in a data sharing project. The Workshop was also guided by the overarching principle that **each company needed to see how to realize individual benefit from the formation of a CDE**. Foundationally, the CDE was formed as a business arrangement that benefitted each participating company economically. This shaped the bottom-up, market-driven approach for the CDE.

In the end, most members of the CDE were able to sustain a business mindset for the Workshop aimed at achieving benefits from horizontal scaling of data and capabilities, something for which there has been no documented experience with multiple competitors in a specific industry. The CDE also reviewed and was motivated by an economic analysis that projected the value of comprehensively using SM and AI/ML/DTs to automate a suite of process control systems. Taken together, the Workshop findings detail and benchmark significant industry-defined economic and operational value opportunities, many untapped, with scaled, more open but managed, and cost-effective data sharing. Consistent Data-First strategies in combination with collaborative data processing and management generate opportunities that are not possible when done independently within site, factory, and company silos. The coalition emphasized these primary ecosystem opportunities:

- Far more valuable collaborative, qualified, categorized, and consistently processed dataset inventories that are AI-Ready for all to use for individual implementation.
- Ecosystem datasets that are prepared and staged for cross-machine, cross-site, cross-company, and supply chain applications that involve data exchange.
- Categorized datasets that can be selected, searched, and studied for greater operational insights.
- Individual site datasets that can be aggregated to build more robust AI models for all sites.
- Data science methods and steps, which are nuanced and extensive in processing AI-Ready data, that can be developed, applied, updated, and sustained far more cost- and

functionally-effective ways as an ecosystem, while also making consistently processed data available with shared processes that allow for updating methods.

- The opportunity for an ecosystem to build a sustained qualified data inventory in which data are continually updated, performance data can be shared, data processing methods are maintained and updated, and shared models are continually improved.

The findings also show that an industry data ecosystem can form around and execute on collaborative interests in increasing the value of data. There are substantial benefits with commitments to governance, while co-ensuring necessary guardrails for security and IP protections. There is sufficient industry benefit for ecosystems to form and move forward, but there remains an important need for developing services and scaling data availability to be able to assign quantitative value to qualified data when:

- Contributed and aggregated in a timely manner.
- Used at various levels of processing, in the sense that qualified AI-Ready data is far more valuable than raw data. Operationally categorized machine data is more valuable than machine-only data.
- Used directly and/or in the form of more robust shared applications, models, or profiles.
- Applied to methods and data science that has been tuned to operational interests and can be reused.

The infrastructure and tools to manage, train, and share the data add value to the data. Additionally, there are other kinds of value when data can be selected and shared with other communities and/or shared broadly on the internet. Further work to quantify the value of data is clearly needed. Data ecosystems provide an important framework for doing so.

2.4. A Line of Sight to a Grand Vision at the Scale of the Internet

This Workshop was a bold, rarely conducted, industry study about the need for and ‘how to do’ AI today as an ecosystem. The Workshop united industry leaders, academia, and government agencies in tackling critical challenges of collaboration and data sharing by using a segment of the semiconductor industry as an important industry use case. The Workshop did not seek a consensus from all stakeholders on all points. Instead, the Workshop focused on the collaborative ecosystem with various stakeholders weighing in with rich perspectives on possibilities for the industry to consider when operating with a different mindset. It also succeeded in building a better understanding of manufacturing between industry and academia.

The Workshop delivered on its promise of addressing recommendations from earlier national studies. It also dug deeply into the value of data based on manufacturers’ views of their data ownership. Far from being just another forum for discussion, it delivered tangible milestones, including demonstrating a shared data and model inventory, building necessary training about qualifying, scaling, and processing AI-Ready data for AI application building, ranking key focus

areas for process monitoring and control, and driving needed granular discussions in which different types of data could be shared or protected in ways specific to the data type. Breaking down barriers to collaborative data sharing laid the groundwork for a transformative future with CDEs when data ownership, data interoperability, and operational interoperability are driven as business decisions.

A grand vision came into sharper focus for the organizing committee as the number of economic value points and the benchmark results of a collaborative Data-First strategy with consistent data processing became clear. The Workshop showed that collaborative data processing scales with data from common machines/unit operations, but in discrete steps based on data scientist/engineer capacity relative to the number of applications, kinds of models, number of machine datasets processed, and the amount and timeliness of data. There is huge value in assimilating new data science methods and IT technologies in operational context together, which also maintains consistency. Also, not all manufacturers benefit from a CDE in the same way since the value of data contributed can be different from the value of data used or the value of the validated tools and applications collaboratively developed by the CDE. There are uncharted opportunities for joint work and new service models with IT and OT providers and equipment builders. From a business perspective, there are multiple ways to increase the value of data and how it is used by forming a CDE.

As defined in this Workshop, the CDE is foundationally held together by individual manufacturer business benefits that can be achieved by working together on qualified data, validated methods, and/or models collaboratively developed. While much remains uncharted, a grand vision takes shape around multiple **CDEs producing qualified, consistent, scalable, categorized, and validated data and models that collectively and continuously increase digital prediction, robustness, and fidelity for manufacturers to use at all scales of value.** Importantly, a CDE does not need to be a closed entity. Without added expense or loss of immediate individual site benefits, a CDE is also positioned to open qualified, discoverable, validated data, and models into the grand database of the internet network to drive quality network effects that leverage data and AI at the scale of the internet. Potential with using Large Language Models (LLMs) to find and analyze against qualified datasets, data information models, and validated application models at network scale is analogous to being able to find qualified content to jump start effective code development.

In summary, there is sufficient opportunity, together with acceptable execution paths forward, to recommend factories and companies to identify and build the value points for ecosystems. The business motivation for a CDE exists today. The report lays out important business principles for how and when to share data for individual company benefit. It encourages collaborative workforce training on data processing and how to value data, to be able to act on and build out all facets of data value. The report also provides a path forward for forming a CDE to define and align governance and shared resources to get started. There remains, however, a key recommendation to continue to work on how to value data when there are many facets with contribution, use, openness, methods, tools, consistency and data interoperability.

The report calls out site and collaborative steps needed to address qualified AI-Ready data and solution validation, implementation, and sustainment that are underpinned by the engagement of the factory workforce. There is a line of sight with this CDE model where trust, collaboration, scale, and value of a Data-First strategy are understood and grow. Over time, we can anticipate greater automation for consistency, new service models, etc., and a shift in workforce training that is more oriented toward Scaled Data Sharing value and opportunities. In the short run, however, the recommendation is to much more strongly emphasize on-the-job training about data sharing, data value, and consistent data processing than exist today. The findings for CDE execution strongly encourage concentrating the data-science oriented steps into a collaborative activity with shared staff. On the other hand, operational data modeling, data qualification and selection, data engineering in operational context, and model validation are specific steps that require direct site staff involvement and require training for execution consistency. These direct collaborative and site tasks are nicely addressed with apprenticeship-type programs.

Importantly and in addition, all staff engaged with any aspect of these applications needs to see themselves involved and need to appreciate and trust the steps in data processing even if they are not involved directly. Training on data processing for direct and indirect involvement is therefore important to the success of a CDE and setting the stage for the future. An advantage of a CDE is that training can be done through collaborative programs that span factories and companies. These programs not only ensure consistency with contribution and use of data, but also build trust, staff engagement, and knowledge sharing among CDE participants that are beneficial to each individual factory and company.

This grand vision is still at its core about breaking down silos to enable **seamless data sharing and collaborative algorithm development** across factories, companies, and stakeholders at a much greater and more valuable scale. As described by this semiconductor industry coalition, the benefits of more open data sharing include:

1. **Ecosystem and Industry-Wide Value that Increases Individual Manufacturer Value:** By establishing data sharing frameworks and infrastructure for reusable, consistently processed, and categorized datasets, the industry can collectively optimize common operational needs (e.g., defect detection, root cause analysis, and advanced process control) to reduce costs and improve quality at a much greater scale.
2. **Innovation Through Collaboration:** Facilitating collaboration with more granular data sharing, while respecting confidentiality (via shared vs. coalition vs. private data categories) creates pathways for collective research and development (R&D) that accelerate innovation (i.e., let the experts build better AI/ML/DT algorithms and DTs without jeopardizing competitive advantages).
3. **Data as an Asset for Cost Sharing:** Positioning data as a valuable shared resource means companies can pool resources to build infrastructure, scale services, train AI models, train people, and solve common challenges. Data sharing is a far more efficient and cost-effective way to tackle the scaling and application of manufacturing data.

4. **Scalable AI Integration:** Building minimal viable data infrastructures and demonstration use cases signals a move toward scalable, AI-driven solutions that can adapt across factories and companies, enabling smarter decision making and more preventive, predictive, and proactive insights.

Based on the achievements of this Workshop, the vision for a CDE is achievable. The knowledge gained is also transferable to help other companies and industries develop the fundamental capability required to utilize SM and AI/ML/DT and build CDEs. Simply stated, the future depends on adopting a Data-First strategy with Scaled Data Sharing as the key to capturing the true value of data.

3. Background on Data Sharing

The economic opportunities and barriers have been extensively explained in a series of government-sponsored industry, academic, and government studies developed since 2020. These studies represent a very large base of manufacturing industry contributions, and they have made the case for integration and change. A clear convergence was reported on the technical, business, workforce, and mindset changes that are required to execute on a vision of collaborative and scalable SM industry ecosystems. The three studies are:

1. *National Strategy for Advanced Manufacturing*, Report by the Subcommittee on Advanced Manufacturing Committee on Technology of the National Science and Technology Council, October 2022.
2. *Towards Resilient Manufacturing Ecosystems through Artificial Intelligence – Symposium Report*, NIST Advanced Manufacturing Series, NIST AMS 100-47, September 2022.
3. *Options for a National Plan for Smart Manufacturing*, National Academies of Science, Engineering and Medicine, Consensus Study Report, 2024.

The 2024 National Academies Consensus Study, “*Foundational Research Gaps and Future Directions for Digital Twins*,” aligns with the elements of a collaborative data ecosystem and the critical role of sensors and observing systems, data acquisition, and data integration.

Digitally, SM defines how **Advanced Sensing, Controls, Platforms and Modeling (ASCPM)** are constructed into digital software systems that use data to describe, diagnose, predict, and prescribe actions to control and optimize operations at scale. AI/ML, physics-based modeling, and DTs are dominant forms of modeling in which data and models are used synchronously to continuously learn and determine the best physical operating controls and management actions.

When integrated with the physical operation, energy, material, facilities, and a data-savvy workforce see significantly improved productivity throughout the industry. Assemblies, parts, and materials are made with greater precision while addressing faster, more variable demand dynamics and drawing value from faster changeovers. Factories, supply chains, and ecosystems operate better, faster, and cheaper with greater agility, flexibility, and resilience in supply and demand trade. The workforce works smarter while AI addresses dimensions and volumes of data that exceed human bandwidth. The workforce additionally draws from the internet to engineer data and build models. Increased trust and confidence with data set a course for automation and autonomy to grow and evolve when economically beneficial. Factories can drive operations closer towards having zero incidents. With the wide adoption of SM throughout the industry, energy and material productivity are significantly increased by using less input.

These studies have confirmed that scaling digital and physical integration depends on data interoperability and consistent data processing to achieve operational interoperability. It is the scaling of operational interoperability that results in the ability to: scale the control and management of industry operations; access untapped product quality and value; accelerate

manufacturing innovation; increase energy, material, and workforce productivity; increase supply chain resilience; and increase U.S. market share and global manufacturing leadership. From an AI implementation standpoint, these physical outcomes require:

- Meaningful valuation of data assets.
- Consistently and securely aggregated, exchanged, and/or opened data accessibility.
- Models that learn and have the right data at the right time.
- Shared tools, infrastructure, and software applications that facilitate AI/ML/DT orchestration and business processes together.
- Digital interconnectedness with trust.
- Scaling with network effects and using AI to search and combine resources to do AI/ML/DT.
- Changing an industry legacy mindset from data siloing to data sharing.

Importantly, productivity only matters if there is demand. A networked/interconnected manufacturing industry scales monetization of productivity, precision, and performance beyond Industry 3.0 in three nested layers that use industry-wide adoption to drive product demand based on value and respond to product demand variations better and faster (see the NIST 2022 report, <https://doi.org/10.6028/NIST.AMS.100-47>):

- Asset and Quality Management on the factory floor, which depends on data access and sharing to power ASCPM systems. These factory floor systems with data interoperability, consistent data processing, and data sharing within and across companies and supply chains are where digital and physical actions are orchestrated for scaled impact.
- Line Operation and Supply Chain Interoperability in which operational data are exchanged between machines and operations or combined into higher-level key performance indicators (KPIs) so that systems-of-systems can be controlled and managed as enterprise systems for better performance.
- Supply Chain Resilience in which data describing material and product capabilities, availabilities, and flows, as well as demand variability, are connected, made visible, and analyzed throughout and across supply chains.

3.1. Economic and Energy Productivity Potential Drawn from Data

As summarized in the National Academies Consensus Study on SM, CESMII's and CyManII's⁶ individual factory and supply chain industry demonstrations and studies have shown a 5 % to 30 % economic benefit across all industry segments (or 15 % on average) for targeted smart applications. Benefits are economic but, depending on application and industry segment, they are based in various combinations, on higher sales productivity, increased margins, cost

⁶ CyManII (Cybersecurity Manufacturing Innovation Institute): see <https://cymanii.org>

avoidance, and/or energy/materials savings. In nearly all cases, energy productivity increases, and energy usage is reduced by approximately 7 % on average. Supply chain benefits generate an additional economic benefit of 15 % to 20 %. Wide industry adoption of these individual applications is key to substantial national and global impact. These economic and energy productivity effects are additive when multiple applications are orchestrated at factory, company, and supply chain scales.

While economic benefit with increased productivity, flexibility, and agility, and the reinvestment potential of using less energy and material are high, there remains little scaling of SM throughout the industry, largely because scaled, shared, and qualified data requirements have not been addressed. A Data-First strategy becomes critical with scale and when real-time manufacturing data is needed at all levels of ASCPM system solutions. The orchestration and scaling of these systems require AI-Ready data available across applications in the amounts, forms, scale, and access necessary to achieve benefits in productivity, jobs, market share, economic gains, and growth. In this study, a Data-First strategy that emphasizes data processing consistency is key to Scaled Data Sharing. Consistency is required with **collecting** (contextualization, qualification/categorization, protecting, and formatting), **preparing** (pooling, base preparation, normalizing, and scaling), and **engineering** (reducing, aggregating, and selecting) data to produce AI-Ready data training, test, and validation datasets for AI/ML/DT model building and implementation. These are qualified operational datasets that reflect years of experience, even when the physics may not be fully understood.

A Data-First strategy addresses the importance of ensuring data contextualization, qualification, preparation, and engineering are consistent and persistent, as the data is used at different scales and in different models and applications. As data are qualified for operational and measurement integrity, it is important too that the data is categorized for operational selectability with key distinguishing features such as machine, product, material, etc. Data-First strategies also address the ability to implement affordable AI/ML/DT solutions, including how to obtain, categorize, scale, use, and reuse AI-Ready data. This includes how to validate, learn, unlearn, transfer learn, and adjust models when using and reusing this invaluable data.

Robust AI/ML models depend on and thrive on “good” data. When applied in control, automation, and autonomous systems, data must be trustworthy. There must be adequate amounts of the right data, the system performance must be verified and validated, and uncertainties must be quantified. The long-held practice of data siloing is the most costly and constraining practice to be addressed. When data are constrained to a machine or operation, the resulting model is limited. Siloed data are not transferable to anything else, and the expense of acquiring and preparing the data is a lost investment because the data is not easily scaled or reused. Siloing within factories, line operations, and supply chains limits the data and locks out the advantages of scale. Siloing among manufacturers, equipment builders, and solution providers inhibits manufacturer access to their own data and inhibits equipment builders and solution providers from optimizing their products, generalizing tools, and aligning measurement and data-generation capabilities. Restricting access to data shortchanges all stakeholders,

undercuts the value of data, and minimizes the potential of what data can tell us about a process, equipment, and product. Of course, data related to IP and trade secrets must be protected and managed, but it also needs to be positioned for negotiation based on a clear value of the data. A Data-First strategy addresses these issues as implementation integration requirements.

In essence, a CDE seeks to transform existing data scaling challenges into shared opportunities, offering a faster, cheaper, and lower risk path to operational excellence and innovation. The SMM semiconductor CDE benchmarked in the Workshop listed the affinities below that brought participants together and motivated consideration of data sharing “**to lift all boats**” to the benefit of each site individually:

- SMMs can contribute to and benefit from resources they cannot afford individually.
- There are large advantages in collectively addressing common, non-proprietary challenges such as noisy, missing, insufficient, or non-varied data.
- Data sharing ensures qualified and categorized data and validated AI solutions that reflect real world complexities, improving performance across the board.
- Collaboration creates a force multiplier effect where collective intelligence drives exponential improvements in the application of SM and AI technologies.

As the Workshop progressed and as security and IP protections were discussed, there was the realization that the data processing consistency that drives standardized and governed approaches also supports greater security and IP protection:

- Security is far better addressed and managed with a standardized and consistently applied data management infrastructure for collecting, contextualizing, qualifying, formatting, selecting, sharing and exchanging data. This is compared to the building and maintaining an ever increasing, unscalable number of one-off data exchanges and interfaces with non-standard and/or proprietary data management conventions, interfaces, and security approaches. The security standards established by NIST are critical for enhancing data security practices, but so is balancing the use of cloud and on-premise data management. Training the workforce in data processing and application emphasized in this report is the opportune time to also train on cybersecurity and especially on cyber resilience.⁷
- IP protections can be worked out with granular evaluation and agreements on different types of data. Much data can be readily shared. It is how the different types of data are combined or merged that become sensitive. Data models, data processing, and governance can be used to manage IP protections while supporting data sharing with trust.

⁷ Wilkes, M. and J. Davis, “Cybersecurity Resilience for Advanced Manufacturing,” *State of US Manufacturing*, US Center for Advanced Manufacturing, (2024), https://issuu.com/automationalley/docs/uscam_digital_book_statemfg2024_final?fr=xKAE9_zU1NQ.

4. Semiconductor Industry Opportunity

This NSF/NIST workshop described in detail in **Appendix B** was organized to address the projected opportunities and known headwinds with data sharing in a format that a semiconductor industry coalition could quantitatively and qualitatively test requirements, capacity, and capability to maximize value. Data sharing begins with a value proposition and defining the affinities that provide the starting motivation. Typical of industry in general, this study began with manufacturing participants strongly concerned with sharing data with any stakeholder for IP, security, and regulatory reasons. The potential of academic support had been effectively unattainable without working data. Equipment builders were equally concerned about sharing data and, like IT/OT providers, wanted to develop their own business models around aggregating data as a service. However, the manufacturers believed in and wanted to understand the full range of opportunity and the value of their data to their respective bottom lines. There had not been any concerted discussion on business capacity or infrastructure to address siloing and to develop and broaden the valuation of data. The details described below for this study, as viewed by this coalition of manufacturers, are provided as one example of the economies and affinities around which a CDE could form.

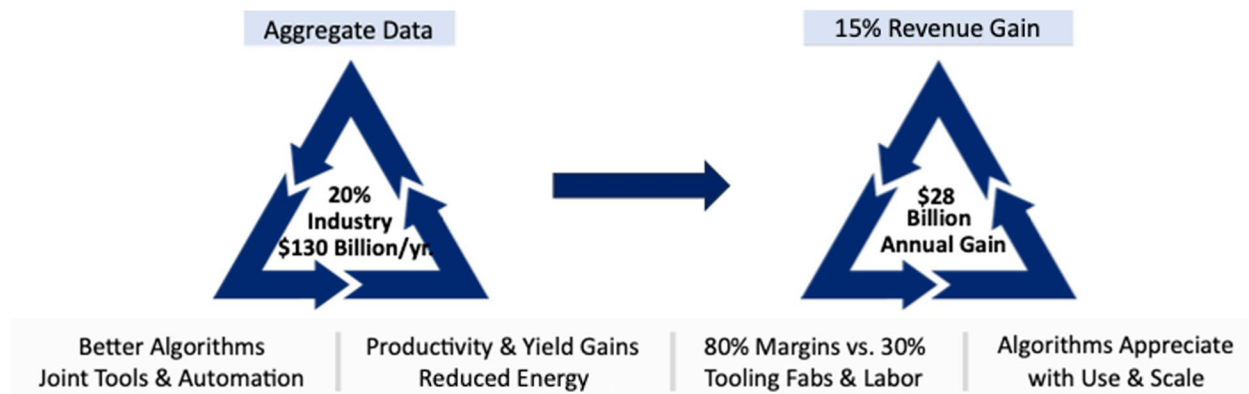
The global semiconductor industry generates an annual revenue of more than \$600 billion, and virtually every innovation of the future will depend on technological advances in this industry. SM and AI/ML/DTs will be essential tools in the development and production of advanced chip technologies. Importantly, existing chip technologies and operations remain vital to the industry with considerable economic headroom to increase manufacturing productivity, precision, and performance with SM. In closing the circle, the semiconductor industry will depend on its own products to address the global economics of materials, energy, facilities, and staffing to increase performance, improve operations, ensure precision and quality, and drive the assimilation of new technologies. The adoption of SM and AI/ML/DT needs to be driven by data, Scaled Data Sharing, and intercompany collaborations, which will enable the semiconductor industry to reshape its fundamental business models to become networked and interconnected.

In the U.S., the application of SM and AI/ML/DT technologies is just entering the early adoption phase. Most companies are pursuing digital transformation with conventional approaches based on incremental, risk-averse positions, and legacy business practices that are not compatible with interconnected data sharing and rapidly expanding IT capabilities. Overcoming these entrenched practices will require early industry demonstrations of SM and AI/ML/DT adoptions that show the financial benefits of data sharing, and that collaboration can produce greater results than those achieved with conventional siloed and data protection approaches.

The largest semiconductor manufacturers such as Taiwan Semiconductor Manufacturing Company (TSMC), Samsung, NVIDIA, and Intel, are aggressively pursuing advanced chip technologies, and SM and AI/ML/DT are expected to be essential tools in these development efforts. While large companies have the resources to utilize SM and AI/ML/DT in advanced manufacturing development, SMMs have limited resources, staffing, and ability to integrate

these technologies into their semiconductor fabs. The integration of SM is disjointed and exacerbated by cost constraints. However, SMMs collectively face common issues in advanced process control, creating an opportune environment for an industry-wide initiative to leverage economies of scale. AI/ML/DT applications are equally valuable to the large base of SMMs, and these companies are a key source of data. AI/ML/DT applications constructed from industry data are also expected to be essential tools across all manufacturers and supply chains.

Motivated by resource limitations and a belief in the collaborative value of SM, AI/ML/DT, and data, Seagate Technology led an effort to organize a group of twelve fab operators into an SMM CDE. As a CDE, the participants represent about 20 % of the industry and more than \$130 billion in annual revenue. The CDE’s initial analysis of using SM, AI/ML/DT, and data sharing to automate process control systems indicates a collective potential gain of \$28 billion from yield improvement, labor efficiency, and capital avoidance. This is illustrated in **Figure 4-1**. Details can be found in **Appendix B**.



\$28 Billion in Capital Avoidance, Labor Efficiency, and Yield Improvement

Figure 4-1. Quantified Potential of SM and AI/ML/DTs Developed by and for the CDE

Despite the potential value of SM and AI/ML/DT adoption in manufacturing, and the emphasis today on the value of real-time manufacturing data as the fuel that powers AI/ML/DT, the NSF Workshop asked **why are we still paying so little attention to the data?** Overemphasis on algorithms without the data is to implement AI/ML/DT with far less than half of what is needed. The value and role of AI/ML/DT in manufacturing operations is to take advantage of excellent product operations resulting from years of experience with machines, operations, and behaviors embedded in current data at the level of fidelity needed to optimize products and operations. This data can also be used to gain insights and build better models and DTs with a systematic way of applying it. However, far too many AI/ML/DT solutions continue to be “engineered” without engineering the data, a trap long proven faulty. Domain physics and operating experience are important to modeling, but equally important is to avoid “over-engineering” by making the model fit only what is known. How to validate, understand, and ensure performance remains

somewhat of an art. Few successful applications address how to get the right coverage, variation, and contribution from data. SM expects AI-Ready data at scale.

The semiconductor industry is a relevant use case for taking advantage of data aggregation in that the industry uses the same machine tools from a small group of equipment builders for similar functions. Products are differentiated in the wafer design and by the materials that are used in the machine processes.

5. Framework for Forming a CDE

The business justification and data affinities that have been described were sufficient for motivating the interest in a cross-company CDE, but insufficient for proceeding to form the CDE and begin acting together. Participants needed to agree on specific organizing principles for staying together, not only to define acceptable execution pathways for data sharing, but also to see the development of an action plan to implement SM and AI/ML/DT in a high value use case. It also became important to move past discussions about the use case and to demonstrate the results that are discussed in subsequent sections of this report. The eight key organizing principles, previously listed in **Table 2** and described in greater detail below, became the centerpiece of agreements and governance for the CDE to take action. The manufacturers agreed to use these as a general framework for forming the CDE:

1. **Assume that existing technology is available to support initial adoption** of SM and AI/ML/DT in the existing SMM semiconductor fabs. Forming a CDE is not an R&D exercise. Demonstrations need to be focused on the use of existing technology under the assumption the technologies are sufficiently production-ready for near-term benefit.
2. **First understand the value of data and the AI/ML/DT as manufacturers** by involving data engineers/scientists from industry, academia, and government. Vendor involvement and their revenue models will be dealt with after the value proposition is understood.
3. **Start by making a commitment to contribute data** as a requirement for company participation. Participants agreed upfront that harvesting the value of currently unused operational data could potentially provide a fast and low-cost pathway to near-term benefit.
4. **Build alignment across business motivations.** Be sure to include benefits beyond production cost savings. These include increased opportunities to compete for local, state, and national investments, accessing the expertise of AI researchers, and aligning investments in training and change management across companies.
5. **Converge on a consistent strategy by presenting a concrete use case** that specifies data integration processes, responsibilities, and workflow. All data types should be acknowledged and listed so that sensitivities and responsibilities can be addressed. A data information model needs to be created to provide a common nomenclature for contextualization and data processing.
6. **Establish the minimum viable infrastructure** needed to reliably collect and contextualize data within and across companies. By quickly deciding on the CESMII SM Profile™ oriented infrastructure with the use case, data processing, aggregation, engineering, modeling, and all associated implementation and integration requirements could be specified to focus the discussion on collaboration and get beyond infrastructure.
7. **Plan beyond commercial sharing of data and models** to make production-relevant data available to researchers with the expertise needed to innovate on ML architectures appropriate for categorizing manufacturing data types.
8. **Quantitatively address the value of data** and the total cost of data use when integrated into factory processes. Value realization depends on manufacturers understanding the

integration of their own data within and across their respective factories and relating it to their products. This is the first step toward business models with equipment builders, vendors, providers, and non-commercial communities.

Overall, these principles became the initial agreements and first demonstrations of CDE governance in the development of an action plan to scale the implementation of SM and AI/ML/DT technologies. Equally important, however, was agreeing on a use case. Early in the Workshop the semiconductor industry participants generated a list of priority control and operational application areas collectively ranked by the companies. These are listed in **Table 3** below. Different manufacturers had different operational priorities and were at different levels of maturity with digitalization. An early governance decision was to settle on virtual metrology as an industry use case of common interest. A virtual metrology ML model constructed from the operating data of etch machine tools, configuration data that captures the product being made and the material used, and metrology data corresponding to individual batch runs was agreed to for Workshop demonstrations and benchmarking. The use case became the centerpiece for working out the specifics of forming the CDE. See **Appendix C** for details.

Table 3. Rank Ordered Priority Application Areas

Advanced Process Control
Fault Detection and Control
Root Cause Analysis
<i>Virtual Metrology (the Workshop Use Case)</i>
Automated Inspection
Energy Management
Materials Management

6. Workshop Conclusions

The overall goal of this Workshop was to build a first-of-its-kind Collaborative Data Ecosystem (CDE) with twelve semiconductor SMMs as participants. The purpose of the CDE was to implement a Data-First strategy for Scaled Data Sharing to power SM and AI/ML/DT systems in semiconductor wafer production, and use actual factory data across multiple machines, factories, and companies to benchmark the performance of these advanced software systems against current best practices. Few, if any, such efforts have resulted in a successful, practical implementation that has been described in the detail of this report.

In general, this Workshop was a bold and rare initiative intended to demonstrate “how to do” SM and AI/ML/DT as an ecosystem to the advantage of a modern manufacturing operation. This effort united industry leaders, academia, and government agencies in tackling the critical challenges of collaboration, data sharing, and a Data-First strategy by using a segment of the semiconductor industry as an important industry demonstration. These efforts developed a procedure to attract a coalition of SMMs, build a CDE committed to data sharing, select a use case in a specific production machine and process, construct AI/ML/DT systems for the machine and process, and use benchmarking to qualify the value and benefits of the project. In the end, the project involved five machines in three different factory locations and individual machines in three different companies. All of this was accomplished by working with the industrial participants in the CDE using existing workforce and known technologies.

The Workshop delivered on its promise of demonstrating and benchmarking recommendations from earlier national studies. Far from being just another forum for discussions, it delivered tangible milestones, including securing funding for demonstrating a shared data and model inventory, ranking convergence on key focus areas for process monitoring and control, and driving needed granular discussions in which different types of data could be shared or protected in ways specific to the data type. Breaking barriers to data sharing laid the groundwork for a transformative future with CDEs. In addition, several overarching lessons were learned in this project:

- The most important lesson is the significance of data in the adoption of advanced software systems. From data capture through data consistency, data aggregation, data contextualization to data sharing, every success in the demonstration project was driven by the availability of consistently processed data. This point also highlights the importance of a Data-First strategy in future business plans for each manufacturer. Without data as a strategic element, companies will continue to manage data as a second-class asset.
- Contrary to many opinions, SM and AI/ML/DT systems can be implemented in a cost-effective manner and provide immediate benefits to a manufacturing operation. Existing technology with which to proceed exists. Focusing on the data processing and engineering first facilitated the algorithm development and validation.

- Focus is essential for the successful implementation of an application. It is important to capitalize on those data resources that can be controlled and to not get distracted with tools or services too early in a decision-making process. Data information modeling provided the vehicle to sort through available data at a granular, data item level. Decisions about IP sensitive data can also be made and approaches for data items are decided.
- For a successful implementation, the representatives from the industry involved in the demonstration project needed to lead the effort. Manufacturers need to see how to realize individual benefits from participation in a CDE. They needed to understand the value of their own data and how value could be increased with collaboration.
- The successful implementation of an application also depends on engaging the staff in working out the site qualification of data, understanding and trusting the steps taken to process the data, and being involved in the engineering and selection of the data for building, validating, and implementing applications.

As the Workshop evolved through the series of five Roundtables, a Grand Vision came into focus for the CDE as the benefits and value of a Data-First strategy became clear. Looking beyond the semiconductor industry, there are multiple data contributions, use and processing economies, and/or ways to increase the value of data around which a CDE might form. CDEs are business-driven entities about providing individual benefits by working together on outcomes from the data. Many business revenue and service models have yet to test the formation of a CDE. The Grand Vision, therefore, took shape around the concept of many collaborative ecosystems producing qualified, consistent, scalable, categorized, and validated data and models that collectively and continuously increase digital prediction, robustness, and fidelity for manufacturers to use at all scales of value.

At its core, this Grand Vision is about breaking down silos to enable seamless data sharing and algorithm collaboration across companies and stakeholders at a much greater scale. As written by the industry participants, the broader benefits included:

- **Ecosystem and Industry-Wide Value that Increases Individual Manufacturer Value:** By establishing data sharing frameworks and an infrastructure for reusable, consistently processed, and categorized datasets, the industry can collectively optimize operations (e.g., defect detection, root cause analysis, and advanced process control) to reduce costs and improve quality at a much greater scale.
- **Innovation Through Collaboration:** Facilitating collaboration with more granular data sharing and protections, while respecting confidentiality (via shared vs. coalition vs. private data categories) creates pathways for collective R&D that accelerate innovation (i.e., let the experts build better AI/ML/DT algorithms and DTs without jeopardizing competitive advantages).
- **Data as an Asset for Cost Sharing:** Positioning data as a valuable shared resource means companies can pool resources to build the infrastructure, scale services, train AI models,

train people, and solve common challenges. Data sharing is a far more efficient and cost-effective way to tackle the scaling and application of manufacturing data.

- **Scalable AI Integration:** Building minimal viable data infrastructures and demonstration use cases signals a move toward scalable, AI/ML/DT-driven solutions that can adapt across factories and companies, driving smarter decision-making and more preventive, predictive, and proactive insights.

Ultimately, the vision for a CDE stems from a set of tangible practices and execution steps for breaking down silos to achieve seamless Scaled Data Sharing and algorithm collaboration across companies and stakeholders. As conceived and tested in this study, the CDE is a bottom-up, business-focused entity that drives the value of its data. It is market-driven and opens new business, revenue, and service opportunities based on the value of data and on the advantages of preparing data and building models together. Acceptable business pathways grew from these principles. Based on this Workshop:

1. The vision for a CDE is achievable. The knowledge gained is also transferable to help other companies and industries develop the fundamental capabilities required to utilize SM and AI/ML/DT and build CDEs. Simply said, the future depends on each manufacturer adopting a Data-First strategy and the CDE facilitating Scaled Data Sharing as the keys to capturing the full value of data. With this approach, the Grand Vision for manufacturing and AI at the scale of the internet becomes a possibility as a market-driven approach.
2. There is a clear recommendation for factories and companies to identify and build the ways to collaboratively increase the value of data and economies for an ecosystem. The business motivation for a CDE exists now. This report provides a path forward for forming a CDE to define and align governance and shared resources to get started. This report articulates and benchmarks significant industry-defined economic and operational value opportunities, many untapped, with scaled, more open but managed, and cost-effective data sharing to encourage factory and company consideration.
3. The recommendation is for this Workshop CDE to take these first findings to a next level of integration and for other potential CDEs to evaluate the model developed from this experience. The report further emphasizes more open data sharing and collaboration to scale, expand, and accelerate the data technologies and tools for AI, ML and DT solutions. It lays out important business principles for how and when to share data for individual company benefit and encourages collaborative workforce training on data processing and how to value data to be able to act on and build out all facets of data value.
4. In the short run, the recommendation is to expand and more strongly emphasize on-the-job training on data sharing, data value, and consistent data processing than exists today. Underpinning a successful CDE that maximizes benefits for all is the engagement and involvement of the factory workforce. There is a line of sight with this CDE model where trust, collaboration, scale, and value of a Data-First strategy are understood and grow.

We can anticipate greater automation for consistency, new service models, better security and IP protections, etc., and a shift in workforce training that is more strongly oriented toward data sharing value and opportunities.

5. The findings for CDE execution strongly encourage concentrating the data science-oriented steps needed in data processing into a collaborative activity with shared staff. Training all staff to do data processing and application building does not address consistency and is not cost effective. Operational data modeling, data qualification and selection, data engineering in an operational context, and model validation are specific steps that require direct staff involvement and require training for execution consistency. These direct tasks are nicely addressed with apprenticeship-type programs. To do so also sets the stage with staff for further automating tasks and understanding security and IP protection methods associated with data processing and a Data-First strategy.

Importantly and in addition, all staff engaged with any aspect of these applications need to see themselves involved. All need to appreciate and trust the steps in data processing even if not doing them directly. Training on data processing with direct and indirect involvement is therefore important to the success of a CDE. An advantage of a CDE is that training can be done through collaborative programs that span factories and companies. These programs not only ensure consistency with contribution and use of data, but also build trust, staff engagement, and knowledge sharing among CDE participants that are beneficial to each individual factory and company.

6. There is a key recommendation to continue work on how to value data quantitatively. This report has laid out key considerations and elements for putting a dollar value on data. The model and/or metrics needed have not been resolved, but a CDE provides a structure with which agreements can be made to suggest, test, and evaluate different cost models.

Appendix A. NSF/NIST Workshop 2023-2025 Participants

Semiconductor Industry Participants

Nabil Alali

Vice President of Fab and Filter Operations
Skyworks Solutions Inc.

Rajesh Appat

Vice President of Business and Technology
Development
Polar Semiconductor

Dan Baseman

Senior Process Engineering Manager
Honeywell

James Bird

Data Scientist
NXP Semiconductors

Brian Coss

FDC Technical Leader/Senior Engineer
Broadcom Inc.

Gregg Damminga

Vice President of Foundry Services
Skywater Technology

Gerry Edwards

Managing Director, Wilmington Operations
Analog Devices

Chris Gillman

Senior Director of Manufacturing
Operations and Automation
Global Foundries

Steve Haumersen

Senior IT Manager, MLOps, AI Research,
GenAI
Seagate Technology

Thomas Huang

Researcher
Seagate Technology

Surya Iyer

President and COO
Polar Semiconductor

Matthew Johnson

Senior Vice President, Wafer FAB
Operations and Recording Head
Development
Seagate Technology

Kumar Karuppana

CHD-FAB Director of Engineering
NXP Semiconductors

Dan Koch

MEMS Operations Leader
Collins Aerospace

Jill Landeis

Senior Manager, IT
Analog Devices

Fu-Yuan "Gavin" Lee

Researcher
Winbond Electronics Corp.

Steve Lenertz

Director, Global Strategy
Collins Aerospace

Dan Malinaric

Corporate Vice President of Fab 4
Operations
Microchip Technology Inc.

Smitha Mathews

Senior Manager
Analog Devices

Kenneth McAvey

Vice President and General Manager, Fab 9
Global Foundries

Rich Molden

Information Technology
Polar Semiconductor

Sunil Narayanan

Senior Director of Applied Intelligence
Solutions
Global Foundries

Evan Ngo

Data Analytics Manager
Polar Semiconductor

Carrie Pelton

Vice President, Wireless Semiconductor
Division
Broadcom Inc.

Derek Poirier

Senior Manager, IT Site Lead
Analog

Tamas Reiter

Senior Data Scientist
Seagate Technology

Amin Shameli

Senior Director of Engineering, Data Science
and AI-Enablement
Skyworks Solutions Inc.

Gary Stinson

Senior Yield Manager
Microchip Technology Inc.

Belay Tumebo

Researcher
Seagate Technology

Juan Velasquez

Senior Director of Industrial and Systems
Engineering
Skyworks Inc

Chao Zhang

Data Scientist
Seagate Technology

University Participants

Elias Bareinboim

Associate Professor, Department of
Computer Science and Director, Causal
Artificial Intelligence (CausalAI) Laboratory
Columbia University

Dragan Djurdjanovic

Professor, Accenture Endowed
Professorship in Manufacturing Systems
Engineering
University of Texas at Austin

George Kesidis

Professor, Computer Science and
Engineering
Pennsylvania State University

Xiaoning "Sarah" Jin

Associate Professor, Mechanical and
Industrial Engineering
Northeastern University

Murat Kocaoglu

Assistant Professor, Department of
Computer Science
Johns Hopkins University

Yin Li

Assistant Professor, Biostatistics & Medical Informatics, Computer Sciences
University of Wisconsin-Madison

Bruno Ribeiro

Associate Professor, Computer Science
Purdue University

Chenhui Shao

Associate Professor, Mechanical Engineering
University of Michigan

Peng “Edward” Wang

Associate Professor, Mechanical and Aerospace Engineering
Case Western Reserve University

Kun Zhang

Associate Professor
Carnegie Mellon University and Mohamed Bin Zayed University of Artificial Intelligence

Catherine Qi Zhao

Associate Professor, Dean’s Fellow
Department of Computer Science & Engineering
University of Minnesota

UCLA/CESMII

Haresh Malkani

Chief Technology Officer
CESMII

Olivia Morales

Senior Solutions Architect
CESMII

Targeted Perspectives

Jian Cao

Associate Vice President for Research
Cardiss Collins Professor of Mechanical Engineering
Director, Northwestern Initiative for Manufacturing Science and Innovation (NIMSI)
Northwestern University

Yu Deng

Principal Research Scientist, Manager
IBM

John Hajdukiewicz

Managing Director
Outcome Design Labs

Johan de Kleer

Distinguished Scientist
Intelligent Systems Laboratory, PARC, SRI

Gordon Shao

Computer Scientist, Life Cycle Engineering Group
National Institute of Standards and Technology

Mark da Silva

Senior Director, SMART Manufacturing
SEMI

Binil Starly

School Director and Professor, School of Manufacturing Systems and Networks
Arizona State University

Benchmarking Team

Andrew Browning

Manager, Research Data and Web
Platforms
Office of Advanced Research Computing
UCLA

Panagiotis Christofides

Professor & Chair
Chemical and Biomolecular Engineering
UCLA

Steve Haumersen

Senior IT Manager, MLOps, AI Research,
GenAI
Seagate Technology

Feiyang Ou

Graduate Student
Chemical and Biomolecular Engineering
UCLA

Julius Suherman

Graduate Student
Chemical and Biomolecular Engineering
UCLA

Mathew Tom

Graduate Student
Chemical and Biomolecular Engineering
UCLA

Henrik Wang

Graduate Student
Chemical and Biomolecular Engineering
UCLA

Hayk Zakaryan

Web Application Developer
Office of Advanced Research Computing
UCLA

Chao Zhang

Data Scientist
Seagate Technology

Program Sponsors

Linkan Bian

Program Director, Advanced Manufacturing
(AM)
National Science Foundation

Satish Bukkapatnam

Program Director, Division of Civil,
Mechanical and Manufacturing Innovation
(CMMI)
National Science Foundation

Sylvia Spengler

Program Director, Division of Information
and Intelligent System (IIS)
National Science Foundation

Reha Uzsoy

Program Director, Division of Civil,
Mechanical and Manufacturing Innovation
(CMMI)
National Science Foundation

Organizing Committee

Sthitie Bom, Co-Chair

Vice President
Seagate Technology

Jim Davis, Co-Chair

Vice Provost Emeritus
Special Advisor on Smart Manufacturing
and Data Science
UCLA

Said Jahanmir

Assistant Director for Federal Partnerships
National Institute of Standards and
Technology

Bruce Kramer

NIST Associate
National Institute of Standards and
Technology

Don Ufford

Advanced Manufacturing Policy Fellow,
Advanced Manufacturing National Program
Office
National Institute of Standards and
Technology

Greg Vogl

Mechanical Engineer, Production Systems
Group
National Institute of Standards and
Technology

Workshop Writers and Contributors

Andrew Browning

Manager, Research Data and Web
Platforms
Office of Advanced Research Computing
UCLA

Dave Dorheim

Lead Writer
DWD Advisors

Shao-Ching Huang

Computational Research Scientist,
Office of Advanced Research Computing
UCLA

Tajendra Singh

Computational Scientist,
Office of Advanced Research Computing
UCLA

Barbara Woltag

Communications and Program
Management
Office of Advanced Research Computing
UCLA

Appendix B. Integrating the Semiconductor Use Case and Workshop Roundtables

B.1. Designed to Build an Industry Ecosystem and Benchmark the Outcomes

The National Science Foundation (NSF) sponsored, and the National Institute of Standards and Technology (NIST) supported, this Workshop (NSF award 2334590), **“Artificial Intelligence Development with Open and Scaled Data Sharing in the Semiconductor Industry,”** to define, demonstrate, benchmark, and generalize a manufacturing business strategy for scaling the value and use of data. Referred to as a **CDE**, a cross-factory, cross-company collaborative was organized to test how to scale the benefits of data sharing by focusing on: (1) a Data-First strategy for qualifying and processing AI-Ready data as a collaborative, (2) using site-qualified and pooled data to build validated AI/ML/DT models that are ready for site implementation, and (3) acting together to provide a line of sight to sustaining a continuous learning and model performance improvement cycle. The Workshop was organized as a practical business and technology management, governance, and execution demonstration on how to implement industrial data sharing. The Workshop was chaired by Sthitie Bom at Seagate Technology and Jim Davis at UCLA, with Bruce Kramer (NIST, formerly NSF), Said Jahanmir (NIST), Greg Vogl (NIST), and Don Ufford (NIST) as active organizing committee members involved in planning and conducting the Roundtables, drawing observations, interpreting participant input, and reviewing benchmark studies. Moreover, the organizing committee was key to encouraging a data sharing mindset, ensuring separation from legacy practices that had become barriers, facilitating the consideration of modern opportunities, and keeping these aligned against moving targets on data processing, testing, and benchmarking.

To organize the Workshop, Seagate Technology (Seagate), supported by UCLA/CESMII,⁸ formed a joint project coalition with 12 small- and medium-sized manufacturers (SMMs) in semiconductor wafer manufacturing to create the semiconductor CDE. The CDE reached sufficient agreement on numerous technical and business considerations that defined execution pathways for Scaled Data Sharing and the collaborative adoption of SM and AI/ML/DT technology for near-term economic improvements in respective wafer production operations. The credibility of the AI/ML use case demonstration drew from the collective interest and commitment of the CDE to develop an implementation plan and a collective willingness to openly report and share Workshop outcomes from an industry demonstration.

B.2. A Unique Workshop on Business and Technical Execution with Data

Overall, the Workshop broke several barriers to the adoption of SM and AI/ML/DT technologies. This report includes discussions, experiences, collaborative agreements, and benchmarked

⁸ CESMII (Collaborative Ecosystems for Smart Manufacturing Innovation Institute) sponsored by the DOE as the 9th national Manufacturing USA Institute in 2017, see <https://www.cesmii.org>

demonstrations specific to the semiconductor manufacturers, but these experiences and outcomes are also fully open for consideration in other industry segments. These have been included and where possible generalized so that other manufacturing industry segments can relate to the many facets of the experience.

Regarding the use case, **Figure B-1** has been widely used to illustrate a typical semiconductor manufacturing operational line. Cooperative engagement on SM and AI/ML/DT solutions for common operational functions was seen by the SMMs as a pathway to increased productivity by working together to draw value from current data in successful operations, derive new insights on factory operations, attract investment for industry specific challenges, and ultimately lower costs. This test CDE focused on a new business model using cross-company data aggregation and algorithm building for mutual competitive benefit. This mindset produced roundtable discussions on the potential benefits of SM and AI/ML/DT systems with investments in “smart” data assets that have fundamental advantages over physical assets such as equipment, machines, and tooling that depreciate over time. SM and AI/ML/DT systems, data aggregation, and algorithms appreciate as usage expands, and once implemented only require nominal investments for scaling and reuse.

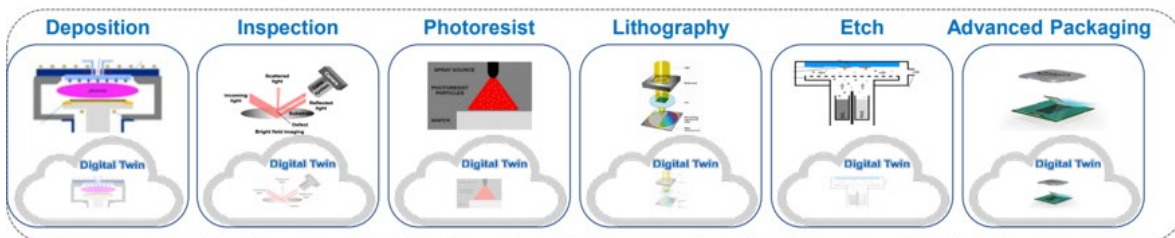


Figure B-1. Semiconductor Manufacturing Process Flow

Figure B-2 shows the specific line operation used in the benchmark use case. In assessing production opportunities for initial SM and AI/ML/DT implementation, roundtable participants ultimately focused on the etching process and the potential to use etch machine data to predict and control wafer thickness. This benchmark focus also has a natural extension to wafer flatness monitoring and automated root cause analysis and other AI/ML/DT solutions of interest. With successful demonstrations of virtual metrology for the etch process, the CDE anticipated that the use of SM and AI/ML/DT can expand across other operations.

Etching was selected as the initial use case because it shares common equipment, processes, and process controls among the companies in the CDE. More specifically, the baseline use case was to predict metrology-measured flatness as pass or fail from etch tool batch run data. This is discussed in detail below.

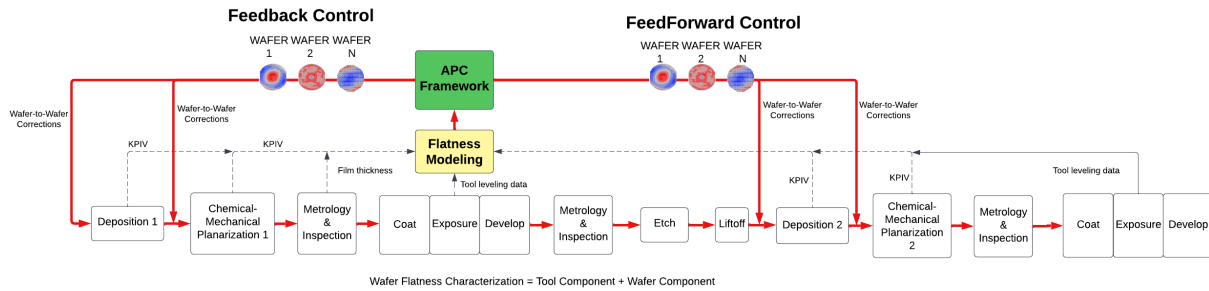


Figure B-2. Factory Line Operation Used for Benchmark Studies

To identify the key questions and organize the roundtables, the organizing committee drew heavily from the 2022 NSF/NIST report.⁹ This report highlights seven technical and business integration constraints that are inhibiting industry collaboration, data sharing, and the use of SM and AI/ML/DT across all industries and all sizes of firms, including the start of small demonstration projects. The largest barrier to robust and installed AI/ML/DT applications is the availability and access to enough of the right data. Additionally, manufacturers have reported that as much as 70 % of the cost of a single implementation can be in the preparation of data needed to build an application. There is a need to scale this investment.

More fundamentally, AI/ML/DT thrives on data. An application constructed from data from a single machine in a single factory is likely less robust and is more limited in use than if there can be data aggregated from other similar operations to broaden the operational range of the data used to build the application. Broader operational behaviors built into an application greatly increase the coverage of the application and increase the likelihood that there can be a prediction by association with a local machine. The effect of aggregating data is to build in and take advantage of the operational experiences across many machines. The ability to aggregate data for greater application robustness goes to improved performance.

Another key argument centered on the ability to use multiple solutions together and at higher enterprise level. Too often today, single applications are built only with the data from the operation to which the application is applied. The result is a limited application since the data cannot be used for anything else, and most of the cost (i.e., about 70 %) for preparing the data is not usable for the next application. If AI/ML/DT solutions are to be proliferated and orchestrated, a focused attention and investment on collecting, preparing, and scaling data for multi-application building is needed. There is a critical need for extending and reusing data for other applications and lowering the cost of data preparation per application.

If these needs could be addressed by a CDE, the expectation was that costs for all members would be substantially reduced, while data availability would be substantially increased. There are also large benefits to sharing data preparation resources, know-how, and methods.

⁹ See <https://www.nist.gov/publications/towards-resilient-manufacturing-ecosystems-through-artificial-intelligence-symposium>

Based on the NIST 2022 Report and industry experience, the CDE believed that successful demonstrations of financial benefits would require collaborating on a single, high value use case for initial implementation, agreements to share the necessary data, and participation of representatives from industry, academia, and government. With an interest in near term benefits, the CDE also premised the effort on existing technology that is available to support initial adoption of SM and AI/ML/DT in existing SMM semiconductor fabs today. R&D is a future pathway for continuing developments and improvements.

B.3. The Workshop Roundtables

The Workshop consisted of a series of five roundtables, targeted discussions with many participants before and after the roundtables, and weekly organizing committee member meetings. Benchmark studies with industrial data and the use case sponsored by Seagate Technology (see **Appendix C**) were used to shape roundtable questions and conduct iterative reviews of findings with participating manufacturing and data science experts. The industrial use case from Seagate made it possible to examine and evaluate the Workshop findings with today's tools and consider challenges with industry data with much needed granularity.

With reference to **Figure 2-1**, Roundtables 1 through 3 were devoted to setting up the CDE, agreeing on the use case, deciding on a minimum viable infrastructure, and addressing key execution requirements. In parallel, benchmarking was set up by going through the steps of building the ML model solution using individual and combined datasets from five etch machines. A baseline of methods was established for processing the data and model performance metrics and baselines for each of the five machines. Roundtables 4 and 5 then focused on data sharing processes as a CDE. The benchmarking focused on testing/demonstrating key aspects of shared data processing and collaborative model building. Each of the five Roundtables challenged conventional approaches. The sequence of targeted questions for the Roundtables was organized to address technical, business risks, and challenges to clear a path of critical elements for the industry CDE to embrace all aspects of data sharing. The specific subject, organization, and format of each roundtable was not fully specified in advance, however. Rather, the organizing committee determined the target questions for each Roundtable based on the questions and discussions from the prior Roundtable and many pre-post Roundtable discussions. The following subsections contain summaries of how the Roundtables proceeded:

B.3.1. Semiconductor Industry Use Case Roundtables 1 and 2 on Oct. 10 and 24, 2023: The Business Case for Increasing Productivity of Metrology

Semiconductor industry representatives were divided into two separate roundtables. These Roundtables established the organizing principles that facilitated broader industry development and adoption of business practices, technologies, existing tools, and collaborative methods. Each roundtable discussed various operations in semiconductor

wafer manufacturing where SM and AI/ML/DT could be applied to achieve a measurable financial benefit. Rank-ordered priorities for improvements in productivity and quality in existing facilities were: (1) predictive maintenance, (2) advanced process control, (3) fault detection and control, (4) root cause analysis, (5) virtual metrology, (6) automated inspection, (7) energy management, and (8) materials management. Since metrology can represent up to 40 % of the total manufacturing cost in a semiconductor manufacturing operation, both roundtable groups focused on this problem area as a strategic area for improvement. Specifically, the use of etch machine data to predict and control metrology outcomes along with automated root cause analysis were selected as the high value use cases to demonstrate the potential impact of SM and AI/ML/DT in semiconductor manufacturing. A business case briefing document was developed to quantify the financial value of focusing on metrology by describing the forms in which financial gains were achieved.

B.3.2. Data Sharing Infrastructure Roundtable 3 on Nov. 14, 2023: Minimum Viable Infrastructure

To prepare AI-Ready data, a standardized infrastructure is needed for systematic collection, contextualization, formatting, and qualifying machine tool and metrology data at each site and for pooling data and supporting consistently applied data and modeling processes. There was also a need to support several upfront Workshop constraints: (1) understanding the value of data by staying separate from vendor service models for now; (2) using existing technology to understand the value of data sharing that can be done today; (3) avoiding delays with technology development or R&D; and (4) being able to openly understand all aspects of data processing. Building on the foundations defined in Roundtables 1 and 2, participants assessed the approach, infrastructure, and SM Profile™ structure for data information models developed by CESMII, as an independent Manufacturing USA national institute. The fact that the infrastructure was usable now and vendor agnostic, and that CESMII was a non-profit industry-driven partnership, also played heavily in moving forward with CESMII SM Profiles™ as the base infrastructure for collecting data and enabling the processing of data for the Workshop study. Collecting input from targeted commercial providers and several data scientists was also sought to ensure the viability of the approach.

The CDE agreed that CESMII's open infrastructure specification and methods were workable and represented the minimal viable infrastructure necessary for application of AI/ML/DT systems for the wafer thickness and flatness use case. CESMII's platform ecosystem includes an SM Innovation Platform™ (SMIP) that consumes SM Profiles for collecting and contextualizing data and interfacing with applications that further process and use the data for operational solutions through a standard API. SM Profiles commit to the kinds of data types that are contextualized as bottom-up based on individual machines or unit operations. This standardized structure and approach to an SM Profile is a key enabler for data processing consistency, reusability, and interoperability. The CESMII ecosystem further supports an SM Profile Designer™ and SM Profile Library™. Building the first SM Profile requires the "lift" of

industry involvement. Once developed and agreed to, the SM Profile can then be used to drive cross-company resolution and become a common, accepted SM Profile. SM Profiles provide data in structured information models to facilitate selection, movement, and sharing of data across factories and companies.

B.3.3. AI Ready Data Roundtable 4 on Feb. 28, 2024: Individual Factory and CDE Responsibilities as Coalition Governed Workflows

This Roundtable was a critical in-person meeting in which participants were brought together to examine the use case in detail, and for the industry and academic participants to level set on the problem being tackled. Of note, the data for the use case was available in detail and could be discussed at levels of granularity which allowed type, condition, processing, IP sensitivity, and anonymization impacts to be considered in detail. Participants reviewed these factors in the context of the wafer flatness use case as well as their own experiences with other industry use cases. The building and use of CESMII SM Profiles™ was also discussed in depth. Detailed discussions also focused on sourcing data from actual factory operations, how to separate data that must be protected as IP from data that can be shared, how to avoid loss of information, what data processing methods should or could be used for the ML metrology model, and defining a collaborative workflow for contributing data to the CDE.

The focus on data from actual factory operations was essential because vast amounts of data are required to develop robust AI models. Research laboratories simply do not have access to enough data to be effective. Every company independently ranked their specific operational datasets as open, company confidential, or collaborative confidential. The roundtable participants analyzed the results and reached agreement on a set of data that could be shared. This was an industry first and a critical milestone that provided a starting point for a subset of companies to meet regularly after the roundtable to determine how the data would be standardized, profiled, and sanitized to begin work on the wafer flatness use case. Participants reviewed an initial collaborative workflow for contributing data to the CDE effort, jointly processing the data into AI-Ready data, using jointly developed datasets to build a collaborative ML model, and implementing and validating the model. This initial workflow was subsequently refined and benchmarked.

B.3.4. Data Application Roundtable 5 on Dec. 9, 2024

Roundtable 5 reviewed the benchmark results for collaboratively processing data and aggregating it for more robust ML model building. Sharing operational data demonstrated numerous, quantifiable economic benefits to individual factory and company operations that are not possible when data is compartmentalized in silos. The Roundtable zeroed in on reviewing a procedure for building a common data information model that could be used across factories and companies for consistent site contextualization and categorization of

data. Importantly, the Roundtable focused on how to quantify the value of data with respect to contribution, processing, use, and time contributed.

Appendix C. Use Case Specifics

The Workshop focused on a virtual metrology application in which operational measurements from an etch machine tool are used to predict whether wafer thickness will pass or fail when measured in a subsequent metrology step. This application was used to investigate the benefits of aggregating data across multiple machines to build more robust AI/ML/DT models and to benchmark a collaborative AI-Ready data preparation, ML model building, and implementation workflow.

Below is a summary of the key aspects of the use case and the benchmark workflow:

1. **Figure C-1** is a sketch of a general ion beam etch (IBE) machine showing the key functions of a typical machine:

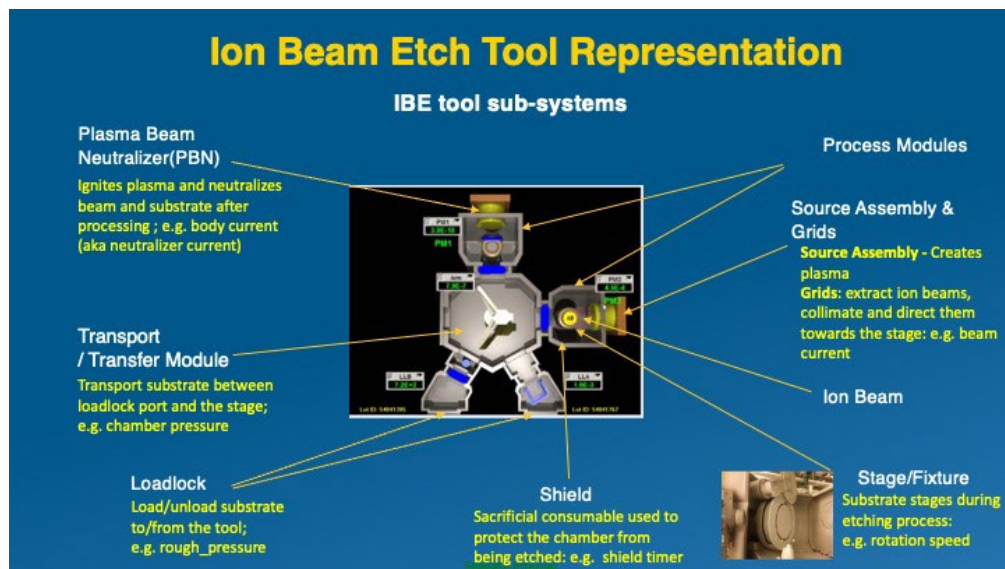


Figure C-1. Schematic of a Typical Etch Tool

2. Seagate and UCLA constructed a Data Information Model based on CESMII's SM Profile™ structure. The data information model and how it was constructed is illustrated in **Figure C-2** as an Entity Relationship Diagram (ERD). The SM Profile organizes the data measurements by the tag name and units that are relevant to wafer thickness. The experience is that an ERD maps directly into an SM Profile.

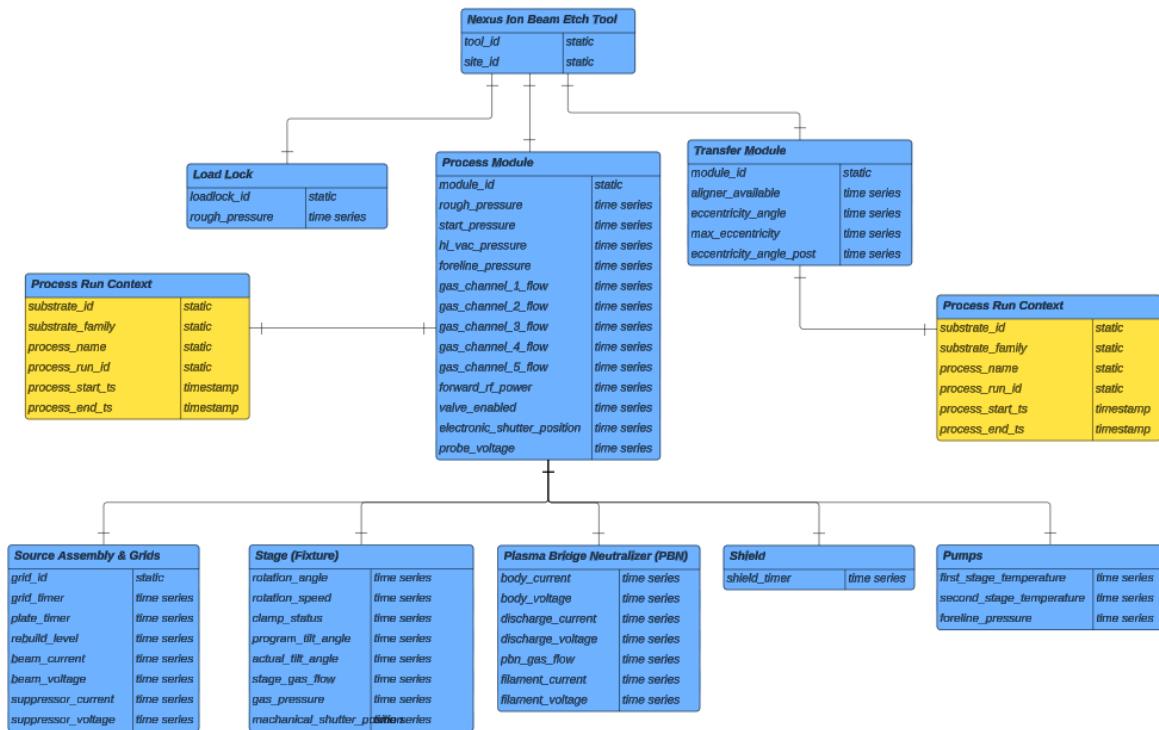


Figure C-2. Etch Tool Information Model: Ion Beam Subsystems and Processing Context

3. Based on prior experience with this metrology classification problem:

- a. Raw data are collected at one second intervals from start to stop of an etch machine batch run for the Process function.
- b. An etch machine batch run is a line operation step that can be used in different manufacturing processing, material, and product contexts.
- c. The raw data collected at one second intervals is qualified and then averaged over the duration of a batch run to form a “batch” dataset to be used to learn Pass/Fail (P/F) outcomes over many batch runs. A batch dataset is categorized by machine line operation step, product, and material. The virtual metrology objective is to predict P/F from the etch operation measurements and use expensive and time-consuming metrology more productively.

4. A typical batch run dataset is illustrated in **Figure C-3** below:

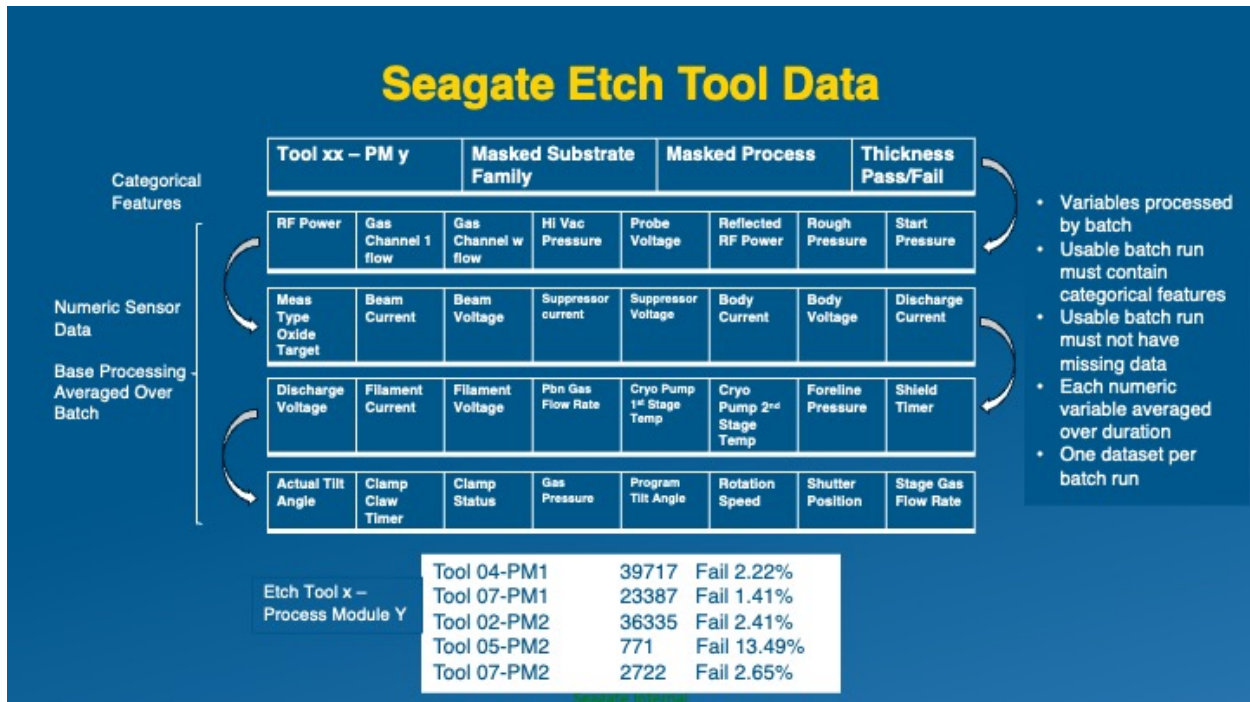


Figure C-3. The Elements of an Etch Machine Dataset – Single Row 37 Features and Variables

As shown in **Figure C-3**, there are four categorical variables and 33 etch tool numeric data measurements with each averaged over the duration of the batch step. The categorical features include:

- Tool xx – PM yy refers to etch tools in different line operations and/or factories. In this use case, there are four distinct etch tools. i.e., xx = 02, 04, 05, or 07. Some of the etch tools have multiple processing capabilities called Process Modules (PMs). For tool 07, data were collected from both modules.
- Effectively, data are collected from five distinct tools allowing the study to consider tools in different line operations and one tool in the same line but with parallel processing.
- Masked Substrate Family refers to the wafer materials involved. Specific substrates were masked in name to protect IP.
- Masked Process refers to the specific process step in the manufacture of the product. Process steps were masked in name only to protect IP. Operationally there are about 70 different processing stages and about 180 products.
- Thickness Pass/Fail captures the Pass/Fail metrology determination.

The data shown in the table forms a single dataset for a single batch run. As noted, there are over 100,000 batch datasets across the five machines and two modules over two years of operation.

Three machines/modules have substantial amounts of data and two do not. Across the five machines, variations in tools, processes, substrates, and operations provide a range of data situations with which to study when and how to aggregate data. There are five process categories offering aggregation potential:

- Same machine, same material, and same operational purpose for the same process step.
- Different machine groups, same material, and same operational purpose for the same process step.
- Same machine, different materials, and different operational purposes in similar, repeated, and subsequent process steps.
- Different machine groups, different materials, and different operational purposes on different process steps.
- Processes involving different machine groups, different materials, and same operational purpose configured using time-based or sensor-based completion detection.

Details concerning the condition of the data were also noted:

- Data from Seagate etch tools (4 tools from the same family), covering roughly 2 years of wafer manufacturing data; about 100k batch runs, with a failure rate between 1 % and 3 %.
- The Pass/Fail classification was provided by Seagate. It is a mapping from an aggregation of six different types of measurements, which are floating point measurements of thickness taken by the metrology tools at different locations on the wafer.
- For the etch machine in which the wafer is processed, 33 tool sensors most relevant to the etching process were used.
- Eight of these 33 sensor readings had missing values with missing rates ranging from 0.03 % to 70 %.

Appendix D. Data Processing Use Case Baseline

While **Appendix C** describes the baseline ML model use case, **Appendix D** describes the baseline data processing. Given the baseline use case and process together, we set out to build a source-to-AI-Ready data process that maximized the value of data sharing by maximizing the economic value points with data processing, data aggregation, data usage, model implementation, and model performance when acting together with appropriate governance.

Figure D-1 is the baseline workflow for data processing that generally tracks bottom-to-top and then left-to-right. In the lower-left-hand corner, the starting point is the four Seagate etch machines in two factories. The first green box (*Step 1*) shows that Seagate had previously collected and stored raw data in the two different factories for each etch machine. *Step 2* reflects the etch machine CESMII SM Profile™ (see **Figure C-2**) that was developed by Seagate engineers and data scientists and used to collect, re-organize, and contextualize the data from individual etch machines. The data from multiple machines across both factories (light blue boxes) were concatenated by machine, forming a raw data machine inventory of the batch run time series data for each variable (*Step 3*). The data for each variable for each machine was then individually averaged across each batch run and contextualized to form an inventory of etch machine batch run data (*Step 4*). These two inventories are shown with heavy black outlines to indicate these multisite/multiline shared inventories. The outcome of *Step 4* is shown in the white box which summarizes the specifics of all the datasets collected (Tool 05-PM2 was left in the list but not used).

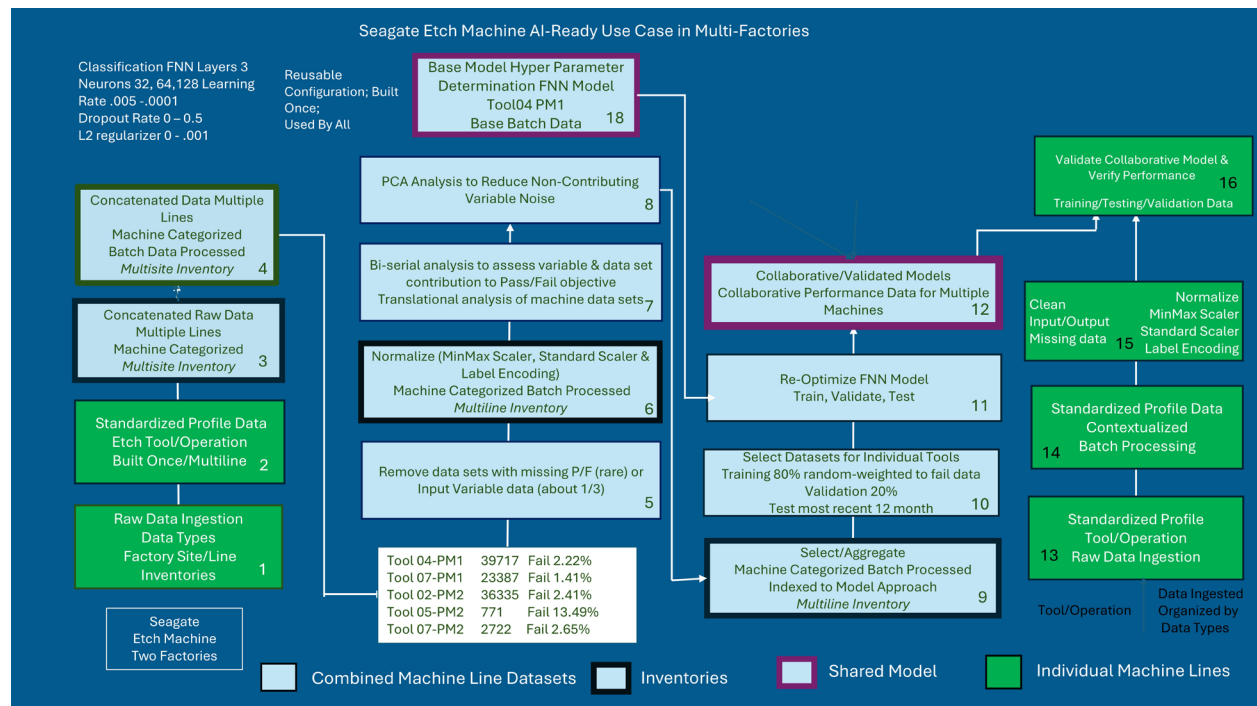


Figure D-1. Workflow of AI-Ready Steps Used for the Etch Machine Benchmark

With the data from all machines consistently collected, contextualized, and pooled, we enter the next phase of preparing AI-Ready data. *Step 5* is about qualifying the data by removing any datasets that do not have an associated P/F metrology metric or substrate feature (rare occurrences). It was, however, necessary to deal with missing data from the machine measurements with some missing data in about 1/3 of the datasets. In this case, the decision was made to simply use the dataset (setting the missing value to 0) so that the ML model would be trained on missing data situations. Normalization (*Step 6*) followed next, which for the use case involved applying a standard scalar and processing the feature data into a usable numeric encoding for ML modeling. Please note that because the data are being processed within a single company, protection of the data for sharing was not needed.

The result of qualifying and normalizing produced a third multi-site shared dataset in *Step 6* (heavy outline) processed for batch-over-batch classification analysis and modeling. The commitment of this inventory to a classification modeling approach came from an early analysis and decision about the nature of the ML model needed for this application. This decision is shown as the light blue box at the top with a heavy purple outline. For this use case, the decision on classification mapped into the use of a deep learning Feedforward Neural Network (FNN) as shown. This is a decision in which the commitment to the FNN was applied to all machines, i.e., a shared decision on a reusable ML model configuration that was built once and used by all lines.

Steps 7 and 8 are important data science methods for assessing the data for noise, contribution, and aggregation impacts on the P/F prediction. For this use case, several assessments were conducted that included biserial analysis of individual variables, principal component analysis (PCA) to reduce non-contributing variables, statistical match, and translation shifts of datasets across machines. The consistent applications of these analyses enabled an optimized selection of aggregated, cross-machine datasets for building a cross-machine ML model shown in *Step 9*. *Step 10* is where the aggregated datasets for each tool are “selected” for training, testing, and validation. This includes any weighting of datasets, i.e., weighted toward fail data, applying methods for randomly selecting training data, and decisions about test and validation data.

Steps 10 and 11 are closely linked because of the tight coupling of learning (training) and validating but separated to bring out the importance of distinguishing between the two processing steps, since validation can produce a need to re-optimize the FNN model, i.e., via transfer learning. *Step 12* is shown with a heavy purple outline to indicate that collaboratively developed ML models with shared, carefully aggregated (where applicable) multimachine datasets are available.

The green boxes on the far right of **Figure D-1** show the implementation steps for a particular etch machine on site. In *Step 13*, the common profile is used to collect, ingest, and contextualize the time series data for the etch machine in its line operation, and in *Step 14* the data are averaged over a batch run. The data are cleaned and normalized as before in *Step 15*. *Step 16*

fine tunes and validates the collaboratively built ML model. If performance measurements are collected in a consistent way, then comparisons of cross-machine performances are possible.

Appendix E. Analysis Details on Benchmark Findings

E.1. Sustaining Data Processing Consistency in Operational Use

As noted in the Executive Summary, numerous benefits to the individual factories within and across companies acting collaboratively were documented. Collectively, these benefits supported key execution pathways for a CDE to meaningfully share data to advantage in multiple ways. The findings from acting as an ecosystem are summarized in **Table 1** (see **Section 2.2**). Findings are based on comparing data processing and model building across five etch machine tools in three factories within Seagate with the baseline of addressing the same data process and modeling steps for each machine independently. **Appendix C** and **Appendix D** provide, respectively, specific detail on the Etch Machine use case, how baselines were established, and the steps taken to process the data used to build an ML classification model for predicting wafer flatness, pass/fail, from etch machine operating data.

From a benchmarking standpoint, the Workshop approach was to develop a baseline set of data processing steps after first considering a standard Feedforward Neural Network (FNN) as a possible ML algorithm for the Pass/Fail classification problem. Workshop participants generally concurred with the data processing methods and steps from a standard problem standpoint, but participants also discussed many ways in which the data for this problem could be approached and how to address operations that change and shift over time. From a model building standpoint, there was considerable discussion about federated ML model training approaches in which data from individual sites could be accessed, site by site, to train the model versus pulling datasets together for various forms of training with pooled datasets. There was also a great deal of discussion about how more open sharing of qualified data (beyond company and coalition confidential) would enable the development, testing, and implementation of data processing methods and model building approaches in more direct support of manufacturing. Notably, there was considerable discussion about reference variables that underlie or group outward facing measured variables in ways that are less sensitive to operational changes. Finally, there was also significant discussion about the pitfalls with processing data, e.g., the pitfalls with noise, bias that could be introduced with different methods for addressing missing data, the tradeoffs with reducing dimensions, the levels of granularity with which to analyze data, the loss of information with masking and anonymization, and the problems of working with poor quality data.

Although generally appreciated by the Workshop participants, what became collectively very clear was: (1) how important the processing of data was to the success of models, (2) there are many ways to process data, (3) there are always going to be new methods and new algorithms, and (4) data processing needs to be consistent to be scaled, whether sharing data to build algorithms or federating access as another form of shared data. Consistency with data processing is essential for trust in data. Categorizing data with respect to sources and processing is also key to data sharing. Renewed awareness of these facts with benchmarked quantification collectively led to a strong Workshop focus on data processing consistency and sustaining it for operational

use. The focus of benchmarking was on achieving multi-data source consistency and showing how benefits accrued from this consistency for factories and companies. There was no focus on a specific method, set of steps, order of workflow, or algorithm, although the workflow tested was reviewed as reasonable.

Benchmarking proceeded with a baseline set of existing, often used methods for data processing and algorithm building. The focus was on how to achieve consistency with one set of methods for one application. This lens of consistency enabled the Workshop to identify and quantify value points when working as a collaborative. It also led to carefully distinguishing what had to be done at a site from what could be done collaboratively. This required benchmarking on how to achieve consistent data contextualization from multiple distributed machines across factories and across companies to avoid each site introducing inconsistencies at the data sources.

Benchmarking made it possible to see execution pathways for scaling the data for other AI/ML/DT applications, for use at larger operational scales, and for broader industry-wide benefit, while maintaining individual manufacturer benefit. Of note, it proved to be exceptionally important to learn from, but not get delayed by, different vendor approaches, business and service models, platforms, and infrastructure so that data processing and algorithm building for benchmarking could be performed unencumbered. This also proved to be a key for individual manufacturers to understand the value of their own data in the context of their operational interests and from the perspective of contributing and using data across a much wider range of data sharing possibilities.

Appendix D details the methods, steps, and workflow baseline used to build the ML metrology model and are shown graphically in **Figure D-1**. **Figure E-1** summarizes the analysis of value points with ecosystem data sharing and an analysis of the results of benchmarking the workflow to achieve maximum data processing consistency. While drawn from the use case, the workflow in **Figure E-1** is described below functionally and not in terms of specific methods:

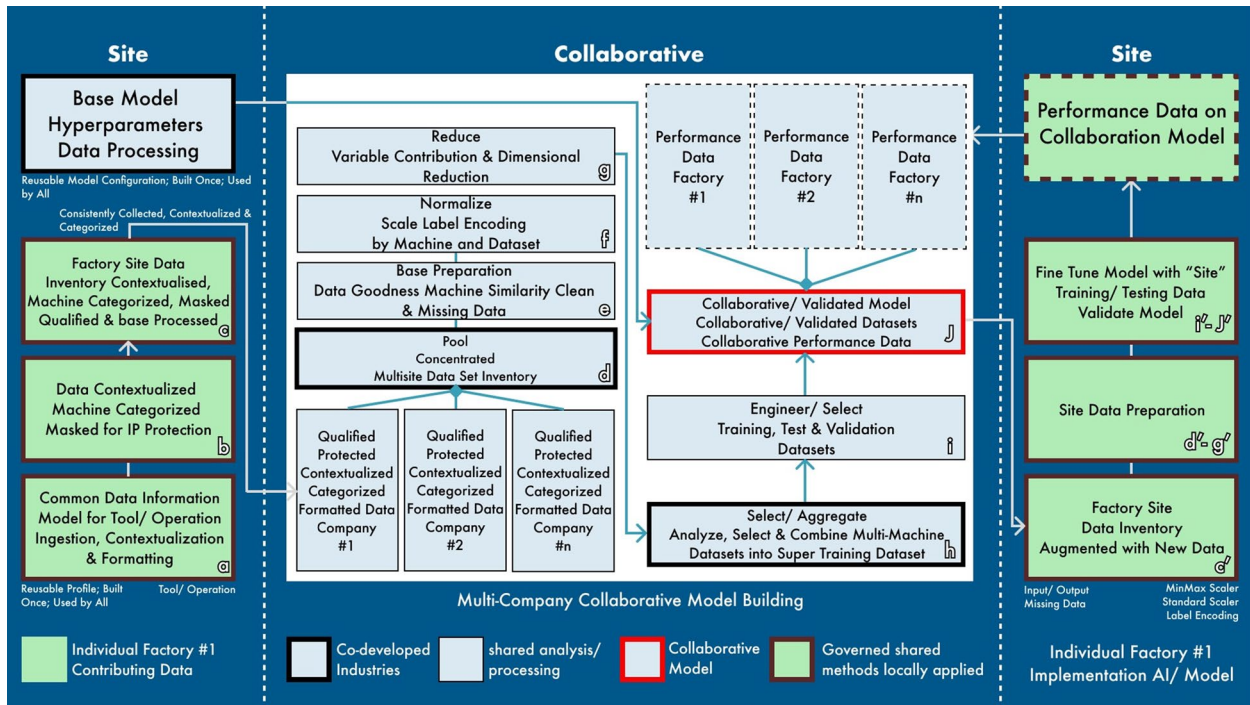


Figure E-1. Analysis of the Benchmark Workflow for Data Processing Consistency

Step a: This ecosystem process is foundationally enabled by a common data information model for the machines and objectives of interest, i.e., each machine and metrology quality measurements. For there to be consistent data processing, data from the relevant tools need to be collected, contextualized, and formatted consistently. As discussed, the CESMII SM Profile™ was used as the IT structure with which to capture and use common data models. As a value point, the common data model needs to be jointly developed and is therefore itself a shared outcome that is applied at each individual site. Note that for model building there is typically more than one data information model in use, and data models from different machines of the manufacturing operation may be used. *The light green filled boxes in Figure E-1 indicate site actions. The heavy brown outline indicates a governed, shared method that was applied at the machine site.*

Step b: The application of common data models ensures that wherever they are used, and there is agreement, the data will be ingested, consistently contextualized, and consistently formatted for any next action with that data. Importantly, the data information model proved to have two additional advantages. A data information model for a machine or operation is, by definition, a dataset that is categorized by machine. It was necessary, however, to further build in categorizations for products, recipes, and materials since these needed to be included in the ML model. The explicitness of the data information model with this type of configuration information provided critical categorizations that can be used to search for specific datasets. This ability was key to additional insights about machines, materials, and products from which data have been pooled.

Additionally, in this benchmarking exercise, companies were willing to openly share machine data, but their product, recipe, and material data were considered important IP and needed to be masked. The data information model proved to be the best location and process step for masking this IP. Product, recipe, and material categorizations were each given code names in the data model that only the manufacturer could translate. For this data, this form of masking was considered adequate protection within the confidentiality bounds of the CDE. The form of masking also made it possible for there to be coalition confidential analyses on categories. *The filled boxes in light green indicate site actions. The heavy brown outline indicates a governed, shared method that was applied at the machine site.*

Step c: A particularly important step to the success of a model is site qualification of the data. Qualification refers to making sure the data is “good” data for the application and model building by considering several questions. Was the machine running within normal expectations? Were the sensors working and were the measurements within expected ranges? Did something change with the product recipe, materials, or product specifications? Is there any base processing of the data at the site that needs to be accounted for to understand the data? “Good” data is required for any use or application, regardless of working only at a site level or as an ecosystem. However, ecosystem consistency requires consistent qualification by all participating sites. **During data processing there are checks on the ‘goodness’ of the data, but in this benchmarking exercise, and not surprisingly, site inspection of data proved to be of critical value.** This led to future considerations about engaging data and operational engineers to develop consistent inspection methods that are extended to others with training and automated inspection tools. *Step c* recognizes that in a factory context, data is now stored on site, or at a company, or in a contracted cloud location. Data from multiple machines are now consistently collected, contextualized, and formatted, and ready to be concatenated (i.e., pooled) and stored together. Among multiple factories and companies in an ecosystem, these steps form consistently qualified, categorized, contextualized, and formatted data for factory and company data inventories across factories in the ecosystem. *This is shown in **Figure E-1** with the three company boxes in the lower left-hand corner of the large white area feeding into box d.*

Step d: While the first steps are necessarily site-specific tasks, optimal sharing is indicated with the steps in the large white box in the center of **Figure E-1**. As shown, datasets from each factory site, now consistently qualified, protected, contextualized, categorized, and formatted, are concatenated into a multi-site inventory, an important shared data resource for the CDE. *This box is outlined in black because this is a co-developed and qualified industry dataset inventory and resource.*

Steps e, f, and g: These are data preparation steps that are applied to machine datasets in succession to achieve consistency for all site datasets for similar machines, similar functions, and similar model objectives. As a practical matter, it proved to be exceedingly

difficult to achieve data consistency across datasets from multiple sources if any of the data preparation steps were conducted in a distributed manner. The shared concatenated multi-site dataset inventory in *Step d* became a key shared data inventory. As discussed below, this shared inventory was also needed for implementation of the collaborative ML metrology model at a particular site.

Step g: This step crosses over between data preparation and data engineering for an application objective. It is important to reduce the data to only those data that contribute to the modeling objective. The extent contributing data reflects an objective determines model performance. Removing non-contributing data significantly helps with reducing unwanted noise.

Step h: This step addresses the analysis, selection, and aggregation of datasets from multiple, similar machines to assemble super training sets for building a collaboratively developed common ML model. As discussed below, this study benchmarked the value of these common ML models and found they performed better for all machines than as individually developed models for each machine. This is a potentially high value step and among the primary motivations for data aggregation.

Steps i and j: These steps address the selection and further engineering of training, test, and validation datasets for building a collaborative super model for a given application objective for all machines. *Box j is outlined in red to indicate shareable, collaboratively developed models from shared datasets.*

Base Model: Building a collaborative, data-centered model requires a decision on the model structure and algorithm to address the application objectives. The base model reflects the type of problem that will be using the data (e.g., classification, regression, or DT), the general data to be applied (e.g., streaming data or batch data), and general sense of the data in an operational context (e.g., normal shifts in operations, product changes, or failure rates). It was typically useful to try one or more algorithms to get a feel for the range of the hyperparameters. *This upfront aspect of model building is shown in the upper left-hand corner of **Figure E-1**. It is filled in light blue and outlined in black because establishing the base model is a shared outcome (i.e., the model can be developed once and used by all).*

Please see the far right-hand side of **Figure E-1** for site implementation of the collaborative ML metrology model. This workflow was confirmed by Seagate Technology staff who evaluated model performance internally:

Steps c' to g': Site data from a particular factory and machine can be accessed from the multi-site dataset inventory in *Step c*. Importantly, this data is augmented with new site data. This site-specific dataset is then prepared with the same steps and methods used in *Steps d to g*.

Steps i' and j' : Site data is used to fine tune, test, and validate the collaborative ML metrology model with site data.

Finally, in **Figure E-1**, note the dashed-line light green box and three dashed-line light blue boxes in the upper right corner. These boxes are intended to project the potential of closing the loop with shared data going back to individual factory sites. If the performance of each collaboration model at each implementation site could be measured consistently, collaborative performance results could be used to continuously improve the model as qualified, protected, contextualized, categorized, and formatted data continues to flow in from CDE sites. Also, analysis of collaborative model performance could be readily related to data subset categories (e.g., machine, material, recipe, and product) to provide new insights. This closed-loop process is a cross-factory/cross-company **Shared Data Value Appreciation Cycle**.

E.2. Benchmark Data Processing Consistency Economics

To estimate the cost-saving benefit from generating consistently processed datasets as a CDE, Seagate's cost experience with data engineers and data scientists was used as a baseline and proxy for cross-company sharing. Staffing costs for processing data consistently across five sites together were benchmarked. After data is qualified as "good" data at each site, pooled processing could be done in multi-site machine application groupings with 1.5 full-time equivalent (FTE) data engineers/scientists for all sites versus 5.5 FTE if each site processed data independently. This is a 3x cost avoidance in staff, which translates into approximately \$1 million in total cost avoidance over a year for multiple applications or about \$200,000 per site. This is also a 4 FTE avoidance in additional headcount to process data for five machines. While not benchmarked quantitatively, there was a reasonable indication that a single data engineer/scientist had the capacity to process data for more than five machines for one application and/or take on additional applications or machines. Furthermore, this staff utilization potential is a conservative estimate because it did not account for the additional data engineer/scientist effort that would have been required to address the difficult task of data consistency reconciliation using pooling data from multiple sites after the data had been independently processed.

This benchmarking effort demonstrated that consistently processed datasets can be generated cost effectively using existing staff if there is willingness to share data processing. Cost avoidance was in the form of scaled, qualified data availability without increased staffing. The benefits of available, consistently processed data are readily realized at each of the five machine sites. This cost avoidance on staff effort and headcount for data processing provided significant justification for data processing within a CDE. However, the staffing investment is still a means to an end. The operational advantages documented with readily available, qualified, and scaled data that is consistently processed drives the overall economic benefit. It is clear, though, that acting as a CDE not only makes it possible to have qualified, scaled data, but also that the qualification and

processing of data can be done better with little increase in staff compared to acting independently.

Based on this benchmarking effort, the industry was quick to recognize that substantially greater economic value was available from cross machine/cross company data processing consistency that is otherwise difficult to achieve when done independently for a given solution. Data processing at an individual production site is still needed to perform data collection, contextualization, qualification, and formatting. Overall, source-to-application consistency required driving data processing actions at individual sites as close to zero as possible while maximizing the processing of the pooled data from all sites. In this benchmark exercise, site actions were reduced to the single manual task of qualifying data. Qualifying the data involved removing any etch machine tool data collected when the machine tool and/or its sensors were not operating within “normal expectations.” In the future, data qualification could be automated and would benefit from doing so, but it was not done in this Workshop. Consistent collection, contextualization, and formatting did require an agreed upon data information model that was jointly developed. A data information model in the CESMII sense ensures data is collected, contextualized, and formatted. It also ensures that categorical features are consistently accounted for to describe the operational context (i.e., material, product, and metrology). Once a shared data information model was applied, the resulting categorized datasets could then be pooled (in multiple ways) and brought together for combined processing and engineering. These consistently qualified and processed cross-machine datasets are far more valuable than independently processed data in the following ways:

1. Ready for cross-machine, cross-operation, cross-factory comparative analysis and optimization.
2. Ready for data aggregation and collaborative AI model building, **as benchmarked in the Workshop.**
3. Availability of data that is reusable for multiple applications.
4. Capable of opening untapped opportunities for cross-operational insights about materials, recipes, and products, and available to seek new models and solutions with more open sharing.
5. Qualified and categorized datasets that can be discovered and selected.
6. Opens the door for assimilating new data processing technologies as co-developed methods with cross-operation validation.

E.3. Benchmark Outcomes with Aggregating Data for Model Robustness

With reference to opportunity #2 in the list above and *Step h* in **Figure E-1**, benchmarking specifically considered the potential for aggregating data from multiple machines into aggregated datasets that would lead to more robust model performance for all machines. Both classification and regression AI/ML models were considered. The Seagate/UCLA benchmark team has

published the details of this study.^{10,11} Many combinations of machine data were considered and compared to individual machine performances when trained only with data from the machine. The study makes clear that aggregating data across machines to build a model for all machines is beneficial when the aggregating opportunities are carefully assessed with respect to potential contributions to operational coverage, variation, and similarity in operating signature. Aggregating data from multiple machines operating too differently from each other or for different model objectives should not be done naively.

Figure E-2 shows two examples (of many) of improvement with data aggregation for the wafer flatness pass/fail classification model. To benchmark performance for this application that is focused on fails, there are four possible outcomes for this classifier model: 1) an actual pass is classified as a pass (true positive), 2) an actual pass is classified as a fail (false negative), 3) an actual fail is classified as a fail (true negative), and 4) an actual fail is classified as a pass (false positive). The true positive and true negative outcomes are trivially good since the classifier model is correct. The false positive, while not ideal, can be addressed with downstream manufacturing procedures. Similarly, if all fails are reevaluated at the metrology machine and manually measured to determine whether they are true fails, then the false negatives will be caught and correctly reclassified as passes. The main impact of incorrect classification of a problem wafer is that as this wafer moves through the process, it wastes resources and time. Thus, a high-performing classification model for this metrology application will have a low false positive rate (FPR) relative to a true positive rate (TPR). The papers in the footnotes provide detailed descriptions of this metrology use case in predicting a Pass/Fail based on wafer flatness.

The Receiver Operating Characteristic (ROC) analysis offers a robust evaluation method by examining the TPR and FPR of the model's predictions on test data at different classification thresholds as shown in **Figure E-2**. The model performance can be evaluated using the Area Under the Curve (AUC) score, where the AUC score is defined as the area under the ROC curve when plotting TPR on the y-axis and FPR on the x-axis across a range of reasonable thresholds. Ideally, a perfect model would achieve a TPR of 100 % (max sensitivity) and an FPR of 0 % (zero false alarms), resulting in an AUC score of 1. Conversely, a model that makes random guesses would have a TPR that equals an FPR of 50 %, resulting in a diagonal ROC line ($y = x$) and an AUC score of 0.5. The ROC to AUC score, therefore, provides a comprehensive measure of model performance across all possible thresholds, making it particularly useful and widely applied for evaluating models on imbalanced datasets.

¹⁰ Ou, F., H. Wang, C. Zhang, M. Tom, S. Bom, J. F. Davis and P. D. Christofides, "Industrial Data-Driven Machine Learning Soft Sensing for Optimal Operation of Etching Tools," *Dig. Chem. Eng.*, **13**, 100195, 2024.

¹¹ Ou, F., J. Suherman, C. Zhang, H. Wang, C. Zhang, S. Bom, J. F. Davis and P. D. Christofides, "Industrial Multi-Machine Data Aggregation, AI-Ready Data Preparation, and Machine Learning for Virtual Metrology in Semiconductor Wafer and Slider Production," *Dig. Chem. Eng.*, **15**, 100242, 2025.

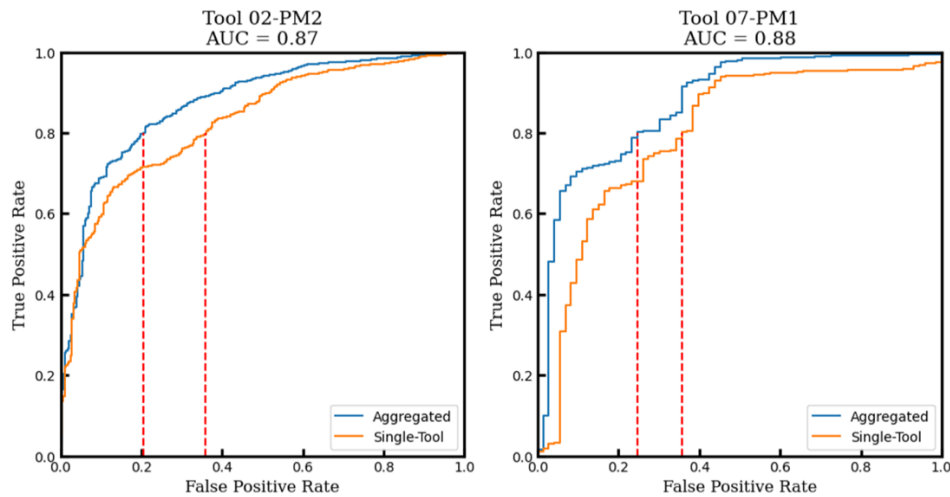


Figure E-2. Receiver Operating Characteristic Curve for Two Tools

For this use case, an FPR in the vicinity of 0.2 for at least a TPR of 0.8 was considered acceptable. As shown in **Figure E-2**, the FPR performance of Tool 02-PM2 improved from 0.38 to 0.2 for a TPR of 0.8 with aggregated data. Similarly, the FPR for Tool 07-PM1 improved from 0.37 to 0.24. Please note that this study focused on benchmarking the improvements. Final performance was not optimized and has subsequently been shown to perform much better when optimized for factory use.

E.4. Benchmarking Site Contextualization, Qualification, Categorization, and Formatting

There are multiple ways to build data information models (e.g., CESMII SM Profiles™ in this study is one approach), but to achieve consistent contextualization at distributed sites, the industry participants in the CDE needed to agree on a common approach for information modeling. The Workshop benchmarked the use of CESMII SM Profiles™ as existing, standards-based, IT structures for capturing a data information model that can be descriptively separate from the source of the data and how the data is used. It is digitally consumed as a template that recognizes several different data types (e.g., streaming, floating point, integer, or nameplate). Once filled in with the operational context of the data, the SM Profile provides instructions on how to ingest, contextualize, and format the data. While the CESMII SM Profile™ template is agnostic to the basis on which a data model is constructed (e.g., structural, behavioral, functional, causal, fault, or workflow), cross-industry development with many partners has converged on a machine or unit operation base level of structural granularity. This granularity tends to reflect components and sensor measurements that are tightly integrated. It also tends to align with how people naturally think in terms of units that comprise physical systems. The CESMII SM Profile™ is a key element in a comprehensive specification that: (1) brings together a format for data upon

ingestion, (2) uses the Profile to define collection and contextualization instructions, and (3) formats the data for application use through a common application programming interface (API). The CESMII SM Profile™ therefore defines a standards-based approach for contextualization and formatting data across machines with a common profile. An example of an SM Profile for an etch machine based on an Entity Relation Diagram can be found in **Figure C-2**.

Benchmarking focused on the ability to build a common SM Profile™ across five machines in three factory locations and across individual machines in three additional companies. Building SM Profiles is a factory site process that requires capturing the function, structure, and behavior of a machine and naming how each machine is used in the context of a manufacturing operation. Every site had described and named the machine elements and measurements differently. Each site entered the process of building a common SM Profile with concerns about sharing data. The process of building the SM Profile had to engage the site's data scientists and operations engineers and overcome this trust issue.

The CDE cross factory/cross company process involved multiple iterations that each participating company undertook independently and then shared. The process was a facilitated process involving site data scientists and machine tool/process engineers converging on functional descriptions, naming common functions, naming sensors, naming categorizations, and agreeing on data types and units. The exercise demonstrated that common profiles are achievable. Benchmarking showed the process becomes faster and easier with each iteration. There were several additional notable outcomes:

1. CESMII's focus on machine and unit operation level SM Profile's proved to be the appropriate level of granularity for building factory-focused SM Profiles for this semiconductor application. This recognition provided a line of sight to building line operation (i.e., multiple machine profiles) by linking machine profiles.
2. Site experts recognized that different data types required different levels of protection (i.e., machine data needed to be treated differently than data about materials, products, and recipes). They used the naming convention requirement for an SM Profile to mask those data that were considered IP.
3. The site experts involved in building and sustaining a common profile were the same people involved in qualifying data. The act of building an SM Profile informed how the data would be qualified.
4. The organizing committee recognized the value of the categorizations afforded by the SM Profile. These included machine type, brand and model, company confidential designations of materials and processing, and machine data types that were built into the ML model. All these features could be searchable given consistent formatting of the data.

With respect to benchmarking, the CESMII SM Profile™ demonstrated an enabling capability for formatting, contextualizing, protecting, qualifying, and categorizing data to build a searchable data inventory that could be consistently prepared and engineered as shared AI-Ready data.

E.5. Benchmarking Operating Model and Governance

Figure E-1 emphasizes a Data Sharing life cycle operational model for a CDE. Roundtable discussions focused on shared inventories, shared analysis, shared collaboratively developed models, and the need for governed methods to achieve consistency. The projected results define a CDE governance structure in which site-enabled data contribution, use validation, and performance are consistently prepared and reported, and a collaboratively developed data inventory is sustained for CDE benefit. Data, methods, models, technology, and workflows can be evaluated, updated, and improved to be better, faster, and cheaper as a CDE recognizes that consistency is a major value opportunity. It was also clear that collaboratively-developed training and change management become shared benefits whose value increases with shared activity.

Full cycle collaboration with meaningful cost sharing, data sharing, and scaled use of data still requires an understanding of how this would be done and a demonstration that it can be done. A demonstration entails delineating value points, executing on the shared usability of a collaborative data and model building workflow cycle, and documenting that costs are reduced and benefits amplified by implementing a collaborative Data-First strategy. Such a sustained workflow cycle needs to align cost share values with AI-Ready data inventories, collaborative model building, and training resources that are consistent, governed, scaled for flexibility, and continuously updated as “living” performance-driven resources.

As shown in **Figure E-3** below, value in data sharing comes from contributing data across multiple applications. Some applications benefit from greater contributions than others. Value to CDE participants also comes from access to the multiple applications and from the consistent methods and tools used together. Data is therefore valued from both contributing and using perspectives. A CDE is therefore expected to operate from a richer affinity base when there are multiple applications, contributors, and multiple users leading to a richer, multi-faceted sharing potential.

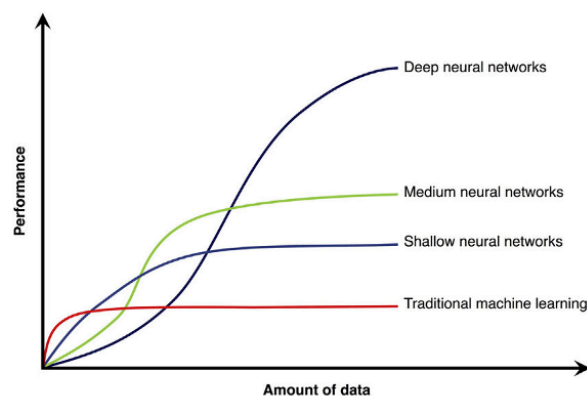


Figure E-3. Projected Data Sharing Performance Requirements with Different Kinds of Applications